citibike project

Assignment # 15

BCS - Tableau

Jose Luis Bracho

Citibike Project

Background:

Congratulations on your new job! As the new lead analyst for the New York Citi Bike Program, you are now responsible for overseeing the largest bike sharing program in the United States. In your new role, you will be expected to generate regular reports for city officials looking to publicize and improve the city program.

Since 2013, the Citi Bike Program has implemented a robust infrastructure for collecting data on the program's utilization. Through the team's efforts, each month bike data is collected, organized, and made public on the Citi Bike Data webpage.

However, while the data has been regularly updated, the team has yet to implement a dashboard or sophisticated reporting process. City officials have several questions on the program, so your first task on the job is to build a set of data reports to provide the answers.

Requested:

The requested task is to aggregate the data found in the Citi Bike Trip History Logs and find two unexpected phenomena.

Design 2-5 visualizations for each discovered phenomenon (4-10 total). You may work with a timespan of your choosing. Optionally, you may merge multiple datasets from different periods.

Design a dashboard for each phenomenon. The dashboards should be accompanied with an analysis explaining why the phenomena may be occurring.

City officials would also like to see one of the following visualizations:

* **Basic:** A static map that plots all bike stations with a visual indication of the most popular locations to start and end a journey with zip code data overlaid on top.
* **Advanced:** A dynamic map that shows how each station's popularity changes over time (by month and year). Again, with zip code data overlaid on the map.
* The map should also be accompanied by a write-up unveiling any trends that were noticed during the analysis.

Create a Tableau story that brings together the visualizations, requested maps, and dashboards.

Data Cleaning:

It was used Excel to be formatting and clean the data given the usefulness of Excel and that the data size allows it; otherwise the alternative would be Python.

The first issue that it was seen with the data concerns with the year people were born where customers. All the registers with birth year before 1939 were deleted. It was considered that a person of 80 years of age can perfectly ride a bicycle. The following picture shows an example of the data filtered:

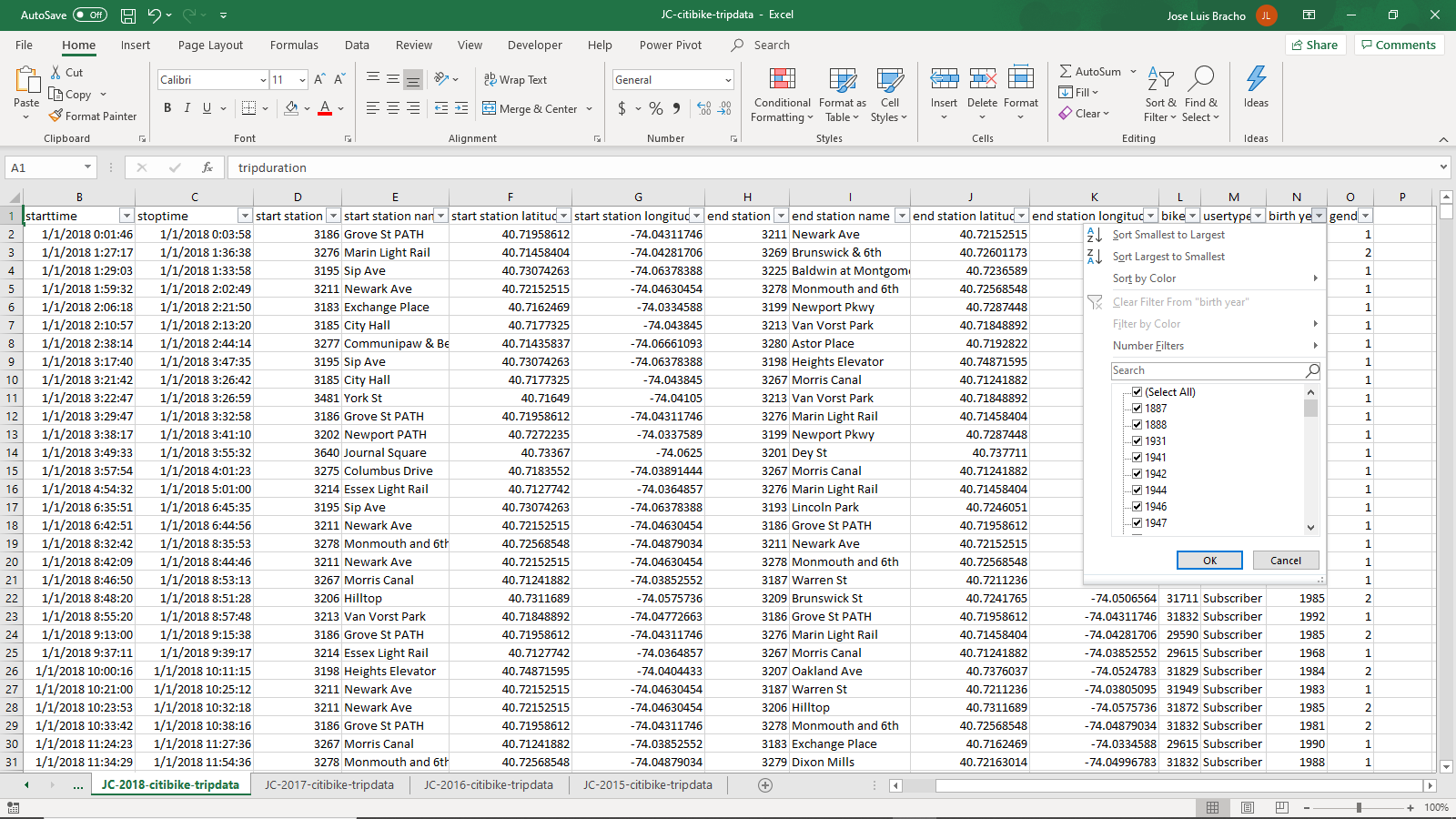


Figure . Data cleaning. Registers with Birth Year before were deleted.

Following the same criteria registers with null or blank information were deleted. See the following picture and table:

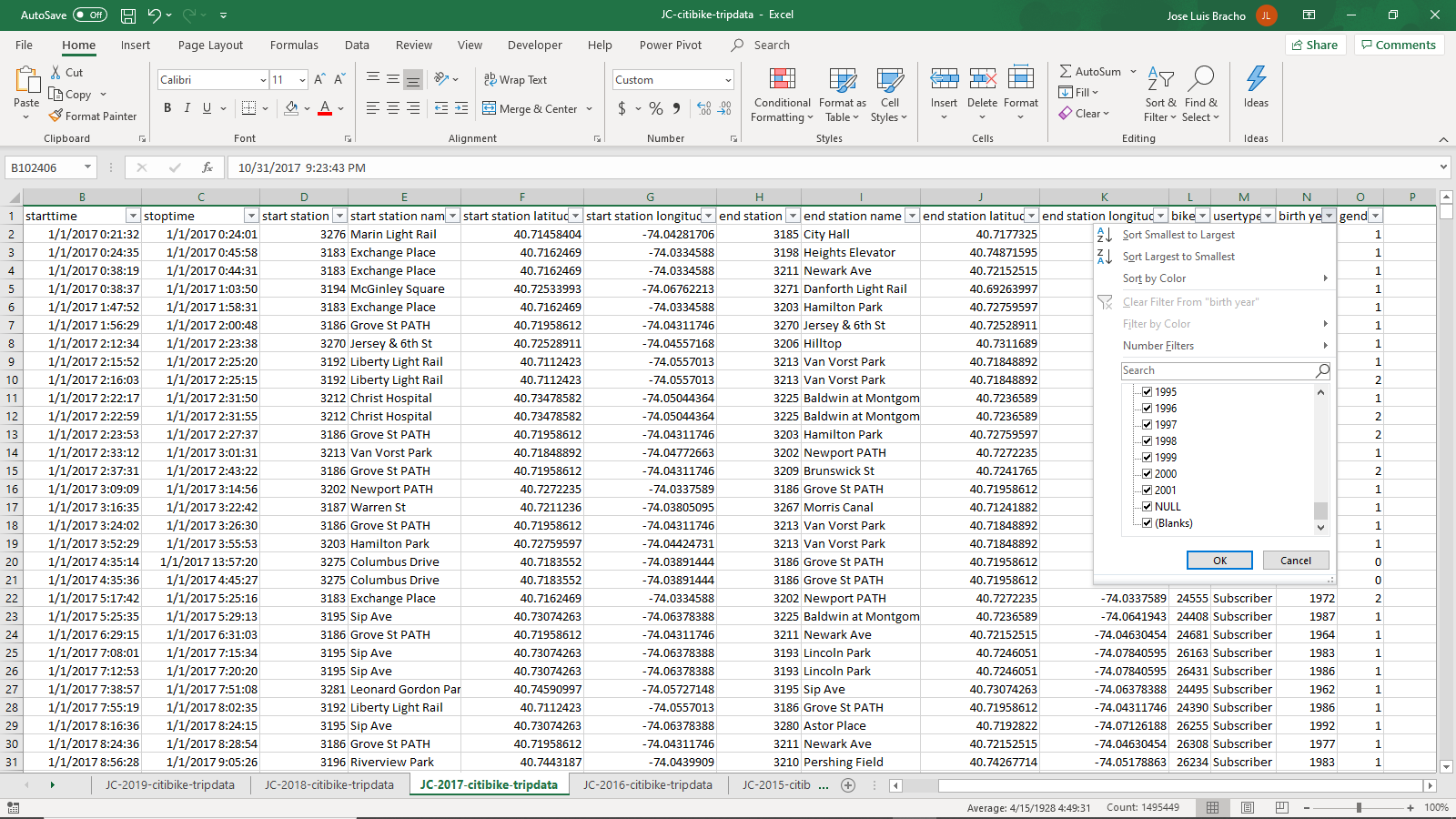


Figure . Data cleaning. Registers with null or blanks were deleted.

Table . Data deleted because null or blanks in Birth Year.

|  |  |  |  |
| --- | --- | --- | --- |
| Year | Null | Blanks | Comments |
| 2019 |  |  |  |
| 2018 |  |  |  |
| 2018 |  |  |  |
| 2017 | 17910 | 1705 | All of them with unknown gender |
| 2016 |  | 18999 | All of them with unknown gender |
| 2015 |  | 5628 | All of them with unknown gender |

Concerning the User Type, data with blanks was deleted.

Table . Data deleted because null or blanks in User Type.

|  |  |  |  |
| --- | --- | --- | --- |
| Year | Null | Blanks | Comments |
| 2019 |  |  |  |
| 2018 |  |  |  |
| 2018 |  |  |  |
| 2017 |  | 1404 |  |
| 2016 |  | 380 |  |
| 2015 |  |  |  |

Because no information in geographical coordinates, data registers was drop.

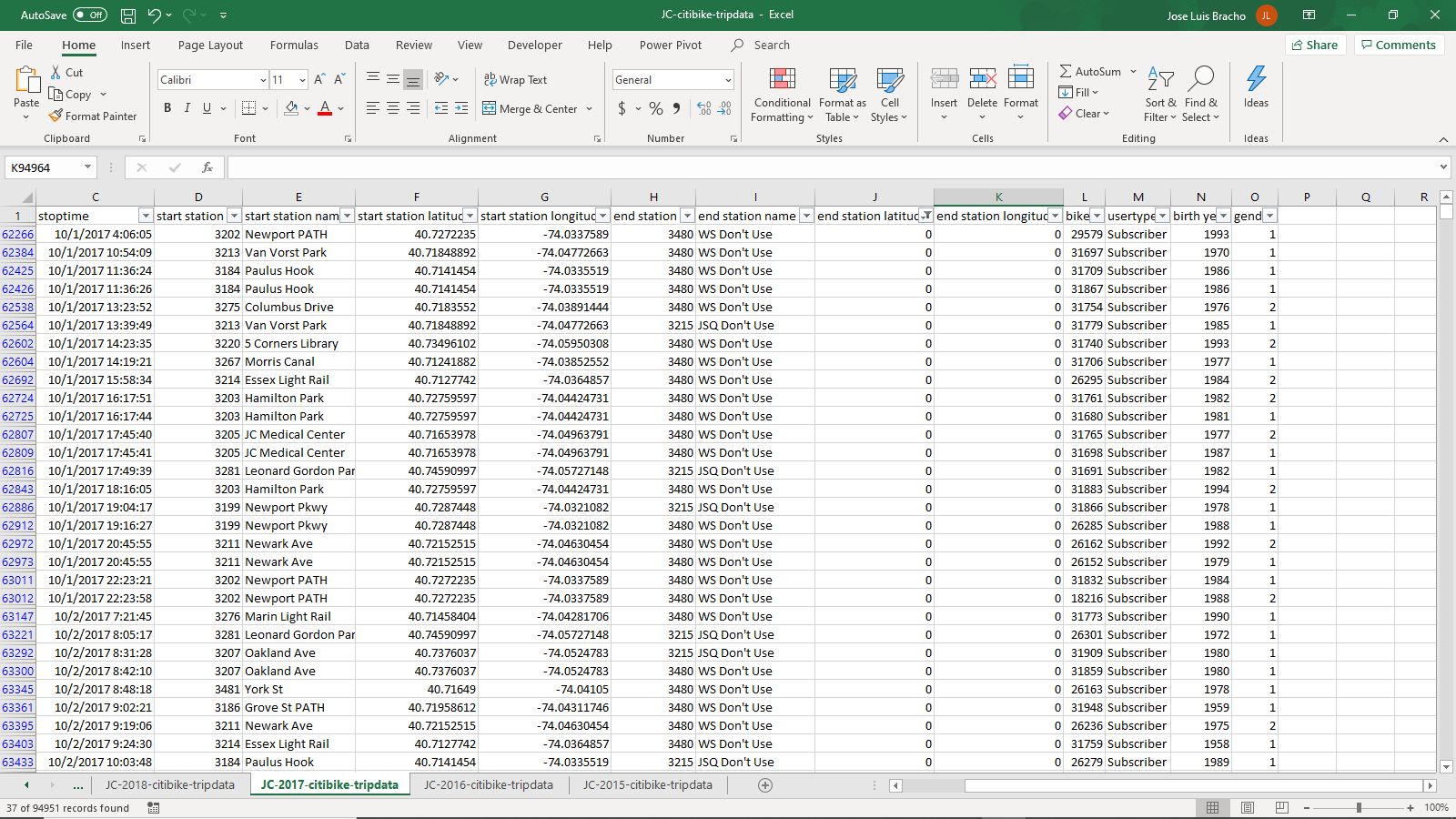


Figure . Data cleaning. Registers with no information in geographical coordinates were deleted.

Table . Data deleted because null or blanks in User Type.

|  |  |  |  |
| --- | --- | --- | --- |
| Year | Null | Blanks | Comments |
| 2019 |  |  |  |
| 2018 |  |  |  |
| 2018 |  |  |  |
| 2017 |  | 692 | JSQ Don't Use, WS Don't Use |
| 2016 |  |  |  |
| 2015 |  |  |  |

Finally, the date format was changed to reflect all the data needed in tableau. The format used was: “mm/dd/YYYY hh:mm:ss”.

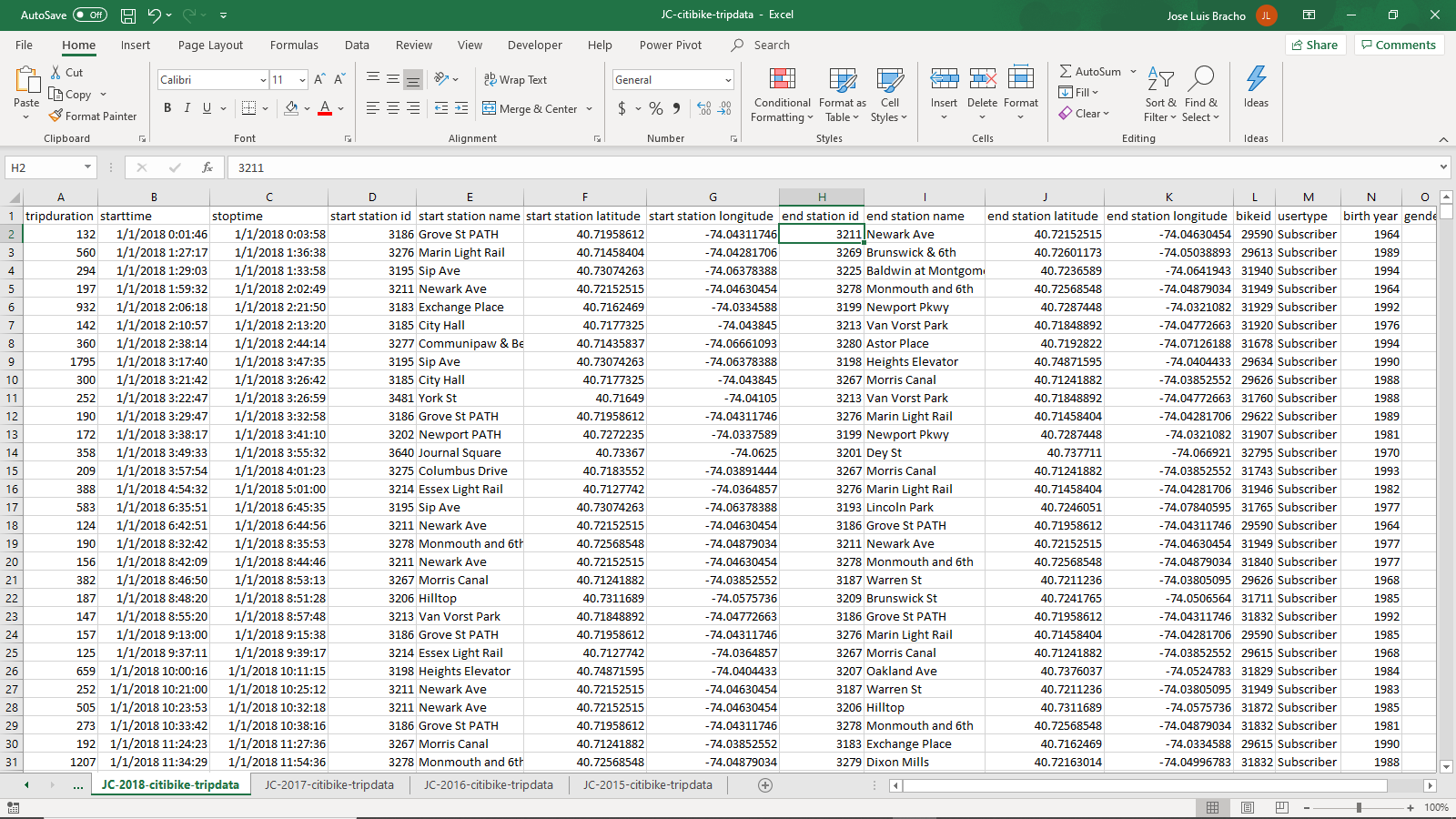


Figure . Data cleaning. Date format was changed.

Data Analysis:

In the City of Jersey were installed 63 stations of the citibike program. A total of valid 1255233 trips were reported and a total of 763271 miles were estimated from the data recorded. During the data exploration were found 3 phenomena in the data: 1) Not all the stations were active, so it was needed to track through the time the working stations; 2) Unexpected distribution between male, female and unknown user was encountered and, 3) The same situation happened to the distribution of user type (this is less unexpected).

The program seems to be a successful one, due to every year the number of subscribers increases, so the miles traveled. On the other hand, it seems that people in a big proportion use the service to go to work and return to home.

In some areas with higher income per capita, the use of the program is greater with a higher number of stations.

Next is a summary of the analysis of each encountered phenomenon:

1. Not all stations were active:

In the following picture it can see that the most active station in Grove St Path, however, the distribution changes in time, so a filter by year and month was applied. This filter allows to observe which station is active at any date. On the other, a filter of user type was applied to the 25 top stations chart as it can be seen the proportion between customer and subscribers in those stations.

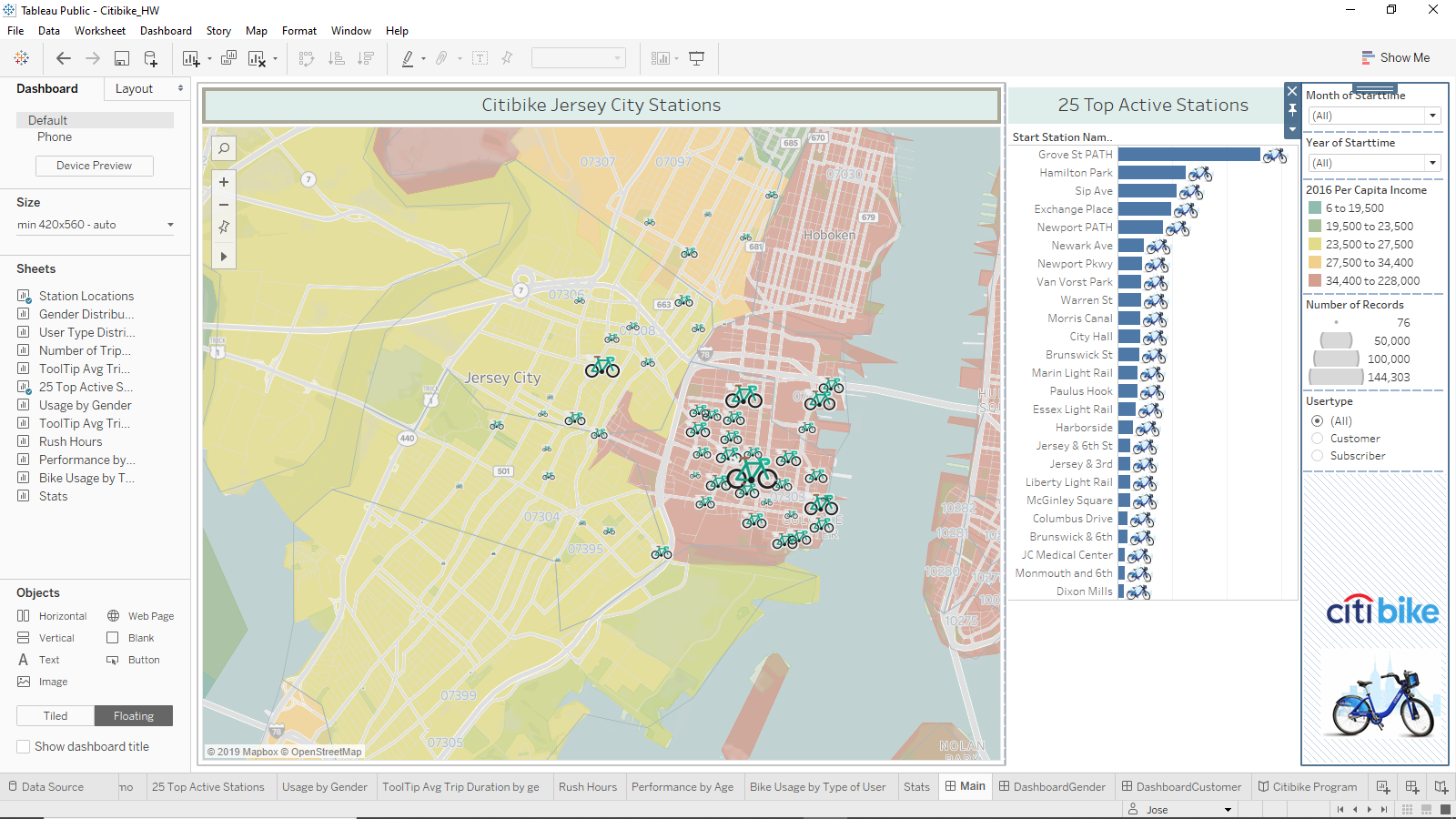


Figure . Map showing the citibike stations in Jersey City.

1. Unexpected distribution in gender:

The second found phenomenon concerns the gender that most uses the program is the male; 74% of the total users. The following dashboard shows the distribution between the different genders. The proportion is higher for men (74.3% of users are men) although the performance is very similar. The performance corresponds to three parameters; the distance traveled, the time using the bike and the miles per hour. Evidently, the distance traveled decreases as age increases. Surprisingly, the minimum elapse time in average in doing by people between 25 – 45 years, so the maximum performance is obtained for people between 25 – 45 years.

Again, it can be selected a desired date to see the number of users by gender/station in the bubbles chart. In the same chart it can be shows the daily distribution of the number of trips.

The reasons why more men use the program are speculative at the time the analysis was made. It could be due to save time and money or as a means for exercising.

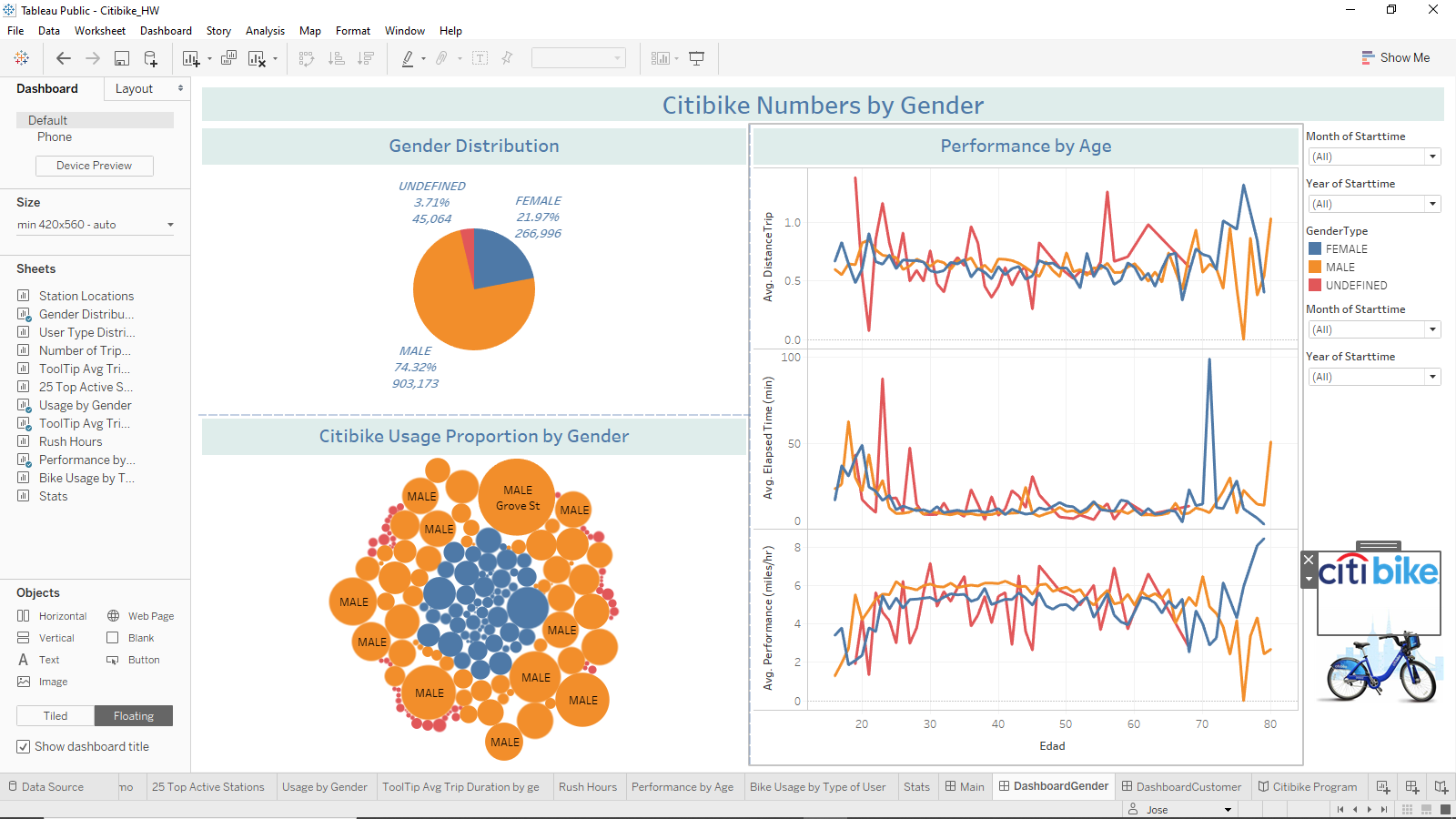


Figure . Dashboard showing the proportion in gender.

1. Unexpected distribution in user type:

The third encountered phenomenon is about the user type proportion. 95% of the users are subscribers and 5% are customers.

In the following dashboard can see the evolution of users’ number through the time. The curve is a cycling one but increasing every year. The cyclicity corresponds to the year seasons, being the most active period the summer one in months of July, August and September.

The last chart shows the minimum proportion of user type customer. This behavior can be due that people uses the program to go to work and for that reason is better to have a subscription to avoid buying the ticket frequently.

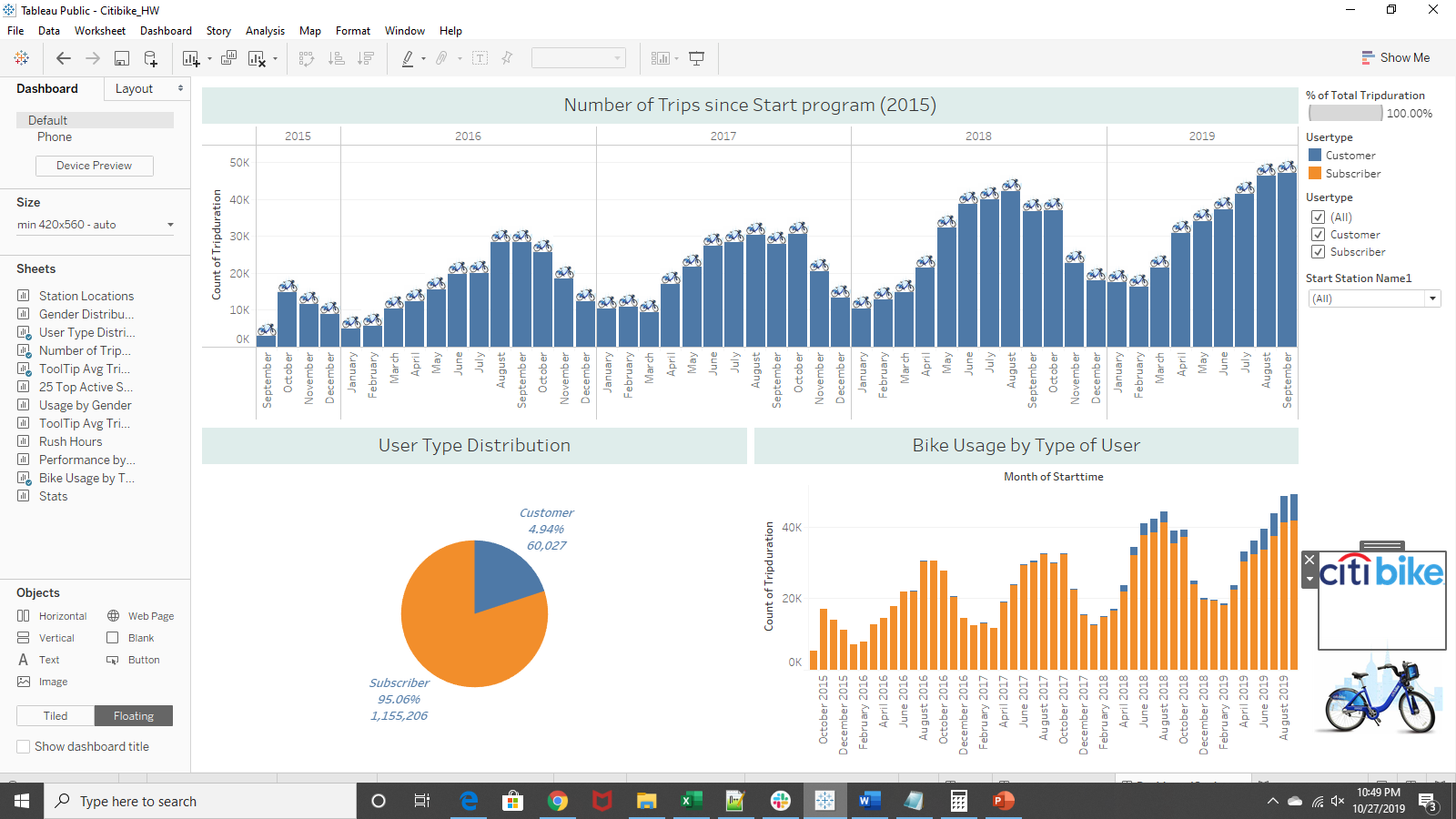


Figure . Dashboard showing the proportion in user type.