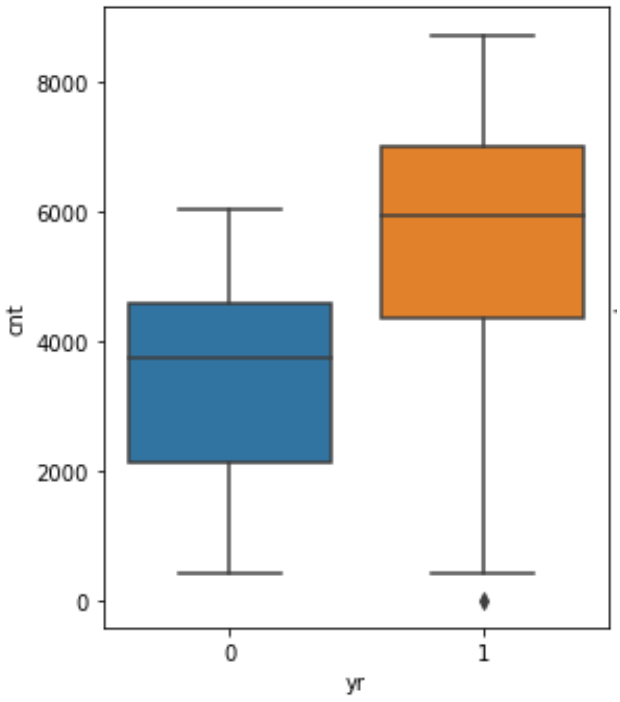
Assignment-based Subjective Questions:

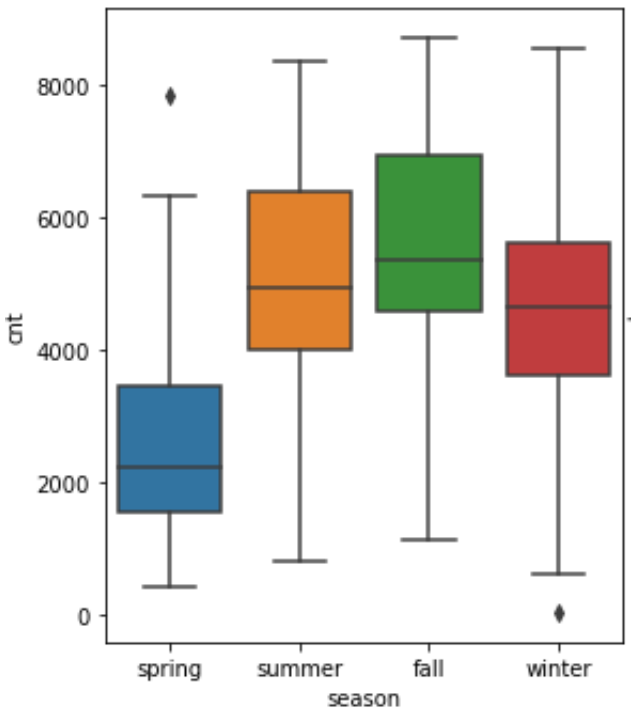
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Solution.

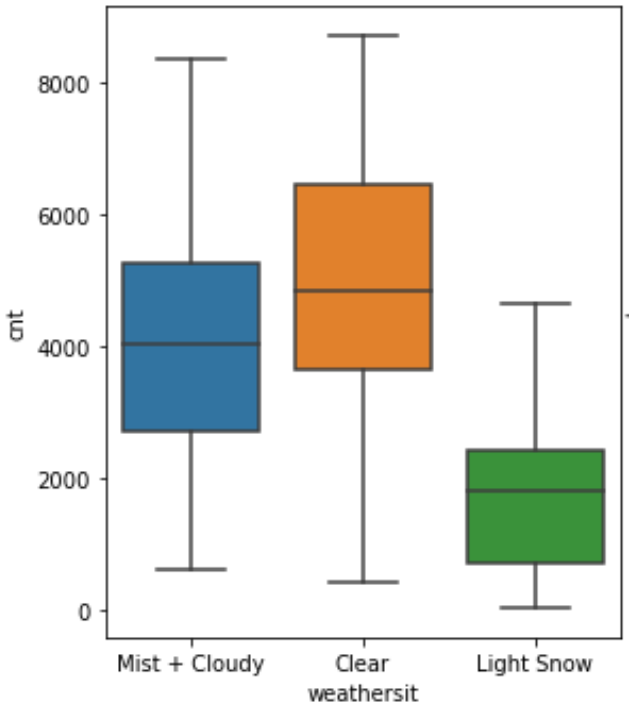
- Yr: Bike Rentals are more in the year 2019 as compared to 2018.



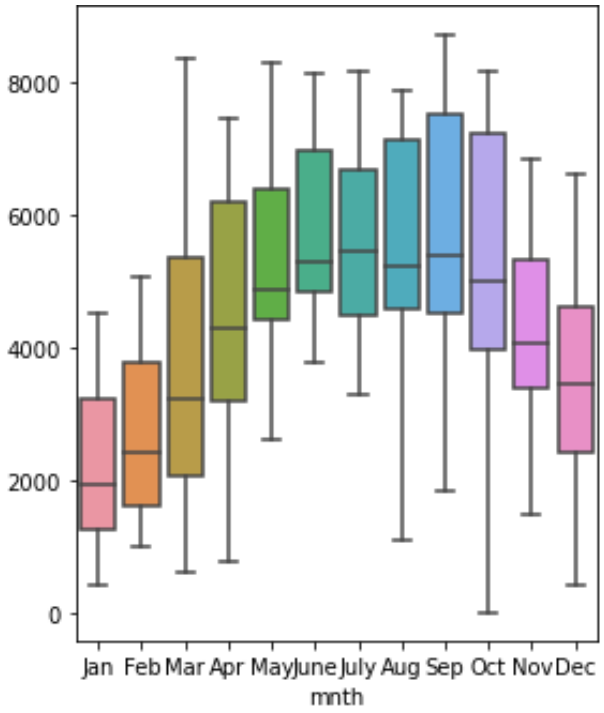
- Season: Bike Rentals are more during the Fall and summer season than in winter or spring.



- Weather: Bike Rentals are more in clear weather than in Light Snow.



- Months: Boom bikes rental business is more in the months of August, September and October.



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

Solution.

drop_first=True is important to use, as it helps in reducing the extra column (first column) created during dummy variable creation. For eg. Season has 4 categorical variables Spring, Summer, Fall and Winter. If we drop first column still, we can represent same information with 3 variables as follow.

000 Spring

001 Summer

010 Fall

100 Winter

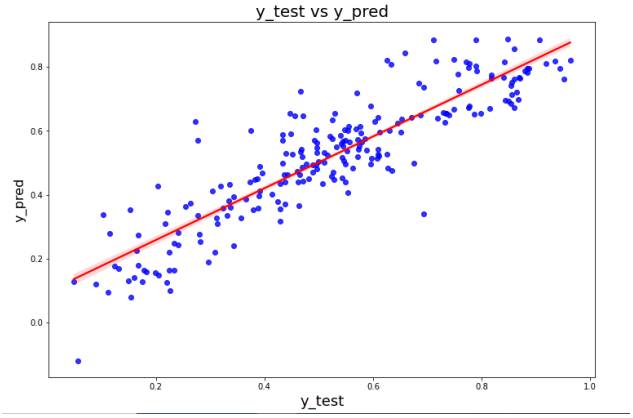
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

Solution. Temperature has the highest correlation. It has 0.63 correlation with cnt.

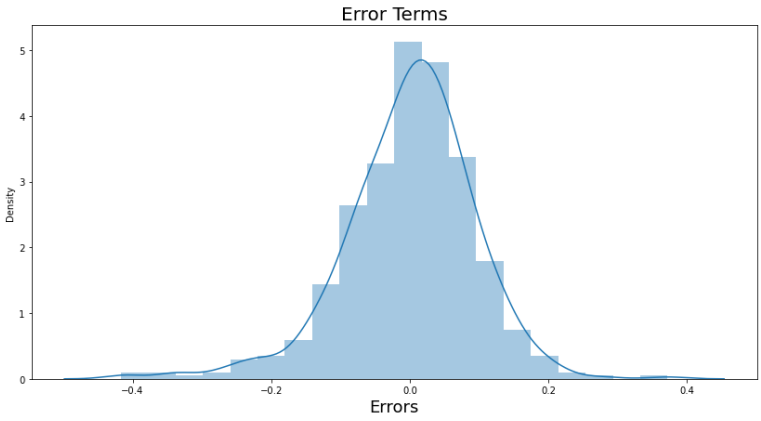
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Solution.

- There should be linear relationship between other variables and target variables.



- The error term should be normally distributed.



- No Multicollinearity: so, value of VIF should be less than 5.
- The model should not be overfitted

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

Solution.

Top 3 features contributing significantly towards explaining the demand of the shared bikes:

- Temperature having the highest positive coefficient 0.5390. It effects the business most.
- Year is second on the list because in 2019 after Covid business had good growth
- The light snow has most negative coefficient -0.2964. When there is light snow the demand of bike dropped down significantly.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Soln.

Linear Regression is a ML algorithm which is based on supervised learning. linear regression is a linear approach for modelling the relationship between a dependent variable and independent variables. It helps us model a target variable based on independent variables.

Linear regression predicts a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y (output).

$$y = \theta_1 + \theta_2.x$$

While training a model we are given:
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)

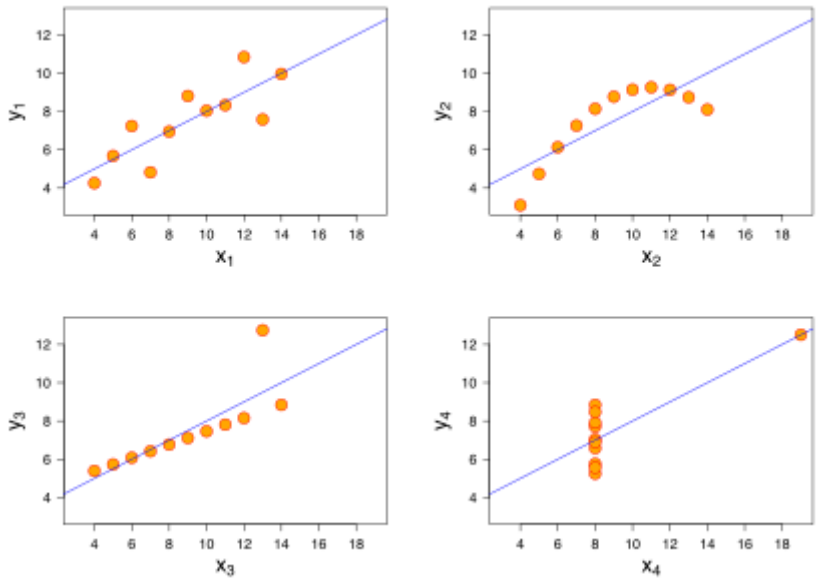
While training the model, it fits the best line to predict the value of y for a given value x. The model gets the best regression fit line by finding the best values of θ_1 and θ_2.
θ_1: intercept
θ_2: coefficient of x
Once we find the best θ_1 and θ_2 values, we will get the best fit line. After that when we are finally using our model for prediction, it will predict the value of y for the input value of x.

Cost Function (J): Cost function(J) of Linear Regression is the Root Mean Squared Error (RMSE) between predicted y value (pred) and true y value (y). The model aims to predict y value such that the error difference between predicted value and true value is minimum.

Gradient Descent: Update the values of θ_1 and θ_2 in order to reduce the Cost function (minimizing RMSE value) and achieve the best fit line the model using Gradient Descent. The idea is to start with random values of θ_1 and θ_2 and then iteratively update the values, reaching minimum cost.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet Consist of four data sets that have nearly same simple descriptive statistics but have very different distributions and appear very different when graphed as shown below.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed. Lets see them one by one.

For this all for set the value of mean, correlaton between x and y, R square value and sample variance is same.

First scatter plot (top left): Appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.

Second graph (top right): While a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant.

Third graph (bottom left): The modelled relationship is linear, but should have a different regression line.

Fourth graph (bottom right): One high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson's R?

Soln. Pearson's r is a numerical summary of strength of the linear relation between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association.

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

Scaling is a step of Data Pre Processing that is applied to independent variables or features of data. It helps to normalize the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm.

It is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale.

Normalization means rescaling the values into a range of [0,1]. Standardization means rescaling data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

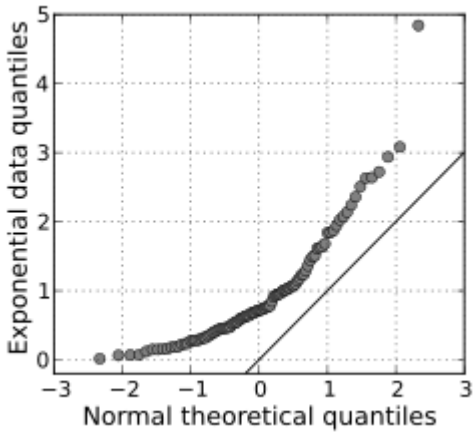
Solution.

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

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

Solution.

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Q-Q Plots are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line as shown in below image.



Use of Q-Q: It is use to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. Also, Q–Q plots is to compare the distribution of a sample to a theoretical distribution.

Importance of Q-Q plot is If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.