AISSMS
INSTITUTE OF INFORMATION TECHNOLOGY
ADDING VALUE TO ENGINEERING

Department of Artificial Intelligence & Data Science

Data Analytics with Python Laboratory

(ADVSE406)

Second Year Artificial Intelligence & Data Science (2023 Course)

Semester- IV

List of Experiments

Sr. No.	Title of Experiment	CO	PO	PSO
1.	Bridge the gap (Python programming using panadas data frame)	CO1	PO1	PSO1
2.	Implement Numpy, Pandas data analysis: Use a sample dataset, clean, process, and analyze the data, extracting meaningful insights using Pandas functionalities.	CO1	PO1	PSO2
3.	Develop a python program that analyzes a dataset containing information about a company's sales, expenses & profit. Implement data cleaning, descriptive statistics & data visualization.	CO1	PO1	PSO2
4.	Design a python program to explore the concepts of sampling & sampling distributions. Generate random samples from a given population analyze their characteristics & demonstrate how the sample mean evolves as more samples are taken. Investigate properties of sampling distributions (Mean & S.D).	CO1	PO2	PSO2
5.	Sampling Distribution: Write a program to generate multiple samples (at least 30) from the population & calculate the mean for each sample.	CO1	PO2	PSO2
6.	Random Sampling: Implement a random sampling process to extract a representative sample from the chosen dataset.	CO1	PO2	PSO2
7.	Develop a python program for hypothesis testing, focusing on comparing means of two samples. Implement a statistical test such as t-test or z-test, to analyze whether there is significant difference between the means.	CO1	PO3	PSO2
8.	Create a python program that performs a two sample hypothesis test & introduce the basics of ANOVA. Allow users to input data for multiple groups, choose a significance level & conduct a two sample t-test for comparing means or ANOVA for comparing means or ANOVA for comparing means across more than two groups.	CO1	PO4	PSO2
9.	Create a python program for multiple regression analysis .Allow users to input a datasets with multiple independent variables, specify a dependent variable & perform multiple regression to model the complex relationship.	CO2	PO2,4	PSO2
10.	Build a python program that allows users to input data for a binary outcome, perform logistic regression, evaluate the models performance using ROC analysis & visualize the ROC curve.	CO2	PO3	PSO2
11.	Design a python program for clustering analysis. Allow users to input a dataset & choose a clustering algorithm. Implement the selected algorithm, visualize the clusters & provide insights into grouping	CO2	PO3	PSO2

	patterns within the data.			
12.	Stock Price Forecasting: Develop a program to perform linear regression on historical stock prices use relevant features like previous day closing price, trading volume &market indices. Access the models ability to predict future stock prices.	CO2	PO4	PSO2
13.	Utilize housing price prediction dataset, split it into training & testing sets, train the model &evaluate its accuracy.	CO2	PO4	PSO2
14.	Student Performance Analysis: Implement linear regression to analyze the relationship between study hours & exam scores in a dataset of students. Visualize the regression line & access how well study hours predict academic performance. (https://www.kaggle.com/datasets/spscientist/students-performance-in-exams)	CO2	PO4	PSO2
15.	Implement Logistic regression to analyze rainfall . (https://www.kaggle.com/code/chandrimad31/rainfall-prediction-7-popular-models)	CO2	PO4	PSO2
16	Content Beyond Practical	CO1,2	PO1,2	PSO1, 2

Bridge the Gap

TITLE: Bridge the gap (Python Programming)

OBJECTIVES: 1. To get familiar with dataset.

2. To learn python library installation of TensorFlow, NumPy, SciPy, Pandas, Matplotlib, SciKit-Learn.

PROBLEM STATEMENT: Perform the following operations using Python on any open-source dataset.

1. Install all the required Python Libraries.
2. Import all the required Python Libraries.
3. Load the Dataset into pandas data frame and solve query based on dataset.

OUTCOME: Students will be able to,

1. understand various libraries in python.
2. use dataset and analyze it.

THEORY-CONCEPT:

Python is the most widely used programming language today. When it comes to solving data science tasks and challenges, Python never ceases to surprise its users. Most data scientists are already leveraging the power of Python programming every day. Python is an easy-to-learn, easy-to-debug, widely used, object-oriented, open-source, high-performance language, and there are many more benefits to Python programming. Python has been built with extraordinary Python libraries for data science that are used by programmers every day in solving problems.

1. TensorFlow

The first in the list of python libraries for data science is TensorFlow. TensorFlow is a library for high-performance numerical computations. It is used across various scientific fields. TensorFlow is basically a framework for defining and running computations that involve tensors, which are partially defined computational objects that eventually produce a value.

How to install tensorflow in jupyter notebook?

pip install tensorflow

2. NumPy

NumPy (Numerical Python) is the fundamental package for numerical computation in Python; it contains a powerful N-dimensional array object. It is a general-purpose array-processing package that provides high-performance multidimensional objects called arrays and tools for working with them. NumPy also addresses the slowness problem partly by providing these multidimensional arrays as well as providing functions and operators that operate efficiently on these arrays.

How to install numpy in jupyter notebook?

pip install numpy

3. SciPy

SciPy (Scientific Python) is another free and open-source Python library for data science that is extensively used for high-level computations. It's extensively used for scientific and technical computations, because it extends NumPy and provides many user-friendly and efficient routines for scientific calculations.

How to install numpy in jupyter notebook?

pip install scipy

4. Pandas

Pandas (Python data analysis) is a must in the data science life cycle. It is the most popular and widely used Python library for data science, along with NumPy in matplotlib. it is heavily used for data analysis and cleaning. Pandas provides fast, flexible data structures, such as data frame CDs, which are designed to work with structured data very easily and intuitively.

How to install panads in jupyter notebook?

pip install pandas

5. Matplotlib

Matplotlib has powerful yet beautiful visualizations. It is a plotting library for Python. Because of the graphs and plots that it produces, it is extensively used for data visualization. It also provides an object-oriented API, which can be used to embed those plots into applications.

How to install matplotlib in jupyter notebook?

pip install matplotlib

6. Scikit-learn

Next in the list of the top python libraries for data science comes Scikit-learn, a machine learning library that provides almost all the machine learning algorithms you might need. Scikit-learn is designed to be interpolated into NumPy and SciPy.

How to install scikit-learn in jupyter notebook?

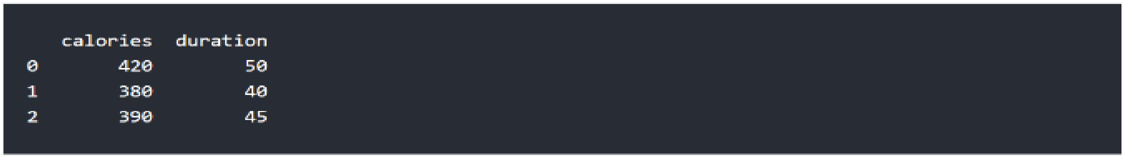
pip install -U scikit-learn

Example:

Create a simple Pandas DataFrame:

```
import pandas as pd
data = {
  "calories": [420, 380, 390],
  "duration": [50, 40, 45]
}
#load data into a DataFrame object:
df = pd.DataFrame(data)
print(df)
```

Output



```
   calories  duration
0       420        50
1       380        40
2       390        45
```

CONCLUSION:

Assignment No. 2

TITLE: Implement Numpy, Pandas data analysis

OBJECTIVES: 1. To get familiar with dataset.

2. To use dataset and analyze it using pandas functionalities.

PROBLEM STATEMENT: Implement Numpy, Pandas data analysis: Use a sample dataset, clean, process, and analyze the data, extracting meaningful insights using Pandas functionalities.

OUTCOME: Students will be able to,

1. understand various libraries in python.
2. use dataset and analyze it.

THEORY-CONCEPT:

Data Wrangling in Python Data Wrangling is the process of gathering, collecting, and transforming Raw data into another format for better understanding, decision-making, accessing, and analysis in less time. Data Wrangling is also known as Data Munging.

Importance of Data Wrangling

Data Wrangling is a very important step. The below example will explain its importance as:

Books selling Website want to show top-selling books of different domains, according to user preference. For example, a new user search for motivational books, then they want to show those motivational books which sell the most or having a high rating, etc. But on their website, there are plenty of raw data from different users. Here the concept of Data Munging or Data Wrangling is used. As we know Data is not Wrangled by System. This process is done by Data Scientists. So, the data Scientist will wrangle data in such a way that they will sort that motivational books that are sold more or have high ratings or user buy this book with these packages of Books, etc. Based on that, the new user will make choice.

Data Wrangling is a crucial topic for Data Science and Data Analysis. Pandas Framework of Python is used for Data Wrangling. Pandas is an open-source library specifically developed for Data Analysis and Data Science. The process like data sorting or filtration, Data grouping, etc. Data wrangling in python deals with the below functionalities:

1. **Data exploration:** In this process, the data is studied, analyzed, and understood by visualizing representations of data.

2. Dealing with missing values: Most of the datasets having a vast amount of data contain missing values of NaN, they are needed to be taken care of by replacing them with mean, mode, the most frequent value of the column or simply by dropping the row having a NaN value.

3. Reshaping data: In this process, data is manipulated according to the requirements, where new data can be added or pre-existing data can be modified.

4. Filtering data: Sometimes datasets are comprised of unwanted rows or columns which are required to be removed or filtered

5. Data exploration: We assign the data, and then we visualize the data in a tabular format.

6. Other: After dealing with the raw dataset with the above functionalities we get an efficient dataset as per our requirements and then it can be used for a required purpose like data analyzing, machine learning, data visualization, model training etc.

Pandas

Pandas (Python data analysis) is a must in the data science life cycle. It is the most popular and widely used Python library for data science, along with NumPy in matplotlib. it is heavily used for data analysis and cleaning. Pandas provides fast, flexible data structures, such as data frame CDs, which are designed to work with structured data very easily and intuitively.

How to install panads in jupyter notebook?

pip install pandas

CONCLUSION:

Assignment No. 3

TITLE: Implement data cleaning, descriptive statistics & data visualization

OBJECTIVES: 1. To learn data cleaning process on different dataset.

2. To Explore different data visualization technique.

3. To learn different descriptive statistics.

PROBLEM STATEMENT: Develop a python program that analyzes a dataset containing information about a company's sales, expenses & profit. Implement data cleaning, descriptive statistics & data visualization.

OUTCOME: Students will be able to,

1. apply different data cleaning process on dataset.
2. Implement different descriptive statistics on dataset.

THEORY-CONCEPT:

Descriptive statistics are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis , and skewness.

For example

The sum of the following data set is 20: (2, 3, 4, 5, 6). The mean is 4 (20/5).

The mode of a data set is the value appearing most often, and the median is the figure situated in the middle of the data set. It is the figure separating the higher figures from the lower figures within a data set. However, there are fewer common types of descriptive statistics that are still very important.

Types of Descriptive Statistics

1. **Central Tendency**

Measures of central tendency focus on the average or middle values of data sets, whereas measures of variability focus on the dispersion of data. These two measures use graphs, tables, and general discussions to help people understand the meaning of the analyzed data. Measures of central tendency describe the center position of a distribution for a data set. A person analyzes the frequency of each data point in the distribution and describes it using the mean, median, or mode, which measures the most common patterns of the analyzed data set

2. Measures of Variability

Measures of variability (or the measures of spread) aid in analyzing how dispersed the distribution is for a set of data. For example, while the measures of central tendency may give a person the average of a data set, it does not describe how the data is distributed within the set.

So, while the average of the data maybe 65 out of 100, there can still be data points at both 1 and 100. Measures of variability help communicate this by describing the shape and spread of the data set. Range, quartiles, absolute deviation, and variance are all examples of measures of variability.

Consider the following data set: 5, 19, 24, 62, 91, 100. The range of that data set is 95, which is calculated by subtracting the lowest number (5) in the data set from the highest (100).

3. Distribution

Distribution (or frequency distribution) refers to the quantity of times a data point occurs. Alternatively, it is the measurement of a data point failing to occur.

Consider a data set: male, male, female, female, female, other. The distribution of this data can be classified as:

- The number of males in the data set is 2.
- The number of females in the data set is 3.
- The number of individuals identifying as other is 1.
- The number of non-males is 4.

Seaborn which is another extremely useful library for data visualization in Python. The Seaborn library is built on top of Matplotlib and offers many advanced data visualization capabilities. Though, the Seaborn library can be used to draw a variety of charts such as matrix plots, grid plots, regression plots etc.,

Downloading the Seaborn Library

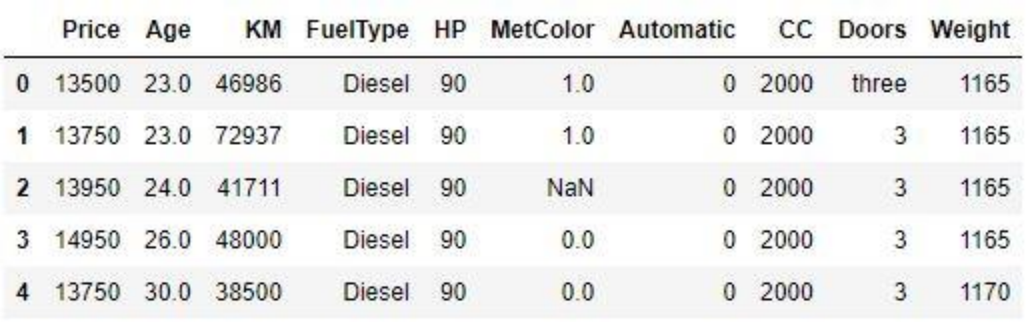
The seaborn library can be downloaded in a couple of ways. If you are using pip installer for Python libraries, you can execute the following command to download the library:

pip install seaborn

Alternatively, if you are using the Anaconda distribution of Python, you can use execute the following command to download the seaborn library:

conda install seaborn

Here we will use the dataset called Toyota Corolla, which is a cars dataset.



	Price	Age	KM	FuelType	HP	MetColor	Automatic	CC	Doors	Weight
0	13500	23.0	46986	Diesel	90	1.0	0	2000	three	1165
1	13750	23.0	72937	Diesel	90	1.0	0	2000	3	1165
2	13950	24.0	41711	Diesel	90	NaN	0	2000	3	1165
3	14950	26.0	48000	Diesel	90	0.0	0	2000	3	1165
4	13750	30.0	38500	Diesel	90	0.0	0	2000	3	1170

Scatter Plot:

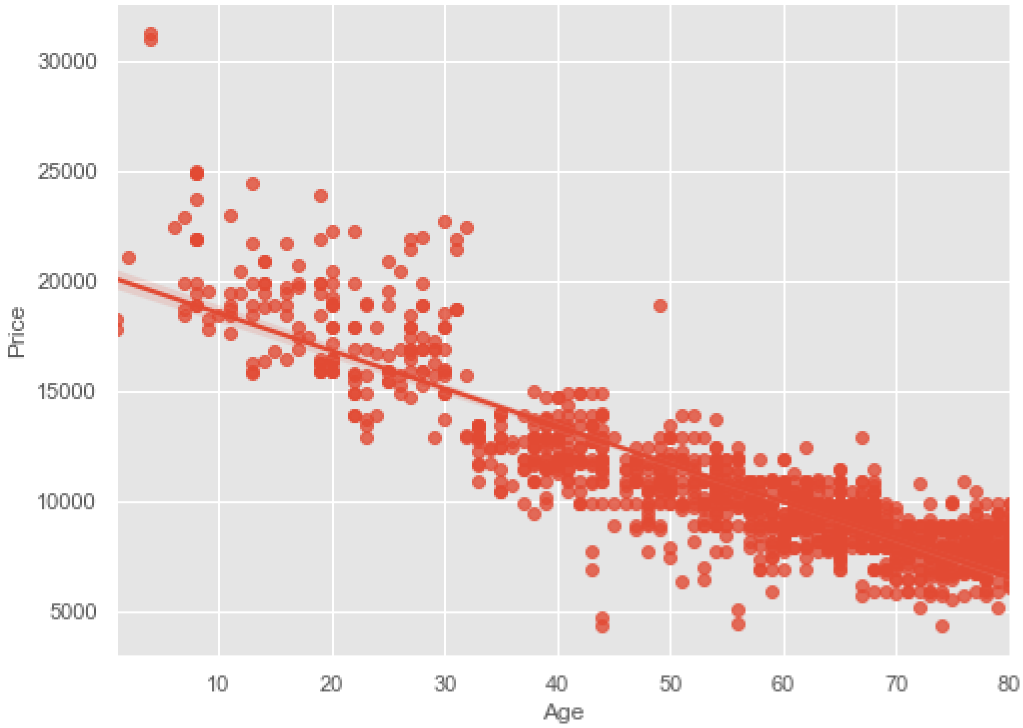
Scatter plots can be used to show a linear relationship between two or three data points using the seaborn library. A Scatter plot of price vs age with default arguments will be like this:

```
plt.style.use("ggplot")

plt.figure(figsize=(8,6))

sns.regplot(x = cars_data["Age"], y = cars_data["Price"])

plt.show()
```



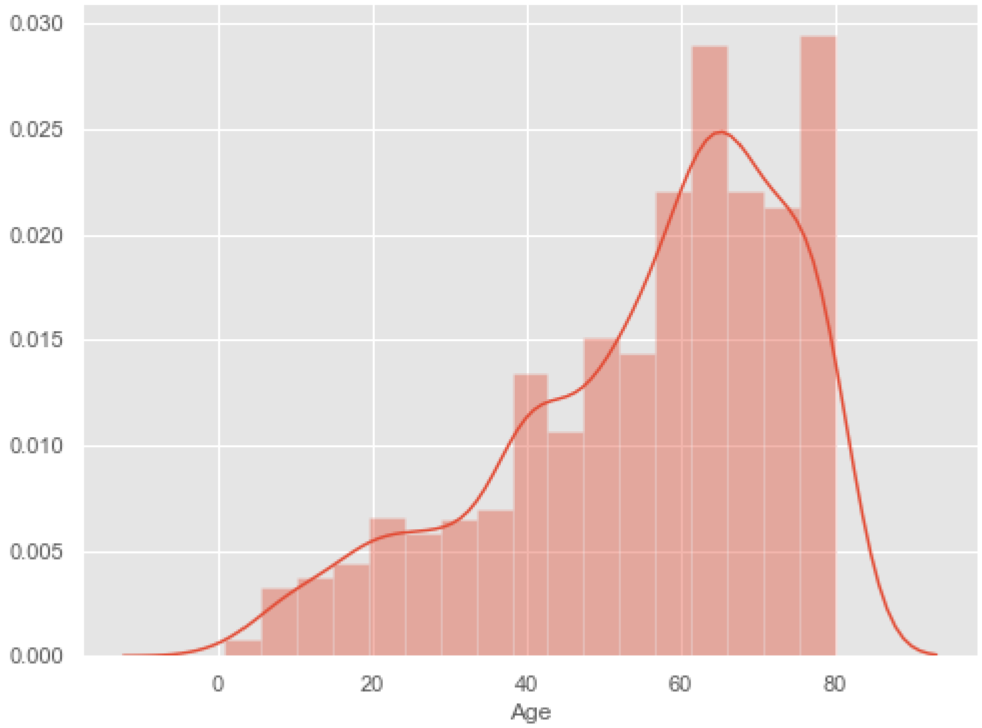
Histogram:

In order to draw a histogram in Seaborn, we have a function called distplot and inside that, we have to pass the variable which we want to include. Histogram with default kernel density estimate:

```
plt.figure(figsize=(8,6))

sns.distplot(cars_data['Age'])

plt.show()
```



For the x-axis, we are giving Age and the histogram is by default include kernel density estimate (kde). Kernel density estimate is the curved line along with the bins or the edges of the frequency of the Ages.

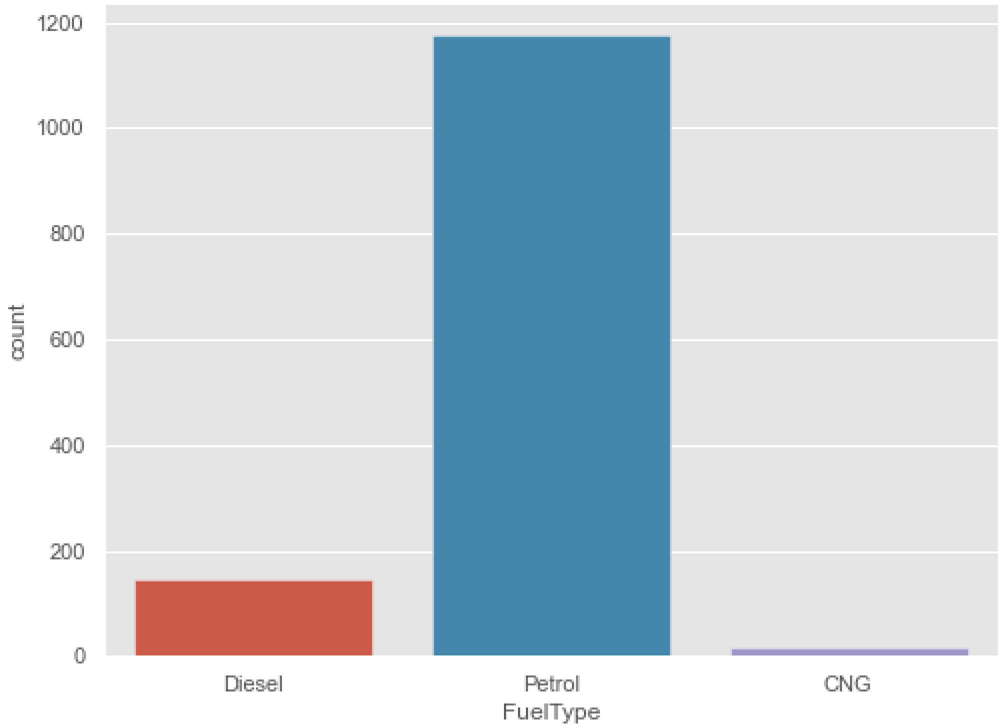
Bar Plot:

Bar plot is for categorical variables. Bar plot is the commonly used plot because of its simplicity and it is easy to understand data through them. You can plot a bar plot in seaborn using the count plot library. It is simple. Let us plot a bar plot of FuelType.

```
plt.figure(figsize=(8,6))

sns.countplot(x="FuelType", data=cars_data)

plt.show()
```



In the y-axis, we have got the frequency distribution of FuelType of the cars.

Grouped Bar Plot:

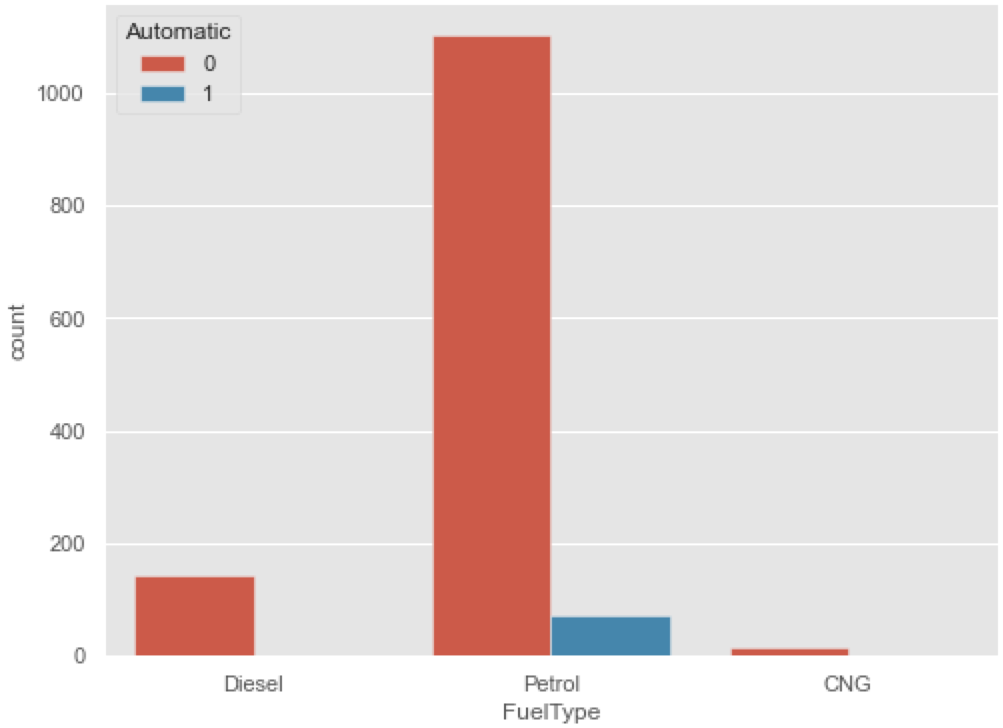
We can plot a bar plot between two variables. That is called grouped bar plot. Let us plot a bar plot of FuelType distributed by different values of the Automatic column.

```
plt.figure(figsize=(8,6))

sns.countplot(x="FuelType", data=cars_data,

            hue="Automatic")

plt.show()
```


Automatic
0
1
count
1000
800
600
400
200
0
Diesel
Petrol
CNG
FuelType

CONCLUSION:

Assignment No. 4

TITLE: Design a python program to explore the concepts of sampling & sampling distributions.

