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?

   Answer

   - The year box plots indicate that more bikes were rented in 2019.
   - The season box plots indicate that more bikes are rented during the season.
   - The working day and holiday box plots indicate that more bikes are rented during normal rented days than on weekends or holidays.
   - The monthly box plots indicate that more bikes are rented during September.
   - The weekday box plots indicate that more bikes are rent on Saturday.
   - The weathersit box plots indicate that more bikes are rent during Clear, Few clouds, Partly cloudy weather.

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

   Answer: it Returns a subsequence containing all but the given number of initial elements, and if drop_first is true, it removes the first column, which is created for the first unique value of a column

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

   Answer: atemp and temp both have the same correlation with the target variable of 0.63, which is the highest among all numerical variables.

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

   Answer:

   if this assumption is met or not is by creating a scatter plot x vs y. If the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variables, and the assumption holds.

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

   Answer:

   1. weathersit_Light_Snow(negative correlation).
   2. yr_2019(Positive correlation).
   3. temp(Positive correlation)

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   Answer:
   Linear regression is one of the very basic forms of machine learning where we train a model to predict the behavior of your data based on some variables. In the case of linear regression, as you can see the name suggests, linear means the two variables, which are on the x-axis and y-axis, should be linearly correlated.
   Mathematically, we can write a linear regression equation as: y=a+bx
   Here, x and y are two variables on the regression line.
   b = Slope of the line
   a = y-intercept of the line
   x = Independent variable from dataset
   y = Dependent variable from dataset

2. Explain Anscombe's quartet in detail.
   Answer:
   Anscombe's quartet comprises four datasets with 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 analyzing it and the effect of outliers on statistical properties

3. What is Pearson's R?
   Answer
   In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Answer
   It is a step of data Pre-Processinthatch is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

S.NO.	Normalisation	Standardisation
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

Answer

If there is a perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this pro,blem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Answer

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to determine if two sets of data come from the same distribution.

A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.