**Big Data Analysis**

**Big Data** is a huge chunk of data that requires DBMS to analyze and take out useful Information, insights from it. This process is called **Big Data Analytics**.

Consider a cellular company which is sending messages to its customers on the behalf of Banks, Income Tax department, Shopping stores, Insurance companies, Utility bills etc etc . this Cellular company is sitting on the ***gold mine of Data***, because it has the details of each of its customer’s Bank Accounts, Buying preferences, insurance payment amount, utility bills amount etc. So if an insurance company get the data from this cellular company .. the data will consist of irrelevant/ dirty data. where names might not be in proper format or there might be some alphanumeric values or simply lot of irrelevant information out of insurance domain. Here comes the part of Big data, to clean this huge chunk of millions of rows we use MatLab/ SAS/ Hive/ Apatche Spark/ pig etc .

Data cleaning is the Pre-Modelling technique with following steps in SAS:

1. Understanding the business and industry scenario
2. Understanding the data and the variable structure
3. Define of the purpose and scope of the project – finalization of key deliverables
4. Identification of redundant/relevant observations and variables
5. Cleaning/Sanity Check

* **Outliers** **– (Maximum 1% of Total data) – Floor Them/Cap them**

Extreme Values can be obtained by **PROC UNIVATRIATE** (2-3% on either side)

-It returns the Extreme values (lowest & highest) so we can detect outliers from it

-It returns the quantile values which tell us the distribution of data from 0% to 100% from

-Here we Can get the insights that on which quantile the data is not normally spread.

-It returns the summary of all variables, with MEAS, MEDIAN and MODE.

-If the data is not uniformly spread, Mean and Median would be very different

-Extreme High values are to be floored by the subsequent normal lower value

-Extreme Low values are to be capped by subsequent higher normal value.

* **Use Scatter Plot** by **PROC GPLOT** **DATA** for data visualization and reporting.
* **Invalid Values** – Make them Valid

**Treatment :**

**DATA** B01.A;SET B01.A;

ACS\_K3 = ABS(ACS\_K3);

IF AVG\_ED = **.** THEN AVG\_ED = **0**;

IF FULL <=**1** THEN FULL = FULL\***100**;**RUN**;

**(Here we have replaced the values containing “. “with “0”**

**and transformed the decimal values in percentages)**

* **Check for Missing Values** –**by NMISS Option**

Missing Value Treatment

**PROC** **MEANS** DATA = B01.m NMISS; **(Get the missing values)**

**RUN**;

**PROC** **MEANS** DATA = B01.m MIN MAX MEAN NMISS;

VAR MEALS;

CLASS MEALCAT; (**Geting the Mean, Max & Min obs wits missing values frequencies in each category of**

**RUN**; **mealcat)**

**DATA** B01.A;SET B01.A;

IF ACS\_K3 =**.** THEN ACS\_K3 =**19.1608040**;

IF ACS\_46 =**.** THEN ACS\_46 =**29.6851385**;

IF MOBILITY =**.** THEN MOBILITY =**18.2531328**;

IF MEALCAT = **1** AND MEALS =. THEN MEALS = **28.36**;

IF MEALCAT = **2** AND MEALS =. THEN MEALS = **66.0468750**;

**RUN**;

**Here we Replaced the missing values by the Mean of Non Missing Values grouped by Mealcat variable**

Many times the data comes in the text format which is to be arranged and cleaned in proper continues / categorical format, this task can be done by the TEXT functions.

Data cleaning is one of the most important but time consuming and tedious job to do. After doing these steps we are good to go for our statistical modeling part.

------------------------------------------------------------------------------------------------------------------------------------------