**CitiBike data analysis workbook using Tableau**

Introduction:

This project we got from the Citibike website <https://www.citibikenyc.com/system-data> to collect the data of the previous bikes used in NY/NJ areas. The monthly data available across the tristate are in millions and I came across more difficulties in loading the data into Tableau, some how after more hardwork all the csv files are loaded into the Tableau and reports are generated as requested.

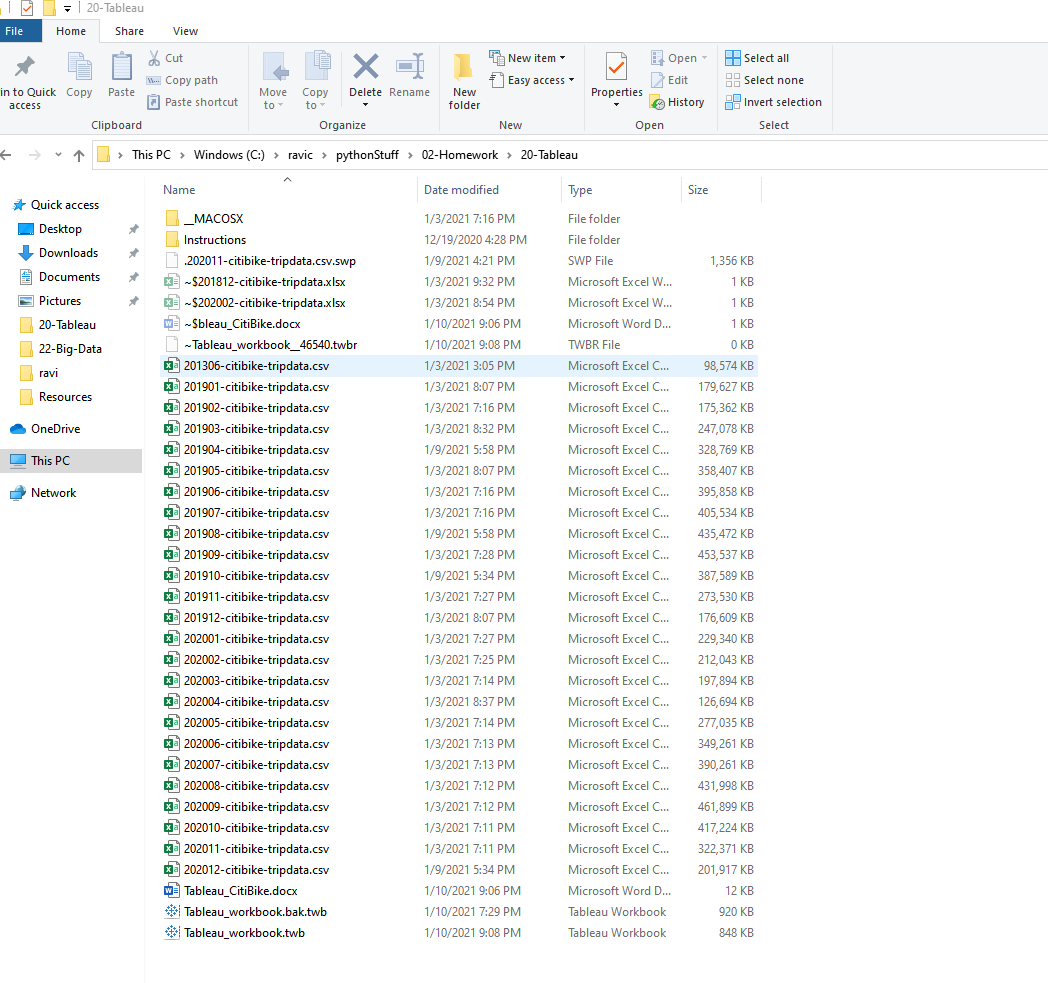
Because of memory issues, we could include much of the coloring features on the report.

If required you could download the csv file and store it locally, since there are more CSV files, could not be uploaded into GIT, <https://s3.amazonaws.com/tripdata/index.html> the files required are 2020 CSV files and 2019 CSV files for all the months.

The data files are used from 2012 to 2019 Jan, almost 18 months of data are used for analysis, some of the report generation took more memory and loading time even for evaluation if they try to load the graphs it will take more time to load.

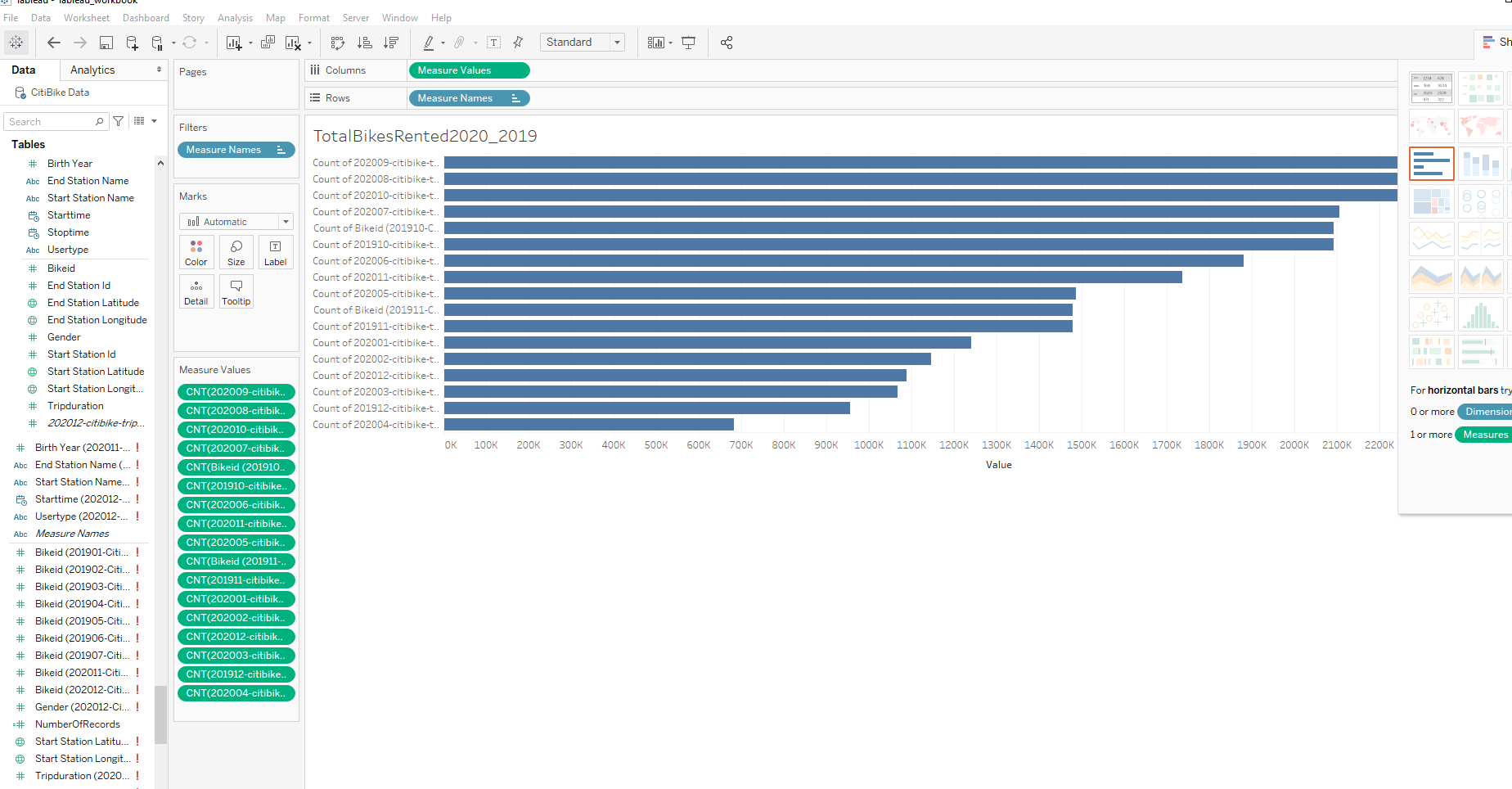
The file “tableau\_workbook.twb” is mapped for data for 2020 and 2011, the csv files are required to download from the site when they load the twb on tableu.

The other file tableau\_workbook\_updated.twb and tableau\_workbook\_updated.twbx mapped only with 4 set of data 202012, 202011, 202010 and 202009 this files are checked into the git hub.

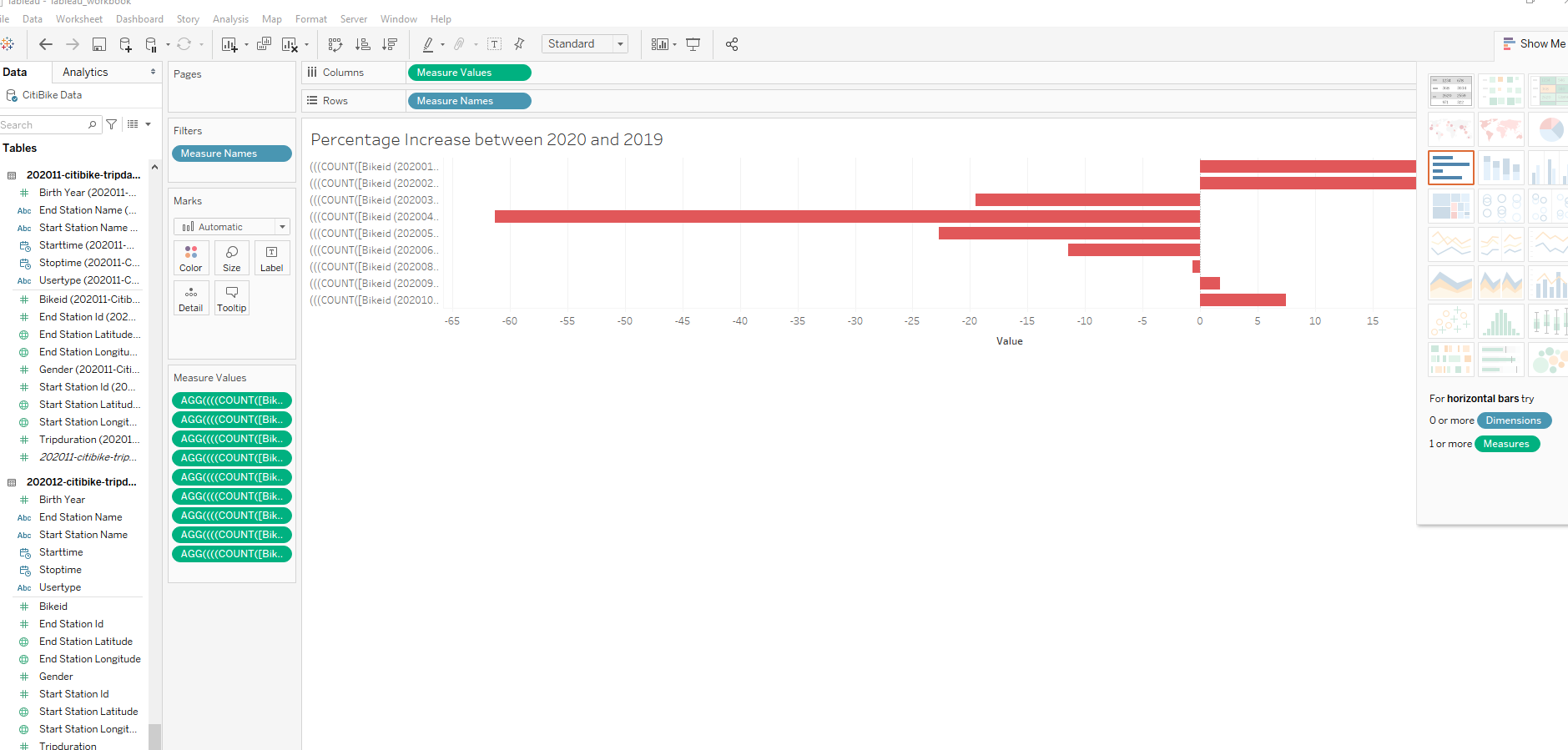


The page load of twb file for tableau will take more time since more csv files are involved, these are the analyzed report

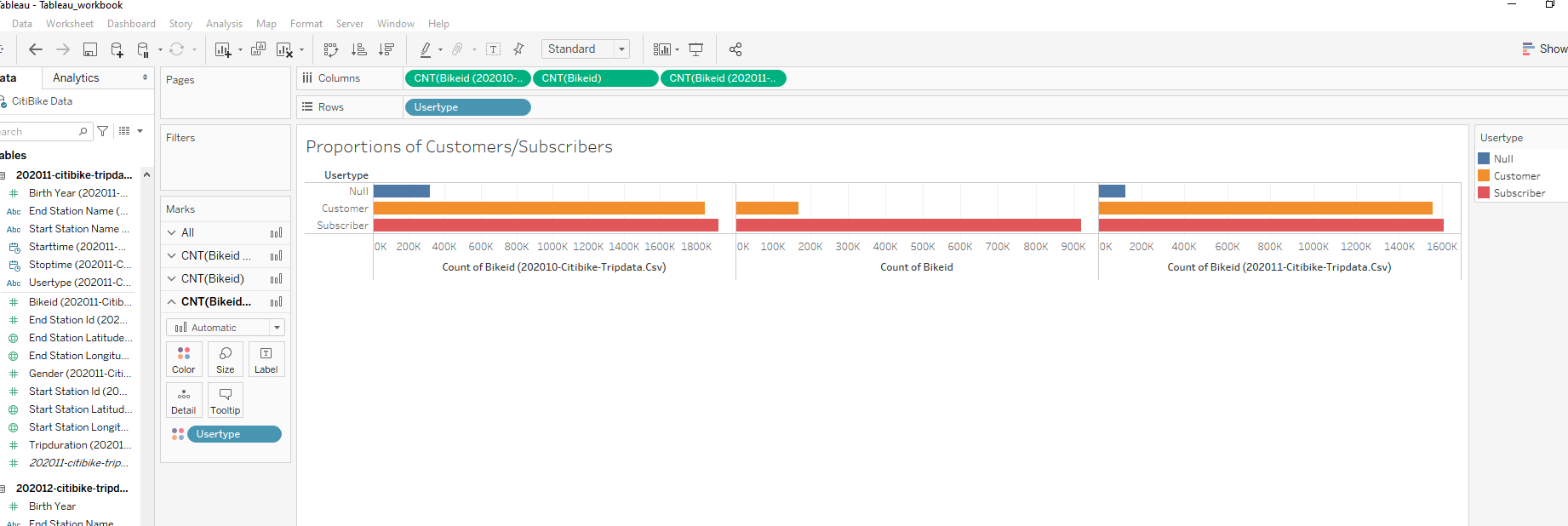
 How many trips have been recorded total during the chosen period?



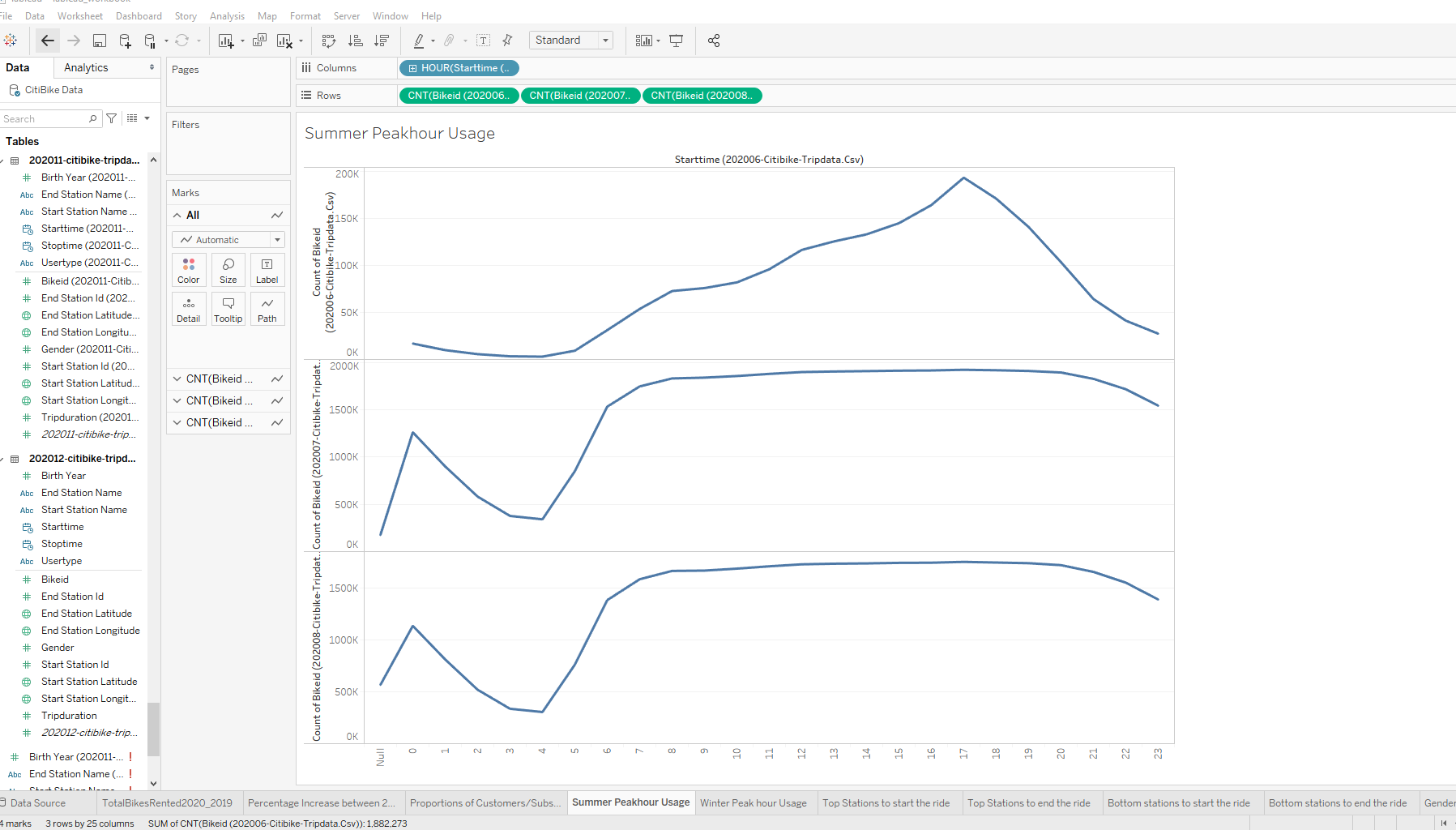
\* By what percentage has total ridership grown?



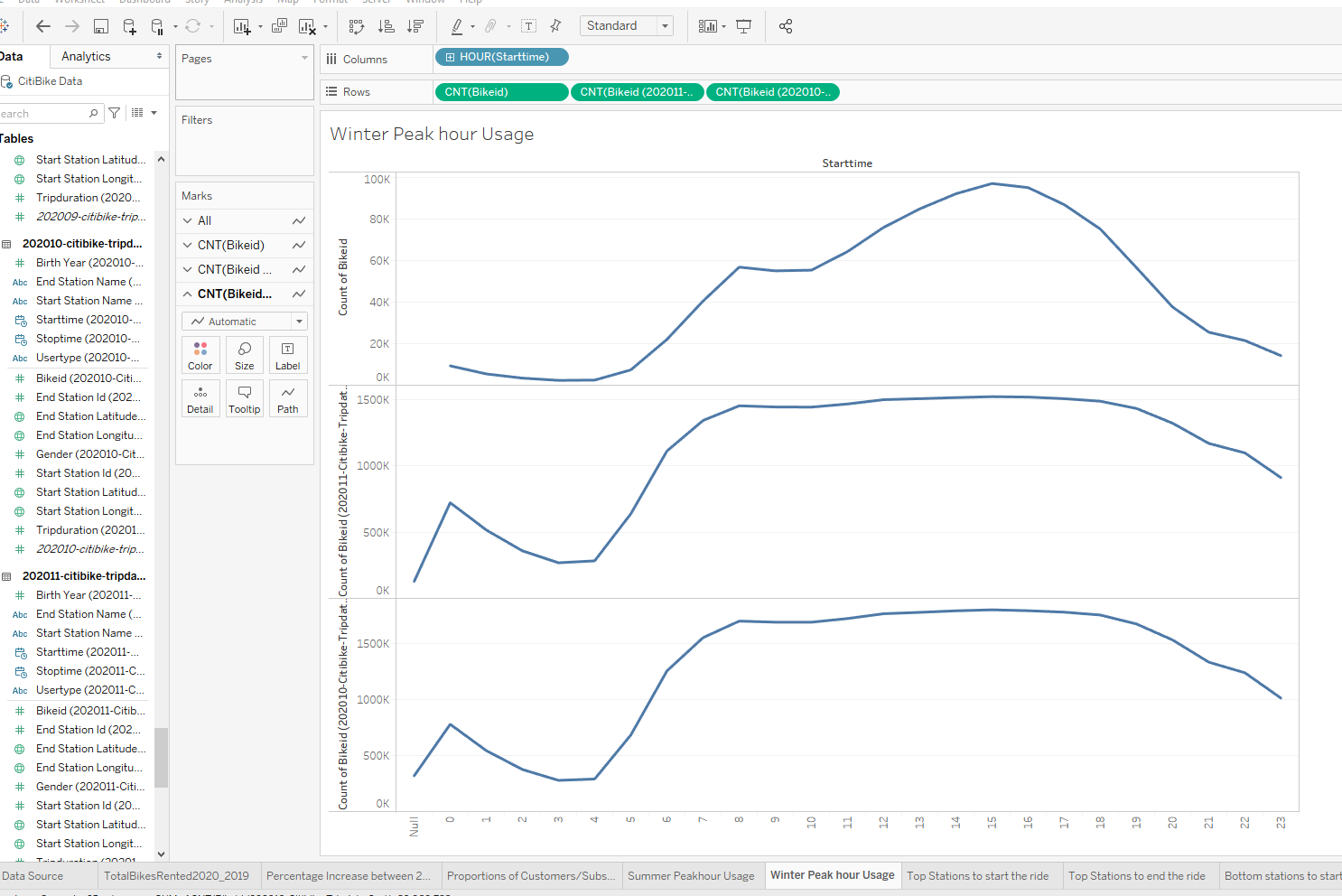
\* How has the proportion of short-term customers and annual subscribers changed?



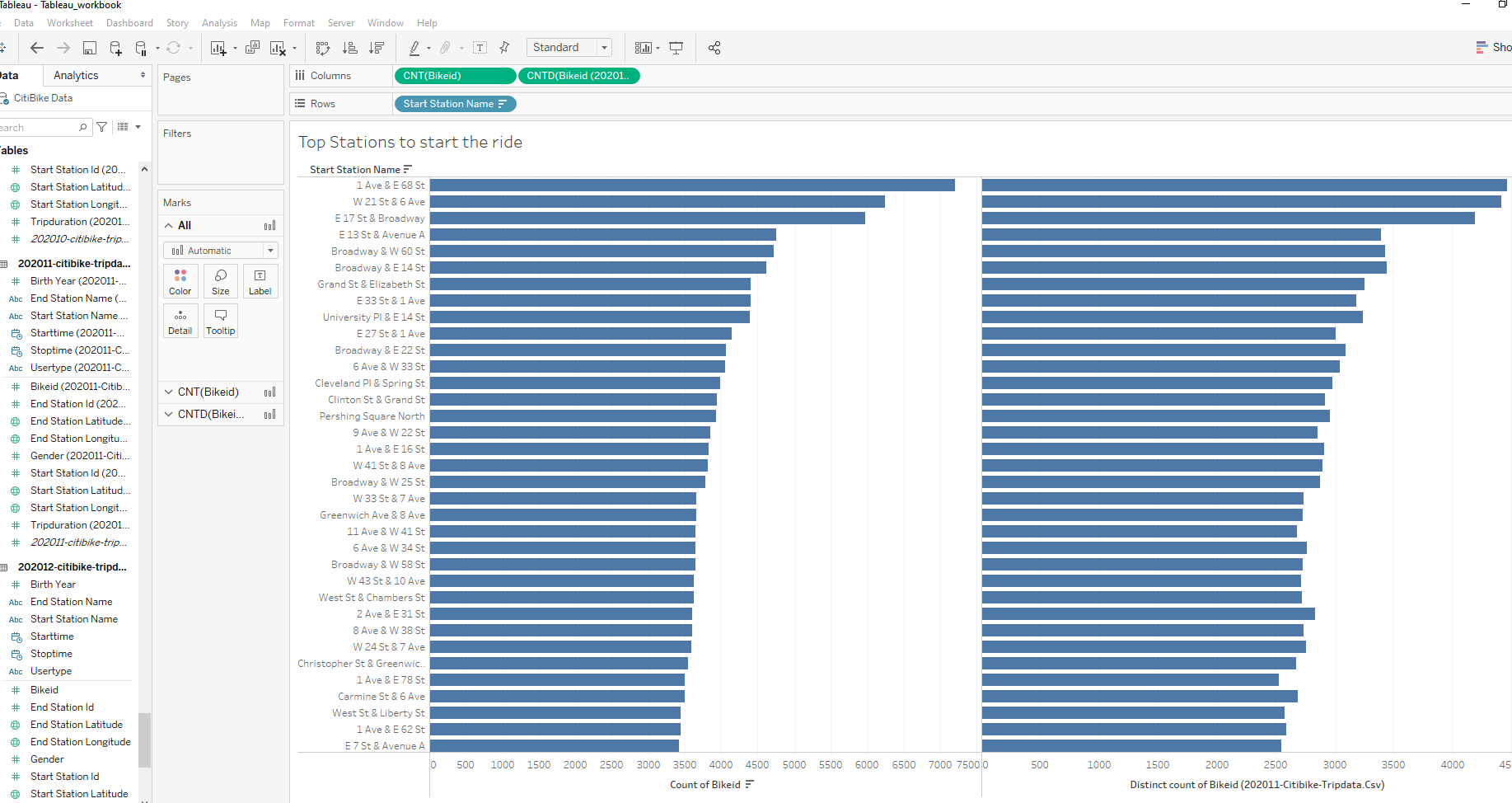
What are the peak hours in which bikes are used during summer months?



