**ADATA 5230 Module 2 Assignment Chaudhary Yog 95 Answer Q1 to Q12**

**Q1. What are the examples to basic plots and distribution plots?**

Answer: Basic plots include bar plots, line plots, and scatter plots. Distribution plots include histograms and box plots.

**Bar Plot:** A bar plot is used to display data with rectangular bars. Used for visualizing categorical data or comparing different categories.

**Line Plot:** A line plot is used to display points connected by a straight line. Used to show trends over time.

**Scatter plots:** Scatter plots are used to visualize the relationship between two continuous. Each data point is represented as a dot on the plot.

**Distribution Plots:** Box plots display the distribution of a dataset's values, showing the median, quartiles, and potential outliers.

**Histogram:** Histograms are used to show the distribution of a single variable. They divide the data into bins and display the frequency of values within each bin.

**Q2. What is meant by the phrase “a picture is worth a thousand words”?**

Answer: A picture is worth a thousand words" which means that a visual representation can convey complex ideas more effectively than a written or verbal description. Data flow systems can be used for the representation of different data which could be general or very specific. The logical elements of a system could include describing the boundaries of a system. System boundary could be described by communicating existing system knowledge to the users involved in the data flow. As a result, many users benefit from data flow diagrams because they are easy to depict by all audiences. flowchart could describe a process throughout multiple departments including manual activities, computer systems, and accounting. Flowcharts are a great tool to monitor waste and inefficiency. An example of a system flowchart could be software installation. This is generally the same with all users but different decision processes along the way could result in different outcomes of the installation. There are key differences between data diagrams and system flowcharts. A system flowchart will show a user the process of data from start to finish. It can also show the different processes along the way that can delay a process or where decision-making could take place. A data diagram will show how data flows through a system. Internal and external sources of information can be a deciding factor in how the diagram moves through the process as a flowchart does not have any external conflict. As a system flowchart shows the steps the data diagram shows the movement of data through a system.

**Q3. Where do visualization techniques fit into the data mining process?**

Answer: Visualization techniques are used in the data mining process to explore and analyze data, identify patterns and trends, and communicate insights to stakeholders.

**Q4. Graphical exploration supports free-form exploration for what purposes?**

Answer: Graphical exploration supports free-form exploration for purposes such as identifying outliers, understanding relationships between variables, and detecting patterns or trends.

Overall, graphical exploration in data analysis offers a flexible way to interact with data, encouraging curiosity, hypothesis generation, and a deeper understanding of complex datasets. It allows analysts to go beyond structured analysis and discover insights that may not be apparent through traditional statistical methods alone.

**Q5. Which Python library for data visualization is the oldest with high flexibility?**

Answer: **Matplotlib** is the oldest Python library for data visualization and offers high flexibility in creating various types of plots. While Matplotlib is known for its flexibility, it's worth noting that it can have a steeper learning curve compared to some more recent Python visualization libraries like Seaborn and Plotly, which offer higher-level abstractions and easier-to-use interfaces for certain types of plots. However, Matplotlib remains a powerful and versatile choice for those who require fine-grained control over their visualizations and are willing to invest time in learning its intricacies. Additionally, it serves as the foundation for many other Python visualization libraries.

**Q6. Does MEDV of Boston Housing data variable follow the Normal Distribution? Explain your answer shortly.**

Answer: The MEDV variable in the Boston Housing data does not follow a normal distribution. It represents median values of owner-occupied homes, which can be influenced by various factors.

Q7. Which variable pairs of Boston Housing data have the highest/lowest correlation?

Answer: The variable pairs with the highest/lowest correlation in the Boston Housing data can be determined by calculating the correlation coefficient between each pair of variables.

**Q8. Why do we rescale some of the variables?**

Answer: Rescaling variables is a data preprocessing technique in data analysis and machine learning. It involves transforming the values of one or more variables. Its method depends on the nature of the data and the specific requirements of the analysis or modeling task. Rescaling variables is done to ensure that they are on a similar scale, which can prevent certain variables from dominating the analysis or affecting the results disproportionately.

**Q9. What is dimensionality of a model?**

Answer: The dimensionality of a model refers to the number of features or variables used to represent the data. The dimensionality of a model is a critical factor in data analysis and machine learning. The choice of the right dimensionality depends on the specific problem, the quality and quantity of data, and the modeling techniques being employed. Properly managing dimensionality is essential for building effective and interpretable models.

**Q10. List down five commonly used summary statistics.**

Answer: Five commonly used summary statistics are mean, median, mode, standard deviation, and range. These statistics offer insights into the central tendency and distribution of datasets, making them fundamental tools for data analysis.

**Q11. What is PCA and its main purpose?**

Answer: PCA (Principal Component Analysis) is a dimensionality reduction technique used to transform a dataset into a lower-dimensional space while retaining most of the original information. Its main purpose is to simplify complex data and identify the most important features or patterns. They also create new variables that are linear combinations of the original variables. Therefore, the weighted average of the original variables. These linear combinations are uncorrelated and only a few of them contain most of the original information. The new variables are called principal components.

**Q12. What is the main reason behind the domination of sodium and potassium in the breakfast cereals example, and how do we correct for this domination?**

Answer: The domination of sodium and potassium in the breakfast cereals example could be due to their higher concentrations compared to other nutrients. To correct this domination, we can normalize the nutrient values by scaling them to a common range or using other normalization techniques.