Task 3

Explanation about task 3

In this task we chose two different data:

1) Data about the 20 years of Olympic history(athletes): their age, height and weight for almost 27,000 athletes

the link for this data is

https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results

2) Data about the population of the cities in India (population of men, population of women, men who graduated, women who graduated for almost 500 different cities in India

the link for this data is

https://www.kaggle.com/zed9941/top-500-indian-cities

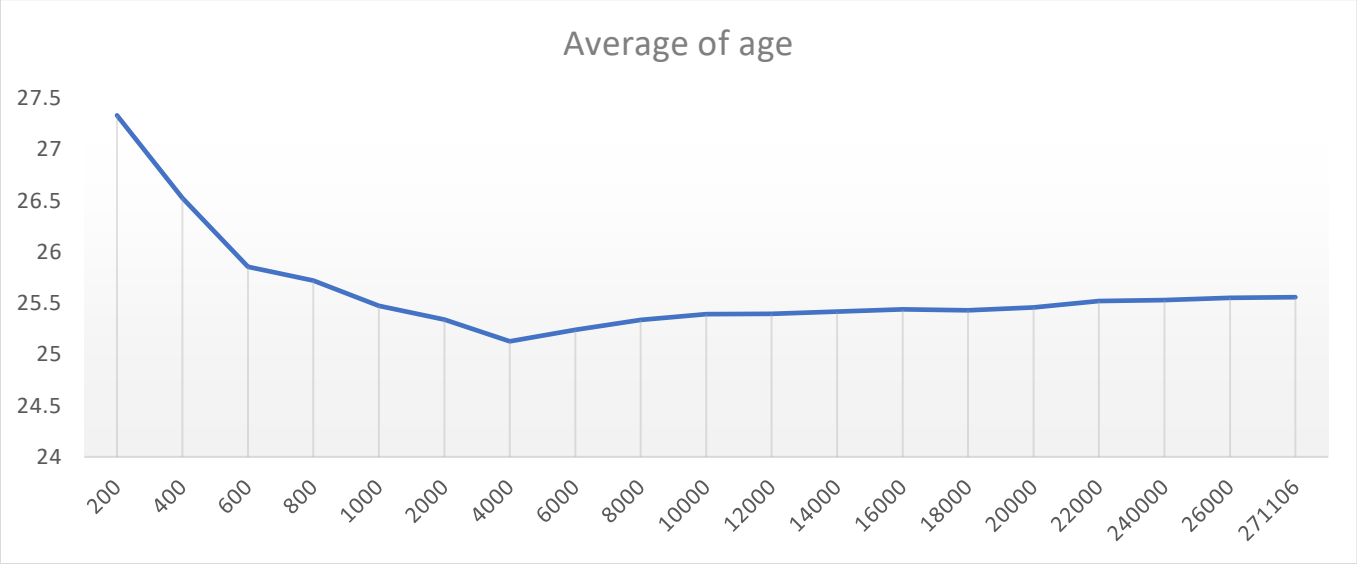
these two data are from Kaggle website and there are real as many data scientists and machine learning developer take this data to use it in their work

we use every Random variable in this data, take samples, take their average and each time we increase the size of the sample until the mean you compute is independent from the sample size

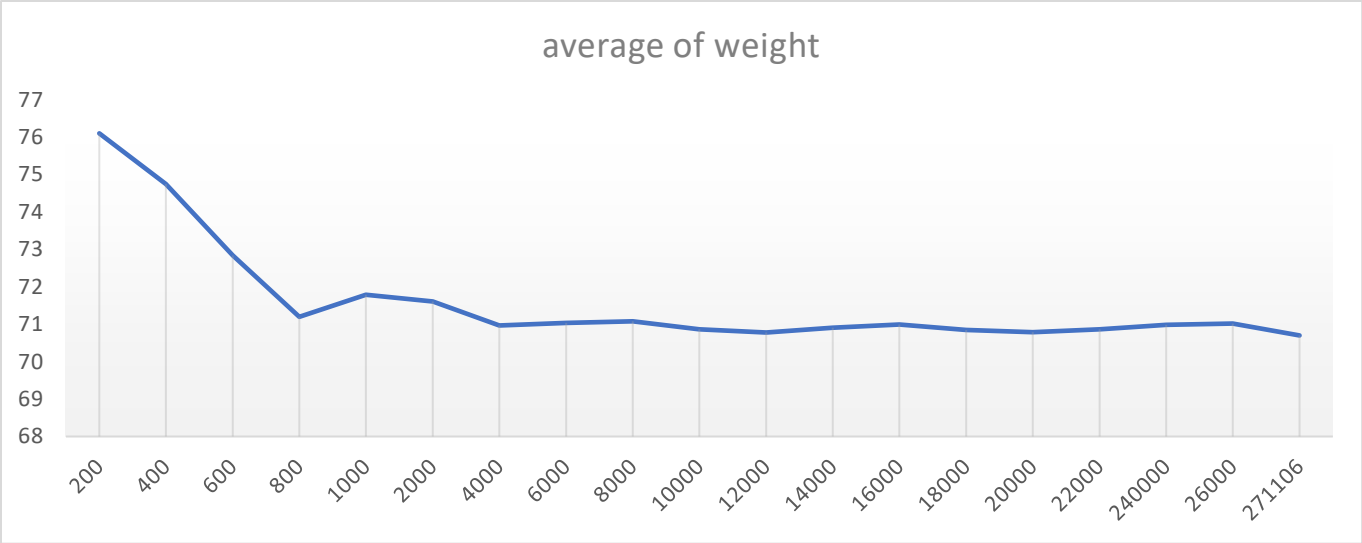
we also include the excel data sheets in our submission and we include the table of the number of samples, the mean of them and the plots which explain the results in this report

Data about the 20 years of Olympic history(athletes)

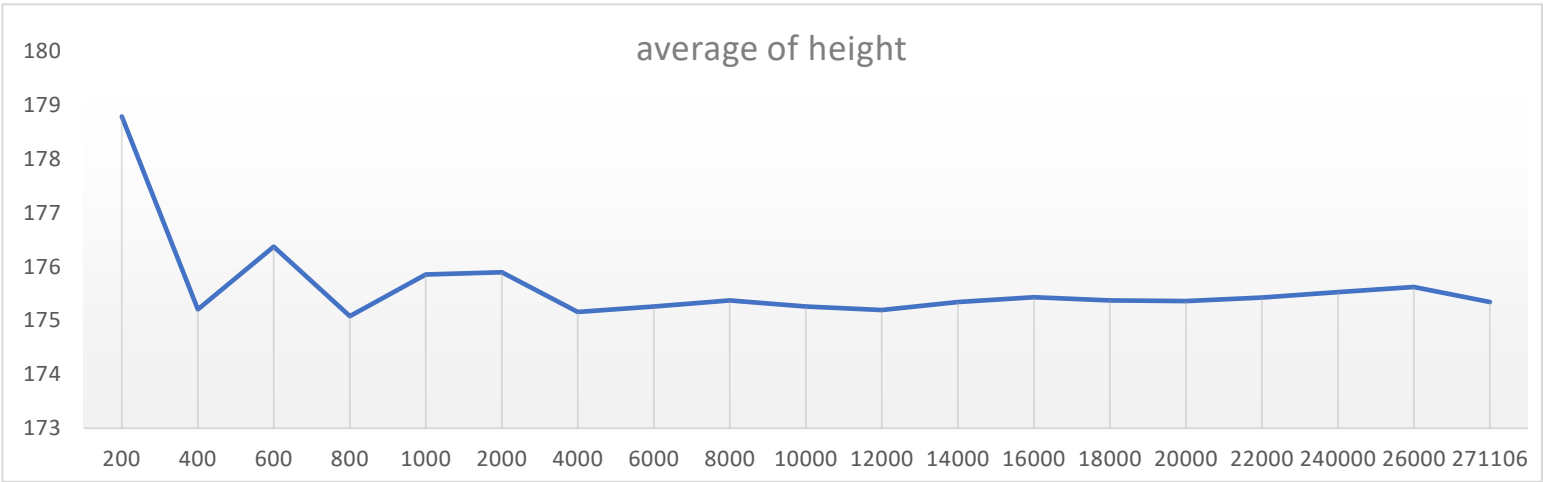
Number of sample	average of age	average of height	average of weight
200	27.33163265	178.7831325	76.10542169
400	26.52432432	175.2069173	74.74914676
600	25.85424354	176.3673469	72.84451902
800	25.72027027	175.0793388	71.20343137
1000	25.47435897	175.8503937	71.79335072
2000	25.34032512	175.8928571	71.6120801
4000	25.12734975	175.1595711	70.97222222
6000	25.2383629	175.2569809	71.03729456
8000	25.33596215	175.3727668	71.08134824
10000	25.39120395	175.2568819	70.86644147
12000	25.39523102	175.1957311	70.78490627
14000	25.41817502	175.3400329	70.9071356
16000	25.43783077	175.4330414	70.99428295



