README file for project 3

1. Name: Ruochen Li, Zeqing Wang
   Uni: rl3292, zw2856

2. List of files: main.py, INTEGRATED-DATASET.csv, example_run.txt

3. Commands required to run the program:
   cd ..
   cd zeqing
   cd cs6111/proj3
   python3 main.py INTEGRATED-DATASET.csv 0.15 0.6

4. **(a)**
   We chose to use the "Parking Violations Issued - Fiscal Year 2023" dataset from the given NYC data website. Link: https://data.cityofnewyork.us/City-Government/Parking-Violations-Issued-Fiscal-Year-2023/pvqr-7yc4

   (b)
   After looking through the description and checking the downloaded file, we realized that the dataset is too big for our purpose of project, it contains more than 13 million rows, we figured that running our program on the full dataset will take an unreasonable amount of time. Therefore, we have written a helper program to randomly choose 16,000 entries from the original dataset, and then mapped them to our INTEGRATED-DATASET file. Although this would make our findings less representative for the data, after a few trials on different data size, we believe that the reduced dataset will still be enough to generate interesting and insightful association rules.

   (c)
   To begin with, the original dataset we found are really rich in terms of information and potential associations, it is a record on open parking and camera violations, and it has provided 43 attributes regarding the violation information, such as state, license type, violation type and etc. From there, we have figured that there has to be interesting associations between these attributes, for example, if a car has violated the parking rules in NY, then it might a specific type of violation such as parking on the street, since intuitively, NY area is known for the hardness to

find parking slot so people might choose to park on the street. While in other county, the violation type might be different. Furthermore, the [NY, no parking on street] association might be related to other attributes as well, such as the license type – as we know that smaller cars are probably going to park on the street more often than trucks in NY area given the traffic.
Moreover, we have also eliminated some columns from the dataset, as we figured that it will significantly reduce the time required to run our program, and reducing them will not influence the our association rules much. The columns that we have removed and reasonings are below:
Reference links:
https://www.nyc.gov/site/finance/vehicles/required-elements-in-a-ticket.page
Summons Number, Plate ID, Meter Number: These columns represent the IDs, which are basically unique given the data size, it won't help minging associations rules since it's almost impossible for them to be in the frequent itemsets.
Registration State: We are focusing on the violations, although this attribute might generate association, it is not likely to be associated with the violation that we are interested in.
Vehicle year: This the specific information regarding the vehicle, although the violation might be related to the vehicle itself, there are other columns (vehicle body type, vehicle make) that could capture the variance and form better association. Vehicle year seems to be a really trivial factor so we have dropped the column.
Issue Date: Although it seems related to the topic, our dataset contains violation information for 2022 and 2023, so unless we set the min sup to a very low value, the date is almost impossible to be in the frequent itemset.
House Number, Street Name, Intersecting Street, Street Code1, Street Code2, Street Code3: We have included other attributes (violation location, state, county), to represent the locations.
Vehicle Expiration Date: We can't think of any relations between expiration date and parking/camera violation.
Issuer Precinct, Issuer Code, Issuer Command, Issuer Squad: These columns are informations of the Issuer, which are not of our interests here.
Time First Observed, Date First Observed: These columns indicate when the officer first observed the vehicle, we have included the column "Violation time" to count for the time of the violation, therefore these columns can be dropped.
Law Section, Sub Division, Violation Legal Code, Violation Post Code: These columns correspond to which specific laws the violation is based on, we have included "Violation code" to address the type of violation so we dropped these columns.

Days Parking In Effect, From Hours In Effect, To Hours In Effect: These columns address the time limitations on the parking sign, and we believe they are not very meaningful since we are looking for associations for the violations.
Issuing Agency: Although it's related to the topic, but the value is the same for all the "market basket", so it won't help to keep this feature.
Feet From Curb: Although we initially thought this might be important to include, it actually refers to the distance a vehicle should be parked away from the curb by the local laws, therefore it is essentially covered by the location columns. So we decided to drop this one as well.
Modified column:
Violation time: The original dataset used the exact time to represent the violation time, however, we think that this representation would be really hard for us to generate the association rules, since it will be almost impossible to be in the frequent datasets. Therefore, we changed this column to represent a range of time: early morning(12am - 6am), morning(6am - 12pm), afternoon(12pm - 6pm), evening(6pm - 12am). We believe the modification will make

Overall, we believe that our dataset is compelling, and our modifications are reasonable for the purpose of association rule mining.

5. Our project mainly contains the following functions:
    - **main():** It takes the args from user prompt and do the output to txt tasks (Step 2, 5)
    - **read_data():** It reads the data from the input INTEGRATED-DATASET file. (Step 1)
    - **apriori():** Our apriori algorithm implementation, it contains three helper functions – find_frequent_itemsets() which finds the frequent itemset by using the min sup; join_candidate() which join the itemsets from itemsets generated from the previous function; prune() which conduct the prune step required by the refined apriori algorithm implementation. Then the algorithm is going to take the itemset from the helper functions, and check and keep the ones above the min sup threshold. (Step 3)
    - **generate_rules():** This function takes the itemsets and compute association rules according to the project requirement: LHS(1+) → RHS(1), and then return the ones above the min confidence. (Step 4)

6. Command line: python3 main.py INTEGRATED-DATASET.csv 0.15 0.6

Brief explanation on the results: After several trials and interpretation on the results, we have chosen min sup = 0.15 and min conf = 0.6 as our parameters, the combination of the parameters ensures the validity of our results, as the threshold is not set too low, without leaving too few confidence rules for us to interpret. We have discovered a few dozen rules in regards of License type, Violation time, Violation In Front Of Or Opposite, Vehicle Body Type, State and etc. I will briefly talk about some example rules that we think are compelling:
[PAS, 21] => [Morning]: This rule indicates that if a violation of "No Parking – Street Cleaning" occurs with the driver's licence type being "Passenger" type, the violation tends to occur in morning hours. We think this rule is in line with our common sense – think of someone going to work in hurry, and makes a stop by the curb to get breakfast but got tickets – pretty typical stories that we would hear from our friends. Knowing this association would hopefully make drivers be more aware of that violation, and help them be more careful especially in the morning hours.
[Q] => [PAS]: This rule indicates that if the violation occurs in "Queens", then the license type of the driver is likely to be "Passenger" type – We found this rule very useful, especially for people who live in Queens and have the type of license – Be careful.
[SUBN, Morning] => [F]: We found this rule very interesting, since it indicates that for Vehicle type beings "SUBN" (Suburban or SUV) that violates the rule in the morning, the violation usually occurs "In the front" from a specific area or object where parking is prohibited (e.g. fire hydrant). There might be some psychologies behind this phenomenna that we don't know of, but we found this rule interesting and surprising.
These are some examples of our association rules generated for parking violations in the NY area, and interpreting them is fun for us, some are helpful and useful for drivers and some are just surprising and interesting phenomena we have discovered. There are many other interesting association rules that we found explanations of. Overall, we think our results of association rules are compelling and backed up by the common sense.