CS433/CS603 Programming Assignment Data Wrangling and Aggregation

Please follow the instructions to complete the following Python programs. For each program, please also provide your own testing cases. Please complete them in either Jupyter notebook or with .py file and submit your programs and running results.

Data wrangling on real estate transaction .csv file. Use our class Jupyter note note9_01_dataWrangling as an example to do the following:

- ✓ With Hierarchical Indexing, rearrange the data to create a first dataframe which will have first level of index of city, second level of index of zip, first level of column of type, second level of column of sq_ft and price, please also rename index and column to make them more understandable, such as change zip to zipcode. Display the first eight rows of data. Do some plotting to reflect this new dataframe.
- ✓ Try stack and unstack with the above created dataframe. Display the first eight rows of data.
- ✓ With Hierarchical Indexing, rearrange the data to create a second dataframe which will have first level of index of city, second level of index of zip, columns of bed, bath and sale_date. Please also rename index and column to make them more understandable, such as change zip to zipcode. Display the first eight rows of data.
- ✓ Reshape the above two newly created dataframes and merge them.

Read the article at http://blog.yhat.com/tutorials/7-Aggregation-and-Grouping.html and use our class Jupyter note note9_02_dataAggregation as an example to do further wrangling and aggregation work on the credit-data-non-null.csv file:

- ✓ Complete the exercise given in the article.
- ✓ Add a new column age_group for the dataframe. Divide the age into groups of every 10 years in ascending order, then rearrange the data to match the age_group categorization. Display the first eight rows of data. Do some plotting to reflect it. Use age_group for data aggregation on revolving_utilization_of_unsecured_lines and monthly_income. Display the result.
- ✓ Regroup data using number_real_estate_loans_or_lines and serious_dlqin2yrs, display the first eight rows of data and do some plotting to reflect this it. Then do count, mean, max, min on revolving_utilization_of_unsecured_lines and monthly_income. Display the result. Then select the top five monthly_income values by group and display it.