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: From the dataset, after plotting the Bar plot for categorical variables
   We can infer that

   i. Plot yr vs cnt:
   Demand of the bikes was increased in 2019 when compared to 2018

   ii. Season vs cnt:
   Demand for bikes higher in season fall, summer and winter

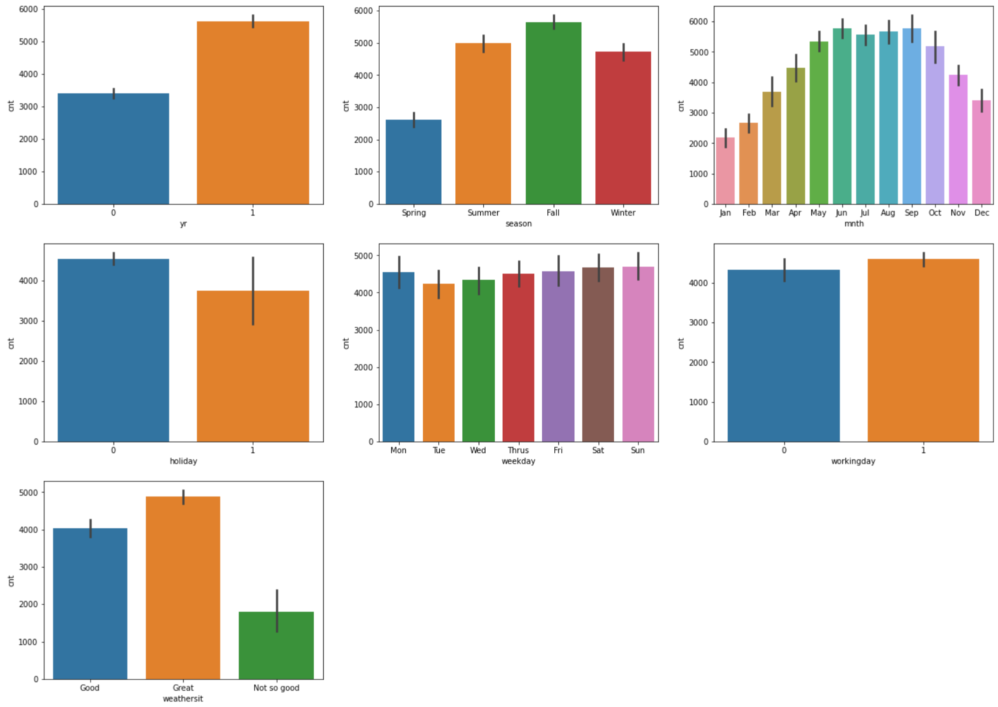
   iii. Mnth vs cnt:
   Demand for bikes is increased over the months.

   iv. Holiday vs cnt
   Holiday does show some effect on cnt.

   v. weekday vs cnt
   weekday does not show any effect on cnt.

   vi. wokingday vs cnt
   wokingday does not show any effect on cnt.

   vii. Weathersit vs cnt:
   Demand for the bikes was increased when there is good weather



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

Ans: For Categorical variables to fit in the model we need to convert he categorical as numerical variables for the we use one-hot encoding. We use get_dummies function from pandas.
So in One hot encoding lets say a column season will have 4 kinds summer, winter, rainy, spring
For this season column when we do one hot encoding it looks like

columns	summer	winter	rainy	spring
summer	1	0	0	0
winter	0	1	0	0
rainy	0	0	1	0
spring	0	0	0	1

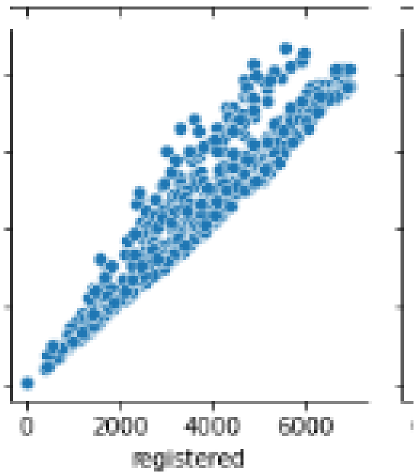
As we can instead of defining 4 levels for 4 kinds of season we can do the same thing at 3 levels and other level can be defined by 0s **which can increase the efficiency of encoding**.

We can define same thing by

columns	Level 1	Level 2	Level 3
summer	0	0	0
winter	1	0	0
rainy	0	1	0
spring	0	0	1

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

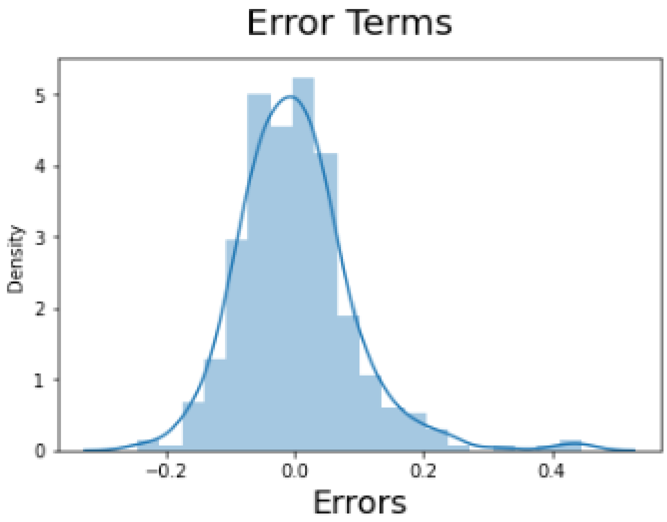
Ans: From the pair plot we can say that **registered vs cnt** shows the highest correlation.



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

Ans. We should perform: Residual Analysis of the train data

- To check if the error terms are also normally distributed which is infact, one of the major assumptions of linear regression, plot the histogram of the error terms.
- The error terms should be normally distributed at 0 with standard deviation as 1.



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.

1	windspeed	3.77
2	casual	3.57
3	workingday	3.27

From the RFE and final model after that we can say that windspeed, casual and working day are the top 3 features contributing significantly towards explaining the demand of the shared bikes

General Subjective Questions

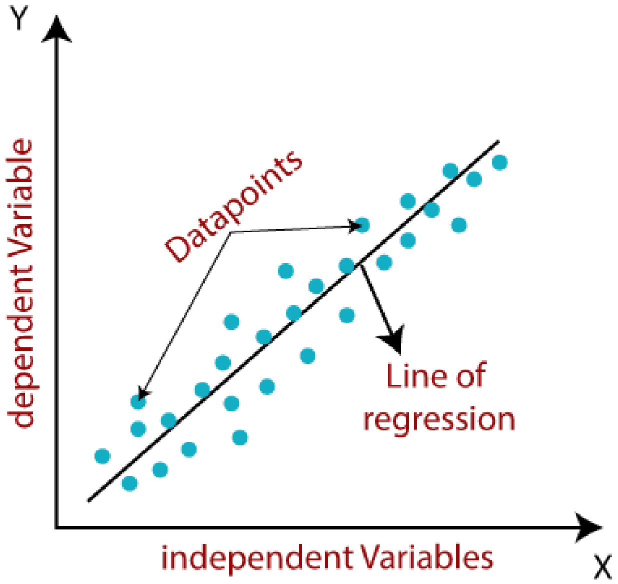
1. Explain the linear regression algorithm in detail. (4 marks)

   Ans.

   - Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

   - Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

   - The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Mathematically, we can represent a linear regression as:
y= B0 + B1*X + ε
Here,

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

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

   Ans:
   Anscombe’s quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

3. What is Pearson’s R? (3 marks)

   Ans.
   In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

   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? (3 marks)

   Ans:

   <u>Scaling:</u>
   Scaling is a geometric change that linearly enlarges or reduces things. A property of objects or rules known as scale invariance is that they remain unchanged when scales of length, energy, or other variables are multiplied by a common factor.
   Scaling law, a law that explains how many natural phenomena exhibit scale invariance.

   <u>Scaling performed because:</u>
   It is a data pre-processing procedure used to normalize data within a specific range by applying it to independent variables. Additionally, it aids in accelerating algorithmic calculations. The majority of the time, the obtained data set includes characteristics that vary greatly in magnitudes, units, and range.

The difference between normalized scaling and standardized scaling

The values of a normalized dataset will always fall between 0 and 1. A standardized dataset will have a mean of 0 and a standard deviation of 1, but the maximum and minimum values are not constrained by any specified upper or lower bounds.

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

   Ans.
   If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

   An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

   Ans.

   Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

   This helps in a scenario of linear regression 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.