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?

   Higher average bike rentals in summer compared to winter and spring.
   Higher average bike rentals in August, May and September compared to other months.
   Higher average bike rentals in mist weather compared to other weather.

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

   By setting drop_first=True in functions like pd.get_dummies(), you drop the first dummy variable, effectively removing the redundant information. This prevents multicollinearity and ensures that your model can be estimated accurately.

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

   Yr , temp, windspeed and holiday showing highest correlation.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   Based on Variance Inflation factor, Residual Analysis , R-squared

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Yr, if we increase the yr then Demand will increase by 0.2504
   In Jun month, demand gets increase by 0.0790
   windspeed, if we increase the windspeed then Demand will decrease by 0.1549

General Subjective Questions

1. Explain the linear regression algorithm in detail
   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 find the best-fitting line that minimizes the distance between the predicted values and the actual values.

   The Model
   The linear regression model is represented by the equation:
   $y = \beta0 + \beta1*x1 + \beta2*x2 + ... + \beta n*xn + \varepsilon$

   Where:

- y is the dependent variable (the value we want to predict)
- x1, x2, ..., xn are the independent variables (the features)
- β0 is the intercept (the value of y when all x's are zero)
- β1, β2, ..., βn are the coefficients (weights) for each independent variable
- ε is the error term (the difference between the predicted and actual values)

The Process

I. **Data Collection:** Gather data with the dependent variable and one or more independent variables.

II. **Data Preparation:** Clean and preprocess the data, handling missing values, outliers, and feature scaling if necessary.

III. **Model Fitting:** Find the values of the coefficients (β0, β1, β2, ...) that minimize the sum of the squared errors (residuals) between the predicted and actual values. This is often done using the Ordinary Least Squares (OLS) method.

IV. **Evaluation:** Assess the model's performance using metrics like R-squared, adjusted R-squared, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is a set of four small datasets that have nearly identical statistical properties (mean, variance, correlation, and linear regression line), yet when visualized, look quite different.

The Purpose

This quartet was created by Francis Anscombe to illustrate the importance of **graphical data analysis** before performing statistical analysis. It highlights the potential pitfalls of relying solely on summary statistics and the necessity of visualizing data to uncover underlying patterns, outliers, and other important features.

Anscombe's Quartet emphasizes the importance of:

- **Exploratory Data Analysis (EDA):** Visualizing data to understand its characteristics before applying statistical methods.
- **Caution with summary statistics:** Using summary statistics alone can be misleading.
- **Considering different visualization techniques:** Multiple plots (scatter plots, histograms, box plots) can provide different insights.

3. What is Pearson's R?

Pearson's correlation coefficient (r) is a statistical measure that quantifies the linear relationship between two continuous variables. It ranges from -1 to 1.

Interpretation of r:

- **r = 1:** Perfect positive correlation (as one variable increases, the other increases proportionally).
- **r = -1:** Perfect negative correlation (as one variable increases, the other decreases proportionally).
- **r = 0:** No linear correlation between the variables.

Key points:

- **Measures linear relationship:** Pearson's correlation only captures linear relationships. Non-linear relationships might not be detected.
- **Sensitive to outliers:** Outliers can significantly impact the correlation coefficient.
- **Correlation does not imply causation:** A high correlation between two variables doesn't necessarily mean one causes the other.
- **Formula to calculate**
  $r=\sum(xi-\bar{}x)(yi-\bar{}y)\sqrt{\sum(xi-\bar{}x)2\sum(yi-\bar{}y)2}$

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

   What - Scaling is a technique used to transform numerical features into a common scale. This is essential when features have different ranges, units, or magnitudes. By scaling, we ensure that all features contribute equally to the model's learning process.

   Why-

- **Improves Algorithm Performance:** Many machine learning algorithms, especially distance-based algorithms (like K-Nearest Neighbors) and gradient-descent-based algorithms (like Linear Regression), converge faster and produce better results when features are on a similar scale.
- **Prevents Dominance of Features:** Features with larger magnitudes can dominate the learning process, leading to biased models. Scaling ensures that all features are given equal importance.
- **Enhances Interpretability:** Scaled features make it easier to compare the relative importance of different features.

Normalized Scaling vs. Standardized Scaling

- **Normalized Scaling (Min-Max Scaling):**
  - Rescales features to a specific range, typically between 0 and 1.
  - Formula: (x - min(x)) / (max(x) - min(x))
  - Suitable when you know the exact minimum and maximum possible values and want to preserve the original distribution shape.
- **Standardized Scaling (Z-score Scaling):**
  - Rescales features to have a mean of 0 and a standard deviation of 1.
  - Formula: (x - mean(x)) / std(x)
  - More robust to outliers and preserves the shape of the original distribution.

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

   A VIF (Variance Inflation Factor) of infinity indicates perfect multicollinearity. This means one or more of your independent variables is a perfect linear combination of other independent variables.

- **Duplicate Variables:** Two or more variables contain identical information.
- **Linear Combinations:** One variable is a linear combination of other variables (e.g., X3 = 2*X1 + X2).
- **Dummy Variable Trap:** Including all levels of a categorical variable without omitting one leads to perfect multicollinearity.

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

   A Q-Q plot (Quantile-Quantile plot) is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

   Use in Linear Regression

   In the context of linear regression, the Q-Q plot is primarily used to assess the normality of the residuals (the differences between the observed and predicted values).

   Why is it important?

- **Normality of residuals** is one of the key assumptions of linear regression.
- A normal distribution of residuals indicates that the model is appropriate for the data.
- Deviations from normality can affect the reliability of statistical inferences.

How to Interpret a Q-Q Plot:

- If the data points closely follow a straight line, it suggests that the data is normally distributed.
- Deviations from the line indicate departures from normality.
- Outliers or heavy tails can be identified visually.