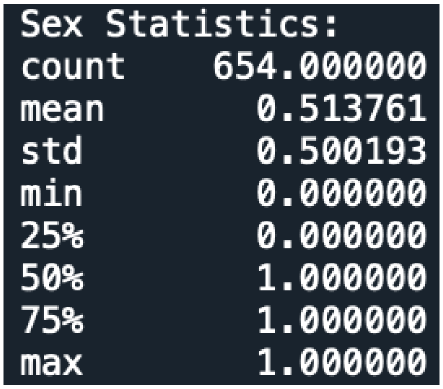
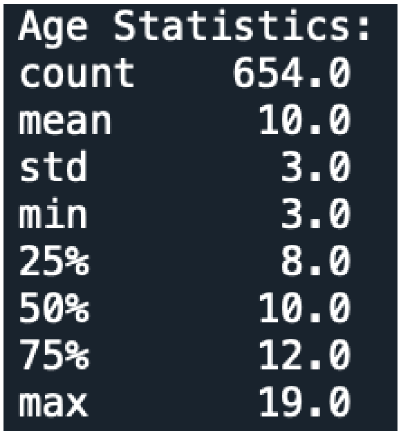
Project 2: Confidence Intervals
Lincoln Nordquist

Part 1:

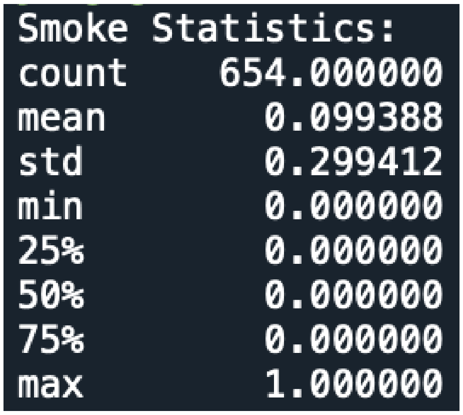
i) Descriptive statistics for each variable



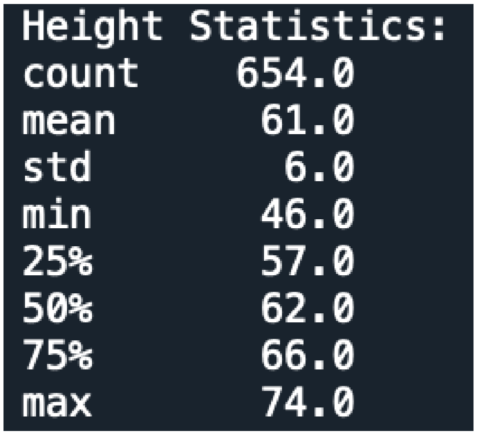
```
Sex Statistics:
count    654.000000
mean       0.513761
std        0.500193
min        0.000000
25%        0.000000
50%        1.000000
75%        1.000000
max        1.000000
```



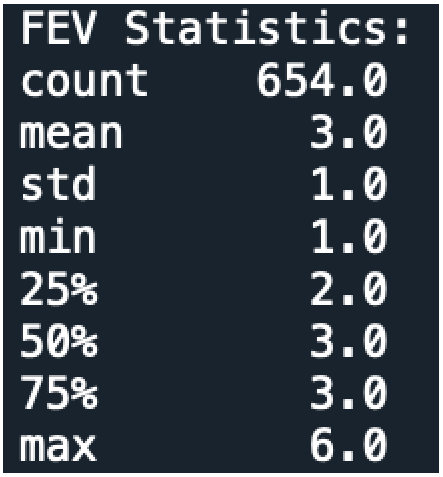
```
Age Statistics:
count    654.0
mean      10.0
std        3.0
min        3.0
25%        8.0
50%       10.0
75%       12.0
max       19.0
```



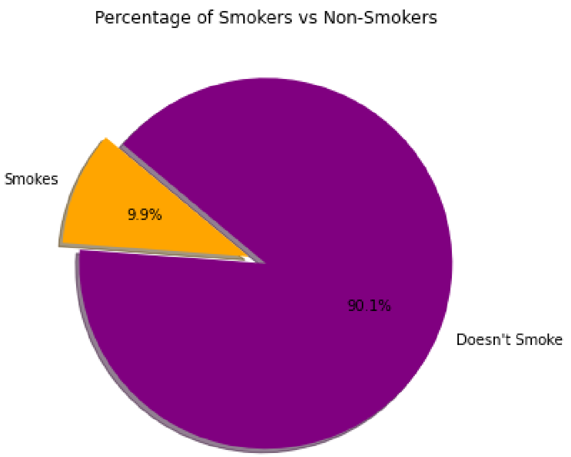
```
Smoke Statistics:
count    654.000000
mean       0.099388
std        0.299412
min        0.000000
25%        0.000000
50%        0.000000
75%        0.000000
max        1.000000
```

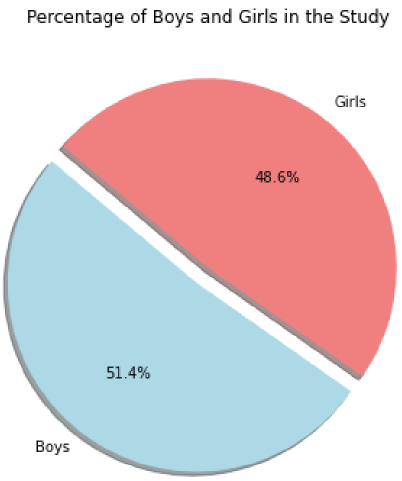


```
Height Statistics:
count    654.0
mean      61.0
std        6.0
min       46.0
25%       57.0
50%       62.0
75%       66.0
max       74.0
```



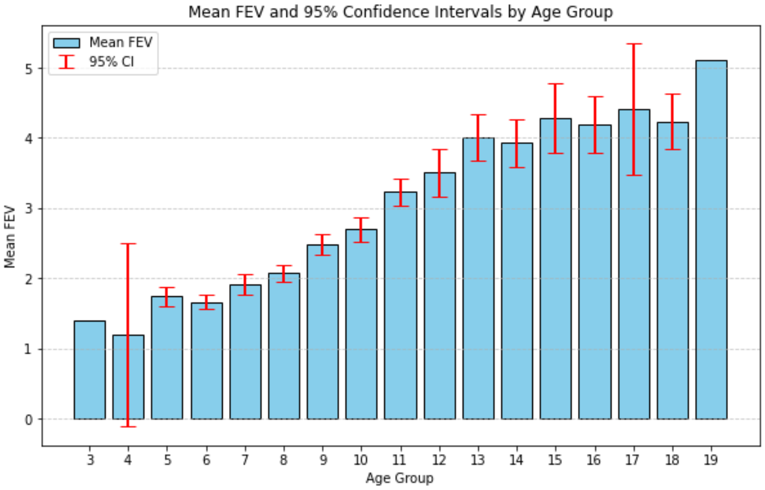
```
FEV Statistics:
count    654.0
mean       3.0
std        1.0
min        1.0
25%        2.0
50%        3.0
75%        3.0
max        6.0
```

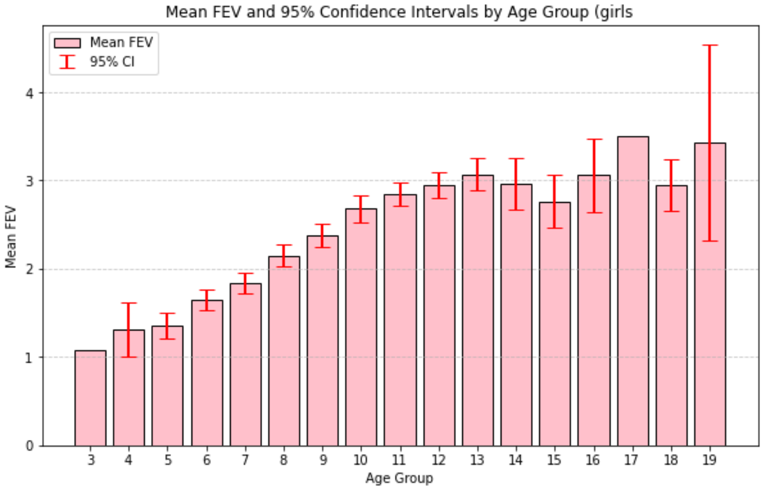




95% confidence intervals for mean of FEV to Age, Height, and Smoking Status (boys and girls)

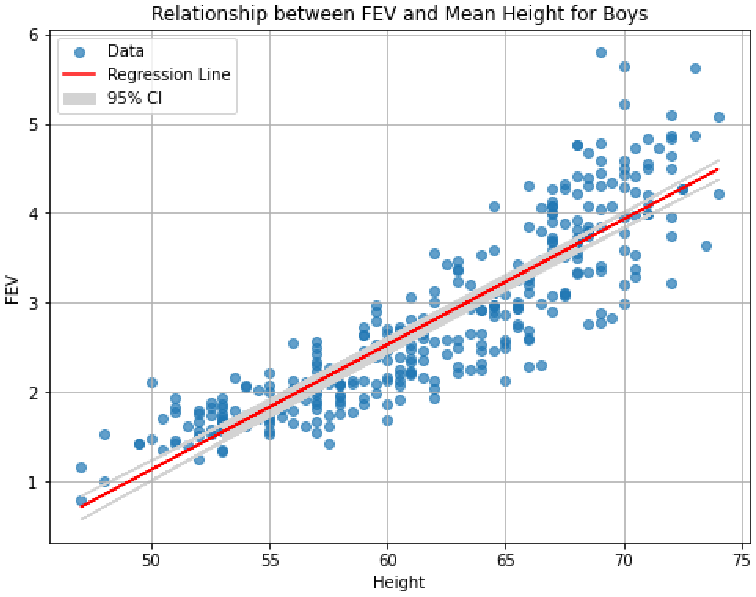
FEV to Age:

