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?

There are three types of Coefficients:

To infer the effect of categorical variables on the dependent variable (cnt), you can analyze the coefficients of these variables from the linear regression model.

Positive Coefficients: A positive coefficient for a categorical variable indicates that the presence of that category increases the dependent variable (cnt).

Negative Coefficients: A negative coefficient for a categorical variable indicates that the presence of that category decreases the dependent variable (cnt).

Magnitude of Coefficients: The magnitude of the coefficients indicates the strength of the effect. Larger absolute values suggest a stronger impact on the dependent variable.

Equation of the linear regression model:

cnt = 0.2220 + (-0.0435 * yr) + (0.0052 * mnth) + (0.4183 * holiday) + (0.0970 * weekday) + (-0.1798 * workingday) + (-0.1836 * weathersit) + (0.0614 * temp) + (0.0341 * atemp) + (0.0939 * hum) + (0.1331 * windspeed) + (-0.0435 * season_2) + (0.0052 * season_3) + (-0.0313 * season_4) + (-0.0922 * yr_1) + (0.0614 * mnth_2) + (0.0341 * mnth_3) + (0.0939 * mnth_4) + (0.1331 * mnth_5) + (-0.0435 * mnth_6) + (0.0052 * mnth_7) + (-0.0313 * mnth_8) + (-0.0922 * mnth_9)

Conclusion

From the analysis of the coefficients, you can infer that:

- Rentals tend to increase in spring, summer, and fall compared to winter.
- The year 2019 saw a significant increase in rentals compared to 2018.
- Holidays have a positive impact on the count of rentals.

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

The **drop_first=True** is used in dummy variable creation to avoid the **dummy variable trap**.

Consider a categorical variable season with four categories: winter, spring, summer, and fall. When creating dummy variables without drop_first=True, you get four dummy variables:

- season_winter
- season_spring
- season_summer
- season_fall

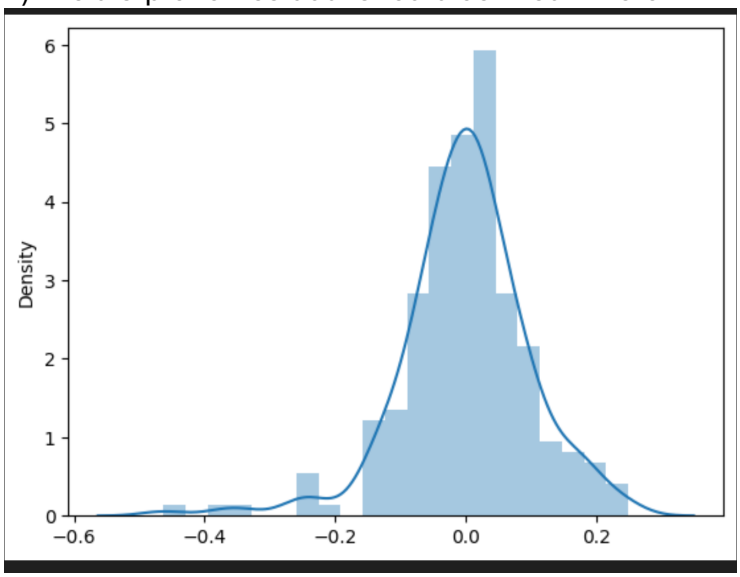
However, these four dummy variables are perfectly multicollinear because knowing the values of any three of them allows you to determine the value of the fourth. For example, if season_winter, season_spring, and season_summer are all 0, then season_fall must be 1.

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

Temperature (temp) has the highest correlation with the target variable (cnt), we could see pair-plot to check the relationships and a correlation matrix to quantify the correlations.

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

1) The distplot of residual should centred in zero



2) We can check R2 value for Test data set using the model created using train data.

The train data R-Square value 0.84
The test data R-Square value 0.78

Multicollinearity: The VIF values should be less than 10, indicating no severe multicollinearity.

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

The top three features which contributing demand of the shared bikes 1) Temperature 2) Season 3) Year

General Subjective Questions

1)Explain the linear regression algorithm in detail.

Linear regression is a method to find the relationship between a dependent variable (target) and one or more independent variables (features).

Key Concepts

1. **Dependent Variable (y)**
2. **Independent Variables (X)**
3. **Linear Equation**: The equation that represents the relationship between the dependent and independent variables.

Simple Linear Regression

In simple linear regression, we have one independent variable. The relationship is modeled using the following linear equation:

[y = beta_0 + beta_1 x + epsilon]

Where:

- (y) is the dependent variable.
- (x) is the independent variable.
- (\beta_0) is the intercept (the value of (y) when (x) is 0).
- (\beta_1) is the slope (the change in (y) for a one-unit change in (x)).
- (\epsilon) is the error term (the difference between the observed and predicted values).

Multiple Linear Regression

In multiple linear regression, we have more than one independent variable. The relationship is modeled using the following linear equation:

[y = beta_0 + beta_1 x1 + beta_2 x2 + ... + beta_n xn + epsilon]

Where:

- (y) is the dependent variable.
- (x1, x2, ... , xn) are the independent variables.
- (beta_0) is the intercept.
- (beta_1, beta_2, ..., beta_n) are the coefficients for the independent variables.
- (\psilon) is the error term.

Steps in Linear Regression

1. **Data Collection**
2. **Data Preprocessing**
3. **Splitting the Data**
4. **Model Training**
5. **Model Evaluation**
6. **Prediction**

Model Training

The goal of training a linear regression model is to find the best-fitting line that minimizes the sum of the squared differences between the observed and predicted values.

Ordinary Least Squares (OLS)

OLS is a method to estimate the coefficients ((beta_0, beta_1, ..., beta_n)) by minimizing the sum of the squared residuals.

2)Explain the Anscombe's quartet in detail.

Anscombe's quartet is a collection of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. It was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and to show how statistical properties can be misleading when not visualized.

Key Points of Anscombe's Quartet

1. **Identical Descriptive Statistics**:
    - Each dataset in the quartet has the same mean, variance, correlation coefficient, and linear regression line.
    - Despite these similarities, the datasets are very different when plotted.
2. **Importance of Visualization**:
    - Anscombe's quartet highlights the necessity of visualizing data to uncover patterns, relationships, and anomalies that are not apparent from summary statistics alone.

3) What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. It quantifies the degree to which two variables are linearly related. The value of Pearson's R ranges from -1 to 1, where:

- **1** indicates a perfect positive linear relationship.
- **-1** indicates a perfect negative linear relationship.
- **0** indicates no linear relationship.

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

Scaling is a data preprocessing technique used to adjust the range of features in a dataset. It ensures that all features contribute equally to the model, preventing features with larger ranges from dominating the learning process. Scaling is particularly important for algorithms that rely on distance calculations, such as k-nearest neighbors (KNN) and support vector machines (SVM).

Why is Scaling Performed?

1. **Improves Model Performance**: Many machine learning algorithms perform better when features are on a similar scale.
2. **Speeds Up Convergence**: Gradient-based optimization algorithms, such as gradient descent, converge faster when features are scaled.
3. **Prevents Dominance**: Features with larger ranges can dominate the learning process, leading to biased models.
4. **Ensures Fair Contribution**: Scaling ensures that all features contribute equally to the model.

Types of Scaling

Normalization (Min-Max Scaling):

Normalization rescales the features to a fixed range, usually [0, 1]
The transformed data will be within the range [0, 1] (or any other specified range).
Useful when you know the minimum and maximum values of the features.

Standardization (Z-score Scaling)
Standardization rescales the features to have a mean of 0 and a standard deviation of 1.
Useful when the data contains outliers.
Centers the data around 0 and scales it based on the standard deviation, but does not necessarily make the data normally distributed.

Conclusion

- **Normalization** rescales the data to a fixed range, typically [0, 1], and is sensitive to outliers.
- **Standardization** rescales the data to have a mean of 0 and a standard deviation of 1, making it less sensitive to outliers but still affected by them.

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

A VIF value can become infinite (or extremely large) due to perfect multicollinearity, which occurs when one predictor variable is a perfect linear combination of one or more other predictor variables. This means that the predictor can be exactly predicted from the other predictors, leading to an undefined or infinite VIF.

Causes of Infinite VIF

- Perfect Multicollinearity:
- Dummy Variable Trap:
- Redundant Variables:

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

A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, such as the normal distribution. It plots the quantiles of the dataset against the quantiles of the theoretical distribution. If the data follows the theoretical distribution, the points on the Q-Q plot will lie approximately along a straight line.

Use and Importance of a Q-Q Plot in Linear Regression

In the context of linear regression, a Q-Q plot is primarily used to assess the normality of the residuals (errors). The assumptions of linear regression include that the residuals are normally distributed. Checking this assumption is crucial because:

1. **Validity of Statistical Tests**: Many statistical tests, such as t-tests for coefficients, assume that the residuals are normally distributed. If this assumption is violated, the results of these tests may not be valid.

2. **Model Diagnostics**: Non-normal residuals can indicate issues with the model, such as missing variables, incorrect functional form, or outliers.
3. **Confidence Intervals and Predictions**: The accuracy of confidence intervals and prediction intervals relies on the normality of residuals.