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 : There are total of 6 categorical variables we can find in the Dataset, namely,
   Season with values as Spring, Summer, Fall and Winter
   Mnth is the month of year
   Holiday shows if that particular day is a holiday or not
   Year the data is of from year 2018 and 2019
   weekday tells the day of the week
   workingday is the variable which captures if particular day is weekday or not
   weathersit has 4 values based on the weather of the day. It can take values as below
   1. Clear, Few clouds, partly cloudy, partly cloudy
   2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
   3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
   4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
   By analysing the above categorical variables, we can infer that,
   Season plays big role in customer renting the bikes and Fall has maximum customers renting and Spring has less customers compared to all the seasons.
   We can see July, September, June and August has the maximum number of customers.
   Bikes are most rented during non-holidays.
   When the weather is clear number of bike rents are high and its low when the Snow/Rainy days.

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

   Ans : By using drop_first= True while creating the dummy variables, it creates n-1 categorical instead of n categorical variables.
   If we create n dummy variables the model becomes redundant and complex.

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

   Ans : By looking at pair-plot temp and atemp both has the highest correlation with the target variable cnt.(both having value of 0.65)

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

   Ans : By creating scatter plots between the predicted values and the actual values we can see the scatter points form almost a straight line and it suggests linearity.

   By plotting residuals (actual - predicted values) against the predicted values we didn’t see any clear pattern which indicates linearity.

We have calculated the VIF (Variance Inflation Factor) values for each of the selected variables and didn't see any high VIF (>5) values which confirms no multicollinearity

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 the final model temp, weathersit (Light Snow, Light Rain) negatively and Yr variables are significantly contributing the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Ans : Linear Regression is a method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between dependent and independent variables to find the best-fitting line such that minimizes the distance between the predicted values and the actual values (or Error).

   Steps Involved in liner regression models are:

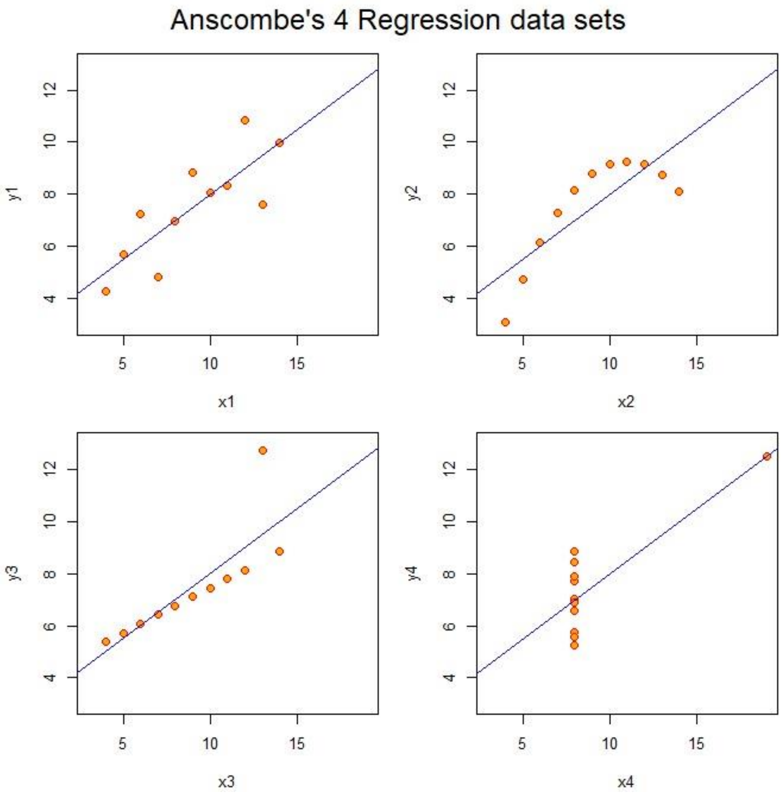
   - Preparing the Data for analysis: Collect and perform EDA on the data.
   - Specifying the Model: Linear regression models are represented by the equation as
     $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$
     where, y is the dependent variable,
     β0 is the constant,
     β1, β2… βn are the coefficients of the independent variables x1, x2…xn
     and ε is the error term.
   - Optimisation: Use method like least squared error to estimate the coefficients of the independent variables.
   - Model evaluation: Assess the performance of the model by using values of R-squared/adjusted R-squared.
   - Prediction: Use the model to predict the dependent value.

   Interpretation of the β coefficient is such that, The unit change in X_i leads to β_i changes in Y assuming all the other independent variables remain constant.

