Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

- For Season : Maximum count was recorded in fall season

- For weather situation : Count is max in weather situation 1 (Clear, Few clouds, Partly cloudy, Partly cloudy) an is min in weather situation 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)

- For year: Significant increase in number of counts in 2019 from 2018.

- For months: Counts in middle of the year is higher than in the beginning and the end of the year.

- For weekday: Median of the Count fairly lie in similar range across week.

- For holiday: Count of total rental bikes is lower on holidays.

- For working day: Count of total rental bikes is higher on working days.

Q2. Why is it important to use **drop_first=True** during dummy variable creation?

If we don't do that, we would have an extra column that adds no new information that the previous columns don't provide when looked together. I will just increase multicollinearity.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

With the target, "temp" and "atemp" have the highest correlation which is equal (0.63).

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Validated the assumption of Linear Regression Model based on below assumptions –

- Normality of error terms : Error terms should be normally distributed. Checked using residual plot we got normal distribution.

- Multicollinearity check : There should be low multicollinearity among variables. Checked VIF values and removed those variable having high VIF.

- Error terms are uncorrelated with each other : No auto correlation. Plotting residuals against the order of residuals.

- Homoscedasticity : There should be no visible pattern in residual values. They should have constant variance.

- Linear relationship between X and Y : Linearity should be visible among variables. Checked using scatter plots.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

- Temperature
- year
- Category 3 of Weathersit which is Light Snow, Light Rain+Thunderstorm+Scattered Clouds, or Scattered clouds+light rain

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

- Linear Regression algorithm, as the name suggests, is a regression algorithm that identifies a linear relation between dependent or the set of dependent variables with the independent variables.
- The aim of this algorithm is to identify a line that explains the trend with minimum error.
- It is highly useful when dependent and independent variables have a significant correlation, i.e., one increases or decreases at the same pace as the other changes.
- Mathematically the relationship can be represented with the help of following equation − Y = mX + c Here, Y is the dependent variable we are trying to predict. X is the independent variable we are using to make predictions. m is the slope of the regression line which represents the effect X has on Y c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.
- Linear regression is of the following two types: Simple Linear Regression and Multiple Linear Regression
- Assumptions - The following are some assumptions about dataset that is made by Linear Regression model –
    - Multi-collinearity – Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
    - Auto-correlation – Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
    - Relationship between variables – Linear regression model assumes that the relationship between response and feature variables must be linear.
    - Normality of error terms – o Error terms should be normally distributed
    - Homoscedasticity – o There should be no visible pattern in residual values.

Q2. Explain the Anscombe’s quartet in detail.

- Anscombe’s quartet contains four datasets that have similar statistical features like mean, variance, correlation between dependent and independent variables, equation of linear regression line, and coefficient of determination.
- However, these datasets have a very different distribution and appear extremely different when graphed.
- This shows that statistics alone can make people take wrong decisions and graphing the data is also important alongside analysis.

Q3. What is Pearson's R?

- Pearson's R is also called the bivariate correlation. It measures the linear correlation between two variables and its value lies between -1 and 1.
- Mathematically, Pearson's R is the covariance of the two variables divided by the product of their standard deviation.
- If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

- Usually, the raw data available to perform statistical analysis on has different range for different features. This is a problem for machine learning problems, and statistical analysis techniques in general, and therefore we need to bring all the data to same or similar range. For this, we use scaling techniques which uses different information from the data and brings it to the same scale.
- Scaling also helps in converging of gradient decent faster.
- Difference between scaling techniques

Normalized Scaling	Standard Scaling
Uses Minimum and Maximum Value	Uses and Mean and Standard Deviation
Ensures scale to be between 0 and 1 or -1 and 1, depending on technique used and requirements	Ensures the mean is 0 and Standard Deviation is 1.
Useful when we don't know about the distribution	Useful when the feature distribution is Normal or Gaussian.
In SkLearn, use MinMaxScaler	In SkLearn, use StandardScaler

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

- As Variance Inflation Factor (VIF) is inversely proportional to (1-R^2), infinite VIF is only possible when R^2 is 1.
- Value of R^2 is going to be 1 if and only if there is a perfect correlation between two independent variables.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

- Q-Q plot stands for Quantile-Quantile plot.
- In simple terms, a Q-Q plot tells you about the distribution the given data follows. A slope of 1 on a Q- Q plot indicates that the given data follows a perfectly normal distribution.
- In practical terms, let's say you got two datasets – set 1 is the training data and set 2 is the test data or the data you want to make predictions on. We want to see that both the datasets follow the same distribution. For this, a Q-Q plot is an ideal way of approaching the problem.
- Use of Q-Q plot: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.
- Importance of Q-Q plot: When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests