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?**
   By creating the boxplot graphs for the categorical variables I could see the following trends:
   a. Day of the week (weekday) doesn't seem to matter as the medians are very close throughout.
   b. Clear weather see the best usage, followed by misty weather, followed by light-snow/light-rain type of weather. Rainy, snowy weather with thunderstorms is a definite no-go, as expected.
   c. The average booking is higher if it is a working day, compared to a holiday.
   d. Fall season sees the maximum bookings, followed by summer and winter. Spring sees the least bookings.
   e. Month-wise, there is a rise in bookings after February and it slowly rises and stabilizes towards June. After October the usage drops again. Maybe this data is for a place in the northern hemisphere, and therefore the usage is more during warmer months.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   Suppose you create dummy variables for a categorical variable car_type which can contain values 0, for hatchback, 1 for sedan and 3 for SUV. Instead of having 3 dummy variables for each type, we can have 2. For example, create variables isHatchback and is Sedan. If both these values are 0s, then it's obvious that the car is an SUV. This way we reduce the number of dummy variables by 1, therefore reducing the possible multicollinearity.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   The variables temp and atemp have a very high collinearity with cnt, at 0.63 each.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   I can create a histogram of the residuals and see if the curve shows a typically normal distribution or not. Mainly the mean should be zero and the errors should be equally distributed around the mean.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   a. temp has a correlation of 0.49. So the warmer the weather, the more the demand.
   b. weathersit_3 has a correlation of -0.29 meaning snow, light-rain kind of weather is when the demand is the least. Of course, weather_sit 4 is the worst, with no demand at all i.e. rains and thunderstorm.
   c. yr has a correlation of 0.23 which is not very predictive since we only have 2 years in the data, 2018 and 2019. But we can still say with some confidence that the demand is increasing as the years go by.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   The drill of using the Linear Regressing (Lnr) algorithm is as follows:

   a. We analyse the data and look for outliers, nulls and duplicates and clean them up.

   b. We look at the categorical variables and is possible and applicable convert them to dummy variables.

   c. We split the dataset into train and test datasets. The train dataset will be slightly larger than the test one and will be used to recursively fine-tune the model and the test dataset will be used to test our findings from the train dataset.

   d. We will scale the dataframe so all the variables are on a similar scale with the mean as 0 and the error normally distributed around the mean.

   e. We then check the VIFs and P-values of each variable in the current model. We remove variables with VIFs >=10 and perform step d. till there are no variables with VIF >= 10 left.

   f. We then check if there are variables with p-values > 0 and recursively remove the one with the highest value and perform step d. till we have all the remaining variables with low VIFs and 0 as p-values. This will be our final model.

   g. Now we take the final model and apply it to the test dataset using the python predict function and see how the spread looks. Hopefully the scatter-plot shows good collinearity between the test set and the predictions.

   h. Finally we check the adjusted R-squared value. Any value >=70% is deemed to be good enough to say we have created a good model.

2. **Explain the Anscombe's quartet in detail.**

   The Anscombe Quartet is a visual experiment that disprove the classical thought that raw data is more easily understandable than graphs that represent the same data. Using 4 graphs, Anscombe went on to show that when used correctly graphs can give a more intuitive and readily understandable view of the data, as compared to just reading the numerical statistics around the same data.

   For this experiment, Ascombe took 4 datasets that, when you look at the numeric statistics of, look almost identical. Any casual observer would deem no much variance among the datasets. However, when you look at them graphically:

   a. The first dataset shows a linear relationship

   b. The second dataset shows more of a non-linear curved relationship

   c. The third graph shows a somewhat similar linearity as the first dataset, but it has an outlier, which causes the regression line to deviate from the datapoints such that the errors are not equally distributed.

   d. The fourth graph shows the datapoints almost towered upright, showing that one high value can produce a high correlation coefficient.

   Nowadays, it is generally accepted that a picture speaks a thousand words. The Ascombe's Quartet experiment aimed to prove this point.

3. **What is Pearson's R?**
   Pearson's R, also called Pearson's correlation coefficient is a value between -1 and 1 that shows the level of collinearity between 2 variables.

   An R score >=0 and <=1 shows positive collinearity. The higher the number the greater the collinearity. For example, Cooldrink sales vs temperature. Increasing temperature gives rise to increased demand for cooldrinks.

   An R score of <=0 and -1 shows negative collinearity. For example, higher temperature lowers the demand for sweaters.

   An R score of 0 says there is absolutely no correlation between the 2 variables. For example, the temperature and sales of fountain pens.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
   Scaling is a technique used to normalize the range of the data we are working with. This help is to speed up the calculations in any ML algorithm while ensuring the resulting coefficients don't vary between the scaled and non-scaled datasets.
   Differences between normalize and standardized scaling are:

   Normalized scaling:
   a. Max and Min values of the dataset are used for scaling
   b. Is used when the features are of different scales
   c. Severely impacted by outliers

   Standardized scaling:
   a. Mean and Std are used for scaling
   b. Used when we need to make the mean 0 and errors equally distributed on either side of the mean
   c. Less impacted by outliers.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   VIF measures how much the variance of a coefficient increases due to collinearity. If the VIF is >10 it means there is severe collinearity. Around 5 means moderate and 1 means no collinearity.

   An Infinite VIF means that there is perfect collinearity between 2 independent variables and therefore we have to drop one of those variables when we are building a Linear Regression Model.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
   A Quantile-Quantile plot helps us compare a dataset's distribution against an expected normal distribution. We can use Q-Q plot to compare the shapes of 2 datasets in terms of

range, skewness and location on the graph.
The Q-Q plot is mainly advantageous if we need to compare datasets of 2 different sizes. For example if we compare rice consumption is Goa vs Kerala, the averages and std will approximately be the same, even though Kerala has a much higher population that Goa. The Q-Q plot for both these states will therefore look very similar.