DATE : November 19, 2020
SUBJECT : **Final Group Project** – Data Mining and Predictive Analytics Essentials
TEAM : **Charry DeAndres**, **Kaniz Syeda** *and* **Shahnaz Jalali**

Introduction

The high-level goal of our project is to develop machine-learning algorithms and create different *predictive* models to predict the price of a house within King County. We are motivated and interested to explore on which features the price of a house be dependent on and choose the best-fit regression model.

Source: King County House Sales dataset

Problem Statement(s):

- Compare 3 models and assess which model is more effective in predicting the price of houses in King County
- How much should we pay to purchase a house in King County that meets the following criteria:
  - 3 bedrooms, 1.75 bathrooms, 1520 sqft_living, 6380 sqft_lot, 1 floor,
  - 730 sqft basement , yr_built = 1948,
  - 1520 Sqft_living15, 6235 sqft_lot15
  - good grade and condition, no waterfront/view, basement
  - Location: lat 47.6950 longitude -122.304

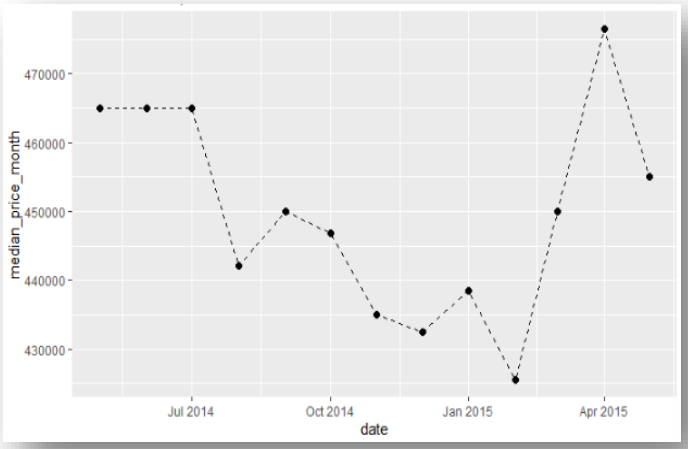
Data Explanation and/or Limitations

The KC house dataset includes homes sold between May 2014 and May 2015 in King County, Washington. Dataset contains information on 21 predictors along with 21,613 observations. There were no missing values in the dataset. Only 4.22% percent (914 total houses) of the house inventory has renovation data. There were 18 houses where number of bedroom is either zero or outside the scale such as 33 rooms, which we considered an outlier and decided to remove these from the dataset. Selected houses with 1 to 9 bedrooms only for modeling. Some of the features had imbalanced data i.e. yr_renovated, sqft_basement, waterfront and view.

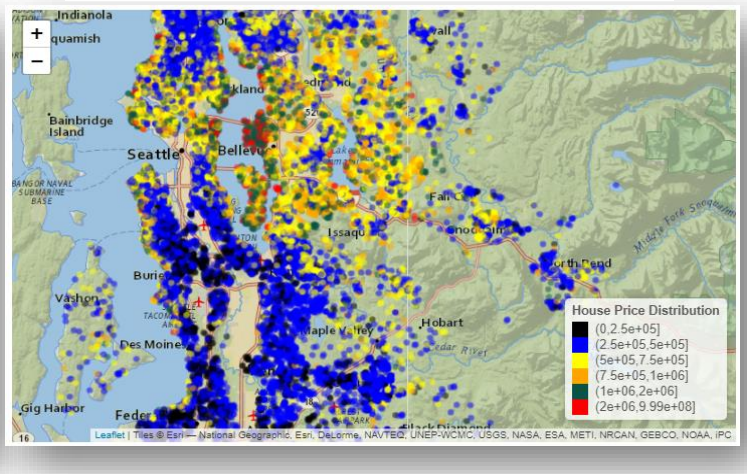
Method / Process / Analysis:

We applied exploratory data analysis techniques to get the information of median price per month over time and the visual below shows the changes of median price in King County over time. The chart is showing a dip in median price in February 2015. Using EDA and ESRI map, below is the visual of latitude and longitude of houses mapping its location to better visualize these houses with its corresponding price. Through visualization, home values on the priciest level appears to be mostly homes in Seattle and the Eastside displaying clusters in yellow to red colors. The most expensive house sold in the dataset is from Zipcode 98117 with a price tag of $7.7M. This house is in the neighborhood of Loyal Heights Ballard in Seattle.

Median Price per Month over Time

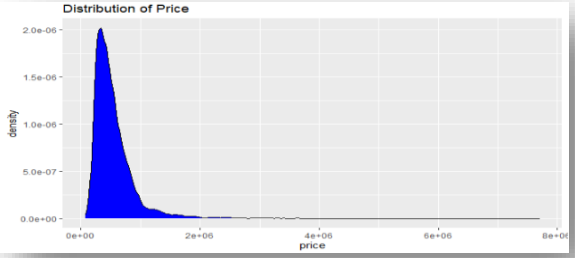


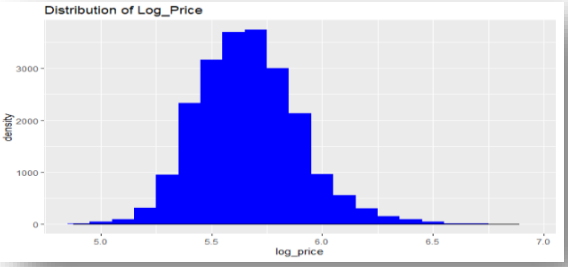
House Price Distribution in King County



We also applied a variety of transformations to the independent variables. The goal was to achieve somewhat normally distributed variables that could describe the variability in price.

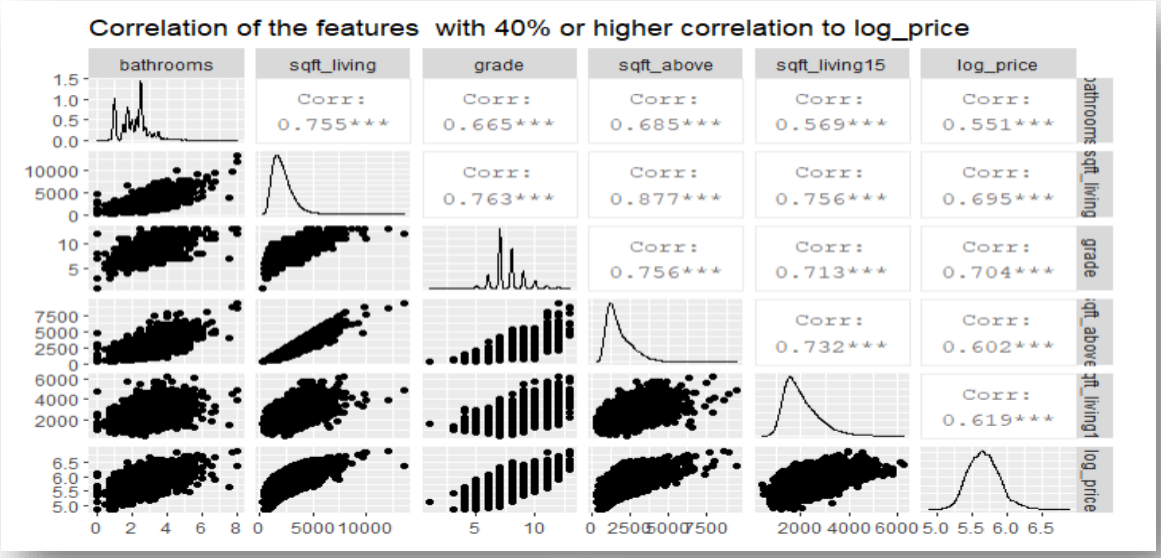
House prices of the dataset is right *skewed*, and "**log**() function" has been applied to the price variable. Using log_price will be more appropriate for linear regression application.

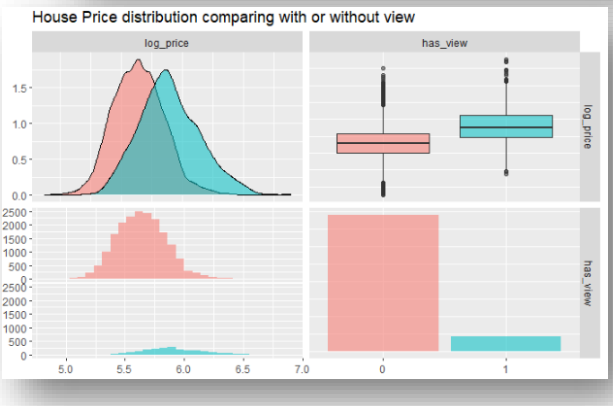




The ggpairs plot below is very informative for the modeling process. It is showing features with correlation values of equal or greater than 40%. Simultaneously, inferred that sqft_living is highly correlated to sqft_living15, sqft_above and grade. Knowing these correlations help us in the selection of features for machine learning algorithms.

Another ggpairs showing distribution of prices comparing houses with or without waterfront or view is helpful to understand having a house with waterfront or view will likely increase the price of your house.





To build machine learning algorithms, dataset has been split into train (80%) and test (20%) sets, applied pre-processing steps using center, scale and removing "nzv" method and divided into features and target (log price). Thus, the algorithms would be trained on one set of data and tested out on a different set of unseen data. We created 2 Linear Regression models using RMSE metrics and 1 Lasso model and applied cross-validation of 10 folds to all models. We tested all models and predicted on test data. "RMSE" and R-Squared" are the two metrics we used to assess the more effective model for predicting house price. Finally, we have compared the results of the models using resampling function and visualized the results using dotplot and bwplot (box plot) functions.

Below are the three (3) models with cross validation of 10-fold

Model_1: Linear regression using full features

Model_2: Linear regression with 5 features based on **varImp** result on model1: sqft_living, lat, grade, yr_built and view

Model_3: Lasso regression with 5 features based on **varImp** result on model1: sqft_living, lat, grade, yr_built and view

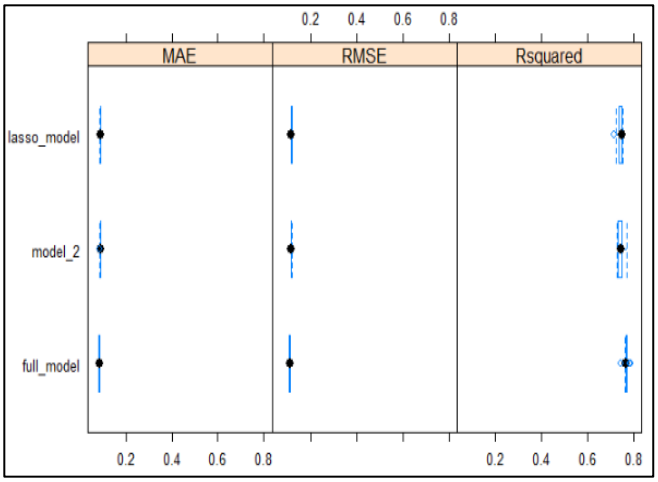
```
Call:
summary.resamples(object = results)

Models: full_model, model_2, lasso_model
Number of resamples: 10

MAE
                  Min.    1st Qu.     Median       Mean    3rd Qu.       Max. NA's
full_model  0.08283883 0.08365640 0.08450896 0.08504871 0.08640845 0.08794587    0
model_2     0.08624566 0.08892105 0.08940319 0.08931914 0.09004473 0.09170558    0
lasso_model 0.08798017 0.08908621 0.09018751 0.09019289 0.09137203 0.09225734    0

RMSE
                  Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
full_model  0.1068754 0.1086037 0.1098660 0.1098532 0.1111073 0.1127750    0
model_2     0.1130817 0.1145190 0.1155221 0.1155816 0.1165735 0.1193105    0
lasso_model 0.1141408 0.1147108 0.1169267 0.1166623 0.1178375 0.1195760    0

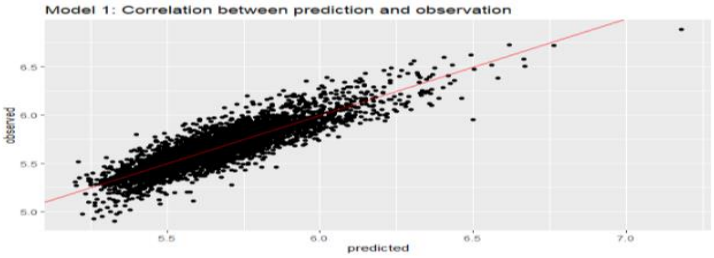
Rsquared
                  Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
full_model  0.7445725 0.7658142 0.7670931 0.7687908 0.7719394 0.7874224    0
model_2     0.7279286 0.7342796 0.7451091 0.7439997 0.7496447 0.7690085    0
lasso_model 0.7160197 0.7381353 0.7479969 0.7423449 0.7510802 0.7526574    0
```



Model Median Summary of RMSE and R-Squared

	RMSE	R-Squared
Full_Model	0.109866	0.7670931
Model_2	0.1155221	0.7451091
Lasso	0.1169267	0.7479969

Plot of Full_Model prediction result



	Price	Log_Price	Variance	
Actual	**$437,942.40**	5.64147	Percent	Dollar
Full Model	$ 430,532.60	5.63401	**2%**	$ 7,409.80
Model 2	$ 458,747.80	5.66157	-5%	$ (20,453.38)
Lasso	$ 452,700.50	5.65581	-3%	$ (15,459.22)

```
# Convert log price to actual price
```{r,warning=FALSE,message=FALSE}
test_house_price = 10^ 5.634006
test_house_price
```

[1] 430532.6
```

Conclusion:

Regression model fits the data well if the differences between the observations and the predicted values are small. After assessing numeric measures of goodness-of-fit, which is the R-squared, we also evaluated the residual plots. The RMSE of Full_model is lower than model_2 and lasso. Similarly, the R-squared of Full_model is higher than model_2 and lasso. We have 76% percent information to make an accurate prediction about the price of a house using this model.

We put all models into test by using the real data (*reference*: problem statement #2) and clearly, Model 1 outperformed both Model 2 and Lasso model. The answer to problem statement #2, based on the result of selected "best-fit-model" - full model, it will costs us $430,532.60 to purchase this property. After comparison and assessment, we conclude that full_model is more effective than model 2 and lasso model in predicting the price of houses in King County.

Attachment: **FinalProject-CharryDeandres-KanisSyeda-ShahnazJalali**.rmd