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. 1) Summer and fall season are actually showing that a greater number of sales happen.
2) If it is a working day then the sales tend to go up.
3) Months from May to September tend to show a rise in the sales.

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

Ans. When you perform dummy variable creation basically it will create n number of column if you don't do drop_first=True. But the concept of dummy variable creation is to create n-1 number of columns. Where n is the number of levels in the column.

Example with drop_first=False:

If there is a column season which has 3 values Summer, Winter and Rainy. So dummy variables will create 3 columns named season_summer,season_winter and season_rainy which will contain binary values. If season_summer is 1 and both the other are 0 then it means it is summer season.

Example with drop_first=True:

Now similarly the values will still be binary but now there will only be 2 columns season_winter and season_rainy. So basically, if season_winter and seaon_rainy have values as 0 then obviously it means that the season is summer.

This is why it is important to do drop_first=True as it will reduce the variables after creation of dummy variables.

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

Ans. Registered and casual are the most obvious variables who are correlated with the target variable but other than that temp and atemp are very correlated 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. The training and testing accuracy are quite close to each other so the model does not overfit. We have removed the multi collinear variables and the error terms follow a normal distribution which are centered around 0. We have evaluated this by visualizing the error rate.

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. According to us the temperature, humidity and windspeed are the most significant variables contributing to the demand of the shared bikes. Which also kind of makes sense you wouldn't want to use a shared bike in a bad weather that is just not convenient.

General Subjective Questions

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

Ans. Linear regression model is a supervised machine learning algorithm which when the target variables is of continuous in nature. The main objective of a linear regression model is to fit a straight line which best fits the data. It assumes that the relationship between the independent and dependant variables is linear, the residuals follow a normal distribution and are centered around 0 and the independent variables are not highly correlated with each other.
Mathematically linear regression looks like this: $y=\beta 0+\beta 1x1+\beta 2x2+\ldots+\beta nxn+\epsilon$

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

Ans. Anscombe's quartet was created by Francis Anscombe which basically was a group of 4 datasets which had same mean and variance and also the same linear regression line y = 3+0.5x.The data point were very different and also different in nature as in there were outlier in some hence the main goal of Anscombe's quartet was that solely relying on the summary stats can be misleading so you should always try to visualise the data in order to understand the patterns, relationships and outlier in the data.

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

Ans. Pearson's relation is a statistical method developed by Karl Pearson that tells the direction and strength of linear relationship between two continuous variables. Pearson's R ranges from -1 to +1. +1 indicates a high linear relation while -1 indicates a negative linear relation and 0 indicates no relation at all.

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

Ans. Scaling basically means bringing a range of values to a certain range of values which are standard across all variables. Meaning when the independent values are way too big with comparison to the other independent variables the coefficients of those variables will be very high which might affect the model and its explainability. This is the reason why scaling is performed. To bring all the independent variables and even dependant variable onto one scale of values so that we can ensure uniformity across coefficients in terms of the weightage they carry.

Normalized Scaling: - We use this by using MinMaxScaler in python which basically converts the values into a scale of 0 to 1. This also handles outliers very well so it is advised to generally use normalized scaling.

Standardized Scaling: - We use this by using StandardScaler in python which basically converts the values into a scale of -1 to 1. It does not perform well with outliers.

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

Ans. This situation happens when the particular variable value of VIF is infinite meaning that the variable is very highly correlated with the target variable or with other independent variables.

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

Ans. Q-Q plot basically means quantile-quantile plot which is a graphical tool used in statistics to verify if a given set of data points follow a certain distribution like normal distribution as example. In this basically first the data is sorted in ascending order then for each data point the quantile is calculated based on the distribution that you are trying to check and then these data points are plotted against it. In linear regression Q-Q plot is use to assessing the assumption of normality of error terms.