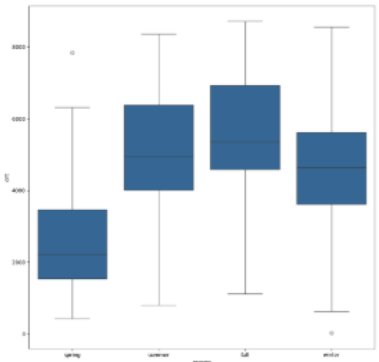
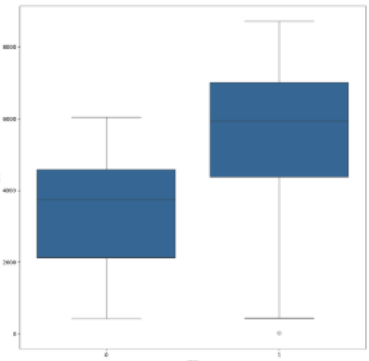
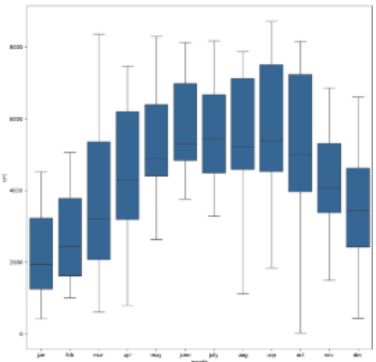
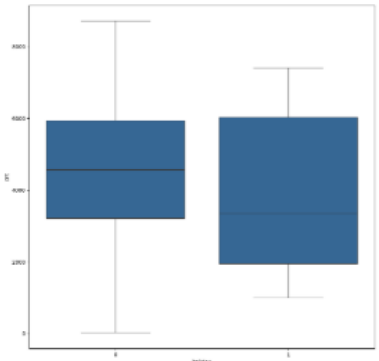
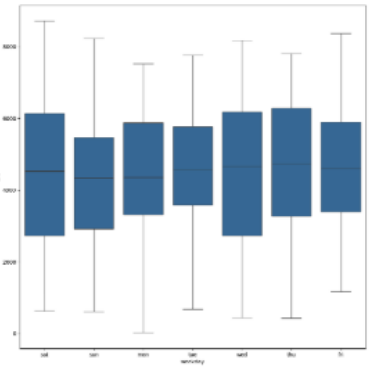
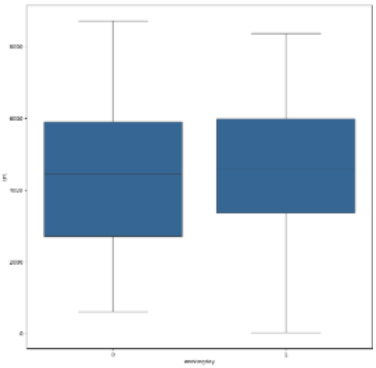
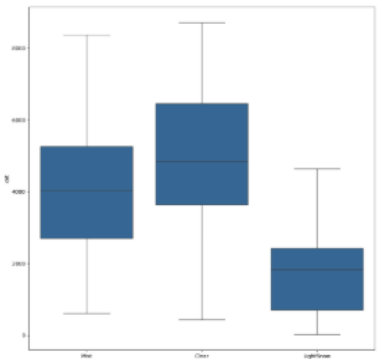
Assignment-based Subjective Questions

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)

Observations :¶

- We can observe that there are no outliers to handle
- Fall season has highest demand
- In year 2019 the demand has grown
- Demand grows from jan to july and then decreases from aug to dec
- In case of holiday the demand decreases
- Weekday has no effect on demand
- Clear weather situation has highest demand

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

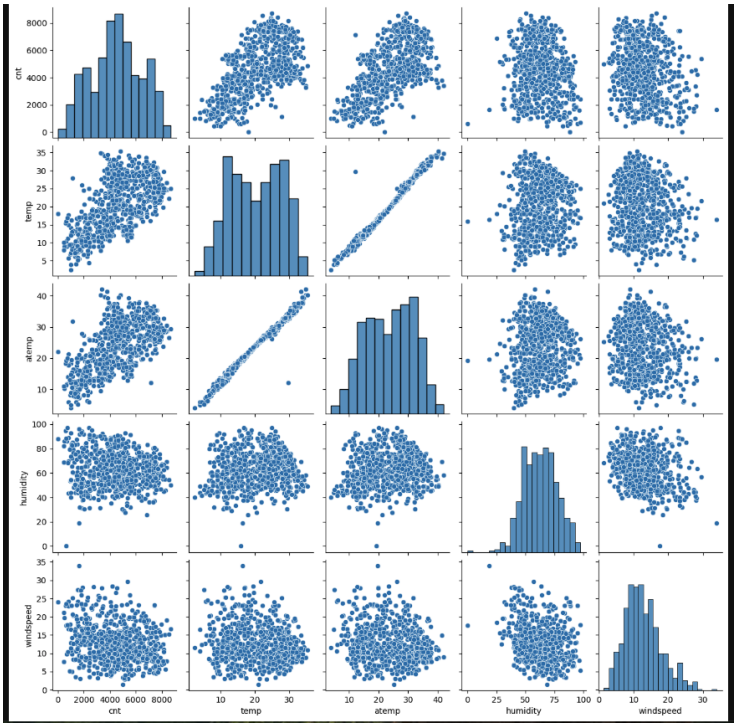
Total Marks: 2 marks (Do not edit)

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

The purpose of a dummy variable is to represent a categorical variable with 'n' levels by creating 'n-1' new columns, where each column indicates the presence of a specific level using a binary value (0 or 1). The parameter drop_first=True is applied to ensure that one level is omitted, resulting in 'n-1' columns. This approach minimizes correlation among the dummy variables**.**
For example, if there are three levels, *drop_first=True* will drop the first column.

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)

The variables 'temp' and 'atemp' show the highest correlation with the target variable 'cnt' compared to the other variables.



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)

Linear regression model is validated on

a. Linear relationship between dependent and independent variable
b. Homoscedasticity: The residuals have constant variance across all levels of the predicted values
c. The residuals are approximately normally distributed.
d. No multicollinearity among independent variables

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 year, season and holiday

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)

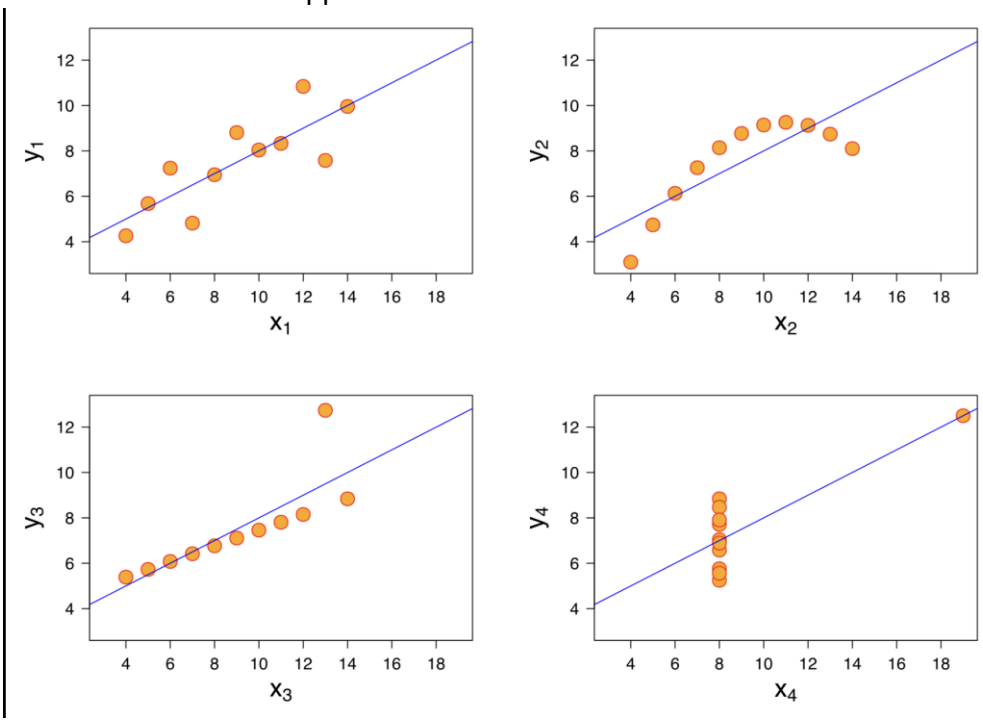
Linear regression is a predictive modeling technique used to determine the relationship between a dependent variable (target) and one or more independent variables (predictors). It establishes a linear relationship, showing how the dependent variable changes with variations in the independent variable(s). When there is only one predictor (x), the model is called simple linear regression, whereas with multiple predictors, it is referred to as multiple linear regression.
The linear regression model generates a straight sloped line that describes the relationship between the variables. This regression line can exhibit either a positive or a negative linear relationship. The primary objective of the linear regression algorithm is to determine the optimal values of coefficients a0 and a1 to produce the best-fit line, minimizing the error.

Techniques like Residual Sum of Squares (RSS), Mean Squared Error (MSE), or the cost function are employed to calculate the best values for a0 and a1, ensuring the line fits the data points with the least error.

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 refers to a group of four datasets that are nearly identical in their basic descriptive statistics, such as mean, variance, and correlation. However, these datasets exhibit distinct patterns and distributions when visualized through scatter plots. Despite sharing similar statistical summaries, the peculiarities within each dataset can mislead a regression model if built without proper analysis.
The quartet was created to emphasize the importance of visualizing data before conducting analysis or building models. It demonstrates how outliers and unique data patterns can

significantly influence statistical properties and model performance, even when the numerical summaries appear consistent across datasets.



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 commonly known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a measure that quantifies the linear relationship between two variables.

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

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 your data so that it fits within a specific range or scale. It is an essential step in data pre-processing, which helps to ensure that the data is appropriately

scaled, enabling algorithms to perform calculations more efficiently. Collected data often includes features with varying magnitudes, units, and ranges. Without scaling, algorithms may give more weight to features with higher values and ignore others, leading to inaccurate modeling.

Difference between Normalization and Standardization:

1. **Normalization** uses the minimum and maximum values of the features, while **Standardization** uses the mean and standard deviation for scaling.
2. **Normalization** is applied when features have different scales, whereas **Standardization** ensures a zero mean and unit standard deviation.
3. **Normalization** scales values within a range, typically (0,1) or (-1,1), while **Standardization** does not restrict values to a specific range.
4. **Normalization** is sensitive to outliers, whereas **Standardization** is less affected by outliers.
5. **Normalization** is preferred when the distribution of data is unknown, while **Standardization** is suited for data that follows a normal distribution.
6. **Normalization** is often referred to as "scaling normalization," whereas **Standardization** is also known as "Z-score normalization."

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) helps assess the relationship between one independent variable and all the other independent variables in a model. The VIF is calculated using a specific formula.

A VIF value greater than 10 is considered high, and a value above 5 should also be carefully examined.

A very high VIF indicates a strong correlation between two independent variables. In cases of perfect correlation, the R^2 value will be 1, which results in a VIF approaching infinity ($1/(1 - R^2) \rightarrow \infty$). To address this issue, one of the correlated variables should be removed from the dataset to resolve the perfect multicollinearity.

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 tool used to compare two probability distributions by plotting their quantiles against each other. It helps determine if a dataset follows a specific theoretical distribution, such as Normal, Exponential, or Uniform.

The Q-Q plot can also be used to assess whether two distributions are similar. If the distributions are

similar, the plot will appear more linear. The assumption of linearity is best verified through scatter plots. Additionally, in linear regression, the assumption that all variables follow a multivariate normal distribution can be checked using histograms or Q-Q plots.

Importance of Q-Q Plot in Linear Regression: In linear regression, when working with both training and testing datasets, a Q-Q plot can confirm whether both datasets come from populations with the same distribution.

Advantages:

- It can be used with smaller sample sizes.
- It helps detect various distributional aspects, such as shifts in location, scale, symmetry, and the presence of outliers.

Uses of Q-Q Plots for Two Datasets:

- To determine if both datasets come from populations with the same distribution.
- To check if the datasets have similar location and scale.
- To compare the shape of the distribution between datasets.
- To assess the tail behavior of the distributions in both datasets.
