Hurtownie danych  
Laboratorium Czw 11:15  
  
Projekt  
  
Kajetan Pynka 254495

Spis treści

[Spis treści 2](#_Toc104414263)

[Etap 1 3](#_Toc104414264)

[1. Zakres realizacji projektu 3](#_Toc104414265)

[1.1. Tytuł projektu 3](#_Toc104414266)

[1.2. Charakterystyka dziedziny problemowej 3](#_Toc104414267)

[1.3. Krótki opis obszaru analizy 3](#_Toc104414268)

[1.4. Problemy i potrzeby 3](#_Toc104414269)

[1.5. Cel przedsięwzięcia 3](#_Toc104414270)

[1.5.1. Oczekiwania 3](#_Toc104414271)

[1.5.2. Zakres analizy – badane aspekty 4](#_Toc104414272)

[1.6. Źródła danych (lokalizacja, format, dostępność) 4](#_Toc104414273)

[2. Profilowanie danych 5](#_Toc104414274)

[2.1. Analiza danych 5](#_Toc104414275)

[2.2. Ocena przydatności danych 8](#_Toc104414276)

[2.3. Definicja typów encji/klas oraz związków 9](#_Toc104414277)

[2.4. Propozycja wymiarów, hierarchii, miar 14](#_Toc104414278)

[2.5. Diagram klas 15](#_Toc104414279)

[3. Utworzenie bazy danych 15](#_Toc104414280)

[Wnioski: 15](#_Toc104414281)

[Etap 2 16](#_Toc104414282)

[Dokumentacja procesu ETL 16](#_Toc104414283)

[Mapa logiczna procesu ETL 21](#_Toc104414284)

[Wnioski: 25](#_Toc104414285)

[Etap 3 26](#_Toc104414286)

[Dokumentacja kostki 26](#_Toc104414287)

[Zaplanowane zestawienia (1.5.2) 30](#_Toc104414288)

[Analiza zestawień 41](#_Toc104414289)

[Analiza w głąb 42](#_Toc104414290)

[Wnioski: 48](#_Toc104414291)

# 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.
7. Zestawienie największej liczby wycieczek 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 | int | 2-84 | W pełni poprawne dane |
| 2 | name | varchar(45) | 9-45 znaków | W pełni poprawne dane, nastąpiła zmiana nazw niektórych stacji |
| 3 | lat | float(6, 4) | 37.3297-37.8048 | W pełni poprawne dane, nastąpiła zmiana położenia niektórych stacji |
| 4 | long | float(7, 4) | -122.419  do -121.8773 | W pełni poprawne dane, nastąpiła zmiana położenia niektórych stacji |
| 5 | dock\_count | int | 11-27 | W pełni poprawne dane |
| 6 | city | varchar(13) | 8-13 znaków | W pełni poprawne dane |
| 7 | installation\_date | datetime | 8/5/2013 – 4/9/2014 | W pełni poprawne dane, format MM/DD/YYYY |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Plik: status.csv | | | | |
| L.p. | Atrybut | Typ danych | Zakres wartości | Uwagi – ocena jakości danych |
| 1 | station\_id | int | 2-84 | W pełni poprawne dane |
| 2 | bikes\_available | int | 0-27 | W pełni poprawne dane |
| 3 | docks\_available | int | 0-27 | W pełni poprawne dane |
| 4 | time | datetime | 2013/08/29 12:06:01 – 2015/08/31 12:06:01 | W pełni poprawne dane, czas w formacie YYYY/MM/DD HH:mm:SS. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Plik: trip.csv | | | | |
| L.p. | Atrybut | Typ danych | Zakres wartości | Uwagi – ocena jakości danych |
| 1 | id | int | 4079-913459 | W pełni poprawne dane |
| 2 | duration | int | 60-17270400 | W pełni poprawne dane, czas mierzony w sekundach |
| 3 | start\_date | datetime | 8/29/2013 9:38 – 8/31/2015 23:26 | W pełni poprawne dane, data w formacie MM/DD/YYYY HH:mm. |
| 4 | start\_station\_name | varchar(45) | 9-45 | W pełni poprawne dane, nazwa niektórych stacji uległa zmianie w czasie |
| 5 | start\_station\_id | int | 2-84 | W pełni poprawne dane |
| 6 | end\_date | datetime | 8/29/2013 9:41 – 8/31/2015 23:39 | W pełni poprawne dane, data w formacie MM/DD/YYYY HH:mm. |
| 7 | end\_station\_name | varchar(45) | 9-45 | W pełni poprawne dane, nazwa niektórych stacji uległa zmianie w czasie |
| 8 | end\_station\_id | int | 2-84 | W pełni poprawne dane |
| 9 | bike\_id | int | 9-878 | W pełni poprawne dane |
| 10 | subscription\_type | varchar(10) | 8-10 znaków | W pełni poprawne dane, przyjmuje zasadniczo dwie wartości: „Subscriber” oraz „Customer” |
| 11 | zip\_code | varchar(11) | 1-11 znaków | Około 1% rekordów posiada wartości puste, około 2% rekordów posiada nieprawidłowe wartości: zdecydownie za mało albo za dużo cyfr, występują też przypadki liter. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Plik: weather.csv | | | | |
| L.p. | Atrybut | Typ danych | Zakres wartości | Uwagi – ocena jakości danych |
| 1 | date | datetime | 8/29/2013 – 8/31/2015 | W pełni poprawne dane, data w formacie MM/DD/YYYY |
| 2 | Max\_temperature\_f | int | 44-102 | Występują 4 rekordy puste (mniej niż 1% wszystkich) |
| 3 | Mean\_temperature\_f | int | 38-84 | Występują 4 rekordy puste (mniej niż 1% wszystkich) |
| 4 | Min\_temperature\_f | int | 25-75 | Występują 4 rekordy puste (mniej niż 1% wszystkich) |
| 5 | Max\_dew\_point\_f | int | 20-68 | Występują 54 rekordy puste (około 1% wszystkich) |
| 6 | Mean\_dew\_point\_f | int | 13-65 | Występują 54 rekordy puste (około 1% wszystkich) |
| 7 | Min\_dew\_point\_f | int | 2-63 | Występują 54 rekordy puste (około 1% wszystkich) |
| 8 | Max\_humidity | int | 24-100 | Występują 54 rekordy puste (około 1% wszystkich) |
| 9 | Mean\_humidity | int | 24-96 | Występują 54 rekordy puste (około 1% wszystkich) |
| 10 | Min\_humidity | int | 4-93 | Występują 54 rekordy puste (około 1% wszystkich) |
| 11 | Max\_sea\_level\_pressure\_inches | float(4,2) | 29.5-30.65 | Występuje 1 rekord pusty (mniej niż 1% wszystkich) |
| 12 | Mean\_sea\_level\_pressure\_inches | float(4,2) | 29.43-30.41 | Występuje 1 rekord pusty (mniej niż 1% wszystkich) |
| 13 | Min\_sea\_level\_pressure\_inches | float(4,2) | 28.98-30.37 | Występuje 1 rekord pusty (mniej niż 1% wszystkich) |
| 14 | Max\_visibility\_miles | int | 5-20 | Występuje 13 rekordów pustych (mniej niż 1% wszystkich) |
| 15 | Mean\_visibility\_miles | int | 4-20 | Występuje 13 rekordów pustych (mniej niż 1% wszystkich) |
| 16 | Min\_visibility\_miles | int | 0-20 | Występuje 13 rekordów pustych (mniej niż 1% wszystkich) |
| 17 | Max\_wind\_speed\_mph | int | 0-128 | Występuje 1 rekord pusty (mniej niż 1% wszystkich) |
| 18 | Mean\_wind\_speed\_mph | int | 0-23 | Występuje 1 rekord pusty (mniej niż 1% wszystkich) |
| 19 | Max\_gust\_speed\_mph | int | 6-114 | 25% rekordów jest pustych |
| 20 | Precipitation\_inches | varchar(4) | 1-4 znaków | Występuje 1 rekord pusty. W 96% przypadków jest to float(4,2) natomiast dla 4% rekordów znak ‘T’ oznaczający nieznaczne opady. |
| 21 | Cloud\_cover | int | 0-8 | Występuje 1 rekord pusty (mniej niż 1% wszystkich) |
| 22 | Events | varchar(17) | 3-17 znaków | 86% rekordów jest pustych, pozostałe posiadają jedną z pięciu wartości: „Rain”, „Fog”, „Fog-Rain”, „Rain-Thunderstorm”, „rain”. |
| 23 | Wind\_dir\_degrees | int | 0-2772 | Występuje 1 rekord pusty (mniej niż 1% wszystkich) |
| 24 | Zip\_code | int | 94041-95113 | Dane w pełni poprawne, występuje pięć kodów pocztowych: 94107, 94063, 94301, 94091, 95113 |

## 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. Ogólnie jednak dane nie są przydatne jeśli o chodzi o założenia projektowe, więc można zignorować ten plik. |
| 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. Kod pocztowy po oczyszczeniu będzie się nadawał do dalszej analizy. |
| 4 | weather.csv | W większości kolumn występują marginalne brakujące dane, nieistotne dane atmosferyczne możemy odrzucić na potrzeby dalszych analiz. Liczba opadów pozostanie jako łańcuch znakowy, należy pamiętać o znaku ‘T’ jako jednej z możliwości tej kolumny. Należy oczyścić kolumnę zdarzeń atmosferycznych i połączyć „Rain” oraz „rain” w jedno zdarzenie. Kod pocztowy do przekształcenia na miasto (wtedy odpowiada miastu ze stacji). |

## 2.3. Definicja typów encji/klas oraz związków

**Encje:**

|  |  |  |  |
| --- | --- | --- | --- |
| Encja: **DIM\_TIME** | | | |
| Nazwa atrybutu | Opis atrybutu | Typ | OBL(+) OPC(-) |
| PK\_TIME | Klucz główny, identyfikator w formie łańcucha znaków | varchar(12) | + |
| Year | Rok zapisany jako liczba całkowita | integer | + |
| Month | Miesiąc zapisany jako liczba całkowita | integer | + |
| Month\_Name | Ciąg znaków reprezentujący nazwę miesiąca | varchar(9) | + |
| Day | Dzień zapisany jako liczba całkowita | integer | + |
| Week\_Day | Ciąg znaków reprezentujący nazwę dnia tygodnia | varchar(9) | + |
| Hour | Godzina zapisana jako liczba całkowita | integer | - |
| Minute | Minuta zapisana jako liczba całkowita | integer | - |

|  |  |  |  |
| --- | --- | --- | --- |
| Encja: **DIM\_STATION** | | | |
| Nazwa atrybutu | Opis atrybutu | Typ | OBL(+) OPC(-) |
| STATION\_ID | Klucz główny, identyfikator pojedynczej stacji | integer | + |
| Name | Nazwa stacji rowerowej zapisana jako łańcuch znaków | varchar(45) | + |
| Lat | Szerokość geograficzna położenia stacji | float(6,4) | + |
| Long | Długość geograficzna położenia stacji | float(7,4) | + |
| Dock\_Count | Liczba możliwych rowerów do zaparkowania w stacji | integer | + |
| City | Ciąg znaków oznaczający nazwę miasta, w której znajduje się stacja | varchar(13) | + |
| Installation\_Date | Klucz obcy, referencja do czasu przechowywanego w DIM\_TIME | varchar(12) | + |

|  |  |  |  |
| --- | --- | --- | --- |
| Encja: **DIM\_BIKE** | | | |
| Nazwa atrybutu | Opis atrybutu | Typ | OBL(+) OPC(-) |
| BIKE\_ID | Klucz główny, identyfikator roweru | integer | + |

|  |  |  |  |
| --- | --- | --- | --- |
| Encja: **DIM\_CUSTOMER** | | | |
| Nazwa atrybutu | Opis atrybutu | Typ | OBL(+) OPC(-) |
| CUSTOMER\_ID | Klucz główny, identyfikator klienta w formie liczby całkowitej | integer | + |
| Cust\_Zip\_Code | Łańcuch znaków reprezentujący kod pocztowy klienta, zły kod pocztowy zastąpiony przez NULL | varchar(11) | - |
| Subscription\_Type | Łańcuch znaków reprezentujący status klienta: „Customer” albo „Subscriber” | varchar(10) | + |

|  |  |  |  |
| --- | --- | --- | --- |
| Encja: **DIM\_WEATHER** | | | |
| Nazwa atrybutu | Opis atrybutu | Typ | OBL(+) OPC(-) |
| WEATHER\_ID | Klucz główny, identyfikator pogody w formie liczby całkowitej | integer | + |
| Measure\_Date | Klucz obcy z referencją do czasu przechowywanego w DIM\_TIME | varchar(12) | + |
| Measure\_City | Ciąg znaków oznaczający miasto, w którym dokonanu pomiaru | varchar(13) | + |
| Max\_Temperature | Liczba całkowita oznaczająca maksymalną osiągniętą temperaturę | integer | - |
| Mean\_Temperature | Liczba całkowita oznaczającą średnią osiągniętą temperaturę | integer | - |
| Min\_Temperature | Liczba całkowita oznaczającą minimalną osiągniętą temperaturę | integer | - |
| Max\_Humidity | Liczba całkowita oznaczająca maksymalną osiągniętą wilgotność | integer | - |
| Mean\_Humidity | Liczba całkowita oznaczająca średnią osiągniętą wilgotność | integer | - |
| Min\_Humidity | Liczba całkowita oznaczająca minimalną osiągniętą wilgotność | integer | - |
| Max\_Pressure | Liczba zmiennoprzecinkowa oznaczająca maksymalne osiągnięte ciśnienie | float(4,2) | - |
| Mean\_Pressure | Liczba zmiennoprzecinkowa oznaczająca średnie osiągnięte ciśnienie | float(4,2) | - |
| Min\_Pressure | Liczba zmiennoprzecinkowa oznaczająca minimalne osiągnięte ciśnienie | float(4,2) | - |
| Max\_Visibility | Liczba całkowita oznaczająca maksymalną widoczność w milach | integer | - |
| Mean\_Visibility | Liczba całkowita oznaczająca średnią widoczność w milach | integer | - |
| Min\_Visibility | Liczba całkowita oznaczająca minimalną widoczność w milach | integer | - |
| Precipitation\_Inches | Łańcuch znaków określający liczbę opadów w calach lub znak ‘T’ gdy opady były niewielkie. | varchar(4) | - |
| Cloud\_Cover | Liczba całkowita oznaczająca zachmurzenie w skali 0-8 | integer | - |
| Events | Zdarzenie atmosferyczne zapisane jako ciąg znaków, jedno z 4 wydarzeń: „Rain”, „Fog”, „Fog-Rain”, „Rain-Thunderstorm” | varchar(17) | - |

|  |  |  |  |
| --- | --- | --- | --- |
| Encja: **FACT\_TRIP** | | | |
| Nazwa atrybutu | Opis atrybutu | Typ | OBL(+) OPC(-) |
| TRIP\_ID | Klucz główny, liczba całkowita oznaczająca identyfikator wycieczki | integer | + |
| Start\_Time | Klucz obcy, referencja czasu przechowywanego w DIM\_TIME | varchar(12) | + |
| End\_Time | Klucz obcy, referencja czasu przechowywanego w DIM\_TIME | varchar(12) | + |
| Start\_Station | Klucz obcy, referencja do identyfikatora stacji z DIM\_STATION | integer | + |
| End\_Station | Klucz obcy, referencja do identyfikatora stacji z DIM\_STATION | integer | + |
| Start\_Station\_Name | Ciąg znaków reprezentujący nazwę stacji startowej | varchar(45) | + |
| End\_Station\_Name | Ciąg znaków reprezentujący nazwę stacji końcowej | varchar(45) | + |
| Trip\_Customer | Klucz obcy, referencja do klienta z DIM\_CUSTOMER | integer | + |
| Trip\_Bike | Klucz obcy, referencja do roweru z DIM\_BIKE | integer | + |
| Trip\_Weather | Klucz obcy, referencja do pogody z DIM\_WEATHER | integer | + |
| Duration | Liczba całkowita oznaczająca długość wycieczki w sekundach | integer | + |

**Związki:**

* DIM\_TIME(1) – (0..\*)DIM\_WEATHER
* DIM\_TIME(1) – (0..\*)DIM\_STATION
* DIM\_TIME(2) – (0..\*)FACT\_TRIP
* DIM\_STATION(2) – (0..\*)FACT\_TRIP
* DIM\_WEATHER(1) – (0..\*)FACT\_TRIP
* DIM\_CUSTOMER(1) – (0..\*)FACT\_TRIP
* DIM\_BIKE(1) – (0..\*)FACT\_TRIP

## 2.4. Propozycja wymiarów, hierarchii, miar

**Wymiary:**

* DIM\_TIME
* DIM\_STATION
* DIM\_BIKE
* DIM\_CUSTOMER
* DIM\_WEATHER

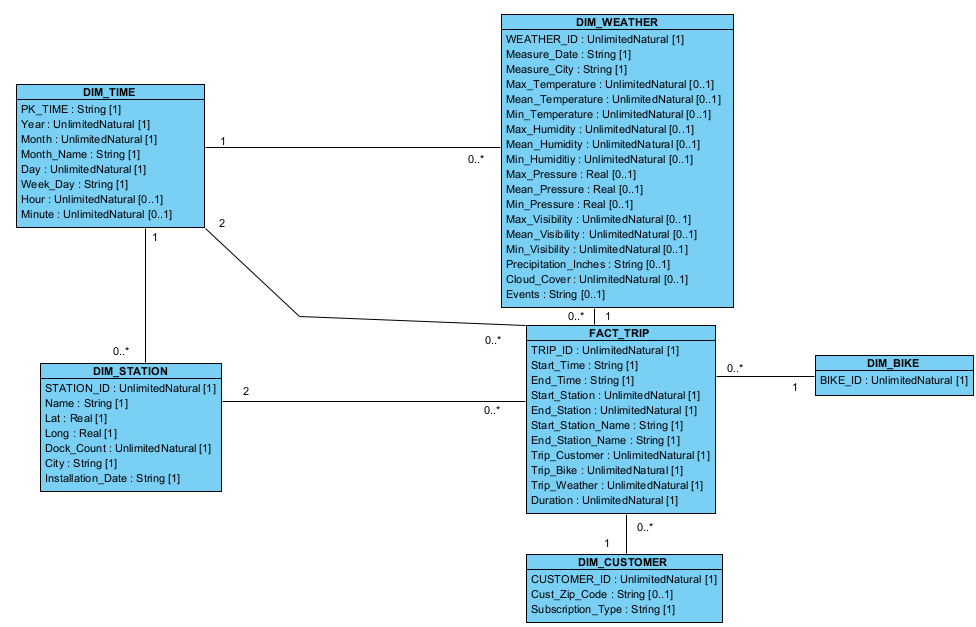
**Hierarchie:**

* DIM\_TIME: Year -> Month -> Day -> Hour -> Minute
* DIM\_STATION: City -> Name

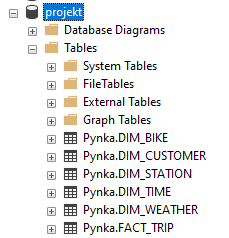
**Miary:**

* Długość wycieczki
* Liczba unikalnych klientów
* Liczba unikalnych rowerów
* Liczba wycieczek

## 2.5. Diagram klas



## 3. Utworzenie bazy danych



Rysunek przedstawia utworzoną strukturę bazy danych.

## Wnioski:

W celu przygotowania odpowiedniej hurtowni danych musimy być świadomi wymagań biznesu by wiedzieć po co nam ma służyć dana hurtownia. Następnie, po dogłębnej analizie danych możemy wstępnie zaplanować naszą hurtownię, tak aby móc w przyszłości wytworzyć kostkę, a co za tym idzie by móc skutecznie dokonywać analizy danych znajdujących się w hurtowni.

# Etap 2

## Dokumentacja procesu ETL

|  |  |
| --- | --- |
| Główny przepływ zadań procesu ETL | |
| **Nazwa zadania** | **Opis zadania** |
| DELETE | Zadanie wykonujące kwerendę SQL, która usuwa wszystkie istniejące w hurtowni tabele faktów i wymiarów. |
| CREATE | Zadanie wykonujące kwerendę SQL, która tworzy w hurtowni tabele faktów oraz wymiarów (samą ich strukturę). |
| INSERT DIM\_STATION | Blok podzadań, który wypełnia danymi wymiar DIM\_STATION. |
| INSERT DIM\_BIKE | Blok podzadań, który wypełnia danymi wymiar DIM\_BIKE. |
| INSERT DIM\_CUSTOMER | Blok podzadań, który wypełnia danymi wymiar DIM\_CUSTOMER. |
| INSERT DIM\_WEATHER | Blok podzadań, który wypełnia danymi wymiar DIM\_WEATHER. |
| CREATE HELPER TABLES | Zadanie wykonujące kwerendę SQL, która tworzy pomocnicze tabele ENUM\_MONTH oraz ENUM\_WEEKDAY oraz wypełnia je danymi. |
| INSERT DIM\_TIME | Blok podzadań, który wypełnia danymi wymiar DIM\_TIME. |
| INSERT FACT\_TRIP | Blok podzadań, który wypełnia danymi tabelę faktów FACT\_TRIP. |
| REFERENCES | Zadanie wykonujące kwerendę SQL, która tworzy więzy integralności (klucze główne i obce) dla tabel faktów i wymiarów. |
| CLEANUP | Zadanie wykonujące kwerendę SQL, która usuwa występujące w danych anomalie lub błędne wartości. |

Dokumentacja podzadań:

|  |  |
| --- | --- |
| DIM\_STATION | |
| **Nazwa zadania** | **Opis zadania** |
| Station\_csv | Zadanie wczytuje wszystkie dane z pliku station.csv. |
| Convert Installation\_Date to DATE | Zadanie rzutuje kolumnę installation\_date na DATE. |
| Format Installation\_Date | Zadanie formatuje datę z kolumny installation\_date postaci MM/dd/yyyy na YYYYMMDDXXXX, gdzie X to po prostu znak ‘X’. |
| DIM\_STATION | Zadanie zapisuje przekształcone dane do tabeli DIM\_STATION. |

|  |  |
| --- | --- |
| DIM\_BIKE | |
| **Nazwa zadania** | **Opis zadania** |
| Trip\_csv | Zadanie wczytuje kolumnę bike\_id z pliku trip.csv |
| Cast bike\_id to int | Zadanie rzutuje kolumnę bike\_id na 4-bitową liczbę całkowitą. |
| Remove duplicate ids | Zadanie usuwa duplikaty wartości w kolumnie bike\_id. |
| DIM\_BIKE | Zadanie zapisuje przekształcone dane do tabeli DIM\_BIKE. |

|  |  |
| --- | --- |
| DIM\_CUSTOMER | |
| **Nazwa zadania** | **Opis zadania** |
| Trip\_csv | Zadanie wczytuje kolumny zip\_code oraz subscription\_type z pliku trip.csv |
| Remove duplicate rows | Zadanie usuwa duplikaty rekordów (zip\_code, subscription\_type). |
| DIM\_CUSTOMER | Zadanie zapisuje przekształcone dane do tabeli DIM\_CUSTOMER. |

|  |  |
| --- | --- |
| INSERT DIM\_WEATHER | |
| **Nazwa zadania** | **Opis zadania** |
| Weather\_csv | Zadanie wczytuje wszystkie dane z pliku weather.csv. |
| Cast date from str to date | Zadanie rzutuje kolumnę date na DATE. |
| Format date | Zadanie formatuje datę z kolumny date z postaci MM/dd/yyyy na YYYYMMDDXXXX, gdzie X to po prostu znak ‘X’. |
| Convert zipcode to city name | Zadanie tworzy nową kolumnę city, w której znajduje się nazwa miasta odpowiadająca jednemu z pięciu kodów pocztowych w polu zip\_code. |
| Convert city from Unicode to varchar | Zadanie rzutuje nazwę miasta z formatu Unicode na format tekstowy z kodem regionu 1250. |
| DIM\_WEATHER | Zadanie zapisuje przekształcone dane do tabeli DIM\_WEATHER. |

|  |  |
| --- | --- |
| INSERT DIM\_TIME | |
| **Nazwa zadania** | **Opis zadania** |
| Station\_csv | Zadanie wczytuje kolumnę installation\_date z pliku station.csv. |
| Weather\_csv | Zadanie wczytuje kolumnę date z pliku weather.csv. |
| Trip\_start\_csv | Zadanie wczytuje kolumnę start\_date z pliku trip.csv. |
| Trip\_end\_csv | Zadanie wczytuje kolumnę end\_date z pliku trip.csv. |
| Cast Installation\_Date to DATE | Zadanie rzutuje kolumnę installation\_date na DATE. |
| Cast Weather\_Date to DATE | Zadanie rzutuje kolumnę date na DATE. |
| Cast Start\_Date to DATETIME | Zadanie rzutuje kolumnę start\_date na DATETIME. |
| Cast End\_Date to DATETIME | Zadanie rzutuje kolumnę end\_date na DATETIME. |
| Create DIM\_TIME columns | Zadanie tworzy z atrybutu installation\_date odpowiednie dla wymiaru DIM\_TIME kolumny. |
| Create DIM\_TIME columns 2 | Zadanie tworzy z atrybutu date odpowiednie dla wymiaru DIM\_TIME kolumny. |
| Create DIM\_TIME columns 3 | Zadanie tworzy z atrybutu start\_date odpowiednie dla wymiaru DIM\_TIME kolumny. |
| Create DIM\_TIME columns 4 | Zadanie tworzy z atrybutu end\_date odpowiednie dla wymiaru DIM\_TIME kolumny. |
| Combine all rows | Zadanie łączy wszystkie rekordy otrzymane z poprzednich czterech zadań. |
| Remove duplicates | Zadanie usuwa duplikaty rekordów powstałych w wyniku wykonania poprzedniego zadania. |
| ENUM\_MONTH | Zadanie wczytuje wszelkie dane z pomocniej tabeli ENUM\_MONTH. |
| Sort enum\_month | Zadanie wprowadza sortowanie kolumn z ENUM\_MONTH na potrzeby późniejszego złączenia kolumn. |
| Append Month\_Name to all rows | Zadanie dołącza do rekordów z zadania „Remove duplicates” kolumnę Month\_Name z ENUM\_MONTH. |
| Sort time rows | Zadanie wprowadza sortowanie kolumn z poprzedniego zadania na potrzeby późniejszego złączenia kolumn. |
| ENUM\_WEEKDAY | Zadanie wczytuje wszelkie dane z pomocniczej tabeli ENUM\_WEEKDAY. |
| Sort enum\_weekday | Zadanie wprowadza sortowanie kolumn z poprzedniego zadania na potrzeby późniejszego złączenia kolumn. |
| Append Weekday\_Name to all rows | Zadanie dołącza do rekordów z zadania „Sort time rows” kolumnę Weekday\_Name z ENUM\_WEEKDAY. |
| DIM\_TIME | Zadanie zapisuje przekształcone dane do tabeli DIM\_TIME. |

|  |  |
| --- | --- |
| INSERT FACT\_TRIP | |
| **Nazwa zadania** | **Opis zadania** |
| Trip\_csv | Zadanie wczytuje wszystkie dane z pliku trip.csv |
| Cast Start\_Date and End\_Date to DATETIME | Zadanie rzutuje kolumny start\_date oraz end\_date na DATE. |
| Format dates and cast Start\_Station\_Id to INT | Zadanie formatuje kolumny start\_date oraz end\_date do postaci yyyyMMddHHmm. Dodatkowo tworzona jest kolumna Weather\_Time, która powstaje z przekształcenia start\_date na format yyyyMMddXXXX gdzie X to po prostu znak ‘X’. Oprócz tego kolumna start\_station\_id jest rzutowana na liczbę całkowitą. |
| Sort trip rows | Zadanie wprowadza sortowanie kolumn z poprzedniego zadania na potrzeby późniejszego złączenia kolumn. |
| DIM\_CUSTOMER | Zadanie wczytuje wszelkie dane z tabeli DIM\_CUSTOMER. |
| Sort customer rows | Zadanie wprowadza sortowanie kolumn z poprzedniego zadania na potrzeby późniejszego złączenia kolumn. |
| Append Customer\_ID to all trip rows | Zadanie dołącza do rekordów z zadania „Sort trip rows” kolumnę Customer\_ID z DIM\_CUSTOMER. |
| Sort trip rows 2 | Zadanie wprowadza sortowanie kolumn z poprzedniego zadania na potrzeby późniejszego złączenia kolumn. |
| DIM\_STATION | Zadanie wczytuje wszelkie dane z tabeli DIM\_STATION. |
| Sort station rows | Zadanie wprowadza sortowanie kolumn z poprzedniego zadania na potrzeby późniejszego złączenia kolumn. |
| Append City to all trip rows | Zadanie dołącza do rekordów z zadania „Sort trip rows 2” kolumnę City z DIM\_STATION. |
| Sort trip rows 3 | Zadanie wprowadza sortowanie kolumn z poprzedniego zadania na potrzeby późniejszego złączenia kolumn. |
| DIM\_WEATHER | Zadanie wczytuje wszelkie dane z tabeli DIM\_WEATHER. |
| Sort weather rows | Zadanie wprowadza sortowanie kolumn z poprzedniego zadania na potrzeby późniejszego złączenia kolumn. |
| Append Weather\_ID to all trip rows | Zadanie dołącza do rekordów z zadania „Sort trip rows 3” kolumnę Weather\_ID z DIM\_WEATHER. |
| FACT\_TRIP | Zadanie zapisuje przekształcone dane do tabeli FACT\_TRIP. |

## Mapa logiczna procesu ETL

Dane źródłowe pochodzą z plików CSV bez dodatkowych metadanych, w związku z czym zapisane są w postaci ciągu znaków (plain text). Przy importowaniu w ramach SSIS posiadają formę DT\_STR(50).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Cel | | | | Źródło | | | Przekształcenie |
| Tabela | **Kolumna** | **Typ danych** | **Plik** | | **Kolumna** |  | |
| FACT\_TRIP | TRIP\_ID | INTEGER | trip.csv | | id | Zwykłe przepisanie | |
| FACT\_TRIP | Start\_Time | VARCHAR(12) | trip.csv | | start\_date | Zmiana formatu z MM/dd/yyyy na yyyyMMddHHmm | |
| FACT\_TRIP | End\_Time | VARCHAR(12) | trip.csv | | end\_date | Zmiana formatu z MM/dd/yyyy na yyyyMMddHHmm | |
| FACT\_TRIP | Start\_Station | INTEGER | trip.csv | | start\_station\_id | Zwykłe przepisanie | |
| FACT\_TRIP | End\_Station | INTEGER | trip.csv | | end\_station\_id | Zwykłe przepisanie | |
| FACT\_TRIP | Start\_Station\_Name | VARCHAR(45) | trip.csv | | start\_station\_name | Zwykłe przepisanie | |
| FACT\_TRIP | End\_Station\_Name | VARCHAR(45) | trip.csv | | end\_station\_name | Zwykłe przepisanie | |
| FACT\_TRIP | Trip\_Customer | INTEGER | - | | - | Nowo utworzony identyfikator dla tabeli DIM\_CUSTOMER. | |
| FACT\_TRIP | Trip\_Bike | INTEGER | trip.csv | | bike\_id | Zwykłe przepisanie | |
| FACT\_TRIP | Trip\_Weather | INTEGER | - | | - | Nowo utworzony identyfikator dla tabeli DIM\_WEATHER. | |
| FACT\_TRIP | Duration | INTEGER | trip.csv | | duration | Zwykłe przepisanie | |
| DIM\_BIKE | BIKE\_ID | INTEGER | trip.csv | | bike\_id | Przepisanie po usunięciu duplikatów. | |
| DIM\_STATION | STATION\_ID | INTEGER | station.csv | | id | Zwykłe przepisanie | |
| DIM\_STATION | Name | VARCHAR(45) | station.csv | | name | Zwykłe przepisanie | |
| DIM\_STATION | Lat | FLOAT(24) | station.csv | | lat | Zwykłe przepisanie | |
| DIM\_STATION | Long | FLOAT(24) | station.csv | | long | Zwykłe przepisanie | |
| DIM\_STATION | Dock\_Count | INTEGER | station.csv | | dock\_count | Zwykłe przepisanie | |
| DIM\_STATION | City | VARCHAR(13) | station.csv | | city | Zwykłe przepisanie | |
| DIM\_STATION | Installation\_Date | VARCHAR(12) | station.csv | | installation\_date | Zmiana formatu z MM/dd/yyyy na yyyyMMddXXXX gdzie X to po prostu znak ‘X’ | |
| DIM\_CUSTOMER | CUSTOMER\_ID | INTEGER | - | | - | Nowo utworzony identyfikator dla każdej unikalnej pary (Cust\_Zip\_Code, Subscription\_Type) | |
| DIM\_CUSTOMER | Cust\_Zip\_Code | VARCHAR(11) | trip.csv | | zip\_code | Zamiana NULL’i na wartość „Unknown” oraz na pewno nieprawidłowych kodów pocztowych na „Incorrect” | |
| DIM\_CUSTOMER | Subscription\_Type | VARCHAR(10) | trip.csv | | subscription\_type | Zwykłe przepisanie | |
| DIM\_TIME | PK\_TIME | VARCHAR(12) | station.csv + trip.csv + weather.csv | | installation\_date + start\_date + end\_date + date | Przekształcenie dat na format yyyyMMddHHmm lub yyyyMMddXXXX gdzie X to po prostu znak ‘X’ dla dat bez podanej godziny. | |
| DIM\_TIME | Year | INTEGER | station.csv + trip.csv + weather.csv | | installation\_date + start\_date + end\_date + date | Wyciągnięcie roku w formie liczbowej z dat | |
| DIM\_TIME | Month | INTEGER | station.csv + trip.csv + weather.csv | | installation\_date + start\_date + end\_date + date | Wyciągnięcie numeru miesiąca w formie liczbowej z dat | |
| DIM\_TIME | Month\_Name | VARCHAR(9) | station.csv + trip.csv + weather.csv | | installation\_date + start\_date + end\_date + date | Wyciągnięcie nazwy miesiąca na podsawie numeru miesiąca i tabeli pomocniczej ENUM\_MONTH | |
| DIM\_TIME | Day | INTEGER | station.csv + trip.csv + weather.csv | | installation\_date + start\_date + end\_date + date | Wyciągnięcie dnia miesiąca w formie liczbowej z dat | |
| DIM\_TIME | Week\_Day | VARCHAR(9) | station.csv + trip.csv + weather.csv | | installation\_date + start\_date + end\_date + date | Wyciągnięcie nazwy dnia tygodnia na podstawie daty i tabeli pomocniczej ENUM\_WEEKDAY | |
| DIM\_TIME | Hour | INTEGER | trip.csv | | start\_date + end\_date | Wyciągnięcie godziny w formie liczbowej z czasu lub NULL dla samych dat | |
| DIM\_TIME | Minute | INTEGER | trip.csv | | start\_date + end\_date | Wyciągnięcie minuty w formie liczbowej z czasu lub NULL dla samych dat | |
| DIM\_WEATHER | WEATHER\_ID | INTEGER | - | | - | Nowo utworzony identyfikator pogody zarejstrowanej w danym dniu i danym mieście | |
| DIM\_WEATHER | Measure\_Date | VARCHAR(12) | weather.csv | | date | Zmiana formatu z MM/dd/yyyy na yyyyMMddXXXX gdzie X to po prostu znak ‘X’ | |
| DIM\_WEATHER | Measure\_City | VARCHAR(13) | weather.csv | | zip\_code | Zmiana kodu pocztowego na odpowiadające miasto | |
| DIM\_WEATHER | Max\_Temperature | INTEGER | weather.csv | | max\_temperature\_f | Zwykłe przepisanie | |
| DIM\_WEATHER | Mean\_Temperature | INTEGER | weather.csv | | mean\_temperature\_f | Zwykłe przepisanie | |
| DIM\_WEATHER | Min\_Temperature | INTEGER | weather.csv | | min\_temperature\_f | Zwykłe przepisanie | |
| DIM\_WEATHER | Max\_Humidity | INTEGER | weather.csv | | max\_humidity | Zwykłe przepisanie | |
| DIM\_WEATHER | Mean\_Humidity | INTEGER | weather.csv | | mean\_humidity | Zwykłe przepisanie | |
| DIM\_WEATHER | Min\_Humidity | INTEGER | weather.csv | | min\_humidity | Zwykłe przepisanie | |
| DIM\_WEATHER | Max\_Pressure | FLOAT(24) | weather.csv | | max\_sea\_level\_pressure\_inches | Zwykłe przepisanie | |
| DIM\_WEATHER | Mean\_Pressure | FLOAT(24) | weather.csv | | mean\_sea\_level\_pressure\_inches | Zwykłe przepisanie | |
| DIM\_WEATHER | Min\_Pressure | FLOAT(24) | weather.csv | | min\_sea\_level\_pressure\_inches | Zwykłe przepisanie | |
| DIM\_WEATHER | Max\_Visibility | INTEGER | weather.csv | | max\_visibility\_miles | Zwykłe przepisanie | |
| DIM\_WEATHER | Mean\_Visibility | INTEGER | weather.csv | | mean\_visibility\_miles | Zwykłe przepisanie | |
| DIM\_WEATHER | Min\_Visibility | INTEGER | weather.csv | | min\_visibility\_miles | Zwykłe przepisanie | |
| DIM\_WEATHER | Precipitation\_Inches | VARCHAR(4) | weather.csv | | precipitation\_inches | Zmiana wartości „T” na „Tiny” | |
| DIM\_WEATHER | Cloud\_Cover | INTEGER | weather.csv | | cloud\_cover | Zwykłe przepisanie | |
| DIM\_WEATHER | Events | VARCHAR(17) | weather.csv | | events | Zmiana NULL’i na wartość „None” oraz wartości „rain” na „Rain” | |

