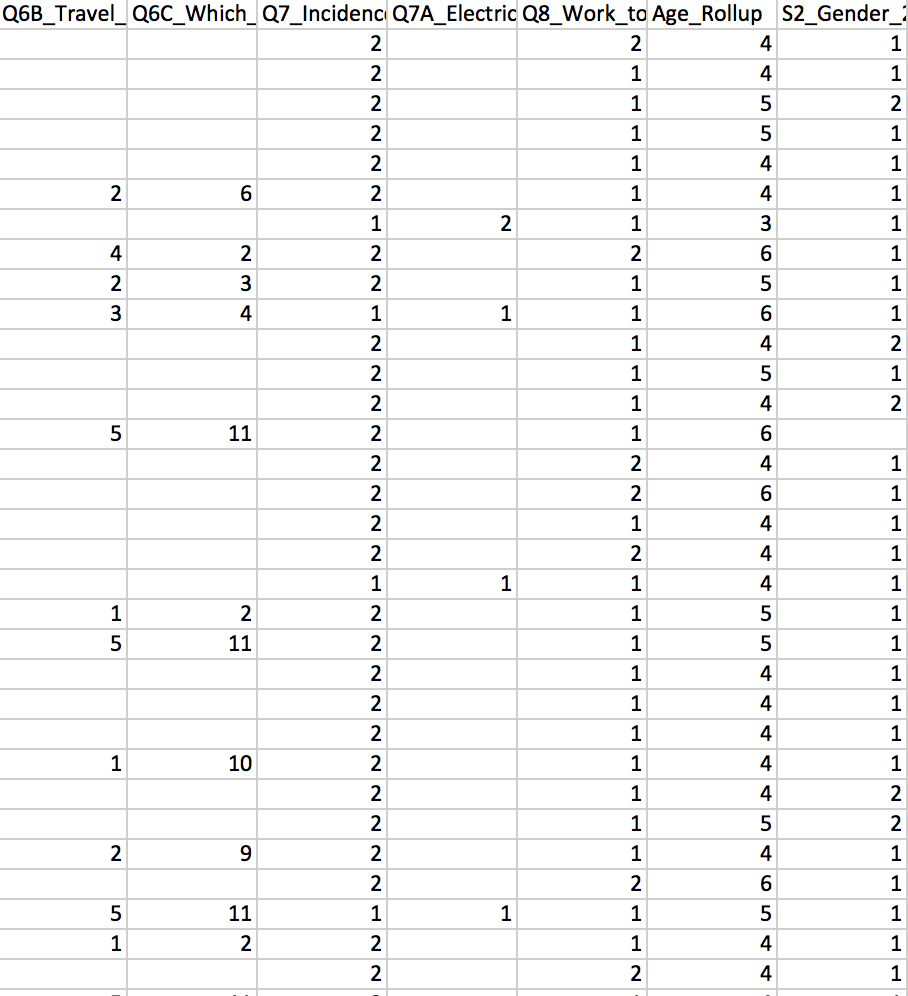
**Decoding Coded SDA Data** (decode\_SDA\_data.py)

****

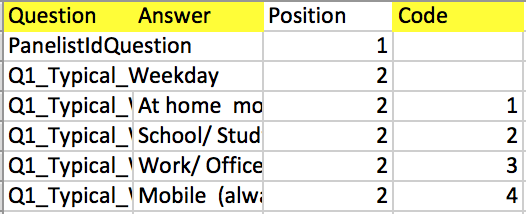
$ python3 decode\_SDA\_data.py –h

-> This will prompt you on what is needed as input.

$ python3 decode\_SDA\_data.py SDA\_coded\_data.csv SDA\_codes.csv new\_file\_name.csv

SDA\_coded\_data.csv: name of the SDA coded dataset (.csv)

SDA\_codes.csv: SDA code file matched to values (.csv)

****Format of data

* 1st column: Question (must match header in SDA coded dataset)
* 2nd column: Answer
* 4th column: Code

new\_file\_name.csv: New file name for the decoded SDA data to be saved in

**-**

After decoding the data, it is now ready to be input into pre-validation.py.

**Pre-validation** (pre-validation.py)

**Loading & Parsing of SDA data**

$ python3.5 pre-validation.py SDA\_data.csv

Output:

Removing any rows with NULL primary key (in column 'S8\_phone\_numbers\_1')...

8 row(s) removed.

Validity Check: Input data is valid.

(validity check according to formats specified in input\_format\_specifications.txt)

**Transform Address Data into Lat-Longs (Optional Task 1)**

**- Example:**

Please select an optional task (1, 2 or 3):

1. Transform Address Data into Lat-Longs

2. Grouping of categories

3. Filter

Press 0 to skip. 1

Please specify columns for:

1) building info/addresses

2) districts (if any) 🡪 district helps to source postcodes from text addresses more accurately. It is optional to input a district column.

Q1\_Typical\_Weekday

Q2\_Profile

Q2A\_Regular\_Employee\_Business\_Owner

Q2B\_Office\_or\_mobile

Q2C\_District

Q2D\_Building\_information

Q4\_Primary\_mode\_of\_transport\_0\_Bus

Q4\_Primary\_mode\_of\_transport\_0\_MRT

Q4\_Primary\_mode\_of\_transport\_0\_CarMotorbike

Q4\_Primary\_mode\_of\_transport\_0\_Taxi

Q4B1\_MRT\_station\_start

Q4B2\_MRT\_station\_alighting\_at

Q5B\_Bus\_number

Q6C\_Which\_major\_expressway

Age\_Rollup

S2\_Gender\_2

S3\_Postcode

S8\_phone\_numbers\_1

Or press 0 to exit transform-address-data function. Q2D\_Building\_information Q2C\_District

Remove rows with blanks, NAs and NILs in Q2D\_Building\_information column? (y/n)

Or press 0 to exit transform-address-data function. y

3 row(s) removed.

Converting textual building info/addresses into postal codes, as derived from google.com...

http://www.google.com/search?q=MoM+Chinatown+singapore+building+address

http://www.google.com/search?q=MoM+Chinatown+singapore+building

………

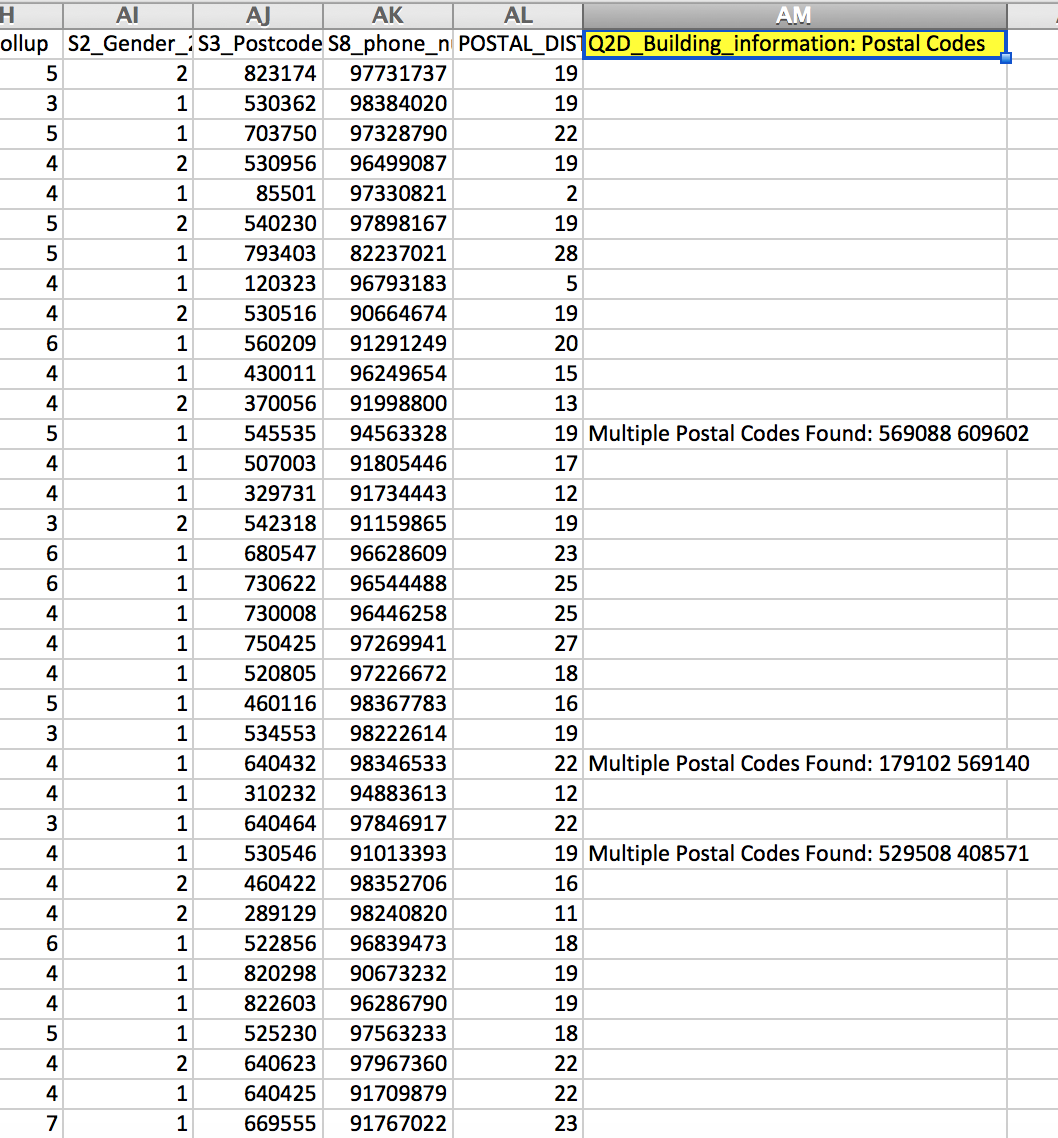
……

…

For rows where no postcodes were found, or where multiple postcodes were found: refer to 'No\_Single\_Postcode\_Found.csv'. User may manually match text addresses to postcodes and upload file back into program to continue data processing.



Example data of No\_Single\_Postcode\_Found.csv:



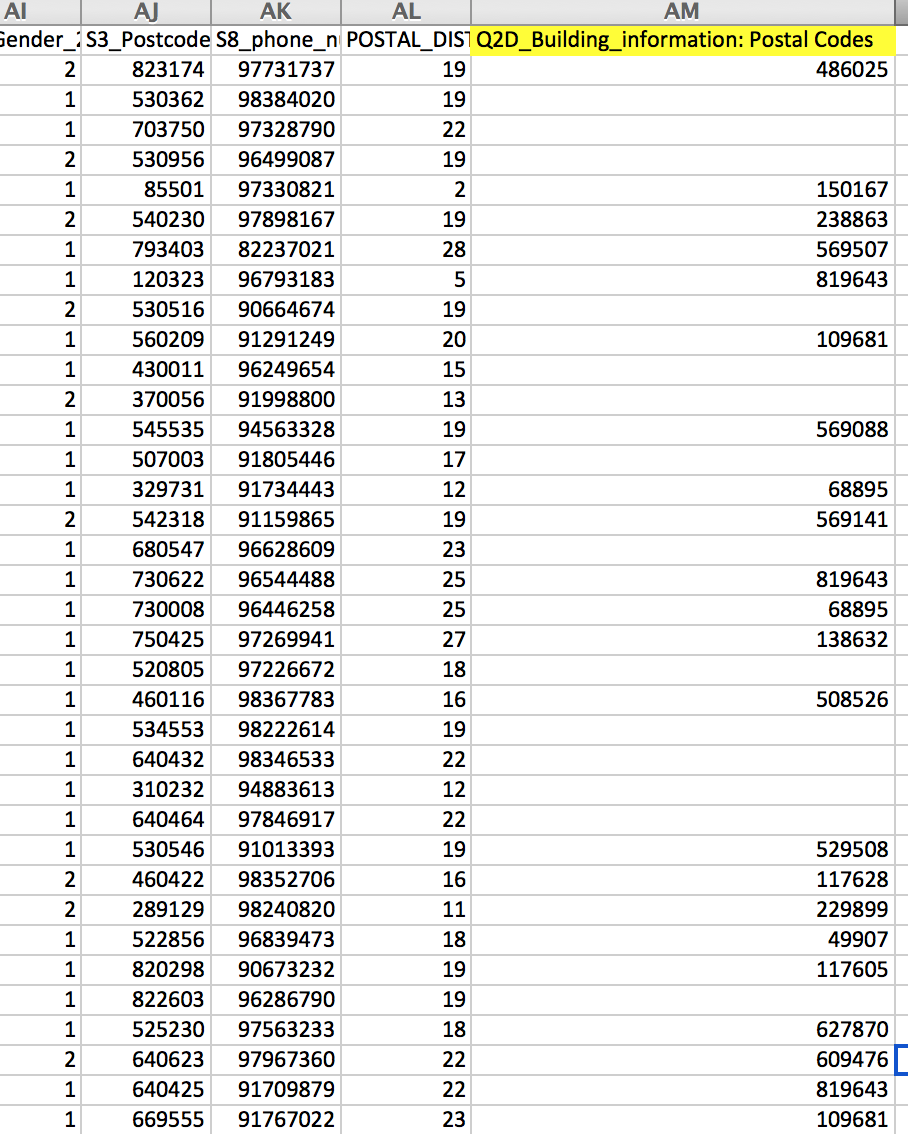
Would you like to input any manually matched rows back into dataset? (y/n)y

Enter the name of the file (.csv) which you would like to combine with the rest of the postal code data. File must be in the same directory as program. (Program assumes that postal codes are in the last column of dataset, and that column headings are included.)No\_Single\_Postcode\_Found.csv



Here, the user modified the No\_Single\_Postcode\_Found.csv file that was output by the program; changes were saved under the same file name and file was input back into the program.

Example of No\_Single\_Postcode\_Found.csv with manually added postcodes:



All rows (original postcode rows, text-to-postcode converted rows, and manually converted rows if any) have been combined.

Deriving latitudes and longitudes for all postcodes...

(Rows with blanks, NAs, NILs and erroneous postcodes in the address column will return empty cells for lat-long)

**Grouping of Categories (Optional Task 2)**

Please select an optional task (1, 2 or 3):

1. Transform Address Data into Lat-Longs

2. Grouping of categories

3. Filter

Press 0 to skip. 2

**(a) Applying Grouping on a Single Column**

Please specify column(s) to apply grouping on:

Q1\_Typical\_Weekday

Q2\_Profile

Q2A\_Regular\_Employee\_Business\_Owner

Q2B\_Office\_or\_mobile

Q2C\_District

Q2D\_Building\_information

Q4\_Primary\_mode\_of\_transport\_0\_Bus

Q4\_Primary\_mode\_of\_transport\_0\_MRT

Q4\_Primary\_mode\_of\_transport\_0\_CarMotorbike

Q4\_Primary\_mode\_of\_transport\_0\_Taxi

Q4B1\_MRT\_station\_start

Q4B2\_MRT\_station\_alighting\_at

Q5B\_Bus\_number

Q6C\_Which\_major\_expressway

Age\_Rollup

S2\_Gender\_2

S3\_Postcode

S8\_phone\_numbers\_1

Or press 0 to exit grouping function. Age\_Rollup

Please specify categories to group.

25-39 years

40-49 years

18-24 years

50-59 years

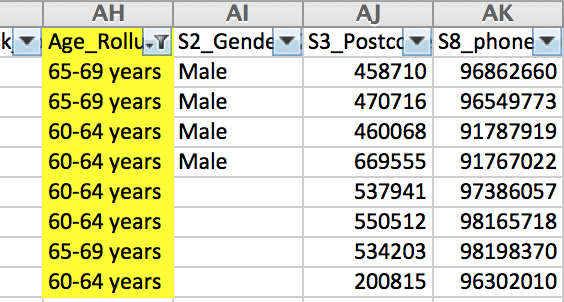
65-69 years

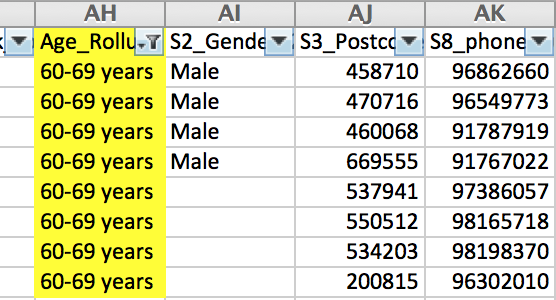
60-64 years

"60-64 years" "65-69 years”  Enclose a category in quotes if the category name contains whitespace

Please specify a name for the new category (any name besides 0). Or press 0 to exit grouping function 60-69 years

8 rows were renamed under new category name.







**(b) Applying Grouping on Multiple Columns**

Please specify column(s) to apply grouping on:

Q1\_Typical\_Weekday

Q2\_Profile

Q2A\_Regular\_Employee\_Business\_Owner

Q2B\_Office\_or\_mobile

Q2C\_District

Q2D\_Building\_information

Q4\_Primary\_mode\_of\_transport\_0\_Bus

Q4\_Primary\_mode\_of\_transport\_0\_MRT

Q4\_Primary\_mode\_of\_transport\_0\_CarMotorbike

Q4\_Primary\_mode\_of\_transport\_0\_Taxi

Q4B1\_MRT\_station\_start

Q4B2\_MRT\_station\_alighting\_at

Q5B\_Bus\_number

Q6C\_Which\_major\_expressway

Age\_Rollup

S2\_Gender\_2

S3\_Postcode

S8\_phone\_numbers\_1

Or press 0 to exit grouping function. Q4\_Primary\_mode\_of\_transport\_0\_CarMotorbike Q4\_Primary\_mode\_of\_transport\_0\_Taxi

Please specify a new category name for records where Columns "Q4\_Primary\_mode\_of\_transport\_0\_CarMotorbike", "Q4\_Primary\_mode\_of\_transport\_0\_Taxi" have values of "0", "0" respectively.

0

Please specify a new category name for records where Columns "Q4\_Primary\_mode\_of\_transport\_0\_CarMotorbike", "Q4\_Primary\_mode\_of\_transport\_0\_Taxi" have values of "Car/ Motorbike", "0" respectively.

Car/ Motorbike

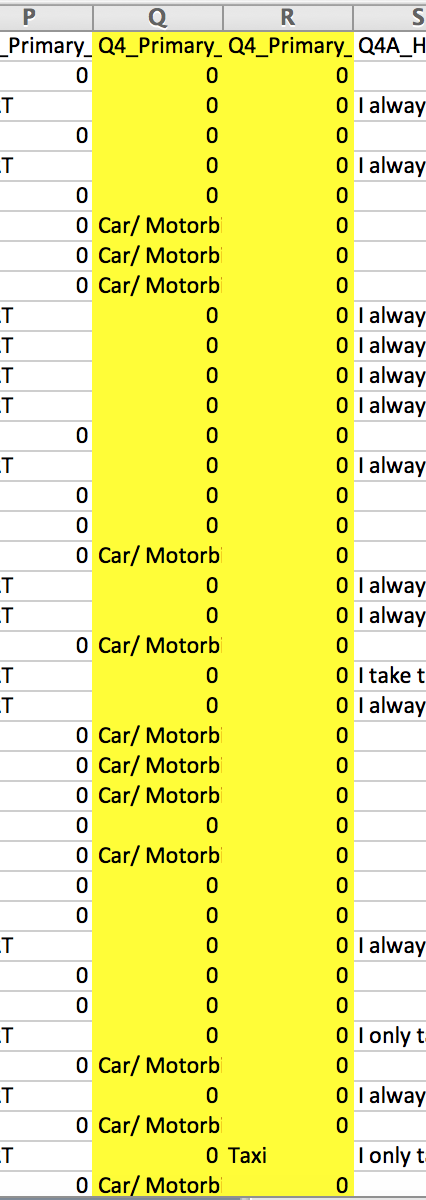
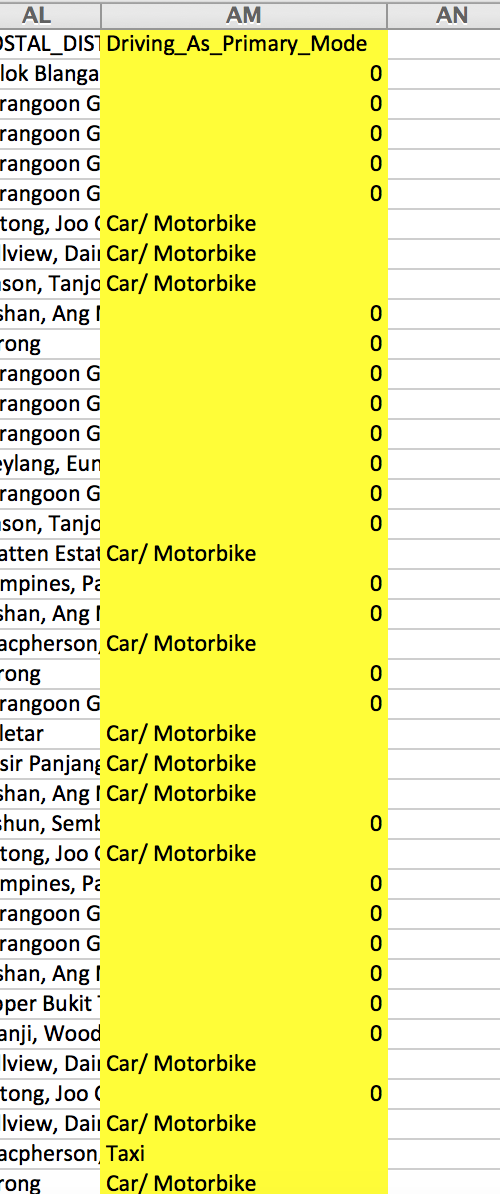
Please specify a new category name for records where Columns "Q4\_Primary\_mode\_of\_transport\_0\_CarMotorbike", "Q4\_Primary\_mode\_of\_transport\_0\_Taxi" have values of "0", "Taxi" respectively.

Taxi

Please specify a new category name for records where Columns "Q4\_Primary\_mode\_of\_transport\_0\_CarMotorbike", "Q4\_Primary\_mode\_of\_transport\_0\_Taxi" have values of "Car/ Motorbike", "Taxi" respectively.

Both

Columns have been combined. Please input a column name for the combined column. Ensure that you do not include whitespace within the column name. Driving\_As\_Primary\_Mode

****



**Filter (Optional Task 3)**

Please select an optional task (1, 2 or 3):

1. Transform Address Data into Lat-Longs

2. Grouping of categories

3. Filter

Press 0 to skip. 3

1. **Exclude blanks, NAs and NILs**

Please select a filter option:

1. Exclude blanks, NAs and NILs

2. Filter based on categories

Press 0 to exit filter function. 1

Please specify a column to filter out blanks, NAs and NILs from:

Q1\_Typical\_Weekday

Q2\_Profile

Q2A\_Regular\_Employee\_Business\_Owner

Q2B\_Office\_or\_mobile

Q2C\_District

Q2D\_Building\_information

Q4\_Primary\_mode\_of\_transport\_0\_Bus

Q4\_Primary\_mode\_of\_transport\_0\_MRT

Q4\_Primary\_mode\_of\_transport\_0\_CarMotorbike

Q4\_Primary\_mode\_of\_transport\_0\_Taxi

Q4B1\_MRT\_station\_start

Q4B2\_MRT\_station\_alighting\_at

Q5B\_Bus\_number

Q6C\_Which\_major\_expressway

Age\_Rollup

S2\_Gender\_2

S3\_Postcode

S8\_phone\_numbers\_1

Press 0 to exit filter function. Q4B1\_MRT\_station\_start

205 row(s) removed.

1. **Filter based on categories**

Please select a filter option:

1. Exclude blanks, NAs and NILs

2. Filter based on categories

Press 0 to exit filter function. 2

1. **1 column specified; Include specific categories**

Please specify column(s) to apply categorical filter on:

Q1\_Typical\_Weekday

Q2\_Profile

Q2A\_Regular\_Employee\_Business\_Owner

Q2B\_Office\_or\_mobile

Q2C\_District

Q2D\_Building\_information

Q4\_Primary\_mode\_of\_transport\_0\_Bus

Q4\_Primary\_mode\_of\_transport\_0\_MRT

Q4\_Primary\_mode\_of\_transport\_0\_CarMotorbike

Q4\_Primary\_mode\_of\_transport\_0\_Taxi

Q4B1\_MRT\_station\_start

Q4B2\_MRT\_station\_alighting\_at

Q5B\_Bus\_number

Q6C\_Which\_major\_expressway

Age\_Rollup

S2\_Gender\_2

S3\_Postcode

S8\_phone\_numbers\_1

Press 0 to exit filter function. Q2C\_District

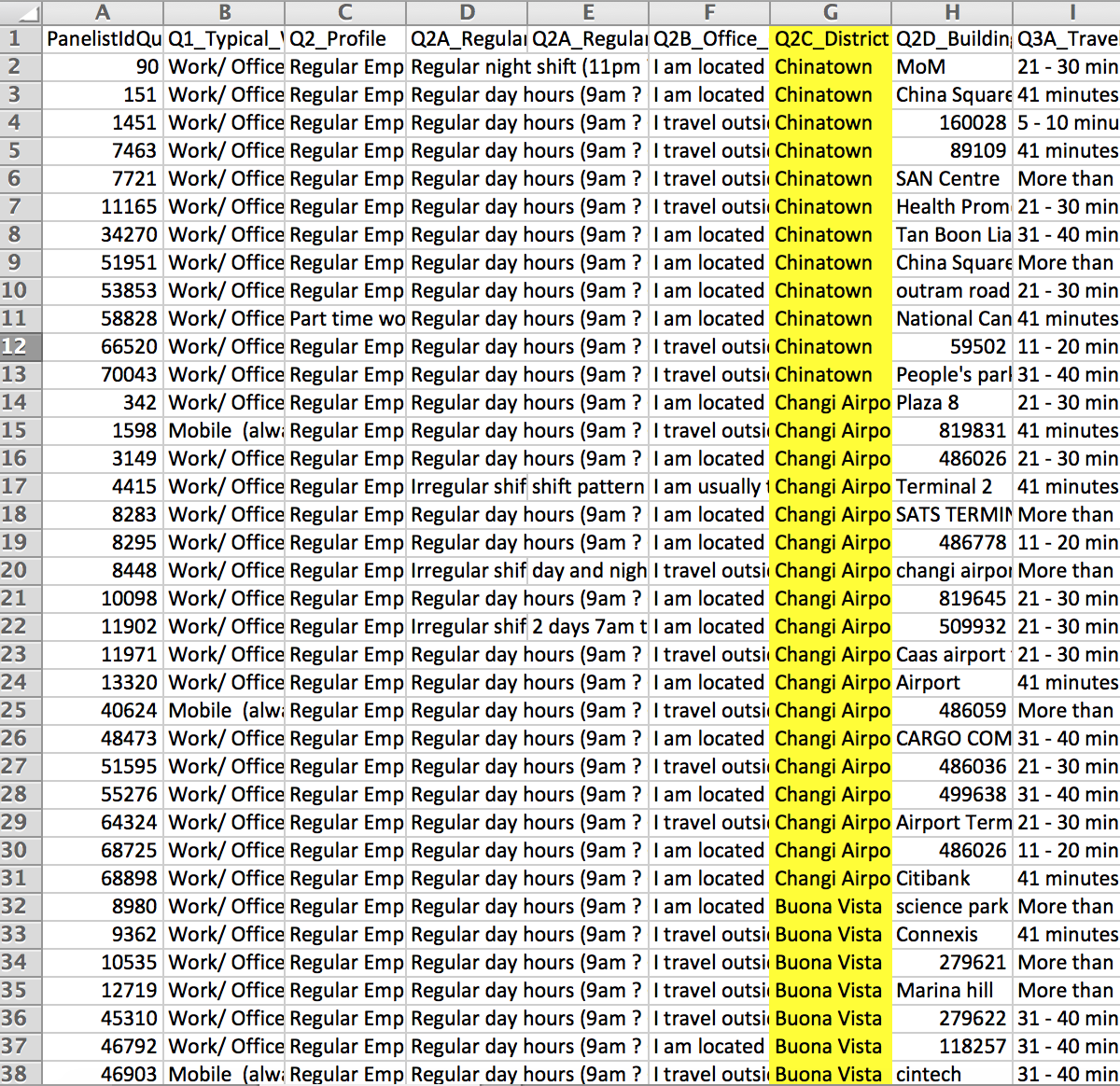
1. Include specific categories or 2. Exclude specific categories? 1

Key in the category(s) by row: (Press 0 to exit filter function) Chinatown;"Changi Airport";"Buona Vista"

375 row(s) removed. 

Only rows with “Chinatown” OR “Changi Airport” OR “Buona Vista” in the Q2C\_District column are included in the dataset.

After applying filter:

****

**(ii) Multiple columns specified; Include specific categories**

Please specify column(s) to apply categorical filter on:

Q1\_Typical\_Weekday

Q2\_Profile

Q2A\_Regular\_Employee\_Business\_Owner

Q2B\_Office\_or\_mobile

Q2C\_District

Q2D\_Building\_information

Q4\_Primary\_mode\_of\_transport\_0\_Bus

Q4\_Primary\_mode\_of\_transport\_0\_MRT

Q4\_Primary\_mode\_of\_transport\_0\_CarMotorbike

Q4\_Primary\_mode\_of\_transport\_0\_Taxi

Q4B1\_MRT\_station\_start

Q4B2\_MRT\_station\_alighting\_at

Q5B\_Bus\_number

Q6C\_Which\_major\_expressway

Age\_Rollup

S2\_Gender\_2

S3\_Postcode

S8\_phone\_numbers\_1

Press 0 to exit filter function. Q1\_Typical\_Weekday Q2B\_Office\_or\_mobile

1. Include specific categories or 2. Exclude specific categories? 1

Key in the category(s) by row: (Press 0 to exit filter function) "Work/ Office" "I am located in the office most of the time";"Mobile (always on the move, travelling to different areas)" "I am usually travelling outside of the office for work / on the road"

112 row(s) removed. 

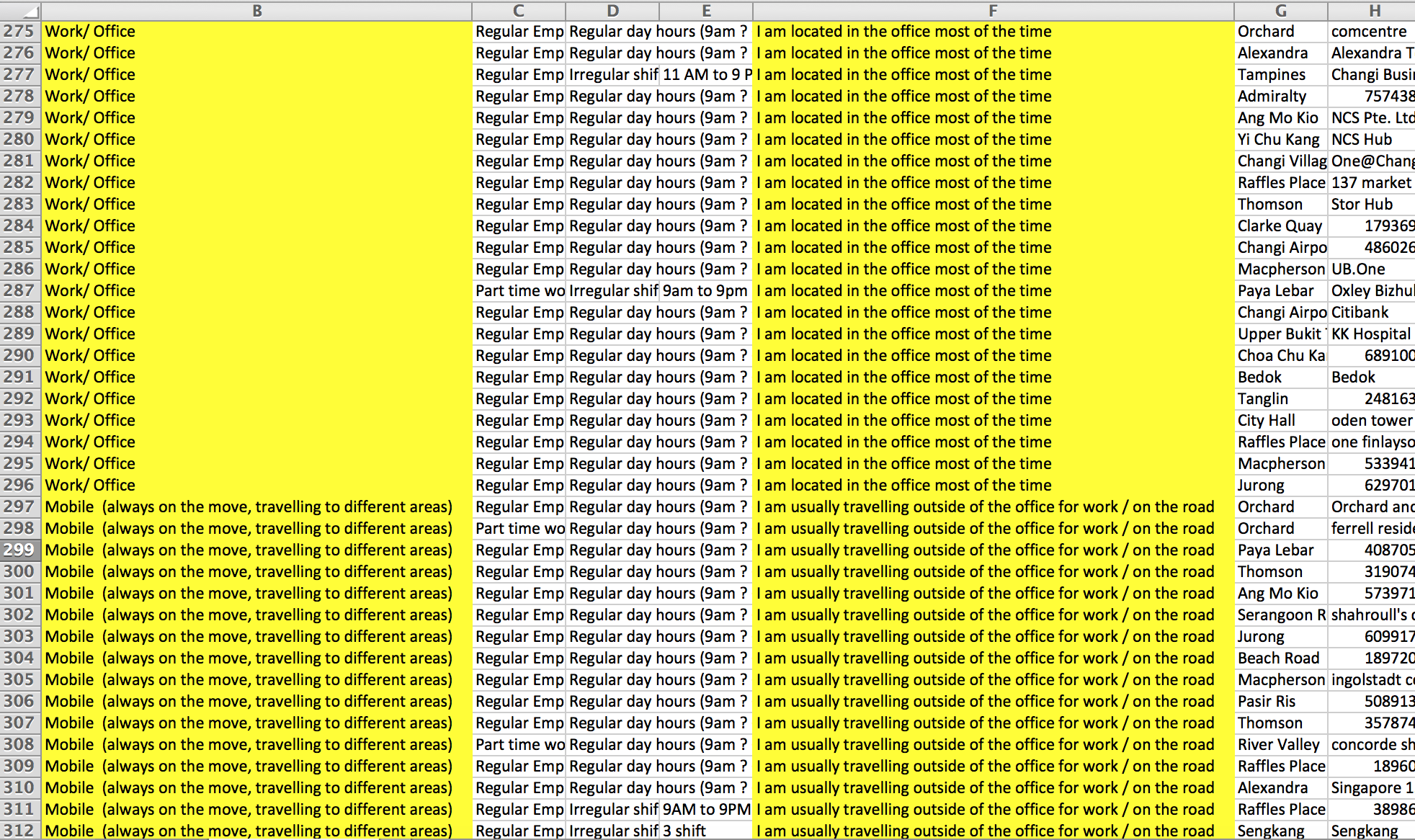
A row is included if:

* It contains “Work/ Office” in Q1\_Typical\_Weekday column AND “I am located in the office most of the time” in Q2B\_Office\_or\_mobile column

OR

* It contains “Mobile (always on the move, travelling to different areas)" in Q1\_Typical\_Weekday column AND “I am usually travelling outside of the office for work / on the road” in Q2B\_Office\_or\_mobile column

* AND: delimited by space
* OR: delimited by “;”

After applying filter:

**(iii) 1 column specified; Exclude specific categories**

Please specify column(s) to apply categorical filter on:

Q1\_Typical\_Weekday

Q2\_Profile

Q2A\_Regular\_Employee\_Business\_Owner

Q2B\_Office\_or\_mobile

Q2C\_District

Q2D\_Building\_information

Q4\_Primary\_mode\_of\_transport\_0\_Bus

Q4\_Primary\_mode\_of\_transport\_0\_MRT

Q4\_Primary\_mode\_of\_transport\_0\_CarMotorbike

Q4\_Primary\_mode\_of\_transport\_0\_Taxi

Q4B1\_MRT\_station\_start

Q4B2\_MRT\_station\_alighting\_at

Q5B\_Bus\_number

Q6C\_Which\_major\_expressway

Age\_Rollup

S2\_Gender\_2

S3\_Postcode

S8\_phone\_numbers\_1

Press 0 to exit filter function. Age\_Rollup

1. Include specific categories or 2. Exclude specific categories? 2

Key in the category(s) by row: (Press 0 to exit filter function) "50-59 years";"40-49 years"

132 row(s) removed. 

Rows with “50-59 years” OR “40-49 years” in the Age\_Rollup column are excluded from the dataset.

**(iv) Multiple columns specified; Exclude specific categories**

Please specify column(s) to apply categorical filter on:

Q1\_Typical\_Weekday

Q2\_Profile

Q2A\_Regular\_Employee\_Business\_Owner

Q2B\_Office\_or\_mobile

Q2C\_District

Q2D\_Building\_information

Q4\_Primary\_mode\_of\_transport\_0\_Bus

Q4\_Primary\_mode\_of\_transport\_0\_MRT

Q4\_Primary\_mode\_of\_transport\_0\_CarMotorbike

Q4\_Primary\_mode\_of\_transport\_0\_Taxi

Q4B1\_MRT\_station\_start

Q4B2\_MRT\_station\_alighting\_at

Q5B\_Bus\_number

Q6C\_Which\_major\_expressway

Age\_Rollup

S2\_Gender\_2

S3\_Postcode

S8\_phone\_numbers\_1

Press 0 to exit filter function. Q4\_Primary\_mode\_of\_transport\_0\_Bus Q4\_Primary\_mode\_of\_transport\_0\_MRT

1. Include specific categories or 2. Exclude specific categories? 2

Key in the category(s) by row: (Press 0 to exit filter function) Bus MRT;0 0

228 row(s) removed. 

A row is excluded if:

* It contains “Bus” in Q4\_Primary\_mode\_of\_transport\_0\_Bus column AND “MRT” in Q4\_Primary\_mode\_of\_transport\_0\_MRT column

OR

* It contains “0” in Q4\_Primary\_mode\_of\_transport\_0\_Bus column AND “0” in Q4\_Primary\_mode\_of\_transport\_0\_MRT column

**Mapping MSISDN/IMSIs and outputting final manipulated SDA dataset**

Please select an optional task (1, 2 or 3):

1. Transform Address Data into Lat-Longs

2. Grouping of categories

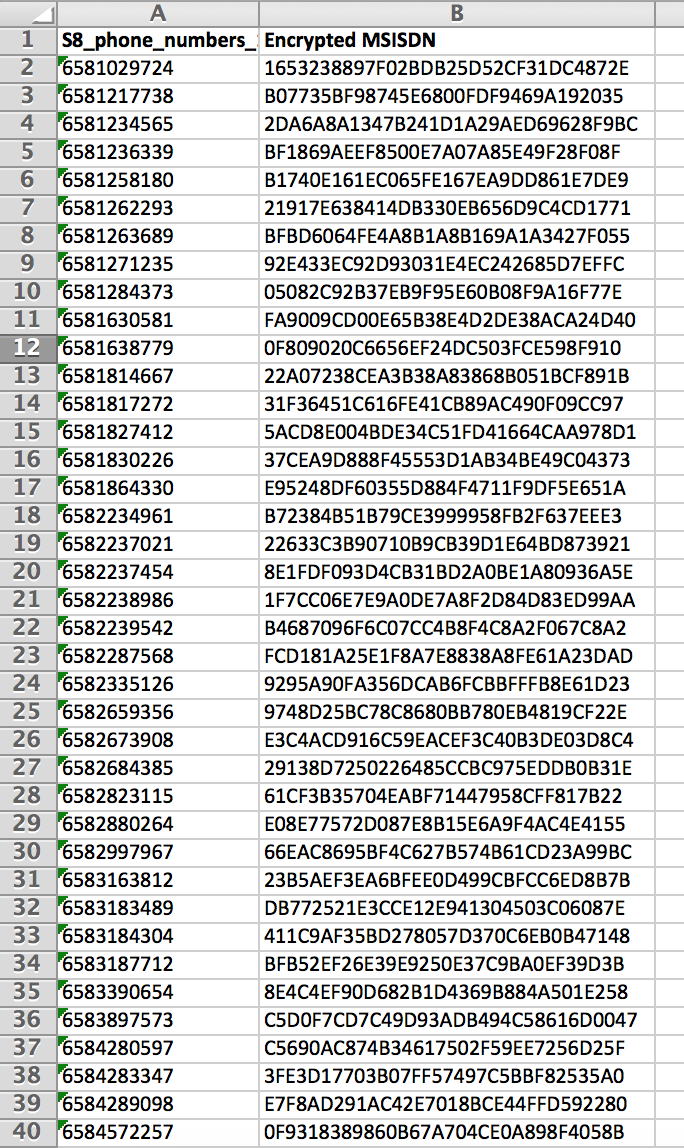
3. Filter

Press 0 to skip. 0

All the remaining rows in this dataset will be used to validate the algorithm output.

Enter the name of the encrypted MSISDN/IMSI file (.csv). File must be in the same directory as program. (Primary key in 1st column, Encrypted MSISDN/IMSI in 2nd column. Column headings are to be included) MSISDN\_mapping\_table.csv





Example of MSISDN/IMSI file

1st column – phone numbers beginning with 65

2nd column – MSISDN/IMSI

Program is now matching encrypted MSISDNs or IMSIs to primary key of each row...

Matching completed. Final SDA validation data stored in 'final\_SDA\_data.csv', format of final data stored in ‘SDA\_data\_format.csv’.

Pre-validation data processing is completed.