Sample Viva Questions of Data Visualisation

UNIT I

1. What is analytics and why is it important?

Answer:
Analytics is the science of examining raw data to draw conclusions and support decision-making. It is important because it helps organizations make data-driven decisions, uncover patterns, and improve performance. Analytics is widely used in fields like marketing, finance, healthcare, and more.

2. Define and differentiate between features and labels in a dataset.

Answer:
Features (or independent variables) are input data used to predict outcomes. The label (or dependent variable) is the output we want to predict. For example, in a house price prediction model, size, location, and number of rooms are features, while the house price is the label.

3. What are the main stages of the analytics process model?

Answer:
The stages are:

1. Problem Definition
2. Data Collection
3. Data Cleaning
4. Exploratory Data Analysis
5. Modeling
6. Evaluation
7. Deployment
   Each step ensures a systematic approach to solving data problems.

4. How does analytics apply to different professional domains?

Answer:
In marketing, it helps with customer segmentation; in finance, with risk modeling; in healthcare, with diagnostics; in HR, with employee performance analysis. Every domain uses analytics to improve efficiency and decision-making using data insights.

5. What are the requirements for building an analytical model?

Answer:
You need high-quality and relevant data, clearly defined objectives, appropriate algorithms, domain knowledge, and proper evaluation metrics. Clean data and feature selection are critical for building accurate and reliable models.

6. Name some common data sources for analytics.

Answer:
Data can come from databases (SQL), APIs, sensors, surveys, web scraping, social media, and public

datasets (e.g., government or research portals). Choosing the right source depends on the problem and required data type.

7. What is sampling and why is it used?

Answer:
Sampling involves selecting a small subset from a large population to analyze. It saves time and resources while allowing accurate estimation of population characteristics, provided the sample is representative and unbiased.

8. Explain the concept of sampling distribution.

Answer:
Sampling distribution is the distribution of a statistic (like mean) calculated from many samples of a population. It shows how the statistic would vary across different samples, and it's used to make inferences about the population.

9. What are the types of data elements in analytics?

Answer:
Data elements are classified as:

- **Numeric** (continuous or discrete)
- **Categorical** (nominal or ordinal)
- **Boolean** (True/False)
- **Date/Time**
  These help in deciding how to handle and model the data.

10. What causes missing values and how are they handled?

Answer:
Missing values may occur due to sensor failure, data entry errors, or incomplete surveys. They are handled by deleting rows/columns or using imputation techniques like mean, median, mode, or model-based filling (e.g., KNN imputation).

11. What is an outlier and how do we treat it?

Answer:
An outlier is a data point significantly different from others. It can distort analysis. Methods like z-score or IQR help detect outliers. Treatment involves removal, transformation (e.g., log), or capping values to reduce their impact.

12. Explain the concept of standardization with examples.

Answer:
Standardization scales data for uniformity.

- **Min-Max Scaling** brings values to a 0–1 range.
- **Z-score Normalization** centers data around a mean of 0 and standard deviation of 1.
  These techniques are useful for machine learning algorithms like KNN or SVM.

13. What is categorization in data processing?

Answer:
Categorization involves converting continuous variables into groups or categories. For example, ages can be categorized into 'Teen', 'Adult', and 'Senior'. It helps simplify data and make models more interpretable.

14. What is segmentation and why is it useful?

Answer:
Segmentation is the process of dividing a dataset into groups with similar characteristics. Techniques like K-means or hierarchical clustering are used. It's useful in marketing (e.g., customer segments), healthcare, and fraud detection.

15. Why is data preprocessing important in analytics?

Answer:
Data preprocessing ensures the data is clean, consistent, and suitable for analysis. It improves model performance, reduces errors, handles missing values, standardizes scales, and makes features more meaningful. It's a crucial step before modeling.

Unit II

1. What is statistical hypothesis testing?

Answer:
Statistical hypothesis testing is a method to decide whether a statement about a population is true based on sample data. It involves two hypotheses:

- **Null hypothesis (H_o)**: No effect or difference.
- **Alternative hypothesis (H_1)**: A real effect exists.
  Tests like t-test, z-test, and chi-square are used to make conclusions about H_o.

2. What is a p-value and how is it interpreted?

Answer:
The p-value is the probability of observing the data (or more extreme) assuming the null hypothesis is true.

- A **small p-value (≤ 0.05)** indicates strong evidence against H_o → reject H_o.
- A **large p-value (> 0.05)** indicates weak evidence → fail to reject H_o.
  It helps in decision-making in hypothesis testing.

3. Define confidence interval. What does a 95% confidence interval mean?

Answer:
A confidence interval gives a range of values within which the population parameter is expected to lie.
A **95% confidence interval** means that if we take 100 samples, approximately 95 of them will contain the true population value.
It provides both estimate and uncertainty.

4. What is correlation?

Answer:
Correlation is a statistical measure that describes the strength and direction of a relationship between two variables.

- **Positive correlation**: As one variable increases, so does the other.
- **Negative correlation**: One increases, the other decreases.
  The correlation coefficient ranges from -1 to 1.

5. What is Pearson's correlation coefficient?

Answer:
Pearson's correlation measures the **linear relationship** between two continuous variables.

- **+1**: Perfect positive linear relationship
- **-1**: Perfect negative linear relationship
- **0**: No linear relationship
  It assumes normally distributed data.

6. Explain Simpson's Paradox with an example.

Answer:
Simpson's Paradox occurs when a trend appears in different groups but reverses when the groups are combined.
For example, a treatment may appear effective in separate age groups but look ineffective when the data is combined, due to unequal group sizes or confounding factors.

7. What is the difference between correlation and causation?

Answer:
Correlation means two variables are related, but it doesn't prove that one causes the other.
Causation means one variable directly affects another.
For example, ice cream sales and drowning incidents may be correlated due to a common cause (summer), not causation.

8. What are some common correlational caveats?

Answer:

- **Outliers** can distort correlation.
- **Non-linear relationships** may have zero correlation even if they're related.
- **Third variables** (confounders) can mislead results.
- **Simpson's Paradox** may hide actual relationships.

9. What is ANOVA and when is it used?

Answer:
ANOVA (Analysis of Variance) is used to compare means of three or more groups.
It tests if at least one group mean is significantly different from others.
If the p-value is low, we reject the null hypothesis that all group means are equal.

10. What are the assumptions of ANOVA?

Answer:

1. The data is normally distributed.
2. The variances across groups are equal (homogeneity of variance).
3. Observations are independent.
   Violating these assumptions may affect the validity of the results.

11. What is the difference between t-test and ANOVA?

Answer:

- **T-test** compares the means of **two groups**.
- **ANOVA** compares the means of **three or more groups**.
  Both test whether group differences are statistically significant.

12. What is Type I and Type II error in hypothesis testing?

Answer:

- **Type I Error (False Positive)**: Rejecting a true null hypothesis.
- **Type II Error (False Negative)**: Failing to reject a false null hypothesis.
  Controlling these errors is important for reliable hypothesis testing.

13. What does statistical significance mean?

Answer:
Statistical significance means that the observed result is unlikely due to chance, under the null hypothesis.
It is typically determined by comparing the **p-value to a significance level ($\alpha = 0.05$)**.
If $p \leq \alpha$, the result is considered statistically significant.

14. When do we use one-tailed and two-tailed tests?

Answer:

- **One-tailed test** is used when the direction of effect is known (e.g., $A > B$).
- **Two-tailed test** is used when checking for any difference (e.g., $A \neq B$).
  Choosing the correct test affects how p-values are interpreted.

15. Can two variables have a high correlation but no causal relationship?

Answer:
Yes. Two variables can move together due to a third hidden factor or coincidence.
For example, ice cream sales and shark attacks both increase in summer. They are correlated due to a common cause (season), not because one causes the other.

Unit III

1. What is data visualization and why is it important?

Answer:
Data visualization is the graphical representation of information and data using plots and charts. It

helps in understanding trends, patterns, and outliers in data easily. Visualization is essential for communicating insights clearly and quickly, especially to non-technical stakeholders.

2. What is Matplotlib and what is it used for in Python?

Answer:
Matplotlib is a popular Python library used for creating static, animated, and interactive plots. It provides functions to create line plots, bar charts, scatter plots, and more. It's highly customizable and is the foundation for other libraries like Seaborn.

3. How do you create a simple line graph using Matplotlib?

Answer:

```
import matplotlib.pyplot as plt

x = [1, 2, 3, 4]

y = [10, 20, 25, 30]

plt.plot(x, y)

plt.title("Line Graph")

plt.show()
```

This plots a basic line connecting the points (1,10), (2,20), etc.

4. What is a bar chart and when is it used?

Answer:
A bar chart represents categorical data with rectangular bars. The height or length of each bar shows the value of the category. It is useful for comparing discrete categories like sales by region or population by country.

5. How do you plot a pie chart using Matplotlib?

Answer:

```
import matplotlib.pyplot as plt

labels = ['A', 'B', 'C']

sizes = [30, 40, 30]

plt.pie(sizes, labels=labels, autopct='%1.1f%%')

plt.title("Pie Chart")

plt.show()
```

A pie chart is useful for showing percentage or part-to-whole relationships.

6. What is a scatter plot and when is it used?

Answer:
A scatter plot shows the relationship between two continuous variables. Each point represents a data pair (x, y). It is used to detect correlations, clusters, and outliers in data.

7. How do you create multiple plots in one figure?

Answer:

```
plt.subplot(1, 2, 1)  # 1 row, 2 columns, 1st plot
plt.plot([1, 2], [3, 4])
plt.subplot(1, 2, 2)  # 2nd plot
plt.plot([1, 2], [4, 3])
plt.show()
```

The subplot() function allows you to display multiple plots in a grid layout.

8. How do you change the size of a figure in Matplotlib?

Answer:

```
plt.figure(figsize=(8, 4))
plt.plot([1, 2, 3], [4, 5, 6])
plt.show()
```

figsize=(width, height) sets the dimensions of the plot in inches.

9. What are legends and how are they added to plots?

Answer:
A legend explains the meaning of different plot elements (like lines or colors).

```
plt.plot(x, y, label='Data 1')
plt.legend()
```

Legends are helpful when multiple datasets are plotted on the same figure.

10. What is Seaborn and how is it different from Matplotlib?

Answer:
Seaborn is a Python visualization library built on top of Matplotlib. It provides a higher-level, more attractive interface for complex visualizations like violin plots, heatmaps, and categorical plots. It supports pandas dataframes natively.

11. How do you plot a distribution plot (histogram) using Seaborn?

Answer:

```
import seaborn as sns
sns.displot([10, 20, 20, 30, 40, 50])
plt.show()
```

displot() shows the frequency distribution of a dataset. It can include KDE (density estimation) for smoother curves.

12. What is relplot() used for in Seaborn?

Answer:
relplot() creates relational plots like scatter and line plots, and supports subplots with col and row parameters for facets.

sns.relplot(x="age", y="income", data=df, kind="scatter")

It is ideal for analyzing relationships between two or more variables.

13. What is catplot() and when is it used?

Answer:
catplot() is used for plotting categorical data. It supports various plot types like bar, box, and strip.

sns.catplot(x="gender", y="score", kind="bar", data=df)

It is helpful for comparing statistics across different categories.

14. What is a count plot in Seaborn?

Answer:
A **count plot** shows the count of observations in each category. It is similar to a bar plot but directly counts frequency.

sns.countplot(x="gender", data=df)

It is useful for categorical frequency analysis.

15. How can you apply color palettes in Seaborn plots?

Answer:
Seaborn provides built-in color palettes for better aesthetics.

sns.set_palette("pastel")

sns.barplot(x="day", y="sales", data=df)

Popular palettes include "deep", "muted", "pastel", "bright", etc., improving plot readability and visual appeal.

