Hurtownie danych
Laboratorium Czw 11:15

Projekt

Kajetan Pynka 254495

Spis treści

Etap 1

1. Zakres realizacji projektu

1.1. Tytuł projektu

Analiza systemu rowerów publicznych Bay Area Bike Share w San Francisco.

1.2. Charakterystyka dziedziny problemowej

System rowerów publicznych oferowany przez przedsiębiorstwa prywatne związany jest z następującymi elementami:

- Utrzymywanie rowerów wykorzystywanych przez klientów w stanie nadającym się do użytku
- Zarządzanie i zapewnianie poprawnego działania stacji rowerowych
- Zbieranie anonimowych danych ze stacji / rowerów czy też od klientów
- Zapewnianie klientom możliwości opłaty roweru ze stacji lub wygodnie z aplikacji mobilnej
- Monitorowanie stanu zapełnienia stacji rowerowych i reagowanie w odpowiednim czasie
- Prowadzenie działu obsługi klienta (telefonicznego / internetowego)

1.3. Krótki opis obszaru analizy

W ramach tego projektu skupię się na danych zebranych i udostępnionych przez byłą firmę Bay Area Bike Share. Przedstawiają one użytkowanie poszczególnych rowerów, stacji rowerowych oraz dane pogodowe w okresie między 29 sierpnia 2013r. a 1 września 2015r. Dane dotyczą stacji znajdujących się w regionie Zatoki San Francisco (pochodzą z różnych miast, a same stacje posiadają informacje o długości i szerokości geograficznej).

1.4. Problemy i potrzeby

- Zoptymalizowanie wykorzystania stacji rowerowych
- Zachęcenie użytkowników do zakupienia subskrypcji
- Analiza wpływu pogody na użytkowanie rowerów
- Analiza przychodów pod kątem: regionu (miasta, stacji), czasu (pory dnia, pory roku)
- Wykorzystanie najdłuższych wycieczek rowerowych do wyznaczenia potencjalnych miejsc nowych stacji

1.5. Cel przedsięwzięcia

1.5.1. Oczekiwania

Wykrycie trendów i korelacji pomiędzy danymi, dostarczenie prognoz na kolejne lata funkcjonowania stacji rowerowych, zaproponowanie kroków do podjęcia w celu zwiększenia zysków czy też wydajności.

1.5.2. Zakres analizy – badane aspekty

1. Sumaryczna długość wycieczek ze względu na dzień tygodnia dla każdej stacji.
2. Liczba wycieczek ze względu na zachmurzenie według miast.
3. Średnia liczba dostępnych rowerów dla stacji ze względu na miesiąc.
4. Procentowy udział klientów niezarejestrowanych oraz subskrybentów ze względu na miasto.
5. Liczba wycieczek podczas mgły ze względu na godzinę i miasto.
6. Procentowe zapełnienie stacji rowerowej nr 66 ze względu na miesiąc i godzinę.
7. Liczba wycieczek dla każdej stacji ze względu na opady (=0 – brak, T-nieznaczne, < 0.20 – średnie, > 0.20 – znaczące).
8. Średnia długość wycieczki dla każdej stacji z San Jose i rodzaju klienta.
9. Sumaryczna liczba minut z brakiem dostępnych rowerów dla każdej stacji ze względu na godzinę.
10. Zestawienie najpopularniejszej stacji docelowej dla każdej stacji ze względu na miesiąc.

1.6. Źródła danych (lokalizacja, format, dostępność)

L.p.	Plik	Typ	Liczba rekordów	Rozmiar [MB]	Opis
1	station.csv	csv	70	0.00565	Łańcuchy znaków w języku angielskim, daty w formacie MM/DD/YYYY, brak znaków specjalnych, liczby całkowite jak i zmiennoprzecinkowe o małej precyzji. Niektóre stacje zmieniły lokalizację i nazwę.
2	status.csv	csv	72.000.000	1990	Małe liczby całkowite mieszczące się w bajcie. Czas w formacie YYYY/MM/DD HH:mm:SS.
3	trip.csv	csv	670.000	80.21	Łańcuchy znaków w języku angielskim, liczby całkowite. Czas w formacie MM/DD/YYYY HH:mm. Rodzaj subskrypcji jako typ wyliczeniowy 2 łańcuchów znakowych: „Subscriber” i „Customer”.
4	weather.csv	csv	3665	0.43806	Data w formacie MM/DD/YYYY. Liczby zmiennoprzecinkowe o małej precyzji, liczby całkowite, łańcuchy znaków w języku angielskim.

