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)

1.) yr (Year): A 1-unit increase in yr increases cnt by 23.46%. As the time passes the count of bike booking will increase year on year.

2.) workingday: On working days, the count increases by 5.48%, holding other factors constant.

3.) windspeed: A unit increase in wind speed reduces cnt by 16.10%.

4.) Season Variables: Spring (season_spring): −0.0717 (negative impact, significant).
Summer (season_summer): 0.0338 (positive impact, significant).
Winter (season_winter): 0.0919 (positive impact, significant).

5.) Weather Conditions:
Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (weathersit_light_snow): Strongly reduces cnt by 29.80%.

Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist (weathersit_mist_cloudy): reduces cnt by 8.31%.

6.) temp : A unit increase in temperature is associated with a 43.44% increase in cnt.

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)

When creating dummy variables for a categorical variable with n categories, n dummy variables are generated by default (one for each category). To generate only n-1 variables, we have to use drop_first=True.

Including all dummy variables means the information is redundant, which could result in multicollinearity.

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 and atemp variables have the highest correlation with the target variable.

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)

1.) Linearity - There is a linear relationship between X and Y — This was verified by residual plot(residuals vs predicted). Residuals are randomly scattered around zero, it implies linearity because the linear regression model assumes that the relationship between the independent variables and the dependent variable is linear.

2.) Homoscedasticity - Error terms have constant variance — Drawing the residual plot(residuals vs predicted), it can be seen that the variance is not following any pattern as the error terms change.

3.) Normality of Errors - Error terms are normally distributed with mean zero — This is verified by drawing a histogram of error terms using sns.displot().

4.) Error terms are independent of each other — Drawing the residual plot(residuals vs predicted),

it can be seen that there is no visible pattern that means error terms are independent of each other.

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)

1.) temp : A unit increase in temperature is associated with a 43.44% increase in cnt.

2.) yr (Year): A 1-unit increase in yr increases cnt by 23.46%. As the time passes the count of bike booking will increase year on year.

3.) Weather Conditions:
Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (weathersit_light_snow): Strongly reduces cnt by 29.80%.

Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist (weathersit_mist_cloudy): reduces cnt by 8.31%.

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 is a statistical method used to model the relationship between a dependent variable (y) and one or more independent variables (X). It aims to predict the value of y based on X by finding the best-fit line (or hyperplane in higher dimensions).

Below is the equation:-

$y=\beta 0 +\beta 1\, x1 +\beta 2\, x2 +\cdots+\beta n\, xn +\epsilon$

Here we have,

y: Dependent (target) variable.

x1 ,x2 ,…,xn : Independent (predictor) variables.

β0 : Intercept (value of y when X=0).

β1 ,β2 ,…,βn : Coefficients of the independent variables.

ϵ: Error term (unexplained variance).

Assumptions for Linear Regression are as given under question number 4.

Method to find coefficients -

The coefficients are estimated using the Ordinary Least Squares (OLS) method, which minimizes the sum of squared residuals (SSR):

SSR= square of (actual value - predicted value) / total number of values

This is done by calculating the partial derivatives of SSR with respect to each coefficient and setting them equal to zero.

Once the model is built, its performance is evaluated using various metrics like R square , Adjusted R square.

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>

Anscombe's quartet in short says "same stats but different graphs". Basically it says that even though two or more data sets have identical summary statistics but the data sets can be very different when visualized graphically. This shows the importance of graphical visualization of data instead of just relying on the summary statistics.

Anscombe's Quartet is a set of four datasets created by the statistician Francis Anscombe in 1973. Each of the four datasets in Anscombe's Quartet consists of an independent variable x and a dependent variable y. The datasets are constructed such that they share nearly identical summary statistics like mean, variance, correlation. But visual characteristics of each data set are different.

1. First Data set portrays Linear relationship.
2. Second Data set represents Non linear relationship.
3. Third Data set represents outlier dominance . Most of the data points form a horizontal line, but one outlier strongly influences the slope of the regression line.
4. Fourth Data set depicts Vertical outlier influence. The x values are constant except for one outlier, which dominates the regression line.

This shows that,

a.) Visualization is very important and we should not just rely on summary statistics.Graphical exploration of data often reveals patterns, anomalies, or relationships that are not obvious from numerical summaries.

b.) Importance of outlier detection - Outliers can have huge impact of statistical measures. So its very important to get rid of unnecessary outliers before starting the analysis.

c.) Limitation of Linear Regression - Linear regression assumes a linear relationship between variables, but this may not be true always.

d.) Data Interpretation should not be done just based on summary statistics but context and visualization are extremely important to interpret data correctly.

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>

Pearson's R - Pearson correlation coefficient is a statistical measure that quantifies the strength and direction of the linear relationship between two variables.

$R=\text{cov}(x,y)/\sigma X\, \sigma Y$,

where $\text{Cov}(X,Y)$: Covariance between X and Y,

σX : Standard deviation of X.

σY : Standard deviation of Y.

Pearson's R ranges from **-1 to +1**, and its value indicates:

+1: Perfect positive linear relationship (as X increases, Y increases proportionally).

0: No linear relationship (but other relationships, like quadratic, might exist).

-1: Perfect negative linear relationship (as X increases, Y decreases proportionally).

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>
Scaling is the process of transforming data into a specific range or distribution. It adjusts the values of features in a dataset to ensure they are comparable and have consistent scales.

Scaling is performed because in models features with large ranges or high numerical values can dominate models, scaling ensures all features contribute equally.

Normalized scaling, scales data to a fixed range like in our assignment we have that range from 0 to 1, whereas Standardized scaling, Scales data to have zero mean and unit variance, this does not have any fixed range.

Formula for Normalized Scaling - (x-min(x)) / max(x)-min(x)
Formula for Standardized scaling - (x-mean)/standard deviation

---

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 square), for perfect multicollinearity R Square will be 1. Thus if we put 1 in place of (R square) in the given formula, the value will become 1/0 which is infinite.
Thus the value of the Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity among the predictors in a regression model. This occurs when one predictor is a perfect linear combination of one or more other predictors, leading to mathematical and computational issues in calculating the VIF.

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>
A Q-Q plot is used to compare the distribution of a dataset to a theoretical distribution. It plots the quantiles of the data against the quantiles of the theoretical distribution. The x-axis represents the quantiles of the theoretical distribution. The y-axis represents the quantiles of the observed data.
Use and importance of a Q-Q plot in linear regression - Q-Q plots are crucial in linear regression to assess one of its key assumptions: the residuals should be normally distributed. If the points align closely with the diagonal, the residuals are approximately normal. Deviations from the line suggest non-normality. This offers a straightforward visual method to assess normality compared to numerical tests

---