Uniwersytet Pedagogiczny im. Komisji Edukacji Narodowej w Krakowie

INSTYTUT BEZPIECZEŃSTWA I INFORMATYKI

Kierunek Informatyka, Specjalność Administracja Systemami Informatycznymi

Piotr Maliga oraz Krzysztof Wszołek

EKSPLORACJA DANYCH ZA POMOCĄ TOPOLOGICZNEJ ANALIZY DANYCH NA PRZYKŁADZIE AMERYKAŃSKO-KANADYJSKIEJ LIGI KOSZYKARSKIEJ NBA

praca Inżynierska

napisana pod kierunkiem

dr. Marcin Żelawski

mgr. Patryk Mazurek

Kraków 2023

Spis treści

Wstęp

Nadrzędnym celem naszej pracy będzie analiza danych Amerykańsko-Kanadyjskiej ligi koszykarskie NBA, w celu wyciągnięcia interesujących wniosków. Naszą analizę, Nasze sprawozdanie podzielimy na dwie części, pierwsza będzie dotyczyć Wprowadzenie do tematu naszej pracy, a druga będzie zawierać analizę danych..

Rozdział 1

Wprowadzenie

Rozwój technologii informatycznych związanych z gromadzeniem i składowaniem danych spowodował, że instytucje, organizacje i przedsiębiorstwa dysponują coraz większymi zbiorami danych. Co więcej tempo przyrostu zbieranych danych i ich zróżnicowanie jest coraz większe. Wpływają na to takie czynniki jak rozwój mediów społecznościowych, multimediów, handel internetowy, a w najbliżej perspektywie rozwój internetu rzeczy. W wielu przypadkach ilość lub charakter danych utrudnia ich bezpośrednie wykorzystanie. Analiza wielkich zbiorów danych, tak zwanych big data, staje się często czynnikiem przewagi konkurencyjnej dla przedsiębiorstw oraz źródłem innowacji.

Aby pozyskać wiedze, na podstawie zebranych danych konieczne jest ich odpowiednie usystematyzowanie i przetworzenie. Praktycznymi problemami są duże ilości danych ich duża zmienność i różnorodność, a także szum informacyjny, dane niepełne, niedokładne, niezweryfikowane. Zebrane dane mogą mieć postać liczbową, tekstową, graficzną lub dowolną inną. Wśród stosowanych w praktyce metod agregacji, manipulacji, analizy i wizualizacji danych raport McKinsey z 2011 roku wymieniał między innymi testy A/B, uczenie maszynowe, metody statystyczne, sieci neuronowe, algorytmy genetyczne, przetwarzanie języka naturalnego, przetwarzanie danych w chmurze. Nowym podejściem do analizy danych jest zauważenie, że dane mają kształt, a kształt ma znaczenie. To problematyka zajmuje się topologiczną analizą danych. Topologia to dział matematyki zajmujący się badaniem własności obiektów, które nie ulegają zmianie nawet po ich zdeformowaniu. Przez deformacje rozumie się dowolne odkształcanie niewymagające rozrywania i łączenia różnych części na przykład rozciąganie czy zginanie. Z punktu widzenia analizy danych kluczowe znaczenie ma to, że obiekty nie są ograniczone do przestrzeni dwu lub

trójwymiarowej. Mogą mieć dowolny wymiar co przekłada się na możliwość analizy dowolnie złożonych danych przy zastosowaniu podobnej metodologii. Przykładowo dane genetyczne mają ponad pół miliona cech wzajemnie ze sobą powiązanych.

Historycznie pierwszym rozważanym naukowo zagadnieniem topologicznym był problem mostów z Królewca rozwiązany przez Eulera w XVIII wieku. Jednak dopiero od kilkunastu lat zaczęto zauważać związki topologii ze zbiorami danych. Dane możemy traktować jako skończoną chmurę punktów w przestrzeni wielowymiarowej. Taka chmura może być traktowana jako próbka wzięta z obiektu geometrycznego, prawdopodobnie zawierająca szum. Topologiczna analiza danych próbuje ustalić własności takiego obiektu. Obecnie topologiczna analiza danych jest ju wykorzystywana komercyjnie, stała si elementem rynku big data. W 2008 roku powstała Ayasdi, firma typu spin-off Uniwersytetu Stanforda, specjalizująca się w analizie danych medycznych i finansowych. Wartość inwestycji typu Venture Capital w Ayasdi w latach 2012–15 przekroczyła 100 mln dolarów, co świadczy o tym, że rynek zauważa potencjał nowego spojrzenia na dane.

1.1 Czym jest analiza danych ?

Na początku naszej pracy zastanówmy się czym tak naprawdę jest analiza danych?

Jak podaje Encyklopedia Zarządzania - *„Analiza danych to proces polegający na sprawdzaniu, porządkowaniu, przekształcaniu i modelowaniu danych w celu zdobycia użytecznych informacji, wypracowania wniosków i wspierania procesu decyzyjnego. Analiza danych ma wiele aspektów i podejść, obejmujących różne techniki pod różnymi nazwami, w różnych obszarach biznesowych, naukowych i społecznych. Praktyczne podejście do definiowania danych polega na tym, że dane to liczby, znaki, obrazy lub inne metody zapisu, w formie, którą można ocenić w celu określenia lub podjęcia decyzji o konkretnym działaniu. Wiele osób uważa, że dane same w sobie nie mają znaczenia – dopiero dane przetworzone i zinterpretowane stają się informacją.”*

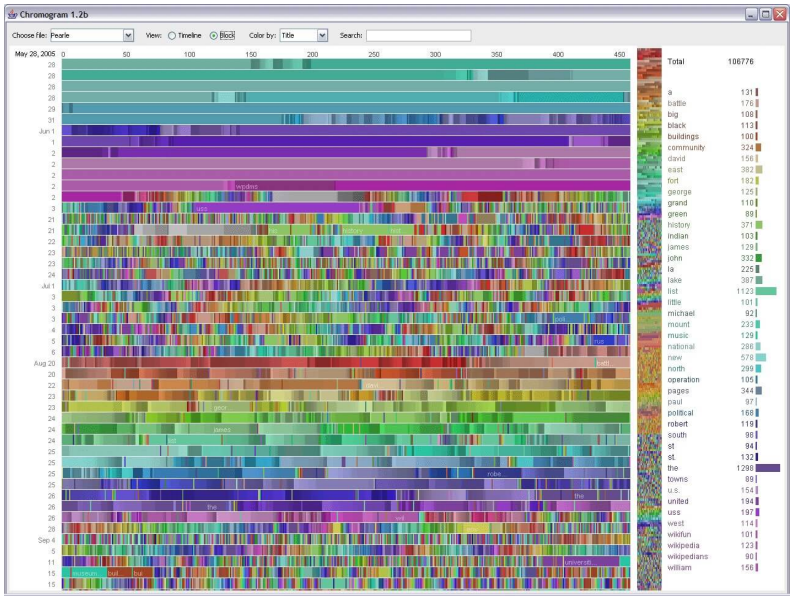
1.2 Metody Analizy danych

Badanie statystyczne jest złożonym procesem, którego celem jest uzyskanie zamierzonych informacji poprzez całokształt czynności badawczych. Zarówno zbieranie, gromadzenie, jak i opracowywanie danych statystycznych to zaledwie tylko część badania, ponieważ po zebraniu materiału statystycznego należy przedstawić wyniki w formie obliczeń, opracowań i analiz. Dogłębne poznanie metod analizy danych statystycznych jest istotne dla każdego, kto chciałby zająć się szeroko pojętą analizą danych. Niezwykle istotne jest poznanie wad każdego sposobu analizy, aby badacz miał świadomość błędów, jakie może popełnić.

Do metod analizy danych statystycznych zaliczamy:

- **analizę wariancji** - (ANOVA) służącą do badania obserwacji zależnych od jednego lub wielu działających jednocześnie czynników; wyjaśnia nam z jakim prawdopodobieństwem wyodrębnione wcześniej czynniki mogą być powodem różnic pomiędzy obserwowanymi przez nas średnimi grupowymi,
- **analizę korelacji** - polega na badaniu, czy dwie zmienne są ze sobą istotnie statystycznie powiązane; sprawdza również, czy jakiekolwiek dwie cechy, atrybuty bądź własności (wyrażone liczbowo) współwystępują ze sobą; obliczony współczynnik zawsze waha się między -1 a 1
- **analizę przeżycia** - badającą procesy, w których interesuje nas czas, jaki upłynie do (pierwszego) wystąpienia pewnego zdarzenia (zdarzeniem tym może być np. śmierć pacjenta lub odejście pracownika z firmy),
- **analizę regresji** - opisującą współzmienność kilku zmiennych poprzez odpowiednie dopasowanie do nich funkcji; analiza ta umożliwia nam przewidywanie nieznanych wartości jednych zmiennych na podstawie znanych wartości innych,
- **analizę czynnikową** - Jej celem jest opisanie zależności pomiędzy zaobserwowanymi, skorelowanymi zmiennymi przy pomocy możliwie jak najmniejszej nieobserwowanej ich liczby nazywanej czynnikami bądź faktorami, które są wzajemnie nieskorelowane,
- **analizę dyskryminacyjną** - rozstrzyga, które zmienne niezależne w najlepszy sposób będą dzielić dany zbiór przypadków na występujące w naturalny sposób grupy, opisane jakościową zmienną zależną,

- **analizę szeregów czasowych** - bada własności szeregów czasowych, a także prognozuje na ich podstawie,
- **analizę kanoniczną** - pozwala badać związki pomiędzy dwoma zbiorami zmiennych.



Ryc. 1.1: Wizualizacja edycji Wikipedii, jako klasyczny przykład Big Data

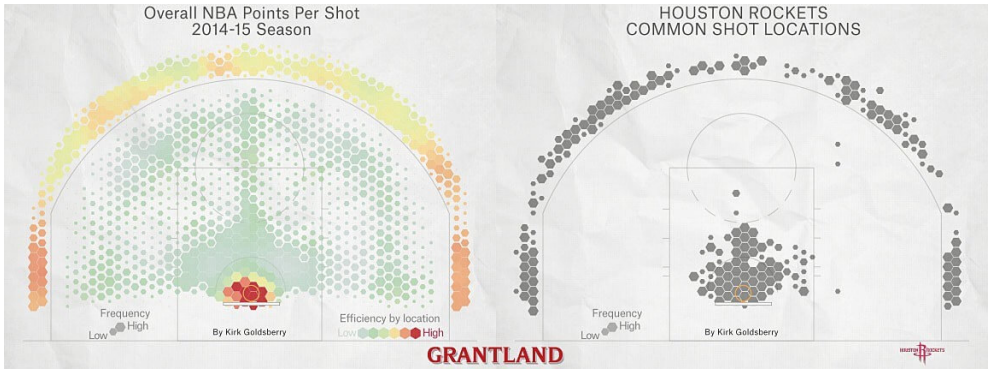
1.3 Metody topologicznej analizy danych

Topologiczna analiza danych to sposób analizy dużych wielkowymiarowych zbiorów danych. Jej założeniem jest poznanie kształtu danych i wyciągnięcie z niego wniosków. Prace omawia podstawowe pojęcia związane z topologicznym kształtem danych. Przedstawia założenia i metody obliczania homologii persystentnej. Jej celem jest analiza chmury punktów danych pozwalająca pogrupować je i zależnie zależności między nimi. Metoda działa w dowolnych przestrzeniach n-wymiarowych i nie wymaga wcześniejszych założeń co do szukanych zależności. Do prezentacji wyników służą wykresy słupkowe i diagramy persystencji.

1.4 Analiza danych w lidze NBA

W koszykówce w ciągu ostatniej dekady zbiory danych poczyniły duże, wpływowe postępy, pomagając poprawić wyniki indywidualne i zespołowe, a także sposób rekrutacji nowych graczy. Przytłaczająca większość drużyn NBA w USA ma teraz swoich pracowników, jako analityków danych, a liga organizuje coroczny hackathon Big Data.

Wśród drużyn ligi NBA, dobrze znamy Houston Rockets. Ich awans ze średniej przeciętnej drużyny do bycia jedną z najlepszych drużyn koszykówki, nie sprowadza się tylko do składu genialnych graczy. Big Data odegrała tutaj ogromną rolę. Wszystko zaczęło się pod koniec lat 90 ubiegłego wieku, kiedy Rockets byli jedną z czterech drużyn NBA, które zainstalowały system śledzenia wideo, który wydobywał surowe dane z gier. Analiza danych wykazała, które rzuty miały najwyższe wskaźniki skuteczności. Były to wsady za dwa punkty i rzuty za trzy punkty. Od sezonu 2008-2009 liczba rzutów za trzy punkty, jako procent wszystkich rzutów w lidze NBA znacznie wzrosła. W sezonie 2017-2018 drużyna Houston Rockets oddała 1184 rzutów za trzy punkty, więcej niż jakikolwiek inny zespół NBA w historii tego sportu. To był główny czynnik, który przyczynił się do pobicia rekordu i zdobycia większości zwycięstw w sezonie.



Ryc. 1.2: Houston shot locations (miejsca, z których najczęściej padają rzuty)

1.5 Maszyny rozrozumieniejące grę, idealne dla Trenera

Od sezonu 2017-2018, Second Spectrum jest oficjalnym dostawcą technologii śledzenia wideo dla drużyn NBA. Kamery zainstalowane na arenach stadionów zbierają trójwymiarowe dane przestrzenne, w tym lokalizację piłki/gracza/sędziego, ruchy zawodników i tym podobne. Technika rozpoznawania wzorców czasoprzestrzennych identyfikuje i wzmacnia dane wyodrębnione z nagrania wideo, aby generować spostrzeżenia dla trenerów i zespołów.

„Są rzeczy, które trenerzy koszykówki chcą wiedzieć, a problem polega na tym, że nie mogą ich znać, ponieważ musieliby oglądać każdą sekundę meczu i pamiętać o tym. A człowiek nie może tego zrobić, ale maszyna może” - Powiedział prezes Second Spectrum, Rajiv Maheswaran w przemówieniu TED w 2015 roku.

Ryc. 1.3: Video tracking system (system śledzenia wideo)

Rozdział 2

Analiza danych

Wprowadzenie

Dane pochodzą z serwisu basketball-reference.com ta strona zawiera oficjalne dane z ligi NBA. W celu wyrównana danych z każdej branej przez nas drużyny weźmiemy pod uwagę graczy, którzy spędził przynajmniej 1000min na boisku podczas danego sezonu. Dla każdego gracza podane są następujące statystyki:

- Imię oraz wiek gracza
- Liczba gier
- Czas gry (min)
- Liczba rzutów
- Rzuty za 3 oraz za 2
- Zbiórki w ofensywie
- Zbiórki w defensywie
- Wszystkie zbiórki (TR)
- Asysty (AS)
- Przechwyty (S)
- Liczba fauli
- Blok (B)
- Liczba zdobytych punktów(PTS)

W celu zbudowana diagramów przedstawiających skuteczność wybranych drużyn wybraliśmy dane, które odpowiadają za główny czynnik zwycięstwa: TR, AS, S, B, PTS. W celu filtracji wszystkich podanych statystyk używaliśmy narzędzia w serwisie basketball-reference.com służącego do modyfikacji tabel oraz metod języka python. Następnie statystyki zostały skompilowane do pliku csv. Owe pliki zostały wprowadzone do programu (wykorzystującego bibliotekę Ripser.py 0.6.4), który na podstawie wprowadzonej przez nas chmury danych na wyjściu tworzył wykres barcode persystencji klas homologii. Finalnie mogliśmy wizualnie przeanalizować wyniki.

Ogólnie rzecz biorąc, trwała homologia zarejestrowana, dla każdego zespołu składa się tylko z 0-wymiarowej i 1-wymiarowej homologia. Są to wymiary, które interpretowaliśmy w sekcji teoretycznej. Zakładając, że oczekujemy, że zespół będzie wyróżniał się, jeśli jego 0-wymiarowy diagram homologii ma długie wskaźniki przeżycia. Geometrycznie oznacza to większy rozrzut w statystykach graczy. Oczekujemy, że zespół będzie wyróżniał się, jeśli jednowymiarowa homologia jest krótkotrwała lub nie istnieje. Analizując diagramy 2 najlepszych drużyn z sezonu 90-91 (Denver Nugets i Golden State Warriors widzimy długie wskaźniki przeżycia w wymiarze 0 i brak klas 1-wymiarowych. Najsłabiej wypadły dwie drużyny Minesota Timberwolws i Sacramento Kings

Diagram wymiarowy Sacramento Kings jest klasycznym przykładem wadliwego zespołu. Wczesna identyfikacja w dolnej część diagramu przedstawia grupę podobnych graczy. Praktyczny wniosek: skład zespołu jest zadowalający lub korzystny, gdy diagram 0-wymiarowy ma długie czasy przeżycia jak w przypadku Denver Nugets czy Golden State Warriors. Wydaje się, że jest to podstawowa cecha, którą można łatwo rozpoznać, analizując dane diagramy. Inną numeryczną cechą diagramu 0-wymiarowego jest średnia długości klas 0, która jest faktycznie obszarem obszar zacieniony. Drugorzędną cechą charakterystyczną jest liczba i długość 1-wymiarowej klasy.

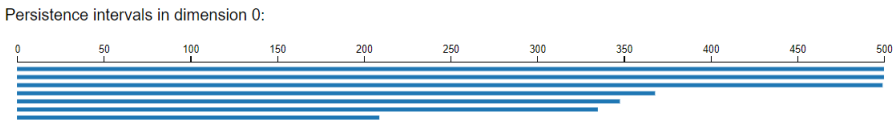
2.1 Proponowane wyjaśnienie odkrytej korelacji

Korzystając z topologii, przechodzimy od lokalnej do globalnej informacji. Wszystkie podane statystyki gracza niekoniecznie będą świadczyć o tym, czy dana drużyna odniesie sukces, czy też nie. Prosta ocena „siły" zawodnika nie jest w stanie zagwarantować sukcesu danego zespołu. W pomiarze rzadkości nie chodzi o obecność gorszych i lepszych graczy w takim samym stopniu jak graczy o zróżnicowanych indywidualnych cechach. Korzystne zróżnicowanie drużyny możemy przedstawić w sytuacji, w której, kiedy w naszej pierwszej-5 mamy zawodnika silnego i wysokiego, który niekoniecznie będzie trafiał najwięcej rzutów, ale przez swoją fizyczność będzie w stanie mocno budować pozycje na boisku, zbierać nietrafione rzuty czy to w ofensywie, czy w defensywie oraz blokować zawodników rzucających z bliskiej pozycji od kosza. Kolejnym potrzebnym ogniwem będzie zawodnik atletyczny szybki, który będzie w stanie w odpowiednich okolicznościach poprowadzić szybką kontrę czy też nadać szybszego tempa gry, jeśli będzie to potrzebne. Kolejnym zawodnikiem będzie gracz, który będzie w stanie celnie podawać, zgarniając przy tym liczne asysty. Potrzebny będzie również gracz mający wysoką skuteczność rzutu za 3, gdyż właśnie w koszykówce głównie pozwalają zbudować szybką przewagę nad drużyną przeciwnika. Oczywiście mamy również wyjątki graczy wybitnych, którzy swoim stylem gry oraz połączeniem większości z powyżej podanych cech są w stanie niejednokrotnie odmienić os całej drużyny. Oczywiście nie jest to recepta na drużynę, która jest w stanie wygrać każdy sezon, oczywiście jest jeszcze współczynnik losowy wiele innych czynników, które mogą wpłynąć w pozytywny czy tęż negatywny sposób. Również w zależności, w jakim stylu trener chcę prowadzić grę, przeciw komu gra to wszystko, może wpłynąć na różne konfiguracje zawodników.

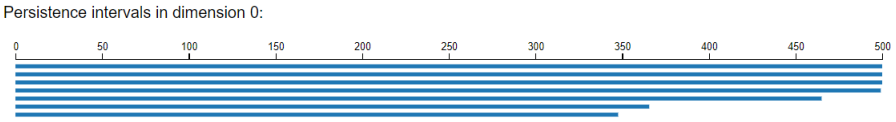
Kolejnym przykładem stojącym za korzystnym zróżnicowaniem zespołu jest przykład przedstawiony przez Muthu Alagappana, studenta inżynierii ze Stanford, przedstawił on swoje badania na konferencji MIT Sloan Sports Analytics Conference. Przeanalizował występy drużyn NBA w sezonie 2011-2012. Sporządził mapę graczy z drużyn korzystających z autorskiego oprogramowania o nazwie Mapper, czyli opracowany przez firmę Ayasdi. Alagappan stwierdził, że im bardziej różnorodna jest drużyna, tym większe sukcesy odnosi w trakcie sezonu.

2.2 Proponowane wykorzystanie analityki

Ogólny pomysł na aplikację do zarządzania zespołem polega na tym, że diagramy trwałości identyfikują braki w obecnym składzie zespołu. Po zidentyfikowaniu danego diagramu menedżer może podjąć próbę usunięcia braków poprzez sprzedanie zawodnika czy zakupienie brakującego. Dana aplikacja mogłaby przyjmować statystyki zawodnika, którego dana drużyna chcę kupić, następnie zestawiać jego dane z resztą danych naszego zespołu w celu analizy diagramu, na podstawie którego selekcjoner mógłby podjąć odpowiednią decyzję. Oto hipotetyczny przykład. Załóżmy, że chcielibyśmy udoskonalić Sacramento Kings poprzez wymianę. Po szybkim eksperymencie, z możliwymi transakcjami, jeden radził sobie wyjątkowo dobrze, podczas gdy inne nie robiły różnicy homologicznej. Sprawdźmy więc, czy wymiana w Minnesota Timberwolves, Feltona Spensera na Horace Granta zamierza wzmocnić zespół



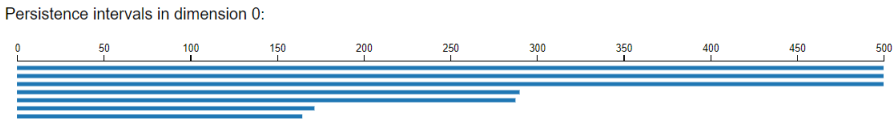
Ryc. 2.1: Minnesota Timberwolves przed zamianą zawodników:



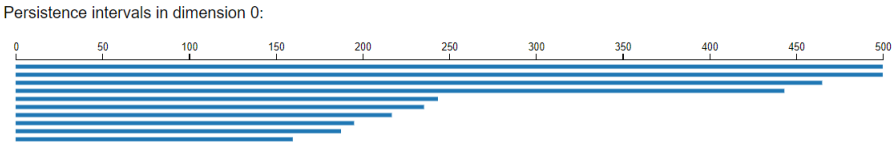
Ryc. 2.2: Minnesota Timberwolves po zamianą zawodników:

Możemy zaobserwować duża zmienne wykresu. Idąc dalej tokiem naszego przykładu manager w takiej sytuacji mógłby opierać decyzje zamiany zawodników na wygenerowanym przez nas wykresie.

Porównanie dwóch najleprzych drużyn z sezonu 90-91

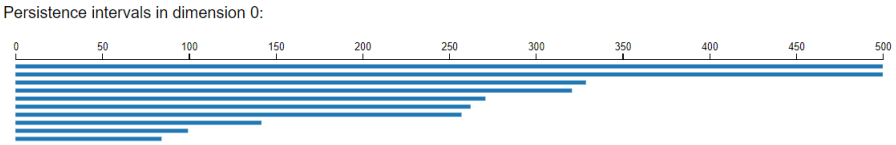


Ryc. 2.3: Golden State Warriors

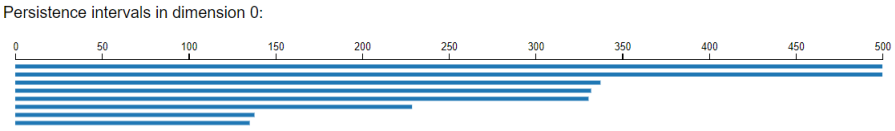


Ryc. 2.4: rys: Denver Nuggets

Porównanie dwóch drużyn które na koniec sezonu były pośrodku tabeli

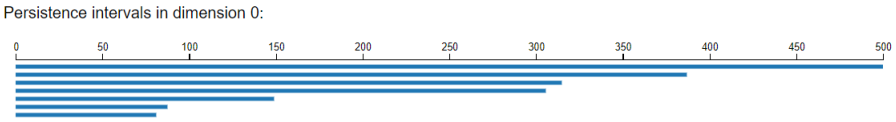


Ryc. 2.5: Los Angeles Lakers

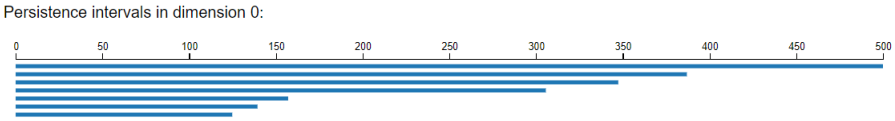


Ryc. 2.6: Orlando Magic

Porównanie dwóch najsłabszych drużyn na koniec sezonu



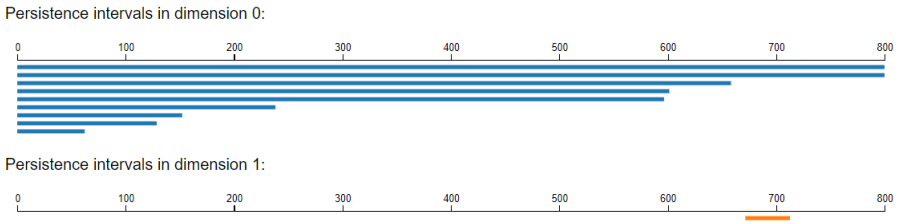
Ryc. 2.7: Minesota Timberwolves



Ryc. 2.8: Sacramento Kings

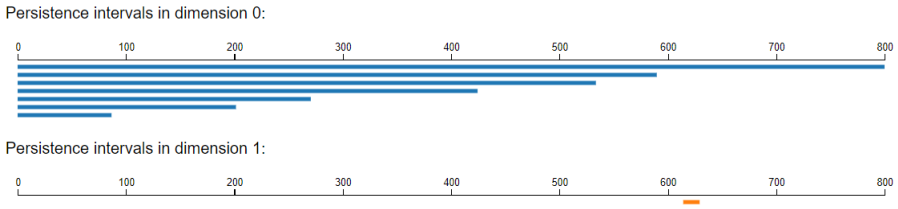
Porównanie 4 sezonów 93-96 drużyny Chicago Bulls. Sezony 93-96 dla fanów byków były bardzo emocjonujące w sezonie 92-93 Chicago Bulls świętowało woje 3 z rzędu zwycięstwo ligi NBA w sezonie 93-94 słynny Michael Jordan wziął rok urlopu od koszykówki co znacznie zmieniło dynamikę drużyny. Wszystkie te zmiany możemy zaobserwować na wykresach poniżej. Najlepiej widać to na wykresie 1 wymiaru gdzie widzimy, że linia życia ulega znacznej zmianie. Kolejny sezon 94-95 był sezonem, w którym Michel Jordan spędzał bardzo mało czasu na boisku. Z racji wyrównania danych jego statystyki nie zostały dodane do chmury punktów. Sezon 95-96 Michel Jordan znowu zaczyna królować na boisku, przywracając silne Chicago Bulls i Zdobywając mistrzostwo ligi. Jak widzimy wykres z sezonu 95 – 96 upodobnił się bardziej do wykresu z sezonu 92 – 93. Oczywiście łączy je to, że w tych sezonach Chicago Bulls była bardzo skuteczną drużyną.

Sezon 92 - 93



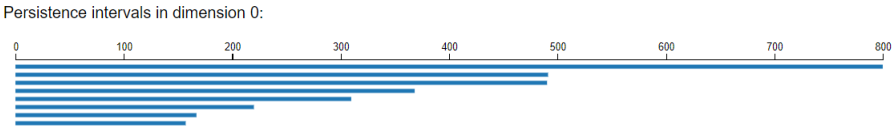
Ryc. 2.9: Sezon 92 - 93

Sezon 93 - 94



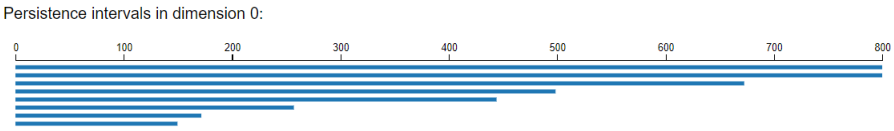
Ryc. 2.10: Sezon 93 - 94

Sezon 94 - 95



Ryc. 2.11: Sezon 94 - 95

Sezon 95 - 96



Ryc. 2.12: Sezon 95 - 96

Spis rysunków

Bibliografia

[1] *Encyklopedia Zarządzania - Analiza Danych*

[2] *Zastosowanie Topologicznej Analizy Danych - Autorstwa: Artura Żuwała*

[3] *Moneyball 2.0: How Missile Tracking Cameras Are Remaking The NBA*

[4] *Analyzing NBA basketball data with R*

[5] *Game of Waveforms*

Książki

[6] ***Lutz Mark*** *- Python - Wprowadzenie*

[7] ***Vaughan Lee*** *- Python z życia wzięty. Rozwiązywanie Problemów za pomocą kilku linii kodu*

Kursy online

[8] ***Kanał o Wszystkim*** *- Kurs Python - Programowanie*

[9] ***Greg Hogg*** *- Complete Beginner's Tutorial to Google Colab*

Prace własne

[10] ***Piotr Maliga*** *- Metody-Badawcze-w-Informatyce-Piotr-Maliga*