JGI JAIN DEEMED-TO-BE UNIVERSITY | Powered By Futurense

Revolutionising B.Tech

JGI JAIN DEEMED-TO-BE UNIVERSITY | Powered By Futurense

Module 4:
Data Exploration

Course Name: **DATA EXPLORATION AND PREPARATION** [22TCSE239]

Total Hours : 09

Table of Content

Aim

To equip students in Exploring data visually and statistically and identify patterns and abnormalities that require further investigation

Objective

A. Identify key steps in data exploration and preparation such as data cleaning, data wrangling. data transformation and data integration.

B. Ability to code in a scalable environment to automate data cleaning and processing tasks.

C. Understand the sources of data quality issues. different types of data, how data is stored and accessed.

D. Explore data visually and statistically and identify patterns and abnormalities that require further investigation.

E. Able to clean and transform data, identify missing or incorrect values, and merge or join datasets as needed.

F. Evaluate the effectiveness and efficiency of data exploration and preparation methods in terms of their impact on downstream analysis and modelling.

1. Introduction to Exploratory Data Analysis:

What is Exploratory Data Analysis

Exploratory Data Analysis is a process of examining or understanding the data and extracting insights dataset to identify patterns or main characteristics of the data.

- Analyze and visualize the data to gain insights.
- Identify patterns, trends, and potential relationships.

EDA is generally classified into two methods, i.e. **graphical** analysis and **non-graphical** analysis.

Technically, The primary motive of EDA is to

- Examine the data distribution
- Handling missing values of the dataset(a most common issue with every dataset)
- Handling the outliers
- Removing duplicate data
- Encoding the categorical variables
- Normalizing and Scaling

Descriptive Statistics for data analysis, Data Distribution Analysis, Correlation Analysis

Data Analysis:

Definition: Data analysis is a broad process involving the examination, cleaning, transformation, and interpretation of data to extract meaningful insights and support decision-making.

Approaches/Methods:

Descriptive Statistics: Summarize and describe the main features of a dataset (mean, median, standard deviation, etc.).

Inferential Statistics: Make inferences and predictions about a population based on a sample of data.

Data Visualization: Use graphical representations to convey patterns, trends, and relationships in the data.

Machine Learning: Employ algorithms and models for predictive modeling and pattern recognition.

Data Distribution Analysis:

Definition: Data distribution analysis specifically focuses on understanding the distribution of values within a dataset, exploring patterns, and characterizing the shape of the distribution.

Approaches/Methods:

Histograms: Visualize the frequency distribution of values.

Quantile-Quantile (Q-Q) Plots: Compare observed quantiles with expected quantiles from a theoretical distribution.

Box Plots: Summarize central tendency, spread, and identify outliers.

Summary Statistics: Calculate measures like skewness, kurtosis, mean, median, etc.

Normality Tests: Assess whether the data follows a normal distribution.

Correlation Analysis:

Definition: Correlation analysis examines the relationship between two or more variables to understand the degree and direction of their association.

Approaches/Methods:

Correlation Coefficients: Compute measures such as Pearson correlation coefficient for linear relationships or Spearman rank correlation for non-linear relationships.

Scatter Plots: Visualize the relationship between variables graphically.

Correlation Matrices: Display the correlation coefficients between multiple pairs of variables.

1.1 Descriptive Statistics for data analysis

Two types of statistical methods are widely used in data analysis: descriptive and inferential

Aspect	Descriptive Statistics	Inferential Statistics
Purpose	Summarize and describe data	Draw conclusions or predictions
Data Sample	Analyzes the entire dataset	Analyzes a sample of the data
Examples	Mean, Median, Range, Variance	Hypothesis testing, Regression
Scope	Focuses on data characteristics	Makes inferences about populations
Goal	Provides insights and simplifies data	Generalizes findings to a larger population
Assumptions	No assumptions about populations	Requires assumptions about populations
Common Use Cases	Data visualization, data exploration	Scientific research, hypothesis testing

- These statistics are fundamental for data analysis and interpretation. Here are some commonly used methods:

1. Measures of Central Tendency:

Mean (Average):

- Represents the arithmetic average of a set of values.
- Calculated as the sum of all values divided by the number of values.
- Affected by extreme values (outliers).

$$\text{Mean} = X_1 + X_2 + X_3 + \ldots + X_n / n$$

Median:

- Represents the middle value when the data is sorted.
- Less sensitive to extreme values compared to the mean.
- Suitable for skewed distributions.

Mode:

- Represents the most frequently occurring value in the dataset.
- A dataset may have no mode, one mode, or multiple modes.

Example:

Data – 10,20,30,40,50 and Number of observations = 5

Mean = [10+20+30+40+50] / 5

Mean = 30

Example:

Odd number of Data – 10,20,30,40,50

Median is 30.

Even the number of data – 10,20,30,40,50,60

Find the middle 2 data and take the mean of those two values.

Here, 30 and 40 are middle values.

Now, add them and divide the result by 2

30+40 / 2 =35

Median is 35

Example:

Data – 1,3,4,6,7,3,3,5,10, 3

Mode is 3, because 3 has the highest frequency (4 times)

2. Measures of Dispersion:

Range:

- Represents the difference between the maximum and minimum values in a dataset.
- Sensitive to outliers. Range=Maximum Value−Minimum Value

Variance:

- Measures the average squared deviation of each data point from the mean.
- Provides a measure of the dataset's spread.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

where:

- σ^2 is the variance,
- n is the number of data points in the set,
- X_i represents each individual data point,
- $\bar{X}$ is the mean of the data set.

Standard Deviation:

- Represents the square root of the variance.
- Provides a more interpretable measure of the spread.
- Sensitive to outliers.

$$\sigma = \sqrt{\sigma^2}$$

Interquartile Range (IQR):

- Represents the range of values between the first quartile (25th percentile) and the third quartile (75th percentile).
- Less sensitive to extreme values than the range.
- IQR=Q3−Q1

3. Measures of Shape:

Skewness:

- Measures the asymmetry of a distribution.
- Positive skewness indicates a right-skewed distribution.
- Negative skewness indicates a left-skewed distribution.

$$g_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$$

where:

- n is the number of observations in the sample,
- x_i is each individual observation,
- $\bar{x}$ is the sample mean,
- s is the sample standard deviation.

Kurtosis:

- Measures the peakedness or flatness of a distribution.
- Positive kurtosis indicates a more peaked distribution
- Negative kurtosis indicates a flatter distribution

$$g_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

where:

- n is the number of observations in the sample,
- x_i is each individual observation,
- $\bar{x}$ is the sample mean,
- s is the sample standard deviation.

4. Percentiles:

Quartiles:

- Divide the data into four equal parts (25th, 50th, and 75th percentiles).
- Useful for understanding the spread of the data and identifying outliers.

$$\cdot\ Q1 = \left(\frac{25}{100}\right)(n+1)$$
$$\cdot\ Q2 = \left(\frac{50}{100}\right)(n+1)$$
$$\cdot\ Q3 = \left(\frac{75}{100}\right)(n+1)$$

Percentile Rank:

- Represents the percentage of values in a dataset that are less than or equal to a particular value.
- Useful for understanding a specific value's relative position in the dataset.

- *PR*=(Number of values below or equal to the given value/Total number of values in the dataset)× 100

5. Summary Statistics:

Count:

The number of observations in the dataset.

Provides the dataset's size.

Sum:

The total of all values in the dataset.

Useful for variables with a clear additive meaning.

Minimum and Maximum:

Identify the smallest and largest values in the dataset.

Useful for understanding the range of values.

6. Frequency Distribution:

Frequency Table:

Lists the frequency (count) of each unique value in a dataset.

Useful for understanding the distribution of categorical variables.

Histogram:

A graphical representation of the frequency distribution for continuous variables.

Helps visualize the shape of the data distribution.

7. Correlation and Covariance:

Covariance:

Measures the degree of joint variability between two variables.

Correlation Coefficient:

Standardized measure of the linear relationship between two variables.

Ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation).

1.2 Data Distribution Analysis

1. **Histograms:**
   Description: Histograms are graphical representations that display the frequency distribution of values in a dataset.
   Method: Divide the range of data into intervals (bins) and count the number of observations in each bin. The resulting bar chart provides a visual representation of the data distribution.
2. **Box Plots (Box-and-Whisker Plots):**
   Description: Box plots provide a visual summary of the central tendency, spread, and identify potential outliers in the data distribution.
   Method: The box represents the interquartile range (IQR), with a line inside indicating the median. Whiskers extend to the minimum and maximum values within a defined range, and outliers may be plotted individually.
3. **Kernel Density Plots:**
   Description: Kernel density plots estimate the probability density function of the data, offering a smooth representation of the distribution.
   Method: Smooth a kernel (usually Gaussian) over each data point and sum the results to create a continuous density plot.
4. **Quantile-Quantile (Q-Q) Plots:**
   Description: Q-Q plots compare the quantiles of the observed data against the quantiles of a theoretical distribution (e.g., normal distribution).
   Method: Plot the observed quantiles against the quantiles expected under a theoretical distribution. Deviations from the diagonal line suggest departures from the assumed distribution.

5. **Summary Statistics:**
   Description: Calculating summary statistics provides numerical measures of the central tendency, spread, skewness, and kurtosis of the data.
   Method: Compute mean, median, mode, variance, standard deviation, skewness, and kurtosis to quantify various aspects of the data distribution.
6. **Normality Tests:**
   Description: Normality tests assess whether the data follows a normal distribution.
   Method: Utilize statistical tests such as the Shapiro-Wilk test or Anderson-Darling test to determine if the data significantly deviates from a normal distribution.
7. **Kernel Density Estimation (KDE):**
   Description: KDE is a non-parametric method for estimating the probability density function of a random variable.
   Method: Smoothly estimate the underlying distribution by placing a kernel at each data point and summing the results.
8. **Empirical Cumulative Distribution Function (ECDF):**
   Description: The ECDF provides a visual representation of the cumulative distribution of the data.
   Method: Plot the cumulative proportion of data points below each value in the dataset.

1.3 Correlation Analysis

1. Pearson Correlation Coefficient

The Pearson correlation coefficient (r) measures the linear correlation between two continuous variables. The formula for the Pearson correlation coefficient is:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Here's an example with two datasets, X and Y:

$$X : [2, 4, 6, 8, 10]$$
$$Y : [1, 3, 5, 7, 9]$$

1. **Calculate Means:**
   $$\bar{X} = \frac{2+4+6+8+10}{5} = 6$$
   $$\bar{Y} = \frac{1+3+5+7+9}{5} = 5$$
2. **Calculate Differences from Means:**
   $$X - \bar{X} : [-4, -2, 0, 2, 4]$$
   $$Y - \bar{Y} : [-4, -2, 0, 2, 4]$$

3. **Calculate Cross-Products:**

   $(X_i - \bar{X})(Y_i - \bar{Y}) : [16, 4, 0, 4, 16]$

4. **Calculate Sums:**

   $\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = 40$

   $\sum_{i=1}^{n}(X_i - \bar{X})^2 = 40$

   $\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 40$

5. **Calculate Pearson Correlation Coefficient:**

   $r = \frac{40}{\sqrt{40 \cdot 40}} = 1$

In this example, the Pearson correlation coefficient (r) is 1, indicating a perfect positive linear correlation between the two datasets X and Y. As one variable increases, the other variable also increases proportionally.

2 Spearman Rank Correlation Coefficient

The Spearman Rank Correlation Coefficient (ρ) is used to assess the monotonic relationship between two variables. The formula for Spearman's rank correlation coefficient is:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

where d_i is the difference between the ranks of corresponding pairs of observations and n is the number of pairs of observations.

Let's consider an example with two datasets, X and Y:

$$X : [10, 5, 8, 3, 7]$$
$$Y : [20, 15, 18, 12, 17]$$

1. **Rank the Data:**

$$X : [4, 2, 3, 1, 2.5]$$
$$Y : [5, 2, 4, 1, 3]$$

2. **Calculate Differences:**

$$d_i = X_i - Y_i$$

$$d_i : [-1, 0, -1, 0, -0.5]$$

3. **Square the Differences:**

$$d_i^2 : [1, 0, 1, 0, 0.25]$$

4. **Sum the Squares of Differences:**

$$\sum d_i^2 = 2.25$$

5. **Calculate Spearman Rank Correlation Coefficient:**

$$\rho = 1 - \frac{6 \times 2.25}{5(5^2 - 1)} = 1 - \frac{13.5}{100} = 0.865$$

In this example, the Spearman rank correlation coefficient (ρ) is approximately 0.865, indicating a strong positive monotonic relationship between the two datasets X and Y. As the values in one dataset increase, the ranks of the corresponding values in the other dataset also tend to increase.

3. Kendall Tau Rank Correlation Coefficient

The Kendall Tau Rank Correlation Coefficient (τ) is used to measure the strength and direction of a monotonic relationship between two variables. The formula for Kendall Tau is:

$$\tau = \frac{(\text{concordant pairs} - \text{discordant pairs})}{\frac{1}{2}n(n-1)}$$

where:

- concordant pairs is the number of pairs with the same order in both variables.
- discordant pairs is the number of pairs with different orders in the two variables.
- n is the number of pairs of observations.

Let's consider an example with two datasets, X and Y:

$X : [3, 1, 4, 2]$
$Y : [2, 4, 1, 3]$

1. **Pairwise Comparisons:**
   - Compare each pair of observations in both variables and determine if they are concordant or discordant.

$$(3, 2) \text{ - Discordant } (3 > 2)$$
$$(1, 4) \text{ - Concordant } (1 < 4)$$
$$(4, 1) \text{ - Discordant } (4 > 1)$$
$$(2, 3) \text{ - Concordant } (2 < 3)$$

   - Count the number of concordant pairs (c) and discordant pairs (d).

$$c = 2, \quad d = 2$$

2. **Calculate Kendall Tau:**

$$\tau = \frac{c-d}{\frac{1}{2}n(n-1)}$$
$$\tau = \frac{2-2}{\frac{1}{2}\times 4\times(4-1)} = -0.5$$

In this example, the Kendall Tau (τ) is -0.5, indicating a negative monotonic correlation between the two datasets X and Y. As one variable increases, the other tends to decrease.

