**Data visualization and analysis based on a case study**

**Dataset:**

* Introduction: For this case study, a dataset is downloaded from Kaggle website [1] called “The Boston Housing Dataset” [2]. In this report, this dataset is used to extract some business related information to highlight different features and analyze the data with help of a data visualization tool. Also, the report presented some key findings from the data analysis.

The Boston Housing Dataset was collected by U.S Census Service and appeared as statistical analysis paper authored by David Harrison Jr. and Daniel L. Rubinfeld [3]. According to the analysis, the researchers were convinced that people are willing to pay more for a clean environment around the housing area [4] [5]. Later, the dataset was first published publicly in 1978 to understand and analyze the housing prices based on various features like crime rate in the suburb area, concentration of Nitrogen Oxide in order to understand the air pollution rate, average number of rooms, age of the property, tax rate, and proportion of the lower status population etc. The dataset contains total 506 housing price samples including total number of 14 features. The dataset can demonstrate how these features can manipulate house prices. Identifying the most important feature will help us predict the house prices and also to understand how these features are co-related to each other. In this report, the dataset is visualized and analyzed to focus on the key relations and importance of these features.

* Description of the Features:
  + Per capita crime rate by town: Number of crimes in a town divided by total population of town, which identifies how secure the area is.
  + Proportion of residential land zoned for lots over 25,000 square feet
  + Proportion of non-retail business acres per town.
  + Charles River dummy variable.
  + Nitrogen oxides concentration, which refers to air pollution level.
  + Average number of rooms per dwelling.
  + Proportion of owner-occupied units built prior to 1940.
  + Weighted mean of distances to five Boston employment centers.
  + Index of accessibility to radial highways.
  + Full value property tax rate per ten thousand United States dollar (10000 USD).
  + Pupil teacher ratio by town, which indicates the education facility around the area.
  + Proportion of black people by town.
  + Lower status of the population (percent).
  + Median value of owner-occupied homes in per $1000s.

**Data Visualization and Analysis:**

* Software/ Tool used: For data visualization purpose Tableu [6] is used, which is an open source software to visualize and analyze data, which is widely used for business intelligence and business analytics.
* Insights:
  + Trend lines based on the relationship between median house price and some important features:

A trend line can show how two features are related to each other. For experimenting we have used exponential trend lines which fits our data points more smoothly than linear trend lines. Figure 1 shows how median price of the households are changing based on some important features like crime rate, lower status population, nitrogen oxide rate, average number of rooms, pupil-teacher ratio and tax rate. We can clearly identify that lower crime rate, small portion of lower status population, lower nitrogen oxide concentration are increasing the median price value as the trend line goes downward when these value increases. Similarly, more number of rooms, lower pupil-teacher ratio and lower tax rate increases the house price. The trend line is going upward in these cases. People are willing to pay more for a household where it is safe with decent environment and other facilities. Also average people status plays an important role on the median price of the household.

