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: The dataset includes the categorical variables: season, year, month, holiday, weekday, and weather situation. From the analysis, we can infer the following:

1. **Fall Season:** The fall season, particularly September, attracts the highest number of active customers.
2. **Sales Trends:** In 2019, sales were higher than in 2018.
3. **Impact of Holidays:** Holidays lead to a decrease in the active customer count.
4. **Weather Effects:** There are no users during heavy rain, while partly cloudy or clear conditions result in the highest user count.

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

Ans: Not using drop_first=True when creating dummy variables would result in multicollinearity among the dummy variables, leading to redundancy. This is not desirable for our analysis.

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

Ans: *atemp* and *temp* has the highest correlation 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: One of the fundamental assumptions of a linear regression model is that the error terms should correspond to a normal curve, when plotted on histogram, which is the same in our analysis

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

Ans: The top three features that directly influence the count are those with the highest coefficients. These include:

1. **Temperature:** Positively influences the count.
2. **Year:** Also has a positive effect.
3. **Snowy and Rainy Weather:** Negatively impacts the count.

General Subjective Questions

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

Ans: Linear regression is an interpolation technique used to predict the correlation between variables and assess how an independent variable is influenced by the dependent variable(s). After conducting exploratory data analysis and cleaning the data, we split the dataset into a training set (used to train the model) and a testing set (used to evaluate the model's accuracy).

Next, we checked for collinearity among the variables and selected the relevant ones for training. We then evaluated the model by examining the R-value and the p-values of the dependent variables, making necessary adjustments by dropping or addressing specific columns. After iterating through these steps of feature elimination, we arrived at a final model.

In line with linear regression assumptions, which state that the error distribution must be normal, we proceeded to test the model using the test dataset. The conclusions drawn from this model will provide valuable insights and predictions for data points within its range.

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

Ans: A regression model is not always necessarily an exact one, it can also be fooled by some (smart) data! In certain cases, there are multiple datasets which are completely different but after training, the regression model looks the same. A group of four such datasets having identical descriptive statistics but with some peculiarities, is the Anscombe's quartet.

3. What is Pearson’s R? (3 marks)

Ans: Pearson's correlation coefficient, or Pearson's R, measures the strength of the correlation between two variables and is commonly used in linear regression. Its value ranges from -1 to +1, where +1 indicates a perfectly positive linear correlation and -1 indicates a perfectly negative linear correlation. Values between these extremes reflect the degree of collinearity between the two variables.

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

Ans: Scaling is essential for a model to function properly with the appropriate range of coefficients. For instance, consider two independent variables, such as price and months, that affect car sales. The price range could be significantly higher than the range for months, which only has 12 values. In this case, scaling the price variable helps prevent decimal errors in the model.

There are two types of scaling:

1. Normalized Scaling: This technique adjusts the data distribution to approximate a Gaussian distribution. It does not have a preset range and is commonly used in neural networks.

2. Standardized Scaling: This involves compressing variable values into a specific range to suit the model, as illustrated in the previous example.

Regarding the value of the Variance Inflation Factor (VIF), it can sometimes be infinite. This occurs when there is perfect multicollinearity among the independent variables, meaning one variable can be expressed as an exact linear combination of others, leading to redundancy in the model.

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

Ans: If there is a perfect correlation between the dependent variable and independent variable(s), the R-squared value comes out to be 1. Hence VIF, which is (1/(1-R^2)) turns out to approach infinity.

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

Ans: A Q-Q plot is a graphical tool used to assess whether two sets of data come from the same statistical distribution. It is especially useful in linear regression when working with separate training and testing datasets. In this context, it's crucial to determine if both datasets originate from the same distribution to ensure the integrity of the model.