Data Challenge

Using historical data to understand the impact of **temperature** on the **Chicago bike share program**

Fiona Strobridge

Chicago Bikes: Goal

Determine the effect of temperature on:

- Bike trip **lengths**
- Bike trip **counts**

Controlling for any relevant seasonality and customer segments

Approach:
- Clean the data
- Carry out exploratory data analysis to understand trends in the data
- Identify relevant customer segments
- Identify seasonality effects
- Understand temperature effects on trip lengths
- Understand temperature effects on trip counts (over appropriate timeframe)

There were two datasets:

- Raw
- Cleaned

This study focuses on the **raw** data; to ensure greater control over imputing missing values and assumptions around what is "relevant" data

The data:
- 1.38 million entries form 06/27/2013 - 10/01/2017
    - ~9k trips per day
- Mean trip ~17 mins; median is ~12 mins; the data is positively skewed
- Longest trips are limited to 1 day

Chicago Bikes: Features

Trip Information	User Information	Location Information	Weather Information
Trip ID	User Type	From Station ID	Temperature
Start Time	Gender	From Station Name	Windchill
Stop Time		Latitude Start	Dewpoint
Trip Duration		Longitude Start	Humidity
		DP Capacity Start	Pressure
		To Station ID	Visibility
		To Station Name	Wind Speed
		Latitude End	Precipitation
		Longitude End	Events
		DP Capacity End	Rain
			Conditions

Chicago Bikes: User Types

Dependent - Single

Discounted Single Ride

$1/trip

one trip up to 30 minutes

See Details

Customer - Daily

Day Pass

$15/day

unlimited 3-hour rides in a 24-hour period

See Details

Subscriber - Annual

Discounted Annual Membership

$49.50/year

unlimited 45-minute rides

Join Now

Chicago Bikes: Cleaning the data

Duplicates:
- 61 rows (out of 1.38 million) were duplicated, so were dropped

Missing Values:
- Values denoted -9999 or -999 were treated as missing (weather data)
- **Gender was missing 27% of the values; this was kept and called "unknown" (i.e. 3 categories of gender: male, female and unknown)**
- Since this focus was on temperature and we still had the weather events information, all weather was dropped, except temperature, events and rain

Non-numerical Values:
- The station name features were dropped (as we have ID and lat/long)

Dates:
- Date, hour, DayOfWeek, Year, Month, Week, Day, DayOfYear, Weekend were added
- Public holidays were also added

UserType:
- 190 of the 1.38 million rows were "dependent" users (i.e. single ride), no meaningful conclusions can be made here; so only **Subscribers** (annual) and **Customers** (daily) were kept

Trip Duration:
- Verified that the trip duration was meaningful data and more accurate than (stop - start) time

Chicago Bikes: EDA

Temperature:
- Varies from **-15 to 96.1 °F**

Trip Duration:
- The maximum value is capped at 24 hours, so there is a slight increase in the data for this value
- In practise, if someone were to use the bike for a full day, it is reasonable to limit this to <16 hours
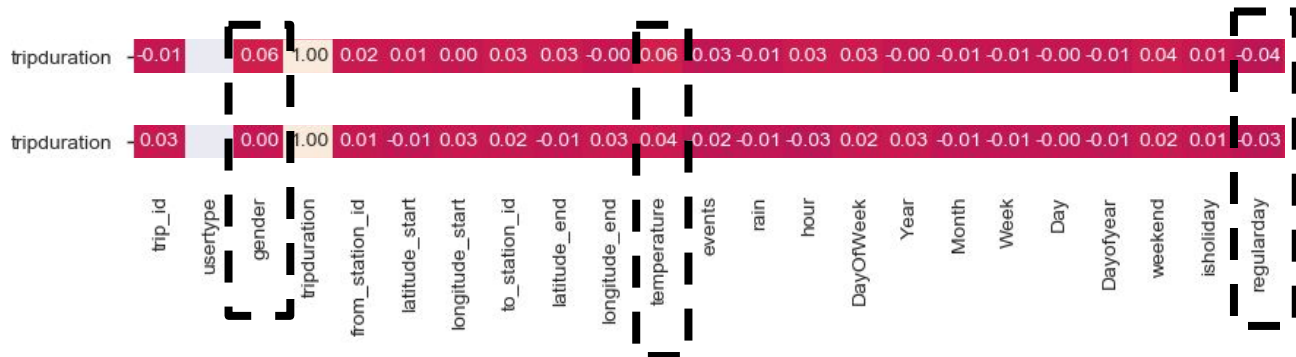- Therefore we are only looking at **trip durations < 16 hours**.

Correlation HeatMap:
- From the correlation coefficients, the factors that influence **trip duration** the most are:
    - User information has the strongest correlation; both **user type** and **gender**
    - Followed by time (i.e. **weekend** / **regular work day**)
    - Weather is also a factor; specifically **temperature**
    - And also the location (specifically the **longitude** start and end)

Chicago Bikes: Trends in the data segmented by User Type

- First look at the highest correlated feature:
  - Subscribers (annual) & Customers (Daily)

Insights
- We only have gender information for the "subscribers" (annual users)
- The **customers** mean trip duration is **29 minutes** compared **12 minutes** for the **subscribers**
  - The subscribers are 73% of the trips and the customers are 27% of the trips
  - 75% of the subscriber trips are by men and 25% are by women
  - Women have a slightly longer trip duration
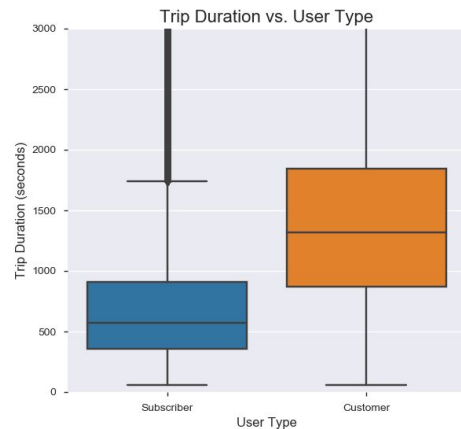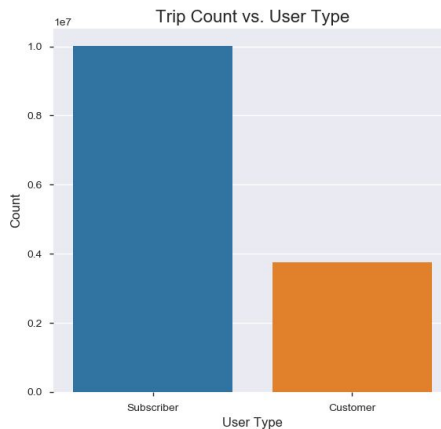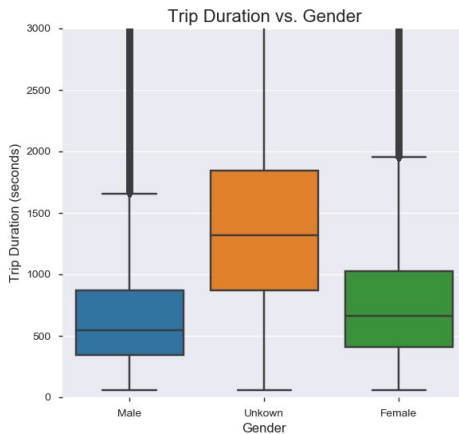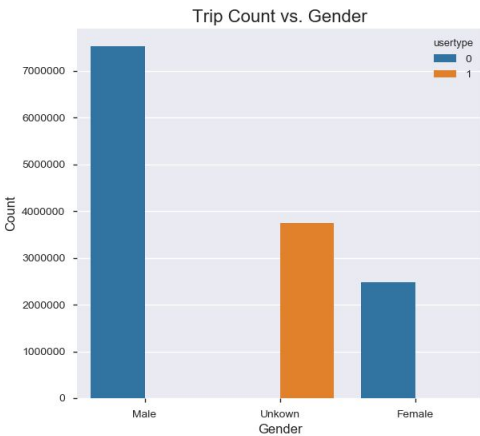
Chicago Bikes: EDA By User Type

Correlation HeatMap:
- From the correlation coefficients, the factors that influence **trip duration** the most are:
  - **Gender** - for the subscribers
  - Weather; specifically **temperature** (more strongly correlated for the **subscribers**)
  - Followed by time (i.e. **weekend** / **regular work day**)
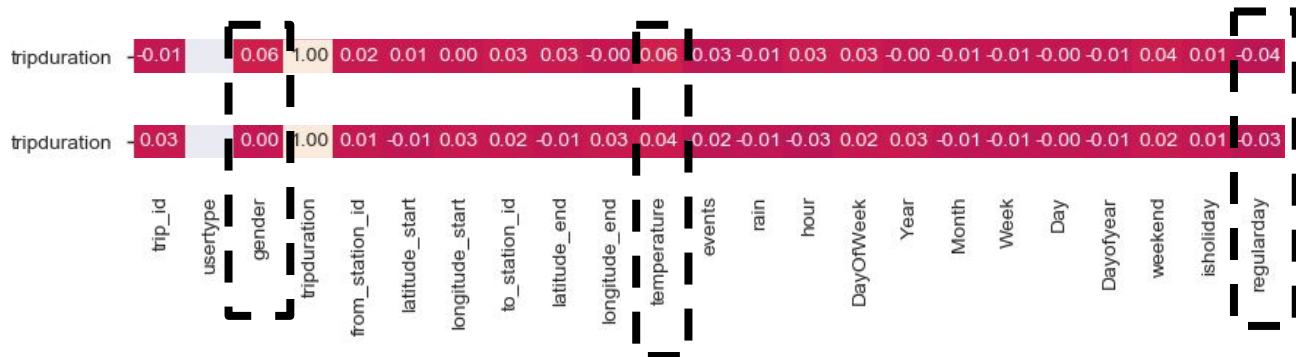
Chicago Bikes: Trends in the data segmented by User Type

Subscribers (Annual)
- Using the bikes more during the working week and between commuter times

Customers (Daily)
- Using the bikes more on the weekend and between 10 am and 10 pm, with a peak around 2 pm

Insights
- This is why the trip duration is correlated to the weekend / weekday, because it it related to the user types, who have different average trip durations, owing to their different use
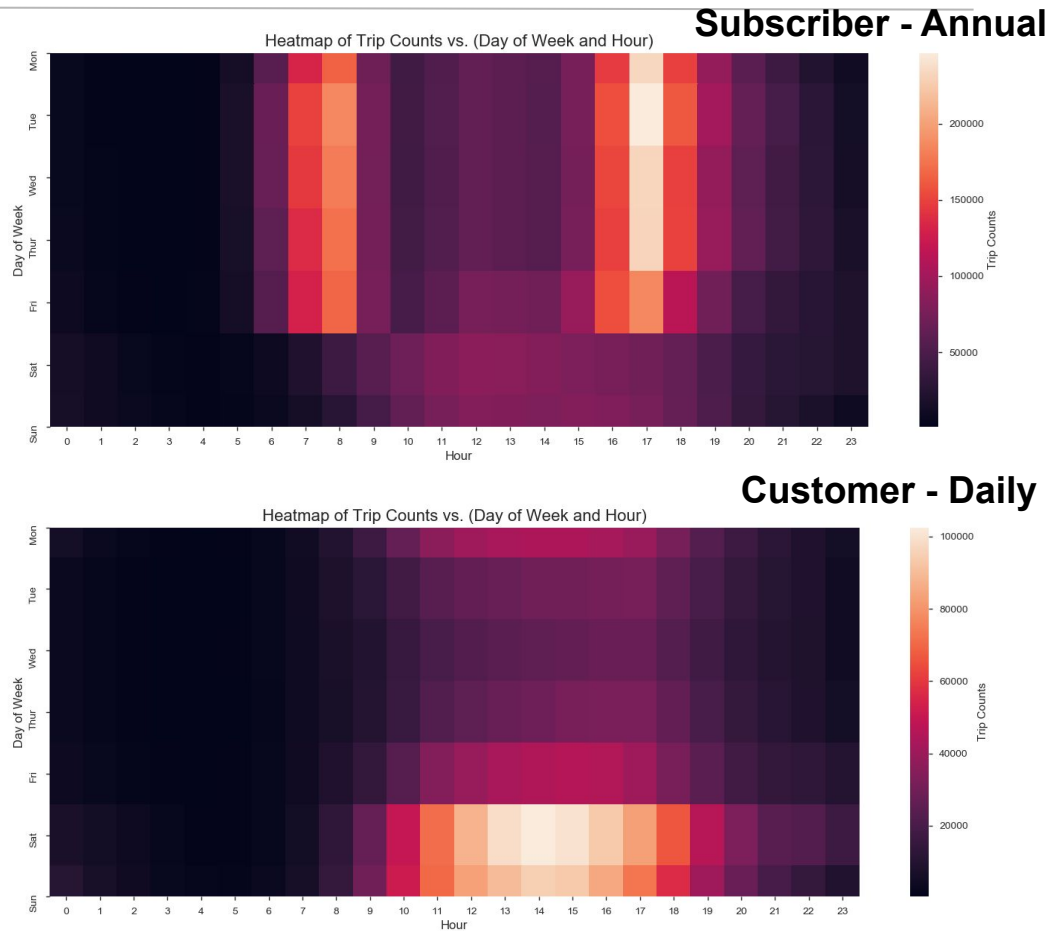
Subscriber - Annual

Heatmap of Trip Counts vs. (Day of Week and Hour)

Customer - Daily

Heatmap of Trip Counts vs. (Day of Week and Hour)

Chicago Bikes: Relationship between Temperature and Average Trip Duration

By User Segments (Subscription Males, Subscription Females and Daily Customers)
- When the temperature increases from 0 to 80 °F the average trip duration:
    - Increases from ~7.5 minutes to 10 minutes for male subscribers
    - Increases from ~8.3 minutes to 12 minutes for female subscribers
    - Increases from 12 minutes to ~22.5 minutes for customers (daily riders)
- For all users, it then decreases at 100 °F
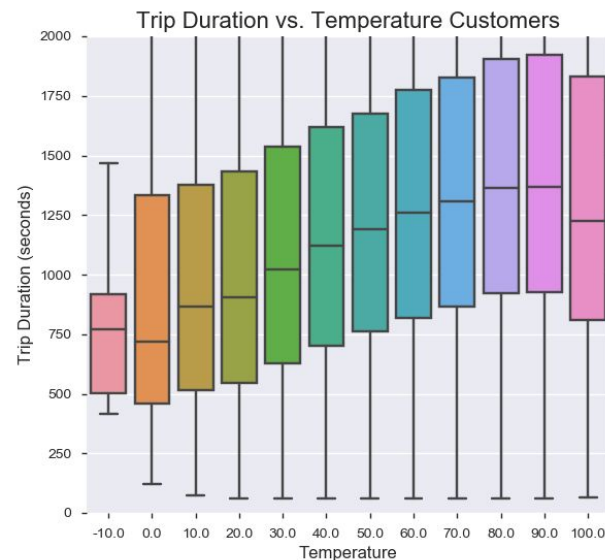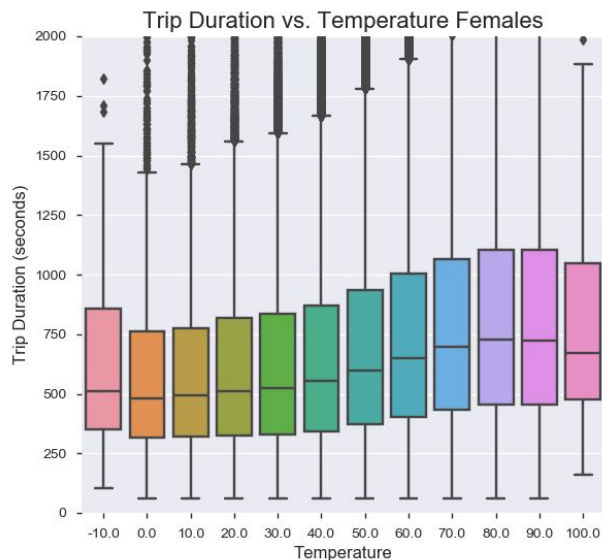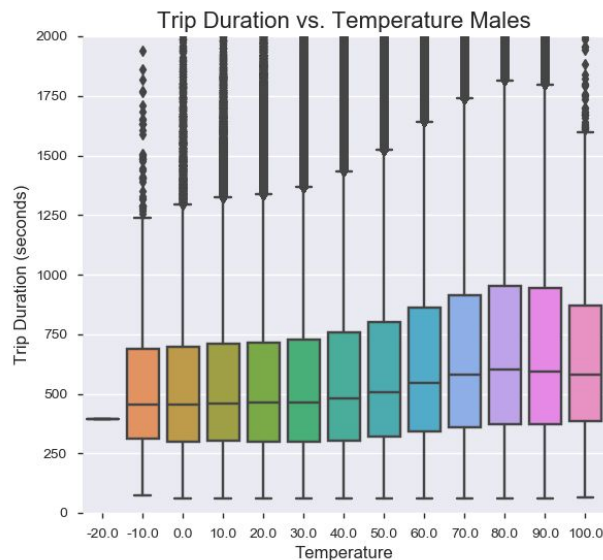
Chicago Bikes: Trends in the data segmented by User Type

Relationship on Temperature and Average Trip Duration by User Segments
- There are more trips at hours between 64 - 88 °F; but this may be because there are more hours at these temperatures
- Therefore we will look at the proportional effect to ride counts vs temperature:
    - Aggregating the data by the hour (since that's the detail we have weather data on)
    - Sorting by 3 user groups:
        - Daily users
        - Male subscribers
        - Female subscribers

Trip Count vs. Temperature
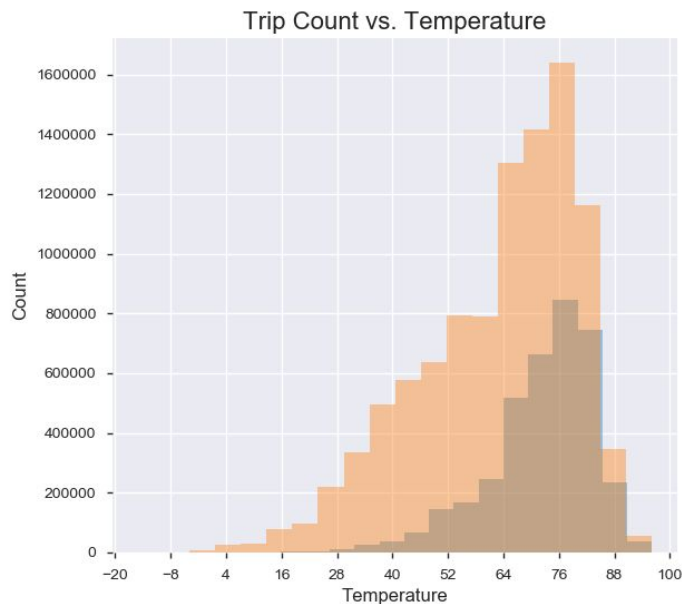
Chicago Bikes: Relationship between Temperature and Number of Trips

Daily Customers
- Are over **3 times** more likely to use the bikes **>75 °F**
- Are **1.5 times** more likely to use the bikes between **65 - 75 °F**
- Are less likely to use the bikes **< 65 °F** (trend increasing with decreasing temperature)

Subscribers
- Are over **2 times** more likely to use the bikes **>75 °F**
- Are **1.5 times** more likely to use the bikes between **65 - 75 °F**
- Are less likely to use the bikes **< 45 °F** (trend increasing with decreasing temperature)
- **Women** have a stronger correlation with temperature than men
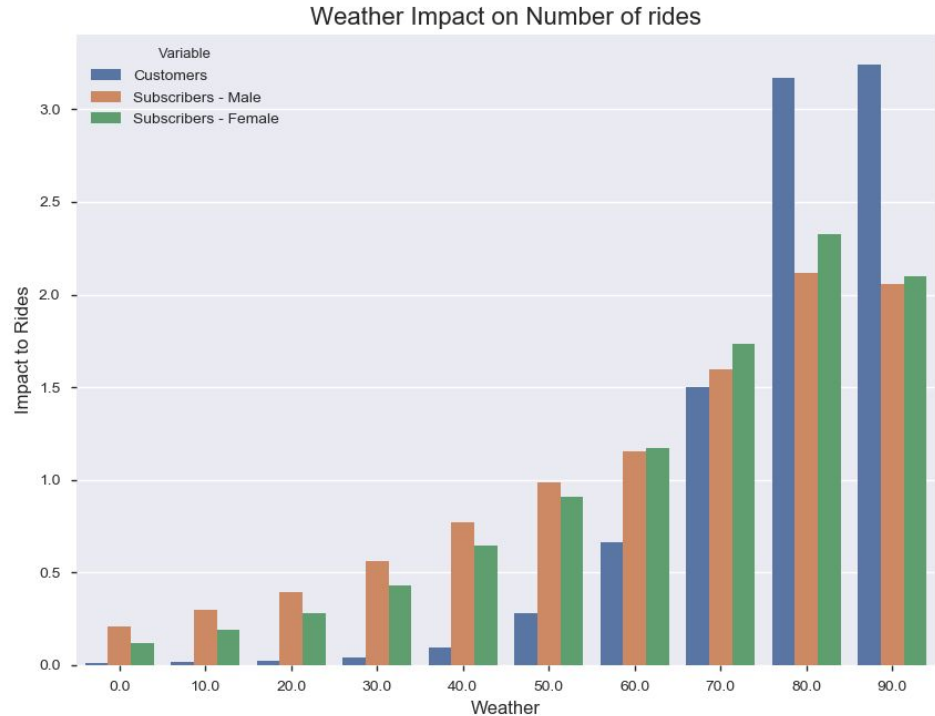
Weather Impact on Number of rides

Next steps:
- Look at the effects:
    - Over time (has this behaviour changed over the years)
    - By weather (specifically rain)
    - At hours of the day (commute time for subscribers and weekends for customers)
    - By locations

Chicago Bikes: Supplemental Slides

Chicago Bikes: Trends in the data before segmentation

Trends over the years:
- Increasing number of trips year over year from 2013 => 2017
- Decreasing duration of the trips
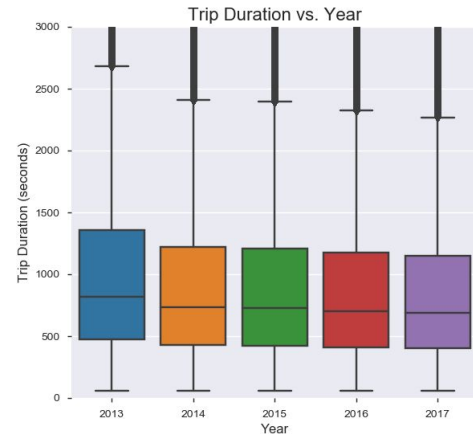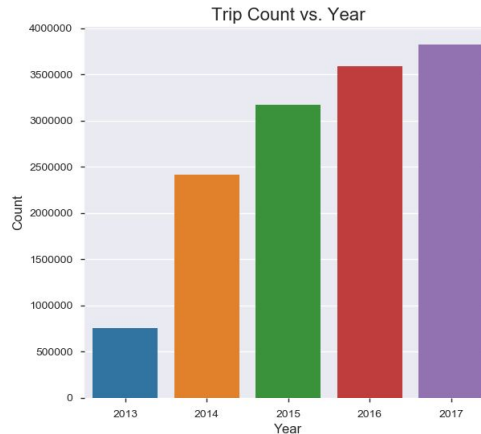
Trends over months:
- Increasing number of trips in the summer months (factor of 10) compared to winter
- Increasing duration of the trips in the summer months (50% increase) from winter
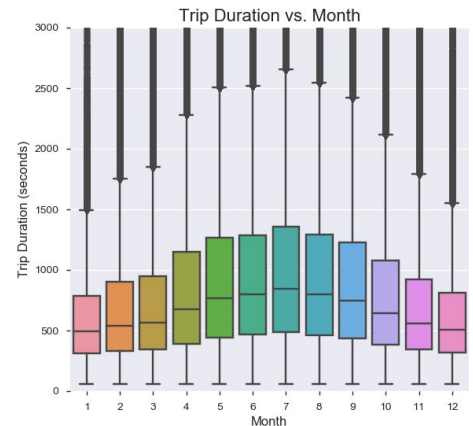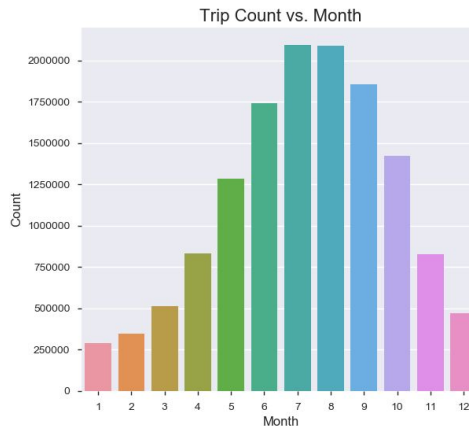
Chicago Bikes: Trends in the data after segmentation

Trends over the years:
- **Increasing** number of trips year over year from 2013 => 2017 for the **subscribers** (annual)
- # of **customer** trips stays **constant**, slight decrease 2015 to 2016
- Duration of the trips stays constant, slight increase for customers
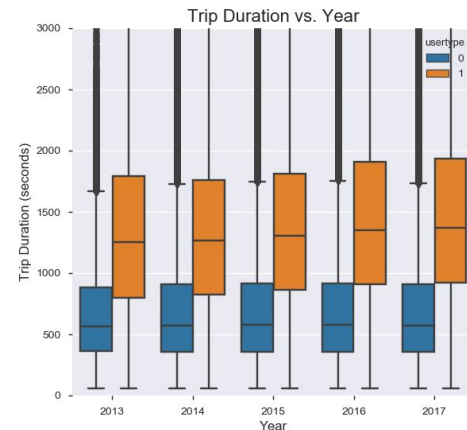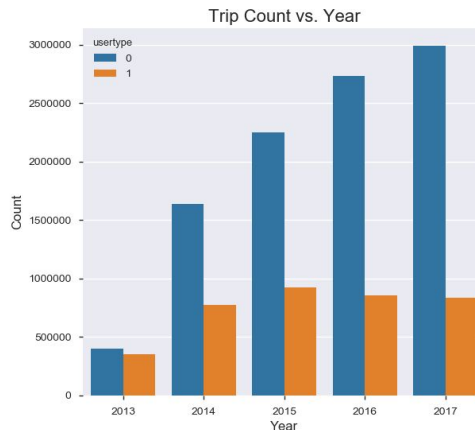
Trends over months:
- Increasing number of trips in the summer months compared to winter
- Very few customer trip in winter
- Slight increase in duration of the trips in the summer months
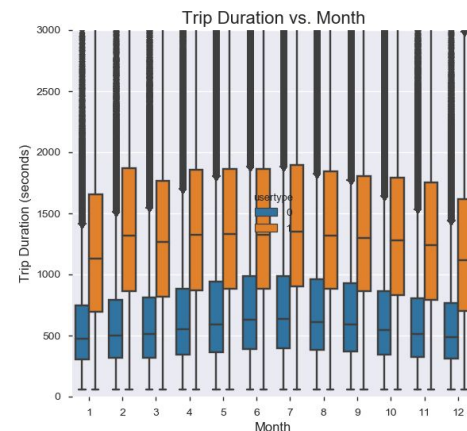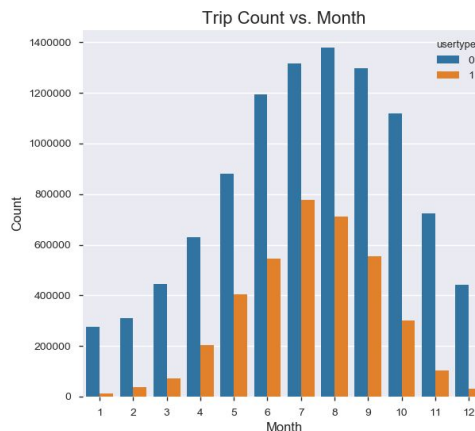
Chicago Bikes: Trends in the data before segmentation
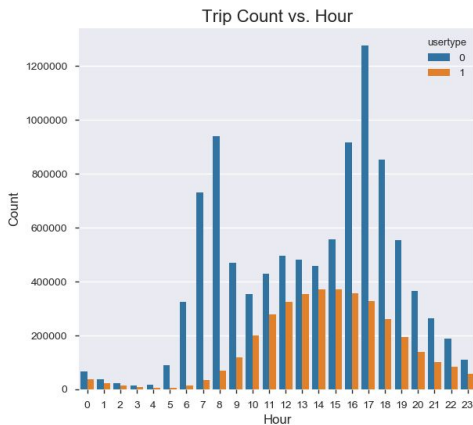
Trends over time:
- During the week, the peak hours are at rush hour
- At the weekend, there is a peak ~2 pm and the trip duration is longer

Heatmap of Trip Counts vs. (Day of Week and Hour)

Trip Count vs. Hour

Trip Duration vs. Hour

Trip Count vs. Day of Week

Trip Duration vs. Day of Week
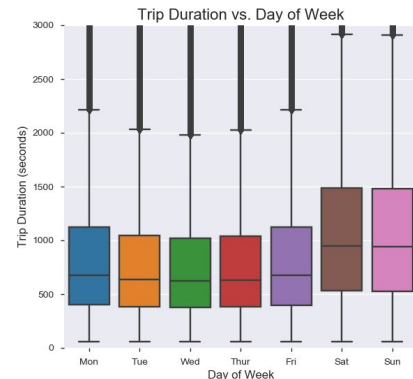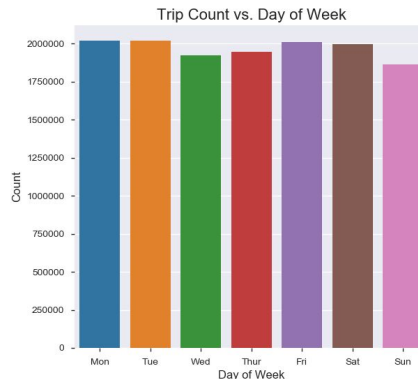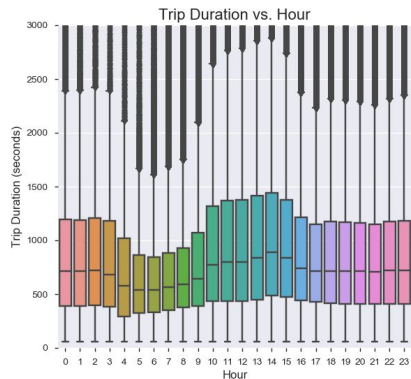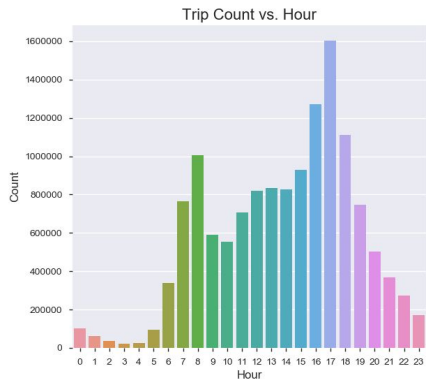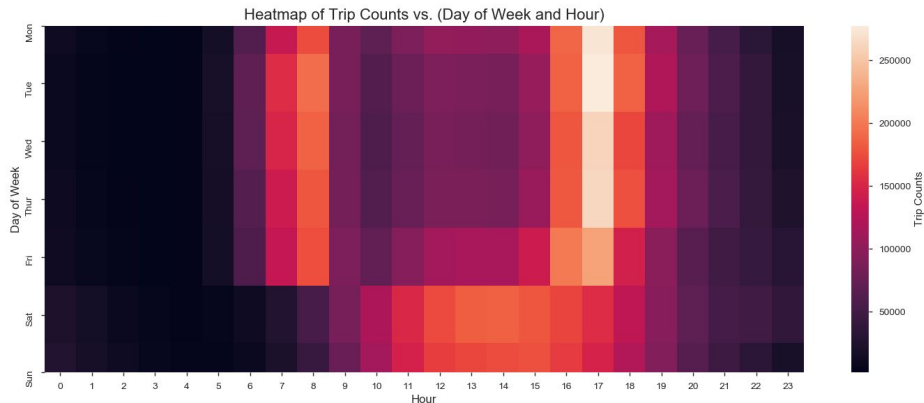
Chicago Bikes: Trends in the data after segmentation

Trip Duration vs Day / Time
- As inferred on previous slides there is a relationship between day of the week and trip duration, but it is more related to the usertypes
- The customers ("daily") users take longer trips in the middle of the day
- At the weekend, there are almost as many daily users as annual users
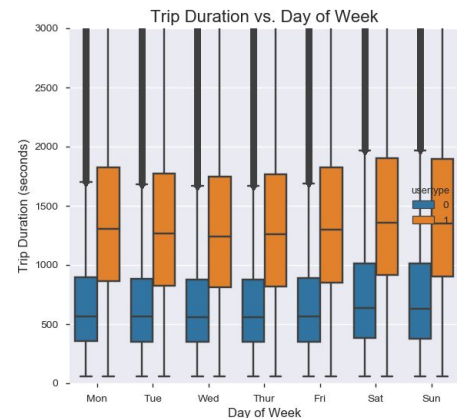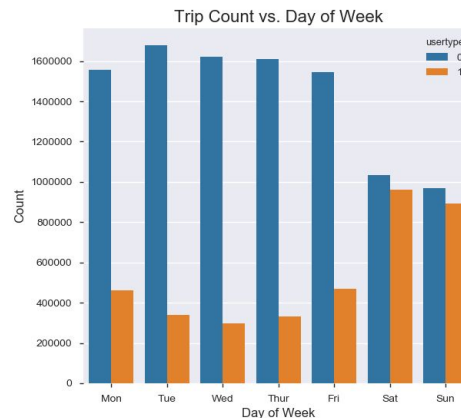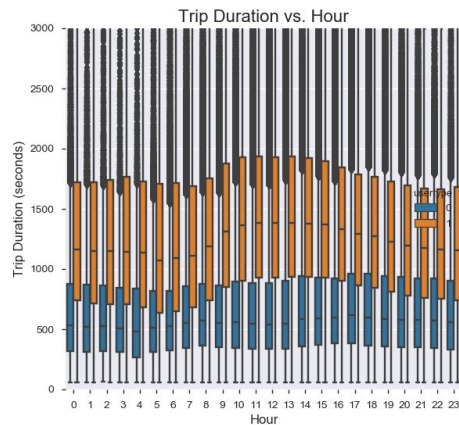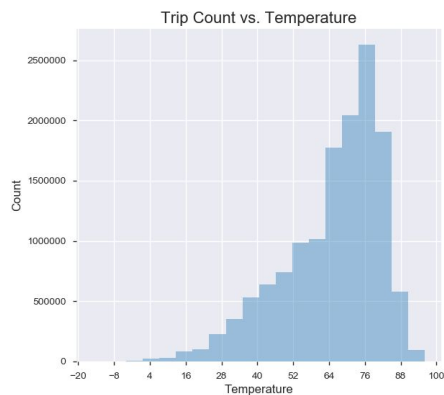
Chicago Bikes: Trends in the data before segmentation

Trends over temperature:
- Increasing number of trips between 64 - 88 °F
- Increasing duration of the trips with increasing temperature

Trends over weather conditions:
- More trips on cloudy days, but this could be because of more days that are cloudy
- Trip duration decreases with snow and sleet
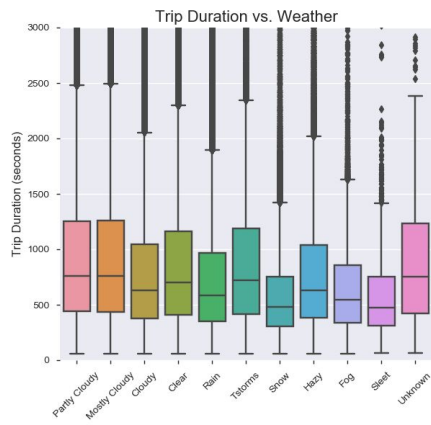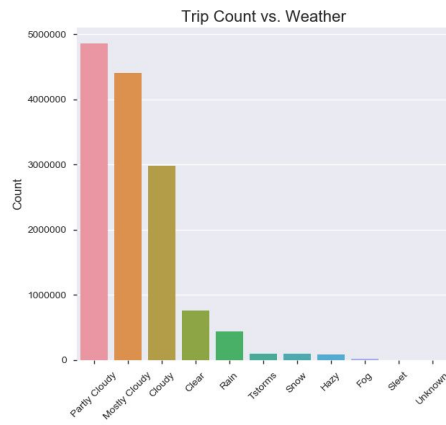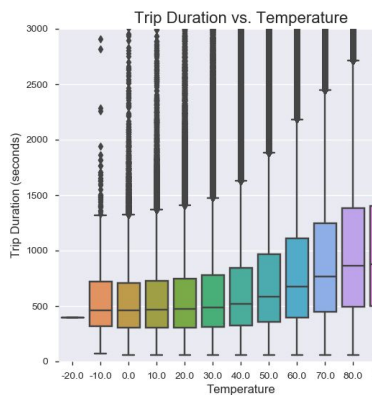
Chicago Bikes: Trends in the data before segmentation

Location dependence:
- There is a strong location dependence - more trips are clustered in the center of the city
- The trip durations are also longer with a start or stop station being in the center
- There is no big difference between start and end station
- There are two "hot pockets" in the center, one is west and one is on the coast
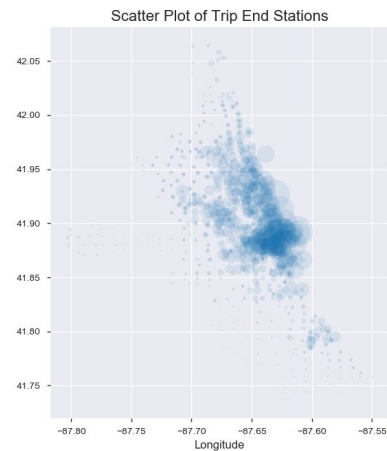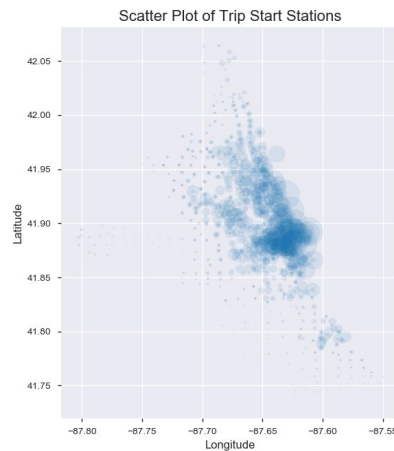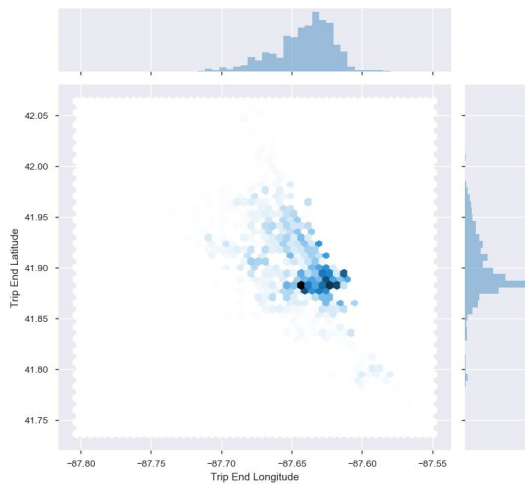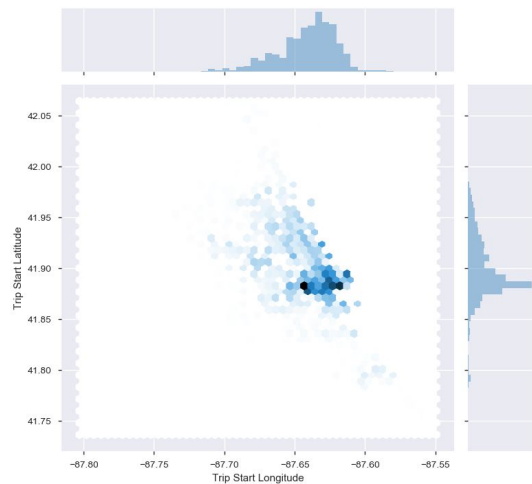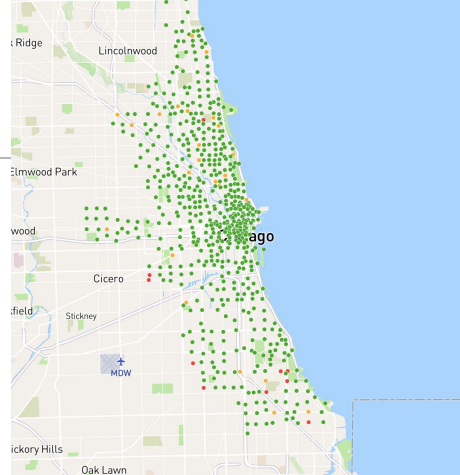
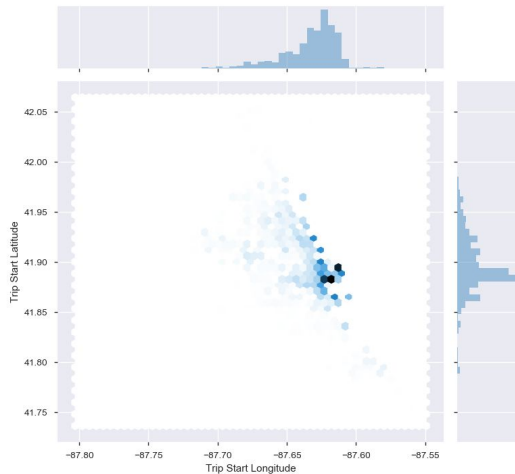# Chicago Bikes: Trends in the data after segmentation

Location dependence:
- Subscribers take more trips from the suburbs west of the city center and go east
- Customers take more trips from the city center and north and south along the coast
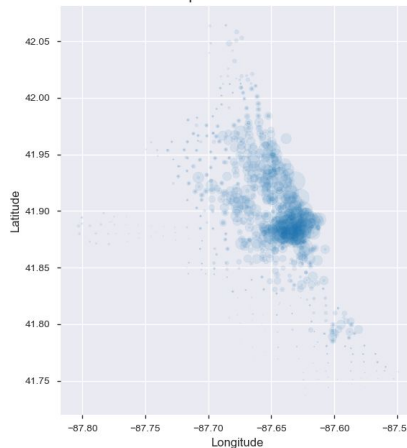- The subscribers take more trips and longer trips in the suburbs

Subscriber - Annual
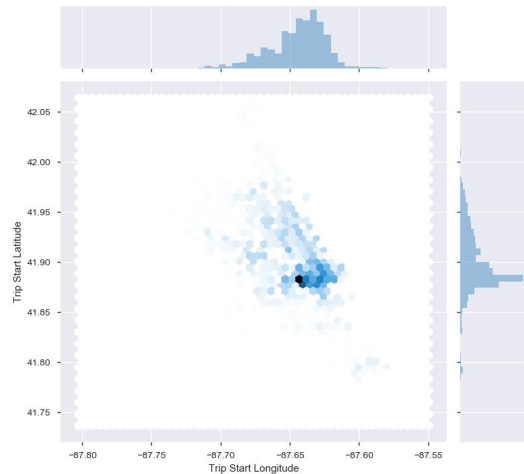
Customer - Daily

Scatter Plot of Trip Start Stations for Subscribers

Scatter Plot of Trip Start Stations for Customers