**Analysis on U.S. Online Shopping Behavior under COVID in relation to Consumer Behavior Factors**

**Abstract**

This study investigates the influences of COVID and other consumer behaviors on the online shopping behavior in the U.S. market. I will first identify if COVID’s presence significantly increased more comparative online shopping behavior (i.e., more people increasing online shopping activities). The second question is to identify what other consumer behavior factors can influence more comparative online shopping behavior despite the impact of COVID and financial fluctuation. This study used data from a national survey of 2 million consumers responses from Prosper Insight & Analytics, and linear regression model as the primary statistical analysis tool to answer the two research questions. The major findings of this study are that COVID significantly increased comparative online shopping. In addition, consumer’s carpooling, overall spending, and the employment behavior seem to show strong association with more comparative online shopping.

**Introduction**

Over the past decade, e-commerce has been growing substantially in its industry (Ho et al., 2007). Hence, it is becoming more important for companies and firms to understand what makes e-commerce more favorable to consumers and what particular type of consumers are likely to increase their online purchases compared to previous time periods (comparative online shopping). Understanding consumers better will greatly contribute to e-commerce firms’ future operation and marketing strategy. To aid in company planning/strategy, it is important to investigate what consumer behaviors are associated with people’s decision to shop more online. In addition to the consumer behavior factors, the recent COVID pandemic seems to also add another level of complexities in affecting people’s shopping preference (Baker, Farrokhnia, Meyer et al., 2020). There are currently many reports on e-commerce sales growth during the pandemic. However, there are lack of published statistical analysis on the behavioral analysis of online shopping and how other consumer behaviors might have contributed to the overall online shopping activities since COVID’s emergence. Therefore, the two main foci of this paper are (1) analyzing COVID’s impact on more comparative online shopping behavior and (2) identify other behavior factors that are associated with an increase in comparative online shopping behavior. By answering the question about the pandemic and consumer behaviors’ influence on more comparative online shopping trend, this paper fills in the current gap in knowledge of behavioral analysis on online shopping. For clarification, more comparative online shopping behavior in this paper represents the percentage of people whose response was doing more online shopping activities than the previous month. For specific variable description, please reference Table 5 in Appendix.

Overall, this paper will first discuss about the historical trend of e-commerce and then dive deeper into recent academic articles’ discussion about the impact of COVID on online shopping to point out the current gap in knowledge. Given the gap in knowledge in consumer’s behavioral analysis, this paper will move forward to answer two questions. 1) Did COVID significantly change people’s online shopping behavior on more comparative online shopping? 2)What consumer behavior factors significantly influenced more comparative online shopping and what were the corresponding impacts? The overall research will be based on using the merged data of monthly consumer behavior survey, COVID statistics, and financial volatility to investigate my research questions. Specific data collection and data cleaning details will be provided in the methods section. Finally, this paper will be discussing the findings of how COVID or consumer behavior factors can influence more comparative online shopping and provide useful model for predicting more comparative online shopping in future analysis.

**Literature Review**

With the advantage of being able to compare price and quality of a product online before purchasing, e-commerce has been growing over the past couple decades (Ho et al., 2007). Additionally, under the COVID situation, the guideline of social distancing for many countries forces people to shop even less in the physical store (Baker, Farrokhnia, Meyer et al., 2020). Alternatively, it is incentivizing people to shop more online (Stanciu et al., 2020). Among the different sectors, the U.S. market is showing a huge decrease in economic activities that require close contact (e.g., restaurant and retail shopping) and a huge increase in online shopping for both nonperishable goods and grocery products (Richards & Rickard, 2020). To better understand the complex relationship between e-commerce and the current pandemic, I have separated this literature reviews into small sections.

First, I will briefly discuss the growth in e-commerce during the past decade or two and what has contributed to the growth in e-commerce. Then, with the pre-existing online shopping trend, I will examine how did the COVID affects people's lives and their consumption style preferences. Next, there will be sections about previous studies’ discussion on factors that impact consumption style (online or physical) in the general sense and how it matches with COVID’s impact on the people to create that growth in online shopping across different countries. Finally, I will identify the current research gap and discuss the reasoning for conducting research on COVID and consumer behavior factor’s influence on comparative online shopping trend.

***Growth in E-Commerce or Online Shopping***

In the beginning of the 21st century, there is already an indication that online shopping is becoming more popular among young people and the overall population in the U.S as studies compare the online shopping trend among different countries and demographics (Comegys et al., 2006). Even in terms of global impact, online shopping expenditures among European countries also showed a clear increase from 2000 to 2004 (Ho et al., 2007). Moreover, the reason behind such a fast-growing e-commerce market in the early 21st century around the world is mainly due to increase in access to the internet and the power of computerization (Ho et al., 2007). With higher accessibility to internet and electronic device, in online shopping does seems to be a global trend over the past decades. However, the trend is not exactly the same across different countries as one study shows Norway's 2004 e-commerce expenditures per capita (measured in US dollars) is roughly $355.79 and Greece had only $1.75 in e-commerce expenditures per capita in the year of 2004 (Comegys et al., 2006). This shows that different countries responded differently to online shopping. So, individual country’s online shopping trends will be an interesting topic for future investigation. Therefore, my study only focuses on analyzing U.S. online shopping behavior instead of the complex global e-commerce trend.

As online shopping shows clear growth during the early 21th century, e-commerce seems to continually expand even during the COVID pandemic. Before the pandemic hit, there are studies showing that e-commerce began to be more efficient and popular in the recent years as firms work toward more robots in the warehouse and taking on roles such as transportation and packing (Bogue, 2016). Amazon's purchase of Whole Foods in 2017 is also showing the expansion of online shopping in different retail sectors such as the grocery sector (Richards & Rickard, 2020). During the pandemic, online shopping in general shows even more growth. Recent studies report that in the U.S., Target e-commerce sales rose 141% every month in the first quarter of 2020, and that online grocery sales were expected to increase 40% by the end of 2020. In comparison, comparable in-store sales at U.S. Target locations increased less than 1% (Chang & Meyerhoefer, 2020). Given the expectation of the report, it is interesting that COVID seems to have pushed the growth of e-commerce and reduced the physical retail sales. According to the report, the growth of online shopping is likely to continue as well. Therefore, I will start the discussion on the reason why COVID has pushed the growth of e-commerce in the following sections.

***COVID’s Behavioral and Economic Impact Transformed Shopping Preferences***

COVID doesn’t directly change consumption patterns, but it creates uncertainties and rules that affect the way people interact with one another which further change one’s consumption behavior. Recent studies examined the government’s policy in terms of COVID response, which the biggest impact is the socio-economic impact instead of destruction in material assets (Hall et al., 2020). To be more specific, the border control, mobility restriction, and social-distancing affect how people live their daily lives (Hall et al. 2020). In Romania, there was a relatively strict measure to stop COVID in terms of social distancing, which the policy leads to more than 80% of people rarely leaving their home and forces everyone to adapt to online grocery shopping (Stanciu at al., 2020). In the U.S., there has been a sharp drop on the use of restaurants, retail, air travel and public transport in mid to late March which is the time when the virus began to spread quickly (Baker, Farrokhnia, Meyer et al., 2020). The shelter-in-place orders create an environment where online shopping is a good choice for people in terms avoiding contact with others (Baker, Farrokhnia, Meyer et al., 2020). Given that pandemic policy also changes people’s day to day behavior, it shapes my research interest to assess how those daily behaviors might have changed people’s behavior on online shopping aside from the health and economic panic.

Other than government policy, the pandemic also creates varieties of concern and uncertainties for the overall economy and households. In the U.S., the uncertainties for the economy are mostly in three fields: stock market volatility, newspaper-based economic uncertainty, and business expectation. The volatility index for S&P 500 increased by 500% from January to March 2020, words related to economic panic appear more often in newspapers, and business expectation survey on future sales uncertain values hit a peak in March 2020 (Baker, Davis, Terry, et al., 2020). Given that health and economic panic is present under COVID, this is why my research design needs to control for those panics in order to find the true association between online shopping behavior and other consumer behaviors. Moving forward, uncertainties in the economy created the consumer behavior of stock-piling, in which people stocked nonperishable foods in the beginning of the pandemic (Richards & Rickard, 2020). In-store grocery shopping increased instantly due to stock-piling, but the consumer turned to more consistent online shopping afterwards. Another interesting finding about consumer behavior is that there is not much reduction for credit card usage during a time where people are trying to save more than usual (Baker, Farrokhnia, Meyer et al., 2020). Credit card usage is also an important consumer behavior that makes online shopping a more preferable shopping method.

***Online Consumption Theories and How COVID’s Consumption Trend Match Theories***

With COVID’s social distancing restriction, consumers seem to have a shift of preference to shop online. So, what is the fundamental mechanism behind the preference of online shopping? There are multiple researches and studies done before and during COVID to discuss the factors that influence online shopping preference. One of the studies indicates that the availability of online payment infrastructure capabilities is likely to accelerate online transaction-making in e-commerce (Comegys et al., 2006). This means that credit card and online grocery shopping websites are essential to the online shopping trends. Because of that, I purposefully include consumer behavior of owning credit card in my own research for investigation. Moving forward, another research uses a logistic model and found that purchasing method, time window, minimum order requirement, and purchasing fees change the likelihood of people shopping online (Grashuis et al., 2020). Therefore, consumer’s online shopping behavior can be impacted by online infrastructures and the cost of buying online. In addition, studies focusing on consumer’s behavior in terms of confidence discover that risks associated with online shopping such as financial risks and product risks can impact consumers’ attitude for online shopping (Moshrefjavdi et al., 2012). In this case, people evaluate risks that are associated with online shopping.

The consumption behavior under COVID on a large scale echoed with these previous findings on factors that affect online consumption. In the case for New Zealand, there is a sharp drop in grocery and liquor sales after the initial stock-piling action and turn to online grocery shopping, which reflects the consumers’ fears of being in proximity to others (Hall et al., 2020). Since people tend to evaluate risk when they make purchases, the risk for close contact overwhelms the risk for online purchases and hence lead to more online shopping. As I mentioned earlier, one of the studies points out that there is not much reduction for credit card usage during a time where people are trying to save more than usual (Baker, Farrokhnia, Meyer et al., 2020). As credit cards can be used for online shopping’s payment method, it is convenient for people to pay online there is no much reduction for credit card usage during a time where people are trying to save more than usual (Baker, Farrokhnia, Meyer et al., 2020). Overall, there is an increase in e-commerce sales across the world. Canada changed from 1.5% of grocery shopping done online to 9% by the third week of the pandemic (Richards & Rickard, 2020). In the case of Taiwan, its online grocery shopping app (Ubox) also experienced an increase in online grocery sales (Chang & Meyerhoefer, 2020). For New Zealand, online shopping sales volumes exceeded the same period of the previous year by up to 25% during the pandemic (Hall et al., 2020). Going beyond the previous model in explaining the decision to shop online, some studies also believe that people around the world will adapt to online grocery shopping long-term, and online grocery sales are expected to reach 30% of total grocery sales by 2025 (Richard & Richards, 2020).

***Summary & Knowledge Gap***

To sum up the findings, I first learn that there is a clear growth in e-commerce across different countries starting from the early 21th century. E-commerce continues to grow with the support of technology advancement. In this current special situation of world pandemic, online shopping is experiencing a boost in their sales performance. The boost in online shopping is likely due to economic uncertainties and social distancing policies where people change their shopping method in order to avoid risk of infection. Finally, the growth in e-commerce is expected to continue in the future as the pandemic provides the opportunity for people to adapt to online shopping.

Among the current studies on how COVID impacts online shopping, there is not much research about how online shopping behavior in the U.S. market. Moreover, there hasn't been much research that incorporates financial volatility, consumer behavior, and COVID statistics to identify a model that measures how consumer behavior factors such as eating more at home and using coupons can influence online shopping behavior, controlling for the impact of health or financial panics. Additionally, there is overwhelming research on the revenue and sales of e-commerce due to COVID’s impact, but there is not much published analysis on the behavioral aspect of what percentage people are doing more comparative online shopping solely due to the impact of pandemic. Therefore, this research will focus on analyzing online shopping behavior in the U.S. market. The main objective to identify COVID’s influence and other consumer behaviors’ influence on the percentage of people conducting more online shopping activities will fill in the current gap of knowledge of consumer behavioral study. Understanding and being able to predict more comparative online shopping behavior will add another perspective for assessing the sustainability of growth in e-commerce in the future (in both scenarios: for if pandemic to ends or if it doesn’t end). The broader impact of this research is to be able to predict comparative online shopping behavior using a consumer behavior survey to allow business analysts to manage supplies of online purchasing products more efficiently, according to consumer demand.

**Method**

Given my research goal is to identify how did the pandemic influence more comparative online shopping and what other consumer behaviors are influencing more comparative online shopping, there are two parts of analysis. The first part of my analysis is to identify if COVID’s presence did significantly impact the online shopping behavior by having a higher percentage of people that responded with more comparative shopping online. In this part of the analysis, I specifically disregard other factors first and focus only on whether the pandemic itself has a significant impact on online shopping. The second part of my research is about identifying other consumer behavior factors that can influence more comparative online shopping behavior despite the impact of COVID and the financial fluctuation (health and economic panic during the pandemic). For this part of the analysis, the goal is to control for pandemic statistics and financial fluctuations variables in my model and see what consumer behavior factors still remain strongly associated with more comparative online shopping behavior. This is not a causation analysis because online shopping as one of the consumer behaviors often affects other consumer behaviors as well.

***Data Cleaning***

Before proceeding to the analysis approach, I want to first provide my method for retrieving and cleaning the data. First, I have retrieved a consumer behavior dataset from Prosper Insights & Analytics and gained permission from the firm to use the dataset and share the results (Prosper Insights & Analytics, 2020). The consumer behavior dataset is a monthly time-series from January 2010 to September 2020, which I will transform by having rows in the data as all the months from January 2010 to September 2020 and each column as the survey questions (n=834). Specifically, each survey question is asking about a particular consumer behavior or shopping preference and the corresponding values represent the percentage of people who responded “yes” to that consumer behavior. Having my consumer behavior dataset in a proper format, I then retrieved and aggregated COVID statistics and volatility index (measure of financial fluctuation) to the consumer behavior dataset as the final dataset for model analysis.

For COVID statistics, I downloaded the national summary data from “The COVID Tracking Project” published by The Atlantic Monthly Group (The COVID Tracking Project, 2020). The dataset contained COVID statistics in a monthly time series format that matches my consumer behavior dataset and it also has high credibility with the source of data being the state level health authorities. For data cleaning, I kept the cumulative confirmed cases, deaths, recovered, hospitalization, and tests on the first date of each month for it to match the monthly row identity of my consumer behavior dataset to merge the two datasets together. In addition, I generated variables such as the percent of positive tests (divide total cases by total tests) and a dummy variable for whether or not COVID had occurred during each time period. As a limitation and reminder for my readers, COVID data has a certain level of uncertainty even with the CDC data, which possess potential accuracy issues (The COVID Tracking Project, 2020).

For the volatility index, I downloaded both the daily and monthly average volatility index of the S&P 500 spreadsheet from the yahoo finance website (Yahoo Finance, 2020). As a note, the volatility index ranges from 0 ~ 100, with 0 being no fluctuation of stock price at all and 100 being highly volatile stock price. I selected the first date of each month on the daily volatility index to match the row identity of my consumer behavior dataset. For the monthly average volatility index, its row identity is already in the first date of each month. I merged the volatility index datasets with consumer behavior dataset and the COVID dataset. To handle missing values in the daily volatility index data (n=46), I manually inserted the nearest day’s value (volatility index doesn’t change much in 1 or 2 days). If there were two nearest days (before and after) to the missing value, I inserted the average of the two days for the missing value.

After I have merged the three datasets together, the full dataset contains too many variables for analysis (even more than the observations). Moreover, a lot of variables are either not consumer behaviors or have issues with many missing values. Therefore, variable reduction process was performed prior to model analysis. Since I conducted variable selection before running the model, one of the limitations to be aware of is the possibility that my variable choice could have led to a biased result. The 13 consumer behaviors variables included in the final dataset were date, more public transportation, amazon prime, increased carpooling, more home meal, more comparative online shopping, more comparative mobile shopping, shopping for more sales, spending less overall, traveling less, using coupon more, no credit card, and employed full time. More comparative mobile shopping will not be used in this project as it is a subcategory of online shopping using a mobile app. I included it in my dataset for possible future research on mobile shopping.

Aside from consumer behavior factors in my final dataset, I also have 7 pandemic variables including COVID dummy variables (0 = Non-COVID time, 1 = COVID time), total confirmed cases, deaths, recovered, hospitalizations, tests, and percentage of positive tests. Additionally, I also have the monthly average S&P 500 volatility index variable and daily volatility index on the first date of each month. To view a full variable description table, including how measures were coded, see the Appendix (Table 5). Raw data is available upon request and permission from Prosper Analytics and from the manuscript author, and complete reproducible code and documentation for the data cleaning and analysis is available in GitHub repository from the author (URL available upon request and details in the Appendix).

***Analysis Approach for COVID’s Influence on More Comparative Online Shopping***

The first part of my question is to identify if COVID’s presence significantly impacted the more comparative online shopping behavior. In other words, I want to see if a higher percentage of people responded to doing more online shopping activities during the period where COVID is around. To do that, I first used ggplot2 package (Wickham, 2016) to look at the overall trend of more comparative shopping online. For statistical approach, I used only the dummy variable of COVID in the linear regression to see what is the average difference of more comparative online shopping before/during COVID and whether the COVID dummy variable is significant or not. In this approach, I was verifying whether COVID’s presence is significant to the change of online shopping behavior.

In addition, since the first recorded cases for COVID happened in my data’s observation of March 1, 2020, I want to compare the average of more comparative online shopping percentage between the time period of 7 months post March 2020 (post COVID time) with 7 months prior to March 2020. To do that, I ran the same linear regression model of only using COVID dummy variable as my input and limited my time period into the most recent 14 months in my dataset (August 2019 – September 2020). In this case, I am measuring the impact of COVID on the most recent trend of more comparative online shopping.

***Analysis Approach for Consumer Behavior Factors on More Comparative Online Shopping***

The second part of my research is about how other consumer behavior factors can influence online shopping behavior despite the impact of possible influence of economic and health panic noted in recent studies (Baker et al. 2020). Since previous consumer behavior studies shows that linear models can describe the impacts of input variables in a simple and straightforward way (Koufaris et al., 2001; Moshrefjavadi et al., 2012), I used linear regression model as my main analysis method on identifying what other consumer behaviors can influence online shopping behavior and their corresponding influence.

But before I proceed with the linear regression model, I first performed some initial data exploration to understand my variables better. Since studies have shown that PCA is very effective with identifying variable correlations (Kriegel et al., 2010), I implemented PCA and correlation plot using Program R’s ggfortify and ggcorrplot package (Horikoshi & Tang, 2018; Kassambara, 2016) to see how my variables are correlated with each other. Then, I looked at summary statistics to check for possible outlier issues among my variables.

Moving into my major analysis tool of linear regression model, I first used the StepAIC method from the MASS package in R (Venables & Ripley, 2002) to build a model on estimating more comparative online shopping from my consumer behavior factors, COVID statistics, and stock market volatility index (reference all the variables in Table 5 of Appendix). This method initially helped me identify the significant factors in estimating more comparative online percentage. Then, I went through model validation process (i.e. checking multicollinearity, normality, heteroskedasticity, and serial correlation) to derive the final linear model and report the significant consumer behavior variables that have influenced more comparative online shopping (Schmidt & Finan 2018; Prost, 2008).

To check and resolve for multicollinearity issues, I used VIF (Variance Inflation Factor) test from the car package in R (Fox & Weisberg, 2019). The VIF test produces a value for all input variables in a model and a high VIF value (general threshold of “greater than 5”) means the input variable exhibits multicollinearity with another input variable. To solve the multicollinearity issue, I ran correlation matrix to see which variables were highly correlated. Then, I tested each variable by leaving one of each pair or triplet of highly correlated variables in the model at a time to check the adjusted R2 value. Finally, for each pair or triplet of highly correlated variables, I kept the one variable that produced the highest adjusted R2 value to solve multicollinearity issues in my model while maintaining the best predictive model possible. Repeating the process, I derived my final linear regression model.

For further model validation, possible linear regression model issues such as normality, heteroskedasticity, and auto-correlation were tested through Shapiro-Wilks test, ncvtest, and Durbin-Watson's test using R’s built-in function and the car package (R Core Team, 2020; Fox & Weisberg, 2019). Since there were only 81 observations (after discounting missing values) in the data, I also looked at residual plot and residual histograms to confirm the validity of my model. Lastly, I recorded important statistics about my models such as RMSE and to show how well my model can make the prediction (Ma & Iqbal, 1983).

Aside from the linear regression model, I also realized the possibility of non-linear relationships between consumer behavior factors and more comparative online shopping. Therefore, I also used random forest method’s variable importance function and single tree plot to confirm the important variables (identified from the final linear regression model) on making decisions for predicting more comparative online shopping (Genuer et al., 2010). The random forest model was implemented by the caret package in R (Kuhn, 2008). The single tree model was implemented also using the caret package and tree structure graphics were produced with rpart.plot package (Milborrow, 2020). Since the linear regression model is my main method for finding how different consumer behavior characteristics influence more comparative online shopping, random forest method and the tree model is just to add additional aspect of how other methods are using consumer behavior factors to estimate more comparative online shopping.

***Research Transparency***

For the sake of transparency on my research, I shared final cleaned datasets and the codes for creating/validating my model in my private GitHub Repository (URL available upon request). Note, I did not share the raw data in the GitHub Repository or Appendix to protect intellectual property of Prosper Insights & Analytics. Data is available upon request, given permission by Prosper Insights & Analytics. The appendix section will contain all the visualizations and tables I have created for either data description, data exploration or model validation to make sure my rationale in the paper can be better understood by the readers to further validate my model. Overall, the approach mentioned in this paper is transparent, valid, and duplicable through reading the supplementary material in the Appendix and by requesting access to the data and GitHub Repository. Finally, all the data cleaning and analysis are performed using R programming language (R Core Team, 2020).

**Result**

To recap, my research question is about COVID’s influence on comparative online shopping trends and other consumer behaviors’ interaction with comparative online shopping. The first part of my question is to identify if COVID’s presence significantly impacted the comparative online shopping behavior. In respect to that question, my hypothesis was that a higher percentage of people will report that they have more comparative online shopping activities since the onset of COVID pandemic. The second part of my research is about if other consumer behavior factors can influence comparative online shopping behavior despite the impact of COVID and financial fluctuation. For this part, my goal is to identify truly significant variables in predicting more comparative online shopping and to provide a valid model for discussion of impact.

***Analysis for COVID’s Influence on More Comparative Online Shopping***

***Identify Trend Overtime***

In order to understand how COVID influences comparative online shopping behavior, I need to first look at the more comparative online shopping trend before and during COVID time (Figure 1). For a reminder, more comparative online shopping trend shows the percentage of consumer across time respond to increasing online shopping activity. Therefore, in addition to the well noticed growth of e-commerce, the more comparative online shopping trend will present whether the rate of growth for e-commerce is slowing down or getting faster. More importantly, Figure 1 will use color to differentiate pre-COVID time and time where COVID exist to see how did COVID changed the trend for more comparative online shopping behavior.

