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?
   Ans. From bivariate analysis of season variable and dependent variable(cnt) it seems demand for bike-sharing is quite high during fall season followed by summer, winter and spring. We can infer that the winter season yields the highest positive effect on the dependent variable.
   Where as the bivariate analysis of whethersit variable and dependent variable(cnt) defines high correlation between dependent variable a weather which is Clear, Few clouds, Partly cloudy, Partly cloudy followed by a weather which is Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist and Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds.

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

   Ans. In order to be slightly less redundant and keep it easy for the algorithm to fit in the model, it is important to use drop_first = True during dummy variable creation.

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

   Ans. By looking at the pairplot temp variable seems to has the highest correlation with the target variable.

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

   Ans. The validation process began from a simple pairplot to see if independent variables exhibit linear relationship with dependent variable and, yes they did.
   Second, checked if the error terms are also normally distributed (which is in fact, one of the major assumptions of linear regression) and found the residuals normally distributed
   Third, check multicollinearity by calculating VIF and eliminate the variables having p-value greater than 0.05 and VIF greater than 5
   Thus, validated that all assumptions of regression are taken care of and safely build the model.

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

   Ans. Based on final model temp, season and weathersit columns are top 3 features contributing significantly towards explaining the demand of shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail

   Ans. Linear regression model is a supervised machine learning model employed to predict the numerical variable.
   Linear regression model is based on the line equation y= mx +c and the parameters of linear model can be derived using least square method or maximum likelihood estimation
   Linear Regression Model can be further divided into:
   - Simple Linear Regression
   - Multiple Linear Regression

   The model is defined in terms of coefficients named beta not and beta one. For example, a problem having inputs X with n variables x1, x2,...,xn will have coefficients beta1, beta2,.... And beta0.
   - ☐ y = beta0 + beta1 * x1 + beta2 * x2 + … + betan*xn

   First, Linear Regression:

   - Quantifies the relationship in the data (which is R square)

   - this needs to be large

   - Determines how reliable that relationship is which is p-value with F

   - this needs to be small

   It uses a simple linear regression to fit a line to the data given

The model begins by drawing a line to calculate the sum of square residuals (which is the sum of distance of each data point from the line
The model just repeating the process by rotating the line a bit and calculate the sum of square residuals
After a bunch of rotations plot the sum of squared residuals and corresponding rotations
Then find the rotation that has least sum of square
That is why the method is called least sum of square
After this, calculates the sum of square around the mean (SS)

- SS = (data- mean)^2
- Variation = SS / n [this is to average the sum square per data]

Then sum the squared residuals around least squared fit

- R^2 = Var(mean) - Var(fit) / Var(mean)

When R^2 is zero means the predictor doesn't explain any of the variation around the mean

Then it calculates the p-value. So, **when the p-value is low enough, we reject the null hypothesis and conclude the observed effect holds.**

2. Explain the Anscombe's quartet in detail.

   Ans. According to the definition given in wikipedia, **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. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties. Anscombe's quartet states the significance of visualising data before employing various algorithms for building models. It says the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

3. What is Pearson's R?

   Ans. Pearson R is a bivariate coefficient that measures correlation between two sets of data. It is a ratio between the covariance of two variables.

4. What is scaling? Why is scaling performed? What is the difference between normalised scaling and standardised scaling?

   Ans. Scaling is a process performed during preparing the data for a machine learning model.
   It is common to have data where scaling differs from variable to variable hence **it is important to perform scaling** for **effective** and efficient **performance of algorithms**
   The difference between normalize and standardise scaling is the method where Normalisation scales each input variable separately to the range

0-1, and Standardization scales each input variable separately by subtracting the mean and dividing by the standard deviation to shift the distribution to have a mean of zero and a standard deviation of one.

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

   Ans. Infinite value of VIF is an indication of perfect correlation between two independent variables which is called perfect collinearity.

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

   Ans. Q-Q plots is an abbreviation of quantile- quantile plot which stands against each quantile.

   Q-Q plots are useful for:
   Identifying if the two population belongs to same distribution
   Matching the error terms of two population
   Identity the skewness of distribution whether the data belongs to one side or other
   It is important in linear regression:
   to check whether the sample size is equal
   To determine the similarities in the populations