FINAL PROJECT

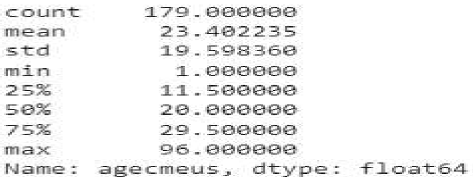
Aya Sabry Mohamed

FEBRUARY 9, 2023

DA ISMAILIA

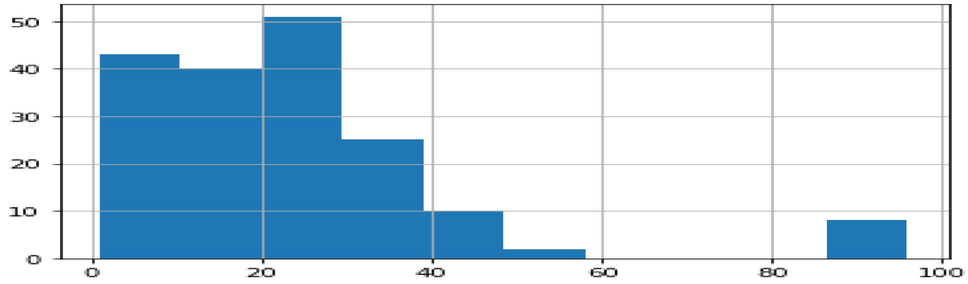
Q1: One of the questions in the 2008 General Social Survey was, if you were born outside the United States, at what age did you permanently move to the United States (AGECMEUS)?

a. Calculate the mean, variance, and standard deviation.



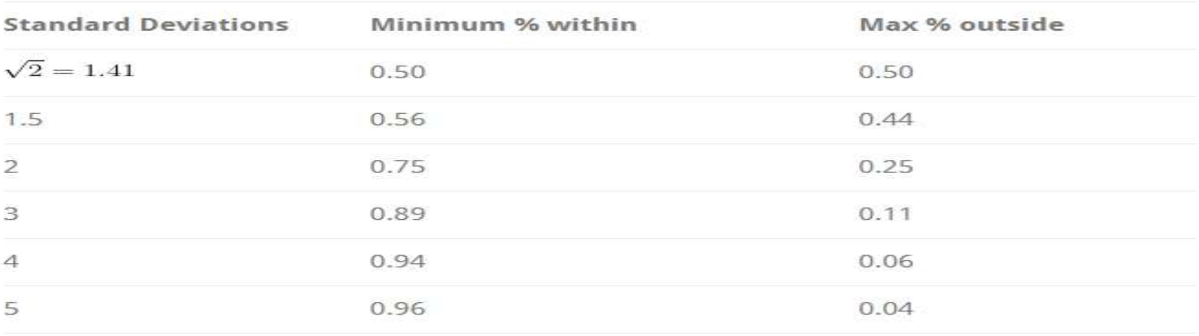
```
count     179.000000
mean       23.402235
std        19.598360
min         1.000000
25%        11.500000
50%        20.000000
75%        29.500000
max        96.000000
Name: agecmeus, dtype: float64
```

b. Draw a histogram.



c. Use the Empirical rule, if applicable, or ChebySheff's Theorem to interpret the mean and standard deviation.

We cannot apply the Empirical Rule because the histogram is not bell shaped. Instead, we must apply Chebysheff's Theorem.

