### **1. Fill-in-the-Blank\**

**Question:** To load a CSV file named data.csv into a Pandas DataFrame, use the function \_\_\_\_\_\_.  
**Answer:** pd.read\_csv()

### **2. Multiple Choice**

**Question:** What method can you use to handle missing values in a Pandas DataFrame?  
a) df.drop()  
b) df.fillna()  
c) df.describe()  
d) df.mean()  
**Answer:** b) df.fillna()

### **3. Fill-in-the-Blank**

**Question:** Write Python code to add a new column 'Revenue' to a DataFrame by multiplying the columns 'Price' and 'Quantity'.  
**Answer:**

python

Copy code

df['Revenue'] = df['Price'] \* df['Quantity']

### **4. Fill-in-the-Blank**

**Question:** Write Python code to save a DataFrame to a CSV file named output.csv.  
**Answer:**

python

Copy code

df.to\_csv('output.csv', index=False)

### **5. Multiple Choice**

**Question:** What does a cosine similarity of 1 between two vectors indicate?  
a) The vectors are orthogonal.  
b) The vectors point in the exact same direction.  
c) The vectors point in opposite directions.  
d) The vectors are identical in magnitude.  
**Answer:** b) The vectors point in the exact same direction.

### **6. Fill-in-the-Blank**

**Question:** Write Python code to calculate cosine similarity between all rows of a DataFrame using Scikit-Learn.  
**Answer:**

python

Copy code

from sklearn.metrics.pairwise import cosine\_similarity

similarity\_matrix = cosine\_similarity(df.values)

### **7. Fill-in-the-Blank**

**Question:** Write Python code to find the cosine distance between two vectors uuu and vvv.  
**Answer:**

python

Copy code

cosine\_distance = 1 - (np.dot(u, v) / (np.linalg.norm(u) \* np.linalg.norm(v)))

### **8. Multiple Choice**

**Question:** Which library provides the LinearRegression class for building regression models?  
a) NumPy  
b) SciPy  
c) Pandas  
d) Scikit-Learn  
**Answer:** d) Scikit-Learn

### **9. Fill-in-the-Blank**

**Question:** Write Python code to fit a simple linear regression model to a DataFrame with columns 'X' and 'Y'.  
**Answer:**

python

Copy code

from sklearn.linear\_model import LinearRegression

model = LinearRegression()

model.fit(df[['X']], df['Y'])

### **10. Fill-in-the-Blank**

**Question:** Write Python code to calculate R2R^2R2 (coefficient of determination) for a regression model in Scikit-Learn.  
**Answer:**

python

Copy code

r2 = model.score(X\_test, y\_test)

### **11. Multiple Choice**

**Question:** Which matrix is used to compute the eigenvectors in PCA?  
a) The original data matrix  
b) The covariance matrix  
c) The diagonal matrix  
d) The singular value matrix  
**Answer:** b) The covariance matrix

### 

### 

### 

### **12. Fill-in-the-Blank**

**Question:** Write Python code to standardize data before applying PCA using Scikit-Learn.  
**Answer:**

python

Copy code

from sklearn.preprocessing import StandardScaler

standardized\_data = StandardScaler().fit\_transform(df)

### **13. Fill-in-the-Blank**

**Question:** Write Python code to get the first principal component from PCA in Scikit-Learn.  
**Answer:**

python

Copy code

first\_component = pca.components\_[0]

### **14. Multiple Choice**

**Question:** Which statistical test is used to compare the means of two paired samples?  
a) Independent t-test  
b) Paired t-test  
c) ANOVA  
d) Chi-square test  
**Answer:** b) Paired t-test

### 

### 

### 

### 

### 

### **15. Fill-in-the-Blank**

**Question:** Write Python code to calculate the confidence interval for a one-sample t-test.  
**Answer:**

python

Copy code

from scipy.stats import t

mean = sample.mean()

sem = sample.std() / np.sqrt(len(sample))

confidence\_interval = t.interval(0.95, len(sample) - 1, loc=mean, scale=sem)

### **16. Multiple Choice**

**Question:** What is the range of values for cosine similarity, and what do the extremes represent?  
a) -1 to 1: 1 (Same direction), 0 (Orthogonal), -1 (Opposite direction).  
b) 0 to 1: 0 (Orthogonal), 1 (Same direction).  
c) 0 to -1: -1 (Same direction), 0 (Orthogonal).  
**Answer:** a) -1 to 1: 1 (Same direction), 0 (Orthogonal), -1 (Opposite direction).

### **17. Fill-in-the-Blank**

**Question:** Write Python code to handle missing data in a DataFrame by filling with the median.  
**Answer:**

python

Copy code

df.fillna(df.median(), inplace=True)

### 

### 

### 

### **18. Short Answer**

**Question:** Explain why centering data is important before applying PCA.  
**Answer:** Centering ensures the first principal component captures maximum variance by removing the mean.

### **19. Short Answer**

**Question:** What does the p-value represent in hypothesis testing?  
**Answer:** The p-value represents the probability of observing the data assuming the null hypothesis is true.

### **20. Fill-in-the-Blank**

**Question:** Write Python code to calculate eigenvalues and eigenvectors of a covariance matrix.  
**Answer:**

python

Copy code

eigenvalues, eigenvectors = np.linalg.eig(cov\_matrix)

### **21. Short Answer**

**Question:** What does the intercept of a regression line represent?  
**Answer:** The intercept is the predicted value of y when x = 0.

### **22. Fill-in-the-Blank**

**Question:** Write Python code to compute the slope of a regression line.  
**Answer:**

python

Copy code

slope = np.cov(x, y, bias=True)[0, 1] / np.var(x)

### **23. Fill-in-the-Blank**

**Question:** Write Python code to calculate the Mean Squared Error (MSE) in Python.  
**Answer:**

python

Copy code

from sklearn.metrics import mean\_squared\_error

mse = mean\_squared\_error(y\_true, y\_pred)

### **24. Short Answer**

**Question:** What is the primary purpose of PCA?  
**Answer:** To reduce dimensionality while retaining as much variance as possible.

### **25. Fill-in-the-Blank**

**Question:** Write Python code to apply PCA and retain only the first two components.  
**Answer:**

python

Copy code

from sklearn.decomposition import PCA

pca = PCA(n\_components=2)

reduced\_data = pca.fit\_transform(data)

### 

### 

### **26. Multiple Choice**

**Question:** What does a residual plot with a clear pattern suggest about the regression model?  
a) The regression model is a good fit.  
b) The regression model is not a good fit.  
c) The residuals are evenly distributed.  
d) The residuals are uncorrelated.  
**Answer:** b) The regression model is not a good fit.

### **27. Fill-in-the-Blank**

**Question:** Write Python code to calculate the predicted yyy-values for a regression line with slope m=3m = 3m=3, intercept b=7b = 7b=7, and x=[1,2,3]x = [1, 2, 3]x=[1,2,3].  
**Answer:**

python

Copy code

y\_pred = [3 \* x\_i + 7 for x\_i in [1, 2, 3]]

### **28. Short Answer**

**Question:** Why is dimensionality reduction important in machine learning?  
**Answer:** Dimensionality reduction simplifies models, reduces computation time, and minimizes overfitting while retaining essential data patterns.

### **29. Multiple Choice**

**Question:** What does the scree plot in PCA analysis represent?  
a) The cumulative variance explained by each component.  
b) The eigenvalues corresponding to each component.  
c) The explained variance ratio of each component.  
d) The data distribution in the principal components.  
**Answer:** b) The eigenvalues corresponding to each component.

### **30. Fill-in-the-Blank**

**Question:** Write Python code to check for missing values in a DataFrame.  
**Answer:**

python

Copy code

missing\_values = df.isnull().sum()

### 

### **31. Short Answer**

**Question:** What is the role of singular vectors in SVD?  
**Answer:** Singular vectors represent the direction of maximum variance in the data for each singular value.

### **32. Multiple Choice**

**Question:** In hypothesis testing, what is the null hypothesis (H0H\_0H0​)?  
a) The hypothesis that there is no effect or difference.  
b) The hypothesis that there is an effect or difference.  
c) The level of significance of the test.  
d) The alternative hypothesis.  
**Answer:** a) The hypothesis that there is no effect or difference.

### **33. Fill-in-the-Blank**

**Question:** Write Python code to calculate the sum of squared residuals in a regression model.  
**Answer:**

python

Copy code

ssr = sum((y - y\_pred)\*\*2)

### **34. Multiple Choice**

**Question:** Which Python library provides a direct implementation of PCA?  
a) NumPy  
b) SciPy  
c) Scikit-Learn  
d) Pandas  
**Answer:** c) Scikit-Learn

### 

### 

### **35. Fill-in-the-Blank**

**Question:** Write Python code to perform a two-sample t-test using SciPy.  
**Answer:**

python

Copy code

from scipy.stats import ttest\_ind

t\_stat, p\_value = ttest\_ind(sample1, sample2)

### **36. Multiple Choice**

**Question:** What does a p-value less than the significance level (α\alphaα) indicate?  
a) Fail to reject the null hypothesis.  
b) Reject the null hypothesis.  
c) The null hypothesis is true.  
d) The alternative hypothesis is false.  
**Answer:** b) Reject the null hypothesis.

### **37. Short Answer**

**Question:** What is the primary difference between PCA and SVD?  
**Answer:** PCA reduces dimensionality by retaining variance, while SVD decomposes a matrix into its constituent components.

### **38. Fill-in-the-Blank**

**Question:** Write Python code to compute the explained variance ratio in PCA.  
**Answer:**

python

Copy code

explained\_variance = pca.explained\_variance\_ratio\_

### 

### **39. Multiple Choice**

**Question:** Which method in Pandas displays a summary of statistics for numerical columns?  
a) df.info()  
b) df.describe()  
c) df.mean()  
d) df.stats()  
**Answer:** b) df.describe()

### **40. Fill-in-the-Blank**

**Question:** Write Python code to calculate the mean absolute error (MAE) for a regression model.  
**Answer:**

python

Copy code

from sklearn.metrics import mean\_absolute\_error

mae = mean\_absolute\_error(y\_true, y\_pred)

### **41. Short Answer**

**Question:** What is the primary purpose of using statistical tests in machine learning?  
**Answer:** To validate hypotheses, evaluate model performance, and ensure the reliability of results.

### **1. Fill-in-the-Blank**

**Question:** To load a CSV file named data.csv into a Pandas DataFrame, use the function \_\_\_\_\_\_.  
**Answer:** pd.read\_csv()

### **2. Multiple Choice**

**Question:** What method can you use to handle missing values in a Pandas DataFrame?  
a) df.drop()  
b) df.fillna()  
c) df.describe()  
d) df.mean()  
**Answer:** b) df.fillna()

### **3. Fill-in-the-Blank**

**Question:** Write Python code to add a new column 'Revenue' to a DataFrame by multiplying the columns 'Price' and 'Quantity'.  
**Answer:**

python

Copy code

df['Revenue'] = df['Price'] \* df['Quantity']

### **4. Fill-in-the-Blank**

**Question:** Write Python code to save a DataFrame to a CSV file named output.csv.  
**Answer:**

python

Copy code

df.to\_csv('output.csv', index=False)

### **5. Multiple Choice**

**Question:** What does a cosine similarity of 1 between two vectors indicate?  
a) The vectors are orthogonal.  
b) The vectors point in the exact same direction.  
c) The vectors point in opposite directions.  
d) The vectors are identical in magnitude.  
**Answer:** b) The vectors point in the exact same direction.

