

OLYMPIC GAMES STATS PROJECT

Data Analysis Project Proposal

28/01/2022

Leonardo Saviane

An abstract composition of various geometric shapes. In the top left, a green-outlined triangle points right. To its right is a solid blue circle. Below the triangle is a blue-outlined circle. In the center is a large orange semi-circle. To the right of the semi-circle is a vertical yellow dashed line. In the bottom left is a large solid orange circle. To its right are four short, curved yellow dashed lines. In the bottom right is a green-outlined square.

- I want to retrieve some insights on the best athletes that the world have ever known, to understand whether some characteristic, such as height, weight, age and country, have some impact on achieving a medal in some sport.
- This can be useful for professional sport institutes who wants to push their athletes more towards some disciplines than another. Eg. If someone is tall maybe is more likely to perform better in sports like high jump or basket.

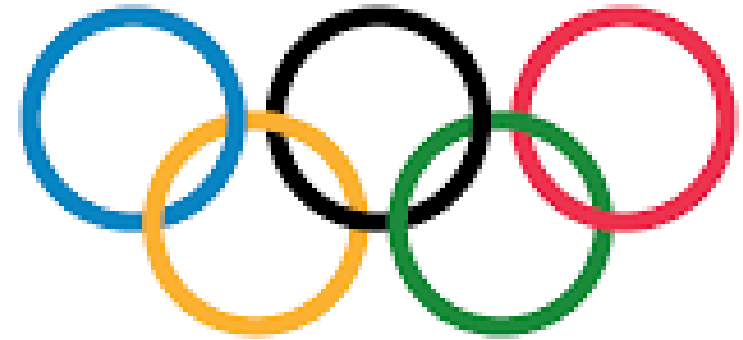
The Dataset and the cleaning

- The dataset used is SportsStats, which is a sports analysis firm partnering with local news and elite personal trainers to provide “interesting” insights to help their partners. Insights could be patterns/trends highlighting certain groups/events/countries, etc.
- Some variable such as “season” (winter and summer, “sex” (male or female) will be transformed in binary variables.
- “Medals” variable, so if the athletes has won a bronze/silver/gold medal, will be transformed in a binary variable as well. My purpose it’s to estimate the likelihood of winning a medal given the other factors.
- Some data are missing, such as height, weight and age for some athletes. Therefore, those value will be dropped to avoid biased estimate. Putting the average in the missing value would lead to possible biased estimate. Since the sample is large enough losing more or less 100000 observations is ok.
- **In the following slide the ERD and an anticipation of the dataset, then the questions I want to answer→**

ENTITY RELATIONAL DIAGRAM

Athletes_events
<ul style="list-style-type: none">• ID• Name• Sex• Age• Height• Weight• Team• NOC• Games• Year• Season• City• Sport• Event• Medal

NOC_regions
<ul style="list-style-type: none">• NOC• Region• notes



Short description of the table 'Athletes_events':

```
1  select
2      count(*) as N_Observations,
3      count(distinct name) as N_Atheltes,
4      count(distinct NOC) as N_Countries,
5      count(distinct Year) as N_Years,
6      count(distinct Sport) as N_Sports,
7      (select count(distinct sport) from athlete_events
8         where Season='Summer') as N_Summer_Sports,
9      (select count(distinct sport) from athlete_events
10         where Season='Winter') as N_Winter_Sports
11  from athlete_events
```

	N_Observations	N_Atheltes	N_Countries	N_Years	N_Sports	N_Summer_Sports	N_Winter_Sports
1	271116	134731	230	35	66	52	17

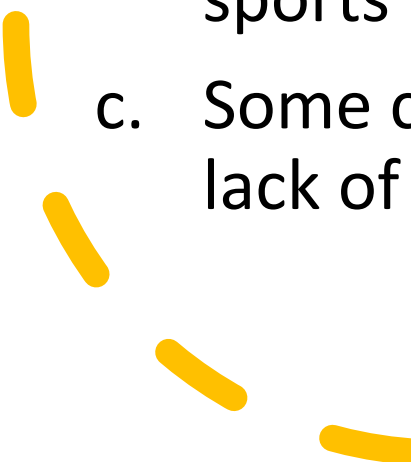
SECTION 1: QUESTIONS TO ANSWER

- Exists a correlation between medals won and country?
- Are Height and Age more relevant for some sports than other?
- Does these discovery change between summer sport and winter sports?



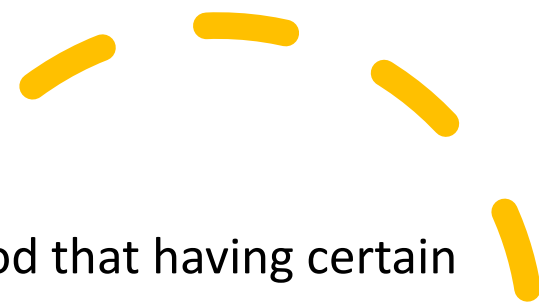
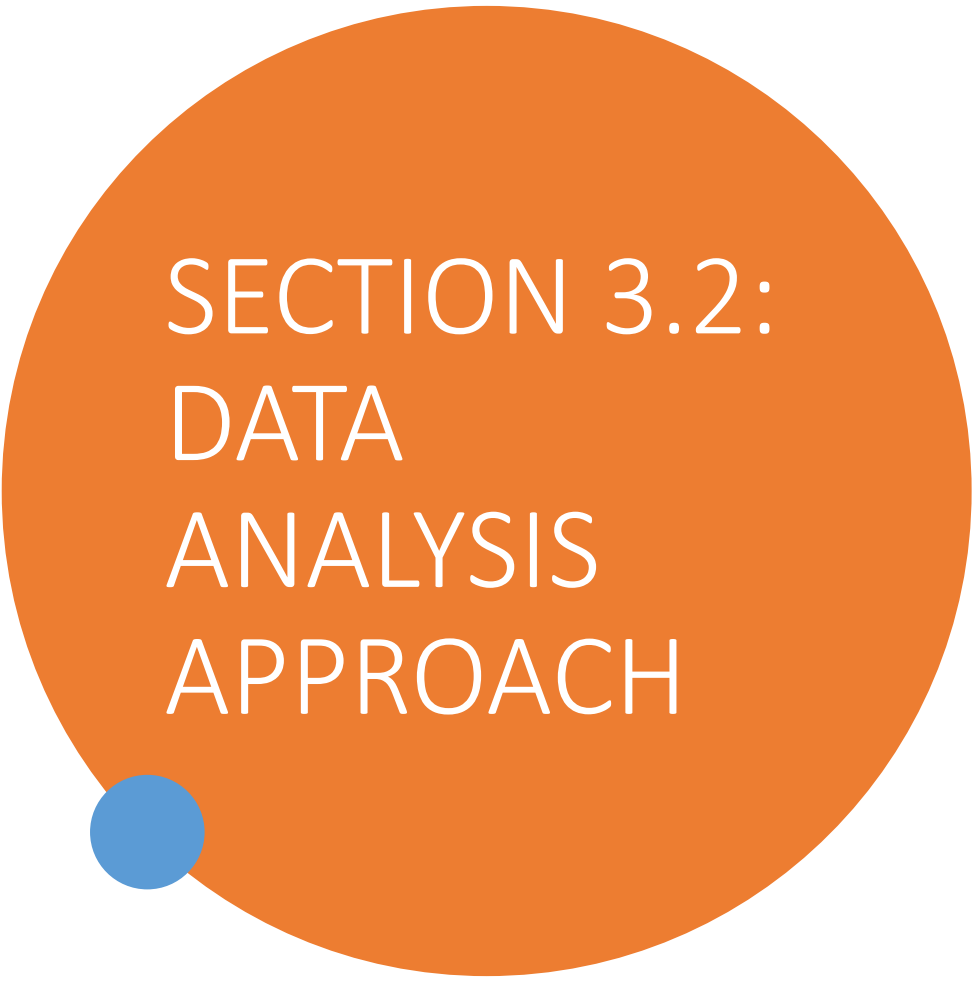


