Football Wages Analysis

Matt Clark - February 2025

This project aimed to analyse the wages of football (soccer) players and determine whether the best players are also the highest paid players.

The data show that overall there is a clear trend that spending more money on wages results in a higher finish in the league. However, there have been a significant number of instances where the highest paying team did not finish 1st in their league. Therefore, the answer to the question of are the best players also the highest paid is usually yes but not always.

Introduction

Data on the amount of money football teams have spent on wages and where teams finished in the league were collected. These were the annual spends over the previous 10 seasons (from 2014-2015 until 2023-2024) and were all converted to GBP. The top 5 leagues across Europe were looked at:

- Premier League, PL (England)
- Serie A, SA (Italy)
- La Liga, LL (Spain)
- Ligue 1, LU (France)
- Bundesliga, BL (Germany)

These data were obtained by web scraping using Python. They were then cleaned and manipulated using SQL (MySQL). They were then visualised using Tableau.

Data Collection

The source of the data for this project is https://fbref.com/en/, a website that contains many football-related data. Those required for this project were the amount of money players were paid by each team as well as the teams final league position for the previous 10 complete seasons. The data weren't in a downloadable format so to obtain them they were scraped using python and turned into CSV files.

Here is a breakdown of the process. The full code is uploaded to the repository.

The first step was to install the required modules. To view all of the data each url had to be navigated. Therefore selenium was used to identify and click buttons.

```
from bs4 import BeautifulSoup as bs
import requests
import pandas as pd
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
import time
```

The data for each country's league were located on different urls. These were listed and then inserted into a list. Iterating over each url using this list made the code more succinct. The abbreviated names of each league and the seasons were also put into lists which were used to name the files to be saved.
The process of obtaining the data used the following steps:

- Open the url with the Chrome web driver
- Click the pop-up to reject cookies
- Navigate the url to the desired seasons data
- Parse the html
- Find the correct table within the html
- Find the column headers and create a data frame with those headers

```
#different urls for each league
PL = "https://fbref.com/en/comps/9/wages/Premier-League-Wages"      #premier league
SA = "https://fbref.com/en/comps/11/wages/Serie-A-Wages"            #serie a
LL = "https://fbref.com/en/comps/12/wages/La-Liga-Wages"            #la liga
BL = "https://fbref.com/en/comps/20/wages/Bundesliga-Wages"         #bundesliga
LU = "https://fbref.com/en/comps/13/wages/Ligue-1-Wages"            #ligue 1
urls = [PL, SA, LL, BL, LU]
leagues = ['PL','SA','LL','BL','LU']
seasons = ['23-24','22-23','21-22','20-21','19-20','18-19','17-18','16-17','15-16','14-15']
```

- Find the rows of data and within them find the individual data points
- Strip (clean-up) the data and insert them into the data frame
- Delete any duplicate rows
- Save the data as a CSV in the desired filepath

This was then repeated for each season and each league, generating 50 CSVs.

```
for i in range(len(urls)):
    url = urls[i]                   #Loop throuh each league using their urls

    #opens the url with the chrome web driver
    driver = webdriver.Chrome()
    driver.get(url)

    #rejects all cookies when page opens up
    no_cookies = driver.find_element(By.XPATH, "//button[@class=' osano-cm-denyAll osano-cm-buttons__button osano-cm-button osano-
    no_cookies.click()

    #waits 1 second to allow page to load
    time.sleep(1)

    for j in range(len(seasons)):

        #clicks previous season button to go back to previous season
        prev_season = driver.find_element(By.XPATH, "//a[@class='button2 prev']")
        prev_season.click()

        time.sleep(1)

        if j == 3:
            popup_close = driver.find_element(By.XPATH, "//div[@id='modal-close']")    #closes pop-up that appears
            popup_close.click()
            time.sleep(1)

        #defines 'soup' as the pages html and 'table' as the specific html for the table of data
        soup = bs(driver.page_source, "html.parser")
        table = soup.find("table", {"id": "squad_wages"})

        #pulls the table headers from the html
        table_titles = table.find_all("th", {"scope": "col"})
        titles = [title.text.strip() for title in table_titles]
```

```
        #creates a data frame with the table headers
        df = pd.DataFrame(columns = titles)

        all_data = []

        for t in table:
            rows = table.find("tbody").find_all("tr")          #finds the rows of data using the tags 'tbody' and 'tr'
            for row in rows:
                cols = row.find_all(["th","td"])               #finds the individual data using the tags 'th' and 'td'
                cols = [col.text.strip() for col in cols]      #cleans the data using strip()
                all_data.append(cols)
                length = len(df)                               #adds the data to the dataframe
                df.loc[length] = cols


        #deletes duplicate rows
        df = df.drop_duplicates()

        #saves file as league + season in that leagues folder e.g. for the first one 'PL 23-24.csv' in the PL folder within t
        folder = r"C:\Users\mattc\OneDrive\Documents\Data\Football Wages\\" + leagues[i] + "\\"
        filename = leagues[i] + " " + seasons[j] + ".csv"
        df.to_csv(folder + filename, index=False)


#closes the web driver browser
driver.quit()
driver = None

time.sleep(1)
```

These raw data were then imported into MySQL to be cleaned and manipulated.

Data Cleaning and Manipulation

The raw data were imported into MySQL. The full code is uploaded to the repository:

- wagescleanup.sql
- winnerstable.sql
- combinedwinnerstable.sql
- wagepercents.sql
- highestwagefinishes.sql

The first step was to clean the data (wagescleanup.sql). The annual wages column contained the amount in 3 different currencies within the same cell, along with symbols, commas, spaces and brackets. This was reduced to just the numbers in a single currency (GBP). The data type was also converted to integer.

```
update lu_23_24
set `Annual Wages` = regexp_replace(`Annual Wages`, '^[^£]*£',''),
    `Annual Wages` = regexp_replace(`Annual Wages`, '\\$.*',''),
    `Annual Wages` = trim(`Annual Wages`),
    `Annual Wages` = cast(replace(`Annual Wages`, ',', '') as unsigned)
;
```

Data before cleaning

Rk	Squad	# Pl	Weekly Wages	Annual Wages	% Estimated	Final Position
1	Manchester Utd	86	Â£ 3,719,423 (â‚¬ 4,435,653, $4,519,922)	Â£ 193,410,000 (â‚¬ 230,653,939, $235,035,9...	3%	2
2	Chelsea	82	Â£ 2,842,538 (â‚¬ 3,389,911, $3,454,313)	Â£ 147,811,988 (â‚¬ 176,275,358, $179,624,2...	2%	4
3	Arsenal	87	Â£ 2,835,577 (â‚¬ 3,381,609, $3,445,853)	Â£ 147,450,000 (â‚¬ 175,843,663, $179,184,3...	1%	8

Data after cleaning

Rk	Squad	# Pl	Weekly Wages	Annual Wages	% Estimated	Final Position
1	Manchester Utd	86	Â£ 3,719,423 (â‚¬ 4,435,653, $4,519,922)	193410000	3%	2
2	Chelsea	82	Â£ 2,842,538 (â‚¬ 3,389,911, $3,454,313)	147811988	2%	4
3	Arsenal	87	Â£ 2,835,577 (â‚¬ 3,381,609, $3,445,853)	147450000	1%	8

The data then needed to be moved into different tables so that the relevant data were together in order to be visualised. First, a table of the winners (ie final position = 1) was created and populated for each league (winnerstable.sql).

```sql
create table lu_winners(           #creates the table with column headers and data types
year int,
team varchar(255),
wage_rank int)
;


insert into lu_winners (year)             #manually add the years
values
(2015),(2016),(2017),(2018),(2019),
(2020),(2021),(2022),(2023),(2024)
;


insert into lu_winners (team, wage_rank)      #add the teams and wage ranks from the exisiting tables
select squad, Rk from lu_23_24                #(change for each season table)
where `Final Position` = 1
;


#the rows don't align so need to be manipulated

with temp as (                            #creates a temporary table in a cte with an added row number
    select *,
    row_number() over() as rownum
    from lu_winners),

temp2 as (                                #creates a second temporary table in a cte where the rows are aligned using the row numbers
    select t1.year, t2.team, t2.wage_rank
    from temp t1
    join temp t2
    on t2.rownum = t1.rownum + 10)

update lu_winners w                  #updates the data in the winners table using the temporary cte table above
join temp2 t
on w.year = t.year
set w.wage_rank = t.wage_rank,
    w.team = t.team
;


delete from lu_winners             #deletes the bottom rows since they're now duplicates
where year is null
;


update lu_winners                        #converts years and wages ranks to number data types
set `Year` = CAST(`Year` AS UNSIGNED),
    `wage_rank` = CAST(`wage_rank` AS UNSIGNED)
;

```

This created 5 tables - one for each league - as below for Ligue 1.

year	team	wage_rank
2015	Paris S-G	1
2016	Paris S-G	1
2017	Monaco	2
2018	Paris S-G	1
2019	Paris S-G	1
2020	Paris S-G	1
2021	Lille	6
2022	Paris S-G	1
2023	Paris S-G	1
2024	Paris S-G	1

They were then combined into a single table (combinedwinnerstable.sql).

```sql
create table winners(            #creates the table with column headers and data types
year int,
pl_team varchar(255),
pl_wage_rank int,
sa_team varchar(255),
sa_wage_rank int,
ll_team varchar(255),
ll_wage_rank int,
bl_team varchar(255),
bl_wage_rank int,
lu_team varchar(255),
lu_wage_rank int)
;


insert into winners (year)              #manually add the years
values
(2015),(2016),(2017),(2018),(2019),
(2020),(2021),(2022),(2023),(2024)
;


update winners w                    #adds the data from the league winners tables
join lu_winners p
on w.year = p.year
set w.lu_team = p.team,
    w.lu_wage_rank = p.wage_rank
;

```

year	pl_team	pl_wage_rank	sa_team	sa_wage_rank	ll_team	ll_wage_rank	bl_team	bl_wage_rank	lu_team	lu_wage_rank
2015	Chelsea	3	Juventus	3	Barcelona	1	Bayern Munich	1	Paris S-G	1
2016	Leicester City	17	Juventus	1	Barcelona	1	Bayern Munich	1	Paris S-G	1
2017	Chelsea	4	Juventus	1	Real Madrid	2	Bayern Munich	1	Monaco	2
2018	Manchester City	3	Juventus	1	Barcelona	1	Bayern Munich	1	Paris S-G	1
2019	Manchester City	3	Juventus	1	Barcelona	1	Bayern Munich	1	Paris S-G	1
2020	Liverpool	5	Juventus	1	Real Madrid	1	Bayern Munich	1	Paris S-G	1
2021	Manchester City	4	Juventus	1	AtlÃ©tico Madrid	3	Bayern Munich	1	Lille	6
2022	Manchester City	3	Inter	2	Real Madrid	1	Bayern Munich	1	Paris S-G	1
2023	Manchester City	2	Milan	5	Barcelona	2	Bayern Munich	1	Paris S-G	1
2024	Manchester City	2	Napoli	5	Real Madrid	1	Leverkusen	4	Paris S-G	1

Another table was created, this one containing the percentage each teams wage spend was out of the total for their league (wagepercents.sql).

```sql
create table wage_pcts(                #create a table containing a column for wage percentages
year integer,
final_position integer,
team varchar(255),
league varchar(255),
wage_pct_league_total float)
;


alter table lu_23_24                     #add a column for wage percentages to each existing table
add column wage_pct_league_total float
;


with total_wages as (                                #create a CTE with the total wages spent across the league
select sum(`Annual Wages`) as total_wages
from lu_23_24)
update                                    #populate the wage percentage column with each teams wage spend as a percentage of the league total
lu_23_24 l,
total_wages t
set l.wage_pct_league_total = round(100.0*l.`Annual Wages`/t.total_wages,1)
;


 insert into                                                    #insert the data from each table into the wage percentage table
 wage_pcts (final_position, team, wage_pct_league_total)
 select `Final Position`, squad, wage_pct_league_total
 from lu_23_24
 ;


 update wage_pcts                 #fill in the year column
 set year = 2024
 where year is null
 ;


 update wage_pcts                 #fill in the season column
 set league = 'LU'
 where league is null
 ;


 select *
 from wage_pcts
 order by league desc, year desc, final_position
 ;
```

Here are the first few rows of the table (991 rows in total)

year	final_position	team	wage_pct_league_total	league
2024	1	Inter	11.6	SA
2024	2	Milan	8.5	SA
2024	3	Juventus	12	SA
2024	4	Atalanta	4.5	SA
2024	5	Bologna	2.9	SA
2024	6	Roma	10.3	SA
2024	7	Lazio	7.3	SA
2024	8	Fiorentina	5.9	SA
2024	9	Torino	3.9	SA
2024	10	Napoli	7.4	SA
2024	11	Genoa	3.1	SA
2024	12	Monza	2.8	SA
2024	13	Hellas Ve...	2.3	SA
2024	14	Lecce	1.4	SA
2024	15	Udinese	2.6	SA
2024	16	Cagliari	3.1	SA

Finally a table containing data about the final positions of the teams who had paid the highest wages was created (highestwagefinishes.sql).

```sql
with ranked as (                        #create a CTE with each team ranked by how much they paid in wages
select
*,
row_number() over(partition by league, year order by wage_pct_league_total desc) as wage_rank
from wage_pcts)

select                                  #output the count of final positions of those who paid the most in wages (i.e. wage rank = 1)
    league,
    count(case when final_position = 1 then 1 end) as 1st,
    COUNT(case when final_position = 2 then 1 end) as 2nd,
    COUNT(case when final_position = 3 then 1 end) as 3rd,
    COUNT(case when final_position = 4 then 1 end) as 4th,
    COUNT(case when final_position = 5 then 1 end) as 5th,
    COUNT(case when final_position = 6 then 1 end) as 6th,
    COUNT(case when final_position = 7 then 1 end) as 7th,
    COUNT(case when final_position = 8 then 1 end) as 8th,
    COUNT(case when final_position = 9 then 1 end) as 9th,
    COUNT(case when final_position = 10 then 1 end) as 10th
from ranked t1
where wage_rank = (
    select MIN(wage_rank)
    from ranked t2
    where t2.league = t1.league
)
group by league
order by league
;
```

	league	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
▶	BL	9	0	1	0	0	0	0	0	0	0
	LL	7	2	1	0	0	0	0	0	0	0
	LU	8	2	0	0	0	0	0	0	0	0
	PL	0	2	3	2	0	2	0	1	0	0
	SA	5	0	1	2	0	0	1	0	0	1

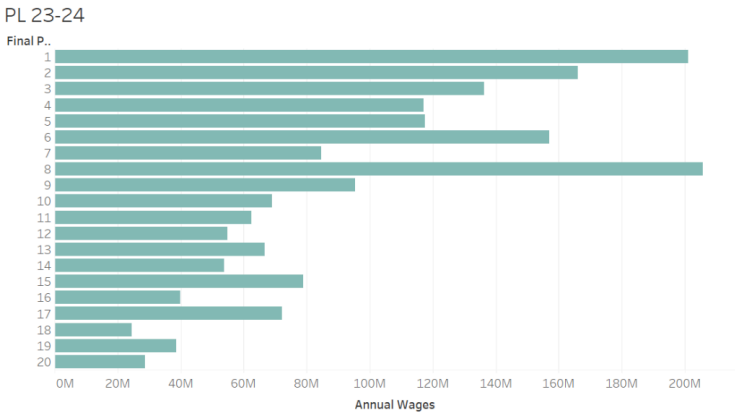
The generated tables were then imported into Tableau to be visualised.

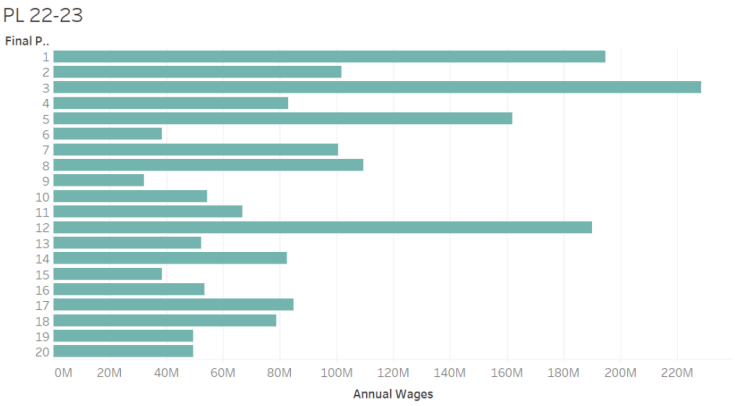
Data Visualisation

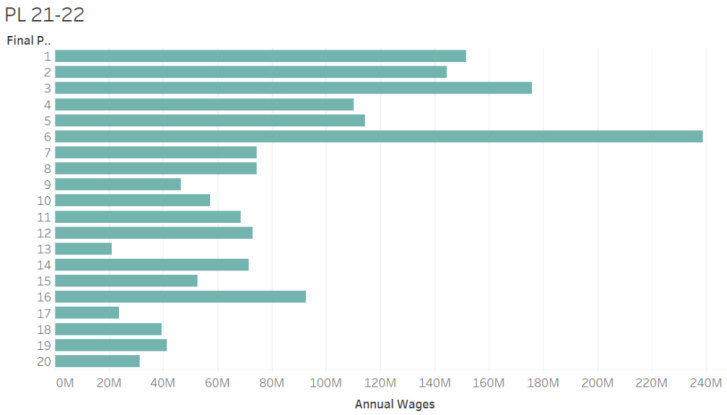
The data tables were imported into Tableau to be visualised. The dashboard is uploaded to the repository.

