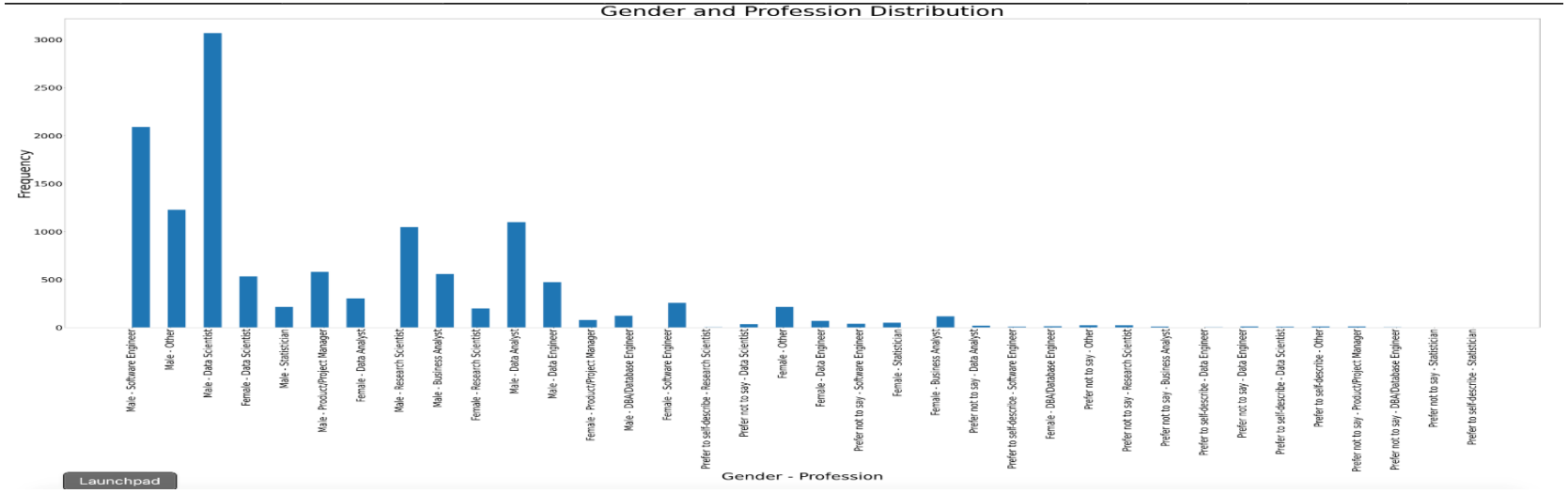
UNIVERSITY OF TORONTO

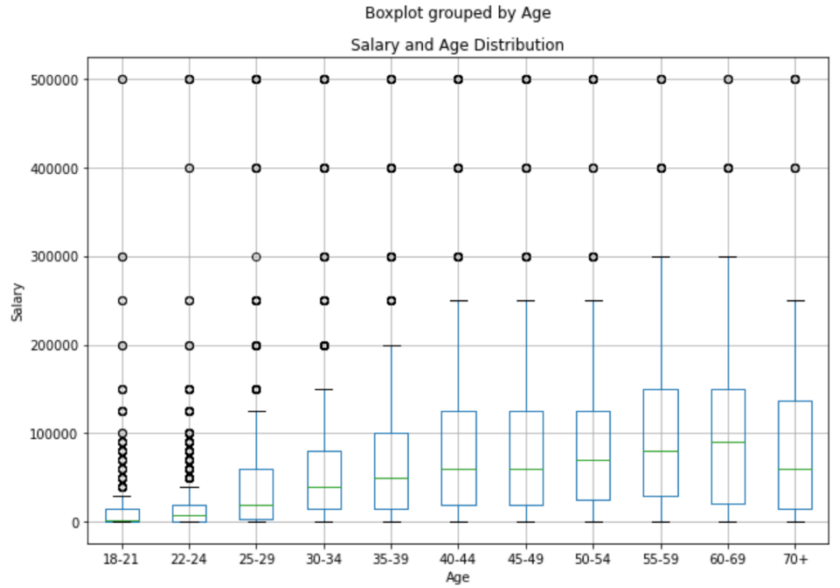
MIE1624H-Introduction to Data Science & Analytics Assignment-1

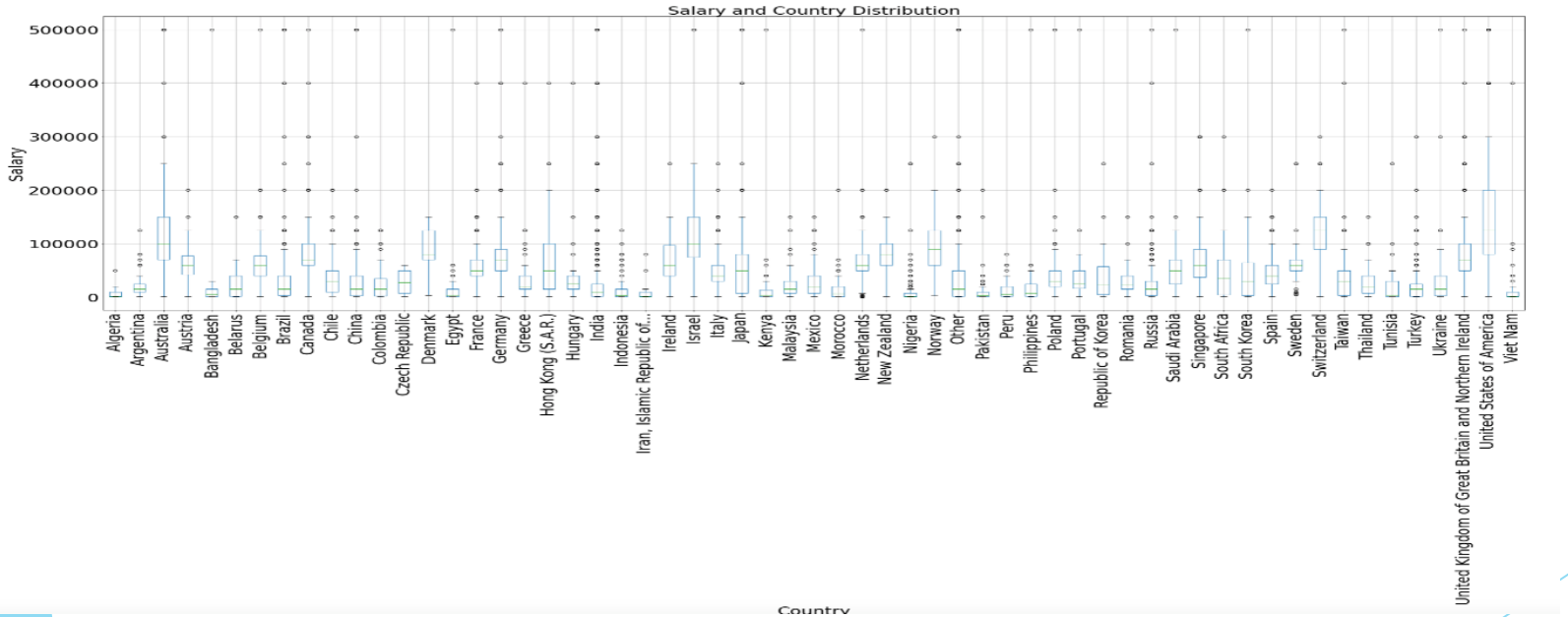
Content Table:

Name- TUHIN RANJAN
Program-MENG MIE
Student No-1006633555

Descriptive Analysis







KEY POINTS FOR Q1:

1. Data from provided dataset is being reoriented to plot graphs for describing the effect of data science profession on the salary structure due to gender , country or age.
2. Histogram and boxplot are being use to demonstrate the conclusions.
3. Gender & Profession Distribution- Males are dominating in field of data science and its oriented profession.
4. Developed Countries have higher job opportunity and higher salary wrt to data science related profession
5. Upto 30 yrs of age, the salary distribution is quite low in data science profession.

Male & Female Vs Salary Distribution Comparison

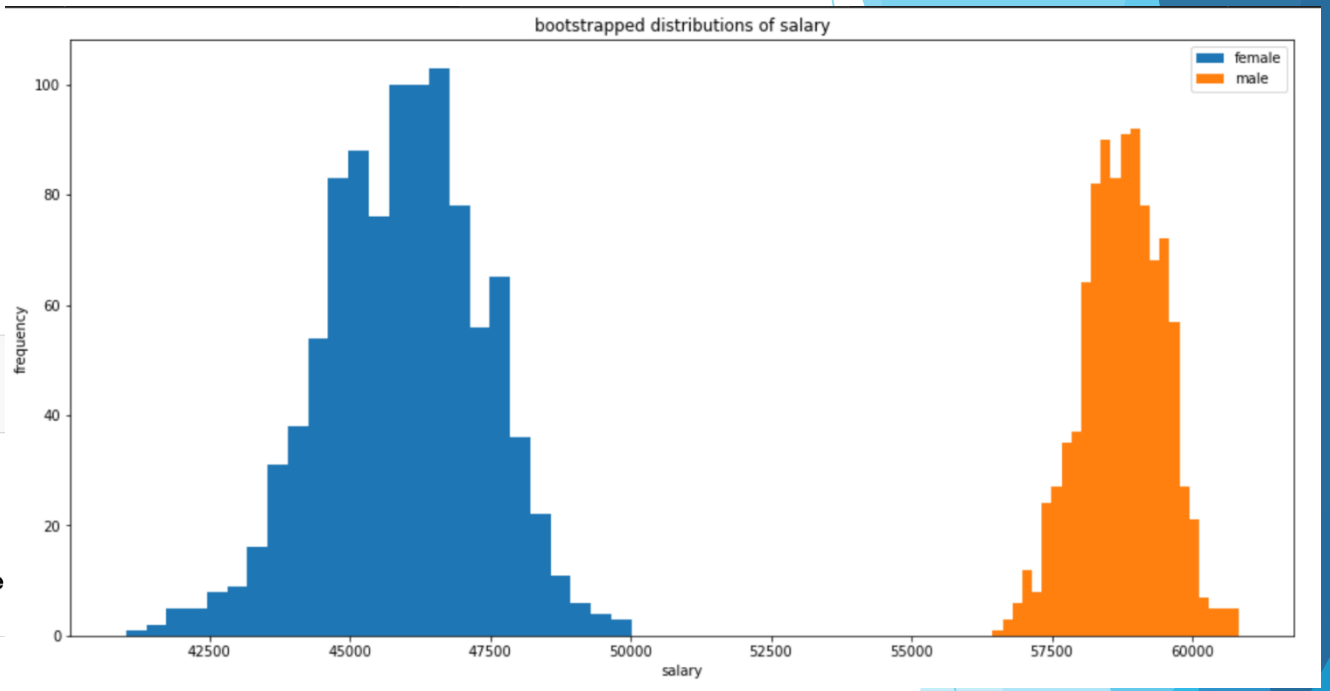
	Salary							
	count	mean	std	min	25%	50%	75%	max
Gender								
Female	1827.0	45933.771210	60253.789591	1000.0	3000.0	20000.0	70000.0	500000.0
Male	10473.0	58709.586556	74920.620048	1000.0	7500.0	30000.0	80000.0	500000.0
Prefer not to say	167.0	70664.670659	95027.634584	1000.0	5000.0	40000.0	90000.0	500000.0
Prefer to self-describe	30.0	109783.333333	145801.097587	1000.0	9375.0	75000.0	125000.0	500000.0

In [46]:
```
#Performing ttest using the groupping of subset created

salary_gender_ttest = pg.ttest(male_salary, female_salary, correction=False)
display(salary_gender_ttest)
```

	T	dof	tail	p-val	CI95%	cohen-d	BF10	power
T-test	6.909348	12298	two-sided	5.108939e-12	[9151.36, 16400.27]	0.17518	5.991e+08	1.0

pval<0.05, thus rejecting null hypothesis as salary does depend on gender, though larger the sample, the more likely the difference of a given size will be, thus will be checking through bootstrap methodolgy.

