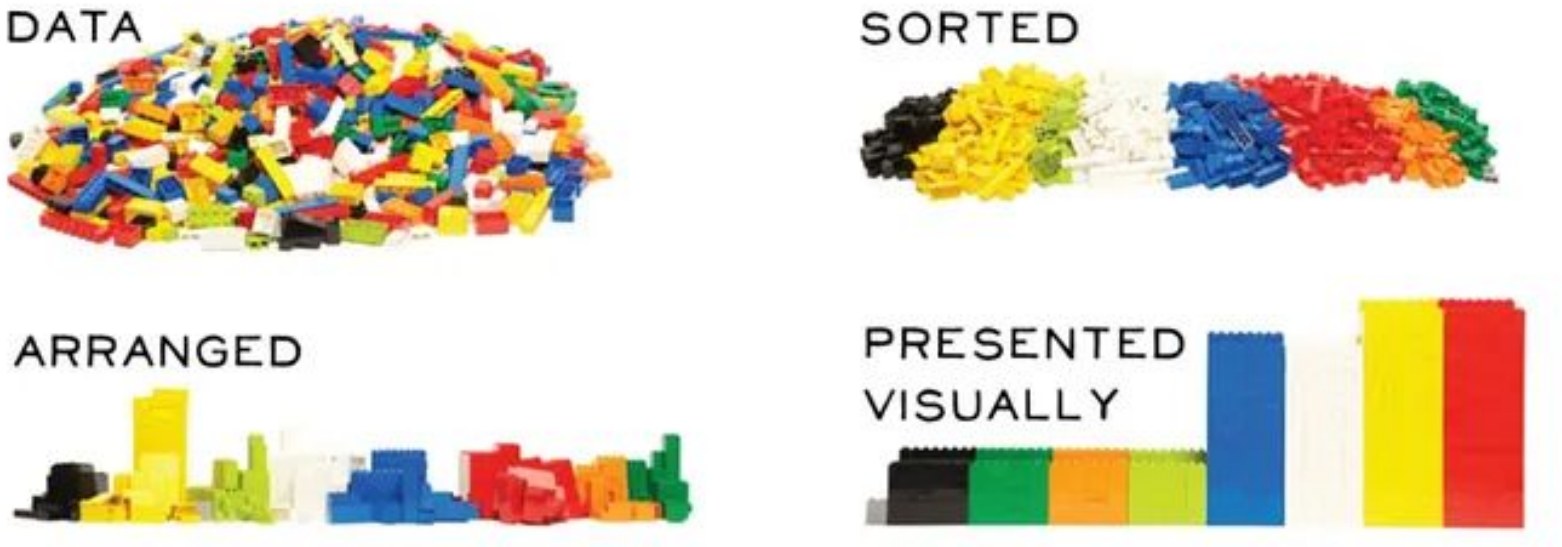
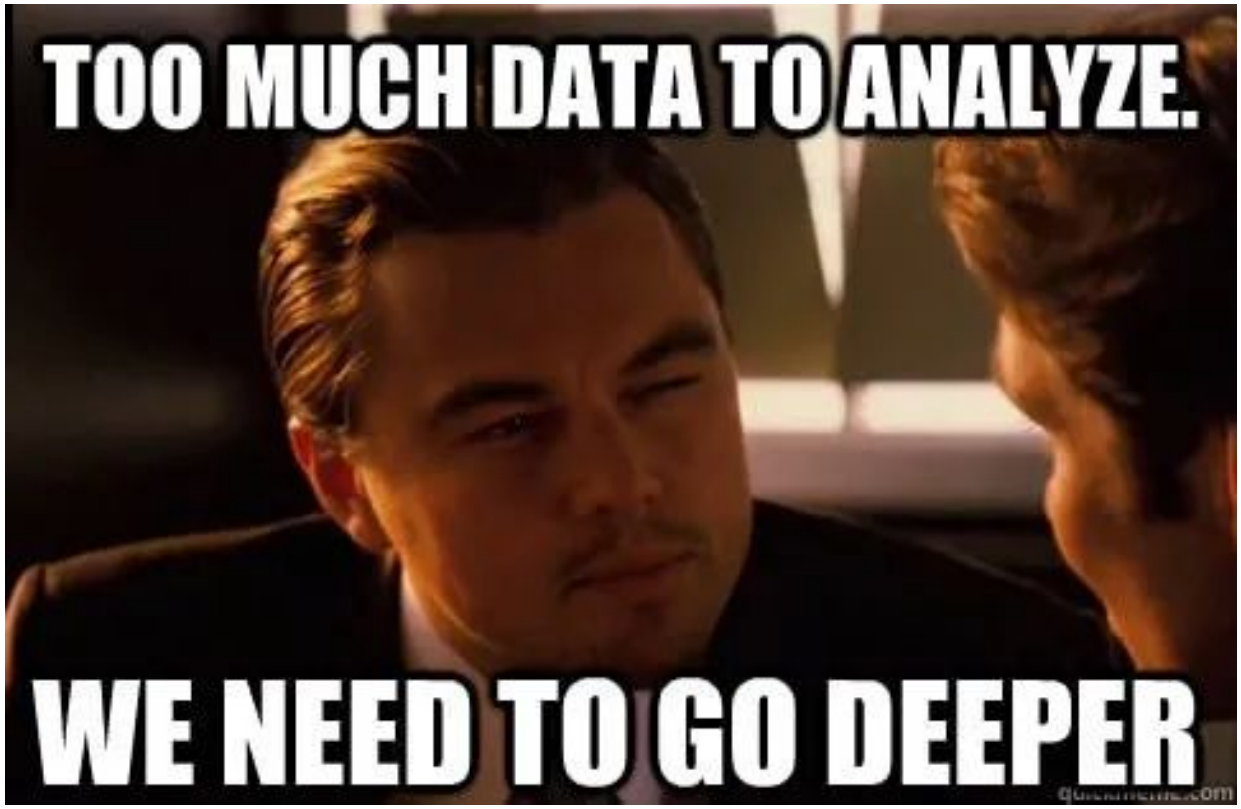
ITDO6014
AI AND DS-1

Module 5: Exploratory Data Analysis

Exploratory Data Analysis



Exploratory Data Analysis



Exploratory Data Analysis

- “Torture the data, and it will confess to anything.”
- — Ronald Coase
- With proper use of data one could rule the entire world as well!
- But, raw, unprocessed data isn’t of much use unless you derive insights from it.
- Exploratory Data Analysis (EDA) is the process of visualizing and analyzing data to extract insights from it. In other words, EDA is the process of summarizing important characteristics of data in order to gain better understanding of the dataset.

Exploratory Data Analysis

- With EDA, you can find anomalies in your data, such as outliers or unusual observations, uncover patterns, understand potential relationships among variables, and generate interesting questions or hypotheses that you can test later using more formal statistical methods.
- Exploratory data analysis is like detective work: you're searching for clues and insights that can lead to the identification of potential root causes of the problem you are trying to solve. You explore one variable at a time, then two variables at a time, and then many variables at a time
- EDA is generally classified into two methods, non-graphical or graphical. And each method can be applied to one variable/column (univariate) or a combination of variables/columns(bivariate).

Exploratory Data Analysis

- Typical Data Formats:

https://www.weirdgeek.com/2018/12/common-file-formats-used-in-data-science/

- CSV
- Text Files
- JSON
- Microsoft Excel File
- SAS
- SQL
- Python Pickle File
- Stata
- HDF5
- HTML
- ZIP
- PDF
- DOCX
- Images
- Google Bigquery

Exploratory Data Analysis

- **Objective of Exploratory Data Analysis:**
- Identifying and removing data outliers
- Identifying trends in time and space
- Uncover patterns related to the target
- Creating hypotheses and testing them through experiments
- Identifying new sources of data

Exploratory Data Analysis

- **Steps Involved in Exploratory Data Analysis (EDA)**
- **1. Data Collection**
- **2. Finding all Variables and Understanding Them**
- **3. Cleaning the Dataset**
- **4. Identify Correlated Variables**
- **5. Choosing the Right Statistical Methods**
- **6. Visualizing and Analyzing Results**

Types of Exploratory Data Analysis

- **1. Univariate Non-Graphical**
- It is the simplest of all types of data analysis used in practice. As the name suggests, uni means only one variable is considered whose data (referred to as population) is compiled and studied. The main aim of univariate non-graphical EDA is to find out the details about the distribution of the population data and to know some specific parameters of statistics. The significant parameters which are estimated from a distribution point of view are as follows:

Types of Exploratory Data Analysis

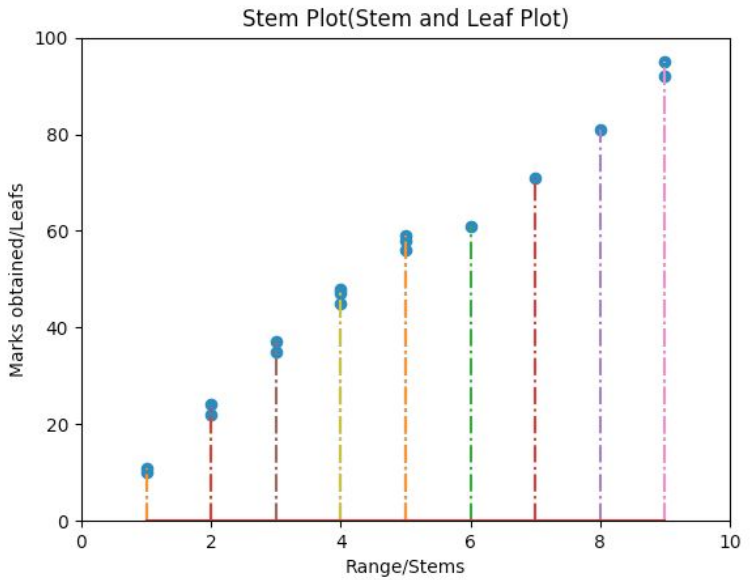
- **Central Tendency:** This term refers to values located at the data's central position or middle zone. The three generally estimated parameters of central tendency are mean, median, and mode. Mean is the average of all values in data, while the mode is the value that occurs the maximum number of times. The Median is the middle value with equal observations to its left and right.

Types of Exploratory Data Analysis

- **Range:** The range is the difference between the maximum and minimum value in the data, thus indicating how much the data is away from the central value on the higher and lower side.
- **Variance and Standard Deviation:** Two more useful parameters are standard deviation and variance. Variance is a measure of dispersion that indicates the spread of all data points in a data set. It is the measure of dispersion mostly used and is the mean squared difference between each data point and mean, while standard deviation is the square root value of it. The larger the value of standard deviation, the farther the spread of data, while a low value indicates more values clustering near the mean.

Types of Exploratory Data Analysis

- **2. Univariate Graphical:**
- **Stem-and-leaf Plots:** This is a very simple but powerful EDA method used to display quantitative data but in a shortened format. It displays the values in the data set, keeping each observation intact but separating them as stem (the leading digits) and remaining or trailing digits as leaves. But histogram is mostly used in its place now.



Types of Exploratory Data Analysis

- **Histograms (Bar Charts):** These plots are used to display both grouped or ungrouped data. On the x-axis, values of variables are plotted, while on the y-axis are the number of observations or frequencies. Histograms are very simple to quickly understand your data, which tell about values of data like central tendency, dispersion, outliers, etc. The simplest fundamental graph is a histogram, which is a bar plot with each bar representing the frequency, i.e., the count or proportion (the ratio of count to the total count of occurrences) for various values.

Types of Exploratory Data Analysis

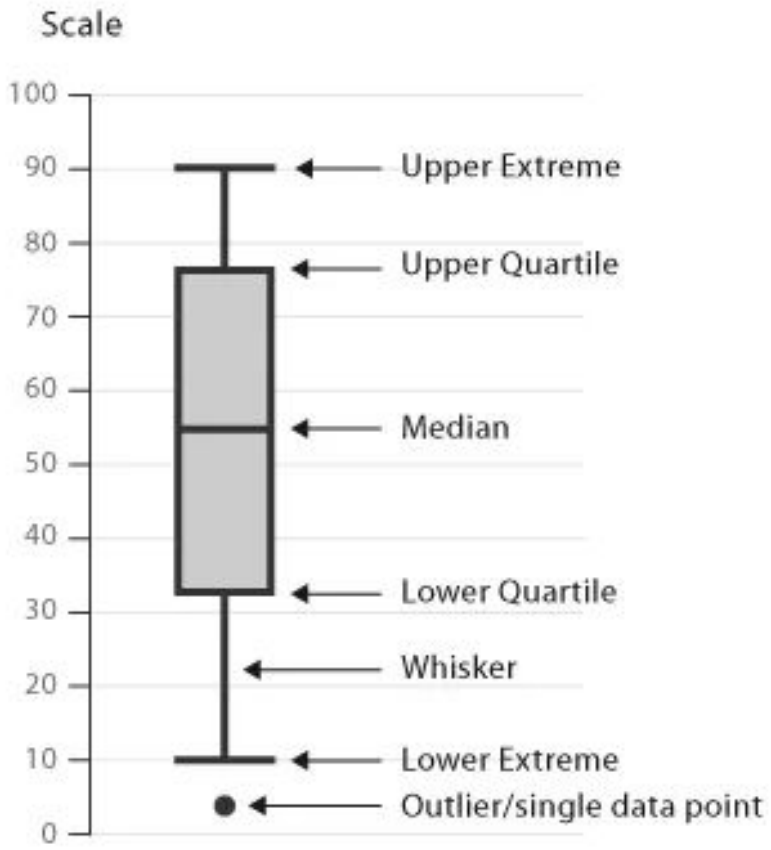
- There are many types of histograms, a few of which are listed below:
- **Simple Bar Charts:** These are used to represent categorical variables with rectangular bars, where the different lengths correspond to the values of the variables.
- **Multiple or Grouped charts:** Grouped bar charts are bar charts representing multiple sets of data items for comparison where a single color is used to denote one specific series in the dataset.

