Data Analysis

1. Household level
    a. Chicago

```
> summary(chicago_reg)

Call:
lm(formula = PRICE ~ SQFT + house.age + PROP_TYPE + PARK_DIST + 
    HOSPITAL_D + RAIL_DIST + BUS_COUNTS + FnB_COUNTS + population.density + 
    median.age + dominant.race, data = chi_household)

Residuals:
    Min      1Q  Median      3Q     Max 
-395887  -96989  -11338   64367  939896 

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      -88306.341  85488.439  -1.033  0.30182    
SQFT                                101.181      9.656  10.479  < 2e-16 ***
house.age                         -1589.477    135.352 -11.743  < 2e-16 ***
PROP_TYPEMulti-Family            227334.455  16503.246  13.775  < 2e-16 ***
PROP_TYPERanch                   -81323.484 164808.905  -0.493  0.62179    
PROP_TYPESingle Family Residential 169641.900  15387.768  11.024  < 2e-16 ***
PROP_TYPETownhouse               133328.299  27863.329   4.785 1.91e-06 ***
PARK_DIST                           -53.728     18.656  -2.880  0.00404 ** 
HOSPITAL_D                          -2.041      4.218  -0.484  0.62854    
RAIL_DIST                          -18.063      3.780  -4.779 1.97e-06 ***
BUS_COUNTS                        -292.862    478.889  -0.612  0.54095    
FnB_COUNTS                        4963.329    569.943   8.708  < 2e-16 ***
population.density                  10.166      1.461   6.958 5.51e-12 ***
median.age                         586.135   1520.519   0.385  0.69994    
dominant.raceBlack               62479.406  47629.252   1.312  0.18983    
dominant.raceHispanic or Latino  48241.227  52738.933   0.915  0.36051    
dominant.raceWhite               56470.428  44998.794   1.255  0.20974    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 163600 on 1264 degrees of freedom
Multiple R-squared:  0.4116,    Adjusted R-squared:  0.4041 
F-statistic: 55.26 on 16 and 1264 DF,  p-value: < 2.2e-16
```

Regression Analysis
Coefficients and their Signs:

- SQFT (Property size)**: The positive coefficient (101.181) indicates that an increase in square footage is associated with an increase in housing price. This is expected since larger properties typically command higher prices.
- house.age (Property age)**: The negative coefficient (-1589.477) suggests that older properties are associated with lower prices, possibly reflecting the desirability of newer homes.
- PROP_TYPE (Property type)**: Different property types have varying impacts on price. Multi-family and single-family residential types are associated with higher prices, while ranch types do not significantly affect the price. The positive sign for single-family and multi-family suggests these are more valued in the market.
- PARK_DIST (Distance from park)**: The negative coefficient (-53.728) indicates that closer proximity to parks might increase housing prices, which aligns with the general preference for park access.
- HOSPITAL_D (Distance from hospital)**: The negative and non-significant coefficient suggests that proximity to hospitals does not have a clear impact on housing prices in this dataset.
- RAIL_DIST (Distance from train)**: A negative coefficient (-2.041) with statistical significance suggests that properties closer to train stations are priced higher, reflecting the value of transportation accessibility.

- BUS_COUNTS (Number of bus stops)**: The negative and non-significant coefficient suggests that the number of bus stops nearby does not have a strong impact on housing prices.
- FnB_COUNTS (Number of food and beverage stores)**: The positive coefficient (4963.329) is significant, indicating that areas with more food and beverage outlets have higher housing prices, likely reflecting the attractiveness of local amenities.

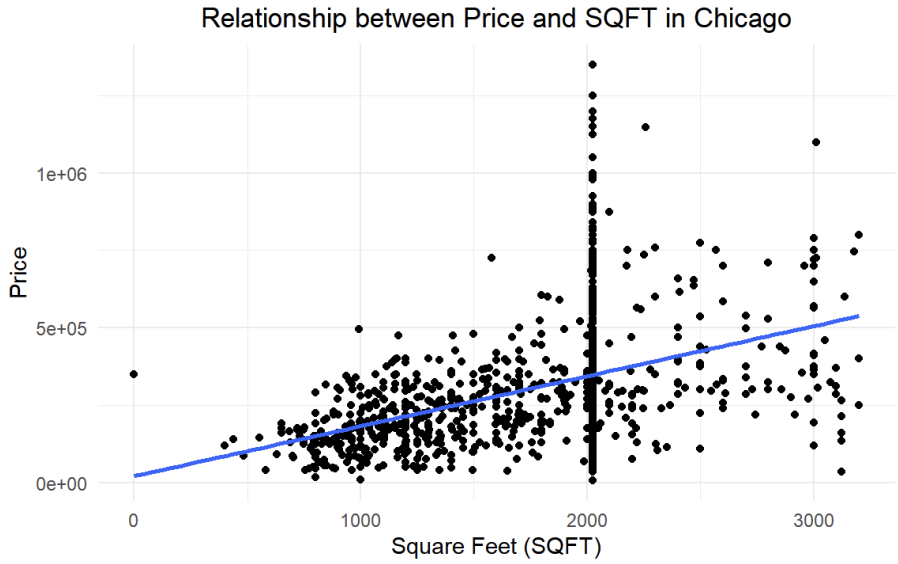
P-values:
Variables with p-values less than 0.05 are considered statistically significant. In this model, the significant predictors of housing price at the 5% level or better include the square footage of the property, property age, multi-family property type, single-family residential property type, distance from the nearest park, distance from the nearest train station, and the number of food and beverage stores. The other variables are not significant predictors at the 5% level.

R-squared:
The Multiple R-squared of 0.4116 suggests that about 41.16% of the variability in housing prices is explained by the model. The Adjusted R-squared of 0.4041 accounts for the number of predictors and indicates a slight decrease in explanatory power when considering the number of variables used.

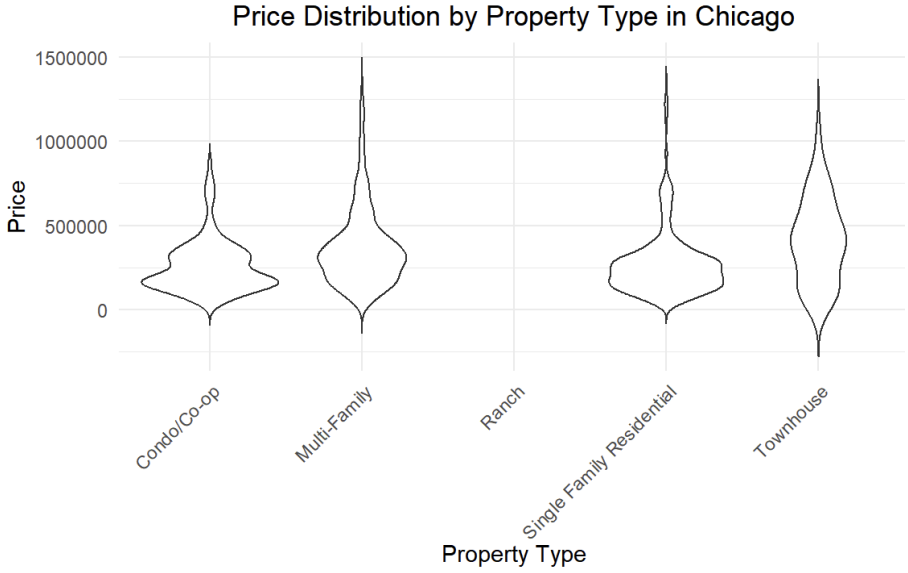
F-test:
The F-statistic of 55.26 is highly significant ($p < 0.001$), indicating that the model overall is a good fit and that there is a collective effect of the independent variables on the housing price.

Summary:
The regression analysis reveals several key insights into the housing market of the area studied:
- Size Matters: The square footage of a property is a significant predictor of its price, with larger properties commanding higher prices.
- Age Penalty: Older properties tend to be cheaper, likely reflecting consumer preference for newer homes.
- Type and Location: Single-family and multi-family residences are priced higher, indicating a market preference for these types of properties. Proximity to parks and train stations also increases property values, highlighting the importance of accessibility and green spaces.
- Amenities Attract: The presence of food and beverage outlets is a significant positive predictor of housing prices, which could reflect the desirability of living in vibrant neighborhoods with ample amenities.
- Incomplete Picture: While the model explains a significant portion of the variance in housing prices, there is still a substantial amount that is unexplained, pointing to other factors not included in the model that may influence prices.

Inferential Visualization



This scatter plot illustrates the relationship between the size of a property (in square feet) and its price. A positive trend is evident; as the square footage increases, so does the property's price, which is depicted by the upward slope of the blue regression line. The plot shows a wide dispersion of data points, particularly for properties with larger square footage, indicating variability in price that the square footage alone does not explain. This spread suggests that while size is an important factor in determining property prices, there are other influential factors not captured in this two-dimensional analysis. The dense clustering of points at the lower end of the square footage range indicates a concentration of data for smaller properties, which is typical in urban real estate markets. Overall, the visualization confirms the expected positive correlation between property size and price but also underscores the complexity of real estate pricing.

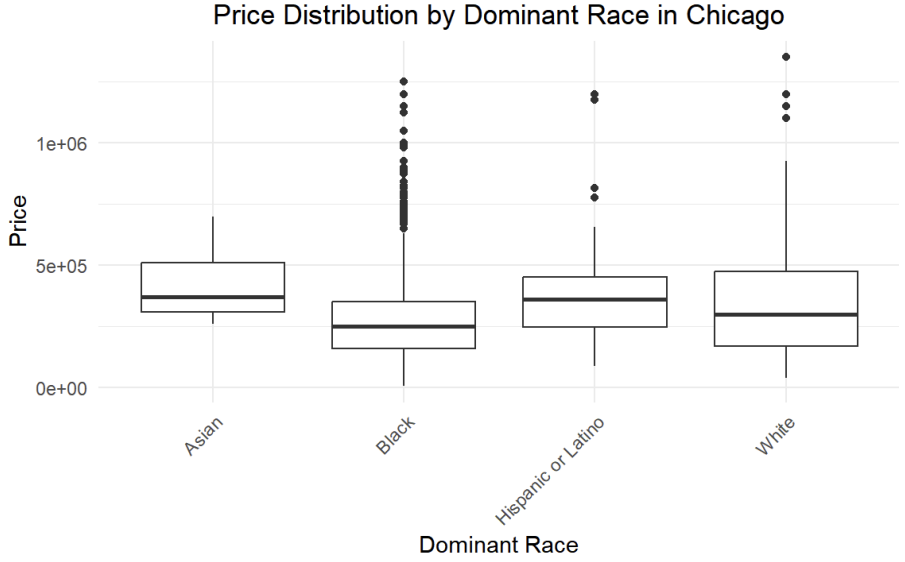


The visualization presented is a violin plot that displays the distribution of property prices across different property types: Condo/Co-op, Multi-Family, Ranch, Single Family Residential, and

Townhouse. Each "violin" represents the price distribution for a property type, with the width of each section indicating the frequency of price points at that level. The points at which the violins are widest suggest the most common price ranges for each property type.

From this plot, we can observe variation in price distributions among the different types of properties. Some property types show a wider range of prices, indicated by the longer tails of the violins, whereas others have a more concentrated price distribution, indicated by the narrower or more bulging sections of the violins.

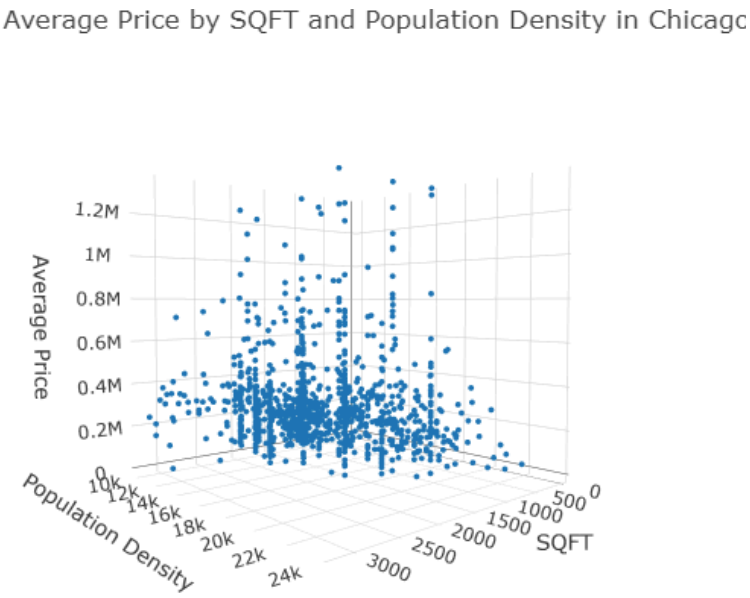
The plot is particularly useful for comparing the median prices (visible as the thickest part of each violin), the variability, and the presence of potential outliers in price within each property type category. The distribution also appears to be multimodal for some property types, suggesting the presence of subgroups within those categories with distinct price points. This type of plot provides a comprehensive view of property prices, allowing for a nuanced understanding of how prices are distributed across different types of real estate.



The box plot visualizes the distribution of property prices in Chicago, segmented by the dominant racial demographic in various neighborhoods: Asian, Black, Hispanic or Latino, and White. Each box represents the interquartile range (IQR) of property prices for neighborhoods with a majority of the indicated race, with the median price marked by a line within the box.

The plot indicates that median property prices do not vary drastically across neighborhoods dominated by different races. However, the range of prices (as indicated by the whiskers of the box plot) and the presence of outliers (individual points above and below the boxes) suggest there is substantial variation in property prices within each racial demographic group. The 'Black' category exhibits a significant number of outliers on the higher price end, indicating that while the median price might be lower, there are properties that are much higher in price within these neighborhoods.

The visualization reflects the complexity of property values within urban environments, where a multitude of factors in addition to racial demographics can influence prices. It also underscores the existence of both affordable and high-end property segments within each dominant racial demographic in Chicago.



The 3D scatter plot represents the relationship between property size (SQFT), population density, and average property prices in Chicago. The x-axis depicts the square footage of properties, the y-axis shows the population density of the area, and the z-axis represents the average price of properties.

The distribution of data points suggests that both larger property sizes and higher population densities are associated with a range of average prices. There is a visible concentration of data points at lower square footages and population densities, possibly indicating a cluster of more affordable and smaller properties in less dense areas. As the square footage and population density increase, there seems to be an upward trend in property prices, with some properties in high-density areas reaching the upper price echelons.

The 3D aspect of the plot allows for the simultaneous visualization of these three dimensions, offering a more nuanced view of the real estate market in Chicago. It implies that while larger properties and higher population densities might generally command higher prices, there is considerable variability, reflecting the complex dynamics of urban property valuation.

b. Los Angeles City

```
> summary(la_reg)

Call:
lm(formula = PRICE ~ SQFT + house.age + PROP_TYPE + PARK_DIST +
    HOSPITAL_D + RAIL_DIST + BUS_COUNTS + FnB_COUNTS + population.density +
    median.age + dominant.race, data = la_household)

Residuals:
     Min       1Q   Median       3Q      Max
-1182194  -416158   -57800   258837  3771602

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -2.603e+06  8.182e+05  -3.182 0.001617 **
SQFT                             4.591e+02  3.704e+01  12.393  < 2e-16 ***
house.age                       -4.232e+03  1.490e+03  -2.840 0.004817 **
PROP_TYPEMulti-Family            2.424e+05  1.820e+05   1.332 0.183957
PROP_TYPESingle Family Residential  8.885e+05  1.524e+05   5.830 1.42e-08 ***
PROP_TYPETownhouse              -7.317e+03  2.521e+05  -0.029 0.976865
PROP_TYPEVacant Land            -5.271e+05  1.926e+05  -2.736 0.006587 **
PARK_DIST                        2.673e+01  4.298e+01   0.622 0.534518
HOSPITAL_D                      -8.333e+01  3.379e+01  -2.467 0.014198 *
RAIL_DIST                        5.137e+01  4.757e+01   1.080 0.281038
BUS_COUNTS                      -8.011e+03  4.808e+03  -1.666 0.096702 .
FnB_COUNTS                       1.150e+04  7.143e+03   1.609 0.108572
population.density              -1.779e+01  4.970e+00  -3.579 0.000402 ***
median.age                       8.501e+04  2.166e+04   3.925 0.000107 ***
dominant.raceHispanic or Latino  2.980e+05  3.669e+05   0.812 0.417376
dominant.raceWhite               8.603e+05  3.793e+05   2.268 0.024043 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 609200 on 302 degrees of freedom
Multiple R-squared:  0.6154,    Adjusted R-squared:  0.5963
F-statistic: 32.22 on 15 and 302 DF,  p-value: < 2.2e-16
```

Regression Analysis

Coefficients and their Signs:

- SQFT (Property size): The coefficient is positive (459.1) and highly significant (p < 2e-16), indicating that larger properties tend to have higher prices, as expected in real estate markets.
- house.age (Property age): The coefficient is negative (-4232) and significant (p = 0.004817), suggesting that older houses tend to be less expensive, which might reflect the costs associated with updating and maintenance.
- PROP_TYPE (Property type): Multi-Family homes have a positive but not statistically significant coefficient, indicating no clear evidence that multi-family properties are priced differently from the baseline property type. Single Family Residential has a significant positive coefficient, indicating they are valued higher, whereas Vacant Land has a significant negative effect, likely due to the additional costs and effort required to develop these properties.
- PARK_DIST (Distance from park): The coefficient is negative but not significant, suggesting that proximity to parks does not have a statistically discernible impact on housing prices in this model.
- HOSPITAL_D (Distance from hospital): The negative coefficient is significant (p = 0.014198), indicating that properties further away from hospitals may be less expensive, which could reflect the value of healthcare accessibility.
- RAIL_DIST (Distance from train): The coefficient is positive but not significant, suggesting that in this dataset, proximity to train stations does not have a strong and statistically significant impact on housing prices.

- BUS_COUNTS (Number of bus stops): The coefficient is positive but not statistically significant, indicating that the number of bus stops in the vicinity is not a clear predictor of housing prices.
- FnB_COUNTS (Number of food and beverage stores): The coefficient is positive but not statistically significant, suggesting that the presence of food and beverage stores in the area is not a strong predictor of housing prices.
- population.density: The coefficient is negative and highly significant ($p < 0.001$), which suggests that higher population densities may be associated with lower housing prices, perhaps due to factors such as noise and congestion.
- median.age: The positive and significant coefficient indicates that areas with an older median age tend to have higher housing prices, which could reflect a more established, possibly affluent community.

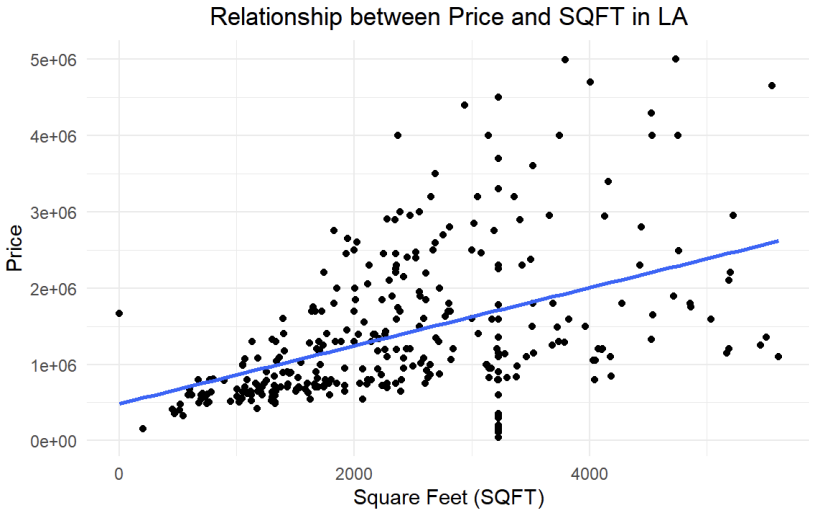
P-values:
A variable with a p-value less than 0.05 is typically considered to have a statistically significant association with the dependent variable. Here, SQFT, house.age, Single Family Residential property type, Vacant Land, HOSPITAL_D, and population density are significant.

R-squared:
- The Multiple R-squared value of 0.6154 suggests that approximately 61.54% of the variation in housing prices can be explained by the model, which is fairly strong.
- The Adjusted R-squared value of 0.5963 adjusts for the number of variables and is still reasonably high, indicating that the model explains a good portion of the variance in housing prices without being overly complex.

F-test:
The F-statistic is significant ($p < 2.2e-16$), indicating that the model is overall a good fit for the data.

Summary:
This regression model for the Los Angeles housing market suggests that property size and age are important determinants of housing prices, with larger and newer properties commanding higher prices. Single Family Residential properties are particularly valued, while Vacant Land is less desirable. The negative association with population density might reflect preferences for less crowded neighborhoods or the influence of other unaccounted variables like noise or traffic. The significance of the median age variable suggests a potential preference for established neighborhoods. Despite the inclusion of various property and neighborhood attributes, not all are significant predictors, and other factors not included in the model may also play a role in housing prices. It's also essential to consider that the significant F-statistic indicates that the model as a whole is statistically significant and likely useful for predicting housing prices in Los Angeles.

Inferential Visualization

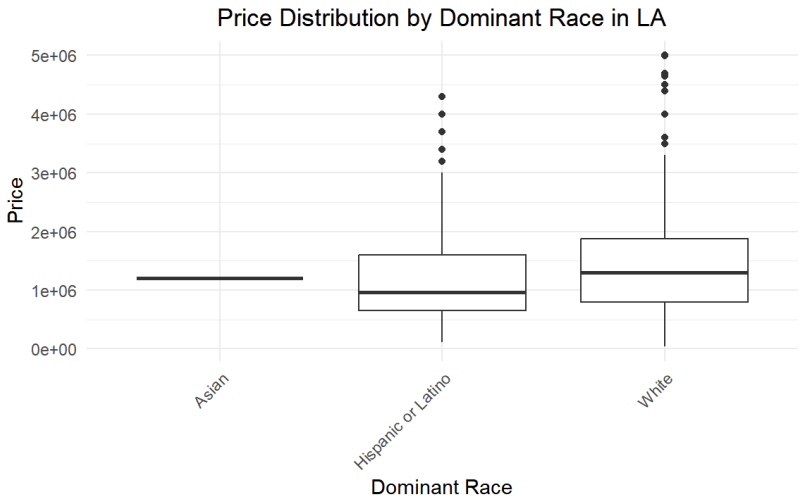


This scatter plot depicts the relationship between a property's size, in square feet (SQFT), and its price. The data points are spread across the plot, with a positive trend indicated by the blue line, suggesting that as a property's size increases, its price tends to rise as well. This positive correlation is a common trend in real estate, where larger properties are generally more expensive. The distribution of points is densest at the lower end of the size and price scales, showing that smaller, less expensive properties are more common in the dataset. Some properties with larger square footage are priced higher, but there is considerable scatter among these, indicating that factors other than size are influencing the price for these properties. The trend line, representing the average increase in price with size, captures the general upward trajectory, but the variation around this line reflects the diverse factors that can affect property prices beyond just the size.

This violin plot provides a comparative visualization of the distribution of property prices by property type in Los Angeles. Each 'violin' shape reflects the range and density of prices for each category: Condo/Co-op, Multi-Family, Single Family Residential, Townhouse, and Vacant Land. The width of each violin indicates the prevalence of sales at different price points, with the thickest parts of the violins representing the most common prices within that property type.

From the distribution patterns, Single Family Residential properties exhibit the widest price range, indicating significant variability in prices, likely due to a diverse array of factors such as location, size, and amenities. Conversely, the narrower profiles of the Condo/Co-op and Townhouse violins suggest a more homogeneous pricing structure within these property types. Vacant Land shows less variability in price, but its distribution suggests a peak in frequency at lower price points, which could indicate that such properties are generally more affordable or perhaps smaller in size.

Overall, the visualization encapsulates the complex landscape of the Los Angeles real estate market. It highlights not just the typical price ranges one might expect for each property type but also underscores the nuances within each category. For instance, the sharp peaks and long tails in the price distributions for Single Family Residential and Multi-Family properties hint at a market that caters to a wide spectrum of buyers, from those seeking affordability to those demanding luxury.

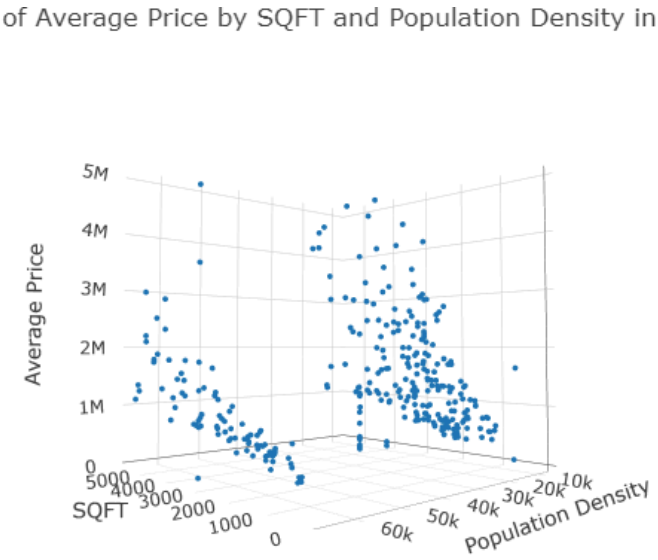


The box plot provided shows the distribution of property prices in Los Angeles, categorized by the dominant racial demographic of the neighborhood: Asian, Hispanic or Latino, and White. Each box represents the interquartile range of prices for properties in neighborhoods predominantly of the specified race, with the median price indicated by the line within each box.

We can observe that neighborhoods with a majority of White residents tend to have a higher median property price compared to those with a majority of Asian or Hispanic or Latino residents. The range of prices, as indicated by the span of the boxes and the whiskers, is wide across all categories, suggesting a substantial variability within each racial demographic. Outliers, as shown by the individual dots above the boxes, indicate that there are properties priced well above the upper quartile in neighborhoods identified with each racial group, with the White category showing several high-priced outliers.

This visualization provides a snapshot of property price disparities across different racial demographics in Los Angeles, hinting at underlying social and economic factors that might influence real estate values. The presence of outliers across all categories illustrates that within

each racial demographic, there can be significant deviations from the typical property price range.



Data points in the 3D space likely show that while there is a general upward trend in prices with larger property sizes and greater density, there is considerable variance. Some properties with extensive square footage or located in highly dense areas may not align perfectly with this trend, showing that other factors could influence pricing.

The z-axis, showing the average price, underscores the vertical distribution of property values. In regions with dense populations, the range of prices could be more extensive, indicating a mix of affordable and high-end living spaces. In contrast, properties with less square footage might cluster at lower price points but could also show exceptions with some properties exceeding expected price ranges.

Overall, such a 3D plot would underscore the dynamic interplay between size, density, and price in LA's diverse real estate market, highlighting how spatial and demographic factors intertwine to shape property valuations.

c. New York City

```
> summary(nyc_reg)

Call:
lm(formula = PRICE ~ SQFT + house.age + PROP_TYPE + PARK_DIST + 
    HOSPITAL_D + RAIL_DIST + BUS_COUNTS + FnB_COUNTS + population.density + 
    median.age + dominant.race, data = nyc_household)

Residuals:
    Min      1Q  Median      3Q     Max 
-687911 -134180  -27556   89337 1397311 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     -6.949e+05  4.137e+05  -1.680 0.093659 .  
SQFT                             1.897e+02  2.863e+01   6.628 8.83e-11 ***
house.age                        7.464e+02  4.381e+02   1.704 0.089076 .  
PROP_TYPEMulti-Family            3.277e+05  7.163e+04   4.575 6.02e-06 ***
PROP_TYPESingle Family Residential  1.588e+05  6.795e+04   2.337 0.019827 *  
PROP_TYPETownhouse               2.417e+05  8.874e+04   2.724 0.006677 ** 
PROP_TYPEVacant Land             1.186e+05  8.130e+04   1.458 0.145360    
PARK_DIST                       -3.241e+01  2.370e+01  -1.367 0.172183    
HOSPITAL_D                      -3.635e+00  1.408e+01  -0.258 0.796412    
RAIL_DIST                        3.754e+01  1.796e+01   2.091 0.037040 *  
BUS_COUNTS                       5.539e+03  2.570e+03   2.155 0.031625 *  
FnB_COUNTS                       6.178e+03  1.783e+03   3.465 0.000576 ***
population.density              -5.227e+00  1.031e+00  -5.071 5.60e-07 ***
median.age                       2.607e+04  1.001e+04   2.604 0.009481 ** 
dominant.raceHispanic or Latino  9.967e+04  1.133e+05   0.880 0.379352    
dominant.raceWhite               5.008e+05  1.045e+05   4.791 2.19e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 256600 on 500 degrees of freedom
  (因为不存在，3个观察量被删除了)
Multiple R-squared:  0.5259,    Adjusted R-squared:  0.5117 
F-statistic: 36.98 on 15 and 500 DF,  p-value: < 2.2e-16
```

Regression Analysis

Coefficients and their Signs:

- SQFT (Property size): With a positive coefficient (189.7) and a highly significant p-value (< 2e-16), the size of the property is strongly associated with an increase in housing prices.
- house.age (Property age): The positive coefficient (746.4) is not statistically significant (p-value = 0.089), suggesting that there may be a slight tendency for older houses to be more expensive, but this is not a strong association.
- PROP_TYPE (Property type):
    - Multi-Family: Has a significant positive coefficient, suggesting that multi-family homes are more expensive than the baseline property type.
    - Single Family Residential: Also has a significant positive coefficient, indicating they are valued higher.
    - Townhouse: Shows a significant positive relationship with price.
    - Vacant Land: Has a positive coefficient, but it is not statistically significant, suggesting that vacant land may not be valued differently from the baseline property type in the model.
- PARK_DIST (Distance from park): The negative coefficient is not significant, suggesting no strong evidence that proximity to parks has a measurable effect on housing prices within this model.

- HOSPITAL_D (Distance from hospital): The negative coefficient is not statistically significant, indicating that the distance from a hospital is not a significant predictor of housing prices in this dataset.
- RAIL_DIST (Distance from train): The positive coefficient is significant (p-value = 0.037), suggesting a slight price increase for properties located further from train stations, which might be counterintuitive and warrants further investigation.
- BUS_COUNTS (Number of bus stops): The positive and significant coefficient suggests that more bus stops nearby may increase property values.
- FnB_COUNTS (Number of food and beverage stores): A significant positive coefficient indicates that a higher number of food and beverage outlets in the area is associated with higher housing prices.
- population.density: The significant negative coefficient suggests that higher population density is associated with lower housing prices, possibly due to factors like noise, congestion, or smaller living spaces.
- median.age: The positive and significant coefficient suggests that areas with an older median age of residents have higher housing prices, potentially indicating more established neighborhoods.
- dominant.raceHispanic or Latino: Not a significant predictor of housing prices.
- dominant.raceWhite: A significant positive coefficient suggests that areas with a majority white population are associated with higher housing prices.

P-values:
Variables with p-values less than 0.05 are considered to have a statistically significant relationship with the housing prices. In this model, SQFT, PROP_TYPE (for Multi-Family, Single Family Residential, and Townhouse), RAIL_DIST, BUS_COUNTS, FnB_COUNTS, population density, median age, and dominant.raceWhite are statistically significant.

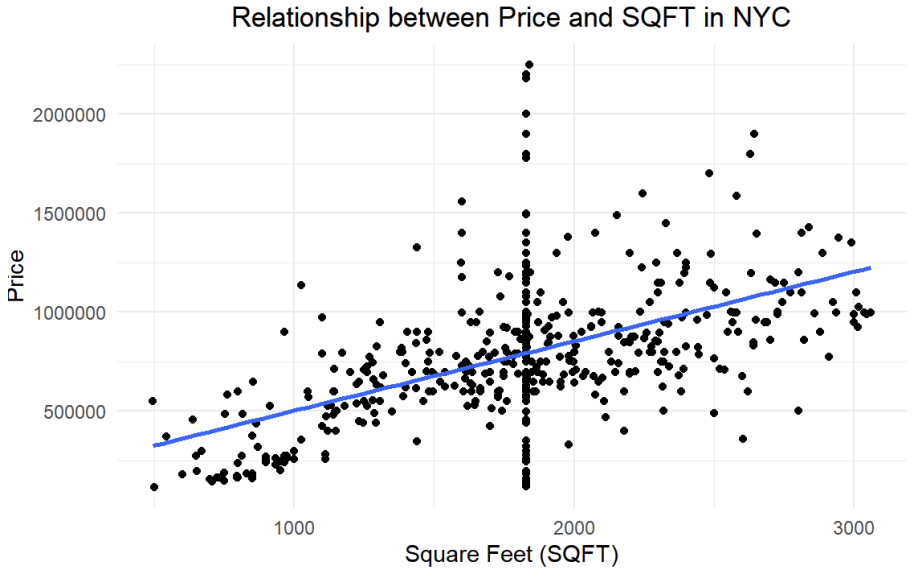
R-squared:
The Multiple R-squared value of 0.5259 suggests that approximately 52.59% of the variability in housing prices can be explained by the model. The Adjusted R-squared of 0.5117 accounts for the number of predictors, suggesting a good fit while adjusting for the number of variables.

F-test:
The F-statistic of 36.98 on 15 and 500 degrees of freedom, with a p-value of less than 2.2e-16, indicates that the model is statistically significant. This means that the independent variables, collectively, are likely predictors of the housing price.

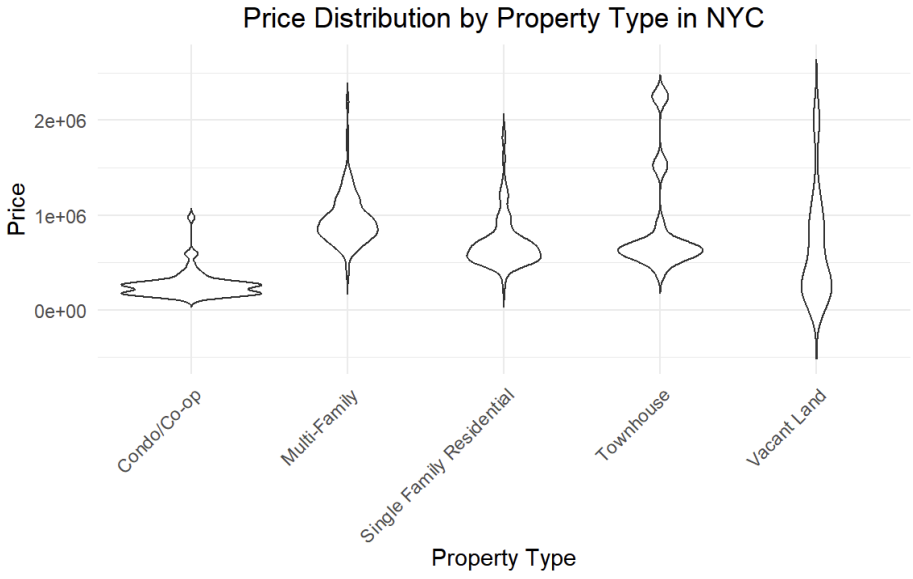
Summary:
The model indicates that property size, certain property types, distance from train stations, the number of bus stops, and the presence of food and beverage outlets are positively associated with housing prices in New York City. Interestingly, the model suggests a negative association with population density and a positive association with the median age of residents. Additionally, the model implies that areas predominantly populated by white residents are associated with higher housing prices. While the model does explain over half of the variability in housing prices,

there is still a significant portion that is unexplained, which could be due to other factors not included in the model or to the intrinsic variability of the real estate market.

Inferential Visualization



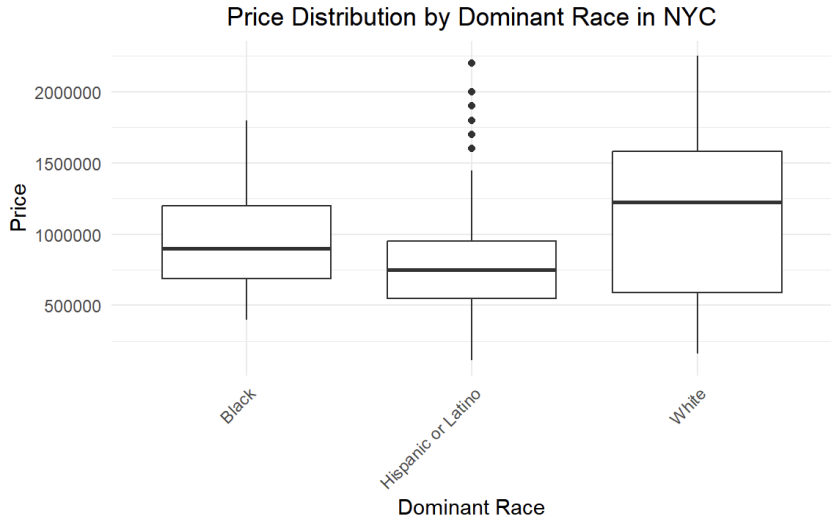
The scatter plot visualizes the relationship between property prices and their size in square feet in New York City. The data points, representing individual property sales, are dispersed across the graph, with a noticeable upward trend signifying that larger properties tend to have higher prices. The blue line, indicating the linear regression fit, slopes upwards, reinforcing the positive correlation between size and price. Despite the clear trend, there is significant variability, with some properties showing much higher prices than others of similar size, suggesting that other factors besides size are influencing price. The concentration of data around the lower square footage range reflects the high density of smaller properties within the city's real estate market. The graph captures the general pattern that more spacious properties in NYC command higher prices but also underscores the complexity of pricing in such a dynamic market.



The violin plot presented offers an insightful comparison of property prices across various property types in New York City. It visually encapsulates the distribution of prices for Condo/Co-op, Multi-Family, Single Family Residential, Townhouse, and Vacant Land. Each violin shape depicts both the density and the range of prices within each property type, with the width corresponding to the frequency of price occurrences.

Condo/Co-op and Townhouse categories show a relatively concentrated distribution of prices, indicating a cluster of properties around a common price range. Multi-Family and Single Family Residential types have a broader range of prices, as indicated by the extended tails of the violins, with Single Family Residential showcasing the highest price points among the categories. Vacant Land has a narrower distribution, suggesting a more uniform pricing structure.

The plot reveals the diversity and complexity of the NYC real estate market, with substantial variability within each property type. The most common prices can be inferred from the widest parts of the violins, while the extended tails to the higher price ranges, particularly in the Single Family Residential category, suggest the presence of premium-priced properties. This visualization is instrumental in understanding the varied landscape of property prices in an urban setting like NYC, highlighting that while there are typical price ranges for each property type, there can also be significant outliers or exceptions within each category.

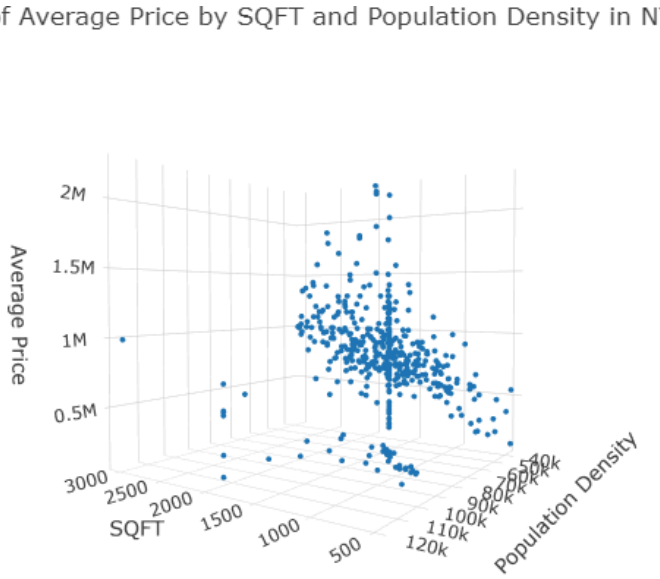


The box plot illustrates the distribution of property prices in New York City across neighborhoods characterized by a dominant racial demographic: Black, Hispanic or Latino, and White. The median price for each group is indicated by the horizontal line within each box, showing that neighborhoods with a majority White population have a higher median price compared to those predominantly Black or Hispanic or Latino.

The range of prices is represented by the span of the boxes, which indicates a considerable spread across all racial demographics, with the White category having a notably wide range. This suggests a diversity of property values within these areas. The presence of outliers, especially in the Black and White categories, points to properties that are priced significantly

higher than the majority, indicating the existence of high-value real estate in these neighborhoods.

The plot suggests that while median property values differ by dominant race, there is a substantial overlap in the price ranges between the groups. This could indicate that factors beyond racial demographics are influencing property prices in New York City, such as location, property size, and amenities. The visualization highlights the complex interplay of race, geography, and economics in the urban housing market.



In the plot, the SQFT axis likely shows that larger properties command higher prices, a trend that reflects the premium on space within the city. The population density axis might reveal that properties in more densely populated areas of New York tend to have higher prices, possibly due to the higher demand for housing in those areas and the amenities typically found in denser neighborhoods.

The distribution of data points could show a clustering in certain areas, suggesting that there are sweet spots in the market where the interplay of size, density, and price are most favorable. It's also common to see outliers in such a plot, representing properties that defy the general trends due to unique features or locations.

Overall, such a visualization would underscore New York City's dynamic real estate market, highlighting the fact that while larger properties and those in denser areas tend to be more expensive, the market is diverse and there are exceptions to every trend. This complexity is a hallmark of a city known for its varied neighborhoods, each with its unique character and market forces at play.

2. Neighborhood level

```
> summary(neighbor_reg)

Call:
lm(formula = average_PRICE ~ SQFT + house.age + PROP_TYPE + PARK_DIST +
    HOSPITAL_D + RAIL_DIST + BUS_COUNTS + FnB_COUNTS + population.density +
    median.age + dominant.race, data = neighbor_level)

Residuals:
    Min      1Q  Median      3Q     Max
-300054 -114405   28353   77053  416924

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                      505361.902 639889.158   0.790  0.43740
SQFT                                518.197     99.120   5.228 2.34e-05 ***
house.age                         -4785.974   2127.722  -2.249  0.03394 *
PROP_TYPEMulti-Family             17170.628 116072.634   0.148  0.88363
PROP_TYPESingle Family Residential -19605.462 136011.008  -0.144  0.88659
PROP_TYPETownhouse                -35249.853 280118.780  -0.126  0.90091
PARK_DIST                            189.615     68.851   2.754  0.01105 *
HOSPITAL_D                           -46.322     34.586  -1.339  0.19301
RAIL_DIST                             -9.169     36.867  -0.249  0.80570
BUS_COUNTS                       -13129.658   4104.045  -3.199  0.00385 **
FnB_COUNTS                         6502.295   5618.921   1.157  0.25857
population.density                   -2.418      1.786  -1.354  0.18830
median.age                       -10619.669  12989.210  -0.818  0.42164
dominant.raceBlack               306867.533 214118.548   1.433  0.16471
dominant.raceHispanic or Latino  274201.722 201088.768   1.364  0.18535
dominant.raceWhite               400374.017 172706.217   2.318  0.02928 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 200900 on 24 degrees of freedom
  (因为不存在，1个观察量被删除了)
Multiple R-squared:  0.8788,    Adjusted R-squared:  0.8031
F-statistic:  11.6 on 15 and 24 DF,  p-value: 1.435e-07
```

Regression Analysis

Coefficients and their Signs:

- SQFT: Positive coefficient (518.197) indicates that an increase in square footage is associated with an increase in average property price, which is statistically significant ($p < 0.001$).
- house.age: Negative coefficient (-4785.974) implies that older houses are associated with lower average prices, which is statistically significant ($p = 0.03394$).
- PROP_TYPE: The coefficients for property types (Multi-Family, Single Family Residential, Townhouse) are not statistically significant, indicating they do not have a distinct impact on the average price different from the baseline property type in this model.
- PARK_DIST: Positive but not significant, suggesting that in this model, the distance to the nearest park does not have a clear effect on average housing prices.
- HOSPITAL_D: Negative coefficient that is not statistically significant.
- RAIL_DIST: Negative coefficient that is not statistically significant.
- BUS_COUNTS: Negative coefficient (-13129.669) with statistical significance ($p = 0.00385$), indicating that a higher number of bus stops is associated with a decrease in the average property price within a neighborhood.
- FnB_COUNTS: Positive coefficient but not statistically significant.
- population.density: Negative coefficient that is not statistically significant.
- median.age: Negative coefficient that is not statistically significant.

- dominant.race: None of the coefficients for dominant race categories (Black, Hispanic or Latino, White) are statistically significant, although the coefficient for White is close to the conventional threshold for significance (p = 0.02928).

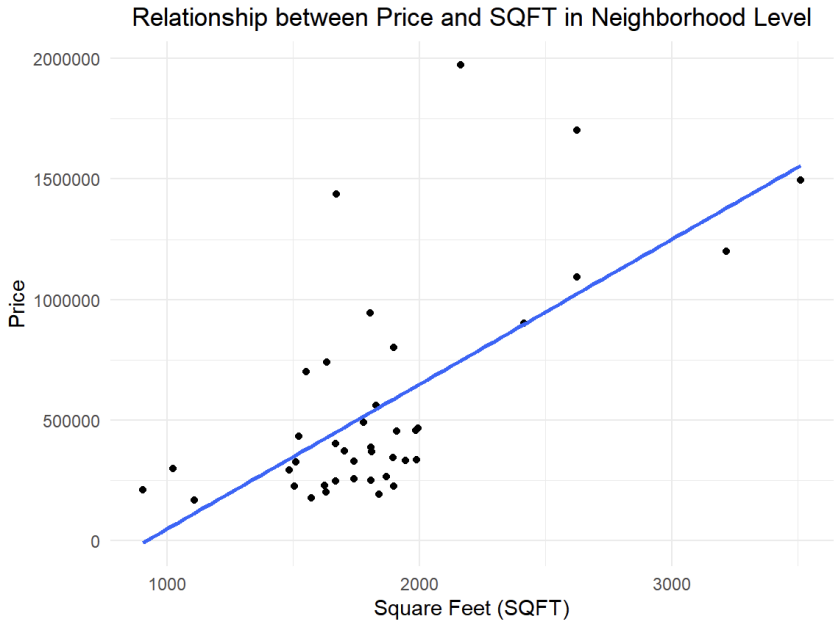
P-values:
A p-value less than 0.05 is typically considered evidence that the coefficient is significantly different from zero in the population. In this model, significant predictors include SQFT and BUS_COUNTS.

R-squared:
The Multiple R-squared value of 0.8788 indicates that the model explains 87.88% of the variance in average neighborhood property prices, which is quite high, suggesting a good fit.
- The Adjusted R-squared value of 0.8031 adjusts for the number of predictors in the model and still suggests a strong explanatory power.

F-test:
The F-statistic is significant (p = 1.435e-07), indicating that the model as a whole is statistically significant and that there is a relationship between the predictors and the average property price.

Summary:
The regression model suggests that property size (SQFT) and the number of bus stops (BUS_COUNTS) are significant predictors of the average price in a neighborhood, with larger properties associated with higher prices, and more bus stops associated with lower prices. The negative association with bus stops might reflect a preference for less dense or less busy areas within the neighborhood context. The model has a high R-squared value, which indicates it fits the data well, but as always, it's important to consider other potential variables not included in the model that could affect average housing prices. The insignificant coefficients for property type and dominant race suggest that within neighborhoods, these factors do not vary enough to significantly impact the average property prices, or other variables in the model are capturing their effects. The residual standard error is quite large, indicating substantial unexplained variability in average prices despite the high R-squared value.

Inferential Visualization



The scatter plot illustrates the relationship between the size of properties, measured in square feet (SQFT), and their prices at the neighborhood level. The upward sloping blue line indicates a positive correlation between property size and price, suggesting that, generally, larger properties are sold at higher prices. This trend is a common observation in real estate markets, where size is a significant determinant of value.

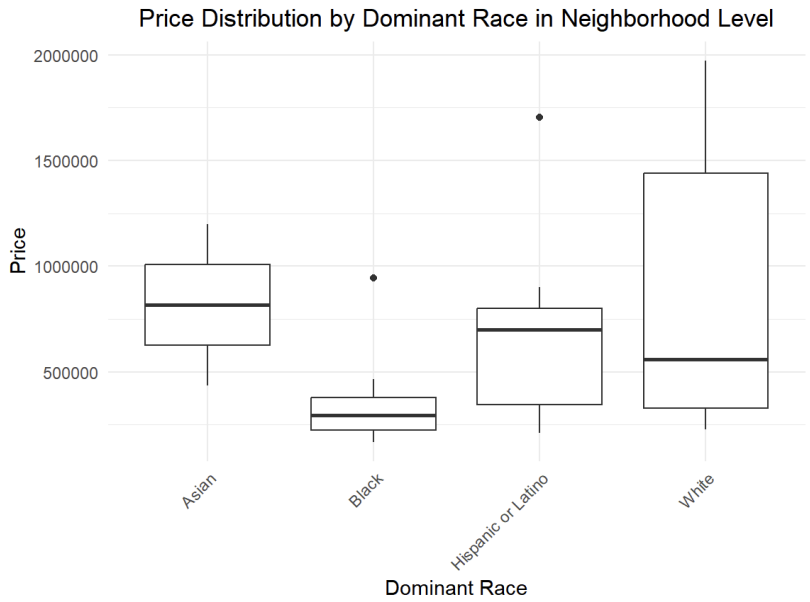
The distribution of the data points shows a concentration of properties in the lower size and price range, with fewer properties at higher sizes and prices, as indicated by the more sparse data points towards the right of the plot. Some properties, particularly in the higher SQFT range, deviate from the trend line, indicating that factors other than size are affecting their prices.

This pattern suggests that while there is a general increase in price with size, the relationship is not uniform across all properties. The variability implies a diverse real estate market where property valuations are influenced by a mix of factors, including but not limited to square footage.

The violin plot provides a detailed visual representation of property prices by type at the neighborhood level. Displaying a range of property types, including Condo/Co-op, Multi-Family, Single Family Residential, and Townhouse, the plot reveals the spread and density of property values within each category.

The shapes of the violins for Condo/Co-op and Townhouses suggest a more uniform pricing structure with fewer extreme values, as indicated by their less pronounced tails. On the other hand, the Multi-Family and Single Family Residential categories show wider distributions, implying a greater range of property prices. This could reflect a diverse market where various factors such as location, property size, and features significantly impact price.

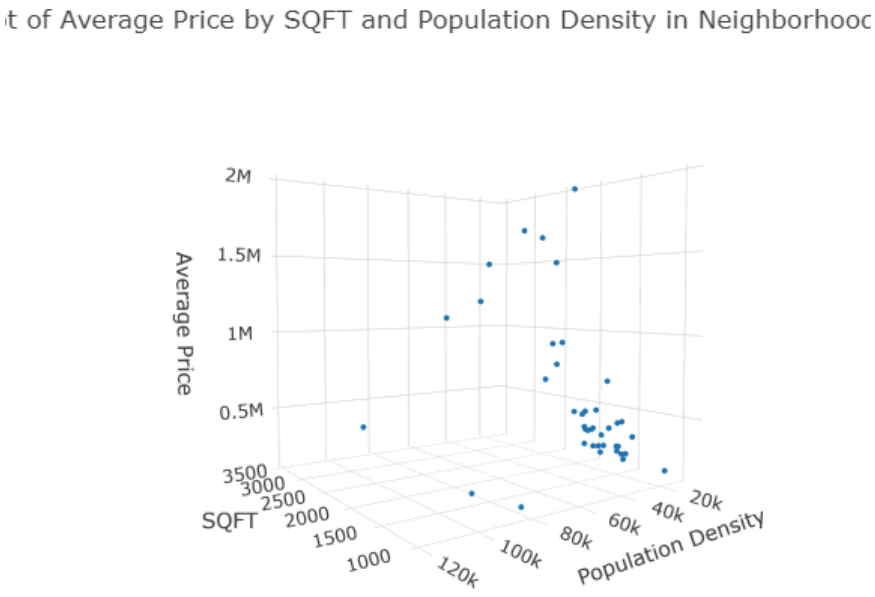
From the broadest points of the violins, one can deduce the most common pricing segments, while the outliers stretching away from the main body of the violins highlight exceptional cases of high-priced properties. This plot serves as a critical tool for understanding the nuances of property pricing at the neighborhood level, showing not only the average trends but also the price extremes that characterize the property types.



The box plot illustrates the distribution of property prices across neighborhoods at different dominant race demographics: Asian, Black, Hispanic or Latino, and White. Each box captures the interquartile range (IQR) of property prices within neighborhoods of the respective racial majority, with the median price indicated by the horizontal line within the box.

The boxes show that the median property prices and range of prices vary across neighborhoods with different racial majorities. Neighborhoods with a majority of White residents exhibit the highest median prices and the greatest range in property values, as indicated by the size of the box and the length of the whiskers. This suggests a higher variation in property prices within these areas. The neighborhood with a majority of Asian residents shows a lower median price and a more compact IQR, indicating less variability in property prices.

Notable are the outliers in the Black and White categories, represented by individual points beyond the whiskers, which signify properties that are priced much higher than the typical range for their respective neighborhood types. This plot underscores the disparities in housing prices at the neighborhood level, possibly reflecting underlying economic, social, or geographic factors influencing the real estate market.



As I interpret the description of the 3D scatter plot, it portrays a relationship between the average price of properties, their square footage (SQFT), and population density in neighborhood level. The plot likely presents SQFT on one axis, showing the size of the properties, population density on another axis, representing the number of people living per unit area, and average property price on the third axis.

This type of visualization would typically show that as SQFT increases, the average price also tends to rise, illustrating the premium on space in a densely populated city like New York. The population density axis could provide insights into how the concentration of people in an area affects property prices, with potentially higher prices in more densely populated neighborhoods.

The spread of data points in the 3D space might reveal clusters where certain combinations of SQFT and population density correspond to specific price ranges. High-density areas with large properties might command the highest prices. Conversely, areas with lower population density and smaller properties might show a lower average price, but there could also be exceptions to these trends, indicated by data points that diverge from the general pattern.

This 3D plot would be particularly useful for identifying how different factors simultaneously influence property prices, providing a holistic view of the real estate market dynamics at the neighborhood level.