# Springboard--DSC Program

# Capstone Project 2: Milestone Report 1

# An Exploration of Housing Sales in Washington D.C. and King County

By: Garrett Yamane

March 21st, 2020

## 1. Introduction

Washington D.C., the capital of the United States of America, is home to many historical landmarks and our beloved government. King County, in Washington state, encompasses a variety of geographically diverse cities, from rural farms to the tech hub of Seattle and Bellevue. Both locations offer so much opportunity, and are places that are highly sought-after to move to. I am from the Bay Area where the cost of living has been increasing more and more throughout Silicon Valley's rapid progression.

When people move, do different housing features impact the price of a house sale? Are prices of houses being sold affected differently by these features based on location? These questions create the foundation of what I would like to explore in Washington D.C. and King County's housing sales.

## 2. Problem Statement

Moving can be a major hurdle for one's life. Whether it is the first time moving out on your own after graduating from college or if you are beginning to start a family and looking to settle down, buying one's own place can be quite expensive. Knowing the housing market and what to expect if you are moving to an unfamiliar place can be a challenging task. But what if there was a way to model housing trends and predict what you may have to pay? One model cannot accurately summarize every housing market, but being able to take two geographically separate locations and compare how house features are weighted differently when building housing price models can prove beneficial to both real estate agents and their clients when deciding where to move and what to buy. The goal of this project is to build individual accurate predictive machine learning model for both Washington D.C. and King County in order to both closely examine what features heavily impact the price of a house sale and to help prospective buyers understand the market they are buying into.

# 3. Target Client

Real estate agents need to be very educated and aware of the housing market when helping clients buy a new place. Not only is it important to know how much a house is going to cost, but it is also helpful to know what factors will increase the price and what features carry the highest value for different locations. People moving into the suburbs may value a place with more bedrooms and living space for raising a family, whereas those moving into a larger city may want a view with newer renovations. Real estate agents for major cities and their clients can benefit from a study that targets two different geographical locations through predictive modeling: King County in Washington State and D.C.

# 4. Data Wrangling

## a) Gathering the Data

The data will come from two different datasets from Kaggle:

- King County, Washington Dataset [1]
- Washington D.C. Dataset [2]
-

There are 21.6K rows in the King County dataset, and 159K rows in the Washington D.C. dataset. However, the King Country dataset only contains house sales between May 2014 and May 2015, whereas the Washington D.C. dataset contains sales dating back to 1995. For

[1] https://www.kaggle.com/harlfoxem/housesalesprediction
[2] https://www.kaggle.com/christophercorrea/dc-residential-properties#DC_Properties.csv

consistency, I only sampled sale dates that both datasets have in common which decreased the number of rows of the D.C. dataset down to around 14K. I also looked at the following features that each dataset has in common:

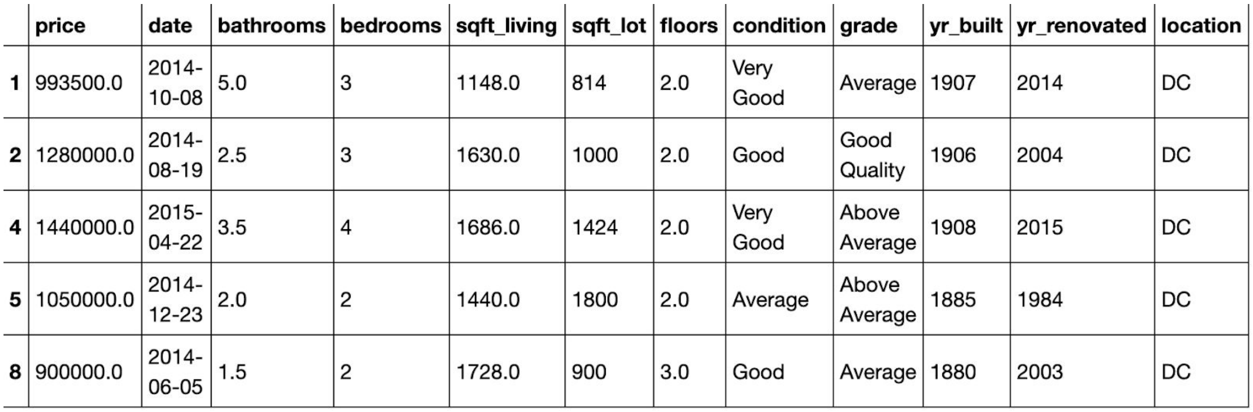
- Price
- Sale Date
- # bathrooms
- # bedrooms
- Living SQFT
- Lot SQFT
- Stories
- Condition
- Grade
- Year Built
- Year Remodeled

## b) Cleaning and Wrangling the data

Ultimately, I wanted to combine the data sets into a single data frame for simplicity. To accomplish this, it was important to not only look at the columns that each data set has in common, but to also make sure that the data in each corresponding column is of the same type prior to merging the two data frames together.

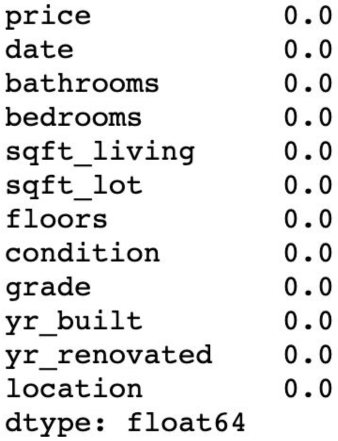
The first step was to modify the Washington D.C. data. Renaming the columns in the Washington D.C. data frame made the merge much easier, so I changed the columns names to match those of the King County data frame. I then filtered down the data set to only contain housing sales between May 2014 and May 2015. I also want to make sure that missing values are accounted for and that I handled these cases prior to the merge. Thus, I removed the rows where the price of the sold housing unit is missing, since this was my dependent variable I built my regression models around. For any rows where the "Year Remodeled", "Year Built", "Living SQFT", or "# Floors" were missing, I imputed the values with the average for the entire column. For rows where the "Condition" or "Grade" was missing, I imputed the values with the label "Missing."

For the next part, I modified the King County data set. The only modification I made was to convert the "Grade" and "Condition" columns from numerical categorical values to the matching String values found in the Washington D.C. data. For example, the values found in the "Condition" column in the King County data set were numbered 1, 2, 3, 4, and 5. However, the grades in Washington D.C. were "Poor", "Average", "Good", "Very Good", and "Excellent." Because of this, I replaced each of the numerical values for the column with the corresponding grades.



| | price | date | bathrooms | bedrooms | sqft_living | sqft_lot | floors | condition | grade | yr_built | yr_renovated | location |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 993500.0 | 2014-10-08 | 5.0 | 3 | 1148.0 | 814 | 2.0 | Very Good | Average | 1907 | 2014 | DC |
| **2** | 1280000.0 | 2014-08-19 | 2.5 | 3 | 1630.0 | 1000 | 2.0 | Good | Good Quality | 1906 | 2004 | DC |
| **4** | 1440000.0 | 2015-04-22 | 3.5 | 4 | 1686.0 | 1424 | 2.0 | Very Good | Above Average | 1908 | 2015 | DC |
| **5** | 1050000.0 | 2014-12-23 | 2.0 | 2 | 1440.0 | 1800 | 2.0 | Average | Above Average | 1885 | 1984 | DC |
| **8** | 900000.0 | 2014-06-05 | 1.5 | 2 | 1728.0 | 900 | 3.0 | Good | Average | 1880 | 2003 | DC |

*Figure 1: A portion of the finalized data frame with labeled columns*



```
price           0.0
date            0.0
bathrooms       0.0
bedrooms        0.0
sqft_living     0.0
sqft_lot        0.0
floors          0.0
condition       0.0
grade           0.0
yr_built        0.0
yr_renovated    0.0
location        0.0
dtype: float64
```

*Figure 2: Missing value percentage for each column in the dataset*

Finally, the Washington D.C. data frame consisted of 7160 entries and the King County data consisted of 21,613 entries. Overall, there were 12 features that described each house sale. Now, the respective columns all are of the same type and have no missing values.

# 5. Exploratory Data Analysis and Initial Findings

Because Washington D.C. and King County, Washington are located on opposite sides of the country, it is very interesting to look at how housing features and prices may differ and what could be in high demand depending on if you want to move to the east or west coast.

## a) Are houses more likely to sell at specific times of the year?

The first thing I wanted to do was to look at when houses were most likely to be sold. Does the time of the year actually impact when houses sell?

