# KATHMANDU UNIVERSITY

## Department of Artificial Intelligence

**Dhulikhel, Kavre**

**Lab Report On: Exploratory Data Analysis**

***Submitted By:***

Name: Aaryan Shakya

Roll No: 20

Subject Code: [AICC 301]

***Submitted To:***

Mr. Sunil Regmi

Lecturer, Department of Artificial Intelligence

6th July 2025

## Objective

1. Understand Data Distributions:
   To compute and interpret summary statistics (mean, median, mode, variance, etc.) to understand the central tendency and dispersion of the dataset.
2. Detect Patterns and Relationships:
   To use graphical methods such as scatter plots, histograms, and boxplots to identify trends, correlations, and outliers in the data.
3. Assess Data Quality:
   To identify missing values, duplicates, and anomalies through both statistical summaries and visualization techniques.
4. Gain Insights for Modeling:
   To derive insights and hypotheses that inform data preprocessing and feature selection for subsequent machine learning or statistical modeling.

## Introduction

Data sources:

1. Databases:
   Structured data stored in relational databases (like SQL) or NoSQL systems and data warehouses. These are common in business and transaction systems.
2. Flat Files:
   Includes CSV, Excel, text, XML, or JSON files. These are easy to use and often used for importing/exporting data.
3. APIs:
   Data collected through web services or application interfaces. Useful for accessing data from online platforms in a structured way.
4. Streaming Data:
   Real-time data from IoT devices, sensors, or live feeds. Often used in monitoring systems, stock markets, or smart devices.
5. Cloud Services:
   Data stored on platforms like AWS, Google Cloud, or Azure. These provide scalable and centralized data storage and access.
6. Manual Input:
   Data entered by users or operators. Though slower and error-prone, it's used in surveys, forms, and certain business operations.
7. Other Sources: Includes unstructured or semi-structured data from social media, RSS feeds, or web scraping. These provide real-time insights and trends.

## Types of Attributes in Data Mining

1. **Qualitative (Categorical) Attributes**
   a. Nominal Attributes
   The values of a nominal attribute are just different names or labels with no inherent ordering. These values simply help to distinguish one object from another. They allow only equality or inequality comparisons (i.e., =, ≠). Examples: Gender, eye color, zip code, employee ID.
   b. Ordinal Attributes
   The values of an ordinal attribute provide enough information to rank or order the data objects. However, the differences between values are not meaningful or measurable. Examples: Grades (A, B, C), customer satisfaction levels (bad, average, good), hardness of minerals.
   c. Binary Attributes
   Binary attributes have only two possible states or categories, usually represented as 0 and 1 (or yes/no, true/false). They can be symmetrical (no importance given to either outcome) or asymmetric (one outcome is more important). Examples: Male/Female, pass/fail, is student (yes/no).

2. **Quantitative (Numeric) Attributes**
   a. Interval Attributes
   The values of an interval attribute are numeric and have meaningful differences between them. However, they do not have a true zero point, so ratios are not meaningful. Examples: Temperature in Celsius or Fahrenheit, calendar dates.
   b. Ratio Attributes
   The values of a ratio attribute have all the properties of interval attributes, but they also have a true zero point. This makes both differences and ratios between values meaningful. Examples: Age, height, weight, income, temperature in Kelvin.
   c. Discrete Attributes
   Discrete attributes take on a finite or countably infinite set of values. They are often the result of counting. Examples: Number of children, number of products sold.
   d. Continuous Attributes
   Continuous attributes can take on any value within a range and are often obtained by measurement. They are typically represented as floating-point numbers. Examples: Height, weight, time, temperature.

3. **Other Data Types in Data Mining**
   a. Biological Sequences
   These are ordered sequences like DNA, RNA, or proteins, where the arrangement of characters carries biological meaning. Example: DNA sequences (ATGC...), amino acid chains.

b. Time Series Data
   Time series data consists of values collected at successive time intervals. It captures trends, cycles, and seasonal patterns. Example: Stock prices, temperature logs, ECG signals.

c. Image Data
   Image data consists of pixel values arranged in a grid, typically representing visual information. Example: MRI scans, satellite photos, face images.

d. Sound (Audio) Data
   Sound data is represented by waveforms or frequencies over time and is used in audio recognition tasks. Example: Voice recordings, music clips.

e. Video Data
   Video data is a sequence of image frames along with optional audio, representing dynamic visual scenes. Example: Surveillance footage, YouTube videos, motion capture.

