Assignment-based Subjective Questions

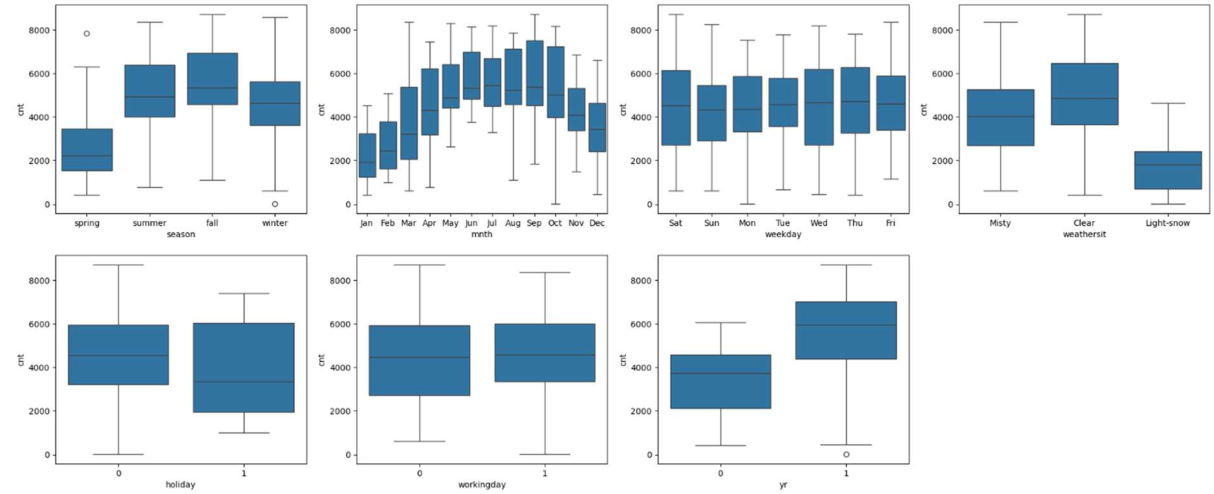
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Analysis on categorical columns using the boxplot and bar plot. Below are the few points we can infer from the visualization

1. Season Vs. Cnt: Fall has highest median, and Spring has lowest.
2. Month Vs. Count: Starting and ending of the year have the lowest count. The Count increased from January till September and then there is a fall in count till December,
3. Weekday Vs. Cnt: The average remained almost similar irrespective of the day.
4. Weather situation Vs. Cnt: Light snow has the lowest count and clear weather has boosted the count and has the highest count.
5. Holiday vs. Cnt: (1) Holiday has very low count and the most bookings are happening during non-holidays.
6. Working day Vs. Cnt: Both working and non-working day averages almost similar in terms of number of bookings.
7. Year Vs. Cnt: 2019 has highest count compared to year 2018.



Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

Dummy variables are created to represent categorical variables numerically.

drop_first=True helps in avoiding :

1. Multicollinearity and
2. To simplify coefficients.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

Temp has the highest correlation with Count.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

Here are some key aspects:

1. Ensured that the residuals follow a normal distribution.
2. Checked for low multicollinearity among the independent variables to ensure the model's stability.
3. Verified that a linear relationship exists between the independent variables and the dependent variable.
4. Ensured the residuals are independent of each other.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

Temp, Light-snow and year are the top 3 contributors in explaining the demand of shared bikes.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

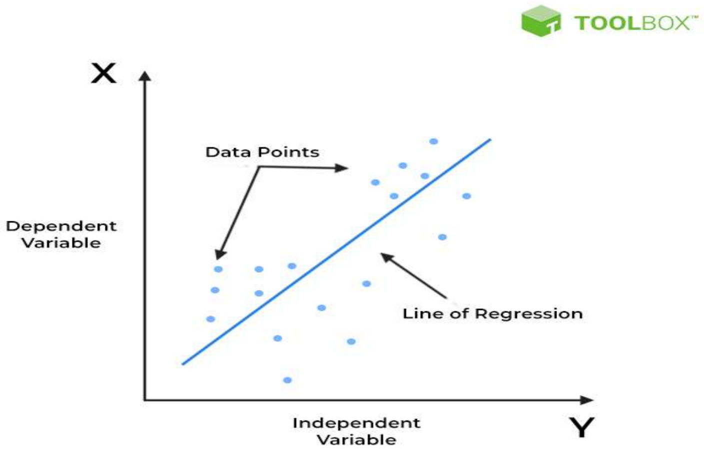
Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.

The independent variable is also the predictor or explanatory variable that remains unchanged due to the change in other variables. However, the dependent variable changes with fluctuations in the independent variable. The regression model predicts the value of the dependent variable, which is the response or outcome variable being analyzed or studied.

Thus, linear regression is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables such as sales, salary, age, product price, etc.

This analysis method is advantageous when at least two variables are available in the data, as observed in stock market forecasting, portfolio management, scientific analysis, etc.

A sloped straight line represents the linear regression model.



X-axis = Independent variable

Y-axis = Output / dependent variable

Line of regression = Best fit line for a model

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's quartet, introduced by statistician Francis Anscombe in 1973, is a set of four datasets, each containing 11 data points (x, y pairs). Despite having identical statistical properties, these datasets display strikingly different patterns when visualized as scatter plots. Anscombe created these datasets to emphasize the dangers of relying solely on summary statistics and to highlight the critical role of data visualization in understanding data.
The importance of Anscombe's quartet is:

1. **Revealing the Power of Visualization:** Visualization uncovers patterns, trends, and anomalies in data that summary statistics alone cannot detect.
2. **Exposing the Limitations of Statistics:** Measures such as means, variances, and correlations can oversimplify or obscure complex relationships within datasets.
3. **Highlighting Contextual Understanding:** Although datasets may appear statistically identical, their visual representations demand different interpretations and analyses.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's r is a numerical measure that summarizes the strength and direction of the linear relationship between two variables. When both variables increase or decrease together, the correlation coefficient is positive. Conversely, when one variable increase while the other decreases, the correlation coefficient is negative.
The Pearson correlation coefficient (r) ranges from -1 to +1:

- r=0: Indicates no linear association between the variables.
- r>0: Represents a positive association, where an increase in one variable corresponds to an increase in the other.
- r<0: Represents a negative association, where an increase in one variable corresponds to a decrease in the other.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

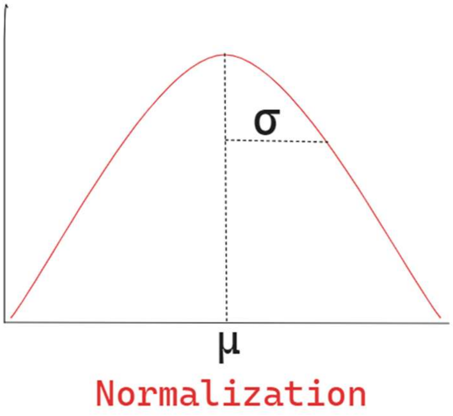
Scaling: Scaling is a preprocessing method used to modify the range and distribution of numerical features in a dataset, ensuring they fall within a specific scale. This prevents any single feature from dominating others due to differences in magnitude. It is especially crucial for machine learning algorithms that are sensitive to the magnitude of data, such as gradient-based models like logistic regression, SVMs, and neural networks.

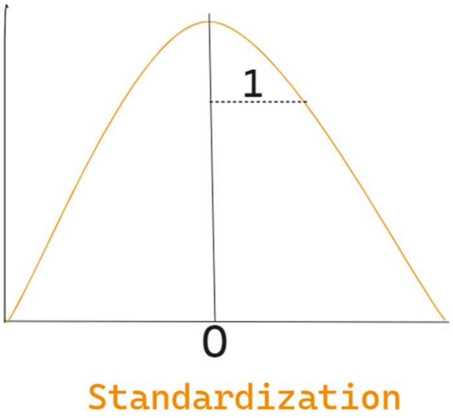
Scaling is performed to ensure that numerical features in a dataset are on a similar range, preventing features with larger magnitudes from dominating those with smaller magnitudes. It improves the performance and efficiency of machine learning models by allowing algorithms that rely on distance measures or gradients, such as K-Nearest Neighbors, Support Vector Machines, and neural networks, to function effectively. Additionally, scaling accelerates convergence in optimization processes like gradient descent and enhances numerical stability in computations. By standardizing the contribution of each feature, scaling ensures that all variables are treated equally, making the model's performance more consistent and reliable. It also minimizes the impact of units or scales of measurement, which is especially important when working with features measured in different units.

Difference between Normalized scaling and standardized scaling:

Aspect	**Normalization (Min-Max Scaling)**	**Standardization (Z-Score Scaling)**
Definition	Transforms data to fit within a specific range, typically [0,1][0, 1][0,1].	Centers data around a mean of 0 and scales it to have a standard deviation of 1.

Aspect	**Normalization (Min-Max Scaling)**	**Standardization (Z-Score Scaling)**
Formula	X′=X−XminXmax−XminX' = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}X′=Xmax−XminX−Xmin	X′=X−μσX' = \frac{X - \mu}{\sigma}X′=σX−μ
Range	Produces values within a defined range, usually between 0 and 1 (or a specified range like [−1,1][-1, 1][−1,1]).	Produces values that are centered at 0, with most falling between -3 and +3 standard deviations.
Sensitivity to Outliers	Highly sensitive to outliers, as it scales based on the minimum and maximum values in the data.	Less sensitive to outliers since it uses mean and standard deviation, though extreme outliers can still have some influence.
Use Case	Ideal for bounded data or when features have different units (e.g., percentages, image pixel values).	Suitable for unbounded data and for models that assume normally distributed features.
Examples of Usage	Used in neural networks, K-Nearest Neighbors (KNN), and when combining features with different ranges.	Used in algorithms like logistic regression, SVMs, and PCA, which rely on normal distribution.
Effect on Data Distribution	Does not change the shape of the data's distribution; only shifts and scales it to the defined range.	Adjusts the data's distribution to standardize it, often assuming it follows a normal distribution.





Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

The Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity among the predictor variables in a regression model.

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

When Xi is perfectly collinear with other predictors, Ri2=1, meaning the predictor can be predicted perfectly by other variables. In those cases, VIF becomes infinite.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

A Q-Q (Quantile-Quantile) plot is a graphical technique used to compare the distribution of a dataset with a theoretical distribution, such as the normal distribution. It plots the quantiles of the dataset against the quantiles of the chosen theoretical distribution. If the points on the plot align closely with a straight line, it suggests that the dataset follows the theoretical distribution. Deviations from the line indicate discrepancies between the observed and expected distributions.

Use and Importance of a Q-Q Plot in Linear Regression:

1. Checking Normality Assumption: One of the key assumptions in linear regression is that the residuals are normally distributed. This assumption is critical because it impacts the validity of hypothesis tests (like t-tests for model coefficients) and the accuracy of confidence intervals. A Q-Q plot helps visually assess whether this assumption holds.
2. Identifying Non-Normality: If the residuals deviate significantly from a normal distribution (i.e., if the points on the Q-Q plot don't follow the straight line), this suggests a violation of the normality assumption. This can undermine the reliability of the regression model’s statistical tests.
3. Spotting Outliers: Outliers in the residuals will appear as points that are far away from the reference line on a Q-Q plot. Identifying these outliers is crucial, as they can disproportionately affect the model's results and predictions. The Q-Q plot serves as a useful tool for detecting these outliers.
4. Model Diagnostics: A Q-Q plot is a valuable diagnostic tool in regression analysis. If the residuals do not follow a normal distribution, it may indicate:
   - Model misspecification (e.g., missing important variables or incorrect model form).
   - The need for data transformations (such as taking the log of variables or adding polynomial terms).
   - Possible issues with nonlinearity or heteroscedasticity (non-constant variance of residuals).