The Chicago parking data had it a number of missing sections as would any public data set. To make the data relevant to the project at hand early on we needed to make sure that there was an actionable time series. We created a time series comprising of a day and hour columns that allowed us to use the data accordingly. Besides creating the time series we also had to use some other libraries to produce the needed info.

First impressions showed that much care was taken in the submission of the data. The initial data set had tickets from the year of 2007 to the spring of 2018. Out of 18 million rows, 90% of the columns were not missing any info. Most of the missing values were in columns that I later removed or they were not relevant for the task at hand. Some of the removed columns included the license plate number and the vehicle make.

One import column that needed to be fixed was the zip code column. There were 1-1.5 million incorrect zip codes initially. Upon early inspection, we found that the zip codes were wrong or alphanumeric based. We fixed the bad zip codes by running a geolocator program to output the proper zip code for the location of the ticket. The correct zip code was not found for every line in the data frame. The years and days that we did not use I did not look for any data on.

The other challenge that came in the data wrangling was producing the latitude and longitude of the address where each ticket was issued. It was nothing that 25 lines of python couldn’t fix there was quite a learning curve that went with learning about dressing the problem. The data cleaned up and worked quite well. I am very happy with the outcome that we produced.