for **age variable**, we recommended to take a sample size **more than or equal 22000** because it's obvious that the mean started to be constant from that size



for **weight variable**, we recommended to take a sample size **more than or equal 10000** because it's obvious that the mean started to be constant from that size

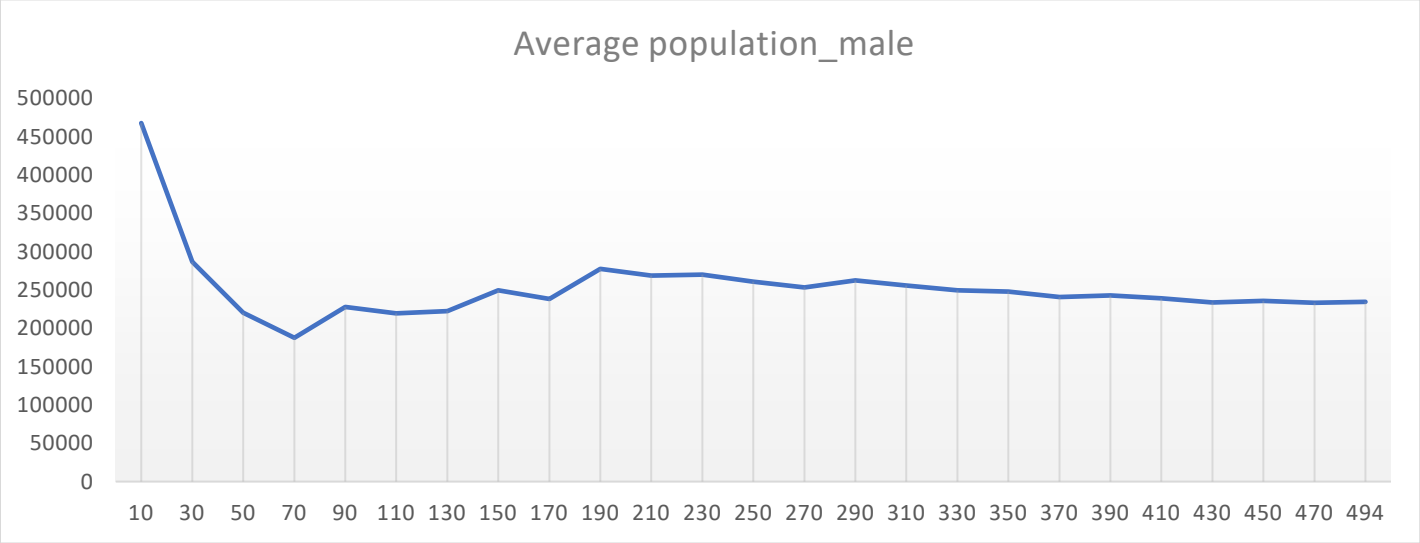


for **height variable**, we recommended to take a sample size **more than or equal 10000** because it's obvious that the mean started to be constant from that size

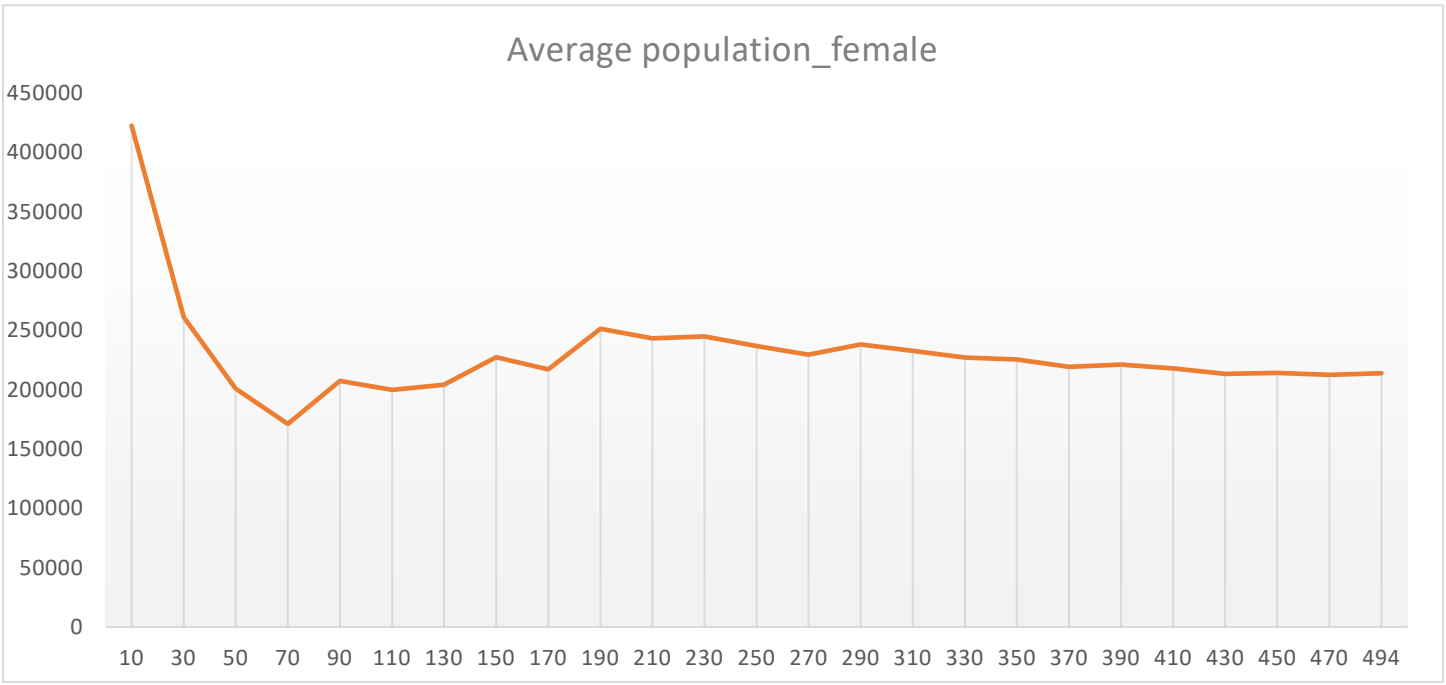
so, as a result: while increasing the sample size the mean will be orbit around a specific value until we reach a constant value for the mean which don't depend on the sample size

