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?

Answer: We can see that rate of bike rentals are more in 2019 compared to 2018. Mostly on holidays bike rental counts are increased. Also when wheather the clear then count is good. In humidity and increase of windspeed, the rental counts is very less.

2. Why is it important to use drop_first=True during dummy variable creation?
Answer: It is important to use drop_first=True during dummy variable creation because it helps to reduce extra column. It also helps to reduce correlations between dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Answer: After watching pair-plot among numerical variables we can say that target variables have strong correlation with temp.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
Answer: I have done feature selection in independent variables to check multicollinearity by getting VIF value. I have checked linear relationship between independent and target variables.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Answer: After building final model , we can say that temp, workingday and seasons are the top 3 features contributing significantly.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear regression is nothing but supervised machine learning algorithm which is used for predictions. There are basically two types in linear regression-

a. Simple Linear Regression- Here we do predictions based on one independent variable
b. Multiple Linear Regression- Here we do predictions based on multiple independent variables

The linear equation we follow is y=mx+c, where x is independent variable, y is target variable and m is slope.
Basically we create a best fit line using residual analysis between actual y value and predicted y value.

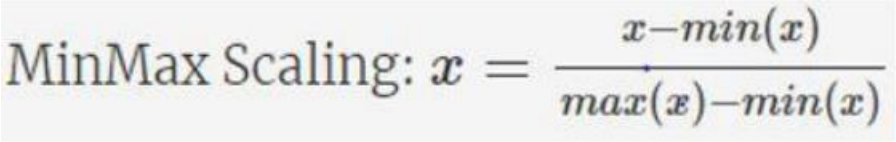
2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet is group of four datasets which are almost identical in descriptive statistics. These are having different distributions and appear differently when we plot scatter plot.
These all four datasets having almost same statistical observations, which provides same statistical information that involves variance and all x and y points in all four datasets.

3. What is Pearson's R?

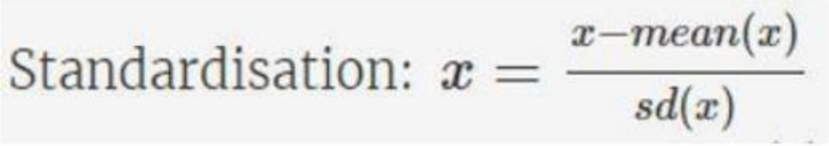
Answer: Pearson's R is used to measure the correlation coefficient. It gives measures between -1 to +1 that measure the strength and direction of the relationship between two variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Answer: We do scaling of variables in data pre-processing. It is useful to normalize the data within a range which will be helpful for further analysis in model building process.
Normalization/Min-Max Scaling: It brings all data in the range of 0 and 1.



$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling: 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).



$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Answer: We get VIF as infinity when there is perfect correlation. It shows a very strong relationship between two independent variables. Since the equation is 1/(1-R^2) , so when we have R^2 as 1 then is leads to get value as infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Answer: Q-Q plot is nothing but Quantile- Quantile plots. These are used to plot the quantiles of a sample distribution against quantiles of theoretical distributions. It helps us to determine whether a dataset is having probability distributions like normal, exponential or uniform