Contents:



# Project Info

**Individual**： Ading

**Topic**: What's the interesting thing in these FIFA World Cup

**Tableau Link**： https://public.tableau.com/views/FIFAPlayers2019andFIFAWorldCup--Ading/FIFA-WorldCup1930-2018FIFAPlayer2019?:language=zh-CN&publish=yes&:display_count=n&:origin=viz_share_link

**Python Code**: `big_data_viz`

# Step1: data source：

1. FIFA-2019.csv

   This Dataset Contains all Information of players who participated in the FIFA-2019 League.
   https://www.kaggle.com/datasets/devansodariya/football-fifa-2019-dataset/download

2. FIFA world cup 1930-2018 (21 csv files in total)
   This Dataset consists of Records from all the previous Football World Cups (1930 to 2018)
   https://www.kaggle.com/datasets/iamsouravbanerjee/fifa-football-world-cup-dataset/download

---

# Step2: Data cleansing ,Dimension selection,Data Restruction

## 1. for the csv file: FIFA-2019.csv

There are 90 Features of a record, containing almost every aspect of performance of a player like player's name, skills, age, club, wage, potential, etc.

```
      Unnamed: 0      ID  ... GKReflexes  Release Clause
5376         5376  201908  ...       12.0           €3.9M
7740         7740  187925  ...       14.0           €1.2M
16081       16081  241552  ...        5.0           €630K
11903       11903  243675  ...       65.0           €1.6M
```

After **cleansing** and **dimension selection**....I got this: `fifa-2019-cleansed.csv`

```
Name,Age,Potential,Body Type,Value,Wage
L. Messi,31,94,Messi,110500000.00,565000.00
Cristiano Ronaldo,33,94,C. Ronaldo,77000000.00,405000.00
Neymar Jr,26,93,Neymar,118500000.00,290000.00
De Gea,27,93,Lean,72000000.00,260000.00
K. De Bruyne,27,92,Normal,102000000.00,355000.00
```

create mysql table: `fifa_player_2019`

```
CREATE TABLE `fifa_player_2019` (
  `id` int NOT NULL AUTO_INCREMENT,
  `name` varchar(255) COLLATE utf8mb4_unicode_ci DEFAULT NULL,
  `age` int DEFAULT NULL,
  `potential` int DEFAULT NULL,
  `body_type` varchar(255) CHARACTER SET utf8mb4 COLLATE utf8mb4_unicode_ci
DEFAULT NULL,
  `value` bigint DEFAULT NULL,
  `wage` bigint DEFAULT NULL,
  PRIMARY KEY (`id`)
) ENGINE=InnoDB AUTO_INCREMENT=18208 DEFAULT CHARSET=utf8mb4
COLLATE=utf8mb4_unicode_ci;
```

Then , **By python**(pymysql) , **I insert these rows** into MySQL db ( dataviz2022.`fifa_player_2019`)

| id | name | age | potential | body_type | value | wage |
|---|---|---|---|---|---|---|
| 1 | L. Messi | 31 | 94 | Messi | 110500000 | 565000 |
| 2 | Cristiano Ronaldo | 33 | 94 | C. Ronaldo | 77000000 | 405000 |
| 3 | Neymar Jr | 26 | 93 | Neymar | 118500000 | 290000 |
| 4 | De Gea | 27 | 93 | Lean | 72000000 | 260000 |
| 5 | K. De Bruyne | 27 | 92 | Normal | 102000000 | 355000 |
| 6 | E. Hazard | 27 | 91 | Normal | 93000000 | 340000 |
| 7 | L. Modrić | 32 | 91 | Lean | 67000000 | 420000 |
| 8 | L. Suárez | 31 | 91 | Normal | 80000000 | 455000 |
| 9 | Sergio Ramos | 32 | 91 | Normal | 51000000 | 380000 |
| 10 | J. Oblak | 25 | 93 | Normal | 68000000 | 94000 |

## 2. for a group of csv files: fifa-1930-2018 (21 files in total)

Take a glance at one of the 21 files : `FIFA - 1930.CSV`

```
Position,Team,Games Played,Win,Draw,Loss,Goals For,Goals Against,Goal
Difference,Points
1,Uruguay,4,4,0,0,15,3,12,8
```

There are 10 fields of these FIFA - XXX.csv, and 13-32 rows for each file.

