

Title of the Assignment: Predict the price of the Uber ride from a given pickup point to the agreed drop-off location.

Perform following tasks:

1. Pre-process the dataset.
2. Identify outliers.
3. Check the correlation.
4. Implement linear regression and random forest regression models.
5. Evaluate the models and compare their respective scores like R2, RMSE, etc.

Dataset Description:

The project is about on world's largest taxi company Uber inc. In this project, we're looking to predict the fare for their future transactional cases. Uber delivers service to lakhs of customers daily. Now it becomes really important to manage their data properly to come up with new business ideas to get best results. Eventually, it becomes really important to estimate the fare prices accurately.

Link for Dataset:<https://www.kaggle.com/datasets/yasserh/uber-fares-dataset>
Objective of the Assignment:

Students should be able to pre-process dataset and identify outliers, to check correlation and implement linear regression and random forest regression models. Evaluate them with respective scores like R2, RMSE etc.

Prerequisite:

1. Basic knowledge of Python
2. Concept of pre-processing data
3. Basic knowledge of Data Science and Big Data Analytics.

Contents of the Theory:

1. Data Pre-processing
2. Linear regression
3. Random Forest regression models
4. Box Plot
5. Outliers
6. Haversine
7. Mathplotlib
8. Mean Squared Error

Data Preprocessing:

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.

Why do we need Data Preprocessing?

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

It involves below steps:

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data
- Splitting dataset into training and test set
- Feature scaling

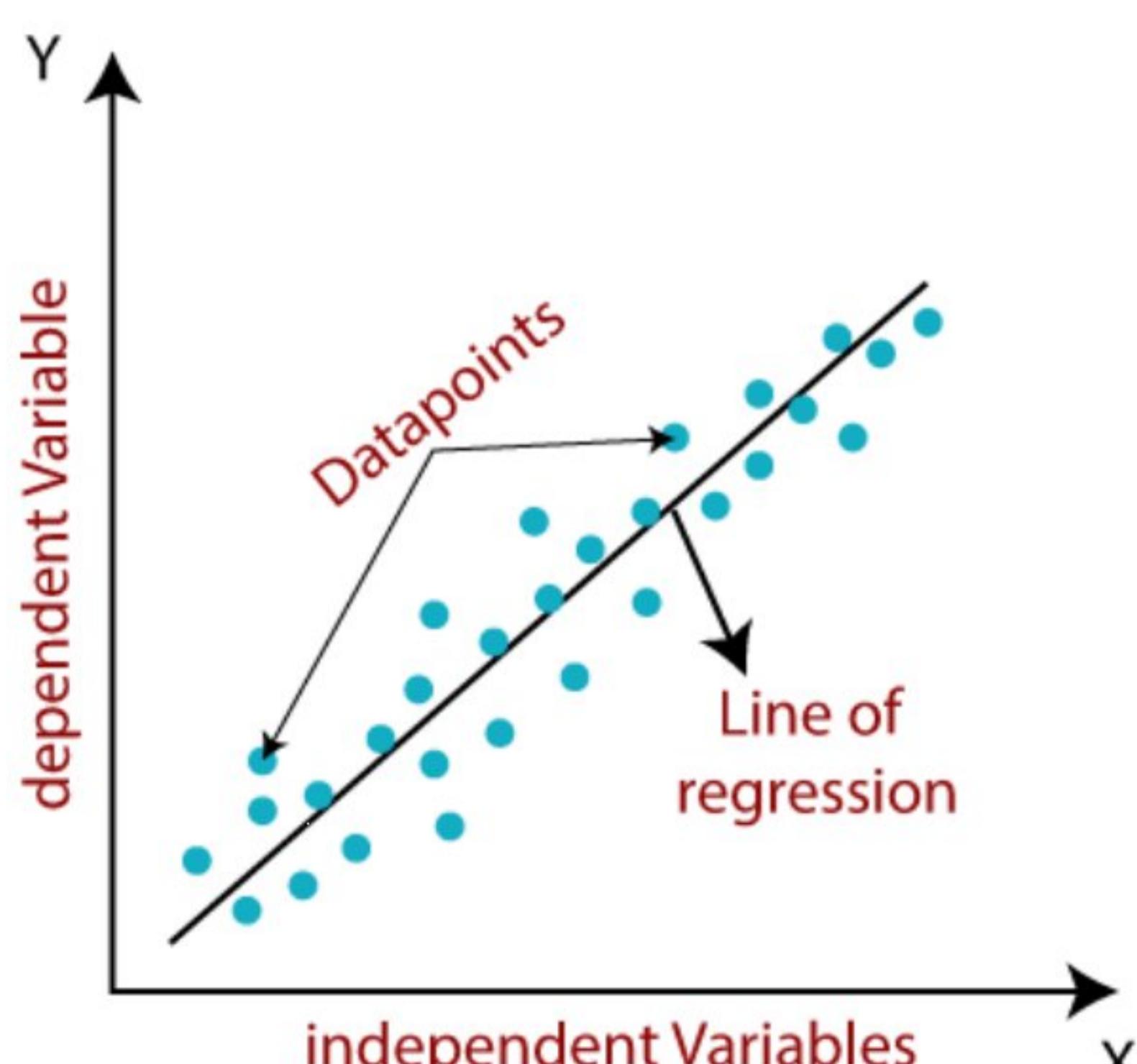
Linear Regression:

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price**, etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (x) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it indicates how the value of the dependent variable is changing according to the value of the independent variable.

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

Random Forest



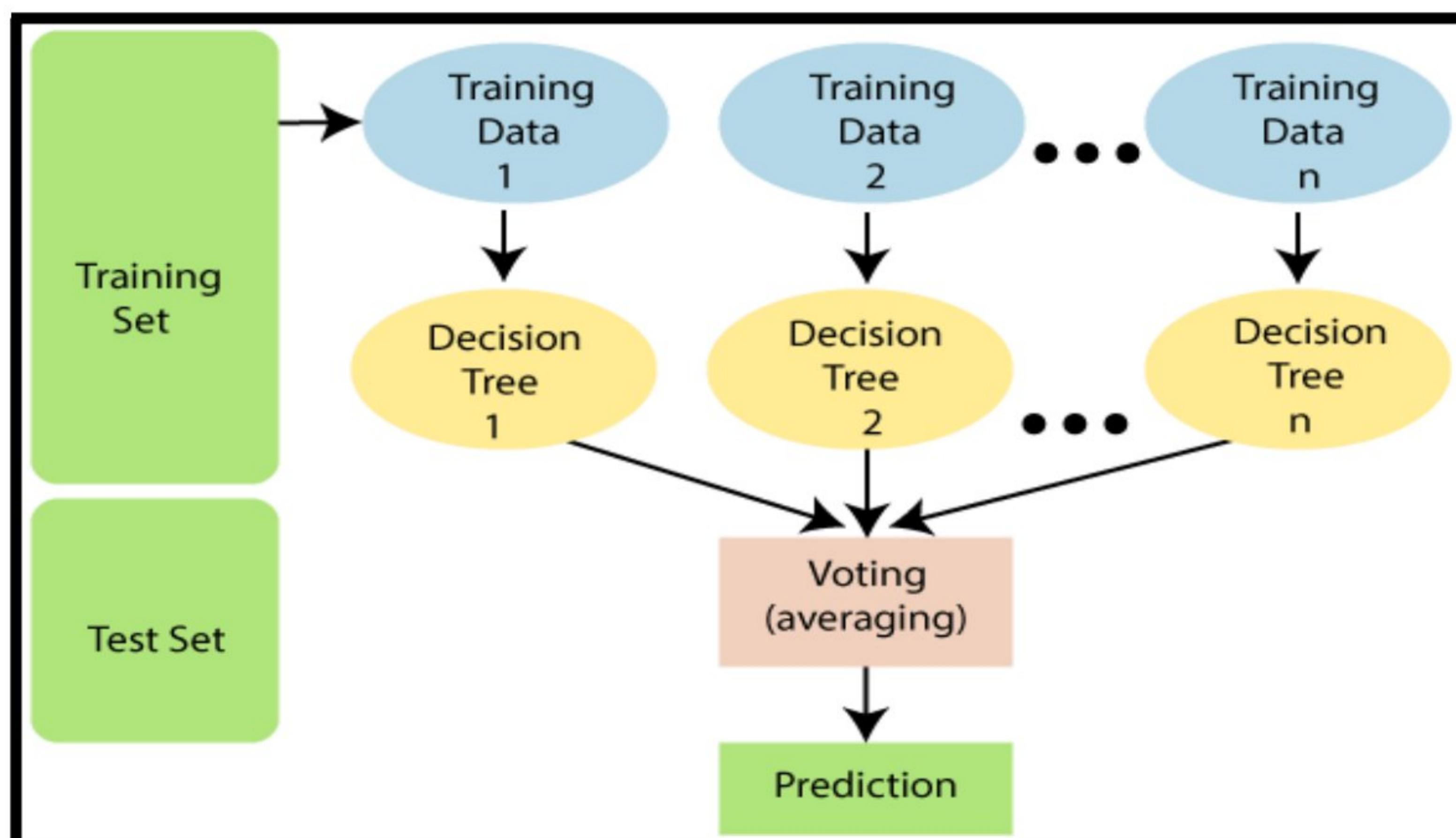
1. x	It is a vector or a formula.
2. data	It is the data frame.
3. notch	It is a logical value set as true to draw a notch.
4. Var width	It is also a logical value set as true to draw the width of the box same as the sample size.
5. names	It is the group of labels that will be printed under each boxplot.
6. main	It is used to give a title to the graph.

Random Forest Regression Models:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning**, which is a process of combining multiple classifier to solve a complex problem and to improve the performance of the model.

As the name suggests, "**Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.**" Instead of relying on one decision tree, the random

prediction from each tree and based on the majority votes of predictions, and it predicts the Final output.



The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Boxplot:

Boxplots are a measure of how well data is distributed across a data set. This divides the data set into three quartiles. This graph represents the minimum, maximum, average, first quartile, and the third quartile in the data set. Boxplot is also useful in comparing the distribution of data in a data set by drawing a boxplot for each of them.

R provides a `boxplot()` function to create a boxplot. There is the following syntax of `boxplot()` function:

`boxplot(x, data, notch, varwidth, names, main)`

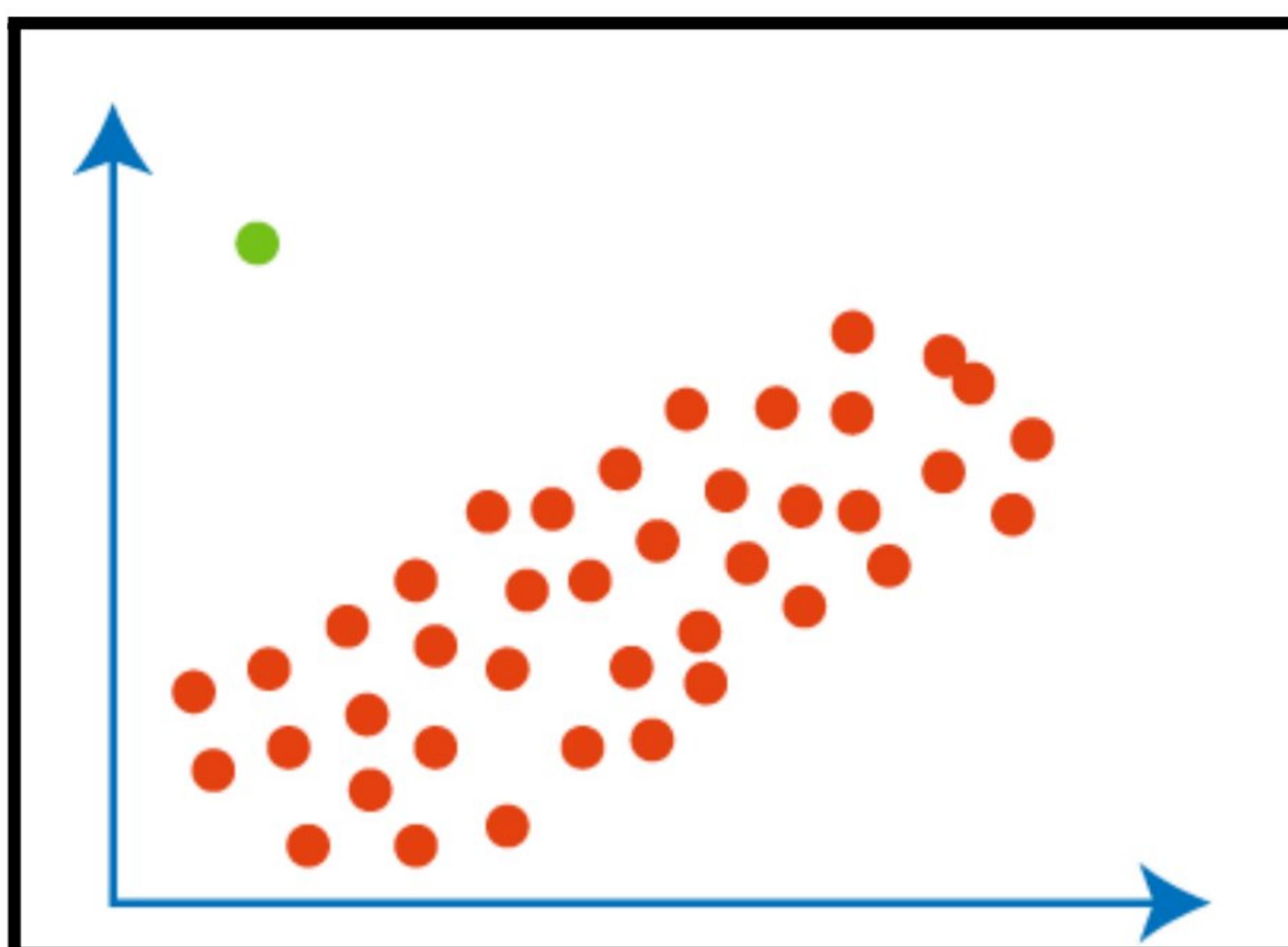
`boxplot(x, data, notch, varwidth, names, main)`



As the name suggests, "outliers" refer to the data points that exist outside of what is to be expected. The major thing about the outliers is what you do with them. If you are going to analyze any task to analyse data sets, you will always have some assumptions based on how this data is generated. If you find some data points that are likely to contain some form of error, then these are definitely outliers, and depending on the context, you want to overcome those errors. The data mining process involves the analysis and prediction of data that the data holds. In 1969, Grubbs introduced the first definition of outliers.

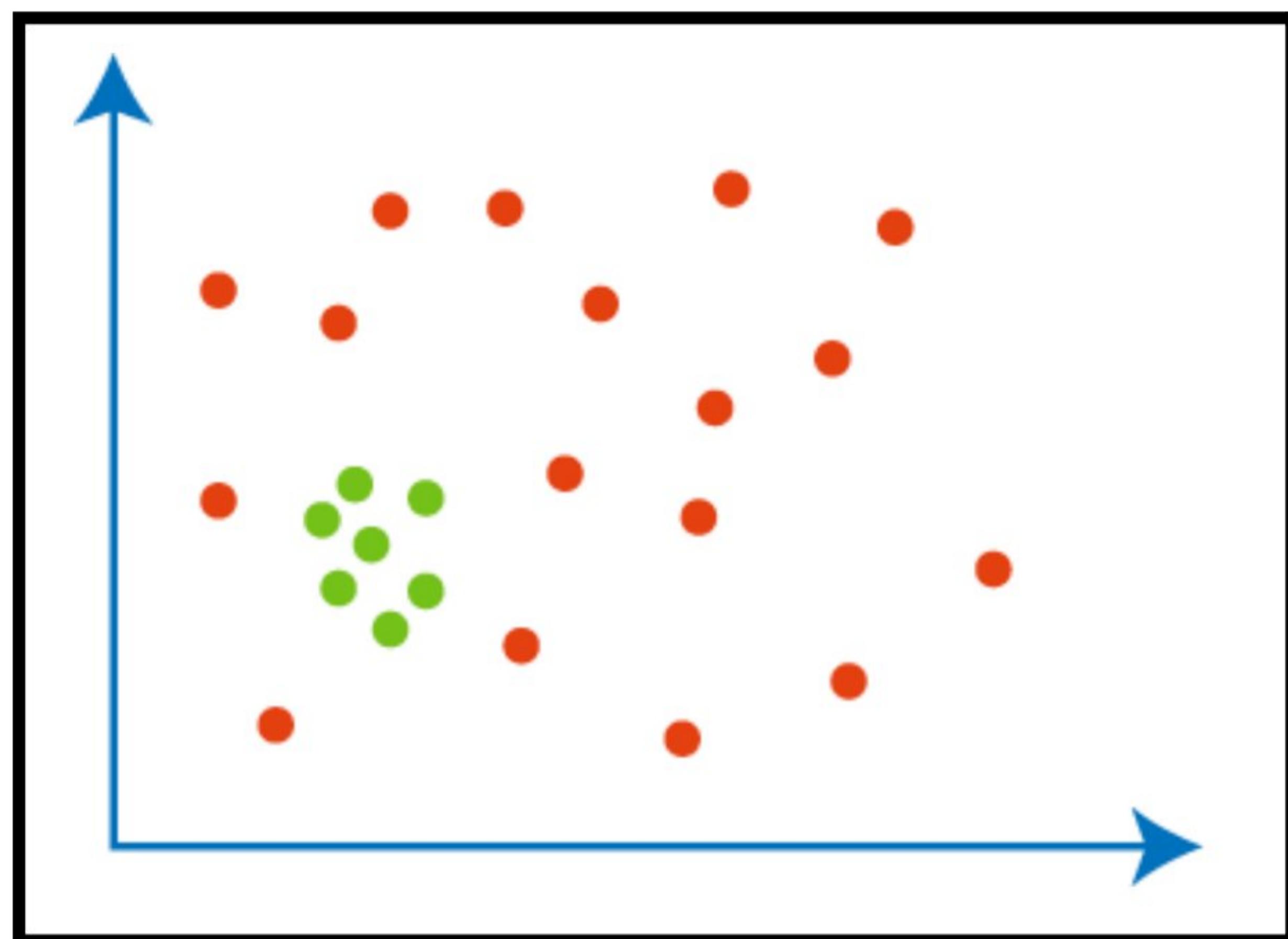
Global Outliers

Global outliers are also called point outliers. Global outliers are taken as the simplest form of outliers. When data points deviate from all the rest of the data points in a given data set, it is known as the global outlier. In most cases, all the outlier detection procedures are targeted to determine the global outliers. The green data point is the global outlier, known as the global



Collective Outliers

In a given set of data, when a group of data points deviates from the rest of the data set is called collective outliers. Here, the particular set of data objects may not be outliers, but when you consider the data objects as a whole, they may behave as outliers. To identify the types of different outliers, you need to go through background information about the relationship between the behavior of outliers shown by different data objects. For example, in an Intrusion Detection System, the DOS package from one system to another is taken as normal behavior. Therefore, if this happens with the various computer simultaneously, it is considered abnormal behavior, and as a whole, they are called collective outliers. The green data points as a whole represent the collective outlier.

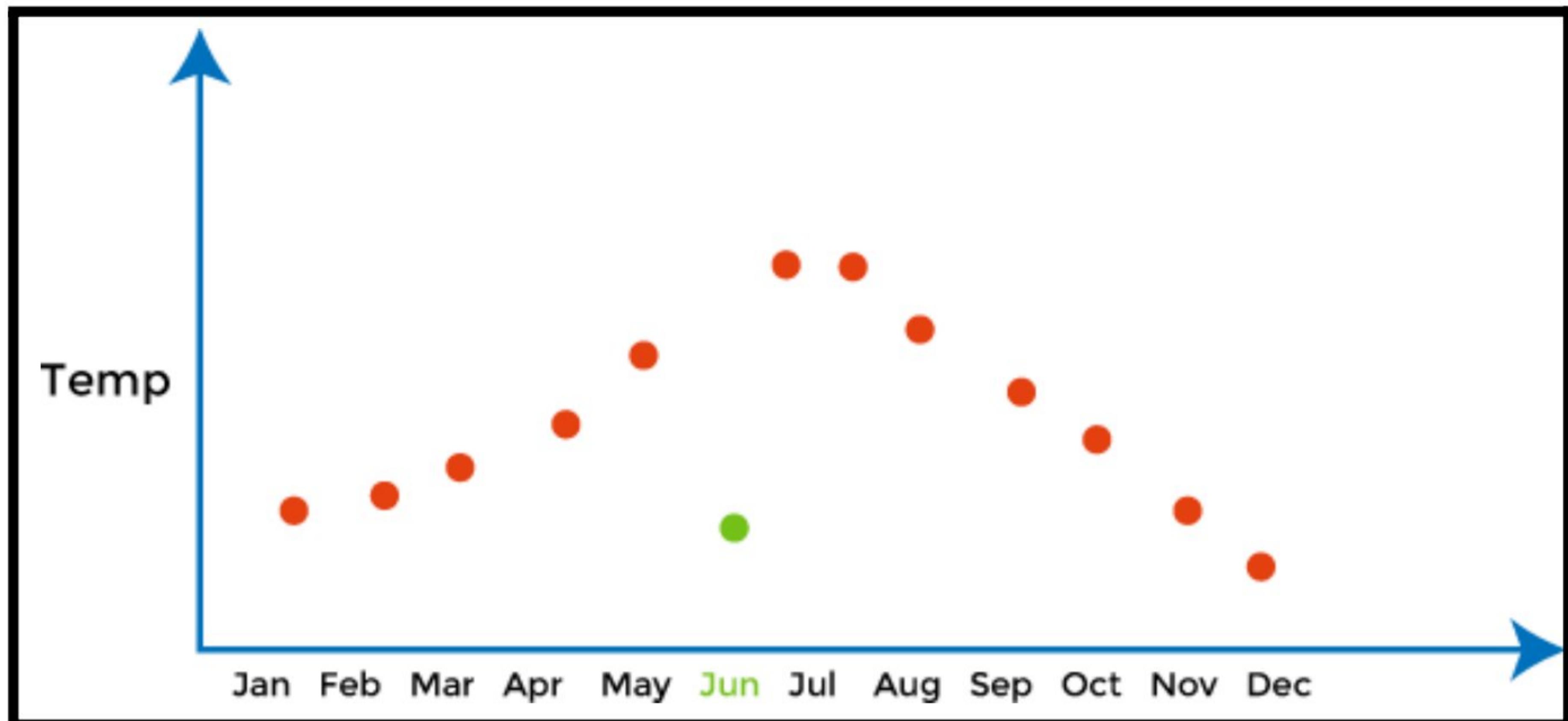


Collective Outliers

In a given set of data, when a group of data points deviates from the rest of the data set is called collective outliers. Here, the particular set of data objects may not be outliers, but when you consider the data objects as a whole, they may behave as outliers. To identify the types of different outliers, you need to go through background information about the relationship between the behavior of outliers shown by different data objects. For example, in an Intrusion Detection System, the DOS package from one system to another is taken as normal behaviour. Therefore, if this happens with the various computer simultaneously, it is considered abnormal behaviour, and as a whole, they are called collective outliers. The green data points as a whole represents the collective outlier.

Contextual Outliers

As the name suggests, "Contextual" means this outlier introduced within a context. For example, in the speech recognition technique, the single background noise. Contextual outliers are also known as Conditional outliers. These types of outliers happen if a data object deviates from the other data points because of any specific condition in a given data set. As we know, there are two types of attributes of objects of data: contextual attributes and behavioural attributes. Contextual outlier analysis enables the users to examine outliers in different contexts and conditions, which can be useful in various applications. For example, A temperature reading of 45 degrees Celsius may behave as an outlier in a rainy season. Still, it will behave like a normal data point in the context of a summer season. In the given diagram, a green dot representing the low-temperature value in June is a contextual outlier since the same value in December is not an outlier.



Haversine:

The Haversine formula calculates the shortest distance between two points on a sphere using their latitudes and longitudes measured along the surface. It is important for use in navigation.

