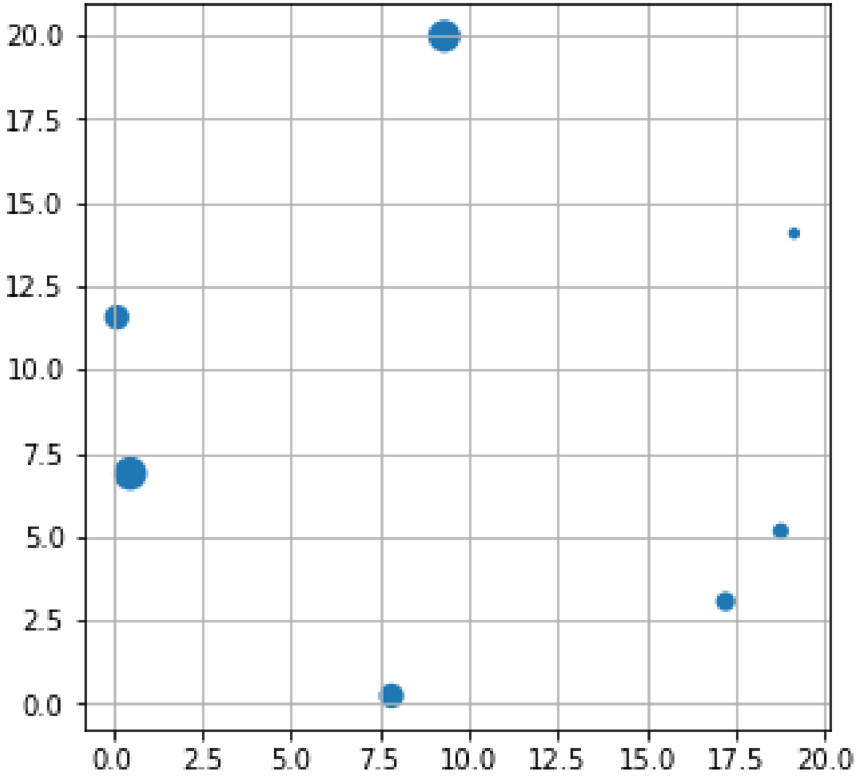
Exercice 1:

Suppose that we have data from 10 samples scattered as the figure above shows.



- How many variables have been used for visual encoding?
- Calculate count, mean, min, max, var, and std from the data.
- Plot the mean, 1 std, 2 std, and 3 std.
- Is there any correlation between the variables? Estimate it and make some conclusions.
- Give an example of where we can find such data behavior in the real world.
- Write the code that generates this figure (including modeling samples).
- Write an additional code that brings some noise up to the data.
- What function(s) do you think best model this data?
- Alter the code using this function if it is necessary.
- Plot the data as a histogram/pie, after mapping the data into three equal bins based on y-axis.
- Generate a boxplot for this data and check if the sample [17,0.5] is an outlier.

Exercise 2:

Consider the following data table,

State	Population	Murder andNonnegligentmanslaughter	Robbery	Aggravated^{2}assault	Burglary	Motorvehicletheft	Arson
Alabama	248431	20.13	177.11	485.85	1216.84	506.78	22.94
Alaska	296188	9.12	262.67	799.49	748.17	1047.98	20.93
Arizona	517898	4.82	124.57	242.99	497.731	266.62	14.38
California	668216	6.63	Null	328.02	497.88	556.25	28.00
Colorado	515864	7.543	153.58	345.25	532.416	611.02	20.35
District of Columbia	693972	16.72	338.77	529.42	260.53	366.73	Null

1. Calculate covariance/correlation over columns and over rows. What do you deduce?
2. Based on that, do you suggest removing any of the variables (crime types)? If yes, what are the reasons?
3. Perform a bivariate analysis (using graphical representation) of (y='MurderandNonnegligentmanslaughter', x = 'Aggravatedassault')
4. In the table above, there are two cells with Null values, what do suggest to resolve the missing? Justify!
5. Is there a possibility of predicting the rate of murder given the rate of Aggravated assault? If yes, what do propose as a model?
6. Do you think that increasing the population certainly increases the crime rate? Prove your answer.
7. Given the following univariate crime rate sub-data:

State	Alabama	Alaska	Arizona	California	Colorado	District of Columbia	Florida	Georgia	Hawaii	Idaho	Illinois	Indiana	Iowa	Kansas	Kentucky	Louisiana	Maryland	Massachusetts	Michigan	Minnesota	Missouri	Nebraska	Nevada	New Jersey
Total	6217.0	6640.0	3509.6	3626.5	3918.9	5104.96	4132.6	4814.4	3053.77	2741.97	4381.65	4673.8	5102.7	6583.5	4457.86	6034.705	6997.68	2758.22	6726.82	5015.52	7248.93	4527.4	3191.2075	2823.7

a) Calculate the mean and standard deviation of the total crime rate.

b) Perform density estimation while assuming that the data follows a gaussian pdf ($f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$)

c) Compare the resulting density plot to the corresponding data histogram (bins = 5)

d) A citizen randomly chooses a state to live within, what is the probability that this state has a low, medium and high rate of crime. (use histogram then confirm with p-value)

Exercise 3:

A set of 20 apples has been collected and the size/weight of each has been measured and listed in the table below.

Size(dm)	Weight(dg)	Class
1.75	4.38	Red
1.98	3.89	Red
1.68	3.47	Red
1.65	3.06	Red
1.68	4.18	Red
2.11	2.04	Yellow
2.37	2.18	Yellow
1.65	0.80	Yellow
1.79	6.30	Yellow
1.44	4.03	Red
1.74	3.83	Red
1.96	2.61	Yellow
1.96	4.18	Red
1.81	4.30	Red
1.50	2.72	Yellow
1.24	4.22	Yellow
1.13	4.39	Yellow
1.33	5.09	Yellow
1.98	3.00	Red
1.55	4.63	Yellow

1. plot the data scatter.
2. PCA aims at finding directions (vectors) that captures maximum amount of variations. Using the power of visualization, estimate the component of the two eigenvectors and their corresponding eigenvalues. What is the expected accuracy if we used only the first component given that the accuracy yielded from the original dataset is 98%.
3. Perform Simple, stratified, and systematic random 10-subset selection. Calculate then compare some descriptive statistics quantities. Plot the expected estimated density from each subset. What do you deduce?

4. Using descriptive statistics, determine what size of the subset should be to considered representative.

Exercise 4: Given the following data table, where variables are important as the same.

time	profit	loss	saving
2.3	199	4	896
1.1	12	2	895
2.4	251	4	898
2.2	158	4	899
2.4	251	32	928
1	10	2	899
2.9	794	5	903
2	100	4	903
1	10	2	902
2.3	199	4	905
1.5	31	3	905
1.7	50	3	906
2.7	501	5	909
1.7	50	3	908
2.7	501	5	911
2.2	158	4	911
2.4	251	4	912
2.6	398	5	914
1.1	12	2	912
2.6	398	5	916

- Using EDA, find out which variable needs to be normalized and what type of normalization should be used.
- Perform the adequate normalization type for each of the variables (if needed).
- To perform a linear regression of a variable, we need to calculate the corresponding bias and weight. Given that:
    - $b = (x - \mu_x)(y-\mu_y) / (x - \mu_x)^2$

- w = μ_y – $(b * \mu_{x)}$

1. Perform linear regression for the variable '**profit'** before and after normalization. Compare the sum of residual errors for both cases.

2. Explain the output (*profit*) in terms of the input (*time*) in each case.