Measures

Saturday, June 17, 2023 11:32 AM

Various Sources / Formats → Data → Exploratory Data Analysis → Plan a ML Model

Exploratory Data Analysis:
- Compute the measures
- Graphs / Visualization

Plan a ML Model ↓↑ Model Building ↓↑ Model Evaluation

Geometric Mean: a, b → $(a\,b)^{1/2}$; $a_1, a_2, \ldots, a_n$ → $(a_1 \cdot a_2 \cdot \ldots \cdot a_n)^{1/n}$

Harmonic Mean: Reciprocal of mean of reciprocals

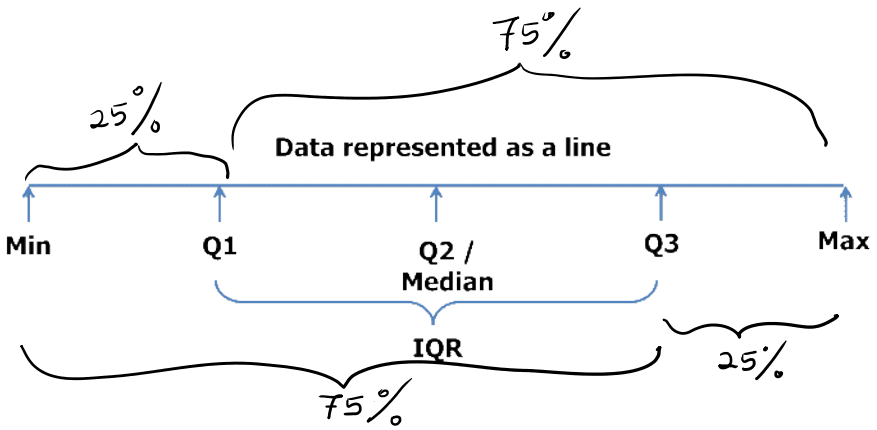
a, b → $\dfrac{1}{\frac{\frac{1}{a}+\frac{1}{b}}{2}} = \dfrac{2ab}{a+b}$

$a_1, a_2, \ldots a_n$ → $\dfrac{1}{\frac{\frac{1}{a_1}+\frac{1}{a_2}+\cdots+\frac{1}{a_n}}{n}}$

1st Quartile :- 25% values in data less than 1st Q.

2nd Quartile :- 50% ——|——— 2nd Q
= median

3rd Quartile :- 75% ——|——— 3rd Q



Quartiles: Divide the data into 4 equal parts
Deciles: Divide the data into 10 equal parts
Percentiles: Divide the data into 100 equal parts

$$\text{Abs deviation about mean} = \frac{\sum |X_i - \text{mean}|}{N}$$ Mean Deviation

$$\text{Variance} = \frac{\sum (X_i - \text{mean})^2}{N}; \quad \text{Standard deviation} = \sqrt{\text{Variance}}$$

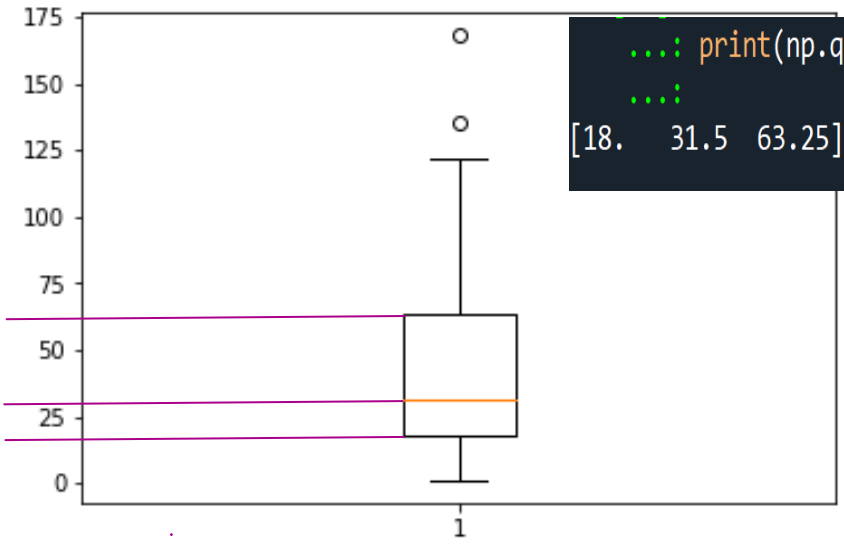
$$\text{mean} = \frac{\sum X_i}{N},$$

$$\sum_{i=1}^{N}(a_i - b_i) = \sum_{i=1}^{N} a_i - \sum_{i=1}^{N} b_i$$

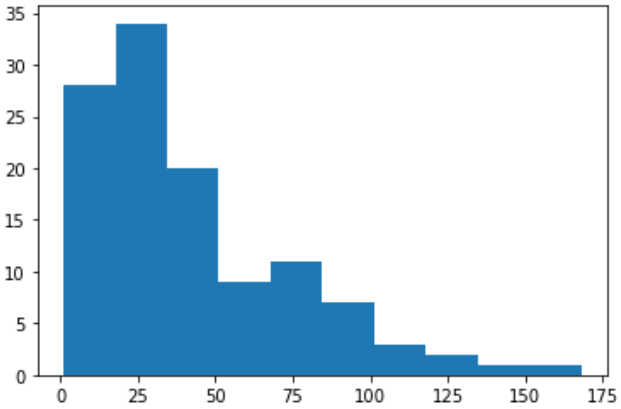
$$\sum_{i=1}^{N} (X_i - \text{mean}) = \sum_{i=1}^{N} X_i - \sum_{i=1}^{N} \text{mean}$$

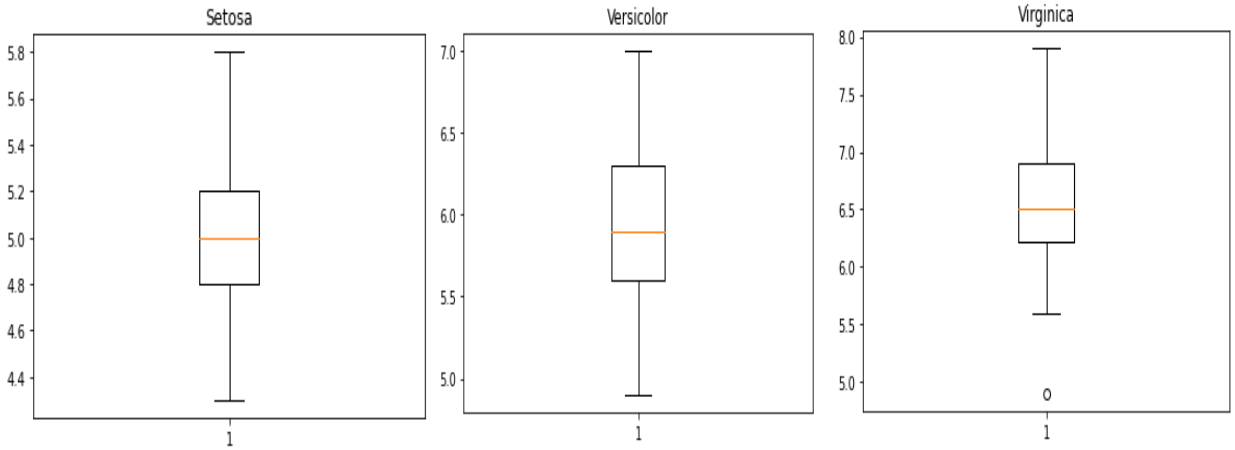
$$= N * \text{mean} - N * \text{mean}$$

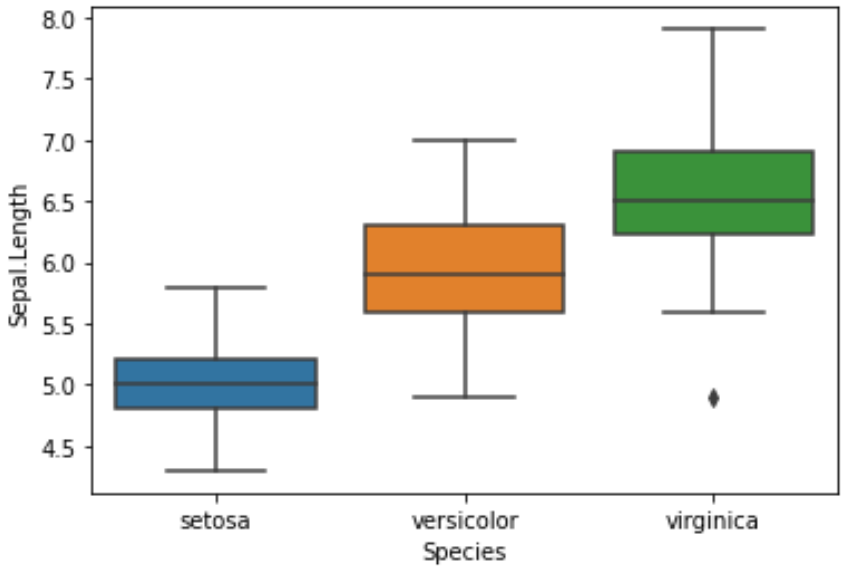
$$= 0$$

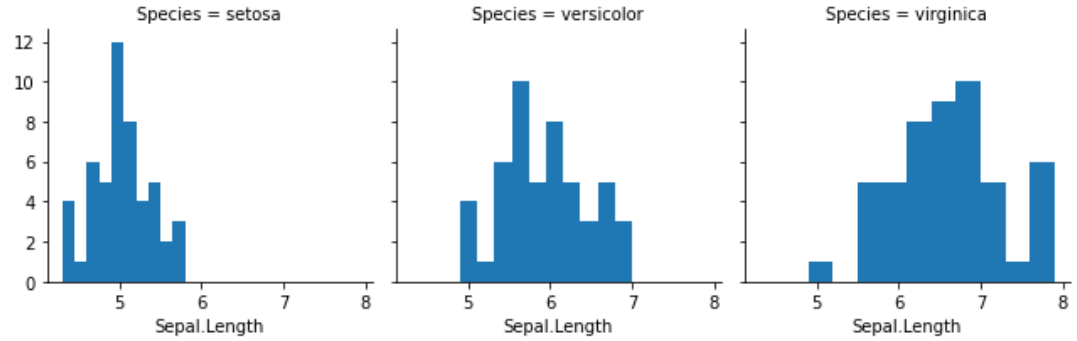


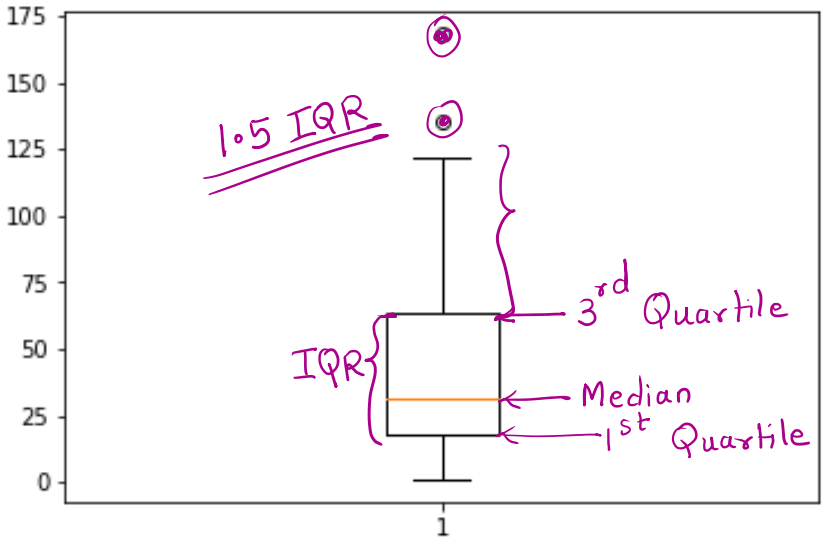

```
...: print(np.quantile(air['Ozone'].dropna(),
...:                   [0.25,0.5,0.75]))
[18.   31.5  63.25]
```


35
30
25
20
15
10
5
0
0
25
50
75
100
125
150
175

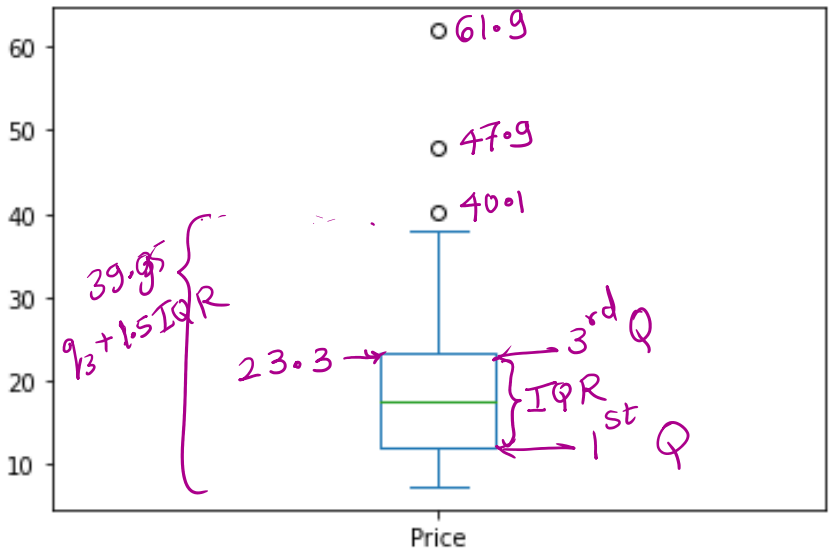

Setosa
Versicolor
Virginica
5.8
5.6
5.4
5.2
5.0
4.8
4.6
4.4
7.0
6.5
6.0
5.5
5.0
8.0
7.5
7.0
6.5
6.0
5.5
5.0
o
1
1
1


8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
Sepal.Length
setosa
versicolor
virginica
Species





$Q_3 - Q_1 = IQR$



Outliers :-
Any point which is 1.5 IQR away from the edge of the box

$1.5IQR = 16.65$

$q_3 + 16.65 = 39.95$

40.1
47.9
61.9

Outliers

Covariance & Correlation

Wednesday, June 21, 2023 3:09 PM

μ : mean

$$\text{Variance} = \frac{1}{N}\sum (x_i - \mu)^2$$

μ_x : mean of X

(X_1, Y_1)
(X_2, Y_2)
$\vdots$
(X_N, Y_N)

$$\text{Var}(X) = \sigma_x^2 = \frac{1}{N}\sum (x_i - \mu_x)^2$$

$$\text{Var}(Y) = \sigma_Y^2 = \frac{1}{N}\sum (Y_i - \mu_Y)^2$$

μ_Y : mean of Y

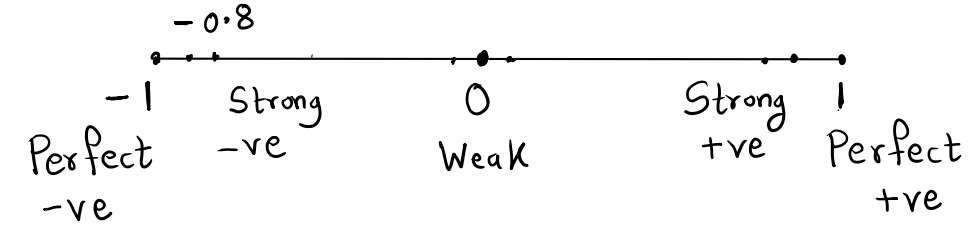
$$\text{Covariance} : \text{Cov}(X,Y) = \frac{1}{N}\sum (x_i - \mu_x)(Y_i - \mu_Y)$$

X	23	56	78	90	109	123
Y	789	896	908	1023	1348	1789

$X_1\ X_2 \ldots X_p$: Variance Covariance matrix

$$\begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1X_2) & \text{Cov}(X_1X_3) & \cdots & \text{Cov}(X_1X_p) \\ \text{Cov}(X_2X_1) & \text{Var}(X_2) & \text{Cov}(X_2X_3) & \cdots & \text{Cov}(X_2X_p) \\ \text{Cov}(X_3X_1) & \text{Cov}(X_3X_2) & \text{Var}(X_3) & & \\ \vdots & & & \ddots & \\ \text{Cov}(X_pX_1) & \text{Cov}(X_pX_2) & \cdots & & \text{Var}(X_p) \end{bmatrix}_{p \times p}$$

ρ : Corr Coefficient



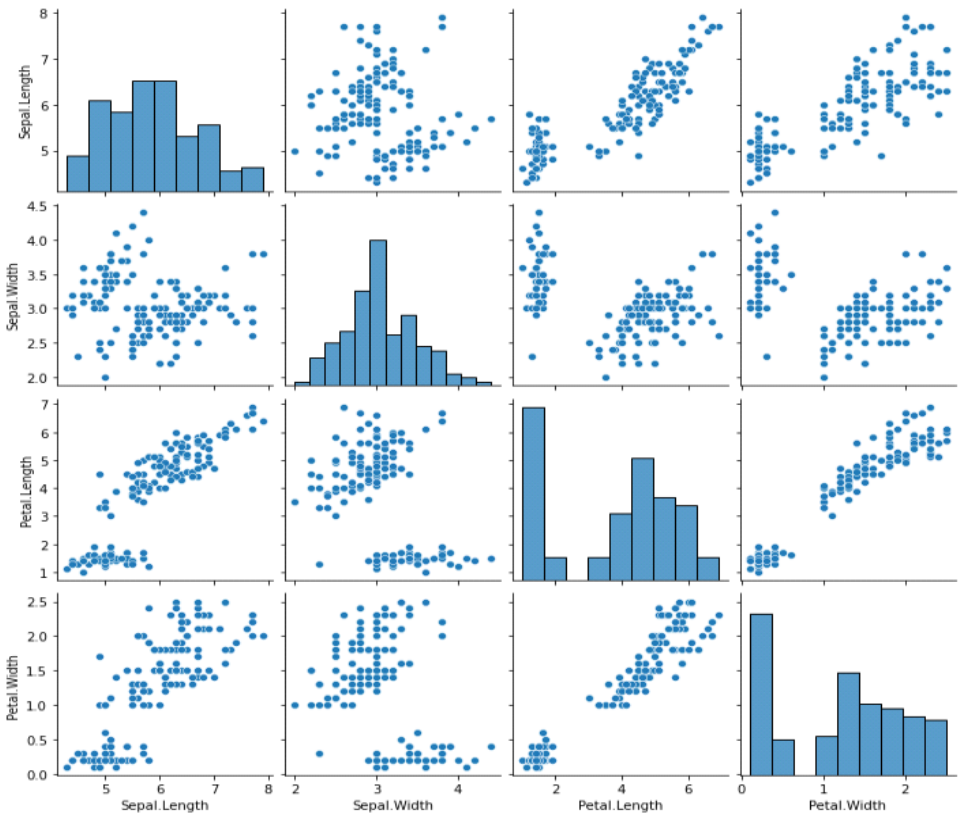
Correlation matrix

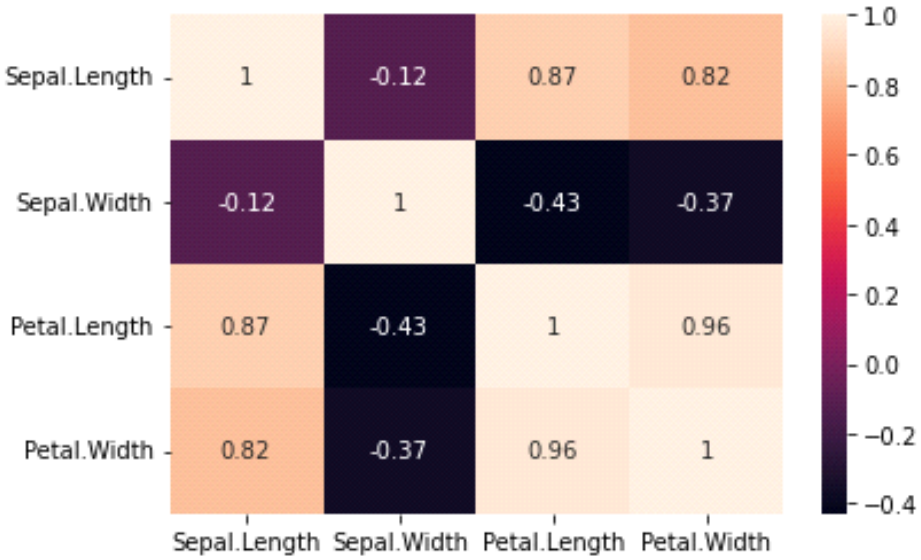
$X_1\ X_2 \ldots X_p$

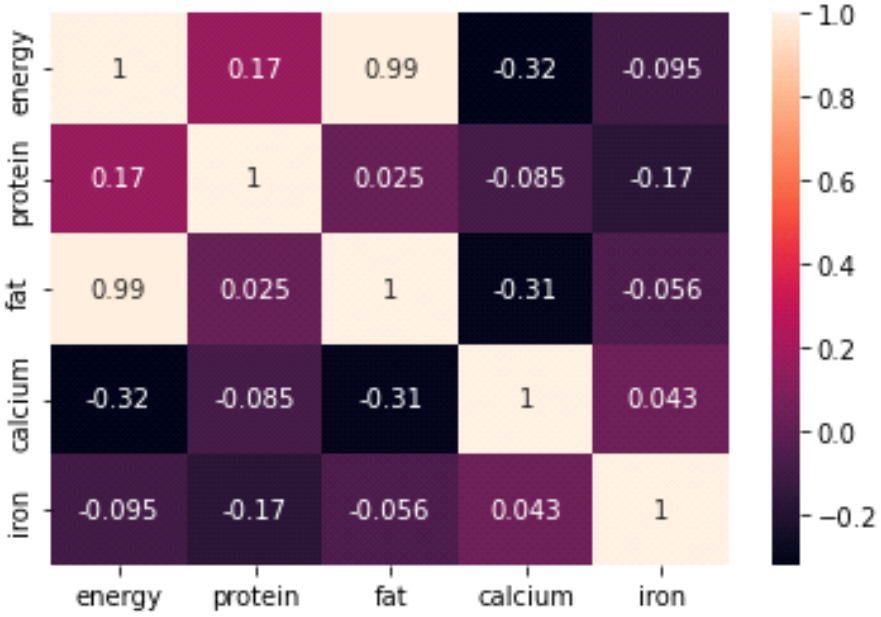
Correlation matrix

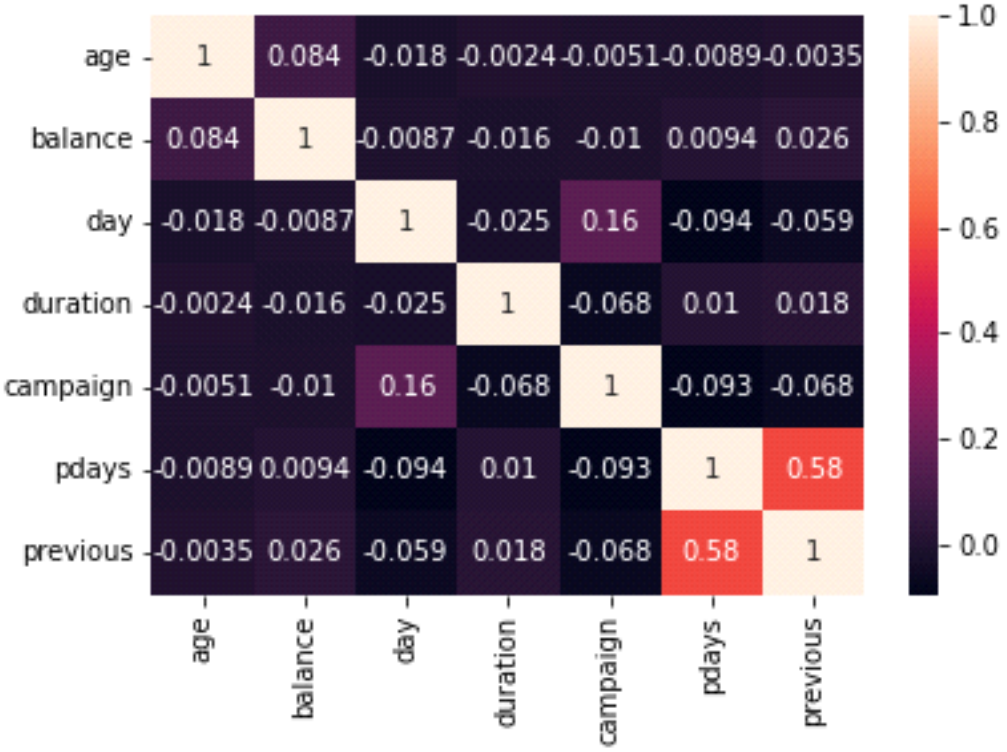
$X_1 \; X_2 \; \ldots \; X_p$

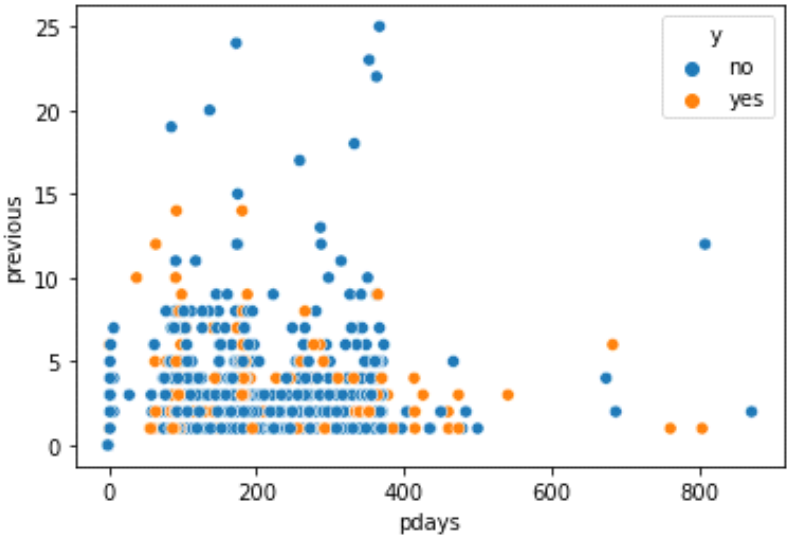
$$\begin{bmatrix} 1 & Corr(X_1 X_2) & Corr(X_1 X_3) & \cdots & Corr(X_1 X_p) \\ Corr(X_2 X_1) & 1 & Corr(X_2 X_3) & \cdots & Corr(X_3 X_p) \\ Corr(X_3 X_1) & Corr(X_3 X_2) & 1 & & \\ & & & \ddots & \\ Corr(X_p X_1) & \cdots & & & 1 \end{bmatrix}$$

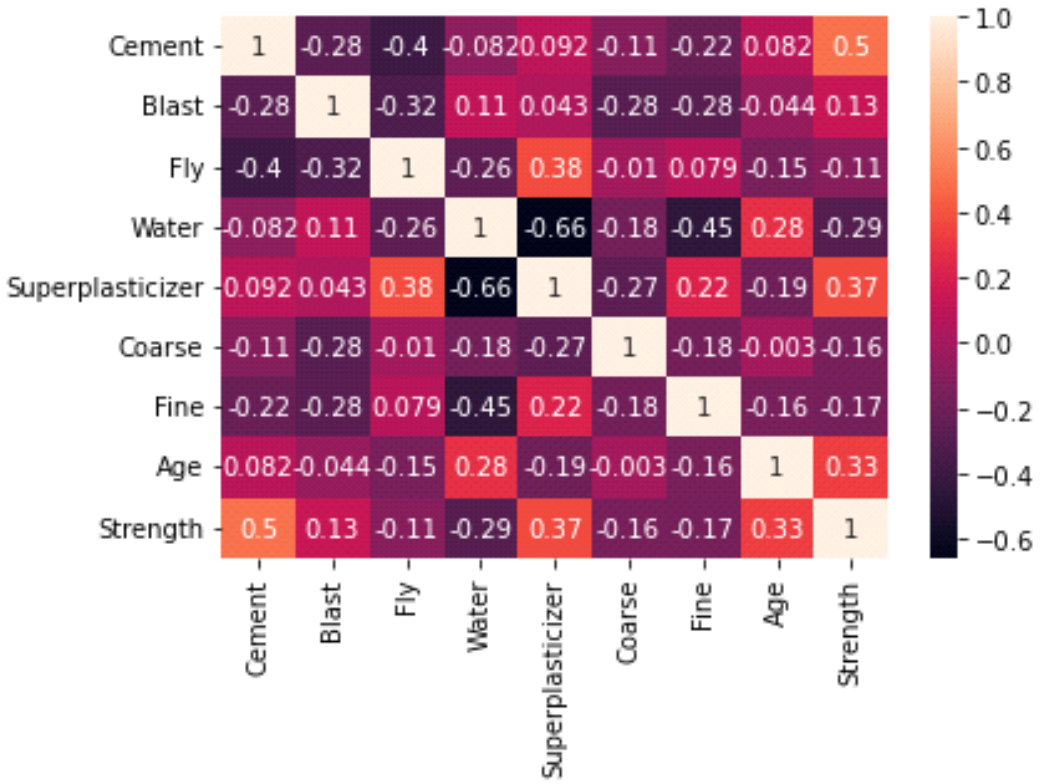



Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
1
-0.12
0.87
0.82
-0.12
1
-0.43
-0.37
0.87
-0.43
1
0.96
0.82
-0.37
0.96
1
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


energy
protein
fat
calcium
iron
1
0.17
0.99
-0.32
-0.095
0.17
1
0.025
-0.085
-0.17
0.99
0.025
1
-0.31
-0.056
-0.32
-0.085
-0.31
1
0.043
-0.095
-0.17
-0.056
0.043
1
energy
protein
fat
calcium
iron
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


age
balance
day
duration
campaign
pdays
previous
1 0.084 -0.018 -0.0024 -0.0051 -0.0089 -0.0035
0.084 1 -0.0087 -0.016 -0.01 0.0094 0.026
-0.018 -0.0087 1 -0.025 0.16 -0.094 -0.059
-0.0024 -0.016 -0.025 1 -0.068 0.01 0.018
-0.0051 -0.01 0.16 -0.068 1 -0.093 -0.068
-0.0089 0.0094 -0.094 0.01 -0.093 1 0.58
-0.0035 0.026 -0.059 0.018 -0.068 0.58 1
1.0
0.8
0.6
0.4
0.2
0.0

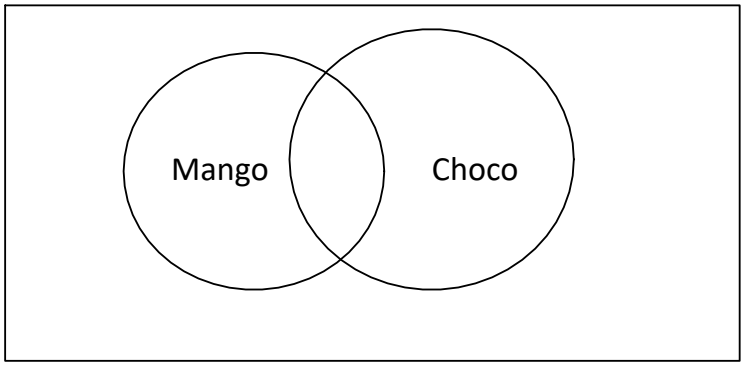

y
no
yes
previous
pdays
0
5
10
15
20
25
0
200
400
600
800


Cement 1 -0.28 -0.4 -0.082 0.092 -0.11 -0.22 0.082 0.5
Blast -0.28 1 -0.32 0.11 0.043 -0.28 -0.28 -0.044 0.13
Fly -0.4 -0.32 1 -0.26 0.38 -0.01 0.079 -0.15 -0.11
Water -0.082 0.11 -0.26 1 -0.66 -0.18 -0.45 0.28 -0.29
Superplasticizer 0.092 0.043 0.38 -0.66 1 -0.27 0.22 -0.19 0.37
Coarse -0.11 -0.28 -0.01 -0.18 -0.27 1 -0.18 -0.003 -0.16
Fine -0.22 -0.28 0.079 -0.45 0.22 -0.18 1 -0.16 -0.17
Age 0.082 -0.044 -0.15 0.28 -0.19 -0.003 -0.16 1 0.33
Strength 0.5 0.13 -0.11 -0.29 0.37 -0.16 -0.17 0.33 1
Cement
Blast
Fly
Water
Superplasticizer
Coarse
Fine
Age
Strength
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

If A and B are independent,
$P(B|A) = P(B)$ & $P(B|A^c) = P(B)$
$P(A|B) = P(A)$ & $P(A|B^c) = P(A)$

Multiplication Theorem:
$P(A \text{ intersection } B) = P(B|A) * P(A) = P(B) * P(A)$

A large-scale survey finds that 80% of college students enjoy eating chocolate ice cream, 65% of college students enjoy eating mango ice cream, and 55% of college student enjoy eating both chocolate and mango ice cream. What proportion of college students enjoys eating chocolate or mango ice cream?



P(M) = 0.65
P(Choco) = 0.8

P(M and Choco) = 0.55

P(M union Choco)
= P(M) + P(Choco) - P(M int Choc)
= 0.65 + 0.8 - 0.55
= 0.9

In a large class, the probability of randomly selecting a woman student is 0.65. The probability of randomly selecting a student who is a woman and who earned a grade A is 0.25. If you randomly select a student who is a woman, what is the probability that she earned a grade A?

0.25 / 0.65 = 0.3846

Xi	Pi
23	0.6250000
45	0.1734694
89	0.2015306

23. The number of children per family was determined and summarized in the following table.

Number of Children	Number of families
1	15
2	31
3	24
4	7

Find the expected number, variance, and standard deviation of the number of children per family.

14. A Canadian business school summarized the gender and residency of its incoming class as follows:

	Residency				
Gender	**Canada**	**United States**	**Europe**	**Asia**	**Other**
Male	125	18	17	50	8
Female	103	8	10	92	4

a. Construct a joint probability table.

b. Calculate the marginal probabilities.

Joint Probability Distribution

	A1	A2	A3	A4
B1	P(A1∩B1)	P(A2∩B1)	P(A3∩B1)	P(A4∩B1)
B2	P(A1∩B2)	P(A2∩B2)	P(A3∩B2)	P(A4∩B2)
B3	P(A1∩B3)	P(A2∩B3)	P(A3∩B3)	P(A4∩B3)

	A1	A2	A3	A4	Marginal
B1	P(A1∩B1)	P(A2∩B1)	P(A3∩B1)	P(A4∩B1)	P(B1)
B2	P(A1∩B2)	P(A2∩B2)	P(A3∩B2)	P(A4∩B2)	P(B2)
B3	P(A1∩B3)	P(A2∩B3)	P(A3∩B3)	P(A4∩B3)	P(B3)
Marginal	P(A1)	P(A2)	P(A3)	P(A4)	1

Conditional	A1	A2	A3	A4
B1	P(B1|A1)	P(B1|A2)	P(B1|A3)	P(B1|A4)
B2	P(B2|A1)	P(B2|A2)	P(B2|A3)	P(B2|A4)
B3	P(B3|A1)	P(B3|A2)	P(B3|A3)	P(B3|A4)
Marginal	P(A1)	P(A2)	P(A3)	P(A4)

Region	Book	DVD	Total
East	56	42	98
North	43	42	85
South	62	37	99
West	100	90	190
Total	261	211	472

a. Find the marginal probabilities that a sale originated in each of the four regions and the marginal probability of each type of sale (book or DVD).

b. Find the conditional probabilities of selling a book given that the customer resides in each region.

16. Use the Civilian Labor Force data in the Excel file *Census Education Data* to find the following:

a. P(unemployed and advanced degree)

b. P(unemployed | advanced degree)

c. P(not a high school grad | unemployed)

d. Are the events "unemployed" and "at least a high school graduate" independent?

17. Using the data in the Excel file *Consumer Transport Survey*, develop a cross-tabulation for Gender and Vehicle Driven; then convert this table into probabilities.

a. What is the probability that a respondent is female? **0.58**

b. What is the probability that a respondent drives an SUV? 0.30

c. What is the probability that a respondent is male and drives a minivan? 0.02

d. What is the probability that a female respondent drives either a truck or an SUV? 0.10

e. If it is known that an individual drives a car, what is the probability that the individual is female? 0.72727273

f. If it is known that an individual is male, what is 0.47619

f. If it is known that an individual is male, what is the probability that he drives an SUV? 0.47619

g. Determine whether the random variable "gender" and the event "vehicle driven" are statistically independent. What would this mean for advertisers?

Gender	Female	Male	All
Vehicle Driven			
Car	0.32	0.12	0.44
Mini Van	0.16	0.02	0.18
SUV	0.10	0.20	0.30
Truck	0.00	0.08	0.08
All	0.58	0.42	1.00

```
    ...:                normalize='index')
Out[18]:
Gender                Female      Male
Vehicle Driven
Car                 0.727273  0.272727
Mini Van            0.888889  0.111111
SUV                 0.333333  0.666667
Truck               0.000000  1.000000
All                 0.580000  0.420000
```

0.4783

0.3044

0.6087

0.2608

13. A survey of shopping habits found the percentage of respondents that use technology for shopping as shown in Figure 5.30. For example, 17.39% only use online coupons; 21.74% use online coupons and check prices online before shopping, and so on.

a. What is the probability that a shopper will check prices online before shopping?

b. What is the probability that a shopper will use a smart phone to save money?

c. What is the probability that a shopper will use online coupons?

d. What is the probability that a shopper will not use any of these technologies?

e. What is the probability that a shopper will check prices online and use online coupons but not use a smart phone? 0.2174

f. If a shopper checks prices online, what is the probability that he or she will use a smart phone? 0.2174

g. What is the probability that a shopper will check prices online but not use online coupons or a smart phone? 0.0435

Binomial

Thursday, June 22, 2023 7:30 PM

In a typical Month, an Insurance agent presents life insurance plans to 40 potential customers. Historically, one in four such customers chooses to buy Life Insurance from this agent. Based on the relevant binomial distribution , answer the following questions :

1. What is the probability that exactly 5 customers will buy life Insurance from this agent in the coming month ?

X : no. of customers who will buy the insurance

p = 0.25, q = 0.75, n = 40

P(X = 5)

```
binom.pmf(5, 40, 0.25)
0.0272317427532459 48
```

2. What is the probability that not more than 10 customers will buy life insurance from this agent in the coming month ?

P(X<=10)

```
binom.cdf(10, 40, 0.25)
0.5839040780287896
```

3. What is the probability that at least 20 customers will buy life insurance from this agent in the coming month ?

P(X>=20) = P(X>19)

```
binom.sf(19, 40, 0.25)
0.0005724311071761386

1 - binom.cdf(19, 40, 0.25)
0.0005724311071760857
```

4. Determine the mean and variance of the number of customers who will buy life insurance from this agent in the coming month.

```
binom.stats(40, 0.25)
(10.0, 7.5)
```

28. If a cell phone company conducted a telemarketing campaign to generate new clients and the probability of successfully gaining a new customer was 0.07, what is the probability that contacting 50 potential customers would result in at least 5 new customers?

p = 0.07, n =50
P(X>4)

```
binom.sf(4, 50, 0.07)
0.27097309000112557
```

1. An airline estimates that 94% of people booked on their flights actually show up. If the airline books 71 people for a flight of which the maximum number of seats is 69, what is the probability that the number of people who show up will exceed the capacity of the plane? Assume Binomial Distribution for the number of flights staying booked.

p = 0.94, n = 71
P(X>69)

```
binom.sf(69,71,0.94)
0.06838377878951435
```

2. The prevalence of a disorder in a certain group of people is 35%. If 20 people from that group are chosen at random, what is the probability that:
   a) None of them have that disorder
   b) 10 of them have that disorder
   c) At most 10 of them have that disorder
   d) At least 14 of them have that disorder

2. A student is applying for Masters course in 8 US Universities and believes that she has in each of the eight universities a constant and independent 0.42 probability of getting selected. Write code to answer the following questions:

(a) What is the probability that she will get call from at least 3 universities? → $P[X \geq 3] = P[X > 2]$ (sf)

(b) What is the probability that she will get calls from exactly 4 universities? → $P[X = 4]$ (pmf)

$n = 8$ $p = 0.42$

Poisson

Thursday, June 22, 2023 8:51 PM

Poisson Examples:

1. Number of accidents happening on a road in a certain period of time
2. Number of defects in an item piece
3. Number of arrivals at a particular counter in a certain period of time
4. Number of calls received by a customer care division in a certain period of time
5. Number of requests processed in a certain period of time

1. The number of calls received per day by a Customer Care division is observed to follow Poisson Distribution with mean calls as 56. Find the following:
   a) Probability that it may get more than 70 calls in a day
   b) Probability that less than 20 calls are received in a day

```
poisson.sf(70, 56)
0.029824687242845115

poisson.cdf(19, 56)
9.647463412493e-09
```

2. The number of customers served at a counter per hour are 4. Find the following:
   a. Probability that more than 5 customers will be served in an hour
   b. Probability that less than 3 customers will be served in an hour

```
poisson.sf(5, 4)
0.2148696129695948

poisson.cdf(2, 4)
0.23810330555354436
```

1. The number of customer returns in a retail chain per day follows a Poisson distribution at a rate of 25 returns per day. Write Python code to answer the following questions:
   (a) Calculate the probability that the number of returns exceeds 30 in a day.
   (b) If the chance of fraudulent return is 0.05, calculate the probability that there will be at least 2 fraudulent returns in any given day.

```
poisson.sf(30, 25)
0.1366911308473363

poisson.sf(1, 0.05)
0.001209104274250291
```

Uniform Distribution

Friday, June 23, 2023 12:36 PM

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{for } a \le x \le b \\ 0, & \text{otherwise} \end{cases}$$

and the cumulative distribution function is

$$F(x) = \begin{cases} 0, & \text{if } x < a \\ \dfrac{x-a}{b-a}, & \text{if } a \le x \le b \\ 1, & \text{if } b < x \end{cases}$$

The expected value and variance of a uniform random variable X are computed as follows:

$$E[X] = \frac{a+b}{2} \tag{5.20}$$

$$\text{Var}[X] = \frac{(b-a)^2}{12} \tag{5.21}$$

35. The time required to play a game of Battleship™ is uniformly distributed between 20 and 60 minutes.

a. Find the expected value and variance of the time to complete the game.

b. What is the probability of finishing within 30 minutes?

c. What is the probability that the game would take longer than 40 minutes?

Uniform (20, 60)

(loc, loc+scale)
loc=20, loc+scale=60
scale=40

Suppose that sales revenue, X, for a product varies uniformly each week between $a = \$1{,}000$ and $b = \$2{,}000$. The density function is $f(x) = 1/(2{,}000 - 1{,}000) = 1/1{,}000$ and is shown in Figure 5.15. Note that the area under the density function is 1, which you can easily verify by multiplying the height by the width of the rectangle.

Suppose we wish to find the probability that sales revenue will be less than $x = \$1{,}300$. We could do this in two ways. First, compute the area under the density function using geometry, as shown in Figure 5.16. The area is $(1/1{,}000)(300) = 0.30$. Alternatively, we could use formula (5.19) to compute $f(1{,}300)$:

$$F(1{,}300) = (1{,}300 - 1{,}000)/(2{,}000 - 1{,}000) = 0.30$$

In either case, the probability is 0.30.

Now suppose we wish to find the probability that revenue will be between \$1,500 and \$1,700. Again, using geometrical arguments (see Figure 5.17), the area of the rectangle between \$1,500 and \$1,700 is $(1/1{,}000)(200) = 0.2$. We may also use formula (5.17) and compute it as follows:

$$\begin{aligned} P(1{,}500 \le X \le 1{,}700) &= P(X \le 1{,}700) - P(X \le 1{,}500) \\ &= F(1{,}700) - F(1{,}500) \\ &= \frac{(1{,}700 - 1{,}000)}{(2{,}000 - 1{,}000)} - \frac{(1{,}500 - 1{,}000)}{(2{,}000 - 1{,}000)} \\ &= 0.7 - 0.5 = 0.2 \end{aligned}$$

Exponential

Friday, June 23, 2023 1:12 PM

The exponential distribution has the density function

$$f(x) = \lambda e^{-\lambda x}, \quad \text{for } x \geq 0$$

and its cumulative distribution function is

$$F(x) = 1 - e^{-\lambda x}, \quad \text{for } x \geq 0$$

The expected value of the exponential distribution is $1/\lambda$ and the variance is $(1/\lambda)^2$.

If number of arrivals at a counter follow Poisson Distribution then the inter-arrival time follows Exponential Distribution

8:00	8:04	8:14	8:14	8:32	8:40	8:44	9:00	9:03

X: number of arrivals in a 15 minutes duration follows Poisson
Y:Inter-arrival time follow Exponential
0, 4, 10, 10, 18, 8, 4, 16, 3

43. The actual delivery time from a pizza delivery company is exponentially distributed with a mean of 28 minutes.

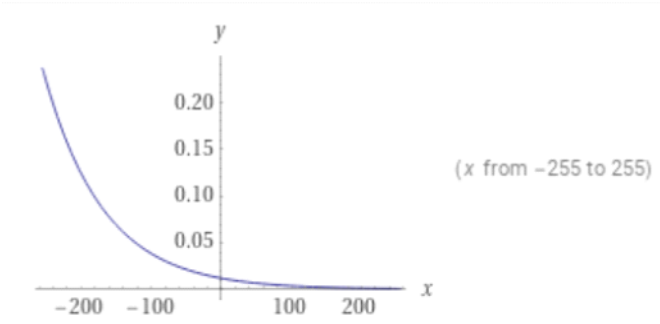
a. What is the probability that the delivery time will exceed 31 minutes?

b. What proportion of deliveries will be completed within 25 minutes?

```
expon.sf(31, loc=1/28, scale=28)
0.3309237320446033

expon.cdf(25, loc=1/28, scale=28)
0.5899932404256827
```

3. The time-of-failure of a machine follows exponential distribution with mean time between failures (MTBF) estimated to be 85 hrs. Write code to answer the following questions:
   (a) Calculate the probability that the system will fail before 85 hrs.
   (b) Calculate the probability that it will not fail up to 150 hrs.




```
expon.cdf(85, loc=1/85, scale=85)
0.6320696377349755

expon.sf(150, loc=1/85, scale=85)
0.17126084522766885
```

Normal

Friday, June 23, 2023 3:14 PM

Example 1:

Suppose that the height of a female in a geographical region is normally distributed with μ = 64 inches and σ = 4 inches.

- What is the probability of finding a woman who will be less than 58 inches tall ?

```
norm.cdf(58, 64, 4)
0.06680720126885807
```

Example 2 :

Suppose the weight of a typical male in a geographical region follows a normal distribution with μ = 180lb and σ = 30lb.

What fraction of males weigh more than 200 pounds?

```
norm.sf(200, 180, 30)
0.2524925375469229
```

37. In determining bike mileage ratings, it was found that the mpg (X) for a certain model is normally distributed, with a mean of 34 mpg and a standard deviation of 1.9 mpg. Find the following:

a. $P(X < 33)$

b. $P(31 < X < 38)$

c. $P(X > 36)$

d. $P(X > 33)$

e. The mileage rating that the upper 6% of bikes achieve.

```
norm.cdf(33, 34, 1.9)
0.2993344071288827

norm.cdf(38, 34, 1.9) - norm.cdf(31, 34, 1.9)
0.9251917315155274

norm.sf(36, 34, 1.9)
0.1462549390919427

norm.sf(33, 34, 1.9)
0.7006655928711173
```

38. The distribution of SAT scores in math for an incoming class of business students has a mean of 610 and standard deviation of 20. Assume that the scores are normally distributed.

a. Find the probability that an individual's SAT score is less than 600.

b. Find the probability that an individual's SAT score is between 590 and 620.

c. Find the probability that an individual's SAT score is greater than 650.

d. What scores will the top 5% of students have?

e. Find the standardized values for students scoring 540, 600, 650, and 700 on the test. Explain what these mean.

```
norm.cdf(600, 610, 20)
0.3085375387259869

norm.cdf(620, 610, 20) - norm.cdf(590, 610, 20)
0.532807207342556

norm.sf(650, 610, 20)
0.022750131948179195

norm.ppf(0.95,610, 20 )
642.8970725390294
```

```
In [26]: import numpy as np

In [27]: values = np.array([540, 600, 650, 700])

In [28]: print("Standardized Values:", (values-610)/20)
Standardized Values: [-3.5 -0.5  2.   4.5]
```

4. As per a survey on use of pesticides among 1000 farmers in grape farming for around 10 acres of grape farmland, it was found that the grape farmers spray 38 liters of pesticides in a week on an average with the corresponding standard deviation of 5 liters. Assume that the pesticide spray per week follows a normal distribution. Write code to answer the following questions:
   (a) What proportion of the farmers is spraying more than 50 liters of pesticide in a week?
   (b) What proportion of farmers is spraying less than 10 liters?
   (c) What proportion of farmers is spraying between 30 liters and 60 liters?

```
norm.sf(50, 38, 5)
0.008197535924596131

norm.cdf(10, 38, 5)
1.07175902583108887e-08

norm.cdf(60, 38, 5) - norm.cdf(30, 38, 5)
0.9451952957565343
```

A fast-food restaurant sells As and Bs. On a typical weekday the demand for As is normally distributed with mean 313 and standard deviation 57; the demand for Bs is normally distributed with mean 93 and standard deviation 22.
A) How many As must the restaurant stock to be 98% sure of not running out of stock on a given day ?
B) How many Bs must the restaurant stock to be 90% sure of not running out on a given day ?
C) If the restaurant stocks 450 As and 150 Bs for a given day, what is the probability that it will run out of As or Bs (or both) that day ? Assume that the demand for As and Bs are probabilistically independent.

$$P(Xa > 450 \cup Xb > 150) = P(Xa > 450) + P(Xb > 150) - P(Xa > 450 \cap Xb > 150)$$
$$= P(Xa > 450) + P(Xb > 150) - P(Xa > 450) * P(Xb > 150)$$

Reason of Simulation with Predictive Modelling

The data obtained may be very small in volume at times. But for better performance, we require it to be bigger. Hence a fictitious data with real world characteristics can be generated with simulation techniques like Monte-Carlo.

10. A metal pistons manufacturer conducts a marketing research and finds that for every 10 pistons made, an average of 12% of its pistons are rejected because they are not correctly sized. Generate 20 random variates for the number of pistons that would be rejected to estimate the minimum and maximum number that might be expected.

Generate a random number set for 200 batches of pistons

6. The exponential distribution of the amount of time a car battery lasts has a mean of 4 years. Generate 20 random variates from this distribution as whole numbers.

Generate data for 200 batteries

3. The weekly demand of a slow-moving product has the following probability mass function:

Demand, *x*	Probability, *f(x)*
0	0.1
1	0.3
2	0.2
3	0.4
4 or more	0

X	Prob	Cumulative Probabilities	Range of Random Numbers
0	0.1	P(X<=0)=0.1	0 - <0.1
1	0.3	P(X<=1)=0.4	0.1 - <0.4
2	0.2	P(X<=2)=0.6	0.4 - <0.6
3	0.4	P(X<=3)=1	0.6 - <1

E(X) = 1.9

0.24074492	0.86864199	0.54745293	0.7191533	Mean
1	3	2	3	2.25

Steps for Monte-Carlo Simulation Method:

1. For the probability distribution, calculate cumulative probabilities
2. Form the range of random numbers based on cumulative probabilities
3. Generate the random numbers from uniform distribution
4. Identify the range of random number and assign the corresponding value to it
5. The values generated in step 4 would be the simulated data

Sick drivers problem

- At a bus terminal every bus should leave with the driver. At the terminus they keep 2 drivers as reserved if any one on scheduled duty is sick and could not come. Following is the probability distribution that driver becomes sick:

No. of Absent Drivers	0	1	2	3	4	5
Probability	0.30	0.20	0.15	0.10	0.13	0.12

Simulate the data for a month and find utilization of reserved drivers. Also find how many days and how many buses cannot run because of non-availability of drivers.

x	cp	range	
0	0.3	0 - <0.3	
1	0.5	0.3 - <0.5	

[0.65626419 0.95490138 0.97087603 0.09783218 0.24074492 0.86864199 0.54745293]

0	0.3	0 - <0.3	
1	0.5	0.3 -<0.5	
2	0.65	0.5 -<0.65	
3	0.75	0.65 -<0.75	
4	0.88	0.75-<0.88	
5	1	0.88 -<1	

[0.65626419 0.95490138 0.97087603 0.09783218
0.24074492 0.86864199 0.54745293]

Sim:
3, 5, 5, 0, 0, 4, 2

12. A research study on the duration of two-month-old babies' smile was carried out. Analysts created the following distribution, which were assumed to be uniform over various intervals, each with a discrete probability (all in seconds).

a	*b*	Probability
1	2	0.12
1	3	0.35
2	3	0.28
2	5	0.14
2	6	0.08
3	6	0.03

Sampling Methods

Many types of sampling methods exist. Sampling methods can be *subjective* or *probabilistic*. Subjective methods include **judgment sampling**, in which expert judgment is used to select the sample (survey the "best" customers), and **convenience sampling**, in which samples are selected based on the ease with which the data can be collected (survey all customers who happen to visit this month). Probabilistic sampling involves selecting the items in the sample using some random procedure. Probabilistic sampling is necessary to draw valid statistical conclusions.

The most common probabilistic sampling approach is simple random sampling. **Simple random sampling** involves selecting items from a population so that every subset of a given size has an equal chance of being selected. If the population data are stored in a database, simple random samples can generally be easily obtained.

1. SRSWOR
2. SRSWR
3. Systematic Random Sampling
   (1,2,….100) -->
   (6,16,26,36,46,56,…96)
4. Stratified Random Sampling : Based on categories. e.g. (Males: 1500, Females: 500) --> (75, 25)
5. Cluster Sampling: Data is divided into different clusters/groups. Some clusters are chosen at random.

12. Using the data in the Excel file *Consumer Transportation Survey*, test the following null hypotheses:

a. Individuals spend at least eight hours per week in their vehicles.

b. Individuals drive an average of 600 miles per week.

c. The average age of SUV drivers is no greater than 35.

a. H0: Hours per week >= 8 Vs H1: Hours per week < 8

```
stats.ttest_1samp(consumer['# of hours per week in vehicle'],
                  popmean=8.0,
                  alternative='less')
## Conclusion: Hours per week may be at least 8.
TtestResult(statistic=0.21908455193438453, pvalue=0.5862528908348619, df=49)
```

b. H0: Miles per week = 600 Vs H1: Miles per week

```
In [6]: stats.ttest_1samp(consumer['Miles driven per week'],
   ...:                   popmean=600,
   ...:                   alternative='two-sided')
   ...: ## Conclusion: Miles pwe week may not be 600
Out[6]: TtestResult(statistic=-2.369407386313186, pvalue=0.02180041862974647,
df=49)
```

c. H0: Age of SUV <= 35 Vs H1: Age of SUV > 35

20. An industry trade publication stated that the average profit per customer for this industry was greater than $4,500. The Excel file *Sales Data* provides data on a sample of customers. Using a test of hypothesis, do the data support this claim or not?

H0: avg profit per customer <= 4500

H1: avg profit per customer > 4500

```
    ...: stats.ttest_1samp(sales['Gross Profit'].dropna(),
    ...:                   popmean=4500, alternative='greater')
    ...: ## Conclusion: avg profit per customer may not be greater than 4500
Out[17]: TtestResult(statistic=-0.3476456590343202,
pvalue=0.6353282443468329, df=59)
```

16. Using the data in the Excel file *Airport Service Times*, determine if the airline can claim that its average service time is less than 2.5 minutes.

H0: Times >= 150 Vs H1: Times < 150

```
    ...: stats.ttest_1samp(airport['Times (sec.)'].dropna(),
    ...:                   popmean=150, alternative='less')
    ...: ## Conclusion: Service time may be less than 2.5 minutes
Out[22]: TtestResult(statistic=-6.426512207272327,
pvalue=1.1126391554084507e-10, df=811)
```

Paired t

Saturday, June 24, 2023 3:31 PM

38. Using the data in the Excel file *Ohio Education Performance*, test the hypotheses that the mean difference in writing and reading scores is zero and that the mean difference in math and science scores is zero. Use the paired-sample procedure.

```
    ...: stats.ttest_rel(ohio['Writing'], ohio['Reading'])
    ...: ## Conclusion: Writing and reading may not be equal
Out[45]: TtestResult(statistic=3.150334504876617,
pvalue=0.0036795066022806064, df=30)

In [46]:
    ...:
    ...: stats.ttest_rel(ohio['Math'], ohio['Science'])
    ...: ## Conclusion: Math and Science may not be equal
Out[46]: TtestResult(statistic=-7.710330061853673,
pvalue=1.3336301941932849e-08, df=30)
```

A group of seven patients of rheumatic heart disease with distention of abdomen due to ascites, affecting breathing capacity were treated. Can we say that treatment has improved breathing capacity? Data is in file Rheumatic.csv

Maximum breathing capacity (L/min for 7 patients)	1	2	3	4	5	6	7
Before Treatment	102	89	32	82	36	56	79
After Treatment	132	116	50	82	61	64	92

```
    ...: stats.ttest_rel(rhum['Before'], rhum['After'],
    ...:                 alternative="less")
    ...: ## Conclusion: Treatment may be effective
Out[53]: TtestResult(statistic=-4.176554325821337,
pvalue=0.0029179069531987257, df=6)
```

33. In the Excel file *Cell Phone Survey*, test the hypothesis that the mean responses for value for the dollar and customer service do not differ by gender.

```
In [77]: cell = pd.read_csv("Cell Phone Survey.csv")
    ...: males = cell[cell['Gender']=='M']
    ...: females = cell[cell['Gender']=='F']
    ...:
    ...: # H0: var_males = var_fem
    ...: # H1: var_males ne var_fem
    ...: stats.bartlett(males['Value for the Dollar'],
    ...:                females['Value for the Dollar'])
    ...: ## Conclusion: Variances may be equal
Out[77]: BartlettResult(statistic=0.008475577925522698,
pvalue=0.9266480631482593)
```

```
    ...: stats.ttest_ind(males['Value for the Dollar'],
    ...:                 females['Value for the Dollar'],
    ...:                  equal_var=True)
    ...: ## Conclusion: Means may be equal
Out[78]: Ttest_indResult(statistic=0.18568833497751558,
pvalue=0.8534404345590003)
```

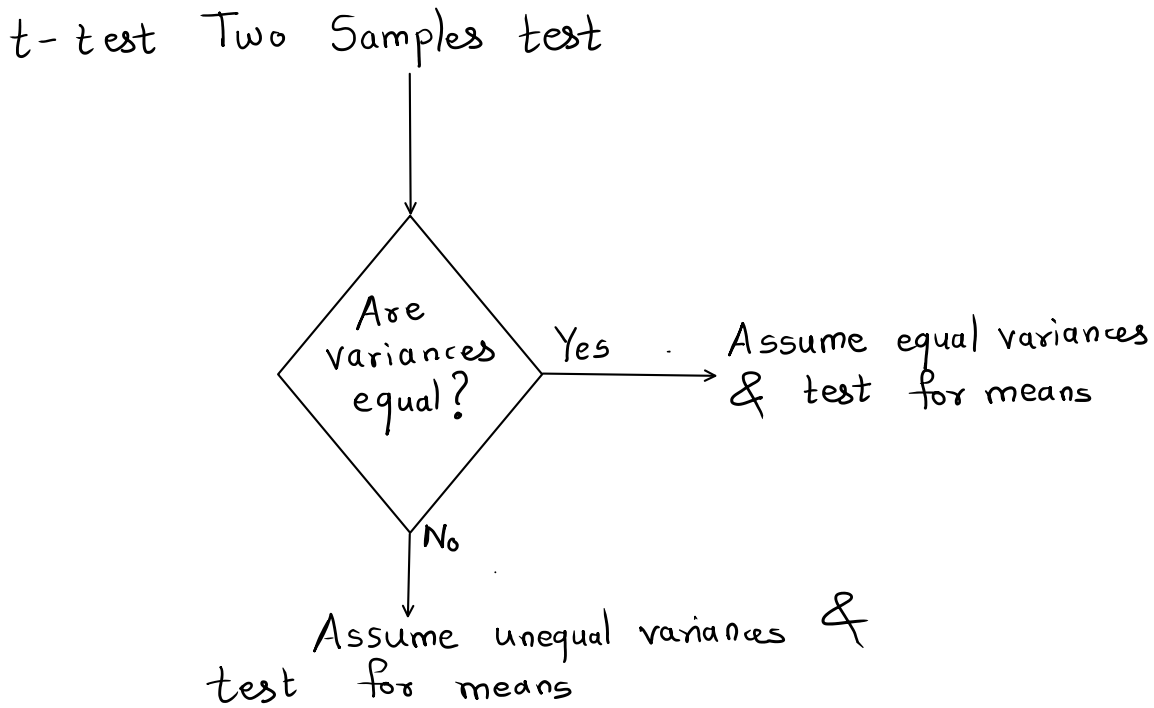
The `Puromycin.csv` data frame has 23 rows and 3 columns of the reaction velocity versus substrate concentration in an enzymatic reaction involving untreated cells or cells treated with Puromycin.

conc : a numeric vector of substrate concentrations (ppm)
rate : a numeric vector of instantaneous reaction rates (counts/min/min)
state: a factor with levels `treated untreated`
Is there any mean population difference in values of **rate** for two different treatments?

The soporific effect of drugs A and B was studied on ten patients separately. The results were assessed for the additional hours of sleep produced by the drugs. Compare soporific effects of the drugs from the data in file Soporific.csv. Test whether variances and means for drug A and drug B are equal or not.

$H_0: \sigma_A^2 = \sigma_B^2 \qquad H_1: \sigma_A^2 \neq \sigma_B^2$

$H_0: \mu_A = \mu_B \qquad H_1: \mu_A \neq \mu_B$



Test for means of 2 independent samples:

1. Parametric: t-test. Assumes that the two populations are Normally Distributed
2. Non-Parametric Test: Mann-Whitney U test. Does not assume any distribution of two populations

Tests of Normality: These tests check whether the distribution of the population is Normal or not
H0: Population is Normal
H1: Population is not Normal

1. Shapiro-Wilk Test
2. Anderson-Darling Test
3. Kolmogorov-Smirnov Test

The `Puromycin.csv` data frame has 23 rows and 3 columns of the reaction velocity versus substrate concentration in an enzymatic reaction involving untreated cells or cells treated with Puromycin.

conc : a numeric vector of substrate concentrations (ppm)
rate : a numeric vector of instantaneous reaction rates (counts/min/min)
state: a factor with levels `treated untreated`
Is there any mean population difference in values of **rate** for two different treatments?

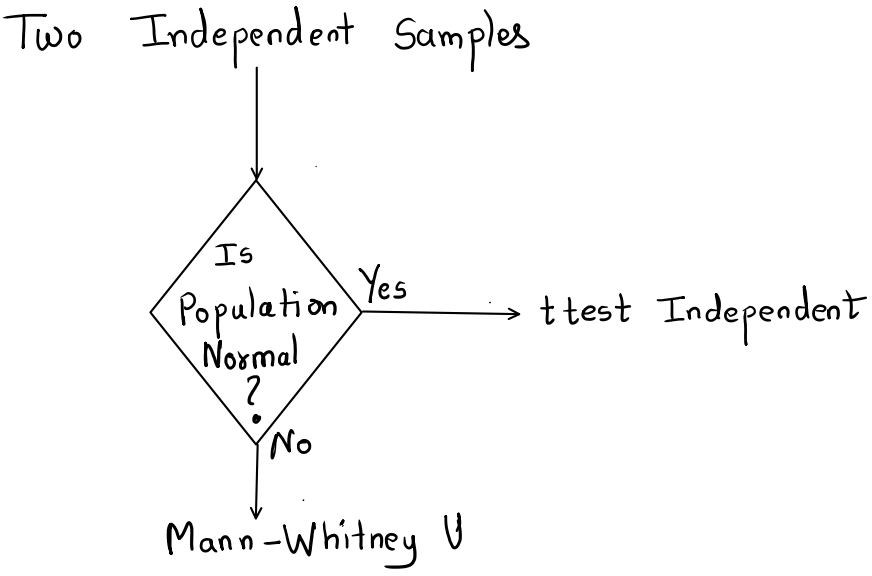
```
stats.mannwhitneyu(treated['rate'],
                   untreated['rate'])
MannwhitneyuResult(statistic=88.0, pvalue=0.18575772936738766)
```

33. In the Excel file *Cell Phone Survey*, test the hypothesis that the mean responses for value for the dollar and customer service do not differ by gender.

```
In [20]: print(stats.shapiro(males['Value for the Dollar']))
ShapiroResult(statistic=0.8909915089607239, pvalue=0.002674493007361889)

In [21]: print(stats.shapiro(females['Value for the Dollar']))
ShapiroResult(statistic=0.87392657995224, pvalue=0.020663034170866013)

In [22]: stats.mannwhitneyu(males['Value for the Dollar'],
    ...:                    females['Value for the Dollar'])
Out[22]: MannwhitneyuResult(statistic=322.0, pvalue=0.7518661276232965)
```



The soporific effect of drugs A and B was studied on ten patients separately. The results were assessed for the additional hours of sleep produced by the drugs. Compare soporific effects of the drugs from the data in file Soporific.csv. Test whether variances and means for drug A and drug B are equal or not.

```
In [32]: print(stats.shapiro(sop['Drug A']))
    ...: print(stats.shapiro(sop['Drug B']))
    ...:
    ...: # H0: distribution_A = distribution_B
    ...: # H1: distribution_A ne distribution_B
    ...: stats.mannwhitneyu(sop['Drug A'],
    ...:                    sop['Drug B'])
    ...: ## Conclusion:  Distributions may be equal
ShapiroResult(statistic=0.8839643597602844, pvalue=0.1448628455400467)
ShapiroResult(statistic=0.9027420878410339, pvalue=0.2347579002380371)
Out[32]: MannwhitneyuResult(statistic=35.0, pvalue=0.2719522794479482)
```

ANOVA

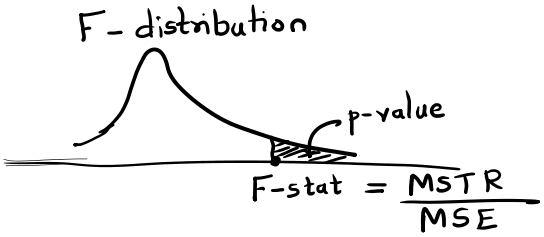
Monday, June 26, 2023 3:16 PM

r: no. of treatments
n: Total obs.

Sources of Variation	Sums of Squares	Degrees of freedom	Mean Square	F Ratio (F-stat)	P-Value
Treatment	SSTR	r – 1	MSTR=SSTR / (r – 1)	MSTR/MSE	
Error	SSE	n – r	MSE = SSE / (n – r)		
Total	SST	n – 1			

MSTR: measure of Between Variation
MSE: measure of within variation

If MSTR > MSE ⇒ Trt unequal
o.w. Trt equal



$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1: Not all μ_i (i = 1,2,3,4) are equal

	sum_sq	df	F	PR(>F)	
Treatments	1551.607762	3.0	18.293252	0.000006	< 0.05
Residual	565.457238	20.0	NaN	NaN	

∴ We reject H_0 at 5% l.o.s.

Conclusion:- Treatments may not be homogeneous

H_0: Group2 $\overset{?}{=}$ Group1

group1	group2	meandiff	p-adj	lower	upper	reject	
I	II	13.0976	0.0014	4.8177	21.3775	True	$\mu_{II} > \mu_I$
I	III	-0.6567	0.9969	-9.6685	8.3552	False	$\mu_{III} = \mu_I$
I	IV	18.1000	0.0001	9.5075	26.6925	True	$\mu_{IV} > \mu_I$
II	III	-13.7543	0.0014	-22.4686	-5.0399	True	$\mu_{II} > \mu_{III}$
II	IV	5.0024	0.3541	-3.2775	13.2823	False	$\mu_{II} = \mu_{IV}$
III	IV	18.7567	0.0001	9.7448	27.7685	True	$\mu_{IV} > \mu_{III}$

```
In [51]: agr.groupby('Treatments')['Yield'].mean()
Out[51]:
Treatments
I      23.716667
II     36.814286
III    23.060000
IV     41.816667
```

$\mu_{II}, \mu_{IV} > \mu_I, \mu_{III}$

A college is trying to determine if there is a significant difference in the mean GMAT score of students from different undergraduate backgrounds who apply to the MBA program. The Excel file *GMAT Scores* contains data from a sample of students. What conclusion can be reached using ANOVA?

	sum_sq	df	F	PR(>F)
Major	2983.945344	2.0	14.947815	0.00002
Residual	3493.423077	35.0	NaN	NaN

group2 − group1

group1	group2	meandiff	p-adj	lower	upper	reject
Business	Liberal Arts	-12.0769	0.0371	-23.5391	-0.6147	True
Business	Sciences	11.4231	0.0093	2.5240	20.3222	True
Liberal Arts	Sciences	23.5000	0.0000	12.6092	34.3908	True

LA < Bus
Sci > B
Sci > LA
Sci > B > LA

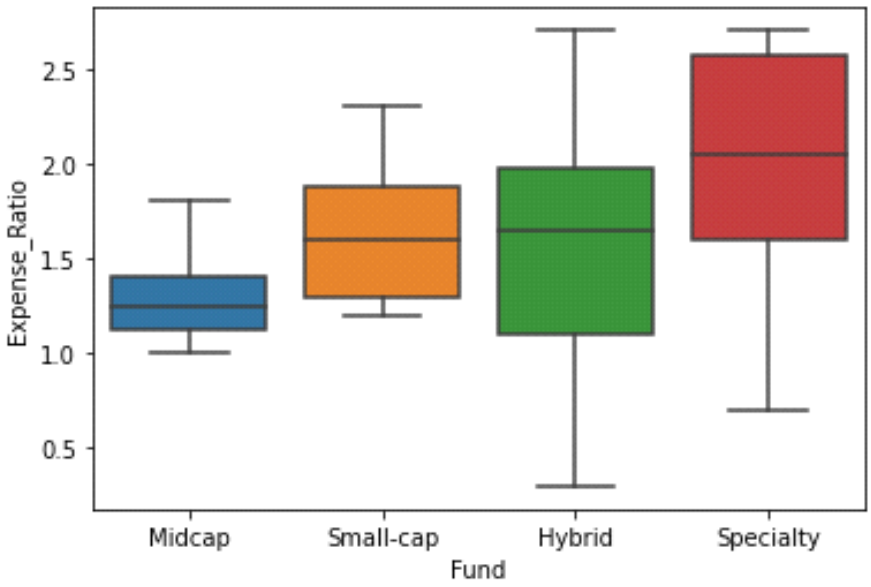
Conclusion: Science mean GMAT > Business mean GMAT > Liberal Art Mean GMAT

- *A magazine reports percentage returns and expense ratios for stock and bond funds.* The data FUNDS.csv are the expense ratios for 10 midcap stock funds, 10 small-cap stock funds, 10 hybrid stock funds, and 10 specialty stock funds.
- Test for any significant difference in the mean expense ratio among the four types of stock funds.

	sum_sq	df	F	PR(>F)
Fund	2.603	3.0	2.94346	0.045936
Residual	10.612	36.0	NaN	NaN

group1	group2	meandiff	p-adj	lower	upper	reject
Hybrid	Midcap	-0.32	0.5578	-0.9739	0.3339	False
Hybrid	Small-cap	0.02	0.9998	-0.6339	0.6739	False
Hybrid	Specialty	0.40	0.3659	-0.2539	1.0539	False
Midcap	Small-cap	0.34	0.5074	-0.3139	0.9939	False
Midcap	Specialty	0.72	0.0262	0.0661	1.3739	True
Small-cap	Specialty	0.38	0.4108	-0.2739	1.0339	False

Spec > Midcap


Expense_Ratio
2.5
2.0
1.5
1.0
0.5
Midcap
Small-cap
Hybrid
Specialty
Fund

47. The cross-tabulation data given below represent the number of males and females in a work group who feel overstressed and those who don't.

Overstressed	Women	Men
No	9	4
Yes	6	9

a. Write the hypotheses for the chi-square test for independence.

b. Find the expected frequencies.

c. Compute the chi-square statistic using a level of significance of 0.05.

Row Attribute: Feeling overstressed

Column Attribute: Gender

H_0: Feeling overstressed or not is independent of Gender

H_1: Feeling overstressed or not dependent on Gender

Overstressed	Women	Men	Total
No	9 $=O_{11}$	4 $=O_{12}$	13 $=R_1$
Yes	6 $=O_{13}$	9 $=O_{14}$	15 $=R_2$
Total	15 $=C_1$	13 $=C_2$	28 $=n$

$E_{11} = \frac{R_1 C_1}{n}$

$E_{12} = \frac{R_1 C_2}{n}$

$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

$= \frac{(9-6.96)^2}{6.96} + \frac{(4-6.03)^2}{6.03} + \frac{(6-8.03)^2}{8.03} + \frac{(9-6.96)^2}{6.96}$

Overstressed	Women	Men
No	E11=13*15/28=6.96	E12=13*13/28=6.03
Yes	E21=15*15/28=8.03	E22=15*13/28=6.96

```
In [43]: obs = np.array([[9,4],
    ...:                [6,9]])

In [44]: chi2_contingency(obs,correction=False)
Out[44]:
Chi2ContingencyResult(statistic=2.3924260355029583, pvalue=0.12192428411669565,
dof=1, expected_freq=array([[6.96428571, 6.03571429],
       [8.03571429, 6.96428571]]))
```

> 0.05

Conclusion: Feeling overstressed may be not be related to gender

49. The following cross-tabulation shows the number of people who rated a customer service representative as friendly and polite based on whether the representative greeted them first.

a. Write the hypotheses for the chi-square test for independence.

b. Find the expected frequencies.

c. Compute the chi-square statistic using a level of significance of 0.01.

d. Find the chi-square critical value and *p*-value and draw a conclusion.

	Friendly/Polite	
Staff Greeting	**No**	**Yes**
No	13	7
Yes	12	22

```
In [45]: obs = np.array([[13,7], [12,22]])

In [46]: print(chi2_contingency(obs,correction=False))
Chi2ContingencyResult(statistic=4.469403651115618, pvalue=0.034507033347004896
dof=1, expected_freq=array([[ 9.25925926, 10.74074074],
       [15.74074074, 18.25925926]]))
```

>0.01

∴ We don't reject Ho at 1 % l.o.s.

Conclusion:- Being friendly/Polite not related to staff greeting.

55. For the data in the Excel file *New Account Processing*, perform chi-square tests for independence to determine if certification is independent of gender and if certification is independent of having prior industry background.

```
In [60]:
    ...: a = pd.crosstab(index=acc['Gender'],
    ...:                 columns=acc['Certified'])
    ...: print(chi2_contingency(a,correction=False))
Chi2ContingencyResult(statistic=0.054727773462326004, pvalue=0.8150318082955932,
dof=1, expected_freq=array([[7.67741935, 6.32258065],
       [9.32258065, 7.67741935]]))
```

```
In [61]:
    ...: b = pd.crosstab(index=acc[' Prior Background'],
    ...:                 columns=acc['Certified'])
    ...: print(chi2_contingency(b, correction=False))
Chi2ContingencyResult(statistic=8.957200152788387, pvalue=0.0027637813584297775,
dof=1, expected_freq=array([[ 6.03225806,  4.96774194],
       [10.96774194,  9.03225806]]))
```

```
In [90]: ctab
Out[90]:
left             0     1
Department
IT             954   273
RandD          666   121
accounting     563   204
hr             524   215
management     539    91
marketing      655   203
product_mng    704   198
sales         3126  1010
support       1674   555
technical     2023   697
```

	Department	left	value
0	IT	0	954
1	RandD	0	666
2	accounting	0	563
3	hr	0	524
4	management	0	539
5	marketing	0	655
6	product_mng	0	704
7	sales	0	3126
8	support	0	1674
9	technical	0	2023
10	IT	1	273
11	RandD	1	121
12	accounting	1	204
13	hr	1	215
14	management	1	91

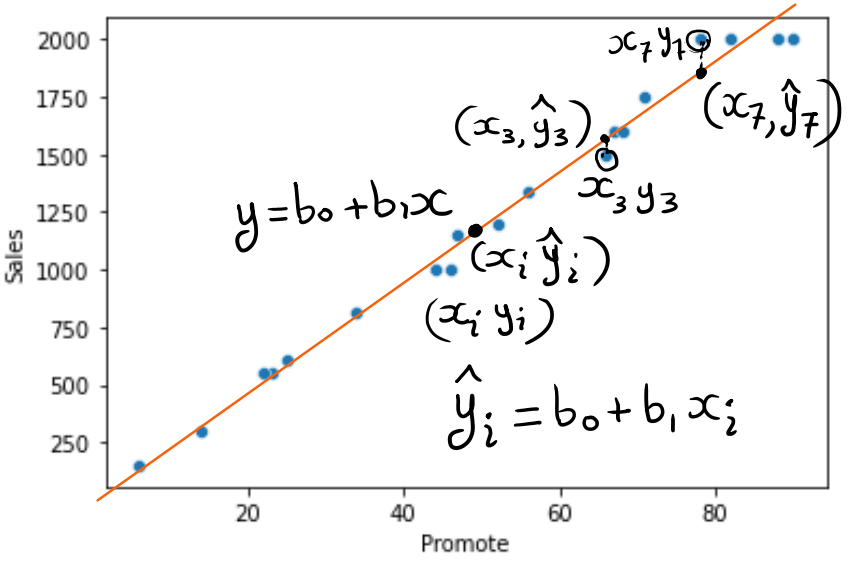
All Tests

Monday, June 26, 2023 2:37 PM

#	Scenario	Assumption	Test	Python Function
1	Testing mean of one population	Normal Distribution of Population	One Sample t-test	ttest_1samp
2	Testing the difference of means in the paired samples	Normal Distribution of Population	Paired t-test	ttest_rel
3	Testing the difference of variances in 2 independent samples	Does not assume Normal Distribution	Bartlett's test	bartlett
4	Testing the difference of means in 2 independent samples	Normal Distribution of Populations	2 independent samples t-test	ttest_ind
5	Testing the difference of distributions in 2 independent samples	Does not assume Normal Distribution	Mann-Whitney's U test	mannwhitneyu
6	Testing difference in means between multiple populations	Normal Distribution of Populations	Analysis of Variance Test (ANOVA)	statsmodels.formula.api.ols & statsmodels.api.stats.anova_lm
7	Testing the independence of attributes in a cross-table	Does not assume Normal Distribution	Chi-Square Test	stats.chi2_contingency

Regression

Tuesday, June June 27, 2023 5:27 PM



$ax + by + c = 0$

(x_1, y_1) $\quad ax_1 + by_1 + c = 0$

$y = mx + c$

? ?

Method of least squares

$y = b_0 + b_1 x$

(c) (m)

$x_1\ y_1$

$x_2\ y_2$

$\vdots$

$x_n\ y_n$

$$\text{Residual} = e_i = y_i - \hat{y}_i$$

$\sum_i (y_i - \hat{y}_i)^2$: Residual Sum of Squares

Optimization Problem :-

Find b_0, b_1 such that

$\sum_i (y_i - \hat{y}_i)^2$ is minimum

$$\hookrightarrow Z = \sum_i (y_i - b_0 - b_1 x_i)^2$$

$$\frac{\partial Z}{\partial b_0} = 0 \qquad \frac{\partial Z}{\partial b_1} = 0$$

Simultaneously

to get best values of b_0, b_1

y : Dependent or Response or

X : Independent variable(s) or Predictors or Features

Response
or
Label

Predictors
or
Features

```
In [17]: print("b0 =", lr.intercept_)
b0 = 5.4858653632529695

In [18]: print("b1 =", lr.coef_)
b1 = [23.50640302]
```

$$Sales = 5.485 + 23.506 * Promote$$

$Promote = 0 \Rightarrow Sales = 5.485$

$$Sales = 5.485 + 23.506 * 100$$

? → 23.506 ; 100 → 1

$$\text{Mean Squared error} = \frac{\sum_i (y_i - \hat{y}_i)^2}{n}$$

$$\hat{y}_i = b_0 + b_1 x \quad \text{(x ← Home)}$$

```
b0 = 23045.63894523328
b1 = [215.21298174]
```
for Home

One hot Encoding / Dummying

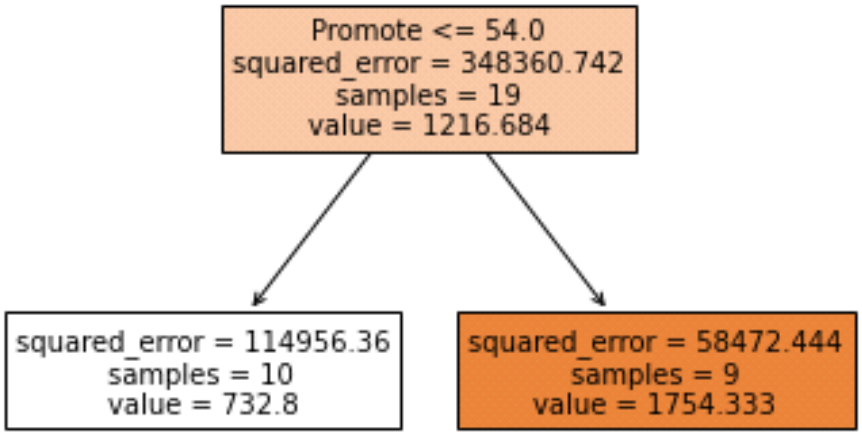
Sex
F
F
M
M
F

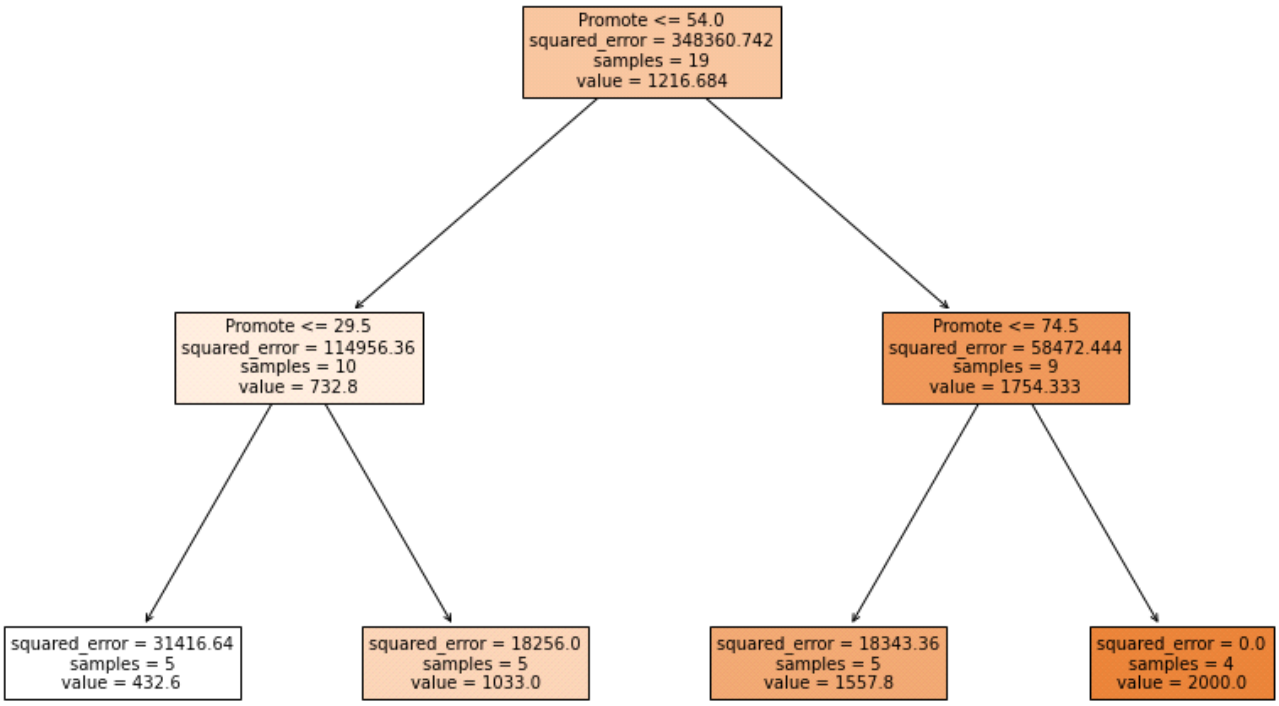
F	M
1	0
1	0
0	1
0	1
1	0

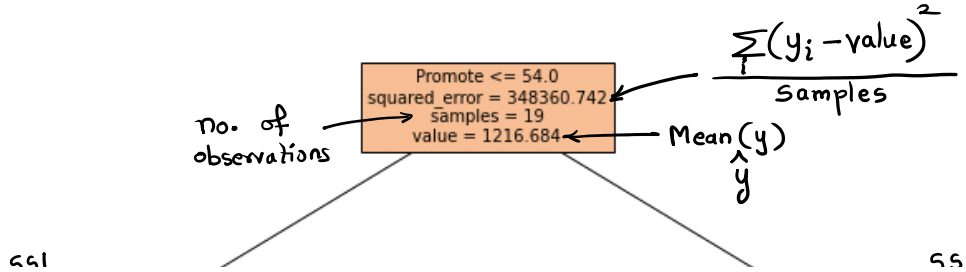
rank	Aso P	Ass P	P
P	0	0	1
Ass P	0	1	0
Aso P	1	0	0
Ass P	0	1	0
Aso P	1	0	0
P	0	0	1
P	0	0	1

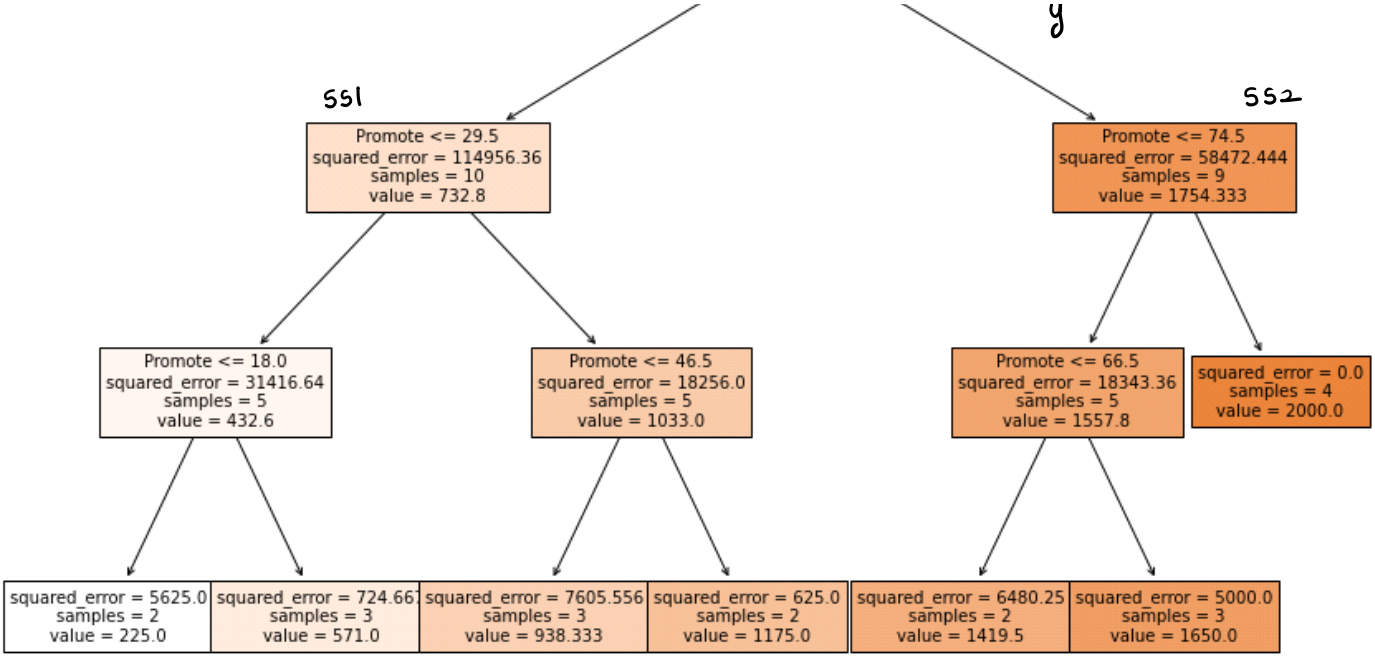
Regression Trees

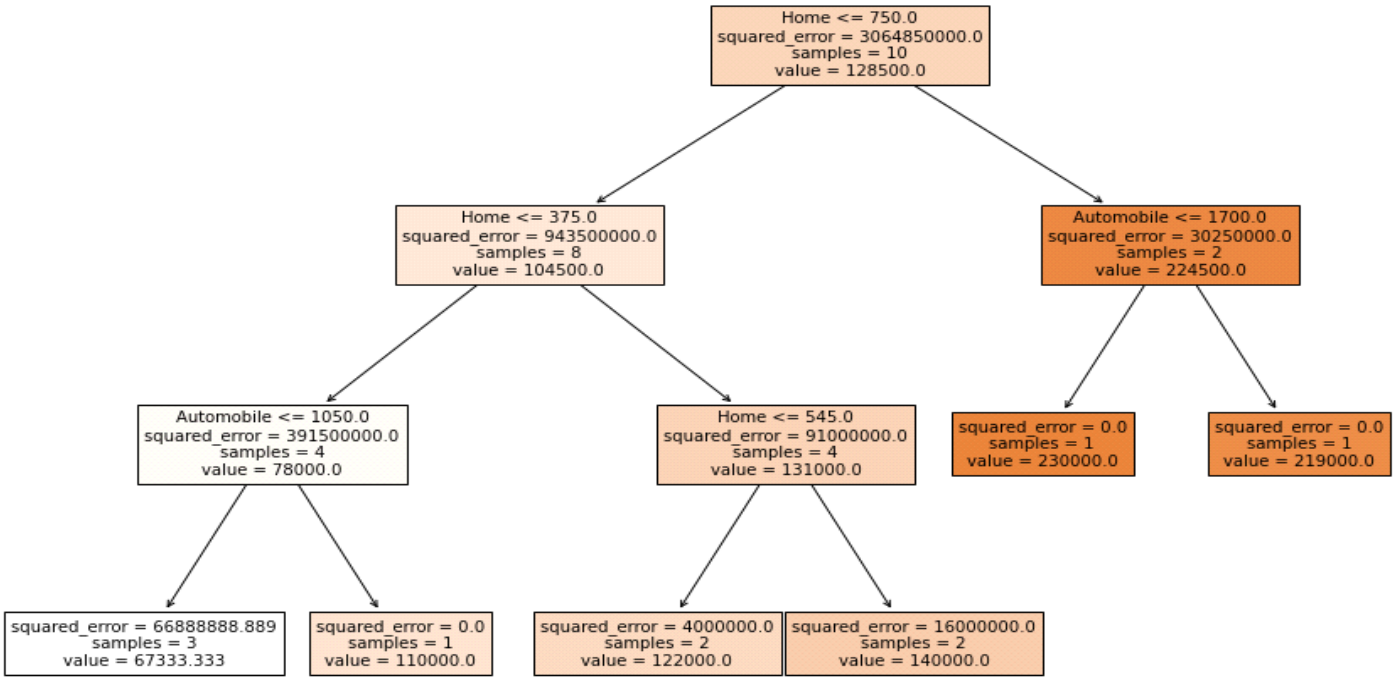
Tuesday, June 27, 2023 7:44 PM





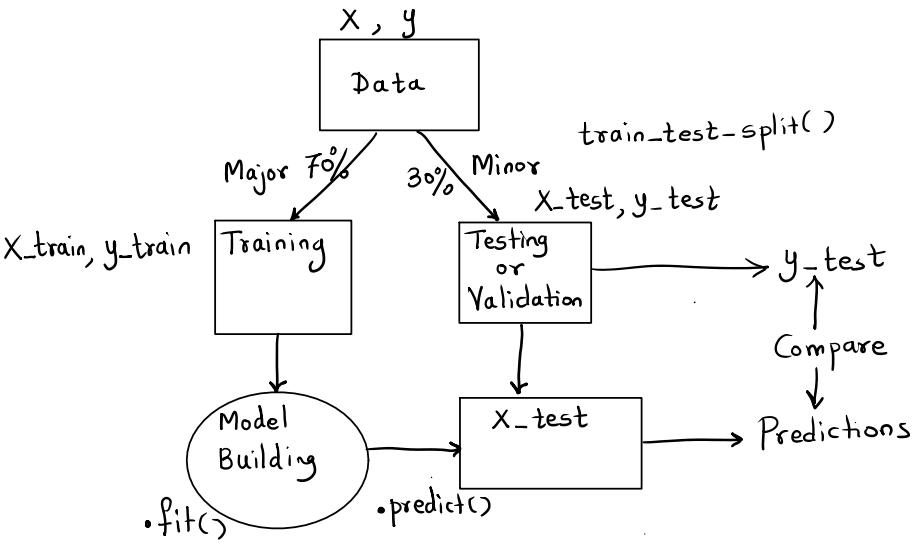


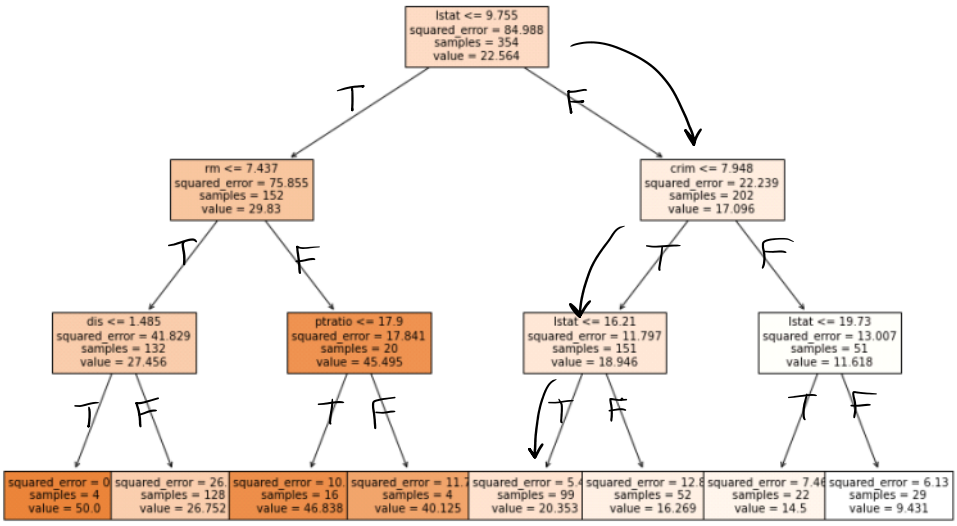

y
ss1
ss2
Promote <= 29.5
squared_error = 114956.36
samples = 10
value = 732.8
Promote <= 74.5
squared_error = 58472.444
samples = 9
value = 1754.333
Promote <= 18.0
squared_error = 31416.64
samples = 5
value = 432.6
Promote <= 46.5
squared_error = 18256.0
samples = 5
value = 1033.0
Promote <= 66.5
squared_error = 18343.36
samples = 5
value = 1557.8
squared_error = 0.0
samples = 4
value = 2000.0
squared_error = 5625.0
samples = 2
value = 225.0
squared_error = 724.661
samples = 3
value = 571.0
squared_error = 7605.556
samples = 3
value = 938.333
squared_error = 625.0
samples = 2
value = 1175.0
squared_error = 6480.25
samples = 2
value = 1419.5
squared_error = 5000.0
samples = 3
value = 1650.0

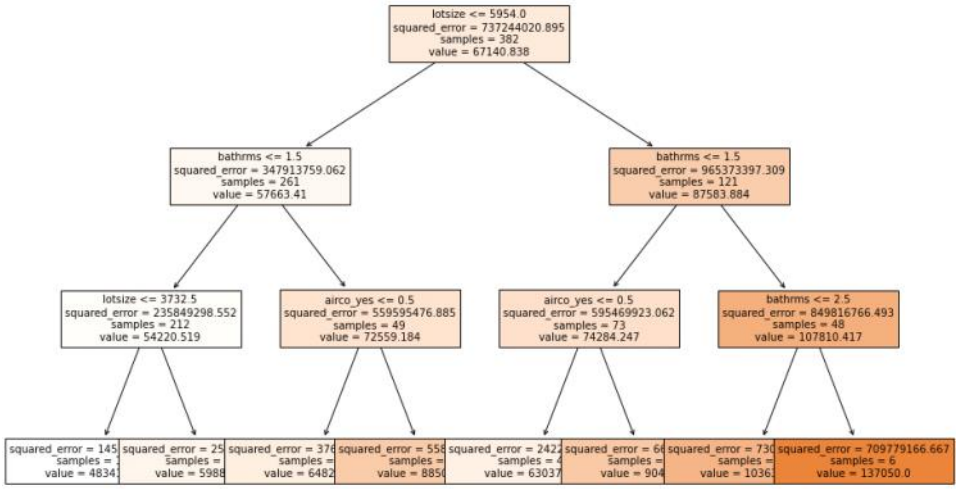

Home <= 750.0
squared_error = 3064850000.0
samples = 10
value = 128500.0
Home <= 375.0
squared_error = 943500000.0
samples = 8
value = 104500.0
Automobile <= 1700.0
squared_error = 30250000.0
samples = 2
value = 224500.0
Automobile <= 1050.0
squared_error = 391500000.0
samples = 4
value = 78000.0
Home <= 545.0
squared_error = 91000000.0
samples = 4
value = 131000.0
squared_error = 0.0
samples = 1
value = 230000.0
squared_error = 0.0
samples = 1
value = 219000.0
squared_error = 66888888.889
samples = 3
value = 67333.333
squared_error = 0.0
samples = 1
value = 110000.0
squared_error = 4000000.0
samples = 2
value = 122000.0
squared_error = 16000000.0
samples = 2
value = 140000.0

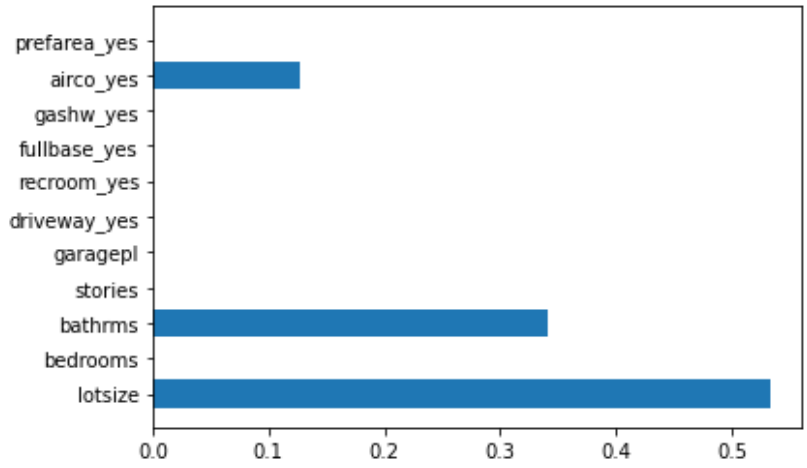
Data Partitioning

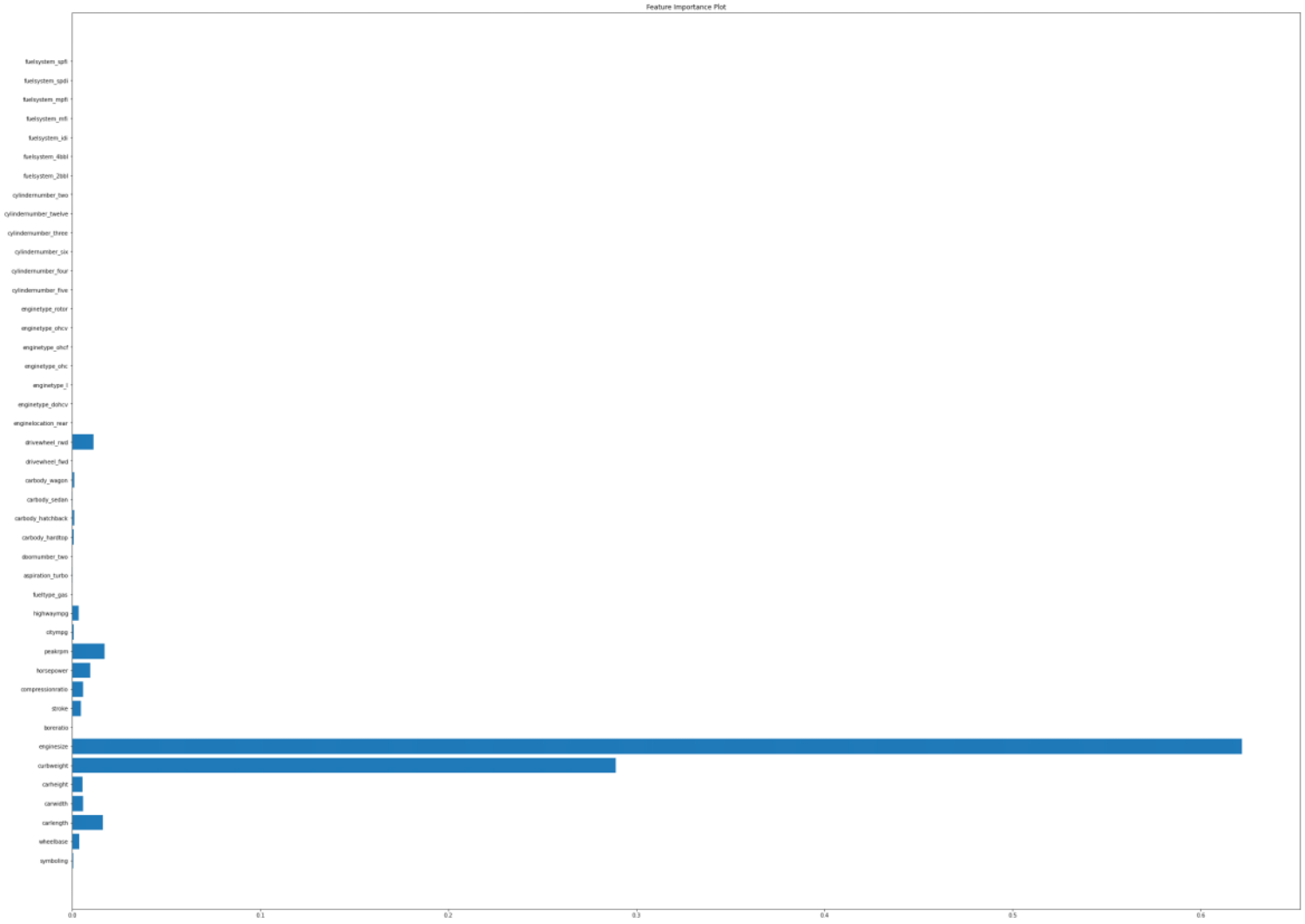
Wednesday, June 28, 2023 2:23 PM






lotsize <= 5954.0
squared_error = 737244020.895
samples = 382
value = 67140.838
bathrms <= 1.5
squared_error = 347913759.062
samples = 261
value = 57663.41
bathrms <= 1.5
squared_error = 965373397.309
samples = 121
value = 87583.884
lotsize <= 3732.5
squared_error = 235849298.552
samples = 212
value = 54220.519
airco_yes <= 0.5
squared_error = 559595476.885
samples = 49
value = 72559.184
airco_yes <= 0.5
squared_error = 595469923.062
samples = 73
value = 74284.247
bathrms <= 2.5
squared_error = 849816766.493
samples = 48
value = 107810.417
squared_error = 709779166.667
samples = 6
value = 137050.0


prefarea_yes
airco_yes
gashw_yes
fullbase_yes
recroom_yes
driveway_yes
garagepl
stories
bathrms
bedrooms
lotsize
0.0
0.1
0.2
0.3
0.4
0.5


Feature Importance Plot
fuelsystem_spfi
fuelsystem_spdi
fuelsystem_mpfi
fuelsystem_mfi
fuelsystem_idi
fuelsystem_4bbl
fuelsystem_2bbl
cylindernumber_two
cylindernumber_twelve
cylindernumber_three
cylindernumber_six
cylindernumber_four
cylindernumber_five
enginetype_rotor
enginetype_ohcv
enginetype_ohcf
enginetype_ohc
enginetype_l
enginetype_dohcv
enginelocation_rear
drivewheel_rwd
drivewheel_fwd
carbody_wagon
carbody_sedan
carbody_hatchback
carbody_hardtop
doornumber_two
aspiration_turbo
fueltype_gas
highwaympg
citympg
peakrpm
horsepower
compressionratio
stroke
boreratio
enginesize
curbweight
carheight
carwidth
carlength
wheelbase
symboling
0.0
0.1
0.2
0.3
0.4
0.5
0.6

Classification Trees

Wednesday, June 28, 2023 6:17 PM

$$gini = \sum_i f_i(1-f_i)$$

A	B
35	65

$f_1 = 0.35 \quad f_2 = 0.65$

$gini = f_1(1-f_1) + f_2(1-f_2)$

$= 0.35(0.65) + 0.65(0.35)$

$= 0.455$

A	B
25	75

$f_1 = 0.25 \quad f_2 = 0.75$

$gini = 0.375$

A	B
10	90

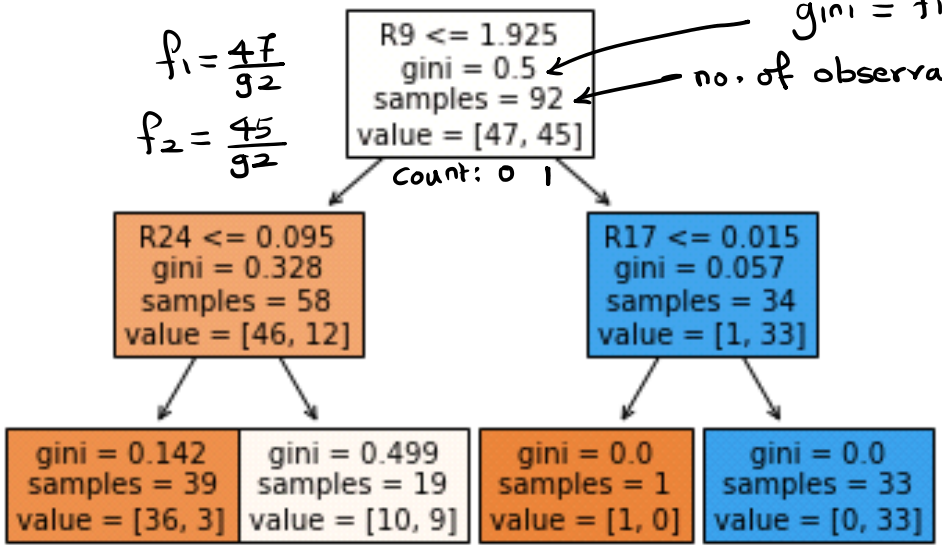
$gini = 0.18$

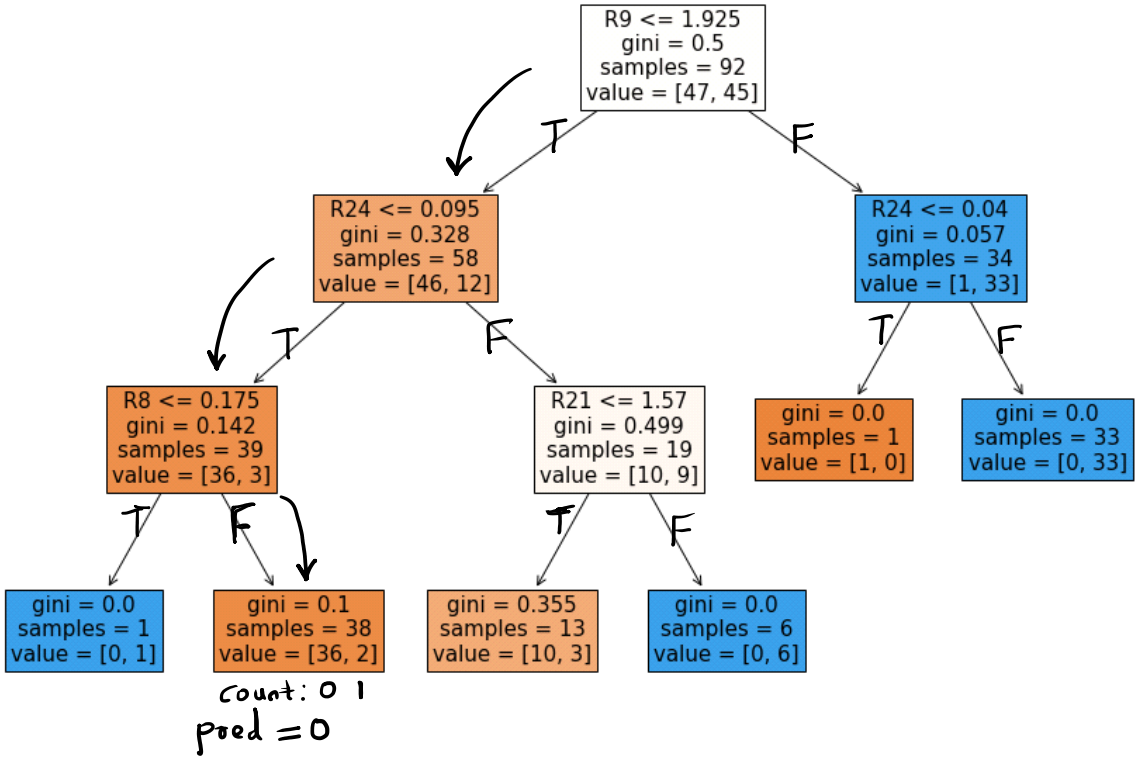
$f_1 = \frac{47}{92}$

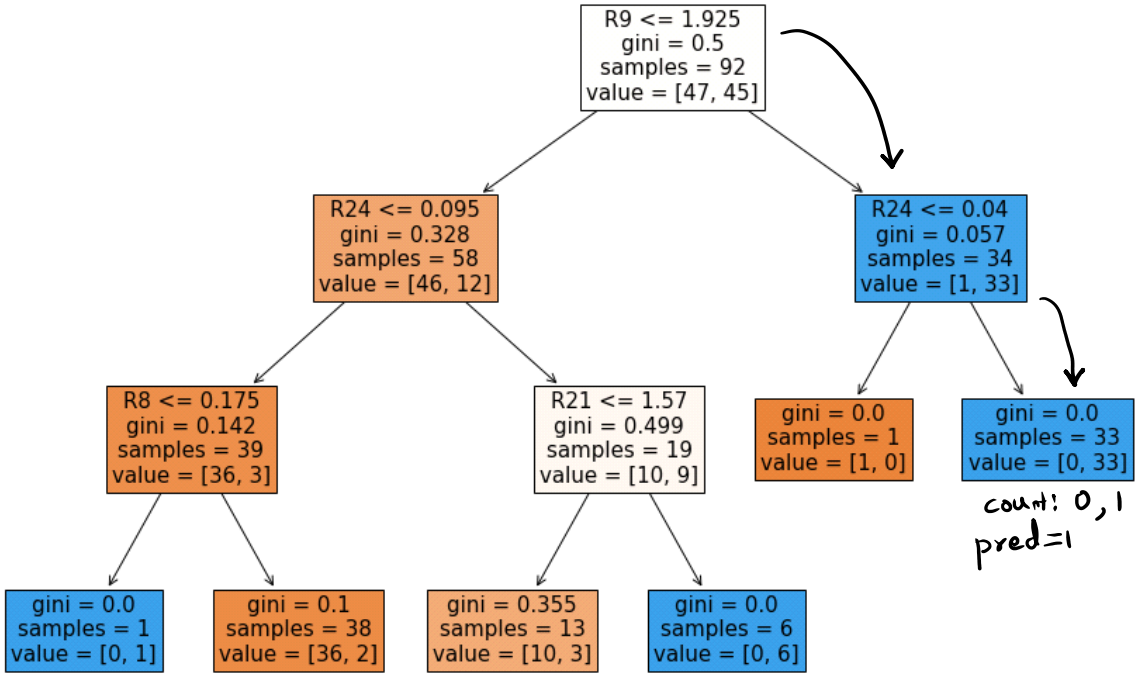
$f_2 = \frac{45}{92}$

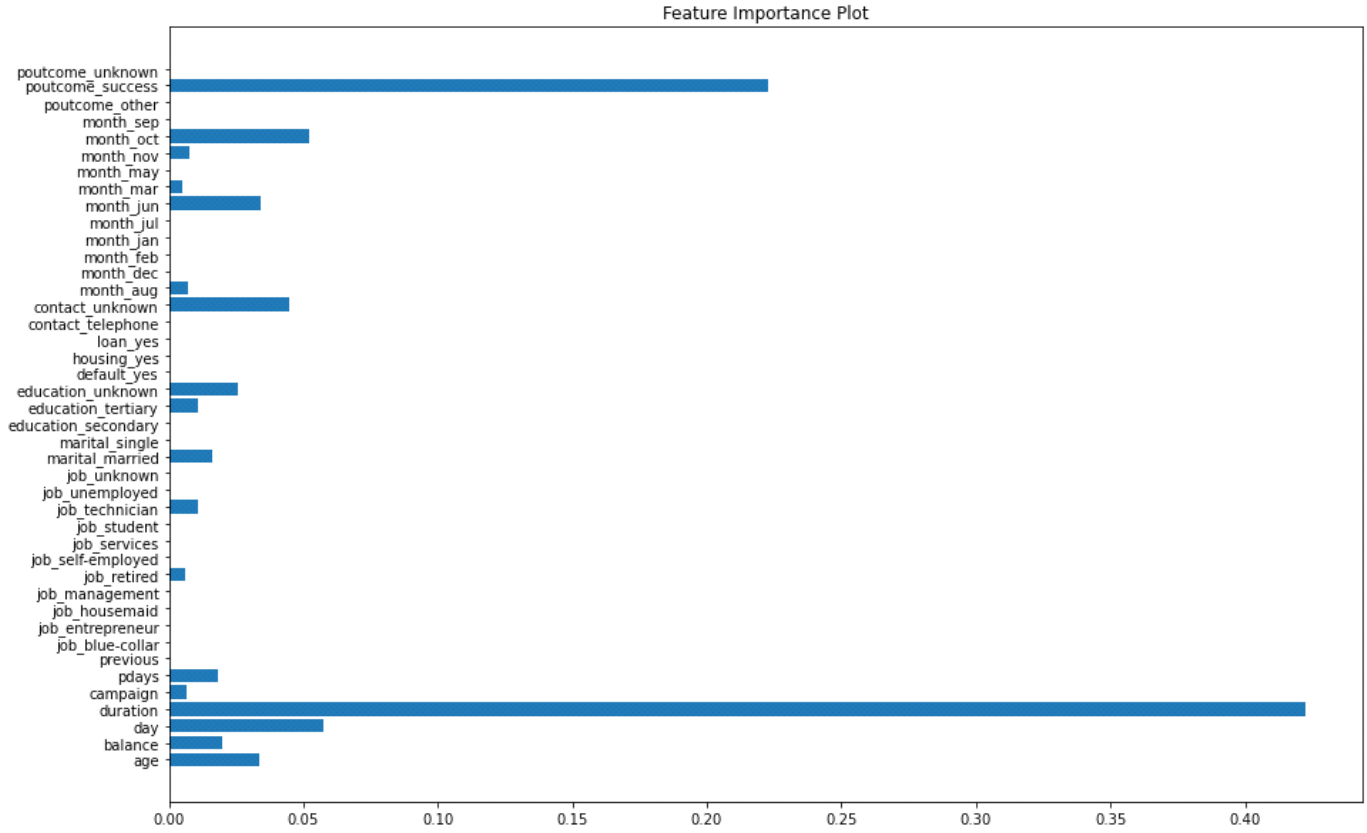
$gini = f_1(1-f_1) + f_2(1-f_2)$

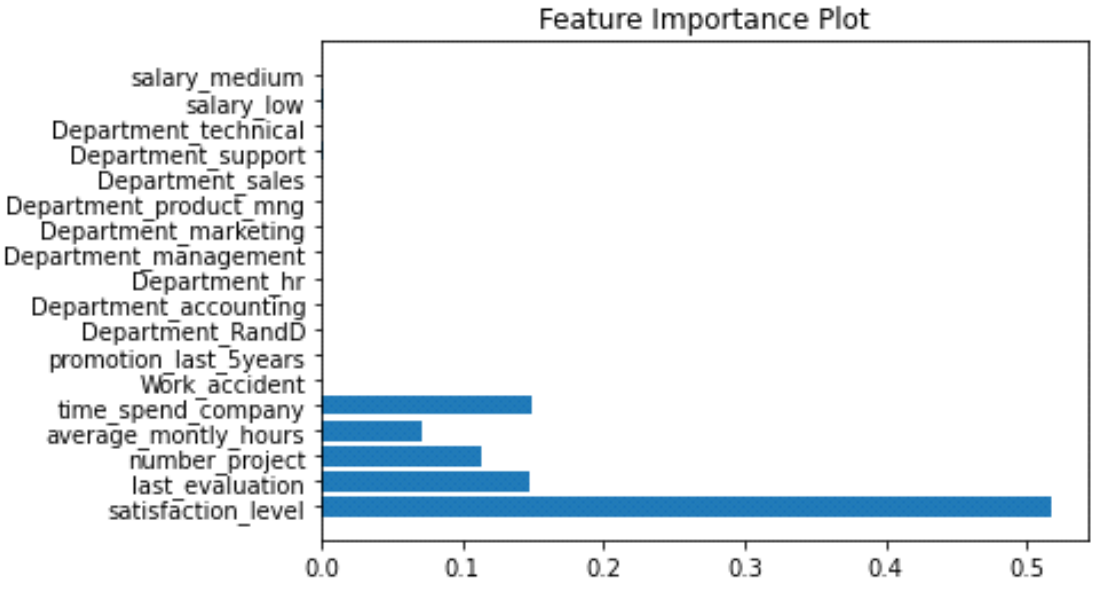
no. of observations




R9 <= 1.925
gini = 0.5
samples = 92
value = [47, 45]
T
F
R24 <= 0.095
gini = 0.328
samples = 58
value = [46, 12]
R24 <= 0.04
gini = 0.057
samples = 34
value = [1, 33]
T
F
T
F
R8 <= 0.175
gini = 0.142
samples = 39
value = [36, 3]
R21 <= 1.57
gini = 0.499
samples = 19
value = [10, 9]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 33
value = [0, 33]
T
F
T
F
gini = 0.0
samples = 1
value = [0, 1]
gini = 0.1
samples = 38
value = [36, 2]
gini = 0.355
samples = 13
value = [10, 3]
gini = 0.0
samples = 6
value = [0, 6]
count: 0 1
pred = 0

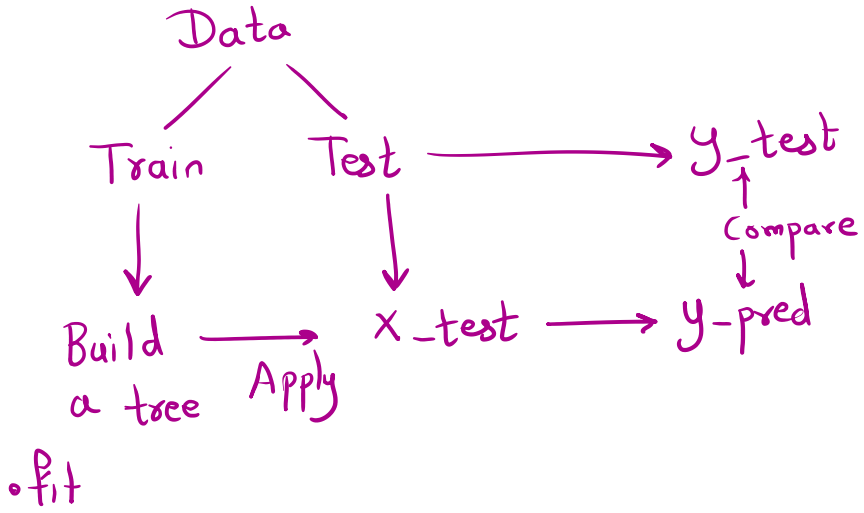

R9 <= 1.925
gini = 0.5
samples = 92
value = [47, 45]
R24 <= 0.095
gini = 0.328
samples = 58
value = [46, 12]
R24 <= 0.04
gini = 0.057
samples = 34
value = [1, 33]
R8 <= 0.175
gini = 0.142
samples = 39
value = [36, 3]
R21 <= 1.57
gini = 0.499
samples = 19
value = [10, 9]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 33
value = [0, 33]
count: 0, 1
pred=1
gini = 0.0
samples = 1
value = [0, 1]
gini = 0.1
samples = 38
value = [36, 2]
gini = 0.355
samples = 13
value = [10, 3]
gini = 0.0
samples = 6
value = [0, 6]


Feature Importance Plot
poutcome_unknown
poutcome_success
poutcome_other
month_sep
month_oct
month_nov
month_may
month_mar
month_jun
month_jul
month_jan
month_feb
month_dec
month_aug
contact_unknown
contact_telephone
loan_yes
housing_yes
default_yes
education_unknown
education_tertiary
education_secondary
marital_single
marital_married
job_unknown
job_unemployed
job_technician
job_student
job_services
job_self-employed
job_retired
job_management
job_housemaid
job_entrepreneur
job_blue-collar
previous
pdays
campaign
duration
day
balance
age
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40


Feature Importance Plot
salary_medium
salary_low
Department_technical
Department_support
Department_sales
Department_product_mng
Department_marketing
Department_management
Department_hr
Department_accounting
Department_RandD
promotion_last_5years
Work_accident
time_spend_company
average_montly_hours
number_project
last_evaluation
satisfaction_level
0.0
0.1
0.2
0.3
0.4
0.5

Trees

Data


Data
Train
Test
y_test
Compare
Build a tree
Apply
X_test
y_pred
.fit

Naïve Bayes

Thursday, June 29, 2023 2:17 PM

Indep

Talks for more than 100 min? (TT >= 100)	Gender	Response
y	male	not bought
n	male	not bought
n	female	not bought
n	female	not bought
n	male	not bought
n	male	not bought
y	male	bought
y	female	bought
n	female	bought
y	female	bought

A & B indep $\Rightarrow P(A \cap B|C) = P(A|C)\,P(B|C)$

$$P(B|TT\geq 100, m) = \frac{P(TT\geq 100 \cap M|B)\,P(B)}{P(TT\geq 100 \cap M|B)\,P(B) + P(TT\geq 100 \cap M|NB)\,P(NB)}$$

Apriori Probabilities

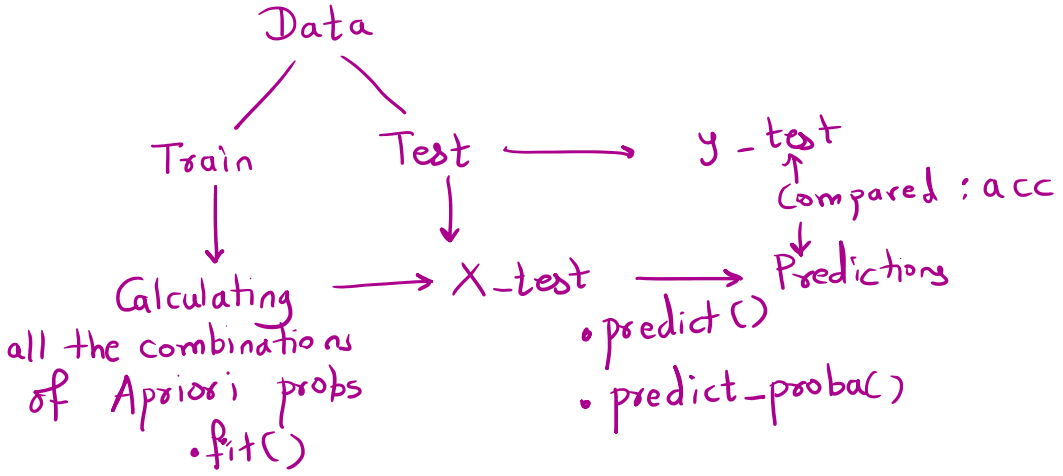
$$= \frac{P(TT\geq 100|B)\,P(M|B)\,P(B)}{P(TT\geq 100|B)\,P(M|B)\,P(B) + P(TT\geq 100|NB)\,P(M|NB)\,P(NB)}$$

$$= \frac{\frac{3}{4}\times\frac{1}{4}\times\frac{4}{10}}{\frac{3}{4}\times\frac{1}{4}\times\frac{4}{10} + \frac{1}{6}\times\frac{4}{6}\times\frac{6}{10}} = 0.529$$

Posterior Probability

$$P(B|TT\geq 100 \cap F) = \frac{P(TT\geq 100|B)\,P(F|B)\,P(B)}{P(TT\geq 100|B)\,P(F|B)\,P(B) + P(TT\geq 100|NB)\,P(F|NB)\,P(NB)}$$

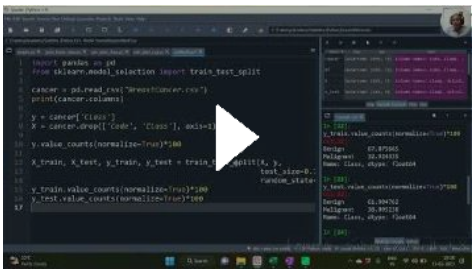
$$= \frac{\frac{3}{4}\times\frac{3}{4}\times\frac{4}{10}}{\frac{3}{4}\times\frac{3}{4}\times\frac{4}{10} + \frac{1}{6}\times\frac{2}{6}\times\frac{6}{10}} = 0.87$$



Discrete NB	Kernel NB

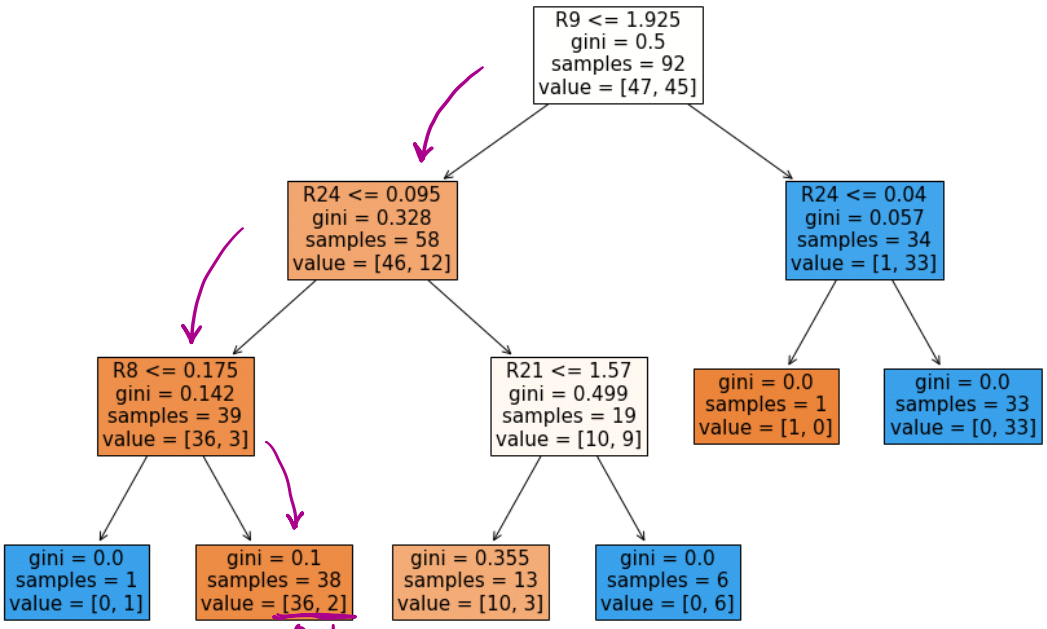
Categorical Features	Numerical Features
Apriori Probabilities are calculated based on counts	Apriori Probabilities are calculated based on the function of Normal Distribution
Bayes Formula	Bayes Formula

Stratification | Why to Stratify? | stratify=y option



Probability Estimation

Thursday, June 29, 2023 5:16 PM



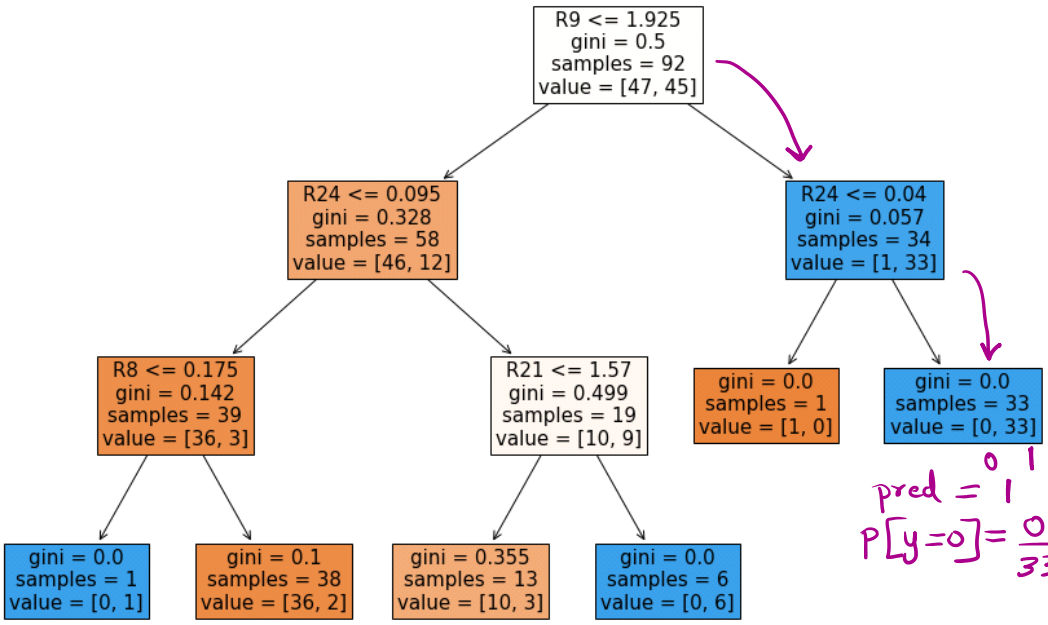
pred = 0

$P[0] = \frac{36}{38}$ $P[1] = \frac{2}{38}$

0.9473684210526315 0.05263157894736842

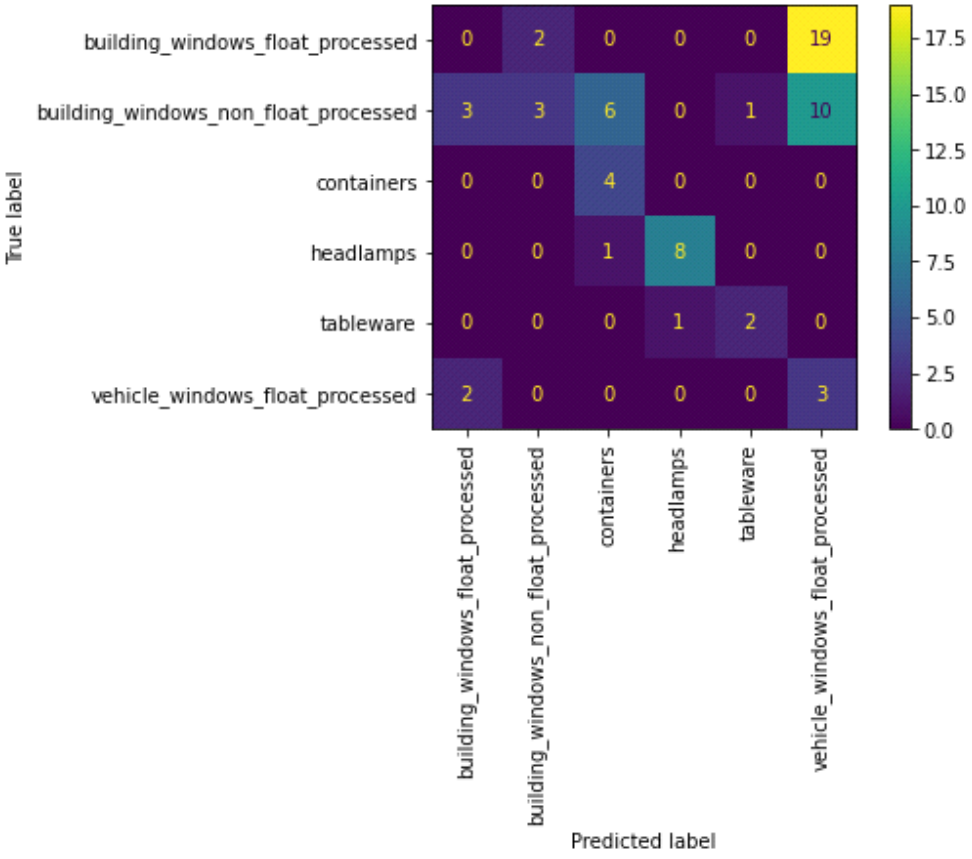
Predicted Probabilities

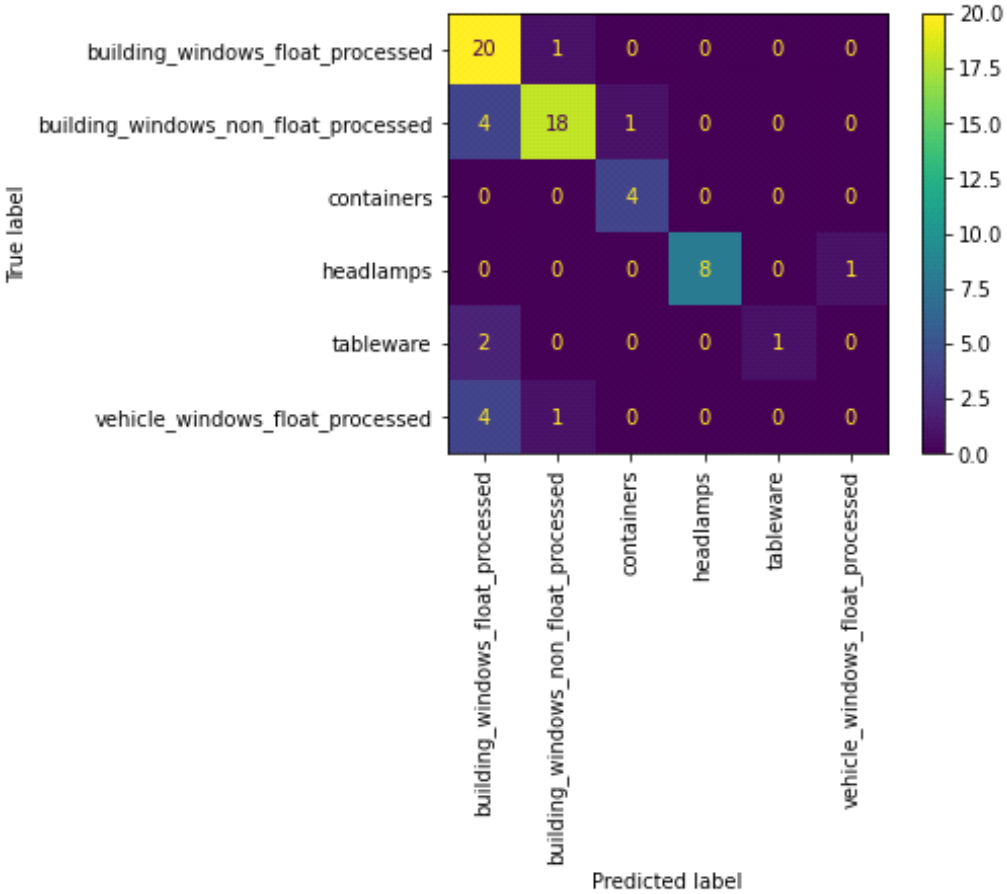
```
In [97]: dtc.predict_proba(X_test)
Out[97]: P[y=0]   P[y=1]
array([[0.94736842, 0.05263158],
       [0.94736842, 0.05263158],
       [0.94736842, 0.05263158],
       [0.94736842, 0.05263158],
       [0.        , 1.        ],
       [0.        , 1.        ],
       [0.94736842, 0.05263158],
       [0.        , 1.        ],
       [0.94736842, 0.05263158],
       [0.        , 1.        ],
       [0.        , 1.        ],
```




```
In [100]: nb.predict(X_test)
Out[100]:
array([0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0,
       0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1], dtype=int64)

In [101]: nb.predict_proba(X_test)
Out[101]: P(y=0)          P(y=1)
array([[9.99977940e-001, 2.20601420e-005],
       [9.90789108e-001, 9.21089206e-003],
       [2.32438774e-002, 9.76756123e-001],
       [3.02827814e-002, 9.69717219e-001],
       [9.99997200e-001, 2.80039369e-006],
       [6.78116258e-010, 9.99999999e-001],
       [2.08929954e-006, 9.99997911e-001],
```


building_windows_float_processed
building_windows_non_float_processed
containers
headlamps
tableware
vehicle_windows_float_processed
True label
Predicted label
0 2 0 0 0 19
3 3 6 0 1 10
0 0 4 0 0 0
0 0 1 8 0 0
0 0 0 1 2 0
2 0 0 0 0 3
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

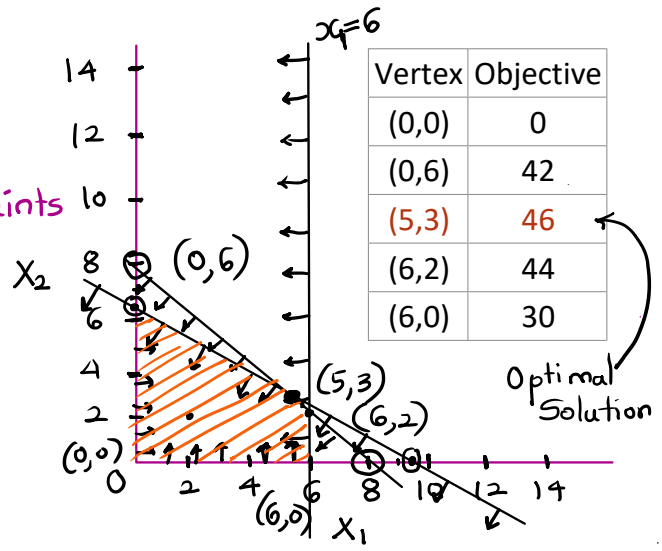

building_windows_float_processed
building_windows_non_float_processed
containers
headlamps
tableware
vehicle_windows_float_processed
True label
Predicted label
20 1 0 0 0 0
4 18 1 0 0 0
0 0 4 0 0 0
0 0 0 8 0 1
2 0 0 0 1 0
4 1 0 0 0 0
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

x_1, x_2 : Decision variables

Max $5x_1 + 7x_2$ → Objective

s.t.
$x_1 \le 6$
$2x_1 + 3x_2 \le 19$
$x_1 + x_2 \le 8$
$x_1, x_2 \ge 0$

Contraints



Vertex	Objective
(0,0)	0
(0,6)	42
(5,3)	46
(6,2)	44
(6,0)	30

Feasible Region

$2x_1 + 3x_2 = 19$

Put $x_1 = 0$ $\quad x_2 = \frac{19}{3} = 6.33$

Put $x_2 = 0$ $\quad x_1 = \frac{19}{2} = 9.5$

$x_1 + x_2 = 8$

Put $x_1 = 0 \Rightarrow x_2 = 8$

$x_2 = 0 \Rightarrow x_1 = 8$

Maximize

$Z = 20x_1 + 10x_2 + 15x_3$

Subject to:

$3x_1 + 2x_2 + 5x_3 \le 55$
$2x_1 + x_2 + x_3 \le 26$
$x_1 + x_2 + 3x_3 \le 30$
$5x_1 + 2x_2 + 4x_3 \le 57$
$x_1, x_2, x_3 \ge 0$

A company that operates 10 hours a day manufactures two products on three sequential processes. The following table summarizes the data of the problem

	Minutes per unit			
Product	Process 1	Process 2	Process 3	Unit Profit
1	10	6	8	Rs. 2/-
2	5	20	10	Rs. 3/-

Determine the optimal mix of the two products.

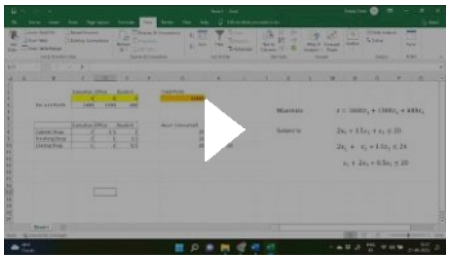
Product ? ? Product
1 2

Let x_1: units of product 1
x_2: units of product 2

Profit $= 2x_1 + 3x_2$
To be maximized

Process 1: $10x_1 + 5x_2 \leq 600$
Process 2: $6x_1 + 20x_2 \leq 600$
Process 3: $8x_1 + 10x_2 \leq 600$
$x_1, x_2 \geq 0$

Linear Optimization in Excel with Solver Add-in | LPP in Excel



A company makes three models of desks, an executive model, an office model and a student model. Each desk spends time in the cabinet shop, the finishing shop and the crating shop as shown in the table:

Type of desk	Cabinet shop (in hrs.)	Finishing shop (in hrs.)	Crating shop (in hrs.)	Profit (in Rs.)
Executive	2	1	1	150
Office	1	2	1	125
Student	1	1	0.5	50
Available hours	16	16	10	

How many of each type of model should be made to maximize profits?

Let
x_1: no. of exe. desk
x_2: no. of office desk
x_3: no. of student desk

Profit: $150x_1 + 125x_2 + 50x_3$

Cabinet: $2x_1 + x_2 + x_3 \leq 16$
Finishing: $x_1 + 2x_2 + x_3 \leq 16$
Crating: $x_1 + x_2 + 0.5x_3 \leq 10$
$x_1, x_2 \geq 0$

1. Mohan-Meakins Breweries Ltd. Has bottling plants, one located at Solan and the other located at Mohan-Nagar. Each plant produces three drinks namely, Whisky A, Beer B and Fruit Juice C. The number of bottles produced per day are as follows:

	Plant At	
Drinks	**Solan**	**Mohan-Nagar**
Whisky A	1500	1500
Beer B	3000	1000
Fruit Juice C	2000	5000

A market survey indicates that during the month of April, there will be a demand of 20,000 bottles of Whisky A, 40,000 bottles of Beer B and 44,000 bottles of Fruit Juice C. The operating costs per day for the plants at Solan and Mohan-Nagar are 600 and 400 rupees per day. For how many days each plant be run in April so as to minimise the production cost, while still meeting the market demand? Formulate the model and provide solution.

x_1 : Solan x_2: Mohan-nagar

Minimize Cost $= 600x_1 + 400x_2$; $0 \leq x_1, x_2 \leq 30$

A : $1500x_1 + 1500x_2 \geq 20000$

B ; $3000x_1 + 1000x_2 \geq 40000$

C ; $2000x_1 + 5000x_2 \geq 44000$

2. The following is the information of two parts A and B manufactured by a certain company per week.

Type of Machine	Time Required per unit		Maximum Time Available
	A	B	
Lathes	12	6	3000
Milling	4	10	2000
Grinding	2	3	900
Profit per unit(in Rs.)	40	100	

Formulate the model and provide solution for maximising the profit.

3. A Company producing three brands of Shampoos has two plants located at two places. Each plant has following production capacities per day:

Plants	Brands (Bottles per day)		
	Fresh	Blossom	Moon
I	3000	1000	2000
II	1000	1000	6000

A market survey indicates that during any particular month there will be minimum demand of 24,000 bottles of Fresh, 16,000 bottles of Blossom and 48,000 bottles of Moon. The operative costs per day of running the plants I and II are 600 monetary units and 400 monetary units respectively. How many days should the company run each plant during the month so that the production cost is minimised while meeting the market demand?

x_1 : Plant I days x_2 : Plant II days

Cost $= 600x_1 + 400x_2$

Fresh : $3000x_1 + 1000x_2 \geq 24000$

Blossom : $1000x_1 + 1000x_2 \geq 16000$

Moon : $2000x_1 + 6000x_2 \geq 48000$

Blossom : [illegible]

Moon : $2000 x_1 + 6000 x_2 \geq 48000$

$0 \leq x_1, x_2 \leq 30$

9. FlowerAurora, a florist and gifts shop, prepares three types of flower bouquets to sell on the convocation day at a local university. The bouquets are made from four types of flowers: Gypsophila, gerberas, roses, and tulips. The number of stalks of each type of flower required, along with the profit model for each bouquet, is listed in the table below:

Types of Flower	T1	T2	T3	Flower Stalks Available
Rose	3	2	5	540
Gerbera	2	4	0	320
Gypsophila	2	1	2	106
Tulip	4	3	4	273
Profit per flower bouquet	$20	$58	$39	

a. Identify the decision variables, objective function, and constraints in simple verbal statements.

4. Valencia Products makes automobile radar detectors and assembles two models: LaserStop and SpeedBuster. The firm can sell all it produces. Both models use the same electronic components. Two of these can be obtained only from a single supplier. For the next month, the supply of these is limited to 4,000 of component A and 3,500 of component B. The number of each component required for each product and the profit per unit are given in the table.

	Components Required/Unit		
	A	B	Profit/Unit
LaserStop	18	6	$124
SpeedBuster	12	8	$136

a. Identify the decision variables, objective function, and constraints in simple verbal statements.

b. Mathematically formulate a linear optimization model.

7. A paper mill produces three grades of paper: X, Y, and Z. The mill has a budget of €100,000 to invest in the production of these three different types of paper. The cost per ton and expected profit over the next two years is given in the table.

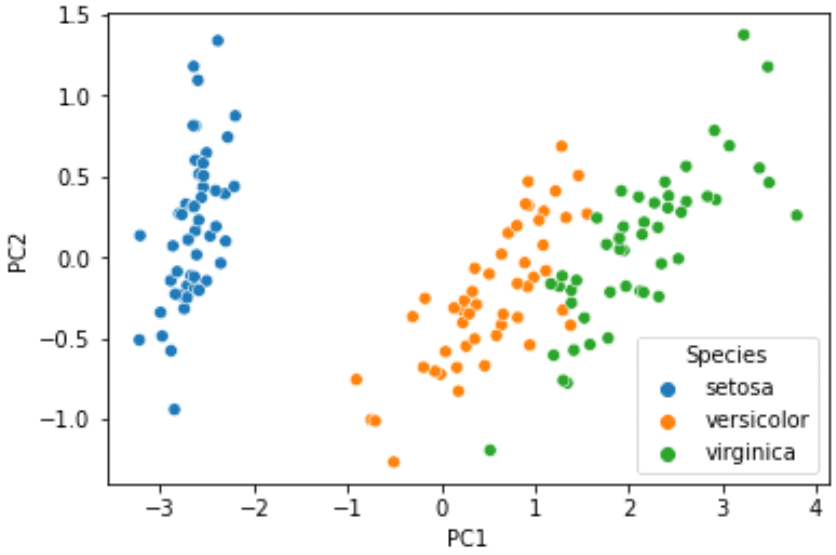
Paper Grade	X	Y	Z
Cost/ton	€10	€12	€7
Profit/ton	€30	€40	€20

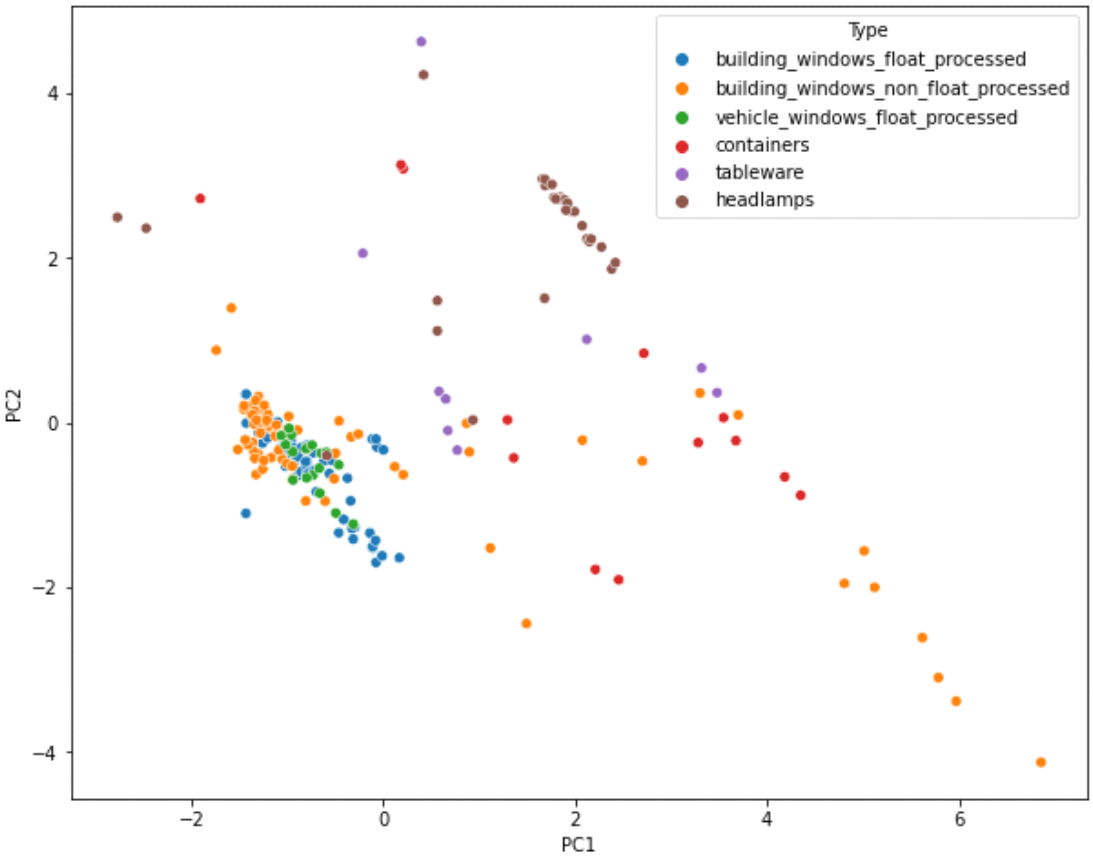
a. Identify the decision variables, objective function, and constraints in simple verbal statements.

b. Mathematically formulate a linear optimization model.

PCA

Friday, June 30, 2023 6:38 PM

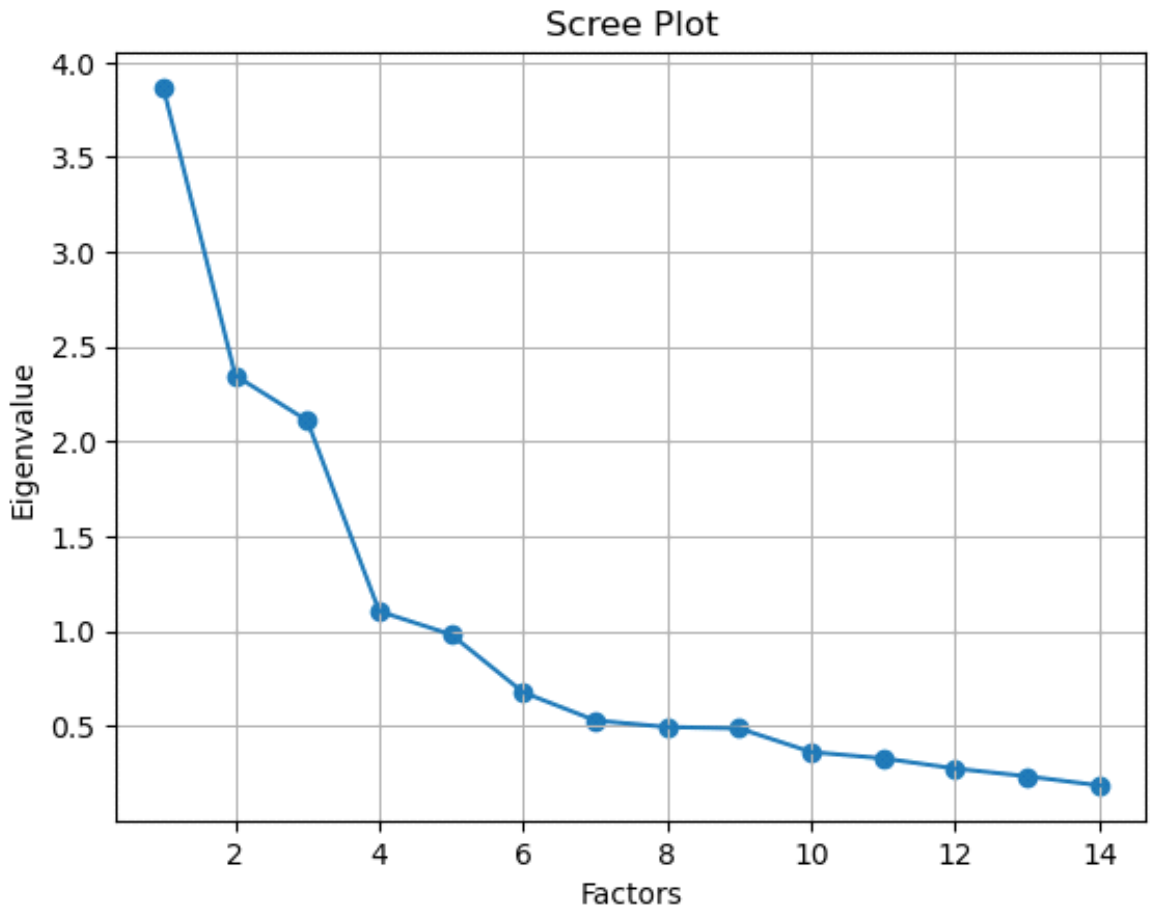




The Variances of the principal components calculated columns are nothing the eigenvalues of the variance covariance matrix.

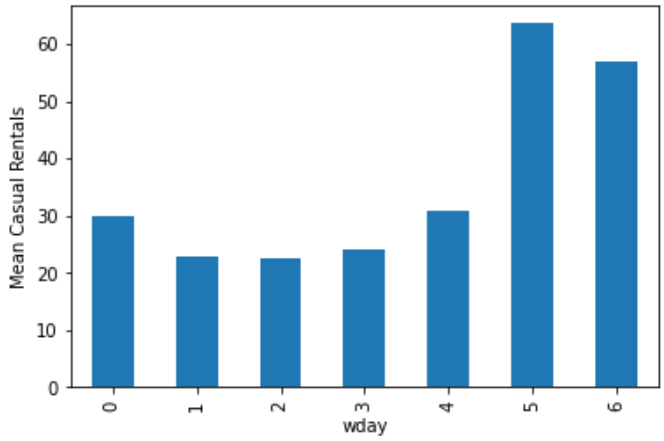
Factor Analysis

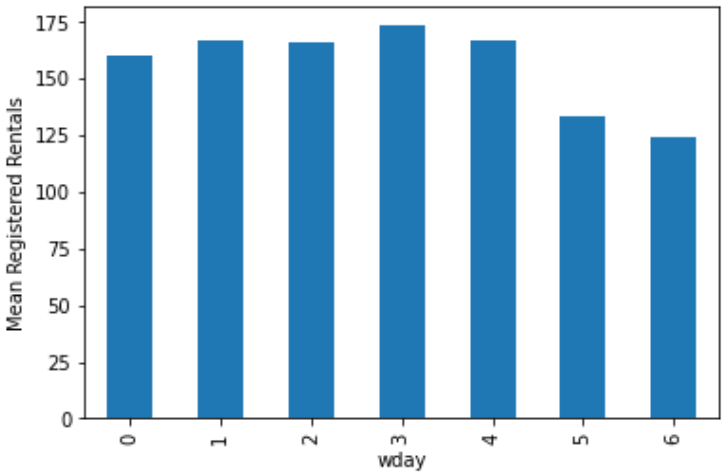
Saturday, July 1, 2023 8:31 AM



Kaggle-Bike

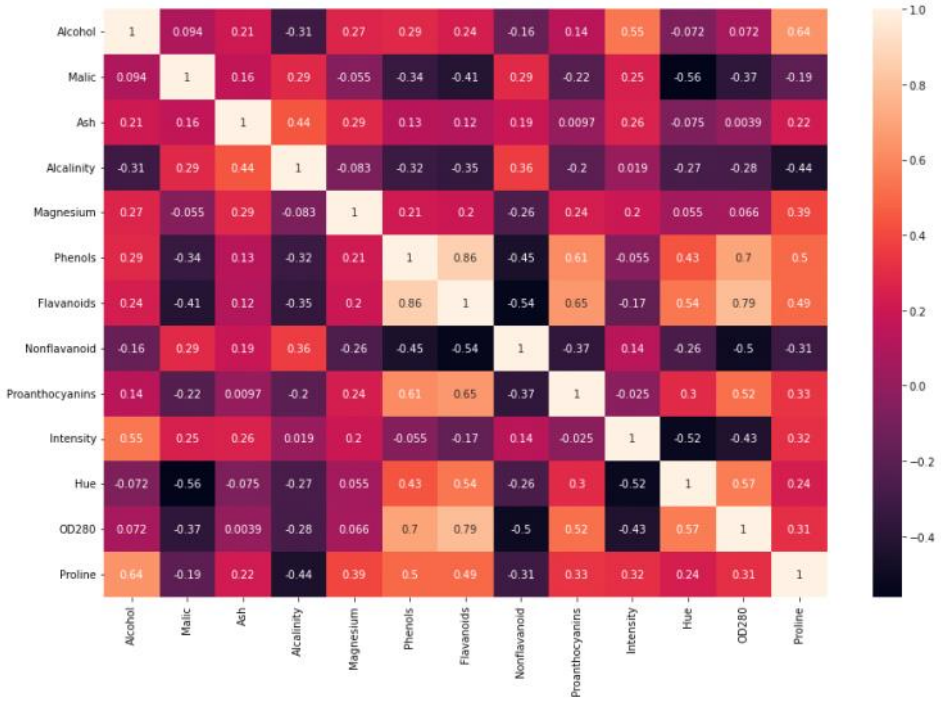
Monday, July 3, 2023 3:57 PM

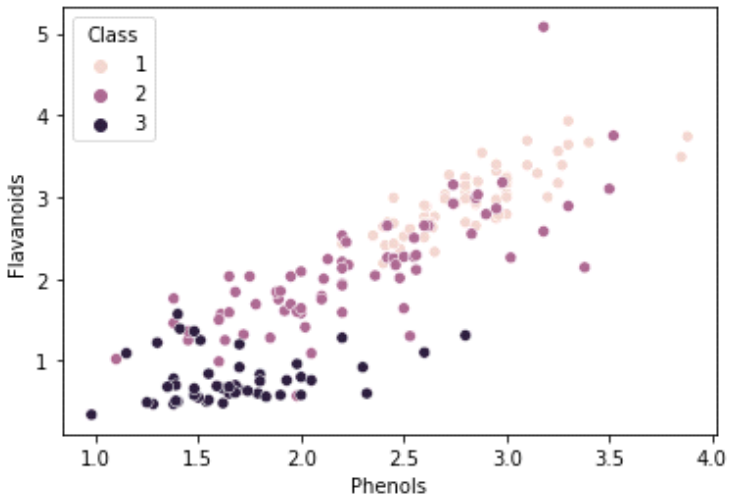


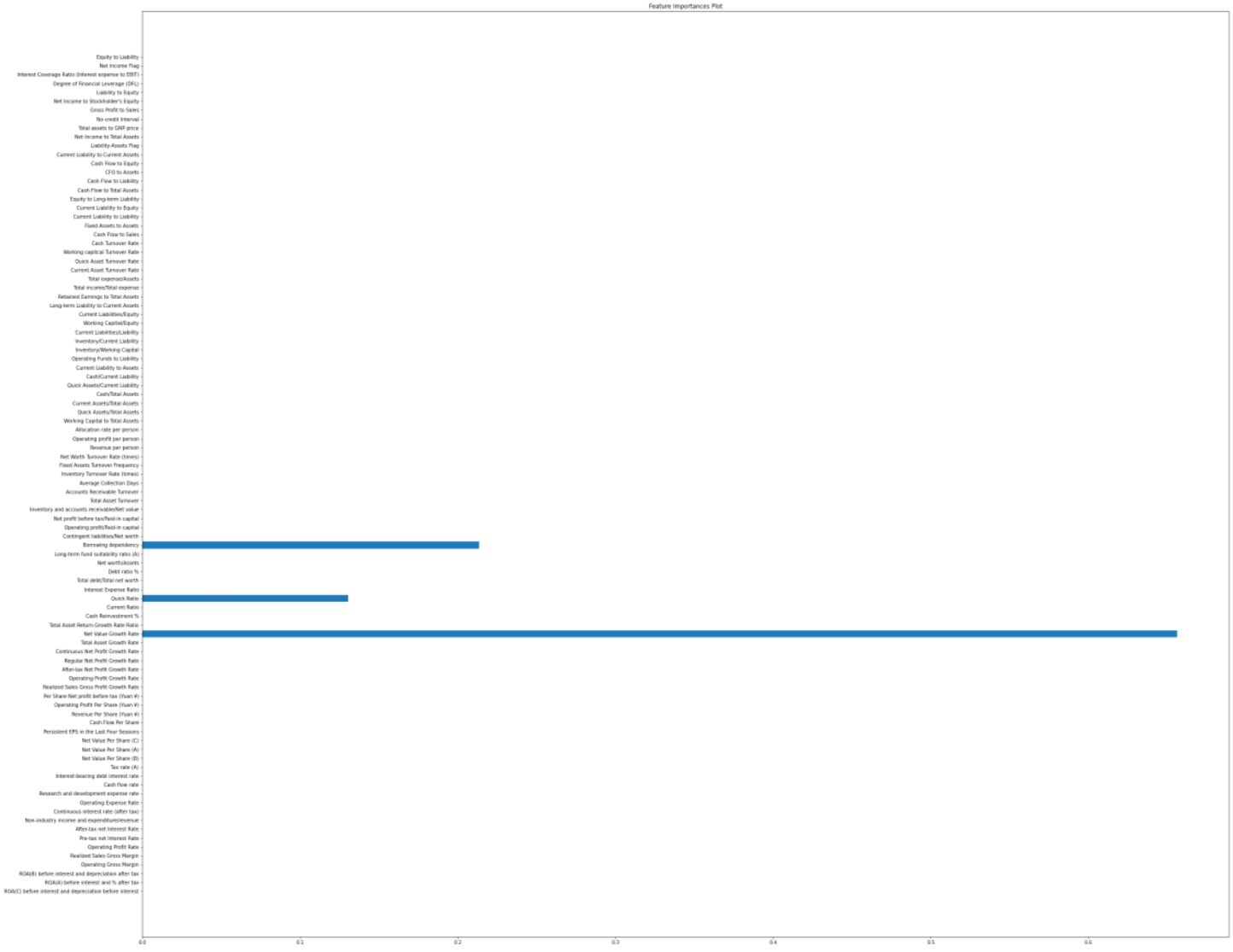


Revision

Wednesday, July 5, 2023 2:14 PM






Feature Importances Plot
Equity to Liability
Net Income Flag
Interest Coverage Ratio (Interest expense to EBIT)
Degree of Financial Leverage (DFL)
Liability to Equity
Net Income to Stockholder's Equity
Gross Profit to Sales
No-credit Interval
Total assets to GNP price
Net Income to Total Assets
Liability-Assets Flag
Current Liability to Current Assets
Cash Flow to Equity
CFO to Assets
Cash Flow to Liability
Cash Flow to Total Assets
Equity to Long-term Liability
Current Liability to Equity
Current Liability to Liability
Fixed Assets to Assets
Cash Flow to Sales
Cash Turnover Rate
Working capital Turnover Rate
Quick Asset Turnover Rate
Current Asset Turnover Rate
Total expenses/Assets
Total income/Total expense
Retained Earnings to Total Assets
Long-term Liability to Current Assets
Current Liabilities/Equity
Working Capital/Equity
Current Liabilities/Liability
Inventory/Current Liability
Inventory/Working Capital
Operating Funds to Liability
Current Liability to Assets
Cash/Current Liability
Quick Assets/Current Liability
Cash/Total Assets
Current Assets/Total Assets
Quick Assets/Total Assets
Working Capital to Total Assets
Allocation rate per person
Operating profit per person
Revenue per person
Net Worth Turnover Rate (times)
Fixed Assets Turnover Frequency
Inventory Turnover Rate (times)
Average Collection Days
Accounts Receivable Turnover
Total Asset Turnover
Inventory and accounts receivable/Net value
Net profit before tax/Paid-in capital
Operating profit/Paid-in capital
Contingent liabilities/Net worth
Borrowing dependency
Long-term fund suitability ratio (A)
Net worth/Assets
Debt ratio %
Total debt/Total net worth
Interest Expense Ratio
Quick Ratio
Current Ratio
Cash Reinvestment %
Total Asset Return Growth Rate Ratio
Net Value Growth Rate
Total Asset Growth Rate
Continuous Net Profit Growth Rate
Regular Net Profit Growth Rate
After-tax Net Profit Growth Rate
Operating Profit Growth Rate
Realized Sales Gross Profit Growth Rate
Per Share Net profit before tax (Yuan ¥)
Operating Profit Per Share (Yuan ¥)
Revenue Per Share (Yuan ¥)
Cash Flow Per Share
Persistent EPS in the Last Four Seasons
Net Value Per Share (C)
Net Value Per Share (A)
Net Value Per Share (B)
Tax rate (A)
Interest-bearing debt interest rate
Cash flow rate
Research and development expense rate
Operating Expense Rate
Continuous interest rate (after tax)
Non-industry income and expenditure/revenue
After-tax net Interest Rate
Pre-tax net Interest Rate
Operating Profit Rate
Realized Sales Gross Margin
Operating Gross Margin
ROA(B) before interest and depreciation after tax
ROA(A) before interest and % after tax
ROA(C) before interest and depreciation before interest
0.0
0.1
0.2
0.3
0.4
0.5
0.6