### **6. Fill-in-the-Blank**

**Question:** Write Python code to calculate cosine similarity between all rows of a DataFrame using Scikit-LLearn.  
**Answer:**

python

Copy code

from sklearn.metrics.pairwise import cosine\_similarity

similarity\_matrix = cosine\_similarity(df.values)

### **7. Fill-in-the-Blank**

**Question:** Write Python code to find the cosine distance between two vectors u and v.  
**Answer:**

python

Copy code

cosine\_distance = 1 - (np.dot(u, v) / (np.linalg.norm(u) \* np.linalg.norm(v)))

### 

### 

### 

### 

### **8. Multiple Choice**

**Question:** Which library provides the LinearRegression class for building regression models?  
a) NumPy  
b) SciPy  
c) Pandas  
d) Scikit-Learn  
**Answer:** d) Scikit-Learn

### **9. Fill-in-the-Blank**

**Question:** Write Python code to fit a simple linear regression model to a DataFrame with columns 'X' and 'Y'.  
**Answer:**

python

Copy code

from sklearn.linear\_model import LinearRegression

model = LinearRegression()

model.fit(df[['X']], df['Y'])

### **10. Fill-in-the-Blank**

**Question:** Write Python code to calculate R2R^2R2 (coefficient of determination) for a regression model in Scikit-Learn.  
**Answer:**

python

Copy code

r2 = model.score(X\_test, y\_test)

### 

### 

### 

### 

### **11. Multiple Choice**

**Question:** Which matrix is used to compute the eigenvectors in PCA?  
a) The original data matrix  
b) The covariance matrix  
c) The diagonal matrix  
d) The singular value matrix  
**Answer:** b) The covariance matrix

### **12. Fill-in-the-Blank**

**Question:** Write Python code to standardize data before applying PCA using Scikit-Learn.  
**Answer:**

python

Copy code

from sklearn.preprocessing import StandardScaler

standardized\_data = StandardScaler().fit\_transform(df)

### **13. Fill-in-the-Blank**

**Question:** Write Python code to get the first principal component from PCA in Scikit-Learn.  
**Answer:**

python

Copy code

first\_component = pca.components\_[0]

### **14. Multiple Choice**

**Question:** Which statistical test is used to compare the means of two paired samples?  
a) Independent t-test  
b) Paired t-test  
c) ANOVA  
d) Chi-square test  
**Answer:** b) Paired t-test

### **15. Fill-in-the-Blank**

**Question:** Write Python code to calculate the confidence interval for a one-sample t-test.  
**Answer:**

python

Copy code

from scipy.stats import t

mean = sample.mean()

sem = sample.std() / np.sqrt(len(sample))

confidence\_interval = t.interval(0.95, len(sample) - 1, loc=mean, scale=sem)

### **16. Fill-in-the-Blank**

**Question:** Write Python code to calculate the mean of a column named 'Marks' in a Pandas DataFrame.  
**Answer:**

python

Copy code

mean\_marks = df['Marks'].mean()

### **17. Fill-in-the-Blank**

**Question:** Write Python code to replace all missing values in the 'Salary' column with the average salary.  
**Answer:**

python

Copy code

df['Salary'].fillna(df['Salary'].mean(), inplace=True)

### 

### 

### **18. Fill-in-the-Blank**

**Question:** Write Python code to calculate the slope of a regression line given points (1, 2), (2, 4), and (3, 6) using NumPy.  
**Answer:**

python

Copy code

import numpy as np

x = np.array([1, 2, 3])

y = np.array([2, 4, 6])

slope = np.cov(x, y, bias=True)[0, 1] / np.var(x, ddof=0)

### **19. Short Answer**

**Question:** Define Mean Squared Error (MSE) in simple terms.  
**Answer:** The Mean Squared Error measures how far the predicted values are from the actual values on average, squared.

### **20. Fill-in-the-Blank**

**Question:** Write Python code to apply PCA and extract the first principal component using Scikit-Learn.  
**Answer:**

python

Copy code

from sklearn.decomposition import PCA

pca = PCA(n\_components=1)

principal\_component = pca.fit\_transform(df)

### 

### 

### 

### **21. Fill-in-the-Blank**

**Question:** Write Python code to calculate the residuals in a linear regression model.  
**Answer:**

python

Copy code

residuals = y\_true - y\_pred

### **22. Short Answer**

**Question:** What is the main purpose of PCA in data analysis?  
**Answer:** To reduce the dimensionality of data while retaining most of its variance.

### **23. Fill-in-the-Blank**

**Question:** Write Python code to compute the explained variance ratio in PCA using Scikit-Learn.  
**Answer:**

python

Copy code

explained\_variance = pca.explained\_variance\_ratio\_

### **24. Multiple Choice**

**Question:** Which method in Pandas displays a summary of statistics for numerical columns?  
a) df.info()  
b) df.describe()  
c) df.mean()  
d) df.stats()  
**Answer:** b) df.describe()

### 

### 

### **25. Fill-in-the-Blank**

**Question:** Write Python code to calculate the Mean Absolute Error (MAE) in Python.  
**Answer:**

python

Copy code

from sklearn.metrics import mean\_absolute\_error

mae = mean\_absolute\_error(y\_true, y\_pred)

### **26. Multiple Choice**

**Question:** What does the scree plot in PCA analysis represent?  
a) The cumulative variance explained by each component.  
b) The eigenvalues corresponding to each component.  
c) The explained variance ratio of each component.  
d) The data distribution in the principal components.  
**Answer:** b) The eigenvalues corresponding to each component.

### **27. Short Answer**

**Question:** What is the primary difference between PCA and SVD?  
**Answer:** PCA reduces dimensionality by retaining variance, while SVD decomposes a matrix into its constituent components.

### **28. Fill-in-the-Blank**

**Question:** Write Python code to calculate the p-value for a one-sample t-test using SciPy.  
**Answer:**

python

Copy code

from scipy.stats import ttest\_1samp

t\_stat, p\_value = ttest\_1samp(sample, population\_mean)

### 

### **29. Multiple Choice**

**Question:** In hypothesis testing, what is the null hypothesis (H0H\_0H0​)?  
a) The hypothesis that there is no effect or difference.  
b) The hypothesis that there is an effect or difference.  
c) The level of significance of the test.  
d) The alternative hypothesis.  
**Answer:** a) The hypothesis that there is no effect or difference.

### **30. Fill-in-the-Blank**

**Question:** Write Python code to calculate the sum of squared residuals in a regression model.  
**Answer:**

python

Copy code

ssr = sum((y\_true - y\_pred)\*\*2)

### **31. Short Answer**

**Question:** What is the role of singular vectors in SVD?  
**Answer:** Singular vectors represent the direction of maximum variance in the data for each singular value.

### **32. Fill-in-the-Blank**

**Question:** Write Python code to standardize a DataFrame and apply PCA with 2 components using Scikit-Learn.  
**Answer:**

python

Copy code

from sklearn.preprocessing import StandardScaler

from sklearn.decomposition import PCA

scaled\_data = StandardScaler().fit\_transform(df)

pca = PCA(n\_components=2)

principal\_components = pca.fit\_transform(scaled\_data)

### **33. Fill-in-the-Blank**

**Question:** Write Python code to calculate the cosine similarity between two vectors using Scikit-Learn.  
**Answer:**

python

Copy code

from sklearn.metrics.pairwise import cosine\_similarity

similarity = cosine\_similarity([u], [v])

### **34. Multiple Choice**

**Question:** What is the purpose of centering data before PCA?  
a) To scale the data to a specific range.  
b) To normalize the variance of the data.  
c) To remove the mean and align data with the origin.  
d) To increase the explained variance.  
**Answer:** c) To remove the mean and align data with the origin.

### **35. Fill-in-the-Blank**

**Question:** Write Python code to visualize the explained variance ratio from PCA in a bar chart.  
**Answer:**

python

Copy code

import matplotlib.pyplot as plt

plt.bar(range(len(pca.explained\_variance\_ratio\_)), pca.explained\_variance\_ratio\_)

plt.xlabel('Principal Components')

plt.ylabel('Explained Variance Ratio')

plt.show()

### **36. Multiple Choice**

**Question:** What does a high p-value in a hypothesis test indicate?  
a) Strong evidence against the null hypothesis.  
b) Weak evidence against the null hypothesis.  
c) The null hypothesis is definitely true.  
d) The alternative hypothesis is true.  
**Answer:** b) Weak evidence against the null hypothesis.

### **37. Short Answer**

**Question:** What does the first principal component represent in PCA?  
**Answer:** It represents the direction of maximum variance in the data.

### **38. Fill-in-the-Blank**

**Question:** Write Python code to generate a scatter plot of the first two principal components after PCA.  
**Answer:**

python

Copy code

plt.scatter(principal\_components[:, 0], principal\_components[:, 1])

plt.xlabel('Principal Component 1')

plt.ylabel('Principal Component 2')

plt.title('PCA Scatter Plot')

plt.show()

### **39. Multiple Choice**

**Question:** What does the Mean Squared Error (MSE) measure in regression analysis?  
a) The variance of the residuals.  
b) The accuracy of the predicted values.  
c) The average squared difference between observed and predicted values.  
d) The slope of the regression line.  
**Answer:** c) The average squared difference between observed and predicted values.

### **40. Short Answer**

**Question:** What does the term "dimensionality reduction" mean in the context of PCA?  
**Answer:** It refers to the process of reducing the number of features in a dataset while retaining as much variance as possible.

### **41. Fill-in-the-Blank**

**Question:** Write Python code to calculate the R-squared value for a regression model manually.  
**Answer:**

python

Copy code

ss\_total = sum((y\_true - y\_true.mean())\*\*2)

ss\_residual = sum((y\_true - y\_pred)\*\*2)

r\_squared = 1 - (ss\_residual / ss\_total)

**21. Multiple Choice**

**Question:** What does a residual plot with a clear pattern suggest about the regression model?  
a) The regression model is a good fit.  
b) The regression model is not a good fit.  
c) The residuals are evenly distributed.  
d) The residuals are uncorrelated.  
**Answer:** b) The regression model is not a good fit.

### **22. Fill-in-the-Blank**

**Question:** Write Python code to calculate the predicted yyy-values for a regression line with slope m=3m = 3m=3, intercept b=7b = 7b=7, and x=[1,2,3]x = [1, 2, 3]x=[1,2,3].  
**Answer:**

python

Copy code

y\_pred = [3 \* x\_i + 7 for x\_i in [1, 2, 3]]

### **23. Short Answer**

**Question:** Why is dimensionality reduction important in machine learning?  
**Answer:** Dimensionality reduction simplifies models, reduces computation time, and minimizes overfitting while retaining essential data patterns.

### **24. Multiple Choice**

**Question:** What does the scree plot in PCA analysis represent?  
a) The cumulative variance explained by each component.  
b) The eigenvalues corresponding to each component.  
c) The explained variance ratio of each component.  
d) The data distribution in the principal components.  
**Answer:** b) The eigenvalues corresponding to each component.

### **25. Fill-in-the-Blank**

**Question:** Write Python code to check for missing values in a DataFrame.  
**Answer:**

python

Copy code

missing\_values = df.isnull().sum()

### **26. Short Answer**

**Question:** What is the role of singular vectors in SVD?  
**Answer:** Singular vectors represent the direction of maximum variance in the data for each singular value.

### 

### 

### **27. Multiple Choice**

**Question:** In hypothesis testing, what is the null hypothesis (H0H\_0H0​)?  
a) The hypothesis that there is no effect or difference.  
b) The hypothesis that there is an effect or difference.  
c) The level of significance of the test.  
d) The alternative hypothesis.  
**Answer:** a) The hypothesis that there is no effect or difference.

### **28. Fill-in-the-Blank**

**Question:** Write Python code to calculate the sum of squared residuals in a regression model.  
**Answer:**

python

Copy code

ssr = sum((y - y\_pred)\*\*2)

### **29. Multiple Choice**

**Question:** Which Python library provides a direct implementation of PCA?  
a) NumPy  
b) SciPy  
c) Scikit-Learn  
d) Pandas  
**Answer:** c) Scikit-Learn

### **30. Fill-in-the-Blank**

**Question:** Write Python code to perform a two-sample t-test using SciPy.  
**Answer:**

python

Copy code

from scipy.stats import ttest\_ind

t\_stat, p\_value = ttest\_ind(sample1, sample2)

### **31. Multiple Choice**

**Question:** What does a p-value less than the significance level (α\alphaα) indicate?  
a) Fail to reject the null hypothesis.  
b) Reject the null hypothesis.  
c) The null hypothesis is true.  
d) The alternative hypothesis is false.  
**Answer:** b) Reject the null hypothesis.

### **32. Short Answer**

**Question:** What is the primary difference between PCA and SVD?  
**Answer:** PCA reduces dimensionality by retaining variance, while SVD decomposes a matrix into its constituent components.

### **33. Fill-in-the-Blank**

**Question:** Write Python code to compute the explained variance ratio in PCA.  
**Answer:**

python

Copy code

explained\_variance = pca.explained\_variance\_ratio\_

### **34. Multiple Choice**

**Question:** Which method in Pandas displays a summary of statistics for numerical columns?  
a) df.info()  
b) df.describe()  
c) df.mean()  
d) df.stats()  
**Answer:** b) df.describe()

### 

### 

### **35. Fill-in-the-Blank**

**Question:** Write Python code to calculate the mean absolute error (MAE) for a regression model.  
**Answer:**

python

Copy code

from sklearn.metrics import mean\_absolute\_error

mae = mean\_absolute\_error(y\_true, y\_pred)

### **36. Short Answer**

**Question:** What is the primary purpose of using statistical tests in machine learning?  
**Answer:** To validate hypotheses, evaluate model performance, and ensure the reliability of results.

-------------------------------------------------------------------------------------------------------------------------------

**Question:** To load a CSV file named data.csv into a Pandas DataFrame, use the function \_\_\_\_\_\_.  
**Answer:** pd.read\_csv()

**Question:** Write Python code to load a CSV file named data.csv and display its first 5 rows.  
**Answer:**python  
Copy code  
import pandas as pd

df = pd.read\_csv('data.csv')

print(df.head())

**Question:** Which method is used to display the data types of each column in a Pandas DataFrame?  
**Answer:** df.dtypes

**Question:** Write Python code to count the number of missing values in each column of a DataFrame.  
**Answer:**python  
Copy code  
missing\_values = df.isnull().sum()

**Question:** Which method is used to fill missing values in a column with a specific value?  
**Answer:** df.fillna(value, inplace=True)

**Question:** Write the formula for cosine similarity between two vectors uuu and vvv.  
**Answer:**Cosine Similarity=u⋅v∣∣u∣∣×∣∣v∣∣\text{Cosine Similarity} = \frac{u \cdot v}{||u|| \times ||v||}Cosine Similarity=∣∣u∣∣×∣∣v∣∣u⋅v​

**Question:** Write Python code to calculate cosine similarity between two vectors uuu and vvv using NumPy.  
**Answer:**python  
Copy code  
import numpy as np

cosine\_similarity = np.dot(u, v) / (np.linalg.norm(u) \* np.linalg.norm(v))

**Question:** What does a cosine similarity value of -1 indicate about two vectors?  
**Answer:** The two vectors are completely opposite in direction.

**Question:** Write the equation of a linear regression line.  
**Answer:** y=mx+by = mx + by=mx+b, where mmm is the slope and bbb is the intercept.

**Question:** Write Python code to calculate the slope of a regression line given arrays xxx and yyy.  
**Answer:**python  
Copy code  
slope = np.cov(x, y, bias=True)[0, 1] / np.var(x, ddof=0)

**Question:** Write Python code to calculate the Mean Squared Error (MSE) of a regression model.  
**Answer:**python  
Copy code  
from sklearn.metrics import mean\_squared\_error

mse = mean\_squared\_error(y\_true, y\_pred)

**Question:** Define the intercept of a regression line in simple terms.  
**Answer:** The intercept is the predicted value of yyy when x=0x = 0x=0.

**Question:** What is the primary goal of PCA?  
**Answer:** To reduce the dimensionality of the dataset while retaining as much variance as possible.

**Question:** Write Python code to apply PCA and retain the first two principal components.  
**Answer:**python  
Copy code  
from sklearn.decomposition import PCA

pca = PCA(n\_components=2)

reduced\_data = pca.fit\_transform(data)

**Question:** What does the scree plot in PCA represent?  
**Answer:** The scree plot shows the eigenvalues corresponding to each principal component.

**Question:** Write Python code to compute the explained variance ratio in PCA using Scikit-Learn.  
**Answer:**python  
Copy code  
explained\_variance = pca.explained\_variance\_ratio\_

**Question:** Write Python code to perform a one-sample t-test in Python using SciPy.  
**Answer:**python  
Copy code  
from scipy.stats import ttest\_1samp

t\_stat, p\_value = ttest\_1samp(sample, popmean=50)

**Question:** What does a p-value represent in hypothesis testing?  
**Answer:** The p-value represents the probability of obtaining the observed results assuming the null hypothesis is true.

**Question:** Write Python code to calculate the chi-square test of independence in SciPy.  
**Answer:**python  
Copy code  
from scipy.stats import chi2\_contingency

chi2, p, dof, expected = chi2\_contingency(contingency\_table)

**Question:** Define the null hypothesis in a t-test.  
**Answer:** The null hypothesis assumes there is no significant difference between the sample mean and the population mean.

**Question:** Write Python code to calculate the correlation matrix of a DataFrame.  
**Answer:**python  
Copy code  
correlation\_matrix = df.corr()

**Question:** Write Python code to group a DataFrame by a column 'Category' and calculate the mean of 'Sales'.  
**Answer:**python  
Copy code  
grouped\_sales = df.groupby('Category')['Sales'].mean()

**Question:** Which method in Pandas is used to concatenate two DataFrames?  
**Answer:** pd.concat()

**Which of the following is NOT a feature of Principal Component Analysis (PCA)?**

* A) Reduces data dimensionality while preserving most of the variance.
* **B) Makes new dimensions correlated.**
* C) Constructs new dimensions using eigenvectors of the covariance matrix.
* D) Captures variance with the first few principal components.

**When performing a one-sample t-test in Python, which library function is used?**

* A) scipy.stats.ttest\_ind
* **B) scipy.stats.ttest\_1samp**
* C) statsmodels.stats.ttest\_1samp
* D) numpy.ttest

**In linear regression, the error measure Mean Squared Error (MSE) is calculated as:**

* **A) The mean of squared differences between predicted and actual values.**
* B) The square root of the average error.
* C) The mean of the absolute differences between predicted and actual values.
* D) None of the above.

**Cosine similarity between two vectors measures:**

* A) Their magnitude.
* **B) The angle between them.**
* C) Their orthogonality.
* D) The sum of their components.

**In Singular Value Decomposition (SVD), the matrix Σ\SigmaΣ contains:**

* A) Left singular vectors.
* B) Right singular vectors.
* C) Eigenvalues.
* **D) Singular values.**

**In a two-sample t-test, the null hypothesis typically states:**

* A) The two samples come from populations with different means.
* **B) The two samples come from populations with the same mean.**
* C) The two samples have equal medians.
* D) The samples are dependent.

**In hypothesis testing, the \_\_\_\_\_\_\_\_\_\_ is the statement being tested, often representing no effect or no difference.**

* **null hypothesis**

**The formula for calculating cosine similarity is:**Cosine Similarity(U,V)=U⋅V∣∣U∣∣∣∣V∣∣\text{Cosine Similarity}(U, V) = \frac{U \cdot V}{||U|| ||V||}Cosine Similarity(U,V)=∣∣U∣∣∣∣V∣∣U⋅V​, where U⋅VU \cdot VU⋅V is the \_\_\_\_\_\_\_\_\_\_ of the vectors UUU and VVV.

* **dot product**

**In PCA, the total variance is preserved because the \_\_\_\_\_\_\_\_\_\_ transformation does not alter the data's total variance.**

* **orthogonal**

**The \_\_\_\_\_\_\_\_\_\_ test is a non-parametric alternative to the two-sample t-test used for comparing medians of two independent samples.**

* **Wilcoxon Rank-Sum**

**In machine learning, a \_\_\_\_\_\_\_\_\_\_ problem involves predicting a continuous-valued output variable, such as temperature or housing price.**

* **regression**

**In vector algebra, two vectors are considered \_\_\_\_\_\_\_\_\_\_ if their dot product is zero.**

* **orthogonal**

**What is the primary advantage of using PCA for dimensionality reduction?**

* A) It minimizes the total variance of the data.
* **B) It captures the most variance with the first few principal components.**
* C) It ensures new dimensions are correlated.
* D) It removes all noise from the data.  
  **Answer: B**

**Which Python library provides functions for calculating cosine similarity?**

* A) pandas
* **B) scikit-learn**
* C) numpy
* D) statsmodels  
  **Answer: B**

**In hypothesis testing, what does a p-value represent?**

* A) The probability that the alternative hypothesis is true.
* B) **The probability of observing the test statistic as extreme as the sample value, assuming the null hypothesis is true.**
* C) The probability of making a Type II error.
* D) The significance level of the test.  
  **Answer: B**

**When calculating a dot product between two vectors, what type of value is returned?**

* A) A vector
* B) A matrix
* **C) A scalar**
* D) A list of scalars  
  **Answer: C**

**In linear regression, the slope of the regression line represents:**

* A) The y-intercept of the line.
* **B) The rate of change of the dependent variable with respect to the independent variable.**
* C) The residual variance in the data.
* D) The prediction error.  
  **Answer: B**

**In Singular Value Decomposition (SVD), which of the following describes the columns of matrix VVV?**

* A) Left singular vectors.
* B) Eigenvalues.
* **C) Right singular vectors.**
* D) Singular values.  
  **Answer: C**

**The Wilcoxon Rank-Sum test is best used when:**

* A) Comparing means of two dependent samples.
* **B) Comparing medians of two independent samples.**
* C) Testing independence between two categorical variables.
* D) Performing a one-sample t-test.  
  **Answer: B**

**Which of the following is an example of a classification problem in machine learning?**

* A) Predicting house prices based on location and size.
* B) Estimating the probability of rain tomorrow.
* **C) Determining if an email is spam or not spam.**
* D) Predicting the temperature in a city.  
  **Answer: C**

**In cosine similarity, a value of 1 means that two vectors:**

* **A) Point in the exact same direction.**
* B) Are orthogonal to each other.
* C) Have no similarity.
* D) Point in opposite directions.  
  **Answer: A**

**In hypothesis testing, a two-tailed test is used when:**

* A) **You are testing for any difference (greater or less) between the null and alternative hypotheses.**
* B) You are only testing for values greater than the null hypothesis.
* C) You are only testing for values less than the null hypothesis.
* D) You are comparing medians instead of means.  
  **Answer: A**

**In PCA, the eigenvalues of the covariance matrix represent:**

* A) The orthogonality of the new dimensions.
* B) The total variance of the data.
* **C) The variance captured by each principal component.**
* D) The covariance between the original variables.  
  **Answer: C**

1. **Which of the following is a key assumption for performing a two-sample t-test?**
   * A) The two samples must be dependent.
   * **B) The two populations must be normally distributed.**
   * C) The population variances must be unknown.
   * D) The samples must be of the same size.  
     **Answer: B**
2. **In hypothesis testing, the risk of rejecting the null hypothesis when it is true is called:**
   * A) Type II error.
   * **B) Type I error.**
   * C) P-value.
   * D) Significance level.  
     **Answer: B**
3. **When using the dot product formula geometrically, the cosine of the angle between two vectors is given by:**
   * A) ∣U∣⋅∣V∣U⋅V\frac{|U| \cdot |V|}{U \cdot V}U⋅V∣U∣⋅∣V∣​
   * **B) U⋅V∣U∣⋅∣V∣\frac{U \cdot V}{|U| \cdot |V|}∣U∣⋅∣V∣U⋅V​**
   * C) U⋅V⋅∣U∣⋅∣V∣U \cdot V \cdot |U| \cdot |V|U⋅V⋅∣U∣⋅∣V∣
   * D) ∣U−V∣U+V\frac{|U - V|}{U + V}U+V∣U−V∣​  
     **Answer: B**
4. **Which of the following correctly describes a rotation matrix in linear algebra?**
   * A) It scales a vector while maintaining its direction.
   * **B) It rotates a vector while preserving its length.**
   * C) It projects a vector onto another vector.
   * D) It computes the dot product of two vectors.  
     **Answer: B**
5. **What is the primary purpose of using Singular Value Decomposition (SVD) in image recognition?**
   * A) To cluster similar images into groups.
   * **B) To reduce the dimensionality of the data while capturing significant patterns.**
   * C) To perform hypothesis testing on image data.
   * D) To increase the resolution of images.  
     **Answer: B**
6. **Which Python function is used to compute the cosine similarity between two vectors?**
   * A) numpy.linalg.norm
   * B) scipy.spatial.distance.euclidean
   * **C) sklearn.metrics.pairwise.cosine\_similarity**
   * D) statsmodels.similarity.cosine  
     **Answer: C**
7. **When performing PCA, the new dimensions (principal components) are:**
   * A) Dependent and correlated.
   * B) Orthogonal but not independent.
   * **C) Orthogonal and uncorrelated.**
   * D) Arbitrary linear combinations of the original variables.  
     **Answer: C**

### **Fill-in-the-Blank Questions**

1. **The \_\_\_\_\_\_\_\_\_\_ error is the risk of failing to reject the null hypothesis when it is false.**
   * **Type II** (nullAltHyp)
2. **In vector algebra, the \_\_\_\_\_\_\_\_\_\_ of two vectors measures their similarity based on orientation rather than magnitude.**
   * **dot product** (cosDistance)
3. **In PCA, the \_\_\_\_\_\_\_\_\_\_ of the covariance matrix represent the variance captured by each principal component.**
   * **eigenvalues** (LandsetPCA)
4. **In SVD, the matrix UUU contains the \_\_\_\_\_\_\_\_\_\_ singular vectors.**
   * **left** (svdMNIST (1))
5. **The \_\_\_\_\_\_\_\_\_\_ test is used to determine whether there is an association between two categorical variables.**
   * **chi-square** (threeStatiscalTests)
6. **In linear regression, the \_\_\_\_\_\_\_\_\_\_ minimizes the sum of squared differences between observed and predicted values.**
   * **least-squares line** (regressionPseudoExample)
7. **The \_\_\_\_\_\_\_\_\_\_ function in Python can calculate the dot product of two vectors.**
   * **numpy.dot** (dotProd2)
8. **In hypothesis testing, a test statistic is computed and compared to a critical value to decide whether to reject the \_\_\_\_\_\_\_\_\_\_ hypothesis.**
   * **null** (nullAltHyp)

**The \_\_\_\_\_\_\_\_\_\_ test is a non-parametric alternative to the paired t-test used to compare medians of paired samples.**

* **Wilcoxon signed-rank** (threeStatiscalTests)

**In PCA, the first principal component captures the \_\_\_\_\_\_\_\_\_\_ amount of variance compared to other components.**

* **maximum** (LandsetPCA)

**When computing the cosine similarity, vectors are compared based on their \_\_\_\_\_\_\_\_\_\_ rather than their magnitude.**

* **orientation** (cosDistance)

**In a regression equation, the \_\_\_\_\_\_\_\_\_\_ term represents the value of the dependent variable when all independent variables are zero.**

* **intercept** (regressionPseudoExample)

**The \_\_\_\_\_\_\_\_\_\_ value in hypothesis testing represents the threshold probability for rejecting the null hypothesis.**

* **significance (alpha)** (nullAltHyp)

**In SVD, the \_\_\_\_\_\_\_\_\_\_ matrix contains the singular values arranged in descending order.**

* **diagonal** (svdMNIST (1))

**The dot product of two orthogonal vectors is always \_\_\_\_\_\_\_\_\_\_.**

* **zero** (dotProd2)

**The \_\_\_\_\_\_\_\_\_\_ test is a hypothesis test used to determine if a sample mean differs from a known population mean.**

* **one-sample t-test** (oneSampleTtest)

**In PCA, the eigenvectors of the covariance matrix represent the \_\_\_\_\_\_\_\_\_\_ directions in the data.**

* **principal** (LandsetPCA)

**In hypothesis testing, a \_\_\_\_\_\_\_\_\_\_ test is used when you are interested in deviations in both directions from the null hypothesis.**

* **two-tailed** (nullAltHyp)

### **Multiple-Choice Questions**

1. **In PCA, which of the following transformations does NOT change the total variance of the data?**
   * A) Non-linear transformation
   * B) Scaling by a constant
   * **C) Orthogonal transformation**
   * D) Random permutation  
     **Answer: C**
2. **Which Python function can be used to perform a one-sample t-test?**
   * A) statsmodels.ttest\_ind
   * B) numpy.stats.ttest
   * **C) scipy.stats.ttest\_1samp**
   * D) pandas.ttest  
     **Answer: C**
3. **What does a cosine similarity of 0 indicate about two vectors?**
   * A) They are identical.
   * **B) They are orthogonal (perpendicular).**
   * C) They point in the same direction.
   * D) They point in opposite directions.  
     **Answer: B**
4. **Which test would you use to compare the medians of two paired samples?**
   * A) Two-sample t-test
   * **B) Wilcoxon signed-rank test**
   * C) Chi-square test
   * D) One-sample t-test  
     **Answer: B**
5. **In the equation for a regression line y=mx+by = mx + by=mx+b, what does bbb represent?**
   * A) Slope of the line
   * **B) Y-intercept**
   * C) The predicted value of yyy
   * D) The residual error  
     **Answer: B**
6. **What is the primary advantage of SVD in machine learning?**
   * A) Simplifies data visualization.
   * **B) Reduces dimensionality while preserving significant data patterns.**
   * C) Increases the number of features for better learning.
   * D) Removes noise completely from the data.  
     **Answer: B**
7. **In vector algebra, two vectors are linearly independent if:**
   * A) Their dot product is zero.
   * B) They have the same direction but different magnitudes.
   * **C) No scalar multiple of one vector equals the other.**
   * D) They are orthogonal.  
     **Answer: C**
8. **The equation A⋅B=∣A∣⋅∣B∣⋅cos⁡(θ)A \cdot B = |A| \cdot |B| \cdot \cos(\theta)A⋅B=∣A∣⋅∣B∣⋅cos(θ) represents:**
   * A) Euclidean distance
   * B) Vector projection
   * **C) Dot product**
   * D) Covariance  
     **Answer: C**
9. **Which test is appropriate for determining whether two categorical variables are associated?**
   * A) Paired t-test
   * B) Wilcoxon signed-rank test
   * **C) Chi-square test**
   * D) Two-sample t-test  
     **Answer: C**
10. **In PCA, the fraction of the total variance explained by a principal component is calculated by:**
    * A) Dividing its eigenvalue by the total number of components.
    * **B) Dividing its eigenvalue by the sum of all eigenvalues.**
    * C) Squaring the eigenvalue and dividing by the total variance.
    * D) Adding the eigenvalue to the total variance.  
      **Answer: B**

### **Multiple-Choice Questions**

1. **Which of the following is NOT a feature of Principal Component Analysis (PCA)?**
   * A) Reduces data dimensionality while preserving most of the variance.
   * B) Captures variance with the first few principal components.
   * C) Constructs new dimensions using eigenvectors of the covariance matrix.
   * D) **Alters the total variance of the data**.  
     **Answer: D**
2. **In PCA, the eigenvalues of the covariance matrix represent:**
   * A) The covariance between the original variables.
   * **B) The variance captured by each principal component.**
   * C) The total variance of the dataset.
   * D) The correlation between principal components.  
     **Answer: B**
3. **What does a cosine similarity of 1 indicate about two vectors?**
   * A) They are orthogonal.
   * B) They point in opposite directions.
   * **C) They are identical in direction.**
   * D) They have the same magnitude.  
     **Answer: C**
4. **Which statistical test is used to compare medians of two paired samples?**
   * A) Two-sample t-test
   * **B) Wilcoxon signed-rank test**
   * C) Chi-square test
   * D) One-sample t-test  
     **Answer: B**
5. **In hypothesis testing, the significance level (α\alphaα) represents:**
   * A) The probability of a Type II error.
   * **B) The threshold for rejecting the null hypothesis.**
   * C) The probability of making a correct decision.
   * D) The risk of failing to reject the null hypothesis.  
     **Answer: B**
6. **In Singular Value Decomposition (SVD), the matrix UUU contains:**
   * A) Eigenvalues of the covariance matrix.
   * B) Right singular vectors.
   * **C) Left singular vectors.**
   * D) Principal components.  
     **Answer: C**
7. **Which machine learning model would be most appropriate for predicting a house price based on its features?**
   * A) Classification
   * **B) Regression**
   * C) Clustering
   * D) Dimensionality reduction  
     **Answer: B**
8. **In regression, a high R-squared value indicates:**
   * A) Poor fit of the model.
   * **B) The model explains a large proportion of the variance.**
   * C) Multicollinearity between predictors.
   * D) Overfitting of the data.  
     **Answer: B**
9. **What does it mean when two vectors are linearly independent?**
   * A) Their dot product is zero.
   * B) They point in the same direction.
   * **C) No scalar multiple of one vector can produce the other.**
   * D) Their cosine similarity is 1.  
     **Answer: C**
10. **The projection of one vector onto another represents:**
    * A) The magnitude of the first vector.
    * B) The direction of the second vector.
    * **C) The component of the first vector in the direction of the second.**
    * D) The angle between the two vectors.  
      **Answer: C**
11. **Which transformation in PCA ensures that new dimensions are uncorrelated?**
    * **A) Orthogonal transformation**
    * B) Non-linear transformation
    * C) Random permutation
    * D) Scaling by a constant  
      **Answer: A**
12. **Which test would be most appropriate for analyzing the relationship between two categorical variables?**
    * A) One-sample t-test
    * B) Wilcoxon rank-sum test
    * **C) Chi-square test**
    * D) Paired t-test  
      **Answer: C**
13. **In cosine similarity, vectors are compared based on:**
    * A) Their magnitudes.
    * **B) Their orientations.**
    * C) Their dot product.
    * D) Their projections.  
      **Answer: B**
14. **What does the slope (mmm) in the regression line y=mx+by = mx + by=mx+b represent?**
    * A) The y-intercept of the line.
    * B) The residual error in the model.
    * **C) The rate of change of the dependent variable with respect to the independent variable.**
    * D) The variance explained by the model.  
      **Answer: C**
15. **Which of the following best describes supervised learning?**
    * A) Clustering data points into groups without prior labels.
    * **B) Using labeled data to predict outcomes.**
    * C) Reducing the dimensionality of a dataset.
    * D) Transforming data into principal components.  
      **Answer: B**

