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?**

There were total 6 categorical variables in the dataset. Those variables are “season” , “workingday” , “weathersit” , “weekday” , “yr” , “holiday” , “mnth”

As we know , our best fit line is

- Best fit line is Count = 0.1430 + 0.2350 x year + 0.0411 x workingday + 0.4890 x temp - 0.1462 x windspeed + 0.0672 x Sep + 0.0561 x Mon - 0.2717 x Light Snow - 0.0787 x (Mist + Cloudy) - 0.0604 x Spring + 0.0592 x Summer + 0.0957 x Winter

- Positive factors:
  - Year
  - Working Day
  - Sep
  - Mon
  - Summer
  - Winter
- Negative factors:
  - Light Snow
  - Mist+Cloudy
  - Spring

2 years data is available and there is increase in counts from year 2018 to 2019.

The counts of rental bikes are higher when there it is working day and mainly Monday as compared to other days.

The counts of rental bikes are higher in the month of September.

The counts of rental bikes are higher in the Summer and Winter seasons as compared to other seasons.

When the weather situation is Light Snow or Mist + Cloudy the counts of bikes is lower and it is an negative factor that affects the count of the rental bikes.

When it is Spring season, then the counts of bikes is lower i.e. it is a negative factor that affects the count of rental bikes.

These are all the categorical variables that will be significant in predicting the demand for shared bikes.

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

   A dummy variable is created to cover the range of values of the categorical variable. The values of each dummy variable are 1 and 0. A 1 represents the presence of each category and a 0 represents the absence. That is, if a categorical variable has three categories, there are three dummy variables.
   Drop_first = True is used when creating a dummy variable to drop the base/reference category. This is to ensure that the model does not contain multicollinearity when all dummy variables are included. A reference category can be easily derived if one row has 0 for all other dummy variables of a given category.

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

"temp" is the variable which has the highest correlation with target variable i.e., 0.63.

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

1. Linear relationship between independent and dependent variables
– Linearity is verified by examining points symmetrically distributed around the diagonal of the actual and predicted plots

2. Error terms are independent of each other , there is no specific pattern in the error terms for the predictions. Therefore, the error terms can be said to be independent of each other.

3. The error term is normally distributed: Histograms and distribution plots are useful for understanding the normal distribution and mean 0 of the error term.

4. The variance of the error term is constant (homogeneous variances):
It can be seen that the variance of the error term is approximately constant. Therefore, it follows the homoscedasticity assumption.

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

Top 3 features are :

1. Temp
2. Light snow (negative indicator)
3. Year

- If the temp is 1 unit and rest all other variables are constant , then the count of the rental bikes goes up by 0.632
- If the light snow is 1 unit and rest all other variables are constant , then the count of the rental bikes goes down by -0.1287
- If the year is 1 unit and rest all other variables are constant , then the count of the rental bikes goes up by 0.378

General Subjective Questions

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

• Linear regression is a method of finding the best linear relationship among independent and dependent variables.
• The algorithm uses the line of best fit to represent the association between the independent and dependent variables.

- There are 2 types of linear regression algorithms
  - Simple Linear Regression – Single independent variable is used.
    - $Y = \beta_0 + \beta_1 X$ is the line equation used for SLR.
  - Multiple Linear Regression – Multiple independent variables are used.
    - $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \in$ is the line equation for MLR.
  - $\beta_0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$
  - $\beta_1, \beta_2, \ldots, \beta_p = Slope\ or\ the\ gradient.$
- Cost functions – The cost functions helps to identify the best possible values for the $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ which helps to predict the probability of the target variable. The minimization approach is used to reduce the cost functions to get the best fitting line to predict the dependent variable. There are 2 types of cost function minimization approaches – Unconstrained and constrained.
  - Sum of squared function is used as a cost function to identify the best fit line. The cost functions are usually represented as
    - The straight-line equation is $Y = \beta_0 + \beta_1 X$
    - The prediction line equation would be $Y_{pred} = \beta_0 + \beta_1 x_i$ and the actual Y is as $Y_i$.
    - $Now\ the\ cost\ function\ will\ be\ J(\beta_1, \beta_0) = \sum (y_i - \beta_1 x_i - \beta_0)^2$
  - The unconstrained minimization is solved using 2 methods
    - Closed form
    - Gradient descent
- While finding the best fit line we encounter that there are errors while mapping the actual values to the line. These errors are nothing but the residuals. To minimize the error squares OLS (Ordinary least square) is used.
  - $e_i = y_i - y_{pred}$ is provides the error for each of the data point.
  - OLS is used to minimize the total $e^2$ which is called as Residual sum of squares.
  - RSS $= = \sum_{i=1}^{n} (y_i - y_{pred})^2$
- Ordinary Lease Squares method is used to minimize Residual Sum of Squares and estimate beta coefficients.

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

Statistics such as variance and standard deviation are usually considered sufficient parameters to understand the variation in some data without looking at each data point. Statistics are great for explaining general trends and aspects of your data.
Francis Anscombe recognized in his 1973 that statistical measures alone are not sufficient to represent datasets. He created several data sets with some identical statistical properties to illustrate the facts.

• Anscombe's quartet means that multiple data sets with many similar statistical properties can differ even when plotted.
• Quartet warns you about the danger of outliers in your data set. Without these outliers, the descriptive statistics in this case would have been very different.

### 3. What is Pearson's R?

The Pearson's R (also known as Pearson's correlation coefficients) measures the strength between the different variables and the relation with each other. The Pearson's R returns values between -1 and 1. The interpretation of the coefficients are:

- -1 coefficient indicates strong inversely proportional relationship.
- 0 coefficient indicates no relationship.
- 1 coefficient indicates strong proportional relationship.

$$r = \frac{n(\Sigma x * y) - (\Sigma x) * (\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2] * [n\Sigma y^2 - (\Sigma y)^2]}}$$

Where:
N = the number of pairs of scores
$\Sigma xy$ = the sum of the products of paired scores
$\Sigma x$ = the sum of x scores
$\Sigma y$ = the sum of y scores
$\Sigma x^2$ = the sum of squared x scores
$\Sigma y^2$ = the sum of squared y scores

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

• Scaling is a data preparation step for regression models. Scaling normalizes these different data types to a specific data range.
• In most cases, characteristic data are collected in the public domain and the interpretation of variables and the units of these variables are kept as open as possible. This leads to greater variability in units and data space. If these datasets are not scaled, it is more likely that the data will be processed without proper unit conversions. Also, the larger the range, the more likely the coefficients are to be biased to compare the variances of the dependent variables.
Scaling affects coefficients only. Forecasts and forecast accuracy remain unaffected after scaling.

Normalization/Min-Max scaling – The Min max scaling normalizes the data within the range of 0 and 1. The Min max scaling helps to normalize the outliers as well.

$$MinMaxScaling: x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Standardization converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.

$$Standardization: x = \frac{x - mean(x)}{sd(x)}$$

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

$$VIF = \frac{1}{1 - R^2}$$

The VIF formula can state when the VIF will be infinite. If the $R^2 = 1$ then the VIF is infinite. The reason for $R^2 = 1$ is that there is a perfect correlation between 2 independent variables.
$R^2 = 1$ it means that 100 percent of variance it can explain.

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

Ans: A quantile-quantile (QQ) plot plots the quantiles of a sample distribution with a theoretical distribution to indicate whether the affected dataset follows a distribution such as normal, uniform, or exponential. used to make decisions. It helps determine whether two data sets follow the same type of distribution. It's also useful to check if the data set errors are normal.