Exploratory Data Analysis

Agenda

- Exploratory Data Analysis Quiz
- Univariate Analysis
- Bivariate / Multivariate Analysis
- Missing Value Treatment
- Outlier Detection and Treatment

Let's begin the discussion by answering a few questions on Exploratory Data Analysis

EDA Quiz

What is the primary purpose of conducting a data overview?

A. To validate statistical hypotheses

B. To create machine learning models

C. To understand the high-level structure and patterns of the data

D. To make predictions without understanding the data

EDA Quiz

What is the primary purpose of conducting a data overview?

A. To validate statistical hypotheses

B. To create machine learning models

C. To understand the high-level structure and patterns of the data

D. To make predictions without understanding the data

Data Overview

Critical for gaining an early understanding of the data and directing subsequent steps in the analytical process

Method	Syntax	Description
Shape of Dataset	`df.shape`	It provides dimensions of the dataset (no. of rows and columns)
Information of the dataset	`df.info()`	It provides essential details such as the total number of non-null values, data types of each column, etc.
Statistical summary of the dataset	`df.describe()`	It returns a statistical summary of the attributes in the data

EDA Quiz

Which statistical measure provides information about the spread or variability of a dataset?

A. Mean

B. Median

C. Mode

D. Standard Deviation

EDA Quiz

Which statistical measure provides information about the spread or variability of a dataset?

A. Mean

B. Median

C. Mode

D. Standard Deviation

Summary Statistics

Summary Statistic	Description
Mean	The average of all values in a numerical attribute
Median	The middle value of a numerical attribute when arranged in ascending / descending order.
Mode	The most frequently occuring value(s) in an attribute (numerical / categorical)
Standard Deviation	The average distance between the mean value and all the values of a numerical attribute

Spread measures the distance between data points in a dataset, while variability measures the degree of diversity or differences within the dataset.

EDA Quiz

What is the primary objective of Univariate Analysis?

A. Analyze how various variables relate to one another

B. Spot patterns and structures in individual variables

C. Make confident predictions about future trends

D. Identify the independent variables that affect a model's prediction

EDA Quiz

What is the primary objective of Univariate Analysis?

A. Analyze how various variables relate to one another

B. Spot patterns and structures in individual variables

C. Make confident predictions about future trends

D. Identify the independent variables that affect a model's prediction

Univariate Analysis

The distribution, spread, and central tendency of a single variable in a dataset are examined without taking into account the relationships with other variables

Plot	Type of variable	Python Function
Histogram	Numerical	`plt.hist()` or `sns.histplot()`
Boxplot	Numerical	`sns.boxplot()`
KDE plot	Numerical	`sns.displot()`
Bar graph	Categorical	`plt.bar()` or `sns.countplot()`

EDA Quiz

When the mean is greater than the median, the distribution becomes:

A. Positively skewed

B. Negatively skewed

C. Symmetric

D. Uniform

EDA Quiz

When the mean is greater than the median, the distribution becomes:

A. Positively skewed

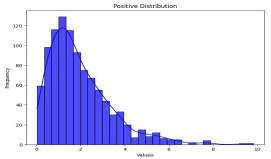
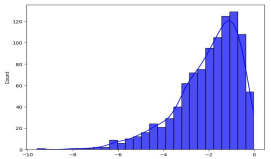
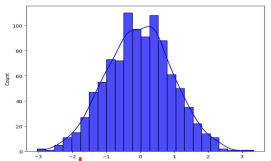
B. Negatively skewed

C. Symmetric

D. Uniform

Skewness

A measure of the deviation of the probability distribution of a variable from its mean.

Type	Description	Sample Graph
Positive Skewness (Right Skewed)	• Majority of the data points are concentrated on the left side. • Mean > Median	
Negative Skewness (Left Skewed)	• Majority of the data points are concentrated on the right side. • Mean < Median	
Symmetric Distribution	• Data is evenly distributed on both sides of the mean. • Mean = Median = Mode	

EDA Quiz

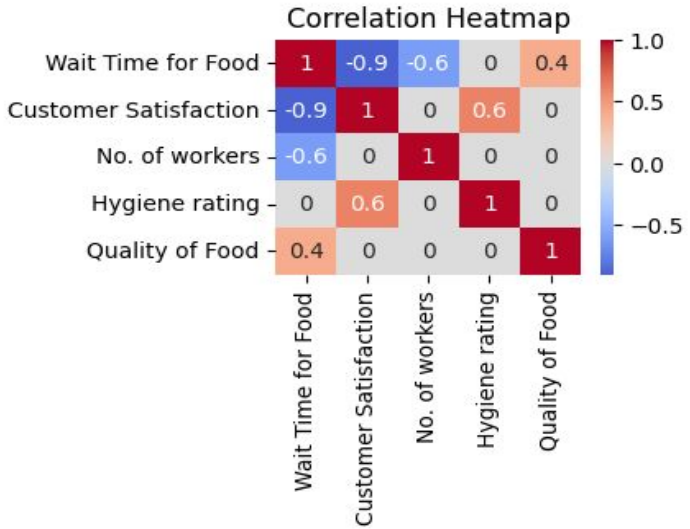
According to the heatmap below, which pair of variables is most correlated with each other?

A. Wait Time for Food and Customer Satisfaction

B. No. of workers and Wait Time for Food

C. Hygiene rating and Customer satisfaction

D. Wait time for Food and Quality of Food



EDA Quiz

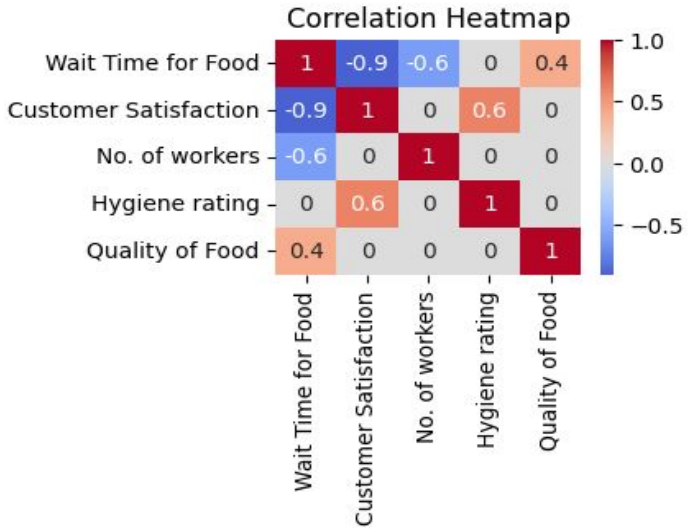
According to the heatmap below, which pair of variables is most correlated with each other?

A. Wait Time for Food and Customer Satisfaction

B. No. of workers and Wait Time for Food

C. Hygiene rating and Customer satisfaction

D. Wait time for Food and Quality of Food



Correlation

A statistical measure of the **association between two variables**

Measures both strength and direction of the relationship between pairs of variables

A **correlation heatmap** displays the correlation coefficients between pairs of variables, using **color intensity to represent the strength and direction of correlations**.

The **strength of the correlation is independent of the direction** - one can have strong positive and negative correlations - **-0.9 correlation is stronger than +0.6**

Same magnitude but different directions of correlation imply variables with opposite relationship - inverse association for -ve correlation and direct association for +ve correlation

EDA Quiz

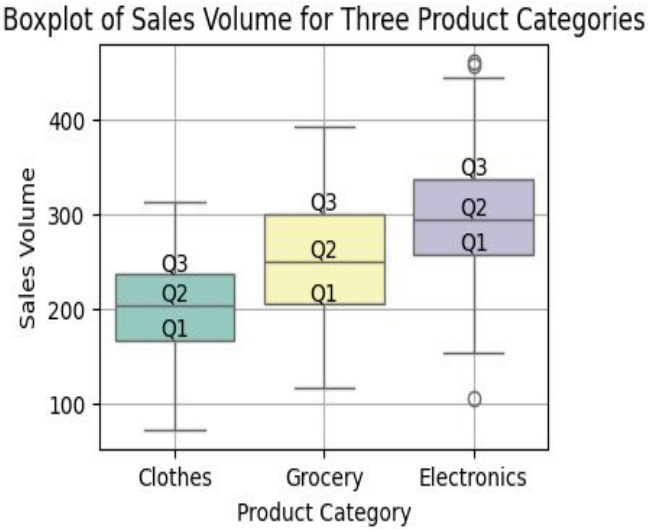
According to the given boxplot, x% of Electronics category has higher Sales volume than y% of Clothes category. What are the values of x and y?

A. 25, 100

B. 50, 75

C. 10, 100

D. 50, 100



EDA Quiz

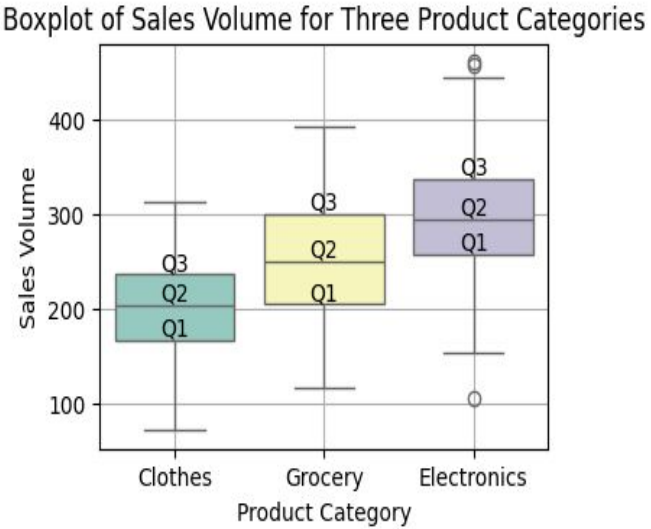
According to the given boxplot, x% of Electronics category has higher Sales volume than y% of Clothes category. What are the values of x and y?

A. 25, 100

B. 50, 75

C. 10, 100

D. 50, 100

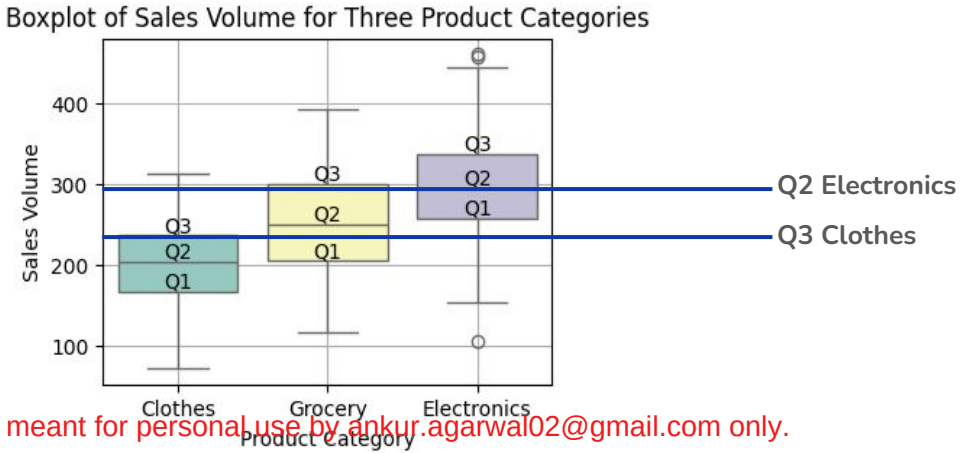


Quartiles and Boxplot

Q1 is the value below which 25% of the data falls, i.e, 25% of data values <= Q1

Q2 (median) is the value below which 50% of the data falls, i.e., it splits the dataset into two equal halves

Q3 is the value below which 75% of the data falls, i.e, 75% of data values <= Q3



EDA Quiz

Consider a dataset containing the columns Work experience (in years) and salary ($), Which of the following methods are generally used to deal with missing values in the salary column?

A. Imputation by Mean

B. Imputation by Median

C. Imputation by Mode

D. Dropping the missing values

EDA Quiz

Consider a dataset containing the columns Work experience (in years) and salary ($), Which of the following methods are generally used to deal with missing values in the salary column?

A. Imputation by Mean

B. Imputation by Median

C. Imputation by Mode

D. Dropping the missing values

Missing Values in Data

Missing values indicate that there is no data for a given variable or observation, and are generally represented as `None` or `NaN` (Not a Number).

The **selection of a treatment technique is influenced by various factors**, including the nature and amount of missing data, the type of analysis, and the study's specific objectives.

Missing Value Treatment

Method	Working
Imputation by Mean	Replaces missing values with the mean of non-missing values in the column
Imputation by Median	Replaces missing values with the median of non-missing values in the column - more suitable when the data is skewed
Imputation by Mode	Replaces missing values with the most frequently occurring value in the column - primarily used for categorical variables
Dropping rows with missing values	Removes rows with missing values from the dataset - appropriate when the missing values are few and dropping them doesn't impact the analysis
Dropping attributes with missing values	Removes attributes with missing values from the dataset - appropriate when the proportion of missing values in the attribute is high and imputation might impact the data distribution

EDA Quiz

In general, data points which are less than Q1 - x * Interquartile Range (IQR) or greater than Q3 + x * IQR are considered to be outliers.
What is the value of x?

A. 1

B. 3

C. 1.5

D. 2

EDA Quiz

In general, data points which are less than Q1 - x * Interquartile Range (IQR) or greater than Q3 + x * IQR are considered to be outliers.
What is the value of x?

A. 1

B. 3

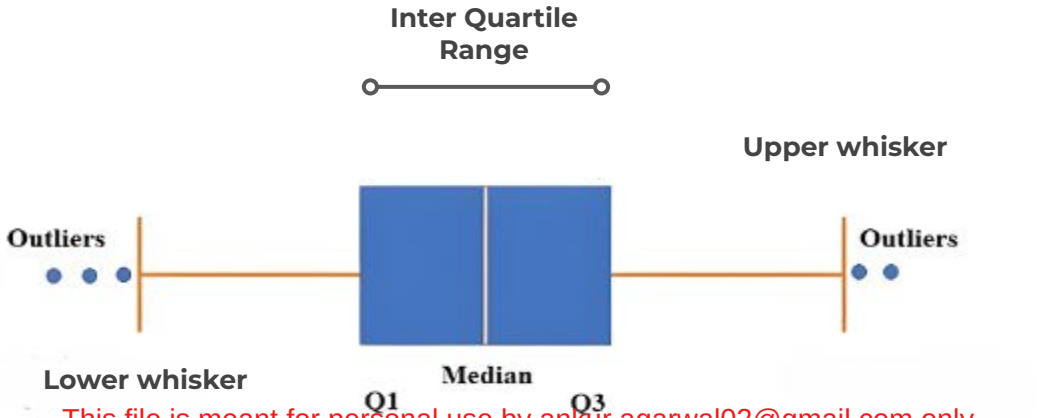
C. 1.5

D. 2

Outlier Detection

Data points that **deviate significantly from the majority of the observations** in a dataset, potentially impacting analysis and modeling

Values less than Q1 - 1.5 * IQR (lower whisker) or greater than Q3 + 1.5 * IQR (upper whisker) are considered as outliers.



Great Learning

Happy Learning !