# Lab 2

Umuthan Ercan
February 28, 2023

**Abstract**

This lab report provides information on how we fulfilled the requirements of part A from Lab 2, which was mainly focused on the implementation of some distribution operations, demonstration of the central limit theorem, and finally exploration of a dataset and execution of some data analysis operations as requested. We used related Python packages and libraries, tried to keep our code clean and readable, and we commented on our code to make it more understandable.

## Introduction

Part A consisted of 4 parts that can be summarized as continuous distribution, normal distribution, implementation of the central limit theorem, and exploration of the "Melbourne Real-Estate Dataset" [2]. We will continue the explanation of our work separately since all four tasks were independent of each other and they require to be described independently.

- Part 1 was based on continuous distribution, and required us to randomly generate 50 values, list and format them, perform mean and standard deviation calculations to have an understanding of the data, describe what we understood, box plot the data to visualize and improve our understanding, detect the outliers if any, compare the actual data with expected, and answer some data related questions. Further details will be provided in the following sections.
- Part 2 was based on normal distribution, and required us to use a data set given for 7 lap times of Terry Vogel, list and format the provided values, randomly select 6 values from each stratum, record the lap times, construct a histogram for the collected values, perform mean and standard deviation calculations to have an understanding of the data, describe what we understood, analyze the distribution, fill some missing values in given sentences and answer a data related question. Further details will be provided in the following sections. • Part 3 was based on the central limit theorem, and required us to perform operations on a data set about the same amount of cookies baked according to different cookie recipes and how long they last, list the provided values, perform mean and standard deviation calculations to have an understanding of the data, randomly select 4 samples containing 5 elements first and perform mean and standard deviation calculations, randomly select 4 samples containing 10 elements and perform mean and standard deviation calculations, construct a histogram for the original population, and reconstruct the histograms for 5 and 10 elements, compare the constructed histograms, describe the data and answer data related questions. Further details will be provided in the following sections.
- Part 4 was based on data analysis, and required us to work on a data set provided called "Melbourne RealEstate Dataset" [2], analyze and observe the correlations in the data set by performing data analysis tasks, and arrive deductions based on the houses, room numbers, prices, regions, sellers and other related labels. Further details will be provided in the following sections.

## Methodology

We used Anaconda as a main suite, and Jupyter Notebook during code implementation. We also used the random, numpy, pandas, and matplotlib.pyplot libraries.

1. In part A.1, we generated 50 random numbers between 0 and 1 using the random module, and joined them in an empty list. After that, we separated the 50 values into 5 lists containing 10 numbers just to fill the required table. We calculated some statistics on the random values including mean, standard deviation, and median calculations, we separated the data into 4 quartiles and calculated values for the first and third quartiles using numpy. Later, we generated histograms of 8 and 5 bins using matplotlib.pyplot.

   We repeated mean, standard deviation, and median calculations for theoretical distribution and we included a function to catch, separately list, and show the outliers if any exist.

   Finally, we compared the data and statistics obtained from the random module and the theoretical distribution.
2. In part A.2, we created a list containing the values of the provided Lab Times data set including the values column by column, so for each lap we would have the values of the 20 races in order before adding the ones from the next lap. Then with nested loops we sample random lap times from each lap (so times from random races for each lap).

   After, we generate a histogram, calculate some statistic values, analyze the distribution and describe the data, and we repeat everything for a theoretical normal distribution comparing the results obtained afterwards.
3. In part A.3, we take a look at the central limit theorem. We again hand-write the given data, and in this case we generate smaller samples of the data given in a similar manner as we did in task 2, randomizing which values we

take from the bigger population. We calculate some statistic values and try what happens changing the sampling size, checking if our distribution tends to look like a normal one as the CTL postulates.

4. In part A.4, we are given a dataset and we are asked to calculate some statistics other than normal means or standard deviations. The dataset is the [2], and here we decided to use a pandas dataframe to store the data. With this data, we have played with other statistic operations and used some more advanced data science techniques such as correlation matrices to check how different attributes' evolutions in a dataset are related. For one particular case in which we wanted to relate the number of houses in a suburb to the average prices of a house in that particular suburb, I declare two dictionaries that store both the number of houses in each suburb and the average prices of a house in each suburb (being the dictionary of the prices depending on the other dictionary, as we need the number of houses in each suburb to calculate the average prices of a house in there).

   After those dictionaries are created, I create a dataframe with the data contained in them so that I can calculate the correlation between both the number of houses in a suburb and the average price of a house in the suburb. We also plot that relation.

# 1 Results

1. In part A.1, we can see that the results for the statistic values, as they depend on a random generation of numbers will vary in each execution, but we can put the latest ones that we obtained.

   The random generated table (random numbers from 0 to 1) was:

```
[0.667, 0.475, 0.066, 0.24, 0.956, 0.007, 0.539, 0.223, 0.651, 0.769]
[0.165, 0.655, 0.768, 0.54, 0.361, 0.678, 0.195, 0.54, 0.245, 0.003]
[0.156, 0.962, 0.09, 0.053, 0.184, 0.763, 0.944, 0.711, 0.796, 0.609]
[0.797, 0.352, 0.722, 0.06, 0.452, 0.456, 0.955, 0.046, 0.247, 0.245]
[0.49, 0.882, 0.249, 0.016, 0.394, 0.825, 0.351, 0.696, 0.384, 0.384]
```

Figure 1: Randomly distributed numbers from 0 to 1.

Moving on to the statistics calculated, the values are:

(a) Mean: 0.4602800000000001.

(b) Standard deviation: 0.2916940890727818.

(c) First quartile is: 0.003.

(d) Third quartile is: 0.454.

(e) The median is: 0.454.

We can see that changing the number of bins in the histogram affects the shape of it and seeing a small number of bins might make us think that we will find certain values that are within that bin, but as we are measuring density of data if in a bin there are a lot of instances in one end of the bin, visually we might think that the instances are spread evenly throughout the bin, but if we increase the number of bins we might be surprised with having nothing where before we had an area covered by a bin.

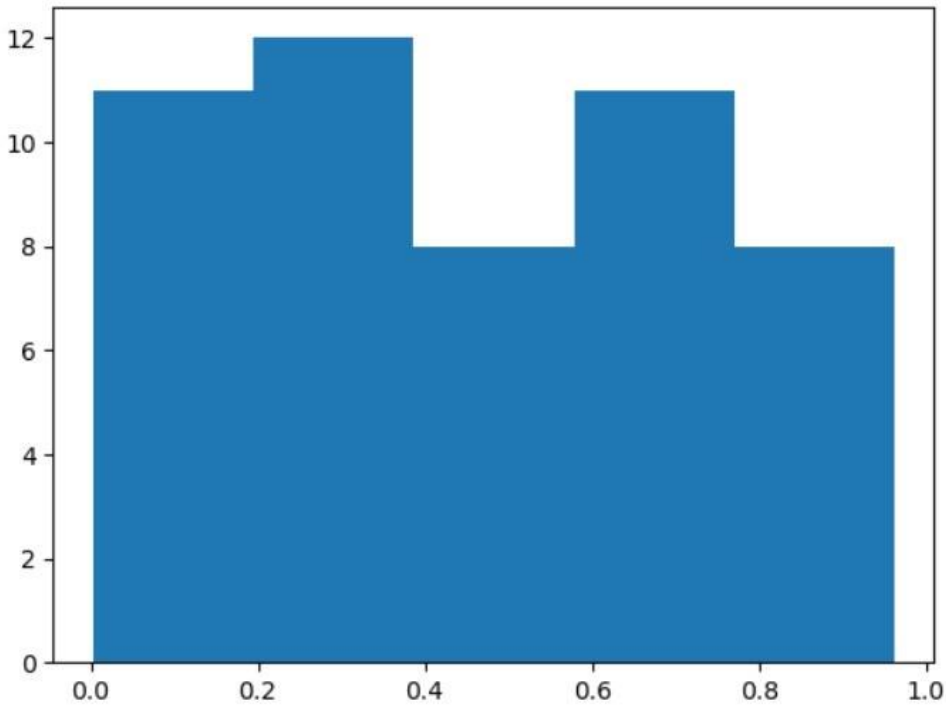


Figure 2: Randomly distributed numbers from 0 to 1, 5 bars histogram.

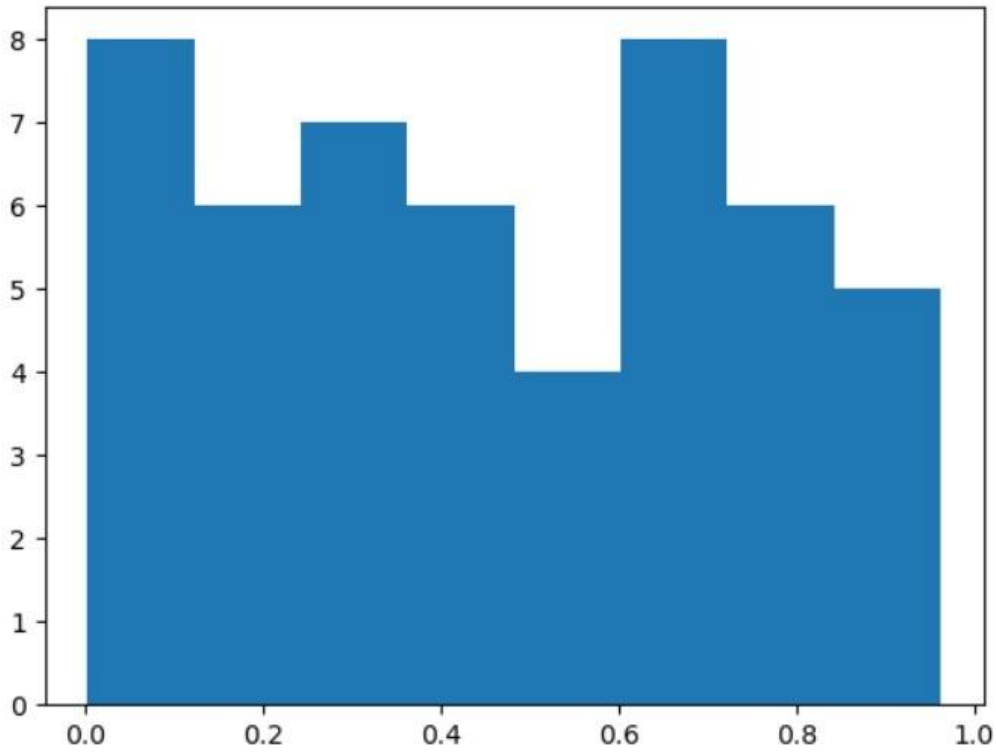


Figure 3: Randomly distributed numbers from 0 to 1, 8 bars histogram.

Moving on to the theoretical distribution, the results are:

(a) Mean: 0.4807220937906957.
(b) Standard deviation: 0.2501018306924198.
(c) First quartile is: 0.0103529878129619.
(d) Third quartile is: 0.47239969526866554.

(e) The median is: 0.47239969526866554.

Also with the requirements set for an element being an outlier we do not find any.

We can see that both theoretical data and empirical data are very similar and therefore all the statistic values calculated are close as well.

2. In part A.2, the calculated statistic values on the data provided are:

   (a) Mean: 129.73809523809524.

   (b) Standard deviation: 2.4981852824030923.

   (c) The IQR goes from 124.0 to 130.0.

   (d) The 15th percentile is: 127.15.

   (e) The 85th percentile is: 132.0.

   (f) The median is: 0.47239969526866554.

   The histogram of the sampled lap times from the provided data is:

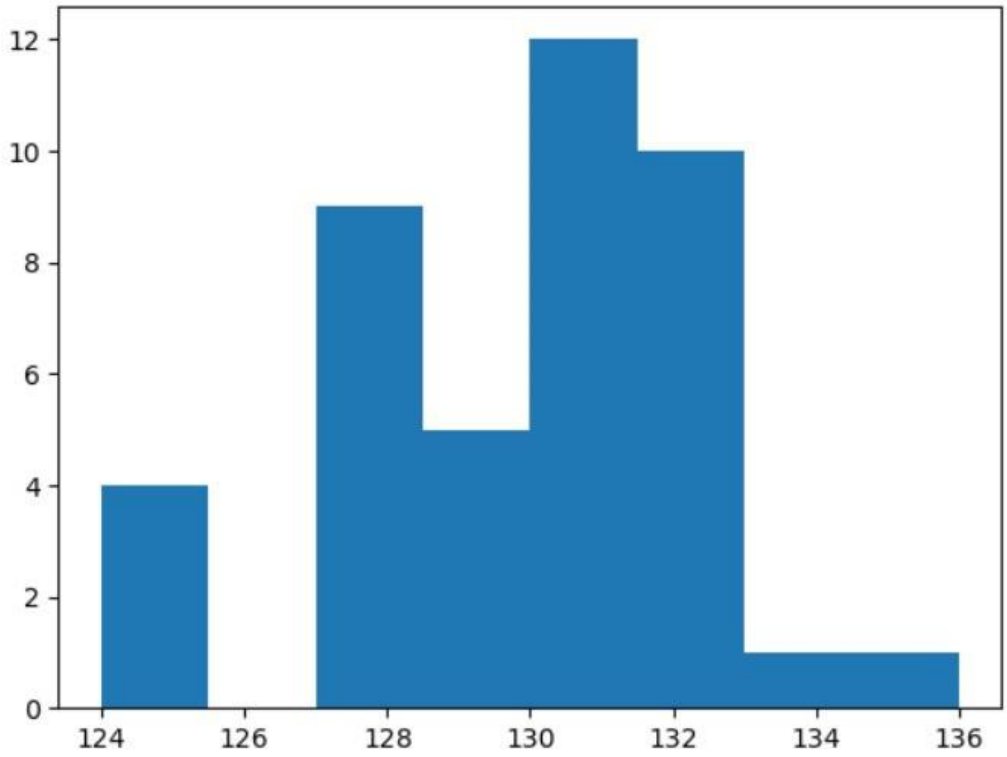


Figure 4: Sampled Lap times Provided: Histogram.

The empirical probability of a lap time being 130s is: 0.5714285714285714. The number of laps which were greater or equal to 130s are: 24.

Generating the data randomly with the random module based on the obtained distribution, we obtain the following statistic results:

(a) The IQR goes from 123.31097519345448 to 130.12245310831264.

(b) IQR = 6.811477914858159.

(c) The 15th percentile is: 127.59617188433448. (d) The 85th percentile is: 131.77689675683496.

(e) The median is: 130.12245310831264.

(f) The theoretical probability of a lap time being 130s is: 0.5238095238095238.

The 85th percentile is a value such if we would order all the values from smallest to biggest, it would be bigger than the first 85 percent of the data.

3. In part A.3, we first calculate the statistic values of the mean and standard deviation over the whole data:

   (a) The mean time of the recipes is: 3.566666666666667.

   (b) The standard deviation of those times is: 2.147608489045949.

   Moving on, we now generate two groups of samples, with different amount of samples so that we can test how the Central Limit Theorem works, we now will focus on the group with 5 samples subgroups:

```
[1, 11, 5, 6, 1]
[6, 3, 1, 4, 2]
[4, 4, 4, 1, 5]
[1, 5, 5, 6, 5]
```

Figure 5: 5-samples-subgroups group.

(a) The mean times for each random selection of samples is: [4.8, 3.2, 3.6, 4.4].

(b) The mean time for all the previous samples of size 5 is: 4.0 The standard deviation for all the previous times with sampling size 5 is: 2.4083189157584592.

Focusing now on the group with 10 samples subgroups:

```
[4, 6, 5, 6, 6, 2, 11, 6, 5, 5]
[5, 5, 1, 2, 3, 6, 1, 1, 6, 4]
[5, 5, 1, 2, 2, 1, 2, 6, 4, 6]
[4, 5, 2, 5, 2, 4, 2, 4, 3, 5]
```

Figure 6: 10-samples-subgroups group.

(a) The mean times for each random selection of samples is: [5.6, 3.4, 3.4, 3.6].

(b) The mean time for all the previous samples of size 5 is: 4.0 The standard deviation for all the previous times with sampling size 5 is: 2.0615528128088303. We can see that this one is much closer to the original one with the whole data.

Now, after all those statistic values have been calculated, we can do histograms to see the distributions of both the original data and sampled one:

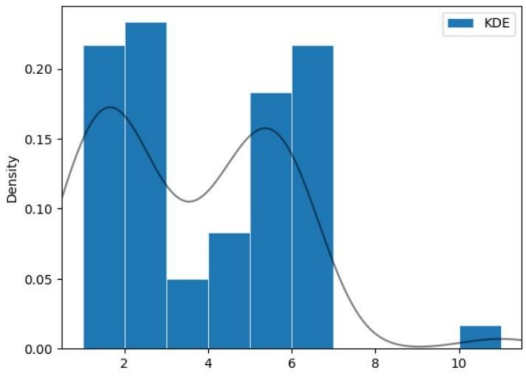


Figure 7: Histogram of the original data.

In figure 7, we can see how the distribution of the original big bulk of data does not look like a normal distribution whatsoever and follows some other distribution of data. It instead has 2 clear bumps.

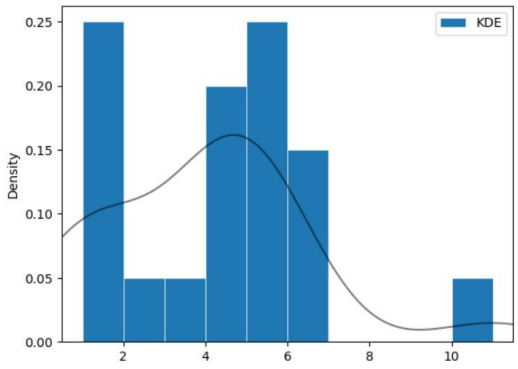


Figure 8: Histogram of the sampled data (10 element subgroups).

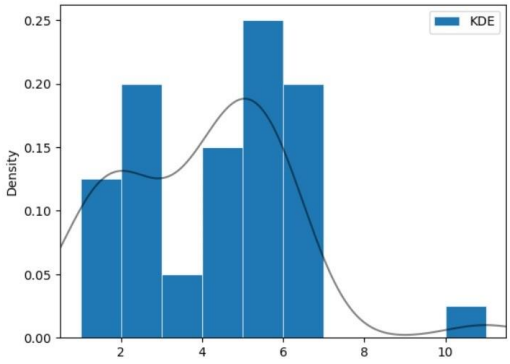


Figure 9: Histogram of the sampled data (5 element subgroups).

We can clearly see how in the histograms for the sampled data, the density function tends to lose the first bump and looks closer to a normal distribution, which is exactly what the Central Limit Theorem explains, "establishes that, in many situations, for identically distributed independent samples, the standardized sample mean tends towards the standard normal distribution even if the original variables themselves are not normally distributed."[1]

4. In part A.4, exploring the Melbourne dataset [2], I mainly focus on three tasks which are: Finding a relation between the distance to the CBD and the prices of houses, The year a house has been built in and the price of it, and last, one a bit more enthusiastic as it uses data not directly provided in the dataset but that is calculable: Finding a relation between the number of houses in a suburb and the average price of a house in that suburb.

