Formal Analysis

Statistical Inference

Are females more intelligent than males?

Introduction

The big question that drove my interest is: Are females more intelligent than males or vice versa? In this report, I tried to capture part of the question by exploring whether there is a difference in the average GPAs of females and males. I estimated the average GPA of females and males by constructing a confidence interval for both. I also performed a test for statistical and practical significance to determine whether gender is associated with GPAs of students or not. In other words, is there convincing evidence that gender impacts the GPA of students?

Dataset

I obtained this dataset from Kaggle datasets (BoraPajo, 2016) . There is a total of 73 females respondents and 48 males respondents. They were asked to give their GPA, their gender, food habits calories per day and income. In the original dataset, they had column for gender response and another for GPA, but for the sake of the study, I modified the columns to have one for female's GPA and the other for male's GPA to make it easier to use it for coding and calculations. The specific data that I will be using in this report can be found here.

I am interested in using this sample data to estimate the average GPA of females and males and examine whether there is a difference in GPA with gender. The variables of interest are gender and GPA. Gender is the independent categorical and dimensionless variable while the GPA is the quantitative dependent variable. I am interested in examining whether the GPA is dependent on the gender and can be changed according to gender differences. [1]

[1] #**variables**: identify variables included in the study with an explanation of the relation between dependent and independent variables.

Methods

I used pandas' package and matplotlib to read and analyze the dataset in python. First, python was used to examine the descriptive statistics for each variable of interest. Table 1. Provides the summary statistics for the GPA of females and the GPA of males (they are computed in Appendix A).

The sample distributions for females and male GPAs are displayed in Figures 1 and 2 (these are created in Appendix B).

Table 1: Summary statistics for GPA by genders

	Female GPAs	Male GPAs
Count	$n = 73$	$n = 48$
Mean	$\bar{x}_f = 3.434$	$\bar{x}_m = 3.388$
Median	3.5	3.4
Mode	3 and 3.5 (both repeated 8 times)	3.3 and 3.5 (both repeated 5 times)
Standard Deviation	$s_f = 0.377$	$s_m = 0.409$
Range	4-2.25 = 1.75	4-2.20 = 1.8

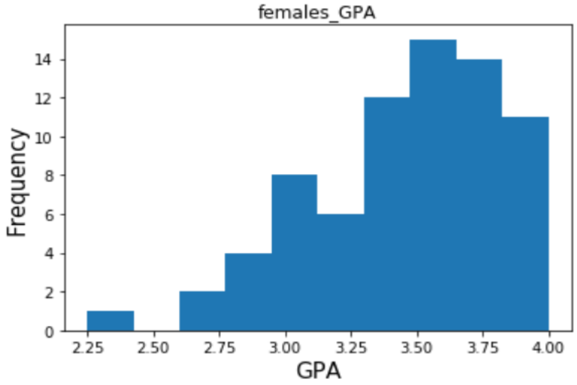


Figure 1: Histogram for Females GPAs

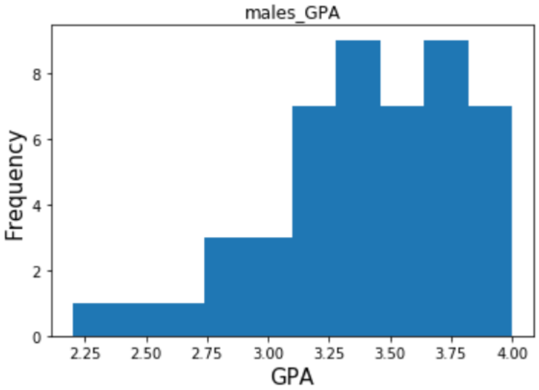


Figure 2: Histogram for Males GPAs

By looking at the mean and median values in Table 1, we find that in both females and males GPA the mean is less than the median value resulting in negative skewness which agrees with what is seen in the two histograms showing that data more concentrated to the right of the histograms (students sampled tend to have GPA higher than the median regardless of the gender). The standard deviation of the male's GPA is slightly larger than that of the females which corresponds to more spread of data in the second one.[2][3]

Using confidence interval, we can get an estimate of the range of values that is likely to contain the unknown parameter of the population using the sample data that we have. Particularly in this case, we set confidence level to 95% to get a plausible range of values that is likely to contain the average GPA of males and females. To ensure the precision of the confidence interval, we need to have normal approximation. So, to ensure that sampling distribution is nearly normal, and the estimate of the standard error is sufficiently accurate, we should check that the following conditions are satisfied before proceeding:

- Independence of observations: We check this by noting that the observations consist of fewer than 10% of the population, but we must also assume that the observations are randomly chosen.

[2] **#descriptivestats**: provide a table that has the summary of the descriptive statistics including mean, median, mode, standard deviation and range. Then made conclusion based on this information.

[3] **#dataviz**: creating a data visualization using python that shows the distribution of females and males GPAs.

