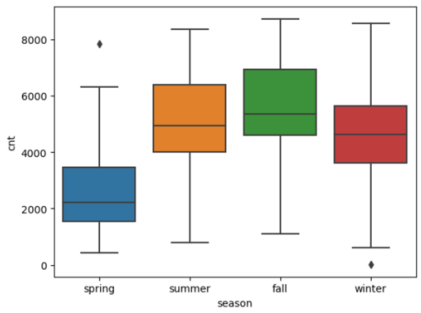
Assignment-based Subjective Questions

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:

Here the dependent variable is the demand of the shared bikes, ie cnt.
Effect of categorical variables on cnt:

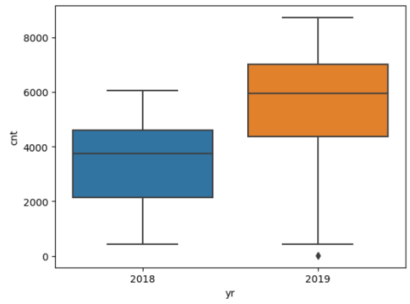
- Season:



Cnt is more during summer and fall and is low in spring.

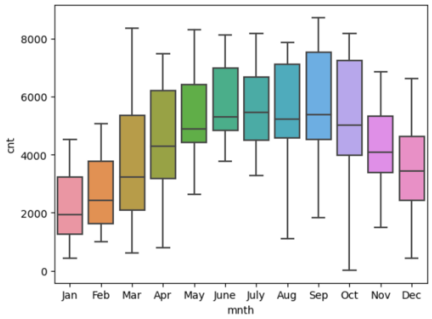
- Yr:

  Year is another categorical variable. According to the dataset, it is either 2018 or 2019. According to the dataset, demand for shared bikes is more during 2019.

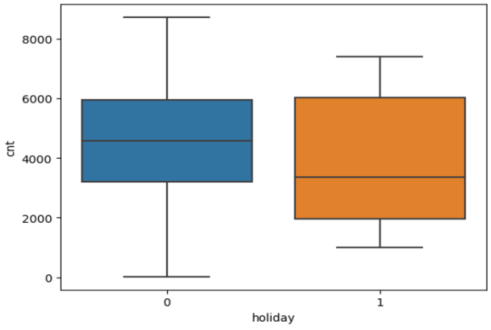


- Mnth:

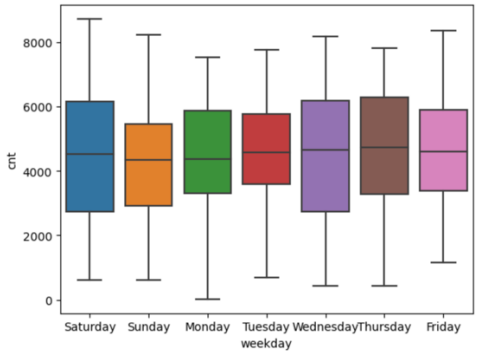
  Demand is more from June till October.



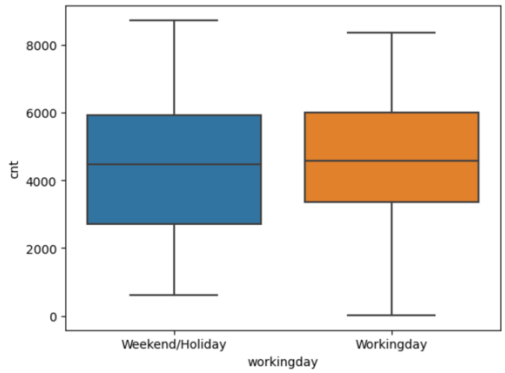
- Holiday:



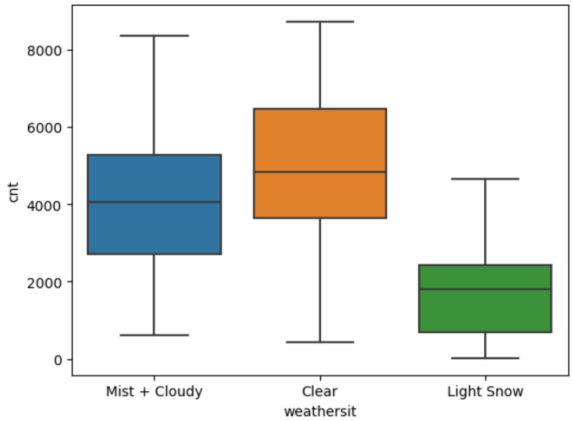
- Weekday
  No significant pattern for weekday



- Workingday
  No significant difference



- Weathersit
  Demand is more in clear weather and when it is cloudy.



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

Answer:

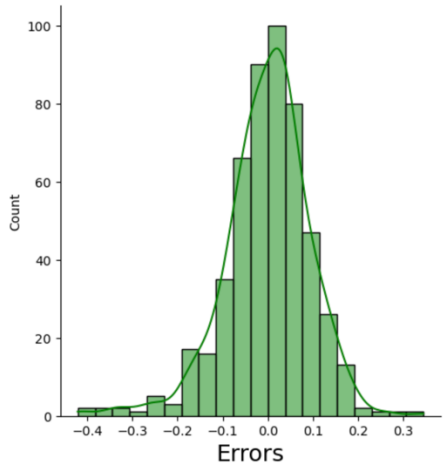
The drop_first parameter specifies whether or not you want to drop the first category of the categorical variable you're encoding. If we do not use drop_first = True, then n dummy variables will be created, and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

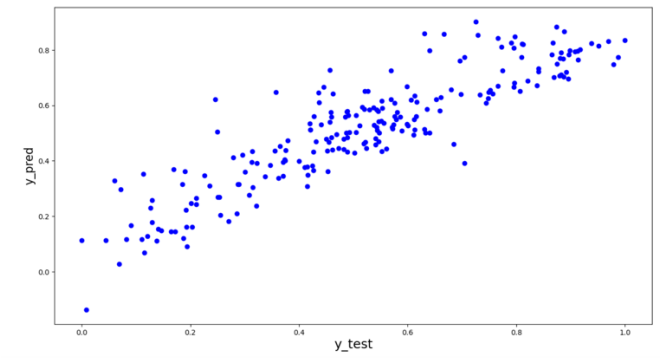
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

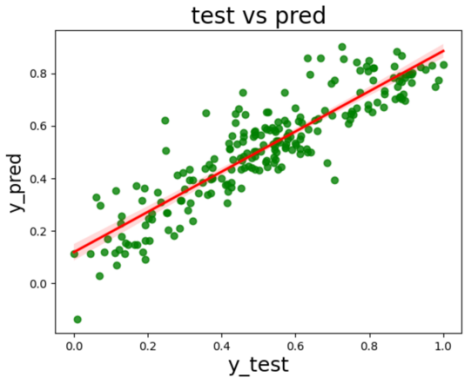
Answer:
Variable ‘temp’ has highest correlation with the target variable ‘cnt’.

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

Answer:
On plotting histogram, it is clear that normal distribution of terms is along with the mean of 0.







On plotting predicted value and test value the points are distributed around the diagonal line of actual vs predicted.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer:
‘temp’: temperature

'season': Season is a factor contributing to the demand of shared bikes.
'yr': Year is another feature that contributes to the demand of bike sharing.

General Subjective Questions

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

Answer:

Regression is a statistical technique that shows an algebraic relationship between two or more variables.

Based on this algebraic relationship (rather than a function), one can estimate the value of a variable, given the values of the other variables.

Linear Regression is one of the most fundamental algorithms in the Machine Learning world which comes under supervised learning. Basically it performs a regression task. Regression models predict a dependent (target) value based on independent variables.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y (output). Hence, the name is Linear Regression.

Simple Linear Regression

When we try to find out a relationship between a dependent variable (Y) and one independent (X) then it is known as Simple Linear Regression.

The mathematical equation can be given as:

$y = \beta 0 + \beta 1 * X$

Where

Y is the response or the target variable

x is the independent feature

$\beta 1$ is the coefficient of x

$\beta 0$ is the intercept

Multiple Linear Regression

This type of analysis allows you to understand the relationship between a continuous dependent variable and two or more independent variables.

The independent variables can be either continuous (like age and height) or categorical (like gender and occupation).

Formulation of multiple linear regression

$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 \ldots\ldots.. \beta pXp + \epsilon$

The goal of the linear regression algorithm is to get the **best values for B0 and B1** to find the best fit line. The best fit line is a line that has the least error which means the error between predicted values and actual values should be minimum.

Random Error (Residuals)

In regression, the difference between the observed value of the dependent variable(**yi**) and the predicted value(**predicted**) is called the residuals.

$\epsilon i = ypredicted - yi$

where $ypredicted = \beta 0 + \beta 1 Xi$

Evaluation Metrics for Linear Regression

These evaluation metrics usually provide a measure of how well the observed outputs are being generated by the model.

The most used metrics are,

- Coefficient of Determination or R-Squared (R2)
- Root Mean Squared Error (RSME) and Residual Standard Error (RSE)

Coefficient of Determination or R-Squared (R2)

R-Squared is a number that explains the amount of variation that is explained/captured by the developed model. It always ranges between 0 & 1 . Overall, the higher the value of R-squared, the better the model fits the data.
Mathematically it can be represented as,

R2 = 1 – (RSS/TSS)

- **Residual sum of Squares (RSS)** is defined as the sum of squares of the residual for each data point in the plot/data. It is the measure of the difference between the expected and the actual observed output.

$$RSS = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$$

- **Total Sum of Squares (TSS)** is defined as the sum of errors of the data points from the mean of the response variable. Mathematically TSS is,

$$TSS = \sum (y_i - \bar{y}_i)^2$$

Considerations of Multiple Linear Regression
All the four assumptions made for Simple Linear Regression still hold true for Multiple Linear Regression along with a few new additional assumptions.

- **Overfitting**: When more and more variables are added to a model, the model may become far too complex and usually ends up memorizing all the data points in the training set. This phenomenon is known as the overfitting of a model. This usually leads to high training accuracy and very low test accuracy.
- **Multicollinearity**: It is the phenomenon where a model with several independent variables, may have some variables interrelated.
- **Feature Selection:** With more variables present, selecting the optimal set of predictors from the pool of given features (many of which might be redundant) becomes an important task for building a relevant and better model.

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

Answer:

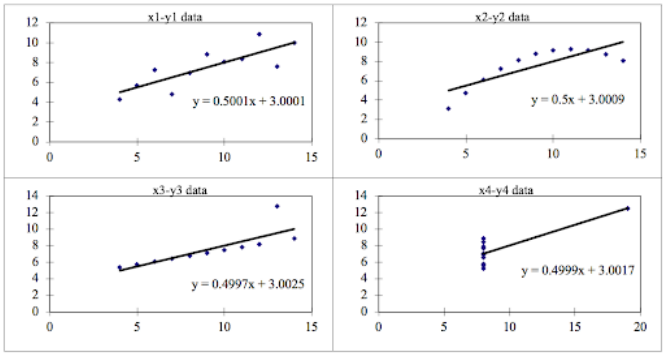
Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics yet have very different distributions and appear very different when graphed.
Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This 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.
We can define these four plots as follows:

Anscombe's Data								
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

The statistical information for these four data sets are approximately similar. We can compute them as follows:

Anscombe's Data								
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89
			Summary Statistics					
N	11	11	11	11	11	11	11	11
mean	9.00	7.50	9.00	7.500909	9.00	7.50	9.00	7.50
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

However, 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:



Anscombe's Quartet Four Datasets

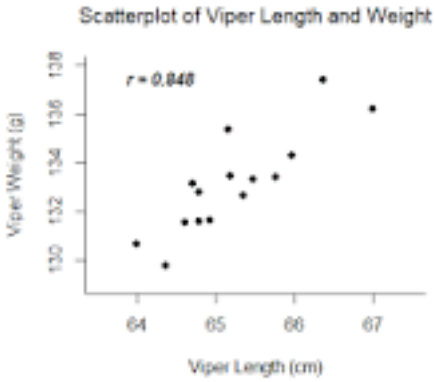
- **Data Set 1:** fits the linear regression model pretty well.
- **Data Set 2:** cannot fit the linear regression model because the data is non-linear.
- **Data Set 3:** shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- **Data Set 4:** shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

- Important points
    - Plotting the data is very important and a good practice before analysing the data.
    - Outliers should be removed while analysing the data.
    - Descriptive statistics do not fully depict the data set in its entirety.

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

Answer:

Person's R or The Pearson correlation coefficient (r) is a way of measuring linear correlation.



The Pearson correlation coefficient, r, it is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.
The further away r is from zero, the stronger the linear relationship between the two variables. The sign of r corresponds to the direction of the relationship. If r is positive, then as one variable increases, the other tends to increase.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   Answer:
   It is a step of Data Pre Processing that is applied to independent variables or features of data. It helps to normalize the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm.
   When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:
   1. Ease of interpretation
   2. Faster convergence for gradient descent methods
   You can scale the features using two very popular method:
   Standardizing and MinMax Scaling
   Standardization is divided by the standard deviation after the mean has been subtracted. Data is transformed into a range between 0 and 1 by normalization, which involves dividing a vector by its length.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   Answer:
   VIF becomes infinite when R^2 becomes 1. Formula for VIF is
   VIF = 1/(1-R^2)

The reason for R^2 to be 1 is that there is a perfect correlation between 2 independent variables.

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

Answer:

A QQ plot is a scatterplot created by plotting two sets of quantiles against one another.

It determines how many values in a distribution are above or below a certain limit.

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 are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

This helps 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.

Few advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior