# Analiza danych dotyczących klubów piłkarskich od sezonu 2015/2016

Nikodem Świerkowski

June 8, 2023

## Contents

# 1 Temat projektu

Celem projektu jest odpowiedzenie na pytanie czy na podstawie zachowania klubu na rynku transferowym można przewidzieć końcowy wynik klubu w lidze krajowej. Zebrane dane dotyczą lat od 2016 (sezon 2015/2016) do 2023 (sezonu 2022/2023). Zachowanie klubu na rynku transferowym zostanie opisane za pomocą:

- wydatków klubu w ciągu całego sezonu
- wpływów do klubu w ciągu całego sezonu
- salda transferowego w sezonie, czyli różnicy między wpływami a wydatkami klubu
- liczbą nowych zawodników, którzy przybyli do klubu
- liczbą zawodników odchodzących z klubu
- suma zawodników, którzy opuścili i przybyli do klubu

Końcowy wynik klubu w lidze będzie opisany za pomocą liczbą punktów zdobytych przez klub w ciągu całego sezonu.

# 2 Zdobycie danych

Dane zostały zdobyte poprzez scrapowanie danych z dwóch stron internetowych. Do uzyskania danych dotyczących finansów klubów oraz lig w konkretnych sezonach została użyta strona internetowa "Transfermarkt" ([1]), natomiast do uzyskania danych dotyczących tabeli wyników końcowych klubów, wykorzystana została strona "Wikipedia" ([2]). Brakujące dane dotyczące tabel końcowych lig, które nie zostały znalezione na wymienionej stronie np. tabela końcowa Chinese Super League w roku 2021, zostały pozyskane ręcznie z podstron strony Transfermarkt. Próba zautomatyzowania pobrania brakujących danych była skazana na niepowodzenie przez dużą wybiórczość, które dokładnie dane należy uzupełnić. Do scrapowania danych użyto bibliotek Pythona - Selenium oraz BeutifulSoup. Selenium została wykorzystana do załadowania strony internetowej, natomiast BeutifulSoup do wyszukania odpowiedniej tabeli na stronie. Dodatkowo zebrano jeszcze dane dotyczące finansowych statystyk lig w poszczególnych sezonach, co pozwoli pokazać punkt odniesienia do wartości związanych z klubami np. ile klub wydał w stosunku do wszystkich wydatków ligi.

# 3 Wyczyszczenie danych

## 3.1 Usunięcie zbędnych danych

Na początku zostały zgromadzone dane dotyczące zachowania klubów na rynku transferowym. Skrypt "count_leagues.py" powstał w celu znalezienia informa-

cji:

- ile jest wierszy powiązanych z konkretną ligą
- ile jest wszystkich recordów
- jakie ligi nie posiadają wierszy związanych ze wszystkimi ośmioma rozpatrywanymi sezonami

Na podstawie tych informacji ustalone zostało, że tabela posiada wiele lig, które nie występują w każdym z sezonów oraz że całkowita liczba lig wynosi 99. Należy jednak zwrócić uwagę, że wiele z tych lig to takie, które posiadają jedynie pojedyncze rekordy albo błędnie nazwane ligi, które nie występują w rzeczywistości np. "Hiszpania". Scrapowanie danych o wynikach klubów z tych wszystkich lig byłoby bardzo czasochłonne i wymagające scrapowania całej tabeli w danym sezonie, tylko dlatego że jedna drużyna akurat znalazła się w zbiorze danych lub mogło okazać się, że liga pod tą nazwą nie istnieje tak jak wspomniana wyżej "Hiszpania". Aby uniknąć takich przypadków dane dotyczące wyników klubów zostały ograniczone do lig, które zawierają nie mniej niż 40 rekordów, co pozostawiło jedynie najważniejsze ligi. Tak przygotowana tabela została zapisana do pliku "repaired_finances.csv".

## 3.2 Stworzenie tabeli dotyczących wyników drużyn

Zgromadzone dane dotyczące tabel końcowych zostały zapisane jako pliki csv w folderach odpowiadających nazwom lig których dotyczą, a do nazw plików została dodana data, która symbolizuje, jakiego sezonu dotyczy dana tabela np. "Ligue_12018.csv" dotyczy najwyższego szczebla rozgrywkowego we Francji w sezonie 2017/2018. Z powodu różnego sposobu zapisania nagłówków w tabelach niemożliwe było stworzenie skryptu łączącego pliki w jeden duży przechowujące wszystkie dane za pomocą skryptu, dlatego zostało zrobione to ręcznie. Tak stworzony plik został nazwany jako: "club_results.csv" i zawiera następujące kolumny: Klub, M (oznaczające liczbę rozegranych meczy przez klub w sezonie), Pkt (liczba zdobytych punktów przez klub), Sezon, Rozgrywki (nazwa ligi).

## 3.3 Ujednolicenie nazw lig

Zebranie danych z dwóch niezależnych źródeł zaskutkowało różnicami w nazewnictwie klubów piłkarskich np. "Arsenal" i "Arsenal FC", albo "Juventus Turyn" i "Juventus". Uniemożliwiało to stworzenie, przy pomocy operatora algebry relacji "LEFT JOIN" jednej tabeli, w której każdemu klubowi z tabeli "repaired_finances.csv" zostałaby przyporządkowana liczba rozegranych meczów oraz liczba zdobytych punktów z tabeli "club_results.csv". W celu ujednolicenia nazw powstał skrypt "repair_clubs_names.py", którego zadaniem było wczytanie wszystkich lig z obu plików, a następnie znalezienie takich samych nazw, lub zawierających w sobie te same słowa, tak jak we wcześniejszym przykładzie "Arsenal" i "Arsenal FC". Skrypt ostatecznie stworzył plik "corresponding_club_names.py", który pełnił rolę swoistego słownika nazw lig z

"club_results.csv" do formy tych z "repaired_finances.csv". Niestety część nazw nie można było wyszukać i przyporządkować do siebie w prosty sposób za pomocą skryptu jak np. "RasenBallsport Leipzig" i "RB Lipsk". Takie przypadki zostały znalezione i poprawione ręcznie.

### 3.4 Połączenie tabel

Następnym krokiem do uprzątnięcia danych było sprowadzenia danych do jednej tabeli. Skrypt "join_two_dataframes.py" wykonuje działanie lewostronnegp złączenia na tabeli pobranej z pliku "repaired_finances.csv" oraz na tabeli z pliku "club_results.csv", z przekształconymi nazwami klubów z pomocą pliku "corresponding_club_names.py". Tak utworzona tabela została zapisana do pliku: "main_table.csv".

### 3.5 Wyczyszczenie danych dotyczących statystyk lig

Kolejnym krokiem było wyczyszczenie danych z tabeli "leagues_finances.csv", w której znajdowały się statystyki dotyczące finansowych sprawozdań lig. Przede wszystkim wartości dotyczące kwot, zostały zapisane w dwóch kolumnach (wartość całkowita w jednej, a w drugiej wartość po przecinku z dopiskiem sugerującym wielkość kwoty jak "mln", "mld" itd. oraz symbol waluty. W wyczyszczonej tabeli kwoty zostały zapisane w jednej komórce bez waluty oraz bez dopisku symbolizującego wielkość kwoty. Innym problem okazał się brak danych dotyczących finansów ligi meksykańskiej - MX Clausura. Jednak w tabeli "repaired_finances.csv" istnieje większość drużyn z ligi meksykańskiej w każdym sezonie, co więcej tych które wydawały najwięcej i były najbardziej aktywne na rynku transferowym. Stąd finanse dotyczące ligi meksykańskiej zostały obliczone, jako suma wydatków/zarobków/sald transferowych klubów z tej ligi w danych sezonie. Zostało przyjęte więc założenie, że nieobecne meksykańskie kluby w pliku "repaired_finances.csv", były na tyle mało aktywne że nie wpłynęłoby to w znaczącym stopniu na wydatki, zarobi czy saldo transferowe całej ligi w poszczególnych sezonach. Dodatkowo plik nie posiadał łącznej sumy punktów jakie zdobyły kluby z lig w poszczególnych sezonach. Zostało to obliczone na podstawie sumy punktów klubów z tabeli "club_results.csv". Ponieważ w tej tabeli występują wszystkie kluby biorące udział w danych rozgrywkach, zdobyte tak dane są jak najbardziej dokładne i żadna wartość nie została utracona. Wszystkie powyższe czynności zostały wykonane w pliku "league_statistics.py" i zapisane do pliku: "league_statistics.csv".

### 3.6 Dodanie atrybutów bezwzględnych

Na końcu należało wyczyścić dane z głównej tabeli oraz dodać atrybuty, które pozwoliły by rozpatrywać je bez względu na ligę, z których pochodzą np. wydatki klubów z Anglii są znacznie większe niż wydatki klubów z Holandii, ale podzielone przez sumę wydatków wszystkich klubów ze swojej ligi, różnicę te stają się możliwe do zaniedbania. Podobnie, jak w przypadku danych dotyczących kwot

z "leagues_finances.csv", zostały one zapisane w notacji z walutą oraz symbolem oznaczającym wielkość kwoty np. "mln". Należało, więc przekształcić kolumnę na wartości będące liczbami całkowitymi. Następnie tabela została rozbudowana o następujące kolumny, w celu pogłębienia analizy:

- suma zawodników przychodzących i odchodzących
- średnia liczba punktów na mecz
- wpływy klubu podzielone przez sumę wpływów wszystkich klubów z ligi
- wydatki klubu podzielone przez sumę wydatków wszystkich klubów z ligi
- liczba punktów zdobyta przez klub w lidzę podzielona przez sumę wszystkich punktów zdobytych przez kluby ligi
- liczba zawodników przychodząca do klubu podzielona przez sumę liczby zawodników przychodzących do klubów ligi
- liczba zawodników odchodzących do klubu podzielona przez sumę liczby zawodników odchodzących do klubów ligi

W późniejszych fragmentach raportu dane uniezależnione od ligi zostaną skrótowo zapisane używając dopisku "względem ligi" np. wydatki względem ligi. Oczywiście będzie to znaczyło, że mówimy o atrybucie: wydatki klubu podzielone przez sumę wydatków wszystkich klubów z ligi.

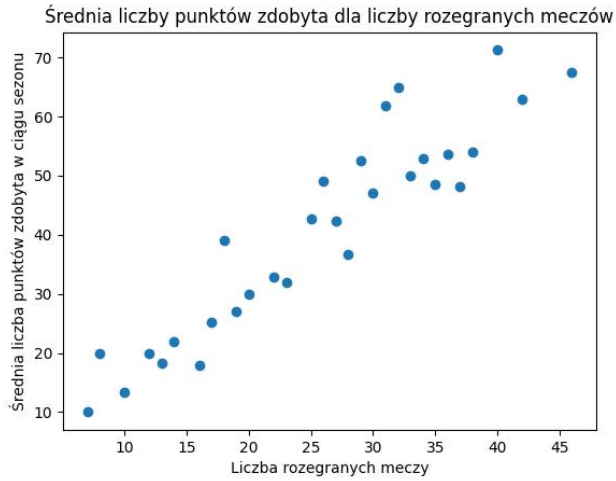
# 4 Wstępna analiza danych

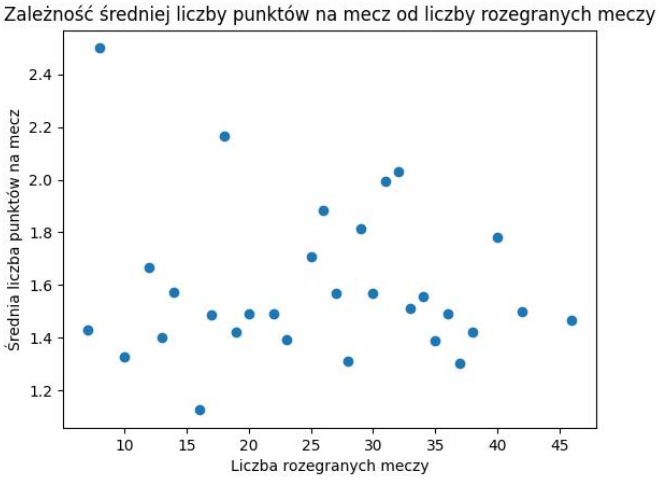
## 4.1 Podstawowe zależności

Należy zwrócić uwagę, że w zbiorze danych występuje wiele spodziewanych zależności, które należy uwzględnić.

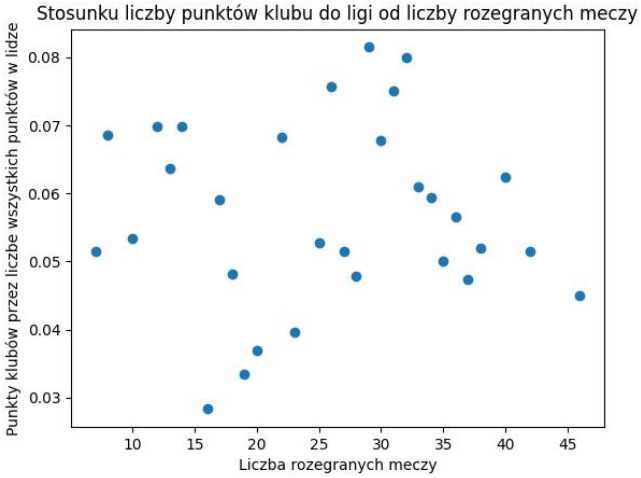
Powyższy wykres prezentuje, że średni wynik punktowy w lidzę jest zależny od liczby meczów rozegranych przez drużynę. Średnio drużyny, które rozegrały więcej meczów mają więcej punktów. Należy, więc znaleźć wartość, która niezależnie od liczby rozegranych przez drużynę meczów, pozwoli nam określić, jaki wynik końcowy osiągnie drużyna.

Potencjalne wartości, jakie możemy więc rozpatrywać to: średnia punktu na mecz oraz liczba zdobytych punktów na mecz do sumy wszystkich punktów zdobytych przez wszystkie drużyny z ligi.




Powyższy wykres prezentuje, że liczba meczów nie ma wpływu na średni wynik drużyn, jeśli będzie rozpatrywać średnią liczbę punktów na mecz, czyli liczba punktów zdobytych w sezonie podzielona przez liczbę rozegranych meczów.

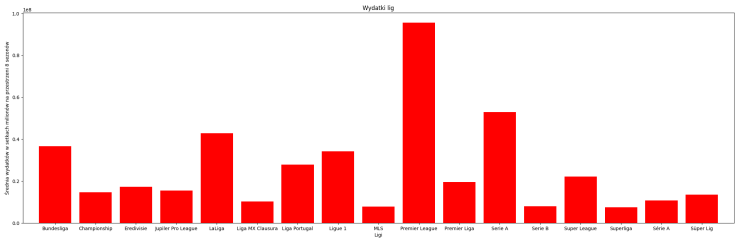



Podobnie jak w powyższym przykładzie, widzimy że stosunek liczby punktów zdobytych przez drużynę do liczby punktów zdobytych przez wszystkie drużyny w lidzę nie zależy od liczby rozegranych meczy.

Na podstawie dwóch powyższych wykresów możemy przyjąć, że końcowy wynik w lidzę możemy użyć któregokolwiek z powyższych atrybutów. Jednak bardziej intuicyjną i łatwiejszą do interpretacji wartością jest liczba punktów na mecz, więc w modelu spróbujemy estymować tą wartość w zależności od pozostałych atrybutów.

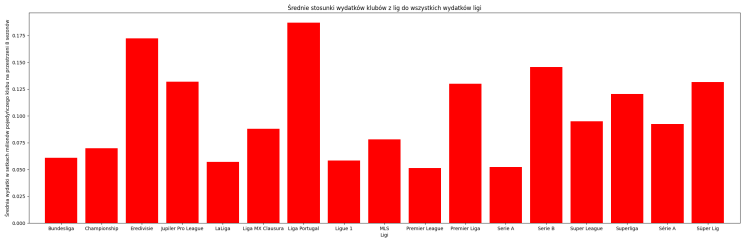
## 4.2 Dane dotyczące wydatków

Dane dotyczące wydatków są bardzo uzależnione od bogactwa ligi, stąd to czy klub wydał dużo czy mało, można jedynie rozpatrzyć, obserwując inne kluby z ligi:

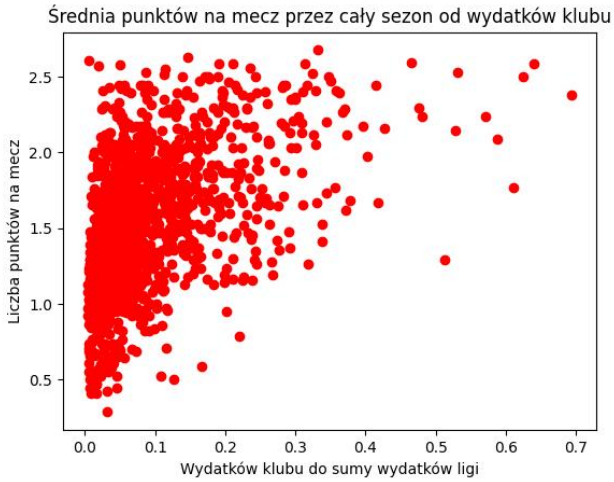



Powyższe wykresy prezentują średnie wydatki klubu danej ligi, przy czym drugi pokazuje średnie wydatki względem ligi.

