|  |
| --- |
|  |
| Raport ze statystyki stosowanej |
| Porównanie skuteczności dryblingów i podań progresywnych wśród zawodników TOP5 lig europejskich w sezonie 2023/2024 |
|  |
| **Antoni Krzak i Wojciech Haładewicz** |
| **Grupa w środy o 15** |

|  |
| --- |
|  |

1. Wstęp

Celem raportu jest przeprowadzenie analizy danych rzeczywistych za pomocą statystyki opisowej. My postanowiliśmy połączyć pracę akademicką z naszymi zainteresowaniami, biorąc pod lupę statystyki piłkarzy nożnych, grających w TOP 5 ligach europejskich (najwyższych szczeblach rozgrywek ligowych w 5 krajach: Anglia, Hiszpania, Włochy, Niemcy, Francja), mianowicie procent udanych dryblingów oraz liczbę podań progresywnych na przestrzeni sezonu 2023/2024, obecnie trwającego, od 08.2023 do stanu na dzień 05.05.2024. Dane zostały pozyskane ze strony FBref, witryny internetowej wyposażonej w bogaty zasób danych dotyczących zmagań na murawie w wielu krajach, na różnych poziomach rozgrywkowych. Link do odpowiednich podstron strony głównej, z których dane są uzyskiwane (pierwszy link - dryblingi, drugi - podania progresywne):

* <https://fbref.com/en/comps/Big5/possession/players/Big-5-European-Leagues-Stats>,
* <https://fbref.com/en/comps/Big5/stats/players/Big-5-European-Leagues-Stats>.

Analiza została przeprowadzona na próbce złożonej z danych dla 532 różnych piłkarzy. Podjęliśmy taką decyzję, gdyż dostępne materiały dotyczą ponad 2500, jednakże, jako że wybrane do analizy dziedziny są domenami zawodników usposobionych ofensywnie, branie pod uwagę chociażby bramkarzy lub środkowych obrońców jedynie utrudniłoby interpretację wyników, dlatego testową próbkę określiliśmy poprzez wybranie do niej wszystkich, którzy dokonali 36 lub więcej prób dryblingu. Tym sposobem uzyskujemy graczy, w których zachowaniu dostrzec można już jakiś trend, czy to skuteczność, czy nieskuteczność owych prób.

Do sporządzenia raportu wykorzystaliśmy możliwości programu Python, a dokładniej bibliotek: BeautifulSoup do przeniesienia danych do programu; numpy oraz scipy do obliczenia wartości cech statystycznych, a także seaborn oraz matplotlib do wygenerowania odpowiednich wykresów.

*Ważne adnotacje*

Warto też na wstępie odnotować, że pomimo wzięcia danych z tego samego okresu dla statystyk analizowanych sportowców, każdy z nich mógł przebywać na boisku inną liczbę minut, a w konsekwencji, mieć różne szanse na poprawienie swoich liczb. Niemniej w tym raporcie skupiamy się na samej ich jakości, a nie ilości, albowiem klasowy zawodnik obroni się jakością bez względu na czas gry.

Za istotny należy uznać również fakt, iż niektórzy zawodnicy rozegrali spotkania w więcej niż jednej ekipie na przestrzeni tego sezonu (na przykład poprzez transfer w zimowym okienku). Z tego powodu oraz, jako że reprezentują oni także swoją ligę, w której grają, w statystykach, przyporządkowujemy ich jako zawodników rozgrywek, w których rozegrali więcej minut na przestrzeni całego sezonu.

1. Przedstawienie danych

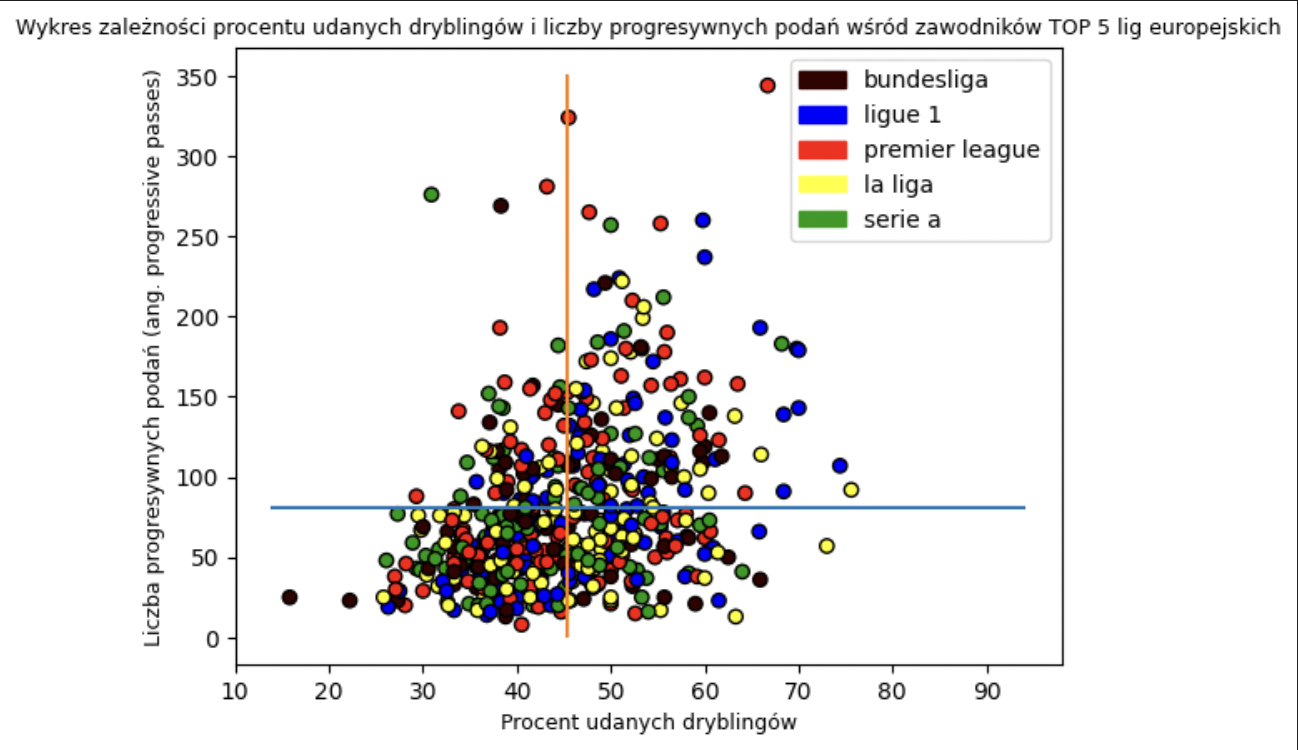
# Słownik

Zanim przejdziemy do samej analizy postanowiliśmy wyjaśnić kilka terminów ze świata piłkarskiego, którymi będziemy się w raporcie posługiwać, a są one niezbędne do jego zrozumienia.

* ***Drybling*** - określenie na bieg oraz wszelkie zmiany jego kierunku, jak również szybkie zwroty ciała z piłką przy nodze dokonywane przez zawodnika.
* ***Skuteczność dryblingu*** - stosunek udanych prób dryblingu do wszystkich podejmowanych prób przez jednego zawodnika, wyrażony w procentach.
* ***Progresywne podanie*** - (definicja z samej strony z danymi) udane, czyli takie które doszło do jednego z partnerów z drużyny, podanie, które albo przesuwa akcję drużyny przy piłce o co najmniej 10 jardów (około 9 metrów) w stronę bramki przeciwnika w porównaniu z najdalszą pozycją piłki w przeciągu ostatnich 6 podań, albo każde podanie, które znajduje adresata w polu karnym przeciwnika. Dodatkowo nie kwalifikują się do tej statystyki podania dokonywane spod własnej bramki oraz z okolic defensywnej tercji boiska (około 40% całej powierzchni murawy, licząc od bramki swojej drużyny). W dużym uproszczeniu, podania, które przyspieszają, dynamizują i intensyfikują akcję drużyny, mogące stworzyć też bezpośrednie zagrożenie pod bramką przeciwnika.
* ***TOP 5 lig i ich nazwy*** - określenie na przyjęte umownie ligi z najwyższym poziomem na świecie. Podamy ich nazwy i państwa, w którym rozgrywki mają miejsce, aby intuicyjnie wiedzieć, z jakim regionem je kojarzyć:
  + - * + Premier League – Anglia,
        + La Liga – Hiszpania,
        + Serie A – Włochy,
        + Bundesliga – Niemcy,
        + Ligue 1 – Francja.
  + ***Pozycje zawodników*** - po krótce omówimy też niektóre pozycje na boisku, dla uzmysłowienia, z jakim typem zawodnika mamy do czynienia:
* wahadłowy - obrońca grający przy linii bocznej, często wychodzący do akcji ofensywnych,
* boczny obrońca - obrońca grający przy linii bocznej, skupiony na grze defensywnej
* skrzydłowy - ofensywny zawodnik grający przy linii bocznej,
* środkowy pomocnik (pivot) - zawodnik grający w środku pola, zrównoważony w grze ofensywnej i defensywnej,
* napastnik (snajper) - zawodnik grający w linii ataku, skupiony na wykorzystywaniu i kreowaniu sytuacji.

# Wykres zależności dwóch statystyk

Znając pojęcia statystyk, których zależność będziemy omawiać, możemy wyjaśnić, dlaczego akurat te rubryki tabel wzięliśmy pod uwagę. Zarówno robienie przewagi w ataku grą z piłką (drybling), jak i „pozbywanie się" jej dla dobra drużyny, mówiąc mniej brutalnie przekazywanie jej (progresywne podania), cechują zawodników kreatywnych, a więc takich, którzy mogą zrobić największą różnicę w meczu i przyczyniają się najbardziej do szans kreowanych przez swoją ekipę. Przedstawienie porównania piłkarzy w ramach tych dwóch statystyk na tle całej reszty pozwala na określenie zarówno słabych i mocnych ogniw w poszczególnych ekipach czy też ligach oraz ocenie rokowań na przyszłość, uwzględniając zebrane dotychczas liczby, pozwalające na korektę treningów całego zespołu lub udzielenie indywidualnych wskazówek konkretnym graczom.

Wyniki zależności procentu udanych dryblingów i liczby progresywnych podań wśród zawodników TOP 5 lig europejskich dla wziętej próbki znajdują się poniżej.

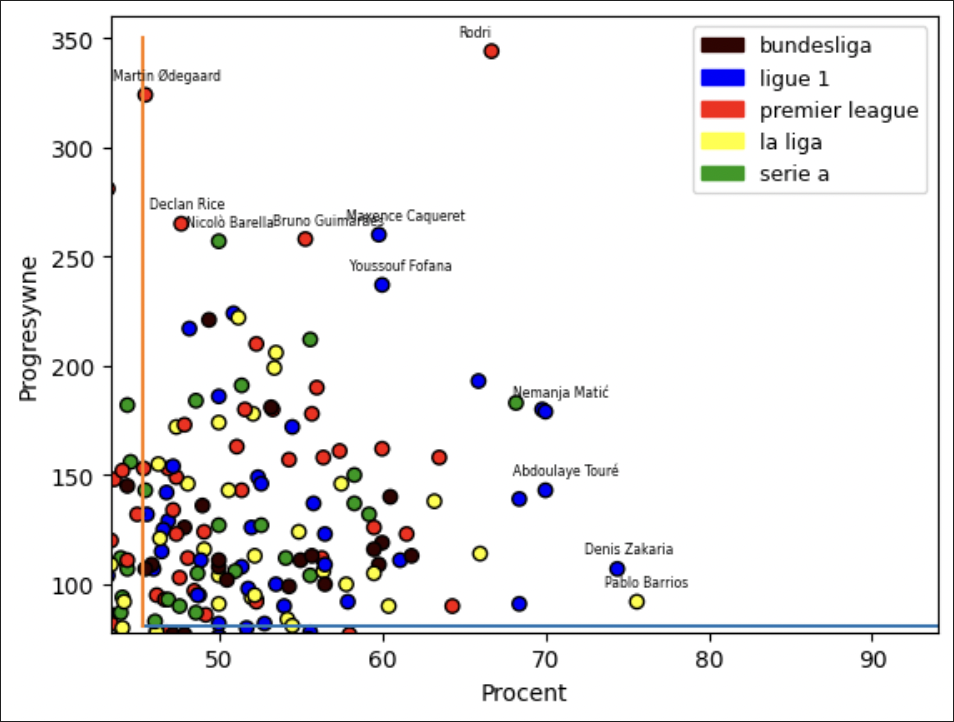
Na wykresie kropki reprezentują danego zawodnika, a ich współrzędne - osiągnięte przez niego statystyki, odpowiednio na osi X skuteczność dryblingów, a na osi Y liczba progresywnych podań. Każda kropka jest również opatrzona odpowiednim kolorem, odzwierciedlającym ligę, w której gra dany zawodnik. Przynależność zawodnika do ligi możemy stwierdzić studiując legendę.

Na wykresie zostały również umieszczone dwie proste przecinające się. Są to linie średnich arytmetycznych: niebieska to średnia liczba podań progresywnych, pomarańczowa - średni procent skutecznych dryblingów, wśród analizowanych zawodników. Zostały one naniesione, gdyż pozwalają na głębsze spojrzenie od ogółu do szczegółu, dzieląc wykres na 4 nierówne ćwiartki, obrazujące konkretne nacechowanie zawodników w odniesieniu do uzyskiwanych liczb wobec średniej. Miejsce przecięcia prostych wyznacza zawodnika posiadającego najbardziej przeciętne, niewyróżniające się spoza ogółu rezultaty.

Z samego wykresu dostrzec można duże rozproszenie danych im bliżej 4 rogów wykresu oraz ogromne zagęszczenie w okolicach przecięcia prostych. Jest to zrozumiałe, gdyż piłkarzom nie jest łatwo osiągnąć wartości skrajne i wykraczające poza średnią, a większość z nich prezentuje kolokwialnie mówiąc umiejętności „wymagane" na tym poziomie profesjonalnej piłki, określane właśnie przez przeciętne ich wartości. Warto też odnotować, że nie ma zauważalnych na pierwszy rzut oka żadnych znaczących trendów w kontekście podziału na ligi, albowiem we wszystkich ćwiartkach znajdują się reprezentanci każdej z rozgrywek. Udowadnia to zróżnicowanie piłkarzy na szczeblu ligowym - każdy zestaw cech znajdzie dla siebie miejsce. Głębsza analiza ukazuje, którzy z nich powinni być nagradzani za owe miejsce, a którzy „karani”. Za godne uwagi uznać można także niewielką liczbę zawodników z bardzo wysoką skutecznością dryblingów, nieprzekraczającą u nikogo nawet 80%. Dzieję się tak, gdyż drybling jest jednym z najtrudniejszych elementów gry do opanowania w futbolu. Praktykowanie go regularnie na najwyższym poziomie zasługuje na uznanie, natomiast tylko nieliczni potrafią być dodatkowo nieuchwytni dla rywali w większości sytuacji. Progresywne podania wiążą się często z bardziej strategicznym podejściem do sytuacji, dlatego skuteczność w tej dziedzinie może być wypracowana dzięki licznym treningom oraz doświadczeniu nabytemu w prawdziwych meczach.

Możemy teraz przejść do omówienia „demografii” poszczególnych ćwiartek, pozwalających na bardziej szczegółową analizę danych.

# Ćwiartka 1 – dobry drybler, dobry asystent



Najpierw omówimy część wykresu, która analityków klubowych do spraw rekrutacji interesuje najbardziej. Jej reprezentanci prezentują ponadprzeciętny poziom w obu omawianych statystykach, obrazujący ich wpływ na grę całej drużyny oraz wartość indywidualną. Zawodnikiem najbardziej odosobnionym wydaje się tutaj Rodri, zawodnik Manchesteru City w Anglii. Uzyskał on w tym sezonie najwyższy współczynnik liczby podań progresywnych spośród omawianej próbki, prezentując ponadto nienajgorsze predyspozycje dryblerskie. Wiedząc, iż 'The Citizens' (przydomek klubu) to jedna z najdroższych i najbardziej utalentowanych ekip na całym świecie, a sam Hiszpan był wymieniany w gronie faworytów do zdobycia Złotej Piłki za poprzedni sezon, dobitnie ilustruje klasę i znaczenie tego zawodnika oraz przypomina trenerom, aby pozostał w klubie za wszelką cenę. Z kolei żółta kropka sygnowana nazwiskiem Pablo Barriosa naświetla go jako gracza świetnie dryblującego, jednakże nienotującego aż tylu podań progresywnych. Taki stan rzeczy, wobec młodego wieku Hiszpana, pozwala sztabowi trenerskiemu rozważyć, czy nie warto poświęcić więcej uwagi temu drugiemu aspektowi, aby stał się on kompletniejszym zawodnikiem, chociażby na wzór wyżej wspomnianego rodaka. Z kolei Martin Ødegaard, zauważalny w lewym górnym rogu, scharakteryzowany może zostać jako znakomity asystent w aspekcie progresywnych podań, ale jego skuteczność dryblingu oscyluje raczej w granicach średniej wartości. Pozwala to na wysnucie wniosków, jakoby piłkarz ten wspaniale nadawał się do gry, gdy ma piłkę i dyryguje całym atakiem, jednakże był niezbyt użyteczny w roli robiącego różnice brawurowymi próbami przejścia rywali.

Dalsze obserwacje ukazują fakt, że większość z zawodników podpisanych występuje na boisku jako środkowy pomocnik lub w roli bardzo do niej zbliżonej. Można to wyjaśnić tym, że to właśnie pivoci są określani mianem najbardziej kreatywnych zawodników na murawie, niejednokrotnie będąc motorem napędowym swojej ekipy. Nietrudno stwierdzić również, że w tej części nie ma wyraźnego lidera, jeśli chodzi o ligę z największą liczbą reprezentantów. Świadczy to o podobnej charakterystyce tych 5 rozgrywek w kontekście kreatywności tych najlepszych piłkarzy.

# Ćwiartka 2 – słaby drybler, dobry asystent

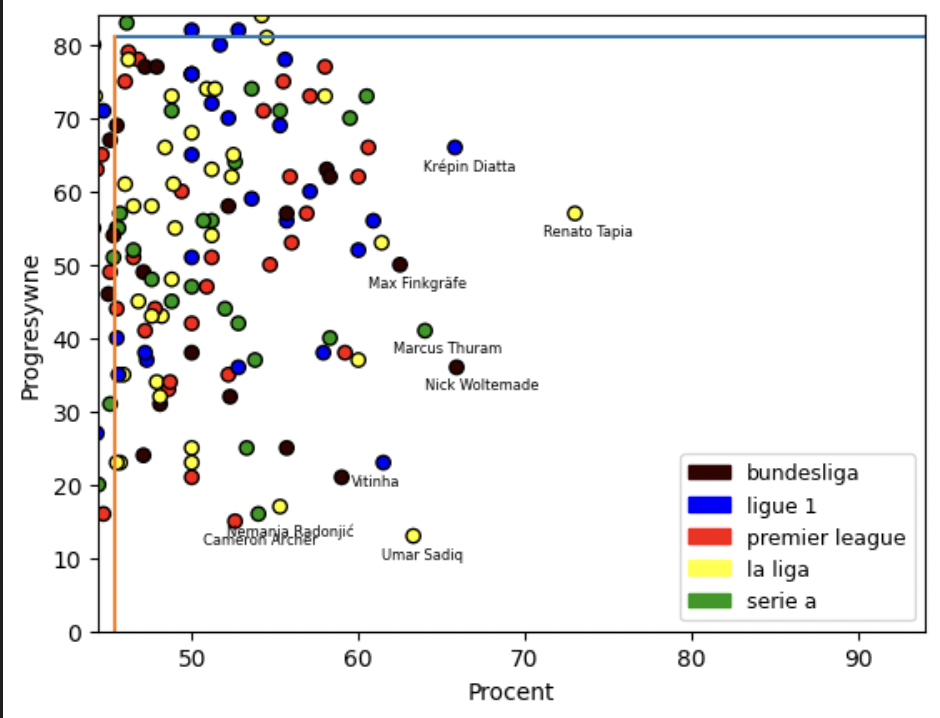
**Obraz zawierający zrzut ekranu, tekst, oprogramowanie, diagram

Opis wygenerowany automatycznie**

Skupmy się teraz na lewej górnej ćwiartce. Jej przedstawicieli określić można mianem świetnych asystentów, co kompensuje w pewnym stopniu braki w skuteczności ich prób dryblingu. Giovanni Di Lorenzo to zawodnik, który od razu rzuca się w oczy jako sztandarowy tego przykład. Osiąga on ponadprzeciętne liczby w przypadku podań progresywnych, jednak zrobienie przewagi szybką akcją z piłką przy nodze to nie jego domena. Z kolei Bruno Fernandesa z Manchesteru United można określić mianem skrzywdzonego, gdyż został pominięty przy omawianiu poprzedniej ćwiartki, pomimo znakomitych statystyk, co spowodowane jest minimalnie niższą skuteczność dryblingów niż wynosi średnia. Z drugiej strony Polak Matty Cash balansuje na granicy lewych ćwiartek, sugerując stosunkowo niewielkie, ale wystarczające liczby, aby jego występy były uznane za akceptowalne przez rekruterów.

Przechodząc do obserwacji ogólniejszej nie sposób nie odnieść wrażenia, że dominują tutaj reprezentanci Premier League. Są zarówno skumulowani przy środku wykresu, jak i rozproszeni, im dalej od niego. Sugerować to może promowanie w tej lidze kreatywności uzyskiwanej poprzez odpowiednie reagowanie na sytuacje boiskowe, polegające raczej na dokonywaniu trafnych decyzji i obsługiwaniu partnerów, niż przysłowiowym „kiwaniu się". Jeśli chodzi o charakterystykę samych zawodników zauważamy, że są to dalej w większości środkowi pomocnicy (choć tym razem ewidentnie usposobieni ofensywnie) tacy jak Bruno Fernandes, James Maddison czy Weston McKennie, lecz znajdują się wśród nich także ich rywale występujący na bokach. Mohamed Salah to przykład skrzydłowego, a Matty Cash i Giovanni Di Lorenzo - wahadłowych. Daje to do zrozumienia, że oczekiwanie tego typu kreatywności wobec innych pozycji niż tylko od liderów środka pola jest jak najbardziej na miejscu na poziomie TOP 5 lig europejskich.

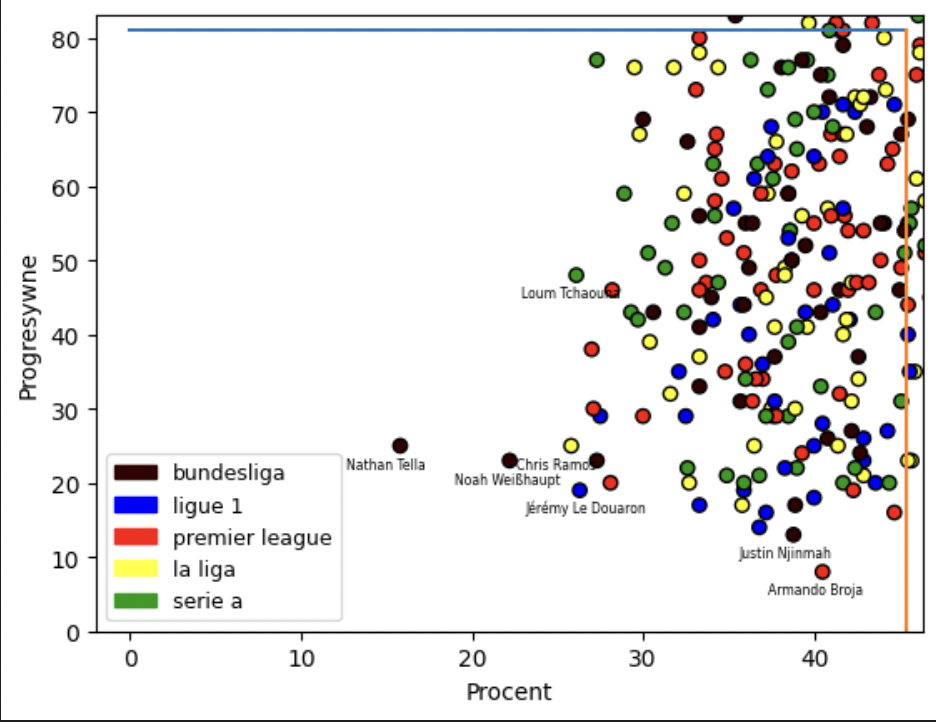
# Ćwiartka 3 – dobry drybler, słaby asystent



Trzecia, prawa dolna ćwiartka ukazuje nam zawodników mogących zostać ogłoszonymi w żargonie piłkarskim „jeźdźcami bez głowy". Duża skuteczność dryblingów przy niskim współczynniku podań progresywnych promuje zawodników szybkich i przebojowych, a niekoniecznie myślących i przewodzących akcji. Jest to klasyczny przypadek tak zwanych „zadaniowców", użytecznych przy konkretnym sposobie robienia przewagi - „kiwaniu się". Na naświetlenie zasługuje Renato Tapia. Peruwiańczyk cechuje się dużą skutecznością dryblingów przy minimalnie niższej liczbie podań progresywnych niż osiągana średnio. Choć nie należy do grona wschodzących gwiazd, mając już 28 lat, wciąż rysuje się na użytecznego i kreatywnego pomocnika w oczach skautów. Z kolei jego rywal ligowy, Umar Sadiq, gra znacznie mniej wszechstronnie. Obie jego statystyki nie prezentują się tak okazale jak Renato Tapii i, choć grają oni na innych pozycjach, obrazuje to trenerom pole do poprawy gry nigeryjskiego snajpera.

Studiując uważniej trendy na wykresie odnotować warto znacznie większe rozproszenie w porównaniu z pozostałymi omawianymi ćwiartkami oraz lekką dominację punktów barwy wskazującej na ligę hiszpańską. Wyjaśnić to można naciskiem stawianym na techniczność rekrutowanych do tych rozgrywek piłkarzy, która uatrakcyjnia mecze i dynamizuje grę. Zauważalnie większe odległości między punktami tłumaczą zróżnicowanie w charakterystyce dryblerów - każdy z nich dobrze gra piłką, kontrast polega na umiejętnościach rozgrywania. Pochylając się nad pozycjami zawodników wyróżnionych stwierdzamy, że próżno szukać wśród nich pivotów. Osoby pokroju Marcusa Thurama, Krépina Diatty czy Vitinhii to napastnicy, a ich niskie liczby w płaszczyźnie podań progresywnych dowodzą, że gracze najbardziej wysuniętej linii formacji piłkarskiej muszą uciekać się częściej do rozwiązania akcji indywidualnie zamiast kreowania szans partnerom podążającym za nimi.

# Ćwiartka 4 – słaby drybler, słaby asystent



Ostatnią, nieomówioną jeszcze ćwiartką jest ta wyznaczająca zawodników notujących liczby, które mogą menadżerów martwić. Słaby drybler i niezbyt kreatywny asystent, tak można określić graczy znajdujących się w tej części. Przynajmniej nie wychodząc poza wykres zaprezentowany. Aby szerzej omówić te wyniki należałoby przedstawić je w perspektywie minut, jakie poszczególni sportowcy spędzili na murawie. Dla przykładu, wybija się tutaj na pierwszy plan postać Nathana Telli, który wydaje się „szorować po dnie” kreatywnych statystyk. Nie możemy się jednak dać zmylić, gdyż wchodząc głębiej w dostępne dane oraz włączając mecze jego zespołu, absolutnego fenomenu tego sezonu, Bayeru Leverkusen, jesteśmy w stanie stwierdzić, że chłopak ten dostaje stanowczo za mało szans (liczba minut odpowiadająca 8 pełnym meczom, przy 32 możliwych do rozegrania), aby udowodnić swoją realną wartość i wymaga dalszej obserwacji w celu określenia czy liczby odzwierciedlają jego faktyczne umiejętności. To samo tyczy się Armando Brojy. Ten napastnik wydaje się nie szukać partnerów z drużyny i liczyć jedynie na swoje próby indywidualne, które i tak często kończą się fiaskiem. W tym przypadku także należy zwrócić uwagę na niewielką liczbę otrzymywanych minut, natomiast pozycja boiskowa pozwala mu na wybieranie tak ryzykownych zagrań jak wchodzenie w drybling.

Same punkty reprezentujące omawiany typ zawodników są skumulowane w okolicach przecięcia średnich, a dominującym kolorem jest ten odpowiedzialny za ligę niemiecką. Małe rozproszenie pokazuje, że mimo dużego zróżnicowania żaden piłkarz nie może sobie pozwolić na znaczące odstawanie od reszty stawki, co stawiałoby go w skrajnie niekorzystnym świetle. Dodatkowo analiza roli wyszczególnionych graczy unaocznia fakt, że większość z nich to właśnie napastnicy. Nie można jednak zapomnieć, że na ogół nie wymaga się od nich liczb w tych dziedzinach, gdyż zwykle kompensują je bramkami i asystami. Z tego powodu, by sprawdzić użyteczność graczy z takowymi niechlubnymi statystykami przy swoim nazwisku, analitycy powinni raczej obejrzeć ich w akcji niż bazować wyłącznie na danych zawartych na niektórych wykresach.

1. Podstawowe metody statystyki

# Tabela cech statystycznych

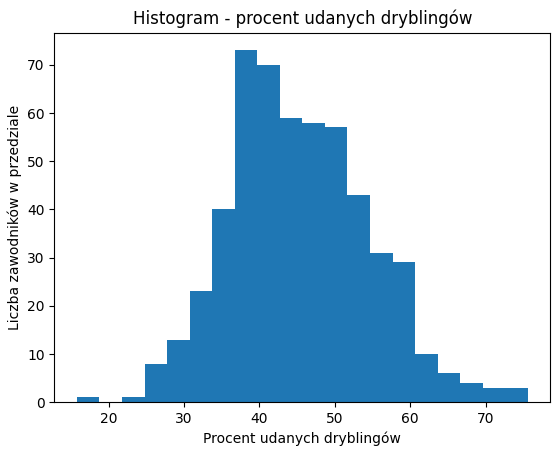
|  |  |  |
| --- | --- | --- |
| Cecha statystyczna | Procent udanych dryblingów | Podania progresywne |
| Średnia arytmetyczna | 45,39 | 81,28 |
| Średnia geometryczna | 44,45 | 66,88 |
| Średnia harmoniczna | 43,49 | 54,14 |
| Mediana | 44,7 | 69 |
| Odchylenie standardowe | 9,19 | 52,37 |
| Współczynnik skośności | 0,335 | 1,506 |
| Kurtoza | 3,107 | 6,048 |

# W tym momencie przejdziemy do analizowania naszych danych osobno, przy pomocy bardziej zaawansowanych metod. Pierwszą z nich będzie interpretacja wartości cech statystycznych obu naszych zbiorów. W powyższej tabelce umieściliśmy wyniki, które otrzymaliśmy, przy pomocy języka programowania python i stworzonych przez nas w nim funkcji, a także funkcji wbudowanych, dostępnych w bibliotekach. Zaczynając od pierwszej grupy pomiarów, dla procentu udanych dryblingów widać, że wszystkie nasze średnie oraz mediana są całkiem blisko siebie, zarówno w liczbach bezwzględnych, jak i względnych. Odzwierciedla się to także w małym odchyleniu standardowym. To, że średnia arytmetyczna jest większa, niż mediana znajduje swoje potwierdzenie w współczynniku skośności – jest on lekko dodatni, czyli nasze dane są lekko prawostronnie skośne. Kurtoza w tym wypadku niewiele przewyższa 3, co oznacza, że nasz rozkład będzie podobny do rozkładu normalnego, z tą różnicą, że widoczny będzie nieduży „szpikulec”, gdzie rozkład normalny ma gładką postać.

# W przypadku podań progresywnych średnie oraz mediana różnią się już wyraźnie. Odchylenie standardowe oraz współczynnik skośności, które są odpowiednio niemal i ponad 5-krotnie większe, niż w przypadku dryblingów, rzucają nam więcej światła na ten fakt – nasze dane są mocno rozproszone oraz mocno prawostronnie skośne. Dodatkowo, kurtoza wyraźnie powyżej 3, sugeruje istnienie wyraźnego „czubka” przy wizualizacji rozkładu tych danych.

# Wizualizacja danych

# Histogramy



W tym momencie przyszła pora na to, żeby zacząć wizualizować statystyki, które analizujemy, na różnego rodzaju wykresach. Pierwszą z naszych danych jest procent udanych dryblingów, który powyżej został przedstawiony na histogramie z 20 przedziałami. Można na nim zauważyć już kilka właściwości, które zostały przez nas wspomniane przy omawianiu tabelki, chociażby lekka skośność prawostronna danych. Widać także, że u większości z analizowanych przez nas zawodników skuteczność ich dryblingów wynosi od około 37% do około 52%, gdzie moda to właśnie przedział mniej więcej od 37% do 40%. Można wnioskować z tego, że przeciętny piłkarz, chętnie podejmujący drybling, na 2-3 próby, które podejmuje, będzie wychodził zwycięsko z jednej.

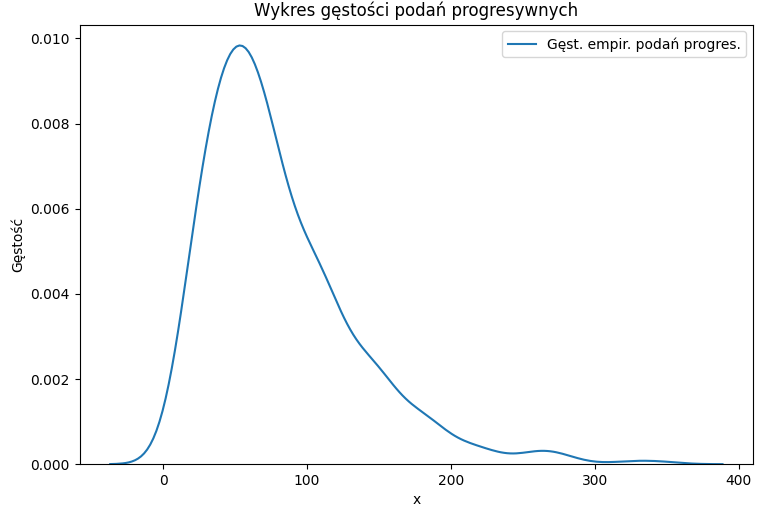
Obraz zawierający tekst, zrzut ekranu, diagram, Wykres

Opis wygenerowany automatycznie

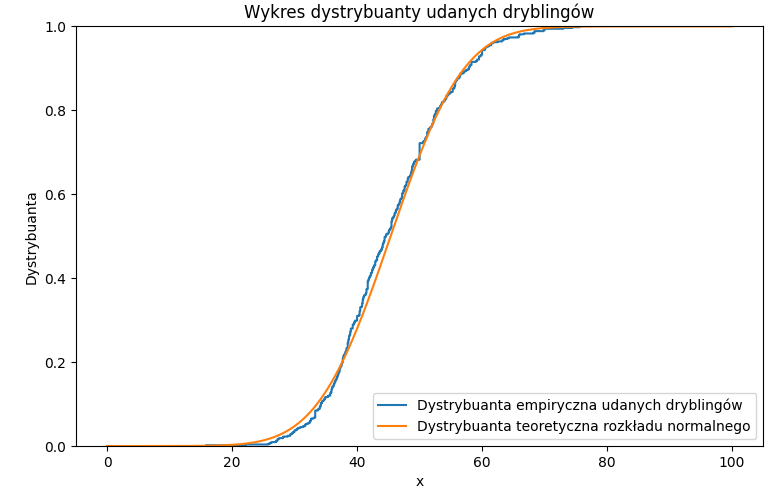
Histogram podań progresywnych wygląda zauważalnie inaczej względem histogramu skuteczności dryblingów. Jeszcze mocniej widać w przypadku tych danych skośność prawostronną, a także szpikulec, który tworzą dane – kolejny raz, fakty wspominane w przypadku tabelki z cechami. Na tym wykresie również mamy przedstawione 20 przedziałów, z czego najwięcej zawodników wpada do przedziałów od około 30 do około 70 podań progresywnych, gdzie modą jest przedział od 43 do 57 podań. Warto zwrócić uwagę na ogon naszego histogramu - widać, że ogromna część zawodników zaliczyła w tym sezonie mniej niż 200 podań, jednak istnieje spora grupa tych, którzy są wyraźnie wyspecjalizowani w tym fragmencie gry i osiągają nawet wartości bliskie 350.

# Gęstości

Gęstość rozkładu dryblingu, bez zaskoczenia, odzwierciedla to co widzieliśmy i w tabelce, i na histogramie - skośność prawostronną czy nagromadzenie danych w bliskiej okolicy średniej. Co ciekawe, zauważyliśmy tutaj duże podobieństwo wizualizacji tego rozkładu do rozkładu normalnego, wobec czego postanowiliśmy nałożyć na nasz wykres właśnie gęstość tego drugiego z parametrami równymi średniej arytmetycznej i odchyleniu standardowemu tych danych. Jak widać, obie funkcje wyglądają bardzo podobnie, jeszcze bliższe podobieństwo psuje jednak wspomniana wcześniej prawostronna skośność, jak i lekki „szpikulec”. Widać tu też, podobnie jak na histogramie, że najprawdopodobniejsza skuteczność dryblingu wśród zawodników mieści się w granicach od około 35% do 50%.



Gęstość rozkładu podań progresywnych wskazuje kolejny raz na dużą skośność prawostronną, a także wizualizuje sporą kurtozę - świadczy o niej duży, ostry "czub" powstający po lewej stronie wykresu. Właśnie ta akumulacja jest najciekawszym elementem tej statystyki. Można na jej podstawie zakładać, że zauważalna większość zawodników wykonała co najwyżej 150 podań progresywnych w tym sezonie. Tutaj niestety nie znaleźliśmy żadnego ze znanych nam rozkładów, który byłby wystarczająco podobny do tej gęstości.

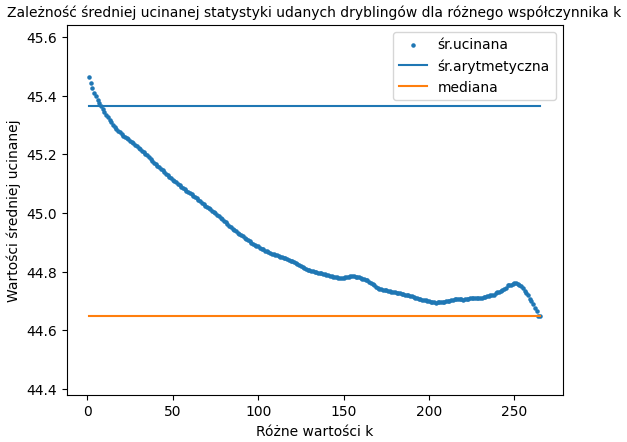
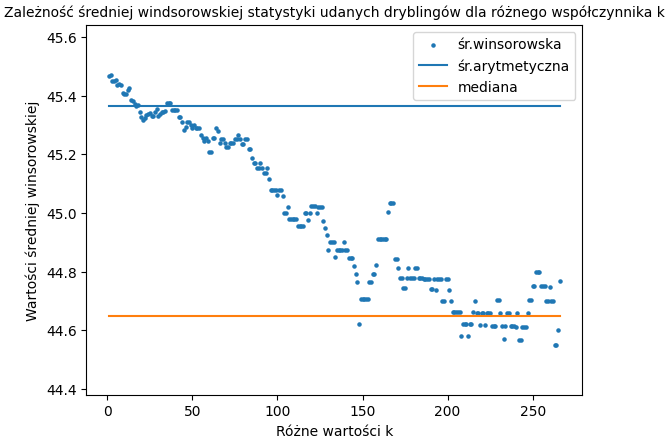
* **Dystrybuanty**

Dystrybuantę procentu udanych dryblingów, podobnie jak gęstość, z powodu podobieństwa porównaliśmy z dystrybuantą rozkładu normalnego. Jak widać na wykresie nasze funkcje pokrywają się w jeszcze większym stopniu niż miało to miejsce przy gęstościach. Widać tutaj największy wzrost w środkowej części wykresu, a wolniejszy w początkowej i końcowej. Momentami, zauważalny jest też jej skokowy wzrost, powiązany z nie aż tak dużą ilością danych, przy małej względnej różnicy między wartościami minimalną i maksymalną.

Obraz zawierający tekst, linia, diagram, Wykres

Opis wygenerowany automatycznie

Dystrybuanta podań progresywnych może natomiast na pierwszy rzut oka przypominać dystrybuantę rozkładu wykładniczego. To co ją jednak odróżnia to początek wykresu, w którym nasza funkcja rośnie raczej powolnie, gdzie w przypadku rozkładu wykładniczego ten wzrost już od samego zera jest bardzo dynamiczny. Później widać głównie podobieństwa, czyli wyraźny, szybki wzrost, który dla wartości podań progresywnych powyżej 100 zaczyna coraz bardziej hamować, aż w końcu dla tych powyżej 250 jest minimalny.

* **Średnie ucinane i winsorowskie**

Na powyższych wykresach porównujemy tym razem średnia ucinaną i winsorowską, dla różnych wartości parametru „k” ze średnią arytmetyczną i medianą. Podobnie jak wcześniej zaczniemy od procentu skuteczności dryblingów, analizowanych zawodników. Widać, że dla małych k te średnie są bliskie średniej arytmetycznej, jednak dla zwiększającego się parametru, bez zaskoczenia, zbiegają one do mediany. W przypadku średniej ucinanej zbieganie jest "gładkie", tj. w przeciwieństwie do średniej winsorowskiej nie mamy tak dużych wahań między niewiele różniącymi się wartościami k. Dla średniej winsorowskiej te wahania mogą być jednak tak widoczne, ze względu na fakt, że porównując do siebie, względnie, medianę i średnią arytmetyczną, są one bardzo blisko siebie - różnice nie przekraczają 2%. Wobec tego nawet małe różnice w naszych średnich ucinanych i winsorowskich będą bardziej widoczne na wykresie, niż gdyby wspomniana różnica była większa. W obu przypadkach widać też, że po wyraźnym zbliżeniu się do wartości mediany w okolicy k=200 obie średnie przez chwilę zaczynają rosnąć, aby później znowu, przed k=250, zacząć maleć i w szybkim tempie dążyć do wartości mediany.

Obraz zawierający tekst, zrzut ekranu, linia, diagram

Opis wygenerowany automatycznieObraz zawierający tekst, zrzut ekranu, linia, Czcionka

Opis wygenerowany automatycznie

Dla podań progresywnych wykresy średniej ucinanej i winsorowskiej wyglądają bardzo podobnie. W tym przypadku jednak można znaleźć kilka różnic - spadek średnich, wraz ze zwiększającym się parametrem k, jest dużo "gładszy", bez zauważalnych bardzo nagłych zmian wartości, a także wzrost w końcowej fazie jest zauważalnie mniejszy. Oba wynikają jednak głównie z faktu, że tutaj średnia arytmetyczna i mediana różnią się o prawie 20%, więc małe bezwzględne różnice liczbowe są dużo mniej zauważalne.

1. Porównanie danych

* Obraz zawierający tekst, zrzut ekranu, diagram, linia

  Opis wygenerowany automatycznie**Wykresy pudełkowe**

Pierwszym z naszych sposobów porównania dwóch rodzajów naszych danych jest spojrzenie na wygląd ich wykresów pudełkowych. Wykres pudełkowy procentu udanych dryblingów wraz z nałożonymi średnimi arytmetyczną, geometryczną i harmoniczną mówi nam kilka rzeczy. Pierwsze co rzuca się w oczy to fakt, że średnia geometryczna na wykresie wygląda jak równa naszej medianie. Widzieliśmy co prawda to już w tabelce, ale tu widać, że mimo różnicy liczbowej, wartości te na wykresie wyglądają na równe. Wszystkie średnie oraz mediana są względnie blisko siebie, co oznacza, że zbiór naszych danych jest całkiem symetryczny, jednak fakt, że średnia arytmetyczna jest trochę większa niż mediana po raz kolejny potwierdza jego lekką skośność prawostronną, o czym już kilkukrotnie wspominaliśmy. Widać też, że mamy bardzo mało wartości skrajnych - zaledwie czterech zawodników znajduje się poza przedziałem wyznaczonym przez wąsy dolny i górny. Z tego grona trójka wyróżnia się na plus, a jeden na minus. Biorąc pod uwagę, że analizowaliśmy ponad 500 zawodników to bardzo mała część. Dodatkowo widać też, że odległość od mediany do górnego wąsa jest większa niż do dolnego, co oznacza, że zawodnicy prezentujący się lepiej niż mediana, mają skłonności do większego odbiegania od wartości środkowej niż Ci słabsi, co ma także odzwierciedlenie w średniej arytmetycznej.

Obraz zawierający tekst, zrzut ekranu, diagram, linia

Opis wygenerowany automatycznie

Wykres pudełkowy podań progresywnych wyróżnia się przede wszystkim dużą ilością wartości skrajnych, szczególnie względem poprzedniego analizowanego, większych od górnego wąsa. Widać, że dużo zawodników wyróżnia się w tej statystyce w sposób pozytywny - notują dużo większe wartości niż średnia. Po raz kolejny średnia geometryczna wygląda na równą medianie. Tym razem jednak różnice bezwzględne między średnimi są dużo większe niż w przypadku poprzedniego wykresu pudełkowego - wynika to ze skośności naszego zbioru danych i tym samym jego asymetryczności. Widoczna jest ona też w odległości wąsów od mediany, wąs górny jest wyraźnie dalej niż dolny. Można wnioskować, że piłkarze przewyższający medianą często dużo mocniej odbiegają od niej, niż Ci wykonujący ich mniej. Innymi słowy, jeżeli piłkarzowi zdarza się wykonywać więcej podań niż ogólny poziom, to jest spore prawdopodobieństwo, że wykonuje on ich bezwzględnie bardzo dużo.

* Obraz zawierający zrzut ekranu, oprogramowanie, Oprogramowanie multimedialne, Wykres

  Opis wygenerowany automatycznie**Heatmapa**

Na powyższym wykresie przedstawiona została tak zwana heatmapa - zależność danych o skuteczności dryblingów oraz liczby podań progresywnych przedstawionych na histogramie w wymiarze 2D, skonstruowany dzięki pythonowej komendzie 'matplotlib.pyplot.hist2d'. Łatwo dostrzec duże podobieństwo do skonstruowanego na początku wykresu punktowego tejże zależności. Po prawej stronie legenda kolorów informuje nas, ile zawodników kwalifikuje się do odpowiedniego kwadratu, na które została podzielona cała plansza. Im jaśniejszy kolor, tym więcej piłkarzy „wpadło" do danej podziałki. Tym samym zauważamy, że największe zagęszczenie panuje lekko poniżej przecięcia się średnich, które były zaznaczone na wcześniejszym wykresie. Podobne spostrzeżenie mogliśmy wysnuć i wtedy. Jednakże taka forma prezentacji danych pozwala na dokładniejszą analizę ilościową wykresów punktowych, uświadamiając nas o kształtującym się rozłożeniu danych ilościowych. Ponadto takowe przedstawienie promuje zawodników wyróżniających się (szczególnie tych na plus), właściwie czarno na białym malując ich wybijanie się ponad stan. Analiza ilościowa daje także informacje na temat właściwego uśrednienia wyników, które mogą się trochę nie zgadzać z wartościami średnich wyliczanymi, obrazując prawdziwą definicję przeciętności i normalności. W końcu, heatmapy takie jak ta udoskonalają porównanie dwóch danych, z której można tłumaczyć ich zachowanie w stosunku do siebie, na przykład tutaj więcej kwadratów w jaśniejszych barwach wyodrębnić można w prawej dolnej części niż w przeciwnej do niej lewej górnej, co by oznaczało, że częściej skuteczny drybling wiąże się z małą liczbą progresywnych podań niż kreatywność uzyskiwana przez progresywne podania z niską jakością dryblingu. Dzięki temu histogram 2D jest użyteczny do celów rozpoznawania trendów i porównywania ich z wynikami dostarczonymi przez analityków klubów.

* **Kowariancja i korelacja**

Na koniec przeanalizujemy jeszcze miary zależności naszych zestawów danych. Podobnie jak w przypadku wcześniej zamieszczonej tabeli z cechami statystycznymi, policzyliśmy je przy pomocy pythona. Po pierwsze kowariancja analizowanych przez nas skuteczności dryblingów i podań progresywnych jest równa 162. Z faktu, że jest ona wyraźnie dodatnia można wywnioskować, że gdy zawodnik lepiej radzi sobie w dryblingach możemy oczekiwać, że będzie także wykonywał więcej podań progresywnych.

Policzona korelacja Pearsona, wykorzystująca naszą kowariancję wynosi natomiast 0,339. Wobec tego nasze dane są wyraźnie skorelowane w sposób liniowy i potwierdza się też fakt, że wzrost jednej ze statystyk najprawdopodobniej odzwierciedla się w pozytywny sposób w drugiej. Jednak jest ona dalej na tyle mała, że zdarzają się wyraźne odchylenia i przypadki skrajne.

Z kolei korelacja Spearmana przyjmuje wartość 0,363. Widać tu, że ta miara, mierząca również inne rodzaje zależności, niż tylko liniową, jest większa - nasze dane są trochę mocniej skorelowane niż sugerowałaby to korelacja Pearsona. Dalej nie jest to bardzo duża korelacja, ale jest ona na pewno wyraźna i pozwala nam o wnioskowaniu na temat wartości jakie ma największe szanse przyjmować jedna statystyka zawodnika, gdy znamy tylko tę drugą.

1. Podsumowanie

Za cel raportu postawiliśmy sobie analizę próbki 532 piłkarzy pod kątem zależności procentu udanych dryblingów oraz liczby podań progresywnych na przestrzeni obecnie trwającego sezonu 2023/2024. Udało się nam sporządzić wykres powiązania tych danych, a dokładniejsze pochylenie się nad wynikami pozwoliło na wyciągnięcie istotnych wniosków:

* Analiza wykresów punktowych umożliwia klasyfikację pewnej grupy podmiotów na bazie kombinacji statystyk i porównania z pozostałymi jednostkami,
* Użycie średnich arytmetycznych na wykresie daje sposobność określenia „środka ciężkości” statystyki punktowej, to znaczy obszaru, w którym duża część badanej próbki będzie się znajdować,
* Same zagłębienie się w sporządzone materiały dało odpowiedź na pytania: którzy zawodnicy błyszczą w tym sezonie swoją kreatywnością, jakie cechy są promowane w różnych ligach piłkarskich, a w dalszych refleksjach doszliśmy do konkluzji, że przy poziomie cyfryzacji dzisiejszego życia przygotowywanie takowych wykresów przez analityków klubowych staje się praktycznie nieodzowne, aby nie stracić z oczu konkurencji.

Następnie obliczyliśmy wartości cech statystycznych par danych z naszej próbki, które umieściliśmy w tabelce, a w niektórych przypadkach wygenerowaliśmy odpowiednie wykresy, by następnie porównać dwie dane przyjęte na samym początku, wizualizując je na wykresach pudełkowych, histogramie 2D oraz obliczając kowariancję, korelacje Pearsona i Spearmana. Po wykonaniu ich udało nam się również stwierdzić, że:

* Analiza wykresów gęstości i dystrybuant odsłoniła przed nami ciekawy fakt; rozkład skuteczności dryblingów zbliżony jest do rozkładu normalnego, natomiast rozkład liczby progresywnych podań do pewnego stopnia przypomina rozkład wykładniczy,
* Wykresy pudełkowe pozwoliły na ponowne zdefiniowanie podziału ogółu zawodników, tym razem w zależności od kwartylów, w jakich ich wyniki się mieszczą, wyodrębnienie pomiarów niemieszczących się w przeważającej części wyników oraz porównanie poszczególnych średnich z medianą i samym wykresem pudełkowym,
* Wyliczenia kowariancji i korelacji udowodniły występowanie zależności liniowej między tymi dwiema danymi, co sugerowałoby sensowność wszelkich naszych rozważań na temat określania kreatywności zawodnika na bazie różnych od siebie statystyk, które analizowaliśmy od początku. Warto jednak wspomnieć, że zdarzają się odchylenia od normy, nie do końca potwierdzające nasze przypuszczenia, jednakże takowe błędy statystyczne zdarzyć się mogą w każdym badaniu.

Przeanalizowanie wszystkich wykresów i wyników wyliczeń daje nam dobre zorientowanie w świecie tego-sezonowej piłki i umożliwia predykcję przyszłych ruchów klubów TOP 5 lig na bazie formy prezentowanej przez poszczególnych zawodników. Jednakże nie pozwala na stwierdzenie, że wiemy już wszystko. Jesteśmy zatem gotowi do poszerzania zarówno swojej wiedzy hobbystycznej, jak i akademickich umiejętności z zakresu stosowania statystyki w przyszłych projektach.