SECTION 2: INITIAL HYPOTHESIS

- a. Not all the countries are the same. Some country tends to have a younger population hence more young athletes and may be advantaged on older population country. (eg. African's country)
 - b. Some other country such as china have an average height smaller than global avarage therefore they will be in disadavantage on sports like: volleyball,basketball, etc.
 - c. Some country will be less likely to win winter sports because the lack of mountain and snow to practice.
- 

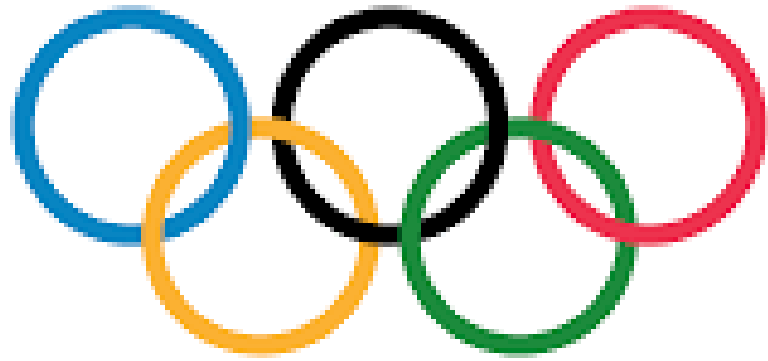
SECTION 3.1: DATA ANALYSIS APPROACH

- a. The first step is to separate Male athletes from Female athletes, since they do not have the same abilities.
- b. The second step is to separate winter sports from summer sports, because since some country can practice freely snow sports, without particular equipments, the latter will be on advantage on the other. Including 'no Snow Country' (NOS) on the high (taller than 175cm) and age (above 30) likelihood to win a medal could lead to bias results.
- c. Analysis by continent, to understand which continent perform better on some sport.
- d. Analysis by sport, to understand which country perform better on some sport



SECTION 3.2: DATA ANALYSIS APPROACH

- a. Probability: What is the likelihood that having certain data allows you to win a medal
- b. A/B Testing and (probably) Regression to understand and test whether some characteristics are useful to win a medal
- c. Graphs will be needed to understand if some country is improving their overall performance and are improving the overall performance (number of medals won increased).



Milestone 2: Descriptive Stats

Data Analysis Project Proposal

30/01/2022

Leonardo Saviane

The Sample transformation

As we can see our dataset has mostly men (70%) and the average age is around 25 years old, max age is around 71 years old this may suggest that further cleaning is needed, for example removing athletes above a certain age threshold (eg.50) to avoid bias estimate and because up to a certain age you cannot compete with younger athletes.

```
sport[['Age', 'Height', 'Weight', 'Sex', 'Win']].describe()
```

	Age	Height	Weight	Sex	Win
count	206165.000000	206165.000000	206165.000000	206165.000000	206165.000000
mean	25.055509	175.371950	70.688337	0.676419	0.146392
std	5.483096	10.546088	14.340338	0.467843	0.353500
min	11.000000	127.000000	25.000000	0.000000	0.000000
25%	21.000000	168.000000	60.000000	0.000000	0.000000
50%	24.000000	175.000000	70.000000	1.000000	0.000000
75%	28.000000	183.000000	79.000000	1.000000	0.000000
max	71.000000	226.000000	214.000000	1.000000	1.000000

Sex indicate with 1 if the athletes is male and 0 if she is female; Win is a binary variable with 1 if the athlete won a medal and 0 if he didn't

The Sample transformation

- Some athletes in 1920 competed at the age of 72(google research), these are clearly outlier that bias the estimate.
- The way people compete change over time, therefore to have interesting insights we should reduce the sample, taking the athletes who competed at games after 1950, without imposing an arbitrary threshold on age directly
- Note that the dataset is parented with local news to they do not take into consideration solely the Olympic games, therefore we should take this into consideration.
- Hence I want to see the sample after the winter Olympics of Moscow in 1980
- I will take into consideration the following Olympics: Rio de Janeiro, London, Beijing, Athina(Athene), Sydney, Atlanta, Barcelona, Seoul, Los Angeles, Moscow, Sochi, Vancouver, Salt Lake City, Nagano, Lillehammer, Albertville, Calgary, Sarajevo, Lake Placid

```
sport_80=pysqldf('''select * from sport where Year>=1980
and City in ('Rio de Janeiro', 'London', 'Beijing','Athina','Sydney','Atlanta','Barcelona')''')
sport_m=pysqldf('''select * from sport_80 where Sex=1 ''')
sport_f=pysqldf('''select * from sport_80 where Sex=0 ''')
sport_80[['Age','Height','Weight','Sex','Win']].describe()
```

