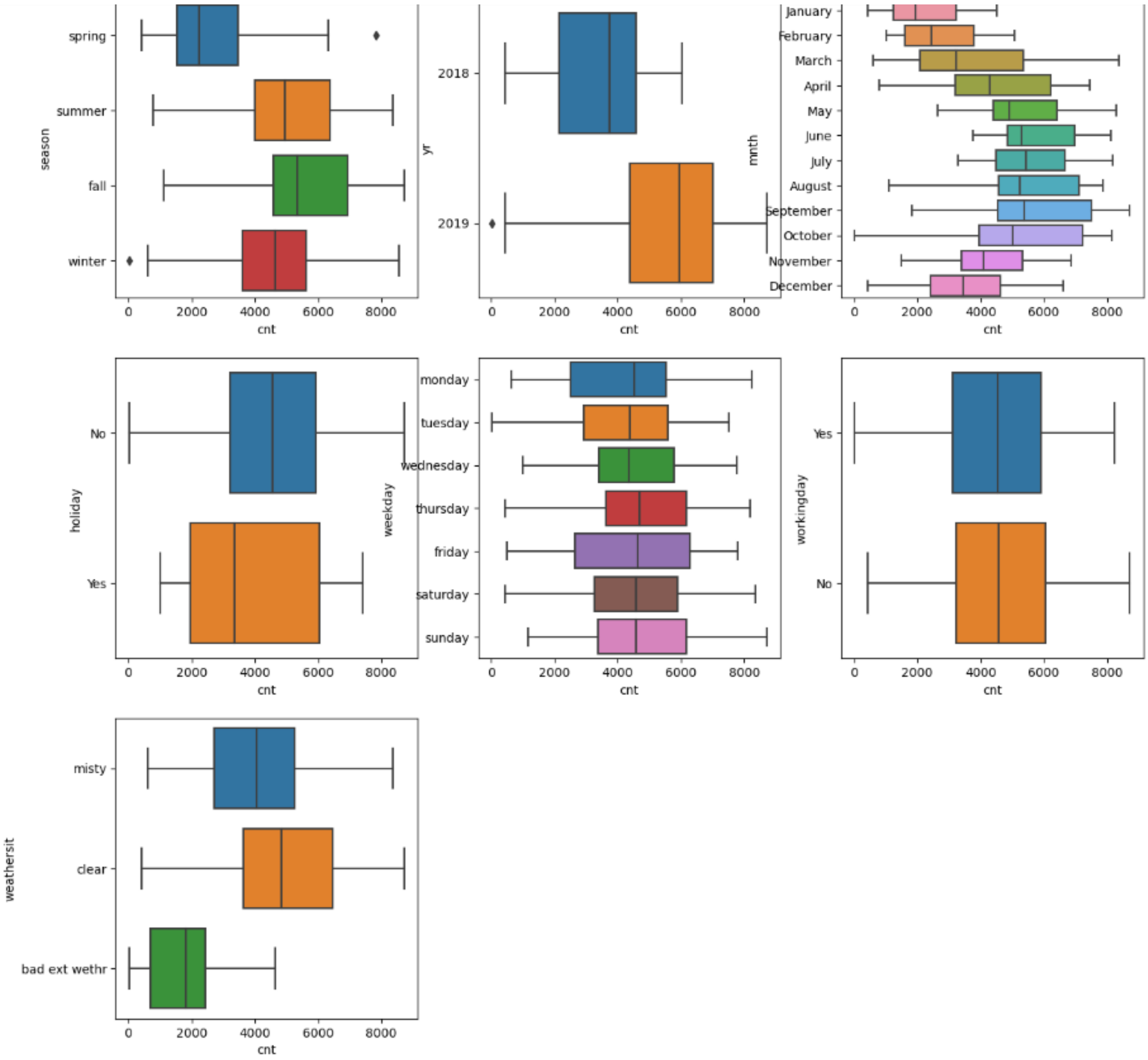
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans.



1. Season: 3:fall has highest demand for rental bikes

2. I see that demand for next year has grown

3. Demand is continuously growing each month till June. September month has highest demand. After September, demand is decreasing

4. When there is a holiday, demand has decreased.

5. Weekday is not giving clear picture abount demand.

6. The clear weathershit has highest demand

7. During September, bike sharing is more. During the year end and beginning, it is less, could be due to extereme weather conditions

2. Why is it important to use drop_first=True during dummy variable creation?

Ans. drop_first=True helps in reducing the extra column created during the dummy variable creation and hence avoid redundancy of any kind.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans. The temp variable has the highest correlation with the target variable

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans. Validated the assumptions of linear regression by checking the VIF, error distribution of residuals and linear relationship between the dependent variable and a feature variable.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

- **Ans. Temperature (temp)** - A coefficient value of '0.4587' indicated that a unit increase in temp variable increases the bike hire numbers by 0.4587 units.
- **yr_2019** - A coefficient value of '0.2381' indicated that a unit increase in yr_2019 variable increases the bike hire numbers by 0.2381 units.
- **season_spring** - A coefficient value of '-0.1115' indicated that, w.r.t Weathersit1, a unit increase in season_spring variable decreases the bike hire numbers by -0.1115 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

ans. Linear regression is a type of supervised machine learning algorithm that learns the linear relationship between a dependent variable and one or more independent variables. The dependent variable is also called the target or the output variable, and the independent variables are also called the features or the input variables. The goal of linear regression is to find the best line or function that can fit the data points and minimize the error between the predicted values and the actual values.

There are two main types of linear regression: simple linear regression and multiple linear regression. Simple linear regression involves only one independent variable and one dependent variable, and it assumes that the relationship between them is linear, which means it can be represented by a straight line. Multiple linear regression involves more than one independent variable and one dependent variable, and it assumes that the relationship between them is linear, which means it can be represented by a plane or a hyperplane.

The equation of a simple linear regression line is:

$$y=a0+a1x+\epsilon$$

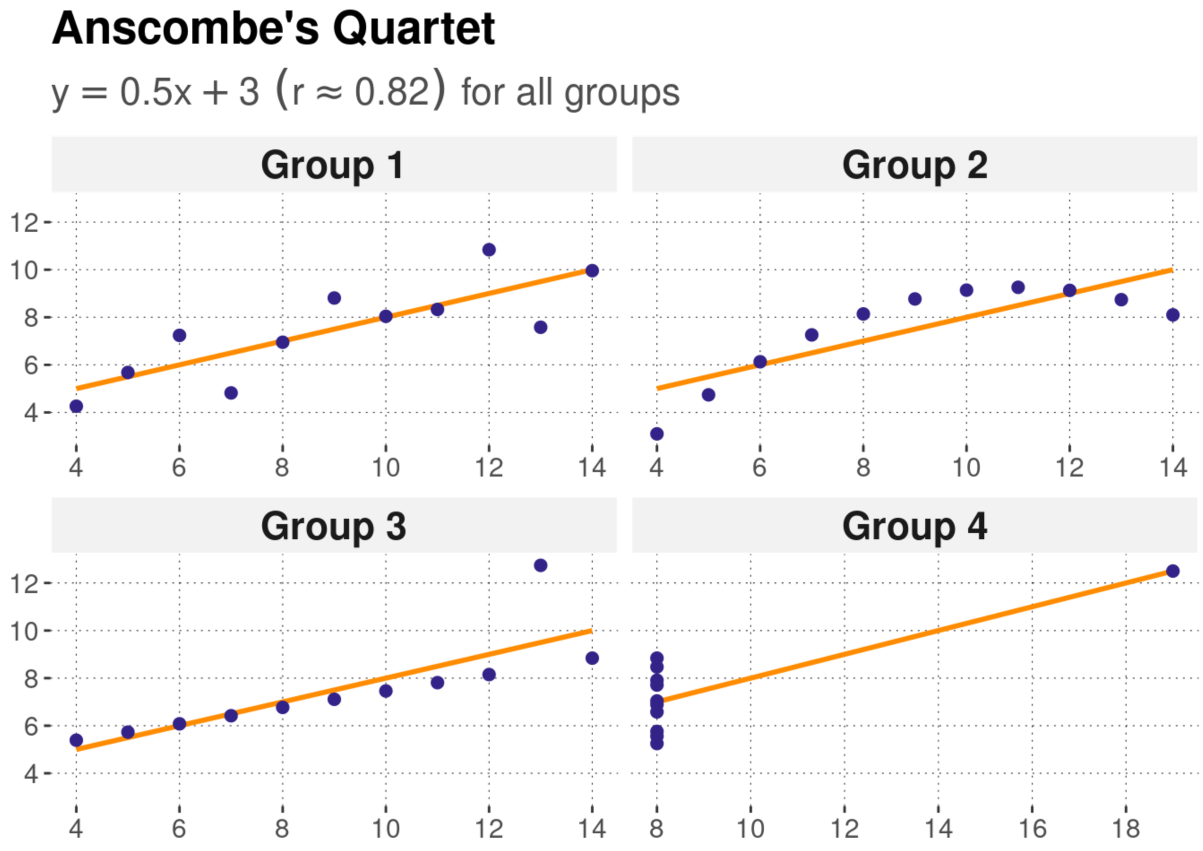
where y is the dependent variable, x is the independent variable, a0 is the intercept, a1 is the slope, and ϵ is the random error. The intercept is the value of y when x is zero, and the slope is the rate of change of y with respect to x. The random error is the difference between the actual value of y and the predicted value of y by the line.

The equation of a multiple linear regression line is:

$$y=a0+a1x1+a2x2+...+anxn+\epsilon$$

2. Explain the Anscombe's quartet in detail.

ans. Anscombe's quartet is a set of four datasets that have nearly identical summary statistics, such as mean, variance, correlation, and linear regression, but have very different distributions and patterns when plotted on a graph.



For all four datasets, the summary statistics are:

- Mean of x: 9
- Mean of y: 7.50
- Variance of x: 11
- Variance of y: 4.12
- Correlation between x and y: 0.816
- Linear regression line: y = 3 + 0.5x
- Coefficient of determination (R-squared): 0.67

However, when we plot the four datasets on a graph, we can see that they have very different shapes and trends. The first dataset (top left) appears to have a simple linear relationship between x and y, with a positive correlation and a small error. The second dataset (top right) has a curved or quadratic relationship between x and y, with a positive correlation but a large error. The third dataset (bottom left) has a linear relationship between x and y, with a positive correlation and a small error, except for one outlier that lowers the correlation and affects the regression line. The fourth dataset (bottom right) has no relationship between x and y, with a zero correlation and a large error, except for one influential point that increases the correlation and determines the regression line.

3. What is Pearson's R?

Ans. Pearson's R, also known as the Pearson correlation coefficient, is a measure of how strong and in what direction the linear relationship is between two variables. *It is a number between -1 and 1, where -1 means a perfect negative correlation, 0 means no correlation, and 1 means a perfect positive correlation*

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans. Scaling is a way of changing the data to have the same scale. This makes the data easier to compare and use, especially when the data has different units, sizes, or shapes. Scaling can also make some machine learning algorithms work better and more accurately, because they care about the scale of the data.

There are many ways of scaling, but two common ones are normalized scaling and standardized scaling. Normalized scaling, or min-max scaling, makes the data between 0 and 1. It does this by taking away the smallest value and dividing by the difference between the biggest and smallest values. Standardized scaling, or z-score scaling, makes the data have 0 mean and 1 standard deviation. It does this by taking away the average value and dividing by the standard deviation.

The difference between normalized scaling and standardized scaling is that normalized scaling keeps the data shape, but may not show some extreme data. Standardized scaling shows the extreme data, but may not keep the data shape. Normalized scaling is good for data that has a known range. Standardized scaling is good for data that has an unknown range or outliers.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans. VIF is short for variance inflation factor, and it shows how much multicollinearity there is among the independent variables in a regression model. Multicollinearity is when some of the independent variables are very related to each other, which can make the regression coefficients less accurate and stable. The VIF is calculated by:

$$VIF = \frac{1}{1 - R^2}$$

where R2 is how well one independent variable can be explained by the other independent variables. The VIF is infinite when R2 is 1, which means that one independent variable is exactly the same as the other independent variables. This means that there is complete multicollinearity, and the regression model is invalid. To prevent infinite VIF values, one should look for multicollinearity among the independent variables and remove or merge the variables that are very related to each other. There are many ways to find and fix multicollinearity, such as correlation matrix, variance decomposition, principal component analysis, ridge regression, etc.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans. A Q-Q plot is a graphical tool that compares the distribution of a data set with a theoretical distribution, such as a normal distribution. It plots the quantiles of the data set against the quantiles of the theoretical distribution. A quantile is a value that divides the data into equal proportions. For example, the median is the 50th quantile that splits the data into two halves.

A Q-Q plot is useful for checking the assumption of normality in linear regression. Normality is an assumption that the error terms in the regression model follow a normal distribution. This assumption affects the validity and reliability of the regression coefficients and the hypothesis tests. A Q-Q plot can help to visually assess if the data is approximately normal or not. If the data is normal, the points in the Q-Q plot should lie close to a straight line. If the data is not normal, the points may deviate from the line in a curved or skewed pattern.

A Q-Q plot can also help to identify outliers or extreme values in the data. Outliers are values that are very different from the rest of the data. They can affect the accuracy and stability of the regression model. A Q-Q plot can show if there are any outliers by looking for points that are far away from the line or the other points.