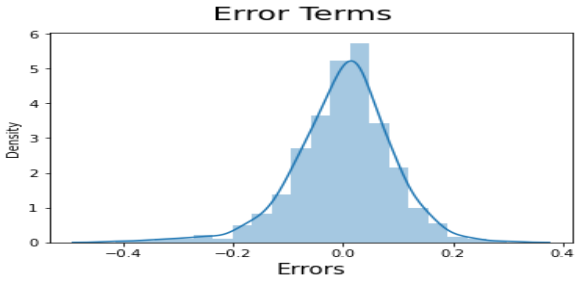
Bike Sharing Assignment: QA

➔ 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?
    - The categorical variables we had: season, Year, Holiday, Weekday, WorkingDay, Month, WeatherSit, the dependent variable was cnt
    - I did visualize the categorical vs Dependent variables and found the below details
    - From the initial analysis, summer, fall season have positive impact, cnt increases in summer, fall and decreases in winter, least is spring
    - Cnt decreases on a holiday
    - Cnt increases on working day
    - Cnt increases in June, July, Aug, Sep months
    - Cnt is higher when the weather is clear/partly cloudy
    - Cnt decreases when its raining/snow
    - Cnt has increased in year 2019 compared to 2018

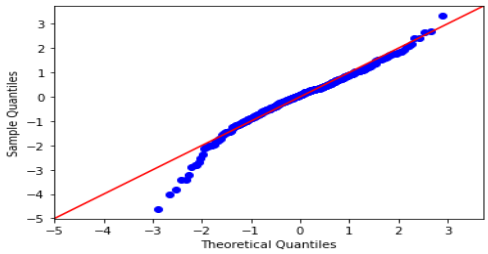
2. Why is it important to use drop_first=True during dummy variable creation?
    - This is One hot encoding is technique which converts categorical data into a form which is understandable by ml model
    - Dummy variable creation is a process where we convert a categorical variable into numerical variables. This process helps us in reducing extra columns created during dummy variable creation, It reduces the correlations created between dummy variables. Drop_first=True will helps in reducing this extra column. If we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables, if drop_first is true it removes the first column which is created for the first unique value of a column.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
    - Registered feature 0.95 – we have removed them as they cannot add value to prediction
    - Temperature feature 0.65 – has positive co relation we can consider this as highest

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
    - After building a model on training set I performed residual analysis, this helps us understand if errors are normally distributed or not
    - 

    - VIF will help us in understanding the co-relation between the variables

  - Variance Inflation Factor explains the severity of multicollinearity 1 / (1 – R Square)
  - VIF should be less than 5
- Homoskedasticity, we can check this via plot between residual vs fitted values
- Plotting the variables will explain the linear curve, absence of this will tell there is a no linearity between independent variable and dependent variable
- I used qq plot to understand the normal distribution of errors



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

- Temp has higher positive impact on demand of bikes
- Season has major impact
- Week day has positive impact

	coef
const	0.1075
summer	0.0737
winter	0.1156
Jan	-0.0453
July	-0.0372
Sep	0.0886
Cloudy	-0.0805
Light Rain	-0.2890
Saturday	0.0590
yr	0.2334
holiday	-0.0573
workingday	0.0470
temp	0.5384
windspeed	-0.1627

-

➔ GENERAL SUBJECTIVE QUESTIONS:

1. Explain the linear regression algorithm in detail.

Answer: Linear Regression is a machine learning algorithm which is based on supervised learning category. It finds a best linear fit relationship on given data, between independent and dependent variables. It creates the best straight-line fitting to the provided data to find the best linear relationship between the independent and dependent variables. Mostly it uses Sum of Squared Residuals Method.
2 types of linear regression:

1. Simple Linear Regression: It explains the relationship between a dependent variable and only one independent variable using a straight line. Formula for this is $Y=\beta0+\beta1X1+\epsilon$

2. Multiple Linear Regression: It shows the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X. Formula for this is $Y=\beta0+\beta1X1+\beta2X2+\ldots+\beta pXp+\epsilon$. This will be fitting a plane rather than a line.

Linear Regression can be used for predictive analysis and modeling.

Assumptions of Linear regression:

1. Errors are normally distributed
2. Errors are independent from one another
3. Constant variance between error terms
4. There should be a linear relationship between x and Y

2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet is set of 4 quadrants, which is often used to explain how important it is to look at the data graphically even before starting to analyze.

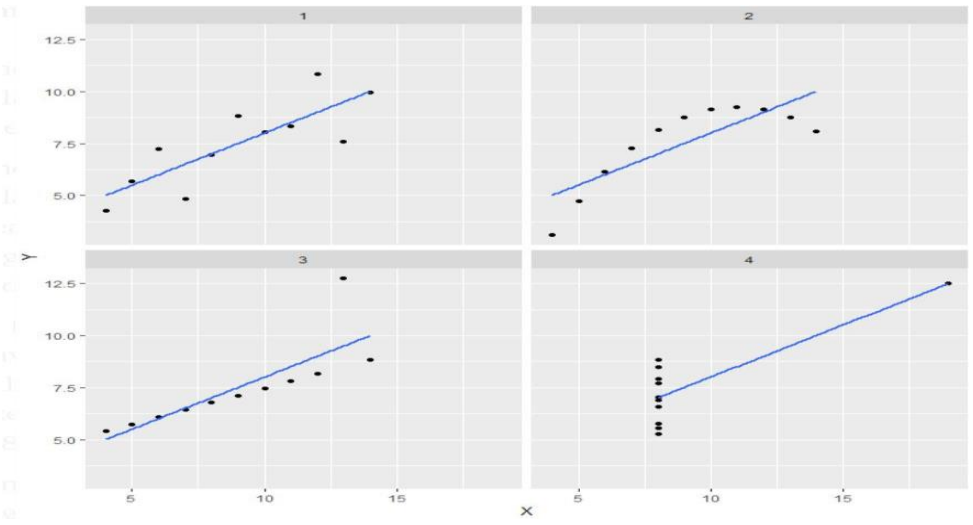
This Quadrants were created in 1973 by Anscombe. It consists of 4 data sets with x and Y values, each set has 11 floats. Statistically if we try to understand the data sets, all of them have same mean, standard deviation and co-relation.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

But when we plot them in a scatter plot, it looks like they are all different as shown in below image



- In the first one(top left), the scatter plot we will see that there seems to be a linear relationship between x and y.
- In the second one(top right) we can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) we can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high point is enough to produce a high correlation coefficient.

3. What is Pearson's R?

Answer: Pearsons R is a correlation coefficient which is a measure of the strength of linear association between 2 variables, it is determined by 'r'. The value of r will always be between -1 and +1, where +1 is defining the positive co-relation and -1 defining the negative co-relation. 0 is no co-relation.

Mathematically, Pearson's correlation coefficient is denoted as the **covariance** of the two variables divided by the product of their **standard deviations**.

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

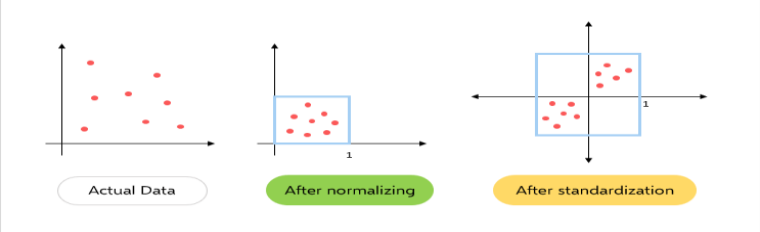
Where:

N	=	number of pairs of scores
Σxy	=	sum of the products of paired scores
Σx	=	sum of x scores
Σy	=	sum of y scores
Σx^2	=	sum of squared x scores
Σy^2	=	sum of squared y scores

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

Answer: Scaling is the process to normalize the data within a particular range. Many times, in our dataset we see that multiple variables are in different ranges. So, scaling is required to bring them all in a single range.

Most used scaling methods are **Normalization** and **Standardization**. Normalization typically scales the values into a range of [0,1]. Standardization typically scales data to have a mean of 0 and a standard deviation of 1.



Formula for normalized scaling:

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

Formula for standardized scaling:

$$x = \frac{x - mean(x)}{sd(x)}$$

We should use Normalization only when the data distribution is unknown or the data doesn't have Gaussian Distribution(probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean)

We should use standardization when data has Gaussian distribution (bell curve data)

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

Answer: VIF is used to measure the multicollinearity. The formula to find it is

$$VIF_i = \frac{1}{1 - R_i^2}$$

'i' refers to the ith variable.

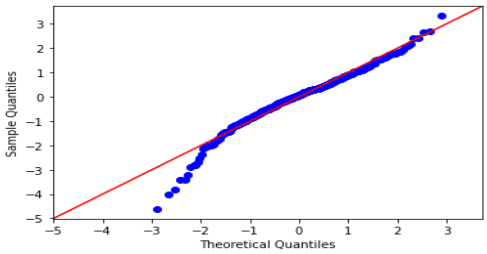
If R-squared value is equal to 1 then the denominator of the above formula becomes 0 and the overall value become infinite. It denotes perfect correlation in variables.

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

Answer: q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

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.

On a Q-Q plot normally distributed data appears as roughly a straight line (although the ends of the Q-Q plot often start to deviate from the straight line).



Above image is the Q-Q plot of residuals, here it tells that residuals are normally distributed. If the points don't lie on the line it means residuals are not normal.