Data about the population of the cities in India

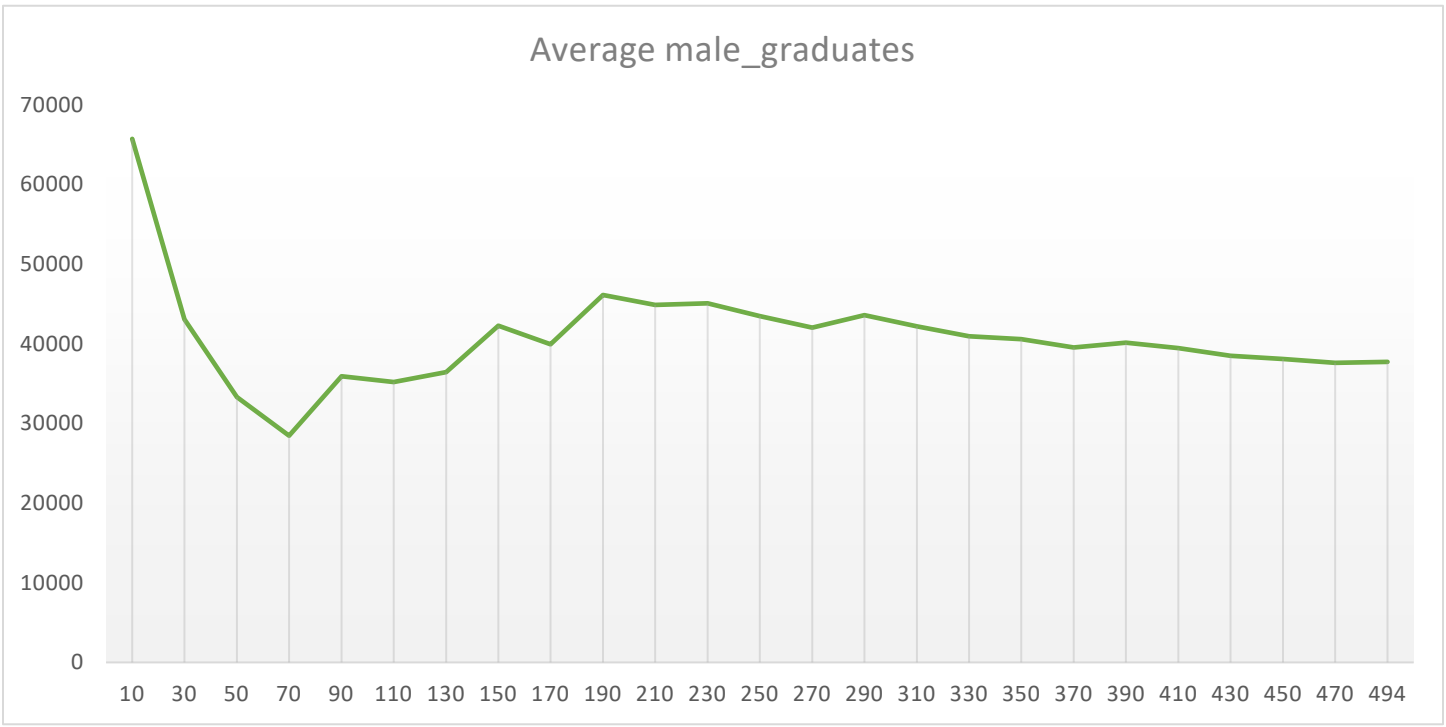
number of samples	Average population_male	Average population_fe	Average male_gradua	Average Female_graduat
10	467738	422560.6	65703.4	49668.3
30	286739.7667	260961.2667	43040.96667	33282
50	220414.32	201037.94	33301.72	24821.1
70	187602.0143	171184.7286	28454.58571	21062.1
90	227922.9222	207372.4667	35911.77778	26391.48889
110	219693.0455	199955.5818	35184.88182	25557.90909
130	222286.2615	204096.9923	36433.5	26897.34615
150	249553.0533	227356.3133	42270.37333	31744.87333
170	238262.5059	217084.4353	39907.08235	29899.39412
190	277606.0526	251481.1263	46104.81579	34793.64737
210	268621.6905	243346.5905	44872.73333	33750.66667
230	269901.213	244897.8348	45082.43043	33928.51304
250	260717.16	236741.708	43444.132	32607.292
270	253127.1778	229531.6	41999.66667	31543.94815
290	262345.0517	238211.0724	43583.2069	33025.84828
310	255959.1613	232851.5774	42188.75161	31962.58065
330	249505.0818	227006.5879	40942.49091	30976.20909
350	247780.7086	225573.0657	40565.88857	30710.76
370	240860.1676	219407.2459	39536.98919	29911.69189
390	242803.6256	221207.8154	40128.13077	30327.29487
410	238945.9268	217797.3585	39432.29024	29779.01707
430	233848.3023	213250.3372	38473.43256	29033.44651
450	235774.9111	214255.1356	38088.6	28705.24
470	233376.8574	212541.5574	37606.27021	28441.22128
494	234346.789	213765.5842	37715.56187	28486.79513



