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)

- Categorical variables need to be converted into numerical representations (e.g., dummy variables or one-hot encoding) before being used in regression models. This ensures that the model can handle them as part of the feature set.

Based on the analysis:

- You can infer which categories have the most or least impact on the dependent variable.
- This information helps in decision-making, such as targeting specific categories or adjusting inputs to optimize outcomes

---

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 during dummy variable creation in pandas is important to avoid a problem called multicollinearity in regression models.

When we using the drop_first=True during dummy variable creation:

1. Avoids the dummy variable trap and prevents multicollinearity.
2. Ensures one category serves as a reference for interpretation.
3. Reduces the number of features, improving efficiency.

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)

By looking into the Pair plot of numerical variable, w.r.t target variable **atemp** and **temp** variable are having 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)

1. Relationship between the Independent and Dependent variables are Linear.because of to plot the residuals against the predicted values.
2. The Residuals(errors) are normally distributed.
3. Errors terms are Independent of each other.
4. Errors terms have constant variance.

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, features contributing significantly to the demand of bikes are, Hum, weathersite, temp and atemp.

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 assumes that there is a linear relationship between the dependent variable (Y) and the independent variable(s) (X). The goal is to fit a straight line (in the case of one independent variable) or a hyperplane (in the case of multiple independent variables) that best describes the relationship.

1.Equation of Linear Regression:

Simple Linear Regression (one independent variable):

- $Y=\beta 0+\beta 1X+\epsilon Y$
  - Y : Dependent variable (target)
  - X : Independent variable (predictor)
  - $\beta 0$: Intercept (value of YYY when $X=0X = 0X=0$)
  - $\beta 1$: Slope (rate of change of YYY with respect to XXX)
  - ϵ : Error term (residuals, representing the difference between the observed and predicted values of Y).

Multiple Linear Regression (multiple independent variables):

- $Y=\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta nXn+\epsilon Y$

2. Objective of Linear Regression

The objective of linear regression is to find the best-fitting line (or hyperplane) by minimizing the error between the actual values and the predicted values.

This is done using the **Least Squares Method**, which minimizes the **Sum of Squared Errors (SSE)**:

$SSE=\sum i=1n(Yi-Y\hat{}i)2$

Where:

- Yi : Actual value of the dependent variable
- Y^ : Predicted value of the dependent variable

3. Steps of the Linear Regression Algorithm

Step 1: Hypothesis Function

The algorithm assumes a linear relationship between X and Y:

$Y\hat{}=\beta 0+\beta 1X1+\beta 2X2+\cdots+\beta nXn$

Step 2: Loss Function

To measure how well the model fits the data, we use a loss function. The most common loss function is the Mean Squared Error (MSE):

$MSE=1n\sum i=1n(Yi-Y\hat{}i)2$

Step 3: Optimization

Using Gradient Descent or a closed-form solution, the algorithm minimizes the MSE to find the optimal values of $\beta 0,\beta 1,\ldots,\beta n$.

Gradient Descent:

The coefficients are updated iteratively:

$\beta j:=\beta j-\alpha*\partial/\partial*\beta j\ MSE$

Where:

- α : Learning rate (controls the step size in each iteration)

- ∂/∂βj*MSE : Gradient (partial derivative of the loss function)

---

4. Key Assumptions of Linear Regression

To ensure the reliability of the model, linear regression makes the following assumptions:

1. **Linearity**: The relationship between X and Y is linear.
2. **Independence**: The observations are independent of each other.
3. **Homoscedasticity**: The variance of residuals is constant across all levels of X.
4. **Normality of Residuals**: The residuals are normally distributed.
5. **No Multicollinearity**: The independent variables are not highly correlated.

---

5. Metrics for Model Evaluation

Once the model is trained, evaluate its performance using these metrics:

R-squared (R2**)**:

- Measures the proportion of variance in Y explained by X.
- R2=1−SSE/STR, where SST is the total sum of squares.
- R2 ranges from 0 to 1. A higher value indicates a better fit.

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 collection of four datasets that have nearly identical statistical properties but differ significantly in their visual representation. It was created by statistician Francis Anscombe in 1973 to highlight the importance of visualizing data before drawing conclusions solely based on statistical measures.

Key Characteristics of Anscombe's Quartet

Each dataset in the quartet contains 11 pairs of (x,y)values and has the following similar statistical properties:

1. **Mean of** x**:** The mean of x is the same (9.0) across all four datasets.
2. **Mean of** y**:** The mean of y is also the same (7.5) across all datasets.
3. **Variance of** x**:** The variance of x is the same for all datasets.
4. **Variance of** y**:** The variance of y is similar across datasets.
5. **Correlation (**r**) between** x **and** y**:** The correlation is approximately 0.816 for all datasets.
6. **Linear regression line:** All datasets yield nearly the same regression line: y=3+0.5x

Purpose of Anscombe's Quartet

Highlight the Importance of Data Visualization:

1. Relying solely on statistical summaries (mean, variance, correlation, etc.) can be misleading.
2. Graphical representation reveals the true nature of the data.

Outliers' Impact:

1. Outliers can have a significant effect on the regression line and statistical metrics.

Limitations of Statistical Measures:

1. Statistical metrics like R2, mean, and correlation can hide critical differences in datasets.

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 a linear relationship between two continuous variables. It is widely used in statistics and data analysis to determine how closely two variables are linearly related.

The formula to calculate Pearson's R is:

r=(∑i=1n(xi−x¯)(yi−y¯))/(∑i=1n(xi−x¯)2∑i=1n(yi−y¯)2)

Where:

- Xi and yi : Individual data points of variables X and Y.
- x¯ and y¯ : Mean of X and Y.
- n: Number of data points.

This formula can also be interpreted as the **covariance between** X **and** Y divided by the product of their standard deviations.

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 a data preprocessing technique used to adjust the range of features in a dataset so that they have comparable magnitudes. This is crucial when building machine learning models, as many algorithms are sensitive to the scale of input data and perform poorly if features have widely varying ranges.

There are two common types of scaling:

1. **Normalization (Min-Max Scaling)**.
2. **Standardization (Z-Score Scaling)**.

---

1. Normalized Scaling (Min-Max Scaling)

Definition: Normalization rescales the features to fit within a specific range, typically **[0, 1]**.

Formula:

- $Xscaled=(X-Xmin)/(Xmax-XminX)$_

Where:

- X = Original feature value.
- Xmin = Minimum value in the feature.
- Xmax = Maximum value in the feature.

When to Use:

- When the distribution of the data is **not Gaussian** (not bell-shaped).
- When you need all features in a fixed range (e.g., neural networks).

---

2. Standardized Scaling (Z-Score Scaling)

Definition: Standardization transforms the data to have a **mean of 0** and a **standard deviation of 1**.

Formula:

- $Xscaled=(X-\mu)/\sigma$

Where:

- X = Original feature value.
- µ = Mean of the feature.
- σ = Standard deviation of the feature.

When to Use:

- When the data follows a Gaussian (normal) distribution.
- When scaling is needed for algorithms like PCA or linear regression.

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)

The Variance Inflation Factor for a predictor Xi measures how much the variance of its regression coefficient βi is inflated due to multicollinearity (correlation with other predictors).

The formula for VIF:

1. VIFi=1/(1−Ri2

Where Ri2 is the coefficient of determination for the regression of Xi on all other predictors.

Perfect Multicollinearity:

1. If a predictor Xi is perfectly explained (100%) by a combination of other predictors, the Ri2 value for that predictor will be **1**.
2. Substituting Ri2=1 into the VIF formula: VIFi=1(1−1)=∞ This happens because the denominator becomes zero, leading to an infinite VIF value.

Perfect Multicollinearity Occurs When:

1. One or more predictor variables are **exactly linear combinations** of others.
2. This often arises when dummy variables are incorrectly created, such as not using `drop_first=True` when encoding categorical variables.

Why is Infinite VIF Problematic?

- High or infinite VIF values indicate that the coefficients in the regression model are unstable and highly sensitive to small changes in the data.
- It makes the interpretation of the model coefficients unreliable because the multicollinearity inflates their standard errors.

How to Handle Infinite VIF?

Check for Redundant Variables:

- Identify predictors that are perfect linear combinations of others.
- For example, if you have dummy variables for categories, drop one category to avoid redundancy.

Use `drop_first=True` When Creating Dummy Variables:

- While encoding categorical variables, ensure that one level is dropped to prevent the "dummy variable trap."

Remove Highly Correlated Predictors:

- Calculate the correlation matrix of predictors.
- If two variables have a correlation close to ±1 , consider removing one of them.

Regularization Techniques:

- Use techniques like **Ridge Regression** (L2 regularization) to handle multicollinearity instead of directly dropping variables.

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 (Quantile-Quantile) plot** is a graphical tool used to compare the distribution of a dataset with a theoretical probability distribution, most commonly the **normal distribution**. It helps to assess whether the data follows the expected theoretical distribution.

- The Q-Q plot plots the **quantiles** of the observed data (sample quantiles) on the Y-axis against the quantiles of a theoretical distribution (e.g., normal quantiles) on the X-axis.
- If the data follows the theoretical distribution, the points will approximately align on a **straight 45-degree line**.

Use and Importance of a Q-Q Plot in Linear Regression

The Q-Q plot is particularly important in linear regression because it helps validate the assumptions underlying the model. Linear regression assumes:

Normality of Residuals:

1. One key assumption in linear regression is that the residuals (differences between observed and predicted values) are **normally distributed**.
2. A Q-Q plot of the residuals can be used to check this assumption. If the residuals are normally distributed, the points on the Q-Q plot will fall approximately along the straight line.

Detecting Deviations from Normality:

1. **Heavy tails**: If the points curve away from the line at the ends, it indicates the presence of outliers or heavy tails in the residual distribution.
2. **Skewness**: If the points consistently deviate above or below the line, the residuals may be skewed.

Model Validation:

1. The Q-Q plot helps determine whether the linear regression model's assumptions are valid. If the residuals deviate significantly from normality, it could mean that:

    1. The model is misspecified (e.g., important variables are missing or nonlinear relationships exist).
    2. The data needs transformation to meet the normality assumption.