4. Partial Correlation Coefficient

The partial correlation coefficient measures the strength and direction of the linear relationship between two variables while controlling for the influence of one or more additional variables. The formula for the partial correlation coefficient ($r_{XY.Z}$) is as follows:

$$r_{XY.Z} = \frac{r_{XY} - r_{XZ} \cdot r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$$

where:

- $r_{XY.Z}$ is the partial correlation coefficient between variables X and Y controlling for Z.
- r_{XY} is the correlation coefficient between X and Y.
- r_{XZ} is the correlation coefficient between X and Z.
- r_{YZ} is the correlation coefficient between Y and Z.

Consider three variables: X (Height), Y (Weight), and Z (Age). We want to calculate the partial correlation coefficient between X and Y while controlling for Z.

1. **Collect Data:**
   - X: [160, 165, 155, 170, 175]
   - Y: [60, 68, 55, 72, 80]
   - Z: [25, 30, 22, 35, 40]
2. **Calculate Correlation Coefficients:**
   - r_{XY} (correlation between X and Y): Suppose $r_{XY} = 0.8$.
   - r_{XZ} (correlation between X and Z): Suppose $r_{XZ} = 0.6$.
   - r_{YZ} (correlation between Y and Z): Suppose $r_{YZ} = -0.5$.
3. **Apply the Formula:**

$$r_{XY.Z} = \frac{0.8 - 0.6 \cdot (-0.5)}{\sqrt{(1 - 0.6^2)(1 - (-0.5)^2)}}$$

4. **Calculate:**

$$r_{XY.Z} \approx \frac{0.8 + 0.3}{\sqrt{(0.64)(0.75)}} \approx \frac{1.1}{\sqrt{0.48}} \approx \frac{1.1}{0.692} \approx 1.587$$

5 Point-Biserial Correlation Coefficient

The Point-Biserial Correlation Coefficient (r_{pb}) is used to measure the strength and direction of the relationship between a binary variable and a continuous variable. Here's the formula for the point-biserial correlation coefficient:

$$r_{pb} = \frac{\bar{Y}_1 - \bar{Y}_0}{s_Y} \sqrt{\frac{n_1 n_0}{n^2 p(1-p)}}$$

where:

- $\bar{Y}_1$ is the mean of the continuous variable for the group with the binary variable equal to 1.
- $\bar{Y}_0$ is the mean of the continuous variable for the group with the binary variable equal to 0.
- s_Y is the standard deviation of the continuous variable.
- n_1 is the sample size of the group with the binary variable equal to 1.
- n_0 is the sample size of the group with the binary variable equal to 0.
- n is the total sample size.
- p is the proportion of the sample with the binary variable equal to 1.

Suppose we have a dataset with a binary variable X indicating whether students participated in an after-school program (1 for participated, 0 for did not participate), and a continuous variable Y representing their exam scores.

$X : [1, 1, 0, 1, 0, 1, 0, 0, 1, 1]$

$Y : [85, 92, 78, 88, 75, 90, 80, 82, 94, 87]$

1. **Calculate Group Means:**
   - Calculate the mean of Y for each group based on the binary variable X.
   - $\bar{Y}_1 = \frac{85+92+88+90+94+87}{6}$
   - $\bar{Y}_0 = \frac{78+75+80+82}{4}$
2. **Calculate Standard Deviation:**
   - Calculate the standard deviation (s_Y) of Y.
   - $s_Y = \sqrt{\frac{\sum_{i=1}^{10}(Y_i - \bar{Y})^2}{10}}$

3. **Calculate Sample Sizes:**
   - Calculate n_1 (number of participants) and n_0 (number of non-participants).
   - $n_1 = 6, n_0 = 4$
4. **Calculate Proportion p:**
   - Calculate the proportion of participants (p).
   - $p = \frac{n_1}{n} = \frac{6}{10} = 0.6$
5. **Apply the Formula:**

   $r_{pb} = \frac{\bar{Y}_1 - \bar{Y}_0}{s_Y} \sqrt{\frac{n_1 n_0}{n^2 p(1-p)}}$
6. **Plug in Values:**
   - Plug in the calculated values into the formula and calculate r_{pb}.

6. Cramér's V

Cramér's V is a measure of association between two categorical variables. It is an extension of the chi-square statistic and is used to quantify the strength and direction of the association between categorical variables. The formula for Cramér's V is as follows:

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(k-1, r-1)}}$$

where:

- χ^2 is the chi-square statistic from the contingency table.
- n is the total number of observations.
- k is the number of rows (categories) in the contingency table.
- r is the number of columns (categories) in the contingency table.

	B1	B2	B3
A1	20	30	10
A2	15	25	5

Let's assume that after calculations, $\chi^2 = 12, n = 100, k = 2$, and $r = 3$.

$$V = \sqrt{\frac{12}{100 \cdot \min(2-1, 3-1)}} = \sqrt{\frac{12}{100 \cdot 1}} = \sqrt{0.12} \approx 0.3464$$

	B1	B2	B3
A1	20	30	10
A2	15	25	5

2. **Calculate Row Totals (T_{row}) and Column Totals (T_{col}):**
   - Sum the values in each row and each column.
3. **Calculate Grand Total (T_{total}):**
   - Sum all the observed frequencies to get the grand total.
4. **Calculate Expected Frequencies (E_{ij}):**
   - Calculate the expected frequency for each cell using the formula: $E_{ij} = \frac{T_{\text{row}} \times T_{\text{col}}}{T_{\text{total}}}$.
5. **Calculate Chi-Square (χ^2):**
   - Sum the squared differences between observed (O_{ij}) and expected (E_{ij}) frequencies, divided by the expected frequency.
   - The formula is: $\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$.

	B1	B2	B3	T_row
A1	20	30	10	60
A2	15	25	5	45
T_col	35	55	15	T_total = 105

Eij

	B1	B2	B3
A1	34.29	53.33	14.29
A2	25.71	41.67	10.71

Calculating for each cell and summing up:

$$\chi^2 = \frac{(20-34.29)^2}{34.29} + \frac{(30-53.33)^2}{53.33} + \frac{(10-14.29)^2}{14.29} + \frac{(15-25.71)^2}{25.71} + \frac{(25-41.67)^2}{41.67} + \frac{(5-10.71)^2}{10.71}$$

After calculation, suppose $\chi^2 \approx 10.45$.

2 Data Exploration Techniques

2. 1 Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data.

The linear equation can be expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

Here:

- Y is the dependent variable.
- $X_1, X_2, \ldots, X_n$ are the independent variables.
- β_0 is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients representing the slopes.
- ε is the error term.

The goal of linear regression is to find the values of $\beta_0, \beta_1, \ldots, \beta_n$ that minimize the sum of squared differences between the predicted values ($\hat{Y}$) and the actual values (Y).

We'll use the following dataset representing the number of hours a student studies (X) and their exam scores (Y):

$$X : [2, 3, 4, 5, 6]$$
$$Y : [60, 70, 75, 85, 90]$$

We want to build a linear regression model to predict exam scores based on the number of hours studied.

Step 1: Calculate Means and Differences

Calculate the means ($\bar{X}$ and $\bar{Y}$) and differences between each data point and the mean:

$$\bar{X} = \frac{2 + 3 + 4 + 5 + 6}{5} = 4$$
$$\bar{Y} = \frac{60 + 70 + 75 + 85 + 90}{5} = 76$$

$$X_i - \bar{X} : [-2, -1, 0, 1, 2]$$
$$Y_i - \bar{Y} : [-16, -6, -1, 9, 14]$$

Step 2: Calculate Slope (β_1) and Intercept (β_0)

Calculate the slope (β_1) and intercept (β_0) using the formulas:

$$\beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$
$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

$$\beta_1 = \frac{(-2 \times -16) + (-1 \times -6) + (0 \times -1) + (1 \times 9) + (2 \times 14)}{(-2^2) + (-1^2) + (0^2) + (1^2) + (2^2)}$$
$$\beta_0 = 76 - \beta_1 \times 4$$

After calculations:

$\beta_1 \approx 5.7, \beta_0 \approx 52.4$

Step 3: Formulate the Linear Model

Substitute the values into the linear equation:

$$\text{Exam Scores} = 52.4 + 5.7 \times \text{Hours Studied}$$

Step 4: Make Predictions

Use the model to make predictions for new values of X.

2. 2 Principal Component Analysis

Principal Component Analysis (PCA) is a powerful technique used in data analysis, particularly for reducing the dimensionality of datasets while preserving crucial information. It does this by transforming the original variables into a set of new, uncorrelated variables called **principal components**

PCA's key aspects:

- Dimensionality Reduction
- Data Exploration and Visualization
- Linear Transformation
- Feature Selection
- Data Compression
- Clustering and Classification

Step 1: Standardize the Dataset

Step 2: Compute the Covariance Matrix

Step 3: Calculate Eigenvectors and Eigenvalues of the Covariance Matrix

Step 4: Sort Eigenvalues and Select Top *k* Eigenvectors

Step 5: Construct the Projection Matrix *W* from Selected Eigenvectors

Step 6: Transform the Original Dataset

Example : Dataset

Feature1	Feature2
4	11
8	4
13	5
7	14

Step 1: Standardize the Dataset

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

$$z = \frac{x - \mu}{\sigma}$$

Z:

Feature1	Feature2
-1.2344268	0.60192927
0.0	-1.08347268
1.5430335	-0.84270097
-0.3086067	1.32424438

Step 2: Compute the Covariance Matrix

covariance between two variables X and Y, which is given by:

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

where:

- X_i and Y_i are the individual sample points indexed with i,
- $\bar{X}$ and $\bar{Y}$ are the sample means of X and Y,
- n is the number of sample points.

$$C = \begin{pmatrix} \mathrm{cov}(x,x) & \mathrm{cov}(x,y) \\ \mathrm{cov}(y,x) & \mathrm{cov}(y,y) \end{pmatrix}$$

1.33333333, -0.81734138
-0.81734138, 1.33333333

Step 3: Calculate Eigenvectors and Eigenvalues of the Covariance Matrix

For a 2×2 matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, this equation can be expanded to:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \lambda \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

Expanding and rearranging gives us the characteristic equation:

$$\begin{pmatrix} a-\lambda & b \\ c & d-\lambda \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

To find λ, we set the determinant of $\begin{pmatrix} a-\lambda & b \\ c & d-\lambda \end{pmatrix}$ to zero:

$$\det(\begin{pmatrix} a-\lambda & b \\ c & d-\lambda \end{pmatrix}) = 0$$

Solving for λ gives us the eigenvalues, and substituting λ back into $Av = \lambda v$ allows us to solve for v, the eigenvectors.

Step 4: Sort Eigenvalues and Select Top *k* Eigenvectors

Step 5: Construct the Projection Matrix *W* from Selected Eigenvectors

Eigenvalues:

- $\lambda_1 = 2.15067471$
- $\lambda_2 = 0.51599195$

These values represent the magnitude of the variance along the directions defined by their corresponding eigenvectors.

Eigenvectors:

- For $\lambda_1 = 2.15067471$, the eigenvector is $\begin{pmatrix} 0.70710678 \\ -0.70710678 \end{pmatrix}$.
- For $\lambda_2 = 0.51599195$, the eigenvector is $\begin{pmatrix} 0.70710678 \\ 0.70710678 \end{pmatrix}$.

$$W = \begin{bmatrix} 0.71 & 0.71 \\ -0.71 & 0.71 \end{bmatrix}$$

Step 6: Transform the Original Dataset

To derive the new dataset, we project the original dataset X onto our selected principal components by multiplying X with the projection matrix W:

$$Y = XW$$

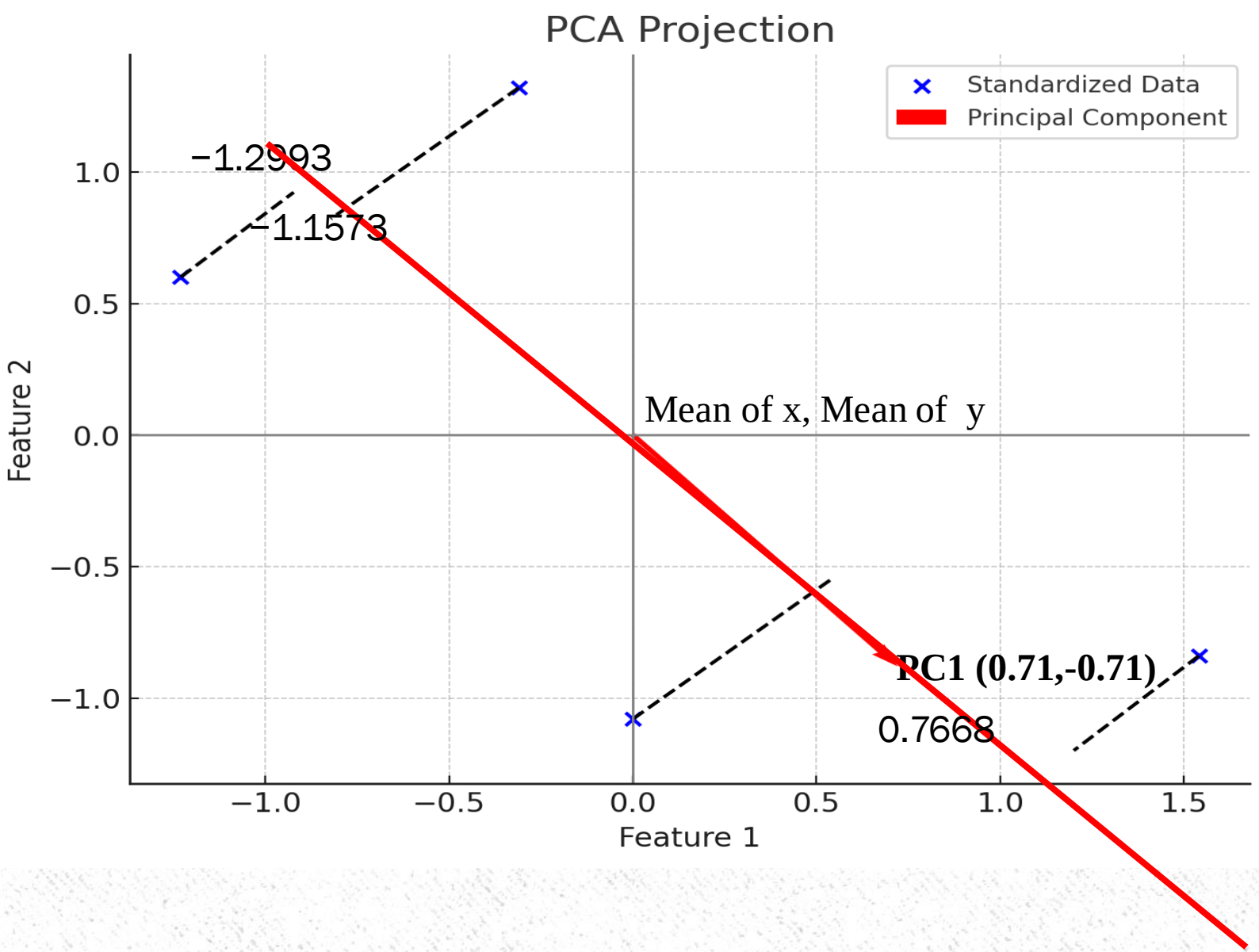
This operation transforms the original data X into the new dataset Y in the reduced dimensional space:

For example, applying this to the first row of X:

$$\begin{bmatrix} -1.23 & 0.60 \end{bmatrix} \begin{bmatrix} 0.71 & 0.71 \\ -0.71 & 0.71 \end{bmatrix} = \begin{bmatrix} Y_{11} & Y_{12} \end{bmatrix}$$

Assuming Y_{11} and Y_{12} are the new coordinates for the first observation in the transformed space.

$$Y = \begin{pmatrix} -1.2993 \\ 0.7668 \\ 1.6898 \\ -1.1573 \end{pmatrix}$$



2. 3 Clustering Techniques

- Clustering is a technique used in unsupervised machine learning to group a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups.
- Several clustering techniques are widely used, each with its own methodology and applications

1. K-Means Clustering

Key Steps:

1. **Initialization:** Randomly select k centroids.
2. **Assignment:** Assign each data point to the nearest centroid based on distance (usually Euclidean distance).
3. **Update:** Update the centroid of each cluster to be the mean of all points in the cluster.
4. **Repeat:** Repeat the assignment and update steps until convergence (when centroids do not significantly change between iterations).

Formula:

- **Centroid Update:** $C_i = \frac{1}{|S_i|} \sum_{x \in S_i} x$, where C_i is the centroid of cluster i, S_i is the set of all points in cluster i, and x are the points.

Application: Customer segmentation in marketing to identify groups of customers with similar behavior for targeted marketing campaigns.

- **Dataset *X*:** (1,2),(1.5,1.8),(5,8),(8,8),(1,0.6),(9,11)
- **Step 1: Initialization**
- Choose *K*=2 initial centroids randomly from the dataset:
- Centroid 1 (*C*1) = (1,2)
- Centroid 2 (*C*2) = (5,8)
- **Step 2: Assignment**
- Assign each point to the nearest centroid. Suppose we use the Euclidean distance for this purpose.

The distance between a point $p = (x_1, y_1)$ and a centroid $c = (x_2, y_2)$ is:

$$d(p, c) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

- After calculating the distances, we assign the points to the nearest centroids. For simplicity, let's say the assignments are:
- Points closer to *C*1: (1,2),(1.5,1.8),(1,0.6)
- Points closer to *C*2: (5,8),(8,8),(9,11)
- **Step 3: Update**
- Update the centroids by calculating the mean of the points in each cluster:
- New *C*1 = Mean of (1,2),(1.5,1.8),(1,0.6) = (1.17,1.47)
- New *C*2 = Mean of (5,8),(8,8),(9,11)= (7.33,9)
- **Step 4: Repeat**
- Repeat the Assignment and Update steps. With the updated centroids, points will be reassigned, and centroids updated again. This process continues until the centroids stabilize and no longer change significantly.
- **Final Clusters**
- After convergence, we might end up with final clusters, for example:
- Cluster 1 around centroid (1.17,1.47)with points (1,2),(1.5,1.8),(1,0.6)
- Cluster 2 around centroid (7.33,9)with points (5,8),(8,8),(9,11)

2. Hierarchical Clustering

1. **Start:** Treat each data point as a single cluster.
2. **Find Closest Clusters:** Calculate the distance between every pair of clusters and identify the pair of clusters that are closest together.
3. **Merge Clusters:** Merge the two closest clusters into one cluster.
4. **Repeat:** Repeat steps 2 and 3 until all data points are in a single cluster or a stopping criterion is met.

Distance Measures:

- **Single Linkage:** $d(A, B) = \min\{\|a - b\| \mid a \in A, b \in B\}$
- **Complete Linkage:** $d(A, B) = \max\{\|a - b\| \mid a \in A, b \in B\}$
- **Average Linkage:** $d(A, B) = \frac{1}{|A| \cdot |B|} \sum_{a \in A} \sum_{b \in B} \|a - b\|$

Application: Gene expression data analysis where researchers can find groups of genes that exhibit similar expression patterns.

3. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

1. **Specify Parameters:** Choose ε (the radius of a neighborhood around a point) and $MinPts$ (the minimum number of points required to form a dense region).
2. **Classify Points:** Label each point as a core point, border point, or noise, based on the number of neighbors within ε and $MinPts$.
3. **Form Clusters:** For each core point, if it is not already assigned to a cluster, create a new cluster, then iteratively add all its density-reachable points to the cluster.

Core Point Condition:

A point p is a core point if $|N_\varepsilon(p)| \geq MinPts$, where $N_\varepsilon(p)$ is the ε-neighborhood of p.

Application: Identifying regions of high density in spatial data, such as identifying regions of a forest that have a high density of certain tree species.

4. Mean Shift Clustering

1. **Choose Bandwidth:** The bandwidth parameter h is critical and determines the size of the region to search through.
2. **Compute Mean Shift:** For each point, compute the mean shift vector that points towards the direction of the highest density of points.
3. **Update Points:** Move each point towards the region of higher density by updating its position to the location of the mean shift vector.
4. **Repeat:** Repeat step 2 and 3 until convergence.

Mean Shift Vector Formula:

$$M(x) = \frac{\sum_{x_i \in N(x)} K(x_i - x) x_i}{\sum_{x_i \in N(x)} K(x_i - x)}$$

where $N(x)$ is the neighborhood of x within bandwidth h, x_i are the points within $N(x)$, and K is the kernel function used to weigh points.

Application: Image segmentation and tracking where mean shift can be used to locate objects in an image by finding dense regions of pixels.

- Imagine we have a small dataset of 1D points: [1,2,3,8,9,10][1,2,3,8,9,10]. We want to cluster these points using Mean Shift with a bandwidth (window size) of 4.
- Step 1: Initialization
- Choose initial centroids. In Mean Shift, every point can initially be considered a centroid. So, we start with centroids at every point: ={1,2,3,8,9,10}*C*={1,2,3,8,9,10}.
- Step 2: Compute Mean Shift and Update Centroids With a bandwidth of 4, for each point, we find points within this range, compute the mean, and update the centroid to this mean. This process is iterated until convergence.
- For the centroid at 1: The points within the bandwidth are [1,2,3]. The mean is 1+2+3/3=2.
- For the centroid at 2: It also covers [1,2,3], so its mean will also be 2.
- For the centroid at 3: Same as above, the mean will be 2.
- For the centroid at 8: The points within the bandwidth are [8,9,10]. The mean is 8+9+10/3=9.
- Similarly, centroids at 9 and 10 will also move to 9.
- After one iteration, our centroids are updated to [2,2,2,9,9,9].
- Step 3: Repeat Until Convergence
- The process of updating centroids is repeated until no centroids change their locations significantly. In our simple example, the centroids have converged after one iteration to two distinct values: 2 and 9.
- Final Clusters
- Given the convergence to centroids at 2 and 9, we have two clusters:
- Cluster 1: [1,2,3]centered around 2.
- Cluster 2: [8,9,10] centered around 9.

Inclass 4

4.1 examine the relationship between exam scores (continuous variable) and attendance (dichotomous variable: attend class = 1, absent = 0) using Point-Biserial Correlation Coefficient Method

Students who attended class (n=1): [80, 85, 78, 90, 95]
Students who were absent (n=0): [65, 70, 60, 75]

4.2 Find the association using Cramér's V Method

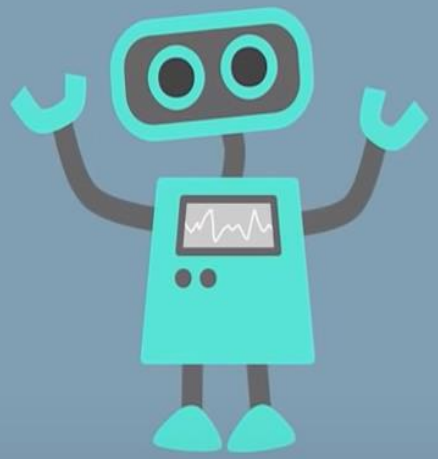
	Packaging A	Packaging B
Men	20	30
Women	15	25

Self Assessments and Activities

Here are some assessments you can use to evaluate students understanding of DATA EXPLORATION AND PREPARATION.

- Quizzes
- In class Participation
- Write up assessments

Did You Know?



A statistician and data scientist known for creating several widely used R packages for data analysis and visualization

HADLEY WICKHAM

Summary

Outcomes: students able to

a. Analyze and interpret descriptive statistics to gain insights.

Terminal Questions

1. Explain Descriptive Statistics for data analysis
2. Explain Data Distribution Analysis
3. Explain Correlation Analysis
4. Explain Linear Regression
5. Explain Principal Component Analysis
6. Explain Clustering Techniques

Reference Links

TEXTBOOKS:

- Python for Data Analysis, Data Wrangling with Numpy, Pandas and Jupyter, Third Edition, O'Reilly.
- Hands on Exploratory Data Analysis using Python, Packt.

REFERENCES:

- Peter Bruce, Andrew Bruce and Peter Gedeck, Practical Statistics for Data Scientist, O'Reilly.
- Avinash Navlani, Armando Fandango and Ivan Idris, Python Data Analysis, Third Edition, Packt.

Thank you