

INTRODUCTION TO DATA MANGEMENT PROJECT REPORT

(Project Semester August-December 2019)

FIFIA 19

Submitted by

Yug Ahuja

Registration No.- **11714531**

Programme- **B.Tech (CSE)**

Section- **KM094**

Course Code- **INT217**

Under the Guidance of

Ms. Vasudha

Discipline of CSE/IT

Lovely School of Computer Science & Engineering

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that Yug Ahuja bearing Registration no. 11714531 has completed INT217 project titled, “**FIFA 19**” under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

Signature and Name of the Supervisor

Designation of the Supervisor

School of Computer Science & Engineering

Lovely Professional University

Phagwara, Punjab.

Date: 20/11/19

DECLARATION

I Yug Ahuja, student of Computer Science & Engineering under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date:20th Nov,2019

Registration No.- 11714531

Name of the student

Yug Ahuja

ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to my data management teacher “Ms Vasudha” for her able guidance and support in completing my project. Her suggestions and instructions have served as major contributor towards the successful completion of project.

Lastly, I would like to thank my classmates in resolving queries and problems.

TABLE OF CONTENT

1. <u>Introduction</u>	6
2. <u>Objectives/Scope of the Analysis</u>	7
3. <u>Source of dataset</u>	8
4. <u>ETL process</u>	9-14
5. <u>Analysis on dataset (for each analysis)</u>	15-25
i. Introduction	
ii. Analysis results	
iii. Visualization	
iv. Results	
6. <u>Bibliography</u>	26

INTRODUCTION

There were times when we were required to analyse large amounts of data and produce easy to read and understand reports. That used to be very hectic work at that times, but pivot tables have made our work easy to maximum extent. Pivot tables allow us to analyse such data and produce reports that meet our business reporting requirements.

I have worked on a project titled as “FIFA 19 player dataset” using pivot table and charts and has made various analysis from a very large data of around 13 Thousand players who were part of football in 2019. A pivot table is a data summarization tool that is used in context of data processing. Pivot table are used to summarize, sort, reorganize, group, count, total or average data stored in a database. It allows its users to transform columns into rows and rows into columns.

Moreover, pivot charts helps to represent the result of analysis in a very catchy manner no matter how large the data just like the FIFA dataset for year 2019 was too big to analyse. A pivot table helps you to extract the significance from a large, detailed data set. It provides n number of features like sorting, filtering etc. In the project analysis like nationality, football club, age, player ID, skills category wise and many more very almost ease with the help of pivot table and charts. Data Analysis is a process of inspecting, transforming and modelling data with goal of discovering useful information, information conclusions and supporting decision making. It has multiple assets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science and social science domains.

OBJECTIVES/SCOPE

Here are the objectives to be completed via this dashboard

- Player Profile
- Player Statistics
- Player Id
- Player Skills
- Player Gameplay
- Player IR rating

SOURCE OF DATASET

<https://www.kaggle.com/karangadiya/fifa19>

This is the source of my data set “FIFA 19”.it has n number of datasets of all fields, sizes, places, departments etc. It is an amazing source of datasets. While other sources of reference were also there like data. World, Wikipedia. **KAGGLE** is an online community of data scientist and machine learner, owned by [Google](#). Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

ETL PROCESS

ETL is defined as a **process** that extracts the data from different RDBMS source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system. **ETL** full-form is Extract, Transform and Load.

In computing, extract, transform, load is the general procedure of copying data from one or more sources into a destination system which represents the data differently from the source or in a different context than the source.

In this project also ETL Process is used using tableau prep for extraction like from csv to excel then for cleaning purpose like (removal of null values or negative values, redundancy, spelling error, grouping etc). then it transformed data is loaded to for further analysis. this is how etl process plays a very important role data analytics. This is how ten analysis are done on different basis.

1. Extraction:

The first step of the ETL process is extraction. In this step, data from various source systems is extracted which can be in various formats like relational databases, No SQL, XML and flat files into the staging area. It is important to extract the data from various source systems and store it into the staging area first and not directly into the data warehouse because the extracted data is in various formats and can be corrupted also. Hence loading it directly into the data warehouse may damage it and rollback will be much more difficult. Therefore, this is one of the most important steps of ETL process.

2. Transformation:

The second step of the ETL process is transformation. In this step, a set of rules or functions are applied on the extracted data to convert it into a single standard format. It may involve following processes/tasks:

- Filtering – loading only certain attributes into the data warehouse.
- Cleaning – filling up the NULL values with some default values, mapping U.S.A, United States and America into USA, etc.
- Joining – joining multiple attributes into one.
- Splitting – splitting a single attribute into multiple attributes.
- Sorting – sorting tuples on the basis of some attribute (generally key-attribute).

3. **Loading:**

The third and final step of the ETL process is loading. In this step, the transformed data is finally loaded into the data warehouse. Sometimes the data is updated by loading into the data warehouse very frequently and sometimes it is done after longer but regular intervals. The rate and period of loading solely depends on the requirements and varies from system to system.

For the FIFA 19 dataset ETL process works as follows:

Step1. Extraction

Data set when extracted from Kaggle which was in csv format.

- It had null values

- It had spelling mistakes

- It had unnecessary columns

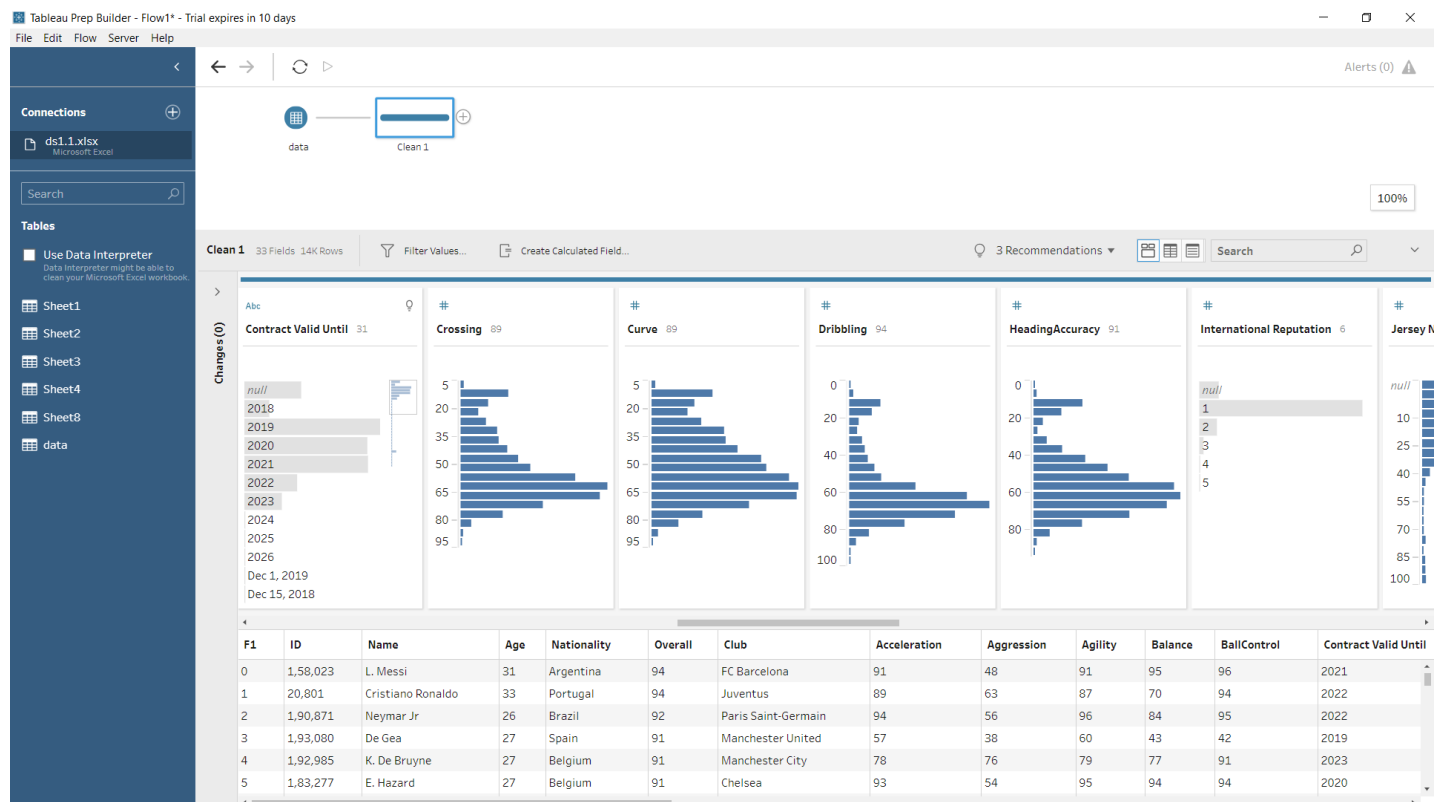
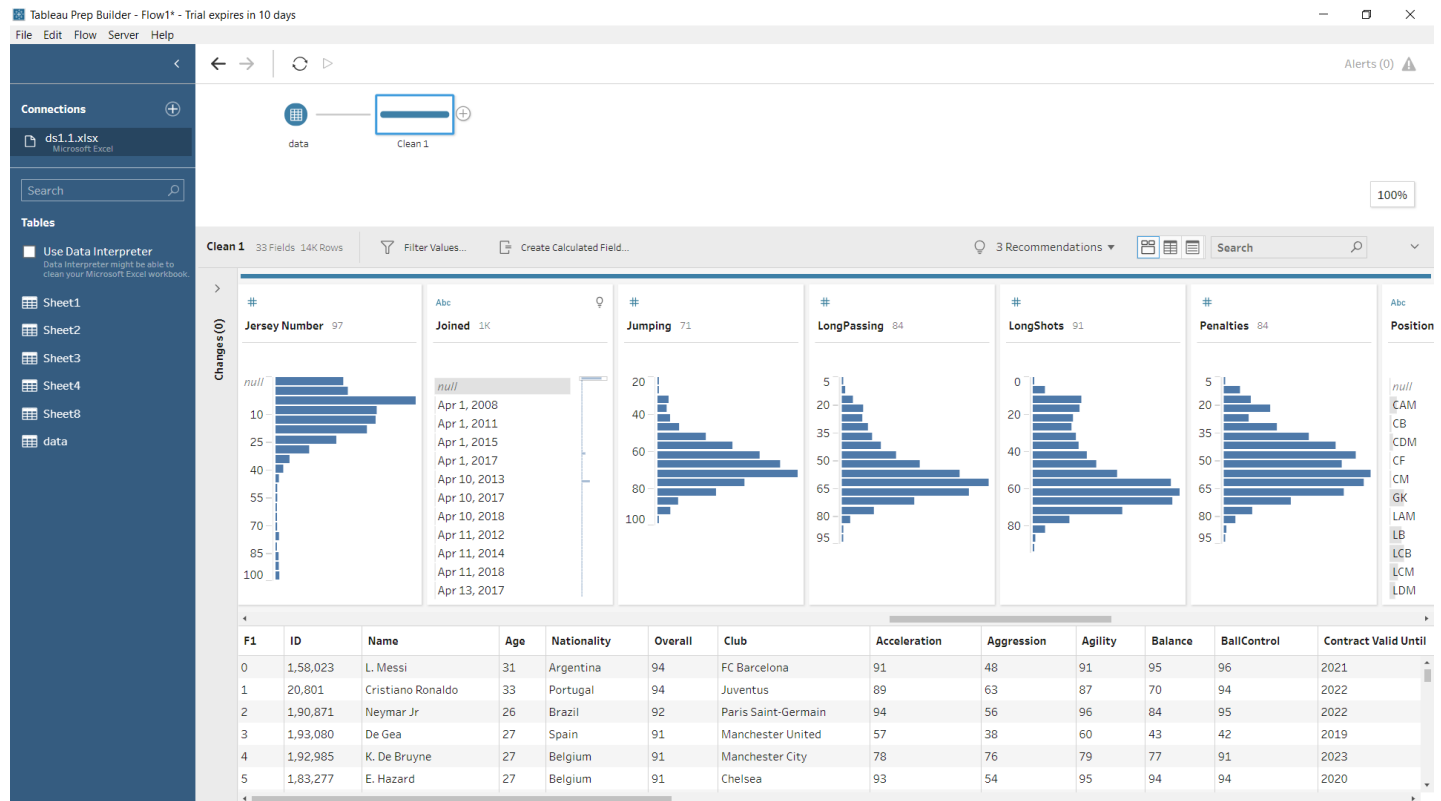
So first the file was extracted

- To make certain modifications

- To clean the data

- To analyse the results

This is how it looked before cleaning



Step 2. Transformation, cleaning has to be done after it is extracted from the source.

Here 3 cleaning is done

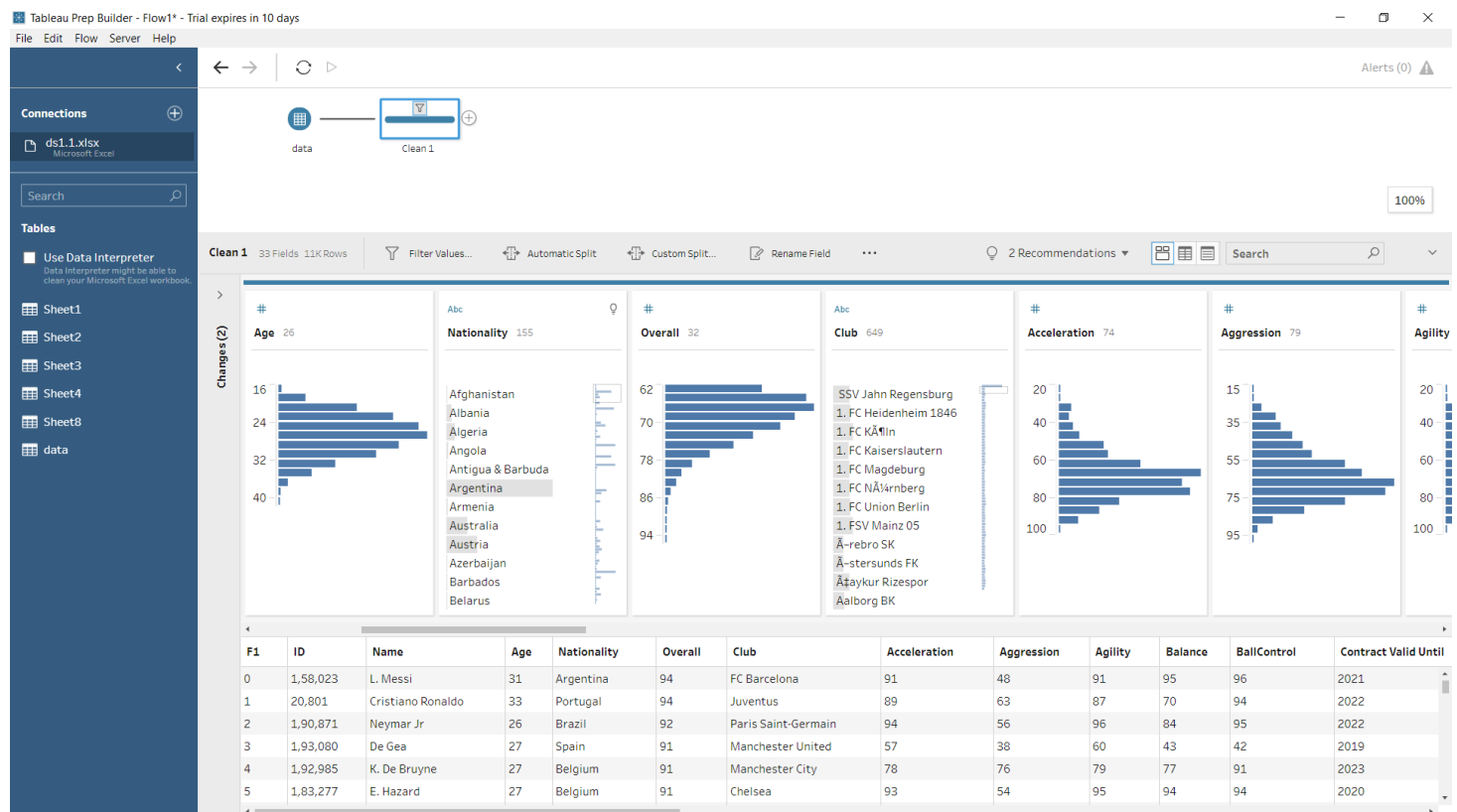
In first it removed unwanted columns

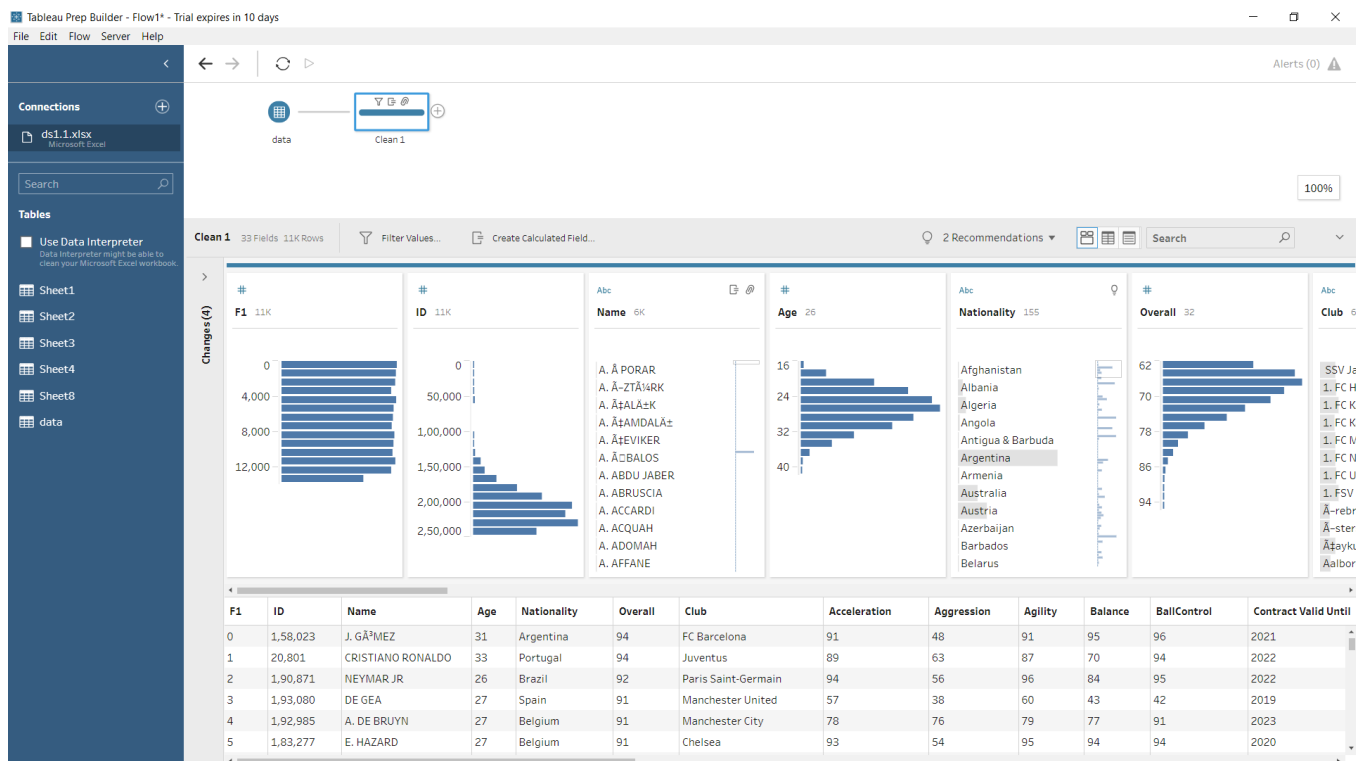
In second it removed null values and entries

In third it corrected spelling mistakes.

Removal of the unnecessary column (location) as it has no use in analysis. For this we just selected the entire column and pressed delete button from keyboard hence deleted the unwanted column from the table.

2. Removal of null values which is very important part of cleaning.





After all the cleaning and modifications are done dataset looks like:

AutoSave On

Yug pro Ds

Yug Ahuja

File Home Insert Page Layout Formulas Data Review View Help Data Streamer Inquire Power Pivot

Clipboard Cut Copy Paste Format Painter

Font Calibri 11 A

Alignment Merge & Center

Number %

Styles Conditional Formatting Format as Table Cell Styles

Cells Insert Delete Format

Editing AutoSum Fill Sort & Filter Find & Select

Ideas Sensitivity

Share Comments

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
	ID	Name	Age	Nationality	Overall	Club	Acceleration	Aggression	Agility	Balance	BallControl	Contract Valid Until	Crossing	Curve	Dribbling	HeadingAccuracy	International
1	158023	L. Messi	31	Argentina	94	FC Barcelona	91	48	91	95	96	2021	77	93	97	70	
2	20801	Cristiano Ronaldo	33	Portugal	94	Juventus	89	63	87	70	94	2022	84	81	88	89	
3	190871	Neymar Jr	26	Brazil	92	Paris Saint-Germain	94	56	96	84	95	2022	79	88	96	62	
4	193080	De Gea	27	Spain	91	Manchester United	57	38	60	43	42	2019	17	21	18	21	
5	192985	K. De Bruyne	27	Belgium	91	Manchester City	78	76	79	77	91	2023	93	85	86	55	
6	183277	E. Hazard	27	Belgium	91	Chelsea	93	54	95	94	94	2020	81	83	95	61	
7	177003	L. Modrić	32	Croatia	91	Real Madrid	80	62	93	94	93	2020	86	85	90	55	
8	176580	L. Suárez	31	Uruguay	91	FC Barcelona	86	87	82	83	90	2021	77	86	87	77	
9	155862	Sergio Ramos	32	Spain	91	Real Madrid	76	88	78	66	84	2020	66	74	63	91	
10	200389	J. Oblak	25	Slovenia	90	Atlético Madrid	43	34	67	49	16	2021	13	13	12	15	
11	192119	T. Courtois	26	Belgium	90	Real Madrid	46	23	61	45	23	2024	14	19	13	13	
12	188545	R. Lewandowski	29	Poland	90	FC Bayern München	77	80	78	78	89	2021	62	77	85	85	
13	182521	T. Kroos	28	Germany	90	Real Madrid	64	60	70	71	90	2022	88	86	81	54	
14	182493	D. Godín	32	Uruguay	90	Atlético Madrid	68	89	58	54	76	2019	55	49	53	92	
15	167495	M. Neuer	32	Germany	90	FC Bayern München	54	29	51	35	48	2021	15	14	30	25	
16	215914	N. Kanté	27	France	89	Chelsea	82	90	82	92	80	2021	68	49	79	54	
17	211110	P. Dybala	24	Argentina	89	Juventus	87	48	91	85	92	2022	82	88	92	68	
18	202126	H. Kane	24	England	89	Tottenham Hotspur	68	76	71	71	84	2024	75	78	80	85	
19	194765	A. Griezmann	27	France	89	Atlético Madrid	88	69	90	80	90	2023	82	84	88	84	
20	192448	M. ter Stegen	26	Germany	89	FC Barcelona	38	43	37	43	18	2022	15	18	17	11	
21	189511	Sergio Busquets	29	Spain	89	FC Barcelona	50	85	66	52	88	2023	62	66	80	68	
22	179813	E. Cavani	31	Uruguay	89	Paris Saint-Germain	75	84	77	59	82	2020	70	77	80	89	
23	178603	M. Hummels	29	Germany	89	FC Bayern München	62	69	63	60	81	2021	64	65	68	87	
24	168542	David Silva	32	Spain	89	Manchester City	69	57	92	90	94	2020	84	82	89	54	
25	153079	S. Agüero	30	Argentina	89	Manchester City	88	65	86	91	89	2021	70	82	89	74	
26	138956	G. Chiellini	33	Italy	89	Juventus	63	92	54	55	57	2020	58	60	58	84	
27	209331	M. Salah	26	Egypt	88	Liverpool	94	63	91	88	88	2023	78	83	89	59	
28	200145	Casemiro	26	Brazil	88	Real Madrid	59	87	62	66	78	2021	52	59	69	76	
29	198710	J. Rodríguez	26	Colombia	88	FC Bayern München	75	64	83	76	90	Jun 30, 2019	90	89	85	62	
30	198219	L. Insigne	27	Italy	88	Napoli	94	34	94	93	93	2022	86	87	90	56	

DASHBOARD Sheet3 Sheet4 Sheet5 Sheet6 Sheet11 data

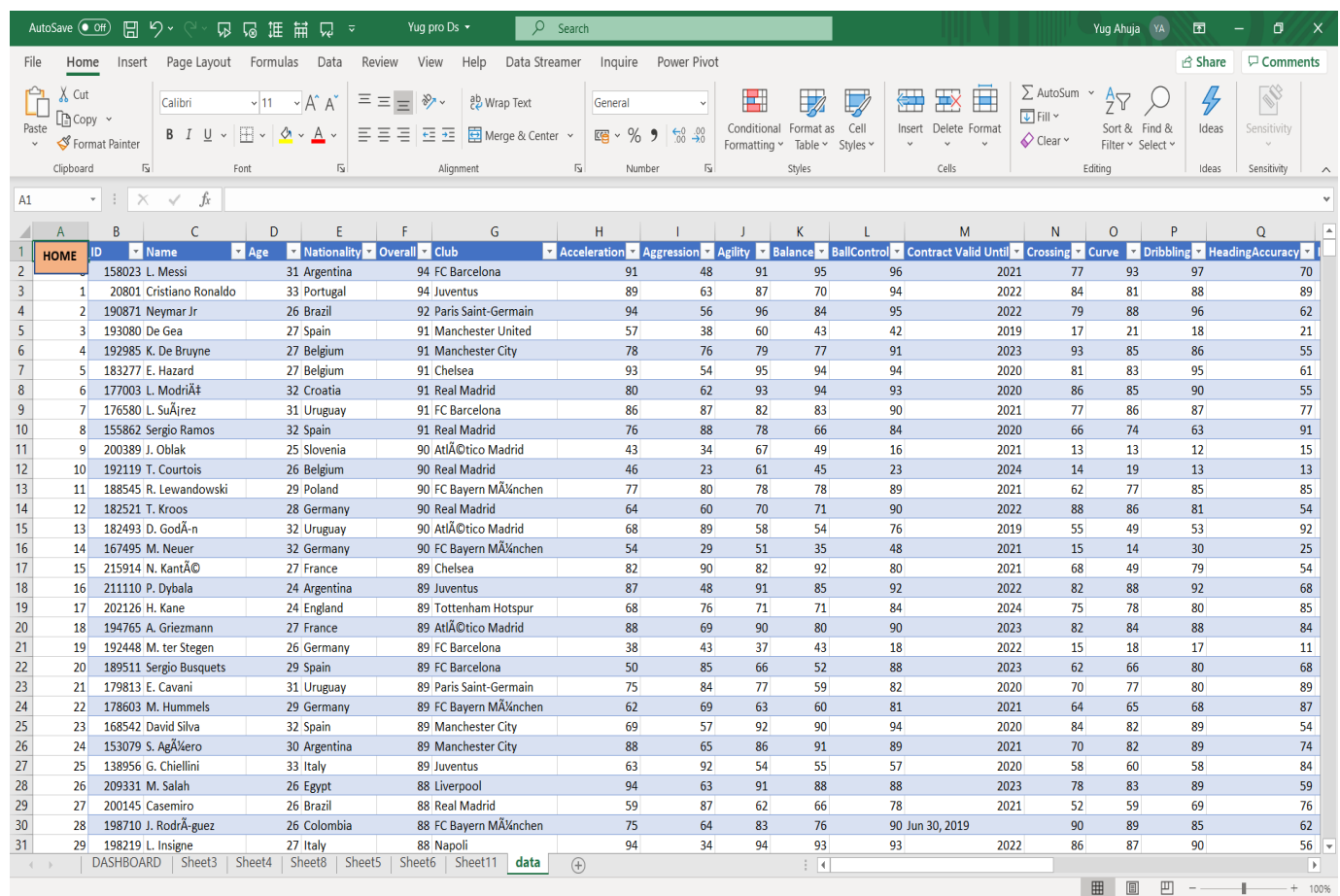
100%

Step 3. After transformation data is loaded and is worked further on analysis

Here further few analysis are done using pivot table and charts in most effective manner.

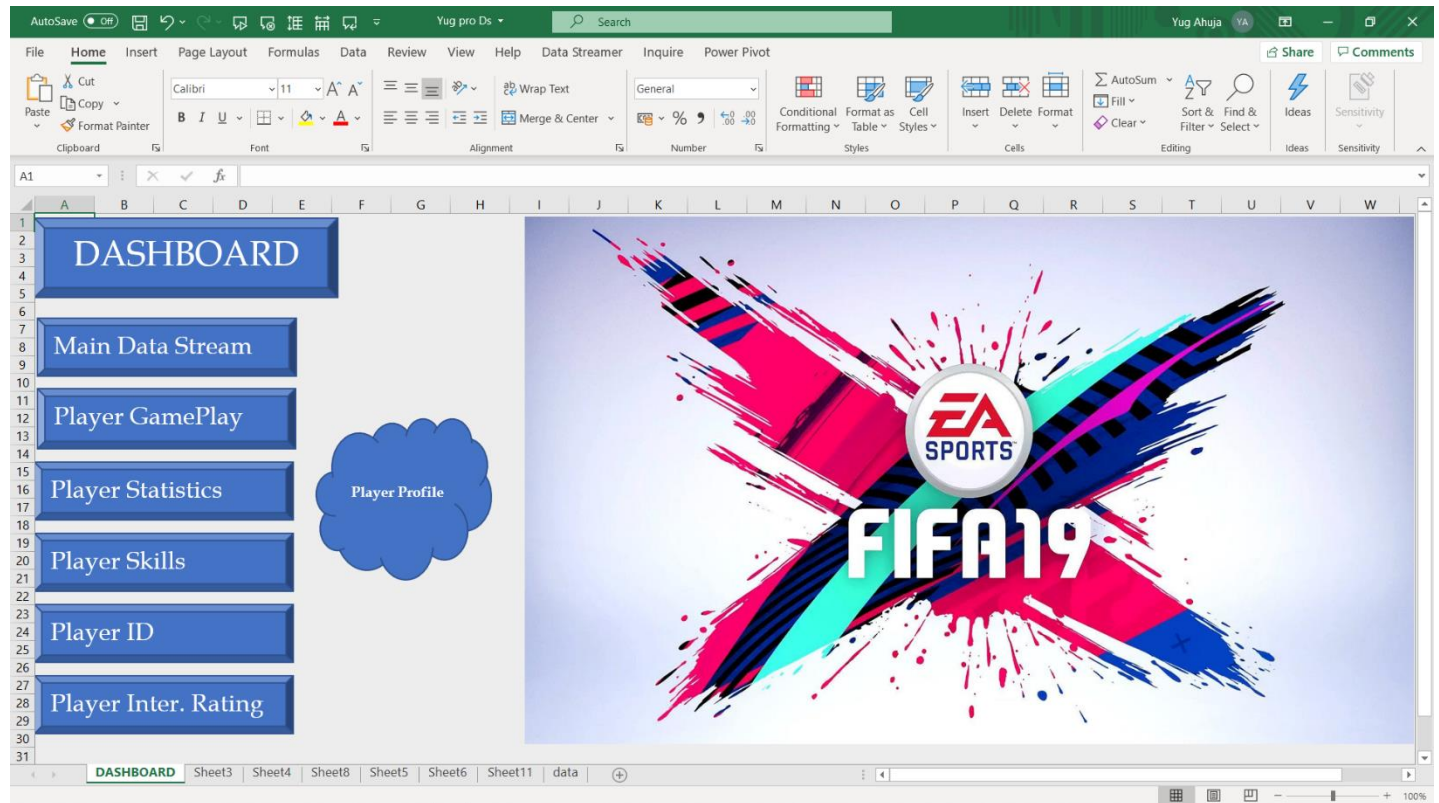
- Player Profile
- Player Statistics
- Player Id
- Player Skills
- Player Gameplay
- Player IR rating

Snapshot of final cleaned Main data stream.



ID	Name	Age	Nationality	Overall	Club	Acceleration	Aggression	Agility	Balance	BallControl	Contract Valid Until	Crossing	Curve	Dribbling	HeadingAccuracy
158023	L. Messi	31	Argentina	94	FC Barcelona	91	48	91	95	96	2021	77	93	97	70
20801	Cristiano Ronaldo	33	Portugal	94	Juventus	89	63	87	70	94	2022	84	81	88	89
190871	Neymar Jr	26	Brazil	92	Paris Saint-Germain	94	56	96	84	95	2022	79	88	96	62
193080	De Gea	27	Spain	91	Manchester United	57	38	60	43	42	2019	17	21	18	21
192985	K. De Bruyne	27	Belgium	91	Manchester City	78	76	79	77	91	2023	93	85	86	55
183277	E. Hazard	27	Belgium	91	Chelsea	93	54	95	94	94	2020	81	83	95	61
177003	L. Modrić	32	Croatia	91	Real Madrid	80	62	93	94	93	2020	86	85	90	55
176580	L. Suárez	31	Uruguay	91	FC Barcelona	86	87	82	83	90	2021	77	86	87	77
155862	Sergio Ramos	32	Spain	91	Real Madrid	76	88	78	66	84	2020	66	74	63	91
200389	J. Oblak	25	Slovenia	90	Atlético Madrid	43	34	67	49	16	2021	13	13	12	15
192119	T. Courtois	26	Belgium	90	Real Madrid	46	23	61	45	23	2024	14	19	13	13
188545	R. Lewandowski	29	Poland	90	FC Bayern München	77	80	78	78	89	2021	62	77	85	85
182521	T. Kroos	28	Germany	90	Real Madrid	64	60	70	71	90	2022	88	86	81	54
182493	D. Godín	32	Uruguay	90	Atlético Madrid	68	89	58	51	76	2019	55	49	53	92
167495	M. Neuer	32	Germany	90	FC Bayern München	54	29	51	35	48	2021	15	14	30	25
215914	N. Kanté	27	France	89	Chelsea	82	90	82	92	80	2021	68	49	79	54
211110	P. Dybala	24	Argentina	89	Juventus	87	48	91	85	92	2022	82	88	92	68
202126	H. Kane	24	England	89	Tottenham Hotspur	68	76	71	71	84	2024	75	78	80	85
194765	A. Griezmann	27	France	89	Atlético Madrid	88	69	90	80	90	2023	82	84	88	84
192448	M. ter Stegen	26	Germany	89	FC Barcelona	38	43	37	43	18	2022	15	18	17	11
189511	Sergio Busquets	29	Spain	89	FC Barcelona	50	85	66	52	88	2023	62	66	80	68
179813	E. Cavani	31	Uruguay	89	Paris Saint-Germain	75	84	77	59	82	2020	70	77	80	89
178603	M. Hummels	29	Germany	89	FC Bayern München	62	69	63	60	81	2021	64	65	68	87
168542	David Silva	32	Spain	89	Manchester City	69	57	92	90	94	2020	84	82	89	54
153079	S. Agüero	30	Argentina	89	Manchester City	88	65	86	91	89	2021	70	82	89	74
138956	G. Chiellini	33	Italy	89	Juventus	63	92	54	55	57	2020	58	60	58	84
209331	M. Salah	26	Egypt	88	Liverpool	94	63	91	88	88	2023	78	83	89	59
200145	Casemiro	26	Brazil	88	Real Madrid	59	87	62	66	78	2021	52	59	69	76
198710	J. Rodríguez	26	Colombia	88	FC Bayern München	75	64	83	76	90 Jun 30, 2019		90	89	85	62
198219	I. Insigne	27	Italy	88	Napoli	94	34	94	93	93	2022	86	87	90	56

Snapshot of DASHBOARD



1. PLAYER PROFILE

Introduction

Under this sheet information about the nationality, football club, age, jersey number and joining date of the particular players are analysed using pivot and slicers.

Specific Requirements, functions and formulas

No such specific requirements, functions or formulae is used as in dataset.

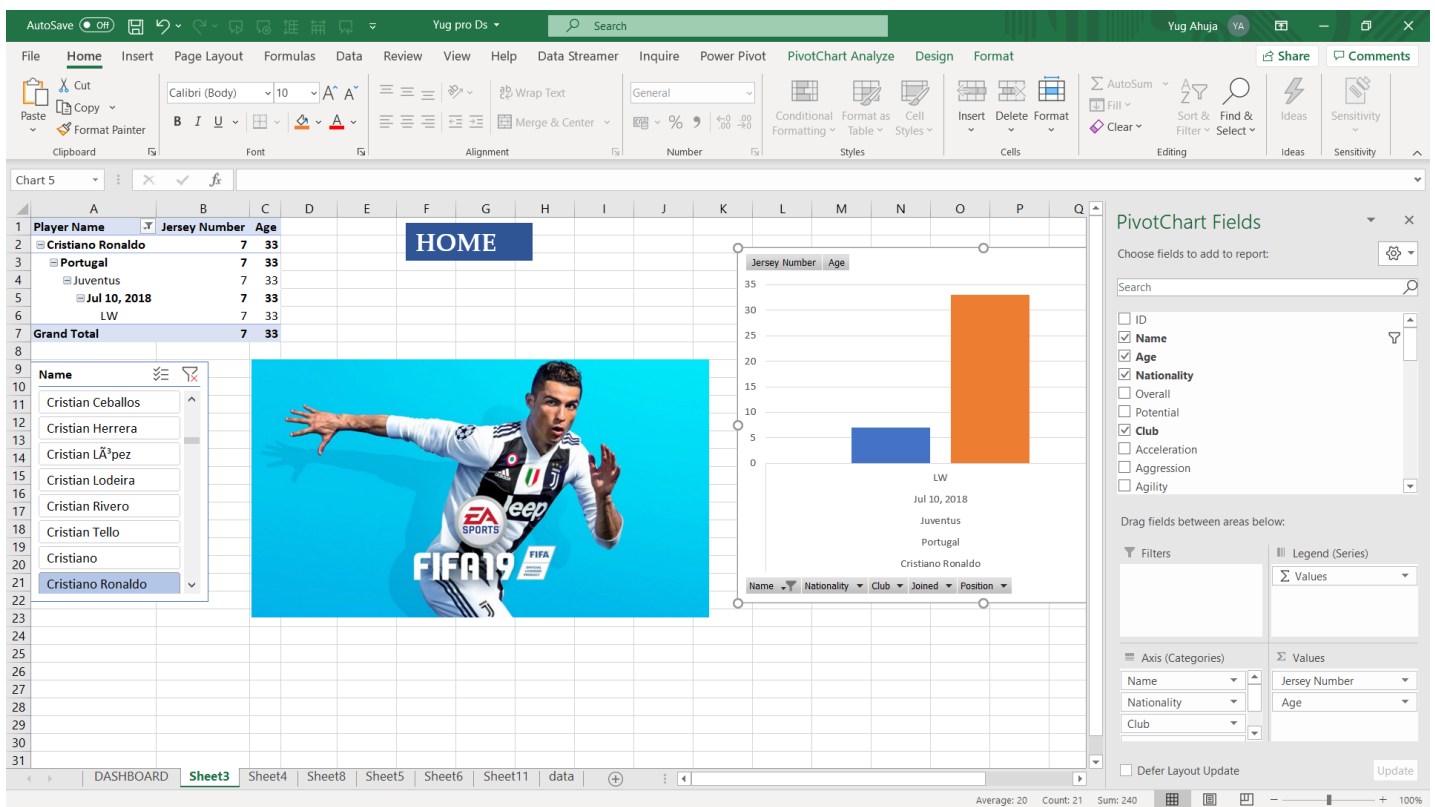
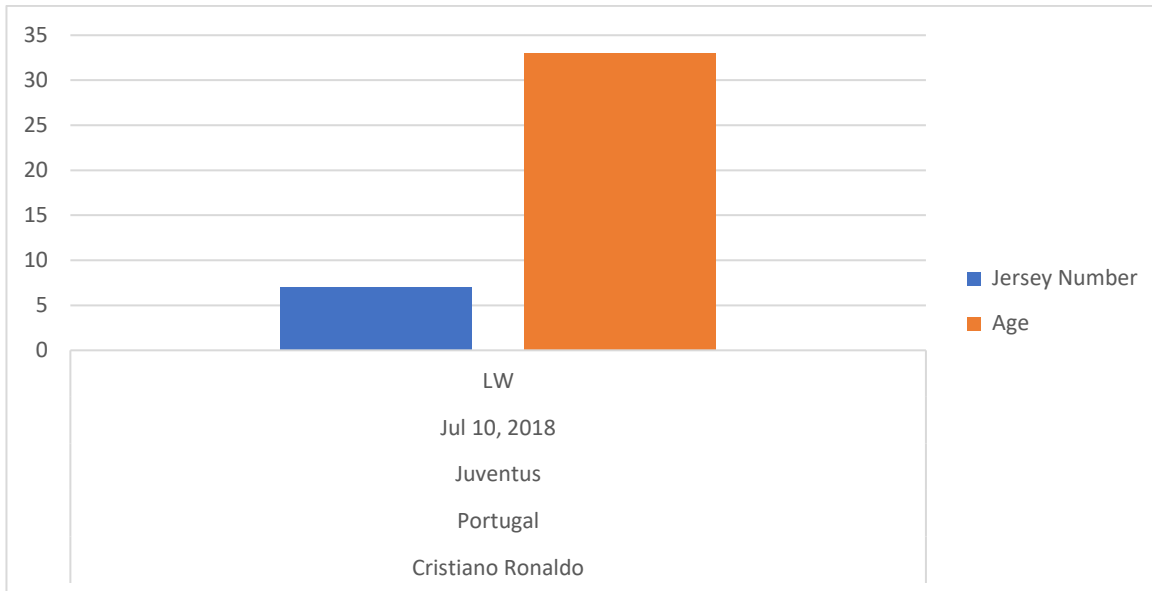
Analysis results

Player Name	Jersey Number	Age
Cristiano Ronaldo	7	33
Portugal	7	33
Juventus	7	33
Jul 10, 2018	7	33
LW	7	33
Grand Total	7	33



Pivoting is used to analyse the data and using slicers we are able to locate information for each player.

Visualization



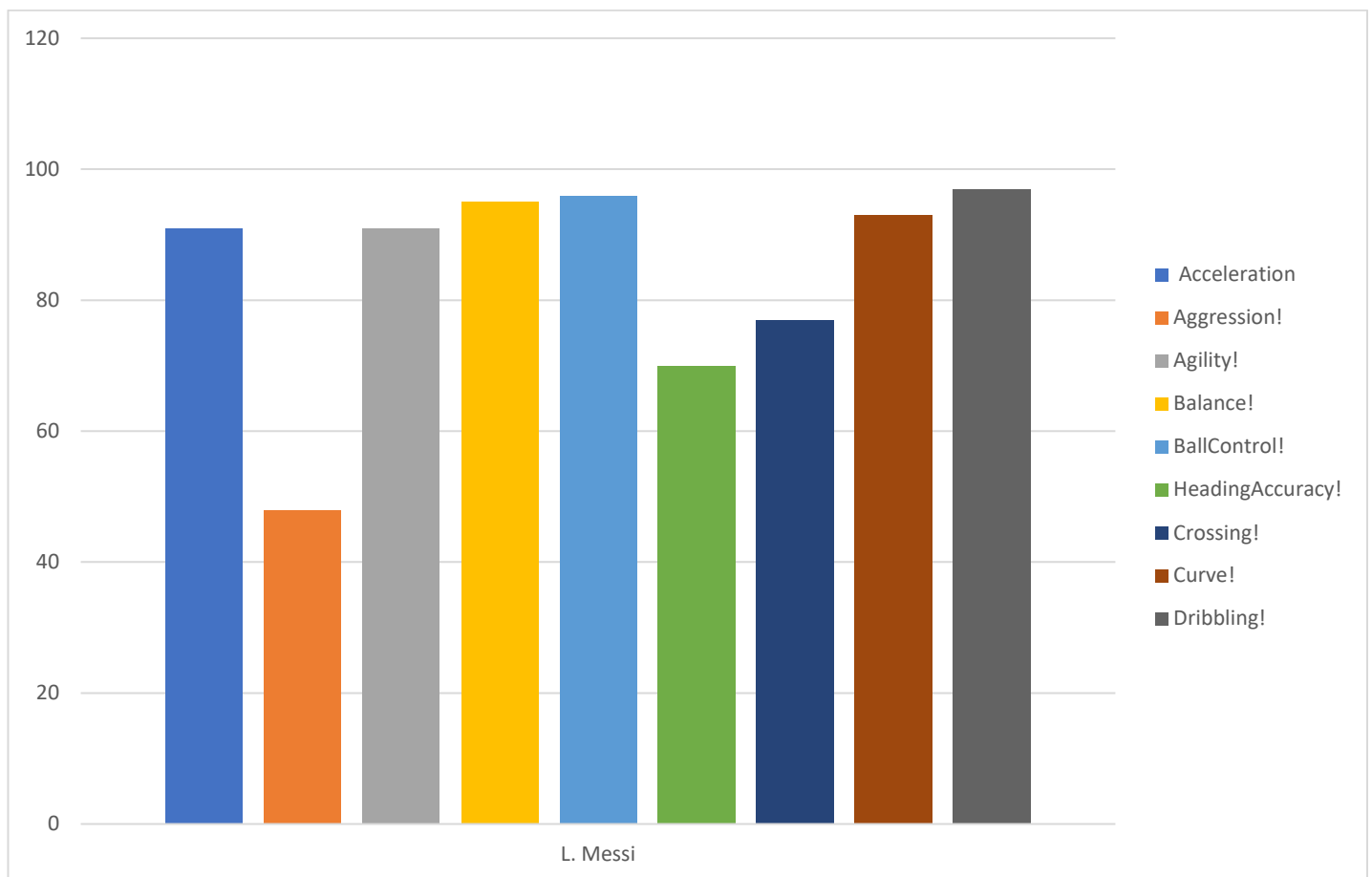
2. PLAYER STATISTICS

Introduction

Under this all the players were analysed according to their performance in the field.

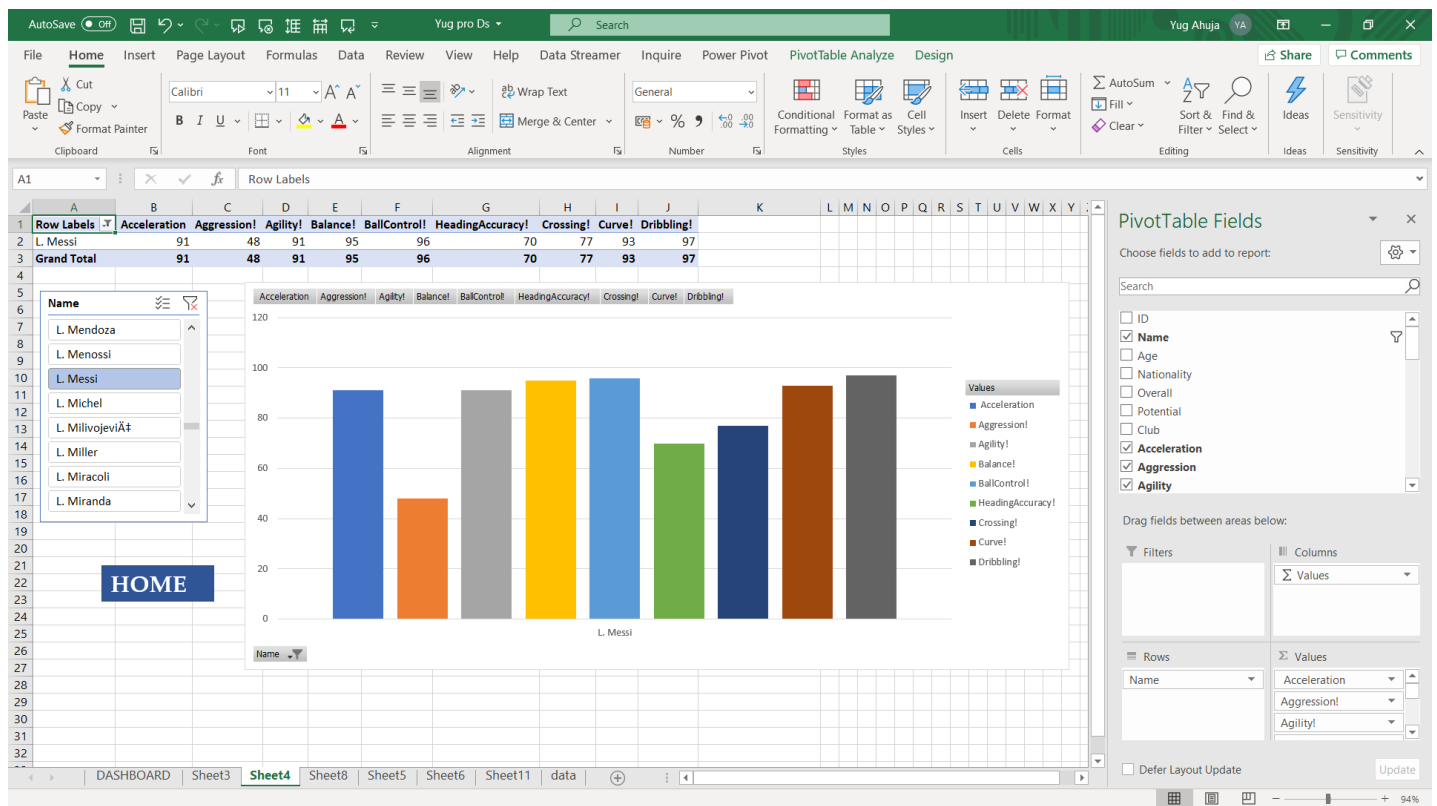
Eg:- Acceleration, Aggression, Agility, Balance, Ball Control, Heading Accuracy, Crossing, Curve, Dribbling.

Analysis results



The performance of the players was analysed with ease. In above chart analysis of player Messi is performed using slicer and pivot charts.

Visualization

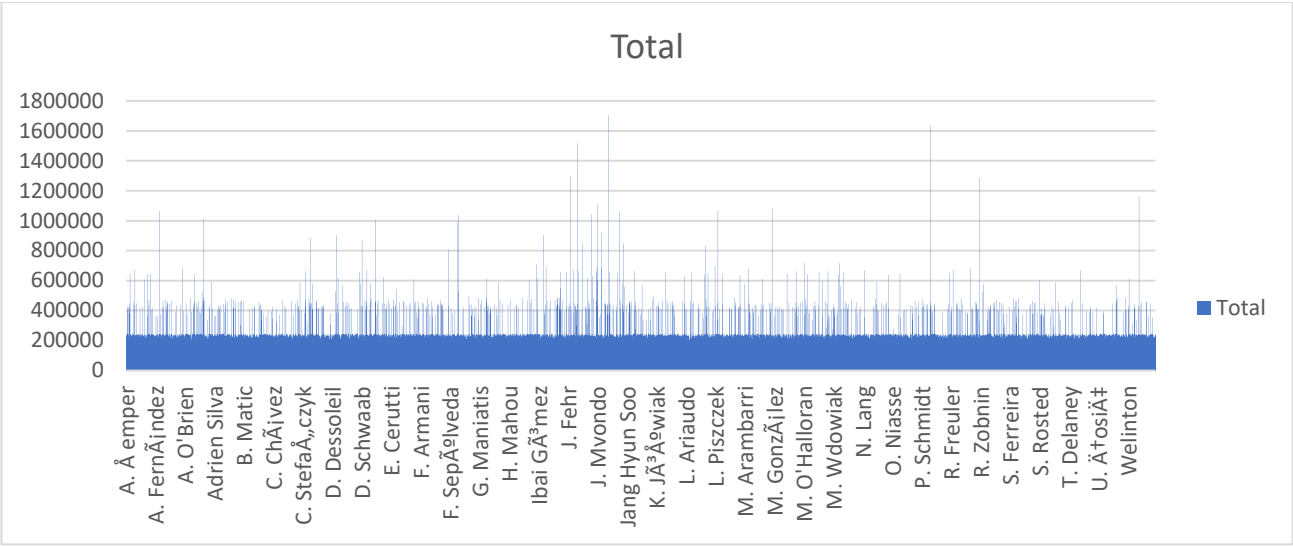


3. PLAYER ID

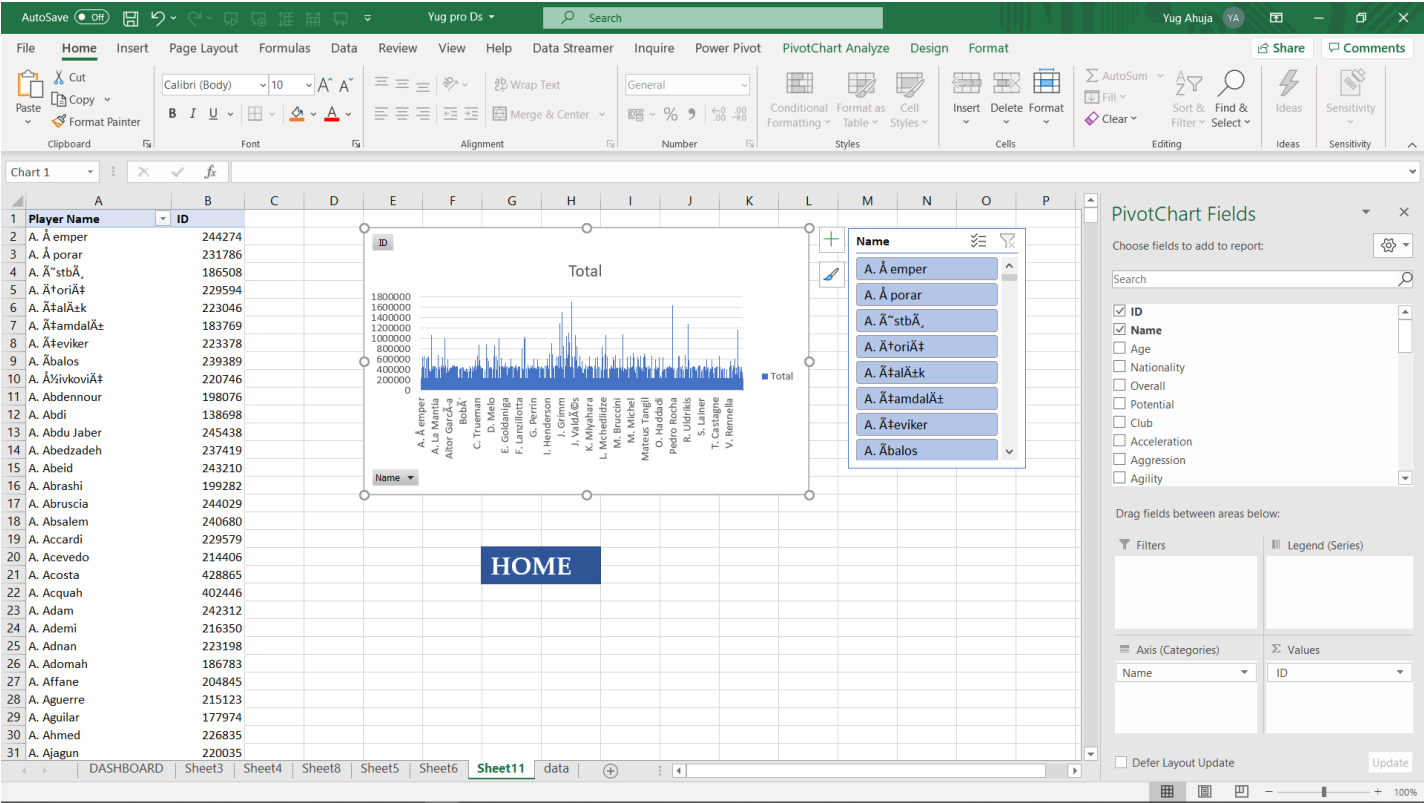
Introduction

Under this analysis each player was provided with unique player ID in order to differentiate between the rest of the players.

Analysis results



Visualization



4. PLAYER SKILLS

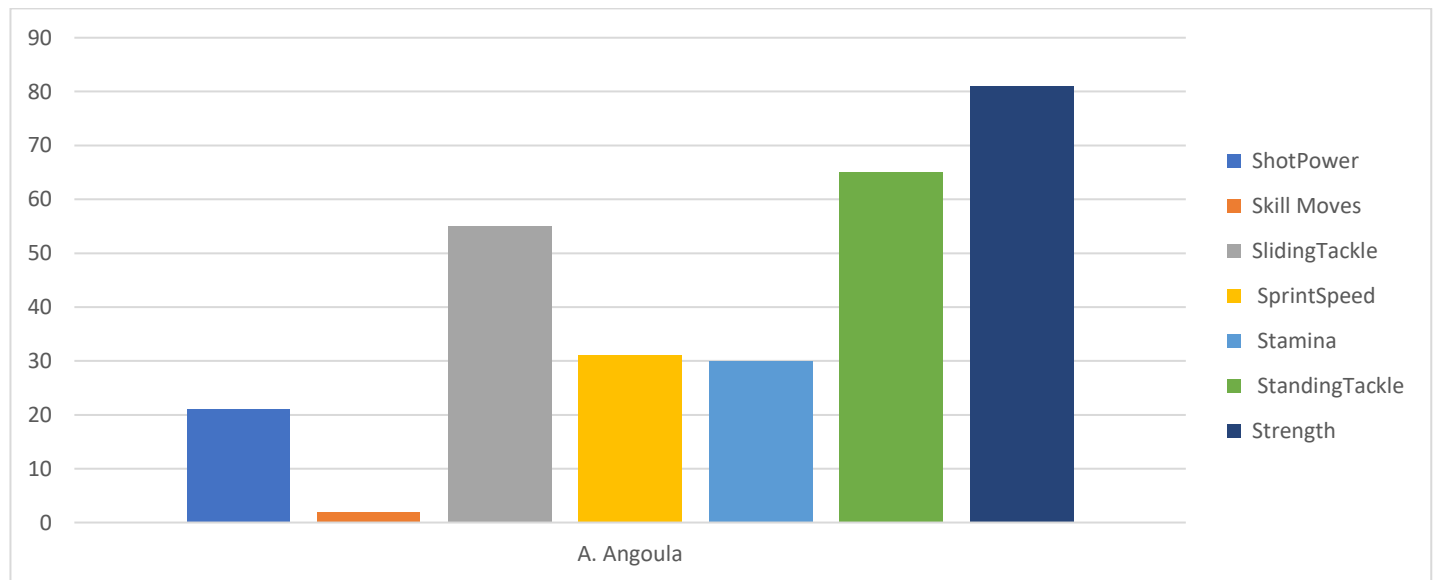
Introduction

The players were analysed on the basis of their skill level.

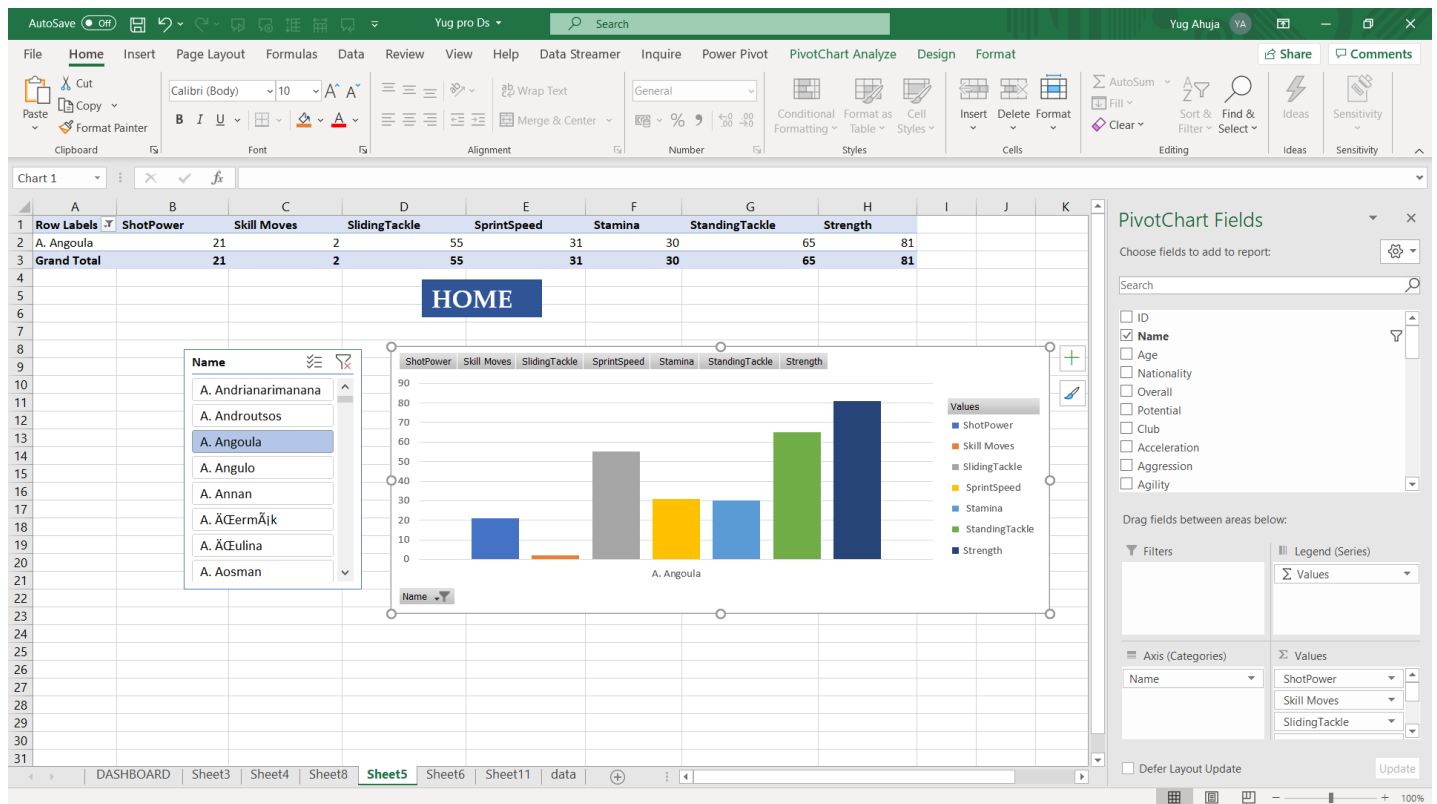
Analysis results

The players were analysed on the basis of the following factors:

- Shot power
- Skill moves
- Sliding tackle
- Sprint speed
- Stamina
- Standing tackle
- Strength



Visualization



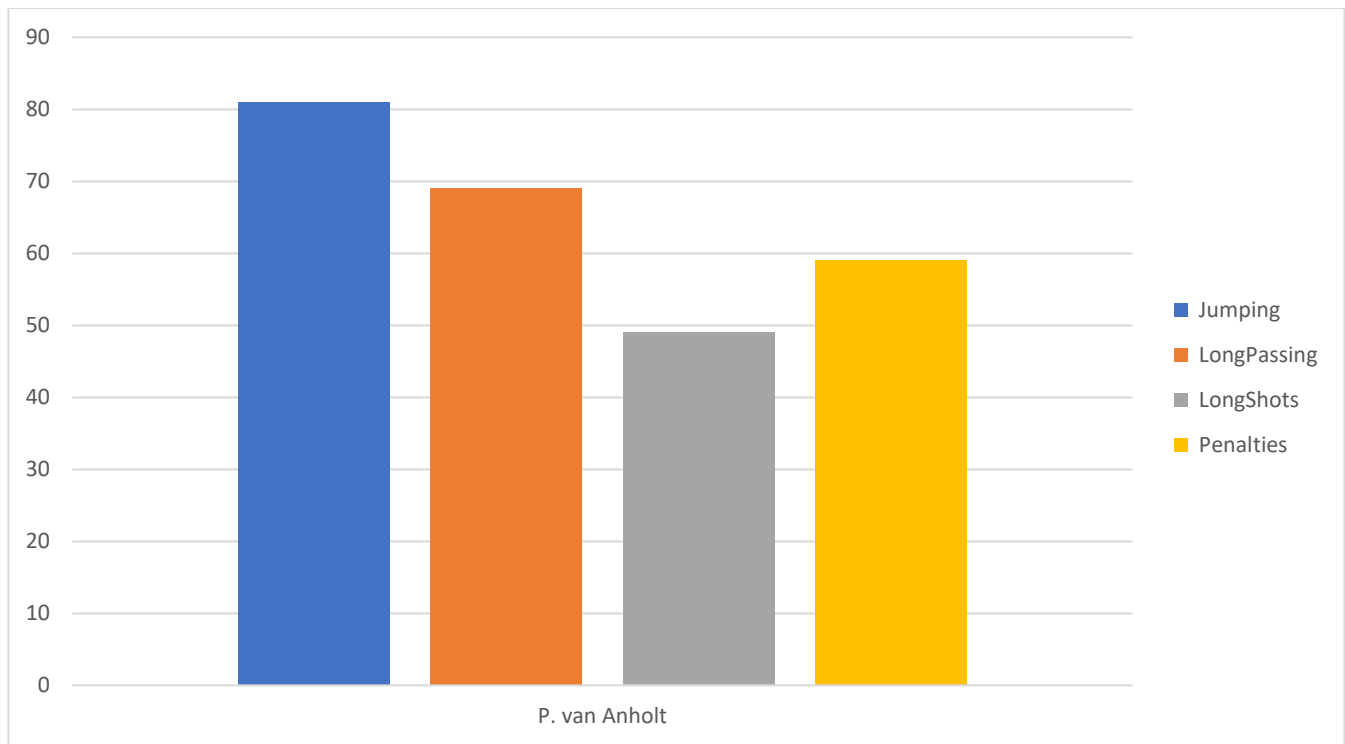
5. PLAYER GAMEPLAY

Introduction

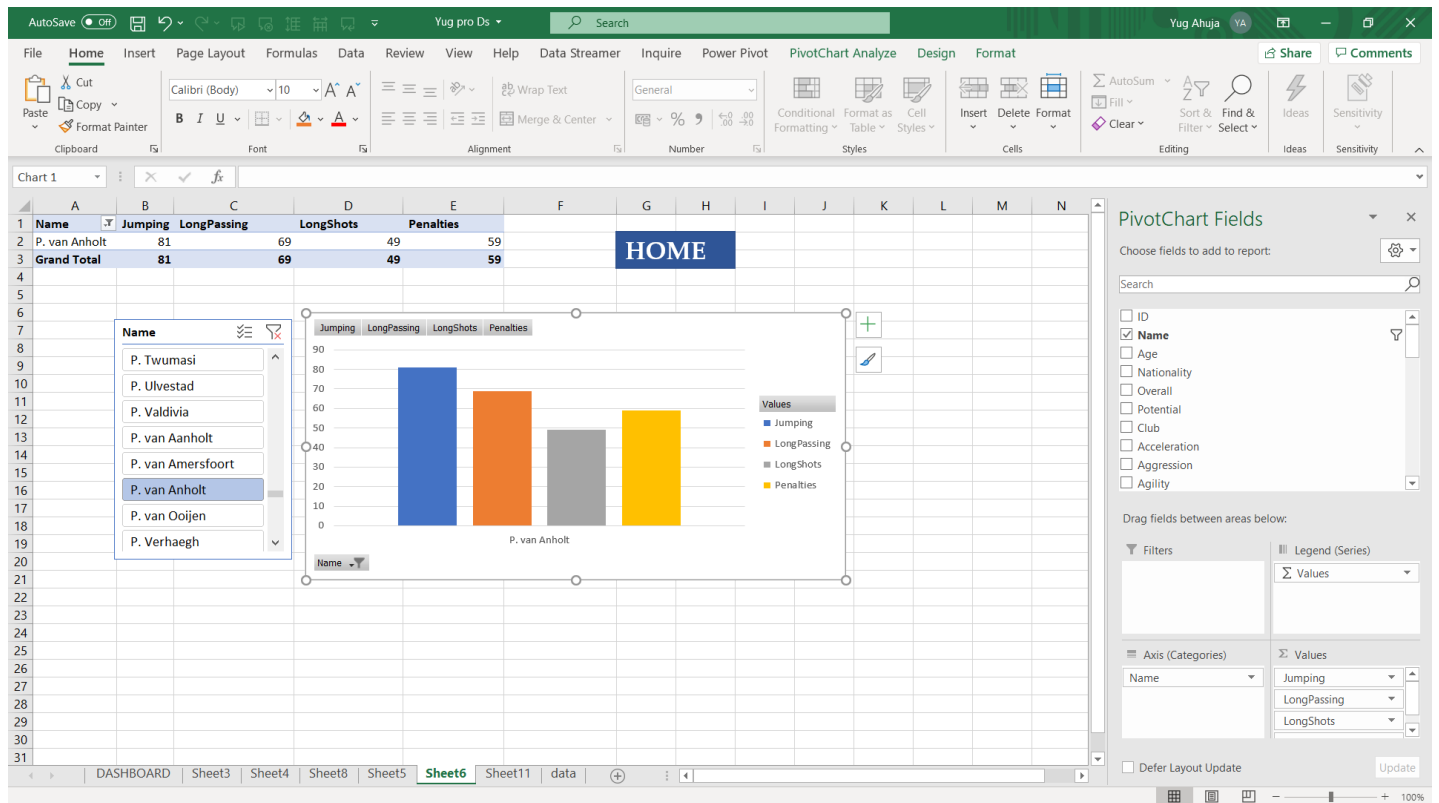
Under this the entire gameplay of the players was analysed based on the following factors:

- Jumping
- Long passing
- Penalties
- Long shots

Analysis results



Visualization



6. PLAYER IR RATING

Introduction

All the players in the data set were rated internationally on the scale of 1-5 on this basis of their performance and personality and their social life.

Analysis results

Row Labels	International Reputation!
L. Messi	5
Grand Total	5

Name

L. Melano

L. Melgarejo

L. Memela

L. Memushaj

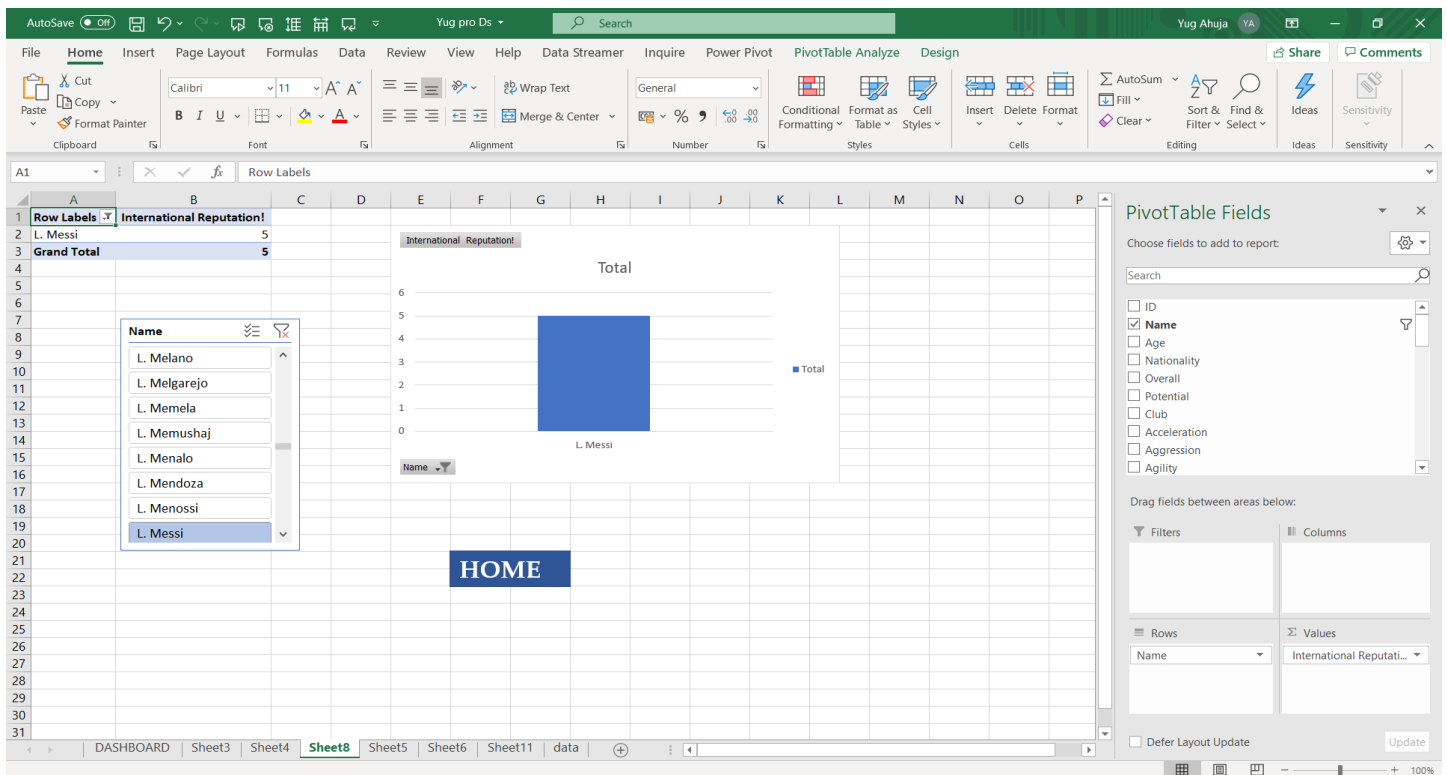
L. Menalo

L. Mendoza

L. Menossi

L. Messi

Visualization



BENEFITS OF PIVOT

A **pivot table** is a data summarization tool that is **used** in the context of data processing. **Pivot tables** are **used** to summarize, sort, reorganize, group, count, total or average data stored in a database. It allows its users to transform columns into rows and rows into columns. a **Pivot Table** is a tool built into Excel that allows you to summarize large quantities of data quickly and easily. Given an input **table** with tens, hundreds, or even thousands of rows, **Pivot Tables** allow you to extract answers to a series of basic questions about your data with minimal effort. Use a **pivot table** to build a list of unique values. Because **pivot tables** summarize data, they can be **used** to find unique values in a field. This is a good way to quickly see all the values that appear in a field and also find typos, and other inconsistencies. pivot table is a tool in data summation that is common in many business software. It is utilized to collect the summary of a specific data set in a compressed technique. It is a very useful tool in consolidating a large quantity of data that is contained in Microsoft Excel. They let the user make a faster organization and drawing of conclusions from data being collected. Pivot table consists of columns, rows, pages, and data fields. These can be moved around and it assists in expanding, isolating, summarizing, and grouping the specific data. And all of these can be accomplished in real time.

Pivot Table Slicers are a visual filter in the form of an interactive button. There are several cool things that you can do with **Pivot Table Slicers**, like customize them, filter them, connect them to multiple **Pivot Tables** plus much more!

It can create instant data

It makes data analyses easier

It summarises data easily

It helps to give quick overview

It is very fast and easy to work with.

LEARNING OUTCOMES

I have learnt a lot from this dashboard project using excel. Outcomes are as follows:

1. Learnt to use pivot technique
2. Learnt to analyse data efficiently which is the main task of data scientists.
3. Pivot charts are very helpful in data representation
4. Also, learnt to manage large datasets.
5. Learnt to make dynamic dashboards.
6. Slicers use make task so much easier.
7. Create, modify, and format PivotTables

BIBLIOGRAPHY

www.kaggle.com

www.data.world.com

www.wikipedia.com