**Figure 1: Trend of More Comparative Online Shopping Response**



*Source: U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020)*

*Note: Graphics created using Program R’s ggplot2 packages (Wickham, 2016)*

We can clearly see that fewer people claim to do more comparative online shopping from 2013 to 2019. I also see a clear increase with people claiming more comparative online shopping activities during COVID time (blue). So, COVID does seem to have some influence on making people shop more online. However, the increase in the percentage of people who claim more comparative online shopping activities during COVID time is not enough to support that the pandemic’s situation makes more people doing comparative shopping online activities in comparison to all historical time periods.

***Statistical Analysis***

For statistical analysis, I ran two linear regression models in R programming language using COVID (dummy variable) as my input variable and more comparative online shopping as my dependent variable using two different data samples (R Core Team, 2020). The first model uses full observations and the second model uses the most recent and balanced observations in terms of pre & during COVID time periods (7 months pre-COVID period and 7 month during-COVID period).

**Table 1: Full Data Model Statistics in Estimate More Comparative Online Shopping %**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 0.231254 | 0.004143 | 55.813 | 2e-16 \*\*\* |
| COVID | -0.025606 | 0.014939 | -1.714 | 0.09 . |

**N = 91 F-statistics = 2.938 . Adj-R2 = 0.021**

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

*Source: U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020)*

*Note: Linear Regression Model using Program R’s built in function (R Core Team, 2020)*

**Table 2: Past 14 Months Model Statistics in Estimate More Comparative Online Shopping %**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 0.175041 | 0.003660 | 47.826 | 4.57e-15 \*\*\* |
| COVID | 0.030607 | 0.005176 | 5.913 | 7.10e-05 \*\*\* |

**N = 14 F-statistics = 34.97 \*\*\* Adj-R2 = 0.723**

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

*Source: U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020)*

*Note: Linear Regression Model using Program R’s built in function (R Core Team, 2020)*

Looking at the full data model statistics in Table 1, the p-value of the COVID dummy variable indicates that pandemic’s existence is not significant in interpreting the % of people that report more comparative online shopping activities in terms of the 5% significance level threshold. However, looking at the recent data of 7 months pre-pandemic period and 7 month during-pandemic period, COVID’s existence is highly significant (0.1%) in interpreting the % of people that report more comparative online shopping activities. On average, about 3% more people reported more comparative online shopping during the 7 months of COVID period compared to the 7 months previous on COVID. Responding to the question of COVID’s impact on comparative online shopping behavior, COVID did seem to significantly increase the percentage of people that conduct more comparative online shopping in recent periods. However, the significance of COVID variable in the full data model suggests that such an increase is not significant in explaining the overall trend on the percentage of people doing more comparative online shopping.

***Analysis for Consumer Behavior Factors on More Comparative Online Shopping***

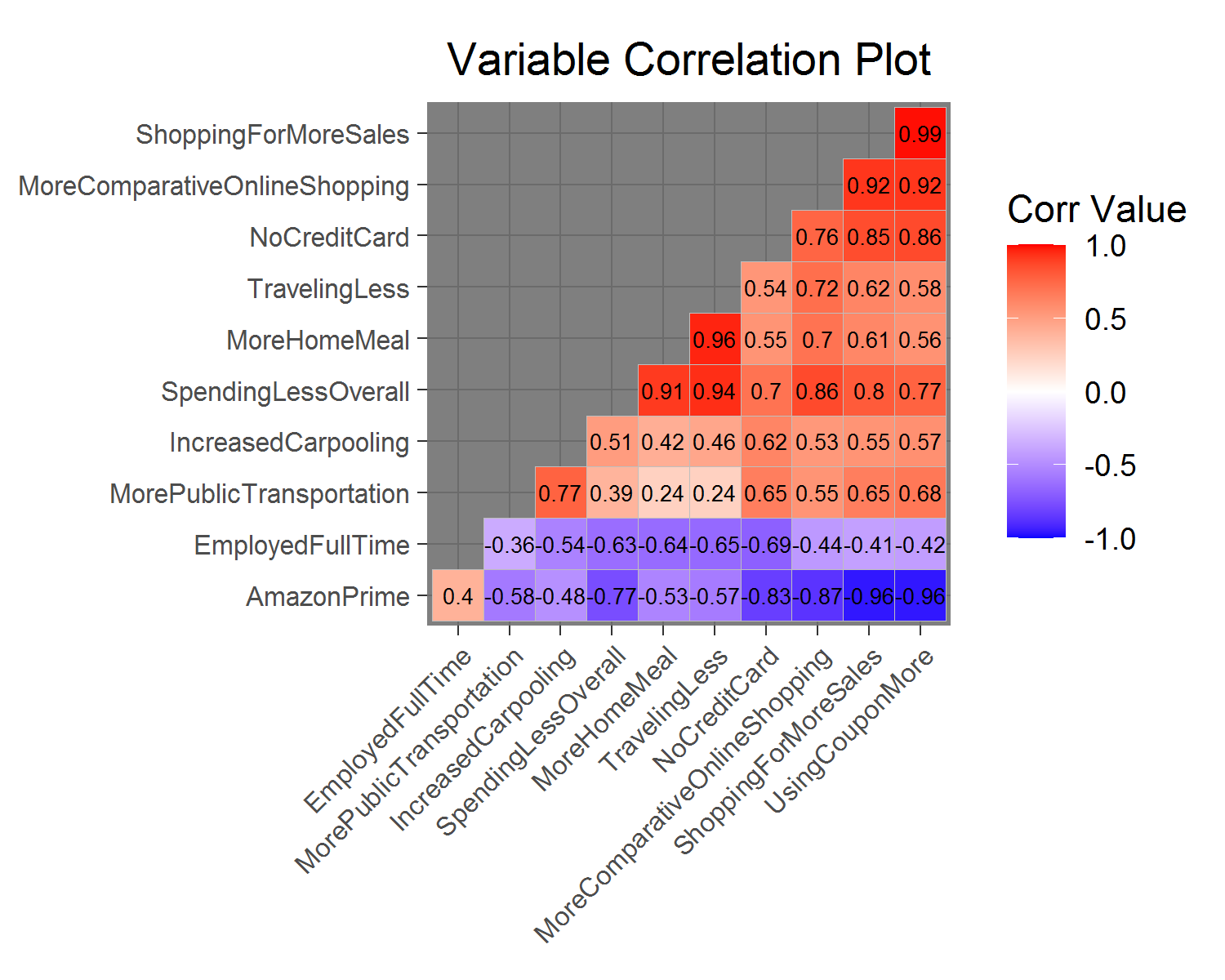
Moving into the analysis on other consumer behaviors’ interaction with more comparative online shopping behavior, the 10-key independent consumer behavior variables for investigate are more public transportation, amazon prime holders, increased carpooling, more home meals, shopping more for sales, spending less overall, traveling less, using coupons more, not owning any credit cards, and full-time employment status. As reminder, those variables are measured as percent of people who responded “yes” to that behavior.

***Early Data Exploration Result***

Before the actual model analysis, I first performed some early data exploration. The goal of performing this early data exploration section is to understand how my independent variables are related with each other (potential multicollinearity issue) and to justify my later model transformation choices. In addition, data exploration such as principal component analysis and correlation plot can also provide me with early insights on what consumer behavioral factors tend to be more associated with more comparative online shopping behavior.

First, I used Program R’sggcorrplot package (Kassambara, 2016) to create a correlation plot on the consumer behavior variables to see how each of those variables are correlated with each other for initial data exploration. Few variables seem to exhibit extremely high correlation (extreme dark color in the correlation plot and correlation value |r| > 0.9) with each other (Figure 2), which might result in multicollinearity issues if my later model were to include those variables. Using coupons more and shopping for more sales variables tend to have strong negative correlation with amazon prime membership. Using coupons more and shopping for more sales are also highly positively correlated with each other. Lastly, traveling less, more home meals, and spending less tend to show extreme positive correlation as well. Out of my independent consumer behavior variables that will be used in my initial model, the variables listed above show strong correlation with each other. Having those variables in my model signifies a possible multicollinearity issue that I must check for in later analysis.

**Figure 2: Consumer Behavior Factors Correlation Plot**



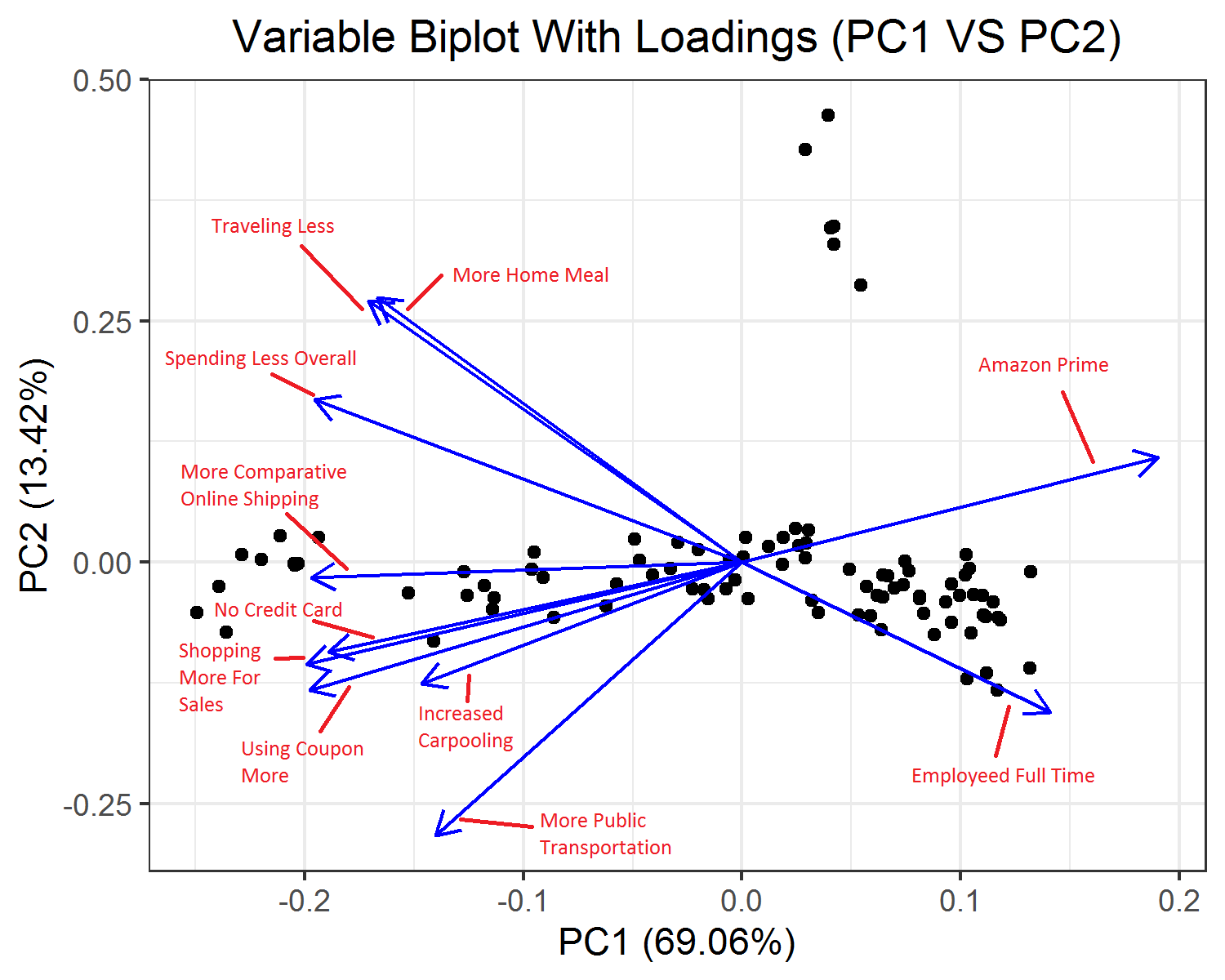
*Source: U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020)*

*Note: Graphics created using Program R’s ggcorrplot packages (Kassambara, 2016)*

In addition, I also put my dependent variable of more comparative online shopping in the correlation plot to have initial investigation on which consumer behaviors are likely to be significant in estimating more comparative online shopping behavior. Looking at correlation plots of Figure 2, it seems like the variable of more comparative online shopping has a strong correlation value (darker color & |r| > 0.8) with variables such as amazon prime membership (r = -0.87), shopping more for sale (r = 0.92), spending less overall (r = 0.86), and using more coupons (r = 0.92). Therefore, these variables are likely to be significant in estimating more comparative online shopping behavior. Specifically, the result indicates a negative correlation between more comparative online shopping and amazon prime membership, while variables such as shopping more for sale, spending less overall, and using more coupons have positive correlations with more comparative online shopping. Further data exploration and validation (e.g., PCA biplots) on these input variables was performed to confirm this early correlation findings.