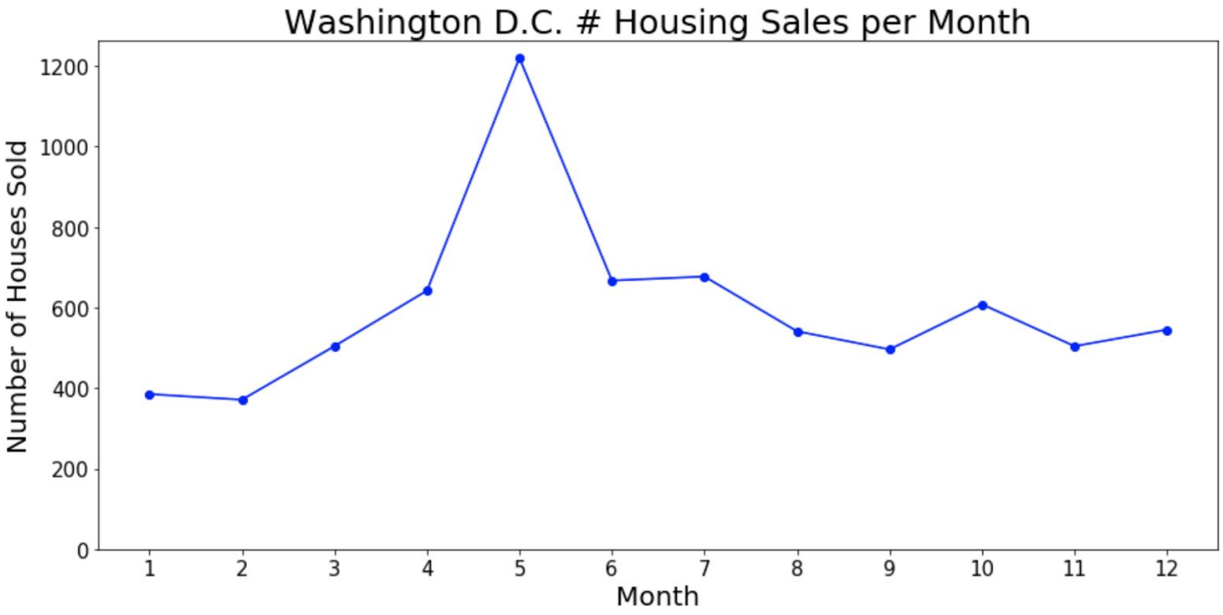


*Figure 3: Washington D.C. House Sales by Month*

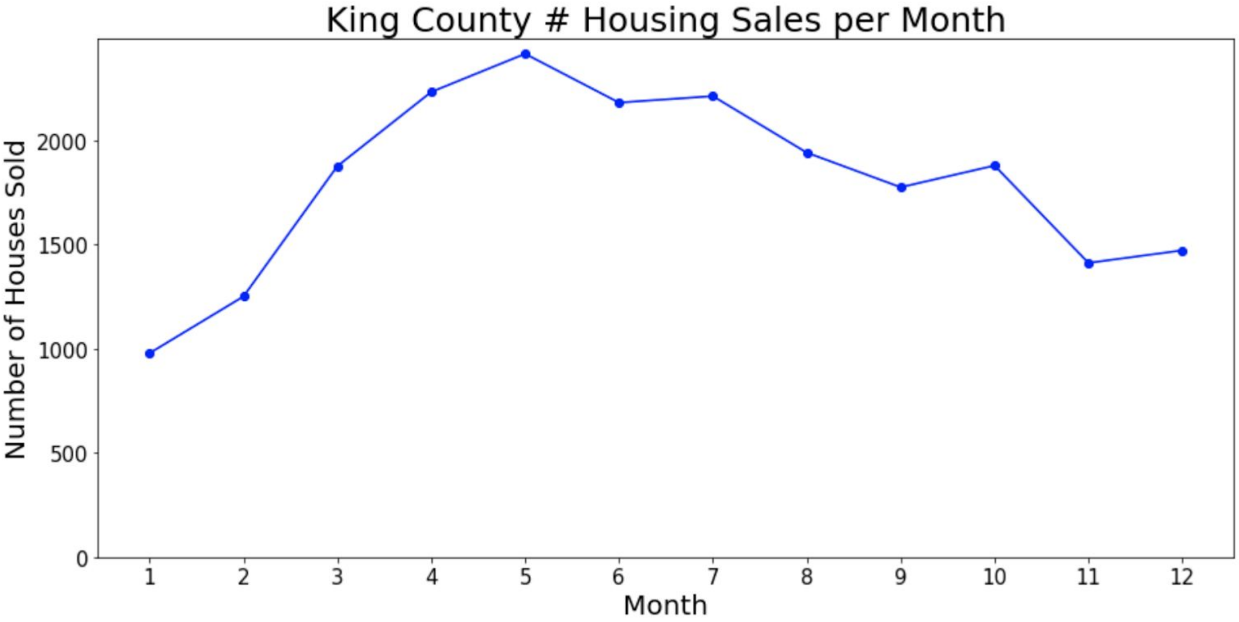
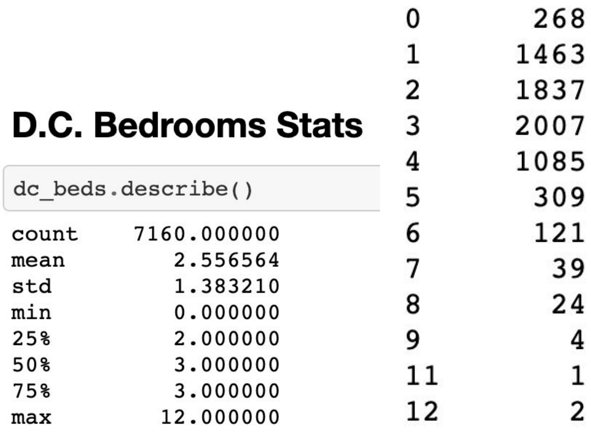


*Figure 4: King County House Sales by Month*

By aggregating the housing sales by month for each data set, figures 4 and 5 highlight that May is the month when most houses are sold for both regions. It is interesting to note that in King County, the immediate months preceding and proceeding May are much closer to the peak number of houses sold compared to those months in Washington D.C. There is an obvious increase in sales in May, with the surrounding months dropping in sales significantly.

## b) Descriptive Statistics for Bedrooms, Bathrooms, Square Feet of Living, and Square Feet of Lot



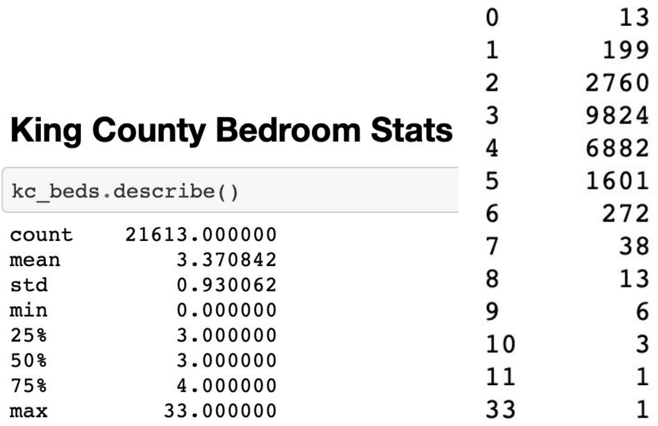
D.C. Bedrooms Stats

```
dc_beds.describe()
```

```
count    7160.000000
mean        2.556564
std         1.383210
min         0.000000
25%         2.000000
50%         3.000000
75%         3.000000
max        12.000000
```

```
0      268
1     1463
2     1837
3     2007
4     1085
5      309
6      121
7       39
8       24
9        4
11       1
12       2
```

*Figure 5: Washington D.C. bedrooms summary statistics and value counts*



King County Bedroom Stats

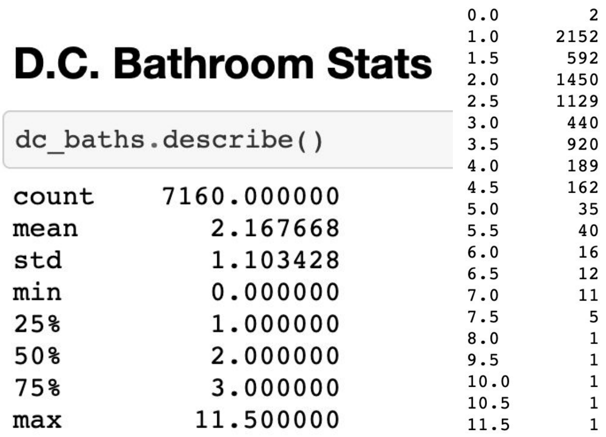
```
kc_beds.describe()
```

```
count    21613.000000
mean         3.370842
std          0.930062
min          0.000000
25%          3.000000
50%          3.000000
75%          4.000000
max         33.000000
```

```
0       13
1      199
2     2760
3     9824
4     6882
5     1601
6      272
7       38
8       13
9        6
10       3
11       1
33       1
```

*Figure 6: King County bedrooms summary statistics and value counts*

Looking at the values shown in figures 5 and 6 helps to distinguish some interesting outliers. It appears that in King County, there is a housing unit that was sold with 33 bedrooms. There are only single cases where certain housing units had a significantly larger number of bedrooms. It is also worth noting that King County has an average of almost 1 more bedroom on average per house sale (which could be due to the outlier of the house with 33 bedrooms).



