In the process of gathering and cleaning data, my task is to obtain the number of automobiles in different counties from the official website of DMV. The original document is a PDF version of data source, and I converted this data source to txt format via Adobe acrobat Xi Pro. After that, I read my data into a data frame in Python. Since the data source shows a full list of numbers of automobile in California, my challenge is to retrieve some certain target counties that are included in our analysis and switch the data into the format that could be executed for later calculation. In order to solve that, I inserted a block of “while” and “if” code where I could locate the comma and avoid it.

Chi:

During the process of gathering data part, I collect the income data. The income data comes from the website [www.city-data.com](http://www.city-data.com). Some problems made the data not easy to collect. First of all, there isn’t a complete data form, I need to search data city by city. Fortunately, most of these web page were set as the same form which let me easier to read and collect. I use bash and regular expression to read the HTML and clean the data, which lead a lot of task. The second problem is that I need to come up with a method to remove the comma in the number of income. After trying a lot of method, finally, I used python loops and conditions to complete the work and it works well.