FREE Online
Citizen Data Scientist Course

You Too Can be a Citizen Data Scientist – No Matter Your Role, Skill or Job Function

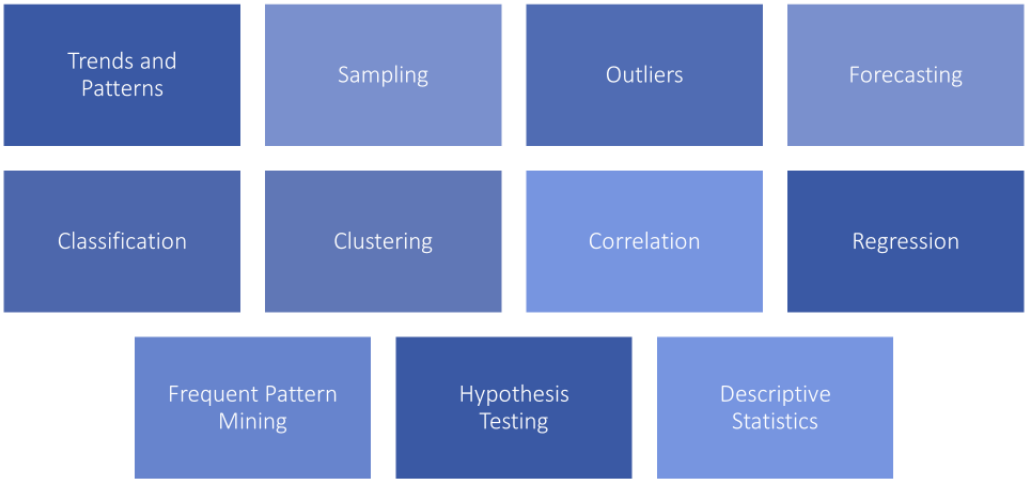
Section 7 - Analytical Techniques - Part 1

Instructor Notes: This supporting documentation includes a complete set of the slides used in the course material. Some sections also include expanded material, articles and documentation to further student understanding of more complex topics.

For your conenience, contact information is displayed at the end of the online course and at the conclusion of the supporting documentation in Section 12. We invite you to contact us with questions, requests or comments.

Section 7 - Analytical Techniques and Algorithms – Part 1

Analytical Techniques And Algorithms

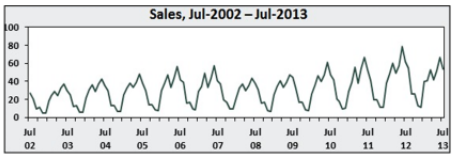


Trends And Patterns

- Trend

Seasonal

When fluctuations repeat over fixed periods of time and are therefore predictable and where those patterns do not extend beyond a one-year period.

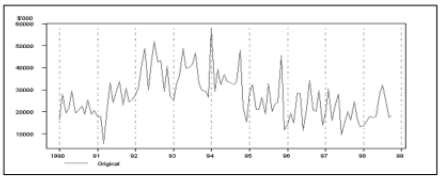


Cyclical

Cyclical patterns occur when fluctuations do not repeat over fixed periods of time and are therefore unpredictable and extend beyond a year.

Random

These fluctuations are short in duration, erratic in nature and follow no regularity in the occurrence pattern in prediction.



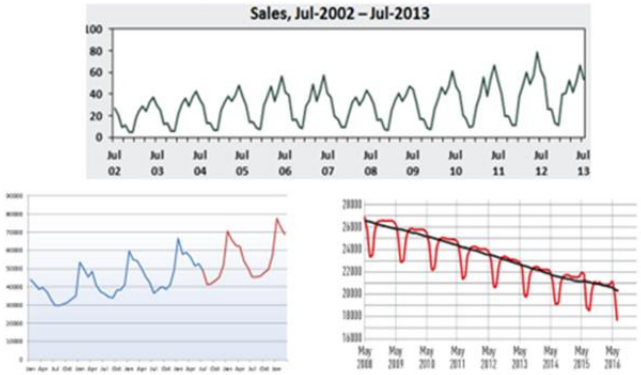
Trends And Patterns – Seasonal

Seasonal Pattern

- Seasonality occurs when the time series exhibits regular fluctuations during the same period every year.

Seasonal Pattern

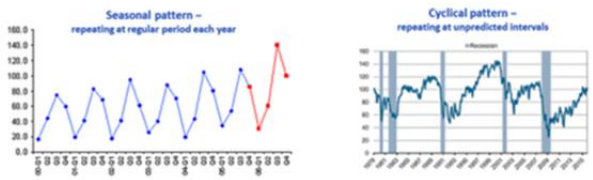
- Seasonality occurs when the time series exhibits regular fluctuations during the same period every year.



