LAB MANUAL
PART A
(PART A: TO BE REFFERED BY STUDENTS)

Experiment No-04

A.1 Aim:
Exploratory Data Analysis and visualization of Social Media Data for business.

Lab Objective	To understand the fundamental concepts of social media networks
Lab Outcome	Collect, monitor, store and track social media data

PART B
(PART B: TO BE COMPLETED BY STUDENTS)

(Students must submit the soft copy as per following segments within two hours of the practical. The soft copy must be uploaded on the Blackboard or emailed to the concerned lab in charge faculties at the end of the practical in case the there is no Black board access available)

Roll. No. C36	Name: Sanskruti Kadam
Class BE-C	Batch: C2
Date of Experiment:	Date of Submission:
Grade:	

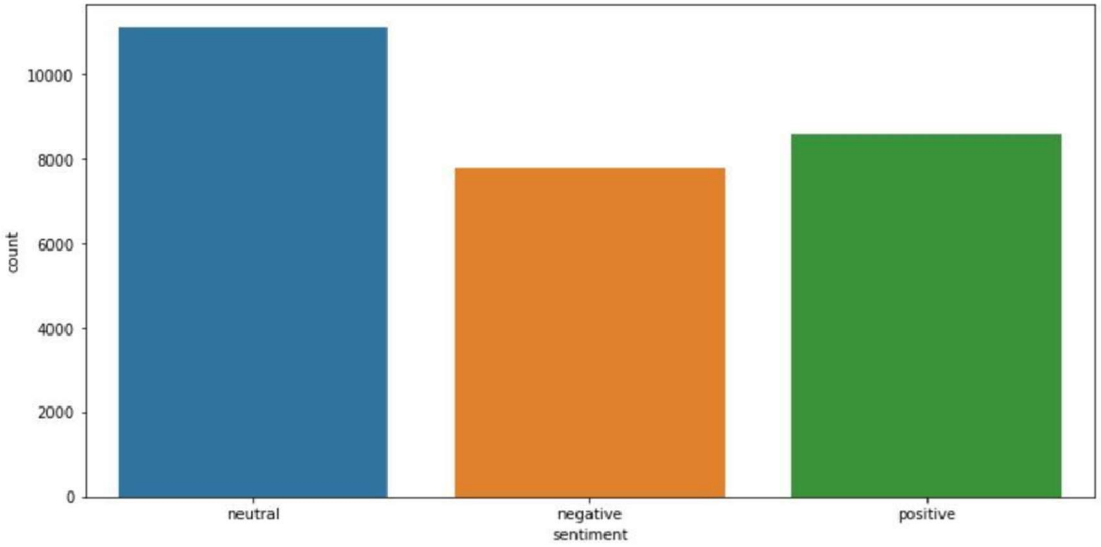
B1. Study the fundamentals of social media platform and implement data cleaning, preprocessing, filtering and storing social media data for business:
(Paste your Search material completed during the 2 hours of practical in the lab here)

- Students need to use the previous social media dataset to perform exploratory data analysis and visualization.

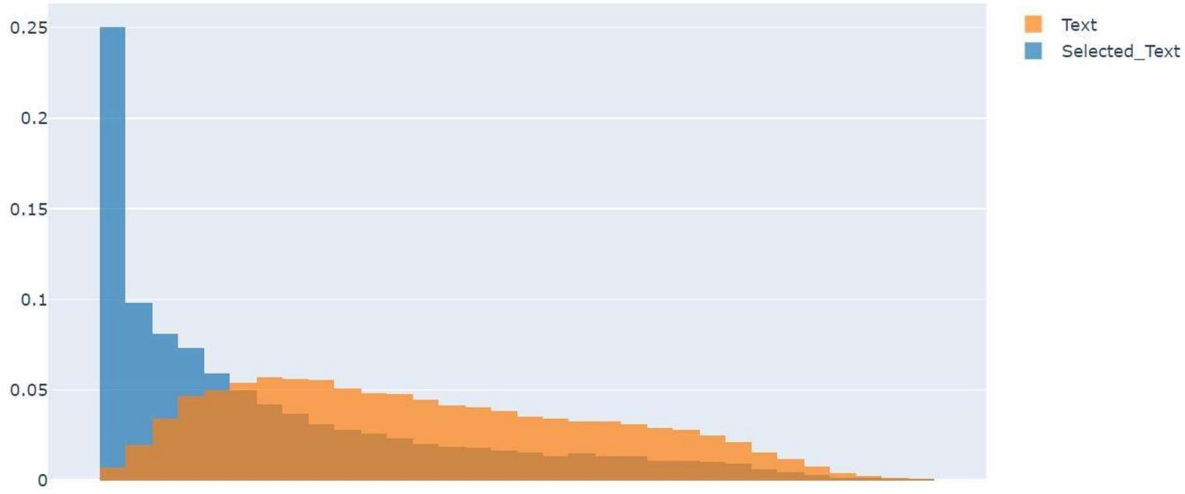
B.2 Input and Output:
(Command and its output)

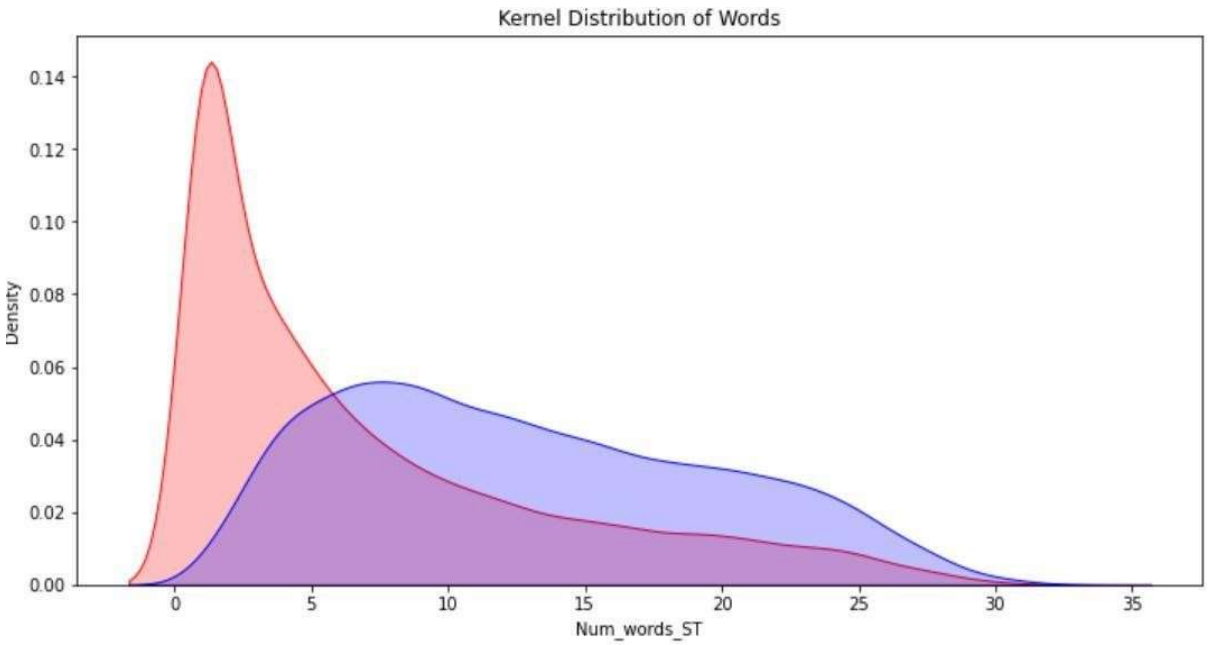
```
sns.countplot(x='sentiment', data=df_train)
```

