Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)
The categorical variables available in the dataset are season, weathersit, workingday, weekday, mnth, yr and holiday.

- Season: Most favorable season for biking is Fall followed by Summer. While least business is during spring time.
- weathersit: weathersit shows clearly that bike demand is on top during 'Clear' weather followed by 'Mist' and then 'Light Snow' time.There is no data available for 'HeavySnow' weather which may be due to the fact that there is almost no demand during that time.
- weekday: It doesn't show any specific trend with target variables. Though bike usage is higher for registered users which means bikes are being used for office commute.
- mnth: Business seems to be better in month with supportive weather conditions (Mar - Oct) as compared to extreme weather months (Jan, Feb, Nov, Dec)
- yr: Bike business has increased with the year from 2018 to 2019 which shows business is growing Y0Y.
- workingday, holiday: Working day (Holiday = 0) shows median (50th percentile) higher as compared to weekend though it is not as clear from workingday variable**.**

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)
Dummy variables are created using one hot encoding for categorical variable values where each dummy variable is being represented by 0's and 1's. 1 represents presence while 0 represents absence of the value.
drop_first=True is used to drop the base/reference category value. Base value can be derived easily keeping 0 in complete row for all other values of that variable category. This helps to avoid multicollinearity for the categorical variable.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

- temp variable has the highest correlation with target variable which is 0.63.
- Not considering casual and registered users as they are part of the target variable ('cnt') itself.
- Similarly, atemp has multicollinearity with temp, windspeed and humidity.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

- **Linear relationship b/w independent and dependent variables**: Use scatter plot to establish that there is a linear trend between independent and dependent variables.
- **Error terms are normally distributed :** Plot distplot of residuals (actual - predicted target

variable), it shall be normal distribution with mean 0.

- **Error terms are independent of each other**: Plot scatterplot for error terms vs predicted value. There shall not be any relationship between them.
- **Error terms have constant variance (homoscedasticity)** : There is constant variance on scatter plot for predicted vs actual values.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

- Temp: Temperature is the most significant feature with a positive impact on Bike business.
- yr : Bike business has increased with the year from 2018 to 2019 which shows business is growing Y0Y.
- weatherit : Weather conditions (Spring, Mist, Snow) has the most negative impact on Bike business.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

- Linear regression algorithm purpose is to find the linear relationship between independent and dependent variables.
- Algorithm use the best fitting regression line to map the relationship between them.
- There are 2 types of linear regression:
- **Simple Linear Regression (SLR)**: Single independent variable is used.
- $Y = \beta 0 + \beta 1X$ is the line equation for SLR.
- **Multiple Linear Regression (MLR)**: Multiple independent variables are used.
- $.Y = \beta 0 + \beta 1X1 + \cdots + \beta pXp + \in$ is the line equation for MLR.
- $\beta 0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$
- $\beta 1, \beta 2, \ldots, \beta p = Slope\ or\ the\ gradient.$
- We can use Python libraries Statsmodel or SKlearn for linear regression.
- Ordinary Least Squares (OLS) method is used to minimize Residual Sum of Squares and estimate beta coefficients.
- Important statistics to consider while finding the best model (F Statistics, R squared, p value, Variance Inflation Factor (VIF) for multicollinearity.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>
Statistics like variance and standard deviation are usually considered good enough parameters to understand the variation of some data. But Francis Anscombe established that only statistical measures are not good enough to depict the data sets. He created several data sets all with several identical statistical properties to

illustrate the fact.

Important points:

- Anscombe's Quartet signifies that multiple data sets with many similar statistical properties could still be different from one another when plotted.
- The dangers of outliers in data sets are warned by the quartet.
- Plotting the data is very important and a good practice before analyzing the data.
- Outliers should be removed while analyzing the data.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

The Pearson's R (or Pearson's correlation coefficients) measures the strength between the different variables and the relation with each other. Pearson's R 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 positive proportional relationship.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

- What - The scaling is the data preparation step for a regression model. The scaling normalized these varied data types to a particular data range.
- Why – Scaling normalizes features with diverse ranges and units to a common scale. This ensures that features with larger scales don't disproportionately influence the model. By standardizing the data, we can improve the model's performance and make the coefficients more interpretable. While scaling impacts the magnitude of coefficients, it doesn't alter the model's predictive power or accuracy.
- 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. $MinMaxScaling: x = (x - \min(x)) / (\max(x) - \min(x))$
- Standardization converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.
  $Standardization: x = (x - mean(x)) / sd(x)$

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

$VIF = 1 / (1 - R \text{ Square})$

The VIF formula clearly signifies if the R Square is 1 then the VIF is infinite. The reason for R Square to be 1 is that there is a perfect correlation between 2 independent variables.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

Q-Q plots are graphical tools used to compare the distributions of two datasets. By plotting the quantiles of one dataset against the quantiles of another (often a theoretical distribution like normal, exponential, or uniform), we can visually assess if they come from the same underlying distribution. In linear regression, Q-Q plots are helpful in determining if the training and test data share similar distributional properties

Interpretations

a. Similar distribution: If all the data points of quantile are lying around the straight line at an angle of 45 degree from x-axis.

b. Y values < X values: If y-values quantiles are lower than x-values quantiles.

c. X values < Y values: If x-values quantiles are lower than y-values quantiles.

d. Different distributions – If all the data points are lying away from the straight line.

Advantages

a. Distribution aspects like loc, scale shifts, symmetry changes and the outliers all can be identified from the single plot.

b. The plot has a provision to mention the sample size as well.
