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)

1) Season

The bike rentals seems to be higher during the fall season. Spring seems to be the lowest in terms of bike rentals

2) Year

The bike rentals seem to be higher in 2019 as compared to 2018. However, this is too preliminary to draw any conclusions

3) Month

The bike rentals seem to be higher from month 5 to month 10 (may to october)

4) Weathersit

The bike rentals are higher when the weather is clear or partly cloudy; with the median higher compared to other weather situations

5) Weekday

Weekday or Weekend doesnt seem to have any impact on the rentals. However, we can check later in the model

6) Holiday

Holiday or Working day also doesnt seem to have any impact on the rental counts. However, we can check later in the model

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

If we do not use drop_first = True, then n dummy variables will be created, and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap. One dummy variable can be easily explained by the other dummy variables(n-1). If we don't use the drop_first flag, it can lead to extra columns, additional data and computation power.

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

The variables temp and atemp have the highest correlation with the target variable i.e. bike rental count.

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

- Residual analysis of the training data: Plotted a histogram of the error terms to check if they are normally distributed with a mean of zero
- Look for patterns in the error residuals

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

The top 3 features are:

- Temperature with 0.52 coefficient
- Year with 0.23 coefficient
- Light rain_Light snow_Thunderstorm (Negative correlation; negative coefficient with -0.28)

General Subjective Questions

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

Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.

The goal of the linear regression algorithm is to get the best values for the coefficients for the independent variables to find the best fit line. The best fit line should have the least error means the error between predicted values and actual values should be minimized. The cost function helps to figure out the best possible values for the coefficients, which provides the best fit line for the data points. Cost function optimizes the regression coefficients or weights and measures how a linear regression model is performing. The cost function is used to find the accuracy of the mapping function that maps the input variable to the output variable. This mapping function is also known as the Hypothesis function.

In Linear Regression, Mean Squared Error (MSE) cost function is used, which is the average of squared error that occurred between the predicted values and actual values.

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

Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. When plotted visually, the data sets have very different distributions so they look completely different from one another when you visualize the data on scatter plots.

Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when we plot these data sets, they look very different from one another.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

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

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation between 2 variables. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. Another way to think of the Pearson correlation coefficient (r) is as a measure of how close the observations are to a line of best fit. It also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, r is negative. When the slope is positive, r is positive.

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

Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. Feature scaling is the process of transforming the features in a dataset so that their values share a similar scale. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. It wont be appropriate to compare the coefficients. Therefore, in order for us to interpret these features on the same scale, we need to

perform feature scaling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

- Normalized scaling or min-max scaling will try to bring all the data in the range of 0 to 1
- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean zero and standard deviation one .

Normalization	**Standardization**
This method scales the model using minimum and maximum values.	This method scales the model using the mean and standard deviation.
Values fall between 0 and 1 or -1/1	Values are not constrained to a particular range
It gets affected by outliers	Doesn't get affected by outliers
Useful when we don't know about the underlying distribution	Useful when the underlying distribution is normal or gaussian distribution

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

If one of the variables have a perfect correlation with other independent variables, then VIF can be infinite. It is better to drop this variable as it can be explained by the other variables. When multicollinearity is perfect (i.e., the variable is equal to a linear combination of other variables), the VIF tends to infinity.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. QQ plots is very useful to determine

- If two populations are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
- Skewness of distribution

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

In summary, A Q-Q plot helps you compare the sample distribution of the variable at hand against any other possible distributions graphically.