Capstone Project : FIFA Market Value & Wage

Problem definition

Nowadays, Football is the most popular sport in the world. The financial in this industry is very massive and rapid growth. So, it's very interesting to analyze and consider for the subject.

Among the cost in the industry, the most spending is related the market transfer value and wage for the football player. As a result, most of the clubs in the world would need to reduce this cost and spend only on the proper contract. Moreover, it's become serious when the FIFA announce the 'Financial Fair Play' rules to limit the spending on buying or hiring.

We craft this project to find the insight of average Player's Market value and Wage base on each dimension such as the Football League or Player Position to help the Management level on making decision for the contract agreement.

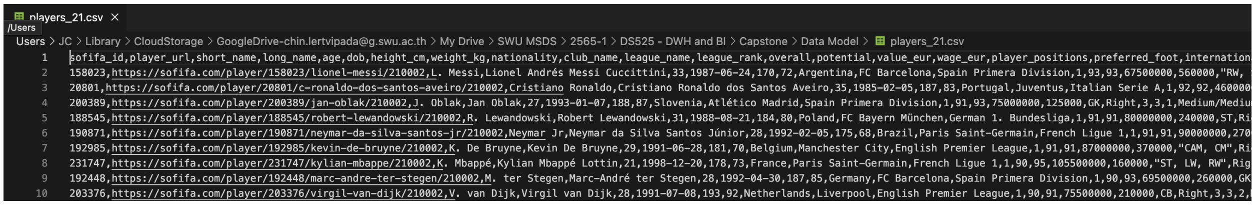
Data source

FIFA 21 complete player dataset :

https://www.kaggle.com/datasets/stefanoleone992/fifa-21-complete-player-dataset

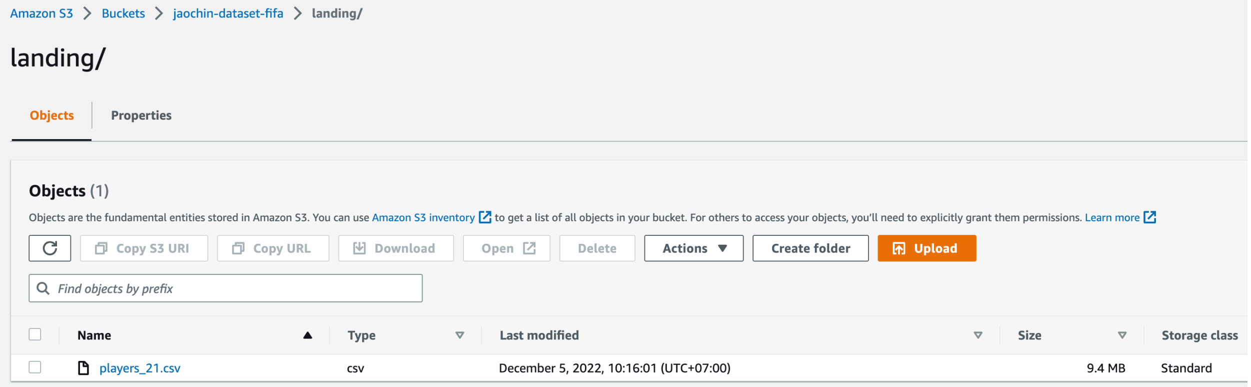
Raw data

players_21.csv :


players_21.csv ×
/Users
Users > JC > Library > CloudStorage > GoogleDrive-chin.lertvipada@g.swu.ac.th > My Drive > SWU MSDS > 2565-1 > DS525 - DWH and BI > Capstone > Data Model > players_21.csv

```
sofifa_id,player_url,short_name,long_name,age,dob,height_cm,weight_kg,nationality,club_name,league_name,league_rank,overall,potential,value_eur,wage_eur,player_positions,preferred_foot,internation
158023,https://sofifa.com/player/158023/lionel-messi/210002,L. Messi,Lionel Andrés Messi Cuccittini,33,1987-06-24,170,72,Argentina,FC Barcelona,Spain Primera Division,1,93,93,67500000,560000,"RW, 
20801,https://sofifa.com/player/20801/c-ronaldo-dos-santos-aveiro/210002,Cristiano Ronaldo,Cristiano Ronaldo dos Santos Aveiro,35,1985-02-05,187,83,Portugal,Juventus,Italian Serie A,1,92,92,460000
200389,https://sofifa.com/player/200389/jan-oblak/210002,J. Oblak,Jan Oblak,27,1993-01-07,188,87,Slovenia,Atlético Madrid,Spain Primera Division,1,91,93,75000000,125000,GK,Right,3,3,1,Medium/Mediu
188545,https://sofifa.com/player/188545/robert-lewandowski/210002,R. Lewandowski,Robert Lewandowski,31,1988-08-21,184,80,Poland,FC Bayern München,German 1. Bundesliga,1,91,91,80000000,240000,ST,Ri
190871,https://sofifa.com/player/190871/neymar-da-silva-santos-jr/210002,Neymar Jr,Neymar da Silva Santos Júnior,28,1992-02-05,175,68,Brazil,Paris Saint-Germain,French Ligue 1,1,91,91,90000000,270
192985,https://sofifa.com/player/192985/kevin-de-bruyne/210002,K. De Bruyne,Kevin De Bruyne,29,1991-06-28,181,70,Belgium,Manchester City,English Premier League,1,91,91,87000000,370000,"CAM, CM",Ri
231747,https://sofifa.com/player/231747/kylian-mbappe/210002,K. Mbappé,Kylian Mbappé Lottin,21,1998-12-20,178,73,France,Paris Saint-Germain,French Ligue 1,1,90,95,105500000,160000,"ST, LW, RW",Rig
192448,https://sofifa.com/player/192448/marc-andre-ter-stegen/210002,M. ter Stegen,Marc-André ter Stegen,28,1992-04-30,187,85,Germany,FC Barcelona,Spain Primera Division,1,90,93,69500000,260000,GK
203376,https://sofifa.com/player/203376/virgil-van-dijk/210002,V. van Dijk,Virgil van Dijk,28,1991-07-08,193,92,Netherlands,Liverpool,English Premier League,1,90,91,75500000,210000,CB,Right,3,3,2,
```

** The raw data file is stored in the AWS S3

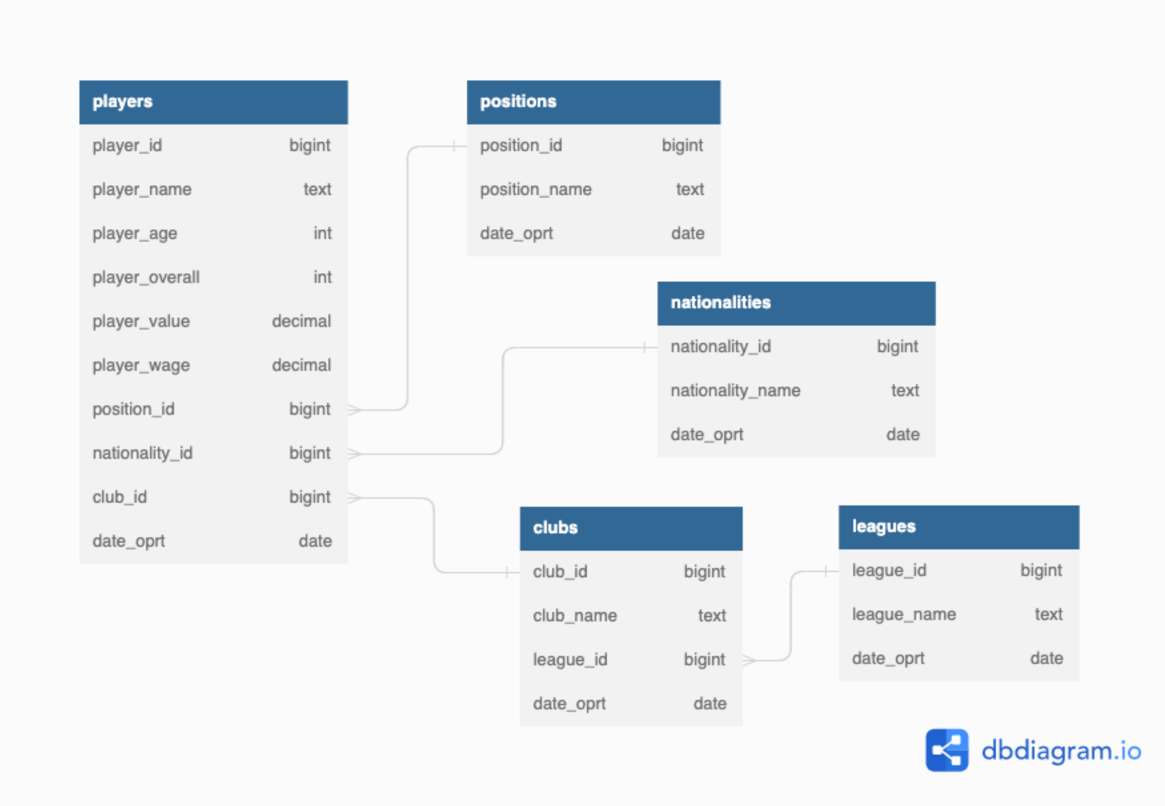


Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name ▲	Type	Last modified	Size	Storage class
□	players_21.csv	csv	December 5, 2022, 10:16:01 (UTC+07:00)	9.4 MB	Standard

Data model

Datalake :



players	
player_id	bigint
player_name	text
player_age	int
player_overall	int
player_value	decimal
player_wage	decimal
position_id	bigint
nationality_id	bigint
club_id	bigint
date_oprt	date

positions	
position_id	bigint
position_name	text
date_oprt	date

nationalities	
nationality_id	bigint
nationality_name	text
date_oprt	date

clubs	
club_id	bigint
club_name	text
league_id	bigint
date_oprt	date

leagues	
league_id	bigint
league_name	text
date_oprt	date

Leagues : Collect the football league information.

Clubs : Collect the football club information. Appear the relationship with Leagues because each club must be in a league.

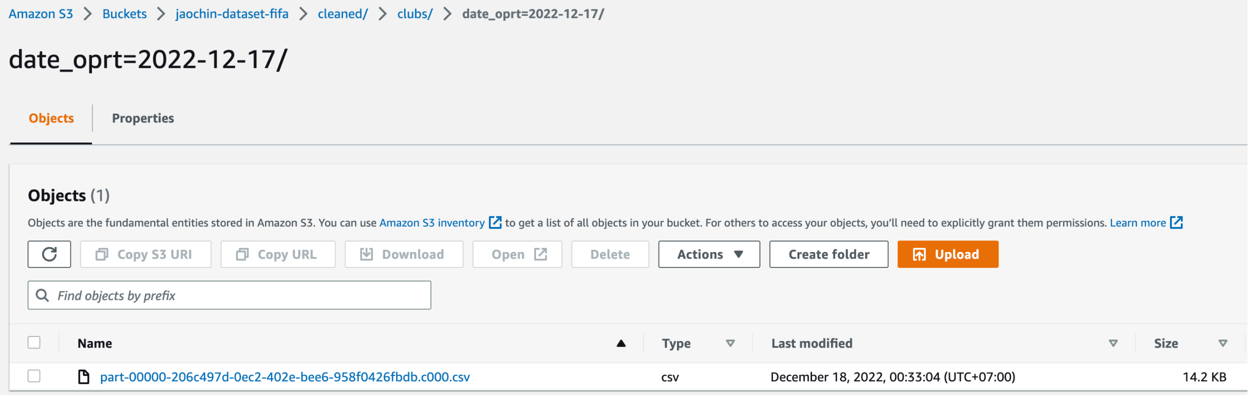
Nationalities : Collect the nationality which belong to existing football player from source data.

Positions : Collect the player position, played for the club.

Players : Collect the player personal information including transfer value and wage. Appear the relationship with Positions, Nationalities and Clubs to identify the dimensions for this player.

DS525 – Capstone Project - Chin Lertvipada - 64199130039

** Datalake tables are stored in csv format in AWS S3 with partitioning on “date_oprt”

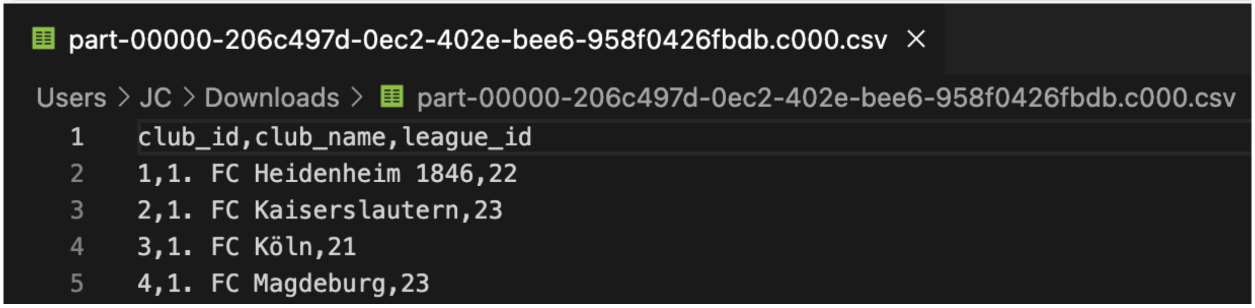


date_oprt=2022-12-17/

Objects (1)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

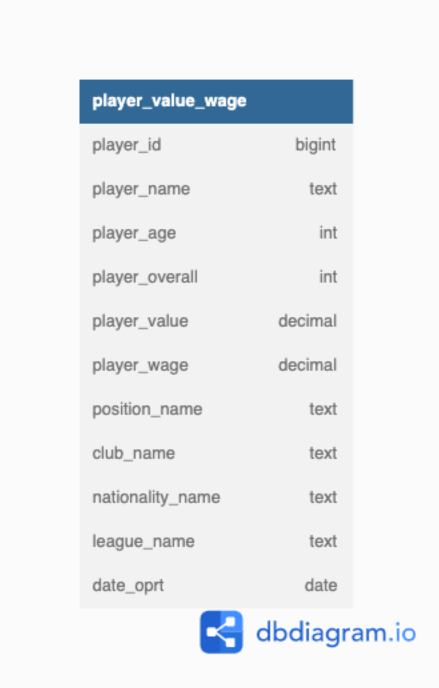
Name	Type	Last modified	Size
part-00000-206c497d-0ec2-402e-bee6-958f0426fbdb.c000.csv	csv	December 18, 2022, 00:33:04 (UTC+07:00)	14.2 KB



part-00000-206c497d-0ec2-402e-bee6-958f0426fbdb.c000.csv
Users > JC > Downloads > part-00000-206c497d-0ec2-402e-bee6-958f0426fbdb.c000.csv

```
club_id,club_name,league_id
1,1. FC Heidenheim 1846,22
2,1. FC Kaiserslautern,23
3,1. FC Köln,21
4,1. FC Magdeburg,23
```

Datawarehouse :

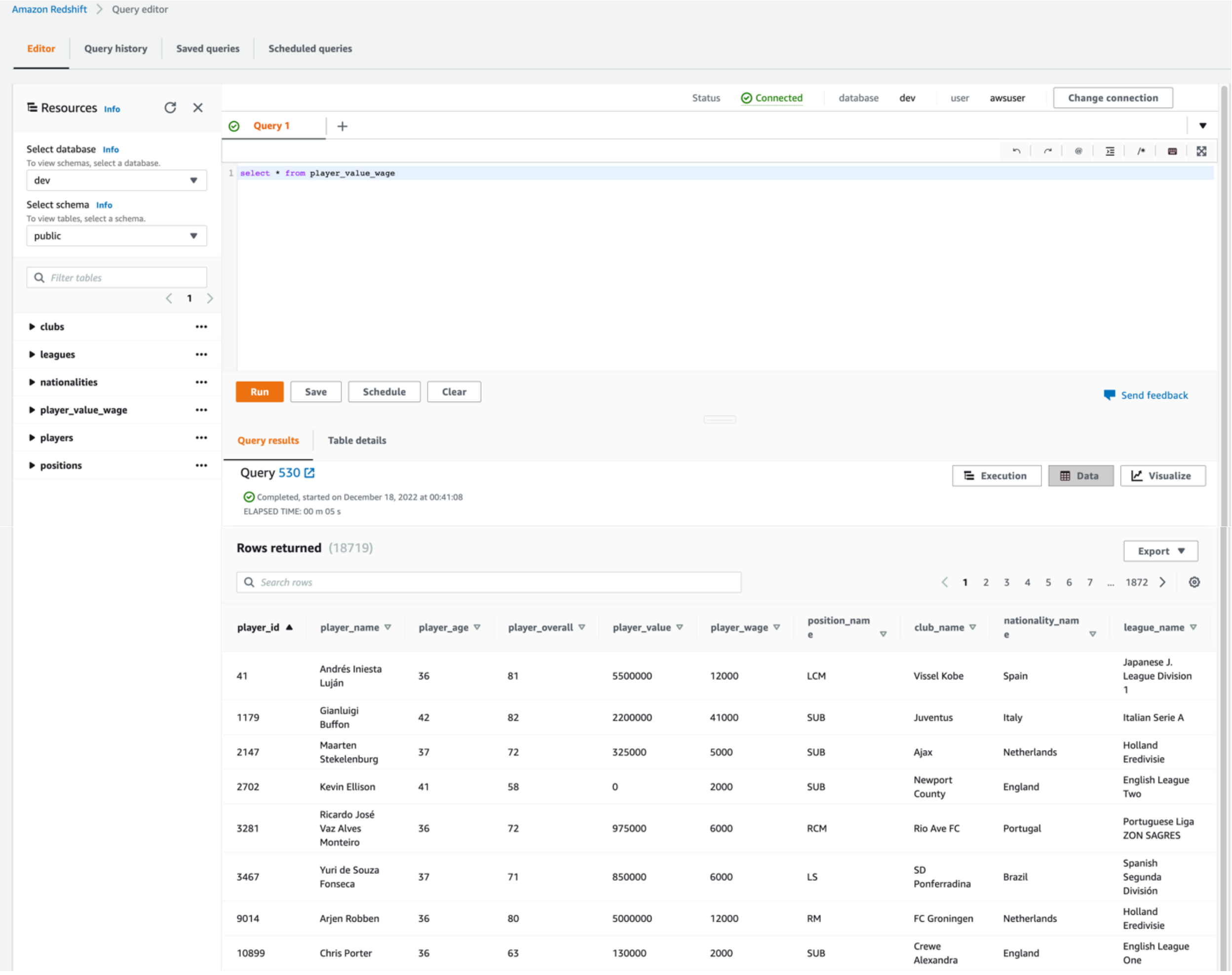


player_value_wage	
player_id	bigint
player_name	text
player_age	int
player_overall	int
player_value	decimal
player_wage	decimal
position_name	text
club_name	text
nationality_name	text
league_name	text
date_oprt	date

Player_value_wage : Collect the information of player including necessary information such as position, club. The purpose of this datawarehouse table is to serve the OLAP processing, Tableau for example.

DS525 – Capstone Project - Chin Lertvipada - 64199130039

** Datawarehouse table is stored in AWS Redshift with "**date_oprt**" as execution date



```
select * from player_value_wage
```

Query 530

Completed, started on December 18, 2022 at 00:41:08
ELAPSED TIME: 00 m 05 s

Rows returned (18719)

player_id	player_name	player_age	player_overall	player_value	player_wage	position_name	club_name	nationality_name	league_name
41	Andrés Iniesta Luján	36	81	5500000	12000	LCM	Vissel Kobe	Spain	Japanese J. League Division 1
1179	Gianluigi Buffon	42	82	2200000	41000	SUB	Juventus	Italy	Italian Serie A
2147	Maarten Stekelenburg	37	72	325000	5000	SUB	Ajax	Netherlands	Holland Eredivisie
2702	Kevin Ellison	41	58	0	2000	SUB	Newport County	England	English League Two
3281	Ricardo José Vaz Alves Monteiro	36	72	975000	6000	RCM	Rio Ave FC	Portugal	Portuguese Liga ZON SAGRES
3467	Yuri de Souza Fonseca	37	71	850000	6000	LS	SD Ponferradina	Brazil	Spanish Segunda División
9014	Arjen Robben	36	80	5000000	12000	RM	FC Groningen	Netherlands	Holland Eredivisie
10899	Chris Porter	36	63	130000	2000	SUB	Crewe Alexandra	England	English League One

Project workflow

1. The AWS Cloud environments setup
    a. S3 to store the *“raw data”* and *“cleaned data”* (Datalake)
    b. Redshift to store the *“OLAP data”* (Datawarehouse)
    c. AWS Credentials for application to access AWS
2. The source data (raw csv file) will be stored in the AWS S3
    a. The source data will be loaded into S3 with other process or team
    b. S3 repository : “jaochin-dataset-fifa/landing/”
3. The Datalake process will load the *“raw data”* and produce the *“cleaned data”* in AWS S3
    a. Data transformation & cleansing by PySpark
    b. Produce 5 tables : Clubs, Leagues, Positions, Nationalities, Players
    c. Output tables are in csv format
    d. Each table (csv file) is partitioned by *“date_oprt”* (execution date)
    e. Example S3 repository : “jaochin-dataset-fifa/cleaned/clubs/date_oprt=2022-12-17/”
4. The Datawarehouse process will load the *“cleaned data”* and produce the *“OLAP data”* in AWS Redshift
    a. Data transformation & load by Python
    b. Monthly schedule at 1st of each month
    c. Load *“cleaned data”* from S3 into Redshift
    d. Load *“cleaned data”* with filtering *“date_oprt”* = execution date (current date)
    e. Merge and transform *“cleaned data”* to produce *“OLAP data”*
    f. Produce datawarehouse table : Player_value_wage
    g. The datawaehouse table is partitioned by *“date_oprt”* (execution date)
5. The dashboard for data visualization using Tableau Desktop
    a. Connect Tableau Desktop to Redshift with information in step **1.b**
    b. Select the database and data which collect the necessary information
    c. Create the dashboard to answer the problem or question definition
    d. Publish the dashboard to the Tableau Public for online access
6. The workflow orchestration using Airflow
    a. The step **1** - **4** should be fully automated and controlled by Airflow
    b. However, due to configuration in many steps are very complicate and consume a lot of time. So, for the moment we decide to use the Airflow to only automate and control for step **4** Datawarehouse process.

Project implementation instruction

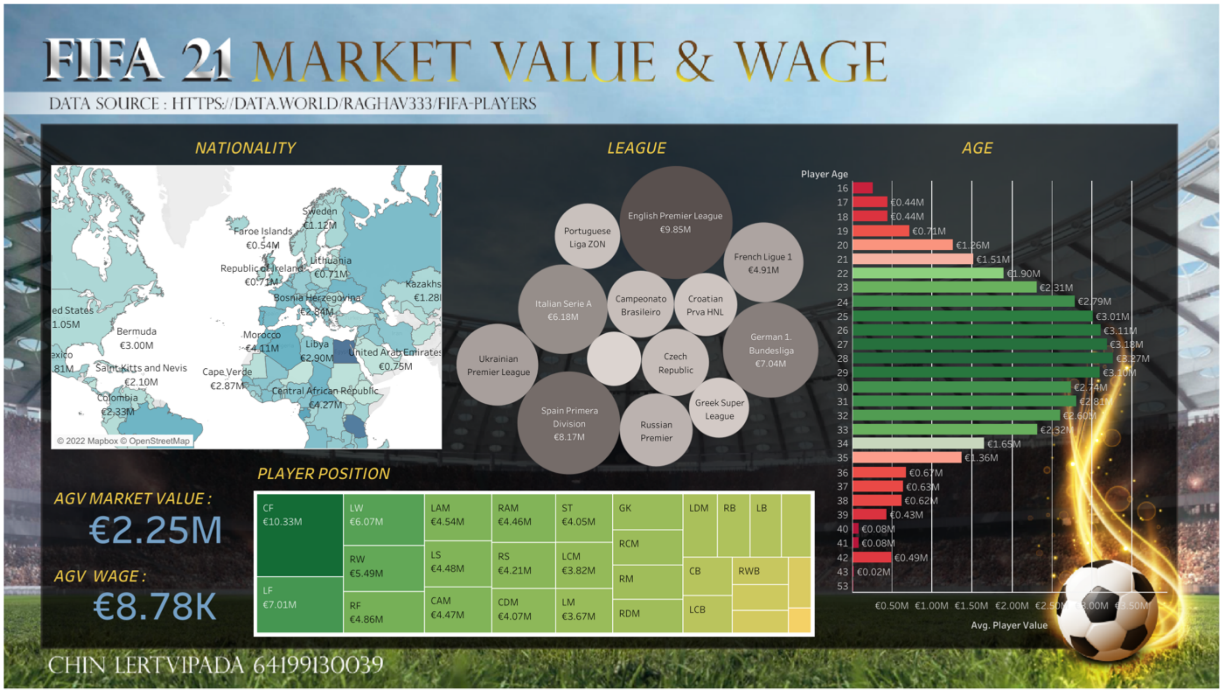
For the full implementation instruction (step-by-step), please find the information here :

https://github.com/chin-lertvipada/swu-ds525/tree/main/chin-capstone#readme

Dashboard

Please find the dashboard online here :

https://public.tableau.com/app/profile/chin.lertvipada/viz/Capstone_csv/FootballMarketValue



The dashboard visualizes “Average Market value & Wage” for football player in many dimensions to answer the questions and solve the problems

- Average Market value & Wage : show the average value & wage among all football player
- Nationality : show the average value & wage for player from their nationality
- League : show the average value & wage by the football league
- Age : show the average value & wage by span of football player age
- Player Position : show the average value & wage by the position that they play for a club

Summary

From the raw data, we create the process and workflow to extract, transform and load (ETL) the data to produce the data standardization for utilization of the data.

We make the flow thru the Datalake process and the Datawarehouse process to build to “OLAP data” which we believe that it will solve the problems and questions.

After that we create the Dashboard to visualize the data to help Management level or any stakeholder for the decision-making process.