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)

To analyze the categorical column, I used the box plots and bar plots to assess the relationship between the categorical columns and the dependent column i.e. count

Below is my finding,

- **Season** –
  - The maximum number of bookings are made in **fall** season which indicates
  - The number of booking increases from spring < summer < fall and then reduces in winter.
  - The booking for each season is increasing year by year.
- **Month** –
  - The highest number of booking is done in month of **September.**
  - The number of bookings are increasing from Jan->Sept and then decreases month by month till December.
  - The number of bookings for each month is increasing year by year.
- **Weekday** –
  - The most number of bookings are made on Fri, Sat
  - This clearly indicates that the customer is opting to make the booking on weekends.
- **Holiday/Workday –**
  - Most number of bookings are made on Holidays in comparison to Workdays.
  - People tend to make lesser bookings on workdays
- **Year**
  - The number of booking is increasing year by year

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

When the dummy variables are created on the categorical columns, For Each Categorical Column where k levels are present, k dummy columns are created. These dummy columns are highly correlated with each other and to reduce this correlation, the first column is suggested to be dropped. This dropped column can be easily derived using the rest of the columns since these are correlated.

So if we use drop_first = false

For k levels, k dummy columns will be created

If we use drop_first= true,

k-1 dummy columns will be created as the first column will be dropped

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

Looking at the pair plot, The columns ‘temp’, ‘atemp’ are highly correlated with the target column count.

Since ‘temp’ and ‘atemp’ are also showing very high correlation i.e. 99%, we have dropped ‘atemp’ column.

Hence we can say that ‘temp’ column is highly correlated with ‘count’ target column.

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

The validation of assumptions are done using the 5 assumptions,

1. **The Normality of Errors**
   - This was achieved by plotting the residuals (y train – y pred) using a distribution plot where the mean was centered to the value of 0 and the distribution is following a normal distribution.
2. **The Multicollinearity**
   - This was achieved by plotting a heatmap on the training set data and there is no strong collinearity between two independent columns. Industry standard of 70% + is considered.
3. **Linear Relationship**
   - By giving an example of CCPR, The ‘temp’, ‘wind-speed’ columns are showing linear relationship.
4. **Homoscedasticity**
   - This is achieved by plotting a scatterplot between residual and count column (Target).
   - There is no visible pattern in residual values
5. **Independence of correlation**
   - There is no auto correlation

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

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes:

1. Temp = 0.432
2. Light_snowrain = -0.288
3. Year = 0.235

General Subjective Questions

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

Linear regression Algorithm is a statical model that is used on the continuous variables and which analyses the relationship between the target variable with the other independent variables.

The relationship between the independent variable vs target variable can be in both positive and negative.

This relationship can be mathematically represented with below equation,

Y=mx +c

Where m = coefficient of x

C = constant

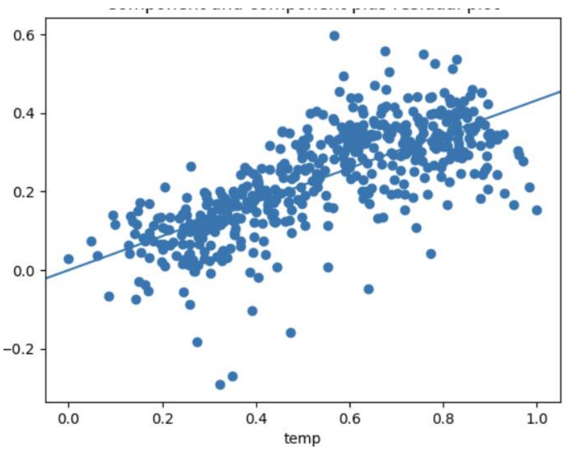
Y = target variable

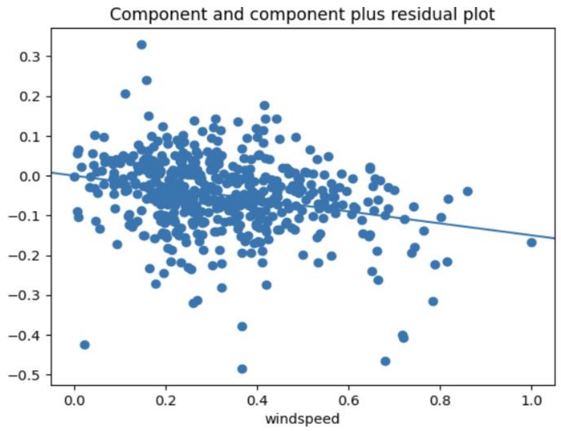
X = independent variable

The Relationship can be both positive and negative.

Below example shows both the relations using the same shared bike assignment

There is a positive relationship between temp and count variables and also there is a negative relationship between the windspeed and count variable as shown below





There are two types of Linear regression:

1. Simple Linear Regression – **SLR**
2. Multiple Linear Regression – **MSR**

There are assumptions that are followed:

1. **Multicollinearity** – There should be no collinearity between the independent variables.
2. **Linear Relationship** – The independent variables must show linear relationship with target variables.
3. **Normality of Errors –** The residuals must be normally distributed and must be centered at mean at 0.
4. **Auto Correlation** – Auto correlation happens when there is a correlation between the residual errors
5. **Homoscedasticity** – There should be no visual pattern between the residuals values.

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

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

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

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

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