Step 1: Preparing for Your Proposal

1. Which client/dataset did you select and why?
   My client selection is Olympics Games (OG), from which I would like to infer women rights across countries and centuries.

2. Describe the steps you took to import and clean the data.
   I've imported data as a local table in the databricks environment and then started to investigate its content. Thus, I've noticed that some instance of the Team field is followed by -1 or -2. This would result in wrong data groups if later different ID are grouped by country. Thus, I've proceeded with its removal.
3. Perform initial exploration of data and provide some screenshots or display some stats of the data you are looking at.
   It follows some code of a) initial exploration of data, b) Team field cleaning and c) stats on males and females OG attendance

a)

```
SELECT * FROM default.athlete_events_5_csv AS SS WHERE TEAM LIKE "Ge%" LIMIT 100
```

b)

```
SELECT *
FROM (
  SELECT *
  FROM default.athlete_events_5_csv
 ) AS ori
INNER JOIN (
  SELECT ID
        , CASE
           WHEN INSTR(TEAM, '-') > 0
           THEN SUBSTRING (TEAM, 0, INSTR(TEAM, '-')-1)
           ELSE TEAM
         END AS TEAM_ERRATUM
  FROM default.athlete_events_5_csv
    ) AS mod
ON ori.ID=mod.ID
```

c)

```
SELECT TEAM
      , COUNT(Sex) AS MALES
FROM (
   SELECT *
     FROM (
      SELECT *
      FROM default.athlete_events_5_csv
      ) AS ori
     INNER JOIN (
      SELECT ID
        , CASE
           WHEN INSTR(TEAM, '-') > 0
           THEN SUBSTRING (TEAM, 0, INSTR(TEAM, '-') -1)
           ELSE TEAM
          END AS TEAM_ERRATUM
      FROM default.athlete_events_5_csv
      ) AS mod
     ON ori.ID=mod.ID
     ) AS sport_stats_erratum
WHERE Sex = "M"
GROUP BY TEAM
SORT BY MALES DESC
```

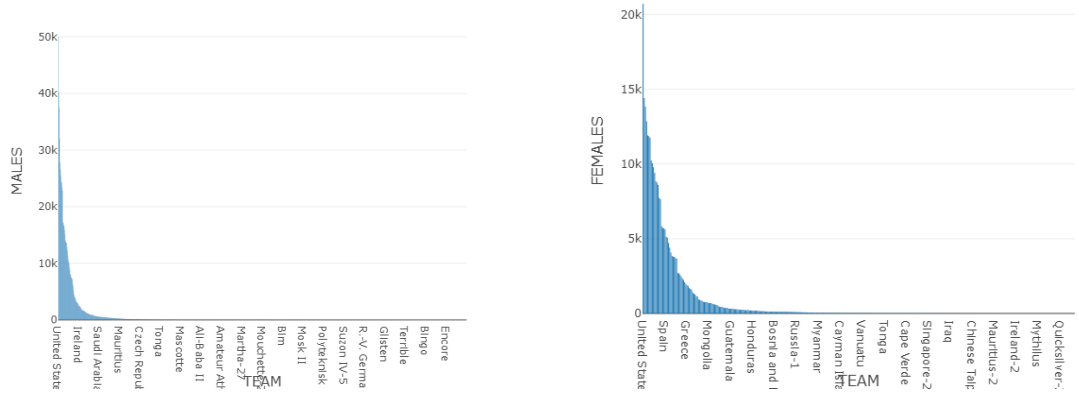

FIGURE 1. Stats on males and females OG attendance

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

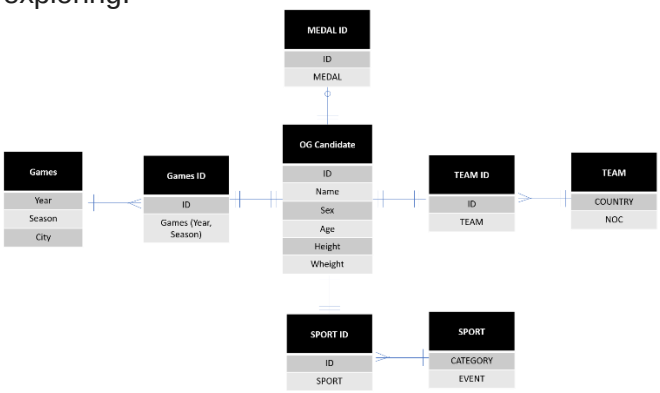

FIGURE 2. ERD for sport stats data

Step 2: Develop Project Proposal

1. 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 aim of my project is to help a non-profit organization on women-rights from Saudi Arabia to promote women equality. This will be achieved by showing the path traced by several societies worldwide, based on Olympic Games attendance. The data available, which show the variation of attendance rate for different countries in the last century may confirm the efforts made by Saudi Arabia in the last years and promote a lot more to empower women in the future. My findings might encourage political awareness and the action needed to allow the cultural transition.

2. Questions
   a) Does today women attendance at the OG show gender-equality across country?
   b) Do historical attendances of women for given countries show milestones in gender-equality?
   c) Can we predict future attendances for Saudi Arabia women equality by looking at the past?
   d) What is the type of events and physiology that Saudi Arabia women are best fit for winning the next OG?

3. Hypothesis
   a) I expect Western countries to have higher rates than Muslim countries.
   b) I expect slope changes in correspondence of women rights milestones achieved by countries across the globe
   c) I expect that a given country attendance started from a relatively low quote and raised proportionally to both internal and external historical achievements. So looking at the slope of past GO we can predict future attendances.
   d) I expect Saudi Arabia women excel in summer events and with body types fitted for athletics

4. Approach:

   a) To answer this question, I need to introduce some basic metrics such as the percentage and the average over several events and we also need to limit the number

of events to look at. As 2012 has been the first OG where women participated from Saudi Arabia, we will count this average of percentages from 2012 on.
b) To answer this question, I need to calculate the I've calculated the attendance percentage of all women both participating worldwide with respect to men and starting from the first year of attendance. For instance, I will need to start counting from OG 1900 in the case of US, and from OG 2012 in the case of Saudi Arabia.
c) For this answer I could base our prediction on the historical trends of Saudi Arabia and compare this with the trend of other country that have already attempted women equality at the OG.
d) Finally, here I will need to look for events and body types of Saudi Arabia women who have won medals during the past OG and promote them accordingly.

WEEK 2

- Provide a summary of the different descriptive statistics you looked at and WHY.

I've looked at percentage and average over several events to calculate women OG attendance per country and per year. I've also looked at minimum and maximum values. I've then looked at slope changes in graphs as identifiers of milestones. Finally, I've looked at Events and Body Mass Index for which Muslim countries have won Olympic Medals in the past Games.

- Submit 2-3 key points you may have discovered about the data, e.g. new relationships? Aha's! Did you come up with additional ideas for other things to review?

After I've noticed that Saudi Arabia women have not gained Olympic Medals in the OGs 2012-2016, I was obliged to extend the statistics on Events and Body Mass Index to other awarded Muslim Countries geo-politically close to Saudi Arabia. Thus, I've limited my exploration to countries where Islam is professed by more than 70 % of their population: Tunisia, Algeria, Turkey, Syria, Iran, Morocco, Egypt, Tajikistan, Uzbekistan, Kazakhstan.

- Did you prove or disprove any of your initial hypotheses? If so, which one and what do you plan to do next?

I've confirmed that Western countries have higher rates than Muslim countries.

I've found slope changes in correspondence of women rights milestones achieved by western countries in the 50's and Muslim countries in the 90's.

Within the dataset of Muslim women who have won medals at the past OGs, I've found that athletics was the best event, with a BMI around 18.

- What additional questions are you seeking to answer?

I'd like to investigate also the top second event and associated body type of Muslim women who have won medals at the past OGs. This will give a diversified sport alternative for Saudi Arabia women that would like to candidate to next Olympics.

WEEK 3

Dive Deeper

Look deeper into the features you are investigating, consider:

- Relationships / Correlation, Pearson Correlation
- Linear Regression for future prediction (if the relationship is linear)
- Textual Analysis for TF-IDF (Term Frequency-Inverse Document Frequency; Row-based and column-based, stop-word removal?

Specify 1-2 correlations you discovered. List the fields that you found to be correlated and describe what you learned from these correlations.

1. I will need to calculate a nonlinear regression to answer question and hypothesis b) of my project:

   b) Do historical attendances of women for given countries show milestones in gender-equality?
   b) I expect slope changes in correspondence of women rights milestones achieved by countries across the globe

   As reported in FIGURE 1, I've calculated the attendance percentage of all women participating worldwide with respect to men and starting from the first year of attendance. For instance, I started counting from OG 1900 in the case of US, and from OG 2012 in the case of Saudi Arabia. For the sake of completeness, I've also calculated the percentage of men which confirms the opposite trend.

