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, our scoring function has taken calculatedfinishedsquarefeet, bedroomcnt and regionidcounty as the variables.
Here, calculatedfinishedsquarefeet and bedroomcnt correlate well with the logerror Other than these two property variables, we have taken regionidcounty to rank the houses based on their county. Here our ranks are based on mean taxvalue of the county - higher the taxvalue, higher the rank.

In our scoring, we have removed the outlier houses and calculated the score for Single Family Residential houses as they account for most number of houses in the entire dataset. (Approximately 2,200,000 houses)

We have used the zscores of these variables to normalize the data so that each feature had a significantly equal role to play, as individual ranges differed a lot for calculatedfinishedsquarefeet and bedroomcnt.

Question 2 - Pairwise Distance Function

The function distanceMetric takes input the zscores of the features of the two houses as input, and calculates and returns the distance between them. The geographic feature used is regionidcounty and property features used are calculatedfinishedsquarefeet and bedroomcnt. The distance is calculated as the sum of difference in zscores of calculatedfinishedsquarefeet, bedroomcnt and regionidcounty features of the houses. If the counties differ geographically, our distance function separates them further apart in the map plot.

Question 3 - Clusters of Distance Function

For clustering, a sample of the data is used, which includes 6000 records. The sampling is done such that after every 350 records a record is chosen. These 6000 records are grouped into 50 clusters. The map is plotted with x axis as latitude and y axis as longitude of the house.
It can be 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 which is provided by Federal Bureau of Investigation (FBI).

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

1. As the postal code provided by Zillow maps one-to-one to the original postal code, records with unique regionidzip codes were chosen.
2. Using Google geocoder, the cities were fetched using the latitude and longitude from the above sample of records.
3. A dictionary was created to map regionidzip to corresponding city.
4. As there were some records that didn't have regionidzip value, latitude and longitude were used to fetch the corresponding city.
5. The entire properties dataset was left joined with the dictionary containing mapping of cities.
6. The newly formed properties dataset was further left joined with the crime statistics dataset using city as the key attribute.

Question 5 - Prediction Model with merged dataset

Linear Regression is done on the merged properties dataset. The features chosen are: calculatedfinishedsquarefeet, bedroomcnt, crime_count. The dataset is split into train and test set in the ratio 95:5. The squared mean error and r2 score came out to be 0.0338089150366 and 0.00107681417878 respectively.

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 future 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

Question 7 - Zillow submission on Kaggle

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