Next, I run Principal Component Analysis (PCA) on my data using Program R’s ggfortify package (Horikoshi & Tang, 2018). Then I use ggplot2 package (Wickham, 2016) to create auto-plot of the PCA with labels of my variables to see how my variables are related in explaining the variance of my dataset.

**Figure 3: 1st and 2nd PCA Variable Biplot**



*Source: U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020)*

*Note: Graphics created using Program R’s ggplot2 packages (Wickham, 2016)*

The first and second Principal Component Analysis explain 82.48% of the total variance in my data set (Figure 3). Moreover, the variable biplot of first and second PCA confirm my findings with correlation plot, which more home meal and traveling less variable explain a very similar aspect (same direction) in my data (it is probably the reason why they are highly correlated in a positive way). The biplot seems to also take a step forward and consider variables such as no credit card, shopping for more sales, more coupon usage, and increased carpooling to explain similar aspect in my data. In addition, those variables listed above and the spending less overall variable are the closest to the more comparative online shopping characteristic in my data, which signifies that they might be important variables for estimating more comparative online shopping. Moreover, PCA biplot suggests that Amazon Prime variable does seem to explain the opposite aspect of more comparative online shopping, in which it confirms the negative correlation between more comparative online shopping and amazon prime membership in correlation plot.

My early data exploration result helped me identified that as no credit card, shopping for more sales, more coupon usage, and increase carpooling seem to be positively influencing more comparative online shopping, while Amazon Prime variable seems to be negatively influencing more comparative online shopping. Moreover, there are possible multicollinearity issues due to some of my key independent variables being highly correlated, which I need to watch out for when validating my model. Before I began my regression test, I also checked summary statistics of my key independent variables which are my consumer behavior variables (Figure 4, Appendix). No variables seemed to have extreme outliers as max and min in each variable are relatively close to its 1st and 3rd quantile value in the summary statistics. So, I was comfortable in moving forward with my key consumer behavior variables.

***StepAIC Linear Regression Model***

For finding the best model with variables that can estimate more comparative online shopping, I first used stepAIC from the mass package (Venables & Ripley, 2002) on all variables including date, consumer behavior variables, COVID statistic variables, and volatility index variables for my initial attempt to select variables for my model. I included the Date variable because of the natural decreasing trend of more comparative online shopping overtime presented in Figure 1 above, which I want to control for possible serial correlation issue (natural trending of time). The presence of COVID statistics and my volatility index was for StepAIC to pick up possible influences of health and economic panic and serve as control variables for behavioral factors.

**Table 3: StepAIC Regression Model in Estimate More Comparative Online Shopping %**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 5.280e-01 | 1.714e-01 | 3.081 | 0.002831 \*\*\* |
| Spending Less Overall | 4.697e-01 | 8.818e-02 | 5.326 | 8.98e-07 \*\*\* |
| Amazon Prime | 2.587e-01 | 4.835e-02 | 5.350 | 8.15e-07 \*\*\* |
| Death | 7.356e-07 | 1.794e-07 | 4.100 | 9.85e-05 \*\*\* |
| Date | -4.195e-05 | 9.529e-06 | -4.403 | 3.28e-05 \*\*\* |
| Employed Full Time | 2.467e-01 | 7.166e-02 | 3.443 | 0.000919 \*\*\* |
| Shopping for More Sales | 3.293e-01 | 9.091e-02 | 3.623 | 0.000510 \*\*\* |
| Traveling Less | -2.858e-01 | 1.080e-01 | -2.647 | 0.009786 \*\* |
| Increased Carpooling | 4.952e-01 | 1.761e-01 | 2.813 | 0.006174 \*\* |
| Total Test Results | -9.434e-10 | 3.786e-10 | -2.492 | 0.014784 \* |
| More Home Meal | -1.649e-01 | 9.247e-02 | -1.783 | 0.078354 . |

**N = 91 F-statistics: 131.3 \*\*\* Adj-R2 = 0.9354**

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

*Source: U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020)*

*Note: StepAIC Regression Model using Mass package (Venables & Ripley, 2002)*

The StepAIC Regression Model selected the variables spending less overall, amazon prime, death, date, employed full time, shopping for more sales, traveling less, increased carpooling, total test results, and more home meal in the model. The Adj-R2 indicated that my initial model explains 93.54% of the total variance in more comparative online shopping behavior, which signifies a highly explanatory model (Table 3). Moreover, using 5% significance level as threshold, this initial StepAIC model shows that spending less overall, amazon prime, employed full time, shopping for more sales, traveling less, and increased carpooling are significant behavioral factors in estimating more comparative online shopping even when COVID statistics and volatility index variable were involved in the variable selection of this model. Finally, this model includes COVID’s death and total test result variable but no volatility index, which shows that the overall health panic in the society is more important to people’s decision to increase their online shopping behavior.

However, this model can’t be used as my final model for interpretation as it exhibits a possible multicollinearity issue. Some of the variables in the model are highly correlated with each other (Figure 2). Therefore, I checked the model with VIF (variance inflation factor) test (See Appendix Table 5) and kept only the best variable from each highly correlated pair or triplet for estimating more comparative online shopping to use in my next model.

***Final Linear Regression Model Validation***

After the model transformation, my final linear regression model contained date, increase carpooling, spending less overall, employed full time, and death (Table 4). I further validated my final regression model by performing Shapiro-Wilk Normality test, Non-Constant Variance Test (ncvTest), and Durbin-Watson’s test using Program R’s car package and the its built-in function (Fox & Weisberg, 2019; R Core Team, 2020) to check for normality, heteroskedasticity, and serial correlation issue of model. In supplement, I also look at the residual histogram and residual plots to double check on the possible model issues.

The histogram for my model’s residual was normally distributed (Appendix Figure 7). To further check for normality test, the p-value of 0.1966 (>0.1) in my Shapiro-Wilks Test failed to reject the null hypothesis of normal distribution of the residual. The residual plot showed randomness of residual (no specific trend; Appendix Figure 8) which signifies that my model is unlikely to have serial correlation and heteroskedasticity issue. The ncvTest’s p-value for my model is 0.80583 (>0.1), which I fail to reject the null hypothesis of constant variance (homoscedastic). I also use Program R’s embedded Durbin-Watson test to test serial correlation. The p-value is 0.82 (>0.1), which I fail to reject the null hypothesis of no serial correlation. By rejecting all of those null hypotheses, the model validation tests are suggesting that my model is valid because it does not seem to exhibit issues with normality, heteroskedasticity, and serial correlation. VIF test were checked again as well (Figure 6, Appendix). The VIF values are all lower than 5, so my model doesn’t seem to exhibit multicollinearity issues.

***Final Linear Regression Model Result***

**Table 4: Final Linear Model in Estimate More Comparative Online Shopping %**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 5.495e-01 | 8.181e-02 | 6.717 | 1.99e-09 \*\*\* |
| Date | -3.210e-05 | 3.313e-06 | -9.691 | 2.16e-15 \*\*\* |
| Spending Less Overall | 1.801e-01 | 3.375e-02 | 5.337 | 7.72e-07 \*\*\* |
| Employed Full Time | 2.890e-01 | 8.229e-02 | 3.512 | 0.000714 \*\*\* |
| Increased Carpooling | 5.608e-01 | 1.831e-01 | 3.063 | 0.002938 \*\* |
| Death | 1.996e-07 | 6.274e-08 | 3.181 | 0.002050 \*\* |

**N = 91 F-statistics: 168.7 \*\*\* Adj-R2 = 0.9084**

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

*Source: U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020)*

*Note: Linear Regression Model using Program R’s built in function (R Core Team, 2020)*

Based on my validated final regression model, the Adj-R2 = 0.9084 indicates that the model explains 90.85% of the total variance in more comparative online shopping behavior, which is still very robust after the removal of some variables. Moreover, the best combination variables to estimate more comparative online shopping were date, spending less overall, employed full time, increased carpooling, and death (COVID’s total fatality). Now, we can move into analyze how my consumer behavior affect the estimation on more comparative online shopping behavior holding impact of date (time variance) and COVID (health panic). One thing to note before the analysis is that my consumer behavior variable and dependent variable are both measured in percentage decimal (1-unit increase = 100% increase). So, I need to first divide the coefficient by 100 and then discuss the impact in percentage increase. According to my model, for every one percentage point increase in the national percentage of people respond that they increased carpooling behavior, spending less overall behavior, and the full-time employment status, the percentage of people respond to purchasing more online will increased by 0.5608%, 0.1801%, and 0.2890% respectively. Even though the number seems small, if we multiply the percentage by the whole US population, it represents a potentially large economic impact. As those consumer behavior variables are having a large economic impact of comparative online shopping, I will provide further analysis on those consumer behavioral factors in the discussion section later. Finally, my model is pretty good with the prediction as RMSE calculated using Program R’s basic function (R Core Team, 2020) is about 0.01155, which means that on average my prediction is only off the actual % more comparative online shopping by 1.115%.

***Random Forest and Tree Model on estimating***

**Figure 9. Random Forest Model Variable Importance Chart**



*Source: U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020)*

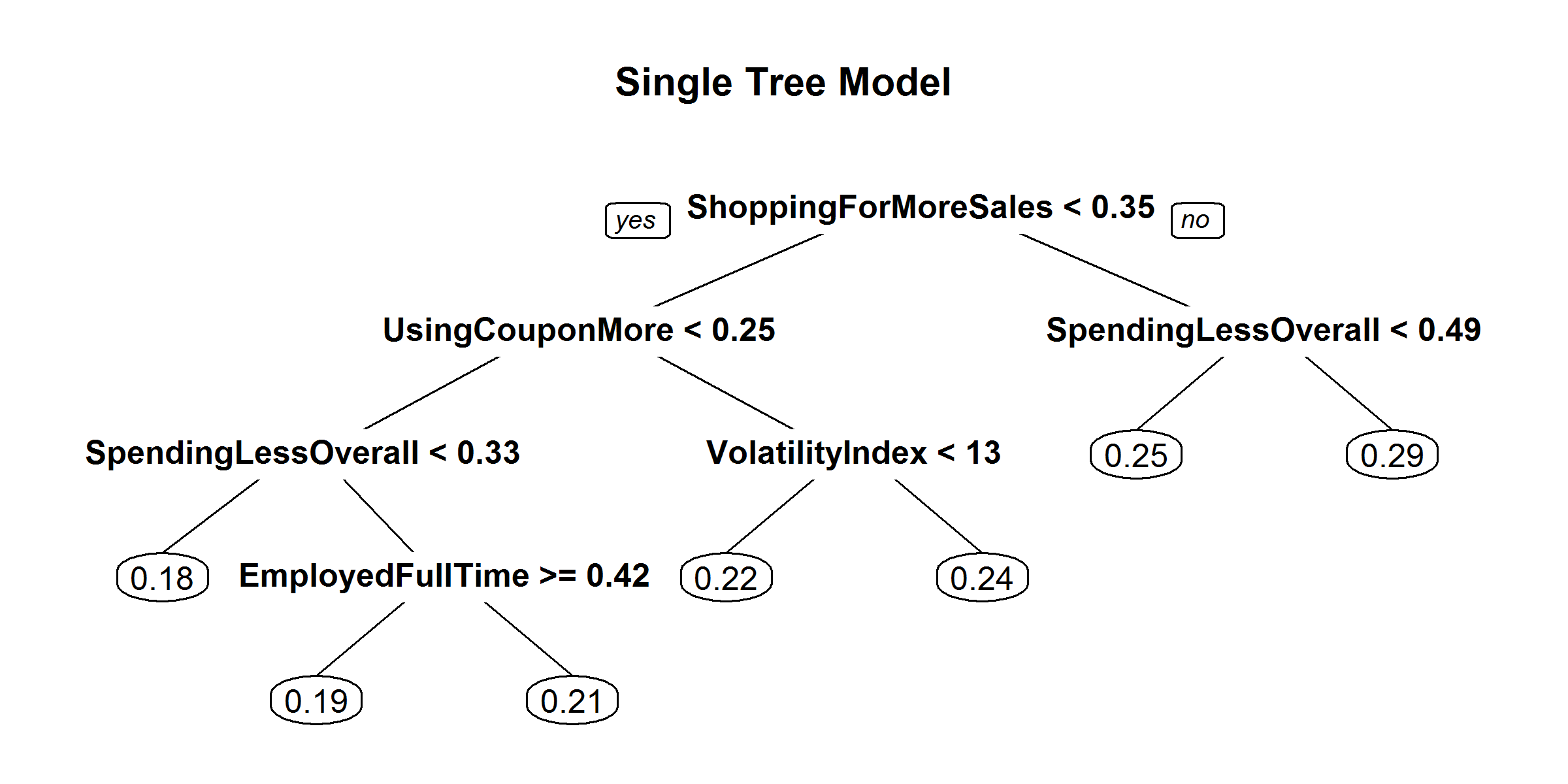
*Note: Graphics created using Program R’s ggplot2 package and caret package(Wickham, 2016; Kuhn, 2008).*

Finally, I will use Random Forest machine learning method, one of the best machine learning methods for prediction (Genuer et al, 2010), to evaluate another model that estimates more comparative online shopping and check to see if the random forest model selects the same variables as I did in the linear model to come up with prediction. I will also look at a single tree model to see how would a typical tree make a prediction on the percentage more comparative online shopping. After I used Program R’s caret package (Kuhn, 2008) to run a random forest model with 2000 trees, I retrieved the random forest variable importance chart (Figure 9). For how I construct my random forest model and retrieve variable importance chart, please reference to statistical analysis code in the Appendix.

According to the random forest model, the RMSE is 0.00578, which means the prediction of random forest for the percentage of people respond to more comparative online shopping is off by 0.578% on average. Comparing RMSE of my random forest model (0.00578) and final linear regression Model (0.01155), Random Forest’s predictive power is better than my linear model, which is as expected because machine learning method are able to capture more variance than simple linear model. However, the RMSE for Random Forest is not extremely better than my linear model, so my linear model is actually also a very good model for primary analysis on more comparative online shopping.

The top five variables that were used most often (scale of 0% to 100%) in estimating percentage of people that responded to more comparative online shopping are using coupon more (100%), shopping for more sales (91%), date (90%), spending less overall (85%), and amazon prime (72%) (Figure 9). Only two of the top five variables were the same as the variables identified in my final linear model. Therefore, random forest seems to pick up other interesting variables in estimating more comparative online shopping. In addition, it is interesting that the random forest model believes “using coupon more” variable is the most important determinant factor for predicting more comparative online shopping behavior, but using coupon more did not appear throughout the process in finding my final best linear regression model. Moving forward, since random forest is basically a more collective method based on many single tree models on different bootstrap samples, it will be also interesting to look at how single tree model make predictions. Figure 10 were created using rpart.plot package (Milborrow, 2020).

**Figure 10. A Single Tree Model for Estimating More Comparative Online Shopping %**



*Source: U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020)*

*Note: Graphics created using Program R’s rpart.plot package and caret package(Milborrow, 2020; Kuhn, 2008).*

The split of tree plot for estimating the percentage of people responding to more comparative online shopping is showing that higher value of shopping more for sale, spending less overall, using coupon more, volatility index, and employed full time will lead to higher percentage of people purchasing more online (Figure 10). Interestingly, I noticed that not all of the top five important variables in the random forest model importance chart were used in this tree model. Therefore, just because certain variables are more important than other variables in making prediction, it doesn’t mean that the top important variables together will make a good prediction. The tree model is utilizing variables that explain different aspects of more comparative online shopping behavior to make a prediction. This is probably why my linear regression model also did not include all top five important variables identified by random forest method.

**Discussion & Conclusion**

**Is COVID Really Influencing Comparative Online Shopping?**

The percentage of people participating in more comparative online shopping is generally decreasing over time (Figure 1). Therefore, although previous research has indicated growth of e-commerce, there is actually less people purchasing more than their regular online purchasing behavior over time. In other words, e-commerce is growing, but the rate of the growth has been slowing as people began to have a more stable online purchasing behavior. This is an interesting finding that adds on to the well-known growth of e-commerce. Given the general decreasing trend of more comparative online shopping, it is interesting that the percentage of people participate in more comparative online shopping did seem to increase during COVID time (Figure 1). Although COVID’s influence is not statistically significant for the overall comparative online shopping trend, it is statistically significant for recent periods. On average, about 3% more people reported more comparative online shopping during the 7 months of pandemic period compared to the 7 months previous on pandemic. Therefore, my findings did support my initial expectation of COVID making people shop online more. The unexpected part is that the increase has been relatively small. This unexpected small increase in more comparative online shopping might be related to some online news article’s idea of “consumers are buying less while online shopping more” (Hartman, 2020).

If COVID’s influence on more comparative online shopping is not that big as of right now, then the next question is whether COVID’s influence on more comparative online shopping will be bigger in the future. I will argue yes to that question from two perspectives. First of all, the increase of % more comparative online shopping hasn’t yet seemed to slow down (Figure 1). In addition, my final linear regression model indicates that COVID’s death count has a statistically significantly positive influence for predicting more comparative online shopping. Therefore, as the pandemic continues and people continue to die from COVID, it will likely influence people to prefer more comparative online shopping. In this case, businesses should prepare for more online shopping activities in the future. This conclusion greatly aligns with the prediction that “Online grocery sales are expected to increase 40% at the end of 2020” from previous studies (Chang & Meyerhoefer, 2020).

**What Consumer Behaviors Influence Comparative Online Shopping and What Does It Implies?**

In addition, my final linear model suggests that the percentage of people that responded to behavior such as increased carpooling, spending less overall, and full-time employed are having positive influence on more comparative online shopping. Among them, increased carpooling (sharing a vehicle) and spending less overall implies people’s action of saving money. Therefore, having those two variables being significant factors for estimating the percentage of people who responded to more comparative online shopping, it actually implies that people will do more comparative online shopping when they are trying to save money. This rationale is confirmed by a consumer opinion in an online newspaper article; the consumer expressed that he/she wants to shop online more during the pandemic because online shopping gives him the chance to compare price and get the best deal (Hartman, 2020). Full-time employment as another important factor in estimating more comparative online shopping is reflecting the overall economic well-being. Having full-time employment as a significant factor in positively influencing more comparative online shopping means that people are willing to have more purchasing activity (including online purchasing activity) when the economy is doing well and people are earning more with full-time jobs.

However, people might wonder about why is there an increase in more comparative online shopping, when there is a high unemployment rate in the current situation? Although many people lost their jobs, they are actually moving toward the behavior of saving money. As highlighted in the results, the impact of increase in the percentage of people who responded to increased carpooling and spending less overall (saving behavior) jointly are bigger than the impact of full-time employment. Therefore, it is reasonable that everyone is still seeing the increase in more comparative online shopping when the unemployment rate is high. To support my finding about saving behavior leading to increased online shopping, previous research suggested a rationale of doing more online shopping during bad economic times because people in general want to save more money (Baker, Farrokhnia, Meyer et al., 2020). This type of rationale might be related to the idea of saving with price comparison on different online platforms (Hartman, 2020). Moving forward, some of my findings were unexpected as it does not confirm some previous studies’ findings presented in my domain review. Some previous articles have indicated that payment infrastructure is important to doing more online shopping (Comegys et al., 2006; Grashuis et al., 2020). However, my final model did not include the consumer behavior variable of no credit card to estimate more comparative online shopping. This is unexpected, but one possible reason for that might be most people in current society have a credit card. So, the payment infrastructure may no longer be a key determining factor for conducting online shopping.

Lastly, the random forest model and the single tree model together seem to identify the behavior of coupon usage as a very important behavioral factor in estimating more comparative online shopping aside from the variable already identified in my linear regression model. The tree model’s split on using the more coupon variable suggested that a higher percentage of people using more coupon seems to imply higher percentage of people doing more comparative shopping online. I think this variable also implies about people’s action of saving because coupon usage helps the consumer save money. The reason why it is not selected from my initial StepAIC model could be that coupon usage explained the same aspect of saving money as spending less overall, which is not necessary to keep both.

**Conclusion**

Overall, my findings on higher percentage of people doing more comparative online shopping during pandemic time compared to right before pandemic time confirmed my initial expectation. As noted in previous research regarding COVID’s impact on online shopping, many e-commerce sectors are experiencing huge growth in sales during pandemic time (Chang & Meyerhoefer, 2020). Therefore, I am not surprised with my result indicating that there is a higher percentage of people doing more comparative online shopping during pandemic time compared to right before pandemic time. For my findings on consumer’s saving behaviors being highly associated with more comparative online shopping behavior, it seems to validate the current market analysis on consumer shopping preference. Many consumers are preferring online shopping due to safety concerns and the ability to save by comparative shopping with different e-commerce platforms (Hartman, 2020). Therefore, consumer’s saving behavior is another big factor aside from the current health panic for incentivizing people to do more online shopping. These are the key findings of my research paper.

In general, the largest contribution of this research is the analysis on how certain consumer behaviors and the current pandemic influence people to perform even more online shopping activities than pre-pandemic time, which adds on to the knowledge of the current existing growth of e-commerce. Moreover, my linear regression model and the random forest model can be a useful tool for predicting the percentage of people that will be doing more comparative online shopping for future analysis.

Although my model has pretty good prediction capability with low RMSE value, there are still a few limitations in my research that I want my readers to be aware of. Given that the consumer dataset is survey based and all COVID data are not 100% accurate, there is a potential accuracy limitation on my model result. In addition, my own judgement was involved when selecting consumer behavior variables and removing variables due to model validation issue. Therefore, there is a certain level of my personal bias potentially involved in my models. Furthermore, this paper’s goal is to provide analysis on what influences more comparative online shopping. Therefore, the information might be more favorable to business as they can use the information to implement marketing strategies such as issue more coupons. The unintended consequence might be seen as favorable or not favorable by the general public. So, please be cautious in implementing policy or marketing strategies based on the finding of this research paper.

For future work, my research can still be improved with more accurate data on more comparative online shopping trend compared to survey responses for improving my model (i.e., gain monthly consumer data from e-commerce firms such as Amazon on what percentage of the users are having more frequent purchases than the previous month). Then, a more accurate comparative online shopping trend can be used to confirm the linear model in this research or build a new model for investigating what consumer behaviors are associated with more comparative online shopping behavior. In addition, as mobile devices began to be more popular among the younger population, shopping online with mobile device seems to be a new trend of shopping style that people are picking up. Prosper’s data also includes a variable for more comparative mobile shopping as a sub-category survey question. Therefore, it will be interesting for future investigation to test if the consumer behavior factors that are associated with more comparative online shopping will also be associated with the more comparative mobile shopping trend.