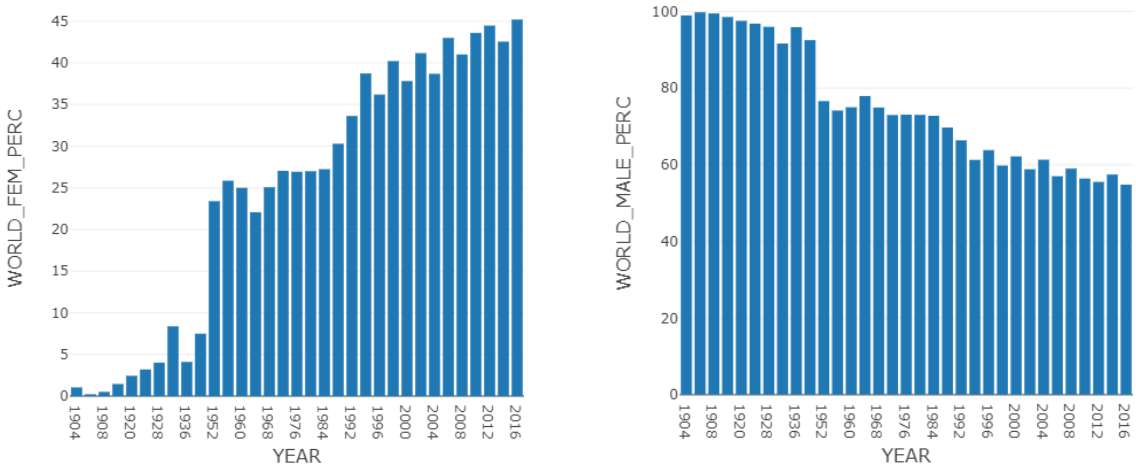

FIGURE 1. Attendance percentage of women (left) and men (right) summed up worldwide starting from OG 1900.

Thus, I've used support vector machine regression algorithm to interpolate the data on the left of Figure 1, whose result is reported in Figure 2 for two different choices of the fitting parameters. Moreover, the data have been split in train and test data and the

respective test score was calculated. The best test score was scanned through parameters via the Grid Search function:

C	γ	Best cross-validation accuracy	Test set score
100	0.001	96%	85%
100	10	5%	3%

TABLE 1. Best cross-validation accuracy and test set score as a result of support vector machine regression algorithm to fit the attendance percentage of women summed up worldwide starting from OG 1900.

From Figure 2 one can extrapolate two main slope changes highlighting two milestones: western countries women rights in the 50s (rights to vote and economic boom after WW2) and Muslim countries women rights in the 90s (Cairo Declaration on Human Rights in Islam). More precisely, the long transition towards women equality in countries has witnessed Moroccan women as the first entering the GO in the 70s, while Saudi Arabia women were amongst the last entering the GO in 2012 as mentioned above.

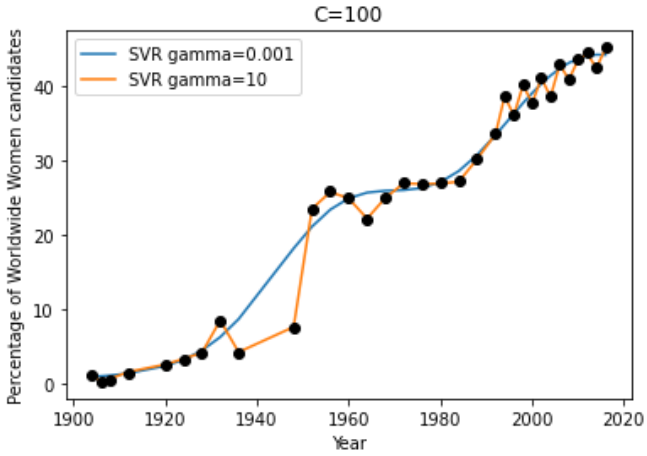

FIGURE 2. Dots- Attendance percentage of women worldwide with respect to men starting from the GO 1900. Solide Lines- SVR fit with parameters γ = 0.001 and C=100 (blue) and γ = 10 and C=100 (orange)

Finally, the fit also help identifying a saturation of attendance percentage towards 45%, which tells that worldwide equality is not reached yet and that some more efforts are needed globally in order to fill this 5% gap in attendance at the next OGs.

2. I will need to calculate a nonlinear decision boundary to answer question and hypothesis d) of my project:
   d) What is the type of events and physiology that Saudi Arabia women are best fit for winning the next OG?
   d) I expect Saudi Arabia women excel in summer events and with body types fitted for athletics

   After I've noticed that Saudi Arabia women have not gained Olympic Medals in the OGs 2012-2016, I was obliged to extend the statistics on Events and Body Mass Index to other awarded Muslim Countries geographically close to Saudi Arabia. Thus, I've limited my exploration to countries where Islam is professed by more than 70 % of their

population: Tunisia, Algeria, Turkey, Syria, Iran, Morocco, Egypt, Tajikistan, Uzbekistan, Kazakhstan. Here is the distribution of the top two sport events that have gained more medals since their first OG attendance.

Sport Event	Number of Medals
Athletics	40
Wrestling	13

TABLE 2. Top two sport events that have gained more medals within the following countries: Tunisia, Algeria, Turkey, Syria, Iran, Morocco, Egypt, Tajikistan, Uzbekistan, Kazakhstan.

This table already helps answer the first question: as athletics and wrestling are the sport events with more medals gained in the past OGs by Muslim Countries geographically close to Saudi Arabia, Saudi Arabia women might have more chances to win in these disciplines. To know what kind of body type Saudi Arabia women should aim to win in these disciplines, I've reported in Figure 3 a nonlinear decision boundary study with features weight/height of the above awarded women and targets Athletics/Wrestling. A support vector machine classification algorithm has been used with parameters C= 100000 and γ=0.001. Best cross-validation accuracy and test score are 100%. Interestingly, there are less points in Figure 3 than number of medals in Table 2. This is because different subcategory sport event medals have been won by the same athlete. For the future women with weight and height in the blue area have then more chances to win in athletics and viceversa women with weight and height in the orange area to win in wrestling.

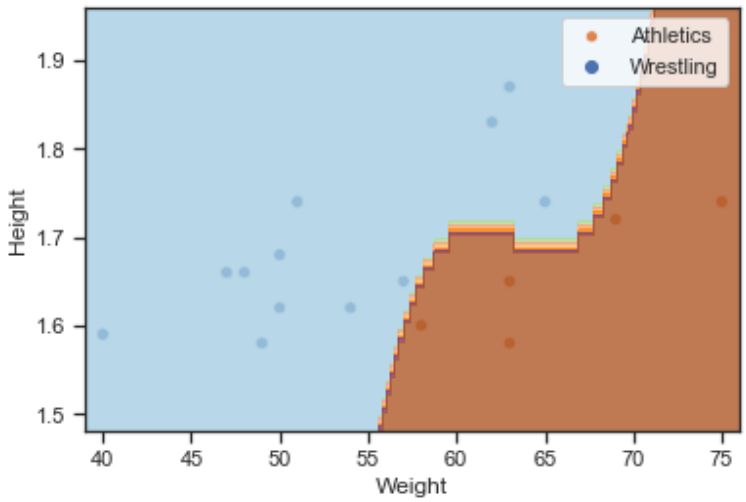

FIGURE 3. Nonlinear decision boundary via support vector machine classification study (C= 100000 and γ=0.001) with features weight/height and targets Athletics/ Wrestling.

Go Broader

Expand the features you are investigating. Look for connections/relationships that you may have initially missed.

1. What jumps out at you now?
2. Use the descriptive stats to point you to features that you may now want to consider.

What key terms did you discover in any text analysis, for whom? Any themes? If you are not analyzing text, summarize what other things you are considering in your analysis?

Another interesting statistic to look at is the percentage of Muslim women participating at GOs, as this can help identifying the areas on Earth where the equality gap is higher and where efforts are more needed. In this regard, Figure 4 reports the attendance rate of women per year from Saudi Arabia or other Muslim countries. For this calculation I've limited my research to countries close to Saudi Arabia (Islam professed by more than 70 %, and overall population more than 10 millions).

As in Figure 2, I've used support vector machine regression algorithm to interpolate the data. The fit confirms the slope change in the 90's and help identifying a saturation of attendance percentage towards 35%. This is 10 % less than the saturation saw worldwide confirming that in this part of the world equality gap is higher. We can then expect that Saudi Arabia attendance rise from the last updated percentage of 25% in 2016 to the average rate of Muslim countries of 35 % as seen in Figure 3.

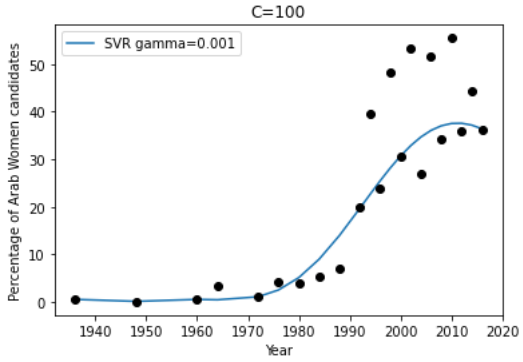

FIGURE 4. Dots- Attendance percentage of women countries neighbour of Saudi Arabia (Islam professed by more than 70 %, ad overall population more than 10 millions). Solide blue Line- SVR fit with parameters g = 0.001 and C=100 (blue)

New Metric

Create 1 or 2 new metrics to track relationships of data you discovered. Explain why you created them.

By looking at US historical attendance we can infer some women rights milestones. As reported in the histogram and its SVR fit in Figure 4, US women attendance follows closely the world attendance reported in Figure 2 for what concern the slope change in the 50s. This follows women rights to vote and the economic boom after WW2. At the same time, one can

notice that US women took 50 years from their first attendance, to see the rate growing up to 25%. Finally, a small negative trend can be noticed in recent years. Thus, equality is not reached yet and some more efforts are needed even in US in order to fill the 7% gap in attendance at the next OGs.

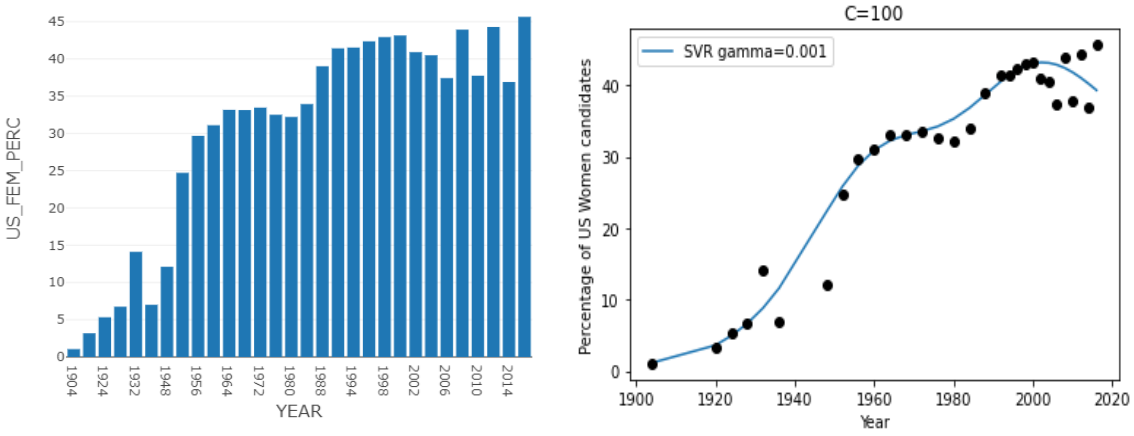

FIGURE 4. Attendance percentage of US women from 1900 and SVR fit with parameters g = 0.001 and C=100.

By looking at Saudi Arabia historical attendance we can instead see that 2012 was their first year of attendance. Interestingly, Saudi Arabia was the last country in the world to achieve the right to vote for women in 2011. Saudi Arabia experienced in the first 4 years of attendance the same rate US had in the first 50 years. We could then infer that Saudi Arabia rate of women attendance is destined to rise more rapidly than in the early years of US women attendance and that an overall rapid grow of women rights is expected in Saudi Arabia in the next years.

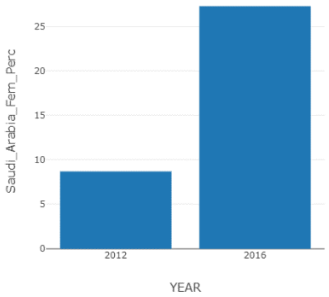

FIGURE 4. Attendance percentage of Saudi Arabia women from 2012.