## Exploratory Data Analysis (EDA)

EDA methods are used to summarize and visualize the important characteristics of data. These can be non-graphical (statistical summaries) or graphical (visual plots), and are applied to:

1. **Univariate Data (Single Variable)**
   a. Non-Graphical Methods:
      - Measures of Central Tendency: Mean, median, mode
      - Measures of Dispersion: Range, variance, standard deviation, interquartile range (IQR)
      - Summary Statistics: Min, max, skewness, kurtosis
      - Frequency Distribution: Tables of value count

   b. Graphical Methods:
      - Histogram: Distribution shape of numerical data
      - Boxplot: Shows median, quartiles, and outliers
      - Bar Chart: For categorical variables
      - Pie Chart: Shows proportion of categories

### 2. Bivariate Data (Two Variables)

   a. Non-Graphical Methods:
      - Correlation Coefficient (Pearson/Spearman): Measures strength of linear/monotonic relationship
      - Covariance: Indicates the direction of relationship
      - Cross-tabulation (Contingency Table): For two categorical variables
      - Difference of Means: For one numeric and one categorical variable

   b. Graphical Methods:
      - Scatter Plot: Relationship between two numeric variables
      - Line Graph: For two continuous variables over time

- Grouped Boxplots: Compare numeric variables across categorical groups
- Stacked Bar Chart: For categorical vs categorical variables

### 3. Multivariate Data (More than Two Variables)

a. Non-Graphical Methods:
   - Correlation Matrix: Pairwise correlation between multiple numeric variables
   - Summary Tables: Group-wise mean, median, etc.
   - Multivariate Statistics: PCA (Principal Component Analysis), MANOVA

b. Graphical Methods:
   - Heatmap: For correlation matrix or multi-variable comparison
   - Pair Plot (Scatterplot Matrix): Plots all combinations of numeric variable pairs
   - 3D Scatter Plot: Shows relationship among three numeric variables
   - Parallel Coordinates Plot: Shows patterns across multiple variables
   - Bubble Chart: Adds a third variable as size to a scatter plot

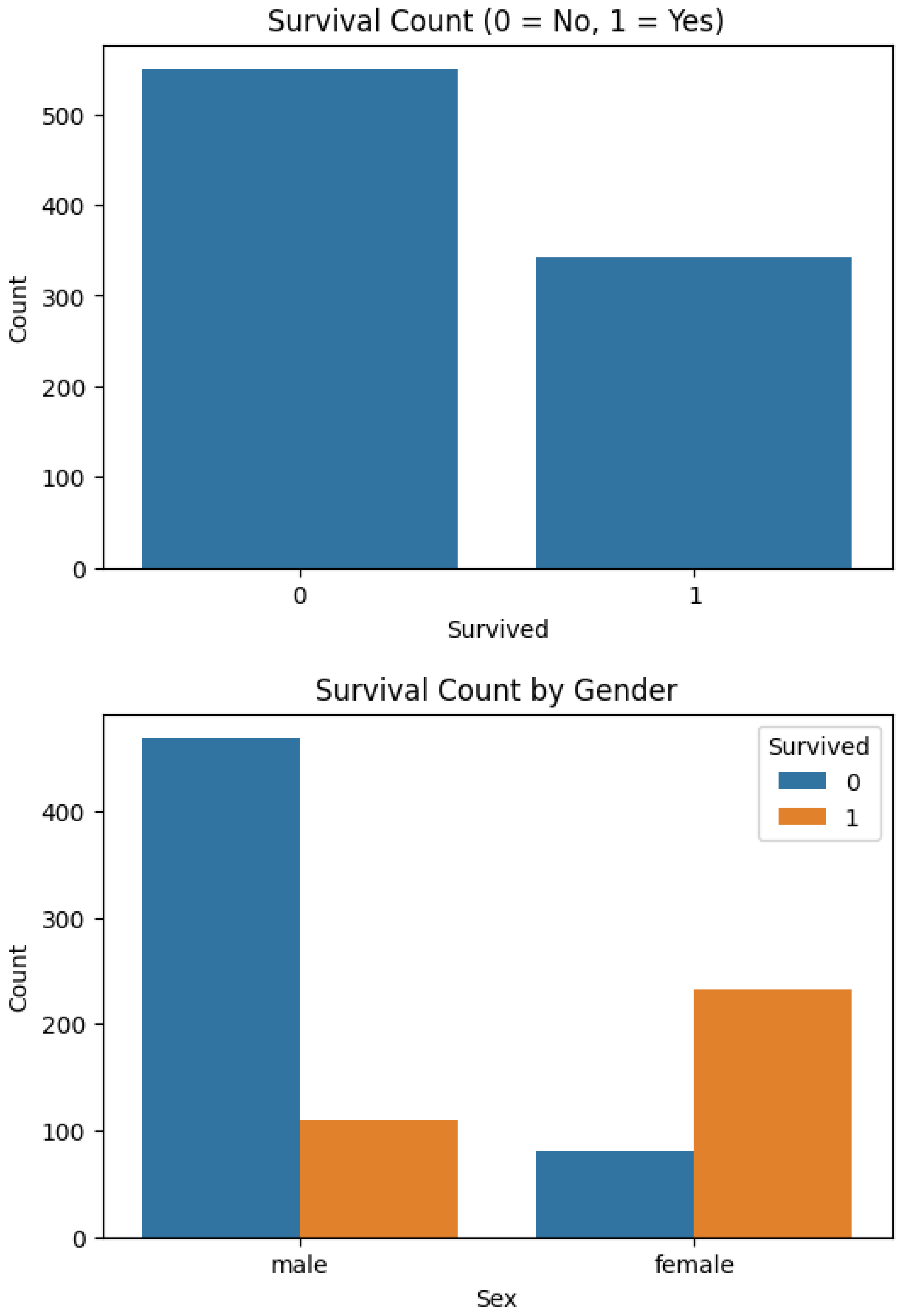
# Experiment

### Statistical Summary and Visualization of a Titanic Dataset

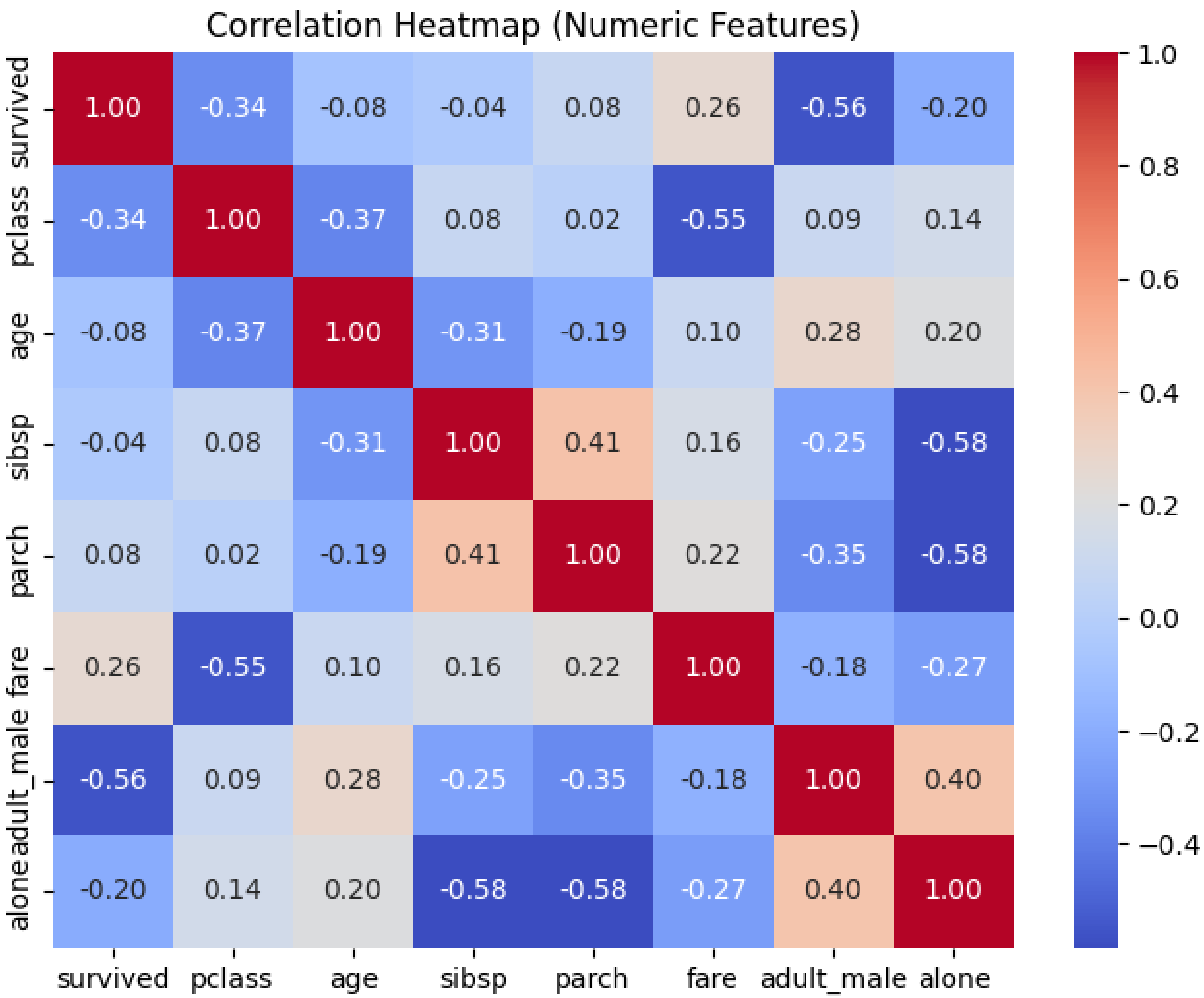
```
# Load the Titanic dataset from seaborn
titanic = sns.load_dataset("titanic")
print(titanic.head())

# Descriptive statistics (only numeric columns)
desc_stats = titanic.describe()
print("Descriptive Statistics:\n", desc_stats)

# Median
medians = titanic.median(numeric_only=True)
print("\nMedian:\n", medians)

# Mode (could return multiple rows if multimodal)
modes = titanic.mode(numeric_only=True)
print("\nMode:\n", modes)

titanic.hist(figsize=(10, 8), edgecolor='black')
plt.suptitle("Histograms of Titanic Numeric Features", fontsize=16)
plt.tight_layout()
plt.show()

plt.figure(figsize=(10, 6))
sns.boxplot(data=titanic[['age', 'fare']])
plt.title("Boxplot of Age and Fare")
plt.show()

plt.figure(figsize=(6, 4))
sns.countplot(x='survived', data=titanic)
plt.title("Survival Count (0 = No, 1 = Yes)")
plt.xlabel("Survived")
plt.ylabel("Count")
plt.show()

plt.figure(figsize=(6, 4))
sns.countplot(x='sex', hue='survived', data=titanic)
plt.title("Survival Count by Gender")
plt.xlabel("Sex")
plt.ylabel("Count")
plt.legend(title='Survived')
plt.show()

plt.figure(figsize=(8, 6))
sns.heatmap(titanic.corr(numeric_only=True), annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Correlation Heatmap (Numeric Features)")
plt.show()
```

**Outcome:**

The exploratory data analysis of the Titanic dataset provided meaningful insights into the demographics and survival patterns of the passengers. Out of a total of **891 passengers**, only **342 survived**, indicating a survival rate of approximately 38%. Descriptive statistics showed that the average age was around 29.7 years, and the average fare paid was about £32.20. The analysis revealed that most passengers were in the third class and had no family members aboard. Histograms and boxplots showed that the **fare variable had significant outliers**, with a maximum fare of over £512, while the median was just £14.45. This large gap suggests that a few wealthy passengers paid extremely high fares, possibly for luxury first-class accommodation, which skewed the distribution. These outliers in fare also correlated with a higher survival rate, highlighting the influence of socioeconomic status during the disaster. Additionally, females and passengers in higher classes had noticeably better survival chances. Overall, the analysis emphasizes the impact of class, gender, and economic status on survival outcomes aboard the Titanic.


Survival Count (0 = No, 1 = Yes)

Survival Count by Gender



Correlation Heatmap (Numeric Features)

| | survived | pclass | age | sibsp | parch | fare | adult_male | alone |
|---|---|---|---|---|---|---|---|---|
| survived | 1.00 | -0.34 | -0.08 | -0.04 | 0.08 | 0.26 | -0.56 | -0.20 |
| pclass | -0.34 | 1.00 | -0.37 | 0.08 | 0.02 | -0.55 | 0.09 | 0.14 |
| age | -0.08 | -0.37 | 1.00 | -0.31 | -0.19 | 0.10 | 0.28 | 0.20 |
| sibsp | -0.04 | 0.08 | -0.31 | 1.00 | 0.41 | 0.16 | -0.25 | -0.58 |
| parch | 0.08 | 0.02 | -0.19 | 0.41 | 1.00 | 0.22 | -0.35 | -0.58 |
| fare | 0.26 | -0.55 | 0.10 | 0.16 | 0.22 | 1.00 | -0.18 | -0.27 |
| adult_male | -0.56 | 0.09 | 0.28 | -0.25 | -0.35 | -0.18 | 1.00 | 0.40 |
| alone | -0.20 | 0.14 | 0.20 | -0.58 | -0.58 | -0.27 | 0.40 | 1.00 |

## Conclusion

In this lab, we successfully explored a sample dataset (Titanic/Iris) using various statistical and graphical techniques. By leveraging the capabilities of **Pandas** for descriptive statistics and **Matplotlib/Seaborn** for data visualization, we gained valuable insights into the structure, distribution, and relationships within the data.

Through **univariate analysis**, we observed individual variable distributions using histograms, box plots, and bar charts. **Bivariate analysis** helped us identify relationships and correlations between two variables via scatter plots and correlation matrices. Further, **multivariate visualizations** like pair plots and heatmaps provided a comprehensive view of interactions across multiple variables.

This experiment enhanced our understanding of data characteristics, emphasized the importance of initial data exploration, and demonstrated how EDA serves as a foundation for further data preprocessing, modeling, and decision-making in data science.