**CitiBike Data Analysis   
(Insights & Recommendations)**

**An Introduction:** [**Citi Bike**](https://www.citibikenyc.com/how-it-works) 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 🚴

Like most organizations, Citi Bike is constantly looking for ways to improve their business model and provide an even better experience for their customers. Through the Citi Bike app, they are able to gather loads of useful data which, when analyzed, reveals great insights into things like user demographics and behavior—for example, when and where people pick up and drop off their bikes and how long the average journey lasts. Such data is extremely valuable as it helps the good people at Citi Bike to understand how the service is being used, and to plan and make decisions accordingly. For example, at what rate is the customer base growing and how many more bikes should they install across the city to accommodate this growth? Where should they install the most bikes? Who should they tailor their marketing and advertising to? Essentially, data helps them to determine where and how their money and efforts can be invested for maximum impact.

**Based on that Dataset that I got from careerfoundry crash course from this link:**

“[What Is Data Analytics? | Free Tutorial for Beginners (careerfoundry.com)](https://careerfoundry.com/en/tutorials/data-analytics-for-beginners/introduction-to-data-analytics/)”

**This link to the dataset directly to download it by making a copy of it, then start cleaning, preparing and analyzing the data and finally show the insights and the recommendation for increasing profits:**

**This is the final form that I made after analyzing the data and get insights from it:** “[New York Citi Bikes\_Analyzed Data - Google Sheets](https://docs.google.com/spreadsheets/d/1AK063HsDwQXB_xlbQr81qkomD-sTPf--Ac-s82D7MEo/edit#gid=1565002328)”

**Key Questions:**

“ Mission is to analyze data collected by Citi Bike and help key stakeholders to make smart, data-driven decisions based on the insights you uncover.”

Here’s what I will seek to investigate:

1. **What are the most popular pick-up locations across the city for Citi Bike rental?**
2. **How does the average trip duration vary across different age groups, and over time?**
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?**
5. **Do factors like weather and user age impact the average bike trip duration?**

**Note:**

(First, we need to prepare this data by cleaning it and make it organized and make more sense to be able to extract useful insights from it easily)

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**“Some key steps in the data cleaning process”**

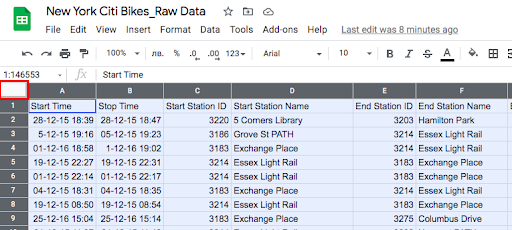
**In “Google Sheets”**

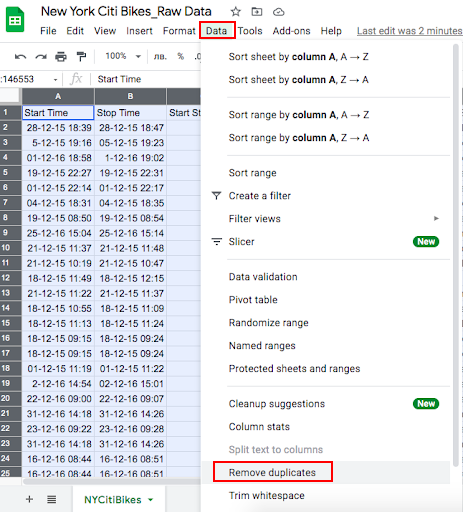
1. **Delete unnecessary columns**. Chances are, your dataset will contain some values that aren’t relevant to your analysis. For example, in an analysis of students’ test scores compared to hours spent studying, things like student ID number and date of birth aren’t relevant. You could simply delete the columns containing this data**. (I didn’t use this step in this dataset)**
2. **Identify and remove duplicates.** Duplicate data tends to occur during the data collection phase, so it’s important to filter them out. I’ll be removing duplicates from my dataset.

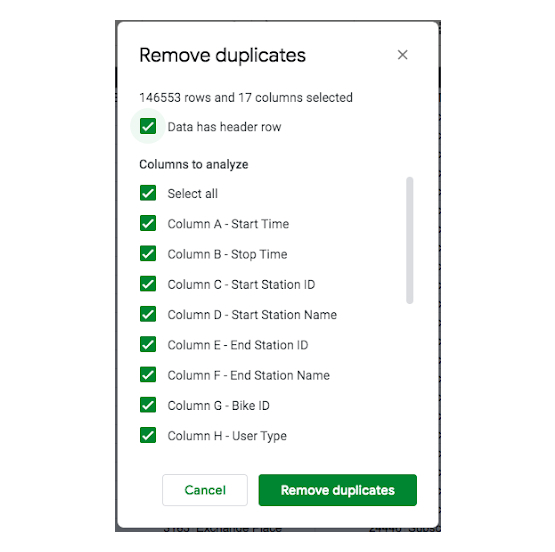
**By following this path of orders:**

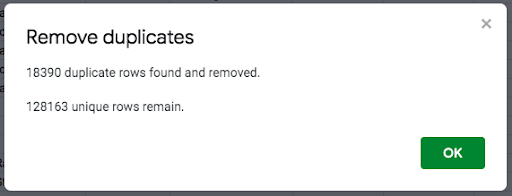
**1- (Select All The Data > Click on Data section > Remove duplicates)” Excel”**

**2- (Select All The Data > Click on Data section > Data cleanup > Remove duplicates)” Google Sheets” (steps in pictures from Google Sheets)**

****

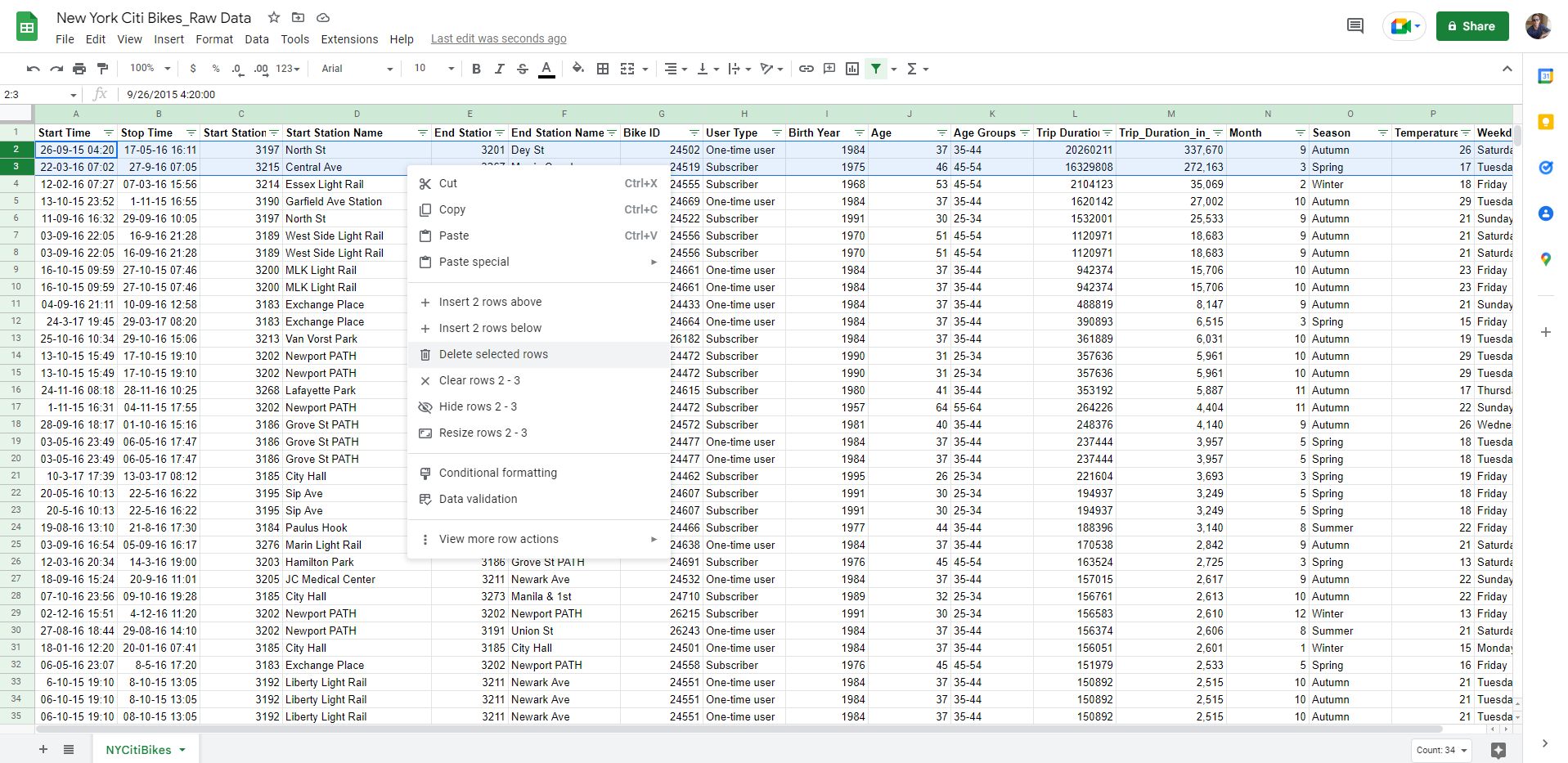
****

****

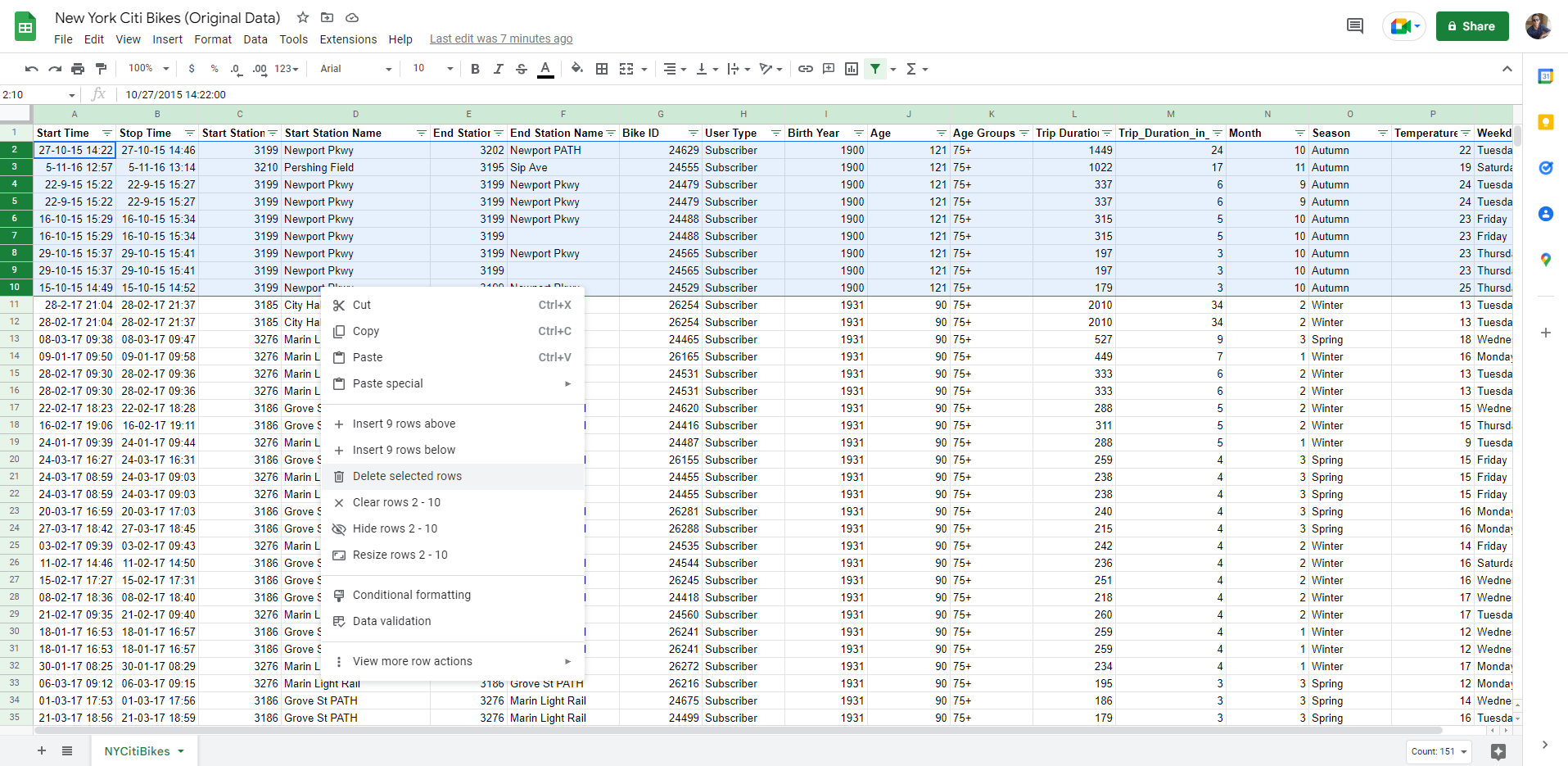
****

**3-Remove unwanted outliers**. [**Outliers**](https://careerfoundry.com/en/blog/data-analytics/what-is-an-outlier/) are values that differ significantly from other values in your data. Outliers may be the result of an error, but that’s not always the case, so approach with caution when deciding whether or not to remove them. (Filter the data first from the Highest or the Lowest)

For Example:

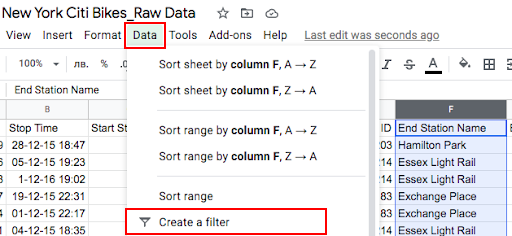
1- when I filter the “trip\_duration\_in\_mins”, from the Highest to the Lowest (Sort Z-A) I discover that there are 2 cells with very extreme values, so I select the rows that contain them and delete them all.

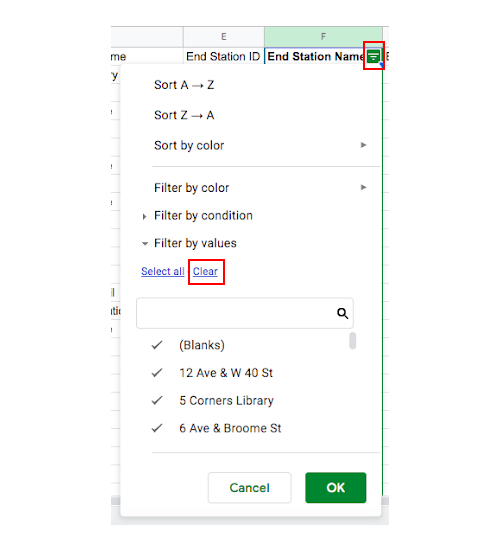
1. **Also, when I checked the age limits from highest to lowest, I discovered that there is an outlier age that is “121 years old” what make no sense at all, so I deleted it’s rows from my dataset**

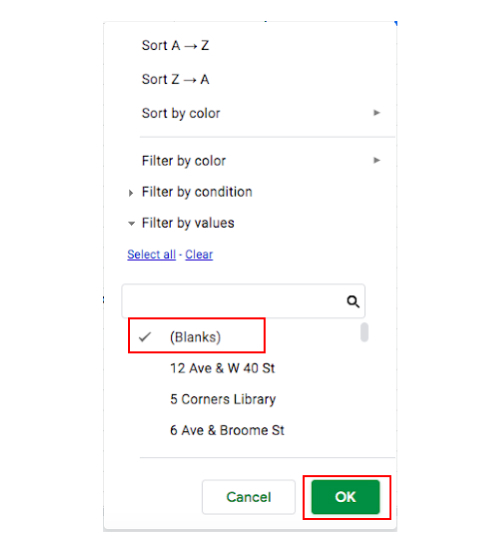
****

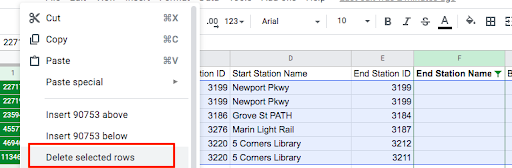
**4-Fix inconsistencies** : As already mentioned, inconsistencies in data include things like typos and irregular naming conventions. You can fix these manually (for example, using the “Find and replace” function in Google Sheets or Microsoft Excel to locate one spelling or convention and replace it with another) or by creating a filter. (Didn’t need to fix any inconsistencies here in this dataset)

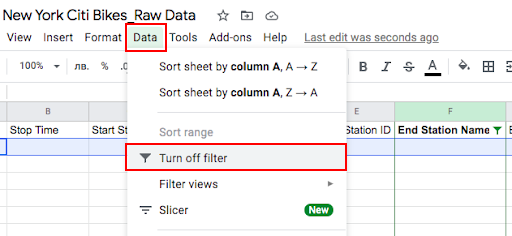
### **5-Identify and remove missing data points: Here I will focus on deleting null or blank cells that doesn’t represent any kind of data. Here, I will use the filter function to show only blank or null cells, then delete them. (This process will be applied on all the columns by one column for each time)**





****

****

****

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**“Now it’s time for the Data Exploring and analyzing process”**

**In “Google Sheets”**

**- You can think of exploratory data analysis as an initial investigation of your dataset where you seek to understand and summarize its main characteristics. EDA**

**- Is useful because it helps you to understand how your data is structured, to spot potential patterns and trends, and to catch any anomalies. EDA is also important for determining if the methods of analysis you are planning to use later on are actually appropriate for your dataset.**

## **- Data exploring process can be summarized in 2 components and these are (***Descriptive statistics and pivot tables*)

A- *Descriptive Statistics:* **help you to summarize or describe the characteristics of your dataset in a meaningful way**. **Also, useful for spotting potential errors or strange occurrences within your dataset.**

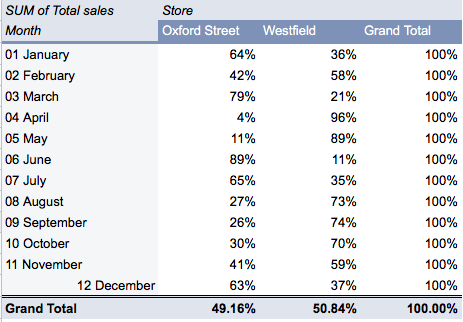
**The three main types of descriptive statistics are:**

**1- Frequency distribution**, which tells you how frequently (i.e. how many times) a certain value occurs within your dataset.

**2- Measures of central tendency**, which estimate the middle or average values within your dataset. **Measures of central tendency** are the **Mean**, **Median**, and **Mode**.

**3- Measures of variability** help you gauge how much variability or “spread” there is within your dataset; in other words, how spread out the values are. **Measures of variability** are **Range**, **Standard Deviation**, and **Variance**.

B- *Pivot Tables****:* A pivot table summarizes large amounts of data in a more digestible, at-a-glance format. It does this by grouping the data in a meaningful way, for example by showing the sum or average values of certain variables.**

**Example:** Let’s imagine you have data for a chain of department stores. Your dataset includes data on the sales made each month in two different store locations. You want to be able to see, at a glance, how each store performed across the year, but it’s impossible to know when faced with thousands of rows of data in a spreadsheet. Here’s where a pivot table can help. You’d add up the total sales for each store for each month of the year, and summarize the data in a pivot table, like so:

**- Basic Questions to Answer (Using Descriptive Statistics):**

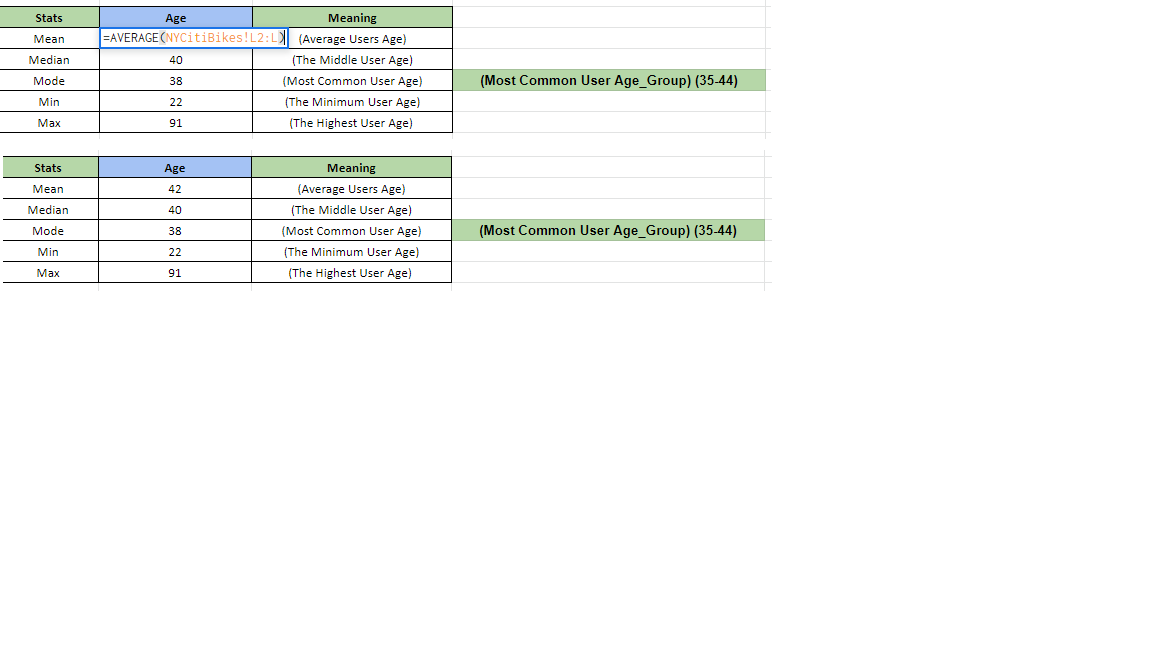
**1-What’s the average age of Citi Bike users?**

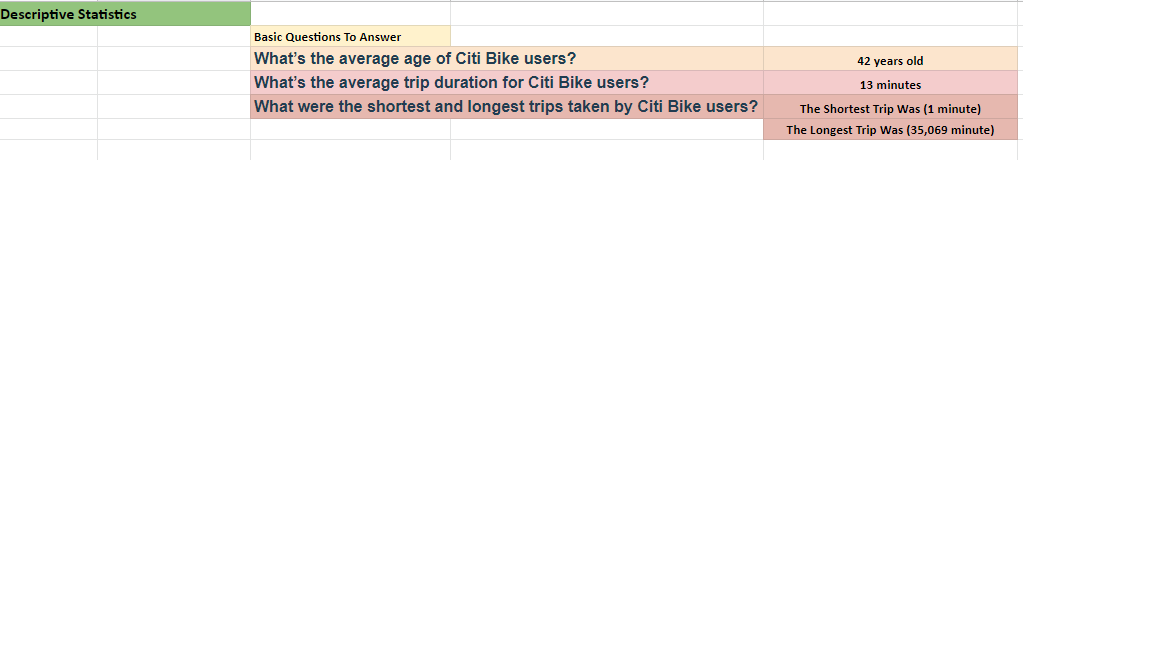
**2-What’s the average trip duration for Citi Bike users?**

1. **What were the shortest and longest trips taken by Citi Bike users?**

**- To answer these questions, we need do a Descriptive Analysis on 2 Columns (Age, Trip\_duration\_in\_mins)**

****

****

**- Then we going to have our simple 3 Questions Answers:**

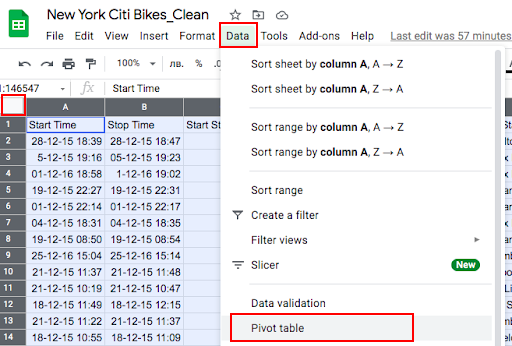
**- Now, to answer the main question of our analysis, I will use pivot tables and charts, and those questions were:**

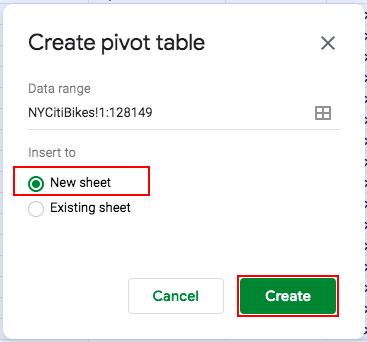
**Q1- What are the most popular pick-up locations across the city for Citi Bike rental?**

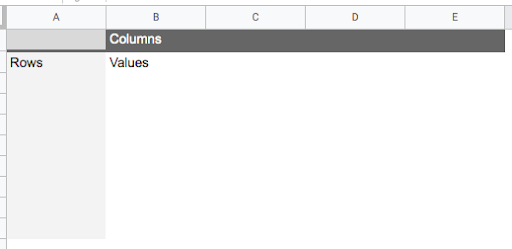
**Q2- How does the average trip duration vary across different age groups, and over time?**

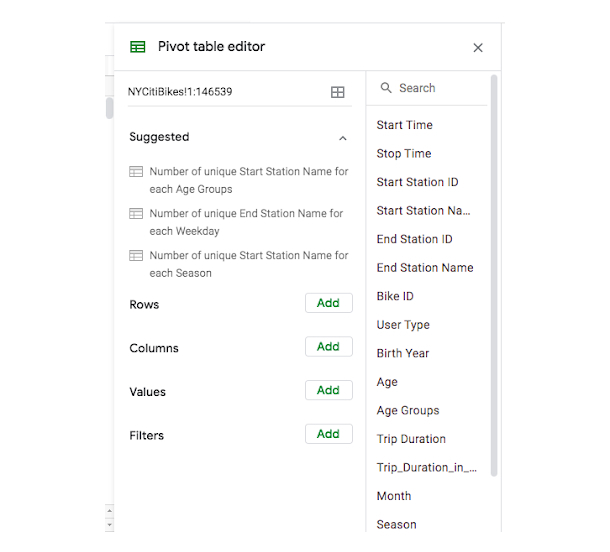
**Q3- Which age group rents the most bikes?**

**Q4- How does bike rental vary across the two user groups (one-time users vs long-term subscribers) on different days of the week?**

**Q5- Do factors like weather and user age impact the average bike trip duration?- Pivot tables here will do a vital role in making the analysis easier as shown here, I will have a sheet for pivot tables processes and each result will be in its own sheet.**

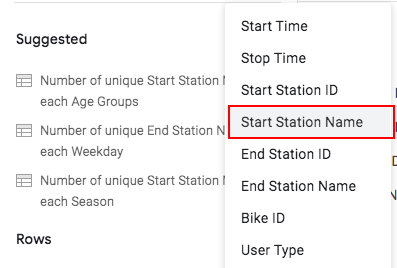
****

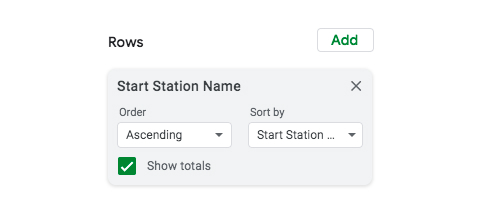
****

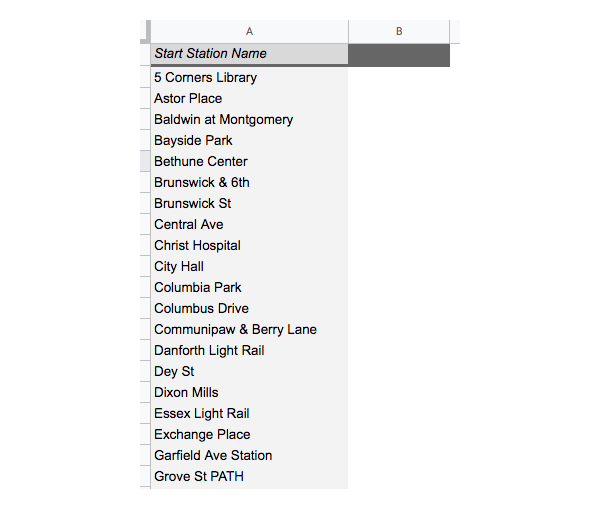
****

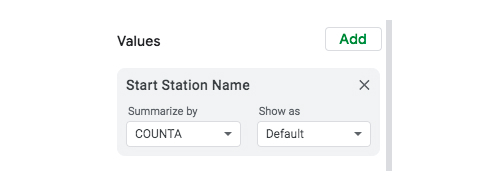
**-Starting with the first question:**

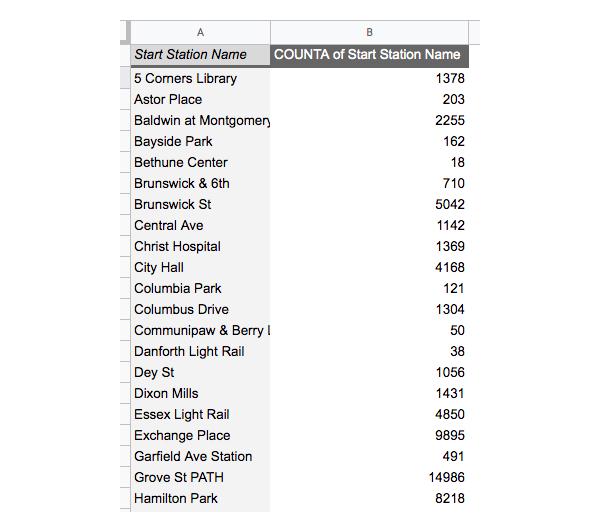
**Q1- What are the most popular pick-up locations across the city for Citi Bike rental?**

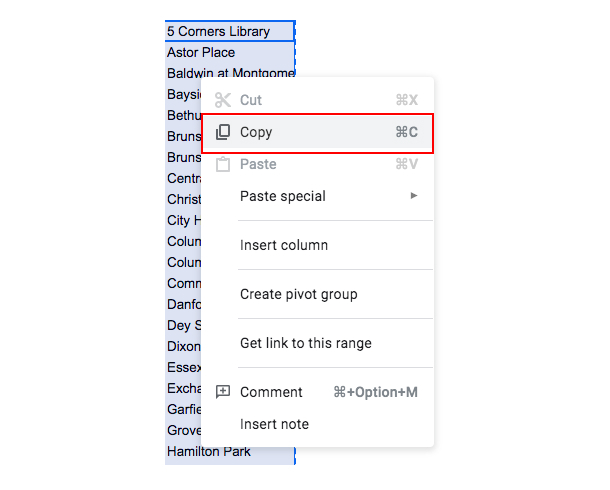
**- in this case, we’re interested in the “Start station name” column. We’ll calculate the frequency (another descriptive statistic) to see how many times each pick-up station occurs within the dataset. This will allow us to see which ones occur most frequently, and are therefore the most popular for Citi Bike pick-up.**

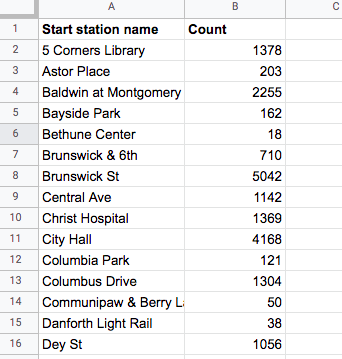
****

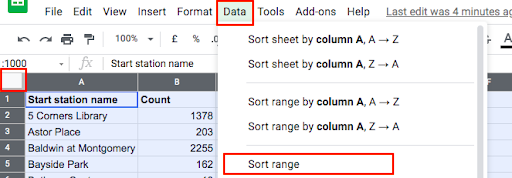
****

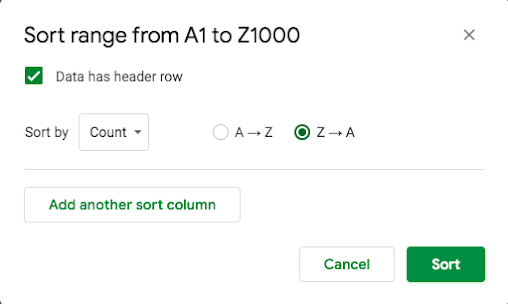
****









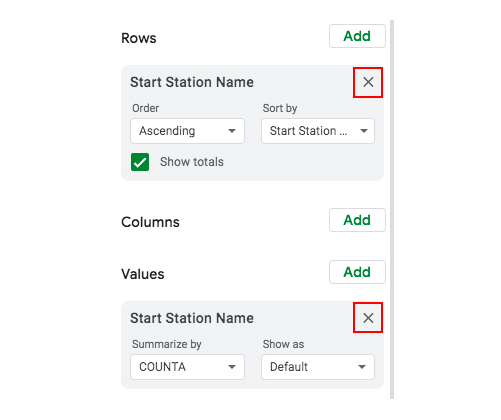


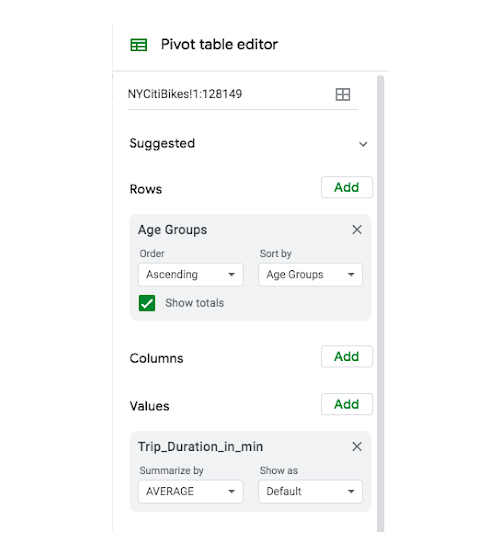
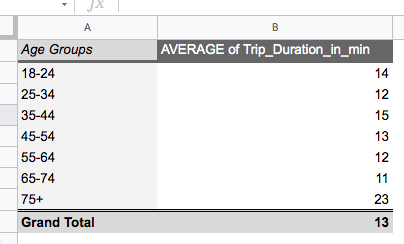


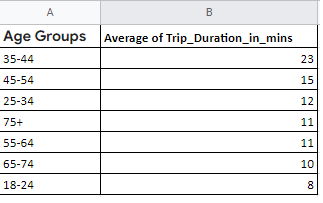
**Q2-** **How does the average trip duration vary across different age groups, and over time?**

**Here, I will get rid of the previous factors I was analyzing and put my new ones.**

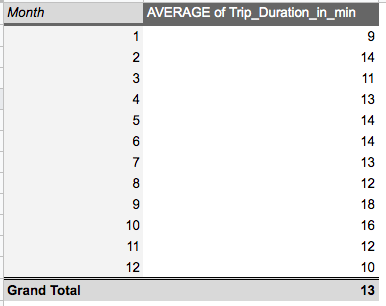
**- First, I will start showing the relation between “Age\_Group” and the “trip\_duration\_in\_mins”**



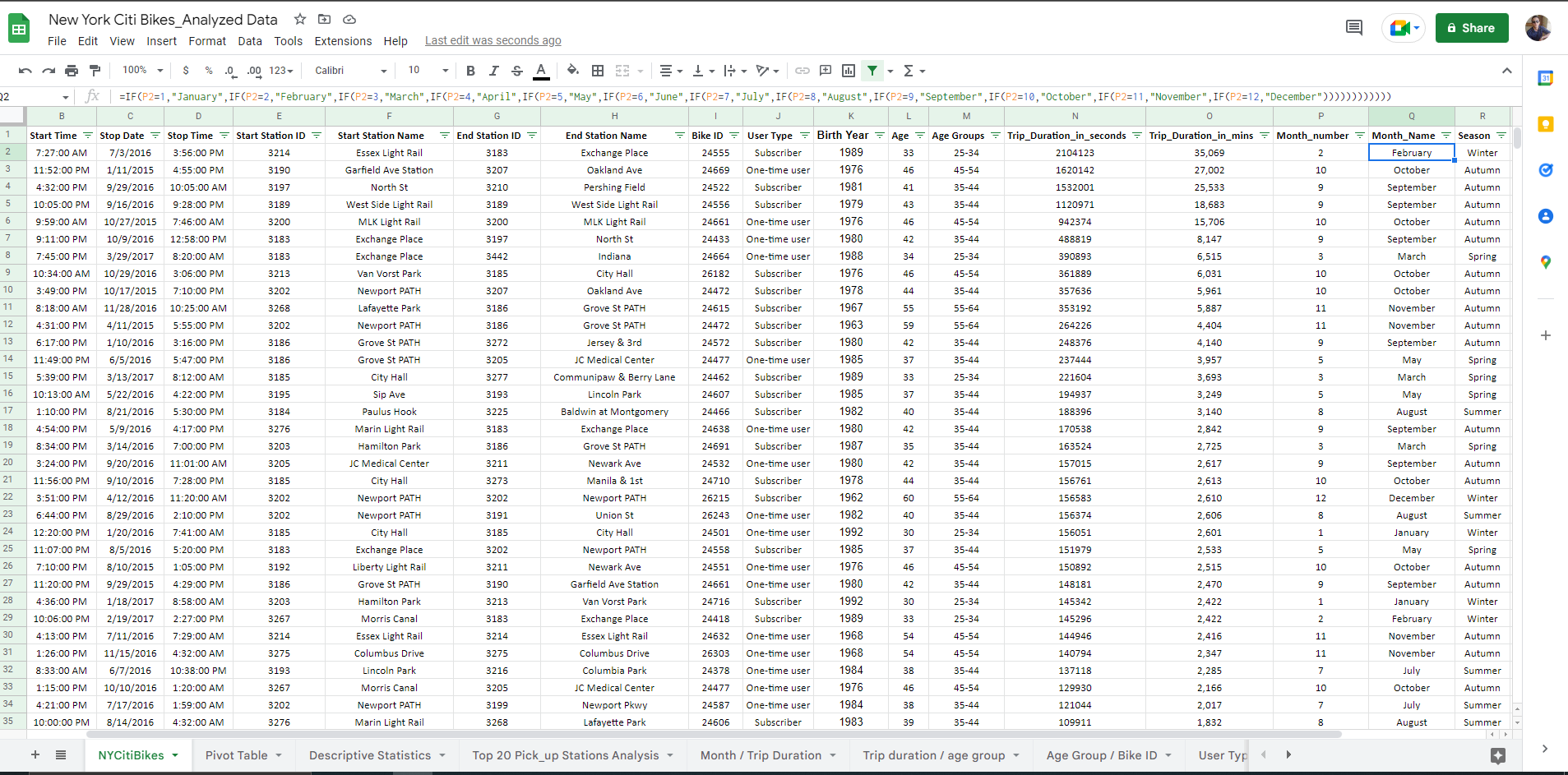


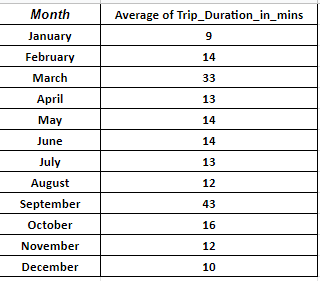
****

**- Secondly, I will start showing the relation between “Month\_Name” and the “trip\_duration\_in\_mins**

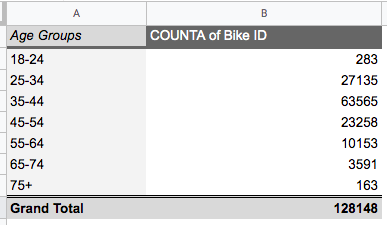


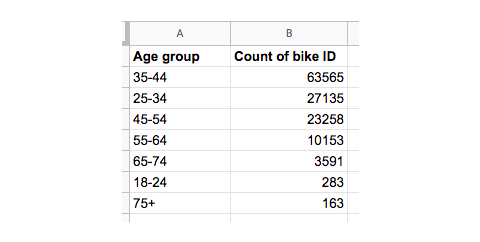
**Note: to make it easier in reading with no effort at all, I will create a new table and call Month\_Name by using this formula in the month\_name column cells**

**“**=IF(P2=1,"January",IF(P2=2,"February",IF(P2=3,"March",IF(P2=4,"April",IF(P2=5,"May",IF(P2=6,"June",IF(P2=7,"July",IF(P2=8,"August",IF(P2=9,"September",IF(P2=10,"October",IF(P2=11,"November",IF(P2=12,"December"))))))))))))**”**

**- And apply it over all the cells in the column, and the result will be like this:**

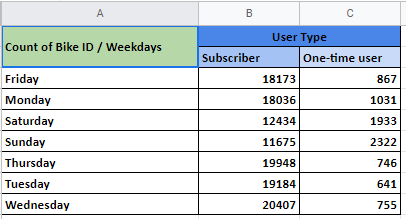
**Q3-** **Which age group rents the most bikes?**



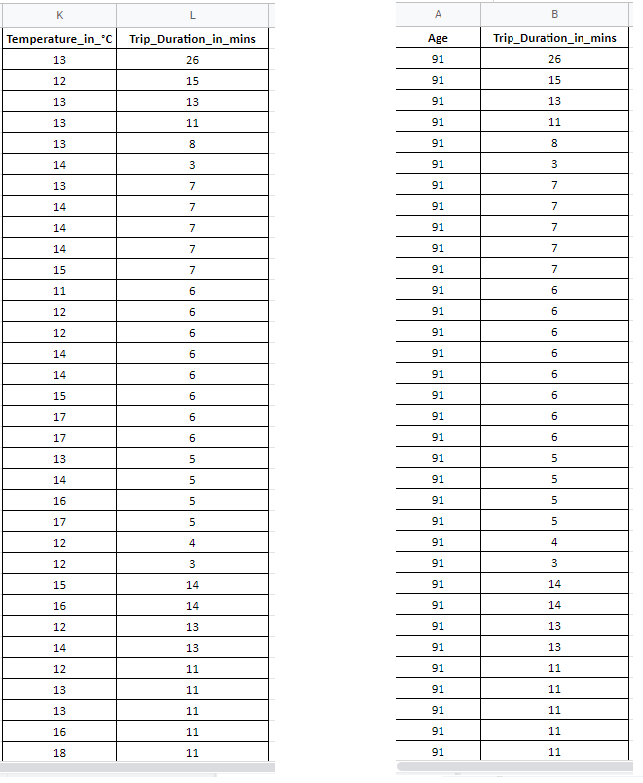
****

**Notice: You can see that users aged 75+ rented the least bikes, The highest number of bikes were actually rented by those in the 35-44 age range.**

**Q4- How does bike rental vary across the two user groups (one-time users vs long-term subscribers) on different days of the week?**



**Q5- Do factors like weather and user age impact the average bike trip duration?**



Note: This is a part of the data as it’s huge and not possible to be contained in a pic, but when we come to charts, it will be understood way better.

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**“Now it’s time for the Data Visualization Based on the previous analyzed Data”**

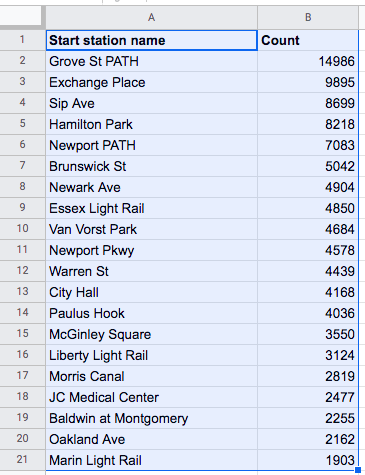
**In “Google Sheets”**

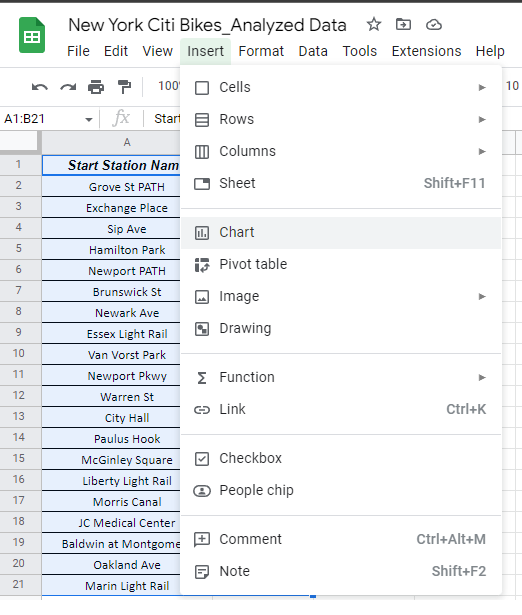
**“In this phase, I will present the analysis for each question in a chart that represent it in an easy way to be understood by clients and regular people.”**

**- Now visualize my questions results in form of charts, each chart fits a different question type.**

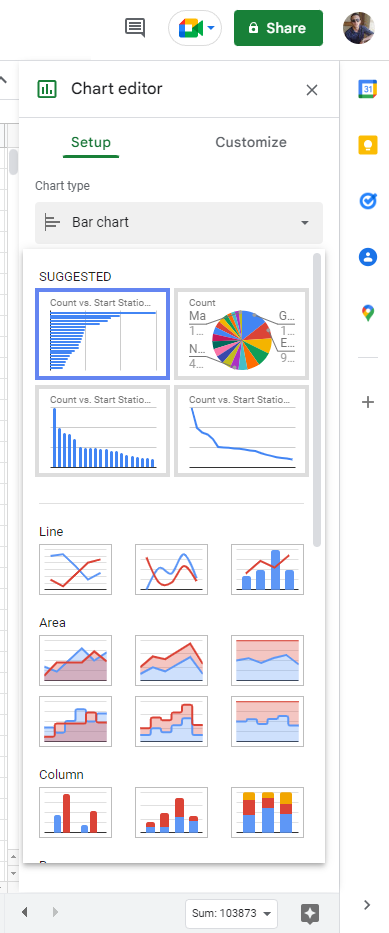
**Q1- What are the most popular pick-up locations across the city for Citi Bike rental?**

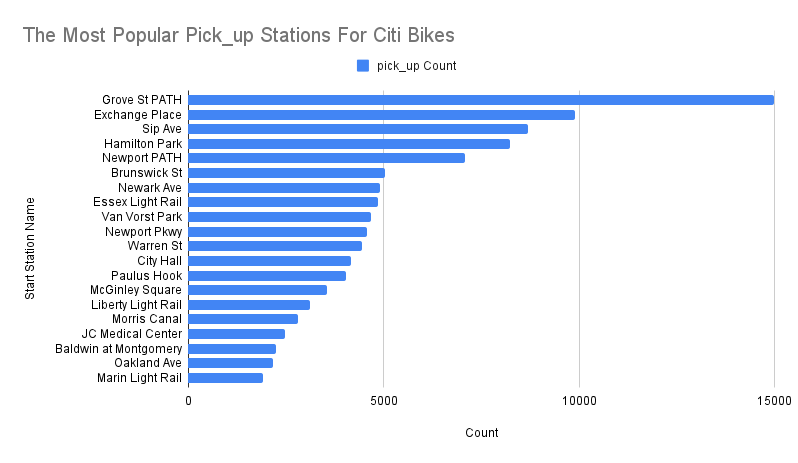
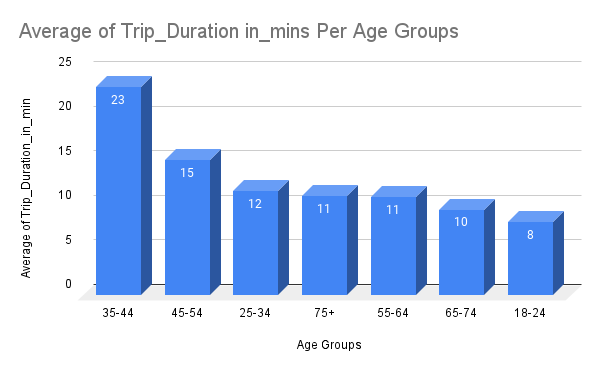
**- I will show the steps for visualization in this question and the rest almost the same**

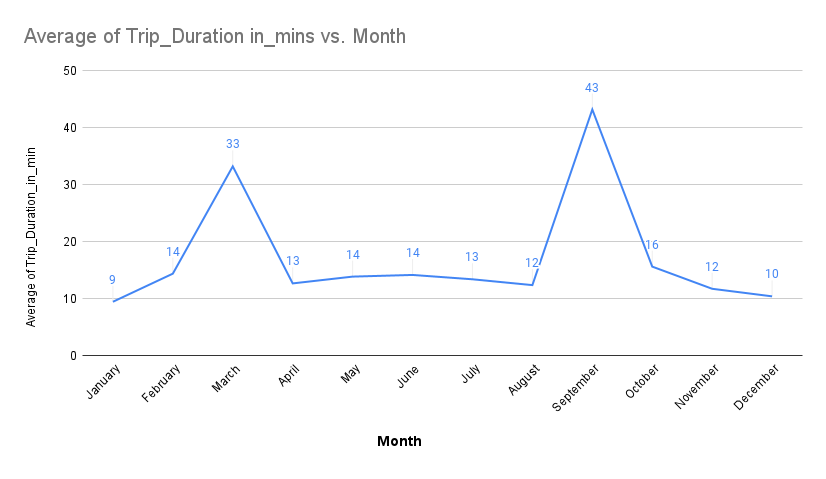
****

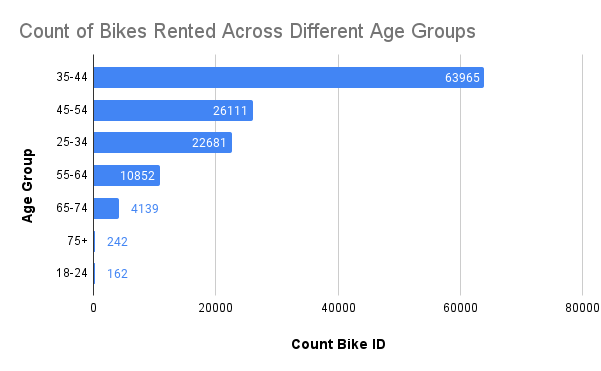


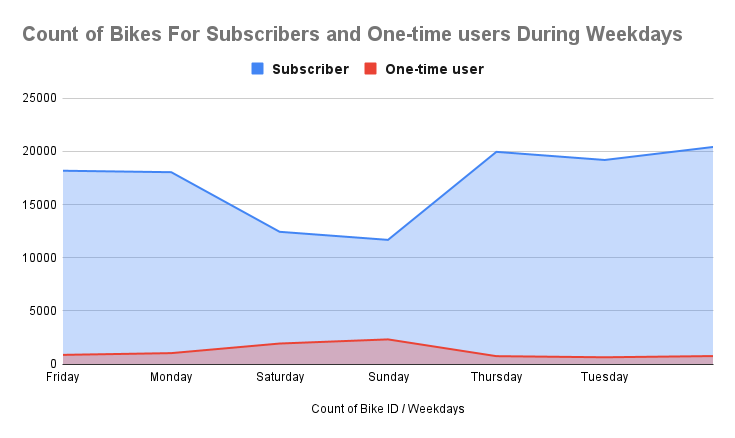
**- I will pick a bar chart as my wanted chart for visualizing this question results.**

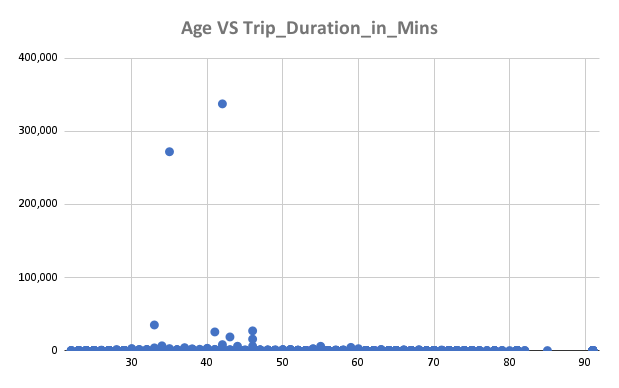
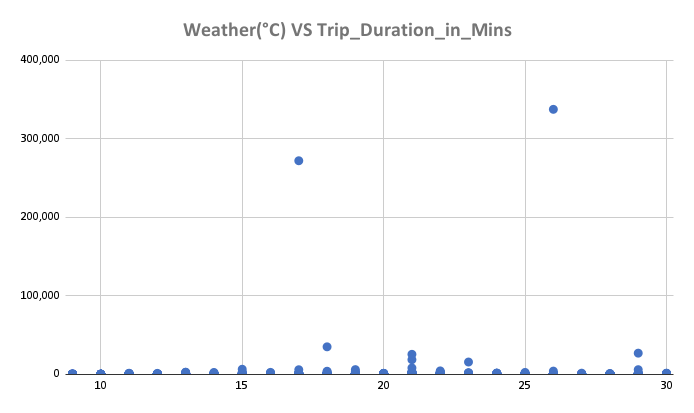


**Q2- How does the average trip duration vary across different age groups, and over time? (This is a Column Chart)**

****

**Q3- Which age group rents the most bikes?**

**Q4- How does bike rental vary across the two user groups (one-time users vs long-term subscribers) on different days of the week?**

**Q5- Do factors like weather and user age impact the average bike trip duration? (Scatter Plot Charts)**

**- Finally, the presentation for this dataset (Insights & Recommendations) will be in a separate PowerPoint File. (the crown of the whole processes)**