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)

Answer: Using bar plot I am able to compare how change in one variable affects dependent variable. Here are some inferences.

- With using BoxPlot it is clear that , there is no much outliers in the given data set.
- During Fall and Summer seasons, bike rental demand is more. It is also observed that rentals are more in Fall and Summer seasons of both years 2018 and 2019.
- Demand is increased in 2019 when compare to 2018.
- Count is more that 5000 during the months from May to October. Also highest bikes are counted in 2019 September.
- Count in all weekdays is nearly same.
- More rental bikes if whether it is clear.

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)

Answer: Setting drop_first=True removes the first dummy variable during creation. This is important because an n-level categorical variable can be effectively represented with n-1 dummy variables. Including the extra dummy variable would introduce redundancy and multicollinearity, which can negatively impact the computation of model 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)

Answer: temp

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)

Answer: To evaluate a model, we can examine the difference between the actual observed target values and the predicted values, known as residuals. Residual analysis helps validate model assumptions through the following methods:

1. **Histogram Plot**: If the histogram of residuals forms a bell-shaped curve, it indicates that the residuals are normally distributed.
2. **Q-Q Plot**: If the residual points align closely with the standard normal distribution line, it suggests that the residuals are normally distributed.
3. **Residuals vs. Fitted Plot**: If the residuals are randomly scattered around zero with no discernible pattern, it indicates that the model is appropriate.

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
- Light snow rain
- Windspeed

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>
Answer:

- Linear regression is a supervised machine learning algorithm used to predict a continuous target variable (YYY) based on one or more independent variables (XXX). It establishes a linear relationship, represented by:
- Y=β0+β1X1+β2X2+⋯+βnXn+ϵY = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilonY=β0+β1X1+β2X2+⋯+βnXn+ϵ
- where β0\beta_0β0 is the intercept, βi\beta_iβi are coefficients, and ϵ\epsilonϵ is the error term.
- The algorithm minimizes the Sum of Squared Errors (SSE) using Ordinary Least Squares (OLS) to find the best-fit line. Predictions are made by plugging values of XXX into the equation.
- Key assumptions include linearity, independence of observations, homoscedasticity (constant variance of residuals), normality of residuals, and no multicollinearity. Violating these can affect model performance.
- The model is evaluated using metrics like R2R^2R2 (variance explained), Adjusted R2R^2R2 (corrected for predictors), and Mean Squared Error (MSE). It is simple, interpretable, and efficient for small datasets but sensitive to outliers and assumes linearity.
- Applications include sales prediction, price estimation, and forecasting trends. It works well when assumptions are met and the relationship is genuinely linear.

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>

Answer: Anscombe's quartet is also known as A Tale of Four Datasets. Each consisting of 11 data points (x,y). While these datasets appear remarkably similar at first glance but they exhibit different statistical points when visualised.

Dataset		Key statistical properties (Example)	Reveals distinct patterns when visualised
1	A simple linear relationship between x and y.	Mean of x,y – [9, 7.5], Variance of x,y - [10, 4], Correlation coefficient – 0.816	A clear linear relationship between x and y.
2	A quadratic linear relationship between x and y, with single outlier.	Mean of x,y – [9, 7.5], Variance of x,y - [10, 4], Correlation coefficient – 0.816	A quadratic relationship with outlier pulling the line away from majority of points.
3	A linear relationship between x and y with constant x value for most points and a single outlier	Mean of x,y – [9, 7.5], Variance of x,y - [10, 4], Correlation coefficient – 0.816	A vertical line with an outlier pulling correlation coefficient towards 1.
4	A linear relationship between x and y with two outliers.	Mean of x,y – [9, 7.5], Variance of x,y - [10, 4], Correlation coefficient – 0.816	A linear relationship with two outliers, one pulling line towards up and another pulling line towards down.

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>

Answer: Pearson's R is a statistical measure that quantifies the linear relationship between two variables it ranges from -1 to 1.

r = 1 Perfect positive relation meaning variables increase or decrease together perfectly.

r = -1 Perfect positive relation meaning one variable increases as the other variable decreases perfectly. r = 0 No correlation between the variables.

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>

Answer:

- Scaling is a technique to transform the data within the fixed range. It is a crucial step which ensures that all variables contribute equally for model learning process.
- Dominant features: When feature have significantly different scales, algorithms might give undue weight for larger values. This can lead to biased modelling.
- Fair Contribution: Scaling ensures that all features contribute to the model's decisions based on their relative importance, rather than their magnitude.
- Standardization, Scaling can standardize features to a common scale, making it easier to compare and interpret results.

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>

Answer: If the VIF value is infinite means there is a multicollinearity exists in the model. It means that predictor variable is highly correlated with other predictors which can lead to unstable and unreliable regression models. This can also happen if there is a numerical instability.

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>

Answer: Quantile-Quantile plot is a visual comparison of distributions. It is particularly useful for assessing whether datasets follows a specific theoretical distribution (eg: Normal distribution, Uniform distribution, Exponential distribution).

Uses:

Normality Testing: To assesses if dataset follows a normal distribution.

Comparing Distributions: To compare two datasets or a datasets with a theoretical distribution.

Outlier Detection: To identify outlier in the dataset.

Interpretation:

Straight Line: If the points fall approximately on a straight line, it suggests that the two distributions are similar in shape.

Deviation from Line: Deviations from the line indicate differences in the distributions. For example, a curve shape might suggest a different distribution or the presence of outliers.
