UCLA Ext -EDA Final

**Data Summary, Oddities and Outliers**

The data I received consisted of two csv files. One with housing data and one with related school data. The housing data contained 683 rows of data and15 columns with information on neighborhoods, property information (sqft, beds, lotsize etc.). Some rows contained missing data in select columns (sqft:2, lotsize:20). The school data contained 52 rows of data and 3 columns with information on the school’s name, size and rating out of 10.

This data was initially examined for any possible errors or outliers. This is done to avoid errant data affecting the analysis. There were a few oddities found in the housing data. For the “baths” column, there were two outliers. One row had a value of 999, and another and a value of 25. The value of 999 seems to be errant and is in a row with other missing data. This row will be kept out of the analysis. The value of 25 seems to be a typo, and the most likely true value is 2.5 baths. However, for the initial analysis, all errant data will have their rows removed from the analysis. A sensitivity analysis will be conducted where these values are imputed to compare. For the “year” column, there are two issues. There is one row with a value of 1495 and a row with a value of 2111. These values are typos, with their intended values likely to be 1945, and 2011 respectively. These rows were removed from the initial analysis. In the “soldprice” column there is one outlier of 664. The most likely true value of this row is 664000, but it too will be kept out of the initial analysis.

**Data Cleaning**

The data was initially cleaned and combined using Python. Python was chosen because it is what I am most familiar with, and the notebook format allows for easy communication of the function of code. The script used to do the cleaning is available and named “*initial\_eda.ipynb*”. First the two datasets were combined into one for ease of analysis. This was done by joining the school names in ‘schools.csv’ to the school names in the ‘housing.csv’. This resulted in a file with all the housing data and the respective elementary, middle and high school’s sizes and ratings. Then the data was inspected for missing values and outliers.

Missing values were found for the “lotsize” and “sqft” columns, 20 and 2 missing values respectively. For the initial eda, these missing values are removed from the analysis. For the sensitivity analysis, these may be imputed.

Outliers and oddities, as described in the section above, were also removed in the data cleaning process. A secondary data set was created with imputed values for the missing columns to be used in the sensitivity analysis.

**One Variable Visuals**

The focus of this analysis will be to look at the data by neighborhood. This is to hopefully give an idea of housing value and related factors in different neighborhoods to hopefully provide insight into those neighborhoods that seem the most valuable to the company.

Initially, I would like to just get an idea of how big each neighborhood is. This may provide information into which neighborhoods may provide the most opportunity for investment. The plot below is a bar plot showing the number of rows (properties) for each neighborhood.

Chart, bar chart

Description automatically generated

As the plot above shows, the neighborhoods with the most properties represented in the data are Orange, Blue, Red, and Green with between 102 and 141 properties. The neighborhoods with the least number of properties are Yellow, Silver, Gold, and Purple with values between 3 and 68, with Purple having 3 and the next highest being Gold with 51. The plot above shows that the Purple neighborhood is a special case, having very few properties. This may indicate that the value of these properties might be affected by differently than similar factors in other neighborhoods.

Now I would Like to get an understanding of the cost of properties in each neighborhood. I think this will be a good insight for the company so they can understand the types of investment required for properties in the respective neighborhoods.

Chart, box and whisker chart

Description automatically generated

Above we begin to get an idea of the range of property values that can be expected for each neighborhood. We can see that most neighborhoods have a decent range of property values, however we can also see that some neighborhoods tend to have a higher range than others. Above we see the top 4 neighborhoods, in terms of average property value, are Gold, Green, Yellow, and Silver with a mean range of $1,640,896 to $1,272,016. The bottom 4 neighborhoods are Blue, Orange, Purple, and Red, with a mean range of $1,229,997 to $960,694.

Next, I wanted to get an idea of the years properties in the dataset were built, and if they differ depending on the neighborhood. I decided to look at year built as that can be an indicator of many factors related to property value, including types of materials used, and overall wear of things like plumbing, heating, and other household amenities.

The plot below is a histogram of year properties were built; color coded by neighborhood. The histogram is stacked so we can better see how neighborhoods may differ.

Chart, histogram

Description automatically generated

Above we can see that most properties in the data set, regardless of neighborhood, were built between 1970 and 1990. We also see that the Red neighborhood tends to skew older, with the oldest property build in 1908, the newest built in 1985, and a mean construction data of 1954. The newest neighborhood is Gold, with the oldest property built in 1974, the newest in 2018, and a mean construction date of 1996.

This graph hopefully provides some insight into the age of the properties that can be found in each neighborhood.

**Two Variable Visuals**

The next step in my analysis is to look at some of the relationships between the different variables in the data. Initial, I am interested in looking at the relationship between school ratings and property values. If there is a relationship that proves strong and positive, I think this could be an important metric in finding good property investments. Namely, by looking for somewhat undervalued properties near high rated schools.

The next visual is a scatter plot of the property values and school ratings, again color coded by the neighborhood to see if there are any notable differences.

Chart, scatter chart

Description automatically generated

Above we can see a clear positive correlation between the average school rating and property value. We can also see separation between the different neighborhoods. At the high end, we see that the Gold neighborhood has some of the highest average school ratings and property values. We also see that the Red and Purple neighborhoods tend to have some of the lowest average school ratings and property values.

We can also see that regardless of average school rating, there is still a large range of property values, and that it is not the determining factor. However, I do believe that there is reason consider average school rating in the investment value of a property.

The next thought I had was to look at the relationship between the year of construction and property value. Since we have looked at the average age of each neighborhood, understanding this relationship might give a greater understanding of the valuation of each neighborhood.

Below is a 2d density plot, showing the relationship between the construction year and the property value, as well as the density of data in this relationship.

Chart

Description automatically generated

Above we can see a slight positive between the year a property was constructed and its value. We also see a high density of properties with a value between $1.25 million and $1.75 million constructed between 1970 and 1990. This reinforces the insight given by the histogram in the previous section.

**Analysis**

For my analysis, I wanted to showcase the relationship between the property value and other factors. First, I wanted to look more closely at the relationship between the average school rating and the property value. To do this, I created a linear regression model and created a graph of the residuals. Below is a table of the model output, which I will describe the significance of the different values shown.



The main values above that we are interested in are the coefficients, Adjusted R-squared, and p-value. For the coefficients, we see we have an intercept of 687556, this means that for any estimation, the initial starting price point is $464,920, the error is a plus or minus of $42,794. Below we see the school rating has a coefficient of $93,804 and an error of $5,699. This means that as the school rating increases by 1, the value of a property increases by approximately $93,804, plus or minus $5,699. The sqft coefficient increases property value by $105.54 per square foot plus or minus $11.08.

The adjust R-squared value essentially represents the percent of which our model can explain the increase or decrease of the property value by only looking at the average school rating. Here we see that our model can account for approximately 36% of the variation in property value.

The next value of interest is the p-value. This essentially tells us the strength of our model. The number represents how close our model predictions are to random chance, the smaller the number the more likely the model’s predictions are better than random chance. Here, our p-value is 2.2e-16, this is a very small number indicating we have a strong model. Normally anything below 0.05 would be considered strong.

Below is a plot of the residuals of the model. The residuals basically show the difference between a predicted value and observed value. The closer to zero the better the model is a predicting.

Chart, scatter chart

Description automatically generated

Above we see that the residuals are linear to the 0 line, but they are distributed above and below it. This means that the relationship is linear and that the model is on average either overpredicting or underprediction property values. We can see some residuals near zero, meaning that occasionally our predictions can be very accurate.

While this model is helpful in better showing the relationship between property value school rating and square footage, for better predictions, more relationships need to be added to the model.

For a final step, I wanted to look at the sold price column as a categorical. That is, is the property value low, medium, or high. This can help better show some of the relationships between property value and other factors.

Below is a scatter plot matrix of “soldprice”, “year”, “sqft” and, “school\_rating” color coded by the KMeans cluster it is in, black, pink and green, representing high, medium, and low property values respectively.

A picture containing calendar

Description automatically generated

Above we can see clear separation in the scatter plots that include the “soldprice”. This helps further illustrate the range of years built, square footage and average school rating found in the different price ranges.

**Sensitivity Analysis**

Initially, all rows with missing data or outliers were removed from the dataset to avoid errant data affecting the analysis. However, now I want to check to make sure that removing those data points isn’t going to throw off the analysis I have done. To do this missing values must be imputed.

This was also done in Python, as mentioned in the first section of this report. One major missing value was “lotsize”, with up to 20 rows missing data. To impute this variable, I found the average percent the built portion of the property is to the lot size. I found that on average, the bulilt portion of a property is about 19%. For each row with missing lot size data, I would look at the square footage value, assume it is 19% of the lot size, and use math to convert from square footage to acres.

For the outliers, most of them seemed to be typos, as outlined in the first section of this report. These outliers had their values replaced by the most likely true value, as specified in the first section.

After imputation was completed, I reran the R analysis I had done previously to see if anything had changed. I compared the visuals and linear model between the two analyses and observed no noticeable change. To further drive this point home, the statistical values of the linear models are nearly identical. To me this means that the removal of those values did not have an affect on the analysis and the information above is valid.