Analiza wyników – zestaw zadań 3

Izabela Karczewska, Natalia Kołodziejczyk, Natalia Kubańska

A screenshot of a computer

Description automatically generated

**Natężenie ruchu rowerowego**

Zestaw danych zawiera dane na temat ruchu rowerowego na moście Fremont Bridge w Seattle. Akcja mierzenia liczby przejazdów rowerowych przez most rozpoczęła się w październiku 2012. Pętle indukcyjne mierzą liczbę przejazdów na zachodniej i wschodniej ścieżce bez względu na kierunek jazdy. Choć nie jest to uwzględnione w danych to większość przejazdów po wschodniej ścieżce jest w kierunku północnym, natomiast na zachodniej ścieżce w kierunku południowym. Indeksem danych jest godzina, w której dokonano zliczeń. Dane zostały ograniczone do roku 2022. Zliczenia występują osobno dla wschodniej i zachodniej strony, kolumna total jest sumą dwóch pozostałych.

A screenshot of a data table

Description automatically generated

**Pogoda**

Dane to połączone raporty pogodowe miesięczne ze stacji NWS Seattle/Tacoma, WA dla całego roku 2022. Statystyki obliczone są dla każdego dnia, data jest zawarta w indeksie danych. Mamy trzy kolumny dotyczące temperatury – temperatura minimalna i maksymalna w trakcie dnia oraz średnia, obliczana jako średnia z wartości dwóch poprzednich kolumn. Następna kolumna dotyczy całkowitej sumy opadów atmosferycznych, w tym śniegu, w milimetrach. Dwie następne zmienne dotyczą śniegu: całkowita ilość opadów w milimetrach oraz głębokość śniegu zmierzona o godzinie 6:00. Dalej kolumny dotyczą średniej i maksymalnej prędkości wiatru w m/s, ostatnia kolumna to średnie zachmurzenie pomiędzy wschodem i zachodem słońca, wyrażone w ułamkach dnia (np. 0.1 oznacza, że zaobserwowane zachmurzenie trwało około 10% godzin dnia).

A screenshot of a data table

Description automatically generated

**Średnią liczbę rowerzystów przejeżdżających przez most w poszczególnych dniach tygodnia w sumie oraz z podziałem na zachodnią i wschodnią stronę mostu**

Na moście łączącym dzielnice Fremont i Queen Anne obserwujemy największy ruch rowerzystów we wtorki, a nieco mniejszy w środy i czwartki. Początkowo może to zaskakiwać, gdyż można by przypuszczać, że większość osób wybiera się na rower rekreacyjnie w weekendy. Jednak fakt, że most jest kluczowym połączeniem dla dojeżdżających do pracy, wyjaśnia ten trend. Zauważalny jest również spadek liczby rowerzystów w poniedziałki i piątki. Istnieje prawdopodobieństwo, że w dni te ze względu na około weekendowość część osób pracuje zdalnie. Dodatkowo, większość ruchu rowerowego koncentruje się na zachodniej stronie mostu.

A screenshot of a computer

Description automatically generated

**Sumaryczna liczba rowerzystów przejeżdżających przez most w poszczególnych dniach roku w sumie oraz z podziałem na zachodnią i wschodnią stronę mostu przez pierwsze dziesięć dni stycznia 2022 roku**

A table with numbers and numbers

Description automatically generated

**Średnia liczba rowerzystów przejeżdżających przez most w poszczególnych godzinach w sumie oraz z podziałem na zachodnią i wschodnią stronę mostu, a także z podziałem na dni robocze i weekend**

A screenshot of a graph

Description automatically generated

# **Wnioski na podstawie wcześniej wklejonych ramek oraz wykresu**

Zaczynając od tego co zostało powiedziane już w poprzednim punkcie, wyróżnia się tendencja do wybierania zachodniej strony mostu. Jej przyczyny nie są jednoznacznie wyjaśnione w ogólnodostępnych źródłach, co prowadzi do różnych hipotez. Możliwe, że zachodnia część mostu jest lepiej przystosowana dla rowerzystów, na przykład poprzez szersze pasy. Inna teoria sugeruje, że usytuowanie przejść na drugą stronę mostu może wpływać na wybór tej trasy, jak również lokalizacja ważnych punktów docelowych dla użytkowników tego szlaku. Nie można wykluczyć również technicznej usterki czujnika na wschodniej stronie mostu. Aby zweryfikować te przypuszczenia, warto by było przeprowadzić analizę porównawczą dla danych z różnych lat oraz dokładniej zbadać lokalizacje, do których najczęściej zmierzają użytkownicy. Ponadto, tendencja ta jest najbardziej wyraźna w dni robocze, zwłaszcza w godzinach szczytu, co może dawać dodatkowe wskazówki co do przyczyn preferencji tej strony mostu.

Ruch rowerowy na moście charakteryzuje się rocznym rozkładem przypominającym bardzo szeroką parabolę. Liczba rowerzystów jest mniejsza na początku roku, wzrasta w kolejnych miesiącach i utrzymuje się na stabilnym poziomie aż do jesieni, kiedy to zaczyna powoli spadać. Taki trend wydaje się być naturalnym wynikiem zmian temperatur, które bezpośrednio wpływają na preferencje co do wyboru środka transportu. Występują także dwie nietypowe zmiany liczby rowerzystów. Pierwsza miała miejsce w czwartek 07.04.2022, kiedy to odnotowano znaczący wzrost, mimo braku świąt czy ważnych wydarzeń. Tego dnia obchodzony jest jednak światowy dzień zdrowia, co mogło przyczynić się do akcji promujących jazdę na rowerze. Drugą nietypową obserwację zanotowano w niedzielę 03.07.2022, jednak nie odnaleziono informacji o wydarzeniu, które mogłoby wyjaśniać ten wzrost. Analiza pogody z tych dni pokazuje, że 07.04.2022 temperatura znacznie wzrosła, co mogło zachęcić do wyboru roweru jako środka transportu, natomiast 03.07.2022 odnotowano spadek temperatury i opady deszczu, co zwykle ogranicza chęć do aktywności na zewnątrz, zwłaszcza w niedzielę.

Tendencje do jazdy na rowerze w poszczególne dni tygodnia została omówiona w poprzedniej części. Analizując wykres, można również zauważyć, że zmiany w natężeniu ruchu rowerowego pomiędzy poszczególnymi dniami są znacznie łagodniejsze po wschodniej stronie mostu.

Przyjrzenie się rozkładowi liczby rowerzystów w poszczególnych godzinach wykazało zdecydowane różnice między dniami powszednimi a weekendami. W dni robocze szczególnie dużo rowerzystów obserwuje się w godzinach porannych 7-9 oraz popołudniowych 16-18, co jest związane z typowymi godzinami dojazdu do pracy czy szkoły oraz powrotami z tych miejsc. Natomiast w weekendy brak jest wyraźnych szczytów w ruchu rowerowym. Liczba rowerzystów układa się w rozkład bliski normalnemu z medianą przypadającą na godziny południowe.­

A screenshot of a graph

Description automatically generated

**Statystyki miesięczne pogody**

Dane dotyczące pogody zostały zagregowane w obrębie poszczególnych miesięcy. Temperatury wahają się bardzo, od niskich, ujemnych, w grudniu do relatywnie wysokich, maksymalnie aż 35 stopni w lipcu. Średnia temperatura wzrasta stopniowo od końca zimy do lata, osiągając szczyt w lipcu i sierpniu, po czym zaczyna spadać wraz z nadejściem jesieni. Lato charakteryzuje się bardzo niskimi opadami. Największe opady występują w grudniu, z czego ponad połowa to opady śniegu. Śnieg pada tylko dwa miesiące pod koniec roku – lekko w listopadzie i następnie w grudniu, a utrzymuje się aż do stycznia. Jeśli chodzi o wiatr jest on chyba najbardziej stabilną zmienną w ciągu roku. Jego wartości wahają się w okolicach 3-4 m/s. Widać nieznaczny wzrost prędkości w okresie jesienno-zimowym. Pokrywa chmur jest zróżnicowana ale ogólnie dość wysoka. Widać znaczne przejaśnienia zaczynające się od lipca. Pod koniec roku jednak zachmurzenie gwałtownie wzrasta i utrzymuje się na poziomie około 80% – jest to zgodne z sezonem zimowym na północnej półkuli.

A screenshot of a graph

Description automatically generated

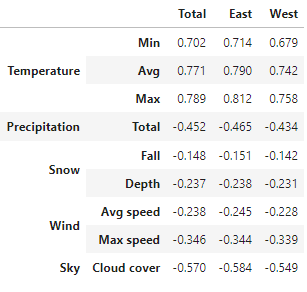
**Wizualizacja pogody**

W Seattle panuje umiarkowanie ciepły klimat. Lata są dość krótkie, ciepłe, suche oraz częściowo zachmurzone. Zimy są bardzo chłodne, mokre i znacznie zachmurzone. Tak więc miesiące zimowe stanowią przeciwieństwo do tych letnich.

Ciepła pora w roku 2022 trwała nieco dłużej niż zwykle. Wzrost temperatury nastąpił w czerwcu i trwał aż do października (zazwyczaj kończy się we wrześniu). Najgorętszym miesiącem był sierpień, kiedy to temperatura średnia wynosiła około 22 stopnie. Pora ta wyraźnie charakteryzuje się dużymi różnicami między temperaturą minimalną i maksymalną, dużo większymi niż w zimie. Może to wynikać z tego, że w stosunku do ciepłego popołudnia noce bywały dość zimne. Okres chłodny trwa nieco dłużej, od listopada do marca, oddzielony jest od ciepłego okresami przejściowymi. Najniższe temperatury występują w grudniu i często sięgają wartości ujemnych. Ogólny kształt rozkładu temperatury przypomina w dużym przybliżeniu szeroką parabolę. Kształt ten w pewnym stopniu odpowiada kształtowi ruchu rowerowego w przeciągu roku. Oznacza to, że istnieje wyraźna zależność pomiędzy tymi dwoma zmiennymi.

W mieście Seattle w ciągu roku widać znaczne sezonowe zmiany pod względem stopnia zachmurzenia. Przejaśnienia zaczynają się w połowie czerwca i pokrywają się z początkiem okresu ciepłego. Taki stan trwa do października. Najbardziej przejrzyste niebo można było zaobserwować na przełomie lipca i sierpnia. Okres ze znacznym zachmurzeniem zaczyna się od października i trwa do samego lata. Zachmurzenie utrzymuje się wtedy na znaczącym poziomie

Spore wahania dotyczą również opadów atmosferycznych. Okres ciepły, wraz z mniejszym zachmurzeniem, jest niezwykle suchy opady praktycznie wtedy nie występują (zwłaszcza w sierpniu). Opady rozpoczynają się w okolicach listopada i występują do czerwca. Widać też wyraźny skok w opadach na przełomie lutego i marca, na odznaczającym się poziomie 80 mm.



A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

**Korelacje liniowe pomiędzy parametrami pogody a liczbą rowerzystów**

Korelacja między zmiennymi jest miernikiem siły ich związku prostoliniowego. Znak korelacji informuje nas o kierunku zależności. Wartość dodatnia oznacza zmiany jednokierunkowe między cechami, natomiast ujemna zmiany różnokierunkowe (wraz ze wzrostem jednej zmiennej druga maleje). Wartość bliska zeru oznacza brak korelacji liniowej. Współczynnik determinacji liniowej informuje o tym, jaka część zmienności zmiennej Y jest wyjaśniona przez zmienność zmiennej X.

Najwyższymi współczynnikami determinacji pomiędzy ilością przejazdów przez most odznaczają się trzy zmienne dotyczące temperatury, z najwyższą wartością dla temperatury maksymalnej. Korelacja dla tych zmiennych jest dodatnia, co oznacza, że wraz ze wzrostem temperatury ilość rowerzystów wzrasta. Obserwacja ta jest zgodna z faktem, że wcześniej stwierdzone zostało wyraźne podobieństwo w kształcie wykresów dla ilości przejazdów i temperatur w skali roku 2022.

Wysoki współczynnik determinacji posiada zachmurzenie, co oznacza że zmienna ta wyjaśnia w dużym stopniu zmienność ilości przejazdów. Widzimy że jest to korelacja ujemna, czyli zależność odwrotna co w przypadku temperatury. Może to być związane z klimatem w Seattle – podczas okresu ciepłego, kiedy występuje więcej rowerzystów zachmurzenie jest mniejsze. W momencie kiedy zachmurzenie zaczyna wzrastać i rozpoczyna się okres chłodny liczba rowerzystów spada.

Ujemną korelację posiada również ilość opadów atmosferycznych. Co jest ciekawe ilość opadów śniegu ma już bardzo niewielki współczynnik determinacji, co oznacza, że jego zmiany nie wpływają na zmiany w ilości przejazdów. Cała zmienność w opadach śniegu zamyka się w okresie zimowym, w miesiącach listopad i grudzień. Wtedy też ilość opadów śniegu wzrasta i maleje. Natomiast liczba rowerzystów w tym okresie już jest niewielka i ma stały trend malejący, niezależny od wzrostów i spadków opadów śniegu. Natomiast sam deszcz wydaje się być zniechęcający dla rowerzystów.

Dla wiatru współczynnik determinacji wydaje się być najmniej znaczący. Jest to również najmniej zróżnicowana charakterystyka pogody w tym mieście. Wydaje się, że wiatr nie jest w stanie aż tak skutecznie zniechęcić potencjalnych rowerzystów.

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a color chart

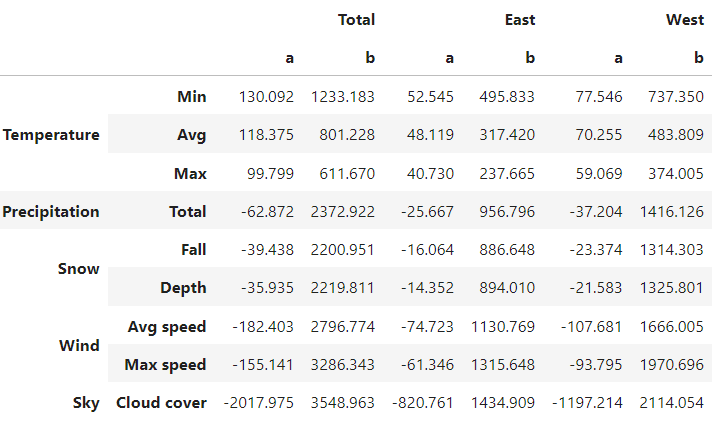
Description automatically generated

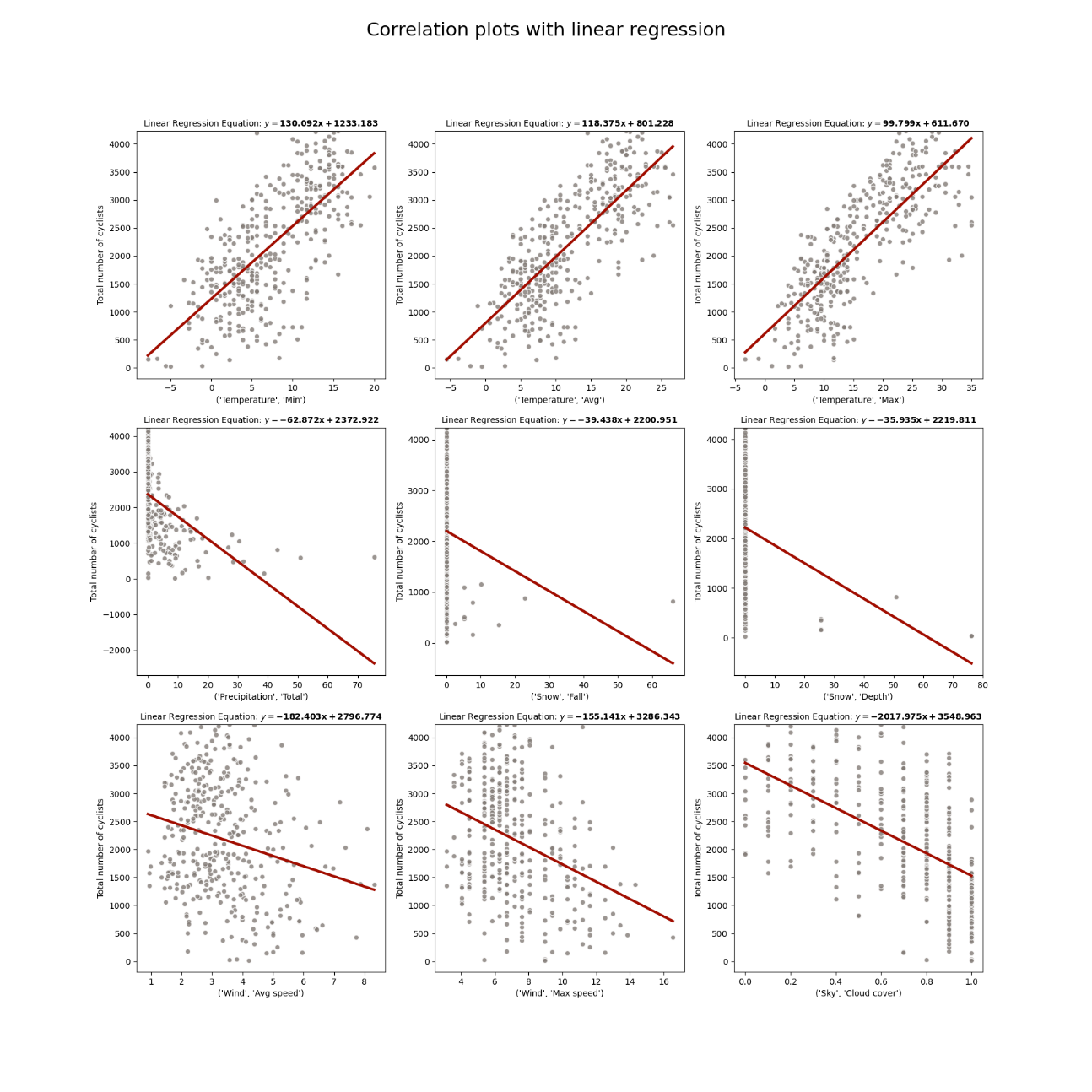
**Korelacje liniowe między parametrami pogody**

Pierwsze co rzuca się w oczy to najwyższe wartości współczynnika determinacji dla trzech zmiennych dotyczących temperatury. Jest to też najbardziej intuicyjny związek, zwłaszcza że możemy zapisać w sposób matematyczny tą zależność – avg = (min+max)/2 (zgodnie z tym co zawarte było w opisie danych). Oznacza to, że wraz ze wzrostem zarówno minimalnej jak i maksymalnej temperatury temperatura średnia będzie rosnąć. Zależność pomiędzy temperaturą minimalną i maksymalną może już nie jest aż tak ścisła, ale wciąż oczywista. Na podobnej zasadzie możemy stwierdzić dużą zależność pomiędzy wartością maksymalną prędkości wiatru a jego średnią.

Możemy też zauważyć, jak połączone są zmienne w kontekście występowania okresów ciepłych i chłodnych. Jak wspomniane zostało wcześniej, pora ciepła w Seattle charakteryzuje się wysoką temperaturą, znikomymi opadami oraz dużym zachmurzeniem, natomiast pora chłodna odznacza się przeciwnymi cechami. Znajduje to potwierdzenie w ujemnej korelacji pomiędzy temperaturą a zachmurzeniem, temperaturą i opadami oraz korelację dodatnią pomiędzy zachmurzeniem a opadami.

Znaczący współczynnik determinacji można zaobserwować też pomiędzy opadami śniegu a głębokością. Jest to dosyć logiczne – im więcej śniegu napada tym więcej go jest na ziemi. Te dwie zmienne są jednak najmniej skorelowane z całą resztą.





**Regresja Liniowa**

Współczynnik zostały dopasowane zarówno do liczby rowerzystów po wschodniej, zachodniej i obu stronach mostu. Widać między nimi tendencję - dla łącznej liczby rowerzystów są największe, a potem kolejno dla zachodniej i wschodniej. Jest to poprawny objaw, ponieważ wcześniej omówiliśmy tendencje co do większej liczby obserwacji po zachodniej stronie. W dalszej części szerzej zostanie omówione dopasowanie regresji dla łącznej liczby rowerzystów po obu stronach mostu. Wnioski są analogiczne, a wykresy, które zostały wygenerowane tylko dla całkowitej liczby rowerzystów dobrze to obrazują.

Uzależnienie liczby rowerzystów od temperatury wydaje się trafnym wyborem – ma pozytywny wpływ na liczbę rowerzystów. Funkcje zarówno dla średniej, maksymalnej i minimalnej temperatury wydają się być dopasowane dobrze.

Między opadami a liczbą rowerzystów widać ujemną relację. Jednak stosowanie regresji liniowej do modelowania tej zależności może być nieoptymalne, co sugeruje potrzebę zastosowania bardziej złożonych metod statystycznych, które lepiej odzwierciedlą nieliniową naturę tego związku.

Objaśnianie ilości opadów śniegu oraz jego głębokości wydaje się być ciężkie do wymodelowania za pomocą regresji. Istnieje ujemna zależność jednak nie jest ona liniowa. Jest to bardziej tendencyjność. Wydaje się, że dane nie są wystarczająco dobrze zbalansowane pod tym kątem (w Seattle nie pada aż tak często śnieg) aby tworzyć model regresji liniowej na ich podstawie.

Średnia prędkość wiatru oraz maksymalna prędkość wiatru mają także negatywny wpływ na liczbę rowerzystów. Dane wskazują, że silniejszy wiatr może zniechęcać rowerzystów, jednak duże zróżnicowanie wartości sugeruje, że prosta regresja liniowa może być tu zbyt uproszczonym podejściem.

Większe zachmurzone wpływa na obniżenie liczby rowerzystów. Analiza zachmurzenia jest mierzona w skali co 10%. Pełniejsze wykorzystanie skali 0-100% mogłoby pozwolić na lepsze zrozumienie tej zależności.

A screenshot of a calculator

Description automatically generated

A graph of a curve

Description automatically generated with medium confidence

**Regresja Krzywoliniowa**

Użycie regresji krzywoliniowej do modelowania zależności między liczbą rowerzystów, a zachmurzeniem i opadami jest lepszym rozwiązaniem w szczególności dla tego drugiego. Funkcja wpasowała się w dane dobrze i odzwierciedla naturę tej zależności, w której od pewnej ilości opadów liczba rowerzystów utrzymuje się na zbliżonym poziomie. W przypadku zachmurzenia, choć zastosowanie modelu krzywoliniowego nie przynosi tak spektakularnych rezultatów, jak w przypadku opadów, nadal jest to bardziej adekwatne podejście niż prosta regresja liniowa.

**Uwagi do analizy**

* do lepszego zrozumienia tendencji przydatne wydaje się również przeanalizowanie danych z innych lat, które mogłyby utwierdzić nas w pewnych założeniach i dałaby bardziej szerokie spojrzenie
* Nie jest to problem bezpośrednio związany z samą analizą, a bardziej rzeczywistość, z którą będziemy musieli się godzić zapewne nie raz w pracy z danymi, ale nasuwa się ciekawość czym była spowodowana tendencja w preferencji zachodniej strony mostu. Ze względu na jego lokalizację ciężko jest mieć jakieś intuicyjnie przypuszczenia na ten temat nie będąc tam nigdy. Przykładowo, gdyby była mowa o moście Dębnickim i okazałoby się, że więcej osób przejeżdża rowerem po lewej stronie (idąc od AGH) nasuwałoby się, że jest to bliżej Wawelu i zazwyczaj więcej osób spędza czas właśnie po tej stronie
* Przydatne mogłoby okazać się stworzeniu również innych modeli regresji, przykładowo wielorakiej, aby zaobserwować łączny wpływ większej liczby czynników