Assignment-based Subjective Questions

1. **From your analysis of categorical variables from the dataset, what can you infer about effect on dependent variable?(3 marks)**

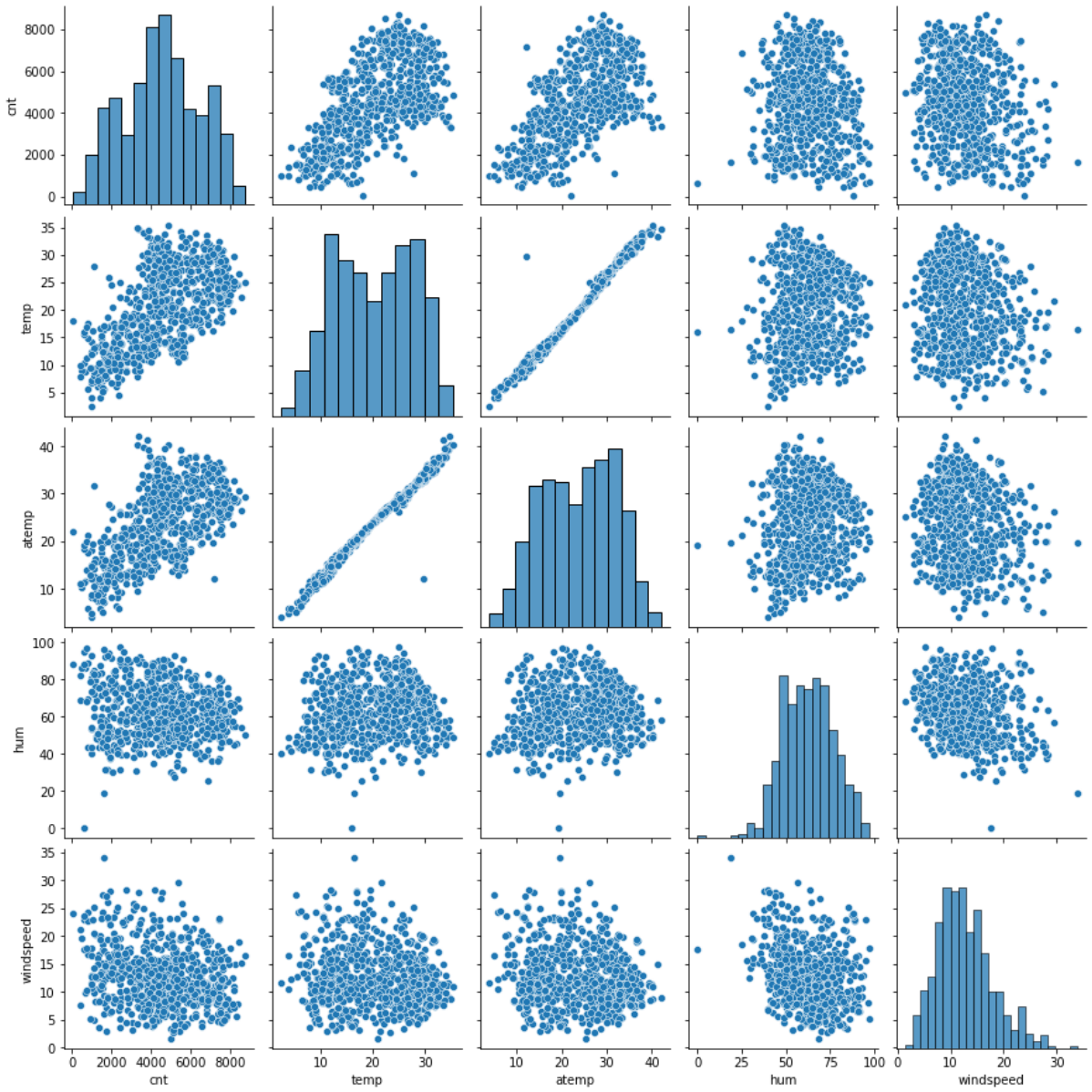
   The categorical variable in the dataset are **season**, **weathersit**, **holiday**, **mnth**, **yr** and **weekday**. These variables had the following effect on our dependant variable:-

   1. **Season -** Spring season have less number of cnt whereas fall had maximum number of cnt. Summer and winter had intermediate number of cnt.
   2. **Weathersit -** Clear, Partly Cloudy' whether shows highest number of cnt whereas heavy rain/ snow shows less or no users
   3. **Holiday** – Number of users reduces on holidays.
   4. **Mnth** – Month of September shows highest number of cnt and December shows less number of cnt.
   5. **Yr -** The number of users for year 2019 are more than 2018.

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

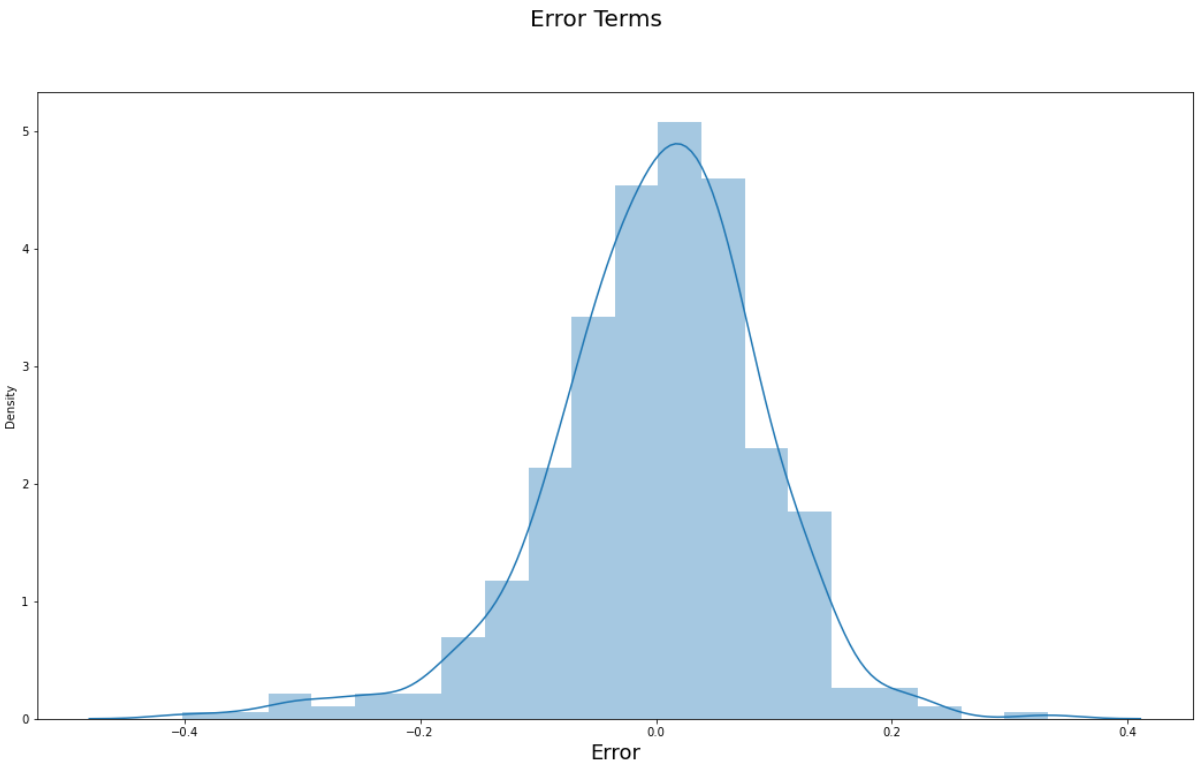
   The creation of dummy variable is necessary to convert categorical variable to numeric variable, however if we create the same number of dummy variable as number of categorical variables than it will lead to Multicollinearity.

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



Variables "**temp**" and "**atemp**" 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)**



We validate our model by plotting a histogram of residuals which help us determine the error terms are normally distributed and are centred are zero.

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

The top 3 features are:-

1. temp with coefficient : 0.5164

2. weathersit Light Snow & Rain with coefficient: -0.2837

3. yr with coefficient : 0.2324

General Subjective Questions

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

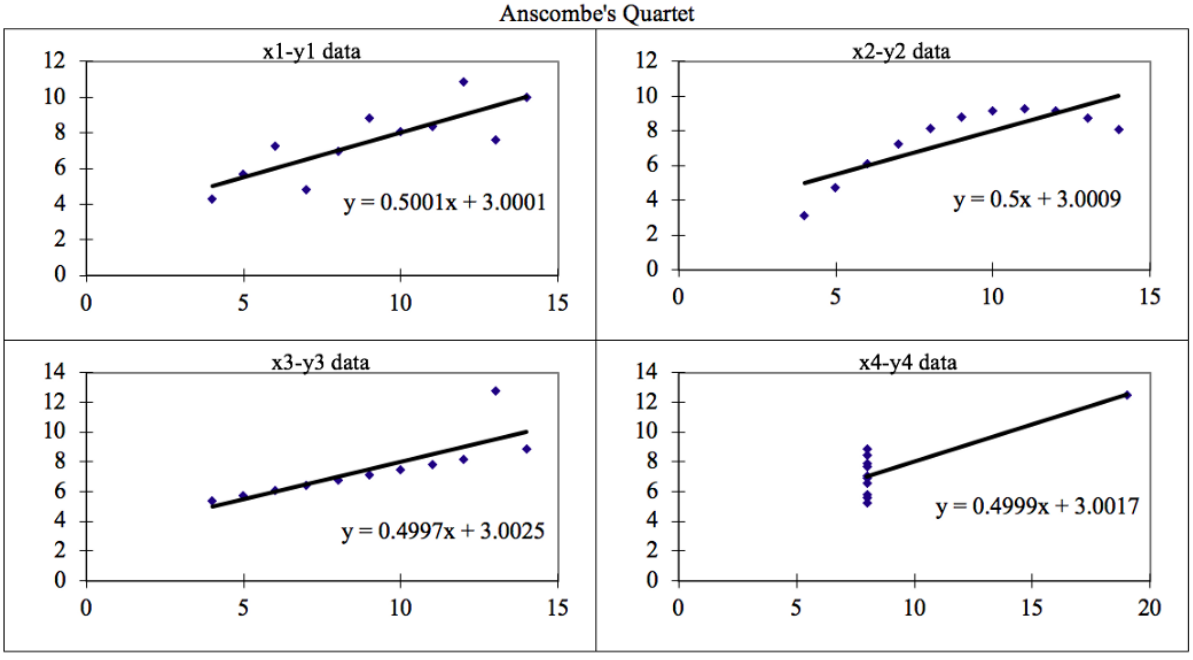
Linear regression is a machine learning algorithm used when there is linear relationship between dependent variable and independent variables. It is represented my equation of straight line i.e “y=mx +c” where y represent the dependent variable, x represent the independent variable.

There are two types of linear regression models namely

1. Simple Regression Model: When the dependent variable is dependent only on a single independent variable. It is represented by “ y = β0 + β1x. where y is dependent variable X is independent variable, β1 is coefficient of x and β0 is the intercept.
2. Multiple Regression Model: When there are more than single independent variable. It is represented by “ y = β0 + β1x1 +β2X2….βnXn. where βs are coefficient of independent variables and n is the number of independent variables.

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

Anscombe’s quartet is group of four identical dataset but have different distribution when represented over graph. It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.



The four datasets can be described as:

1. Dataset 1 fits the linear regression model pretty well.
2. Dataset 2 is unable to fit linear regression model on data quite well as the data is non-linear.
3. Dataset 3 shows the outliers involved in the dataset which cannot be handled by linear regression model.
4. Dataset 4 shows the outliers involved in the dataset which cannot be handled by linear regression model.

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

Pearson's R is a statistic that measures the linear correlation between two variables. It has numerical value that lies between -1.0 to +1.0. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations**.** It is calculated by below formula:

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

Where,
N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

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

Scaling is a step of data pre-processing which is applied to independent variables to normalise the data within a particular range. It also helps in speeding up the calculation in an algorithm.Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

There are two ways of scaling namely:

1. Normalising or Min/Max scaling:

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

2. Standardization Scaling:

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

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

VIF is variance inflation factor and it is calculated by formula:

(VIF) =1/(1-R^2) when R^2 is equal to 1 then then VIF value becomes infinity.

R^2 value will be equal to 1 when the dependent variable is completely explained by independent variable. i.e perfect correlation.

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

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. 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

Importance in Linear Regression:

1. When we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.
2. It can be used with sample sizes also
3. Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.