Hurtownie danych  
Laboratorium Czw 11:15  
  
Projekt  
  
Kajetan Pynka 254495

Spis treści

[Spis treści 2](#_Toc101999220)

[Etap 1 3](#_Toc101999221)

[1. Zakres realizacji projektu 3](#_Toc101999222)

[1.1. Tytuł projektu 3](#_Toc101999223)

[1.2. Charakterystyka dziedziny problemowej 3](#_Toc101999224)

[1.3. Krótki opis obszaru analizy 3](#_Toc101999225)

[1.4. Problemy i potrzeby 3](#_Toc101999226)

[1.5. Cel przedsięwzięcia 3](#_Toc101999227)

[1.5.1. Oczekiwania 3](#_Toc101999228)

[1.5.2. Zakres analizy – badane aspekty 4](#_Toc101999229)

[1.6. Źródła danych (lokalizacja, format, dostępność) 4](#_Toc101999230)

[2. Profilowanie danych 5](#_Toc101999231)

[2.1. Analiza danych 5](#_Toc101999232)

[2.2. Ocena przydatności danych 5](#_Toc101999233)

[2.3. Definicja typów encji/klas oraz związków 5](#_Toc101999234)

[2.4. Propozycja wymiarów, hierarchii, miar 5](#_Toc101999235)

[2.5. Diagram klas 6](#_Toc101999236)

[3. Utworzenie bazy danych 7](#_Toc101999237)

[Wnioski: 7](#_Toc101999238)

# 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. Procentowy udział klientów niezarejestrowanych oraz subskrybentów ze względu na miasto.
4. Liczba wycieczek podczas mgły ze względu na godzinę i miasto.
5. Liczba wycieczek dla każdej stacji ze względu na opady (=0 – brak, T-nieznaczne, < 0.20 – średnie, > 0.20 – znaczące).
6. Średnia długość wycieczki dla każdej stacji z San Jose i rodzaju klienta.
7. Zestawienie najpopularniejszej stacji docelowej dla każdej stacji ze względu na miesiąc.
8. Liczba wycieczek dla klientów zamieszkujących pod każdym kodem pocztowym ze względu na miesiąc.
9. Sumaryczna długość wycieczek dla każdego roweru ze względu na godzinę.
10. Liczba unikalnych klientów rozpoczynających lub kończących wycieczkę 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 |
| 1 | id |  |  |  |
| 2 | name |  |  |  |
| 3 | lat |  |  |  |
| 4 | long |  |  |  |
| 5 | dock\_count |  |  |  |
| 6 | city |  |  |  |
| 7 | installation\_date |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Plik: status.csv | | | | |
| L.p. | Atrybut | Typ danych | Zakres wartości | Uwagi – ocena jakości danych |
| 1 | station\_id |  |  |  |
| 2 | bikes\_available |  |  |  |
| 3 | docks\_available |  |  |  |
| 4 | time |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Plik: trip.csv | | | | |
| L.p. | Atrybut | Typ danych | Zakres wartości | Uwagi – ocena jakości danych |
| 1 | id |  |  |  |
| 2 | duration |  |  |  |
| 3 | start\_date |  |  |  |
| 4 | start\_station\_name |  |  |  |
| 5 | start\_station\_id |  |  |  |
| 6 | end\_date |  |  |  |
| 7 | end\_station\_name |  |  |  |
| 8 | end\_station\_id |  |  |  |
| 9 | bike\_id |  |  |  |
| 10 | subscription\_type |  |  |  |
| 11 | zip\_code |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Plik: weather.csv | | | | |
| L.p. | Atrybut | Typ danych | Zakres wartości | Uwagi – ocena jakości danych |
| 1 | date |  |  |  |
| 2 | Max\_temperature\_f |  |  |  |
| 3 | Mean\_temperature\_f |  |  |  |
| 4 | Min\_temperature\_f |  |  |  |
| 5 | Max\_dew\_point\_f |  |  |  |
| 6 | Mean\_dew\_point\_f |  |  |  |
| 7 | Min\_dew\_point\_f |  |  |  |
| 8 | Max\_humidity |  |  |  |
| 9 | Mean\_humidity |  |  |  |
| 10 | Min\_humidity |  |  |  |
| 11 | Max\_sea\_level\_pressure\_inches |  |  |  |
| 12 | Mean\_sea\_level\_pressure\_inches |  |  |  |
| 13 | Min\_sea\_level\_pressure\_inches |  |  |  |
| 14 | Max\_visibility\_miles |  |  |  |
| 15 | Mean\_visibility\_miles |  |  |  |
| 16 | Min\_visibility\_miles |  |  |  |
| 17 | Max\_wind\_speed\_mph |  |  |  |
| 18 | Mean\_wind\_speed\_mph |  |  |  |
| 19 | Max\_gust\_speed\_mph |  |  |  |
| 20 | Precipitation\_inches |  |  |  |
| 21 | Cloud\_cover |  |  |  |
| 22 | Events |  |  |  |
| 23 | Wind\_dir\_degrees |  |  |  |
| 24 | Zip\_code |  |  |  |

## 2.2. Ocena przydatności danych

|  |  |  |
| --- | --- | --- |
| L.p. | Plik | Ocena jakości danych |
| 1 | station.csv | Brak pustych pól, wszystkie są poprawne. Dla niektórych stacji zmieniła się nazwa oraz położenie (nie jest to problemem ponieważ dalej obowiązuje ten sam identyfikator). |
| 2 | status.csv | W pełni poprawne dane, 3 niewielkie liczby całkowite wraz z czasem co do sekundy (w praktyce każdy zapis był dokonywany co minutę). Format YYYY/MM/DD HH:mm jest akceptowalny. NIEPRZYDATNE |
| 3 | trip.csv | Wszystkie kolumny, poza jedną, są wypełnione poprawnymi danymi i są w pełni użyteczne. Należy pamiętać o tym, że również występuję tu kwestia zmiany nazw niektórych stacji. |
| 4 | weather.csv |  |

## 2.3. Definicja typów encji/klas oraz związków

Związki:

* Czas-Pogoda
* Czas-Wycieczka
* Stacja-Wycieczka
* Stacja-Czas
* 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

Diagram

Description automatically generated

## 3. Utworzenie bazy danych

## Wnioski: