Spatial and Demographic Effects on Theft Distribution Across Los Angeles

Abstract

Using the full 2020 Los Angeles Police Department crime dataset, we examined how theft risk varied across the city when conditioned on a crime having occurred. After cleaning the data and adding population-based “Area Type” and population per square mile “Area Density” indicators, we compared nested logistic regression models. Results showed that raw population size had no independent effect on theft odds, while higher population density was consistently associated with *lower* odds of a reported crime being theft, suggesting that denser environments may offer stronger informal guardianship. Demographically, each additional decade of victim age increased theft odds by approximately 5%. Incidents involving men, Black, or Hispanic victims were less likely to be theft than those involving women or White victims, whereas Asian victims experienced a higher proportion of theft. Overall, population density was the strongest predictor of Los Angeles theft in 2020, highlighting the value of density-aware models in urban crime prevention.

Introduction

The COVID-19 pandemic brought profound disruptions to daily life, reshaping economic, social, and institutional structures. Lockdowns, job losses, and increased social isolation caused by the pandemic are important factors that led to shifts in crime patterns. While overall crime in the United States declined by 9% during 2020, specific categories of violent and property crime increased in certain regions. For instance, in Los Angeles, homicides rose significantly by 36% in the same year (Haskell, 2021). Property crime also shifted, with vehicle thefts in California rose by 19.6%, with 180,939 vehicles stolen at an estimated total value of $1.6 billion (California Highway Patrol, 2021). These incidents make Los Angeles an important focus for understanding the geography behind rising theft rate. We came up with the research question: *Which areas of Los Angeles are more dangerous in terms of theft, and how does theft risk vary across age, gender, and racial groups?* To answer this, our study applies statistical methods to examine theft patterns across neighborhoods in Los Angeles, aiming to identify how theft was distributed across this area in 2020 and what socioeconomic or geographic factors may explain these trends.

Methods

The dataset that will be used in this project is a collection of crime incidents in the City of Los Angeles from 2020, sourced from Data.gov. When a crime occurs, Los Angeles Police Department (LAPD) documents the incident using standardized paper reports, and these reports are later transcribed into digital format. The original dataset includes 28 variables, but we chose to work with the LAPD divisions where the crime was found, the crime code, and the victim descriptors (Age, Sex, and Descent) for each reported incident. To ensure consistency, instances with missing or uninterpretable age, race or sex were excluded. We divided Victim Descent into five broad groups—White, Black, Hispanic, Asian, and Other Races; Age was treated as a discrete variable, and Sex was classified as male or female. Spatial context was measured using two new variables created from LAPD division data: Area.Type, based on total division population (categorized as rural <100,000; suburban 100,000–200,000; urban >200,000), and Area.Density**,** calculated by dividing population by division area and grouped into low, moderate, high, and very high (thresholds at 6,000, 9,000, and 14,700 people/mile^2) (Appendix, Table 3). Both columns were added to the dataset manually using population and area data from the LAPD Organization Chart (n.d.). We also examined each crime code closely and divided them into 6 main types of crime: Theft, Assault, Burglary, Vandalism, Robbery, Rape, and Others.

To explore potential patterns, we used bivariate data visualizations—conditional bar charts for categorical variables and boxplots for quantitative variables—to examine the relationship between theft and various explanatory variables, including both individual (age, sex, descent) and area-specific variables (population, population density) (Appendix, Figures B–E). Originally, we included two conditional bar charts: one overall chart and one adjusted for population. While we expected population size to directly or inversely relate to theft odds, the pattern did not align with this assumption. As a result, we established a more area-specific variable: population density.

Based on the perceived effect sizes observed in these visualizations, we built logistic regression models by sequentially adding variables from strongest to weakest associations with theft. Greater variation in theft odds across categories indicated stronger association. Using likelihood ratio tests to compare nested models, we assessed whether newly added variables were worthwhile predictors after adjusting for previously included ones. In our models, 'Unpopulated

Areas’, ‘Female’, and ‘White’ served as reference groups (Appendix, Figures F & H). Finally, we interpreted the coefficients to identify factors associated with increased theft odds.

Models

According to crime-opportunity theory, theft risk is influenced by the presence of suitable targets and the absence of capable guardians, both shaped by factors like population size and density (McKee, n.d.). More residents mean more potential targets and offenders, while higher density reflects how closely people live and interact. In geographically large areas, what matters is not just population size, but how concentrated that area is and how frequently they interact. Since LAPD divisions vary widely in size, with some being six times larger than others, areas with similar populations can differ significantly in density. Using both ‘Area.Type’ (population size) and ‘Area.Density’ (crowdedness) in logistic regression models allows us to separate the effects of population size from population density.

The first model only controls for population:

$$\Pr(Theft) = \frac{1}{1+e^{-(\widehat{\beta_0}+\widehat{\beta_1}Vict.Age+\widehat{\beta_2}Male+\widehat{\beta_3}Black+\widehat{\beta_4}\text{ Hispanic}+\widehat{\beta_5}Asian+\widehat{\beta_6}OtherRace+\widehat{\beta_7}Populated+\widehat{\beta_8}VeryPopulated)}}$$

The second model controls for both population and area size (Area.Density):

$$\Pr(Theft) = \frac{1}{1+e^{-(\widehat{\beta_0}+\widehat{\beta_1}Vict.Age+\widehat{\beta_2}Male+\widehat{\beta_3}Black+\widehat{\beta_4}\text{ Hispanic}+\widehat{\beta_5}Asian+\widehat{\beta_6}OtherRace+\widehat{\beta_7}Moderate+\widehat{\beta_8}High+\widehat{\beta_9}VeryHigh)}}$$

Model	$\widehat{\beta_0}$	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$	$\widehat{\beta_4}$	$\widehat{\beta_5}$	$\widehat{\beta_6}$	$\widehat{\beta_7}$	$\widehat{\beta_8}$	$\widehat{\beta_9}$
Population	-0.69*	0.005*	-0.017*	-0.62*	-0.69*	0.21*	-0.1*	-0.005	0.12*	-
Density	-0.56*	0.005*	-0.022*	-0.61*	-0.69*	0.21*	-0.1*	-0.03*	-0.05*	-0.20*

Note that coefficients with p-value <0.001 has 1* and coefficients with p-value >0.05 has 0*.

Results

Exploratory Data Analysis (EDA): The unconditioned bar chart of crimes by LAPD divisions does not reveal many insights other than each area having either theft or assault as their most prominent crime (Appendix, Figure B). After adjusting for population size, theft rates are highest in highly populated urban areas, lower in sparsely populated rural areas, but lowest in populated suburban areas. Though a pattern emerges, there is not a clear direct or inverse relationship (Appendix, Figure C). To address the potential ambiguity population-based area suggest, we also examine spatial effects while controlling for population density. An inverse relationship emerges between population density and theft rates (Appendix, Figure D). Given these observed patterns, we incorporate both categorization of area type (Population and Population Density) into multivariable logistic regression models (Appendix, Figure F and H) to determine whether they remain an independent predictor or prove statistically negligible.

Demographic controls: Given a crime has occurred and other predictors remain constant (Appendix, Figure I), each ten-year increase in victim age raises theft odds by about 5% ($\widehat{\beta_1}$ ≈ 0.0054). Male victims are 2% ($\widehat{\beta_2}$ ≈ -0.020, OR ≈ 0.98) less likely, in terms of odds, to be involved in thefts than female victims. Racial effects are pronounced: Black ($\widehat{\beta_3}$ ≈ -0.61) and Hispanic ($\widehat{\beta_4}$ ≈ -0.69) victims show roughly 45–50% lower odds than Whites, whereas Asian victims ($\widehat{\beta_5}$ ≈ 0.21) show 24% higher odds. However, because the numbers of thefts on a per capita basis of Black and Hispanic are lower than White's, their low theft odds suggest that Black and Hispanic individuals are more likely to be involved in many violent crimes (non-theft crimes). Conversely, the higher theft rate per capita among Asian residents suggests that their overall crime burden is relatively light, meaning violent incidents are less frequent and theft comprises a larger share of the crimes they experience.

Spatial effect: In the population model, the "Populated" category is not statistically significant (p-value = 0.52 > 0.05) when compared to the "Unpopulated" category. By contrast, predictors in the density-based model remain statistically significant. While both models capture victim demographics effects on theft-odds equally well (conditional on a crime having occurred), the density model yields the more dependable set of predictors. This aligns with the EDA finding that theft rate falls as area density increases in Los Angeles.

Discussion

Our analysis shows that, in 2020, Los Angeles theft risk was driven more by population density than by total population. After adjusting for the effects of demographic and geography, theft odds declined as population density increased, reversing the weak U-shaped pattern seen when divisions were grouped only by raw population. Densely populated urban areas had fewer thefts per resident, whereas low-density suburban divisions had the lowest total theft count. These findings align with crime-opportunity theory, which states highly trafficked areas might benefit from stronger guardianship (e.g surveillance, policing), whereas low-density suburbs might be vulnerable to property crime due to lower security.

As for victim demographic, our data covers only crimes that have already happened, so the model asks: *When a crime occurs, which traits increase or decrease the odds of that crime being theft?* Our data shows that older victims are more often involved in theft-related incidents, which may reflect their generally higher wealth and less physically able to deter theft. In contrast, younger individuals experience a higher share of violent crime, possibly due to greater exposure to riskier environments such as nightlife. Women are slightly more likely than men to be victims of theft, which may be linked to carrying handbags or personal items in more accessible ways. Men experience more assaults and robberies, likely reflecting higher engagement in public-risk situations. For Black and Hispanic victims, theft makes up a smaller share of total victimization because a larger portion involves violent crimes, which may reflect broader exposure to high-crime environments or systemic factors such as concentrated disadvantage and aggressive policing.

Despite over a million observations, several important limitations prevent our findings from fully capturing crime in Los Angeles. Our analysis focuses only on theft, chosen for its highest frequency. While useful as a representative property crime, theft is only one part of the broader urban crime landscape and might not accurately reflect the "dangerousness" of areas. If assault or robbery follow different geographic patterns, our conclusions about population density and neighborhood risk could misrepresent the true nature of danger in areas where violent crimes are more common. Future research should extend this density-based approach to other crimes—such as assault, burglary, robbery, and vandalism—using a multinomial model to compare patterns. We also plan to apply a Chi-square Goodness of Fit test to assess whether Los Angeles' theft distribution significantly differs from national trend, helping to determine whether Los Angeles stands out as a particularly dangerous area compared to the U.S.

References

California Highway Patrol. (2021). 2020 California vehicle theft facts. https://www.chp.ca.gov

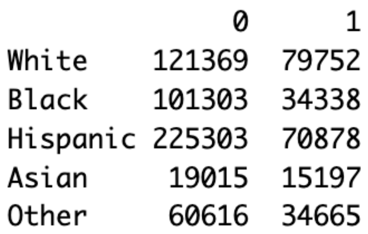
Haskell, J. (2021, January 26). Los Angeles saw overall crime drop in 2020, but homicides increased 36%. ABC7 Los Angeles. https://abc7.com/los-angeles-crime-pandemic-homicides/10012319/

Los Angeles Police Department. (n.d.). *Crime data from 2020 to present*. Data.gov. https://catalog.data.gov/dataset/crime-data-from-2020-to-present (Our dataset was collected from data.gov, based on LAPD report to FBI)

Los Angeles Police Department. (n.d.). *LAPD organization chart*. LAPD Online. https://www.lapdonline.org/lapd-organization-chart/

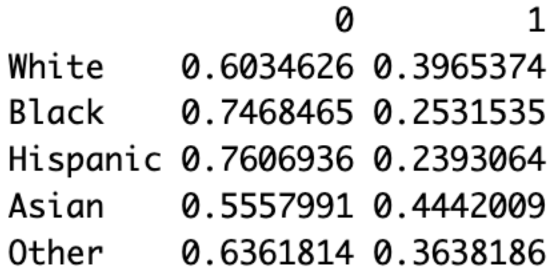
McKee, A. J. (n.d.). *Opportunity theory | Definition*. Doc McKee. Retrieved May 14, 2025, from https://docmckee.com/cj/docs-criminal-justice-glossary/opportunity-theory-definition/

Appendix



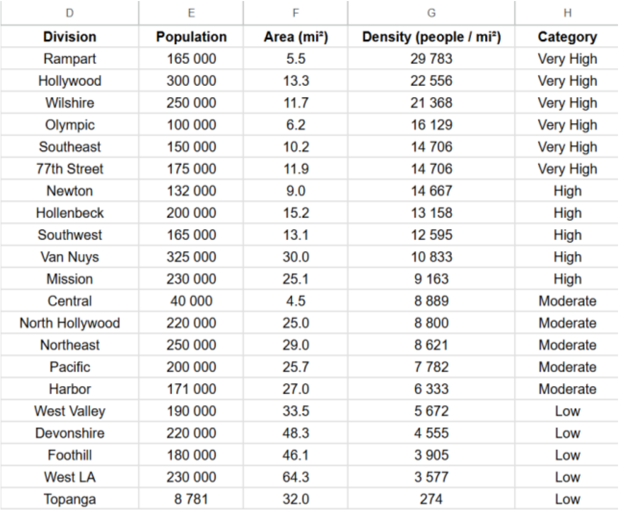
```
              0      1
White    121369  79752
Black    101303  34338
Hispanic 225303  70878
Asian     19015  15197
Other     60616  34665
```

Table 1: Frequency table of violent crimes (0) and theft (1) counts by Victim Descent.



```
                 0         1
White    0.6034626 0.3965374
Black    0.7468465 0.2531535
Hispanic 0.7606936 0.2393064
Asian    0.5557991 0.4442009
Other    0.6361814 0.3638186
```

Table 2: Proportion table of violent crimes (0) and theft (1) by Victim Descent.



D	E	F	G	H
Division	**Population**	**Area (mi²)**	**Density (people / mi²)**	**Category**
Rampart	165 000	5.5	29 783	Very High
Hollywood	300 000	13.3	22 556	Very High
Wilshire	250 000	11.7	21 368	Very High
Olympic	100 000	6.2	16 129	Very High
Southeast	150 000	10.2	14 706	Very High
77th Street	175 000	11.9	14 706	Very High
Newton	132 000	9.0	14 667	High
Hollenbeck	200 000	15.2	13 158	High
Southwest	165 000	13.1	12 595	High
Van Nuys	325 000	30.0	10 833	High
Mission	230 000	25.1	9 163	High
Central	40 000	4.5	8 889	Moderate
North Hollywood	220 000	25.0	8 800	Moderate
Northeast	250 000	29.0	8 621	Moderate
Pacific	200 000	25.7	7 782	Moderate
Harbor	171 000	27.0	6 333	Moderate
West Valley	190 000	33.5	5 672	Low
Devonshire	220 000	48.3	4 555	Low
Foothill	180 000	46.1	3 905	Low
West LA	230 000	64.3	3 577	Low
Topanga	8 781	32.0	274	Low

Table 3: Table showing population count (people) and population density (people/mile²) in each of the 21 LAPD areas.