   Before continuing with the individual tasks, for the first two, I calculated a correlation matrix that gives us the relation between different attributes in a dataset, so if the value is high that means that they are highly related and that one attribute either negatively or positively is proportional to the other one. We have not pre-processed the data but rather just saw what could proportionally mean a strong relationship; in order to see that clearly it is always better to normalize the data so that the range of both attributes is in the same magnitude, but it was not really necessary to see the relationships, and also a bit more fun to work with because we can also disprove relations that appear due to just having similar magnitudes.

   Without further due, lets examine the tasks one by one:

   (a) The correlation between the distance to CBD and the price of the houses shows some negative relation, meaning that the smaller the distance is the bigger the price would be. To prove that I have plotted such relation in a scatter plot.

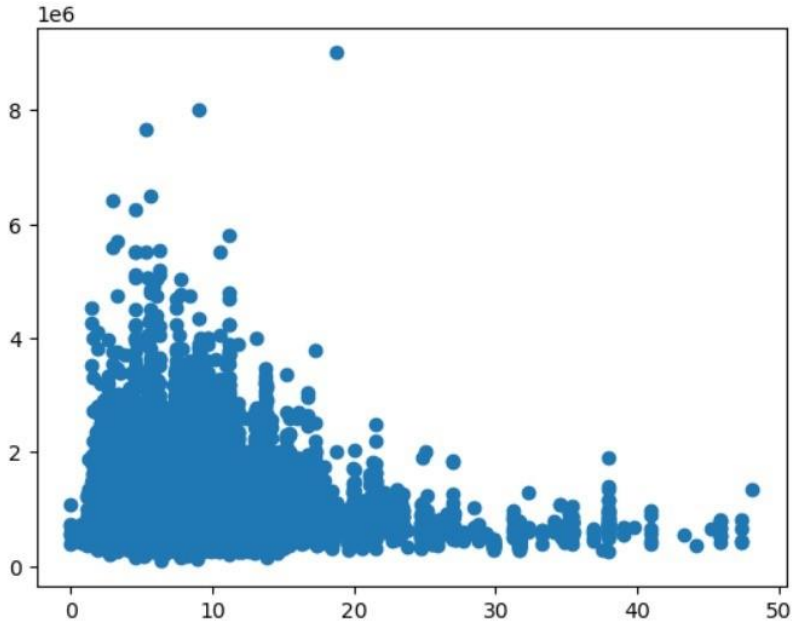


Figure 10: Relation between the distance to the CBD (horizontal axis) and the price of the house (vertical axis).

We can observe how this relation is proved by figure 10, as the prices clearly go up the closest the house is to the CBD.

(b) The correlation between the year that the house was built in, and the price of the house, shows a strong negative dependence as well, but in the next scatterplot we will disprove that relation.

