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)

Inference from categorical variables:

1. Season:

- Observation: 32% of bookings occurred in season3, followed by season2 (27%) and season4 (25%).
- Conclusion: Season appears to be a strong predictor of bike rentals, as the distribution shows that different seasons significantly impact the number of bookings. This feature should definitely be kept in the model.

2. Month (mnth):

- Observation: Months 5-9 showed higher bookings (10% each), with medians above 4000.
- Conclusion: Month also seems to play an important role, with specific months showing higher bookings. This indicates a seasonal pattern or trend, making month a valuable predictor for the model.

3. Weather (weathersit):

- Observation: 67% of bookings occurred during weathersit1, followed by weathersit2 (30%).
- Conclusion: Weather conditions have a noticeable impact on the bookings, with certain weather conditions likely influencing the number of rentals. This feature is a strong predictor and should be kept in the model.

4. Holiday:

- Observation: 97.6% of bookings occurred on non-holidays.
- Conclusion: Holiday is highly imbalanced, with most bookings happening on non-holidays. This creates a potential bias and suggests that holiday might not be a useful predictor for the model unless additional information or adjustments are made.

5. Weekday:

- Observation: Bookings were spread evenly across weekdays (13.5% - 14.8%), with similar medians.
- Conclusion: Weekday seems to have minimal influence on the number of rentals, with no strong patterns emerging across the days of the week. It could be kept in the model, but it may not be very predictive.

6. Working Day:

- Observation: 69% of bookings occurred on working days, with medians close to 5000.
- Conclusion: Working Day has a significant impact on the number of bookings, with more rentals occurring on working days. This suggests that working day is a strong predictor and should be included in the model.

Recommendations:

- Keep: Season, Month, Weather (weathersit), and Working Day as they show strong predictive potential.
- Consider Removing: Holiday, due to its high imbalance.

- Minimal Impact: Weekday may not add significant value, but it could still be included for completeness, or tested for feature selection during model evaluation.

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 the dummy variable trap by eliminating redundancy and multicollinearity. This ensures that the model's independent variables remain linearly independent.

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 pair plot shows that temp, atemp have highest correlation with target variable 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)

After constructing the model, we validated the assumptions of Linear Regression:

1. **Linearity**: We examined residual vs. fitted plots to ensure there were no discernible patterns.
2. **Normality**: We checked that the residuals followed a normal distribution using Q-Q plots.
3. **Homoscedasticity**: We confirmed that the residuals exhibited constant variance.
4. **Multicollinearity**: We assessed the Variance Inflation Factor (VIF) values, ensuring they were below 5.
5. **Error Independence**: We used the Durbin-Watson statistic, where a value near 2 indicates independence.

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)

As per our final Model, the top 3 predictor variables that influences the bike booking are:

- **Temperature (temp) .**
- **Weather Situation 3 (weathersit_3)**
- **Year (yr)**

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 models the relationship between a dependent variable and one or more independent variables using a linear equation:

Y=β0+β1X1+⋯+βnXn+ϵY = \beta_0 + \beta_1X_1 + \dots + \beta_nX_n + \epsilonY=β0+β1X1+⋯+βnXn+ϵ

- YYY is the dependent variable, X1,X2,…X_1, X_2, \dotsX1,X2,… are the independent variables, and β0,β1,…\beta_0, \beta_1, \dotsβ0,β1,… are the model coefficients.
- The goal is to minimize the sum of squared errors (SSE) between the observed and predicted values.

The coefficients are typically found using Ordinary Least Squares (OLS). After training, the model can predict YYY for new inputs.

Key Assumptions:

1. **Linearity**: The relationship is linear.
2. **Independence**: Residuals are independent.
3. **Homoscedasticity**: Constant variance of residuals.
4. **Normality**: Residuals are normally distributed.

Linear regression can be simple (one predictor) or multiple (multiple predictors).

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

Anscombe's Quartet consists of four datasets with nearly identical summary statistics (mean, variance, correlation, and regression line) but vastly different distributions. Purpose:

- Highlights the importance of data visualization to avoid misleading conclusions from statistical measures alone.

Key Insights:

1. Datasets have the same mean and variance for xxx and yyy.
2. Correlation and linear regression are identical.

3. Visualizing the data reveals distinct patterns (e.g., outliers, nonlinear trends).

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's R, also known as the Pearson correlation coefficient, measures the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1:

- 1 indicates a perfect positive correlation,
- -1 indicates a perfect negative correlation,
- 0 indicates no linear correlation.

It is used to understand how closely related two variables are.

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 transforming the features of a dataset so that they are on a comparable scale. This ensures that all features contribute equally to the analysis or machine learning model, especially in algorithms sensitive to the magnitude of data.

Why is Scaling Performed?

Scaling is performed for several reasons:

1. **Ensure Equal Contribution**: Some algorithms, like distance-based models (e.g., k-NN, SVM), are sensitive to the magnitude of the data. Without scaling, features with larger values can dominate the model, leading to biased results.

2. **Convergence in Optimization**: Algorithms like gradient descent in linear regression and neural networks may converge faster when features are scaled, since the gradients are more balanced.

3. **Improve Model Performance**: Scaling can improve the performance of certain machine learning algorithms (e.g., k-means clustering, PCA) that rely on the distances between data points.

Difference Between Normalized Scaling and Standardized Scaling

Normalized Scaling (Min-Max Scaling)

- **Definition**: Normalization rescales the feature to a fixed range, typically [0, 1]. This is done by subtracting the minimum value of the feature and dividing by the range (difference between maximum and minimum values).

Xnormalized=X−XminXmax−XminX_{\text{normalized}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}Xnormalized=Xmax−XminX−Xmin

- **When to Use**: Normalization is useful when the data has a known range or when you need to bound the features within a fixed range for algorithms that require it (e.g., neural networks, k-NN).

- **Effect**: It changes the scale of the data while preserving the shape of the original distribution.

Standardized Scaling (Z-Score Scaling)

- **Definition**: Standardization transforms the data by subtracting the mean and dividing by the standard deviation of the feature. This scales the data to have a **mean of 0** and a **standard deviation of 1**.

Xstandardized=X−μσX_{\text{standardized}} = \frac{X - \mu}{\sigma}Xstandardized=σX−μ

where μ\muμ is the mean and σ\sigmaσ is the standard deviation.

- **When to Use**: Standardization is typically used when the data does not have a fixed range or when you need the data to have a normal distribution (for example, when applying algorithms that assume normally distributed data, like linear regression).

- **Effect**: It changes the scale of the data and is useful for algorithms sensitive to data spread, such as linear regression and PCA.

Summary of Differences:

Property	**Normalized Scaling**	**Standardized Scaling**
Formula	Xnormalized=X−XminXmax−XminX_{\text{normalized}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}Xnormalized=Xmax−XminX−Xmin	Xstandardized=X−μσX_{\text{standardized}} = \frac{X - \mu}{\sigma}Xstandardized=σX−μ
Result	Rescales features to a specific range [0, 1]	Rescales features to have mean = 0, standard deviation = 1
When to Use	When features have known min and max values or require bounding	When data is normally distributed or has no fixed range
Effect on Data	Preserves the shape, but not the spread	Shifts data distribution and rescales the spread

Conclusion:

- **Normalization** is used to scale data within a fixed range [0, 1], while **Standardization** is used to center the data with a mean of 0 and a standard deviation of 1.
- The choice between normalization and standardization depends on the nature of the data and the machine learning algorithm being used.

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 be infinite if there is perfect multicollinearity between two or more independent variables. This means that one feature can be exactly predicted by a linear combination of other features in the model. When this happens**:**

1. Determinant of the correlation matrix becomes zero, leading to a division by zero in the VIF formula.
2. Perfect correlation between variables makes the model unstable, as the variance of the regression coefficients increases infinitely.

In practice, this signals that one or more predictors are redundant and should be removed from the model to avoid collinearity issues.

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 (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a specific theoretical distribution, typically the normal distribution. In the plot, the quantiles of the dataset are plotted against the quantiles of the normal distribution. If the data follows a normal distribution, the points on the plot will lie approximately on a straight line.

Use and Importance in Linear Regression:

- **Normality Check**: In linear regression, residuals (errors) should ideally follow a normal distribution. A Q-Q plot helps assess this assumption by visually checking if the residuals align with the normal distribution.
- **Model Validation**: If the residuals deviate significantly from the straight line, it may indicate non-normality, which could lead to invalid conclusions in regression analysis. This helps in diagnosing potential problems in the model.