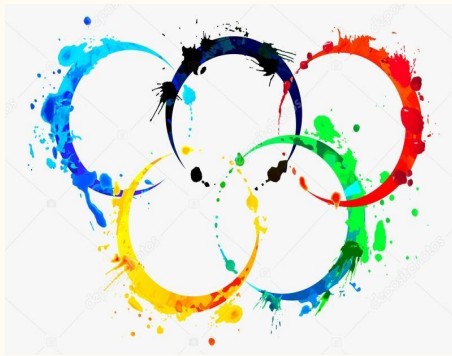# Project Proposal

—

Data Analysis Project Proposal
Faranak Fard

Step 1: Proposal Preparation

Which dataset?

And why?

I selected SportsStats dataset. I have always been curious to learn more about Olympic games stats. Also, prediction in sports using data mining techniques looks an interesting research area.

Importing Data

The dataset is downloaded from dropbox:

SportsStats (Olympics Dataset - 120 years of data)

There are two CSV files. I am using Python and Pandas in Jupyter Notebook to explore SportsStats. I name two CSV files as "nocr" and "atev" dataframes.

```python
import pandas as pd
```

```python
df_nocr = pd.read_csv('./SportsStats/noc_regions.csv')
df_atev = pd.read_csv('./SportsStats/athlete_events.csv')
```

Exploring Dataset: nocr

df_nocr consists of three attributes:

- NOC (a three-letter code)
- region (name of the country)
- notes

There are 230 entries. Knowing that there are 195 countries in the world today, there should be some duplications.

Also, 3 entries have Null value for region!

Exploring Dataset: nocr

Here is the entries with Null value for region:

A list of regions represented with more than one NOC:

	NOC	region	notes
168	ROT	NaN	Refugee Olympic Team
208	TUV	NaN	Tuvalu
213	UNK	NaN	Unknown

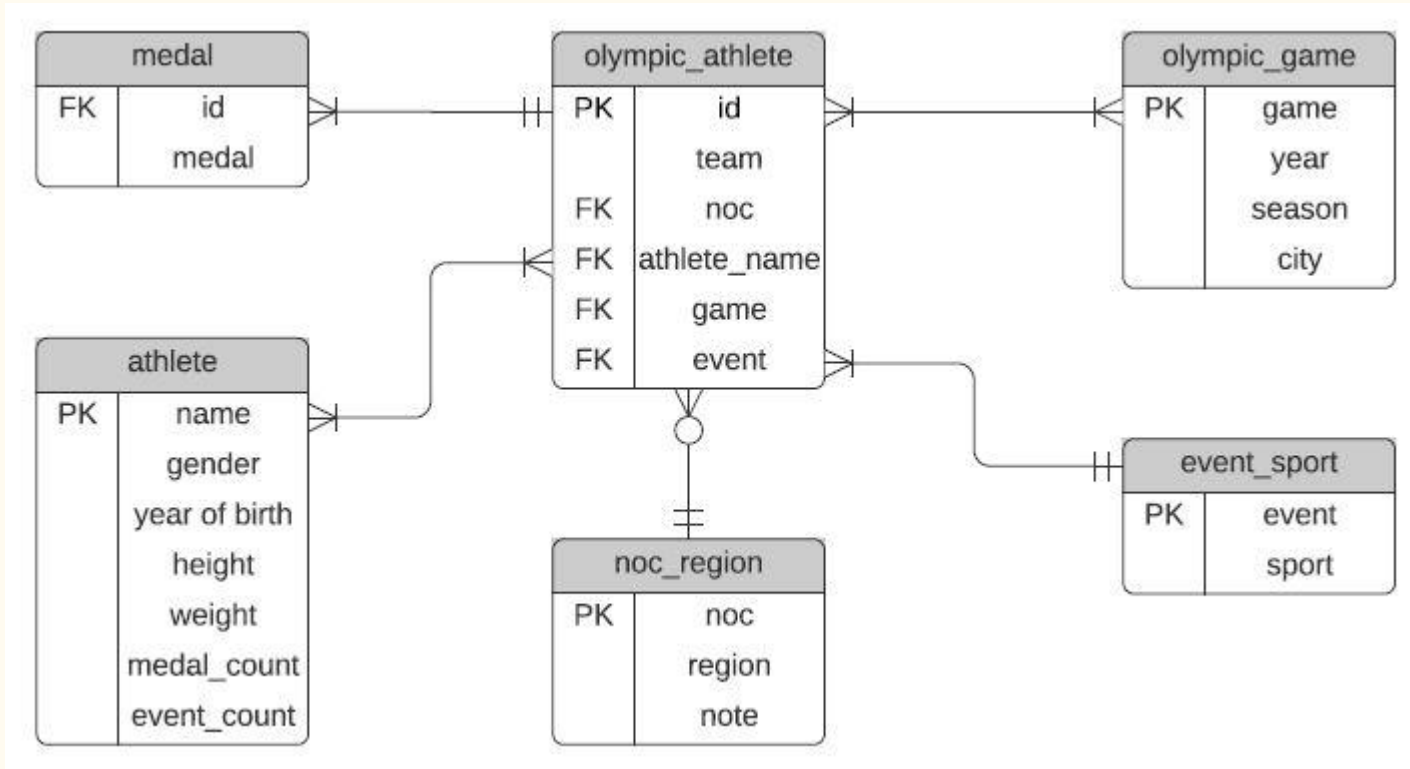| region	NOC
Australia	2
Canada	2
China	2
Czech Republic	3
Germany	4
Greece	2
Malaysia	3
Russia	3
Serbia	3
Syria	2
Trinidad	2
Vietnam	2
Yemen	3
Zimbabwe	2

Exploring Dataset: atev

df_atev has 15 attributes:

```
Index(['ID', 'Name', 'Sex', 'Age', 'Height', 'Weight', 'Team', 'NOC', 'Games',
       'Year', 'Season', 'City', 'Sport', 'Event', 'Medal'],
      dtype='object')
```

271,116 entries. Three columns of Age, Height, and Weight have missing values. Also the attribute medal has Null values representing "gained no medal".

Proposed ERD

Step 2:
Project Proposal Development

What to expect in Step 2?

1. Questions to Answer

2. Initial Hypothesis (or Hypotheses)

3. Data Analysis Approach

Section 1: Questions to Answer

1. What countries have had the most number of :
    a. Attendees in games?
    b. Women athletes?
    c. Medals?
2. About sports:
    a. How many are they?
    b. Which ones existed from begining?
    c. Which ones are the most popular?
3. About athletes:
    a. Who had the most medals, in which sports?
    b. How is the age range?
    c. How is the portion of men to women?
    d. What are popular sports in different age range?
    e. What is the longest years of appearance in games for an athlete?

Section 2: Initial Hypothesis (or Hypotheses)

1. Russia, USA, and China will have the most number of athletes, women, and medals.
2. I believe there are many sports have been added to the list in the past 50 years.
3. I think the age range is between 15 to 40. And women have a lower (but growing) portion of men.

Section 3: Data Analysis Approach

1. I will looking mostly at frequency:
   a. For athletes' names
   b. For gained medals in each country
   c. For sports
2. I will also look for:
   a. Rate of women in years
   b. Portion of women to men

Step 3:

The End