UNIVERSITÀ degli STUDI di CATANIA

EUROPE TOP 5 FOOTBALL LEAGUES

BIG DATA ANALYSIS

MUHAMMAD HASSAN ALI | 1000023619
SYED MUHAMMAD ZAFFAR | 1000023608

Prof. Orazio Tomarchio

Table of Contents

Introduction and Business Questions

This report analyzes stats about the matches that played throughout the five big Leagues of Europe (La Liga, English Premier League, BundesLiga, Serie A and Ligue 1) by the end of each season from 2014 to 2018. In the Business Questions chapter, we asked questions related to the matches that played among all the European leagues mentioned above. First of all, we wanted to understand the statistical summary, about the matches played by all the league's teams. Then, we came up with the following questions:

a. Which team won the most league titles in Europe's top 5 leagues from 2014 to 2018?
b. Which team won league title with the highest points?
c. Which team has scored most goals in Europe top 5 leagues since 2014?
d. Which team has spent most on transfers since 2014?
e. Which team has earned most from transfers since 2014?
f. Which team wins most away and home matches?
g. Which team conceded less yellow and red cards?
h. Which team conceded most yellow and red cards?
i. Which team conceded most fouls since 2014?

In the Dataset chapter we analyze the structure of our datasets and in ETL chapter we discuss how data was cleaned, treated and manipulated in order to obtain the final dataset used for the dashboard creation. The Dimensional Fact Model chapter explains how our final dataset was conceptually structured. The Dashboard chapter discusses how our dashboards were made, while the Answers chapter explains how our dashboards answer our original questions. In the Conclusion chapter we make final considerations about our work and how it could be improved in the future.

Dataset

The report is based on data relating to football matches among Europe's top five leagues from 2014 to 2018. From the data source https://www.kaggle.com/ogrofratesi/football-data.

Description

The dataset contains statistical summary data by the end of each season from 2014 to 2018 for the five big Leagues of Europe:

- La Liga
- English Premier League
- BundesLiga
- Serie A
- Ligue 1

The dataset is organized in 3 csv files:

Football_Data.csv (484 obs): Contains data about each game.

Transfer_team.csv (935 obs): Contains data with the information about transfers of each team.

stats_per_game.csv (1000 obs): Contains information for every team match.

The original tables are shown below;

Football_Data.csv

Field_Name	Type	Description
League	String	League's names
Year	Date	Year from 2014 to 2018
position	Integer	Team's ranks in respective league by year
Team	String	Team's name
matches	Integer	No. of matches held by each league
wins	Integer	No. of matches wins by each team in respective league
draws	Integer	No. of matches draws by each team in respective league
loses	Integer	No. of matches loses by each team in respective league
scored	Integer	Goals scored
missed	Integer	Goals missed
pts	Integer	Over all points
xG	Decimal	expected goals metric, it is a statistical measure of the quality of chances created and conceded

xG_diff	Decimal	difference between actual goals scored and expected goals
npxG	Decimal	expected goals without penalties and own goals
xGA	Decimal	expected goals against
xGA_diff	Decimal	difference between actual goals missed and expected goals against
npxGA	Decimal	expected goals against without penalties and own goals
npxGD	Decimal	difference between "for" and "against" expected goals without penalties and own goals
ppda_coef	Decimal	passes allowed per defensive action in the opposition half (power of pressure)
oppda_coef	Decimal	opponent passes allowed per defensive action in the opposition half (power of opponent's pressure)
deep	Decimal	passes completed within an estimated 20 yards of goal (crosses excluded)
deep_allowed	Decimal	opponent passes completed within an estimated 20 yards of goal (crosses excluded)
xpts	Decimal	expected points
xpts_diff	Decimal	difference between actual and expected points
0VS0	Decimal	Matches were the final result was 0-0
H_Shots	Integer	Total shots in home
A_Shots	Integer	Total shots in away
HS_OnTarget	Integer	Total shots on target(Home)
AS_OnTarget	Integer	Total shots on target(Away)
HYellow	Integer	Yellow cards(Home)
AYellow	Integer	Yellow cards(Away)
HRed	Integer	Red cards(Home)
ARed	Integer	Red cards(Away)
HCorners	Integer	Total Corners (Home)

ACorners	Integer	Total Corners (Away)
HFouls	Integer	Fouls committed (Home)
AFouls	Integer	Fouls committed (Away)
H_Rival_Shots	Integer	Shots against (Home)
A_Rival_Shots	Integer	Shots against (Away)
H_RS_OnTarget	Integer	Shots on target against (Home)
A_RS_OnTarget	Integer	Shots on target against (Away)
H_Rival_Fouls	Integer	Fouls against (Home)
A_Rival_Fouls	Integer	Fouls against (Away)
H%Ov_Ov	Decimal	The percentage for all the matches were the team won both halfs at home
A%Ov_Ov	Decimal	The percentage for all the matches were the team won both halfs at away
H%LoseR	Decimal	% of all the matches were the team loses the first half, the percentage of those that end up winning (Home)
A%LoseR	Decimal	Of all the matches were the team loses the first half, the percentage of those that end up winning (Away)
H%DrawR	Decimal	Of all the matches were the team draw the first half, the percentage of those that end up winning (home)
A%DrawR	Decimal	Of all the matches were the team draw the first half, the percentage of those that end up winning (Away)
Sec	Integer	How many matches in a row does the team had before the first Draw
Draw_Last9	Integer	How many matches does the team Draw in the last 9 games of the season
Transfer_E	Decimal	Team Transfer Expediture
Transfer_I	Decimal	Team Transfer Income

%Ov_Ov	Decimal	The percentage for all the matches were the team won both halfs
%LoseR	Decimal	% of all the matches were the team loses the first half, the percentage of those that end up winning
%DrawR	Decimal	% of all the matches were the team draw the first half, the percentage of those that end up winning
Corners	Integer	Over all corners(Home & Away)
Shots	Integer	Over all shots(Home & Away)
Yellow	Integer	Over all yellow cards(Home & Away)
Red	Integer	Over all red cards(Home & Away)
Fouls	Integer	Over all fouls committed(Home & Away)
S_OnTarget	Integer	Over all shots on target (Home & Away)

Stat_per _game

Field_name	Type	Description
league	String	Seasons name
year	Date	Duration(2014-2018)
h_a	Char	Home or Away
xG	Decimal	expected goals metric, it is a statistical measure of the quality of chances created and conceded
xGA	Decimal	expected goals against
npxG	Decimal	expected goals without penalties and own goals
npxGA	Decimal	expected goals against without penalties and own goals
deep	Integer	passes completed within an estimated 20 yards of goal (crosses excluded)

deep_allowed	Integer	opponent passes completed within an estimated 20 yards of goal (crosses excluded)
scored	Integer	Goals scored
missed	Integer	Goals missed
xpts	Decimal	expected points
result	String	Match result (win-w,draw-d or lose-l)
date		
wins	Integer	No. of matches wins
draws	Integer	No. of matches draws
loses	Integer	No. of matches loses
pts	Integer	No. of points
npxGD	Decimal	difference between "for" and "against" expected goals without penalties and own goals
ppda_coef	Decimal	passes allowed per defensive action in the opposition half (power of pressure)
oppda_coef	Decimal	opponent passes allowed per defensive action in the opposition half (power of opponent's pressure)
team	String	Team's names
xG_diff	Decimal	expected goals without penalties and own goals
xGA_diff	Decimal	difference between actual goals missed and expected goals against
xpts_diff	Decimal	difference between actual and expected points

Transfer_team.csv

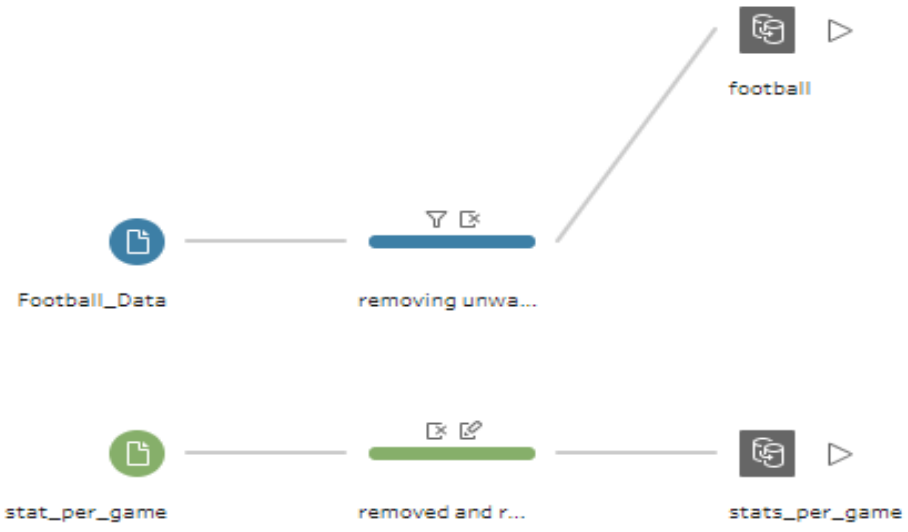
Field_name	**Type**	**Description**
ID	Integer	Player's ID
Name	String	Player's name
Position	Integer	Player's position in field while playing
Age	Integer	Player's age
Team_from	String	Player belonging team before transfer

League_from	String	Player belonging league before transfer
Team_to	String	Player belonging team after transfer
League_to	String	Player belonging league before transfer
Season	Integer	Season year(2014-2018)
Market_value	Integer	Player's market value before transfer
Transfer	Integer	Player's market value after transfer

ETL

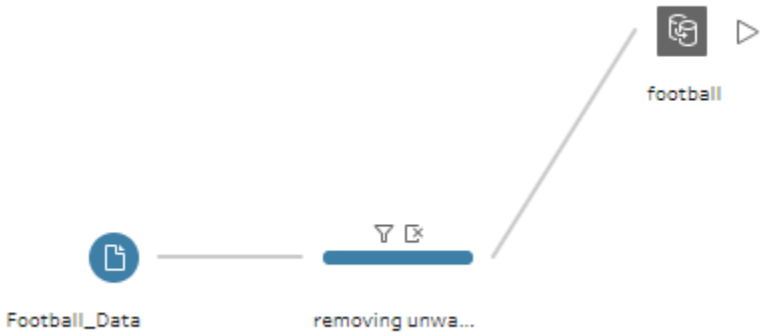
Introduction and Overall ETL Structure

As our data set contained three different tables and all the tables are already well organized, therefore we just go through the cleaning stage by removing unnecessary fields, null values and renaming some fields where required in order to finalize our dataset while performing ETL phase. We performed the ETL phase with Tableau Prep.



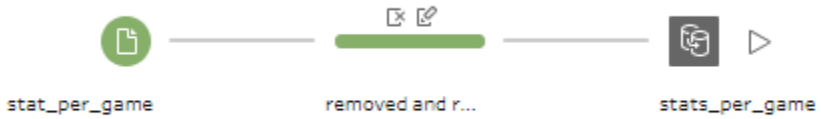
Cleaning

Football_Data



Here we cleaning our data set table by removing un-necessary fields and null values which were present in our dataset.

Stat_per_game



Here we also removed un-necessary fields and null values from our data set while performing cleaning processed and renamed some field as well in our data set where required.

Transfer_team.csv

As table transfer_team was already well organized and cleaned, therefore here cleaning was not required.

Dimensional Fact Model

The dimensional fact model (DFM) is an ad hoc and graphical formalism specifically devised to support the conceptual modeling phase in a DW project. DFM is extremely intuitive and can be used by analysts and non-technical users as well. Thus, the DFM is a graphical conceptual model, specifically devised for multidimensional design, in order to:

• lend effective support to conceptual design;

• create an environment in which user queries may be formulated intuitively;

• make communication possible between designers and end users with the goal of formalizing requirement specifications;

• build a stable platform for logical design;

• provide clear and expressive design documentation. The conceptual representation generated by the DFM consists of a set of fact schemata. Fact schemata model facts, measures, dimensions, and hierarchies. Besides these basic elements, the DFM includes a large set of constructs for expressing the multitude of conceptual nuances that characterize actual modeling scenarios in projects of small to large complexity (Figure 1)

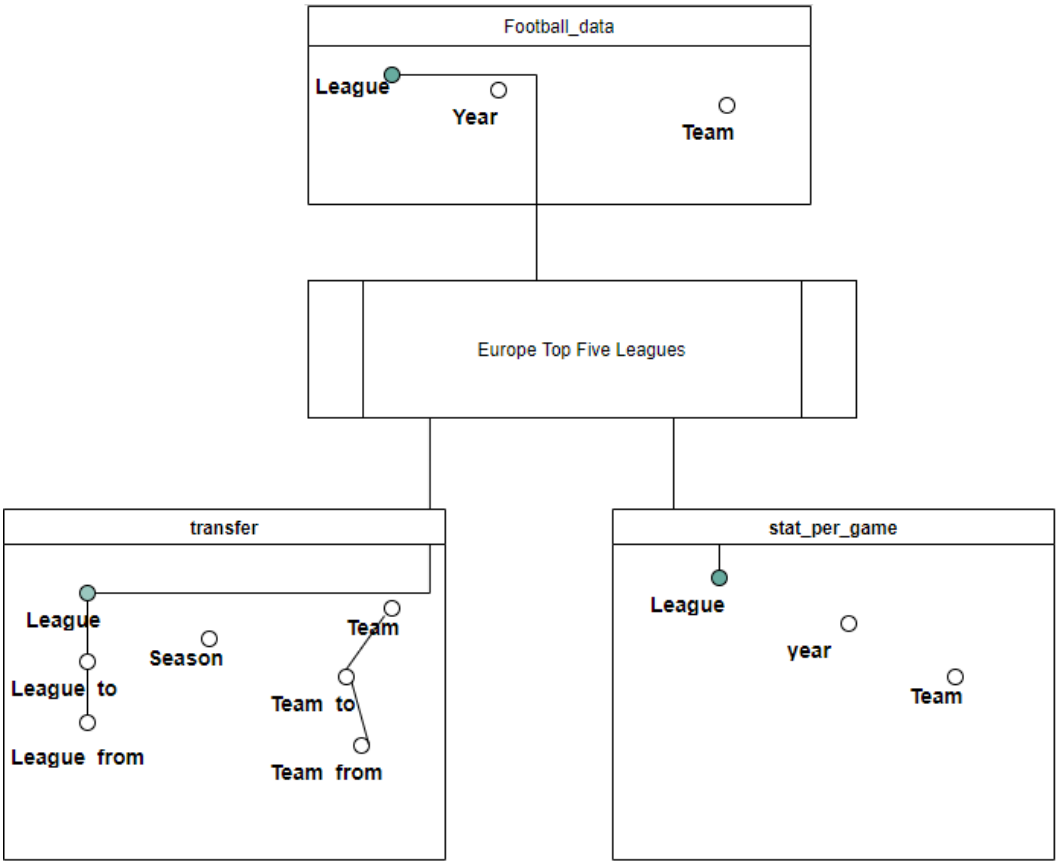
The following attributes have been defined for each hierarchy:

● Season → year

● Team → Team_to →Team_from

● League → League_to → League_from

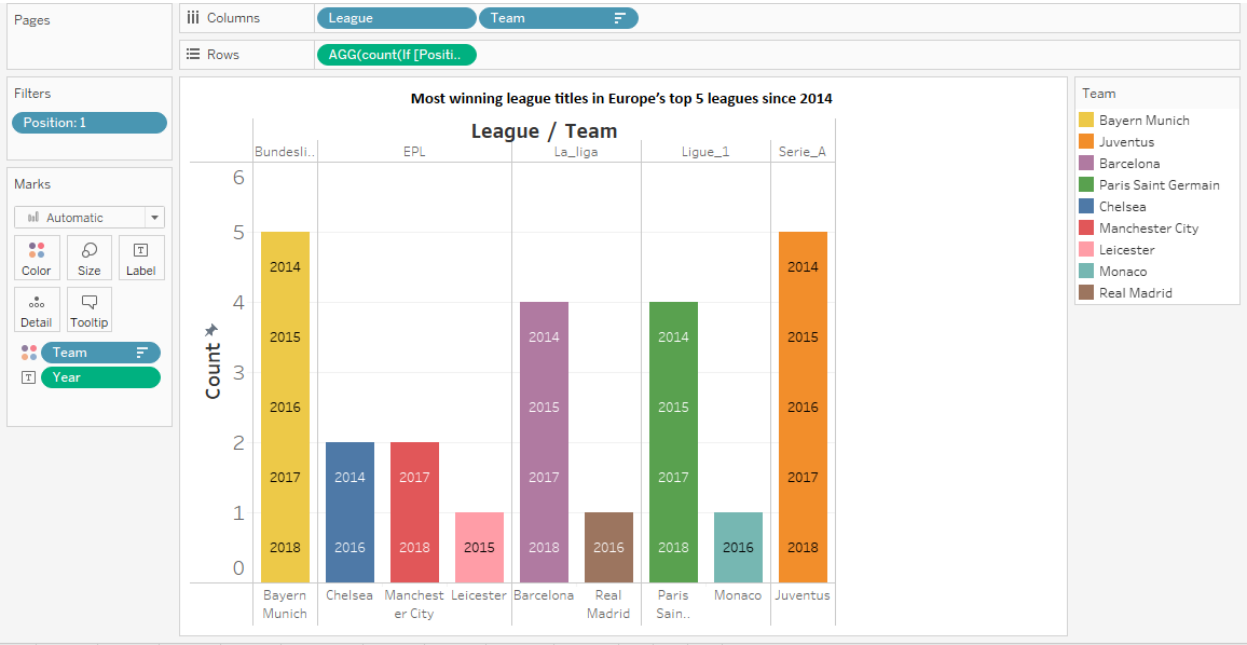
fig-1



Dashboard

In order to answer the proposed business questions, different dashboards were built for each query, and each containing different information according to the business questions that we proposed.

a.

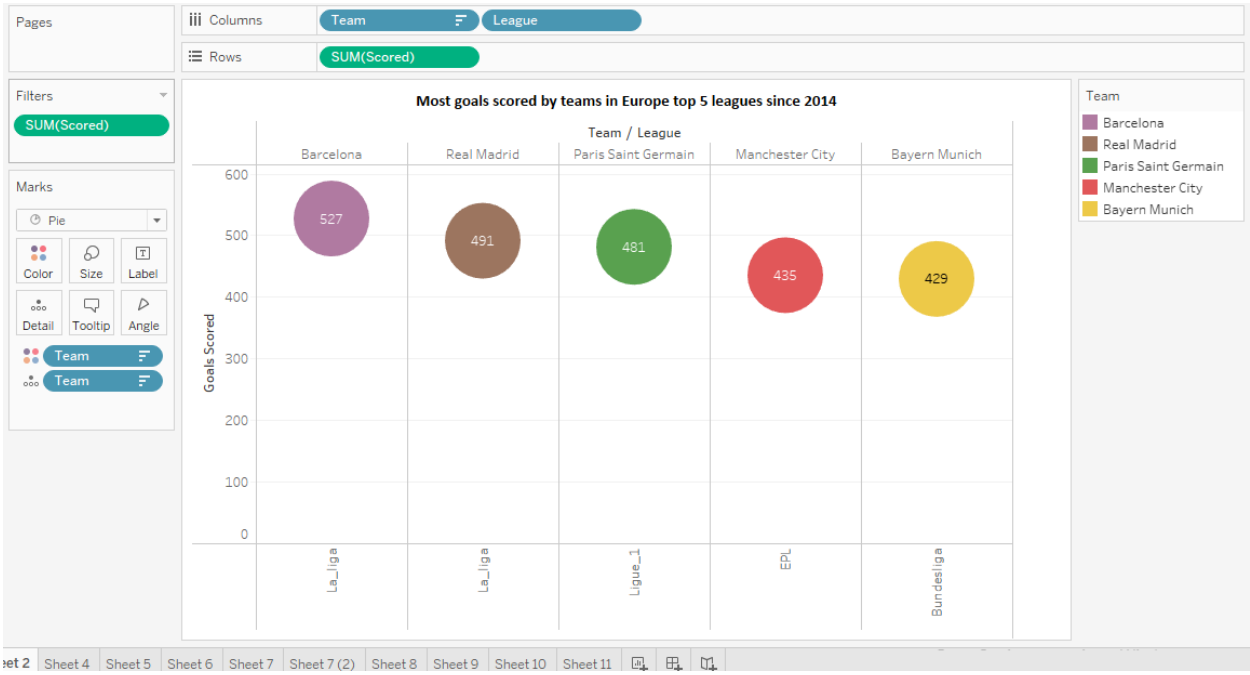


In the above dashboard we can see that the team Bayern Munich (German league_Bundesliga's team) and juventus (Italian league_Serie_A's team) both teams won the most titles among all the teams in the Europe top five leagues since 2014.

b.

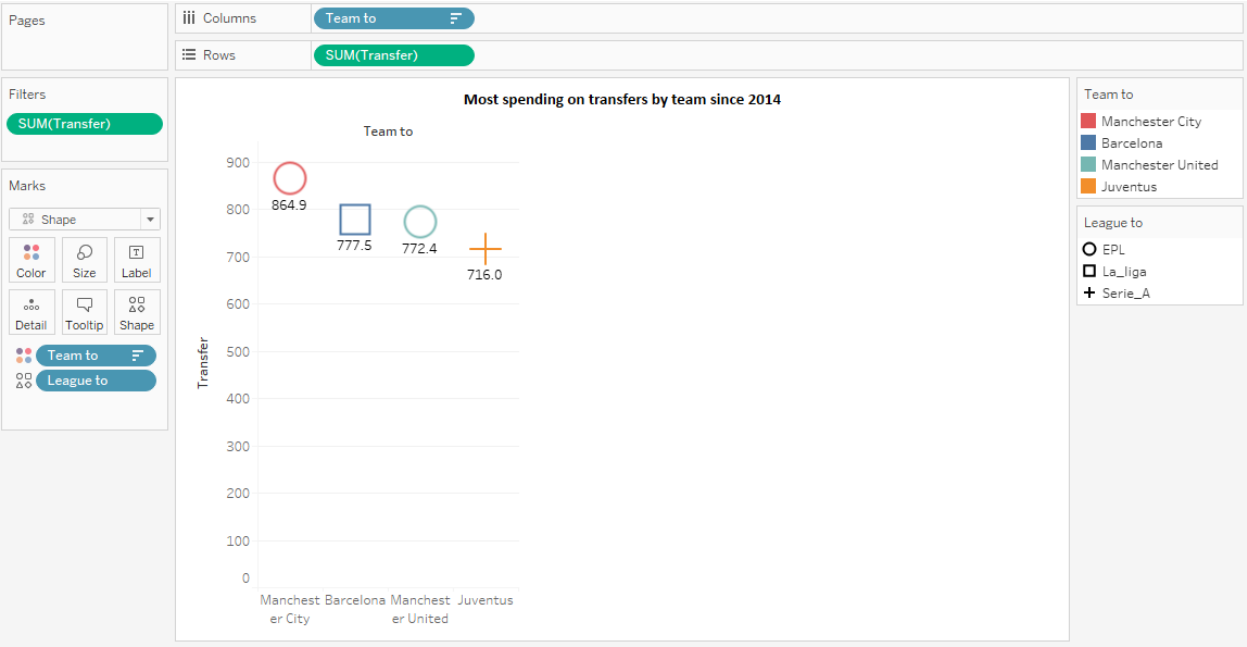
Team Manchester United won the league titles with highest points among all the Europe top five league's teams since 2014.

c.



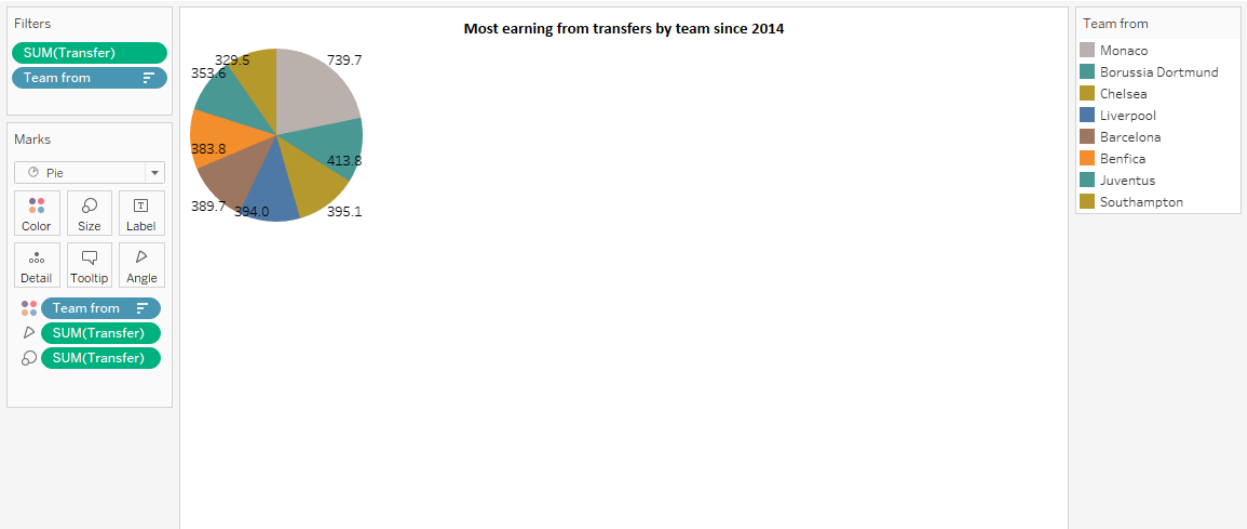
Barcelona scored most goals during the seasons from 2014 to 2018 among all the top five Europe league's teams.

d.



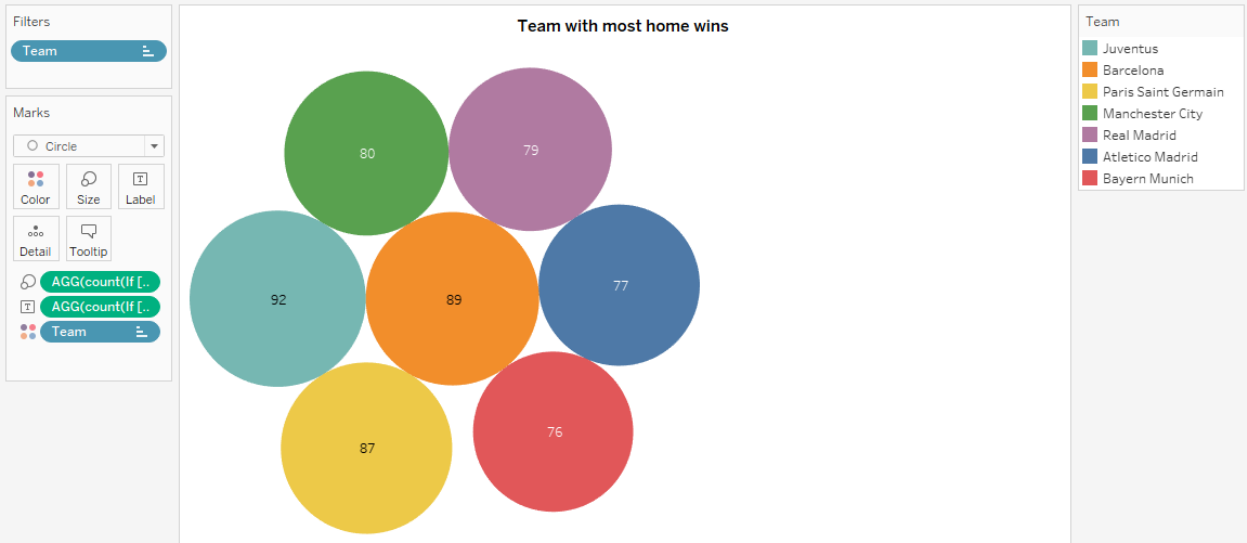
The team Manchester City spending most on transfer throughout all the seasons since 2014.

e.



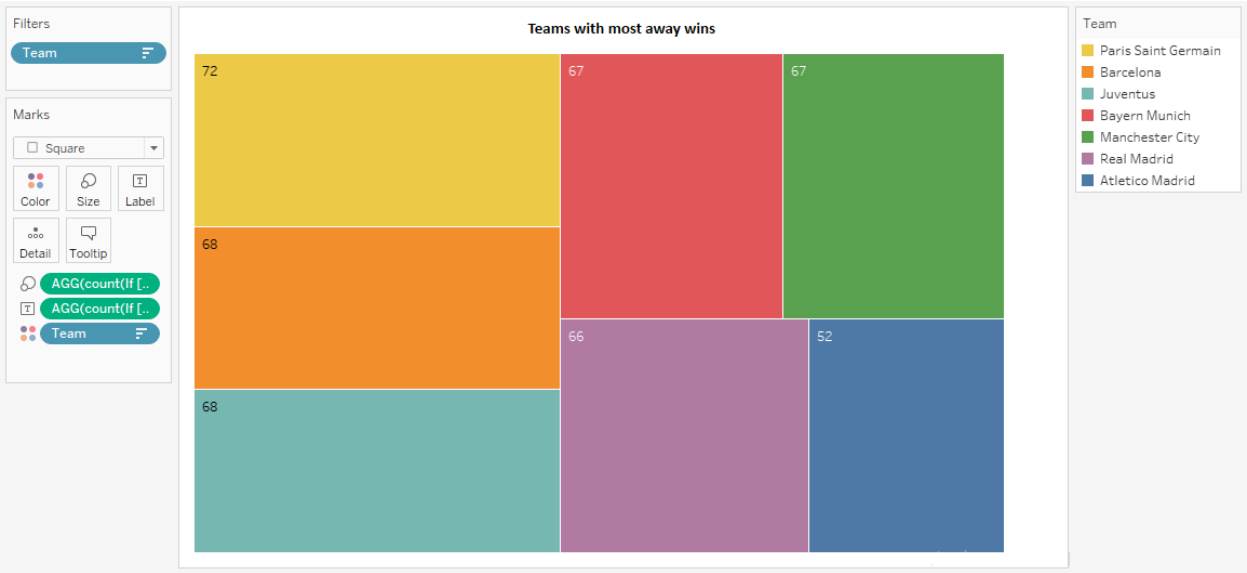
Team Monaco earned the most from transfers throughout all the seasons since 2014.

f1.



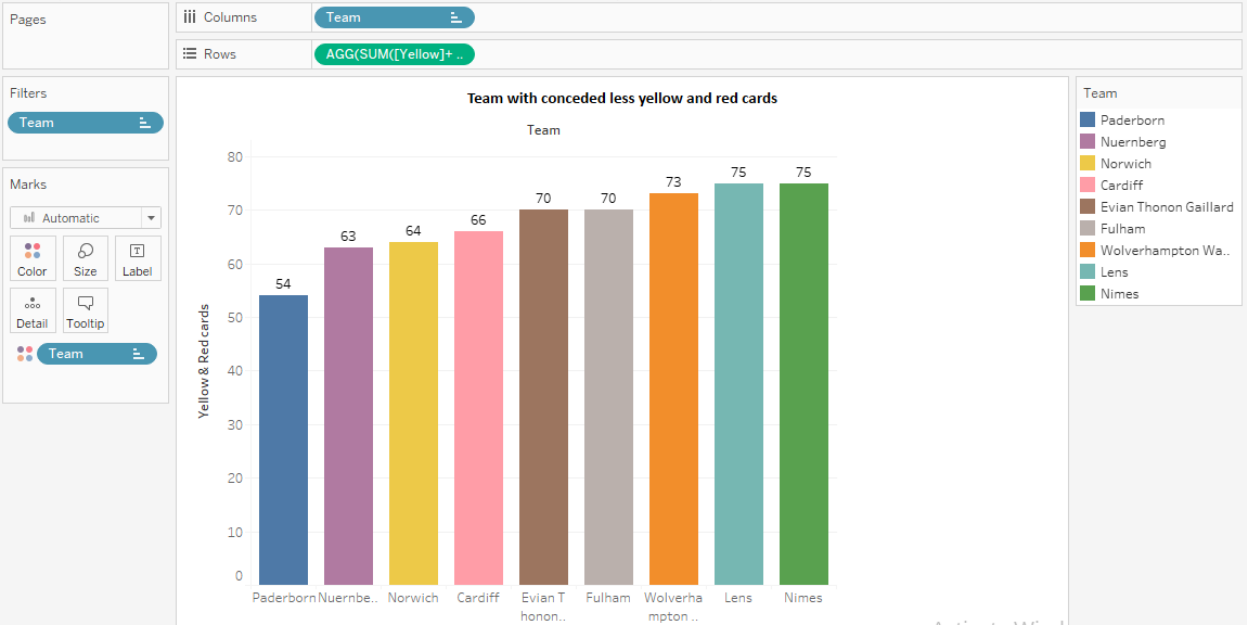
Team Juventus wins most home matches throughout all the seasons since 2014.

f2.



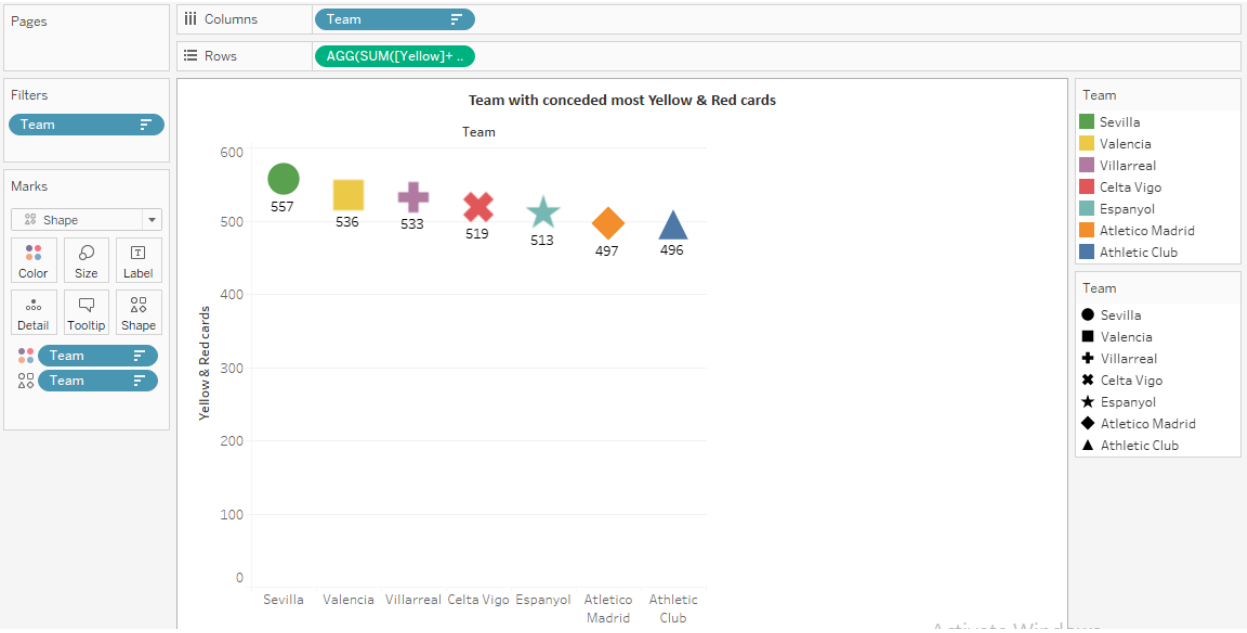
Team Paris Saint Germain(PSG) wins most home matches throughout all the seasons since 2014.

g1.



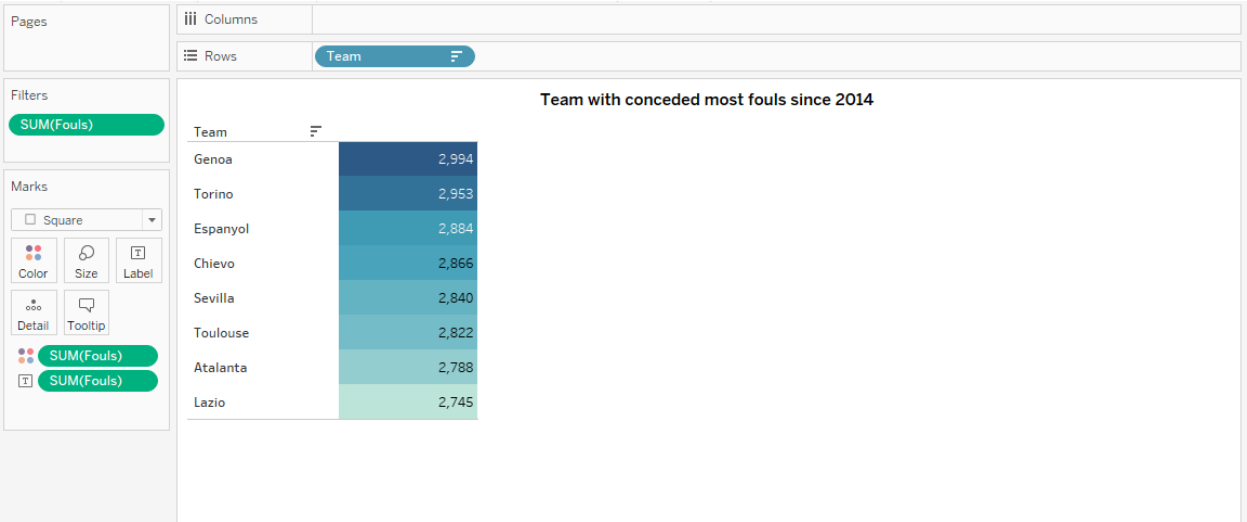
Team Paderborn played most fair matches by conceded the less yellow and red cards throughout all the seasons matches since 2014.

g2.



Team Sevilla played mostly unfair matches by conceded most yellow and red cards throughout all the seasons matches since 2014.

h.



Team Genoa conceded most fouls throughout all the seasons matches since 2014.

Conclusion

After report analysis we concluded that;

- Since 2014, both Bayern Munich and Juventus have won the most league titles out of all the teams in Europe's top five leagues.
- Since 2014, Manchester United has won the league titles with the most points among all of the Europe's top five leagues.
- Between 2014 and 2018, Team Barcelona (la liga league) scored the most goals in the top five Europe’s leagues.
- Team Manchester City has spent the most money on transfers through all the seasons since 2014.
- Since 2014, Team Monaco has made the most money through transfers across all seasons.
- Since 2014, Juventus has won the most home matches in the Europe’s top five leagues.
- Since 2014, the team Paris Saint-Germain (PSG) has won the most home matches in Europe’s top five leagues.
- Team Paderborn played most fair matches by conceded the less yellow and red cards throughout all the seasons matches since 2014.
- Team Sevilla played mostly unfair matches by conceded most yellow and red cards throughout all the seasons matches since 2014.
- Since 2014, Genoa has been the team that has conceded the most fouls in all of its matches.