1. **What is data visualisation and why is it important in data analysis?**

Data visualisation is the graphical representation of data to identify patterns, trends, and insights. It is important because it makes complex data more accessible, helping analysts and decision-makers quickly understand information, communicate findings effectively, and identify anomalies or relationships within the data.

2. **Briefly describe the difference between univariate and bivariate analysis in the context of data visualisation.**

Univariate analysis examines a single variable's distribution, often visualised using histograms or bar charts. Bivariate analysis, on the other hand, explores the relationship between two variables, typically visualised using scatter plots or line charts to identify correlations or trends between them.

3. **What type of variables is a bar chart typically used to visualise? Provide an example.**

A bar chart is used to visualise categorical variables. For example, a bar chart can be used to show the distribution of different car brands sold in a dealership, where each bar represents a specific brand and its frequency of sales.

4. **Explain the difference between a simple bar chart, a multiple (or grouped) bar chart, and a stacked bar chart.**

A simple bar chart displays a single category with bars representing the frequency of occurrences. A multiple bar chart compares different groups across the same categories, with bars grouped side by side. A stacked bar chart displays the total value for each category, with segments stacked within each bar to show the contribution of subcategories.

5. **For what purpose is a pie chart used in data visualisation? What kind of data is suitable for a pie chart?**

A pie chart is used to show proportions of a whole. It is best suited for categorical data where the values represent parts of a total, such as showing the percentage distribution of market share among different companies.

6. **What type of variables are used to produce a scatter plot? Give an example of how a scatter plot can be used.**

A scatter plot visualises the relationship between two continuous variables. For example, it could be used to plot the correlation between hours studied and exam scores to see if there's a trend between study time and performance.

7. **What information does a histogram display and what type of variable is best suited for a histogram?**

A histogram shows the distribution of a continuous variable by grouping data into bins. It is best used for continuous variables, such as income or age, where you want to see the frequency of different ranges of values.

8. **What is a box plot used for, and what kind of information can be gleaned from it?**

A box plot is used to display the spread and skewness of a continuous variable, showing the median, quartiles, and potential outliers. It helps identify the range, central tendency, and variability of the data.

9. **Where in SPSS can you find the Chart Builder?**

In SPSS, the Chart Builder can be found under the "Graphs" menu. You can select "Chart Builder" from the options to create a variety of visualisations.

10. **What is one advantage of using the "Analyze -> Descriptive Statistics -> Frequencies" option in SPSS over using the Chart Builder?**

The "Frequencies" option in SPSS provides a quick summary of the data, showing the frequency distribution, percentages, and cumulative percentages. This is useful for a quick analysis of categorical data without needing to create visual graphs, and it can also provide a more detailed numerical output alongside visuals.

Skewness and Kurtosis: A Comprehensive Study Guide

Quiz

Answer the following questions in 2-3 sentences each.

Explain what skewness measures in a data distribution and why it's important.

Skewness measures the asymmetry of a data distribution. It is important because it helps understand how values are spread around the mean, indicating if the data is skewed to the left (negative skew), right (positive skew), or is symmetrical.

Describe the characteristics of a negatively skewed distribution. What does this tell you about the mean, median, and mode?

A negatively skewed distribution has a long tail extending to the left on the number line. This means the mean is less than the median, and the mode is greater than both the mean and median.

What is kurtosis, and what aspect of a data distribution does it quantify?

Kurtosis measures the peak or height of a data distribution. It quantifies the extent to which a frequency distribution is peaked compared to a normal distribution

Differentiate between a platykurtic and a leptokurtic distribution. What do these terms indicate about the shape of the curve?

A platykurtic distribution has a lower peak and flatter shape than a normal distribution (kurtosis value < 3). A leptokurtic distribution has a higher peak and sharper shape (kurtosis value > 3).

How does adding significantly higher values than the mean influence the skewness of a distribution?

Adding significantly higher values than the mean increases the mean, resulting in positive skewness, because the majority of extreme values are higher than the mean.

How does adding significantly lower values than the mean influence the skewness of a distribution?

Adding significantly lower values than the mean decreases the mean, resulting in negative skewness, because the majority of extreme values are smaller than the mean.

Explain how skewness can be used in finance or economics.

Skewness in finance can help to show that stock prices are more likely to rise slowly but crash suddenly. This can allow investors to adjust their portfolios to mitigate these risks.

How is kurtosis commonly used in engineering and physical sciences?

Kurtosis can allow engineers to identify data distributions that are subject to extreme values or volatility, such as in signal processing or structural analysis.

Which part of the distribution is skewness most sensitive to?

Skewness is most sensitive to changes in the tails of the distribution.

Which parts of the distribution are kurtosis most sensitive to?

Kurtosis is most sensitive to changes in the centre and shoulders of the distribution.

Glossary of Key Terms

Data Visualisation: The graphical representation of information and data.

SPSS: Statistical Package for the Social Sciences, a software package used for statistical analysis.

Chart Builder: A tool within SPSS used to create various types of graphs and charts.

Variable: A characteristic or attribute that can be measured or counted (e.g., car height, car body type).

Categorical Variable: A variable that represents categories or groups (e.g., car body type, engine type).

Numerical Variable: A variable that represents numerical values (e.g., car height, car width).

Univariate Analysis: Analysing a single variable at a time.

Bivariate Analysis: Analysing the relationship between two variables.

Bar Chart: A chart that uses bars to represent the frequency or count of categorical data.

Pie Chart: A circular chart that shows the proportion of different categories within a whole.

Scatter Plot: A graph that plots two numerical variables against each other to visualise their relationship.

Histogram: A chart that displays the distribution of a numerical variable by grouping data into bins.

Box Plot: A chart that displays the distribution of a numerical variable, showing the median, quartiles, and outliers.

Frequency Distribution: A summary of how often each value (or set of values) occurs in a dataset.

Outlier: A data point that is significantly different from other data points in a dataset.

Mean: The average of a set of numbers.

Standard Deviation: A measure of the spread of data around the mean.

Gridlines: Lines added to a graph to make it easier to read the values.

Data Labels: Labels on a chart that show the actual values of the data points.

Automatic Bins: Default groups into which a histogram's variable will be sorted.

1. A company collects customer data and uses statistical methods to analyze purchasing patterns. This approach is part of which phase in the analytical workflow?

A. Data collection

B. Data preparation

C. Data analysis

D. Data visualization

Answer: C

2. An analyst wants to explore a dataset to uncover patterns without any prior hypothesis. What is the most suitable technique for this?

A. Supervised learning

B. Exploratory data analysis (EDA)

C. Predictive modeling

D. Time-series forecasting

Answer: B

3. A company collects vast amounts of data from customer interactions, but it struggles to derive meaningful insights. What should be their next step?

A. Ignore the data and continue with traditional methods

B. Improve the data preparation and cleaning process

C. Continue collecting more data without analysis

D. Skip visualization and go straight to modeling

Answer: B

4. An analyst uses machine learning to build a predictive model for customer churn based on past customer behavior. What phase of the analytical workflow is this?

A. Data collection

B. Data preparation

C. Modeling

D. Reporting

Answer: C

5. A company wants to understand current market trends based on historical data and future predictions. Which of the following would be the most effective tool?

A. Descriptive analysis

B. Predictive analysis

C. Prescriptive analysis

D. Exploratory data analysis (EDA)

Answer: B

6. A business leader wants to understand the effects of a new marketing campaign on sales.

Which part of the analytical workflow should be prioritized?

A. Data collection

B. Predictive modeling

C. Descriptive analytics

D. Visualization

Answer: B

7. When preparing data for analysis, which of the following steps is most crucial to ensure the

accuracy of the results?

A. Visualization

B. Data cleaning and transformation

C. Reporting

D. Model testing

Answer: B

8. What is the role of prescriptive analytics in the analytical workflow?

A. It predicts future outcomes

B. It suggests actions to optimize business outcomes based on predictive analytics

C. It summarizes data for reporting purposes

D. It visualizes trends and relationships

Answer: B

9. A retail company has been using Excel to analyze sales data, but as data grows larger, they

adopt SPSS for better analysis. What trend does this shift represent?

A. Transition from manual to automated workflows

B. Shift from predictive to prescriptive analysis

C. Increased use of qualitative data

D. Transition from offline to online tools

Answer: A

10. Which of the following file formats can SPSS import data from?

A. Excel (.xls, .xlsx)

B. CSV (.csv)

C. SQL Databases

D. All of the above

Answer: D