Linear_Regression_Subjective_Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:

From my analysis of categorical variables like season, weathersit, and workingday, I found that:

- Season significantly affects bike demand, with more rentals during summer and fall compared to winter and spring. This is likely due to better weather conditions during warmer months.
- Weather conditions (weathersit) also play an important role. Clear weather is associated with higher bike rentals, while rainy or misty weather leads to fewer rentals.
- Workingday: Weekdays generally have higher bike rentals, likely because people use bikes for commuting, whereas weekends show slightly lower rentals.

---

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

Answer:

Using drop_first=True during dummy variable creation helps avoid the dummy variable trap, which occurs when multicollinearity is introduced into the model because of redundant information. Dropping the first category ensures that the model does not suffer from perfect multicollinearity, making the regression coefficients interpretable and the model more stable.

---

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

Answer:

From the pair plot, registered users had the highest correlation with the target variable cnt (total bike rentals), as it directly contributes to the total number of rentals. Among independent variables, **temperature (temp)** showed a strong positive correlation with the target variable.

---

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

Answer:

I validated the assumptions of linear regression by:

- Linearity: I checked for a linear relationship between features and the target variable through scatter plots and the high R-squared value.
- Normality of residuals: I plotted the residuals and checked for normal distribution around zero.
- Multicollinearity: I calculated the VIF (Variance Inflation Factor) to detect multicollinearity. Any variables with high VIF values were either dropped or adjusted.

---

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

Answer:

Based on the final model, the top 3 contributing features were:

1. Temperature (temp): Higher temperatures led to more bike rentals.
2. Year (yr): As bike-sharing gained popularity over time, more rentals were observed in later years.
3. Season (summer): Summer months showed a significant positive effect on bike rentals.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal is to fit a line that best predicts the dependent variable based on the input features. The equation for simple linear regression is:

$Y = \beta_0 + \beta_1 X + \varepsilon$

Where:

- Y is the dependent variable (what we're predicting),
- X is the independent variable (input),
- β_0 is the intercept (value of Y when X is 0),
- β_1 is the slope of the line (change in Y for a unit change in X),
- ε is the error term.

The algorithm minimizes the sum of squared errors (SSE) to find the best-fit line. In the case of multiple linear regression, the equation extends to:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$

---

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's Quartet is a collection of four datasets that have nearly identical statistical properties, yet look very different when graphed. It was constructed by the statistician **Francis Anscombe** in 1973 to demonstrate the importance of graphing data before analyzing it, as relying solely on summary statistics can be misleading.

Statistical Properties of Anscombe's Quartet

Each of the four datasets in the quartet has:

- The same **mean** of both x and y.
- The same **variance** of both x and y.
- The same **correlation coefficient** between x and y (≈ 0.816).
- The same **linear regression** equation ($y = 3 + 0.5x$).
- The same **coefficient of determination** ($R^2 \approx 0.67$).

Despite these similarities, the datasets are quite different when plotted. Here's a detailed explanation of each dataset:

Dataset 1

- A typical dataset that fits well with the linear regression model. There are no outliers or patterns that suggest a need for caution.

Dataset 2

- A dataset where the x-values are all the same except for one point, which gives the impression of a strong linear relationship. However, this relationship is driven by just one outlier.

Dataset 3

- This dataset consists of data points that follow a nonlinear relationship. A linear regression model is not appropriate here, despite the summary statistics suggesting otherwise.

Dataset 4

- This dataset is extreme in that almost all of the x-values are identical, with one outlier distorting the correlation. A simple graph would reveal that a linear model is not suitable.

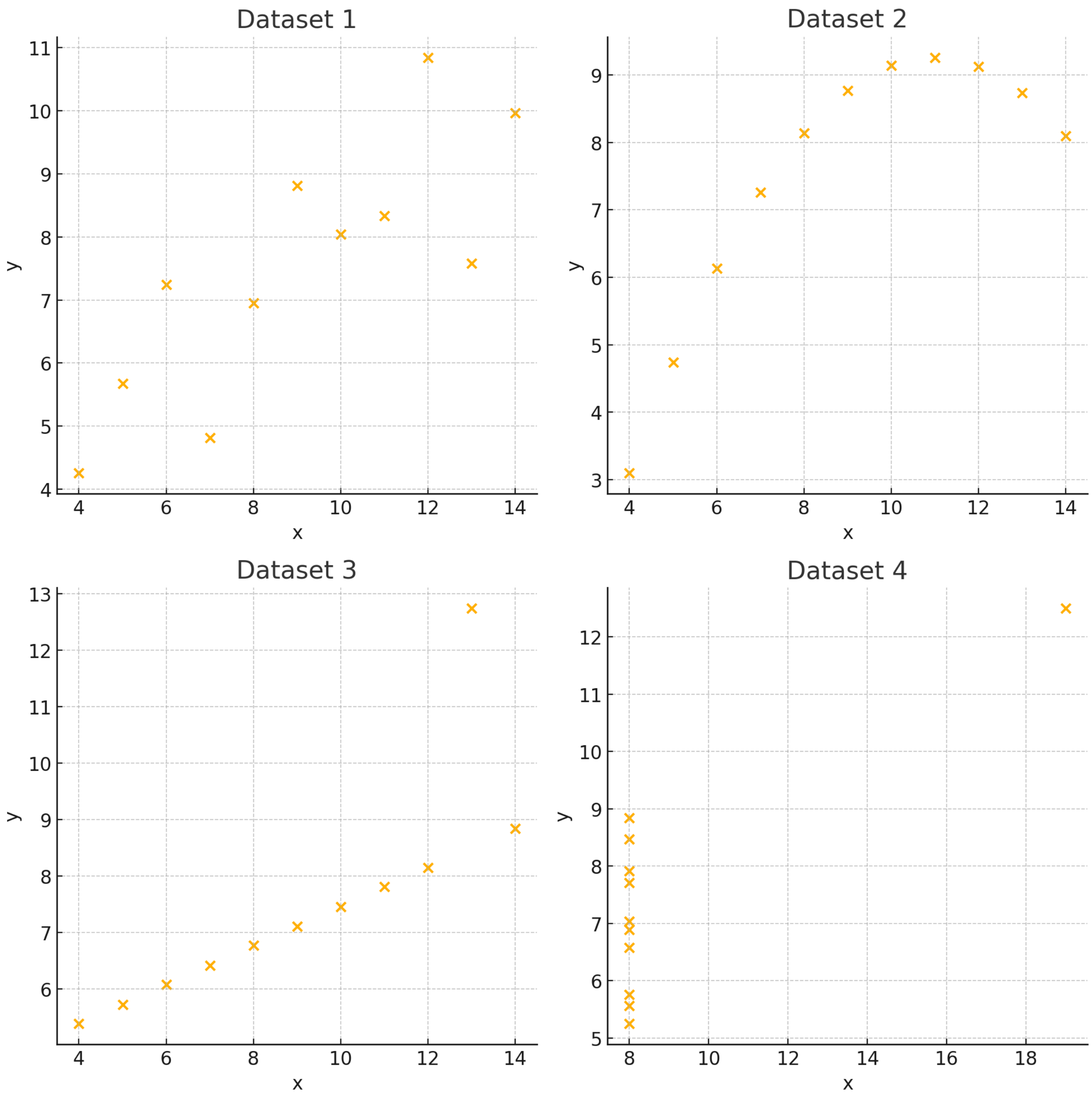
Why is Anscombe's Quartet Important?

It illustrates that relying solely on summary statistics like the mean, variance, and correlation coefficient can be misleading. Without plotting data, we might miss key insights, such as patterns, clusters, or outliers that significantly affect the interpretation of the data.

Here is the visual representation of Anscombe's Quartet. As you can see, despite having similar statistical properties, each dataset behaves very differently when graphed. This highlights the importance of visualizing data rather than relying solely on summary statistics for interpretation.

- **Dataset 1** shows a clear linear trend.
- **Dataset 2** has one outlier that strongly influences the correlation.
- **Dataset 3** follows a non-linear pattern.
- **Dataset 4** shows an outlier that distorts the appearance of the linear relationship.

Each of these datasets offers insights that would be missed if we looked only at the summary statistics like the mean, variance, or regression line



3. What is Pearson’s R?

Answer:

Pearson’s R, also known as the Pearson correlation coefficient, measures the strength and direction of the linear relationship between two variables. The value 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.

Pearson’s R is given by:

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

r = correlation coefficient
x_i = value of x-variable in a sample
x^ = mean of values of the x-variable
y_i = values of the y-variable in a sample
y^ = mean of the values of y variable.

---

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

Answer:

Scaling is the process of transforming data so that it fits within a particular range, which is especially important when the range of values in a dataset varies greatly. It ensures that each feature contributes equally to a model’s performance by adjusting the numerical range of the data without distorting differences in the data’s range. Many machine learning algorithms perform better when features are on a similar scale.

Why is Scaling Performed?

1. **Improve Model Performance**: Some machine learning algorithms (like k-nearest neighbors, support vector machines, and gradient descent-based algorithms like logistic and linear regression) are sensitive to the scale of the data.
2. **Faster Convergence**: Algorithms like gradient descent converge faster when features are scaled, as large variations in feature magnitudes can lead to slow optimization.
3. **Avoid Dominance of Large Values**: Features with large ranges can dominate the learning process, leading to biased models if not scaled.

Two Common Scaling Techniques: Normalization and Standardization

Aspect	**Normalization**	**Standardization**
Definition	Adjusting data to a specific range, usually [0, 1].	Centering data around the mean with unit variance.
Range of Transformed Data	Data is rescaled to a range between [0, 1] or [-1, 1].	No fixed range, but typically, the data is centered around 0 with a standard deviation of 1.
Sensitive to Outliers?	Yes, outliers can significantly impact the range.	Less sensitive to outliers since it uses mean and standard deviation.
When to Use?	Useful when the data is bounded (e.g., pixel values, or when features vary within a limited range).	Useful when the data follows a Gaussian distribution or when you want all features to have comparable distributions.

---

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

Answer:

An infinite VIF value occurs when there is **perfect multicollinearity** between the independent variables, meaning one variable is a perfect linear combination of others. This causes issues because the model cannot differentiate between the collinear variables. To avoid this, such variables should be removed or combined.

---

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

Answer:

A **Q-Q (Quantile-Quantile) plot** is used to compare the distribution of residuals to a normal distribution in linear regression. It plots the quantiles of the residuals against the quantiles of a theoretical normal distribution. If the residuals are normally distributed, the points in the Q-Q plot will lie approximately along a straight line. This is important for validating the normality assumption of residuals in linear regression.