LINEAR REGRESSION – SUBJECTIVE QUESTIONS

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? (3 marks)

Answer:

The categorical variables in the given 'day' dataset - season, weathersit, holiday, mnth, yr and weekday were visualized using a boxplot. Their effect on the dependent variable can be summarized as below:

- Season – Boxplot indicated the least value of cnt in spring season whereas fall had maximum value of cnt. Summer and winter had intermediate value of cnt.
- Weathersit – Highest cnt was observed when weather situation was clear partly cloudy. There's no user during heavyrain_ice_thunderstrmmist as the weather indicates it to be highly unfavorable.
- Holiday – Rentals reduced during holiday
- Yr – Number of increased in 2019 when compared to 2018
- Mnth – September month saw highest number of rentals whereas December month saw least primarily due to extreme weather situation as observed in Weathersit.

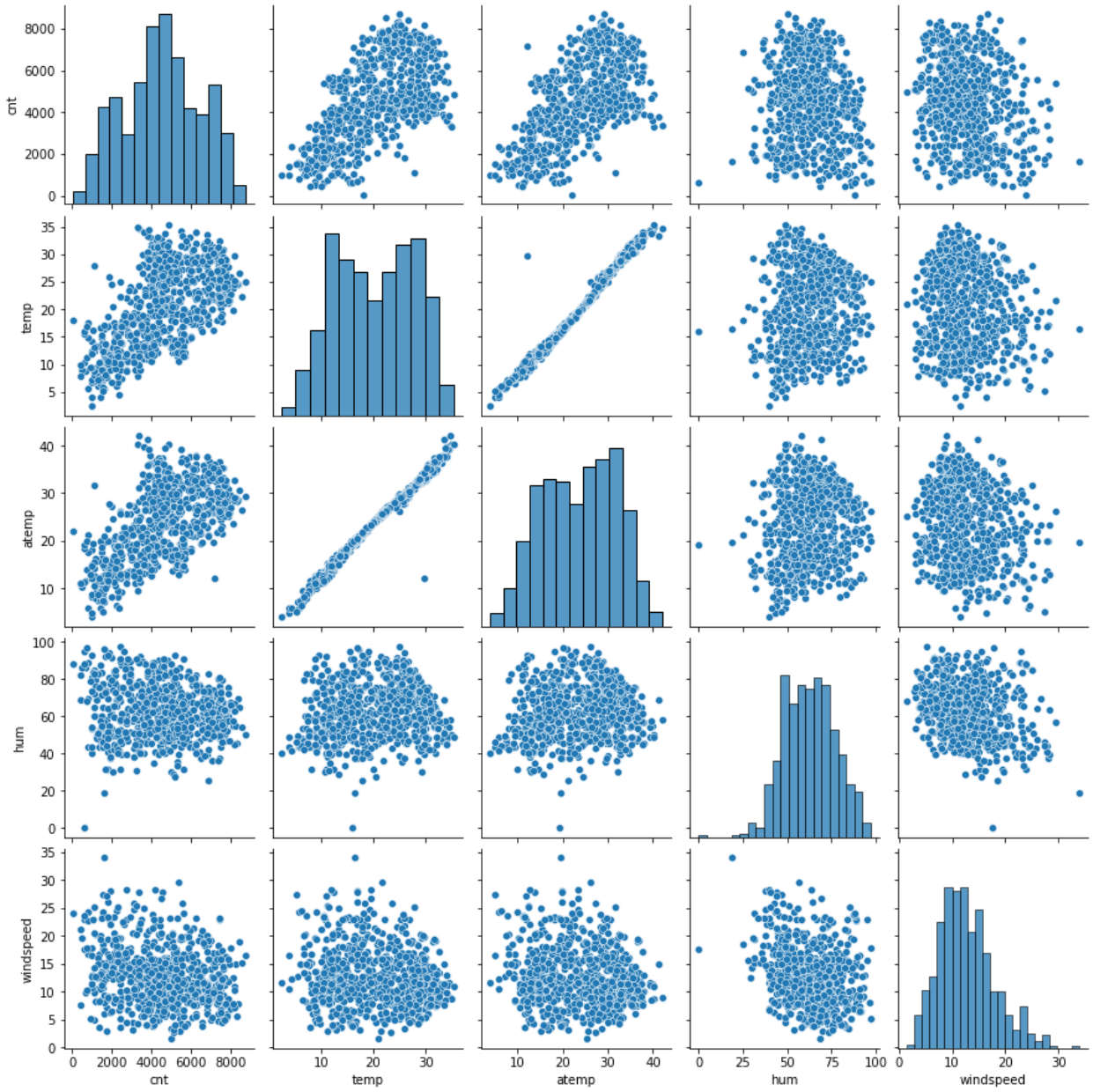
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Answer:

It's important to use drop_first=True during dummy variable creation as it helps in reducing the extra column created. It also 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? (1 mark)

Answer:



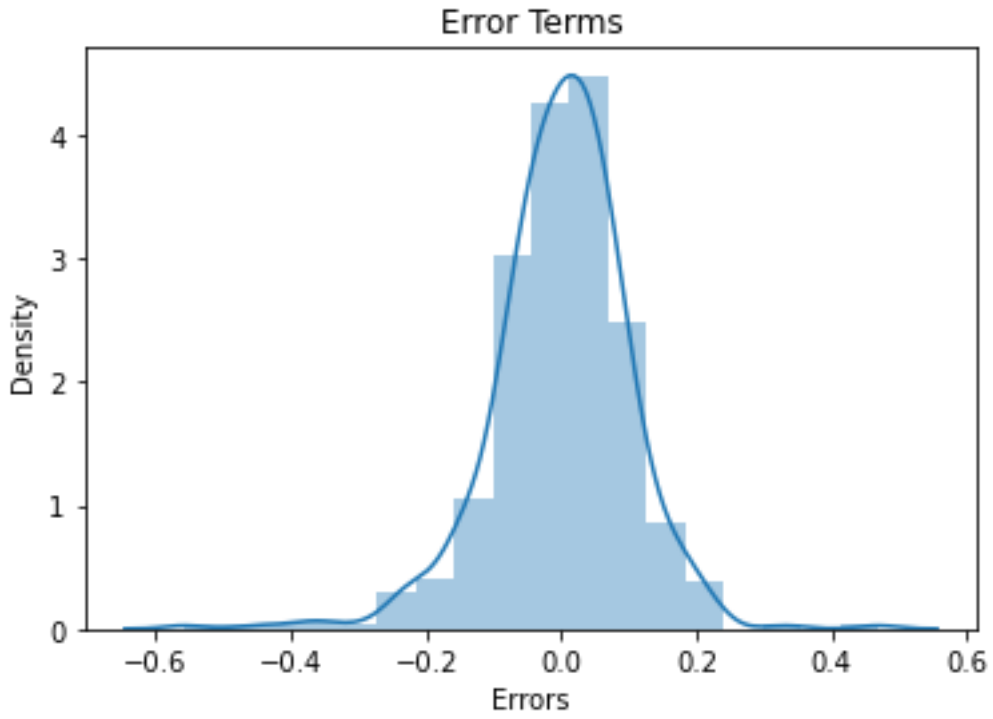
From the pairplot above, Variables temp and atemp are two numerical variables which are highly correlated with the target variable (cnt)

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:

Residual distribution should follow a normal distribution and centered around mean = 0. We validated the assumptions of Linear Regression after building the model on the

training set by plotting a distplot of residuals to check if residuals follow a normal distribution. The diagram below show that residuals are distributed at mean = 0.



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

Answer:

Top three features contributing significantly towards explaining the demand of the shared bikes are yr, weathersit and mnth.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer:

Linear regression is an analysis that assesses whether one or more predictor variables explain the dependent (criterion) variable. It's a type of supervised learning algorithm and is used for predictive analysis model. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis),

consequently called linear regression. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.

Fig 1.1

The above figure 1.1 presents the linear relationship between the dependent variable and independent variables. When the value of x (independent variable) increases, the value of y (dependent variable) is likewise increasing. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best. Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b\,(slope) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a\,(intercept) = \frac{n\sum y - b\,(\sum x)}{n}$$

where,

b = slope of the line

a = y-intercept of the line

x = independent variable from dataset

y = dependent variable from dataset

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:

Anscombe's quartet was developed by the statistician *Francis Anscombe* in 1973 to signify to signify both the importance of plotting data before analyzing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x, y) points. The basic thing to analyze about these data-sets is that they all share the same descriptive statistics (mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.

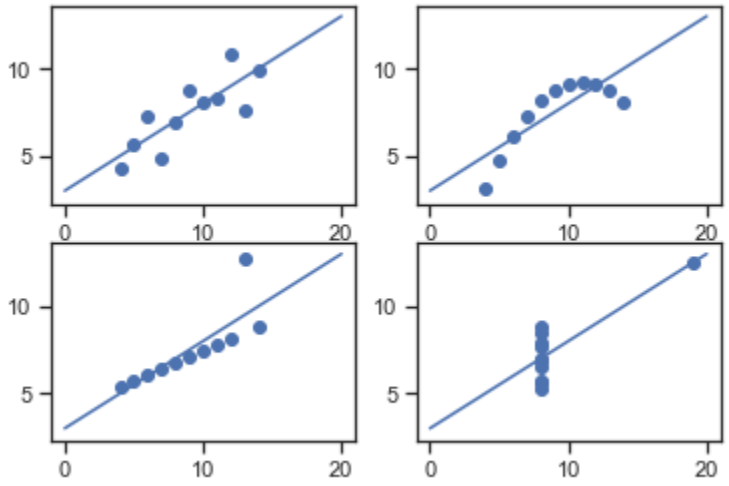


Fig: Graphical Representation of Anscombe's Quartet

3. What is Pearson's R? (3 marks)

Answer:

Pearson's r is a numerical summary of the strength of the linear association between the variables. It's value ranges from +1 to -1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

Scaling is a step carried out in data pre-processing stage which is applied to independent variables to normalize the data within a particular range. Normally the dataset contains features highly varying in magnitudes, units and range and if scaling is not performed, we will end up in incorrect modelling.

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling. sklearn.preprocessing.MinMaxScaler helps implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization on the other hand, is another scaling technique which replaces the values by their Z scores. It brings all the data into a standard normal distribution which has mean (μ) zero and standard deviation one(σ). **sklearn.preprocessing.scale** helps to implement standardization in python.

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

VIF provides how much the variance of the coefficient estimate is being inflated by collinearity. VIF is infinite when there's a perfect correlation which shows a perfect correlation between two independent variables.

In case of perfect correlation, we get R2 = 1, which leads to 1/(1-R2) infinity. This problem can be resolved by dropping one of the variables from the dataset that is causing perfect multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer:

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. It helps in determining if two data sets come from populations with a common distribution. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another

Q-Q plot is useful in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.