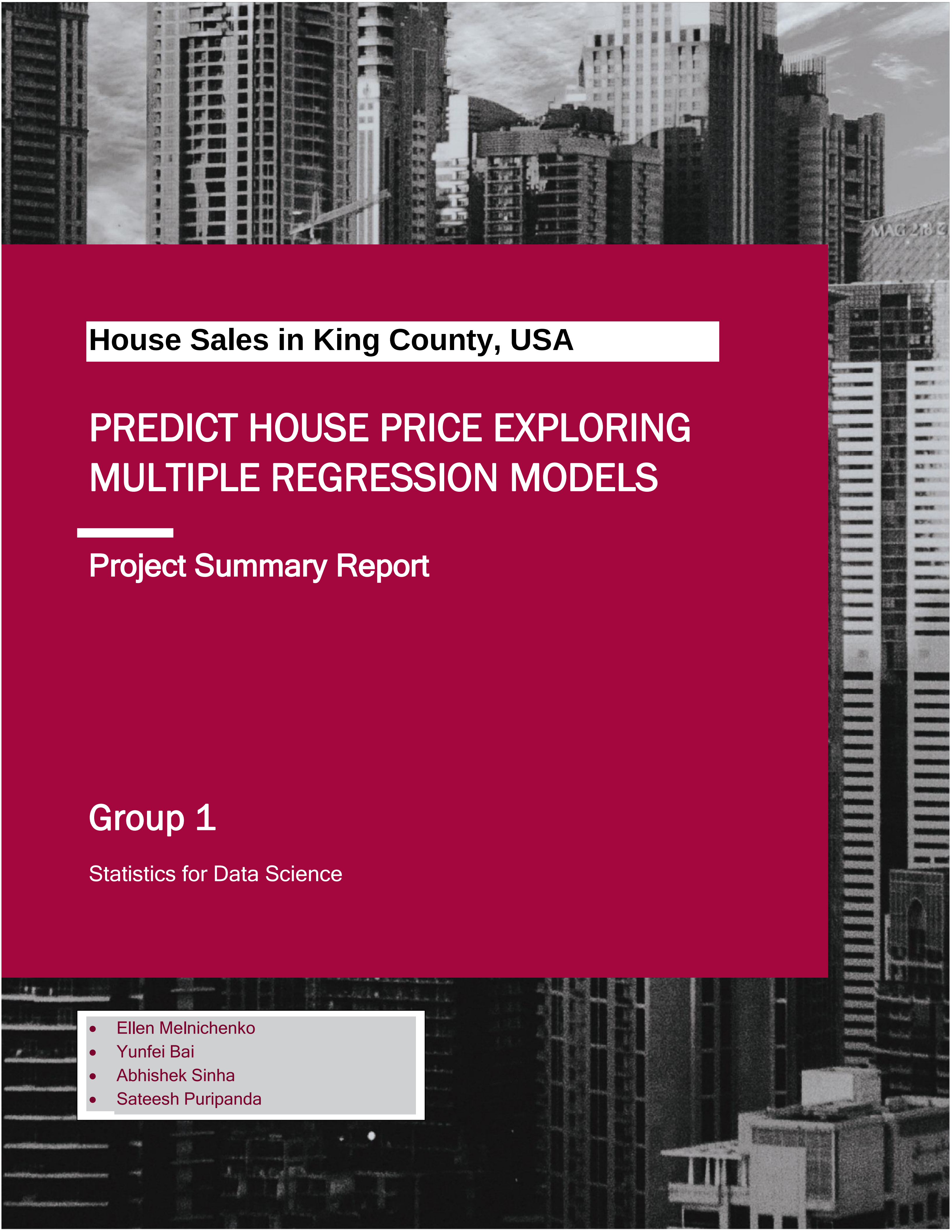


House Sales in King County, USA

PREDICT HOUSE PRICE EXPLORING MULTIPLE REGRESSION MODELS

Project Summary Report

Group 1

Statistics for Data Science

- Ellen Melnichenko
- Yunfei Bai
- Abhishek Sinha
- Sateesh Puripanda

Objective:

- Comparative Analysis of different Regression Models to predict housing prices in King County, USA.

Our aim is to study dataset and build regression models, to predict the house sale price, with predictors (X variables) and Housing Price (Y response variable), identify relationship between confounding predictors, select the features using backward elimination, calculate intercept, coefficients and R^2, valid the predictions with test data, compare the prediction models among algorithms of linear regression, Random Forest, Gradient Boosting Regression models.

Data Preparation

This dataset contains house sale prices for King County, which includes Seattle, and includes home prices between May 2014 and May 2015 (https://www.kaggle.com/harlfoxem/housesalesprediction).

Dataset actually is very clean. Duplicated ids, distinct values and nulls have been checked in the dataset. There are no null values in the whole dataset. Duplicated ids are found in “id” column which is reasonable, because each distinct id indicate one property and the property can resale during this period of time.

Exploratory Data Analysis:

Some insights are coming from temporal nature of the dataset. For instance, sales (number of houses sold) display the following behavior (the area of dot is proportional to number or sold houses):

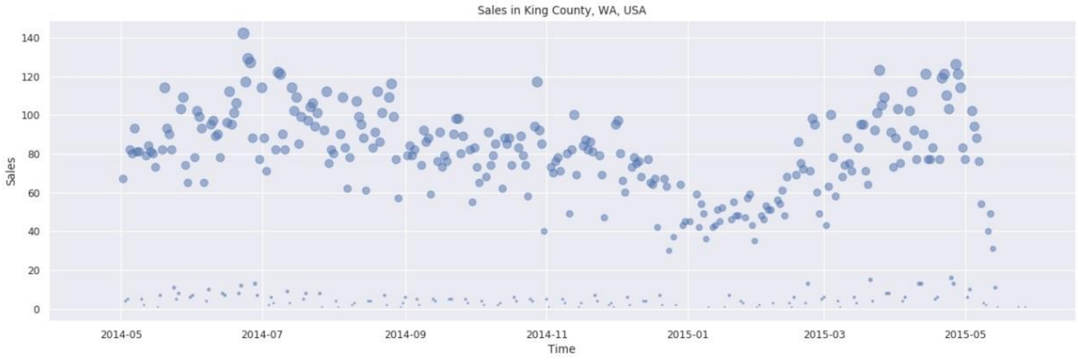


Fig.1 Number of houses sold in King County vs Date.

Real estate business shows seasonality: activity is at minimum at the beginning of the year, bursting up in May, July and August. Similar seasonality observed for average over month prices: the lowest average price corresponds with lowest sales in February, while two peaks in April and June correspond with maximum sales, indicating high market.

Fig.2 Average monthly house price in King County.

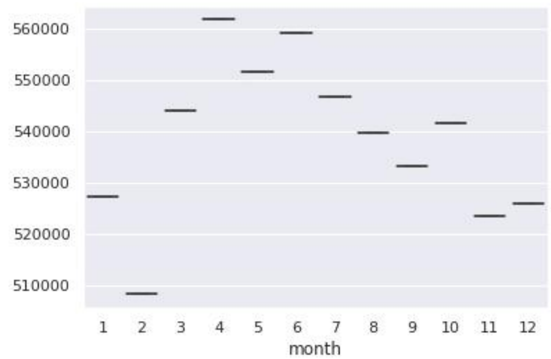
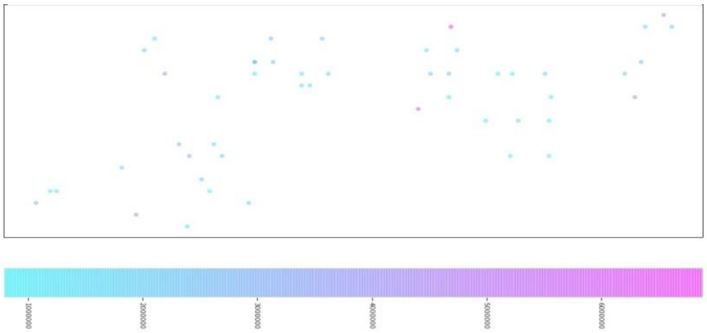