- The sample size is large: We have n = 73 > 30/ n = 48 > 30 which meets our rule of thumb.
- The population distribution is not strongly skewed: We do not have information about the population distribution, but by looking at the sample distribution in Figure1and 2, we see that there are no prominent outliers. Thus, we can assume that this condition is met.

Since the three conditions are met, the central limit theorem ensures that the distribution of the sample mean is approximated by a normal model.[4]

To compute the confidence interval for the mean females and males' GPAs we use the general formula: [point estimate $\pm$z*SE]. The interval values for females GPA for a 95% confidence is [3.347, 3.521], the interval for males GPA is [3.271, 3.505] .We compute the standard error using the formula: $SE =$standard deviation /$\sqrt{sample\ size}$. Bessel's correction is used to reduce the bias and make more accurate estimation. The full calculation can be found in Appendix C.[5]

The research question is: Is there a difference in the average GPA between Females and Males students? To address this question, we calculate significance statistics test. The significance level chosen is $\alpha = 0.05$because we have no preference on whether to reject the null hypothesis or not, in other words, we are not worried about either type of errors.

Even though we have sample size more than 30, we will use t-distribution because we are given sample data not population data. We also check that each sample meets the conditions of using t-distribution:

1-independence: observations consist of fewer than 10% of the population, but we must also assume that the observations are randomly chosen.

2-normality/skewness: justified above in the confidence interval part, we have relatively large sample size >30.

[4] **#distributions**: identifying the sampled distribution of the variables and discussed the conditions that will allow us to use the central limit theorem that allows the approximation of the distribution into normal model.

[5] **#confidenceintervals**: applying thorough calculations of the confidence interval, and clearly interpret its meaning.

Therefore, we are justified in proceeding with the t-distribution for inference and will conduct difference between means.

The hypotheses are:

- Null hypothesis: $\bar{x}_f - \bar{x}_m = 0$ (there is no difference between females and males average GPAs)
- Alternative hypothesis: $\bar{x}_f - \bar{x}_m \neq 0$(there is a difference between females and males' average GPAs; either males or females is higher or lower than the other. Thus, the test is 2 tailed because we are looking for a difference in either direction)

To assess statistical significance, we first compute the T-score using the usual formula for a difference of means test, $T = \frac{point\ estimate - null\ value}{standard\ error}$ with SE=sqrt(s1**2/n1 + s2**2/n2). We calculate the degrees of freedom of the t-distribution as df =n-1 , where n is the least value between the two sample sizes and in this case it would be 48-1=47, Then the t-score is converted into p-value. Finally, we compare between p-value and significance level. (See Appendix D for the calculation in Python).

To assess practical significance, we need a measure of effect size. Here, we choose Hedge's g as our measure because the bias is reduced by using Hedge's g particularly with smaller sample sizes. Computing Hedge's g requires the pooled standard deviation for which the formula is sqrt((s1**2*(n1-1) + s2**2*(n2-1))/(n1+n2-2)) See Appendix D contains the calculation of Cohen's d=(x2 - x1)/SDpooled which is then use to compute Hedge's g= Cohensd*(1-(3/4*(n1+n2)-9)).[6]

[6] **#significance**: clearly identify significance level, tails, calculation of t and p values, beside the statistical and practical significance with well-justified interpretation of them.

Results and Conclusions

The 95% confidence interval for the females GPA is [3.347, 3.521] , while the interval for males GPA is [3.271 , 3.505](outputted in Appendix C), which provides a plausible range of values for females and males average GPA. We are about 95% confident that the average GPA for the females population will fall between [3.347, 3.521] and 95% confident that the average GPA of males population will be larger than 3.271 but less than 3.505. which means that if repeated sample were taken 95% of the interval would contain the population mean of the GPA.

In Appendix D, we see the results for the test of statistical and practical significance. The T-score of 0.4981 results in a two-tailed p-value of 0.6207 >0.05. Thus, we conclude that we fail to reject the null hypothesis which means the claim of the null hypothesis is valid, thus it is valid that there is no difference between the average GPAs of females and males. Additionally, Cohen's g is -0.094., the negative sign indicates tells the direction of the effect, and according to the rule of thumb the 0.0942< 0.2 , we conclude that the effect size is small which indicates trivial difference between the two groups.

These conclusions are inductive because we inferred that there is no difference or just trivial between the average GPA of females and males population in general based on results from samples that are extremely small compared to the number of the population of the two group (inductive generalization).Yet it does not mean that we do not have evidence to support our reasoning and argument because we have evidence by calculating not only statistical but also practical significance; however, if there is more evidence discovered, there is a chance that hypothesis can be refuted. Furthermore, we can not only infer the difference between the intelligence of genders based on GPA only (one variable of many). [7]

[7] #**induction**: effectively analyze inductive reasoning with detailed explanation and identify its type.

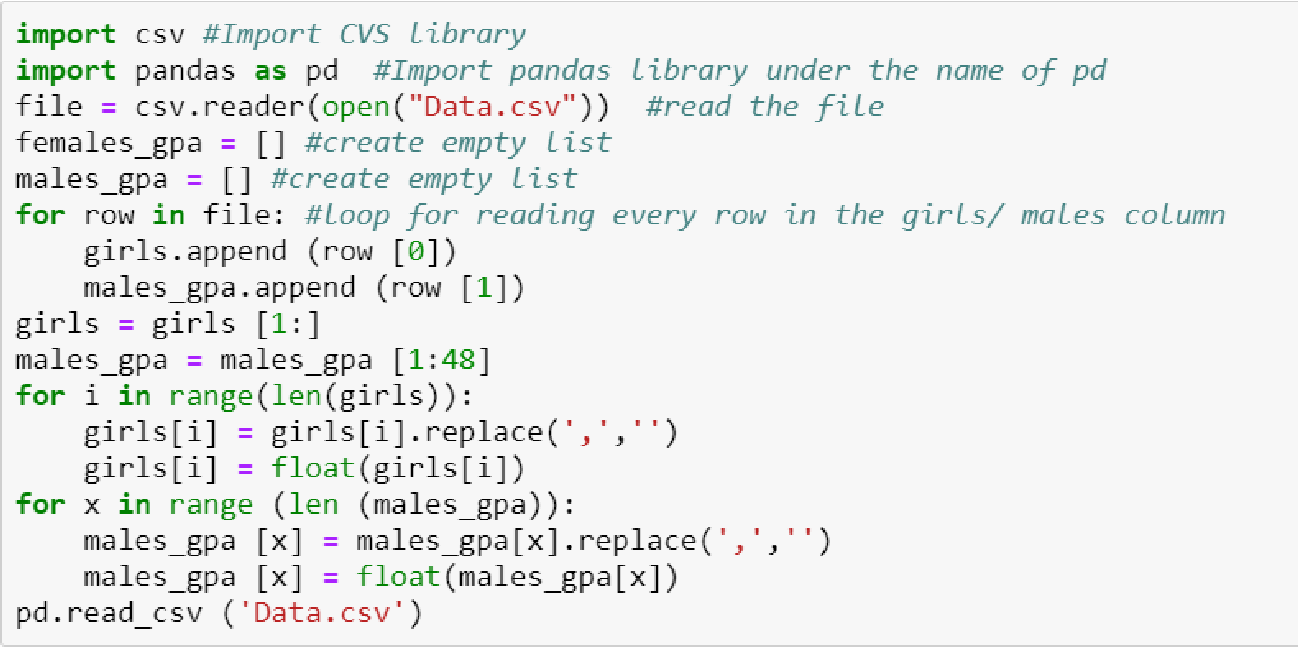
References

BoraPajo. (2016). Food choices: College students' food and cooking preferences. Retrieved from https://www.kaggle.com/borapajo/food-choices

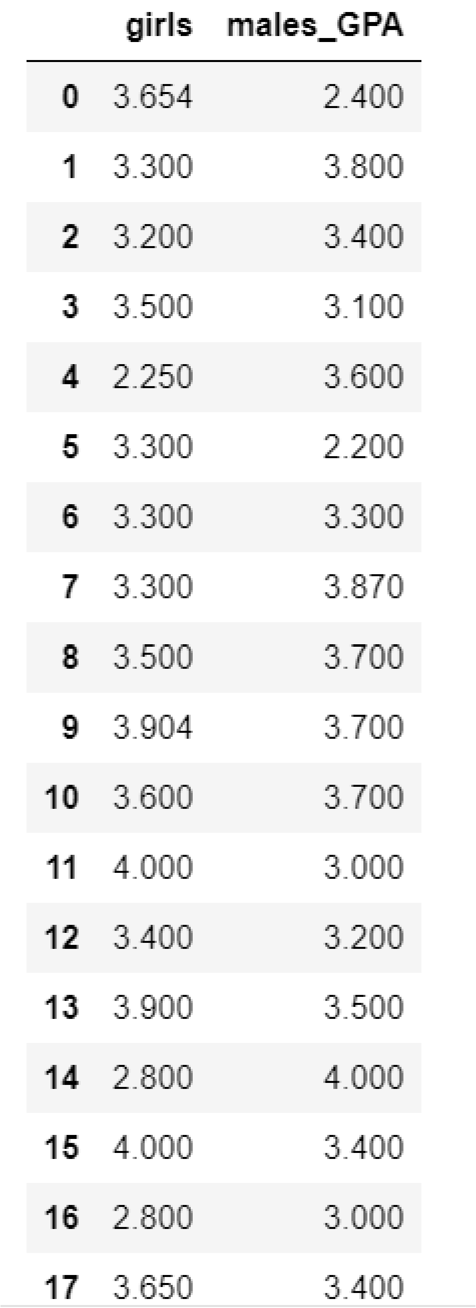
Appendix

The full Jupyter notebook file can be accessed here and the data can be found here. Here, the online calculator used can be found.

