

**This assignment will be on Data Preparation**. The dataset needed for this are present on the server in the following location.

**Z:\ Foundation Exercises\** **Assignments\Class11 - Data Exploration and Preparation\Data Preparation**

**Data Prep Exercise 1: Missing Data and Outliers**

In the csv file **DataPrep1.csv:**

* How many variables have missing values (frequency distribution)
* If all missing value records are deleted, what is the total loss of records?
* If you decide to update the missing values with mean, what is the impact on standard deviation of each variable?
* If you decide to impute missing values, how could you potentially impute minutes used4?
* How many outliers can you identify for each variable?
* If you replaced the top two extreme values by the mean for any one of the variables, how would the standard deviation change?

**Data Prep Exercise 2: Data Preparation**

Use the data in the csv file DataPrep\_Final.csv to answer the following questions:

An auto manufacturer wants to estimate a price range for a new model of a car that is being developed. It has collated data on prices of other cars being sold in the market, and attributes of each model to use as a guide in pricing its new model. The dataset is attached here:

* How could the data be used to arrive at a suggested price?
* How would you prepare this dataset to be used in a model?
* What model form would be appropriate?
* How many valid records exist?
* What additional variables would you create?
* Instead of using the distinct values of some of the variables, it may be easier to create categories like High Medium Low or Above Average and Below Average. Supposing you wanted to understand if Height and Weight by categories – Small, Medium, Large, how would you create them?
* You think that in general length is not a positive driver of pricing, but some cars that perhaps belong to more luxury segments that may not be true. How will you create variables to capture this?
* Create an interaction variable between length and width
* You think that fuel efficiency, especially on highways is a positive driver of pricing premium, but it may not have the same impact at every price point. How will you test that?
* What follow-up questions would you have for the manufacturer on the data provided?