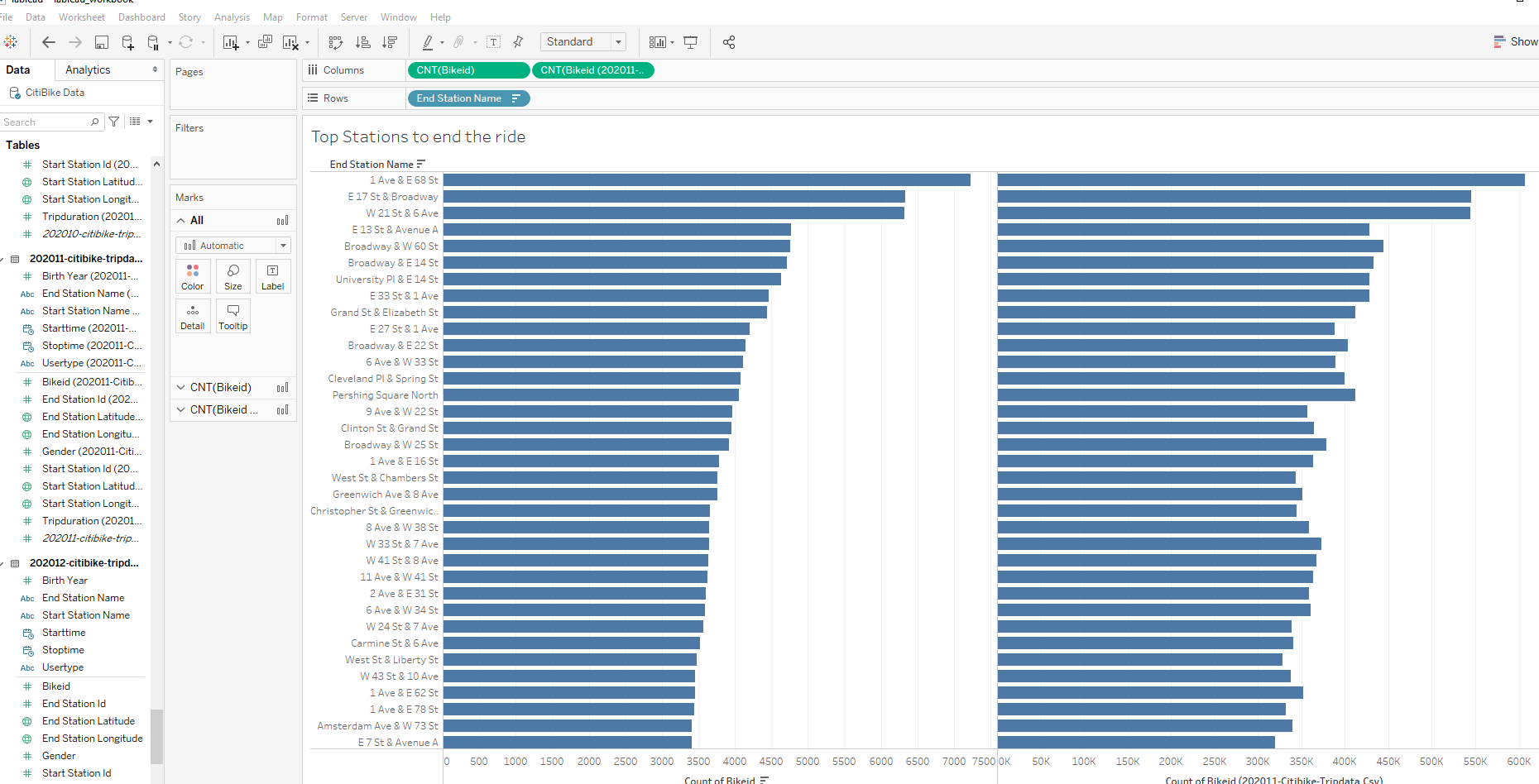
\* What are the peak hours in which bikes are used during winter months?



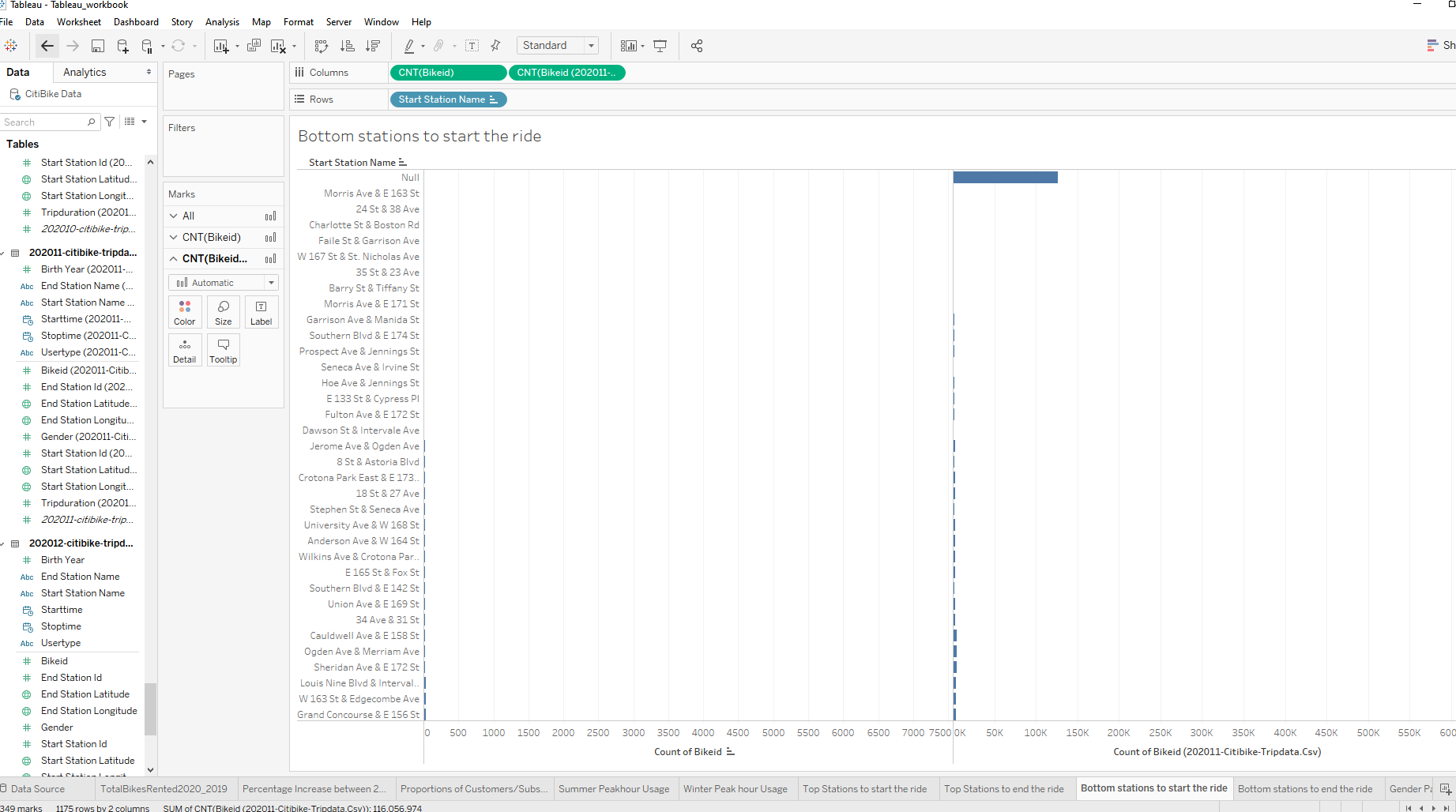
\* Today, what are the top 10 stations in the city for starting a journey? (Based on data, why do you hypothesize these are the top locations?)