	Age	Height	Weight	Sex	Win
count	136875.000000	136875.000000	136875.000000	136875.000000	136875.000000
mean	25.285728	175.841571	70.983039	0.622466	0.142729
std	5.387866	10.893256	14.992473	0.484772	0.349797
min	12.000000	127.000000	28.000000	0.000000	0.000000
25%	22.000000	168.000000	60.000000	0.000000	0.000000
50%	25.000000	176.000000	70.000000	1.000000	0.000000
75%	28.000000	183.000000	80.000000	1.000000	0.000000
max	71.000000	226.000000	214.000000	1.000000	1.000000

NB. Mean age I expected to decrease, instead it increased

Checking differences between Sexes

wemen

```
sport_f[['Age', 'Height', 'Weight', 'Sex', 'Win']].describe()
```

	Age	Height	Weight	Sex	Win
count	51675.000000	51675.000000	51675.000000	51675.0	51675.000000
mean	24.397775	168.552143	60.574446	0.0	0.155530
std	5.511821	8.932364	10.523527	0.0	0.362412
min	12.000000	132.000000	28.000000	0.0	0.000000
25%	21.000000	163.000000	54.000000	0.0	0.000000
50%	24.000000	168.000000	60.000000	0.0	0.000000
75%	28.000000	174.000000	66.000000	0.0	0.000000
max	63.000000	213.000000	167.000000	0.0	1.000000

men

```
sport_m[['Age', 'Height', 'Weight', 'Sex', 'Win']].describe()
```

	Age	Height	Weight	Sex	Win
count	85200.000000	85200.000000	85200.000000	85200.0	85200.000000
mean	25.824284	180.262711	77.295998	1.0	0.134965
std	5.238486	9.511485	13.724913	0.0	0.341688
min	12.000000	127.000000	37.000000	1.0	0.000000
25%	22.000000	174.000000	68.000000	1.0	0.000000
50%	25.000000	180.000000	76.000000	1.0	0.000000
75%	28.000000	186.000000	85.000000	1.0	0.000000
max	71.000000	226.000000	214.000000	1.0	1.000000

With respect to previous analysis (not reported in the slides) mean age for men and wemen increased, although median and quantiles are constant.

Checking differences between Sexes with A/B TEST

wemen

```
sport_f[['Age', 'Height', 'Weight', 'Sex', 'Win']].describe()
```

	Age	Height	Weight	Sex	Win
count	51675.000000	51675.000000	51675.000000	51675.0	51675.000000
mean	24.397775	168.552143	60.574446	0.0	0.155530
std	5.511821	8.932364	10.523527	0.0	0.362412
min	12.000000	132.000000	28.000000	0.0	0.000000
25%	21.000000	163.000000	54.000000	0.0	0.000000
50%	24.000000	168.000000	60.000000	0.0	0.000000
75%	28.000000	174.000000	66.000000	0.0	0.000000
max	63.000000	213.000000	167.000000	0.0	1.000000

men

```
sport_m[['Age', 'Height', 'Weight', 'Sex', 'Win']].describe()
```

	Age	Height	Weight	Sex	Win
count	85200.000000	85200.000000	85200.000000	85200.0	85200.000000
mean	25.824284	180.262711	77.295998	1.0	0.134965
std	5.238486	9.511485	13.724913	0.0	0.341688
min	12.000000	127.000000	37.000000	1.0	0.000000
25%	22.000000	174.000000	68.000000	1.0	0.000000
50%	25.000000	180.000000	76.000000	1.0	0.000000
75%	28.000000	186.000000	85.000000	1.0	0.000000
max	71.000000	226.000000	214.000000	1.0	1.000000

With respect to previous analysis (not reported in the slides) mean age for men and wemen increased, although median and quantiles are constant. Moreover as you can see below the two groups are heterogenous in weight and height Therefore..perfmance.

```
t_stat, p_val= ss.ttest_ind(weight_m['Weight'],weight_f['Weight'])
print('Difference in Weight; t_stat: ',t_stat, 'p_value: ', p_val)

t_stat, p_val= ss.ttest_ind(height_m['Height'],height_f['Height'])
print('Difference in Height; t_stat: ',t_stat, 'p_value: ', p_val)
```

Difference in Weight; t_stat: 237.78583480326142 p_value: 0.0

Difference in Height; t_stat: 225.90664378364585 p_value: 0.0

Are there State differences?

- We may ask ourself if there are any differences between two continent or state. Here i take two sample countries that i think are representative of their own continent.
- By running an A/B test the null hyphotetis is rejected, therefore there is a difference between the two countries.
- I assume that such difference exists between each continent and may exists between countries.
- Therefore I will create a set of dummies for each state, it will be useful for the regression

```
height_ita=pysqldf('''select Height from sport_80 where NOC="ITA" and Sex=1 ''')
height_chn=pysqldf('''select Height from sport_80 where NOC="CHN" and Sex=1 ''')
height_ita.describe()
```

Height	
count	3369.000000
mean	180.266845
std	8.870825
min	152.000000
25%	174.000000
50%	180.000000
75%	186.000000
max	215.000000

```
height_chn.describe()
```

Height	
count	2121.000000
mean	177.689769
std	11.732913
min	148.000000
25%	170.000000
50%	178.000000
75%	185.000000
max	226.000000

```
t_stat, p_val= ss.ttest_ind(height_ita['Height'],height_chn['Height'])
t_stat, p_val
```

```
(9.229752559109919, 3.791512826226679e-20)
```

