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?

Answer:

- Season: Demand for shared bike increases in summer and fall as compared to winter and spring.
- Year: Demand increasing with year after covid.
- Month: Demand is high in June to September.
- Holiday: Demand decreases on holidays.
- Working day: Demand increases on working day.
- Weathersit: Demand increases on a clear weather day.
- Dteday: Demand increases with starting of the month but gradually decreases and get constant for rest of the month.
- Weekday: Demand increase from monday to thursday and decreases gradually through friday and weekends.

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

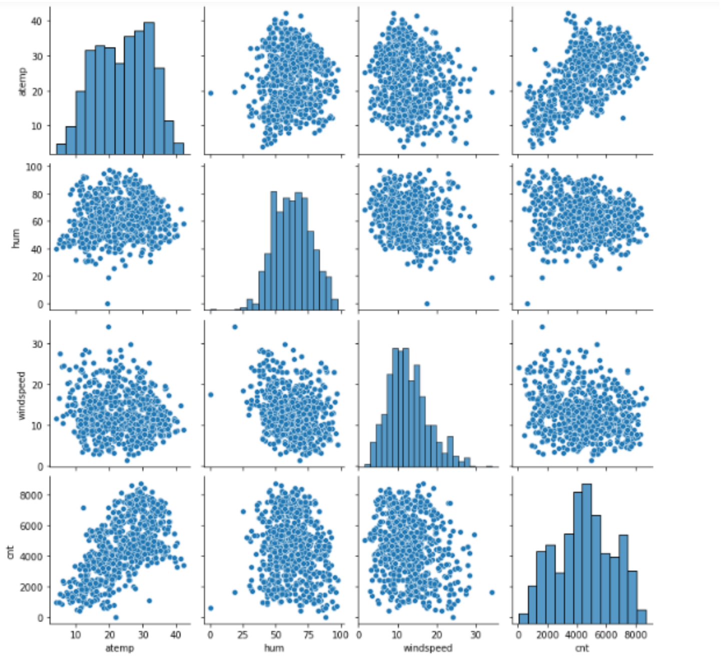
Answer:

- To reduce the dummy variable to n-1 level. It helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Answer:

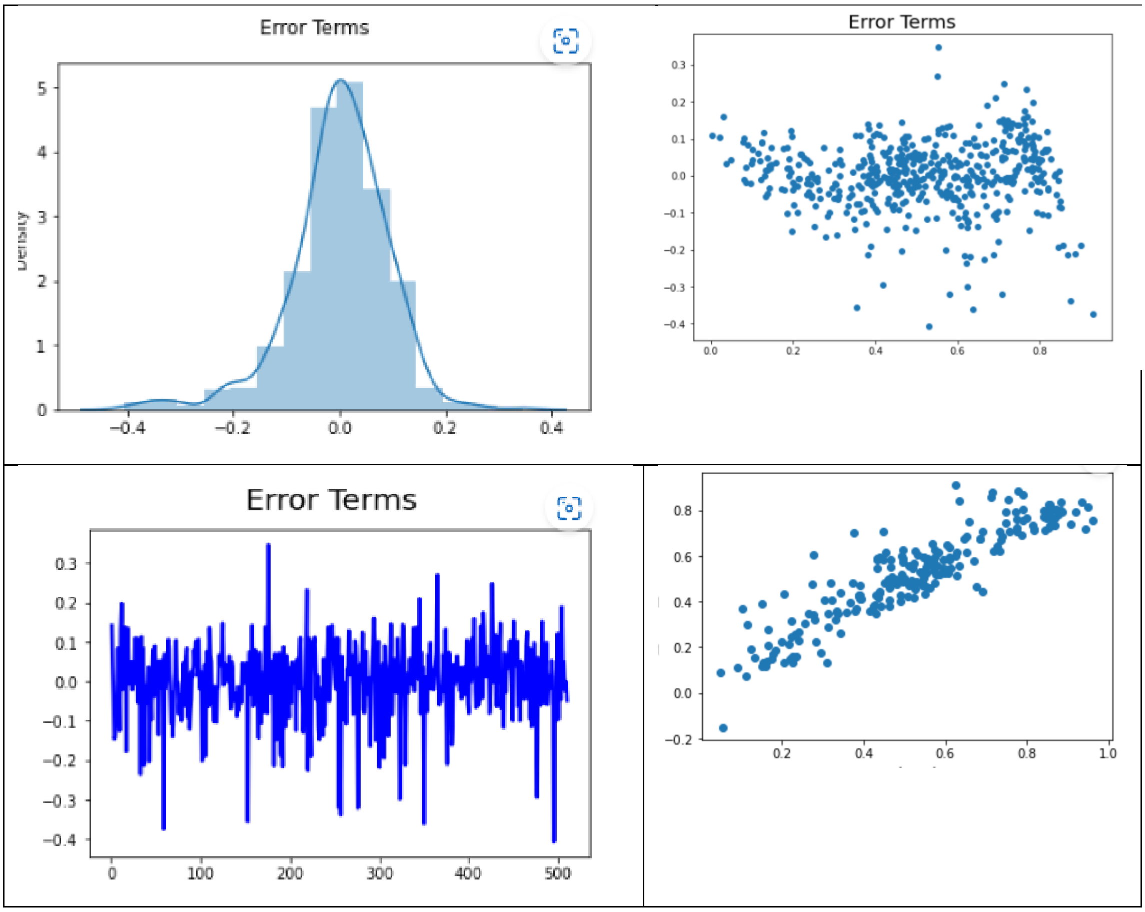
- atemp has the highest correlation with the target variable as can be seen in the figure below.



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

Answer:

- By performing Residual Analysis:



- error terms have normal distribution with mean=0.
- there is no pattern in error terms, which confirms homoscedasticity.
- line plot shows residuals are independent.
- Linearity.

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

Answer:

- atemp
- light rain
- year

General Subjective Questions

1. Explain the linear regression algorithm in detail.

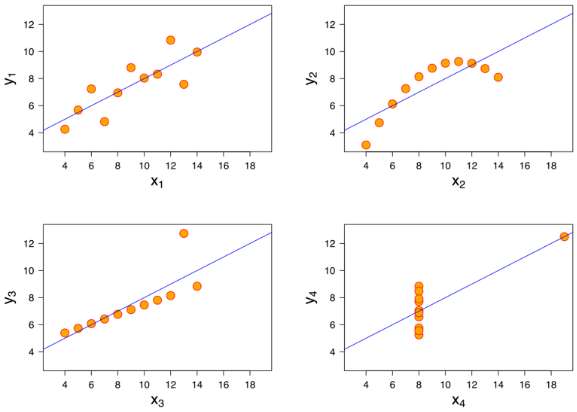
Answer:

- Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.
- By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ1 and θ2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).
- Various libraries used are statsmodel, sklearn and scipy, etc.
- The regression has five key assumptions:
    - Linear relationship
    - Multivariate normality
    - No or little multicollinearity
    - No auto-correlation
    - Homoscedasticity

2. Explain the Anscombe's quartet in detail.

Answer:

- Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.
- It illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.



3. What is Pearson’s R?

Answer:

- In statistics, the Pearson correlation coefficient also known as Pearson's r is a measure of linear correlation between two sets of data.
- It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.
- As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations.

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

Answer:

- It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
- Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
- Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

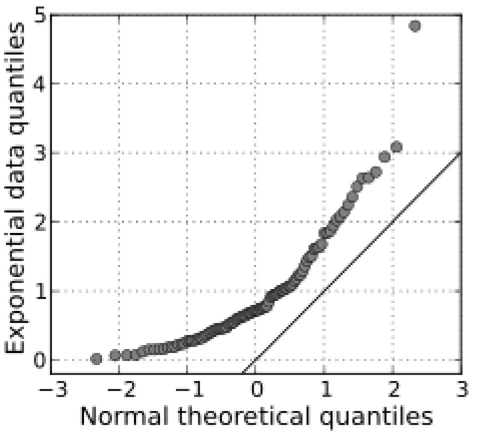
Answer:

- If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
- An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Answer:

- Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.
- A Q-Q plot showing the 45 degree reference line:


Exponential data quantiles
5
4
3
2
1
0
−3
−2
−1
0
1
2
3
Normal theoretical quantiles