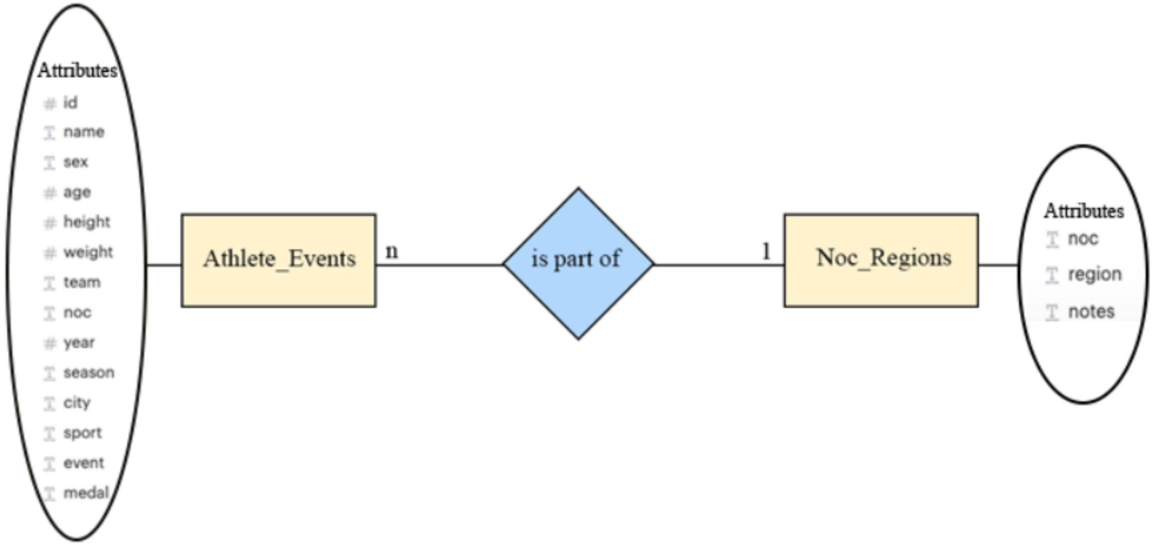
Olympics Data Set

About

- We are trying to draw insights from datasets containing info on Olympic athletes in the last 120 years
- Presented to: SportStat US client trying to understand how to improve winter sport participation and success.
- Hypothesis:
    - 1. Athletes for certain sports should fall within an optimal physical characteristics range for age, weight, and height.
    - 2. Certain countries have advantages in winter sports due to their climates.
    - 3. More sport participation leads to more medals

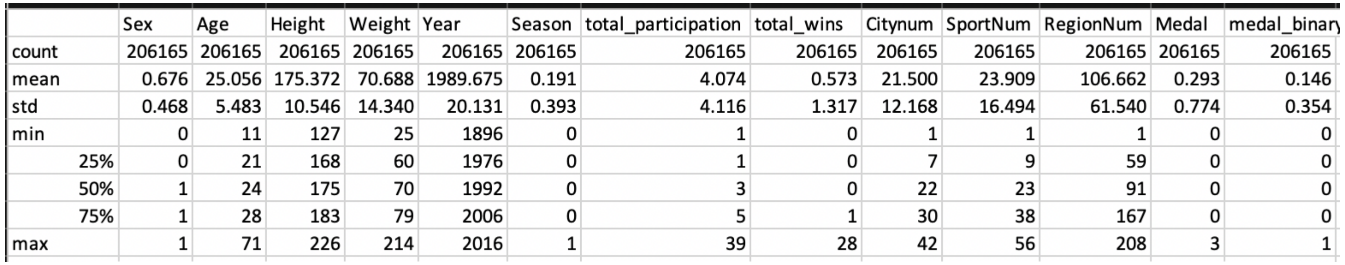
ER Diagram


Attributes
id
name
sex
age
height
weight
team
noc
year
season
city
sport
event
medal
Athlete_Events
n
is part of
1
Noc_Regions
Attributes
noc
region
notes

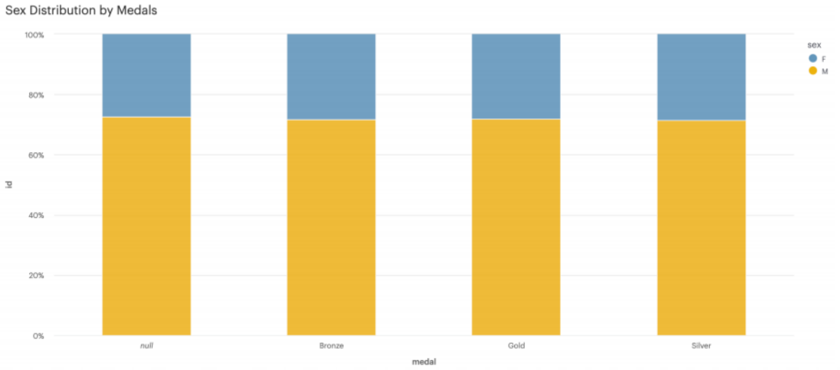
Data Cleaning

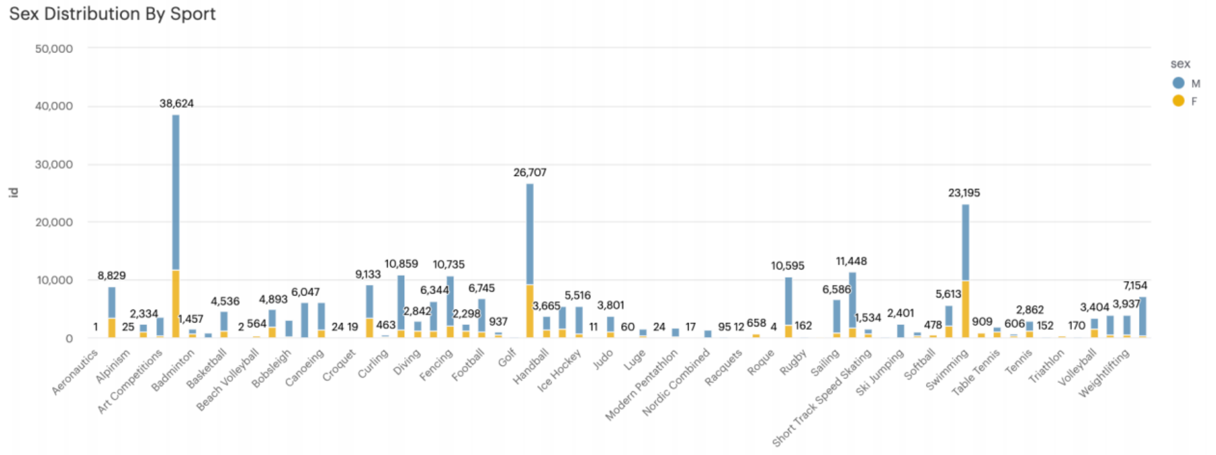
1. Joining NOC and athlete information
2. Making sure NOC and regions are consistent
3. Removing nulls in age, height, and weight since they don't greatly impact data
4. Changing qualitative information to numerical classifiers: Sex, Sport, Region, Medal, Season
5. Adding metrics: medal_binary, total_participation, total_wins
6. Dropping unused columns: notes, Games, NOC, Team, Event

Data Exploration



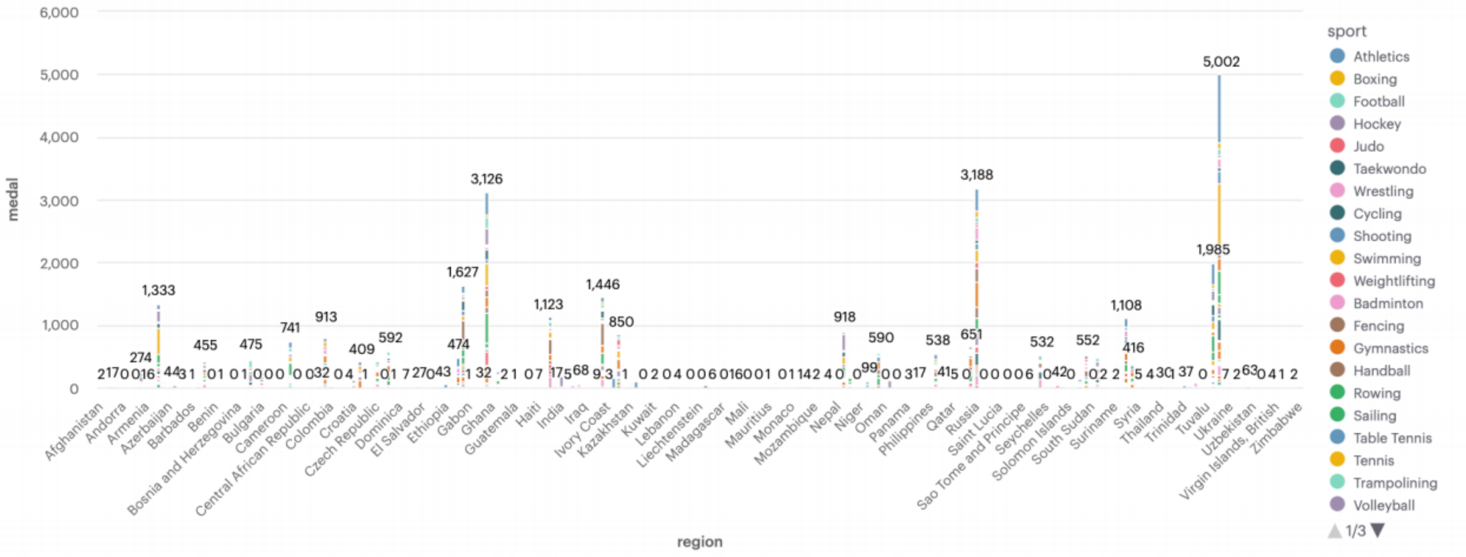
	Sex	Age	Height	Weight	Year	Season	total_participation	total_wins	Citynum	SportNum	RegionNum	Medal	medal_binary
count	206165	206165	206165	206165	206165	206165	206165	206165	206165	206165	206165	206165	206165
mean	0.676	25.056	175.372	70.688	1989.675	0.191	4.074	0.573	21.500	23.909	106.662	0.293	0.146
std	0.468	5.483	10.546	14.340	20.131	0.393	4.116	1.317	12.168	16.494	61.540	0.774	0.354
min	0	11	127	25	1896	0	1	0	1	1	1	0	0
25%	0	21	168	60	1976	0	1	0	7	9	59	0	0
50%	1	24	175	70	1992	0	3	0	22	23	91	0	0
75%	1	28	183	79	2006	0	5	1	30	38	167	0	0
max	1	71	226	214	2016	1	39	28	42	56	208	3	1



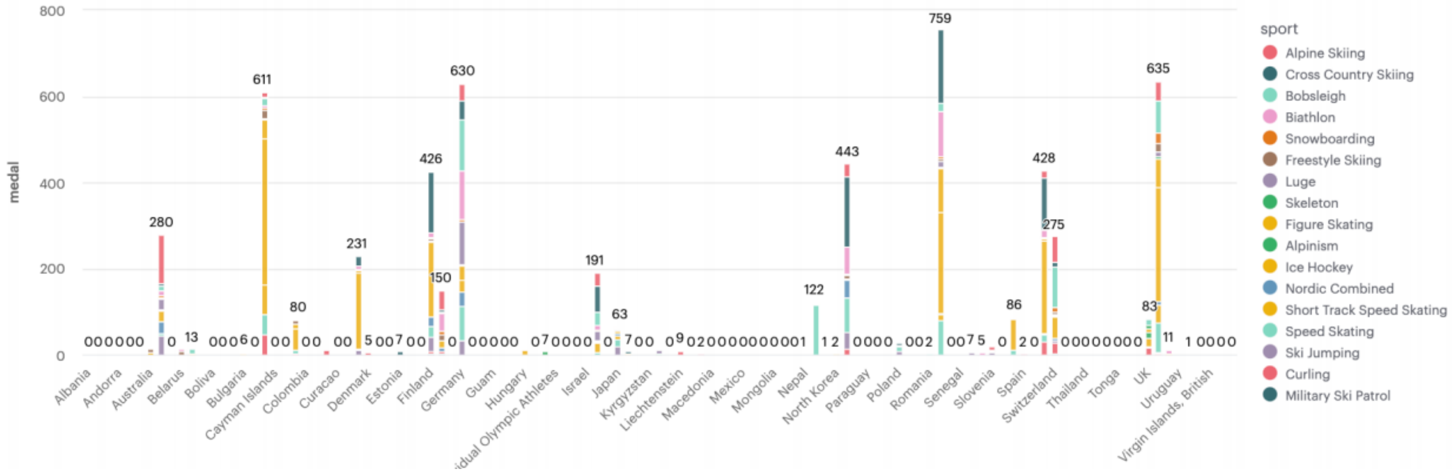


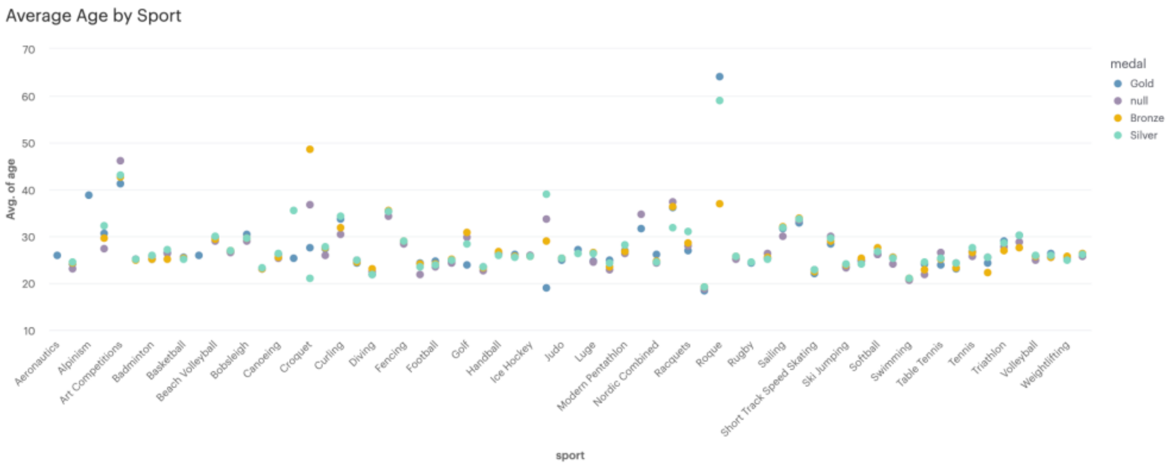
M:F Ratio is about equal

Summer Global Medal Distribution

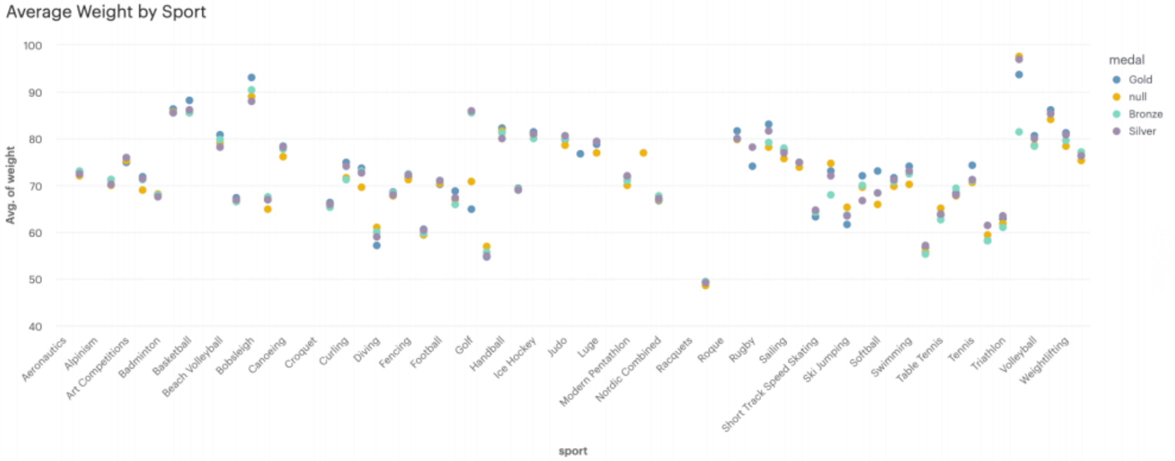


Winter Global Medal Distribution



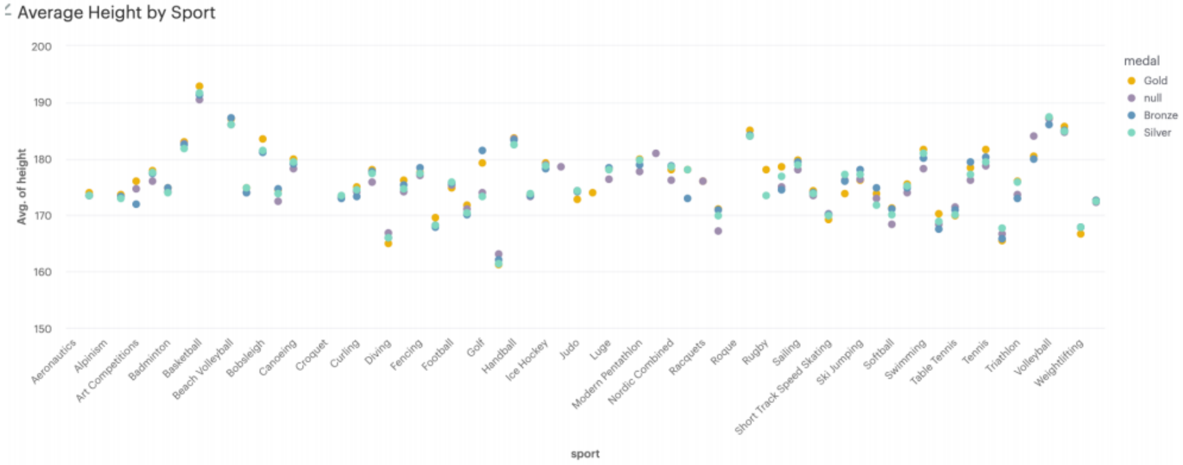


Average Age by Sport

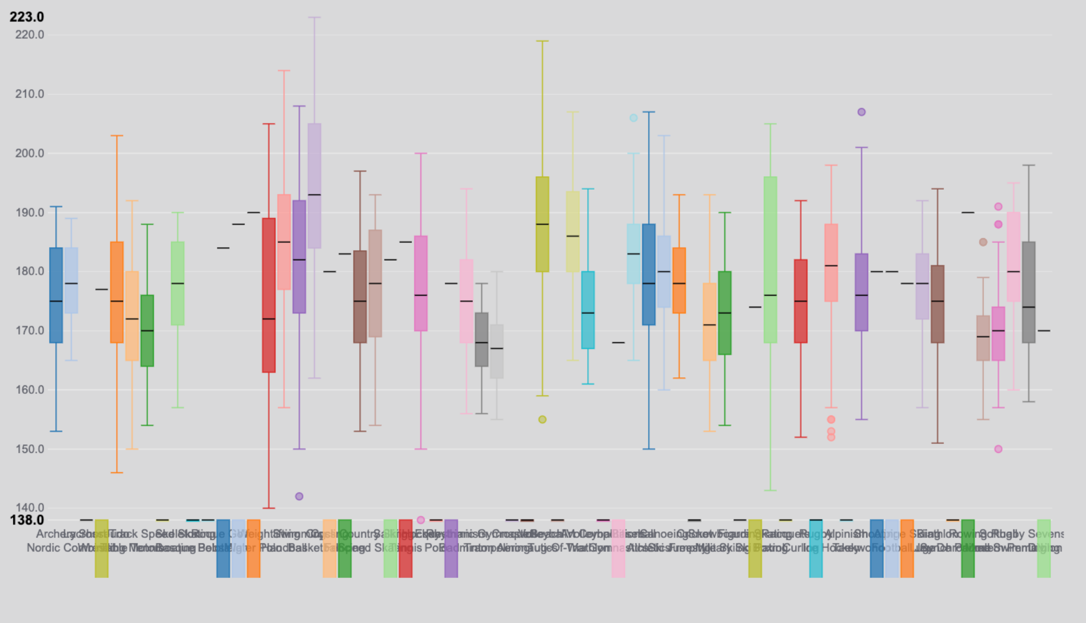


