DJUWITA CARNEY, MAY 13, 2020

Predictive Modeling

For DC Housing Price

DJUWITA CARNEY, MAY 13, 2020

DC Living

Taxation without representation

DJUWITA CARNEY, MAY 13, 2020

Problem Statement

Creating a predictive model to estimate DC housing price based on the Kaggle Dataset

Datasets:

CSV file from Kaggle consists of 28900 lines, 46 columns

EDA:

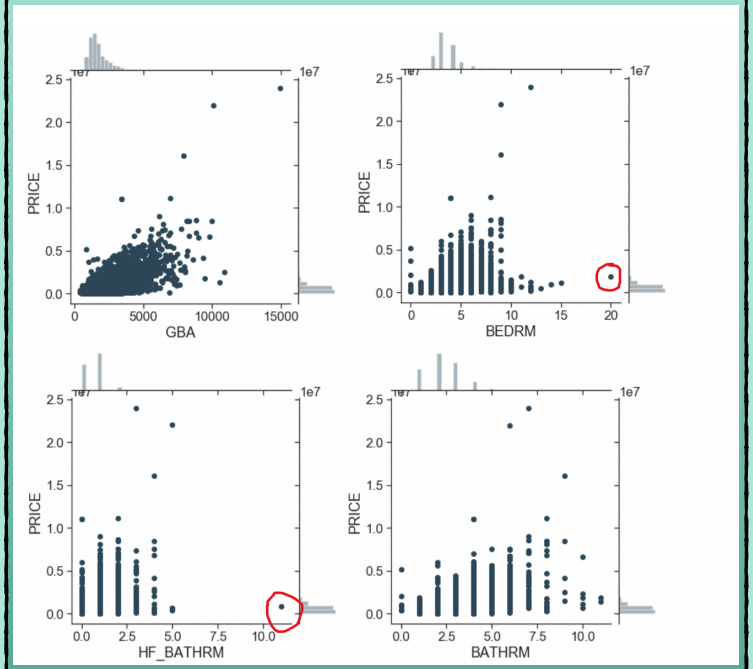
- Check fo missing data
- Irrelevant variables elimination
- Outliers handling
- Check for other anomalies:
  - Unrealistic maximum price in SE Quadrant
  - Unrealistically low prices
  - Data type error in SALEDATE

EDA:

- No missing data
- Irrelevant variables:'QUALIFIED', 'SALE_NUM', 'GIS_LAST_MOD_DTTM', 'SOURCE', 'CITY', 'ZIPCODE', 'NATIONALGRID','LATITUDE', 'LONGITUDE', ASSESSMENT_NBHD','ASSESSMENT_SUBNBHD', 'CENSUS_TRACT','CENSUS_BLOCK', 'WARD', 'SQUARE','X', 'Y'

EDA:

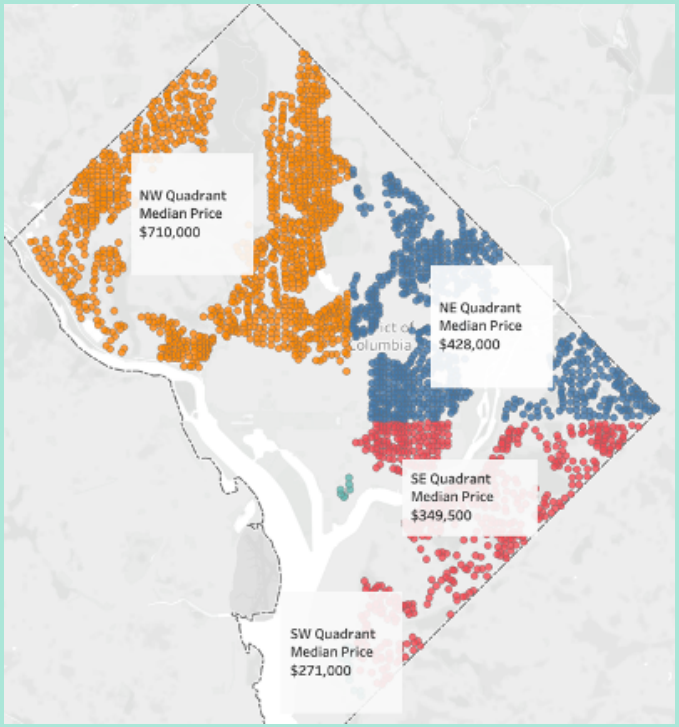
- Outliers handling

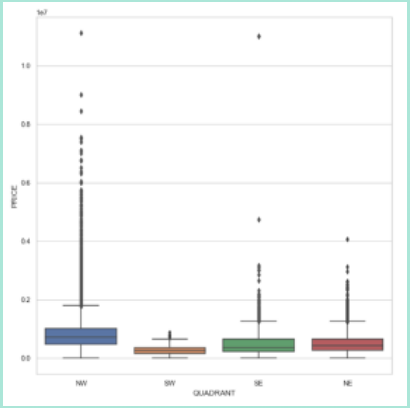


DJUWITA CARNEY, MAY 13, 2020

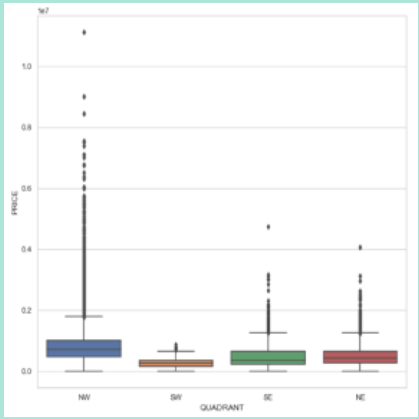
EDA:

- Unrealistic maximum price in SE Quadrant





Original data



Internet adjusted data

DJUWITA CARNEY, MAY 13, 2020

EDA:

- Unrealistic minimum prices

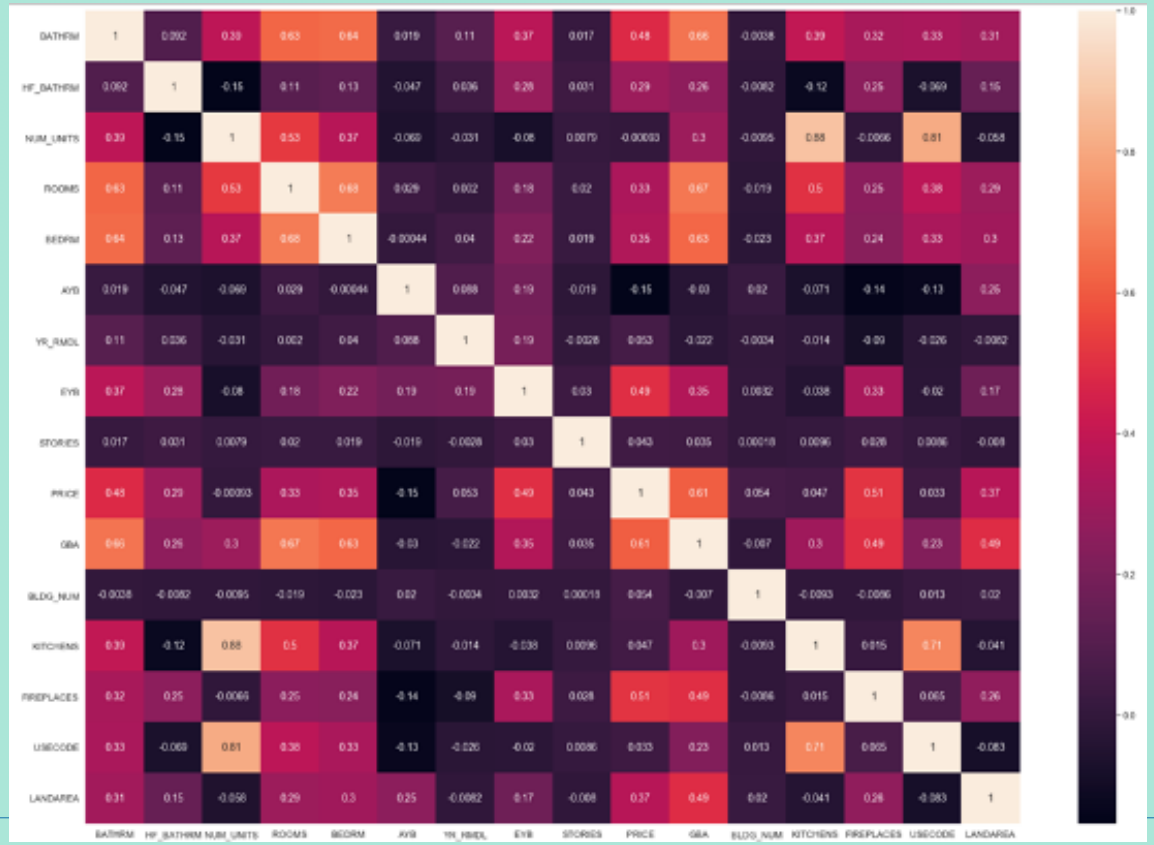
QUADRANT	PRICE
NE	250
NW	10
SE	250
SW	1

- Data type error for SALEDATE

DJUWITA CARNEY, MAY 13, 2020

Analysis:

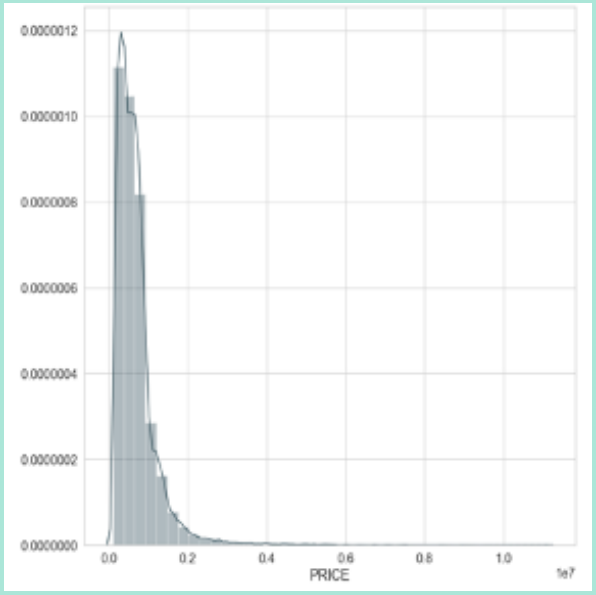
- Correlation coefficient between variables



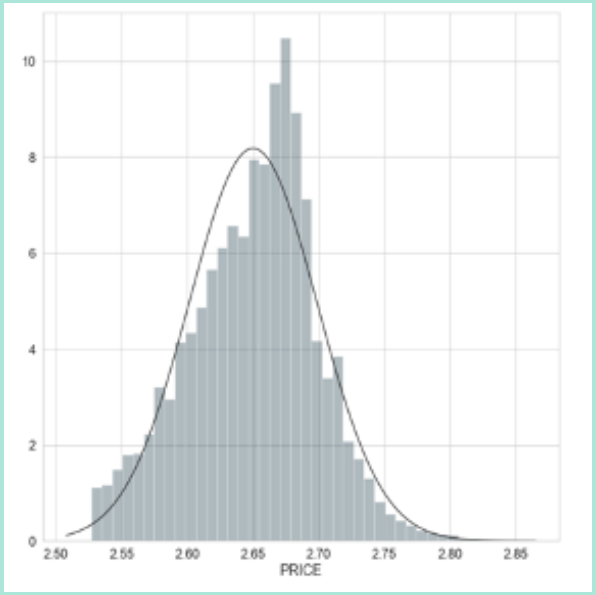
USE PCA to eliminate Correlation coefficients

Analysis:

- Check for normality



Original price distribution

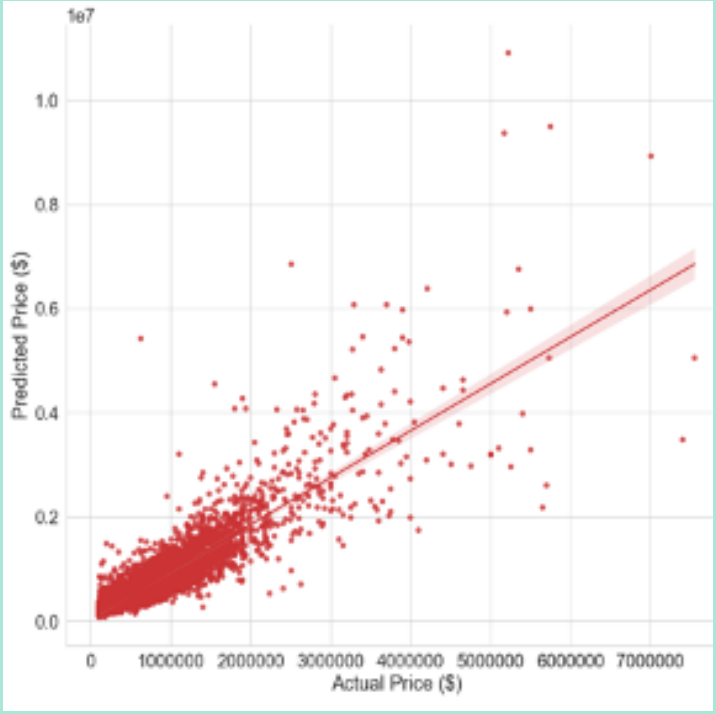


Normalized price distribution

Analysis:

- Create dummies ====> The final number of predictors is 162
- A lot of variables, use Elasticnet penalty to avoid overfitting
- Use PCA to eliminate dependency among predictive variables
- Split into 60% training data and 40% test data
- Linear modeling is used, fit , predict
- Check R^2 and plot the actual price vc prediction

Results:



Actual vc predicted price

Train score = 0.7925

Test score = 0.7026

DJUWITA CARNEY, MAY 13, 2020

Conclusion:

The house price in Washington DC was predicted based on the predictive variables such as building area, number of rooms, location, built year, renovated year and sold year, among other variables both numerical and categorical. Linear Regression model was applied resulted in the training and test data with the scores closed to **0.8**.

The model predicts low and medium prices relatively well. However, it does not perform as well for predicting higher prices. This might be due to the nature of data showing a few houses with very high prices that is not necessarily correlated with the most important variables.

References:

1. https://www.kaggle.com/christophercorrea/dc-residential-properties

2 .Redfin.com

3. https://seaborn.pydata.org/tutorial.html

4. https://matplotlib.org/3.2.1/contents.html

5. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

6. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html