And I'd like to correlate them into one file, so I need to add a new field called year.

Then I write python codes to achieve it and create a mysql table named `fifa_1930_2018` to store the data structurally.

Create SQL:

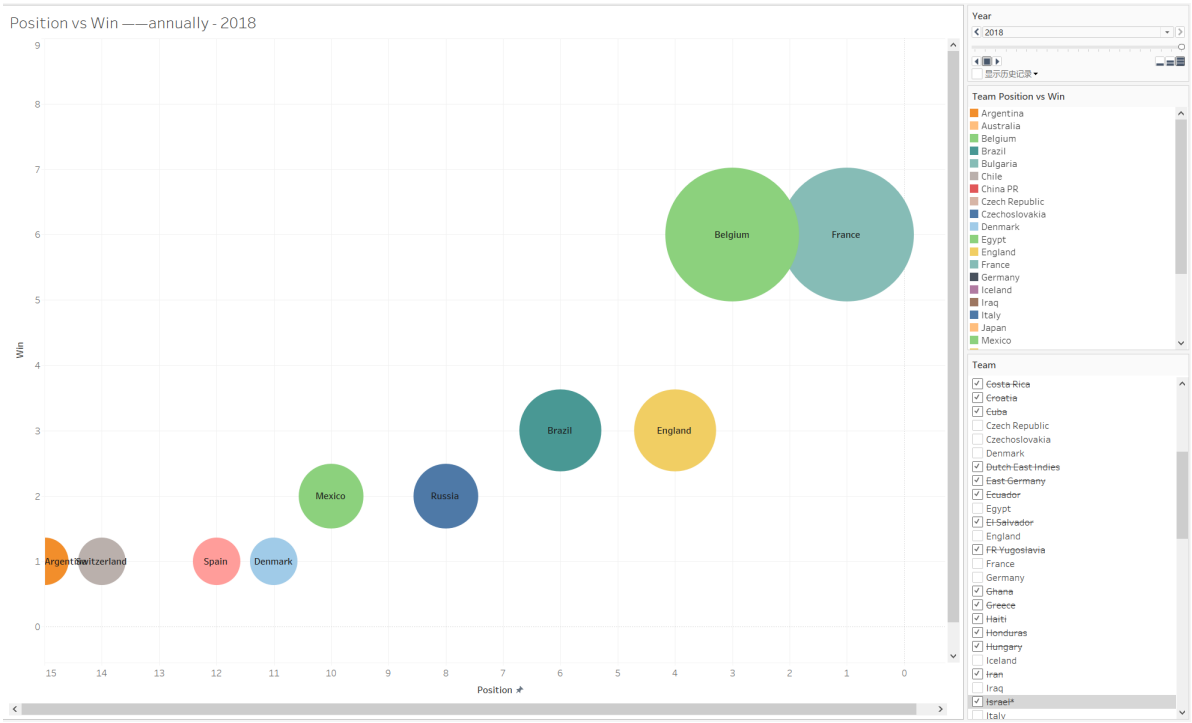
```
CREATE TABLE `fifa_1930_2018` (
  `position` smallint DEFAULT NULL COMMENT 'team rank',
  `team` varchar(255) COLLATE utf8mb4_unicode_ci DEFAULT NULL,
  `games_played` smallint DEFAULT NULL COMMENT 'number of games that a team in the field',
  `win` smallint DEFAULT NULL,
  `draw` smallint DEFAULT NULL,
  `loss` smallint DEFAULT NULL,
  `goals_for` int DEFAULT NULL,
  `goals_against` int DEFAULT NULL,
  `goals_difference` varchar(5) COLLATE utf8mb4_unicode_ci DEFAULT NULL,
  `points` int DEFAULT NULL,
  `year` year DEFAULT NULL
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_unicode_ci;
```

| position | team | games_played | win | draw | loss | goals_for | goals_against | goals_difference | points | year |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Uruguay | 4 | 4 | 0 | 0 | 15 | 3 | 12 | 8 | 1930 |
| 2 | Argentin | 5 | 4 | 0 | 1 | 18 | 9 | 9 | 8 | 1930 |
| 3 | United St | 3 | 2 | 0 | 1 | 7 | 6 | 1 | 4 | 1930 |
| 4 | Yugoslav | 3 | 2 | 0 | 1 | 7 | 7 | 0 | 4 | 1930 |
| 5 | Chile | 3 | 2 | 0 | 1 | 5 | 3 | 2 | 4 | 1930 |
| 6 | Brazil | 2 | 1 | 0 | 1 | 5 | 2 | 3 | 2 | 1930 |
| 7 | France | 3 | 1 | 0 | 2 | 4 | 3 | 1 | 2 | 1930 |
| 8 | Romania | 2 | 1 | 0 | 1 | 3 | 5 | -2 | 2 | 1930 |
| 9 | Paraguay | 2 | 1 | 0 | 1 | 1 | 3 | -2 | 2 | 1930 |
| 10 | Peru | 2 | 0 | 0 | 2 | 1 | 4 | -3 | 0 | 1930 |

# Step3: Data Viz

I create 4 worksheets and 1 dashboard

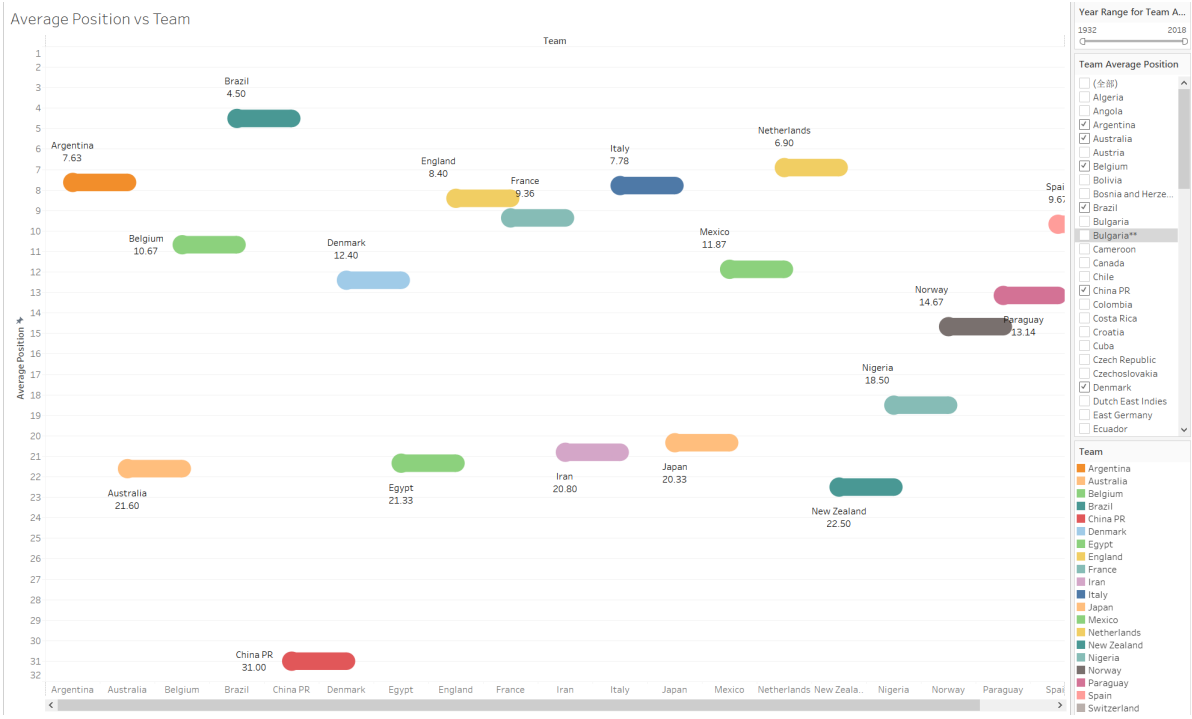
1. `Position vs Win - annually`




Position vs Win –2018

User can filter teams and year to customize data viz

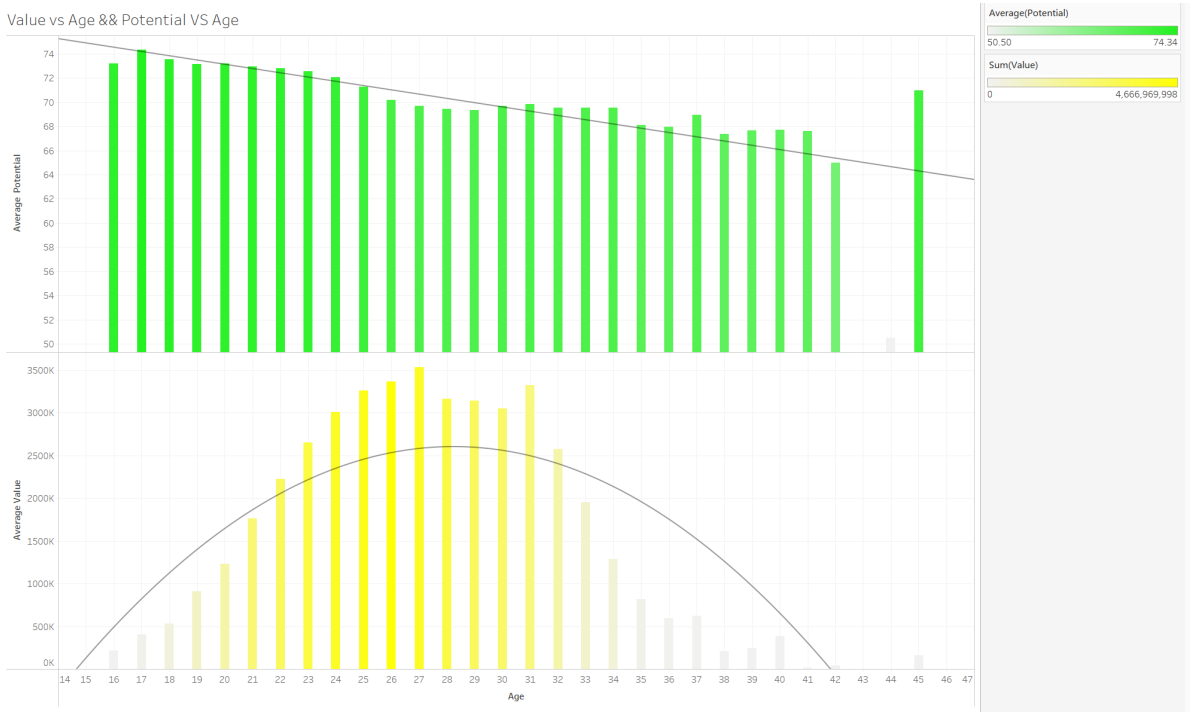
2. `Average Position vs Team`




User can filter teams and year range to customize data viz

We can see from the range of 1932 to 2018 that Brazil takes the top average position and China PR takes the embarrassing average position.

3. `Value vs Age && Potential VS Age`




We can observe that (from the green bar graph) the Age is inversely proportional [1] to the Potential of the player

A player shows average peak potential at the age of 17 , so I should have started to play football/soccer at the age of 17

And there is an outlier in the viz, at the age of 44, the player ‘s potential is suddenly decreasing .Because there one player that is 44 years old in the dataset.

---

We can observe that (from the golden bar chart) at the age between 22 to 31, players are most valuable. And after 31 years old , a player’s value  is sharply dropping.

Well this is an average situation. There are still many football star who are older than 31 but still being valuable such as Lionel Messi (34 ), C Ronaldo (37) and Zlatan Ibrahimovic (41)

4. `Team Position Trend`


Here is plotting team position trend. User can select different team (country) and year to customize data viz

We can see that at 2002, team China PR (the red point at the bottom) (Chinese Men's National Soccer Team) finally showed up.

It occurs to me that this is the only time that China PR team ever showed up in World Cup. That's upset.

---

1. adj. 成比例的；相称的，协调的；（数）成常比的 ↩