

SportsStats

Capstone Project

Saurabh Joshi



Table of Contents

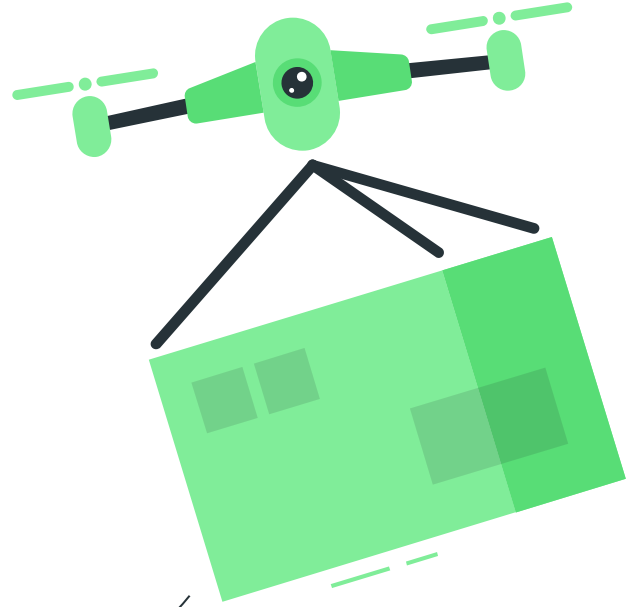
1

Preparing for Your Proposal

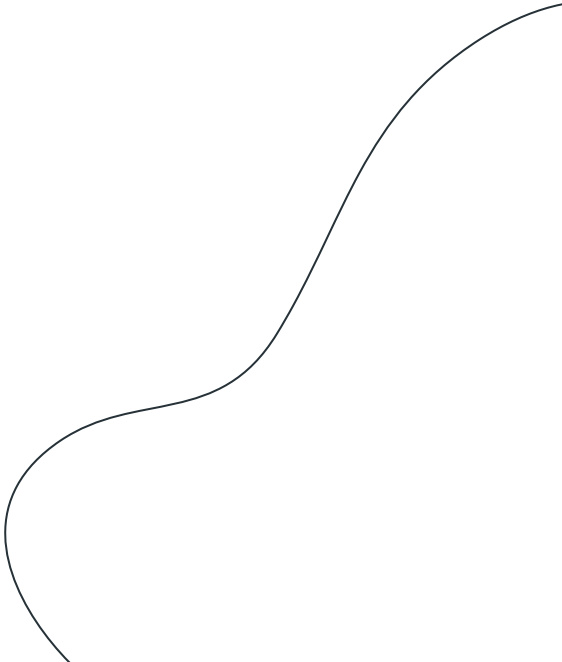
2

Developing Project Proposal

Preparing for Your Project Proposal



Which client did you select and why?

- I choose the SportsStats client. I selected this project since I have strong knowledge in sports and have interests in working in ports analytics reports.
 - With an analysis of the dataset, I can find patterns and hidden insights for players, hidden insights.
- 

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

- For importing the data, I used pandas to read the CSV files.
- Used in-build pandas `to_sql()` to store it in MySQL dataset.
- Since the dataset has NaN values, I didn't clean it, because that would be a falsification of data!

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

```
[6] ▶ MI
country = pd.read_sql('SELECT DISTINCT * FROM regions;', con = engine)

[7] ▶ MI
country
```

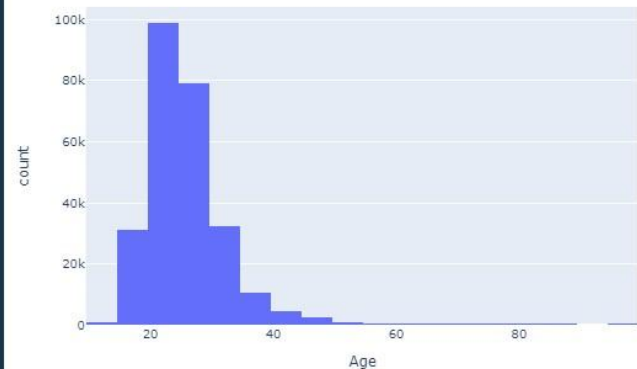
	NOC	region	notes
0	AFG	Afghanistan	None
1	AHO	Curacao	Netherlands Antilles
2	ALB	Albania	None
3	ALG	Algeria	None
4	AND	Andorra	None
...
225	YEM	Yemen	None
226	YMD	Yemen	South Yemen
227	YUG	Serbia	Yugoslavia
228	ZAM	Zambia	None
229	ZIM	Zimbabwe	None

230 rows x 3 columns

```
[8] > %ML
age_distribution = pd.read_sql('SELECT `Age` FROM athlete_events', con = engine)
```

```
[9] > %ML
plt.figure(figsize=(10,15))

fig = px.histogram(age_distribution, x = 'Age', nbins = 30)
fig.show()
```

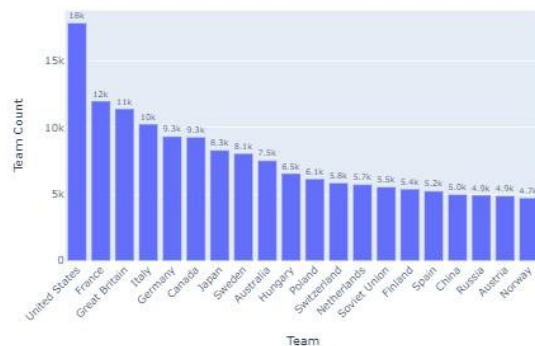


<Figure size 720x1080 with 0 Axes>

```
Team = pd.read_sql('''SELECT `Team`, COUNT(`Team`) AS `Team Count`
FROM athlete_events
GROUP BY `Team`
ORDER BY COUNT(`Team`) DESC
LIMIT 20;''',
con = engine
)
```

```
[11] > %ML
plt.figure(figsize=(15, 10))

fig = px.bar(Team, y = 'Team Count', x = 'Team', text='Team Count')
fig.update_traces(texttemplate = '%(text:.2s)', textposition = 'outside')
fig.update_layout(uniformtext_minsize = 8, uniformtext_mode = 'hide', xaxis_tickangle=-45)
fig.show()
```



<Figure size 1080x720 with 0 Axes>

```
''' USA teams with Sport = Football '''
```

```
[14] > %>% MI
      USA_Football = pd.read_sql(
      ...
      SELECT DISTINCT(`Name`), `Sex`, regions.NOC, `Games`, `Year`
      FROM regions
      INNER JOIN athlete_events
      ON
      regions.NOC = athlete_events.NOC
      WHERE athlete_events.NOC = 'USA'
      AND `Sport` = 'Football'
      ORDER BY `Games` ASC;
      ''', con = engine
    )
```

```
[15] > %>% MI
      USA_Football.head(15)
```

	Name	Sex	NOC	Games	Year
0	Peter Joseph Ratican	M	USA	1984 Summer	1984
1	Joseph J. Brady	M	USA	1984 Summer	1984
2	Alexander Cudmore	M	USA	1984 Summer	1984
3	Louis John Menges	M	USA	1984 Summer	1984
4	Martin Thomas Dooling	M	USA	1984 Summer	1984
5	Johnson	M	USA	1984 Summer	1984
6	Edward B. Dierkes	M	USA	1984 Summer	1984
7	Thomas Thurston "Tom" January	M	USA	1984 Summer	1984
8	John Hartnett January	M	USA	1984 Summer	1984
9	Harry Francis Tate	M	USA	1984 Summer	1984
10	Leo Anthony O'Connell	M	USA	1984 Summer	1984
11	Charles James Hicks January, Jr.	M	USA	1984 Summer	1984
12	Henry Wood Janeson	M	USA	1984 Summer	1984
13	Charles Albert Bartliff	M	USA	1984 Summer	1984
14	Thomas Joseph "Tom" Cooke	M	USA	1984 Summer	1984

```
[16] > %>% MI
      USA_Football.tail(15)
```

