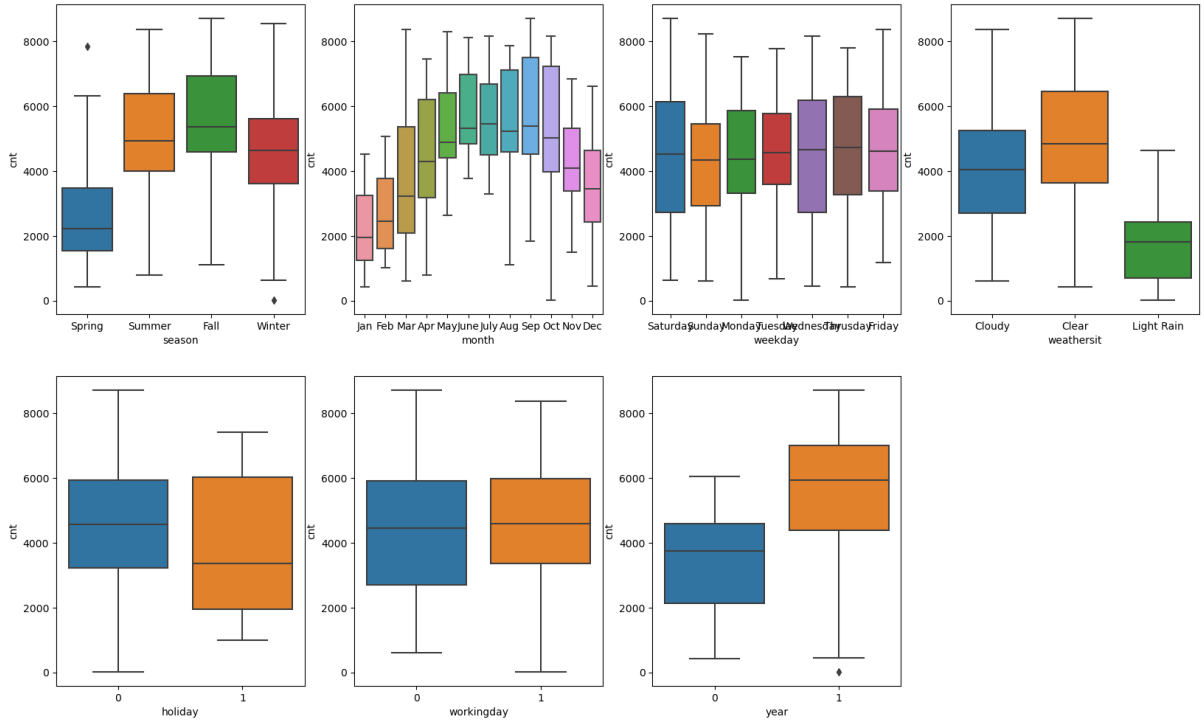
Assignment – Based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

Answer: –

I have done the analysis on the basis of Categorical Columns using Boxplot. Below are the few points which is observed from the Visualization –



- Bookings tend to peak during the fall season. And the number of counts has been increased in 2019 as compared 2018.
- The months of June, July, August, and September consistently see a high number of bookings. In 2019, there was a higher number of booking counts than in 2018.
- The number of bookings tends to be high when the weather is clear. And There were more bookings in 2019 compared to 2018.
- We noticed that booking numbers tend to be lower when it's not a holiday, which makes sense as people may prefer to stay home and spend time with family during holidays. But, in 2019 the number of bookings is more as compared to 2018.
- The count remains almost the same whether it's a working or non-working day. However, there has been an increase in the number of counts from 2018 to 2019.
- The overall booking numbers showed a significant increase in 2019.

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

Answer: –

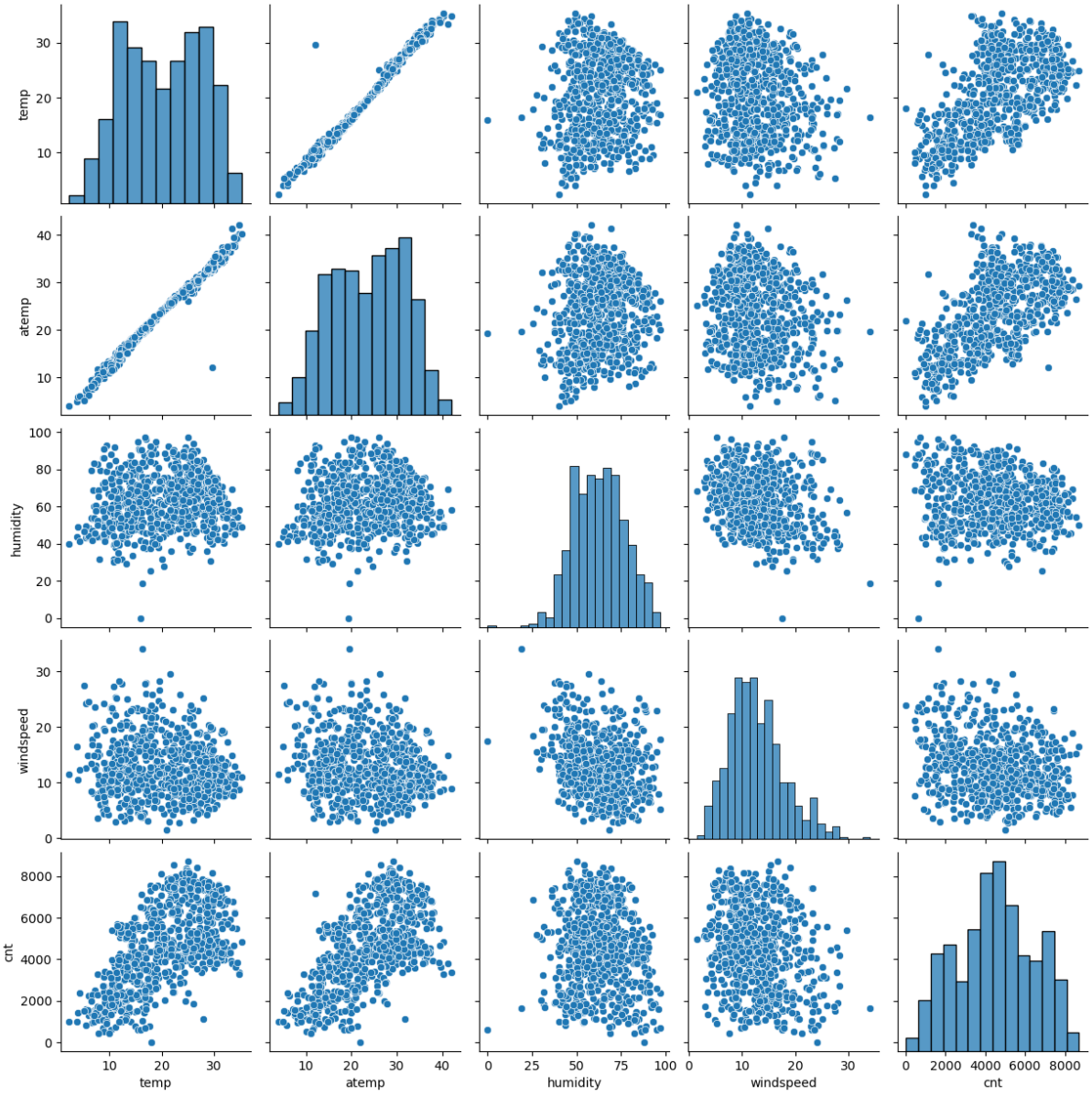
While creating dummy variable it is important to use drop_first=True to avoid the multicollinearity issues in Multiple Linear Regression Analysis and ensure the independence of predictor variables.

Having k columns for k levels of a categorical variable is useful, but since one level introduces redundancy with a separate column, it's best to drop one column and retain k-1 columns to represent k levels effectively.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

Answer: –

The pair plot indicates that the variables "temp" and "atemp" have 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? (3 marks)**

Answer: –

We have validated the assumptions of Linear Regression after building the model on the training set are as follows: -

- Error Terms: - Error Terms should be normally distributed.
- Independence in Residuals: - No Autocorrelation
- Homoscedasticity: - There is not any kind of visible pattern.
- Multicollinearity: - There should be insignificant multicollinearity on variables.
- Linear Relation: - Linear Relation should be visible on variables.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

Answer: -

Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are as follows: -

- year
- workingday
- temp

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

Answer: -

Linear regression is a mostly used in statistical technique for modelling the relationship between the dependent variables and the independent variables.

Linear regression assumes that the relationship between the predictor variables and the response variable is linear.

The main goal of Linear regression is to find the best fit line or hyperplane for multiple predictors. That explains the relationship between the independent variables (X) and the dependent variables (y).

Mathematically, the multiple linear regression model can be represented as:

$\mathrm{y} = \beta_0 + \beta_1\mathrm{X}_1 + \beta_2\mathrm{X}_2 + . \ . \ . \ . + \beta_\mathrm{n}\mathrm{X}_\mathrm{n}$

Where,

$\mathrm{y} =$ dependent variable

$\mathrm{X}_1, \mathrm{X}_2, . \ . \ . \ . , \mathrm{X}_\mathrm{n} =$ Independent Variable / predictors

$\beta_1, \beta_2, \ . \ . \ . \ . , \beta_\mathrm{n} =$ coefficients / slopes

$\beta_0 =$ intercept

Assumptions:

Assumptions about the dataset that is made by Linear Regression model are as follows –

- **Linearity**: There should be a linear relationship between the independent variables and the dependent variable.
- **Independence**: The observations should be independent of each other.
- **Homoscedasticity**: The variance of the residuals should be constant across all levels of the independent variables.
- **Normality of Residuals**: The residuals should follow a normal distribution.
- **No Multicollinearity**: The independent variables should not be highly correlated with each other.

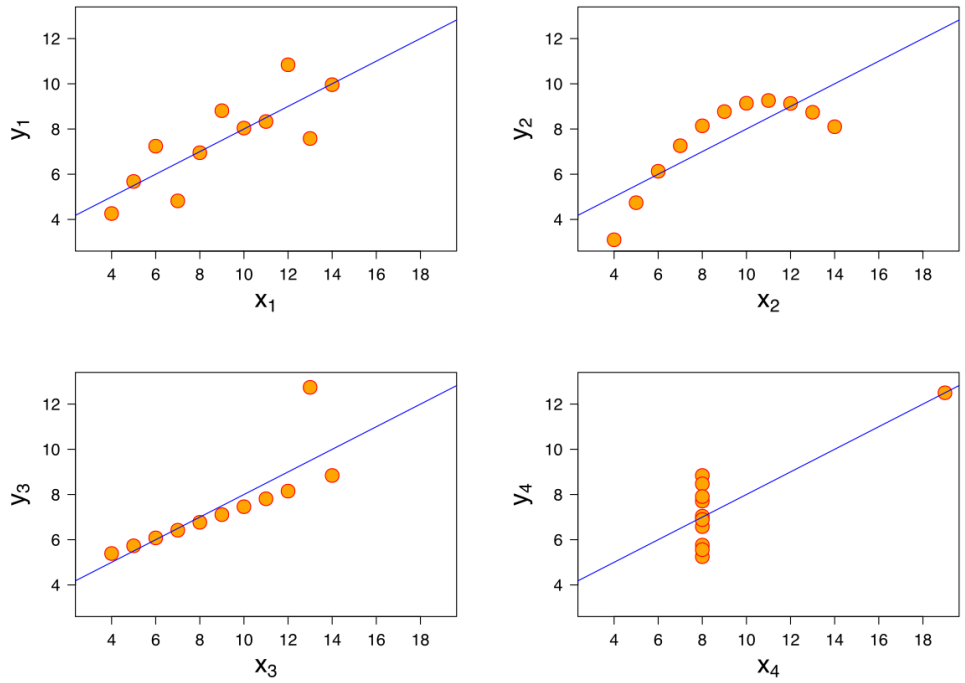
★ ***Once the model is trained and evaluated, it can be used to make predictions on new dataset.***

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer: -

Anscombe's quartet comprises four datasets, each containing eleven (x, y) points. Despite their similar simple descriptive statistics, these datasets exhibit markedly different distributions and graphical appearances.

Francis Anscombe created them in 1973 to underscore the importance of graphing data before analysis and to highlight the influence of outliers and influential observations on statistical properties. This demonstration challenges the misconception among statisticians that numerical calculations are precise while graphs are merely rough representations.



3. What is Pearson's R? (3 marks)

Answer: -

Pearson's R is a statistical measure that indicates the strength and direction of the linear relationship between two continuous variables.

The range of Pearson's R is from -1 to 1.

- If $R = 1$, it indicates that it is a perfect positive linear relationship between the variables. It means if one variable increase then the other variable also increases.

- If $\mathrm{R} = -1$, it indicates that it is a perfect negative linear relationship between the variables. It means if one variable increase then the other variable also decreases.
- If $\mathrm{R} = 0$, it means no linear relationship between variables.

Formula of Pearson's R: -

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

$r = corrletion\ coefficient$
$x_i = Values\ of\ the\ x - variable\ in\ a\ sample$
$\bar{x} =$ mean of the values of the $\mathrm{x} -$ variable
$y_i = Values\ of\ the\ y - variable\ in\ a\ sample$
$\bar{y} =$ mean of the values of the $\mathrm{y} -$ variable

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer: -

Scaling is a method used to make sure all the features in your data are on the same scale or range. This is important because if the features have very different magnitudes or values, some features might have a bigger influence on the model just because their values are bigger, even if they're not actually more important. By scaling the features, we ensure that they all contribute equally to the model, regardless of their original units or sizes. This is done during data preprocessing to make sure our machine learning algorithms work well with the data.

The key difference between normalized scaling and standardized scaling lies in how they adjust the scale of the features. Normalized scaling maintains the original range of values but transforms them to fall within the range [0, 1], while standardized scaling standardizes the values to have a mean of 0 and a Standard Deviation [SD] of 1, irrespective of the original range.

5. You might have observed that sometimes the value of VIF in infinite. Why does this happen? (3 marks)

Answer: -

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 2, this means that the variance of the model coefficient is inflated by a factor of 2 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation, we get R-squared = 1, which lead to $\frac{1}{1-\mathrm{R}^2}$ infinity. To solve this, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

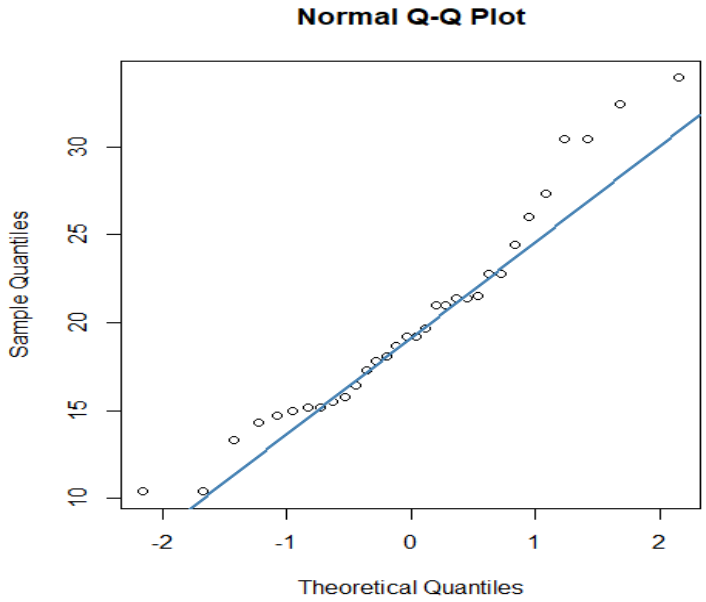
6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

Answer: -

Q-Q plots are also known as Quantile-Quantile plots. they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

Use of Q-Q Plot: -

Q-Q plots basically used to find the which type of distribution for a random variable whether it is a Gaussian distribution, Uniform distribution, exponential distribution, or a Pareto distribution. It means that we can tell the type of distribution using the power of Q-Q plot.



A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a normal Q-Q plot when both sets of quantiles truly come from normal distributions.