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 – We have almost 7 categorical variables such as season, weathersit, holiday, weekday, working day, year and month where we see season has an impact where the rental count decreases in spring season. When there is bad weather like rain and snow it affects the booking which is obvious. In case of holiday, we see good number of bookings for both the section when it is holiday or not but during holiday the amount of booking is more showing a good business growth. During the weekday it seems not that much affecting as it shows almost similar graph. And for non-working days it has almost same mean but the cluster is bigger on the non-working day. The trend for month seems going higher and then coming down saying that there is a high demand during some particular month in a year and it also shows the growth of service increasing year on year.

In a whole if we see it shows that the demand is high on favourable circumstances like if it is a holiday with good weather condition and right season, for rest of the section it does not show complete down business but it shows a drop in the demand.

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

Ans – In case of dummy variable creation it has been seen that columns can be represented in binary value format such as gender or days in a week it says for example if it is Sunday then it cannot be any of the day. So, when we create dummy variable let say for weekday which has 7 days it can create 7 dummy variables where only one weekday value would be 1 rest would be 0 consisting of 7 columns but the same thing, we can represent in 6 columns because all the columns value is 0 it means the value is the weekday which is not present in the columns mentioned. Same goes with gender if it is not male it can be taken to be female.

So, because of this reason we put **drop_first=true** which creates **(n-1)** columns

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

Ans – Looking at the pair plot we see temperature having the highest correlation with the target variable

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

Ans – There are some statistical values which we should verify to confirm that the model is good enough. They are F-Statistic which should be high at the same time probability of F-statistic should very small as close to zero. It should have a positive R square and adjusted R square along with it the p value should not be greater the 0.05 and the VIF should be in a range what we have set it for.

For a matter of fact, we can segregate the combination of VIF and P value into four combination such as

1. High P value High VIF
2. Low P value Low VIF
3. High P value Low VIF
4. Low P value High VIF

Here we see the first and the second ones are easy to handle but in case of 3rd and 4th we should always consider the high P value and then consider the high VIF.

For this model the above things were considered properly.

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 – The top three features contributing could be

1. Temperature
2. Season
3. Weather/Windspeed

General Subjective Questions

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

Ans – Linear Regression is the basically the first machine learning algorithm which is a supervised way of learning. It can be said as a statistical way which we use to determine the dependent variable with the help of one or many independent variables, there can be mainly examples in this regard like predicting the marks, house prices demand in bike rental or and subscription demand on the OTT platform.

The Linear Regression is basically two types of: -

1. Single Linear Regression
2. Multiple Linear Regression.

Single LR is basically based on the fact that there would be only one only one independent variable upon which the unknown or the dependent variable will be predicted.

On the other hand, in case of multiple LR we have multiple independent variable which may add up to get a good prediction for the dependent variable. In case of multiple LR it might be possible that some R square gets added up to boost up the resultant R square but some cannot so choose a set of the independent variable for the final model is something we do to get the best set of variables which would build a good model.

The equation for the LR is y = c + mx

The equation for the multiple LR is y = c + mx + m1x1 + m2x2 + m3x3 + ……

After we have built the model there comes lot of steps such as residual calculation which tells us what's the difference between what we have and what we got. Which is also useful to know and calculate the R square which represents the model accuracy.

The formula for R square is $R^2 = 1 - (RSS / TSS)$

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

Ans - Anscombe's quartet tells us that it is not always significant that getting an exact definition mathematically proves that the thought process is correct as visualization can bring in some perspective to the study and solutioning which we were not able to see via the calculation.

So Anscombe's quartet tells that when there are 4 datasets taken which looks to be similar and has all the descriptive statistic's similar such as (mean, variance, correlation).

Even though they have the stats similar when we plot it shows that the graph representation is different.

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

For the above sample data taken If we calculate the statistical measure we get: -

