**Citi Bike Data Analysis**

*Google Data Analytics Professional Certificate Case Study*

My name is Abdullahi Adeyemi Dawodu, I recently just completed the program Google Data Analytics professional certificate. As part of the course project, students are required to complete a data analysis case study to proof or show case their skills based on what they have learnt so far during the period of the course. In this case study, I hold the position of a junior data analyst working at Citi Bike-New York city official bike sharing.

**Scenario**

Citi Bike is the largest bike-share program in the United States with 20,000 bikes and over 1300 pick-up stations across Manhattan, Brooklyn, Queens, the Bronx, and Jersey City. As stated on our 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 the customers. Through the Citi Bike app, we 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. Myself and other data analyst have been tasked by stakeholders to look into where and how their money and efforts can be invested for maximum impact. They want to know 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?

In order to achieve these set goals, the following steps will be used to process the data:

* Ask phase
* Prepare phase
* Process phase
* Analyze phase
* Share phase
* Act phase

1. **Ask phase**
2. What are the most popular pick-up locations the city for Citi Bike rental?
3. How does the average trip duration vary across different age groups?
4. Which age group rents the most bikes?
5. How does bike rental vary across the two user groups (one-time users vs long-term subscribers) on different days of the week?
6. Do factors like weather and user age impact the average bike trip durations?
7. **Prepare phase**

Citi Bike is a real company. However, the project is purely hypothetical. I’ve sourced the data from a free data site: Kaggle.com and customized it to fit the scope of our project, adding in extra variables as well as excluding other intangible variables. During the analysis, I stored original copy of the data on a secured hard drive, and worked with copies of the data on my PC.

The data included the following fields:

Start time: the date and time the bike was checked-out

Stop time: the date and time the bike was checked-in

Start Station ID: This is the unique identification number for the start station

Start Station Name: the name of the station at the start of the trip

End Station ID: a unique identification number for the end station.

End Station name: the name of the station at the end of the trip

Bike ID: a unique identification number on each bike

User Type: contains the different set of people that makes use of the subscribers and one-time users.

1. **Process Phase**

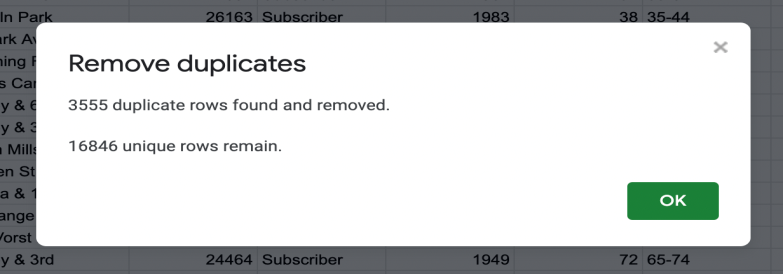
**Excel:** I used this for the initial cleaning and processing of the data. I also used it to analyse the average ride length per month, mode number of rides per month.

**Tableau:** I used this to create two visualizations from excel spreadsheets. As the platform lends itself to drag and drop functionality and allows one to create simple yet clear visuals and join data from various sources.

**Excel Cleaning and Manipulation**

The datasets contains 146553 rows and 17 columns.

First step – I removed the duplicates the datasets contained: over 3555 duplicate rows.



I filtered the dataset for blanks. This is simply an easy way to highlight all the rows underneath the data that are empty. I deleted them so they don’t pull through into our pivot tables later on. (Having an empty row of about 128164-146553).

Next, I discovered that some rows in our data set contain some missing data. In column F (End Station Name) which was later deleted.

1. **Analyze Phase**

Next, a basic table was created titled ‘Descriptive Statistics’. Which include all the descriptive statistics we calculated for each of our two variables of interest: Trip duration and user age.

First, I calculated the mean value our first variable of interest: trip duration in minutes. Which will answer our question – Do Citi Bike users rent bikes for long durations or do they tend to take shorter trips?

