Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Answer:
1. We can see a low trend in choosing bicycles during the spring, which also shows that the season is a good indicator of dependent variables.
2. We observe a tendency in the months, which may be related to the season and the weather. It means it is a reliable prediction.
3. It is evident that individuals prefer cars far less when it is snowing, and the weather is a good predictor of dependent variables.
4. The number of user increase as the year increase. This is a good sign of progress.

Question 2. Why is it important to use drop_first=True during dummy variable creation?
Answer: When making dummy variables, the base/reference category is eliminated by using the drop_first = True setting. This is done to prevent multi-collinearity from being included in the model if all dummy variables are used. Where all of the other dummy variables for a specific category are equal to 0, the reference category can be simply determined and hence avoid redundancy of any kind.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Answer: The temp and the target variable are most correlated.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set?
Answer: By examining the VIF, the residual error distribution (constant variance, normal distribution, independent), and the linear relationship between the dependent variable and a feature variable, it was possible to validate the assumptions of linear regression.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Answer: Temperature, Weather and Year and holiday

General Subjective Questions:

Question1. Explain the linear regression algorithm in detail.
Answer: An ML approach used for supervised learning is linear regression. Based on the provided independent variable(s), it aids in forecasting a dependent variable (target). A dependent variable and the other independent variables are typically connected linearly by the regression approach. Simple linear regression and multiple linear regression are the two forms of linear regression. When a single independent variable is used to predict the value of the target variable, simple linear regression is used. When several independent factors are used to

forecast the numerical value of the target variable, this is known as multiple linear regression. Regression lines are linear graphs that depict the connection between dependent and independent variables. A positive linear relationship is when the dependent variable is on the Y-axis along with the independent variable in the X-axis. However, if the dependent variable value decreases with an increase in the independent variable value increase in the X-axis, it is a negative linear relationship.

Simple Linear Regression: This type of regression analysis is used when there is just one independent variable and one dependent variable.
A straight-line equation can be used to represent the relationship: y = mx + b

Multiple Linear Regression: This statistical method is known as multiple linear regression when there are multiple independent variables.
The resulting equation is: y = b0 + b1*x1 + b2*x2 + ... + bn*xn

Limitations and Assumptions:

1. A linear relationship between the independent and dependent variables is what linear regression presupposes.
2. The residuals—differences between observed and predicted values—are assumed to be normally distributed and to have a fixed variance.
3. When there is multicollinearity (high correlation) among the independent variables or when the connections between the variables are not truly linear, linear regression may not work well.

2. Explain the Anscombe's quartet in detail.
Answer: Anscombe's quartet is a group of four tiny datasets with almost identical simple descriptive statistics but with remarkably diverse distributions and visual appearances on graphs. The statistician Francis Anscombe developed this dataset in 1973 to highlight the value of data visualisation before making decisions based only on summary statistics. It is a potent reminder that summary statistics by themselves can be deceptive and may not accurately reflect the nature of the data.

Basic summary statistics (mean, variance, correlation coefficient, etc.) for each dataset would reveal that they are highly similar or virtually identical, however, summary statistics can be misleading. Due to this, a person can assume that the datasets are nearly identical when in fact, they have quite different underlying patterns.

Anscombe's quartet emphasises the importance of data visualisation and why it matters. The variations in these datasets' patterns are seen when you graph them. It emphasises how crucial it is to produce scatterplots, histograms, or other data visualisations in order to comprehend data.

Question 3. What is Pearson’s R?
Answer: The Pearson correlation coefficient, commonly abbreviated as "r" or "Pearson's r," is a statistical indicator that expresses the magnitude and direction of a linear relationship between two continuous variables. It is one of the most used techniques for determining how linearly connected two variables are. Between -1 and 1, Pearson's correlation coefficient has the following interpretations:

Positive Correlation ($r > 0$): A strong positive linear association is shown if r is positive and close to 1. This implies that as one variable rises, the other usually follows suit. The strength of the positive connection increases as r approaches 1.

No Correlation ($r = 0$): If r is close to 0, it indicates that the variables have little to no linear relationship. There is no consistent correlation between changes in one variable and changes in the other. It's crucial to keep in mind, though, that variables may still be connected by other statistical relationships or non-linear means.

A high negative linear correlation ($r < 0$) is indicated by a negative correlation coefficient, which is negative and close to one. In this situation, one variable tends to decrease as the other one rises. The negative correlation is stronger the closer r gets to -1.

Question 4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Answer: Scaling is a pre-processing method used to standardise the independent feature variables in the dataset within a predetermined range.

The dataset may contain a number of features that range widely in high magnitudes and units. There will be some discrepancy in the units of all the characteristics included in the model if scaling is not done on this data, which results in erroneous modelling.

- Normalization/Min-Max scaling - Using this technique, values between 0 and 1 are normalised. Additionally, normalising the outliers is aided by the Min max scaling.

$$MinMaxScaling{:}x = x - \mathrm{min}(x)/\mathrm{max}(x) - \mathrm{min}(x)$$

- Standardisation results in a standard normal distribution with a mean of 0 and a standard deviation of 1 for all the data points.

$$Standardization{:}x = x - mean(x)/sd(x)$$

Question 5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Answer: When the two independent variables are perfectly correlated, VIF has an infinite value. In this instance, the R-squared value is 1. This leads to VIF infinity as VIF equals to 1/(1-R2). According to this idea, multi-collinearity is an issue, and one of these variables must be eliminated in order to create a useful regression model.

Question 6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Answer: To assess whether a dataset in question follows a certain distribution, such as a normal, uniform, or exponential distribution, the quantile-quantile (Q-Q) plot is used to plot quantiles of a sample distribution with a theoretical distribution. It enables us to determine whether the distribution of two datasets is the same. It is also useful to determine whether or not the errors in the dataset are typical.