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?**
    a. Categorical variables had a very strong impact towards our dependent variable. In fact they were included in the top 5 predictor features. The sub features within the larger ‘weather condition’ and ‘season’ were important features our model learned to use as predictors.
        i. light_snow
            1. If there is light snow, it’s likely that we'll see a decrease in the demand for bike rentals.
        ii. yr
            1. The year in which an individual decides to register for a bike share has an incredible impact as a predictor of demand. It is our largest predictor with a positive correlation to demand count. As this ride sharing company continues to grow, it makes sense that demand would increase as well as popularity and word of mouth increases.
        iii. spring
            1. Interestingly enough, during spring there is a slight negative correlation with increased bike rental demand. This could be due to the fact that many users are on vacation and not using it to ride to work.
        iv. winter
            1. Last but not least, there is slightly higher demand for bike share during the winter, but lower demand if it's snowing! During the winter, it’s possible that there is a slight increase in bike share demand due to the large majority of cars on the street.
2. **Why is it important to use drop_first=True during dummy variable creation?**
    a. It is important to use drop_first=True because it helps reduce the creation of an extra column during dummy variable creation. It reduces the correlations created among dummy variables which is a step we care about when optimizing our model.
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
    a. The feature ‘atemp‘ has the highest correlation with the target variable ‘cnt‘.
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
    a. We validated our assumptions in several steps. We first looked at the r squared score of a baseline model before optimizing. After attaining a baseline we leveraged RFE modeling to recursively eliminate features that are not needed. After identifying certain features, I looked at VIF scores for each feature of the model to identify if there was any multicollinearity among features. We were able to remove a few additional features in this way. I looked to ensure that the P values of each feature were low, the T values were greater than 2 and less than

-2 and the F-statistic score was high. Lastly I evaluated a new model based on a specific set of features on its r squared and adjusted r squared score.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
    a. 'temp'
    b. Weather situation of 'light_snow'
    c. Season of 'spring'

General Subjective Questions

1. Explain the linear regression algorithm in detail.
    a. Linear regression is a supervised machine learning algorithm that performs regression tasks. A regression task is one that predicts values that are continuous in nature. In particular a linear regression task is one that finds a linear relationship, or best fit line, between an input and an output. If the equation of a line is y = 0_1 + 0_2(X), when training a model, we are trying to find the best values for 0_1 and 0_2 such that when we add those to the equation, and include input X, we can a value of Y that is close to what the true data shows. On each model training iteration, the model tries to minimize a cost function. Specifically, the linear regression model leverages a cost function called root mean squared error which is the square root of the mean of the square of all errors. It can be quite a mouthful but its a rather simple cost function and works wonders. After our model has been trained, and found its best fit line, it can use the learned 0_1 and 0_2 values to predict new values of Y given new values of X.
2. Explain the Anscombe's quartet in detail.
    a. Anscombe's quartet is a group of four data sets which can fool general descriptive statistics analysis leveraging variance and mean, but have very different distributions and appear different when plotted on scatter plots. It was developed by a statistician Francis Anscombe to illustrate the importance of plotting graphics before analyzing or model building. An important consideration for our linear regression work here is that a linear regression model can only find linear relationships. It can be the case that certain data has the right mean, and variance, but when plotting it on a scatter plot, we can see that the relationship is anything but linear.
3. What is Pearson's R?
    a. Pearson's correlation coefficient is a test that measures a statistical relationship between two continuous variables. It uses a method of covariance. What's great is that it gives us information about the magnitude of a correlation and the direction of a relationship. A correlation coefficient value that is negative (closer to -1) indicates a negative correlation and one that is positive (closer to 1) indicates a positive correlation.
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
    a. Scaling is the process of transforming your data in a way that makes it easier for your model to learn and understand the problem it's trying to solve. There are

many ways to scale your data. Two ways in which we've learned over the last module are scaling using a standard scalar and a normalized scalar like minmax scalar. A standard scalar rescales your data in a way that transforms the mean of the data to 0. This will transform your data into negative and positive values. A minmax scalar, a kind of normalized scaling process, re-scales your data between its max and min values. Other ways to scale your data between 0 and 1 values.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
    a. If we have an infinite VIF value, it indicates a perfect correlation between two different variables. In this case, we need to remove one of the variables to reduce multicollinearity. Increased multicollinearity indicates that one of the features is a linear combination of another variable.
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
    a. Quantile-quantile plots are two different quantile plots plotted against one another. We would use a plot like this to identify if two different datasets come from the same distribution. If the two datasets do come from a similar distribution and are linearly related, you'll see them fall similarly along a 45 degree angle.