# 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)

Here are some key insights derived from the analysis of categorical variables:

→ Median bike rentals have shown a consistent increase year over year.

→ Fall records the highest median rentals, which aligns with optimal weather conditions during this season.

→ Monthly rental trends reflect seasonal variations.

→ Bike rentals are higher on non-holidays compared to holidays.

→ There is no significant difference in median rentals between working and non-working days.

→ Clear weather conditions are the most favorable for bike rentals.

# 

**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 is important because it helps eliminate redundancy by removing one of the dummy variables for each categorical feature. This prevents multicollinearity, a situation where dummy variables are highly correlated, which can negatively impact model performance and interpretability.

**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 analyzing the pair plot among numerical variables, the variable 'temp' shows the highest correlation with the 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)

To validate the assumptions of Linear Regression after building the model on the training set, a simple approach is to create a scatter plot of X vs. Y. If the data points align along a straight line, it indicates a linear relationship between the dependent and independent variables, confirming that the linearity assumption holds

**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)

# → Atemp - 0.5458

→ yr\_2019 - 0.2506

→ mnth\_3 - 0.0426

# 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: An Overview

Linear Regression is a supervised learning algorithm used for regression tasks, where the goal is to predict a continuous target variable based on one or more independent variables. It is widely used to identify relationships between variables and for forecasting.

Different regression models vary based on:

The type of relationship between dependent and independent variables.

The number of independent variables used in the model.

Hypothesis Function of Linear Regression

The equation of a simple linear regression model is:

𝑌=𝑚𝑋+𝑐

where:

Y = Predicted target variable.

X = Input feature (independent variable).

m = Coefficient (slope) of X,represents how much Y changes with X

c = Intercept, the value of Y when X=0

Training the Model

When training the model, we are given:

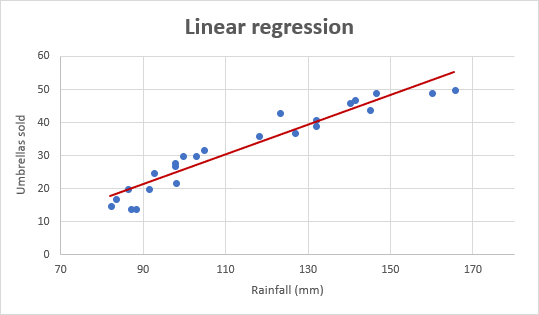
X: Input training data (can be univariate or multivariate).

Y: Corresponding target values (labels) in a supervised learning setup.

The goal of training is to find the optimal values of m and c that minimize the prediction error and best fit the data. The model determines these values by minimizing the difference between predicted and actual values using a method like Ordinary Least Squares (OLS) or Gradient Descent.

Once the best-fit regression line is found, the model can be used to make predictions:

Given a new input X, the model computes 𝑌 using the learned values of m and c.



**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 set of four datasets created by Francis Anscombe to illustrate the importance of data visualization in statistics. Each dataset has identical statistical properties, including the same mean, variance, correlation, and linear regression line. However, when visualized, the datasets show very different patterns:**

**Dataset 1: A perfect linear relationship.**

**Dataset 2: A nonlinear relationship that does not fit a straight line.**

**Dataset 3: A linear relationship with one outlier influencing the results.**

**Dataset 4: A nearly constant relationship with one extreme point that distorts the correlation.**

**This quartet demonstrates that summary statistics (mean, variance, correlation) can be misleading and highlights the necessity of plotting data before drawing conclusions. It encourages statisticians to visualize data to uncover hidden patterns, outliers, or anomalies that could affect analysis.**

**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 called the correlation coefficient) is a statistical measure that quantifies the strength and direction of a linear relationship between two variables. It ranges from -1 to 1:

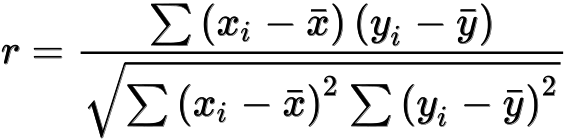
→ 1 indicates a perfect positive linear relationship

→ 1 indicates a perfect negative linear relationship

→ 0 indicates no linear relationship

A value closer to 1 or -1 implies a stronger linear correlation, while a value closer to 0 suggests a weaker or no linear correlation.

Pearson’s R is computed as the covariance of the two variables divided by the product of their standard deviations:



Where:

Xi and Yi are the individual data points.

Xˉ and Yˉ are the means of the variables.

Pearson’s R is sensitive to outliers, so extreme values can distort the correlation. It assumes a **linear relationship** and works best when the data is normally distributed.

**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 the process of adjusting the values of features in your data to a common scale without distorting differences in the ranges of values. It's crucial for ensuring that all features contribute equally to the model, especially when their ranges differ significantly.

* **Normalized Scaling (Min-Max Scaling)**: Rescales the features to a specified range, usually **0 to 1**. It works by subtracting the minimum value of the feature and dividing by the range (max - min). This method is sensitive to outliers because outliers can push the values to extreme ends.

Formula:

Xnorm​=X−min(X)/max(X)−min(X)

* **Standardized Scaling (Z-score Scaling)**: Standardization rescales the data so that it has a **mean of 0** and a **standard deviation of 1**. This is achieved by subtracting the mean of the feature and dividing by its standard deviation. It’s less sensitive to outliers than normalization and is commonly used when data follows a Gaussian (normal) distribution or when you want to compare features with different units.

Formula:

Xstd​=X−μ​/σ

where μ is the mean and σ is the standard deviation.

### ****Key Differences:****

* **Normalization**: Scales data to a fixed range (typically 0-1). It's useful when the features have different ranges and you need them on a comparable scale, like for distance-based algorithms (k-NN, SVM).
* **Standardization**: Scales data to have a mean of 0 and a standard deviation of 1. It's preferred when the data has outliers or when features have different units or scales, especially for algorithms that assume normality (e.g., Linear Regression, Logistic 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)

If there is perfect correlation between two independent variables, the **Variance Inflation Factor (VIF)** becomes infinite. This occurs because, in the case of perfect correlation, the **R²** value is 1, leading to a VIF calculation of 1−R21​, which results in infinity. This indicates **perfect multicollinearity**, where one variable can be perfectly predicted from another.

To address this issue, one of the correlated variables must be dropped from the dataset to resolve the multicollinearity problem. An infinite VIF value suggests that the corresponding variable is a linear combination of other variables, which also shows infinite VIFs, causing instability in regression models.

**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 plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. It plots the quantiles of the data against the quantiles of the reference distribution. If the points fall roughly along a straight line, it suggests that the data follows that distribution.

Use and Importance in Linear Regression:

Checking Normality of Residuals: In linear regression, a key assumption is that the residuals (the differences between observed and predicted values) are normally distributed. A Q-Q plot helps visually check if this assumption holds. If the residuals form a straight line in the Q-Q plot, it indicates that they follow a normal distribution.

Assessing Model Validity: If the residuals deviate significantly from the straight line (i.e., exhibit a pattern or curve), it may indicate problems with the model, such as:

→ Non-normality of errors, which could suggest that a different model or transformation of data might be needed.

→ Outliers or extreme values affecting the model's assumptions.

Improving Model Performance: By using the Q-Q plot, you can identify issues with residuals early and decide whether to transform variables, apply different regression techniques, or address outliers, all of which can improve model performance.

Why It's Important:

Validity Check: Ensures that the assumptions behind linear regression are satisfied, particularly the normality of residuals, leading to more reliable and valid results.

Model Diagnostics: Helps detect issues like skewness, heavy tails, or outliers, which could impact the accuracy of regression predictions.

In simple terms, a Q-Q plot is like a "reality check" for your regression model, making sure the residuals behave the way they should.