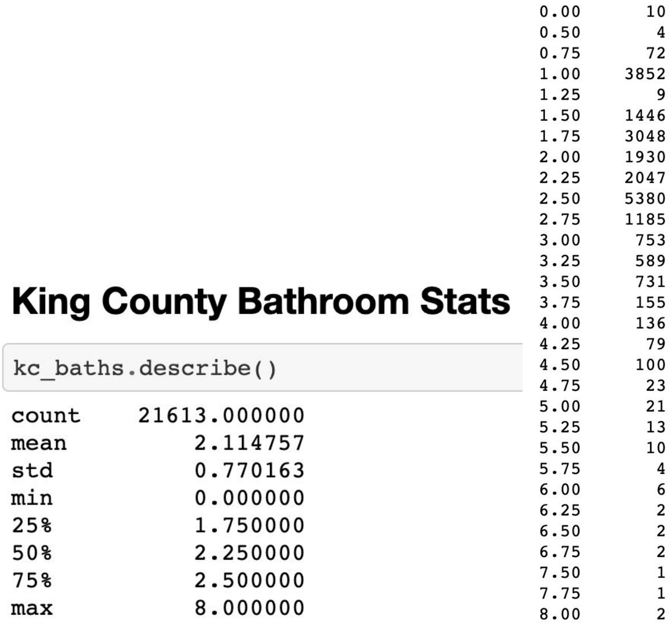
## D.C. Bathroom Stats

```
dc_baths.describe()
```

```
count    7160.000000
mean        2.167668
std         1.103428
min         0.000000
25%         1.000000
50%         2.000000
75%         3.000000
max        11.500000
```

```
0.0        2
1.0     2152
1.5      592
2.0     1450
2.5     1129
3.0      440
3.5      920
4.0      189
4.5      162
5.0       35
5.5       40
6.0       16
6.5       12
7.0       11
7.5        5
8.0        1
9.5        1
10.0       1
10.5       1
11.5       1
```

*Figure 7: Washington D.C. bathroom summary statistics and value counts*



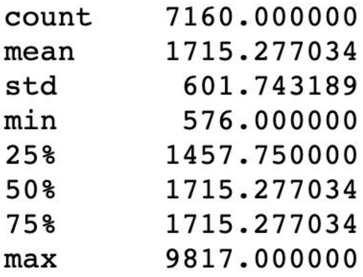
## King County Bathroom Stats

```
kc_baths.describe()
```

```
count    21613.000000
mean         2.114757
std          0.770163
min          0.000000
25%          1.750000
50%          2.250000
75%          2.500000
max          8.000000
```

```
0.00      10
0.50       4
0.75      72
1.00    3852
1.25       9
1.50    1446
1.75    3048
2.00    1930
2.25    2047
2.50    5380
2.75    1185
3.00     753
3.25     589
3.50     731
3.75     155
4.00     136
4.25      79
4.50     100
4.75      23
5.00      21
5.25      13
5.50      10
5.75       4
6.00       6
6.25       2
6.50       2
6.75       2
7.50       1
7.75       1
8.00       2
```

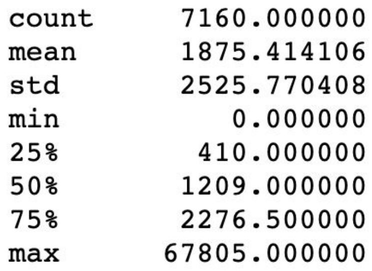
*Figure 6: King County. bedrooms summary statistics and value counts*



```
count    7160.000000
mean     1715.277034
std       601.743189
min       576.000000
25%      1457.750000
50%      1715.277034
75%      1715.277034
max      9817.000000
```

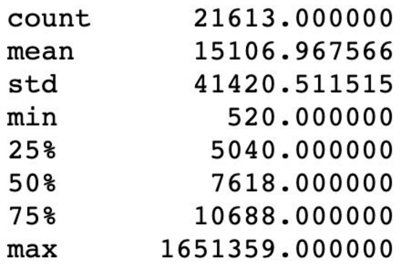
*Figure 7: King County living square feet statistics*

In the D.C. data set, the mean sqft_living is around 1715 per house sold. This number is 300 feet smaller than the average size home sold in King County, with King County's average living square feet having a larger maximum and smaller minimum.



```
count     7160.000000
mean      1875.414106
std       2525.770408
min          0.000000
25%        410.000000
50%       1209.000000
75%       2276.500000
max      67805.000000
```

*Figure 8: Washington D.C. lot square feet summary statistics*



```
count      21613.000000
mean       15106.967566
std        41420.511515
min          520.000000
25%         5040.000000
50%         7618.000000
75%        10688.000000
max      1651359.000000
```

*Figure 9: King County lot square feet summary statistics*

It appears that the average lot size for housing units in King County was much lower than those in Washington D.C. There was a much larger range of property lot sizes found in King Country, ranging from 520 square feet up to 1651359 square feet. These could be large outliers to the rest of the data set, and it is important to look into them when building my model. Washington D.C. is a total of 61.05 square miles, whereas King County is 2,307 square miles. This leaves a lot more room for larger plots of land (such as farmland) to be sold.

### c) What are the grade and condition distributions of houses sold?

For context, condition refers to the overall quality or build of the house. The better the condition, the more well-maintained the housing unit is. Low condition scores means that the housing unit is approaching a condition that might require reconstruction. Grade, on the other hand, refers to the evaluation of the construction materials and level of craftsmanship used to build the house. The higher the score, the higher the quality. For example, a score of 13 (the highest score) is mansion-level.

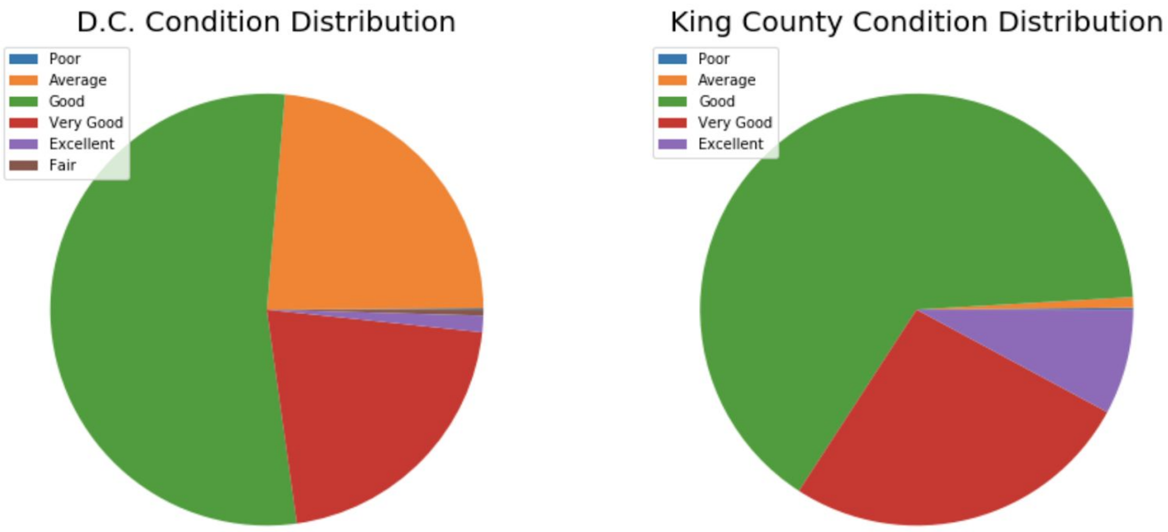


*Figure 10: Condition categorical distribution for Washington D.C. and King County*