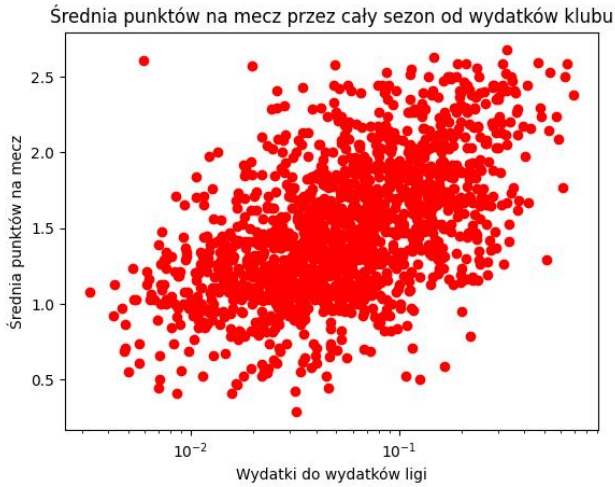
Na pierwszym wykresie widzimy że dla różnych lig średnie wydatki klubu są bardzo rozbieżne. Znacznie mniejsze rozbieżności występują jeśli rozpatrzymy



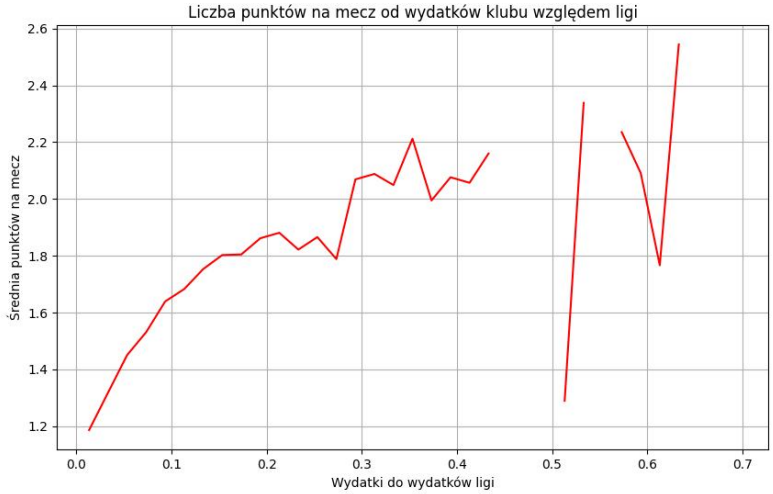

średnią wydatków klubu podzieloną przez sumę wydatków wszystkich klubów w lidzę. Dlatego dla modelu będziemy wykorzystywać wartości z kolumny "Wydatki do wydatków ligi", ponieważ jest on znacznie mniej uzależniony od ligi.




Na powyższym wykresie możemy zobaczyć jak średnia punktów za mecz zależy od wartości wydatków. Możemy zaobserwować korelację, że wraz ze wzrostem wydatków względem ligi, rośnie średnia liczba punktów na mecz. Zwróćmy jednak uwagę na duże zagęszczenie punktów dla wydatków względem ligi należących dla przedziału między 0, a 0.2. Aby przyjrzeć się tym wartościom rozpatrzmy powyższy wykres w skali logarytmicznej.




Na podstawie powyższego wykresu widzimy praktycznie liniowy wzrost średniej liczby punktów od wydatków względem ligi. Ponieważ wykres jest w skali logarytmicznej, możemy zakładać, że zależność ta jest zależnością logarytmiczną.

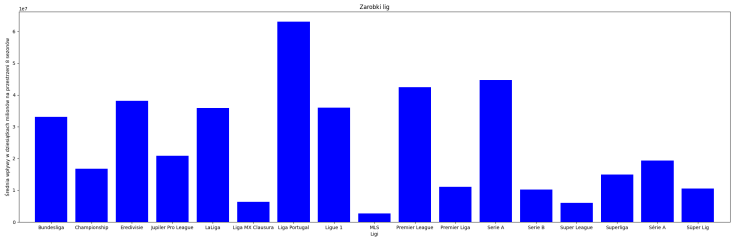



Jeśli podzielimy argumenty na tzw. kubełki o wielkości 0.02 i dla każdego kubełka wyliczymy średnią, wykres plot utwierdzi nas w przekonaniu o zależności logarytmicznej między liczbą punktów na mecz a wydatkami względem ligi. Wprawdzie, wartości powyżej 0.5 wydatku względem ligi, może pokazywać załamanie tego trendu, należy jednak pamiętać, że są to wartości skrajne (ponieważ mało który klub odpowiada za ponad połowę wydatków ligi) i przez

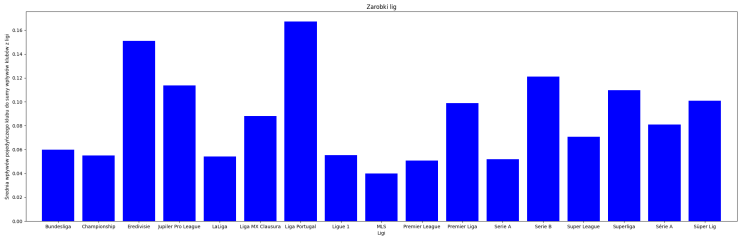
brak wystarczającej liczby danych nie należy traktować ich jako trendu.

## 4.3 Dane dotyczące wpływów

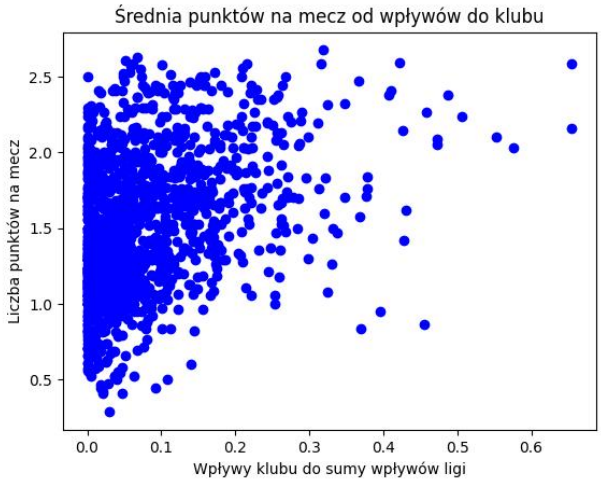
Wpływy, podobnie jak wydatki, są bardzo uzależnione od bogactwa ligi. Dlatego należy odpowiedzieć, czy też w modelu nie korzystniej byłoby korzystać z danej uwzględnionej o dane dotyczące wpływów ligi.



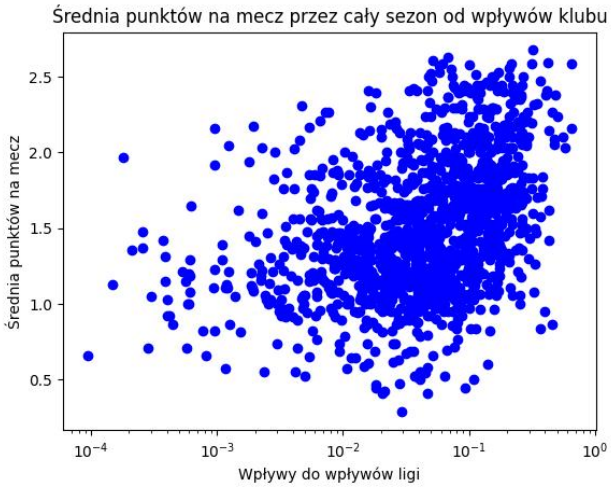

Powyższy wykres słupkowy (prezentujący średnie wpływy klubu z danej ligi) jest znacznie mniej zróżnicowany niż jego odpowiednik dla wydatków, jednak wciąż możemy zaobserwować sporą rozbieżność średnich wpływów do klubu dla różnych lig.




Jeśli zaprezentujemy średnie wpływy klubu podzielone przez wpływy całej ligi, możemy zwrócić uwagę że wartości na wykresie są bliżej siebie, co również utwierdza w przekonaniu, że w przypadku budowy modelu to właśnie wartości z kolumny “Wpływy względem wpływów ligi” powinny być brane pod uwagę.




Wykres typu scatter prezentuje wyraźną korelację. Wraz ze wzrostem wpływów klubu względem wpływów ligi rośnie liczba punktów na mecz. Podobnie, jak w przypadku wydatków duże zagęszczenie danych dla wpływów względem ligi mniejszych niż 0.2, dlatego rozpatrzmy ten wykres w skali logarytmicznej.

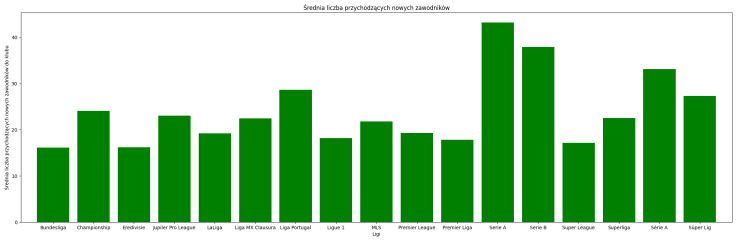



W tym wykresie widzimy znaczny wzrost między 0.01 a 0.1 wpływów względem ligi. Nie jesteśmy już w stanie jednoznacznie określić typu funkcji, ale widzimy, że wzrost jest szybszy niż w przypadku wydatków.

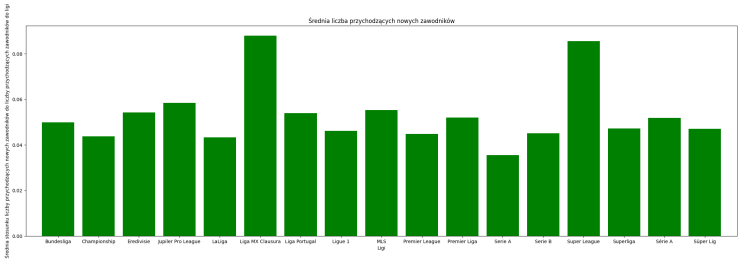

Wykres typu plot, stworzony jako średnie wartości dla kubełków o wielkości 0.2, sugeruje że funkcja rośnie niemal liniowo. Podobnie jak w przypadku wydatków, zakrzywienia krzywej dla wpływów względem ligi większych 0.35 są spowodowane małą liczbą danych do wpływów z względem ligi większych niż 0.35.

## 4.4 Dane dotyczące zawodników przychodzących

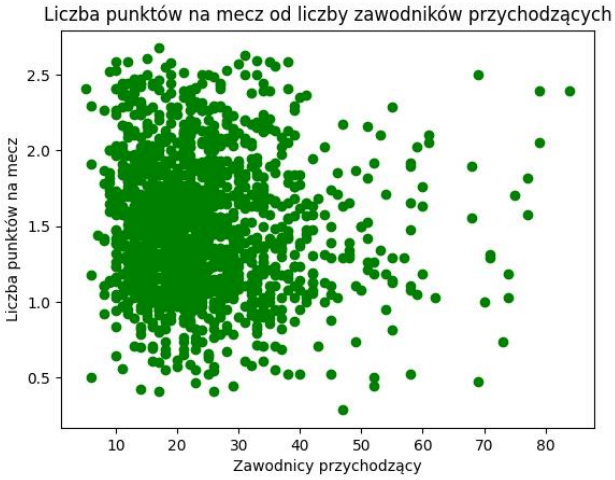
Możemy się spodziewać że dane dotyczące liczby zawodników przychodzących nie będą wymagały uwzględnienia w skali ligi, ponieważ w większości krajów kadry piłkarskie mogą mieć określoną liczbę zarejestrowaną zawodników. Rozpatrzmy jednak czy aby na pewno.



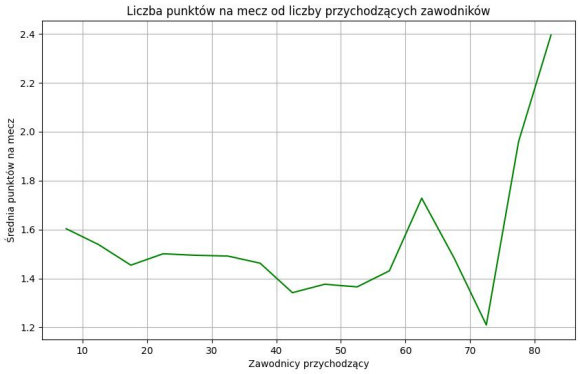

Jak widzimy powyższy wykres średniej liczby zawodników przychodzących na sezon dla danej ligi są zgodnie z oczekiwaniami w miarę zbliżone do siebie.




Uwzględnienie liczby przychodzących zawodników do wszystkich klubów ligi niewiele zmieniło. Możemy więc rozpatrywać dane zwykłej liczby przychodzących zawodników.



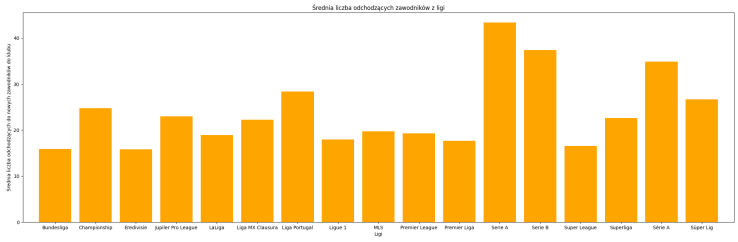

Wykres typu scatter nie pokazuje nam żadnej korelacji między liczbą przychodzących zawodników do klubu, a zdobytą liczbą punktów na mecz. Możemy więc przypuszczać, że dane te są niezależne.

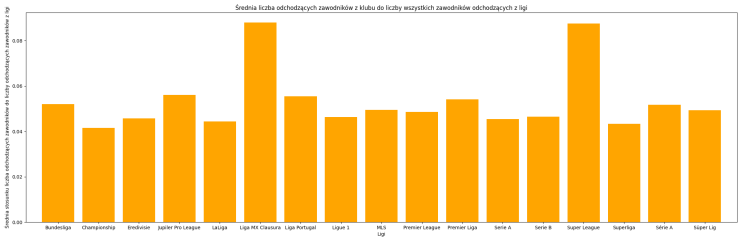



Wykres typu plot, średniej liczby punktów na mecz dla kubełków o wielkości 5, pokazuje że bez względu na zmianę liczbę zawodników przychodzących wartość liczby punktów na mecz się nie zmienia. Ponownie skrajne wartości trochę przeczą wyciągniętym wnioskom, ale trzeba pamiętać, że za zbyt mała liczba klubów w ciągu jednego sezonu ściąga ponad 60 zawodników, by wyciągnąć z tych wartości miarodajne wnioski.

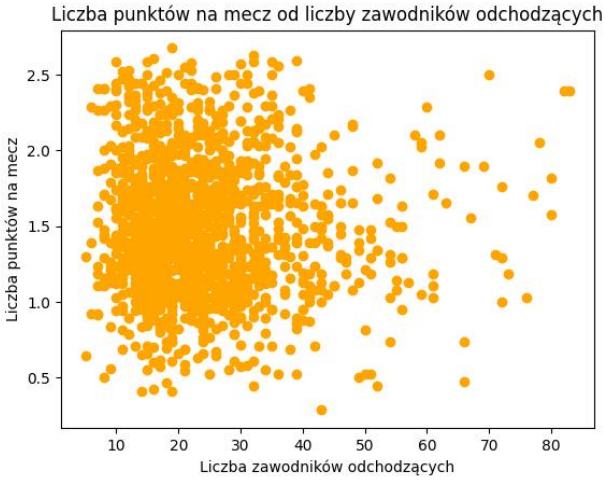
## 4.5 Dane dotyczące zawodników odchodzących

Podobnie jak w przypadku liczby zawodników przychodzących, rodzaj ligi nie powinien mieć wpływ na liczbę zawodników odchodzących z klubu. Rozpatrzmy dla tego wykres średniej liczby odchodzących zawodników dla danej ligi:

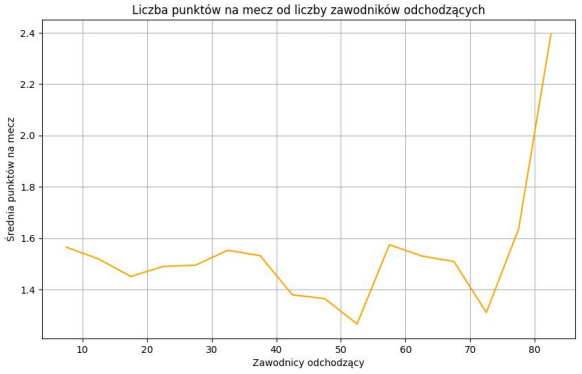


Jak widzimy uwzględnienie danych ligi nie wpływa w sposób znaczący na odchylenie standardowe, dlatego możemy rozpatrywać liczbę odchodzących bez uwzględniania wartości ligi.



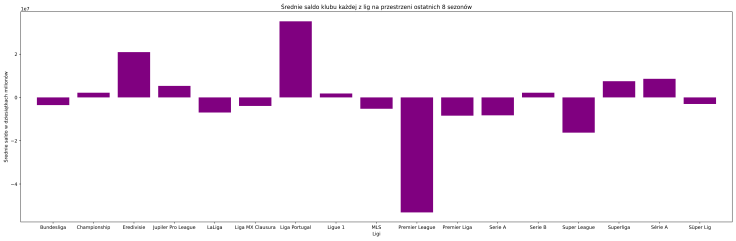

Wykres typu scatter nie pozwala nam wyciągnąć wniosków dotyczących zależności między liczbą punktów na mecz, a liczbą zawodników odchodzących. Sugeruje nam to, że może nie występować korelacja między tymi atrybutami.



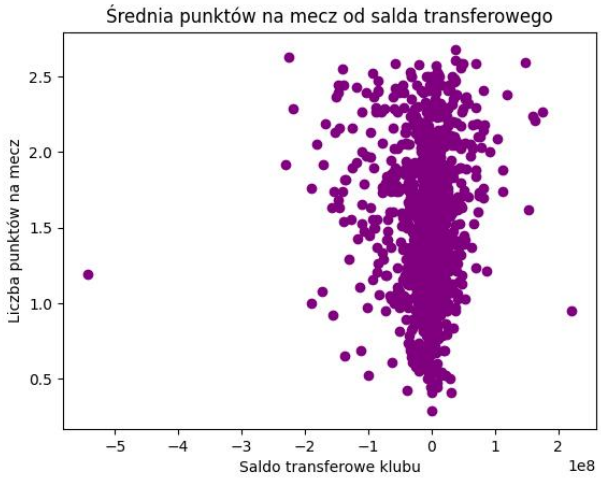

Wykres typu plot, dla kubełków o wielkości 5, potwierdza wnioski wyciągnięte z wykresu typu scatter dla liczby zawodników odchodzących. Możemy przyjąć, że nie istnieje korelacja między liczbą punktów na mecz, a liczbą zawodników odchodzących.

