Swarnika Singh
Div: D15B Roll no: 65

EXPERIMENT-4

AIM: Experiment to perform exploratory data analysis and data visualization using python

Theory*:*

Exploratory Data Analysis :Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics, often with visual methods.It involves uncovering patterns, trends, relationships, and anomalies in the data to gain insights and guide further analysis.

Objectives:

- Understand the structure and distribution of the data.
- Identify missing values, outliers, and errors in the dataset.
- Explore relationships and dependencies between variables.
- Formulate hypotheses and generate insights to guide subsequent analysis.

Key Techniques:

- Summary Statistics: Calculating descriptive statistics such as mean, median, mode, standard deviation, minimum, maximum, and quartiles.
- Dispersion- Dispersion, also known as variability or spread, refers to the extent to which data points in a dataset are spread out or dispersed from the central tendency. It quantifies the degree of variability or diversity within the dataset-

Variance: The average of the squared differences from the mean. It measures the average deviation of data points from the mean.
Standard Deviation: The square root of the variance. It represents the average distance of data points from the mean.
Range: The difference between the maximum and minimum values in a dataset. It provides a measure of the spread of values in the dataset.
Interquartile Range (IQR): The range between the first quartile (25th percentile) and the third quartile (75th percentile). It represents the middle 50% of the data and is less sensitive to outliers compared to the range.

- Central tendency: Central tendency refers to the typical or central value around which data points in a dataset tend to cluster. It provides a summary measure of the "center" of the data distribution-

Mean: The arithmetic average of a set of values. It is calculated by summing all values and dividing by the number of values.
Median: The middle value of a dataset when it is ordered from least to greatest. It divides the dataset into two equal parts, with half of the values below and half above the median.
Mode: The value that appears most frequently in a dataset. It represents the most common value or values in the dataset.

Benefits:

- Helps in understanding the data before performing more complex analyses or modeling.
- Provides insights into patterns and trends that may not be apparent from raw data.
- Guides data preprocessing and feature engineering steps.
- Assists in identifying potential hypotheses or research questions for further investigation.

Data Visualization (DV)-Data Visualization (DV) is the graphical representation of data to communicate insights, patterns, and trends effectively.It involves creating visual representations such as charts, graphs, and maps to convey complex information in an intuitive manner.

Objectives:

- Provide a visual overview of the data and its main characteristics.
- Enable users to explore and interact with the data to gain insights.
- Communicate findings and results to stakeholders in a clear and understandable way.

Key Techniques:

- Charts and Plots: Includes histograms, bar charts, line plots, scatter plots, box plots, and pie charts.
- Geographic Visualization: Maps and geospatial visualizations to represent data across geographical regions.
- Interactive Visualization: Tools and techniques that allow users to interact with visualizations and explore data dynamically.
- Dashboarding: Aggregating multiple visualizations into a single dashboard for comprehensive analysis.

Benefits:

- Facilitates understanding complex datasets and patterns more intuitively.
- Enhances communication and storytelling by presenting data in a visually appealing manner.
- Encourages exploration and discovery by allowing users to interact with the data.
- Supports decision-making processes by providing actionable insights derived from data analysis.

Overall, Exploratory Data Analysis and Data Visualization are indispensable components of the data analysis process, enabling analysts and stakeholders to derive meaningful insights and make informed decisions based on data. They play a crucial role in transforming raw data into actionable knowledge.

CODE-

1. Descriptive analysis - statistical measures of data (Central tendency)-

```
[15] import pandas as pd

     # Load your dataset
     df = pd.read_csv('passwordscom.csv')  # Replace 'your_dataset.csv' with the path to your dataset

     # Descriptive analysis - statistical measures of data (Central tendency)
     central_tendency = df.describe()

     # Display statistical measures
     print("Statistical measures of data (Central tendency):\n", central_tendency)
```

```
Statistical measures of data (Central tendency):
                length     num_chars   num_digits    num_upper     num_lower  \
count  10000.000000  10000.000000  9998.000000  9999.000000  10000.000000
mean       6.651300      5.030300     1.617524     0.025303      5.005000
std        1.370947      2.804098     2.972068     0.322925      2.809727
min        3.000000      0.000000     0.000000     0.000000      0.000000
25%        6.000000      4.000000     0.000000     0.000000      4.000000
50%        7.000000      6.000000     0.000000     0.000000      6.000000
75%        8.000000      7.000000     1.000000     0.000000      7.000000
max       16.000000     13.000000    12.000000     8.000000     13.000000

        num_special   num_vowels  num_syllables
count  10000.000000  9998.000000   10000.000000
mean       0.003400     1.805961       1.606600
std        0.119958     1.242344       0.681383
min        0.000000     0.000000       0.000000
25%        0.000000     1.000000       1.000000
50%        0.000000     2.000000       2.000000
75%        0.000000     3.000000       2.000000
max        6.000000    10.000000       6.000000
```

2. Descriptive analysis - statistical measures of data (Dispersion)

```
7] import pandas as pd

   # Load your dataset
   df = pd.read_csv('passwordscom.csv')  # Replace 'your_dataset.csv' with the path to your dataset

   # Filter out non-numeric columns
   numeric_columns = df.select_dtypes(include=['number'])

   # Descriptive analysis - statistical measures of data (Dispersion)
   dispersion_measures = {
       'Variance': numeric_columns.var(),
       'Standard Deviation': numeric_columns.std(),
       'Range': numeric_columns.max() - numeric_columns.min(),
       'Interquartile Range (IQR)': numeric_columns.quantile(0.75) - numeric_columns.quantile(0.25)
   }

   # Convert dictionary to DataFrame for better visualization
   dispersion_df = pd.DataFrame(dispersion_measures)

   # Display statistical measures of dispersion
   print("Statistical measures of data (Dispersion):\n", dispersion_df)
```

```
Statistical measures of data (Dispersion):
                Variance  Standard Deviation  Range  Interquartile Range (IQR)
length          1.879496            1.370947   13.0                        2.0
num_chars       7.862968            2.804098   13.0                        3.0
num_digits      8.833191            2.972068   12.0                        1.0
num_upper       0.104281            0.322925    8.0                        0.0
num_lower       7.894564            2.809727   13.0                        3.0
num_special     0.014390            0.119958    6.0                        0.0
num_vowels      1.543419            1.242344   10.0                        2.0
num_syllables   0.464283            0.681383    6.0                        1.0
```

3. Correlation between attributes-

```
[18] import pandas as pd

     # Load your dataset
     df = pd.read_csv('passwordscom.csv')  # Replace 'your_dataset.csv' with the path to your dataset

     # Compute the correlation matrix
     correlation_matrix = df.corr()

     # Display the correlation matrix
     print("Correlation between attributes:\n", correlation_matrix)
```

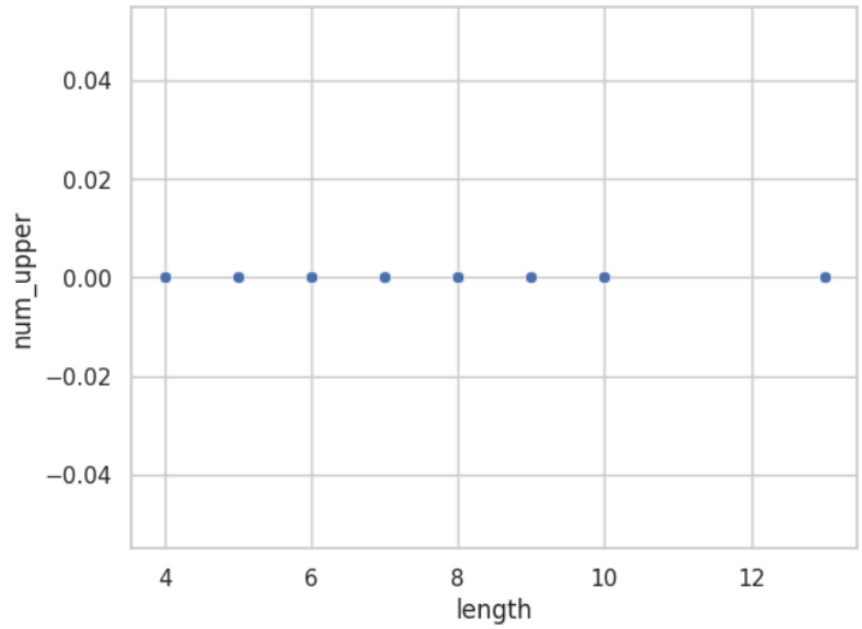
```
Correlation between attributes:
                length  num_chars  num_digits  num_upper  num_lower  \
length        1.000000   0.116878    0.351494   0.019701   0.114379
num_chars     0.116878   1.000000   -0.887967   0.040134   0.993385
num_digits    0.351494  -0.887967    1.000000  -0.028689  -0.882889
num_upper     0.019701   0.040134   -0.028689   1.000000  -0.074873
num_lower     0.114379   0.993385   -0.882889  -0.074873   1.000000
num_special  -0.006777  -0.041931   -0.003926  -0.002221  -0.041592
num_vowels    0.057753   0.803433   -0.729924  -0.055825   0.808240
num_syllables 0.234272   0.590230   -0.446859   0.012518   0.587610

               num_special  num_vowels  num_syllables
length           -0.006777    0.057753       0.234272
num_chars        -0.041931    0.803433       0.590230
num_digits       -0.003926   -0.729924      -0.446859
num_upper        -0.002221   -0.055825       0.012518
num_lower        -0.041592    0.808240       0.587610
num_special       1.000000   -0.036512      -0.049706
num_vowels       -0.036512    1.000000       0.506381
num_syllables    -0.049706    0.506381       1.000000
<ipython-input-18-ee192f826782>:7: FutureWarning: The default value of
  correlation_matrix = df.corr()
```

Visualization-

1.Scatterplot-

```
#plot scatter plot using seaborn
import seaborn as sns
sns.set(style='whitegrid')
x_axis = df['length'].head(50)
y_axis = df['num_upper'].head(50)
sns.scatterplot(x=x_axis, y=y_axis, data=df)
```

```
<Axes: xlabel='length', ylabel='num_upper'>
```



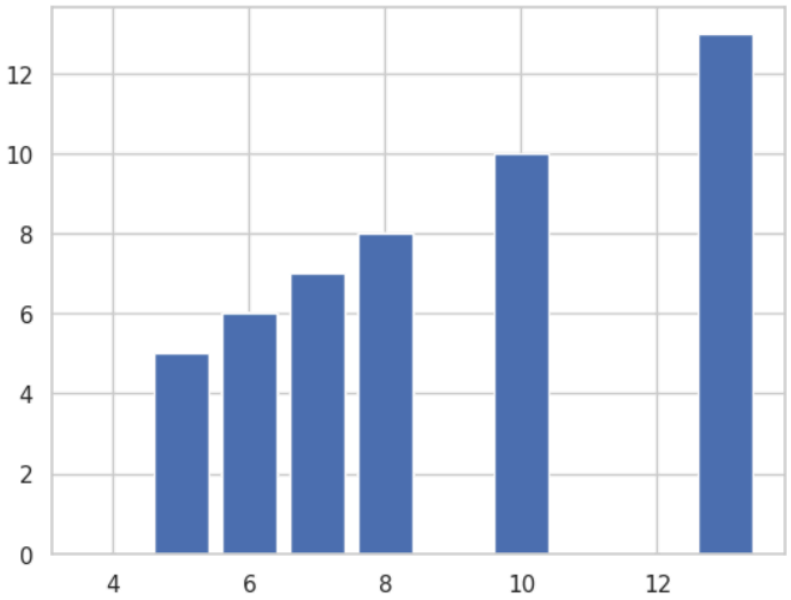
Inference: In a scatterplot of "num_upper" and password length attributes in a dataset of common passwords, the dispersion of points indicates the relationship between password length and the count of uppercase characters, offering insights into password complexity and potential patterns in password creation strategies.

2.Bar Graph using contingency table:

```
[ ]  #setting DataFrame
     df = pd.DataFrame(df)

     #setting the variables
     x_axis = df['length']
     y_axis = df['num_chars']

     #Figure size
     from matplotlib import pyplot as plt
     fig = plt.figure(figsize=(50,7))
```

```
<Figure size 5000x700 with 0 Axes>
```

```
[ ]  # create bar graph, contingency table using any 2 features.
     plt.bar(x_axis[0:50], y_axis[0:50])
```

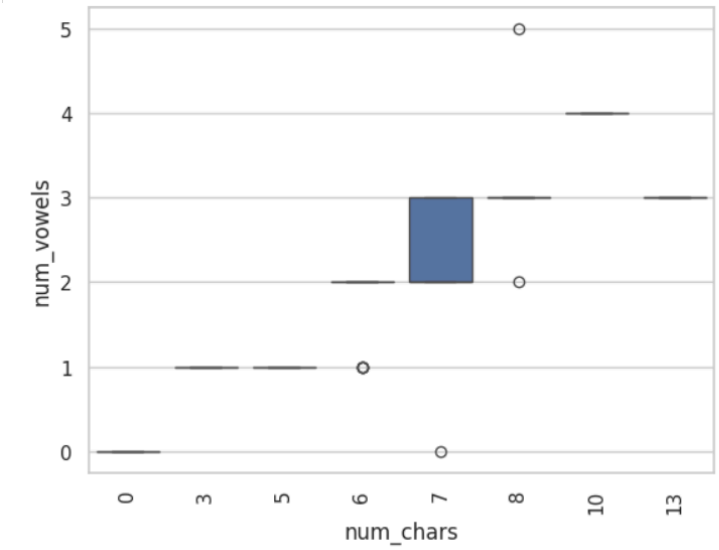


Inference: bar graph derived from a contingency table comparing the length and "num_chars" attributes of common passwords, differences in frequency distributions reveal varying patterns in password composition, aiding in understanding password complexity and potential security implications.

3.Box Plot

```
[23] #plot boxplot using seaborn.
     x_axis = df['num_chars'].head(50)
     y_axis = df['num_vowels'].head(50)
     sns.boxplot(y=y_axis, x=x_axis);
     plt.xticks(rotation = 90)
```

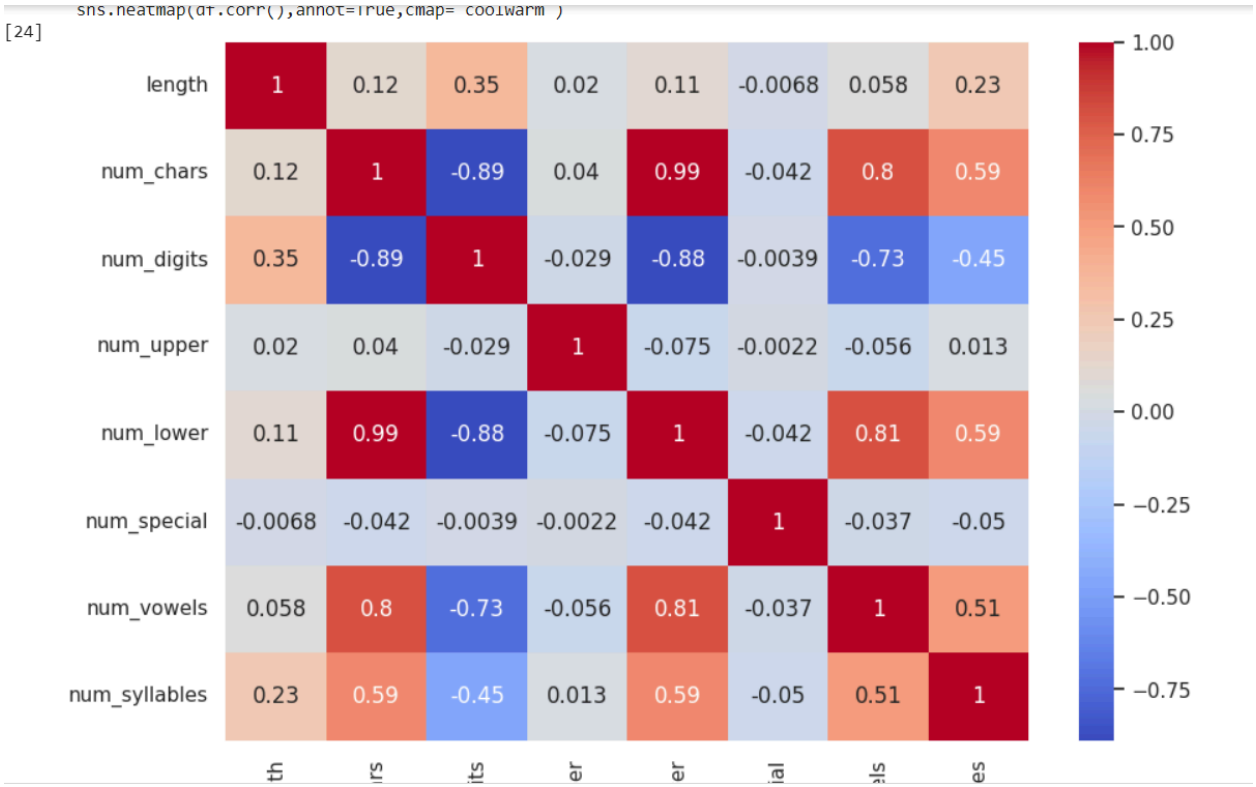
```
([0, 1, 2, 3, 4, 5, 6, 7],
 [Text(0, 0, '0'),
  Text(1, 0, '3'),
  Text(2, 0, '5'),
  Text(3, 0, '6'),
  Text(4, 0, '7'),
  Text(5, 0, '8'),
  Text(6, 0, '10'),
  Text(7, 0, '13')])
```



Inference : box plot of "num_chars" and "num_vowels" in a dataset of common passwords provide insights into the distribution, variability, and potential relationships between password length and vowel count.

4.HeatMap

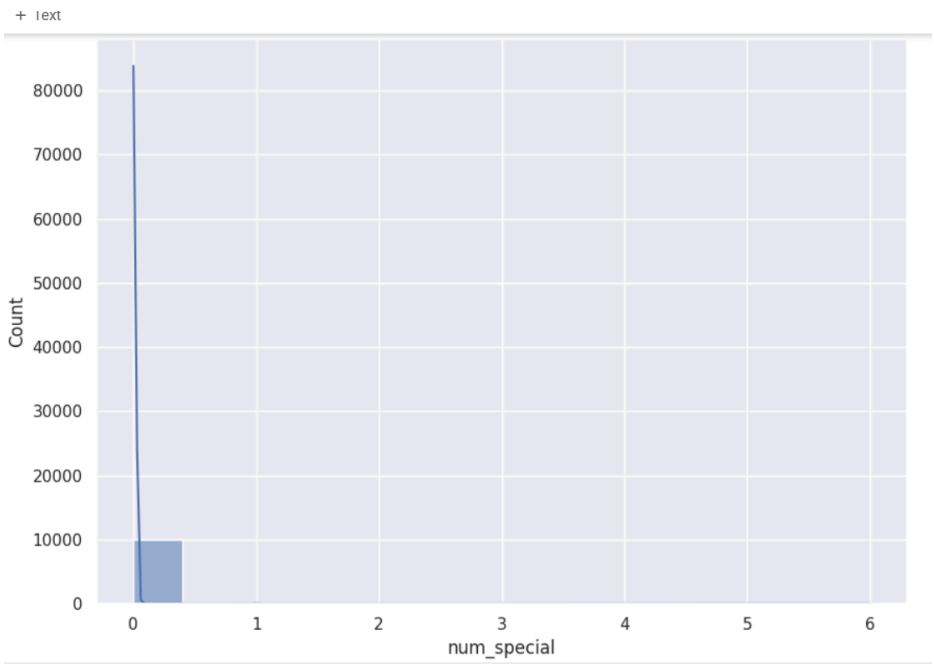
```
#Heatmap
sns.set(rc={"figure.figsize":(10,7)})
sns.heatmap(df.corr(),annot=True,cmap="coolwarm")
plt.show()
```



Inference: heatmap displaying the relationship between various attributes of common passwords, darker shades indicate stronger correlations, revealing potential patterns or dependencies between different password characteristics, aiding in understanding password usage trends and potential security vulnerabilities.

5. Histogram

```
[25] #create histogram
     import pandas as pd
     import seaborn as sns
     from matplotlib import pyplot as plt
     sns.histplot(df['num_special'],kde=True)
     plt.show()
```



Inference: From the histogram of the "num_special" attribute in a dataset of common passwords, the distribution of special characters indicates the prevalence of certain characters in passwords, potentially reflecting common patterns or preferences in password creation, and highlighting the importance of considering special character usage in password security policies.

Conclusion: We successfully understood and implemented exploratory data analysis and different data visualization techniques.