Fig.3 Geo distribution of the houses sold in King County 98039 zip code area



Some interesting conclusions might be drawn from max prices data: two of the most expensive houses are located in “98039” zip code area. Unlike “98039” with 49 sales, the highest number of sold houses related to “98103” zip code area with 599 sales over a year. Does it witnessing the most expensive neighborhood in the city? Let’s check map representation (Python Cartopy library) of data for the neighborhood above (Fig.3). In fact, better indicator for high end neighborhood is more likely to be an average income.

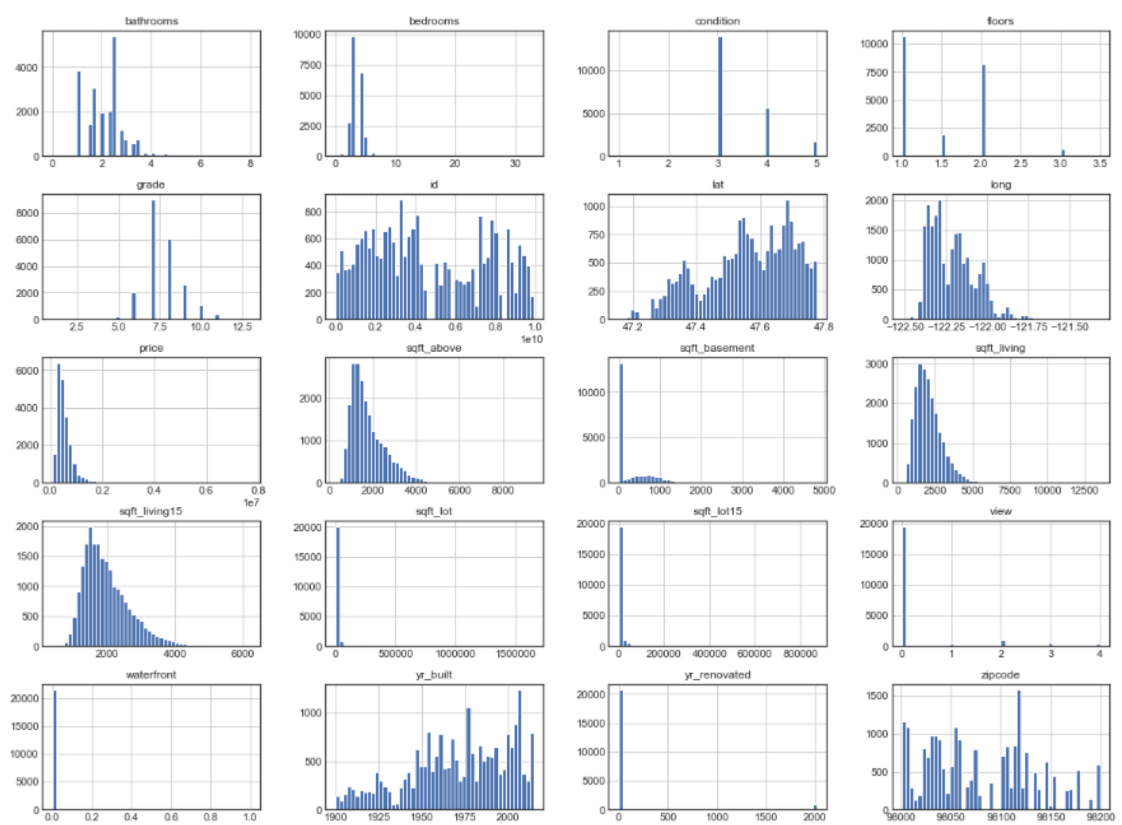
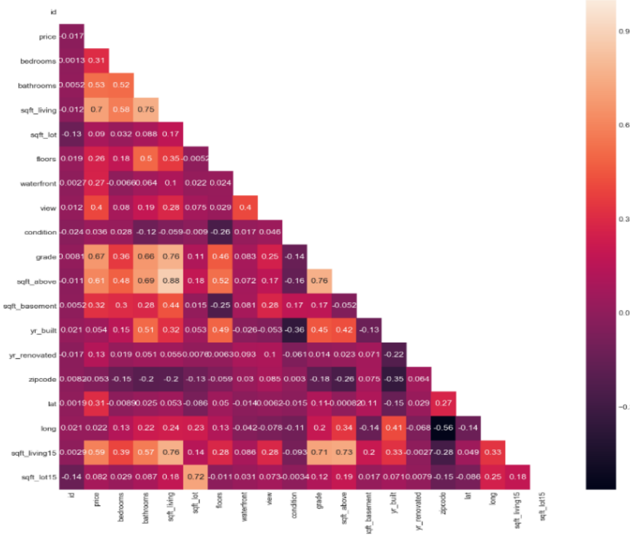


Fig.4 Histogram of variables in dataset

In descriptive analysis, we further plot histogram of most variables in the dataset in Fig 4. Most continues variables like sqft_living, lot, have decent distribution. And only some discrete variables like waterfront or yr_renovated have extreme skew distribution. Future we plot out correlation heat map for variables to check relationship between given two variables from dataset. In the Fig.5, some variables like sqft_living, grade, sqft_living15, sqft_above are highly correlated to price, our target variable. And sqft_living and sqft_above are highly correlated with each other, which suggests we may consider to take one out in the modeling step.

Use longitude and latitude in our dataset and cartopy library in python, we also plot the price range on the King county map in Fig.6. The dark blue and pink dots indicate high price area whereas the light blue ones indicate areas of lower house price. Compared with google maps, we can see the high price areas are mainly districts in Seattle and Bellevue, and lower price areas concentrate in Kent.



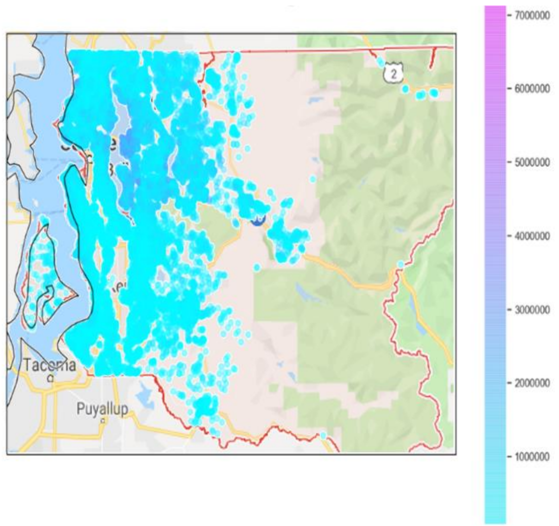


Fig.5 Correlation heatmap

Fig.6 Price range on the map of King County

Feature Engineering before feed into models:

Most variables by now are numerical and ready to be fed into models except for two. Date is in string format, so we engineer it to be numerical by transforming it into number of days since a reference date "1900-01-01" and then can be convenient to put into any models. Another column is zipcode. In King county dataset, we have 70 distinct values for it and because we don't want to be bias on any particular zipcode, so we created 70 dummy variables, each for a zipcode in the dataset. We also drop columns of "id", "lat" and "long", which are obviously irrelevant to our target price. After this, now we have a ready dataset to perform predicting processes.

Regression Models:

Before any modeling process, we first split 15% of the data as validation set. Then the rest of the dataset we split 70% for training and 30% for testing/cross validation.

Before we fit into our first model, a partial residual plot and partial regression plot are draw to check the linearity of each independent variable to target variable and notice any outliners that need to be concern. In Fig. 7, even we notice there are a little amount of outliners that might affecting the overall linearity, the most of our independent variables more or less follow some linear relationship with target variable, price.

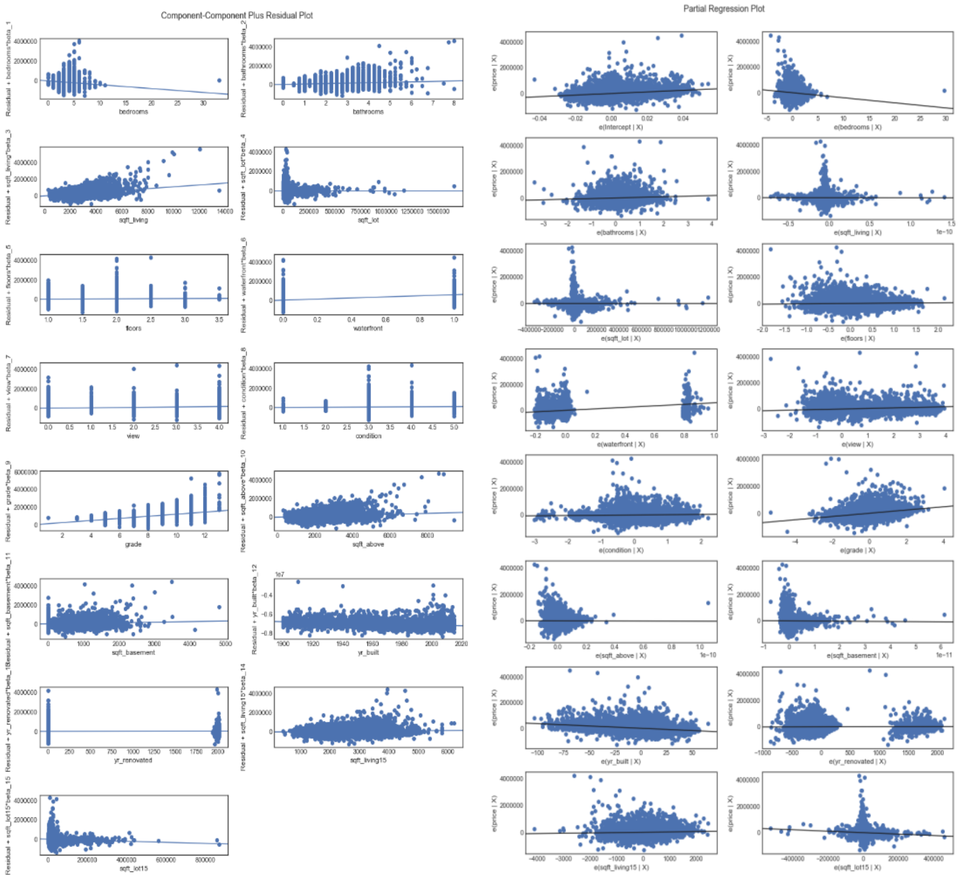
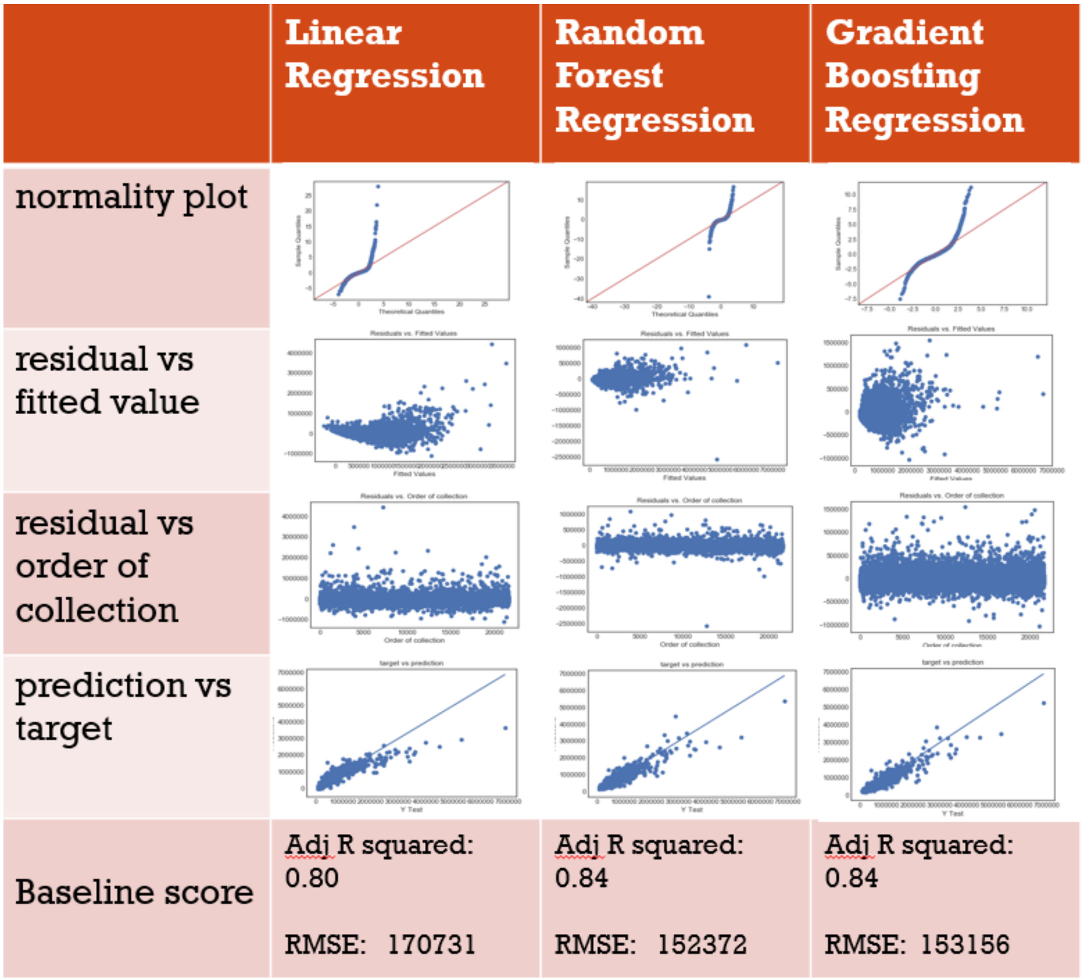


Fig.7 Partial residual and partial regression plot

We fit the engineered data into three algorithms, multiple linear regression, random forest regression and gradient boosting regression. Backward elimination approach and cross validation are combined together to get us best models by reducing features, on all three regression models. Then best models are used to predict the validation dataset (the 15%). Surprisingly, there is almost no change or even worse in scores in terms of R square and RMSE, compared with baseline models without any feature reduction. With this fact, we will use the baseline models instead of those best models getting from backward elimination approach.

Of all three models, we compared prediction results in multiple aspects, as shown in Fig. 8. Out from these three regression approach, linear regression has worst score. Also noticed in the normality plot and residual vs. fitted value plot, multiple linear regression model cannot handle outliners very well. Second, both random forest regression and gradient boosting regression provide us very good score, with adjusted R squared of 0.84. The RMSE of these two models also much smaller than the multiple linear regression, which indicate both random forest regression and gradient boosting regression can predict price quite well from our dataset. Third, even both random forest and gradient boosting can predict price well, the gradient boosting can handle the outliners much better, according its normality plot and residual vs. order of collection plot.



	Linear Regression	Random Forest Regression	Gradient Boosting Regression
normality plot			
residual vs fitted value			
residual 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

Sample Quantiles
Theoretical Quantiles
Residuals vs. Fitted Values
Fitted Values
Residuals vs. Order of collection
Order of collection
target vs prediction
Y Test

Fig.8 Comparison between different regression models.

Conclusion:

A very detailed time series analysis and descriptive analysis for house price in King County was performed. Also after carefully feature engineering, we did a thoroughly model exploration among three regression models: multiple linear regression, random forest regression and gradient boosting regression, with backward elimination approach plus cross validation to give us best models. Compared among three algorithms, both gradient boosting regression and random forest regression give us very good prediction and gradient boosting algorithm can handle outliners better then random forest. The multiple linear regression still give us a fairly good score, with adjusted R squared 0.8, however, still less than those two models and also can’t handle outliners very well.