



Analiza danych o użyciu rowerów miejskich / skuterów elektrycznych

Eksploracja Danych, 2023

ANNA GUT, JAKUB JANICKI, MIKOŁAJ ZASADA

Analiza danych o użyciu rowerów miejskich / skuterów elektrycznych

Anna Gut, Jakub Janicki, Mikołaj Zasada

Czerwiec, 2023



Streszczenie

Celem tego projektu jest przeprowadzenie analizy danych dotyczących użytkowania rowerów i miejskich skuterów elektrycznych w celu uzyskania zaawansowanych statystyk ruchu jednośladów oraz przedstawienia wyników wizualizacyjnych na mapie. Projekt wykorzystuje różnorodne zbiory danych związane z lokalizacją, przemieszczaniami i ostatnią aktywnością pojazdów. Poprzez głębową analizę tych danych, badane będą różne aspekty, takie jak trendy dobowe w użytkowaniu jednośladów, rozkład wielkości przemieszczeń oraz lokalizowanie obszarów o największej i najmniejszej aktywności transportowej. Wyniki analizy danych zostaną zaprezentowane w formie wizualizacji na mapie, co pozwoli na łatwą interpretację zebranych informacji. Przedstawienie tych wyników może dostarczyć cennych wskazówek i perspektyw dla planowania rozwoju infrastruktury rowerowej oraz skuterów elektrycznych w danym obszarze.

Spis treści

1 Wstęp	4
1.1 Opis zbioru danych	4
1.1.1 City Bikes Trips 2015 - 2017 Jersey City	4
1.1.2 City Bikes Trips 2016 Philadelphia	5
1.1.3 E-Scooter Trips 2020 Chicago	5
1.1.4 E-Scooter Trips 2018-2019 Louisville	5
1.2 Preprocessing	6
1.3 Cel i zakres prac	6
1.4 Przegląd literatury i dostępnych rozwiązań	7
2 Wykorzystane biblioteki i algorytmy	8
2.1 Narzędzia	8
2.2 Biblioteki	8
2.3 Algorytmy	9
3 Analiza danych	10
3.1 Liczba wynajmów na przestrzeni czasu	11
3.2 Liczba wynajmów w danym dniu tygodnia	12
3.3 Liczba wynajmów w danej godzinie	13
3.4 Liczba wynajmów w zależności od wieku wynajmującego	14
3.5 Liczba wynajmów w zależności od płci wynajmującego	15
3.6 Czas przejazdu w zależności od daty	16
3.7 Czas przejazdu w danym dniu tygodnia	17
3.8 Czas przejazdu w danej godzinie dnia	18
3.9 Liczba przejazdów w pierwszych i ostatnich piętnastu dniach wynajmu	19
3.10 Średni czas przejazdu w pierwszych i ostatnich piętnastu dniach wynajmu	19
3.11 Ilość wynajmów w danym dniu tygodnia w pierwszych i ostatnich piętnastu dniach	20
3.12 Ilość wynajmów w danej godzinie dnia w pierwszych i ostatnich piętnastu dniach	21
3.13 Procentowa ilość tras o takiej samej stacji końcowej i początkowej	22
4 Wizualizacje tras podróży	23
4.1 Liczba rozpoczętych podróży w zależności od lokalizacji	23
4.2 Wizualizacja różnic ilości przyjazdów i odjazdów w zależności od pory dnia	25

4.3 Lokalizacje z największą liczbą rozpoczętych/zakończonych podróży	30
4.4 Mapy przepływu	35
5 Model predykcyjny	39
5.1 Analizowane parametry przejazdu	39
5.2 Porównanie wyników predykcji	40
6 Podsumowanie	42
6.1 Obserwacje	42
6.1.1 Zależność liczby przejazdów od pory roku	43
6.1.2 Zależność liczby oraz czasu przejazdów od dnia tygodnia	44
6.1.3 Zależność liczby oraz czasu przejazdów od godziny	44
6.1.4 Zależność liczby przejazdów od wieku oraz płci wynajmującego	45
6.1.5 Procentowa ilość tras o takiej samej stacji końcowej i początkowej	45
6.1.6 Identyfikacja obszarów przejazdu o największej aktywności	46
6.2 Wnioski	47
6.3 Kod źródłowy	47

1 Wstęp

Jako temat badań wybrano *Analiza danych o użyciu rowerów / miejskich skuterów elektrycznych*. Praca wykonywana będzie w oparciu istniejące zbiory danych opisujących parametry korzystania z usługi wynajmu krótkoterminowego miejskich pojazdów dwukołowych. Celem będzie analiza zbioru danych oraz stworzenie możliwe skutecznego i wydajnego modelu predykcyjnego umożliwiającego klasyfikacje niektórych parametrów przejazdów.

1.1 Opis zbioru danych

Analizie oraz porównaniu podlega kilka zbiorów danych. Przedstawiono ich podział ze względu na poniższe parametry:

- Okres, w którym zbierano dane
- Ilość rekordów
- Prezentowane kolumny

1.1.1 City Bikes Trips 2015 - 2017 Jersey City

Zawiera informacje o wypożyczanych rowerach miejskich w Jersey City. (1)

- Okres od 21.09.2015 do 31.03.2017
- 735502 rekordów
- Kolumny: ID, czas trwania (w sekundach), dzień i godzina rozpoczęcia, dzień i godzina zakończenia, ID stacji startu, nazwa stacji startu, szerokość geograficzne stacji startu, długość geograficzna stacji startu, ID stacji końcowej, nazwa stacji końcowej, szerokość geograficzna stacji końcowej, długość geograficzna stacji końcowej, ID roweru, typ użytkownika (subscriber, customer), rok urodzenia użytkownika, płeć użytkownika

1.1.2 City Bikes Trips 2016 Philadelphia

Zawiera informacje o wypożyczanych rowerach miejskich w Philadelphii. (2)

- Okres od 01.04.2016 do 30.06.2016
- 170824 rekordów
- Kolumny: ID, czas trwania (w sekundach), dzień i godzina rozpoczęcia, dzień i godzina zakończenia, ID stacji startu, szerokość geograficzna stacji startu, długość geograficzna stacji startu, ID stacji końcowej, szerokość geograficzna stacji końcowej, długość geograficzna stacji końcowej, ID roweru, ilość dni na które wykupiony został pakiet, kategoria tripu (round trip, one way), nazwa wykupionego planu, typ roweru

1.1.3 E-Scooter Trips 2020 Chicago

Zawiera informacje o wypożyczanych hulajnogach w Chicago (3)

- Okres od 12.08.2020 do 26.11.2020
- 630816 rekordów
- Kolumny: ID, dzień i godzina rozpoczęcia, dzień i godzina zakończenia, dystans (w kilometrach), czas trwania (w sekundach), typ roweru, ID stacji startu, ID stacji końcowej, nazwa stacji startu, nazwa stacji końcowej, szerokość geograficzna stacji startu, długość geograficzna stacji startu, obydwie współrzędne geograficzne stacji startu, szerokość geograficzna stacji końcowej, długość geograficzna stacji końcowej, obydwie współrzędne geograficzne stacji końcowej

1.1.4 E-Scooter Trips 2018-2019 Louisville

Zawiera informacje o wypożyczanych hulajnogach w Louisville (4)

- Okres od 09.08.2018 do 27.10.2019
- 434582 rekordów
- Kolumny: ID, dzień rozpoczęcia, godzina rozpoczęcia, dzień zakończenia, godzina zakończenia, czas trwania (w minutach), dystans (w kilometrach), szerokość geograficzna stacji startu, długość geograficzna stacji startu, szerokość geograficzna stacji końcowej, długość geograficzna stacji końcowej, dzień tygodnia, godzina

1.2 Preprocessing

Na każdy z nich nałożono filtr, który wyklucza trasy z drogą krótszą niż 150 metrów oraz czasem trwania krótszym niż 1.5 minuty. Odrzucono także rekordy z wartościami pustymi w kluczowych kolumnach takich jak: czas trwania, dystans, współrzędna końcowa, współrzędna początkowa.

1.3 Cel i zakres prac

Celem tego projektu było przeprowadzenie kompleksowej analizy danych dotyczących użytkowania rowerów i miejscowych skuterów elektrycznych. Zakres prac obejmował gromadzenie danych z różnych źródeł, takich jak zbiory danych udostępnione przez miasta Chicago, Louisville, Jersey City oraz Philadelphia. Zebrane dane obejmowały informacje o lokalizacji, przemieszczeniach i ostatniej aktywności pojazdów.

Następnie przeprowadzono analizę danych w celu uzyskania zaawansowanych statystyk ruchu jednośladow, takich jak trendy dobowe oraz rozkład wielkości przemieszczeń. Ponadto, dokonano identyfikacji obszarów o największej i najmniejszej aktywności transportowej w badanych miastach.

W kolejnym etapie projektu przeprowadzono wizualizację przepływu danych na mapie, aby w bardziej czytelny sposób przedstawić zebrane informacje. Wykorzystano różne techniki wizualizacji, takie jak mapy cieplne, wykresy czasowe oraz interaktywne mapy, aby w pełni uchwycić dynamikę i charakterystykę ruchu jednośladow w badanych obszarach.

Dodatkowo, jako część projektu, stworzono model predykcyjny, który mógłby przewidywać przyszłe trendy użytkowania rowerów i skuterów elektrycznych na podstawie zebranych danych. Model ten opierał się na zaawansowanych technikach analizy danych i uczenia maszynowego, takich jak regresja czy algorytmy klasifikacji.

W rezultacie, ten projekt pozwolił na głębsze zrozumienie i wykorzystanie danych dotyczących użytkowania jednośladow, wizualizację i prezentację wyników oraz rozwinięcie modelu predykcyjnego, który mógłby mieć praktyczne zastosowanie w planowaniu i optymalizacji systemów miejskiej mobilności.

1.4 Przegląd literatury i dostępnych rozwiązań

Przegląd literatury oraz dostępnych rozwiązań w temacie analizy danych dotyczących użytkowania rowerów i miejscowych skuterów elektrycznych wykazuje, że jest to obszar intensywnie badany ze względu na rosnące zainteresowanie ekologicznymi środkami transportu i rozwojem infrastruktury miejskiej mobilności.

W literaturze naukowej można znaleźć wiele publikacji dotyczących analizy danych rowerowych i skuterowych. Często stosowane metody obejmują analizę trendów użytkowania w różnych porach dnia, sezonach i dniach tygodnia, co pozwala na identyfikację wzorców i preferencji użytkowników. Analizuje się również czynniki wpływające na popularność jednośladów, takie jak pogoda, dostępność stacji czy koszty korzystania z usług.

W zakresie dostępnych rozwiązań technologicznych istnieje wiele platform i narzędzi do analizy danych miejskiej mobilności. Przykłady to platformy takie jak Google Maps, które oferują funkcje śledzenia trasy, szacowania czasu podróży rowerem oraz analizy popularnych tras rowerowych w danym obszarze. Istnieją także dedykowane aplikacje mobilne dla użytkowników jednośladów, które zbierają dane o lokalizacji, czasie i długości podróży, a następnie analizują je w celu udostępnienia statystyk i rekomendacji.

W kontekście modelowania predykcyjnego istnieją różne podejścia, takie jak modele regresji, modele szeregow czasowych, metody uczenia maszynowego i sztuczne sieci neuronowe. Celem tych modeli jest prognozowanie przyszłego użytkowania jednośladów na podstawie historycznych danych, co może mieć zastosowanie w planowaniu i optymalizacji systemów transportowych.

Pośród najbardziej popularnych artykułów naukowych w badanym temacie możemy wyróżnić:

- *Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns* (5) - praca ta analizuje różne aspekty systemów rowerów publicznych, w tym trendy użytkowania, czynniki wpływające na popularność, profile użytkowników i wpływ na transport miejski
- *Shifting to Shared Wheels: Factors Affecting Dockless Bike-Sharing Choice for Short and Long Trips* (6) - w tej publikacji badane są intencje użytkowników dotyczące korzystania z systemów rowerów miejskich w porównaniu do tradycyjnych konkurencyjnych środków transportu - samochodów prywatnych, autobusów i pieszych.

2 Wykorzystane biblioteki i algorytmy

W ramach tego projektu wykorzystano różne narzędzia, biblioteki i algorytmy do przeprowadzenia analizy danych, wizualizacji przepływu oraz tworzenia modelu predykcyjnego. Dzięki nim możliwe było uzyskanie wglądu w trendy, preferencje użytkowników oraz przewidywanie przyszłego użytkowania rowerów i miejscowych skuterów elektrycznych.

2.1 Narzędzia

- Jupyter Notebook (7) - został użyty do tworzenia interaktywnych notatników, w których przeprowadzono analizę danych i wizualizację wyników
- Python IDEs (Integrated Development Environments) - takie jak PyCharm (8) czy Visual Studio Code (9), były wykorzystane do tworzenia i zarządzania skryptami analizy danych

2.2 Biblioteki

- Pandas (10) - została użyta do manipulacji i przetwarzania danych, takich jak filtrowanie, agregacja czy grupowanie
- NumPy (11) - została wykorzystana do wykonywania operacji numerycznych i obliczeń na danych
- Matplotlib (12) i Seaborn (13) - te biblioteki wizualizacyjne umożliwiły tworzenie różnorodnych wykresów, histogramów, map cieplnych i innych wizualizacji danych
- Plotly (14) - ta biblioteka została wykorzystana do tworzenia interaktywnych wykresów i wizualizacji danych.
- Scikit-learn (15) - jest popularną biblioteką do uczenia maszynowego w języku Python. Była używana do tworzenia modelu predykcyjnego
- GeoPy (16) - ta biblioteka została wykorzystana do obsługi operacji geolokalizacji i odwzorowania danych geograficznych

2.3 Algorytmy

- DecisionTreeClassifier (17)

DecisionTreeClassifier to algorytm drzewa decyzyjnego stosowany w zadaniach klasyfikacji. Opiera się na tworzeniu drzewa decyzyjnego, gdzie węzły reprezentują cechy, a krawędzie reprezentują reguły decyzyjne. Algorytm podejmuje decyzje na podstawie wartości cech, które są przekazywane przez drzewo, aż do osiągnięcia liścia, który zawiera przypisaną etykietę klasy. Drzewa decyzyjne są łatwe do interpretacji i mogą być skuteczne w rozwiązywaniu problemów klasyfikacji

- KNeighborsClassifier (18)

KNeighborsClassifier to algorytm k-najbliższych sąsiadów stosowany w zadaniach klasyfikacji. Opiera się na idei, że obiekty o podobnych cechach sąsiedztwa mają tendencję do przypisywania tych samych klas. Algorytm szuka k najbliższych sąsiadów dla nowego punktu danych i przypisuje mu klasę, która jest najczęściej występująca wśród tych sąsiadów. Algorytm k-najbliższych sąsiadów jest prosty w implementacji i może być skuteczny w problemach klasyfikacji

- RandomForestClassifier (19)

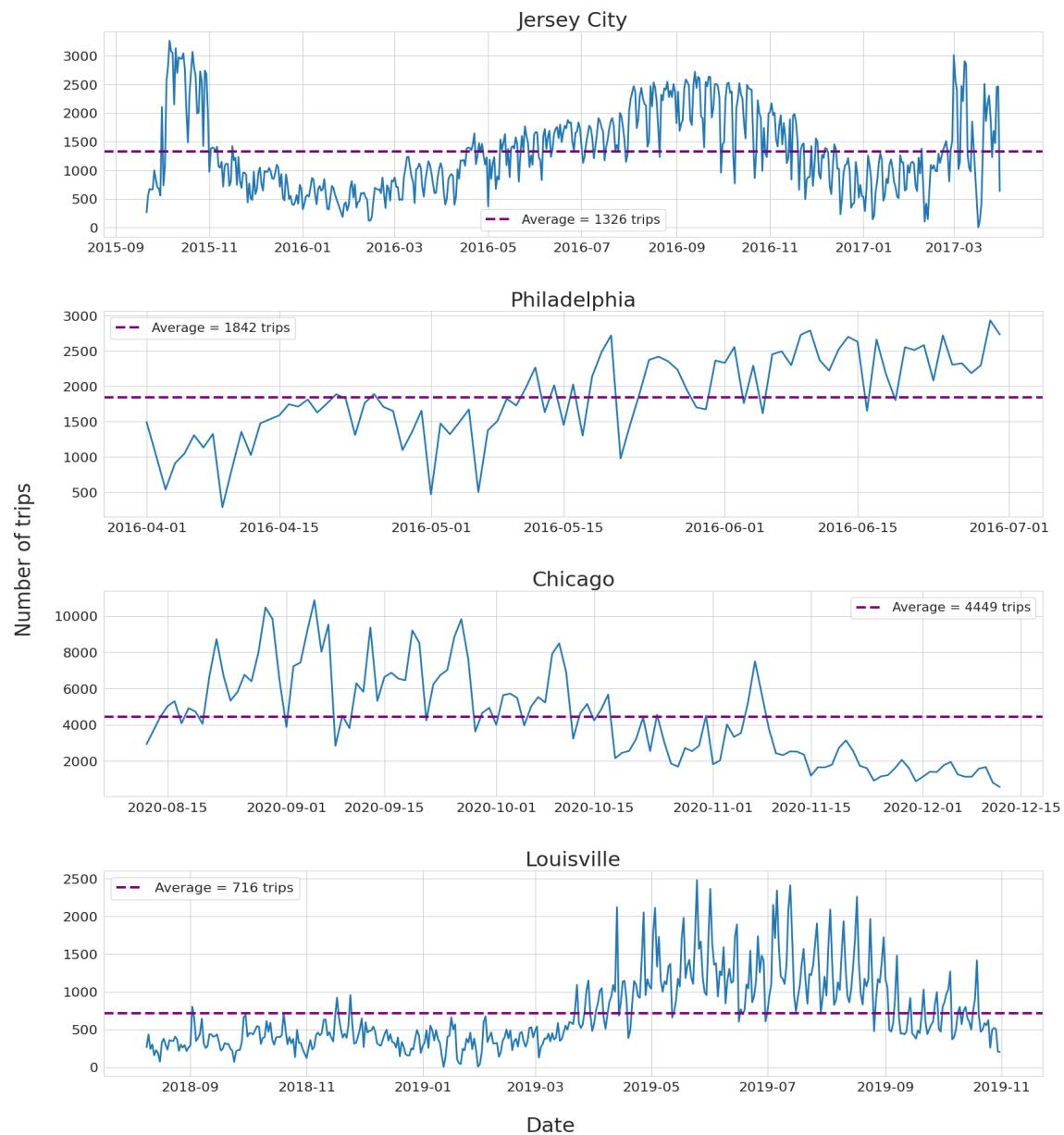
RandomForestClassifier to algorytm lasów losowych, który łączy wiele drzew decyzyjnych w celu poprawy skuteczności klasyfikacji. Algorytm losowo wybiera podzbior cech i danych treningowych, a następnie tworzy wiele drzew decyzyjnych. Kiedy następuje klasyfikacja, każde drzewo oddzielnie przewiduje wynik, a ostateczna predykcja jest dokonywana na podstawie głosowania większościowego. Las losowy jest odporny na overfitting i ma zastosowanie w zadaniach klasyfikacji, gdzie ma wiele cech i obserwacji

3 Analiza danych

Sekcja zawiera analizę danych oraz przedstawienie szczegółowych prawidłowości, które zostały podczas niej zaobserwowane.

3.1 Liczba wynajmów na przestrzeni czasu

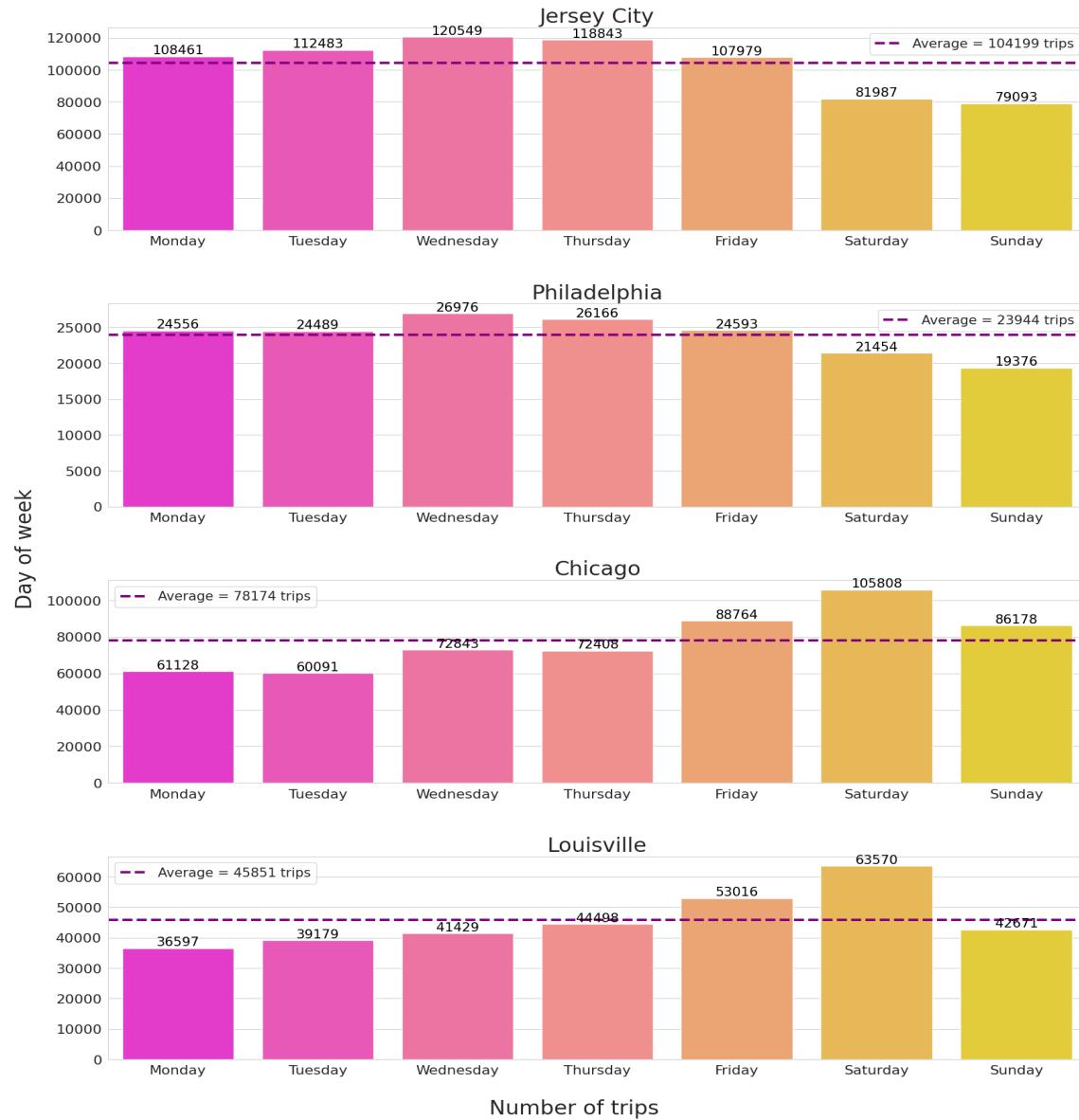
Na wykresie przedstawiono liczbę wynajmów w okresie zbierania danych oraz średnią dzienną wartość.



Rysunek 1: Wykres zależności liczby wynajmów od czasu dla określonych miast

3.2 Liczba wynajmów w danym dniu tygodnia

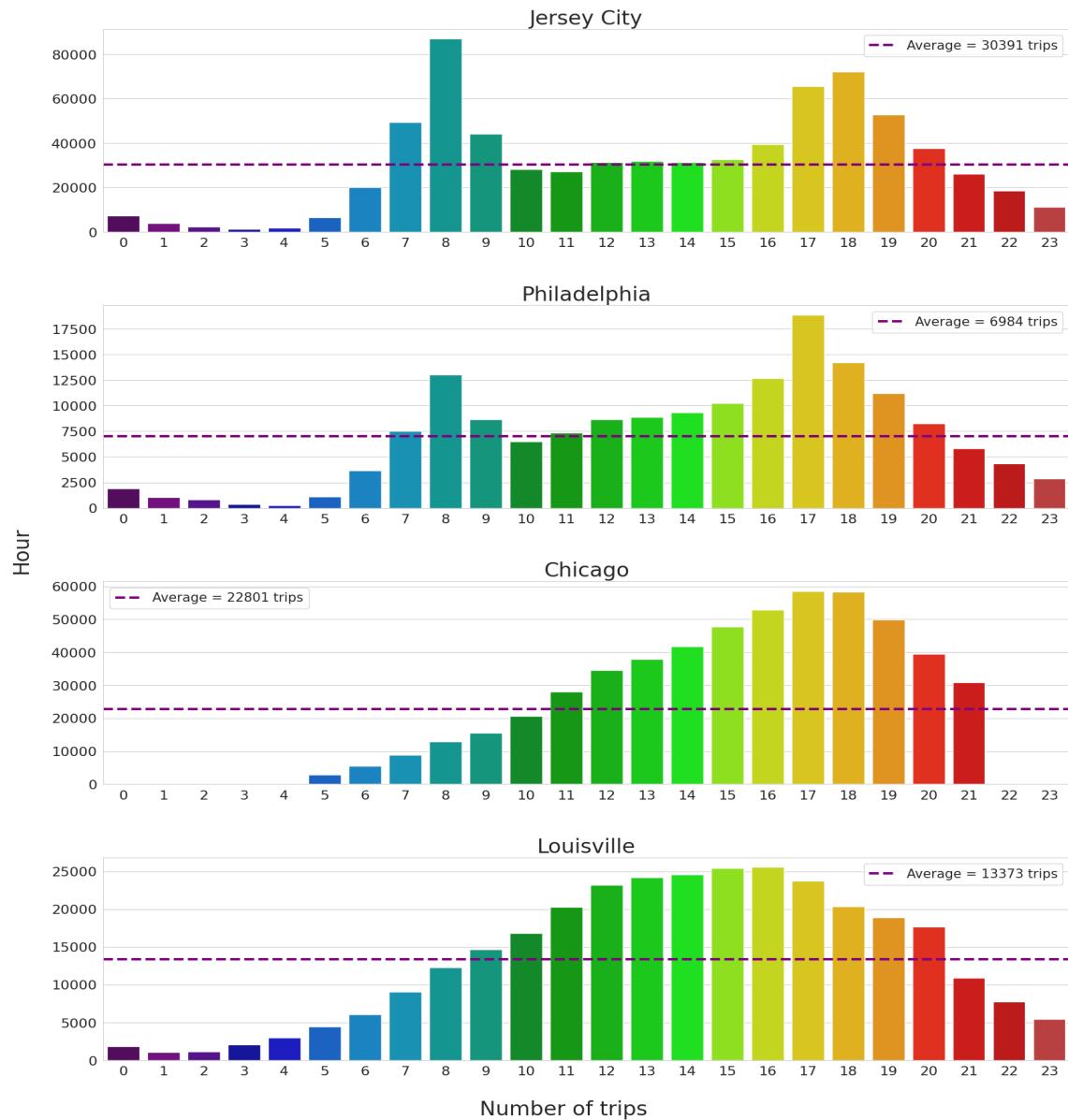
Na wykresie przedstawiono skumulowaną liczbę wynajmów w danym dniu tygodnia oraz średnią wartość danych.



Rysunek 2: Wykres zależności liczby wynajmów od dnia tygodnia dla określonych miast

3.3 Liczba wynajmów w danej godzinie

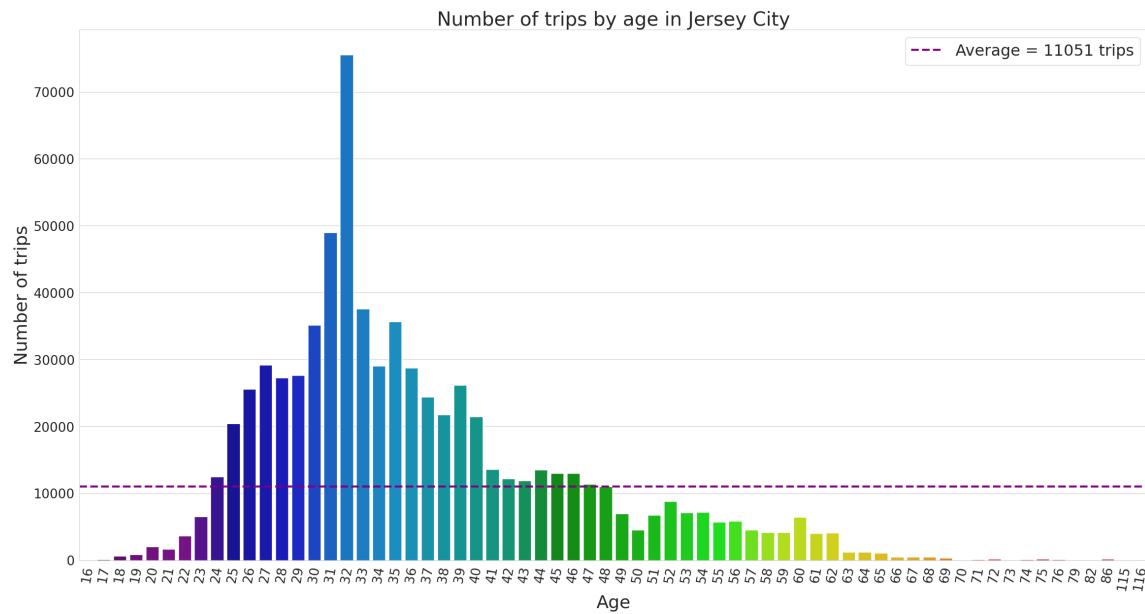
Na wykresie przedstawiono skumulowaną liczbę wynajmów w danej godzinie oraz średnią wartość danych.



Rysunek 3: Wykres zależności liczby wynajmów od danej godziny dla określonych miast

3.4 Liczba wynajmów w zależności od wieku wynajmującego

Na wykresie przedstawiono liczbę wynajmów w zależności od wieków wynajmującego oraz średnią wartość danych dla Jersey City.



Rysunek 4: Wykres zależności liczby wynajmów od wieku dla Jersey City

3.5 Liczba wynajmów w zależności od płci wynajmującego

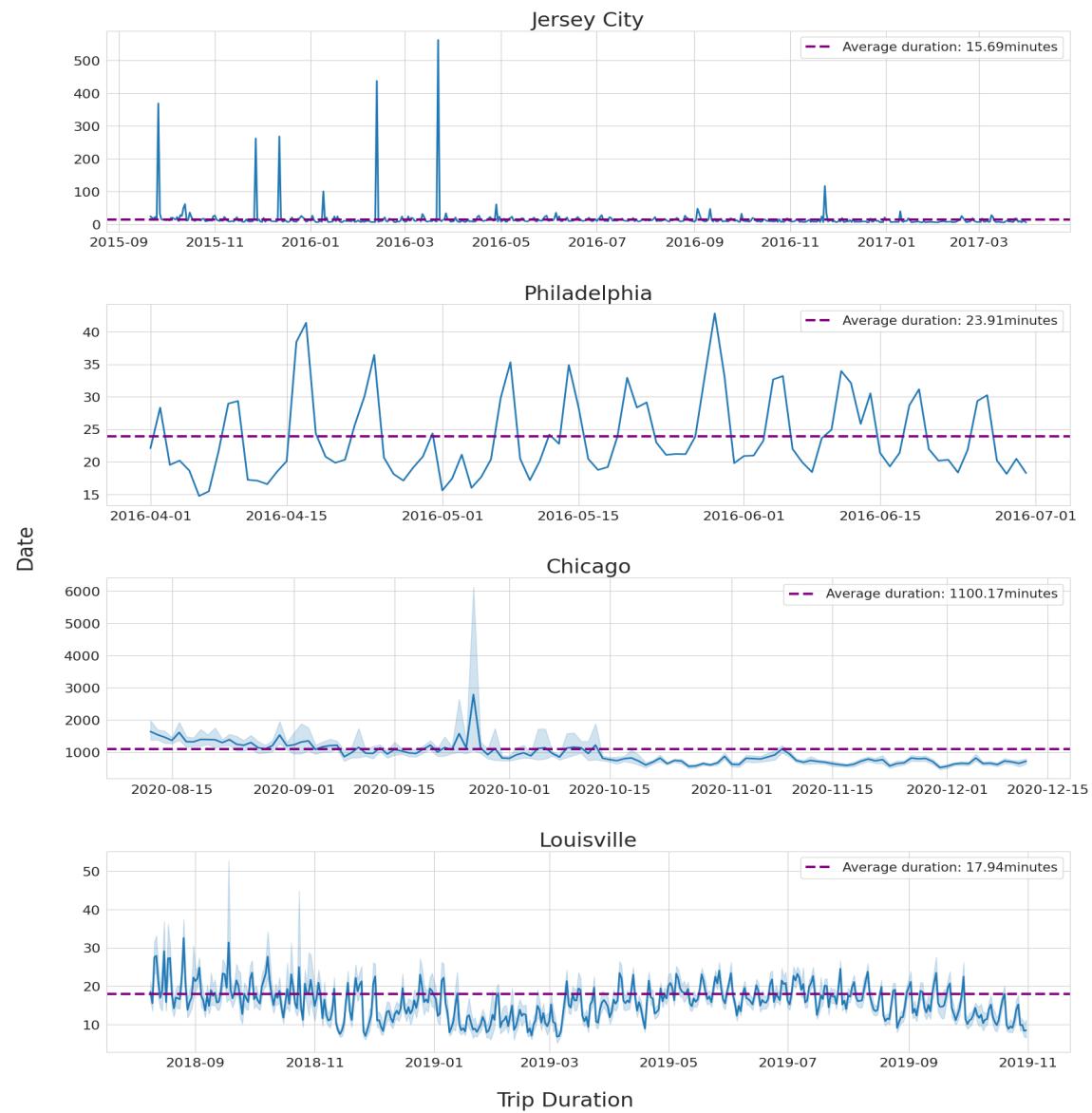
Na wykresie przedstawiono liczbę wynajmów w zależności od płci wynajmującego dla Jersey City.



Rysunek 5: Wykres zależności liczby wynajmów od płci dla Jersey City

3.6 Czas przejazdu w zależności od daty

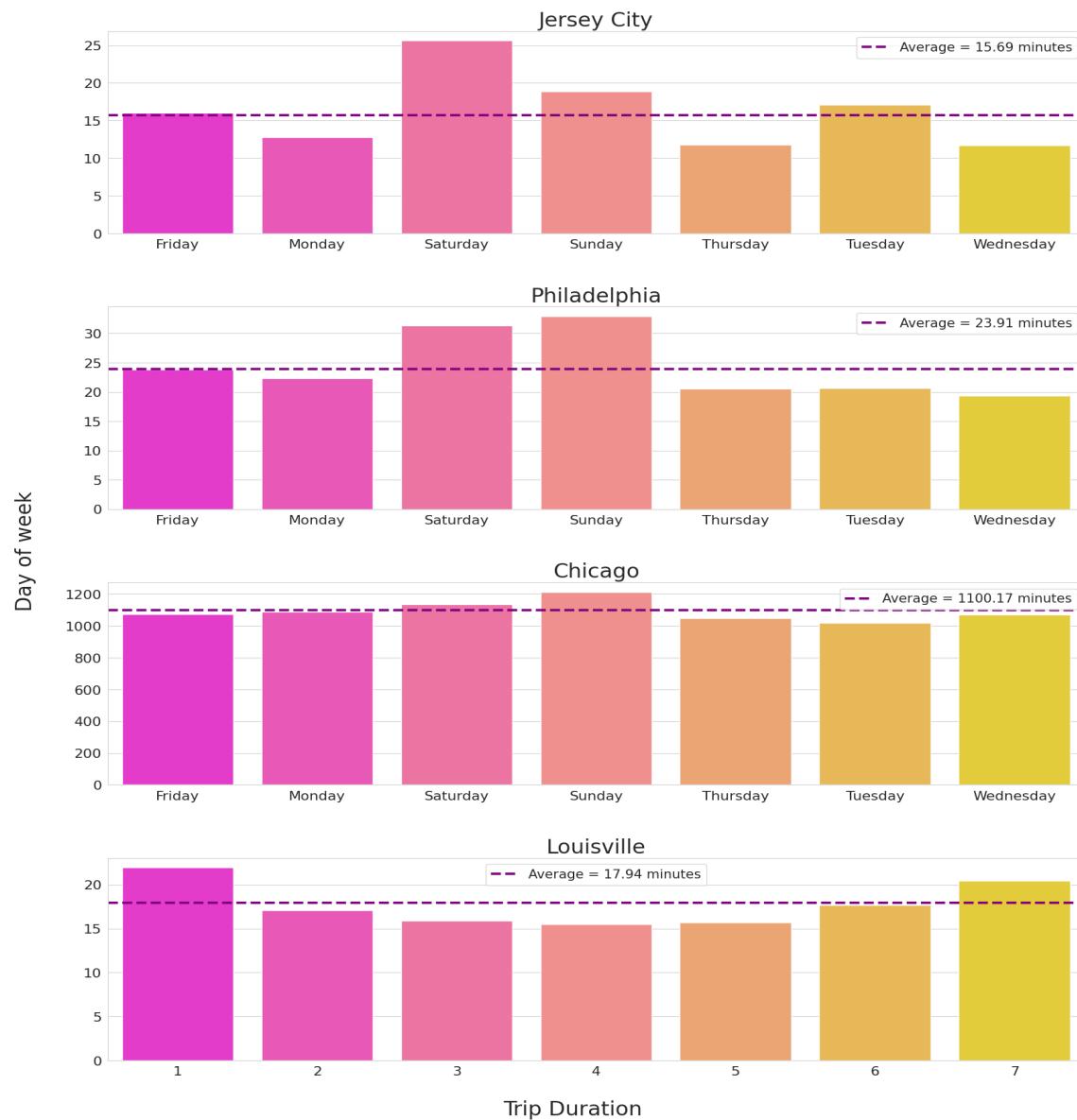
Na wykresie przedstawiono czas przejazdu w zależności od daty oraz średnią wartość danych dla określonych miast.



Rysunek 6: Wykres zależności czasu przejazdu od daty

3.7 Czas przejazdu w danym dniu tygodnia

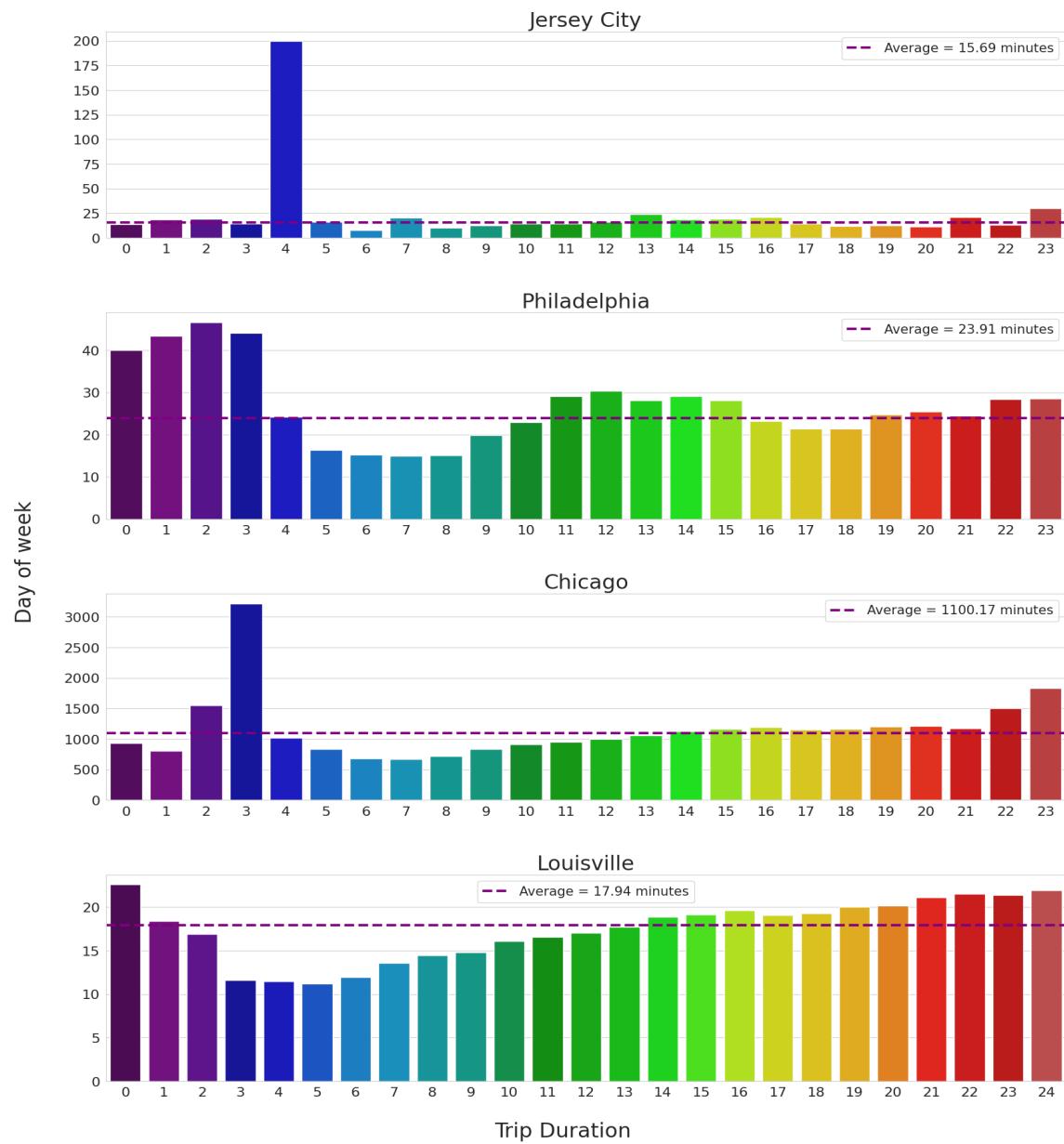
Na wykresie przedstawiono czas przejazdu w zależności od dnia tygodnia dla określonych miast.



Rysunek 7: Wykres zależności czasu przejazdu od dnia tygodnia

3.8 Czas przejazdu w danej godzinie dnia

Na wykresie przedstawiono czas przejazdu w danej godzinie dnia oraz średnią wartość danych dla określonych miast.



Rysunek 8: Wykres zależności czasu przejazdu od godziny dnia

3.9 Liczba przejazdów w pierwszych i ostatnich piętnastu dniach wynajmu

W tabeli przedstawiono liczbę przejazdów w pierwszych i ostatnich piętnastu dniach wynajmu dla określonych miast.

Miasto	Pierwsze 15 dni	Ostatnie 15 dni
Jersey City	13089	22596
Philadelphia	15258	33857
Chicago	79187	28729
Louisville	3824	7840

Tabela 1: Liczba przejazdów w pierwszych i ostatnich piętnastu dniach

3.10 Średni czas przejazdu w pierwszych i ostatnich piętnastu dniach wynajmu

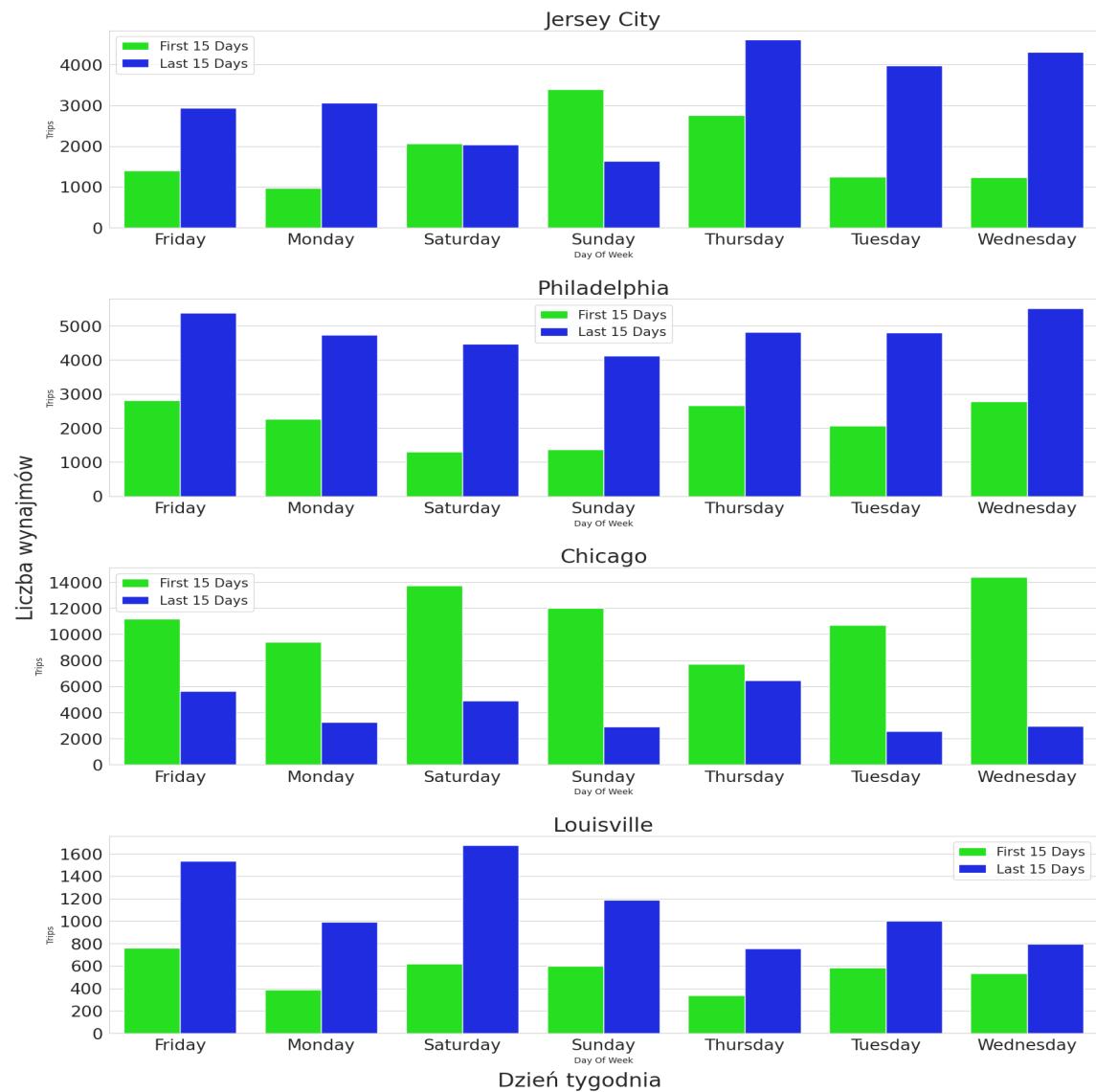
W tabeli średni czas przejazdu w pierwszych i ostatnich piętnastu dniach wynajmu dla określonych miast.

Miasto	Pierwsze 15 dni	Ostatnie 15 dni
Jersey City	44.27	10.39
Philadelphia	19.77	22.62
Chicago	24.1	12.45
Louisville	21.34	13.71

Tabela 2: Średni czas przejazdu w minutach w pierwszych i ostatnich piętnastu dniach

3.11 Ilość wynajmów w danym dniu tygodnia w pierwszych i ostatnich piętnastu dniach

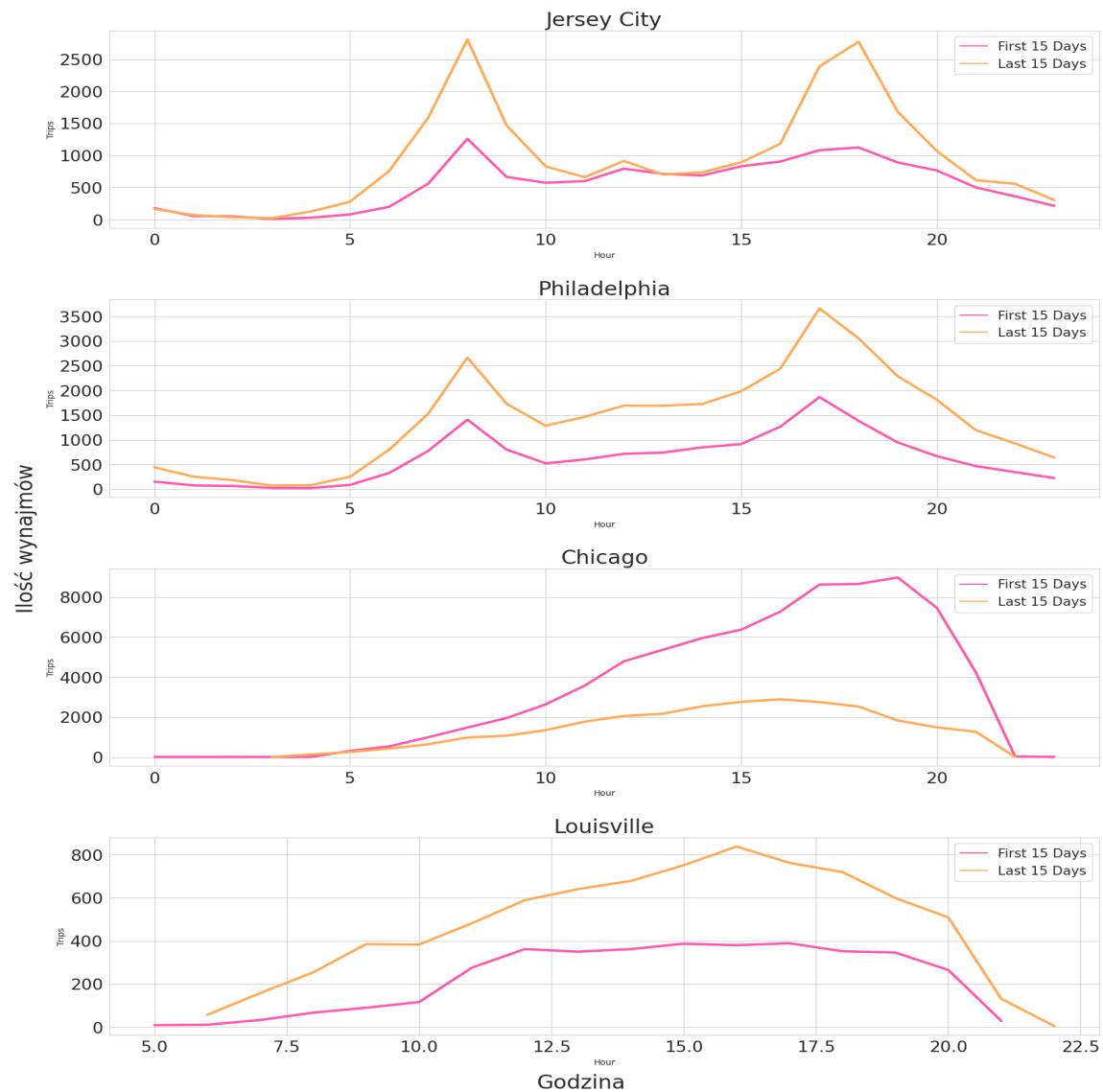
Na wykresie przedstawiono ilość wynajmów w danym dniu tygodnia w pierwszych i ostatnich piętnastu dniach dla określonych miast.



Rysunek 9: Wykres ilość wynajmów w danym dniu tygodnia w pierwszych i ostatnich piętnastu dniach

3.12 Ilość wynajmów w danej godzinie dnia w pierwszych i ostatnich piętnastu dniach

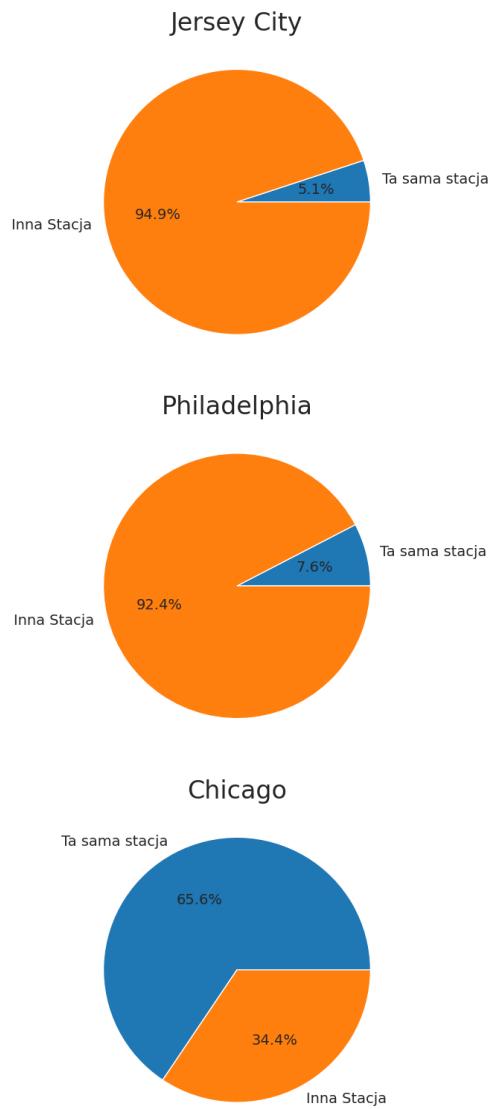
Na wykresie przedstawiono ilość wynajmów w danej godzinie dnia w pierwszych i ostatnich piętnastu dniach dla określonych miast.



Rysunek 10: Wykres ilość wynajmów w danej godzinie dnia w pierwszych i ostatnich piętnastu dniach

3.13 Procentowa ilość tras o takiej samej stacji końcowej i początkowej

Na wykresie przedstawiono procentową ilość tras o takiej samej stacji końcowej i początkowej dla określonych miast.



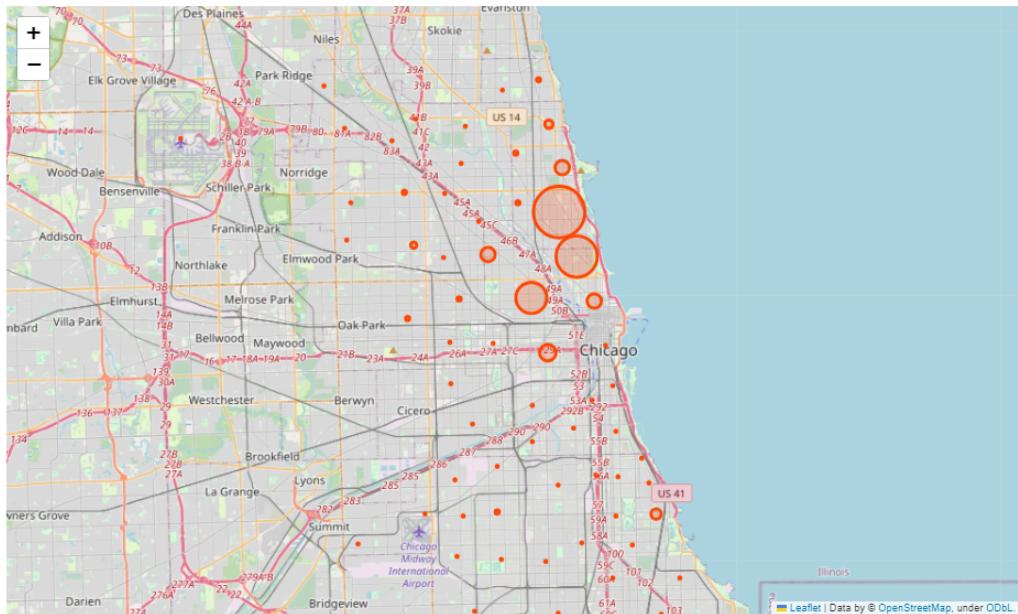
Rysunek 11: Wykres procentowej ilości tras o takiej samej stacji końcowej i początkowej

4 Wizualizacje tras podróży

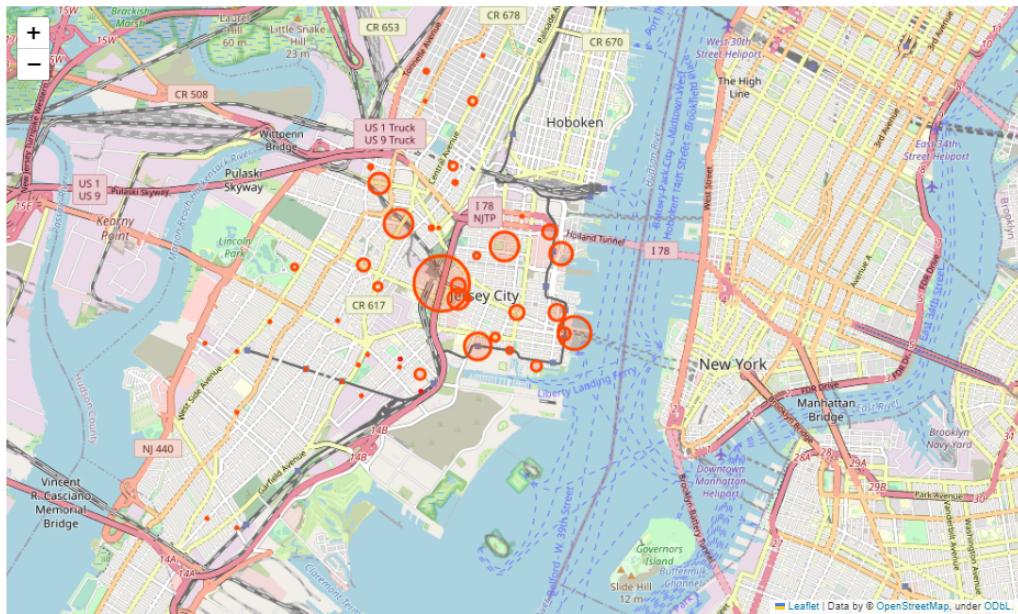
W tej sekcji skupimy się na prezentacji interaktywnych map, które przedstawiają trasy podróży na podstawie dostępnych danych o lokalizacji i czasie. Dzięki tym wizualizacjom będziemy mogli zapoznać się z dynamicznymi i realistycznymi obrazami przemieszczania się rowerów i skuterów elektrycznych w badanych obszarach.

4.1 Liczba rozpoczętych podróży w zależności od lokalizacji

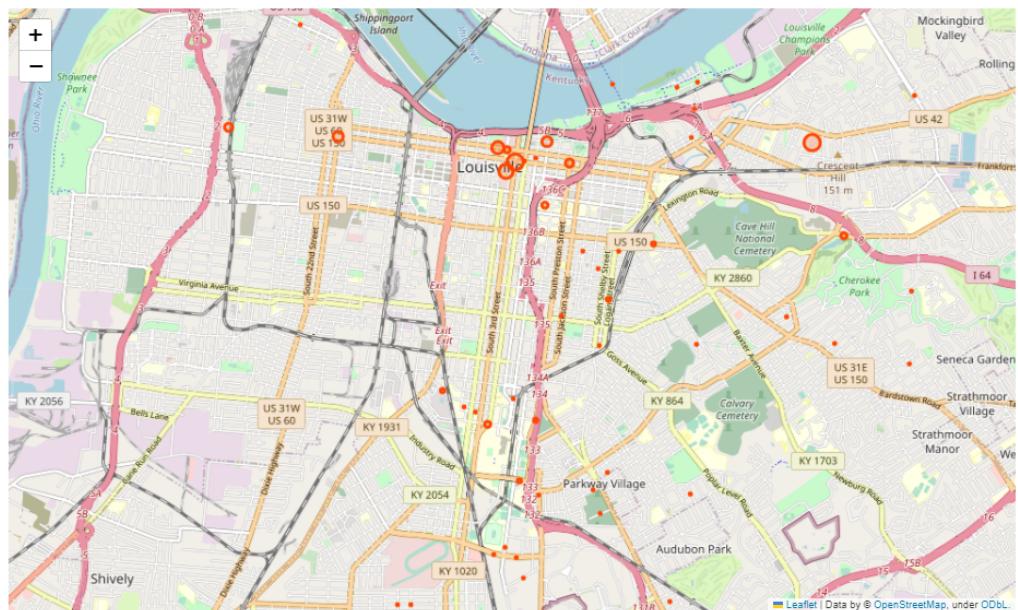
Poniższe mapy przedstawiają miejsca, w których rozpoczynały się podróże oraz ich natężenie.



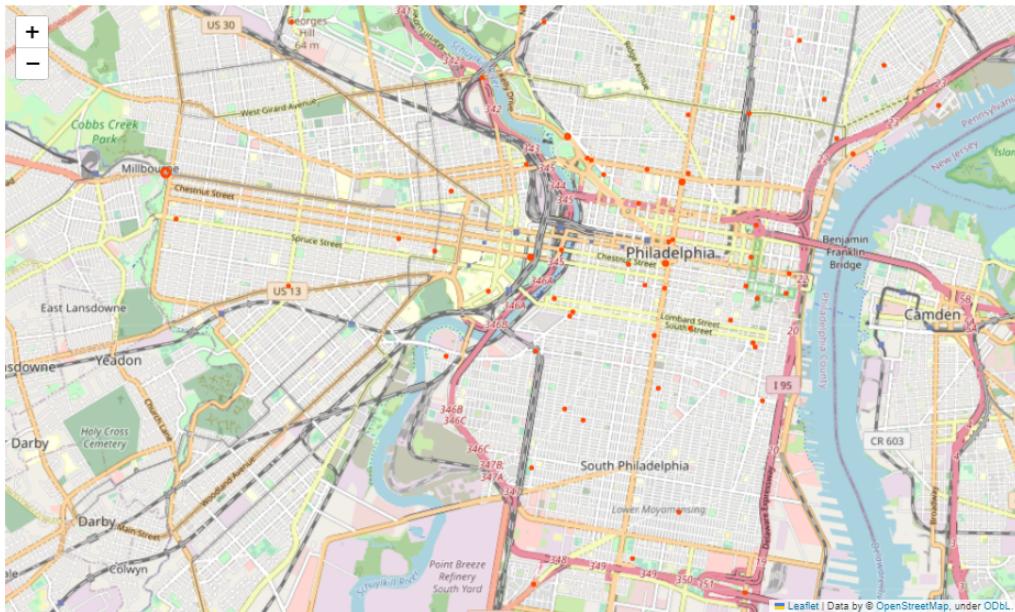
Rysunek 12: Mapa startów przejazdu w Chicago



Rysunek 13: Mapa startów przejazdu w Jersey City



Rysunek 14: Mapa startów przejazdu w Louisville

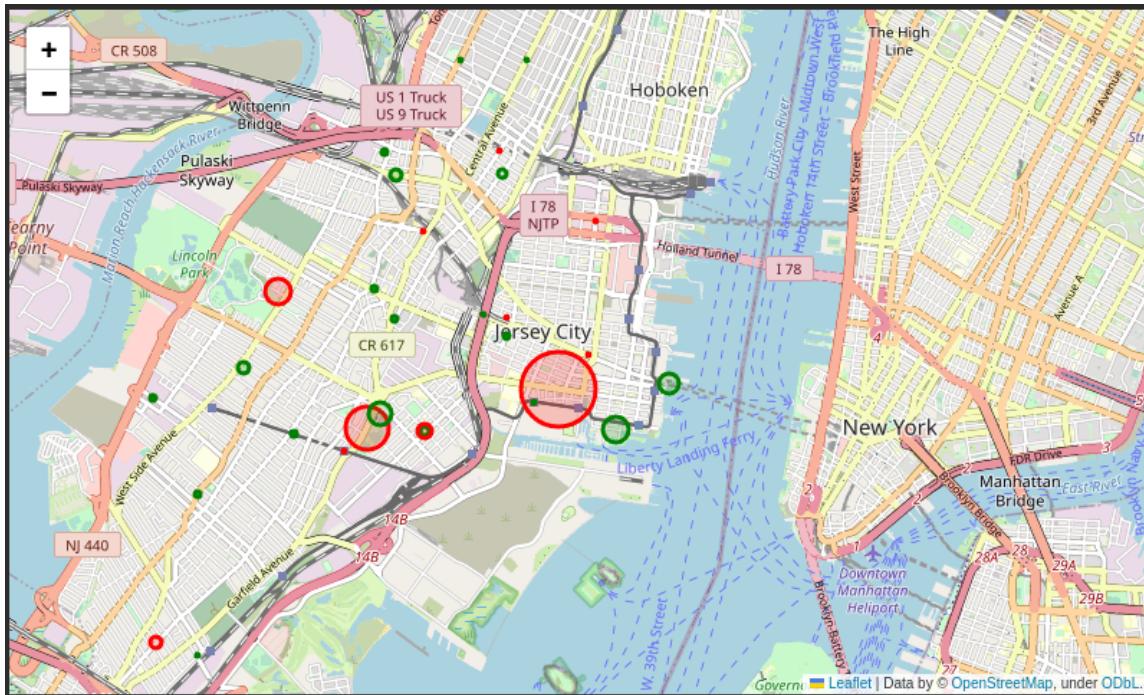


Rysunek 15: Mapa startów przejazdu w Filadelfii

4.2 Wizualizacja różnic ilości przyjazdów i odjazdów w zależności od pory dnia

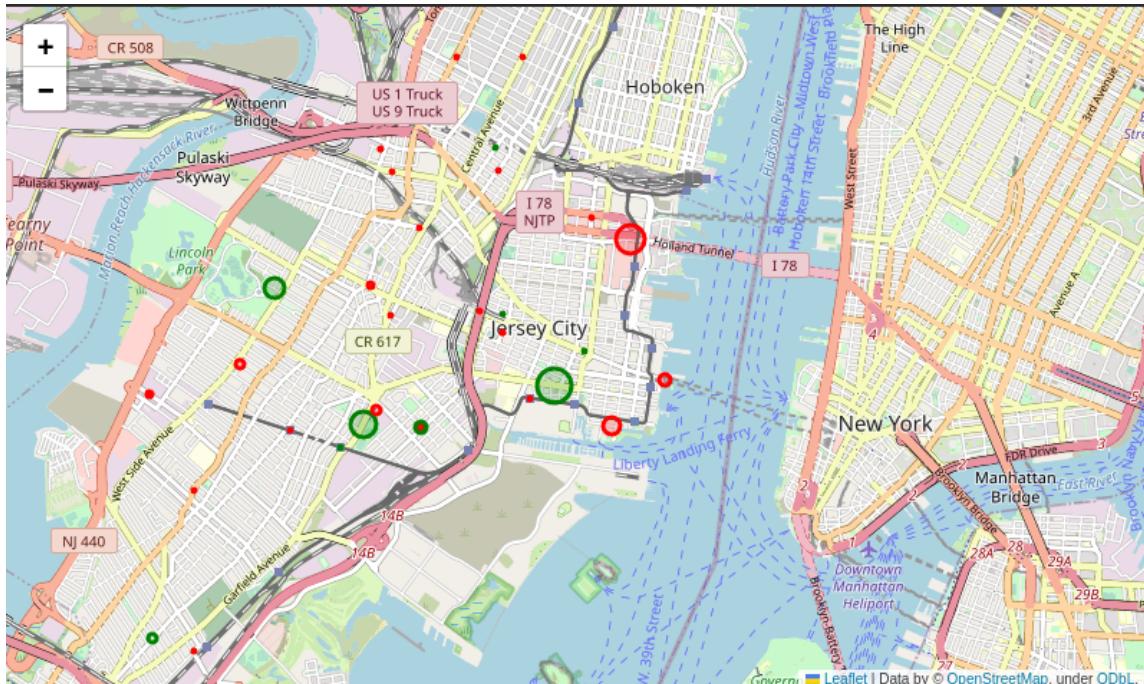
Przedstawione poniżej wykresy pokazują różnice w ilości przyjazdów i odjazdów w zależności od pory dnia. Zielonym kolorem zaznaczone są miejsca, w których więcej jest odjazdów. Wielkość markera jest wprost proporcjonalna do obliczonej różnicy.

1. Jersey City godziny poranne



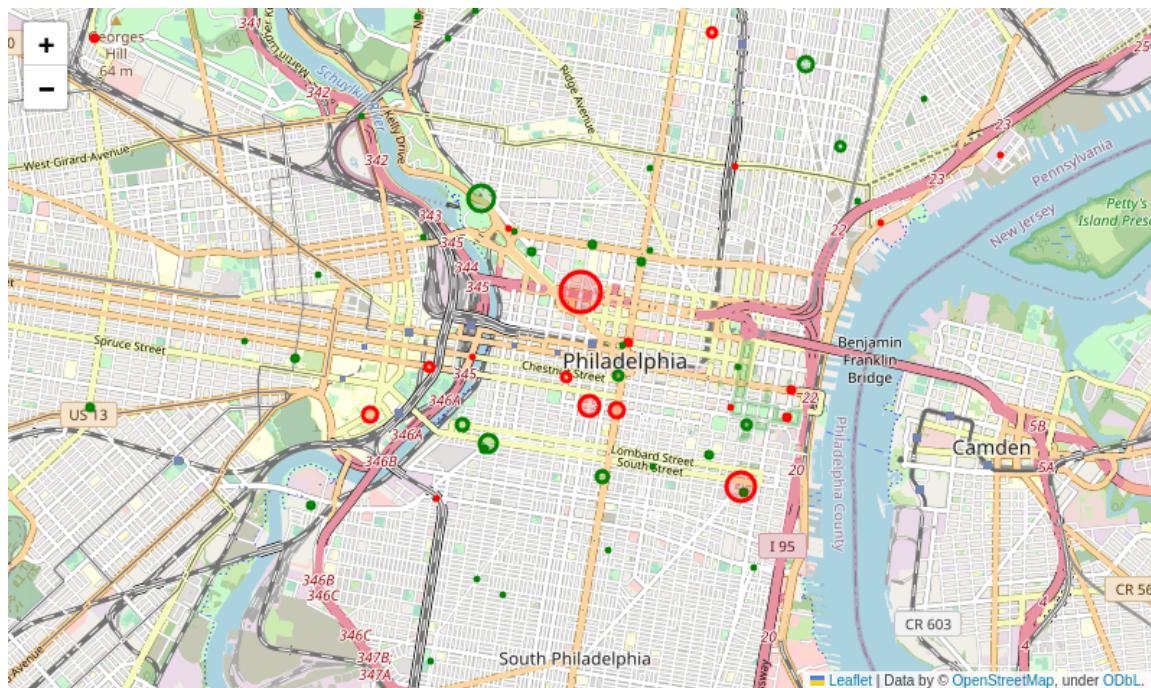
Rysunek 16: Mapa lokalizacji przyjazdów i odjazdów w Jersey City w godzinach porannych

2. Jersey City godziny popołudniowe



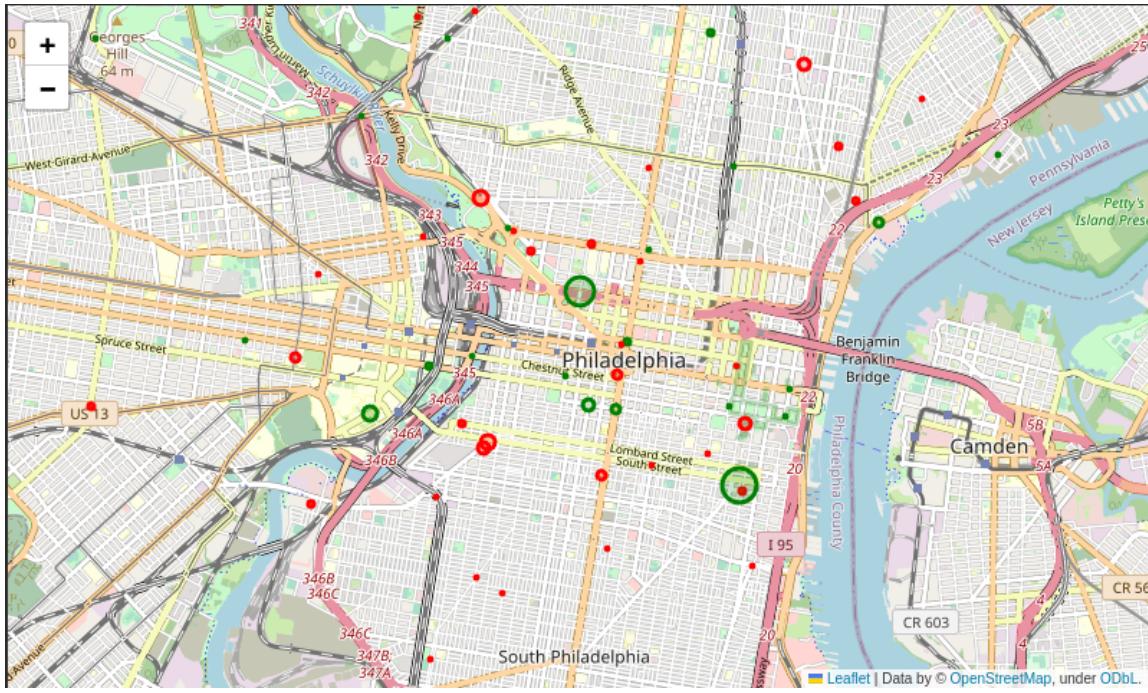
Rysunek 17: Mapa lokalizacji przyjazdów i odjazdów w Jersey City w godzinach wieczornych

3. Philadelphia godziny poranne



Rysunek 18: Mapa lokalizacji przyjazdów i odjazdów w Philadelphia w godzinach porannych

4. Philadelphia godziny wieczorne

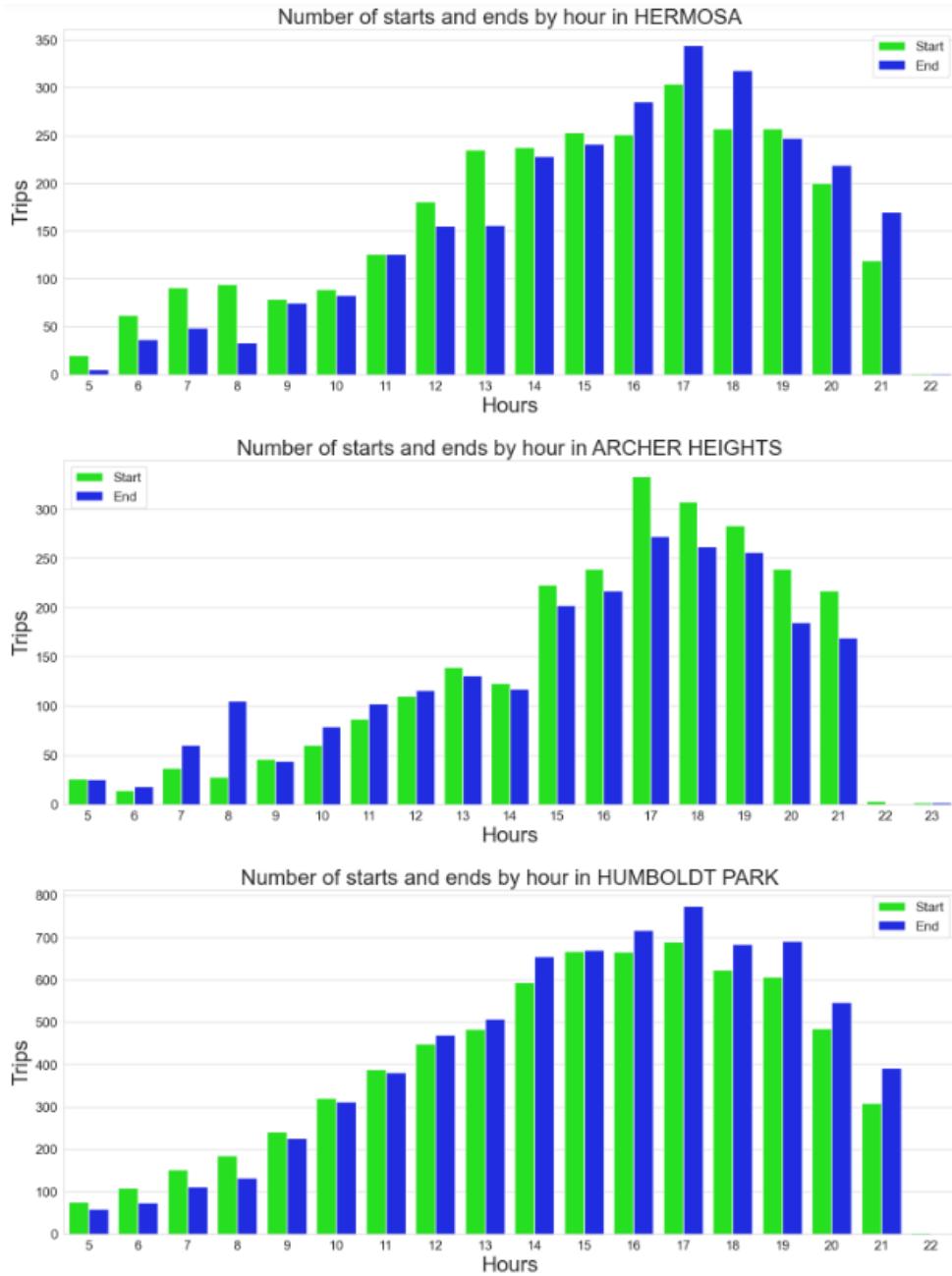


Rysunek 19: Mapa lokalizacji przyjazdów i odjazdów w Philadelphia w godzinach wieczornych

4.3 Lokalizacje z największą liczbą rozpoczętych/zakończonych podróży

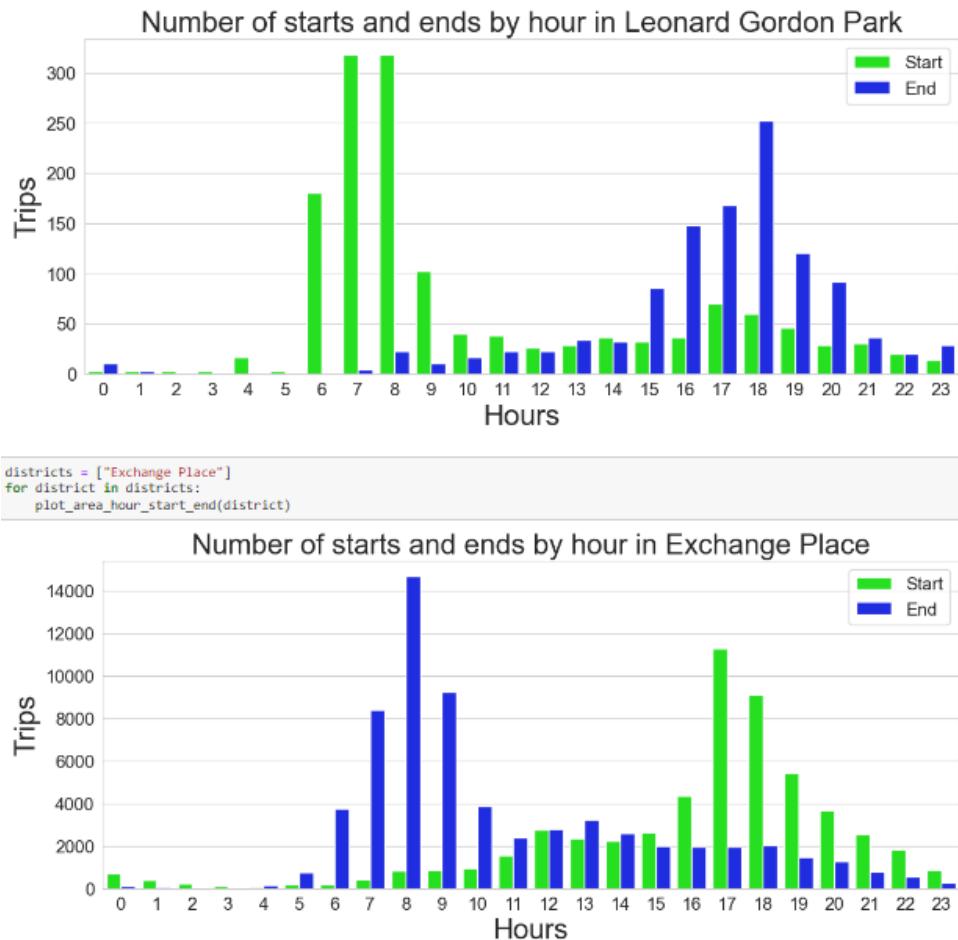
Przedstawione poniżej wykresy ilustrują liczbę rozpoczętych i zakończonych podróży dla miejsc o największym natężeniu ruchu.

1. Chicago



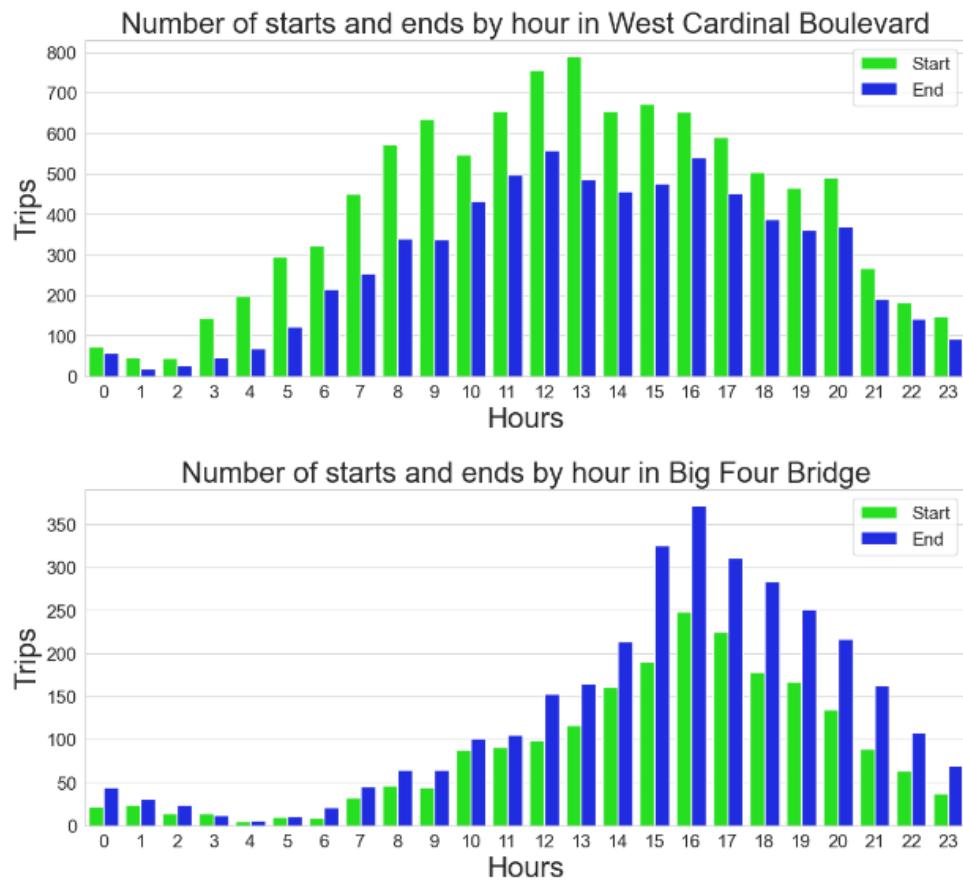
Rysunek 20: Mapa przepływu dla Chicago

2. Jersey City



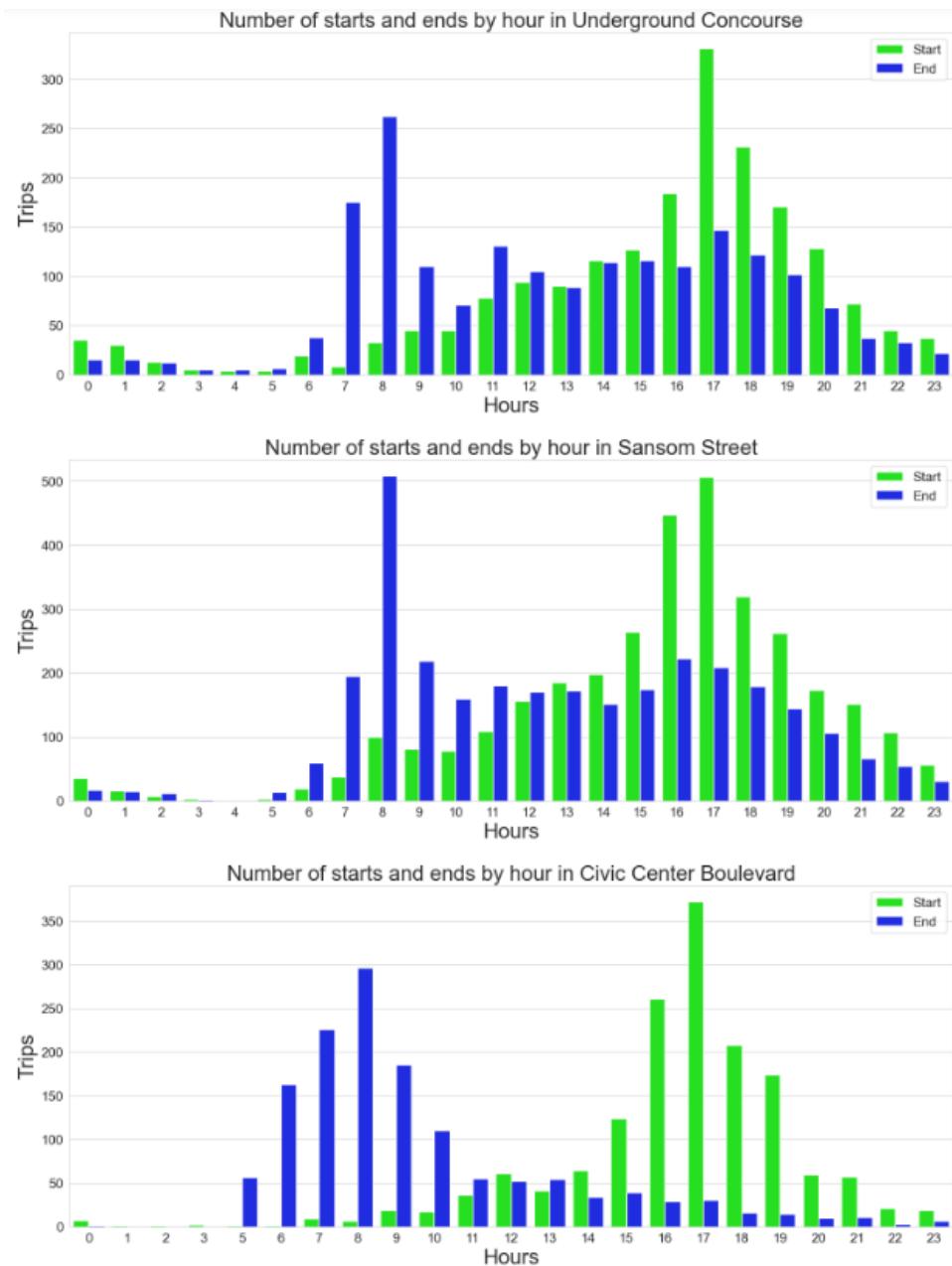
Rysunek 21: Liczba rozpoczętych i zakończonych podróży dla stacji w Jersey City

3. Louisville



Rysunek 22: Liczba rozpoczętych i zakończonych podróży dla stacji w Louisville

4. Philadelphia

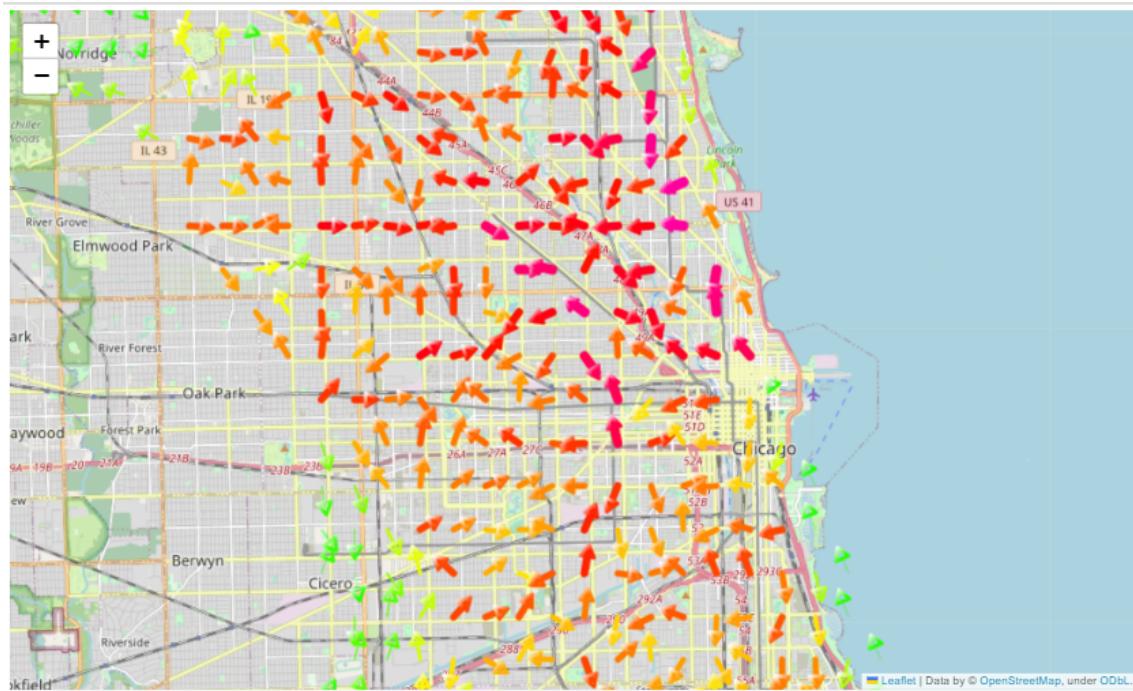


Rysunek 23: Liczba rozpoczętych i zakończonych podróży dla stacji w Philadelphii

4.4 Mapy przepływu

Poniższe mapy przedstawiają kierunki podróży oraz ich natężenie.

1. Chicago



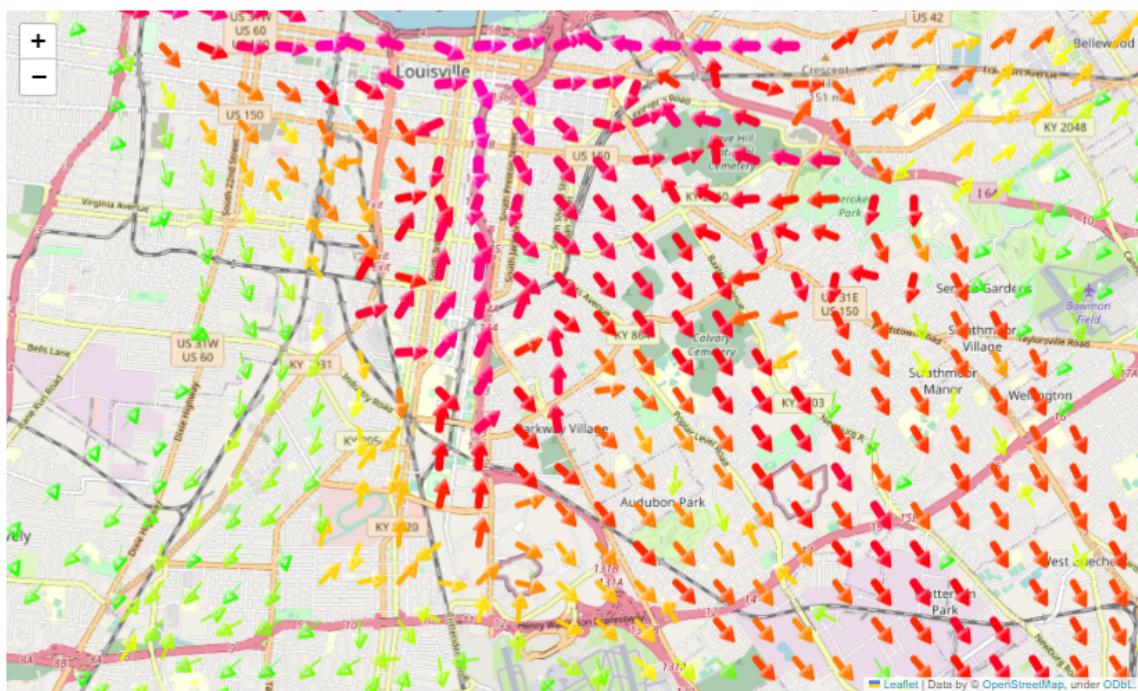
Rysunek 24: Mapa przepływu dla Chicago

2. Jersey City



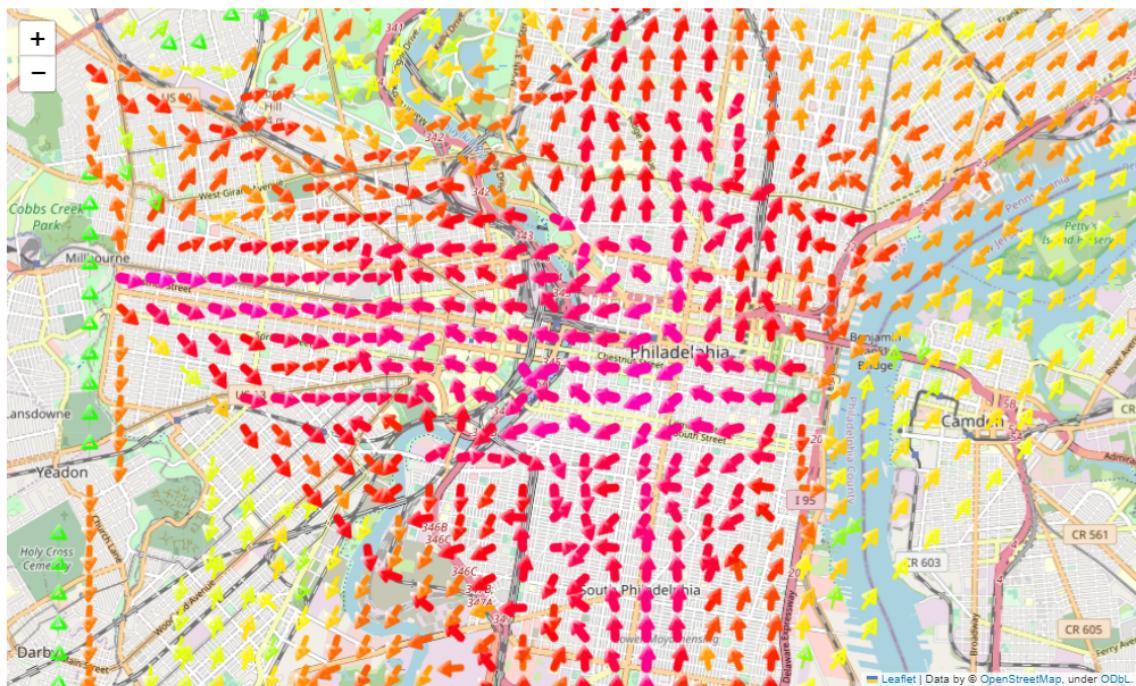
Rysunek 25: Mapa przepływu dla Jersey City

3. Louisville



Rysunek 26: Mapa przepływu dla Louisville

4. Philadelphia



Rysunek 27: Mapa przepływu dla Philadelphii

5 Model predykcyjny

Predykcja parametrów przejazdów pojazdów dwukołowych w wynajmie krótkoterminowym jest zagadnieniem pozwalającym na lepsze dostosowanie infrastruktury na potrzeby klienta oraz zaoszczędzenie zasobów. Dla przykładu, predykcja ilości wynajętych pojazdów pozwala na odpowiednio wcześnie umieszczenie ich w określonym miejsciu z określonym zapotrzebowaniem. W tym przypadku konieczne jest także uwzględnienie się pojazdów przyjezdzących na daną stację. Predykcja innych parametrów przejazdu takich jak czas, prędkość, dystans pozwala na odpowiednie dostosowanie infrastruktury np. dostosowanie baterii w pojazdach elektrycznych, bądź dostosowanie oferty abonamentu wynajmu pod daną grupę klientów.

5.1 Analizowane parametry przejazdu

Przygotowane zostały przez nas dwa modele predykcyjne, które opierały się na następujących parametrach:

a szacowanie wartości parametru opisującego nazwę/ID stacji końcowej przejazdu w zależności od parametrów:

- czas trwania przejazdu
- nazwa/ID stacji początkowej
- dzień tygodnia, w którym odbywał się przejazd
- godzina przejazdu

b szacowanie wartości parametru opisującego czas trwania przejazdu w zależności od parametrów:

- pokonany dystans
- nazwa/ID stacji początkowej
- nazwa/ID stacji końcowej
- dzień tygodnia
- godzina przejazdu

Cechy kategoryczne: nazwa stacji początkowej, nazwa stacji końcowej oraz dzień tygodnia były uprzednio przygotowywane poprzez kodowanie za pomocą *LabelEncoder*, aby zamienić ich wartości na numeryczne reprezentacje.

Przewidywany czas trwania przejazdu był klasyfikowany jako krótki (poniżej 10 minut), średni (pomiędzy 10, a 25 minut) lub długi (powyżej 25 minut).

5.2 Porównanie wyników predykcji

Do predykcji parametrów *Nazwa/ID stacji końcowej* oraz czasu trwania przejazdu zastosowano opisane we wcześniejszej sekcji klasyfikatory: *DecisionTreeClasifier*, *KNeighboursClassifier* oraz *RandomForestClassifier*. Poniższa tabela przedstawia wyniki klasyfikacji uzyskane dla każdego miasta w zależności od wykorzystanego algorytmu.

a Predykcja stacji końcowej

- Chicago

Klasyfikator	Dokładność	Śr. precyzja	Śr. czułość	Śr. miara F1
DecisionTreeClasifier	0.53	0.53	0.53	0.53
KNeighboursClassifier	0.21	0.17	0.21	0.19
RandomForestClassifier	0.64	0.64	0.64	0.64

Tabela 3: Uzyskane wyniki klasyfikacji dla miasta Chicago

- Jersey City

Klasyfikator	Dokładność	Śr. precyzja	Śr. czułość	Śr. miara F1
DecisionTreeClasifier	0.76	0.76	0.76	0.76
KNeighboursClassifier	0.43	0.41	0.43	0.42
RandomForestClassifier	0.77	0.77	0.77	0.77

Tabela 4: Uzyskane wyniki klasyfikacji dla miasta Jersey City

- Louisville

Klasyfikator	Dokładność	Śr. precyzja	Śr. czułość	Śr. miara F1
DecisionTreeClasifier	0.14	0.14	0.14	0.14
KNeighboursClassifier	0.15	0.14	0.15	0.14
RandomForestClassifier	0.19	0.17	0.19	0.18

Tabela 5: Uzyskane wyniki klasyfikacji dla miasta Louisville

- Philadelphia

Klasyfikator	Dokładność	Śr. precyzja	Śr. czułość	Śr. miara F1
DecisionTreeClasifier	0.23	0.23	0.23	0.23
KNeighboursClassifier	0.15	0.16	0.15	0.15
RandomForestClassifier	0.25	0.25	0.25	0.25

Tabela 6: Uzyskane wyniki klasyfikacji dla miasta Philadelphia

b Predykcja czasu trwania przejazdu

- Chicago

Klasyfikator	Dokładność	Śr. precyzja	Śr. czułość	Śr. miara F1
DecisionTreeClasifier	0.58	0.58	0.58	0.58
KNeighboursClassifier	0.65	0.64	0.65	0.64
RandomForestClassifier	0.65	0.64	0.65	0.64

Tabela 7: Uzyskane wyniki klasyfikacji dla miasta Chicago

- Jersey City

Klasyfikator	Dokładność	Śr. precyzja	Śr. czułość	Śr. miara F1
DecisionTreeClasifier	0.86	0.85	0.86	0.85
KNeighboursClassifier	0.83	0.82	0.83	0.82
RandomForestClassifier	0.86	0.85	0.86	0.85

Tabela 8: Uzyskane wyniki klasyfikacji dla miasta Jersey City

- Louisville

Klasyfikator	Dokładność	Śr. precyzja	Śr. czułość	Śr. miara F1
DecisionTreeClasifier	0.67	0.67	0.67	0.67
KNeighboursClassifier	0.71	0.69	0.71	0.69
RandomForestClassifier	0.75	0.72	0.75	0.73

Tabela 9: Uzyskane wyniki klasyfikacji dla miasta Louisville

- Philadelphia

Klasyfikator	Dokładność	Śr. precyzja	Śr. czułość	Śr. miara F1
DecisionTreeClasifier	0.70	0.70	0.70	0.70
KNeighboursClassifier	0.64	0.64	0.64	0.64
RandomForestClassifier	0.68	0.68	0.68	0.68

Tabela 10: Uzyskane wyniki klasyfikacji dla miasta Philadelphia

6 Podsumowanie

W tej sekcji przedstawimy podsumowanie uzyskanych wyników naszej analizy danych dotyczącej użytkowania rowerów i miejscowych skuterów elektrycznych. Przez przeprowadzenie zaawansowanej analizy, tworzenie wizualizacji oraz modelu predykcyjnego, otrzymaliśmy cenne informacje i wnioski, które mogą przyczynić się do lepszego zrozumienia wzorców ruchu jednośladów w badanych obszarach.

6.1 Obserwacje

W trakcie analizy danych dotyczących wykorzystania rowerów i elektrycznych hulajnog w różnych miastach i okresach czasu, skupiliśmy się na różnych aspektach, takich jak średnia liczba przejazdów w ciągu dnia, miesiące z najwyższą i najniższą liczbą przejazdów, godziny i dni tygodnia z największą i najmniejszą liczbą przejazdów oraz miejsca o największej aktywności transportowej.

6.1.1 Zależność liczby przejazdów od pory roku

Analizując zależności między porą roku a liczbą przejazdów, można zauważać pewne wzorce i tendencje. Pora roku, wraz ze zmianami warunków atmosferycznych i temperaturą, może mieć wpływ na preferencje i zachowania użytkowników jednośladów.

W przypadku rowerów miejskich w mieście Jersey City, obserwuje się, że miesiące jesienne (październik i listopad) w latach 2015 i 2016 oraz marzec 2017 miały liczbę przejazdów powyżej średniej. Natomiast miesiące zimowe (grudzień-luty) w obu badanych okresach miały liczbę przejazdów poniżej średniej.

W przypadku rowerów miejskich w Filadelfii w 2016 roku, obserwuje się, że lato (maj-lipiec) miało liczbę przejazdów powyżej średniej. Natomiast w kwietniu, wcześniejszą wiosną, liczba przejazdów była poniżej średniej.

W przypadku hulajnog w Chicago w 2020 roku, okres letni (sierpień-październik) miał liczbę przejazdów powyżej średniej. Natomiast listopad i grudzień miały liczbę przejazdów poniżej średniej.

W przypadku hulajnog w Louisville w latach 2018-2019, okres letni (kwiecień-październik) miał liczbę przejazdów powyżej średniej. Natomiast wczesna wiosna (marzec) i zima (styczeń-luty) miały liczbę przejazdów poniżej średniej..

Podsumowując, pora roku i warunki atmosferyczne mogą mieć istotny wpływ na liczbę tripów w różnych miesiącach. Okresy z cieplejszą pogodą i bardziej stabilnymi warunkami są zazwyczaj związane z większą aktywnością użytkowników jednośladów. Jednak warto zauważać, że preferencje i zachowania mogą różnić się w zależności od lokalizacji geograficznej, klimatu i innych czynników specyficznych dla danego obszaru.

6.1.2 Zależność liczby oraz czasu przejazdów od dnia tygodnia

Analizując dane dotyczące liczby oraz czasu przejazdów w zależności od dnia tygodnia dla różnych miast, można wyciągnąć następujące wnioski:

W przypadku rowerów miejskich w Jersey City oraz Filadelfii, dni robocze (poniedziałek - piątek) charakteryzowały się większą liczbą przejazdów w porównaniu do weekendowych dni (sobota i niedziela). Największą ilość przejazdów odnotowano w środy i czwartki, natomiast soboty i niedziele miały najniższą liczbę przejazdów.

W przypadku hulajnog w Chicago oraz Louisville, zauważono, że dni weekendowe (piątek - niedziela) miały większą liczbę przejazdów niż dni robocze (poniedziałek - czwartek). Największą ilość przejazdów odnotowano w piątki i soboty, podczas gdy poniedziałki i wtorki miały najmniejszą liczbę przejazdów.

Dla jednośladów Jersey City, Filadelfii oraz Chicago, najdłużej trwające przejazdy odbywały się w dni weekendowe (piątek-niedziela), a najkrótsze w poniedziałki, wtorki i czwartki. Nieco inaczej wyglądało to dla Louisville, w którym najdłuższe przejazdy odbywały się w środy i piątki, a najkrótsze w soboty i niedziele.

6.1.3 Zależność liczby oraz czasu przejazdów od godziny

Wnioski z analizy danych dotyczących liczby oraz czasu przejazdów w zależności od godziny dla różnych miast są następujące:

Dla rowerów miejskich w Jersey City i Filadelfii, godziny szczytu (7-9, 12-20) charakteryzowały się większą liczbą przejazdów, a godziny poza szczytem (0-6, 10-11, 21-0) miały mniejszą liczbę przejazdów. Godziny 8, 18 i 17 odnotowały największą liczbę przejazdów, podczas gdy godziny 3, 4 i 2 miały najmniejszą liczbę przejazdów.

W Chicago i Louisville, godziny od 9 do 21 miały większą liczbę przejazdów, a godziny od 21 do 8 miały mniejszą liczbę przejazdów. Godziny 17, 18 i 16 odnotowały największą liczbę przejazdów, podczas gdy godziny od 23 do 4 miały najmniejszą liczbę przejazdów.

Dla wszystkich badanych miast, najdłuższe przejazdy odbywały się w godzinach nocnych (0-4), a najkrótsze w godzinach wczesnoporannych (5-8).

Analiza danych wskazuje, że liczba oraz czas przejazdów różni się w zależności od godziny w poszczególnych miastach. Godziny szczytu często charakteryzują się większym zapotrzebowaniem na środki transportu, takie jak rowery miejskie czy hulajnogi.

6.1.4 Zależność liczby przejazdów od wieku oraz płci wynajmującego

Jednym zbiorem, który zawierał informacje o płci oraz wieku wynajmującego były dane dla przejazdów rowerami miejskimi w Jersey City. Analiza tych danych wskazała na największą popularność rowerów miejskich wśród osób w przedziale wiekowym 25-40 lat, a najmniejszą wśród osób w wieku poniżej 18 i powyżej 65 roku życia. Ponadto, około 85% procent osób wypożyczających rowery stanowili mężczyźni.

6.1.5 Procentowa ilość tras o takiej samej stacji końcowej i początkowej

Wykresy zostały stworzone dla miast Jersey City, Filadelfia, Chicago. Dane dla miasta Louisville zawierają dokładniejszą lokalizację miejsca początkowego oraz startowego. Powodem tego może być brak wymagania, żeby pojazdy były odkładane w wyznaczonych miejscach. Z tego względu też analiza tego zbioru danych pod tym kątem jest zbędna, ponieważ lokalizacje się nie powtarzają. Dane dla miast Jersey City oraz Filadelfii wskazują na znaczącą przewagę odkładania roweru w innej stacji. Przez stację rozumiemy tutaj dzielnicę, więc powodem takiej prawidłowości może być gęste rozmieszczenie stacji/mały obszar miejski. Dane dla Chicago wskazują na znaczną przewagę odkładania roweru w tym samym dystrykcie. Miasto Chicago jest większe, więc nawet dłuższe dystanse mogą być przeprowadzone w ramach jednego dystryktu. Ważnym w analizie tego problemu jest dokładność lokalizacji oraz rozróżnienie na lokalizację końcową trasy a lokalizacje dzielnicy, w którym znajduje się końcowa stacja.

6.1.6 Identyfikacja obszarów przejazdu o największej aktywności

Podane obszary w poszczególnych miastach charakteryzują się największą aktywnością. Oto obszary o największej aktywności w poszczególnych miastach:

- a Jersey City: Christopher Columbus Drive, Exchange Place, Sip Avenue, McWilliams Place
- b Filadelfia: Market Street, Walnut Street, South Broad Street, North Broad Street
- c Louisville: South 3rd Street, North Clifton Avenue, Link Downtown Skywalk System, East Main Street
- d Chicago: Lake View, Lincoln Park, West Town, Near West Side

W tych obszarach można spodziewać się większego ruchu i aktywności, być może ze względu na centralne położenie, dostępność komunikacji publicznej, atrakcje turystyczne, sklepy lub biura.

6.2 Wnioski

Podczas analizy danych, zauważaliśmy interesujące trendy dobowe w użyciu rowerów i skuterów elektrycznych. Określiliśmy godziny szczytu, w których obserwowało największą aktywność transportową, co może być istotne przy planowaniu zasobów i infrastruktury. Ponadto, przyjrzaliśmy się rozkładowi wielkości przemieszczeń, co pozwoliło nam zidentyfikować zarówno krótkie jak i długie trasy preferowane przez użytkowników.

Wizualizacje odbytych podróży na mapach pozwoliły nam zobaczyć miejsca, w których rozpoczęły się trasy oraz ich natężenie. Przez analizę tych danych, zidentyfikowaliśmy obszary o największej i najmniejszej aktywności transportowej w badanych miastach. Odkryliśmy popularne trasy, które były intensywnie wykorzystywane przez użytkowników, co może sugerować potrzebę dalszego rozwoju infrastruktury w tych obszarach.

Dodatkowo, przeprowadziliśmy modelowanie predykcyjne, które miało na celu przewidywanie parametru *End Station ID/Name* na podstawie cech takich jak czas trwania podróży, stacja początkowa, dzień tygodnia i godzina. Nasze modele oparte na algorytmach *DecisionTreeClassifier*, *KNeighborsClassifier* i *RandomForestClassifier* osiągnęły satysfakcyjne, jak na specyfikę danych zbiorów, wyniki predykcji.

Przeprowadzono także predykcję czasu trwania przejazdu w zależności od dystansu, stacji początkowej, stacji końcowej, dnia tygodnia oraz godziny przejazdu. Zależnie od miasta i użytej metody klasyfikacji otrzymaliśmy zróżnicowane wyniki.

Podsumowując, nasza analiza danych, wizualizacje podróży oraz model predykcyjny dostarczyły nam wartościowych informacji dotyczących użytkowania rowerów i skuterów elektrycznych. Uzyskane wyniki mogą być wykorzystane przez lokalne władze oraz organizacje transportowe do podejmowania lepiej ugruntowanych decyzji związanych z planowaniem infrastruktury, alokacją zasobów oraz doskonaleniem systemów jednośladów w badanych miastach.

6.3 Kod źródłowy

Kod źródłowy został umieszczony w repozytorium na GitHub:

<https://github.com/JanickiJ/E-Scooter-Data-Analysis>

W repozytorium znajdują się również analizowane w pracy zbiory danych.

Bibliografia

- [1] Lyft, Citi Bike Trip Histories. 2018; <https://ride.citibikenyc.com/system-data>.
- [2] Indego, Anonymized Indego trip data. 2018; <https://www.rideindegoo.com/about/data/>.
- [3] Raza, M. Electric -Scooter Trips in Chicago. 2020; <https://www.kaggle.com/datasets/razamh/escooters-trips>.
- [4] Morley, B. City Lousiville escooter trip data. 2018/19; <https://www.kaggle.com/datasets/busielmorley/city-lousiville-escooter-trip-data>.
- [5] Mattfeld, D. Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns. **2011**,
- [6] Politis, I. Shifting to Shared Wheels: Factors Affecting Dockless Bike-Sharing Choice for Short and Long Trips. **2020**,
- [7] Fernando Pérez, B. G. Jupyter Notebook. 2014; <https://jupyter.org/>.
- [8] JetBrains, PyCharm. 2010; <https://www.jetbrains.com/pycharm/>.
- [9] Microsoft, Visual Studio Code. 2015; <https://code.visualstudio.com/>.
- [10] McKinney, W. Pandas 2.0.2. 2008; <https://pandas.pydata.org/>.
- [11] Oliphant, T. NumPy 1.24.3. 2023; <https://numpy.org/>.
- [12] Hunter, J. D. Matplotlib 3.0.3. 2019; <https://matplotlib.org/>.
- [13] Waskom, M. Seaborn 0.12. 2012; <https://seaborn.pydata.org/>.
- [14] Johnson, A.; Parmer, J.; Parmer, C.; Sundquist, M. Plotly 5.15.0. 2020; <https://plotly.com/>.
- [15] Cournapeau, D. Scikit-learn 1.2.2. 2023; <https://scikit-learn.org/stable/>.
- [16] Esmukov, K. GeoPy 2.3.0. 2014; <https://geopy.readthedocs.io/en/stable/>.
- [17] Scikit-learn, A decision tree classifier. <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>.

- [18] Scikit-learn, Classifier implementing the k-nearest neighbors vote. <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>.
- [19] Scikit-learn, A random forest classifier. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>.