## 4.6 Dane dotyczące salda transferowego

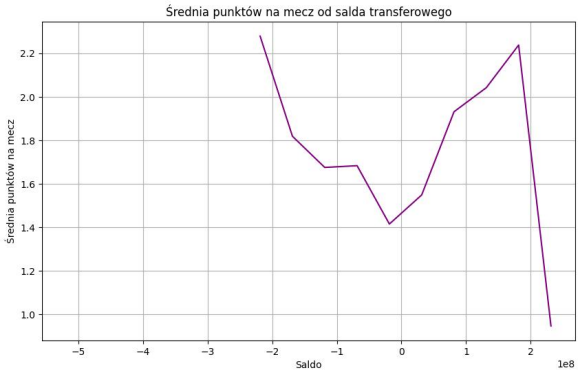
Saldo transferowe rozumiane, jako różnica wpływów do klubu i wydatków klubu, jest problematyczne do zapisu. Możemy oczekiwać, że różnice między ligami będą wyraźne. Natomiast, przez to że saldo może być zarówno dodatnie jak i ujemne nie mamy prostego mechanizmu do uwzględnienia wartości salda ligi.




Powyższy wykres, średnich wartości salda klubu danej ligi pokazuje, że odchylenie standardowe jest dość duże i zależne od ligi, z której pochodzi klub.



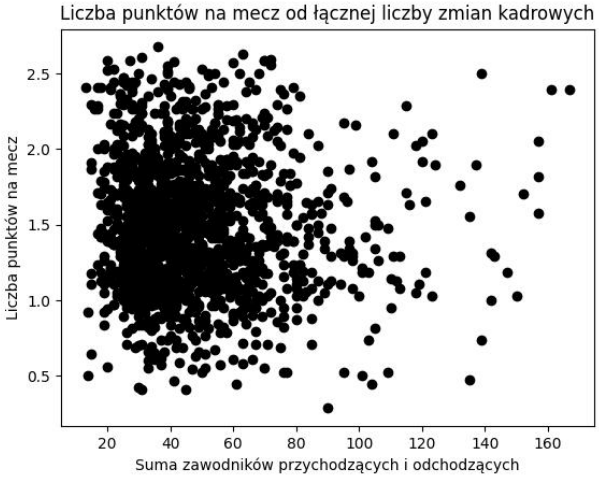

W przeciwieństwie do poprzednich danych dla salda, wykres typu scatter nie okazuje się łatwy do interpretacji, przez duże zagęszczenie wartości.



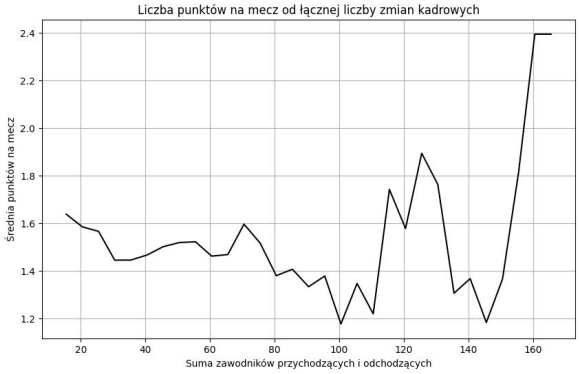

Na podstawie wykresu plot, o kubełku równym pięćdziesiąt milionów, możemy potwierdzić wnioski, wyciągnięte z wykresu typu scatter. Średnio kluby, których saldo jest bliskie zeru osiągają gorsze wyniki niż te których salda są większe lub mniejsze. Tak, jak w przypadku wcześniejszych wykresów typu plot, należy nie brać pod uwagę skrajnych wartości (tutaj dla salda powyżej 200 milionów) przez zbyt małą liczbę klubów o takich saldach.

## 4.7 Dane dotyczące sumy liczb zawodników przychodzących i odchodzących

Z powodu braku korelacji między liczbą punktów na mecz od liczby zawodników przychodzących i liczbą zawodników odchodzących, możemy wnioskować że nie będzie istniała korelacja między liczbą punktów na mecz, a sumą tych dwóch wartości.



Zgodnie z oczekiwaniami z wykresu typu scatter, nie możemy wyciągnąć wniosków o korelacji między średnią liczbą punktów na mecz, a sumą zawodników przychodzących i odchodzących do klubu.




Wykres typu plot, dla kubełku równym 5, potwierdza wnioski i nie prezentuję korelacji między oboma rozpatrywanymi atrybutami. Tak jak we wcześniejszych przykładach nie należy wyciągać wniosku dla łącznej liczby zmiany zawodników większej niż 100 z powodu braku wystarczającej liczby klubów, które w ciągu sezonu sprowadziły i oddały tak dużą liczbę zawodników.

# 5 Przygotowanie modelu

## 5.1 Podzielenie danych

Aby móc faktycznie wytrenować i sprawdzić skuteczność modelu, dane zostały podzielone. Siedemdziesiąt pięć procent losowo wybranych danych zostało wybrane z głównej tabeli i posłużyły jako dane treningowe dla modelu, natomiast pozostałe pełniły rolę danych testowych. Tak rozdzielone dane zostały jeszcze znormalizowane z użyciem obiektu StandardScaler z biblioteki sklearn.

## 5.2 Wytrenowanie modelu

W analizie danych ustaliliśmy, że część danych nie ma wpływu na średnią liczbę punktów drużyny na mecz, dlatego nie zostały uwzględnione w modelu. Innym przypadkiem jest saldo transferowe. Po wstępnej analizie danych widać, że istnieje pewna korelacja między nim a wynikiem końcowym klubów, jednak dana ta w dużym stopniu jest uzależniona ligi lub może wynikać z innych powodów niż przewidujemy. Nie wiemy, czy uwzględnienie jej poprawi wynik modelu, czy też pogorszy, stąd decyzja o wytrenowaniu modeli dla dwóch przypadków. Jednego, którego argumentami będą stosunek wydatków klubu względem ligi oraz wpływ

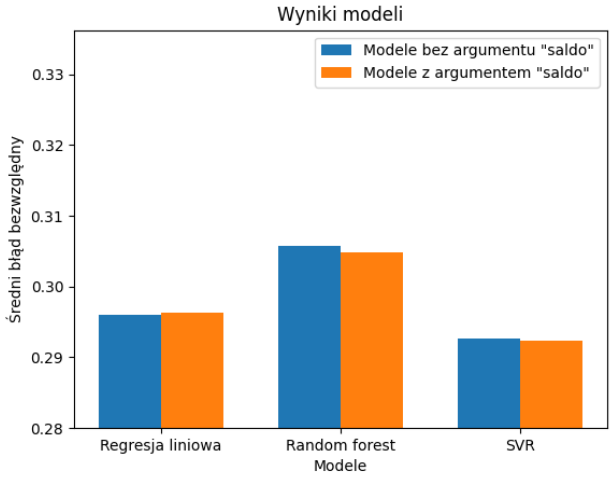
klubu względem ligi. Natomiast drugi poza dwoma wymienionym, będzie traktował jako argument również saldo transferowe klubu. Do wyznaczenia krzywej regresji zostały wybrane następujące modele:

- regresja liniowa - zależność między przynajmniej jedną zmienną niezależną, a zmienną zależną są liniowe, co pozwala nam sądzić, że wynik regresji liniowej będzie bardzo dobry.
- random forest - model oparty na kombinacji wielu drzew decyzyjnych, może lepiej sprawdzić się w przypadku argumentów, których zależności nie są liniowe, więc las losowy może być najlepszym modelem, sposród tych wykorzystujących atrybut "saldo".
- svr - może łączyć zalety obu modeli i dobrze dopasowywać się zarówno do zależnościach między zmiennymi liniowymi i nieliniowymi.

Jako wynik modelu zostanie wykorzystany średni błąd bezwzględny (ang. mean absolute error) co pozwoli łatwo zinterpretować wyniki modeli. W tym przypadku będzie on oznaczał, o ile średnio myli się model estymujący średnią punktów na mecz w ciągu całego sezonu.

## 5.3 Wyniki modeli

Wytrenowane modele otrzymały następujące wyniki:



| | Średni bezwzględny błąd | |
|---|---|---|
| Model | bez argumentu "saldo transferowe" | z argumentem "saldo transferowe" |
| Liniowa regresja | 0.296 | 0.296 |
| Random forest | 0.306 | 0.305 |
| SVR | 0.293 | 0.292 |

Modele osiągnęły podobne wyniki, jednak spośród wszystkich to svr dla argumentów z saldem transferowym osiągnął najmniejszy średni błąd bezwzględny. Należy jednak zwrócić uwagę, że różnice między modelami uwzględniającymi saldo transferowe, a tymi które nie uwzględniają są minimalne, co potwierdza obawy o sensowność używania tego atrybutu w modelu. Prognozy, mówiące że różnice między modelami typu random forest będą najbardziej widoczne okazały się trafne, jednak różnica ta okazała się niemal symboliczna.

### 5.4 Określenie jakości modelu

Dla wszystkich modeli w przybliżeniu średni bezwzględny błąd wynosi 0.3. Dla lig, w których w sezonie rozgrywa się trzydzieści osiem kolejek (czyli w najpopularniejszym formacie rozgrywek), oznacza to że model myliłby się średnio o trochę mniej niż 12 punktów, jeśli rozpatrujemy końcową liczbę punktów klubu w lidzę. Oznacza to że za pomocą modelu nie bylibyśmy precyzyjnie wskazać miejsca które zajmie drużyna w lidzę, ale możliwe by było określenie o jaką część tabeli klub może walczyć np. europejskie puchary lub walka o utrzymanie.

## 6 Wnioski

Zebrany zbiór danych pozwala na estymację końcowego wyniku w lidze. Możemy więc potwierdzić tezę, że zachowanie klubu na rynku transferowym ma wyraźny wpływ na końcową liczbę zdobytych punktów. Zwróćmy uwagę, że rozpatrywane dane były dosyć ubogie. Rozpatrywane zostały przede wszystkim kwoty wydatków oraz wpływów względem innych klubów w lidzę. W zbiorze danych nie było żadnego atrybutu, który pozwoliłby określić "jakości transferów", np. różnica między kwotą piłkarza zapłaconą a szacowaną, albo jakość dyrektora sportowego, która pozwoliłaby oszacować, jak często klub dokonuje udanych lub nieudanych transferów. Pokazuje to jak istotna w sukcesie drużyny piłkarskiej jest sytuacja finansowa klubu. Warto również zwrócić uwagę na brak korelacji między liczbą zawodników przychodzących, odchodzących i ich sumą, a końcowym wynikiem. Intuicyjne byłoby przyjęcie, że duża liczba zmian kadrowych wpłynie negatywnie na postawę drużyny. Jednak jak się okazuje nie ma to zazwyczaj większego wpływu na wynik klubu w lidzę.

W przyszłości można by rozbudować model o dodatkowe informację, jak wspomnianą różnicę między kwotą zapłaconą za piłkarza a szacowaną wartością rynkową, a także zacząć uwzględniać wyniki w klubu w poprzednim sezonie co pozwoliłoby lepiej ukazywać przyjęty wycinek rzeczywistości np. klub zajmujący

pierwsze miejsce w tabeli w poprzednim sezonie, nawet nie dokonując transferów prawdopodobnie w kolejnym sezonie nadal bedzie w czołówce ligi, jeśli chodzi o końcowy wynik. Na podstawie zdobytych danych możliwa byłaby również budowa modelu klasyfikacji, który przewidywałby na podstawie zachowania klubu na rynku transferowym, jaki rejon tabeli końcowej zajmie klub.

# Źródła


[1] Transfermarkt. *Saldo transferowe*. URL: `https://www.transfermarkt.pl/statistik/transfersalden`. Accessed: 15.05.2023.

[2] Wikipedia. URL: `https://pl.wikipedia.org/wiki/Wikipedia:Strona_g%C5%82%C3%B3wna`. Accessed: 15.05.2023.