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. 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

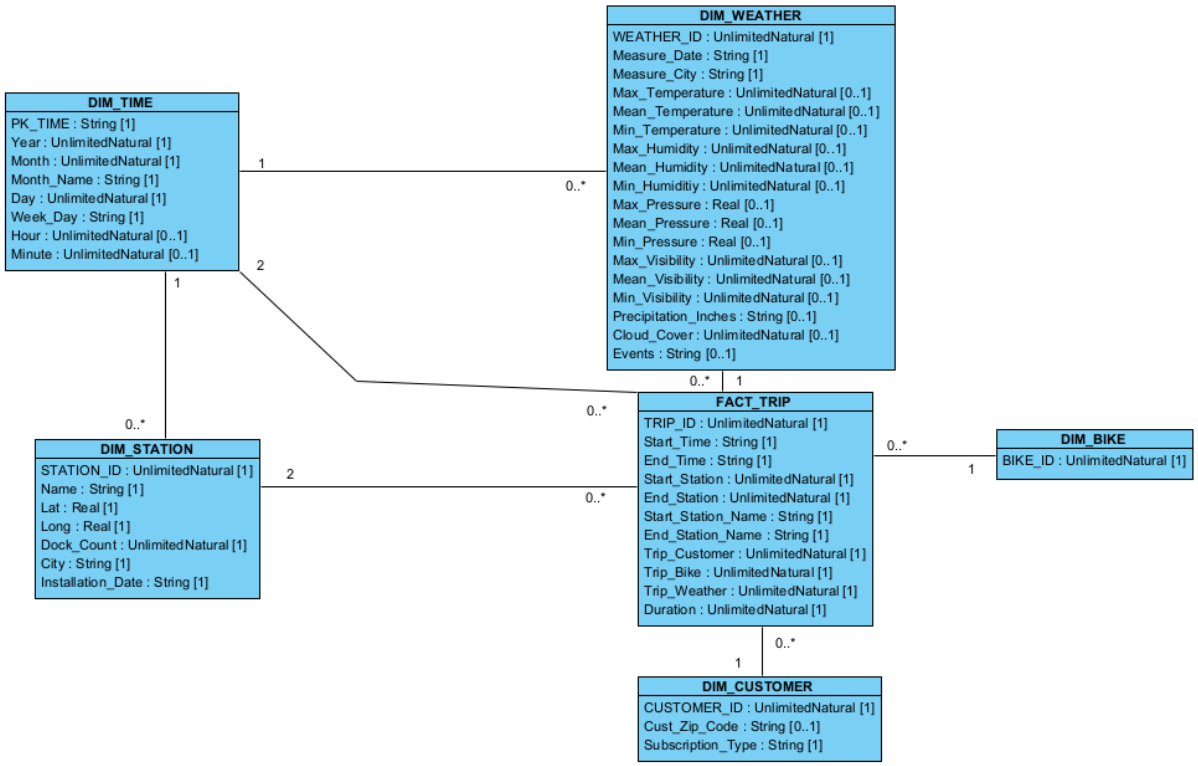
Hierarchie:

- DIM_TIME: Year -> Month -> Day -> Hour -> Minute
- DIM_STATION: City -> Name

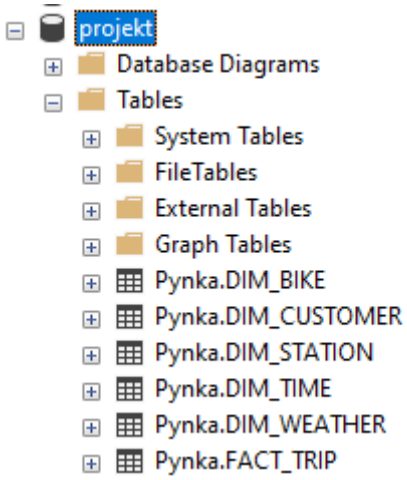
Miary:

- 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 zarejestrowanej 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_m iles	Zwykłe przepisanie
DIM_WEATHER	Precipitation_Inches	VARCHAR(4)	weather.csv	precipitation_in ches	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”.

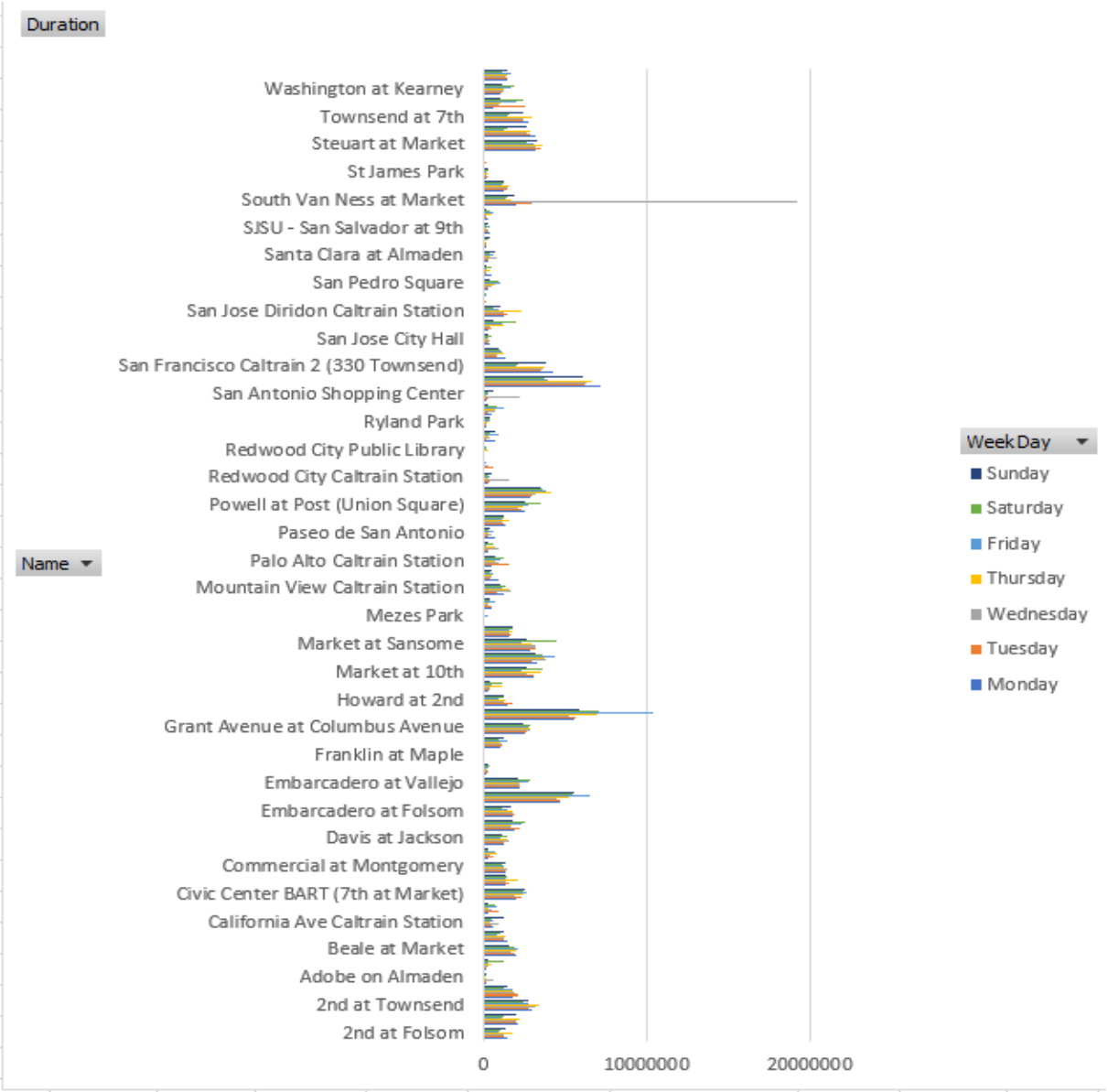
Atrybut	Opis
City	Typ wyliczeniowy reprezentujący miasto, w którym usytowana jest stacja rowerowa: „San Jose”, „San Francisco”, „Redwood City”, „Palo Alto” oraz „Mountain View”.
Dock Count	Liczba całkowita reprezentująca pojemność rowerów w doku stacji.
Installation Date	Ciąg znaków reprezentujący datę 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)

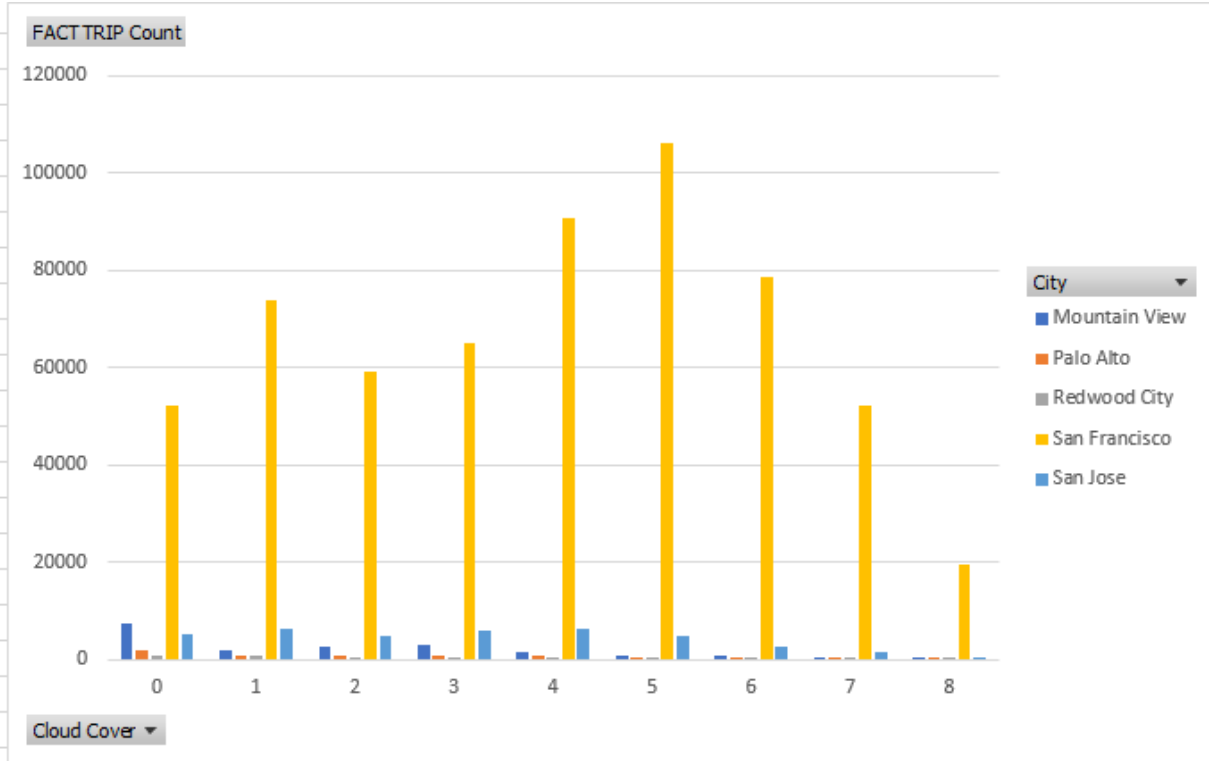
1. Sumaryczna długość wycieczek ze względu na dzień tygodnia dla każdej stacji.

Duration	Column Labels							
Row Labels	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	Grand Total
2nd at Folsom	1500522	1296294	1220777	1778222	885365	1070375	1361468	9113023
2nd at South Park	2115971	2025587	1969966	2216589	1131958	1224636	1963866	12648573
2nd at Townsend	2954445	2805614	3178866	3355315	2779451	2410457	2698817	20182965
5th at Howard	1733597	2148530	1863538	1784402	1786280	1220583	1433210	11970140
Adobe on Almaden	132103	152345	557365	124822	93302	202005	151811	1413753
Arena Green / SAP Center	158894	175380	224873	492769	305996	1243445	321184	2922541
Beale at Market	1992018	1837519	1682701	1963338	2100290	1932833	1583801	13092500
Broadway St at Battery St	1457766	1274687	1239930	1377168	784759	1059675	1227933	8421918
California Ave Caltrain Station	605144	453721	917072	416860	589512	457525	1214054	4653888
Castro Street and El Camino Real	231518	973358	526030	287835	761600	746518	273850	3800709
Civic Center BART (7th at Market)	2002277	2286521	1845732	2384391	2636317	2455250	2586507	16196995
Clay at Battery	1307110	1579282	1346299	2117814	1362541	1472825	1307680	10493551
Commercial at Montgomery	1390119	1324428	1425499	1295456	1250794	1172121	1326101	9184518
Cowper at University	258542	561045	414017	829042	721762	310276	268232	3362916
Davis at Jackson	1268467	1230099	1584671	1459751	1078622	1484154	1102996	9208760
Embarcadero at Bryant	1916833	2166029	1632153	1699422	2330009	2517320	1814960	14076726
Embarcadero at Folsom	1813367	1862439	1822540	1771088	1415508	1167301	1635549	11487792
Embarcadero at Sansome	4735203	4670159	4451597	5240628	6477247	5463581	5532634	36571049
Embarcadero at Vallejo	2203080	2264803	2164094	2218294	2759650	2878420	2152218	16640559
Evelyn Park and Ride	125696	288035	244094	168688	280177	418789	282098	1807577



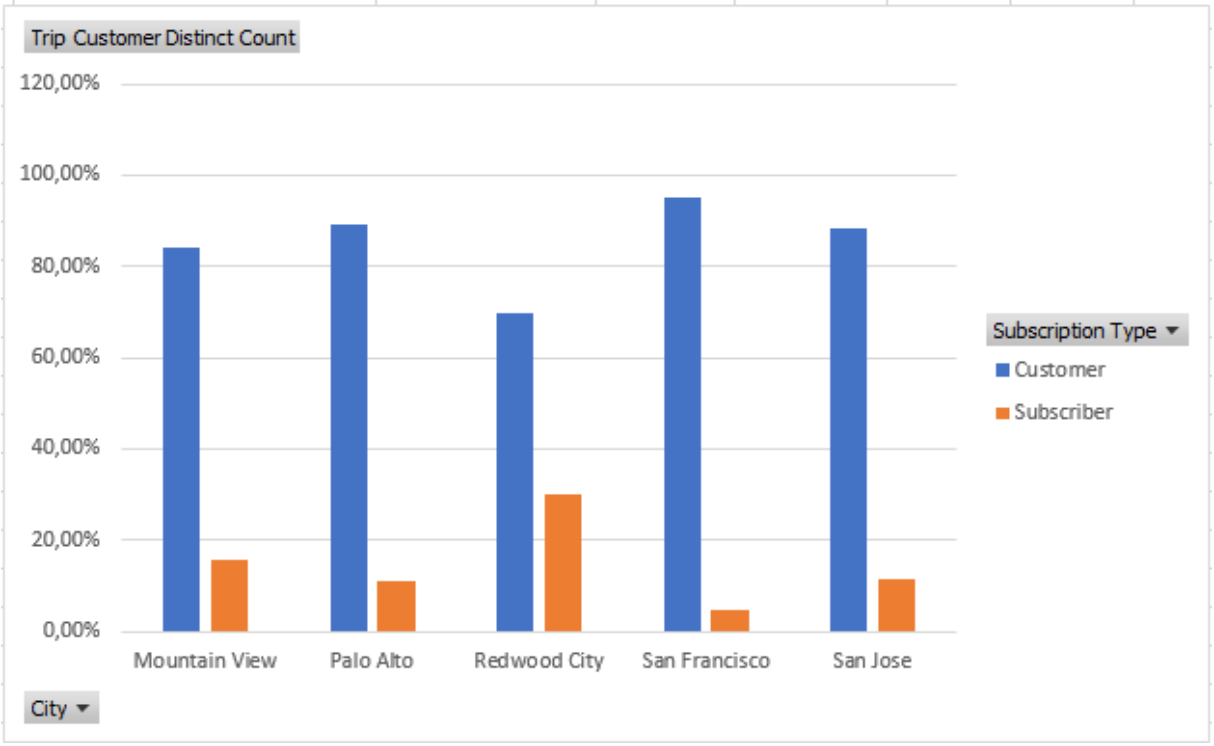
2. Liczba wycieczek ze względu na zachmurzenie według miast.

FACT TRIP Count	Column Labels					
Row Labels	Mountain View	Palo Alto	Redwood City	San Francisco	San Jose	Grand Total
0	7465	1915	702	52193	5289	67564
1	1804	852	689	73976	6274	83595
2	2526	966	568	59251	4775	68086
3	2870	817	387	65025	5912	75011
4	1535	742	350	90597	6355	99579
5	773	590	354	106175	4854	112746
6	641	428	191	78725	2550	82535
7	244	264	103	52040	1515	54166
8	224	63	65	19617	89	20058
Grand Total	**18082**	**6637**	**3409**	**597599**	**37613**	**663340**



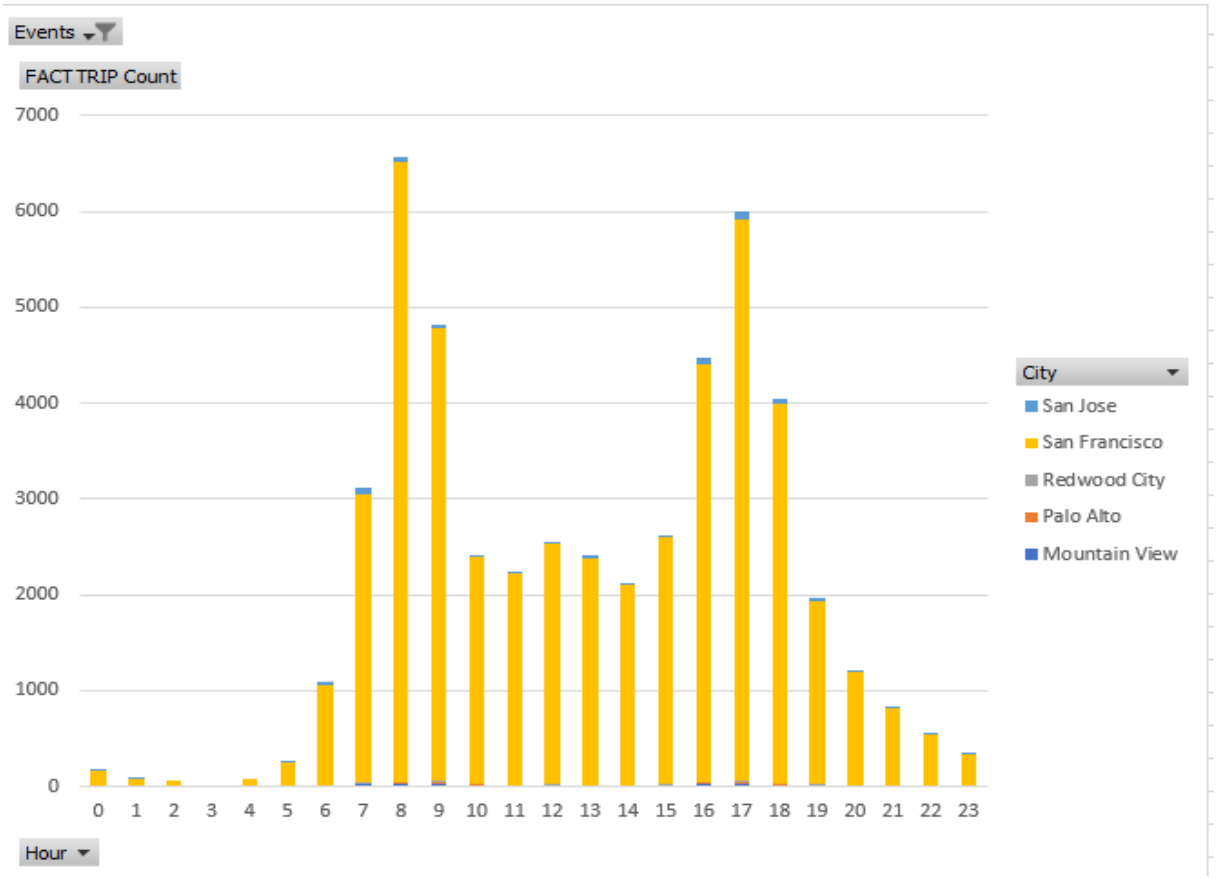
3. Procentowy udział klientów niezarejestrowanych oraz subskrybentów ze względu na miasto.

Trip Customer Distinct Count	Column Labels		
Row Labels	Customer	Subscriber	Grand Total
Mountain View	84,22%	15,78%	100,00%
Palo Alto	89,07%	10,93%	100,00%
Redwood City	69,90%	30,10%	100,00%
San Francisco	95,26%	4,74%	100,00%
San Jose	88,46%	11,54%	100,00%
Grand Total	**95,38%**	**4,62%**	**100,00%**



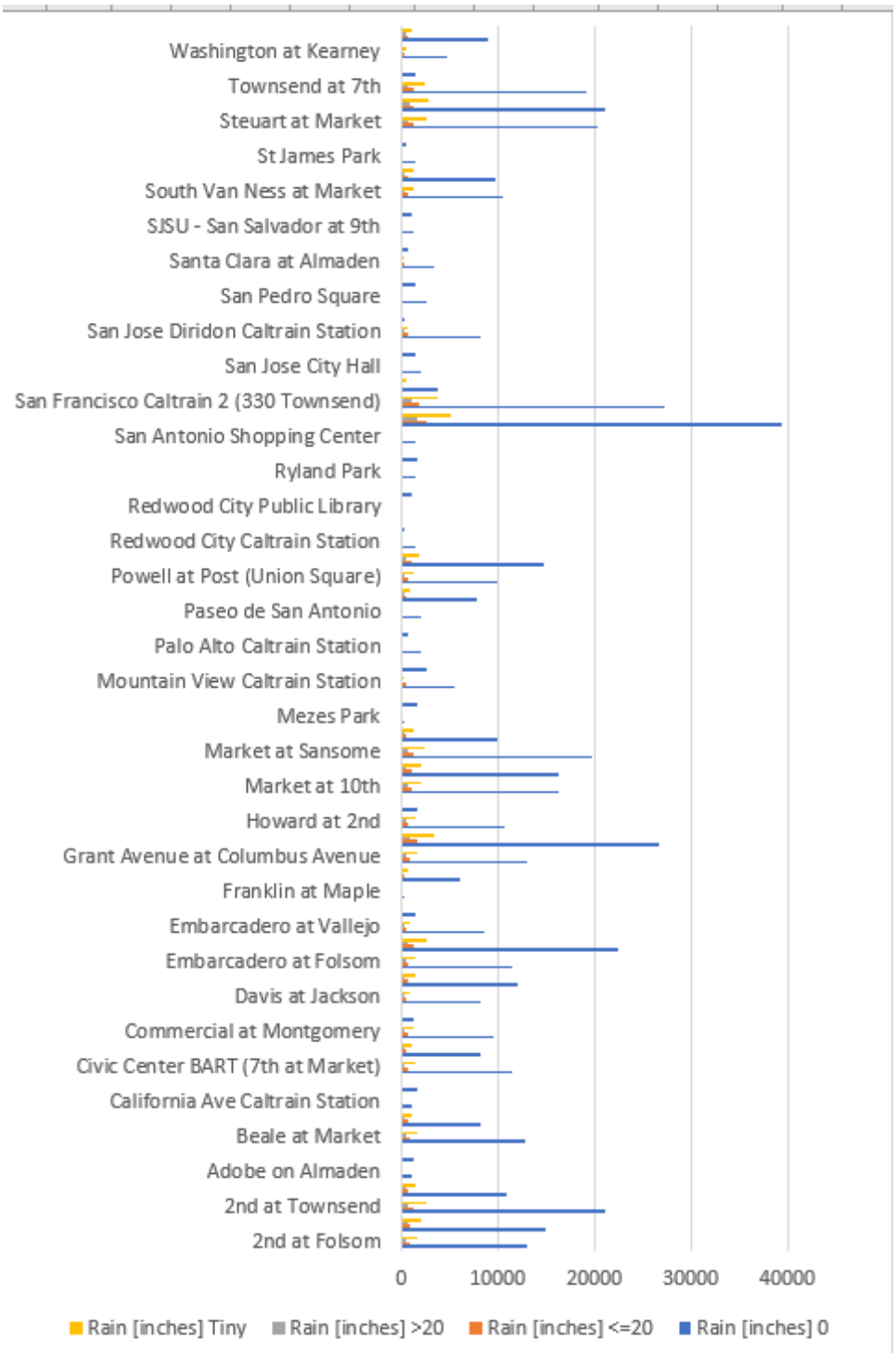
4. Liczba wycieczek podczas mgły ze względu na godzinę i miasto.

Events	(Multiple Items)					
FACT TRIP Count	**Column Labels**					
Row Labels	**Mountain View**	**Palo Alto**	**Redwood City**	**San Francisco**	**San Jose**	**Grand Total**
0		1	1	167	2	171
1			1	75	1	77
2		2	1	57		60
3	1			19		20
4				86		86
5	5			254	3	262
6	7			1060	30	1097
7	30	1	12	3007	67	3117
8	30	12	6	6463	50	6561
9	33	19	4	4734	33	4823
10	14	8	4	2364	15	2405
11	4	7	2	2210	26	2249
12	9	7	9	2502	28	2555
13	6	6	4	2369	32	2417
14	3	7	4	2086	27	2127
15	8	9	10	2577	16	2620
16	29	11	11	4350	66	4467
17	33	21	5	5852	84	5995
18	19	5	3	3976	44	4047
19	8	8	6	1919	21	1962
20	2	3	5	1180	9	1199
21	2	2	3	812	11	830
22	1		1	539	14	555
23	1	2		328	11	342
Grand Total	**245**	**131**	**92**	**48986**	**590**	**50044**



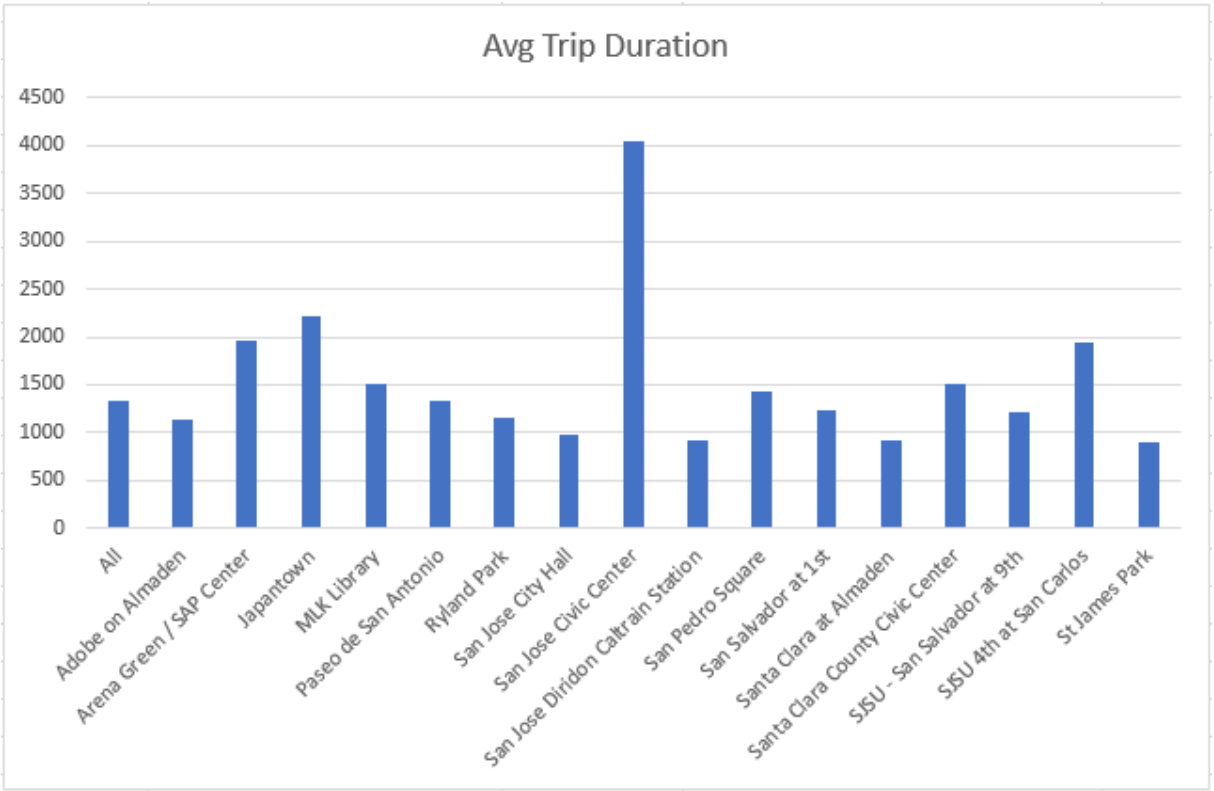
5. Liczba wycieczek dla każdej stacji ze względu na opady (=0 – brak, T-nieznaczne, < 0.20 – średnie, > 0.20 – znaczące).

FACT TRIP Count	Rain [inches]				
Station Name	0	<=20	>20	Tiny	Grand Total
2nd at Folsom	12927	858	534	1551	15870
2nd at South Park	14951	944	597	1952	18444
2nd at Townsend	21124	1195	689	2655	25663
5th at Howard	10874	732	423	1371	13400
Adobe on Almaden	1095	61	23	70	1249
Arena Green / SAP Center	1271	101	24	91	1487
Beale at Market	12721	801	469	1627	15618
Broadway St at Battery St	8175	669	330	1133	10307
California Ave Caltrain Station	978	12	7	5	1002
Castro Street and El Camino Real	1736	145	50	103	2034
Civic Center BART (7th at Market)	11410	677	392	1420	13899
Clay at Battery	8123	464	281	1015	9883
Commercial at Montgomery	9584	586	354	1282	11806
Cowper at University	1319	24	9	8	1360
Davis at Jackson	8190	472	287	954	9903
Embarcadero at Bryant	12095	740	389	1425	14649
Embarcadero at Folsom	11398	704	399	1468	13969
Embarcadero at Sansome	22372	1309	641	2612	26934
Embarcadero at Vallejo	8647	472	277	963	10359
Evelyn Park and Ride	1418	129	31	84	1662
Franklin at Maple	214	7	0	0	221
Golden Gate at Polk	5997	348	195	730	7270
Grant Avenue at Columbus Avenue	13082	910	501	1598	16091
Harry Bridges Plaza (Ferry Building)	26616	1550	845	3301	32312
Howard at 2nd	10641	661	422	1363	13087
Japantown	1667	107	26	92	1892
Market at 10th	16348	1151	588	2010	20097
Market at 4th	16229	1014	572	1992	19807
Market at Sansome	19669	1170	699	2443	23981
Mechanics Plaza (Market at Battery)	9981	540	289	1256	12066


Washington at Kearney
Townsend at 7th
Steuart at Market
St James Park
South Van Ness at Market
SJSU - San Salvador at 9th
Santa Clara at Almaden
San Pedro Square
San Jose Diridon Caltrain Station
San Jose City Hall
San Francisco Caltrain 2 (330 Townsend)
San Antonio Shopping Center
Ryland Park
Redwood City Public Library
Redwood City Caltrain Station
Powell at Post (Union Square)
Paseo de San Antonio
Palo Alto Caltrain Station
Mountain View Caltrain Station
Mezes Park
Market at Sansome
Market at 10th
Howard at 2nd
Grant Avenue at Columbus Avenue
Franklin at Maple
Embarcadero at Vallejo
Embarcadero at Folsom
Davis at Jackson
Commercial at Montgomery
Civic Center BART (7th at Market)
California Ave Caltrain Station
Beale at Market
Adobe on Almaden
2nd at Townsend
2nd at Folsom
0
10000
20000
30000
40000
Rain [inches] Tiny
Rain [inches] >20
Rain [inches] <=20
Rain [inches] 0

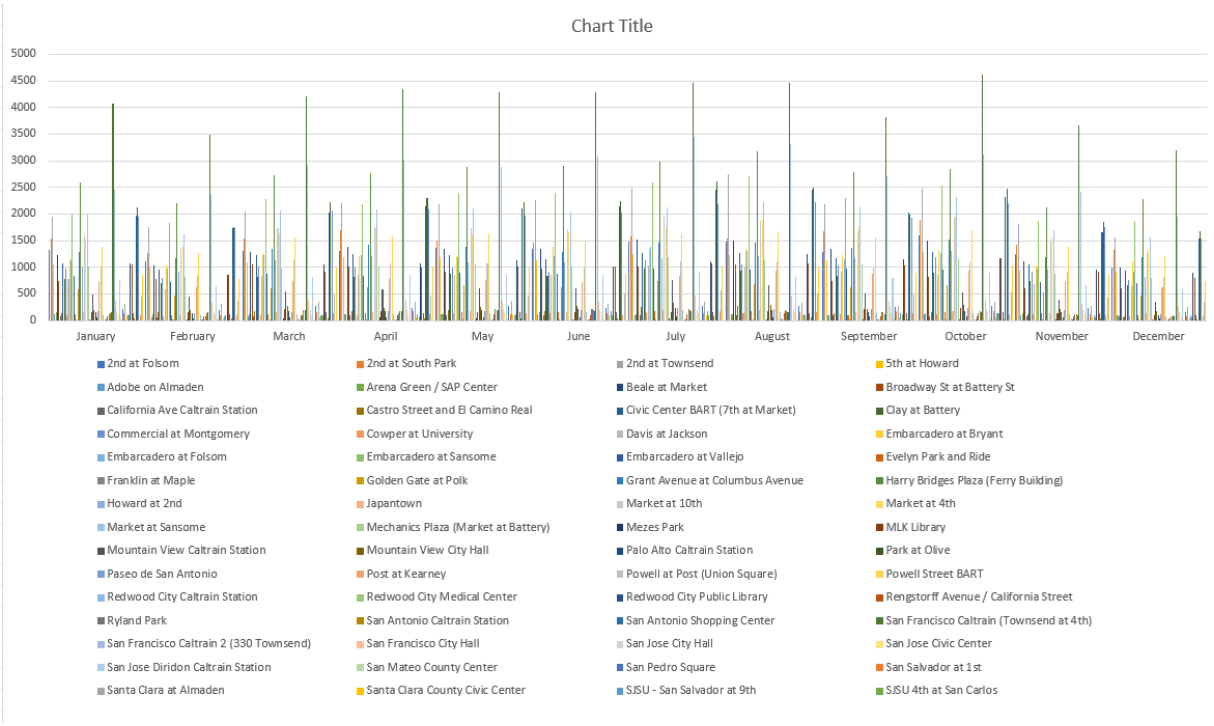
6. Średnia długość wycieczki dla każdej stacji z San Jose.

	Avg Trip Duration
All	1341,525988
Adobe on Almaden	1131,907926
Arena Green / SAP Center	1965,394082
Japantown	2217,245243
MLK Library	1508,321606
Paseo de San Antonio	1341,6507
Ryland Park	1153,886976
San Jose City Hall	980,9415229
San Jose Civic Center	4046,16031
San Jose Diridon Caltrain Station	917,3311988
San Pedro Square	1423,423489
San Salvador at 1st	1224,310304
Santa Clara at Almaden	920,7547613
Santa Clara County Civic Center	1501,613689
SJSU - San Salvador at 9th	1222,046543
SJSU 4th at San Carlos	1947,099138
St James Park	909,2752843



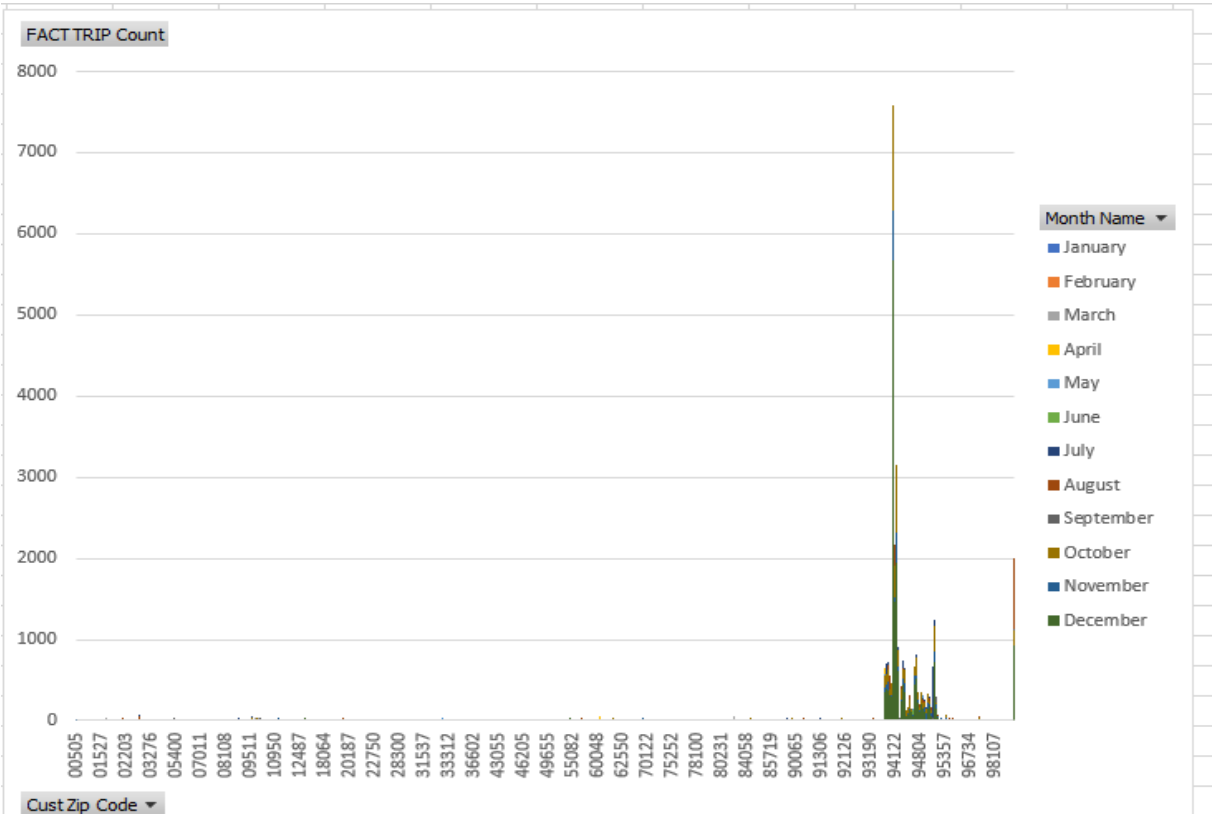
7. Zestawienie największej liczby wycieczek dla każdej stacji ze względu na miesiąc.

	January	February	March	April	May	June	July	August	September	October	November	December
2nd at Folsom	1335	1110	1298	1307	1372	1345	1483	1479	1294	1599	1251	997
2nd at South Park	1534	1269	1540	1702	1508	1469	1583	1542	1673	1886	1418	1320
2nd at Townsend	1943	1755	2048	2203	2181	2263	2485	2756	2186	2479	1809	1555
5th at Howard	1054	1001	1089	1194	1185	1136	1255	1227	1122	1294	931	912
Adobe on Almaden	121	67	88	113	124	109	99	122	103	118	93	92
Arena Green / SAP Center	151	125	125	120	125	152	116	127	104	135	130	77
Beale at Market	1235	1025	1296	1384	1341	1353	1516	1509	1348	1503	1106	1002
Broadway St at Battery St	736	777	1050	1013	911	972	1016	1057	743	827	601	604
California Ave Caltrain Station	76	56	82	95	90	89	105	98	97	85	79	51
Castro Street and El Camino Real	146	155	178	179	193	174	255	267	146	148	114	79
Civic Center BART (7th at Market)	1081	964	1233	1245	1223	1149	1275	1273	1180	1292	1056	928
Clay at Battery	786	704	821	826	880	837	967	941	837	892	735	657
Commercial at Montgomery	975	804	1008	1002	998	922	1128	1061	1038	1198	913	759
Cowper at University	89	74	107	106	137	143	150	135	118	123	109	71
Davis at Jackson	772	588	831	864	861	856	1033	1006	908	867	675	642
Embarcadero at Bryant	1169	1042	1263	1216	1206	1375	1338	1354	1229	1313	1039	1105
Embarcadero at Folsom	1138	976	1236	1232	1181	1203	1389	1302	1150	1276	976	910
Embarcadero at Sansome	1993	1825	2276	2186	2392	2390	2588	2710	2293	2527	1875	1879
Embarcadero at Vallejo	841	714	870	845	903	887	973	952	984	959	724	707
Evelyn Park and Ride	122	104	124	134	161	162	172	181	107	159	132	104
Franklin at Maple	20	9	28	12	18	9	22	23	28	26	17	9
Golden Gate at Polk	577	461	608	630	667	624	734	684	630	665	521	469
Grant Avenue at Columbus Avenue	1297	1172	1348	1427	1394	1284	1459	1454	1364	1518	1184	1190



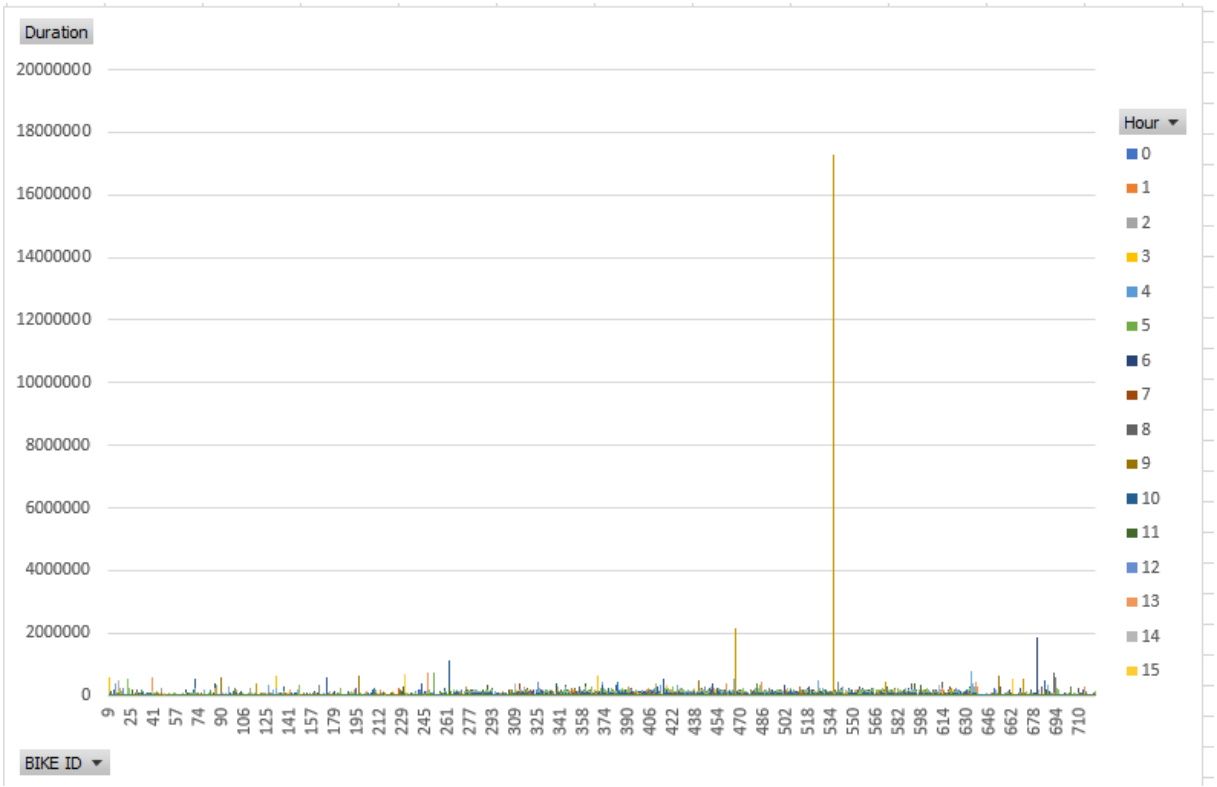
8. Liczba wycieczek dla klientów zamieszkujących pod każdym kodem pocztowym ze względu na miesiąc.

FACT TRIP Count	Column Labels												
Row Labels	January	February	March	April	May	June	July	August	September	October	November	December	Grand Total
00505			2										2
00506							10				1		11
00507			1						2				3
00510		4				4							8
00514				2								1	3
00517											1		1
00521		2						1			2	4	9
00553				1									1
00555								1					1
00556				2									2
00562						1							1
00570						6							6
00571			1										1
00577				11									11
00580								3					3
00591						2							2
00598			8	3		4							15
00600								1					1
00603		4											4
00604			9		1					4		2	16
00607		3											3
00610			4										4
00612										1			1
00614												4	4
00617							6						6
00626								1					1
00638		1	6	4	1								12
00646		1						2					3
00650					3	1	3						7
00652							1						1
00664						2							2



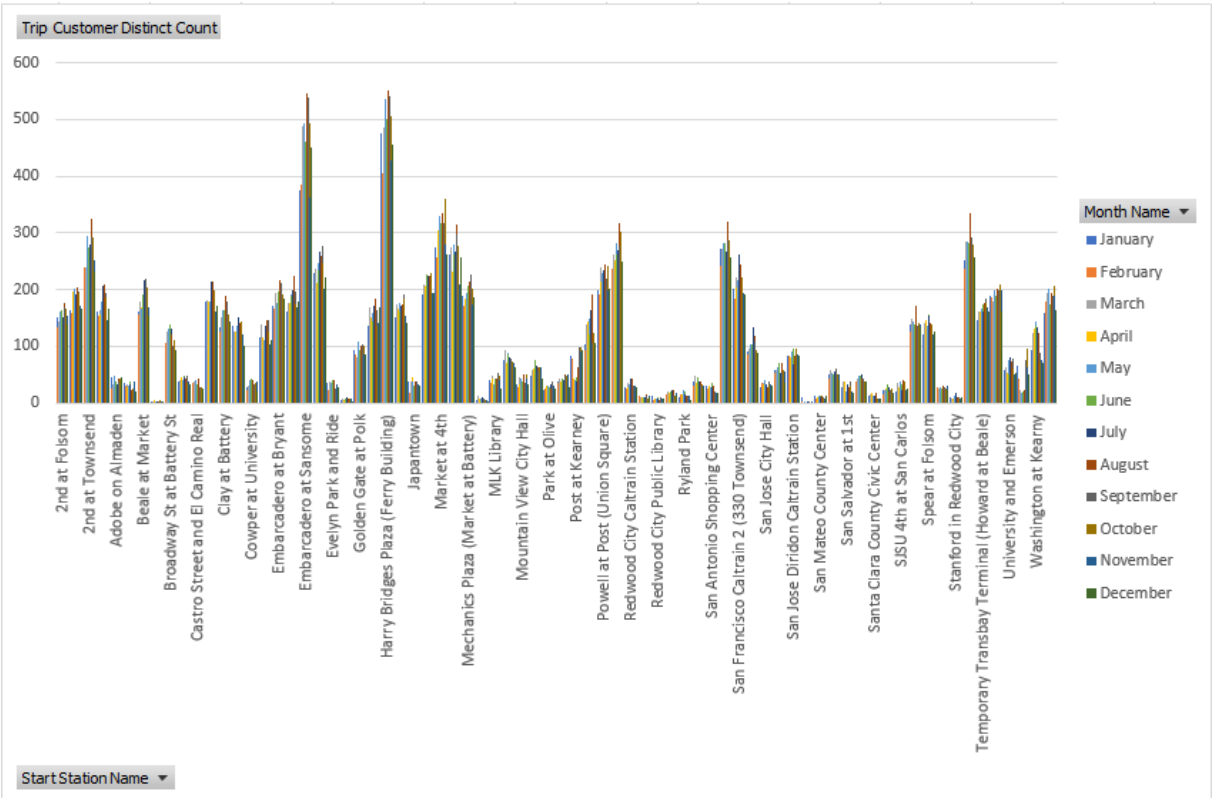
9. Sumaryczna długość wycieczek dla każdego roweru ze względu na godzinę.

Duration	Column Labels																							
Row Labels	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Grand Total
9			4262		602338			694	15807	12438	36860	3663	10358	9588	10657	15252	17984	20081	81122	11608	20357	2299	6734	731	8205	891038
10		38525					466	5202	10335	9101	115976	9276	4868	13467	8049	8570	37653	23704	17357	175275	4944	48265	7789	1472	4695	544989
11		409	830	128783			960	970	6579	10064	4609	1553	19031	4570	1540	11445	14219	22685	8557	66998	5192	2681	1469	1008	859	315011
12		246	1571				474	3017	2765	10312	10643	4753	224866	112207	7349	2562	4509	28674	22969	173053	11113	17280	64839	3989	4505	711696
13			676				514	1345	5023	9623	23565	17721	13347	14944	17388	22568	29969	377714	11679	16249	6792	1802	6021	874		577814
14							670	207	9141	8757	19611	3973	24451	21697	13240	29344	82191	7976	34164	28888	8832	2563	97089	1019	2972	396785
15						681	263	1980	7390	29903	4255	4311	18779	36636	9442	506638	23446	9358	11673	63942	88354	11111	2095		892	831149
16		3429	1502	940		798	4663	18575	49792	84647	73359	101382	78484	53141	200792	40246	153724	69610	232294	62452	27375	15346	16656	8703	5439	1303349
17			2119	154	73313		804	5335	16455	38939	14063	5805	21553	12711	33638	196869	2251	14628	25541	20296	19835	764	901	2317	1115	509406
18		958	131443				289	2415	10311	23536	12742	9678	3348	11091	17136	35241	6316	94825	33785	9722	7309	3702	66194	920	1793	482754
19		1283	3598				407	2101	13264	11777	7491	17535	6593	254671	10884	28239	3966	9333	68515	23414	5564	52585	2289	1988	1833	527330
20			595				3322	1841	5284	18207	9489	28905	15959	26392	12164	10235	20113	32782	25164	9747	8680	10970	12944	2394	4085	259272
21		501		8376		2896	266	1247	7894	18186	12327	73742	7137	19358	26413	14394	33120	11586	16692	11033	5382	4863	4924	1632	867	282836
22		977			584		294	2032	9091	39254	71628	16308	8249	64521	13203	26981	23309	30694	519987	16815	10280	73087	722	4040		932056
23		1870					257	9612	14563	9018	7841	5534	5717	10194	8617	16064	10627	14803	269316	16808	6740	3997	2089	2407	2431	418505
24		170						906	12397	7985	17221	10093	3918	40238	31588	24640	29548	4570	17711	69714	14952	1399	5214	783	486	293533
25		628	1020				1707	2261	12873	18400	10528	4636	183974	13688	16155	19125	14115	17288	19780	74151	10525	23411	3399	4873	785	453322
26		38986	1775	20899				937	3478	3020	4890	3126	1382	35208	19618	12908	3313	13996	5196	3121	4164		432	1316	198	177963
27		208					245	3691	9988	44445	8183	23120	36347	52115	8573	13958	19264	17520	23013	21372	104664	9073	339	7176	304	403598
28			3062					1153	14217	25366	14527	44273	22025	21054	93059	14946	20563	16402	17476	15500	11084	2359	1132	3930	459	342587
29		2179	530		1051	1136	3633	12698	43400	205775	66964	88944	45228	115553	98188	88819	58113	93163	91682	80119	26639	15361	10270	65528	2341	1217314
30		74923	994			39751	781	4994	10788	7787	2642	22080	6574	7562	13562	21526	9241	10747	12680	13665	8729	4204	79107	4057	1429	357823
31		300	77724			499	807	4297	14762	14350	18768	22265	9186	38821	10897	11454	19006	16252	16329	22767	95556	3832	3483	2944	859	405158
32		10934	1745	10490			213116	784	8721	12376	3886	52928	8660	8364	45905	13624	4559	12418	16925	118275	5429	4742	3459	2041		559381
33		5120	484	540	1104	598	3832	12965	51169	96694	93760	104338	95768	143048	91746	52068	81207	106028	117425	84474	24819	16815	59409	6009	1453	1250873
34		1177						232	171	416			1178	2817		1004	176	952	504	2396	694	1144		720		13581



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

Trip Customer Distinct Count	Column Labels												
Row Labels	January	February	March	April	May	June	July	August	September	October	November	December	Grand Total
2nd at Folsom	152	134	143	156	162	164	152	177	169	166	153	139	573
2nd at South Park	165	158	186	196	202	192	192	204	197	193	172	166	698
2nd at Townsend	211	238	240	261	294	274	279	324	291	291	253	231	1139
5th at Howard	162	154	165	155	178	177	207	209	193	186	147	166	845
Adobe on Almaden	45	26	34	36	49	39	33	43	37	43	32	45	185
Arena Green / SAP Center	36	30	34	27	30	38	32	23	23	26	38	20	166
Beale at Market	161	157	180	169	167	191	216	218	219	203	169	154	783
Broadway at Main	4	4	3	5	3	2	4	2	5	5	4	4	18
Broadway St at Battery St	92	105	126	117	130	138	132	121	102	110	93	94	492
California Ave Caltrain Station	38	26	40	45	41	47	45	43	47	36	38	32	234
Castro Street and El Camino Real	35	36	37	37	40	33	40	44	29	27	24	26	131
Civic Center BART (7th at Market)	178	157	182	178	180	172	214	215	193	199	161	171	1018
Clay at Battery	133	125	150	160	163	164	157	189	178	157	144	142	800
Commercial at Montgomery	136	114	125	125	135	139	151	141	139	143	122	102	560
Cowper at University	29	24	30	29	41	44	35	41	33	36	37	21	187
Davis at Jackson	116	93	138	115	110	124	135	145	146	125	104	111	519
Embarcadero at Bryant	171	167	193	159	177	196	201	217	212	191	155	185	986
Embarcadero at Folsom	162	154	176	177	191	190	199	225	196	175	169	179	753
Embarcadero at Sansome	375	384	487	418	494	461	515	546	538	494	362	451	2835
Embarcadero at Vallejo	229	220	237	211	246	233	268	260	277	246	201	221	1490
Evelyn Park and Ride	35	24	38	30	36	40	36	40	25	28	33	29	127
Franklin at Maple	5	5	9	6	6	8	10	9	6	7	7	4	38
Golden Gate at Polk	93	86	82	78	108	94	100	100	104	101	85	87	495
Grant Avenue at Columbus Avenue	136	148	169	152	158	143	172	185	165	137	141	168	1062

