**Goal**

Analyze CitiBike usage data in New York City to determine if a similar program would be profitable in Des Moines, IA to visualize trends available on Tableau Public. The public workbook can be seen here on Tableau Public.

**Data Cleaning**

Before doing any visualizations, I first used Python and the Pandas library to clean the data. In a Jupyter Notebook, I concatenated monthly data files into one, and then dropped rows where the trip duration was under 90 seconds and the starting and ending stations were the same. This was done to weed out any instances of a faulty bike.

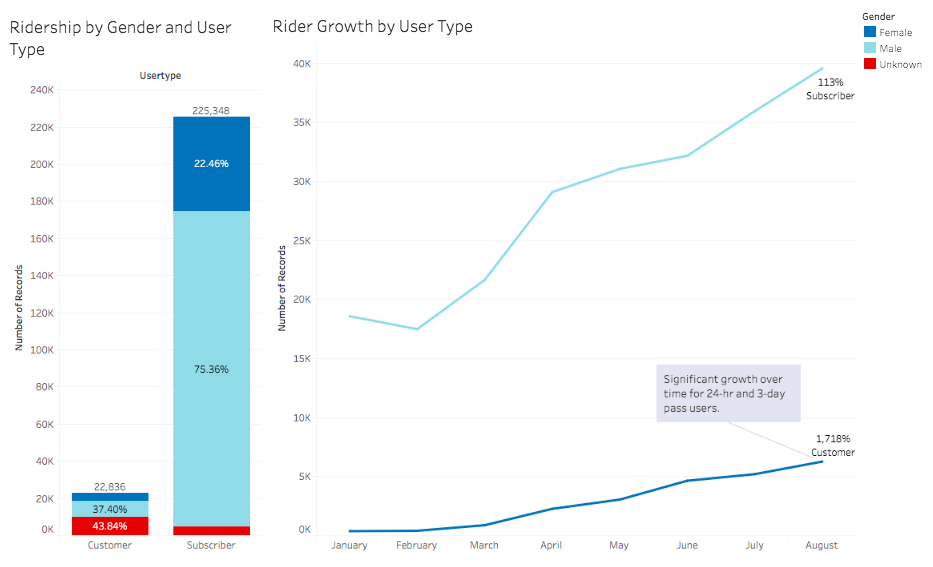
Then, trip duration was converted from seconds to minutes, which is personally easier to understand at first glance, and all trips over 24 hours (1440 minutes) dropped. Under CitiBike’s current pricing model, bikes can be used an unlimited amount of times (under the day passes and subscriptions) but in 30- or 45-minute intervals. Trips longer than that, are charged an extra $2.50 or $4.00 every 15 minutes. Therefore, all bikes that were not docked for over 24-hours were most likely stolen, and not legitimate rides. This was still a conservative assumption, as bike rides up to 24 hours would still cost the user between $232.50 and $376.

Next, I converted the gender column (which was originally numerically coded) into strings using a list comprehension, and added an extra column for ride id to uniquely identify each ride. Then, I calculated the distance for each trip, and the ages of each rider. Finally, I dropped all rows where riders were above the age of 90, assuming these were fake ages (some riders even gave ages over 130), and divided the riders into age bins. Like trip duration, I again made a conservative estimate, as it is unlikely that users in their 80s would be riding bikes in a large city.

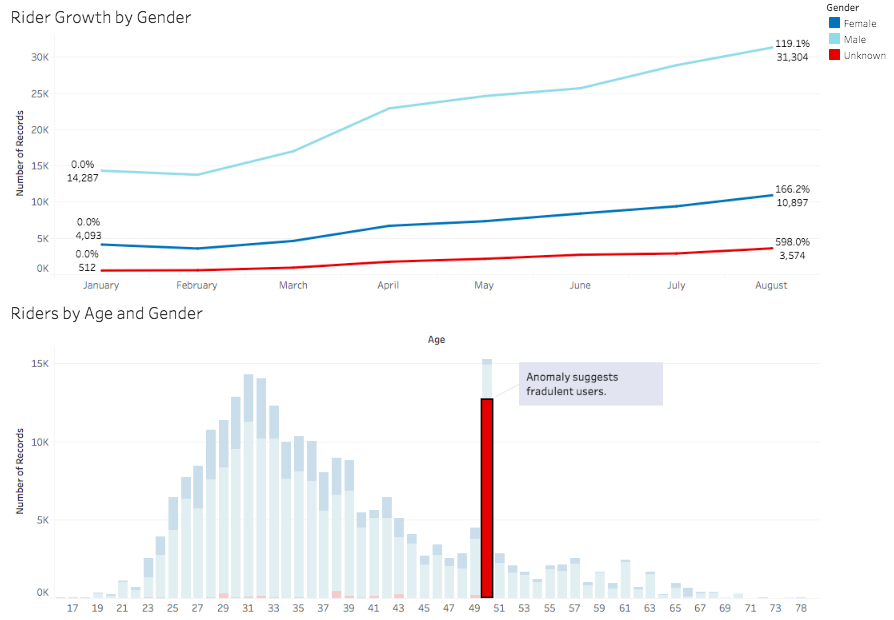
In a separate notebook, I reshaped the cleaned data. Originally, each row represented one ride, with the starting and ending information in one row. Because I wanted to visualization each bike path, I need to plot two sets of coordinates at the same time. So, I divided the data so that the starting information and ending information would be in separate rows. Though visualizations were made with this data, it was not included in the final analysis.

**Analysis**

My analysis focused on user types (customer or subscriber) and gender (male, female, unknown). Though annual subscribers make up over 90% of total riders, customers (1-day and 3-day rentals) have a higher growth rate (1718%) than subscribers (113%). This could signal a growing interest in short-term rentals for users who want to experience the CitiBike product without being tied down for a whole year, or perhaps even a growing tourist market.

[](https://github.com/lorijta92/tableau-citibike/blob/master/Assets/ss-1.png?raw=true)

Despite efforts to filter out fraudulent records from the data set, I still found possible evidence of user fraud. Looking at age and gender, males make up the majority of users, and most users are between the ages of 30-33. However, you’ll see a large spike in the amount of users aged 50, with over 82% of those users identifying as “unknown”. This deviates greatly from the rest of the user demographic trends. Filtering out those users, females show a higher growth rate than males.

[](https://github.com/lorijta92/tableau-citibike/blob/master/Assets/ss-2.png?raw=true)

**Recommendations**

* Continue marketing campaigns to encourage short-term rentals.
* Continue marketing campaigns to encourage female ridership.
* Check all bikes with trip durations under 90 seconds and the same starting and ending stations for any needed repairs.

There is steady ridership growth across user types and gender, suggesting a healthy demand for CitiBikes. Current year-to-date growth rates also suggest a growing market for short-term rentals and female users. Therefore, recommendations would be to focus marketing campaigns towards those users.

Assuming that users use a day pass or annual subscription to its max capacity (48 rides in 24 hours for day pass users, 11,680 rides per year for subscribers), each ride costs $0.25 under a day pass, and $0.01 under an annual subscription. Therefore, short-term users are more profitable per ride, especially when also considering that “time overages” are also charged at a higher rate ($4 per 15 minutes) than subscribers ($2.50 per 15 minutes).

Females also make up 51% of the Jersey City population [source](http://worldpopulationreview.com/us-cities/jersey-city-population/). Yet, only 40% of Jersey City females use CitiBike, leaving over 80,000 potential local users.

**Further Analysis**

For further analysis, I would look at the bike and station ids of all the rows that were dropped from this data set for any trends. Are the same bikes being stolen? Or are bikes being stolen from the same station? Do either these bikes or stations need to be serviced to prevent theft? I would also look into why there are so many “unknown” 50-year old users. Are these just default parameters of the app (gender set to unknown and birth year set to 1969)? Or are there and commonalities in the records of these users?