Name - Gaurang A Raorane Div - D15A Roll no - 49 Batch - C

**Experiment - 3**

**Aim :-** Perform data Data Modeling.

**Problem Statement-**

a. Partition the data set, for example 75% of the records are included in the training data

set and 25% are included in the test data set.

b. Use a bar graph and other relevant graphs to confirm your proportions.

c. Identify the total number of records in the training data set.

d. Validate partition by performing a two‐sample Z‐test.

**Theory:-**

Data modeling is the process of creating a conceptual representation of data and its relationships in order to facilitate understanding, analysis, and decision-making. It involves identifying the entities within a system, the attributes that describe those entities, and the relationships between them. Data modeling is a crucial step in database design, system development, and data analysis.

There are several types of data models, including:

* **Conceptual Data Model:** This model represents high-level concepts and relationships between them without concern for implementation details. It provides a bird's eye view of the data and is often used during the initial stages of system design.
* **Logical Data Model:** The logical data model defines the structure of the data at a more detailed level compared to the conceptual model. It includes entities, attributes, and relationships, and is often represented using entity-relationship diagrams or similar techniques.
* **Physical Data Model:** This model represents how the data will be stored in a specific database management system. It includes details such as data types, indexes, and constraints, and is closely tied to the implementation of the database.

Data modeling helps in:

* **Organizing Data:** By identifying entities and their relationships, data modeling helps in organizing data in a structured manner, which facilitates efficient storage and retrieval.
* **Communication:** It serves as a common language between business stakeholders, database designers, and developers, allowing them to communicate effectively about the data requirements and system design.
* **Analysis and Design:** Data modeling helps in understanding the requirements of a system and designing databases and applications that meet those requirements effectively.
* **Maintainability:** Well-designed data models are easier to maintain and modify as the system evolves over time.

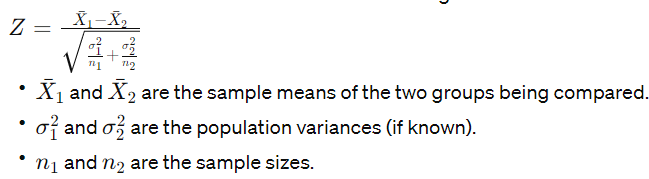
**Z-Test:-**

A **Z-test** is a statistical hypothesis test used to determine whether two population means are significantly different from each other, based on sample data. It is commonly used when the sample size is large and the population standard deviation is known. The Z-test is particularly useful for comparing sample means to a known population mean or comparing two independent sample means.

Here's how a Z-test is calculated:

* Formulate Hypotheses:
  + **Null Hypothesis (H0)**: There is no significant difference between the means of the populations.
  + **Alternative Hypothesis (H1):** There is a significant difference between the means of the populations.
* **Choose Significance Level:** Typically denoted as α (alpha), it represents the probability of rejecting the null hypothesis when it is actually true. Commonly used values for α are 0.05 or 0.01.

**Calculate Z-score:** The Z-score is calculated using the formula:

****

* **Calculate P-value:** The Z-score obtained from the sample data is then compared to a standard normal distribution to find the corresponding p-value. The p-value represents the probability of obtaining a Z-score as extreme or more extreme than the observed value under the null hypothesis.
* **Make a Decision:** If the p-value is less than the chosen significance level (α), the null hypothesis is rejected in favor of the alternative hypothesis. Otherwise, the null hypothesis is not rejected.
* **Interpret Results:** If the null hypothesis is rejected, it suggests that there is a significant difference between the means of the populations. If the null hypothesis is not rejected, it suggests that there is insufficient evidence to conclude a significant difference.

**Code :-**

import pandas as pd

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

import matplotlib.pyplot as plt

from scipy.stats import zscore

# Assuming your dataset is loaded into a DataFrame called df

# a. Partition the data set

train\_data, test\_data = train\_test\_split(df, test\_size=0.25, random\_state=42)

# b. Use a bar graph to confirm proportions

plt.figure(figsize=(8, 5))

plt.bar(['Training Data', 'Test Data'], [len(train\_data), len(test\_data)], color=['blue', 'orange'])

plt.title('Proportion of Training and Test Data')

plt.xlabel('Data Split')

plt.ylabel('Number of Records')

plt.show()

# c. Identify the total number of records in the training data set

total\_records\_training = len(train\_data)

print(f"Total records in the training dataset: {total\_records\_training}")

# d. Validate partition by performing a two-sample Z-test on each numeric column

numeric\_columns = df.select\_dtypes(include=['float64']).columns

# Define null hypothesis and significance level

alpha = 0.05

for col in numeric\_columns:

