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. We made the following conclusions:
weathersit_3 - We see rentals go down by `1726` units during weathersit_3 as compared to weatherset_1 (reference), everything else being constant
weathersit_2 - We see rentals go down by `344` units during weathersit_2 as compared to weatherset_1 (reference), everything else being constant
season_spring - We see that the spring season has a negative impact on bike rentals
We see rentals go down by 635 units in the spring season when compared to the fall season (reference), everything else kept constant
season_summer - We see that the summer season has a positive impact on bike rentals
We see rentals go down by 262 units in the spring season when compared to the fall season (reference), everything else kept constant
season_witer - We see that the winter season has a positive impact on bike rentals
We see rentals go down by 696 units in the spring season when compared to the fall season (reference), everything else kept constant
Holiday - We see that rentals go down by 596 units on public holidays, everything else kept constant

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

A. Using drop_first= True lets us make the very first alphabetically sorted level as our reference variable. It is not always necessary to use drop_first= True flag. If we want to make a different level our reference variable, then that level could be dropped manually too. But dropping one of the levels is important for dummy encoding.

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

A. We see that feature registered has the highest correlation coefficient. But since we're not using that feature in our modelling, we can see temp and atemp features have the highest correlation coefficients at 0.63

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

A. We perform a residual analysis and check if the error in prediction follows a normal distribution by plotting the errors. We can further check the homoscedasticity with a scatter plot of residuals. We can further confirm that there exists a linear relationship between dependent and independent variables. We also make sure that multi-collinearity is within acceptable ranges via VIF scores.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
A. 1. weathersit_3 has the highest impact on rental demand - negative impact

- We see rentals go down by `1726` units during weathersit_3 as compared to weatherset_1 (reference), everything else being constant

2. yr(year) also has a significant impact on rental demand - positive impact

- We see that with each added unit to yr(year), rentals go up by 1000 units, everything else kept constant

3. temp(temperature) also has a significant impact on rental demand - positive demand

- We see rentals go up 985 units, with each added temp unit, with everything else kept constant

General Subjective Questions

1. Explain the linear regression algorithm in detail
   A. linear regression is a linear approach for modelling the relationship between a dependent variable and one or more independent variables. It is a type of supervised learning algorithm. It involves fitting a statistically optimal straight line through a dataset. That straight line can then be used to make prediction on new unseen data or even draw inferences about the independent variables from the model. The easiest form of linear regression equation with one dependent and one independent variable can be defined by the formula y = c + b*x, where x is the independent variable, b is the coefficient and c is the constant.

2. Explain the Anscombe’s quartet in detail.
   A. Anscombe’s quartet represents four different datasets that have exactly the same summary statistics. Seeing this may make us believe that one type of model would be good enough to fit all our four datasets. But when we plot these four datasets independently, we see that these four datasets are very different and cannot be modelled using the same algorithm. This highlights the importance of visualising our data before modelling.

3. What is Pearson’s R?
   A. In Statistics, Pearson's Correlation Coefficient is also referred to as Pearson's r. Pearson's Correlation or Pearson's r is an effective way of measuring association between two numeric variables. A high Pearson's Correlation Coefficient would imply there being a positive associations between the two numeric variables and a negative Pearson's Correlation Coefficient would imply there being a negative associations between the variables.

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

A. Scaling is a process that could be done while preparing the dataset. It involves normalizing the independent features to make the data follow a small range. Although not necessary, when performed scaling helps in modelling features comprising of different units as it eliminates the unit factor and makes the data fit in a range. More importantly, it allows algorithms optimization like gradient descent to find the right coeffiecients/weights much faster. Standardization replaces the values by their Z scores, whereas normalization is performed by subtracting the minimum value from the data point and dividing it by the difference of max(x) and min(x), hence forcing the new values to be in a 0-1 range.

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

A. An infinite value suggests a perfect correlation between the variables in question. It indicates that the corresponding variable can be expressed completely/perfectly by a linear combination of other variables. Like in our assignment we saw features holiday, is_weekend and workingday when modelled together had an infinite VIF.

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

A. Q-Q Plots, also referred to as Quantile-Quantile plots are plots of two quantiles against each other. They can help us determine if a dataset follows any particular type of probability distribution like a normal, uniform, etc. They are also important if we want to tell if two datasets are coming from the same distribution. Understanding the probability distribution of our datasets lets us pick the best modelling technique, which is very important in any type of predictive analysis. Q-Q plots can also tell if the train and test data are coming from the same distribution if we've received them independently.