Exploratory Data Analysis (EDA) & Data Visualization

Introduction to EDA

- Definition of EDA:-Understanding the data before making any decisions
- Importance in Data Science:-Detecting Missing value,Finds Outliers,Improves Data Quality for Better Models,Saves Time and Prevents Costly Mistakes

cont…..

- Goals of EDA: Understanding Data, Detecting Anomalies

Volume of Data

- Small Data
  Ex:- excel sheet,CSV,JSON,TEXT(Kaggle)
- Medium Data
  Ex:- MySQL, PostgreSQL, or NoSQL(HDFC,SBI,BOI,UNION BANK)
- Big Data
- Ex. Hadoop, Apache Spark, and AWS

variables required for the analysis(Types of Variable)

- **Dependent Variable (Target Variable)**

Def:- Where the output is available is called dependent Variable

Ex.House Price Prediction

cont...........

- **Independent Variables (Predictor Variables)**
- Def:- Where the Input is available is called dependent Variable

Ex.**Marketing Spend , Ad Clicks**

open and paid data sources

- Free Data Sources

  Ex. Kaggle.com, NASA Open Data

- Paid Data Sources(Premium)

  Ex. IBM MarketScan for Health Industry data

Describe the metadata

- **Def**:-Data About data is called as metadata
- **Data Identification**

  :- **what the dataset contains**.

- **Data Quality**

  :-Missing values, data format**. Data Processing & Automation**

- **Data Processing & Automation**

  :- ETL pipelines for understand data

Data Validation: Tools and Processes

- Data validation:- the process of ensuring that data is **accurate, complete, and consistent** before it is used for analysis
- Tools:- EDA Technique

Role of EDA in Data Science

- Data Understanding

  Ex. EDA
- Feature Engineering

  Ex. PCA
- Model Preparation

  Ex. ML Algorithm

Types of Data

- Structured vs. Unstructured Data

  Ex.

  Structured:-Tabular data

  Unstructured:- Image,Video,Audio

cont……….

- Numerical vs. Categorical Data

- Numerical:-

:-**Continuous:** Height (170.5 cm), Temperature (36.8°C)

:-**Discrete:** Number of employees (50), Number of orders (120)

Categorical Data

- **Nominal:** Eye color (Blue, Brown, Green), Car brand (Toyota, Honda, BMW)

- **Ordinal:** Shirt size (S, M, L, XL), Exam Grades (A, B, C, D, F)

Data Cleaning in EDA

- Handling Missing Values
- Removing Duplicates
- Dealing with Outliers

Handling Missing Data

- Mean, Median, Mode Imputation
- Dropping Missing Values

Outlier Detection Techniques

- Boxplot Method
- Z score Method
- IQR Method

Summary Statistics

- Mean, Median, Mode
- Variance

Measuring Data Distribution

- Skewness & Kurtosis
- Normal Distribution

Introduction to Data Visualization

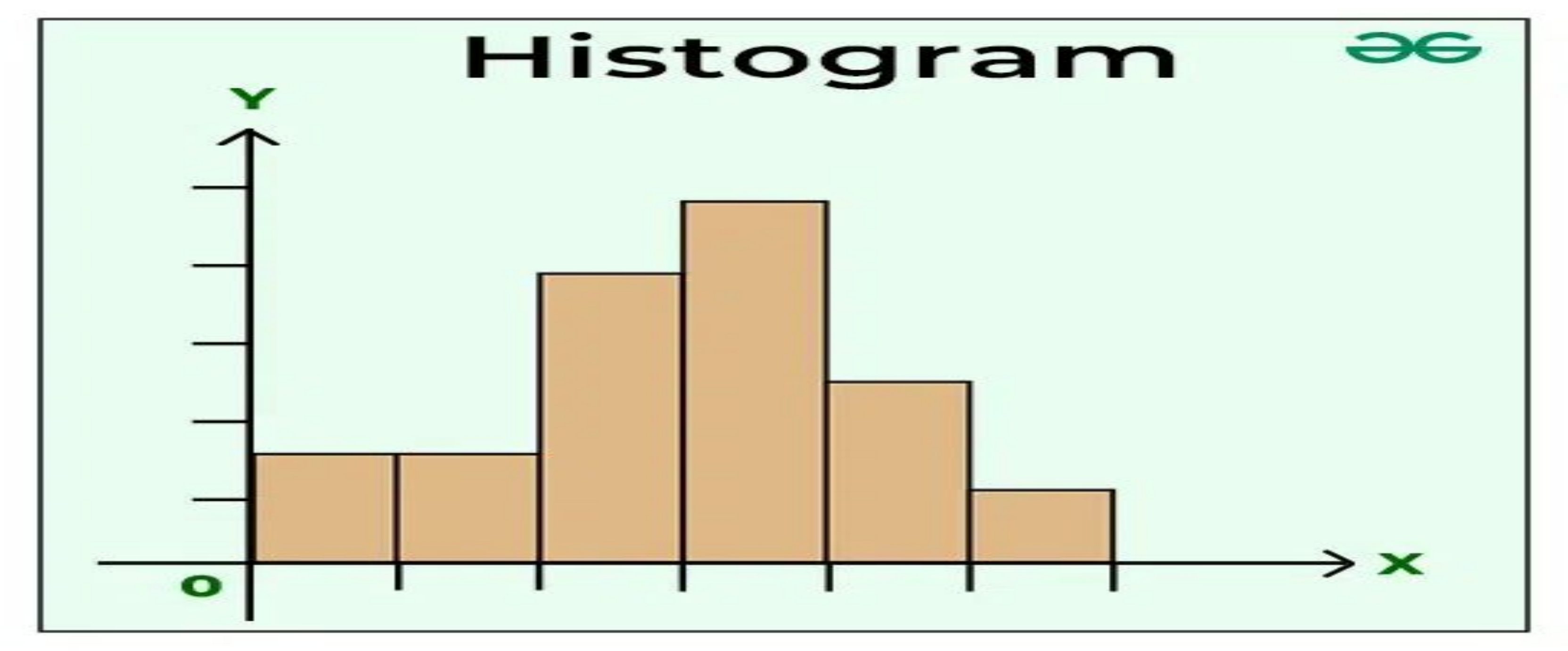
- Importance of Data Visualization
- Storytelling with Data

Types of Data Visualization

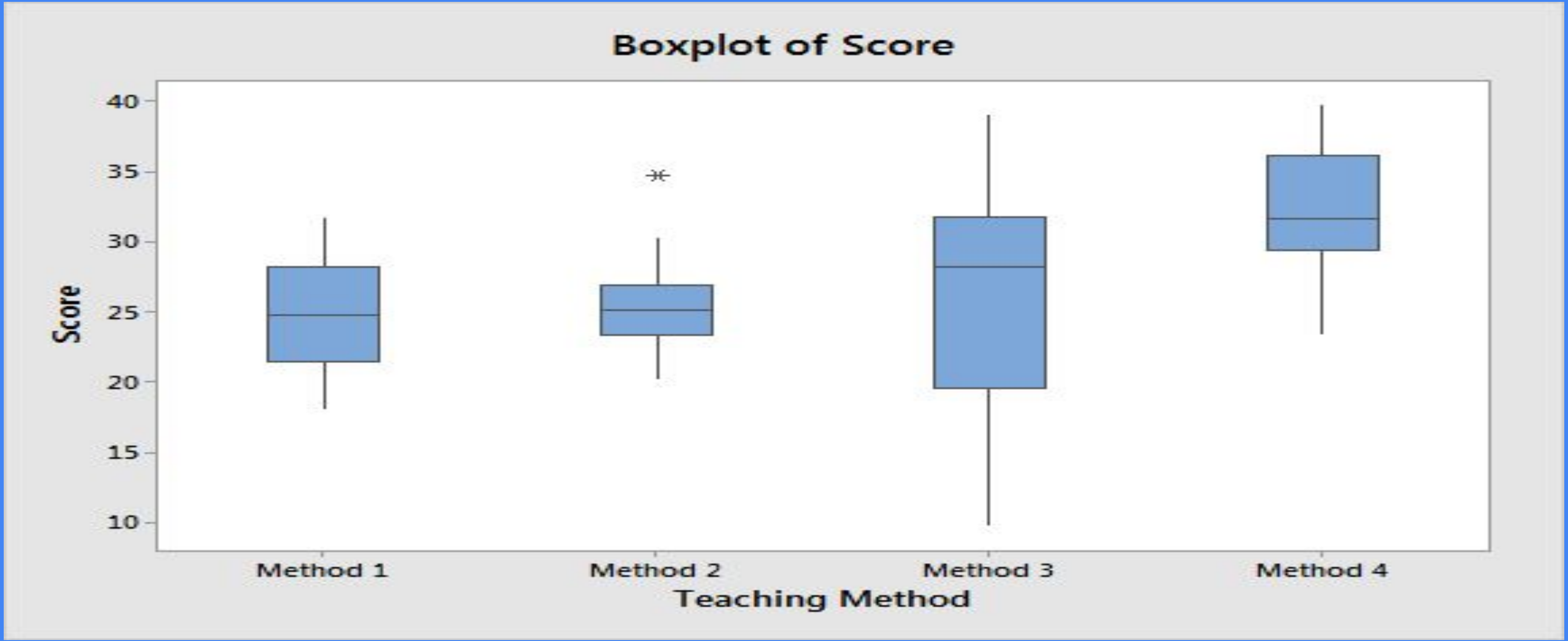
- Univariate, Bivariate, Multivariate Analysis

Univariate Analysis

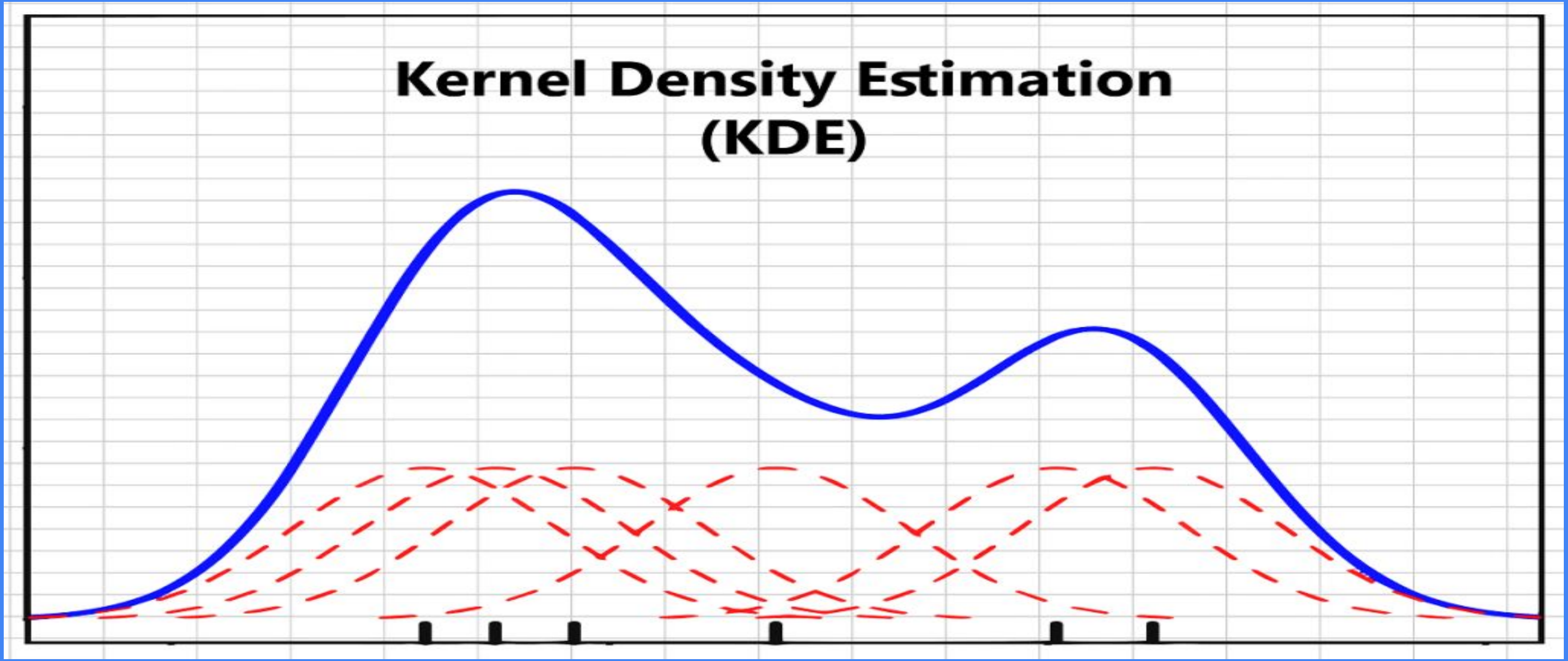
- Histograms



- Box plots

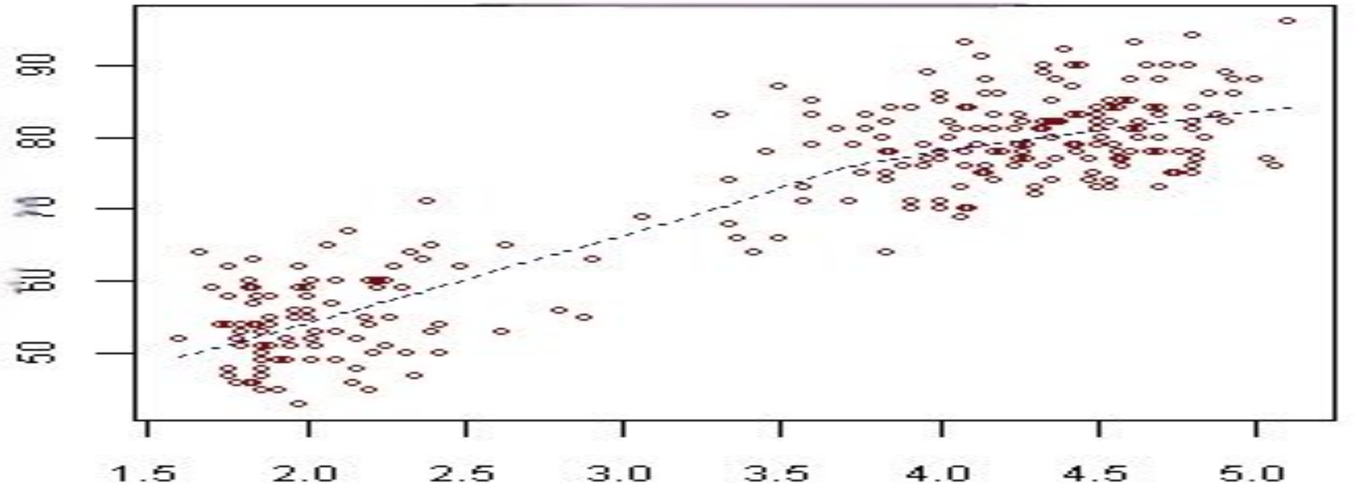

Boxplot of Score
40
35
30
25
20
15
10
Score
Method 1
Method 2
Method 3
Method 4
Teaching Method

- KDE Plots



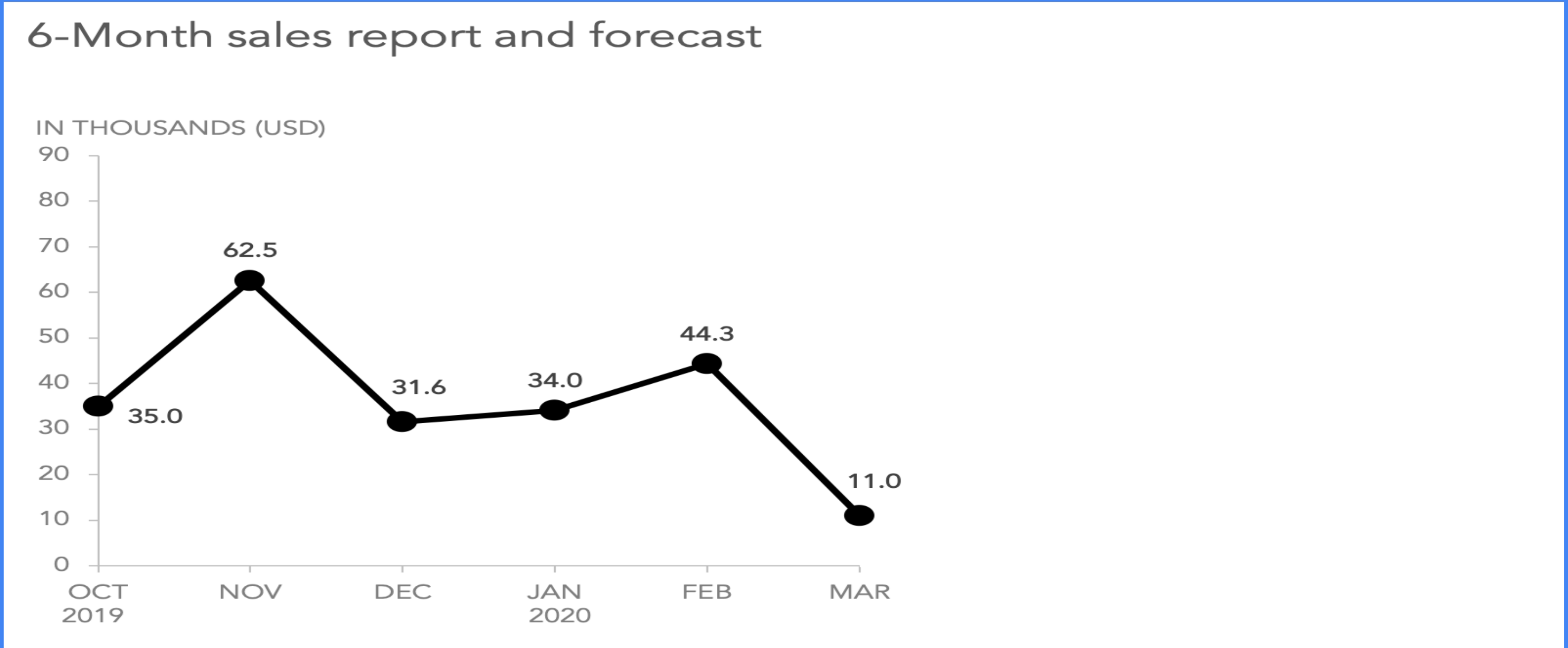
Bivariate Analysis

- Scatter Plots



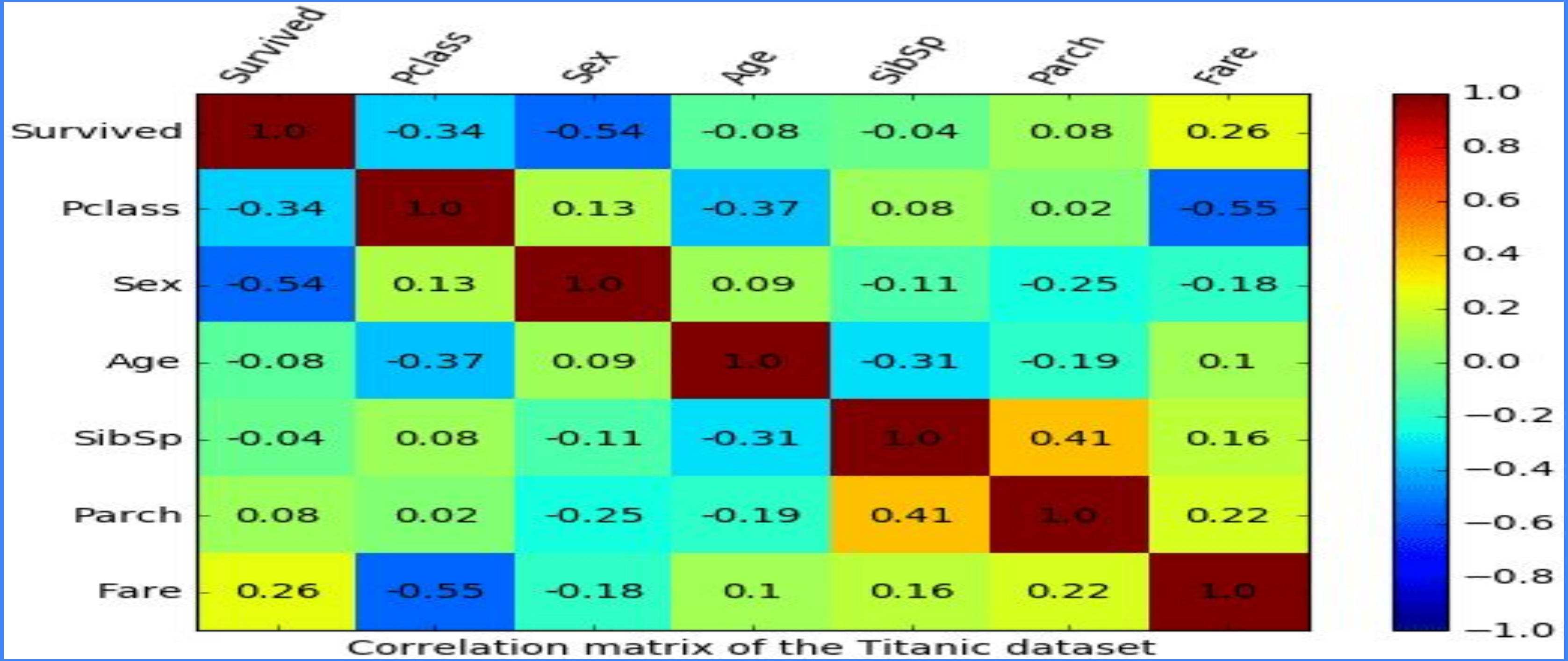
cont……

- Line Graphs



cont....

- Correlation Matrix

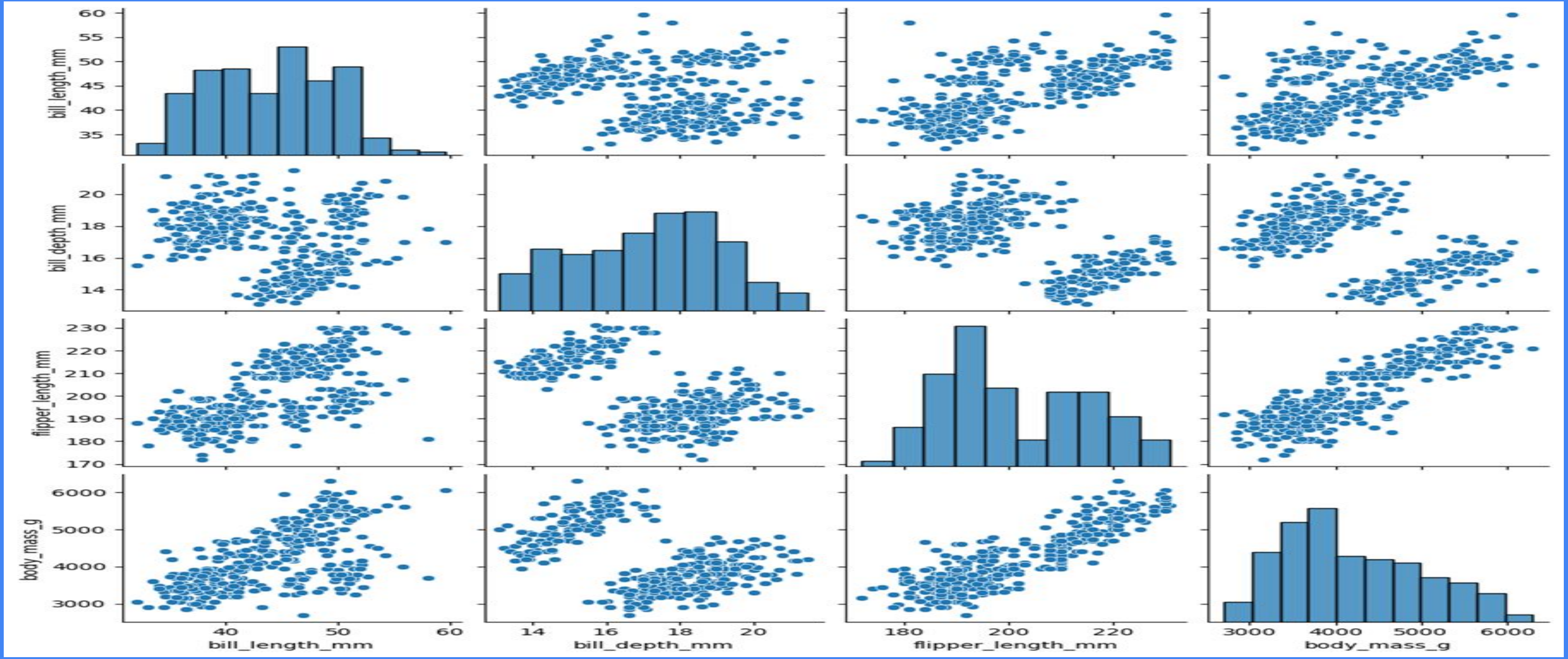


	Survived	Pclass	Sex	Age	SibSp	Parch	Fare
Survived	1.0	-0.34	-0.54	-0.08	-0.04	0.08	0.26
Pclass	-0.34	1.0	0.13	-0.37	0.08	0.02	-0.55
Sex	-0.54	0.13	1.0	0.09	-0.11	-0.25	-0.18
Age	-0.08	-0.37	0.09	1.0	-0.31	-0.19	0.1
SibSp	-0.04	0.08	-0.11	-0.31	1.0	0.41	0.16
Parch	0.08	0.02	-0.25	-0.19	0.41	1.0	0.22
Fare	0.26	-0.55	-0.18	0.1	0.16	0.22	1.0

Correlation matrix of the Titanic dataset

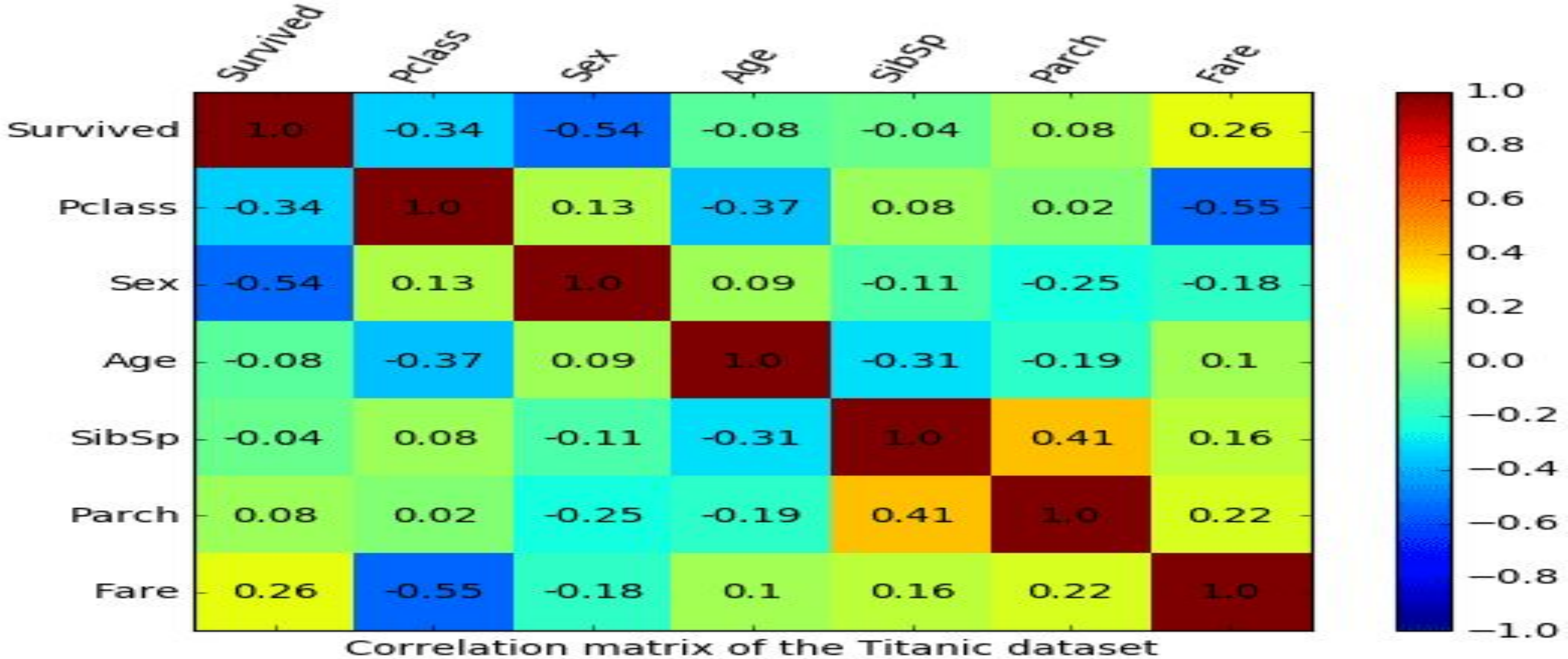
Multivariate Analysis

- Pair Plots



cont……….

- Heatmaps

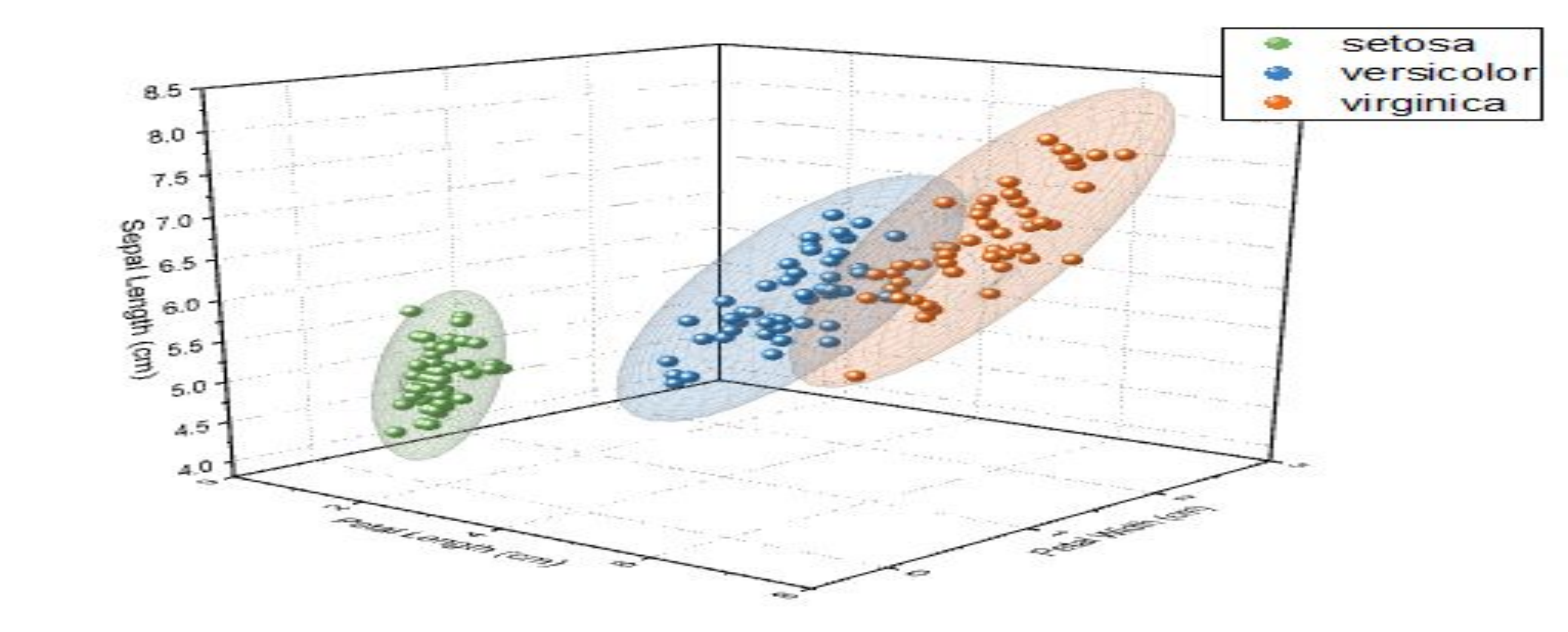


	Survived	Pclass	Sex	Age	SibSp	Parch	Fare
Survived	1.0	-0.34	-0.54	-0.08	-0.04	0.08	0.26
Pclass	-0.34	1.0	0.13	-0.37	0.08	0.02	-0.55
Sex	-0.54	0.13	1.0	0.09	-0.11	-0.25	-0.18
Age	-0.08	-0.37	0.09	1.0	-0.31	-0.19	0.1
SibSp	-0.04	0.08	-0.11	-0.31	1.0	0.41	0.16
Parch	0.08	0.02	-0.25	-0.19	0.41	1.0	0.22
Fare	0.26	-0.55	-0.18	0.1	0.16	0.22	1.0

Correlation matrix of the Titanic dataset

cont.........

- 3D Scatter Plots



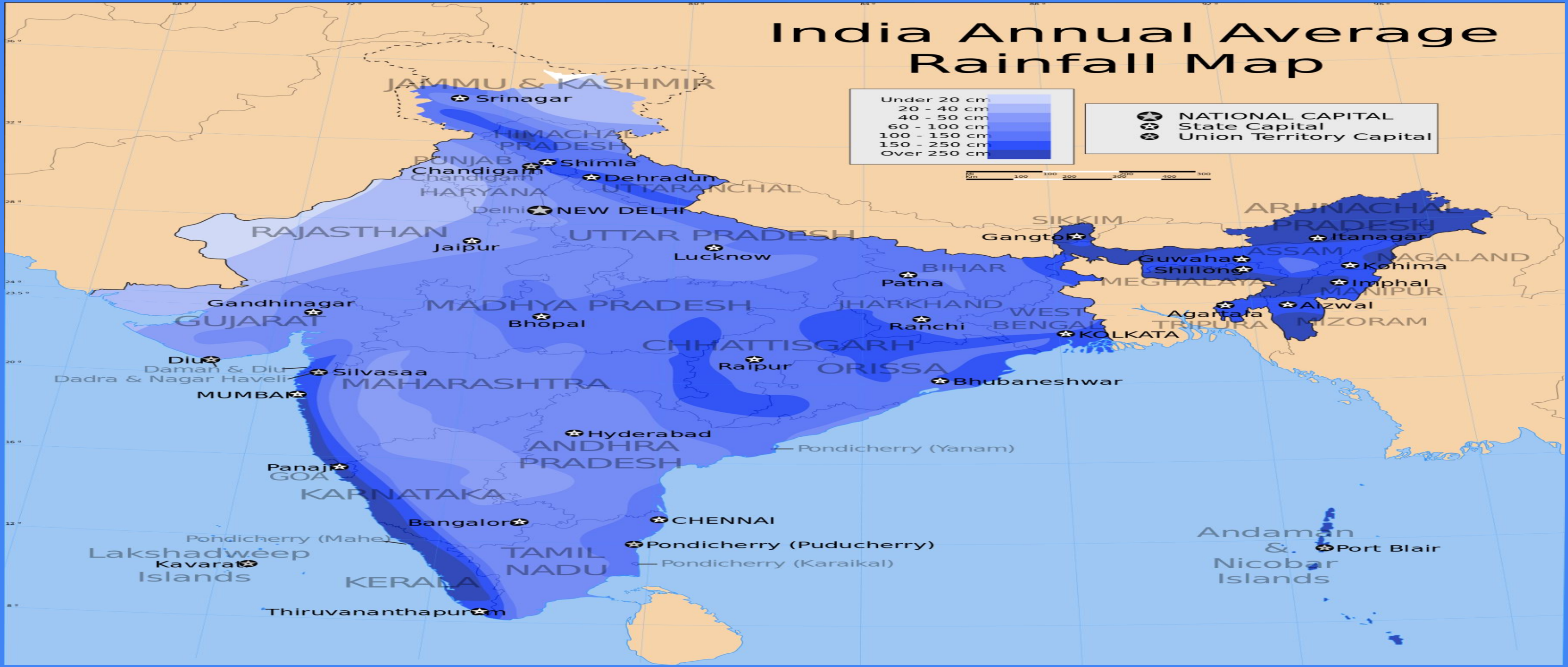
Advanced Visualization Techniques

- Interactive Visualizations

  Ex. Power Bi S/w

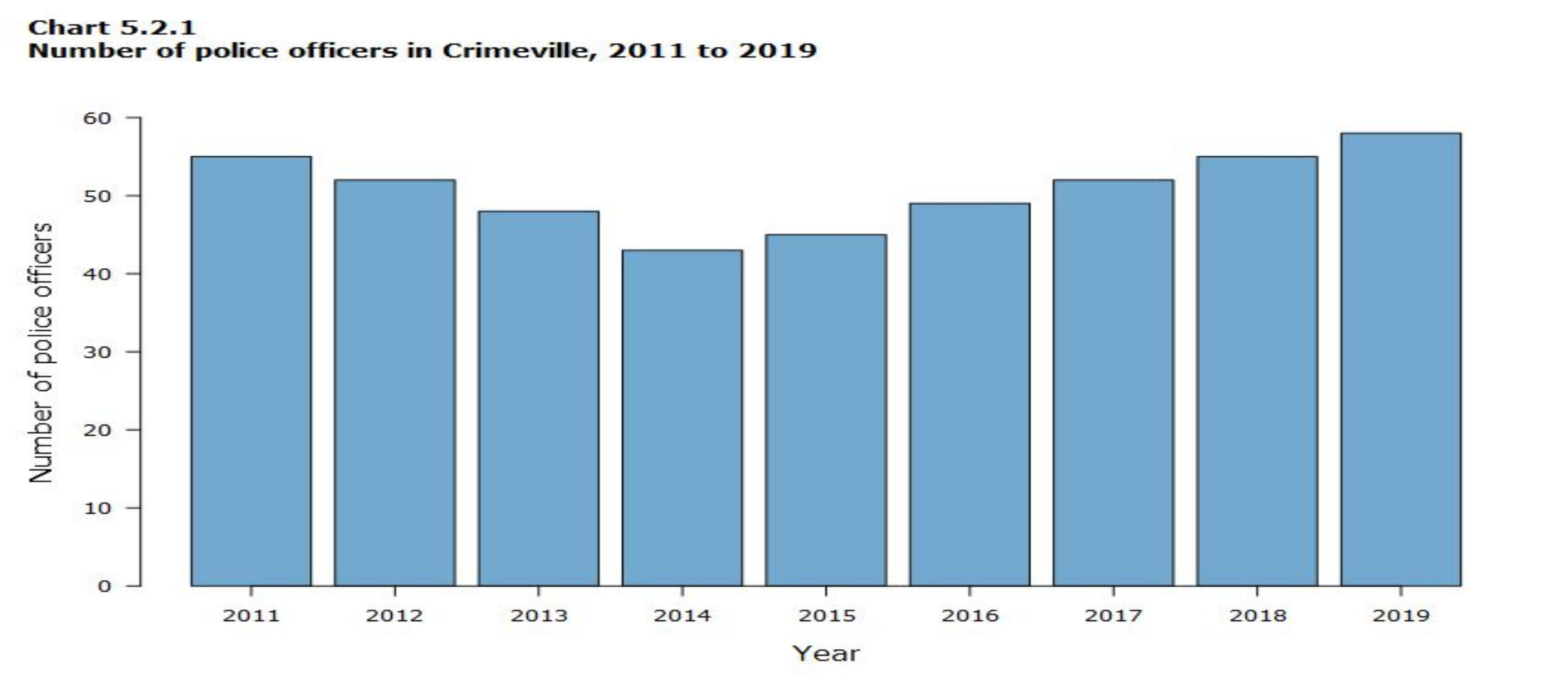
cont.....

- Geospatial Data Visualization



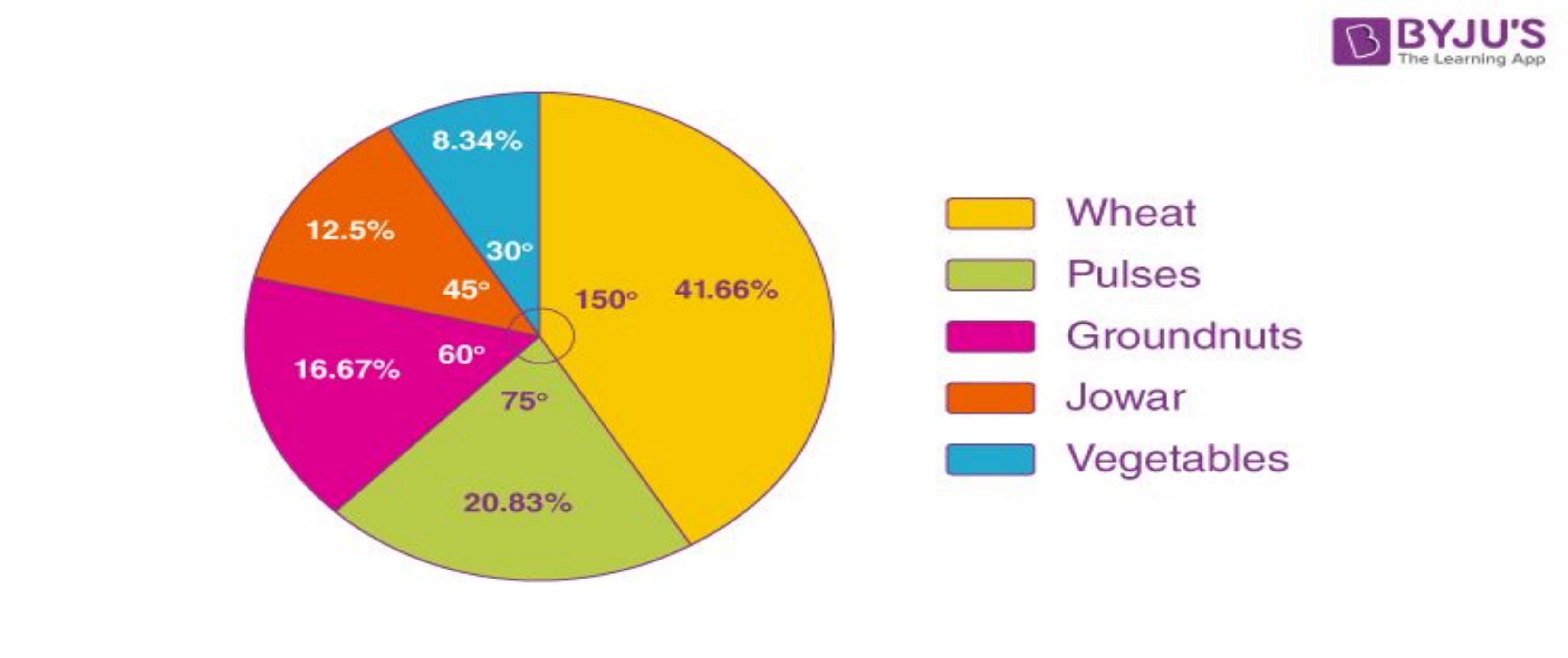
Visualizing Categorical Data

- Bar Charts



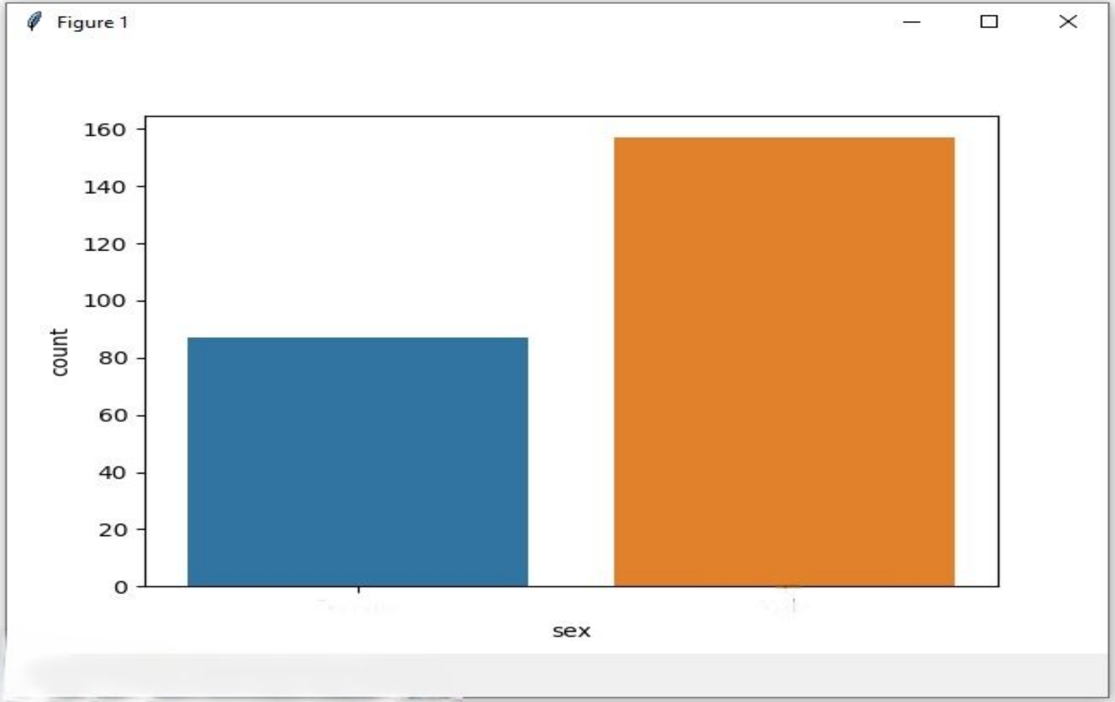
cont……….

- Pie Charts



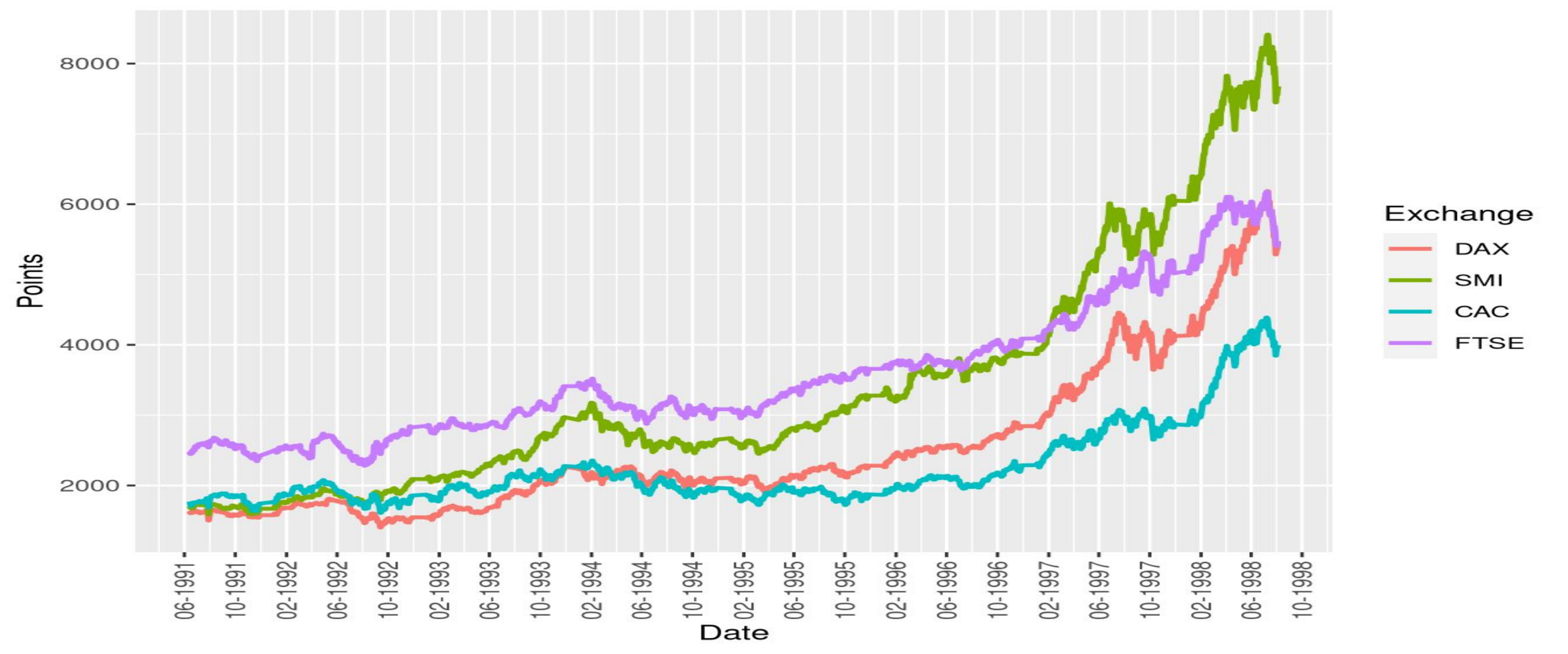
cont..

- Count Plots

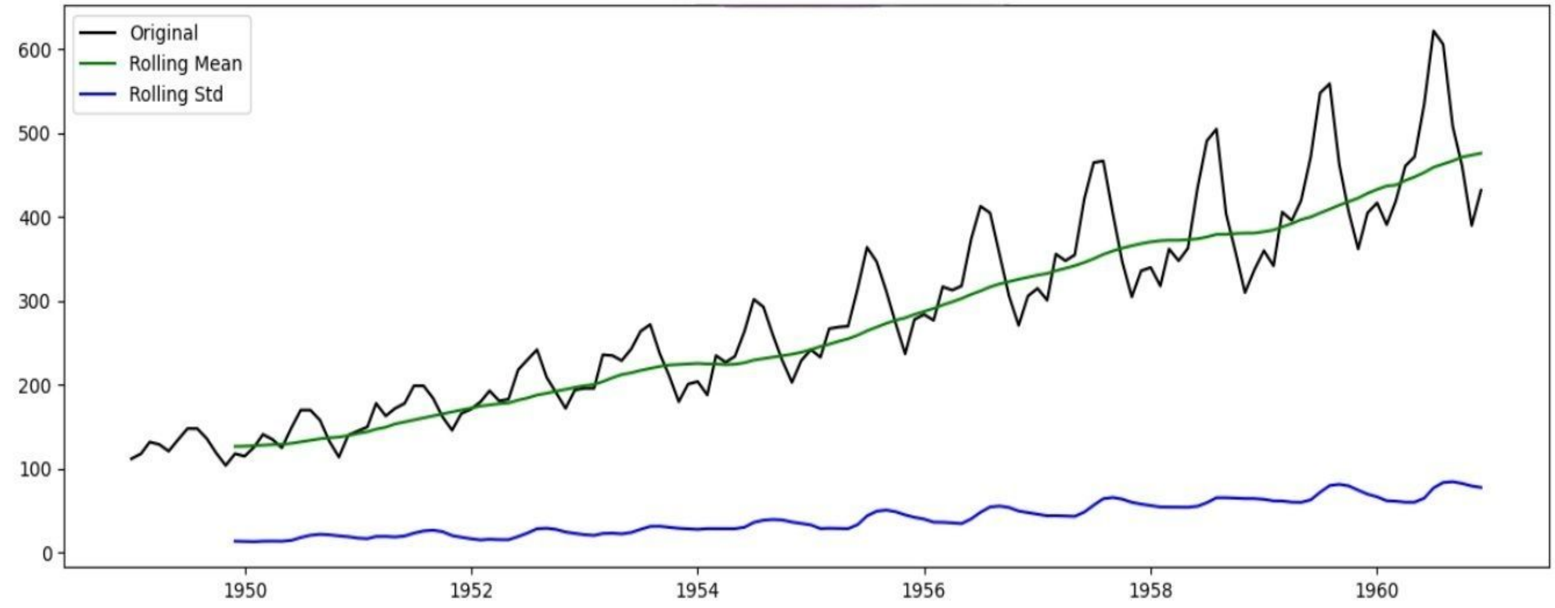


Time Series Data Visualization

- Line Plots

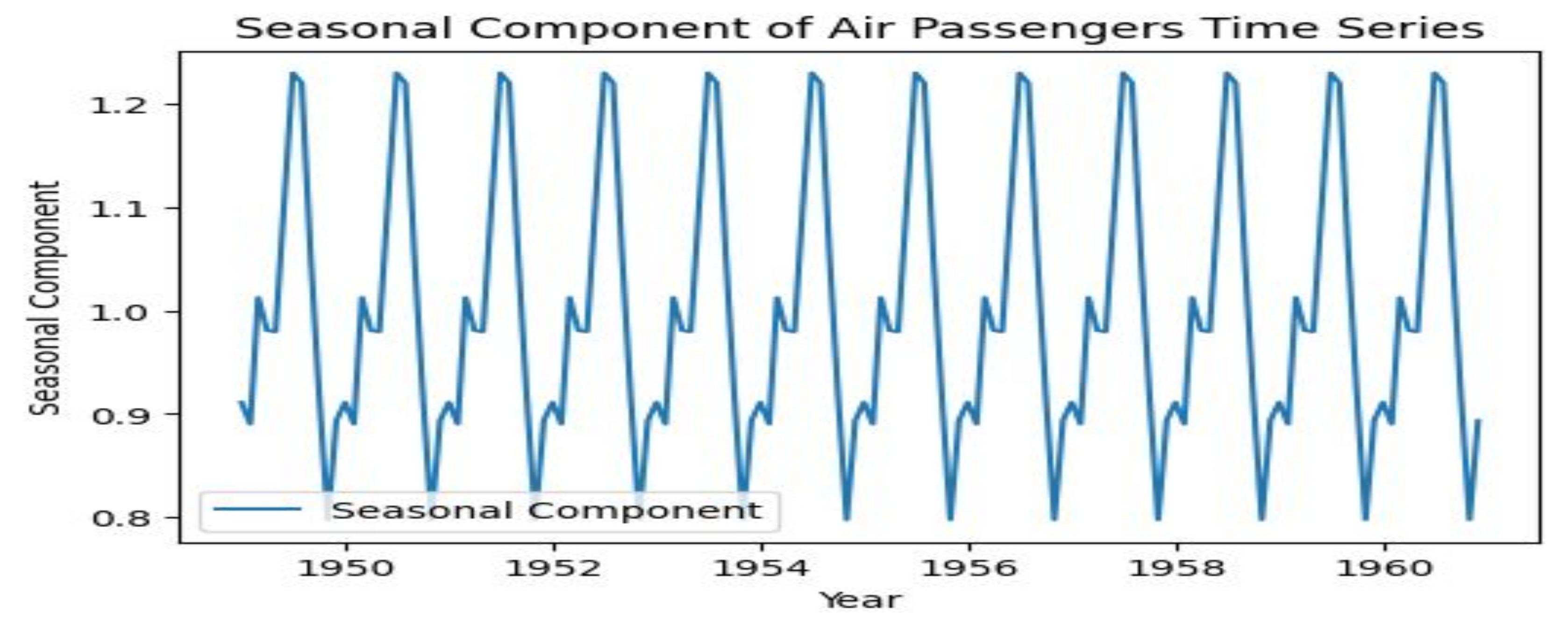


- Rolling Averages



cont….

- Seasonality Trends



Data Visualization Libraries in Python

- Pandas Visualization
- Matplotlib
- Seaborn

Advanced Libraries

- Plotly
- Tableau

Creating Interactive Dashboards

- Using Streamlit & Plotly

Case Study: EDA on a Real Dataset

- Data Cleaning
- Feature Engineering
- Visualization Insights

Common Challenges in EDA

- Handling Large Datasets
- Dealing with Noisy Data
- Choosing the Right Visuals

Best Practices for EDA & Visualization

- Use Meaningful Visuals
- Avoid Misleading Graphs
- Keep It Simple & Clear

Q&A Session

- Open Floor for Questions