# Data-Driven Decision Making - Assignment Project

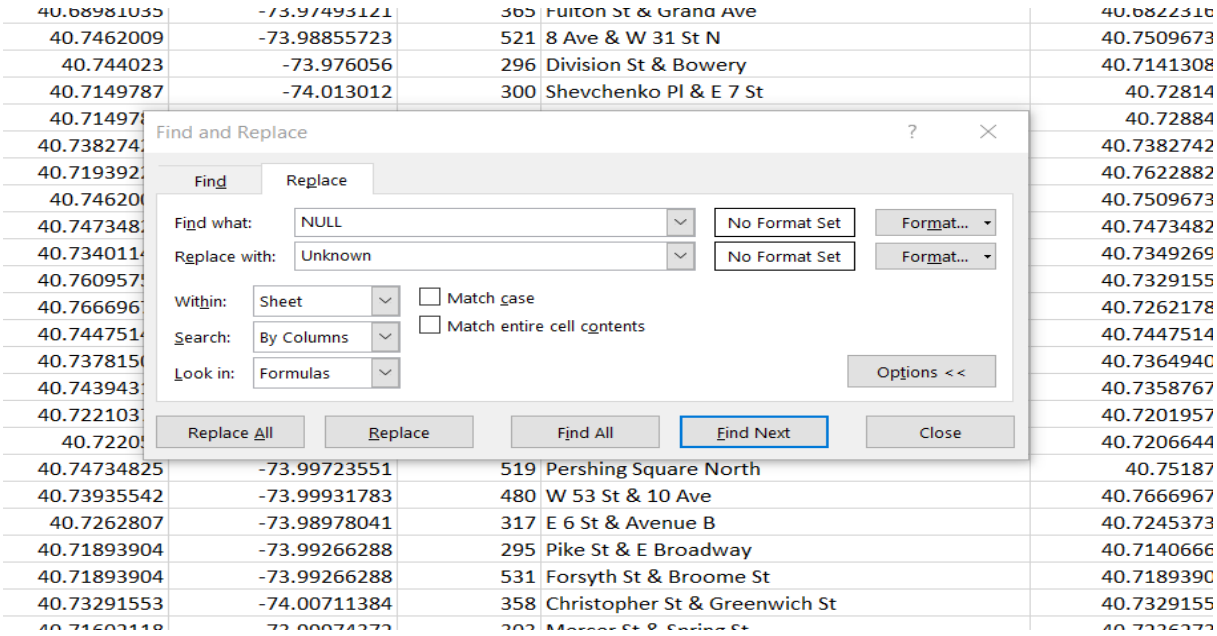
201306-citibike-tripdata file contains a descriptive analysis of the CitiBike company in New York in the period of June 2013. The file contains columns:

- Trip duration in seconds (how long the bike trip lasted from start to finish)
- Start time (the date and time of day where the trip started)
- Stop time (the date and time of day where the trip lasted)
- Start station ID (the ID number of the station which the trip started)
- Start station name (the name of the station which the trip started)
- Start station latitude (value of the vertical (north/south) axis in the coordinate plane, used in maps to locate)
- Start station longitude (value of horizontal (east/west) axis in the coordinate plane, used in maps to locate)
- End station ID (the ID number of the station which the trip ended)
- End station name (the name of the station which the trip ended)
- End station latitude (value of the vertical (north/south) axis in the coordinate plane, used in maps to locate)
- End station longitude (value of horizontal (east/west) axis in the coordinate plane, used in maps to locate)
- Bike ID (the ID of the bike that was used)
- User Type (if the person using the bike is a customer or a subscriber/annual member)
- Birth Year (the birthdate of the person using the bike)
- Gender (the gender of the person using the bike)

Each row in the data represents a single use of the bike from start to finish and the description of the user (the person that rents the bikes).

**Missing values (NULL):**

There were some missing values in the data (in the thousands!) that had the value “NULL”. I changed the value “NULL” and replaced it with “Unknown”, using ctrl+H:



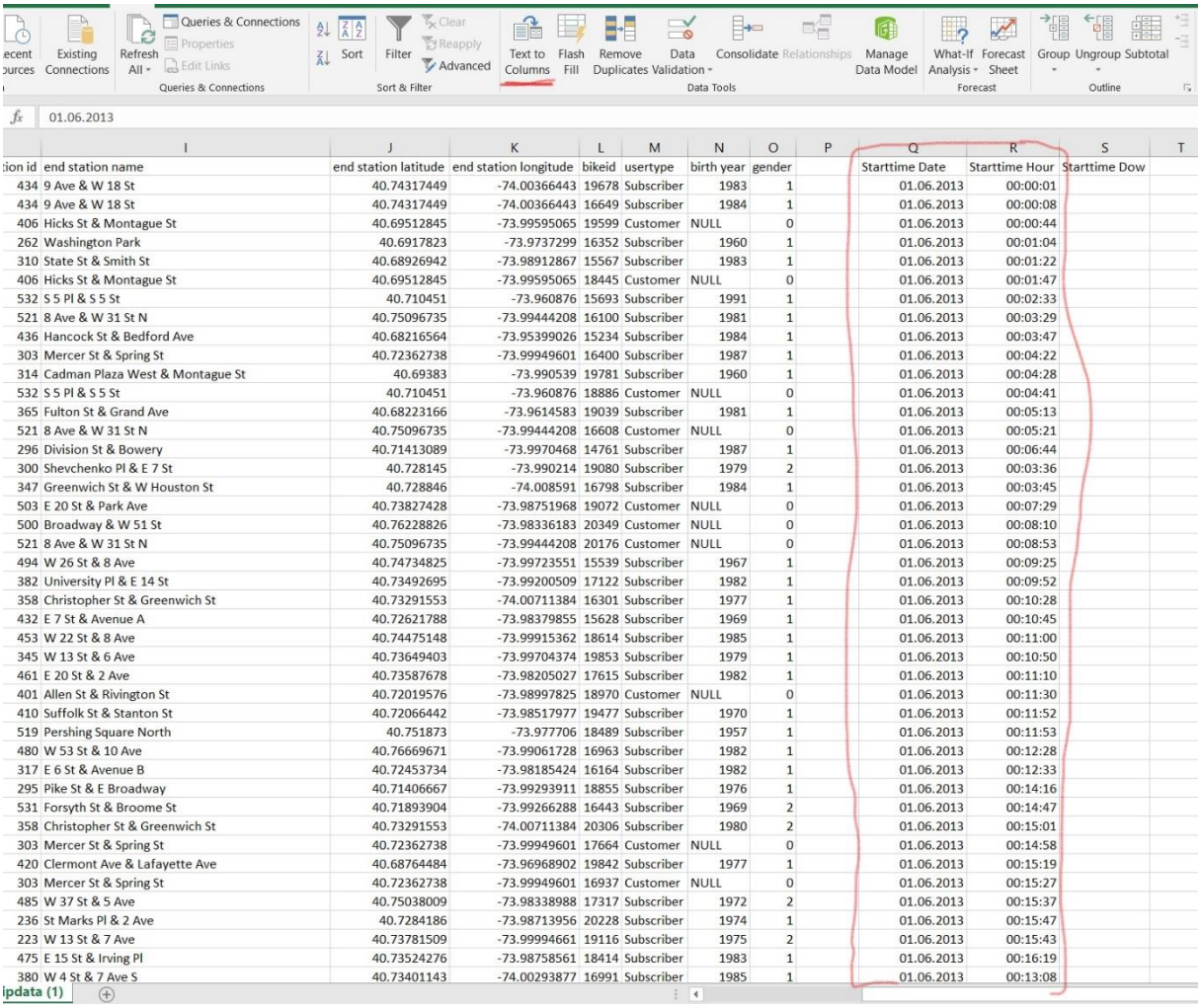

| | | | | |
|---|---|---|---|---|
| 40.68981035 | -73.97493121 | 365 | Fulton St & Grand Ave | 40.6822316 |
| 40.7462009 | -73.98855723 | 521 | 8 Ave & W 31 St N | 40.7509673 |
| 40.744023 | -73.976056 | 296 | Division St & Bowery | 40.7141308 |
| 40.7149787 | -74.013012 | 300 | Shevchenko Pl & E 7 St | 40.72814 |
| 40.71497 | | | | 40.72884 |
| 40.738274 | | | | 40.7382742 |
| 40.719392 | | | | 40.7622882 |
| 40.74620 | | | | 40.7509673 |
| 40.747348 | | | | 40.7473482 |
| 40.734011 | | | | 40.7349269 |
| 40.760957 | | | | 40.7329155 |
| 40.766696 | | | | 40.7262178 |
| 40.744751 | | | | 40.7447514 |
| 40.737815 | | | | 40.7364940 |
| 40.743943 | | | | 40.7358767 |
| 40.722103 | | | | 40.7201957 |
| 40.7220 | | | | 40.7206644 |
| 40.74734825 | -73.99723551 | 519 | Pershing Square North | 40.75187 |
| 40.73935542 | -73.99931783 | 480 | W 53 St & 10 Ave | 40.7666967 |
| 40.7262807 | -73.98978041 | 317 | E 6 St & Avenue B | 40.7245373 |
| 40.71893904 | -73.99266288 | 295 | Pike St & E Broadway | 40.7140666 |
| 40.71893904 | -73.99266288 | 531 | Forsyth St & Broome St | 40.7189390 |
| 40.73291553 | -74.00711384 | 358 | Christopher St & Greenwich St | 40.7329155 |
| 40.71602118 | -73.99974272 | 302 | Mercer St & Spring St | 40.7226272 |

Figure 1: Replacing “NULL” with “Unknown”.

Why I choose to have the missing values is because I believe the missing values is still a value that can be analyzed. What I mean by that is that I fund no missing values or “NULL” in the “start station name” column, but a lot in the “end station name” column (also in other columns like end station latitude/longitude) . It could mean that rented bikes from start stations did not end in an end station, but rather in unknown places in New York city, where people just ended their biketrip in unknown locations. Or people just rented a bike and never actually go nowhere. It could be a lot of reasons. Maybe keeping the missing values will make the analysis difficult. Maybe I have to delete the missing values. I will try to figure this out in this assignment. So long KISS (Keep it simple stupid) 😊

**Assignment 1:**

To create three new variables (columns), the first thing that I did was to use “text to columns” inside the “data” tab to split column B, since column B has both the date and time in the same column. (Note that before I used “text to columns” I copied column B and pasted it in column Q). Figure 1 below shows the result.




| ion id | end station name | end station latitude | end station longitude | bikeid | usertype | birth year | gender | Starttime Date | Starttime Hour | Starttime Dow |
|---|---|---|---|---|---|---|---|---|---|---|
| 434 | 9 Ave & W 18 St | 40.74317449 | -74.00366443 | 19678 | Subscriber | 1983 | 1 | 01.06.2013 | 00:00:01 | |
| 434 | 9 Ave & W 18 St | 40.74317449 | -74.00366443 | 16649 | Subscriber | 1984 | 1 | 01.06.2013 | 00:00:08 | |
| 406 | Hicks St & Montague St | 40.69512845 | -73.99595065 | 19599 | Customer | NULL | 0 | 01.06.2013 | 00:00:44 | |
| 262 | Washington Park | 40.6917823 | -73.9737299 | 16352 | Subscriber | 1960 | 1 | 01.06.2013 | 00:01:04 | |
| 310 | State St & Smith St | 40.68926942 | -73.98912867 | 15567 | Subscriber | 1983 | 1 | 01.06.2013 | 00:01:22 | |
| 406 | Hicks St & Montague St | 40.69512845 | -73.99595065 | 18445 | Customer | NULL | 0 | 01.06.2013 | 00:01:47 | |
| 532 | S 5 Pl & S 5 St | 40.710451 | -73.960876 | 15693 | Subscriber | 1991 | 1 | 01.06.2013 | 00:02:33 | |
| 521 | 8 Ave & W 31 St N | 40.75096735 | -73.99444208 | 16100 | Subscriber | 1981 | 1 | 01.06.2013 | 00:03:29 | |
| 436 | Hancock St & Bedford Ave | 40.68216564 | -73.95399026 | 15234 | Subscriber | 1984 | 1 | 01.06.2013 | 00:03:47 | |
| 303 | Mercer St & Spring St | 40.72362738 | -73.99949601 | 16400 | Subscriber | 1987 | 1 | 01.06.2013 | 00:04:22 | |
| 314 | Cadman Plaza West & Montague St | 40.69383 | -73.990539 | 19781 | Subscriber | 1960 | 1 | 01.06.2013 | 00:04:28 | |
| 532 | S 5 Pl & S 5 St | 40.710451 | -73.960876 | 18886 | Customer | NULL | 0 | 01.06.2013 | 00:04:41 | |
| 365 | Fulton St & Grand Ave | 40.68223166 | -73.9614583 | 19039 | Subscriber | 1981 | 1 | 01.06.2013 | 00:05:13 | |
| 521 | 8 Ave & W 31 St N | 40.75096735 | -73.99444208 | 16608 | Customer | NULL | 0 | 01.06.2013 | 00:05:21 | |
| 296 | Division St & Bowery | 40.71413089 | -73.9970468 | 14761 | Subscriber | 1987 | 1 | 01.06.2013 | 00:06:44 | |
| 300 | Shevchenko Pl & E 7 St | 40.728145 | -73.990214 | 19080 | Subscriber | 1979 | 2 | 01.06.2013 | 00:03:36 | |
| 347 | Greenwich St & W Houston St | 40.728846 | -74.008591 | 16798 | Subscriber | 1984 | 1 | 01.06.2013 | 00:03:45 | |
| 503 | E 20 St & Park Ave | 40.73827428 | -73.98751968 | 19072 | Customer | NULL | 0 | 01.06.2013 | 00:07:29 | |
| 500 | Broadway & W 51 St | 40.76228826 | -73.98336183 | 20349 | Customer | NULL | 0 | 01.06.2013 | 00:08:10 | |
| 521 | 8 Ave & W 31 St N | 40.75096735 | -73.99444208 | 20176 | Customer | NULL | 0 | 01.06.2013 | 00:08:53 | |
| 494 | W 26 St & 8 Ave | 40.74734825 | -73.99723551 | 15539 | Subscriber | 1967 | 1 | 01.06.2013 | 00:09:25 | |
| 382 | University Pl & E 14 St | 40.73492695 | -73.99200509 | 17122 | Subscriber | 1982 | 1 | 01.06.2013 | 00:09:52 | |
| 358 | Christopher St & Greenwich St | 40.73291553 | -74.00711384 | 16301 | Subscriber | 1977 | 1 | 01.06.2013 | 00:10:28 | |
| 432 | E 7 St & Avenue A | 40.72621788 | -73.98379855 | 15628 | Subscriber | 1969 | 1 | 01.06.2013 | 00:10:45 | |
| 453 | W 22 St & 8 Ave | 40.74475148 | -73.99915362 | 18614 | Subscriber | 1985 | 1 | 01.06.2013 | 00:11:00 | |
| 345 | W 13 St & 6 Ave | 40.73649403 | -73.99704374 | 19853 | Subscriber | 1979 | 1 | 01.06.2013 | 00:10:50 | |
| 461 | E 20 St & 2 Ave | 40.73587678 | -73.98205027 | 17615 | Subscriber | 1982 | 1 | 01.06.2013 | 00:11:10 | |
| 401 | Allen St & Rivington St | 40.72019576 | -73.98997825 | 18970 | Customer | NULL | 0 | 01.06.2013 | 00:11:30 | |
| 410 | Suffolk St & Stanton St | 40.72066442 | -73.98517977 | 19477 | Subscriber | 1970 | 1 | 01.06.2013 | 00:11:52 | |
| 519 | Pershing Square North | 40.751873 | -73.977706 | 18489 | Subscriber | 1957 | 1 | 01.06.2013 | 00:11:53 | |
| 480 | W 53 St & 10 Ave | 40.76669671 | -73.99061728 | 16963 | Subscriber | 1982 | 1 | 01.06.2013 | 00:12:28 | |
| 317 | E 6 St & Avenue B | 40.72453734 | -73.98185424 | 16164 | Subscriber | 1982 | 1 | 01.06.2013 | 00:12:33 | |
| 295 | Pike St & E Broadway | 40.71406667 | -73.99293911 | 18855 | Subscriber | 1976 | 1 | 01.06.2013 | 00:14:16 | |
| 531 | Forsyth St & Broome St | 40.71893904 | -73.99266288 | 16443 | Subscriber | 1969 | 2 | 01.06.2013 | 00:14:47 | |
| 358 | Christopher St & Greenwich St | 40.73291553 | -74.00711384 | 20306 | Subscriber | 1980 | 2 | 01.06.2013 | 00:15:01 | |
| 303 | Mercer St & Spring St | 40.72362738 | -73.99949601 | 17664 | Customer | NULL | 0 | 01.06.2013 | 00:14:58 | |
| 420 | Clermont Ave & Lafayette Ave | 40.68764484 | -73.96968902 | 19842 | Subscriber | 1977 | 1 | 01.06.2013 | 00:15:19 | |
| 303 | Mercer St & Spring St | 40.72362738 | -73.99949601 | 16937 | Customer | NULL | 0 | 01.06.2013 | 00:15:27 | |
| 485 | W 37 St & 5 Ave | 40.75038009 | -73.98338988 | 17317 | Subscriber | 1972 | 2 | 01.06.2013 | 00:15:37 | |
| 236 | St Marks Pl & 2 Ave | 40.7284186 | -73.98713956 | 20228 | Subscriber | 1974 | 1 | 01.06.2013 | 00:15:47 | |
| 223 | W 13 St & 7 Ave | 40.73781509 | -73.99994661 | 19116 | Subscriber | 1975 | 2 | 01.06.2013 | 00:15:43 | |
| 475 | E 15 St & Irving Pl | 40.73524276 | -73.98758561 | 18414 | Subscriber | 1983 | 1 | 01.06.2013 | 00:16:19 | |
| 380 | W 4 St & 7 Ave S | 40.73401143 | -74.00293877 | 16991 | Subscriber | 1985 | 1 | 01.06.2013 | 00:13:08 | |



Figure 2: Making columns “starttime date” and “starttime hour”.

To make the last column, column S “starttime dow” I used formula =weekday and added column Q in the formula. I got number 7, which means Saturday (in excel 1 is from Sunday to 7 that is Saturday, as

you probably know). I would like to have the text of a day instead of numbers, so I used “format cells”, then custom inside category and typed dddd, as shown in the next page.

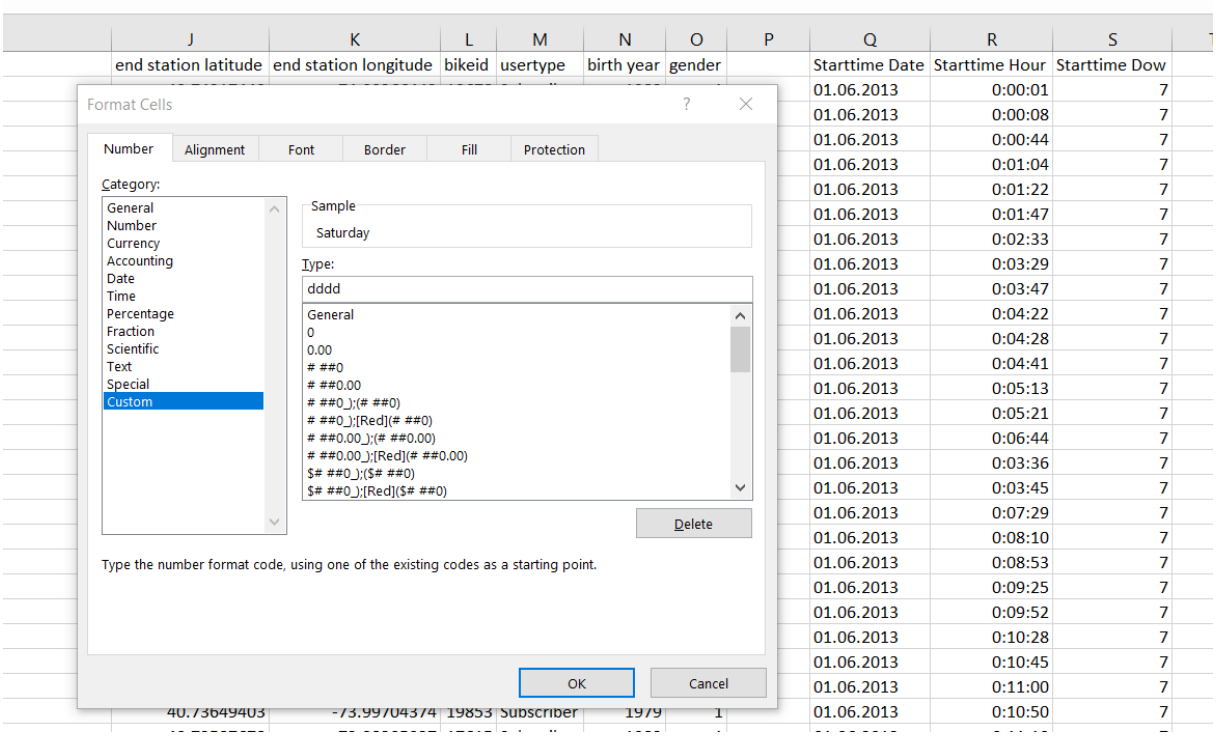



Figure 3: Using “format cells” to replace 7 to Saturday.

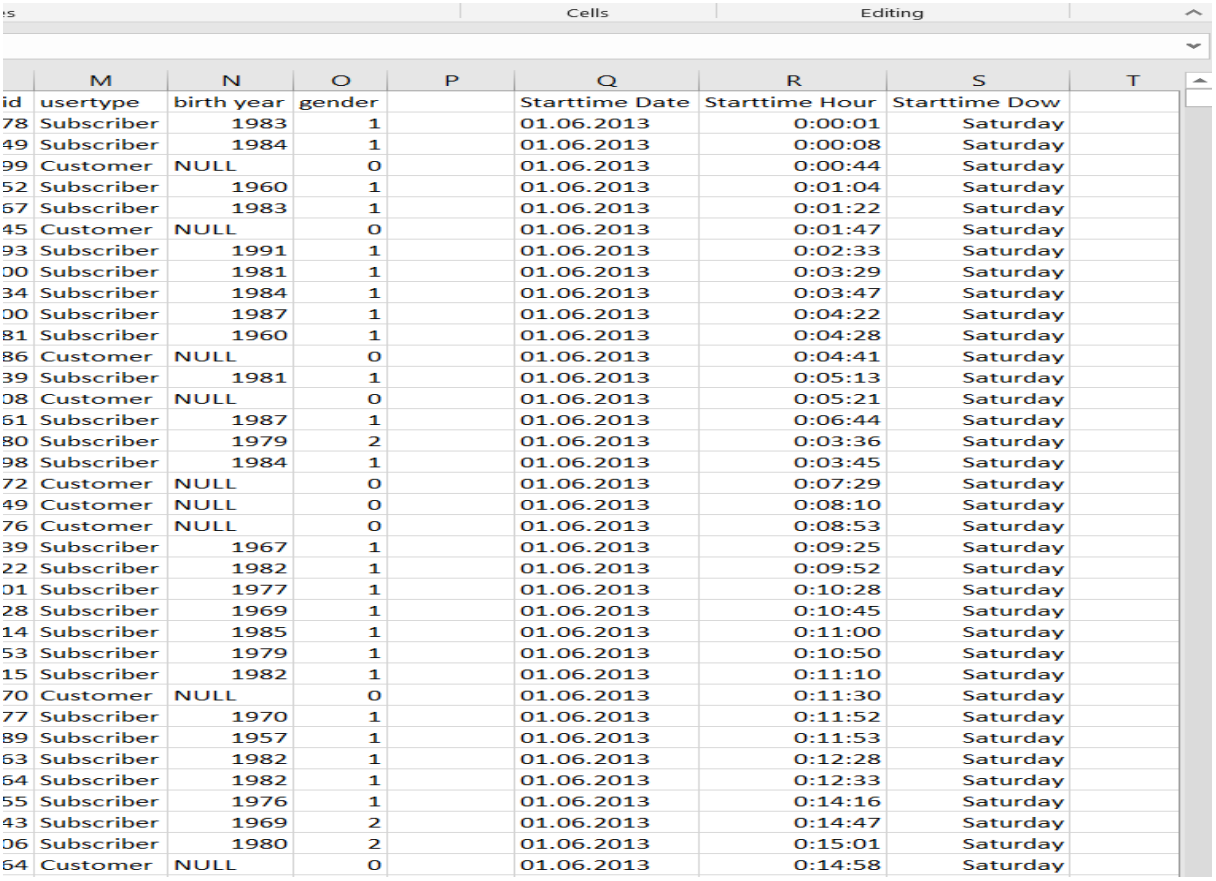



Figure 4: Saturday instead of 7.

**24hour clock:**

The hour on both column “starttime hour” and “stoptime hour” I chose to use a 24hour clock instead of using a.m. and p.m. I tried to use a.m. and p.m., but my excel for some reason got stuck every time I tried to use a.m./p.m., and my whole sheet got weird and just kept loading. But from the information on Kaggle, it seems that they used a 24hour clock on their data.

**Assignment 2:**

To look at which day of the week rentals are most popular and least popular, I used a chart to visualize a total of bike rentals from 01.06.2013 to 01.07.2013. To find the values, I used =COUNTIF.

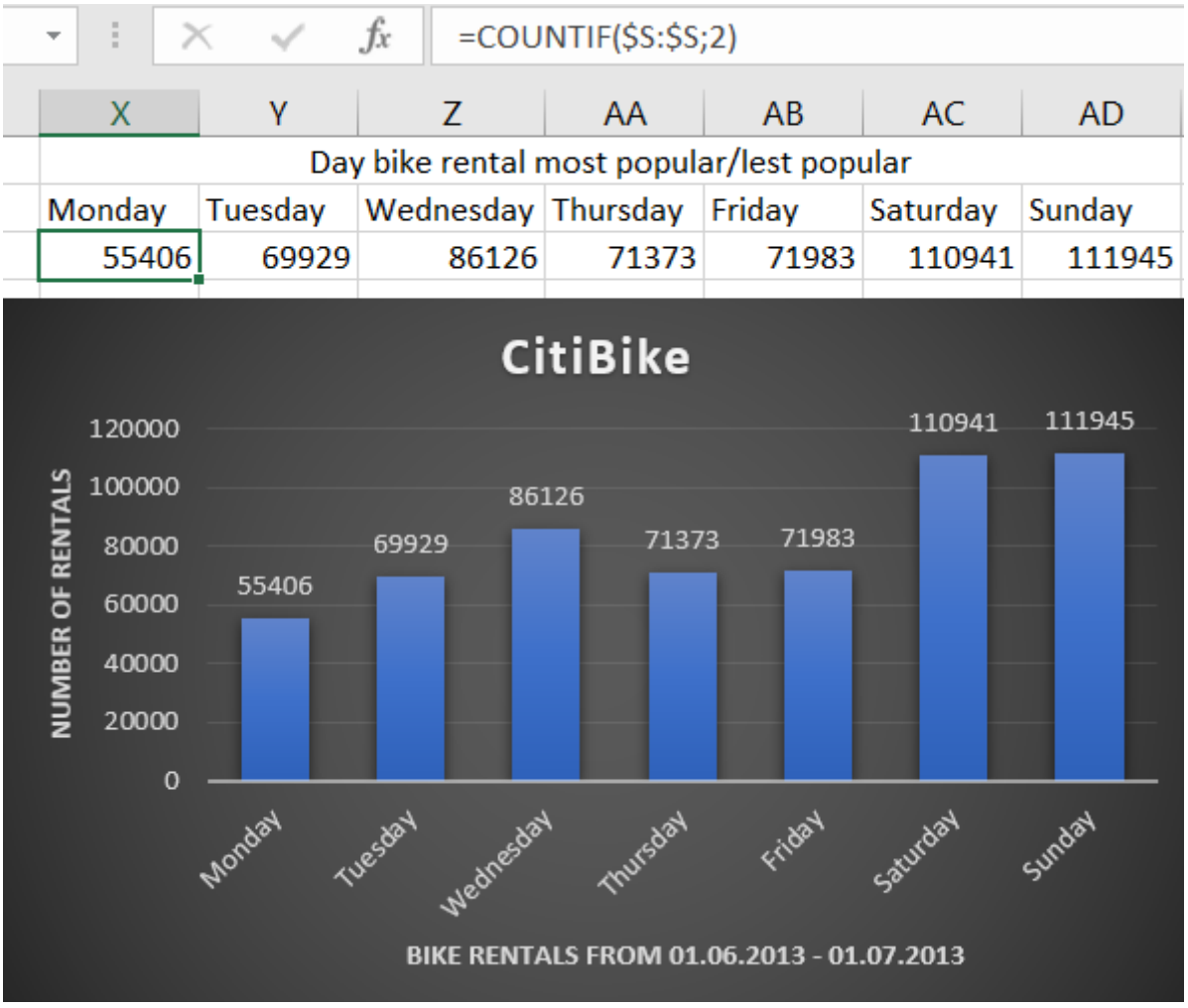



Figure 5: Rentals most popular/ least popular from 01.06.13 to 01.07.13.

As the chart shows:

- Most popular days for renting a bike is in the weekends, Saturdays and Sundays. It could be because people regularly have the day off from work/school, and that people has more time to do recreational purposes or visiting and meeting friends and family.
- Least popular days for renting a bike is in the start of a week, Mondays and Tuesdays. It could be because people are at work or school. It could also mean that people has done

recreational activities in the weekend and is likely not that active on Mondays and Tuesdays. They rather want to stay at home.

The average number of rented bikes on weekends compare to the average number of rented bikes during working days is shown below by using formula =average. The average number from Monday to Friday is 70963 rented bikes. On weekends from Saturday to Sunday the average rented bikes are 111443.

- Though there are only two days in a weekend, the average amount of people renting bikes is still higher than in working days which has five days total.

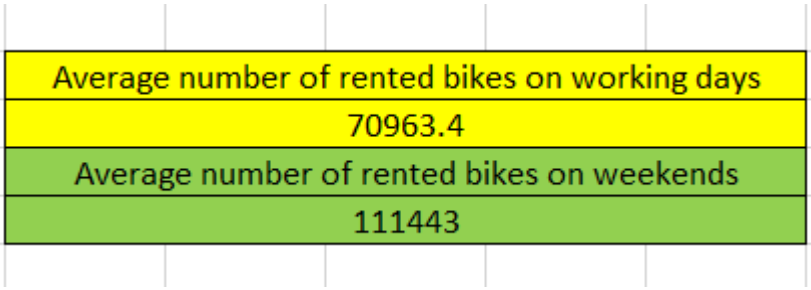



Figure 6: Compared average rented bikes on weekends (Saturdays to Sundays) vs. working days (from Mondays to Fridays)

Now, which hour in a day is the most popular and the least popular time to rent a bike in New York City from 01.06.13 to 01.07.13? To find that out, I created a chart using bins with a duration of 2 hours between each graph (except the first and last hour in a day). I used formula =FREQUENCY($R$2:$R$577704;AB22:AB34).

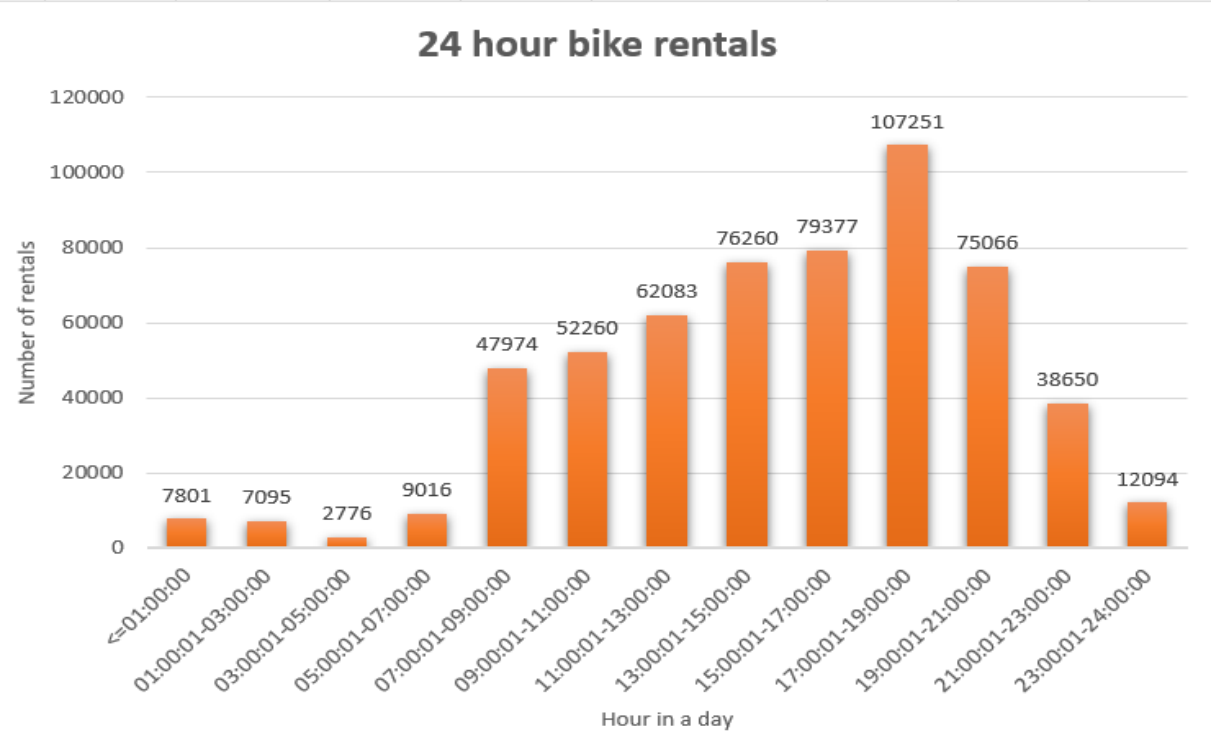
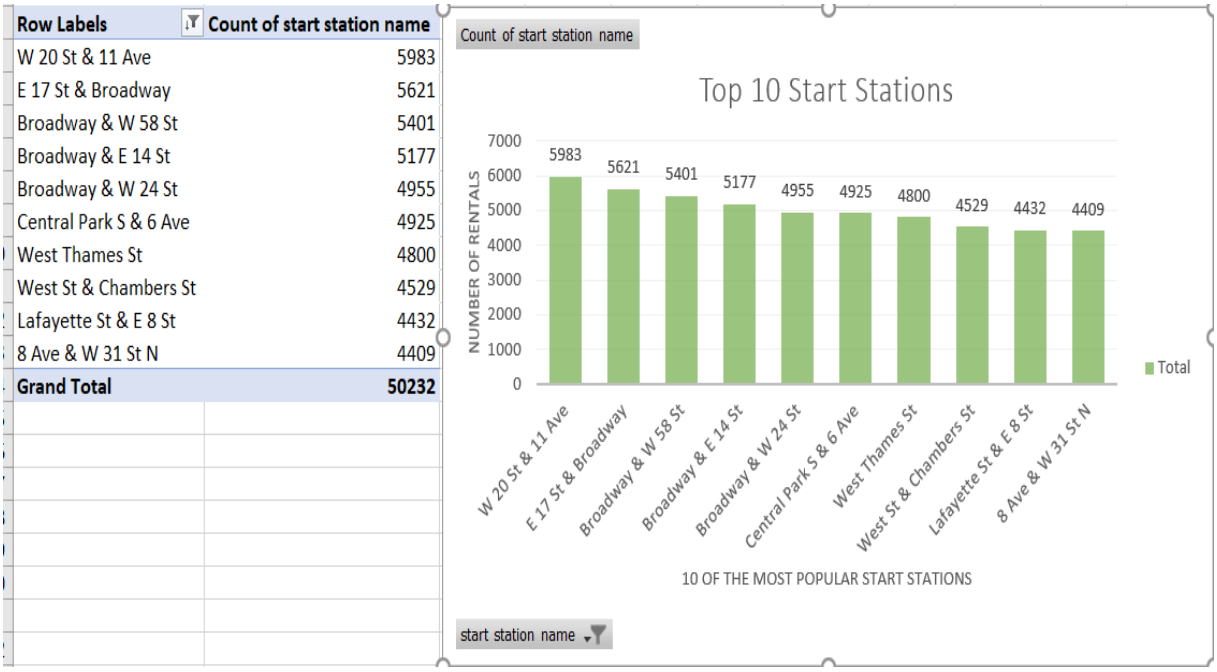



Figure 7: Showing time of day rental bikes are used.

The chart shows:

- The most popular time people renting bikes was between 17:00:01 to 19:00:00 in the evening with a total of 107251 rental bikes. This could be the time when people are finish at work to either travel home or meet family and friend etc.
- The least popular time was between 03:00:01 to 05:00:00 in the morning with a total of 2776 rental bikes. This explanation could be that people normally sleeps at that time of day.

The most popular start stations for bike rentals is shown below. I used a PivotTable to find the top 10 most popular start stations.



| Row Labels | Count of start station name |
|---|---|
| W 20 St & 11 Ave | 5983 |
| E 17 St & Broadway | 5621 |
| Broadway & W 58 St | 5401 |
| Broadway & E 14 St | 5177 |
| Broadway & W 24 St | 4955 |
| Central Park S & 6 Ave | 4925 |
| West Thames St | 4800 |
| West St & Chambers St | 4529 |
| Lafayette St & E 8 St | 4432 |
| 8 Ave & W 31 St N | 4409 |
| **Grand Total** | **50232** |



Figure 8: Top 10 start stations

The least popular start stations for bike rentals is also shown below.

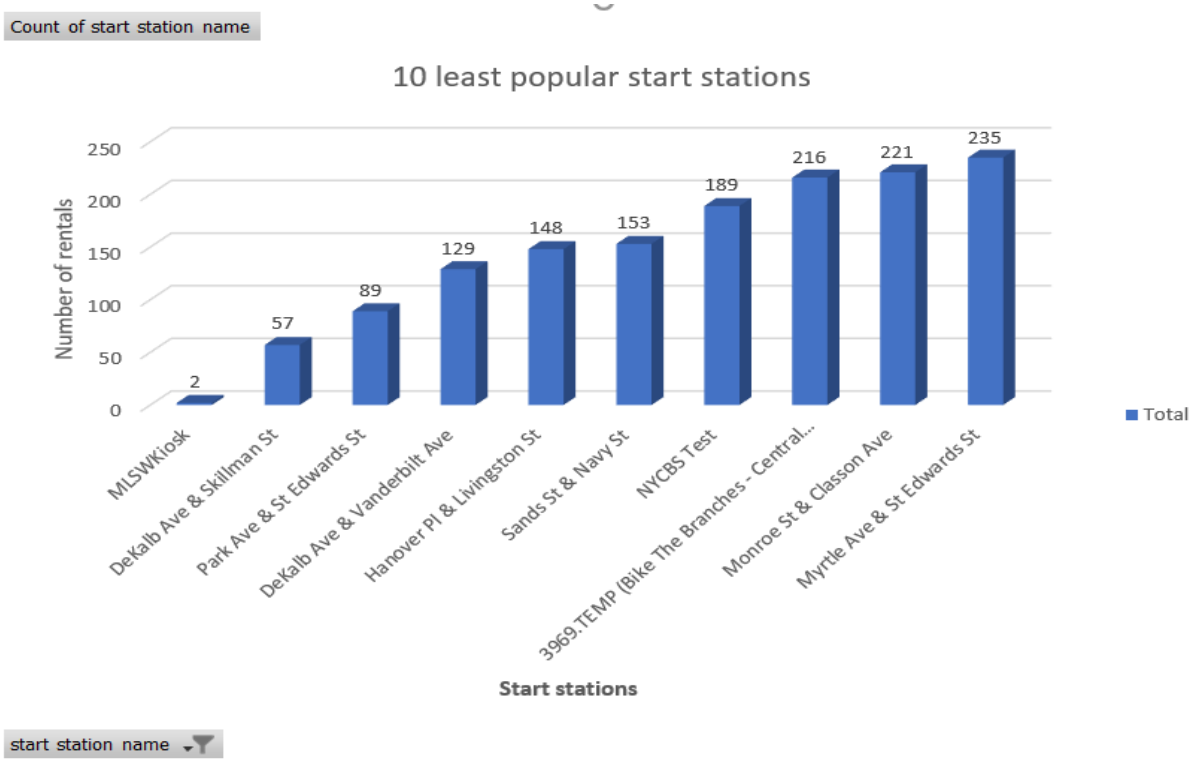



Figure 9: 10 least popular start stations for bike rentals

Figure 8 and figure 9 shows:

- The most popular start station to rent a bike was at West 20 Street & 11 Avenue with 5983 between 01.06.13 to 01.07.13.
- The least popular start station to rent a bike was at MLSWKiosk with only 2 bikes between 01.06.13 to 01.07.13.

It's quite the difference between the most and the least. Maybe the area in West 20 Street & 11 Avenue is a well-known station to rent a bike. Maybe it's located in an area where other transportations meet, like a bus station or a boat pier. It could be a lot of reasons. Why the MLSWKiosk has a low rental result in that same period could be that the location is at an area where most people ether rides the bus, subway, car etc. Maybe the kiosk is located in a far-off area.

The most popular destinations for people how rented bikes are shown below:

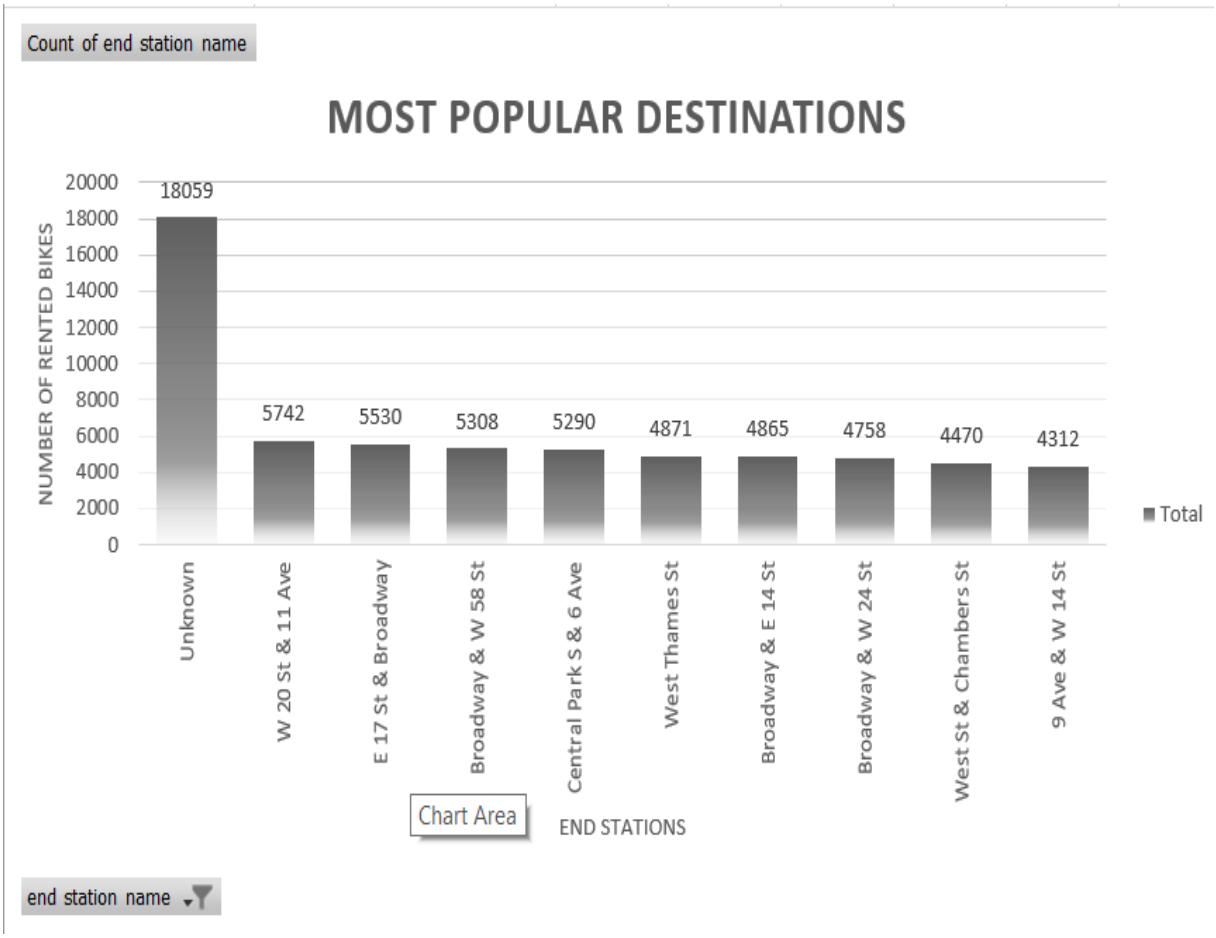



Figure 10: Most popular destination for rental bikes

What the chart shows:

- Destination “Unknown” shows, by far, the most frequent value in the chart. It can be that most of the time people how rents bikes don’t travel to an end station, but rather leave the bikes in unknown places. I hope CitiBike in New York had some kind of tracking devices on their bikes in 2013. If not, this could be costly business to run if 18059 bikes went missing!
- The most popular destination, or end station was West 20 street & 11 Avenue with 5742 rented bikes. It’s the most popular destination and the most popular start station (as shown in figure 10).
- The second most popular destination was East 17 & Broadway with 5530 rented bikes. The same address was also the second most popular start station (as shown in figure 10).
- As I said earlier, the reason why I chose to have the missing values in the data “unknown” is because it maybe has a logical meaning.

The least popular destination is shown below:

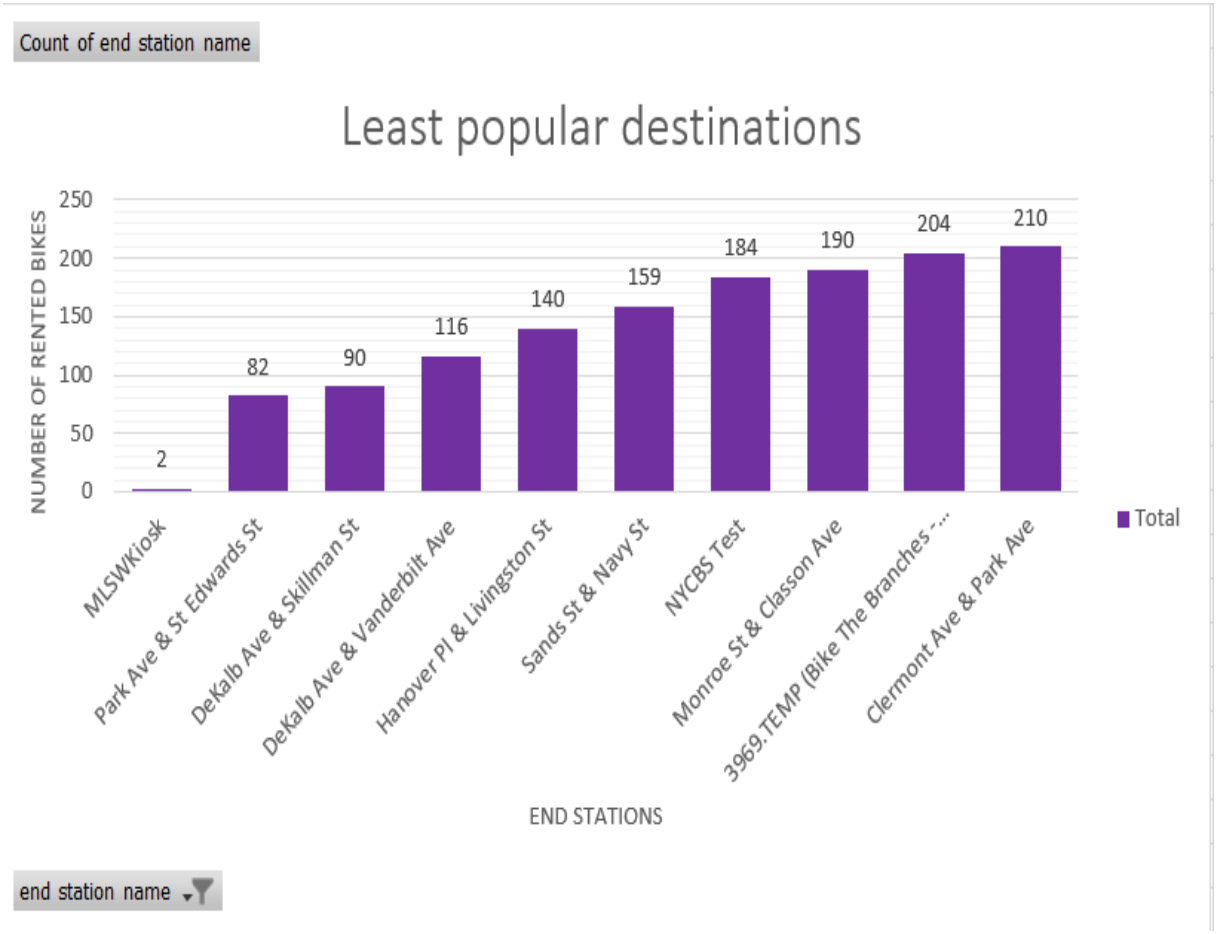



Figure 11: Top 10 least popular destinations between 01.06.13 to 01.07.13.

What the chart shows:

- MLSWKiosk was the least popular destination for people renting bikes with a value of 2. Like in figure 9, MLSWKiosk was the least popular start station with also 2 bikes from 01.06.13 to 01.07.13.
- Park Ave & St. Edwards Street was the second least popular destination, followed by DeKalb Avenue & Skillman Street. The same address was the second least popular start station to rent a bike, as shown in figure 10.

Are start stations and end stations usually the same or different?

That question was hard to figure out. After getting a tip from a fellow student in Discord, I used the station IDs to see if the ID numbers starts and ends matches between the same rows on the sheet. I'm not quite sure if I understood the question completely, but I figured out a solution:

- To find both matching values and not matching values in the same rows I used the =IF function as shown below.

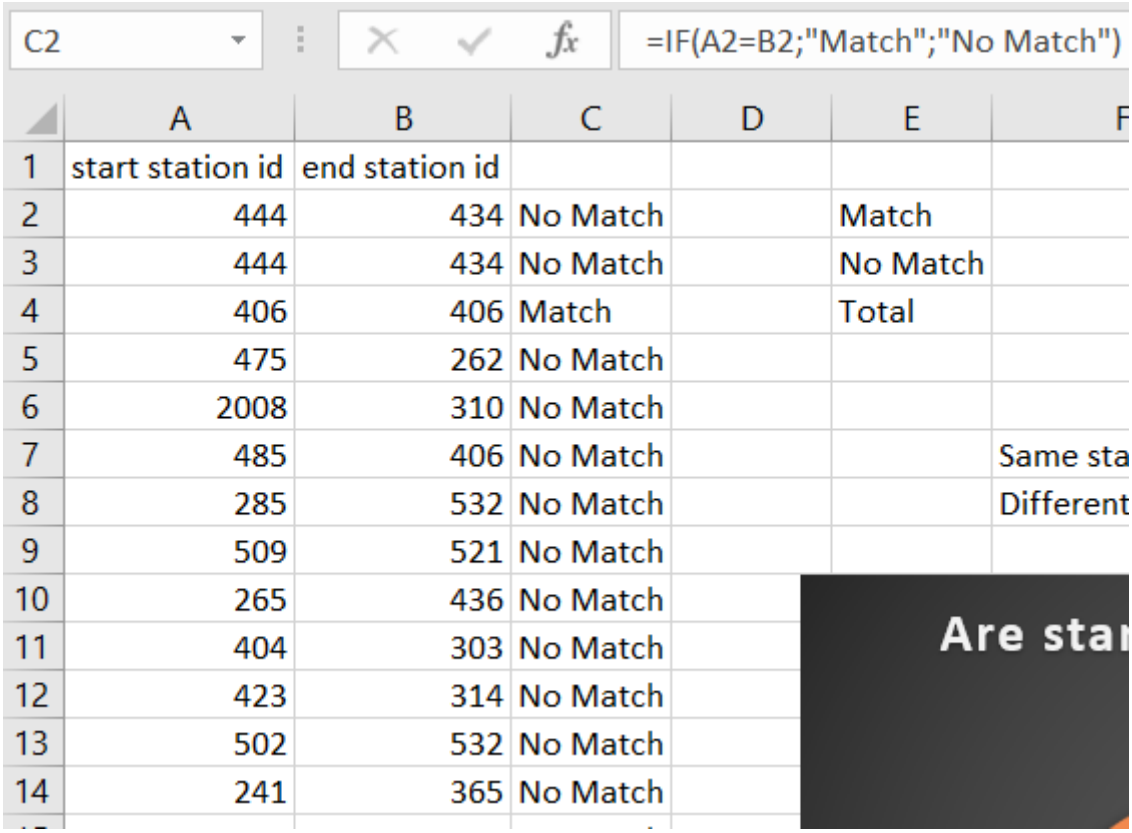



Figure 12: Finding if start stations and end stations are the same or different using =IF

- After finding thousands of "Match" and "No Match" inside column A and B, I then summarized "Match" and "No Match" by using the =COUNTIF function as shown below:

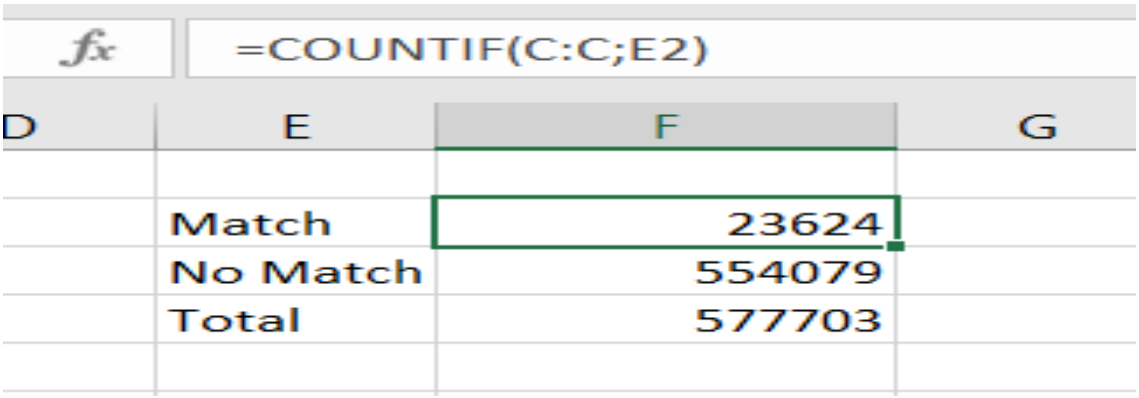



Figure 13: Using =COUNTIF to sum "Match" and "No Match".

- So now I got my values that I was looking for. The word "Match" got replaced with "Same station" and "No Match" got replaced with "Different station/unknown".
- Then I made a chart showing if the destination for people renting bikes ended at the same place as the starting station, or if the destination ended in a different station/unknown location.

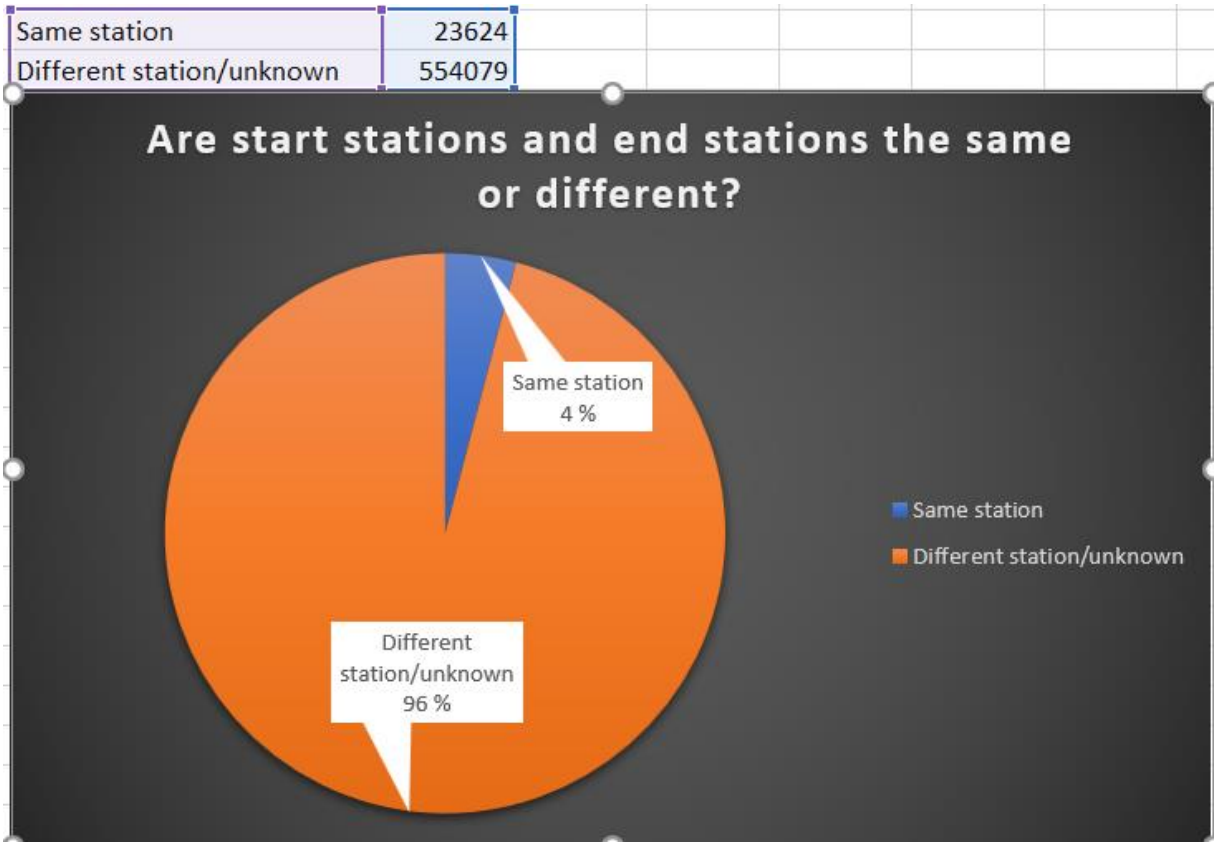
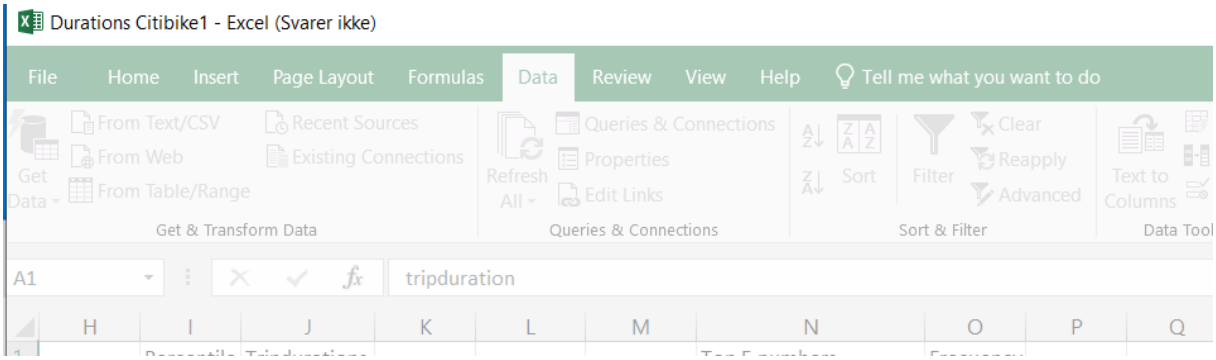



Figure 14: Showing if the destination for people renting bikes ended at the same place as the starting station, or if the destination ended in a different station/unknown location in %.

What the pie chart is showing:

- 96 % of the people who rented bikes usually did not end their destination in the same station where they started from. The bikes ended in other stations or in an unknown area.
- Only 4 % of trips with rented bikes ended at the same station where the bikes started from.

**Note:** *"In this point of assignment my excel started to load a lot and sometimes crashed. I tried to delete all rows and columns that where empty, even changed to a new workbook and copy/paste single columns for analysis, but it kept crashing. Instead of copying charts directly from excel to word, I took a screenshot of every chart and copied it in word. I spent hours of valuable time waiting, fixing, finding ways to make excel run better. Sometimes I had to do calculations all over again. I also had a hard time using the Data Analysis Toolpak because it just kept loading and crashing when I tried to find some results of the data using the tool".*




This "scene" (picture above) became more and more frequent.

**Assignment 3:**

By the help of the Data Analysis Toolpak in excel I got (finally….) some important values to use in this assignment:

| C | D |
|---|---|
| *tripduration* | |
| | |
| Mean | 1372.570279 |
| Standard Error | 11.87797129 |
| Median | 874 |
| Mode | 547 |
| Standard Deviation | 9028.063033 |
| Sample Variance | 81505922.12 |
| Kurtosis | 78921.68157 |
| Skewness | 224.0203653 |
| Range | 3876418 |
| Minimum | 61 |
| Maximum | 3876479 |
| Sum | 792937968 |
| Count | 577703 |
| | |

Figure 15: Using the Data Analysis Toolpak in excel.

- The values calculated by the Toolpak already shows me that the values inside column "trip durations" has some extreme variations.
- Sample variance has a variability of 81505922. This tells me the degree of spread in the data. The variance is so different in the relation to the mean value of 1372,57.
- The standard deviation is quite high. High standard deviation means that the values in the column are spread out.
- The range between minimum and maximum has a range of 3876418.

- Skewness looks like its skewed to the right/positive.
- As I can remember, if the mean is greater than the median, the distribution of the "trip durations" column has to be skewed positively/right side.
- The most frequent value in the column is 547. This could mean that most rented bike duration has a duration of 547 seconds. What I mean about that is that people who rent bikes has an average use of 547 seconds. Said differently, the average trip lasts about 9 minutes.

It looks like the data shows some extreme outliers. I think it would be difficult to make a histogram based on the extreme outliers in the "trip durations" column.

**Outliers:**

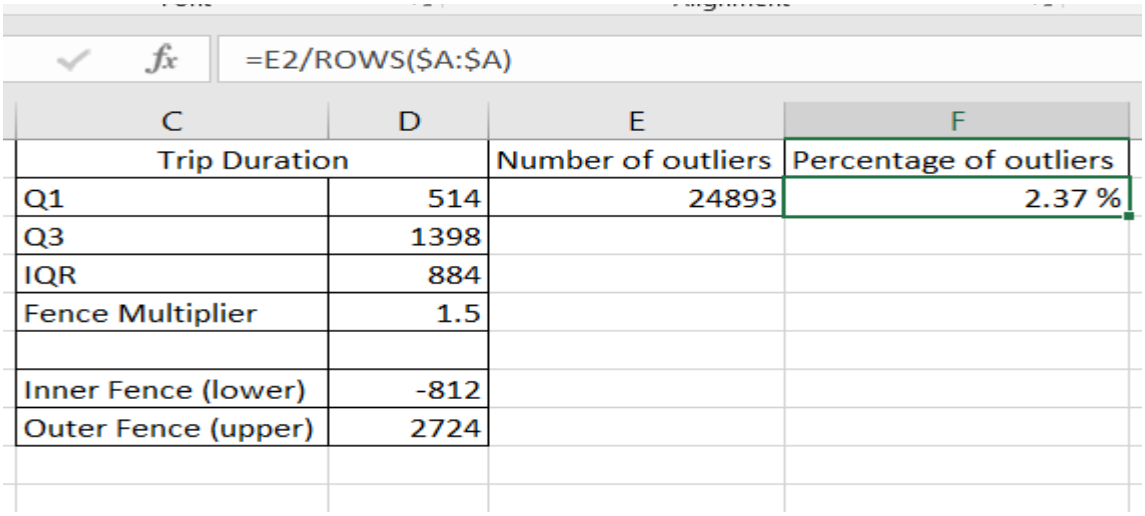
Investigating on the outliers, I checked the "Custom Sort" and got this information:

| Q | R | S | U | V |
|---|---|---|---|---|
| tripduration | Starttime Date | Starttime Hour | Stoptime Date | Stoptime Hour |
| 3876479 | 04.06.2013 | 04:51:57 | 19.07.2013 | 01:39:56 |
| 2742506 | 02.06.2013 | 13:11:18 | 04.07.2013 | 06:59:44 |
| 1860069 | 19.06.2013 | 12:32:51 | 11.07.2013 | 01:14:00 |
| 1331178 | 05.06.2013 | 11:42:36 | 20.06.2013 | 21:28:54 |
| 1239509 | 23.06.2013 | 20:36:50 | 08.07.2013 | 04:55:19 |
| 1048424 | 14.06.2013 | 15:40:53 | 26.06.2013 | 18:54:37 |
| 865531 | 19.06.2013 | 18:49:01 | 29.06.2013 | 19:14:32 |
| 764344 | 12.06.2013 | 14:47:54 | 21.06.2013 | 11:06:58 |
| 661180 | 16.06.2013 | 19:17:53 | 24.06.2013 | 10:57:33 |
| 625137 | 13.06.2013 | 18:14:29 | 20.06.2013 | 23:53:26 |
| 606864 | 16.06.2013 | 13:39:29 | 23.06.2013 | 14:13:53 |
| 576471 | 03.06.2013 | 19:49:15 | 10.06.2013 | 11:57:06 |
| 565786 | 05.06.2013 | 08:07:58 | 11.06.2013 | 21:17:44 |
| 563884 | 23.06.2013 | 06:14:17 | 29.06.2013 | 18:52:21 |
| 531099 | 08.06.2013 | 16:19:05 | 14.06.2013 | 19:50:44 |
| 528411 | 16.06.2013 | 19:01:24 | 22.06.2013 | 21:48:15 |
| 521552 | 22.06.2013 | 18:13:29 | 28.06.2013 | 19:06:01 |
| 504367 | 08.06.2013 | 22:17:05 | 14.06.2013 | 18:23:12 |
| 451931 | 09.06.2013 | 11:45:18 | 14.06.2013 | 17:17:29 |
| 394499 | 15.06.2013 | 21:27:47 | 20.06.2013 | 11:02:46 |
| 354318 | 05.06.2013 | 12:41:59 | 09.06.2013 | 15:07:17 |
| 340149 | 03.06.2013 | 22:53:28 | 07.06.2013 | 21:22:37 |
| 328384 | 02.06.2013 | 20:16:37 | 06.06.2013 | 15:29:41 |
| 315372 | 09.06.2013 | 12:39:46 | 13.06.2013 | 04:15:58 |
| 302588 | 05.06.2013 | 14:12:10 | 09.06.2013 | 02:15:18 |
| 295345 | 05.06.2013 | 07:01:37 | 08.06.2013 | 17:04:02 |

Figure 16: Checking for outliers

- Trip duration 3876479 on column Q: There is one individual who rented a bike from 04.06.13 to 19.07.13. That's 46 days.
- The second longest rented bike was from 02.06.13 to 04.07.13. That's 33 days.
- That's a long way from the average rented bike duration of 547 seconds (9 minutes).

Using the IQR method:



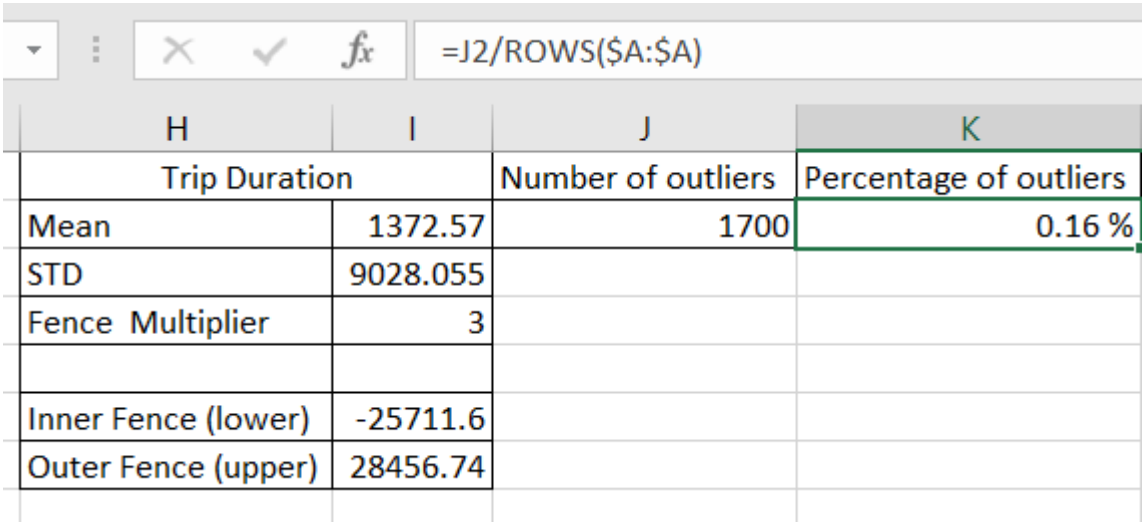

| C | D | E | F |
|---|---|---|---|
| Trip Duration | | Number of outliers | Percentage of outliers |
| Q1 | 514 | 24893 | 2.37 % |
| Q3 | 1398 | | |
| IQR | 884 | | |
| Fence Multiplier | 1.5 | | |
| | | | |
| Inner Fence (lower) | -812 | | |
| Outer Fence (upper) | 2724 | | |

Figure 17: Detecting outliers using IQR

Figure 17 shows:

- The IQR method detected 24893 outliers in column “trip duration”.
- The number of outliers is just 2.37 % of the whole data in the column.

Using the z-score method:




| H | I | J | K |
|---|---|---|---|
| Trip Duration | | Number of outliers | Percentage of outliers |
| Mean | 1372.57 | 1700 | 0.16 % |
| STD | 9028.055 | | |
| Fence Multiplier | 3 | | |
| | | | |
| Inner Fence (lower) | -25711.6 | | |
| Outer Fence (upper) | 28456.74 | | |

Figure 18: Detecting outliers using z-score

Figure 18 shows:

- The z-score method detected 1700 outliers in column “trip duration”.
- The number of outliers is just 0.16 % of the whole data in the column.

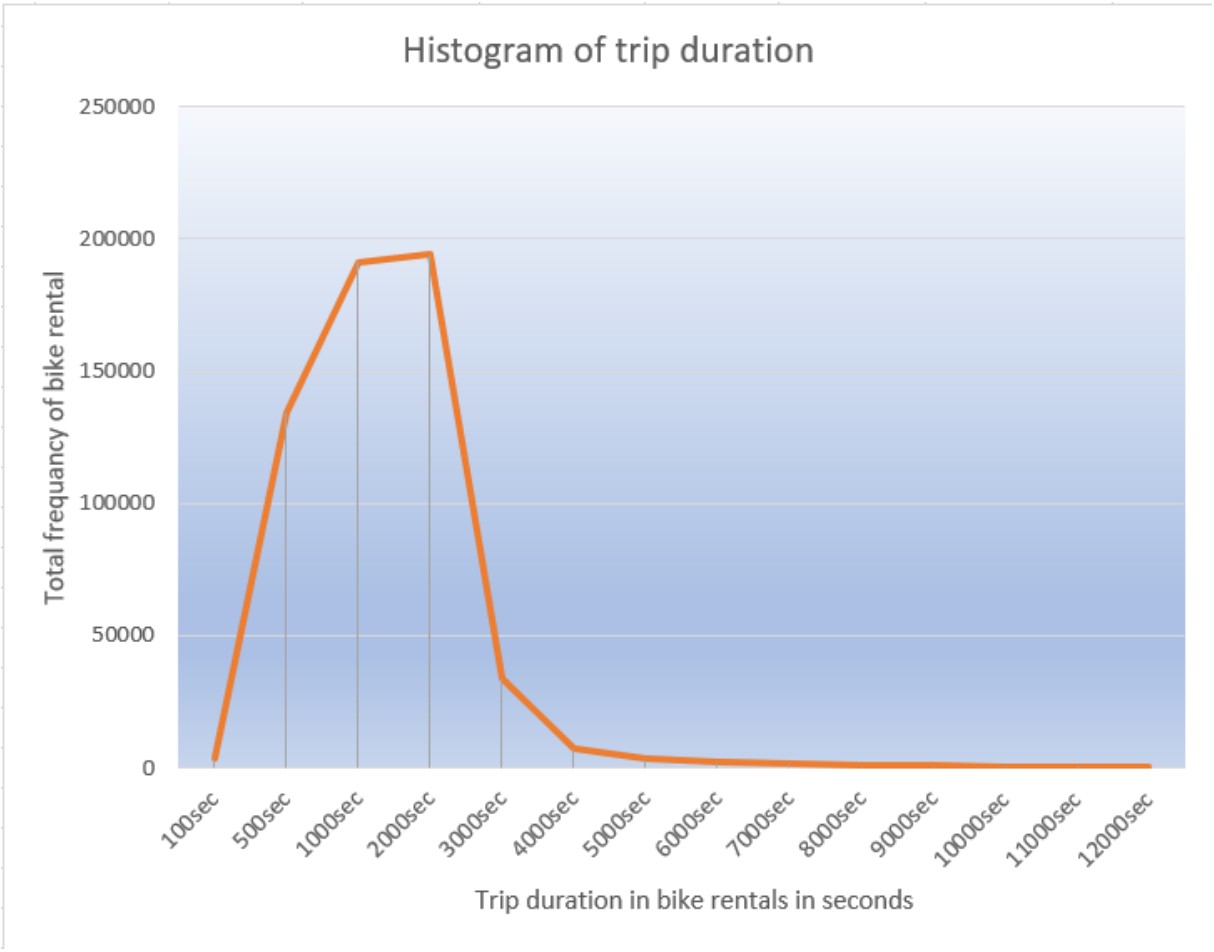
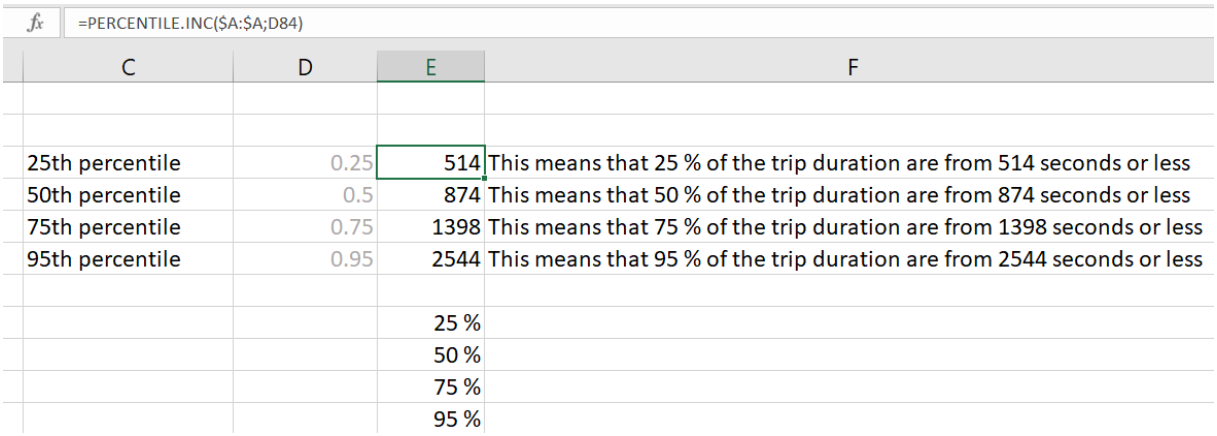



Figure 19: Histogram

The line chart shows:

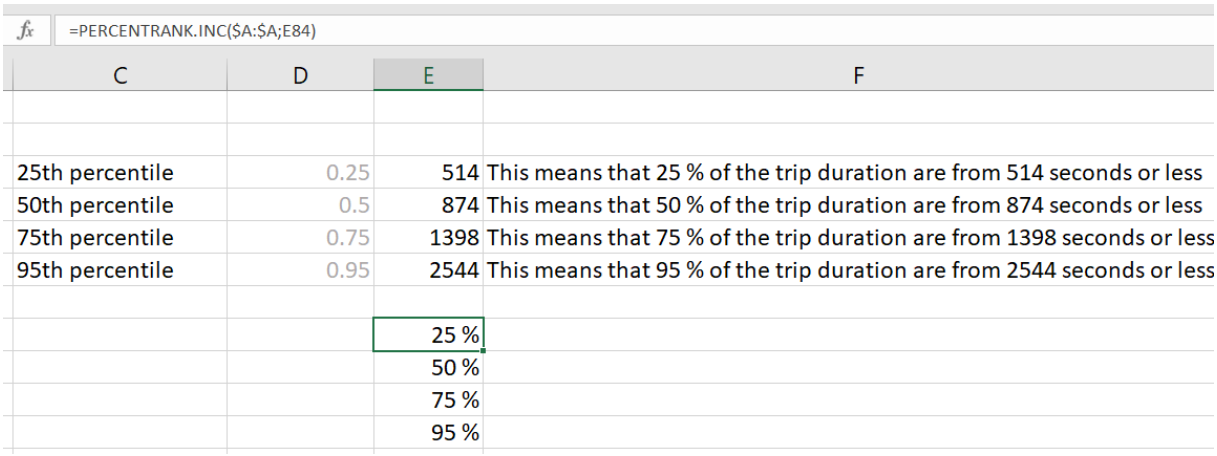
- People who rents bikes usually rents it between a duration of 100 seconds (1,5 minute) to 3000 seconds (50 minutes). This is the most used duration time out of 577703 times a bike was used/rented between 01.06.13 to 01.07.13.
- Histogram shows a right skewness
- The mean is 1372.57 and higher than median 874
- As the line shows, mean better represents the central tendency of the distribution of the chart.
- Using =percentile.inc like the figure below to show if the histogram matches the percentile.

Using this histogram was based on the data in figure 20 (next page), using percentile and finding out where the trip duration was most common.




| | C | D | E | F |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | 25th percentile | 0.25 | 514 | This means that 25 % of the trip duration are from 514 seconds or less |
| | 50th percentile | 0.5 | 874 | This means that 50 % of the trip duration are from 874 seconds or less |
| | 75th percentile | 0.75 | 1398 | This means that 75 % of the trip duration are from 1398 seconds or less |
| | 95th percentile | 0.95 | 2544 | This means that 95 % of the trip duration are from 2544 seconds or less |
| | | | | |
| | | | 25 % | |
| | | | 50 % | |
| | | | 75 % | |
| | | | 95 % | |

Figure 20: Percentile of 25th, 50th, 75th and 95th.

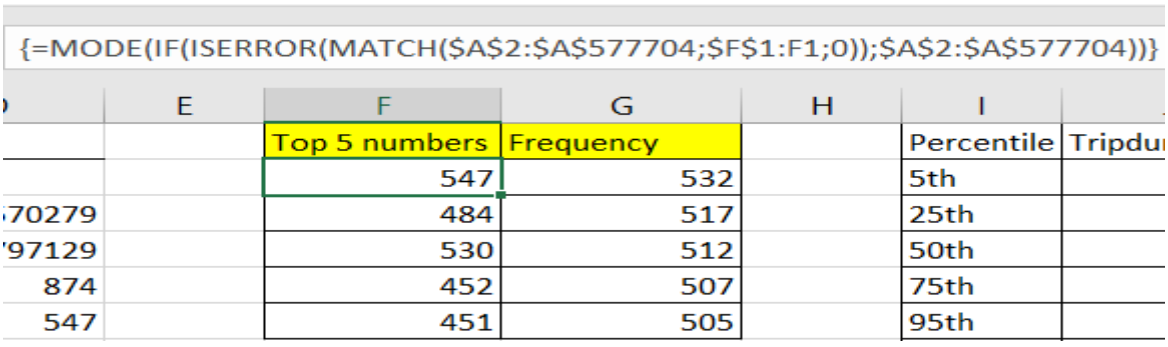



| | C | D | E | F |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | 25th percentile | 0.25 | 514 | This means that 25 % of the trip duration are from 514 seconds or less |
| | 50th percentile | 0.5 | 874 | This means that 50 % of the trip duration are from 874 seconds or less |
| | 75th percentile | 0.75 | 1398 | This means that 75 % of the trip duration are from 1398 seconds or less |
| | 95th percentile | 0.95 | 2544 | This means that 95 % of the trip duration are from 2544 seconds or less |
| | | | | |
| | | | 25 % | |
| | | | 50 % | |
| | | | 75 % | |
| | | | 95 % | |

Figure 21: Using =percentrank.inc to check if the formula and the value matches the %.

The most frequent numbers in column Trip Duration is shown below:

I used some formulas that didn't made the result I wanted, so I had to google it and used the formula in this site: https://www.exceldemy.com/how-to-find-most-frequent-numbers-excel/. Before this I used formulas such as =mode.sngl, =isnumber, =max, =vlookup etc. After finding the values that I was looking for, I started making a chart to show the result.




| | E | F | G | H | I | |
|---|---|---|---|---|---|---|
| | | Top 5 numbers | Frequency | | Percentile | Tripdu |
| | | 547 | 532 | | 5th | |
| 70279 | | 484 | 517 | | 25th | |
| 97129 | | 530 | 512 | | 50th | |
| 874 | | 452 | 507 | | 75th | |
| 547 | | 451 | 505 | | 95th | |

Figure 20: Finding the most frequent trip durations on rented bikes.

Using the top 10 most frequent numbers in a chart:

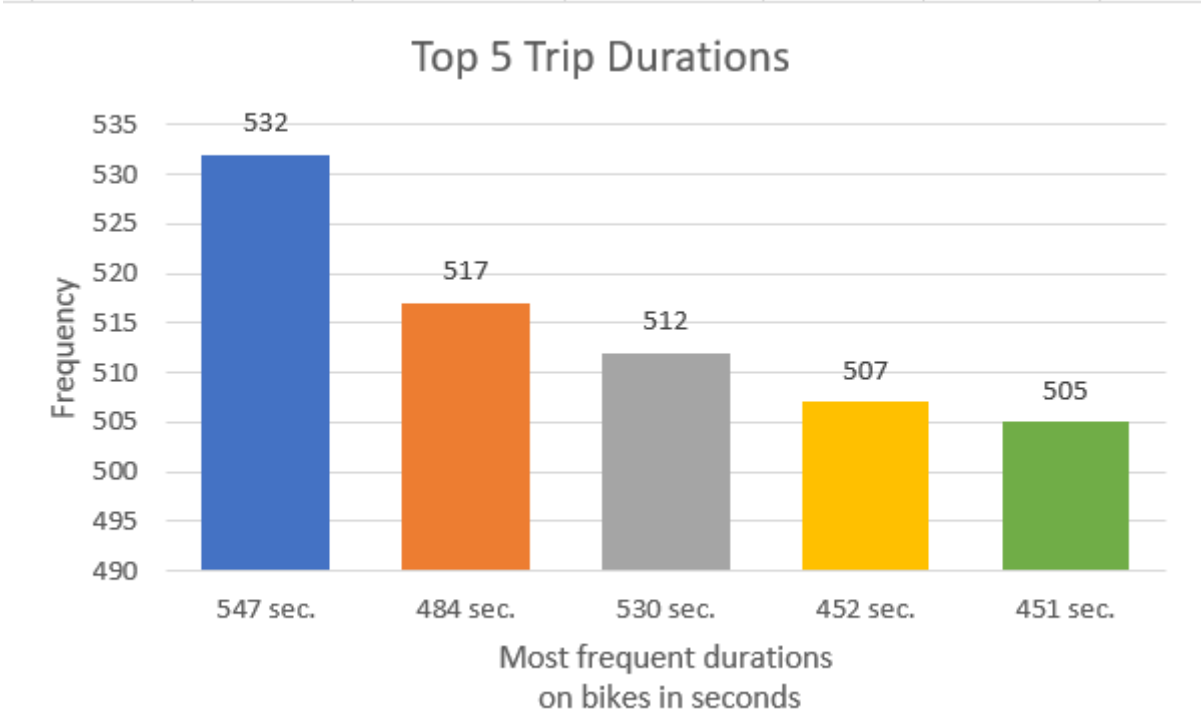



Figure 21: Top 5 trip durations

The chart shows:

- That using a rented bike for 547 seconds was the most common time a rental lasted.
- 532 people had the same amount of time spent renting a bike.

To understand this better, I could try to rather use minutes instead of seconds as shown below:

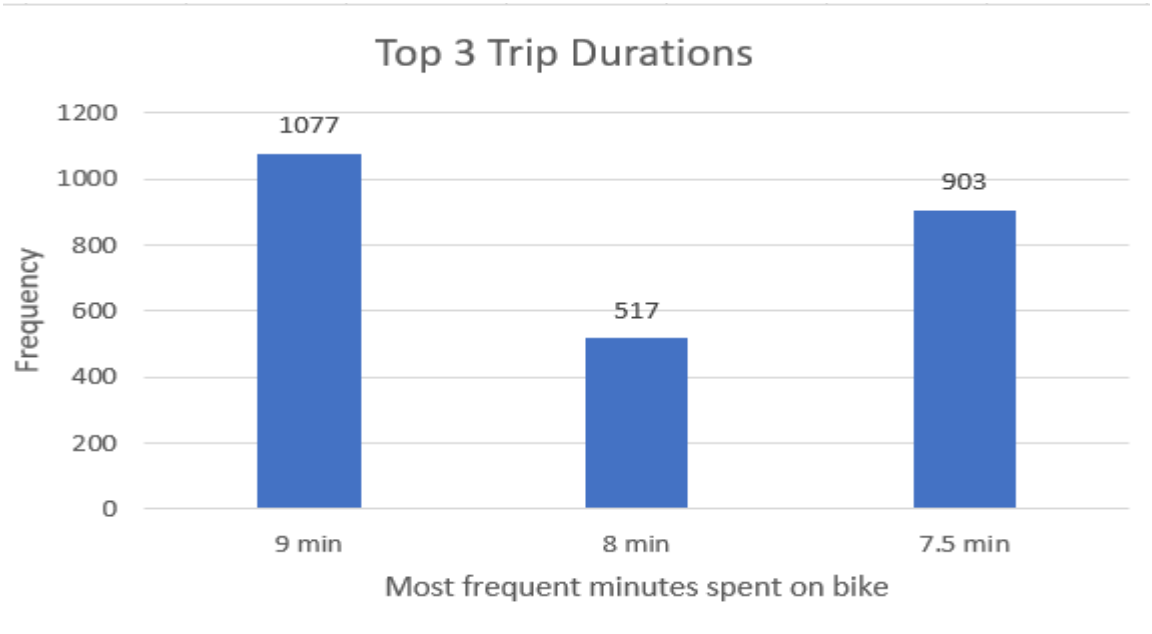


Figure 22: Simpler chart of most frequent minutes spent renting bikes

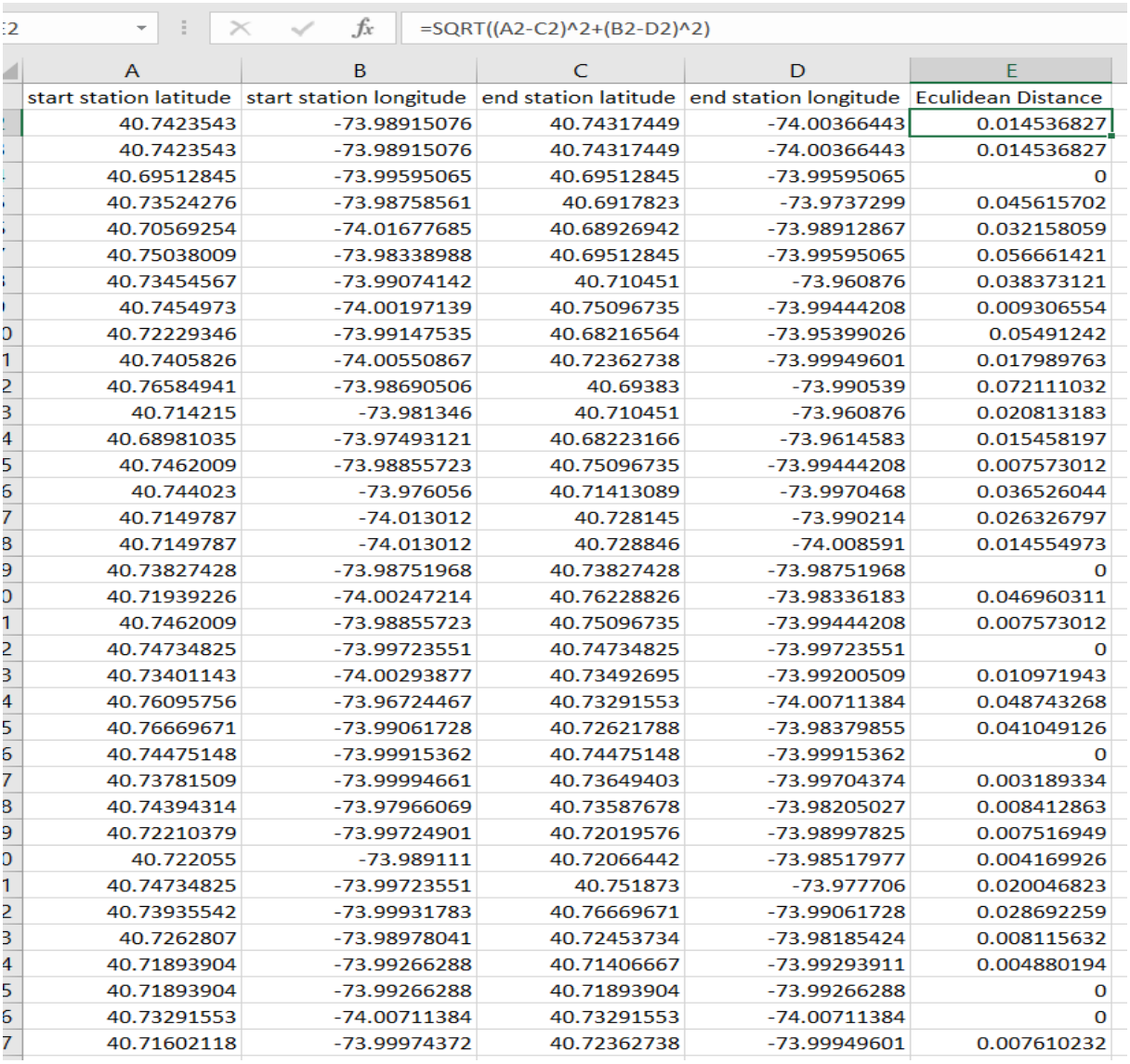
Note:

- The 9-minute bar is a combine value of the 1st and 3rd frequent duration in figure 21. The total is wrong and it should be 1044 total, not 1077.
- The 8-minute bar is the second bar in figure 21.
- The 7.5 minute bar is a combine value of the 2nd , 4th and 5th bar in figure 21.

**Assignment 4:**

Finding the Euclidean distance between column “start stations” to “end station”. This calculates the shortest possible distance between two locations.

- I made a new column called Euclidean distance:




| | A | B | C | D | E |
|---|---|---|---|---|---|
| | start station latitude | start station longitude | end station latitude | end station longitude | Eculidean Distance |
| 2 | 40.7423543 | -73.98915076 | 40.74317449 | -74.00366443 | 0.014536827 |
| 3 | 40.7423543 | -73.98915076 | 40.74317449 | -74.00366443 | 0.014536827 |
| 4 | 40.69512845 | -73.99595065 | 40.69512845 | -73.99595065 | 0 |
| 5 | 40.73524276 | -73.98758561 | 40.6917823 | -73.9737299 | 0.045615702 |
| 6 | 40.70569254 | -74.01677685 | 40.68926942 | -73.98912867 | 0.032158059 |
| 7 | 40.75038009 | -73.98338988 | 40.69512845 | -73.99595065 | 0.056661421 |
| 8 | 40.73454567 | -73.99074142 | 40.710451 | -73.960876 | 0.038373121 |
| 9 | 40.7454973 | -74.00197139 | 40.75096735 | -73.99444208 | 0.009306554 |
| 10 | 40.72229346 | -73.99147535 | 40.68216564 | -73.95399026 | 0.05491242 |
| 11 | 40.7405826 | -74.00550867 | 40.72362738 | -73.99949601 | 0.017989763 |
| 12 | 40.76584941 | -73.98690506 | 40.69383 | -73.990539 | 0.072111032 |
| 13 | 40.714215 | -73.981346 | 40.710451 | -73.960876 | 0.020813183 |
| 14 | 40.68981035 | -73.97493121 | 40.68223166 | -73.9614583 | 0.015458197 |
| 15 | 40.7462009 | -73.98855723 | 40.75096735 | -73.99444208 | 0.007573012 |
| 16 | 40.744023 | -73.976056 | 40.71413089 | -73.9970468 | 0.036526044 |
| 17 | 40.7149787 | -74.013012 | 40.728145 | -73.990214 | 0.026326797 |
| 18 | 40.7149787 | -74.013012 | 40.728846 | -74.008591 | 0.014554973 |
| 19 | 40.73827428 | -73.98751968 | 40.73827428 | -73.98751968 | 0 |
| 20 | 40.71939226 | -74.00247214 | 40.76228826 | -73.98336183 | 0.046960311 |
| 21 | 40.7462009 | -73.98855723 | 40.75096735 | -73.99444208 | 0.007573012 |
| 22 | 40.74734825 | -73.99723551 | 40.74734825 | -73.99723551 | 0 |
| 23 | 40.73401143 | -74.00293877 | 40.73492695 | -73.99200509 | 0.010971943 |
| 24 | 40.76095756 | -73.96724467 | 40.73291553 | -74.00711384 | 0.048743268 |
| 25 | 40.76669671 | -73.99061728 | 40.72621788 | -73.98379855 | 0.041049126 |
| 26 | 40.74475148 | -73.99915362 | 40.74475148 | -73.99915362 | 0 |
| 27 | 40.73781509 | -73.99994661 | 40.73649403 | -73.99704374 | 0.003189334 |
| 28 | 40.74394314 | -73.97966069 | 40.73587678 | -73.98205027 | 0.008412863 |
| 29 | 40.72210379 | -73.99724901 | 40.72019576 | -73.98997825 | 0.007516949 |
| 30 | 40.722055 | -73.989111 | 40.72066442 | -73.98517977 | 0.004169926 |
| 31 | 40.74734825 | -73.99723551 | 40.751873 | -73.977706 | 0.020046823 |
| 32 | 40.73935542 | -73.99931783 | 40.76669671 | -73.99061728 | 0.028692259 |
| 33 | 40.7262807 | -73.98978041 | 40.72453734 | -73.98185424 | 0.008115632 |
| 34 | 40.71893904 | -73.99266288 | 40.71406667 | -73.99293911 | 0.004880194 |
| 35 | 40.71893904 | -73.99266288 | 40.71893904 | -73.99266288 | 0 |
| 36 | 40.73291553 | -74.00711384 | 40.73291553 | -74.00711384 | 0 |
| 37 | 40.71602118 | -73.99974372 | 40.72362738 | -73.99949601 | 0.007610232 |

Figure 23: Finding the Euclidean distance using formula =SQRT.

NOTE: I had to make a new sheet in fear of my excel crashing again. If not, I would have made a new column inside the sheet 201306-citibike-tripdata.

NOTE: When calculating the Euclidean and Manhattan distance I noticed that there where data that could affect assignment 5.

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| | 40.69475701 | -73.99052739 | 40.72938685 | -73.97772429 | 0.036920796 | 0.04743294 |
| | 40.71911552 | -74.00666661 | 40.7262807 | -73.98978041 | 0.018343488 | 0.02405138 |
| | 40.73221853 | -73.98165557 | 40.75455731 | -73.96592976 | 0.027318898 | 0.03806459 |
| | 40.71260486 | -73.96264403 | 40.722055 | -73.989111 | 0.028103481 | 0.03591711 |
| | 40.7262807 | -73.98978041 | 40.73221853 | -73.98165557 | 0.010063342 | 0.01406267 |
| | 40.69089272 | -73.99612349 | 40.69165183 | -73.9999786 | 0.003929137 | 0.00461422 |
| | 40.69239502 | -73.99337909 | 40.68382604 | -73.97632328 | 0.01908738 | 0.02562479 |
| | 40.69089272 | -73.99612349 | 40.69165183 | -73.9999786 | 0.003929137 | 0.00461422 |
| | 40.756014 | -73.967416 | 40.735238 | -74.000271 | 0.038872782 | 0.053631 |
| | 40.7284186 | -73.98713956 | 40.74780373 | -73.9734419 | 0.023736241 | 0.03308279 |
| | 40.73935542 | -73.99931783 | 0 | 0 | 84.47244592 | 114.7386733 |
| | 40.72362738 | -73.99949601 | 40.76370739 | -73.9851615 | 0.042566247 | 0.05441452 |
| | 40.75513557 | -73.98658032 | 40.73971301 | -73.99456405 | 0.017366499 | 0.02340629 |
| | 40.75513557 | -73.98658032 | 40.74195138 | -74.00803013 | 0.025177713 | 0.034634 |
| | 40.73261787 | -73.99158043 | 40.70569254 | -74.01677685 | 0.036875913 | 0.05212175 |
| | 40.69196035 | -73.96536851 | 40.69363137 | -73.96223558 | 0.003550712 | 0.00480395 |
| | 40.759107 | -73.959223 | 40.7462009 | -73.98855723 | 0.032047847 | 0.04224033 |
| | 40.7643971 | -73.97371465 | 40.73587678 | -73.98205027 | 0.029713485 | 0.03685594 |
| | 40.76341379 | -73.99667444 | 40.76269882 | -73.99301222 | 0.003731359 | 0.00437719 |
| | 40.71602118 | -73.99974372 | 40.71602118 | -73.99974372 | 0 | 0 |
| | 40.71406667 | -73.99293911 | 40.71625008 | -74.0091059 | 0.016313564 | 0.0183502 |
| | 40.71406667 | -73.99293911 | 40.73971301 | -73.99456405 | 0.025697766 | 0.02727128 |
| | 40.68764484 | -73.96968902 | 40.68851534 | -73.9647628 | 0.005002541 | 0.00579672 |
| | 40.75019995 | -73.99093085 | 40.746745 | -74.007756 | 0.017176215 | 0.0202801 |
| | 40.72066442 | -73.98517977 | 40.71602118 | -73.99974372 | 0.015286213 | 0.01920719 |
| | 40.71260486 | -73.96264403 | 40.71260486 | -73.96264403 | 0 | 0 |
| | 40.72362738 | -73.99949601 | 40.72679454 | -73.99695094 | 0.004063039 | 0.00571223 |
| | 40.74290902 | -73.97706058 | 40.73587678 | -73.98205027 | 0.00862261 | 0.01202193 |
| | 40.74290902 | -73.97706058 | 0 | 0 | 84.45466315 | 114.7199696 |
| | 40.7172274 | -73.98802084 | 40.70862144 | -74.00722156 | 0.021041155 | 0.02780668 |
| | 40.72405549 | -74.00965965 | 40.71602118 | -73.99974372 | 0.012762281 | 0.01795024 |
| | 40.71260486 | -73.96264403 | 40.73291553 | -74.00711384 | 0.048888519 | 0.06478048 |
| | 40.69308257 | -73.97178913 | 40.71117416 | -74.00016545 | 0.033652952 | 0.04646791 |
| | 40.71625008 | -74.0091059 | 40.73543934 | -73.99453948 | 0.024091664 | 0.03375568 |
| | 40.74395411 | -73.99144871 | 40.76009437 | -73.99461843 | 0.01644856 | 0.01930998 |

Figure 24: Missing values 1.

The red line shows missing data from column C and D. That's missing data of "end station latitude" and "end station longitude". Without that data, the Euclidean and Manhattan distance values are useless. Because it doesn't show the true value of the distance.

Like in the yellow line, the start and end distance are identical. No wonder the Euclidean and Manhattan distances are at value 0, and the mode tells me it's the most frequent value in the data.

Let's look at some sheets including/adding some important columns so we can have more insights of the data. I'll go back to sheet 201306-citibike-tripdata.

| | H | I | J | K | L | Q | R | T | U |
|---|---|---|---|---|---|---|---|---|---|
| 1 | end station id | end station name | end station latitude | end station longitude | bikeid | Starttime Date | Starttime Hour | Stoptime Dat | Stoptime Hour |
| 58 | Unknown | Unknown | Unknown | Unknown | 17470 | 01.06.2013 | 00:11:04 | 01.06.2013 | 00:20:11 |
| 104 | Unknown | Unknown | Unknown | Unknown | 15090 | 01.06.2013 | 00:40:27 | 01.06.2013 | 00:49:24 |
| 122 | Unknown | Unknown | Unknown | Unknown | 16257 | 01.06.2013 | 00:47:51 | 01.06.2013 | 00:55:43 |
| 213 | Unknown | Unknown | Unknown | Unknown | 20106 | 01.06.2013 | 01:32:55 | 01.06.2013 | 01:35:28 |
| 291 | Unknown | Unknown | Unknown | Unknown | 18792 | 01.06.2013 | 02:28:10 | 01.06.2013 | 02:42:11 |
| 304 | Unknown | Unknown | Unknown | Unknown | 15752 | 01.06.2013 | 02:40:31 | 01.06.2013 | 02:55:35 |
| 317 | Unknown | Unknown | Unknown | Unknown | 17106 | 01.06.2013 | 02:54:35 | 01.06.2013 | 03:06:59 |
| 336 | Unknown | Unknown | Unknown | Unknown | 16461 | 01.06.2013 | 03:22:48 | 01.06.2013 | 03:38:02 |
| 356 | Unknown | Unknown | Unknown | Unknown | 17283 | 01.06.2013 | 03:52:23 | 01.06.2013 | 04:04:00 |
| 375 | Unknown | Unknown | Unknown | Unknown | 18327 | 01.06.2013 | 04:50:53 | 01.06.2013 | 05:01:13 |
| 451 | Unknown | Unknown | Unknown | Unknown | 19912 | 01.06.2013 | 07:21:49 | 01.06.2013 | 11:14:14 |
| 497 | Unknown | Unknown | Unknown | Unknown | 18188 | 01.06.2013 | 07:50:57 | 01.06.2013 | 08:05:36 |
| 500 | Unknown | Unknown | Unknown | Unknown | 17548 | 01.06.2013 | 07:51:49 | 01.06.2013 | 08:06:09 |
| 513 | Unknown | Unknown | Unknown | Unknown | 19315 | 01.06.2013 | 07:55:40 | 01.06.2013 | 08:18:44 |
| 522 | Unknown | Unknown | Unknown | Unknown | 15555 | 01.06.2013 | 07:59:35 | 01.06.2013 | 17:42:21 |
| 525 | Unknown | Unknown | Unknown | Unknown | 18257 | 01.06.2013 | 08:00:35 | 01.06.2013 | 08:05:46 |
| 539 | Unknown | Unknown | Unknown | Unknown | 16781 | 01.06.2013 | 08:08:05 | 01.06.2013 | 12:13:15 |
| 616 | Unknown | Unknown | Unknown | Unknown | 15760 | 01.06.2013 | 08:33:05 | 01.06.2013 | 08:37:33 |
| 630 | Unknown | Unknown | Unknown | Unknown | 15353 | 01.06.2013 | 08:38:57 | 01.06.2013 | 08:45:33 |
| 658 | Unknown | Unknown | Unknown | Unknown | 15721 | 01.06.2013 | 08:48:24 | 01.06.2013 | 09:02:36 |

Figure 25: Missing values 2.

The figure shows:

- That even though column “end station ID”, “end station name”, “end station latitude, and “end station longitude” are missing, the data still shows values in column “trip duration”, “start date”, “stop time date” etc., as shown in figure 25.
- I counted 18059 missing rows in column H, I, J and K.
- The missing data in columns H, I, J and K makes it difficult to calculate the distance. Data is useless when calculating the distance since there is no data that tells us where in New York the bike trip ended. We have only the date and time that shows us when the bike rent ended.

Let’s look at the gray line in figure 25 as an example:

- The person renting the bike from 07:59:35 in the morning till 17:42:21 in the afternoon/evening in 01.06.13, had a trip duration of 34966 sec. That’s over 9 hours bike rent. Still, we don’t know the distance of the trip.
- Could be that the trip didn’t ended in an end station, but some random place in the city.

I think that the missing values in distance should not be in the data in finding the values of the distance in assignment 5. The 18059 rows will not be included. Because as shown in figure 23 and 24, the data of distance will be misleading.

If we look again in figure 24, the yellow lines are also values that should not be in the data in finding the distance, since the person who rented the bike never left the start station. I’ll show some examples under:

| A | D | E | F | G | H | I | J | K | Q | R | T | U |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ripduration | start station id | start station name | start station latitude | start station longitude | end station id | end station name | end station latitude | end station longitude | Starttime Date | Starttime Hour | Stoptime Date | Stoptime Hour |
| 695 | 444 | Broadway & W 24 St | 40.7423543 | -73.98915076 | 434 | 9 Ave & W 18 St | 40.74317449 | -74.00366443 | 01.06.2013 | 00:00:01 | 01.06.2013 | 00:11:36 |
| 693 | 444 | Broadway & W 24 St | 40.7423543 | -73.98915076 | 434 | 9 Ave & W 18 St | 40.74317449 | -74.00366443 | 01.06.2013 | 00:00:08 | 01.06.2013 | 00:11:41 |
| 2059 | 406 | Hicks St & Montague St | 40.69512845 | -73.99595065 | 406 | Hicks St & Montague St | 40.69512845 | -73.99595065 | 01.06.2013 | 00:00:44 | 01.06.2013 | 00:35:03 |
| 123 | 475 | E 15 St & Irving Pl | 40.73524276 | -73.98758561 | 262 | Washington Park | 40.6917823 | -73.9737299 | 01.06.2013 | 00:01:04 | 01.06.2013 | 00:03:07 |
| 1521 | 2008 | Little West St & 1 Pl | 40.70569254 | -74.01677685 | 310 | State St & Smith St | 40.68926942 | -73.98912867 | 01.06.2013 | 00:01:22 | 01.06.2013 | 00:26:43 |
| 2028 | 485 | W 37 St & 5 Ave | 40.75038009 | -73.98338988 | 406 | Hicks St & Montague St | 40.69512845 | -73.99595065 | 01.06.2013 | 00:01:47 | 01.06.2013 | 00:35:35 |
| 2057 | 285 | Broadway & E 14 St | 40.73454567 | -73.99074142 | 532 | S 5 Pl & S 5 St | 40.710451 | -73.960876 | 01.06.2013 | 00:02:33 | 01.06.2013 | 00:36:50 |
| 369 | 509 | 9 Ave & W 22 St | 40.7454973 | -74.00197139 | 521 | 8 Ave & W 31 St N | 40.75096735 | -73.99444208 | 01.06.2013 | 00:03:29 | 01.06.2013 | 00:09:38 |
| 1829 | 265 | Stanton St & Chrystie St | 40.72229346 | -73.99147535 | 436 | Hancock St & Bedford Ave | 40.68216564 | -73.95399026 | 01.06.2013 | 00:03:47 | 01.06.2013 | 00:34:16 |
| 829 | 404 | 9 Ave & W 14 St | 40.7405826 | -74.00550867 | 303 | Mercer St & Spring St | 40.72362738 | -73.99949601 | 01.06.2013 | 00:04:22 | 01.06.2013 | 00:18:11 |
| 1316 | 423 | W 54 St & 9 Ave | 40.76584941 | -73.98690506 | 314 | Cadman Plaza West & Montague St | 40.69383 | -73.990539 | 01.06.2013 | 00:04:28 | 01.06.2013 | 00:26:24 |
| 1456 | 502 | Henry St & Grand St | 40.714215 | -73.981346 | 532 | S 5 Pl & S 5 St | 40.710451 | -73.960876 | 01.06.2013 | 00:04:41 | 01.06.2013 | 00:28:57 |
| 386 | 241 | DeKalb Ave & S Portland Ave | 40.68981035 | -73.97493121 | 365 | Fulton St & Grand Ave | 40.68223166 | -73.9614583 | 01.06.2013 | 00:05:13 | 01.06.2013 | 00:11:39 |
| 924 | 486 | Broadway & W 29 St | 40.7462009 | -73.98855723 | 521 | 8 Ave & W 31 St N | 40.75096735 | -73.99444208 | 01.06.2013 | 00:05:21 | 01.06.2013 | 00:20:45 |
| 1233 | 527 | E 33 St & 2 Ave | 40.744023 | -73.976056 | 296 | Division St & Bowery | 40.71413089 | -73.9970468 | 01.06.2013 | 00:06:44 | 01.06.2013 | 00:27:17 |
| 512 | 309 | Murray St & West St | 40.7149787 | -74.013012 | 300 | Shevchenko Pl & E 7 St | 40.728145 | -73.990214 | 01.06.2013 | 00:03:36 | 01.06.2013 | 00:12:08 |
| 505 | 309 | Murray St & West St | 40.7149787 | -74.013012 | 347 | Greenwich St & W Houston St | 40.728846 | -74.008591 | 01.06.2013 | 00:03:45 | 01.06.2013 | 00:12:10 |
| 833 | 503 | E 20 St & Park Ave | 40.73827428 | -73.98751968 | 503 | E 20 St & Park Ave | 40.73827428 | -73.98751968 | 01.06.2013 | 00:07:29 | 01.06.2013 | 00:21:22 |
| 1818 | 257 | Lispenard St & Broadway | 40.71939226 | -74.00247214 | 500 | Broadway & W 51 St | 40.76228826 | -73.98336183 | 01.06.2013 | 00:08:10 | 01.06.2013 | 00:38:28 |
| 682 | 486 | Broadway & W 29 St | 40.7462009 | -73.98855723 | 521 | 8 Ave & W 31 St N | 40.75096735 | -73.99444208 | 01.06.2013 | 00:08:53 | 01.06.2013 | 00:20:15 |
| 899 | 494 | W 26 St & 8 Ave | 40.74734825 | -73.99723551 | 494 | W 26 St & 8 Ave | 40.74734825 | -73.99723551 | 01.06.2013 | 00:09:25 | 01.06.2013 | 00:24:24 |
| 626 | 380 | W 4 St & 7 Ave S | 40.73401143 | -74.00293877 | 382 | University Pl & E 14 St | 40.73492695 | -73.99200509 | 01.06.2013 | 00:09:52 | 01.06.2013 | 00:20:18 |
| 219 | 305 | E 58 St & 3 Ave | 40.76095756 | -73.96724467 | 358 | Christopher St & Greenwich St | 40.73291553 | -74.00711384 | 01.06.2013 | 00:10:28 | 01.06.2013 | 00:14:07 |

Figure 25: Missing values 3

The figure shows:

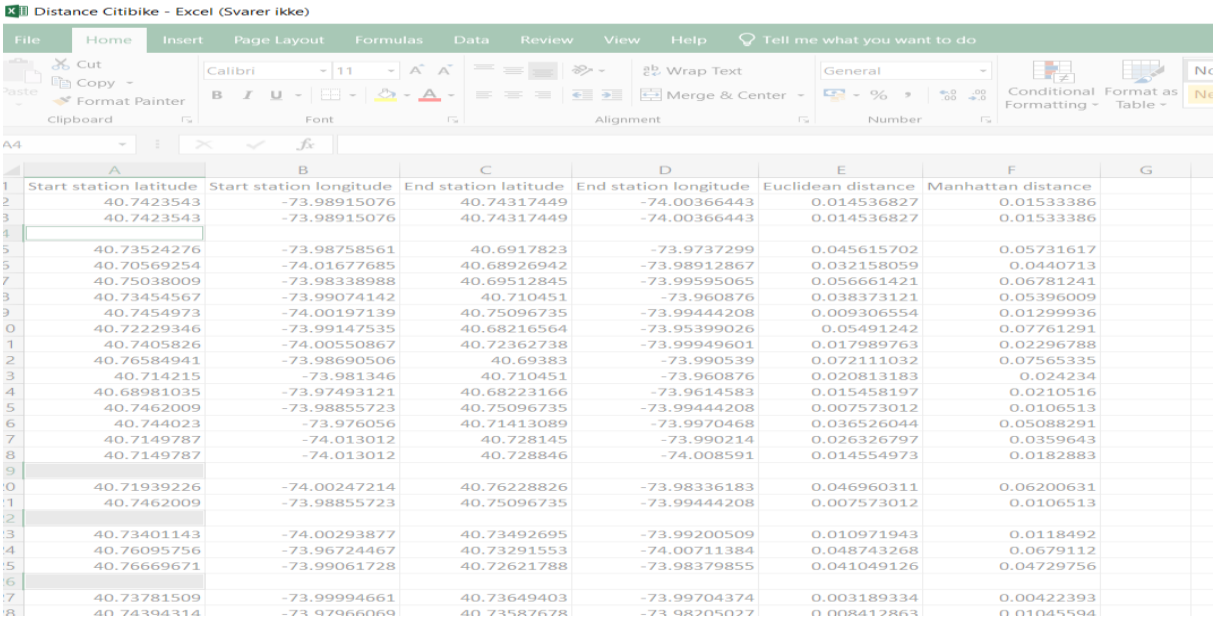
NOTE: On figure 24 the yellow lines are now replaced with red lines. I had some issues loading excel again and had to repeat the analysis.

- The two red lines shows rented bikes with exactly the same start latitude/longitude as in end latitude/longitude.
- The upper red line in figure 25: Even though rented time lasted from 00:00:44 till 00:35:03 nighttime with a trip duration of 2059 (34 minutes), the bike never leaved the start station. The figure shows start station and end station are the same.
- I counted 23625 after getting value of 0 in calculating both Euclidean and Manhattan distance.

This mean that those data can't be in the calculation of distance since I believe the trip never left the start station. Trip duration only show the amount of time spent renting the bike.

This leaves me to not include 38443 of 577703 data observations in the citibike-tripdata in finding the different values in distance.

NOTE: I had big issues in excel when filling the blank rows. I used a lot of time finding ways to make it work. Like in trip durations, I copied and pasted the distance in a new workbook, but it kept crashing and loading to infinity…




| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Start station latitude | Start station longitude | End station latitude | End station longitude | Euclidean distance | Manhattan distance | |
| 2 | 40.7423543 | -73.98915076 | 40.74317449 | -74.00366443 | 0.014536827 | 0.01533386 | |
| 3 | 40.7423543 | -73.98915076 | 40.74317449 | -74.00366443 | 0.014536827 | 0.01533386 | |
| 4 | | | | | | | |
| 5 | 40.73524276 | -73.98758561 | 40.6917823 | -73.9737299 | 0.045615702 | 0.05731617 | |
| 6 | 40.70569254 | -74.01677685 | 40.68926942 | -73.98912867 | 0.032158059 | 0.0440713 | |
| 7 | 40.75038009 | -73.98338988 | 40.69512845 | -73.99595065 | 0.056661421 | 0.06781241 | |
| 8 | 40.73454567 | -73.99074142 | 40.710451 | -73.960876 | 0.038373121 | 0.05396009 | |
| 9 | 40.7454973 | -74.00197139 | 40.75096735 | -73.99444208 | 0.009306554 | 0.01299936 | |
| 10 | 40.72229346 | -73.99147535 | 40.68216564 | -73.95399026 | 0.05491242 | 0.07761291 | |
| 11 | 40.7405826 | -74.00550867 | 40.72362738 | -73.99949601 | 0.017989763 | 0.02296788 | |
| 12 | 40.76584941 | -73.98690506 | 40.69383 | -73.990539 | 0.072111032 | 0.07565335 | |
| 13 | 40.714215 | -73.981346 | 40.710451 | -73.960876 | 0.020813183 | 0.024234 | |
| 14 | 40.68981035 | -73.97493121 | 40.68223166 | -73.9614583 | 0.015458197 | 0.0210516 | |
| 15 | 40.7462009 | -73.98855723 | 40.75096735 | -73.99444208 | 0.007573012 | 0.0106513 | |
| 16 | 40.744023 | -73.976056 | 40.71413089 | -73.9970468 | 0.036526044 | 0.05088291 | |
| 17 | 40.7149787 | -74.013012 | 40.728145 | -73.990214 | 0.026326797 | 0.0359643 | |
| 18 | 40.7149787 | -74.013012 | 40.728846 | -74.008591 | 0.014554973 | 0.0182883 | |
| 19 | | | | | | | |
| 20 | 40.71939226 | -74.00247214 | 40.76228826 | -73.98336183 | 0.046960311 | 0.06200631 | |
| 21 | 40.7462009 | -73.98855723 | 40.75096735 | -73.99444208 | 0.007573012 | 0.0106513 | |
| 22 | | | | | | | |
| 23 | 40.73401143 | -74.00293877 | 40.73492695 | -73.99200509 | 0.010971943 | 0.0118492 | |
| 24 | 40.76095756 | -73.96724467 | 40.73291553 | -74.00711384 | 0.048743268 | 0.0679112 | |
| 25 | 40.76669671 | -73.99061728 | 40.72621788 | -73.98379855 | 0.041049126 | 0.04729756 | |
| 26 | | | | | | | |
| 27 | 40.73781509 | -73.99994661 | 40.73649403 | -73.99704374 | 0.003189334 | 0.00422393 | |
| 28 | 40.74394314 | -73.97966069 | 40.73587678 | -73.98205027 | 0.008412863 | 0.01045594 | |

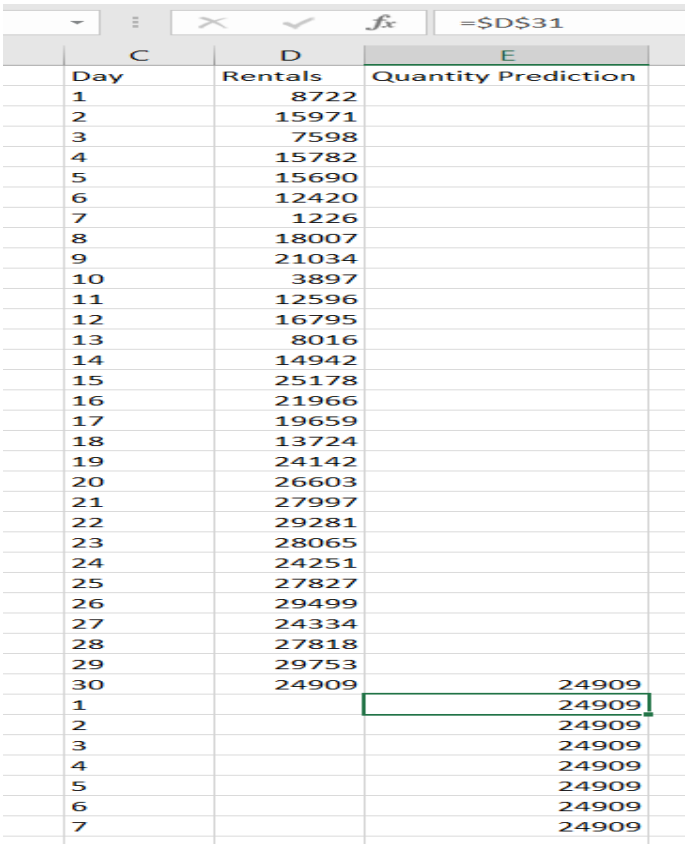
***Note:***

*This was a big issue. I spent hours in figuring how to get excel to work with me. It just kept crashing and loading. It stopped completely when filling the empty rows (figure above) after deleting the values/missing values that I believed didn't fit in finding the distance. I used all the tips in Discord, changed to new workbook, copy/paste single columns for analysis, etc. etc., but it kept crashing. I even changed computers. I used a lot of time on this. Did all the work finding the right way to analyze the distance and couldn't continue my work. I'm so sorry I didn't finish assignment 5 because of this issue.*

**Assignment 6:**

**A naïve prediction:**

In this assignment I made a new sheet called predictions. First prediction was to make a naïve prediction. I made three columns, Day and Rentals (as shown in the assignment pdf), then made a column called Quantity Prediction.



=$D$31

| | C | D | E |
|---|---|---|---|
| | Day | Rentals | Quantity Prediction |
| | 1 | 8722 | |
| | 2 | 15971 | |
| | 3 | 7598 | |
| | 4 | 15782 | |
| | 5 | 15690 | |
| | 6 | 12420 | |
| | 7 | 1226 | |
| | 8 | 18007 | |
| | 9 | 21034 | |
| | 10 | 3897 | |
| | 11 | 12596 | |
| | 12 | 16795 | |
| | 13 | 8016 | |
| | 14 | 14942 | |
| | 15 | 25178 | |
| | 16 | 21966 | |
| | 17 | 19659 | |
| | 18 | 13724 | |
| | 19 | 24142 | |
| | 20 | 26603 | |
| | 21 | 27997 | |
| | 22 | 29281 | |
| | 23 | 28065 | |
| | 24 | 24251 | |
| | 25 | 27827 | |
| | 26 | 29499 | |
| | 27 | 24334 | |
| | 28 | 27818 | |
| | 29 | 29753 | |
| | 30 | 24909 | 24909 |
| | 1 | | 24909 |
| | 2 | | 24909 |
| | 3 | | 24909 |
| | 4 | | 24909 |
| | 5 | | 24909 |
| | 6 | | 24909 |
| | 7 | | 24909 |

Figure 26: Making a naïve prediction of rental bikes.

- To find values inside coulum rentals, I used =countif
- Finding prediction on 01.07.13 to 07.07.13 I used =$D$31.
- Since the number of rental bikes in 30.06.13 was 24909, the naïve prediction will predict that the same amount of bikes will be rented from 01.07.13 to 07.07.13.
- The chart is shown below.

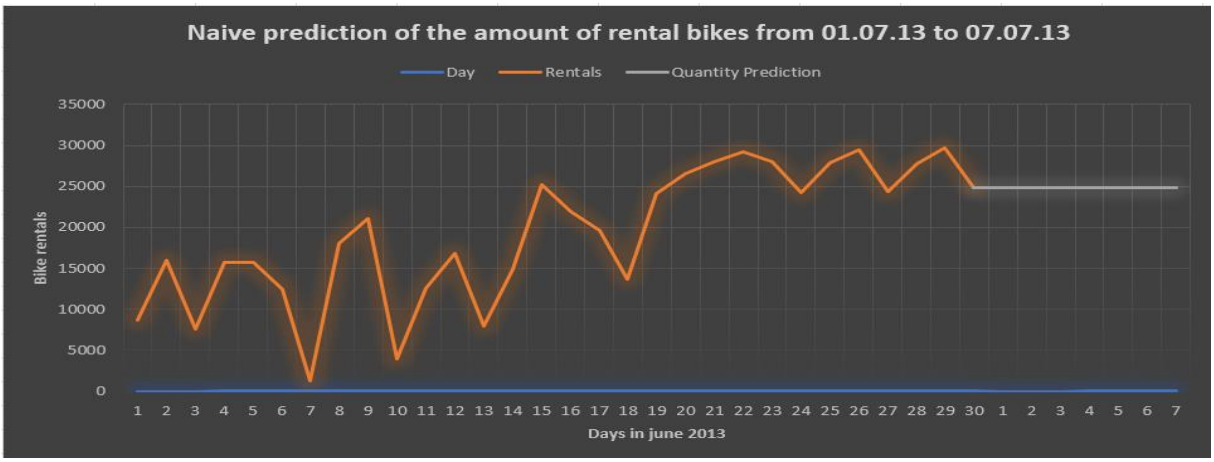


Figure 27: Naïve prediction.

The chart shows:

- That the prediction value from 01.07. to 07.07. as the same value and therefore a straight line.
- This shows a prediction but maybe not quite realistic.

**The average forecast:**

This prediction focuses on the average or the mean value. Finding the mean value using =average on column D “Rentals”, gives me a value of 19257. This will be the value of prediction the first 7 days in juli.

| | | | |
|---|---|---|---|
| 5.2013 | 26 | 29499 | |
| 5.2013 | 27 | 24334 | |
| 5.2013 | 28 | 27818 | |
| 5.2013 | 29 | 29753 | |
| 5.2013 | 30 | 24909 | 24909 |
| 5.2013 | 1 | | 19256.73 |
| 5.2013 | 2 | | 19256.73 |
| 5.2013 | 3 | | 19256.73 |
| 5.2013 | 4 | | 19256.73 |
| 5.2013 | 5 | | 19256.73 |
| 5.2013 | 6 | | 19256.73 |
| 5.2013 | 7 | | 19256.73 |
| 5.2013 | | | |
| 5.2013 | | | |
| 5.2013 | | | |

Figure 28: Making an average forecast of rented bikes.

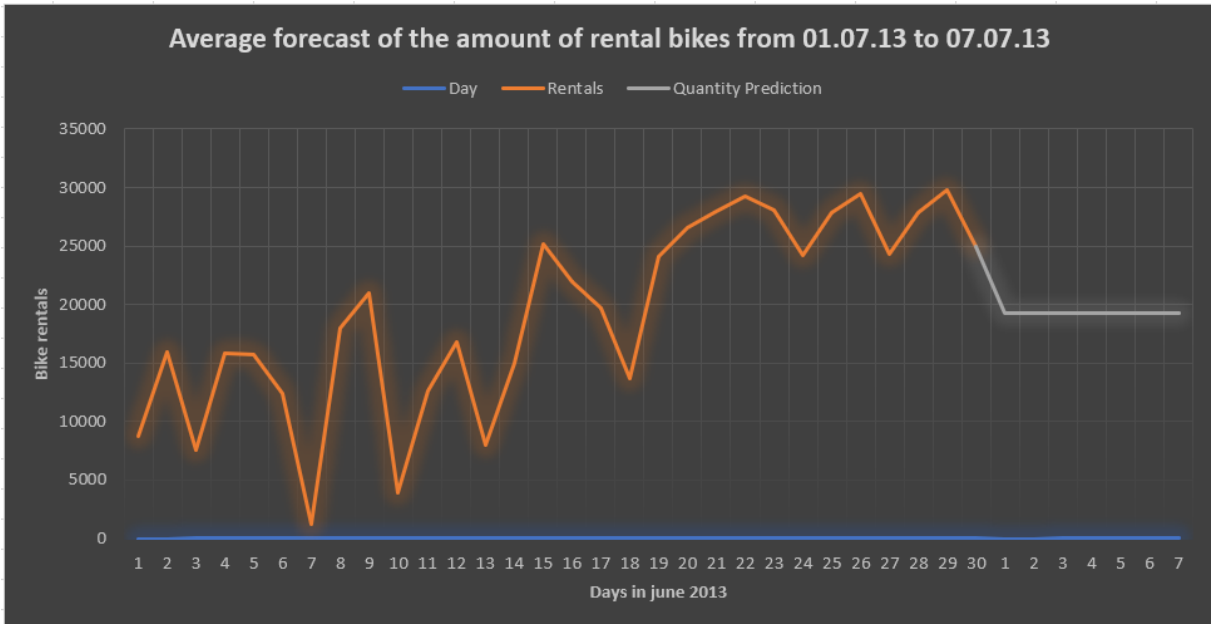



Figure 29: Average forecast.

The chart shows:

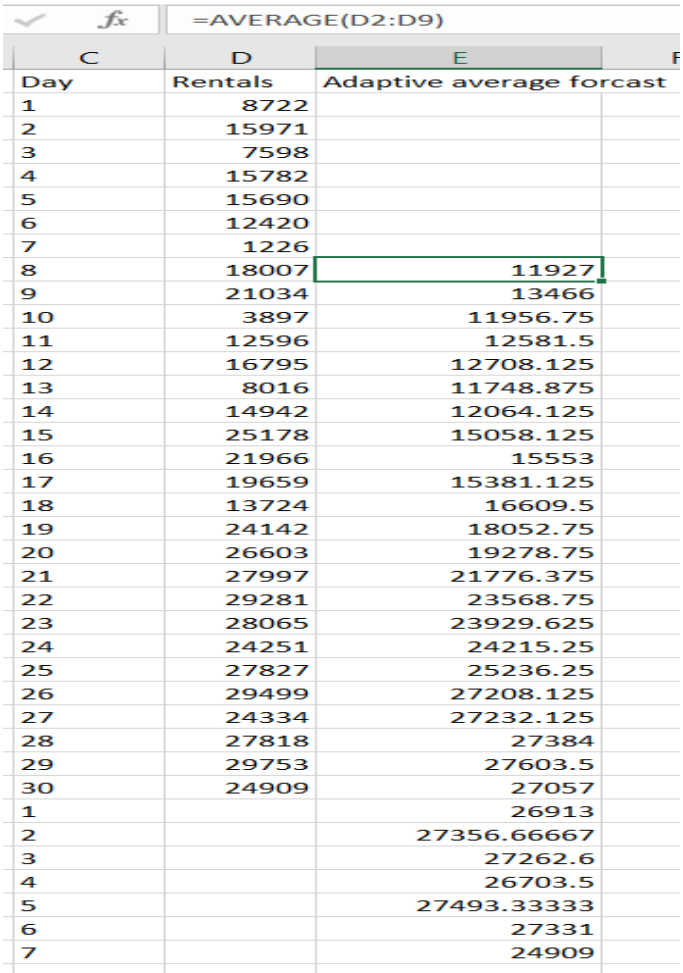
- A straight line like the naïve prediction, but with a different value.
- The same as chart figure 27, this maybe not quite realistic.

Conclusion of the two charts:

Both the naïve prediction and the average forecast on predicting forward values represents little qualitative insights. Both predictions shows us just a linear line of value even though the lines are of some difference.

**Adaptive average forecast:**

In this forecast I started to find the average of the first 7 values (7 days) in column D. Then I made a new column called “adaptive average forecast using =average starting from day 8 all the way to the date 07.07.




| C | D | E |
|---|---|---|
| Day | Rentals | Adaptive average forcast |
| 1 | 8722 | |
| 2 | 15971 | |
| 3 | 7598 | |
| 4 | 15782 | |
| 5 | 15690 | |
| 6 | 12420 | |
| 7 | 1226 | |
| 8 | 18007 | 11927 |
| 9 | 21034 | 13466 |
| 10 | 3897 | 11956.75 |
| 11 | 12596 | 12581.5 |
| 12 | 16795 | 12708.125 |
| 13 | 8016 | 11748.875 |
| 14 | 14942 | 12064.125 |
| 15 | 25178 | 15058.125 |
| 16 | 21966 | 15553 |
| 17 | 19659 | 15381.125 |
| 18 | 13724 | 16609.5 |
| 19 | 24142 | 18052.75 |
| 20 | 26603 | 19278.75 |
| 21 | 27997 | 21776.375 |
| 22 | 29281 | 23568.75 |
| 23 | 28065 | 23929.625 |
| 24 | 24251 | 24215.25 |
| 25 | 27827 | 25236.25 |
| 26 | 29499 | 27208.125 |
| 27 | 24334 | 27232.125 |
| 28 | 27818 | 27384 |
| 29 | 29753 | 27603.5 |
| 30 | 24909 | 27057 |
| 1 | | 26913 |
| 2 | | 27356.66667 |
| 3 | | 27262.6 |
| 4 | | 26703.5 |
| 5 | | 27493.33333 |
| 6 | | 27331 |
| 7 | | 24909 |

Figure 30: Making adaptive average forecast

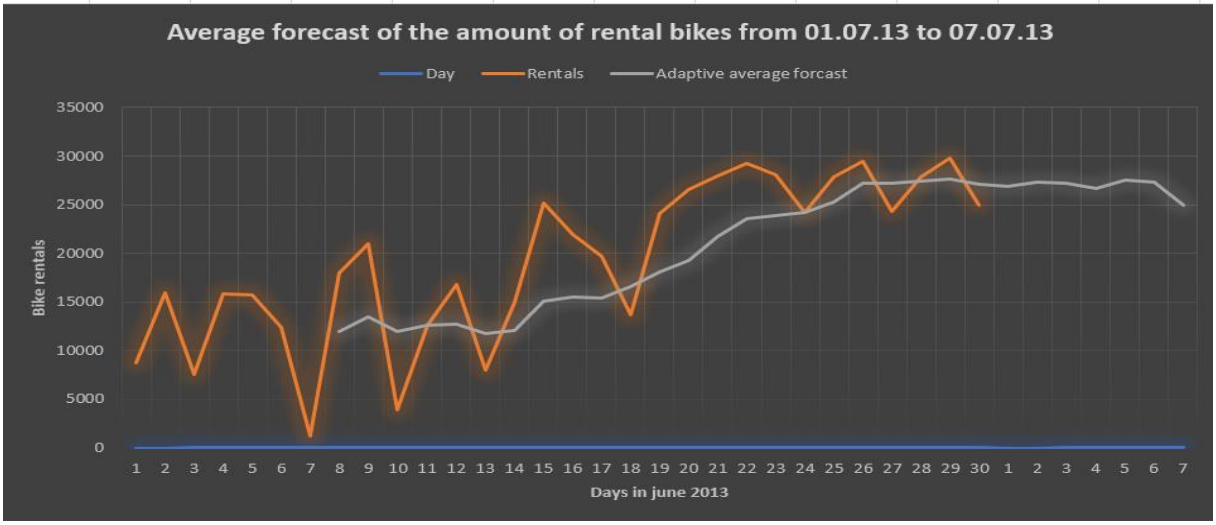


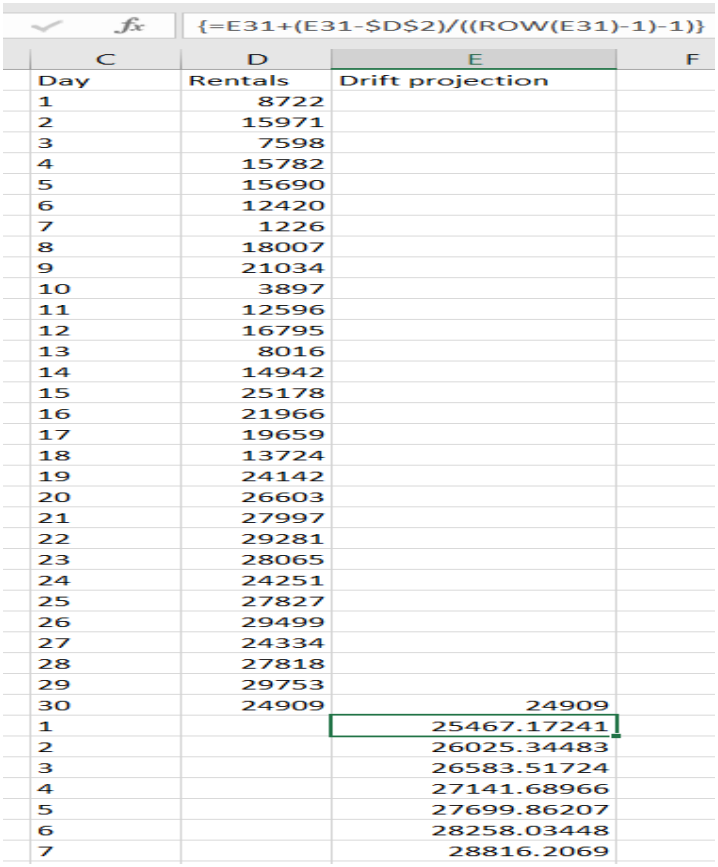

Figure 31: Adaptive mean forecast

The chart shows:

- By finding the first 7 days average value, the chart shows us two lines, where the adaptive average forecast (gray line) predicted the past, while the orange shows us what actually happened. The gay line also predicts the first 7 days in July.

**Drift projection:**

This prediction is to show the trending values of past data.




| | C | D | E | F |
|---|---|---|---|---|
| | Day | Rentals | Drift projection | |
| | 1 | 8722 | | |
| | 2 | 15971 | | |
| | 3 | 7598 | | |
| | 4 | 15782 | | |
| | 5 | 15690 | | |
| | 6 | 12420 | | |
| | 7 | 1226 | | |
| | 8 | 18007 | | |
| | 9 | 21034 | | |
| | 10 | 3897 | | |
| | 11 | 12596 | | |
| | 12 | 16795 | | |
| | 13 | 8016 | | |
| | 14 | 14942 | | |
| | 15 | 25178 | | |
| | 16 | 21966 | | |
| | 17 | 19659 | | |
| | 18 | 13724 | | |
| | 19 | 24142 | | |
| | 20 | 26603 | | |
| | 21 | 27997 | | |
| | 22 | 29281 | | |
| | 23 | 28065 | | |
| | 24 | 24251 | | |
| | 25 | 27827 | | |
| | 26 | 29499 | | |
| | 27 | 24334 | | |
| | 28 | 27818 | | |
| | 29 | 29753 | | |
| | 30 | 24909 | 24909 | |
| | 1 | | 25467.17241 | |
| | 2 | | 26025.34483 | |
| | 3 | | 26583.51724 | |
| | 4 | | 27141.68966 | |
| | 5 | | 27699.86207 | |
| | 6 | | 28258.03448 | |
| | 7 | | 28816.2069 | |

Figure 32: Making a drift projection chart.

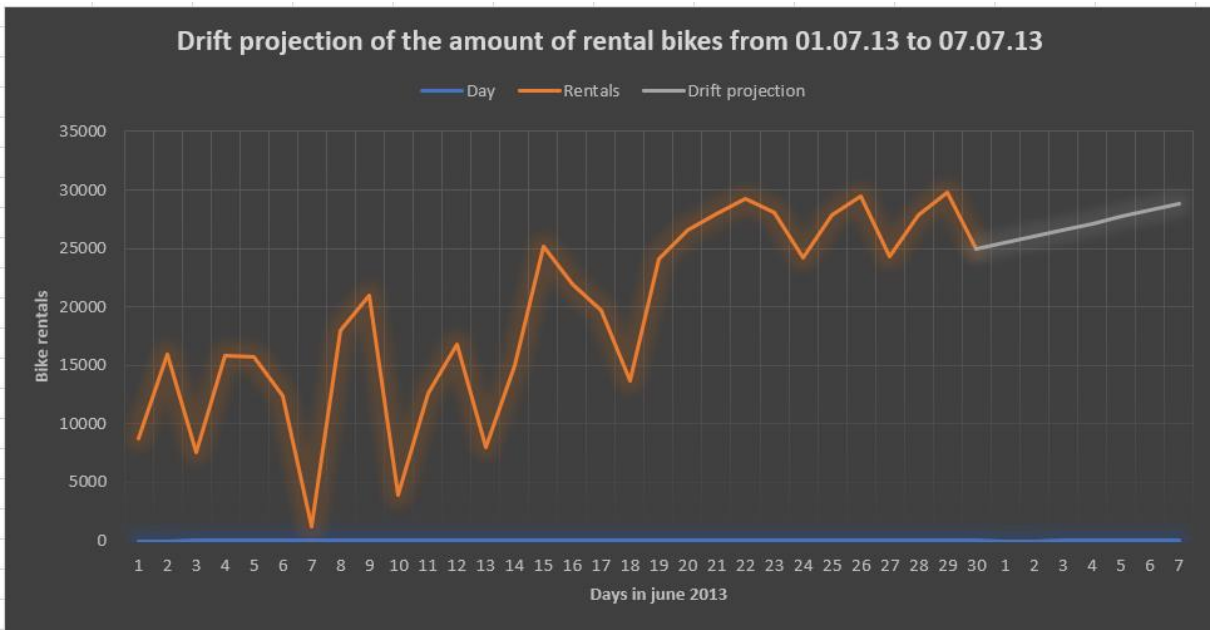



Figure 33: Drift projection chart.

The chart shows:

- An upwards line trending value.
- The trend of increased rented bikes in the past will make the drift projection go up. This is because the prediction follows the trend.

**Naïve seasonal method:**

To make the naïve seasonal method I copied the values from date 01.06 to 07.06. and pasted it from 01.07. to 07.07.

=D2

| C | D | E |
|---|---|---|
| Day | Rentals | Naive sesonal predic |
| 1 | 8722 | |
| 2 | 15971 | |
| 3 | 7598 | |
| 4 | 15782 | |
| 5 | 15690 | |
| 6 | 12420 | |
| 7 | 1226 | |
| 8 | 18007 | |
| 9 | 21034 | |
| 10 | 3897 | |
| 11 | 12596 | |
| 12 | 16795 | |
| 13 | 8016 | |
| 14 | 14942 | |
| 15 | 25178 | |
| 16 | 21966 | |
| 17 | 19659 | |
| 18 | 13724 | |
| 19 | 24142 | |
| 20 | 26603 | |
| 21 | 27997 | |
| 22 | 29281 | |
| 23 | 28065 | |
| 24 | 24251 | |
| 25 | 27827 | |
| 26 | 29499 | |
| 27 | 24334 | |
| 28 | 27818 | |
| 29 | 29753 | |
| 30 | 24909 | 24909 |
| 1 | | 8722 |
| 2 | | 15971 |
| 3 | | 7598 |
| 4 | | 15782 |
| 5 | | 15690 |
| 6 | | 12420 |
| 7 | | 1226 |

Figure 34: Making a naïve seasonal chart.

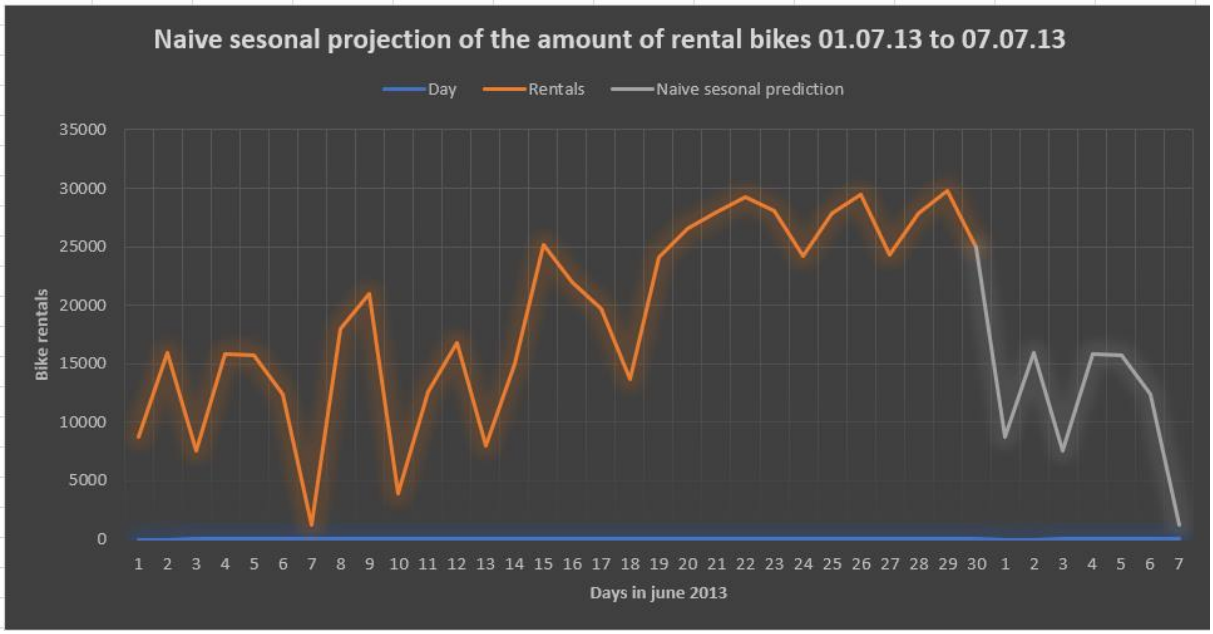



Figure 35: Naïve seasonal prediction.

The chart shows:

- The prediction shows us the same value as in the values from 01.06. to 07.06.
- The prediction shows us a falling rate of rented bikes. But if it follows the same pattern as in June, the line will go up the following day.

**Assignment 7:**

Using a linear trend and exponential trend inside the charts.

**Naïve seasonal projection:**

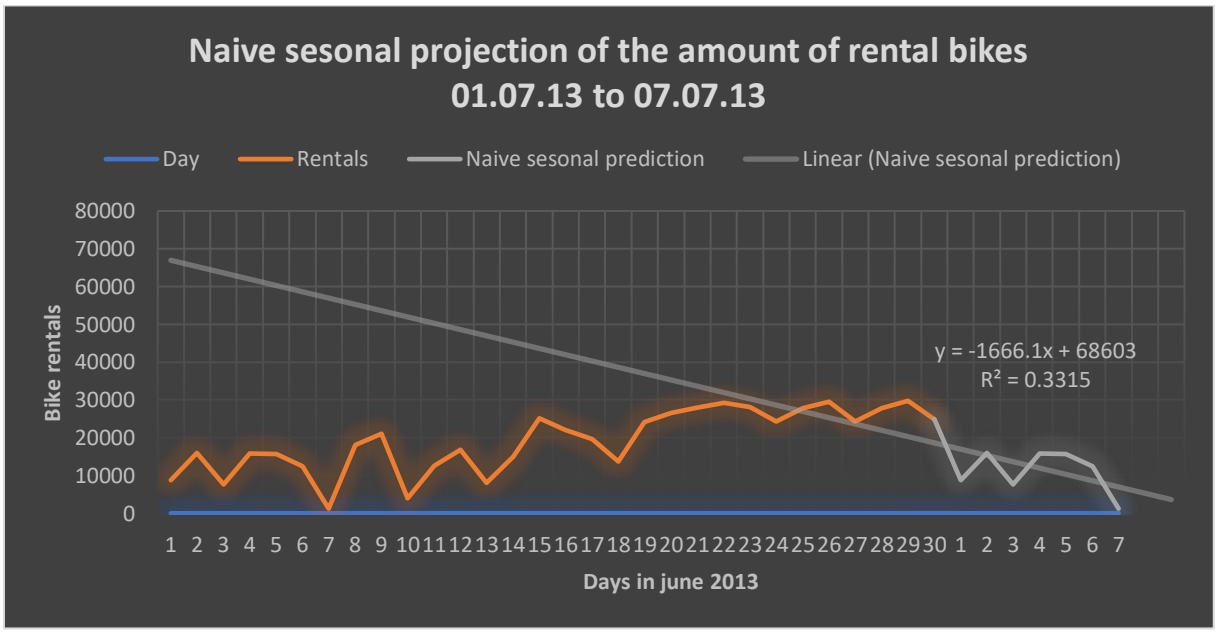


Figure 36: Linear trend telling us that the value is decreasing.

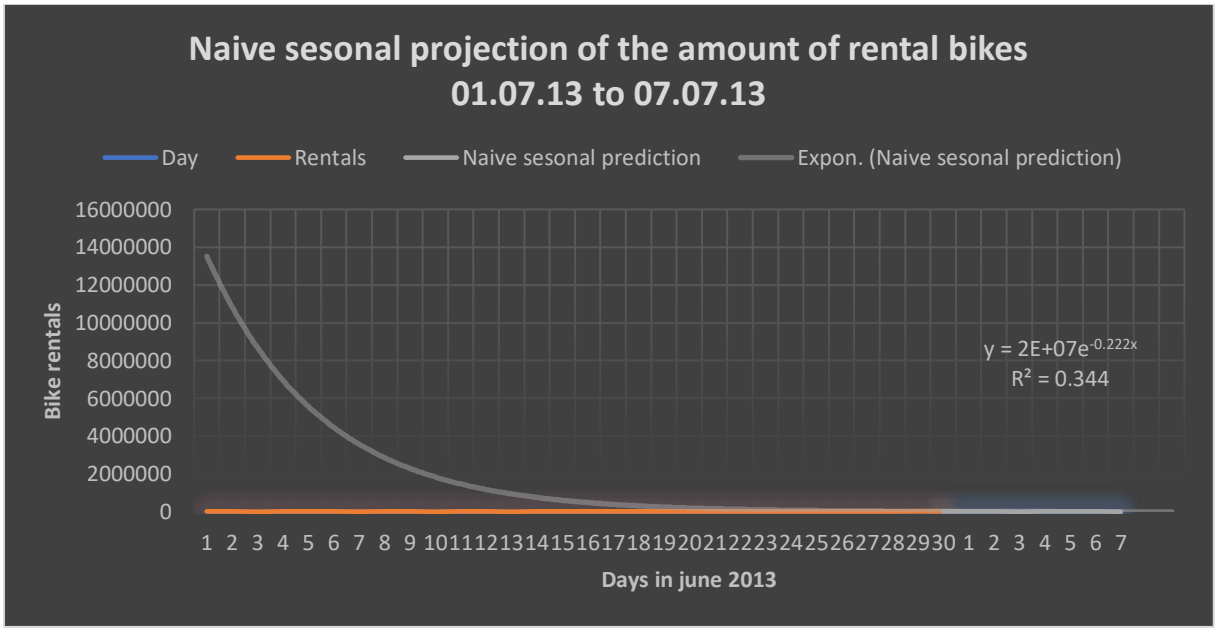



Figure 37: Exponential trend showing an increasing high rate of downfall in values. Not quite sure if I did this right….

**Drift projection:**

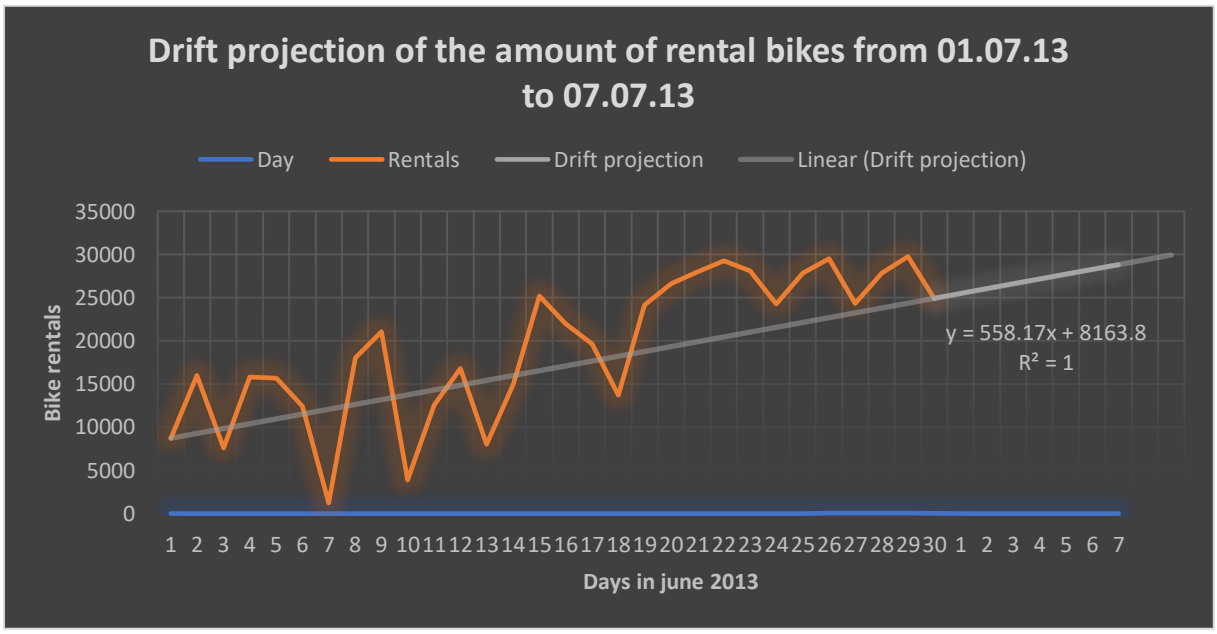



Figure 38: Linear line shows that the values are increasing with a steady rate.

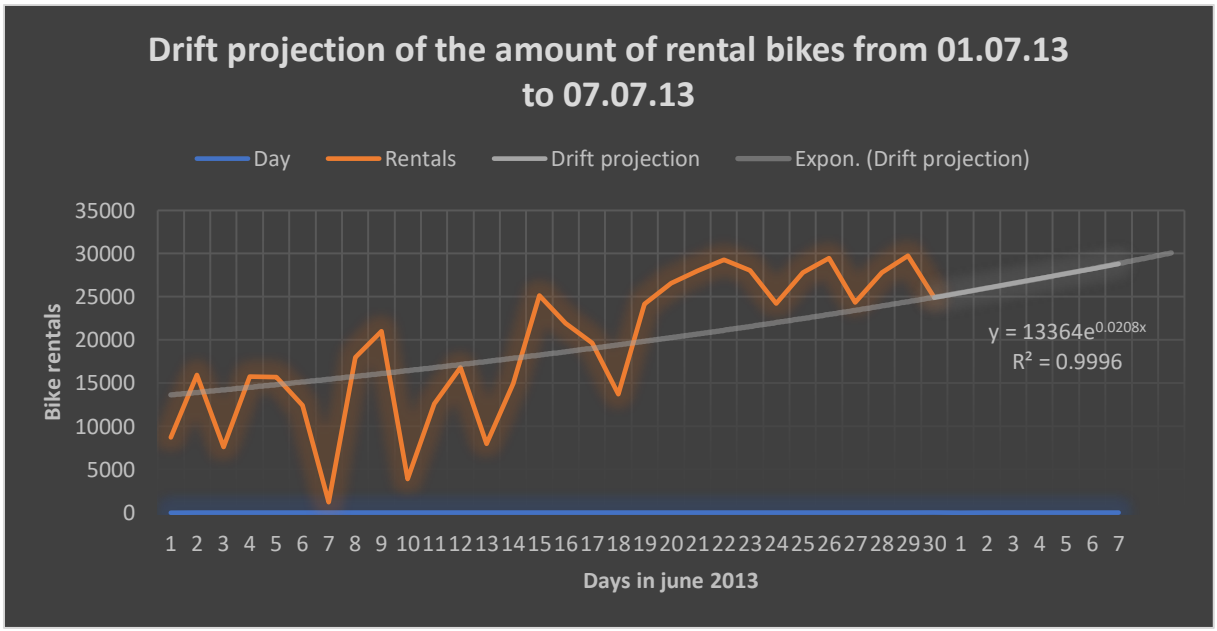



Figure 39: Exponential trend showing a rise in rates.

**Adaptive forecast:**

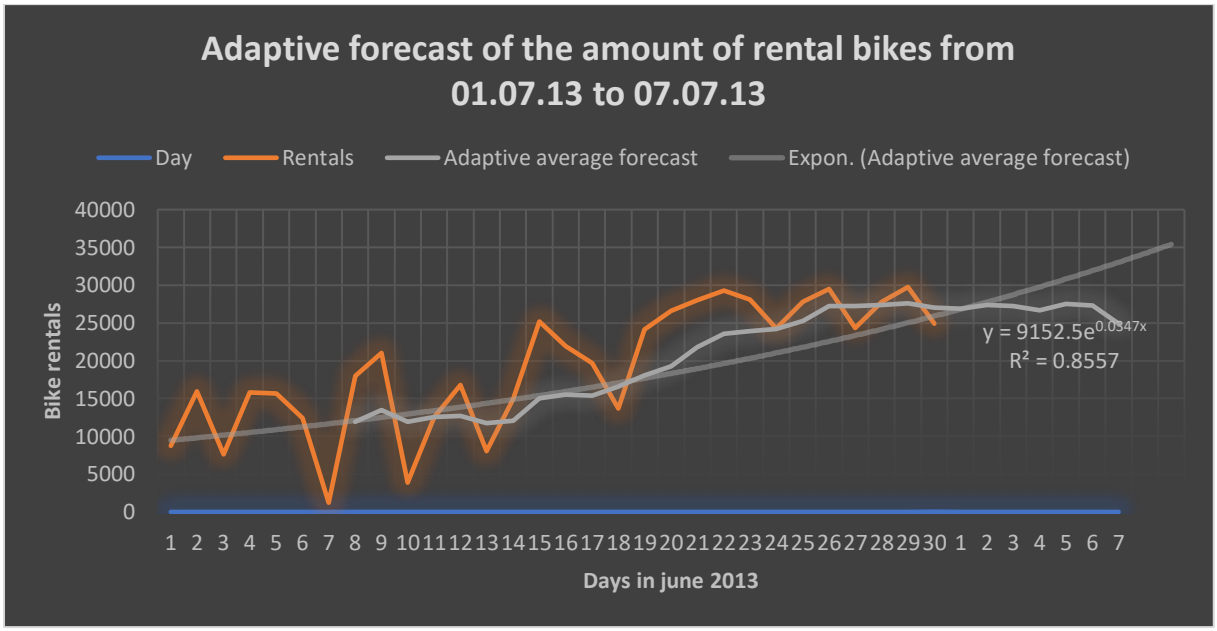



Figure 40: Exponential trend showing an increase in values.

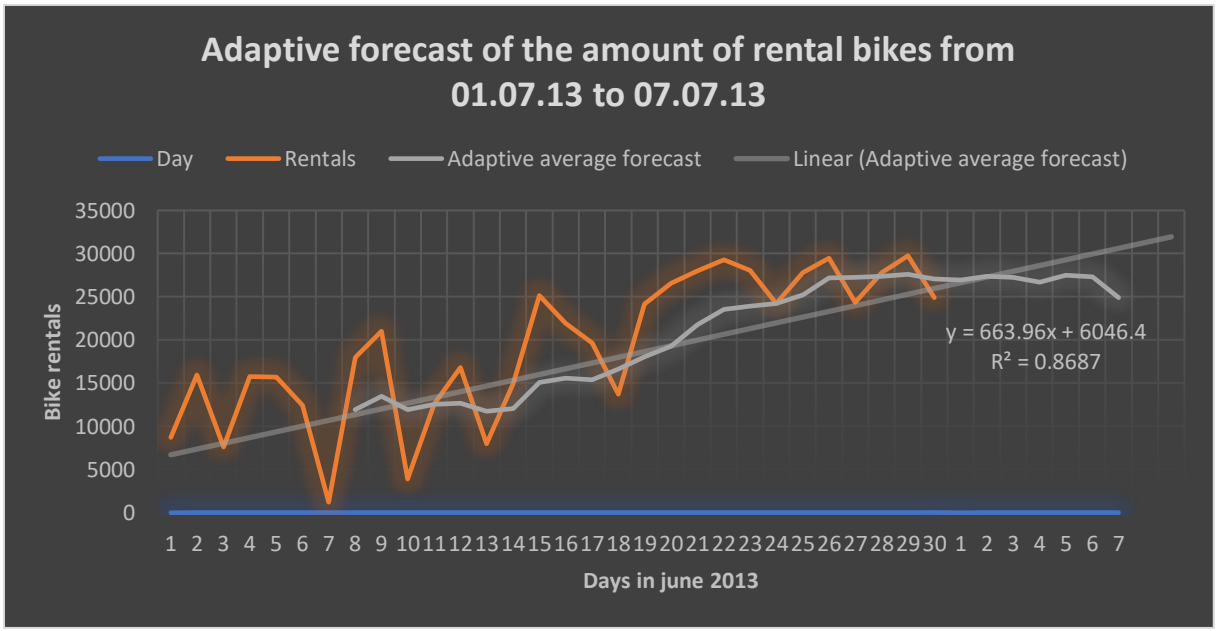



Figure 41: Like in exponential trend, the linear trend also shows a steady increase in values.

**Average forecast:**

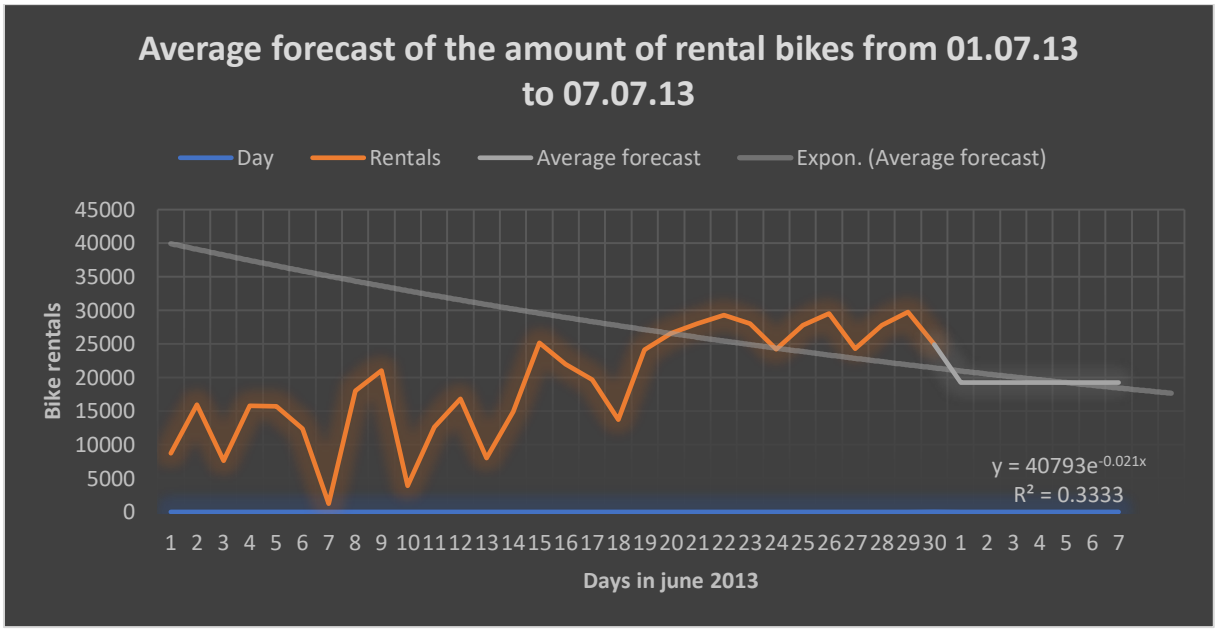



Figure 42: Exponential trend that shows a decrease in values.

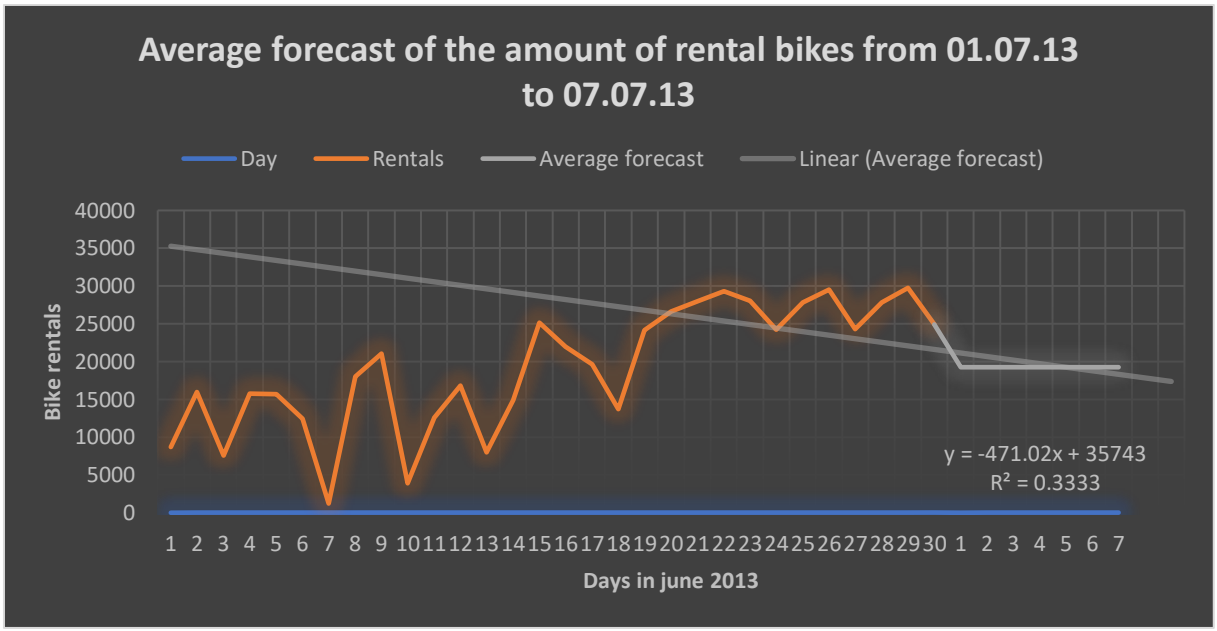



Figure 43: Like in figure 42, the linear trend shows a steady decreasing rate.

**Naïve prediction:**

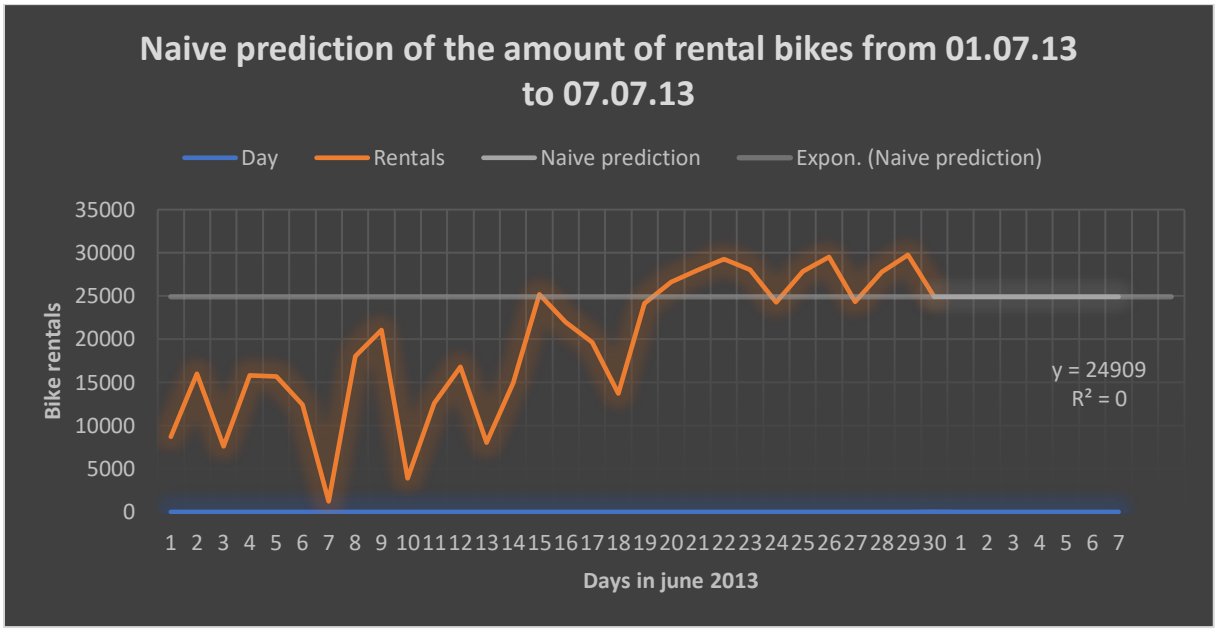



Figure 44: Exponential trend is not suitable in this chart because the naïve prediction forces the trendline to be at a zero value.

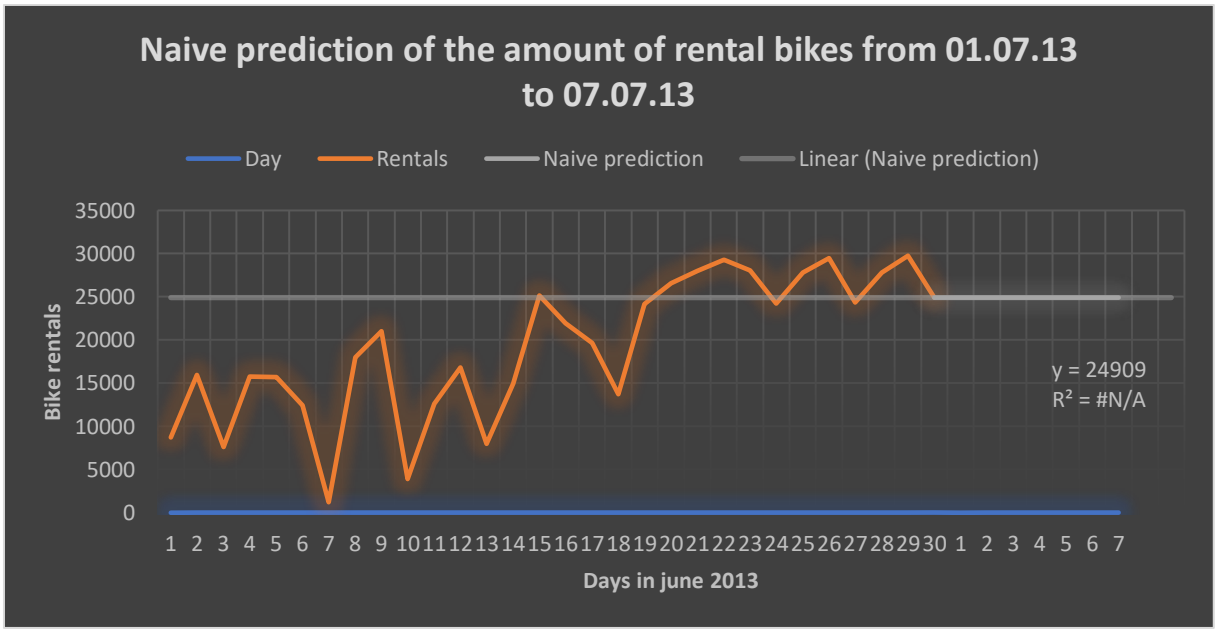



Figure 45: The tend line is constant and does not show us a decrease or increase because of the prediction.