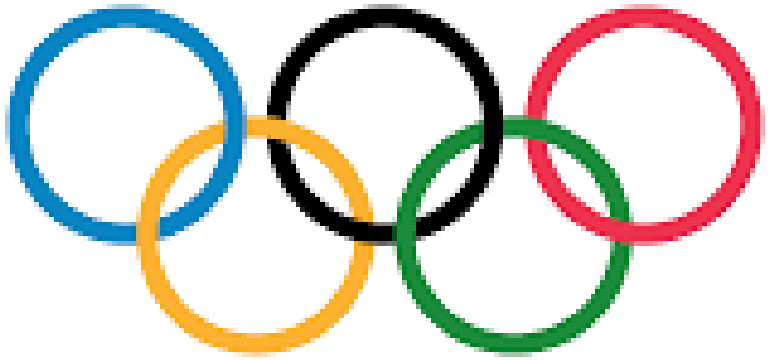


OLYMPIC GAMES STATS PROJECT

Data Analysis Project Proposal

28/01/2022

Leonardo Saviane

Why Sports

- I want to retrieve some insights on the best athletes that the world have ever known, to undestand 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/ore medal, will be transformed in a binary variable as well. My purpose it’s to estimate the likelyhood 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 loosing 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→**

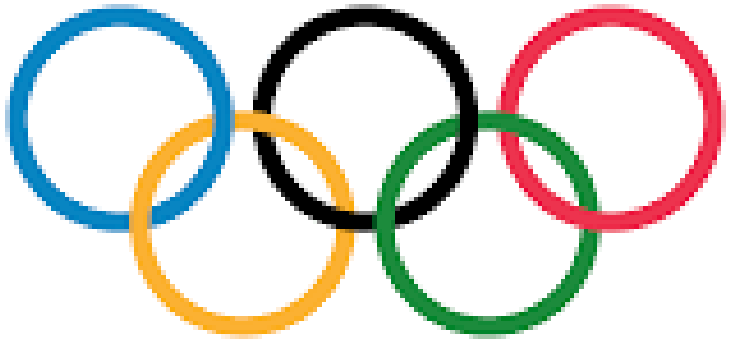
ENTITY RELATIONAL DIAGRAM

Athletes_events

- ID
- Name
- Sex
- Age
- Height
- Weight
- Team
- NOC
- Games
- Year
- Season
- City
- Sport
- Event
- Medal

NOC_regions

- 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.

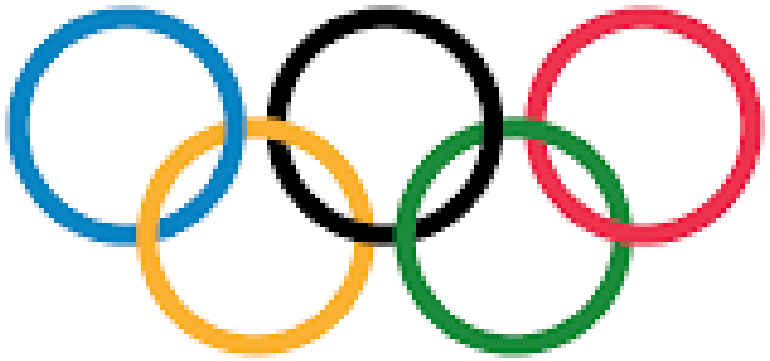
b. The second step is to separate winter sports from summer sports, because since some country can practice freely snow sports, without particular equipments, the latter will be on advantage on the other. Including ‘no Snow Country’(NOS) on the hight (taller than 175cm) and age (above 30) likelyhood 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 it the likelyhood that having certain data allows you to win a medal
b. A/B Testing and (probably) Regression to understand and test wheter some charateristic are usefull to win a medal
c. Graphs will be needed to understand if some country are improving their overall performance and are improving the overall performance (number of medal 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 atheletes 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.
- Thefore 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

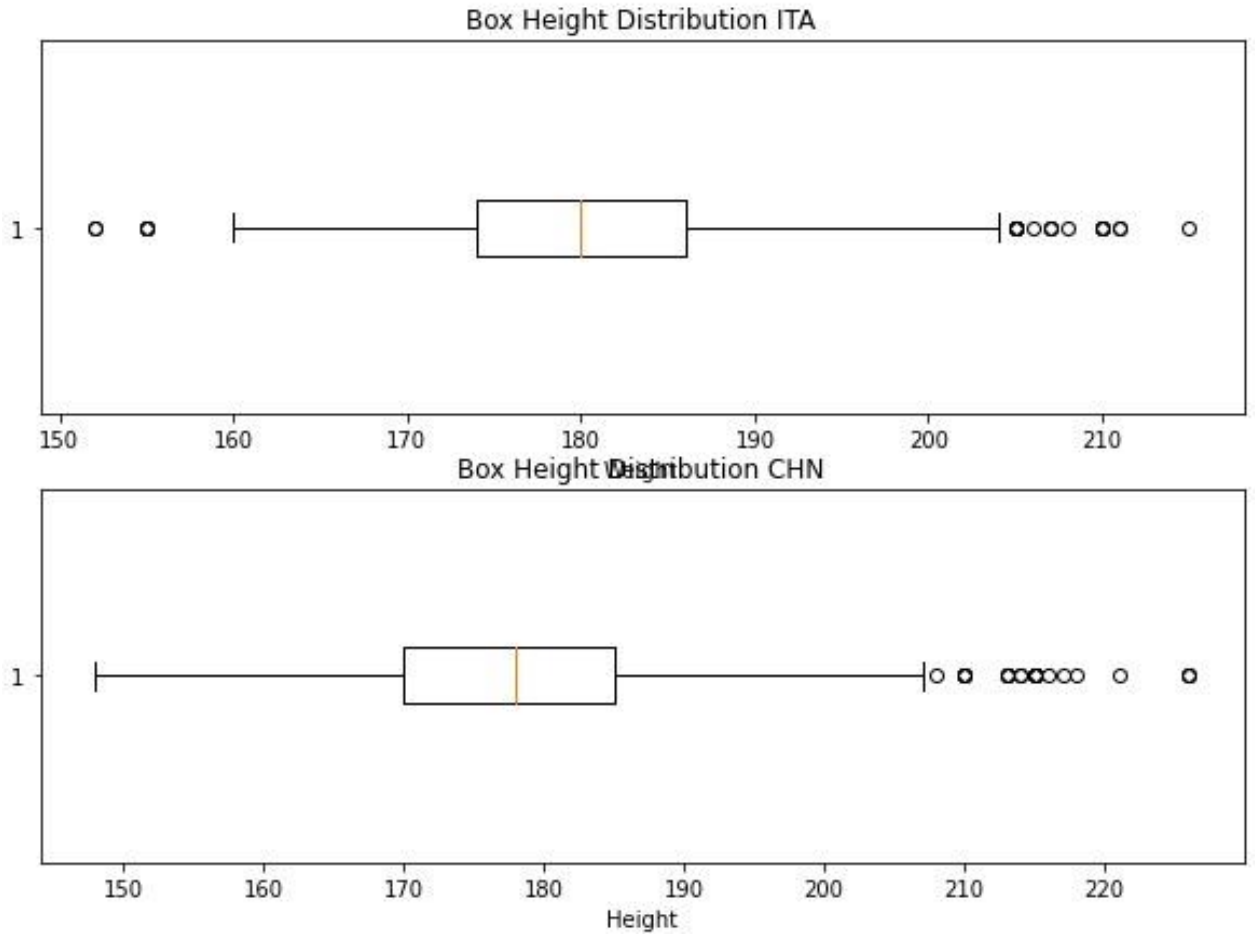
```
t_stat, p_val= ss.ttest_ind(height_ita['Height'],height_chn['Height'])
t_stat, p_val
```

```
(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 beween Winter and Summer Olympics? A/B Test and distributions graphs

- Is there a difference weight and height between winter olympcs and summer olympics?
- Since we have seen that there is a difference between man and wemen, 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! **exept 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

SQL CODE ABOVE

```
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)
```

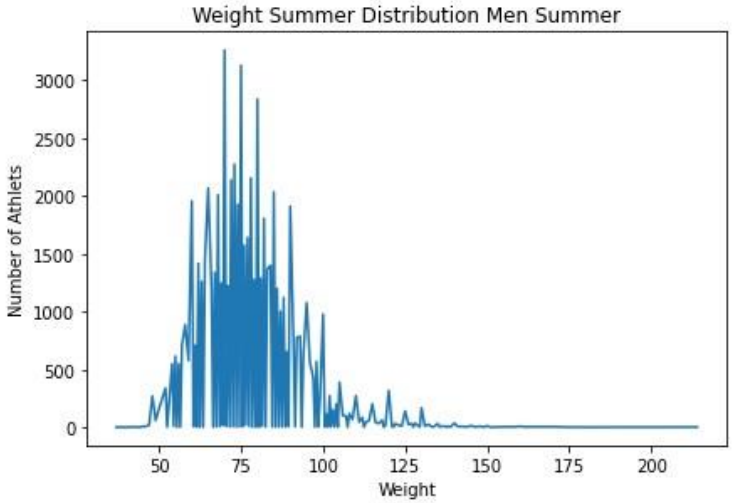
```
t_stat, p_val= ss.ttest_ind(weight_m_s['Weight'],weight_m_w['Weight'])
t_stat, p_val
```

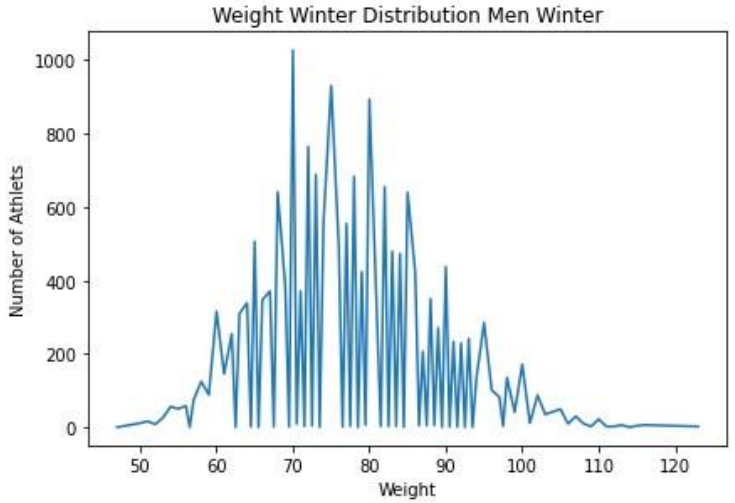
```
(0.7012859803826186, 0.4831264765792309)
```

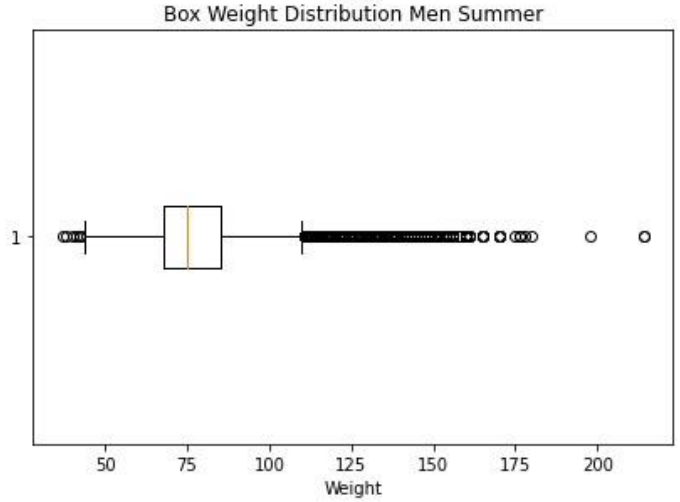
```
t_stat, p_val= ss.ttest_ind(weight_f_s['Weight'],weight_f_w['Weight'])
t_stat, p_val
```

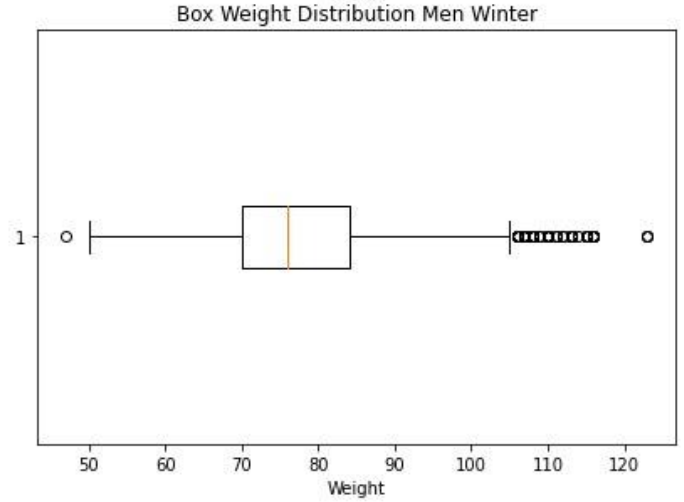
```
(7.293597248171901, 3.0610024058333367e-13)
```

Weight and Height differences beween 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
```

