CS506 Final Project: PlaceMe & Housing in Great Boston

For this final report, we found three topics that connect with Boston housing. In order to relate the information that can help PlaceMe Living, we provide detailed data and analysis associated with each topic.

Part1: Housing price V.S. Housing Market.

Background:

Why:

The first part is about the housing market of Boston. We found 2 useful rates to describe the house sources in Boston. Through availability and vacancy rate, we hope that we can understand which areas of Boston are easier to rent and which areas are worth signing a contract with the local owner. At the same time, house prices in Boston are relatively high compared to other cities. If the company signs a contract with someone but does not rent the house out, it will definitely cause big losses to the company. Therefore, we want to dig into the relationship between vacancy rate and availability rate in order to gain a deeper understanding of the company's focus on the regions and the impact of the two rates toward the prices.

How:

We collect data from Boston pads. From the data we collected, find out the relationship between Real Time Availability Rate, Real Time Vacancy Rate and Housing Prices.

Experiment:

Real Time Availability Rate：(Total Apartments Currently Vacant (Not in the market) + Apartments Set to Become Available on a Later Day) / total apartment in database

Real Time Vacancy Rate：(Total apartments currently available to rent(In the market) + all of those becoming available in the future) / total number of apartments

Correlation between 1b1b price and Real Time Availability Rate: 0.1065869552
Correlation between 1b1b price and Real Time Vacancy Rate: -0.4392974535

Explanation:

Correlation Coefficient Explanation:

Related Level:	Range:
Low	-0.1 < R < 0.1
Middle	-0.6 < R < -0.1 or 0.1 < R < 0.6
High	-1 < R < -0.6 or 0.6 < R <1

By the definition above, we can see the Correlation between 1b1b price and Real Time Vacancy Rate: -0.4392974535. It's a middle rate. The negative means that when the vacancy rate decreases, the price of the house will increase. However, we can see that the correlation of 1b1b price and Real Time Availability Rate is just 0.1065869552 which is low related (This is happening because the availability rate is not related to houses in the market now).

What Can PlaceMe Do:

Maybe we can rent the house in the area with a high current Vacancy Rate + low current Availability Rate, because the rent in this area may increase in the future.

What to Improve In the Future:

1: Try correlation analysis on other data.
2: Try other ways to find relationships(e.g:Spearman Correlation Coefficient)

Part2: Vocab Frequency Analysis to Reviews.

Background:

Why:

Our second task is to do a sentiment analysis on customer reviews of PlaceMe and their competitors: Bedly and Bungalow). Thus, we plan to analyze the advantages of PlaceMe and the competitors from the reviews and want PlaceMe to keep their advantage and learn from other companies, and then we will analyze the negative reviews so that we hope they can make some corrections.

How:

We collected tenant reviews of PlaceMe and its competitors, Bedly and Bungalow, from Google, Facebook and Yelp. We then learned that PlaceMe has the highest star rating and we decided to analyze the advantages of PlaceMe through vocabulary analysis in the review. Because there are few PlaceMe negative reviews, we then focus on collecting and reviewing negative reviews in its competitors, hoping that Placeme can be vigilant to these contents.

Note*: We want to do a sentiment analysis through building a model originally. However, since there are too few data entries and most of the reviews are positive. So the data is not balanced and we decided to do a vocab frequency analysis to see what PlaceMe can improve.

Experiment:

We wrote a python program to transfer the reviews we collected into vocabs and their frequencies. For all the vocabs, we clean the punctuations and lowercase all the characters.

Explanation:

The vocabulary above are some meaningful vocabs which have high frequency in the collected datas. So we believe that they are useful for identifying what behavior is liked/hated by customers.

PlaceMe Positive Reviews:

PlaceMe Review
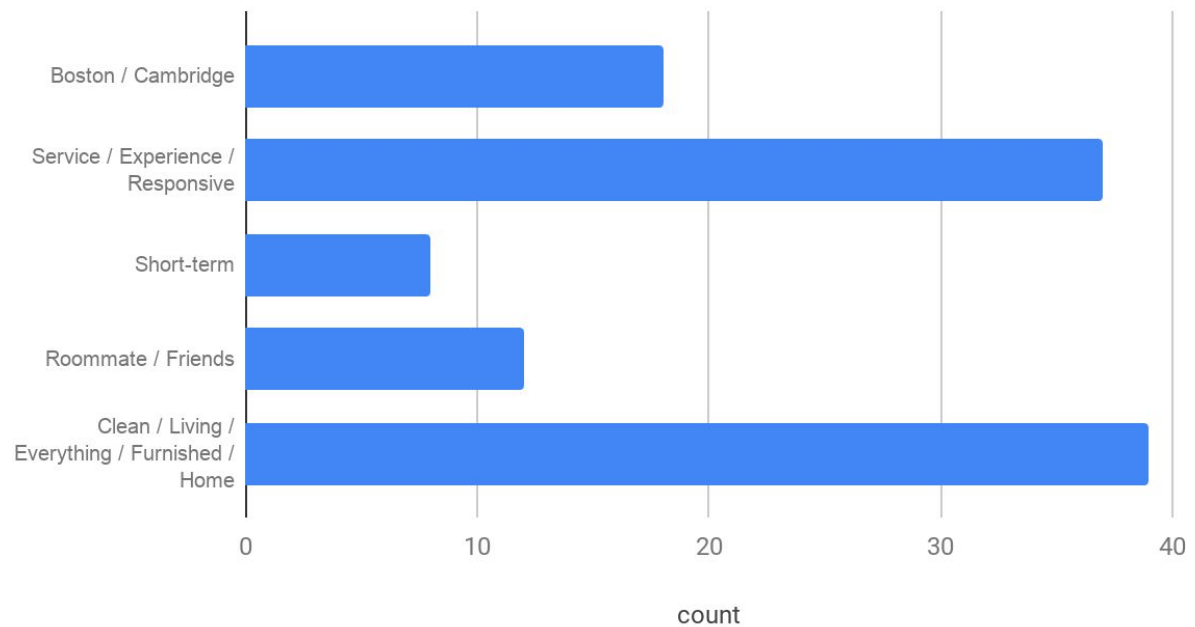
There are 43 positive reviews we found for PlaceMe. From all these reviews, here are some vocabs with high frequency worth notifying.

"Boston/Cambridge": 18 -- PlaceMe have their major services in Boston, especially in Cambridge.

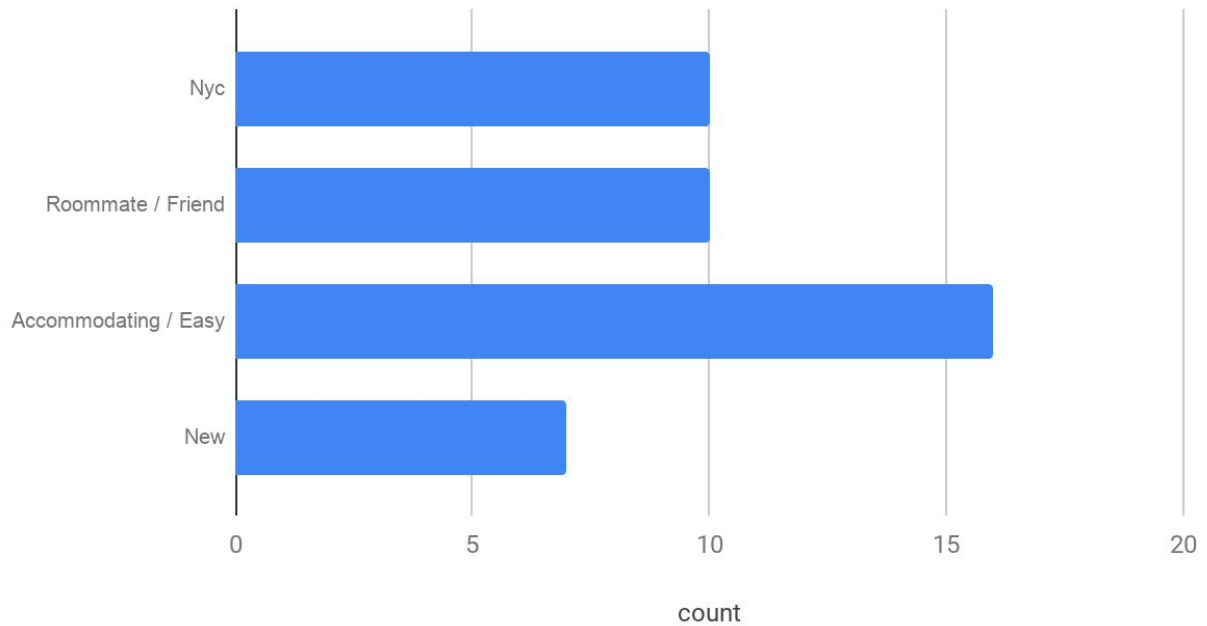"Service/Experience/responsive": 37 -- Good pre/after rent services

"Short-term": 8 -- Allow short-term rent

"Roommate/friends":12 -- PlaceMe provide matched Roommate/Friendships

"Clean/ Living/Everything/furnished/home":39 -- PlaceMe provides all the things for customers and makes them feel like home.

Bedly Positive Reviews:

Bedly Review

There are 25 positive reviews we found for Bedly. From all these reviews, here are some vocabs with high frequency worth notifying.

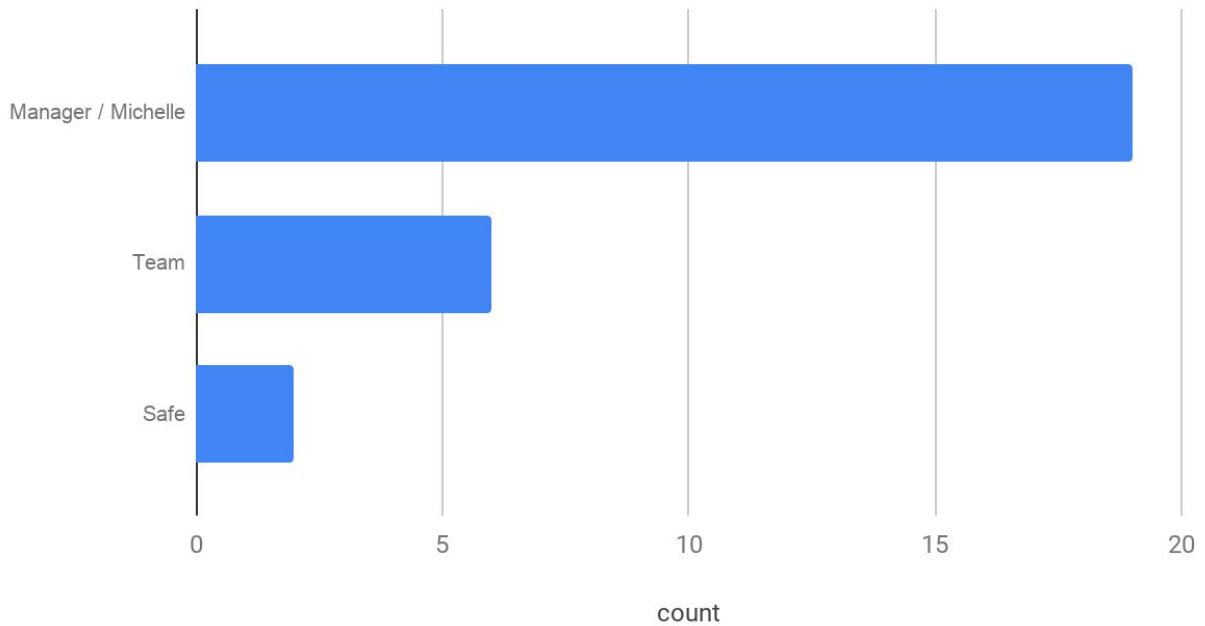"Nyc": 10 -- Bedly have their major services in New York City
"Roommate/Friend": 10 -- Bedly provide matched Roommate/Friendships
"accommodating/easy": 16 -- Quick/humanized services
"New": 7

Bungalow Positive Reviews:

Bungalow Review

There are 11 positive reviews we found for Bungalow. From all these reviews, here are some vocabs with high frequency worth notifying.
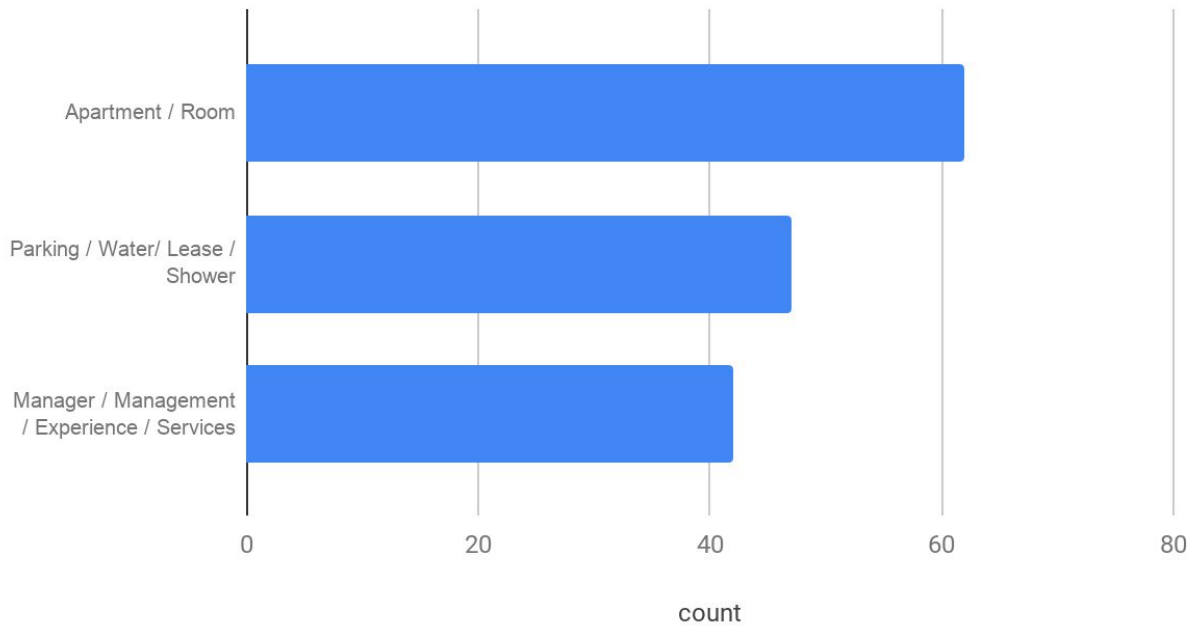
"Manager/Michelle": 19 -- Good managers(Especially a manager called Michelle)
"Team": 6 -- Good response and nice teams
"Safe": 2 -- It just mensions twice but we did not see this field in other companies' reviews. We believe safety is also important.

All Negative Reviews:

All Negative Review

There are 45 Negative reviews. From all these reviews, here are some vocabs with high frequency worth notifying.

"Apartment/Room": 62 -- Bad quality of apartments.
"Parking/Water/Lease/Shower":47 -- Lacking facilities.
"Manager/Management/Experience/Services":42 -- Bad services.

For all other vocabs in the reviews, you can check the result CSV files.

What Can PlaceMe Do:

1: Keep the good things.
2: Learn from the good reviews of other companies.
3: Avoid the bad things in the negative reviews.

What To Improve In the Future:

      1: Collect more review and maybe conduct a survey to get feedback from customers directly

      2: Try other methods to analyze the data.

Part3: Possible Customer and their expected prices.

Why:

If we can help PlaceMe Living select the appropriate target and consumer, it will greatly reduce the company's expenditure on advertisement. Therefore, we want to find customers who match the ideal price of the corresponding house from the analysis of the dataset and PlaceMe could advertise them.

How:

We select variables and create a dataset from ipums. We choose 22 variables that we think might correlate to their willingness to pay for rent. We had a readme file explaining labels and associated numbers we used in this dataset.

Experiment:

The first thing to do is to process data:

      We only use data in MA and cities we choose are Boston, Cambridge and Brookline. Secondly, we drop data who has rent 0 because they have their own house and do not want to rent. Now we end up having a dataset with 13844 data entries and 15 columns. Each of the entries is a person with 15 variables to describe their state.

      We use the rents as labels. According to the rent distribution, we decided to put them into 4 classes:
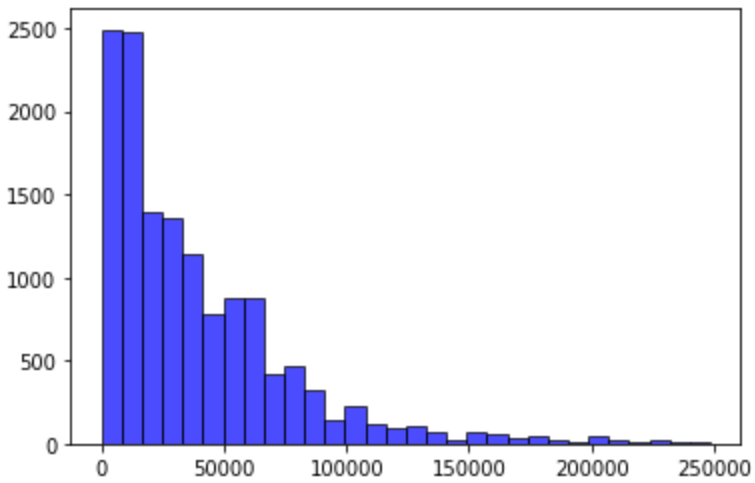
      0: rent = $0- $1300

      1: rent = $1300-$1900

      2: rent = $1900-$3000

      3: rent > $3000

And we are going to predict which class a person is in with the other 14 columns.
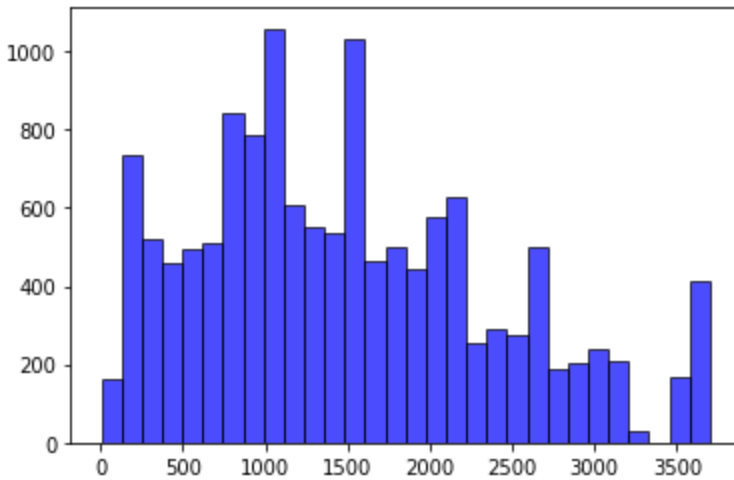
Correlation of Each Field:
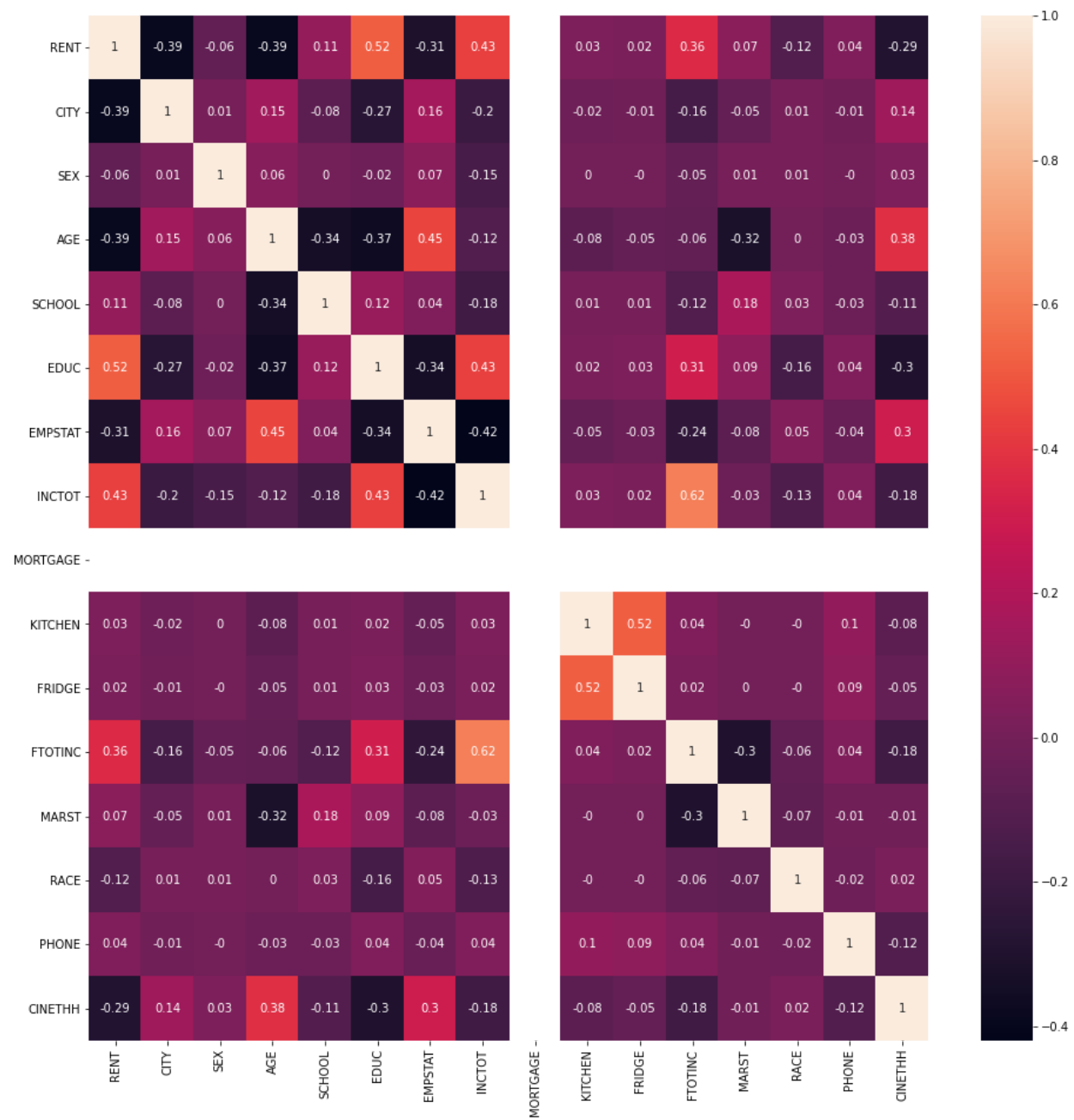
Income:

Rent:

Methods:	Accuracy:
Logistic Regression	0.5953250547845143
Random Forest	0.6172388604821037
MLP((20,15))	0.5931336742147553
SVM (C = 1.5)	0.6092037983929875
LinearSVM(C=3)	0.5997078159240321

KNN(K=47)	0.6055514974433893
SVM + Logistic Regression	0.6166064981949458
SVM + Random Forest	0.5971119133574008
SVM + SVM	**0.6252707581227437**
SVM + LinearSVM	0.6151624548736462
LogisticRegression + LogisticRegression	0.6209386281588448
LogisticRegression + SVM	**0.6259927797833935**
LogisticRegression + Random Forest	0.5891696750902528
LogisticRegression + LinearSVM	0.6166064981949458

Explanation:

The best result is given by the combination of 2 models (**LogisticRegression + SVM**). We first use Logistic Regression to train on the ["INCTOT","AGE"] columns of the data since they are different based on people. We then use SVM to train based on the rest data columns with just multiple classes because they are on the same scale. The data columns including: ['CITY', 'SEX', 'SCHOOL', 'EDUC', 'GRADEATT', 'EMPSTAT', 'FAMSIZE', 'MORTGAGE', 'KITCHEN', 'FRIDGE', 'FTOTINC', 'MARST', 'RACE', 'PHONE', 'CINETHH']

What can placeme do:

- Advertisement on school housing websites for certain house sources.
- Recommend/email certain house sources for people moving to boston.

What To Improve In the Future:

1: Find data entries with higher correlation to improve dataset quality.
2: Try other more complex machine learning models.

Reference:

https://bostonpads.com/2019-boston-apartment-rental-market-report/

https://usa.ipums.org/usa/