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)

From my dataset, I analyzed the effect of the categorical variables on bike demand, and here's what I found:

1. Weather Situation (weathersit_3):
   - Based on my analysis, I observed that the weather situation variable (representing conditions like light snow or light rain) has a negative effect on bike rentals, with a coefficient of -0.3070. This suggests that, as I would expect, poorer weather conditions discourage people from renting bikes. Specifically, light snow or rain leads to a decrease in demand, likely due to discomfort or safety concerns related to biking in such weather.
2. Season 4 (season_4):
   - In my analysis, I found that the winter season (represented by the season_4 variable) is associated with a slight increase in bike rentals compared to spring (the reference category), with a coefficient of 0.128744. This was an interesting finding because, although winter is generally seen as a less bike-friendly season, I suspect that factors like seasonal promotions, better infrastructure management, or milder winter conditions in certain regions may contribute to higher bike rentals during this season.

In summary, from my dataset, I concluded that weather conditions and seasonality play an important role in influencing bike demand. The negative impact of poor weather (such as light snow or rain) reduces rentals, while the winter season surprisingly shows a moderate positive effect, which could be attributed to various external factors.

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)

It is important to use drop_first=True during dummy variable creation to avoid the problem of multicollinearity. Here's why:

1. Multicollinearity: When creating dummy variables from a categorical feature, you convert each category into a separate binary variable. If all categories are included in the model, the dummy variables can become highly correlated with each other. This creates perfect multicollinearity, where one variable can be perfectly predicted by the others, which causes issues with model estimation (e.g., inflated standard errors or inaccurate coefficient estimates).
2. Reference Category: By setting drop_first=True, I ensure that one of the categories (typically the first one) is dropped, and the remaining categories are used as dummy variables. The dropped category becomes the reference category, against which the other categories are compared. This approach avoids multicollinearity because the dropped category doesn't

contribute a separate variable, but its effect is implicitly captured in the intercept term of the regression model.

3. Interpretability: Dropping the first category makes the model coefficients easier to interpret. For the remaining categories, the coefficient reflects how the outcome variable (bike demand, in your case) changes relative to the dropped reference category. This avoids the issue of redundant information and makes it clear how each category contributes to the model's predictions.

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)

From my analysis of the model, temperature (temp) would likely have the highest correlation with the target variable (bike demand). The positive coefficient indicates that as temperature increases, so does bike demand, and this variable probably shows a strong positive correlation in the pair-plot. Year (yr) may also have a notable positive correlation, but temperature seems to be the most influential numerical 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)

Summary of Assumption Validation:

- Linearity: Verified using scatterplots and residuals plots.
- Independence: Checked using the Durbin-Watson test and residuals patterns over time.
- Homoscedasticity: Validated using residuals vs. fitted values plot.
- Normality of Errors: Checked with a histogram and Q-Q plot, and possibly the Shapiro-Wilk test.
- No Multicollinearity: Assessed using the Variance Inflation Factor (VIF).
- Model Specification: Checked using residuals plots to ensure no patterns suggesting missing variables.

By validating these assumptions, I ensure the linear regression model is well-specified and the results are reliable for making predictions and inferences.

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)

Based on the final model, the top 3 features contributing significantly towards explaining the demand for shared bikes are:

1. Temperature (temp):

    - Contribution: Temperature has the strongest positive relationship with bike demand (coefficient = 0.5636). As the temperature increases, the demand for bikes increases significantly, making it the most important predictor in the model. This is intuitive, as warmer weather encourages more outdoor activities, such as biking.
2. Weather Situation 3 (weathersit_3):
    - Contribution: Weather Situation 3 (representing light snow or light rain) has a notable negative relationship with bike demand (coefficient = -0.3070). This variable significantly impacts the demand by discouraging bike rentals during unfavorable weather conditions. The negative coefficient indicates that poor weather conditions, such as light snow or rain, decrease bike rentals.
3. Year (yr):
    - Contribution: The year variable has a positive relationship with bike demand (coefficient = 0.2308), indicating that, over time, the demand for bikes has increased. This suggests that awareness of shared bike services and their popularity have grown, contributing to higher bike demand as time progresses.

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 and one or more independent variables. The goal is to fit a line (or hyperplane in higher dimensions) that best predicts the target variable.

Key Concepts:

1. Equation: For simple linear regression with one predictor, the equation is:

    Y=β0+β1X+ϵY = \beta_0 + \beta_1X + \epsilonY=β0+β1X+ϵ

    Where:

    - YYY = Dependent variable (target)
    - XXX = Independent variable (predictor)
    - β0\beta_0β0 = Intercept
    - β1\beta_1β1 = Slope (coefficient)
    - ϵ\epsilonϵ = Error term

2. Goal: Minimize the difference between the observed values (YYY) and the predicted values (Y^\hat{Y}Y^) by adjusting the model parameters β0\beta_0β0 and β1\beta_1β1.
3. Fitting the Model: This is typically done using the Ordinary Least Squares (OLS) method, which minimizes the residual sum of squares (RSS) to find the optimal values for the coefficients.

4. Prediction: After training, the model can predict new values by applying the learned coefficients to the input features, using the equation:

   Y^=β0+β1X\hat{Y} = \beta_0 + \beta_1XY^=β0+β1X

Assumptions:

- Linearity: There should be a linear relationship between predictors and target.
- Independence: Residuals should be independent.
- Homoscedasticity: Constant variance of residuals.
- Normality: Residuals should be normally distributed.

Linear regression is widely used for predicting numerical outcomes, but its performance depends on meeting these assumptions.

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 a set of four datasets, each consisting of 11 pairs of values for two variables, X and Y. The datasets were created by statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before analyzing it, as well as to show how summary statistics (like mean, variance, and correlation) can be misleading.

Key Features of Anscombe's Quartet:

1. Identical Summary Statistics: Despite having very similar summary statistics (mean, variance, correlation, and regression lines), the datasets are visually very different from each other.
2. Purpose: The quartet highlights the fact that statistical measures like the mean, variance, and correlation do not always reveal the true underlying patterns or relationships in data. Visual inspection of data (e.g., using scatter plots) is essential for understanding the nature of the data.
3. Differences: The four datasets each have different distributions:
   - Dataset 1: Linear relationship with no outliers.
   - Dataset 2: Non-linear relationship, still with a quadratic pattern.
   - Dataset 3: A linear relationship but with a significant outlier that influences the results.
   - Dataset 4: A vertical line, showing a perfect correlation but no variation in the dependent variable.

Anscombe's Quartet emphasizes the importance of visualizing data to avoid misinterpretation based on summary statistics 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.

- Formula: r=∑(Xi−X¯)(Yi−Y¯)∑(Xi−X¯)2∑(Yi−Y¯)2r = \frac{\sum{(X_i - \bar{X})(Y_i - \bar{Y})}}{\sqrt{\sum{(X_i - \bar{X})^2} \sum{(Y_i - \bar{Y})^2}}}r=∑(Xi−X¯)2∑(Yi−Y¯)2∑(Xi −X¯)(Yi−Y¯) Where:
    - XiX_iXi and YiY_iYi are individual data points of the two variables.
    - X¯\bar{X}X¯ and Y¯\bar{Y}Y¯ are the means of the X and Y variables.

Interpretation:

- r = 1: Perfect positive linear relationship.
- r = -1: Perfect negative linear relationship.
- r = 0: No linear relationship.
- 0 < r < 1: Positive correlation (as one variable increases, the other tends to increase).
- -1 < r < 0: Negative correlation (as one variable increases, the other tends to decrease).

Pearson’s R measures only linear relationships and assumes that both variables are continuous, normally distributed, and have homoscedasticity.

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 is the process of transforming the features of a dataset so that they have specific properties, such as the same scale or range. It is typically done to ensure that all variables contribute equally to the model, especially when the features have different units or scales.

Why is Scaling Performed?

- Consistency: Models may perform poorly or behave erratically if variables with different units or scales are used together. For instance, in algorithms like k-NN, SVM, and gradient descent-based methods, features with larger scales can dominate the model's performance.
- Improved Convergence: Some algorithms (e.g., gradient descent) converge faster when features are scaled because it ensures that all features are treated equally in terms of magnitude.
- Interpretability: Scaling helps in comparing the importance of features when they have different units (e.g., height in meters and weight in kilograms).

Difference Between Normalized Scaling and Standardized Scaling:

- Normalized Scaling (Min-Max Scaling):
  - Rescales the data to a fixed range, usually [0, 1]. It is calculated using: Xscaled=X−min⁡(X)max⁡(X)−min⁡(X)X_{\text{scaled}} = \frac{X - \min(X)}{\max(X) - \min(X)}Xscaled=max(X)−min(X)X−min(X)
  - This method is sensitive to outliers and should be avoided if outliers are present.
- Standardized Scaling (Z-score Scaling):
  - Centers the data around the mean (0) and scales it by the standard deviation, making the data have a mean of 0 and a standard deviation of 1: Xscaled=X−μσX_{\text{scaled}} = \frac{X - \mu}{\sigma}Xscaled=σX−μ
  - This method is less sensitive to outliers and is preferred when data follows a Gaussian distribution or needs to be transformed to comparable units.

In summary, normalized scaling is used to scale data to a specific range, while standardized scaling centers the data and adjusts its spread to have a mean of 0 and a standard deviation of 1.

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)

Variance Inflation Factor (VIF) is used to measure the degree of multicollinearity in a regression model. It quantifies how much the variance of a regression coefficient is inflated due to correlation with other predictor variables.

Why Does VIF Become Infinite?

VIF becomes infinite when there is perfect multicollinearity between two or more predictor variables. This happens when one or more predictor variables are exactly or nearly linearly dependent on another. In such cases, the model cannot distinguish between the variables because their values are perfectly correlated.

- Perfect Multicollinearity: When two or more predictors are perfectly correlated (i.e., one variable is a linear combination of others), the determinant of the correlation matrix of the predictors becomes zero. This results in a division by zero in the VIF formula, leading to an infinite VIF value.

Formula for VIF:
VIF=11−R2\text{VIF} = \frac{1}{1 - R^2}VIF=1−R21

Where R2R^2R2 is the coefficient of determination obtained by regressing the predictor variable on all the other predictors. If R2=1R^2 = 1R2=1 (perfect correlation), the VIF will be infinite.

Conclusion: Infinite VIF occurs due to perfect multicollinearity, where predictor variables are highly correlated, making it impossible for the regression model to separate their individual effects on the dependent 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 plot (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It plots the quantiles of the observed data against the quantiles of the expected distribution (often normal distribution). If the data points in the Q-Q plot lie approximately along a straight line, it suggests that the data follows the expected distribution.

Use of Q-Q Plot:

- Assessing Normality: The primary use of a Q-Q plot is to check if the residuals (errors) of a regression model follow a normal distribution. Since linear regression assumes normally distributed residuals for valid statistical inference, a Q-Q plot helps to visually assess this assumption.
- Identifying Deviations: If the data points deviate significantly from the straight line, it indicates departures from normality, such as skewness, kurtosis, or the presence of outliers.

Importance of Q-Q Plot in Linear Regression:

- Assumption Checking: In linear regression, one of the key assumptions is that the residuals should be normally distributed. The Q-Q plot is crucial for visually verifying this assumption.
- Model Diagnostics: A Q-Q plot helps in identifying whether the model might be misspecified (e.g., non-linearity, outliers, or heteroscedasticity), allowing for improvements in the model.

Conclusion: A Q-Q plot is an essential diagnostic tool in linear regression for validating the normality of residuals, which is important for ensuring valid hypothesis tests and reliable predictions.