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?

The count of the total rental bike is dependent on humidity, season, atemp, weekday and working day.

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

Using drop_first=True during dummy variable creation is important to avoid multicollinearity issues in regression analysis and to ensure proper interpretation of the regression coefficients. When creating dummy variables for categorical variables with multiple levels, dropping the first level (category) removes redundant information.

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

Registered had 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?

The assumption of Linear Regression after building the model on the training set is done using residual analysis and by checking normality of the residuals.

Residual analysis is a crucial step in validating the assumptions of a linear regression model and evaluating its performance. It involves examining the residuals, which are the differences between the observed values and the predicted values generated by the regression model. Residual analysis helps assess whether the model adequately captures the relationships between the independent variables and the dependent variable, and whether the assumptions of linear regression are met. Assessing the normality of residuals is crucial because violating this assumption can lead to biased coefficient estimates, incorrect standard errors, and unreliable statistical inferences.

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

Humidity, Working day and winter season are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Question

1. Explain the linear regression algorithm in detail.

Linear regression is a fundamental statistical technique used for modeling the relationship between a dependent variable (target) and one or more independent variables (features). It assumes that there exists a linear relationship between the independent variables and the dependent variable.

The objective of linear regression is to fit a linear equation to the observed data points, minimizing the difference between the predicted values and the actual values (residuals). In simple linear regression, with one independent variable and one dependent variable, the linear relationship between them is represented as:

$y = \beta 0 + \beta 1 * x + \varepsilon$

y is the dependent variable

x is the independent variable

β0 is the y-intercept (the value of y when x is 0).

β1 is the slope of the line (the change in y for a unit change in x).

ε is the error term representing the difference between the observed and predicted values.

The coefficients (β0 and β1) are estimated from the training data using an optimization algorithm such as Ordinary Least Squares (OLS). OLS minimizes the sum of squared differences between the observed and predicted values. During the training phase, the linear regression model learns the coefficients (β0 and β1) that best fit the training data. This involves finding the values of the coefficients that minimize the residual sum of squares (RSS) or mean squared error (MSE). After training, the model's performance is evaluated using metrics such as R-squared. Once trained, the model can be used to make predictions on new or unseen data.

Assumptions:

1. Linearity: The relationship between the independent and dependent variables is linear.
2. Independence: The residuals (errors) are independent of each other.
3. Homoscedasticity: The variance of the residuals is constant across all levels of the independent variables.

4. Normality: The residuals follow a normal distribution.
5. No Multicollinearity: The independent variables are not highly correlated with each other

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics but appear very different when graphed. It was created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data and the limitations of relying solely on summary statistics.

Anscombe's quartet consists of four datasets, each containing 11 pairs of x and y values. Despite their small size, these datasets were carefully constructed to have similar summary statistics but different underlying relationships between the variables.

Description of the datasets:

- Dataset I: It has a linear relationship between x and y.
- Dataset II: It also has a linear relationship between x and y, but with an outlier that significantly affects the regression line.
- Dataset III: It has a non-linear relationship between x and y, resembling a quadratic curve.
- Dataset IV: It consists of an apparent relationship between x and y, except for one point that completely alters the regression line.

Despite the differences in the underlying relationships, all four datasets have nearly identical summary statistics:

- Mean of x: 9.0
- Variance of x: 11.0
- Mean of y: 7.5
- Variance of y: 4.12
- Correlation between x and y: Approximately 0.816 (in each dataset)
- Linear regression line: y = 3 + 0.5x

When plotted, the four datasets illustrate the importance of visualizing data. While their summary statistics are the same, the actual data points and their relationships are vastly different when viewed graphically. For example:

- Dataset I show a clear linear relationship.
- Dataset II appears linear but is heavily influenced by an outlier.
- Dataset III exhibits a non-linear relationship.
- Dataset IV has no apparent relationship until the outlier is considered.

Anscombe's quartet highlights the limitations of relying solely on summary statistics such as means, variances, and correlations. Even when these statistics are identical, the underlying data can vary significantly. Visual examination of data through plotting is essential for understanding patterns, relationships, and outliers.

3. What is Pearson’s R?

Pearson's correlation coefficient, often denoted as Pearson's R or simply as r, is a measure of the linear relationship between two variables. It quantifies the strength and direction of the linear association between two continuous variables.

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

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Range: Pearson's R ranges from -1 to 1.

- A value of 1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable increases proportionally.
- A value of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.
- A value of 0 indicates no linear relationship between the variables.

Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations.

Pearson's R is symmetric, meaning that the correlation between variables x and y is the same as the correlation between variables y and x.

It is affected by outliers in the data, particularly in small sample sizes.

Pearson's R measures only linear relationships. It may not capture non-linear relationships between variables.

The magnitude of Pearson's R indicates the strength of the relationship:

- Close to 1 or -1: Strong linear relationship.
- Close to 0: Weak or no linear relationship.

- The sign of Pearson's R (+ or -) indicates the direction of the relationship:
- Positive: Both variables increase or decrease together.
- Negative: One variable increase as the other decreases.

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

Scaling is a preprocessing step used in machine learning and data analysis to transform the features (variables) of a dataset to a similar scale. The purpose of scaling is to ensure that all features contribute equally to the analysis and to improve the performance and convergence of certain machine learning algorithms.

The reason behind doing scaling:

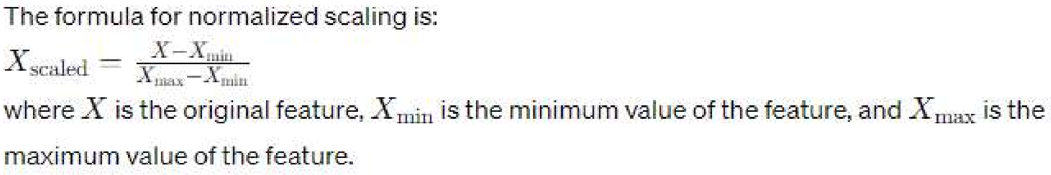
Equal Importance: Many machine learning algorithms, such as k-nearest neighbors (KNN), support vector machines (SVM), and neural networks, compute distances or similarities between data points. If the features are on different scales, features with larger scales may dominate the algorithm's computations, leading to biased or suboptimal results.

Faster Convergence: Gradient-based optimization algorithms, such as gradient descent, converge faster when features are on a similar scale. Unequal scales can cause the optimization process to take longer to reach the minimum or maximum of the objective function.

Regularization: Regularization techniques, such as ridge regression and lasso regression, penalize the magnitude of coefficients. Scaling the features ensures that the regularization penalty is applied uniformly across all features.

Normalized Scaling:

Normalized scaling (also known as min-max scaling) rescales the features to a fixed range, usually between 0 and 1. It preserves the original distribution of the data while ensuring that all features have the same scale. Normalized scaling is sensitive to outliers because it uses the range of the data. Formula:



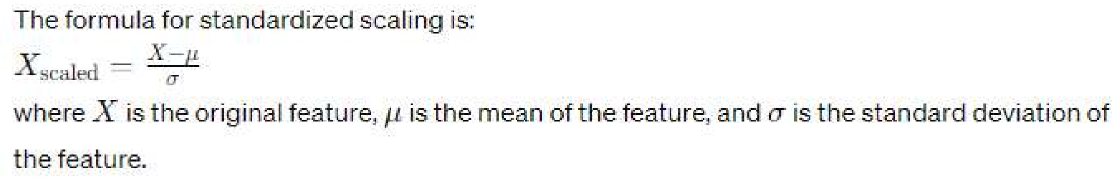
The formula for normalized scaling is:

$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$

where X is the original feature, $X_{\min}$ is the minimum value of the feature, and $X_{\max}$ is the maximum value of the feature.

Standardized Scaling:

Standardized scaling (also known as z-score normalization) standardizes the features to have a mean of 0 and a standard deviation of 1. It transforms the data to have a standard normal distribution. Standardized scaling is robust to outliers because it uses the mean and standard deviation, which are less affected by outliers compared to the range. Formula:



The formula for standardized scaling is:

$X_{\text{scaled}} = \frac{X-\mu}{\sigma}$

where X is the original feature, μ is the mean of the feature, and σ is the standard deviation of the feature.

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

Yes, sometimes the value of the Variance Inflation Factor (VIF) can become infinite. This occurs when one or more independent variables in a regression model are perfectly linearly dependent on other independent variables. In other words, one variable can be expressed as a perfect linear combination of other variables in the model.

VIF measures the degree of multicollinearity among the independent variables in a regression model. It is calculated as the ratio of the variance of the estimated coefficient of a variable in the presence of multicollinearity to the variance of the coefficient when there is no multicollinearity. When one independent variable is a perfect linear combination of other independent variables, it results in perfect multicollinearity. This means that the variance of the estimated coefficient for the redundant variable becomes zero because its relationship with other variables can be precisely determined without any error. In the calculation of VIF, when the variance of the coefficient for a variable becomes zero due to perfect multicollinearity, dividing by these zero variance results in an infinite VIF value.

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

A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to compare the distribution of a sample of data to a theoretical distribution, typically the normal distribution. It helps in assessing whether the data follows a particular distribution or to identify deviations from that distribution.

Use of Q-Q Plot:

Assessing Normality: One common use of a Q-Q plot is to assess whether a set of data follows a normal distribution. In a Q-Q plot, the quantiles of the sample data are plotted against the quantiles of a theoretical normal distribution. If the points fall approximately along a straight line, it suggests that the sample data is normally distributed.

Identifying Distributional Differences: Q-Q plots can also be used to compare the distribution of the sample data to other theoretical distributions, such as the uniform distribution or the exponential distribution. Deviations from the straight line indicate differences between the sample distribution and the theoretical distribution.

Importance in Linear Regression:

Residual Analysis: In linear regression, Q-Q plots are often used to assess the normality of the residuals. Residuals are the differences between the observed values and the values predicted by the regression model. Normality of residuals is an assumption of linear regression, and violations of this assumption can affect the validity of statistical inference and predictions.

Diagnostic Tool: Q-Q plots serve as a diagnostic tool to identify departures from normality in the residuals. If the residuals follow a normal distribution, the points in the Q-Q plot will form a straight line. Deviations from the straight line indicate non-normality, which may suggest the need for further investigation or transformation of the data.

Assumption Checking: Assessing the normality of residuals through Q-Q plots is an essential step in checking the assumptions of linear regression. Along with other diagnostic plots and tests, Q-Q plots help ensure that the assumptions of linearity, independence, homoscedasticity, and normality are met.