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

By: Garrett Yamane

Springboard--DSC Program

Capstone Project 2

Problem Statement

- **Target Client**
  - Real estate agents and prospective house buyers
- **Goals**
  - Build individual accurate predictive machine learning model for both Washington D.C. and King County
  - Closely examine what housing features heavily impact the price of a house sale
  - Help prospective buyers understand the market they are buying into

Data Acquisition and Wrangling

- Sources:
  - Washington D.C. data (7160 rows)
  - King County data (21613 rows)
- Filtered D.C. data to only contain sales from May 2014-May 2015 and by the columns both sets had in common
- Converted columns to have same type and imputed values in columns where missing values were present

```
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
```

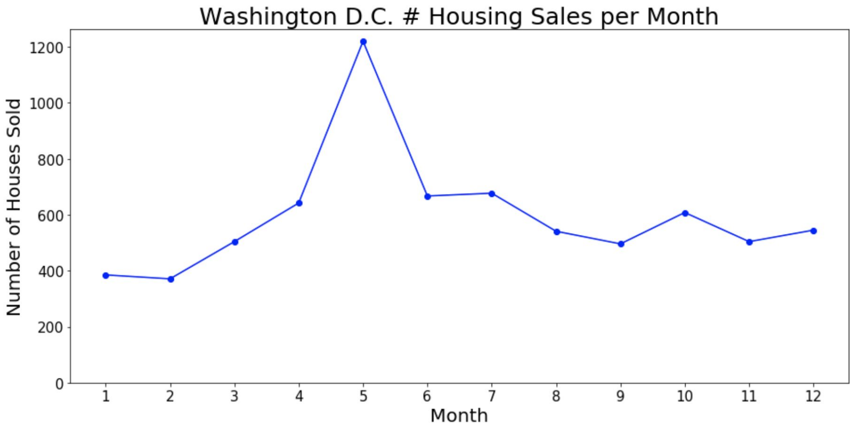
Missing Value % in Final Data Frame

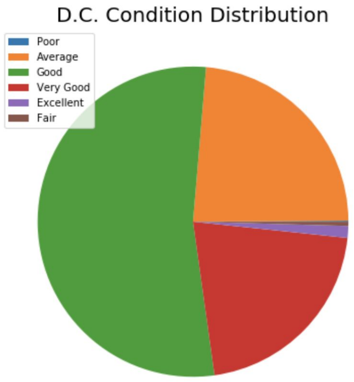
Snapshot of Final Data Frame for Washington D.C

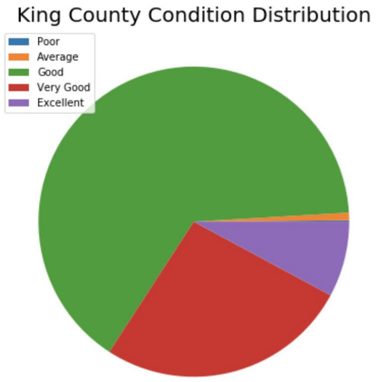
	price	date	bathrooms	bedrooms	sqft_living	sqft_lot	floors	condition	grade	yr_built	yr_renovated	location
1	993500.0	2014-10-08	5.0	3	1148.0	814	2.0	Very Good	Average	1907	2014	DC
2	1280000.0	2014-08-19	2.5	3	1630.0	1000	2.0	Good	Good Quality	1906	2004	DC
4	1440000.0	2015-04-22	3.5	4	1686.0	1424	2.0	Very 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

Initial Findings

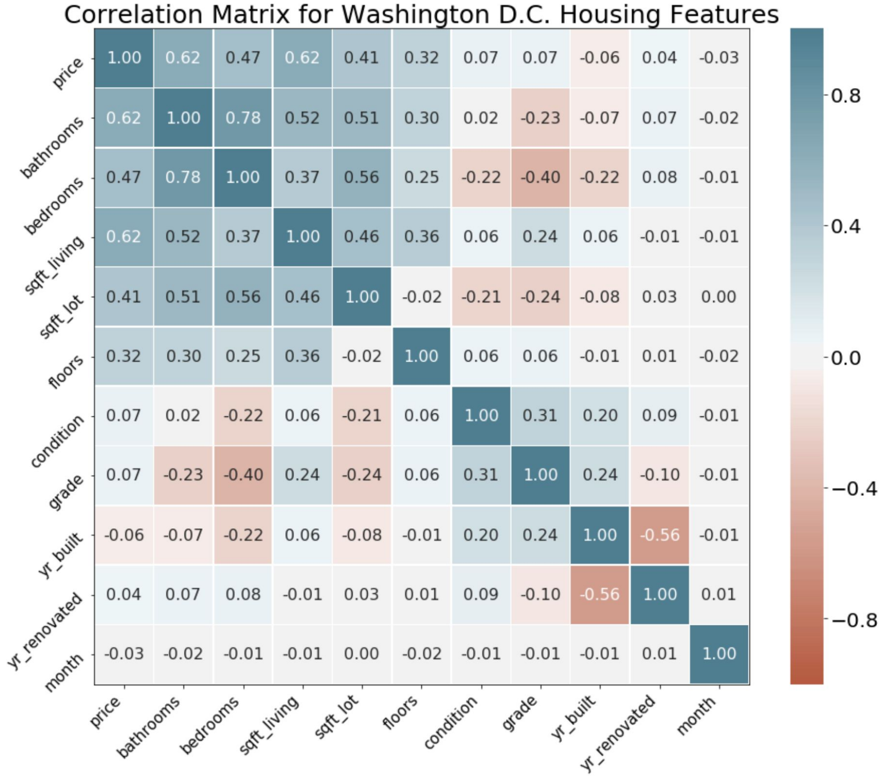
- Focus Questions:
  - Are houses more likely to sell at specific times of the year?
  - Does condition and grade affect the selling price?
  - What features are highly correlated?

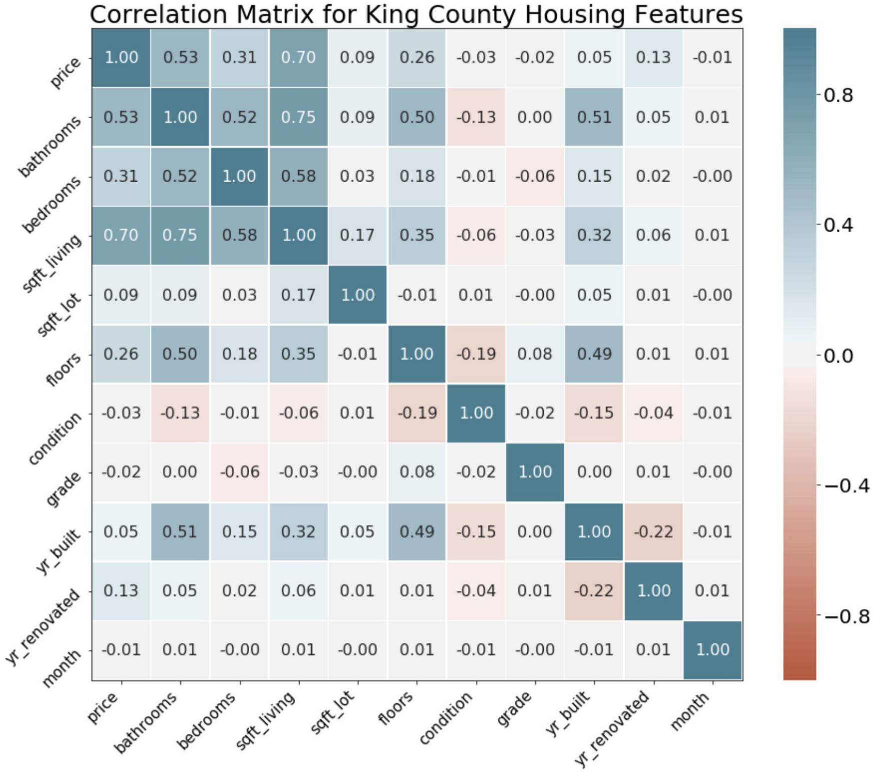






Correlation Heatmap for Housing Features





Application of Inferential Statistics

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

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

Washington D.C. sale price summary statistics

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

King County sale price summary statistics

Application of Inferential Statistics: Results

- Results: *p-value* was less than 0.05
- 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.

```
# 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
```

Baseline Model: Linear Regression

- First, I built out a linear regression model using all housing features for each data set
- Below are the model coefficients for each

Washington D.C. Model Coefficients ¶

```
# Washington D.C. linear regression model coefficients
lm_dc.params
```

```
Intercept        0.006840
bathrooms        0.415863
bedrooms        -0.014473
sqft_living      0.348688
sqft_lot         0.072633
floors           0.065963
condition        0.044293
grade            0.091308
yr_built        -0.094210
yr_renovated    -0.034802
month           -0.007449
dtype: float64
```

King County Model Coefficients

```
# King County linear regression model coefficients
lm_kc.params
```

```
Intercept       -0.002498
bathrooms        0.139652
bedrooms        -0.170795
sqft_living      0.751963
sqft_lot        -0.036162
floors           0.079607
condition       -0.000993
grade           -0.014047
yr_built        -0.258198
yr_renovated     0.011617
month           -0.022038
dtype: float64
```

Feature Selection: Lasso Regression

- Benefit of Lasso Regression: zeros out coefficients for features with a high penalty

	features	estimatedCoefficients
0	bathrooms	0.400750
1	bedrooms	0.000000
2	sqft_living	0.351308
3	sqft_lot	0.059128
4	floors	0.049667
5	condition	0.034085
6	grade	0.086690
7	yr_built	-0.073350
8	yr_renovated	-0.007563
9	month	-0.000000

Washington D.C. Lasso Regression Model Coefficients

	features	estimatedCoefficients
0	bathrooms	0.114235
1	bedrooms	-0.141313
2	sqft_living	0.749540
3	sqft_lot	-0.021016
4	floors	0.070464
5	condition	0.000000
6	grade	-0.000000
7	yr_built	-0.238361
8	yr_renovated	0.018938
9	month	-0.013999

King County Lasso Regression Model Coefficients

Baseline Model Extension: Random Forest

- Random Forest Regressor benefits:
  - Avoids overfitting on training data
  - Aggregates multiple independently built models into a single optimal model
- OOB Error
  - Main metric used to determine the strength of each random forest model
- Hyperparameter Tuning
  - *max_features*: how many random different features to consider in order to decide how to proceed down the decision tree
  - *n_estimators*: the number of trees that are made in the forest

Random Forest Models: Washington D.C.

Random Forest Models: King County

Random Forest: Feature Importance

	Features	Importance Value
0	bathrooms	0.234591
1	bedrooms	0.070417
2	sqft_living	0.237226
3	sqft_lot	0.099992
4	floors	0.039018
5	condition	0.032830
6	grade	0.108064
7	yr_built	0.095940
8	yr_renovated	0.039032
9	month	0.042889

Washington D.C. Random Forest Regressor Feature Importance Values

	Features	Importance Value
0	bathrooms	0.156772
1	bedrooms	0.037921
2	sqft_living	0.360064
3	sqft_lot	0.088141
4	floors	0.025835
5	condition	0.019109
6	grade	0.150021
7	yr_built	0.105644
8	yr_renovated	0.016768
9	month	0.039724

King County Random Forest Regressor Feature Importance Values

Evaluation of the Baseline and Extension Models

- **R-squared Value**: Proportion of the variance explained by the model
  - D.C. Linear Regression R-squared value: 0.51
  - KC Linear Regression R-squared value: 0.56
  - D.C. Random Forest R-squared value: 0.82
  - KC Random Forest R-squared value: 0.65
- **Mean Absolute Value:** Average magnitude of the errors for a set of predictions by a model
  - D.C. Linear Regression MAE: 0.4
  - KC Linear Regression MAE: 0.44
  - D.C. Random Forest MAE: 0.25
  - KC Random Forest MAE: 0.36

Conclusion

- Big Takeaways
    - Models that offer more flexibility to fit a dataset like the Random Forest model made more accurate and precise predictions for both Washington D.C. and King County
    - Building individual models for geographically different locations highlighted housing features that carried more "importance" or affected the price heavier
    - No one model can encompass all housing data

Future Work and Recommendations

- Future Work
    - Feature Engineering
        - Build out customized features from the raw data to see how R-squared and MAE is affected
    - In-depth hyperparameter tuning for Random Forest Regressors
        - Increase range of *n_esimators*
        - Test a wider variety of *max_features* for each model

- Recommendations
    - Prospective Buyers
        - Target focus on what is a hot commodity when it comes to buying a house in their area of interest
        - Some features can heavily impact the sale price: should always keep in mind what is most important to them and if it is worth the price difference
    - Real estate agents
        - Monitor housing sale trends
        - Study how specific geographic locations have been changing and if housing prices are increasing/decreasing due to these changes