Matplotlib:

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

Mean Squared Error;

The **Mean Squared Error (MSE)** or **Mean Squared Deviation (MSD)** of an estimator measures the average of error squares i.e. the average squared difference between the estimated values and true value. It is a risk function, corresponding to the expected value of the squared error loss. It is always non – negative and values close to zero are better. The MSE is the second moment of the error (about the origin) and thus incorporates both the variance of the estimator and its bias.

Conclusion:

In this way we have explored Concept correlation and implement linear regression and random forest regression models.

Assignment Questions:

1. What is data preprocessing?
2. Define Outliers?
3. What is Linear Regression?
4. What is Random Forest Algorithm?
5. Explain: pandas, numpy?

Assignment No. 1

```
In [1]: #Importing the required libraries
```

```
import pandas as pd  
import numpy as np  
import seaborn as sns  
import matplotlib as plt
```

```
In [2]: #Importing the dataset
```

```
df = pd.read_csv("uber.csv")
```

```
In [3]: # 1.Pre-process the dataset
```

```
df.head()
```

```
Out[3]:
```

	Unnamed: 0	key	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
0	24238194	2015-05-07 19:52:06.0000003	7.5	2015-05-07 19:52:06 UTC	-73.999817	40.738354	-73.999512	40.723217	1
1	27835199	2009-07-17 20:04:56.0000002	7.7	2009-07-17 20:04:56 UTC	-73.994355	40.728225	-73.994710	40.750325	1
2	44984355	2009-08-24 21:45:00.00000061	12.9	2009-08-24 21:45:00 UTC	-74.005043	40.740770	-73.962565	40.772647	1
3	25894730	2009-06-26 08:22:21.0000001	5.3	2009-06-26 08:22:21 UTC	-73.976124	40.790844	-73.965316	40.803349	3
4	17610152	2014-08-28 17:47:00.000000188	16.0	2014-08-28 17:47:00 UTC	-73.925023	40.744085	-73.973082	40.761247	5

```
In [4]: #To get the required information of the dataset
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 200000 entries, 0 to 199999  
Data columns (total 9 columns):  
 #   Column           Non-Null Count  Dtype     
 ---  --    
 0   Unnamed: 0        200000 non-null  int64    
 1   key              200000 non-null  object    
 2   fare_amount      200000 non-null  float64   
 3   pickup_datetime  200000 non-null  object    
 4   pickup_longitude 200000 non-null  float64   
 5   pickup_latitude   200000 non-null  float64   
 6   dropoff_longitude 199999 non-null  float64   
 7   dropoff_latitude  199999 non-null  float64   
 8   passenger_count   200000 non-null  int64    
 dtypes: float64(5), int64(2), object(2)  
 memory usage: 13.7+ MB
```

```
In [5]: #To get number of columns in the dataset
```

```
df.columns
```

```
Out[5]: Index(['Unnamed: 0', 'key', 'fare_amount', 'pickup_datetime',  
               'pickup_longitude', 'pickup_latitude', 'dropoff_longitude',  
               'dropoff_latitude', 'passenger_count'],  
              dtype='object')
```

```
In [6]: #To drop unnamed column as it isn't required
```

```
df = df.drop(['Unnamed: 0', 'key'], axis=1)
```

```
In [7]: df.head()
```

	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
0	7.5	2015-05-07 19:52:06 UTC	-73.999817	40.738354	-73.999512	40.723217	1
1	7.7	2009-07-17 20:04:56 UTC	-73.994355	40.728225	-73.994710	40.750325	1
2	12.9	2009-08-24 21:45:00 UTC	-74.005043	40.740770	-73.962565	40.772647	1
3	5.3	2009-06-26 08:22:21 UTC	-73.976124	40.790844	-73.965316	40.803349	3
4	16.0	2014-08-28 17:47:00 UTC	-73.925023	40.744085	-73.973082	40.761247	5

```
In [8]: # To get the total (Rows,Columns)
df.shape
```

```
Out[8]: (200000, 7)
```

```
In [9]: #To get the type of each column
df.dtypes
```

```
Out[9]: fare_amount      float64
pickup_datetime        object
pickup_longitude       float64
pickup_latitude        float64
dropoff_longitude      float64
dropoff_latitude       float64
passenger_count        int64
dtype: object
```

```
In [10]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200000 entries, 0 to 199999
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype  
 ---  --  
 0   fare_amount      200000 non-null  float64 
 1   pickup_datetime  200000 non-null  object  
 2   pickup_longitude 200000 non-null  float64 
 3   pickup_latitude  200000 non-null  float64 
 4   dropoff_longitude 199999 non-null  float64 
 5   dropoff_latitude  199999 non-null  float64 
 6   passenger_count  200000 non-null  int64  
dtypes: float64(5), int64(1), object(1)
memory usage: 10.7+ MB
```

```
In [11]: # To get statistics of each columns
df.describe()
```

	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
count	200000.000000	200000.000000	200000.000000	199999.000000	199999.000000	200000.000000
mean	11.359955	-72.527638	39.935885	-72.525292	39.923890	1.684535
std	9.901776	11.437787	7.720539	13.117408	6.794829	1.385997
min	-52.000000	-1340.648410	-74.015515	-3356.666300	-881.985513	0.000000
25%	6.000000	-73.992065	40.734796	-73.991407	40.733823	1.000000
50%	8.500000	-73.981823	40.752592	-73.980093	40.753042	1.000000
75%	12.500000	-73.967154	40.767158	-73.963658	40.768001	2.000000
max	499.000000	57.418457	1644.421482	1153.572603	872.697628	208.000000

```
In [12]: # Filling Missing values  
df.isnull().sum()
```

```
Out[12]: fare_amount      0  
pickup_datetime      0  
pickup_longitude      0  
pickup_latitude       0  
dropoff_longitude     1  
dropoff_latitude      1  
passenger_count       0  
dtype: int64
```

```
In [13]: df['dropoff_latitude'].fillna(value=df['dropoff_latitude'].mean(),inplace=True)  
df['dropoff_longitude'].fillna(value=df['dropoff_longitude'].median(),inplace=True)
```

```
In [14]: df.isnull().sum()
```

```
Out[14]: fare_amount      0  
pickup_datetime      0  
pickup_longitude      0  
pickup_latitude       0  
dropoff_longitude     0  
dropoff_latitude      0  
passenger_count       0  
dtype: int64
```

```
In [19]: df.head()
```

```
Out[19]: fare_amount  pickup_datetime  pickup_longitude  pickup_latitude  dropoff_longitude  dropoff_latitude  passenger_count  hour  day  month  year  dayofweek  
0          7.5  2015-05-07  
           19:52:06+00:00      -73.999817      40.738354      -73.999512      40.723217          1   19    7     5  2015        3  
1          7.7  2009-07-17  
           20:04:56+00:00      -73.994355      40.728225      -73.994710      40.750325          1   20   17     7  2009        4  
2         12.9  2009-08-24  
           21:45:00+00:00      -74.005043      40.740770      -73.962565      40.772647          1   21   24     8  2009        0  
3          5.3  2009-06-26  
           08:22:21+00:00      -73.976124      40.790844      -73.965316      40.803349          3   8   26     6  2009        4  
4         16.0  2014-08-28  
           17:47:00+00:00      -73.925023      40.744085      -73.973082      40.761247          5   17   28     8  2014        3
```

```
In [20]: # drop the column 'pickup_datetime' using drop()  
# 'axis= 1' drops the specified column  
df = df.drop('pickup_datetime',axis= 1)
```

```
In [21]: df.head()
```

```
Out[21]: fare_amount  pickup_longitude  pickup_latitude  dropoff_longitude  dropoff_latitude  passenger_count  hour  day  month  year  dayofweek  
0          7.5      -73.999817      40.738354      -73.999512      40.723217          1   19    7     5  2015        3  
1          7.7      -73.994355      40.728225      -73.994710      40.750325          1   20   17     7  2009        4  
2         12.9      -74.005043      40.740770      -73.962565      40.772647          1   21   24     8  2009        0  
3          5.3      -73.976124      40.790844      -73.965316      40.803349          3   8   26     6  2009        4  
4         16.0      -73.925023      40.744085      -73.973082      40.761247          5   17   28     8  2014        3
```

```
In [22]: df.dtypes
```

```
Out[15]: fare_amount      float64
pickup_datetime    object
pickup_longitude   float64
pickup_latitude    float64
dropoff_longitude  float64
dropoff_latitude   float64
passenger_count    int64
dtype: object
```

```
In [16]: # Column pickup_datetime is in wrong format(object).
# Convert it to DateTime Format
df.pickup_datetime = pd.to_datetime(df.pickup_datetime, errors='coerce')
```

```
In [17]: df.dtypes
```

```
Out[17]: fare_amount      float64
pickup_datetime    datetime64[ns, UTC]
pickup_longitude   float64
pickup_latitude    float64
dropoff_longitude  float64
dropoff_latitude   float64
passenger_count    int64
dtype: object
```

```
In [18]: # To segregate each time of date and time
df= df.assign(hour= df.pickup_datetime.dt.hour,
              day= df.pickup_datetime.dt.day,
              month= df.pickup_datetime.dt.month,
              year= df.pickup_datetime.dt.year,
              dayofweek= df.pickup_datetime.dt.dayofweek)
```

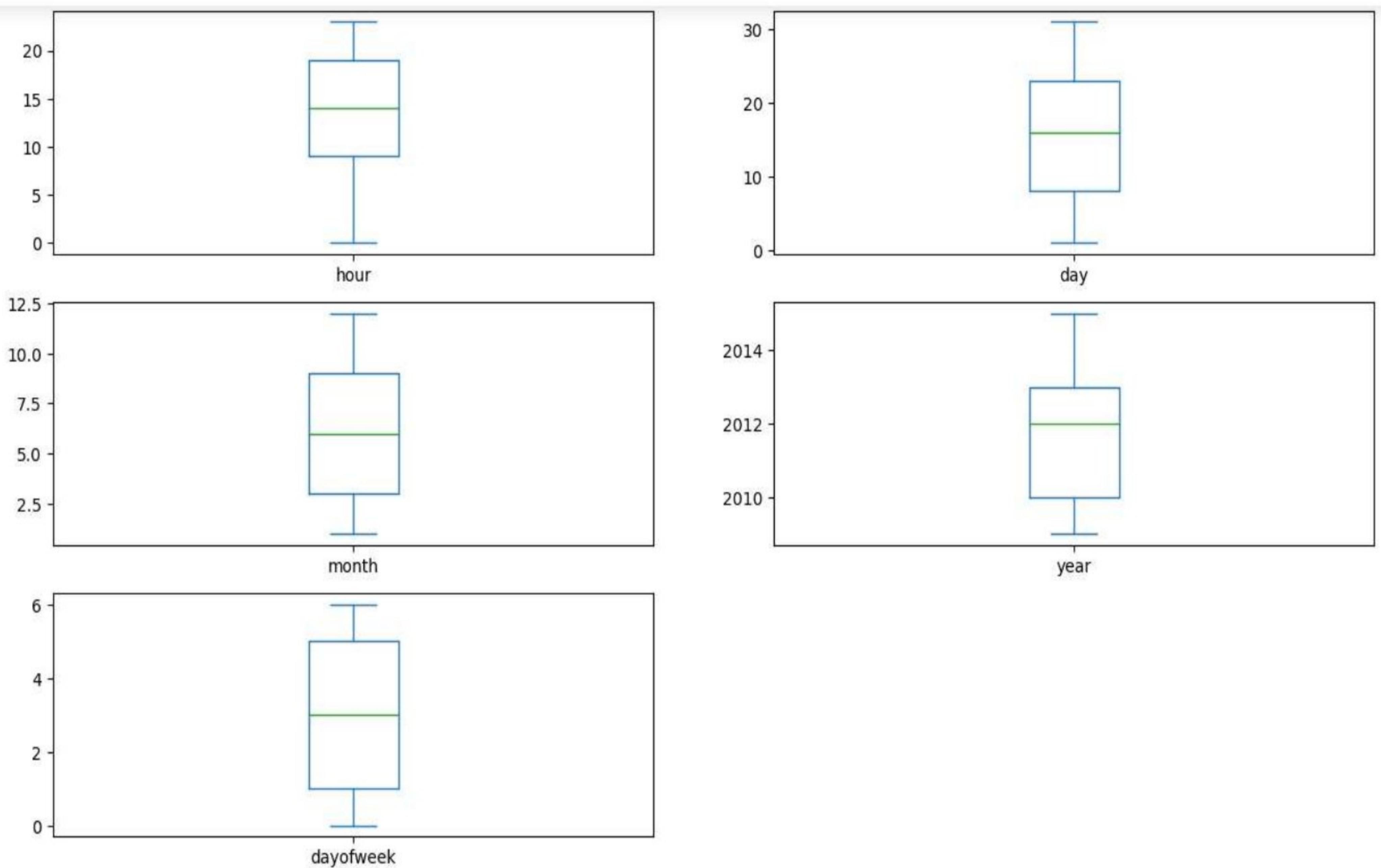
```
In [19]: df.head()
```

```
Out[19]: fare_amount  pickup_datetime  pickup_longitude  pickup_latitude  dropoff_longitude  dropoff_latitude  passenger_count  hour  day  month  year  dayofweek
0          7.5  2015-05-07 19:52:06+00:00       -73.999817        40.738354       -73.999512        40.723217           1   19    7     5  2015         3
1          7.7  2009-07-17 20:04:56+00:00       -73.994355        40.728225       -73.994710        40.750325           1   20   17     7  2009         4
2         12.9  2009-08-24 21:45:00+00:00       -74.005043        40.740770       -73.962565        40.772647           1   21   24     8  2009         0
3          5.3  2009-06-26 08:22:21+00:00       -73.976124        40.790844       -73.965316        40.803349           3   8    26     6  2009         4
4         16.0  2014-08-28 17:47:00+00:00       -73.925023        40.744085       -73.973082        40.761247           5   17   28     8  2014         3
```

```
In [20]: # drop the column 'pickup_datetime'using drop()
# 'axis= 1' drops the specified column
df= df.drop('pickup_datetime',axis= 1)
```

```
In [21]: df.head()
```

```
Out[21]: fare_amount  pickup_longitude  pickup_latitude  dropoff_longitude  dropoff_latitude  passenger_count  hour  day  month  year  dayofweek
0          7.5       -73.999817        40.738354       -73.999512        40.723217           1   19    7     5  2015         3
1          7.7       -73.994355        40.728225       -73.994710        40.750325           1   20   17     7  2009         4
2         12.9       -74.005043        40.740770       -73.962565        40.772647           1   21   24     8  2009         0
3          5.3       -73.976124        40.790844       -73.965316        40.803349           3   8    26     6  2009         4
4         16.0       -73.925023        40.744085       -73.973082        40.761247           5   17   28     8  2014         3
```



In [24]: # Using the InterQuartile Range to fill the values

```
def remove_outlier(df1,col):
    Q1= df1[col].quantile(0.25)
    Q3= df1[col].quantile(0.75)
    IQR= Q3-Q1
    lower_whisker= Q1-1.5*IQR
    upper_whisker= Q3+1.5*IQR
    df1[col]= np.clip(df1[col],lower_whisker, upper_whisker)
    return df1

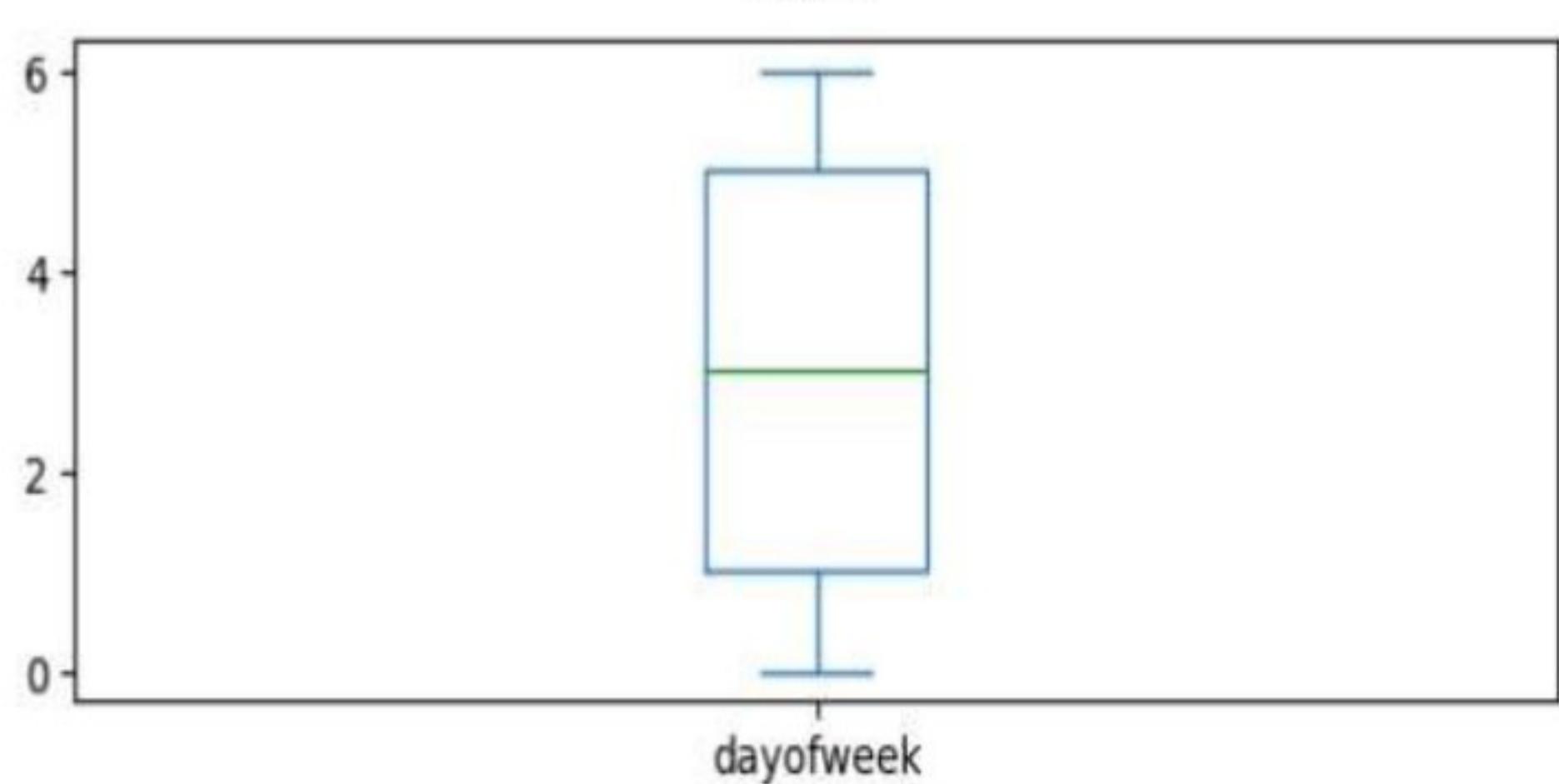
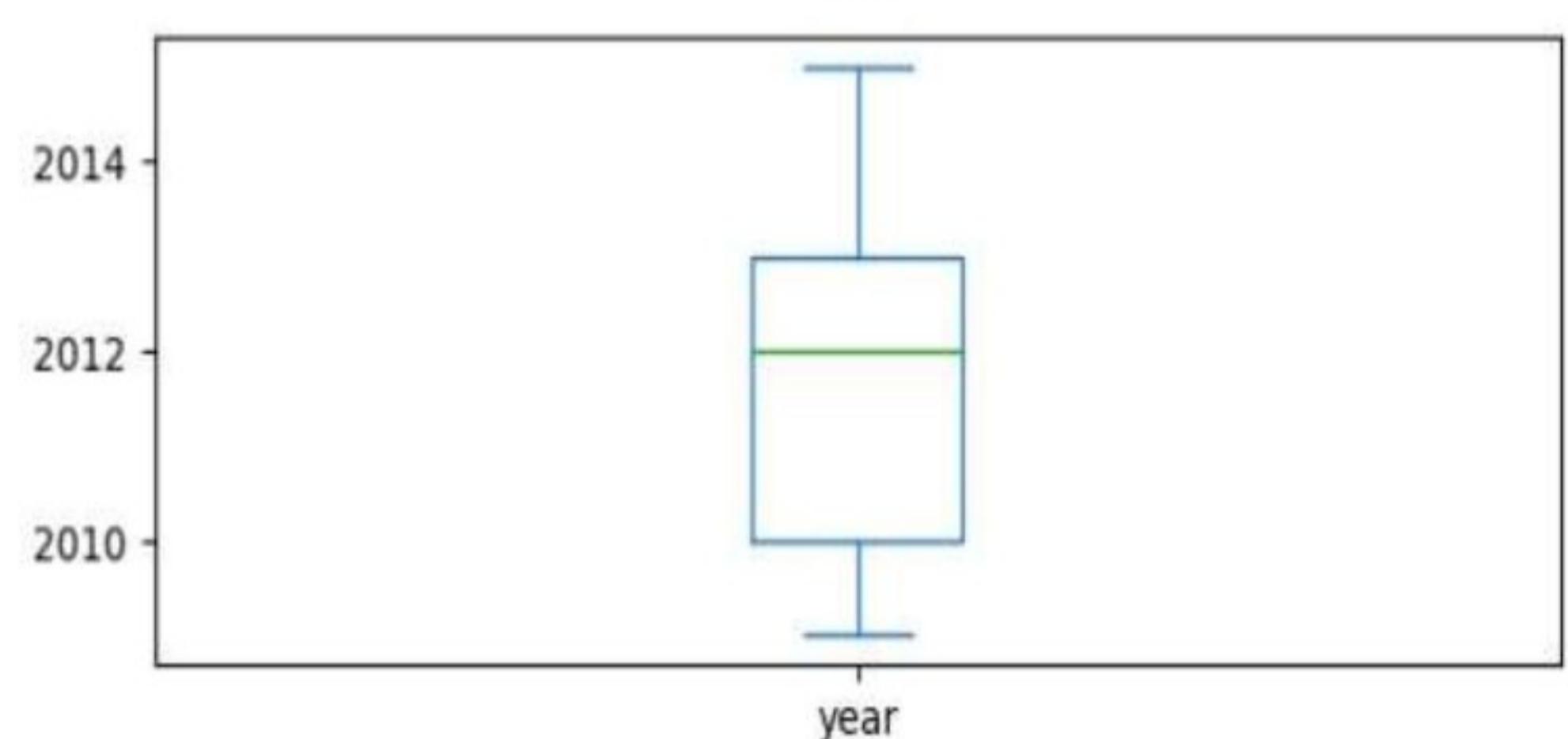
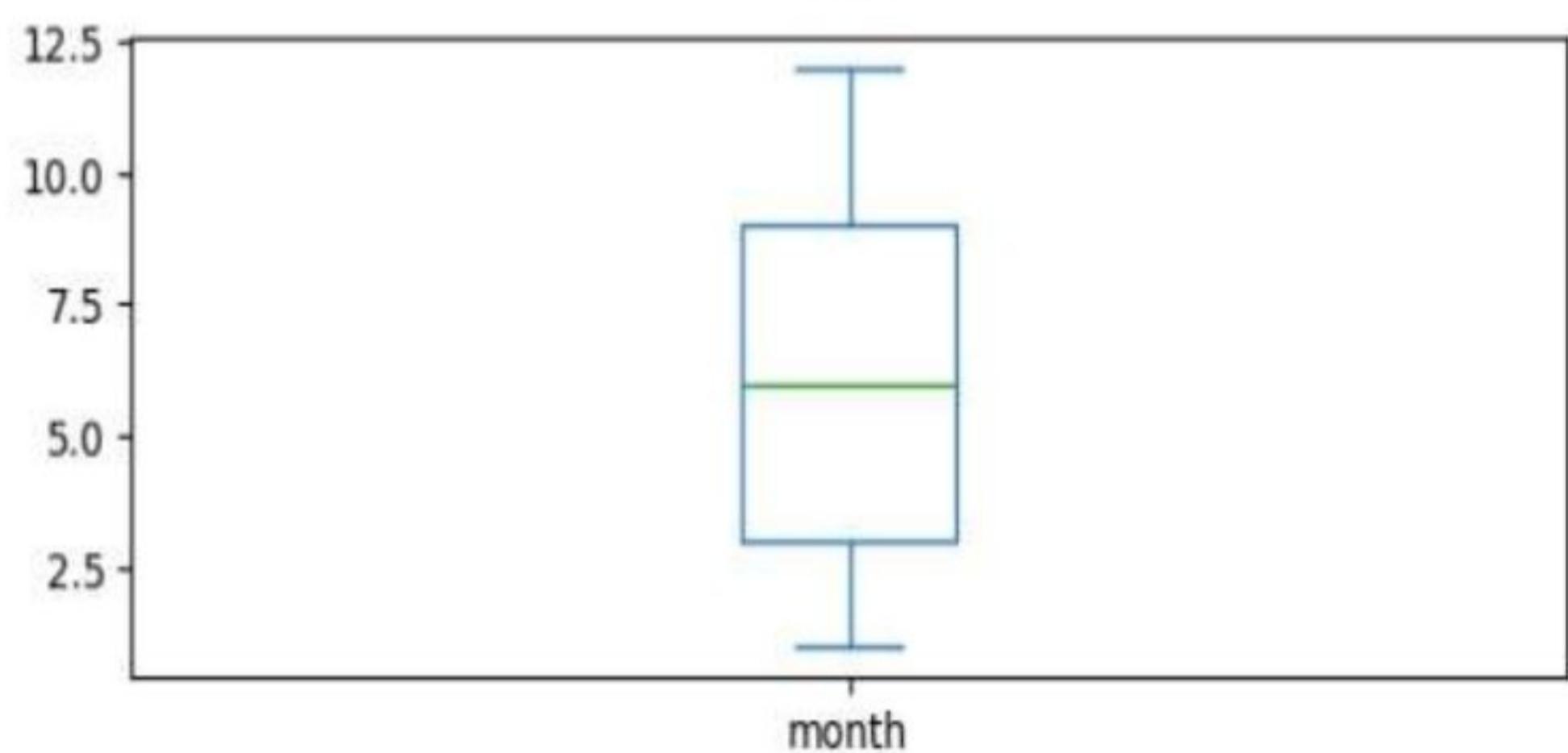
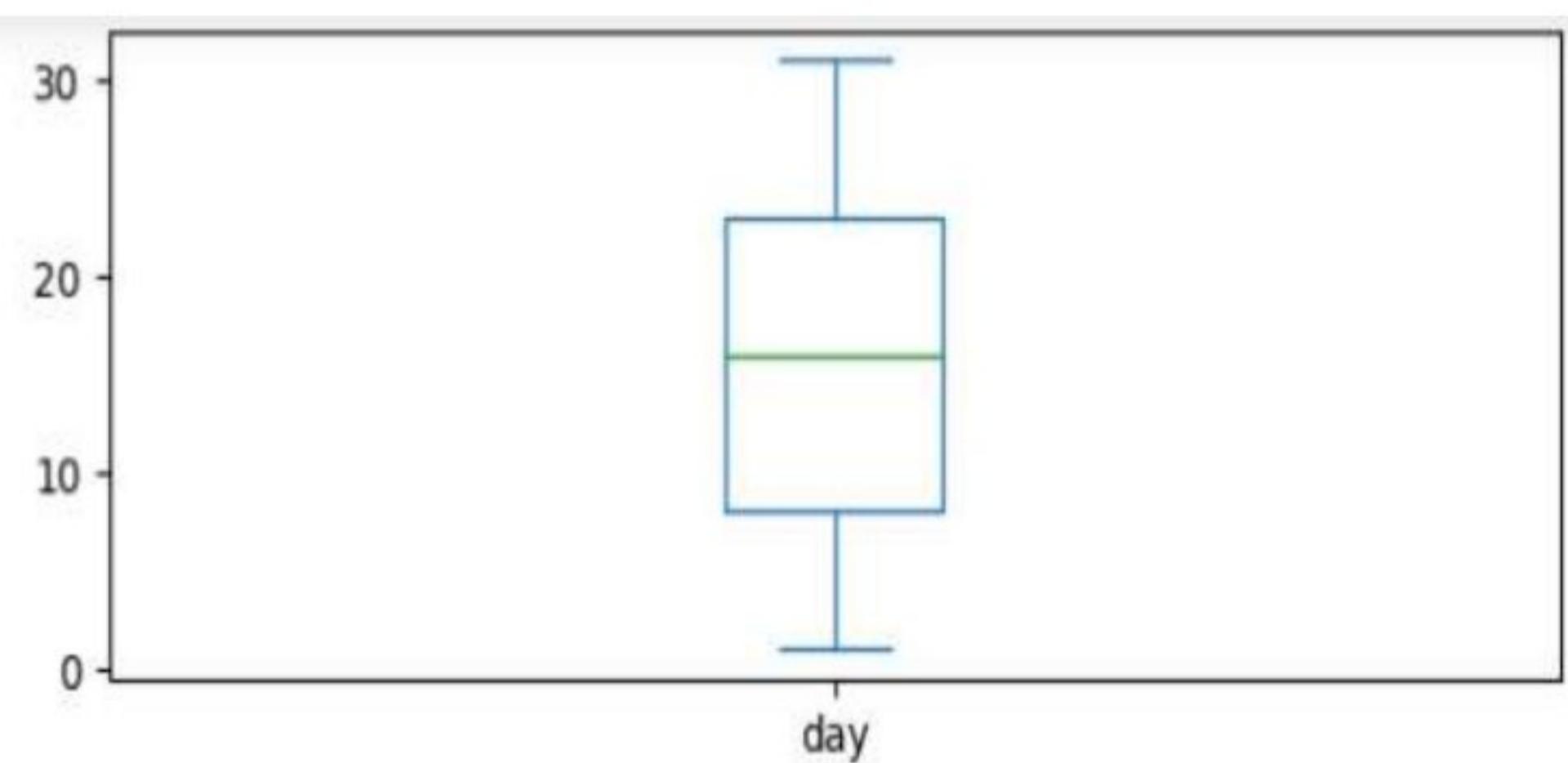
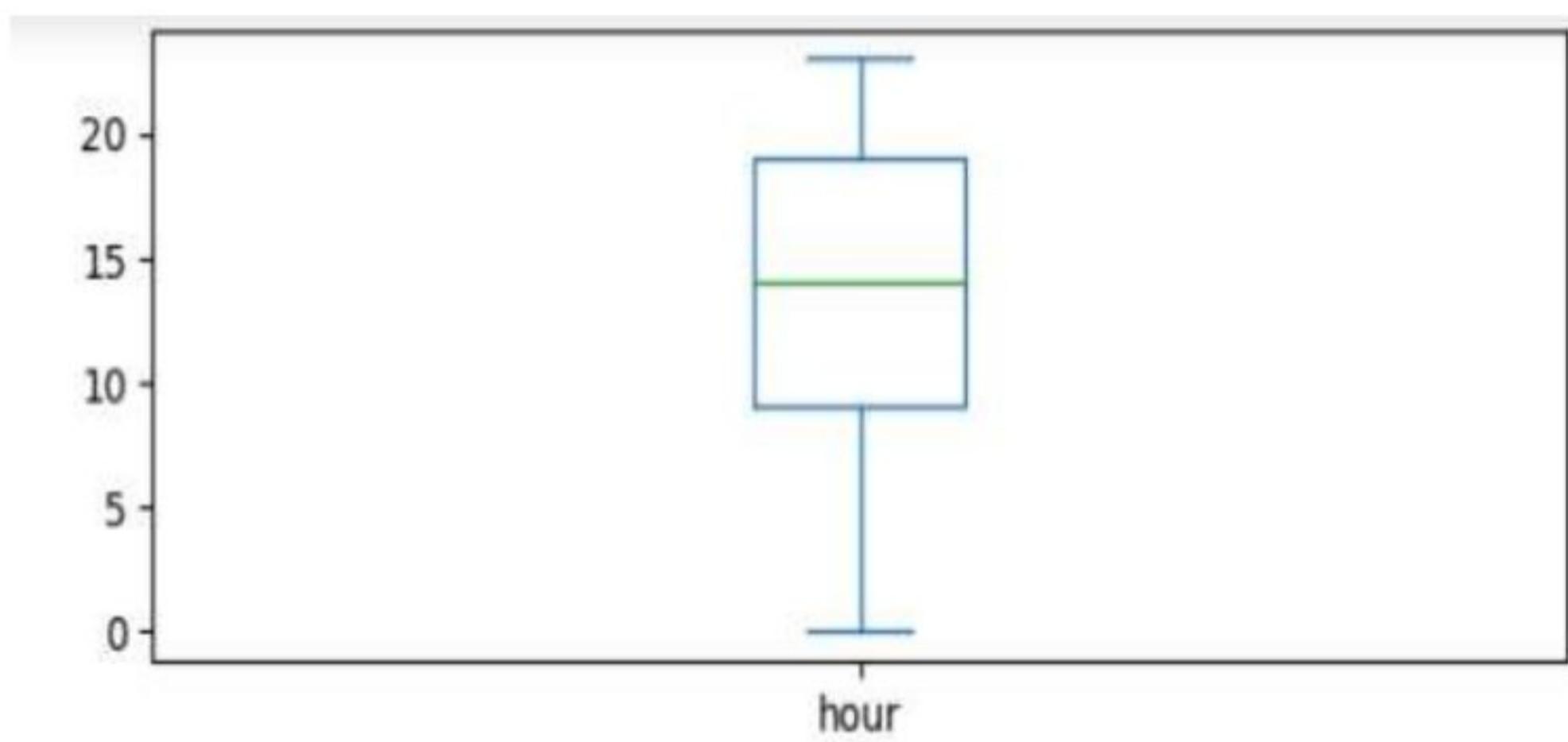
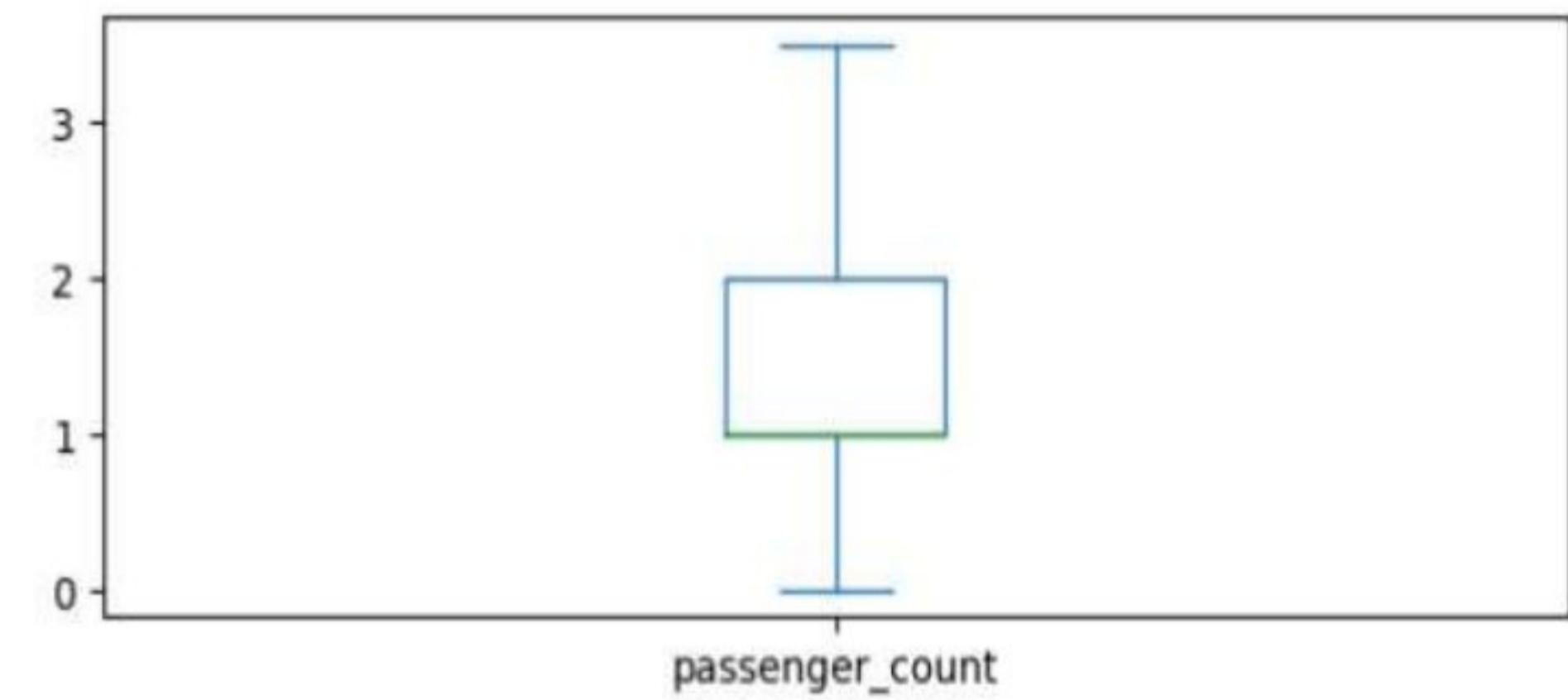
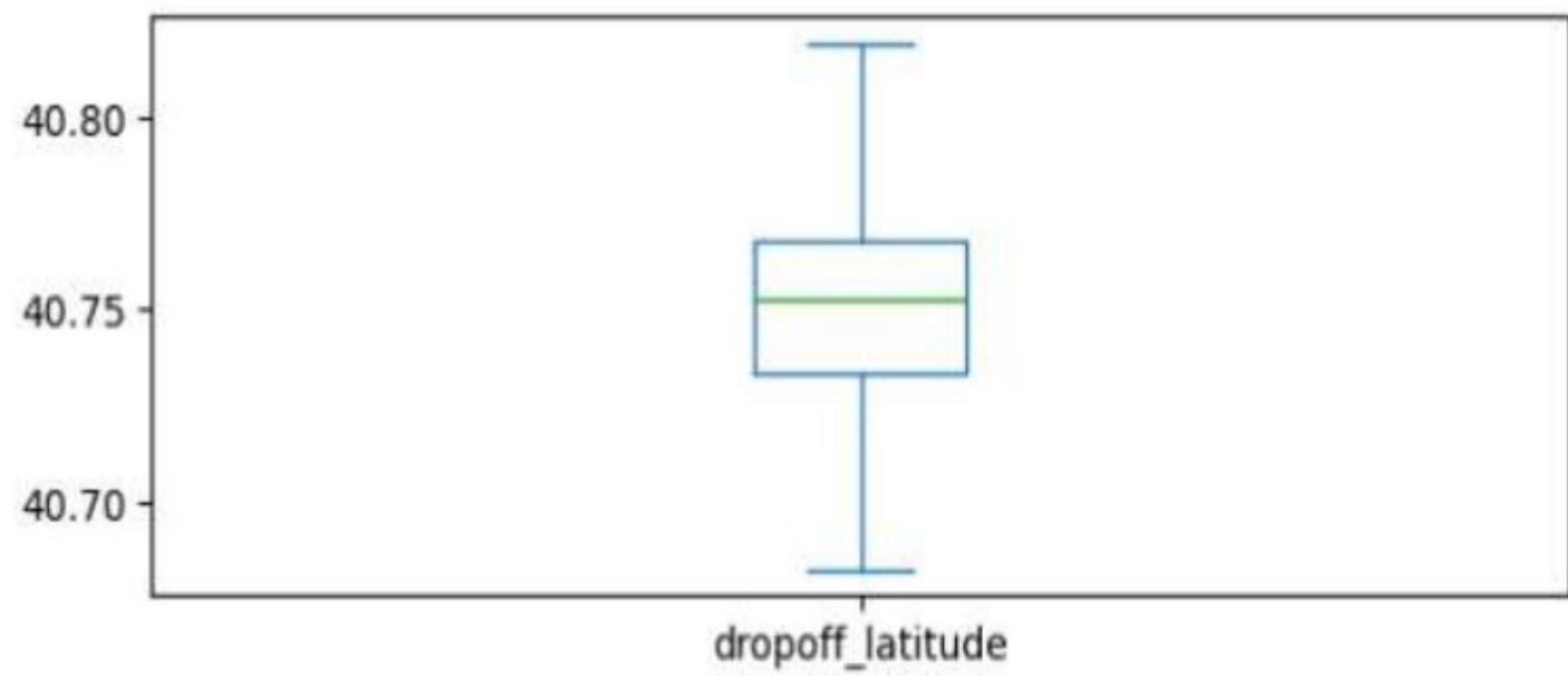
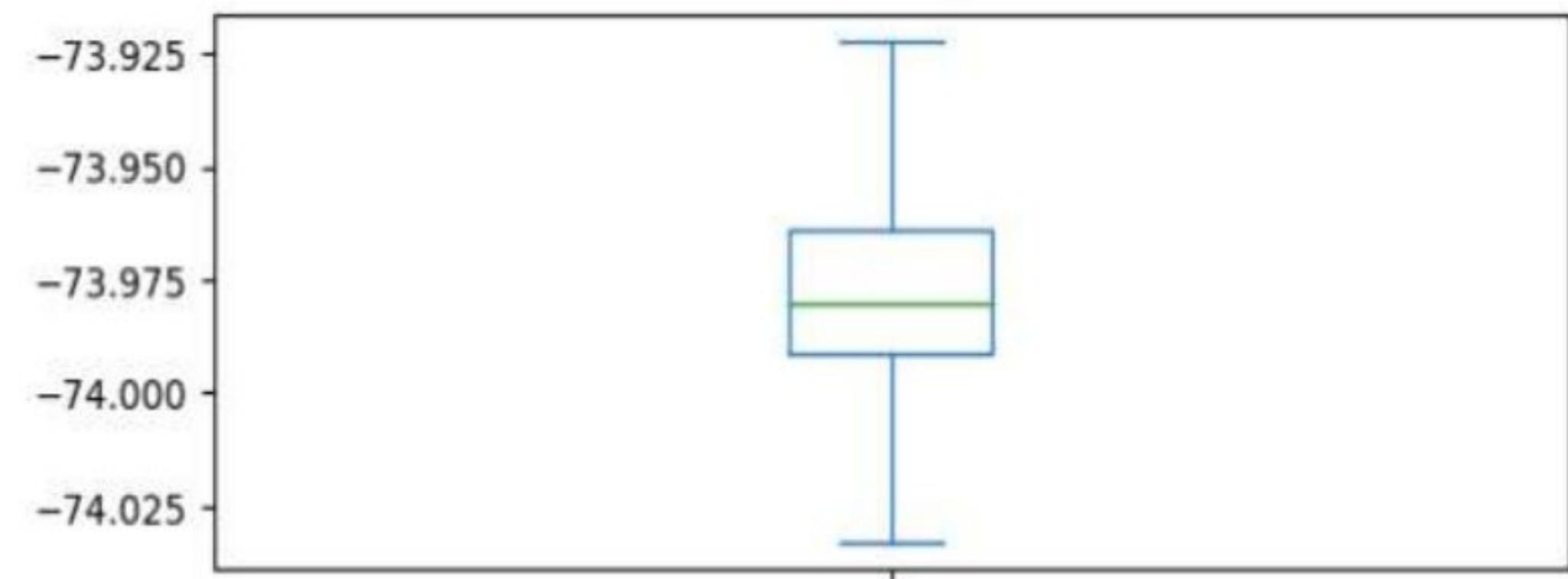
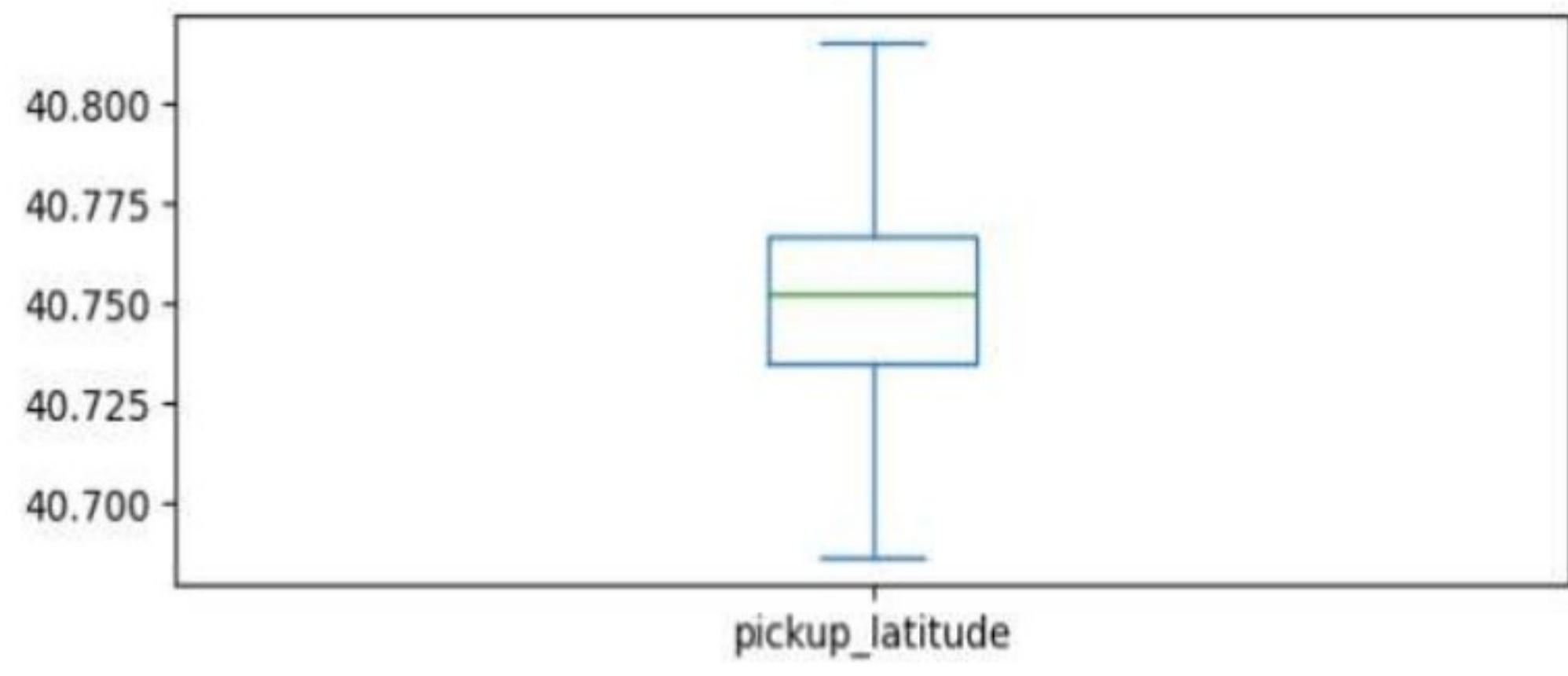
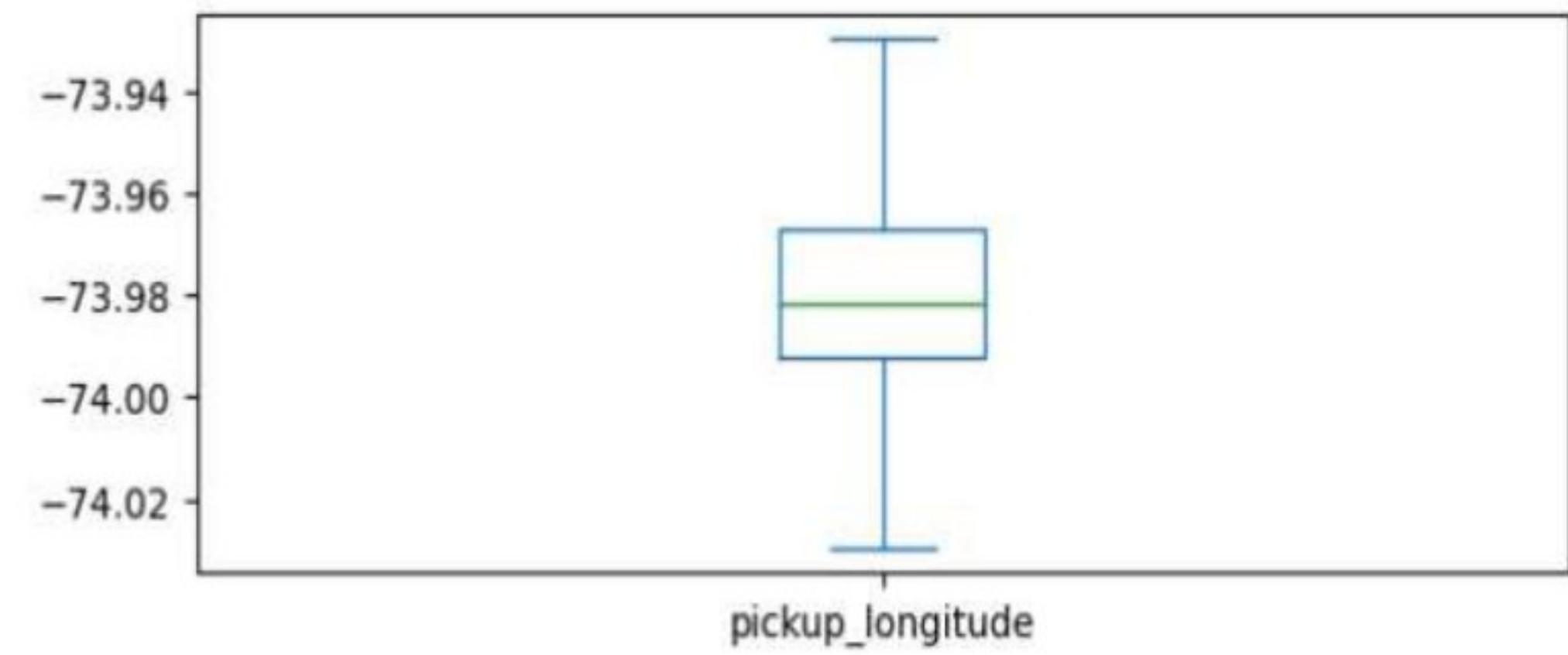
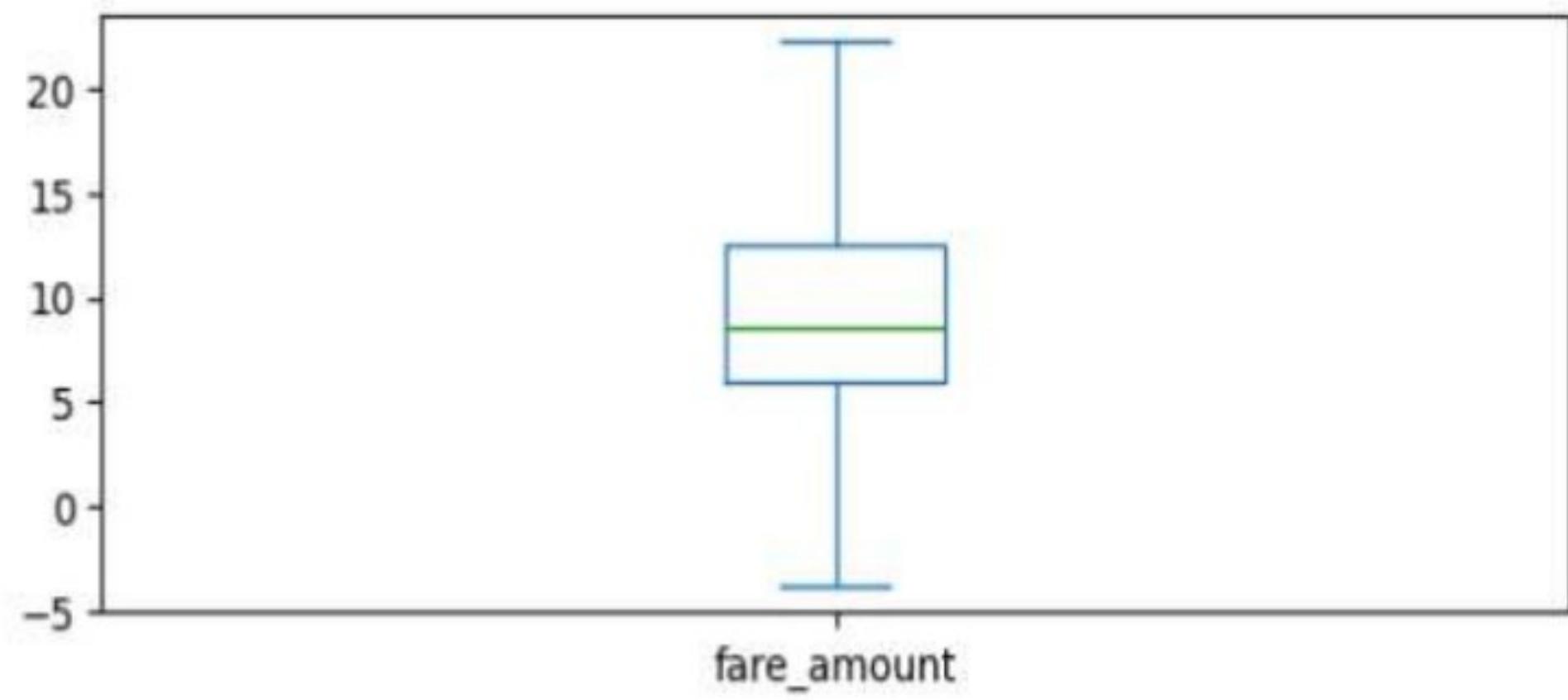
def treat_outlier_all(df1,col_list):
    for c in col_list:
        df1= remove_outlier(df1, c)
    return df1
```

In [25]: df = treat_outlier_all(df, df.iloc[:,0::])

In [26]: #Boxplot shows that dataset is free from outlier

```
df.plot(kind= "box", subplots = True, layout= (7,2), figsize= (15,20))
```

fare_amount	AxesSubplot(0.125,0.786098;0.352273x0.0939024)
pickup_longitude	AxesSubplot(0.547727,0.786098;0.352273x0.0939024)
pickup_latitude	AxesSubplot(0.125,0.673415;0.352273x0.0939024)
dropoff_longitude	AxesSubplot(0.547727,0.673415;0.352273x0.0939024)
dropoff_latitude	AxesSubplot(0.125,0.560732;0.352273x0.0939024)
passenger_count	AxesSubplot(0.547727,0.560732;0.352273x0.0939024)
hour	AxesSubplot(0.125,0.448049;0.352273x0.0939024)
day	AxesSubplot(0.547727,0.448049;0.352273x0.0939024)
month	AxesSubplot(0.125,0.335366;0.352273x0.0939024)
year	AxesSubplot(0.547727,0.335366;0.352273x0.0939024)
dayofweek	AxesSubplot(0.125,0.222683;0.352273x0.0939024)
dtype: object	



```
In [27]: pip install haversine
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: haversine in c:\users\dhanshree\appdata\roaming\python\python39\site-packages (2.8.0)
Note: you may need to restart the kernel to use updated packages.
```

```
In [30]: # Calculate the distance using Haversine to calculate between to points,
```

```
import haversine as hs
travel_dist = []
for pos in range(len(df['pickup_longitude'])):
    long1,lat1,long2,lat2 = [df['pickup_longitude'][pos],df['pickup_latitude'][pos],df['dropoff_longitude'][pos],df['dropoff_latitude'][pos]]
    loc1= (lat1, long1)
    loc2= (lat2, long2)
    c =hs.haversine(loc1,loc2)
    travel_dist.append(c)

print(travel_dist)
df['dist_travel_km']=travel_dist
df.head
```

```
IOPub data rate exceeded.
The notebook server will temporarily stop sending output
to the client in order to avoid crashing it.
To change this limit, set the config variable
`--NotebookApp.iopub_data_rate_limit`.
```

Current values:

```
NotebookApp.iopub_data_rate_limit=1000000.0 (bytes/sec)
NotebookApp.rate_limit_window=3.0 (secs)
```

```
Out[30]: <bound method NDFrame.head of      fare_amount  pickup_longitude  pickup_latitude  dropoff_longitude \
```

0	7.50	-73.999817	40.738354	-73.999512
1	7.70	-73.994355	40.728225	-73.994710
2	12.90	-74.005043	40.740770	-73.962565
3	5.30	-73.976124	40.790844	-73.965316
4	16.00	-73.929786	40.744085	-73.973082
...
199995	3.00	-73.987042	40.739367	-73.986525
199996	7.50	-73.984722	40.736837	-74.006672
199997	22.25	-73.986017	40.756487	-73.922036
199998	14.50	-73.997124	40.725452	-73.983215
199999	14.10	-73.984395	40.720077	-73.985508

0	40.723217	1.0	19	7	5	2015	3
1	40.750325	1.0	20	17	7	2009	4
2	40.772647	1.0	21	24	8	2009	0
3	40.803349	3.0	8	26	6	2009	4
4	40.761247	3.5	17	28	8	2014	3
...
199995	40.740297	1.0	10	28	10	2012	6
199996	40.739620	1.0	1	14	3	2014	4
199997	40.692588	2.0	0	29	6	2009	0
199998	40.695415	1.0	14	20	5	2015	2
199999	40.768793	1.0	4	15	5	2010	5

```
dist_travel_km
0          1.683325
1          2.457593
2          5.036384
3          1.661686
4          4.116088
...
199995    0.112210
199996    1.875053
199997    8.919323
199998    3.539720
199999    5.417791
```

[200000 rows x 12 columns]>

```
In [31]: df = df.loc[(df.dist_travel_km>=1) | (df.dist_travel_km<=130) ]
print("Remaining observation:", df.shape)
```

Remaining observation: (200000, 12)

```
In [32]: incorrect_coordinates = df.loc[(df.pickup_latitude>90)|(df.pickup_latitude<-90)|
                                         (df.dropoff_latitude>90)|(df.dropoff_latitude<-90)|
                                         (df.pickup_longitude>180)|(df.pickup_longitude<-180)|
                                         (df.dropoff_longitude>90)|(df.dropoff_longitude<-90)]
```

```
In [33]: df.drop(incorrect_coordinates, inplace = True, errors = 'ignore')
```

```
In [34]: df.head()
```

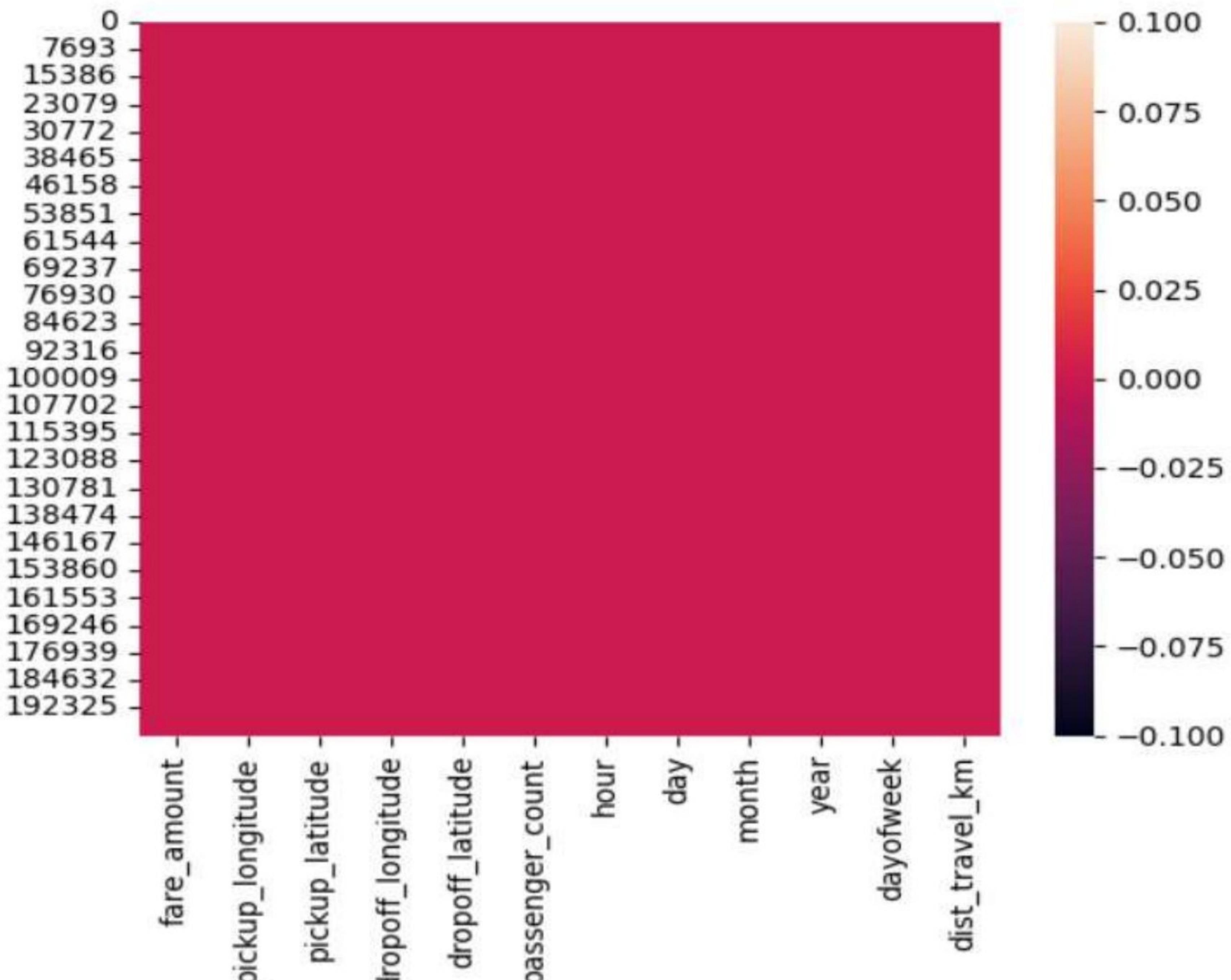
```
Out[34]: fare_amount  pickup_longitude  pickup_latitude  dropoff_longitude  dropoff_latitude  passenger_count  hour  day  month  year  dayofweek  dist_travel_km
0          7.5        -73.999817      40.738354       -73.999512      40.723217           1.0   19   7   5  2015      3     1.683325
1          7.7        -73.994355      40.728225       -73.994710      40.750325           1.0   20  17   7  2009      4     2.457593
2         12.9        -74.005043      40.740770       -73.962565      40.772647           1.0   21   24   8  2009      0     5.036384
3          5.3        -73.976124      40.790844       -73.965316      40.803349           3.0   8   26   6  2009      4     1.661686
4         16.0        -73.929786      40.744085       -73.973082      40.761247           3.5  17   28   8  2014      3     4.116088
```

```
In [35]: df.isnull().sum()
```

```
Out[35]: fare_amount      0
pickup_longitude      0
pickup_latitude       0
dropoff_longitude     0
dropoff_latitude      0
passenger_count       0
hour                   0
day                    0
month                  0
year                   0
dayofweek              0
dist_travel_km        0
dtype: int64
```

```
In [36]: sns.heatmap(df.isnull())
```

```
Out[36]: <AxesSubplot:>
```



```
In [37]: corr = df.corr()  
corr
```

```
Out[37]:
```

	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count	hour	day	month	year
fare_amount	1.000000	0.154069	-0.110842	0.218675	-0.125898	0.015778	-0.023623	0.004534	0.030817	0.14127
pickup_longitude	0.154069	1.000000	0.259497	0.425619	0.073290	-0.013213	0.011579	-0.003204	0.001169	0.01019
pickup_latitude	-0.110842	0.259497	1.000000	0.048889	0.515714	-0.012889	0.029681	-0.001553	0.001562	-0.01424
dropoff_longitude	0.218675	0.425619	0.048889	1.000000	0.245667	-0.009303	-0.046558	-0.004007	0.002391	0.01134
dropoff_latitude	-0.125898	0.073290	0.515714	0.245667	1.000000	-0.006308	0.019783	-0.003479	-0.001193	-0.00960
passenger_count	0.015778	-0.013213	-0.012889	-0.009303	-0.006308	1.000000	0.020274	0.002712	0.010351	-0.00974
hour	-0.023623	0.011579	0.029681	-0.046558	0.019783	0.020274	1.000000	0.004677	-0.003926	0.00215
day	0.004534	-0.003204	-0.001553	-0.004007	-0.003479	0.002712	0.004677	1.000000	-0.017360	-0.01217
month	0.030817	0.001169	0.001562	0.002391	-0.001193	0.010351	-0.003926	-0.017360	1.000000	-0.11585
year	0.141277	0.010198	-0.014243	0.011346	-0.009603	-0.009749	0.002156	-0.012170	-0.115859	1.00000
dayofweek	0.013652	-0.024652	-0.042310	-0.003336	-0.031919	0.048550	-0.086947	0.005617	-0.008786	0.00611
dist_travel_km	0.786385	0.048446	-0.073362	0.155191	-0.052701	0.009884	-0.035708	0.001709	0.010050	0.02229

```
In [41]: import matplotlib.pyplot as plt  
fig, axis = plt.subplots(figsize=(10,6))  
sns.heatmap(df.corr(), annot=True)
```

```
Out[41]: <AxesSubplot:>
```

```
In [69]: #Metrics evaluation for Random Forest  
R2_Random =r2_score(y_test,y_pred)  
R2_Random
```

```
Out[69]: 0.7965792384965629
```

```
In [70]: MSE_Random = mean_squared_error(y_test,y_pred)  
MSE_Random
```

```
Out[70]: 6.036187541793321
```

```
In [71]: RMSE_Random= np.sqrt(MSE_Random)  
RMSE_Random
```

```
Out[71]: 2.4568653894329096
```

Group B**Assignment No:2**

Title of Assignment: Classify the email using the binary classification method. Email Spam detection has two states:

- a) Normal State – Not Spam,
- b) Abnormal State – Spam.

1. Basic knowledge of Python
Use K-Nearest Neighbors and Support Vector Machine for classification. Analyze their performance.

Dataset Description: The csv file contains 5172 rows, each row for each email. There are 3002 columns. The first column indicates Email name. The name has been set with numbers and not recipients' name to protect privacy. The last column has the labels for prediction : 1 for spam, 0 for not spam. The remaining 3000 columns are the 3000 most common words in all the emails, after excluding the non-alphabetical characters/words. For each row, the count of each word(column) in that email(row) is stored in the respective cells. Thus, information regarding all 5172 emails are stored in a compact dataframe rather than as separate text files.

Link:<https://www.kaggle.com/datasets/balaka18/email-spam-classification-dataset-csv>

Objective of the Assignment:

Students should be able to classify email using the binary Classification and implement email spam detection technique by using K-Nearest Neighbors and Support Vector Machine algorithm.

Prerequisite:

1. Data Preprocessing:
2. Concept of K-Nearest Neighbors and Support Vector Machine for classification.

Contents of the Theory:

1. Data Preprocessing
2. Binary Classification
3. K-Nearest Neighbours
4. Support Vector Machine
5. Train, Test and Split Procedure

Data Preprocessing:

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.

Why do we need Data Preprocessing?

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

It involves below steps:

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data
- Splitting dataset into training and test set
- Feature scaling

Assignment No. 2

```
In [1]: import pandas as pd
```

```
In [5]: df = pd.read_csv("emails.csv")
```

```
In [6]: df.shape
```

```
Out[6]: (5172, 3002)
```

```
In [7]: df.head()
```

```
Out[7]:
```

	Email No.	the	to	ect	and	for	of	a	you	hou	...	connevey	jay	valued	lay	infrastructure	military	allowing	ff	dry	Prediction
0	Email 1	0	0	1	0	0	0	2	0	0	...	0	0	0	0	0	0	0	0	0	0
1	Email 2	8	13	24	6	6	2	102	1	27	...	0	0	0	0	0	0	0	1	0	0
2	Email 3	0	0	1	0	0	0	8	0	0	...	0	0	0	0	0	0	0	0	0	0
3	Email 4	0	5	22	0	5	1	51	2	10	...	0	0	0	0	0	0	0	0	0	0
4	Email 5	7	6	17	1	5	2	57	0	9	...	0	0	0	0	0	0	0	1	0	0

5 rows × 3002 columns

```
In [8]: X = df.drop(['Email No.', 'Prediction'], axis=1)  
y = df['Prediction']
```

```
In [9]: X.shape
```

```
Out[9]: (5172, 3000)
```

```
In [10]: X.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 5172 entries, 0 to 5171  
Columns: 3000 entries, the to dry  
dtypes: int64(3000)  
memory usage: 118.4 MB
```

```
In [11]: X.dtypes
```

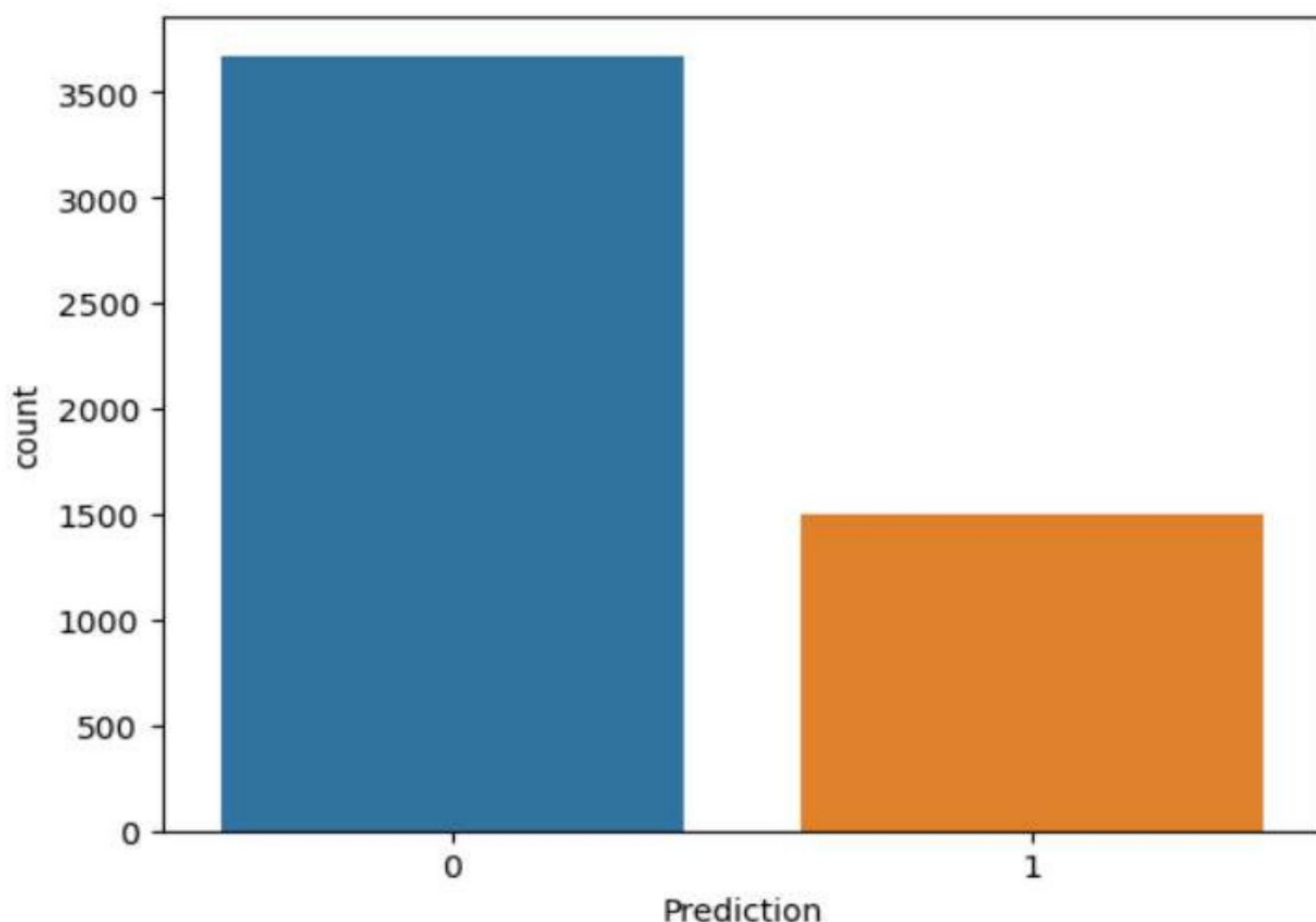
```
Out[11]: the          int64  
        to           int64  
        ect          int64  
        and          int64  
        for          int64  
        ...  
        infrastructure  int64  
        military      int64  
        allowing      int64  
        ff            int64  
        dry            int64  
Length: 3000, dtype: object
```

```
In [12]: set(X.dtypes)
```

```
Out[12]: {dtype('int64')}
```

```
In [14]: import seaborn as sns  
sns.countplot(x=y)
```

```
Out[14]: <AxesSubplot:xlabel='Prediction', ylabel='count'>
```



```
In [15]: y.value_counts()
```

```
Out[15]: 0    3672  
1    1500  
Name: Prediction, dtype: int64
```

```
In [17]: from sklearn.preprocessing import MinMaxScaler  
scaler = MinMaxScaler()  
x_scaled = scaler.fit_transform(X)
```

```
In [18]: x_scaled
```

```
Out[18]: array([[0.        , 0.        , 0.        , 0.        , ..., 0.        , 0.        ,  
   0.        ],  
   [0.03809524, 0.09848485, 0.06705539, ..., 0.        , 0.00877193,  
   0.        ],  
   [0.        , 0.        , 0.        , 0.        , ..., 0.        , 0.        ,  
   0.        ],  
   ...,  
   [0.        , 0.        , 0.        , 0.        , ..., 0.        , 0.        ,  
   0.        ],  
   [0.00952381, 0.0530303 , 0.        , 0.        , ..., 0.        , 0.00877193,  
   0.        ],  
   [0.1047619 , 0.18181818, 0.01166181, ..., 0.        , 0.        ,  
   0.        ]])
```

```
In [19]: from sklearn.model_selection import train_test_split  
X_train, X_test, y_train, y_test = train_test_split(x_scaled, y, random_state=0, test_size = 0.30)
```

```
In [20]: x_scaled.shape
```

```
Out[20]: (5172, 3000)
```

```
In [21]: X_train.shape
```

```
Out[21]: (3620, 3000)
```

```
In [22]: X_test.shape
```

```
Out[22]: (1552, 3000)
```

```
In [23]: from sklearn.neighbors import KNeighborsClassifier  
knn = KNeighborsClassifier(n_neighbors =5)  
knn.fit(X_train, y_train)
```

```
Out[23]: KNeighborsClassifier()
```

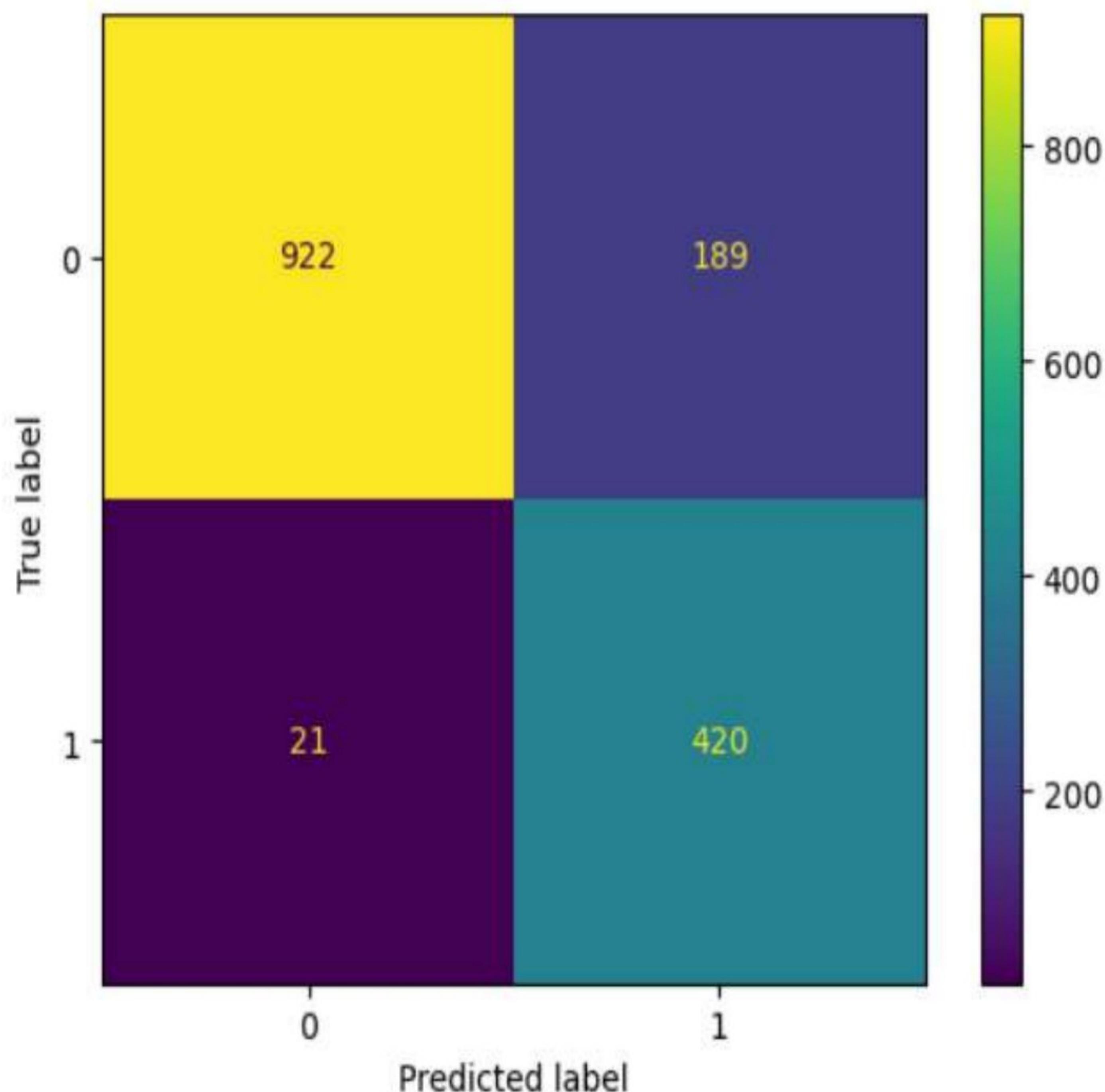
```
In [24]: y_pred = knn.predict(X_test)
```

```
C:\ProgramData\Anaconda\lib\site-packages\sklearn\neighbors\_classification.py:228: FutureWarning:  
    ions (e.g. `skew`, `kurtosis`), the default behavior of `mode` typically preserves the axis it a  
s behavior will change: the default value of `keepdims` will become False, the `axis` over which  
eliminated, and the value None will no longer be accepted. Set `keepdims` to True or False to av  
mode, _ = stats.mode(_y[neigh_ind, k], axis=1)
```

```
In [29]: from sklearn.metrics import ConfusionMatrixDisplay, accuracy_score, classification_report
```

```
In [30]: ConfusionMatrixDisplay.from_predictions(y_test, y_pred)
```

```
Out[30]: <sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x2ab10f28d60>
```



```
In [31]: y_test.value_counts()  
Out[31]: 0    1111  
          1    441  
          Name: Prediction, dtype: int64
```

```
In [32]: accuracy_score(y_test, y_pred)  
Out[32]: 0.8646907216494846
```

```
In [33]: print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.98	0.83	0.90	1111
1	0.69	0.95	0.80	441
accuracy			0.86	1552
macro avg	0.83	0.89	0.85	1552
weighted avg	0.90	0.86	0.87	1552

```
In [34]: import numpy as np  
import matplotlib.pyplot as plt
```

```
In [36]: error =[]  
for k in range(1,41):  
    knn = KNeighborsClassifier(n_neighbors =k)  
    knn.fit(X_train, y_train)  
    y_pred = knn.predict(X_test)  
    error.append(np.mean(y_pred != y_test))
```

```
In [37]: error
```

```
Out[37]: [0.10824742268041238,  
0.10502577319587629,  
0.11855670103092783,  
0.11082474226804123,  
0.13530927835051546,  
0.12886597938144329,  
0.15914948453608246,  
0.15528350515463918,  
0.17719072164948454,  
0.17010309278350516,  
0.19974226804123713,  
0.19652061855670103,  
0.21520618556701032,  
0.21198453608247422,  
0.22809278350515463,  
0.22551546391752578,  
0.23904639175257733,  
0.23646907216494845,  
0.2538659793814433,  
0.25193298969072164,  
0.2654639175257732,  
0.26417525773195877,  
0.27448453608247425,  
0.27512886597938147,  
0.28865979381443296,  
0.2867268041237113,  
0.3015463917525773,  
0.3002577319587629,  
0.3086340206185567,  
0.30605670103092786,  
0.3131443298969072,  
0.3125,  
0.31894329896907214,
```

```
In [38]: knn = KNeighborsClassifier(n_neighbors =1)
knn.fit(X_train, y_train)
```

```
Out[38]: KNeighborsClassifier(n_neighbors=1)
```

```
In [39]: y_pred = knn.predict(X_test)
```

```
C:\ProgramData\Anaconda\lib\site-packages\sklearn\neighbors
 ions (e.g. `skew`, `kurtosis`), the default behavior of `mo
  s behavior will change: the default value of `keepdims` wil
  eliminated, and the value None will no longer be accepted.
    mode, _ = stats.mode(_y[neigh_ind, k], axis=1)
```

```
In [40]: accuracy_score(y_test, y_pred)
```

```
Out[40]: 0.8917525773195877
```

```
In [42]: from sklearn.svm import SVC
svm = SVC(kernel = 'linear')
svm.fit(X_train, y_train)
```

```
Out[42]: SVC(kernel='linear')
```

```
In [43]: y_pred =svm.predict(X_test)
```

```
In [44]: accuracy_score(y_test, y_pred)
```

```
Out[44]: 0.9755154639175257
```

Group B

Assignment No:3

Title of the Assignment: Given a bank customer, build a neural network-based classifier that can determine whether they will leave or not in the next 6 months

Dataset Description: The case study is from an open-source dataset from Kaggle. The dataset contains 10,000 sample points with 14 distinct features such as CustomerId, CreditScore, Geography, Gender, Age, Tenure, Balance, etc.

Link for Dataset: <https://www.kaggle.com/barelydedicated/bank-customer-churn-modeling>

Perform the following steps:

1. Read the dataset.
2. Distinguish the feature and target set and divide the data set into training and test sets.
3. Normalize the train and test data.
4. Initialize and build the model. Identify the points of improvement and implement the same.
5. Print the accuracy score and confusion matrix (5 points).

Objective of the Assignment:

Students should be able to distinguish the feature and target set and divide the data set into training and test sets and normalize them and students should build the model on the basis of that.

Prerequisite:

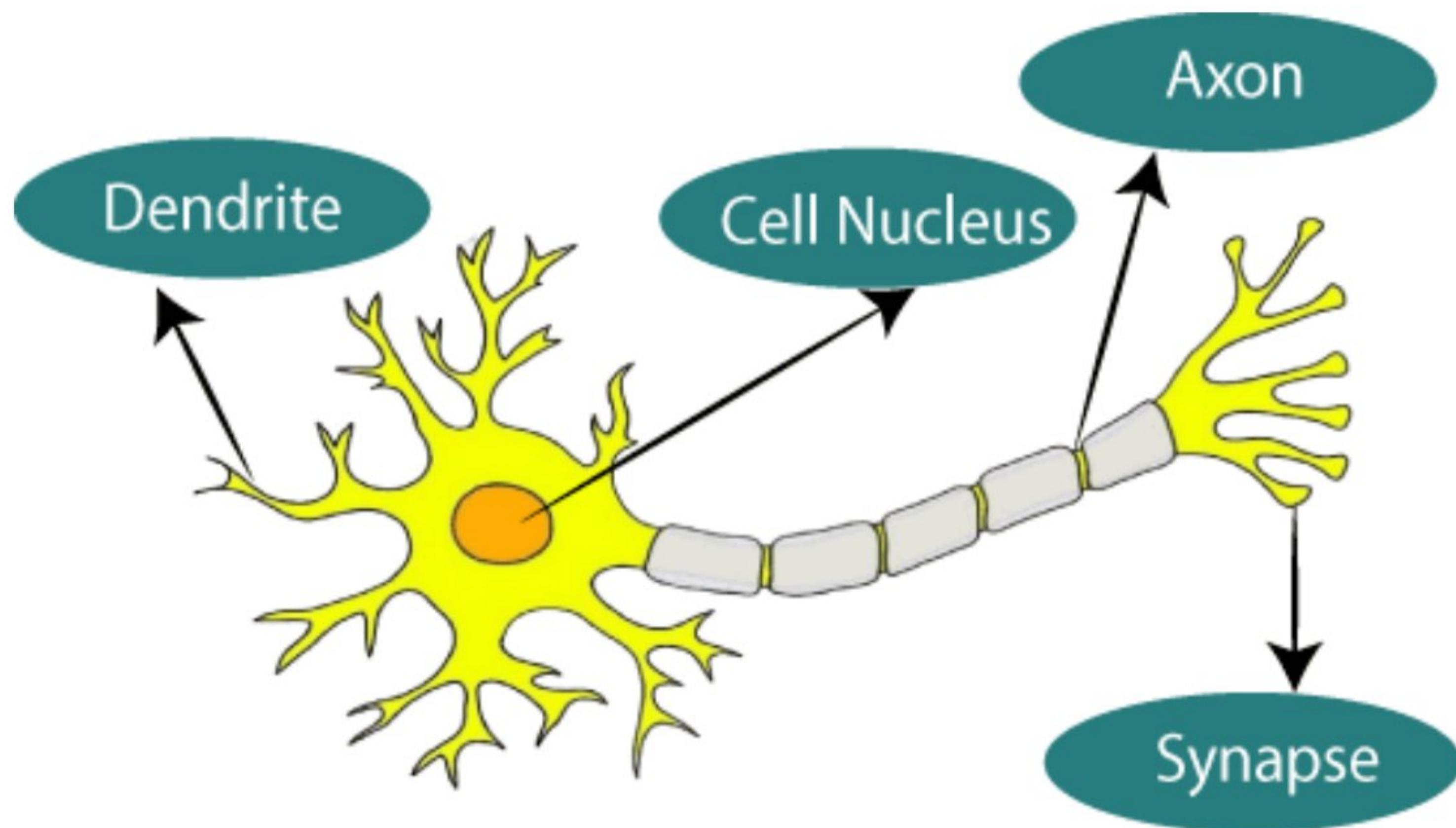
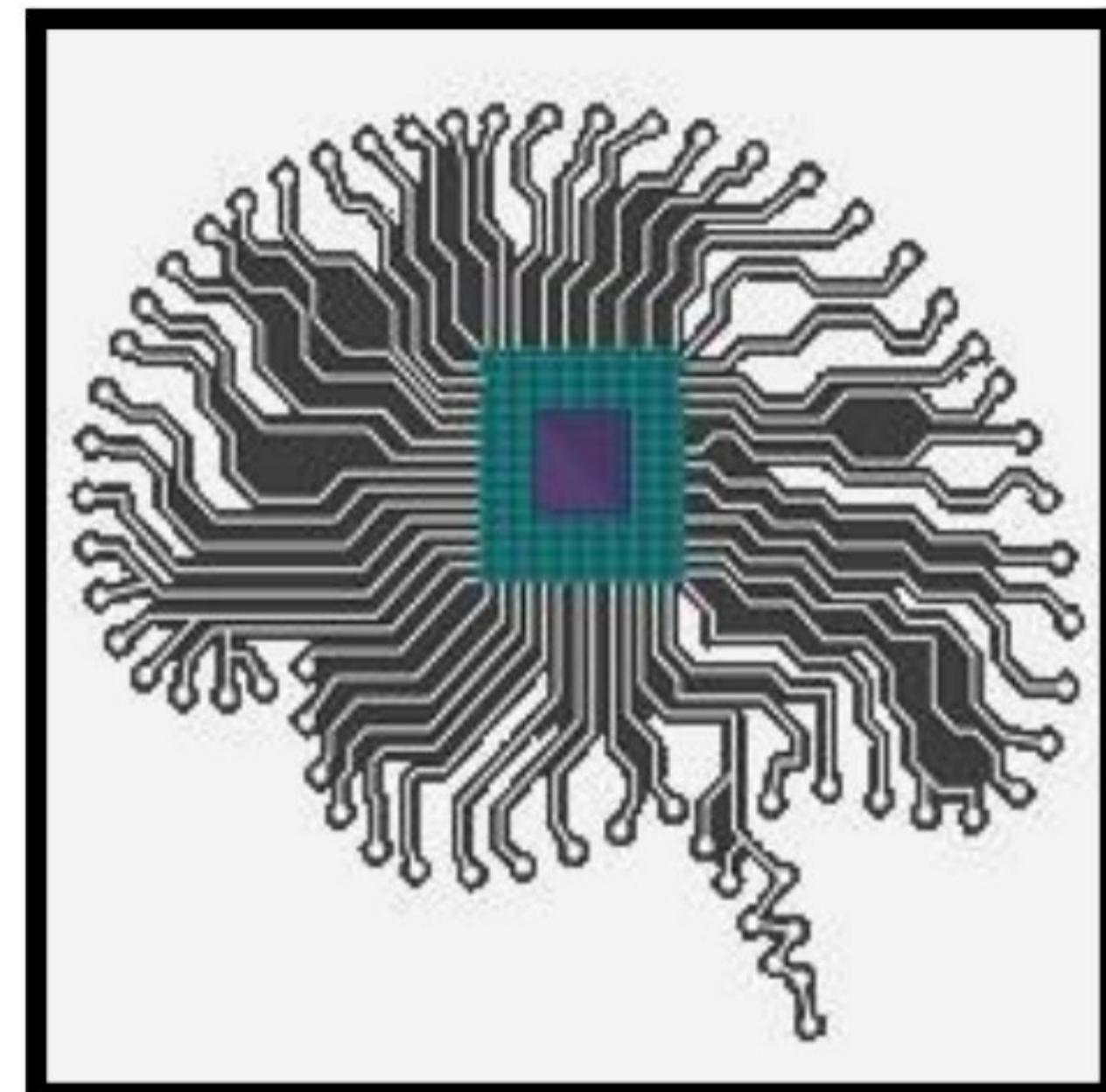
1. Basic knowledge of Python
2. Concept of Confusion Matrix

Contents of the Theory:

1. Artificial Neural Network
2. Keras
3. tensorflow
4. Normalization
5. Confusion Matrix

Artificial Neural Network:

The term "Artificial Neural Network" is derived from Biological neural networks that develop the structure of a human brain. Similar to the human brain that has neurons interconnected to one another, artificial neural networks also have neurons that are interconnected to one another in various layers of the networks. These neurons are known as nodes.



The given figure illustrates the typical diagram of Biological Neural Network.

The typical Artificial Neural Network looks something like the given figure.

Dendrites from Biological Neural Network represent inputs in Artificial Neural Networks, cell nucleus represents Nodes, synapse represents Weights, and Axon represents Output.

Relationship between Biological neural network and artificial neural network:

An **Artificial Neural Network** in held of **Artificial intelligence** where it attempts to mimic the network of neurons makes up a human brain so that computers will have an option to understand things and make decisions in a human-like manner. The artificial neural network is designed by programming computers to behave simply like interconnected brain cells.

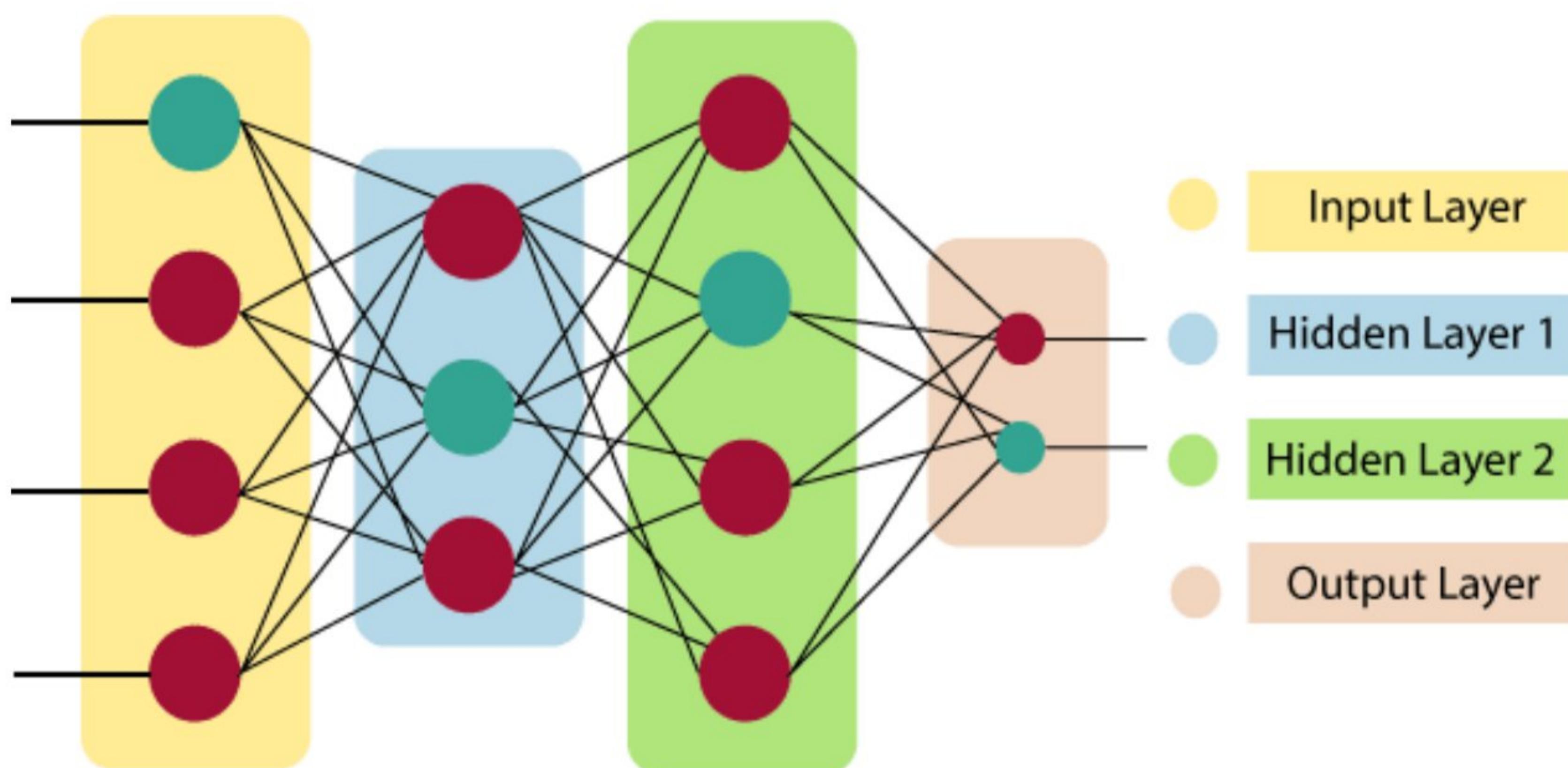
There are around 1000 billion neurons in the human brain. Each neuron has an association point somewhere in the range of 1,000 and 100,000. In the human brain, data is stored in such a manner as to be distributed, and we can extract more than one piece of this data when necessary, from our memory parallelly. We can say that the human brain is made up of incredibly amazing parallel processors.

We can understand the artificial neural network with an example, consider an example of a digital logic gate that takes an input and gives an output. "OR" gate, which takes two inputs. If one or both the inputs are "On," then we get "On" in output. If both the inputs are "Off," then we get "Off" in output. Here the output depends upon input. Our brain does not perform the same task. The outputs to inputs relationship keep changing because of the neurons in our brain, which are "learning."

The architecture of an artificial neural network:

To understand the concept of the architecture of an artificial neural network, we have to understand what a neural network consists of. In order to design a neural network that consists of a large number of artificial neurons, which are termed units arranged in a sequence of layers. Let's look at various types of layers available in an artificial neural network.

Artificial Neural Network primarily consists of three layers:



Input Layer:

As the name suggests, it accepts inputs in several different formats provided by the programmer.

Hidden Layer: The hidden layer presents in-between input and output layers. It performs all the calculations to find hidden features and patterns.

Output Layer:

The input goes through a series of transformations using the hidden layer, which finally results in output that is conveyed using this layer.

The artificial neural network takes input and computes the weighted sum of the inputs and includes a bias. This computation is represented in the form of a transfer function.

$$\sum_{i=1}^n w_i * x_i + b$$

It determines weighted total is passed as an input to an activation function to produce the output. Activation functions choose whether a node should fire or not. Only those who are fired make it to the output layer. There are distinctive activation functions available that can be applied upon the sort of task we are performing.

Keras:

Keras is an open-source high-level Neural Network library, which is written in Python and is capable enough to run on Theano, TensorFlow, or CNTK. It was developed by one of the Google engineers, Francois Chollet. It is made user-friendly, extensible, and modular for facilitating faster experimentation with deep neural networks. It not only supports Convolutional Networks and Recurrent Networks individually but also their combination.

It cannot handle low-level computations, so it makes use of the **Backend** library to resolve it. The backend library acts as a high-level API wrapper for the low-level API, which lets it run on TensorFlow, CNTK, or Theano.

Initially, it had over 4800 contributors during its launch, which now has gone up to 250,000 developers. It has a 2X growth ever since every year it has grown. Big companies like Microsoft, Google, NVIDIA, and Amazon have actively contributed to the development of Keras. It has an amazing industry interaction, and it is used in the development of popular likes Netix, Uber, Google, Expedia, etc.

Tensorflow:

TensorFlow is a Google product, which is one of the most famous deep learning tools widely used in the research area of machine learning and deep neural network. It came into the market on 9th November 2015 under the Apache License 2.0. It is built in such a way that it can easily run on multiple CPUs and GPUs as well as on mobile operating systems. It consists of various wrappers in distinct languages such as Java, C++, or Python.

Normalization:

Normalization is a scaling technique in Machine Learning applied during data preparation to change the values of numeric columns in the dataset to use a common scale. It is not necessary for all datasets in a model. It is required only when features of machine learning models have different ranges.

Mathematically, we can calculate normalization with the below formula:

$$X_n = (X - X_{\min}) / (X_{\max} - X_{\min})$$

Where,

- X_n = Value of Normalization
- X_{\max} = Maximum value of a feature
- X_{\min} = Minimum value of a feature

Example: Let's assume we have a model dataset having maximum and minimum values of feature as mentioned above. To normalize the machine learning model, values are shifted and rescaled so their range can vary between 0 and 1. This technique is also known as Min-Max scaling. In this scaling technique, we will change the feature values as follows:

Case1-If the value of X is minimum, the value of Numerator will be 0; hence Normalization will also be 0.

$$X_n = (X - X_{\min}) / (X_{\max} - X_{\min}) \text{ ----- formula}$$

Put $X = X_{\min}$ in above formula, we get;

$$X_n = X_{\min} - X_{\min} / (X_{\max} - X_{\min}) X_n = 0$$

Case2-If the value of X is maximum, then the value of the numerator is equal to the denominator; hence Normalization will be 1.

$$X_n = (X - X_{\min}) / (X_{\max} - X_{\min}) \text{ Put } X = X_{\max} \text{ in above formula, we get;}$$

$$X_n = X_{\max} - X_{\min} / (X_{\max} - X_{\min}) X_n = 1$$

Case3-On the other hand, if the value of X is neither maximum nor minimum, then values of normalization will also be between 0 and 1.

Hence, Normalization can be dened as a scaling method where values are shifted and rescaled to maintain their ranges between 0 and 1, or in other words; it can be referred to as Min-Max scaling technique.

Normalization techniques in Machine Learning

Although there are so many feature normalization techniques in Machine Learning, few of them are most frequently used. These are as follows:

- **Min-Max Scaling:** This technique is also referred to as scaling. As we have already discussed above, the Min-Max scaling method helps the dataset to shift and rescale the values of their attributes, so they end up ranging between 0 and 1.

- **Standardization scaling:**

Standardization scaling is also known as **Z-score** normalization, in which values are centered around the mean with a unit standard deviation, which means the attribute becomes zero and the resultant distribution has a unit standard deviation. Mathematically, we can calculate the standardization by subtracting the feature value from the mean and dividing it by standard deviation.

Hence, standardization can be expressed as follows:

$$X' = \frac{X - \mu}{\sigma}$$

Here, μ represents the mean of feature value, and σ represents the standard deviation of feature values.

However, unlike Min-Max scaling technique, feature values are not restricted to a specific range in the standardization technique.

This technique is helpful for various machine learning algorithms that use distance measures such as **KNN, K-means clustering, and Principal component analysis**, etc. Further, it is also important that the model is built on assumptions and data is normally distributed.

When to use Normalization or Standardization?

Which is suitable for our machine learning model, Normalization or Standardization? This is probably a big confusion among all data scientists as well as machine learning engineers. Although both terms have the almost same meaning choice of using normalization or standardization will depend on your problem and the algorithm you are using in models.

1. Normalization is a transformation technique that helps to improve the performance as well as the accuracy of your model better. Normalization of a machine learning model is useful when you don't know feature distribution exactly. In other words, the feature distribution of data does not follow a **Gaussian**(bell curve) distribution. Normalization must have an abounding range, so if you have outliers in data, they will be affected by Normalization.

Further, it is also useful for data having variable scaling techniques such as **KNN, artificial neural networks**. Hence, you can't use assumptions for the distribution of data.

2. Standardization in the machine learning model is useful when you are exactly aware of the feature distribution of data or, in other words, your data follows a Gaussian distribution. However, this does not have to be necessarily true. Unlike Normalization, Standardization does not necessarily have a bounding range, so if you have outliers in your data, they will not be affected by Standardization.

Further, it is also useful when data has variable dimensions and techniques such as **linear regression, logistic regression, and linear discriminant analysis**.

Example: Let's understand an experiment where we have a dataset having two attributes, i.e., age and salary. Where the age ranges from 0 to 80 years old, and the income varies from 0 to 75,000 dollars or more. Income is assumed to be 1,000 times that of age. As a result, the ranges of these two attributes are much different from one another.

Because of its bigger value, the attributed income will organically influence the conclusion more when we undertake further analysis, such as multivariate linear regression. However, this does not necessarily imply that it is a better predictor. As a result, we normalize the data so that all of the variables are in the same range.

Further, it is also helpful for the prediction of credit risk scores where normalization is applied to all numeric data except the class column. It uses the **tanh transformation** technique, which converts all numeric features into values of range between 0 to 1.

Confusion Matrix:

The confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data. It can only be determined if the true values for test data are known. The matrix itself can be easily understood, but the related terminologies may be confusing. Since it shows the errors in the model performance in the form of a matrix, hence also known as an **error matrix**. Some features of Confusion matrix are given below:

- For the 2 prediction classes of classifiers , the matrix is of 2*2 table, for 3 classes, it is 3*3 table, and so on.

- The matrix is divided into two dimensions, that are **predicted values** and **actual values** along with the total number of predictions.
- Predicted values are those values, which are predicted by the model, and actual values are the true values for the given observations.
- It looks like the below table:

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

The above table has the following cases:

- True Negative:** Model has given prediction No, and the real or actual value was also No.
- True Positive:** The model has predicted yes, and the actual value was also true.
- False Negative:** The model has predicted no, but the actual value was Yes, it is also called as **Type-II error**.
- False Positive:** The model has predicted Yes, but the actual value was No. It is also called a **Type-I error**.

Need for Confusion Matrix in Machine learning

- It evaluates the performance of the classification models, when they make predictions on test data, and tells how good our classification model is.
- It not only tells the error made by the classifiers but also the type of errors such as it is either type-I or type-II error.
- With the help of the confusion matrix, we can calculate the different parameters for the model, such as accuracy, precision, etc.

Example: We can understand the confusion matrix using an example.

Suppose we are trying to create a model that can predict the result for the disease that is either a person has that disease or not. So, the confusion matrix for this is given as:

n = 100	Actual: No	Actual: Yes	
Predicted: No	TN: 65	FP: 3	68
Predicted: Yes	FN: 8	TP: 24	32
	73	27	

From the above example, we can conclude that

The table is given for the two-class classifier, which has two predictions "Yes" and "NO." Here, Yes denotes that patient has the disease, and No denotes that patient does not have that disease.

- The classifier has made a total of **100 predictions**. Out of 100 predictions, **89 are true predictions**, and **11 are incorrect predictions**.
- The model has given prediction "yes" for 32 times, and "No" for 68 times. Whereas the actual "Yes" was 27, and actual "No" was 73 times.

Calculations using Confusion Matrix:

We can perform various calculations for the model, such as the model's accuracy, using this matrix. These calculations are given below:

- **Classification Accuracy:** It is one of the important parameters to determine the accuracy of the classification problems. It denotes how often the model predicts the correct output. It can be calculated as the ratio of the number of correct predictions made by the classifier to all number of predictions made by the classifiers. The formula is given below:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

- **Misclassification rate:** It is also termed as Error rate, and it denotes how often the model gives the wrong predictions. The value of error rate can be calculated as the number of incorrect predictions to all number of the predictions made by the classifier. The formula is given below:

$$\text{Error rate} = \frac{FP + FN}{TP + FP + FN + TN}$$

- **Precision:** It can be denoted as the number of correct outputs provided by the model

or out of all positive classes that have predicted correctly by the model, how many of them were actually true. It can be calculated using the below formula

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

Recall: It is defined as the out of total positive classes, how our model predicted correctly. The recall must be as high as possible.

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

F-measure: If two models have low precision and high recall or vice versa, it is difficult to compare these models. So, for this purpose, we can use F-score. This score helps us to evaluate the recall and precision at the same time. The F-score is maximum if the recall is equal to the precision. It can be calculated using the below formula:

⊗

$$\text{F-measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

Other important terms used in Confusion Matrix:

- **Null Error rate:** It denotes how often our model would be incorrect if it always predicted the

$$\text{F-measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

majority class. As per the accuracy paradox, it is said that "*the best classifier has a higher error rate than the null error rate.*"

- **ROC Curve:** The ROC is a graph displaying a classifier's performance for all possible thresholds. The graph is plotted between the true positive rate (on the Y-axis) and the false Positive rate (on the x-axis).

Conclusion:

In this way we build a neural network-based classifier that can determine whether they will leave or not in the next 6 months

Assignment Questions:

- 1) What is Normalization?
- 2) What is Standardization?
- 3) Explain Confusion Matrix ?
- 4) Define the following: Classification Accuracy, Misclassification Rate, Precision.
- 5) One Example of Confusion Matrix?

Assignment No. 3

```
In [3]: #Importing the Libraries
```

```
import pandas as pd  
import numpy as np  
import seaborn as sns  
import matplotlib.pyplot as plt
```

```
In [4]: df = pd.read_csv("Churn_Modelling.csv")
```

```
In [5]: #Preprocesssing  
df.isnull()
```

Out[5]:

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary
0	False	False	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False	False	False
...
9995	False	False	False	False	False	False	False	False	False	False	False	False	False
9996	False	False	False	False	False	False	False	False	False	False	False	False	False
9997	False	False	False	False	False	False	False	False	False	False	False	False	False
9998	False	False	False	False	False	False	False	False	False	False	False	False	False
9999	False	False	False	False	False	False	False	False	False	False	False	False	False

10000 rows x 14 columns

```
In [6]: df.isnull().sum()
```

```
Out[6]: RowNumber      0  
CustomerId      0  
Surname        0  
CreditScore     0  
Geography       0  
Gender          0  
Age             0  
Tenure          0  
Balance         0  
NumOfProducts    0  
HasCrCard       0  
IsActiveMember   0  
EstimatedSalary  0  
Exited          0  
dtype: int64
```

```
In [7]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   RowNumber        10000 non-null   int64  
 1   CustomerId      10000 non-null   int64  
 2   Surname          10000 non-null   object  
 3   CreditScore      10000 non-null   int64  
 4   Geography         Geography       object  
 5   Gender            10000 non-null   object  
 6   Age               10000 non-null   int64  
 7   Tenure            10000 non-null   int64  
 8   Balance           10000 non-null   float64 
 9   NumOfProducts    10000 non-null   int64  
 10  HasCrCard        10000 non-null   int64  
 11  IsActiveMember   10000 non-null   int64  
 12  EstimatedSalary  10000 non-null   float64 
 13  Exited           10000 non-null   int64  
dtypes: float64(2), int64(9), object(3)
memory usage: 1.1+ MB
```

```
In [8]: df.dtypes
```

```
Out[8]: RowNumber        int64
CustomerId        int64
Surname          object
CreditScore       int64
Geography         object
Gender            object
Age               int64
Tenure            int64
Balance           float64
NumOfProducts     int64
HasCrCard         int64
IsActiveMember   int64
EstimatedSalary  float64
Exited            int64
dtype: object
```

```
In [9]: df.columns
```

```
Out[9]: Index(['RowNumber', 'CustomerId', 'Surname', 'CreditScore', 'Geography',
   'Gender', 'Age', 'Tenure', 'Balance', 'NumOfProducts', 'HasCrCard',
   'IsActiveMember', 'EstimatedSalary', 'Exited'],
  dtype='object')
```

```
In [10]: #Dropping the unnecessary columns
```

```
df= df.drop(['RowNumber', 'CustomerId', 'Surname'], axis =1)
```

```
In [11]: df.head()
```

Out[11]:

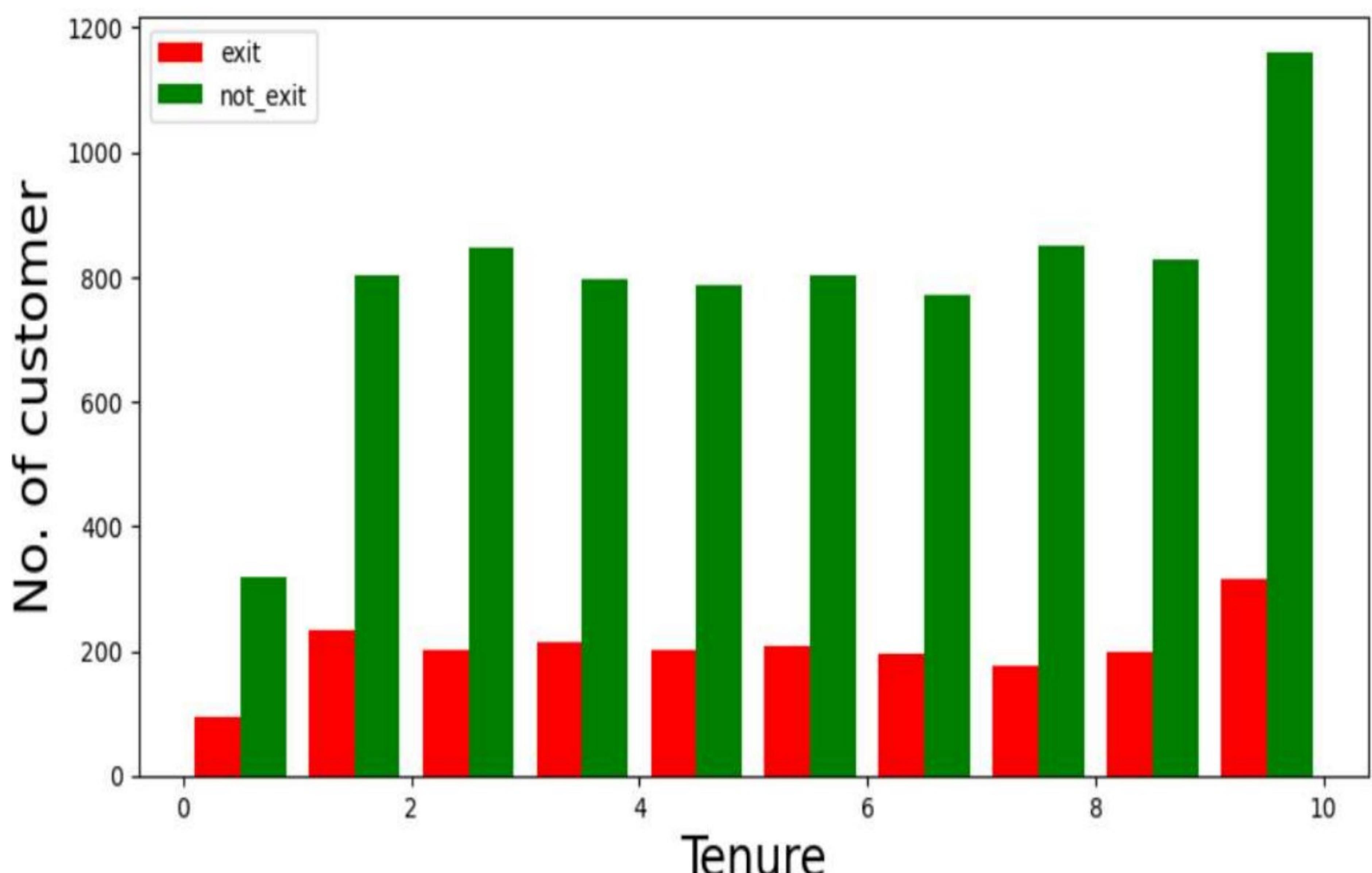
	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

```
In [12]: def visualization (x, y, xlabel):
```

```
    plt.figure(figsize=(10,5))
    plt.hist([x, y], color=['red', 'green'], label = ['exit','not_exit'])
    plt.xlabel(xlabel,fontsize=20)
    plt.ylabel("No. of customer", fontsize=20)
    plt.legend()
```

```
In [13]: df_churn_exited = df[df['Exited']==1]['Tenure']
df_churn_not_exited = df[df['Exited']==0]['Tenure']
```

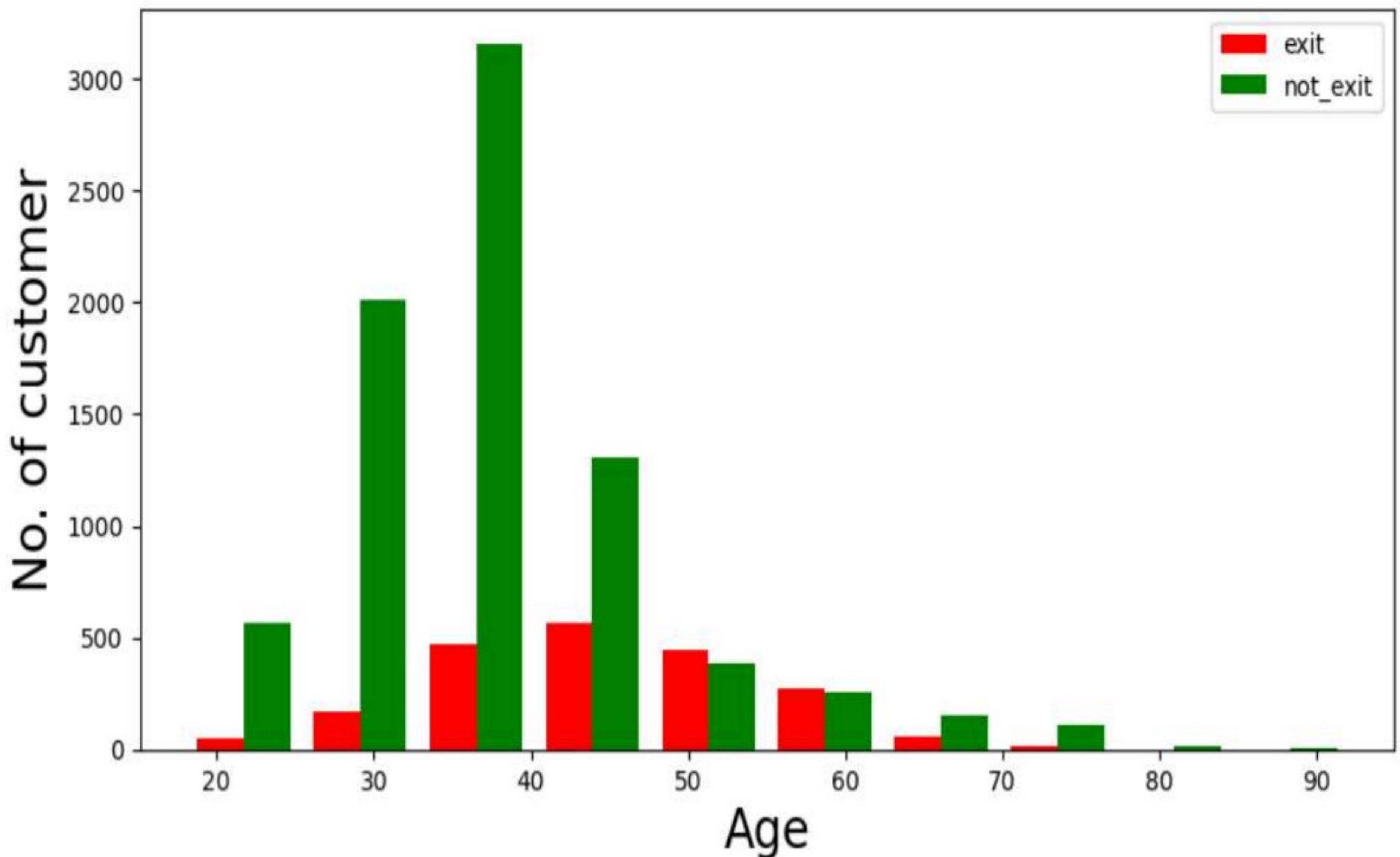
```
In [14]: visualization(df_churn_exited,df_churn_not_exited, "Tenure")
```



```
In [15]: df_churn_exited2 = df[df['Exited']==1]['Age']
```

```
df_churn_not_exited2 = df[df['Exited']==0]['Age']
```

```
visualization(df_churn_exited2,df_churn_not_exited2, "Age")
```



```
In [16]: X = df[['CreditScore', 'Gender', 'Age', 'Tenure', 'Balance', 'NumOfProducts', 'HasCrCard', 'IsActiveMember', 'EstimatedSalary', 'Exited']]
states = pd.get_dummies(df['Geography'], drop_first = True)
gender = pd.get_dummies(df['Gender'], drop_first = True)
```

```
In [17]: df = pd.concat([df, gender, states], axis=1)
```

```
In [18]: #Splitting the training and testing Dataset
df.head()
```

Out[18]:

	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	Male	Germany	Spain
0	619	France	Female	42	2	0.00	1	1	1	101348.88	1	0	0	0
1	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0	0	0	1
2	502	France	Female	42	8	159660.80	3	1	0	113931.57	1	0	0	0
3	699	France	Female	39	1	0.00	2	0	0	93826.63	0	0	0	0
4	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0	0	0	1

```
In [19]: X = df[['CreditScore', 'Age', 'Tenure', 'Balance', 'NumOfProducts', 'HasCrCard', 'IsActiveMember', 'EstimatedSalary', 'Male', 'Germany', 'Spain']]
y = df['Exited']
```

```
In [20]: from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.30)
```

```
In [21]: #Normalizing the values with mean as 0 and Standard Deviation as 1
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

```
In [22]: X_train
```

```
Out[22]: array([[ 0.38628108,  0.48664962,  1.04235774, ...,  0.90244757,
   1.72155741, -0.57339125],
 [ 0.80193718,  0.67749795, -0.68454375, ...,  0.90244757,
  1.72155741, -0.57339125],
 [ 0.81232858, -0.18131955, -0.68454375, ...,  0.90244757,
 -0.58086939, -0.57339125],
 ...,
 [-0.48659671,  1.15461879,  1.04235774, ...,  0.90244757,
 -0.58086939, -0.57339125],
 [-0.67364195,  0.48664962,  0.69697744, ...,  0.90244757,
  1.72155741, -0.57339125],
 [ 1.84107741, -0.37216788, -1.72068464, ..., -1.10809762,
 -0.58086939,  1.74400987]])
```

```
In [23]: X_test
```

```
Out[23]: array([[-0.33072568,  0.29580128, -1.02992404, ..., -1.10809762,
 -0.58086939, -0.57339125],
 [-0.39307409,  0.29580128,  1.73311833, ..., -1.10809762,
 -0.58086939, -0.57339125],
 [ 1.30072449,  1.25004295,  1.04235774, ...,  0.90244757,
 -0.58086939, -0.57339125],
 ...,
 [ 0.27197566, -0.75386455, -0.68454375, ...,  0.90244757,
 -0.58086939, -0.57339125],
 [ 0.83311138,  0.00952878,  0.35159714, ..., -1.10809762,
 -0.58086939,  1.74400987],
 [-0.73599037, -0.18131955,  0.00621685, ...,  0.90244757,
 -0.58086939,  1.74400987]])
```

```
In [24]: from sklearn.neural_network import MLPClassifier
```

```
ann = MLPClassifier(hidden_layer_sizes=(100,100,100),random_state = 0,max_iter =100, activation ='relu')
ann.fit(X_train,y_train)
```

```
C:\ProgramData\Anaconda\lib\site-packages\sklearn\neural_network\_multilayer_perceptron.py:692: ConvergenceWarning: Stochastic Optimizer: Maximum iterations (100) reached and the optimization hasn't converged yet.
  warnings.warn(
```

```
Out[24]: MLPClassifier(hidden_layer_sizes=(100, 100, 100), max_iter=100, random_state=0)
```

```
In [25]: y_pred =ann.predict(X_test)
```

