Assignment-based Subjective

Q-1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer: Here are some of the inferences I made from my analysis of categorical variables from the dataset on the dependent variable (Count)

1. Fall has the highest median, which is expected as weather conditions are most optimal to ride bike followed by summer.

2. Median bike rents are increasing year on as year 2019 has a higher median then 2018, it might be due the fact that bike rentals are getting popular and people are. becoming more aware about environment.

3. Overall spread in the month plot is reflection of season plot as fall months have higher median, 4. People rent more on non-holidays compared to holidays, so reason might be they

prefer to spend time with family and use personal vehicle instead of bike rentals.

5. Overall median across all days is same but spread for Saturday and Wednesday is bigger may be evident that those who have plans for Saturday might not rent bikes as it a non-working day.

6. Working and non-working days have almost the same median although spread is bigger for non-working days as people might have plans and do not want to rent bikes because of that

7. Clear weather is most optimal for bike renting, as temperate is optimal, humidity is

less, and temperature is less.

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

Answer:
1. To avoid multicollinearity (if we don't drop, dummy variables will be correlated) and affects the model adversely

2. To avoid redundant features

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

Answer: Count (target variable) has significantly high correlation with temperature (TEMP)

Q-4) How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: Residual errors follow normal distribution maintain linear relation between dependant variable (test and predicted)

Q-5) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Answer:

1. Temperature (0.4354)
2. Weather Situation –LIGHT AND SNOWY (0.2837)
3. Year (0.2461)

General Subjective Questions

Q-1) Explain the linear regression algorithm in detail.

Answer: The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line. The straight
line is plotted on the scatter plot of these two points.
The standard equation of the regression line is given by the following expression: $Y = \beta_0 + \beta_1 X$

You also learnt an alternative way of checking the accuracy of your model, which is R2 statistics. R2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. In general term, it provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes. Overall, the higher the R-squared, the better the model fits your data.

Q-2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

Q-3) What is Pearson's R?
Answer: Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

Q-4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. 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. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

Answer: If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead 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.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Answer: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.
If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.