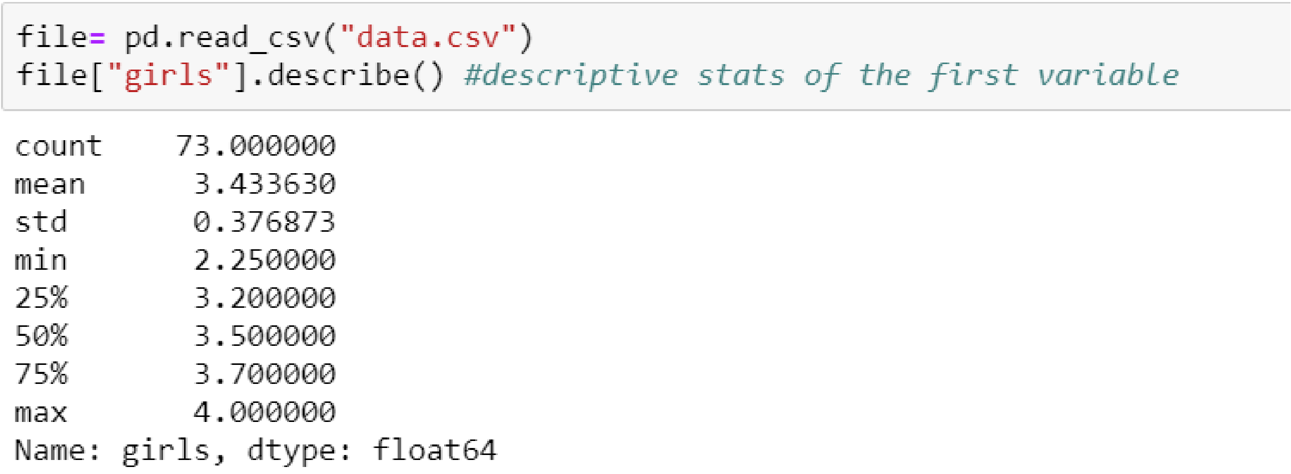
Appendix A: Import and Analyze Data



```
import csv #Import CVS library
import pandas as pd  #Import pandas library under the name of pd
file = csv.reader(open("Data.csv"))  #read the file
females_gpa = [] #create empty list
males_gpa = [] #create empty list
for row in file: #loop for reading every row in the girls/ males column
    girls.append (row [0])
    males_gpa.append (row [1])
girls = girls [1:]
males_gpa = males_gpa [1:48]
for i in range(len(girls)):
    girls[i] = girls[i].replace(',','')
    girls[i] = float(girls[i])
for x in range (len (males_gpa)):
    males_gpa [x] = males_gpa[x].replace(',','')
    males_gpa [x] = float(males_gpa[x])
pd.read_csv ('Data.csv')
```

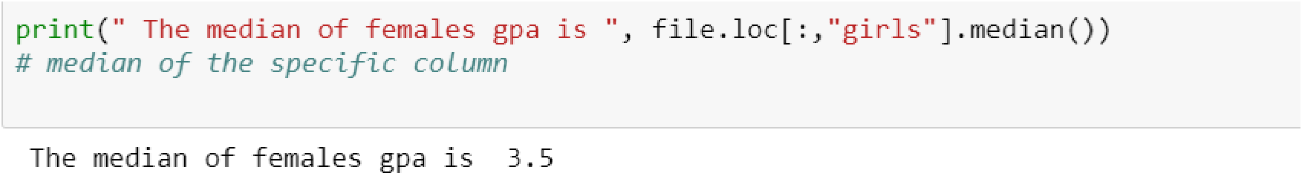


	girls	males_GPA
0	3.654	2.400
1	3.300	3.800
2	3.200	3.400
3	3.500	3.100
4	2.250	3.600
5	3.300	2.200
6	3.300	3.300
7	3.300	3.870
8	3.500	3.700
9	3.904	3.700
10	3.600	3.700
11	4.000	3.000
12	3.400	3.200
13	3.900	3.500
14	2.800	4.000
15	4.000	3.400
16	2.800	3.000
17	3.650	3.400



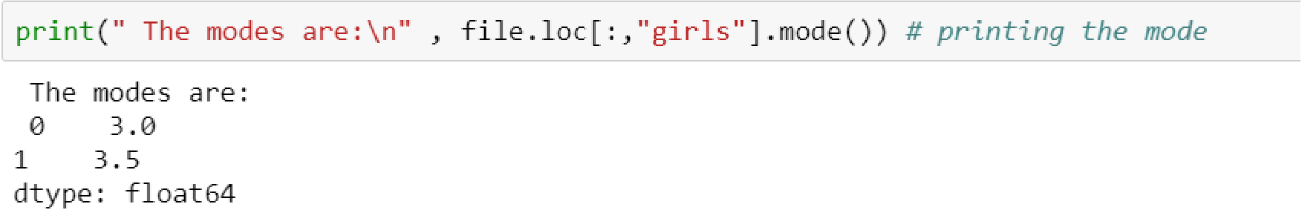
```
file= pd.read_csv("data.csv")
file["girls"].describe() #descriptive stats of the first variable
```

```
count    73.000000
mean      3.433630
std       0.376873
min       2.250000
25%       3.200000
50%       3.500000
75%       3.700000
max       4.000000
Name: girls, dtype: float64
```



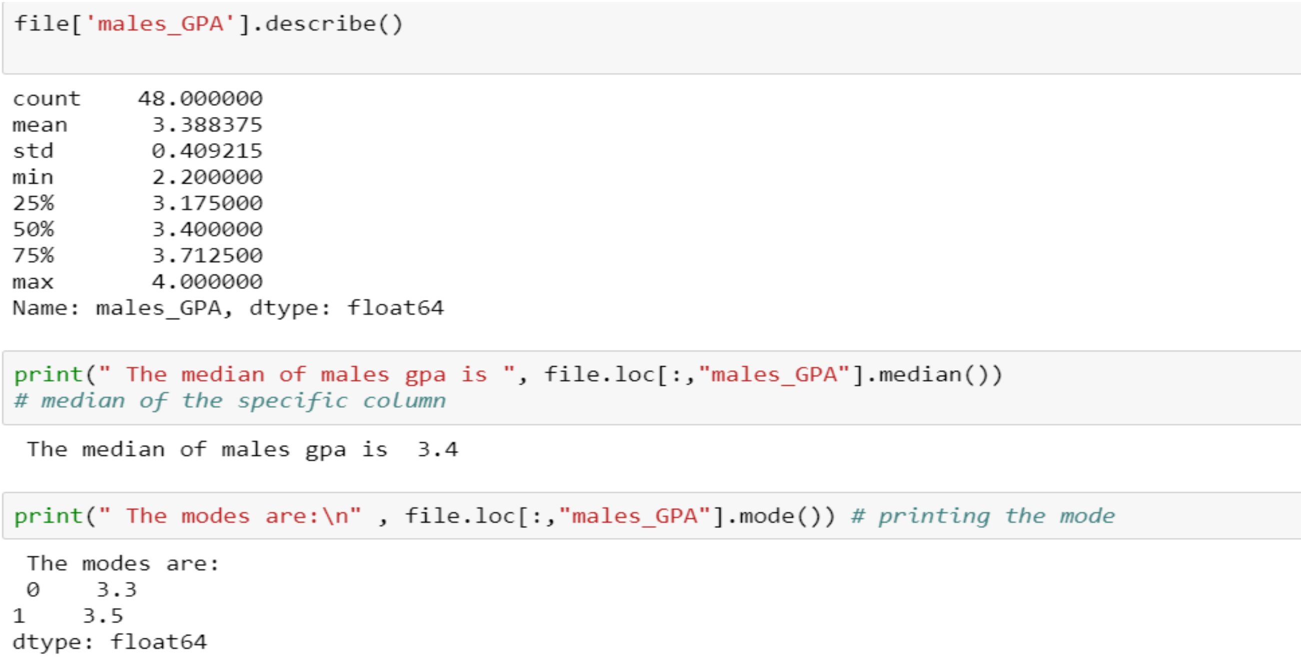
```
print(" The median of females gpa is ", file.loc[:,"girls"].median())
# median of the specific column
```

```
 The median of females gpa is  3.5
```



```
print(" The modes are:\n" , file.loc[:,"girls"].mode()) # printing the mode
```

```
 The modes are:
 0    3.0
1    3.5
dtype: float64
```



```
file['males_GPA'].describe()
```

```
count    48.000000
mean      3.388375
std       0.409215
min       2.200000
25%       3.175000
50%       3.400000
75%       3.712500
max       4.000000
Name: males_GPA, dtype: float64
```

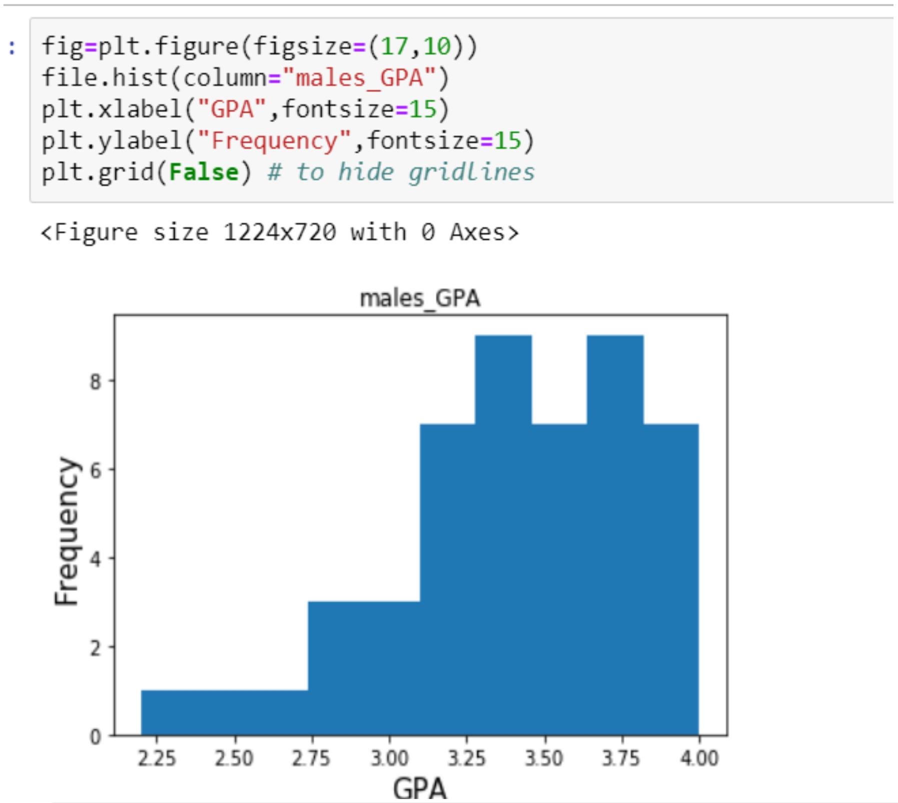
```
print(" The median of males gpa is ", file.loc[:,"males_GPA"].median())
# median of the specific column
```

```
 The median of males gpa is  3.4
```

```
print(" The modes are:\n" , file.loc[:,"males_GPA"].mode()) # printing the mode
```

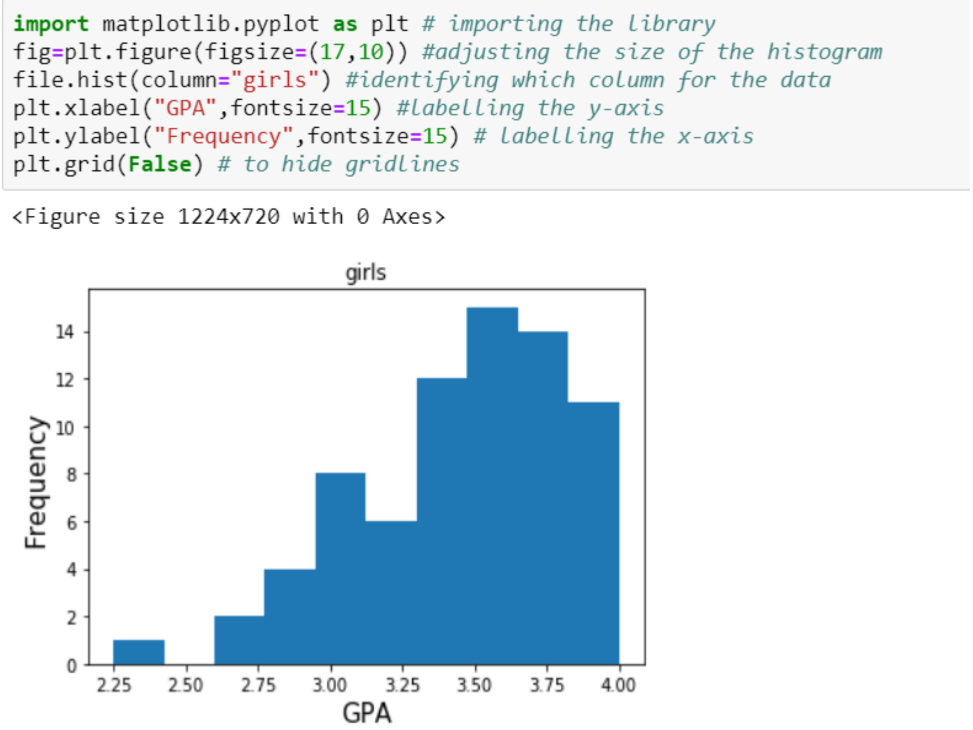
```
 The modes are:
 0    3.3
1    3.5
dtype: float64
```

Appendix B: Visualize Data



```
fig=plt.figure(figsize=(17,10))
file.hist(column="males_GPA")
plt.xlabel("GPA",fontsize=15)
plt.ylabel("Frequency",fontsize=15)
plt.grid(False) # to hide gridlines
```

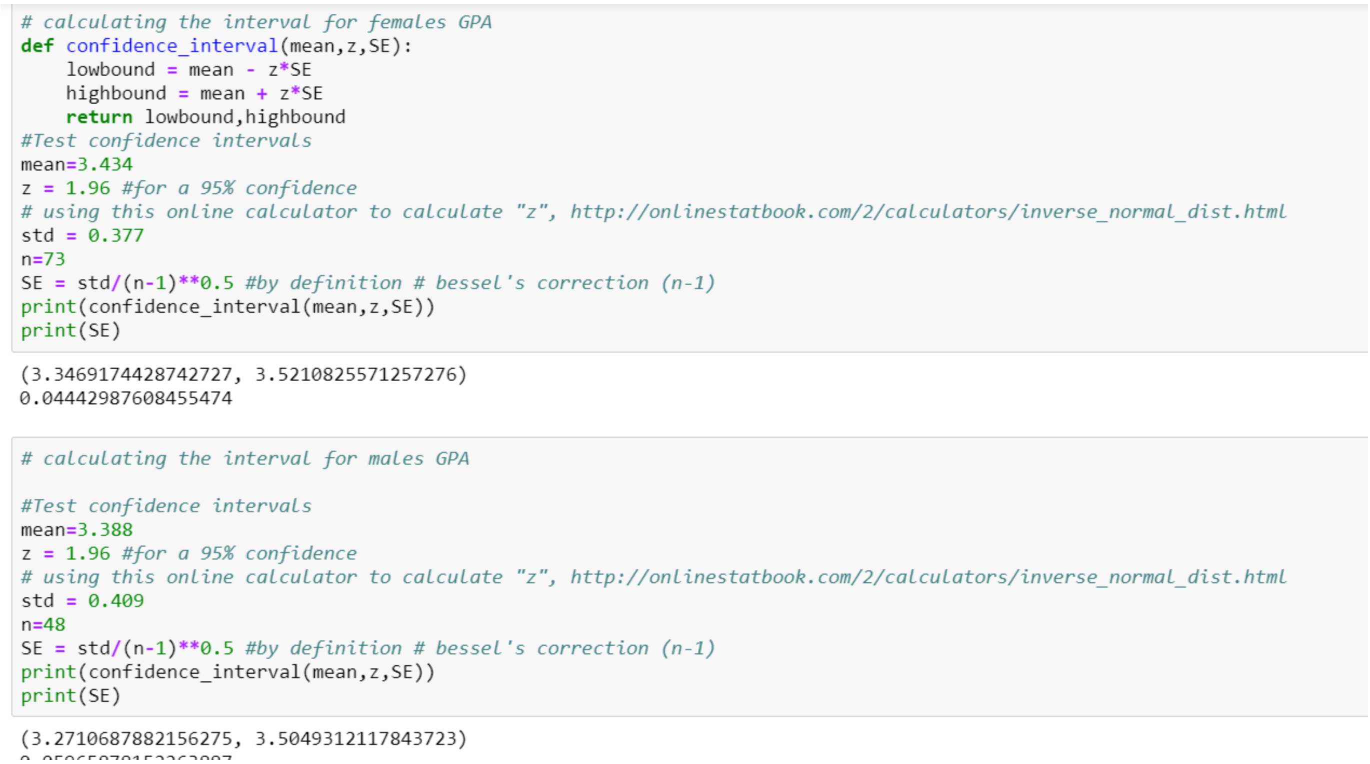
```
<Figure size 1224x720 with 0 Axes>
```



```
import matplotlib.pyplot as plt # importing the library
fig=plt.figure(figsize=(17,10)) #adjusting the size of the histogram
file.hist(column="girls") #identifying which column for the data
plt.xlabel("GPA",fontsize=15) #labelling the y-axis
plt.ylabel("Frequency",fontsize=15) # labelling the x-axis
plt.grid(False) # to hide gridlines
```

```
<Figure size 1224x720 with 0 Axes>
```

Appendix C: Confidence Interval



```
# calculating the interval for females GPA
def confidence_interval(mean,z,SE):
    lowbound = mean - z*SE
    highbound = mean + z*SE
    return lowbound,highbound
#Test confidence intervals
mean=3.434
z = 1.96 #for a 95% confidence
# using this online calculator to calculate "z", http://onlinestatbook.com/2/calculators/inverse_normal_dist.html
std = 0.377
n=73
SE = std/(n-1)**0.5 #by definition # bessel's correction (n-1)
print(confidence_interval(mean,z,SE))
print(SE)
```

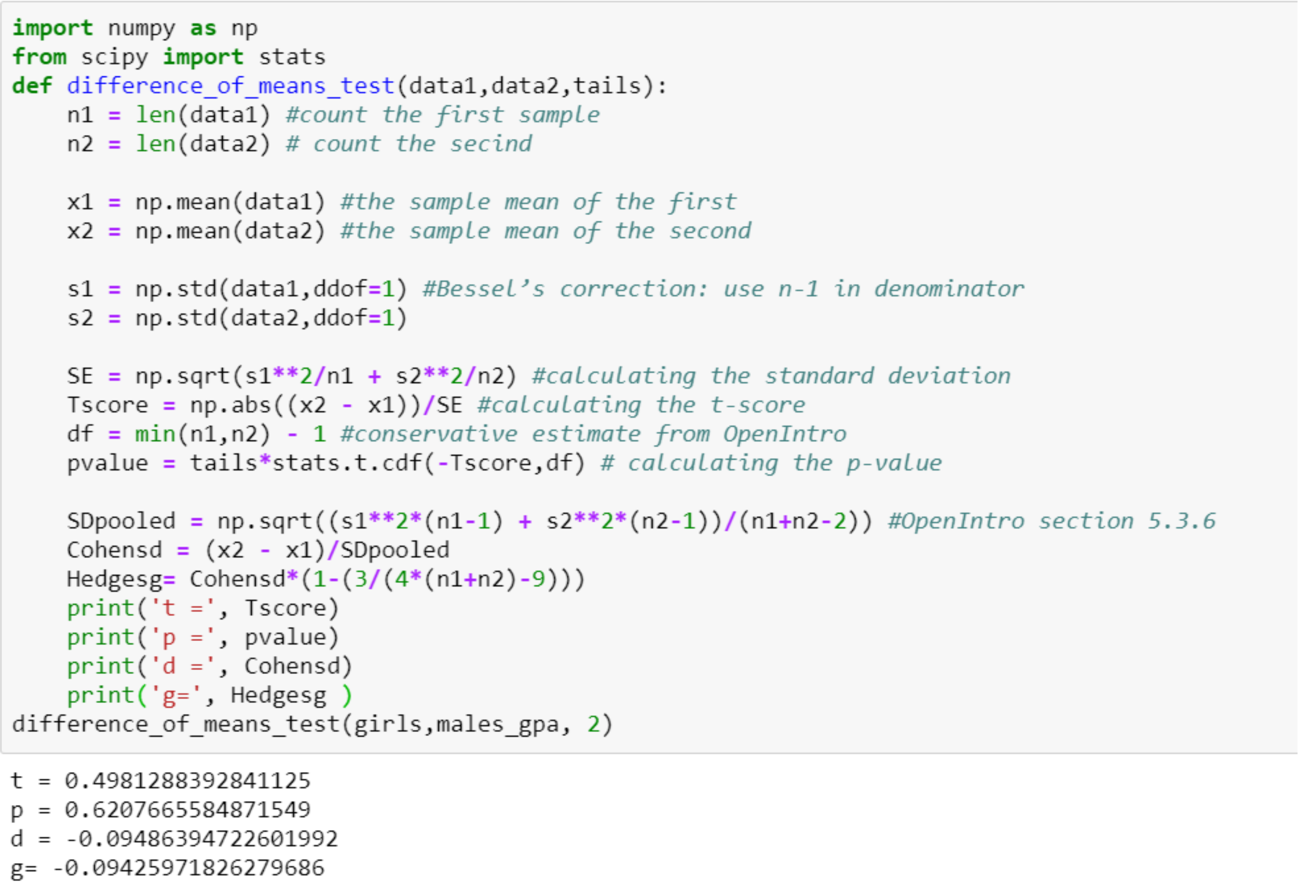
```
(3.3469174428742727, 3.5210825571257276)
0.04442987608455474
```

```
# calculating the interval for males GPA

#Test confidence intervals
mean=3.388
z = 1.96 #for a 95% confidence
# using this online calculator to calculate "z", http://onlinestatbook.com/2/calculators/inverse_normal_dist.html
std = 0.409
n=48
SE = std/(n-1)**0.5 #by definition # bessel's correction (n-1)
print(confidence_interval(mean,z,SE))
print(SE)
```

```
(3.2710687882156275, 3.5049312117843723)
0.05965878153363887
```

Appendix D: Difference of Means Test



```
import numpy as np
from scipy import stats
def difference_of_means_test(data1,data2,tails):
    n1 = len(data1) #count the first sample
    n2 = len(data2) # count the secind

    x1 = np.mean(data1) #the sample mean of the first
    x2 = np.mean(data2) #the sample mean of the second

    s1 = np.std(data1,ddof=1) #Bessel’s correction: use n-1 in denominator
    s2 = np.std(data2,ddof=1)

    SE = np.sqrt(s1**2/n1 + s2**2/n2) #calculating the standard deviation
    Tscore = np.abs((x2 - x1))/SE #calculating the t-score
    df = min(n1,n2) - 1 #conservative estimate from OpenIntro
    pvalue = tails*stats.t.cdf(-Tscore,df) # calculating the p-value

    SDpooled = np.sqrt((s1**2*(n1-1) + s2**2*(n2-1))/(n1+n2-2)) #OpenIntro section 5.3.6
    Cohensd = (x2 - x1)/SDpooled
    Hedgesg= Cohensd*(1-(3/(4*(n1+n2)-9)))
    print('t =', Tscore)
    print('p =', pvalue)
    print('d =', Cohensd)
    print('g=', Hedgesg )
difference_of_means_test(girls,males_gpa, 2)
```

```
t = 0.4981288392841125
p = 0.6207665584871549
d = -0.09486394722601992
g= -0.09425971826279686
```