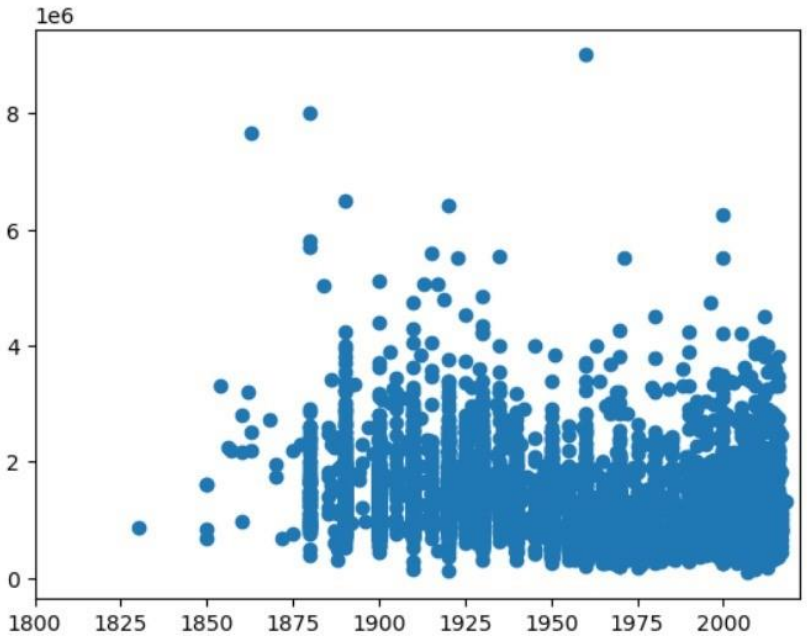


Figure 11: Year in which the house was built (horizontal axis) and the price of the house (vertical axis).

We can observe how this relation is disproved by figure 11, as the prices do not clearly go up as the houses get newer, that is probably due to monumental buildings such as castles.

(c) The last task that we have worked in, is the relation between the number of houses in a suburb and the average price of the houses, we wanted to check if there exists an "exclusivity factor" that affects house prices just by them being more exclusive in the zone that they are located:

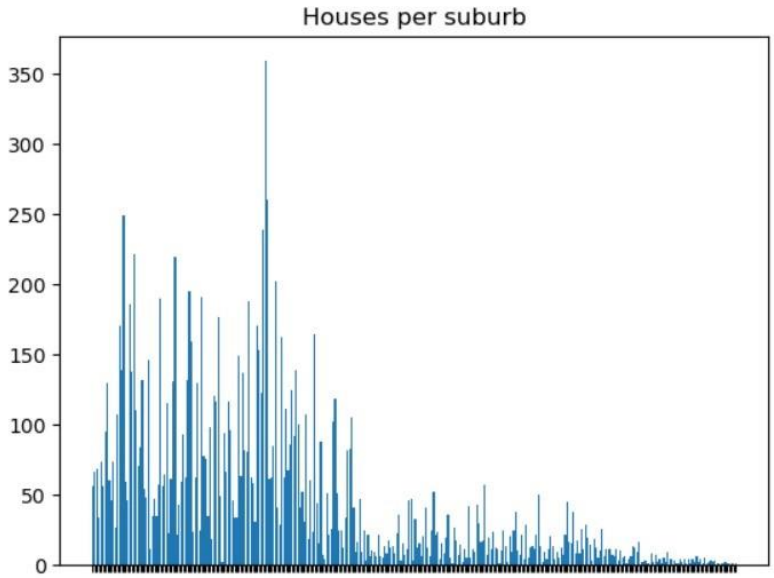


Figure 12: Number of houses (vertical axis) in each suburb (horizontal axis)).

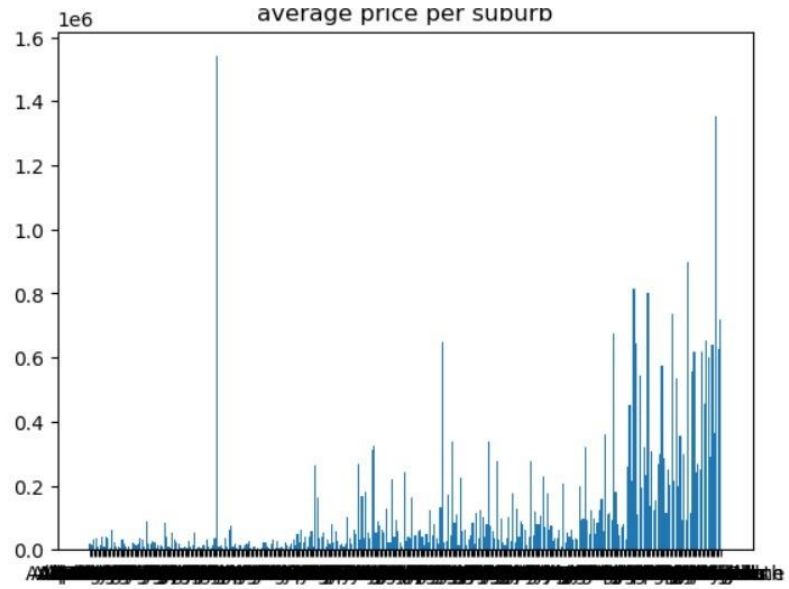


Figure 13: Average price of houses (vertical axis) in each suburb (horizontal axis)).

So far the results are promising, here an inverse relation is easy to detect, but let's go further and calculate the correlation between the number of houses in a suburb and the average price of a house in a suburb. After, let's also plot the relation in a scatter plot to see how it looks like:

| | Number of houses | Avg price |
|---|---|---|
| Number of houses | 1.000000 | -0.385093 |
| Avg price | -0.385093 | 1.000000 |

Figure 14: Correlation between the number of houses in a suburb and the average price of a house in a suburb.

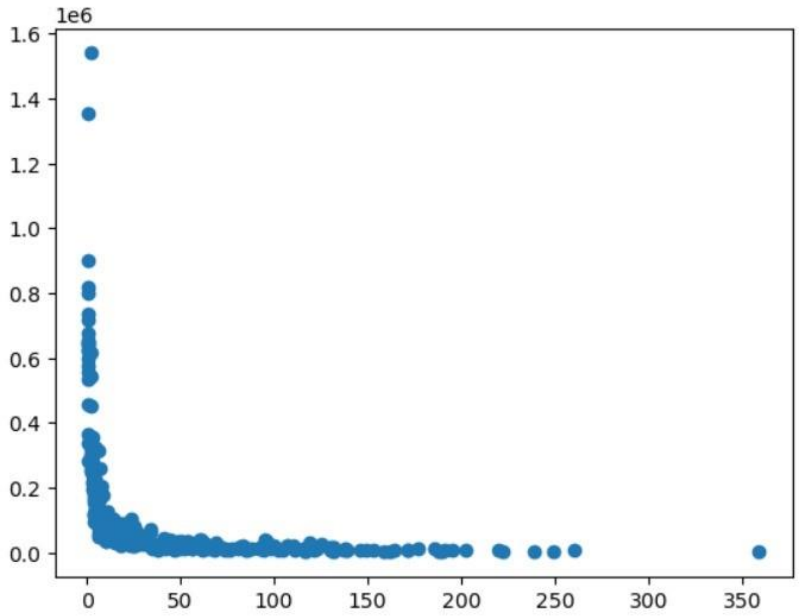


Figure 15: Number of houses in a suburb (horizontal axis) and the average price of a house in a suburb (vertical axis).

# 2 Discussion

As a discussion, it is interesting to mention that the relation presented in the last case which I believe is a really successful finding, and the relation between the year the houses were built in with the price of the houses, have very similar values of correlation.

Yes, even if it seems impossible as the last relation was presented to be a really strong and clear one as saw in figure 15, the correlation value was really close to that of the relation between the years and the price.

Thinking about this, what comes to mind is that maybe we should always double check what we believe is proved by just an algorithm, so we believe that it is a good practice to use more than one method always that something wants to be either proved or disproved.

# 3 Conclusion

From our perspective, the first lab was parallel to the course syllabus and a good exercise for understanding better how statistical distributions work and getting some proper experience with other data science algorithms and methods that will come in handy for both real-life work situations and the project.

We are happy with what we have learnt in this lab, but something that took quite more time than expected was getting all the results from each task related to statistics, especially having to introduce manually the data from some datasets.

# References

[1] ”central limit theorem”. https://en.wikipedia.org/wiki/Central_limit_theorem. Accessed: 28.02.2022.

[2] ”melbourne real estate dataset”. https://www.kaggle.com/dansbecker/melbourne-housing-snapshot. Accessed: 20.02.2022.