Exploratory data analysis - DC Residential properties

For this data analysis, we are examining variables used predict real estate prices. The price data is not normally distributed, so I used graphs and applied nonparametric significance tests to find whether or not other variables could be beneficial in predicting price. Many variables showed an impact on price, and many variables were related to geographical location.

Variables that were challenging to work with included census block and square. These are nominal categorical variables with thousands of categories. While the values are numbered, the numbers don’t have a consistent correlation to geographical space they represent or or have a relational value. These geographical variables do have an impact on prices similar to quadrant, ward, neighborhood, and sub-neighborhood. I graphed medians per census block and square to get an idea of the variety in median prices per area.

The largest categorical geographical variable is Quadrant. After viewing violin and box plots for Quadrant, I applied the Kruskal Wallis test and found that the prices in Quadrant were from different theoretical distributions. Ward, Neighborhood, Subneighborhood, Census Tract and Zipcode, which are approximate geographical subdivisions of Quadrant, graphically appear to be significant in predicting price. The values are more extreme in the smaller categories, so the differences may be more significant in predicting price.

Sale year also had a strong impact on sale price. When graphing the median sale price per year, the positive correlation between sale date and price was clearest. When I applied spearman’s correlation coefficient for a monotonic relationship across all observations, there was a moderate positive monotonic relationship.

Other significant variables with multiple categories included Rooms, Stories, and Grade, all of which had stronger graphical evidence for a difference in distributions than Quadrant, which we proved to be from different distributions. For binary categorical variables, I applied the Mann Whitney Wilcoxon test. I found that the prices for Qualified and Unqualified buyers were from different theoretical distributions.