Sample Q & A for Viva

Topic 1: Preliminary Preparation and Analysis of Data - Descriptive Statistics

1. Q: How do you get the sum of a list `height = [2,4,6,8,10] in Python?

   A: sum(height)

2. Q: Define an outlier.

   A: An outlier is a data point that is significantly different from the rest of the data.

3. Q: Name two common methods to handle missing values.

   A: Imputation and deletion.

4. Q: What is a histogram used for?

   A: To represent the frequency distribution of numerical data.

5. Q: What does standard deviation measure?

   A: The amount of variation or dispersion in a dataset.

6. Q: Explain the difference between mean and median.

   A: The mean is the average of all data points, while the median is the middle value when data is ordered.

7. Q: How does one identify an outlier using the IQR method?

   A: Outliers are identified as values below Q1 - 1.5IQR or above Q3 + 1.5IQR.

8. Q: What is keyword for defining a function in python?

   A: def

9. Q: What error will you encounter if you define a variable as weight = [67, 73] and then try to print it using print(wight) in Python?

   A: **Name Error**: name 'wight' is not defined

10. Q: How can you visually assess the shape of a data distribution?

    A: By using a histogram or a box plot.

11. Q: What are keywords in Python, and can you provide some examples?

    A: In Python, keywords are reserved words that have special meanings and cannot be used as identifiers (e.g., variable names, function names). These keywords are part of the Python language syntax and are used to define the structure and flow of the code.

Ex: def, class, if, import

` import keyword

print(keyword.kwlist)`

Topic 2: Inferential Statistics, Testing Population, Sample, and Estimation

1. Q: What is a p-value?

   A: The probability of obtaining the observed results assuming the null hypothesis is true.

2. Q: Define a Type I error.

   A: Rejecting the null hypothesis when it is actually true.

3. Q: What is the purpose of a confidence interval?

   A: To estimate the range within which a population parameter lies with a certain level of confidence.

4. Q: What does ANOVA test for?

   A: Differences in means across multiple groups.

5. Q: What does the correlation coefficient measure?

   A: The strength and direction of the linear relationship between two variables.

6. Q: Explain the difference between z-test and t-test.

   A: Z-test is used when the population variance is known and the sample size is large, whereas t-test is used when the population variance is unknown and/or the sample size is small.

7. Q: What is the null hypothesis in a chi-square test for independence?

   A: There is no association between the categorical variables.

8. Q: Describe the significance of the F-statistic in ANOVA.

   A: It compares the variance between group means to the variance within groups to determine if the means are significantly different.

9. Q: What is the impact of increasing the sample size on the margin of error?

   A: Increasing the sample size decreases the margin of error, making the estimate more precise.

10. Q: How is the correlation coefficient different from causation?

    A: Correlation measures association, not causation; it does not imply that changes in one variable cause changes in another.

Topic 3: Regression - Predictive Analytics

1. Q: What does R^2 represent in regression analysis?

   A: The proportion of variance in the dependent variable explained by the independent variables.

2. Q: What is the purpose of a regression line?

   A: To model the relationship between the dependent and independent variables.

3. Q: Define multicollinearity.

   A: A situation in which two or more predictor variables are highly correlated.

4. Q: What is the intercept in a regression equation?

   A: The expected value of the dependent variable when all independent variables are zero.

5. Q: What does the slope of a regression line indicate?

   A: The change in the dependent variable for a one-unit change in the independent variable.

6. Q: Explain the assumption of homoscedasticity in regression.

   A: The variance of the residuals is constant across all levels of the independent variables.

7. Q: What is the difference between R^2 and Adjusted R^2?

   A: Adjusted R^2 accounts for the number of predictors in the model and adjusts for the degrees of freedom.

8. Q: How can you detect multicollinearity in a regression model?

   A: By checking Variance Inflation Factors (VIFs) or correlation matrices.

9. Q: What is the purpose of residual plots in regression diagnostics?

   A: To check for non-linearity, unequal error variances, and outliers.

10. Q: Describe the impact of omitted variable bias in regression analysis.

    A: It leads to biased and inconsistent estimates of the coefficients, affecting the validity of the model.

Topic 4: Limited Dependent Variable Models - Classification Analysis

1. Q: What is logistic regression used for?

   A: To model the probability of a binary outcome.

2. Q: Define a confusion matrix.

   A: A table used to evaluate the performance of a classification model by showing true positives, false positives, true negatives, and false negatives.

3. Q: What is an ROC curve?

   A: A graph that shows the performance of a classification model at all classification thresholds.

4. Q: What does AUC stand for in model evaluation?

   A: Area Under the Curve.

5. Q: What is the purpose of a decision tree?

   A: To make decisions based on a series of rules derived from the data.

6. Q: Explain the difference between logistic regression and linear regression.

   A: Logistic regression models binary outcomes, while linear regression models continuous outcomes.

7. Q: How is the cutoff value in logistic regression determined?

   A: It can be determined based on the desired balance between sensitivity and specificity, often using ROC analysis.

8. Q: What are the assumptions of logistic regression?

   A: No multicollinearity, linear relationship between the logit and predictors, and independent errors.

9. Q: Describe how to handle class imbalance in classification.

   A: Techniques include resampling methods (oversampling minority class, under sampling majority class), using different evaluation metrics, or applying algorithms like SMOTE.

10. Q: What is the Gini index in the context of decision trees?

    A: A measure of impurity used to split nodes in a decision tree, with lower values indicating purer nodes.

Topic 5: Multivariate Analysis and Business Analytics Applications

1. Q: What is the goal of Principal Component Analysis (PCA)?

   A: To reduce data dimensionality while preserving as much variance as possible.

2. Q: Define cluster analysis.

   A: A technique used to group similar observations into clusters based on their characteristics.

3. Q: What does factor analysis aim to identify?

   A: Underlying relationships between variables by grouping them into factors.

4. Q: What is conjoint analysis used for?

   A: To understand consumer preferences and the value they place on different attributes of a product.

5. Q: Explain portfolio optimization.

   A: The process of selecting the best mix of assets to achieve desired returns while minimizing risk.

6. Q: How does PCA handle correlated variables?

   A: It transforms correlated variables into a set of uncorrelated components.

7. Q: Describe the difference between hierarchical and K-means clustering.

   A: Hierarchical clustering builds a tree of clusters, while K-means clustering partitions data into a predetermined number of clusters.

8. Q: How is the eigenvalue used in factor analysis?

   A: To determine the number of factors to retain, with factors having eigenvalues greater than 1 considered significant.

9. Q: What is the main advantage of using conjoint analysis in market research?

   A: It provides insights into the trade-offs consumers make and the relative importance of product attributes.

10. Q: Explain the concept of the efficient frontier in portfolio optimization.

    A: It represents the set of optimal portfolios that offer the highest expected return for a defined level of risk.

Topic 6: Visualization - Perceptual Mapping for Business

1. Q: What is the purpose of a bar chart?

   A: To display and compare the frequency or count of different categories.

2. Q: Define data storytelling.

   A: The process of using data visualizations to communicate a narrative that conveys insights and findings.

3. Q: What type of data is best represented by a pie chart?

   A: Categorical data showing proportions or percentages.

4. Q: How can data visualization enhance data interpretation?

   A: By making complex data more accessible and understandable through visual representation.

5. Q: Describe a scenario where a scatter plot is more useful than a bar chart.

   A: When visualizing the relationship between two continuous variables to identify patterns or correlations.

6. Q: What is the significance of using color effectively in data visualization?

   A: Color helps to differentiate data points, highlight important information, and improve readability.

7. Q: Explain the difference between a heat map and a choropleth map.

   A: A heat map uses color intensity to represent data values on a grid, while a choropleth map uses color shading to represent data values across geographical regions.

Topic 7: Time Series Analysis

1. Q: What is a time series?

A: A sequence of data points recorded at successive points in time.

2. Q: Define seasonality in time series data.

A: Regular, predictable patterns or cycles in data that occur at specific intervals (e.g., monthly, yearly).

3. Q: What is the purpose of decomposition in time series analysis?

A: To separate a time series into trend, seasonal, and residual components.

4. Q: What does ARIMA stand for?

A: Auto Regressive Integrated Moving Average.

5. Q: What is the significance of the lag in time series analysis?

A: It represents the time difference between observations used to identify relationships over time.

6. Q: Explain the concept of stationarity in time series analysis.

A: A stationary time series has constant mean, variance, and autocorrelation over time, essential for modelling and forecasting.

7. Q: Describe the difference between additive and multiplicative decomposition.

A: Additive decomposition assumes the components (trend, seasonality, residual) are added together, while multiplicative assumes they are multiplied.

8. Q: What is the purpose of the autocorrelation function (ACF)?

A: To measure the correlation between observations at different lags in a time series.

9. Q: How does the differencing operation help in making a time series stationary?

A: By subtracting the previous observation from the current one, it removes trends and stabilizes the mean of the series.

10. Q: Explain the use of ARCH and GARCH models in financial time series analysis.

A: ARCH (Autoregressive Conditional Heteroskedasticity) models volatility clustering, while GARCH (Generalized ARCH) models' volatility with a longer memory effect.

Topic 8: Applications Development

1. Q: What is the purpose of a virtual environment in Python?

   A: To create an isolated environment for project dependencies, ensuring consistent development setups.

2. Q: Define Streamlit.

   A: A Python library for creating interactive web applications quickly and easily.

3. Q: What is a framework in software development?

   A: A platform for developing software applications, providing a structure and common functionalities.

4. Q: What does the command `pip install virtualenv` do?

   A: It installs the `virtualenv` package, used for creating virtual environments in Python.

5. Q: What is the purpose of a requirements.txt file?

   A: To list all the dependencies required for a project, which can be installed using `pip`.

6. Q: Describe how to deploy a Streamlit app.

   A: By using platforms like Streamlit Cloud, Heroku, or deploying on a web server with Docker and reverse proxy.