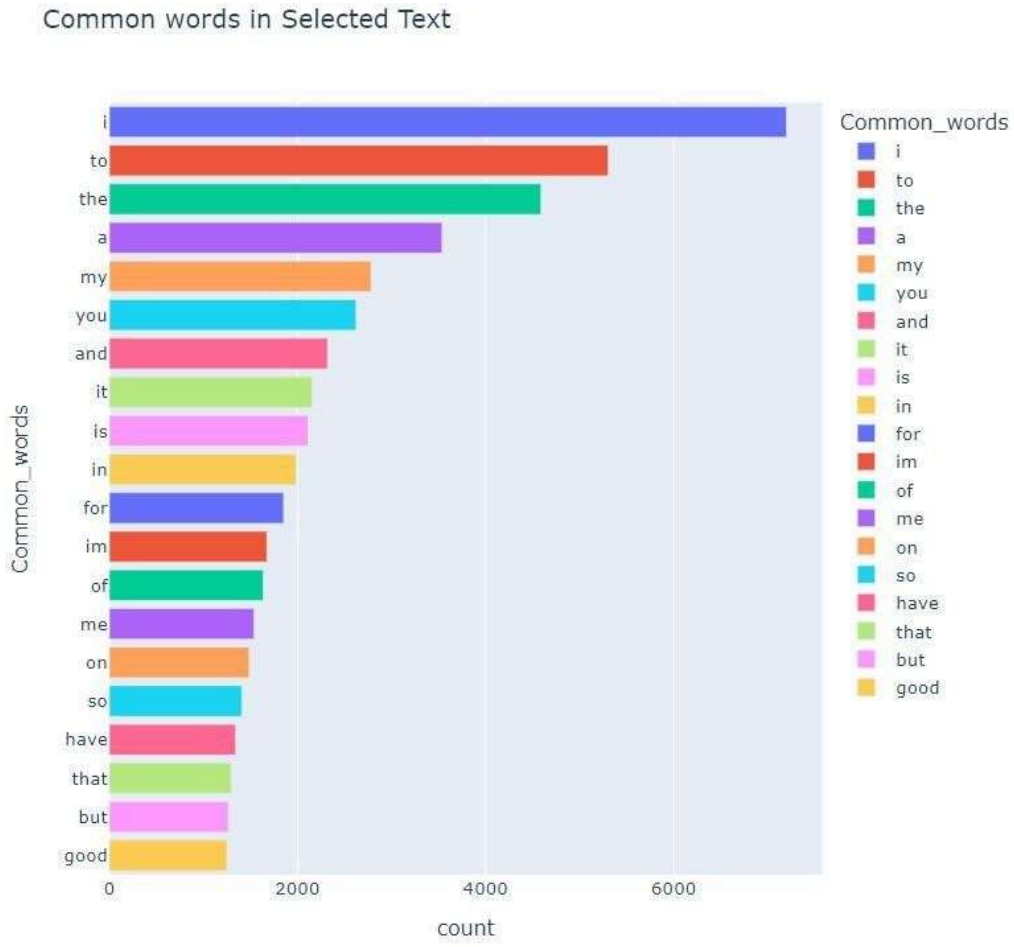
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8b0457b0a0>
```

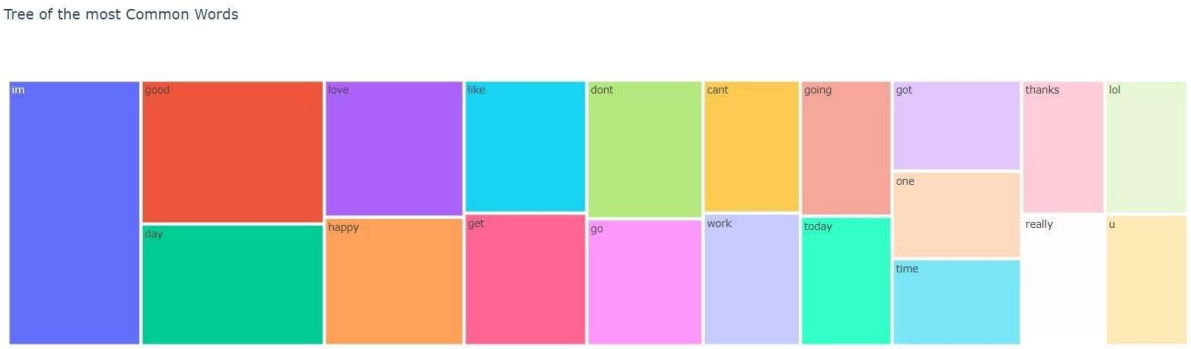


Distribution of Number of words




Kernel Distribution of Words
Density
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
5
10
15
20
25
30
35
Num_words_ST


Common words in Selected Text
Common_words
i
to
the
a
my
you
and
it
is
in
for
im
of
me
on
so
have
that
but
good
0
2000
4000
6000
count
Common_words
i
to
the
a
my
you
and
it
is
in
for
im
of
me
on
so
have
that
but
good



B.3 Observations and learning:

(Students are expected to comment on the output obtained with clear observations and learning for each task/ sub part assigned)

Distribution of the data: Visualization can help to identify the distribution of the data. For example, a histogram can be used to show the frequency distribution of a variable. Observations can be made about the shape of the distribution, the presence of outliers, and whether the distribution is symmetric or skewed.

Correlations between variables: Visualization can help to identify correlations between variables. For example, a scatter plot can be used to show the relationship between two variables. Observations can be made about the direction and strength of the correlation, the presence of any outliers or influential points, and whether the correlation is linear or nonlinear.

Trends over time: Visualization can help to identify trends over time. For example, a line chart can be used to show the trend of a variable over a period of time. Observations can be made about the direction and magnitude of the trend, whether there are any seasonal patterns or fluctuations, and whether the trend is linear or nonlinear.

Group comparisons: Visualization can help to compare groups of data. For example, a bar chart can be used to compare the mean values of a variable for different groups. Observations can be made about the differences in mean values, the presence of any outliers or influential points, and whether there are any significant differences between the groups.

Spatial patterns: Visualization can help to identify spatial patterns in the data. For example, a map can be used to show the distribution of a variable across different geographic locations. Observations can be made about any spatial patterns or clusters, the presence of any outliers, and whether there are any relationships between the variable and geographic features.

Overall, visualization can help to reveal patterns and insights in the data that may not be immediately apparent from raw data or statistical analyses. Observations made during a data visualization experiment can provide valuable insights for further analysis and decision-making.

B.4 Conclusion:

(Students must write the conclusion as per the attainment of individual outcome listed above and learning/observation noted in section B.3)
Hence data visualization for exploratory data analysis was performed for twitter data.

B.5 Question of Curiosity
(To be answered by student based on the practical performed and learning/observations)
Q1. What is EDA? Explain its importance?
Ans:- EDA, or exploratory data analysis, is a critical step in the data analysis process that involves using visualizations and statistical methods to explore and summarize data. EDA allows analysts to gain a deeper understanding of the data and identify patterns, trends, and relationships that may not be immediately apparent through raw data alone.

The importance of EDA lies in its ability to provide insights and inform further analysis. By exploring and summarizing the data, EDA can help analysts to:

Identify patterns and trends: EDA can help analysts to identify patterns and trends within the data, such as seasonality or long-term trends. This information can be used to make informed decisions and predictions.

Detect outliers and anomalies: EDA can help analysts to identify outliers and anomalies within the data that may indicate errors or data quality issues. By detecting and addressing these issues, analysts can improve the accuracy and reliability of the data.

Understand the distribution of data: EDA can help analysts to understand the distribution of data, such as its mean, median, and standard deviation. This information is important for selecting appropriate statistical methods and models.

Explore relationships between variables: EDA can help analysts to explore relationships between variables, such as correlations or causations. This information is important for identifying potential predictors and variables that may be relevant for further analysis.

Overall, EDA is an important first step in the data analysis process that provides a foundation for further analysis and modeling. By gaining a deeper understanding of the data through EDA, analysts can make more informed decisions and draw more accurate conclusions from the data.

Q2. What is the importance of visualization?
Ans:-Visualization is an important tool in data analysis and communication. Here are some reasons why visualization is important:

Provides insights: Visualization allows us to see patterns, trends, and relationships within the data that may not be immediately apparent through raw data alone. It helps us to better understand the data and identify important features that may inform further analysis.

Improves communication: Visualization is an effective way to communicate complex data to others, such as stakeholders, clients, or colleagues. By presenting data in a clear and visually appealing way, we can effectively convey our findings and recommendations.

Facilitates decision-making: Visualization helps us to make better decisions by presenting data in a way that is easy to understand and interpret. By visualizing data, we can quickly identify trends and outliers, compare different groups or variables, and identify potential relationships or correlations.

Encourages exploration: Visualization encourages exploration and discovery by allowing us to interact with the data and explore different scenarios or hypotheses. It helps us to ask better questions and uncover new insights.

Supports data quality assurance: Visualization helps us to identify data quality issues, such as outliers or missing values, that may require further investigation and cleaning. By detecting and addressing these issues, we can improve the accuracy and reliability of the data.

Overall, visualization is an essential tool in data analysis and communication. By presenting data in a clear and visually appealing way, we can gain insights, make better decisions, and communicate our findings more effectively.

Q3. Explain the steps involved in EDA?
Ans:- The steps involved in EDA, or exploratory data analysis, typically include the following:

Data collection: The first step in EDA is to collect the data that will be analyzed. This may involve collecting data from various sources, such as databases, surveys, or experiments.

Data cleaning: Once the data has been collected, the next step is to clean and prepare the data for analysis. This may involve handling missing values, removing duplicates, correcting data format errors, and resolving data conflicts.

Data exploration: The next step is to explore the data using visualizations and statistical methods. This may involve creating histograms, scatterplots, boxplots, or other visualizations to identify patterns and trends within the data.

Data transformation: If necessary, the data may be transformed to improve its distribution or reduce the effects of outliers. This may involve applying logarithmic or exponential transformations or using other techniques to normalize the data.

Data reduction: If the data is large or complex, it may be necessary to reduce its size or dimensionality using techniques such as principal component analysis or clustering.

Statistical modeling: Once the data has been explored and transformed, statistical models can be built to test hypotheses or make predictions. This may involve regression analysis, classification, or clustering.

Model validation: Finally, the models and results are validated using various metrics and tests to ensure that they are accurate and reliable. This may involve cross-validation, hypothesis testing, or other statistical tests.

Overall, EDA is an iterative process that involves collecting, cleaning, exploring, and modeling data to gain insights and inform further analysis. The specific steps involved may vary depending on the nature of the data and the research question.

Q4. What is the Difference between Univariate, Bivariate, and Multivariate analysis? Ans:- Univariate, bivariate, and multivariate analysis are three types of statistical analysis that are used to analyze data.

Univariate analysis involves analyzing a single variable in isolation. The goal of univariate analysis is to describe the distribution of the variable and identify any patterns or trends that may be present. Common techniques used in univariate analysis include measures of central tendency (such as the mean or median) and measures of variability (such as the standard deviation or range).

Bivariate analysis involves analyzing the relationship between two variables. The goal of bivariate analysis is to explore the relationship between the variables and identify any patterns or trends that may be present. Common techniques used in bivariate analysis include correlation analysis and regression analysis.

Multivariate analysis involves analyzing three or more variables simultaneously. The goal of multivariate analysis is to explore the relationships between the variables and identify any patterns or trends that may be present. Common techniques used in multivariate analysis include principal component analysis and factor analysis.

In summary, the main difference between univariate, bivariate, and multivariate analysis is the number of variables that are analyzed. Univariate analysis focuses on a single variable, bivariate analysis focuses on the relationship between two variables, and multivariate analysis focuses on the relationships between three or more variables. Each type of analysis can provide valuable insights into the data and help researchers to better understand the patterns and trends that are present.

Q5. During the data preprocessing step, how should one treat missing/null values? How will you deal with them?
Ans:- Missing or null values are a common issue in datasets and can cause problems during data analysis if not handled appropriately. Here are some common techniques for treating missing or null values during data preprocessing:

Deletion: One option is to simply delete any rows or columns that contain missing values. This can be done using listwise deletion (deleting entire rows with missing values) or pairwise deletion (deleting only the values that are missing). However, this can result in loss of information and may not be appropriate for all datasets.

Imputation: Another option is to impute, or fill in, the missing values with an estimated value. Common imputation techniques include mean imputation (replacing missing values with the mean of the available data), regression imputation (using regression analysis to predict missing values), or multiple imputation (creating multiple imputations based on the available data and combining them). Imputation can help to preserve the sample size and reduce bias in the data.

Domain knowledge: Depending on the nature of the data and the reason for the missing values, domain knowledge can be used to estimate or infer the missing values. For example, if a survey respondent failed to answer a question about their age, their age could potentially be inferred from other demographic information or external sources.

Ignore: In some cases, it may be appropriate to ignore missing values if they represent a very small proportion of the overall dataset and do not have a significant impact on the analysis. However, this approach should be used with caution and it is important to assess the potential impact of ignoring missing values on the analysis results.

Overall, the choice of how to treat missing or null values will depend on the nature of the data and the research question. It is important to carefully consider the potential impact of each approach on the analysis results and choose the approach that is most appropriate for the specific situation.