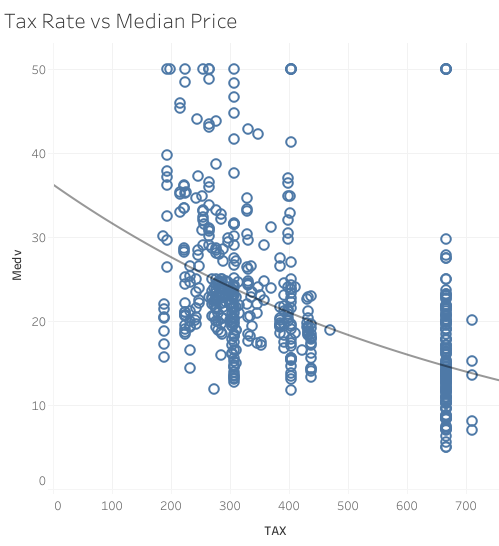
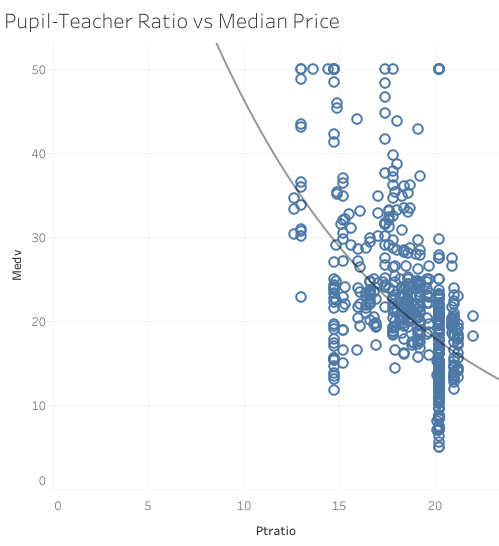
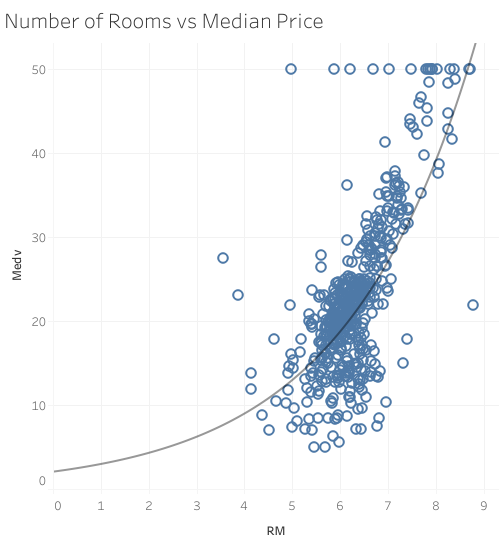
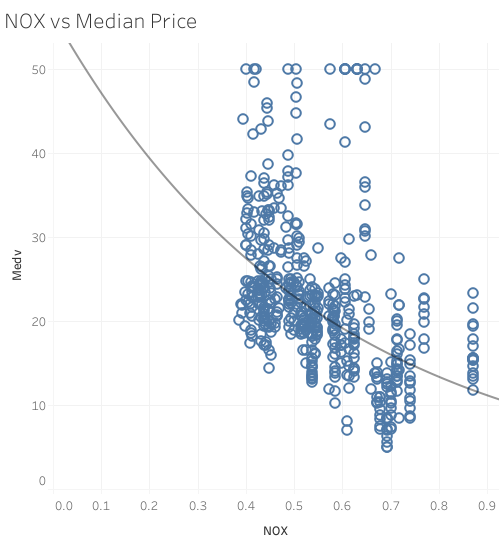
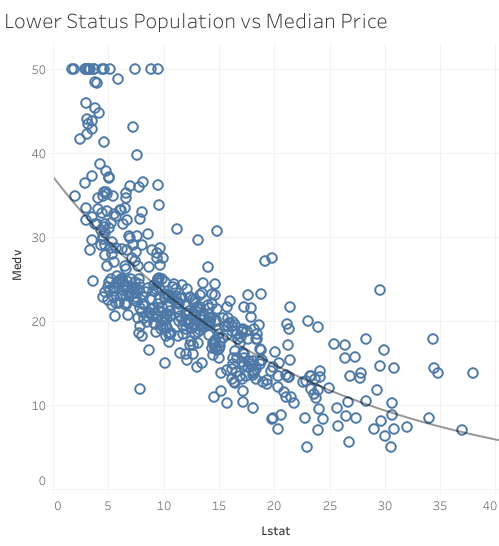
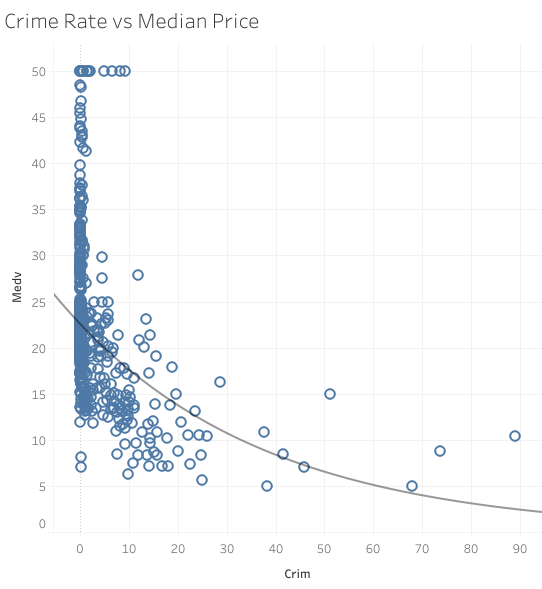


Figure 1: Trend Lines based on different features

* + Average house features based on price:

Figure 2 and 3 shows how all the other features vary between high and low prices. It shows an average value when a range of house price is selected. From the data, we can say that people who want to spend low price ($5000 - $7500) for a household will get comparatively old houses with higher crime rate in the neighbor, also where lower status population is higher. But expensive households will increase these features. More price will provide comparatively new houses, which has greater accessibility to the highways and lesser distance from employee centers. Also pupil teacher ratio is higher, so the facilities of the high priced households are better than low priced households. Figure 3 shows the highest range of price where the features vary from figure 2.

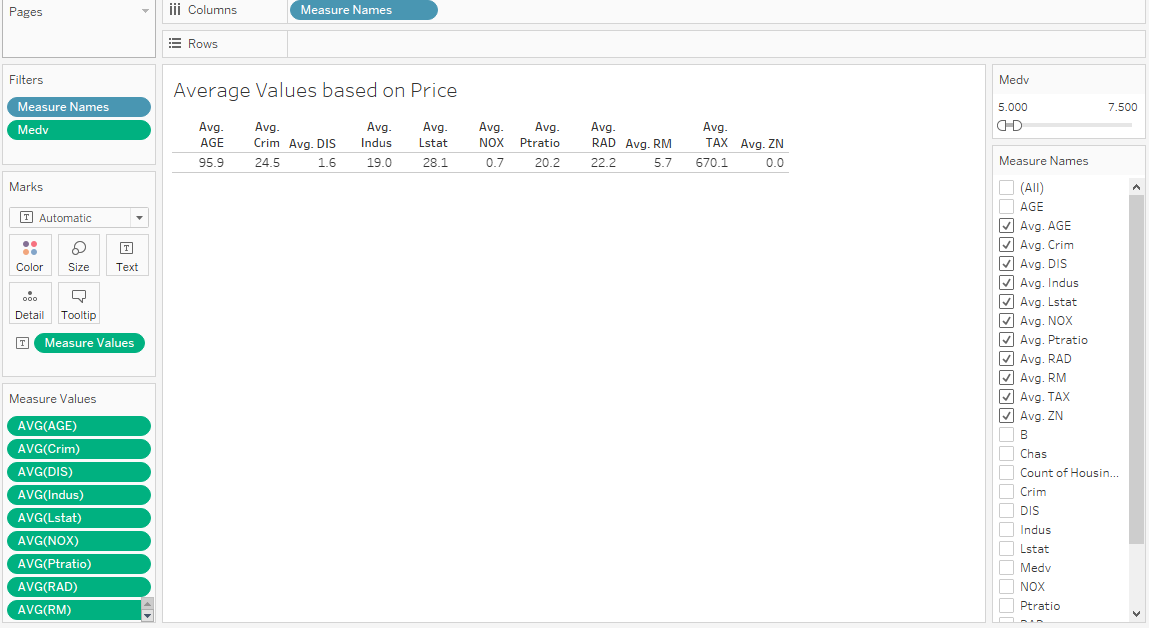


Figure 2: Average features of a household for low price ($5k-$7.5k)

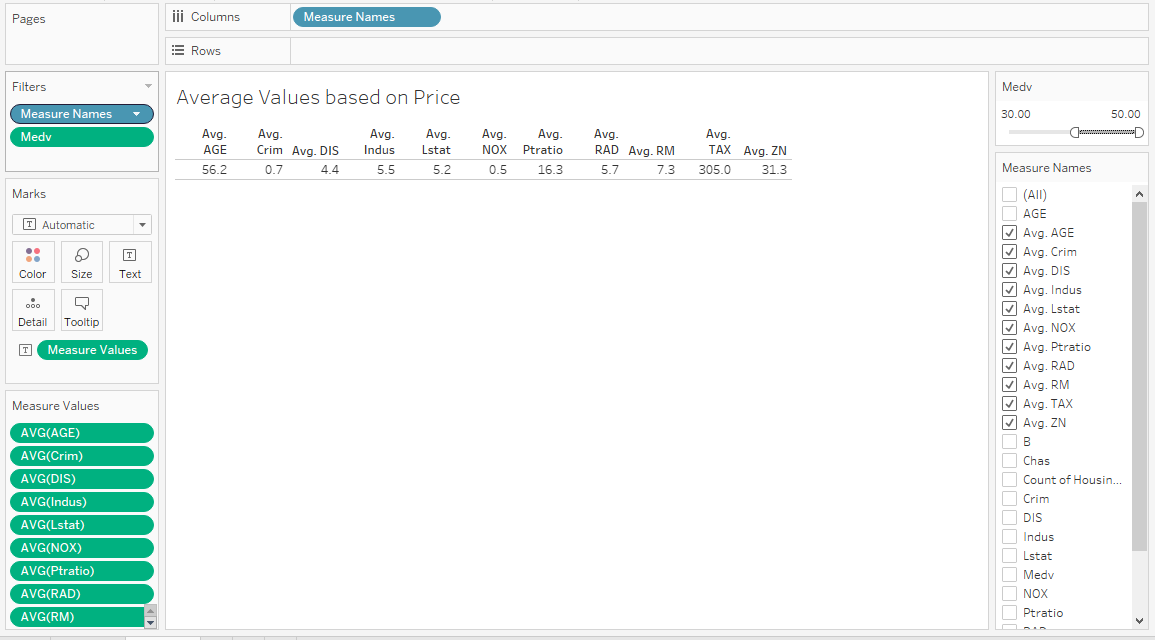
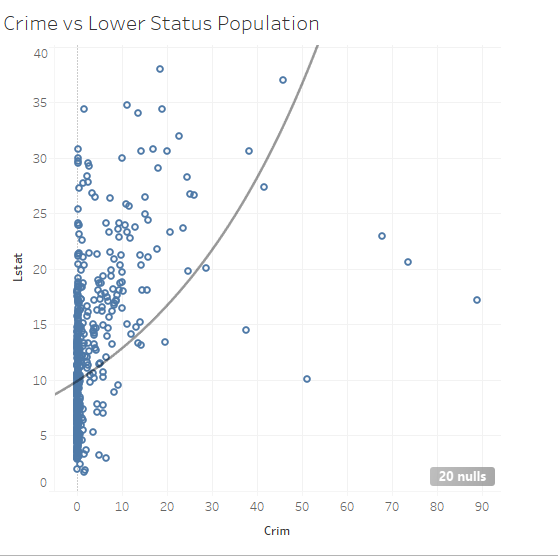
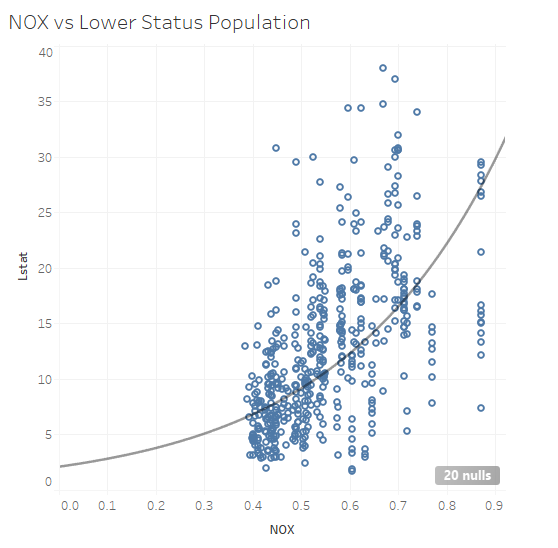


Figure 3: Average features of a household for high price ($30k-$50k)

* + Impact of lower status population:

Figure 4 shows there are more crime rate and air pollution where lower status population percentage is higher. They also tend to live in old houses of the town where other facilities are less. That’s why it also reduced the median price of the households. So, we can say that lower status population is one of the influencing feature for lower household prices based on the data. Lower status population percentage, crime rate, air pollution, median house price and other facilities all are co-related features.



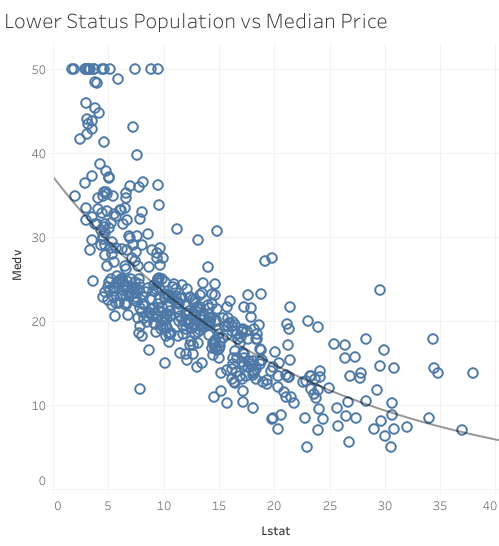
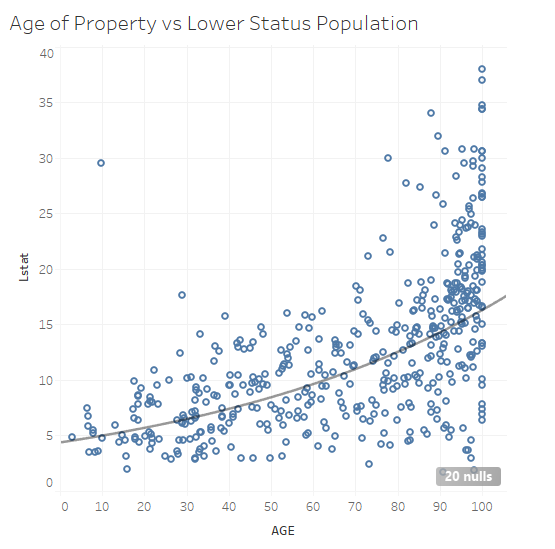
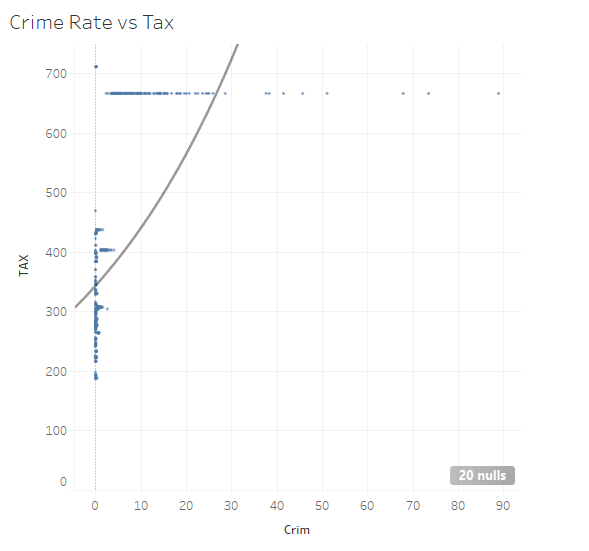
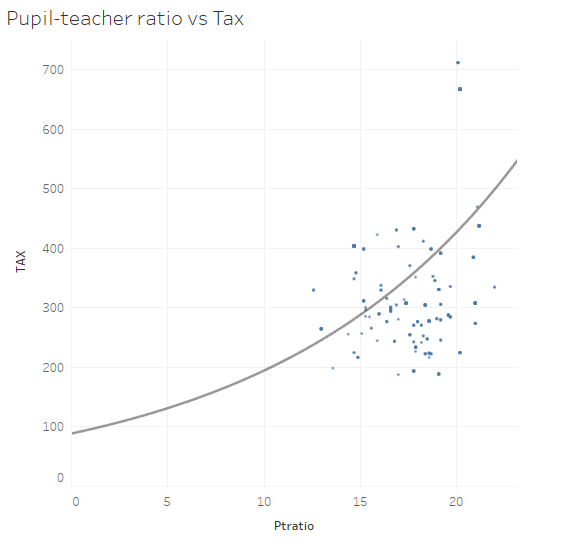
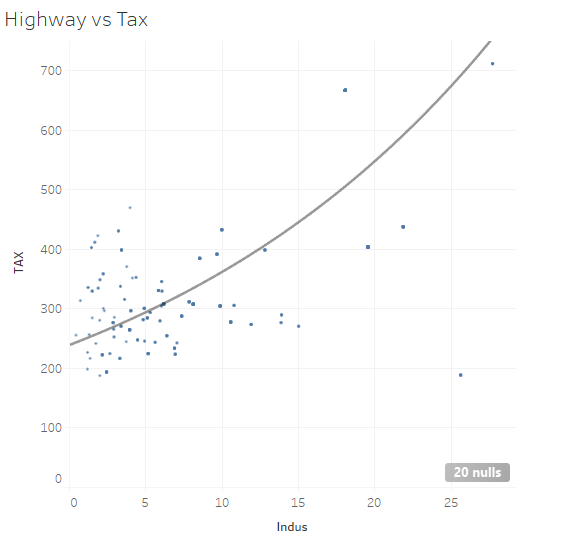
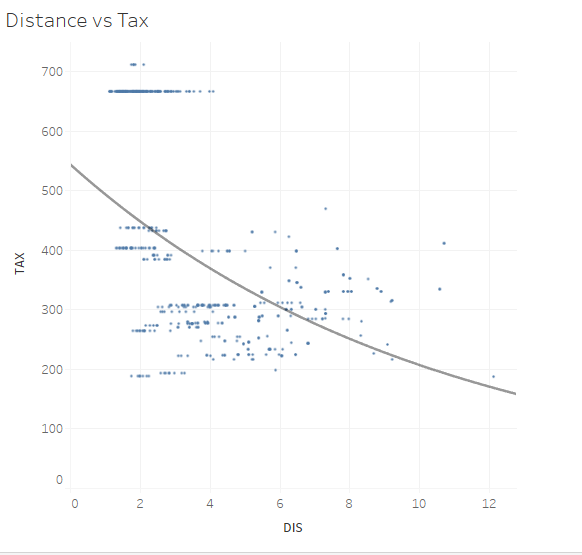


Figure 4: Impact of lower status population

* + Tax Rate: Households far from employee centers have less tax rate than the nearer ones. People who live near highways and have more access to the highways have to pay more tax. Households with better pupil teacher ratio also has a greater tax rate. It suggests that higher facilities increase tax rate. Hence, higher crime rate reduces place value and other benefits, tax rate are also reduced. Tax rates mostly depend on the facilities and benefits in the surrounding area.



* Recommendations: From the data, we can draw a few conclusions. The house prices largely depend on the surrounding environment of the area, like how safe the neighborhood is, the crime rate and air pollution rate or in this case the rate of concentration of nitrogen oxide in the air. Lower status people are more likely to decrease the quality of the environment by committing more crime and polluting environment, tax rates are dependent on the facilities of the area such as how close it is from employment centers or the pupil teacher ratio. And all of the features are co-related to each other. From all of these conclusions, we can suggest a real estate company when selling a house hold to emphasis on the surrounding environment as well as the neighborhood and then fix the price for it. By improving on these features, customers will also get benefitted because then real-estate services will be more customer oriented.

**Conclusion:**

The chosen dataset shows very realistic examples of real estate household data for a particular time period. Though the time period is back dated and the dataset is not that very large, but from the data, we can determine that prices of a household is not only dependent on the house itself, rather its surrounding environments and other facilities play very important role in this. The dataset will help any real estate company to understand these important features so that they can improve them in order to make profitable business decision and more customer oriented pricing policy.

**References:**

[1] Kaggle Datasets, Accessed on: Aug. 27, 2020. [Online]. Available: https://www.kaggle.com/datasets

[2] Boston housing dataset, Accessed on: Aug. 27, 2020. [Online]. Available: <https://www.kaggle.com/altavish/boston-housing-dataset>

[3] Harrison, David & Rubinfeld, Daniel, Hedonic housing prices and the demand for clean air. Journal of Environmental Economics and Management (1978), Journal of Environmental Economics and Management

[4] D. A. Belsley, E. Kuh, R. E. Welsch, Regression Diagnostics: Identifying Influential Data and Sources of Collinearity (1980), Wiley Eds.

[5] O. W. Gilley, On the Harrison and Rubinfeld Data (1996), Journal of Environmental Economics and Management

[6] Tableau, Accessed on: Aug. 27, 2020. [Online]. Available: <https://www.tableau.com/>