**NY bike trips analysis**

Spis treści

[Introduction 1](#_Toc158137271)

[Used technology 1](#_Toc158137272)

[Data manipulations 1](#_Toc158137273)

[Analysis 1](#_Toc158137274)

[Conclusion 11](#_Toc158137275)

# Wstęp

W tym projekcie zamierzamy przeanalizować podróże na wypożyczonych rowerach na Manhattanie. Dysponujemy danymi z 3 dni (19/07/2013 - 21/07/2013) dotyczącymi przejazdów rowerowych w Nowym Jorku. Niestety, w zbiorze danych brakuje informacji o pierwszych 16 godzinach piątku i ostatnich 10 godzinach niedzieli (w sumie jest to 26 godzin, więc przegapiliśmy więcej niż jeden dzień). Dane te otrzymałem od uniwersytetu w ramach jednego z projektów i muszę je przeanalizować, więc postanowiłem zrobić to w Pythonie. Zawierają one informacje o czasie trwania podróży, czasie rozpoczęcia, czasie zakończenia, identyfikatorze stacji początkowej, nazwie stacji początkowej, szerokości geograficznej stacji początkowej, długości geograficznej stacji początkowej, identyfikatorze stacji końcowej, nazwie stacji końcowej, szerokości geograficznej stacji końcowej, długości geograficznej stacji końcowej, identyfikatorze roweru, typie użytkownika, roku urodzenia, płci. Nie wykorzystamy wszystkich tych danych do naszej analizy, ale tylko niektóre z nich (czas trwania podróży, czas rozpoczęcia, czas zakończenia, typ użytkownika, rok urodzenia, płeć). Pod koniec tego projektu będziemy w stanie stworzyć profil przeciętnego użytkownika rowerów publicznych dzięki danym, które zebraliśmy od ludzi..

# Używane technologie

* Python
* Pandas
* Matplotlib
* NumPy
* DateTime

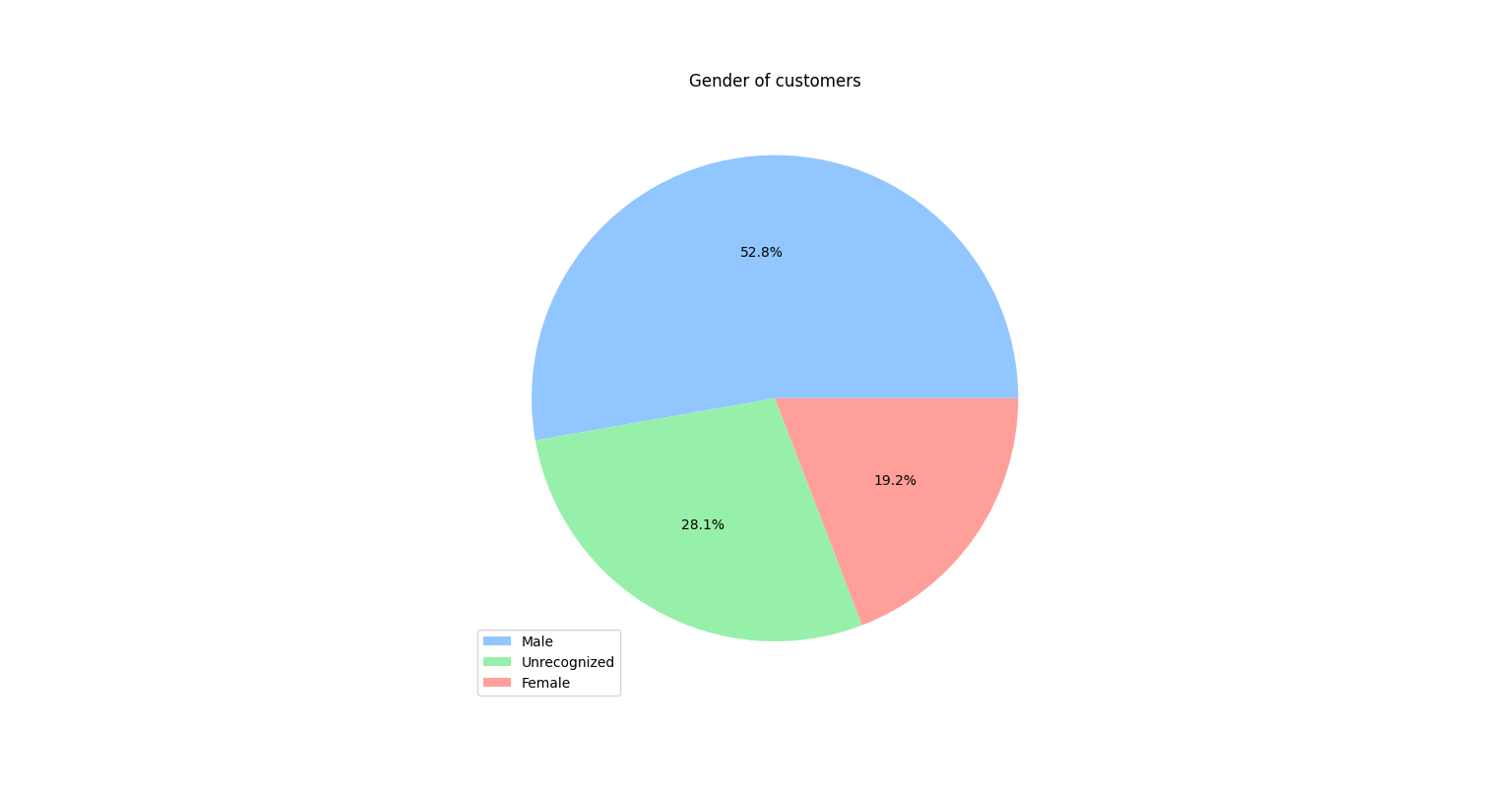
# Manipulacje danymi

Analiza rozpoczęła się od filtrowania danych tylko dla Manhattanu (filtrowanie zostało wykonane przy użyciu innego pliku wejściowego, który zawiera wszystkie identyfikatory stacji Manhattan). Kod użyty do tego celu można znaleźć w pliku "Filter\_station\_only\_Manhattan.py". Surowy zbiór danych zawierał 50 000 przejazdów, ale po użyciu funkcji uzyskano 42 777. Następnym krokiem było odfiltrowanie danych tylko dla mojej stacji i mojego roweru. ID stacji to 388, ID roweru to 16027. Zbiór danych zawierający tylko identyfikator mojej stacji zawiera informacje o 360 przejazdach. Zbiór danych zawierający tylko identyfikator mojego roweru zawiera informacje o 8 przejazdach.

# Ansliza

Analiza rozpoczęła się po kilku wstępnych manipulacjach danymi. W pierwszym etapie przeanalizujemy zestaw danych zawierający tylko moją stację, a w drugim etapie przeanalizujemy niektóre konkretne dane dotyczące rowerów. Cały kod, który został użyty do tej analizy można sprawdzić w pliku "Anslysis\_my\_station.py".

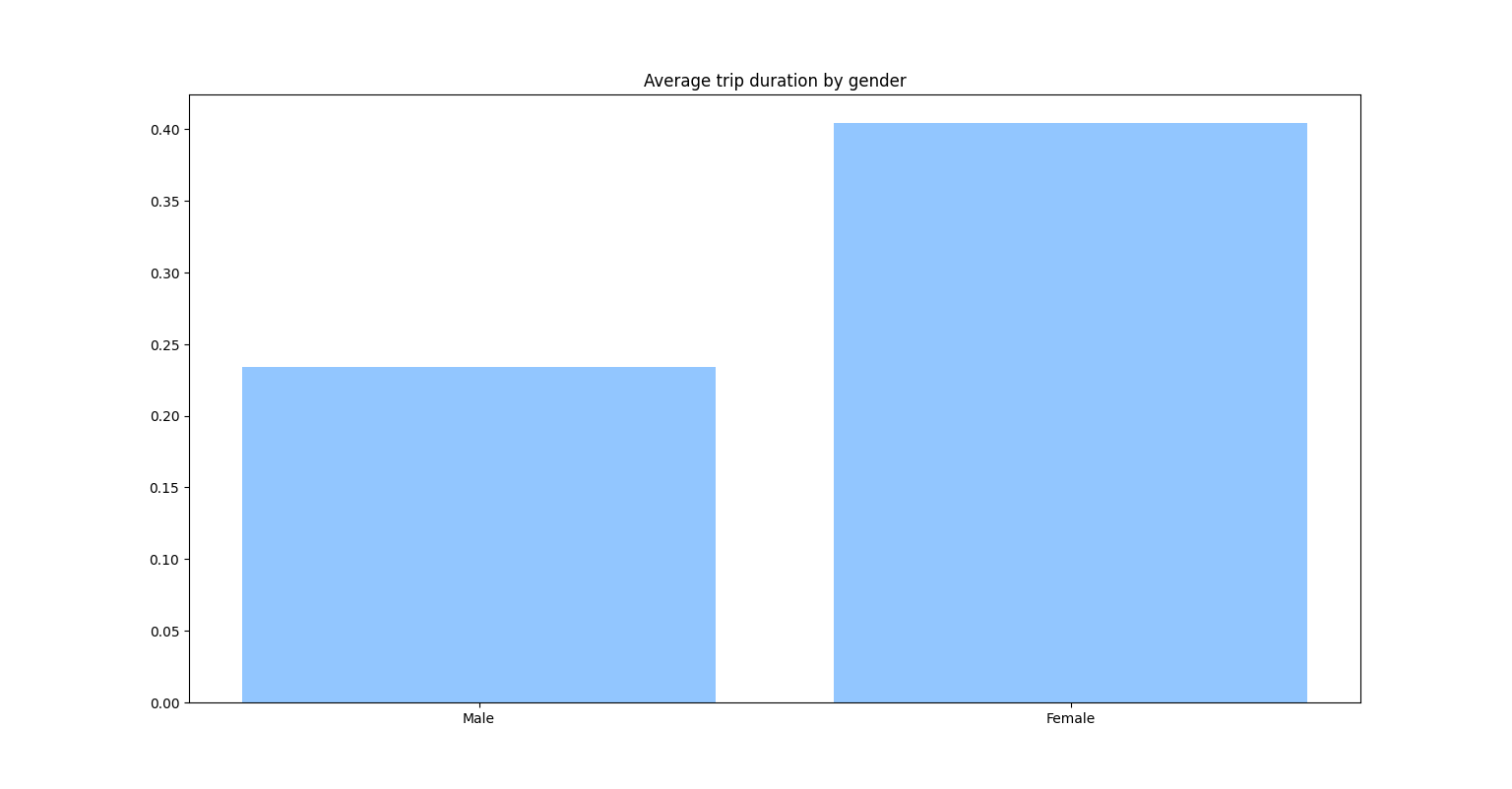
Analizę chciałbym rozpocząć od znajomości użytkowników. Mamy kilka danych na ten temat, więc zacznijmy od płci. Jak widzimy na poniższym wykresie, wśród osób korzystających z tej stacji przeważali mężczyźni (52,8%). 28,1% użytkowników nie określiło swojej płci, a 19,2% stanowiły kobiety. Oznacza to, że mężczyźni ponad dwukrotnie częściej korzystają z usług rowerów publicznych niż kobiety



Rysunek 1. Płeć klienta

Źródło: opracowanie własne

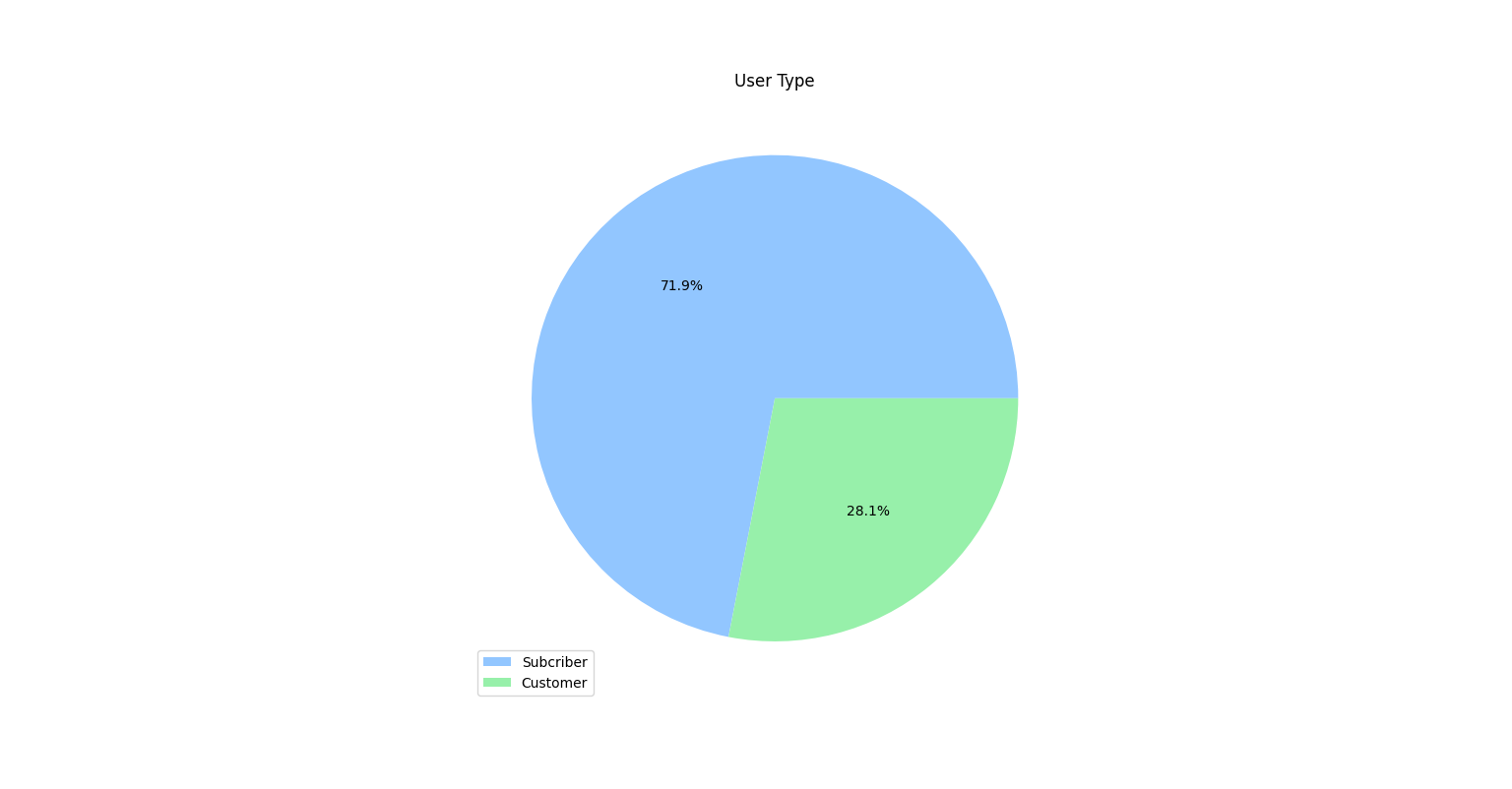
Mając dane na temat płci użytkowników, możemy przeanalizować średni czas trwania podróży kobiet i mężczyzn. Jak widzimy, jest on o prawie 10 minut dłuższy w przypadku kobiet. Tak więc liczba podróży jest wyższa dla mężczyzn, a średni czas trwania dla kobiet



Rysunek 2. Średni czas trwania podróży według płci

Źródło: opracowanie własne

Kolejne dane, które możemy wywnioskować, to typ użytkowników. Istnieją 2 typy: Subskrybent (użytkownik, który zasubskrybował usługi) i Klient (użytkownik, który nie zasubskrybował usług). W naszych danych mamy 71,9% subskrybentów i 28,1% klientów. Może to oznaczać, że ludzie bardzo często korzystają z rowerów, co oznacza, że systemy rowerów publicznych są popularne, dobrze zorganizowane i rozwinięte. Potwierdzenie tych słów można zobaczyć na poniższym wykresie:



Rysunek 3. Typ użytkownika

Źródło: opracowanie własne

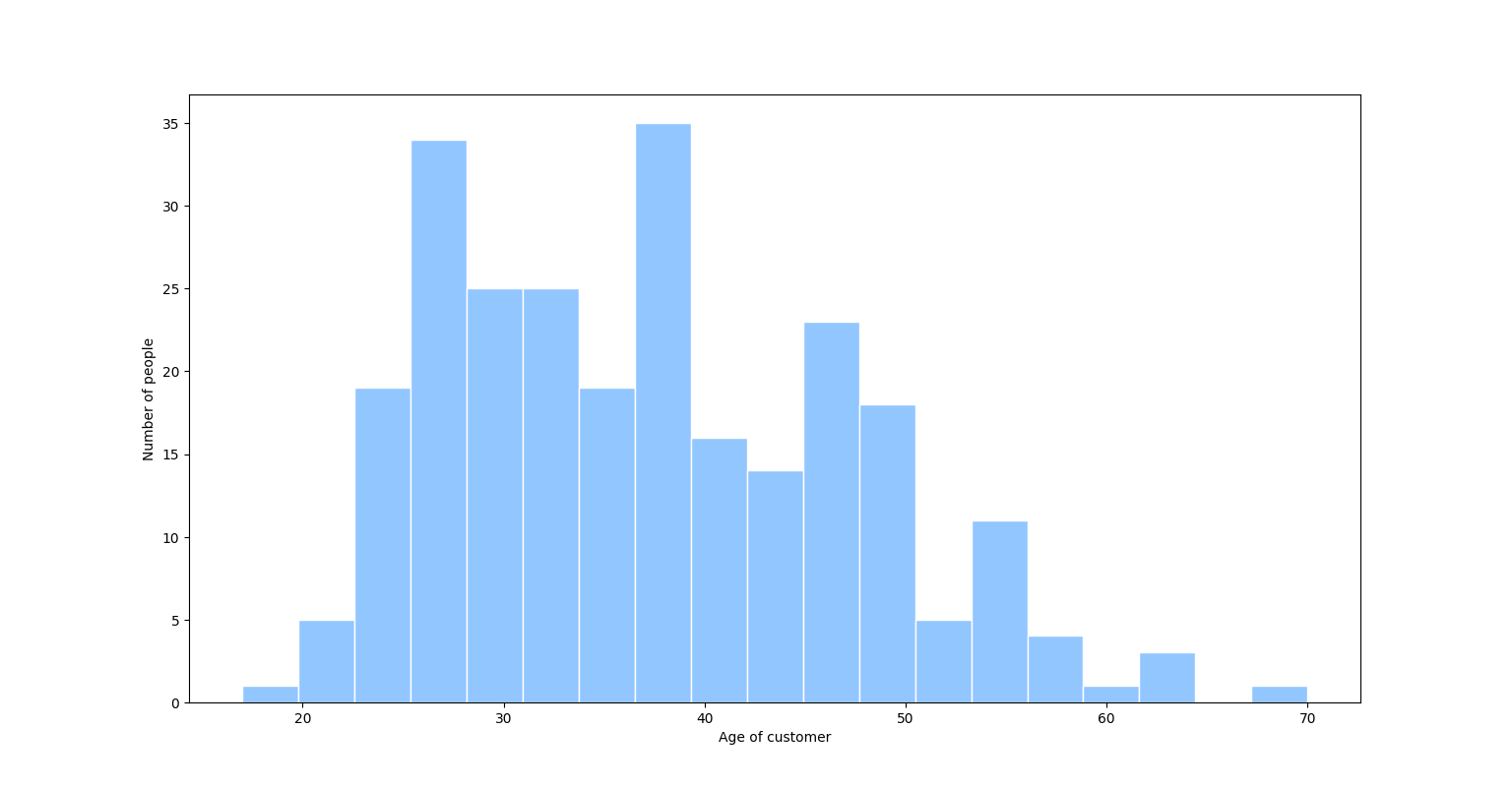
Ostatnią informacją o użytkownikach był rok urodzenia, który możemy łatwo przeliczyć na wiek w momencie podróży (2013 - rok urodzenia). Ale najpierw porozmawiajmy o konkretnych liczbach. Większość klientów jest w wieku od 20 do 50 lat (89% z wyłączeniem nierozpoznanych użytkowników). Tylko 1 użytkownik był w wieku 10-20 lat. 4 użytkowników miało 60-70 lat, a 1 70-80 lat. 101 użytkowników nie określiło swojej daty urodzenia. Jako wstępny wniosek możemy powiedzieć, że wypożyczalnia rowerów jest używana częściej przez osoby starsze, do podróżowania do / z pracy, do / ze sklepów itp. Dane te można zobaczyć w poniższej tabeli:

|  |  |
| --- | --- |
| Wiek | |
| 10-20 | 1 |
| 20-30 | 67 |
| 30-40 | 95 |
| 40-50 | 69 |
| 50-60 | 22 |
| 60-70 | 4 |
| 70-80 | 1 |
| Unrecognised | 101 |

Tabela 1. Wiek użytkownika

Źródło: opracowanie własne

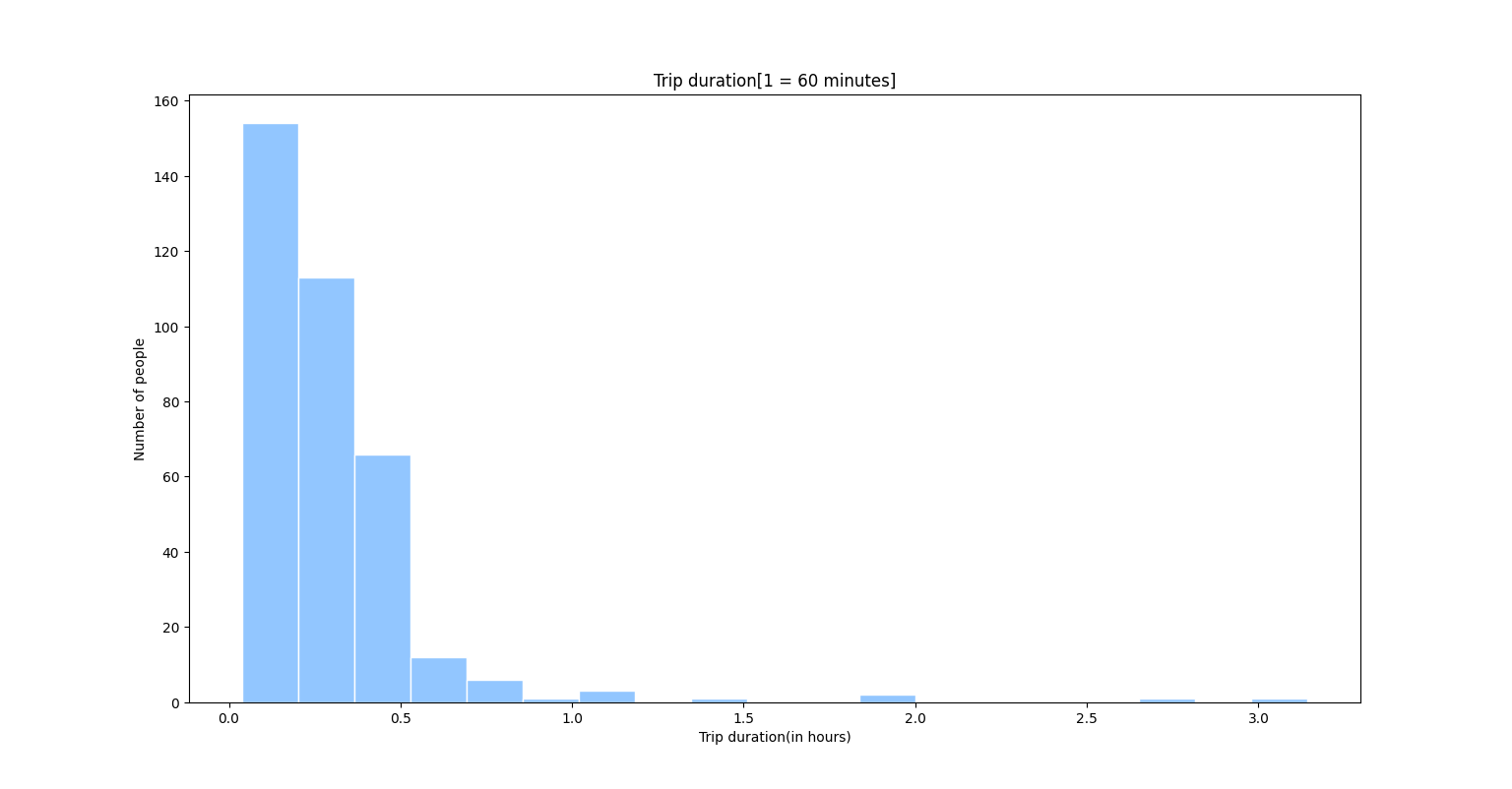
Mamy więc ich wiek i możemy zbudować histogram pokazujący ich rozkład. Jak widzimy, większość klientów jest w wieku od 30 do 60 lat, ze średnią wieku 37 lat, odchyleniem standardowym 9,98 (co jest dość dużo, ¼ średniej, oznacza to, że rowery są poszukiwane wśród osób w każdym wieku) i wariancją 99,6. Patrząc na histogram, możemy powiedzieć, że ma on dodatnie nachylenie, a szczyt przypada na osoby w wieku <40 lat.



Rysunek 4. Typ Rozkład wieku użytkowników

Źródło: opracowanie własne

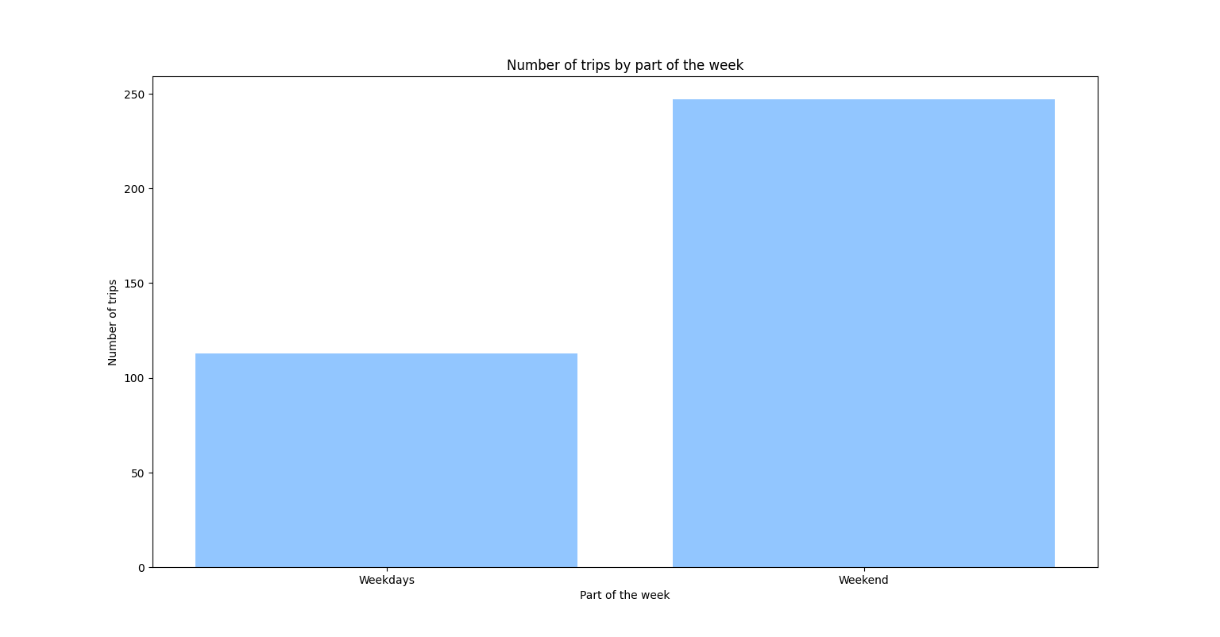
Po zapoznaniu się z użytkownikami możemy przystąpić do analizy danych dotyczących podróży. Po pierwsze, sprawdźmy liczbę podróży według czasu ich trwania w określonych przedziałach czasowych i zbudujmy histogram na podstawie tych danych. Jak widzimy na poniższym wykresie, większość podróży (92,5%) trwała <30 minut (1 = 1 godzina = 60 minut, więc 0,5 = 30 minut). Byli jednak użytkownicy, którzy wypożyczyli rower na 90, 100, 170, a nawet 200 minut (ale liczba podróży o takim czasie trwania jest niewielka). Najprawdopodobniej osoby te korzystały z rowerów w celach rekreacyjnych, a większość podróży odbywała się w konkretnych celach, takich jak powrót do domu z pracy lub pójście do sklepu, ponieważ Manhattan jest bardzo dużym obszarem. Średni czas trwania podróży wynosi 20 minut, z odchyleniem standardowym 20 minut i wariancją 5 minut. Oznacza to, że klienci nie używają rowerów do podróży na długich dystansach, ponieważ średnia prędkość początkującego rowerzysty wynosi 15-20 km / h, więc ludzie używają ich do podróży na 4,5-6 km. Potwierdza to moje wcześniejsze teorie.



Rysunek 5. Czas trwania podróży

Źródło: opracowanie własne

Ponadto przeanalizujemy i porównamy dane dla weekendów i dni powszednich, części dni i różnych dni. Zacznijmy od części tygodnia. Mamy dni powszednie i weekendy. W zbiorze danych tylko piątek jest dniem powszednim, a sobota i niedziela to weekendy. Oznacza to, że porównywanie części tygodnia jest złym pomysłem, ponieważ wynik nie będzie reprezentatywny, ale zrobię to. Po pierwsze, porównajmy liczbę podróży w dni powszednie i weekendy. Przypomnijmy, że mamy tylko drugą połowę piątku i pierwszą połowę niedzieli. Jak widzimy na poniższym wykresie, liczba przejazdów w piątek wynosi nieco ponad 100, a w weekendy około 250.

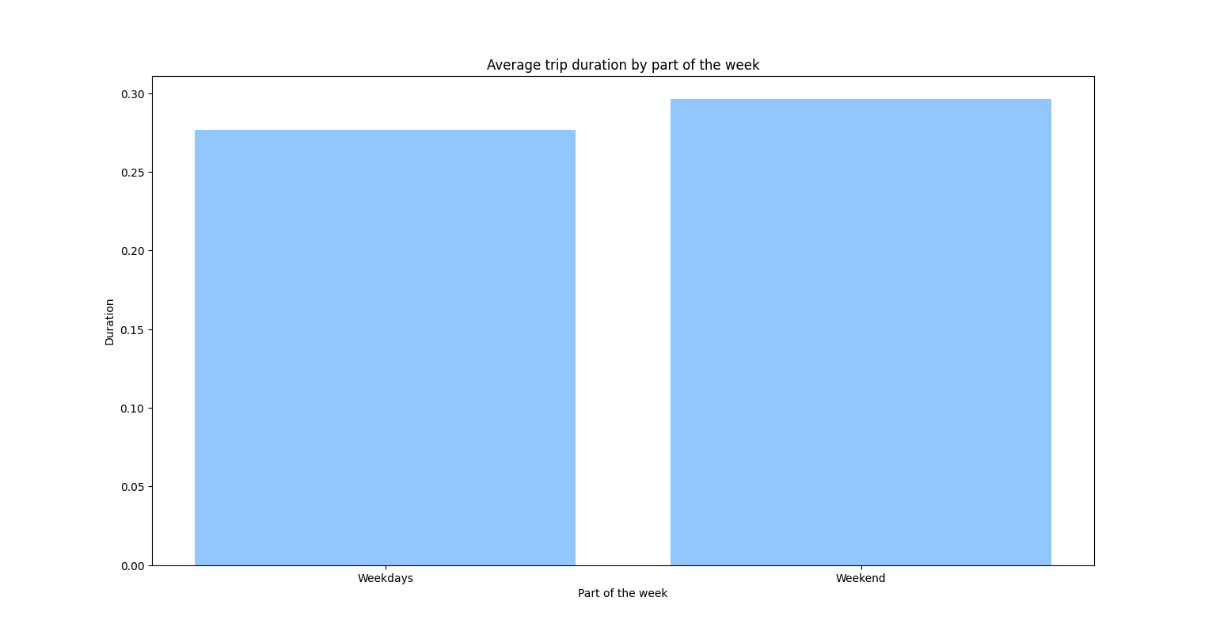


Rysunek 6. Liczba podróży w zależności od części tygodnia

Źródło: opracowanie własne

Nie są to reprezentatywne dane, ale możemy spróbować obliczyć liczbę podróży w dni powszednie dzięki danym z piątku. Mamy tylko ostatnie 16 godzin piątku, co oznacza, że aby uzyskać pełne dane, musimy pomnożyć je przez 1,5. W rezultacie otrzymujemy 170 przejazdów. W tygodniu jest 5 dni, więc wynik należy pomnożyć przez 5. W ten sposób otrzymamy 850 przejazdów (wynik ten może być reprezentatywny tylko wtedy, gdy trend wypożyczania rowerów pozostaje taki sam przez cały tydzień). W ten sam sposób uzupełnimy datę dla weekendu. Mamy pierwsze 14 godzin niedzieli, więc musimy pomnożyć sumę przez 1,715. W rezultacie otrzymujemy 100 przejazdów w niedzielę. W sobotę wykonano 189 przejazdów, a suma wynosi 289. Teraz możemy obliczyć średnią dla dni powszednich i weekendów. Wynosi ona 170 w dni powszednie i 145 w weekendy. Tak więc rower publiczny jest najbardziej poszukiwany w weekendy. Wynik ten potwierdza moje wcześniejsze teorie.

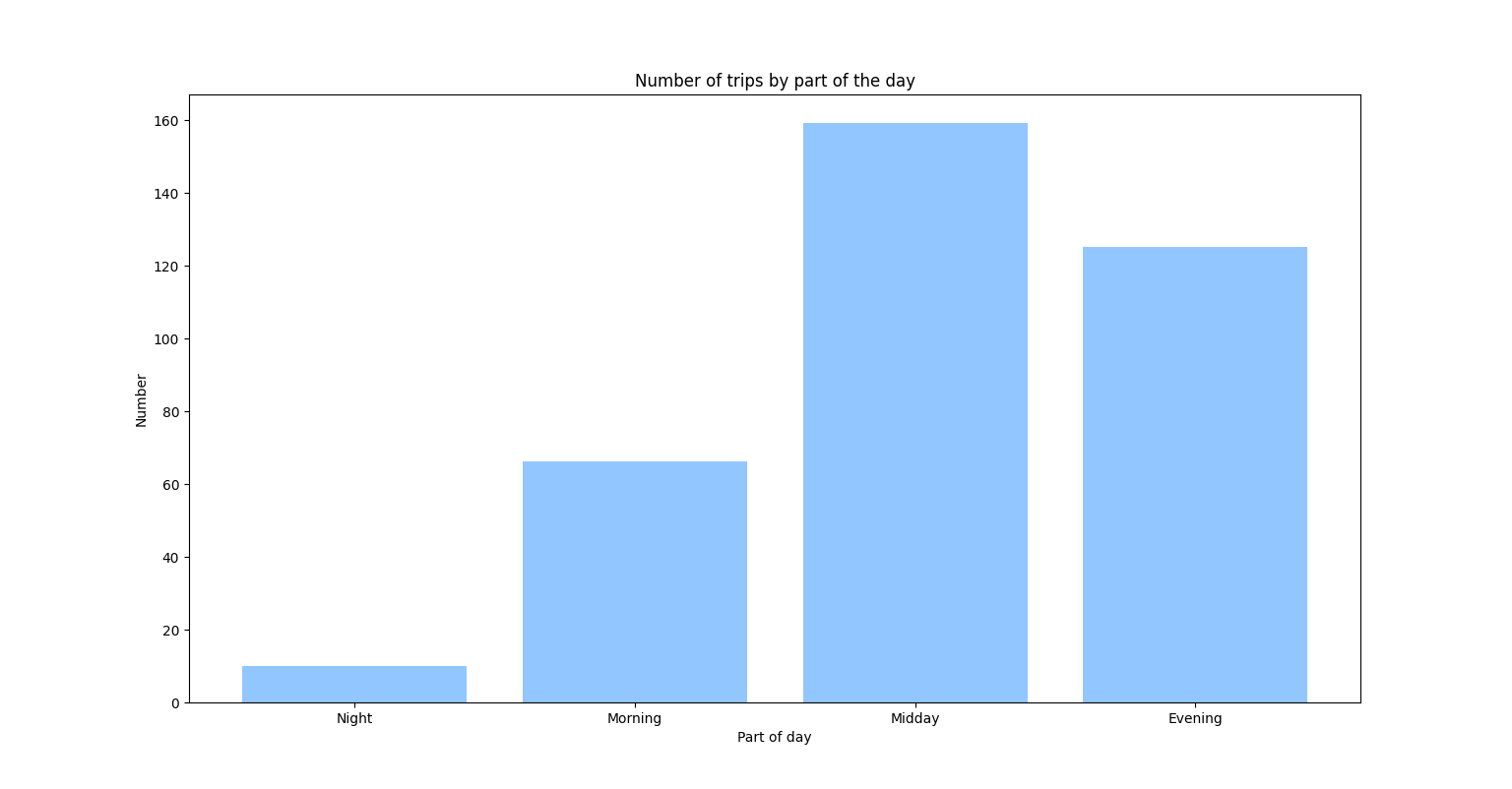
Następnie sprawdźmy i porównajmy średni czas trwania podróży w dni powszednie (możemy zaakceptować dane tylko dla piątku jako dla wszystkich dni) i weekendy. Na poniższym wykresie widzimy, że średni czas trwania podróży w weekendy jest nieco dłuższy niż w dni powszednie. Jest to tylko 1,2 minuty, więc możemy to zaokrąglić i stwierdzić, że czas trwania podróży jest taki sam.



Rysunek 7. Średni czas trwania podróży według części tygodnia

Źródło: opracowanie własne

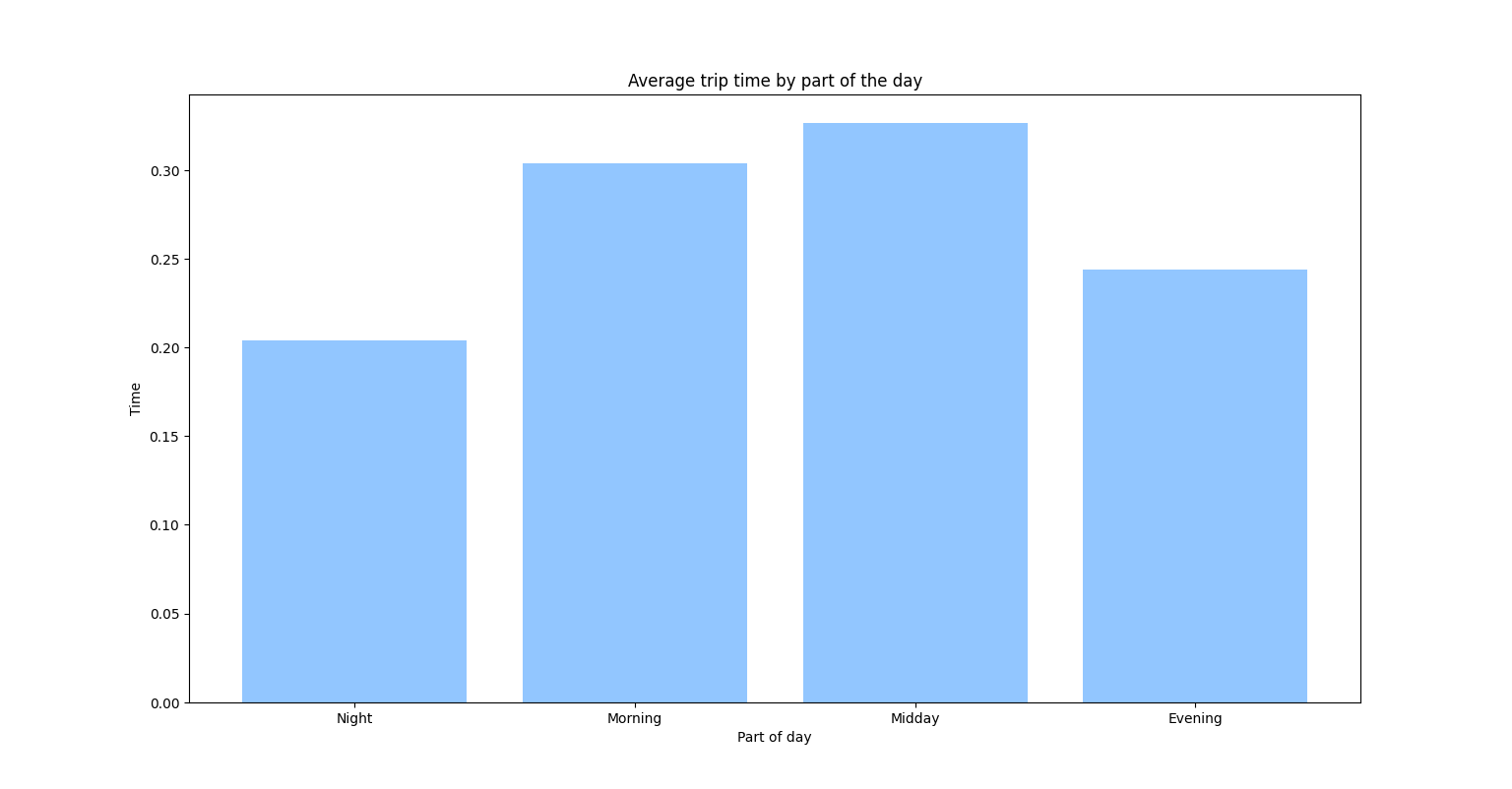
A teraz przejdźmy do analizy podróży w częściach dnia. Zacznijmy od tygodnia. Pierwszym wskaźnikiem jest liczba podróży w ciągu dnia. Jak widzimy, najwięcej podróży odbyło się w południe (159), następnie wieczorem (125), rano (66) i w nocy (10).



Rysunek 8. Liczba podróży w podziale na pory dnia (dla tygodnia)

Źródło: opracowanie własne

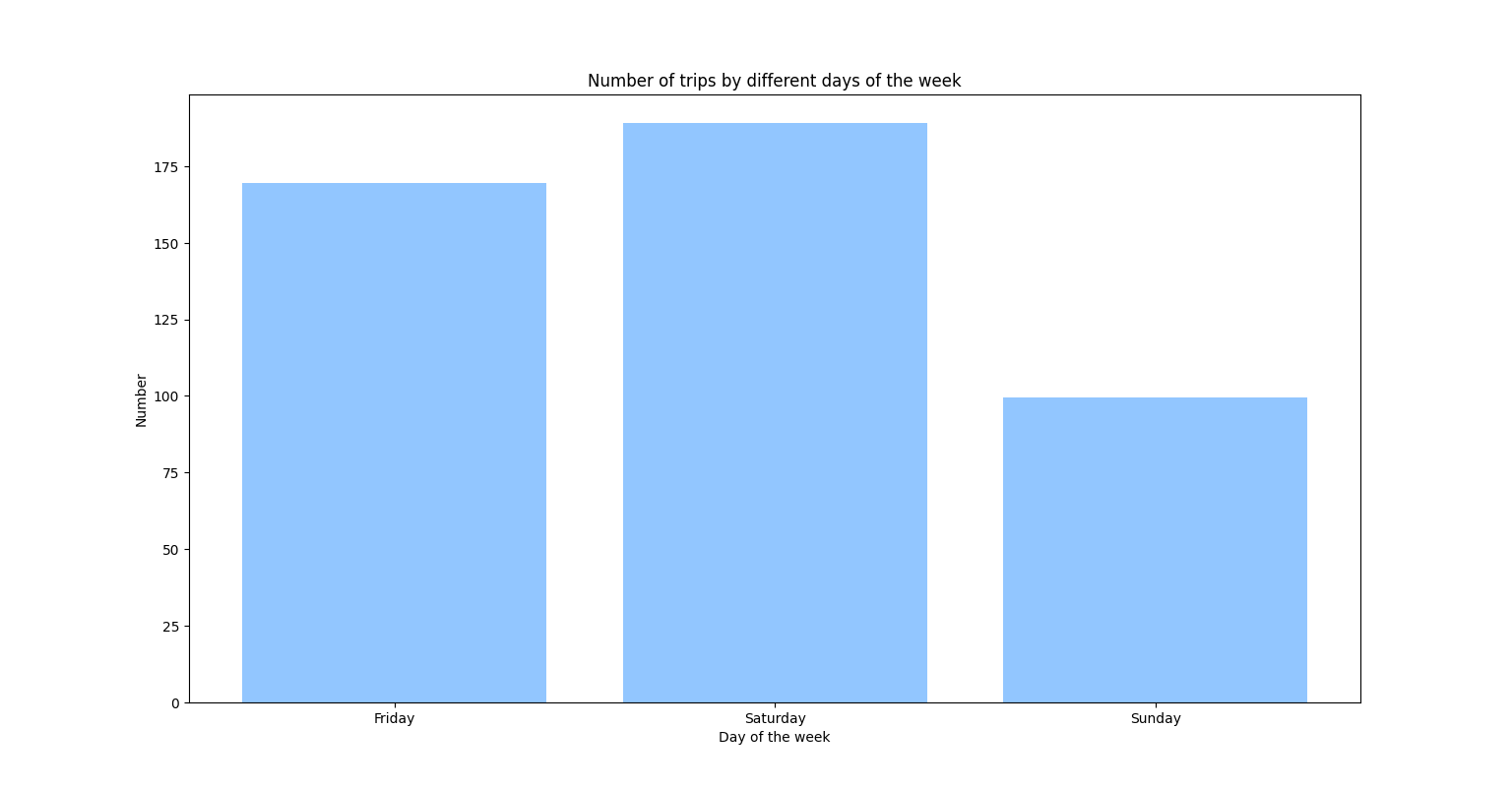
Drugim wskaźnikiem jest średni czas trwania podróży. Jeśli spojrzymy na wykres, zobaczymy, że różnica między porami dnia nie jest duża. Szczyt występuje w południe (19,2 minuty), następnie rano (18 minut), wieczorem (14,5 minuty) i w nocy (12 minut).



Rysunek 9. Średni czas trwania podróży według części dnia (dla tygodnia)

Źródło: opracowanie własne

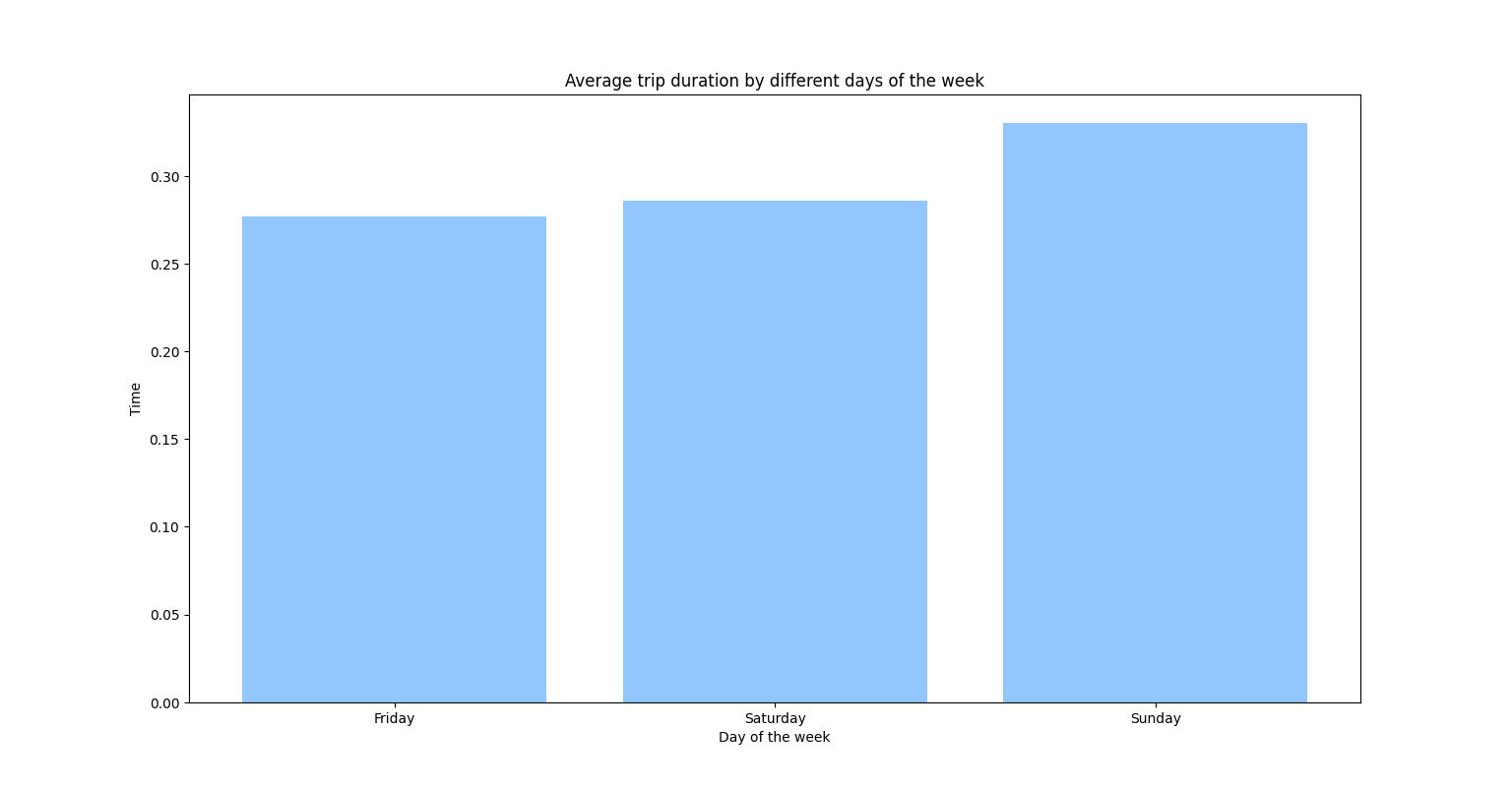
Następnie porównajmy liczbę przejazdów w różnych dniach tygodnia. Mamy dane tylko dla 3 dni (piątek i niedziela nie są pełne). Dlatego możemy zbudować bezużyteczny wykres i przeanalizować go. Ale z drugiej strony możemy pamiętać, że obliczyliśmy tę liczbę dla całego dnia. Tak więc w piątek było 170 przejazdów, a w niedzielę 100. Najwięcej podróży odbyło się w niedzielę. Oznacza to, że rowery są najbardziej potrzebne w weekendy, co nie działa na korzyść mojej teorii, że ludzie używają ich do dojazdów do pracy (ale nie można wyciągać wniosków tylko z 3 niepełnych dni danych).



Rysunek 10. Liczba podróży w poszczególnych dniach tygodnia

Źródło: opracowanie własne

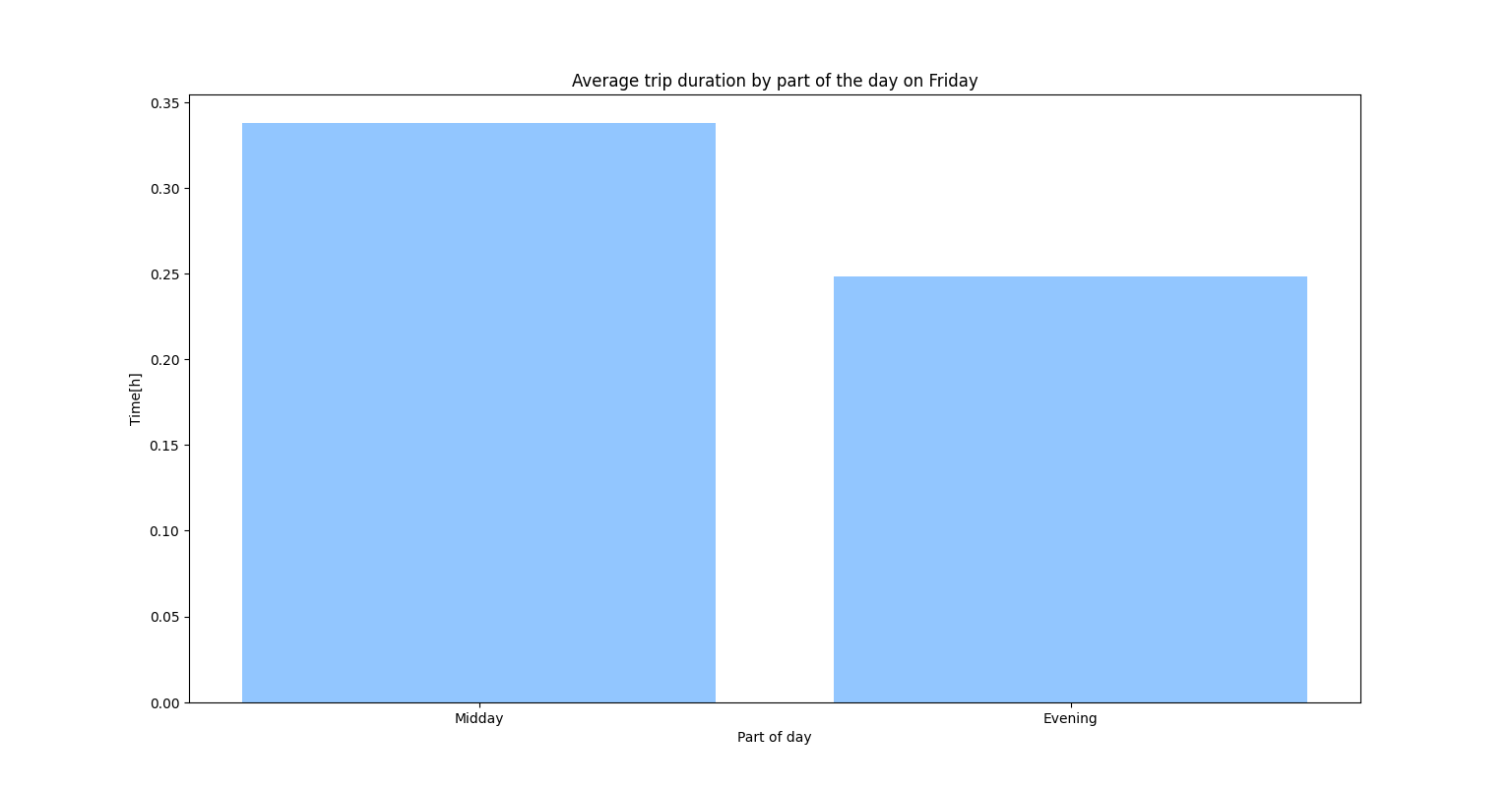
Jeśli jednak sprawdzimy średni czas trwania podróży, zobaczymy inne informacje. Średni czas trwania podróży był dłuższy w niedzielę (20 minut), a następnie w sobotę (17 minut) i piątek (16,5 minuty). Ale ta różnica nie jest aż tak duża. Oznacza to, że możemy powiedzieć, że czas trwania podróży jest taki sam.



Rysunek 11. Liczba podróży w poszczególnych dniach tygodnia

Źródło: opracowanie własne

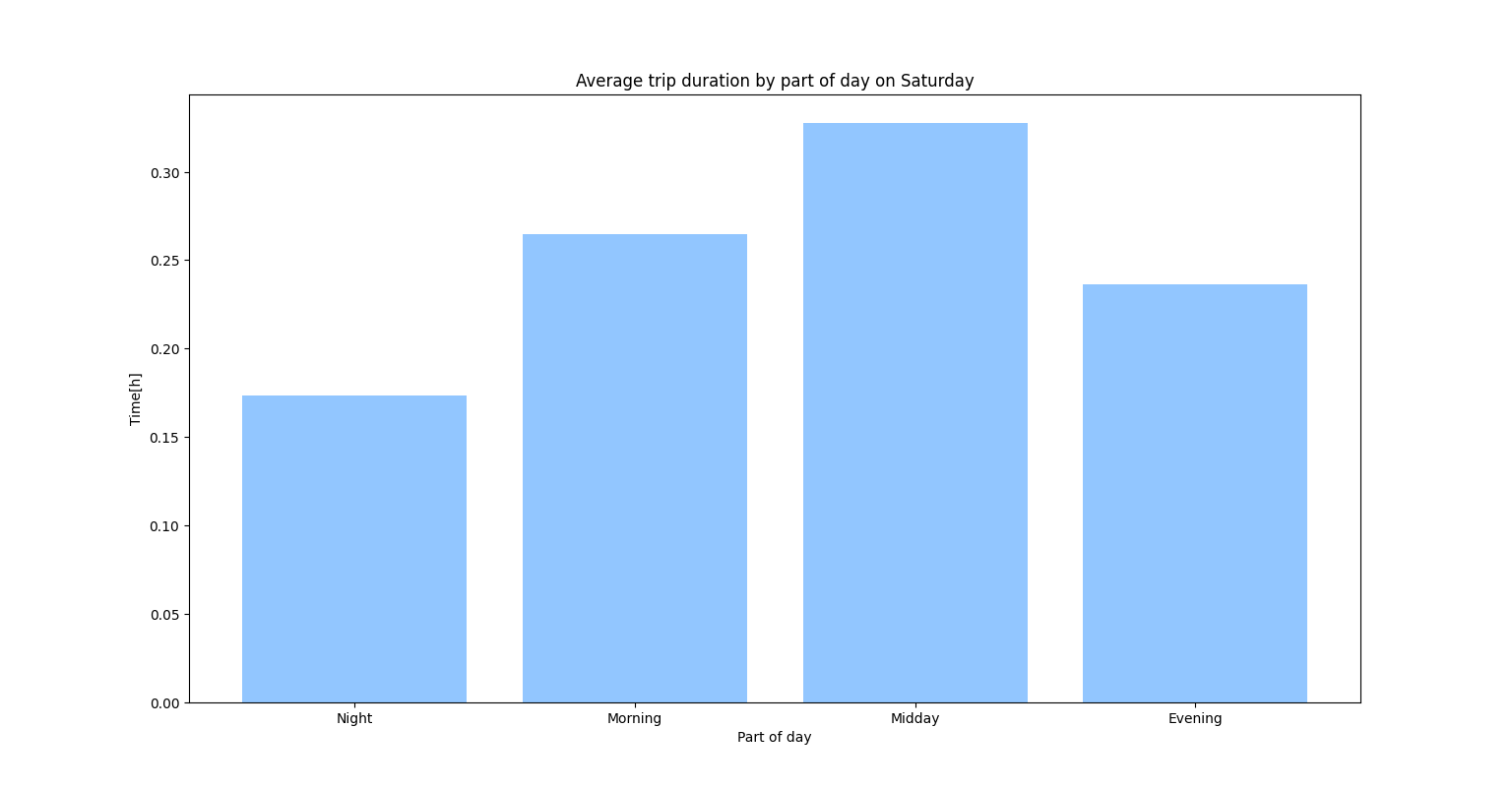
We wstępnej części naszej analizy przyjrzymy się średniemu czasowi podróży w zależności od pory dnia osobno dla piątku, soboty i niedzieli, a następnie porównamy ich wyniki. Zacznijmy od piątku. Niestety, możemy analizować tylko południe i wieczór. Jak widzimy, czas trwania podróży w południe jest o około 30% dłuższy



Rysunek 12. Średni czas trwania podróży według części dnia w piątek

Źródło: opracowanie własne

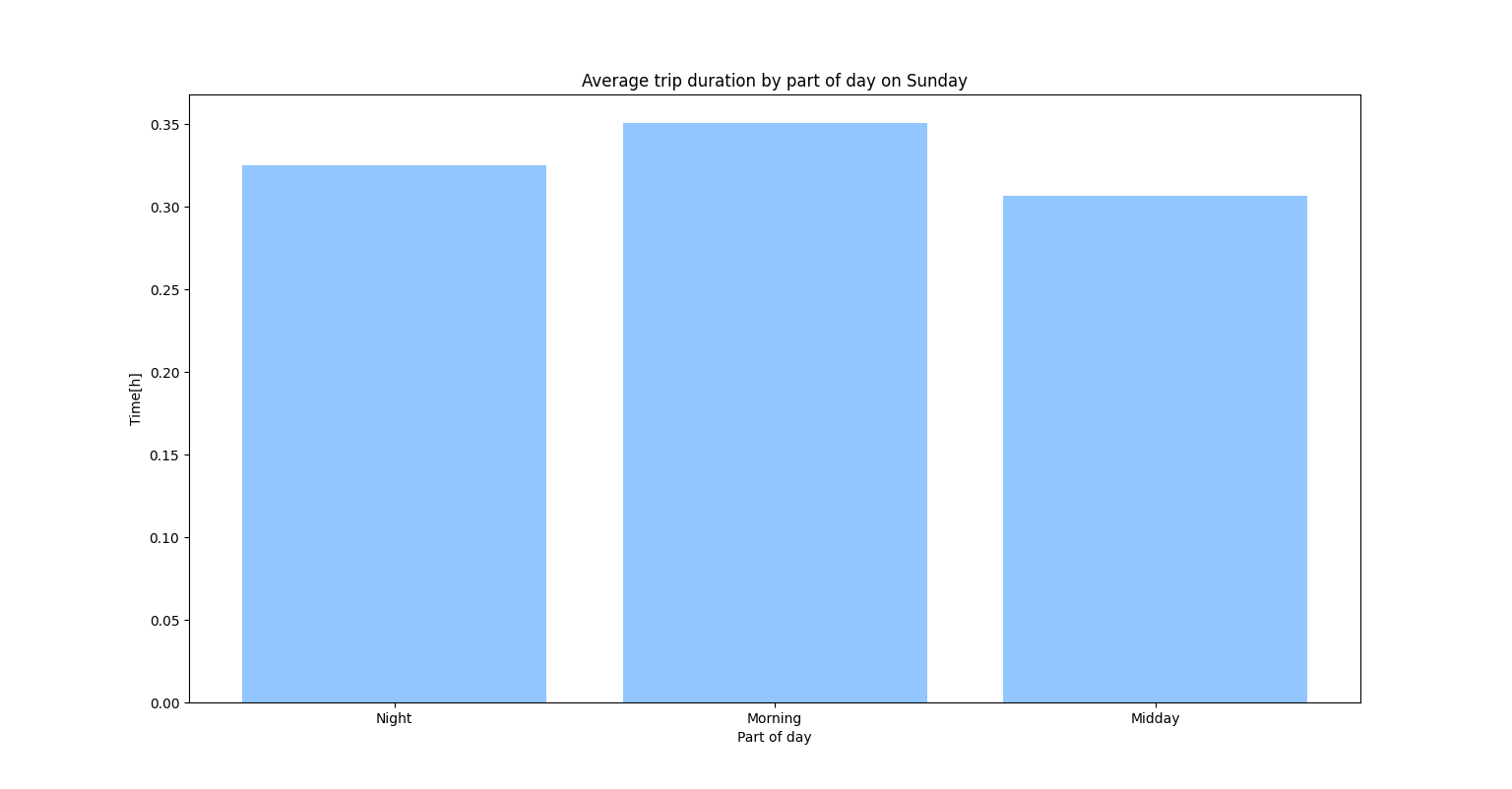
Następnie przyjrzyjmy się tym samym danym dla soboty. Tutaj trend się utrzymuje i średni czas trwania podróży jest dłuższy w środku dnia (ale o kilka minut krótszy niż w piątek), następnie rano, potem wieczorem (podobnie jak w piątek, różnica między wieczorem a środkiem dnia wynosi około 20-30%) i najkrótszy w nocy



Rysunek 13. Średni czas trwania podróży według części dnia w sobotę

Źródło: opracowanie własne

Pozostaje więc przeanalizować dane tylko dla niedzieli. Niestety nie mamy dostępnych danych po godzinie 14:00, więc nie jest możliwe przeanalizowanie średniego czasu trwania podróży w godzinach wieczornych, a tym samym nie dowiemy się, czy trend spadku aktywności w godzinach wieczornych utrzymuje się. Jak widać na wykresie, najdłuższe podróże w niedzielę nie były takie same jak w inne dni, co jest zaskakujące. Na pierwszym miejscu pod względem czasu trwania jest poranek, następnie noc, a dopiero potem środek dnia (należy jednak zaznaczyć, że dysponujemy danymi tylko dla 3 godzin tej części dnia)

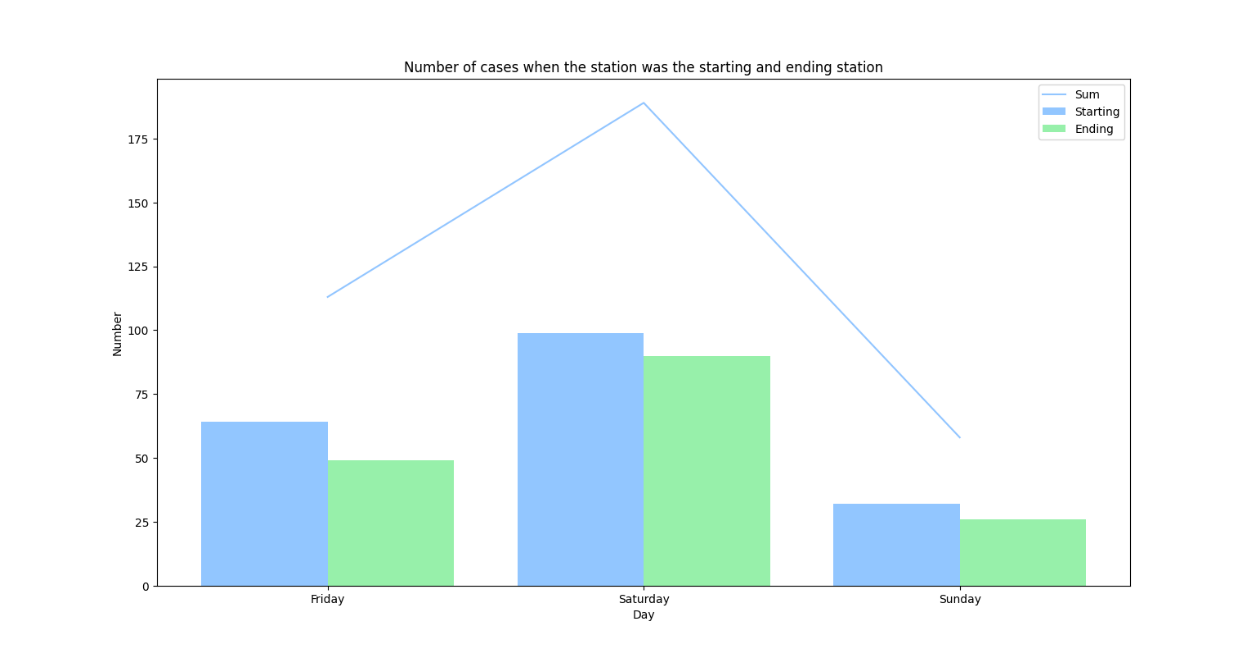


Rysunek 14. Średni czas trwania podróży według części dnia w niedzielę

Źródło: opracowanie własne

Możemy stwierdzić, że w piątek i sobotę dane są mniej więcej takie same, zarówno pod względem czasu trwania, jak i części dnia. Ale w niedzielę dane są radykalnie różne, co jest bardzo zaskakujące. Możemy wysnuć teorię, że w niedzielę ludzie, zmęczeni tygodniem pracy, odpoczywają i zaczynają poranek od roweru, co jest bardzo godne pochwały, ponieważ jest to sport, który można nazwać prawdziwym odpoczynkiem dla naszego mózgu

Ostatnią analizą dla stacji jest liczba przypadków, w których stacja była stacją początkową i końcową. Jak możemy zaobserwować na poniższym wykresie, więcej osób rozpoczęło wypożyczenie roweru ze stacji niż je zakończyło. Różnica między nimi wyniosła 30 podróży, co stanowi prawie 10% całkowitej liczby podróży (przypomnę, że było 360 podróży związanych z tą stacją).



Rysunek 15. Liczba przypadków, w których stacja była stacją początkową i końcową

Źródło: opracowanie własne

Zakończyliśmy analizę stacji i możemy rozpocząć analizę roweru.

Jedyną rzeczą, której nie mogliśmy przeanalizować w stacji i ewentualnie w rowerze, był czas przestoju. Jest to czas, w którym rower stał na różnych stacjach bez poruszania się. Aby to obliczyć, używamy następującego wzoru:

Gdzie:

– czas przestoju

– całkowity czas (46,4 godziny)

– całkowity czas pracy (3,11 godziny)

Po dokonaniu obliczeń możemy powiedzieć, że czas przestoju roweru wyniósł 43 godziny i 32 minuty. Może to oznaczać trzy rzeczy:

- wszystkie stacje, na których zatrzymał się rower, były bardzo niepopularne wśród użytkowników

- rower był zepsuty przez długi czas

- na stacjach jest tak wiele rowerów, że jeden z nich może stać bezczynnie przez 43 i pół godziny.

Jeśli przyjrzymy się tym trzem opcjom, to pierwsza wydaje mi się najmniej prawdopodobna, trzecia najbardziej prawdopodobna, a druga coś pomiędzy.

# 

# Wniosek

W tym projekcie przeanalizowaliśmy udostępnianie rowerów w rejonie Manhattanu. Na potrzeby analizy dane były dostępne przez mniej niż trzy dni (46 godzin), w tym w piątek, sobotę i niedzielę. Dlatego analizy nie można nazwać reprezentatywną, ale postaram się podsumować niektóre wyniki. Przeanalizowaliśmy dwie rzeczy, stację i rower (przeanalizowano tylko czas bezczynności, ponieważ wszystkie inne dane są już dostępne w stacji). Podsumowując wszystko, co zostało powiedziane w tym artykule, wypożyczanie rowerów jest bardzo popularne w Nowym Jorku, ponieważ w ciągu tych 46 godzin wypożyczono 50 000 rowerów. Pokazuje to, że ludzie lubią uprawiać sport i przebywać na świeżym powietrzu, a z drugiej strony nie lubią stać w korkach, które są bardzo powszechne w Nowym Jorku. Rowery są najbardziej poszukiwane w dni powszednie (obliczyliśmy to na podstawie danych z piątku). Jeśli przyjrzymy się częściom dnia, najpopularniejsze okresy dla rowerów przypadają między 12 a 18 godziną. Jeśli spojrzymy na dane dotyczące podróży, zobaczymy, że średni czas trwania podróży nie jest zbyt długi, 20 minut, a biorąc pod uwagę średnią prędkość roweru, będzie to 4,5-6 km na podróż. Oczywiście były też osoby, które korzystały z roweru przez godzinę, dwie lub trzy, ale było ich bardzo niewiele. Sugeruje to, że wypożyczanie rowerów jest bardziej popularne w przypadku krótkich podróży. Jeśli chodzi o płeć użytkowników, w 52% przypadków są to mężczyźni, a tylko w 19 przypadkach kobiety. Średnia wieku wynosi 37 lat, a większość użytkowników mieści się w przedziale wiekowym 30-50 lat. Były też osoby młodsze (10-20 lat) i starsze (70-80 lat). Na początku analizy stworzyłem teorię na temat użytkowników, a dalsza analiza wzmocniła moją wiarę w nią. Dlatego w wyniku tej pracy chciałbym stworzyć portret przeciętnego użytkownika rowerów publicznych. W 50% przypadków jest to mężczyzna w wieku od 30 do 50 lat (średnio 37 lat), który używa roweru, aby dostać się z punktu A do punktu B. Osoba ta najprawdopodobniej dojeżdża do pracy, jeśli wypożycza rower rano, lub wraca z pracy, jeśli wypożycza rower wieczorem. Lub jeśli wypożycza go w weekendy (nie jest to obowiązkowe kryterium, ludzie mogą udać się do sklepu, który znajduje się do 5 km od ich domu). Ponieważ ludzie wolą korzystać z roweru, są bardziej wysportowani i rzadziej chorują