# 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)

 **dentify Categorical Variables**: Recognize factors like gender, location, etc.

 **Examine Descriptive Stats**: Look at frequency counts and proportions to understand data balance.

#  **Visualize**: Use box plots or bar charts to observe relationships between categories and the dependent variable.

 **Statistical Tests**:

* **Chi-Square** for categorical outcomes.
* **ANOVA** or **t-tests** for continuous outcomes.

 **Modeling**:

* **Logistic Regression** for binary dependent variables.
* **Linear Regression** for continuous dependent variables.
* **Decision Trees** to assess categorical variable importance.

#  **Interpretation**: Significant differences or relationships would indicate an effect of the categorical variable on the dependent variable.

# 

**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 to avoid multicollinearity in regression models. When we create dummy variables for a categorical variable, each category gets transformed into a separate column. However, including all dummy variables can lead to perfect multicollinearity, where one variable can be perfectly predicted by others. This causes issues in regression models, as it makes it difficult to estimate coefficients reliably.

By setting drop\_first=True, one category is dropped, serving as the reference category. This ensures that the remaining dummy variables are not perfectly correlated, allowing the regression model to estimate the effects of the categorical variable correctly without multicollinearity.

**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)

# The numerical variable with the highest correlation to the target variable can be identified by visually inspecting the pair-plot. The variable that shows the strongest linear relationship (either positive or negative) with the target variable indicates the highest correlation. For precise measurement, the Pearson correlation coefficient can also be calculated.

**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. **Linearity**: Verify that there is a linear relationship between the independent and dependent variables by examining scatter plots of the predicted vs. actual values. If the relationship appears linear, the assumption is satisfied.
2. **Independence of Errors**: Check for autocorrelation of residuals (errors) using the **Durbin-Watson test**. A value close to 2 suggests no autocorrelation.
3. **Homoscedasticity**: Ensure constant variance of residuals by plotting the residuals against the fitted values. If the plot shows a random spread, the assumption holds. A funnel or pattern indicates heteroscedasticity.
4. **Normality of Errors**: Use a **Q-Q plot** or perform a **Shapiro-Wilk test** to check if the residuals are normally distributed. If the points in the Q-Q plot lie along a straight line, the residuals are normally distributed.
5. **No Multicollinearity**: Check **Variance Inflation Factor (VIF)** for each predictor. A VIF above 5-10 suggests multicollinearity, which could affect the model’s reliability.

By validating these assumptions, I can ensure the linear regression model is appropriate for the data.

**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)

To identify the top 3 features contributing significantly to explaining the demand for shared bikes based on the final linear regression model, I would typically look at:

1. **Coefficients**: Features with larger absolute coefficients have a stronger impact on the target variable (demand for shared bikes). A positive coefficient indicates a direct relationship, and a negative coefficient indicates an inverse relationship with demand.
2. **P-values**: Features with smaller p-values (typically less than 0.05) are statistically significant in predicting the target variable. Features with low p-values and large coefficients are important contributors to the model.
3. **Standardized Coefficients (Beta Weights)**: These values help compare the relative importance of features with different scales. Higher absolute values of standardized coefficients indicate stronger influence on the target variable.

By analyzing the **coefficients**, **p-values**, and **standardized coefficients**, I would identify the top 3 features that contribute most significantly to explaining the demand for shared bikes.

# 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)

# <Your answer for Question 6 goes here>

Linear regression is a statistical method used to model the relationship between a dependent (target) variable and one or more independent (predictor) variables by fitting a linear equation to the observed data. Here’s a detailed explanation of the linear regression algorithm:

### 1. ****Objective of Linear Regression:****

The goal of linear regression is to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the difference between the observed values and the predicted values of the target variable.

### 2. ****Linear Regression Equation:****

For a simple linear regression (one predictor), the model assumes the relationship between the independent variable XXX and the dependent variable YYY is linear and can be expressed as:

Y=β0+β1X+ϵY = \beta\_0 + \beta\_1 X + \epsilonY=β0​+β1​X+ϵ

* **YYY**: Dependent variable (target).
* **XXX**: Independent variable (predictor).
* **β0\beta\_0β0​**: Intercept term (the value of YYY when X=0X = 0X=0).
* **β1\beta\_1β1​**: Slope term (the change in YYY for a unit change in XXX).
* **ϵ\epsilonϵ**: Error term (the difference between the observed and predicted values of YYY).

In **multiple linear regression**, where there are multiple predictors, the equation extends to:

Y=β0+β1X1+β2X2+⋯+βnXn+ϵY = \beta\_0 + \beta\_1 X\_1 + \beta\_2 X\_2 + \dots + \beta\_n X\_n + \epsilonY=β0​+β1​X1​+β2​X2​+⋯+βn​Xn​+ϵ

* **X1,X2,…,XnX\_1, X\_2, \dots, X\_nX1​,X2​,…,Xn​**: Multiple independent variables.
* **β1,β2,…,βn\beta\_1, \beta\_2, \dots, \beta\_nβ1​,β2​,…,βn​**: Coefficients for each predictor.

### 3. ****Assumptions of Linear Regression:****

To ensure the model is valid and produces reliable estimates, linear regression makes several assumptions:

* **Linearity**: The relationship between the independent variables and the target is linear.
* **Independence of errors**: The residuals (errors) are independent of each other.
* **Homoscedasticity**: The variance of errors is constant across all levels of the independent variables.
* **Normality of errors**: The residuals are normally distributed.
* **No multicollinearity**: Independent variables should not be highly correlated with each other.

### 4. ****Training the Model:****

The core task of linear regression is to estimate the values of the coefficients (β0,β1,…,βn\beta\_0, \beta\_1, \dots, \beta\_nβ0​,β1​,…,βn​) that minimize the difference between the observed and predicted values. This is achieved by minimizing the **cost function**, typically the **Mean Squared Error (MSE)**:

MSE=1N∑i=1N(Yi−Y^i)2\text{MSE} = \frac{1}{N} \sum\_{i=1}^{N} (Y\_i - \hat{Y}\_i)^2MSE=N1​i=1∑N​(Yi​−Y^i​)2

Where:

* YiY\_iYi​ is the actual value of the dependent variable for the iii-th observation.
* Y^i\hat{Y}\_iY^i​ is the predicted value for the iii-th observation.
* NNN is the number of data points.

The MSE measures the average squared difference between the actual and predicted values, and minimizing this error helps find the optimal coefficients.

The coefficients (β\betaβ) are typically estimated using **Ordinary Least Squares (OLS)**, which solves for β\betaβ by taking the derivative of the MSE with respect to each coefficient and setting it to zero.

### 5. ****Making Predictions:****

Once the coefficients (β\betaβ) are estimated, the linear regression model can be used to make predictions for new data:

Y^=β0+β1X1+β2X2+⋯+βnXn\hat{Y} = \beta\_0 + \beta\_1 X\_1 + \beta\_2 X\_2 + \dots + \beta\_n X\_nY^=β0​+β1​X1​+β2​X2​+⋯+βn​Xn​

Where Y^\hat{Y}Y^ is the predicted value for the target variable based on new input values X1,X2,…,XnX\_1, X\_2, \dots, X\_nX1​,X2​,…,Xn​.

### 6. ****Model Evaluation:****

After fitting the linear regression model, it's important to evaluate its performance:

* **R-squared (R2R^2R2)**: This statistic measures how well the model explains the variance in the dependent variable. R2R^2R2 ranges from 0 to 1, with 1 indicating a perfect fit.

R2=1−∑i=1N(Yi−Y^i)2∑i=1N(Yi−Yˉ)2R^2 = 1 - \frac{\sum\_{i=1}^{N} (Y\_i - \hat{Y}\_i)^2}{\sum\_{i=1}^{N} (Y\_i - \bar{Y})^2}R2=1−∑i=1N​(Yi​−Yˉ)2∑i=1N​(Yi​−Y^i​)2​

Where:

* + YiY\_iYi​ is the actual value of the dependent variable.
  + Y^i\hat{Y}\_iY^i​ is the predicted value.
  + Yˉ\bar{Y}Yˉ is the mean of the actual values.
* **Mean Absolute Error (MAE)** and **Root Mean Squared Error (RMSE)**: These are other metrics to evaluate the prediction accuracy by measuring the average absolute or squared difference between the predicted and actual values.

# Linear regression is a simple yet powerful algorithm used for predicting a continuous target variable based on one or more independent variables. By fitting a linear equation, it estimates the coefficients that minimize the error between predicted and actual values. Validating assumptions and evaluating the model's performance ensures that the predictions are reliable and meaningful.

**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)

# <Your answer for Question 7 goes here>

Anscombe's quartet is a famous set of four datasets that were created by the statistician **Francis Anscombe** in 1973. The quartet is designed to demonstrate the importance of visualizing data before analyzing it, especially when using statistical methods like correlation and regression. Despite all four datasets having the same basic summary statistics (mean, variance, correlation, etc.), the data exhibit very different patterns when plotted. This highlights how summary statistics alone can be misleading.

### The Four Datasets in Anscombe’s Quartet:

Anscombe's quartet consists of four small datasets, each with 11 data points. These datasets have the following properties in common:

* The **mean** of xxx and yyy are identical for each dataset.
* The **variance** of xxx is identical for each dataset.
* The **correlation** between xxx and yyy is identical for each dataset.
* The **regression line** (best fit line) for each dataset is the same.

### The Datasets:

1. **Dataset I (Linear Relationship)**:
   * This dataset shows a perfect linear relationship between xxx and yyy, meaning that the data points lie along a straight line.
   * The regression line will be a good fit, showing a clear linear trend.
2. **Dataset II (Linear Relationship with Outlier)**:
   * This dataset is similar to Dataset I, except for one extreme outlier.
   * The regression line is still roughly the same as in Dataset I, but the outlier may significantly affect the model, which is why visualizing the data is important.
3. **Dataset III (Non-linear Relationship)**:
   * This dataset shows a clear **curved relationship** between xxx and yyy, which is not linear.
   * The regression line is a poor fit to the data, even though the summary statistics (mean, variance, correlation) are still the same as in the other datasets.
   * A linear model would not capture the true relationship here.
4. **Dataset IV (Vertical Line with Outlier)**:
   * In this dataset, most of the data points are aligned vertically, but there is one outlier that skews the data.
   * The regression line appears almost horizontal, again not fitting the data well. However, the summary statistics might suggest that the data follows a linear trend, which is misleading.

**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)

# <Your answer for Question 8 goes here>

**Pearson’s R**, also known as the **Pearson correlation coefficient**, is a statistical measure that quantifies the **linear relationship** between two continuous variables. It provides a value between **-1** and **+1**, indicating the strength and direction of the relationship.

### Formula:

The Pearson correlation coefficient (rrr) is calculated using the following formula:

r=∑(Xi−Xˉ)(Yi−Yˉ)∑(Xi−Xˉ)2∑(Yi−Yˉ)2r = \frac{\sum (X\_i - \bar{X})(Y\_i - \bar{Y})}{\sqrt{\sum (X\_i - \bar{X})^2 \sum (Y\_i - \bar{Y})^2}}r=∑(Xi​−Xˉ)2∑(Yi​−Yˉ)2​∑(Xi​−Xˉ)(Yi​−Yˉ)​

Where:

* XiX\_iXi​ and YiY\_iYi​ are the individual data points of the variables XXX and YYY.
* Xˉ\bar{X}Xˉ and Yˉ\bar{Y}Yˉ are the means of variables XXX and YYY, respectively.

### Interpretation of Pearson’s R:

* **r=+1r = +1r=+1**: A perfect positive linear relationship between the two variables. As one variable increases, the other increases proportionally.
* **r=−1r = -1r=−1**: A perfect negative linear relationship. As one variable increases, the other decreases proportionally.
* **r=0r = 0r=0**: No linear relationship between the variables. However, there may still be a non-linear relationship.
* **0<r<10 < r < 10<r<1**: A positive linear relationship, with the strength of the relationship increasing as rrr approaches 1.
* **−1<r<0-1 < r < 0−1<r<0**: A negative linear relationship, with the strength of the relationship increasing as rrr approaches -1.

**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)

# <Your answer for Question 9 goes here>

# **Scaling** is the process of transforming features (or variables) in a dataset to a standard scale or range. In machine learning and statistics, scaling is often necessary because different features in the dataset may have different units or magnitudes, which can affect the performance of certain algorithms. Scaling ensures that all features contribute equally to the model's learning process.

 **Equal Contribution**: Many algorithms, such as gradient descent, k-nearest neighbors (KNN), and support vector machines (SVM), rely on distance-based metrics or require optimization over the features. If the features have different units or scales, features with larger numerical values may dominate the learning process, leading to biased or suboptimal results.

 **Improved Convergence**: Scaling ensures that all features are treated equally during optimization. For gradient-based algorithms, it can improve the convergence rate, making training faster and more stable.

 **Consistency in Units**: Scaling standardizes the measurement of features. For example, features with units like meters and kilograms may need to be scaled to a consistent range or standard deviation to be compared meaningfully.

 **Avoiding Mathematical Issues**: Some algorithms assume that the data is centered around zero or within a particular range. If the data is not scaled properly, it might cause numerical instability or incorrect results.

| **Aspect** | **Normalization (Min-Max Scaling)** | **Standardization (Z-Score Scaling)** |
| --- | --- | --- |
| **Range** | Scales the data to a fixed range (e.g., [0, 1] or [-1, 1]) | Centers data around 0 with a standard deviation of 1 |
| **Formula** | Xnormalized=X−XminXmax−XminX\_{\text{normalized}} = \frac{X - X\_{\text{min}}}{X\_{\text{max}} - X\_{\text{min}}}Xnormalized​=Xmax​−Xmin​X−Xmin​​ | Xstandardized=X−μσX\_{\text{standardized}} = \frac{X - \mu}{\sigma}Xstandardized​=σX−μ​ |
| **Effect on Distribution** | Data is compressed into a fixed range, possibly distorting relationships between values | Data is centered and scaled, preserving the underlying distribution |
| **When to Use** | When features have known min-max range or when using distance-based algorithms (e.g., KNN, neural networks) | When data is normally distributed or when features have different units and scales |
| **Handling Outliers** | Sensitive to outliers, as they affect the min and max values | Less sensitive to outliers compared to normalization, but can still be influenced |

**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)

# <Your answer for Question 10 goes here>

The **Variance Inflation Factor (VIF)** measures how much the variance of an estimated regression coefficient increases when your predictors are correlated. It is used to detect multicollinearity in a regression model.

The formula for VIF for a predictor XiX\_iXi​ is:

VIF(Xi)=11−Ri2VIF(X\_i) = \frac{1}{1 - R\_i^2}VIF(Xi​)=1−Ri2​1​

Where:

* Ri2R\_i^2Ri2​ is the coefficient of determination obtained by regressing the predictor XiX\_iXi​ on all the other predictors.

A **VIF** can be infinite when the predictor XiX\_iXi​ is **perfectly linearly related** to the other predictors in the model, i.e., when there is **perfect multicollinearity**. This occurs if:

1. **Perfect Correlation Between Predictors**: If one of the predictors is an exact linear combination of other predictors, then Ri2R\_i^2Ri2​ will be **1**. In this case, the denominator in the VIF formula becomes 1−1=01 - 1 = 01−1=0, causing the VIF to approach infinity.
2. **Singular Matrix**: In linear regression, the matrix of predictors (often denoted as XXX) must be invertible for the model to compute the coefficients. If there is perfect multicollinearity, the matrix becomes **singular** (i.e., non-invertible), meaning that one or more predictors can be expressed as a linear combination of others. This leads to an infinite VIF for the collinear predictor.

### Common Causes of Infinite VIF:

* **Duplicate Variables**: Having identical or nearly identical variables in the dataset.
* **Linear Combinations**: If one predictor is a simple multiple or linear combination of another.
* **Dummy Variable Trap**: Including both a categorical variable and a full set of dummy variables for that categorical variable in the model can cause perfect multicollinearity.

### How to Handle Infinite VIF:

* **Remove or combine highly correlated predictors**: Eliminate one of the collinear variables.
* **Check for perfect linear relationships**: Ensure that predictors are not linear combinations of others.
* **Regularization**: Use regularization techniques such as Ridge or Lasso regression, which can handle multicollinearity by adding penalty terms to the regression.

**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)

# <Your answer for Question 11 goes here>

A **Q-Q (Quantile-Quantile) plot** is a graphical tool used to assess if a dataset follows a particular theoretical distribution, most commonly the **normal distribution**. The plot compares the quantiles of the sample data against the quantiles of a specified theoretical distribution.

* **X-axis**: Quantiles of the theoretical distribution (e.g., normal distribution).
* **Y-axis**: Quantiles of the observed data.

 **Theoretical Quantiles**: The quantiles are values that split the probability distribution into equal intervals. For a normal distribution, these quantiles are based on the z-scores.

 **Sample Quantiles**: The sample data points are sorted, and their corresponding quantiles are calculated.

 **Plotting**: The points are plotted, and if the sample data follows the specified distribution, the points should approximately lie on a straight line (typically a 45-degree line or identity line).

* If the points fall along the straight line, this suggests that the data follows the expected distribution (e.g., normal distribution).
* Deviations from the straight line indicate departures from the assumed distribution (e.g., skewness, heavy tails, etc.).

In the context of **linear regression**, the **Q-Q plot** is primarily used to check one of the key assumptions of linear regression: that the residuals (errors) of the model follow a **normal distribution**.

Here’s how a Q-Q plot is used and why it is important:

**1. Checking Normality of Residuals:**

* In linear regression, one of the assumptions is that the residuals (the differences between the observed and predicted values) should be **normally distributed** for the model to be valid. This is important because hypothesis tests (like t-tests) for the regression coefficients assume normality of residuals.
* A Q-Q plot helps visualize this assumption. If the residuals are normally distributed, the points in the Q-Q plot will follow a straight line.

**2. Detecting Skewness or Kurtosis:**

* If the Q-Q plot shows significant deviation from the straight line, it could indicate **skewness** (if the plot bends away from the line towards one side) or **kurtosis** (if the plot has heavy tails).
* Skewness indicates that the data is not symmetrically distributed around the mean. Kurtosis indicates that the data has outliers or extreme values that might affect the regression model.

**3. Detecting Outliers:**

* In a Q-Q plot, **outliers** will appear as points that are far away from the straight line, especially at the ends of the plot (i.e., the tails of the distribution).
* Outliers can have a significant effect on the results of linear regression, so detecting them early through a Q-Q plot can help in model diagnostics and improvements.

**4. Model Assumption Verification:**

* The Q-Q plot is a simple yet effective tool for visually validating the assumption of normality in residuals. If the residuals are not normally distributed, it could suggest the need for a transformation of the dependent or independent variables, or that a different model (e.g., non-linear regression) might be more appropriate.