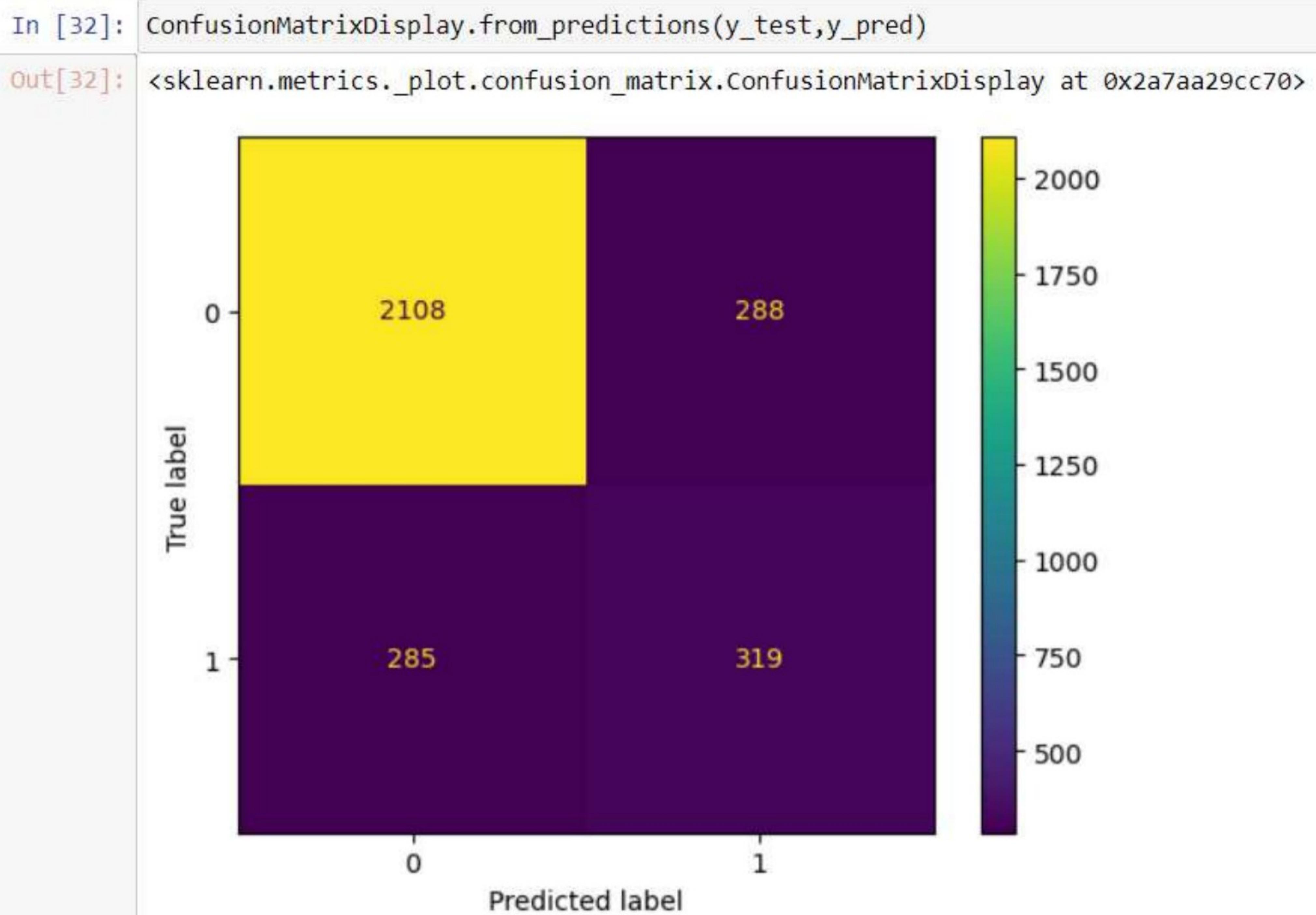
```
In [26]: y_pred
```

```
Out[26]: array([1, 0, 0, ..., 0, 1, 0], dtype=int64)
```

```
In [30]: from sklearn.metrics import ConfusionMatrixDisplay,accuracy_score, classification_report
```

```
In [31]: y_test.value_counts()
```

```
Out[31]: 0    2396
         1    604
Name: Exited, dtype: int64
```



```
In [39]: print(accuracy_score(y_test,y_pred))
```

0.809

```
In [40]: print(classification_report(y_test,y_pred))
```

	precision	recall	f1-score	support
0	0.88	0.88	0.88	2396
1	0.53	0.53	0.53	604
accuracy			0.81	3000
macro avg	0.70	0.70	0.70	3000
weighted avg	0.81	0.81	0.81	3000

Group B

Assignment No:4

Title of the Assignment: Implement K-Nearest Neighbors algorithm on diabetes.csv dataset. Compute confusion matrix, accuracy, error rate, precision and recall on the given dataset.

Dataset Description: We will try to build a machine learning model to accurately predict whether or not the patients in the dataset have diabetes or not?

The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

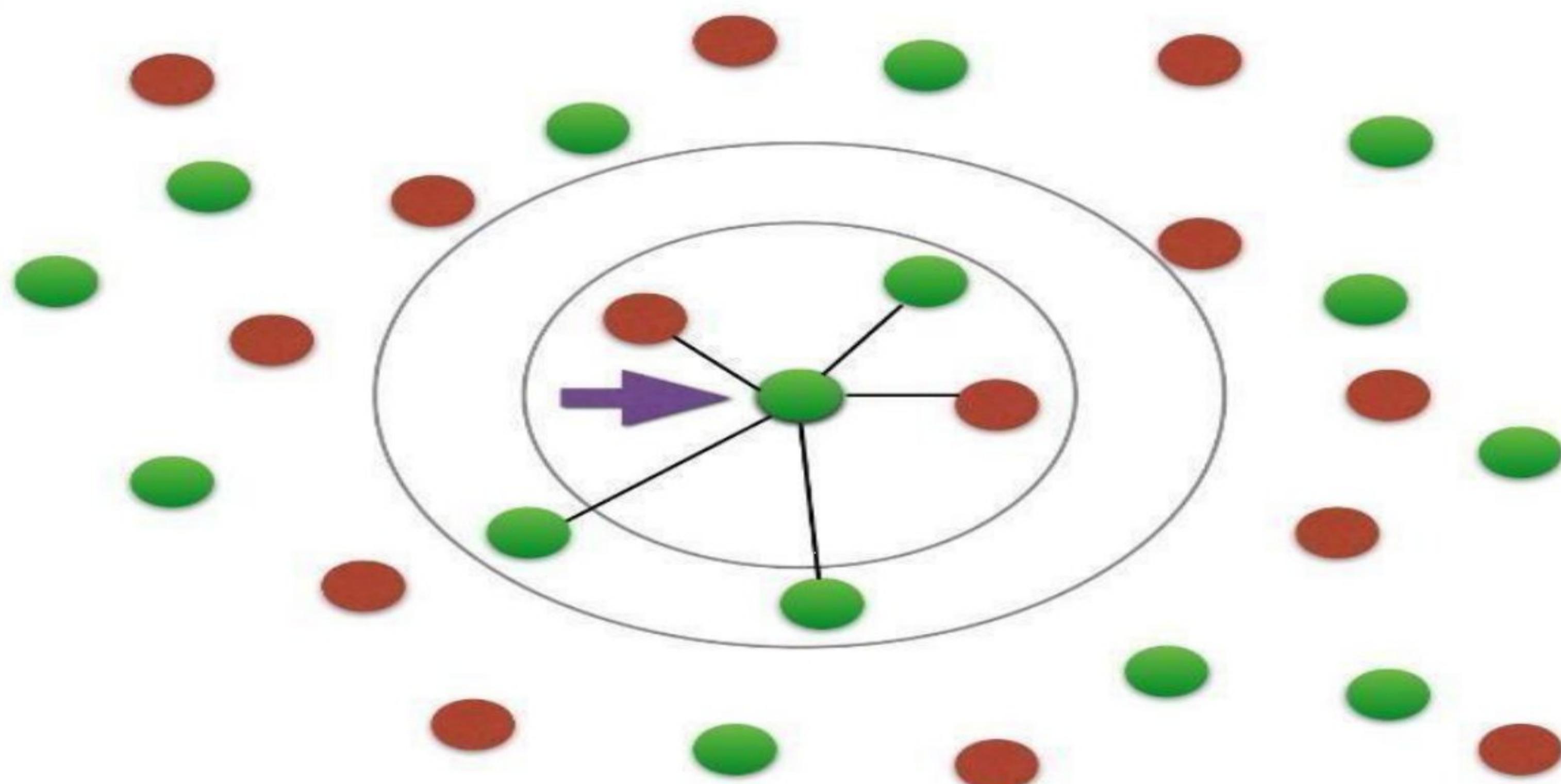
Link for Dataset: [Diabetes predication system with KNN algorithm | Kaggle](#)

Objective of the Assignment:

Students should be able to pre-process dataset and identify outliers, to check correlation and implement KNN algorithm and random forest classification models. Evaluate them with respective scores like confusion matrix, accuracy score, mean_squared_error, r2_score, roc_auc_score, roc_curve etc.

Prerequisite:

1. Basic knowledge of Python
2. Concept of Confusion Matrix
3. Concept of roc_auc curve.
4. Concept of Random Forest and KNN algorithms



k-Nearest-Neighbors (k-NN) is a supervised machine learning model. Supervised learning is when a model learns from data that is already labeled. A supervised learning model takes in a set of input objects and output values. The model then trains on that data to learn how to map the inputs to the desired output so it can learn to make predictions on unseen data.

k-NN models work by taking a data point and looking at the ‘k’ closest labeled data points. The data point is then assigned the label of the majority of the ‘k’ closest points.

For example, if $k = 5$, and 3 of points are ‘green’ and 2 are ‘red’, then the data point in question would be labeled ‘green’, since ‘green’ is the majority (as shown in the above graph).

Scikit-learn is a machine learning library for Python. In this tutorial, we will build a k-NN model using Scikit- learn to predict whether or not a patient has diabetes.

Reading in the training data

For our k-NN model, the first step is to read in the data we will use as input. For this example, we are using the diabetes dataset. To start, we will use Pandas to read in the data. I will not go into detail on Pandas, but it is a library you should become familiar with if you’re looking to dive further into data science and machine learning.

```
Import pandas as pd      #read in the data using pandas
df = pd.read_csv('data/diabetes_data.csv')      #check data has been
read in properly
df.head()
```

	pregnancies	glucose	diastolic	triceps	insulin	bmi	dpf	age	diabetes
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Next, let's see how much data we have. We will call the 'shape' function on our dataframe to see how many rows and columns there are in our data. The rows indicate the number of patients and the columns indicate the number of features (age, weight, etc.) in the dataset for each patient.

```
#check number of rows and columns in dataset df.shape
```

```
Op → (768,9)
```

We can see that we have 768 rows of data (potential diabetes patients) and 9 columns (8 input features and 1 target output).

Split up the dataset into inputs and targets

Now let's split up our dataset into inputs (X) and our target (y). Our input will be every column except 'diabetes' because 'diabetes' is what we will be attempting to predict. Therefore, 'diabetes' will be our target.

We will use pandas 'drop' function to drop the column 'diabetes' from our dataframe and store it in the variable 'X'. This will be our input.

```
#create a dataframe with all training data except the target column
```

```
X = df.drop(columns=['diabetes'])#check that the target variable has been removed
```

```
X.head()
```

	pregnancies	glucose	diastolic	triceps	insulin	bmi	dpf	age
0	6	148	72	35	0	33.6	0.627	50
1	1	85	66	29	0	26.6	0.351	31
2	8	183	64	0	0	23.3	0.672	32
3	1	89	66	23	94	28.1	0.167	21
4	0	137	40	35	168	43.1	2.288	33

We will insert the 'diabetes' column of our dataset into our target variable (y).

```
#Separate target values
```

```
y = df['diabetes'].values
```

```
#view target values y[0:5]
```

```
array([1, 0, 1, 0, 1])
```

Split the dataset into train and test data

Now we will split the dataset into training data and testing data. The training data is the data that the model will learn from. The testing data is the data we will use to see how well the model performs on unseen data.

Scikit-learn has a function we can use called ‘train_test_split’ that makes it easy for us to split our dataset into training and testing data.

```
from sklearn.model_selection import train_test_split#split dataset into train and test data X_train,  
X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1,
```

‘train_test_split’ takes in 5 parameters. The first two parameters are the input and target data we split up earlier. Next, we will set ‘test_size’ to 0.2. This means that 20% of all the data will be used for testing, which leaves 80% of the data as training data for the model to learn from. Setting ‘random_state’ to 1 ensures that we get the same split each time so we can reproduce our results.

Setting ‘stratify’ to y makes our training split represent the proportion of each value in the y variable. For example, in our dataset, if 25% of patients have diabetes and 75% don’t have diabetes, setting ‘stratify’ to y will ensure that the random split has 25% of patients with diabetes and 75% of patients without diabetes.

Building and training the model Next, we have to build the model. Here is the code:

Next, we have to build the model. Here is the code:

```
from sklearn.neighbors import KNeighborsClassifier# Create KNN classifier knn = KNeighbors  
Classifier(n_neighbors = 3)# Fit the classifier to the data knn.fit(X_train,y_train)
```

First, we will create a new k-NN classifier and set ‘n_neighbors’ to 3. To recap, this means that if at least 2 out of the 3 nearest points to a new data point are patients without diabetes, then the new data point will be labeled as ‘no diabetes’, and vice versa. In other words, a new data point is labeled with by majority from the 3 nearest points.

We have set ‘n_neighbors’ to 3 as a starting point. We will go into more detail below on how to better

select a value for ‘n_neighbors’ so that the model can improve its performance.

Next, we need to train the model. In order to train our new model, we will use the ‘fit’ function and pass in our training data as parameters to fit our model to the training data.

Testing the model

Once the model is trained, we can use the ‘predict’ function on our model to make predictions on our test data. As seen when inspecting ‘y’ earlier, 0 indicates that the patient does not have diabetes and 1 indicates that the patient does have diabetes. To save space, we will only show print the first 5 predictions of our test set.

```
#show first 5 model predictions on the test data knn.predict(X_test)[0:5]
```

```
array([ 0,  0,  0,  0,  1])
```

We can see that the model predicted ‘no diabetes’ for the first 4 patients in the test set and ‘has diabetes’ for the 5th patient.

Now let’s see how accurate our model is on the full test set. To do this, we will use the ‘score’ function and pass in our test input and target data to see how well our model predictions match up to the actual results

```
#check accuracy of our model on the test data knn.score(X_test, y_test)
```

0.66883116883116878

Our model has an accuracy of approximately 66.88%. It’s a good start, but we will see how we can increase model performance below.

Congrats! You have now built an amazing k-NN model!

k-Fold Cross-Validation

Cross-validation is when the dataset is randomly split up into ‘k’ groups. One of the groups is used as the test set and the rest are used as the training set. The model is trained on the training set and scored on the test set. Then the process is repeated until each unique group has been used as the test set.

For example, for 5-fold cross validation, the dataset would be split into 5 groups, and the model would be trained and tested 5 separate times so each group would get a chance to be the test set. This can be seen in the graph below.

Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Metric 1
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Metric 2
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Metric 3
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Metric 4
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Metric 5

Training data
Test data

4- fold cross validation (image credit)

The train-test-split method we used in earlier is called ‘holdout’. Cross-validation is better than using the holdout method because the holdout method score is dependent on how the data is split into train and test sets. Cross-validation gives the model an opportunity to test on multiple splits so we can get a better idea on how the model will perform on unseen data.

In order to train and test our model using cross-validation, we will use the ‘cross_val_score’ function with a cross-validation value of 5. ‘cross_val_score’ takes in our k-NN model and our data as parameters. Then it splits our data into 5 groups and fits and scores our data 5 separate times, recording the accuracy score in an array each time. We will save the accuracy scores in the ‘cv_scores’ variable.

To find the average of the 5 scores, we will use numpy’s mean function, passing in ‘cv_score’. Numpy is a useful math library in Python

```
from sklearn.model_selection import cross_val_score import numpy as np#create a new KNN model
knn_cv = KNeighborsClassifier(n_neighbors=3)#train model with cv of 5
cv_scores = cross_val_score(knn_cv, X, y, cv=5)#print each cv score (accuracy) and average them
print(cv_scores)
print('cv_scores mean:{}'.format(np.mean(cv_scores)))

[ 0.68181818  0.69480519  0.75324675  0.75163399  0.68627451]
cv_scores mean:0.7135557253204311
```

Using cross-validation, our mean score is about 71.36%. This is a more accurate representation of how our model will perform on unseen data than our earlier testing using the holdout method.

Hypertuning model parameters using GridSearchCV

When built our initial k-NN model, we set the parameter ‘n_neighbors’ to 3 as a starting point with no real logic behind that choice.

Hypertuning parameters is when you go through a process to find the optimal parameters for your model to improve accuracy. In our case, we will use GridSearchCV to find the optimal value for ‘n_neighbors’.

GridSearchCV works by training our model multiple times on a range of parameters that we specify. That way, we can test our model with each parameter and figure out the optimal values to get the best accuracy results.

For our model, we will specify a range of values for ‘n_neighbors’ in order to see which value works best for our model. To do this, we will create a dictionary, setting ‘n_neighbors’ as the key and using numpy to create an array of values from 1 to 24.

Our new model using grid search will take in a new k-NN classifier, our param_grid and a cross-validation value of 5 in order to find the optimal value for ‘n_neighbors’

```
from sklearn.model_selection import GridSearchCV#create new a knn model
knn2 = KNeighborsClassifier()#create a dictionary of all values we want to test for n_neighbors
param_grid = {'n_neighbors': np.arange(1, 25)}#use gridsearch to test all values for n_neighbors
knn_gscv = GridSearchCV(knn2, param_grid, cv=5)#fit model to data
knn_gscv.fit(X, y)
```

After training, we can check which of our values for ‘n_neighbors’ that we tested performed the best. To do this, we will call ‘best_params_’ on our model.

```
#check top performing n_neighbors value knn gscv.best_params  
{'n_neighbors': 14}
```

We can see that 14 is the optimal value for ‘n neighbors’. We can use the ‘best_score_’ function to check the accuracy of our model when ‘n_neighbors’ is 14. ‘best_score_’ outputs the mean accuracy of the scores obtained through cross-validation.

```
#check mean score for the top performing value of n_neighbors knn_gscv.best_score
```

0.7578125

By using grid search to find the optimal parameter for our model, we have improved our model accuracy by over 4%!

Code <https://www.kaggle.com/code/shrutimech1earn/step-by-step-diabetes-classification-knn-detailed>

Conclusion:

In this way we build a neural network-based classifier that can determine whether they will leave or not in the next 6 months.

Assignment No. 4

```
In [1]: import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn import metrics

In [2]: df =pd.read_csv("diabetes.csv")

In [3]: df.columns
Out[3]: Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
       'BMI', 'Pedigree', 'Age', 'Outcome'],
       dtype='object')

In [4]: df.isnull().sum()
Out[4]: Pregnancies      0
Glucose          0
BloodPressure    0
SkinThickness    0
Insulin          0
BMI              0
Pedigree         0
Age              0
Outcome          0
dtype: int64

In [6]: x = df.drop('Outcome', axis=1)
y = df['Outcome']

In [8]: from sklearn.preprocessing import scale
x= scale(x)
#split into train and test
x_train, x_test, y_train, y_test= train_test_split(x,y, test_size=0.3,random_state=42)

In [9]: from sklearn.neighbors import KNeighborsClassifier
knn =KNeighborsClassifier(n_neighbors= 7)
knn.fit(x_train,y_train)
y_pred=knn.predict(x_test)

In [10]: print("Confusion Matrix:")
cs= metrics.confusion_matrix(y_test,y_pred)
print(cs)

Confusion Matrix:
[[123  28]
 [ 37  43]]

In [11]: print('Accuracy:',metrics.accuracy_score(y_test,y_pred))
Accuracy: 0.7186147186147186

In [15]: total_misclassified = cs[0,1]+ cs[1,0]
print(total_misclassified)
total_examples =cs[0,0]+ cs[0,1]+ cs[1,0]+ cs[1,1]
print(total_examples)
print("Error_rate",total_misclassified/total_examples)
print("Error_rate",1-metrics.accuracy_score(y_test,y_pred))
```

```
65
231
Error_rate 0.2813852813852814
Error_rate 0.2813852813852814
```

```
In [16]: print("Precision Score",metrics.precision_score(y_test,y_pred))
Precision Score 0.6056338028169014
```

```
In [18]: print("Recall Score",metrics.recall_score(y_test,y_pred))
Recall Score 0.5375
```

```
In [19]: print("Classification Report",metrics.classification_report(y_test,y_pred))
```

Classification Report	precision	recall	f1-score	support
0	0.77	0.81	0.79	151
1	0.61	0.54	0.57	80
accuracy			0.72	231
macro avg	0.69	0.68	0.68	231
weighted avg	0.71	0.72	0.71	231

Group B

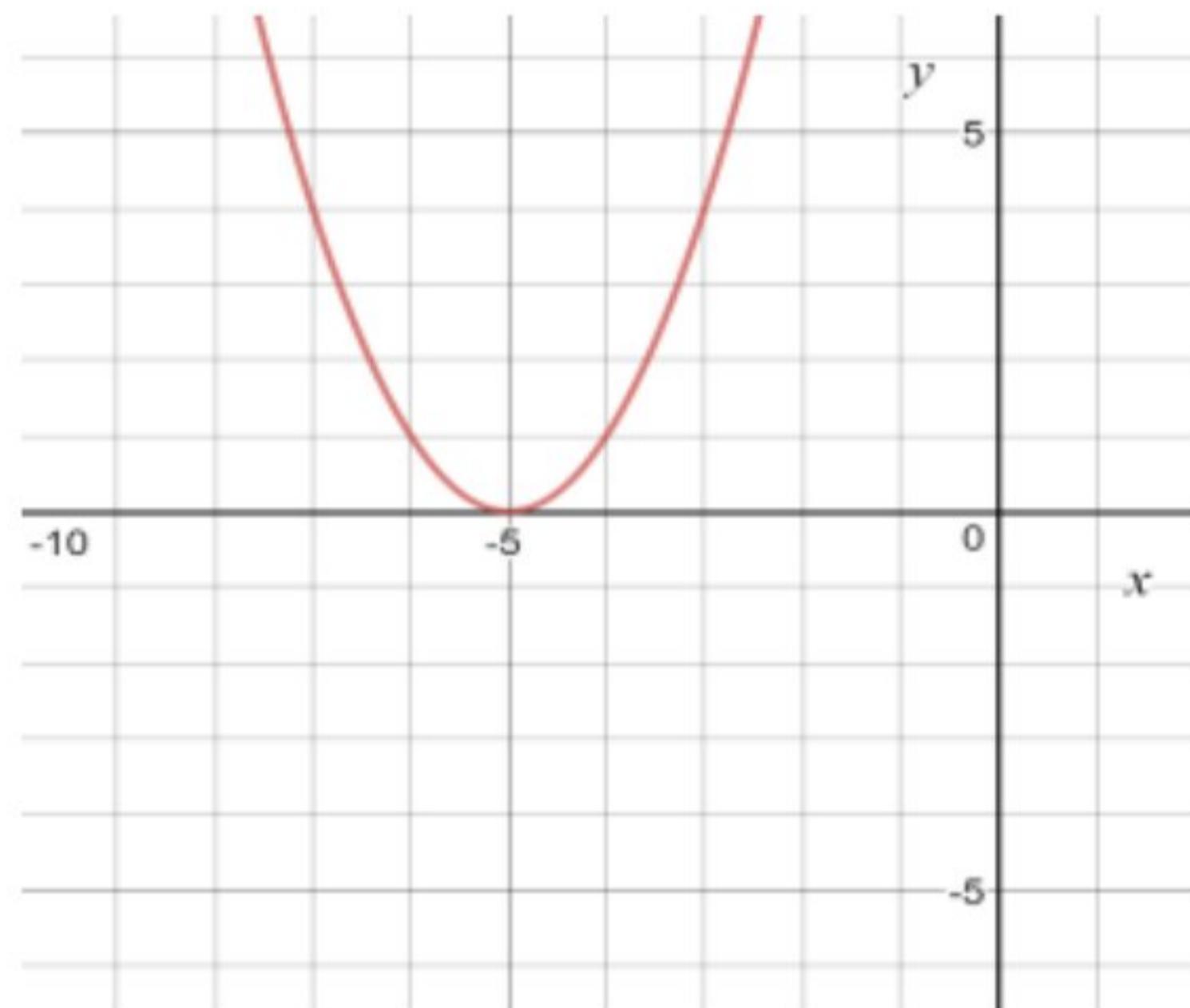
Assignment No:5

Title of Assignment: Implement Gradient Descent Algorithm.

