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?

- The demand of bikes is high in the months from may-oct probably due to seasons summer and fall
- Higher demand in the year 2019
- Higher demand during weather sit `Clear, Few clouds, Partly cloudy, Partly cloudy`

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

- Reduces the need for an extra column created during summy variable creations and hence reduces correlation among other dummy variables. Example seasons variable will have 3 columns instead of 4.

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

- Since I dropped temp variable, out of the three windspeed, atemp and humidity, ‘atemp’ has the highest correlation with count.

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

- To validate the assumptions of linear regression after building the model is by creating a scatter plot between the features and target variable.

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

- Temperature (0.552)
- weathersit : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (-0.264)
- year (0.256)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- Linear regression is used to predict or estimate a continuous target variable.
- It is a supervised learning model
- There are two types of linear regression models. Simple linear with one independent variable and multiple linear with more than one predictor variable
- Regression line is given by: $Y= \beta 0 + \beta 1*X$ where Y and X are dependent and predictor variables respectively.

- The aim of linear regression is to find a best-fit line by minimizing the residuals and finding optimal values for parameters beta.
- Residuals or cost function are optimized using different methods like differentiation and gradient descent

2. Explain the Anscombe's quartet in detail.
   - Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points.
   - It was introduced to explain the importance of graphs for analyzing and the effect of outliers and other influential observations on statistical properties
   - The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets

3. What is Pearson's R?
   - The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset.The Pearson correlation coefficient (r) is a way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   - Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing.
   - Eg: variable humidity has values between 0-97 and windspeed is between 1-34 then scaling is used for comparison of the two variables on the same scale by using for instance using the method normalisation by making the values of the two variables between 0 and 1.
   - Scaling helps with better interpretation data and doesn't after the precision of the model and the p-value of the feature. It helps with faster convergence in gradient descent
   - Normalisations makes the min value of the variable 0 and max value as 1 whereas standardisation makes mean of the variable 0 and standard deviation as 1.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   - This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared =1, which leads to 1/(1-R-squared) to infinity.
   - To solve the above issue, column causing multicollinearity between the variables should be dropped.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   - Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or

Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

- This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.