**Experiment 3**

**Shashwat Tripathi**

**D15A Batch C  
RollNo: 64**

**AIM : Perform data Data Modeling.**

**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 graph 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 partitioning**

Data partitioning is a technique used in databases to divide large sets of data into smaller, more manageable pieces called partitions. This process improves performance, scalability, and reliability of data storage and retrieval. There are several types of data partitioning strategies:

List Partitioning: Data is divided into partitions based on a predefined list of values for a specific column in the table. Each partition contains rows that match a particular value in the list.

Hash Partitioning: Data is divided into partitions based on a hash function applied to a specific column in the table. The hash function generates a value that is used to assign each row to a specific partition. This type of partitioning is useful when there is no obvious range or list to partition on.

Range Partitioning: Data is divided into partitions based on ranges of values. For example, a table of dates could be partitioned so that each partition contains data for a specific month or year.

Round-Robin Partitioning: Data items are evenly cyclically distributed across partitions, ensuring basic load balancing and even distribution.

Composite Partitioning: Combines multiple partitioning methods or criteria. It involves combining partitioning keys or attributes to determine data placement across partitions.

Hybrid Partitioning: Combines horizontal and vertical partitioning. Tuples are assigned to different partitions using horizontal partitioning, and the attributes of the tuples are partitioned and assigned to different partitions using vertical partitioning.

The choice of partitioning strategy depends on the dataset’s properties, access patterns, and the needs of the particular application or system. For instance, geographically distributed databases may use geo-partitioning, dividing data based on geographical regions for efficient spatial queries and analysis.

**Hypothesis testing**

Hypothesis testing is a statistical method used to make inferences or draw conclusions about the population based on a sample of data. It is a way to test the validity of a claim or idea, often referred to as a hypothesis, about a population parameter.

Here are the key components of hypothesis testing:

Null Hypothesis (H0): This is the hypothesis that the researcher wants to disprove. It usually represents the status quo or no effect. For example, "Men are, on average, not taller than women".

Alternative Hypothesis (Ha or H1): This is the hypothesis that suggests a change or difference. It is the opposite of the null hypothesis. For example, "Men are, on average, taller than women".

Test Statistic: A value calculated from the sample data without any unknown parameters, often used to summarize the sample for comparison purposes.

P-Value: The probability of obtaining the observed data, or data more extreme, assuming the null hypothesis is true. A low p-value indicates strong evidence against the null hypothesis 34.

Significance Level (α): The threshold below which the null hypothesis is rejected. Common choices are 0.01, 0.05, and 0.1. If the p-value is less than the significance level, the null hypothesis is rejected.

One-Tailed and Two-Tailed Tests: One-tailed tests are used when the direction of the difference is known, such as when testing if a new drug is better than a placebo. Two-tailed tests are used when the direction of the difference is unknown, such as when testing if a new drug has a different effect than the current standard drug.

The hypothesis testing process generally follows these steps:

Formulate the null and alternative hypotheses.

Choose a significance level.

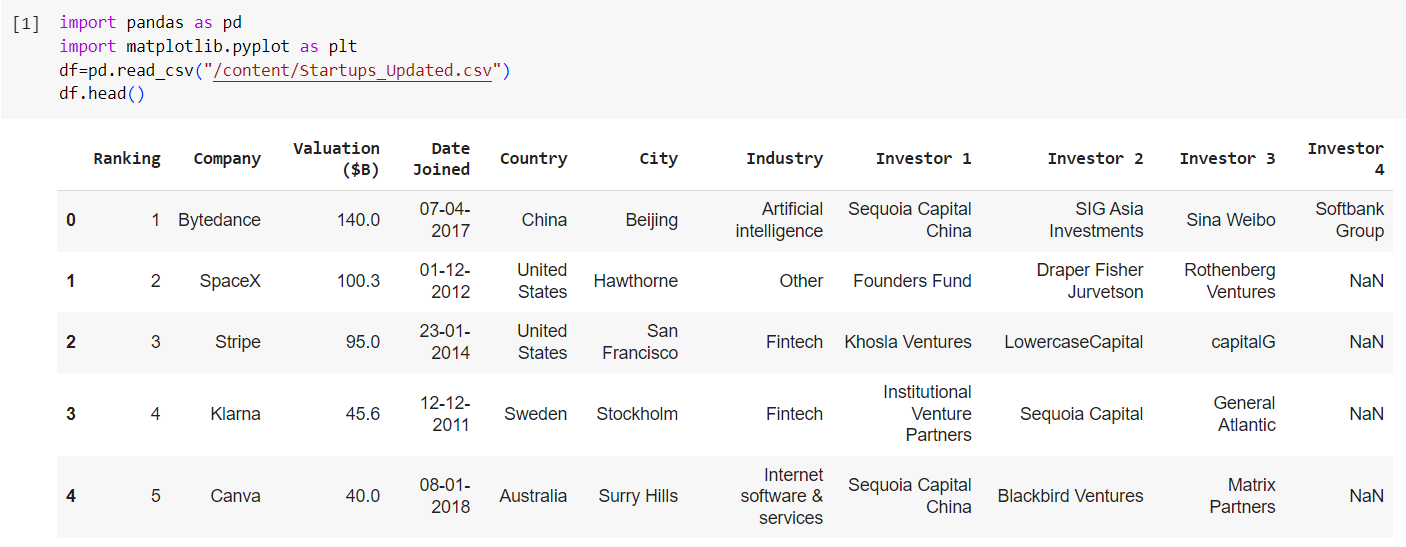
Calculate the test statistic and find the corresponding p-value.

Decide whether to reject or fail to reject the null hypothesis based on the p-value and the chosen significance level.

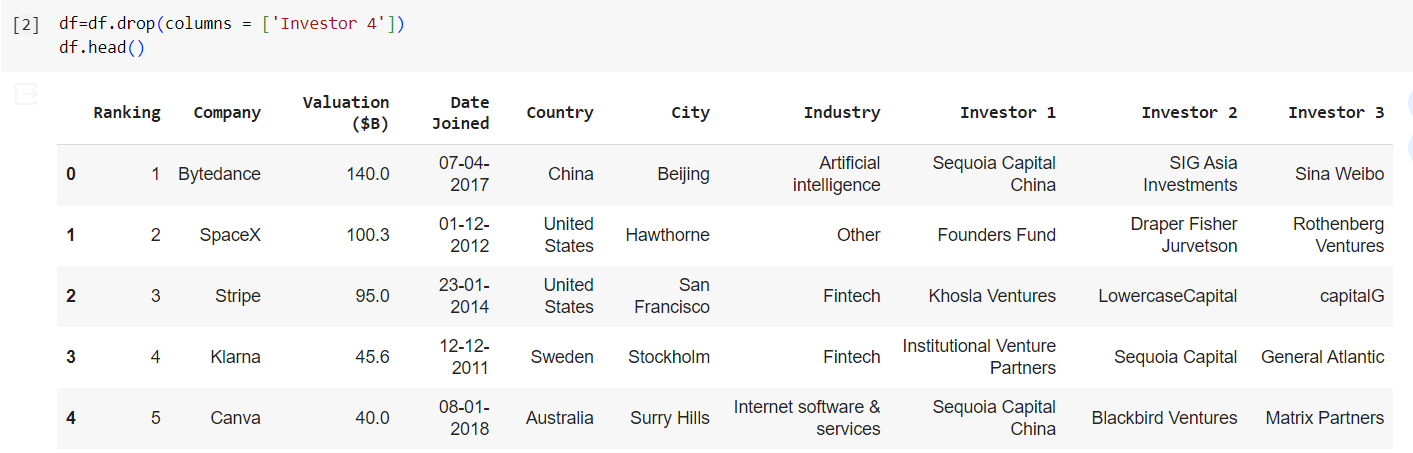
Interpret the results in the context of the research question.

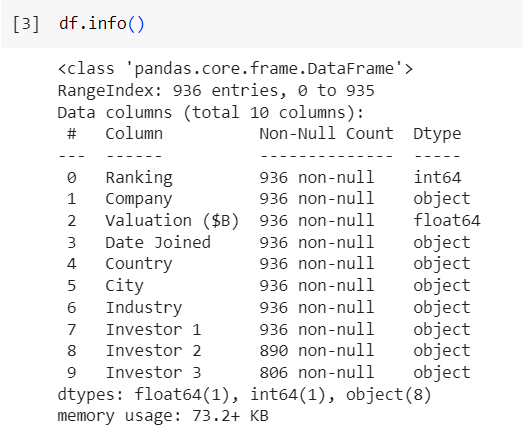
It's important to note that hypothesis testing involves a trade-off between Type I and Type II errors. A Type I error occurs when the null hypothesis is rejected when it is actually true, leading to a false positive. A Type II error occurs when the null hypothesis is not rejected when it is actually false, leading to a false negative.

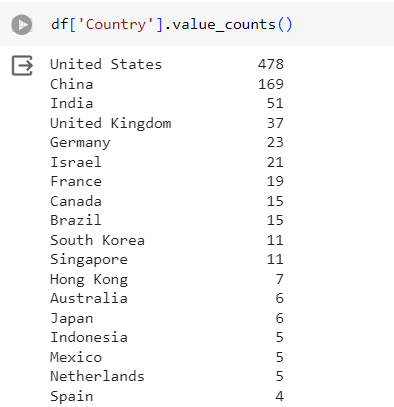
**OUTPUT :**

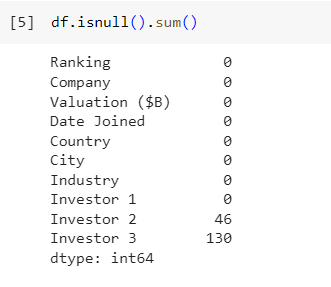
****

**Dropping investor 4.**

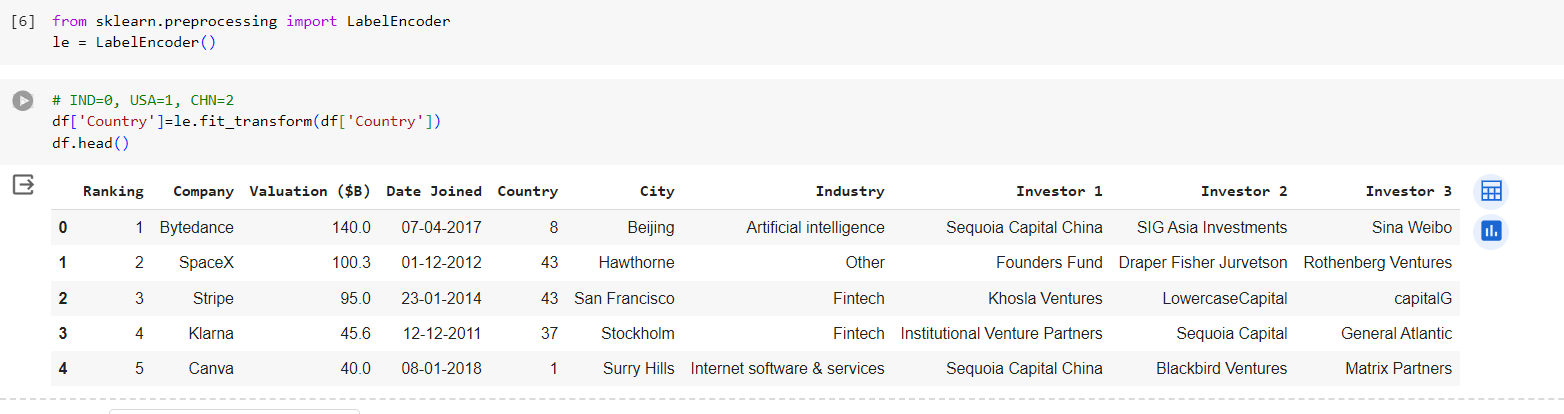
****

****

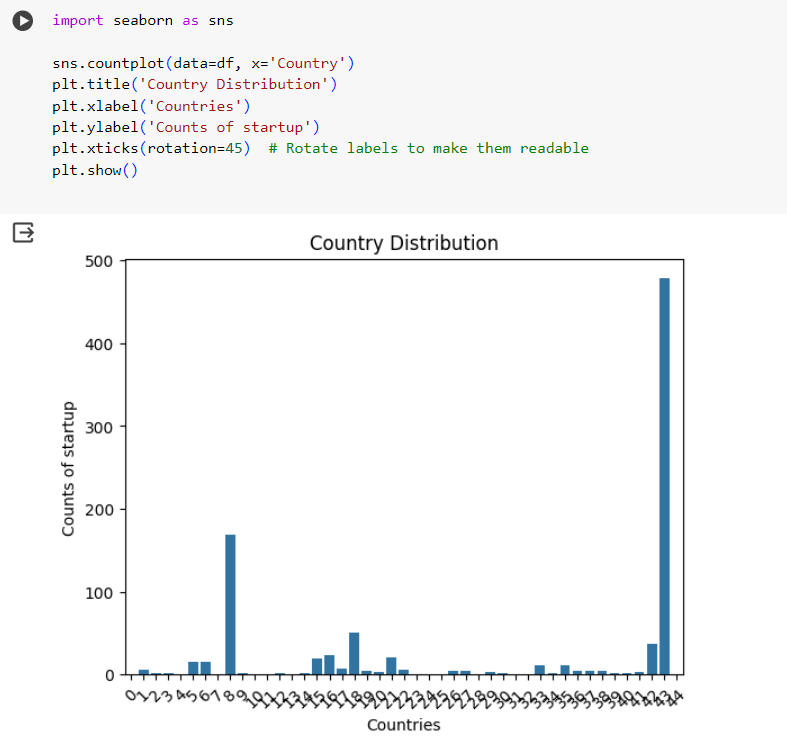
****

****

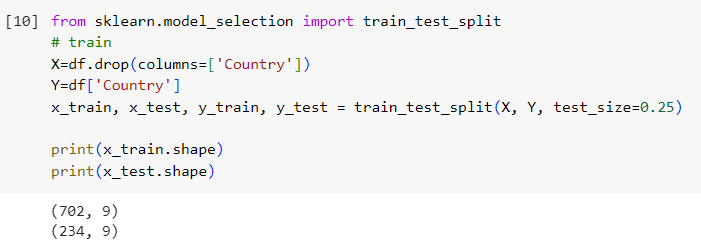
**Assigning country codes based on alphabetical order  
From list of 44 countries alphabetically USA is 2nd last hence 43 is the code.**

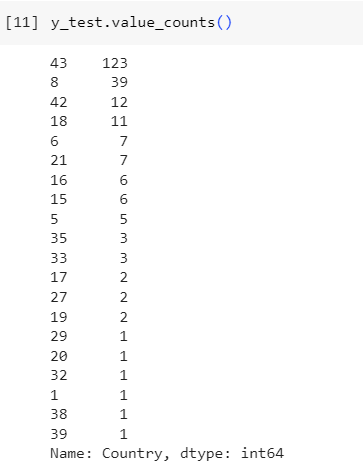
****

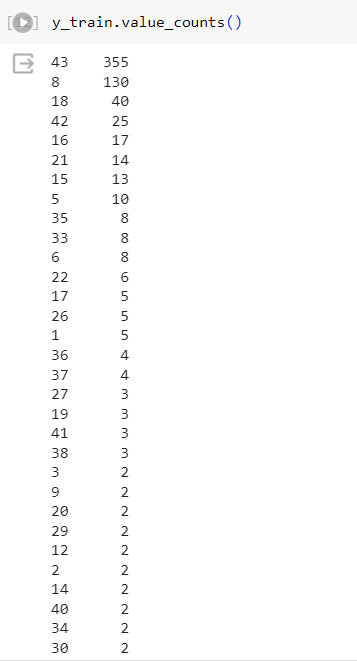
**Visualization of data**

****

**Dividing into training and testing data**

****

****

****

**The hypothesis part**

**If p\_value < 0.05 null hypothesis is rejected and alternative hypothesis is accepted**

**Else if p\_value >= 0.05 null hypothesis is accepted.**

****

**CONCLUSION : Thus, as my null hypothesis is accepted, it shows that there is similarity between my training and testing data.**