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?**

Ans: Based on my observation on categorical variables I can say that

- positive correlation between days of the week represented with the ‘weekday’ column which we later transformed into dummy variables to feed into our model.
- there is a positive correlation between column ‘yr’ (0.59) and ‘cnt’.
- ‘weathersit’ has a correlation with ‘cnt’. We can see that in winters the demand for bikes is very low.
- In our final model we have seen there is a correlation between ‘workingday’ and ‘cnt’

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

Ans: Because we can represent the same amount of information with K-1 variables for K types of categories in any categorical variable. Hence, we drop first column to reduce over information in system and reduce complexity

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

Ans: ‘temp’ has the highest correlation with target variable ‘cnt’ after looking at the pair-plot

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Ans: We were able to validate our assumption based on considering three things. First, we checked the R-Square score of the model which was above 80% which is a decent value for prediction. Then we checked the VIF values of all the variable and fine-tuned the model till we got the values below 5. Finally, we performed the residual analysis to check the error terms.

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

Ans: ‘temp’, ‘light_snow_rain’ which is a dummy of ‘weathersit’ and finally we have ‘yr’ as the most significant variables contributing towards explaining the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail?

Ans: Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

The aim is to find the best-fitting line that describes the relationship between two variables. The line is determined by minimizing the sum of the squared differences between the predicted values and the actual values.

Linear regression has some assumptions such as 1. Linearity of residuals, 2. Independence of residuals, 3. Normal distribution of residuals, 4. The equal variance of residuals.

2. Explain the Anscombe's quartet in detail?

Ans: Anscombe's Quartet is nothing but a group of four data sets which are nearly identical in simple statistics, but there are some strange or unusual features in the dataset that fools the regression model if it is built using them. They have very different distributions and appear very differently when plotted on scatter plots.

It shows the importance of data visualization and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model.

3. What is Pearson's R?

Ans: Pearson correlation coefficient, also known as Pearson R statistical test, measures the strength between the different variables and their relationships. Therefore, whenever any statistical test is conducted between the two variables, it is always a good idea for the person analysing to calculate the value of the correlation coefficient to know how strong the relationship between the two variables is.

It can range from the value +1 to the value -1, where +1 indicates the perfect positive relationship between the variables considered, -1 indicates the perfect negative relationship between the variables considered, and 0 value indicates that no relationship exists between the variables considered.

The Pearson correlation coefficient shows the relationship between the two variables calculated on the same interval or ratio scale. In addition, It estimates the relationship strength between the two continuous variables.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Ans: Scaling is the technique of normalizing all the numerical variables to common level giving you a sense of comparison. It helps us to compare the coefficient of different ranges on normalized levels.

Scaling is performed since all the variables when feeded to the model show large coefficient values if the value of the variable is large, creating the illusion that some variables affect more than others. It is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale.

Normalized scaling - In this approach, the data is scaled to a fixed range - usually 0 to 1. Formula used to perform this scaling is Xsc= (X−Xmin) / (Xmax−Xmin). For this we use Min-Max scaling.

Standardized scaling - Standardization is another scaling method where the values are centered around the mean with a unit standard deviation.In this the attribute becomes zero, and the resultant distribution has a unit standard deviation. Formula for the same is, Standardized value = X – μ / σ

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

Ans : This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which leads to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. We had observed this in our data set when we considered the registered and casual variable in the data set which we later dropped from the entire analysis.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?**

Ans : Q-Q plots are also known as Quantile-Quantile plots. It plots the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential

In linear regression it helps us to check the distribution of the error terms or prediction error. If there is a significant deviation from the mean, we might want to check the distribution of our feature variable and consider transforming them into a normal shape. It also help us to check if two populations are of the same distribution