Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the data-set, what could you infer about their effect on the dependent variable?

- For the categorical variables

- In the year 2019, the count of total rental bikes were more when compared with the rental bikes in yr 2018

- In spring season, the renting of bikes were low when as compared with other seasons. Fall season had the most bookings and summer was also a good season as seen for booking sales

- The bookings in each month were gradually increasing untill October and there was a sudden mid drop for November and December. Majority bookings were done from June to October

- As for weather conditions, there were no bookings made in extreme conditions. And had most bookings in the clear conditions and then some bookings were made in misty conditions as well

- For holiday and working days, there is no specific comparison as to when bookings are made more, its more like these conditions do not much affect on the bookings

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

- We use drop_first=True during dummy variable creation to remove the redundant column which reduces complexity and also avoids multicollinearity. Also when the column is removed, its presence is then shown the values in other columns.

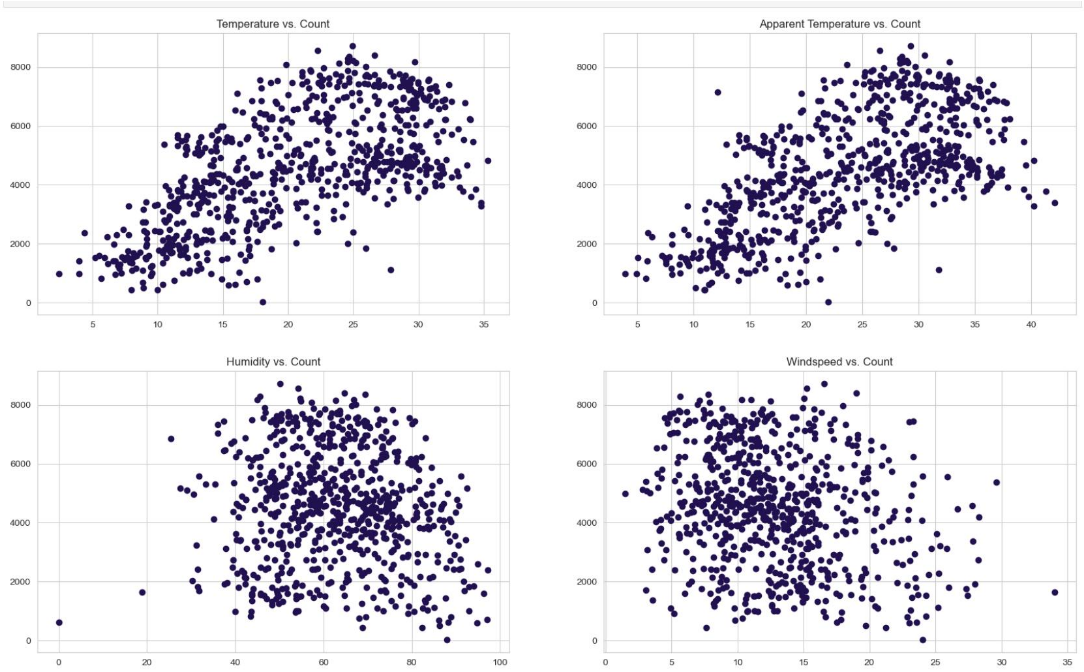
And it comes down from N levels to n-1 dummy columns to represent the information.

With drop_first=True, the first column is dropped and if we want to know the dropped column's presence, the rows with n-1 columns will all have 0.

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

- The temp has the highest correlation with the target variable cnt. Here's a scatter plot for the same

It also has the highest coefficient in the final equation.



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

➢

a. By checking the error terms distribution - it is a normal distribution and mean is close to 0.

b. Checking the relationship between y_train and y_train_predicted which tells the graph is linear

c. Checking multicollinearity by measuring VIF(variation inflation factors). If VIF >5 or 10 means there is a multicollinearity

d. Ensuring the generalization of model by testing it on test sets to see the case of over-fitting.

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

➢ Based on final model the equation which we get is
cnt = 0.13 + 0.23 x yr - 0.09 x holiday + 0.51 x temp - 0.14 x windspeed + 0.10 x summer + 0.13 x winter - 0.08 x mist_and_clouds - 0.28 x stormy_condition + 0.05 x August + 0.11 x September

Features which are contributing significantly are:
- temperature (temp)
- calendar year (yr)
- winter season (winter)

General Subjective Questions

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

- ➢ Linear Regression model is a way to statistically know the relationship between 1 dependent variable and one or more independent variables. Linear regression can be Single LR or Multiple LR depending upon the no of independent variables. It is used to predict the upcoming business problems based on past facts and figures.

Types:
A. Single Linear Regression Model

- this is the case when there is 1 independent variable , and the model is represented as

$y = \beta_0 + \beta_1 x + \varepsilon$

B. Multiple Linear Regression Model

– this is the case when there is more than 1 independent variable , and the model is represented as

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$

Steps we follow while building Linear Regression Models

- ● Data Understanding and Preparation - It includes cleaning the data like outliers, null values. Also if there is a need for scaling we do that to make the comparison on even terms, which can be done either by using MinMaxScaling or Standardization processes.

- ● Training the Model - Splitting data into training and test sets. And then fitting the linear model to training sets using OLS method

- ● Making predictions and Model Evaluations - Once the model is created, we make predictions and evaluate the model's performance in test sets so as to confirm if the model is working fine with making good predictions or not

- ● Assumptions - Need to check if assumptions are followed or not which includes linearity, independence of errors, normality of residuals and homoscedasticity.

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

- ➢ It is a set of 4 datasets which have similar statistical properties like mean, median, variance and correlation but when visualized, they seem different. This shows that datasets with similar statistics can have different patterns when graphed

With this we can say that

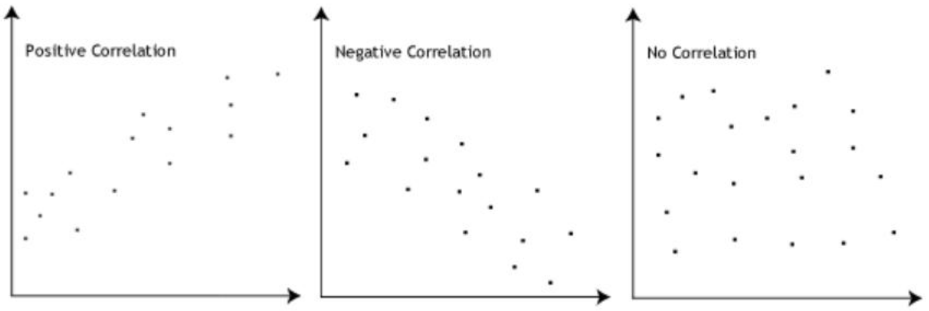
- ● -> visualization is important in understanding data other than statistics
- ● -> Outliers can influence results if not taken care of
- ● -> A high correlation coefficient doesn't mean a strong relationship

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

- ➢ Pearson's R is the measure of linear relationship between 2 variables. It is indicated by 'r' and ranges from -1 to 1.

Here;

1 indicates - a perfect positive linear relationship
0 indicates - there is no relationship or any kind of dependency between 2 variables
-1 indicates - a perfect negative linear relationship



Also note that:
It is valuable only for continuous variables
It does not work with outliers

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

- Scaling is a part/process used in Model Building especially during preprocessing of the data.

It refers to the changes made in the values of a variable and have them under a specific range or distribution so as to compare them with other variables while building model.
It provides equal weightage and helps perform models in a better way on a similar scale.

Types of scaling:
1. Normalized Scaling - It scales the values in the range of 0 to 1.
Its calculated as:
X(normalised) = X - min(X) / max(X) - min(X)

2. Standardized Scaling - It scales the values to have a mean 0 and standard deviation 1
Its calculated as:
X(standardized) = X- mean(X) / SD(X)

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

- VIF also known as Variance Inflation Factor, it helps in determining multicollinearity in multiple regression model where more than 1 factor are used to build model. It quantfies how much the variance of the estimated regression coefficients is increased due to multicollinearity.

Formula of VIF
VIF = 1 / 1- R^2

This R^2 is obtained by regressing that factor against all other independent variables.

Now, if the VIF is infinite, it indicates perfect multicollinearity between variables i.e. one or more variables are perfectly correlated with each other.
We can also say that there is a redundant information - one variable is expressing as a perfect linear combination of others

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 (Quantile-Quantile) plot is a graphical tool used to assess normality of residuals also termed as normal distribution. It compares data between observed data and expected data distributions.

Its uses:

1. To check normality of residuals. If they are normally distributed, points on Q-Q should be a straight line

And if they deviate from the straight line, it indicates non normality due to reason like non linear relationships or outliers.

2. To evaluate model's assumptions like linearity or constant variance

3. To find outliers when visualized in plot

They are thus important for Model Evaluating and improving its accuracy.