Assignment Questions and Answers

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?**

Ans. Season, yr, mnth, holiday, weekday, weathersit are categorical variables in the dataset. From the analysis, it can be inferred that

- Fall is the season to get maximum active customers (September being the month). 2019 observed more sale than 2018.
- Holidays affect the active count which drops.
- During heavy rain, there are no users whereas partly cloudy/clear sky saw the maximum count.

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

Ans. The Dummy Variable Trap occurs when there is multicollinearity, meaning that one of the dummy variables can be perfectly predicted from the others. This happens because the sum of all dummy variables for a categorical feature will always equal 1.

Dropping the first dummy reduces the number of variables in the model, which can simplify the model and improve computational efficiency, especially when dealing with a large number of categories.

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

Ans. *atemp* and *temp* has the highest correlation with the target variable

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Ans. One of the fundamental assumptions of a linear regression model is that the error terms should correspond to a normal curve, when plotted on histogram. We spotted the same. Hence, this assumption is validated.

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

Ans. The top 3 features directly influencing the count are the features with highest coefficients. These are: Temp, Year (positively influencing) and snowy and rainy weather (negatively influencing).

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans. Linear regression is a statistical method used to model the relationship between a dependent variable (also known as the target or outcome variable) and one or more independent variables (also known as predictors, features, or explanatory variables). The goal is to fit a linear equation to observed data that can be used to predict the dependent variable based on the values of the independent variables.

Types of Linear Regression

1.Linear Regression: Models the relationship between one dependent variable and one independent variable.

2.Equation:

$y=\beta0+\beta1x+\epsilon$

1. y is the dependent variable.
2. x is the independent variable.
3. β0 is the y-intercept (constant term).
4. β1 is the slope of the line (coefficient of the independent variable).
5. epsilon ϵ is the error term (residuals).

Assumptions of Linear Regression

To produce reliable estimates and predictions, linear regression relies on several assumptions:

1.Linearity: The relationship between the dependent and independent variables is linear.

2.Independence: Observations are independent of each other, and the residuals (errors) are independent.

3.Homoscedasticity: The variance of the residuals is constant across all levels of the independent variables.

4.Normality: The residuals of the model are normally distributed.

5.No Multicollinearity: The independent variables are not highly correlated with each other.

2. Explain the Anscombe’s quartet in detail.

Ans. Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, such as mean, variance, correlation, and linear regression results, yet they differ significantly when graphed. The quartet was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and to illustrate how statistical properties can be misleading without visual exploration.

Key Points of Anscombe’s Quartet:

1.Identical Statistical Properties:

1. All four datasets have the same mean for the x and y values.
2. They have the same variance for x and y.
3. Each dataset has the same correlation coefficient between x and y.
4. The linear regression line (y = mx + c) is nearly the same for all four datasets.

2.Visual Differences:

1. **Dataset 1:** Shows a typical linear relationship between x and y, which would be expected based on the regression analysis.
2. **Dataset 2:** The data is more curvilinear, indicating a non-linear relationship despite the linear regression line.
3. **Dataset 3:** Contains an outlier, which heavily influences the regression line, leading to misleading interpretations if only the statistics are considered.
4. **Dataset 4:** All x-values are the same except for one, creating a vertical line. The single differing point (an outlier) forces the regression line to fit in a misleading way.

Importance of Anscombe’s Quartet:

•**Visual Exploration:** The quartet illustrates the crucial role of visualizing data before jumping to conclusions based on summary statistics. By plotting the data, one can identify patterns, outliers, or structures that simple statistics might miss.

•**Misleading Conclusions:** If one relies solely on statistical summaries without visual inspection, one might draw incorrect or oversimplified conclusions about the data.

•**Teaching Tool:** Anscombe’s Quartet is widely used in statistics education to teach the importance of graphical analysis and to caution against the over-reliance on summary statistics.

3. What is Pearson’s R?

Ans. **Pearson's R**, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two continuous variables. It quantifies the degree to which two variables are related, ranging from -1 to 1.

Key Points about Pearson's R:

•**Value Range**:

- **1**: Perfect positive linear relationship.
- **0**: No linear relationship.
- **-1**: Perfect negative linear relationship.

•**Interpretation**:

- **Positive correlation**: As one variable increases, the other tends to increase.
- **Negative correlation**: As one variable increases, the other tends to decrease.
- **Strength of the relationship**:
  - |r| close to 1 indicates a strong linear relationship.
  - |r| close to 0 indicates a weak linear relationship.

Formula for Pearson's R:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

- r = Pearson correlation coefficient
- xi and yi = Individual data points
- x¯ and y¯ = Mean of the x and y variables

Assumptions of Pearson's R:

1.Linearity: The relationship between the two variables should be linear.

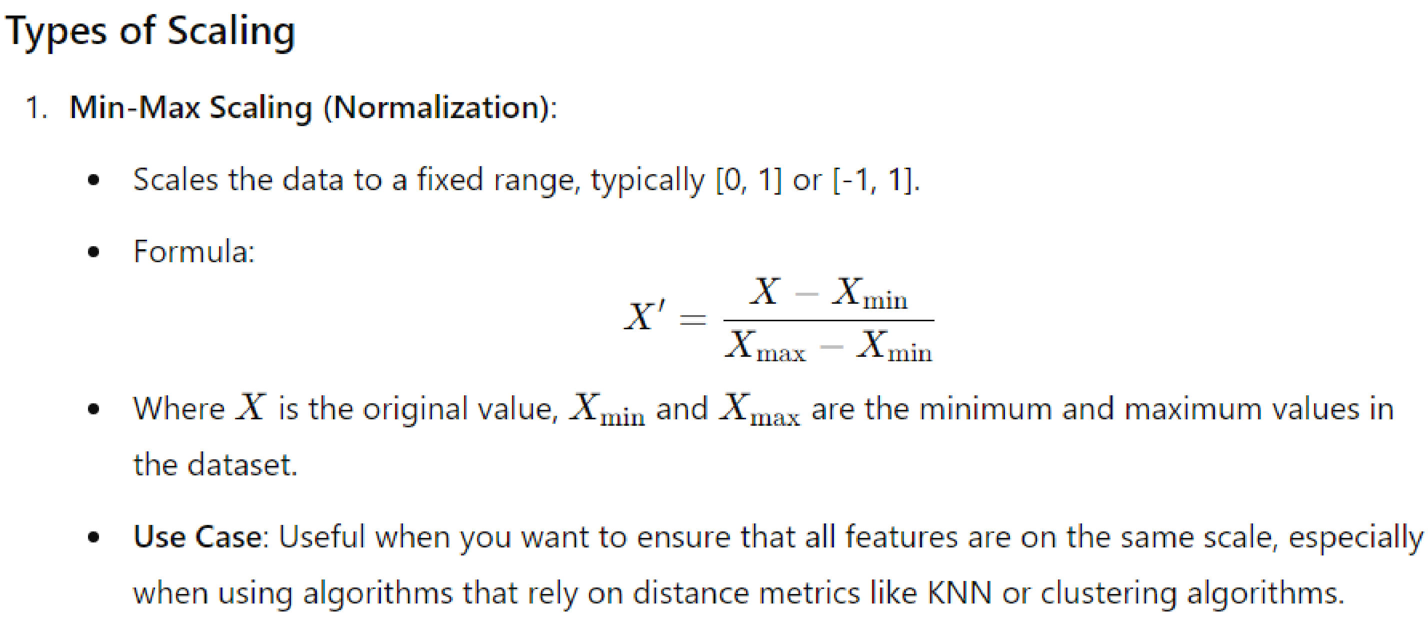
2.Continuous Variables: Both variables should be continuous.

3.Normality: The variables should follow a normal distribution (this assumption is more important for significance testing rather than the calculation of the coefficient itself).

4.Homoscedasticity: The variance of one variable should be constant at all levels of the other variable.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans. **Scaling** is a data preprocessing technique used to standardize or normalize the range of independent variables or features in your dataset. It's especially important when you are working with algorithms that are sensitive to the scale of the input data, such as gradient descent-based algorithms, support vector machines (SVM), k-nearest neighbors (KNN), and principal component analysis (PCA).

