In this project, I am going to work at Citi Bike New York City’s official bike-sharing system.

**‘Citi Bike’** is the largest bike-share program in the United States, with 20,000 bikes and over 1,300 pick-up stations across Manhattan, Brooklyn, Queens, the Bronx, and Jersey City. As stated on their website, the service was designed for quick trips with convenience in mind, offering a fun and affordable way to get around town. Users can sign up for annual membership, or buy a short-term pass through the Citi Bike app. Once they’ve joined, they simply locate a nearby bike, ride around as they please, and return it to a nearby station once they’re done.

I am going to analyse the data to find out a few queries. These are the following queries:

1. What are the most popular pick-up locations across the city for NY Citi Bike rental?
2. How does the average trip duration vary across different age groups?
3. Which age group rents the most bikes?
4. How does bike rental vary across the two user groups (one-time users vs long-term subscribers) on different days of the week?

Let’s get started. I have collected the data from an online school named ‘Career Foundry’. First I focus on the raw data which is in an excel spreadsheet. There is a total of 20400 rows and 17 columns. This data set contains a lot of informations. When they pick-up, where they return, the Bike ID, station ID, station name, season, duration, user type, age, birthdate etc. I have try to show you sampe data figure. Please go to the next page to see the data.

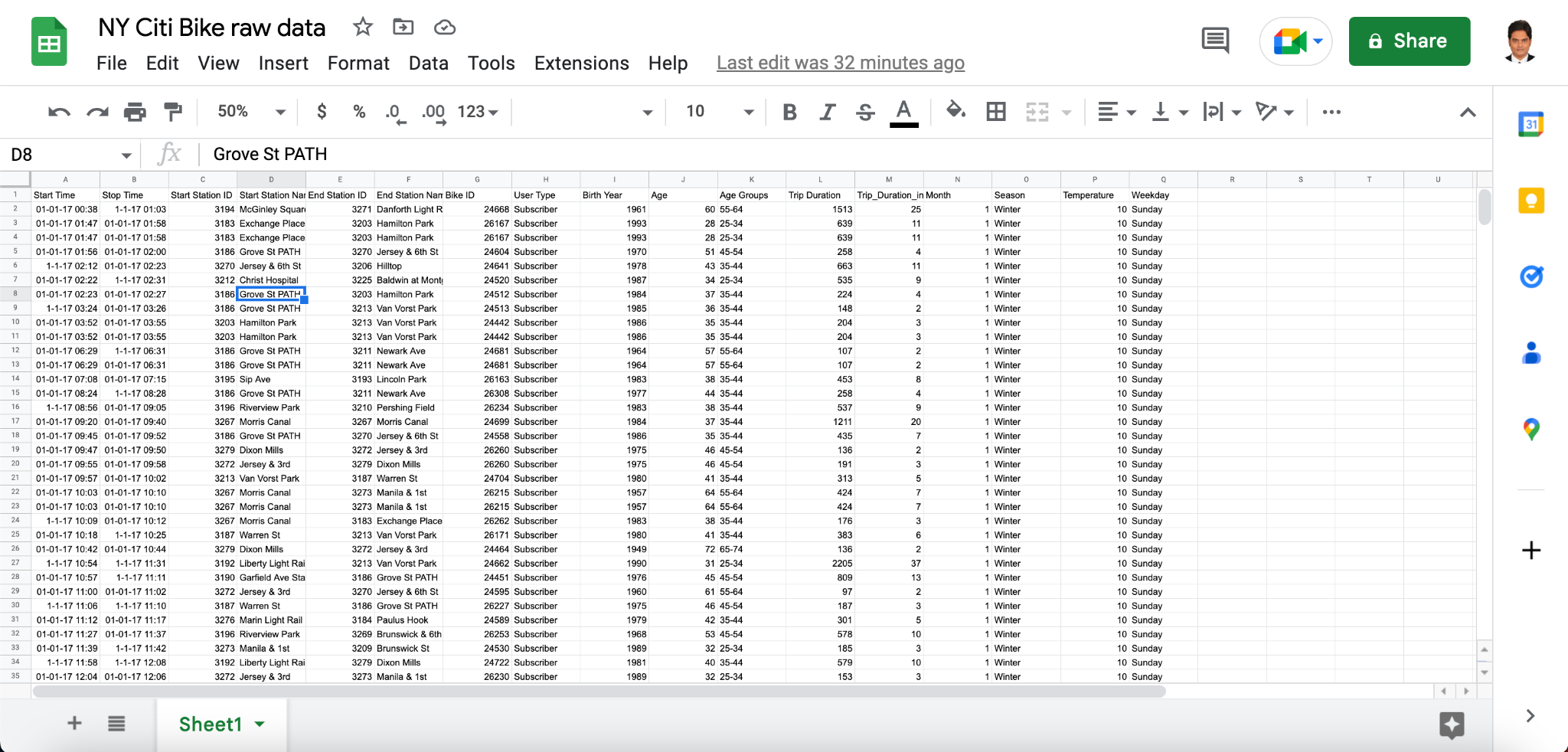


Figure 1

**Data cleaning**

Before I start analysing I need to clear the data. Because when I collect data from different sources, there are few dirty data in the data set that can be disrupted my analysis. That’s why firstly I need to clean the unnecessary data from the data set. To do that I have focused mainly on two steps.

1. **Identifying and removing duplicates**
2. **Identifying and handling missing data points**

**Identifying and removing duplicates**

Let’s get started to explain how I can do that.

**Step 1:** First of all, I am going to find out all the duplicates data and remove all the duplicates data from my dataset. Select the entire dataset by clicking on the gray rectangle to the left of column A and above row 1.

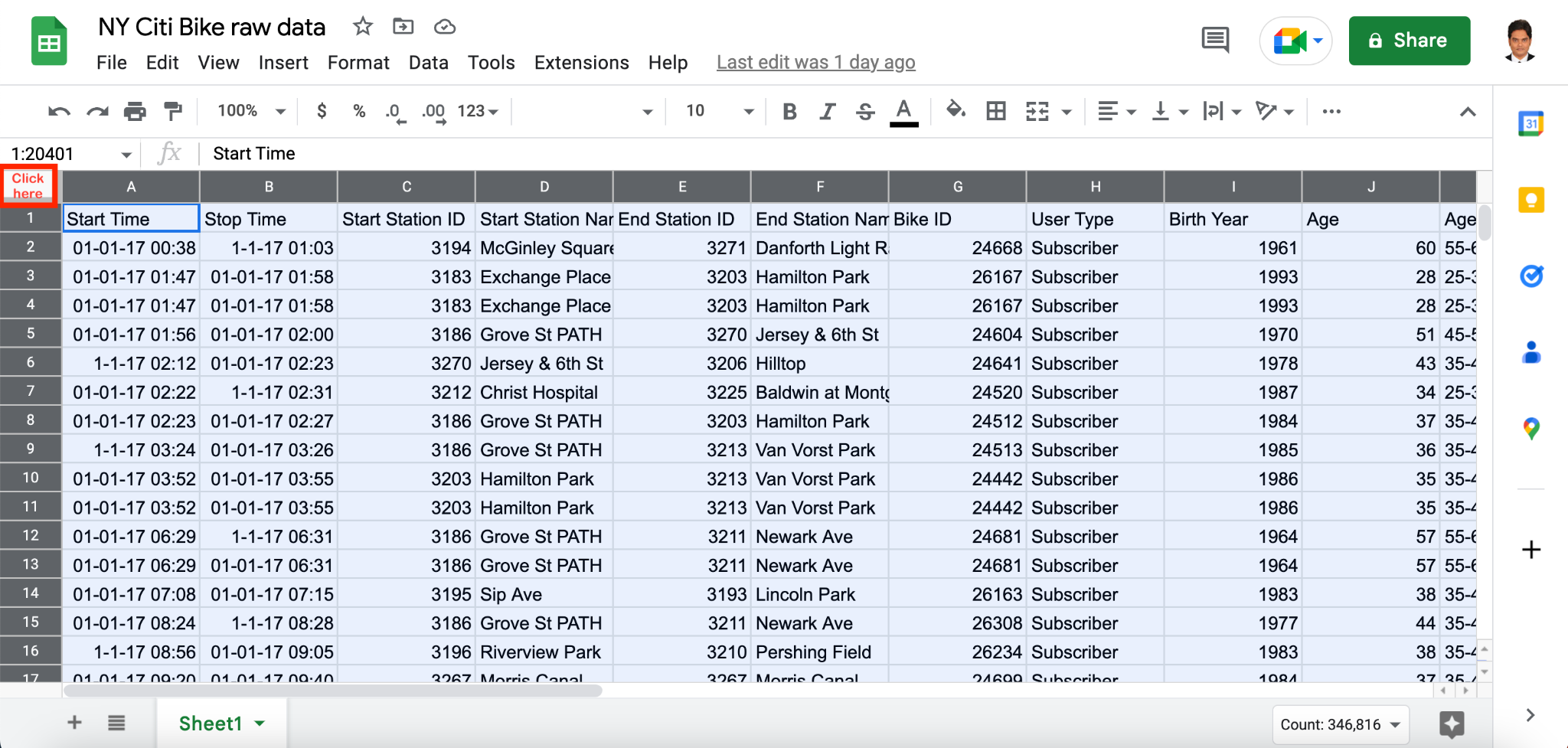


Figure 2

My entire dataset now be highlighted blue.

**Step 2:** Now with my highlighted dataset, click on “Data” in the toolbar and select “Remove duplicates” from the dropdown menu.

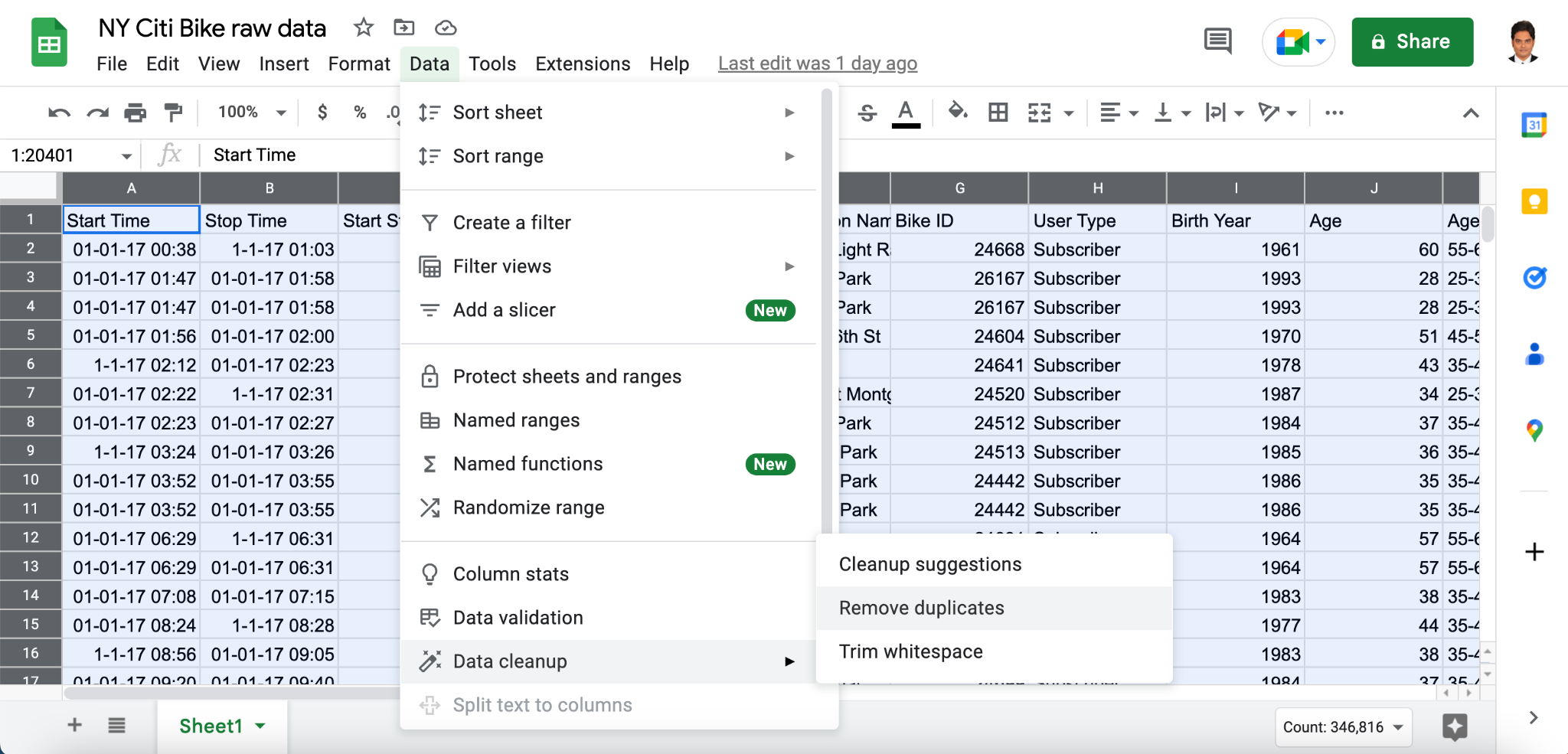


Figure 3

The following window will pop up:

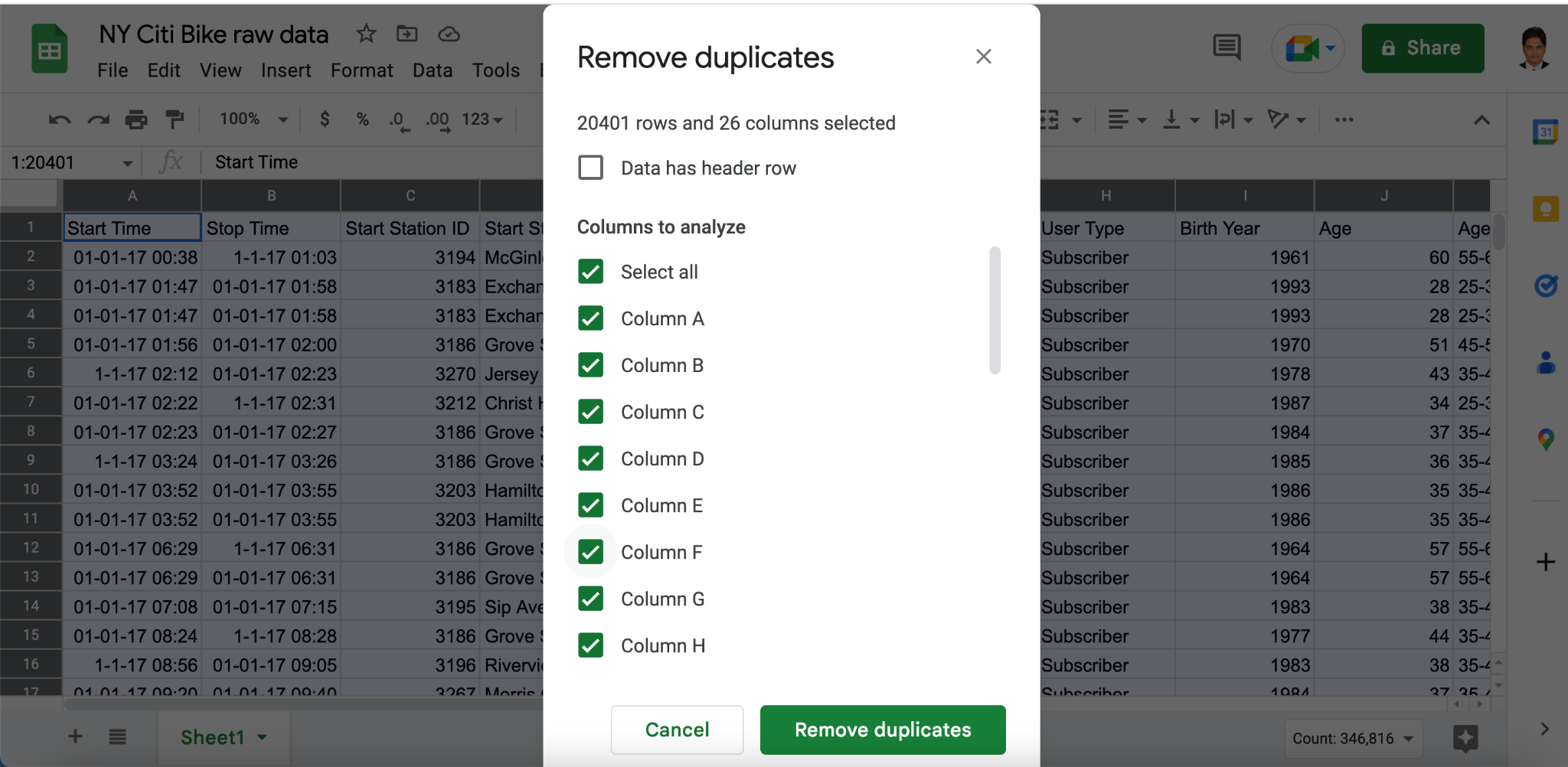


Figure 4

I want to search the entire dataset for duplicates, so I leave all checkboxes selected and click “Remove duplicates.”The dataset contained over 3,500 duplicate rows! So I get this message

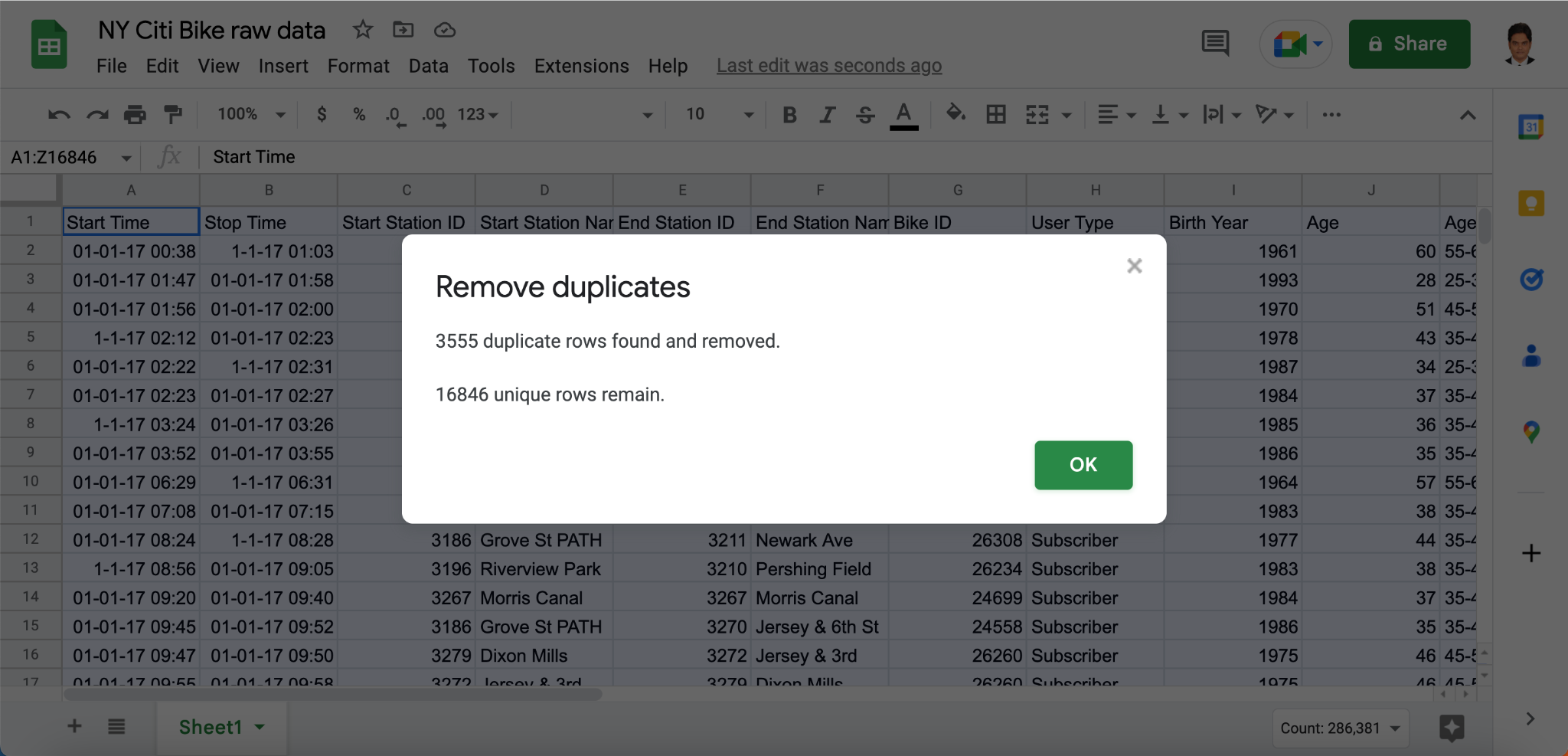


Figure 5

I have successfully removed all those thousands of duplicates rows from my dataset.

**Step 3:** Now I have empty rows in my dataset. I am going to remove those, too. Go to the “Data” menu in the toolbar and select “Create a filter” from the dropdown menu. There is a funnel icon appear next to each of my column headings.

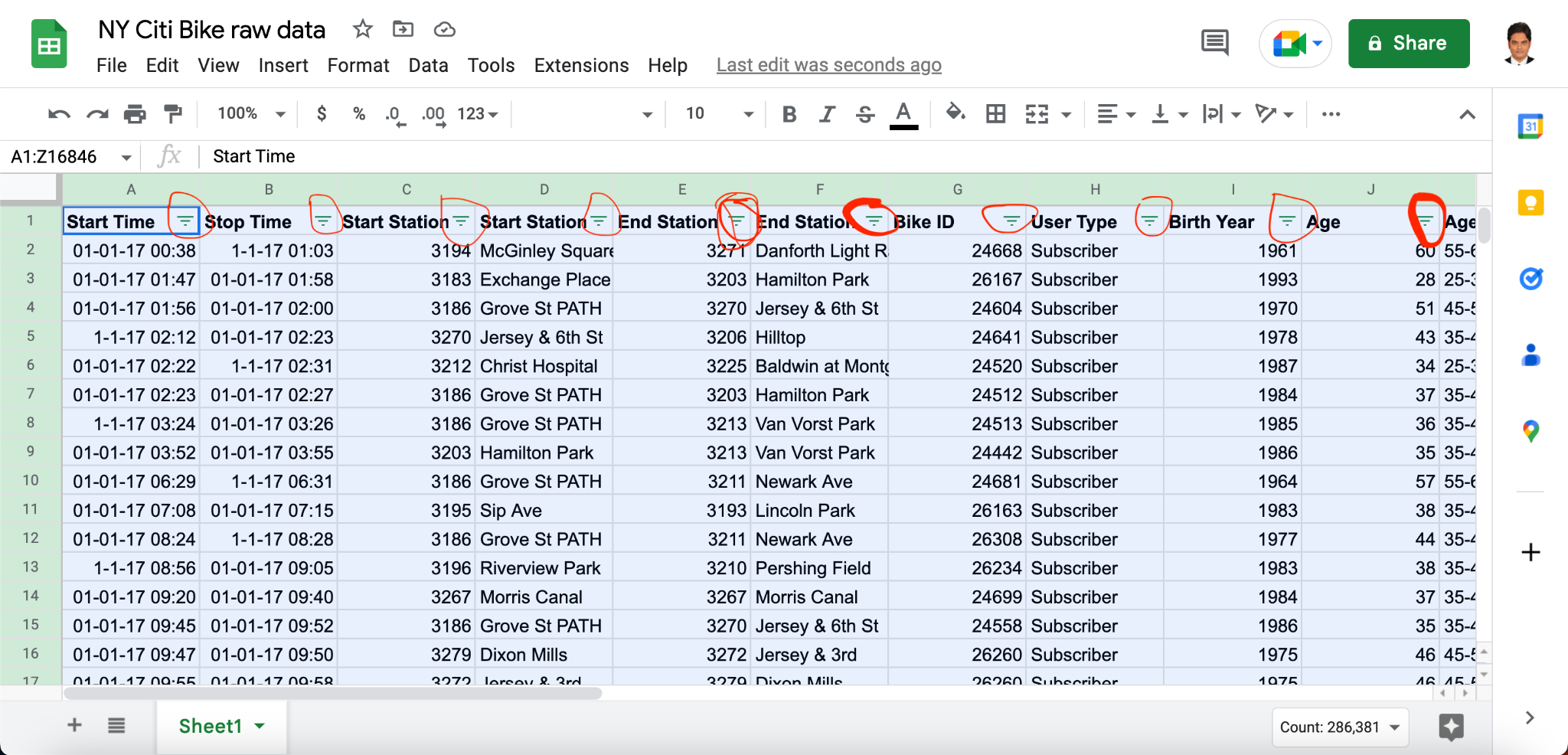


Figure 6

I am going to filter column A for blanks. This is simply an easy way to highlight all the rows underneath my data that are empty (the ones I want to delete). I am deleting them now so they don’t pull through into our pivot tables later.