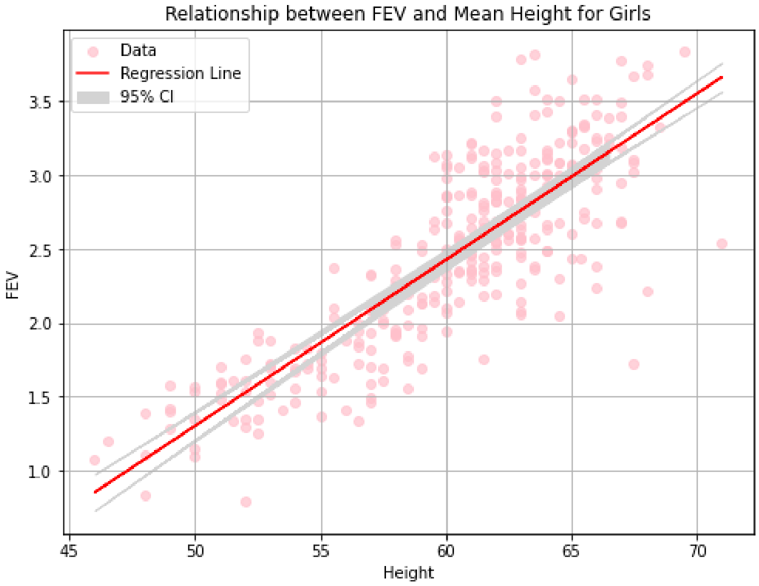




We can see a clear positive correlation between age and mean FEV in both graphs. The male data, however, seems to peak a bit higher than the female's. I have also included a visualization of the 95% confidence interval.

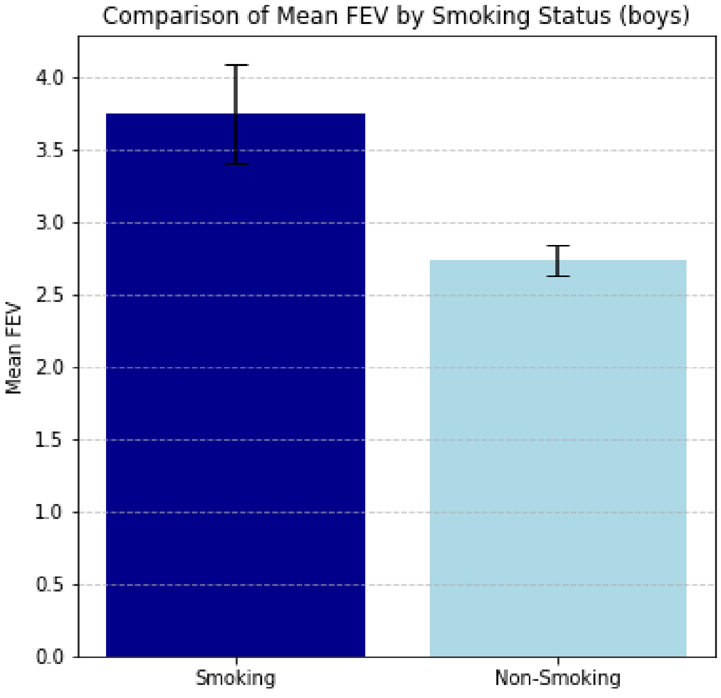
FEV to Height:

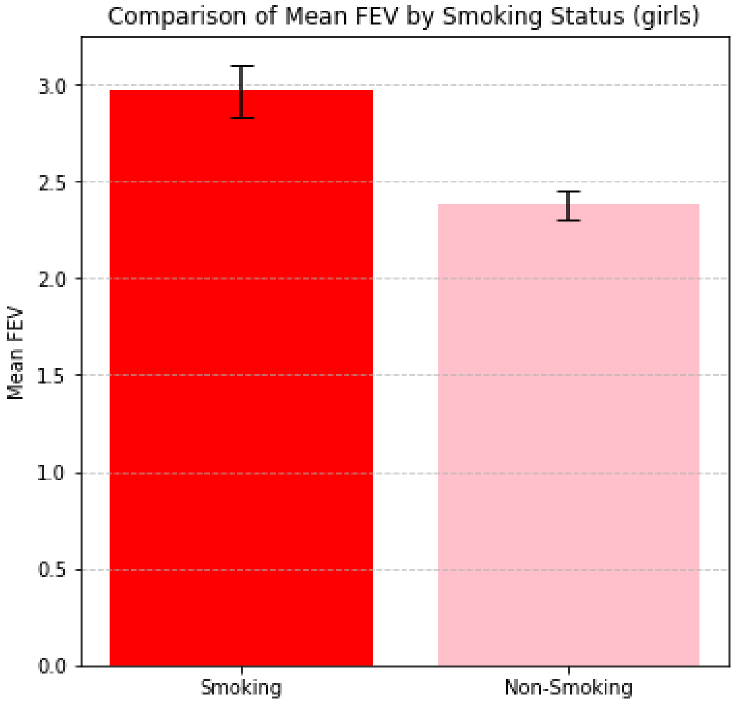




We can also see a positive correlation between height and FEV among boys and girls. This is to be expected, as people tend to get taller with age, and we already have seen a positive correlation between FEV and age. I decided to go with a scatter plot, since there are so many different heights to account for. I have included not only the 95% confidence interval, but the regression line as well for visualization.

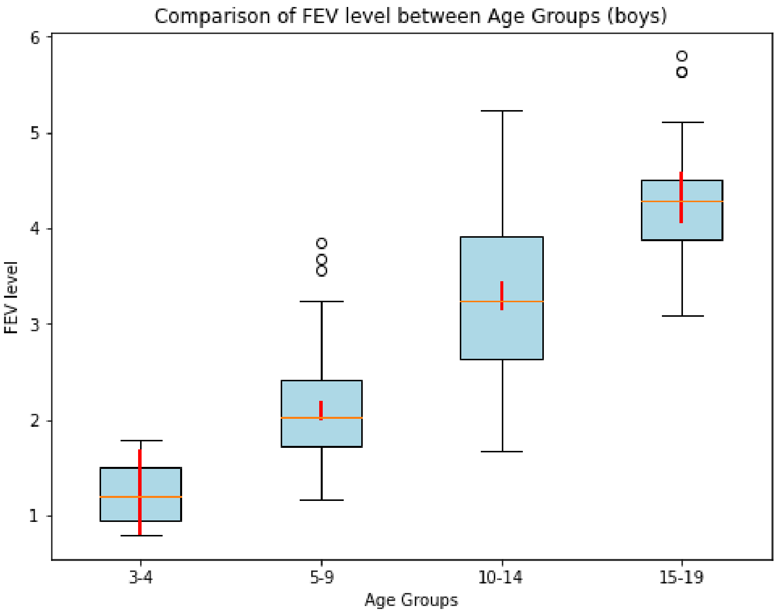
FEV to Smoke:

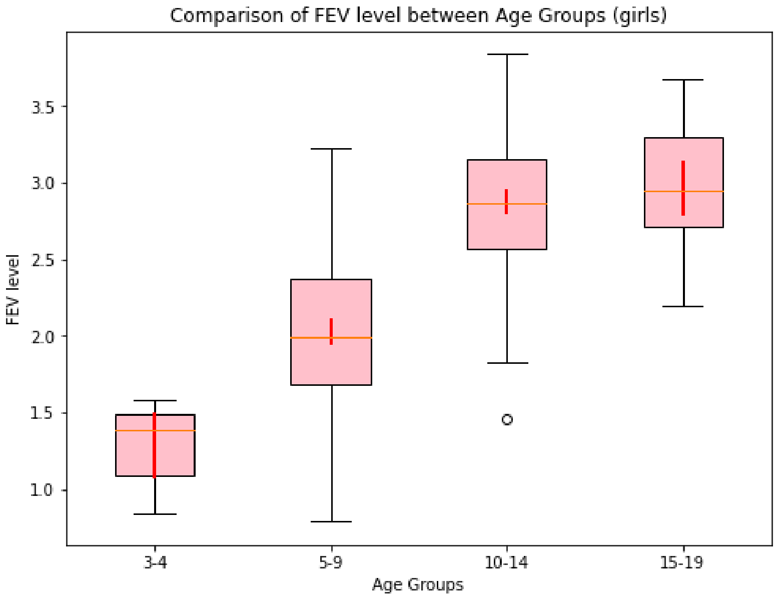




The two graphs above represent the comparison of mean FEV with smokers vs non-smokers. As we can see, the people who smoked had a higher FEV level. This could be due to other factors not being accounted for, such as smokers being older (on average) than non-smokers.

