Assignment-based Subjective Questions

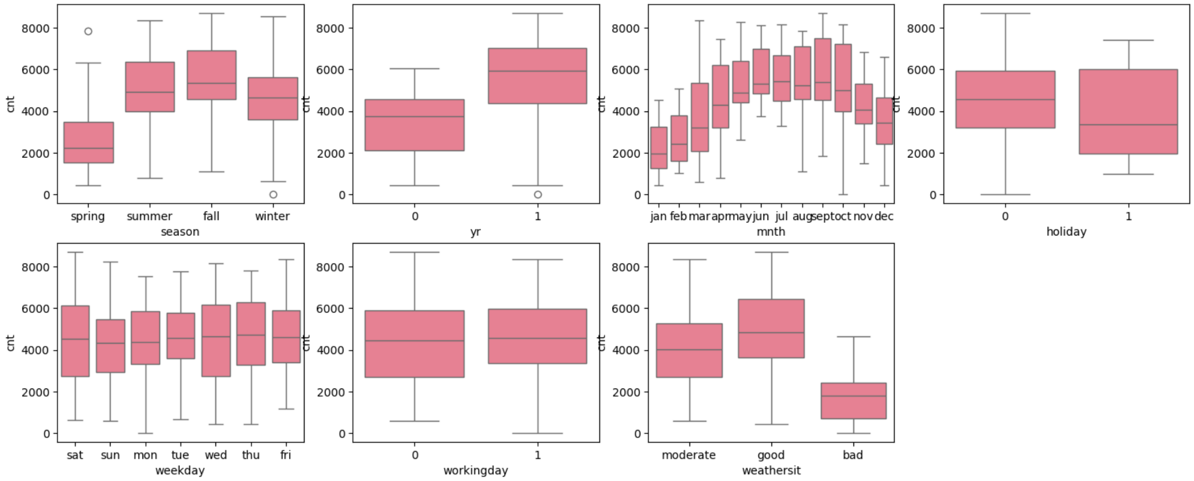
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?   (Do not edit)

Total Marks: 3 marks (Do not edit)

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

Following is the effect on the dependent variable of categorical variables

- Bike sharing count is more in fall and summer season
- Bike sharing count is more in 2019 compared to 2018
- Bike sharing count is more in the month of September and June
- Bike sharing count is more on non holiday days
- Bike sharing count is more on weekdays compared to weekends
- Bike sharing count is more on working days compared to non working days
- Bike sharing count is more on good weather days compared to bad weather days



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)

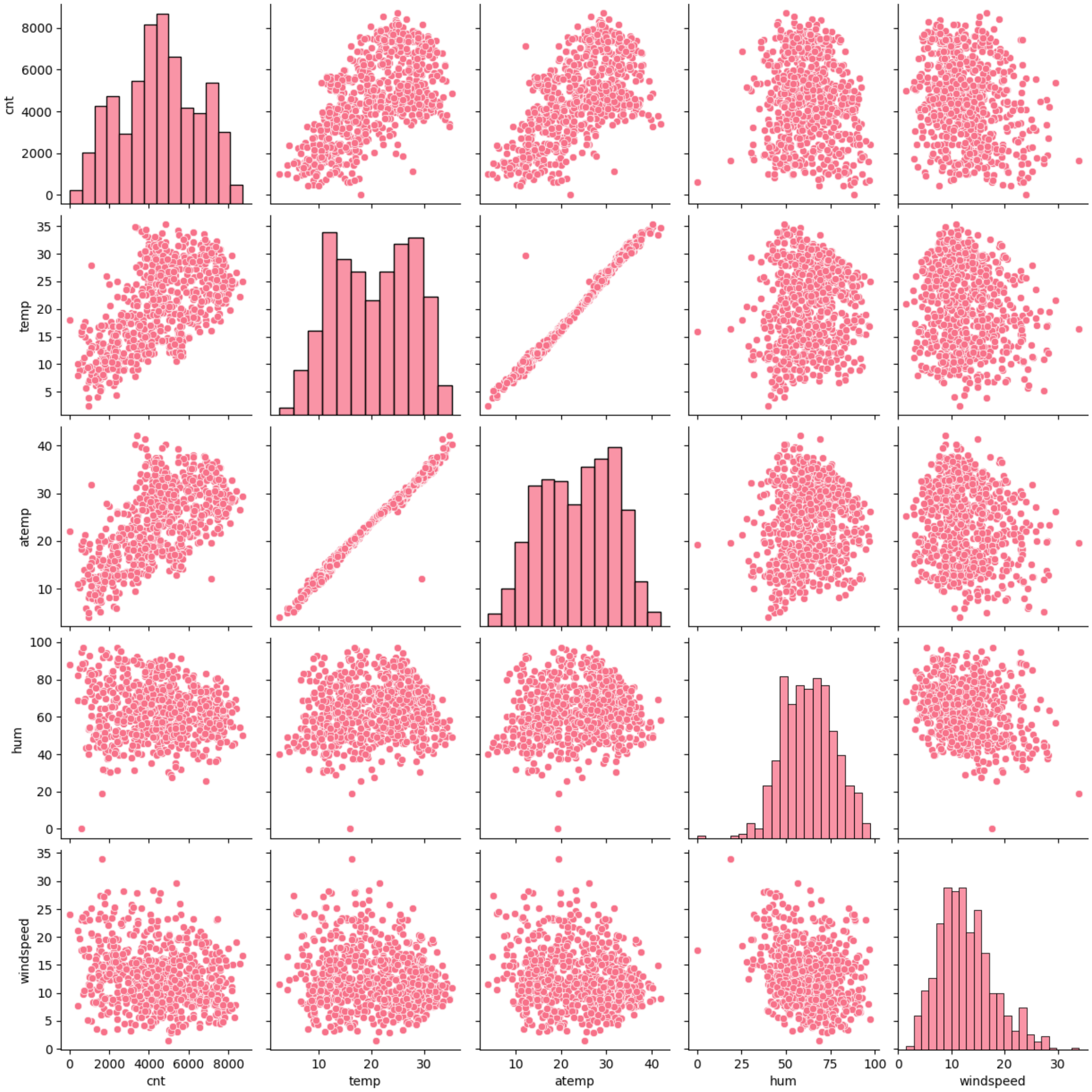
- If we do not set the parameter `drop_first = True`, then `n` dummy variables will be created. This results in multicollinearity because the `n` dummy variables are correlated with each other, leading to what is known as the Dummy Variable Trap.
- To avoid this issue, it is important to create only `k-1` dummy variables, which allows us to eliminate the extra column that would otherwise be created during the process.

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 and atemp has the highest correlation with the target variable



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)

- Establishing a linear relationship between response and predictor variables is essential.
- Ensuring normality of the error distribution, which means the error terms should follow a normal distribution, is also important.
- The errors must exhibit constant variance, known as homoscedasticity.
- Additionally, it is important to minimize multicollinearity among features, indicated by low Variance Inflation Factor (VIF).

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)

Top 3 features are

- Temp
- Season
- Year

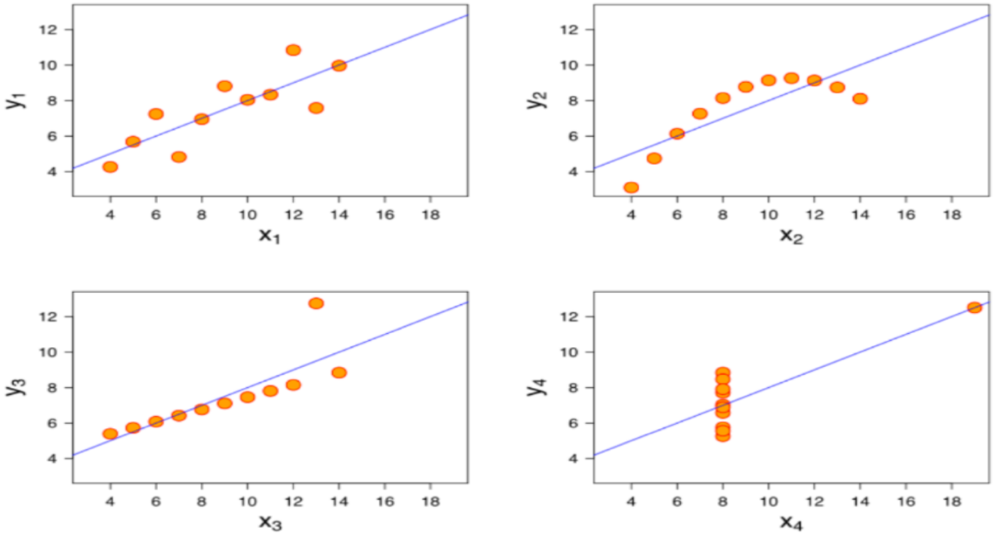
General Subjective Questions

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)

- The regression technique aims to establish a linear relationship between a dependent variable and one or more independent variables. There are two types of linear regression: simple linear regression and multiple linear regression.
- Simple linear regression is used when a single independent variable is employed to predict the value of the target variable. In contrast, multiple linear regression involves the use of multiple independent variables to predict the numerical value of the target variable. The linear line that represents the relationship between the dependent and independent variables is known as the regression line.
- The following is an example of a resulting linear regression equation: Y=β0+β1X1+β2X2+...+βpXp+ϵ Where y is the dependent variable, and X1, X2, and so on, are the explanatory variables. The coefficients (β1, β2, and so on) explain the correlation of the explanatory variables with the dependent variable. The sign of the coefficients (+/-) designates whether the variable is positively or negatively correlated. β0 is the intercept that indicates the value of the dependent variable assuming all explanatory variables are 0.

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 consists of four data sets that share nearly identical simple descriptive statistics. However, each dataset contains unique characteristics that can mislead regression models if they are built without thorough examination. The distributions of the datasets vary significantly, and when plotted on scatter plots, they reveal distinct patterns. Anscombe's Quartet was created to emphasize the importance of visualizing data before conducting analysis and model building, as well as to highlight how outliers can influence statistical properties. Despite having similar statistical summaries, the four data sets produce different visual representations, underscoring the need for careful inspection in statistical analysis.



.

- The first dataset fits a linear regression model because it exhibits a clear linear relationship between X and Y.
- The second dataset does not show a linear relationship between X and Y, indicating that it is unsuitable for a linear regression model.
- The third dataset contains some outliers that cannot be addressed by a linear regression model.
- The fourth dataset includes a high leverage point, which leads to a significantly high correlation coefficient.

In conclusion, regression algorithms can be misleading, so it is essential to perform data visualization before constructing a machine learning model.

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)

- In statistics, the Pearson correlation coefficient is a measure that quantifies the linear correlation between two sets of data. It is calculated as the ratio of the covariance of the two variables to the product of their standard deviations. This formula normalizes the covariance, resulting in a value that always falls between −1 and 1.
- It's important to note that, like covariance, the Pearson correlation coefficient only captures linear relationships between variables and does not account for other types of correlations or relationships.
- For instance, one might expect that the age and height of a group of high school teenagers would yield a Pearson correlation coefficient that is significantly greater than 0 but less than 1, as a value of 1 would indicate a perfect correlation, which is unrealistic.

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 refers to the process of transforming data to fit within a specific range. It is a crucial step in data pre-processing that helps ensure algorithms perform efficiently. When data is collected, its features can vary in magnitude, units, and range. If scaling is not applied, algorithms may disproportionately prioritize features with higher values, leading to inaccurate modelling.

Difference Between Normalization and Standardization

Normalization and **Standardization** are two different techniques for scaling data:

- **Normalization** uses the minimum and maximum values of the features. It is typically applied when the features are on different scales. Normalization scales values to a range between (0, 1) or (-1, 1). One drawback is that normalization is sensitive to outliers.

- **Standardization** uses the mean and standard deviation of the features. This method ensures that the data has a mean of zero and a standard deviation of one. Standardization is preferred when the data has a normal distribution and is not influenced by outliers. It's not restricted to a particular range.

In summary, normalization is referred to as scaling normalization, while standardization is often called Z-score normalization. Normalization is ideal when the distribution of the data is unknown, whereas standardization is suitable for data that 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 value of the Variance Inflation Factor (VIF) becomes infinite when there is a perfect correlation between two independent variables. In this situation, the R-squared value is 1, which results in VIF being computed as infinity, since VIF equals 1/(1-R^2). This indicates that there is a problem of multicollinearity, and one of these variables needs to be removed to create a valid regression model.

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 plot, or Quantile-Quantile plot, is a graphical method used to compare two probability distributions by plotting their quantiles against each other. This tool helps us assess whether a given set of data might originate from a specific theoretical distribution, such as a Normal, Exponential, or Uniform distribution.

Additionally, a Q-Q plot can determine whether two distributions are similar. When the distributions are similar, the Q-Q plot will tend to be more linear. This linearity can be further assessed using scatter plots. Furthermore, linear regression analysis assumes that all variables are multivariate normal, which can be evaluated using histograms or Q-Q plots.

Importance of Q-Q Plots in Linear Regression:
In linear regression, when we have training and test datasets, a Q-Q plot can be created to confirm whether both datasets come from populations with the same distribution.

Advantages of Q-Q Plots:
- They can be applied to sample sizes of any scale.
- They allow for the detection of various distributional aspects, such as shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

Use of Q-Q Plots on Two Datasets:
A Q-Q plot can be used to check:
- If both datasets come from populations with a common distribution.
- If both datasets share a common location and common scale.
- If both datasets have similar distribution shapes.
- If both datasets exhibit similar tail behaviour.
