Assignment-based Subjective

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

The Categorical variables in the dataset are: **Season, Yr, Month, Weekday, Holiday and Weathersit.** There effects on the Target / Dependent variable are as follows:

- **Season**: Boxplot clearly shows that value of cnt is highest for Fall and lowest for Spring. Summer and Winter values are in between them. This shows people are opting more for bikes in Fall season.
- **Yr**: Bike demand was more in 2019 than 2018.
- **Month**: Number of bike rentals is highest in month of September and lowest in month of January.
- **Weekday**: From boxplot, it shows that people opted bikes more on Saturday than any other day of the week.
- **Holiday**: Bike remand tends to reduce on a Holiday.
- **Weathersit**: People have opted for bikes more on a Clear/few clouds weather and almost none on a Heavy rain / snow which seems obvious.

Q2. Why is it important to use drop_first=True during dummy variable creation? (2 marks)

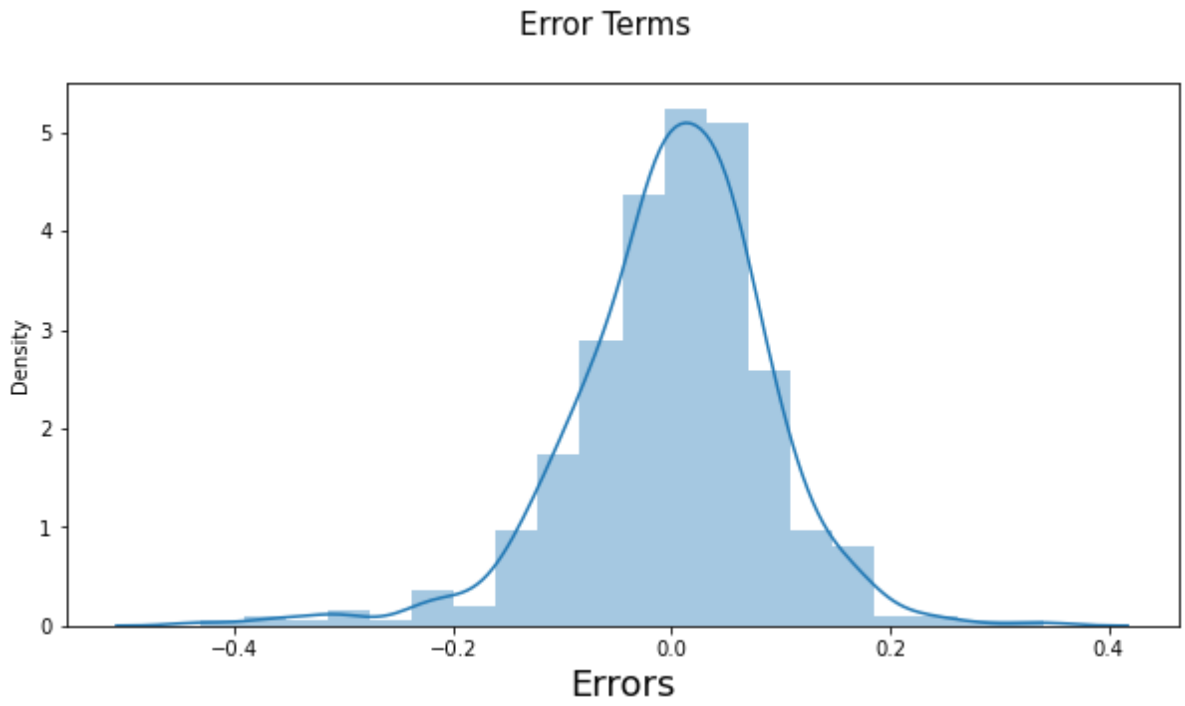
Dropping first column while creating dummy variables is done to avoid chances of getting high correlation between the variables. If there is high correlation between the independent variables then it will adversely affect the predicting capability of model and it becomes worse when cardinality is low. Another reason to remove the first column that it restrict Multicollinearity between the dummy variables as well.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Looking at the pairplot, ‘temp’ and ‘atemp’ variables have shown highest correlation with the target variable ‘cnt’.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

This is achieved by plotting Distplot of the Error Terms distribution which is as follows as found in our analysis:



The error terms distribution should be a Normal Distribution which mean centred around 0. If this condition satisfies then it validates that our model stands on the assumptions of the Linear Regression.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

The top three features which contributes most in explaining the demand of shared bikes are:

a) **Temp** – It has the coefficient value of 0.4332290
b) **Yr** – It has the coefficient value of 0.235222
c) **weathersit_Light_Rain_Snow** – It has coefficient value of -0.291838

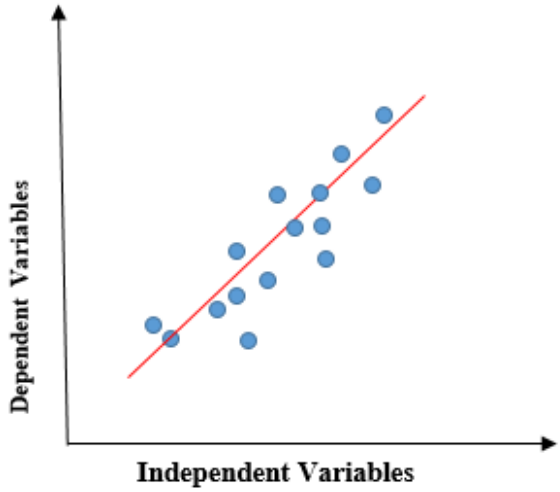
General Subjective Questions

Q1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is a supervised machine learning modelling method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression.

The linear regression model gives a sloped straight line describing the relationship within the variables.

Below graph shows the linear relationship between X and Y variables:



The Linear Regression equation is :

$$y = mx + c$$

where,

y → Dependent /. Target variable
m → Slope
c → Intercept
x → Independent / Predictor varible

In above equation, there is a single predictor variable so it is called as **Single Linear Regression** but when there are more than one predictor variables then it is called as **Multiple Linear Regression**. This is shown in equation as follows:

$$\textit{observed data} \rightarrow \quad y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p + \varepsilon$$

$$\textit{predicted data} \rightarrow \quad y' = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

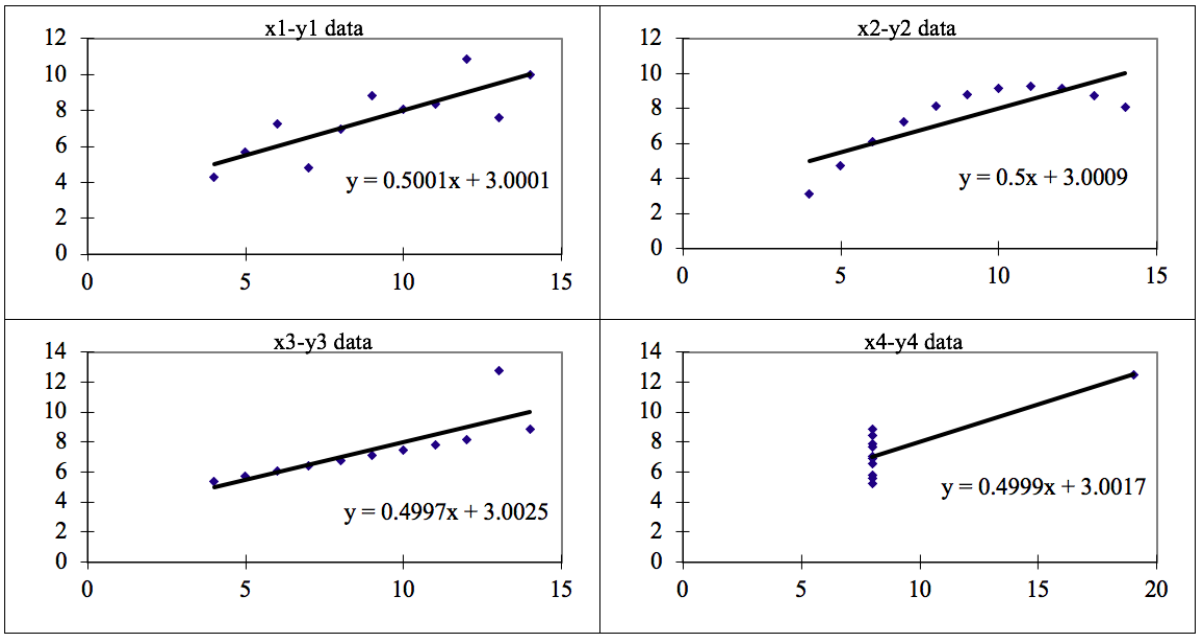
$$\textit{error} \rightarrow \quad \varepsilon = y - y'$$

Error shows the difference between the Predicted values and the actual / observed values.

Q2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet can be defined as a set of four dataset which have very similar behaviour as per statistics values such as Mean, Variance etc. but when they are visualised to see the data pattern then it would come out that only one of them is valid for Linear Regression modelling and other are just to fool the model.

Below is the representation of such four datasets:



The four datasets can be described as:

1. **Dataset 1:** Fits Linear Regression.
2. **Dataset 2:** Not a good fit as many data points are showing divergence behaviour.
3. **Dataset 3:** Not a good fit due to outliers.
4. **Dataset 4:** Not a good fit due to many outliers.

So, this concept focussed on the importance of plotting the dataset to see the actual data distribution pattern and then only one should go for modelling. Modelling can't be started just by seeing the statistical behaviour of the dataset.

This concept was coined by **Francis Anscombe** in 1973.

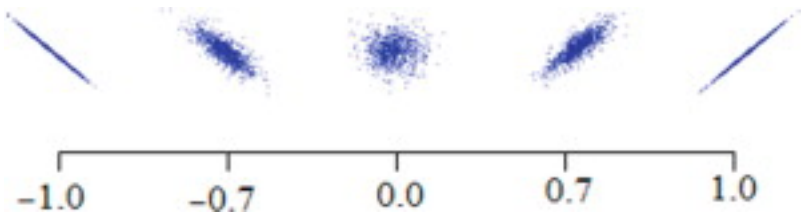
Q3. What is Pearson's R? (3 marks)

Pearson's R method is used to determine correlation between numerical variables. It value ranges from -1 to +1. It is represented by symble 'r'. Values of 'r' is interpreted as follows:

r = 0 → no correlation.

0 < r < +1 → Positive correlation
-1 < r < 0 → Negative correlation.

This pattern can be visualised as follows:



Pattern at +1 and -1 shows perfect correlation thus Clear Linear distribution among data.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is done to normalise the values of the variables / columns. This is done to bring all variables values in a particular range so that while doing modelling, outliers or high values don't get higher weightage and lower values should get ignored. This is done at Data Preparation stage. Due to scaling, model is able to utilise all values of variables correctly and predict the target variable values efficiently.

Scaling can be done through two methods:

1. **Normalization Scaling** – This method is performed when given dataset doesn't follow Gaussian distribution. Here, Average and Min value of dataset is used in below formula:

$$X_{new} = \frac{X - X_{mean}}{X_{max} - X_{min}}$$

2. **Standardised Scaling** – This is done when data follows Gaussian distribution but this is not always true. Also, there is no range here so outliers are not affected by this method of scaling. Here , z-value is used in below formula:

$$X_{new} = \frac{X - X_{mean}}{\sigma}$$

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

VIF is calculated to find the multicollinearity between variables used for prediction. If there is high value of VIF then it shows those columns/variables have multicollinearity and thus are correlated. Such variables are removed as they can reduce the significance of model.

Now, if VIF value is infinite then it shows a perfect correlation of that variable among the other variables i.e. value of 'R' getting 1 and it is case of extreme Multicollinearity. In this case, that variable would be removed as VIF lower than 5 is often considered as suitable for modelling.

VIF formula is given as below:

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

So, if R is 1, then VIF = 1 / (1-1) = Infinite.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Q-Q plot is known as Quantile-Quantile plots. In this, plot is drawn between quantile of a sample distribution against the quantile of a theoretical distribution. This helps to determine whether the sample follows which sort of distribution such Normal, Uniform, Exponential etc.

It helps to understand:

- If the two datasets have same distribution.

- If the residuals follow Normal Distribution which is an important assumption in Linear regression.
- Skewness of the dataset
- If the two datasets have common location and scale
- Tells tail behaviour of two datasets.