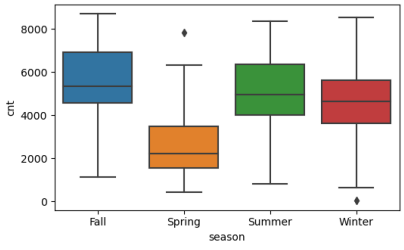
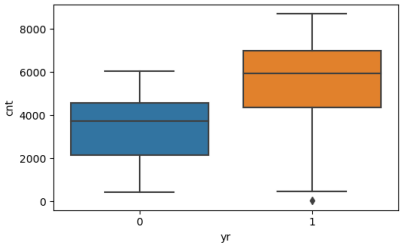
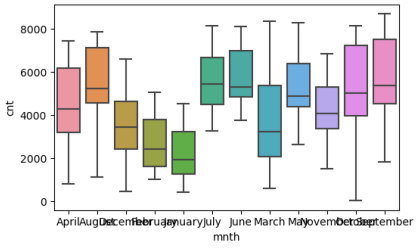
Assignment-based Subjective Questions

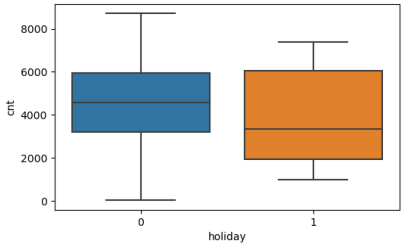
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

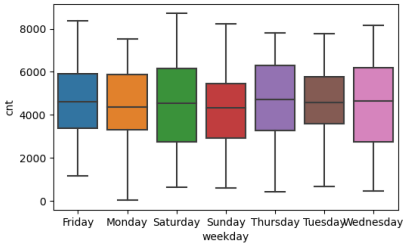
Answer:

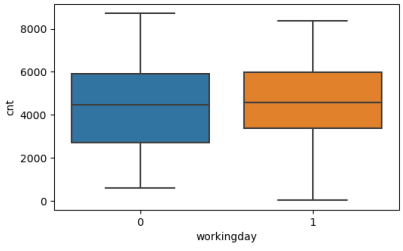


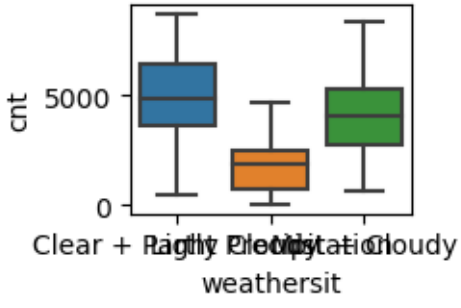












Observations:

1. *From weekday and working day categories are showing similar demands.*
2. *From yr-1 (2019) has higher demand than that in previous year (2018)*
3. *Demand gets increases from January and reaches maximum till July which started falling. Demand is much from May to October.*
4. *There is high demand during fall season.*
5. *During holiday demand decreases.*
6. *Most bikes were rented in misty and partly cloudy weather and least when light rains or snow.*

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

Answer:

It is important use drop_first=True as it helps to reduce extra column during dummy variable creation.