Types of Exploratory Data Analysis

- **Percentage Bar Charts:** These are bar graphs that depict the data in the form of percentages for each observation. The following image shows a percentage bar chart with dummy values.

Types of Exploratory Data Analysis

- **Box plots**
- Box plot shows us the median of the data, which represents where the middle data point is. The upper and lower quartiles represent the 75 and 25 percentile of the data respectively. The upper and lower extremes shows us the extreme ends of the distribution of our data. Finally, it also represents outliers, which occur outside the upper and lower extremes.



Types of Exploratory Data Analysis

- **3. Multivariate Non-Graphical**
- The multivariate non-graphical exploratory data analysis technique is usually used to show the connection between two or more variables with the help of either cross-tabulation or statistics.
- For categorical data, an extension of tabulation called cross-tabulation is extremely useful. For two variables, cross-tabulation is preferred by making a two-way table with column headings that match the amount of one variable and row headings that match the amount of the opposite two variables, then filling the counts with all subjects that share an equivalent pair of levels.

Types of Exploratory Data Analysis

- **4. Multivariate Graphical:**
- Graphics are used in multivariate graphical data to show the relationships between two or more variables. Here the outcome depends on more than two variables, while the change-causing variables can also be multiple.
- Some common types of multivariate graphics include:
- **Scatter Plot**
- The essential graphical EDA technique for two quantitative variables is the scatter plot, so one variable appears on the x-axis and the other on the y-axis and, therefore, the point for every case in your dataset. This can be used for bivariate analysis.

Types of Exploratory Data Analysis

- **D) Bubble Chart**
- Bubble charts scatter plots that display multiple circles (bubbles) in a two-dimensional plot. These are used to assess the relationships between three or more numeric variables. In a bubble chart, every single dot corresponds to one data point, and the values of the variables for each point are indicated by different positions such as horizontal, vertical, dot size, and dot colors.
- **E) Heat Map**
- A heat map is a colored graphical representation of multivariate data structured as a matrix of columns and rows. The heat map transforms the correlation matrix into color coding and represents these coefficients to visualize the strength of correlation among variables. It assists in finding the best features suitable for building accurate Machine Learning models.

Types of Exploratory Data Analysis

- Extremely useful external resources:
- https://www.cuemath.com/data/types-of-statistics/
- https://www.cuemath.com/data/descriptive-statistics/
- https://towardsdatascience.com/mean-median-mode-which-central-tendency-measure-to-use-when-9fb3ebbe3006
- https://www.analyticsvidhya.com/blog/2021/04/3-central-tendency-measures-mean-mode-median/
- https://www.cuemath.com/data/measures-of-dispersion/
- https://www.cuemath.com/data/measures-of-central-tendency/
- https://towardsdatascience.com/statistics-02-measuring-and-visualizing-the-spread-of-data-2fc31d928830
- https://medium.com/swlh/the-art-of-exploratory-data-analysis-eda-94a24320d3bd

Correlation and Covariance

- **Correlation:**
- https://medium.com/analytics-vidhya/correlation-and-machine-learning-fee0ffc5faac
- https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/types-of-correlation.html
- https://www.datacamp.com/tutorial/tutorial-datails-on-correlation
-

Correlation and Covariance

- **Covariance, variance vs. covariance vs. correlation:**
- **https://www.turing.com/kb/covariance-vs-correlation**
- **https://www.simplilearn.com/covariance-vs-correlation-article#:~:text=Covariance%20is%20an%20indicator%20of,strongly%20two%20variables%20are%20related.&text=The%20value%20of%20covariance%20lies,of%20%2D%E2%88%9E%20and%20%2B%E2%88%9E.**
- **https://towardsdatascience.com/statistics-in-python-understanding-variance-covariance-and-correlation-4729b528db01**

Degree of Freedom

- https://www.geeksforgeeks.org/degrees-of-freedom-formula/ **(very important)**
- https://medium.com/analytics-vidhya/an-introduction-of-degrees-of-freedom-in-machine-learning-and-statistics-8453d765d95e

Statistical Methods for Evaluation

- https://www.wallstreetmojo.com/statistical-analysis/
- https://www.indeed.com/career-advice/career-development/types-of-statistical-analysis
- https://www.simplilearn.com/what-is-statistical-analysis-article
- https://www.wallstreetmojo.com/hypothesis-testing/
- https://www.questionpro.com/blog/types-of-sampling-for-social-research/
- https://www.questionpro.com/blog/determining-sample-size/

ANOVA Test

- https://www.cuemath.com/anova-formula/
- https://www.statisticshowto.com/tables/f-table/