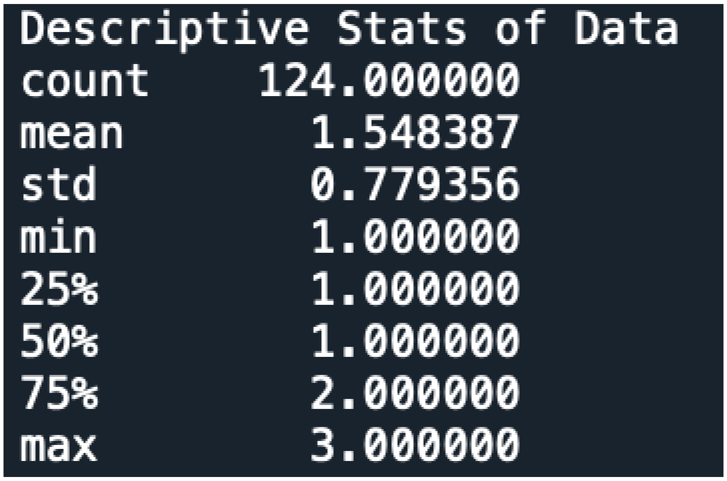
Patterns of Growth of FEV by Age Group (boys and girls)



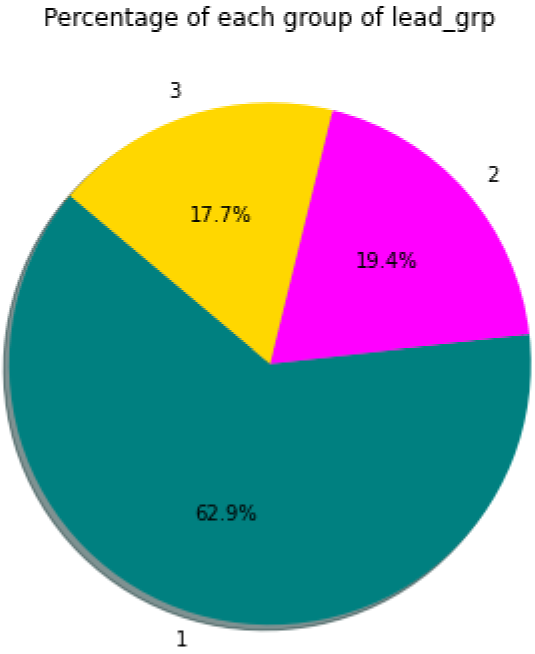


This is essentially the same graph that I used for project 1, except this time, we can see the confidence intervals. As we can see in both graphs, the range of the confidence intervals seems to change quite a bit, and is very inconsistent with each age group.

Part 2:

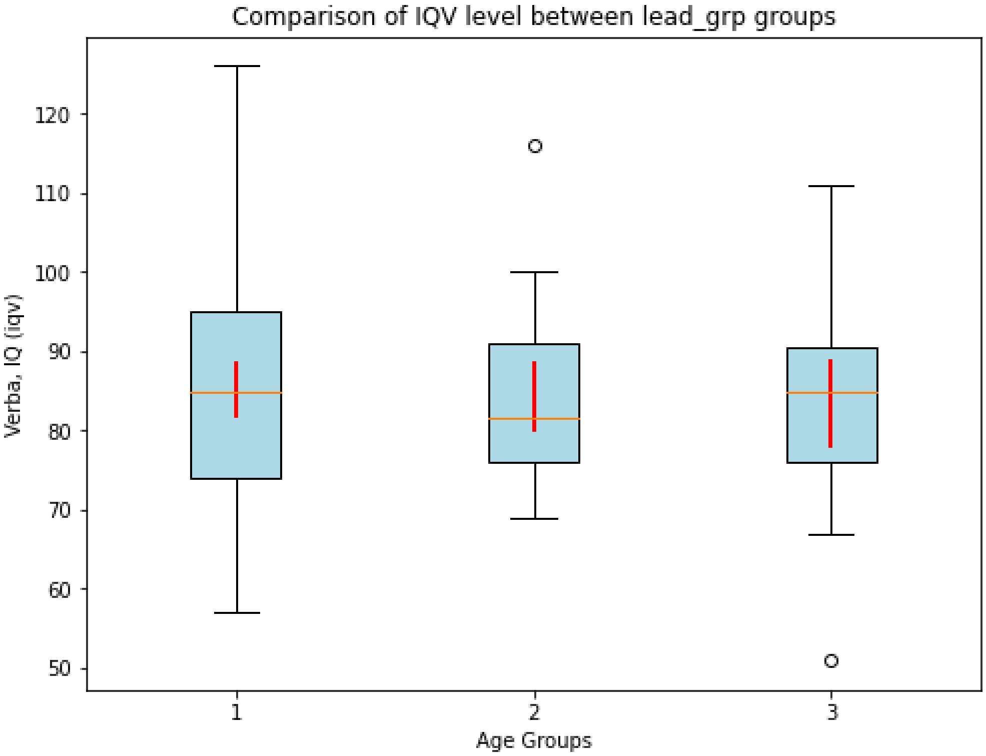


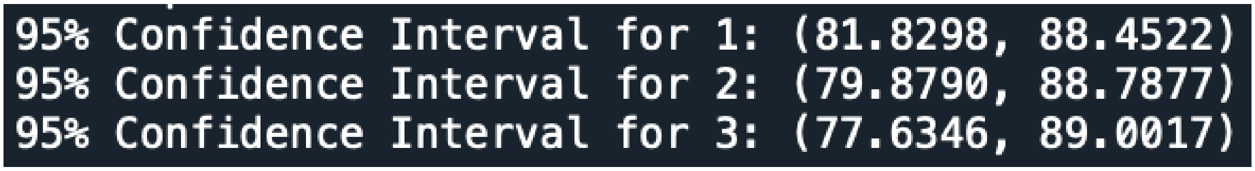
```
Descriptive Stats of Data
count    124.000000
mean       1.548387
std        0.779356
min        1.000000
25%        1.000000
50%        1.000000
75%        2.000000
max        3.000000
```



As we can see from the pie chart, an overwhelming majority of the population had a blood-lead level < 40 μg/100 mL in both 1972 and 1973.

Distribution of Verbal IQ (iq-v)

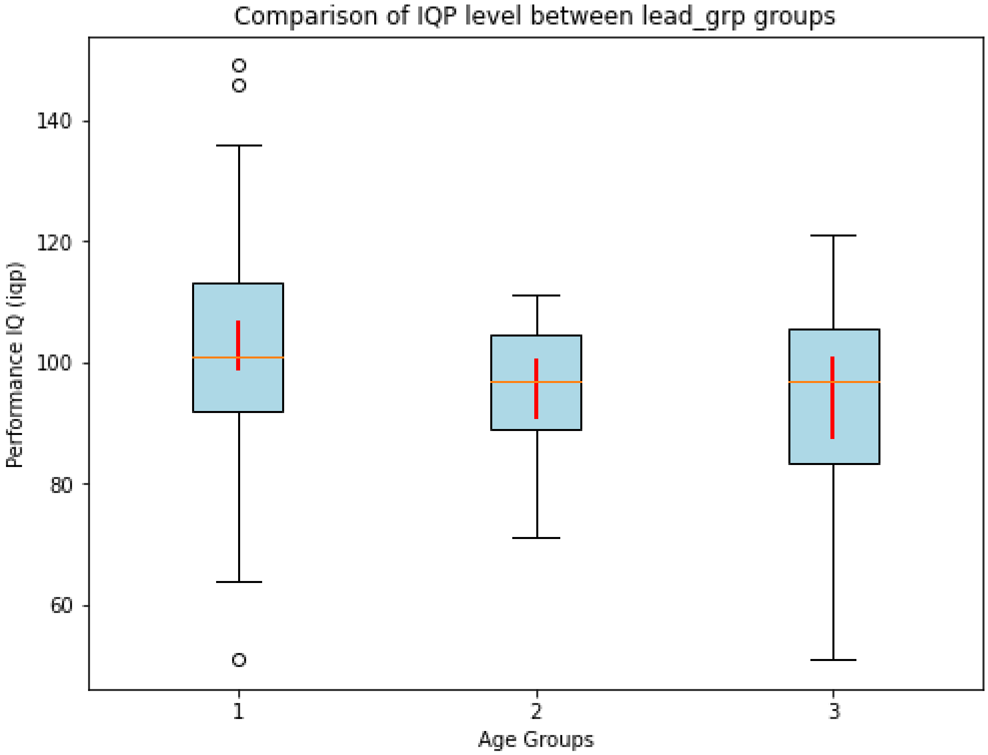


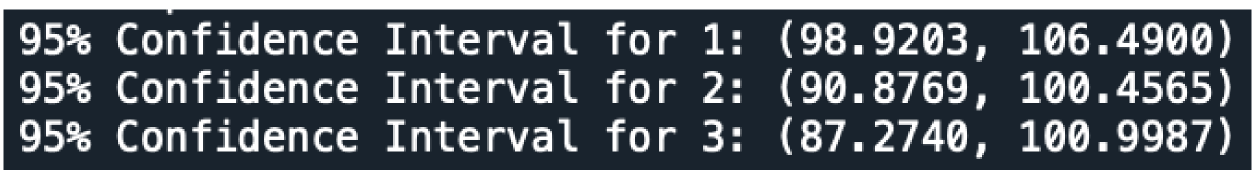


```
95% Confidence Interval for 1: (81.8298, 88.4522)
95% Confidence Interval for 2: (79.8790, 88.7877)
95% Confidence Interval for 3: (77.6346, 89.0017)
```

As we can see from the graph above, all three groups had very similar averages, standard deviations, and confidence intervals. It’s very interesting to see such similarities between the data sets. One thing to note is that group 1 has the shortest range of confidence interval, followed by group 2, and then group 3.

Distribution of Performance IQ (iq-p)

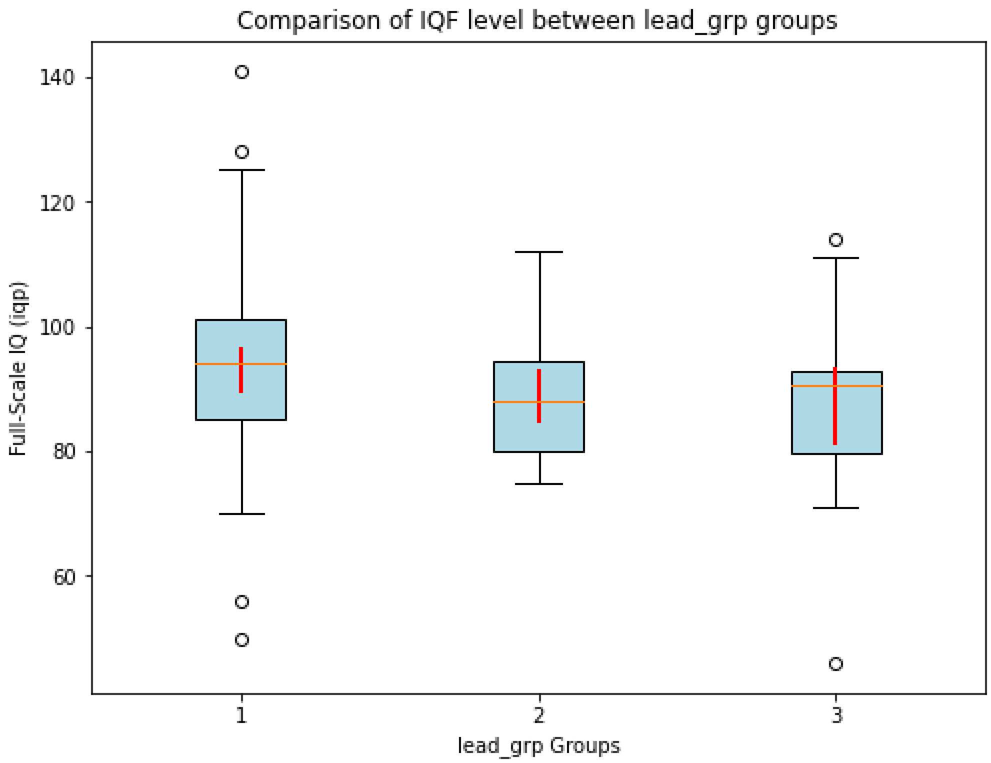


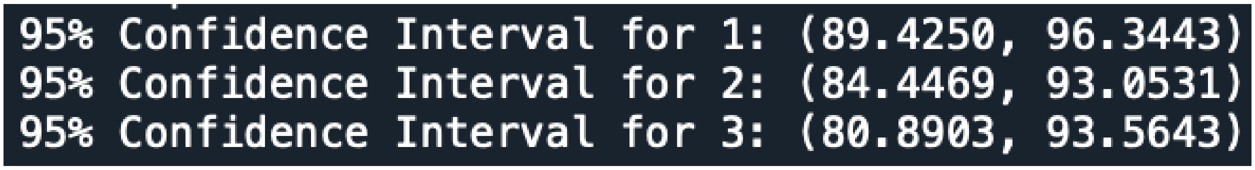


```
95% Confidence Interval for 1: (98.9203, 106.4900)
95% Confidence Interval for 2: (90.8769, 100.4565)
95% Confidence Interval for 3: (87.2740, 100.9987)
```

Similar to the iqv, these three data groups for performance IQ also have very similar sets of data. Group 1 seems to have the most outliers, and group 2 seems to have the smallest standard deviation.

Distribution of Full-Scale IQ (iq-f)



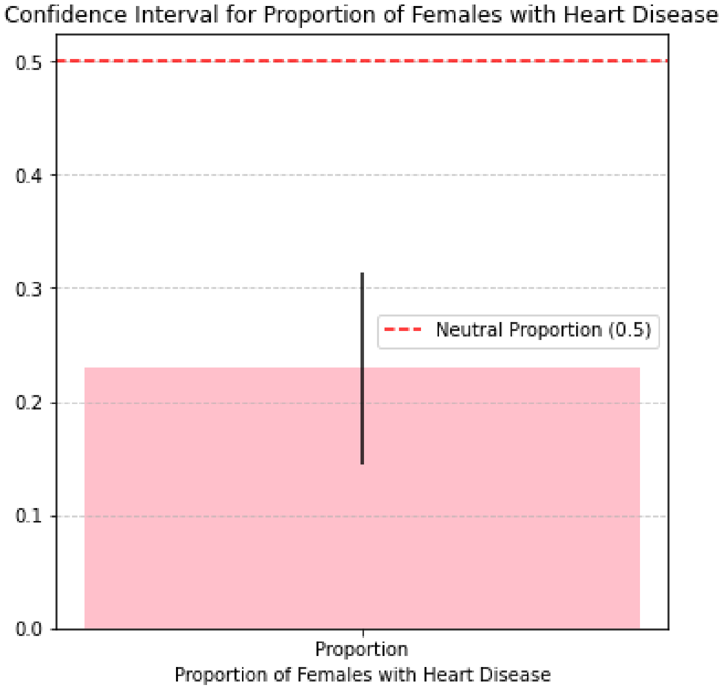


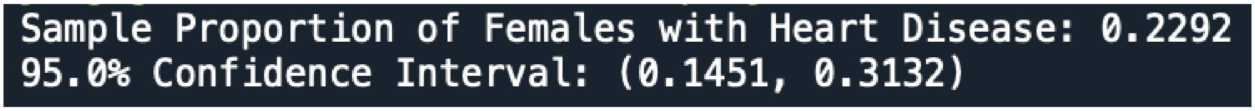
```
95% Confidence Interval for 1: (89.4250, 96.3443)
95% Confidence Interval for 2: (84.4469, 93.0531)
95% Confidence Interval for 3: (80.8903, 93.5643)
```

This set of data contains lots of outliers, especially with group 1. Despite these outliers, all three groups seem to have very similar means as well, along with similar confidence intervals.

Part 3:

Confidence Interval for female population with heart disease:

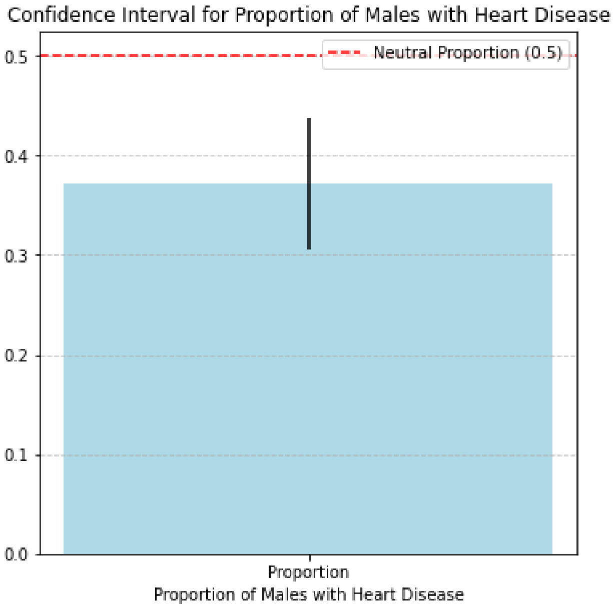


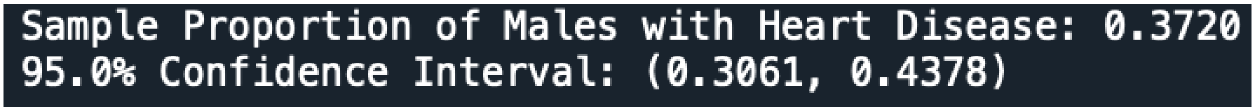


```
Sample Proportion of Females with Heart Disease: 0.2292
95.0% Confidence Interval: (0.1451, 0.3132)
```

The biggest thing I noticed right off the bat, is that the confidence interval is significantly lower than the neutral proportion. This indicates that the observed proportion is significantly different than the expected proportion.

Confidence Interval for male population with heart disease:





```
Sample Proportion of Males with Heart Disease: 0.3720
95.0% Confidence Interval: (0.3061, 0.4378)
```

The men's data is surprisingly much different than the women's data. As we can see, the men have a much higher confidence interval, indicating that the observed proportion is more similar to the expected proportion.