	Name	Sex	NOC	Games	Year
286	Tobin Powell Heath	F	USA	2016 Summer	2016
287	Alexandra Patricia "Alex" Morgan	F	USA	2016 Summer	2016
288	Rebecca Elizabeth "Becky" Sauerbrunn	F	USA	2016 Summer	2016
289	Crystal Alyssia Dunn	F	USA	2016 Summer	2016
290	Carli Anne Lloyd (-Hollins)	F	USA	2016 Summer	2016
291	Lindsey Michelle Horan	F	USA	2016 Summer	2016
292	Megan Anna Rapinoe	F	USA	2016 Summer	2016
293	Christen Annemarie Press	F	USA	2016 Summer	2016
294	Julie Beth Johnston (-Ertz)	F	USA	2016 Summer	2016
295	Alexandra Linsley "Allie" Long	F	USA	2016 Summer	2016
296	Whitney Elizabeth Engen	F	USA	2016 Summer	2016
297	Kelley Maureen O'Hara	F	USA	2016 Summer	2016
298	Hope Amelia Solo (-Stevens)	F	USA	2016 Summer	2016
299	Meghan Elizabeth Klingenberg	F	USA	2016 Summer	2016
300	Mallory Diane Pugh	F	USA	2016 Summer	2016

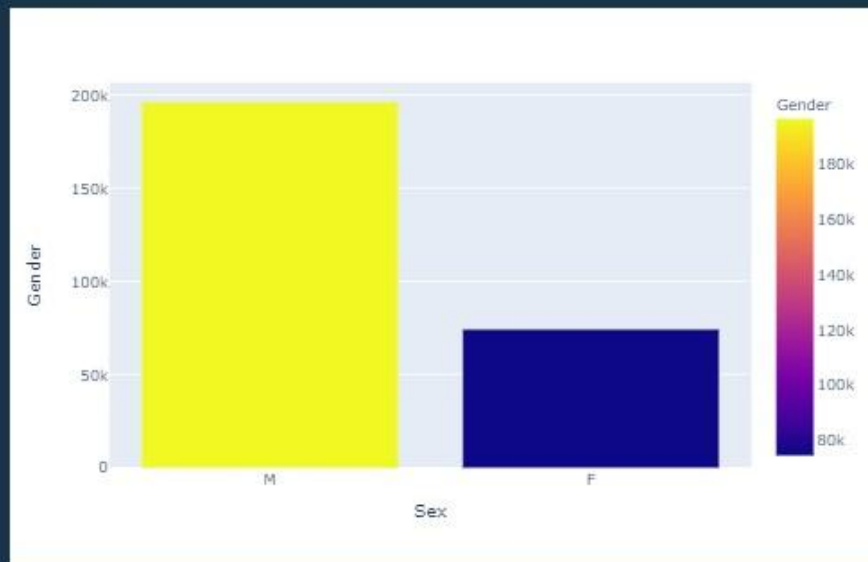
''' Gender Visualization '''

[17]

```
Gender = pd.read_sql(
    '''
    SELECT `Sex`, COUNT(`Sex`) AS `Gender`
    FROM athlete_events
    GROUP BY `Sex`;
    ''', con = engine
)
```

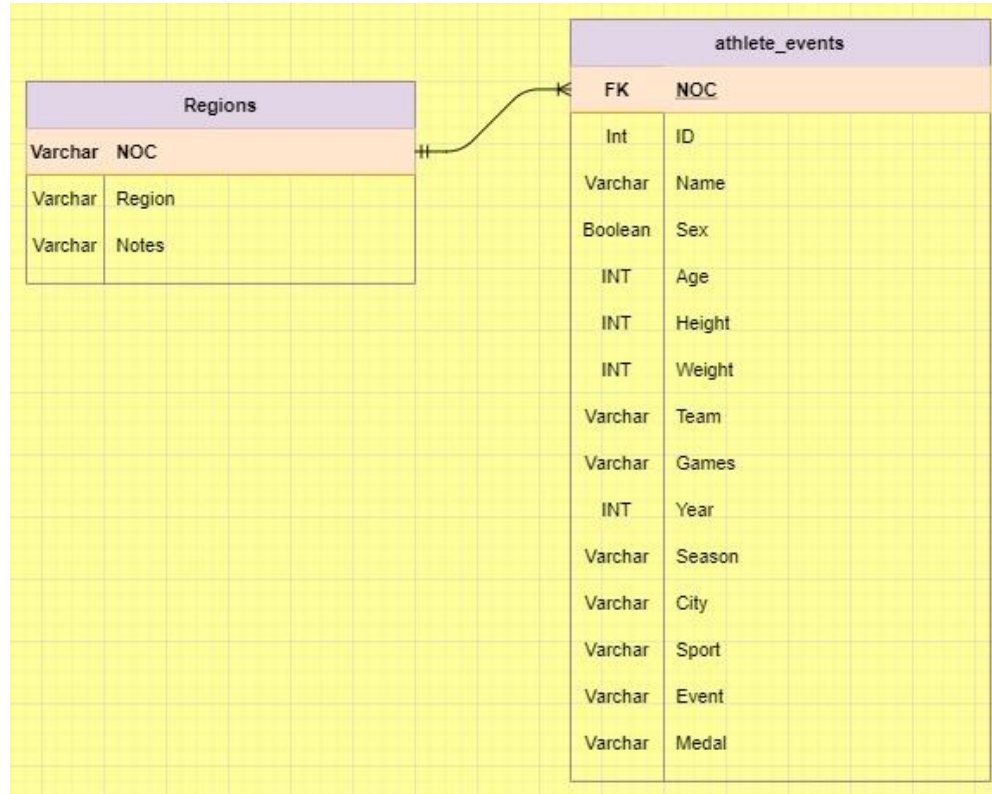
[27]

```
fig = px.bar(Gender, x = 'Sex', y = 'Gender', color = 'Gender')
fig.show()
```

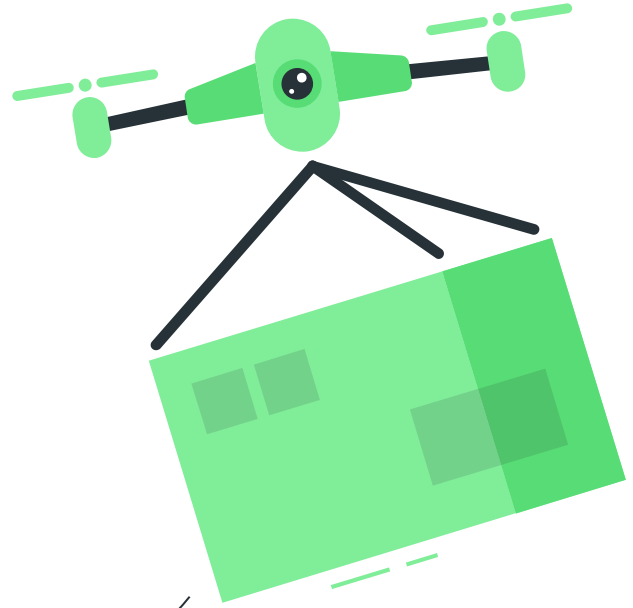


[-]

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



Develop Project Proposal



Description




My Project targets getting past Sports patterns and analysing it. Total team medals.

Get to know more insights on the data, such as when the first event was organized and in which city/country.

This analysis will not only help Sports Coaches to identify patterns and records, but it will also help the SportsStats firm aid in their clients' decision-making.

My audience for the projects would not be limited to Coaches/Trainers but also players who will be able to see their records/performance in past events.



Questions



- When was the first season ever conducted
- How many total medals were distributed
- Age Distribution
- Which sports were in the First Game
- All of the athlete events conducted
- Which country had highest number of Players

.



Hypothesis

- Women have higher number of Medals
- Year > 1956 will have the Highest number of Events
- More people have participated in Football
- 3 Teams will have medals > 40
- People with Age > 40 have received medal in any of the events