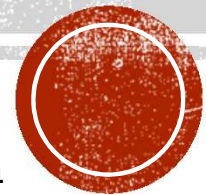


# HOUSE PRICE PREDICTION IN KING COUNTY, USA

Project Group 1

Yunfei Bai, Elena Melnichenko, Sateesh Puripanda, Abhishek Sinha



# FIRST LOOK

- 19 predictor variables to predict Housing Price, along with 21613 observations
  - 15% of data is kept aside to be used for validation. Rest 85% data has been divided into 70:30 to train and test the model.
- No missing values and duplicate records
- Some parameters clearly do not affect prediction, e.g. id, longitude and latitude
- Some parameters are likely to be cross correlated, e.g.
  - Sqft living/ Sqft living 15 (same with Sqft lot, evening, international)
- In the confounding sets sqft\_living and grade have high correlation with price than other variables.

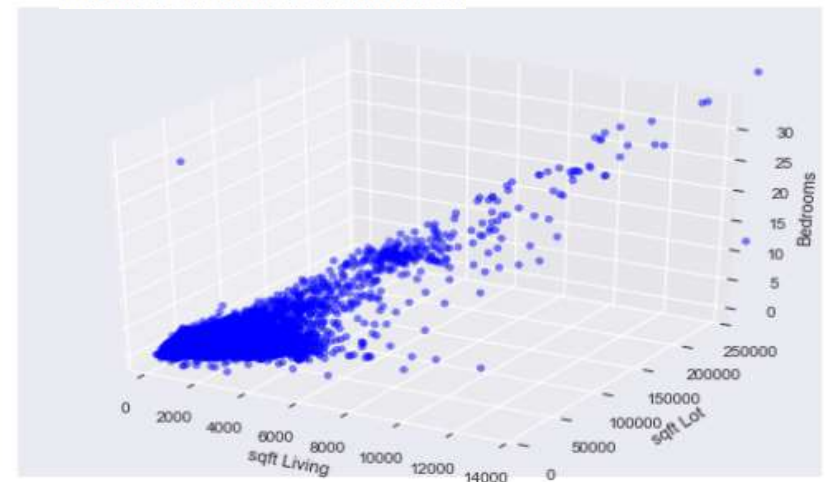
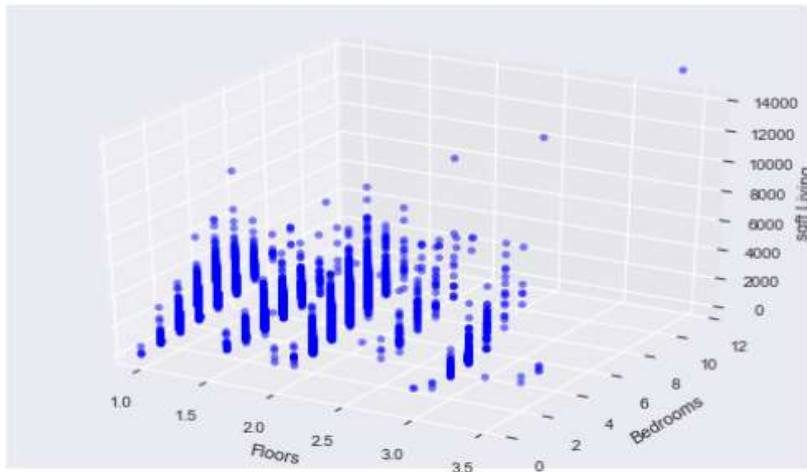


# EXPLORING DATA

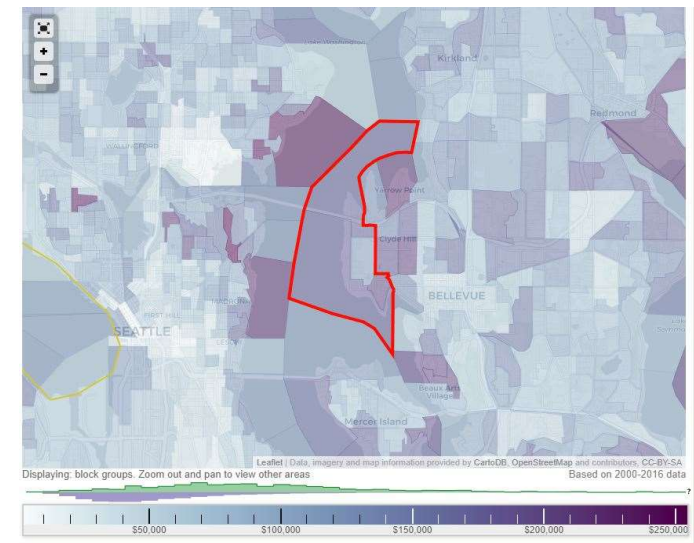
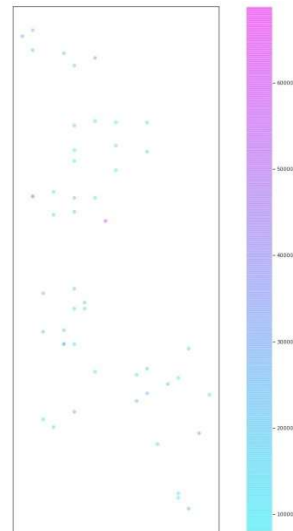
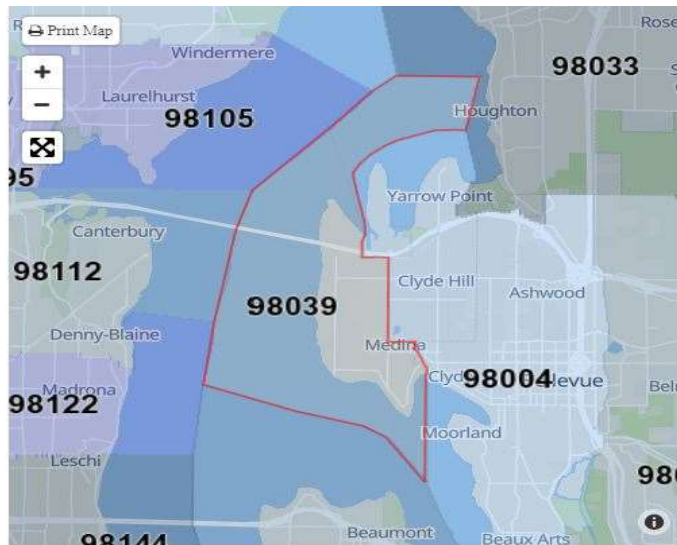
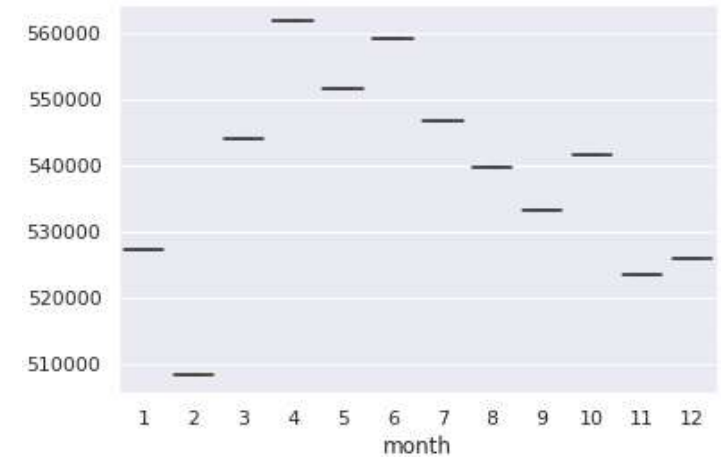
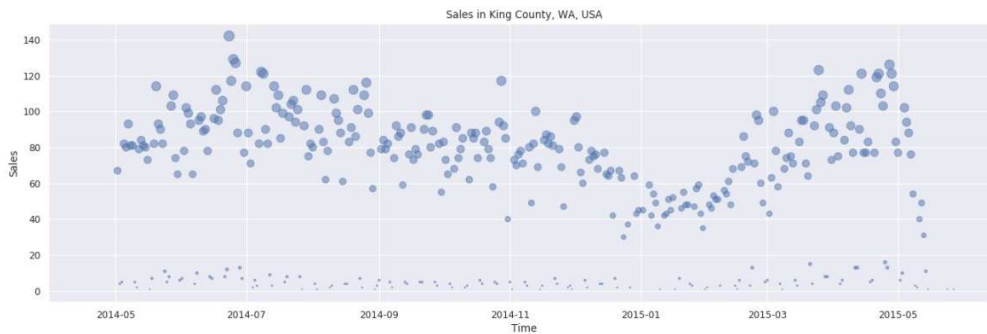
1. Missing values check
2. Converting categorical values to numerical
3. Excluding parameters that will not help for the model
4. Correlation between features

```
-----Check unique values-----
id          21436
date        372
price       4028
bedrooms    13
bathrooms   30
sqft_living 1038
sqft_lot    9782
floors       6
waterfront  2
view         5
condition   5
grade       12
sqft_above  946
sqft_basement 306
yr_built    116
yr_renovated 70
zipcode     70
lat         5034
long        752
sqft_living15 777
sqft_lot15  8689
index       21613
dtype: int64
```

```
-----Check Nulls-----
id          0
date        0
price       0
bedrooms    0
bathrooms   0
sqft_living 0
sqft_lot    0
floors       0
waterfront  0
view         0
condition   0
grade       0
sqft_above  0
sqft_basement 0
yr_built    0
yr_renovated 0
zipcode     0
lat         0
long        0
sqft_living15 0
sqft_lot15  0
dtype: int64
```