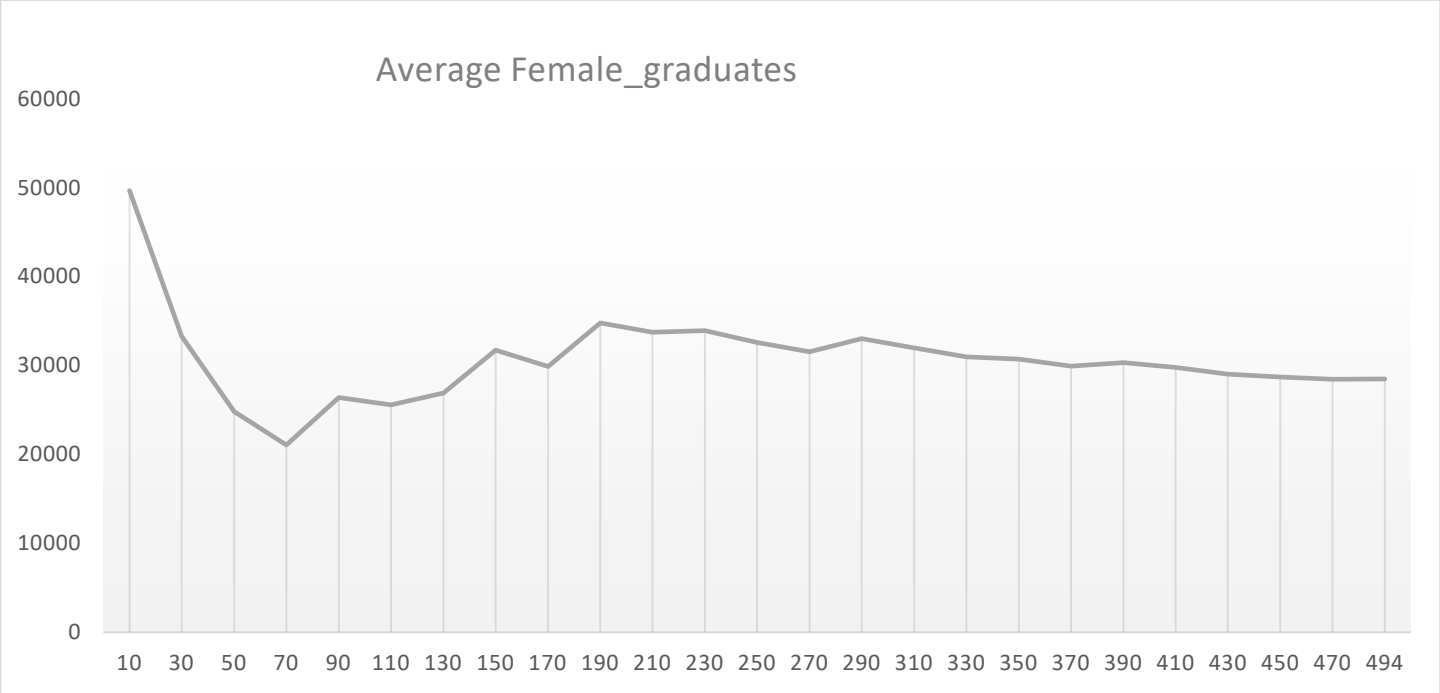
for **population male variable**, we recommended to take a sample size **more than or equal 430** because it's obvious that the mean started to be constant from that size



for **population female variable**, we recommended to take a sample size **more than or equal 430** because it's obvious that the mean started to be constant from that size



for **male graduate's variable**, we recommended to take a sample size **more than or equal 450** because it's obvious that the mean started to be constant from that size



for **female graduate's variable**, we recommended to take a sample size **more than or equal 450** because it's obvious that the mean started to be constant from that size

so, **as a result**: increasing the sample lead to a different value of mean until a specific number of sample (which depend on the variable) where the mean is independent from the sample size