For this mid term I used a credit card dataset that contains fraudulent and unfraudulent data. This assignment asked me to clean and structure this dataset. Unfortunately, this dataset is clean already, I should have known it’s the most popular one on Kaggle. However, in my code I did the steps I would have taken to clean this dataset to further analysis. Overall my biggest struggle was to take examples from the book and make them work in python 3.

Some of the challenged I came across was zip. In python 2 this works as a list but in python 3 it does not. I had to add a tail list on top of zip to denote a list. When using Dictreader\reader I had to make sure I used RT or R and a Unicode of utf8 to import the data. RB is binary and would not work with the dataset giving me errors. Printing needed to be wrapped around parentheses and some packages needed to be installed but eventually I learned more about how python 2 and 3 works.

With this dataset finding duplicates was difficult as the data consisted of timed transactions, so duplicates were not there. I did my best to try and find duplicates in data using the examples in the book that were not examples but uses of the example data. Overall, I think I would have had an easier time picking a dataset that was not already ready for analysis but raw data that had not been cleaned. I would have shown better results and I will be more selective with my dataset choice for the final project.