# TIME, LOCATION, MONEY

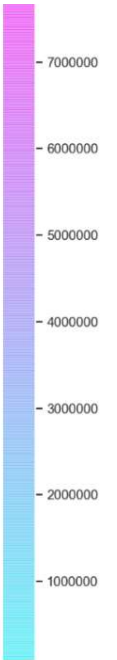
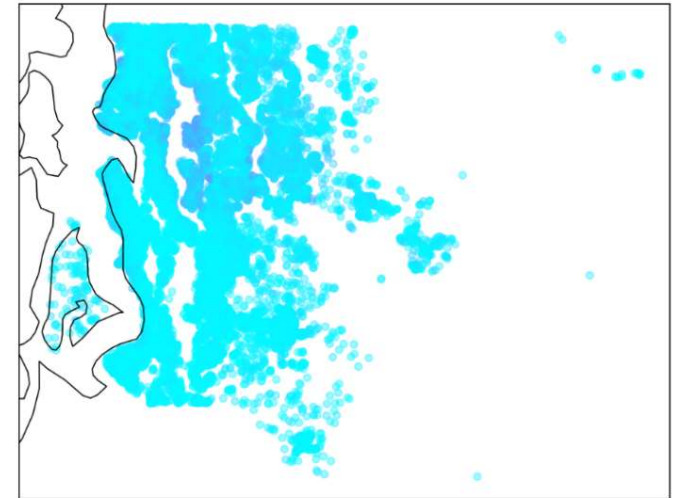
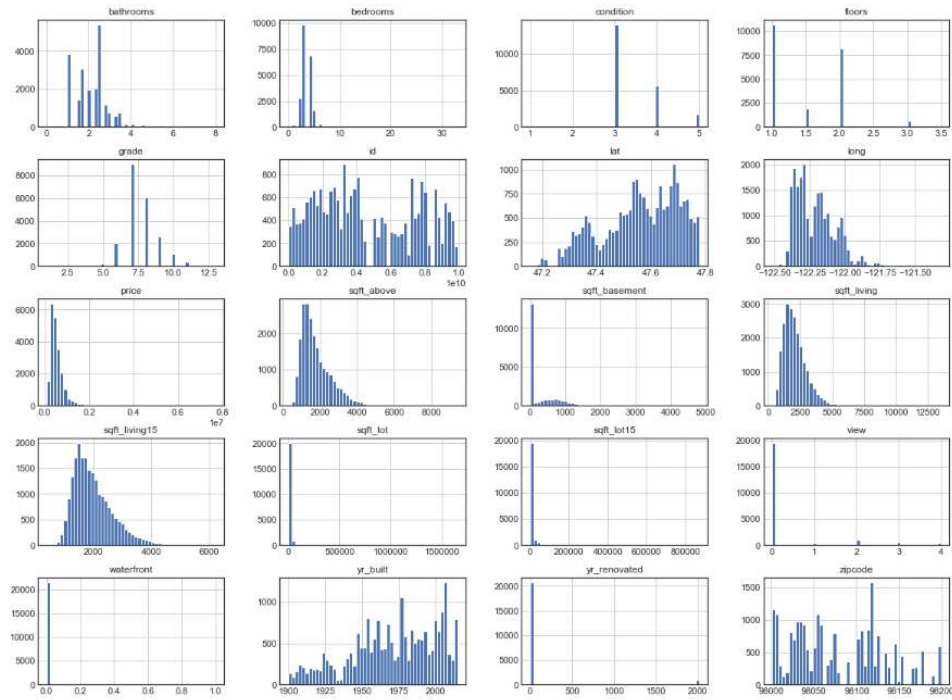


House sale prices for King County between May 2014 and May 2015  
<https://www.kaggle.com/harlfoxem/housesalesprediction>

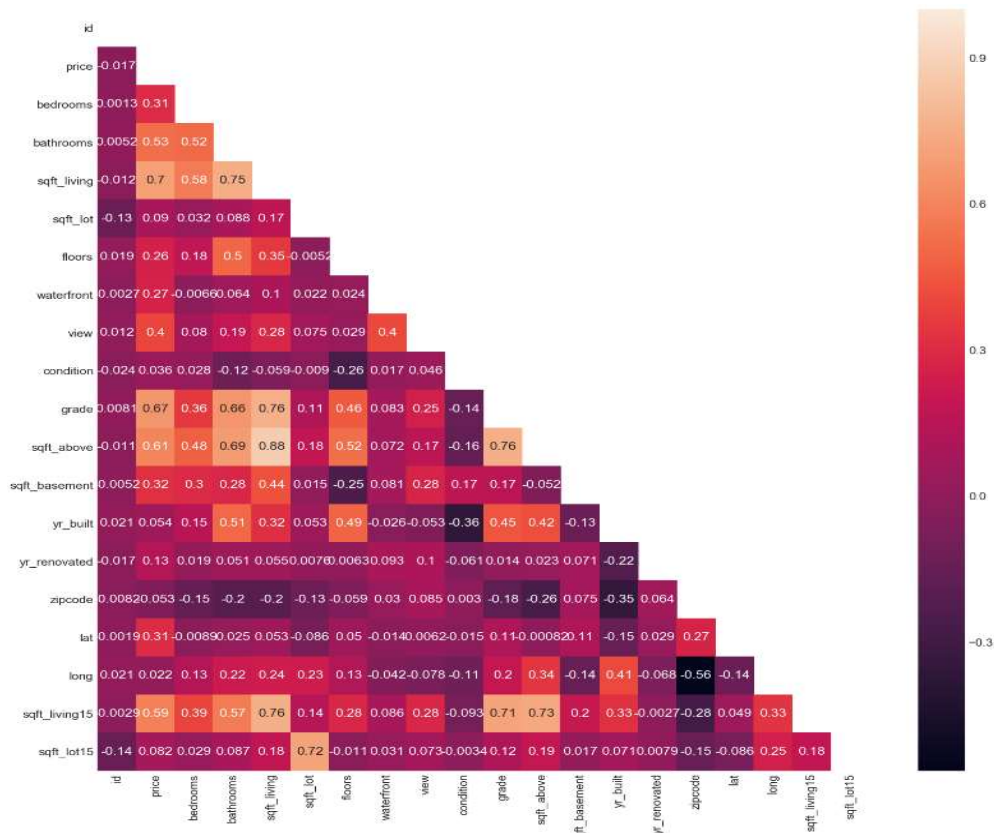


# DESCRIPTIVE ANALYSIS

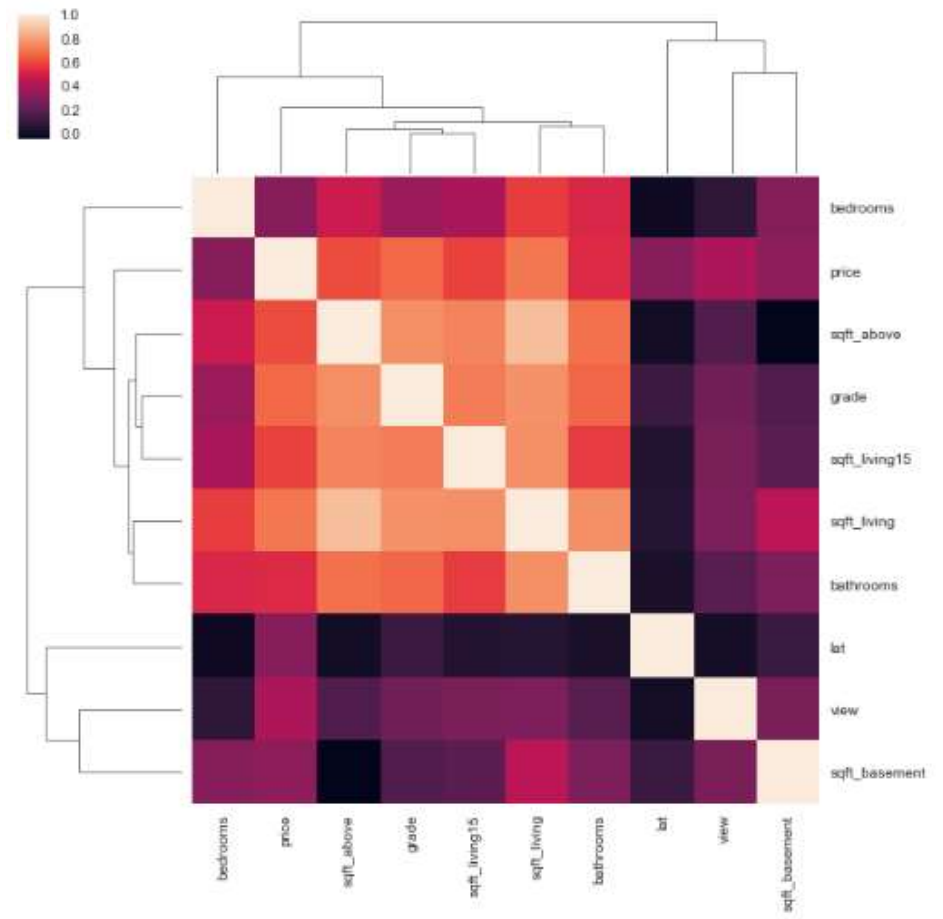
## ■ histogram



# DESCRIPTIVE ANALYSIS



Correlation heatmap



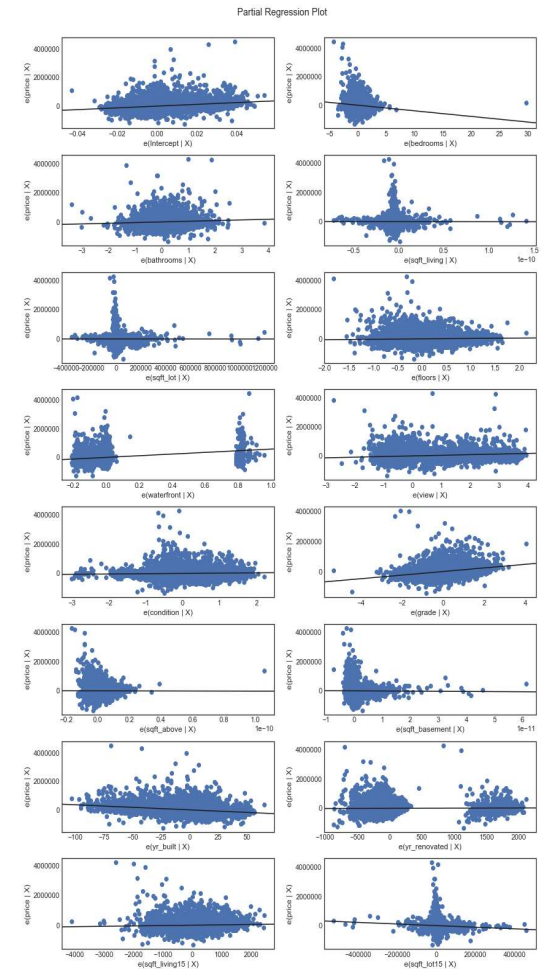
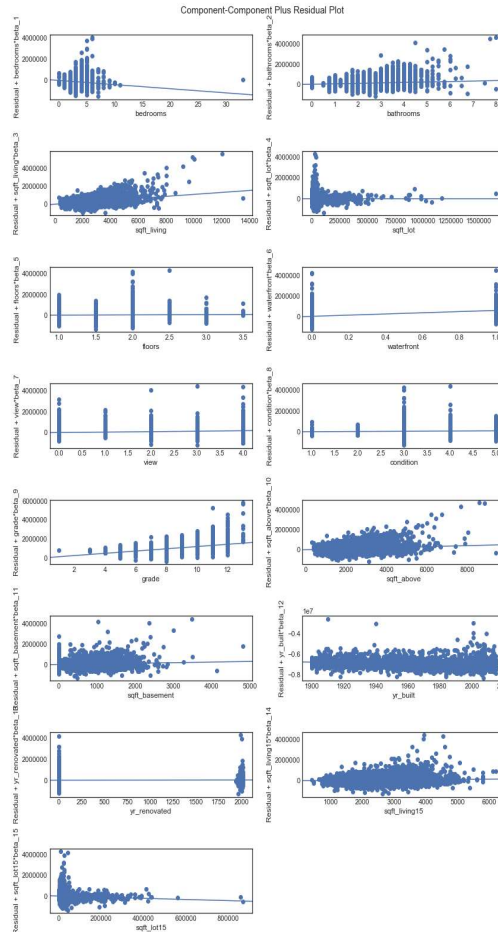
# FEATURE ENGINEERING

## ■ Numerical Variables:

- Date => days since 1900-01-01
- Bedrooms, bathrooms, sqft\_living, sqft\_lot, floors, waterfront, view, condition, grade, sqft\_above, sqft\_basement, yr\_built, yr\_renovated, sqft\_living15, sqft\_lot15

## ■ Categorical variables:

- zipcode : since we have 70 distinct zipcodes in this King county dataset, we categorized it into 70 dummy variables.



# PREDICTION MODEL

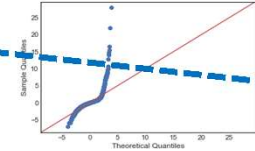
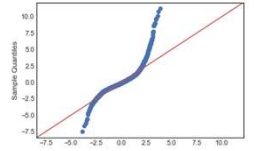
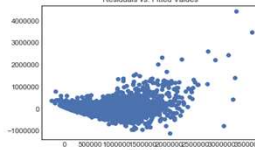
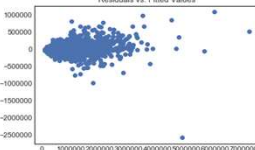
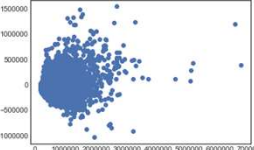
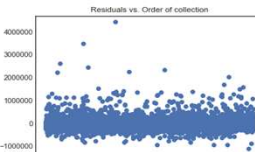
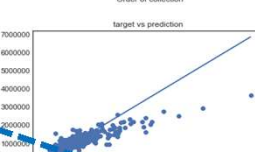
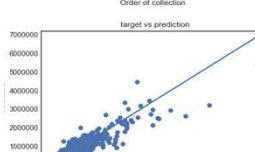
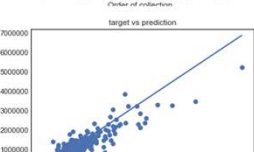
- First we feed features into multi-linear regression model
- Then we feed data into another two algorithm, Random Forest Regression Model and Gradient Boosting Regression Model.
- All three algorithm we used “backward elimination + cross validation” to get the best model.

A

Gradient Boosting Regression handed the outliers pretty well

B

Both Random Forest Regression and Gradient Boosting Regression give much better scores

	Linear Regression	Random Forest Regression	Gradient Boosting Regression
normality plot			
residue vs fitted value			
residue vs order of collection			
prediction vs target			
Baseline score	Adj R squared: 0.80 RMSE: 170731	Adj R squared: 0.84 RMSE: 152372	Adj R squared: 0.84 RMSE: 153156



**THANK YOU**

