# Data Analysis

## SportsStats

By: Kay Royo

### I. Objective

The main objective of this project is to provide an extensive data analysis for the client, *SportsStats,* by extracting valuable insights from the historical Olympics game dataset. The historical Olympics Games dataset is analyzed to provide useful insights to the client's partners, such as local news and personal trainers. Specifically, this dataset is used to identify patterns and trends that highlight certain groups, events, countries, and more for the purpose of developing a news story or discovering key health insights. This data analysis project uses the historical Olympic Games dataset containing 120 years of data from Athens 1896 to Rio 2016. This dataset contains information about the different Olympics Games held in over 200 National Olympics Committees (NOC) regions such as the professional athletes who participated in each game and the different events that were part of each game that are listed below. Another supplemental dataset that is used for this project contains information about the different NOC regions. A dataset containing information about the different host cities where the different Olympic Games were held in the past is also used. This project focuses on exploring how the Olympics Games have changed over the years, which can be used to predict its future outlook and outcome. The prediction of the future of the Olympics Games might be useful for the audiences that this project aims to target including media outlets, advertisers, athletes, prospective sponsors, and the general public interested in the Olympics games.

### II. Hypotheses

- **Assumption 1:** There is a significant difference between the performance of home and away teams.
  - Home teams are more familiar with the playing venue, have less travel time, and more psychological support from the home fans.
- **Assumption 2:** Athlete's age, height, and weight play an important role in athlete performance.
  - Aerobic (or endurance) athletic performance declines with age since the way our body uses oxygen changes as we get older.
  - Reduced body fat improves endurance, speed, and agility.
  - Taller athletes perceive distances better than shorter athletes.
- **Assumption 3:** Male acquire more medals in some sports than female athletes.
  - Since biologically men have physical advantage over women in sports, then there should be more male medalists than female medalists.

## III. Approach

- The average number of medals is used as a metric to measure performance for the following since there are different numbers of participants in each group. The average number of gold medals is also used as an additional performance metric to ensure that we see the results based on the best possible performance in the Olympics which is represented by a gold medal.
  - Host versus visiting teams' performance comparison
  - Male versus female performance in different sports
- The total number of medals (metric 1) and gold medals (metric 2) are also used for the following.
  - Relationship of age, height, and weight to athlete performance

## IV. Technical Challenges

- The original dataset has a significant number of null values in the Age, Height, and Weight column which was replaced with an average value for each column.
- The original dataset contains duplicate rows that needed to be removed.
- The original datasets did not include some regions.
- There are literal NA strings in the data that were converted to null values.
- There are some limitations of pandasql (SQLite) that make some of the SQL queries somewhat slower to execute but still manageable.

## V. Data Exploration

- **Descriptive Statistics**
  - Number of distinct items in each column in the data

| IDs | Names | Sex | Teams | NOCs | Regions | Sports | Events | Games | Cities | Countries | Years | Season | Medals |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 135571 | 134732 | 2 | 1184 | 230 | 207 | 66 | 765 | 51 | 42 | 23 | 35 | 2 | 3 |

  - Age, height, and weight of male and female athletes

| | **Age** | | **Height** (cm) | | **Weight** (kg) | |
|---|---|---|---|---|---|---|
| | Male | Female | Male | Female | Male | Female |
| **Min** | 10 | 11 | 127 | 127 | 28 | 25 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Max** | 97 | 74 | 226 | 213 | 214 | 167 |
| **Median** | 25 | 23 | 175 | 167 | 69 | 58 |
| **Mean** | 26 | 24 | 179 | 168 | 76 | 60 |
| **25th percentile** | 22 | 20 | 172 | 162 | 67 | 54 |
| **50th percentile** | 25 | 23 | 179 | 168 | 74 | 59 |
| **75th percentile** | 29 | 27 | 185 | 173 | 83 | 65 |
| **Standard deviation** | 6 | 6 | 9 | 9 | 13 | 10 |

- Number of medals per game (host vs. visiting teams)

| Average medals | home team > away team | home team < away |
|---|---|---|
| Number of games | 41 | 10 |

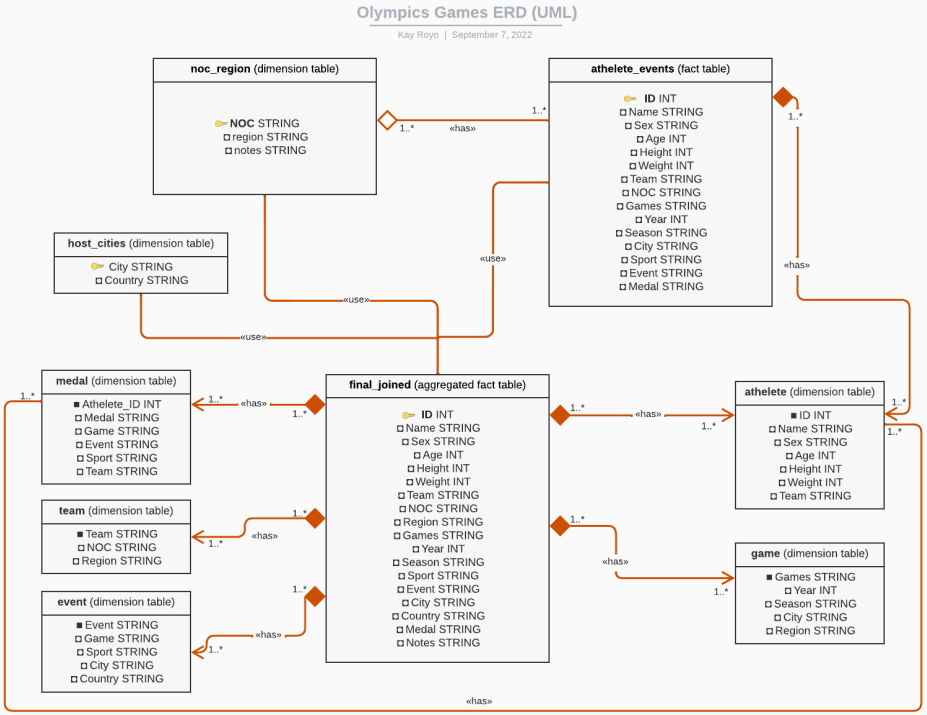
| Average gold medals | home team > away team | home team < away |
|---|---|---|
| Number of games | 36 | 15 |

- Number of medals per sport (male vs. female athletes)

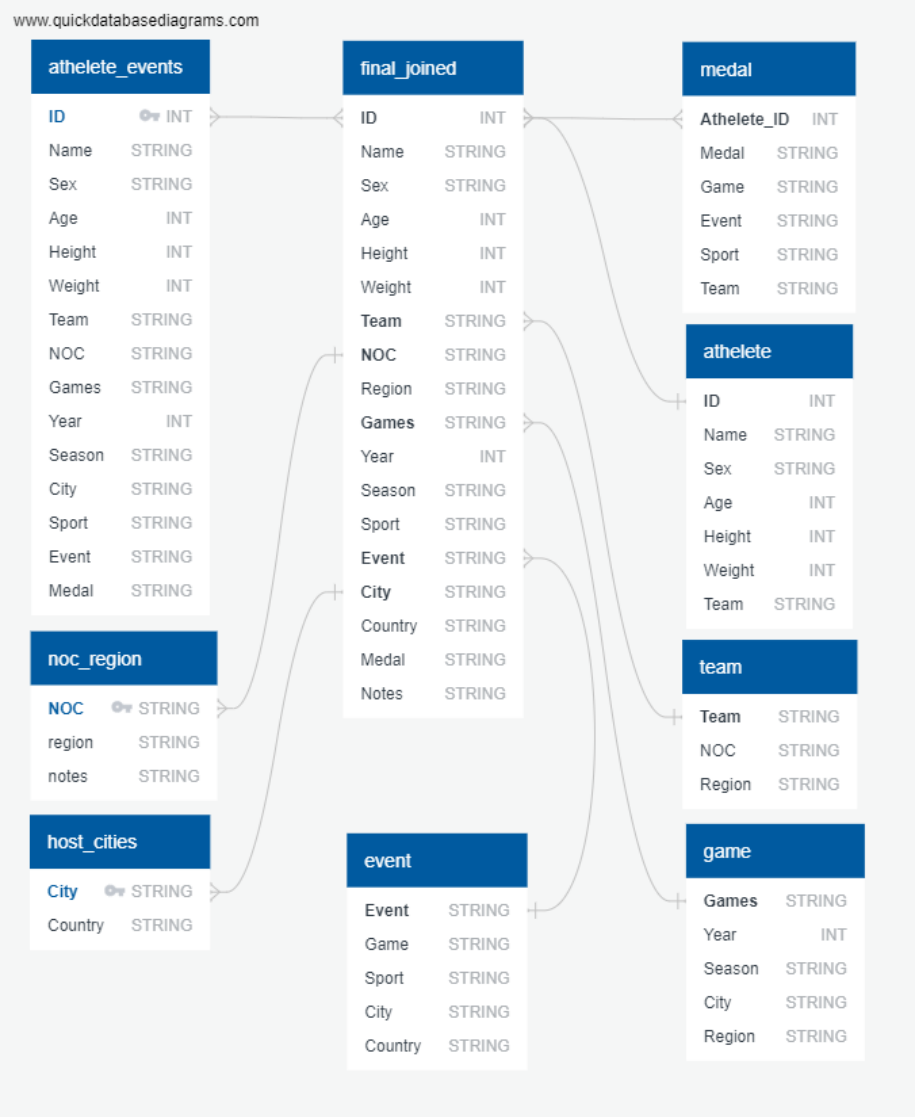
| Average medals | male > female | female > male |
|---|---|---|
| Number of sports | 18 | 31 |
| Sports | Archery, Art Competitions, Badminton, Bobsleigh, Croquet, Curling, Fencing, Figure Skating, Golf, Gymnastics, Luge, Modern Pentathlon, Motorboating, Rugby Sevens, Sailing, Shooting, Ski Jumping, Table Tennis | Alpine, Skiing, Athletics, Basketball, Beach Volleyball, Biathlon, Boxing, Canoeing, Cross Country Skiing, Cycling, Diving, Equestrianism, Football, Freestyle, Skiing, Handball, Hockey, Ice Hockey, Judo, Rowing, Short Track Speed Skating, Skeleton, Snowboarding.Speed Skating, Swimming, Taekwondo, Tennis, Trampolining, Triathlon, Volleyball, Water Polo, Weightlifting, Wrestling |

| | | |
|---|---|---|
| Average gold medals | male > female | female > male |
| Number of games | 20 | 29 |
| Sports | Archery, Art Competitions, Badminton, Bobsleigh, Croquet, Curling, Equestrianism, Fencing, Figure Skating, Golf, Gymnastics, Luge, Modern Pentathlon, Motorboating, Rugby Sevens, Sailing, Shooting, Ski Jumping, Table Tennis, Wrestling | Alpine Skiing, Athletics, Basketball, Beach Volleyball, Biathlon, Boxing, Canoeing, Cross Country Skiing, Cycling, Diving, Football, Freestyle Skiing, Handball, Hockey, Ice Hockey, Judo, Rowing, Short Track Speed Skating, Skeleton, Snowboarding, Speed Skating, Swimming, Taekwondo, Tennis, Trampolining, Triathlon, Volleyball, Water Polo, Weightlifting |

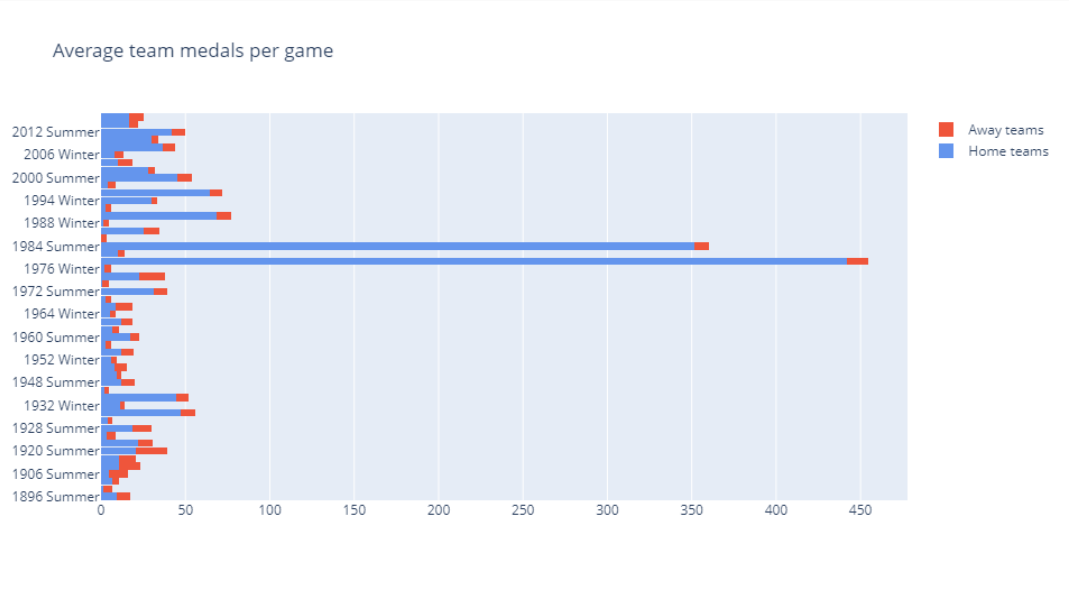
- **Visualizations**
  - **ERD**
    - UML notation

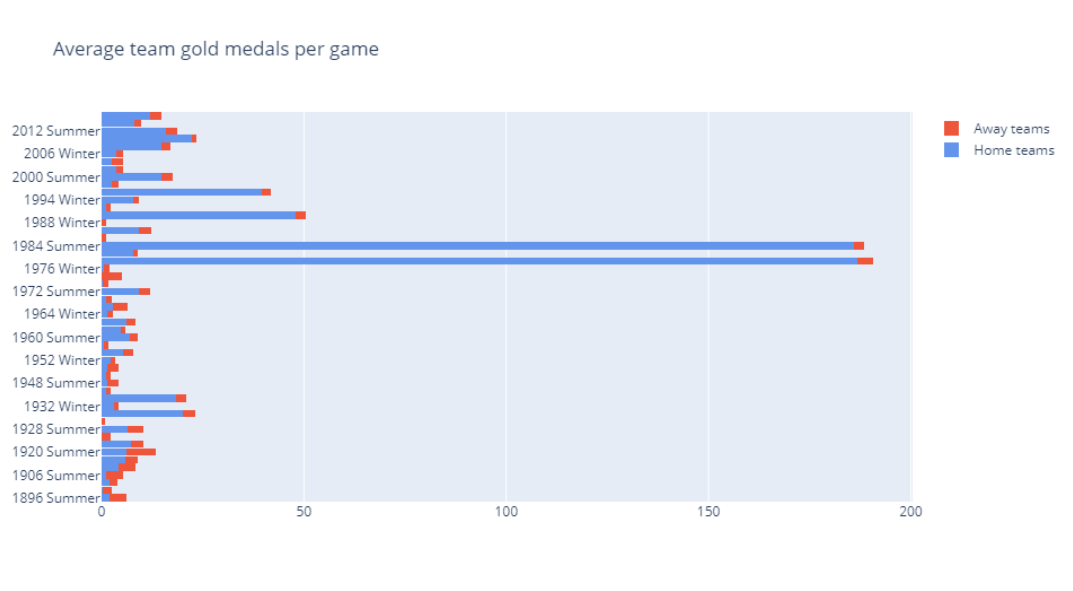


- Crow’s foot notation

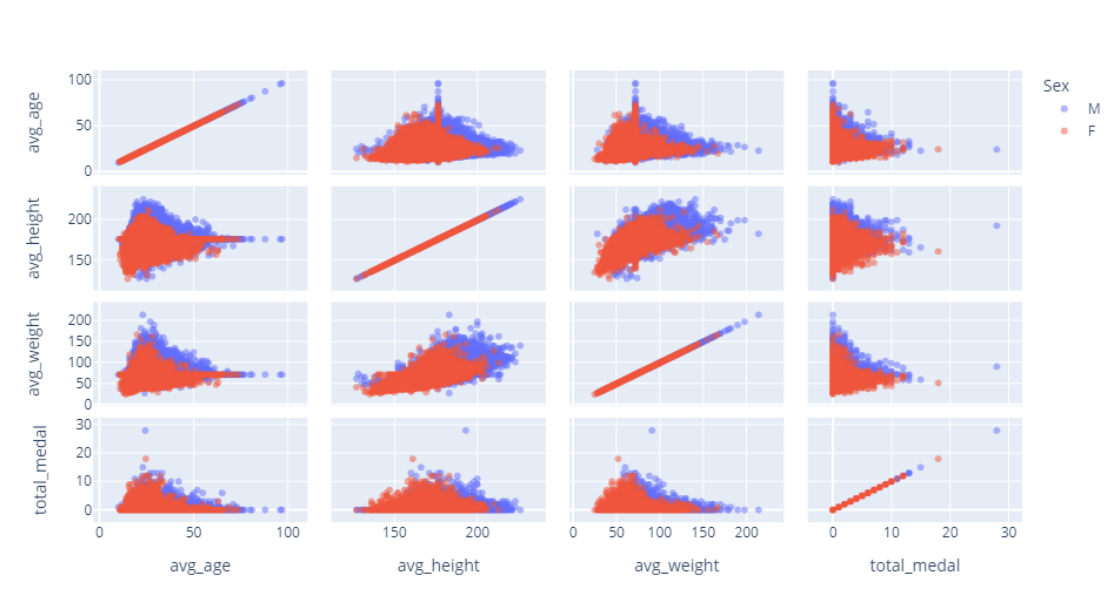


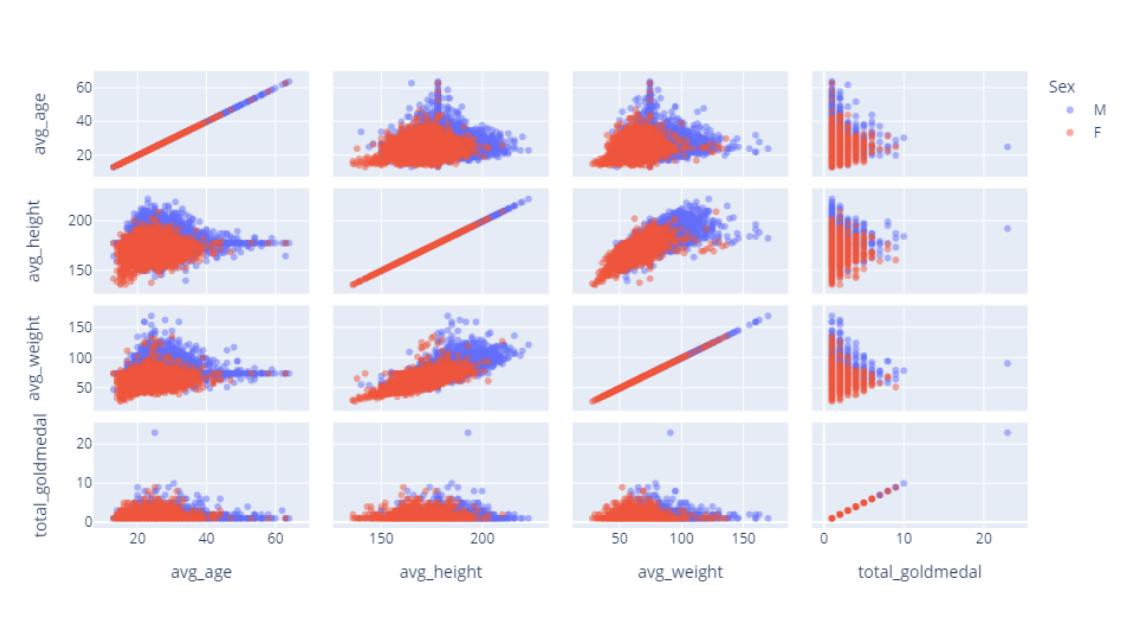

- **Plots**
  - **Assumption 1**



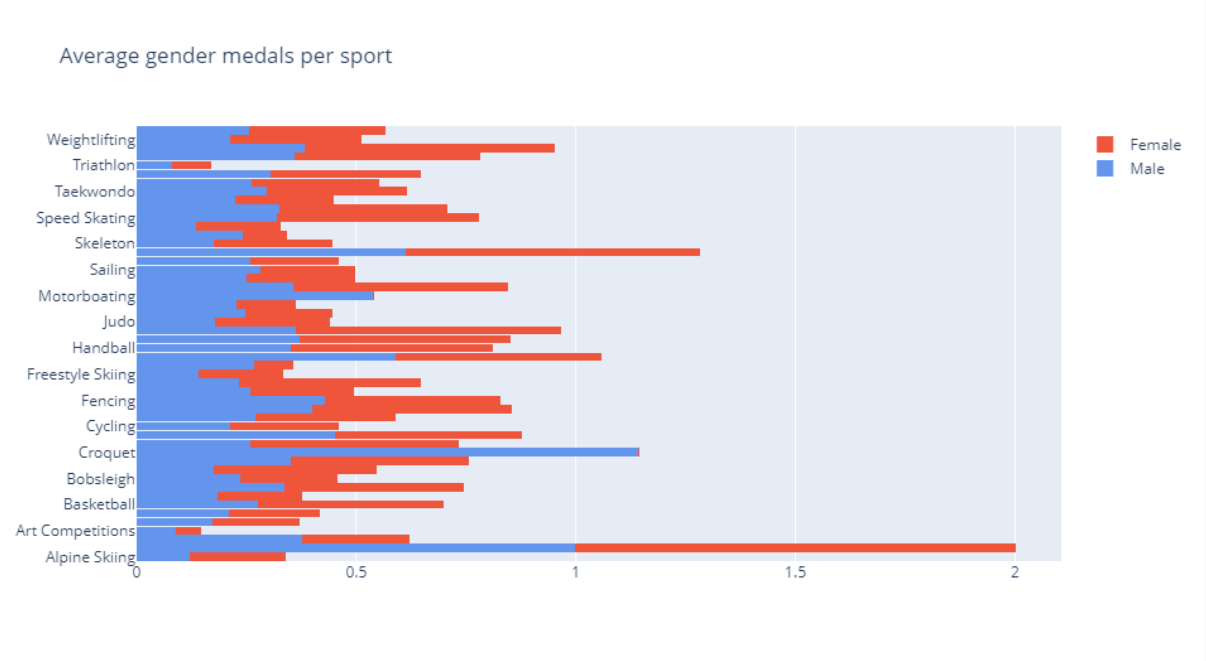



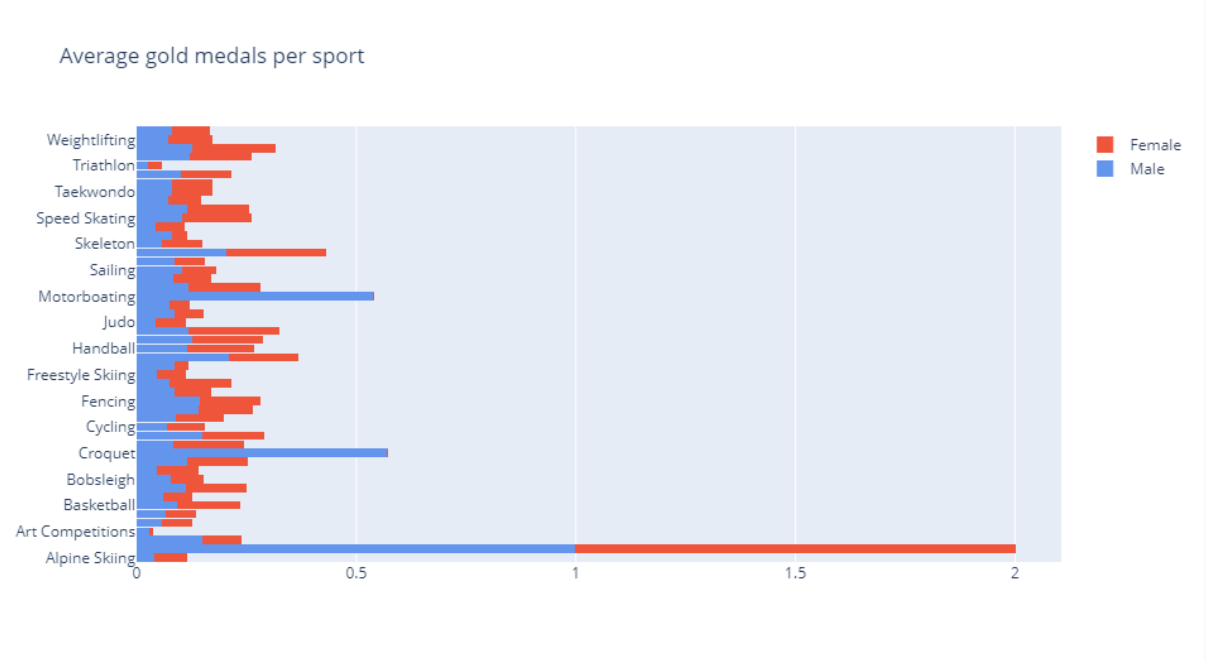

- **Assumption 2**




- **Assumption 3**




- **Correlations (Pearson) and Regressions**
  - Athlete age, height, weight, medals

| *Pearson correlation* | Average age | Average height | Average weight | Medals |
|---|---|---|---|---|
| Average age | 1.000000 | 0.076059 | 0.141231 | 0.046408 |
| Average height | 0.076059 | 1.000000 | 0.762834 | 0.043841 |
| Average weight | 0.141231 | 0.762834 | 1.000000 | 0.044606 |
| Medals | 0.046408 | 0.043841 | 0.044606 | 1.000000 |

- Athlete age, height, weight, gold medals

| *Pearson correlation* | Average age | Average height | Average weight | Gold medals |
|---|---|---|---|---|
| Average age | 1.000000 | 0.075681 | 0.132929 | 0.027825 |
