**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:** The categorical variables in the assignment include "season," "workingday," "weathersit," "weekday," "yr," "holiday," and "mnth." Analyzing these variables reveals valuable insights:

**Season:**

* Summer and fall emerge as the most favorable seasons for biking, suggesting that targeting higher bike usage through strategic advertising during these periods could be beneficial.
* Spring exhibits a significantly lower consumption ratio.

**Workingday:**

* "Workingday" distinguishes between weekdays and weekends/holidays.
* Registered users prefer bikes on working days, while casual users show a preference for non-working days. Understanding the behaviors of these user segments could inform targeted strategies to increase overall bike rentals.

**Weekday:**

* While the "cnt" column doesn't reveal a significant pattern with weekdays, analyzing "registered" and "casual" users separately unveils higher bike usage on working days for registered users and the opposite for casual users.

**Yr:**

* The dataset spans two years, with an observable increase in bike rentals from 2018 to 2019.

**Weathersit:**

* Clean/few clouds days are the most favorable weather conditions for biking.
* Even on lightly rainy days, registered user counts remain high, indicating potential daily commuting patterns.

**Holiday:**

* Casual users tend to use bikes more on holidays, highlighting a notable difference when compared to registered users.

**Mnth:**

* Bike rental ratios are higher in June, July, August, September, and October.
* The 75th quantile experiences growth in the months identified as favorable for biking.
* Regarding the importance of using drop\_first=True during dummy variable creation:

When employing one-hot encoding to create dummy variables for categorical features, the parameter drop\_first=True becomes crucial. This setting ensures that only (n-1) dummy variables are created for a categorical variable with n categories, mitigating the risk of multicollinearity.

Multicollinearity, characterized by high correlation between independent variables in a regression model, can lead to unreliable parameter estimates. If all levels of a categorical variable are included as dummy variables without dropping one level, perfect multicollinearity arises, impacting the model's stability and interpretability.

By excluding the first level (reference category), drop\_first=True maintains linear independence among the dummy variables. This choice enhances the reliability of the regression model, facilitating more accurate estimation and interpretation.

1. Why is it important to use drop\_first=True during dummy variable creation? (2 mark)

**Ans:** It is important to use drop\_first=True during dummy variable creation because it is unnecessarily increasing the number of variables if we fail to drop the first variable.

**Example:** consider a column having three different values. If we create dummy variables, three different variables will be created for three different variables.

**1 0 0** 🡪 value 1

**0 1 0** 🡪 value 2

**0 0 1** 🡪 value 3

Instead we an do the same thing using two variables

**0 0** 🡪 value 1

**1 0** 🡪 value 2

**0 1** 🡪 value 3

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

**Ans:** The variable "temp" exhibits the strongest correlation with the target variable, registering a correlation coefficient of 0.63.

It's essential to note that the "casual" and "registered" variables contribute to the target variable, as their values are aggregated to form the overall target. Therefore, the correlation analysis doesn't explicitly consider these two variables to avoid redundancy in assessing their impact.

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

**Ans:**

After constructing the linear regression model on the training set, several validation steps were undertaken to ensure the robustness of the model and validate its underlying assumptions. These steps include:

**1. Residual Analysis**

Examining the residuals (the differences between predicted and actual values) is crucial. A scatter plot of residuals against predicted values helps assess whether there is a pattern, indicating potential issues like non-linearity or heteroscedasticity.

**2. Normality of Residuals**

Checking the normality of residuals is essential. A histogram or a Q-Q plot of the residuals is often employed to verify if they follow a normal distribution, which is an assumption of linear regression.

**3. Homoscedasticity**

Ensuring homoscedasticity, or constant variance of residuals across all levels of predictors, is critical. A plot of residuals against fitted values aids in detecting heteroscedasticity.

**4. Independence of Residuals**

The independence of residuals is fundamental. Autocorrelation or patterns in the residuals over time or across observations could indicate a violation of this assumption.

**5. Multicollinearity**

Assessing multicollinearity among predictor variables is necessary. High correlation among predictors can destabilize coefficient estimates. Variance Inflation Factor (VIF) calculations are commonly used for this purpose.

**6. Outlier Detection**

Identifying and addressing outliers is crucial. Outliers can disproportionately influence the model, and their impact should be evaluated using techniques such as leverage plots or Cook's distance.

**7. Linearity**

Confirming that the relationship between predictors and the response variable is linear is essential. Scatter plots of predicted vs. actual values help assess linearity.

These validation steps collectively ensure that the linear regression model meets its assumptions and can reliably generalize to new, unseen data. Adjustments or transformations may be applied as needed based on the outcomes of these validation checks.

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:**

The three most influential variables are:

1. **Weathersit:**

The pivotal factor among environmental conditions is temperature, significantly impacting the business positively. Conversely, other weather elements such as rain, humidity, wind speed, and cloudiness exert a negative influence on the business.

**2. Yr**

The year-over-year growth appears to be organic, influenced by inherent geographical characteristics.

**3. Season**

The winter season emerges as a key determinant in driving the demand for shared bikes.

**General Subjective Questions**

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

**Ans:** Linear regression is a statistical technique employed to model the association between two variables by presuming a linear link between the independent and dependent variables. The objective is to identify the optimal line that minimizes the sum of squared variances between predicted and observed values. Widely utilized in fields such as economics and finance, this method facilitates the analysis and prediction of data trends. It can be expanded to encompass multiple linear regression, which involves several independent variables, and logistic regression, particularly useful for addressing binary classification problems.

**Assumptions of simple linear regression**

1. Linear relationship between X and y.
2. Normal distribution of error terms.
3. Independence of error terms.
4. Constant variance of error terms.

Transitioning from Simple Linear Regression (SLR) to Multiple Linear Regression (MLR) introduces new factors that necessitate careful consideration:

1. **Overfitting Concerns:**

Overfitting arises when the model becomes excessively complex, performing exceptionally well on the training data but failing to generalize effectively to testing data. Striking a balance in model complexity is crucial to avoid this issue.

2. **Multicollinearity Assessment:**

Identifying multicollinearity becomes crucial in MLR, involving an examination of potential dependencies among the various independent variables. This helps in recognizing and addressing redundancy within the variable set, ensuring the model's stability and interpretability.

3. **Feature Selection Deliberations:**

In MLR, the challenge of feature selection emerges from the plethora of variables. Determining which features are most relevant and influential becomes pivotal. This process involves discarding redundant features and those that do not contribute significantly to the predictive capability of the model.

**Residuals**

Residuals represent the discrepancy between actual and predicted values in a regression model. When data points deviate significantly from the regression line, the residuals are high, leading to an elevated cost function. Conversely, if the data points closely align with the regression line, the residuals are small, resulting in a lower cost function.

**Gradient Descent**

Gradient Descent serves as a technique to minimize the Mean Squared Error (MSE) by computing the gradient of the cost function. In the context of a regression model, gradient descent is employed to adjust the coefficients of the line, systematically reducing the cost function. This process involves the initial random selection of coefficient values, followed by iterative updates to approach the minimum of the cost function.

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

**Ans:**

Anscombe's Quartet stands as a quintessential illustration highlighting the crucial role of data visualization. Developed by the statistician Francis Anscombe in 1973, this quartet comprises four distinct datasets, each consisting of eleven (x, y) points. The key insight lies in the fact that despite sharing identical descriptive statistics—such as mean, variance, and standard deviation—these datasets exhibit diverse graphical representations. This emphasizes the importance of visually exploring data before delving into statistical analyses, as distinct patterns and behaviors emerge through graphical depictions that may not be apparent solely through quantitative measures. Anscombe's Quartet serves as a compelling reminder of the limitations of relying solely on numerical summaries and underscores the value of incorporating visualizations in data analysis.

**Example**

Consider four datasets, each with the following (x, y) pairs:

Dataset I:

x = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]

y = [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68]

Dataset II:

x = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]

y = [9.14, 8.14, 8.74, 8.77, 9.26, 8.1, 6.13, 3.1, 9.13, 7.26, 4.74]

Dataset III:

x = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]

y = [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73]

Dataset IV:

x = [8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8]

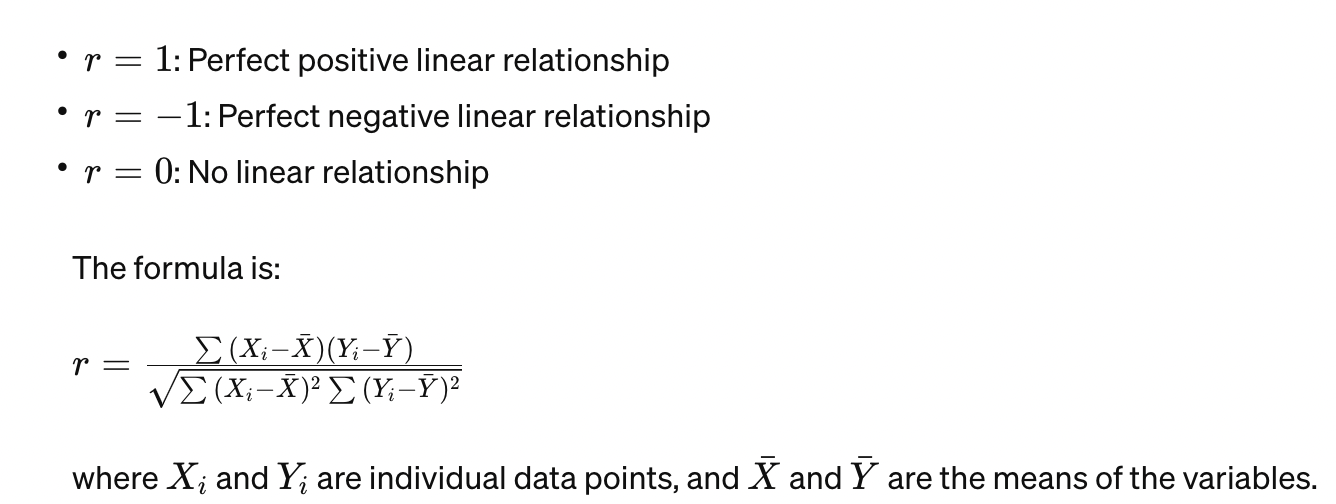
y = [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.5, 5.56, 7.91, 6.89]

Now, if we calculate the means, variances, and other descriptive statistics for the x and y values in each dataset, you'll find that they are nearly identical across all four sets. However, if we were to plot these datasets, you'd observe distinct graphical patterns. This showcases how relying solely on summary statistics might overlook important nuances in the data.

For instance, Dataset I exhibits a linear relationship, Dataset II shows a curved pattern, Dataset III has an outlier that strongly influences the regression line, and Dataset IV displays a vertical line where the x-values are constant but y-values vary widely. These visual differences emphasize the significance of data visualization in gaining a comprehensive understanding of the underlying patterns in the data.

3. What is Pearson’s R? (3 marks)

**Ans: P**earson’s Correlation Coefficient, often denoted as r, measures the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where:



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

**Ans:** Scaling in the context of data and statistics refers to the process of transforming variables or features of different scales or units into a standard range. The primary reasons for scaling are to ensure that the variables contribute equally to the analysis, prevent certain variables from dominating due to their larger scales, and to improve the performance and convergence of machine learning algorithms.

**Normalization**

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Normalization equation

Here, Xmax and Xmin are the maximum and the minimum values of the feature, respectively.

When the value of X is the minimum value in the column, the numerator will be 0, and hence X’ is 0

On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator, and thus the value of X’ is 1.If the value of X is between the minimum and the maximum value, then the value of X’ is between 0 and 1.

Standardization is another scaling method where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero, and the resultant distribution has a unit standard deviation.

Standardization equation

Where is the mean of the feature values and is the standard deviation of the feature values.

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

**Ans:** The occurrence of an infinite Variance Inflation Factor (VIF) is associated with perfect multicollinearity in a multiple regression analysis. The VIF is a metric used to assess the extent of multicollinearity among independent variables, and a high VIF value indicates a strong correlation between predictors. However, the VIF can become infinite when one independent variable is a perfect linear combination of others.

The formula for calculating the VIF is

**​**

is the coefficient of determination obtained by regressing the concerned independent variable against all other independent variables. If is equal to 1, the denominator becomes zero, resulting in an undefined VIF.

Here's an elaboration with an example:

Let's consider a dataset with three independent variables: X1, X2 and X3. Suppose X3 is a perfect linear combination of X1 and X2, meaning X3 can be precisely predicted by a linear combination of X1 and X2.

X3=a X1  + b X2

In this scenario, when we calculate the VIF for X3, the in the denominator of the VIF formula becomes 1. The VIF equation then transforms into , resulting in an undefined and infinite VIF for ​ X3.

.

This situation highlights the issue of perfect multicollinearity, where one variable is a duplicate or a constant multiple of others. To address this problem, it is essential to identify and resolve the multicollinearity issue through techniques such as variable transformations, eliminating redundant variables, or using regularization methods in regression analysis. Handling perfect multicollinearity ensures stable and reliable VIF values for assessing the relationships between independent variables.

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

**Ans:**

The Quantile-Quantile (Q-Q) plot is a graphical tool utilized for evaluating whether a given dataset is reasonably consistent with a specified theoretical distribution, such as the normal, exponential, or uniform distribution. Additionally, it aids in determining if two datasets originate from populations sharing a common distribution. This becomes particularly useful in the context of linear regression, where distinct training and test datasets are employed, and the Q-Q plot serves to confirm the similarity in the distributions of these datasets.

**Advantages of Q-Q plots include:**

**a) Applicability to Sample Sizes:**

Q-Q plots are versatile and can be effectively employed with varying sample sizes. This adaptability allows for the examination of distributions across datasets of different scales and magnitudes.

**b) Detection of Distributional Characteristics:**

Q-Q plots are powerful tools for detecting various distributional aspects, including shifts in location, alterations in scale, changes in symmetry, and the presence of outliers. By visually assessing how quantiles of the observed data compare to those of the theoretical distribution, analysts can identify subtle deviations and characteristics within the dataset.

In the context of linear regression, the Q-Q plot provides a visual means to verify whether both the training and test datasets are drawn from populations with comparable distributions. This confirmation is vital for ensuring that the assumptions underlying the regression analysis are met and that the model's performance can be expected to generalize well to new, unseen data.