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)

- From the box plot graphs we can infer that there is a higher amount of bike rentals during summer and fall. There is a higher amount of bike rentals during 2019. There is a higher amount of bike rentals from May to September. Holiday does not seem to make a really big difference on bike rentals on whether it is a holiday or not. There is not a high variation on bike rentals that can explained by weekdays. Additionally, there is not really big difference on if it's a working day or not. There does seem to make a difference if the weather type is clear and few clouds or there is Light Snow, Light Rain + Thunderstorm + Scattered.

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

- It's to avoid multicollinearity. For categorical variables you need to convert them into a binary format so that it can be examined per category instead of assigning values to the category that could skew the data.

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

with the target variable? (1 mark)

- Temp and atemp both have a correlation value of .63 with respect to 'cnt'

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

training set? (3 marks)

- I checked my adjusted r2 value at the end of the model to get a value of 0.77 which shows that the model is an okay fit based on the assumptions built into the model.

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes? (2 marks)

- Temperature (temp) - A coefficient value of .5687 indicated that a unit increase in temp variable increases the bike hire numbers by 0.5687 units.
- Weather Situation 3 (weathersit_3) - A coefficient value of -0.2541 indicated that a unit increase in Weathersit3 variable (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) compared to Weathersit1 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) decreases the bike hire numbers by .2541 units.
- Year (yr) - A coefficient value of .2337 indicated that a unit increase in yr variable increases the bike hire numbers by 0.2337 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
    a. In simple linear regression there is one independent variable and one dependent variable. The relationship can be shown through y = mx+ b where y is the dependent variable, m is the slope of the line calculated by y2-y1/x2-x1, x is the independent variable, and b is the intercept.
    b. Multiple linear regression is when there are multiple independent variables and one dependent variable. The relationship can be shown as y = b + m1 * x1 + m2 * x2 + ... + mn * xn.
    c. The whole point of linear regression is to minimize the sum of squared differences (r2) on your formulated equation. You will take a dataset, split it into a train and test set (70/30 or 80/20), and then train the model with the 70% data to get an equation. Then you test that equation on the test set to see how it best fits the data. You can see if it fits the test data set by checking the adjusted r2 value.
    d. The steps are to clean the data, see if there is even any linear relationship between variables, train the model, remove any outliers and standardize the data, evaluate the model, and then make predictions. There is more to the data cleaning and modeling because you need to convert any categorical variables to dummy variables and can only remove one variable at a time that has high collinearity.
2. Explain the Anscombe's quartet in detail. (3 marks)
    a. It's when there are 4 small datasets that have almost identical statistics but look very different from each other when graphed. Anscombe's quartet is important because it shows how important data visualization is and you can't just base conclusions off of summary statistics.
3. What is Pearson's R? (3 marks)
    a. It is a coefficient range that goes from -1 to 1. A -1 value shows that there is a perfectly linear negative relationship between the dependent and independent variable. So that as one variable decreases the other increases. A 1 value shows that there is a perfectly positive linear relationship between the dependent and independent variable. So as one variable increases so does the other one. This coefficient value only works when there is a normal distribution in the data and no extreme outliers.

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

and standardized scaling (3 marks)

1. Scaling is the process of standardizing or normalizing the independent variables before fitting the regression model.
2. Scaling is performed to ensure that all variables have the same scale or unit of measurement. When you have predictors with different scales or units, the coefficients associated with those predictors can become difficult to interpret. Scaling puts all variables on a similar scale, allowing you to compare the impact of each variable more easily. Additionally, scaling makes the coefficients of the regression model more interpretable.
3. Normalized scaling or min max scaling is when you transform the datapoints given by a formula of x-xmin/xmax-xmin. This is helpful when you want to keep the original range of the variables.

Standardized scaling or z score scaling is when you transform the variables to have a mean and standard deviation. The formula for that is given by X-mean/standard deviation

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

(3 marks)

- It's when there is a severe case of multicollinearity between independent variables. This can happen because when two or more independent variables are highly correlated, meaning that they move together very closely, their relationships can be expressed almost perfectly as linear combinations of each other. So it's like when 2 variables show the same line almost.

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

(3 marks)

- A q-q plot, quantile-quantile plot, is when a dataset will have a normal distribution. It's used to compare the quantiles of the observed data to the quantiles of the specified distribution. By using a Q-Q plot, you can visually see the normal distribution assumption and see potential problems in your regression model. It is important to visualize data and the Q-Q plot can help do that.