MILESTONE REPORT

**INTRODUCTION:**

The purpose of this report on D. C. Residential Prices is do predictive modeling on housing costs in DC. This knowledge can be used by government or NGOs in studying sustainable housing and segregation by concatenating this dataset with census data. Price predictions can also be used by real estate developers and investors. The data is from a Kaggle competition, “D. C. Residential Properties”, provided by Open Data DC at <https://www.kaggle.com/christophercorrea/dc-residential-properties>. So far, I have cleaned the data and completed exploratory statistical analysis to prepare for machine learning applications.

#### **DATA CLEANING:**

#### STEPS:

-I combined the bathroom and half-bathroom columns into one column.

-I removed a column “Unnamed” which was equivalent to the index.

-I removed “State” and “City” because values were identical for each sale.

-I removed “Fulladdress” because we had “latitutde” and “longitude” columns, and missing data would be difficult to fill accurately.

-I removed “Nationalgrid” because we had “latitutde” and “longitutde” columns.

-I removed ‘X’ and ‘Y’ because they were synonymous with ‘LATITUDE and ‘LONGITUDE’

-I corrected some rounding errors in various columns.

#### OUTLIERS:

-Specific outliers:

-Stories: 250, 275, 826

-Year remodeled: 20

-Structure: ‘default’

-Numerical data:

-Didn’t remove values outside of Q1-1.5IQR or Q3+1.5IQR for most columns because the data was normal.

-For GBA and LIVING\_GBA, I removed outliers according to technique above.

-For LANDAREA and PRICE, the data was nonparametric and skewed right. I used fences at Q1-1.5IQR or Q3+4.5IQR to keep some of the higher values, keeping with the nature of the data.

-Categorical data:

-All categories and distributions seemed reasonable.

#### MISSING VALUES:

-Numerical data:

-Sorted data by location (x/y coordinates).

-Rolling mean with window of 500 so that no column had more than 1% missing data.

-Dropped remaining rows.

-Categorical data:

-Grouped variables by neighborhood.

-Replaced NaN values with the mode of that column in that neighborhood.

-Price:

-removed all observation with missing price, since the goal of the project is to predict price.

Exploratory data analysis - DC Residential properties

**STATISTICAL ANALYSIS:**

INTRODUCTION:

This analysis focuses on finding variables that have a significant impact on 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.

GEOGRAPHICAL VARIABLES:

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:

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 linear relationship.

OTHER VARIABLES:

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.

**MOVING FORWARD:**

I will apply machine learning applications to explore which models best predict outcomes using linear regression, logistic regression, clustering, and random forest applications. I will continue to re-evaluate my approach and reflect upon how the machine learning outcomes match or challenge my exploratory data analysis.