Average Weight by Sport

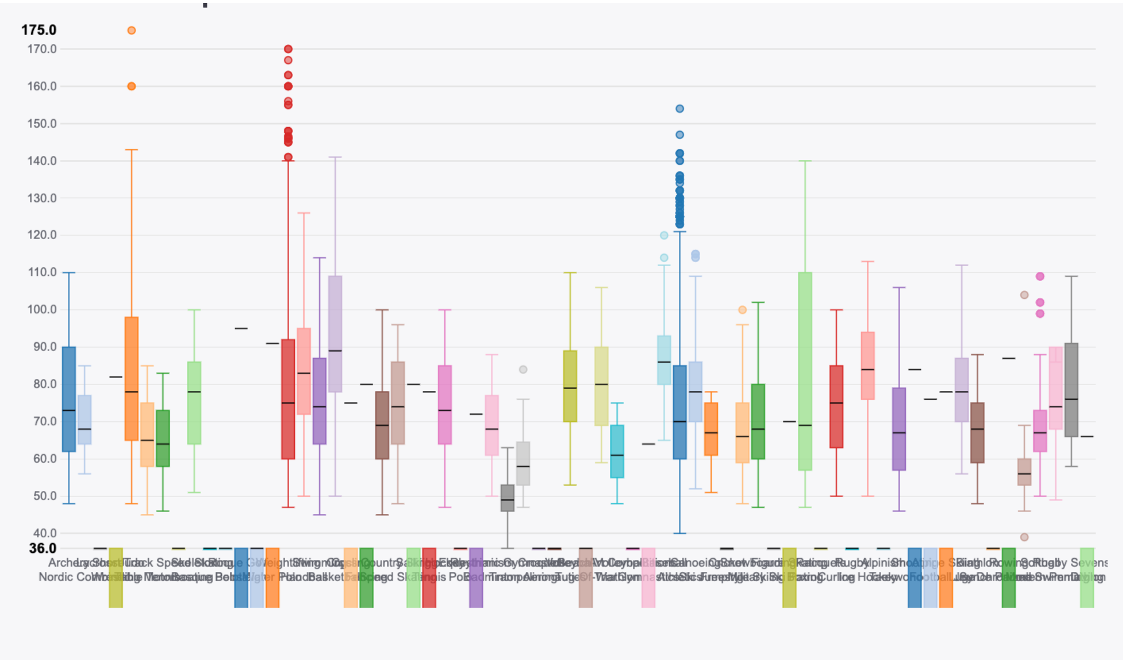
Age, height, and weight seem to be clustered around certain points for each sport.



Average Height by Sport

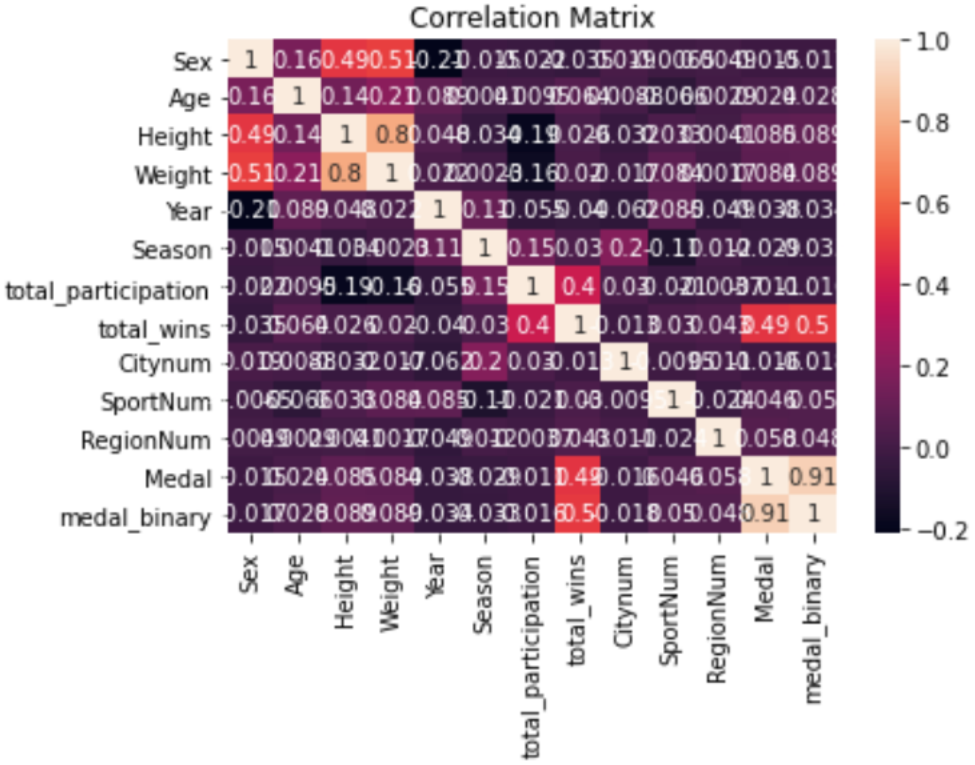


Heights of Winners by Sports Box Plot



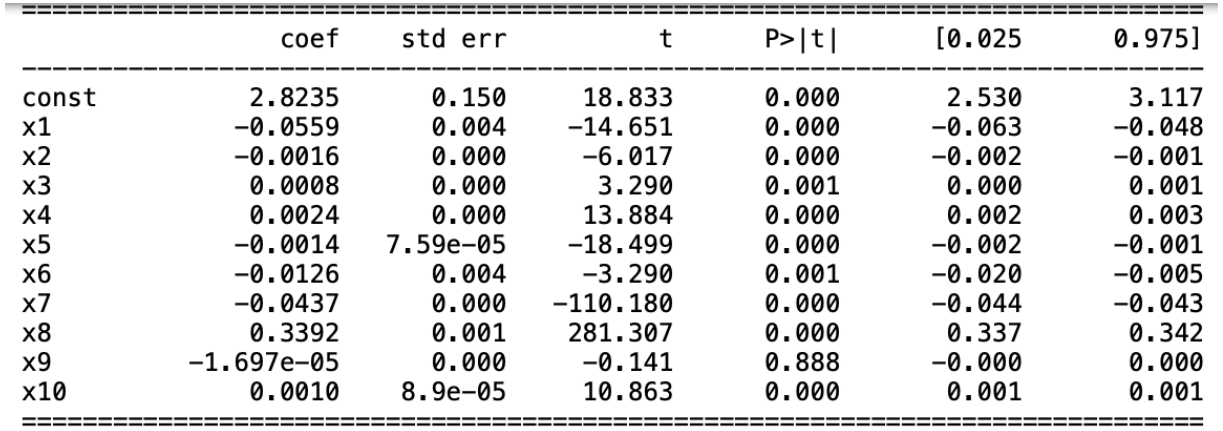
Weights of Winners by Sports Box Plot

Correlations



Linear Regression Model

- Base RMSE: 0.772
- New model: 0.567
- Using a LR model predicting the Medal attribute, and all the variables from Sex to RegionNum, found that only SportNum was not significant factor in finding Medal.
- CityNum had the greatest coefficient



```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.8235      0.150     18.833      0.000       2.530       3.117
x1            -0.0559      0.004    -14.651      0.000      -0.063      -0.048
x2            -0.0016      0.000     -6.017      0.000      -0.002      -0.001
x3             0.0008      0.000      3.290      0.001       0.000       0.001
x4             0.0024      0.000     13.884      0.000       0.002       0.003
x5            -0.0014   7.59e-05    -18.499      0.000      -0.002      -0.001
x6            -0.0126      0.004     -3.290      0.001      -0.020      -0.005
x7            -0.0437      0.000   -110.180      0.000      -0.044      -0.043
x8             0.3392      0.001    281.307      0.000       0.337       0.342
x9         -1.697e-05      0.000     -0.141      0.888      -0.000       0.000
x10            0.0010    8.9e-05     10.863      0.000       0.001       0.001
==============================================================================
```

Limitations

- Missing variables
- Low R^2 value
- More specific research and studies necessary to make meaningful suggestions

Conclusions

- Hypothesis & Suggestions:
  - 1. Athletes for certain sports should fall within an optimal physical characteristics range for age, weight, and height.
    - This is true, and our sport client should select & train their athletes to fall into the ideal range to improve chances of winning.
    - Can be supplemented by further research into specific sport and working with nutritionists and trainers.
  - 2. Certain countries have advantages in winter sports due to their climates.
    - Is true, especially for winter sports. Can consider hiring trainers from winter sports powers like Canada and Finland or sending athletes abroad.
    - Should study the correlation between home game country & participant's country
  - 3. More sport participation leads to more medals
    - Not really correlated to medals, and low coefficient in LR.
    - Athletes should focus on maximizing their physical ability for one event.