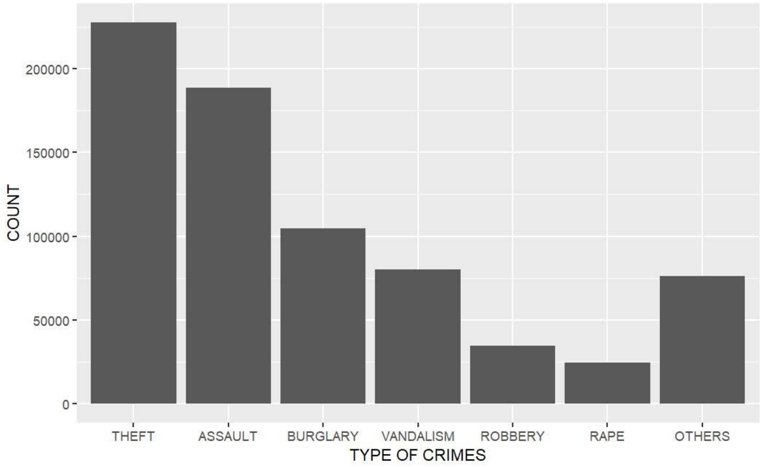


Figure A: The number of different types of crime across Los Angeles

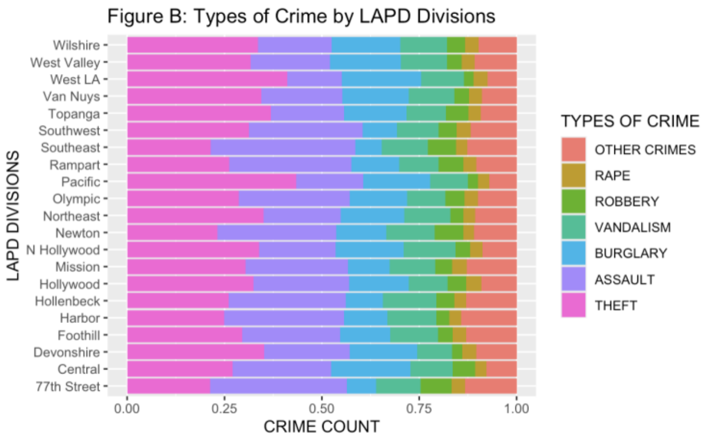


Figure B: Conditional bar chart of crime proportion across 21 LAPD areas

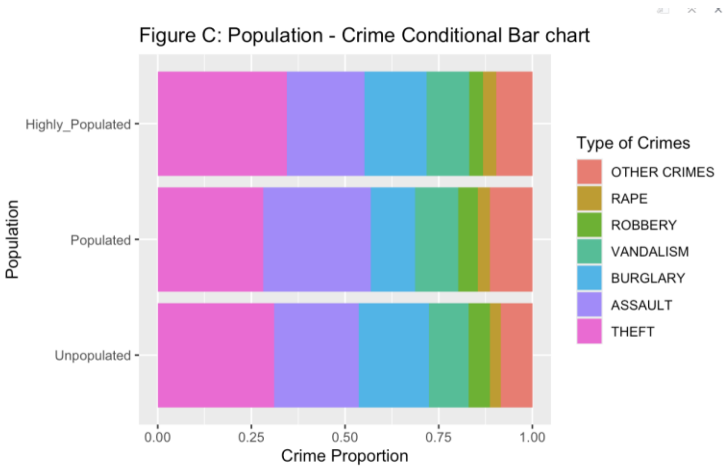


