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)

   Ans:

   Seasonal Impact on Bike Rentals:

   The mean bike rental count is higher during **summer** and **fall**. Rental count gradually increases from January to July, and it decreases steadily starting August till December.

   Weather Conditions and Bike Rentals:

   Higher bike rental counts are observed under below weather conditions.

   - 1: Clear, Few clouds, partly cloudy, Partly cloudy
   - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

   Holiday:

   Rental count is high during non-holiday.

   Year:

   Rental count is high in 2019.

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

   Ans:

   Please find the reason below for removing 1st column during dummy variable creation.

   Avoiding multicollinearity:

   - Multicollinearity occurs when one predictor variable in a model can be linearly predicted from the others with a substantial degree of accuracy. This can cause problems in regression models, making it difficult to determine the effect of each predictor.

   - When you create dummy variables for a categorical feature with (n) categories, you end up with (n) binary variables. However, these (n) variables are not independent because they sum to 1. This introduces perfect multicollinearity.

   - By setting drop_first=True, you drop the first category and create (n-1) dummy variables instead of (n). This reduces redundancy and avoids multicollinearity.

   Interpretability:

   The coefficients of the remaining dummy variables can be interpreted relative to the dropped category. This makes the model easier to interpret.

   By dropping the first category, you ensure that the dummy variables are independent, and the model remains stable and interpretable.

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

Ans:

By looking into Pair plot temperature has the highest co relation with target variable bike share count.
(Ignoring casual and registered user are they sum up to rental count)

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

Ans:

Assumptions of Linear Regression
Linearity: The relationship between the independent and dependent variables is linear.
Independence: The residuals (errors) are independent.
Homoscedasticity: The residuals have constant variance.
Normality: The residuals are normally distributed.
No Multicollinearity: The independent variables are not highly correlated with each other.

Common metrics for evaluating the performance of a linear regression model include:

R-squared ((R^2)): Measures the proportion of the variance in the dependent variable that is predictable from the independent variables. Values range from 0 to 1, with higher values indicating better fit.

Adjusted R-squared: Adjusts the (R^2) value for the number of predictors in the model. It penalizes the addition of irrelevant predictors.

Mean Squared Error (MSE): The average of the squared differences between the observed and predicted values.

Root Mean Squared Error (RMSE): The square root of the MSE, providing a measure of the average magnitude of the errors.

Mean Absolute Error (MAE): The average of the absolute differences between the observed and predicted values.

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:

Year , temperature has positive co relation , Weather situation 3 has negative co-relation (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) they are contributing significantly and having high co-efficient.

General Subjective Questions

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

Linear Regression is a fundamental statistical and machine learning algorithm used to model the relationship between a dependent variable (target) and one or more independent variables (features). The goal is to find the best-fitting linear equation that describes the relationship between the variables.

Types of Linear Regression
Simple Linear Regression: Involves one independent variable.
Multiple Linear Regression: Involves two or more independent variables.
The Linear Regression Model
The linear regression model can be expressed as:

[y = beta_0 + beta_1* x_1 + beta_2 *x_2 + ... + beta_n *x_n + epsilon]

where:

(y) is the dependent variable.
(x_1, x_2,..., x_n) are the independent variables.
(beta_0) is the intercept.
(beta_1, beta_2,..., beta_n) are the coefficients.
(epsilon) is the error term (residual).

Steps in Linear Regression

Data Collection: Gather the data that includes the dependent variable and the independent variables.
Data Preprocessing: Clean the data, handle missing values, and perform exploratory data analysis (EDA).
Feature Selection: Choose the relevant features that have a significant impact on the dependent variable.
Model Training: Fit the linear regression model to the training data.
Model Evaluation: Evaluate the model's performance using appropriate metrics.
Prediction: Use the trained model to make predictions on new data.

Model Training

Simple Linear Regression

For a single independent variable (x), the model is:

[y = beta_0 + beta_1 * x + epsilon]

The objective is to find the best-fitting line by minimizing the sum of the squared differences between the observed values and the predicted values. This is known as the Ordinary Least Squares (OLS) method.

The coefficients (beta_0) and (beta_1) are estimated using the following formulas:

[beta_1 = frac{sum (x_i - bar{x})(y_i - bar{y})}{sum (x_i - bar{x})^2}]

[beta_0 = bar{y} - beta_1 bar{x}]

where (bar{x}) and (bar{y}) are the means of the independent and dependent variables, respectively.

Multiple Linear Regression

For multiple independent variables (x_1, x_2,...., x_n), the model is:

[y = beta_0 + beta_1 *x_1 + beta_2*x_2 +...+ beta_n*x_n + epsilon]

The coefficients are estimated using matrix operations. The formula for the coefficients is:

[{beta} = ({X}^T {X})^{-1} {X}^T{y}]

where:
({X}) is the matrix of independent variables.
({y}) is the vector of the dependent variable.
(beta) is the vector of coefficients.

2. Explain the Anscombe's quartet in detail. (3 marks)
Ans:
Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics yet appear very different when graphed. These datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and to show how statistical properties can be misleading if not visualized.

The Four Datasets

Each dataset consists of 11 (x, y) points. Despite their similarities in summary statistics, the datasets exhibit very different distributions and relationships when plotted. Here are the key statistics that are nearly identical across all four datasets:

Mean of x: 9
Mean of y: 7.5
Variance of x: 11
Variance of y: 4.125
Correlation between x and y: 0.816
Linear regression line: (y = 3 + 0.5x)

The Four Plots

Dataset I: This dataset appears to have a simple linear relationship with a small amount of random noise.

Dataset II: This dataset forms a perfect quadratic relationship (a parabola), which is not captured by the linear regression line.

Dataset III: This dataset includes an outlier that significantly influences the linear regression line, making it appear as though there is a strong linear relationship.

Dataset IV: This dataset has a vertical line of points with one outlier. The linear regression line is heavily influenced by this single outlier.

Lessons from Anscombe's Quartet

Importance of Visualization: The quartet underscores the necessity of visualizing data before drawing conclusions from statistical analyses. Graphs can reveal patterns, relationships, and anomalies that summary statistics alone cannot.

Misleading Statistics: Descriptive statistics like mean, variance, and correlation can be misleading if not complemented with visual inspection. Different datasets can have identical statistical properties but very different distributions and relationships.

Outliers and Influential Points: Outliers can significantly affect statistical measures and regression models. Visualizing data helps identify such points and understand their impact.

Model Appropriateness: The quartet demonstrates that the same statistical model (e.g., linear regression) may not be appropriate for all datasets. Visual inspection can guide the selection of suitable models.

3. What is Pearson's R? (3 marks)
Ans:
The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation.
It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Between 0 and 1 Positive correlation When one variable changes, the other variable changes in the same direction.
0 There is no relationship between the variables. Between 0 and –1 when one variable changes, the other variable changes in the opposite direction.

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

Ans:
Scaling is a data preprocessing technique used to adjust the range of feature values in a dataset. This is particularly important in machine learning and statistical modeling, where the scale of the features can significantly impact the performance and convergence of algorithms.

Why Scaling is Performed

Algorithm Performance: Many machine learning algorithms, such as gradient descent-based methods (e.g., linear regression, logistic regression, neural networks), perform better and converge faster when the features are on a similar scale.
Distance-Based Algorithms: Algorithms like K-Nearest Neighbors (KNN) and clustering algorithms (e.g., K-Means) rely on distance metrics. Features with larger ranges can dominate the distance calculations, leading to biased results.
Improved Interpretability: Scaling can make the model coefficients more interpretable, especially when comparing the importance of different features.

Key Differences Between Normalized and Standardized Scaling is Normalized scaling transforms data to a fixed range (e.g., [0, 1]) and Standardized Scaling transforms data to have a mean of 0 and a standard deviation of 1.

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

(3 marks)

Ans:

The **Variance Inflation Factor (VIF)** is a measure used to detect the presence and severity of multicollinearity in a regression model. Multicollinearity occurs when two or more predictors in the model are highly correlated, which can inflate the variance of the coefficient estimates and make the model unstable. VIF values above 5 are often considered a strong hint that trying to reduce the multicollinearity of the regression might be worthwhile.

Why VIF Can Be Infinite

An infinite VIF value indicates perfect multicollinearity, meaning that one predictor variable is a perfect linear combination of one or more other predictor variables. This situation makes it impossible to estimate the regression coefficients uniquely because the predictors are linearly dependent.

Example Scenario:

Consider a simple linear regression model with two predictor variables, (X_1) and (X_2). If (X_2) is a perfect linear combination of (X_1), we can express (X_2) as:

[X_2 = a + bX_1]

where (a) and (b) are constants.

Demonstration

Let's assume we have the following data:

Observation	(X_1)	(X_2)
1	1	3
2	2	5
3	3	7
4	4	9
5	5	11

In this case, (X_2) can be perfectly predicted by (X_1) using the equation:

[X_2 = 2 + 2X_1]

This indicates perfect multicollinearity because (X_2) is an exact linear combination of (X_1).

Calculation of VIF

The VIF for (X_1) is calculated as:

{VIF}(X_1) = 1/{1 - R^2_{X_1}}

where (R^2_{X_1}) is the coefficient of determination from regressing (X_1) on all other predictors (in this case, (X_2)).

Since (X_2) is perfectly collinear with (X_1), the (R^2_{X_1}) will be 1. Therefore:

{VIF}(X_1) = {1}/{1 - 1} = infty

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

(3 marks)

Ans:

Use of Q-Q Plots in Linear Regression:

In the context of linear regression, Q-Q plots are primarily used to assess whether the residuals (the differences between observed and predicted values) follow a normal distribution. This is important because one of the key assumptions of linear regression is that the residuals are normally distributed.

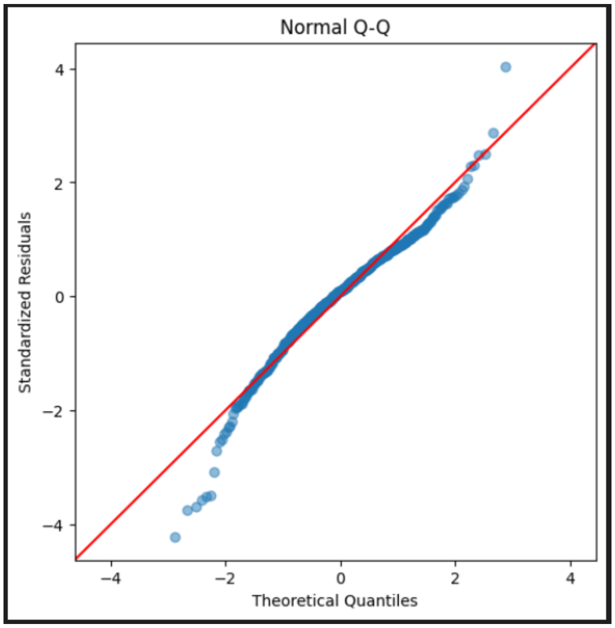
Importance of Q-Q Plots

1. **Assumption Checking**: Linear regression assumes that the residuals are normally distributed. A Q-Q plot helps in visually checking this assumption. If the residuals are normally distributed, the points on the Q-Q plot will lie on a straight line.
2. **Identifying Deviations**: Q-Q plots can help identify deviations from normality, such as skewness or kurtosis. Deviations from the straight line indicate that the residuals may not be normally distributed, which can affect the validity of the regression model.
3. **Model Diagnostics**: By examining the Q-Q plot, one can diagnose potential issues with the model. For instance, systematic deviations from the straight line might suggest that the model is missing key predictors or that there are outliers affecting the model.

How to Interpret a Q-Q Plot

1. **Straight Line**: If the points lie on or near the line (y = x), the residuals are approximately normally distributed.
2. **S-Shaped Curve**: If the points form an S-shaped curve, it indicates that the residuals have heavy tails (leptokurtic distribution).
3. **Inverted S-Shaped Curve**: If the points form an inverted S-shaped curve, it suggests that the residuals have light tails (platykurtic distribution).

4. **Systematic Deviations**: If the points systematically deviate from the line, it may indicate skewness or other forms of non-normality.


Normal Q-Q
Standardized Residuals
Theoretical Quantiles