### **Fill-in-the-Blank Questions**

**In PCA, the \_\_\_\_\_\_\_\_\_\_ of the covariance matrix represent the variance captured by the principal components.  
Answer: eigenvalues**

**The \_\_\_\_\_\_\_\_\_\_ test is a non-parametric alternative to the paired t-test used for comparing medians of paired samples.  
Answer: Wilcoxon signed-rank**

**A \_\_\_\_\_\_\_\_\_\_ is a hypothesis test used to determine if a sample mean differs from a known population mean.  
Answer: one-sample t-test**

**In Singular Value Decomposition (SVD), the matrix Σ\SigmaΣ contains the \_\_\_\_\_\_\_\_\_\_ values of the matrix.  
Answer: singular**

**A \_\_\_\_\_\_\_\_\_\_ model is used in machine learning to predict continuous values such as housing prices or temperatures.  
Answer: regression**

**In vector algebra, the \_\_\_\_\_\_\_\_\_\_ of two vectors measures their similarity based on orientation rather than magnitude.  
Answer: dot product**

**The \_\_\_\_\_\_\_\_\_\_ of a regression line represents the value of the dependent variable when all independent variables are zero.  
Answer: intercept**

**In hypothesis testing, the \_\_\_\_\_\_\_\_\_\_ is the threshold probability for rejecting the null hypothesis.  
Answer: significance level (alpha)**

**Two vectors are considered \_\_\_\_\_\_\_\_\_\_ if their dot product is zero.  
Answer: orthogonal**

**The \_\_\_\_\_\_\_\_\_\_ similarity is a metric commonly used in text analysis to compare the orientation of word frequency vectors.  
Answer: cosine**

### **Fill-in-the-Blank Questions**

1. The \_\_\_\_\_\_\_\_\_\_ of a covariance matrix represent the directions (principal components) in which the data varies the most.
   * **Answer:** Eigenvectors

Write Python code to calculate the dot product of two vectors uuu and vvv using Scikit-Learn.  
python  
Copy code  
from sklearn.metrics.pairwise import cosine\_similarity

dot\_product = cosine\_similarity([u], [v])[0][0] \* (np.linalg.norm(u) \* np.linalg.norm(v))

1. In hypothesis testing, a \_\_\_\_\_\_\_\_\_\_ test evaluates differences in both directions, above and below the mean.
   * **Answer:** Two-tailed

Write Python code to calculate the projection of vector uuu onto vector vvv:  
python  
Copy code  
projection = (np.dot(u, v) / np.dot(v, v)) \* v

1. The \_\_\_\_\_\_\_\_\_\_ in PCA is the ratio of the variance explained by each principal component to the total variance.
   * **Answer:** Explained variance ratio

Write Python code to perform a chi-square goodness-of-fit test:  
python  
Copy code  
from scipy.stats import chisquare

chisq\_stat, p\_value = chisquare(f\_obs, f\_exp)

### **Multiple Choice Questions**

1. Which of the following is a key assumption of the Wilcoxon signed-rank test? a) Samples are independent.  
   b) Data must be normally distributed.  
   c) Data are paired or dependent.  
   d) The sample size must be the same in both groups.
   * **Answer:** c) Data are paired or dependent.
2. In Singular Value Decomposition (SVD), the diagonal entries of the Σ\SigmaΣ matrix represent:  
   a) Variance of the data.  
   b) Eigenvalues of the covariance matrix.  
   c) Singular values of the original matrix.  
   d) Projections of data points onto principal components.
   * **Answer:** c) Singular values of the original matrix.
3. Which statistical test compares the means of two independent samples when the variances are assumed to be unequal?  
   a) Paired t-test  
   b) Independent two-sample t-test  
   c) Welch's t-test  
   d) Chi-square test
   * **Answer:** c) Welch's t-test
4. What does a p-value of 0.02 indicate in a hypothesis test with α=0.05\alpha = 0.05α=0.05?  
   a) Fail to reject the null hypothesis.  
   b) Reject the null hypothesis.  
   c) The null hypothesis is true.  
   d) The alternative hypothesis is false.
   * **Answer:** b) Reject the null hypothesis.
5. What does the scree plot in PCA visually represent?  
   a) The cumulative variance explained by each principal component.  
   b) The eigenvalues corresponding to each principal component.  
   c) The projection of data onto the principal components.  
   d) The orthogonality of principal components.
   * **Answer:** b) The eigenvalues corresponding to each principal component.

### **Short Answer Questions**

1. **Question:** Define the null hypothesis in a chi-square test of independence.
   * **Answer:** The null hypothesis states that there is no association between the two categorical variables being tested.
2. **Question:** What is the geometric interpretation of the dot product of two vectors?
   * **Answer:** The dot product measures the projection of one vector onto another, scaled by the cosine of the angle between them.
3. **Question:** Explain why PCA components are orthogonal.
   * **Answer:** PCA components are derived from the eigenvectors of the covariance matrix, which are orthogonal by definition, ensuring no correlation between components.
4. **Question:** What does a Type I error mean in hypothesis testing?
   * **Answer:** A Type I error occurs when the null hypothesis is rejected despite being true.

### **Python Coding Questions**

1. **Question:** Write Python code to calculate the residuals in a linear regression model.

python

Copy code

residuals = y\_true - y\_pred

1. **Question:** Write Python code to calculate the cosine similarity between all columns of a DataFrame.

python

Copy code

cosine\_sim = cosine\_similarity(df.T)

1. **Question:** Write Python code to calculate the length (magnitude) of a vector vvv:

python

Copy code

magnitude = np.sqrt(np.dot(v, v))

### **Conceptual Questions**

1. **Question:** What is the purpose of standardizing data before applying PCA?
   * **Answer:** Standardization ensures that all features have the same scale, preventing features with larger variances from dominating the principal components.
2. **Question:** Why is SVD particularly useful in image compression?
   * **Answer:** SVD identifies the most significant patterns in image data by retaining only the largest singular values, reducing storage requirements while preserving essential features.

### **Fill-in-the-Blank Questions**

1. In PCA, the \_\_\_\_\_\_\_\_\_\_ matrix is used to calculate eigenvalues and eigenvectors.
   * **Answer:** Covariance

Write Python code to calculate the total variance explained by the first two principal components in PCA.  
python  
Copy code  
total\_variance\_explained = sum(pca.explained\_variance\_ratio\_[:2])

1. The \_\_\_\_\_\_\_\_\_\_ statistic in a chi-square test measures the discrepancy between observed and expected frequencies.
   * **Answer:** Chi-square

Write Python code to calculate the correlation coefficient between two columns in a Pandas DataFrame.  
python  
Copy code  
correlation = df['col1'].corr(df['col2'])

1. In hypothesis testing, a \_\_\_\_\_\_\_\_\_\_ is the range of values where the null hypothesis is not rejected.
   * **Answer:** Confidence interval

Write Python code to generate random samples from a standard normal distribution using NumPy.  
python  
Copy code  
samples = np.random.normal(loc=0, scale=1, size=100)

1. The \_\_\_\_\_\_\_\_\_\_ of a vector represents its magnitude or length.
   * **Answer:** Norm

Write Python code to calculate the eigenvalues of a covariance matrix using NumPy.  
python  
Copy code  
eigenvalues, \_ = np.linalg.eig(cov\_matrix)

### **Multiple Choice Questions**

1. Which of the following describes the principal components in PCA?  
   a) Dependent on the original features.  
   b) Mutually orthogonal and uncorrelated.  
   c) Correlated with the original features.  
   d) Mutually dependent.
   * **Answer:** b) Mutually orthogonal and uncorrelated.
2. In a one-sample t-test, the null hypothesis assumes:  
   a) The sample mean equals the population mean.  
   b) The sample mean is greater than the population mean.  
   c) The sample mean is less than the population mean.  
   d) The sample mean has no variance.
   * **Answer:** a) The sample mean equals the population mean.
3. Which of the following is NOT an assumption of the chi-square test?  
   a) Data must be categorical.  
   b) Expected frequencies in each category must be at least 5.  
   c) Observations must be dependent.  
   d) The sample must be randomly selected.
   * **Answer:** c) Observations must be dependent.
4. Which property of vectors does the dot product measure?  
   a) Orthogonality  
   b) Magnitude  
   c) Directional alignment  
   d) Covariance
   * **Answer:** c) Directional alignment
5. In linear regression, the residuals should:  
   a) Have a mean of zero.  
   b) Be positively correlated with the independent variable.  
   c) Be negatively correlated with the independent variable.  
   d) Always have the same magnitude.
   * **Answer:** a) Have a mean of zero.
6. What is the purpose of the Σ\SigmaΣ matrix in Singular Value Decomposition (SVD)?  
   a) To store eigenvectors of the covariance matrix.  
   b) To represent singular values that capture data variance.  
   c) To project data onto the principal components.  
   d) To normalize the original data.
   * **Answer:** b) To represent singular values that capture data variance.
7. Which test is appropriate for comparing observed frequencies with expected frequencies in categorical data?  
   a) One-sample t-test  
   b) Two-sample t-test  
   c) Chi-square goodness-of-fit test  
   d) Wilcoxon signed-rank test
   * **Answer:** c) Chi-square goodness-of-fit test
8. Which Python library provides functions for performing statistical tests like t-tests and chi-square tests?  
   a) NumPy  
   b) SciPy  
   c) Pandas  
   d) Scikit-Learn
   * **Answer:** b) SciPy

### **Short Answer Questions**

1. **Question:** Why are principal components in PCA orthogonal?
   * **Answer:** They are derived from the eigenvectors of the covariance matrix, which are orthogonal by definition.
2. **Question:** Define the residual sum of squares (RSS) in linear regression.
   * **Answer:** RSS is the sum of the squared differences between the observed and predicted values of the dependent variable.
3. **Question:** What does a singular value in SVD represent?
   * **Answer:** It represents the magnitude of variance captured by the corresponding singular vector.
4. **Question:** What is the main assumption of the Wilcoxon rank-sum test?
   * **Answer:** The data in the two samples are independent and can be ranked.
5. **Question:** What does the cosine similarity measure between two vectors?
   * **Answer:** The cosine similarity measures the orientation of two vectors, indicating how aligned they are.
6. **Question:** What is the primary purpose of a confidence interval in hypothesis testing?
   * **Answer:** To provide a range of values within which the population parameter is likely to fall with a specified level of confidence.

### **Python Coding Questions**

1. **Question:** Write Python code to calculate the mean squared error (MSE) manually.

python

Copy code

mse = np.mean((y\_true - y\_pred)\*\*2)

1. **Question:** Write Python code to calculate the first two principal components of a dataset after PCA.

python

Copy code

principal\_components = pca.components\_[:2]

1. **Question:** Write Python code to calculate the p-value for a chi-square test in Python using SciPy.

python

Copy code

from scipy.stats import chi2

p\_value = 1 - chi2.cdf(chi\_stat, df)

1. **Question:** Write Python code to standardize a NumPy array.

python

Copy code

standardized\_array = (array - np.mean(array)) / np.std(array)

1. **Question:** Write Python code to create a covariance matrix from a dataset.

python

Copy code

covariance\_matrix = np.cov(dataset.T)

### **Fill-in-the-Blank Questions**

1. The \_\_\_\_\_\_\_\_\_\_ test is used to determine whether two categorical variables are independent.
   * **Answer:** Chi-square

Write Python code to calculate the eigenvectors of a covariance matrix using NumPy.  
python  
Copy code  
\_, eigenvectors = np.linalg.eig(cov\_matrix)

1. In hypothesis testing, the \_\_\_\_\_\_\_\_\_\_ error occurs when the null hypothesis is incorrectly rejected.
   * **Answer:** Type I

Write Python code to calculate the variance of a column named "Scores" in a Pandas DataFrame.  
python  
Copy code  
variance = df['Scores'].var()

1. The \_\_\_\_\_\_\_\_\_\_ statistic in a t-test measures the difference between the sample mean and the population mean relative to the sample's variability.
   * **Answer:** t-statistic

Write Python code to calculate the Frobenius norm of a matrix using NumPy.  
python  
Copy code  
frobenius\_norm = np.linalg.norm(matrix, 'fro')

1. In PCA, the eigenvectors of the covariance matrix form the \_\_\_\_\_\_\_\_\_\_ onto which the data is projected.
   * **Answer:** Principal components

### **Multiple Choice Questions**

1. Which of the following is an assumption of PCA?  
   a) Features must be standardized.  
   b) Data must be categorical.  
   c) Data must be paired.  
   d) Data must be non-linear.
   * **Answer:** a) Features must be standardized.
2. In hypothesis testing, a Type II error refers to:  
   a) Rejecting the null hypothesis when it is true.  
   b) Accepting the null hypothesis when it is false.  
   c) Failing to reject the null hypothesis when it is false.  
   d) Failing to reject the null hypothesis when it is true.
   * **Answer:** c) Failing to reject the null hypothesis when it is false.
3. Which of the following is NOT a property of singular vectors in SVD?  
   a) They are orthogonal.  
   b) They correspond to the directions of maximum variance.  
   c) They depend on the eigenvalues of the covariance matrix.  
   d) They are linearly dependent.
   * **Answer:** d) They are linearly dependent.
4. What does the determinant of the covariance matrix represent in PCA?  
   a) The total variance in the dataset.  
   b) The cumulative variance explained by the principal components.  
   c) The volume of the ellipsoid described by the data distribution.  
   d) The correlation between variables.
   * **Answer:** c) The volume of the ellipsoid described by the data distribution.
5. Which of the following metrics is best suited for comparing the similarity of high-dimensional vectors?  
   a) Euclidean distance  
   b) Manhattan distance  
   c) Cosine similarity  
   d) Correlation coefficient
   * **Answer:** c) Cosine similarity
6. Which of the following Python libraries provides functions for hypothesis testing?  
   a) pandas  
   b) numpy  
   c) scipy  
   d) sklearn
   * **Answer:** c) scipy
7. In SVD, the left singular vectors (UUU) represent:  
   a) Patterns in the columns of the original matrix.  
   b) Patterns in the rows of the original matrix.  
   c) The variance of the original matrix.  
   d) The eigenvalues of the covariance matrix.
   * **Answer:** b) Patterns in the rows of the original matrix.
8. Which of the following is an example of a two-tailed hypothesis test?  
   a) Testing if the mean is greater than a specified value.  
   b) Testing if the mean is less than a specified value.  
   c) Testing if the mean is different from a specified value.  
   d) Testing if the mean is equal to a specified value.
   * **Answer:** c) Testing if the mean is different from a specified value.

### **Short Answer Questions**

1. **Question:** What is the main advantage of the Wilcoxon rank-sum test over the two-sample t-test?
   * **Answer:** The Wilcoxon rank-sum test is non-parametric and does not assume normality of the data.
2. **Question:** How is the explained variance in PCA calculated?
   * **Answer:** It is the ratio of each eigenvalue to the sum of all eigenvalues of the covariance matrix.
3. **Question:** What does the trace of a covariance matrix represent in PCA?
   * **Answer:** The trace of the covariance matrix represents the total variance in the dataset.
4. **Question:** What is the difference between cosine similarity and Euclidean distance?
   * **Answer:** Cosine similarity measures directional alignment, while Euclidean distance measures absolute distance between points.
5. **Question:** Why is centering the data important before applying PCA?
   * **Answer:** Centering ensures that the data's mean is zero, which is required for the covariance matrix to correctly capture the variance relationships.

### **Python Coding Questions**

1. **Question:** Write Python code to calculate the cumulative explained variance in PCA.

python

Copy code

cumulative\_variance = np.cumsum(pca.explained\_variance\_ratio\_)

1. **Question:** Write Python code to calculate the determinant of a covariance matrix.

python

Copy code

determinant = np.linalg.det(cov\_matrix)

1. **Question:** Write Python code to calculate the L2 norm of a vector using NumPy.

python

Copy code

l2\_norm = np.linalg.norm(vector, ord=2)

1. **Question:** Write Python code to create a heatmap of the correlation matrix of a DataFrame.

python

Copy code

import seaborn as sns

import matplotlib.pyplot as plt

sns.heatmap(df.corr(), annot=True, cmap='coolwarm')

plt.show()

1. **Question:** Write Python code to perform a one-tailed t-test in SciPy.

python

Copy code

from scipy.stats import ttest\_1samp

t\_stat, p\_value = ttest\_1samp(sample, popmean=50, alternative='greater')

### **Conceptual Questions**

1. **Question:** Why does PCA work best with features that are on similar scales?
   * **Answer:** Features on different scales can disproportionately influence the principal components, leading to biased results.
2. **Question:** What does the rank of a matrix indicate in SVD?
   * **Answer:** The rank indicates the number of non-zero singular values and corresponds to the dimensionality of the data's subspace.
3. **Question:** How does SVD differ from PCA in terms of computation?
   * **Answer:** PCA uses the covariance matrix for eigen decomposition, while SVD operates directly on the data matrix, making it suitable for high-dimensional datasets.
4. **Question:** Why is cosine similarity preferred over Euclidean distance in high-dimensional spaces?
   * **Answer:** Cosine similarity focuses on the angle between vectors and is less sensitive to vector magnitude, which can be misleading in high dimensions.
5. **Question:** What is the main advantage of using PCA for image compression?
   * **Answer:** PCA reduces the dimensionality while retaining the most significant patterns, enabling efficient storage and reconstruction of images.

### **1. Conceptual Questions**

#### **1.1 Reading CSV Files**

**Question:** What is the correct way to read a CSV file in Python into a DataFrame, and check for missing values?

* **Correct Answer:** **(A)** df = pd.read\_csv('file.csv') and df.isnull().sum()

#### **1.2 Cosine Similarity**

**Question:** What does a cosine similarity of -1 between two vectors represent?

* **Correct Answer:** **(C)** They point in exactly opposite directions.

#### **1.3 PCA and SVD**

**Question:** In PCA, the first principal component:

* **Correct Answer:** **(B)** Is always orthogonal to the second principal component.

#### **1.4 Linear Regression**

**Question:** Which of the following measures the error of a regression model?

* **Correct Answer:** **(A)** Mean Squared Error (MSE)

#### **1.5 Statistical Testing**

**Question:** Which hypothesis test is appropriate to check whether the mean of a sample differs significantly from a known population mean?

* **Correct Answer:** **(B)** One-sample t-test

### **2. Python Coding Questions**

#### **2.1 Reading CSV Files**

**Question:** Fill in the blank to complete the code snippet to read a CSV file and print data types for each column:

python

Copy code

import pandas as pd

df = pd.read\_csv('data.csv')

print(df.\_\_\_\_())

* **Correct Answer:** **(A)** dtypes

#### **2.2 Finding Cosine Similarity**

**Question:** Write Python code to compute cosine similarity between two vectors u=[1,2,3]\mathbf{u} = [1, 2, 3]u=[1,2,3] and v=[4,5,6]\mathbf{v} = [4, 5, 6]v=[4,5,6].

**Answer:**

python

Copy code

from numpy import dot

from numpy.linalg import norm

u = [1, 2, 3]

v = [4, 5, 6]

cosine\_similarity = dot(u, v) / (norm(u) \* norm(v))

print(cosine\_similarity)

#### **2.3 Linear Regression**

**Question:** Fill in the blank to compute MSE for a linear regression model:

python

Copy code

from sklearn.metrics import mean\_squared\_error

y\_true = [3, -0.5, 2, 7]

y\_pred = [2.5, 0.0, 2, 8]

mse = mean\_squared\_error(y\_true, y\_pred)

print(\_\_\_\_)

* **Correct Answer:** **(A)** mse

#### **2.4 Applying PCA**

**Question:** Write Python code to apply PCA to reduce a dataset with 5 features to 2 principal components. Assume the dataset is stored in a DataFrame df.

**Answer:**

python

Copy code

from sklearn.decomposition import PCA

import pandas as pd

# Example dataset

df = pd.DataFrame({

'Feature1': [1, 2, 3],

'Feature2': [4, 5, 6],

'Feature3': [7, 8, 9],

'Feature4': [10, 11, 12],

'Feature5': [13, 14, 15]

})

# Applying PCA

pca = PCA(n\_components=2)

reduced\_data = pca.fit\_transform(df)

print(reduced\_data)

#### **2.5 Performing Statistical Tests**

**Question:** Fill in the blank to perform a one-sample t-test:

python

Copy code

from scipy.stats import ttest\_1samp

data = [2.3, 3.1, 2.8, 3.5, 3.0]

pop\_mean = 3.0

t\_stat, p\_value = ttest\_1samp(data, \_\_\_\_)

print(t\_stat, p\_value)

* **Correct Answer:** **(B)** 3

### **3. Advanced Coding Application**

#### **3.1 PCA and SVD Application**

**Question:** Given a dataset of grayscale images (16x16), use Python to:

1. Apply SVD to one image, display the top 3 singular vectors, and reconstruct the image with reduced dimensions.
2. Write code to perform PCA on the entire dataset and plot the explained variance ratio.