Are there State differences? Box-plot ita-chn

```
fig, axs = plt.subplots(2, figsize=(10, 7))

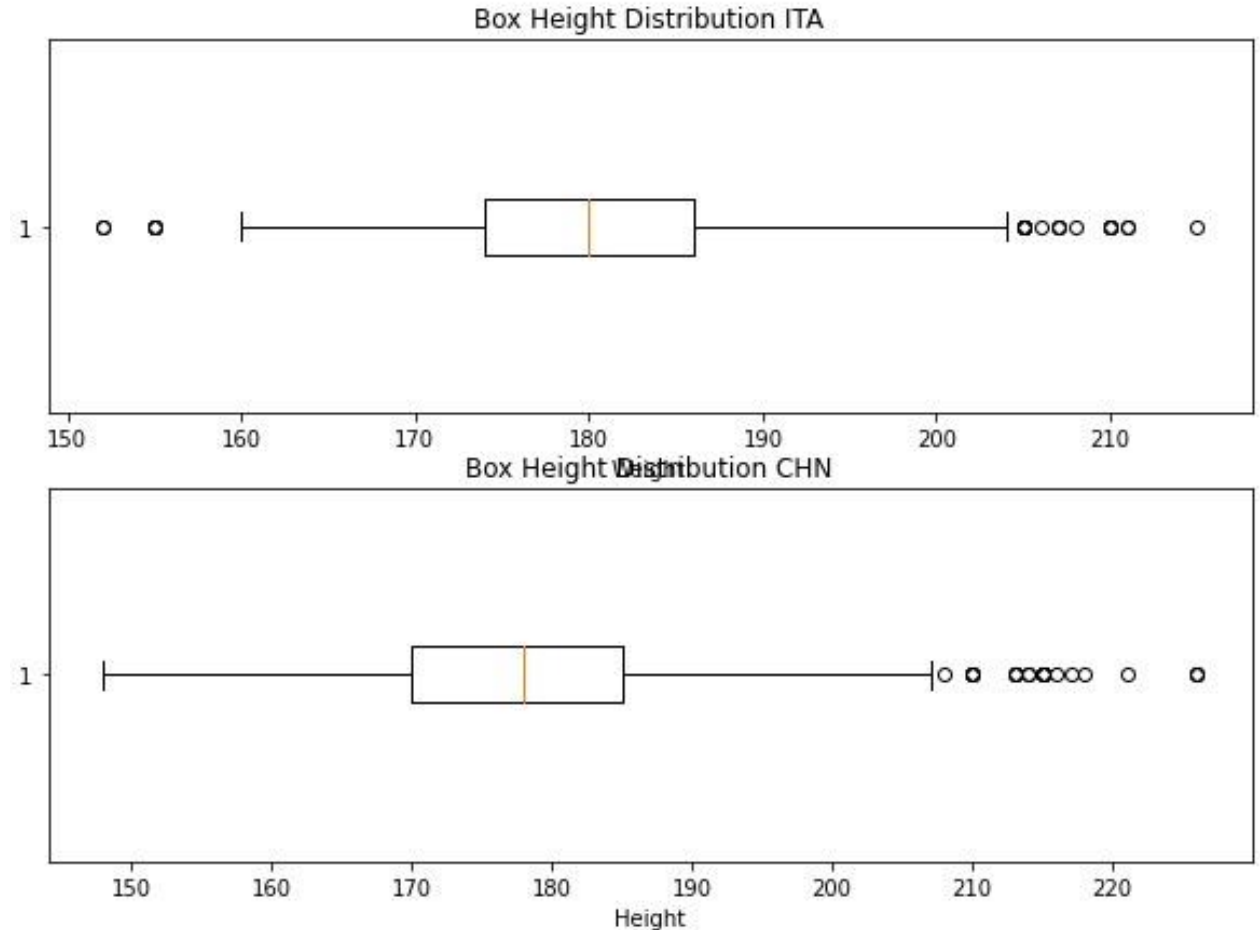
##BOX PLOT
axs[0].boxplot(height_ita['Height'], vert=False)

axs[0].set_xlabel('Weight') # add to x-label to the plot
axs[0].set_title('Box Height Distribution ITA') # add title to the plot

axs[1].boxplot(height_chn['Height'], vert=False)

axs[1].set_xlabel('Height') # add to x-label to the plot
axs[1].set_title('Box Height Distribution CHN') # add title to the plot
```

Codes and results



Weight and Height differences between Winter and Summer Olympics? A/B Test and distributions graphs

- Is there a difference weight and height between winter olympics and summer olympics?
- Since we have seen that there is a difference between man and women, I further explore this analysis by diving the two sexes
- A/B test shows that between Season=1(summer) and winter there are weight and height differences! **except for male weight which is the same in the two groups!**
- You will see many outlier in the box plot for male in summer Olympics

```
height_count_m_s=pysqldf('''select Height,Count(Height) as 'Height_Count' from sport_80 where Sex=1 and Season = 1 group by Height ''')
height_count_f_s=pysqldf('''select Height,Count(Height) as 'Height_Count' from sport_80 where Sex=0 and Season = 1 group by Height ''')

height_m_s=pysqldf('''select Height from sport_80 where Sex=1 and Season = 1 ''')
height_f_s=pysqldf('''select Height from sport_80 where Sex=0 and Season = 1 ''')
```

```
height_count_m_w=pysqldf('''select Height,Count(Height) as 'Height_Count' from sport_80 where Sex=1 and Season = 0 group by Height ''')
height_count_f_w=pysqldf('''select Height,Count(Height) as 'Height_Count' from sport_80 where Sex=0 and Season = 0 group by Height ''')

height_m_w=pysqldf('''select Height from sport_80 where Sex=1 and Season = 0 ''')
height_f_w=pysqldf('''select Height from sport_80 where Sex=0 and Season = 0 ''')
```

A/B TESTING

```
t_stat, p_val= ss.ttest_ind(height_m_s['Height'],height_m_w['Height'])
t_stat, p_val
```

```
(10.537528545769307, 6.017725959006674e-26)
```

```
t_stat, p_val= ss.ttest_ind(height_f_s['Height'],height_f_w['Height'])
t_stat, p_val
```

```
(20.84581328231301, 4.1439688067478515e-96)
```

```
t_stat, p_val= ss.ttest_ind(weight_m_s['Weight'],weight_m_w['Weight'])
t_stat, p_val
```

