Project 03

Bay Area Bike Share

Agenda

- Dataset Overview
- Data Cleaning
- Star Schema Approach
- Analysis
- Conclusion

Why Bay Area Bike Share?

- Transaction level data
- Compare and contrast prior exploration of bike share data via Unix and other tools with SQL and the star schema approach
- Identify whether trends apparent in other bike share city data sets are also present in the Bay Area

About the Dataset

- Sources: 2013 Bay Area Bike Share, Weather Underground
  - Trip: Transaction level ride data, 600K+ observations
  - Station: Bike station info, 70 observations
  - Weather: Hourly weather data, 3,665 observations

- Type of Variables:
  - Trip: Date, Duration, Location, Bike Info, Type of Rider
  - Station: Location, Number of Docks
  - Weather: Temperature, Humidity, Wind, Weather Events

- Due to computational limitations of datanotebook.org for certain functions, we choose to sample 250,000 observations from the Trip data set to ensure that anyone using datanotebook.org could reproduce our analysis
  - Complete Cases of Ridership + Weather

Data Cleaning

- Consistency Between / Among Data Sets:
  - Missing Zip Codes
  - Zip Codes Structure: 5 vs 10
  - Zip Code Formatting: Whitespace
  - Categorical Weather Value Formatting: Rain vs rain

- Dealing with Null Values:
  - Missing Weather Data Updates
    - Integer nulls as -999
    - Categorical nulls as ‘Null’

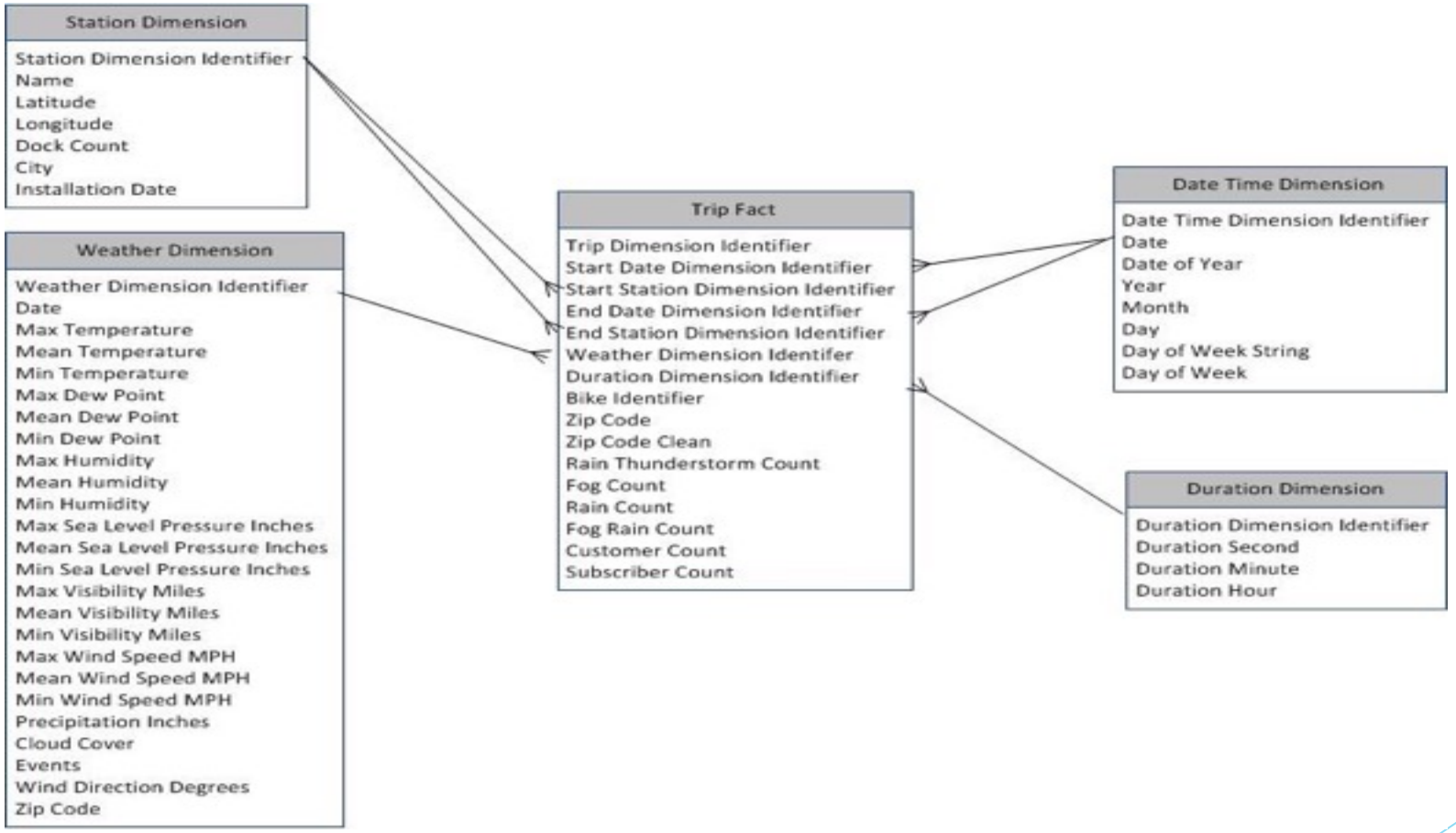
Creating & Populating Tables

- Database tables for trip, station, and weather
- Populated the database tables by pulling in data from the raw csv files
  - Count checks to verify correct number of observations
- INFORMATION_SCHEMA metadata database to extract a schema from a pgsql instance to better understand each table

Steps to create Star Schema

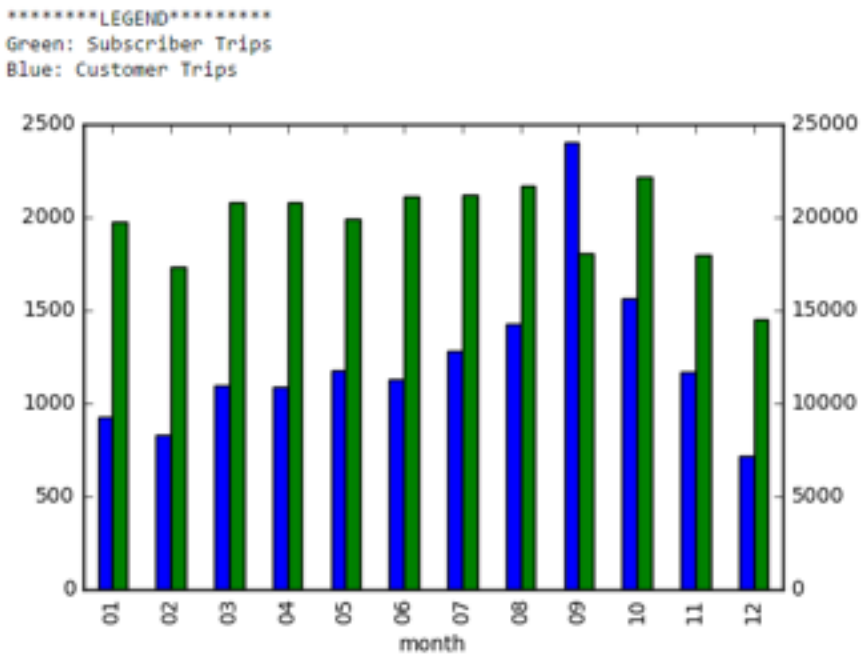
- The tables include station, weather, date-time, and duration.
- The fact table is unique by date, duration, start station, end station. It is also unique by two additional fields: bike identifier and zip code.
- To gather some numeric data, we converted the weather event column into a binary variable which counts the number of occurrences of rain storms, fog, rain, and fog-rain and the subscription_type column to count the subscribers versus casual riders.
- The numeric values in the weather table are not included in the fact table because they are not additive values and are aggregated to mean, max, min, etc.
- After this, we created our fact and dimension tables and populated them using INSERT and UPDATE statements.
- Also, to improve efficiency of our analysis, we created additional columns such as year, month and day.

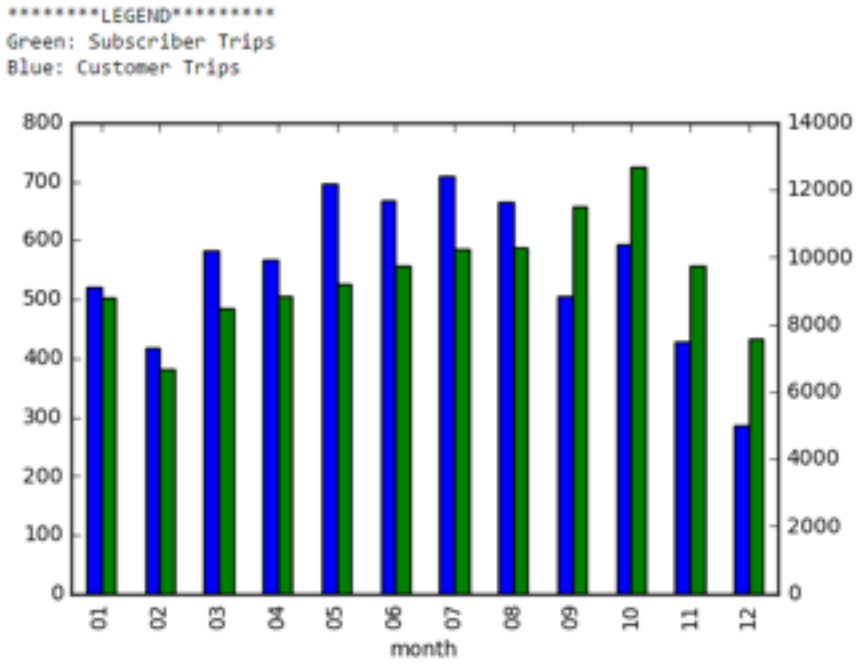
Star Schema


Station Dimension
Station Dimension Identifier
Name
Latitude
Longitude
Dock Count
City
Installation Date
Weather Dimension
Weather Dimension Identifier
Date
Max Temperature
Mean Temperature
Min Temperature
Max Dew Point
Mean Dew Point
Min Dew Point
Max Humidity
Mean Humidity
Min Humidity
Max Sea Level Pressure Inches
Mean Sea Level Pressure Inches
Min Sea Level Pressure Inches
Max Visibility Miles
Mean Visibility Miles
Min Visibility Miles
Max Wind Speed MPH
Mean Wind Speed MPH
Min Wind Speed MPH
Precipitation Inches
Cloud Cover
Events
Wind Direction Degrees
Zip Code
Trip Fact
Trip Dimension Identifier
Start Date Dimension Identifier
Start Station Dimension Identifier
End Date Dimension Identifier
End Station Dimension Identifier
Weather Dimension Identifer
Duration Dimension Identifier
Bike Identifier
Zip Code
Zip Code Clean
Rain Thunderstorm Count
Fog Count
Rain Count
Fog Rain Count
Customer Count
Subscriber Count
Date Time Dimension
Date Time Dimension Identifier
Date
Date of Year
Year
Month
Day
Day of Week String
Day of Week
Duration Dimension
Duration Dimension Identifier
Duration Second
Duration Minute
Duration Hour

Analysis – Rider Seasonality

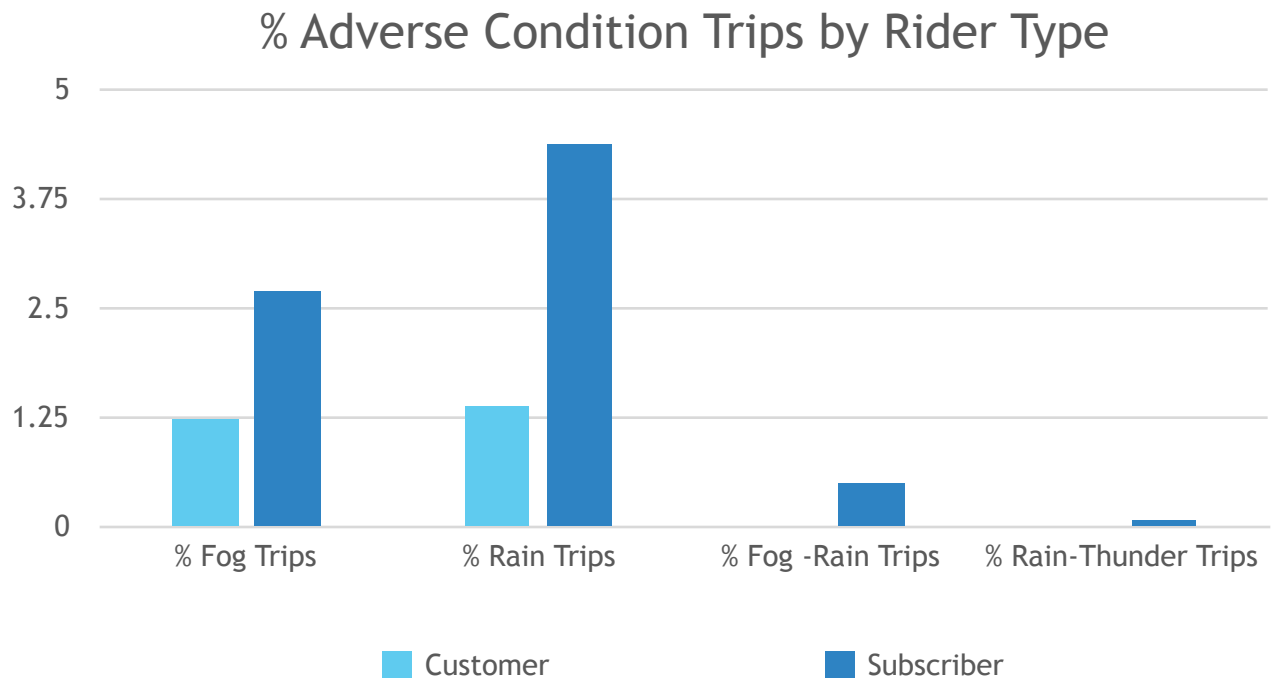
- Exploring the month variable shows that ridership does tend to be higher in the summer months for the full dataset.
  - There was unevenly distributed data for each of the years, so this may be skewing results.
- To isolate seasonality, only include 2014 since a full year exists
- Seasonality of trips differs for customers versus subscribers. Generally, over the summer months customer trips surpass subscriber trips.





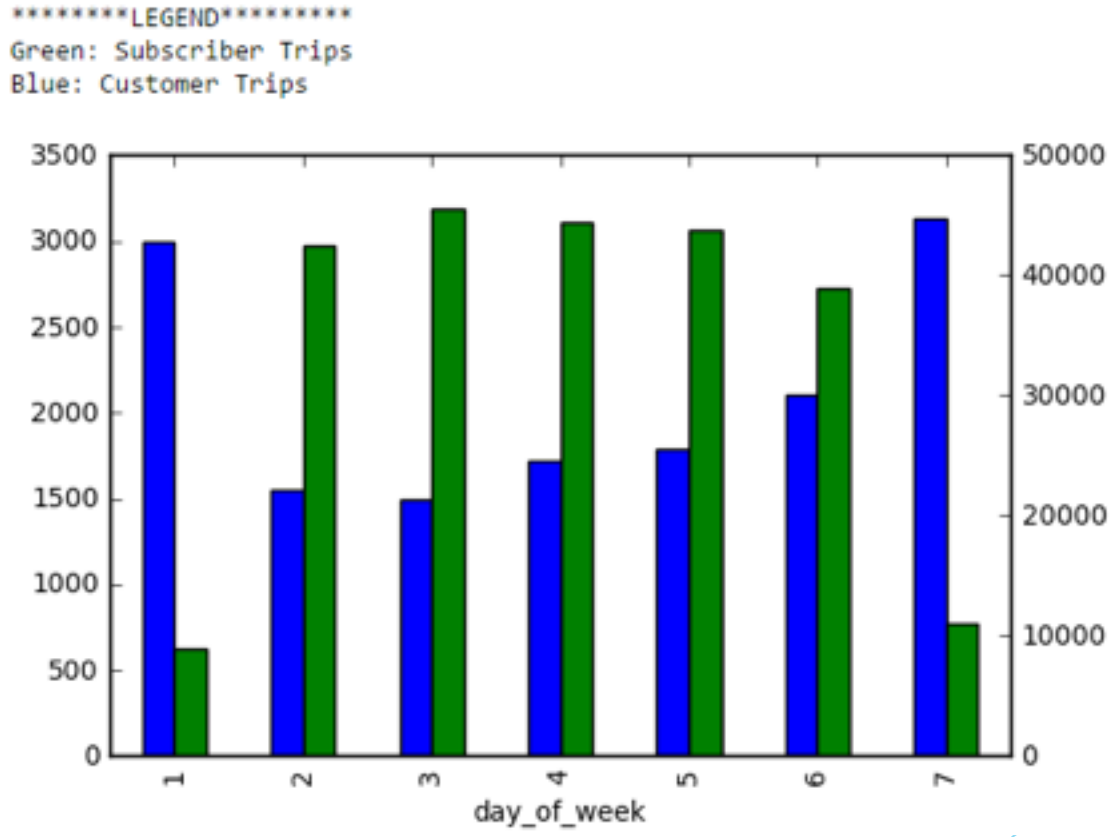
Analysis – Adverse Weather

- Subscribers are much more likely to ride in adverse conditions compared to customers as a percentage of total trips each respective category
- The total percentage of trips in the rain and thunder for a customer was .00676% compared to .0897% for subscribers.
- This suggests that subscribers sometimes don't have a choice but to ride in inclement weather, if they have to commute, etc.



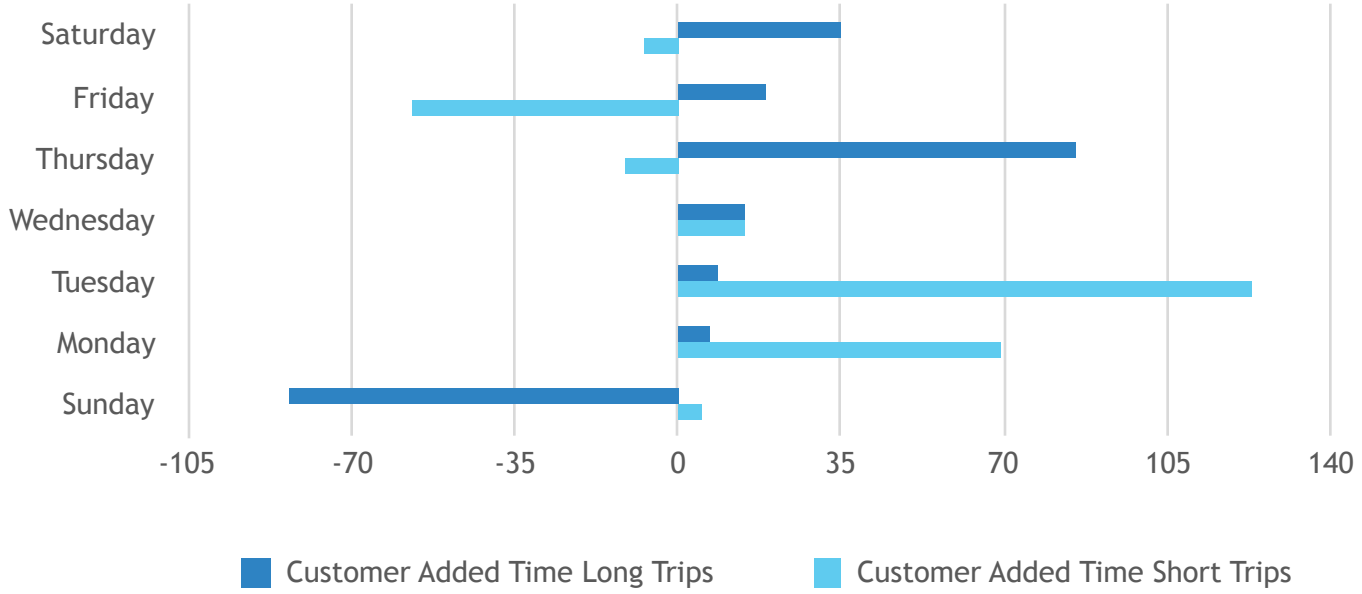
Analysis – Days of Week

- Investigating the day of week variable shows that ridership during weekdays far surpasses ridership on weekends. However, when splitting the data between subscribers and non subscribers, it is clear that non-subscriber ridership increases on the weekends, while subscriber ridership drastically decreases on weekends.
- Among subscribers, ridership on Friday is lower than any other weekday.



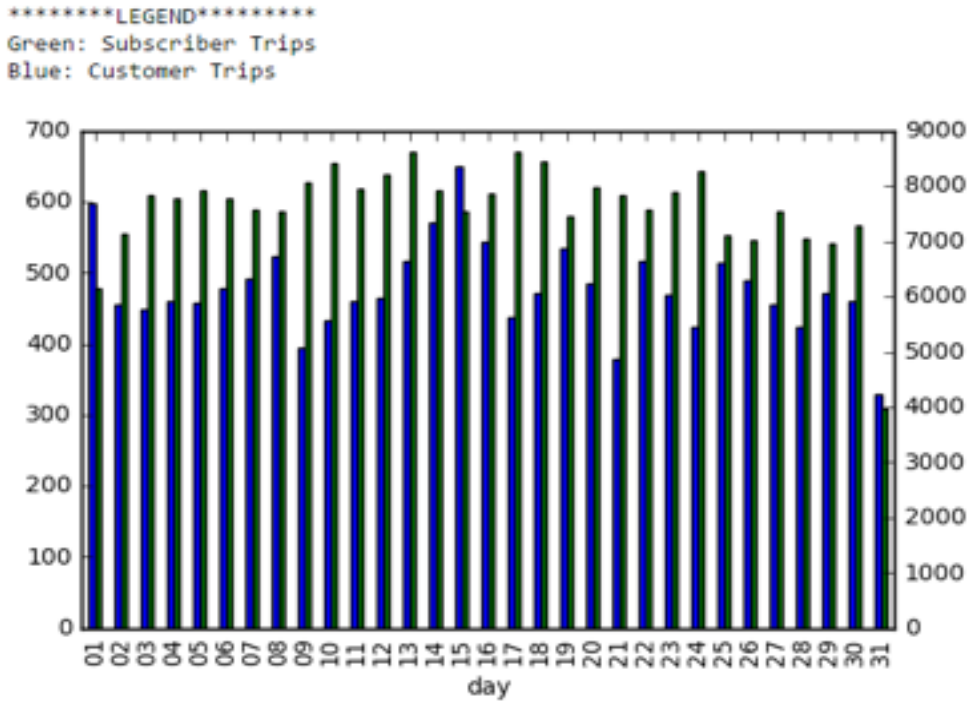
Analysis – Trip Duration

- When limiting the trip time to 20 minutes, there is not really much of a material difference in trip times in casual versus subscribers
  - Random distribution of differences
- By exceeding the trip limit to 90 minutes, we found out that weekday trips are shorter for Subscribers compared to Casual riders. This could suggest that subscribers are more efficient when riding than Casual commuters.



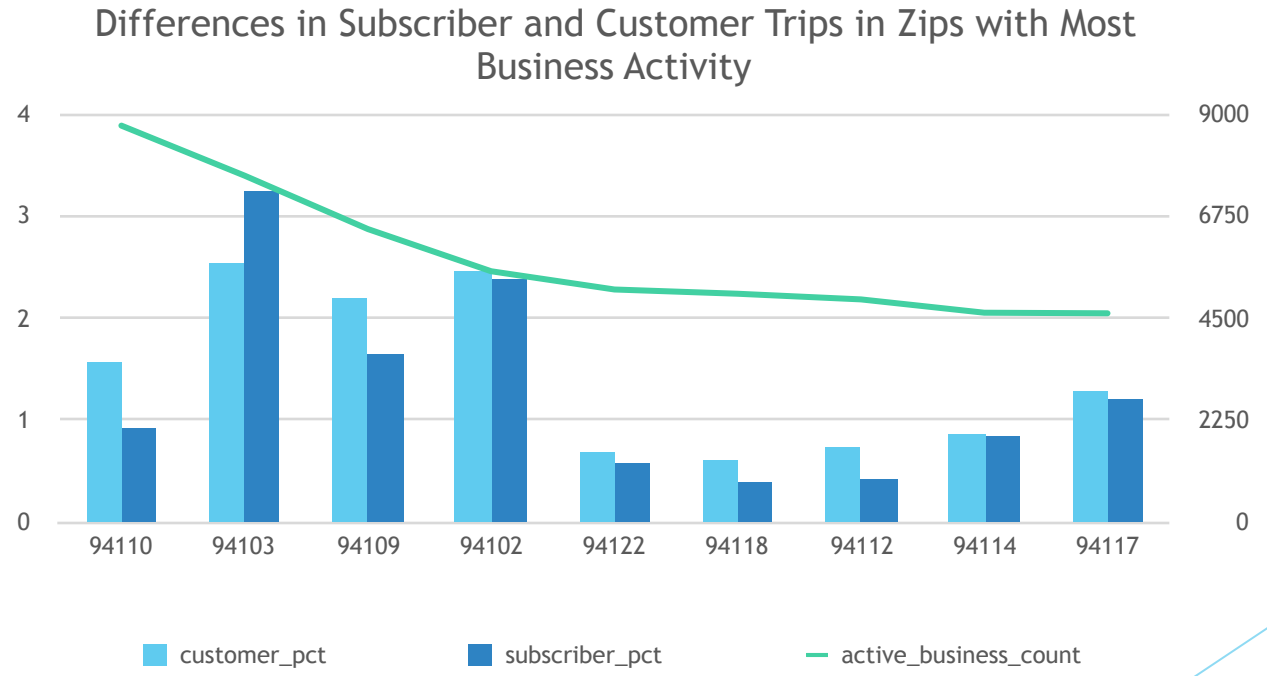
Analysis – Day of Month

- At last, we investigated relationship between day of month and ridership.
- For subscribers, it looks like ridership is fairly consistent throughout the month, with slightly lower ridership at the beginning and end of the month.
- However, it is interesting to see that the 1st-2nd and the 25th-30th days of the month appear to have lower ridership.
- For non-subscribers, ridership by day of the month is a bit more random, with peaks and valleys occurring throughout the month.



Analysis – Active Businesses

- Additional dataset from SFData.gov which had rows for active businesses in SF in addition to address information
  - Summarized by zip code and counted active businesses
- Zip codes with top 10 most active businesses
- Could be used to distinguish zip codes with higher percentages of businesses which would appeal to the customer versus zip codes which attract more subscribers (94107 and 94103) which attract more subscriber rides



Conclusion

- Customers and Subscribers show different patterns in use, most specifically by day of week
  - Should SF Bikeshare designate bikes to just subscribers vs customers?

- Future steps to improve analysis:
  - More accurate zip code information for trip dataset
  - Expand weather information for additional zip codes
  - Expand characteristic data for subscriber
    - Low, Medium, High Usage Rate
  - Observe rider activity in comparison to public transportation issues / traffic
    - Example – Safe Track's effect on DC Capital Bikeshare

Thank You

Lee Eyler, Divya Gorwara, Eugene Hwang and Marissa Wiener