OBJECTIVES:

1. Develop a Python program to delve into the principles of sampling and sampling distributions.
2. Implement random sampling techniques to generate samples from a specified population.

PROBLEM STATEMENT: Design a python program to explore the concepts of sampling & sampling distributions. Generate random samples from a given population analyze their characteristics & demonstrate how the sample mean evolves as more samples are taken. Investigate properties of sampling distributions (Mean & S.D).

OUTCOME: Students will be able to,

1. Design and implement a Python program to explore the concepts of sampling and sampling distributions.
2. Generate random samples from a specified population and analyze their characteristics.

THEORY-CONCEPT:

What Is a Sampling Distribution?

A sampling distribution is a concept used in statistics. It is a probability distribution of a statistic obtained from a larger number of samples drawn from a specific population. The sampling distribution of a given population is the distribution of frequencies of a range of different outcomes that could possibly occur for a statistic of a population. This allows entities like governments and businesses to make more well-informed decisions based on the information they gather.

How Sampling Distributions Work?

Data allows statisticians, researchers, marketers, analysts, and academics to make important conclusions about specific topics and information. It can help businesses make decisions about their future and boost their performance, or it can help governments plan for services needed by a group of people.

A lot of data drawn and used are actually samples rather than populations. A sample is a subset of a population. Put simply, a sample is a smaller part of a larger group. As such, this smaller portion is meant to be representative of the population as a whole.

ampling distributions (or the distribution of data) are statistical metrics that determine whether an event or certain outcome will take place. This distribution depends on a few different factors, including the sample size, the sampling process involved, and the population as a whole. There are a few steps involved with sampling distribution. These include:

- Choosing a random sample from the overall population
- Determine a certain statistic from that group, which could be the standard deviation, median, or mean
- Establishing a frequency distribution of each sample
- Mapping out the distribution on a graph

Once the information is gathered, plotted, and analyzed, researchers can make inferences and conclusions. This can help them make decisions about what to expect in the future. For instance, governments may be able to invest in infrastructure projects based on the needs of a certain community or a company may decide to proceed with a new business venture if the sampling distribution suggests a positive outcome.

For example, Let's say a medical researcher wants to compare the average weight of all babies born in North America from 1995 to 2005 to those from South America within the same time period. Since they cannot draw the data for the entire population within a reasonable amount of time, they would only use 100 babies in each continent to make a conclusion. The data used is the sample and the average weight calculated is the sample mean.

Now suppose they take repeated random samples from the general population and compute the sample mean for each sample group instead. So, for North America, they pull data for 100 newborn weights recorded in the U.S., Canada, and Mexico as follows:

Four 100 samples from select hospitals in the U.S.

Five 70 samples from Canada

Three 150 records from Mexico

The researcher ends up with a total of 1,200 weights of newborn babies grouped in 12 sets. They also collect sample data of 100 birth weights from each of the 12 countries in South America.

The average weight computed for each sample set is the sampling distribution of the mean. Not just the mean can be calculated from a sample. Other statistics, such as the standard deviation, variance, proportion, and range can be calculated from sample data. The standard deviation and variance measure the variability of the sampling distribution.

CONCLUSION:

Assignment No. 5

TITLE: Sampling Distribution**.**

OBJECTIVES:

1. Generate random samples from a given population to understand the concept of sampling and its application in statistical analysis.
2. Explore the properties of sampling distributions by analyzing the distribution of sample means obtained from multiple samples.

PROBLEM STATEMENT: Sampling Distribution: Write a program to generate multiple samples (at least 30) from the population & calculate the mean for each sample.

OUTCOME: Students will be able to, understanding of the concept of sampling distributions and their practical implications in data analysis and decision-making processes.

THEORY-CONCEPT:

What Is a Sampling Distribution?

A sampling distribution is a concept used in statistics. It is a probability distribution of a statistic obtained from a larger number of samples drawn from a specific population. The sampling distribution of a given population is the distribution of frequencies of a range of different outcomes that could possibly occur for a statistic of a population. This allows entities like governments and businesses to make more well-informed decisions based on the information they gather.

How Sampling Distributions Work?

Data allows statisticians, researchers, marketers, analysts, and academics to make important conclusions about specific topics and information. It can help businesses make decisions about their future and boost their performance, or it can help governments plan for services needed by a group of people.

A lot of data drawn and used are actually samples rather than populations. A sample is a subset of a population. Put simply, a sample is a smaller part of a larger group. As such, this smaller portion is meant to be representative of the population as a whole.

ampling distributions (or the distribution of data) are statistical metrics that determine whether an event or certain outcome will take place. This distribution depends on a few different factors, including the sample size, the sampling process involved, and the population as a whole. There are a few steps involved with sampling distribution. These include:

- Choosing a random sample from the overall population
- Determine a certain statistic from that group, which could be the standard deviation, median, or mean
- Establishing a frequency distribution of each sample
- Mapping out the distribution on a graph

Once the information is gathered, plotted, and analyzed, researchers can make inferences and conclusions. This can help them make decisions about what to expect in the future. For instance, governments may be able to invest in infrastructure projects based on the needs of a certain community or a company may decide to proceed with a new business venture if the sampling distribution suggests a positive outcome.

CONCLUSION:

Assignment No. 6

TITLE: Random Sampling.

OBJECTIVES:

1. Develop an understanding of random sampling as a method for selecting a subset of data from a larger dataset.
2. Design a systematic and unbiased approach to extract a representative sample from the chosen dataset.

PROBLEM STATEMENT: Random Sampling: Implement a random sampling process to extract a representative sample from the chosen dataset.

OUTCOME: Students will be able to, obtain a representative subset of data from the chosen dataset, ensuring it accurately reflects the characteristics of the entire population.

THEORY-CONCEPT:

Simple random sampling is a statistical method in which everyone in a population has an equal chance of being selected into a sample. The sample represents a smaller and more manageable portion of the people that can be studied and analyzed. It's a fundamental technique to gather data and make inferences about a population.

Simple random sampling is a technique where every item in the population has an even chance and likelihood of being selected. Here, the selection of items entirely depends on luck or probability. Therefore, this sampling technique is also a method of chance.

The sample size in a simple random sampling method should ideally be more than a few hundred so that it can be applied appropriately. This method is theoretically simple to understand but difficult to implement practically. Working with a large sample size isn't an easy task, and it can sometimes be challenging to find a realistic sampling bias frame.

Simple Random Sampling Methods

Researchers follow these methods to select a simple random sample:

1. They prepare a list of all the population members initially, and each member is marked with a specific number (for example, if there are nth members, then they will be numbered from 1 to N).

2. Researchers from this population choose random samples using random number tables and random number generator software. Researchers prefer random number generator software, as no human interference is necessary to generate samples.