**References**

Alboukadel Kassambara (2016). ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2'. R package version 0.1.1. <http://www.sthda.com/english/wiki/ggcorrplot>

Baker, S., Bloom, N., Davis, S., & Terry, S. (2020). COVID-Induced Economic Uncertainty. *NATIONAL BUREAU OF ECONOMIC RESEARCH,* 1-16. doi:10.3386/w26983

Baker, S., Farrokhnia, R., Meyer, S., Pagel, M., & Yannelis, C. (2020). How Does Household Spending Respond to an Epidemic? Consumption During the 2020 COVID-19 Pandemic. *NATIONAL BUREAU OF ECONOMIC RESEARCH,* 1-34. doi:10.3386/w26949

Bogue, R. (2016). Growth in e-commerce boosts innovation in the warehouse robot market. *Industrial Robot: An International Journal,* *43*(6), 583-587. doi:10.1108/ir-07-2016-0194

Chang, H., &amp; Meyerhoefer, C. (2020). COVID-19 and the Demand for Online Food Shopping Services: Empirical Evidence from Taiwan. NATIONAL BUREAU OF ECONOMIC RESEARCH, 1-29. doi:10.3386/w27427

Comegys, C., Hannula, M., & Väisänen, J. (2006). Longitudinal comparison of Finnish and US online shopping behaviour among university students: The five-stage buying decision process. *Journal of Targeting, Measurement and Analysis for Marketing,* *14*(4), 336-356. doi:10.1057/palgrave.jt.5740193

Fox J, Weisberg S (2019). An R Companion to Applied Regression, Third edition. Sage, Thousand Oaks CA. <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>

Genuer, R., Poggi, J., & Tuleau-Malot, C. (2010). Variable Selection Using Random Forests. *Pattern Recognition Letter,* *31*(14), 2225-2236. <https://doi.org/10.1016/j.patrec.2010.03.014>

Grashuis, J., Skevas, T., & Segovia, M. S. (2020). Grocery Shopping Preferences during the COVID-19 Pandemic. *Sustainability,* *12*(13), 5369th ser. doi:10.3390/su12135369

Horikoshi M, Tang Y (2018). *ggfortify: Data Visualization Tools for Statistical Analysis Results*. [https://CRAN.R-project.org/package=ggfortify](https://cran.r-project.org/package=ggfortify).

Hall, M. C., Prayag, G., Fieger, P., & Dyason, D. (2020). Beyond panic buying: Consumption displacement and COVID-19. *Journal of Service Management,* *Ahead-of-print*(Ahead-of-print). doi:10.1108/josm-05-2020-0151

Hartman, M. (2020, May 15). *Consumers are buying less while shopping online more*. <https://www.marketplace.org/2020/05/14/consumers-are-buying-less-while-shopping-online-more/>.

Ho, S., Kauffman, R. J., & Liang, T. (2007). A growth theory perspective on B2C e-commerce growth in Europe: An exploratory study. *Electronic Commerce Research and Applications,* *6*(3), 237-259. doi:10.1016/j.elerap.2006.06.003

Horikoshi M, Tang Y (2018). *ggfortify: Data Visualization Tools for Statistical Analysis Results*. [https://CRAN.R-project.org/package=ggfortify](https://cran.r-project.org/package=ggfortify).

Koufaris, M., Kambil, A., & Labarbera, P. A. (2001). Consumer Behavior in Web-Based Commerce: An Empirical Study. *International Journal of Electronic Commerce*, *6*(2), 115–138. https://doi.org/10.1080/10864415.2001.11044233

Kriegel, H.-P., Kröger, P., Schubert, E., & Zimek, A. (2008). A General Framework for Increasing the Robustness of PCA-Based Correlation Clustering Algorithms. *Lecture Notes in Computer Science Scientific and Statistical Database Management*, 418–435. https://doi.org/10.1007/978-3-540-69497-7\_27

Kuhn, M. (2008). Caret package. Journal of Statistical Software, 28(5)

Ma, C., & Iqbal, M. (1983). Statistical comparison of models for estimating solar radiation on inclined surgaces. *Solar Energy*, *31*(3), 313–317. https://doi.org/10.1016/0038-092x(83)90019-1

Milborrow, S. (2020). Package ‘rpart.plot’. Plot rpart models: An enhanced version of plot.rpart. <https://www.idg.pl/mirrors/CRAN/web/packages/rpart.plot/rpart.plot.pdf>.

Moshrefjavadi, M. H., Dolatabadi, H. R., Nourbakhsh, M., Poursaeedi, A., & Asadollahi, A. (2012). An Analysis of Factors Affecting on Online Shopping Behavior of Consumers. *International Journal of Marketing Studies,* *4*(5). doi:10.5539/ijms.v4n5p81

Prost, L., Makowski, D., & Jeuffroy, M.-H. (2008). Comparison of stepwise selection and Bayesian model averaging for yield gap analysis. *Ecological Modelling*, *219*(1-2), 66–76. https://doi.org/10.1016/j.ecolmodel.2008.07.026

Prosper Insights & Analytics. (2020). *U.S. 18+ Consumer Behavior 2010-present* [Data set]. Prosper Insights & Analytics. <https://drive.google.com/file/d/1IQMzud29sBCpEgLRCYl62T7Fqin1-TPf/view?usp=sharing>

Richards, T. J., & Rickard, B. (2020). COVID‐19 impact on fruit and vegetable markets. *Canadian Journal of Agricultural Economics/Revue Canadienne D'agroeconomie,* *68*(2), 189-194. doi:10.1111/cjag.12231

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Retrieved from [https://www.R-project.org/](https://www.r-project.org/)

Schmidt, A. F., & Finan, C. (2018). Linear regression and the normality assumption. *Journal of Clinical Epidemiology*, *98*, 146–151. https://doi.org/10.1016/j.jclinepi.2017.12.006

Stanciu, S., Radu, R. I., Sapira, V., Bratoveanu, B. D., & Florea, A. M. (2020). Consumer Behavior in Crisis Situations. Research on the Effects of COVID-19 in Romania.*Annals of the University Dunarea deJosof Galati: Fascicle: I, Economics & Applied Informatics*,*26*(1).

The COVID Tracking Project. (2020). *COVID Summary Data for the United State* [Data set]. The Atlantic Monthly Group. <https://covidtracking.com/data/download>

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*, Fourth edition. Springer, New York. ISBN 0-387-95457-0, <http://www.stats.ox.ac.uk/pub/MASS4/>

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, [https://ggplot2.tidyverse.org](https://ggplot2.tidyverse.org/).

Yahoo Finance. (2020). CBOE Volatility Index (^VIX) [Data set]. Yahoo Finance. <https://finance.yahoo.com/quote/%5EVIX/history?p=%5EVIX>

**Appendix A - Supplementary Figures and Tables**

**Figure 1: Trend of More Comparative Online Shopping Response**



*Source: U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020)*

*Note: Graphics created using Program R’s ggplot2 packages (Wickham, 2016)*

**Table 1: Full Data Model Statistics in Estimate More Comparative Online Shopping %**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 0.231254 | 0.004143 | 55.813 | 2e-16 \*\*\* |
| COVID | -0.025606 | 0.014939 | -1.714 | 0.09 . |

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Source: *U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020)*

*Note: Linear Regression Model using Program R’s built in function (R Core Team, 2020)*

**Table 2: Past 14 Months Model Statistics in Estimate More Comparative Online Shopping %**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 0.175041 | 0.003660 | 47.826 | 4.57e-15 \*\*\* |
| COVID | 0.030607 | 0.005176 | 5.913 | 7.10e-05 \*\*\* |

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Source: *U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020)*

*Note: Linear Regression Model using Program R’s built in function (R Core Team, 2020)*

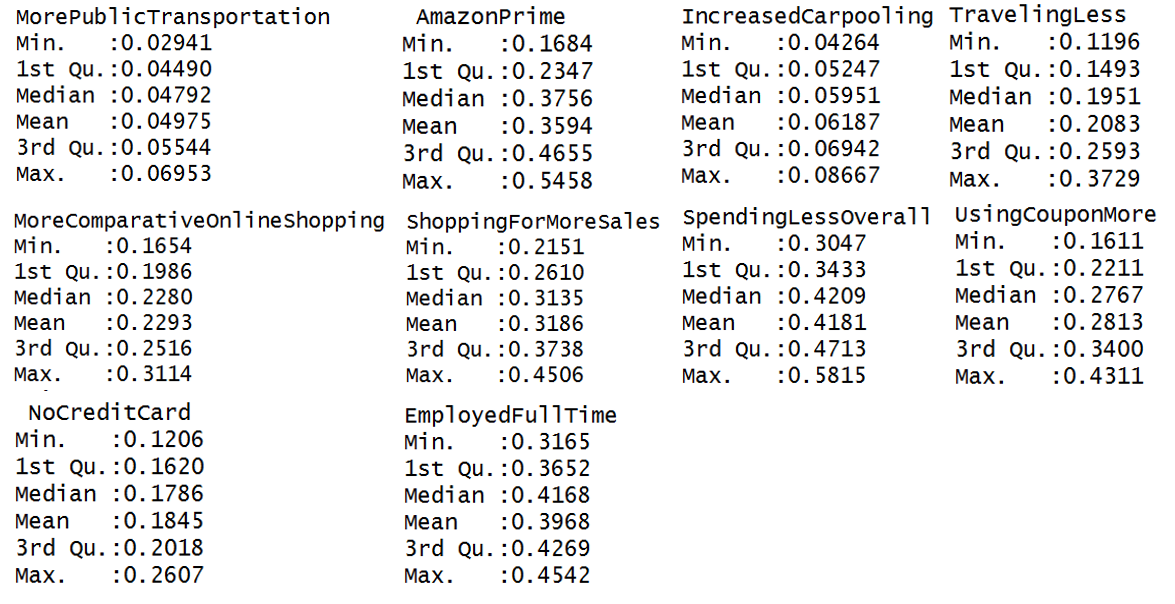
**Figure 2: Consumer Behavior Factors Correlation Plot**



*Source: U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020)*

*Note: Graphics created using Program R’s ggcorrplot packages (Kassambara, 2016)*

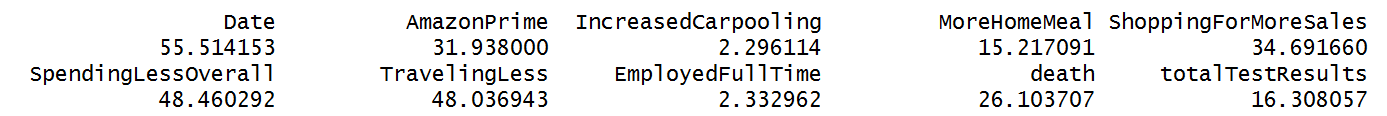
**Figure 4: Consumer Behavior Factors Summary Statistics Matrix**



Source: *U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020)*

*Note: Summary Statistics using Program R’s built in function (R Core Team, 2020)*

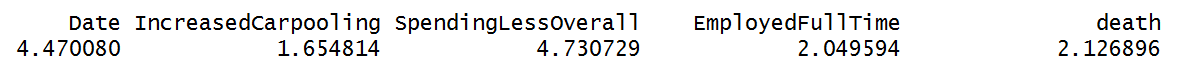
**Figure 5: VIF Matrix on Initial Model**



*Source: U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020)*

*Note: VIF generated using Program R’s car package (Fox & Weisberg, 2019)*. *Tests were performed on StepAIC Regression Model.*

**Figure 6: VIF Matrix on Final Model**



*Source: U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020)*

*Note: VIF generated using Program R’s car package (Fox & Weisberg, 2019)*. *Tests were performed on Final Regression Model.*

