

**Data 1200**

**Introduction to Data Analysis**

**Assignment #1**

**Preliminary Analysis**

Submitted by:

Name: **Nithin Thomas**

Student number:**100804269**

Submitted to:

**Prof. Omar Al Trad**

**PART B  
ASSIGNMENT SUMMARY**

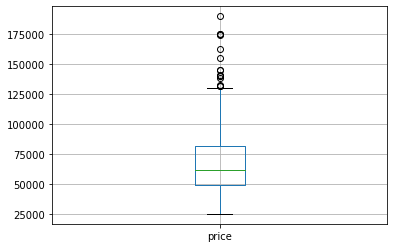
**Purpose and Assignment Objective:**

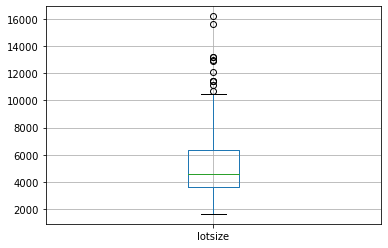
**Present and explain/justify four (4) key insights from the key statistics output that was completed in Part (a). For example: Mean, Standard Deviation, etc.**

1. **Basic Statistical Details**



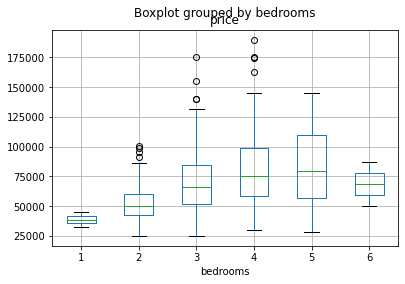
* Standard Deviation of **'price'** and **'lotsize'** is **high** which clearly indicates that each data points in the dataset is **away from the mean value** of the corresponding attributes. Here **high standard deviation** indicates that **datapoints are more spread out** i.e. in case of **'price'** and **'lotsize**’. While **other attributes datapoints** are **clustered around the mean value**.

1. **Boxplot for Price and Lot Size**

****

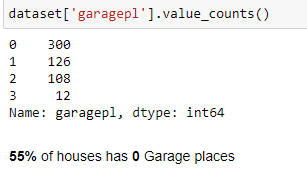
* As predicted by the Standard Deviation values, it’s clear from the above bloxplot that **'price' and 'lotsize'** has **outliers which will affect the mean value of the attributes**.

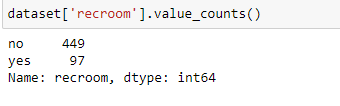
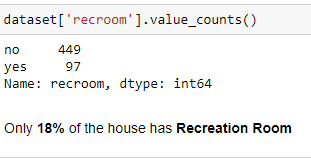
1. **Bedroom vs Price**

****

* **Average price** of houses **increases** with the **increase in no. of bedrooms** until 5 then it reduces as per the available dataset.

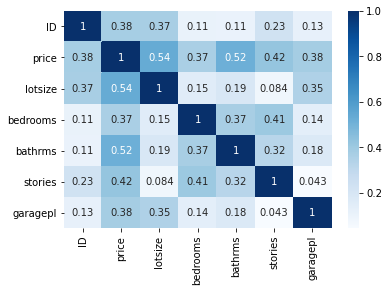
1. **General Insights**

****



* **68%** of the houses **do not have Air-conditioning.**
* Only **18%** of the house has **Recreation Room**
* **55%** of houses has **NO** Garage.

1. **Correlation matrix**



* Correlation is the best between **lot size** and **price**.
* Correlation is least between **garage places** and **stories**

**PART C**

**ASSIGNMENT SUMMARY**

**Purpose and Assignment Objective:**

Using the CRISP-DM model create an outline how you would develop an algorithm to generate insights for your favorite product or service.

# **CONTENTS**

[1. Introduction 1](#_Toc84018015)

[2. Business Understanding 1](#_Toc84018016)

[2.1 Business Problem 2](#_Toc84018017)

[3. Data Understanding 2](#_Toc84018018)

[4. Data Preparation 2](#_Toc84018019)

[4. Modeling 4](#_Toc84018020)

[5. Evaluation of the Model 4](#_Toc84018021)

[6. Deployment 4](#_Toc84018022)

[7. REFERENCE 6](#_Toc84018023)

**1. Introduction**   
In this project, we are required to analysis data with **CRISP-DM** process. The CRISP-DM process is below.

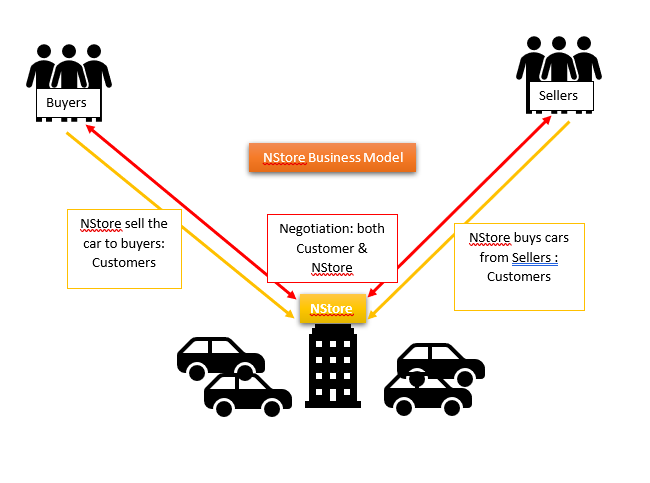
#### CRISP-DM (Cross-Industry Standard Process for Data Mining)

* Business Understanding
* Data Understanding
* Data Preparation
* Modeling
* Evaluation
* Deployment

So, I will be analyzing data using the CRISP-DM Model and come up with a model for prediction.

# **2. Business Understanding**

NStore is an automobile car resale dealer. Their main source of income for NStore is commission and profit gained in the resale process. It is important for NStore to understand the market value of the car because it will affect the revenue of the company. Because if the quoted price of the car is high then the sales will drop due to overpricing and if the quoted price by the company is less than the company will lose its margin on the sales. So, it is important to their business and I expect it to be one of their KPIs.



*Figure 2.1 NStore Business Model*

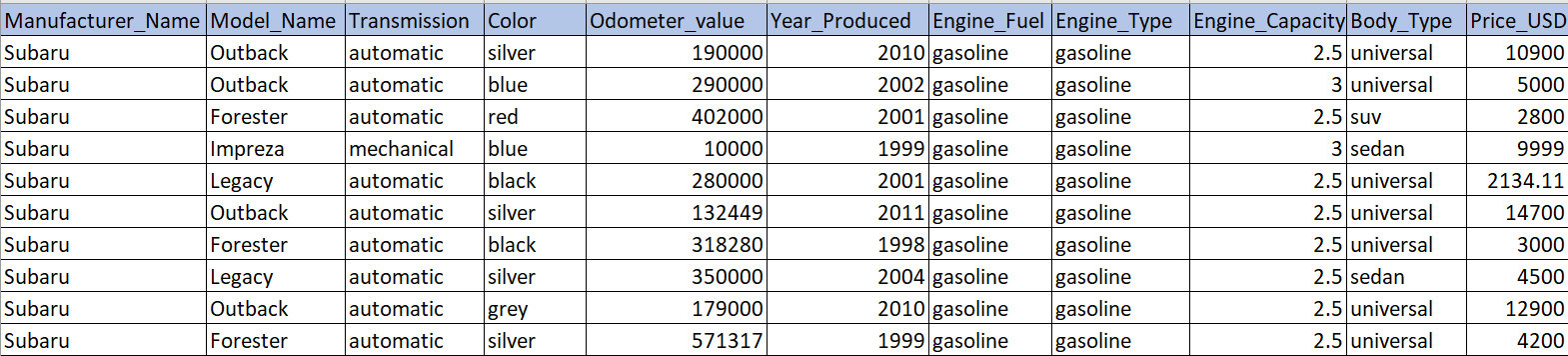
## **2.1 Business Problem**

The first step in CRISP-DM process is to understand the business problem. Profitability of the NStore has reduced drastically this financial year so company wants to develop an algorithm to predict the price of the car models depending on the car attributes. More precise, Company si trying to find answers for the following 3 business questions which will directly affects the profitability of the company:

* Is the price of a car related to car body type?
* Is the price of a car related to gear transmission type?
* Is the price of a car related to odometer reading?
* Is the price of a car related to type of engine type and capacity?
* Can the price of a car be predicted based in its attribute with reasonable accuracy?

# **3. Data Understanding**

The data used for the analysis is taken from NStore sales database. Below is a sample of the data:



*Figure 3.1: Sample Data*

**The above sample contains all columns. All the attributes are listed in the above screenshot Manufacturer\_Name”, “Model\_Name”, “Transmission”, “Color”, “Odometer\_value”, “Year\_Produced”, “Engine\_Fuel”, Engine\_Type”, “Engine\_Capacity”, "Body\_Type”. The last column is the “Price\_USD”, company want to predict the price based on the other attributes.**

# **4. Data Preparation**

In this step, we clean the available data by removing missing values, outliers and by standardizing the data. There are many steps that one should follow to complete the data preparation phase.

Identify the cells containing NA’s and replacing or filling them with appropriate data. Ex. In this dataset certain rows contains NA’s as recorded as Body type which can be identified and replaced using the attributes Manufacturer\_Name” and“Model\_Name”.

Similar to NA Treatment, null values also have to be treated separately because of the following reasons: -

1. The first and foremost step being the **NA treatment**. Normally the data at hand is not clean and always have NA. We must identify such values and appropriately fill or replace them. 
   1. Treating Continuous data

Mean/Median/Mode can be used to replace the missing values accordingly.

For example, price of the some of the models are not recorded which can be replaced by the mean value of the similar models.

* 1. Treating categorical Data

Categorical data contribute significantly for the model creation so its important to fill those correctly otherwise we will end up in with a wrong model.

Example, “Transmission” should be mentioned as “Unavailable” wherever not list because same model and engine type has different transmission type which significantly affect the price of the models.

1. **Outlier Treatment**

All data has outliers treating of outliers are very important. Listed below are some of the methods of treating the outliers: -

* Trimming/removing the outlier

We remove the data points having outliers, although it is not recommended.

* Quantile based flooring and capping

The data points that are lesser than the 10th percentile is replaced with the 10th percentile value and the data points that are greater than the 90th percentile are replaced with 90th percentile value.

* Mean/Median Imputation

Replacing the outliers with the mean value, as this outlier will drastically affect the mean value of the data.

1. **Standardization or Scaling of the data**

Standardization comes into picture when continuous independent variables are measured at different scales which means these variables do not give equal contribution to the analysis. So, the idea is to rescale an original variable to have equal range and/or variance lying between -1 to 1.

Methods of Standardization: -

1. Z score
2. Min-Max Scaling
3. Standard Deviation Method
4. **Feature Engineering**

The goal of feature engineering is simply to make your data better suited to the problem at hand.

Feature engineering is performed to:

* improve a model's predictive performance
* reduce computational or data needs
* improve interpretability of the results

For example, we are trying to predict the Price of square plots of land from the Length of one side. Fitting a linear model directly to Length gives poor results: the relationship is not linear. If we square the Length feature to get 'Area', however, we create a linear relationship. Adding Area to the feature set means this linear model can now fit a parabola. Squaring a feature, in other words, gave the linear model the ability to fit squared features.

# **4. Modeling**

To come up with a model for predicting the price based on the other features, we use different ML algorithm. The idea is to create a model which predict the price accurately. There are different algorithms available choosing the best model is very critical for the prediction. The accuracy of the model should be also checked with the already available data. Combinations of different algorithms can also be used for the final model for increasing the accuracy of the predicted dependent variable.

Using various Machine Learning tools, numerous tests are then executed with available data. The potential model is tested against measured data to see its actual performance. ML tools are capable of identifying new patterns which a human cannot recognize easily.

# **5. Evaluation of the Model**

Evaluation of the model is caried using techniques like Accuracy, Sensitivity, Specificity, F-Score, AUC, R-Sq, Adj R-SQ, RMSE (Root Mean Square Error) etc. This is the phase when you should explain how your model would help the business. Based on the business requirement the evaluation criteria are designed. This is also the phase where the benefits of the model is discussed, it should be realistic and avoid technical jargon that people outside of the data world could understand.

# **6. Deployment**

Created model will be only useful if the customer can access its results. The complexity of this phase varies widely. This final phase has four tasks:

* **Plan deployment**: Develop and document a plan for deploying the model.
* **Plan monitoring and maintenance**: Develop a thorough monitoring and maintenance plan to avoid issues during the operational phase (or post-project phase) of a model.
* **Produce final report**: The project team documents a summary of the project which might include a final presentation of data mining results.
* **Review project**: Conduct a project retrospective about what went well, what could have been better, and how to improve in the future.

The business case is presented to the management with the help of PPTs, documents etc.

# **7. REFERENCE**

* <https://towardsdatascience.com/crisp-dm-methodology-for-your-first-data-science-project769f35e0346c>
* <https://www.datascience-pm.com/crisp-dm-2/>
* <https://www.proglobalbusinesssolutions.com/six-steps-in-crisp-dm-the-standard-data-mining-process/>
* <https://www.kaggle.com/yaginun/crisp-dm-process-on-the-airbnb-dataset>
* <https://towardsdatascience.com/crisp-dm-ready-for-machine-learning-projects-2aad9172056a>