

**Introduction of Data Analytics**

**(DATA 1200)**

**Assignment #1**

**Preliminary Analysis**

**Submitted by: Kartik Sojitra**

**(100723768)**

**To Professor: Sam Plati**

**Part A:**

1. **Preliminary analysis:**

The housing dataset consists of dependent and independent variables. Price is a dependent variable that differs according to the features of the house. The Lot size of the house has a higher influence on the housing price compared to other features as depicted in the dataset. The two histograms of price and lot size are positively skewed. From this, it can be estimated that with an increase in the lot size there is a substantial elevation in the prices of the house.

**2) Effect of Independent variable in housing price estimation:**

Lot size, Bedroom, bathroom, basement arethe independent variables that affect the housing price. The housing price is highly affected by the lot size of the house followed by bedrooms. Bigger the lot size higher is the price of the house. The number of bedrooms also has a significant effect on the housing price. Bathrooms and garage have a minimalistic effect on the prices. From the histogram, the standard deviation can be estimated. 65% of the data points fall in the mean standard deviation of the dataset. There are no outliers in the histogram. Thus, it can be concluded that the data points have a considerable impact on the pricing.

(Plati, Analytical Tools, 2019)

**Part B:**

**CRISP-DM model for Medical Billing: iPatientCare**

**1. Business understanding:**

iPatientCare is a healthcare company that works for Physicians by collecting and analyzing the patient data and submitting it to the insurance companies. It also provides medical billing software to the physicians. The prime goal of the company is to generate maximum revenue of the physicians by minimizing various constraints- timely arrival of patient records, first claim passing ratio, efficiency of medical coder. Constraints can be solved by creating a proper workflow pattern that provides quick submission of the patient records, an increase in the efficiency of a medical coder by providing additional training, increase the first claim passing ratio.

**2. Data Understanding:**

This phase consists of collecting the patient data and its processing. The data can be collected from the physician's office. The data must be accurate and should be cross verified by the audit department before submitting it to the insurance company. The data piling must be reduced by increasing the workforce and changing the workflow pattern making the employees/ medical coder more productive. Also, the data inflow can be increased by providing proper treatment to the patients. Thus, generating higher revenue for the physicians. The daily submissions of the medical data to the insurance company also influence the revenue generation process. The time taken by the insurance company to process the data is an independent variable. The cost that is generated from the initial stage of data collection to the submission stage of the data including the hardware software-related services, also significantly influences the revenue.

**3. Data Preparation:**

In the data preparation step, we prepare and clean the provided data. Based on the dependent and independent variables such as cost of analyzing medical data, personnel training, first claim passing ratio, time of medical data submission, the data can be managed and prepared by using packages in R and Python. The medical data can be analyzed using various graphical representation techniques and determining any Outliers if present.

**4. Modelling:**

Modeling consists of an implementation of different algorithms and evaluating them. Different algorithms are created based upon the business requirements and desired outcome, all those algorithms are executed and analyzed. The one that works is selected for the project. Also, changes can be made in the algorithms and reimplemented.

**5. Evaluation of the model:**

Evaluation of model is undertaken by the company using different algorithms such as Regression, Naïve Bayes, Classification and Regression Trees (CART). The algorithm depends on the type of data to be evaluated, its size and the quality of the data. The selection of the algorithm highly depends on the business requirements and outcome, type of algorithm model used.

**6. Deployment:**

In this stage, the final execution of the model is done. The model that passes the evaluation is presented to the business company and then it undergoes validation processes. After the validation process is completed the model is finally published.

(Plati, 2019)

# References:

Plati, S. (2019). Analytical Tools.

Plati, S. (2019). Predictive modelling overview.