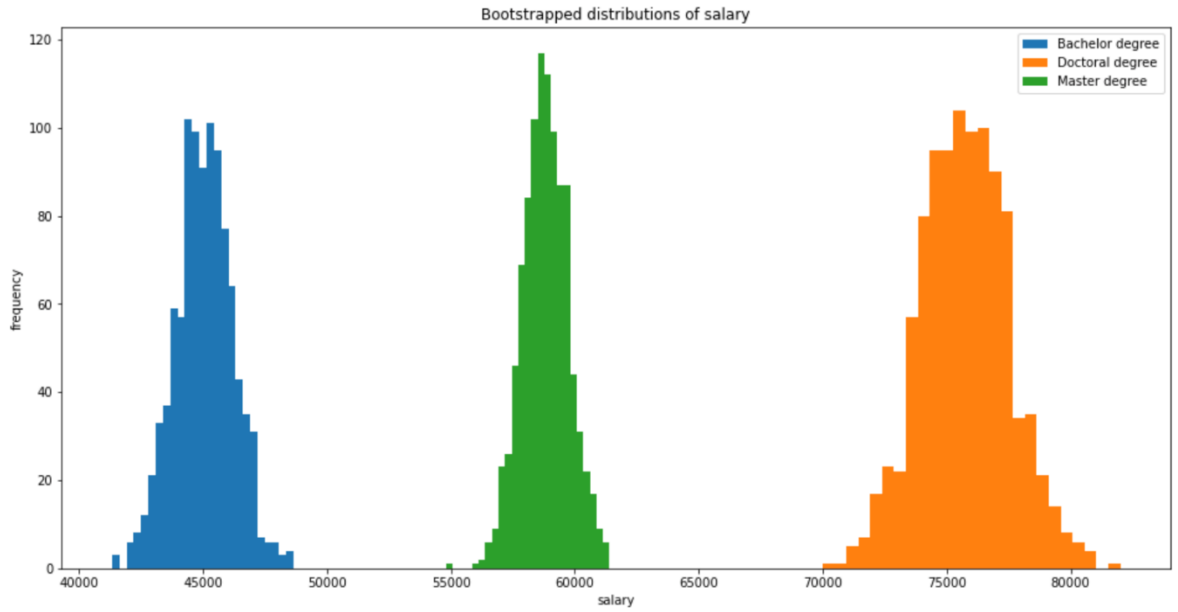


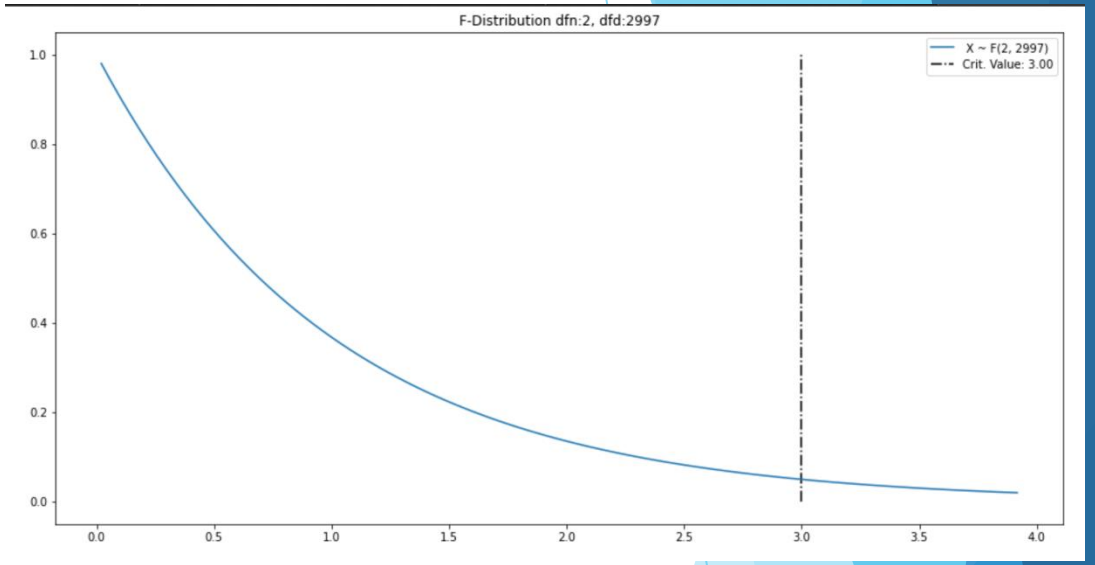
KEY POINTS FOR Q2:

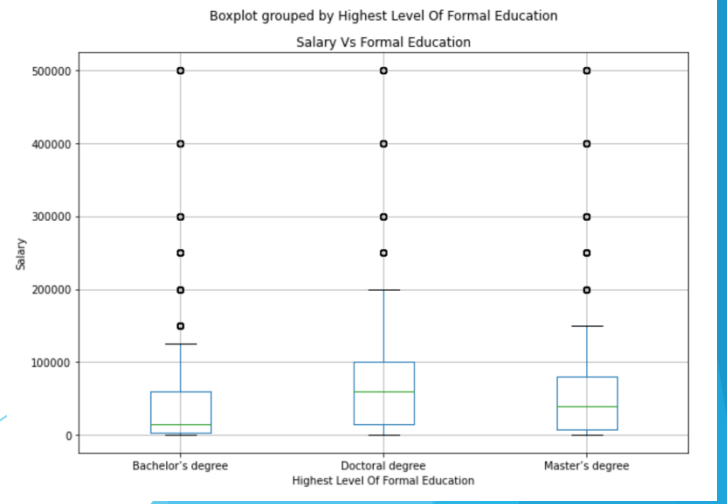
1. Data is cleaned up to find the statistical significant difference of mean salary between male & female.
2. Descriptive data for both groups found out and ttest is being performed to check if p<0.05 or less than 5%.
3. ttest gave the value for p=5.108939e-12 < 0.05, thus rejecting the null hypothesis. Since larger the sample number more confidence of interval will be, so executing bootstrap with resampling of data to 1000 times.
4. After executing bootstrap, it been concluded that p=0.0<0.05 through Hypothesis test and rejecting the null hypothesis.
5. Means of salary between male and female is statically significant or salary is depending upon the gender.

UNIVERSITY OF TORONTO

Highest Level of Formal Education Vs Salary Distribution Comparison





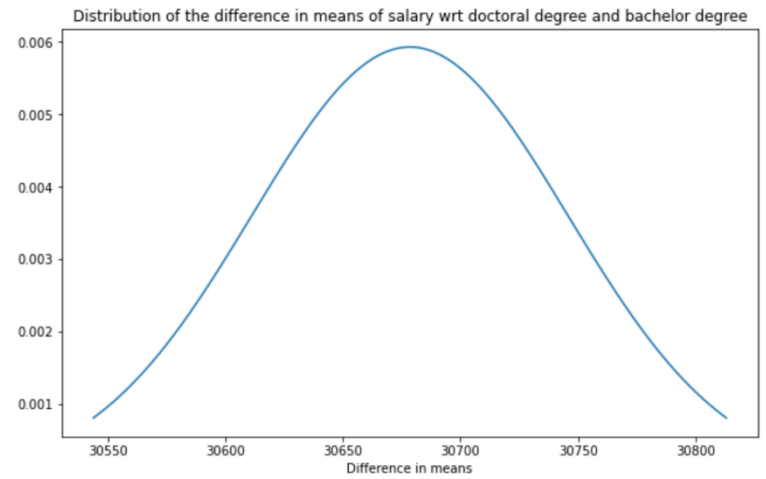


KEY POINTS FOR Q3:

1. Data is cleaned up to find the statistical significant difference of mean salary between the highest level of formal education i.e., bachelor degree, master degree and doctoral degree.
2. Descriptive data for both groups found out and also represented as boxplot.
3. Since there are 3 mean values need to compare so, rejecting ttest method and executing bootstrap method for higher resample size of 1000. Here we used ANOVA method to check the null hypothesis.
4. After executing bootstrap, it been concluded that p=0.0<0.05 through Hypothesis test and f value=131162.24882062527> F critical value, thus rejecting the null hypothesis.
5. Means of salary between bachelor, masters and doctoral degree is statically significant or salary is depending upon the highest level of formal education which is directly proportional.

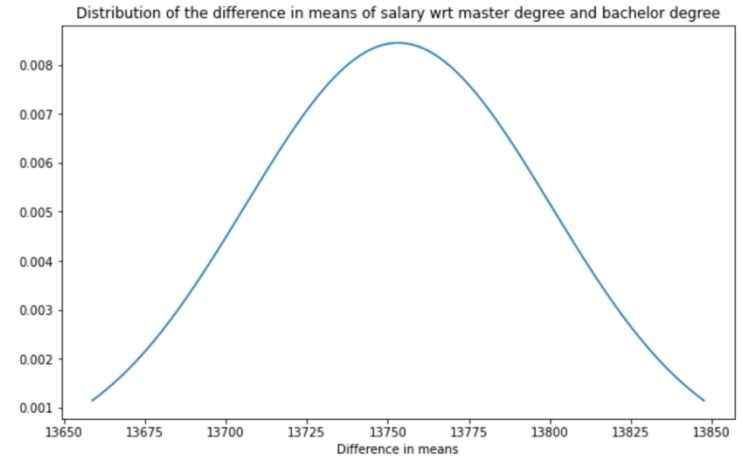
Differences In Means Plots

```
mu_degree1,sigma1,mu_degree1 - 2*sigma1, mu_degree1 + 2*sigma1 #value between
```



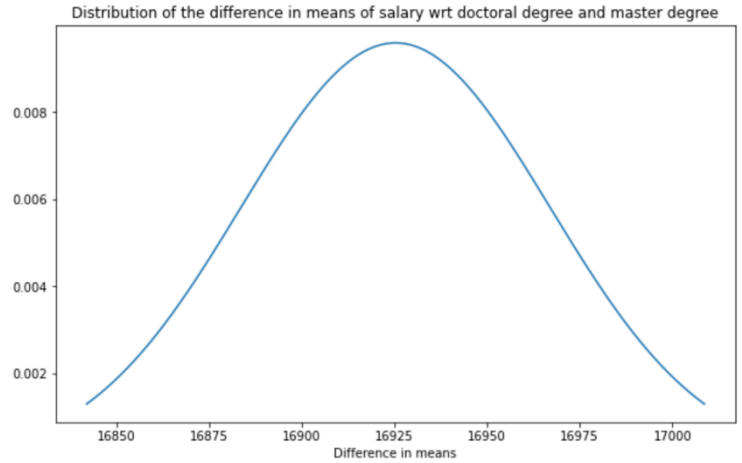
: (30678.488101136936, 67.29075444383322, 30543.90659224927, 30813.069610024602)

```
mu_degree3,sigma3,mu_degree3-2*sigma3,mu_degree3+2*sigma3 #value between the con
```



: (13753.200055008667, 47.21200725736024, 13658.776040493947, 13847.624069523386)

```
mu_degree2,sigma2,mu_degree2 - 2*sigma2, mu_degree2 + 2*sigma2 #value between th
```

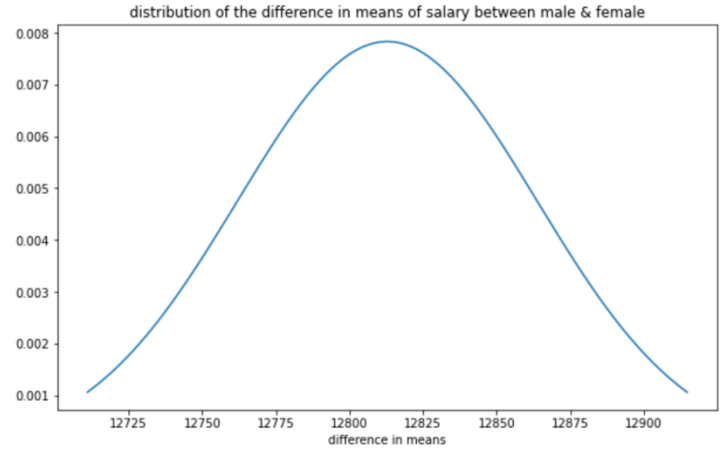


(16925.28804612827, 41.65195253472916, 16841.984141058812, 17008.591951197726)

KEY POINTS FOR Q2 & Q3:

1. Difference in means have been plotted to check the normal distribution nature of the means differences.
2. For Male & Female Means Salary differences, the value ~12812, variance~50.
3. For Master & Bachelor Means Salary differences, the value, mu ~13753, variance~47.
4. For Doctoral & Master Means Salary differences, the value,mu ~16925, variance~41.
5. For Doctoral & Bachelor Means Salary differences, the value,mu ~30678, variance~67.

```
mu,sigma,mu - 2*sigma, mu + 2*sigma #value between the confidence interval of 95%
```



51]: (12812.8674700618, 50.94955462416671, 12710.968360813466, 12914.766579310133)