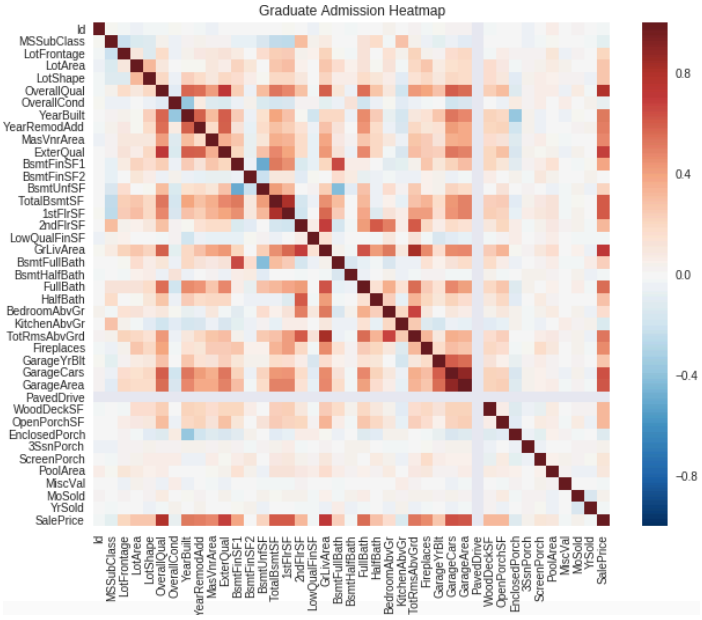
House Price Prediction

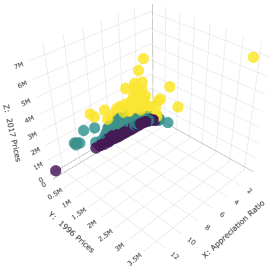
Housing price is one of the most concerned economic topics. In different countries, there are a lot of elements can influence the house price, including house size, location, room numbers, etc. In order to predict the house price, we not only can utilize the time series to forecast the future price of the house but also predict the house price based on the house features. I utilized two datasets to explore the house price prediction. I used time series forecasting for the Zillow data and used traditional machine learning model for the House Prices data (https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data)

The first step is to read data and clean data. There are some 'nan' within each dataset. Based on properties of variables, fill these 'nan' as mean, or zero, or just delete the whole row that with nan in it. Meanwhile, in order to figure out the correlation between each variable, I changed some variable from categorical into numerical.

Secondly, utilizing EDA to figure more relationship between each variable.



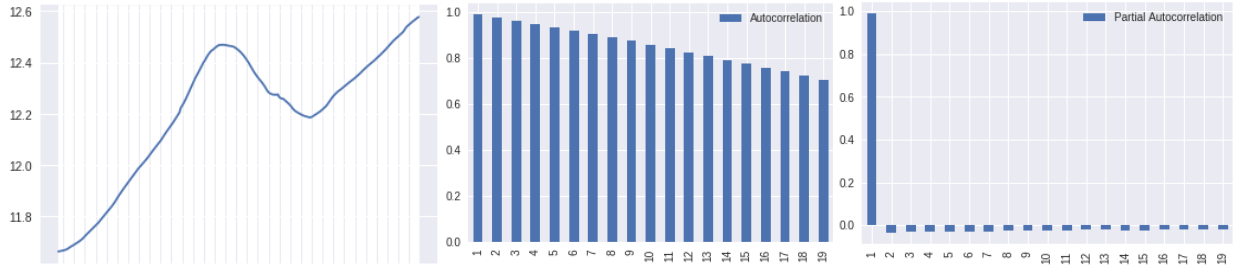
(The Left Picture) This is a heat map of House Prices Data. According to this map, we can clearly figure out the correlation between each variable, with the darker of the red, there is a stronger positive relationship between these two variables. Meanwhile, with the darker of the blue, there is a stronger negative relationship between these two variables. Based on this heat map, I chose to obtain several variables (which correlation with Sale Price is stronger than 0.2) for my traditional machine learning modeling in the end.



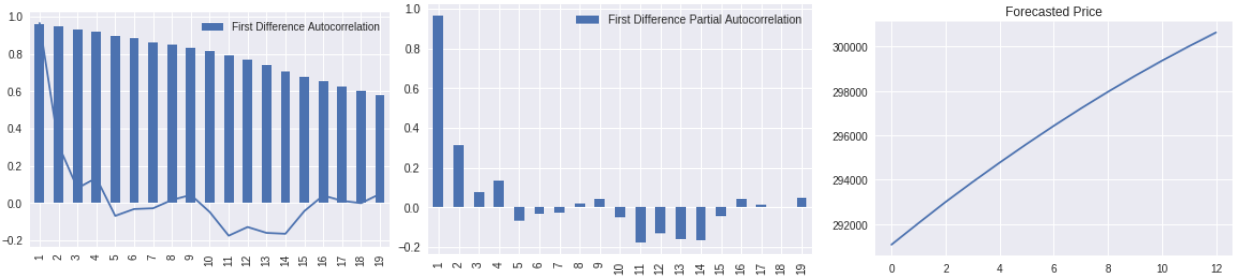
(The Right Picture) This is a 3-dimensions Scatter Plots about the 30-year history USA Real Estate Prices, which colored based on clusters. And the clustering is based on the Size Rank and Price. According to this plot, we can clearly figure out that as time goes on, the price of the house is increasing.

Thirdly, using Time Series Forecasting to predict the mean price of the US house property, and using traditional machine learning model (Random Forest, Decision Tree, Logistic Regression, and Gaussian Naïve Bayes) to predict the house prices.

Because the Zillow Data provided us all properties prices from 1969 to 2018, hence we can utilize time series to predict the future price of the house. In this case, I chose to predict the average house price from 2019/01 to 2019/12.



(Left: log mean price of the house in the US based on time series, Middle: bar-chart of the autocorrelation, Right: bar-chart of the partial autocorrelation.)



(Left: first difference autocorrelation, Middle: first difference partial autocorrelation, Right: The prediction result of the average house price in US from January 2019 to December 2019, and according to this graph, we can get the price of house will be increased in the following year)

House Prices Data provided plenty of variables about a house, including its location, room size, room numbers, year sold, etc. After filtering the variable, I kept 22 variables for the final prediction. And after comparing four model's prediction score, I chose to use Random Forest for my final prediction.

	Model	Score
0	Random Forest	99.860
1	Decision Tree	99.860
3	Gaussian Naive Bayes	57.400
2	Logistic Regression	52.740

In conclusion. The time series forecasting is a good machine learning model for predicting the future price or future trend of an object. The traditional machine learning models are useful for those data which including a lot of various variables in it. And then, we can figure out the correlation among these variables and to do more prediction about the price of the house which has similar features as others. All in all, both of these methods are useful for the price prediction, and we just need to define what kind of result we want to get before we choose the models.

Here is my Colab Link of this homework:
https://colab.research.google.com/drive/1p-vX3Peidbf9tQjXXPXlFom6D3eFsfsw