Two approaches aim to minimize any biases in the process of this method:

01. Method of lottery

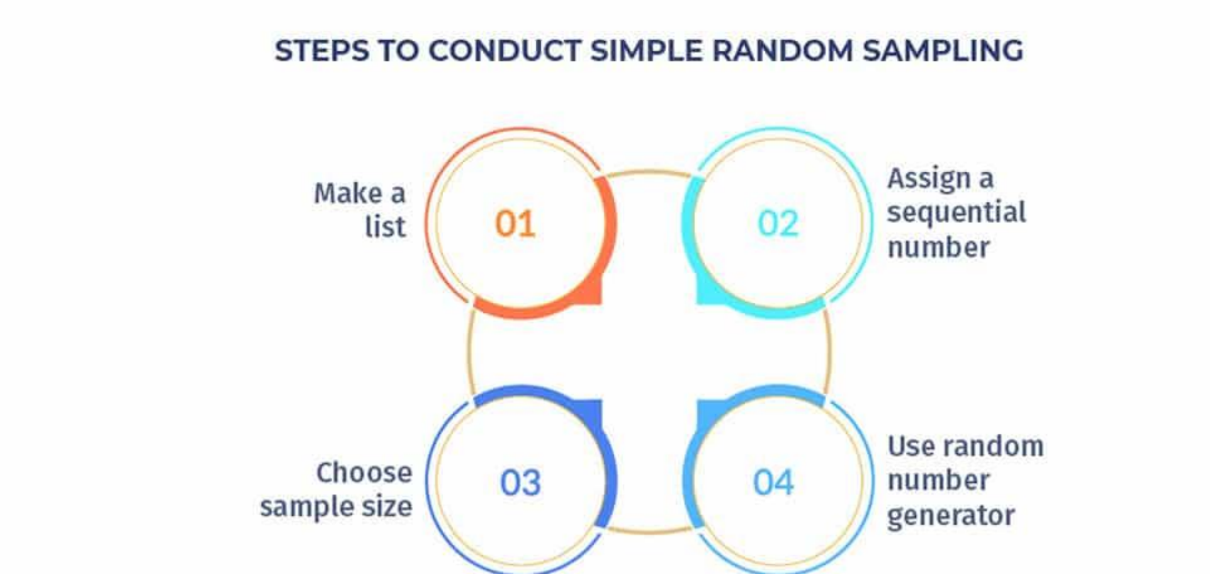
Using the lottery method is one of the oldest ways and is a mechanical example of a random sample. Researchers draw numbers from the box randomly to choose samples. In this method, the researcher gives each member of the population a number.

02. Use of random numbers

Using random numbers is an alternative method that also involves numbering the population.

Simple Random Sampling Steps

Simple random sampling is a crucial method in statistical analysis for drawing unbiased conclusions about a population. Below are the steps to perform simple random sampling to select a sample of 100 employees out of a total of 500 in an organization.


STEPS TO CONDUCT SIMPLE RANDOM SAMPLING
Make a list
01
02
Assign a sequential number
Choose sample size
03
04
Use random number generator

Step 1: Make a List

To start simple random sampling, first, make a complete list of all 500 employees in the organization. It’s important that the list includes the names of every employee to guarantee that each person is considered.

A precise and thorough list is crucial to ensure the sampling accurately reflects the entire population.

Step 2: Assign a Sequential Number

After creating the list of employees, the next thing to do is give each employee a number in order. This is your sampling frame (the list from which you draw your sample). This numbering helps organize the list, making identifying each person in the group easier.

Every employee should have their own number, starting from 1 and going up to n, which is the total number of employees in the organization.

Step 3: Choose Sample Size

Selecting the right sample size is important in simple random sampling. In this situation, we’ve chosen a sample of 100 employees from a total population of 500. It’s essential to pick a sample size that’s large enough for dependable results but still practical for analysis.

Step 4: Use a Random Number Generator

To choose a sample from the group, use a random number generator. First, find the total number of people (Step 2) and decide how many we want in our sample (Step 3).Then, use a random number table or generator to create 100 different random numbers between 1 and 500. These numbers match the order given to each employee, which helps you pick who will be in the sample.

This method ensures that each employee has an equal opportunity for selection, maintaining fairness and impartiality in sample selection. It is important to note that Simple Random Sampling is just one of many sampling methods available, and it may not always be the best option for your specific research needs.

CONCLUSION:

Assignment No. 7

TITLE: Develop a python program for hypothesis testing.

OBJECTIVES:

1. Gain a comprehensive understanding of hypothesis testing, specifically focusing on comparing means of two samples, as a fundamental statistical technique for inference.
2. Design and develop a Python program capable of implementing hypothesis tests such as the t-test or z-test, which are commonly used for comparing means.

PROBLEM STATEMENT: Develop a python program for hypothesis testing, focusing on comparing means of two samples. Implement a statistical test such as t-test or z-test, to analyze whether there is significant difference between the means.

OUTCOME: Students will be able to, Utilization of statistical tests such as t-test or z-test to evaluate whether there exists a significant difference between the means of the two samples, promoting deeper understanding of hypothesis testing methodologies.

THEORY-CONCEPT:

What is Hypothesis Testing?

Hypothesis testing is an essential part in inferential statistics where we use observed data in a sample to draw conclusions about unobserved data — often the population.

Implication of hypothesis testing:

1. clinical research: widely used in psychology, biology and healthcare research to examine the effectiveness of clinical trials
2. A/B testing: can be applied in business context to improve conversions through testing different versions of campaign incentives, website designs etc
3. feature selection in machine learning: filter-based feature selection methods use different statistical tests to determine the feature importance
4. college or university: well, if you major in statistics or data science, it is likely to appear in your exams …

we have two types of hypotheses: null and alternate.

A null hypothesis is one's default belief or argument about a subject matter. In the case of the earth's shape, the null hypothesis was that the earth was flat.

An alternate hypothesis is a belief or argument a person might try to establish. Aristotle and

Eratosthenes argued that the earth was spherical.

The following steps explain how we can test a hypothesis:

Step #1 - Define the Null and Alternative Hypotheses

Before making any test, we must first define what we are testing and what the default assumption is about the subject. For example, we'll be testing if the average weight of 10-year-old children is more than 32kg.

Our null hypothesis is that 10 year old children weigh 32 kg on average. Our alternate hypothesis is that the average weight is more than 32kg. Ho denotes a null hypothesis, while H1 denotes an alternate hypothesis.

Ho = 32

H1 = 32

Step #2 - Choose a Significance Level

The significance level is a threshold for determining if the test is valid. It gives credibility to our hypothesis test to ensure we are not just luck-dependent but have enough evidence to support our claims. We usually set our significance level before conducting our tests. The criterion for determining our significance value is known as p-value.

A lower p-value means that there is stronger evidence against the null hypothesis, and therefore, a greater degree of significance. A p-value of 0.05 is widely accepted to be significant in most fields of science. P-values do not denote the probability of the outcome of the result, they just serve as a benchmark for determining whether our test result is due to chance. For our test, our p-value will be 0.05.

Step #3 - Collect Data and Calculate a Test Statistic

You can obtain your data from online data stores or conduct your research directly. Data can be scraped or researched online. The methodology might depend on the research you are trying to conduct.

We can calculate our test using any of the appropriate hypothesis tests. This can be a T-test, Z-test, Chi-squared, and so on. There are several hypothesis tests, each suiting different purposes and research questions.

T-test is used for comparison of two sets of data when we don't know the population standard deviation. It's a parametric test, meaning it makes assumptions about the distribution of the data. These assumptions include that the data is normally distributed and that the variances of the two groups are equal. In a more simple and practical sense, imagine that we have test scores in a class for males and females, but we don't know how different or similar these scores are. We can use a t-test to see if there's a real difference.

The Z-test is used for comparison between two sets of data when the population standard deviation is known. It is also a parametric test, but it makes fewer assumptions about the distribution of data. The z-test assumes that the data is normally distributed, but it does not assume that the variances of the two groups are equal. In our class test example, with the t-test, we can say that if we already know how spread out the scores are in both groups, we can now use the z-test to see if there's a difference in the average scores.

The Chi-squared test is used to compare two or more categorical variables. The chi-squared test is a non-parametric test, meaning it does not make any assumptions about the distribution of data. It can be used to test a variety of hypotheses, including whether two or more groups have equal proportions.

Step #4 - Decide on the Null Hypothesis Based on the Test Statistic and Significance Level

After conducting our test and calculating the test statistic, we can compare its value to the predetermined significance level. If the test statistic falls beyond the significance level, we can decide to reject the null hypothesis, indicating that there is sufficient evidence to support our alternative hypothesis.