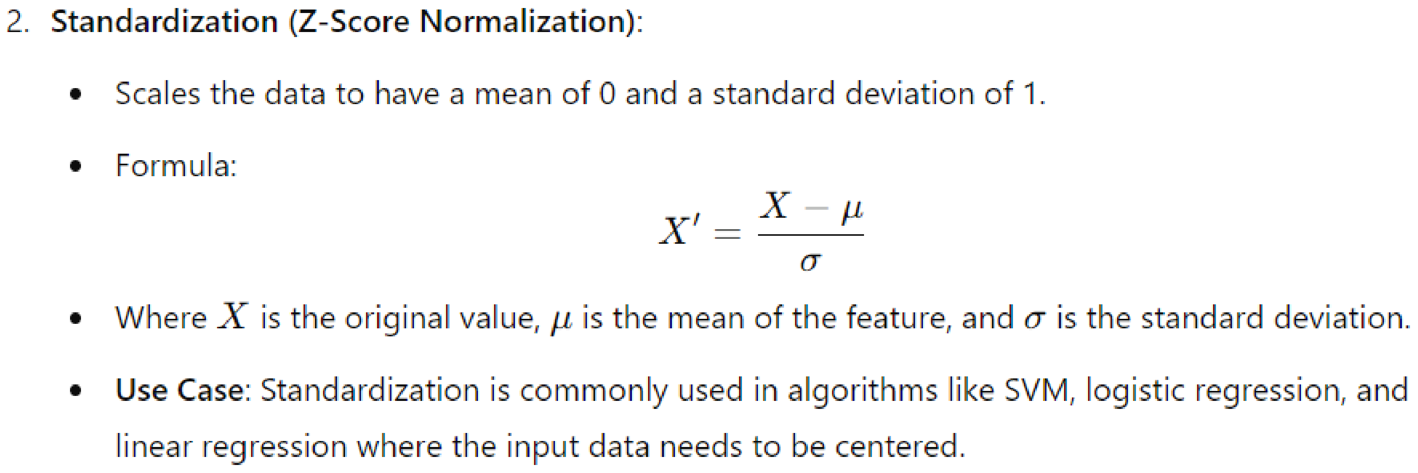


Types of Scaling

1. **Min-Max Scaling (Normalization)**:
   - Scales the data to a fixed range, typically [0, 1] or [-1, 1].
   - Formula:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

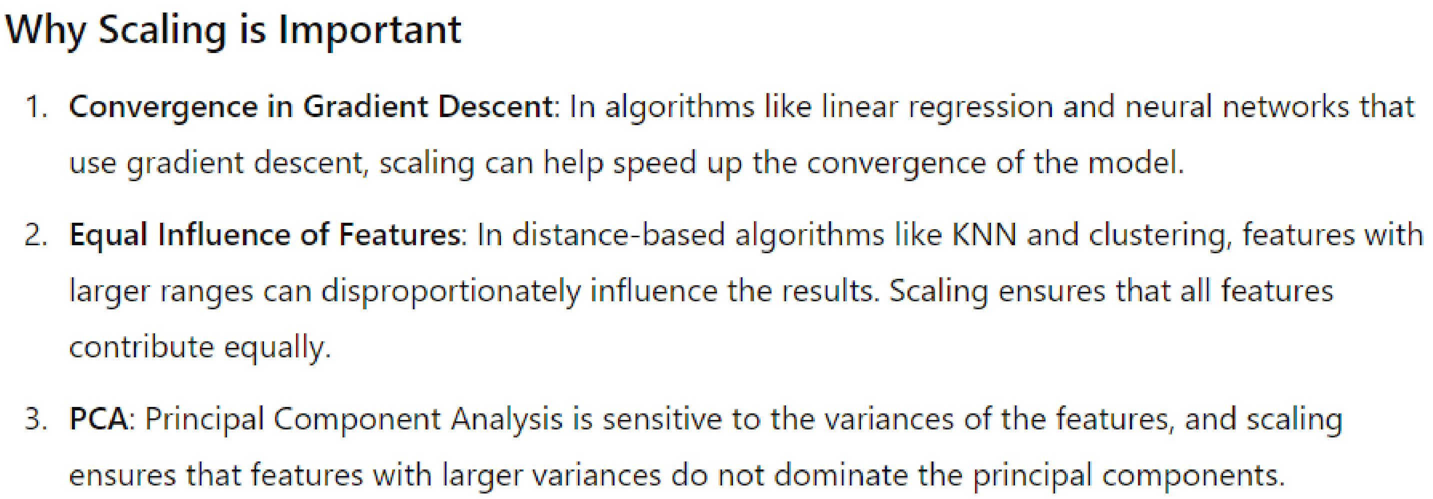
   - Where X is the original value, $X_{\min}$ and $X_{\max}$ are the minimum and maximum values in the dataset.
   - **Use Case**: Useful when you want to ensure that all features are on the same scale, especially when using algorithms that rely on distance metrics like KNN or clustering algorithms.



2. **Standardization (Z-Score Normalization)**:
   - Scales the data to have a mean of 0 and a standard deviation of 1.
   - Formula:

$$X' = \frac{X - \mu}{\sigma}$$

   - Where X is the original value, μ is the mean of the feature, and σ is the standard deviation.
   - **Use Case**: Standardization is commonly used in algorithms like SVM, logistic regression, and linear regression where the input data needs to be centered.



Why Scaling is Important

1. **Convergence in Gradient Descent**: In algorithms like linear regression and neural networks that use gradient descent, scaling can help speed up the convergence of the model.
2. **Equal Influence of Features**: In distance-based algorithms like KNN and clustering, features with larger ranges can disproportionately influence the results. Scaling ensures that all features contribute equally.
3. **PCA**: Principal Component Analysis is sensitive to the variances of the features, and scaling ensures that features with larger variances do not dominate the principal components.

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

Ans. Infinite VIF Value :

The Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in a regression model. It quantifies how much the variance of a regression coefficient is inflated due to the presence of multicollinearity among the independent variables. A VIF value becomes infinite when perfect multicollinearity is present, meaning that one independent variable is an exact linear combination of one or more other independent variables.

Cause of Infinite VIF :

An infinite VIF occurs when the correlation between one independent variable and a combination of the other independent variables is perfect (correlation coefficient of 1 or -1). In this case, the model cannot distinguish between the perfectly correlated variables, and the variance of the affected variable's coefficient is infinitely inflated. Mathematically, this happens when the determinant of the matrix (X'X) used to compute VIF is zero, leading to a division by zero.

Implication and Solution :

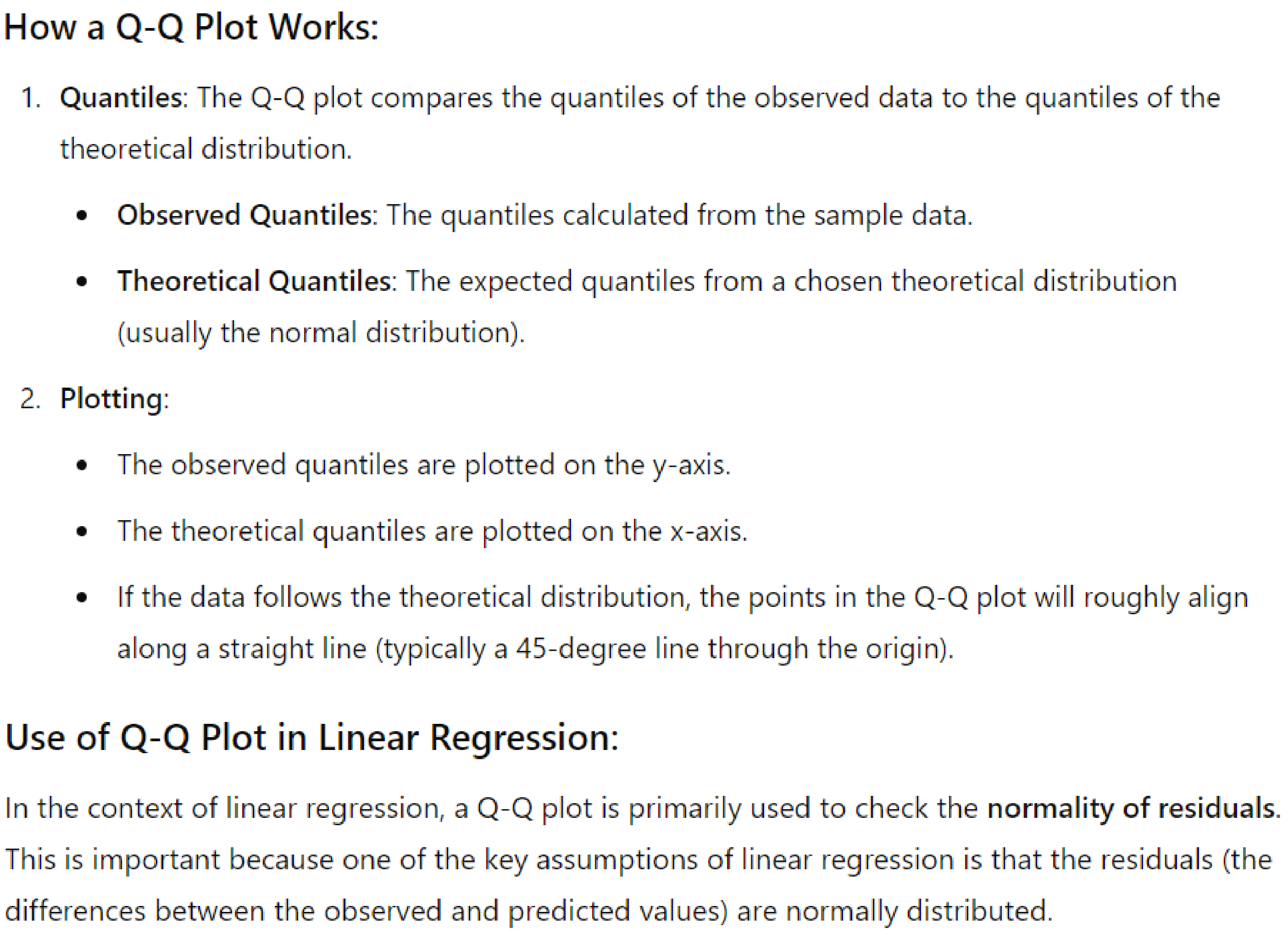
•**Implication:** An infinite VIF indicates severe multicollinearity, which can destabilize the regression coefficients, making them highly sensitive to changes in the model. This undermines the reliability of the model and can lead to incorrect interpretations.

•**Solution:** To address infinite VIF, you may need to:

- **Remove one of the perfectly correlated variables** from the model.
- **Combine the correlated variables** into a single feature through techniques like Principal Component Analysis (PCA).
- **Rethink the model design** to avoid including redundant predictors.

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

Ans. A **Q-Q plot** (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically a normal distribution. It helps to assess whether the data follows a particular distribution, making it a valuable diagnostic tool in statistical analysis, especially in linear regression.



How a Q-Q Plot Works:

1. **Quantiles**: The Q-Q plot compares the quantiles of the observed data to the quantiles of the theoretical distribution.
   - **Observed Quantiles**: The quantiles calculated from the sample data.
   - **Theoretical Quantiles**: The expected quantiles from a chosen theoretical distribution (usually the normal distribution).
2. **Plotting**:
   - The observed quantiles are plotted on the y-axis.
   - The theoretical quantiles are plotted on the x-axis.
   - If the data follows the theoretical distribution, the points in the Q-Q plot will roughly align along a straight line (typically a 45-degree line through the origin).

Use of Q-Q Plot in Linear Regression:

In the context of linear regression, a Q-Q plot is primarily used to check the **normality of residuals**. This is important because one of the key assumptions of linear regression is that the residuals (the differences between the observed and predicted values) are normally distributed.

Importance of Q-Q Plot in Linear Regression:

1. **Validating Model Assumptions**:
   - The linear regression model assumes that the residuals are normally distributed. A Q-Q plot provides a visual check for this assumption. If the residuals are not normally distributed, it can affect the validity of confidence intervals, hypothesis tests, and predictions.
2. **Identifying Outliers and Deviations**:
   - The Q-Q plot can reveal outliers and systematic deviations from normality, such as skewness (asymmetry in the data) or kurtosis (heaviness of the tails).
3. **Improving Model Accuracy**:
   - By detecting non-normality, a Q-Q plot can guide you to transform variables, apply different regression techniques, or use robust methods that do not rely on the normality assumption.
4. **Residual Analysis**:
   - Alongside other residual plots (e.g., residuals vs. fitted values, residuals vs. predictors), the Q-Q plot is an essential part of the residual analysis that helps diagnose and improve linear regression models.

Example of Q-Q Plot Interpretation:

- **Normality (Good Fit)**: If the Q-Q plot shows a straight line, this suggests that the residuals follow a normal distribution, supporting the use of linear regression and validating the results.
- **Heavy Tails**: If the points deviate upward or downward at the ends (tails), it indicates heavy tails in the distribution (i.e., more extreme values than expected in a normal distribution).
- **Skewness**: If the points curve away from the line, it suggests that the residuals are skewed. For instance, a right-skew would show points deviating below the line on the left side and above the line on the right side.

Conclusion:

The Q-Q plot is a crucial diagnostic tool in linear regression for checking the normality of residuals. It helps ensure that the assumptions underlying linear regression are met, leading to more reliable model estimates and predictions. If the Q-Q plot reveals significant deviations from normality, it may indicate the need for model adjustments or the use of alternative statistical methods.