Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

From the analysis, it is clear that the bike rentals are positively depended on categorical variables summer because of the convenience of weather in this season also it depends on positively on days like Sunday, Monday and Tuesday and also on June month

The count is negatively dependent on inconvenient weather conditions like humidity, windspeed, snow, cold month December etc

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Ans: If drop_first=True is not used, all the categories are considered while creating dummy variables. Which causes high collinearity between features and high VIF (Variance Inflation factor)

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans: Registered. Correlation value of 0.95

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans:

1. From the plot of train and test error terms, it is clear that error terms are normally distributed
2. There is no multicollinearity since VIF is < 5

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans: Registered, temperature and monday

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans:

Linear Regression is a statistical method used for modeling the relationship between a dependent variable (target) and one or more independent variables (features). The primary goal is to find the linear equation that best predicts the target variable based on the features.

Types of Linear Regression

1. **Simple Linear Regression:** Involves one independent variable. The relationship is modeled with a straight line.

   y=β0+β1x+ϵy

Where:

    - y -> the dependent variable.
    - x -> the independent variable.
    - β0 is the intercept (constant term).
    - β1 is the coefficient (slope) representing the effect of x on y.
    - ϵ is the error term.

2. **Multiple Linear Regression:** Involves multiple independent variables. The equation becomes:

   y=β0+β1x1+β2x2+...+βnxn+ϵy

   Where x1,x2,...,xnx_1, x_2, ..., x_nx1,x2,...,xn are the independent variables.

Assumptions of Linear Regression

For linear regression to provide valid results, several assumptions must be met:

1. **Linearity:** The relationship between the dependent and independent variables is linear.
2. **Independence:** The residuals (errors) are independent of each other.
3. **Homoscedasticity:** The residuals have constant variance at all levels of the independent variables.
4. **Normality:** The residuals of the model are normally distributed.
5. **No multicollinearity:** Independent variables should not be highly correlated with each other.

Steps in the Linear Regression Algorithm

1. **Data Collection:** Gather the data that contains the dependent and independent variables.
2. **Data Preprocessing:**
    - Handle missing values and outliers.
    - Encode categorical variables if needed.
    - Scale or normalize features if necessary.
3. **Splitting the Dataset:** Split the dataset into training and testing sets to evaluate the model's performance.
4. **Model Fitting:** Use a method (e.g., Ordinary Least Squares) to find the best-fitting line by minimizing the sum of squared residuals (the difference between observed and predicted values).
    - The coefficients (β\betaβ) are estimated using the formula: β=(XTX)−1XTy\beta = (X^T X)^{-1} X^T yβ=(XTX)−1XTy Where XXX is the matrix of features, and yyy is the vector of target values.
5. **Model Evaluation:**
    - **R-squared:** Measures how well the independent variables explain the variance in the dependent variable.

  - **Adjusted R-squared:** Adjusts R-squared for the number of predictors, useful for multiple regression.
  - **Mean Squared Error (MSE):** Measures the average of the squares of the errors.
  - **Residual analysis:** Evaluate the residuals to check assumptions (linearity, homoscedasticity, etc.).

6. **Making Predictions:** Use the fitted model to predict the dependent variable for new data.

Advantages of Linear Regression

- **Simplicity:** Easy to understand and implement.
- **Interpretability:** The coefficients provide clear insight into the relationships between variables.
- **Efficiency:** Computationally efficient for large datasets.

Disadvantages of Linear Regression

- **Linearity Assumption:** The algorithm assumes a linear relationship, which may not hold in all cases.
- **Sensitivity to Outliers:** Outliers can disproportionately influence the model.
- **Multicollinearity:** If independent variables are highly correlated, it can lead to unreliable coefficient estimates.

2. Explain the Anscombe’s quartet in detail. (3 marks)

Anscombe’s Quartet is a collection of four datasets that were created by the statistician Francis Anscombe in 1973. The quartet is famous for illustrating the importance of graphical representation in statistics. Despite having nearly identical statistical properties (such as mean, variance, and correlation), each dataset has very different distributions and relationships when visualized. This demonstrates how relying solely on statistical measures can be misleading.

The Datasets

The four datasets in Anscombe's Quartet are:

1. Dataset I
2. Dataset II
3. Dataset III
4. Dataset IV

Each dataset consists of 11 pairs of x and y values.

Statistical Properties

Despite having the same mean xxx, mean yyy, variance of xxx, variance of yyy, and correlation between xxx and yyy, the datasets differ greatly in their distributions and relationships:

- **Mean of xxx:** 9
- **Mean of yyy:** 7.5
- **Variance of xxx:** 11
- **Variance of yyy:** 4.12
- **Correlation (Pearson's r):** Approximately 0.82 for all datasets.

Visualizing the Datasets

The power of Anscombe's Quartet lies in the visual differences among the datasets. Each dataset can be plotted to reveal the underlying patterns:

1. **Dataset I:** Linear relationship with no outliers.
2. **Dataset II:** Non-linear relationship (quadratic).
3. **Dataset III:** Non-linear relationship with one outlier that heavily influences statistics.
4. **Dataset IV:** Linear relationship with a clear outlier that doesn't affect the linearity.

3. What is Pearson's R? (3 marks)

Ans:

Pearson's R, also known as Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is a widely used method for understanding the degree to which two variables are related.

Interpretation of Pearson's R

- **Range:** The value of Pearson's R ranges from -1 to +1.
  - **+1:** Perfect positive linear correlation. As one variable increases, the other variable also increases.
  - **0:** No correlation. There is no linear relationship between the two variables.
  - **-1:** Perfect negative linear correlation. As one variable increases, the other variable decreases.

Strength of the Correlation

The strength of the correlation can be interpreted as follows:

- **0 to ±0.1:** No or negligible correlation
- **±0.1 to ±0.3:** Weak correlation
- **±0.3 to ±0.5:** Moderate correlation
- **±0.5 to ±0.7:** Strong correlation
- **±0.7 to ±1.0:** Very strong correlation

Assumptions of Pearson's R

For Pearson's R to be valid, certain assumptions should be met:

1. **Linearity:** The relationship between the two variables should be linear.
2. **Normality:** Both variables should be approximately normally distributed, especially for small sample sizes.
3. **Homoscedasticity:** The variance of one variable should be similar across the range of the other variable.
4. **Continuous Scale:** Both variables should be measured on a continuous scale (interval or ratio)

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans: Scaling is process of converting all the features to similar scale. It improves the performance of the model since it ensures all the features contribute equally to the distance calculation.

1. **Normalized scaling = (X – Xmin)/(Xmax – Xmin)**
   Standardized scaling = (X−μ)/σ
2. **Range of scaled values for normalized scaling is [0,1]**
   Range of scaled values for standardized scaling Mean=0, SD=1
3. **Sensitivity to Outliers is high in normalized scaling and low in Standardized scaling.**
4. **Easier to interpret in normalized scaling and it is less intuitive in standaridized scaling.**

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans:

VIF is used to quantify how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors. If its value is high, it indicates that

1. **There is high collinearity between features. VIF is infinite when the collinearity between two feature is perfect and the model can't distinguish between two variables.**
2. **If there are duplicate features in the data.**
3. **If two features too similar Eg temp and atemp**
4. **If all categories are considered while creating dummy variables.**

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans:

A **Q-Q plot** (Quantile-Quantile plot) is a graphical tool used to compare the quantiles of a dataset against the quantiles of a theoretical distribution, often the normal distribution. It helps in assessing whether a dataset follows a specific distribution.

How Q-Q Plots Work

1. **Quantiles:** A quantile divides a dataset into intervals of equal probability. For example, the median is the 50th percentile.
2. **Plotting:** In a Q-Q plot:
   - The x-axis represents the quantiles of the theoretical distribution (e.g., standard normal distribution).
   - The y-axis represents the quantiles of the observed data.
3. **Interpretation:**
   - **Straight Line:** If the points lie approximately on a straight diagonal line (y = x), it suggests that the data follows the theoretical distribution (often normal).
   - **Deviation from Line:** If points deviate from the line, it indicates that the data does not follow the theoretical distribution. For example:
     - An S-shaped curve could indicate heavy tails.
     - A concave or convex shape may suggest a different distribution.

Importance of Q-Q Plots in Linear Regression

In the context of linear regression, Q-Q plots are essential for validating the assumptions underlying the regression model:

1. **Normality of Residuals:**
   - One of the key assumptions of linear regression is that the residuals (the differences between observed and predicted values) should be normally distributed.
   - A Q-Q plot can be used to visually assess the normality of residuals. If the residuals form a straight line in the Q-Q plot, the normality assumption holds. If not, it suggests that the residuals may be skewed or exhibit heavy tails.
2. **Detection of Outliers:**
   - Q-Q plots can help identify outliers that may influence the regression model. Outliers will appear far from the straight line, indicating that they deviate from the expected distribution.
3. **Model Diagnostics:**
   - By evaluating the normality of residuals through Q-Q plots, you can determine if you need to transform your dependent variable or use a different modeling approach. If the residuals are not normally distributed, it may lead to biased estimates and invalid statistical inferences.