| Average height | 0.075681 | 1.000000 | 0.785820 | -0.028184 |
| Average weight | 0.132929 | 0.785820 | 1.000000 | -0.036213 |
| Gold medals | 0.027825 | -0.028184 | -0.036213 | 1.000000 |

- total medal = average age + average height + average weight

```
Intercept:
 -0.29632281634169494
Coefficients:
 [0.00537322 0.00213988 0.00105983]
                            OLS Regression Results
==============================================================================
Dep. Variable:            total_medal   R-squared:                       0.004
Model:                            OLS   Adj. R-squared:                  0.004
Method:                 Least Squares   F-statistic:                     178.2
Date:                Fri, 16 Sep 2022   Prob (F-statistic):          2.72e-115
Time:                        16:06:44   Log-Likelihood:            -1.4780e+05
No. Observations:              135571   AIC:                         2.956e+05
Df Residuals:                  135567   BIC:                         2.957e+05
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.2963      0.048     -6.114      0.000      -0.391      -0.201
avg_age        0.0054      0.000     15.242      0.000       0.005       0.006
avg_height     0.0021      0.000      6.347      0.000       0.001       0.003
avg_weight     0.0011      0.000      4.349      0.000       0.001       0.002
==============================================================================
Omnibus:                   137690.654   Durbin-Watson:                   0.009
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         15800883.137
Skew:                           4.824   Prob(JB):                         0.00
Kurtosis:                      55.001   Cond. No.                     4.77e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.77e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

- total gold medal = average age + average height + average weight

```
Intercept:
 1.3080802087573147
Coefficients:
 [ 0.00437958  0.00024712 -0.00245168]
                            OLS Regression Results
==============================================================================
Dep. Variable:        total_goldmedal   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     8.356
Date:                Sun, 18 Sep 2022   Prob (F-statistic):           1.52e-05
Time:                        14:04:47   Log-Likelihood:                -11695.
No. Observations:               10425   AIC:                         2.340e+04
Df Residuals:                   10421   BIC:                         2.343e+04
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.3081      0.173      7.575      0.000       0.970       1.647
avg_age        0.0044      0.001      3.372      0.001       0.002       0.007
avg_height     0.0002      0.001      0.203      0.839      -0.002       0.003
avg_weight    -0.0025      0.001     -2.709      0.007      -0.004      -0.001
==============================================================================
Omnibus:                    12234.226   Durbin-Watson:                   0.036
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          3492112.221
Skew:                           5.835   Prob(JB):                         0.00
Kurtosis:                      91.900   Cond. No.                     4.63e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.63e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

- **Key Learnings**
  - Some athletes with distinct IDs share the same name. Thus, IDs are used when aggregating information for every athlete.
  - There are more male athletes than female athletes in the history of the Olympics games.
  - The sport Aeronautics only had one sport event with one participant. Therefore, sports with only one gender participant are not included in the table used to compare the number of medals earned by each gender group in each sport.
  - Some athletes participated in multiple sports.
  - Athletes typically only participate in one Olympic Game.
  - There is a 99.7% difference between the number of female and male athletes in the history of the Olympics.

## VI. Insights

- **Hypothesis 1**

  Based on the results shown in the previous section using both performance metrics, there are more Olympics games where the average medals and specifically gold medals obtained by home teams altogether is greater than that of away teams, which proves the assumption that home teams generally perform better due to various factors that give them an advantage over away teams.

- **Hypothesis 2**

  Using correlations, it appears that age, height, and weight do not have a strong, positive correlation with the total number of medals (gold+silver+bronze) and gold medals earned by athletes. However, using multiple linear regression with total medals as the dependent variable, the p-values show that predictors age, height, and weight are all statistically significant using the significance level 0.05. Meanwhile, the multiple regression with dependent variable total gold medals shows that only age and weight are

statistically significant using the same significance level. Therefore, we can conclude that overall age, height, and weight individually do not have a significant influence on the total number of medals or gold medals earned by athletes. However, it appears that altogether age, height, and weight are useful predictors for the total number of medals while age and weight together are useful predictors for the number of gold medals earned.

- **Hypothesis 3**

  Initially, the purpose of this hypothesis was to compare the performance of male and female athletes in every sport by comparing their performance using the total number of medals earned by each gender group in each sport. However, the majority of the sport events in the data are not mixed gender. Therefore, it would not be logical to compare male and female athletes' performance when they play against each other in the same sport event. The new hypothesis then focuses on determining whether male athletes earn more medals than female athletes in certain sports without making an inference on their performance. Based on the results there are more types of sports where female athletes have earned more medals (gold, bronze, and silver) and gold medals alone than male athletes in general.

- **New metrics**

  In order to ensure that the results obtained using the first metric for the three hypotheses stated above, total or average number of medals, make sense, the total or average number of gold medals is introduced as a new metric since gold medals specifically represent the best possible performance in every Olympic game.

## VII. Recommendations and Actions

- Insights summary
  - Overall, it appears that home teams will continue to have an advantage over away teams, not just in the Olympics games.
  - Age, height, and weight individually do not have a strong influence on the number of medals earned by athletes in the Olympics but altogether they can be useful predictors.
  - Even if there are more male than female athletes in the Olympics overall, female athletes have greater average medals in a larger number of sports than male athletes. Although we cannot conclude if female athletes perform better in sports where they receive more average medals using this metric, we know that the gender gap in the Olympics is not extreme.

- Recommendations
  - Recommendations for clients

    Based on the insights stated above, the client can advise news outlets to stop the spread of the idea that there's an extreme gender gap in the Olympics games and advise away teams in the future Olympics games to prepare better than home teams. Since age, height, and weight don't individually influence the performance of an athlete in the Olympics, the client can encourage athletes from different age, height, and weight range to give it a try.

  - Next step or actions
    - Compare the number of medals obtained by male and female athletes in mixed-gender sport events.
    - Investigate how age, height, and weight are correlated with the number of medals in a specific sport or sport event.

- Investigate gender gap in specific sports and how it changed from the first game to the most recent game.