Problem Statement: Develop a program in Python to create a Gradient Descent.

Example by hand :

Question : Find the local minima of the function $y=(x+5)^2$ starting from the point $x=3$



Solution : We know the answer just by looking at the graph. $y = (x+5)^2$ reaches its minimum value when $x = -5$ (i.e when $x=-5$, $y=0$). Hence $x=-5$ is the local and global minima of the function.

Now, let's see how to obtain the same numerically using gradient descent.

Step 1 : Initialize $x = 3$. Then, find the gradient of the function, $dy/dx = 2*(x+5)$.

Step 2 : Move in the direction of the negative of the gradient ([Why?](#)). But wait, how much to move? For that, we require a learning rate. Let us assume the **learning rate → 0.01**

Step 3 : Let's perform 2 iterations of gradient descent

Step 4: We can observe that the X value is slowly decreasing and should converge to -5 (the local minima). However, how many iterations should we perform?

Let us set a precision variable in our algorithm which calculates the difference between two consecutive "x" values. If the difference between x values from 2 consecutive iterations is lesser than the precision we set, stop the algorithm!

Step 5: We can observe that the X value is slowly decreasing and should converge to -5 (the local minima).

However, how many iterations should we perform?

Let us set a precision variable in our algorithm which calculates the difference between two consecutive "x" values. If the difference between x values from 2 consecutive iterations is lesser than the precision we set, stop the algorithm!

Gradient descent in Python :

Step 1 : Initialize parameters

```
cur_x = 3 # The algorithm starts at x=3 rate = 0.01 # Learning rate  
precision = 0.000001 #This tells us when to stop the algorithm previous_step_size = 1 #  
max_iters = 10000 # maximum number of iterations iters = 0 #iteration counter  
df = lambda x: 2*(x+5) #Gradient of our function
```

Step 2 : Run a loop to perform gradient descent :

i. Stop loop when difference between x values from 2 consecutive iterations is less than 0.000001 or when number of iterations exceeds 10,000

```
while previous_step_size > precision and iters < max_iters: prev_x = cur_x #Store current x value  
in prev_x
```

```
cur_x = cur_x - rate * df(prev_x) #Grad descent previous_step_size = abs(cur_x - prev_x) #Change  
in x iters = iters+1 #iteration count
```

```
print("Iteration",iters,"X value is",cur_x) #Print
```

```
iterations print("The local minimum occurs at", cur_x)
```

Facilities:

Google Colab , Jupiter notebook

Input:

To find the local minima of the given function from its starting point.

Output:

Finds the optimal solution by taking a step in the direction of the maximum rate of decrease of the function.

Conclusion:

Hence, we have successfully studied implementation of Gradient Descent Algorithm successfully.

In [7]:

```
x = 2
lr = 0.01
precision = 0.000001
previous_step_size = 1
max_iter = 10000
iters = 0
gf = lambda x: (x+3) ** 2
```

In [8]:

```
1 import matplotlib.pyplot as plt
```

In [9]:

```
1 gd = []
```

In [10]:

```
1 while precision < previous_step_size and iters < max_iter:
2     prev = x
3     x = x - lr * gf(prev)
4     previous_step_size = abs(x - prev)
5     iters += 1
6     print('Iteration:', iters, 'Value:', x)
7     gd.append(x)
```

```
Iteration: 1 Value: 1.75
Iteration: 2 Value: 1.524375
Iteration: 3 Value: 1.31967530859375
Iteration: 4 Value: 1.133079360877005
Iteration: 5 Value: 0.9622559108439301
Iteration: 6 Value: 0.8052611918137536
Iteration: 7 Value: 0.6604610644345152
Iteration: 8 Value: 0.5264713123921045
Iteration: 9 Value: 0.4021113132208596
Iteration: 10 Value: 0.28636769934540596
Iteration: 11 Value: 0.1783655727923978
Iteration: 12 Value: 0.07734549564927831
Iteration: 13 Value: -0.017355057346650715
Iteration: 14 Value: -0.10631676588600673
Iteration: 15 Value: -0.19005079247993095
Iteration: 16 Value: -0.26900893796835756
Iteration: 17 Value: -0.34359205977732477
Iteration: 18 Value: -0.41415709122610556
Iteration: 19 Value: -0.4810229267146679
Iteration: 20 Value: -0.544752916720202
```

In [12]:

```
1 print('Local Minima:', x)
```

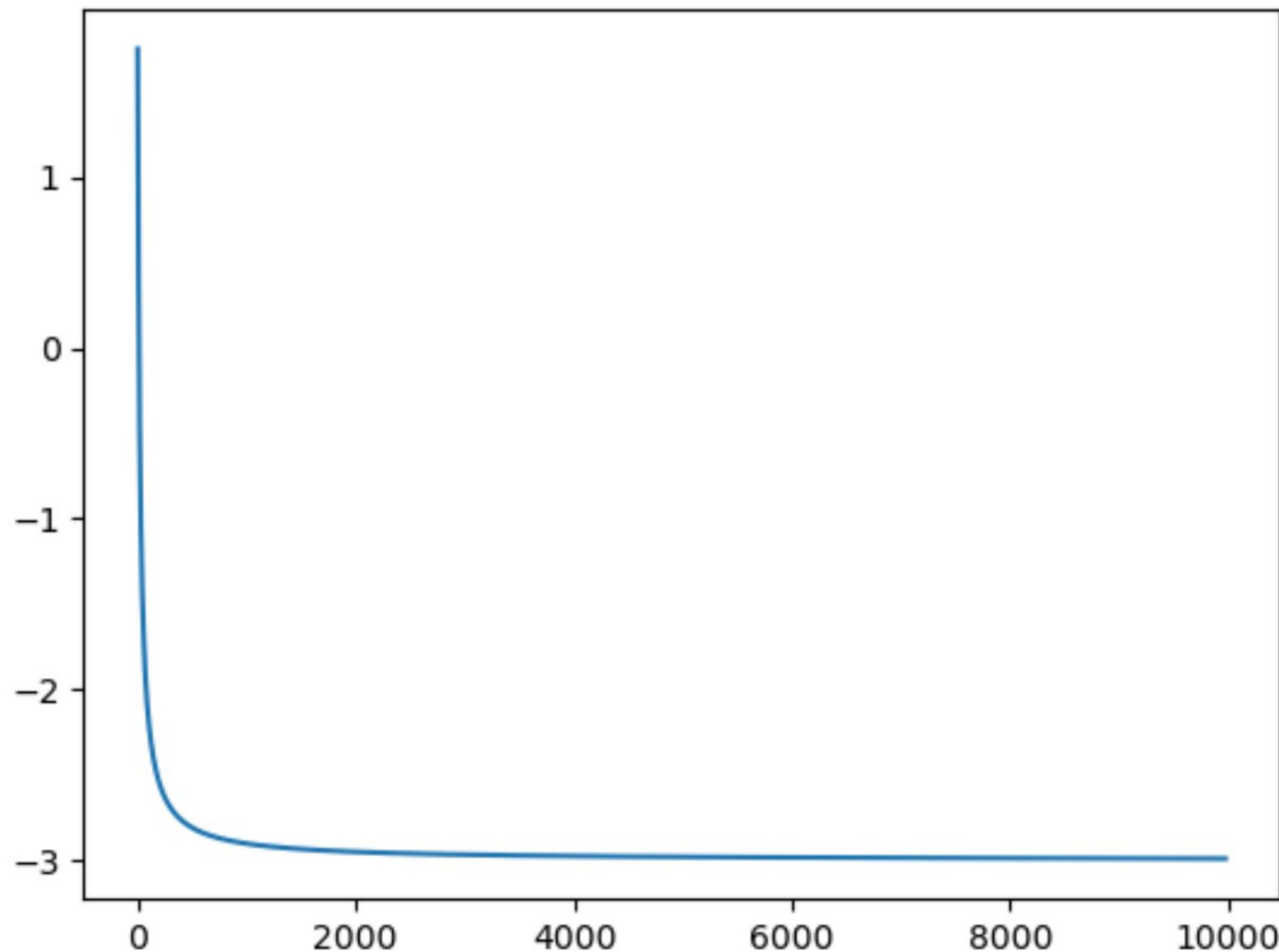
```
Local Minima: -2.990001240409911
```

In [15]:

```
1 plt.plot(gd)
```

Out[15]:

```
[<matplotlib.lines.Line2D at 0x206d090a1f0>]
```



In []:

```
1
```

Assignment No.6

Title of the Assignment: Implement K-Means clustering/ hierarchical clustering on sales_data_sample.csv dataset. Determine the number clusters using the elbow method.

```
In [1]: 1 import pandas as pd
2 import numpy as np
3 import seaborn as sns
4 import matplotlib.pyplot as plt
5
```

```
In [2]: 1 from sklearn.cluster import KMeans
2 from sklearn.decomposition import PCA
```

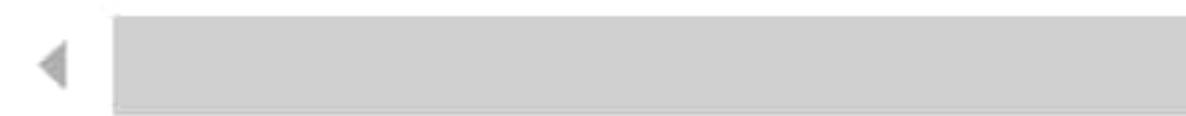
```
In [8]: 1 df = pd.read_csv("sales_data_sample.csv", encoding ="Latin-1")
```

```
In [11]: 1 df.head()
```

Out[11]:

	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDA
0	10107	30	95.70		2 2871.00	2/24/200
1	10121	34	81.35		5 2765.90	5/7/2003 0
2	10134	41	94.74		2 3884.34	7/1/2003 0
3	10145	45	83.26		6 3746.70	8/25/200
4	10159	49	100.00		14 5205.27	10/10/200

5 rows × 25 columns



```
In [12]: 1 df.shape
```

Out[12]: (2823, 25)

```
In [13]: 1 df.describe()
```

Out[13]:

	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES
count	2823.000000	2823.000000	2823.000000	2823.000000	2823.000000
mean	10258.725115	35.092809	83.658544	6.466171	3553.889072
std	92.085478	9.741443	20.174277	4.225841	1841.865106
min	10100.000000	6.000000	26.880000	1.000000	482.130000
25%	10180.000000	27.000000	68.860000	3.000000	2203.430000
50%	10262.000000	35.000000	95.700000	6.000000	3184.800000
75%	10333.500000	43.000000	100.000000	9.000000	4508.000000
max	10425.000000	97.000000	100.000000	18.000000	14082.800000



```
In [15]: 1 df.info()
```

In [15]: 1 df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2823 entries, 0 to 2822
Data columns (total 25 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   ORDERNUMBER      2823 non-null    int64  
 1   QUANTITYORDERED 2823 non-null    int64  
 2   PRICEEACH        2823 non-null    float64 
 3   ORDERLINENUMBER 2823 non-null    int64  
 4   SALES            2823 non-null    float64 
 5   ORDERDATE         2823 non-null    object  
 6   STATUS            2823 non-null    object  
 7   QTR_ID            2823 non-null    int64  
 8   MONTH_ID          2823 non-null    int64  
 9   YEAR_ID           2823 non-null    int64  
 10  PRODUCTLINE       2823 non-null    object  
 11  MSRP              2823 non-null    int64  
 12  PRODUCTCODE       2823 non-null    object  
 13  CUSTOMERNAME     2823 non-null    object  
 14  PHONE              2823 non-null    object  
 15  ADDRESSLINE1      2823 non-null    object  
 16  ADDRESSLINE2      302  non-null    object  
 17  CITY               2823 non-null    object  
 18  STATE              1337 non-null    object  
 19  POSTALCODE         2747 non-null    object  
 20  COUNTRY            2823 non-null    object  
 21  TERRITORY          1749 non-null    object  
 22  CONTACTLASTNAME   2823 non-null    object  
 23  CONTACTFIRSTNAME  2823 non-null    object  
 24  DEALSIZE           2823 non-null    object  
dtypes: float64(2), int64(7), object(16)
memory usage: 551.5+ KB
```

In [16]: 1 df.isnull().sum()

```
In [16]: 1 df.isnull().sum()
```

```
Out[16]: ORDERNUMBER          0
QUANTITYORDERED        0
PRICEEACH              0
ORDERLINENUMBER        0
SALES                  0
ORDERDATE              0
STATUS                 0
QTR_ID                 0
MONTH_ID               0
YEAR_ID                0
PRODUCTLINE            0
MSRP                   0
PRODUCTCODE            0
CUSTOMERNAME           0
PHONE                  0
ADDRESSLINE1            0
ADDRESSLINE2            2521
CITY                   0
STATE                  1486
POSTALCODE             76
COUNTRY                0
TERRITORY              1074
CONTACTLASTNAME        0
CONTACTFIRSTNAME       0
DEALSIZE                0
dtype: int64
```

```
In [17]: 1 df.dtypes
```

```
STATUS                object
QTR_ID                int64
MONTH_ID              int64
YEAR_ID               int64
PRODUCTLINE           object
MSRP                  int64
PRODUCTCODE           object
CUSTOMERNAME          object
PHONE                 object
ADDRESSLINE1           object
ADDRESSLINE2           object
CITY                  object
STATE                 object
POSTALCODE            object
COUNTRY               object
TERRITORY             object
CONTACTLASTNAME       object
CONTACTFIRSTNAME      object
DEALSIZE               object
dtype: object
```

```
In [18]: 1 df_drop = ['ADDRESSLINE1', 'ADDRESSLINE2', 'STATUS', 'POSTALCODE', 'CITY']
```

```
In [19]: 1 df = df.drop(df_drop, axis=1)
```

```
In [21]: 1 df.isnull().sum()
```

```
In [21]: 1 df.isnull().sum()
```

```
Out[21]: ORDERNUMBER      0  
QUANTITYORDERED      0  
PRICEEACH          0  
ORDERLINENUMBER      0  
SALES              0  
ORDERDATE           0  
QTR_ID              0  
MONTH_ID             0  
YEAR_ID              0  
PRODUCTLINE          0  
MSRP                 0  
PRODUCTCODE          0  
COUNTRY              0  
DEALSIZE              0  
dtype: int64
```

```
In [22]: 1 df.dtypes
```

```
Out[22]: ORDERNUMBER      int64  
QUANTITYORDERED      int64  
PRICEEACH            float64  
ORDERLINENUMBER      int64  
SALES                float64  
ORDERDATE             object  
QTR_ID                int64  
MONTH_ID               int64  
YEAR_ID                int64  
PRODUCTLINE             object  
MSRP                  int64  
PRODUCTCODE             object  
COUNTRY                object  
DEALSIZE                object  
dtype: object
```

```
In [26]: 1 df['COUNTRY'].unique()  
2
```

```
Out[26]: array(['USA', 'France', 'Norway', 'Australia', 'Finland', 'Austria', 'UK',  
   'Spain', 'Sweden', 'Singapore', 'Canada', 'Japan', 'Italy',  
   'Denmark', 'Belgium', 'Philippines', 'Germany', 'Switzerland',  
   'Ireland'], dtype=object)
```

```
In [27]: 1 df['PRODUCTLINE'].unique()
```

```
Out[27]: array(['Motorcycles', 'Classic Cars', 'Trucks and Buses', 'Vintage Cars',  
   'Planes', 'Ships', 'Trains'], dtype=object)
```

```
In [28]: 1 df['DEALSIZE'].unique()
```

```
Out[28]: array(['Small', 'Medium', 'Large'], dtype=object)
```

```
In [30]: 1 productline = pd.get_dummies(df['PRODUCTLINE'])  
2 Dealsize = pd.get_dummies(df['DEALSIZE'])
```

```
In [35]: 1 df = pd.concat([df, productline, Dealsize], axis=1)
```

```
In [35]: 1 df = pd.concat([df, productline,Dealsize],axis=1)
```

```
In [36]: 1 df_drop = ['COUNTRY', 'PRODUCTLINE', 'DEALSIZE']
2 df = df.drop(df_drop, axis =1 )
```

```
In [37]: 1 df[ 'PRODUCTCODE' ] = pd.Categorical(df[ 'PRODUCTCODE' ]).codes
```

```
In [40]: 1 df.drop('ORDERDATE', axis = 1, inplace=True)
```

```
In [41]: 1 df.dtypes
```

```
Out[41]: ORDERNUMBER           int64
QUANTITYORDERED        int64
PRICEEACH              float64
ORDERLINENUMBER         int64
SALES                  float64
QTR_ID                 int64
MONTH_ID                int64
YEAR_ID                 int64
MSRP                   int64
PRODUCTCODE             int8
Classic Cars            uint8
Motorcycles              uint8
Planes                  uint8
Ships                   uint8
Trains                  uint8
Trucks and Buses        uint8
Vintage Cars            uint8
Large                   uint8
Medium                  uint8
Small                   uint8
dtype: object
```

```
In [43]: 1 distortions = []
2 K = range(1,10)
3 for k in K:
4     kmeanModel = KMeans(n_clusters=k)
5     kmeanModel.fit(df)
6     distortions.append(kmeanModel. inertia_)
```

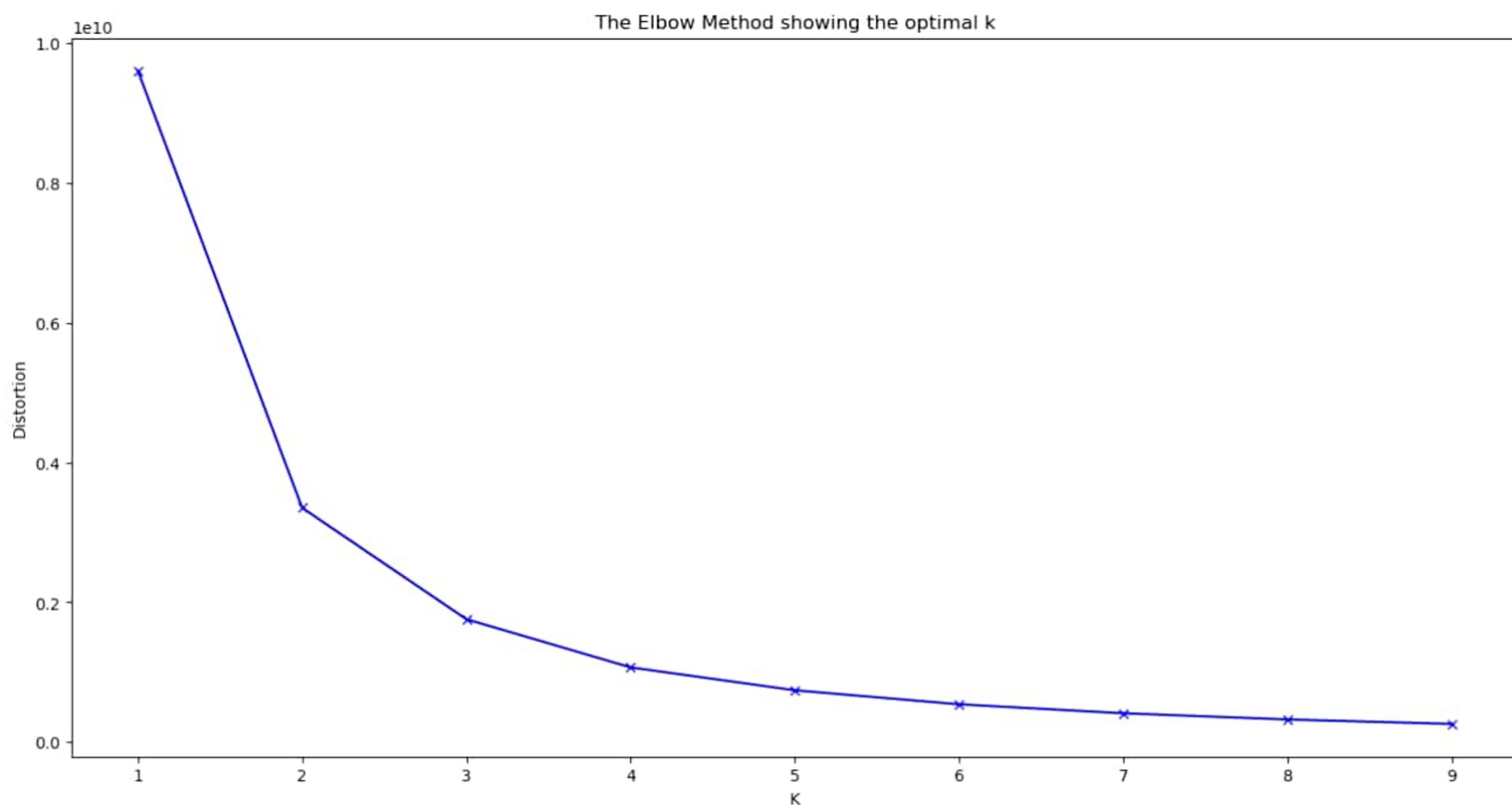
```
In [44]: 1 plt.figure(figsize=(16,8))
```

In [44]:

```

1 plt.figure(figsize=(16,8))
2 plt.plot(K, distortions, 'bx-')
3 plt.xlabel('K')
4 plt.ylabel('Distortion')
5 plt.title('The Elbow Method showing the optimal k')
6 plt.show()

```



In [45]:

```
1 x_train = df.values
```

In [46]:

```
1 x_train.shape
```

Out[46]:

(2823, 20)

In [49]:

```

1 model = KMeans(n_clusters=3, random_state=2)
2 model = model.fit(x_train)
3 predictions = model.predict(x_train)

```

In [51]:

```
1 unique, counts = np.unique(predictions, return_counts=True)
```

In [52]:

```
1 counts = counts.reshape(1,3)
```

In [55]:

```
1 counts_df = pd.DataFrame(counts, columns=['Cluster', 'Cluster2', 'Cluster3'])
```

In [57]:

```
1 counts_df.head()
```

Out[57]:

	Cluster	Cluster2	Cluster3
0	1083	1367	373

In [59]:

```
1 pca = PCA(n_components=2)
```

```
In [59]: 1 pca = PCA(n_components=2)
2 reduced_X = pd.DataFrame(pca.fit_transform(x_train), columns=['PCA1', 'PCA2'])
3 reduced_X.head()
```

Out[59]:

	PCA1	PCA2
0	-682.790370	-151.271539
1	-787.939342	-136.994834
2	330.482091	-125.876905
3	192.812426	-114.565402
4	1651.330150	-103.067424

```
In [ ]: 1
```