2. Profilowanie danych

2.1. Analiza danych

Plik: station.csv				
L.p.	Atrybut	Typ danych	Zakres wartości	Uwagi – ocena jakości danych

Plik: status.csv				
L.p.	Atrybut	Typ danych	Zakres wartości	Uwagi – ocena jakości danych

Plik: trip.csv				
L.p.	Atrybut	Typ danych	Zakres wartości	Uwagi – ocena jakości danych

Plik: weather.csv				
L.p.	Atrybut	Typ danych	Zakres wartości	Uwagi – ocena jakości danych

2.2. Ocena przydatności danych

L.p.	Plik	Ocena jakości danych
1	station.csv	
2	status.csv	
3	trip.csv	
4	weather.csv	

2.3. Definicja typów encji/klas oraz związków

Związki:

- Czas-Pogoda
- Czas-Wycieczka
- Stacja-Wycieczka
- Czas-Status
- Stacja-Status

2.4. Propozycja wymiarów, hierarchii, miar

Wymiary:

- Czas (dzień, miesiąc, rok, godzina, minuta)
- Stacja (Nazwa stacji, długość i szerokość geograficzna, miasto, pojemność rowerowa, data instalacji)
- Długość wycieczki
- Identyfikator roweru
- Rodzaj subskrypcji

- Pogoda (data, max/min/średnia temperatura, suma opadów, zachmurzenie, zdarzenie atmosferyczne, max/min/średnia widoczność, max/min/średnia wilgotność, max/min/średnie ciśnienie)

Hierarchie:

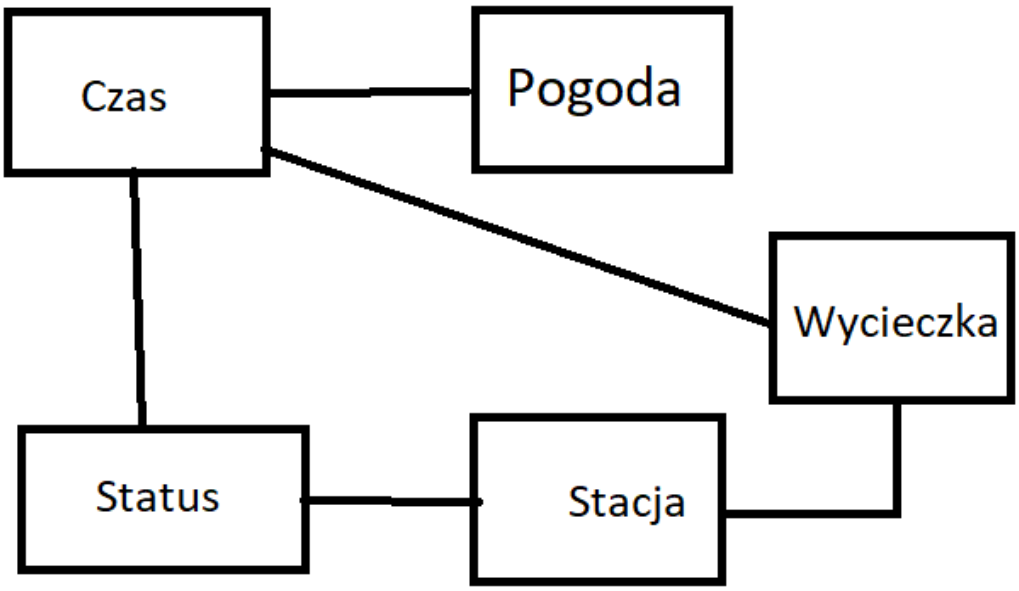
- Czas: Rok -> Miesiąc -> Dzień -> Godzina -> Minuta
- Położenie: Miasto -> Nazwa stacji

Miary:

- Długość wycieczki
- Liczba dostępnych rowerów
- Liczba dostępnych doków
- Rodzaj klienta (nieaddytywna)

2.5. Diagram klas

TYMCZASOWO: Status i wycieczka to dwie tabele faktów u mnie prawda?



3. Utworzenie bazy danych

Wnioski: