# ETL-Project

### Our Team:

- Aditya Bhatnagar
- Jonathan Little
- Michael Legg
- Rosario Chiovaro

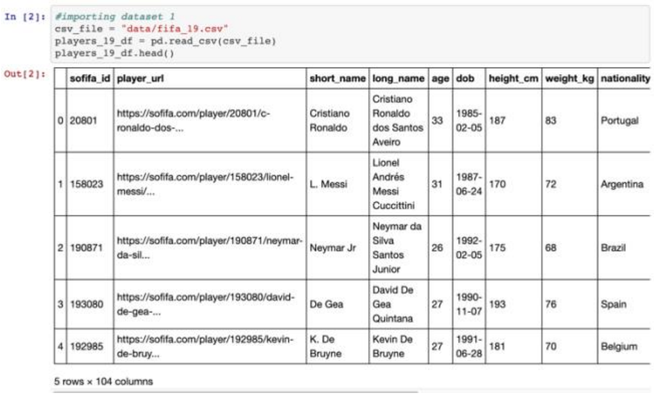
### Project:

Comparing real life season statistics of soccer players to their individual ratings in the corresponding FIFA video game year.

### Extract:

The data was extracted from the source (link below) in the form of CSV's to be later transformed and loaded on a clean database using Pandas and Postgres.

- https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset (CSV)
- https://data.world/cclayford/statbunker-football-statistics (CSV)

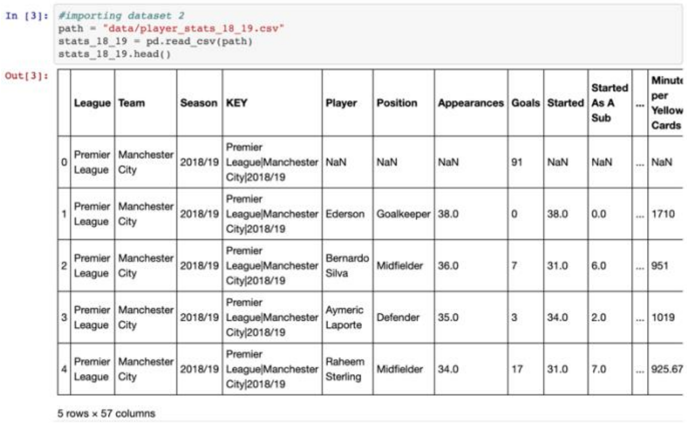


```
In [2]: #importing dataset 1
        csv_file = "data/fifa_19.csv"
        players_19_df = pd.read_csv(csv_file)
        players_19_df.head()
```

Out[2]:

| | sofifa_id | player_url | short_name | long_name | age | dob | height_cm | weight_kg | nationality |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 20801 | https://sofifa.com/player/20801/c-ronaldo-dos-... | Cristiano Ronaldo | Cristiano Ronaldo dos Santos Aveiro | 33 | 1985-02-05 | 187 | 83 | Portugal |
| 1 | 158023 | https://sofifa.com/player/158023/lionel-messi/... | L. Messi | Lionel Andrés Messi Cuccittini | 31 | 1987-06-24 | 170 | 72 | Argentina |
| 2 | 190871 | https://sofifa.com/player/190871/neymar-da-sil... | Neymar Jr | Neymar da Silva Santos Junior | 26 | 1992-02-05 | 175 | 68 | Brazil |
| 3 | 193080 | https://sofifa.com/player/193080/david-de-gea-... | De Gea | David De Gea Quintana | 27 | 1990-11-07 | 193 | 76 | Spain |
| 4 | 192985 | https://sofifa.com/player/192985/kevin-de-bruy... | K. De Bruyne | Kevin De Bruyne | 27 | 1991-06-28 | 181 | 70 | Belgium |

5 rows × 104 columns

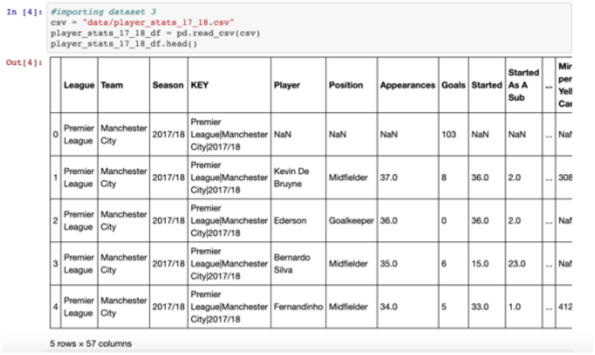


```
In [3]: #importing dataset 2
        path = "data/player_stats_18_19.csv"
        stats_18_19 = pd.read_csv(path)
        stats_18_19.head()
```

Out[3]:

| | League | Team | Season | KEY | Player | Position | Appearances | Goals | Started | Started As A Sub | ... | Minute per Yellow Cards |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Premier League | Manchester City | 2018/19 | Premier League\|Manchester City\|2018/19 | NaN | NaN | NaN | 91 | NaN | NaN | ... | NaN |
| 1 | Premier League | Manchester City | 2018/19 | Premier League\|Manchester City\|2018/19 | Ederson | Goalkeeper | 38.0 | 0 | 38.0 | 0.0 | ... | 1710 |
| 2 | Premier League | Manchester City | 2018/19 | Premier League\|Manchester City\|2018/19 | Bernardo Silva | Midfielder | 36.0 | 7 | 31.0 | 6.0 | ... | 951 |
| 3 | Premier League | Manchester City | 2018/19 | Premier League\|Manchester City\|2018/19 | Aymeric Laporte | Defender | 35.0 | 3 | 34.0 | 2.0 | ... | 1019 |
| 4 | Premier League | Manchester City | 2018/19 | Premier League\|Manchester City\|2018/19 | Raheem Sterling | Midfielder | 34.0 | 17 | 31.0 | 7.0 | ... | 925.67 |

5 rows × 57 columns

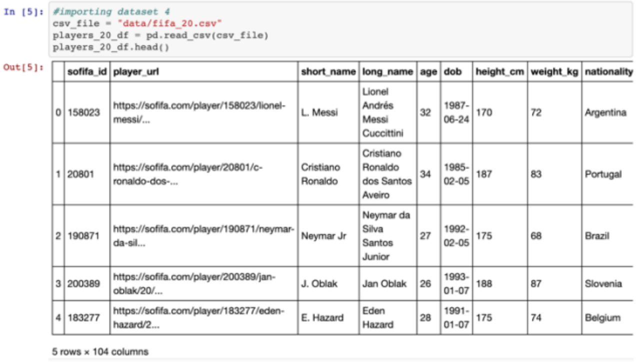


```
In [4]: #importing dataset 3
        csv = "data/player_stats_17_18.csv"
        player_stats_17_18_df = pd.read_csv(csv)
        player_stats_17_18_df.head()
```

Out[4]:

| | League | Team | Season | KEY | Player | Position | Appearances | Goals | Started | Started As A Sub | ... | Mi per Yell Ca |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Premier League | Manchester City | 2017/18 | Premier League\|Manchester City\|2017/18 | NaN | NaN | NaN | 103 | NaN | NaN | ... | Na |
| 1 | Premier League | Manchester City | 2017/18 | Premier League\|Manchester City\|2017/18 | Kevin De Bruyne | Midfielder | 37.0 | 8 | 36.0 | 2.0 | ... | 308 |
| 2 | Premier League | Manchester City | 2017/18 | Premier League\|Manchester City\|2017/18 | Ederson | Goalkeeper | 36.0 | 0 | 36.0 | 2.0 | ... | Na |
| 3 | Premier League | Manchester City | 2017/18 | Premier League\|Manchester City\|2017/18 | Bernardo Silva | Midfielder | 35.0 | 6 | 15.0 | 23.0 | ... | Na |
| 4 | Premier League | Manchester City | 2017/18 | Premier League\|Manchester City\|2017/18 | Fernandinho | Midfielder | 34.0 | 5 | 33.0 | 1.0 | ... | 412 |

5 rows × 57 columns



```
In [5]: #importing dataset 4
        csv_file = "data/fifa_20.csv"
        players_20_df = pd.read_csv(csv_file)
        players_20_df.head()
```

Out[5]:

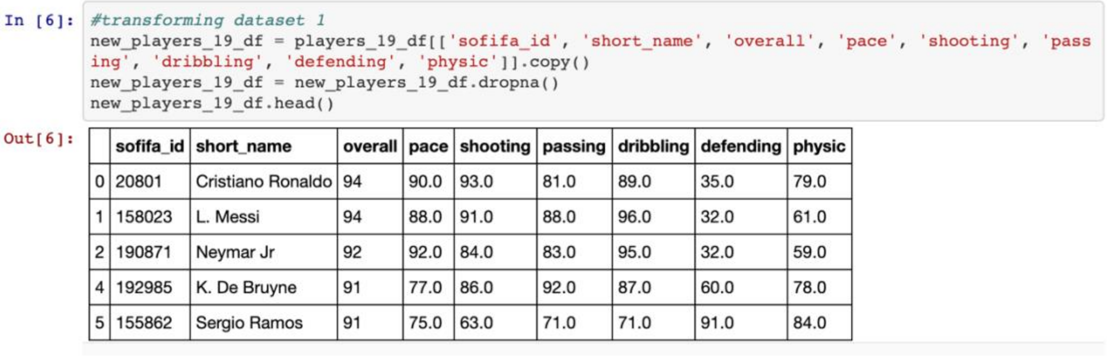
| | sofifa_id | player_url | short_name | long_name | age | dob | height_cm | weight_kg | nationality |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 158023 | https://sofifa.com/player/158023/lionel-messi/... | L. Messi | Lionel Andrés Messi Cuccittini | 32 | 1987-06-24 | 170 | 72 | Argentina |
| 1 | 20801 | https://sofifa.com/player/20801/c-ronaldo-dos-... | Cristiano Ronaldo | Cristiano Ronaldo dos Santos Aveiro | 34 | 1985-02-05 | 187 | 83 | Portugal |
| 2 | 190871 | https://sofifa.com/player/190871/neymar-da-sil... | Neymar Jr | Neymar da Silva Santos Junior | 27 | 1992-02-05 | 175 | 68 | Brazil |
| 3 | 200389 | https://sofifa.com/player/200389/jan-oblak/20/... | J. Oblak | Jan Oblak | 26 | 1993-01-07 | 188 | 87 | Slovenia |
| 4 | 183277 | https://sofifa.com/player/183277/eden-hazard/2... | E. Hazard | Eden Hazard | 28 | 1991-01-07 | 175 | 74 | Belgium |

5 rows × 104 columns

## Transform:

Primarily the data cleaning required the below activities:

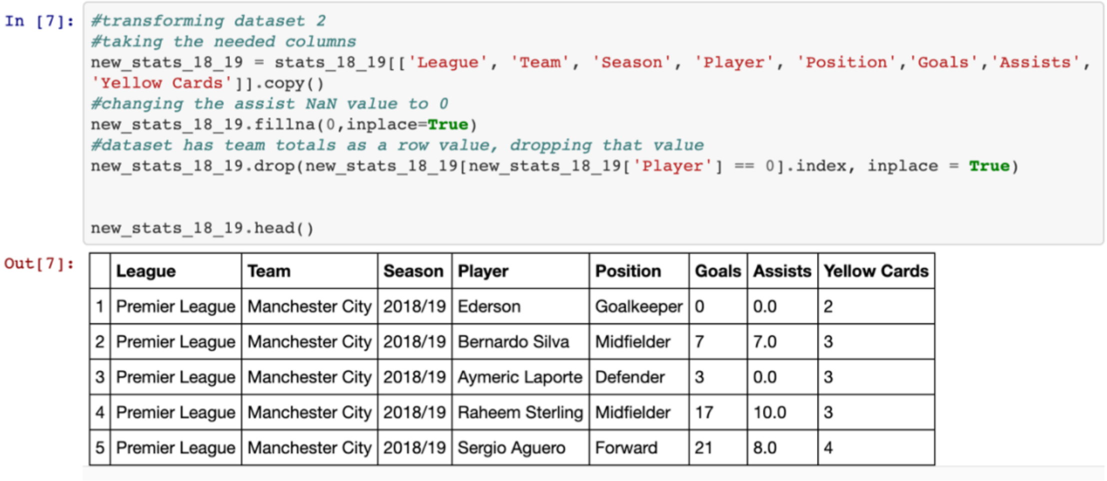
- Selecting the necessary columns
- Dropping null (NaN) values
- Changing null (NaN) values to 0



In [6]:
```
#transforming dataset 1
new_players_19_df = players_19_df[['sofifa_id', 'short_name', 'overall', 'pace', 'shooting', 'pass
ing', 'dribbling', 'defending', 'physic']].copy()
new_players_19_df = new_players_19_df.dropna()
new_players_19_df.head()
```

Out[6]:

| | sofifa_id | short_name | overall | pace | shooting | passing | dribbling | defending | physic |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 20801 | Cristiano Ronaldo | 94 | 90.0 | 93.0 | 81.0 | 89.0 | 35.0 | 79.0 |
| 1 | 158023 | L. Messi | 94 | 88.0 | 91.0 | 88.0 | 96.0 | 32.0 | 61.0 |
| 2 | 190871 | Neymar Jr | 92 | 92.0 | 84.0 | 83.0 | 95.0 | 32.0 | 59.0 |
| 4 | 192985 | K. De Bruyne | 91 | 77.0 | 86.0 | 92.0 | 87.0 | 60.0 | 78.0 |
| 5 | 155862 | Sergio Ramos | 91 | 75.0 | 63.0 | 71.0 | 71.0 | 91.0 | 84.0 |

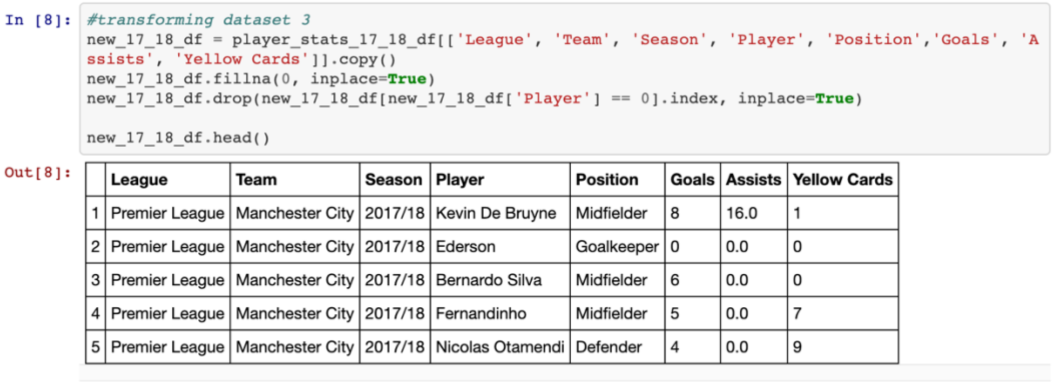


In [7]:
```
#transforming dataset 2
#taking the needed columns
new_stats_18_19 = stats_18_19[['League', 'Team', 'Season', 'Player', 'Position','Goals','Assists',
'Yellow Cards']].copy()
#changing the assist NaN value to 0
new_stats_18_19.fillna(0,inplace=True)
#dataset has team totals as a row value, dropping that value
new_stats_18_19.drop(new_stats_18_19[new_stats_18_19['Player'] == 0].index, inplace = True)


new_stats_18_19.head()
```

Out[7]:

| | League | Team | Season | Player | Position | Goals | Assists | Yellow Cards |
|---|---|---|---|---|---|---|---|---|
| 1 | Premier League | Manchester City | 2018/19 | Ederson | Goalkeeper | 0 | 0.0 | 2 |
| 2 | Premier League | Manchester City | 2018/19 | Bernardo Silva | Midfielder | 7 | 7.0 | 3 |
| 3 | Premier League | Manchester City | 2018/19 | Aymeric Laporte | Defender | 3 | 0.0 | 3 |
| 4 | Premier League | Manchester City | 2018/19 | Raheem Sterling | Midfielder | 17 | 10.0 | 3 |
| 5 | Premier League | Manchester City | 2018/19 | Sergio Aguero | Forward | 21 | 8.0 | 4 |

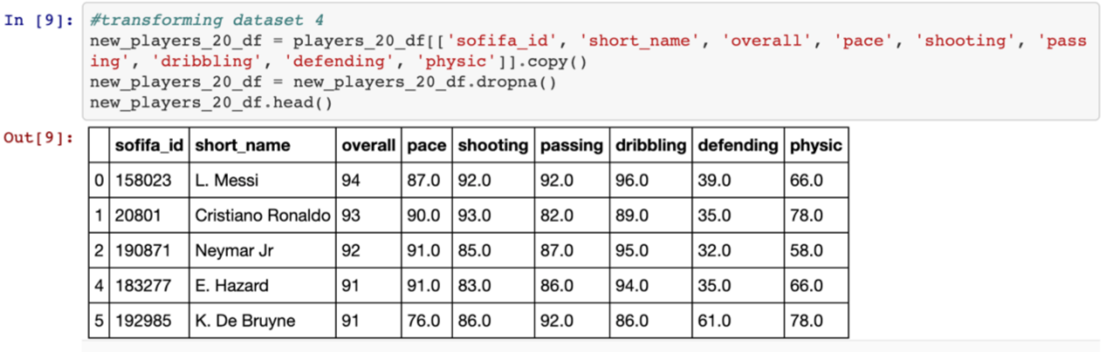


In [8]:
```
#transforming dataset 3
new_17_18_df = player_stats_17_18_df[['League', 'Team', 'Season', 'Player', 'Position','Goals', 'A
ssists', 'Yellow Cards']].copy()
new_17_18_df.fillna(0, inplace=True)
new_17_18_df.drop(new_17_18_df[new_17_18_df['Player'] == 0].index, inplace=True)

new_17_18_df.head()
```

Out[8]:

| | League | Team | Season | Player | Position | Goals | Assists | Yellow Cards |
|---|---|---|---|---|---|---|---|---|
| 1 | Premier League | Manchester City | 2017/18 | Kevin De Bruyne | Midfielder | 8 | 16.0 | 1 |
| 2 | Premier League | Manchester City | 2017/18 | Ederson | Goalkeeper | 0 | 0.0 | 0 |
| 3 | Premier League | Manchester City | 2017/18 | Bernardo Silva | Midfielder | 6 | 0.0 | 0 |
| 4 | Premier League | Manchester City | 2017/18 | Fernandinho | Midfielder | 5 | 0.0 | 7 |
| 5 | Premier League | Manchester City | 2017/18 | Nicolas Otamendi | Defender | 4 | 0.0 | 9 |



In [9]:
```
#transforming dataset 4
new_players_20_df = players_20_df[['sofifa_id', 'short_name', 'overall', 'pace', 'shooting', 'pass
ing', 'dribbling', 'defending', 'physic']].copy()
new_players_20_df = new_players_20_df.dropna()
new_players_20_df.head()
```

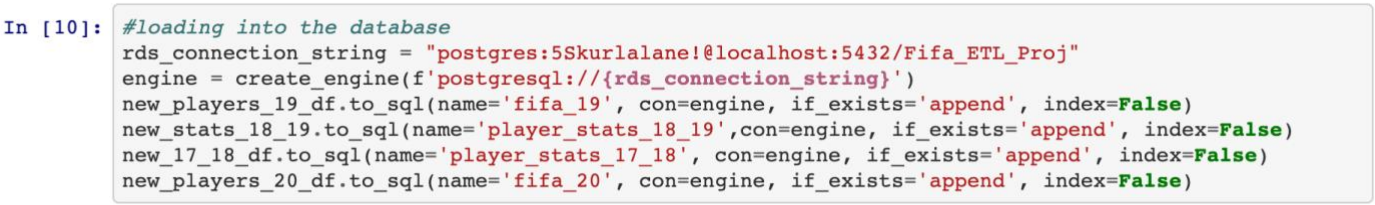
Out[9]:

| | sofifa_id | short_name | overall | pace | shooting | passing | dribbling | defending | physic |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 158023 | L. Messi | 94 | 87.0 | 92.0 | 92.0 | 96.0 | 39.0 | 66.0 |
| 1 | 20801 | Cristiano Ronaldo | 93 | 90.0 | 93.0 | 82.0 | 89.0 | 35.0 | 78.0 |
| 2 | 190871 | Neymar Jr | 92 | 91.0 | 85.0 | 87.0 | 95.0 | 32.0 | 58.0 |
| 4 | 183277 | E. Hazard | 91 | 91.0 | 83.0 | 86.0 | 94.0 | 35.0 | 66.0 |
| 5 | 192985 | K. De Bruyne | 91 | 76.0 | 86.0 | 92.0 | 86.0 | 61.0 | 78.0 |

## Load:

The data was loaded on a SQL database using the procedure shown as below and by creating respective database on PostgresQL

can be used effectively to query the data to compare FIFA video game attributes to real life statistics.



In [10]:
```
#loading into the database
rds_connection_string = "postgres:5Skurlalane!@localhost:5432/Fifa_ETL_Proj"
engine = create_engine(f'postgresql://{rds_connection_string}')
new_players_19_df.to_sql(name='fifa_19', con=engine, if_exists='append', index=False)
new_stats_18_19.to_sql(name='player_stats_18_19',con=engine, if_exists='append', index=False)
new_17_18_df.to_sql(name='player_stats_17_18', con=engine, if_exists='append', index=False)
new_players_20_df.to_sql(name='fifa_20', con=engine, if_exists='append', index=False)
```