Trends and Patterns – Cyclical

Cyclical Pattern

- Definition : **"Patterns are called cyclical when fluctu**
  unpredictable and extend beyond a year."



As can be seen in the images above, seasonality is observed at regular period of time every year and hence is predictable while cyclical pattern is unpredictable and does not show up at regular period of time

Trends and Patterns - Random

Random Pattern - Example

- Irregular economic fluctuations result from unusual events, such as **floods, strikes, civil strife, large bankruptcies and terrorist incidents**. The impact of these fluctuations is usually limited to a certain industry or market. For example, **a flood may affect the distribution capability within a specific region**. Major natural disasters, **such as the 2011 Japanese earthquake, can affect the supply chains of several industries**.

Sampling

Popular Sampling Methods

Simple Random Sampling:

Here, the selection is purely based on a chance and every item has an equal chance of getting selected

Lottery system is an example of simple random sampling

Stratified Random Sampling:

Here, the population data is divided into subgroups known as Strata

The members in each of the subgroup formed have similar attributes and characteristics in terms of demographics, income, location etc.

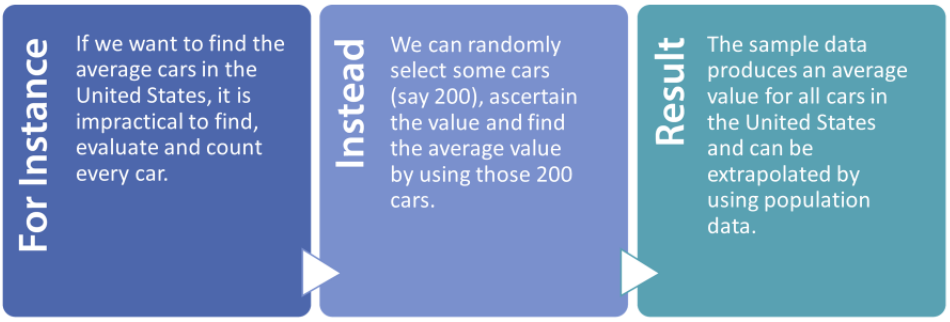
A random sample from each of these subgroups is taken in proportion to the subgroup size relative to the population size

These subsets of subgroups are then added to from a final stratified random sample

Higher statistical precision is achieved through this method due to low variability within each subgroup, also less sample size is required for this method of sampling when compared to simple random sampling

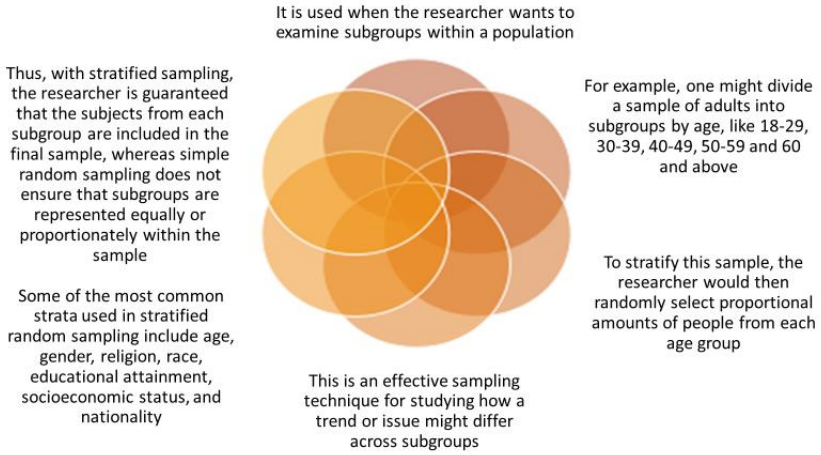
Sampling – Simple Random Sampling

Example : Simple Random Sampling



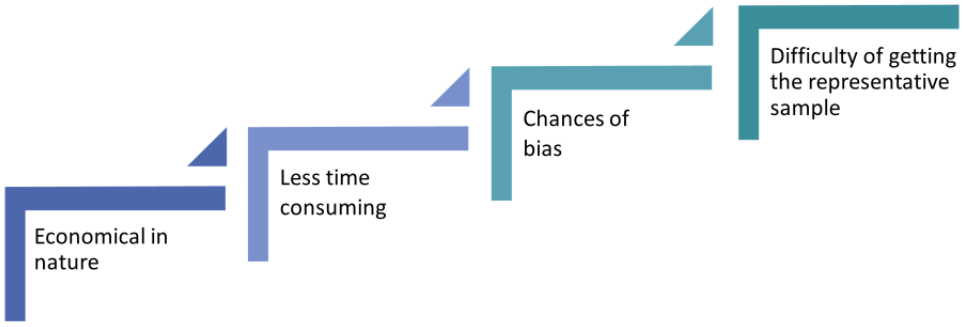
Sampling – Stratified Random Sampling

Example : Stratified Random Sampling



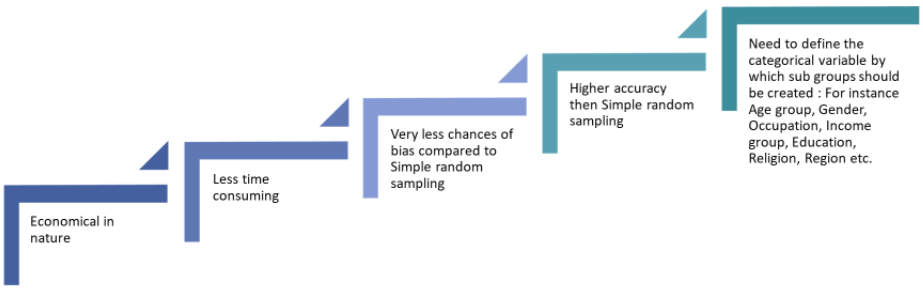
Sampling – Simple Random Sampling

Pros and Cons : Simple Random Sampling



Sampling – Stratified Random Sampling

Pros and Cons : Stratified Random Sampling

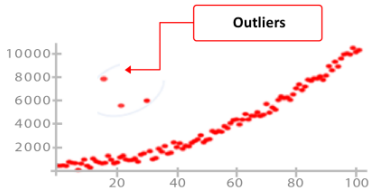


Outlier

What is an Outlier?

An outlier is an element of a data set that distinctly stands out from the rest of the data

In other words, they are the observations lying outside overall pattern of distribution as shown in the figure



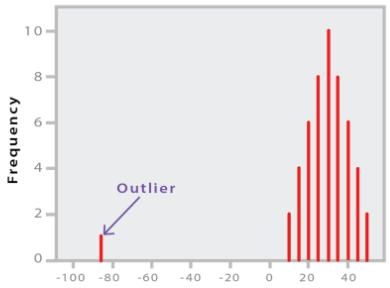
Outlier

Detecting Outliers

With Histogram

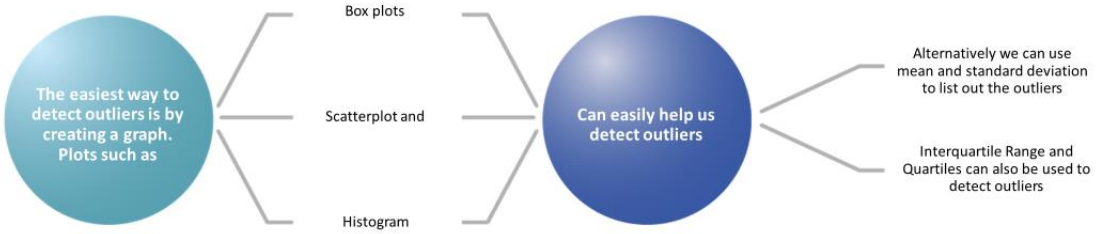
A univariate outlier is a data point that consists of an extreme value on one variable.

If you look at the Histogram, you can notice that there is one value that lies far to the left side of all the other data. This data point is an outlier.



Outlier

How to Detect Outliers



Outlier

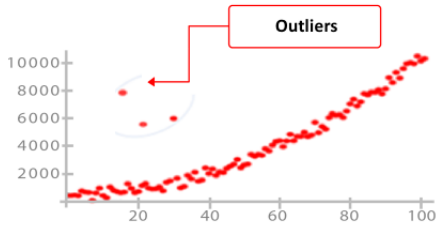
How to Handle Outliers

1. We can either remove the outliers altogether from the selected dataset or we can replace them by recommended statistical measure that is percentiles
2. It is general practice to replace lower and upper outliers with 5th and 95th percentile values respectively
3. But, in case of domains demanding high accuracy and no loss of data, we can use 1st and 99th percentile values to replace lower and upper outliers respectively
4. Thus, it is at sole discretion of an analyst which approach to select and apply

Outlier

What is an Outlier?

An outlier is an element of a data set that distinctly stands out from the rest of the data



In other words, they are the observations lying outside overall pattern of distribution as shown in the figure