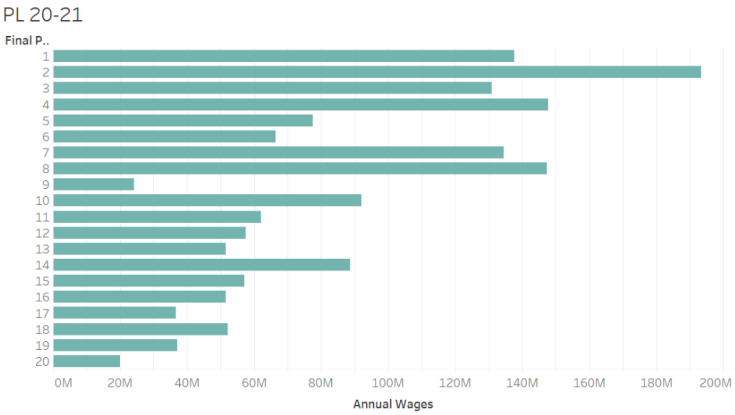
First, the annual wage totals of each team in order of their final position for the previous 4 full seasons for each league.

Premier League (England)



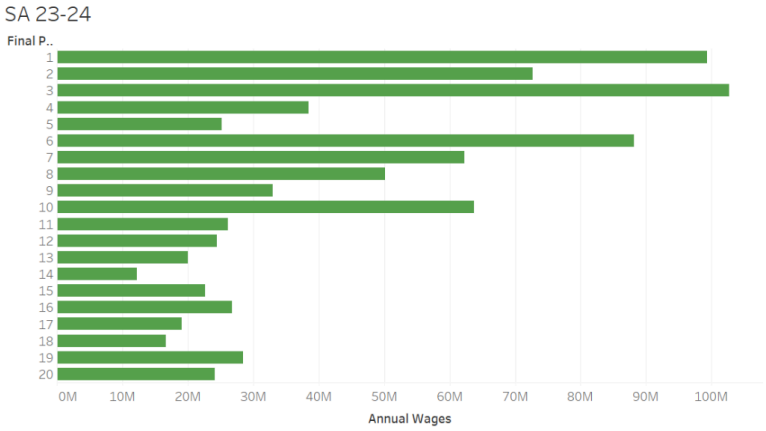


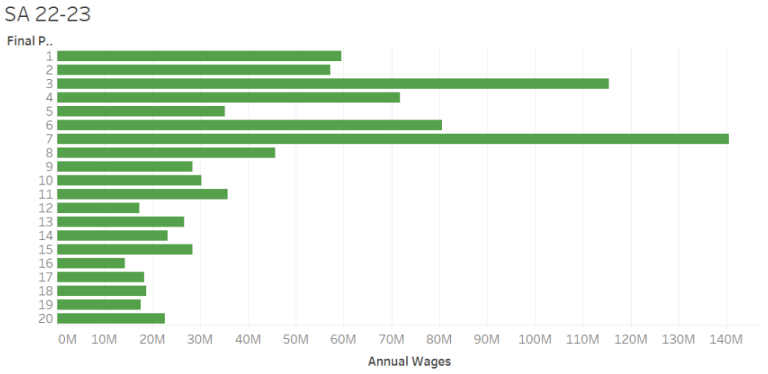


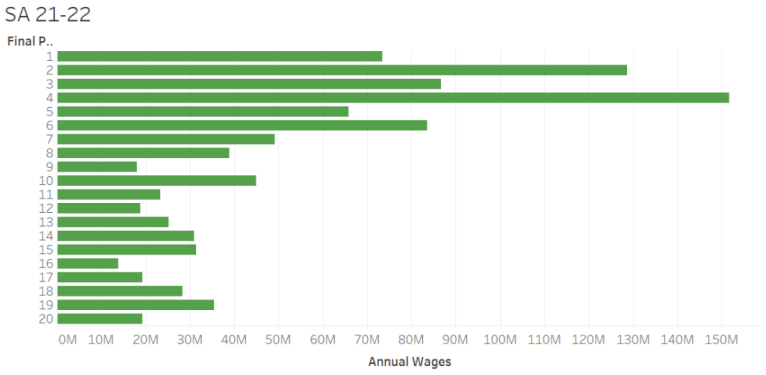


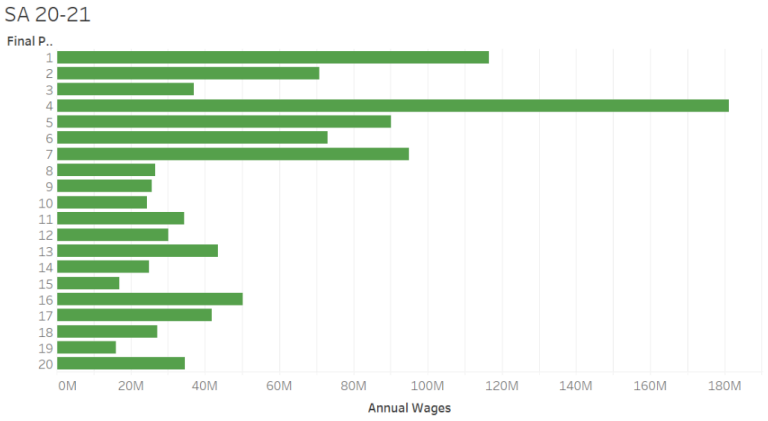
In the 3 of the last 4 premier league seasons there is a general trend that the teams finishing higher have a higher wage spend. In the 22-23 season there is quite a large deviation from this trend with many teams finishing higher than others despite having spent less. Also, in the 23-24 and 21-22 season there are outliers - the team with the highest wage spend finished in 8th and 6th respectively.