**Answer:**

python

Copy code

import numpy as np

import matplotlib.pyplot as plt

from sklearn.decomposition import PCA

# Example image (16x16 grayscale)

image = np.random.random((16, 16))

# SVD

U, S, Vt = np.linalg.svd(image, full\_matrices=False)

print("Top 3 singular vectors:", U[:, :3])

# Reconstruct the image using reduced dimensions

rank = 3

reconstructed = np.dot(U[:, :rank], np.dot(np.diag(S[:rank]), Vt[:rank, :]))

plt.imshow(reconstructed, cmap='gray')

plt.title("Reconstructed Image")

plt.show()

# Example dataset for PCA

dataset = np.random.random((100, 256)) # 100 images, flattened to 256 pixels each

pca = PCA()

pca.fit(dataset)

# Plot explained variance ratio

plt.plot(np.cumsum(pca.explained\_variance\_ratio\_))

plt.xlabel('Number of Components')

plt.ylabel('Explained Variance')

plt.title('PCA Explained Variance')

plt.show()

### **4. Conceptual Questions**

#### **4.1 Data Analysis**

**Question:** Which function is used in pandas to find the number of missing values in a DataFrame?

* (A) df.missing\_values()
* (B) df.isnull().sum()
* (C) df.countna()
* (D) df.missing()

#### **4.2 Cosine Similarity**

**Question:** If the cosine similarity between two vectors is 0.85, what does this indicate?

* (A) The vectors are almost perpendicular.
* (B) The vectors are highly similar in direction.
* (C) The vectors are pointing in opposite directions.
* (D) The vectors are not similar at all.

#### **4.3 PCA**

**Question:** What is the relationship between PCA and SVD?

* (A) PCA is a special case of SVD applied to the covariance matrix.
* (B) SVD is a specific application of PCA for image data.
* (C) PCA and SVD are unrelated mathematical techniques.
* (D) PCA requires SVD to work only in low dimensions.

#### **4.4 Regression Models**

**Question:** In linear regression, which of the following would NOT improve model accuracy?

* (A) Adding more data points.
* (B) Including irrelevant features.
* (C) Normalizing the data.
* (D) Removing outliers.

#### **4.5 Statistical Tests**

**Question:** A researcher wants to test if two groups have different medians. Which test should they use?

* (A) Paired t-test
* (B) Wilcoxon Rank-Sum test
* (C) One-sample t-test
* (D) Chi-Square test

### **5. Python Coding Questions**

#### **5.1 Handling Missing Data**

**Question:** Fill in the blanks to drop rows in a DataFrame with any missing values:

python

Copy code

import pandas as pd

df = pd.read\_csv('data.csv')

df\_cleaned = df.\_\_\_\_(how='any')

Answer options:

* (A) drop\_rows
* (B) dropna
* (C) fillna
* (D) remove\_na

#### **5.2 Cosine Similarity**

**Question:** Write Python code to calculate the cosine similarity between two vectors using scipy.

**Answer:**

python

Copy code

from scipy.spatial.distance import cosine

u = [1, 2, 3]

v = [4, 5, 6]

cosine\_similarity = 1 - cosine(u, v)

print(cosine\_similarity)

#### **5.3 Linear Regression**

**Question:** Complete the code to fit a simple linear regression model using sklearn and calculate the R-squared value:

python

Copy code

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import r2\_score

X = [[1], [2], [3], [4], [5]]

y = [2.2, 2.8, 3.6, 4.5, 5.1]

model = LinearRegression()

model.fit(X, y)

y\_pred = model.predict(X)

r2 = r2\_score(y, y\_pred)

print(\_\_\_\_)

Answer options:

* (A) model
* (B) r2\_score
* (C) r2
* (D) y\_pred

#### **5.4 PCA**

**Question:** Write Python code to determine the percentage of variance explained by the first two principal components of a dataset.

**Answer:**

python

Copy code

from sklearn.decomposition import PCA

import pandas as pd

# Example dataset

df = pd.DataFrame({

'Feature1': [1, 2, 3],

'Feature2': [4, 5, 6],

'Feature3': [7, 8, 9]

})

pca = PCA(n\_components=2)

pca.fit(df)

explained\_variance = sum(pca.explained\_variance\_ratio\_) \* 100

print(f"Percentage of variance explained by first two components: {explained\_variance}%")

#### **5.5 Statistical Tests**

**Question:** Write Python code to perform a paired t-test to compare two related samples.

python

Copy code

from scipy.stats import ttest\_rel

sample1 = [20, 21, 22, 23, 24]

sample2 = [21, 22, 23, 24, 25]

t\_stat, p\_value = ttest\_rel(\_\_\_\_, \_\_\_\_)

print(t\_stat, p\_value)

Answer options:

* (A) sample1, sample1
* (B) sample1, sample2
* (C) sample2, sample1
* (D) sample2, sample2

### **6. Advanced Application**

#### **6.1 PCA on Image Data**

**Question:** You are given a 3-channel image dataset. Write Python code to:

1. Flatten the images.
2. Perform PCA to reduce them to a single principal component.
3. Reconstruct the data using only the first component.

**Answer:**

python

Copy code

import numpy as np

from sklearn.decomposition import PCA

# Example dataset: 3-channel image flattened to 2D

images = np.random.random((10, 3 \* 256)) # 10 images, each with 3 channels (256 pixels each)

# Apply PCA

pca = PCA(n\_components=1)

reduced\_data = pca.fit\_transform(images)

reconstructed = pca.inverse\_transform(reduced\_data)

print("Original shape:", images.shape)

print("Reconstructed shape:", reconstructed.shape)

### **7. Coding Questions with Answers**

#### **7.1 Reading and Analyzing DataFrames**

**Question:** Write Python code to count the number of missing values in each column of a DataFrame.

python

Copy code

import pandas as pd

df = pd.read\_csv('data.csv')

missing\_values = df.isnull().sum()

print(missing\_values)

* **Correct Answer:** **(B)** isnull().sum()

#### **7.2 Linear Regression**

**Question:** Write Python code to predict new values using a trained linear regression model.

python

Copy code

from sklearn.linear\_model import LinearRegression

X\_train = [[1], [2], [3]]

y\_train = [2, 4, 6]

model = LinearRegression()

model.fit(X\_train, y\_train)

# Predict value for X\_test

X\_test = [[4]]

y\_pred = model.predict(X\_test)

print(y\_pred)

* **Correct Answer:** **(B)** predict

#### **7.3 Cosine Similarity**

**Question:** Write Python code to calculate the cosine distance between two vectors using scipy.

python

Copy code

from scipy.spatial.distance import cosine

u = [1, 0, -1]

v = [-1, 0, 1]

cosine\_distance = cosine(u, v)

print(cosine\_distance)

* **Correct Answer:** **(A)** cosine

#### **7.4 PCA Explained Variance**

**Question:** Write Python code to display the explained variance ratio of each principal component in a dataset.

python

Copy code

from sklearn.decomposition import PCA

import pandas as pd

df = pd.DataFrame({

'Feature1': [1, 2, 3],

'Feature2': [4, 5, 6],

'Feature3': [7, 8, 9]

})

pca = PCA(n\_components=3)

pca.fit(df)

print("Explained variance ratio:", pca.explained\_variance\_ratio\_)

* **Correct Answer:** **(A)** explained\_variance\_ratio\_

#### **7.5 One-Sample t-Test**

**Question:** Write Python code to perform a one-sample t-test to check if the sample mean differs from the population mean.

python

Copy code

from scipy.stats import ttest\_1samp

sample = [10, 12, 14, 16, 18]

pop\_mean = 15

t\_stat, p\_value = ttest\_1samp(sample, pop\_mean)

print(t\_stat, p\_value)

* **Correct Answer:** **(A)** sample, pop\_mean

#### **7.6 SVD**

**Question:** Write Python code to compute the SVD of a matrix and print the singular values.

python

Copy code

import numpy as np

A = np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9]])

U, S, Vt = np.linalg.svd(A)

print("Singular values:", S)

* **Correct Answer:** **(B)** A

#### **7.7 Statistical Testing**

**Question:** Write Python code to perform a Chi-Square test of independence.

python

Copy code

from scipy.stats import chi2\_contingency

# Contingency table

data = [[10, 20, 30], [6, 9, 17]]

chi2, p, dof, expected = chi2\_contingency(data)

print(f"Chi2: {chi2}, p-value: {p}, DOF: {dof}")

* **Correct Answer:** **(A)** data

#### **7.8 PCA Data Reconstruction**

**Question:** Write Python code to reconstruct data from PCA using the first two principal components.

python

Copy code

from sklearn.decomposition import PCA

import numpy as np

# Example dataset

data = np.random.random((10, 5)) # 10 samples, 5 features

pca = PCA(n\_components=2)

reduced\_data = pca.fit\_transform(data)

# Reconstruct the original dataset

reconstructed\_data = pca.inverse\_transform(reduced\_data)

print(reconstructed\_data)

* **Correct Answer:** **(B)** inverse\_transform

#### **7.9 Filling Missing Values**

**Question:** Write Python code to fill missing values in a DataFrame with the mean of each column.

python

Copy code

import pandas as pd

df = pd.DataFrame({

'A': [1, 2, None],

'B': [None, 3, 4]

})

df\_filled = df.fillna(df.mean())

print(df\_filled)

* **Correct Answer:** **(B)** fillna

#### **7.10 Statistical Tests**

**Question:** Write Python code to perform a two-sample t-test for two independent samples.

python

Copy code

from scipy.stats import ttest\_ind

sample1 = [1.1, 2.2, 3.3, 4.4]

sample2 = [5.5, 6.6, 7.7, 8.8]

t\_stat, p\_value = ttest\_ind(sample1, sample2)

print(t\_stat, p\_value)

* **Correct Answer:** **(B)** sample1, sample2

### **1. Reading and Analyzing DataFrames**

#### **1.1 Count Missing Values**

python

Copy code

import pandas as pd

# Example dataset with missing values

data = {'A': [1, 2, None], 'B': [4, None, 6], 'C': [7, 8, 9]}

df = pd.DataFrame(data)

# Count missing values in each column

missing\_values = df.isnull().sum()

print("Missing values per column:")

print(missing\_values)

#### **1.2 Drop Rows with Missing Values**

python

Copy code

df\_cleaned = df.dropna()

print("DataFrame after dropping rows with missing values:")

print(df\_cleaned)

#### **1.3 Fill Missing Values with Column Mean**

python

Copy code

df\_filled = df.fillna(df.mean())

print("DataFrame after filling missing values with column mean:")

print(df\_filled)

### **2. Cosine Similarity**

#### **2.1 Compute Cosine Similarity Using NumPy**

python

Copy code

import numpy as np

u = np.array([1, 2, 3])

v = np.array([4, 5, 6])

cosine\_similarity = np.dot(u, v) / (np.linalg.norm(u) \* np.linalg.norm(v))

print(f"Cosine similarity: {cosine\_similarity}")

#### **2.2 Compute Cosine Distance Using SciPy**

python

Copy code

from scipy.spatial.distance import cosine

u = [1, 2, 3]

v = [4, 5, 6]

cosine\_distance = cosine(u, v)

print(f"Cosine distance: {cosine\_distance}")

### **3. Linear Regression**

#### **3.1 Fit a Linear Regression Model**

python

Copy code

from sklearn.linear\_model import LinearRegression

# Training data

X\_train = [[1], [2], [3], [4]]

y\_train = [2.5, 3.6, 4.8, 6.1]

# Create and fit the model

model = LinearRegression()

model.fit(X\_train, y\_train)

# Print model coefficients

print("Intercept:", model.intercept\_)

print("Coefficient:", model.coef\_)

#### **3.2 Predict New Values**

python

Copy code

X\_test = [[5], [6]]

y\_pred = model.predict(X\_test)

print("Predicted values:", y\_pred)

#### **3.3 Calculate Mean Squared Error (MSE)**

python

Copy code

from sklearn.metrics import mean\_squared\_error

y\_true = [2.5, 3.6, 4.8, 6.1]

y\_pred = [2.4, 3.7, 4.9, 6.0]

mse = mean\_squared\_error(y\_true, y\_pred)

print(f"Mean Squared Error: {mse}")

### **4. PCA**

#### **4.1 Apply PCA to Reduce Dimensions**

python

Copy code

from sklearn.decomposition import PCA

import pandas as pd

# Example dataset

df = pd.DataFrame({

'Feature1': [1, 2, 3],

'Feature2': [4, 5, 6],

'Feature3': [7, 8, 9]

})

pca = PCA(n\_components=2)

reduced\_data = pca.fit\_transform(df)

print("Reduced data:")

print(reduced\_data)

#### **4.2 Plot Explained Variance**

python

Copy code

import matplotlib.pyplot as plt

explained\_variance\_ratio = pca.explained\_variance\_ratio\_

plt.bar(range(len(explained\_variance\_ratio)), explained\_variance\_ratio)

plt.xlabel('Principal Component')

plt.ylabel('Variance Explained')

plt.title('PCA Explained Variance')

plt.show()

#### **4.3 Reconstruct Data from PCA**

python

Copy code

reconstructed\_data = pca.inverse\_transform(reduced\_data)

print("Reconstructed data:")

print(reconstructed\_data)

### **5. SVD**

#### **5.1 Compute SVD of a Matrix**

python

Copy code

import numpy as np

A = np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9]])

U, S, Vt = np.linalg.svd(A)

print("U matrix:")

print(U)

print("Singular values:")

print(S)

print("Vt matrix:")

print(Vt)

#### **5.2 Approximate a Matrix with Reduced Rank**

python

Copy code

rank = 2 # Choose rank

A\_approx = U[:, :rank] @ np.diag(S[:rank]) @ Vt[:rank, :]

print("Approximated matrix:")

print(A\_approx)

### **6. Statistical Tests**

#### **6.1 One-Sample t-Test**

python

Copy code

from scipy.stats import ttest\_1samp

data = [2.3, 2.5, 2.7, 3.0, 3.5]

pop\_mean = 3.0

t\_stat, p\_value = ttest\_1samp(data, pop\_mean)

print(f"T-statistic: {t\_stat}, P-value: {p\_value}")

#### **6.2 Paired t-Test**

python

Copy code

from scipy.stats import ttest\_rel

before = [1.5, 2.3, 2.9, 3.4]

after = [1.7, 2.5, 3.0, 3.6]

t\_stat, p\_value = ttest\_rel(before, after)

print(f"T-statistic: {t\_stat}, P-value: {p\_value}")

#### **6.3 Two-Sample t-Test**

python

Copy code

from scipy.stats import ttest\_ind

group1 = [1.1, 2.2, 3.3, 4.4]

group2 = [2.1, 3.2, 4.3, 5.4]

t\_stat, p\_value = ttest\_ind(group1, group2)

print(f"T-statistic: {t\_stat}, P-value: {p\_value}")

#### **6.4 Chi-Square Test**

python

Copy code

from scipy.stats import chi2\_contingency

# Contingency table

data = [[10, 20, 30], [15, 25, 35]]

chi2, p, dof, expected = chi2\_contingency(data)

print(f"Chi2: {chi2}, P-value: {p}, Degrees of Freedom: {dof}")

print("Expected frequencies:")

print(expected)

### **7.1 DataFrame Operations**

#### **7.1.1 Filter Rows Based on a Condition**

python

Copy code

import pandas as pd

# Example dataset

df = pd.DataFrame({

'Name': ['Alice', 'Bob', 'Charlie', 'David'],

'Age': [25, 30, 35, 40],

'Score': [88, 92, 95, 70]

})

# Filter rows where Score > 90

filtered\_df = df[df['Score'] > 90]

print("Filtered DataFrame:")

print(filtered\_df)

#### **7.1.2 Group and Aggregate Data**

python

Copy code

# Group by 'Age' and calculate the mean Score

grouped = df.groupby('Age').mean()

print("Grouped DataFrame with Mean Score:")

print(grouped)

### **7.2 Cosine Similarity**

#### **7.2.1 Compute Cosine Similarity for Multiple Pairs**

python

Copy code

import numpy as np

from sklearn.metrics.pairwise import cosine\_similarity

# Example vectors

X = np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9]])

# Compute cosine similarity between all pairs

similarity\_matrix = cosine\_similarity(X)

print("Cosine similarity matrix:")

print(similarity\_matrix)

### **7.3 Linear Regression**

#### **7.3.1 Create a Synthetic Dataset for Regression**

python

Copy code

from sklearn.datasets import make\_regression

import matplotlib.pyplot as plt

# Generate synthetic data

X, y = make\_regression(n\_samples=100, n\_features=1, noise=10)

# Visualize the data

plt.scatter(X, y)

plt.title("Synthetic Regression Data")

plt.xlabel("Feature")

plt.ylabel("Target")

plt.show()

### **7.4 PCA**

#### **7.4.1 PCA on Image Data**

python

Copy code

from sklearn.decomposition import PCA

import numpy as np

# Example: 16x16 grayscale image

image = np.random.random((16, 16))

# Flatten image and apply PCA

image\_flattened = image.flatten().reshape(1, -1)

pca = PCA(n\_components=5)

reduced\_image = pca.fit\_transform(image\_flattened)

print("Reduced image representation:", reduced\_image)

### **7.5 SVD**

#### **7.5.1 Reconstruct an Image Using SVD**

python

Copy code

import numpy as np

import matplotlib.pyplot as plt

# Create a grayscale image (16x16 matrix)

image = np.random.random((16, 16))

# Compute SVD

U, S, Vt = np.linalg.svd(image)

# Reconstruct the image using the top 5 singular values

rank = 5

reconstructed\_image = U[:, :rank] @ np.diag(S[:rank]) @ Vt[:rank, :]

# Plot the original and reconstructed images

plt.subplot(1, 2, 1)

plt.title("Original Image")

plt.imshow(image, cmap='gray')

plt.subplot(1, 2, 2)

plt.title("Reconstructed Image (Rank 5)")

plt.imshow(reconstructed\_image, cmap='gray')

plt.show()

### **7.6 Statistical Tests**

#### **7.6.1 Wilcoxon Rank-Sum Test (Mann-Whitney U Test)**

python

Copy code

from scipy.stats import mannwhitneyu

group1 = [1.1, 2.2, 3.3, 4.4]

group2 = [2.1, 3.2, 4.3, 5.4]

stat, p\_value = mannwhitneyu(group1, group2)

print(f"Statistic: {stat}, P-value: {p\_value}")

#### **7.6.2 Confidence Intervals for t-Test**

python

Copy code

from scipy.stats import ttest\_1samp

data = [2.1, 2.5, 2.8, 3.0, 3.5]

pop\_mean = 3.0

t\_stat, p\_value = ttest\_1samp(data, pop\_mean)

# Calculate 95% confidence interval

from numpy import mean, std, sqrt

n = len(data)

conf\_interval = (mean(data) - 1.96 \* (std(data) / sqrt(n)),

mean(data) + 1.96 \* (std(data) / sqrt(n)))

print(f"Confidence Interval: {conf\_interval}")

### **7.7 Additional DataFrame Tasks**

#### **7.7.1 Pivot Tables**

python

Copy code

# Create a pivot table to summarize the data

pivot\_table = df.pivot\_table(values='Score', index='Age', aggfunc='mean')

print("Pivot Table:")

print(pivot\_table)

#### **7.7.2 Detect and Drop Duplicates**

python

Copy code

# Add duplicate rows

df = df.append({'Name': 'Alice', 'Age': 25, 'Score': 88}, ignore\_index=True)

# Detect duplicates

duplicates = df.duplicated()

print("Duplicated rows:")

print(df[duplicates])

# Drop duplicates

df\_no\_duplicates = df.drop\_duplicates()

print("DataFrame after removing duplicates:")

print(df\_no\_duplicates)

### **7.8 Advanced PCA and SVD**

#### **7.8.1 Visualize Principal Components**

python

Copy code

from sklearn.decomposition import PCA

import matplotlib.pyplot as plt

# Example dataset

X = np.random.random((50, 3))

pca = PCA(n\_components=2)

X\_pca = pca.fit\_transform(X)

# Scatter plot of the first two principal components

plt.scatter(X\_pca[:, 0], X\_pca[:, 1])

plt.title("First Two Principal Components")

plt.xlabel("PC1")

plt.ylabel("PC2")

plt.show()

#### **7.8.2 Low-Rank Approximation Using SVD**

python

Copy code

# Approximate a matrix with top-k singular values

k = 2

A\_low\_rank = U[:, :k] @ np.diag(S[:k]) @ Vt[:k, :]

print("Low-rank approximation:")

print(A\_low\_rank)

### **7.9 Machine Learning Preprocessing**

#### **7.9.1 Normalize Features**

python

Copy code

from sklearn.preprocessing import StandardScaler

# Example dataset

X = [[1, 2, 3], [4, 5, 6], [7, 8, 9]]

scaler = StandardScaler()

X\_scaled = scaler.fit\_transform(X)

print("Normalized features:")

print(X\_scaled)

#### **7.9.2 Train-Test Split**

python

Copy code

from sklearn.model\_selection import train\_test\_split

# Example dataset

X = [[1], [2], [3], [4], [5]]

y = [2, 4, 6, 8, 10]

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

print("Training data:", X\_train, y\_train)

print("Test data:", X\_test, y\_test)

Your dataframe has two columns, one with "Age" and one with "Income".

Write 1-3 lines of python code to extract (filter) your dataframe to only those rows where "Age" is above 40 AND "Income" is above $100,000.

import pandas as pd

df = pd.DataFrame({'Age': [30, 45, 50], 'Income': [80000, 120000, 150000]})

filtered\_df = df[(df['Age'] > 40) & (df['Income'] > 100000)]