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?

   Answer: From analysis of the categorical variables from the dataset, I could infer that-

   - The number of people using bike sharing on holiday is less than non-holiday. Average number of bike sharing is slightly more on a working day.
   - There is no significant impact of weekday on bike sharing count.
   - A good weather situation attracting more bike sharing and it is least on bad weather.
   - Average demand for bike sharing is gradually increasing from January to July and gradually decreases from july to december. July shows the highest average bike sharing. This might be a correlation of season.
   - From 2018 to 2019, the demand for bike sharing has drastically increased.
   - Demand for bike sharing is highest in the fall season and is least on spring season.

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

   Answer: If a categorical feature is having N distinct values, we can create N dummy variables using the values of the feature. But all the values of the feature can be represented by N-1 dummy variable. The get_dummies function by default returns N dummy variables which can be covered with N-1 variables as well. So to get N-1 variable, it is important to use **drop_first=True.** It also helps in reducing the correlations among dummy variables.

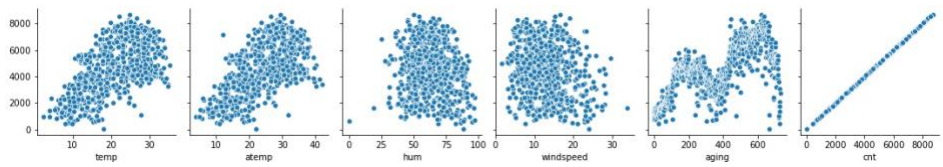
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

   Answer: Looking at the pair-plot among the numerical variables, the column temp has the highest correlation with target variable cnt.

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

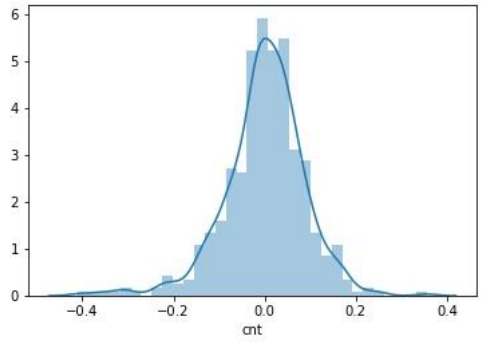
   Answer: Validated the assumptions of Linear Regression based on following metrics-

   1. Linearity



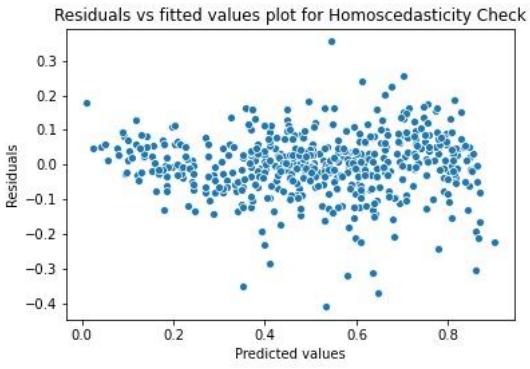
   The above pairplot shows linear relationship of temp with target variable cnt.

2. Mean of residuals



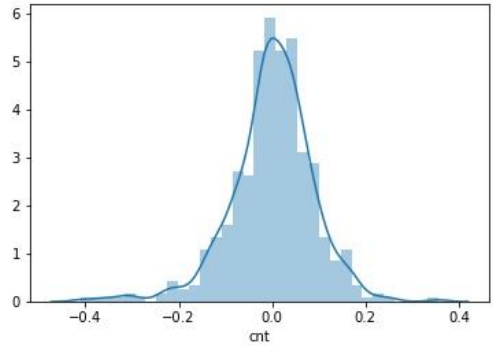
The mean of residuals is found to be -1.040017745126801e-16 which very close to 0. This can be observed in the above displot as well

3. Check for Homoscedasticity



In the above scatterplot, there is no definite patterns observed. So it can be inferred that Homoscedasticity obtained.

4. Check for Normality of error terms/residuals



The above distplot shows that the error terms follows a normal distribution with mean equal to 0.

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

Answer: The top 3 columns contributing towards demand are temp, yr and workingday

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear regression algorithm is a machine learning model in which we try to predict values of a target variable based on one or more independent variables. Here the independent variables have a linear relationship with the target variable. The machine learning model is trained on a training dataset to create the linear relationship which can be expressed as function y

$y=m_1x_1+m_2x_2+...+m_nx_n+c$

Here, y is the predicted value

m_n is coefficient value

c is the intercept value

x_n is independent variable

For example, if there is a linear relationship between advertisement and sales, we can predict amount of sales based on amount of fund invested in sales using linear regression.

2. Explain the Anscombe's quartet in detail.

Answer: A group of four dataset which have nearly equal descriptive statistic but greatly differ when plotted on scatter plot is called Anscombe's quartet. The peculiarities in the dataset gives unexpected prediction if built on a linear regression model.

This shows us the importance of visualising the data before applying any machine learning model on it so that distribution of the data such as linearity, diversity, outliners can be observed.

For example, let us consider following dataset which have 4 sets of 11 data points

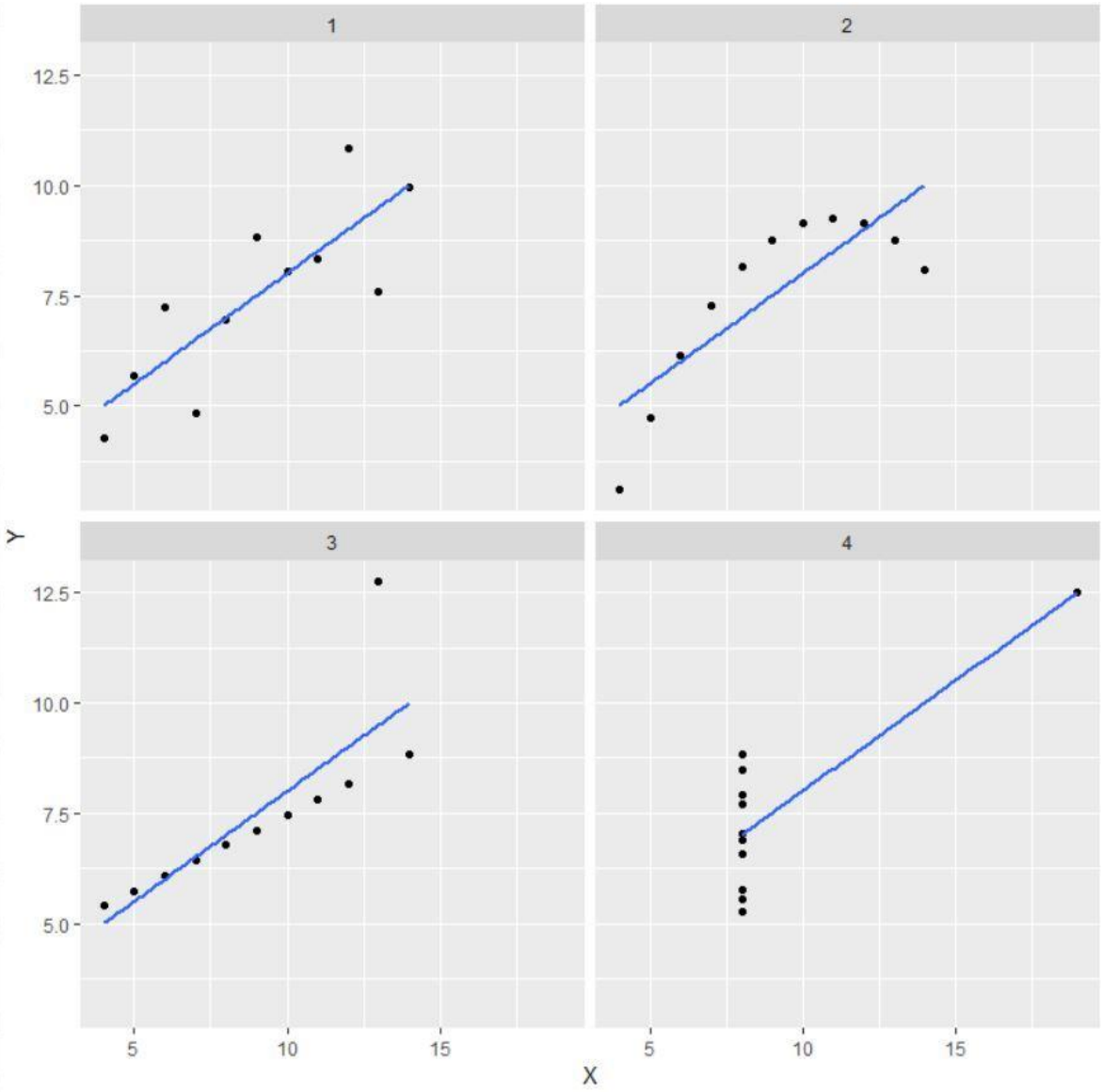
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

Analysing the descriptive statistics mean, standard deviation, and correlation between x and y of the dataset are calculated as below.

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

The descriptive statistics are nearly equal for all x and y. But when visualised in a scatterplot, we observed that the distribution of the datapoints are totally different.



Here it can be observed that-

1. Dataset1 show a linear relationship between x1 and y1. This can be used for linear regression
2. In dataset2, relationship between x2 and y2 is non-linear. Hence linear regression model would be fit.
3. Dataset3 have outliers which can not be handled by linear regression model
4. Dataset4 also have outliers which can not be handled by linear regression model

3. What is Pearson's R?

Answer: The strength of the linear association between the variables can be expressed as a numerical summary. When the variable tends to go up and down together, the correlation coefficient will be positive. If the variable move in opposite direction, the correlation coefficient will be negative. This can be expressed by Person's R. It is the ratio between the covariance of two variables and the product of their standard deviations.

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?

Answer: Scaling is a method to normalize the range of independent variables or features of data during the data preprocessing.

Machine learning algorithms while calculating association among features are biased towards numerically larger values if they are not normalized. So scaling is an important data preprocessing step to be performed in machine learning.

Normalized scaling rescales the values into a range of [0,1] or sometimes [-1,1]. Normalization is useful when there are no outliers as it cannot cope up with them. For example, we would normalize age and not incomes as the age is close to uniform but few people might have high incomes which would create outliers.

Normalization can be expressed as below-

X_new = (X - X_min)/(X_max - X_min)

Standardized scaling scales the data to have a mean of 0 and standard deviation of 1 i.e. unit variance. It does not get affected by outliers because there is no predefined range of transformed features.

Mathematically it can be expressed as below-

X_new = (X - mean)/Std

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

Answer: Yes, it is possible to have value of VIF as infinite. When a variable can be expressed exactly as a linear combination of other variables, the VIF value would be infinite. This shows a perfect correlation among the variables.

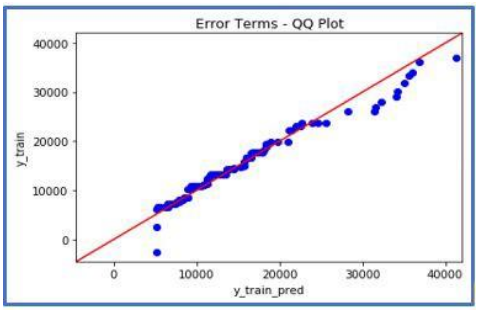
We know that VIF = 1/(1-R2). In the case of perfect correlation, we get R2=1. So 1/(1-R2) is infinite.

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

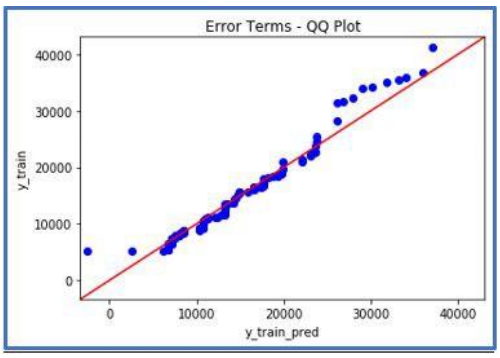
Answer: Q-Q plot abbreviated for quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution. It is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below interpretation of the two datasets are possible-

1. Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis.
2. Y-values < X-values: If y-quantiles are lower than the x-quantiles.



3. X-values < Y-values: If x-quantiles are lower than the y-quantiles.



4. Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis

Q-Q plot is used in linear regression when we separately receive the training and test datasets and we want to confirm that both the datasets are from populations with same distribution. It can be used to check if the dataset –

1. come from populations with a common distribution
2. have common location and scale
3. have similar distributional shapes
4. have similar tail behaviour

It can be used even if sample sizes are not equal. Using Q-Q plots, many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.