Acquiring the Chicago taxi data set was straightforward – I was able to download it directly from the City of Chicago’s website. Since it was a large download (40GB), I added wget functionality to Git Bash and downloaded via that instead of attempting to download it from the browser.

Since such a large dataset would not be able to fit into memory, I decided to set up a PostgreSQL database on my computer to house the data. The inspiration from this was taken from a previous analysis already done on the data by Todd Schneider (<http://toddwschneider.com/posts/chicago-taxi-data/>). The raw data was first loaded into a temporary PostgreSQL database. Unique Taxi IDs were standardized for readability. Then the raw data was typecasted into appropriate data types into a working table.

Since I decided to work solely with 2016 data due to the size of the dataset, I then created another table to house only 2016 taxi trips to speed up queries. Additional to that, since I am focusing mainly on the taxi ID, trip start time, and trip fare, I created another table which housed only those variables. I did this because from my research, I found that SQL queries will load the whole row into memory – even if you are only using certain columns. I did this to cut down on my disk I/O to speed up queries.

There are a number of missing or odd values in the dataset due to bad/incomplete data collection –

* Duration of trip and length of trip are sometimes 0
* Geographical location is missing
* Missing fare amounts
* Extremely large fare amounts (in the thousands of dollars)

Since I am focusing on the taxi ID, trip start time, and trip fare, most of these issues did not affect me.

For the missing and extremely large fare amounts, I took these actions:

* Excluded trips with missing fare amounts
* Excluded trips with a fare amount greater than $200
  + I choose $200 as the exclusion point because from a quick analysis of the mean and standard deviation, I found that around $200 is a little over five standard deviations from the mean. Since I am concerned with typical taxi trips, and since the data collection of the extremely large trips are in question, I feel comfortable with proceeding with this cutoff limit.

These two cleaning steps reduced the number of trips in 2016 from 19,878,210 to 19,873,687 – a reduction of 4,523 trips – which is 0.023% of total trips counted for in 2016.