Peer-graded Assignment: Milestone 1: Project Proposal and Data Selection/Preparation

You are a data scientist working for a data analytics firm. Your firm has explored a multitude of data sources and is tasked with providing key insights that your clients can make actionable. Your manager has asked you to provide some data analytics guidance for one of the firm's clients. In a typical scenario, you would iteratively work with your client to understand the data wanting to be analyzed. Having a solid understanding of the data and any underlying assumptions present is crucial to the success of a data analysis project. However, in this case, you will need to do a little more of the "heavy lifting".

To begin, you will prepare a project proposal detailing:

1. The questions we are wanting to answer,
2. initial hypothesis about the data relationships, and
3. the approach you will take to get your answers.

Review criteria

The project proposal you will develop will guide you where you want to go, but may change along the way; and that's OK! To kick things off you will need to:

- Select your client
- Import your dataset
- Explore and understand your data
- Develop an Entity Relationship Diagram (ERD)

For this milestone, you will upload a PDF version of the two key steps needed in developing your project proposal:

1. Preparing for Your Project Proposal
2. Develop Your Project Proposal

Step 1: Preparing for Your Proposal

You will document your preparation in developing the project proposal. This includes:

1. Which client/dataset did you select and why?

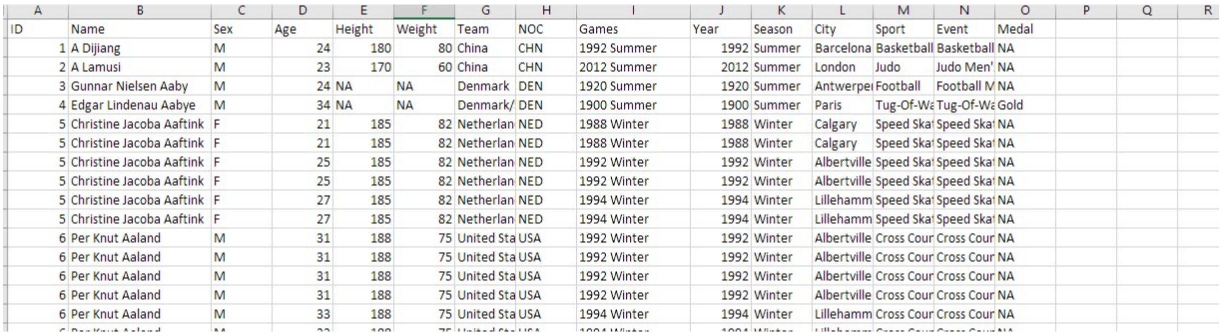
   The sports Data set. I found the possibility of examining the outcomes of medalists and the participants countries of origins, sex, and sporting event type interesting. The imaginary client(s) would be for olympic event organizers and team sponsors for the relevant participants of a given country.

2. Describe the steps you took to import and clean the data.

   Importing the data was a trivial process although even when initially examining the data in the .csv files , I noticed that there were some data points missing which would need to be accounted for.

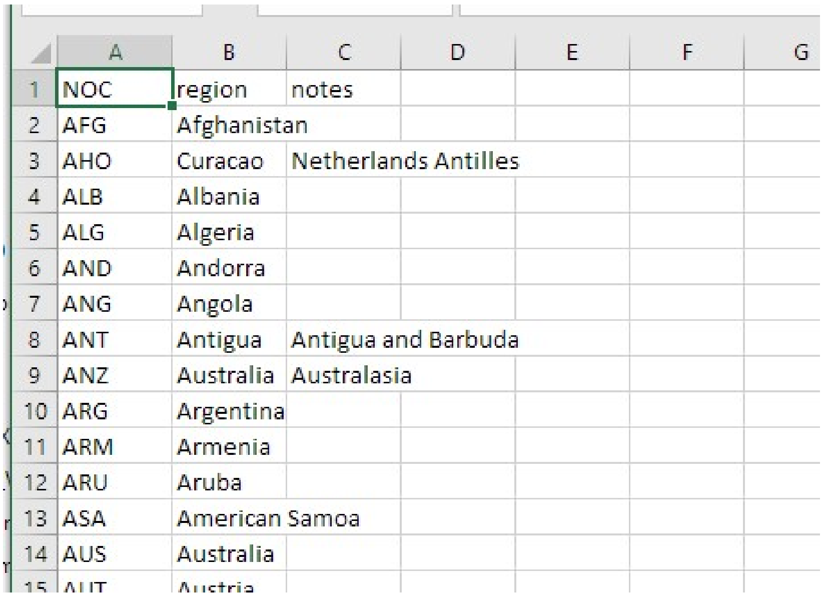
3. Perform initial exploration of data and provide some screenshots or display some stats of the data you are looking at.

   A screenshot of the main .csv file showing the headers and sample data values.



ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
1	A Dijiang	M	24	180	80	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball	NA
2	A Lamusi	M	23	170	60	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men'	NA
3	Gunnar Nielsen Aaby	M	24	NA	NA	Denmark	DEN	1920 Summer	1920	Summer	Antwerpei	Football	Football M	NA
4	Edgar Lindenau Aabye	M	34	NA	NA	Denmark/	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-Wa	Tug-Of-Wa	Gold
5	Christine Jacoba Aaftink	F	21	185	82	Netherlan	NED	1988 Winter	1988	Winter	Calgary	Speed Skat	Speed Skat	NA
5	Christine Jacoba Aaftink	F	21	185	82	Netherlan	NED	1988 Winter	1988	Winter	Calgary	Speed Skat	Speed Skat	NA
5	Christine Jacoba Aaftink	F	25	185	82	Netherlan	NED	1992 Winter	1992	Winter	Albertville	Speed Skat	Speed Skat	NA
5	Christine Jacoba Aaftink	F	25	185	82	Netherlan	NED	1992 Winter	1992	Winter	Albertville	Speed Skat	Speed Skat	NA
5	Christine Jacoba Aaftink	F	27	185	82	Netherlan	NED	1994 Winter	1994	Winter	Lillehamm	Speed Skat	Speed Skat	NA
5	Christine Jacoba Aaftink	F	27	185	82	Netherlan	NED	1994 Winter	1994	Winter	Lillehamm	Speed Skat	Speed Skat	NA
6	Per Knut Aaland	M	31	188	75	United Sta	USA	1992 Winter	1992	Winter	Albertville	Cross Cour	Cross Cour	NA
6	Per Knut Aaland	M	31	188	75	United Sta	USA	1992 Winter	1992	Winter	Albertville	Cross Cour	Cross Cour	NA
6	Per Knut Aaland	M	31	188	75	United Sta	USA	1992 Winter	1992	Winter	Albertville	Cross Cour	Cross Cour	NA
6	Per Knut Aaland	M	31	188	75	United Sta	USA	1992 Winter	1992	Winter	Albertville	Cross Cour	Cross Cour	NA
6	Per Knut Aaland	M	33	188	75	United Sta	USA	1994 Winter	1994	Winter	Lillehamm	Cross Cour	Cross Cour	NA

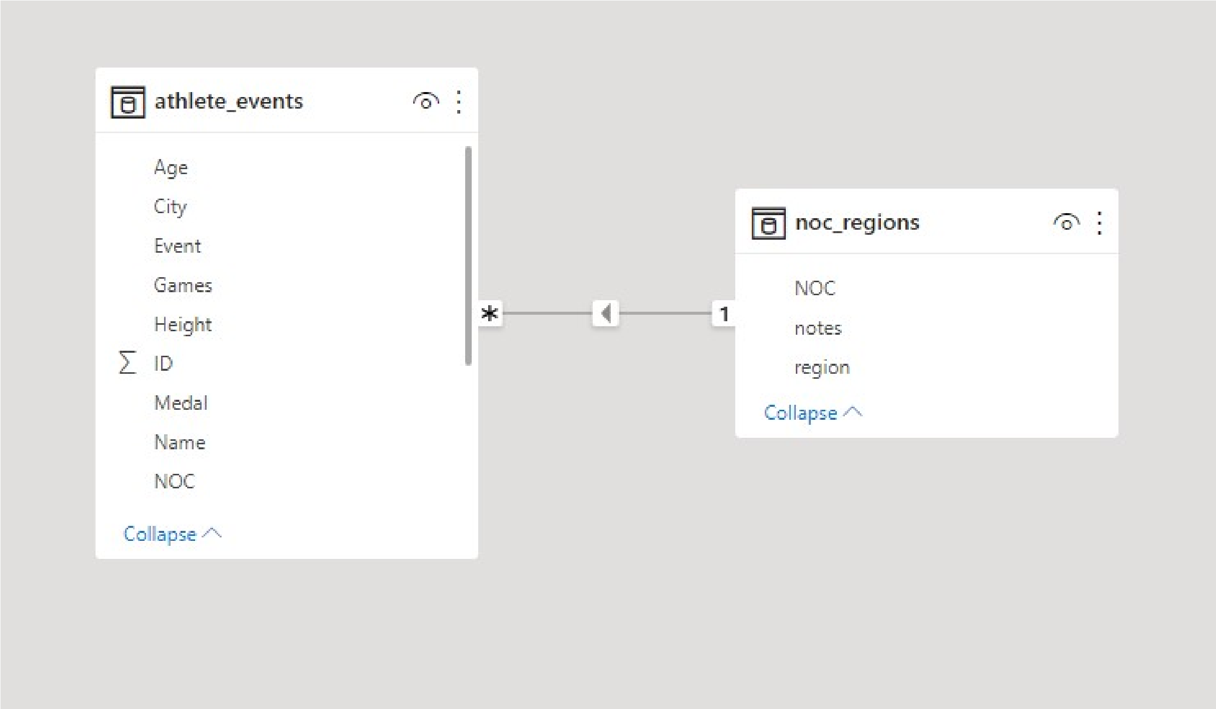
A screen shot of a sample of the second data, focusing on regions and some additional notes of participating teams, set can be seen below. Not certain if this second data set will be of much use as it is likely to be entirely redundant with the first. I may create additional tables from the existing data if it proves beneficial.



NOC	region	notes
AFG	Afghanistan	
AHO	Curacao	Netherlands Antilles
ALB	Albania	
ALG	Algeria	
AND	Andorra	
ANG	Angola	
ANT	Antigua	Antigua and Barbuda
ANZ	Australia	Australasia
ARG	Argentina	
ARM	Armenia	
ARU	Aruba	
ASA	American Samoa	
AUS	Australia	
AUT	Austria	

4. Create an ERD or proposed ERD to show the relationships of the data you are exploring.

   This is the initial ERD based on the data “as is”. I may change this diagram later, for example, if I create additional tables that are subsets of the data. Most likely I will want to use the ID column as the Key value to connect additional tables. I need to do explore the data more to see if additional tables are necessary/ beneficial.



Step 2: Develop Project Proposal

In this step, you will need to include the following:

Description

Write a 5-6 sentence paragraph describing your project; include who might be interested to learn about your findings. Who might be your audience?

> *The intended audience would be client(s) representing olympic event organizers and/or team sponsors for the relevant participants of a given country. Additionally, it might the results might be of interest to sports science researchers.*

Questions

Create 2-3 questions that you want to answer with the data:

1. *To determine if there exists an expected relationship between sporting event gold medalists and country participant origin (i.e., will expected winners be from host countries where that sport is popular / more common such as winter sports).*
2. *Are there any conclusions that can be drawn from examining how male and female athletes perform, and if so what conclusions and what might be possible reasons for those?*
3. *Are there any trends in countries that obtain the max- min of medals and can any conclusions be drawn from those.*
4. *Are there any trends with respect to time that can be observed?*

- This will be easier to answer once you've had an opportunity to look at the data and do some initial exploration.
- Don't get carried away on the analysis piece at this stage as there will be more analysis later.
- Do focus on key data elements that are present. For instance: What are they, when are they, who are they about? Do they connect? How do they connect? Jot down ideas as you brainstorm.

Hypothesis

What are your initial hypotheses about the data?

Write 2-3 assumptions about the data that you'll want to go back to prove or disprove. You will want to keep them in front of you as you look at the data to keep them or change them. You may see relationships that you want to explore and will develop a "belief" about the data.

The assumption is that countries that naturally experience extended winters and have accessible regions to engage in winter sports (mountains etc.) will have a higher number of medalists in winter sports.

The assumption that within a gender group, that the results for number of awards will likely follow larger trends, such as stated in the first assumption.

- Start documenting what you think you can tell from the data.
- What pops up as interesting to you? Most likely it will be interesting to others as well.
- Use the discussion boards to discuss with others about your client and the data to brainstorm together.

Approach

Describe in 5-6 sentences what approach you are going to take in order to prove (or disprove) your hypotheses. Think about the following in your answer:

- What features (fields/columns) are you going to look at first?
- Is there a relationship that exists that you want to explore?
- What metric/ evaluation measure will you use?

One initial hypothesis I wish to test is the idea that countries which have geographical and environmental traits that naturally encourage specific sport events will result in those countries earning more medals (i.e. winter sports).

I will examine medalists by country of origin, date, and event type first.
I want to examine the relationship between country and medalists count.
I will use basic statistical methods (avg, mean, mode etc and if relevant do additional analysis such as regression etc., this might, for example, be interesting to see if indeed a regression model would have predicted the historical outcomes from the data set.)

Metrics will include numbers and types of medals won, countries, years and gender of awardees.