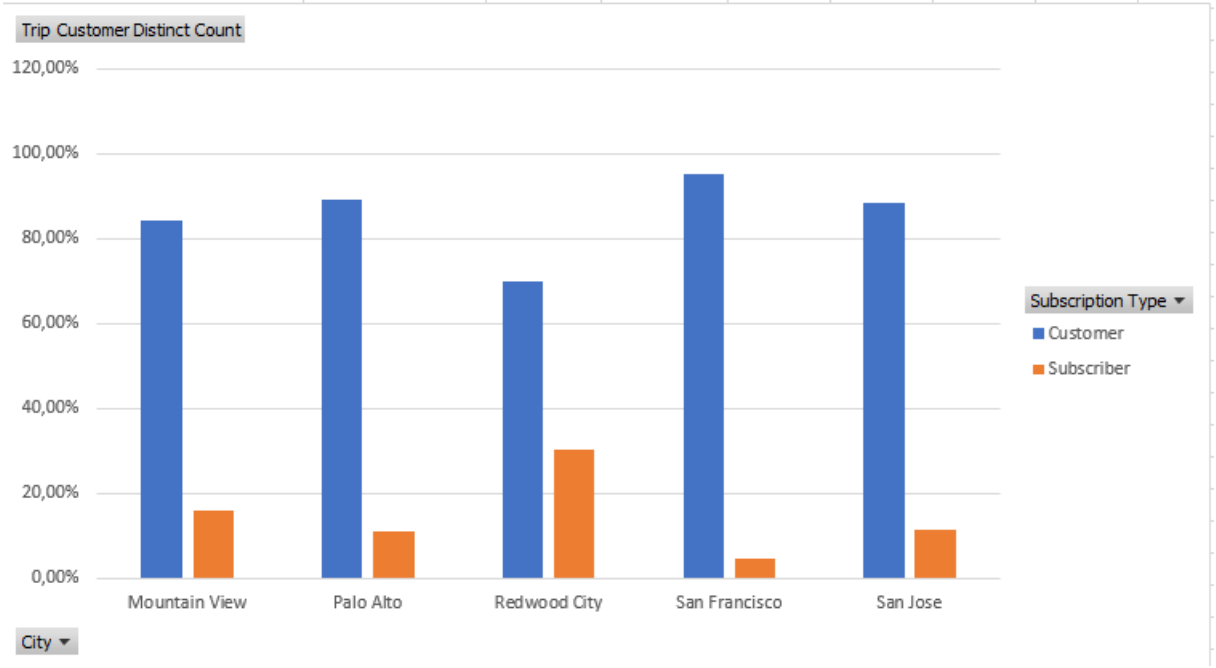


Analiza zestawień

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 prawdopodobnie 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.

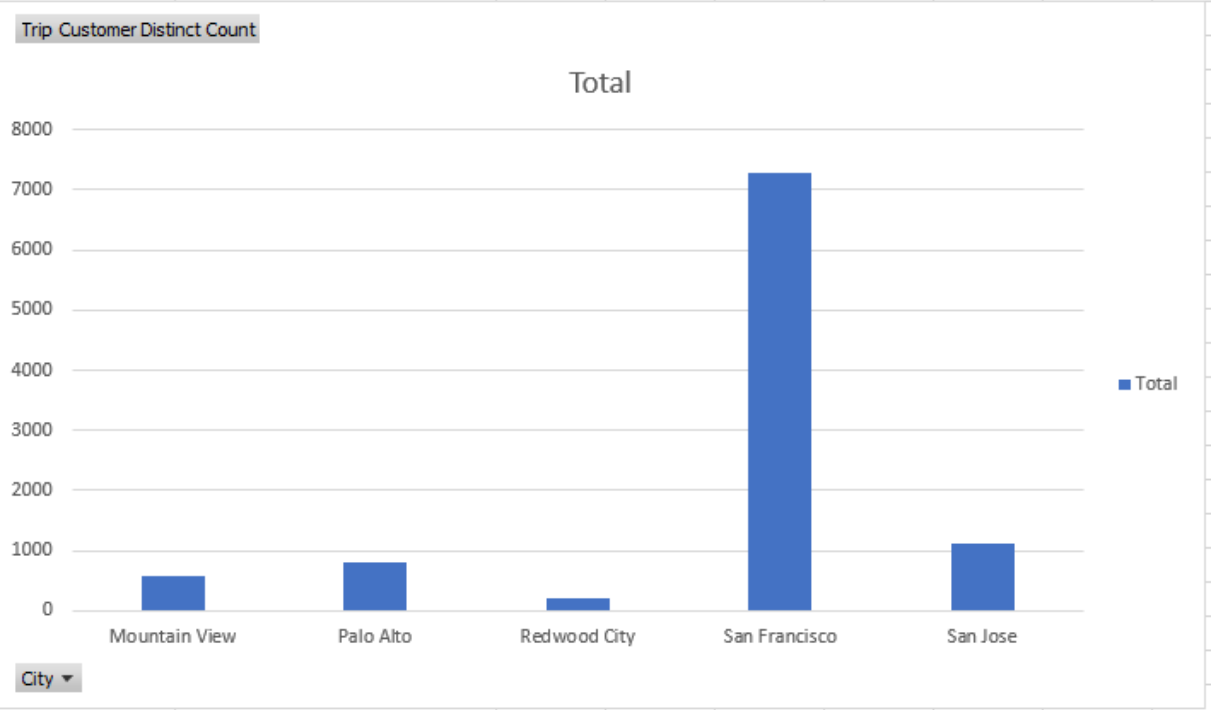
Analiza w głąb

Trip Customer Distinct Count	Column Labels		
Row Labels	**Customer**	**Subscriber**	**Grand Total**
Mountain View	84,22%	15,78%	100,00%
Palo Alto	89,07%	10,93%	100,00%
Redwood City	69,90%	30,10%	100,00%
San Francisco	95,26%	4,74%	100,00%
San Jose	88,46%	11,54%	100,00%
Grand Total	**95,38%**	**4,62%**	**100,00%**



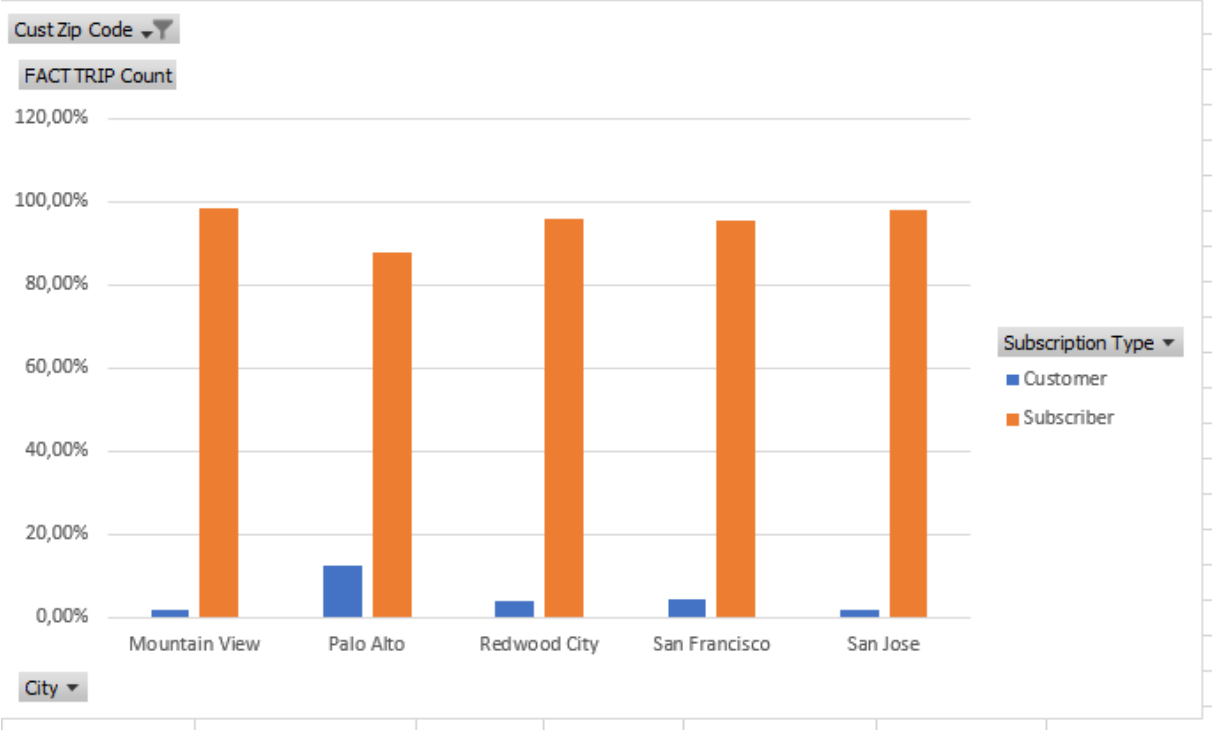
Rozpoczynam analizę z zestawienia trzeciego wykonanego w ramach punktu 1.5.2. W tabel i na wykresie przedstawiony został procentowy udział subskrybentów oraz „przypadkowych" klientów ze względu na unikalną liczbę klientów dla każdego miasta. Jak widać na wykresie, ewidentnie największy udział subskrybentów 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.

Row Labels	Trip Customer Distinct Count
Mountain View	564
Palo Alto	796
Redwood City	206
San Francisco	7278
San Jose	1109
Grand Total	**7766**



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.

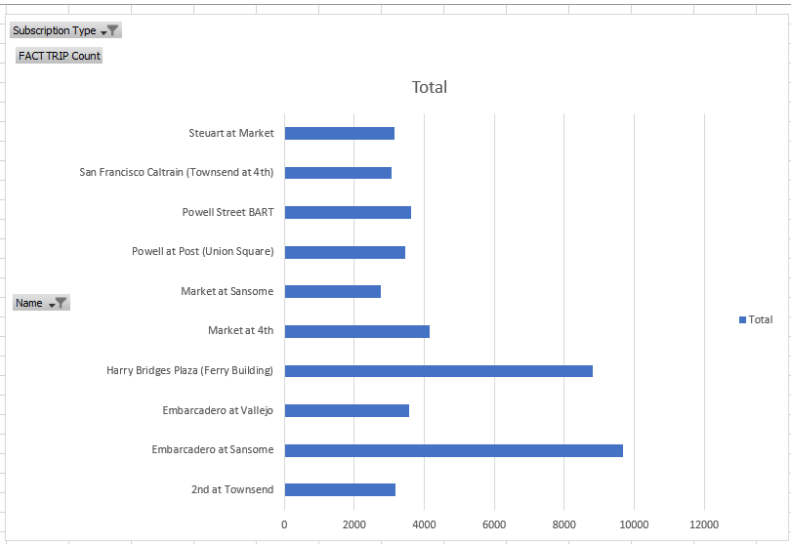
Cust Zip Code	(Multiple Items)		
FACT TRIP Count	**Column Labels**		
Row Labels	**Customer**	**Subscriber**	**Grand Total**
Mountain View	1,74%	98,26%	100,00%
Palo Alto	12,22%	87,78%	100,00%
Redwood City	3,99%	96,01%	100,00%
San Francisco	4,41%	95,59%	100,00%
San Jose	1,88%	98,12%	100,00%
Grand Total	**4,36%**	**95,64%**	**100,00%**



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).

Subscription Type	Customer

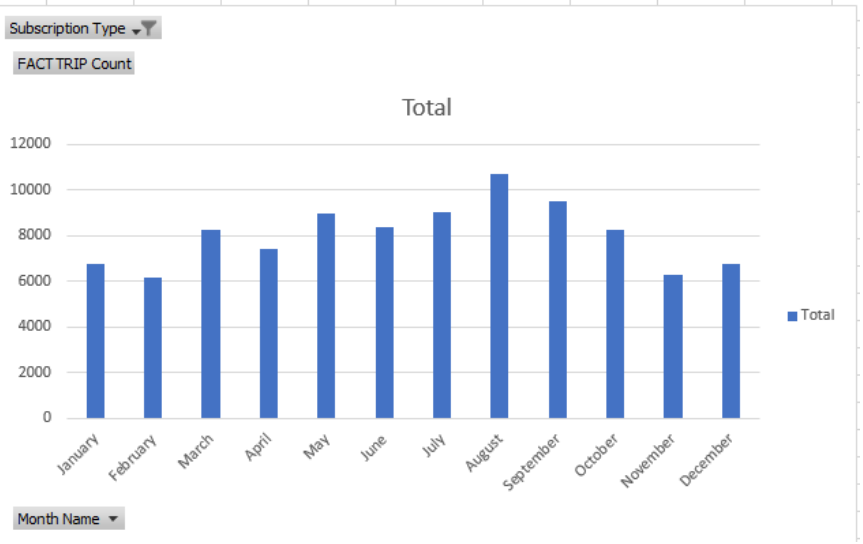
Row Labels	FACT TRIP Count
2nd at Townsend	3181
Embarcadero at Sansome	9666
Embarcadero at Vallejo	3577
Harry Bridges Plaza (Ferry Building)	8810
Market at 4th	4163
Market at Sansome	2768
Powell at Post (Union Square)	3444
Powell Street BART	3622
San Francisco Caltrain (Townsend at 4th)	3074
Steuart at Market	3133
Grand Total	**45438**



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.

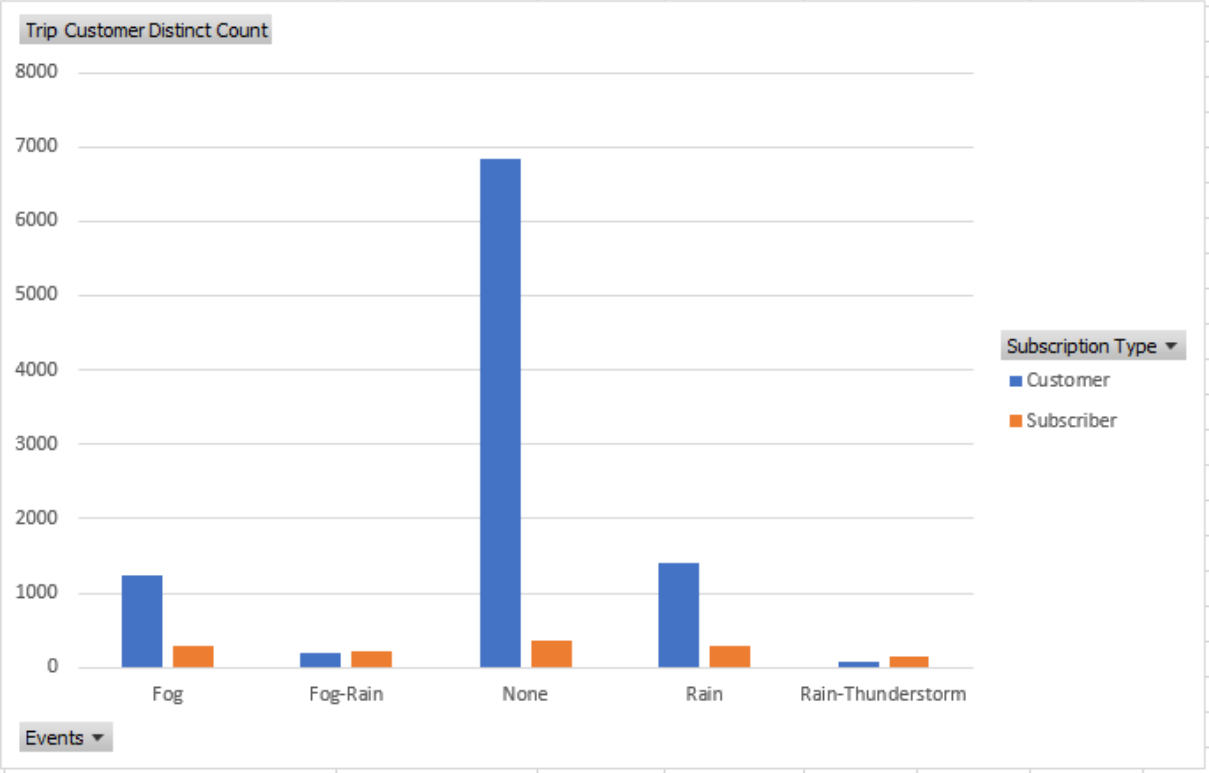
Subscription Type	Customer

Row Labels	FACT TRIP Count
January	6747
February	6191
March	8248
April	7435
May	8991
June	8388
July	9052
August	10714
September	9528
October	8250
November	6289
December	6761
Grand Total	**96594**



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.

Trip Customer Distinct Count	Column Labels		
Row Labels	Customer	Subscriber	Grand Total
Fog	1233	278	1511
Fog-Rain	202	214	416
None	6848	355	7203
Rain	1399	291	1690
Rain-Thunderstorm	82	151	233
Grand Total	**7407**	**359**	**7766**



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).