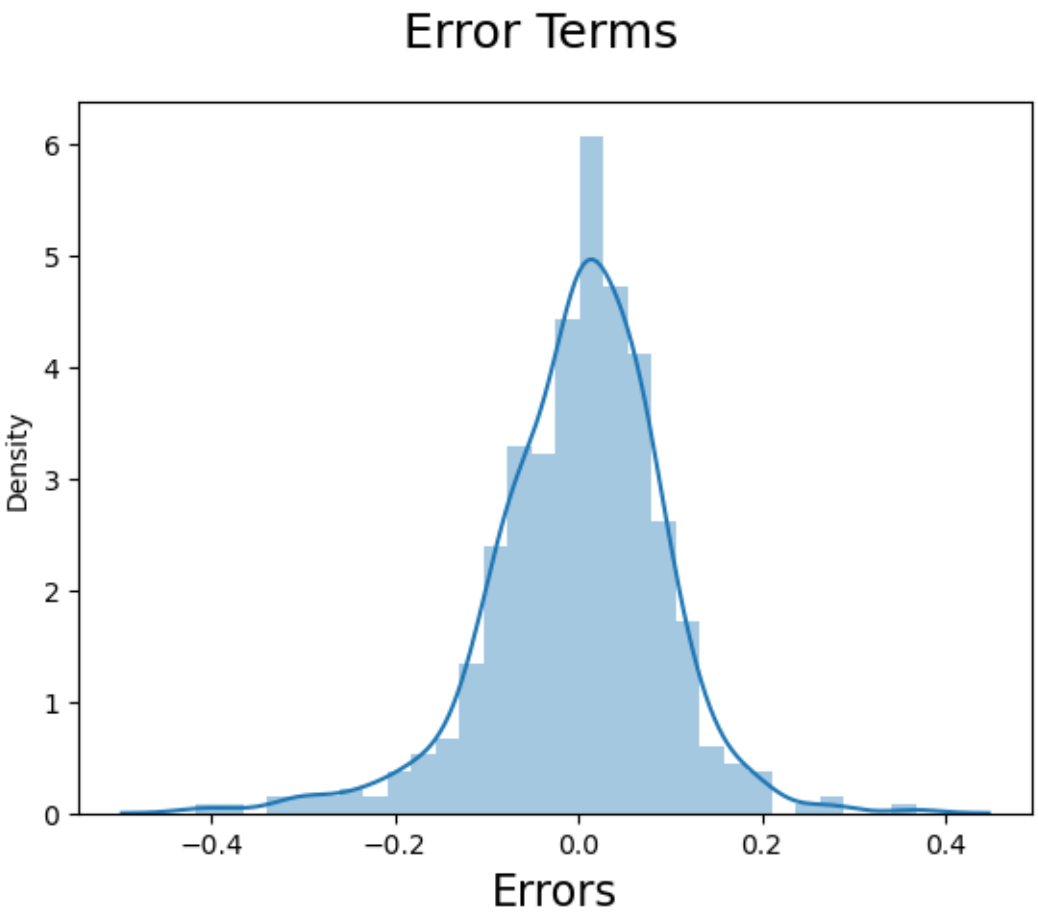
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer:

From pair plot analysis for Numerical variables, "atemp" has the highest corelation (0.63)with the target variable "cnt"

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

Answer:



- *Checked if the error term is equally distributed*
- *Checked data for homoscedasticity.*
- *Verifying all the variables p-value<0.05 and Vif <5*

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

Answer:

- *Yr - Year-2019* ***(Positive Corelation)***
- *weathersit_Light Precipitation - Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (****Negative Corelation)***
- *atemp - feeling temperature in Celsius****(Positive Corelation)***

General Subjective Questions

1.Explain the linear regression algorithm in detail.

Answer:

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

b = Slope of the line

a = y-intercept of the line

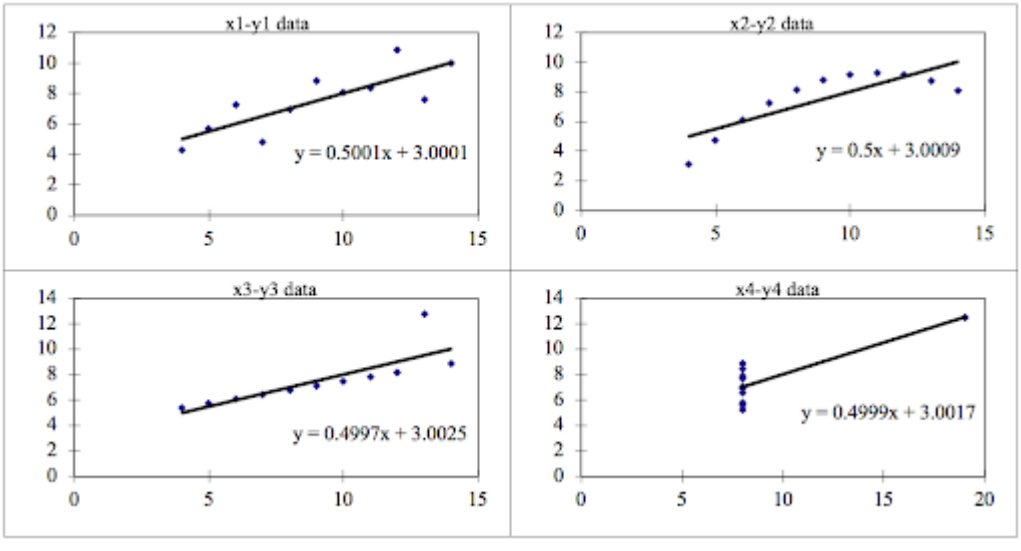
x = Independent variable from dataset

y = Dependent variable from dataset

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.
when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



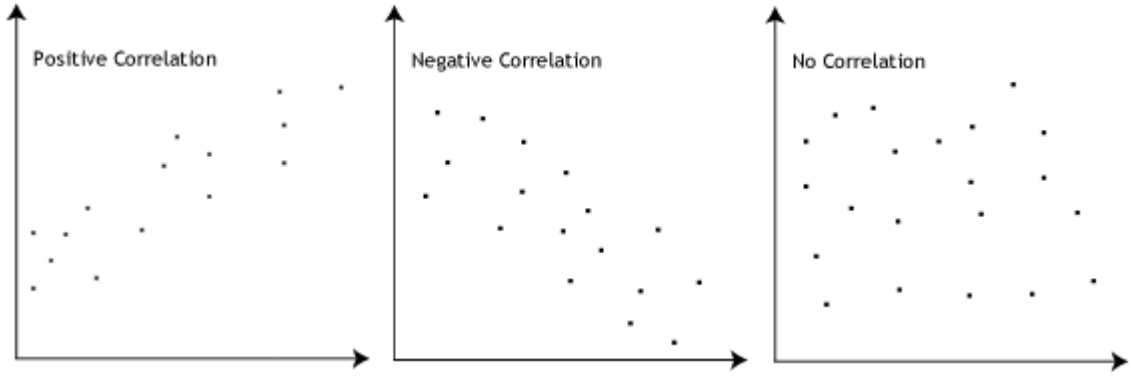
We can describe the four data sets as:
ANSCOMBE'S QUARTET FOUR DATASETS
***Data Set 1**: fits the linear regression model pretty well.*
***Data Set 2**: cannot fit the linear regression model because the data is non-linear.*
***Data Set 3**: shows the outliers involved in the data set, which cannot be handled by the linear regression model.*
***Data Set 4**: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.*

3. What is Pearson's R?

Answer:

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their

standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.
The Pearson's correlation coefficient varies between -1 and +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
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association



Pearson r Formula

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

Here,

=correlation coefficient

=values of the x-variable in a sample

=mean of the values of the x-variable

=values of the y-variable in a sample

=mean of the values of the y-variable

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

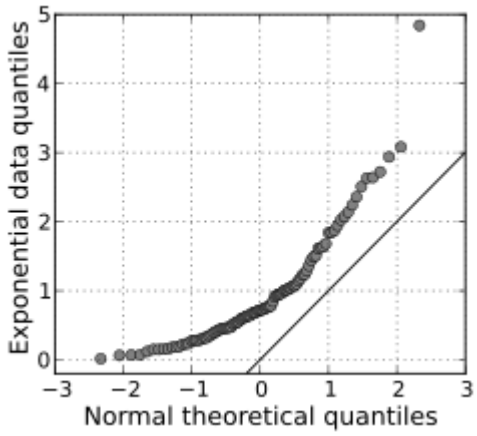
Answer: *Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).*

S.NO.	*Normalisation*	*Standardisation*
1.	*Minimum and maximum value of features are used for scaling*	*Mean and standard deviation is used for scaling.*

S.NO.	Normalisation	Standardisation
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.
A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.