18/8/2022 Question paper solution

Q1) Attempt any Eight of the following: [8x 1 = 8]

a) Define volume characteristic of data in reference to data science

Ans. Volume of Data Defined The volume of data refers to the size of data sets that an organization has collected to be analyzed and processed.

b) Give examples of semi structured data.

Ans. HTML, Email , CSV, XML, and JSON CSV, XML, and JSON are the examples of semi structured data.

c) Define Data Discretization.

Ans. Data discretization is characterized as a method of translating attribute values of continuous data into a finite set of intervals with minimal information loss. Data discretization facilitates the transfer of data by substituting interval marks for the values of numeric data.

d) What is a quartile?

Ans. Quartiles are position indicators that divide a sequence of numbers into 4 equal parts A quartile divides data into three points—a lower quartile, median, and upper quartile.

e) List different types of attributes.

Ans. Nominal Attributes, Binary Attributes, Ordinal Attributes, Numeric Attributes, Discrete Attribute, Continuous Attribute these are the different types of attributes in data science.

f) Define Data object.

Ans. A data object is a collection of one or more data points that create meaning as a whole.

g) What is Data Transformation?

Ans. The data transformation is the process of converting the data set formats to other required formats and forms.

h) Write the tools used for geospatial data.

Ans. Shapely python library, geopandas python library, gdal(geospatial data abstraction library), Fiona library, Rasterio library these tolls are used for geospatial data.

i) State the methods of feature selection.

Ans. Univariate selection, Recursive feature elimination, Stepwise forward selection, Stepwise backward selection, Combination of forward and backward elimination, Decision tree induction These are the methods for feature selection in data science.

j) List any two libraries used in Python for data analysis.

Ans. NumPy, SciPy, Pandas , Matplotlib , Scikit-learn are libraries used in data analytics.

04-02-2023

Q2) Attempt any FOUR of the following [4x 2= 8]

a) Explain any two ways in which data is stored in files.

Ans. i) **XLSX**: The XLSX file is Microsoft Excel Open XML Format Spreadsheet file. This is used to store any type of data but it's mainly used to store financial data and to create mathematical models etc.
ii) **CSV**: the CSV is stand for Comma-separated values. as-well-as this name CSV file is use comma to separated values. In CSV file each line is a data record and Each record consists of one or more than one data fields, the field is separated by commas.
iii) **ZIP**: ZIP files are used an data containers, they store one or more than one files in the compressed form. it widely used in internet After you downloaded ZIP file, you need to unpack its contents in order to use it.
iv) **TXT**: TXT files are useful for storing information in plain text with no special formatting beyond basic fonts and font styles. It is recognized by any text editing and other software programs
v) **HTML**: HTML is stand for stands for Hyper Text Markup Language is use for creating web pages. we can read html table in python pandas using read_html() function.
vi) **PDF**: pdf stands for Portable Document Format (PDF) this file format is use when we need to save files that cannot be modified but still need to be easily available

b) Explain role of statistics in data science

Ans. Statistics plays a crucial role in data science as it provides methods to analyze, summarize and make inferences from data. It helps in understanding the underlying patterns and relationships in the data, creating predictive models, and making data-driven decisions. Statistics provides techniques to clean, pre-process, and transform data into a form suitable for analysis. It also provides tools for hypothesis testing, estimation of parameters, and evaluation of model performance. Additionally, it helps in identifying outliers, detecting and handling missing values, and performing dimensionality reduction. Overall, statistics forms the foundation of data science and helps in making sense of the vast amounts of data that is generated in the modern world.

c) Explain two methods of data cleaning for missing values.

Ans. 1.Mean/Median Imputation: In this method, missing values are replaced with either the mean or median of the non-missing values in the same column. This method is suitable for continuous variables and assumes that the missing values are missing at random and have a similar distribution to the non-missing values.

2.Multiple Imputation: In this method, multiple datasets are created by imputing the missing values using a statistical model such as regression or clustering. Each dataset is analyzed and the results are combined to account for the uncertainty in imputed values. This method is more robust and provides more accurate results compared to mean/median imputation, especially when the missing values are not missing at random.

d) Explain any two tools in data scientist toolbox.

Ans. 1.Jupyter Notebook: Jupyter Notebook is a web-based interactive computational environment that allows data scientists to create and share documents that contain live code, equations, visualizations, and narrative text. It provides an easy way to perform data analysis and prototyping, and is widely used in data science and machine learning projects.

2.Pandas: Pandas is a fast, flexible, and powerful open-source data analysis and data manipulation library for Python. It provides data structures for efficiently storing large datasets and tools for working with them, including data cleaning, filtering, grouping, and aggregating. Pandas makes it easy to manipulate and analyze large datasets, and is a essential tool in the data scientist's toolbox.

A word cloud (also known as a tag cloud) is a visual representation of the frequency of words in a text corpus. The size and color of each word in the cloud indicate its frequency in the text. Word clouds provide a quick and simple way to understand the most commonly used words in a large text, and can be used to identify patterns and trends in the data. They are widely used in the fields of text mining, natural language processing, and data visualization. Word clouds can be generated using a variety of tools and software, and are often used as a first step in the analysis of large text datasets.

04-02-2023

Q3) Attempt any TWO of the following: [2 x 4 = 8]

a) Explain data seience life eyele with suitable diagram.

Ans. The data science life cycle is a systematic process that outlines the steps involved in solving a data-related problem, from acquiring and cleaning the data to deploying the final solution. A typical data science life cycle includes the following steps:

1.Problem Definition: Defining the business problem to be solved using data.

Data Collection: Gathering the relevant data needed to solve the problem.

2.Data Cleaning and Pre-processing: Removing missing values, outliers, and inconsistencies, and transforming the data into a format suitable for analysis.

3.Exploratory Data Analysis (EDA): Analyzing the data to understand its distribution, relationships, and patterns.

4.Model Development: Selecting appropriate algorithms and models to build a solution.

5.Model Evaluation: Evaluating the performance of the model, and fine-tuning it if needed.

6.Model Deployment: Deploying the final solution in a production environment.

7.Monitoring and Maintenance: Monitoring the performance of the deployed solution and updating it as necessary.

b) Explain concept and use of data visualisation.

Ans. Data visualization is the process of transforming data and information into graphical forms, such as charts, graphs, and maps, to help people understand and interpret the data more effectively. The main goal of data visualization is to communicate complex information in a simple and intuitive way that can be easily comprehended by a wide audience. It helps to identify patterns, trends, and relationships in the data that might be difficult to detect otherwise.

Data visualization is used in a variety of fields, including business, science, medicine, and engineering, to help decision-makers make informed decisions, communicate insights to stakeholders, and present information in an engaging and memorable way. It is also used for exploratory data analysis, to discover patterns in the data, and for presentation and reporting, to share findings and insights with others.

c) Calculate the variance and standard deviation for the following data. X: 14 9 13 16 25 7 12.

Ans. The variance and standard deviation are two common measures of dispersion in a set of data. The variance is a measure of how spread out the data is, while the standard deviation is the square root of the variance and provides a measure of the average deviation of the data from the mean.

To calculate the variance and standard deviation for the data set X = [14, 9, 13, 16, 25, 7, 12], we first need to find the mean.

Mean (μ) = (14 + 9 + 13 + 16 + 25 + 7 + 12) / 7 = 14

Next, we subtract the mean from each data point and square the results to find the deviations from the mean.

Deviations from the mean:

(14 - 14)^2 = 0

(9 - 14)^2 = 25

(13 - 14)^2 = 1

(16 - 14)^2 = 4

(25 - 14)^2 = 121

(7 - 14)^2 = 49

(12 - 14)^2 = 4

Next, we sum the squared deviations and divide by the number of data points minus one (n-1) to find the variance:

Variance (σ^2) = (0 + 25 + 1 + 4 + 121 + 49 + 4) / (7 - 1) = 95.43

Finally, we find the standard deviation by taking the square root of the variance:

Standard deviation (σ) = √(95.43) = 9.77

So, the variance and standard deviation for the data set X = [14, 9, 13, 16, 25, 7, 12] are **95.43** and **9.77**, respectively.

24-02-2023

04) Attempt any TWO of the following: [2 x 4 = 8]

a) Write a short note on hypothesis testing.

Ans. Hypothesis testing is a statistical method used to test a claim or assumption about a population based on a sample of data. It is a crucial tool in decision-making, especially in the field of statistics and scientific research.

In hypothesis testing, a null hypothesis and an alternative hypothesis are formulated. The null hypothesis represents the status quo and is usually the opposite of what the researcher wants to prove. The alternative hypothesis represents the researcher's claim or assumption.

A test statistic is calculated from the sample data, and its distribution under the null hypothesis is determined. The test statistic is then compared to a critical value or a p-value, which is the probability of observing a test statistic as extreme or more extreme than the one calculated, assuming the null hypothesis is true.

If the p-value is less than a pre-determined level of significance (often denoted by α), the null hypothesis is rejected, and the alternative hypothesis is accepted. If the p-value is greater than α, the null hypothesis is not rejected, and no conclusion is drawn about the alternative hypothesis.

In summary, hypothesis testing is a method used to make inferences about a population based on a sample of data and helps to make informed decisions by testing claims and assumptions.

b) Differentiate between structured data and unstructured data.

Ans.

Feature	Structured Data	Unstructured Data
Definition	Data that is organized in a well-defined and predictable manner, such as a database table.	Data that does not have a well-defined format, such as text, images, and audio.
Examples	Customer records, financial transactions, product information	Social media posts, email, customer reviews
Format	Numerical or categorical data	Text, images, audio, video
Ease of Processing	Easy to process and analyze	Difficult to process and analyze
Storage	Stored in databases or spreadsheets	Stored in large unstructured data stores such as Hadoop or NoSQL databases
Data Structuring	Has a well-defined structure, such as rows and columns	Has no defined structure

c) Explain data visualization libraries in Python.

Ans. Data visualization is an important aspect of data analysis, as it helps to better understand and communicate insights from data. There are several libraries in Python that can be used to create different types of visualizations, including:

1. **Matplotlib**: This is one of the most widely used data visualization libraries in Python. It is a 2D plotting library that can be used to create a wide range of visualizations, including line plots, scatter plots, bar plots, histograms, and more.
2. **Seaborn:** This library is built on top of Matplotlib and provides a high-level interface for creating visualizations. Seaborn provides a range of visualization options, including heat maps, violin plots, and box plots.

04-05-2023

3. **Plotly**: This is an interactive data visualization library that allows users to create dynamic and interactive visualizations. It supports a wide range of visualization types, including bar charts, line charts, scatter plots, and more.
4. **Bokeh:** This library provides a high-level interface for creating interactive visualizations. It is well suited for large datasets, as it provides efficient rendering for large data sets.
5. **ggplot:** This is a data visualization library that is inspired by the syntax of the R programming language. It provides a simple syntax for creating visualizations and supports a wide range of visualization types.

In conclusion, these libraries provide a wide range of options for creating different types of visualizations in Python. Whether you are looking to create simple plots or interactive visualizations, there is a library in Python that can meet your needs.

04-02-2023

Q5) Attempt any ONE of the following: [1 x 3 = 3]

a) i) Define data science. [1]
ii) Explain any one technique of data transformation. [2]

Ans. **i) Definition of Data Science:** Data science is an interdisciplinary field that combines statistical methods, computer science, and domain expertise to extract insights and knowledge from data. It involves the collection, cleaning, processing, and analysis of large and complex data sets, with the goal of uncovering patterns, trends, and insights that can inform decision-making.

ii) One Technique of Data Transformation: Normalization is a technique in data transformation that rescales the values in a data set to a standard scale, typically between 0 and 1. The purpose of normalization is to remove the impact of scaling differences in the data and to make the data more comparable and interpretable.

Normalization can be performed using a number of methods, including Min-Max Normalization, Z-Score Normalization, and Decimal Scaling. In Min-Max Normalization, the data is scaled such that the minimum value is 0 and the maximum value is 1. In Z-Score Normalization, the data is rescaled such that the mean of the data is 0 and the standard deviation is 1. In Decimal Scaling, the data is rescaled by dividing each value by a power of 10.

Normalization is a critical step in many data analysis and machine learning processes, as it helps to remove the impact of scaling differences in the data and makes the data more comparable and interpretable.

b) i) Write any two applications of data science. [1]
ii) Explain any one type of outliers in detail. [2]

Ans.

i) Applications of Data Science:

1. **Healthcare:** In healthcare, data science is used to analyze large and complex data sets, such as patient records and medical images, to gain insights into disease progression and treatment outcomes. This helps healthcare professionals to make more informed decisions and improve patient outcomes.
2. **Marketing:** Data science is widely used in marketing to analyze customer data, such as purchase history and demographic information, to better understand customer behaviour and preferences. This information can then be used to target marketing campaigns more effectively and improve customer engagement.

ii) Types of Outliers: An outlier is a data point that is significantly different from the other data points in a data set. There are several types of outliers, including:

1. **Univariate Outliers:** These are outliers in a single variable data set. They can be identified using a range of methods, including the use of box plots, scatter plots, and Z-score calculations.
2. **Multivariate Outliers:** These are outliers that occur in multiple variable data sets. They can be identified using methods such as Mahalanobis Distance, which calculates the distance between a data point and the mean of the data set in a multivariate space.
3. **Collective Outliers:** These are outliers that occur in groups, rather than as individual data points. They can be identified using methods such as cluster analysis, which groups similar data points together and identifies clusters that are significantly different from the rest of the data set.

In this answer, we will explain Univariate Outliers in detail:
Univariate Outliers are outliers in a single variable data set. They can be identified using a range of methods, including the use of box plots, scatter plots, and Z-score calculations.
A box plot is a simple way to visualize univariate outliers. It displays the median, first and third quartiles, and outliers of a data set in a single plot. Outliers are typically identified as data points that fall outside of the whiskers of the box plot.
Z-score calculations can be used to identify outliers in a data set. Z-scores represent the number of standard deviations that a data point is from the mean of the data set. Outliers can be identified as data points with a Z-score that is significantly different from 0, typically Z-scores greater than 3 or less than -3 are considered outliers.

Q4-22-2023

Scatter plots can also be used to visualize univariate outliers. Outliers can be identified as data points that fall far from the majority of the data points in the scatter plot.

04-02-2023