Figure C: Conditional bar chart of crime proportion and population count

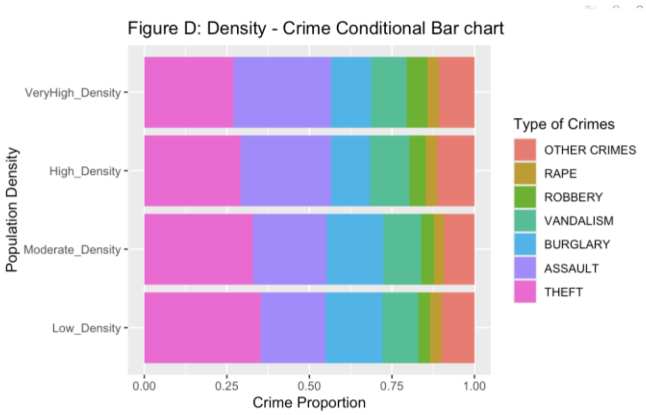


Figure D: Conditional bar chart of crime proportion and population density

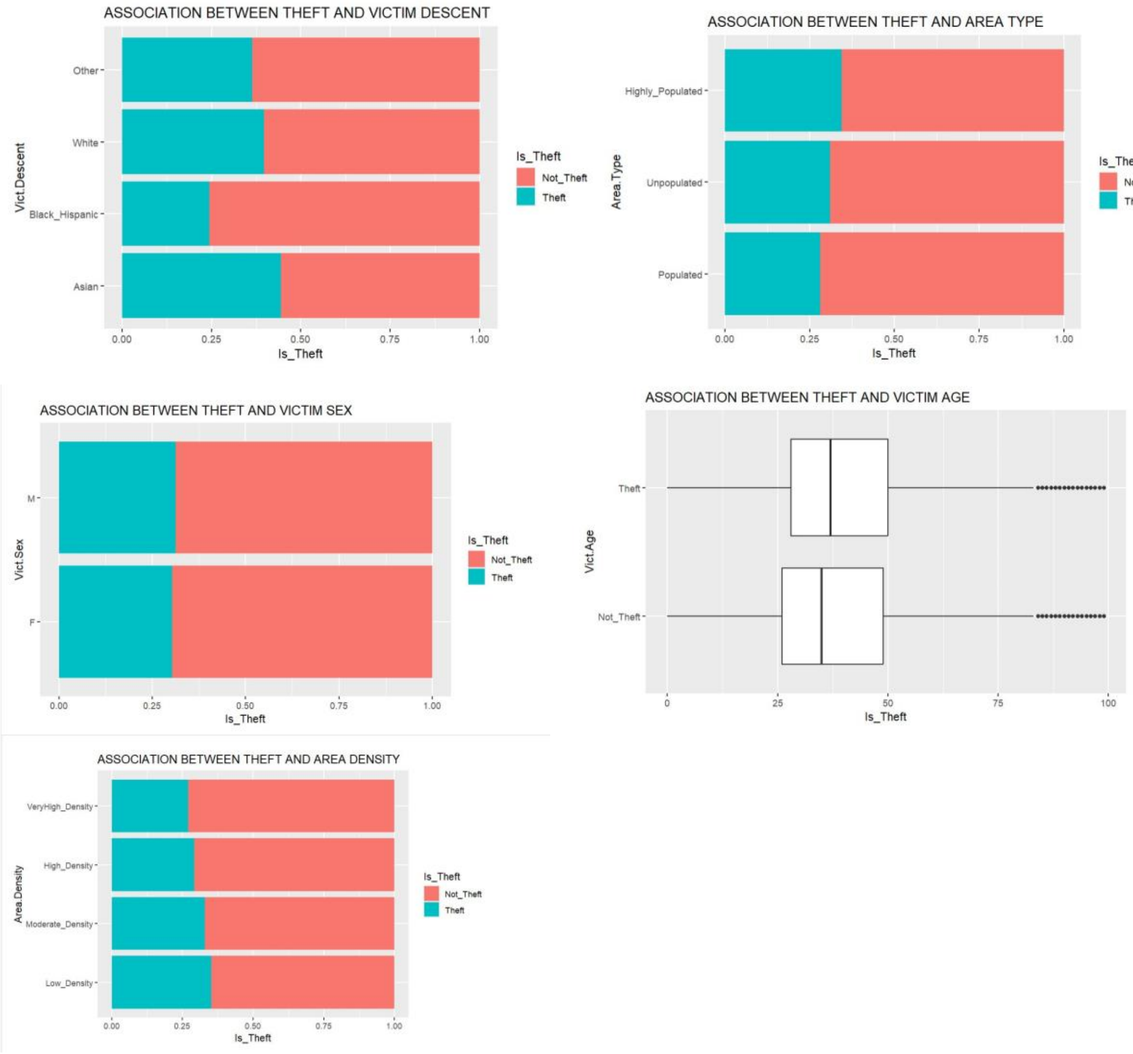
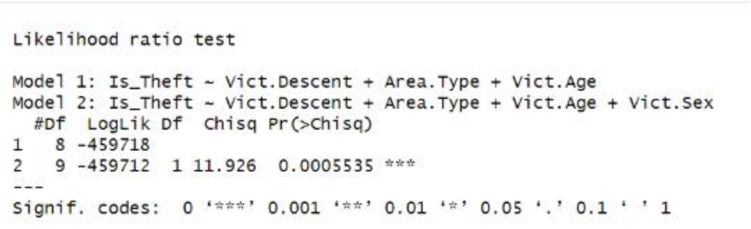
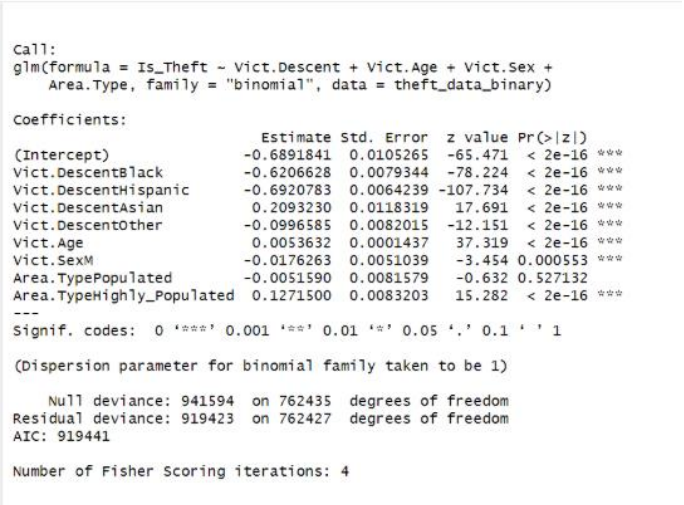


Figure E: Conditional bar charts and box plots showing association between Theft and Vict.Descent, Area.Type, Vict.Sex, Vict.Age and Area.Density



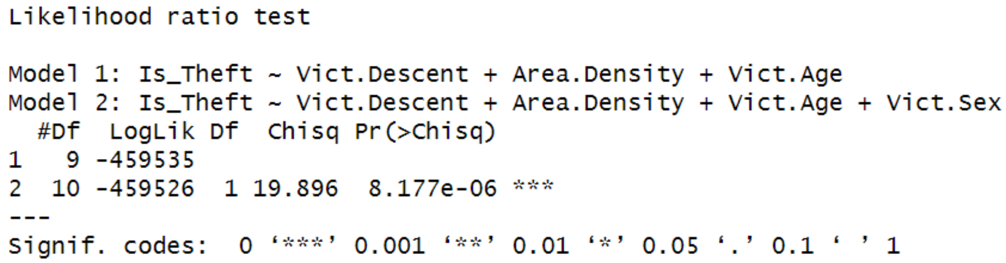
```
Likelihood ratio test

Model 1: Is_Theft ~ Vict.Descent + Area.Type + Vict.Age
Model 2: Is_Theft ~ Vict.Descent + Area.Type + Vict.Age + Vict.Sex
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   8 -459718
2   9 -459712  1 11.926  0.0005535 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure F: Logistic regression model with Area.Type. Previous nested models were compared using Likelihood Ratio Test



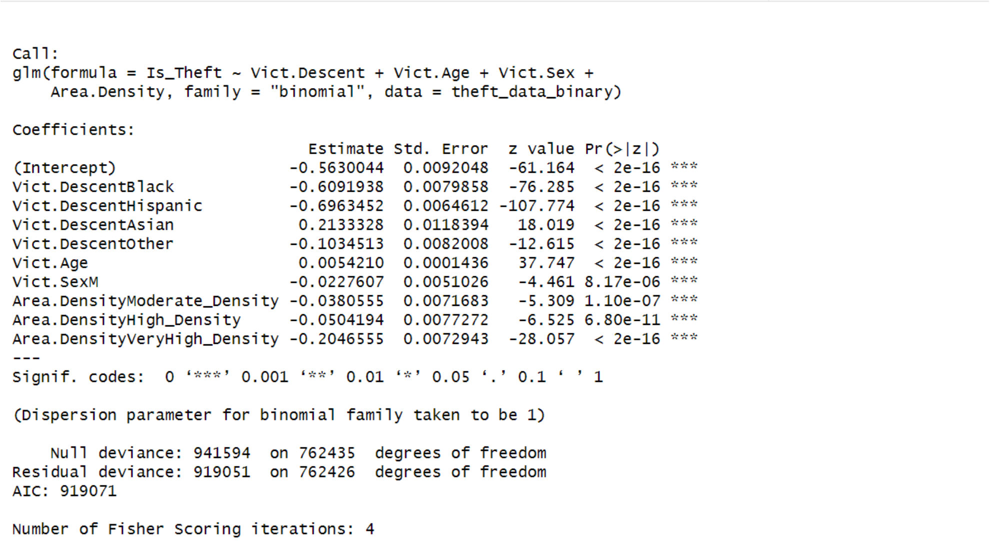
```
Call:
glm(formula = Is_Theft ~ Vict.Descent + Vict.Age + Vict.Sex + 
    Area.Type, family = "binomial", data = theft_data_binary)

Coefficients:
                         Estimate Std. Error  z value Pr(>|z|)    
(Intercept)            -0.6891841  0.0105265  -65.471  < 2e-16 ***
Vict.DescentBlack      -0.6206628  0.0079344  -78.224  < 2e-16 ***
Vict.DescentHispanic   -0.6920783  0.0064239 -107.734  < 2e-16 ***
Vict.DescentAsian       0.2093230  0.0118319   17.691  < 2e-16 ***
Vict.DescentOther      -0.0996585  0.0082015  -12.151  < 2e-16 ***
Vict.Age                0.0053632  0.0001437   37.319  < 2e-16 ***
Vict.SexM              -0.0176263  0.0051039   -3.454 0.000553 ***
Area.TypePopulated     -0.0051590  0.0081579   -0.632 0.527132    
Area.TypeHighly_Populated  0.1271500  0.0083203   15.282  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 941594  on 762435  degrees of freedom
Residual deviance: 919423  on 762427  degrees of freedom
AIC: 919441

Number of Fisher Scoring iterations: 4
```

Figure G: Table of coefficients for the logistic regression model with Area.Type



```
Likelihood ratio test

Model 1: Is_Theft ~ Vict.Descent + Area.Density + Vict.Age
Model 2: Is_Theft ~ Vict.Descent + Area.Density + Vict.Age + Vict.Sex
  #Df  LogLik Df  Chisq Pr(>Chisq)    
1   9 -459535                         
2  10 -459526  1 19.896  8.177e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure H: Logistic regression model with Area.Density. Previous nested models were compared using Likelihood Ratio Test



```
Call:
glm(formula = Is_Theft ~ Vict.Descent + Vict.Age + Vict.Sex + 
    Area.Density, family = "binomial", data = theft_data_binary)

Coefficients:
                              Estimate Std. Error  z value Pr(>|z|)    
(Intercept)                 -0.5630044  0.0092048  -61.164  < 2e-16 ***
Vict.DescentBlack           -0.6091938  0.0079858  -76.285  < 2e-16 ***
Vict.DescentHispanic        -0.6963452  0.0064612 -107.774  < 2e-16 ***
Vict.DescentAsian            0.2133328  0.0118394   18.019  < 2e-16 ***
Vict.DescentOther           -0.1034513  0.0082008  -12.615  < 2e-16 ***
Vict.Age                     0.0054210  0.0001436   37.747  < 2e-16 ***
Vict.SexM                   -0.0227607  0.0051026   -4.461 8.17e-06 ***
Area.DensityModerate_Density -0.0380555  0.0071683   -5.309 1.10e-07 ***
Area.DensityHigh_Density    -0.0504194  0.0077272   -6.525 6.80e-11 ***
Area.DensityVeryHigh_Density -0.2046555  0.0072943  -28.057  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 941594  on 762435  degrees of freedom
Residual deviance: 919051  on 762426  degrees of freedom
AIC: 919071

Number of Fisher Scoring iterations: 4
```

Figure I: Table of coefficients for the logistic regression model with Area.Density