Understanding Data" Assignment 1 Crash Course in Data Quiz Questions
Theory Part Questions :

Question: What is the purpose of exploratory data analysis (EDA)?
Answer: The purpose of EDA is to analyze and summarize the main characteristics of a dataset, often using graphical and statistical methods, to uncover patterns, trends, and relationships within the data.

Question: Why is it important to handle outliers during data analysis?
Answer: Outliers can significantly impact statistical measures and modeling, leading to inaccurate results. Handling outliers is crucial to ensure the robustness and reliability of data analysis.

Question: What does the term "pandas" refer to in Python data analysis?
Answer: "Pandas" is a Python library for data manipulation and analysis. It provides data structures like DataFrames and Series, making it easier to work with structured data.

Question: How can you check for missing values in a pandas DataFrame?
Answer: The isnull() function can be used to identify missing values in a pandas DataFrame. Applying sum() to the result gives the count of missing values in each column.

Question: What is the purpose of the seaborn library in Python?
Answer: Seaborn is a data visualization library in Python that provides a high-level interface for drawing attractive and informative statistical graphics. It works well with pandas DataFrames.

Question: What does the term "boxplot" represent in data visualization?
Answer: A boxplot, or box-and-whisker plot, provides a visual summary of the distribution of a dataset. It displays the median, quartiles, and potential outliers.

Question: What is the primary purpose of the pd.crosstab() function in pandas?
Answer: The pd.crosstab() function in pandas is used to compute a cross-tabulation of two or more factors, providing a frequency table of the variables' relationships.

Question: Why is it important to convert categorical variables into numerical representations in machine learning?
Answer: Machine learning algorithms typically require numerical input. Converting categorical variables into numerical representations allows these algorithms to process and learn from the data effectively.

Question: What is the role of the sns.pointplot() function in seaborn?
Answer: sns.pointplot() is used to show point estimates and confidence intervals as a function of one categorical variable and another variable. It is particularly useful for visualizing relationships between categorical and numeric variables.

Question: How can autocorrelation be assessed in time series data?
Answer: Autocorrelation in time series data can be assessed using a lag plot or by calculating the autocorrelation function (ACF). A positive correlation at a specific lag indicates a pattern in the data repeating at that interval.

Understanding Data" Assignment 1 Crash Course in Data Quiz Questions
Worked Example -1 Part Questions :

1. Question:
What is the mean age in the dataset?

A) 35.0
B) 38.08
C) 47.0
D) 60.0
Answer: B) 38.08

2. Question:
What does the skewness value for the 'age' column suggest about its distribution?

A) Positively skewed
B) Normally distributed
C) Negatively skewed
D) No skewness
Answer: D) No skewness

3. Question:
In the BMI range classification, which category corresponds to a BMI between 25.0 and 29.9?

A) Underweight
B) Healthy Weight
C) Overweight
D) Obese
Answer: C) Overweight

4. Question:
What percentage of people in the dataset have a BMI higher than 26?

A) 30.5%
B) 45.8%
C) 75.2%
D) 92.4%
Answer: B) 45.8%

5. Question:
What does the box plot for 'BMI' reveal about the data?

A) The median BMI is around 18.5.
B) The data has outliers above 47.
C) The data is left-skewed.
D) The majority have a BMI below 30.

Answer: B) The data has outliers above 47.

6. Question:
How many missing values are there in the 'bloodpressure' column?

A) 0
B) 10
C) 50
D) 1340
Answer: A) 0

7. Question:
What is the typical range for the 'claim' amount?

A) 0 - 5000
B) 5000 - 10000
C) 10000 - 20000
D) 20000 - 40000
Answer: C) 10000 - 20000

8. Question:
Which region has the highest number of insurance claims?

A) Northeast
B) Northwest
C) Southeast
D) Southwest
Answer: A) Northeast

9. Question:
What does the bar plot for 'Gender vs Claim' suggest?

A) Women have more claims than men.
B) Men have more claims than women.
C) Claims are equal for both genders.
D) Claims are not related to gender.
Answer: B) Men have more claims than women.

10. Question:
In the joint plot 'BloodPressure and Gender on Claim', what does the scatter plot reveal?

A) No correlation between blood pressure and claim.
B) Higher blood pressure correlates with higher claims.
C) Lower blood pressure correlates with higher claims.
D) Blood pressure has no effect on gender.
Answer: B) Higher blood pressure correlates with higher claims.

Understanding Data" Assignment 1 Crash Course in Data Quiz Questions
Worked Example -2 Part Questions :

1. What is the most common fuel type among the cars in the dataset?

A) Petrol
B) Diesel
C) CNG
D) Electric
Answer: B) Diesel
2. In which location are the car prices generally higher according to the dataset?

A) Mumbai
B) Coimbatore
C) Jaipur
D) Kolkata
Answer: B) Coimbatore
3. Which transmission type tends to have a higher average price for cars?

A) Automatic
B) Manual
Answer: A) Automatic
4. Among the given brands, which one has the highest average price for cars?

A) Lamborghini
B) Mercedes-Benz
C) Land Rover
D) Audi
Answer: A) Lamborghini
5. How many unique locations are there in the dataset?

A) 10
B) 11
C) 12
D) 13
Answer: B) 11
6. What is the correlation between the 'Year' and the 'Car Age' variables?

A) Positive
B) Negative
C) No correlation
Answer: A) Positive
7. Which brand has the highest number of unique models in the dataset?

A) Maruti
B) Hyundai
C) Honda

D) Toyota
Answer: A) Maruti
8. What is the average price of cars with a seating capacity of 7?

A) Rs. 8.5 Lakh
B) Rs. 9.7 Lakh
C) Rs. 10.3 Lakh
D) Rs. 11.2 Lakh
Answer: B) Rs. 9.7 Lakh
9. Which variable has the strongest positive correlation with the 'Power' variable?

A) Engine
B) Price
C) Mileage
D) Car Age
Answer: A) Engine
10. What percentage of entries have missing values in the 'New_Price' column?

A) 78.4%
B) 86.1%
C) 92.3%
D) 95.2%
Answer: B) 86.1%