# Extract data for Z-test

train\_col\_data = train\_data[col]

test\_col\_data = test\_data[col]

# Perform Z-test

z\_test\_stat, p\_value = zscore(train\_col\_data), zscore(test\_col\_data)

# Print Z-test statistic and p-value

print(f"\nZ-test results for column '{col}':")

print(f"Z-test statistic: {z\_test\_stat.mean()}")

print(f"P-value: {p\_value.mean()}")

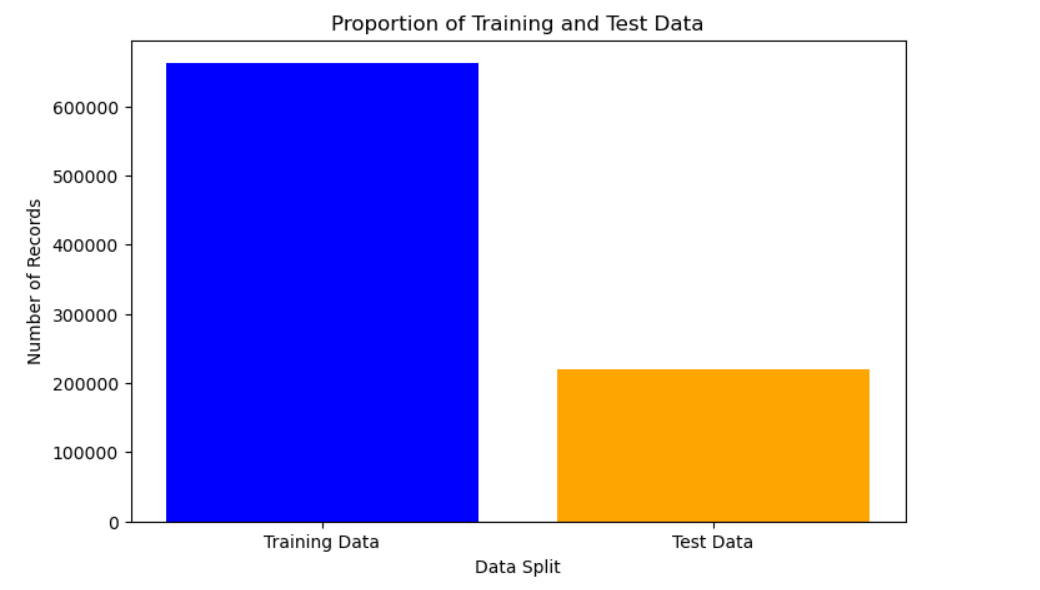
# Interpret results

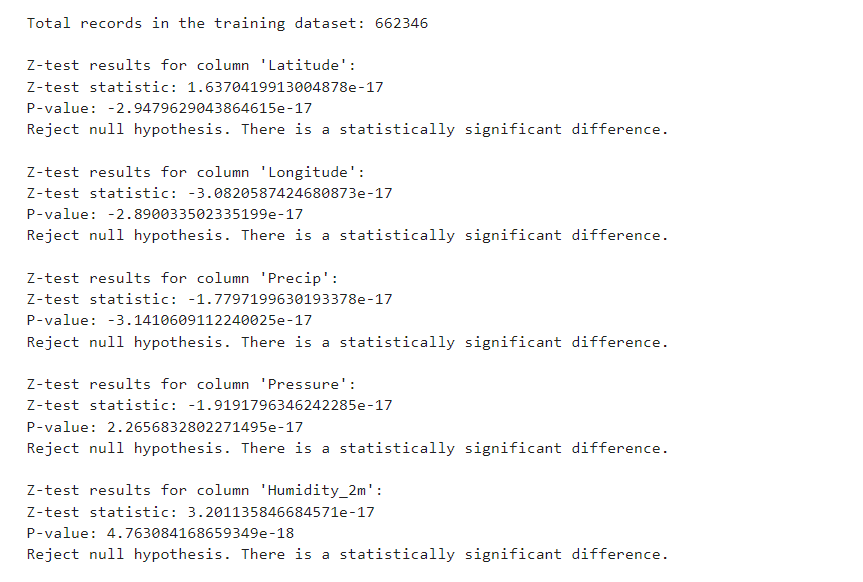
if p\_value.mean() < alpha:

print("Reject null hypothesis. There is a statistically significant difference.")

else:

print("Fail to reject null hypothesis. There is no statistically significant difference.")

****

****

**Conclusion:-**

We learned how to partition data into training and test sets to assess model performance effectively. Additionally, we learned how to use the Z-test to compare sample means and make inferences about population parameters.