**A checklist for future data cleaning exercises.**

1. **Standardise units:** Ensure all observations under a variable have a common and consistent unit, e.g. convert lbs to kgs, miles/hr to km/hr, etc.

2. **Scale values if required:**  Make sure the observations under a variable have a common scale

3. **Standardise precision** for better presentation of data, e.g. 4.5312341 kgs to 4.53 kgs.

4. **Remove outliers:** Remove high and low values that would disproportionately affect the results of your analysis.

5. **Remove extra characters** like such as common prefix/suffix, leading/trailing/multiple spaces, etc. These are irrelevant to analysis.

6. **Standardise case:** There are various cases that string variables may take, e.g. UPPERCASE, lowercase, Title Case, Sentence case, etc.

7. **Standardise format:** E.g. 23/10/16 to 2016/10/23, Narendra Modi", etc.

**FilterData :A checklist for future data cleaning exercises.**

1. **Deduplicate data:** Remove identical rows, remove rows where some columns are identical

2. **Filter rows:** Filter by segment, filter by date period to get only the rows relevant to the analysis

3. **Filter columns:** Pick columns relevant to the analysis

4. **Aggregate data:** Group by required keys, aggregate the rest

**Univeriate Analysis:**

**Data Description:** The first step is to understand what it contains.

**Types of Variables**

**A. Difference between ordered and unordered categorical variables**

1. **Ordered ones** have some kind of ordering. Some examples are

Salary = High-Medium-low

Month = Jan-Feb-Mar etc.

2. **Unordered ones** do not have the notion of high-low, more-less etc. Example:

Type of loan taken by a person = home, personal, auto etc.

Organisation of a person = Sales, marketing, HR etc.

Apart from the two

**B. Quantitative variables**. These are simply numeric variables which can be added up, multiplied, divided etc. For example, salary, number of bank accounts, runs scored by a batsman, the mileage of a car etc.

**Unordered Categorical Variables - Univariate Analysis**

To summarise, the major takeaways from this lecture are:

* Plots are immensely helpful in identifying hidden patterns in the data
* It is possible to extract meaningful insights from unordered categorical variables using rank-frequency plots
* Rank-frequency plots of unordered categorical variables, when plotted on a log-log scale, typically result in a power law distribution

# **Basis of Segmentation**

In the last lecture, you have seen some glimpses of segmented univariate analysis. Now, let’s go deeper into the segmentation process. The entire segmentation process can be divided into four parts:

1. Take raw data
2. Group by dimensions
3. Summarise using a relevant metric such as mean, median, etc.
4. Compare the aggregated metric across groups/categories

Difference between correlation and causation.

Causation means due to this reason impact on result is different. For example. If parents have more childs that it dones’t means that child score less. (Correlation used). If parents have more childs and their condition is not well than child will not get all factors to score more.(Causations)