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)

To analyze the effect of categorical variables on the dependent variable, boxplots were plotted and below trends were observed:

1. Bike demand is highest in fall (season 3)
2. Bike demand is highest in sunny weather (weathersit 1: Clear, Few clouds, Partly cloudy, Partly cloudy). Stormy (weathersit 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog) has no bike demand at all.
3. Bike demand rises from Feb to June.
4. Bike demand has grown in next year
5. Demand decreases on holidays
6. More demand on working days

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)

Using drop_first=True helps avoid multicollinearity and allows for clearer interpretation of the regression coefficients, making the model more robust and interpretable. With p−1 dummy variables, the coefficients for the remaining dummies indicate how much each category differs from the reference category.

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

---

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)

The assumptions made for building the model were:

1. Error Terms were normally distributed – A seaborn distplot was plotted for the error terms, and it showed a normally distributed curve.
2. Error Terms have a mean at 0 – In the same distplot, the mean was observed to occur around 0.
3. The predicted and actual training data should have similar trend – A scatterplot was plotted against y_train and y_train_pred. The data points almost coincided.
4. The independent variables should not be too highly correlated with each other – The VIFs of variables used in model building had VIF less than 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)

1. Temp
2. Weather situation (1: Clear, Few clouds, Partly cloudy, Partly cloudy)
3. year

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 machine learning algorithm used to model the relationship between a dependent variable (target) and
one or more independent variables (features). The goal is to find the best-fitting straight line that predicts the target variable.

The equation is typically written as:

$y = mx + b$

where y is the predicted value, m is the slope (coefficient) indicating the relationship's strength, x is the feature, and b is the y-intercept.

The model is trained using a dataset by minimizing the difference between predicted and actual values,
usually through a method called "Ordinary Least Squares."
Once trained, the model can make predictions on new data.
Key assumptions include linearity, independence, and normality of errors, which should be checked to ensure model validity.

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 is a set of four datasets created by statistician Francis Anscombe to illustrate the importance of data visualization.
Each dataset contains the same statistical properties: they all have the same mean, variance, and correlation coefficient.

However, when plotted, they reveal very different patterns.

1. Dataset I: Linear relationship with no outliers.
2. Dataset II: Strong linear relationship but includes one outlier.
3. Dataset III: Curved relationship, indicating a non-linear trend.
4. Dataset IV: Linear relationship, again impacted by an outlier.

The key lesson from Anscombe's quartet is that summary statistics alone can be misleading. Visualizing data is crucial for understanding its underlying structure and relationships, emphasizing that different datasets can behave very differently despite similar statistics.

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, or Pearson correlation coefficient, measures the strength and direction of the linear relationship between two continuous variables.
It ranges from -1 to 1:

- 1: Perfect positive correlation—when one variable increases, the other does too.
- -1: Perfect negative correlation—when one variable increases, the other decreases.
- 0: No correlation—no linear relationship exists.

The formula is:
$r = n \sum X Y - \sum X \sum Y \, (n \sum X 2 - (\sum X) 2) \cdot (n \sum Y 2 - (\sum Y) 2)$. The terms in that formula are: n = the number of data points, i.e., (x, y) pairs, in the data set. $\sum X Y$ = the sum of the product of the x-value and y-value for each point in the data set.

While Pearson's R is useful for assessing linear relationships, it's important to note that it doesn't capture non-linear relationships. Additionally, outliers can significantly affect the value, so data visualization is recommended to complement its interpretation.

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 refers to the process of adjusting the range and distribution of feature values in a dataset. It is crucial in linear regression because features with different scales can lead to biased coefficients, affecting model performance.

Why Scale?:
Scaling helps in:
- Improving convergence speed in optimization algorithms.
- Ensuring that each feature contributes equally to the distance calculations in models.

Types of Scaling:

1. Normalized Scaling (Min-Max Scaling): Rescales the features to a fixed range, usually [0, 1]. The formula is:

X' = X - min(X)/max(X) - min(X)

2. Standardized Scaling** (Z-score Scaling): Centers the features around the mean with a standard deviation of 1. The formula is:

X' = X - mu / sigma

where mu is the mean and sigma is the standard deviation.

Both methods help improve the performance of linear regression models by ensuring consistent scales across features.

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) can become infinite when there is perfect multicollinearity among the independent variables in a regression model. This occurs when one independent variable can be expressed as an exact linear combination of other variables.

Reasons for Infinite VIF:

1. Exact Linear Relationships: If, for example, you have two independent variables where one is a direct multiple of the other. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

2. Redundant Variables: Including a variable that doesn't provide any new information (e.g., duplicate data) will also lead to infinite VIF values.

When VIF is infinite, it indicates that the model cannot reliably estimate the coefficients for those variables, making it essential to address multicollinearity through variable selection or transformation.

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 (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset against a theoretical distribution, commonly the normal distribution. In a Q-Q plot, the quantiles of the dataset are plotted against the quantiles of the reference distribution.

Uses and Importance in Linear Regression:

1. Normality Check: In linear regression, one key assumption is that the residuals (errors) are normally distributed. A Q-Q plot helps visually assess this assumption. If the points in the plot fall approximately along a straight line, it suggests that the residuals are normally distributed.

2. Identifying Deviations: Deviations from the line indicate departures from normality, such as skewness or kurtosis. This can alert you to potential issues with the model, such as non-normal errors, which can affect hypothesis testing and confidence intervals.

3. Model Validity: By ensuring that the residuals are normally distributed, a Q-Q plot helps validate the model, leading to more reliable predictions and statistical inferences.

Overall, Q-Q plots are essential for diagnosing assumptions in linear regression, contributing to the model's robustness and interpretability.
