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 below are the few inferences from the dataset

1. More no bikes have been demanded in the year 2019
2. Demand has been increased from 2018 to 2019
3. Fall seasons is having high number bike rental.
4. Bike rental in 2019 has increased for every season compared to 2018.
5. May to oct is having number of bike registration.
6. Bike registration in 2019 has increased for every month compared to 2018.
7. People prefer bike rental when weather situations is good.
8. Bike registration in 2019 has increased for every weather situation 2018.

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)

- The reason to drop the column during the dummy variable creation as it will become redundant and avoids multicollinearity by dropping one category per variable.
- By doing so will make the model more stable and easier to interpret and if we didn't drop the first column then the model could face issues with inflated variance and unreliable coefficients.

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)

The 'temp' and 'atemp' variables have highest correlation when compared to the rest of the target variables as 'cnt'.

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)

- Linearity: Plotted predicted values against actual values to verify a linear relationship and checked residual plots for random distribution around zero.
- Homoscedasticity: Reviewed residual vs. predicted values plots to ensure residuals had constant variance across predictions, avoiding funnel-shaped patterns.
- Normality of Residuals: Used a histogram to check if residuals followed a normal distribution.
- Independence of Errors: Checked for autocorrelation, especially in time-series data, to

confirm residuals were independent

- No Multicollinearity: Calculated Variance Inflation Factor (VIF) for each predictor to detect multicollinearity, ensuring VIF values were generally below 5.

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)

The top 3 features that contributed significantly are

1. temp
2. yr
3. season_winter, season_summer

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)

- Linear Regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors), assuming that this relationship can be represented as a straight line.
- Linear Regression is a type of Supervised Machine Learning.
- There are two types of linear regressions.
- Simple Linear Regression: Involves one independent variable and one dependent variable.
- ex: y=b0+b1x, where b0 interceptors and b1 is coefficient or slop is x. x is the predictor.
- Multiple Linear Regression: Involves two or more independent variables predicting one dependent variable
- ex: y= b0+b1x1+b2x2 + ...+bnxn, where b0 is interceptor and b1, b2,b3,...bn is
- coefficient/slopes of x1, x2, x3 ...xn predictors.
- In Linear Regression Target variable is a continuous value. So linear regression is finding a fitted line (the fitted plane in case of multiple linear regression) so that the sum of error between the target value and the predicted value is minimum
- Linearity: The relationship between variables is linear.
- Independence: Observations are independent.
- Homoscedasticity: Residuals have constant variance.
- Normality: Residuals are normally distributed (important for hypothesis testing).
- No Multicollinearity: Independent variables should not be highly correlated with each other.
- Loss Function: The objective is to minimize the difference between actual and predicted values using Mean Squared Error (MSE): MSE=1n∑i=1n(yi−y^i)2MSE
-

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)

- Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending on the summary statistics.
- It also emphasizes the importance of using data visualization to spot trends, outliers and other crucial details that might be obvious from summary statistics alone.
- Visualizing data helps reveal underlying structures, trends, and anomalies that may not be evident through numerical analysis alone.

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)

- Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1, where:
  - 1 indicates a perfect positive linear relationship, meaning as one variable increases, the other variable also increases proportionally.
  - -1 indicates a perfect negative linear relationship, meaning as one variable increases, the other variable decreases proportionally.
  - 0 suggests no linear relationship between the variables.
- Pearson's R is calculated using the covariance of the two variables divided by the product of their standard deviations

---

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)

- **Scaling** : It refers to the process of adjusting the range and distribution of data values to ensure that different features contribute equally to the model's performance.
- In optimization algorithms (like gradient descent), scaling helps speed up convergence, making it easier for the algorithm to find optimal parameters.
- It ensures that features with larger ranges do not disproportionately influence the model, promoting fair comparisons between features.
- **Normalized Scaling** (Min-Max Scaling):
  - Transforms features to a fixed range, usually [0, 1].
  - Formula: X'=(X−Xmin)/(Xmax−Xmin)
  - Useful when the distribution is not Gaussian and when all features should be constrained within a specific range.
- **Standardized Scaling** (Z-Score Normalization):
  - Transforms features to have a mean of 0 and a standard deviation of 1.
  - Formula: X'=(X−μ)/σ
  - Suitable for Gaussian distributions, as it retains the relative relationships of the data

while centering it around zero.

- Each scaling method serves specific purposes depending on the nature of the data and the requirements of the modeling algorithm.

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)

- An infinite Variance Inflation Factor (VIF) arises due to perfect multicollinearity among predictors in the model.
- This situation occurs when one predictor is an exact linear combination of other predictors, rendering it completely redundant.
- In such cases, VIF cannot be computed because the denominator of the VIF formula (which includes $1-R2$ of the predictor on other predictors) becomes zero.
- To resolve the multicollinearity issue, it is essential to remove or adjust the redundant variable.

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)

- A Q-Q (Quantile-Quantile) plot is a graphical method for comparing the distribution of data against a theoretical distribution, typically the normal distribution. In linear regression, it is essential for evaluating the normality assumption of residuals.
- The Q-Q plot displays the quantiles of the residuals against the quantiles of a normal distribution. A straight-line pattern suggests that the residuals are approximately normal.
- Normal residuals allow for reliable hypothesis testing using t-tests and F-tests in regression analysis. Significant deviations from normality in the plot may indicate model issues, suggesting the need for transformations or adjustments to improve prediction accuracy and the validity of confidence intervals.
