

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

# Load dataset
df = pd.read_csv("/content/archive (4).zip")

# Descriptive statistics
df.describe()
```

	math score	reading score	writing score
count	1000.00000	1000.00000	1000.00000
mean	66.08900	69.169000	68.054000
std	15.16308	14.600192	15.195657
min	0.00000	17.000000	10.000000
25%	57.00000	59.000000	57.750000
50%	66.00000	70.000000	69.000000
75%	77.00000	79.000000	79.000000
max	100.00000	100.000000	100.000000



Import Libraries and Load Dataset

Interpretation:

This code imports the Python libraries required to test the data and carry out statistical tests. Pandas is a tool that is used to manipulate data, NumPy is utilized to perform arithmetic operations, Matplotlib to visualize, and inferential tests, which are done in SciPy. The data is loaded into a DataFrame which is denoted asdf. Description The describe () function provides distribution statistics such as a mean, standard deviation, minimum, maximum, and quartiles. This gives a summary of the performance of students in academic settings as well as assists to know the general scores distribution.

```
# Correlation Test (Inferential Analysis)
corr, p_value = stats.pearsonr(df['math score'], df['reading score'])
corr, p_value

(np.float64(0.8175796636720541), np.float64(1.7877531099061433e-241))
```

Correlation Test (Inferential Test)

Interpretation:

This code gives a Pearson correlation test between mathematics and reading scores. o The strength of the relationship and direction is measured by the coefficient of correlation (r). o The relationship is statistically significant based on the value of p. When the p-value is lower than the 0.05, the relationship is significant. A very positive correlation will show that the students who excel in mathematics would also excel in reading. This implies that there is a relationship between academic skills in different subjects.

```
# Check Missing Values
df.isnull().sum()

# Check Duplicate Records
df.duplicated().sum()

np.int64(0)
```

Diagnostic Measures (Missing Values and Duplicates)

Interpretation:

This code verifies quality of data. o Isnull().sum() determine the presence of any missing value in the data. o duplicated().sum() checks for repeated records. The value that represents the zero missing values and the zero duplicates prove that the data is clean. This makes sure that statistical findings will not be biased or distorted as a result of missing or multiple data.

```
# Outlier Detection using IQR (Math Score)

Q1 = df['math score'].quantile(0.25)
Q3 = df['math score'].quantile(0.75)
IQR = Q3 - Q1

df_filtered = df[
    (df['math score'] >= Q1 - 1.5 * IQR) &
    (df['math score'] <= Q3 + 1.5 * IQR)
]
```

Detection of Outliers using IQR method.

Interpretation:

This code identifies and rejects possible outliers in mathematics scores that are dealt with with the Interquartile Range (IQR) method. o Q1 indicates the 25 th percentile. o Q3 is the 75 th percentile. o IQR is used to measure the dispersion of the central 50 percent of the data. Any score that is not within the range of $1.5 \times \text{IQR}$ can be counted as an outlier. The elimination of the extreme values will aid in making results more accurate statistically and make sure that the results will not be affected by the odd cases.

```
# Hypothesis Testing (Test Preparation Effect)

completed = df[df['test preparation course'] == 'completed']['math score']
none = df[df['test preparation course'] == 'none']['math score']

t_stat, p_value = stats.ttest_ind(completed, none)
t_stat, p_value

if p_value < 0.05:
    print("Reject H0: Test preparation significantly affects math performance.")
else:
    print("Fail to reject H0: No significant difference found.")

Reject H0: Test preparation significantly affects math performance.
```

Independent Samples T-test (Hypothesis Testing)

Interpretation:

To evaluate Mathematic score difference between two groups, this code shall conduct independent samples t-test: 1. Learners that attended the test preparation course. 2. Learners who failed to finish the course. The t-statistic is used to measure the difference between a group of means whereas the p-value is used to indicate statistical significance. 1. If the p-value is less than 0.05: 2. The null-hypothesis (H₀) is dismissed. This implies that preparation of tests will statistically affect math performance. 1. If the p-value is greater than 0.05: 2. The null hypothesis is not rejected. This implies that it is not notable between the groups. At this level, the $p < 0.05$ value will show that the students who had gone through the preparation course got the mathematics scores that were significantly higher.