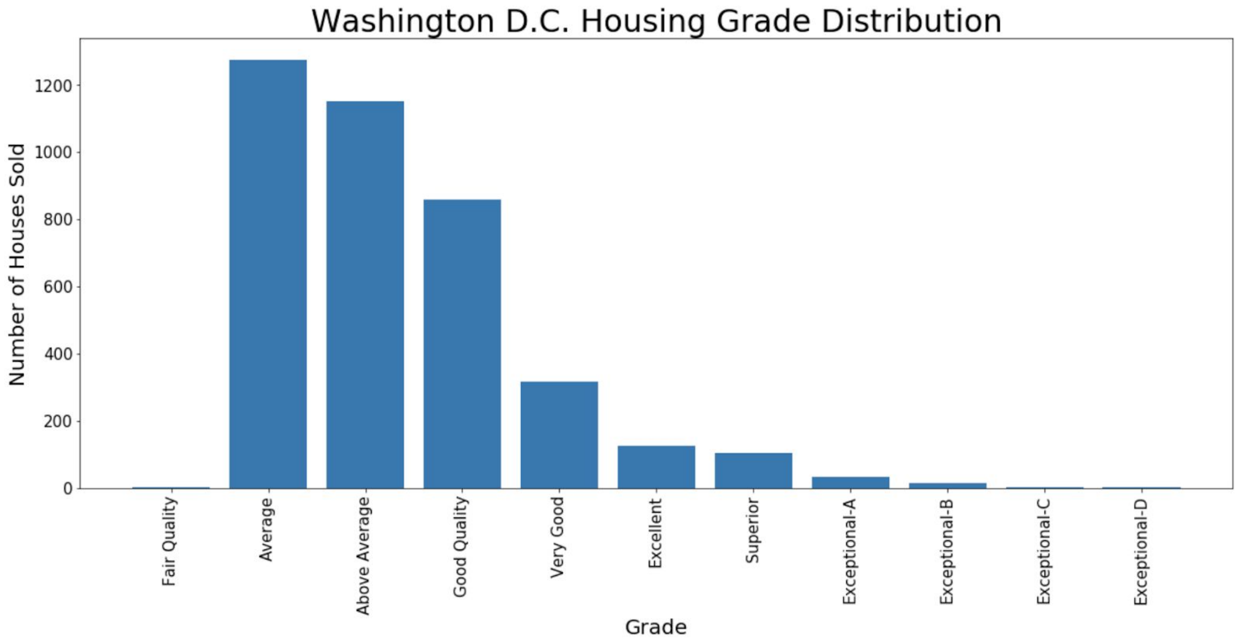


*Figure 11: Washington D.C. housing grade categorical distribution*

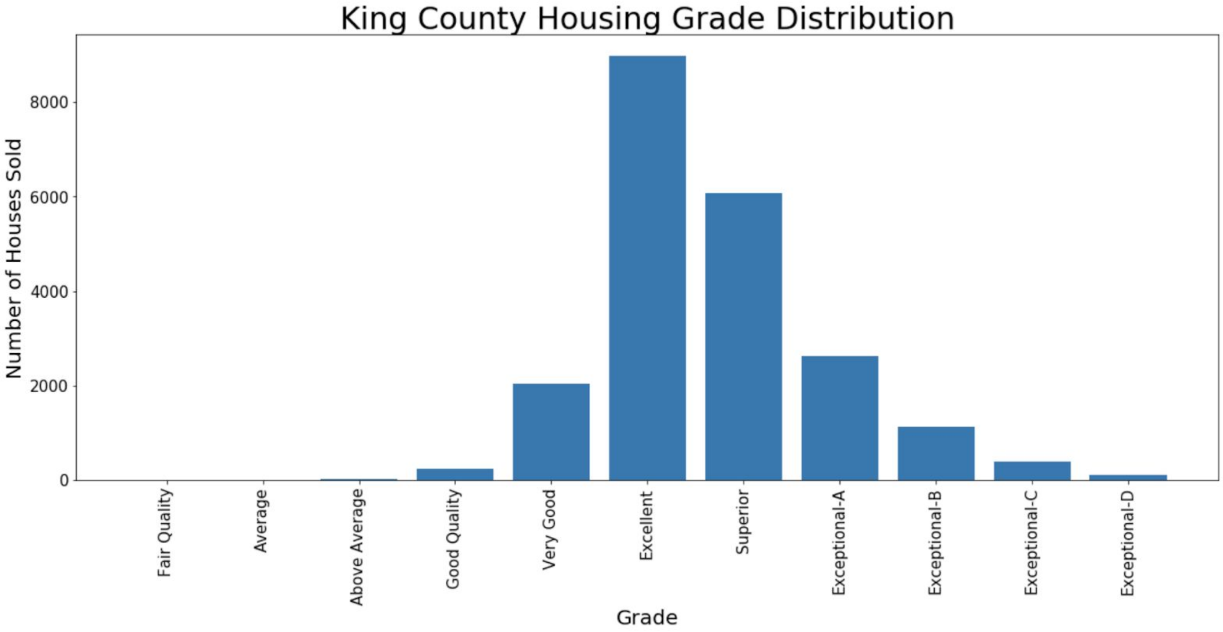


*Figure 12: King County housing grade categorical distribution*

The two distributions really highlight some important housing quality features between Washington D.C. and King County. Firstly, King County has much less 'Average' condition houses sold and much more 'Good' condition houses sold compared to D.C. As a result, on average each house sold in King County was in better shape compared to those in D.C. Additionally, there is a clear visual distinction between the Washington D.C. Grade distribution and King County Grade distribution. Houses appear to be much nicer and high-end compared to those in Washington D.C. with the majority of houses sold being graded either "Excellent" or "Superior" compared to the majority of Washington D.C. houses being either "Average" or "Above Average".

## d) Does condition affect the selling price?

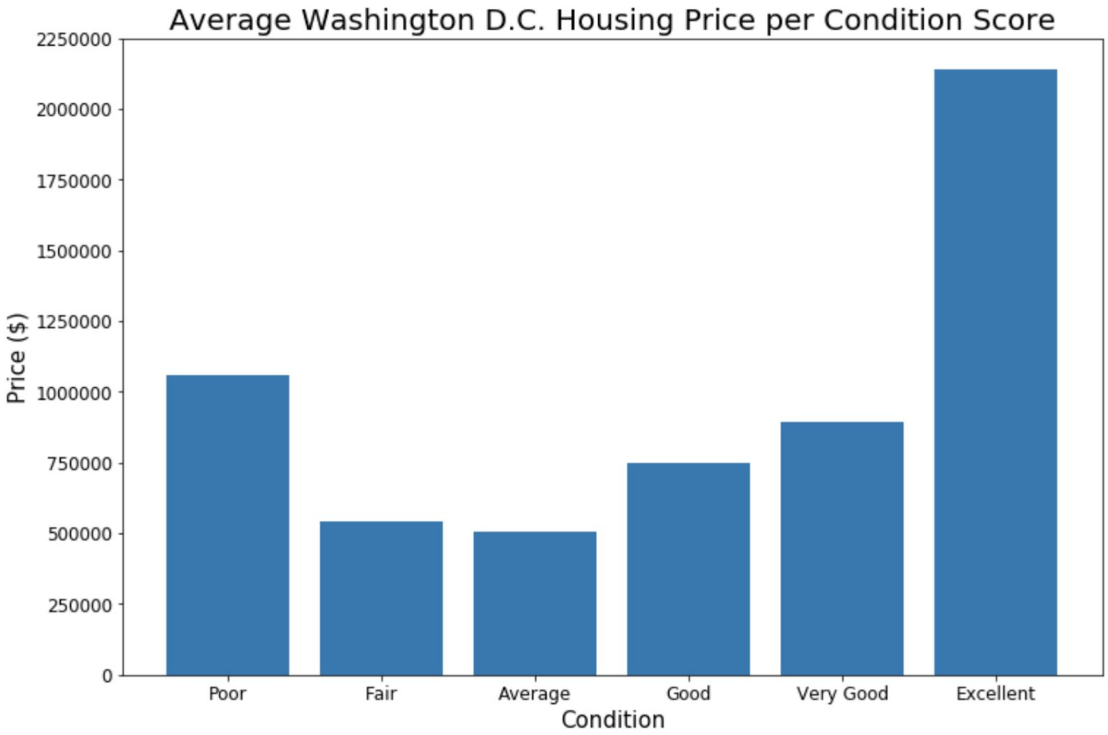


*Figure 13: Washington D.C. Housing Price per Condition Score*

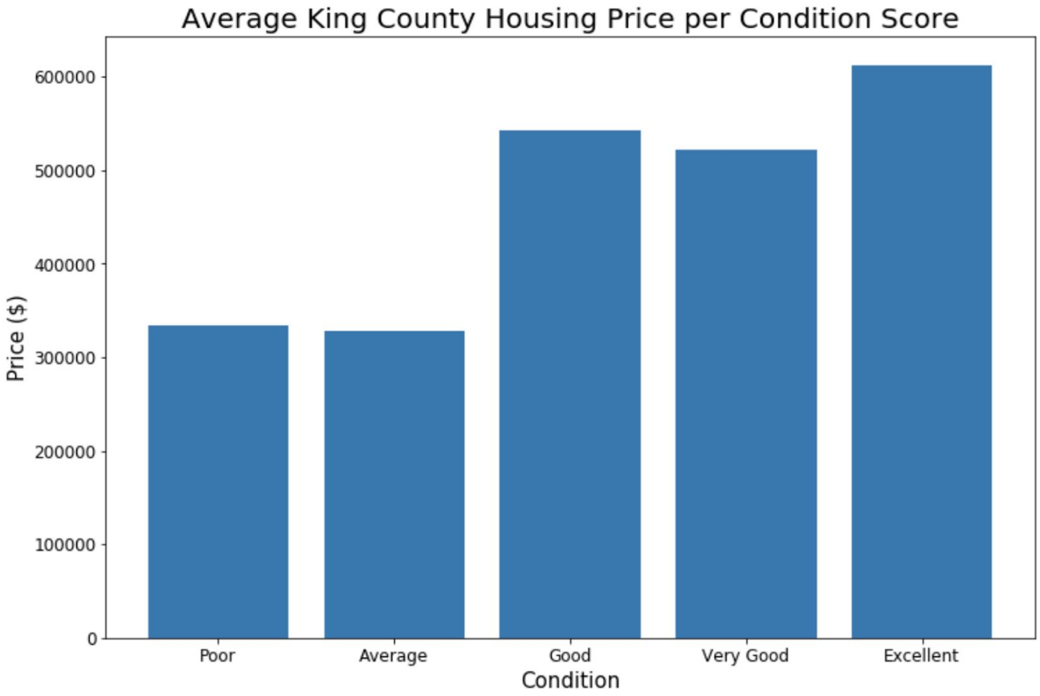


*Figure 14: King County Housing Price per Condition Score*

One of the most surprising things from the above graphs is the average price for 'Poor' condition houses in Washington D.C. The average price for 'Poor' condition houses in Washington D.C. was actually higher than any other condition other than the highest "Excellent" one. Unlike King County, which follows an expected pattern of higher prices for better housing conditions, the Washington D.C. data is unexpected here.

## e) Does grade affect the selling price?

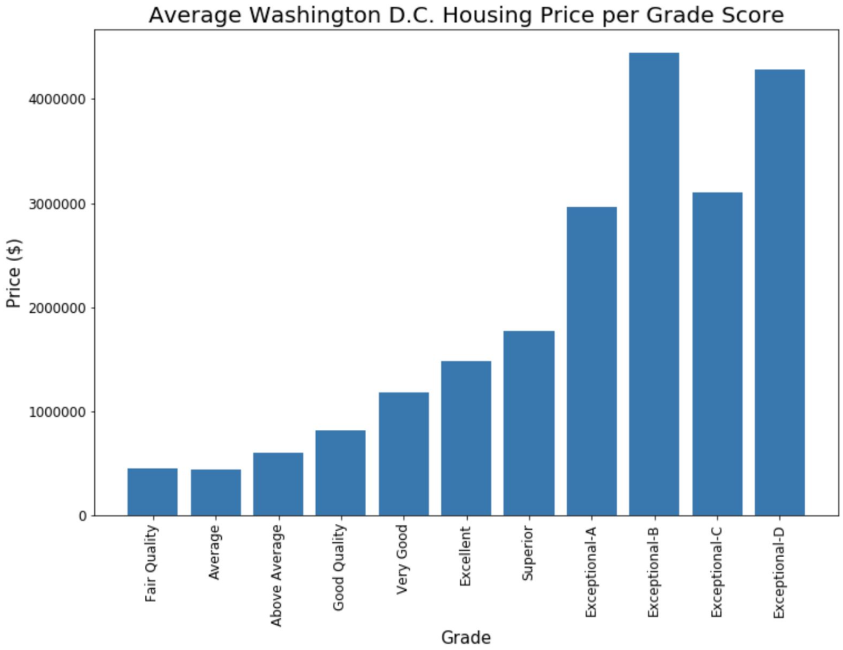


*Figure 15: Washington D.C. Housing Price per Grade Score*

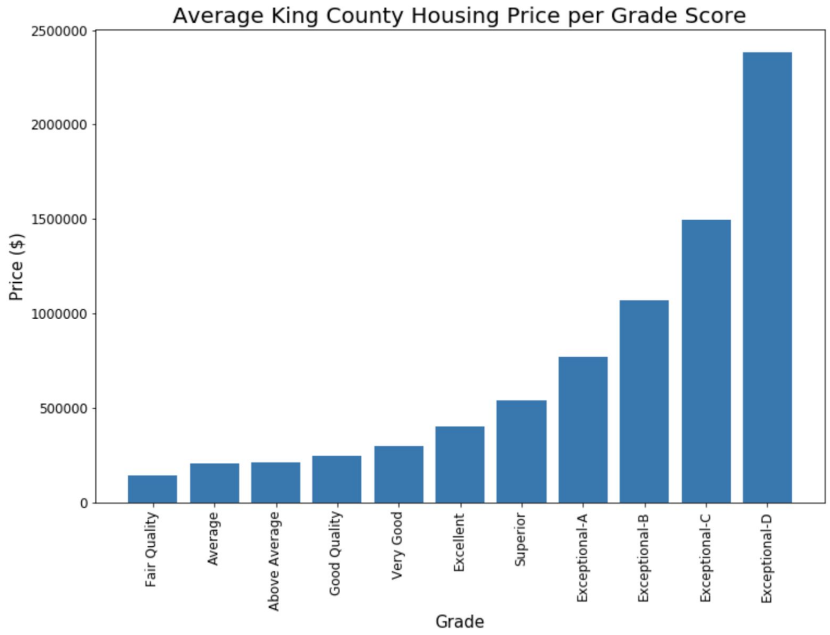
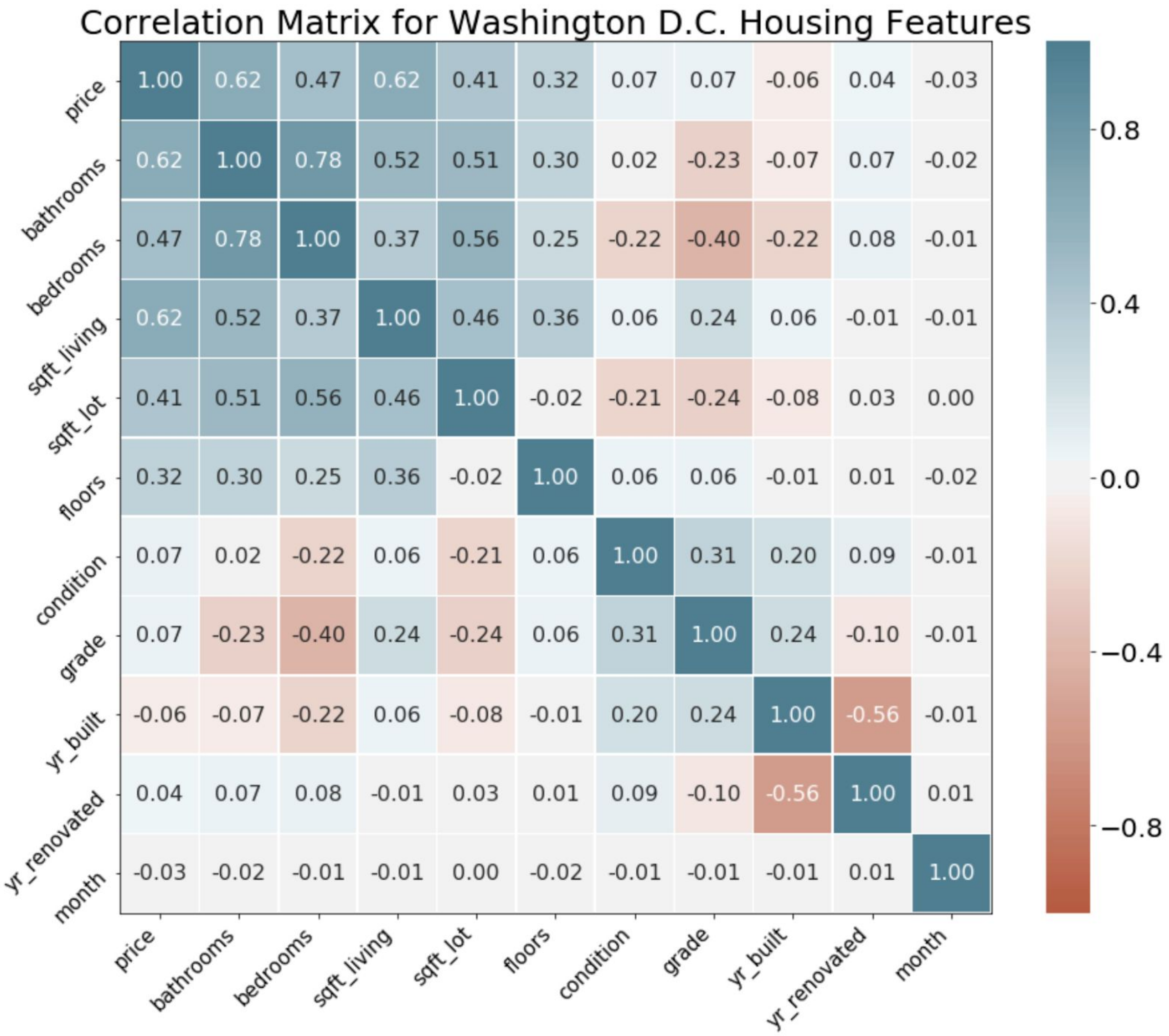


*Figure 16: King County Housing Price per Grade Score*

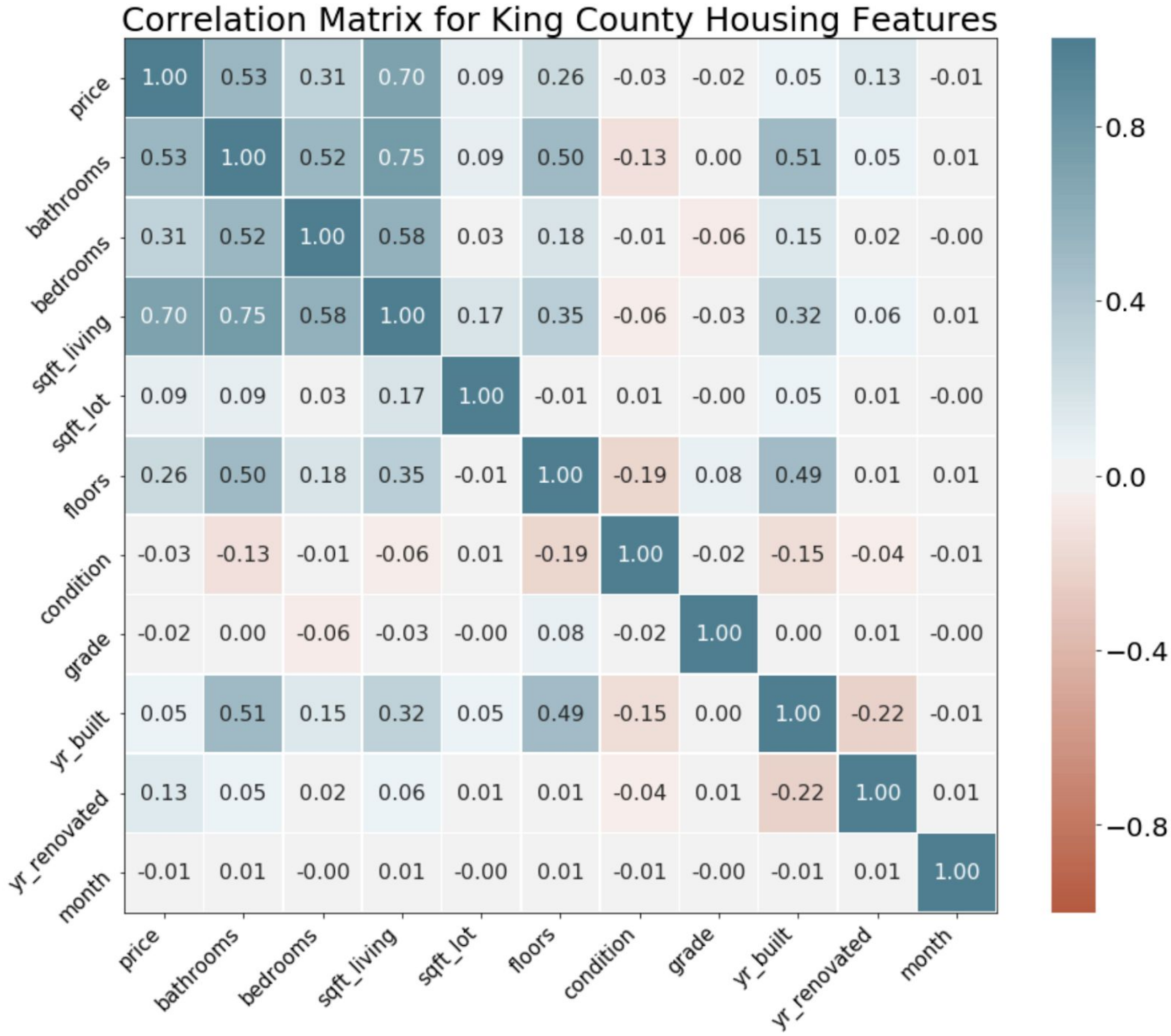
For both charts, it appears that the better grade a house received, the higher the average selling price for the house would be. There is one interesting thing about the Washington D.C. selling price for "Exceptional-C" grade. This grade is the second-highest grade a house can receive, yet there is a dip in the average selling price compared to grades lower than it.

## f) Correlation Heatmap for Housing Features



| | price | bathrooms | bedrooms | sqft_living | sqft_lot | floors | condition | grade | yr_built | yr_renovated | month |
|---|---|---|---|---|---|---|---|---|---|---|---|
| price | 1.00 | 0.62 | 0.47 | 0.62 | 0.41 | 0.32 | 0.07 | 0.07 | -0.06 | 0.04 | -0.03 |
| bathrooms | 0.62 | 1.00 | 0.78 | 0.52 | 0.51 | 0.30 | 0.02 | -0.23 | -0.07 | 0.07 | -0.02 |
| bedrooms | 0.47 | 0.78 | 1.00 | 0.37 | 0.56 | 0.25 | -0.22 | -0.40 | -0.22 | 0.08 | -0.01 |
| sqft_living | 0.62 | 0.52 | 0.37 | 1.00 | 0.46 | 0.36 | 0.06 | 0.24 | 0.06 | -0.01 | -0.01 |
| sqft_lot | 0.41 | 0.51 | 0.56 | 0.46 | 1.00 | -0.02 | -0.21 | -0.24 | -0.08 | 0.03 | 0.00 |
| floors | 0.32 | 0.30 | 0.25 | 0.36 | -0.02 | 1.00 | 0.06 | 0.06 | -0.01 | 0.01 | -0.02 |
| condition | 0.07 | 0.02 | -0.22 | 0.06 | -0.21 | 0.06 | 1.00 | 0.31 | 0.20 | 0.09 | -0.01 |
| grade | 0.07 | -0.23 | -0.40 | 0.24 | -0.24 | 0.06 | 0.31 | 1.00 | 0.24 | -0.10 | -0.01 |
| yr_built | -0.06 | -0.07 | -0.22 | 0.06 | -0.08 | -0.01 | 0.20 | 0.24 | 1.00 | -0.56 | -0.01 |
| yr_renovated | 0.04 | 0.07 | 0.08 | -0.01 | 0.03 | 0.01 | 0.09 | -0.10 | -0.56 | 1.00 | 0.01 |
| month | -0.03 | -0.02 | -0.01 | -0.01 | 0.00 | -0.02 | -0.01 | -0.01 | -0.01 | 0.01 | 1.00 |

*Figure 17: Washington D.C. housing feature correlation matrix*



| | price | bathrooms | bedrooms | sqft_living | sqft_lot | floors | condition | grade | yr_built | yr_renovated | month |
|---|---|---|---|---|---|---|---|---|---|---|---|
| price | 1.00 | 0.53 | 0.31 | 0.70 | 0.09 | 0.26 | -0.03 | -0.02 | 0.05 | 0.13 | -0.01 |
| bathrooms | 0.53 | 1.00 | 0.52 | 0.75 | 0.09 | 0.50 | -0.13 | 0.00 | 0.51 | 0.05 | 0.01 |
| bedrooms | 0.31 | 0.52 | 1.00 | 0.58 | 0.03 | 0.18 | -0.01 | -0.06 | 0.15 | 0.02 | -0.00 |
| sqft_living | 0.70 | 0.75 | 0.58 | 1.00 | 0.17 | 0.35 | -0.06 | -0.03 | 0.32 | 0.06 | 0.01 |
| sqft_lot | 0.09 | 0.09 | 0.03 | 0.17 | 1.00 | -0.01 | 0.01 | -0.00 | 0.05 | 0.01 | -0.00 |
| floors | 0.26 | 0.50 | 0.18 | 0.35 | -0.01 | 1.00 | -0.19 | 0.08 | 0.49 | 0.01 | 0.01 |
| condition | -0.03 | -0.13 | -0.01 | -0.06 | 0.01 | -0.19 | 1.00 | -0.02 | -0.15 | -0.04 | -0.01 |
| grade | -0.02 | 0.00 | -0.06 | -0.03 | -0.00 | 0.08 | -0.02 | 1.00 | 0.00 | 0.01 | -0.00 |
| yr_built | 0.05 | 0.51 | 0.15 | 0.32 | 0.05 | 0.49 | -0.15 | 0.00 | 1.00 | -0.22 | -0.01 |
| yr_renovated | 0.13 | 0.05 | 0.02 | 0.06 | 0.01 | 0.01 | -0.04 | 0.01 | -0.22 | 1.00 | 0.01 |
| month | -0.01 | 0.01 | -0.00 | 0.01 | -0.00 | 0.01 | -0.01 | -0.00 | -0.01 | 0.01 | 1.00 |

*Figure 18: King County housing feature correlation matrix*

Creating a correlation matrix between different housing features will help visualize which features have a high impact on the target and other features. One interesting column to look into is the 'price' column to see which other features have a high correlation with the dependent variable I will be examining. One distinct feature I would like to point out is 'sqft_lot'. In the Washington D.C. matrix, there is a value of 0.41, whereas in the King County matrix this value is 0.09. It is interesting that there is a large difference between the effect the lot square footage has on housing sales in these different locations. It could be that Washington D.C. doesn't have as many houses with a large plot of land, so land can be more valuable when pricing a house compared to King County where there is a wide range of lot sizes.

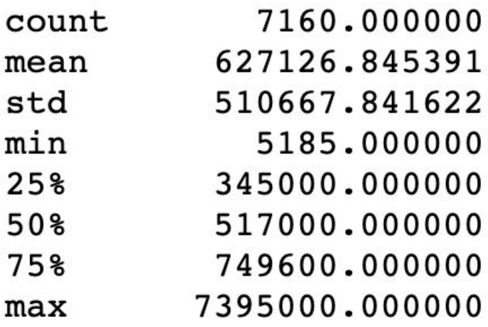
# 6. Applications of Inferential Statistics

## a) Average Housing Price: Is there a statistical significant difference between the average housing sale price between Washington D.C. and King County?

The code that I used for parts of this test can be seen in my notebook [3]. I compared the difference of means between the housing sale prices in Washington D.C and King County by testing the following hypotheses:

- $H_0$ : The true mean housing sale price between the Washington D.C. and King County are the same
- $H_1$ : The true mean housing sale price between the Washington D.C. and King County are not the same

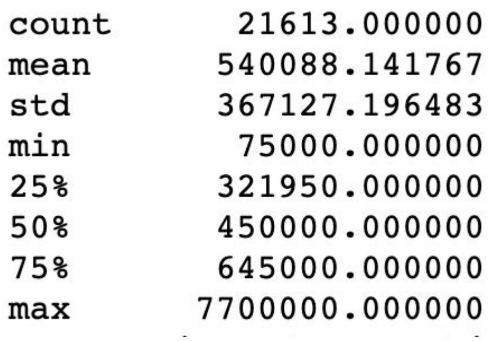
For this test, I assumed an alpha-level of 0.05.



```
count       7160.000000
mean      627126.845391
std       510667.841622
min         5185.000000
25%       345000.000000
50%       517000.000000
75%       749600.000000
max      7395000.000000
```

*Figure 19: Washington D.C. sale price summary statistics*

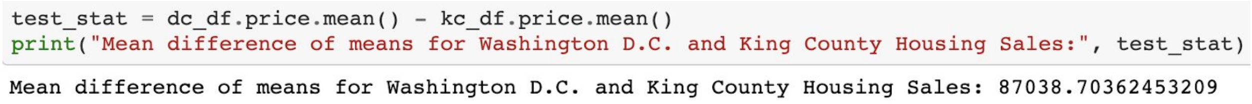
[3] http://onlinestatbook.com/mobile/tests_of_means/difference_means.html



```
count       21613.000000
mean       540088.141767
std        367127.196483
min         75000.000000
25%        321950.000000
50%        450000.000000
75%        645000.000000
max       7700000.000000
```

*Figure 20: King County sale price statistics*

**i) Compute the test statistic: Mean of the sampling distribution of the difference between means**



```
test_stat = dc_df.price.mean() - kc_df.price.mean()
print("Mean difference of means for Washington D.C. and King County Housing Sales:", test_stat)
```

```
Mean difference of means for Washington D.C. and King County Housing Sales: 87038.70362453209
```

In order to test for the difference of means between the Washington D.C. and King County housing prices data sets, there are 3 assumptions that I am going to make:

1. The Washington D.C. and King County housing data sets have the same variance
2. Each housing price population is normally distributed
3. Each housing price is sampled independently from each other value. This assumption means that each housing unit sold is for one value only

**ii) Calculate the Standard Error of the test statistic**

The formula for the variance of the sampling distribution of the mean is:

$$\sigma_M^2 = \frac{\sigma^2}{N}$$

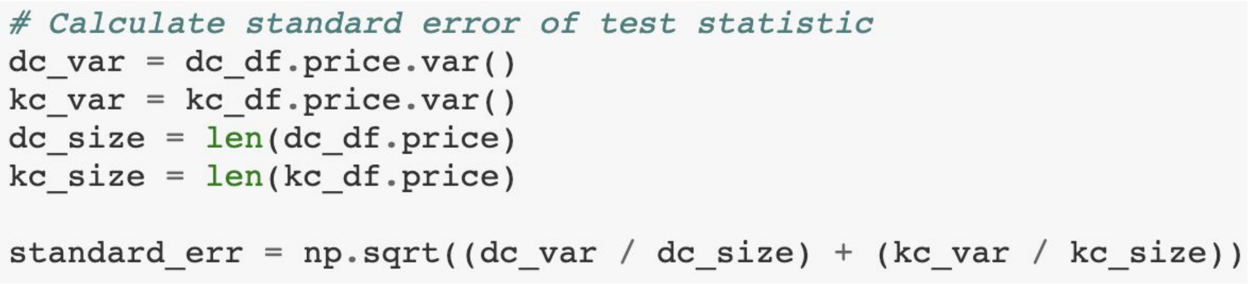
where σ is the standard deviation and N is the sample size. Because the Washington D.C. and King County are different populations and have different sample sizes, we need to distinguish between them via subscripts to represent each population:

$$\sigma^2_{M_1-M_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Using the above formulas, I can use the following formula to calculate the standard error of the difference of means between the two populations.:

$$\sigma_{M_1-M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

The following is the code running these calculations:

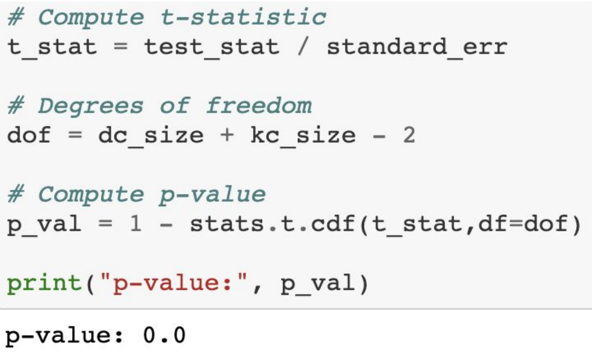


```
# Calculate standard error of test statistic
dc_var = dc_df.price.var()
kc_var = kc_df.price.var()
dc_size = len(dc_df.price)
kc_size = len(kc_df.price)

standard_err = np.sqrt((dc_var / dc_size) + (kc_var / kc_size))
```

Now that I had the standard error, I plugged the value it into the equation below to get the t-statistic and used this to get the probability (p-value) of getting a t as large or larger than the t-statistic or as small or smaller than -(t-statistic). :

$$t = \frac{\text{statistic - hypothesized value}}{\text{estimated standard error of the statistic}}$$



```
# Compute t-statistic
t_stat = test_stat / standard_err

# Degrees of freedom
dof = dc_size + kc_size - 2

# Compute p-value
p_val = 1 - stats.t.cdf(t_stat,df=dof)

print("p-value:", p_val)
```

```
p-value: 0.0
```

Since the p-value was less than 0.05, I could confidently reject my null hypothesis and therefore conclude that the observed difference in the means is statistically significant--informally, this means that the observed difference is likely not to be due to chance..

# 7. Linear Relationships Between Non-categorical Housing Features and Housing Prices

By examining a linear regression model for each individual non-categorical housing features with the housing sale price, I could determine which features are statistically significant predictors of housing sale prices. I went through each non-categorical feature and plotted the feature against the price to first see any obvious correlations between the two.

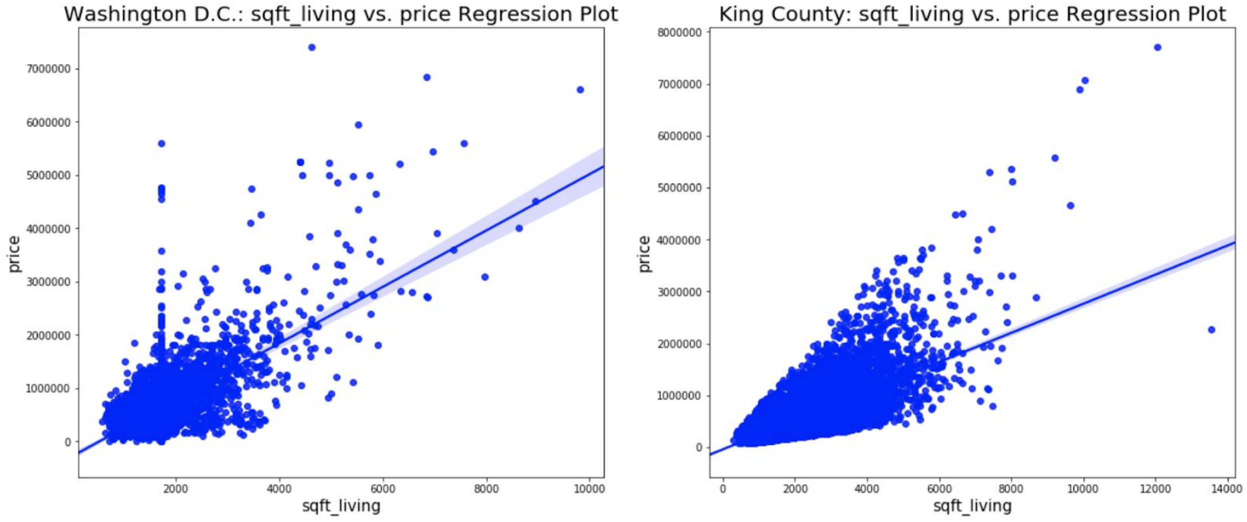
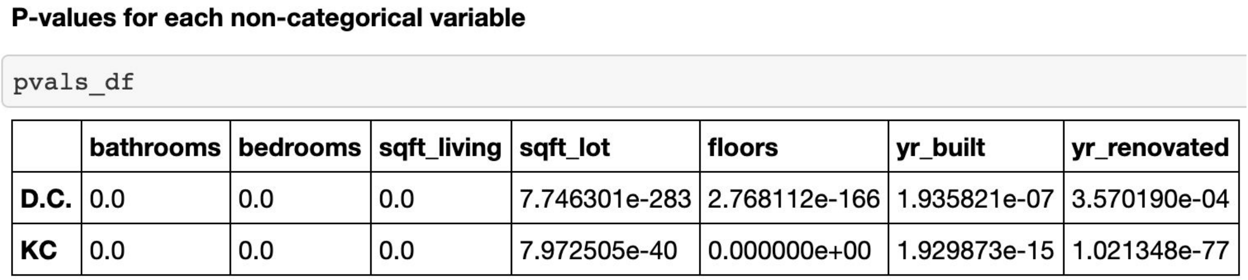


*Figure 21: Square Feet Linear vs. Price Linear Relationship*

The above plot is one example of a linear regression model being built for each data set. By isolating each individual non-categorical feature with the price, the models could reveal some patterns and effects that the individual features had on the dependent variable I was examining. From above, it appears that there is a clear positive correlation between the living square feet and price. For each feature, I pulled the p-value of the resulting linear regression model to help determine if the non-categorical feature was a significant predictor for price:



**P-values for each non-categorical variable**

```
pvals_df
```

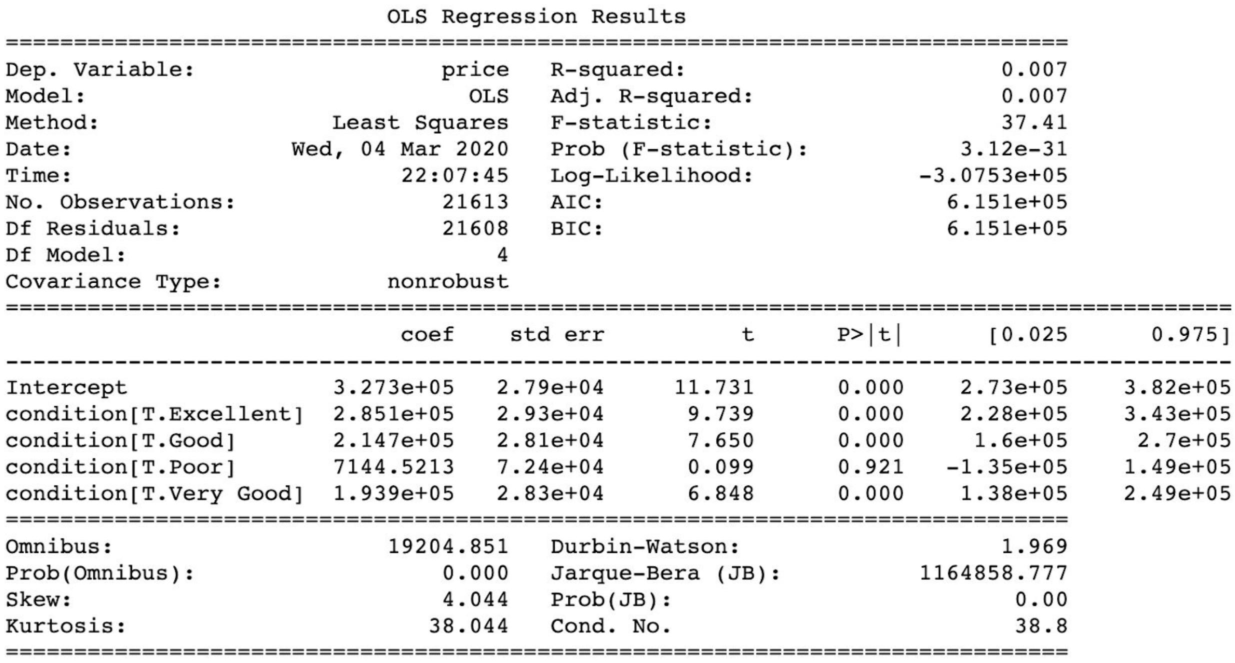
| | bathrooms | bedrooms | sqft_living | sqft_lot | floors | yr_built | yr_renovated |
|---|---|---|---|---|---|---|---|
| **D.C.** | 0.0 | 0.0 | 0.0 | 7.746301e-283 | 2.768112e-166 | 1.935821e-07 | 3.570190e-04 |
| **KC** | 0.0 | 0.0 | 0.0 | 7.972505e-40 | 0.000000e+00 | 1.929873e-15 | 1.021348e-77 |

*Figure 21: P-values for each non-categorical feature in the Washington D.C. and King County data sets*

Because the p-value for each non-categorical variable in both the Washington D.C. data set and King County data set is less than the alpha value of 0.05, each variable is deemed to be a statistically significant predictor for price.

# 8. Linear Regression Models for Categorical Features on Price

Now that I had looked at each non-categorical feature, now it was time to look at the categorical features: grade and condition. For each of these, I used *statsmodels "ols"* function to build the linear regression model. This function was very helpful because it automatically transforms each categorical variable to a dummy variable to be usable with predicting price (a continuous variable). Below is an example of the model built with the *grade* feature for the King County data set:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.007
Model:                            OLS   Adj. R-squared:                  0.007
Method:                 Least Squares   F-statistic:                     37.41
Date:                Wed, 04 Mar 2020   Prob (F-statistic):           3.12e-31
Time:                        22:07:45   Log-Likelihood:            -3.0753e+05
No. Observations:               21613   AIC:                         6.151e+05
Df Residuals:                   21608   BIC:                         6.151e+05
Df Model:                           4
Covariance Type:            nonrobust
============================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Intercept               3.273e+05   2.79e+04     11.731      0.000    2.73e+05    3.82e+05
condition[T.Excellent]  2.851e+05   2.93e+04      9.739      0.000    2.28e+05    3.43e+05
condition[T.Good]       2.147e+05   2.81e+04      7.650      0.000     1.6e+05     2.7e+05
condition[T.Poor]       7144.5213   7.24e+04      0.099      0.921   -1.35e+05    1.49e+05
condition[T.Very Good]  1.939e+05   2.83e+04      6.848      0.000    1.38e+05    2.49e+05
==============================================================================
Omnibus:                    19204.851   Durbin-Watson:                   1.969
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1164858.777
Skew:                           4.044   Prob(JB):                         0.00
Kurtosis:                      38.044   Cond. No.                         38.8
==============================================================================
```

*Figure 22: OLS linear regression model using condition as the predictor*

The above model works by using one of the *condition* categories as a baseline. From there, each other category is compared to this baseline model. For example, the condition category "average" is used as the baseline. "Poor" condition above has a p-value of .921. This p-value means that we cannot say that this category is a statistically significant predictor for "average" houses sold, whereas every other category has a p-value that is below the threshold I am using. The coefficients for each category also help distinguish how additional *conditions* affect the baseline model. I used this for both the Washington D.C. and King County data set on both the *grade* and *condition* features to see how each affects the price individually.

## 9. Next Steps

Following the baseline Linear Regression model, for the next step I will create a Random Forest Regressor. Random Forest is an ensemble learning method that utilizes decision trees to decide which features to place higher importance on to make predictions for the housing prices. This extension helps illustrate how different types of models can affect the prediction power and reveal other intricacies in the data.