CSE 519 - Data Science Fundamentals
HW3 - More on Kaggle Challenge

Charuta Pethe - 111424850
Deven Shah - 111482331
Neha Mane - 111491083

Question 1 - Desirability Scoring Function

In order to rank the houses based on desirability, we have considered the following variables in our scoring function:

- calculatedfinishedsquarefeet
- bedroomcnt
- regionidcounty
- taxvaluedollarcnt

Of these, calculatedfinishedsquarefeet and bedroomcnt are used directly in the scoring function. On the other hand, as regionidcounty is a categorical variable, we assigned a rank to each regionidcounty using taxvaluedollarcnt. We calculated the mean tax value for each county. Higher the mean tax value, higher the rank of the county.

Before scoring the houses, we went through the descriptions of each type of property, and saw that they were not comparable to each other. For instance, we found that an office complex with zero rooms but small area is likely to be costlier than a residential apartment with many rooms and relatively larger area. Hence, we calculated the scores only for Single Family Residential houses, as they account for most number of houses in the entire dataset (~ 2,200,000 houses). We also removed the outliers before scoring all the houses.

The individual ranges of calculatedfinishedsquarefeet and bedroomcnt are extremely different. If these variables were used in raw form, the score would be distorted in favor of the variable with larger values. Therefore, in the scoring function, we have used the Z-scores of these variables to normalize the data, so that each feature plays an equally significant role in the calculation.

Question 2 - Pairwise Distance Function

We have defined the function distanceMetric which takes the records of two houses, calculates the pairwise distance between them and returns it. We calculate the pairwise distance as the summation of:

1. Difference in Z-scores of bedroomcnt of two houses
2. Difference in Z-scores of calculatedfinishedsquarefeet of two houses
3. Difference in rank of the counties(regionidcounty) of the two houses

Further, we add 1 to the distance if the zipcodes of the two houses are different.

Geographic features used:

- regionidcounty
- regionidzip

Property features used:

- calculatedfinishedsquarefeet
- bedroomcnt

Question 3 - Clustering using Pairwise Distance Function

We ran the clustering algorithm on incrementally increasing subsets of the entire dataset. The limiting number of records on which the algorithm gave an output in a reasonable amount of time was 6000. Hence, we decided to run the algorithm on a subset of data which contained diverse houses. Therefore, we selected 1 record after every 350 records for clustering.

We experimented with different numbers of clusters as parameters, and we found that for larger number of clusters, the clustering was taking long time. However, for 50 clusters, the clustering was done in a reasonable amount of time.

Also, for all the experimental values of the number of clusters, we observed closely similar outputs. This is because our pairwise distance function calculates the distance based on the difference in the attributes mentioned above, and houses that are more dissimilar to each other are segregated into different clusters. Moreover, we observed that even though latitude and longitude have not been used as geographical parameters in the distance function, the clusters reflect that similar houses are also located geographically close to each other.

Question 4 - Merging Properties with New Dataset

Our external dataset includes crime statistics of every city in the state of California. It includes the population and counts of various crimes over the year 2015 in the state of California. This dataset is taken from the Kaggle website, and is provided by the Federal Bureau of Investigation (FBI).

The idea behind choosing this dataset was to include crime statistics as a negative component for the desirability of each house. In order to merge this external dataset, we did the following pre-processing:

1. We obtained a list of unique regionidzip values, and further created a custom list of records in the data. This list contained 1 record per unique regionidzip.
2. Using the latitude and longitude of this subset of the dataset, we found the actual postal code, city and county with the help of Google Geocoder.
3. We then created 3 dictionaries with regionidzip as the key, and postal code, city and county as values.
4. Using these dictionaries, we updated the original dataset to show postal code, city and county of each house.
5. For those houses which did not have regionidzip, we determined the postal code, city and county using the latitude and longitude.
6. The original dataset was left joined with the crime dataset on the column city.

We implemented Linear Regression to predict the log error using the updated dataset. The features chosen were: calculatedfinishedsquarefeet, bedroomcnt and crime_count. The dataset is split into train and test set in the ratio 95:5. The mean squared error and R2 score came out to be 0.0338089150366 and 0.00107681417878 respectively.

As compared to our previous model, which did not include the external dataset, this model did not turn out to be an improvement. The old model gave us a mean squared error of 0.0169875170064 and R2 score of 0.00312482685843, which was better than the score after merging the properties with the external dataset.

Question 5 - Prediction model with merged dataset

We implemented linear regression and lasso regression to predict the logerror. The features that we have used for our prediction model are calculatedfinishedsquarefeet, bedroomcnt and crime_count. The reasons for choosing these features are:

1. Calculatedfinishedsquarefeet - As observed from previous questions and assignment, calculatedfinishedsquarefeet proved to be one of the most correlated feature with the logerror and it seems logical to have better and costlier houses with increasing calculatedfinishedsquarefeet.
2. Bedroomcnt - This feature also had high correlation with logerror.
3. Crime_count - Using the merged dataset from question 4, we have used the crime count within each city to see if the overall pricing deteriorates by increase in crime count and how that affects the logerror prediction.

Furthermore, we have precomputed the missing values as:

1. Calculatedfinishedsquarefeet - We choose to fill the missing values with mean as the data is fairly evenly distributed, and replacing the missing values with the mean of the column means keeping the distribution the same without distorting it much.
2. Bedroomcnt - We chose to fill the missing values with the mode as the range of the values was very short for bedroomcnt and mean would have given a decimal value when the bedroomcnt is supposed to be a natural number.
3. Crime_count - In case of crime_count, we observed that the crime_count was in a similar range for most cities, only the city of Los Angeles has a very high crime count. So instead of using the mean and distorting the distribution we chose to use the median.

In order to evaluate the prediction model we split the dataset into train and test set in the ratio 95:5. Following gives an insight into the results obtained from linear and lasso regression.

Linear regression:

After using linear regression to predict the logerror values we observe that mean error and r2 values as follows:

1. mean squared error = 0.0338089150366
2. r2 score = 0.00107681417878

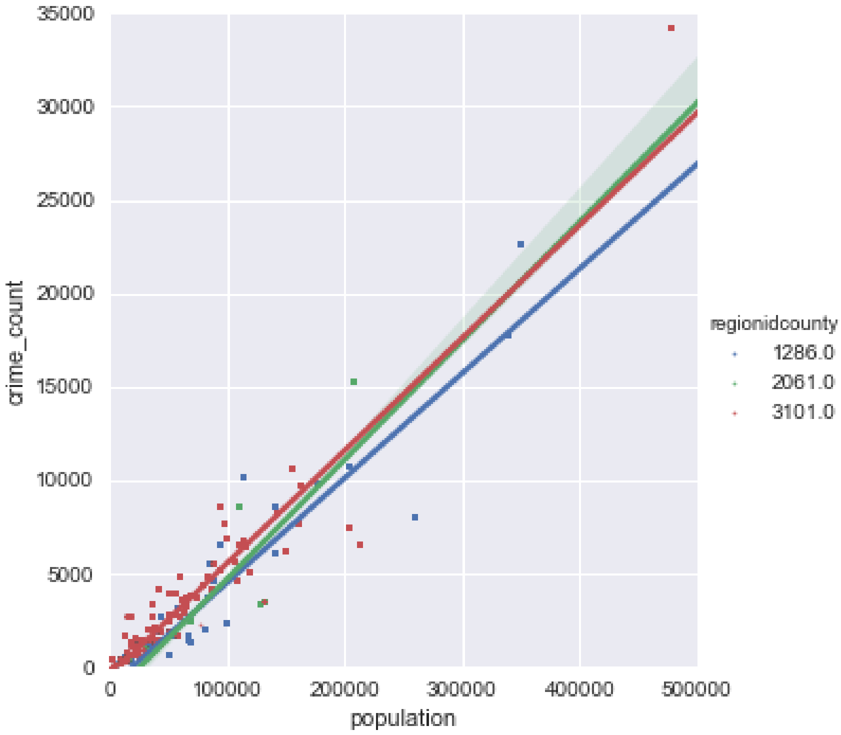
Lasso regression:

1. mean squared error = 0.027
2. r2 score = -0.00089

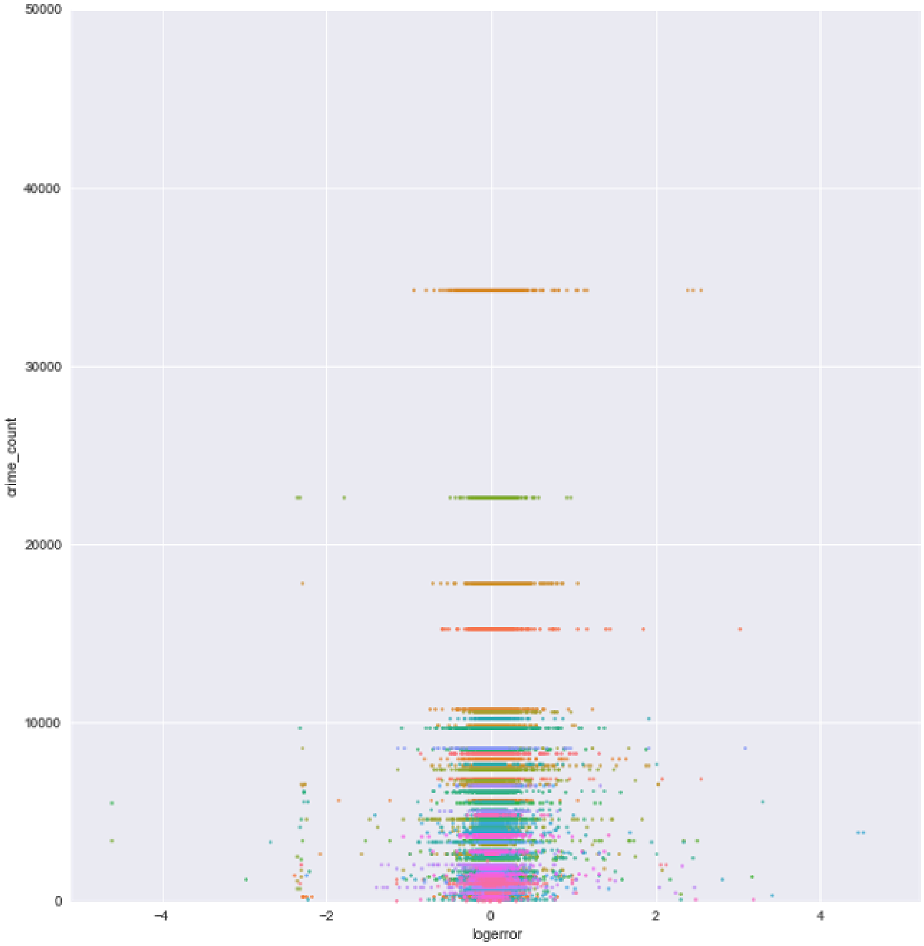
This tells us how linear regression performs better over lasso regression. Even though lasso has better mean squared error, it has a negative r2 score and this gives linear regression an edge over it. Hence, we have used linear regression as our final model to predict the logerror for the submission file and submit it to Kaggle.

Interesting Experiences:

1. While using the crime_count and population from the merged dataset, we plotted the data and observed that they had a high correlation of 0.99. Hence, it seemed redundant to use both of these features in our prediction model. So, we chose to use crime_count and drop population values.



2. Secondly, we found that as the crime_count increases the log error tends to become closer to zero, i.e. its variance becomes less as shown in the graph below.



3. While implementing Lasso regression, we found that the coefficient value of bedroomcnt came out to be 0.

4. We had expected our model to perform better after merging the crime dataset with the properties dataset and using it to train the model. However, this did not happen possibly because crime count does not influence the buying preference of people as much as other factors do. For eg. People prefer to buy houses in Los Angeles in spite of its higher crime count.

Question 6 - Permutation test on p-value

Based on the predictions calculated in Question 5, the mean squared error is taken as a benchmark to compare it with permutations of logerror of the testing dataset.
The permutation of the log error is done using random shuffling.
The following results were obtained on permuting the log error of the testing dataset

Number of Permutations	p-value
100	0.000
500	0.006
1000	0.009

These tests show that our model performs very well.

Question 7 - Zillow submission on Kaggle

The best rank we could achieve was 1998. However, now our rank became 2452 on the leaderboard.