Assignment based subjective questions:

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

A1: Based on my analysis of the categorical variables in the data set, the following were observed by me:

- fall season attracts more bikers and hence has recorded higher bookings.
- the period from May to September has seen most number of bike sharing bookings, possibly because may to june is a holiday season and hence many people might travel a lot.
- Thursday to Sunday have more bookings as compared to first three days of the week.

- clear weather attracts more bookings.
- bookings are less when its not a holiday as people might like to spend time at home.
- 2019 has seen more bookings overall as compared to 2018.
- similar bookings were found whether it is a working day or not a working day.

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

A2: drop_first=True is used because of the fact that a categorical variable with N-levels can be easily described with N-1 dummy variables. So by dropping one of the N- dummy variables being created, we can reduce the complexity of the dataset.

For example, if a categorical variable named "Car brands" has three categories namely ['Tata', 'Mahindra', 'Maruti'] by creating dummy variables only for Mahindra and Maruti, we can easily map all the three brands as

Mahindra	10
Maruti	01
Tata	00

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

A3: From the heatmap or correlation plot, 'temp' has the highest relationship with the target variable.

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

A4: After building the model, I validated the model by observing the distribution of residuals/error between y_test_predicted and y_test and checking whether if the distribution is normal or not.

I also checked for multicollinearity among the variables from the final linear regression model

I also did the homoscedasticity test to see there is no specific pattern between residual and target variables.

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

A5: The following 3 variables are having highest effect on the final model

1. temp
2. winter
3. sep

General Subjective Questions:

Q1. Explain the linear regression algorithm in detail.

A1: Linear regression is a model where the dependent and independent variables are correlated using a linear equation of the form

$Y = mX+c$

There may be only one dependent variable in which case the regression is called simple linear regression

If there are more than one dependent variable, then the linear regression is called multiple linear regression.

Multiple linear regression can be represented as

$Y = m_1X_1+m_2X_2+\ldots+c$

Depending on the value of coefficients of dependent variables, we can assess whether a variable can have positive impact (+ve coefficient) or negative impact (-ve coefficient) on the target variable

Some of the assumptions made by a linear regression model are

1. there should be no multi-collinearity among dependent variables
2. Linear regression model should have very little to no auto-correlation among dependent variables
3. The residual/error terms should be normally distributed with a mean of 0.
4. There should not be any visible trend between residuals and target variable.

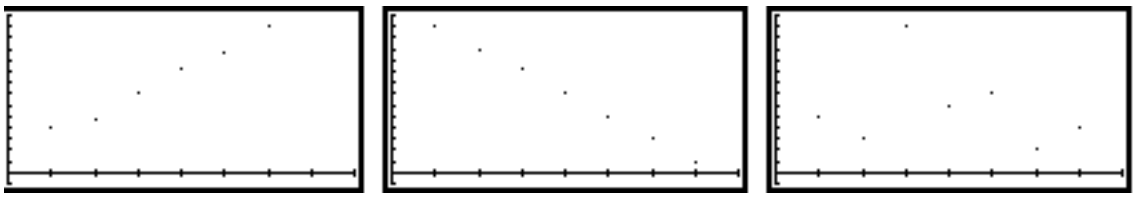
Q 2. Explain the Anscombe's quartet in detail.

A2: Anscombe's quartet is a compilation of 4 data sets, each having a set of 11 (x,y) pairs that look identical with similar descriptive statistics but show entirely different plots when plotted. The Anscombe's quartet is developed by French scientist Francis Anscombe to illustrate the importance of plotting a data to understand its behaviour and distribution.

It was developed to show that not all numerical calculations are exact and that graphs are sometimes more realistic.

Q3. What is Pearson's R?

A3: Pearson's R is a value associated with the strength of the association between two variables. It usually takes values between [-1, +1] with any value below 0 indicating negative correlation and any value above 0 indicating positive correlation. Value 0 indicates no correlation among variables. The following figure shows Pearson's R significance in deciding how the points are distributed about the line.



R > 0 R<0 R=0

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

A4: Data scaling is a technique to handle such attributes in the dataset whose values are peculiarly high in such a way that, their coefficient can cloud the significance of coefficients of other attributes. Scaling usually converts all the data in such a way that the model doesn't weigh any particular variable higher than the other variables just because of its values in the dataset. Thus by doing this, the coefficient and significance of numerical variables will be on par with categorical variables and the final model turns out to be reliable. Scaling is again done by two ways

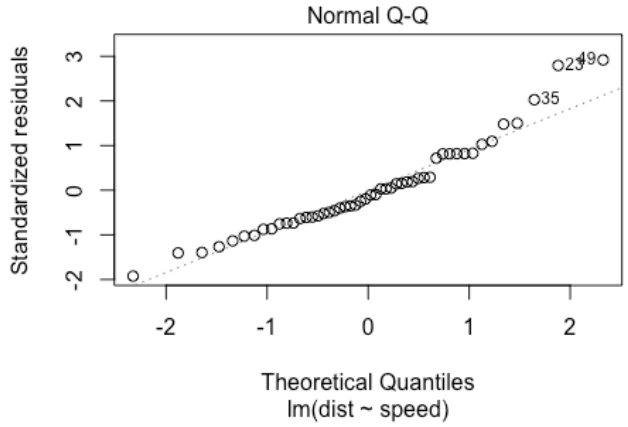
MinMax scaling or Normalized scaling	Standardized scaling
1. Scales values between [0,1] or [1,1]	Scales without any bounds
2. Minimum and maximum values of the particular attribute are used for scaling.	Mean and standard deviation of the data are used for scaling
3. It can be done using MinMax scaler from scikitlearn.	It can be done using StandardScaler from scikitlearn.
4. It is used when we cannot have features at different scales	Can be used when different features can be scaled in different range of values.

Q 5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

A5: VIF = infinity implies perfect correlation. This happens if any two variables are so perfectly correlated. This can be because of the R^2 value might be equal to 1 and hence VIF = 1/(1-R^2) becomes infinity. this can be avoided by dropping one of the two variables that is causing perfect correlation.

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

A6: A Q-Q plot or quantile-quantile plot is a plot between quantiles from two samples where our focus should be on the "y=x" line or the 45^{o} line which is indicates the distribution of the points. If all the points lie along the line, the data can be assumed that the data has been taken from population with same distribution. The following diagram shows the normal Q-Q plot



This is particularly used in linear regression in times where we get the training and test datasets separately. We can use this QQ plot to check whether both datasets are from population with same distribution or not.

It is mainly used to check

1. If data sets have come from same population.
2. If datasets have common scale.
3. Whether the datasets have similar distribution or not.
4. Whether the datasets don't have any outliers or are skewed.