## Wnioski:

Zautomatyzowany proces ETL pozwala nam rekonstruować hurtownię danych na życzenie co jest szczególnie użyteczne w przypadku zmiany w strukturze/źródle danych. Odpowiednie rozplanowanie i podzielenie tegoż procesu na podzadania, wraz z dołączoną dokumentacją opisującą poszczególne zadania, umożliwia szybką modyfikację wykonywanych operacji (pozostawiając cały proces nadal w formie zautomatyzowanej). Dodatkowo możemy posłużyć się mapą logiczną procesu ETL by szybko zorientować się w jakiś sposób dane źródłowe są przekształcane na wynikowe tabele faktów czy wymiarów.

# Etap 3

## Dokumentacja kostki

Wymiar DIM CUSTOMER – reprezentuje klienta, który korzysta z usług Bay Area Bike Share. Ze względu na ochronę prywatności rejestrowany był jedynie status subskrypcji (czy ją posiadał czy była to jednorazowa wycieczka) oraz dobrowolnie podany kod pocztowy.

|  |  |
| --- | --- |
| Atrybut | Opis |
| Cust Zip Code | Kod pocztowy podany przez klienta (mógł być podany niewłaściwy ale nie ma sposobu weryfikacji tegoż przypadku). Tam gdzie kod pocztowy był ewidentnie niewłaściwy nastąpiła zamiana na „INCORRECT”. |
| CUSTOMER ID | Liczba całkowita identyfikująca poszczególnego klienta. |
| Subscription Type | De facto „rodzaj” klienta: „Subscriber” jeśli posiadał wykupioną subksrypcję oraz „Customer” jeśli był to jednorazowy zakup. |

Wymiar DIM BIKE – reprezentuje rower, który dostępny był do wypożyczenia przez klientów Bay Area Bike Share. W udostępnianych danych firma zamieściła tylko informację o numerze roweru.

|  |  |
| --- | --- |
| Atrybut | Opis |
| BIKE ID | Liczba całkowita identyfikująca rower. |

Wymiar DIM WEATHER – reprezentuje ogólnie rozumiane warunki pogodowe i atmosferyczne dla danego dnia oraz danego miasta w regionie zatoki San Francisco.

|  |  |
| --- | --- |
| Atrybut | Opis |
| Cloud Cover | Liczba całkowita w przedziale 0-8 obustronnie domkniętym, reprezentuje skalę zachmurzenia gdzie 0 – czyste niebo, 8 – zupełny brak słońca. |
| Events | Typ wyliczeniowy informujący o występujących zjawiskach atmosferycznych: None – brak zjawisk, Rain – sam deszcz, Fog – sama mgła, Fog-Rain – deszcz z mgłą, Rain-Thunderstorm – burza z deszczem. |
| Max Humidity | Liczba całkowita reprezentująca maksymalną procentową wilgotność powietrza. |
| Max Pressure | Liczba zmiennoprzecinkowa reprezentująca maksymalne ciśnienie n.p.m. w calach rtęci. |
| Max Temperature | Liczba całkowita reprezentująca maksymalną temperaturę w stopniach Fahrenheit’a. |
| Max Visibility | Liczba całkowita reprezentująca maksymalną widoczność w milach. |
| Mean Humidity | Liczba całkowita reprezentująca średnią procentową wilgotność powietrza. |
| Mean Pressure | Liczba zmiennoprzecinkowa reprezentująca średnie ciśnienie n.p.m. w calach rtęci. |
| Mean Temperature | Liczba całkowita reprezentująca średnią temperaturę w stopniach Fahrenheit’a. |
| Mean Visibility | Liczba całkowita reprezentująca średnią widoczność w milach. |
| Measure City | Typ wyliczeniowy reprezentujący miasto, w którym dokonano pomiaru: „San Jose”, „San Francisco”, „Redwood City”, „Palo Alto” oraz „Mountain View”. |
| Measure Date | Ciąg znaków reprezentujący datę pomiaru pogody. |
| Min Humidity | Liczba całkowita reprezentująca minimalną procentową wilgotność powietrza. |
| Min Pressure | Liczba zmiennoprzecinkowa reprezentująca minimalne ciśnienie n.p.m. w calach rtęci. |
| Min Temperature | Liczba całkowita reprezentująca minimalną temperaturę w stopniach Fahrenheit’a. |
| Min Visibility | Liczba całkowita reprezentująca minimalną widoczność w milach. |
| Precipitation Inches | Ciąg znaków reprezentujący liczbę opadów w calach, gdzie 0 – brak opadów, 0.01 lub więcej to odpowiednia liczba opadów a „Tiny” gdy opadów było mniej niż 0.01 cala. |
| WEATHER ID | Liczba całkowita identyfikująca stan pogody z danego dnia dla danego regionu. |

Wymiar DIM TIME – reprezentuje czas zarejestrowany w ramach warunków pogodowych czy też startu/końca wycieczek.

|  |  |
| --- | --- |
| Atrybut | Opis |
| Day | Liczba całkowita reprezentująca dzień. |
| Hour | Liczba całkowita reprezentująca godzinę. |
| Minute | Liczba całkowita reprezentująca minutę. |
| Month | Liczba całkowita reprezentująca miesiąc. |
| Month Name | Ciąg znaków reprezentujący nazwę miesiąca. |
| PK TIME | Ciąg znaków identyfikujący dany rekord czasu. |
| Week Day | Ciąg znaków reprezentujący nazwę dnia tygodnia. |
| Year | Liczba całkowita reprezentująca rok. |

Wymiar DIM STATION – reprezentuje stację rowerową, która udostępnia klientom możliwość wypożyczenia rowerów postawionych w „dokach”.

|  |  |
| --- | --- |
| Atrybut | Opis |
| City | Typ wyliczeniowy reprezentujący miasto, w którym usytowana jest stacja rowerowa: „San Jose”, „San Francisco”, „Redwood City”, „Palo Alto” oraz „Mountain View”. |
| Dock Count | Liczba całkowita reprezentująca pojemność rowerów w doku stacji. |
| Installation Date | Ciąg znaków reprezentujący datę instalacji stacji rowerowej. |
| Lat | Liczba zmiennoprzecinkowa reprezentująca szerokość geograficzną, na której położona jest stacja rowerowa. |
| Long | Liczba zmiennoprzecinkowa reprezentująca długość geograficzną, na której położona jest stacja rowerowa. |
| Name | Ciąg znaków reprezentujący nazwę stacji rowerowej. |
| STATION ID | Liczba całkowita identyfikująca stację rowerową. |

|  |
| --- |
| **Miary:** |
| Duration – funkcja agregacyjna: suma, suma długości wycieczek w sekundach |
| FACT TRIP Count – funkcja agregacyjna: policz, liczba wycieczek |
| Trip Customer Distinct Count – funkcja agregacyjna: policz unikalne, liczba unikalnych klientów wycieczek |
| Trip Bike Distinct Count – funkcja agregacyjna: policz unikalne, liczba unikalnych rowerów wycieczek |

## Zaplanowane zestawienia (1.5.2)

1. Sumaryczna długość wycieczek ze względu na dzień tygodnia dla każdej stacji.

Table

Description automatically generated

Timeline

Description automatically generated with medium confidence

1. Liczba wycieczek ze względu na zachmurzenie według miast.

Table

Description automatically generated

Chart

Description automatically generated

1. Procentowy udział klientów niezarejestrowanych oraz subskrybentów ze względu na miasto.

Graphical user interface, text, application

Description automatically generated

Chart, bar chart

Description automatically generated

1. Liczba wycieczek podczas mgły ze względu na godzinę i miasto.

Graphical user interface, application, table, Excel

Description automatically generated

Chart, histogram

Description automatically generated

1. Liczba wycieczek dla każdej stacji ze względu na opady (=0 – brak, T-nieznaczne, < 0.20 – średnie, > 0.20 – znaczące).

Table

Description automatically generated

Chart, bar chart

Description automatically generated

1. Średnia długość wycieczki dla każdej stacji z San Jose.

Table

Description automatically generated

Chart, bar chart

Description automatically generated

1. Zestawienie największej liczby wycieczek dla każdej stacji ze względu na miesiąc.

Table

Description automatically generated with low confidence

Timeline

Description automatically generated

1. Liczba wycieczek dla klientów zamieszkujących pod każdym kodem pocztowym ze względu na miesiąc.

A picture containing text, large

Description automatically generated

Chart

Description automatically generated with medium confidence

1. Sumaryczna długość wycieczek dla każdego roweru ze względu na godzinę.

Table

Description automatically generated

Graphical user interface, application, table, Excel

Description automatically generated

1. Liczba unikalnych klientów rozpoczynających lub kończących wycieczkę dla każdej stacji ze względu na miesiąc.

Table

Description automatically generated with low confidence

A picture containing text, writing implement, stationary, pencil

Description automatically generated

## Analiza zestawień

1. Zdecydowanie najbardziej wyróżniającym się dniem tygodnia jest środa dla stacji „South Van Ness at Market”, gdzie suma długości wycieczek osiąga wartość maksymalną i przewyższająca wszystkie pozostałe dni wszystkich stacji. Dla pozostałych wyników zasadniczo występują zbliżone proporcje.
2. Ogólnie najwięcej wycieczek odbywa się w San Francisco. Gdy zachmurzenie jest ogromne (7 lub 8) to niezależnie od miasta odbywa się najmniej wycieczek. W skali 0-6 zachmurzenia w zasadzie występuje ogólnie dużo wycieczek (najwięcej w przypadku San Francisco występuje dla stanów 4 lub 5).
3. Zdecydowanie największy procentowy udział subskrybentów występuje w mieście Redwood City. Najmniejszy procent zanotowany został w San Francisco. Może wynikać to z bezwzględnej liczby klientów poszczególnych miast.
4. Najwięcej wycieczek rowerowych podczas mgły występowało w porach porannych (7-9) oraz popołudniowych/wieczornych (16-18). Prawdopodobnie wynika to z godzin szczytu i naturalnego okresu pracy, a więc nie zależy od mgły.
5. Wyniki generalnie podobne do tych z zestawieniego pierwszego. Żeby analizować opady trzeba by było ograniczyć się do kilku miast lub konkretnego miasta (gdyż za bardzo dominuje w danych stacja „San Francisco Caltrain (Townsend at 4th)”).
6. Zdecydowanie najdłuższe średnio wycieczki zaczynały się na stacji „San Jose Civic Center”. Wyprzedza praktycznie dwukrotnie kolejną z kolei stację. Klienci z reszty stacji prowadzą wycieczki podobnych długości.
7. Zasadniczo dla każdej stacji występuje dość podobna proporcja między kolejnymi miesiącami. Danych jest jednak za dużo i należałoby przeprowadzić analizę pojedynczych stacji.
8. Około 75% kodów pocztowych klientów znajduje się w przedziale 94100-94105 co odpowiada bezpośrednim miastom z rejonu zatoki San Francisco.
9. Rower 535 jest kilkunastokrotnie dłużej wykorzystywany od godziny 21 niż jakikolwiek inny rower. Wynika to prawdopodbnie z tego, że jeden klient wynajął rower na bodajże cały rok (prawdopodobnie ze stacji „South Van Ness at Market” w środę o 21 przez co w tych regionach danych odnotowujemy tak wysoki skok).
10. Najwięcej unikalnych klientów odnotowano na stacjach „Harry Bridges Plaza (Ferry Building)” oraz „Embarcadero at Sansome”. Obie te stacje niezależnie od miesiąca wyprzedzają wszystkie pozostałe. Ogólnie na przestrzeni wszystkich stacji liczba unikalnych klientów jest dość równomiernie rozłożona pomiędzy miesiącami.

## Analiza w głąb

Chart, bar chart

Description automatically generated

Rozpoczynam analizę z zestawienia trzeciego wykonanego w ramach punktu 1.5.2. W tabel i na wykresie przedstawiony został procentowy udział subskrybentów oraz „przypadkowych” klientów ze względu na unikalną liczbę klientów dla każdego miasta. Jak widać na wykresie, ewidentnie największy udział subskrybentów odnotowano w mieście Redwood City, natomiast najmniejszy w samym San Francisco. Jako hipotezę przyjmuję podobną liczbę subskrybentów we wszystkich miastach natomiast dużo większą liczbę przypadkowych klientów San Francisco.

Chart, bar chart

Description automatically generated

Na powyższym wykresie i tabeli przedstawione zostało zestawienie liczby unikalnych klientów korzystających z usług stacji rowerowych dla każdego miasta. Jak widać dla San Francisco liczba klientów jest kilkudziesięciokrotnie wyższa niż dla Redwood City. W związku z tym liczba subskrybentów jest ostatecznie większa w San Francisco, natomiast stanowi mniejszy procent ogółu klientów ze względu na wspomnianą wcześniej dysproporcję.

Chart, bar chart

Description automatically generated

Na powyższym wykresie i tabeli przedstawione zostało procentowe zestawienie liczby wycieczek dla każdego miasta, z podziałem na subskrybentów i przypadkowych klientów, uwzględniając jedynie klientów z kodów pocztowych między 94100-94105 (są to kody pocztowe odpowiedające tym miastom). Jak widać w zasadzie większość klientów pochodzących z regionu zatoki San Francisco posiada subskrybcję rowerową – najwyższy odsetek dla miasta Mountain View, 98.26%, natomiast najniższy dla Palo Alto jednak jest to i tak 87,78%. Świadczy to o tym, że większość przypadkowych klientów to osoby spoza regionu zatoki San Francisco (być może zwiedzający).

Graphical user interface, application

Description automatically generated

Na powyższym wykresie i tabeli przedstawione zostało zestawienie dziesięciu stacji, z których liczbowo najwięcej odbyło się wycieczek przypadkowych klientów. Zdecydowanie dominują w tym zestawieniu dwie stacje: „Harry Bridges Plaza (Ferry Building)” oraz „Embarcadero at Sansome”. Ogólnie śledząc położenie i znaczenie geograficzne tych miejsc (co po części sugerują same nazwy), można dojść do wniosku iż faktycznie bardzo dużo wycieczek startuje w miejscach interesujących dla turystów.

Chart, bar chart

Description automatically generated

Na powyższym wykresie i tabeli przedstawione zostało zestawienie liczby wycieczek przypadkowych klientów ze względu na miesiąc. W zależności od tego jak spojrzymy na uzyskane wyniki to możemy uznać, że żaden miesiąc nie jest ewidentnie najgorszy bądź najlepszy, lub że zdecydowanie najlepszy jest sierpień a najgorszy luty. Ogólnie więcej wycieczek odbywa się w okresie między marcem a październikiem a mniej w okresie zimowym.

Chart

Description automatically generated

Na powyższym wykresie i tabeli przedstawione zostało zestawienie liczby unikalnych klientów ze względu na rodzaj subskrypcji oraz wydarzenia pogodowe/atmosferyczne. Zdecydowałem się umieścić to zestawienie ze względu na interesujące wyniki co do subskrybentów. Oczywistym jest, że dla idealnych warunków atmosferycznych odnotowano największą liczbę unikalnych klientów (zarówno subskrybentów jak i przypadkowych klientów). Co ciekawe jednak, w przpadku gdy pojawiał się deszcz lub mgła to liczba przypadkowych klientów drastycznie spada (nawet kilkunastokrotnie), podczas gdy liczba subskrybentów korzystających z usług sieci rowerowych maleje o około 1/3. W przypadku dni burzowych dochodzi nawet do sytuacji gdy więcej wycieczek odbywało subskrybentów niż przypadkowych klientów. Pokazywać to może jak wielu przypadkowych klientów wybiera się na wycieczki podczas ładnej pogody (prawdopodobnie dla relaksu lub by zwiedzać), natomiast dla jak wielu subskrybentów rower jest normalnym środkiem transportu niezależnie od pogody.

## Wnioski:

Odpowiednio przygotowana i przeprocesowana kostka, wraz z dołączoną dokumentacją, pozwala analitykom w bardzo prosty sposób tworzyć wszelakie zestawienia na podstawie danych znajdujących się w hurtowni. Nie wymaga to nawet od niego żadnej znajomości struktury hurtowni czy umiejętności programistycznych (z myślą o SQL). Przygotowaną kostkę można wykorzystać w wielu programach umożliwiających analizę danych (nawet w zwykłym MS Excel’u w którym wygenerowane zostały tabele i wykresy zamieszczone w ramach etapu trzeciego).