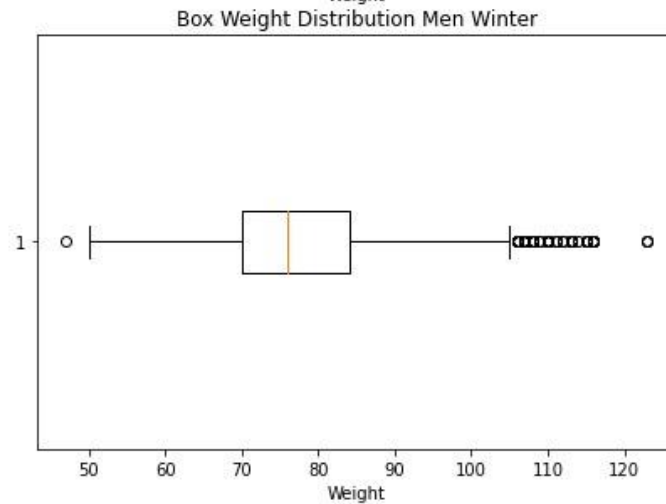
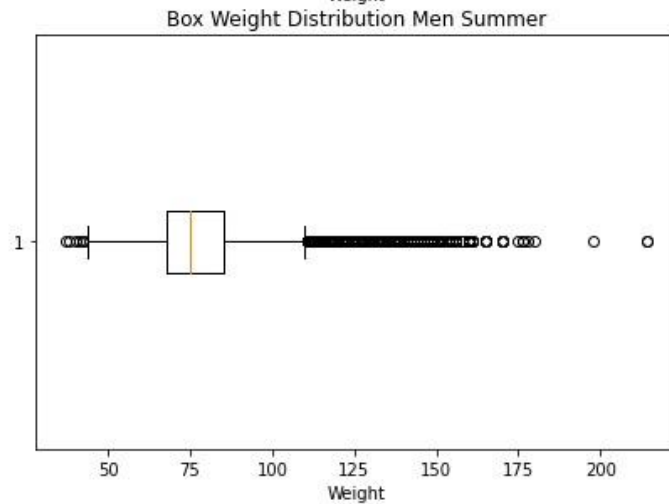
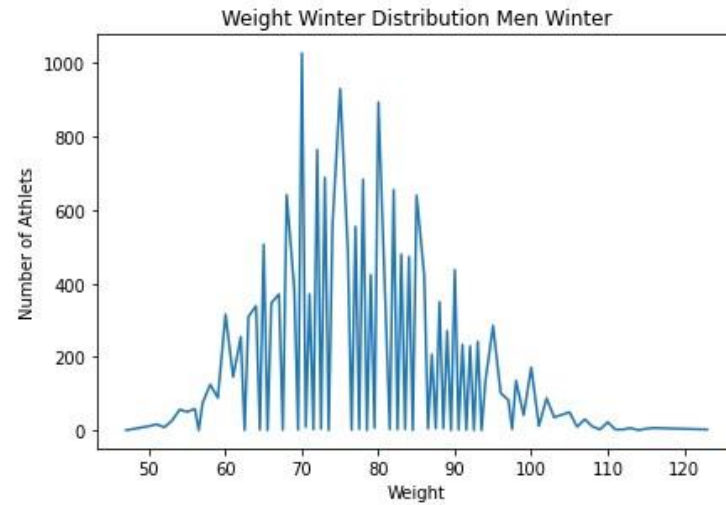
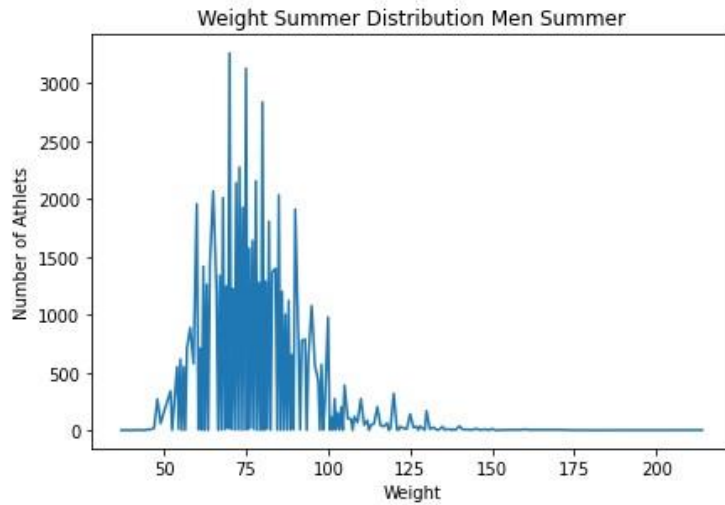
```
(0.7012859803826186, 0.4831264765792309)
```

```
t_stat, p_val= ss.ttest_ind(weight_f_s['Weight'],weight_f_w['Weight'])
t_stat, p_val
```

```
(7.293597248171901, 3.0610024058333367e-13)
```

SQL CODE ABOVE

Weight and Height differences between Winter and Summer Olympics? A/B Test and distributions graphs



```
weight_m_s.describe()
```

	Weight
count	67371.000000
mean	77.312961
std	14.441964
min	37.000000
25%	68.000000
50%	75.000000
75%	85.000000
max	214.000000

```
weight_m_w.describe()
```

	Weight
count	17829.000000
mean	77.231897
std	10.585654
min	47.000000
25%	70.000000
50%	76.000000
75%	84.000000
max	123.000000

```
fig, axs = plt.subplots(2,2,figsize=(15,10))

##DISTRIBUTION GRAPHS
axs[0,0].plot(weight_count_m_s['Weight'],weight_count_m_s['Weight_Count'])

axs[0,0].set_xlabel('Weight') # add to x-label to the plot
axs[0,0].set_ylabel('Number of Athlets') # add y-label to the plot
axs[0,0].set_title('Weight Summer Distribution Men Summer') # add title to the plot

axs[0,1].plot(weight_count_m_w['Weight'],weight_count_m_w['Weight_Count'])

axs[0,1].set_xlabel('Weight') # add to x-label to the plot
axs[0,1].set_ylabel('Number of Athlets') # add y-label to the plot
axs[0,1].set_title('Weight Winter Distribution Men Winter') # add title to the plot

##BOX PLOT
axs[1,0].boxplot(weight_m_s['Weight'],vert=False)

axs[1,0].set_xlabel('Weight') # add to x-label to the plot
axs[1,0].set_title('Box Weight Distribution Men Summer') # add title to the plot

axs[1,1].boxplot(weight_m_w['Weight'],vert=False)

axs[1,1].set_xlabel('Weight') # add to x-label to the plot
axs[1,1].set_title('Box Weight Distribution Men Winter') # add title to the plot
```

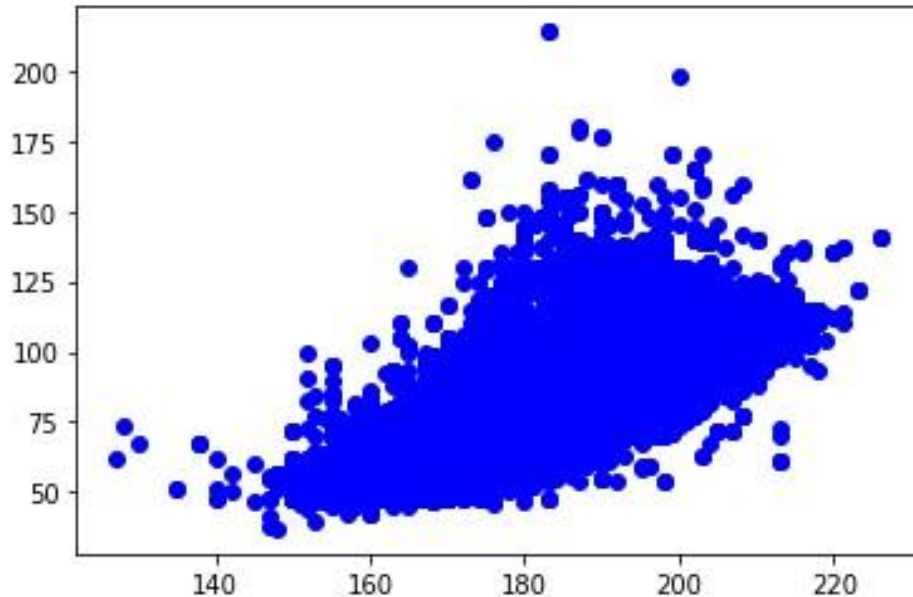
Count of the observations and box plot for male weight in summer and winter

EXTRA: Scatter plot between height and weight

The following graphs show a positive correlation between Height and Weight, respectively for Man and Women

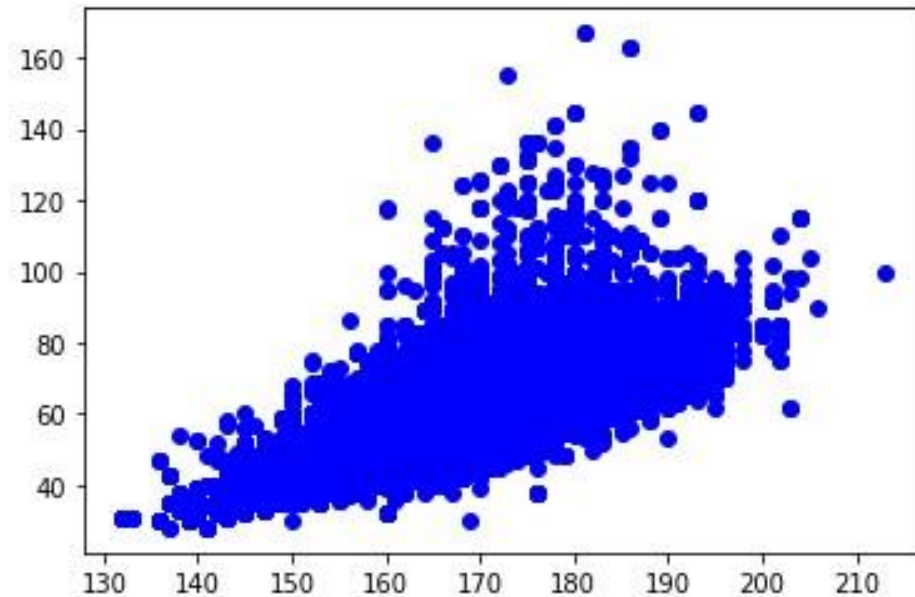
```
plt.scatter(height_m, weight_m, color='blue')
```

```
<matplotlib.collections.PathCollection at 0x26952b53ca0>
```



```
plt.scatter(height_f, weight_f, color='blue')
```

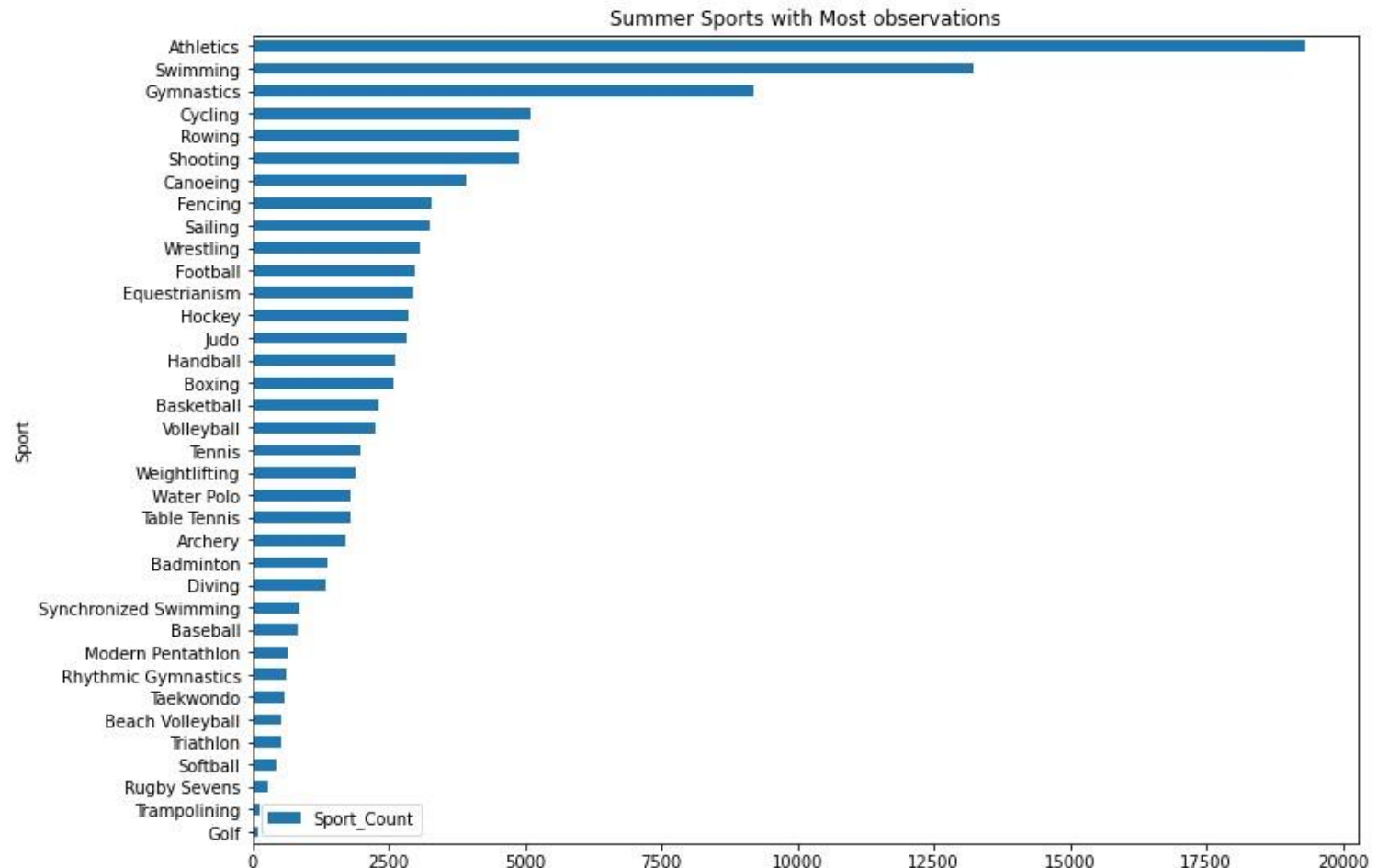
```
<matplotlib.collections.PathCollection at 0x2695487c400>
```



