Bike Sharing Subjective Questions

Submitter: G Raghu Vamshi

Assignment-based Subjective Questions and Answers:

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

Answer: Based on the analysis with target_variable as cnt (Count of Total Rental bikes):

- The season of the summer and fall has high median where bike rental has increased.
- On year 2019, the bike rental has increased than 2018
- The month of may-oct has high median where bike rental has increased.
- The bike rental on holiday is less than the working days.
- There is no bike rental impact based on weekday.
- The weather of 'Mist & Cloudy' and 'Clear' has more bike rentals.

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

Answer: It is important to use the **drop_first=True** where it helps to reduce the extra column created during dummy variable creation.

Example:

As per the data, We have 4 seasons (spring, summer, fall, winter), while we create the dummy variables it is enough to 3 seasons instead of having 4 seasons as dummy variable. The reason it is obvious none of the 3 season is true it is expected to be 4^{th} season, so we no need to waste a column. If we have n-unique value of categorical variable, the n-1 is level is enough for the dummy variable creation.

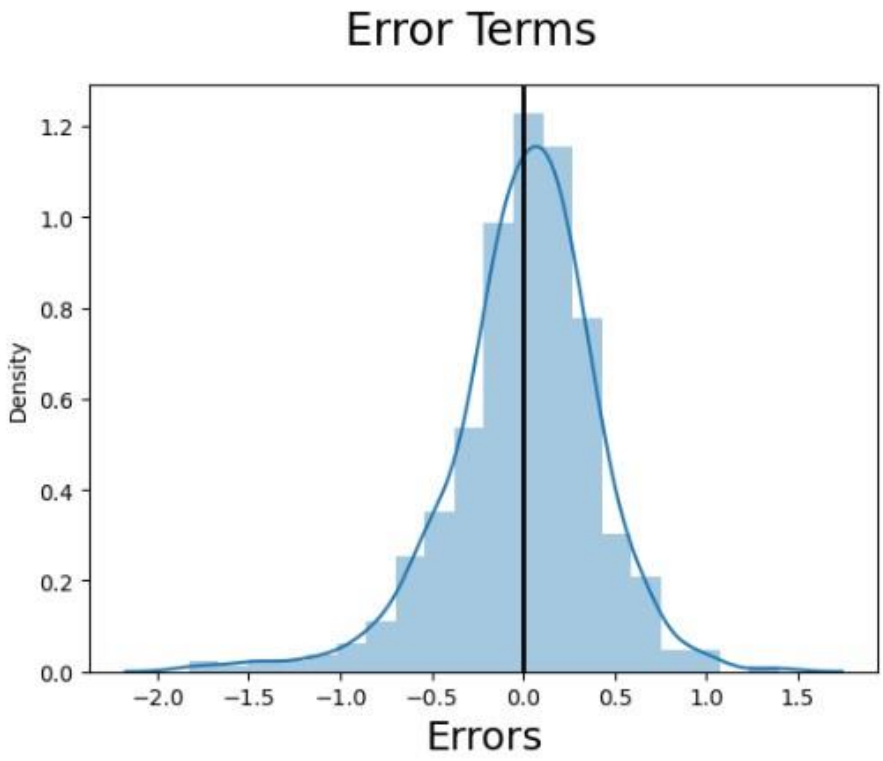
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer: By dropping ‘registered’ and ‘casual’ numerical variable, the ‘atemp’ of 0.63 has the highest correlation with the target_variable ‘cnt’.

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

Answer:

- Validate the linearity using the residual by plotting the distribution where the difference between observed and predicted values. The plot is normally distributed with a mean value = 0.



- Validate the R_Square value for observed vs Predicated value is high to refer high level of correlation, which is around (0.804).
- Examine the Residuals using Q-Q plot and confirm it is normally distributed.

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

Answer: Top 3 features that are contributing significantly:

- Temperature – temp: It has the coefficient of 0.603 refers to increase in temp will increase the bike rentals by 0.603 units.
- Year – yr: It has the coefficient of 0.509 refers to increase of this variable will increase the bike rentals by 0.509 units.
- Season - Season_winter: It has the coefficient of 0.262 refers to increase of this variable will increase the bike rentals by 0.262 units.

General Subjective Questions and Answers:

1. Explain Linear regression algorithm in detail?

Answer: Linear regression algorithm is used to statistical and machine learning techniques which are used for modeling the relationship between a dependent variable (target) and one or more independent variables (features).

- **Types of Regression:**
  - **Simple Linear :** Modeling the relationship between One dependent and One Independent Variable. ○ **Multiple Linear :** Modeling the relationship between One dependent and multiple Independent Variable.
- **Model Training:**
  - It is to process the best fitting linear equation which represents the relationship between the dependent and independent variables.
  - **Steps include** Data Collection, Data Preprocessing, Feature Selection, Training the Model
- **Assumption between the relationship:** Linear, Multicollinearity, Normality
- **Model Evaluation:** R-Square, VIF (Variance Inflation Factor)

2. Explain the Anscombe's quartet in detail?

Answer: Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different.

There are four datasets in Anscombe's quartet:

- Dataset I: A simple Linear relationship.
- Dataset II: A Non-Linear relationship.
- Dataset III: A simple Linear relationship with an outlier.
- Dataset II: More of outlier dataset with no simpler relationship.

3. What is Pearson's R?

Answer: A statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables.

The relationship refers based on the Pearson's r:

- Positive linear relationship when r = 1.
- Negative linear relationship when r = -1.
- No linear relationship when r = -1.

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

   Answer: Scaling is a preprocessing technique used in data analysis and machine learning to transform the data to make it fall within the common scale/range.

   Type of scaling techniques:

   - **Normalized/Min-Max Scaling:** To transform data into specified range, typically (0 and 1).
   - **Standardized/Z-score Scaling:** To standardize the data into standard normal distribution. Transform the data to have the mean of 0 and standard deviation of 1.

5. You might have observed that sometimes that VIF value is infinite. Why does this happen? **Answer:**

   - VIF value = infinite, shows a perfect correlation between two independent variables.
   - **Happens when:**
     i. Multicollinearity problem between the independent variables in the model.
     ii. Perfect correlation where R_square =1, which leads to 1/(1-R2) infinity.

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

   Answer:

Q-Q plot (Quantile – Quantile plot) which contains x-axis represents the quantiles of the theoretical distribution and the y-axis represents the quantiles of the observed data.

Importance:

- To examine the assumption of residuals.
- To examine the outlier detection
- To examine the residuals
- Compare the distribution between the datasets.