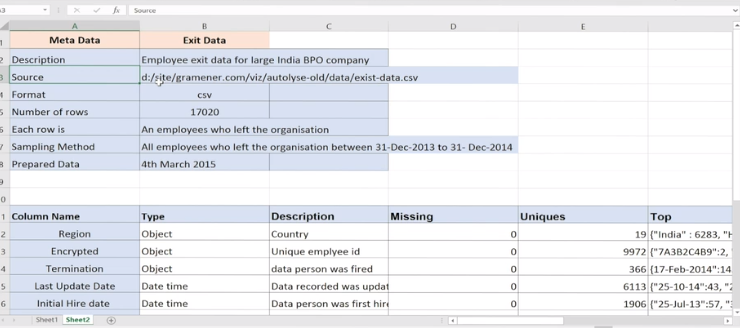
Data Description

Given a data set, the first step is to understand what it contains. Information about a data set can be gained simply by looking at its metadata. Metadata, in simple terms, is the data that describes the each variable in detail. Information such as the size of the data set, how and when the data set was created, what the rows and variables represent, etc. are captured in metadata.



**Types of Variables**

You learnt the difference between **ordered** and **unordered categorical variables**-

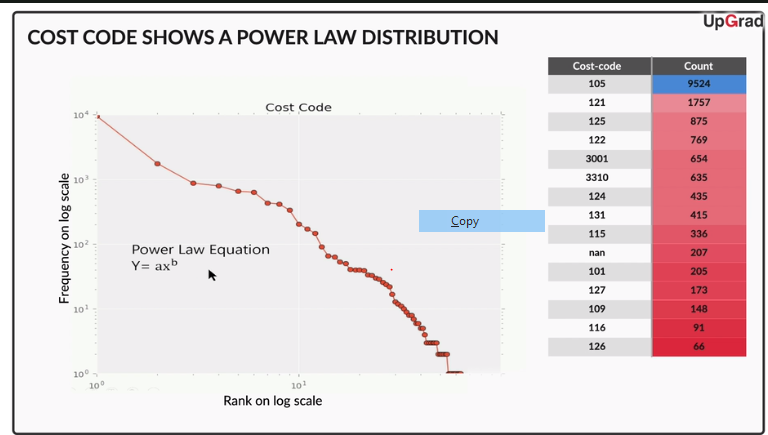
* **Ordered** ones have some kind of ordering. Some examples are
  + Salary = High-Medium-low
  + Month = Jan-Feb-Mar etc.
* **Unordered** ones do not have the notion of high-low, more-less etc. Example:
  + Type of loan taken by a person = home, personal, auto etc.
  + Organisation of a person = Sales, marketing, HR etc.

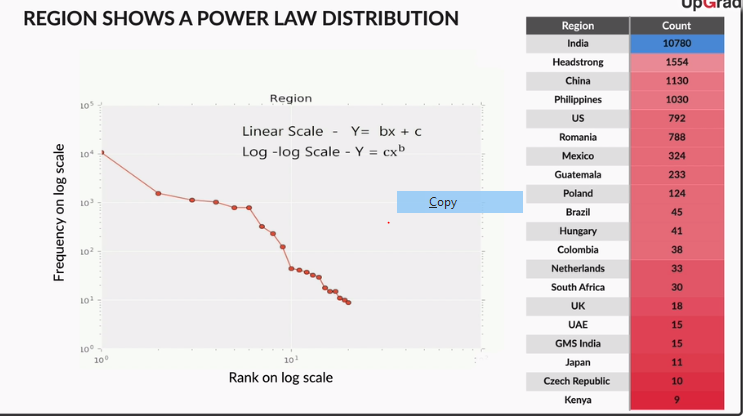
Apart from the two types of categorical variables, the other most common type is **quantitativevariables**. These are simply numeric variables which can be added up, multiplied, divided etc. For example, salary, number of bank accounts, runs scored by a batsman, the mileage of a car etc.

So far, we have discussed the following types of variables:

1. **Categorical variables**
   * Unordered
   * Ordered
2. Quantitative Numerical Variables.

**Unordered Categorical Variables - Univariate Analysis**





It is important to note that **rank-frequency plots** enable you to extract meaning even from seemingly trivial **unordered categorical variables** such as country, name of an artist, name of a github user etc.

**Why plotting on a log-log scale helps**

The objective of using a log scale is to make the plot readable by changing the scale. For example, the first ranked item had a frequency of 29000, the second-ranked had 3500, the seventh had 700 and most others had very low frequencies such as 100, 80, 21 etc.  The range of frequencies is too large to fit on the plot.

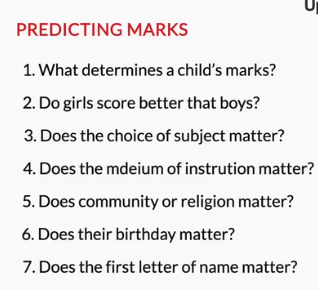
Plotting on a log scale compresses the values to a smaller scale which makes the plot easy to read.