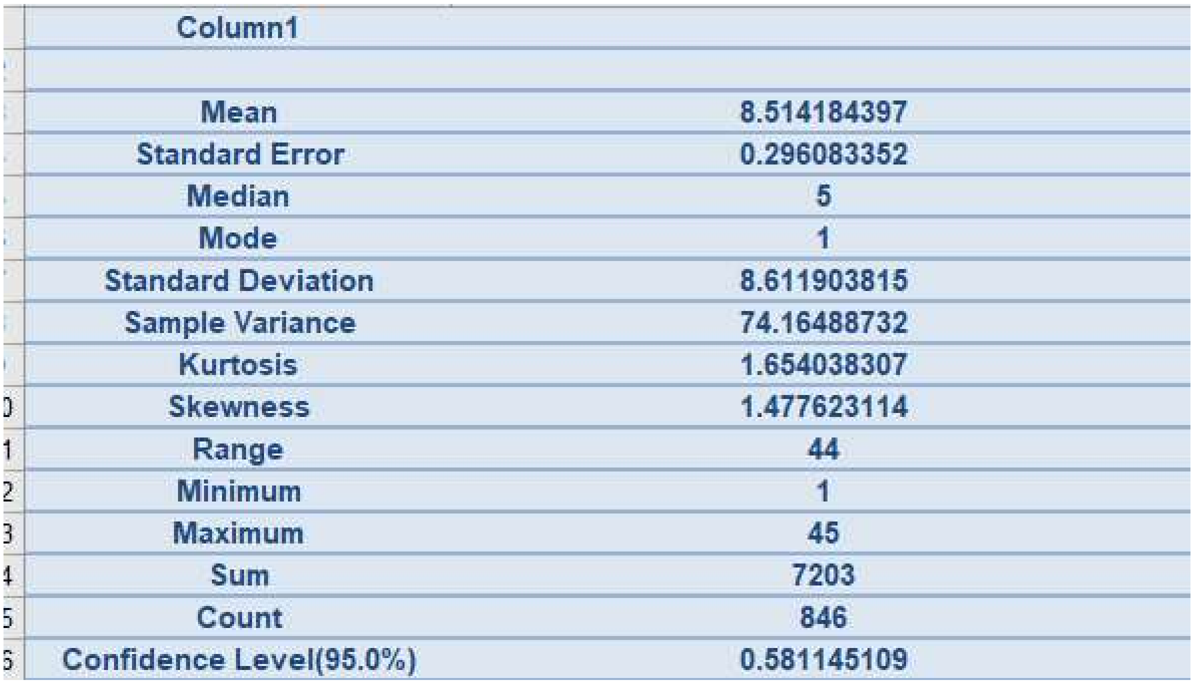


Standard Deviations	Minimum % within	Max % outside
$\sqrt{2} = 1.41$	0.50	0.50
1.5	0.56	0.44
2	0.75	0.25
3	0.89	0.11
4	0.94	0.06
5	0.96	0.04

From : https://statisticsbyjim.com/basics/chebyshevs-theorem-in-statistics/

23-(1.4* 19) =less than 0 ~1
23+(1.4* 19) =~50
But 75% of data at 29
So that more than 50% acc. To table above lies between
mean+ 1.4*std
And mean – 1.4*std

Q2: Estimate with 95% confidence the mean number of years with current employer (CUREMPYR).



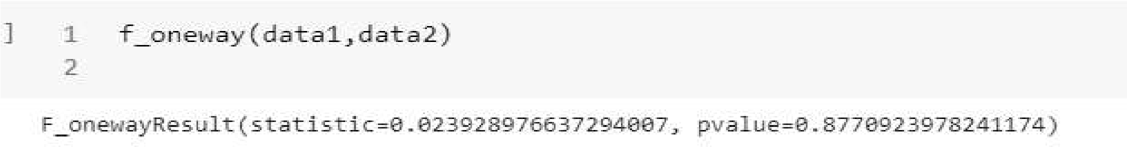
Column1	
Mean	8.514184397
Standard Error	0.296083352
Median	5
Mode	1
Standard Deviation	8.611903815
Sample Variance	74.16488732
Kurtosis	1.654038307
Skewness	1.477623114
Range	44
Minimum	1
Maximum	45
Sum	7203
Count	846
Confidence Level(95.0%)	0.581145109

Therefore, the confidence interval is **8.5 ± 0.58**, which is equal to the **range 7.92 and 9.08** (years).

Q3 Estimate with 90% confidence the proportion of Americans whose income is at least $75,000 (INCOME06).

(INCOME06 column doesn’t exist).

Q4: Is there sufficient evidence to conclude that people who work for the government (WRKGOVT: 1 = Government, 2 = Private) work fewer hours (HRS)?

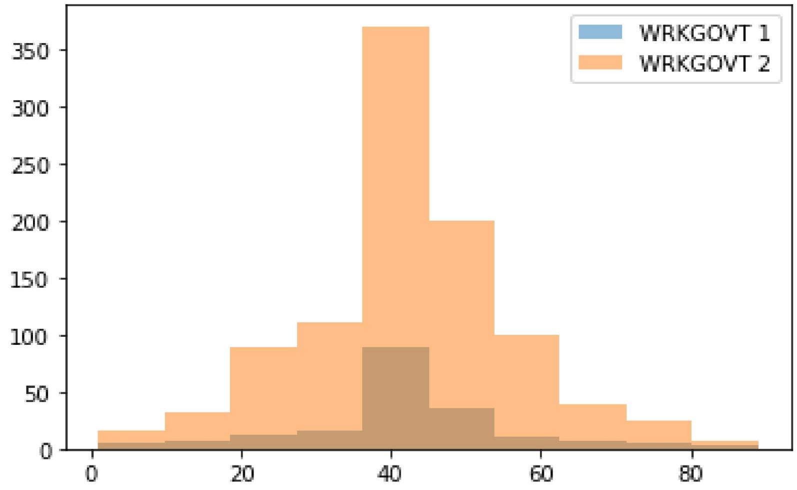


```
]    f_oneway(data1,data2)
```

```
F_onewayResult(statistic=0.023928976637294007, pvalue=0.8770923978241174)
```

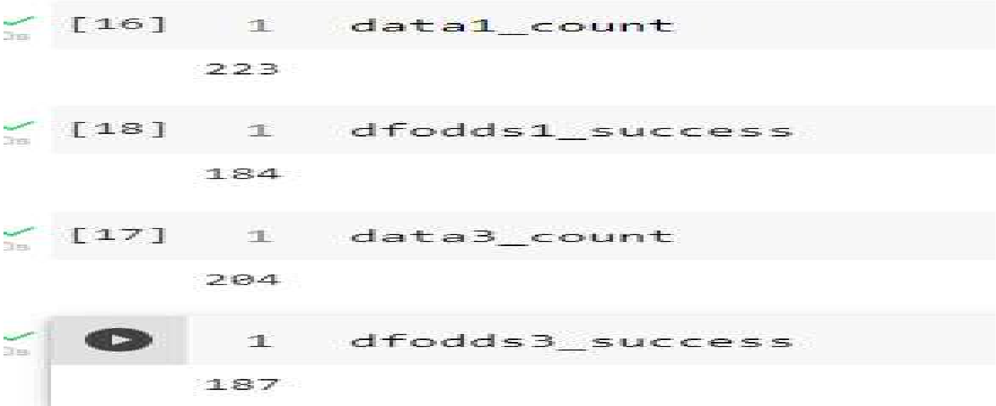
The variables that make up the variable (WRKGOVT) are,
0 = Public, 1 = Private.
HRS is an additional factor.
Here, WRKGOVT is nominal, while the response variable (HRS) contains interval data.
Therefore, the population variance test is used first, followed by either an equal or unequal t-test.
p-value=0.8770923978241174.

Here p_value > 0.05 then we accept null hypothesis.
By distributions it seems to be equal.



Q5: For each of the following variables, conduct a test to determine whether Democrats and Republicans (PARTY 1 = Democrat, 3 = Republican) differ in their correct answers to the following questions:

Correct answers to ODDS1: A doctor tells a couple that there is one chance in four that their child will have an inherited disease. Does this mean that if the first child has the illness, the next three will not? 1 = Yes, 2 = No. Correct answer: No. Democrats and Republicans (PARTY 1 = Democrat, 3 = Republican) differ in their correct answers to the following questions

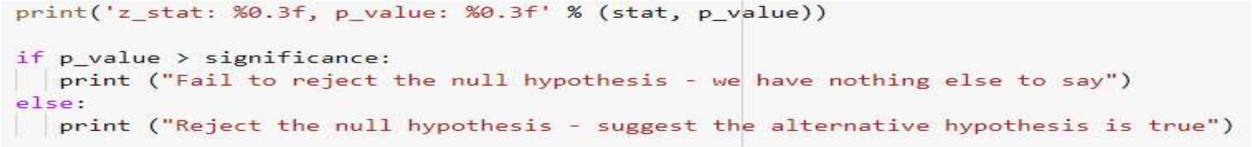


```
[16]  data1_count
223

[18]  dfodds1_success
184

[17]  data3_count
204

      dfodds3_success
187
```

We conduct test for proportion, `significance = 0.025`

For democrat n=223,success=184

For Republication n=204, success=187



```
print('z_stat: %0.3f, p_value: %0.3f' % (stat, p_value))

if p_value > significance:
    print ("Fail to reject the null hypothesis - we have nothing else to say")
else:
    print ("Reject the null hypothesis - suggest the alternative hypothesis is true")
```

```
z_stat: -2.799, p_value: 0.005
Reject the null hypothesis - suggest the alternative hypothesis is true
```

So that We can infer Democrats and Republicans (PARTY 1 = Democrat, 3 = Republican) differ in their correct answers to the following questions

Q6: Can we infer from the data that the proportion of Americans earning at least $75,000 is greater in 2008 than in 2006 (INCOME06)?

(INCOME06 column doesn't exist).

Q7: Can we infer from the data that the majority of Americans support capital punishment form urderers ? (CAPPUN: 1 = Favor, 2 = Oppose).

```
z_stat: 10.530, p_value: 0.000
Reject the null hypothesis - suggest the alternative hypothesis is true
```

Q8: Is there enough evidence to infer that differences in the amount oftelevision watched (TVHOURS) differ between classes (CLASS)?

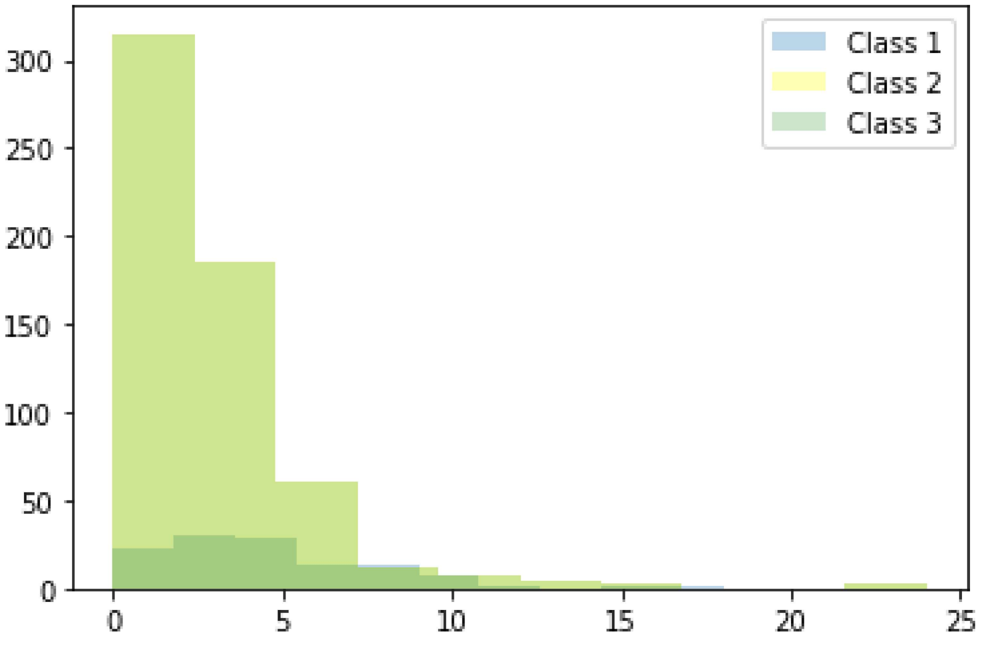
H0=U1=U2=U3   ---for each class

H1: one mean at least differ

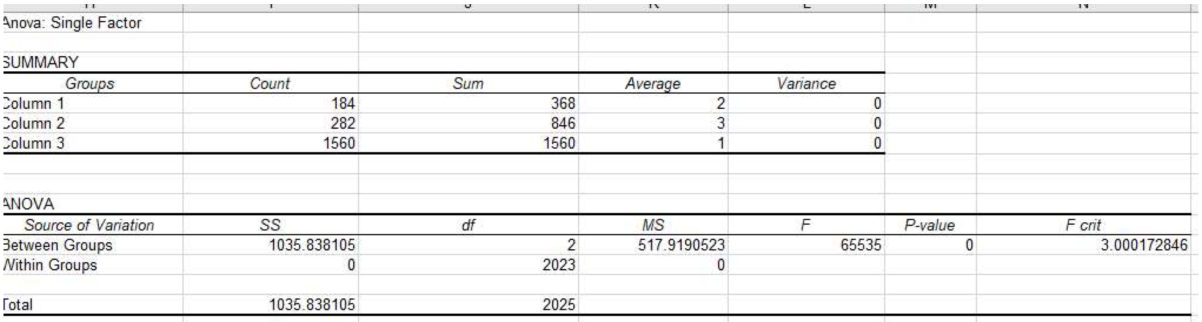
-TV-hours worked are independent for the three classes, So we use ANOVA to test the above null hypothesis.

P value <.05 so we reject null hypothesis and infer that our data provides enough statistical evidence to conclude that differences in number of TV-hours worked (TVHOURS) exist between the three classes.

```
F_onewayResult(statistic=14.961160297459875, p value=3.772788921510866e-
07)
```



Q9: Do the data provide enough statistical evidence to conclude that differences in number of hours worked (HRS) exist between the three races (RACE)?



Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
Column 1	184	368	2	0
Column 2	282	846	3	0
Column 3	1560	1560	1	0

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1035.838105	2	517.9190523	65535	0	3.000172846
Within Groups	0	2023	0			
Total	1035.838105	2025				

H0=U1=U2=U3

H1: one mean at least differ

-hours worked are independent for the three races, So we use ANOVA to test the above null hypothesis.

P value <.05 so we reject null hypothesis and infer that our data provides enough statistical evidence to conclude that differences in number of hours worked (HRS) exist between the three races (RACE)

Q10: Is there sufficient evidence to conclude that less than 50% of Americans support gun laws (GUNLAW)?
Yes , as we reject null hypothesis
z_stat: 25.686, p_value: 0.000
Reject the null hypothesis - suggest the alternative hypothesis is true.
So there is evidence that less than 50% of Americans
support gun laws (GUNLAW)

Q11: Can we infer from the data that Democrats and Republicans (PARTYID: 0, 1 = Democrat, 5, 6 = Republican) differ in their position on whether the government should reduce income differences between rich and poor (EQWLTH)?

```
z_stat: 7.566, p_value: 0.000
Reject the null hypothesis - suggest the alternative hypothesis is true.
```

There is enough evidence to conclude that Democrats and Republicans differ in their positions

Q12: How does income affect a person's response to the question, Should the government improve the living conditions of poor people (HELPPOOR)? Test the relationship between income (INCOME) and (HELPPOOR) to answer the question.

Column1	*INCOME*	*HELPPOOR*
INCOME	1	
HELPPOOR	**0.218136707**	1

Here correlation = 0.22 which is weak correlation so the government can improve the living conditions of poor people and still search about other factors.

Q13: It seems reasonable to assume that the more one works, the greater the income. Test this assumption by analyzing the relationship between hours worked per week (HRS) and income (INCOME).

We test the hypotheses of the regression model.

Let the independent variable be HRS and the dependent variable be INCOME.

Null hypothesis: $H_0 : \beta_1 = 0$

That is, there is no linear relationship between INCOME and HRS.

Alternative hypothesis: $H_1 : \beta_1 \neq 0$

That is, there is positive relationship between INCOME and HRS.

Rejection rule:If *p*-value is lesser than the level of significance $(\alpha = 0.05)$, then reject the null hypothesis (H_0).

	INCOME	*HRS*
INCOME	**1**	
HRS	**0.018145**	**1**

Q14: Use the General Social Survey of 2008 to conduct a regression analysis of

income (INCOME) using the following dependent variables:

Age (AGE)Years of education (EDUC) Hours of work per week (HRS), Spouse's hours of work (SPHRS), Occupation prestige score (PRESTG80),Number of children (CHILDS)

Number of family members earning money (EARNRS), Years with current employer (CUREMPYR)

a. **Test the model's validity**

F	*Significance F*
16.10843	1.19772E-19

P value ~=0 so that the model is valid

b. **Test each of the slope coefficients**

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-81019.62897	20641.08	-3.92516	0.000108
AGE	804.6921138	256.1326	3.141701	0.001855
EDUC	5836.283404	915.8009	6.372874	7.34E-10
HRS	-24.43943272	183.7107	-0.13303	**0.894261**
PRESTG80	485.541341	187.8375	2.584902	0.010233
SPHRS	183.397205	189.5318	0.967633	**0.33404**
CHILDS	-1825.759269	1657.797	-1.10132	0.271679
EARNRS	10263.08888	3208.921	3.198299	0.001537
CUREMPYR	502.5566574	257.9544	1.948239	0.052358

c. **Interpret the coefficients.**

Most of independent values with p value less than 0.05 except hours per week and Spouse's hours of work (SPHRS) so we should delete this variable from our regression as they didn't obey regression rule.