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)

Season: The highest bike usage was observed during the fall season, while the spring season had the lowest number of users.

Month: Bike bookings peaked between May and October, with a steady rise from the beginning of the year, reaching a peak in mid-year, and then declining towards the year's end.

Weekday: Sunday had the lowest number of users, whereas the demand increased progressively from Monday to Friday.

Weather Condition: People preferred riding bikes the most when the weather was clear.

Working Day: The number of bookings remained almost the same on both working and non-working days.

Holiday: More bikes were rented on non-holidays compared to holidays.

Year: Bike rentals saw a higher demand in 2019 compared to the previous year.

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)

If we create dummy variables for a categorical feature with k unique categories, we get k binary columns. However, one of these columns can be predicted from the others, leading to redundant information in the model.

This issue is called the dummy variable trap, which can cause multicollinearity in regression models, making it difficult for the model to estimate coefficients accurately**.**

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)

Temperature Variable has the highest correlation

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)

Validation of Linear Regression Assumptions

After building the Linear Regression model on the training set, it is essential to validate its assumptions to ensure reliability and accuracy. The following key assumptions were tested:

1. Linearity

- **Objective:** To check if there is a linear relationship between independent and dependent variables.
- **Method:** A scatter plot of **predicted values vs. actual values** was examined.

- **Validation:** The presence of a linear pattern indicates that the assumption holds.

2. No Multicollinearity

- **Objective:** To ensure independent variables are not highly correlated with each other.
- **Method:** Variance Inflation Factor (**VIF**) was calculated for each independent variable.
- **Validation:** VIF values below **5** indicate that multicollinearity is not a concern.

3. Homoscedasticity

- **Objective:** To check if the variance of residuals remains constant across all levels of independent variables.
- **Method:** A **Residuals vs. Fitted Values** plot was analyzed.
- **Validation:** Randomly scattered residuals without a distinct pattern confirm homoscedasticity.

4. Normality of Residuals

- **Objective:** To ensure residuals follow a normal distribution.
- **Method:** A **Q-Q Plot** or **Shapiro-Wilk Test** was conducted.
- **Validation:** If the Q-Q plot shows points lying along a **45-degree line**, the residuals are normally distributed.

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
Hum
workingday

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)

1. Introduction to Linear Regression

Linear Regression is a **supervised learning algorithm** used for predicting continuous values based on independent variables. It assumes a linear relationship between the dependent variable (**Y**) and one or more independent variables (**X**).

2. Mathematical Representation

For **Simple Linear Regression**, the relationship between X and Y is represented as:

Y=β0+β1X+ϵY = \beta_0 + \beta_1 X + \epsilon

Where:

- YY is the dependent variable (target/output)
- XX is the independent variable (predictor/input)
- β0\beta_0 is the intercept (constant term)
- β1\beta_1 is the coefficient (slope of the line)
- ϵ\epsilon is the error term

For **Multiple Linear Regression**, the equation extends to:

Y=β0+β1X1+β2X2+...+βnXn+ϵY = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon

3. Steps Involved in Linear Regression

1. **Data Collection:** Gather historical data containing independent and dependent variables.
2. **Data Preprocessing:** Handle missing values, remove outliers, and perform feature scaling if necessary.
3. **Splitting Data:** Divide the dataset into **training** and **testing** sets.
4. **Model Training:** Use **Ordinary Least Squares (OLS)** to estimate coefficients that minimize the sum of squared errors.
5. **Model Evaluation:** Assess model performance using metrics such as **R-squared**, **Mean Squared Error (MSE)**, and **Root Mean Squared Error (RMSE)**.
6. **Prediction:** Use the trained model to predict values for new inputs.

4. Assumptions of Linear Regression

- **Linearity:** The relationship between X and Y should be linear.
- **No Multicollinearity:** Independent variables should not be highly correlated.
- **Homoscedasticity:** The variance of residuals should remain constant.
- **Normality of Residuals:** Residuals should be normally distributed.
- **No Autocorrelation:** Residuals should be independent over time.

5. Advantages & Disadvantages

Advantages:
✓ Simple to implement and interpret
✓ Computationally efficient
✓ Works well for linearly related data

Disadvantages:
✗ Sensitive to outliers
✗ Assumes linearity, which may not always hold
✗ Prone to overfitting if too many independent variables are included

By understanding these concepts, Linear Regression can be effectively used in predictive modeling and statistical analysis.

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

1. Introduction to Anscombe's Quartet

Anscombe's Quartet is a set of **four datasets** that have nearly identical **statistical properties** but differ significantly in their **graphical representations**. It was created by **Francis Anscombe** in 1973 to demonstrate the importance of **visualizing data** rather than relying solely on summary statistics.

2. Statistical Similarities

Each dataset in the quartet has approximately:

- The same **mean** for both X and Y variables.
- The same **variance** for both X and Y variables.
- The same **correlation coefficient** (~0.816).
- The same **linear regression equation** (Y = 3 + 0.5X).

Despite these identical statistics, the datasets are structurally different when visualized.

3. The Four Datasets

1. **Dataset 1:** A standard linear relationship where a linear regression model is appropriate.
2. **Dataset 2:** A **non-linear** relationship where a quadratic model fits better than a linear one.
3. **Dataset 3:** A case where an **outlier** greatly affects the regression line.
4. **Dataset 4:** A scenario where all data points are nearly the same except for one extreme outlier, making the regression unreliable.

4. Importance of Anscombe's Quartet

- Highlights the need for **data visualization** before drawing conclusions.
- Demonstrates that **summary statistics alone can be misleading**.
- Encourages the use of **scatter plots** to better understand data patterns.
- Shows the **impact of outliers** on regression models.

By studying Anscombe's Quartet, we learn that statistical summaries should be complemented with graphical analysis for better decision-making.

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 (Pearson Correlation Coefficient)

1. Definition

Pearson's R, also known as the **Pearson correlation coefficient (PCC)**, measures the **linear relationship** between two variables. It quantifies the degree to which changes in one variable predict changes in another.

2. 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} \sqrt{\sum (Y_i - \bar{Y})^2}}

Where:

- rr = Pearson's correlation coefficient
- Xi,YiX_i, Y_i = Individual data points of variables X and Y
- X¯,Y¯\bar{X}, \bar{Y} = Mean of X and Y

3. Interpretation

- **r = 1** → Perfect **positive correlation** (as X increases, Y increases)
- **r = -1** → Perfect **negative correlation** (as X increases, Y decreases)
- **r = 0** → No correlation (no linear relationship)
- **0 < r < 1** → Weak to strong **positive correlation**
- **-1 < r < 0** → Weak to strong **negative correlation**

4. Assumptions of Pearson's R

- The relationship between variables is **linear**.
- Variables are **normally distributed**.
- Data points are **independent**.
- There are **no significant outliers**.

5. Use Cases

- Understanding relationships in datasets.
- Feature selection in machine learning.
- Analyzing dependencies in finance, healthcare, and research.

Pearson's R is a valuable statistical tool for assessing correlations but should always be used with **scatter plots** to confirm linearity.

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 in Machine Learning

1. What is Scaling?

Scaling is a preprocessing technique used to **transform numerical features** into a specific range, ensuring that all features contribute equally to a model. It is essential when dealing with machine learning algorithms that rely on distance-based calculations, such as **KNN, SVM, and Gradient Descent-based models**.

2. Why is Scaling Performed?

- Prevents features with **larger magnitudes** from dominating the model.
- Speeds up **convergence in optimization algorithms** like Gradient Descent.
- Improves the **performance and accuracy** of models.
- Essential for **distance-based algorithms** like K-Means and KNN.

3. Normalized Scaling vs. Standardized Scaling

Feature	**Normalization (Min-Max Scaling)**	**Standardization (Z-score Scaling)**
Definition	Rescales features to a fixed range (0 to 1 or -1 to 1).	Centers data around mean (0) with unit variance (1).
Formula	X′=X−XminXmax−XminX' = \frac{X - X_{min}}{X_{max} - X_{min}}	X′=X−μσX' = \frac{X - \mu}{\sigma}
When to Use?	When data distribution is **not Gaussian** or has outliers.	When data follows a **Gaussian distribution**.
Effect on Outliers	**Sensitive to outliers** as it rescales values based on min/max.	**Less sensitive to outliers** since it uses mean and standard deviation.

4. Conclusion

- **Normalization** is useful when features have **different scales** and need to be compared directly.
- **Standardization** is preferred for **normal distribution**-based models.
- The choice depends on the **dataset and 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)

1. Understanding VIF (Variance Inflation Factor)

Variance Inflation Factor (VIF) measures the **degree of multicollinearity** among independent variables in a regression model. A high VIF indicates strong correlation between predictors, which can negatively impact the model's accuracy.

2. When Does VIF Become Infinite?

VIF becomes **infinite** when there is **perfect multicollinearity**, meaning:

- One predictor variable is an **exact linear combination** of one or more other predictors.
- There is **redundant information** in the dataset due to duplicate or highly correlated features.
- The **determinant of the correlation matrix** is **zero**, making matrix inversion impossible during VIF calculation.

3. Causes of Infinite VIF

- **Perfect correlation (r = ±1) between two or more independent variables**.
- **Including dummy variables with perfect multicollinearity** (e.g., not using drop_first=True when creating dummies).

- **Redundant features** (e.g., including both temperature in Celsius and Fahrenheit in the same model).

4. Solutions to Avoid Infinite VIF

- **Remove one of the highly correlated variables**.
- **Use Principal Component Analysis (PCA) or Feature Selection techniques**.
- **Ensure proper encoding of categorical variables**.

Infinite VIF is a strong indicator of multicollinearity issues, which should be addressed before building a reliable regression model.

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)

Q-Q Plot: Definition, Use, and Importance in Linear Regression

1. What is a Q-Q Plot?

A **Q-Q (Quantile-Quantile) plot** is a graphical tool used to compare the **distribution of a dataset** to a theoretical distribution (e.g., normal distribution). It helps assess whether a variable follows a specific distribution by plotting the quantiles of the sample data against the quantiles of the theoretical distribution.

2. How to Interpret a Q-Q Plot?

- **Straight diagonal line:** Data follows the theoretical distribution.
- **Upward or downward curve:** Indicates skewness in the data.
- **S-shaped curve:** Suggests heavier or lighter tails compared to the theoretical distribution.
- **Outliers deviating from the line:** Indicate extreme values that may impact analysis.

3. Importance of a Q-Q Plot in Linear Regression

- **Checks Normality of Residuals:** Linear regression assumes that residuals are normally distributed. A Q-Q plot helps verify this assumption.
- **Detects Skewness and Kurtosis:** Helps identify whether residuals exhibit non-normal behavior.
- **Identifies Outliers:** Points deviating from the line indicate potential outliers that may affect model performance.
- **Validates Model Assumptions:** Ensures that errors are randomly distributed, which is crucial for unbiased coefficient estimation.

4. Conclusion

A Q-Q plot is an essential diagnostic tool in linear regression, ensuring that the normality assumption holds for better model reliability and accurate statistical inference.