I used the AVERAGE formula to sum together all our trip duration data, and then divide it by the number of values. Using the following formula:

=AVERAGE(NYCitiBikes! M2:M146547)

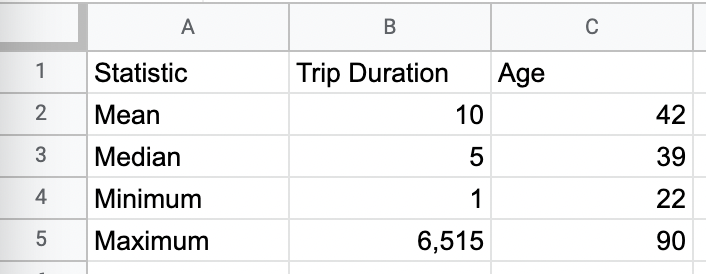
Next, I calculated the median value for our “Trip Duration” variable. Which telss us the middle value in our dataset.

=MEDIAN(NYCitiBikes! M2:M146547)

I then calculated the minimum and maximum values for our “Trip Duration” variable. Which shows us how much there is within our dataset – in other words, how spread out our trip duration values. Is there a huge difference in terms of the shortest trip (the minimum value) and the longest trip (the maximum value)?

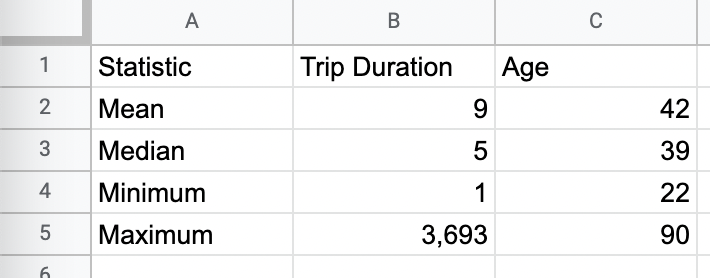
=MIN(NYCitiBikes! M2:M146547)

=MAX(NYCitiBikes! M2:M146547)



You’ll notice that the maximum values for our two variables, “Trip duration” and “User age,” are rather unrealistic.  It’s highly unlikely that someone rode a Citi Bike for 6,515 minutes (that’s about 109 hours, or 4.5 days!). Something’s clearly gone awry!

I did another iteration of data cleaning in order to remove those erroneous values. To remove these outliers from our dataset and stop them skewing your results, I deleted the entire row associated with them.



After this second iteration of data cleaning, we can see that the oldest NY Citi Bike user is 90, and the longest trip was 3,693 minutes (about 2.6 days!)—that’s still pretty impressive, but much more realistic than our previous calculations.

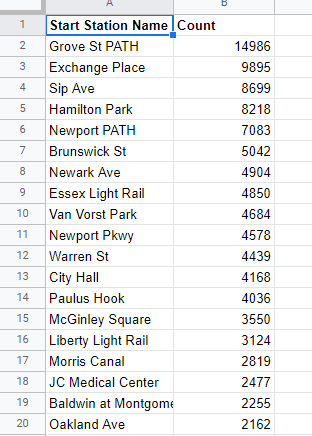
Having calculated some descriptive statistics for our datasets, the next step was answering those key questions with another super powerful tool in Excel: “The Pivot Table” – with this we can easily pinpoint the top 20 pick-up locations.

Task 2.1

The pivot table helps to answer the first question in our analyses – What are the most popular pick up locations in the City for Citi Bike rental.

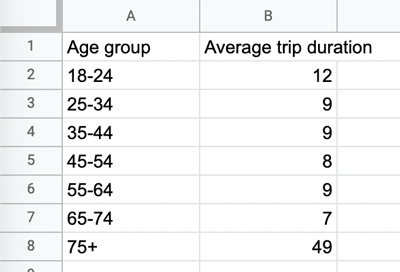
The first question we want to answer is **What are the most popular pick-up locations across the city for NY Citi Bike rental?** In this case, we’re interested in the “Start station name” column. I calculated the frequency (another descriptive statistic) to see how many times each pick-up station occurs within the dataset. This will allow us to see which ones occur most frequently, and are therefore the most popular for NY Citi Bike pick-up.

I added the “Start station name” variable under the “Rows” section of our pivot table and then count the number of occurrences for each station in the “Values” column.



Task 2.2

Created a pivot table to see how bike trip duration varies across different age groups and overtime.

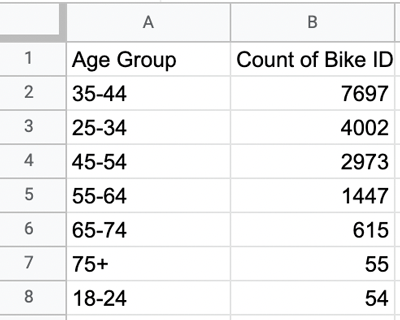


The resulting pivot table gives us a clear at a glance insight into how bike trip durations vary across different age groups. We can see that those in 75+ category tend to take the longest trips on average, while 65-74 year olds take the shortest trips on average.

Task 2.3

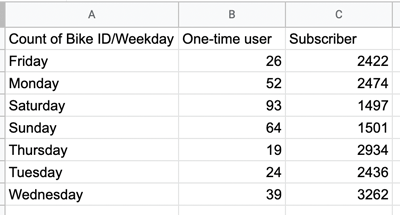
Based on one of the pivot tables I created previously (average trip duration per age group), you might jump to the conclusion that the 75+ age group makes up the biggest segment of our NY Citi Bike customer base. They take the longest trips on average, after all! But does this mean that they rent more bikes than users in other age groups do?

Not necessarily. In order to answer our third question—**which age group rents the most bikes?**—we need to create another pivot table. This time, I looked at the number of unique bike rentals (calculated using the variable “Bike ID”) and summarizing the count by user age (using the “Age group” variable).



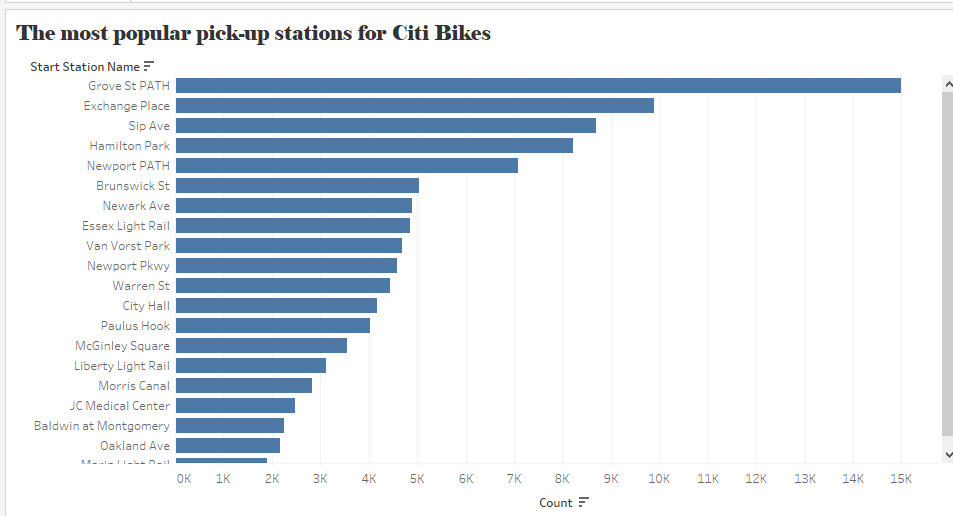
You can see that users aged 75+ rented the least bikes (they just took longer trips on them) The highest number of bikes were actually rented by those in the 35 – 44 age range.

Lastly, I created a pivot table to see how bike rental varies across the two user groups (one-time users vs. long-term subscribers) on different days of the week



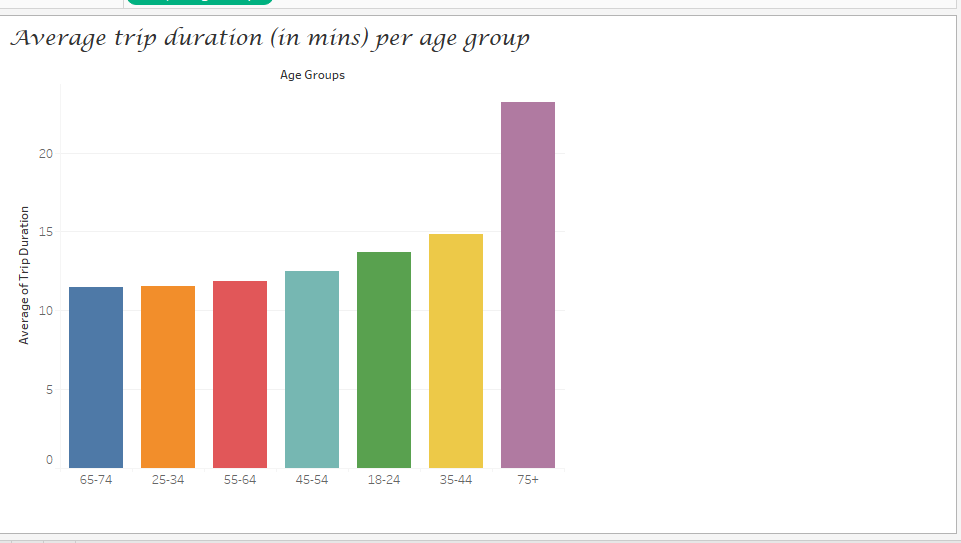
Most Citi Bike users are long term subscribers.

1. **Share Phase:**
2. What are the most popular Citi Bike pick-up locations?

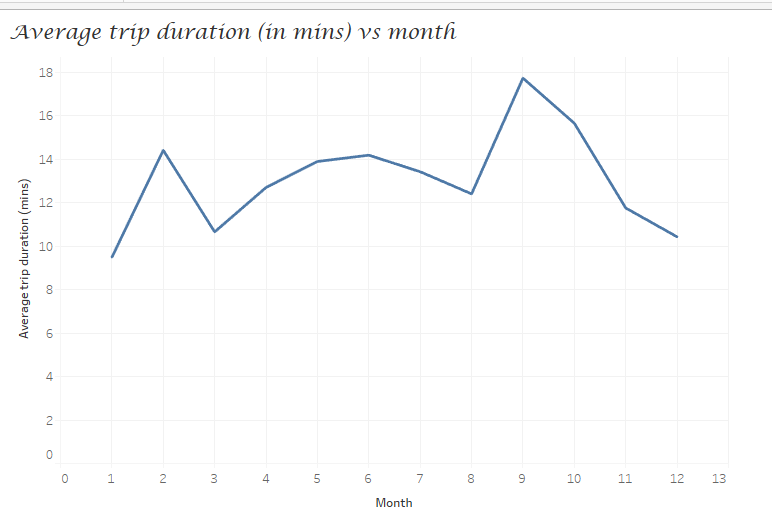
****

Key finding: It is clear from this visualization that Grove St. Path is the most popular pick-up station

1. How does the average trip duration vary across different age groups, and over time?

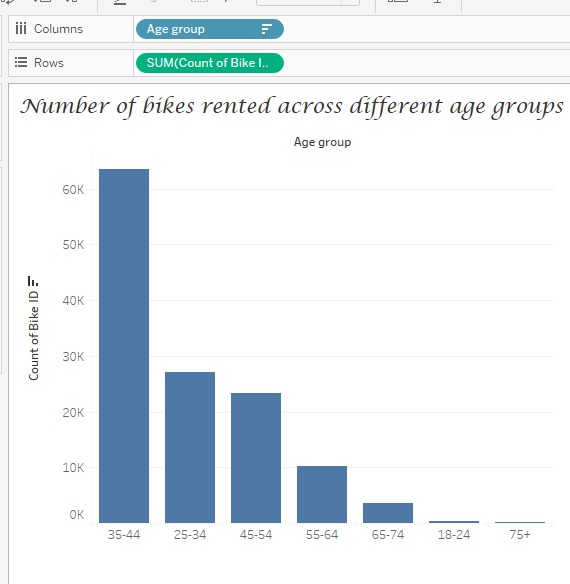


Key finding: Over 75+ year olds take the longest trips (on average). 65- 74 and 25-34 year olds take the shortest trips (on average)



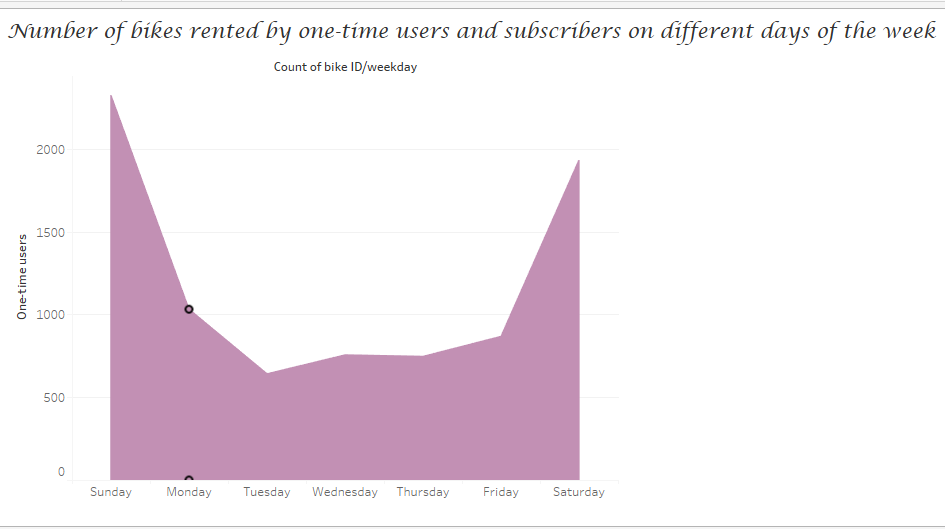
On average, Citi Bike users took the longest trips in September and the shortest trips in January.

1. Which age group rents the most bikes?



Key finding: Here, we can see that age 35-44 year olds rented the most bikes.

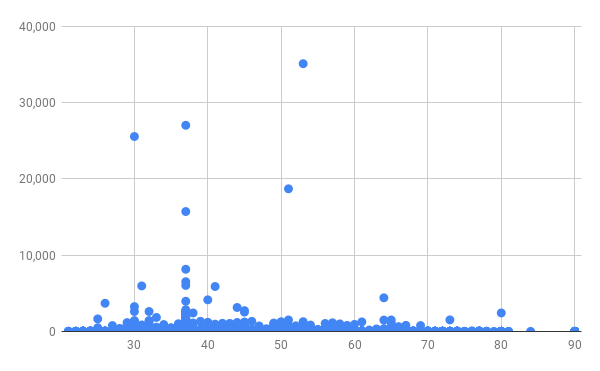
1. How does bike rental vary across the two user groups (one-time users vs long-term subscribers) on different days of the week?

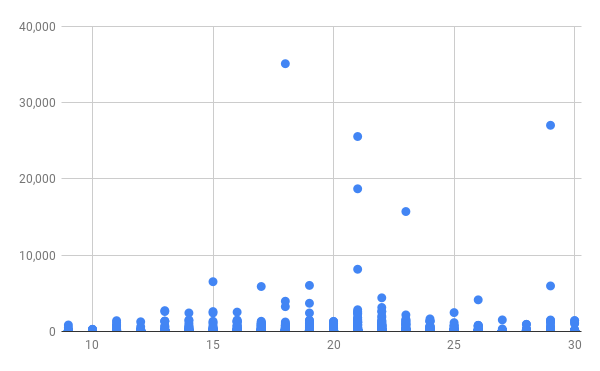


Key findings: Citi Bike customer base is predominantly made up of long-term subscribers. Subscribers are more active during the week. One-time users are more active on weekends.

1. Do factors like weather and age impact the average bike trip duration?

No relationship between user age and trip duration.



Warmer temperatures equals longer Citi Bike rides. 

Summary of findings:

* ***Top 5 pick-up locations for bikes:***
  + *Grove St Path, Sip Ave, Newport Path, Newark Ave, Van Vorst Park*
* ***Customer base:*** 
  + *Mostly long-term subscribers who are more active during the week*
  + *One-time users more active at weekends*
  + *Most bikes rented by 35-44 year olds*
* ***Citi Bike customer behavior:***
  + *75+ year olds take longest average trips, but rent the least bikes*
  + *65-74 and 25-34 year olds take the shortest trips on average*
  + *Longer trips in warmer temperatures → September = longest bike rides*

1. **Actions**

*Product recommendations:*

* *Install more bikes at Grove St Path, Sip Ave, Newport Path, Newark Ave, Van Vorst Park.*
* *Make more bikes available in autumn, as this is when Citi Bike users take the longest trips (which means bikes will be unavailable for longer)*

*Marketing recommendations:*

* *The Citi Bike customer base is mostly long-term subscribers aged between 35-44, who are most active during the week. This tells us that they are probably people who live in New York and use Citi Bikes to commute. Marketing and advertising campaigns should therefore target this particular demographic.*