DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

B.Tech. – Artificial Intelligence and Data Science

Anna University Regulation: 2021

AD3301 – Data Exploration and Visualization

II Year / III Semester

QUESTION BANK

Unit I - Exploratory Data Analysis

QUESTION BANK

AD3301 – DATA EXPLORATION AND VISUALIZATION

UNIT I EXPLORATORY DATA ANALYSIS

EDA fundamentals – Understanding data science – Significance of EDA – Making sense of data – Comparing EDA with classical and Bayesian analysis – Software tools for EDA - Visual Aids for EDA- Data transformation techniques-merging database, reshaping and pivoting, Transformation techniques - Grouping Datasets - data aggregation – Pivot tables and cross-tabulations.

PART – A

1. **Define Exploratory Data Analysis (EDA)?**

   EDA is the process of examining and visualizing data to uncover patterns, trends, and insights before more advanced analyses.

2. **What is the significance of EDA in data science?**

   EDA is crucial in data science as it helps identify patterns, outliers, and data quality issues, providing a foundation for further analysis.

3. **Differentiate EDA from classical statistical analysis?**

   EDA focuses on visual exploration, while classical statistical analysis involves hypothesis testing and parameter estimation.

4. **Why is making sense of data important in EDA?**

   Making sense of data involves extracting meaningful information, enabling informed decisions and insights.

5. **Compare EDA with Bayesian analysis?**

   EDA is non-parametric and exploratory, while Bayesian analysis incorporates prior knowledge and updates probabilities based on new data.

6. **Name two software tools commonly used for EDA?**

   Pandas and Matplotlib are commonly used tools for EDA in Python.

7. **Define data transformation techniques in EDA?**

   Data transformation techniques include normalization, scaling, and handling missing values to prepare data for analysis.

8. **What is the purpose of merging databases in EDA?**

   Merging databases combines datasets based on common identifiers to create a unified dataset for analysis.

9. **Differentiate between reshaping and pivoting in EDA?**

   Reshaping transforms data between wide and long formats, while pivoting reorganizes data to create a new structure.

10. **Define data aggregation in EDA?**

    Data aggregation involves summarizing grouped data using functions like sum, mean, orcount.

11. **How do pivot tables aid in EDA?**

    Pivot tables facilitate multidimensional analysis and summarization of data in a tabularformat.

12. **What visual aids are commonly used in EDA?**

    Histograms, box plots, scatter plots, and heatmaps are common visual aids in EDA for understanding data distributions and relationships.

13. **Define the concept of grouping datasets in EDA?**

    Grouping datasets involves creating subsets based on certain criteria, enabling focused analysis on specific segments.

14. **Why is cross-tabulation useful in EDA?**

    Cross-tabulation is useful in EDA for displaying the frequency distribution of variables in a contingency table.

15. **Name a transformation technique in EDA for handling outliers?**

    Winsorizing is a transformation technique that involves replacing extreme values with less extreme values to handle outliers.

16. **Define the term "data normalization" in EDA?**

    Data normalization in EDA is the process of rescaling variables to a standard range, typically between 0 and 1.

17. **What is the role of visual aids like violin plots in EDA?**

    Violin plots display the distribution of data, providing insights into both central tendency and spread.

18. **Define the concept of data scaling in EDA?**

    Data scaling in EDA involves transforming variables to have a similar scale, preventing dominance by certain features.

19. How does EDA contribute to data science projects?

EDA contributes by providing an initial understanding of data, guiding subsequent modeling and analysis decisions.

20. Why are pivot tables and cross-tabulations useful in summarizing data?

Pivot tables and cross-tabulations provide a concise summary of data, making it easier to identify patterns and trends across different dimensions.

PART – B

1. Explain the Purpose of EDA
2. Differentiate EDA from Classical Analysis
3. Illustrate Visual Aids in EDA
4. Describe Data Transformation in EDA
5. Explore the Significance of Grouping Datasets and how it aids in focused analysis.
6. Explain the Role of Data Aggregation
7. Illustrate the Application of Pivot Tables
8. Compare EDA with Bayesian Analysis:

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

B.Tech. – Artificial Intelligence and Data Science

Anna University Regulation: 2021

AD3301 – Data Exploration and Visualization

II Year / III Semester

QUESTION BANK

Unit II - Visualizing Using MATPLOTLIB

QUESTION BANK

AD3301 – DATA EXPLORATION AND VISUALIZATION

UNIT II VISUALIZING USING MATPLOTLIB

Importing Matplotlib – Simple line plots – Simple scatter plots – visualizing errors – density and contour plots – Histograms – legends – colors – subplots – text and annotation – customization – three dimensional plotting - Geographic Data with Basemap - Visualization with Seaborn.

PART – A

1. **Define Matplotlib**

Matplotlib is a Python library for creating static, interactive, and animated visualizations in a variety of formats.

2. **Differentiate between Line Plot and Scatter Plot**

A line plot connects data points with straight lines, while a scatter plot displays individual data points without connecting lines.

3. **What is the purpose of visualizing errors in plots?**

Visualizing errors helps assess the uncertainty or variability associated with data points, providing insights into the reliability of the measurements.

4. **Define Density Plot**

A density plot visualizes the distribution of a continuous variable, providing a smoothed representation of its underlying probability density.

5. **How do histograms represent data?**

Histograms display the distribution of a dataset by dividing it into bins and counting the frequency of data points within each bin.

6. **Explain the role of legends in plots**

Legends in plots provide labels to distinguish between different data series or elements within a plot.

7. **Differentiate between Colors and Colormaps in Matplotlib**

Colors refer to individual colors used in plots, while colormaps are color schemes that map data values to a range of colors.

8. **Define Subplots in Matplotlib**

Subplots allow the creation of multiple plots within the same figure, enabling side-by-side comparisons of different datasets.

9. **What is the purpose of text and annotation in plots?**

Text and annotation in plots help add explanatory labels or comments to highlight specific features or points of interest.

10. How can plots be customized in Matplotlib?

Plots in Matplotlib can be customized by adjusting parameters such as line styles, markers, axis labels, and title font sizes.

11. Define Three-Dimensional Plotting in Matplotlib

Three-dimensional plotting in Matplotlib involves creating visualizations with three axes, providing depth along with width and height.

12. What is Geographic Data Visualization with Basemap?

Basemap is a Matplotlib toolkit for plotting 2D data on maps, facilitating the visualization of geographic data.

13. Explain the significance of Visualization with Seaborn

Seaborn is a statistical data visualization library that simplifies the creation of aesthetically pleasing and informative visualizations.

14. Differentiate between Line Plots and Contour Plots

Line plots show the relationship between two variables with lines, while contour plots represent three-dimensional data using contours or isolines.

15. Define the term "Scatter Plot Matrix" in Matplotlib

A Scatter Plot Matrix is a grid of scatter plots that allows the simultaneous visualization of relationships between multiple pairs of variables.

16. Why is color important in data visualization?

Color in data visualization helps convey information, distinguish between different elements, and highlight patterns within the data.

17. What role do subplots play in complex visualizations?

Subplots allow for the organization of multiple plots within a single figure, aiding in the creation of complex visualizations and comparisons.

18. Define the term "Colormap" in Matplotlib

A colormap is a range of colors used in a plot to represent variations in data values, helping to convey information visually.

19. How does Matplotlib handle 3D plotting?

Matplotlib provides tools for creating three-dimensional plots, allowing visualization of data in three dimensions.

20. What is the advantage of using Seaborn for visualization?

Seaborn simplifies the process of creating aesthetically pleasing statistical visualizations, providing a high-level interface for complex plots.

PART – B

1. Explain the Process of Creating a Simple Line Plot in Matplotlib?
2. Illustrate the Use of Scatter Plots in Visualizing Relationships?
3. Describe the Role of Density Plots in Data Visualization?
4. Illustrate the Construction and Interpretation of Histograms?
5. Describe the role of legends in plots and provide examples to illustrate how they contribute to making plots more informative and interpretable?
6. Describe the Customization Options in Matplotlib?
7. Provide a step-by-step illustration of how to create subplots in Matplotlib?
8. Explain the purpose of using subplots and their significance in complex visualizations?
9. Explain the Process of Geographic Data Visualization with Basemap?

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

B.Tech. – Artificial Intelligence and Data Science

Anna University Regulation: 2021

AD3301 – Data Exploration and Visualization

II Year / III Semester

QUESTION BANK

Unit III - Univariate Analysis

QUESTION BANK

AD3301 – DATA EXPLORATION AND VISUALIZATION

UNIT III UNIVARIATE ANALYSIS

Introduction to Single variable: Distributions and Variables - Numerical Summaries of Level and Spread - Scaling and Standardizing – Inequality - Smoothing Time Series.

PART – A

1. **Define Single Variable?**

   A single variable represents an individual characteristic, attribute, or measurement that can take different values in a dataset.

2. **Differentiate between Level and Spread in Data?**

   Level refers to the central tendency or average of a dataset, while spread represents the variability or dispersion of the data.

3. **What are Numerical Summaries of Level and Spread?**

   Numerical summaries of level include measures like mean and median, while those of spread involve measures like range, variance, and standard deviation.

4. **Define Scaling and Standardizing?**

   Scaling adjusts the numerical range of a variable, while standardizing transforms it to have a mean of 0 and a standard deviation of 1.

5. **Differentiate between Scaling and Standardizing?**

   Scaling changes the range of values, while standardizing rescales to have a mean of 0 and a standard deviation of 1.

6. **Explain the Concept of Inequality in Data?**

   Inequality in data refers to the uneven distribution of values, where some values occur more frequently than others.

7. **Define Time Series Smoothing?**

   Time series smoothing involves removing noise or short-term fluctuations from a dataset to highlight underlying trends.

8. **Differentiate between Mean and Median?**

   The mean is the average value, while the median is the middle value in a sorted dataset.

9. **What is the Purpose of Scaling Variables?**

   Scaling variables is done to bring them to a comparable numerical range, avoiding dominance by variables with larger magnitudes.

10. Define Interquartile Range (IQR)?

The Interquartile Range (IQR) is the range between the first and third quartiles, representing the central 50% of the data.

11. Differentiate between Range and Standard Deviation?

Range is the difference between the maximum and minimum values, while standard deviation measures the average deviation of values from the mean.

12. Explain the Role of Boxplots in Data Visualization?

Boxplots visually represent the distribution of data, showing median, quartiles, and potential outliers.

13. Define Skewness in a Distribution?

Skewness measures the asymmetry of a distribution; positive skewness indicates a tail to the right, and negative skewness indicates a tail to the left.

14. What is Z-Score in Standardization?

Z-Score represents the number of standard deviations a data point is from the mean in a standardized distribution.

15. Define the Coefficient of Variation (CV)?

The Coefficient of Variation (CV) is the ratio of the standard deviation to the mean, providing a relative measure of variability.

16. Differentiate between Time Series and Cross-Sectional Data?

Time series data is collected over time, while cross-sectional data is collected at a single point in time.

17. Explain the Concept of Kernel Density Estimation?

Kernel Density Estimation is a non-parametric method to estimate the probability density function of a random variable, providing a smooth curve.

18. Define Outliers in a Dataset?

Outliers are data points that deviate significantly from the rest of the dataset and may distort statistical analyses.

19. Explain the Importance of Scaling in Machine Learning?

Scaling is important in machine learning to ensure that all features contribute equally to model training, preventing bias towards variables with larger scales.

20. What is the Purpose of Time Series Smoothing Techniques?

Time series smoothing techniques aim to reveal underlying trends by reducing noise and short-term fluctuations in data.

PART – B

1. Explain the Significance of Numerical Summaries in Single Variable Analysis?
2. Illustrate the Process of Scaling and Standardizing Variables?
3. Describe the Role of Interquartile Range (IQR) in Data Analysis?
4. Explain the Concept of Skewness and Its Impact on Distributions?
5. Illustrate the Application of Kernel Density Estimation in Data Visualization?
6. Describe the Importance of Outlier Detection in Data Analysis?
7. Explain the Purpose of Time Series Smoothing Techniques?
8. Provide a step-by-step illustration of calculating the Coefficient of Variation (CV)?
9. Explain how it offers a relative measure of variability and its interpretation in different scenarios?

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

B.Tech. – Artificial Intelligence and Data Science

Anna University Regulation: 2021

AD3301 – Data Exploration and Visualization

II Year / III Semester

QUESTION BANK

Unit IV - Bivariate Analysis

QUESTION BANK

AD3301 – DATA EXPLORATION AND VISUALIZATION

UNIT IV BIVARIATE ANALYSIS

Relationships between Two Variables - Percentage Tables - Analyzing Contingency Tables - Handling Several Batches - Scatterplots and Resistant Lines – Transformations.

PART – A

1. **Define Relationships between Two Variables?**
   Relationships between two variables refer to the association or correlation between their respective values in a dataset.
2. **Differentiate between Percentage Tables and Contingency Tables?**
   Percentage tables display proportions, while contingency tables show the distribution of two categorical variables.
3. **What is the Purpose of Analyzing Contingency Tables?**
   Analyzing contingency tables helps examine the relationship between two categorical variables and identify patterns or dependencies.
4. **Define Handling Several Batches in Data Analysis?**
   Handling several batches involves managing and analyzing data that is collected or processed in multiple groups or sets.
5. **Differentiate between Scatterplots and Resistant Lines?**
   Scatterplots visually represent the relationship between two numerical variables, while resistant lines are fitted lines that minimize the impact of outliers.
6. **Explain the Significance of Transformations in Data Analysis?**
   Transformations modify the scale or distribution of data, aiding in making relationships more linear or meeting statistical assumptions.
7. **What is the Purpose of Percentage Tables in Data Analysis?**
   Percentage tables express values as a percentage of the total, providing a clearer understanding of the relative contribution of each category.
8. **Define Contingency Tables in Statistics?**
   Contingency tables organize and display the frequency distribution of two categorical variables, showing how they are related.
9. **How do Scatterplots Facilitate Relationship Visualization?**
   Scatterplots visually display the relationship between two numerical variables, allowing for the observation of patterns, trends, or correlations.

10. **Explain the Concept of Resistant Lines in Regression?**

    Resistant lines in regression minimize the impact of outliers, providing a more accurate representation of the overall relationship between two variables.

11. **Define Transformation Techniques in Data Analysis?**

    Transformation techniques modify the structure or scale of data to meet statistical assumptions or enhance analysis.

12. **Differentiate between Batch Handling and Data Aggregation?**

    Batch handling involves managing data in separate groups, while data aggregation combines and summarizes information across groups.

13. **What is the Role of Scatterplots in Identifying Patterns?**

    Scatterplots visually reveal patterns or trends in data, aiding in the identification of relationships between two variables.

14. **Define the Purpose of Percentage Tables in Categorical Data Analysis?**

    Percentage tables are useful in categorical data analysis to express the distribution of categories as a percentage of the total, providing a relative comparison.

15. **Explain the Significance of Handling Several Batches in Experimental Design?**

    Handling several batches is important in experimental design to account for variations introduced by different conditions, ensuring robust and generalizable results.

16. **Define the Concept of Data Transformation in Regression Analysis?**

    Data transformation in regression alters the form or distribution of data to meet assumptions, improve model fit, or enhance interpretability.

17. **How do Contingency Tables Aid in Statistical Inference?**

    Contingency tables assist in statistical inference by providing a structured overview of the relationship between two categorical variables.

18. **What is the Purpose of Resistant Lines in Outlier-Prone Data?**

    Resistant lines are valuable in outlier-prone data as they reduce the impact of extreme values, ensuring a more reliable interpretation of the relationship between variables.

19. **Define the Role of Scatterplots in Exploratory Data Analysis?**

    Scatterplots play a crucial role in exploratory data analysis by visually uncovering potential relationships, outliers, or patterns between two variables.

20. **How do Transformations Improve Linearity in Regression?**

    Transformations in regression analysis can improve linearity, making the relationship between variables more suitable for linear modeling and interpretation.

PART – B

1. Explain the importance of relationships between two variables in data analysis?
2. Illustrate the construction and interpretation of percentage tables?
3. Describe the process of analyzing contingency tables and its applications?
4. Explain the challenges and strategies in handling several batches in data analysis?
5. Illustrate the role of scatterplots in visualizing relationships?
6. Explain how patterns and trends are identified through scatterplot analysis?
7. Describe the characteristics and applications of resistant lines in regression?
8. Explain the concept of transformations in data analysis and provide examples?
9. Illustrate the application of batch handling in experimental design?

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

B.Tech. – Artificial Intelligence and Data Science

Anna University Regulation: 2021

AD3301 – Data Exploration and Visualization

II Year / III Semester

QUESTION BANK

Unit V - Multivariate and Time Series Analysis

QUESTION BANK

AD3301 – DATA EXPLORATION AND VISUALIZATION

UNIT V MULTIVARIATE AND TIME SERIES ANALYSIS

Introducing a Third Variable - Causal Explanations - Three-Variable Contingency Tables and Beyond - Longitudinal Data – Fundamentals of TSA – Characteristics of time series data – Data Cleaning – Time-based indexing – Visualizing – Grouping – Resampling.

PART – A

1. **Define Causal Explanations in Data Analysis?**
   Causal explanations seek to establish a cause-and-effect relationship between variables, explaining how changes in one variable influence another.
2. **Differentiate Between Two-Variable and Three-Variable Contingency Tables?**
   Two-variable contingency tables display the relationship between two categorical variables, while three-variable contingency tables incorporate a third categorical variable.
3. **Explain the Significance of Longitudinal Data in Analysis?**
   Longitudinal data tracks observations over time, providing insights into trends and changes in variables over different periods.
4. **Describe the Fundamentals of Time Series Analysis (TSA)?**
   Time Series Analysis (TSA) involves examining and modeling data points collected over time to understand patterns and trends.
5. **Define Characteristics of Time Series Data?**
   Time series data is characterized by observations collected sequentially over time, displaying trends, seasonality, and potential cyclical patterns.
6. **Explain the Importance of Data Cleaning in Time Series Analysis?**
   Data cleaning in time series analysis involves identifying and rectifying errors or anomalies to ensure the accuracy and reliability of temporal patterns.
7. **What is Time-Based Indexing in Time Series Data?**
   Time-based indexing involves using time as the index for organizing and accessing data in time series datasets.
8. **Illustrate the Process of Visualizing Time Series Data:**
   Visualization of time series data involves creating plots or charts to visually represent patterns, trends, and fluctuations over time.
9. **Explain the Concept of Grouping in Time Series Analysis?**
   Grouping in time series analysis involves categorizing data into meaningful subsets based on certain characteristics, facilitating more focused analysis.

10. **What is Resampling in the Context of Time Series Analysis?**

    Resampling involves changing the frequency of time series data, such as aggregating daily data into weekly or monthly intervals for a broader perspective.

11. **Define the Term "Third Variable" in Data Analysis?**

    A third variable in data analysis is an additional factor considered alongside two primary variables to explore relationships and potential causal effects.

12. **Differentiate Between Cross-Sectional and Longitudinal Data?**

    Cross-sectional data is collected at a single point in time, while longitudinal data tracks observations over a period, allowing the analysis of changes over time.

13. **Explain the Role of Causal Explanations in Predictive Modeling?**

    Causal explanations are valuable in predictive modeling as they provide insights into the factors influencing predictions, enhancing model interpretability.

14. **Illustrate the Application of Three-Variable Contingency Tables?**

    Three-variable contingency tables display the joint distribution of three categorical variables, enabling a comprehensive analysis of relationships among them.

15. **Describe the Essentials of Time Series Analysis?**

    Time series analysis involves examining patterns, trends, and seasonal variations in sequential data, providing insights into temporal dynamics.

16. **What are the Characteristics of Seasonality in Time Series Data?**

    Seasonality in time series data refers to recurring patterns or fluctuations that follow a regular time interval, often associated with specific seasons or time periods.

17. **Explain the Purpose of Time-Based Indexing in Data Analysis?**

    Time-based indexing organizes data chronologically, facilitating efficient retrieval and analysis of time series data.

18. **Illustrate the Cleaning Steps in Time Series Data?**

    Data cleaning in time series involves steps like handling missing values, smoothing outliers, and addressing seasonality to ensure accurate analysis.

19. **Define the Concept of Resampling Frequency in Time Series Analysis?**

    Resampling frequency in time series analysis refers to the interval at which data points are aggregated or disaggregated, affecting the granularity of analysis

20. Explain the Role of Grouping in Visualizing Time Series Data?

Grouping in time series visualization involves organizing data into subsets to observe and compare patterns within specific groups or categories.

PART – B

1. Describe how introducing a third variable enhances the understanding of relationships between two other variables. Provide examples to illustrate its impact?
2. Illustrate the Process of Establishing Causal Explanations?
3. Describe the Application of Three-Variable Contingency Tables?
4. Explain the construction and interpretation of three-variable contingency tables?
5. Explain the Significance of Longitudinal Data in Trend Analysis?
6. Illustrate the Steps of Time Series Analysis?
7. Describe the Characteristics of Time Series Data?
8. Explain the Cleaning Procedures for Time Series Data?
9. Illustrate the Importance of Resampling in Time Series Analysis?