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)

As per my analysis, the categorical variables I see in the dataset as Season, yr, month, holiday, weekday, workingday, weathersit.

All the categorical variables has significant effect on the model and there by on the dependent variable cnt due to the quite high correlation with the cnt variable which means that different conditions of independent variables affect the bike rentals in the problem terms.

The season does affect the bike rentals numbers as from the boxplot of data visualization it is evident that fall which is pleasant weather the bike rentals are highest and like wise the rentals are lowest in spring.

The weather situation also affects the bike rentals as more the weather is clear more the rental numbers of bikes.

There is also a rise in popularity of the business seen from 2018 to 2019 as the number increased from 2018 in 2019 so that even has an effect more popularity as inference.

Holidays can be inferred straight forward more rentals and that is a common sense even.

Mnth has high correlation with weather and is evident from the box plot and coefficients as well that fall time of June, July, August September has high bike rentals numbers.

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

It is important to use the drop_first=True during the dummy variable's creation as if we include all the variables in the dummy variable creation the multicollinearity is perfectly fit which is harmful for our predictions because say if we have 3 variables and we include all the 3 variables then knowing 2 will default make 3rd variable predicable default as only 3 values are possible say for those categorical variables. This is the reason why we use drop first as True and drop 1 variable so that there is no perfect multicollinearity and remove redundancy thereby.

Prediction of model becomes more reliable and genuinely true as well I think.

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

The highest correlation or linear relationship can be seen on holiday and workingday with cnt dependent variable.

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

I used the residual analysis technique to verify the assumptions that the model is fit to do with Linear Regression with a normal distribution of the plotted values for the training set against the predicted values.

Also, while calculating the VIF and eliminating the variables I made sure that the multicollinearity is addressed and the quality of my predictions is meaningful and accurate enough on test set as learnt from training set.

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 factors season, weather situation and temperature. I infer this based on the coefficient values, p values and VIF values of the 3 variables.

General Subjective Questions

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

The goal for any linear regression algorithm as per me is finding out the linear relationship between the dependent and independent variable.

Linear Regression Algorithm as per the learnt concepts is divided into steps for building itself.

Firstly, the data preparation is key in starting to build the algorithm.

Secondly, Visualizing the data after data preparation will give a fair idea of the data for further building the algorithm and checking if the data is fit for the purpose or not.

Thirdly, scaling the variables and removing the redundancy is important.

Fourthly, training the model and building the model itself.

Fifth, is to do the residual analysis on the predicted values to evaluate the validations on the model built.

Sixth step is to do the model evaluation with r square value study and finally doing some assumptions check on the data set selected to train the model if there is anything wrong in the r square value of predicted versus trained data set values.

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

As my research the statistician Francis Anscombe designed four datasets to illustrate the importance of visualising the data in graph format rather than solely relying on the statistical summary of variables.

Graphical data visualization can show us if there are any outliers or anomalies which in statistical summary may not be visible straight forward.

Mean, variance and correlation do not always say the true story of any datasets is what Anscombe meant with his analysis and inferences.

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

Pearson R ranges between -1 and 1 and indicates the linear relationship strength of the variables to fit into a straight line.

-1 is perfect negative correlation while 1 is perfect positive correlation and 0 means no linear relationship.

Pearson R may mislead when there are outliers.

The formula to calculate Pearson R is Covariance between variables divide by product of standard deviations of those variables.

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

Scaling is standardization of values when compared to analysed amongst different variables. The coefficient values can vary drastically if scaling is not done.

Say, for example we have a variable of housing as area which has values very high in scale of say from 500 to 10000 but on the other hand there is variable of price whose scale is even higher than area and comparing these may give totally false inferences. In such case both of

these variables need to be scaled on a similar scale say between 0 and 1 with min max scaling method to infer the statistical measures rightly.

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

In my assignment I had found such an issue with temp and atemp variables when the multicollinearity between them was perfectly aligning and hence after removal of atemp variable I was able to get the VIF values with logical values.
I feel both the variables were able to fully explain the values of cnt variable as prediction hence this may have happened.

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

Q-Q plot I did not study or know more on the concept but as per the research on the internet it is more to explain the normal distribution of data.
A Q-Q plot is used to diagnose the fit of the regression model. Non-normality in residuals can suggest model inadequacies or the need for transformation of the response variable.
Deviations from the Q-Q line can also highlight potential outliers or influential data points that might affect the model.
Validating the assumptions of normality and identifying any deviations is crucial for ensuring the reliability and validity of the regression analysis results. Non-normal residuals can impact hypothesis tests and confidence intervals derived from the model.