A close-up of a chart

Description automatically generated

Statistical is dealing with quantitative data.

Statistical analysis is the application of statistical methods to a sample of data in order to develop an understanding of what that data represents.

* A sample is a representative selection drawn from a total population.
* Population is a discrete group of people or things that can be identified by at least one common characteristic for purpose of data collection and analysis.
* Statistical methods are mainly used to ensure that data is interpreted correctly and apparent relationships are meaningful.

2 Types of statistics :

1. Descriptive statistics : to summarize information about the sample

* Objectives is to make it easier to understand and visualize raw data without making conclusions regarding any hypotheses that were made.
* Enables you to present data in a meaningful way
* Allowing simpler interpretation of the data. Data is described using summary charts, table and graphs.
* Ex : English test score in a specific class of 25 students. Calculate the summary statistics and produce a graph.
* Common measures of descriptive analysis :
* Central Tendency , locating the center of data sample ex. Mean, median and mode.
* Dispersion, measure of variability in dataset ex. Variance(how far away data points from the center-distribution of values, lower variability means consistent values in a dataset), Standard deviation(How tightly your data is clustered around the mean), Range (distance between the smallest and largest values in your datasets).
* Skewness, measure of whether the distribution of values is symmetrical around a central value or skewed left/right. Ex. Correlation and scatterplots to assess the relationship between paired data.

1. Inferential statistics : to make inferences or generalization about the broader population.

* Takes data from a sample to make inferences about the larger population from which the sample was drawn.
* Help draw generalization that apply the results of the sample to the population as a whole.
* Common measures of inferential analysis :
  + Hypothesis Testing, for studying the effectiveness of a vaccine by comparing outcomes in a control group
  + Confidence Intervals, incorporate the uncertainty and sample error to create a range of values the actual population value is likely to fall within.
  + Regression analysis, incorporate hypothesis test that help determine whether the relationships observed in the sample data actually exist in the population rather than just the sample.

SOFTWARE to perform statistical data analysis :

1. Statistical analysis system (SAS)
2. Statistical Package for the social sciences (SPSS)
3. Stat Soft.

Statistics combined with Data Mining together help in better decision-making by providing measures and methodologies and also identifying patterns that help differences between random noise and significant findings.

WHAT is data Mining…

**DATA MINING** is the process of extracting knowledge from data. It involves the use of pattern recognition, statistical analysis and mathematical techniques. The goal is to identify correlations in data, find patterns and variations, understand trends and predict probabilities.

* Pattern recognition is the discovery of regularity’s or commonalities in data. Example analysis of user behavior using user log data.
* A trend , is the general tendency of a set of data to change over time. Season time temperature a trend.

Data mining applications example is :

* profiling customer behaviors , needs and disposable income in order to offer targeted campaigns.
* Financial institutions tracking customer transactions for unusual behaviors and flagging fraudulent transactions
* Helping investigation agencies deploy police force where the likelihood of crime is higher
* Aligning supply and logistic with demand forecasts.

Data Mining techniques :

1. Classification – classifying attributes into target categories.
2. Clustering – involves grouping data into clusters ,so they can be treated as groups.
3. Anomaly or outlier detection – Finding patterns in data that are not normal or unexpected.
4. Association rule mining – Establishing a relationship between two data events (Buy coffee and buy milk)
5. Sequential patterns – Tracing a series of events that take place in a sequence (traces customer shopping trail from the time they log in an online store to the time they log out)
6. Affinity grouping – Discovering co-occurrence in relationships (cross selling and up selling their products by recommending products to people based on the purchase history of other people who purchased the same item)
7. Decision trees – building classifications model in the form of a tree structure with multiple branches.
8. Regression – Identifying the relationship between two variables (causal or correlational)

Tools for data Mining :

1. Spreadsheet – creating pivot table to showcase specific aspects of your data.
2. R
3. Python
4. IBM SPSS – Requires a license for use.
5. IBM Watson Studio – available through a web browser on the public cloud/private cloud and as a desktop app.
6. SAS – graphical user interface for non-technical users.

DATA ANALYSIS PROCESS, start with understanding the problem and it ends with communicating the findings in ways that impact decision making. As a data analysts need to tell the story with your data by visualizing the insights clearly and creating a structured narrative explicitly targeted at your audience.

A diagram of visual data

Description automatically generated

Understanding the information needs of your audience will help you decide what, and how much information is essential to enable a better understanding of your findings.

1. Begin your presentation by demonstrating your understanding of the business problem to your audience.
2. The next step is designing your communication is to structure and organize your presentation for maximum impact (Reference your data, state your assumptions, identify the best formats for presenting – summary, fact sheet or report )
3. A thousand word essay will not have the same impact as a visual in creating a clear mental image in the minds of your audience.
4. A powerful visualization(Graph, Chart, Diagram) tells a story through the graphical depiction of facts and figures.
5. Present data within a narrative and support the narrative with visual to built TRUST, UNDERSTANDING and Relate ability.

Data Visualization is the discipline of communicating information through the use of visual elements such as graphs, charts and maps. The goals is to make information easy to comprehend, interpret and retain. Common types of graphs :

* Bar Charts - Great for comparing related data sets or parts of a whole. Ex. Number of populations in the world.
* Column Charts - Compare values side-by-side, to show change over time. Ex. Showing how page views and user session time on website is changing month to month basis (over time).
* Pie Charts - show the breakdown of an entity into its sub parts and the proportion of the sub parts in relation to one another. Each portion of the pie represent a static value or category and the sum of all categories is equal to hundred percent.
* Line Charts - Display trends. They are great for showing how a data value is changing in a relation to a continuous variable. Can be used for understanding trends, pattern and variations in data. Ex, how the sales of the multiproducts change over time .
* **DASHBOARDS -** organize and display reports and visualizations coming from multiple data sources into a single graphical interface. Dashboard use to monitor daily progress or the overall health of a business function or even a specific process.

**Data Visualization techniques :**

1. Contextualize data

* Involves using benchmarks and trends. Ex. Adding an industry benchmark provide valuable context when presenting sales data. This helps the audience understand performance relative to industry standards.
* Creating an emotional connection by making data relatable. Ex when showcasing customer satisfaction data, use smiling face for satisfied customers and vice versa.
* Highlighting anomalies ex. Website visitors anomalies put a noted when there is a outage or damage on our problems graph.

1. Present data effectively

* Use graphs, maps charts, or diagrams to represent your data. Read again COMMONT Types of graphs.
* Providing a tabular view in addition to the visual representation.

1. Improve the clarity of visualization

* Conveying only one central point through each chart. Ex. Don’t put revenue and expense in the same chart.
* Using labels also helps bring clarity to your visualization. Ex. Barchart with LABELS.
* Using annotations , to an exceptionally graphs with a note explaining that it corresponds to.
* Using legends are necessity in charts. Ex. Moreover same with LABELS
* Tooltips are helpful hints that appear when you need them, ex. Showing the numerical value of datapoint when you hover your mouse pointer over it.

1. Ensure consistency and visual appeal

* Using pictures in addition to regular charts or graphs. Ex. Weather chart can use cloud, sun, rainy of the data.
* Using consistent color to bring in consistency, can easily associate that specific color with specific purposes.
* Using consistent fonts and layouts.

1. Engage your audience

* Using interactivity through filters and slicers
* Transparency and credibility to see the actual data,

The key elements of a data findings report (whether you want it to be a brief report of five or fewer pages, or a longer document running more than 100 pages. The structure of the report depends on the length of the document. A brief report is more to the point and summarizes key findings. A detailed report incrementally builds the argument and contains details about other relevant works, data analysis methodology, data sources, intermediate findings, and the main results) are .:

* Cover page
  + will have the title of your report, your name, and then the date.
* Executive summary
  + will briefly explain the details of the project and should be considered a stand-alone section. You will take this information from the main points of your report.
  + Executive summary typically consists of about three paragraphs. However, the executive summary could be longer for larger documents running to a few hundred pages.
* Table of contents
  + will contain a list of the sections and subsections of your report to give your audience an overview of the contents. This section also enables the audience to go directly to a specific section that may be more important to them.
* Introduction
  + explains the nature of the analysis, states the problem, and lists the questions that were to be answered by performing the analysis.
  + The introductory section helps set the context for the audience who might be new to the topic and need to be gently introduced to the subject matter before being immersed in intricate details.
* Methodology
  + explains the data sources used in the analysis and outlines the plan for the collected data. For example, did the analysis involve extracting data from various sources such as website analytics?
  + If you have collected new data, explain the data collection exercise in some detail. You will refer to your choice of variables, data, and methods and how they helped you answer the questions.
* Results
  + would contain charts and graphs that would substantiate the results and highlight the more complex or crucial findings. By providing this interpretation of data, you can give a detailed explanation to the audience and convey how it relates to the problem stated in the introduction.
* Discussion
  + The results section is followed by the discussion section, where you craft your main arguments by building on the results you have presented earlier.
  + This is the section where you rely on the power of narrative to enable numbers to communicate your findings and implications.
  + You refer the audience back to the questions that were to be answered through this study.
* Conclusion
  + This final section should reiterate the problem given in the introduction and provide an overall summary of the findings. It should also state the outcome of the analysis and if any other steps would be taken in the future.
* Appendix
  + would contain information that did not fit into the main body of the report, but you believed it was still important enough to include. This type of information could include resources or references, such as original data.

https://author-ide.skills.network/render?token=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJtZF9pbnN0cnVjdGlvbnNfdXJsIjoiaHR0cHM6Ly9jZi1jb3Vyc2VzLWRhdGEuczMudXMuY2xvdWQtb2JqZWN0LXN0b3JhZ2UuYXBwZG9tYWluLmNsb3VkL0lCTVNraWxsc05ldHdvcmstQkkwMTAxRU4tQ291cnNlcmEvbGFicy9Nb2R1bGU1UmVhZGluZ2VmZmVjdGl2ZV9jb21tdW5pY2F0aW9uX29mX2JpX2luc2lnaHRzLm1kIiwidG9vbF90eXBlIjoiaW5zdHJ1Y3Rpb25hbC1sYWIiLCJhZG1pbiI6ZmFsc2UsImlhdCI6MTcxMTQyNjA1Mn0.vk4KekhsSAFLFTBN74QKvJ6ftxdFc8\_8U3WbIO579VM

* Some of the best practices for accurately conveying your message include:
  + Making sure that charts and graphs are not too small and are clearly labeled, You can test the visualizations by sitting at different distances, similar to your audience. If you cannot see the data clearly, consider redesigning.
  + Using data only as supporting evidence,
  + Sharing only one point from each chart or graph, if you are supplying too much information, it can be confusing
  + Eliminating data that does not support the key message, you will keep the presentation clear and concise by eliminating this unnecessary data and highlighting only data points that support your key ideas.