On the other contrary, if the test statistic does not exceed the significance level, we fail to reject the null hypothesis, signifying that we do not have enough statistical evidence to conclude in favor of the alternative hypothesis.

Step #5 - Interpret the Results

Depending on the decision made in the previous step, we can interpret the result in the context of our study and the practical implications. For our case study, we can interpret whether we have significant evidence to support our claim that the average weight of 10 year old children is more than 32kg or not.

CONCLUSION:

Assignment No. 8

TITLE: Create a python program that performs a two-sample hypothesis test & introduce the basics of ANOVA.

OBJECTIVES:

1. Implement functionality within the program to enable users to specify the significance level for hypothesis testing.
2. Introduce users to the basics of Analysis of Variance (ANOVA) as a statistical method for comparing means across multiple groups.

PROBLEM STATEMENT Create a python program that performs a two-sample hypothesis test & introduce the basics of ANOVA. Allow users to input data for multiple groups, choose a significance level & conduct a two-sample t-test for comparing means or ANOVA for comparing means or ANOVA for comparing means across more than two groups.

OUTCOME: Students will be able to, develop practical skills in statistical hypothesis testing and apply them to real-world datasets or research scenarios.

THEORY-CONCEPT:

What is a Two Sample T Hypothesis Test?

A two sample t hypothesis tests also known as independent t-test is used to analyze the difference between two unknown population means. The Two-sample T-test is used when the two small samples (n< 30) are taken from two different populations and compared. The underlying chart makes use of the T distribution.

Assumptions of Two Sample T Hypothesis Tests

The sample should be randomly selected from the two population

Samples are independent to each other

Two sample sizes must me less than 30

Samples collected from the population are normally distributed

When Would You Use a Two Sample T Hypothesis Tests?

The two-sample t test most likely used to compare two process means, when the data is having one nominal variable and one measurement variable. It is a hypothesis test of means. Use two sample Z test if the sample size is more than 30.

The two sample hypothesis t tests is used to compare two population means, while analysis of variance (ANOVA) is the best option if more than two group means to be compared.

Two sample T hypothesis tests are performed when the two group samples are statistically independent to each other, while the paired t-test is used to compare the means of two dependent or paired groups.

Steps to Calculate Two Sample T Hypothesis Test (Equal Variance)

i. State the claim of the test and determine the null hypothesis and alternative hypothesis
ii. Determine the level of significance
iii. Calculate degrees of freedom
iv. Find the critical value
v. Calculate the test statistics
vi. Make a decision, the null hypothesis will be rejected if the test statistic is less than or equal to the critical value
vii. Finally, Interpret the decision in the context of the original claim.

What is Analysis of Variance (ANOVA)?

ANOVA stands for Analysis of Variance. It is a statistical method used to analyze the differences between the means of two or more groups or treatments. It is often used to determine whether there are any statistically significant differences between the means of different groups.

ANOVA compares the variation between group means to the variation within the groups. If the variation between group means is significantly larger than the variation within groups, it suggests a significant difference between the means of the groups.

ANOVA calculates an F-statistic by comparing between-group variability to within-group variability. If the F-statistic exceeds a critical value, it indicates significant differences between group means.

Example of How to Use ANOVA

A researcher might, for example, test students from multiple colleges to see if students from one of the colleges consistently outperform students from the other colleges. In a business application, an R&D researcher might test two different processes of creating a product to see if one process is better than the other in terms of cost efficiency.

The type of ANOVA test used depends on a number of factors. It is applied when data needs to be experimental. Analysis of variance is employed if there is no access to statistical software resulting in computing ANOVA by hand. It is simple to use and best suited for small samples. With many experimental designs, the sample sizes have to be the same for the various factor level combinations.

ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample t-tests. However, it results in fewer type I errors and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources. It is employed with subjects, test groups, between groups and within groups.

One-Way ANOVA Versus Two-Way ANOVA

There are two main types of ANOVA: one-way (or unidirectional) and two-way. There also variations of ANOVA. For example, MANOVA (multivariate ANOVA) differs from ANOVA as the former tests for multiple dependent variables simultaneously while the latter assesses only one dependent variable at a time. One-way or two-way refers to the number of independent variables in your analysis of variance test. A one-way ANOVA evaluates the impact of a sole factor on a sole response variable. It determines whether all the samples are the same. The one-way ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups.

A two-way ANOVA is an extension of the one-way ANOVA. With a one-way, you have one independent variable affecting a dependent variable. With a two-way ANOVA, there are two independents. For example, a two-way ANOVA allows a company to compare worker productivity based on two independent variables, such as salary and skill set. It is utilized to observe the interaction between the two factors and tests the effect of two factors at the same time.

CONCLUSION:

Assignment No. 9

TITLE: Regression Analysis.

OBJECTIVES:

1. Implement functionality within the program to enable users to specify a dependent variable of interest for regression modeling.
2. Introduce users to the principles of multiple regression analysis as a statistical method for modeling the relationship between multiple independent variables and a single dependent variable.

PROBLEM STATEMENT: Create a python program for multiple regression analysis. Allow users to input a dataset with multiple independent variables, specify a dependent variable & perform multiple regression to model the complex relationship.

OUTCOME: Students will be able to,

1. apply data analytics in given dataset.
2. implement linear regression model in given dataset.

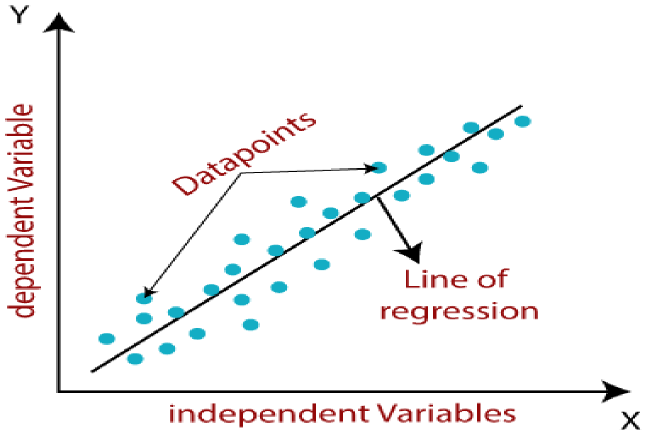
THEORY-CONCEPT:

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

Linear-regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions. Linear regression can be applied to various areas in business and academic study.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Mathematically, we can represent a linear regression as:

$y = a_0 + a_1 x + \varepsilon$

Here,

Y= Dependent Variable (Target Variable)

X= Independent Variable (predictor Variable)

a0= intercept of the line (Gives an additional degree of freedom)

a1 = Linear regression coefficient (scale factor to each input value).

ε = random error

Finding the best fit line:

When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error. The different values for weights or the coefficient of lines (a0, a1) gives a different line of regression, so we need to calculate the best values for a0 and a1 to find the best fit line, so to calculate this we use cost function.

For Linear Regression, we use the **Mean Squared Error (MSE)** cost function, which is the average of squared error occurred between the predicted values and actual values. It can be written as:

$$\text{MSE}= 1\frac{1}{N}\sum_{i=1}^{n}(y_i \ - (a_1x_i + a_0))^2$$

For the above linear equation, MSE can be calculated as:

Where,

N=Total number of observation
Y_i = Actual value
$(a_1x_i+a_0)$ = Predicted value.

Types of Linear Regression

Linear regression can be further divided into two types of the algorithm:

i. **Simple Linear Regression**

   If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

ii. **Multiple Linear regression**

   If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Examples of linear-regression

1. **Evaluating trends and sales estimates**

You can also use linear-regression analysis to try to predict a salesperson's total yearly sales (the dependent variable) from independent variables such as age, education, and years of experience

2. **Assess risk in an insurance company**

Linear regression techniques can be used to analyze risk. For example, an insurance company might have limited resources with which to investigate homeowners' insurance claims; with linear regression, the company's team can build a model for estimating claims costs. The analysis could help company leaders make important business decisions about what risks to take.

3. Sports analysis

Linear regression is not always about business. It is also important in sports. For instance, you might wonder if the number of games won by a basketball team in a season is related to the average number of points the team scores per game. A scatterplot indicates that these variables are linearly related. The number of games won and the average number of points scored by the opponent are also linearly related. These variables have a negative relationship. As the number of games won increases, the average number of points scored by the opponent decreases. With linear regression, you can model the relationship of these variables. A good model can be used to predict how many games teams will win.

CONCLUSION:

Assignment No. 10

TITLE: Logistic Regression.

OBJECTIVES: 1. To learn logistic regression.

2. To learn confusion matrix.

PROBLEM STATEMENT: Build a python program that allows users to input data for a binary outcome, perform logistic regression, evaluate the model's performance using ROC analysis & visualize the ROC curve.

OUTCOME: Students will be able to,

1. apply data analytics in given dataset.
2. implement linear regression model in given dataset.
3. compute confusion matrix for given dataset

THEORY:

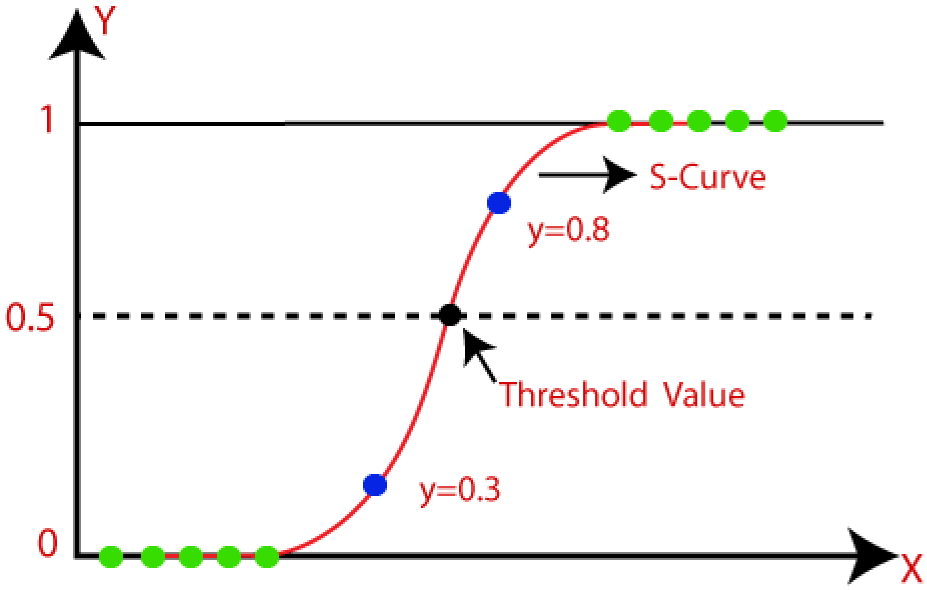
Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much like the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:



Logistic regression is a commonly used model in various industries such as banking, healthcare because when compared to other classification models, the logistic regression model is easily interpreted.

Assumptions for Logistic Regression:

- The dependent variable must be categorical in nature.
- The independent variable should not have multi-collinearity.

Logistic Regression Equation:

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

- We know the equation of the straight line can be written as:

$$y= b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

- In Logistic Regression y can be between 0 and 1 only, so for this let us divide the above equation by (1-y):

$$\frac{y}{1-y}\text{; 0 for y= 0, and infinity for y=1}$$

- But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

The above equation is the final equation for Logistic Regression.

Application of Logistic Regression

1. **Credit scoring**

ID Finance is a financial company that makes predictive models for credit scoring. They need their models to be easily interpretable. They can be asked by a regulator about a certain decision at any moment. Data preprocessing for credit scoring modeling includes such a step like reducing correlated variables. It is difficult if you have more than 15 variables in your model. For logistic regression, it is easy to find out which variables affect the result of the predictions more and which ones less. It is also possible to find the optimal number of features and eliminate redundant variables with methods like recursive feature elimination.

At the final step, they can export prediction results to an Excel file, and analytic even without technical skills can get insights from this data.

At some point, ID finance refused the use of third-party statistical applications and rewrote their algorithms for building models in Python. This has led to a significant increase in the speed of model development. But they did not abandon logistic regression in favor of more complex algorithms. Logistic regression is widely used in credit scoring and it shows remarkable results.

2. **Medicine**

Medical information is gathered in such a way that when a research group studies a biological molecule and its properties, they publish a paper about it. Thus, there is a huge amount of medical data about various compounds, but they are not combined into a single database.

Miroculus is a company that develops express blood test kits. Its goal is to identify diseases that are affected by genes, such as oncology diseases. The company entered into an agreement with

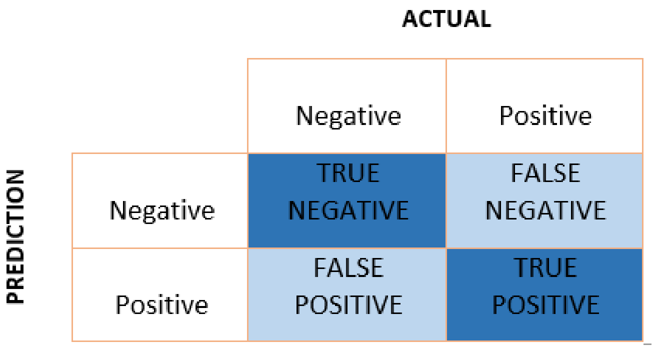
Microsoft to develop an algorithm to identify the relationship between certain micro-RNA and genes.

The developers used a database of scientific articles and applied text analysis methods to obtain feature vectors. The text was split into the sentences, the entities were extracted, labeled data generated from known relations, and after several other text transformation methods, each sentence was converted into a 200-dimensional vector. After converting the text and extracting the distinguishing features, a classification was made for the presence of a link between microRNA and a certain gene. Algorithms such as logistic regression, support vector machine, and random forest were considered as models. Logistic regression was selected because it demonstrated the best results in speed and accuracy.

Logistic regression is well suited for this data type when we need to predict a binary answer. Is there a connection between the elements or not? Thanks to this algorithm, the accuracy of a quick blood test has been increased.

What is a confusion matrix?

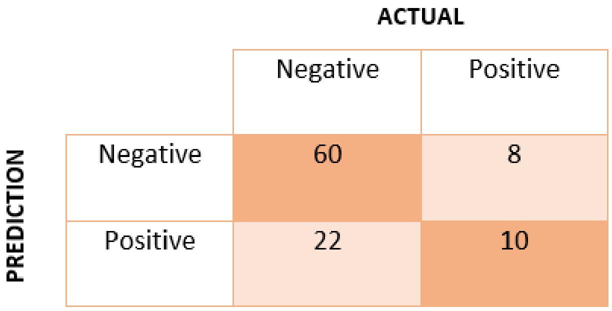
It is a matrix of size 2×2 for binary classification with actual values on one axis and predicted on another.



		ACTUAL	
		Negative	Positive
PREDICTION	Negative	TRUE NEGATIVE	FALSE NEGATIVE
	Positive	FALSE POSITIVE	TRUE POSITIVE

EXAMPLE

A machine learning model is trained to predict tumor in patients. The test dataset consists of 100 people.



		ACTUAL	
		Negative	Positive
PREDICTION	Negative	60	8
	Positive	22	10

Confusion Matrix for tumor detection

True Positive (TP) — model correctly predicts the positive class (prediction and actual both are positive). In the above example, **10 people** who have tumors are predicted positively by the model.

True Negative (TN) — model correctly predicts the negative class (prediction and actual both are negative). In the above example, **60 people** who do not have tumors are predicted negatively by the model.

False Positive (FP) — model gives the wrong prediction of the negative class (predicted-positive, actual-negative). In the above example, **22 people** are predicted as positive of having a tumor, although they do not have a tumor. FP is also called a **TYPE I** error.

False Negative (FN) — model wrongly predicts the positive class (predicted-negative, actual-positive). In the above example, **8 people** who have tumors are predicted as negative. FN is also called a **TYPE II** error.