Unit IV

1. What is Tkinter in Python?

Answer:
Tkinter is Python's standard library for creating GUI (Graphical User Interface) applications. It provides a variety of widgets like buttons, labels, entry boxes, and menus. Tkinter is lightweight and comes pre-installed with Python, making it a popular choice for beginners.

2. How do you create a simple Tkinter window?

Answer:

python

import tkinter as tk

```
root = tk.Tk()

root.title("My Window")

root.geometry("300x200")

root.mainloop()
```

This creates a basic GUI window titled "My Window" with a size of 300x200 pixels.

3. What are widgets in Tkinter?

Answer:
Widgets are GUI components like buttons, labels, entry fields, radio buttons, etc. They are the building blocks of a GUI application. Widgets are created using Tkinter classes like Label, Button, Entry, etc., and are added to the window.

4. How do you create a layout using Tkinter?

Answer:
Tkinter provides three layout managers:

- pack() – simple placement
- grid() – row/column-based layout
- place() – coordinate-based placement
  These are used to position widgets within a window.

5. How do you create a button in Tkinter?

Answer:

```
import tkinter as tk

def say_hello():

   print("Hello!")

root = tk.Tk()

btn = tk.Button(root, text="Click Me", command=say_hello)

btn.pack()

root.mainloop()
```

The button displays "Click Me" and calls the say_hello() function on click.

6. What is the use of Checkbuttons and Radiobuttons in Tkinter?

Answer:

- **Checkbutton** allows selecting multiple options.
- **Radiobutton** allows selecting one option from a group.
  They are useful for capturing user preferences or choices in GUI forms.

7. How do you create a listbox in Tkinter?

Answer:

```
listbox = tk.Listbox(root)

listbox.insert(1, "Item 1")

listbox.insert(2, "Item 2")

listbox.pack()
```

A listbox allows the user to select one or more items from a list.

8. What are Menus in Tkinter and how do you create them?

Answer:
Menus provide a drop-down interface for navigation (like File, Edit, Help).

```
menu = tk.Menu(root)

root.config(menu=menu)

file_menu = tk.Menu(menu)

menu.add_cascade(label="File", menu=file_menu)

file_menu.add_command(label="Exit", command=root.quit)
```

Menus help organize commands in a user-friendly way.

9. What are dialog boxes in Tkinter?

Answer:
Dialog boxes are pop-up windows used to interact with the user. Tkinter provides built-in dialog boxes for file selection, alerts, and input using tkinter.messagebox, filedialog, etc.
Example: messagebox.showinfo("Title", "Message").

10. What is a database and why is it used in Python programs?

Answer:
A database is an organized collection of data. In Python, databases are used to store, retrieve, and manipulate data persistently. SQLite is commonly used due to its simplicity and being built into Python via the sqlite3 module.

11. How do you connect a Python program to a database?

Answer:

```
import sqlite3

conn = sqlite3.connect('data.db')

cursor = conn.cursor()
```

This creates or connects to a SQLite database named data.db.

12. How do you create a table in SQLite using Python?

Answer:

```
cursor.execute("CREATE TABLE IF NOT EXISTS users (id INTEGER, name TEXT)")
conn.commit()
```

This command creates a table users with two fields: id and name.

13. How do you insert data into a database table using Python?

Answer:

```
cursor.execute("INSERT INTO users (id, name) VALUES (?, ?)", (1, 'Alice'))
conn.commit()
```

The ? placeholders prevent SQL injection and allow data binding.

14. How do you retrieve data from a database using Python?

Answer:

```
cursor.execute("SELECT * FROM users")
rows = cursor.fetchall()
for row in rows:
    print(row)
```

This fetches and prints all records from the users table.

15. How do you update and delete data from a table in Python?

Answer:

python

```
cursor.execute("UPDATE users SET name = 'Bob' WHERE id = 1")
cursor.execute("DELETE FROM users WHERE id = 1")
conn.commit()
```

Use SQL UPDATE and DELETE statements to modify or remove records from the table.