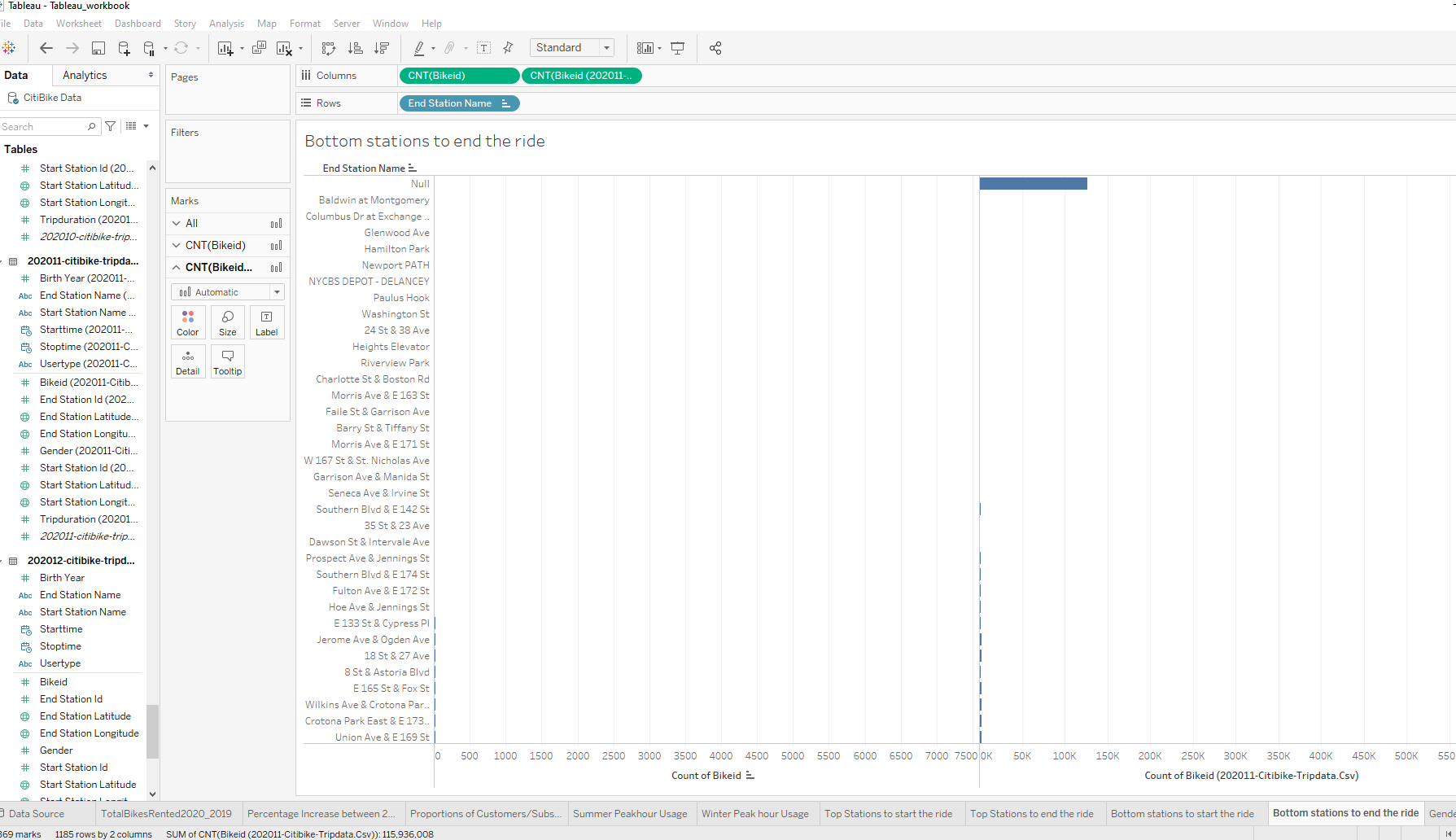
\* Today, what are the top 10 stations in the city for ending a journey? (Based on data, why?)



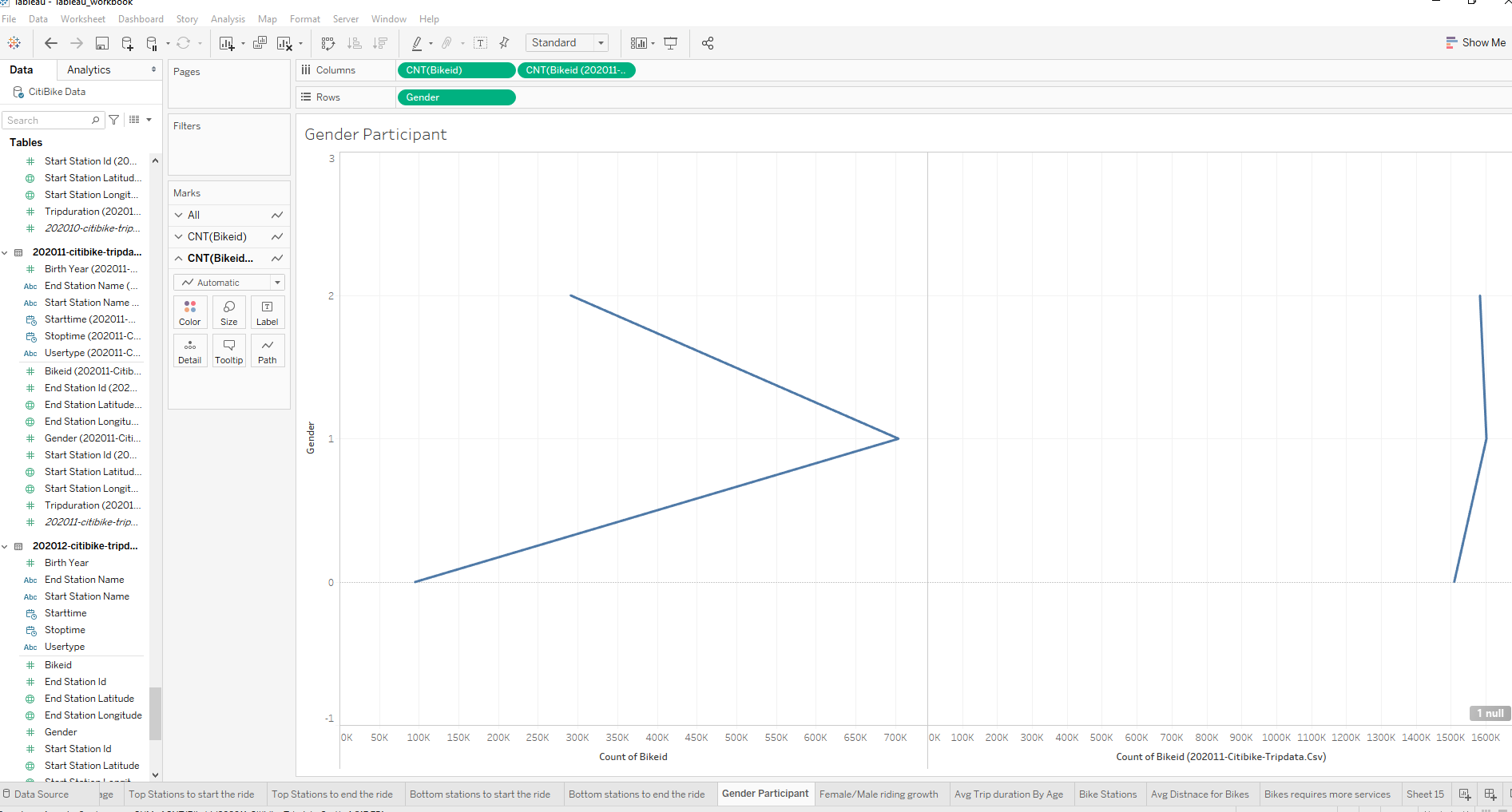
\* Today, what are the bottom 10 stations in the city for starting a journey? (Based on data, why?)



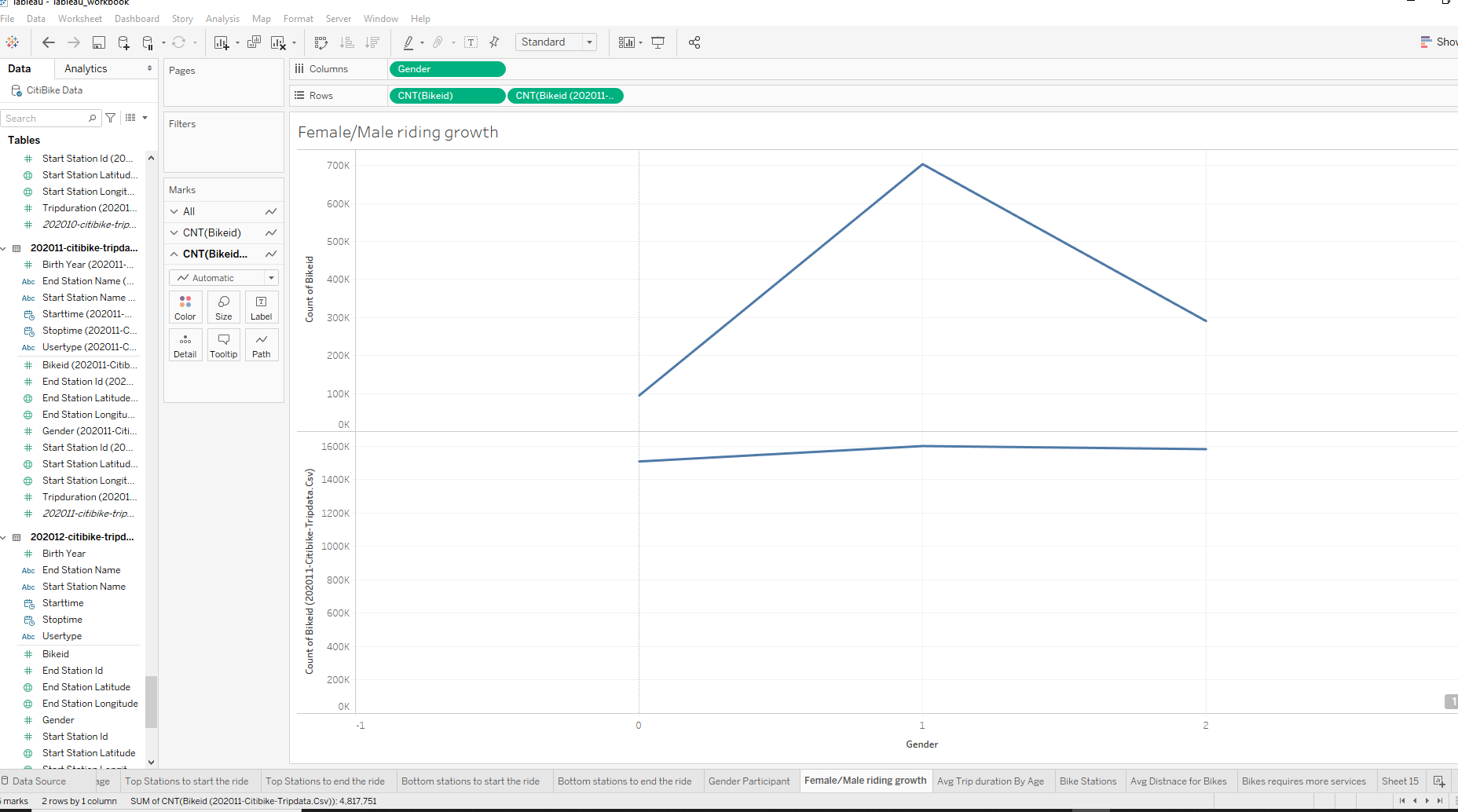
\* Today, what are the bottom 10 stations in the city for ending a journey (Based on data, why?)



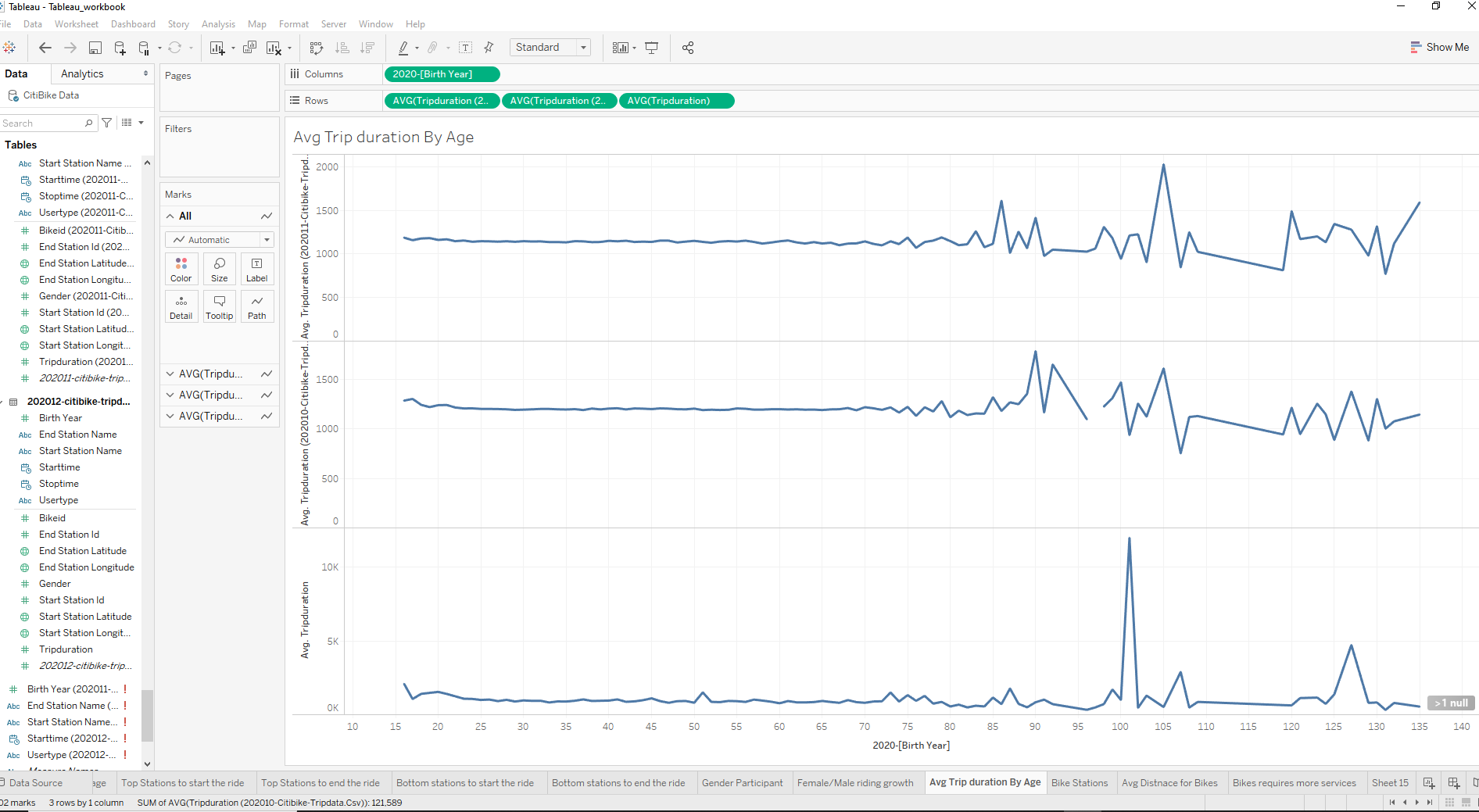
\* Today, what is the gender breakdown of active participants (Male v. Female)?



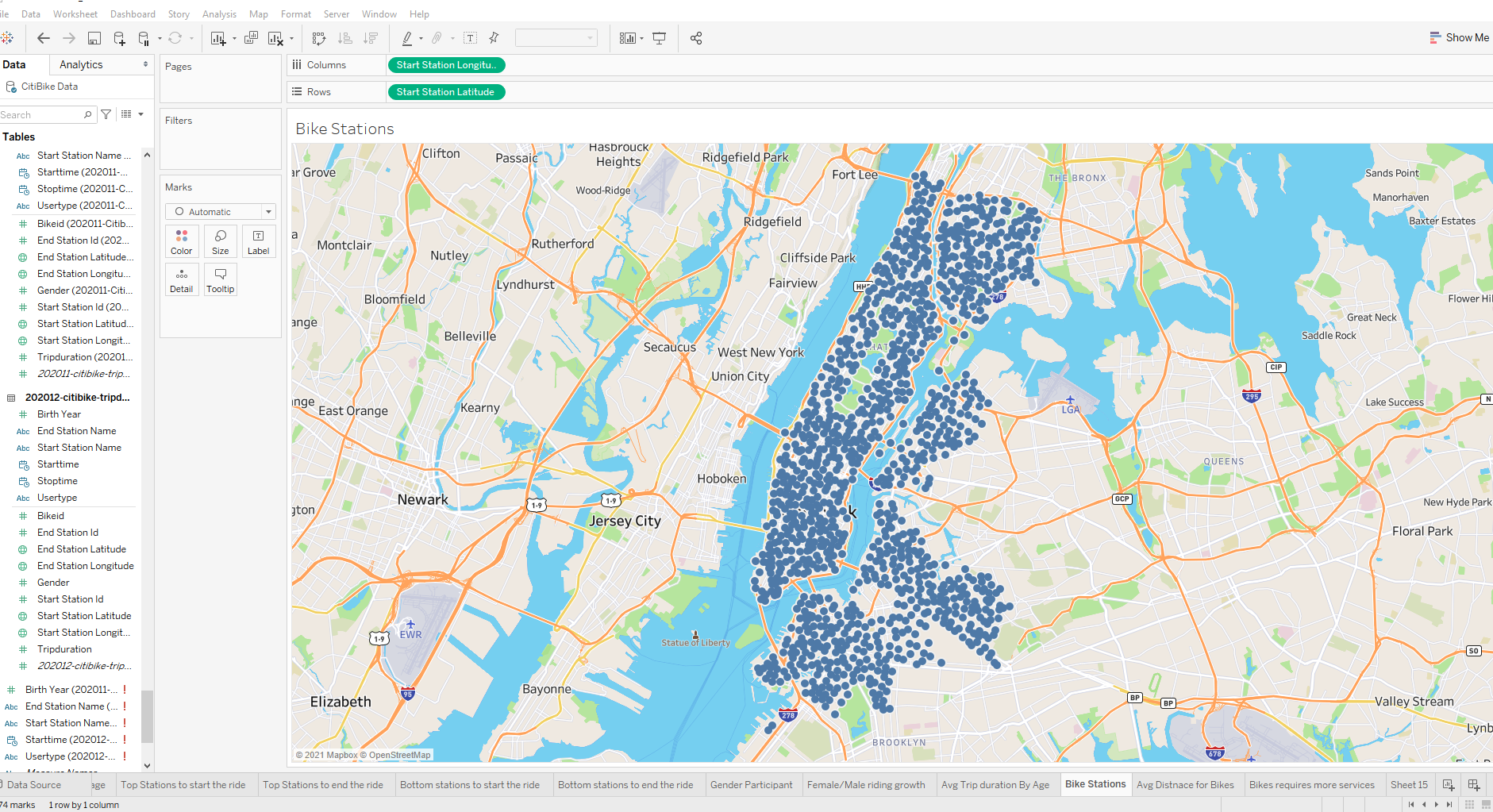
 How effective has gender outreach been in increasing female ridership over the timespan?



\* How does the average trip duration change by age?



Station Maps:



\* What is the average distance in miles that a bike is ridden?

The distance is calculated between two coordinates using

function distance(lat1, lon1, lat2, lon2) {

var p = 0.017453292519943295; // Math.PI / 180

var c = Math.cos;

var a = 0.5 - c((lat2 - lat1) \* p)/2 +

c(lat1 \* p) \* c(lat2 \* p) \*

(1 - c((lon2 - lon1) \* p))/2;

return 12742 \* Math.asin(Math.sqrt(a)); // 2 \* R; R = 6371 km

}

But this involves more calculations on Tableau and on existing resource constraint on the cpu/memory the general assumption of 1 miles for 600 seconds (trip duration) is taken and total distance travelled by the bike is taken



\* Which bikes (by ID) are most likely due for repair or inspection in the timespan?



\* How variable is the utilization by bike ID?m,

The total distance travelled as trip duration is taken as variable of utilization for the bikes, the bike ID that travelled more than the average limits are taken for the bikes for more servicing.

Attached the csv file … “Tableau\_workbook.twb” file in the Git