Assignment-based Subjective Questions
Submitted by: Rishi Garhwal

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Sol:

As per the analysis of categorical variables from the dataset, following inferences can be drawn:

- Year- Their seems an increase in count of bike rentals in year 2019 compared to 2018.
- Weathersit – Clear weather is an important aspect as count of bike rentals is more during clear weather.
- Season – Fall and Summer are more favourable for bike rentals than spring.

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

Sol:

- To avoid multicollinearity, redundant features and affects the model adversely.

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

Sol:

- Count (target Variables) has significantly high correlation with temperature (temp).

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

Sol:

- Maintains linear relation between dependent variable.
- Residual errors follow normal distribution.

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

Sol:

As per the final Model, the top 3 features that influences the bike booking are:

- Temperature (0.4233)
- Weather (-0.3136)
- Year (0.2320)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Sol:

- Linear Regression is a machine learning algorithm which is based on supervised learning category. It finds a best linear-fit relationship on any given data, between independent (Target) and dependent (Predictor) variables.

  Linear regression is of the 2 types:

  i. Simple Linear Regression: It explains the relationship between a dependent variable and only one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

  ii. Multiple Linear Regression: It shows the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X. It fits a 'hyperplane' instead of a straight line.

2. Explain the Anscombe's quartet in detail.

Sol:

- Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

3. What is Pearson's R?

Sol:

- Pearson's R is a correlation coefficient which is a measure of the strength of a linear association between two variables and it is denoted by 'r'. It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. A higher absolute value of the correlation coefficient indicates a stronger relationship between variables.
- In other words, if the value is in the positive range, then it shows that the relationship between variables is correlated positively, and both the values decrease or increase together. On the other hand, if the value is in the negative range, then it shows that the relationship between variables is correlated negatively, and both the values will go in the opposite direction.
- Pearson's Correlation Coefficient formula is as follows,

$$\mathbf{r} = \frac{\mathbf{n(\sum xy) - (\sum x)(\sum y)}}{\sqrt{\mathbf{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}}$$

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

Sol:

- Scaling is the process to normalize the data within a particular range. Many times, in our dataset we see that multiple variables are in different ranges. So, scaling is required to bring them all in a single range.
- The two most discussed scaling methods are **Normalization** and **Standardization**. Normalization typically scales the values into a range of [0,1]. Standardization typically scales data to have a mean of 0 and a standard deviation of 1 (unit variance).

  Formula of Normalized scaling:

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

Formula of Standardized scaling:

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

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

Sol:

- VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.
- The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1 - R_i^2}$$

Where, ‘i’ refers to the ith variable.

- If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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

Sol:

- The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

- A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.
- **Use of Q-Q plot in Linear Regression:** The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.
- **Importance of Q-Q plot: Below are the points:**
  - i. The sample sizes do not need to be equal.
  - ii. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.
  - iii. The q-q plot can provide more insight into the nature of the difference than analytical methods.