Which is the sport that counts the highest number of athletes?

```
obv_sport=pysqldf('''select Sport,Count(Sport) as 'Sport_Count' from sport_80 group by Sport ''')  
  
obv_sport_s=pysqldf('''select Sport,Count(Sport) as 'Sport_Count' from sport_80 Where Season = 1 group by Sport ''')  
  
obv_sport_s.sort_values(by='Sport_Count', ascending=True, inplace=True)  
  
obv_sport_s.set_index('Sport', inplace=True)  
  
obv_sport_s.plot.barh(figsize=(12,9)).set_title('Summer Sports with Most
```

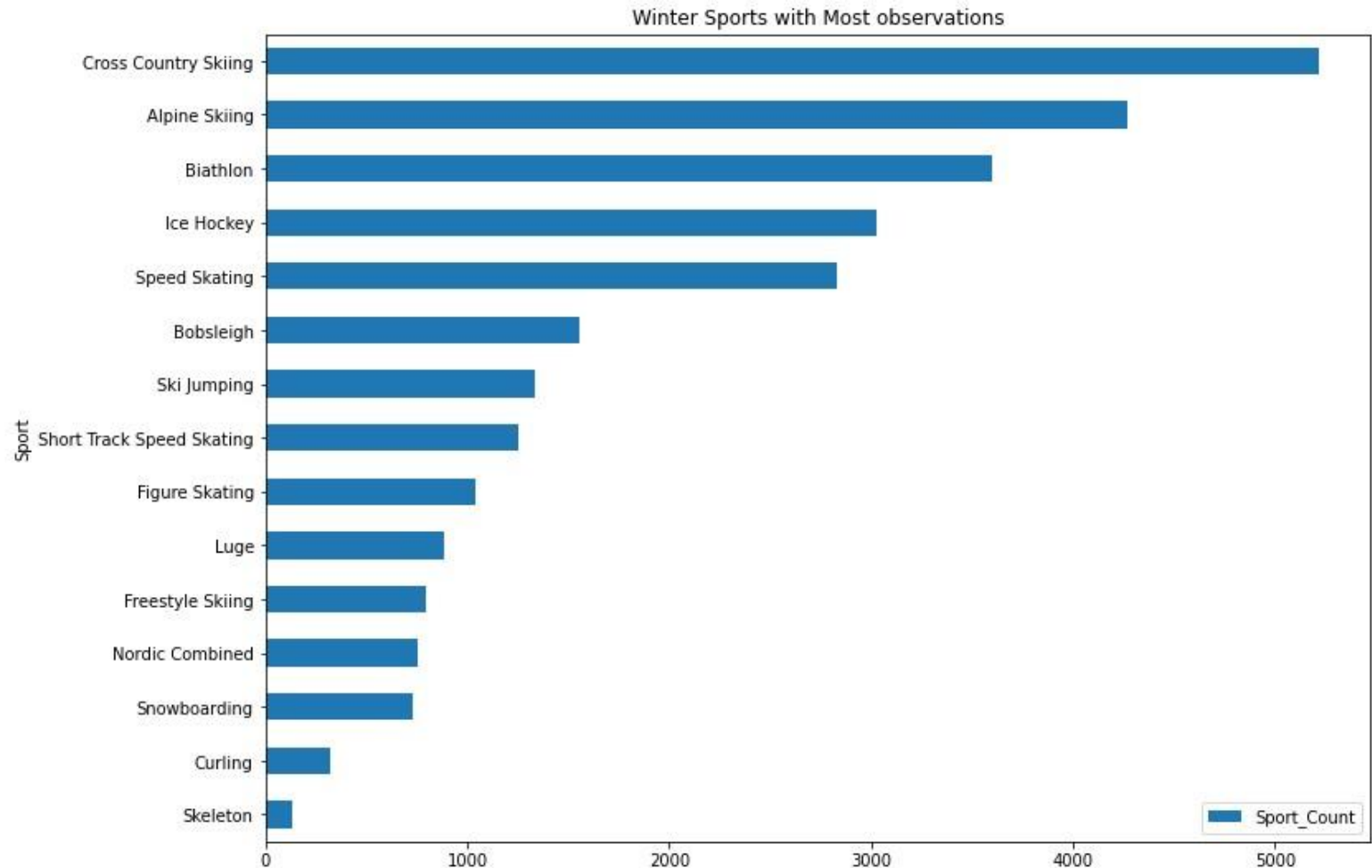
Se we can see the most popular sport among the Olympic summer games is Athelitics with counted more and 20000 atheltes throughout the years



Which is the sport that counts the highest number of athletes?

```
obv_sport_w=pysqldf('''select Sport,Count(Sport) as 'Sport_Count' from sport_80 Where Season = 0 group by Sport ''')
obv_sport_w.sort_values(by='Sport_Count', ascending=True, inplace=True)
obv_sport_w.set_index('Sport', inplace=True)
obv_sport_w.plot.barh(figsize=(12,9)).set_title('Winter Sports wi
```

Se we can see the most popular sport among the Olympic Winter games is Cross country Skiing with counted more and 5000 atheltes throughout the years



LAST: Is there difference in Height and Weight between Sport? A/B Testing

- If there exists a difference in height and weight between each sport, the group of athletes must be separated in the regression because otherwise they will bias the estimate.
- AB test works perfectly here, in the subcategory of the men I will take two sports randomly(Gymnastics and Athletics) and perform the test on height and weight.
- I will assume that the results it's representative of each category.

```
Sport_tt1=pysqldf(''select Height,Weight from sport_80 Where Sport = 'Gymnastics' and Sex=1 ''')
Sport_tt2=pysqldf(''select Height,Weight from sport_80 Where Sport = 'Athletics' and Sex=1 ''')

t_stat, p_val= ss.ttest_ind(Sport_tt1['Height'],Sport_tt2['Height'])
print('Height T-stat: ', t_stat,'Height p-value: ', p_val)

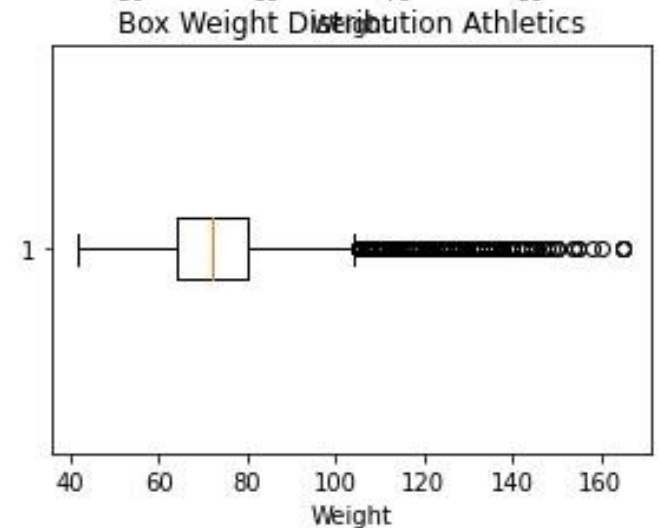
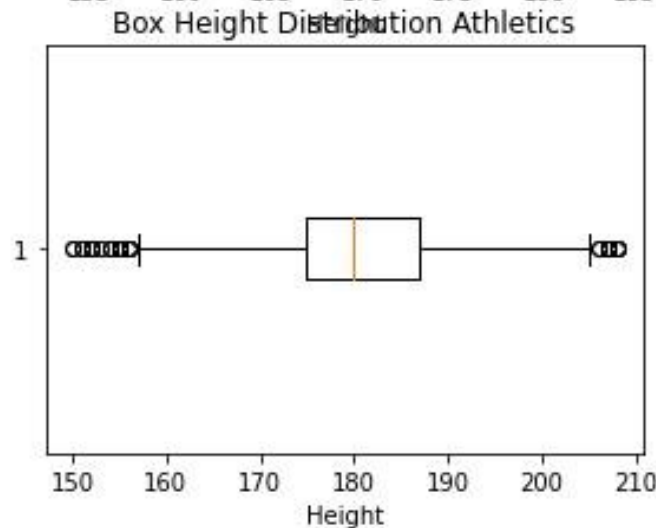
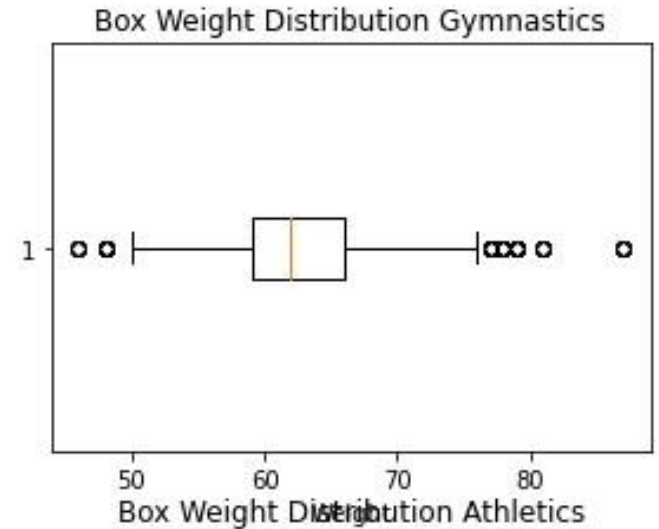
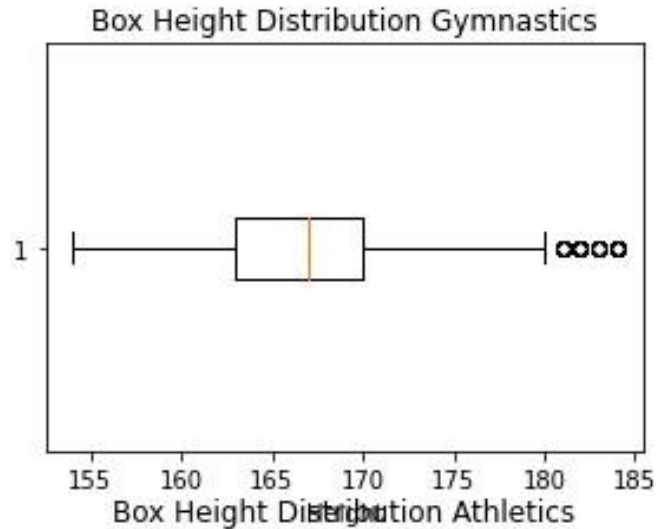
t_stat, p_val= ss.ttest_ind(Sport_tt1['Weight'],Sport_tt2['Weight'])
print('Weight T-stat: ', t_stat,'Weight p-value: ', p_val)
```

```
Height T-stat: -111.3967094784561 Height p-value: 0.0
Weight T-stat: -53.347663315055506 Weight p-value: 0.0
```

The two groups are strongly different!
Null hypothesis is rejected!

LAST: Is there difference in Height and Weight between Sport? A/B Testing

- As we can see from the box plot the median value differs a lot from the other.
- Athletics weights have a lot of outlier which may bias estimate



In the next slide the findings to implement in the last part!

CONCLUSION

Milestone 2: Descriptive Stats

- Most of the hypothesis i had made at beginning can be tested only in the next part.
- Since there is a difference between sexes they should be tested separately in the regression to avoid bias estimate
- Similarly for Winter and Summer Sport, athletes have different weight and height therefore we shouldn't confound them.
- Again, I tested difference between athletes of different sport and there is a statistically significant difference.
- Country as well, but we can create a dummy for each country(whether an athlete belongs to it or not) and see in the regression if belonging to a particular country makes you more likely to win a medal.
- To see if someone is a potential athlete we have to take into consideration all these things in the next part.
- Since there are many sports i will explore only the 4 most popular sports: Cross Country Skiing, Gymnastics, Swimming, Athletics