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: Below are my inferences from my analysis of categorical variables from the dataset.

- People take more bikes in Fall, followed by summer, spring and lastly winter
- People tend to have shared bikes in clear weather when compared to any other weather
- People take bikes differently in every month but it can be deduced that it also follows same relation as season
- People tend to take more bikes on working days than holidays or weekend
- Amongst weekdays, People tend to take bikes on 0,1 and 6th day of the week
- There are more numbers of shared bikes in 2019 than 2018
- The spread of shared bikers are more in the the middle months when compared to starting or ending months of the year.

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

Ans: It is important to use drop_first = True during dummy variable creation because it is advisable to reduce the number of levels by 1 to get right analysis and observations. The extra level of variable is redundant in the analysis.

Example: Furnished status – Furnished, unfurnished, semi furnished

Furnished	Unfurnished	Semi furnished
0	0	1
0	1	0
1	0	0

Furnished and Unfurnished are really important so they should be considered.

If Furnished is 1, others have to be 0

If unfurnished is 1, other have to be 0

But if both furnished and unfurnished are 0, then semi furnished has to be 1

In here the semi furnished becomes redundant so dummy variable in respect to it can be removed

Same can be deduced in case of the assignment in case of season and weathersit columns

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

Ans: The temp variable has the highest correlated variable with the target variable as it has positive regression line and highest correlation coefficient of .63. Also, it is giving the most change in target variable as per the model

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

Ans: a. Check for VIF of the model which is <5 to avoid high multicollinearity.

b. Residual distribution of the model should be normal.

c. Check for DW value. If DW =2, means there is no auto collinearity.

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: 3 features:

1. yr.
2. temp
3. weathersit_3(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)

yr. and temp have positive correlation with co-efficient as .2319 and .4933 respectively

weathersit_3 has negative correlation with co-efficient as -.2479

But they are having a significant change in dependent variable within change of 1 unit of the dependent variable.

General Subjective Questions

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

Ans: Linear Regression is the method of relationship of target variable and independent variables.

1. Loading the data set
2. Checking for categorical and numerical values
3. Creating dummy variables and removing extra columns for best data preparation
4. Checking for the description of the model
5. Plot the correlation of the dependent variable and independent variables
6. Rescale the variables
7. Split the dataset in train and test datasets with 70-30, 80-20 rule respectively
8. Build the model and train the model by checking the p value and VIF
9. Check for residuals and predict the model
10. Then test the model
11. Check the R squared after training model and testing model
12. Compare the predicted model data with test data and check for assumption
13. Also check with RFE is required and check on error terms distribution and plot of the predicted and observed value to get an idea about the regression line

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

Ans: It is a data set group (x, y) having same mean, std. deviation, and regression line, but which are qualitatively different. It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.

It shows multiple data sets with many similar statistics different across different quarters.

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

Ans: It's linear correlation coefficient which is standardized slope of the regression line. It is to understand difference between predicted and observed values. It checks for strength between 2 variables. Value goes from -1 to 1

a. -1 value means negative correlation
b. 0 value means no correlation
c. 1 value means positive correlation

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

Ans. Scaling is the preprocessing step to make the model to learn and understand problem. Scaling is performed to bring all independent variables in fixed range of values of similar magnitude.

Normalization	Standardization
Scales to min and max values	Scales to mean and standard deviation
Values fall within [0,1] and [-1,1]	No range
It is helpful when distribution is unclear	It is helpful when distribution is constant

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

Ans: Infinite VIF happens when there is perfect correlation in between two independent variables. In the case of perfect correlation, we have R squared = 1, leading (1/1-Rsquared) to infinity. This is there because one or more variables are creating multicollinearity in the model.

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

Ans: It is a scatter plot to check whether the residuals in a model are following a normal distribution. It is plotting of quantiles of 2 distributions with respect to one-another. It confirms that training and test sets are from same distributions. It fits the model across the regression line.