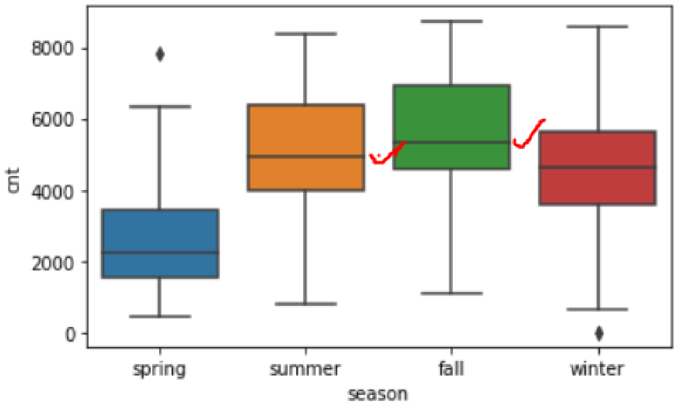
Assignment-based Subjective Questions

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

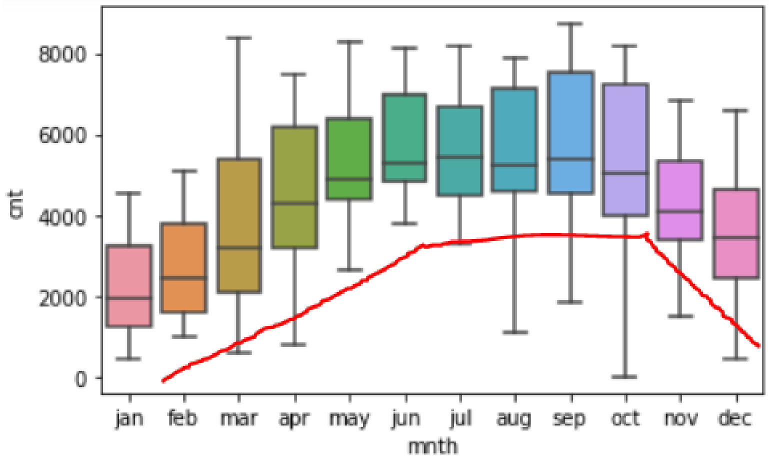
Ans: Effect of various categorical variables on dependent variable ‘cnt’ are as below

a) Impact of weather condition

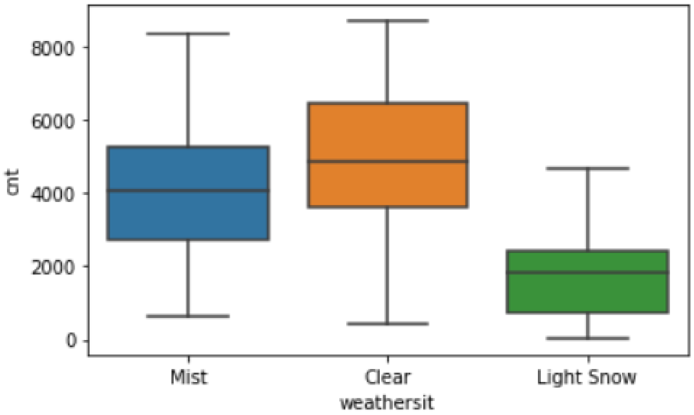
- In summer and Fall season, demand of shared bikes are high as compared to other season.



- Demand of shared bikes increases every month in Q1(Jan-Mar) and Q2(Apr-Jun), Remain high in Q3(Jul-Sep) then starts declining in Q4(Oct-Dec).
  The demands are lowest in winter which runs from Dec1 till Feb 28/29 in US.

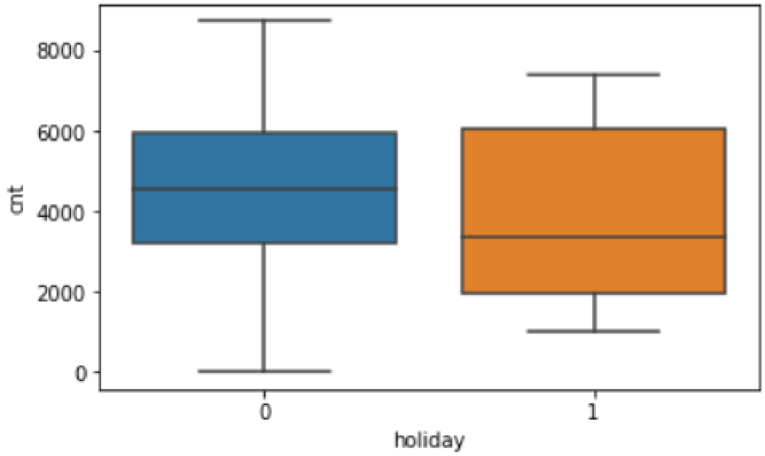


- Demand of shared bikes are very high in Clear weather.



b) Impact of Holiday

The demands are less on holiday as compared to working day. Probably office going commuters are using the shared bikes most or people preferring travel by own vehicle on holidays or just staying at home or out of station on holiday.

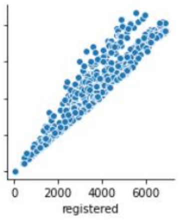


2. Why is it important to use **drop_first=True** during dummy variable creation?

Ans: For categorical variable of level n, the dummy variable needed to represent the values numerically in n-1. The last column can be dropped which can be represented by combination of values in all other n-1 columns.

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

Ans: registered is having highest correlation with the target variable cnt.



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

Ans:

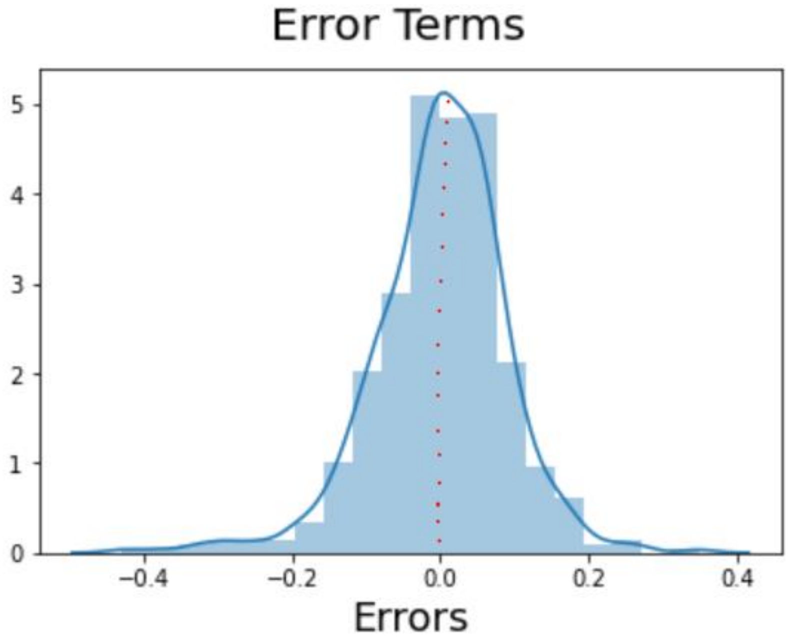
Assumptions of linear regression

- Linear relationship between independent variable X and target variable y
- Error terms are normally distributed having mean centered around 0

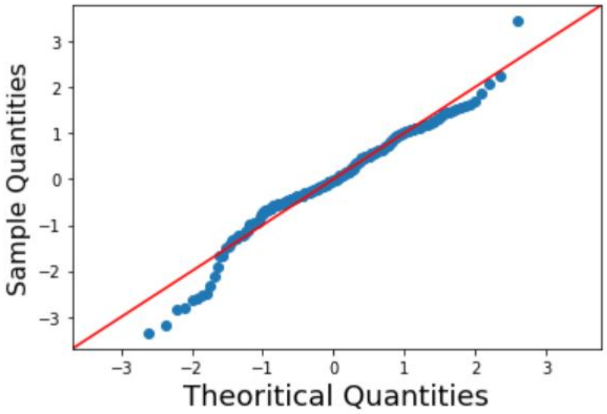
  Error Term = Y_actual – Y_predicted
- Error terms are independent
- Error terms have constant variance

1. The above assumption can be verified by checking the distribution of error terms.

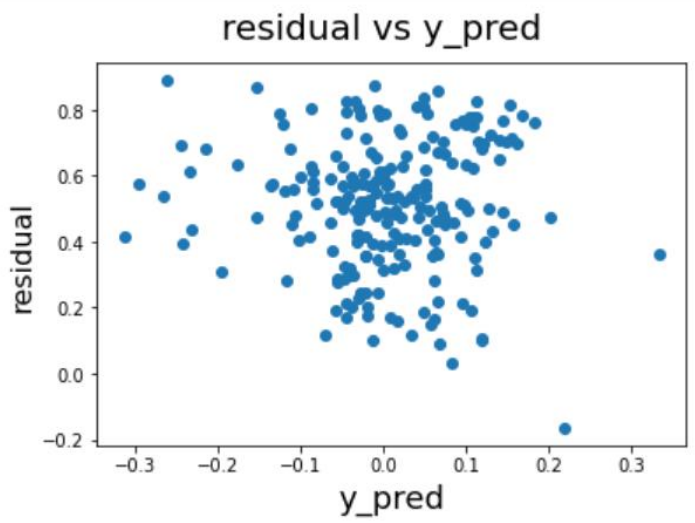
   This can be achieved by plotting histogram of error term. If the result is close to normal distribution curve having mean around 0, that means the assumptions are correct.



2. Another way to validate normal distribution in dataset is to plot QQ plot for the residual.



3. If no pattern observed in residual vs y_predicted plot, then it represents error term are indepdendent



4.

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

Ans:

The top 3 features contributing significantly towards explaining the demand of the shared bikes are

i) **atemp (actual feel temperature)**
The beta coefficient of 'atemp' is + 0.4436 which means that the demand of shared bikes increases with increase in temperature in US.

ii) **Light Snow**
The beta coefficient of 'Light Snow' is - 0.2888 which means that the demand of shared bikes decreases with snow.

iii) **Windspeed**
The beta coefficient of 'windspeed' is - 0.1347 which means that the demand of shared bikes decreases with increase in windspeed.

Looking at all the 3 major factor, it is evident that more favorable the weather outside, the demand of shared bike also grows.

General Subjective Questions

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

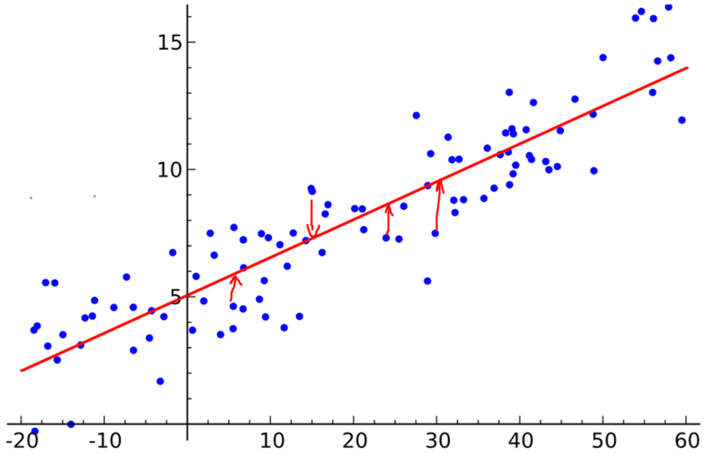
Ans:

Regression in the method of modelling a dependent target variable from independent predictor variables. Linear regression is one such algorithm to build model in machine learning space. When one target variable y is in linear relationship with one or many predictor variables X, the line can be expressed by

$y = mx + c$ for single predictor variable x

or $y = c + B1 * X1 + B2 * X2 + B3 * X3 + \ldots\ldots\ldots. + Bn * Xn$

where c is constant.



Linear Regression

The end goal is to fit the dotted lines as close as red line which can be achieved by below algorithms

a) **Ordinary Least Squares (OLS):** This method tries to fit the best curve by minimizing the sum of the squares of the residuals made in the results of each individual equation where residuals is difference between y-actual and y-predicted.

This method is present in ‘statsmodels’ library.

b) **Recursive Feature Elimination** (RFE): This algorithm works on the idea to repeatedly construct a model by eliminating the worst performing feature and repeating the process with the rest of the features until optimum result is obtained.

This method is present in Sci Kit library.

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

Ans: Anscombe’s quartet explains the importance of data visualization before building the model.

Anscombe’s quartet is defined by group of 4 datasets which are very identical when represented statistically but they have very different distribution. It is identified when we visualize data through Scatter plot and check the distribution.

These anomalies lead to fool the model if built hence it is very important to visualize the data, find such anomalies before proceeding with data modeling using linear regression.

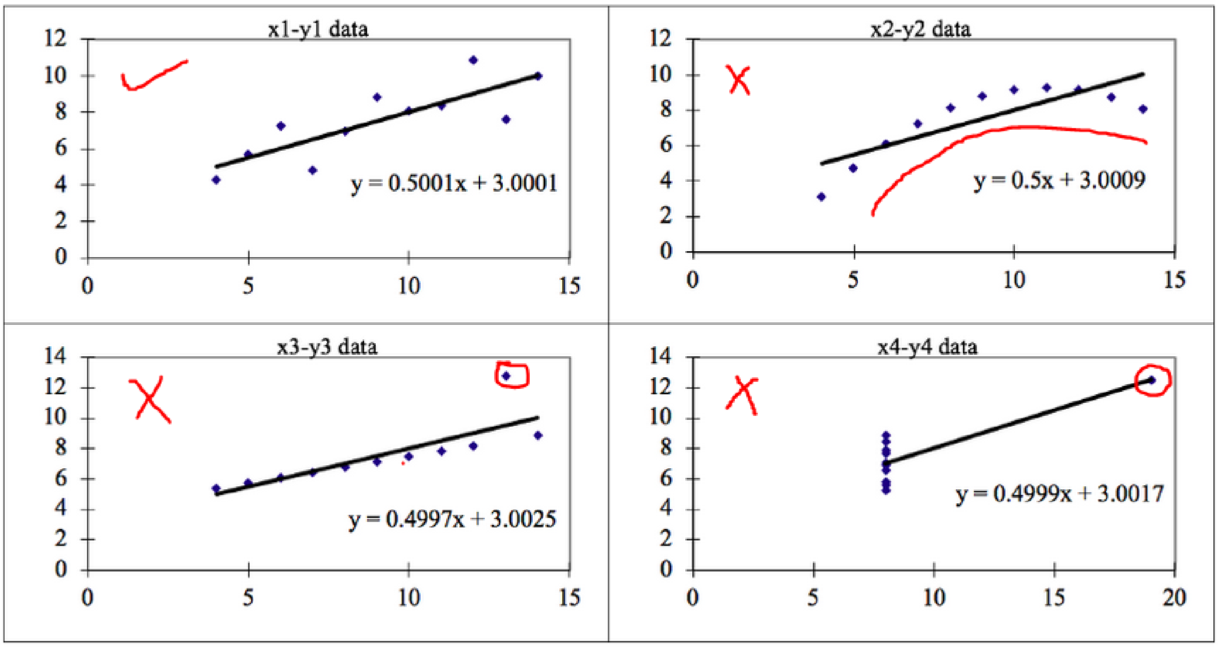
The four datasets can be described as:

Dataset 1: this **fits** the linear regression model pretty well.

Dataset 2: this **could not fit** linear regression model on the data quite well as the data is non-linear.

Dataset 3: shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

Dataset 4: shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

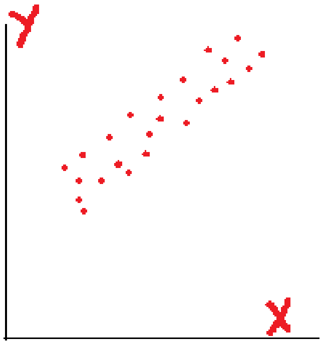


The slope all 4 dataset is ~0.5 and c is ~3 but when represent in graph, only the first one shows linearly distributed.

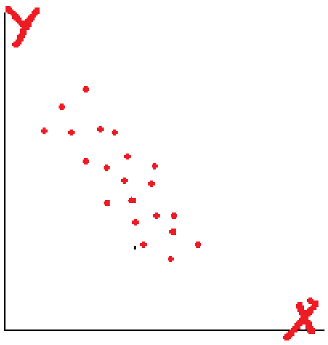
3. What is Pearson's R? (3 marks)

Ans: Pearson's R or Pearson's Correlation Coefficients is the measure of linear correlation between the two datasets. As it tells the relation between two datasets, it is also called bivariate relation.

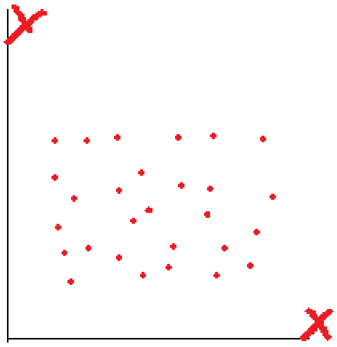
- If Pearson's R can be positive, negative or 0. It varies between -1 to +1.
- When the increases in one variable causes the increase in another, they are called positively correlated and value of Pearson's R varies between 0 to +1.
- When the increases in one variable causes the decline in another, they are called negatively correlated and value of Pearson's R varies between -1 to 0.
- When the increases or decrease in one variable doesn't impact other variable, they are not in any relationship and are independent.



+ve correlation



-ve correlation



No correlation

Pearson’s r is calculated by

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

Where

R or r = correlation coefficient

Xi = value of x variable in sample

X^ = mean of values of x variables

Yi = value of y variable in sample

Y^ = mean of values of y variables

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

Ans:

Scaling is a technique to standardize the independent features present in the data in a fixed range. This is needed for

- Easy interpretation and comparison between beta coefficients of various features
- Help gradient descent to work faster and as values are scaled within certain range.

Normalized Scaling	Standardized Scaling
Rescale the value into the range of 0 to 1	Rescales data to have mean 0 and standard deviation of 1
Formula: $x = \frac{x - min(x)}{max(x) - min(x)}$	Formula: $x = \frac{x - mean(x)}{sd(x)}$

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

Ans:

In 2 scenario we get infinite VIF

1) When there is perfect collinearity between two indepdent variables. In this case R is 1.

   The formula of VIF is $VIF_i = \frac{1}{1-R_i^2}$

   When R=1, VIF = 1/0 = infinite

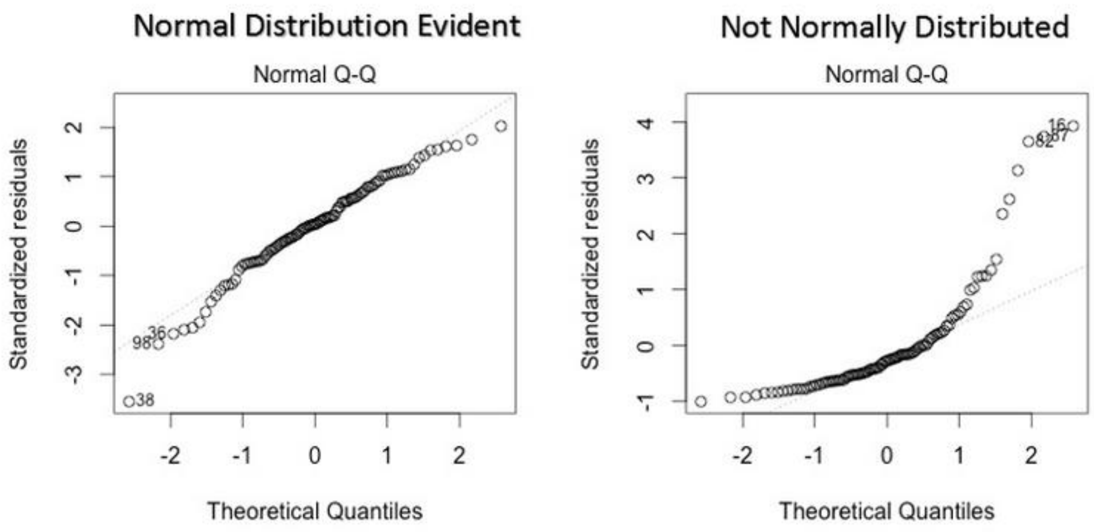
2) When column **const** is not removed from the dataset before calculating VIF, it show infinite for few columns.

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

Ans: Q-Q plot is plot of quantiles of two dataset against each other. If the two dataset are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line.

Use: Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc.

Q-Q plot is used to check the distribution of error terms are normally distributed to validate the assumption of linear regression.


Normal Distribution Evident
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Not Normally Distributed
Normal Q-Q
Standardized residuals
Theoretical Quantiles