**Predicting Housing Prices in a College Town in Virginia**

**Abstract:**
As on-ground student housing in our local college town becomes increasingly unreliable, students are looking for off-campus housing for their final two years in school. The intensifying competitiveness of the market has put a financial burden on many students looking for accommodations, especially when houses are becoming less available, and leasing apartments has become a game of luck. Understanding the dynamics of the housing market will help students navigate the challenging terrain of housing affordability. It will serve as a resource for sellers and real estate owners, equipping them with the knowledge needed to accurately price their properties. Students could use this data to factor in what they are looking for in housing and accurately predict what price they should be paying, preventing students from overpaying and falling into financial debt. The prediction model displays an R-squared value of .9375 and future research should focus on adding additional explanatory variables.

**Introduction:**

As on-campus student housing in the town becomes increasingly unreliable, students are turning to off-campus options, creating a competitive market. This poses financial challenges, especially with fewer available options and uncertain leases. We hope our data will answer the following questions: **Do housing units farther from the campus, measured in miles from the Central Campus, tend to have lower prices? Do housing units with amenities, specifically in-unit laundry, have a higher price? How does housing type (apartment or full house) affect monthly rent in dollars?**

Each data observation corresponds to a single housing unit. Different apartment units within the same building with varying numbers of bedrooms and bathrooms are treated separately. This is indicated by a number next to the address, with the highest number representing the number of floorplans examined. Our data was compiled within a Google Sheet with all of our explanatory variables, the name of the data observation, and our response variable. It was representative of a small sample of the town housing and did not include the entire population. This was in part because of a need for more available data. There are also many data points within the population, making it unreasonable to collect data on the whole population within a concise time frame. We did our best to take a representative sample in all aspects of the data and randomly selected most housing units. We made a few manipulations to the data besides listing parking spots as qualitative and asking if they had it instead of listing the number of parking spaces a unit provides in quantitative form like we initially thought we would. Our data is objective and easily quantifiable, though slight inaccuracies may exist in pricing due to seasonal fluctuations and hidden costs.

**Table 1. Variable codebook**

| Variable Name | Abbreviated Name | Description |
|---|---|---|
| "Address" | Address | The address of the unit, obtained from websites. If there were multiple apartment types in one, each floor plan was numbered next to the address. |
| "Type..Apartment.House" | Type (Apartment/House) | If the housing unit is listed as an apartment or house on the website. It is coded 1 if a house and 0 if it is an apartment (base level). |
| "Square.Footage..sq..ft" | Square Footage | The square footage of the entire housing unit, measured in square feet and found on websites |
| "Distance.From.Campus..mi.." | Distance From Campus | Distance from a central point on campus, measured in miles using the ideal walking route found through Google Maps. |
| "X..Bedrooms" | Number of Bedrooms | The number of usable bedrooms in the listing. |
| "X..Bathrooms" | Number of Bathrooms | The number of usable bathrooms in the listing. We included full bathrooms as adding one and half bathrooms as adding 0.5 to the total. |
| "Parking.spot..Yes.No" | Parking Spot | The number of parking spots included with the housing unit in the listing. |
| "In.unit.laundry" | In-Unit Laundry | If there is laundry included within the unit, listed as yes or no. Coded 1 if yes and 0 if no (base level). |
| "Pool" | Pool | If there is the use of a pool included with the rental of the unit, listed as yes or no. Coded 1 if yes or 0 no (base level). |
| "Price.Sold.Per.Month" | Price Sold Per Month | The price of the housing unit per month, listed in US dollars on the website. This is the total price, not the price per person. It is our response variable |

**Method and Analysis:**

First, the model was built by adding quantitative variables and quantitative interactions using correlation plots and r values. Based on our intuition, there should be an interaction between the number

of bedrooms and square footage because more bedrooms take up more square footage. We did a t-test of this interaction and found that it was significant so it remained in the model. No higher-order variables were needed because none of the scatter plots were curvilinear (Figure B). So, the final model considering only quantitative variables was B0 + B1Bedrooms + B2Bathrooms + B3SquareFootage + B4BedroomSquareFootage since only Bedrooms, Bathrooms, and Square Footage had a high R-value. All of the t-tests in this section proved the variables were significant and should be included in the model.

Next, the qualitative variables and qualitative interactions were added. The boxplots only showed a mean difference for hosing type and laundry so those are the variables we added (Figure B). There was no crossing between the interaction plots so no interaction model was added. The final model from this step was B0 + B1Bedrooms + B2Bathrooms + B3SquareFootage +B4BedroomSquareFootage + B5Type + B6Laundry, where B5 = 0 if Apartment, 1 if house and B6 = 0 if no in-unit laundry and 1 if in-unit laundry. Lastly, the quantitative versus qualitative interactions were added. No interactions were added because none of our scatterplots of a qualitative and quantitative variable showed significant differences in the qualitative variable between the different quantitative variable levels (no separation of colors) (Figure C). The final model from this step was B0 + B1Bedrooms + B2Bathrooms + B3SquareFootage +B4SquareFootage+ B5Type + B6Laundry.

Then, multicollinearity was checked (Figure D). As shown in the multicollinearity plots, where all the variables are plotted against each other, the strongest pairwise relationship is between the Number of Bedrooms and the Square Footage with an r-value of around 0.5. There is no evidence of severe multicollinearity because all the VIFs were less than 10 and the average VIF is greater than 3. Next, the residual assumptions: lack of fit, constant variance, normality, and independence were checked (Figure E). Lack of fit was not violated because residual plots were linear with no pattern. Constant variance was not violated because the residual vs. fitted plot showed no fanning or semi-circle pattern when plotted. There was more clustering at lower prices but nothing significant enough to be a violation of constant variance. Normality was not violated because the QQ plot showed a mostly linear relationship, and the histogram of residuals was unimodal and symmetric with only 1 noticeable outlier. We do not consider the outlier to be severe enough to violate that assumption. Independence was not violated because there was no time series data within our observations. Outliers and influential observations were analyzed in the deleted studentized residuals vs. predicted values plot, using a threshold of the absolute value of 2 (Figure F). Observations 5, 16, and 17 were shown to be below the threshold, so they were removed from the final model and were used in our additional technique testing below.

For the additional technique, we decided to utilize external model validation. External model validation is crucial for assessing the town housing market because it ensures that regression models accurately predict real-world outcomes. Validation techniques build confidence and mitigate risks associated with relying on the model by evaluating predictive accuracy and identifying model limitations. To test the external model validity, we fit a new subset model removing the influential observations that we identified above. Specifically, we removed observations 5, 16, and 17. We then compared the new test statistics to the model without removing those observations. The p-values were very similar but the RSME was less, at 531.6881 instead of 580.7561 like it was before removing the influential observations. This was a positive sign because the lower the RSME, the better the model and the predictions it outputs because it shows the model is a better fit for the data.

Lastly, to assess the final model, the stepwise regression and global F-test processes were used (Figure G). The stepwise regression output did not remove any quantitative variables. A global F test was then used to test the validity of our final model.

**Results:**

Our final model produced a p-value of 2.2e-16, much smaller than 0.05, leading us to reject the null hypothesis and conclude that the final model is adequate for the data (Figure H). The large R-squared value of 0.9375 suggests that the model explains a substantial portion of the variance in the dependent variable. The f-statistic indicated that overall, the model is statistically significant, meaning the predictors have a collective effect on the price. The final prediction equation was Price.Sold.Per.Month = -15.9216 + 276.0221Bedrooms + 655.5136Bathrooms + 0.7326SquareFootage - 44.7936BedroomSquareFootage - 412.0681Type + 234.5429Laundry.

**Table 2. Summary of Model Coefficients**

| | Estimated Slope | P-Value |
|---|---|---|
| **Intercept** | -15.9216 | 0.962988 |
| **Bedrooms** | 276.0221 | 1.12e-05 |
| **Bathrooms** | 655.5136 | 2.67e-10 |
| **Square Footage** | 0.7326 | 1.47e-08 |
| **Type of Housing Unit (House)** | - 44.7936 | 0.053300 |
| **In Unity Laundry (Yes)** | - 412.0681 | 0.488409 |
| **BedroomsBathrooms (Interaction)** | 234.5429 | 0.000897 |

**Conclusion:**

Plugging in an observation helps assure us that the data is accurate at predicting the price per month because we can compare it to the actual price per month that the housing unit sold for. We plugged in the data for observation #1 (Camden Plaza- 230 14th Street NW #2), and our model gave us a final value of 1502.5204. Compared to the actual price sold per month of 1399, this is fairly accurate, with a residual of -103.5204. However, there are some limitations to our analysis due to the subjective decisions we had to make. For example, in our normality assumption analysis, the histogram of residuals was unimodal and symmetric with 1 noticeable outlier. We decided not to transform the variable since the histogram was normal, but our results could have been different if we had decided to square root or log it to transform to get a more symmetrical distribution. Adding on to that, there was an argument that the constant variance did show slight fanning from the left side of the model to the right but we decided that it was not extreme enough to be considered a severe violation. In the future, we would want to add more variables as predictors, including distance from different locations around the campus, such as the School of Engineering or the Medical Center. We could also expand our data by including observations from off-campus apartments and houses all around the town or observations from the on-campus student housing options. Our final model listed above is accurate at predicting the price of a housing unit in the town. There is potential for modification but the current model is adequate for use as shown through our rigorous testing and model-building process

**Appendix A:** References

Background:


1. Bunescu, O. (2024, January 31). *Student Housing Trends Shaping the Market in 2024*. Multi-Housing News.
2. *Town Housing Market: House Prices & Trends*. Redfin. (n.d.). https://www.redfin.com/Town/3867/VA/Town/housing-market
3. Sandlow, H. (2023, August 3). *Town Home Prices Hit "All-Time High", Slowing Market*. The Daily Progress. https://dailyprogress.com/news/local/business/real-estate/Town-home-prices-hit-all-time-high-slowing-market/article_e1a4b280-317a-11ee-a7ba-8f27ceb76eb3.html https://www.multihousingnews.com/student-housing-trends-shaping-the-market-in-2024/


Data:


4. Google. (n.d.). Google Maps. https://maps.google.com/
5. *Houses for Rent In Town, VA*. Apartments.com. (n.d.). https://www.apartments.com/houses/Town-va/?bb=moup-h5r3Hnzh2wM
6. *Real Estate, Apartments, Mortgages & Home Values*. Zillow. (n.d.). http://www.zillow.com
7. *Institution Off Campus Student Apartments & Housing - Town, VA One, 2,3 and 4 bedroom apartments*. Lark on Main. (2023, October 17). https://larkonmain.com/


Supplemental Code and Analysis Help:


8. Robert Kabacoff Quick-R: Subsetting Data. www.statmethods.net/management/subset.html.
9. Riederer, Yihui Xie Christophe Dervieux, Emily. 5.4 Control the Size of Plots/Images | R Markdown Cookbook. 29 Feb. 2024, bookdown.org/yihui/rmarkdown-cookbook/figure-size.html.

**Appendix B:**

**Figure A.** Histogram of Response Variable

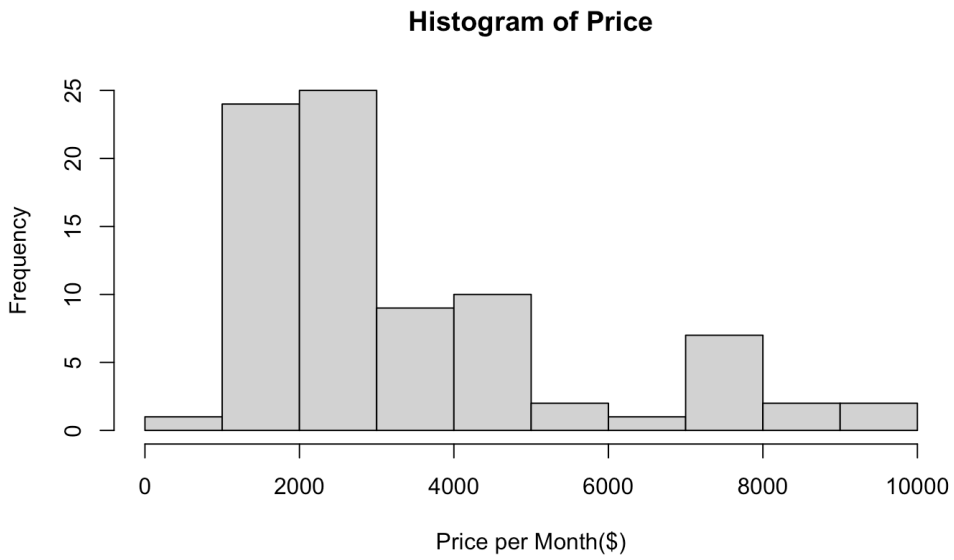


**Figure B.** EDA for Explanatory Variables (quantitative in the scatterplots and qualitative in the boxplots)

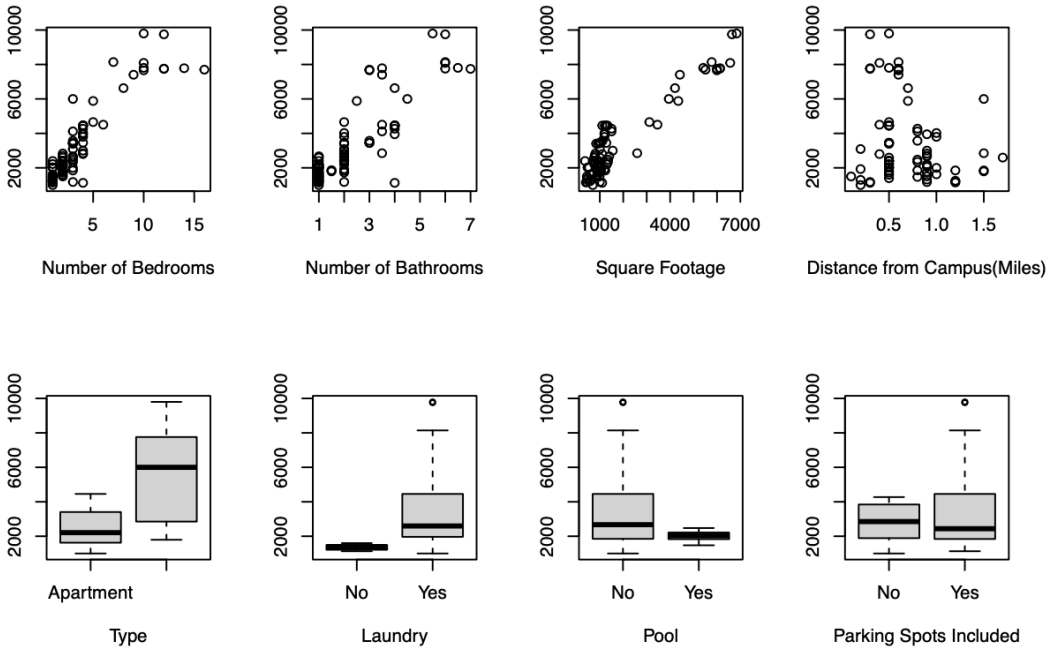

**Figure C.** Interaction Plots

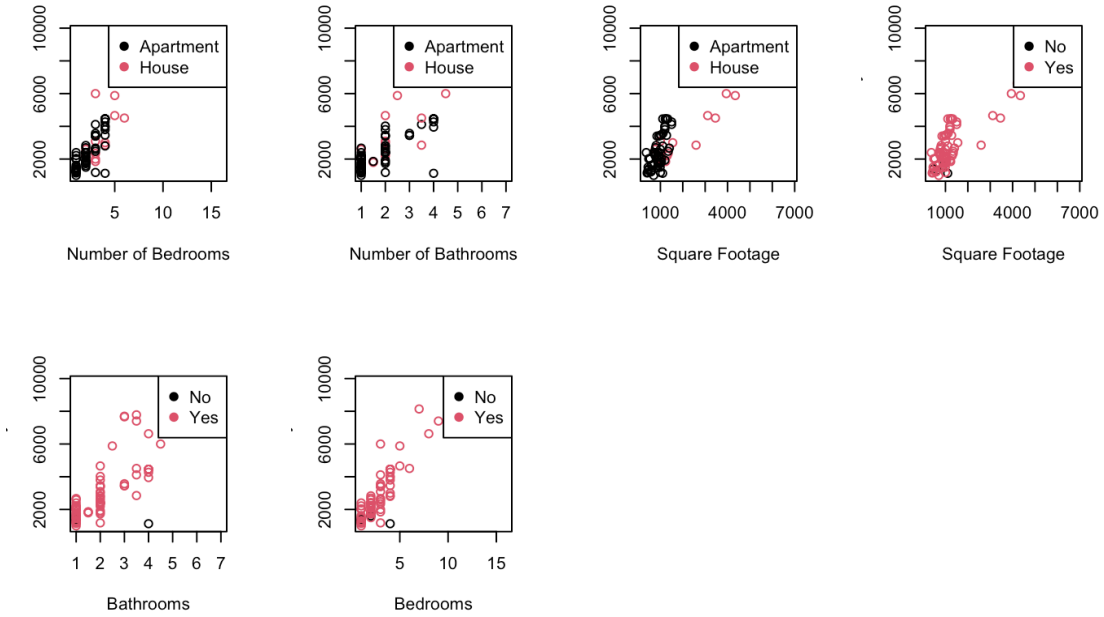


**Figure D.** Multicollinearity Plots

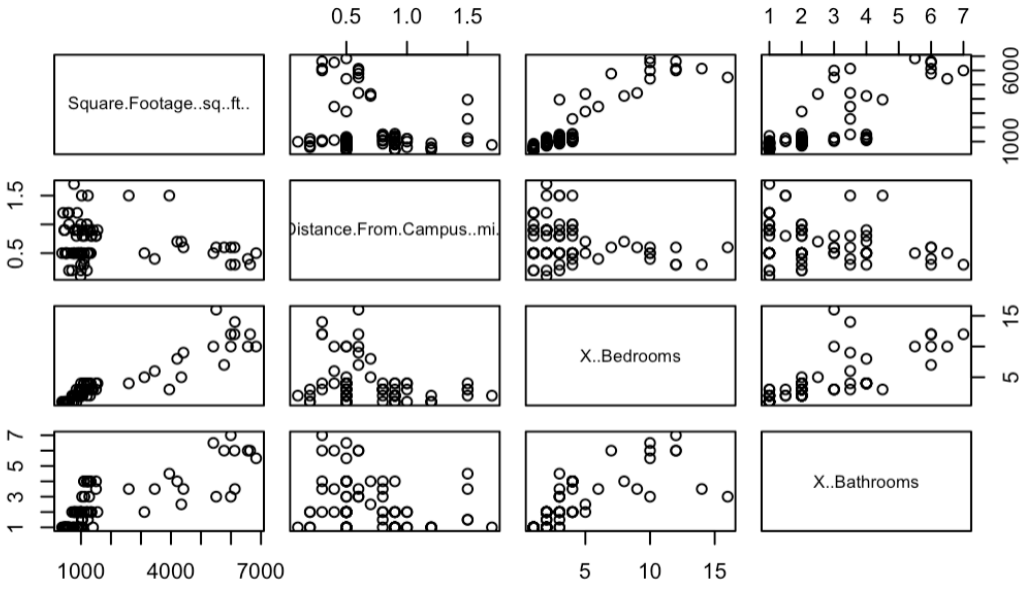

**Figure E**. Residual Assumptions

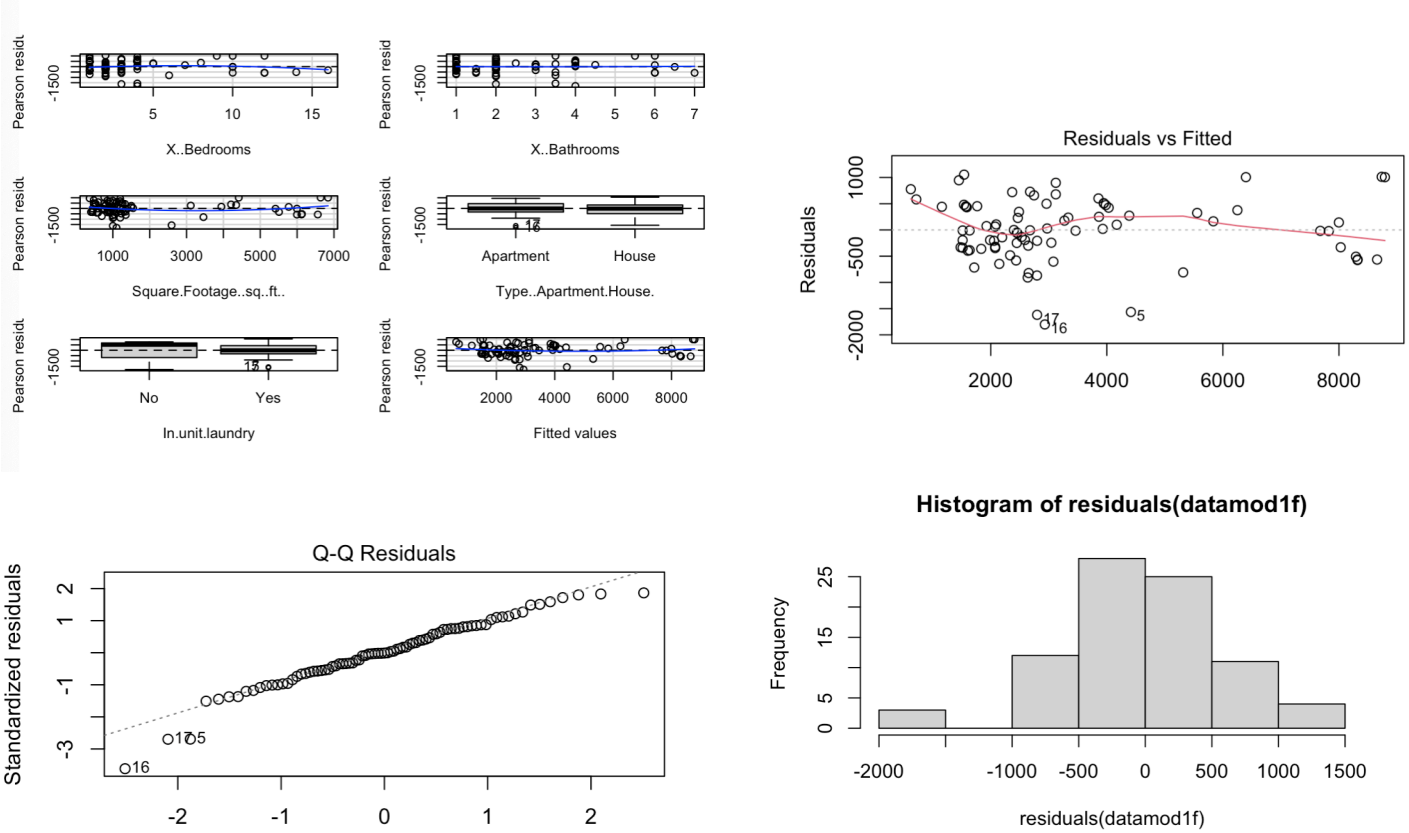


**Figure F.** Influential Observations

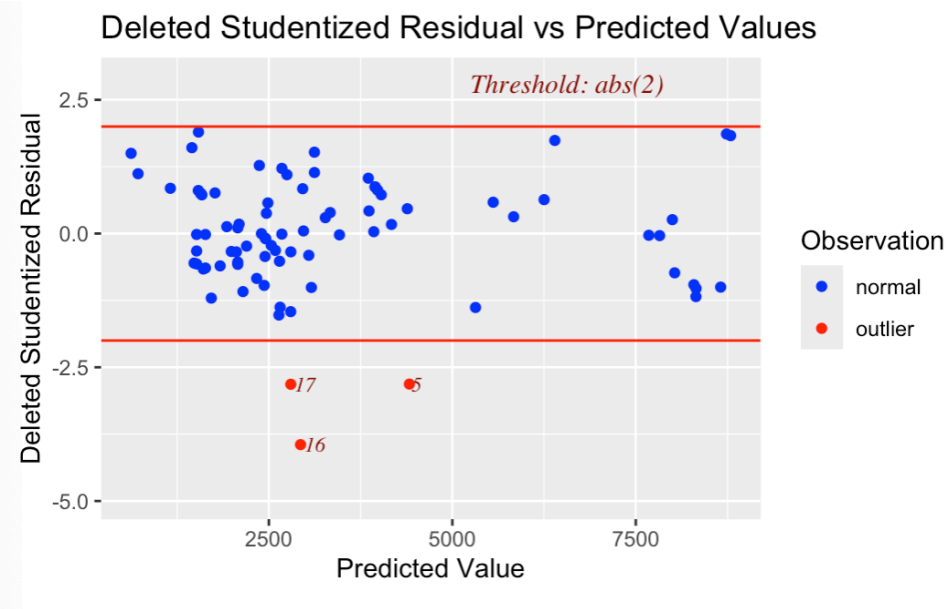

**Figure G.** Stepwise Regression

```
                                    Stepwise Summary
-------------------------------------------------------------------------------------------
--
Step    Variable                          AIC         SBC         SBIC        R2       Adj.
R2
-------------------------------------------------------------------------------------------
--
 0      Base Model                     1517.897    1522.735    1281.020    0.00000
0.00000
 1      Square.Footage..sq..ft.. (+)   1350.389    1357.645    1107.327    0.87027
0.86866
 2      X..Bathrooms (+)               1325.055    1334.730    1072.068    0.90667
0.90433
 3      X..Bedrooms (+)                1320.111    1332.205    1070.528    0.91416
0.91090
 4      X..Bedrooms:X..Bathrooms (+)   1315.538    1330.051    1070.177    0.92069
0.91663
-------------------------------------------------------------------------------------------
--
```

**Figure H.** Final Model F-Test

```
     Min       1Q   Median       3Q      Max
-1643.59  -266.06   -16.28   330.39  1106.28

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -15.9216   341.9596  -0.047 0.962988
X..Bedrooms                 276.0221    58.6296   4.708 1.12e-05 ***
X..Bathrooms                655.5136    89.9487   7.288 2.67e-10 ***
Square.Footage..sq..ft..      0.7346     0.1156   6.353 1.47e-08 ***
Type..Apartment.House.House -412.0681   209.8680  -1.963 0.053300 .
In.unit.laundryYes          234.5429   336.8536   0.696 0.488409
X..Bedrooms:X..Bathrooms    -44.7936    12.9500  -3.459 0.000897 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 555.9 on 75 degrees of freedom
Multiple R-squared:  0.9421,    Adjusted R-squared:  0.9375
F-statistic: 203.4 on 6 and 75 DF,  p-value: < 2.2e-16
```