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)

I have done the analysis on all the categorical variables using bar and box plots

1. Season – The bookings are higher in fall season and least in spring season. Summer and winter almost have similar amount of bookings intermediate between the other two seasons.
2. Year – The bookings have increased year on year. 2019 saw more bookings than 2018
3. Month – The highest bookings are happening in the September month, least in January. The bookings are pretty high in the months through Jun-Aug
4. Holiday – The bookings are more on days when there is no holiday
5. Weekday – The bookings are almost same through all the days. There is no significant difference
6. Workingday – The bookings are more on working days than non-working days
7. Weathersit – The bookings are more when the weather is good, its very less when the weather is bad.

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)

We create Dummy variables for categorical variables with more than 2 categories.
So, when we have n categories, we would need only n-1 dummy variables

If incase we don't drop the extra column, it creates multicollinearity in the model.
Multicollinerality makes one of our assumptions of the model false.
To prevent this we drop 1 column from dummy variables (keep only n-1)variables

We use – drop_first = True for the same

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

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. We compare the y-test values and y-pred values
2. Plot the residuals to check Homoscedasticity
3. Multicollinearity
4. Linearity
5. Independence of residuals
6. Residuals are normally distributed and have mean 0

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)

Temperature, Year – Positive significance
Windspeed – Negative Significance

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 supervised machine learning algorithm used for predicting a continuous target variable (y) based on one or more predictor variables (X). It assumes a linear relationship between the predictors and the target.

It aims to find the best-fitting line that well describes the relation between all the predictor variables and the target avriables.

When we have only 1 independent/predictor variable – Simple Linear Regression

When we have more than 1 independent/predictor variable – Multiple Linear Regression

Assumptions of Linear Regression

- Linearity
- Independence
- Homoscedasticity
- Normality
- No Multicollinearlity or very little

To achieve these, we follow the method of OLS - OLS finds the values of the coefficients that minimize the sum of the squared differences between the observed and predicted values

Once the model is built, it can be evaluated on –

- R-squared
- Adjusted R-squared
- MSE
- RMSE
- Probability(F-Statistic)

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 was created by statistician Francis Anscombe in 1973 to illustrate the importance of plotting data before you analyze it and build your model

- Anscombe's quartet is a set of four datasets that, despite having nearly identical simple descriptive statistics (mean, variance, correlation, linear regression line), have very different distributions and scatter plots.
- Dataset 1 - Appears to follow a simple linear relationship
- Dataset 2 - Shows a clear curved relationship
- Dataset 3 - Shows a perfect linear relationship except for one outlier
- Dataset 4 - Shows that one point is exerting a strong influence, with all other x values being the same

This experiment highlights that

- Visualization is essential
- Limitations of Correlation and linear Regression
- Impact of outliers
- Model Slection

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. It is widely used in data analysis to understand how closely two variables are associated.

It describes –

- Strength – How close the data points cluster around a straight line

- Direction – If the relation is positive or negative

The pearson value ranges from -1 to +1

-1 : Strong negative linear relationship
0 : No linear relationship
1 : Strong positive linear 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>

Scaling means ensuring or transforming all the continuous variables to similar scale.
If all the continuous variables are on same scale – the interpretations or predictions provided by the model will be unbiased and wont be dominated by the larger values.

Its important because–

- Prevent feature dominance
- Faster convergence
- Improved model performance
- Regularization

Scaling can be done in 2 methods

- Normalized - MinMax
- Standardized - ZScore

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>

The Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity in a dataset. Perfect multicollinearity occurs when one predictor (independent variable) is an exact linear combination of one or more other predictors.

You can address it by

- Remove perfectly correlated variables
- Combine collinear variables
- Regularization
- Verify Data

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 a graphical tool used to assess whether a dataset follows a specific theoretical distribution, most commonly a normal distribution. It plots the quantiles of the observed data against the quantiles of the theoretical distribution. It is particularly useful for assessing whether a variable or residuals from a model follow a specific distribution

---

In linear regression, a Q-Q plot is typically used to check the assumption of normality of residuals. The assumptions underlying linear regression include:

- The residuals (errors) should follow a normal distribution.
- This normality assumption is crucial for the validity of hypothesis tests (e.g., ttt-tests for coefficients) and confidence intervals.
