Sportstat Analysis Report

Tianjiao Yang

07.28.2023

Table of Contents

Preparation of Proposal

- Client/Dataset Selection
  - Dataset: SportsStats
  - Why: This is Olympics Dataset with 120 years' records. This dataset listed in details of athletes and events information. From this dataset, we could perform serials of analysis to find patterns and trends related to some specific games or discover insights inside of athletes. This may help teams or coaches to improve their trainings or adjust budgets.
  - Potential audience: sport teams, coaches, sport firms, personal trainers

Preparation of Proposal

- Data Import and Cleaning
  - Import the csv file using pandas library in python

- Data cleaning
  - There are a lot missing value in age, height, weight, medal variables. Since in some cases, age height, weight may not be critical to determine whether a medal is received, and of course not receiving medals should be the case of most athletes. Thus, we keep these missing values for exploratory data analysis, we may delete or substitute them in future analysis

Preparation of Proposal

- Data Exploration
  - Perform exploratory data analysis to identify the basic properties of dataset. For example, we check the distribution, unique values and counts of each variable.

```python
for column in athlete_events.columns:
    print("Variable:", column)
    print(athlete_events[column].value_counts())
    print()
```

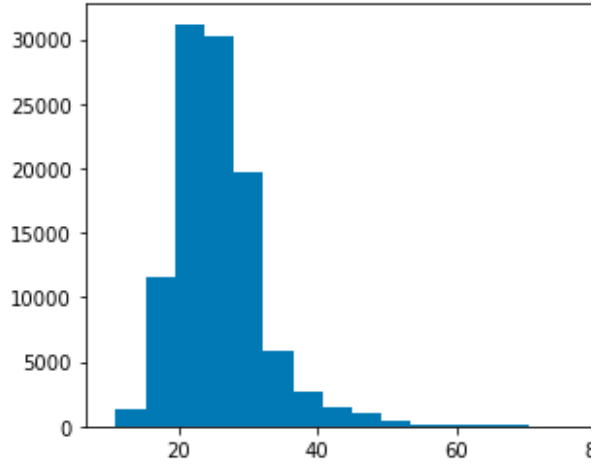
```
Variable: Sex
M    80956
F    28878
Name: Sex, dtype: int64

Variable: Age
23.0    8860
24.0    8748
22.0    8340
25.0    8026
21.0    7848
        ...
88.0       3
73.0       3
96.0       1
84.0       1
75.0       1
Name: Age, Length: 69, dtype: int64
```

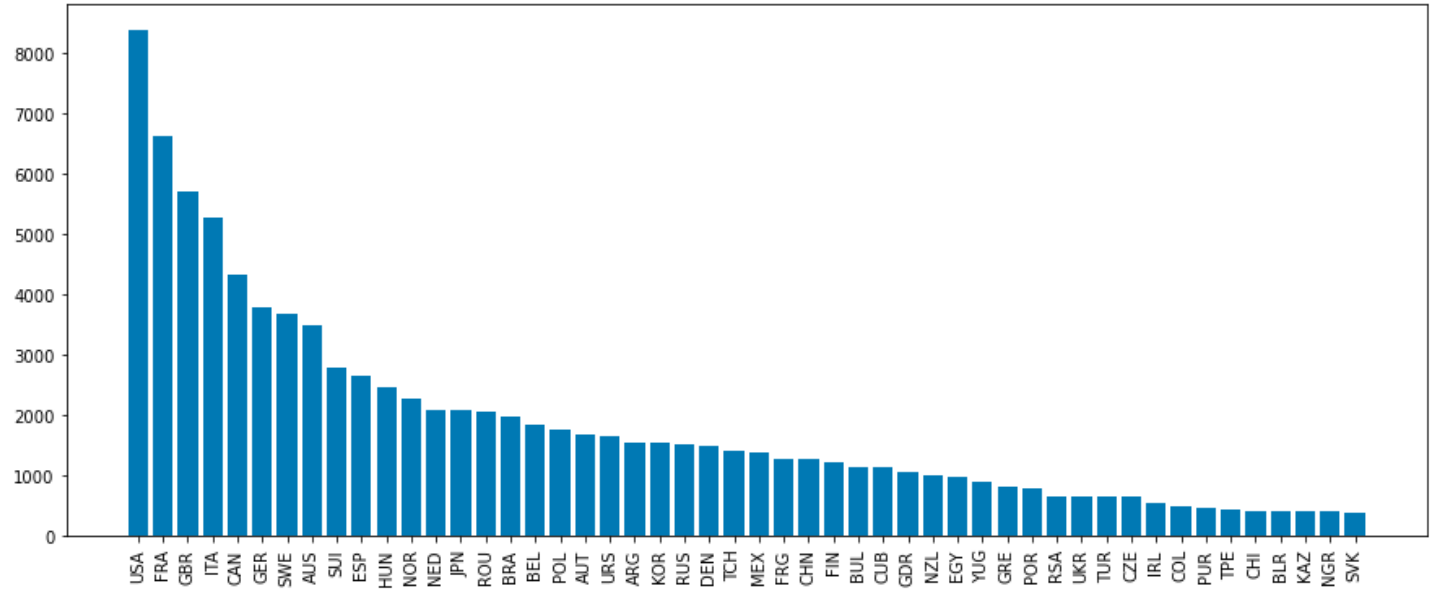
Preparation of Proposal

- we plot the histograms and bar charts to see the distributions of variables.

```
plt.hist(athlete_events['Age'], bins=20)
plt.show()
```

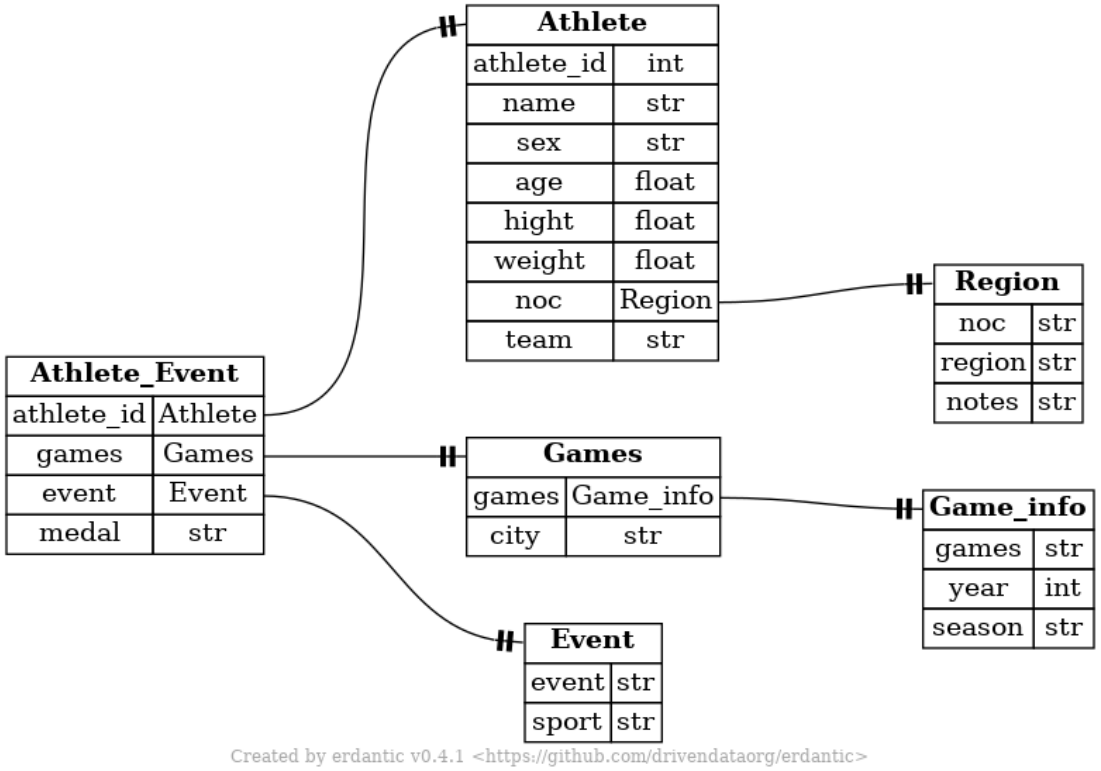



```
plt.xticks(rotation=90)
plt.show()
```



Preparation of Proposal

- Entity Relationship Diagram (ERD)



Development of Proposal

- This project is aim to find patterns and trends between different variables and the medal receipt.
- Sports teams and coaches would be interested in these findings to help adjust their training and budget.
- My audience could be SportsStat firms, sports teams, coaches or even independent trainer who would like to find patterns and trends based on their historical performance and get actionable recommendations for training or business purpose.

Development of Proposal

- Potential Questions

1. From which region did the athletes most come from?
2. Do men or women have higher rate of winning a medal?
3. What are most popular games among Olympic Games?
4. What could be key factors contribute to winning a medal?

Development of Proposal

- Hypothesis

1. There is no difference among genders in the chance of winning a medal
2. Athletes come from different regions evenly
3. There is no difference among NOCs at chances of medal acquisition

4. Number of events are increasing by time

5. Younger people (age<=25) have higher rate of medal acquisition than older people (age>25)

Development of Proposal

- Approach
  - Hypothesis 1

Build metric containing gender, number of medals and perform A/B testing to see if there is a difference of the chance of winning a medal

  - Hypothesis 2

Identify the distribution of athletes' regions and find out if they are evenly distributed

  - Hypothesis 3

Build metric containing NOC and number of medals, compare the chance of winning a medal among different NOC.

  - Hypothesis 4

Identify the number of events each year and make a plot to see if there is a change in the number by time

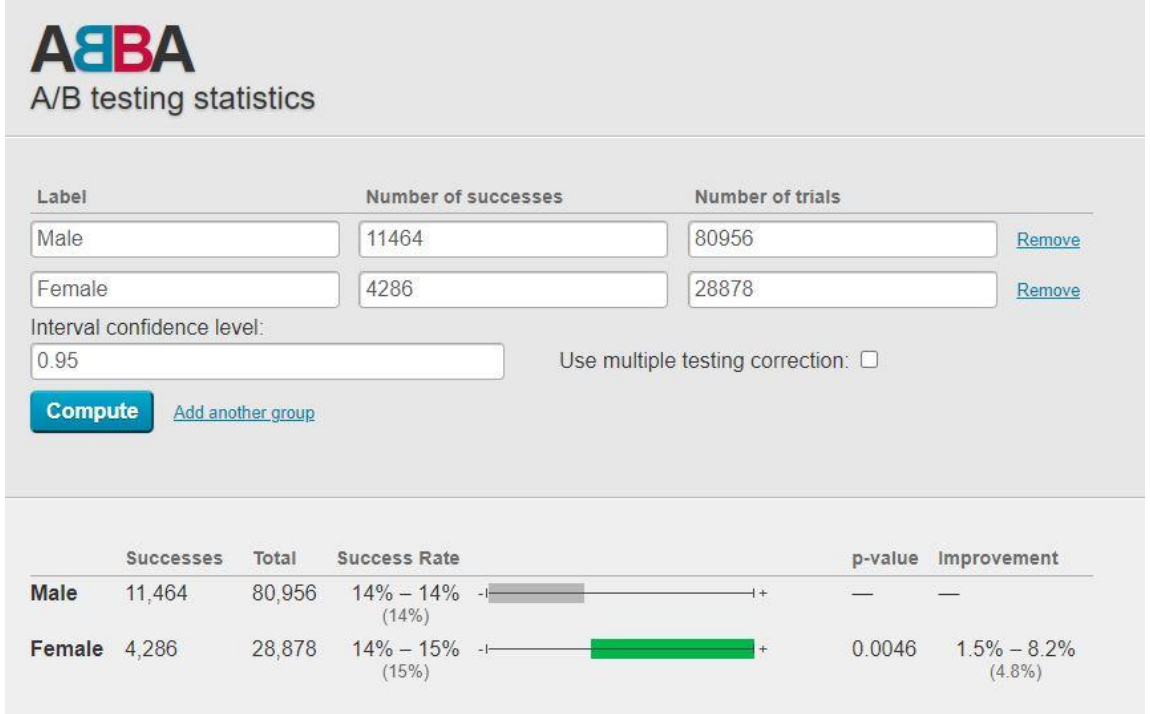
  - Hypothesis 5

Build metric containing age and number of medals earned by each age group (<=25, and >25), then see if there is a difference of number of medals earned and more possibly chances of winning a medal among different age group

Descriptive Analysis and Hypothesis Testing

- Analysis for gender

	Sex	eventnumber	medalnumber
0	F	28878	4286
1	M	80956	11464



- Conclusion:

AB test result showed that female has higher chance of winning a medal (14%) than male (15%) at a p-value of 0.0046

Descriptive Analysis and Hypothesis Testing

- Region analysis 1-Test if athletes come from diffrent regions evenly.

Table 1. numbers of athletes from different regions

	athletes	region
0	4381	USA
1	3227	France
2	2940	UK
3	2901	Germany
4	2451	Italy
...	...	...
201	3	Kosovo
202	2	Burkina Faso
203	1	Kiribati
204	1	Lesotho
205	1	South Sudan

Table 2. top 10 regions having most athletes

	athletes	region
0	4381	USA
1	3227	France
2	2940	UK
3	2901	Germany
4	2451	Italy
5	2128	Canada
6	1809	Australia
7	1723	Sweden
8	1653	Russia
9	1314	Spain

Descriptive Analysis and Hypothesis Testing

- Region analysis 1

- Conclusion

1. Apparently, numbers of athletes from different regions are very deferent.
2. We could see the top 10 regions are from Europe, North america, and Australia.
3. None of regions in Asia nor Africa has population of athletes ranked in top 10.
4. From this we could conclude that Olympic games are not as popular in north America, Europe, and Australia compared to Asia and Africa
5. This could be caused by the economic, political or other reason or the habit of people regarding sports.

Descriptive Analysis and Hypothesis Testing

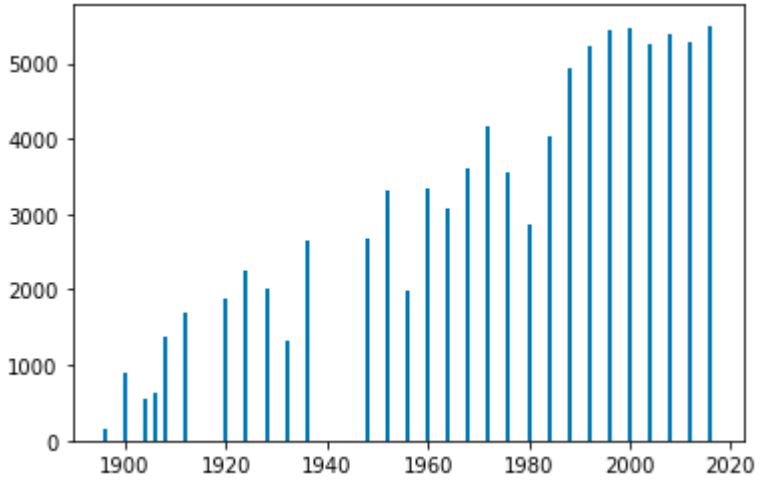
- Region analysis 2 -Test if athletes coming from different regions have the same chance of medal acquisition.
- Conclusion
  - Different regions have different rate of winning medals
  - Russia and USA athletes are at leading position to win medals with a chance of higher than 30% to win a medal.
  - This may because that Russia and USA have more advanced techniques and training system.
  - More over they may have better culture history regarding Olympic games
  - Economics factors may also influence the result
  - Russia and USA could have more experiences too.

	region	medalnumber	rate
0	Russia	1116	0.326125
1	USA	2542	0.302944
2	Pakistan	63	0.238636
3	Germany	1453	0.235685
4	Norway	482	0.210756
5	Jamaica	74	0.198925
6	China	269	0.180295
7	Netherlands	371	0.178023
8	Sweden	650	0.176919
9	Denmark	264	0.176825
10	UK	987	0.173219
11	Australia	607	0.171324
12	Hungary	410	0.165858
13	Paraguay	14	0.162791
14	Serbia	174	0.156616
15	Cuba	178	0.156415
16	Croatia	49	0.155556
17	Ethiopia	30	0.152284
18	Italy	789	0.149460
19	South Korea	227	0.146357

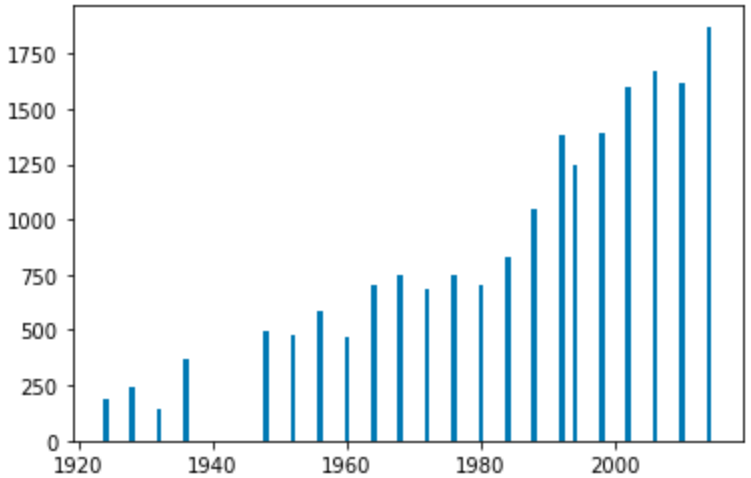
Descriptive Analysis and Hypothesis Testing

- Number of events analysis

Summer season events number



Winter season events number



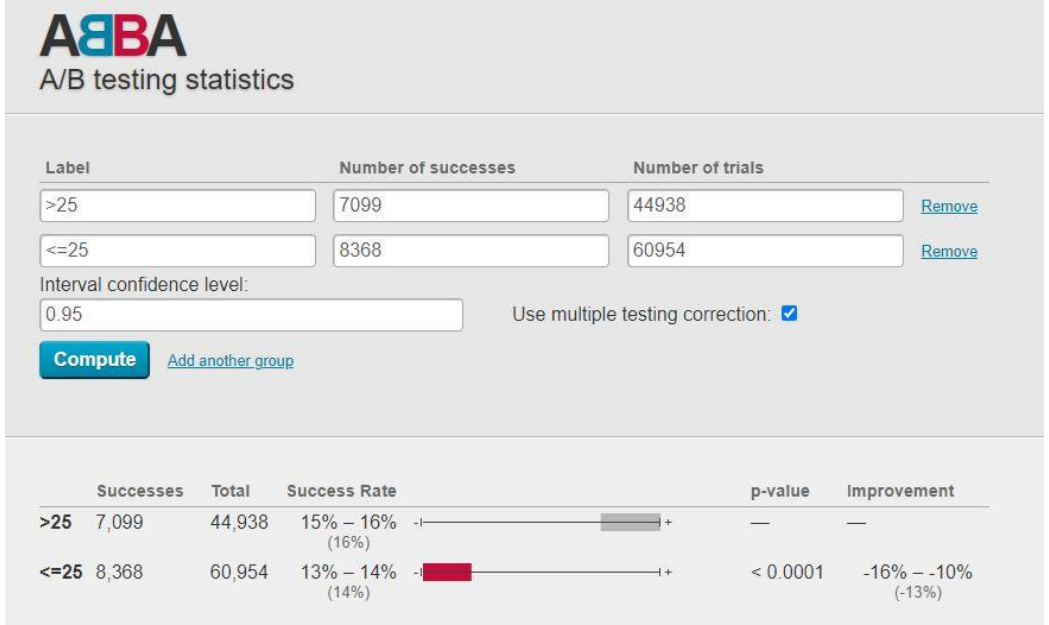
- Conclusion
  - summer season tends to have much more events than winter season
  - For summer season, number of events keeps increasing until 1992, after that number of events kept at similar level at aroud 5300
  - For winter season, number of events increases by time

Descriptive Analysis and Hypothesis Testing

- Age group analysis

Test if younger people (age<=25) have higher rate of winning medals than older people (age>25)

	age_25	eventnumber	medalnumber
0	Oder than 25	44938	7099
1	Undefined	3942	283
2	Younger than or equal to 25	60954	8368



- Conclusion

Athletes younger or at age of 25 have a lower rate of winning a medal (14%) than athletes older than 25 (16%), this could come from the reason that older athletes would have more experience than younger athletes

Deeper Investigation

- Exploring the Relationship between Athletes' Physical Conditions and Performance Achievements

Conduct a comprehensive analysis on the association between athletes' physical attributes, including age, gender, height, weight, and the outcome of medal acquisition

```
# subset selection
subdf=pysqldf("select id, sex, age, height, weight, medal from athlete_events;")
```

```
# eliminate nulls except for medal
subdf_c=pysqldf("select * from subdf where sex is not null and age is not null and height is not null and weight is not null; ")
```

Deeper Investigation

- Transform categorical variables into integers for future logistic regression

	ID	Sex	Age	Height	Weight	Medal	Sex_c	Medal_c
0	1	M	24.0	180.0	80.0	None	2	0
1	2	M	23.0	170.0	60.0	None	2	0
2	5	F	21.0	185.0	82.0	None	1	0
3	5	F	21.0	185.0	82.0	None	1	0
4	5	F	25.0	185.0	82.0	None	1	0
...	...	...	...	...	...	...	...	...
82248	55526	M	24.0	180.0	62.0	None	2	0
82249	55526	M	28.0	180.0	62.0	None	2	0
82250	55527	M	27.0	175.0	61.0	None	2	0
82251	55533	M	29.0	194.0	80.0	None	2	0
82252	55533	M	29.0	194.0	80.0	None	2	0

Deeper Investigation

- Perform logistic regression to discover the relationship between age, gender, height, weight and medal acquisition

```
X=subdf_c.drop(['ID','Sex','Medal','Medal_c'], axis=1)
y=subdf_c['Medal_c']
```

```
X_train,X_test, y_train, y_test=train_test_split(X, y, test_size=0.2, random_state=30)
```

```
logreg=LogisticRegression()
logreg.fit(X_train, y_train)
```

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=None, solver='lbfgs', tol=0.0001, verbose=0,
                   warm_start=False)
```

```
y_pred=logreg.predict(X_test)
```

Deeper Investigation

- Check p-value to see the significance level of independent variables

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                Medal_c   No. Observations:                65802
Model:                          Logit   Df Residuals:                    65797
Method:                           MLE   Df Model:                            4
Date:                Fri, 28 Jul 2023   Pseudo R-squ.:                 0.01942
Time:                        03:43:38   Log-Likelihood:                -26190.
converged:                       True   LL-Null:                       -26709.
Covariance Type:            nonrobust   LLR p-value:                3.156e-223
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -6.3546      0.250    -25.450      0.000      -6.844      -5.865
Age            0.0072      0.002      3.459      0.001       0.003       0.011
Height         0.0260      0.002     15.028      0.000       0.023       0.029
Weight         0.0112      0.001      9.093      0.000       0.009       0.014
Sex_c         -0.6247      0.029    -21.691      0.000      -0.681      -0.568
==============================================================================
```

Deeper Investigation

- Conclusion

1. The age, height, weight, and gender of athletes play a crucial role in predicting medal-winning performances.
2. Age, height, and weight exhibit a positive correlation with medal-winning outcomes, which is understandable given that advancing age often accompanies increased strength, skills, and experience. Similarly, height and weight positively impact attributes such as strength, speed, and other physiological responses.
3. Interestingly, gender displays an inverse relationship with medal-winning success. This finding corroborates our previous analysis, which suggested that female athletes appear to have a higher likelihood of achieving medal-worthy performances compared to their male counterparts.

Conclusion and Recommendation

- Conclusion

1. Females exhibit a higher likelihood of attaining medals in various sporting events.
2. Disparities in both the quantity of athletes and the probability of medal acquisition are evidently discernible across diverse regions.
3. Athletes aged 25 and above demonstrate an increased probability of securing medals.
4. The number of events during the summer season experienced consistent growth until 1992, after which it stabilized at its peak level. Conversely, the winter season witnessed a continual expansion of events over time.

Conclusion and Recommendation

- Recommendation

1. Enhancing athletes' years of experience can significantly bolster their prospects of medal acquisition.
2. Establishing comprehensive training programs in North America or Europe offers athletes access to superior resources, fostering the refinement of their strengths and skills.
3. In order to foster the advancement of sports on a global scale, it is imperative to allocate greater attention to the development of sports in Asia and Africa regions, providing guidance and assistance to enhance their sporting infrastructure and overall performance.