**Figure 7: Final Linear Model Residual Histogram**



*Source: U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020)*

*Note: Graphics created using Program R’s ggplot2 packages (Wickham, 2016). Residuals generated from Final Regression Model.*

**Figure 8. Final Linear Model Residual Histogram**



*Source: U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020)*

*Note: Graphics created using Program R’s ggplot2 packages (Wickham, 2016). Residuals generated from Final Regression Model.*

**Figure 9. Random Forest Model Variable Importance Chart**



*Source: U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020)*

*Note: Graphics created using Program R’s ggplot2 package and caret package(Wickham, 2016; Kuhn, 2008).*

**Figure 10. A Single Tree Model for Estimating More Comparative Online Shopping %**



*Source: U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020)*

*Note: Graphics created using Program R’s rpart.plot package and caret package(Milborrow, 2020; Kuhn, 2008).*

**Table 3: StepAIC Regression Model in Estimate More Comparative Online Shopping %**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 5.280e-01 | 1.714e-01 | 3.081 | 0.002831 \*\*\* |
| Spending Less Overall | 4.697e-01 | 8.818e-02 | 5.326 | 8.98e-07 \*\*\* |
| Amazon Prime | 2.587e-01 | 4.835e-02 | 5.350 | 8.15e-07 \*\*\* |
| Death | 7.356e-07 | 1.794e-07 | 4.100 | 9.85e-05 \*\*\* |
| Date | -4.195e-05 | 9.529e-06 | -4.403 | 3.28e-05 \*\*\* |
| Employed Full Time | 2.467e-01 | 7.166e-02 | 3.443 | 0.000919 \*\*\* |
| Shopping for More Sales | 3.293e-01 | 9.091e-02 | 3.623 | 0.000510 \*\*\* |
| Traveling Less | -2.858e-01 | 1.080e-01 | -2.647 | 0.009786 \*\* |
| Increased Carpooling | 4.952e-01 | 1.761e-01 | 2.813 | 0.006174 \*\* |
| Total Test Results | -9.434e-10 | 3.786e-10 | -2.492 | 0.014784 \* |
| More Home Meal | -1.649e-01 | 9.247e-02 | -1.783 | 0.078354 . |

**N = 91 F-statistics: 131.3 \*\*\* Adj-R2 = 0.9354**

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

*Source: U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020)*

*Note: StepAIC Regression Model using Mass package (Venables & Ripley, 2002)*

**Table 4: Final Linear Model in Estimate More Comparative Online Shopping %**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 5.495e-01 | 8.181e-02 | 6.717 | 1.99e-09 \*\*\* |
| Date | -3.210e-05 | 3.313e-06 | -9.691 | 2.16e-15 \*\*\* |
| Spending Less Overall | 1.801e-01 | 3.375e-02 | 5.337 | 7.72e-07 \*\*\* |
| Employed Full Time | 2.890e-01 | 8.229e-02 | 3.512 | 0.000714 \*\*\* |
| Increased Carpooling | 5.608e-01 | 1.831e-01 | 3.063 | 0.002938 \*\* |
| Death | 1.996e-07 | 6.274e-08 | 3.181 | 0.002050 \*\* |

**N = 91 F-statistics: 168.7 \*\*\* Adj-R2 = 0.9084**

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

*Source: U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020)*

*Note: Linear Regression Model using Program R’s built-in function (R Core Team, 2020)*

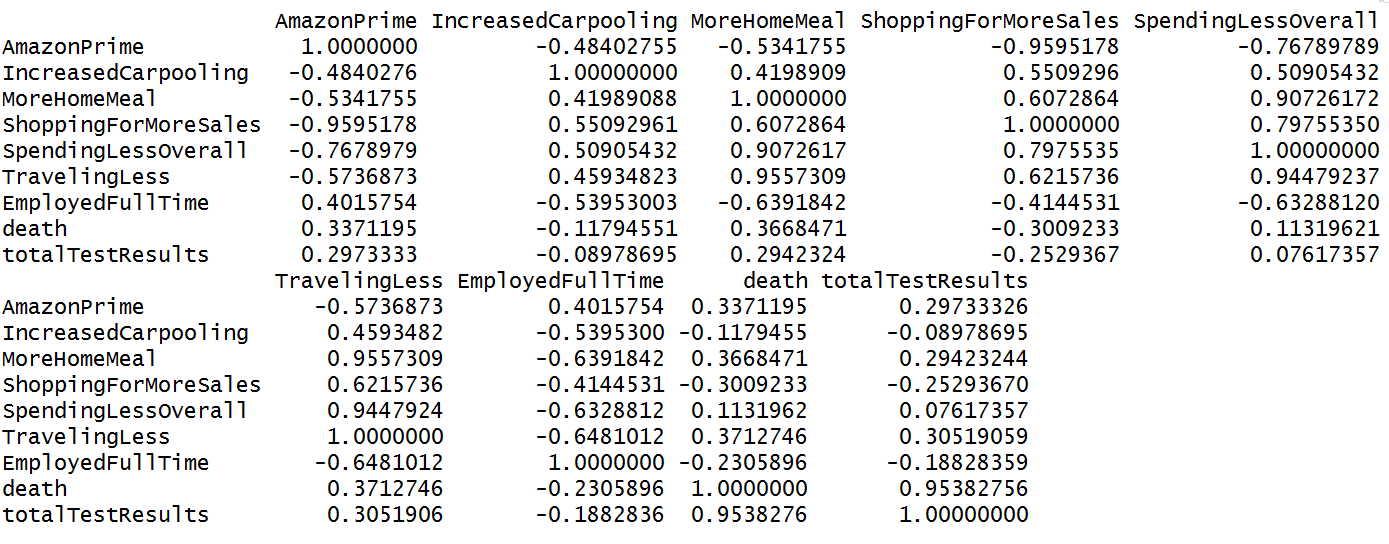
**Table 5: Variables Description**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Variable Description** | **Coded/Measure** |
| **Consumer Behavior Variables (Survey Response)** | | |
| More Comparative Online Shopping | Percentage of total people respond in the consumer behavior survey saying that they shop more online compare to the previous month. | Measured in a number between 0 and 1 to represent % |
| Date | The date for when the survey response was recorded. Since it is monthly recorded data, the date is the first day of each month. | Measured in a number between 0 and 1 to represent % |
| More Public Transportation | Percentage of total people respond in the consumer behavior survey saying that they take more public transportation. | Measured in a number between 0 and 1 to represent % |
| Amazon Prime | Percentage of total people respond in the consumer behavior survey saying that they purchased amazon prime membership. | Measured in a number between 0 and 1 to represent % |
| Increased Carpooling | Percentage of total people respond in the consumer behavior survey saying that they are doing more carpooling. | Measured in a number between 0 and 1 to represent % |
| More Home Meal | Percentage of total people respond in the consumer behavior survey saying that they cook more meals at home. | Measured in a number between 0 and 1 to represent % |
| Shopping for More Sales | Percentage of total people respond in the consumer behavior survey saying that they are shopping for more sales. | Measured in a number between 0 and 1 to represent % |
| Spending Less Overall | Percentage of total people respond in the consumer behavior survey saying that they are spending less overall. | Measured in a number between 0 and 1 to represent % |
| Traveling Less Overall | Percentage of total people respond in the consumer behavior survey saying that they are traveling less overall. | Measured in a number between 0 and 1 to represent % |
| Using Coupon More | Percentage of total people respond in the consumer behavior survey saying that they are using more coupons. | Measured in a number between 0 and 1 to represent % |
| No Credit Card | Percentage of total people respond in the consumer behavior survey saying that they don’t have a credit card. | Measured in a number between 0 and 1 to represent % |
| Employed Full Time | Percentage of total people respond in the consumer behavior survey saying that they are employed full-time. | Measured in a number between 0 and 1 to represent % |
| **COVID-19 Statistics** | | |
| Death | Total cumulative deaths due to COVID-19 | Measured in person |
| Cumulative Hospitalization | Total cumulative hospitalization due to COVID-19 pandemic | Measured in person |
| Positive | Total cumulative positive COVID-19 cases | Measured in person |
| Recovered | Total cumulative recovered COVID-19 cases | Measured in person |
| Total Test Result | Total tests being done for COVID-19 | Measured in person |
| COVID | This is a (0 1) dummy variable that represent whether the pandemic is present or not. | Dummy Variable  0 = No COVID-19  1 = Yes COVID-19 |
| COVID Positive Rate | Percentage of total tests that are positive generate by total positive cases divide by total test results for COVID-19 | Measured in a number between 0 and 1 to represent % |
| **Financial Fluctuation Variables** | | |
| Volatility Index | A number range from 0 to 100 that serves as a standard to measure the volatility of stock market. 0 represent not volatile (no fluctuation) and 100 represent extreme volatility. | Measured in continuous values |
| Monthly Volatility Index Mean | A number range from 0 to 100 that takes the mean of volatility index for each month. | Measured in continuous values |

*Source: U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020),* *COVID National Summary Data (The COVID Tracking Project, 2020), and Volatility Index of S&P 500 (Yahoo Finance, 2020)*

*Note: These are the initial variables to build my final model on in estimate More Comparative Online Shopping.*

**Figure 11: Variable Correlation Matrix for Initial StepAIC Model**



Source: *U.S. 18+ Consumer Behavior 2010-present Data (Prosper Insights & Analytics, 2020)*

*Note: correlation matrix using Program R’s built-in function (R Core Team, 2020*

**Appendix B - Supplementary Code for Analysis**

The attached code below come from my ConsumerBehavior\_StatisticalAnalysis.Rmd file, which it includes all the codes I have used for analysis. The codes were splits into three sections: Data and r-packages preparation, COVID Analysis, and Consumer Behavior Analysis. COVID Analysis section contain codes that tackle my first research question about how COVID influence more comparative online shopping behavior. The Consumer Behavior Analysis section contain codes that tackle the objective of identifying behaviors that associate with more comparative online shopping behavior. Each section can be identified with a bolded/underlined header. Within each section, there are sub-headers that explain the role of each code chunks. The codes below did not include any data cleaning codes explained in the method section. The data cleaning codes are on my private GitHub Repository and available upon request.

***paste in analysis code from Rmd file (courier font)***

**#Data and r-packages preparation**

**##Install Packages for Plotting and Running Regression**

```{r}

#install.packages("ggplot2")

#install.packages("dplyr")

#install.packages("MASS")

#install.packages("ggfortify")

#install.packages("plotly")

#install.packages("ggcorrplot")

#install.packages("car")

#install.packages("caret")

#install.packages("rpart.plot")

#install.packages("randomForest")

```

**##Read Dataset**

```{r}

FinalModel = read.csv("FinalModelData.csv")[-1]

FinalModel$Date = as.Date(as.character(FinalModel$Date))

```

**##Rename some variables to shorten variable names**

```{r}

colnames(FinalModel)[2] = "MorePublicTransportation"

colnames(FinalModel)[4] = "IncreasedCarpooling"

colnames(FinalModel)[5] = "MoreHomeMeal"

colnames(FinalModel)[6] = "MoreComparativeOnlineShopping"

colnames(FinalModel)[7] = "MoreComparativeMobileShopping"

colnames(FinalModel)[8] = "ShoppingForMoreSales"

colnames(FinalModel)[9] = "SpendingLessOverall"

colnames(FinalModel)[10] = "TravelingLess"

colnames(FinalModel)[11] = "UsingCouponMore"

colnames(FinalModel)[20] = "MonthlyVolatilityIndexMean"

colnames(FinalModel)

```

**#COVID Analysis**

**##Loading Packages for graphing ggplots**

```{r}

library(ggplot2)

library(dplyr)

```

**##Generate Dataset Pre and During Covid + GGPLOT**

```{r}

During\_Covid\_Dataset = FinalModel[123:129,]

Pre\_Covid\_Dataset = FinalModel[1:123,]

ggplot()+

geom\_line(aes(x=Date, y=MoreComparativeOnlineShopping, color = “green”), data = Pre\_Covid\_Dataset)+

geom\_line(aes(x=Date, y=MoreComparativeOnlineShopping, color = “red”), data = During\_Covid\_Dataset)+

ggtitle(“People’s Response with More Online Shopping Over Years”)+

xlab(“Years”)+

ylab(“% of Reponse For More Comparative Online Shopping”)+

labs(colour = “Legend”)+

scale\_color\_manual(labels = c(“PreCOVID”, “DuringCOVID”), values = c(“red”, “blue”)) +

scale\_x\_date(date\_breaks = “years” , date\_labels = “20%y”)+

theme\_bw()+

theme(plot.title = element\_text(hjust = 0.5), text=element\_text(size=10, family=”TT Times New Roman”))

####Save high resolution figure

ggsave(“Trend1.tiff”, units=”in”, width=7, height=4, dpi=300, compression = ‘lzw’)

```

**##Covid Model**

```{r}

CovidModel = lm(MoreComparativeOnlineShopping ~ COVID, data = FinalModel)

summary(CovidModel)

```

**##Different Sample for Covid Model (More Recent Period Analysis)**

```{r}

Differensample = FinalModel[116:129,]

CovidModel\_differensample = lm(MoreComparativeOnlineShopping ~ COVID, data = Differensample)

summary(CovidModel\_differensample)

```

**#Consumer Behavior Analysis**

**##Data Exploration**

**###Loading Package for StepAIC**

```{r}

library(MASS)

```

**###Load packages for PCA & Correlation Plot**

```{r}

library(ggfortify)

library(plotly)

library(ggcorrplot)

```

**###Correlation Plot**

```{r}

corr <- round(cor(na.omit(FinalModel[,c(2,3,4,5,6,8,9,10,11,12,13)])), 2)

ggcorrplot(corr, tl.cex = 8, lab\_size = 2.3, title = "Variable Correlation Plot", legend.title = "Corr Value", hc.order = TRUE, type = "lower",lab = TRUE,ggtheme = ggplot2::theme\_dark())+theme(plot.title = element\_text(hjust = 0.5))

####Save high resolution figure

ggsave("correlation.tiff", units="in", width=5, height=4, dpi=300, compression = 'lzw')

```

**###Correlation Matrix for Specific Correlation Values**

```{r}

corr

```

**###PCA Plot**

```{r}

PCAModel = na.omit(FinalModel[,c(2,3,4,5,6,8,9,10,11,12,13)])

ConsumerBehavior.pca = prcomp(PCAModel[,c(1:11)], scale.=TRUE)

autoplot(ConsumerBehavior.pca, loadings = TRUE, loadings.colour = 'blue', loadings.label = TRUE, loadings.label.size = 3, main = "Variable Biplot With Loadings (PC1 VS PC2)") + theme\_bw() + theme(plot.title = element\_text(hjust = 0.5))

tiff(file="PCA.tiff", width=5, height=4, units="in", res=300, compression = 'lzw')

autoplot(ConsumerBehavior.pca, loadings = TRUE, loadings.colour = 'blue', loadings.label = FALSE, loadings.label.size = 3, main = "Variable Biplot With Loadings (PC1 VS PC2)") + theme\_bw() + theme(plot.title = element\_text(hjust = 0.5))

dev.off()

```

**###Summary Statistics for My Consumer Behavior Independent Variables**

```{r}

summary(FinalModel)

```

**##Linear Regression Model**

**###Use StepAIC for Selecting Initial Model**

```{r}

full.model <- lm(MoreComparativeOnlineShopping ~ .-MoreComparativeMobileShopping, data = FinalModel)

step.model <- stepAIC(full.model, direction = "both", trace = FALSE)

summary(step.model)

```

**###Model Validation and modify model**

**###Check Multicollinearity Using VIF Test (car package)**

```{r}

library(car)

vif(step.model)

```

**### Check Multicollinearity Using Correlation Matrix**

```{r}

cor(na.omit(FinalModel[,c(3,4,5,8,9,10,13,14,18)]))

```

*Note: Among my independent variables, the highly correlated (r > 0.8 or r < -0.8) variable pairs/triples are AmazonPrime & ShoppingForMoreSales, MoreHomeMeal & SpendingLessOverall & TravelingLess, death & totalTestResults.*

**###Regression trials for variable removal decision among the highly correlated variables (keeping the variable that provide me with higher R^2 value).**

**###AmazonPrime VS ShoppingForMoreSales**

```{r}

summary(lm(formula = MoreComparativeOnlineShopping ~ Date +

IncreasedCarpooling + MoreHomeMeal +

ShoppingForMoreSales + SpendingLessOverall +

TravelingLess + EmployedFullTime + death +

totalTestResults, data = FinalModel))

summary(lm(formula = MoreComparativeOnlineShopping ~ Date +

AmazonPrime + IncreasedCarpooling + MoreHomeMeal +

SpendingLessOverall +

TravelingLess + EmployedFullTime + death +

totalTestResults, data = FinalModel))

```

*Note: Keeping AmazonPrime seem to be a better choice because higher R^2 value and more variable remain significant. Moving forward with AmazonPrime.*

**###MoreHomeMeal VS SpendingLessOverall VS TravelingLess**

```{r}

summary(lm(formula = MoreComparativeOnlineShopping ~ Date +

AmazonPrime + IncreasedCarpooling + MoreHomeMeal +

EmployedFullTime + death +

totalTestResults, data = FinalModel))

summary(lm(formula = MoreComparativeOnlineShopping ~ Date +

AmazonPrime + IncreasedCarpooling +

SpendingLessOverall + EmployedFullTime + death +

totalTestResults, data = FinalModel))

summary(lm(formula = MoreComparativeOnlineShopping ~ Date +

AmazonPrime + IncreasedCarpooling +

TravelingLess +

EmployedFullTime + death +

totalTestResults, data = FinalModel))

```

*Note: It seems like keeping SpendingLessOverall is better to keep as it generate higher R^2 value.*

**###Death VS totalTestResults**

```{r}

summary(lm(formula = MoreComparativeOnlineShopping ~ Date +

AmazonPrime + IncreasedCarpooling +

SpendingLessOverall + EmployedFullTime + death, data = FinalModel))

summary(lm(formula = MoreComparativeOnlineShopping ~ Date +

AmazonPrime + IncreasedCarpooling +

SpendingLessOverall + EmployedFullTime + totalTestResults, data = FinalModel))

```

*Note: Death is better to keep as it allows for higher R^2 and all my variables are highly significant.*

**###Current Model & Check Multicollinearity Again**

```{r}

CurrentModel = lm(formula = MoreComparativeOnlineShopping ~ Date +

AmazonPrime + IncreasedCarpooling +

SpendingLessOverall + EmployedFullTime + death, data = FinalModel)

summary(CurrentModel)

vif(CurrentModel)

```

*Note: It seems like Date and AmazonPrime continue to be highly correlated with each other creating multicollinearity issue. Therefore, I might need to remove one of them.*

**###AmazonPrime & Date**

```{r}

summary(lm(formula = MoreComparativeOnlineShopping ~

AmazonPrime + IncreasedCarpooling +

SpendingLessOverall + EmployedFullTime + death, data = FinalModel))

summary(lm(formula = MoreComparativeOnlineShopping ~ Date +

IncreasedCarpooling +

SpendingLessOverall + EmployedFullTime + death, data = FinalModel))

```

*Note: Move Forward with Date variable instead of AmazonPrime variable because it is a better model in terms of R-square. Date was also able to detrend serial correlation while Amazon Prime might just due to random correlation.*

**##Final Model & Model Validation Tests: Heteroskedasticity, Normality, Serial Correlation, and VIF**

```{r}

###Final Model

FinalRegressionModel = lm(formula = MoreComparativeOnlineShopping ~ Date + IncreasedCarpooling + SpendingLessOverall + EmployedFullTime + death, data = FinalModel)

summary(FinalRegressionModel)

###Statistical Tests for Model Validation

vif(FinalRegressionModel)

shapiro.test(FinalRegressionModel$res)

ncvTest(FinalRegressionModel)

durbinWatsonTest(FinalRegressionModel)

###Residual Histogram

Res = data.frame(FinalRegressionModel$res)

Res %>%

ggplot(aes(x = FinalRegressionModel.res)) +

geom\_histogram(colour = 'white', fill = 'seagreen')+

ggtitle("Linear Regression Residual Histogram")+

xlab("Residuals")+

ylab("Count")+

theme\_bw()+

theme(plot.title = element\_text(hjust = 0.5), text=element\_text(size=10, family="TT Times New Roman"))

####Save high resolution figure

ggsave("Normality.tiff", units="in", width=5, height=3, dpi=300, compression = 'lzw')

###Residual Plot

Res = cbind(Res, FinalRegressionModel$fitted.values)

Res %>%

ggplot(aes(FinalRegressionModel$fitted.values, FinalRegressionModel.res)) +

geom\_point()+

ggtitle("Linear Regression Residual Plot")+

xlab("Fitted Values")+

ylab("Residuals")+

theme\_bw()+

theme(plot.title = element\_text(hjust = 0.5), text=element\_text(size=10, family="TT Times New Roman"))

####Save high resolution figure

ggsave("Residual.tiff", units="in", width=5, height=3, dpi=300, compression = 'lzw')

```

**###Calculate RMSE to measure model performance**

```{r}

FinalModel\_RMSETest = FinalModel[39:129,] ###Avoid NA Errors

FinalRegressionModelPredict = predict(FinalRegressionModel, FinalModel\_RMSETest)

sqrt(mean((FinalRegressionModelPredict-FinalModel\_RMSETest$MoreComparativeOnlineShopping)^2))

```

**##Tree Model and Random Forest Model to confirm important behavior variables for estimating more comparative online shopping behavior**

```{r}

###Machine Learning Can't deal with missing value, so I will remove the missing value first.

FinalModel1 = na.omit(FinalModel)

library(caret)

library(rpart.plot)

trctrl = trainControl(method = "repeatedcv", number = 10, repeats = 5)

DecisionTreeModel = train(MoreComparativeOnlineShopping ~ .-MoreComparativeMobileShopping,

data = FinalModel1,

method = "rpart",

tuneLength = 10,

trControl = trctrl)

DecisionTreeModel

prp(DecisionTreeModel$finalModel, varlen = 0, main = "Single Tree Model")

###Save high resoltuion figure

tiff(file="Tree.tiff", width=8, height=4, units="in", res=300, compression = 'lzw')

prp(DecisionTreeModel$finalModel, varlen = 0, main = "Single Tree Model")

dev.off()

```

**###Random Forest**

```{r}

trcontrol = trainControl(method="oob")

RandomForestModel = train(MoreComparativeOnlineShopping ~ .-MoreComparativeMobileShopping,

data = FinalModel1,

method="rf",

tuneLength = 7,

trControl=trcontrol,

ntree=2000,

importance = TRUE)

```

**###Random Forest Variable Importance Chart**

```{r}

Importance = varImp(RandomForestModel)

Importance$importance

plot(Importance, main = "Variable Importance Chart")

###Save high resoltuion figure

tiff(file="Importance.tiff", width=7, height=5, units="in", res=300, compression = 'lzw')

plot(Importance, main = "Variable Importance Chart")

dev.off()

```

**###Calculate RMSE to measure model performance**

```{r}

RFPredict = predict(RandomForestModel, FinalModel\_RMSETest)

sqrt(mean((RFPredict-FinalModel\_RMSETest$MoreComparativeOnlineShopping)^2))

```

*Note: The RMSE for Random Forest changed slightly every time I make a prediction, but it should not be too much off from what I have reported in the paper.*