Serie A (Italy)



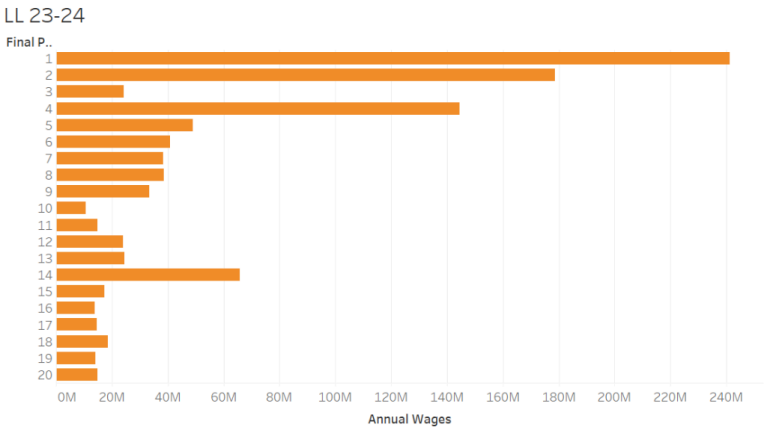


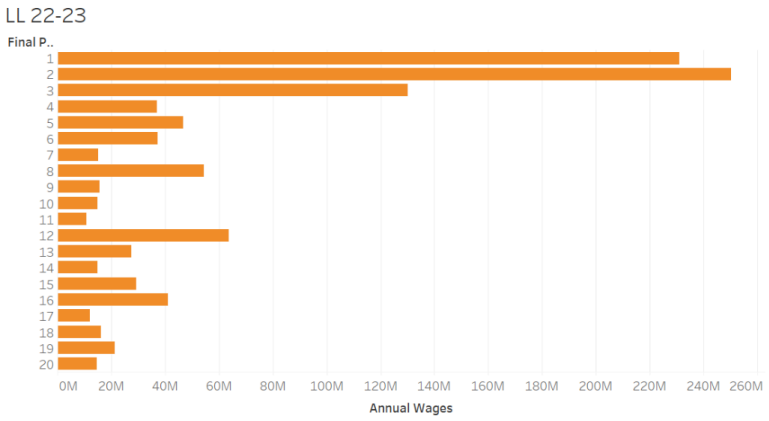


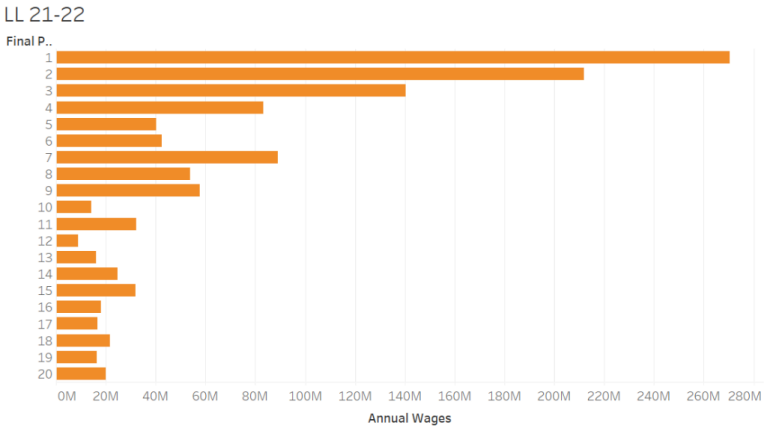


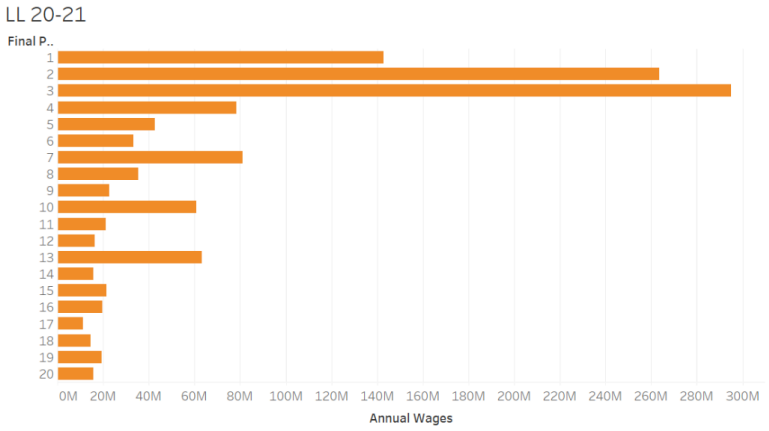
In the last 4 Serie A seasons, there is a rough trend of the teams finishing higher having a higher wage spend. There are 2 groups that emerge - the top 6 or 7 and the rest. Generally the teams finishing lower than 6th or 7th have significantly lower wage spends than those above 6th or 7th. Within those 2 groups, the teams aren't consistently in order, with many teams finishing higher than teams that have spent more than them. There are also some outliers, such as the teams that finished 5th in the 23-24 and 22-23 seasons having far lower wage spends than the 3 teams below and 4 teams above them.

La Liga (Spain)



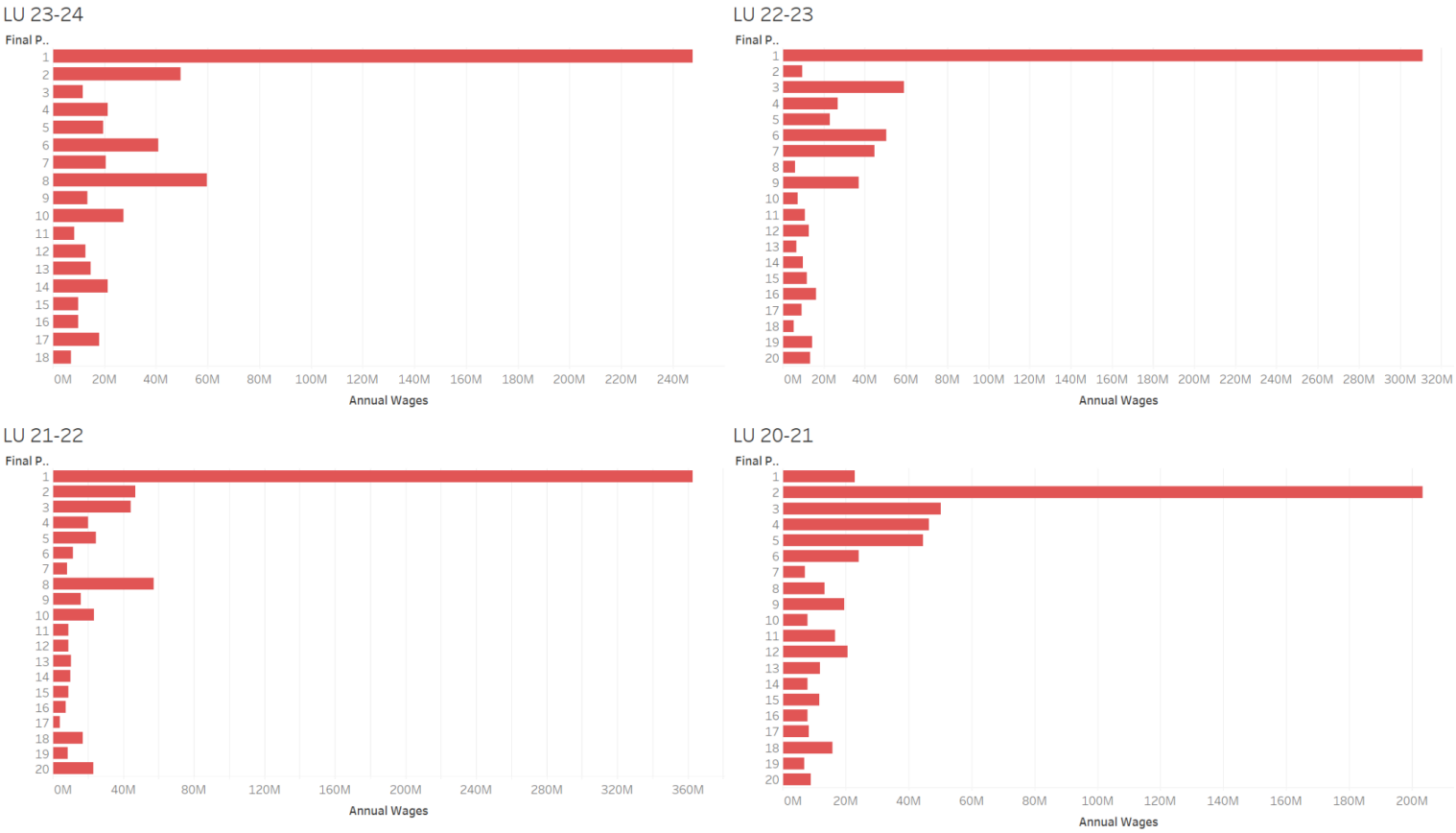






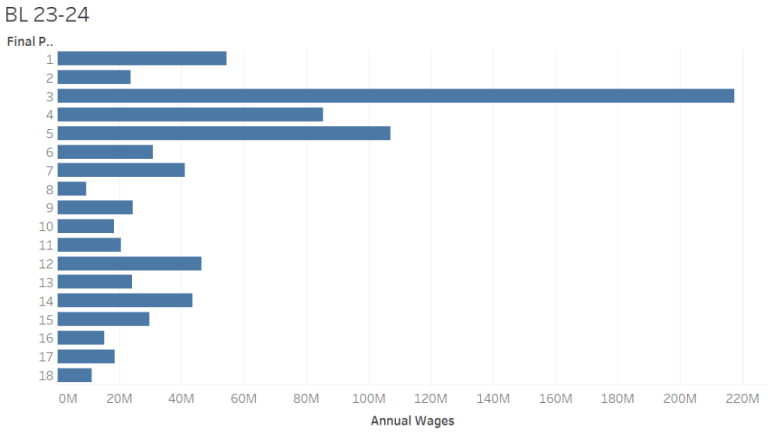
In the previous 4 La Liga seasons, there is a clear group. The teams finishing in the top 3 have the 3 highest wage spends by far, with the exception of 23-24 where the team finishing 3rd having a much lower wage spend and the team finishing 4th having a significantly higher wage spend. Amongst the teams below 3rd, there is a rough trend of a higher wage spend resulting in a higher finish. However it's not entirely consistent and there are several outliers.

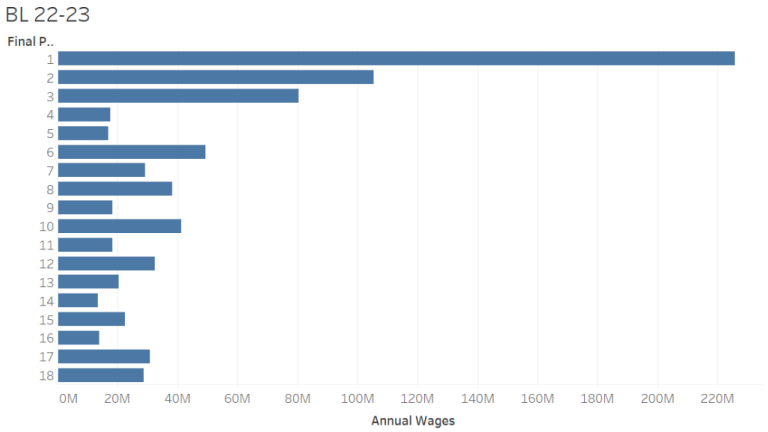
Ligue 1 (France)

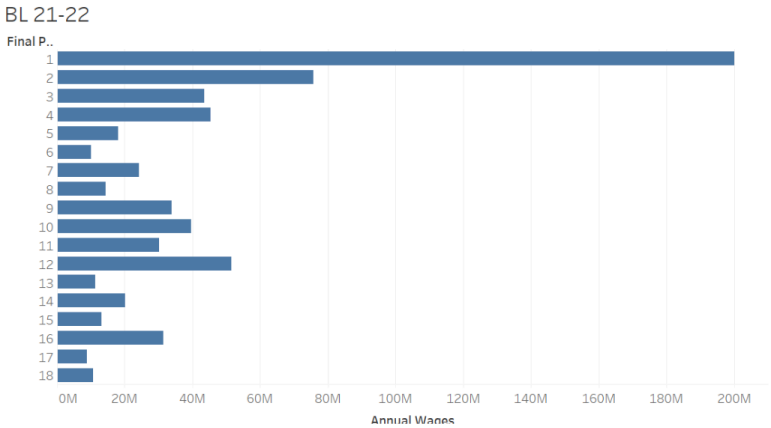


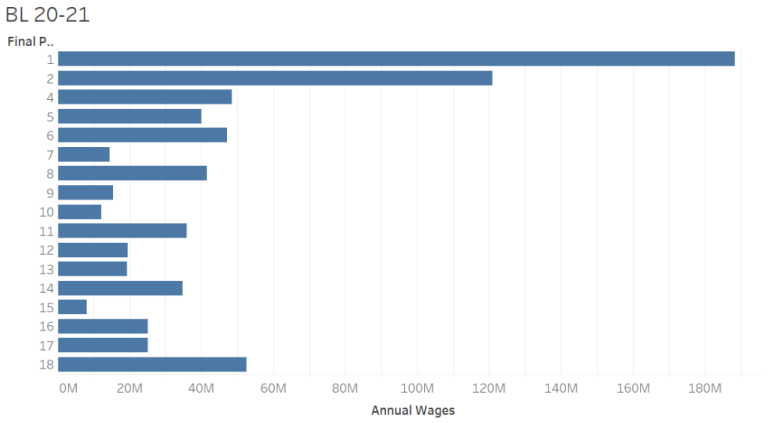
In 3 of the previous 4 Ligue 1 seasons, the team finishing 1st has the highest wage spend by far. In the other, 20-21, the team finishing 2nd had the highest wage spend by far. Of the remaining teams, there is a general trend showing that a higher wage spend results in a higher league finish, particularly in 20-21. However, there are some notable outliers. In 21-22, the teams finishing 18th and 20th have higher wage spends than all teams up to and including 11th. In 22-23, the team finishing 2nd have a far lower wage spend than the 5 teams directly below them.

Bundesliga (Germany)



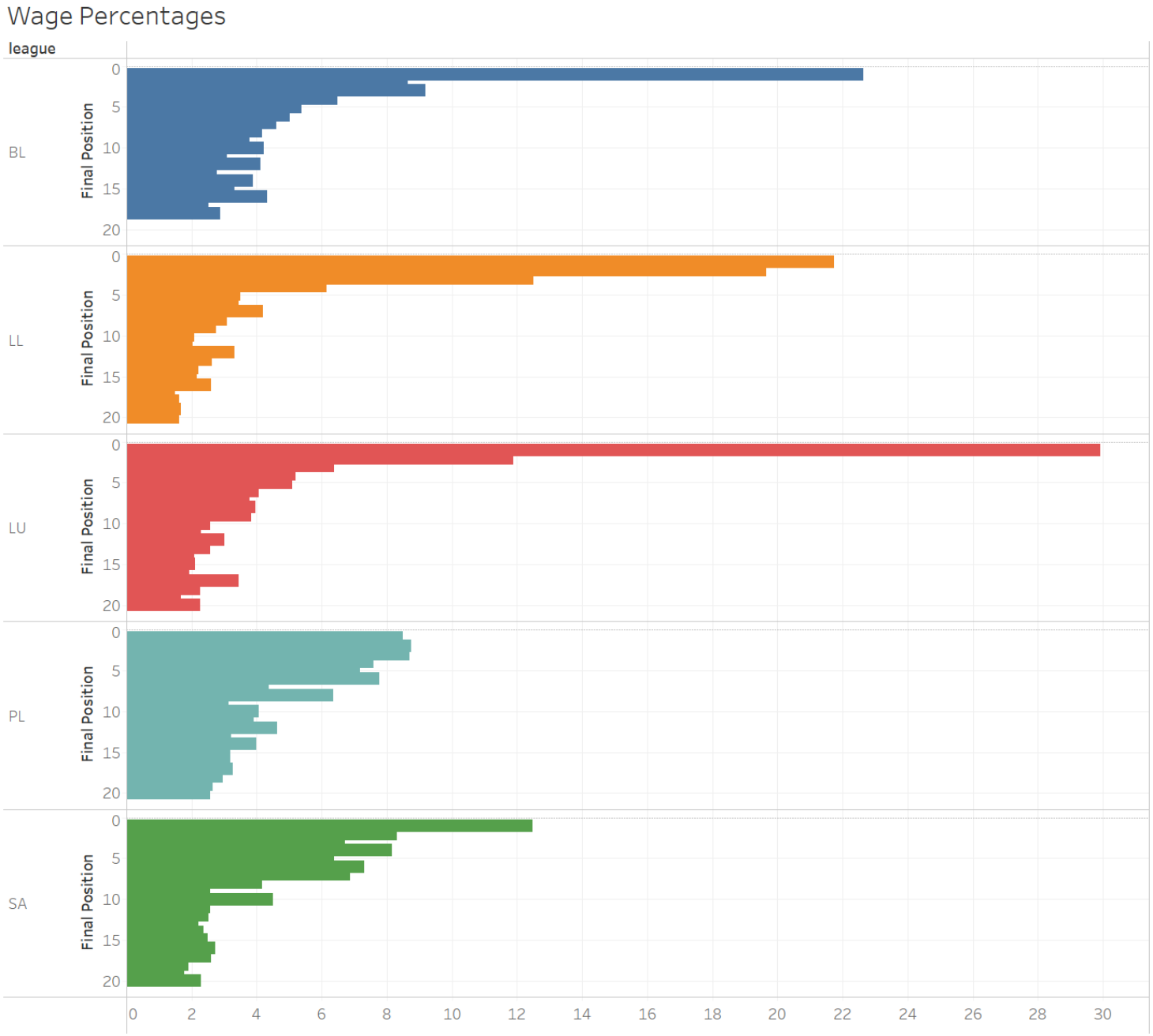






In the most recent Bundesliga season, 23-24, the teams are in a quite inconsistent order. The 4 highest wages spends finished in the top 5, however the teams finishing in 1st and particularly 2nd are far lower than the following 3 finishing teams. Below that, many teams finished above teams that spent more than them. In the preceding seasons, the teams finishing in the top positions had the highest wage spends and were roughly in order. Below the top 4 or 5 finishing teams, the order is much less consistent. Particularly in 20-21, where the team finishing 18th (last) had a higher spend than all teams other than those that finished 1st and 2nd. Also, the teams finishing 7th, 9th and 10th had lower wage spends than all teams that finished below them other than 1 (15th).

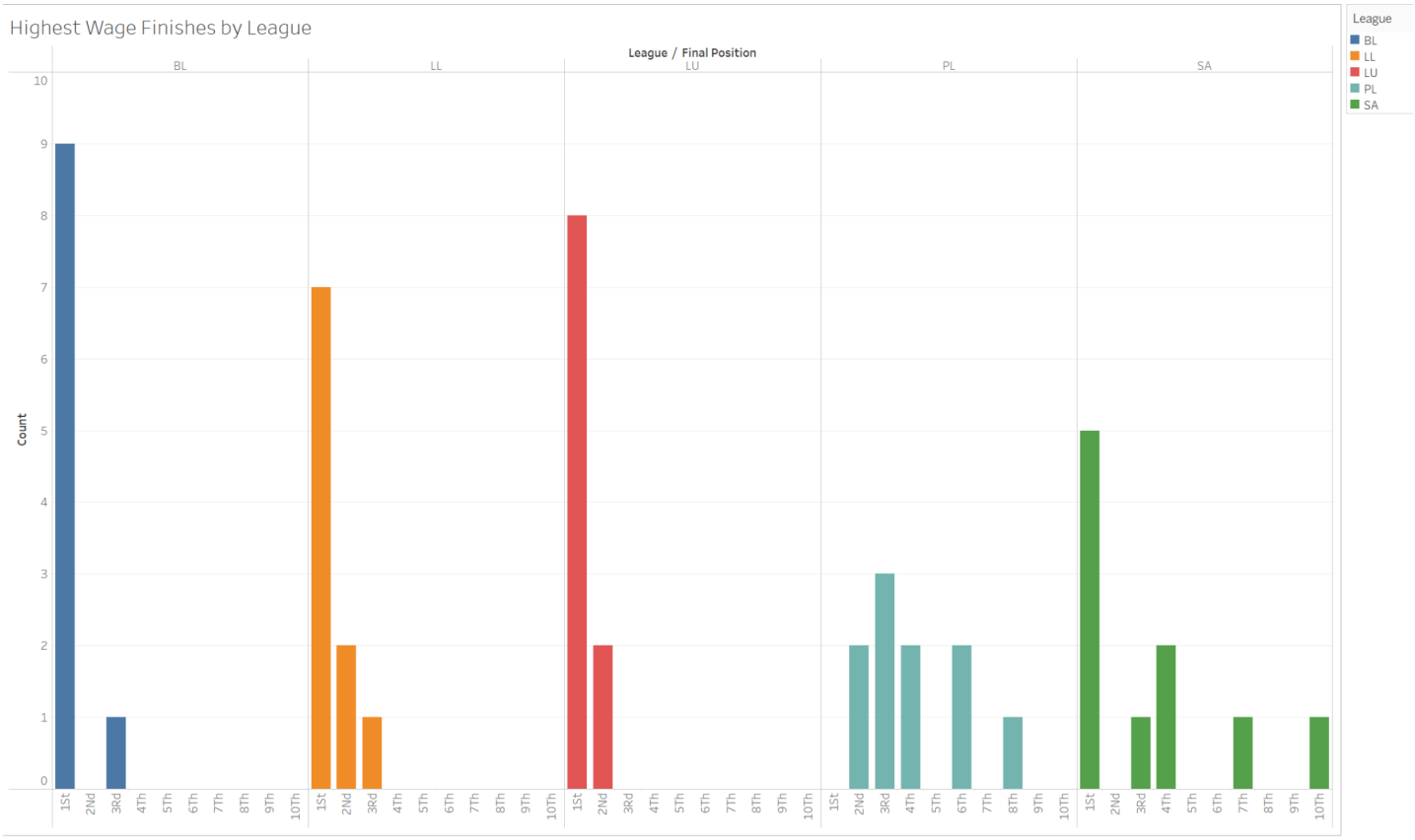
Next, each teams wage spend was converted to a percentage of the total spend of the league and they were plotted against their finishing positions. The previous 10 seasons data were combined. This shows a much clearer trend - spending more results in a higher finishing position.



In 3 of the leagues (Bundesliga, La Liga and Ligue 1) there is a very clear disparity between the teams finishing at the top and at the bottom. Particularly in Ligue 1, where the top team had almost 30% of the leagues wage spend on average, with the 2nd team having almost 12% and every other team having below 7%. In Serie A there is a similar trend but far less exaggerated. It's also clear to see the 2 groups which emerged in the graphs above, with the top 7 teams all having over 7% of the leagues wage spend each and all the rest having below 5%. In the Premier League, there is even less disparity between the top and the bottom. It's also

interesting to note that the team finishing 1st has a lower wage spend on average than those finishing 2nd and 3rd.

Finally, the finishing positions of the team with the highest wage spend were counted over the previous 10 seasons. This shows that generally spending the most on wages results in finishing 1st.



Again, in 3 of the leagues (Bundesliga, La Liga and Ligue 1) there is a clear trend. The highest spending team finished 1st in most of the previous 10 seasons and never finished lower than 3rd. In the Premier league, it is much more spread. In fact, the team with the highest spend didn't finish 1st in any of the previous 10 seasons, the most common finish was 3rd and the lowest finish was 8th. In Serie A, the highest spending team finished 1st in 5 of the previous 10 seasons. However, in the other 5 seasons the finishing positions were also quite spread, with the lowest being 10th.

Conclusion

In general, the data show that spending more money on wages results in a higher finishing position in each of the 5 leagues. In 3 of the leagues (Bundesliga, La Liga and Ligue 1) it is very common for the highest paying team to finish 1st, with this also being the case in half of the previous 10 seasons in Serie A. However, in the Premier League, the highest paying team hasn't finished 1st in any of the previous 10 seasons. This project aimed to analyse whether the highest paid football players are also the best players. The data show that the answer to this is that usually they are but not always. There have been a significant number of instances where the highest paying team hasn't finished 1st in their league, including at least 1 instance in each of the 5 leagues over the previous 10 seasons.