2. Explain the Anscombe's quartet in detail.

   Ans : Anscombe's quartet is a set of four data sets that have the same statistical properties such as mean, variance, and correlation-between x and y-and yet they look drastically different when graphed. This shows the importance of visual exploration of data is also very important.

Below the Anscombe's quartet datasets
Dataset 1: This is a typical linear relationship. The points are closely clustered around the regression line, indicating a strong positive linear correlation between x and y
Dataset 2: Quadratic relationship, not linear.
Dataset 3: Linear relationship, one outlier.
Dataset 4: Perfect linear relationship with constant x value.



The datasets above are very different when we plot it and see it visually, but if we just check the statistics all 4 datasets give the same number. Without visual interpretation of the datasets, we may end up with misleading conclusions.

3. What is Pearson's R?

Ans : Pearson's R is a statistical technique that is used to measure the strength and direction of relation between two variables that are linearly related. It is also known as a correlation coefficient. It ranges from -1 to +1.
Let x and y be two linearly related variables, then R can be calculated using the formula

$$R = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

The direction of the relation can be identified using a scatter plot, but the magnitude of the relation is calculated using R. A sign of the R tells the direction of the relationship. When the R value is 0, we say that there is no relation between X and Y. And when R is exactly 1, it can be said that there exists perfect correlation between X and Y, and in a similar way, we say there exists exactly negative correlation when R is -1. When R is greater than 0.5, it is interpreted as there exists a strong positive correlation, and when R is less than -0.5, we can say that there exists a strong negative relation between two variables.

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

   Ans : Scaling is one of the preprocessing techniques applied to numerical data so that a common range or scale is achieved.
   This provides for a better performance algorithm
   There are two types of scaling namely standardisation and MinMax scaling.
   Normalized Scaling or Min-Max Scaling, and it scales all features into a common range typically in the scale from 0 to 1.
   Formula: (x - min(x)) / (max(x) - min(x))
   Standardized Scaling (Z-Score Scaling):
   Scale data such that the mean is 0 and the standard deviation is 1.
   Formula: (x - mean(x)) / std(x)

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

Ans : VIF is the Variance Inflation Factor. It would become infinite, in case of an extreme form of multicollinearity exists. Therefore, one or more than one predictor variables with respect to regression models are perfectly correlated to each other.

There can be the following reasons for which VIF may be infinite:

Perfect Correlation: If two or more predictors are perfectly correlated (say one is a constant multiple of another), their VIF will be infinite. In such a situation, the variance of the coefficient for one variable becomes infinite as it is trying to describe the variation that already exists due to another perfectly correlated variable.

Not creating the dummy variables correctly: In creating the dummies of a categorical variable, if all levels are taken without dropping one, the multicollinearity is perfect because sum of all the dummies for each observation must add up to 1.

Redundant Predictors: When there are redundant predictors in your model, for instance, two predictors measuring the same thing, it also leads to high VIF, up to infinity.

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

   Ans : The Q-Q plot or quantile – quantile plot is a plot which is used to determine whether two data set is coming from same population or not. A q-q plot is a plot of quantile of one data set plotted against the quantile of another data set. A straight line with 45^0 is also plotted for a reference purpose. This line is known as Q- Q line. If most of the points falls near to the Q – Q line, then we can say that both data sets from the same population.

The procedure to draw the q- q plot:

- Collect the data. Collect the numerical random data from the population of the interest for which the Q-Q plot is required.

- Sort the data. Sort the data in ascending or descending order. This step is required for proper quantile calculations.

- Choose a theoretical distribution: Choose the distribution of the population (commonly taken as Normal distribution) and obtain the quantiles of the distribution. Theoretical quantiles can be obtained using standard normal table.

- Plotting: Obtain the scatter plot by taking sorted values on the y axis and theoretical quantiles on the x-axis.

Interpretation: If most of the points falls on a straight line then it can be concluded that data fits the assumed distribution.

Use of a Q-Q Plot in Linear Regression:

- Check for the normality of residuals: The important assumption of linear regression is errors are from normal distribution. Normality assumption is importance for test for significance of the regression coefficient, goodness of fit of overall model.

- Detecting the outlier: Q-Q plots can help discover outliers by displaying data points that deviate significantly from the distribution's predicted trend. Outliers might appear in a plot as points that depart from the predicted straight line.

Importance of a q-q plot

- Model Validity and improvement: Assists in validating the assumption of normally distributed errors, which is necessary for drawing appropriate conclusions from the regression model. If the Q-Q plot deviates from normality, consider using transformations or alternative models to improve the fit.

- Assumption Testing: The Q-Q plot is a diagnostic tool for determining if the normalcy assumption holds, which impacts confidence intervals and p-values.