Assignment-based Subjective

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

Below are the 5 categorical variables in the dataset.

a. **Season:** Demand for the bikes seem to be very less during spring season and it seems to be on the higher side in fall season.
b. **Yr:** The demand has risen considerably during 2019 compared to 2018.
c. **Mnth:** Demand seems to be high in September month and low in the month of January.
d. **Weekday:** There is not much of a difference in the demand during various days of a week. So, weekdays might not affect the demand.
e. **Weathersit:** The demand for bikes is very low during Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds days and high during Clear, Few clouds, partly cloudy, Partly cloudy days.

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

If there is a categorical variable with n levels we need only n-1 columns to represent this categorical variable. Drop_first=True helps in removing that one extra column. For example, if we have a categorical variable bloodgroup with values A, B, AB and O. We can represent these with only 3 columns. As the table below shows:

	A	B	AB
A	1	0	0
B	0	1	0
AB	1	1	1
O	0	0	0

In the above table we don't need to represent O separately because all 0s will be the implicit representation. In this case we use drop_first=true

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

Temp and atemp variables seem to have highest correlation with the target variable by looking at the pair-plot for numerical variables.

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

The below are the assumptions of linear regression:

a. **There is a linear relationship between X and y**: For this we have created a scatter plot and seen that some of the numerical variables have linear relationship with the Dependant variable Y.
b. **Residuals are normally distributed with mean 0.0**: We can validate this using Residual Analysis. Creating the distplot for the errors, which is the difference between actual

value of y and predicted value of y. If the distplot is showing a graph with normal distribution around the mean of 0.0 then this assumption is valid.

c. **Error terms are independent of each other**: Created a scatter plot between the actual values of y and errors to see if there is any relationship between each other. There is no relationship, so it is a valid assumption.
d. **There should not be any perfect Multicollinearity**: We have checked the vif values for all the selected features and made sure only the features with < 5 VIF values are selected.

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

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

1. Temp
2. Hum
3. windspeed

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning algorithm. It computes the linear relationship between a dependant variable and one or more independent variables. If there is only a single independent variable then it is called Simple Linear Regression, if there are more it is called Multiple Linear Regression.

1. The core assumption is that there is a linear relationship between the dependant variable(target) and independent variables (features).

2. Mathematically it is represented as below:

   $y=b0+b1\cdot x1+b2\cdot x2+\ldots+bn\cdot xn+\epsilon$

   y -> independent variable

   b0 -> intercept

   b1, b2, b3 …-> coefficients

   x1, x2, x3 … -> features

3. The goal of linear regression is to find the values of the coefficients that minimize the sum of squared differences between the predicted and actual values. This is least squares method.

4. **Assumptions in linear regression**
   a. Linear regression assumes a linear relationship between the target and the features.
   b. Observations should be independent of each other.
   c. Residuals (difference bwt actual and predicted values) should have constant variance across all levels of features.
   d. Residuals should be normally distributed.
   e. Features or predictor variables should not be highly correlated with each other.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet tells us that it is always important to plot your data. Let us say we have identical set of 4 datasets with same summary statistics like mean, variance, Rsquared, correlation coefficients, line of best fit. Ideally since all these statistics are the same for all the 4 data sets we could assume that when plotted they will be similar. But when we actually plot the data it could be the case that these are very different graphs. Hence Anscombe's quarter states that it is highly important to always plot your data rather than relying on summary statistics alone.

3. What is Pearson's R?

Pearson correlation measures and analyses the linear relationship between two variables. We can determine how strong the correlation is and in which direction the correlation goes. We can determine how large the linear relationship between two variables is using Pearson coefficient r. r value always lies between -1 and 1

Calculation : r = covariance/(std deviation X * std deviation Y)

Below table explains how the value of r can be used to determine the strength of the correlation.

Pearson's R value	**Strength of the correlation**
0	No correlation
-1	Negative correlation
1	Positive correlation

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

A dataset can have different kinds of data with different units and magnitudes. Like if the dataset has age and height as two different fields, then unit and magnitude of age and heights are different. In such cases we need to perform scaling. It is extremely important to rescale these features so that all of them are on a comparable scale. If we don't have comparable scales, then some of the coefficients obtained by fitting the regression model might be very large and some might be very small. Here are some commonly used scaling techniques.

Normalised scaling: Used to scale down the values of the features between 0 and 1. It is also called as MinMaxScaler.

Standardised scaling: Here all features will be transformed in such a way that the mean is 0 and standard deviation is 1.

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

VIF is Variation Inflation Factor. It tells how other variables are explaining the variance of a variable. If VIF of a variable is high it means that other variables can explain the variance of this variable, hence this variable is not needed in the model, we can drop it.

The formula for VIF = 1/(1-R^2)

If R value is low, then VIF will decrease. If R value is high, then VIF will increase. VIF becoming infinite means R^2 value is 1. This means R value is 1. This means there is a perfect correlation between two independent variables and one of them needs to be dropped from the dataset which is causing multicollinearity.

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

Q-Q plot or Quantile-Quantile plot is used to access if a dataset follows a normal distribution.

In context of linear regression, Q-Q plot is used to check the assumption of normality of residuals. Residuals are the differences between observed values and values predicted by the linear regression model. Normality of residuals is an important assumption in linear regression as violating this assumption can affect the accuracy of the inferences and predictions made by the model.

By creating Q-Q plot we can visually assess the normality of the residuals. If the points on the Q-Q plot deviate from the straight line, it suggests the residuals do not follow normal distribution. They can also help identify potential outliers in the residuals as outliers can have significant impact on the regression model.