This happens because log(x) is a much smaller number than x. For example, log(10) = 1, log(100) = 2, log(1000) = 3 and so on. Thus, log(29000) is now approx. 4.5, log(3500) is approx. 3.5 and so on. What was earlier varying from 29000 to 1 is now compressed between 4.5 and 0, making the values easier to read on a plot.

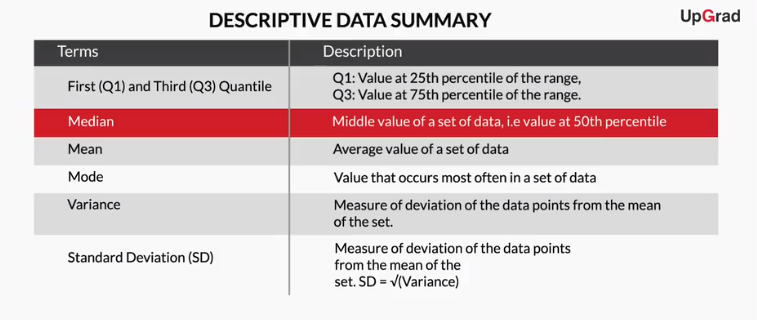
To summarise, the major takeaways from this lecture are:

* 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

whenever you have a continuous or an ordered categorical variable, make sure you plot a histogram or a bar chart and observe any unexpected trends in it.



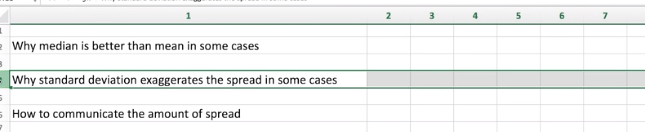
# Quantitative Variables - Univariate Analysis



Mean and median are single values that broadly give a representation of the entire data. As Anand stated very clearly, it is very important to understand when to use these metrics to avoid doing inaccurate analysis.

While mean gives an average of all the values, median gives a typical value that could be used to represent the entire group. As a simple rule of thumb, always question someone if someone uses the mean, since median is almost always a better measure of ‘representativeness’.

Let’s now look at some other summary descriptive statistics such as mode, interquartile distance, standard deviation, etc.



Summary:

Let's summarize what you learnt:

1. **Metadata description** describes the data in a structured way. You should make it a habit of creating a metadata description for whatever data set you are working on. Not only will it serve as a reference point for you, it will also help other people understand the data better and save time.
2. **Distribution plots** reveal interesting insights about the data. You can observe various visible patterns in the plots and try to understand how they came to be.
3. **Summary metrics** are used to obtain a quantitative summary of the data. Not all metrics can be used everywhere. Thus, it is important to understand the data and then choose what metric to use to summarise the data.

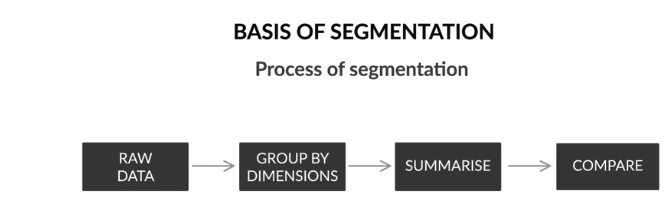
In the next session, you will learn segmented univariate analysis.

Segmented Univariate Analysis

Basis of Segmentation:

The entire segmentation process can be divided into four parts:

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



To summarize, the standard process of segmented univariate analysis is as follows:

* Take raw data
* Group by dimensions
* Summarize using a relevant metric like mean, median, etc.
* Compare the aggregated metric across groups/categories

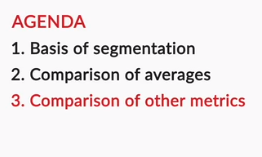
So with this, you have now performed segmented univariate analysis on a few variables, but what if you have a large number of variables in your data set? How would you go about analyzing and explaining the results of hundreds of categorical variables to your client? Let’s see what such a table would look like in the next lecture.

# Quick Way of Segmentation

One way of solving this problem is to make a table with the categorical variables on one axis and the numeric variables (or measures/facts) on the other.

# Comparison of Averages

“Don’t blindly believe in the averages of the buckets — you need to observe the distribution of each bucket closely and ask yourself if the difference in means is significant enough to draw a conclusion. If the difference in means is small, you may not be able to draw inferences. In such cases, a technique called hypothesis testing is used to ascertain whether the difference in means is significant or due to randomness.“ Don’t worry if you do not get the concept of hypothesis correctly, It will be dealt separately in hypothesis module.



In conclusion, the three steps of segmented univariate analysis are as follows:

* Basis of segmentation
* Comparison of averages
* Comparison of other metrics

Besides finding the segments and comparing the metrics, your primary focus should be on understanding the results arising from the segments.