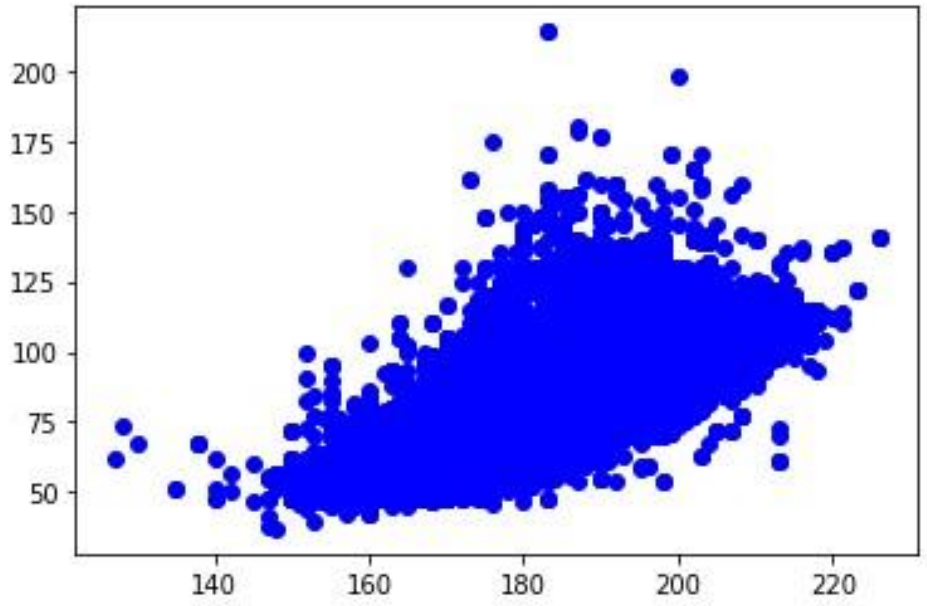
Count of the observations and box plot for male weight in summer and winter

EXTRA: Scatter plot between height and weight

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

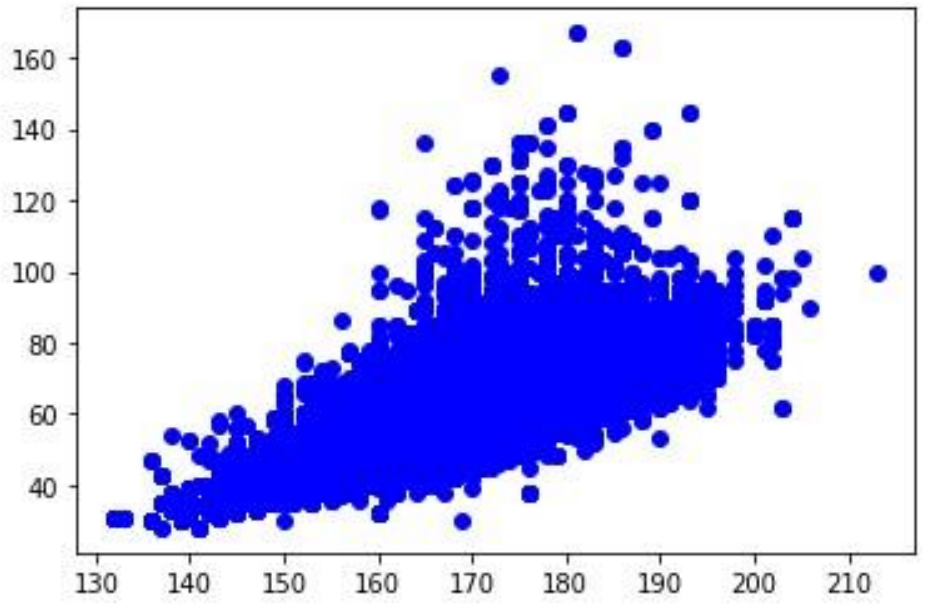
```
plt.scatter(height_m, weight_m,  color='blue')
```

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




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

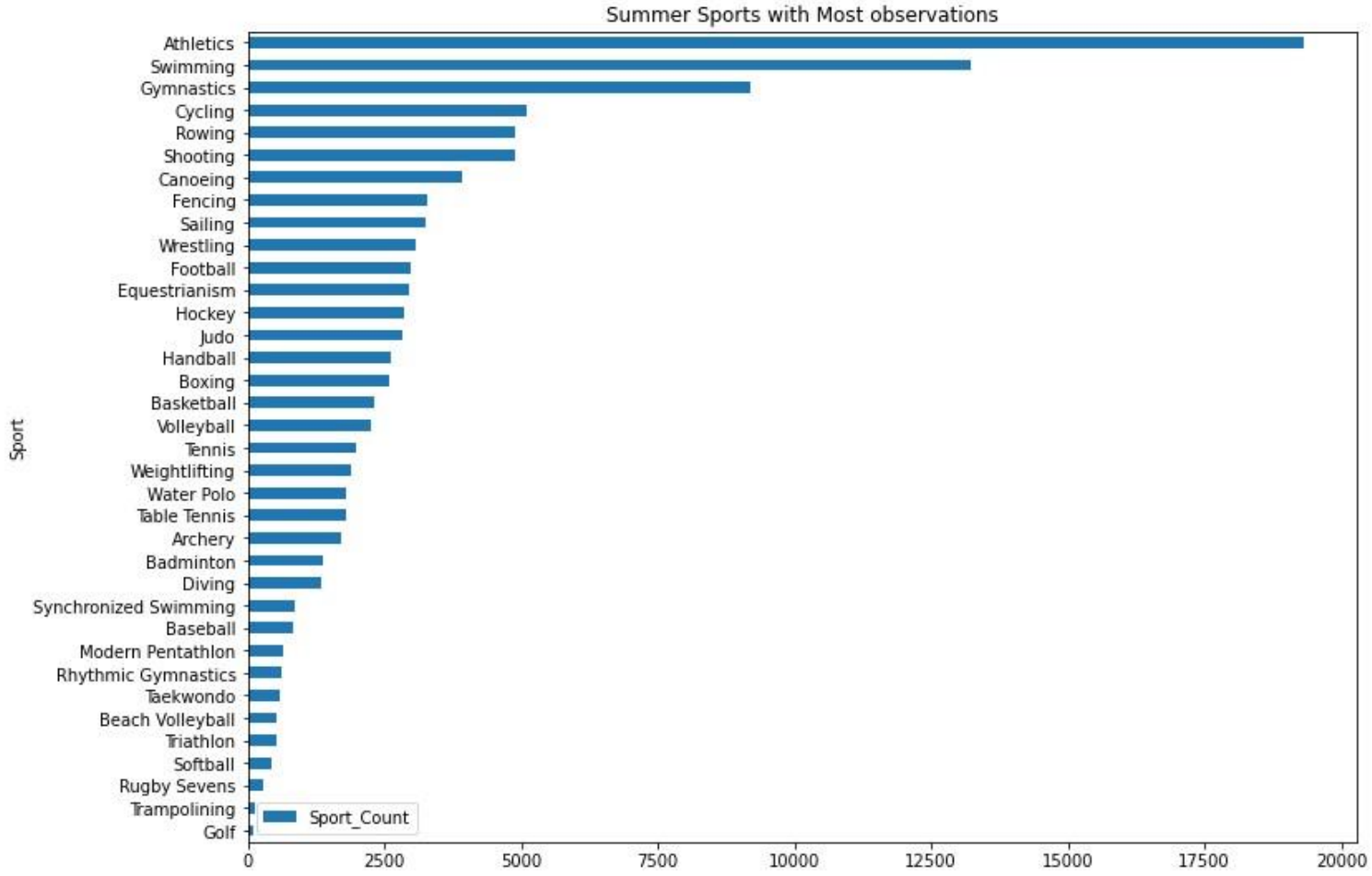
```
<matplotlib.collections.PathCollection at 0x2695487c400>
```



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
```

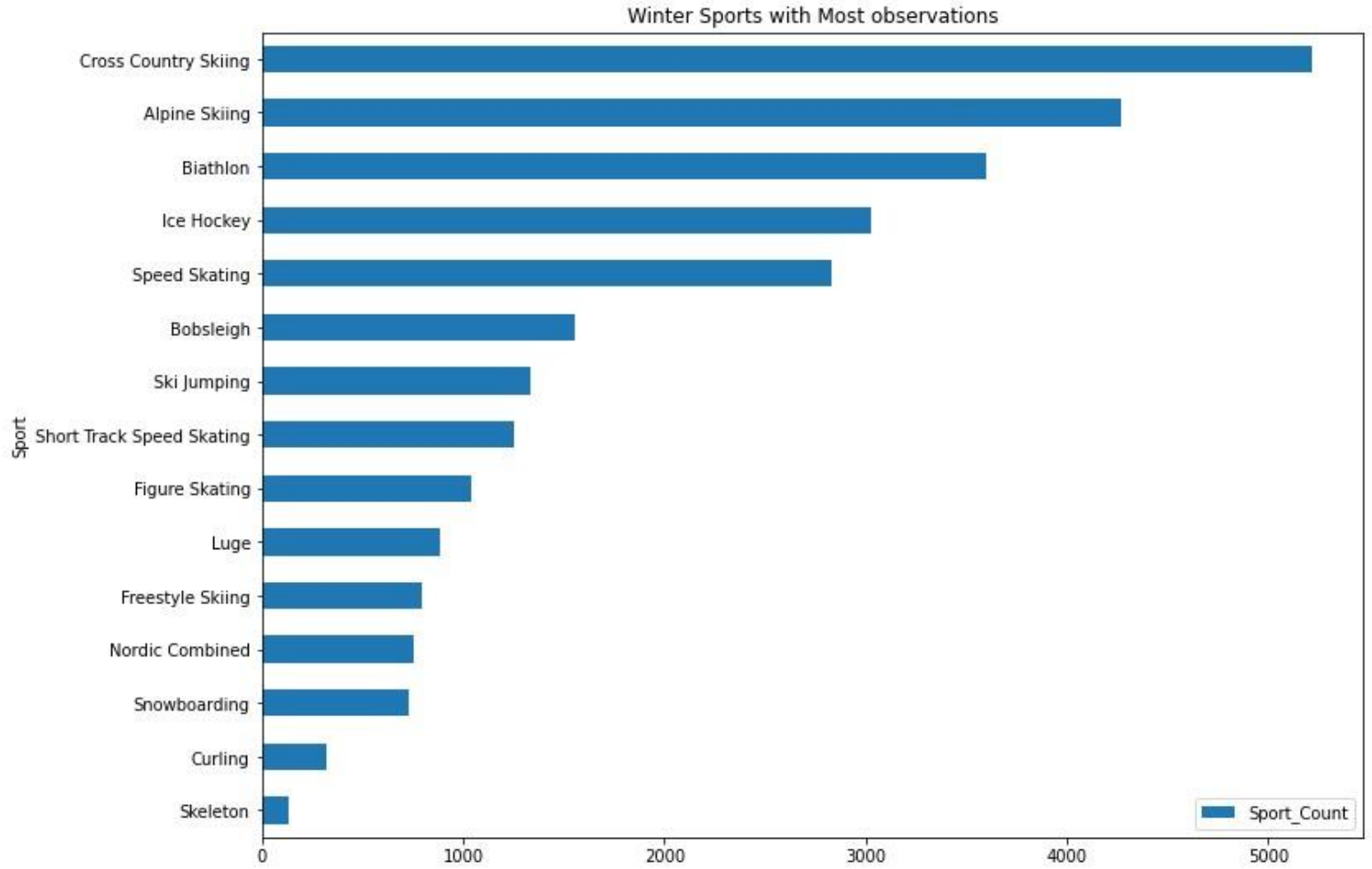
Se we can see the most popular sport among the Olympic summer games is Athelitics with counted more and 20000 atheltes throughout the years



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 becouse 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 perfom the test on height and weight.
- I will assume that the results it’s rapresentative 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)
```

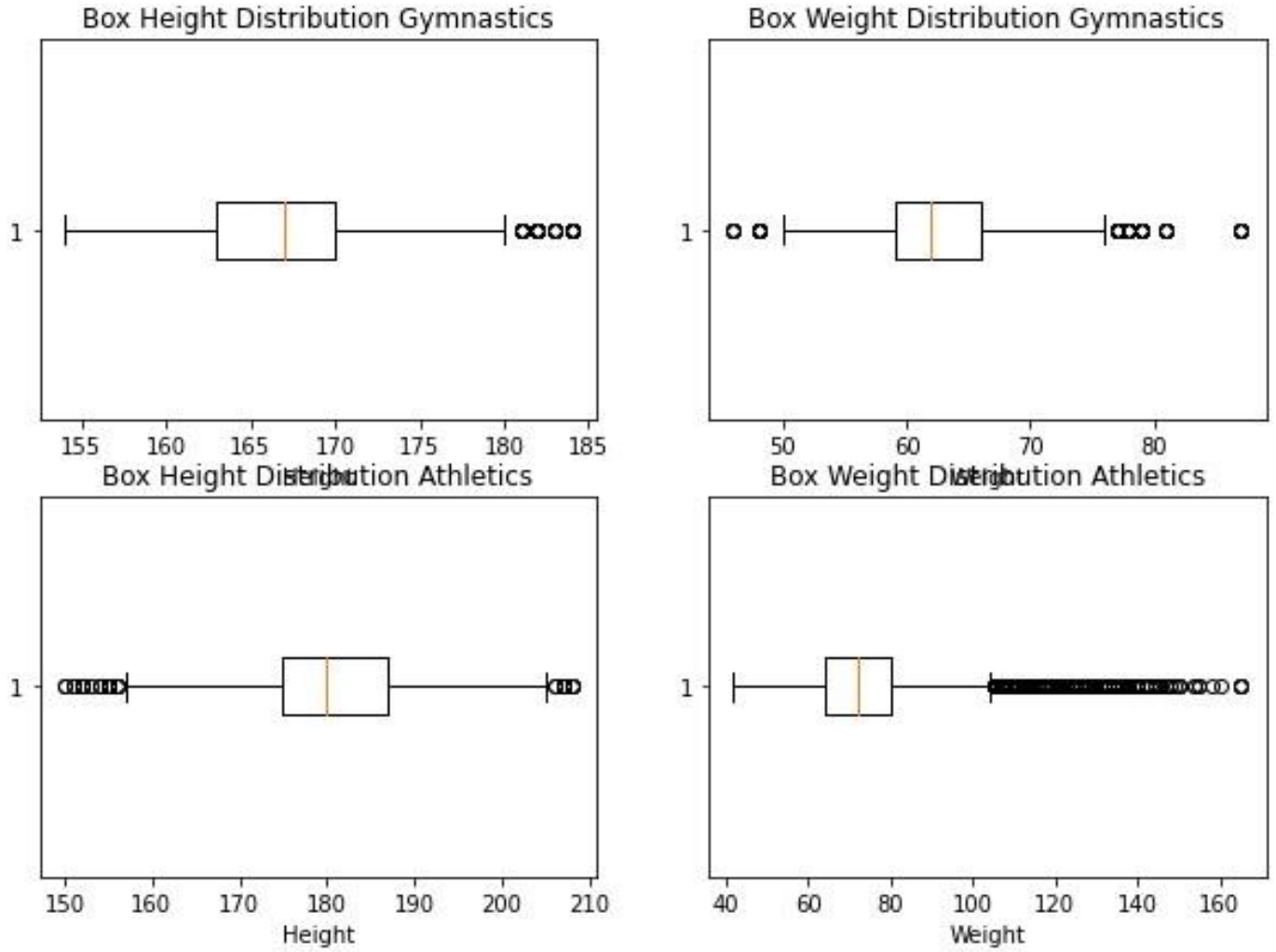
```
Height T-stat:  -111.3967094784561 Height p-value:  0.0
Weight T-stat:  -53.347663315055506 Weight p-value:  0.0
```

The two groups are strongly different!
Null hyphotesis 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 differes a lot from the other.
- Athletics weights have a lot of outlier which may bias estimate

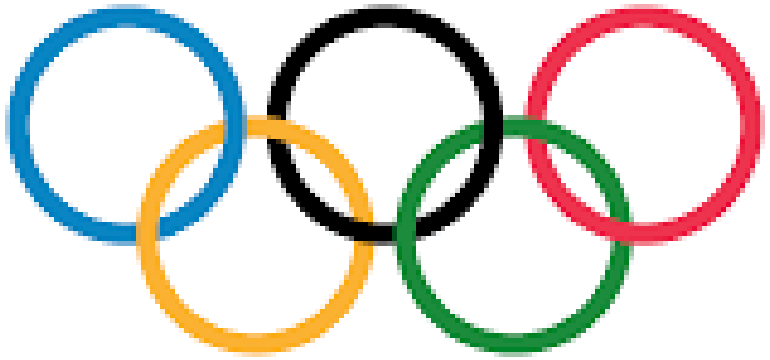
In the next slide the findings to implement in the last part!



CONCLUSION Milestone 2: Descriptive Stats

- Most of the hypothesis i had made at beginning can be tested only in the next part.
- AB test comes in handy for testing differences between groups and it have been used to test if the groups are similar
- 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 significative difference.
- Since are many sports i will explore only the 4 most popular sport among the athletes: Cross Country Skiing (most popular among winter sports), Gymnastics(third to last most popular sport among summer sports), Swimming(second to last most popular sport among summer sports), Atheletics(most popular sport among summer sports).
- To see if someone is a potential athletes we have to take into consideration all these things in the next part.
- Since are many sports i will explore only the 4 most popular sport among the athletes: Cross Country Skiing (most popular among winter sports), Gymnastics(third to last most popular sport among summer sports), Swimming(second to last most popular sport among summer sports), Atheletics(most popular sport among summer sports).

Codes: https://github.com/LeonSavi/Projects_Leonardo_Saviane.git



Milestone 3: Beyond Descriptive Stats

Data Analysis Project Proposal

30/01/2022

Leonardo Saviane

CODES:
https://github.com/LeonSavi/Projects_Leonardo_Saviane.git

PREPERARING FOR REGRESSION

- One of the first thing that we have to do for our regression is to create a dummy for each country, which means if a an athletes belong to country(=1) or not(=0)
- Then through SQL we get the data we need to put in our regression and watch the coefficient
- NOC_ITA will be removed to avoid collinearity, the intercept/costant will capture italy

Now I want to create binary variable for each NOC, in order to see the influence of each country

the following variables will be removed from sport_80: ID, Name, City,Year, NOC_ITA(to avoid collinearity), costant will capture ITALIAN observations

```
dummies=pd.get_dummies(sport_80[['NOC']])

dummies=dummies.drop(['NOC_ITA'],axis=1)

dataset=sport_80.drop(['NOC','ID','Name','City','Year'],axis=1)

dataset = pd.concat([dataset,dummies],axis=1)

dataset.head(5)
```

	Age	Height	Weight	Sport	Sex	Season	Win	NOC_AFG	NOC_AHO	NOC_ALB	...	NOC_VAN	NOC_VEN	NOC_VIE	NOC_VIN	NOC_YAR	NOC_YEM	NOC_YMD	NOC_YUG	NOC_ZAM	NOC_ZIM
0	24.0	180.0	80.0	Basketball	1	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	23.0	170.0	60.0	Judo	1	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	21.0	185.0	82.0	Speed Skating	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	21.0	185.0	82.0	Speed Skating	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	25.0	185.0	82.0	Speed Skating	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 224 columns

```
reg1_m_x=pysqldf('''select * from dataset Where Sport = 'Cross Country Skiing' and Sex = 1 ''')
reg1_m_y=pysqldf('''select Win from dataset Where Sport = 'Cross Country Skiing' and Sex = 1 ''')

reg1_f_x=pysqldf('''select * from dataset Where Sport = 'Cross Country Skiing' and Sex = 0 ''')
reg1_f_y=pysqldf('''select Win from dataset Where Sport = 'Cross Country Skiing' and Sex = 0 ''')

reg1_m_x.drop(['Win','Sport','Season'],axis=1,inplace=True)
reg1_f_x.drop(['Win','Sport','Season'],axis=1,inplace=True)
reg1_m_x.head()
```

	Age	Height	Weight	Sex	NOC_AFG	NOC_AHO	NOC_ALB	NOC_ALG	NOC_AND	NOC_ANG	...	NOC_VAN	NOC_VEN	NOC_VIE	NOC_VIN	NOC_YAR	NOC_YEM	NOC_YMD	NOC_YUG	NOC_ZAM	NOC_ZIM
0	31.0	188.0	75.0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	31.0	188.0	75.0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	31.0	188.0	75.0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	31.0	188.0	75.0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	33.0	188.0	75.0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 221 columns

REGRESSION SET-UP(1)

- While running the regression all country variables seemed to not be statistically significant, R2 was low and F statistic as well. This means that there were too many variables.
- To explore country effect on the probability of winning a medal more complex model are need. Some model that can exploit the difference between groups.
- **In this project I stick with multiple linear regression (OLS)**
- To avoid the problem I said i take into consideration only some of the variables such as Age, Weight, Height. The intercept is added in the model.

```
reg1_m_x=pysqldf('''select Age,Weight,Height from dataset Where Sport = 'Cross Country Skiing' and Sex = 1 ''')

## X is the input variables (or independent variables)
X = reg1_m_x
## y is the target/dependent variable
y = reg1_m_y
## add an intercept (beta_0) to our model
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()
predictions = model.predict(X)

# Print out the statistics
model.summary()
```

REGRESSION SET-UP(2)

- This is the correction I made for each sex and sport.
- And these are the code used for the OLS regression

MEN

OLS Regression Results

Dep. Variable:	Win	**R-squared:**	0.011
Model:	OLS	**Adj. R-squared:**	0.010
Method:	Least Squares	**F-statistic:**	8.549
Date:	Sun, 30 Jan 2022	**Prob (F-statistic):**	1.21e-05
Time:	17:28:40	**Log-Likelihood:**	-389.46
No. Observations:	2295	**AIC:**	786.9
Df Residuals:	2291	**BIC:**	809.9
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.2911	0.206	-1.412	0.158	-0.695	0.113
Age	0.0065	0.001	4.743	0.000	0.004	0.009
Weight	0.0014	0.002	0.760	0.448	-0.002	0.005
Height	0.0008	0.002	0.522	0.602	-0.002	0.004

Omnibus:	1264.762	**Durbin-Watson:**	1.508
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	6288.098
Skew:	2.785	**Prob(JB):**	0.00
Kurtosis:	8.893	**Cond. No.**	6.15e+03

WOMEN

OLS Regression Results

Dep. Variable:	Win	**R-squared:**	0.022
Model:	OLS	**Adj. R-squared:**	0.021
Method:	Least Squares	**F-statistic:**	22.14
Date:	Sun, 30 Jan 2022	**Prob (F-statistic):**	3.57e-14
Time:	17:28:14	**Log-Likelihood:**	-146.30
No. Observations:	2918	**AIC:**	300.6
Df Residuals:	2914	**BIC:**	324.5
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.8380	0.169	-4.962	0.000	-1.169	-0.507
Age	0.0058	0.001	5.279	0.000	0.004	0.008
Weight	0.0009	0.001	0.703	0.482	-0.002	0.003
Height	0.0039	0.001	3.006	0.003	0.001	0.006

Omnibus:	1894.144	**Durbin-Watson:**	1.744
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	14206.977
Skew:	3.218	**Prob(JB):**	0.00
Kurtosis:	11.684	**Cond. No.**	6.99e+03

Results Cross Country Skiing

- As we can see for man, weight and height seems to not be statistically significant this can be due to small sample (2500 observation each) .
- R-squared and F-test are not very encuraging on the results.
- Probably a non linear regression would give better results

Results Gymnastics

- Except for the age the other variable are statistically significant
- As we expected smaller people are more likely to win a medal (coefficient negative on weight and height)
- The smaller you are the better

MEN

OLS Regression Results

Dep. Variable:	Win	**R-squared:**	0.009
Model:	OLS	**Adj. R-squared:**	0.008
Method:	Least Squares	**F-statistic:**	15.40
Date:	Sun, 30 Jan 2022	**Prob (F-statistic):**	5.67e-10
Time:	17:30:56	**Log-Likelihood:**	-116.19
No. Observations:	5313	**AIC:**	240.4
Df Residuals:	5309	**BIC:**	266.7
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.6951	0.129	5.372	0.000	0.441	0.949
Age	-0.0011	0.001	-1.023	0.306	-0.003	0.001
Weight	-0.0015	0.001	-1.688	0.092	-0.003	0.000
Height	-0.0030	0.001	-3.131	0.002	-0.005	-0.001

Omnibus:	3673.557	**Durbin-Watson:**	1.742
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	32838.161
Skew:	3.450	**Prob(JB):**	0.00
Kurtosis:	13.037	**Cond. No.**	6.85e+03

WOMEN

OLS Regression Results

Dep. Variable:	Win	**R-squared:**	0.008
Model:	OLS	**Adj. R-squared:**	0.008
Method:	Least Squares	**F-statistic:**	10.84
Date:	Sun, 30 Jan 2022	**Prob (F-statistic):**	4.35e-07
Time:	17:31:25	**Log-Likelihood:**	-171.85
No. Observations:	3881	**AIC:**	351.7
Df Residuals:	3877	**BIC:**	376.7
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.3035	0.112	2.703	0.007	0.083	0.524
Age	0.0015	0.002	1.001	0.317	-0.001	0.004
Weight	-0.0031	0.001	-3.055	0.002	-0.005	-0.001
Height	-0.0008	0.001	-0.883	0.377	-0.003	0.001

Omnibus:	2608.201	**Durbin-Watson:**	1.561
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	21328.329
Skew:	3.347	**Prob(JB):**	0.00
Kurtosis:	12.331	**Cond. No.**	4.46e+03

MEN

OLS Regression Results

Dep. Variable:	Win	**R-squared:**	0.033
Model:	OLS	**Adj. R-squared:**	0.032
Method:	Least Squares	**F-statistic:**	67.15
Date:	Sun, 30 Jan 2022	**Prob (F-statistic):**	1.06e-42
Time:	17:40:52	**Log-Likelihood:**	-1886.5
No. Observations:	5993	**AIC:**	3781.
Df Residuals:	5989	**BIC:**	3808.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.0801	0.123	-8.782	0.000	-1.321	-0.839
Age	0.0039	0.001	3.475	0.001	0.002	0.006
Weight	0.0038	0.001	3.907	0.000	0.002	0.006
Height	0.0052	0.001	5.586	0.000	0.003	0.007

Omnibus:	2223.010	**Durbin-Watson:**	1.396
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	6003.746
Skew:	2.080	**Prob(JB):**	0.00
Kurtosis:	5.597	**Cond. No.**	5.32e+03

WOMEN

OLS Regression Results

Dep. Variable:	Win	**R-squared:**	0.041
Model:	OLS	**Adj. R-squared:**	0.040
Method:	Least Squares	**F-statistic:**	102.5
Date:	Sun, 30 Jan 2022	**Prob (F-statistic):**	5.30e-65
Time:	17:40:31	**Log-Likelihood:**	-1820.6
No. Observations:	7224	**AIC:**	3649.
Df Residuals:	7220	**BIC:**	3677.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.8752	0.107	-8.202	0.000	-1.084	-0.666
Age	0.0024	0.001	2.197	0.028	0.000	0.005
Weight	0.0053	0.001	7.580	0.000	0.004	0.007
Height	0.0028	0.001	3.625	0.000	0.001	0.004

Omnibus:	3070.015	**Durbin-Watson:**	1.415
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	10083.866
Skew:	2.278	**Prob(JB):**	0.00
Kurtosis:	6.569	**Cond. No.**	5.94e+03

Results Swimming

- All coefficient are statistically significant
- F-statistics are high and encouraging
- The hight and weight have a posive impact on the probability of winning a medal this suggest that taller people are in advantage wrt to the other.
- Although i think that this relationship is non-linear

MEN

OLS Regression Results

Dep. Variable:	Win	**R-squared:**	0.006
Model:	OLS	**Adj. R-squared:**	0.006
Method:	Least Squares	**F-statistic:**	22.79
Date:	Sun, 30 Jan 2022	**Prob (F-statistic):**	1.05e-14
Time:	17:42:39	**Log-Likelihood:**	-1040.9
No. Observations:	11189	**AIC:**	2090.
Df Residuals:	11185	**BIC:**	2119.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.2283	0.067	-3.399	0.001	-0.360	-0.097
Age	-0.0013	0.001	-2.292	0.022	-0.002	-0.000
Weight	0.0004	0.000	2.021	0.043	1.3e-05	0.001
Height	0.0017	0.000	4.107	0.000	0.001	0.003

Omnibus:	6992.686	**Durbin-Watson:**	1.641
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	48270.535
Skew:	3.148	**Prob(JB):**	0.00
Kurtosis:	10.994	**Cond. No.**	5.30e+03

WOMEN

OLS Regression Results

Dep. Variable:	Win	**R-squared:**	0.004
Model:	OLS	**Adj. R-squared:**	0.004
Method:	Least Squares	**F-statistic:**	11.24
Date:	Sun, 30 Jan 2022	**Prob (F-statistic):**	2.35e-07
Time:	17:43:01	**Log-Likelihood:**	-1556.0
No. Observations:	8135	**AIC:**	3120.
Df Residuals:	8131	**BIC:**	3148.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.2771	0.085	-3.243	0.001	-0.445	-0.110
Age	0.0016	0.001	2.316	0.021	0.000	0.003
Weight	0.0004	0.000	1.055	0.291	-0.000	0.001
Height	0.0018	0.001	3.205	0.001	0.001	0.003

Omnibus:	4334.156	**Durbin-Watson:**	1.641
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	20643.256
Skew:	2.740	**Prob(JB):**	0.00
Kurtosis:	8.556	**Cond. No.**	4.78e+03

Results Athletics

- Weight is not statistically significant for the regression.
- This suggest that it should be removed and not taken into consideration
- Fat people tend to be in disadvantage anyway in this sport
- Weirdly age has a negative impact on probability of victory for man but it is positve for women

```
X_train, X_test, y_train, y_test = train_test_split( reg1_m_x, reg1_m_y, test_size=0.2, random_state=4)
print ('Train set:', X_train.shape,  y_train.shape)
print ('Test set:', X_test.shape,  y_test.shape)

LR = LogisticRegression(C=0.01, solver='liblinear').fit(X_train,y_train.values.ravel())

yhat_lr = LR.predict(X_test)
yhat_lr_prob = LR.predict_proba(X_test)

print('JACCARD: ', jaccard_score(y_test, yhat_lr,pos_label=0)) #not winning a medal
print('F1: ', f1_score(y_test, yhat_lr, average='weighted'))
print('LogLoss: ', log_loss(y_test, yhat_lr_prob))
```

MACHINE LEARNING: PREDICTION MODEL

- I suggest a **logistic model** which works very good with binary dependent variable
- I will use Sklearn library of python to run this kind of regression
- If the logloss is low the prediction model is good and can be used to forecast!
- I will use a train sample (80% of the sample) and a train set (20%), obtained randomly through «train_test_split»

LOGISTIC REGRESSION MODEL

- For the sake of brevity I will report some of the scores, the results are similar also for the other models!
- As you can see the model seem to predict very well if an Athlete **won't** win a medal:

ATHLETICS WOMEN

```
Train set: (8951, 221) (8951, 1)
Test set: (2238, 221) (2238, 1)
JACCARD:  0.9186773905272565
F1:  0.8797395039236328
LogLoss:  0.27210371220593615
```

ATHLETICS MEN

```
Train set: (6508, 221) (6508, 1)
Test set: (1627, 221) (1627, 1)
JACCARD:  0.9084204056545789
F1:  0.864827928861493
LogLoss:  0.296838593714039
```

THE LOG LOSS FUNCTION SI LOW

JACCARD AND F1 SCORE ARE HIGH!

GYMNASTIC WOMEN

```
Train set: (3104, 221) (3104, 1)
Test set: (777, 221) (777, 1)
JACCARD:  0.9395109395109396
F1:  0.9102096693337569
LogLoss:  0.2105092609797363
```

GYMNASTIC MEN

```
Train set: (4250, 221) (4250, 1)
Test set: (1063, 221) (1063, 1)
JACCARD:  0.9275634995296331
F1:  0.8927063060382804
LogLoss:  0.253696171549855
```

CONCLUSIONS:

- My purpose as stated earlier is to find, for potential gym institute, a model to understand if somone has the potential to win a medal (binary variable) given age, height, weight, country(binary variable for each country whether an athlete belong to that country).
- I found out from a previous descriptive analysis that there is a difference between sexes they should be tested separately in the regression to avoid bias estimate.
- moreover i found out previously (always by mean of a AB test) that each country has athletes with height and weight different from another country.
- The first challenge that i faced was in the linear regression, it seemed that t-statistics for country variables is not statistically significat therefore i remove them and considered solely the following variable: age, weight, height and the intercepet.
- i run 4 regression (linear OLS) for each sex, in every regression i used 4 subsample of athletes, that is: Cross country skii , gymnastic, swimming, athletics athletes.
- the most interesting findings are:
- for gymnastic weight and height coefficient are negative: this is in line with the idea that the smaller you are the better you perform, both for male and female. althought the p-value for the Weight variable is close to 0.95 percentile. age coef is positive (male coef: 0.0015); this is a difficult sport and experince helps improving the performance up to a certain age.
- similar result for cross country skiing; age, height, weight have negative coefficient. p-value of height is statistically significant.
- In swimming all variable are positive and statistically significant!; eg for male the coefficient of weight and height are: 0.0053 and 0.0028, the taller you are the easier is to win.
- In atheltics all variable are statistically significant for both group, variable age for man seems to have a negative impact on the probability for the athelete to win a medal (-0.0013) for women is positive (0.0016)
- I estimated also a ***logistic regression*** which is very usefull when the dependent variable is binary. I did on the full sample for each of the 8 cases listed above, using the sklearn library and training test and test set (the latter is 20% of the sample). It turned out that the model is very good in predicting if an athlete is not going to win a medal! LOG LOSS value in all the cases is close to 0.3 and jaccard and f1 score is close to 0.9!
- **CODES: https://github.com/LeonSavi/Projects_Leonardo_Saviane.git**