Project – Milestone I – Xiangyu Pei

• Description of the dataset and its features.

The dataset is coming from a "House Prices: Advanced Regression Techniques" which is a project of Kaggle competition. It contains both training and testing data which are including the detail information of each transaction houses (eg. Id, Street, LotArea, Utilities, Neighborhood etc.) out of 81 and 80 attributes separately.

• The analytical questions to be answered

1. What factors are correlated with the final house price.
2. What is the relationship between Year built and garage built?
3. When is the best-selling month for houses?
4. The number of houses sold on georgical?
5. Predict the house price in the next few years on different locations because house price would be quite different on different locations.

• Features of the data you will represent

The following features of the data I would like to use, which including **house sale prices, sale condition, year sold, month sold, location, garage, neighborhood, building style, overall quality, roof condition etc**. to approach the analytical questions as mentioned.

• Analytical questions to be answered by your visualizations

1. I would like to use scatter plot to check what factors are highly correlated with the final house price.
2. I would like to use **temporal data** to show the relationship between "year built" and "garage built".
3. I would like to use bar chart to show the best-selling month for houses.
4. **Geographical data** to demonstrate the relationship between "numbers of house sold" and "location".
5. I would like to use **temporal data** to predict the "time" and "house sold price" in the near future.

Notes:

A. I will create a model base on the training model with random forest and gradient boosting, then I will test the testing data in the model I just created for further data visualization.
B. I will create a Shiny R to demonstrate all my visualization data in a better way.

• Questions should reflect on how the user is going to interact with the Data

This model will be built for the house price prediction so that all users can easily check the house price in the near future. The final decisions also included the correlated factors and the georgical data of the house. The best-selling time would be demonstrating to all users as well. Therefore, all of those details are valuable for investment companies to make their decisions.