Average Value of x = 9

Average Value of y = 7.50

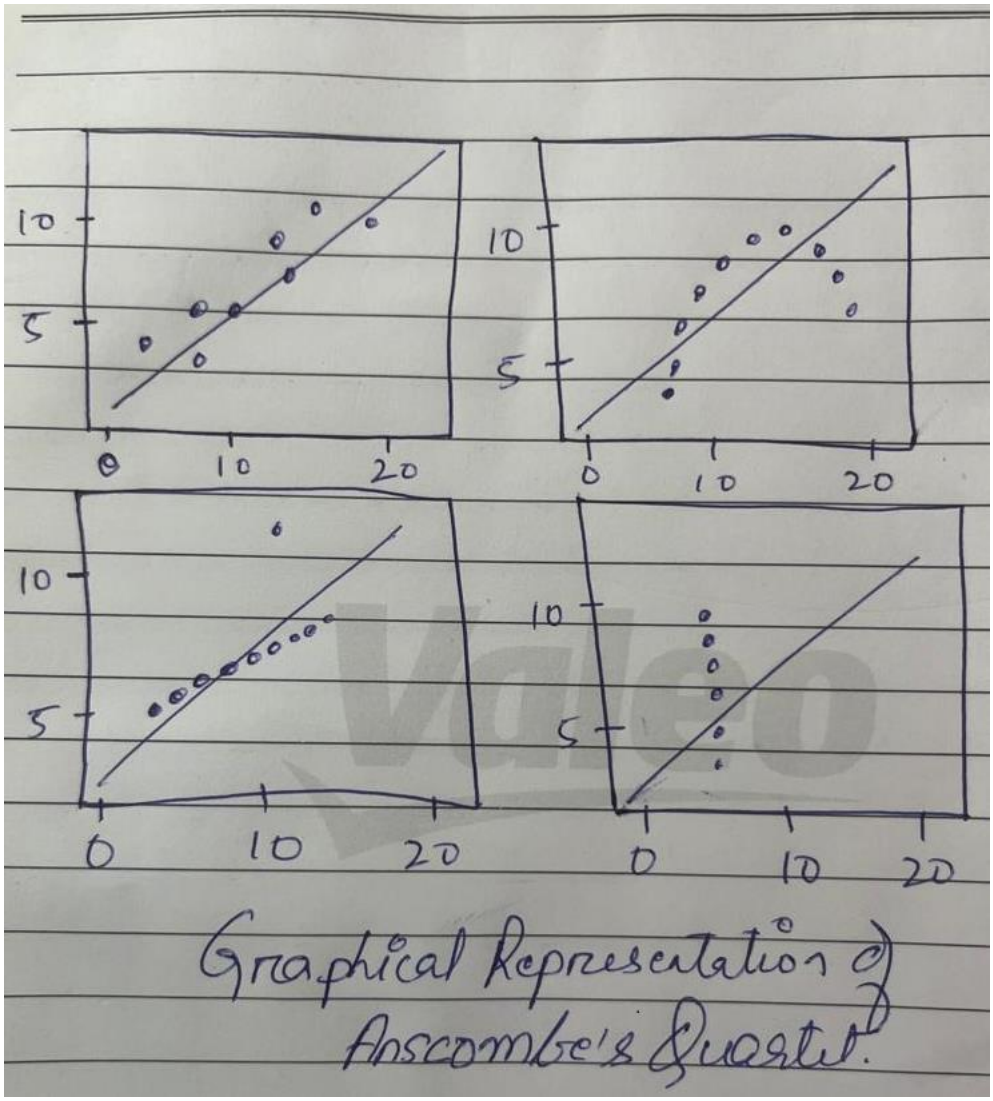
Variance of x = 11

Variance of y =4.12

Correlation Coefficient = 0.816

Linear Regression Equation: y = 0.5 x + 3

But when we plot the graph for the above dataset, we get the below result.



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

Pearson's correlation coefficient (r) is a measure of the linear association of two variables it is mostly used in finding the relationship between two variables.
Correlation analysis usually starts with a graphical representation of the relation of data pairs using a scatter diagram when we want to visualize the relation. The values of correlation coefficient vary from –1 to +1.

Positive values of correlation coefficient indicate a tendency of one variable to increase or decrease together with another variable. Negative values of correlation coefficient indicate a tendency that the increase of values of one variable is associated with the decrease of values of the other variable and vice versa. Values of correlation coefficient close to zero indicate a low association between variables, and those close to –1 or +1 indicate a strong linear association between two variables. The square of the correlation coefficient is the coefficient of determination, which gives the proportion of the variation in one variable that can be explained from the variation of the other variable.

So, we can say that

1. If value is +1 It has very high positive correlation.

2. If value is +0.6 It has strong positive correlation.

3. If value is +0.3 It has positive correlation.

4. If value is 0 It has no correlation.

5. If value is -0.3 It has strong negative correlation.

6. If value is -0.6 It has strong negative correlation.

7. If value is -1 It has very high negative correlation.

Also, if it shows high correlation not always it should be taken as proof for fact rather it should be considered as a part which shows relation which can be used to find some more insights.

It is normally used in case of finding the best fit line where we see which all the data points are not properly correlated.

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

Ans – Scaling here means making the values of columns to land in a particular unit or range so that everything can become comparable.

For example, we cannot compare kilometres with miles or rupees to dollars.

Scaling is performed so that all the values for all the continuous columns they come in a particular range which is -1 to 1 in case of Min max scaling which is mostly preferred. So that the further model development and the comparison performed by the libraries to build will be right.

The basic difference between normalized scaling and standardized scaling is that in case of standardized scaling the outliers in the data might not be impacted by this way of scaling which in case of normalization gets impacted also standardized does not have any range restrictions but normalized have range restrictions.

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

Ans- Yes we have observed high VIF among the temperature, atemp and humidity variable.

We have observed this and the reason for having high VIF is that when the variables are not orthogonal in nature, they persist to have high VIF.

For example, if we see temp and atemp they both represent almost the same thing so they had a very high VIF and we also saw when we released one the VIF went down.

This also means that when we have this type correlated variables removing one of them is a good way to form a good model as they do not add up anything together.

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

Ans – In one line if we have to say we can state QQ plot as 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.

The main use and importance of QQ plot is by doing we can determine and it helps to determine if the underlying dataset follows any particular type of distribution such as normal, uniform distribution etc.