With the help of these four values, we can calculate True Positive Rate (TPR), False Negative Rate (FPR), True Negative Rate (TNR), and False Negative Rate (FNR).

$$TPR = \frac{TP}{Actual\ Positive} = \frac{TP}{TP + FN}$$

$$FNR = \frac{FN}{Actual\ Positive} = \frac{FN}{TP + FN}$$

$$TNR = \frac{TN}{Actual\ Negative} = \frac{TN}{TN + FP}$$

$$FPR = \frac{FP}{Actual\ Negative} = \frac{FP}{TN + FP}$$

Precision, Recall

Both precision and recall are crucial for information retrieval, where positive class mattered the most as compared to negative. While searching something on the web, the model does not care about something **irrelevant** and **not retrieved** (this is the true negative case). Therefore, only TP, FP, FN are used in Precision and Recall.

Precision

Out of all the positive predicted, what percentage is truly positive.

$$Precision = \frac{TP}{TP + FP}$$

The precision value lies between 0 and 1.

Recall

Out of the total positive, what percentage are predicted positive. It is the same as TPR (true positive rate).

$$Recall = \frac{TP}{TP + FN}$$

ROC Curve

The ROC curve is a graphical representation of a model’s ability to distinguish between classes. It plots the True Positive Rate (Sensitivity) against the False Positive Rate (1 — Specificity) for different classification thresholds.

Interpreting the ROC Curve

A perfect classifier would hug the top-left corner of the ROC space, indicating high sensitivity and specificity. A random guess would result in a diagonal line from the bottom-left to the top-right, indicating an AUC of 0.5. The closer the curve is to the top-left corner, the better the model’s performance.

CONCLUSION:

Assignment No. 11

TITLE: Design a python program for clustering analysis.

OBJECTIVES:

1. Develop a Python program capable of performing clustering analysis, a technique used to identify inherent patterns or groupings within a dataset.
2. Enable users to input datasets containing relevant features for clustering analysis, ensuring flexibility and applicability across various domains.

PROBLEM STATEMENT: Design a python program for clustering analysis. Allow users to input a dataset & choose a clustering algorithm. Implement the selected algorithm, visualize the clusters & provide insights into grouping patterns within the data.

OUTCOME: Students will be able to, Implement the selected clustering algorithm on the input dataset, efficiently grouping similar data points into clusters based on specified parameters.

THEORY:

Cluster analysis, also known as clustering, is a method of data mining that groups similar data points together. The goal of cluster analysis is to divide a dataset into groups (or clusters) such that the data points within each group are more similar to each other than to data points in other groups. This process is often used for exploratory data analysis and can help identify patterns or relationships within the data that may not be immediately obvious. There are many different algorithms used for cluster analysis, such as k-means, hierarchical clustering, and density-based clustering. The choice of algorithm will depend on the specific requirements of the analysis and the nature of the data being analyzed.

Cluster Analysis is the process to find similar groups of objects in order to form clusters. It is an unsupervised machine learning-based algorithm that acts on unlabelled data. A group of data points would comprise together to form a cluster in which all the objects would belong to the same group.

The given data is divided into different groups by combining similar objects into a group. This group is nothing but a cluster. A cluster is nothing but a collection of similar data which is grouped together.

For example, consider a dataset of vehicles given in which it contains information about different vehicles like cars, buses, bicycles, etc. As it is unsupervised learning there are no class labels like Cars, Bikes, etc for all the vehicles, all the data is combined and is not in a structured manner.

ow our task is to convert the unlabeled data to labelled data and it can be done using clusters.

The main idea of cluster analysis is that it would arrange all the data points by forming clusters like cars cluster which contains all the cars, bikes clusters which contains all the bikes, etc. Simply it is the partitioning of similar objects which are applied to unlabeled data.

Applications Of Cluster Analysis:

- It is widely used in image processing, data analysis, and pattern recognition.
- It helps marketers to find the distinct groups in their customer base and they can characterize their customer groups by using purchasing patterns.
- It can be used in the field of biology, by deriving animal and plant taxonomies and identifying genes with the same capabilities.
- It also helps in information discovery by classifying documents on the web.

Clustering Methods

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

1. **Partitioning Method:** It is used to make partitions on the data in order to form clusters. If "n" partitions are done on "p" objects of the database then each partition is represented by a cluster and n < p. The two conditions which need to be satisfied with this Partitioning Clustering Method are:
    - One objective should only belong to only one group.
    - There should be no group without even a single purpose.

In the partitioning method, there is one technique called iterative relocation, which means the object will be moved from one group to another to improve the partitioning

2. **Hierarchical Method:** In this method, a hierarchical decomposition of the given set of data objects is created. We can classify hierarchical methods and will be able to know the purpose of classification on the basis of how the hierarchical decomposition is formed. There are two types of approaches for the creation of hierarchical decomposition, they are:
    - Agglomerative Approach: The agglomerative approach is also known as the bottom-up approach. Initially, the given data is divided into which objects form separate groups. Thereafter it keeps on merging the objects or the groups that are close to one another which means that they exhibit similar properties. This merging process continues until the termination condition holds.
    - Divisive Approach: The divisive approach is also known as the top-down approach. In this approach, we would start with the data objects that are in the same cluster. The group of individual clusters is divided into small clusters by continuous iteration. The iteration continues until the condition of termination is met or until each cluster contains one object.

Once the group is split or merged then it can never be undone as it is a rigid method and is not so flexible. The two approaches which can be used to improve the Hierarchical Clustering Quality in Data Mining are: –

- One should carefully analyze the linkages of the object at every partitioning of hierarchical clustering.
- One can use a hierarchical agglomerative algorithm for the integration of hierarchical agglomeration. In this approach, first, the objects are grouped into micro-clusters. After grouping data objects into microclusters, macro clustering is performed on the microcluster.

3. **Density-Based Method:** The density-based method mainly focuses on density. In this method, the given cluster will keep on growing continuously as long as the density in the neighbourhood exceeds some threshold, i.e, for each data point within a given cluster. The radius of a given cluster has to contain at least a minimum number of points.
4. **Grid-Based Method:** In the Grid-Based method a grid is formed using the object together,i.e, the object space is quantized into a finite number of cells that form a grid structure. One of the major advantages of the grid-based method is fast processing time and it

is dependent only on the number of cells in each dimension in the quantized space. The processing time for this method is much faster so it can save time.

5. **Model-Based Method:** In the model-based method, all the clusters are hypothesized in order to find the data which is best suited for the model. The clustering of the density function is used to locate the clusters for a given model. It reflects the spatial distribution of data points and also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. Therefore it yields robust clustering methods.
6. **Constraint-Based Method:** The constraint-based clustering method is performed by the incorporation of application or user-oriented constraints. A constraint refers to the user expectation or the properties of the desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. The user or the application requirement can specify constraints.

CONCLUSION:

Assignment No. 12

TITLE: Design a python program for linear Regression.

OBJECTIVES:

1. To develop a program capable of performing linear regression on historical stock prices using relevant features such as previous day closing price, trading volume, and market indices.

PROBLEM STATEMENT: Stock Price Forecasting: Develop a program to perform linear regression on historical stock prices use relevant features like previous day closing price, trading volume &market indices. Access the models ability to predict future stock prices.

OUTCOME: Students will be able to, Comparison of the performance of linear regression with other forecasting techniques.

THEORY:

Stock price forecasting involves predicting future price movements based on historical data and relevant features. Linear regression is a statistical method commonly used for this purpose, as it can capture linear relationships between independent variables (such as previous day closing price, trading volume, and market indices) and the dependent variable (future stock prices). The model estimates coefficients for each feature, allowing for the prediction of future stock prices based on observed historical data.

Experiment Setup:

Data Collection: Historical stock price data along with relevant features such as previous day closing price, trading volume, and market indices will be collected from financial data sources or APIs.

Data Preprocessing: The collected data will be preprocessed to handle missing values, scale numerical features, and encode categorical variables if necessary.

Feature Selection: Relevant features such as previous day closing price, trading volume, and market indices will be selected for modeling.

Model Development: A linear regression model will be developed using the selected features to predict future stock prices.

Model Evaluation: The trained model will be evaluated using appropriate performance metrics such as mean squared error (MSE), root mean squared error (RMSE), and R-squared to assess its predictive accuracy.

Prediction: The model will be used to predict future stock prices for a specified time period.

Comparison: The performance of the linear regression model will be compared to alternative forecasting methods such as time series models or machine learning algorithms.

CONCLUSION: