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)

Ans The categorical variable in the dataset were *season, weathersit, holiday, mnth, yr* and *weekday*. These were visualized using a boxplot. These variables had the following effect on our dependent variable:-

1. The season variable is highly correlated to mnth and cnt variable and negative correlation with weekday.
2. **Weathersit** - There are no users when there is heavy rain/ snow. Highest count was seen when the weathersit was “Clear, Partly Cloudy”.
3. Holiday has a negative effect on the target variable.
4. **Mnth** - September saw highest no of rentals while December saw least.
5. **Yr** - The number of rentals in 2019 was more than 2018.

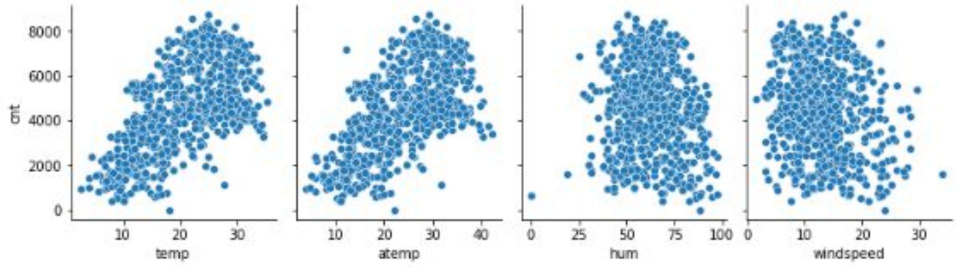
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

We use drop_first = True while creating dummies because the value we are dropping in regression can be found with the remaining dummy columns itself. This is explained in detail in the example. So to avoid redundancy we are dropping a column.

Example:

If there is a categorical variable containing the colour of ball with the values of Red, Blue and Green. When we create dummies to this column so that Red, Blue and Green columns will be created. Here while creating the dummies we are using drop_first = True due to which one of the columns of Red, Blue, Green will be dropped. It is done because, if a row has Green as value the value in the dummy variable of Red and Blue would be 0 which gives it obvious that the value of Green dummy variable would be 1. So its value can be easily interpreted just the two columns. It is not necessary to have a separate column to determine its value using another variable. And it is advisable to create a dummy variable of n-1 columns which has n values. This is simply to avoid redundancy.

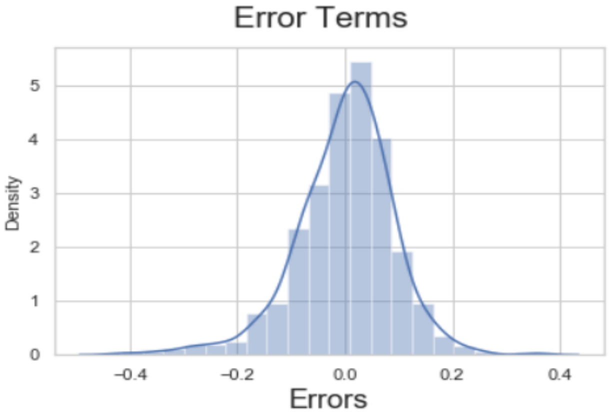
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

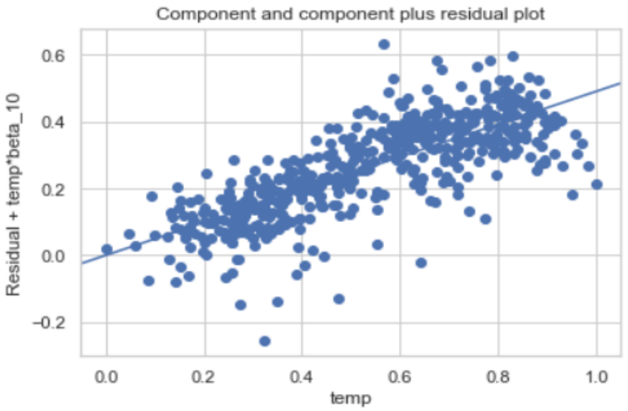


As per the above pair-plot “temp” and “atemp” are the two numerical variables which are highly correlated with the target variable (cnt).

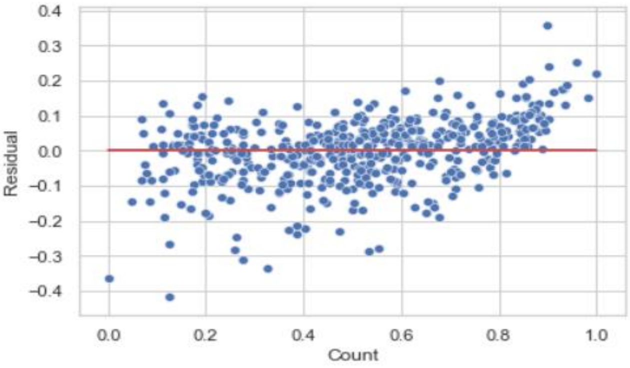
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

The major assumptions of Linear Regression are,
1. X and Y should display some sort of a linear relationship. Otherwise, there is no use of fitting a linear model between them.
2. Error terms are normally distributed with mean (not X and Y)
3. Error terms are independent of each other.
4. Error terms have constant variance





The assumptions done while building the linear regression model. Once theregression model is built on the training set, we use the distplot using seaborn to verify the error terms are normally distributed using the values of y_train and y_train_pred. From the below distplot it is evident that the error terms are normally distributed with mean at 0.



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

The top 3 features are:
1. **temp**
2. **yr**
3. **weathersit_Light Snow & Rain**

These variables are having positive relation with the cnt variable say, 1 unit increase in yr contributes 0.233570 , 1 unit increate in weathersit contributes -0.284199 and 1 unit increase in temp contributes0.490988 towards the cnt variable.

General Subjective Questions

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

Linear regression is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable. Here are the types of regressions:

1. Simple Linear Regression
2. Multiple Linear Regression
3. Logistic Regression
4. Polynomial Regression

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Example:

We are running a sales promotion and expecting a certain number of count of customers to be increased now what we can do is we can look the previous promotions and plot if over on the chart when we run it and then try to see whether there is an increment into the number of customers whenever we rate the promotions and with the help of the previous historical data we try to figure it out or we try to estimate what will be the count or what will be the estimated count for our current promotion this will give us an idea to do the planning in a much better way about how many numbers of stalls maybe we need or how many increase number of employees we need to serve the customer. Here the idea is to estimate the future value based on the historical data by learning the behaviour or patterns from the historical data. In some cases, the value will be linearly upward that means whenever X is increasing Y is also increasing or vice versa that means they have a correlation or there will be a linear downward relationship.

Linear regression is used to predict a quantitative Y from the predictor variable X. Mathematically, we can write a linear regression equation as:

The equation for MLR will be:

$$y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{ip}+\epsilon$$

ϵ = The model's error term (also known as the residuals)

where, for i=n observations:

y_i = Dependent variable

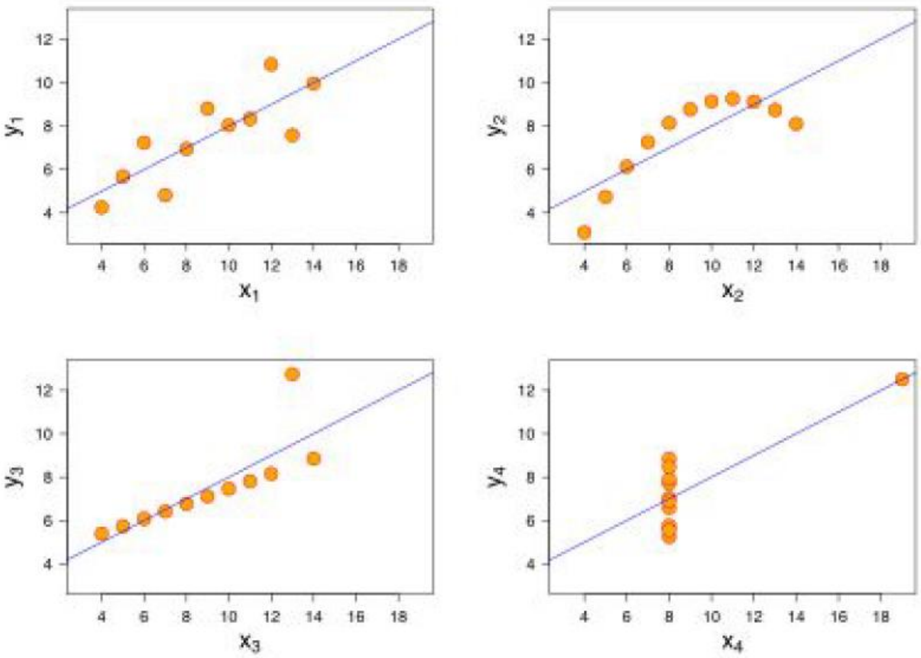
x_i = Explanatory variables

β_0 = y-intercept (constant term)

β_p = Slope coefficients for each explanatory variable

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

Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.



- ➢ The first scatter plot (top left) appears to be a simple linear relationship.
- ➢ The second graph (top right) is not distributed normally, while there is a relation between them, it's not linear.
- ➢ In the third graph (bottom left), the distribution is linear, but should have a different regression line the calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- ➢ Finally, the fourth graph (bottom right) shows an example when 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.

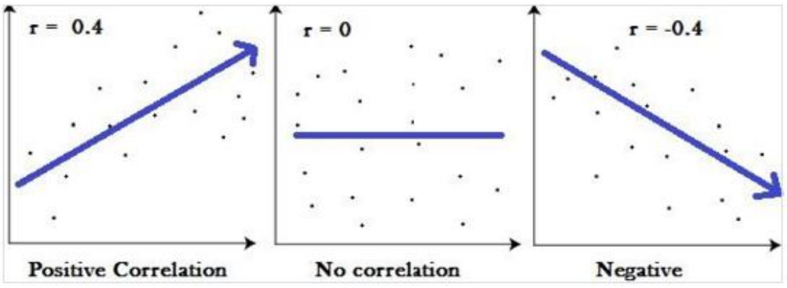
3. What is Pearson's R?

(3 marks)

he **Pearson product-moment correlation coefficient** is a measure of the strength of the linear relationship between two variables. It is referred to as Pearson's correlation or simply as the correlation coefficient. If the relationship between the variables is not linear, then the correlation coefficient does not adequately represent the strength of the relationship between the variables.

The symbol for Pearson's correlation is "ρ" when it is measured in the population and "r" when it is measured in a sample. Because we will be dealing almost exclusively with samples, we use r to represent Pearson's correlation unless otherwise noted. Pearson's r can

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$



r = 1 means the data is perfectly linear with a positive slope.
r = -1 means the data is perfectly linear with a negative slope.
r = 0 means there is no linear association.

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

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

- Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.
- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

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

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

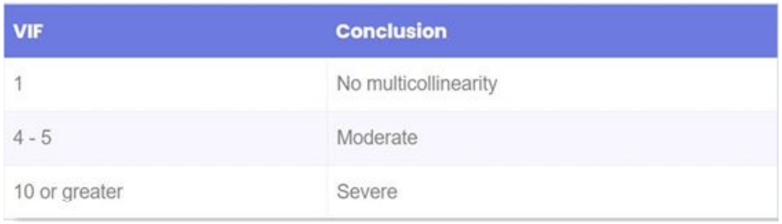
The variance inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model. The extent to which a predictor is correlated with the other predictor variables in a linear regression can be quantified as the R-quared statistic of the regression where the predictor of interest is predicted by all the other predictor variables.

The variance inflation for a variable is then computed as:

$$VIF = \frac{1}{1 - R^2}$$

If all the independent variables are orthogonal to each other, then VIF = 1.0.
• 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 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.
The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity. This is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.

An **infinite VIF** value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.



VIF	Conclusion
1	No multicollinearity
4 - 5	Moderate
10 or greater	Severe

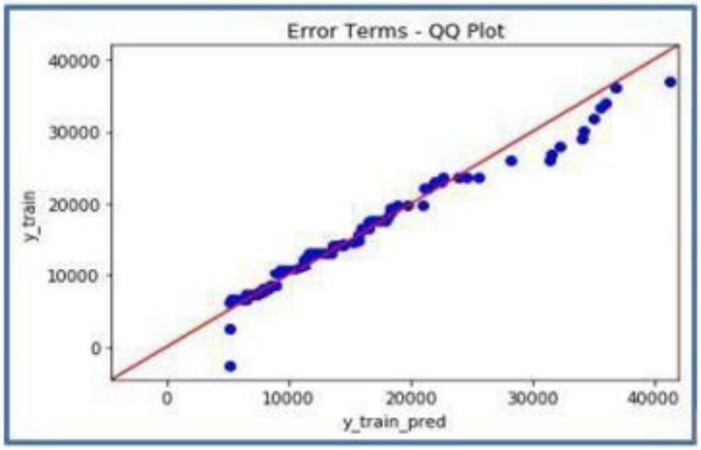
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(3 marks)

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions. 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. The q-q plot is used to answer the following questions:

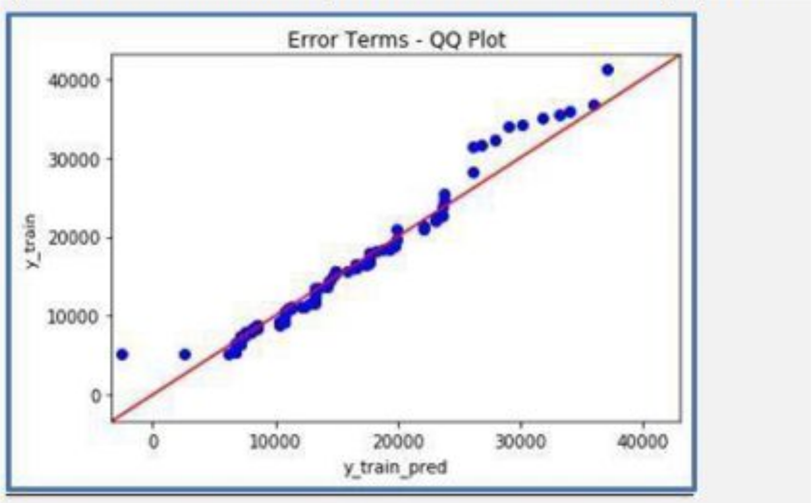
❖ Do two data sets come from populations with a common distribution?
❖ Do two data sets have common location and scale?
❖ Do two data sets have similar distributional shapes?
❖ Do two data sets have similar tail behavior?

nterpretation:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
Below are the possible interpretations for two data sets.
a) **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

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



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



d) D**ifferent distribution**: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis