Assignment 1: Basic data analysis and simulating probability distributions

Name: Min Wu.

Introduction:

In this assignment, some basic analysis of numerical data is studied using statistical libraries in Python such as NumPy. Three scenarios are studied which represents the most common models used in statistics and data science. For example, Bernoulli distribution, Normal distribution and Geometric distribution. In this assignment, synthetic data were generated by simulating the models and were plotted using the histogram, scatter methods.

Part 1: Real estate prices

Load the “houses.csv” file using

-obj=pd.read_csv('houses.csv',sep=',',header=None)

Print the properties (mean, median, standard deviation, minimum, and maximum) of the second column

```
Price
Max: 48465717.00, Min: 150.00, Mean: 174386.75, Median: 129000.00, Standard Deviation: 351461.64
```

Plot a histogram that shows the distribution of the prices.

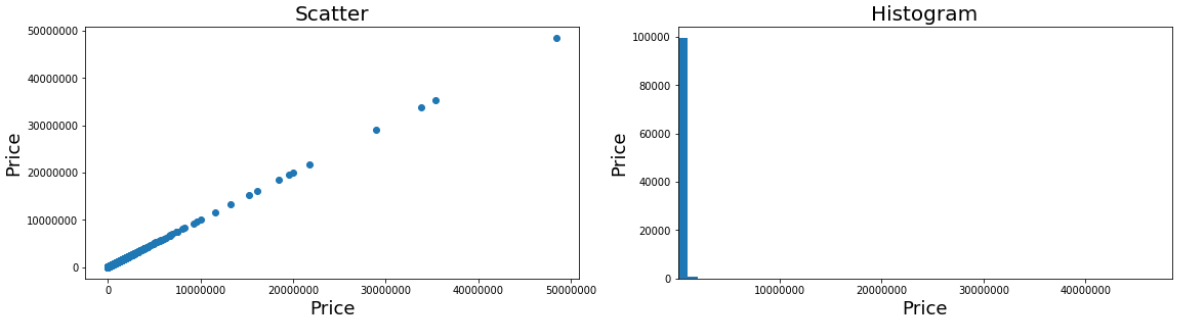


Fig. The distrubtion of the prices. Left panel: scatter plot by price vs price; right panel: histogram plot.

The prices are mainly located in the range of 1.5e2-1e7, but the total range of the data is 1.5e2-4.8e7. Most prices are concentrated in one certain bin. The Histogram plot is thus meaningless because of the large range distribution with a high concentration in a certain place.

What can you do to make it more informative?

The prices are mainly localized in the range of 0-1e7 seeing the scatter plot. The total range can be divided in two part with threshold price = 1e7

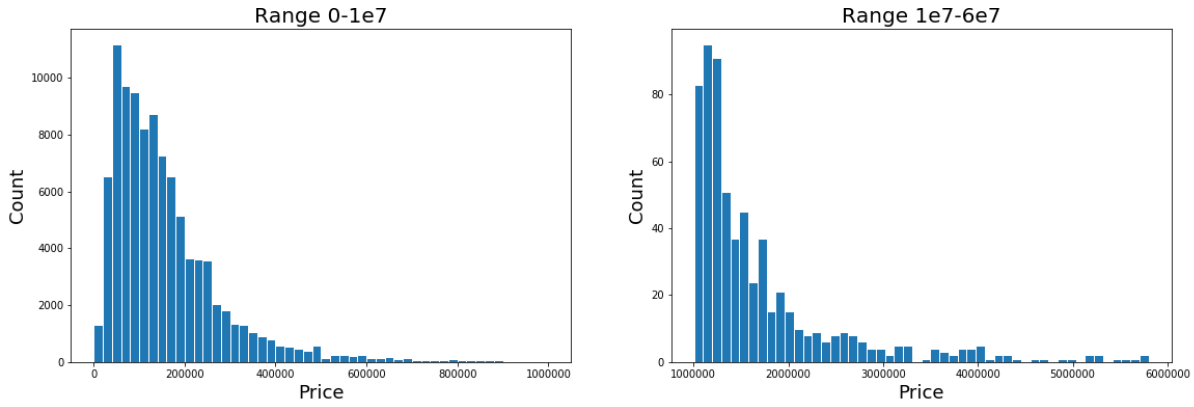


Fig. the histogram plot of price distribution in the range of 0-1e7 and 1e7-6e7

Is real estate more expensive in London?

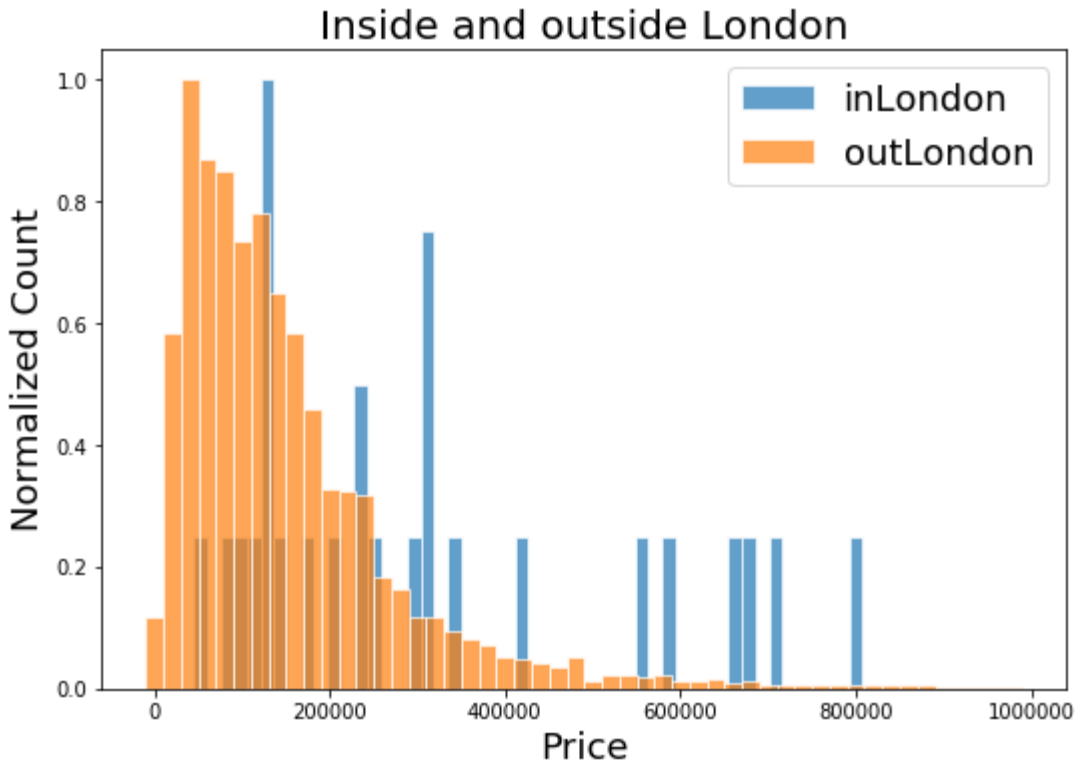


Fig. The price distributions of estate inside and outside London using histogram plot.

Because the counts for London is much smaller than the other cities. The normalized counts are thus employed and the prices range is 0-1e7. The price outside of London is mainly localized in the range 5e4-2e6, but the prices inside London largely spread out which may due to different locations.

Optional task. Make a plot that shows the average price per year.

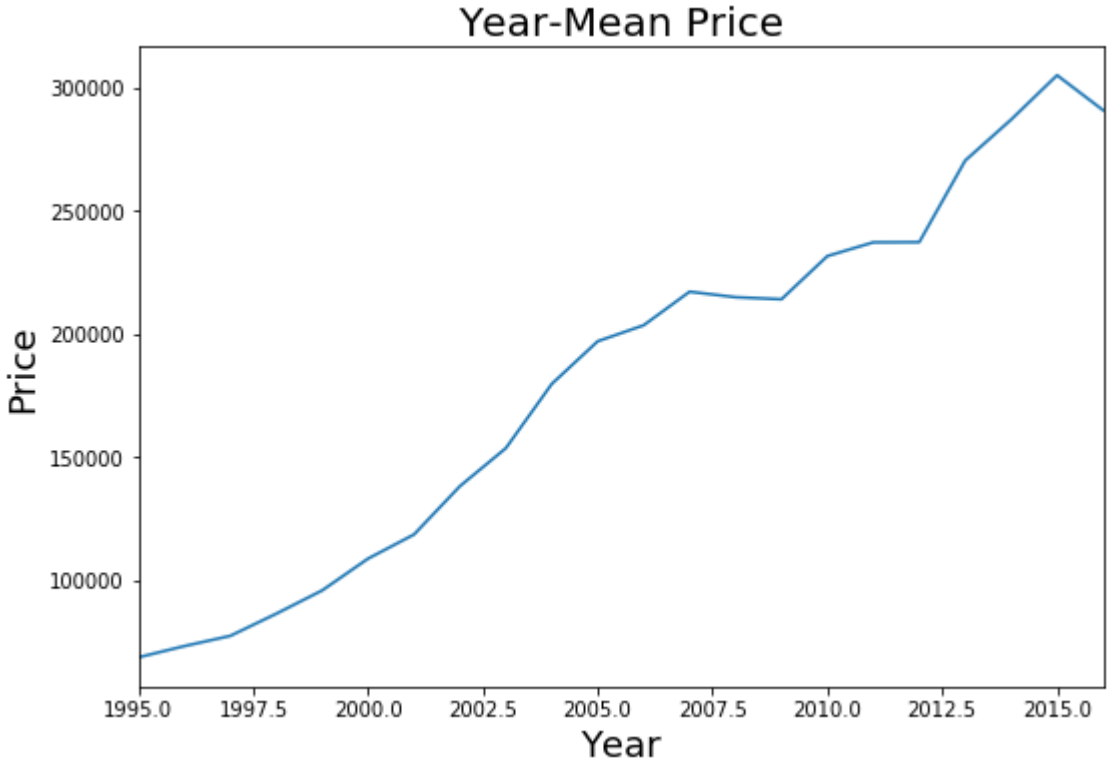
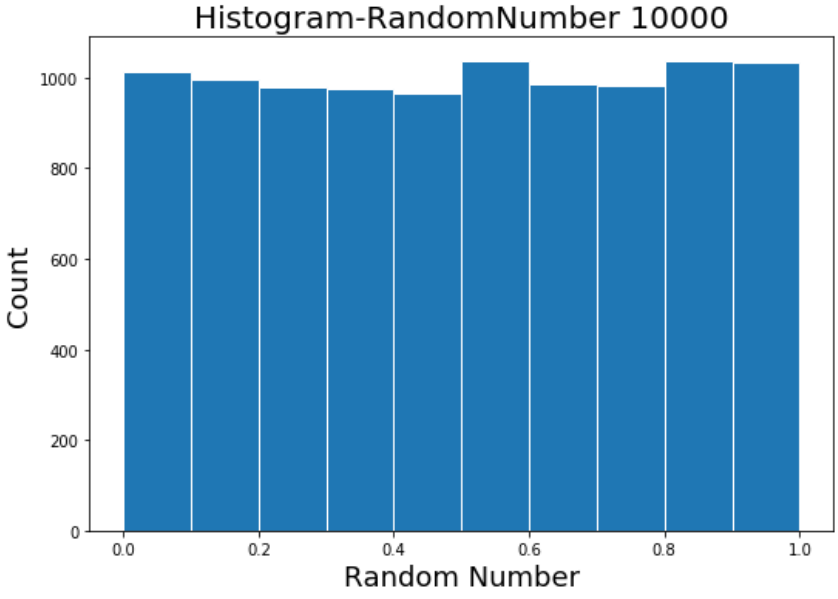


Fig. Average price vs year.

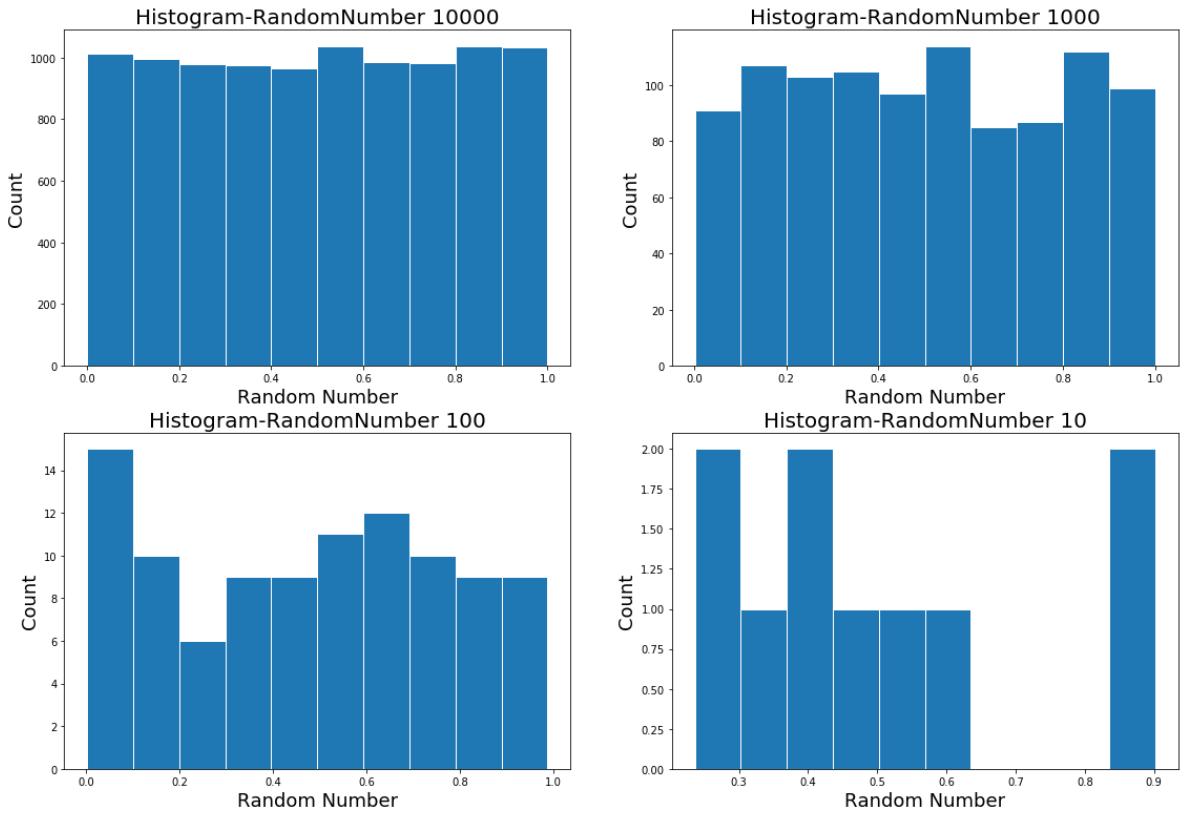
Part 2: Generating random numbers (quick detour).

Generate a set of random numbers using the function `rand` and plot its histogram. What is the shape of this histogram and why?



The random data spread out in the histogram. In each bin, the number of random numbers is almost same.

Investigate how the shape of the histogram is affected by the number of random numbers you have generated.



The different numbers of random numbers are plotted. The histogram for 1000 and 100 random numbers are in a similar shape to the 10000 random numbers. However, when the number of random numbers is decreased to 10, the random number in some bins are missing.

Instead of using `rand` (which corresponds to a *uniform* distribution), generate numbers using some other distribution and plot a histogram. What is the shape now? For instance, with `normal`, the normal (or Gaussian) distribution, you should get the familiar bell shape.

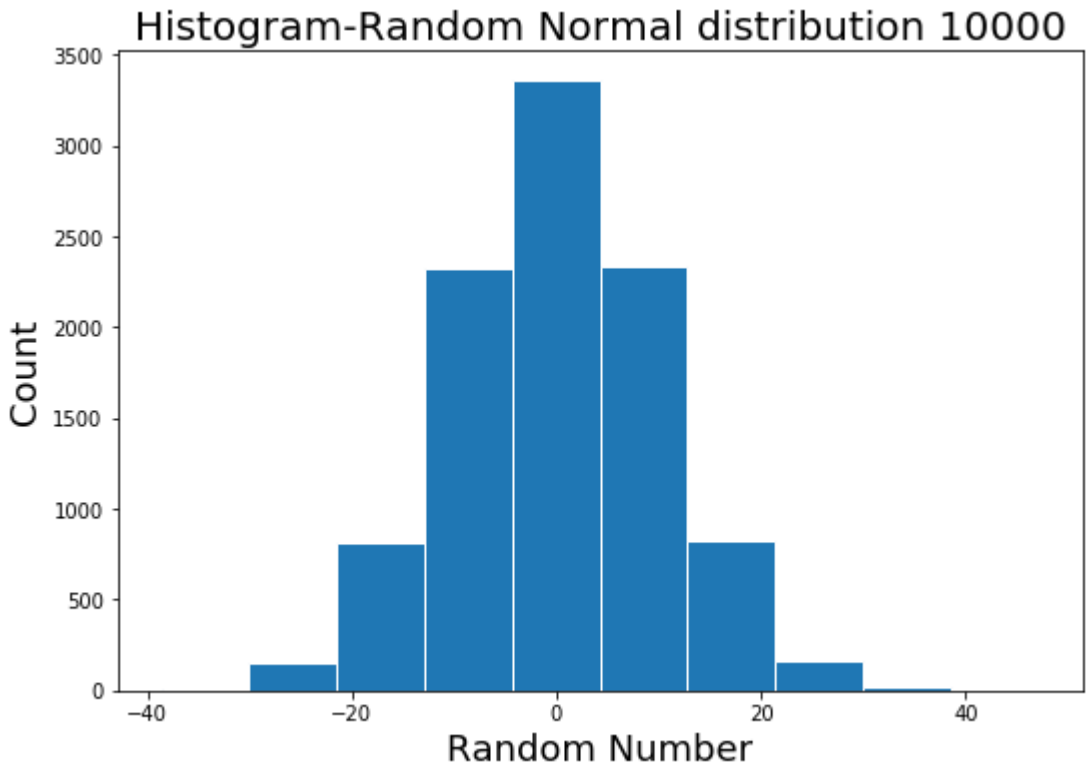


Fig. Histogram plot of 10000 random number with normal distribution.

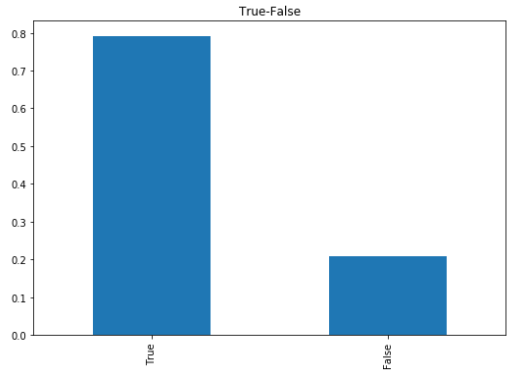
Part 3: Simulating probabilistic models

(a) Modeling a student at an exam

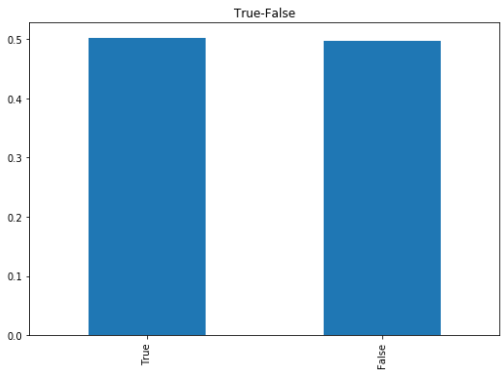
Answering a single question

Write a Python function that simulates that the student answers a single question either correctly or incorrectly with the probability of a correct answer is an input. The function simulates a random variable with a *Bernoulli* distribution.Run this function with two different probabilities (0.8 and 0.5), results are plotted. When the probability is 0.8, the results show 80% correct and 20% incorrect. When the probability is 0.5, the results show 50% correct and 50% incorrect.

p=0.8



p=0.5



How many correctly answered questions?

Write another function simulates an scenario where the student answers a fixed set of questions which are equally difficult.

Input: the number of questions and the probability of a correct answer.

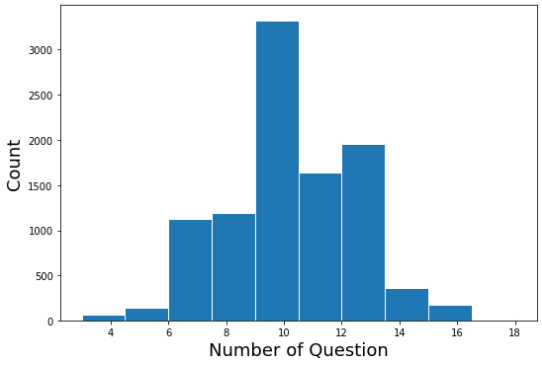
Return. the number of correctly answered questions.

Run the function 10,000 times with two sets of parameters as following:

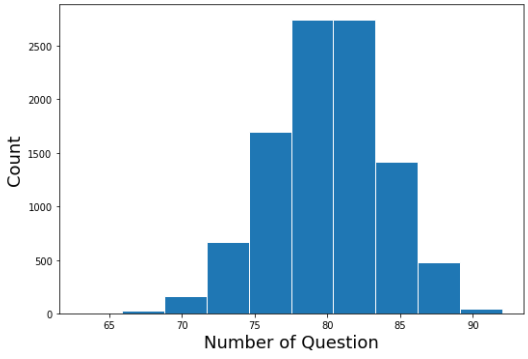
(20 questions, 0.5 probability)

(100 questions, 0.8 probability)

N=20,p=0.5



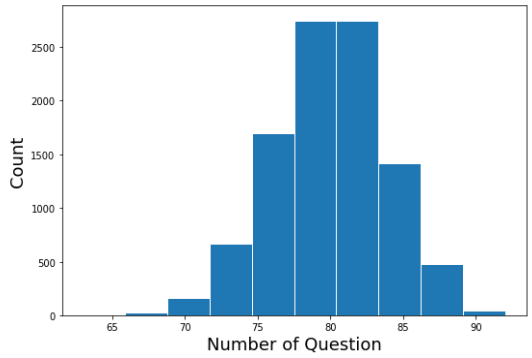
N=100,p=0.8



When the probability of a correct answer is 0.5 with 20 questions, the highest number of correct answered questions are at 10. When the probability of a correct answer is 0.8 with 100 questions, the highest number of correct answered questions are at 80. The results are reasonable.

Investigating the distribution

Run the simulations 10,000 with the probability of a correct answer is 0.8 and 20 questions.



(b) The persistent student

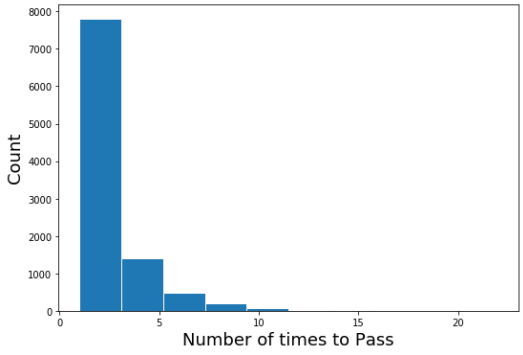
Write a function simulates a scenario where a student takes an exam repeatedly, until passing.

Input: probability of passing a single exam

Return: the number of attempts the student needed before passing

Investigating the distribution

Simulate this model 10,000 times with the probability 0.4 to pass a single exam. The result is plotted using a histogram.



Note: This type of scenario corresponds to the *geometric* distribution.

(c) An unusual village

Write a Python function generates the height and weight of a random inhabitant of Normlösa. Use the following process:

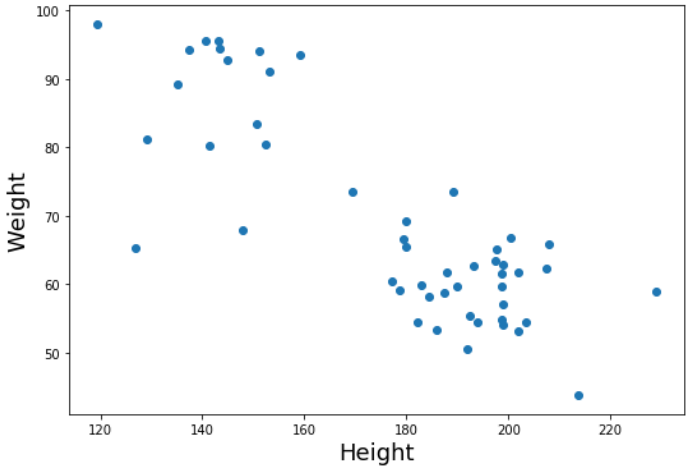
- first, randomly select the gender of the villager; the proportion of males in this village is about 40%.

- then draw random numbers from a Gaussian distribution (normal distribution) for the height and weight of the person;

    for males, the mean height is 140 and the height standard deviation is 15; the mean weight is 90 and the weight standard deviation is 10;

    for females, the mean height is 195 and the height standard deviation is 10; the mean weight is 60 and the weight standard deviation is 5.

Generate a dataset consisting of height–weight pairs for 50 Normlösa inhabitants. Make a scatterplot of the height–weight data.



Let's pretend for a moment that you have been given the data points (the list of height–weight pairs) but you have no information about how they were generated. Could you think of a way to reconstruct the parameters you used in the code previously? For example, that the proportion of males is 40%, that the mean weight of a female is 60 kilograms, etc.

First, plot the data using histogram which could show the distributions (for example gender, weight and height), Using different sets of parameters to fit the data and find out the best fitting results which could represent the data.

Discussion:

In the real word, it is more complicated than the scenarios in this assignment. The simplified assumptions are applied in the scenarios, for example, in the first scenarios, all questions are equally difficult.