Assignment 1

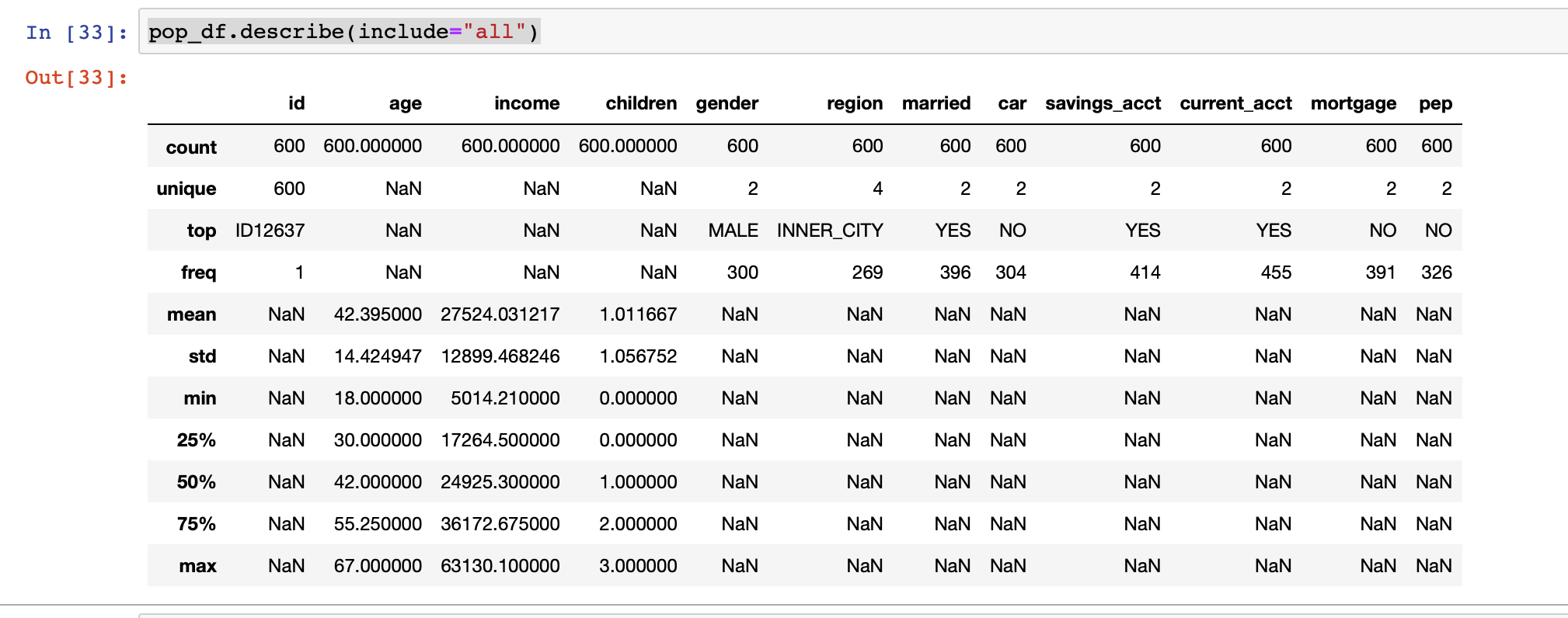
DSC 478

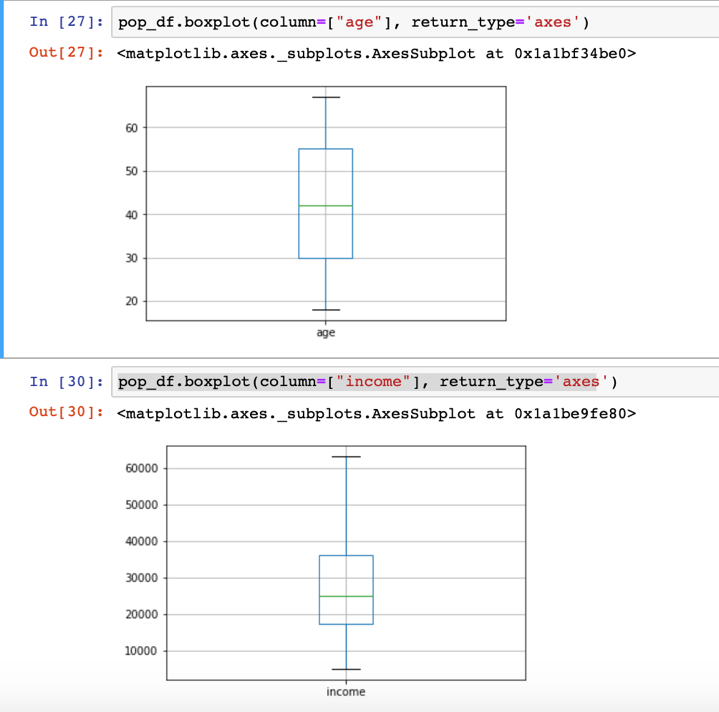
Chaonan Shi

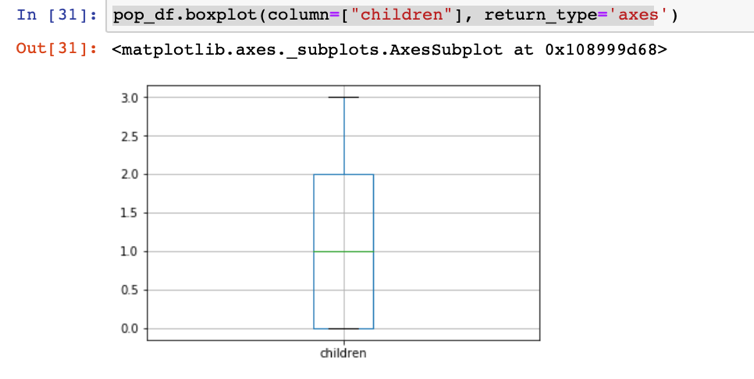
1901412

1. Explore the general characteristics of the data as a whole: examine the means, standard deviations, and other statistics associated with the numerical attributes; show the distributions of values associated with categorical attributes; etc.

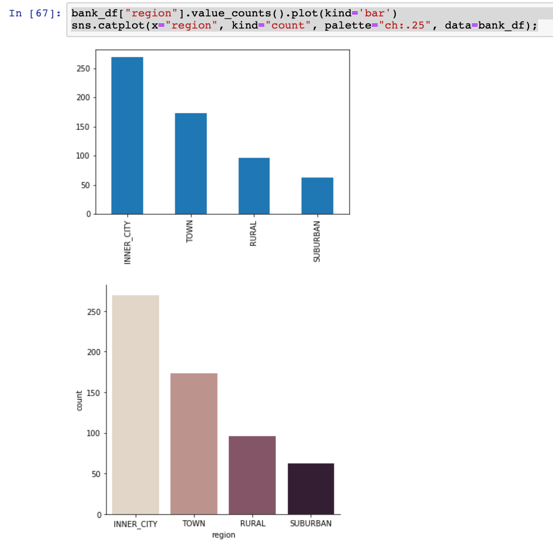
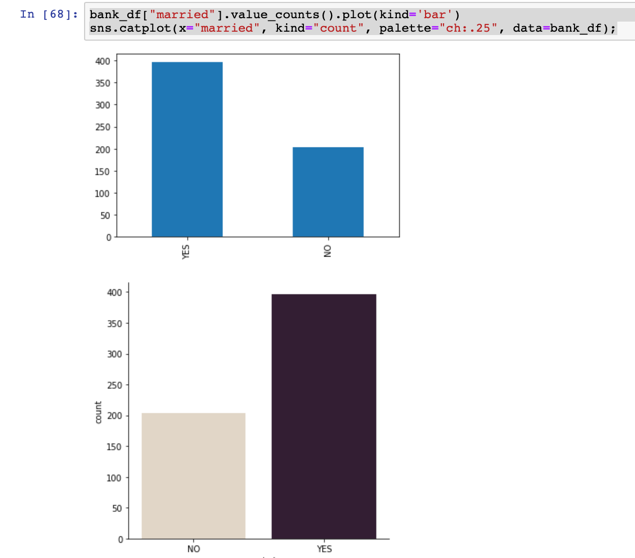
Basic statistical results for continuous features:

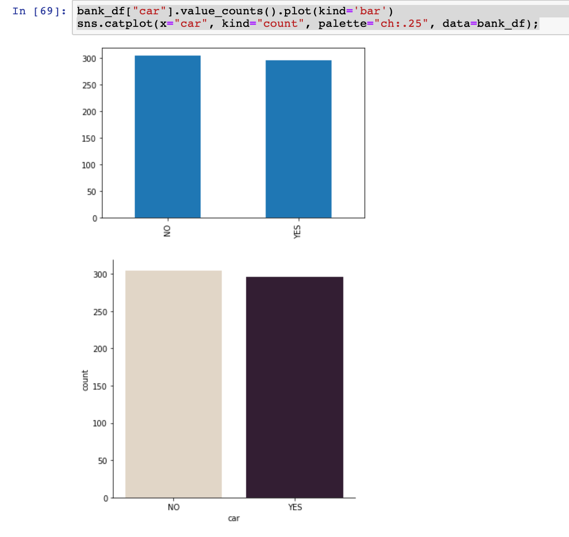
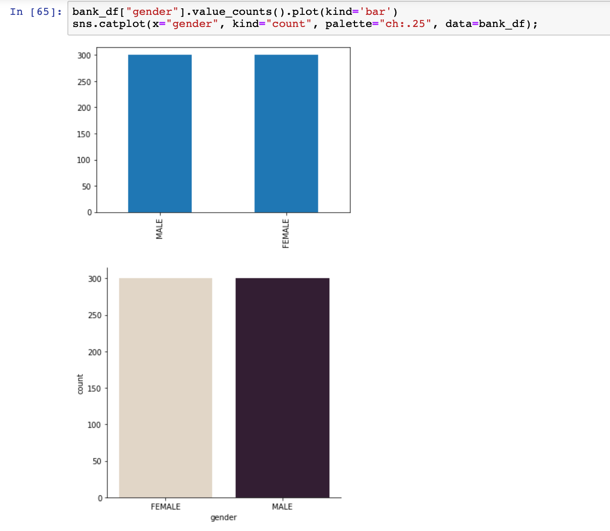


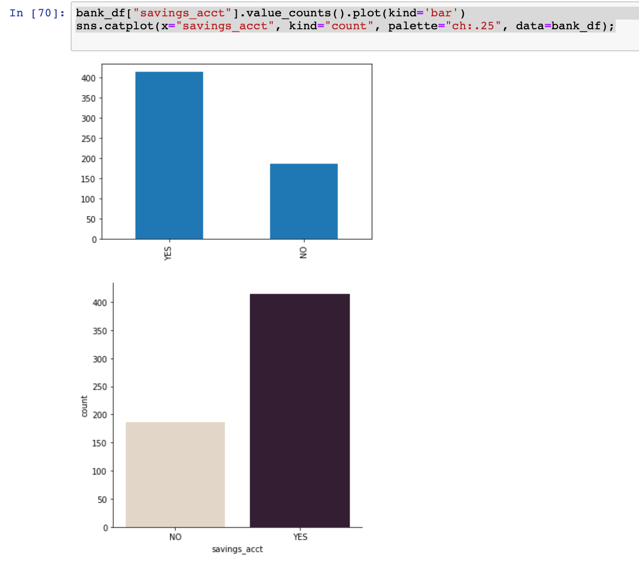
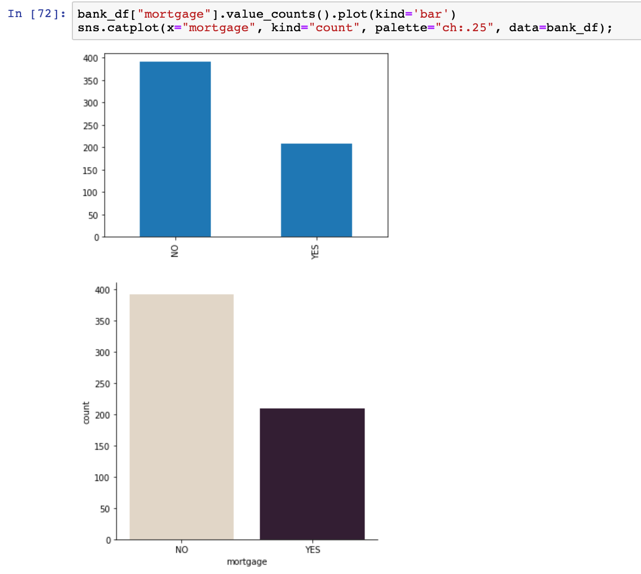


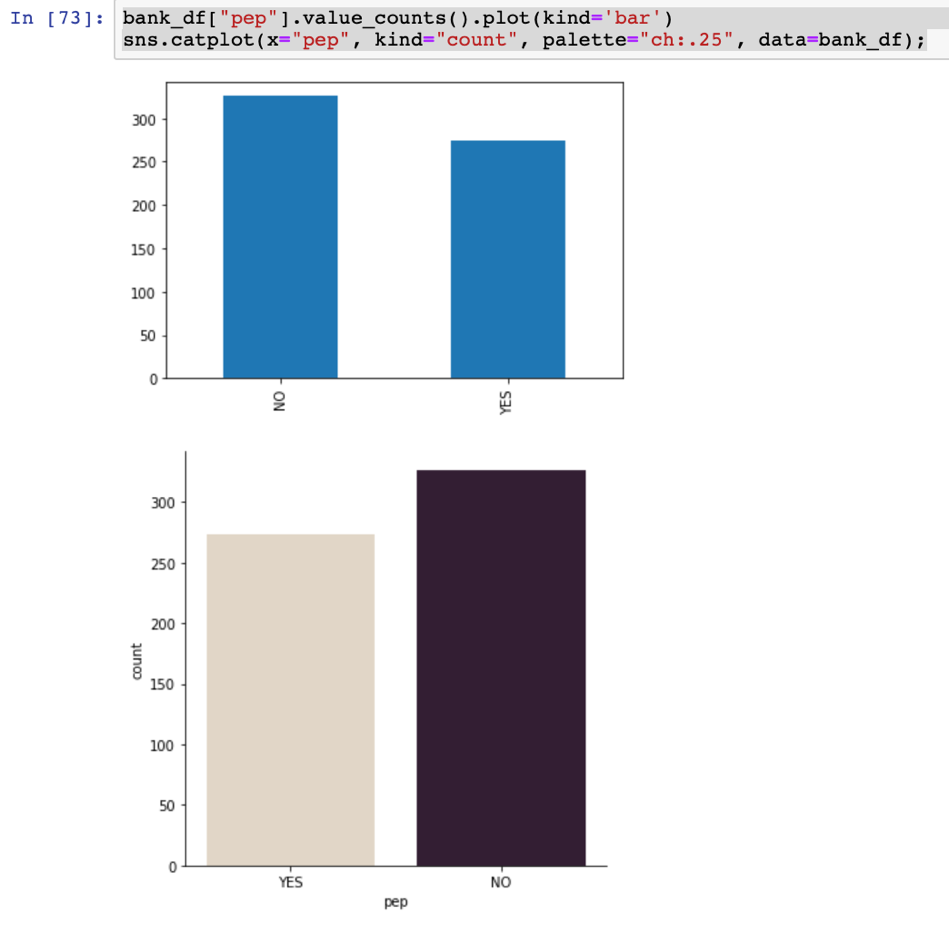


Basic statistical results for categorical features:

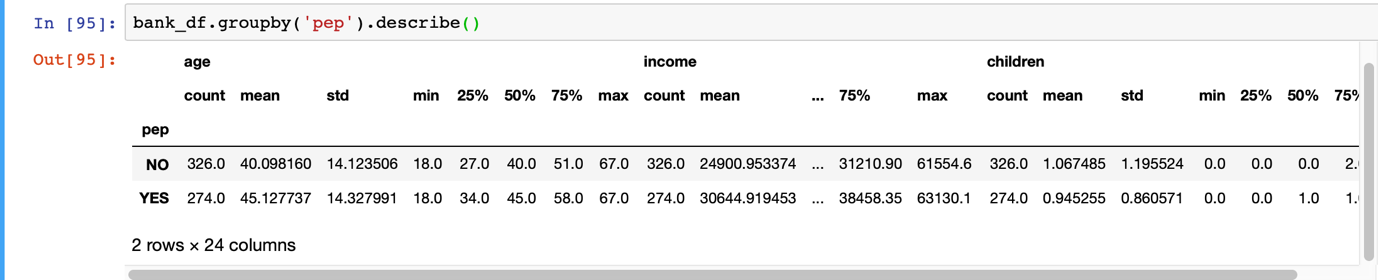


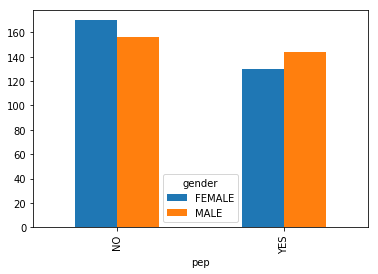
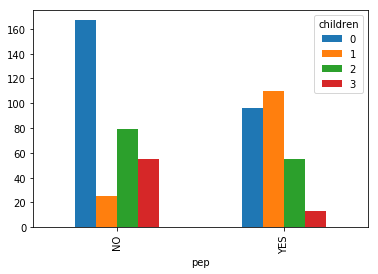


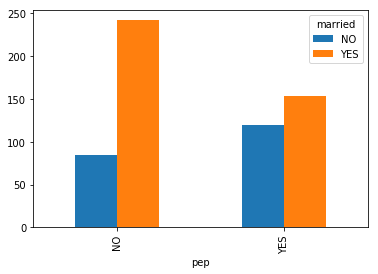
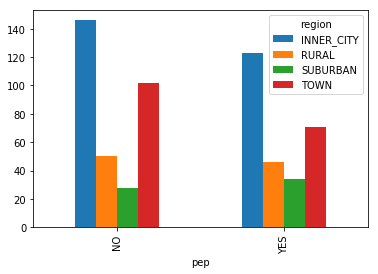
 

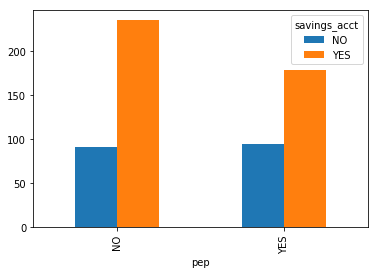
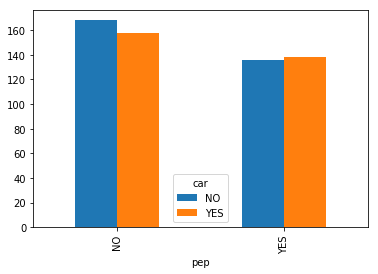


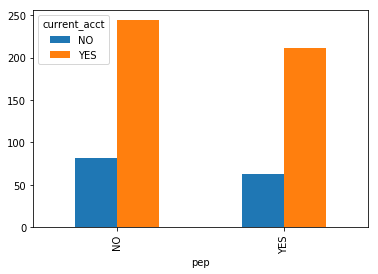
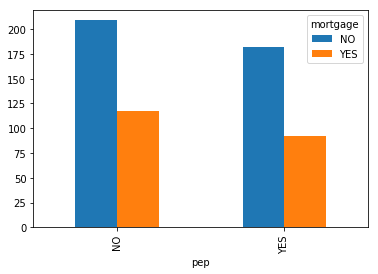
2.Suppose that the hypothetical bank is particularly interested in customers who buy the PEP (Personal Equity Plan) product. Compare and contrast the subsets of customers who buy and don't buy the PEP. Compute summaries (as in part 1) of the selected data with respect to all other attributes. Can you observe any significant differences between these segments of customers? Discuss your observations.



****

****

****

**** 

**Analysis:**

From above analysis, the number of answer of “No” have 16% more than the number of answer of “Yes” (326-274)/326;

Moreover, the standard deviation of answer “No” has less than the answer “Yes”, which means answer “No” has less range of answer “Yes”;

For the categorical variables ‘children’, level ‘0’ has more answer for “No”; For answer “yes”, most of children with level ‘1’;

For the categorical variables ‘gender’, more male answer for “No”;

For the categorical variables ‘car, people who own the car have same level to buy pep, but for the people who do not own the car, they are more likely do not buy pep;

For the categorical variables ‘region’, level with ‘INNER\_CITY’ has more people for answer both “yes” and “No”;

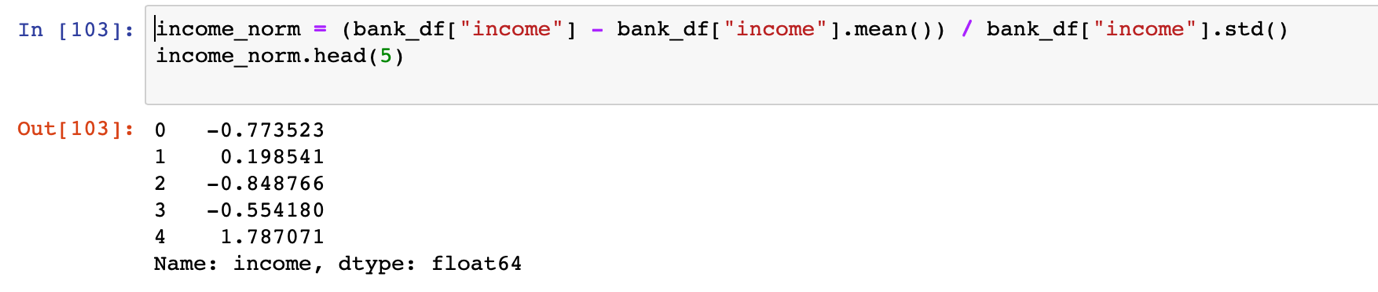
For the categorical variables ‘married’, married people more likely to do not buy pep;

For the categorical variables ‘savings\_acct’, people who do not have savings\_acct have more likely to do not buy pep;

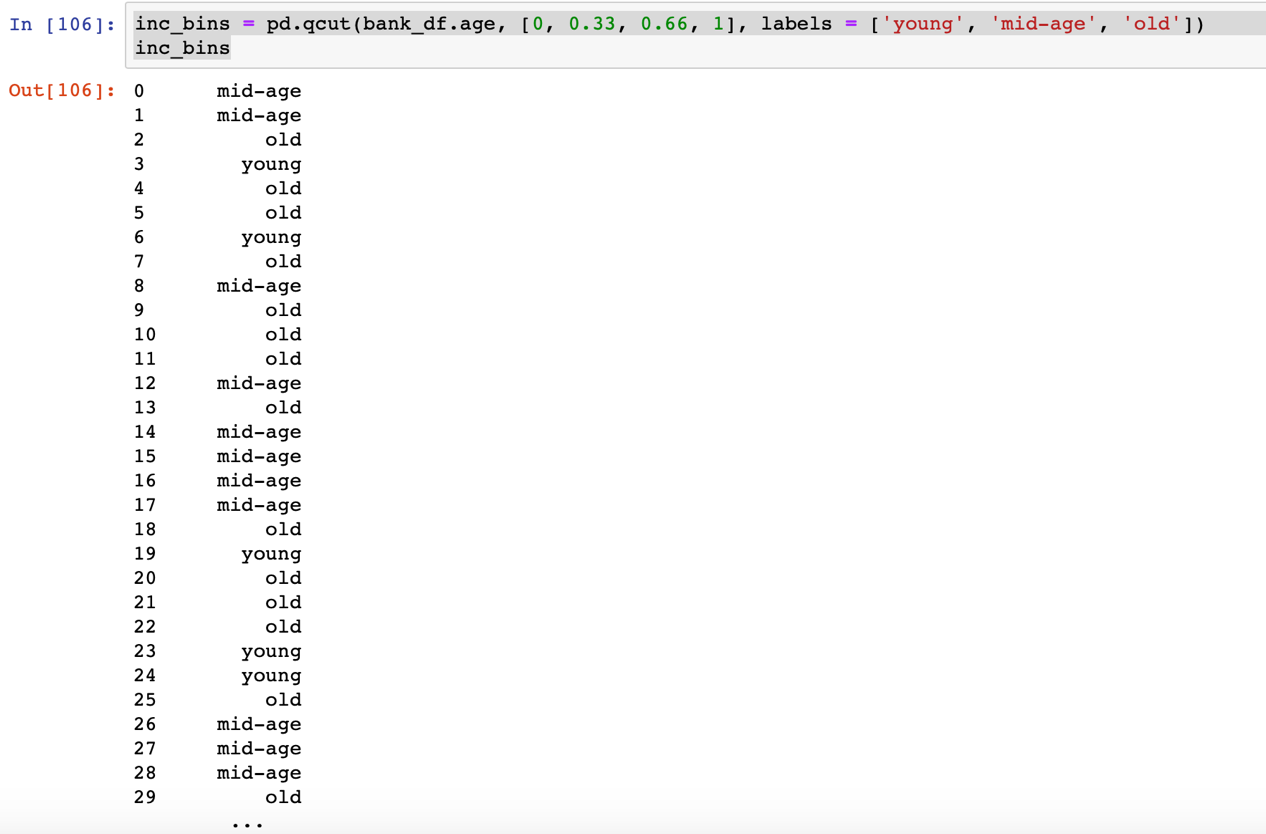
For the categorical variables ‘current\_acct’, people who do have ‘current\_acct’ have more likely to do not buy pep;

For the categorical variables ‘mortgage’, people who do not have ‘mortgage’ have more likely to do not buy pep;

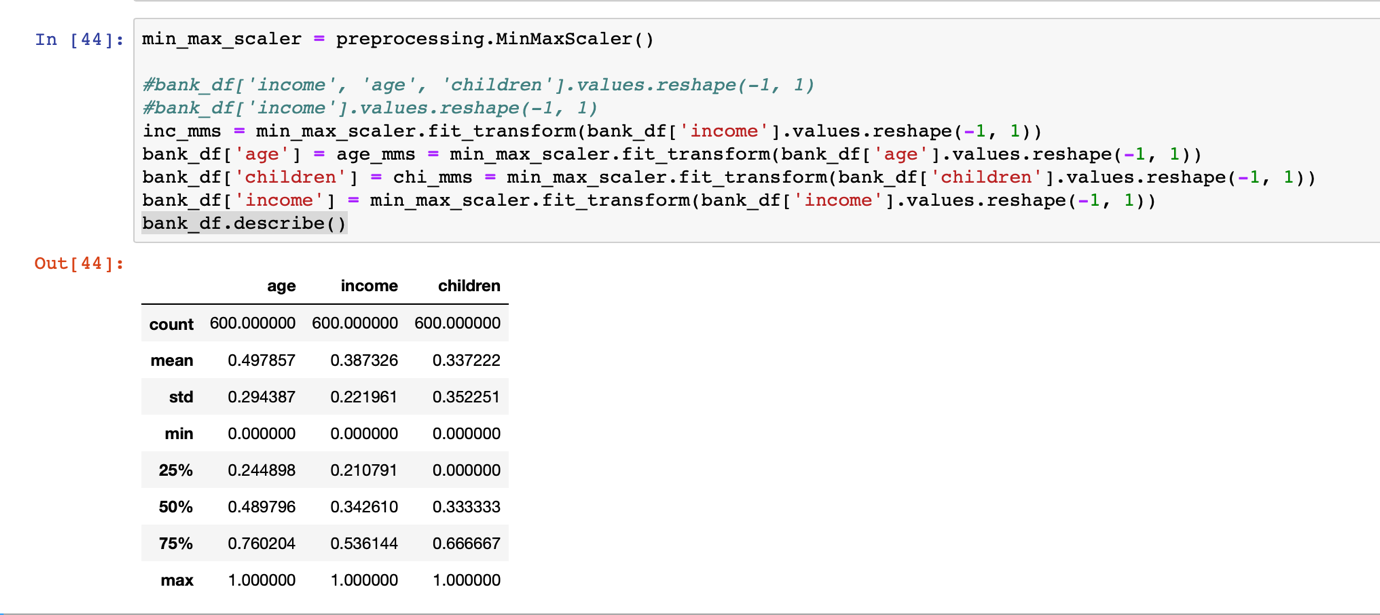
3.Use z-score normalization to standardize the values of the income attribute. [Do not change the original income attribute in the table.]



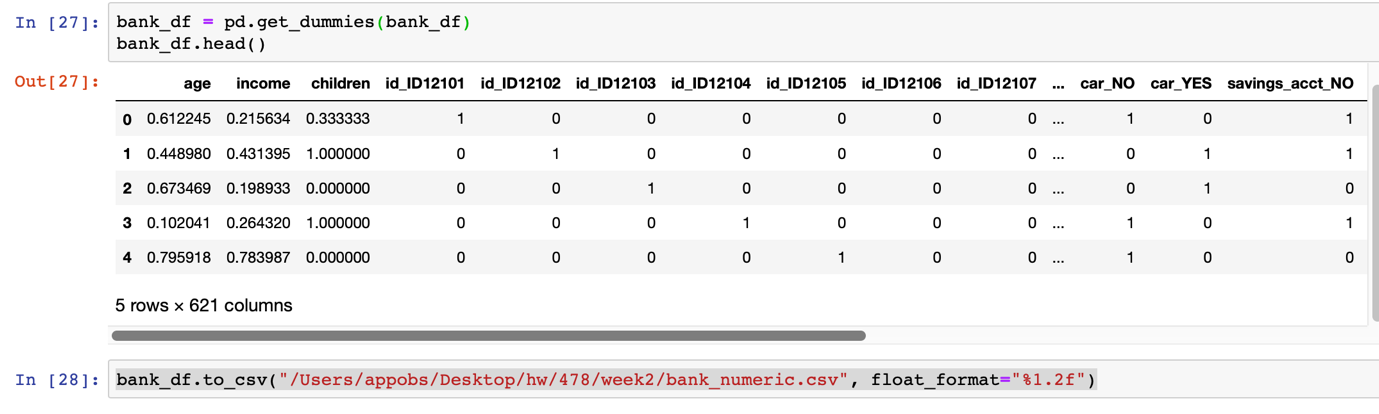
4.Discretize the age attribute into 3 categories (corresponding to "young", "mid-age", and "old"). [Do not change the original age attribute in the table.]



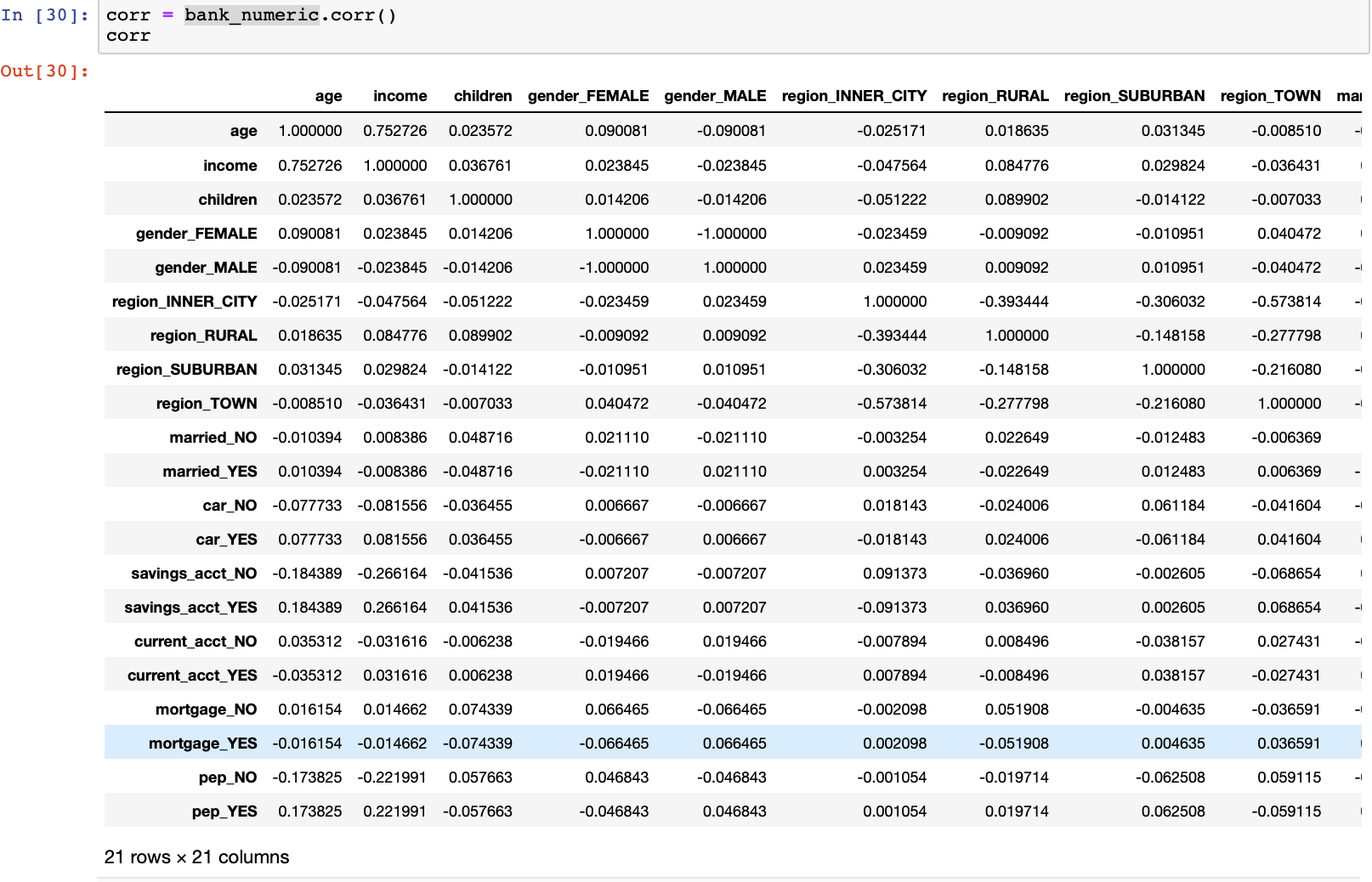
5.Use Min-Max Normalization to transform the values of all numeric attributes  (income, age, children) in the original table (before the transformations in parts 3 and 4 above) onto the range 0.0-1.0.

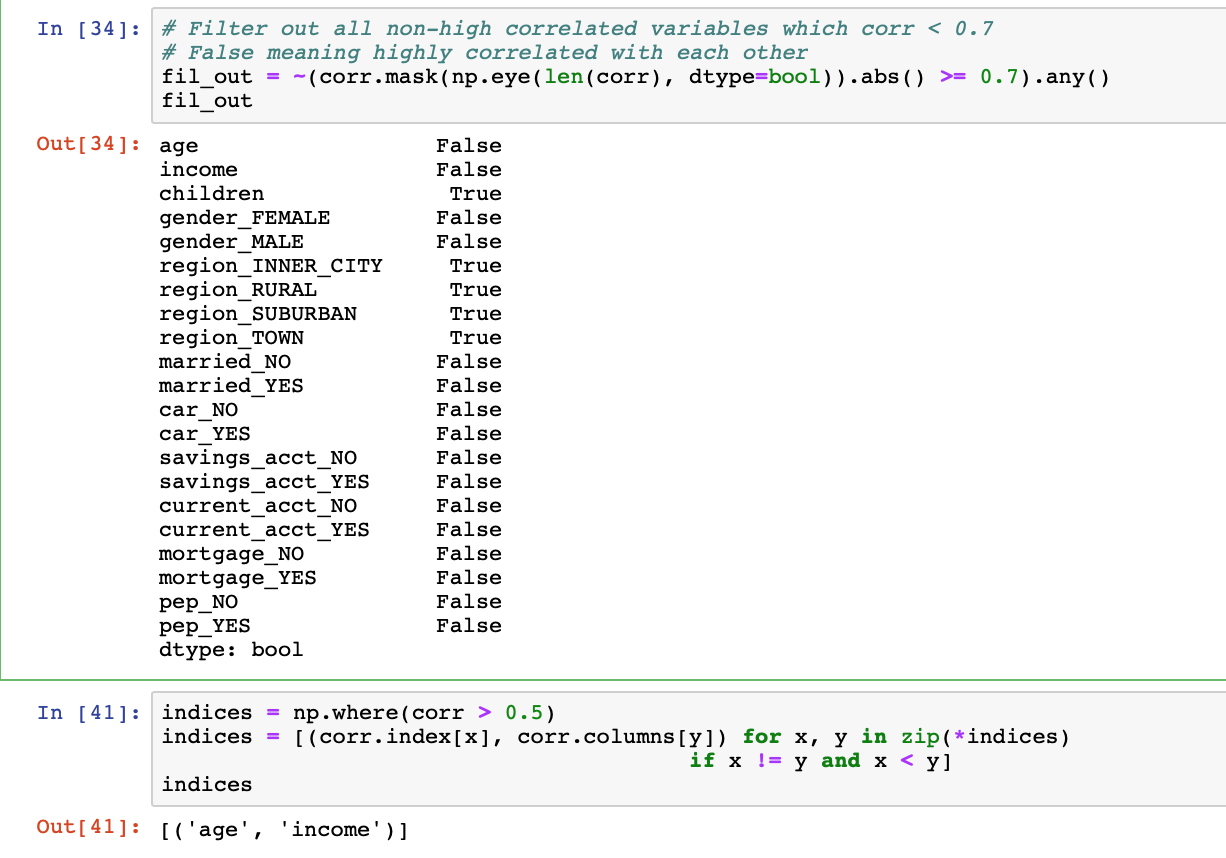


6.Convert the table (after normalization in part 5) into the standard spreadsheet format. Note that this requires converting each categorical attribute into multiple binary ("dummy") attributes (one for each values of the categorical attribute) and assigning binary values corresponding to the presence or not presence of the attribute value in the original record). The numeric attributes should remain unchanged. Save this new table into a file called bank\_numeric.csv and submit it along with your assignment. [Hint: you might consider using the get\_dummies for Pandas data frames.]



7.Using the standardized data set (of the previous part), perform basic correlation analysis among the attributes. Discuss your results by indicating any significant positive or negative correlations among pairs of attributes. You need to construct a complete Correlation Matrix. Be sure to first remove the Customer ID column before creating the correlation matrix. [Hint:you can create the correlation matrix by using the corr() function in Pandas, try at least two corr methods and compare them].



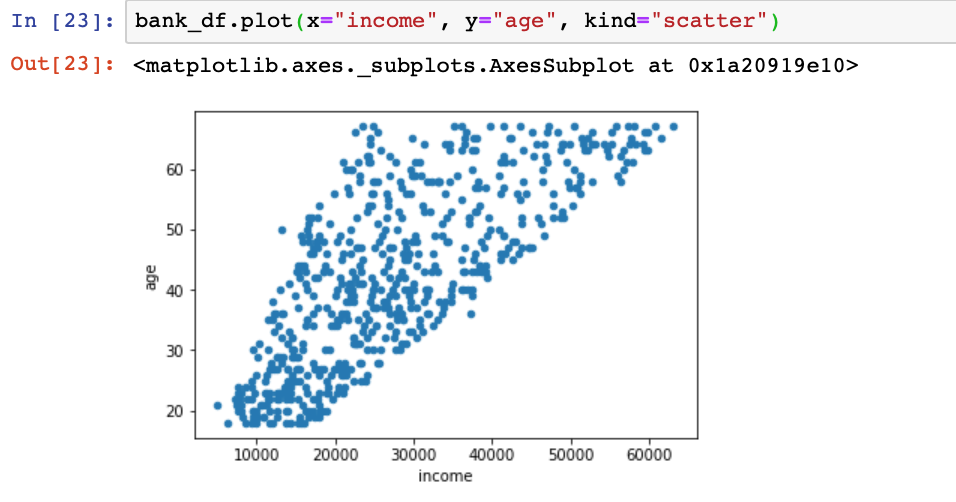


**Analysis:**

From above analysis, we can summarize that variables: age and income are highly correlated with each other; Which meaning people who have higher age also have income in this case;

Moreover, since we created many level of dummies, so they also highly correlated with each other.

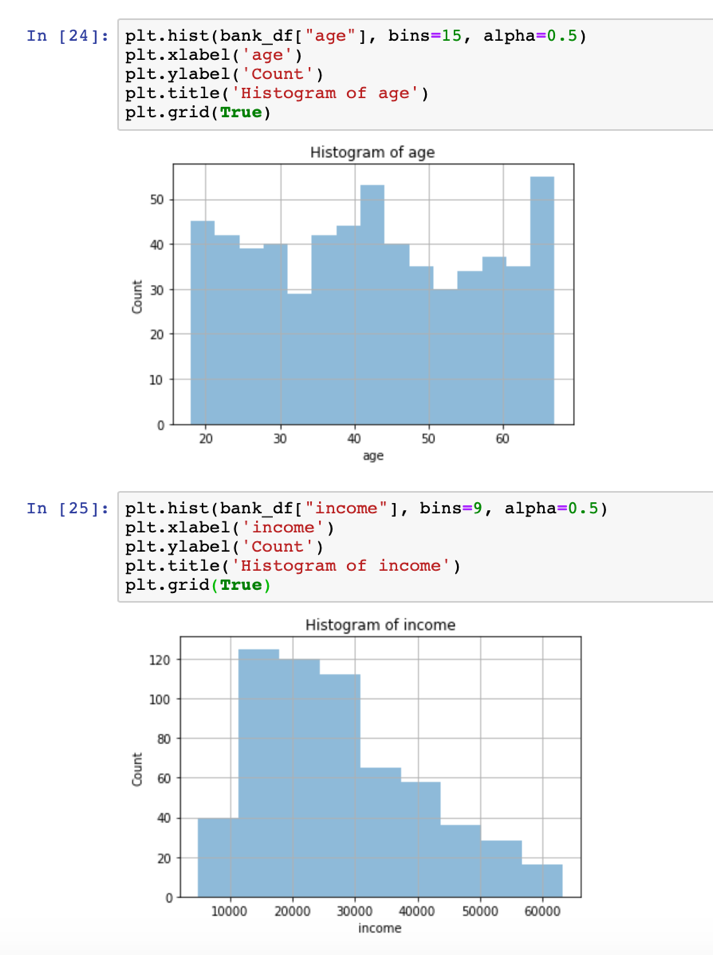
8.Using Matplotlib library and/or ploting capabilties of Pandas, create a scatter plot of the (non-normalized) Income attribute relative to Age. Be sure that your plot contains appropriate labels for the axes. Do these variables seem correlated?



**Analysis:**

From above graph, we can summarize that variables: age and income are highly correlated with each other. Again, it makes sense since higher age people do have higher income as well.

9.Create histograms for (non-normalized) Income (using 9 bins) and Age (using 15 bins).

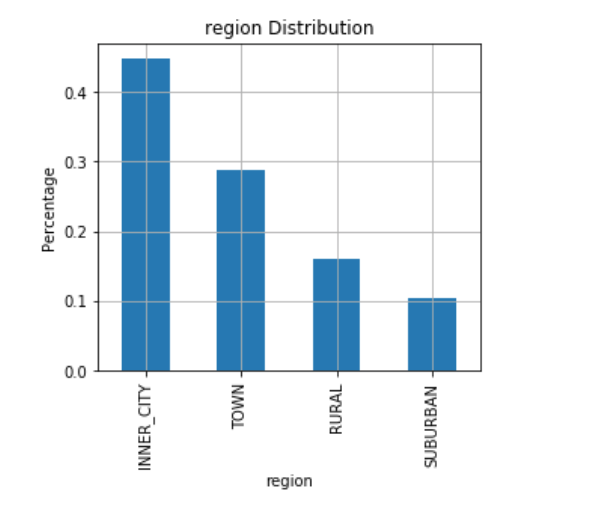


**Analysis:**

From above graph, we can summarize that variable age is not close to normal distribution, so that we need to transform/normalize it to make it more stationary;

For the variable income, it does close to normal distribution but more close to ‘right skew’; in this case, we can summarize that the mood > median > mean for income. Also it makes sense since in the real world, number of people who have lower income more than people who have higher income.

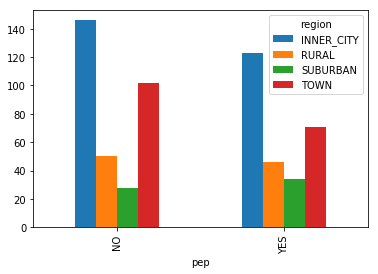
10.Using a bargraph, plot the distribution of the values of the region attribute.



**Analysis:**

From above graph, we can summarize that for variable ‘region’, over 40% of records is level with ‘INNER\_CITY’ ; and the following levels are: TOWN, RURAL, and SUBURBAN with roughly 19%, 16%, 11% from high to low, respectively;

11.Perform a cross-tabulation of the region attribute with the pep attribute. This requires the aggregation of the occurrences of each pep value (yes or no) separately for each value of the region attribute. Show the results as a 4 by 2 (region x pep) table with entries representing the counts. [Hint: you can either use Numpy or use aggregations fucntions in Pandas such as groupby() and  cross-tab().] Then, either using Matplotlib directly or the plot() function in Pandas create a bar chart graph to visualize of the relationships between these sets of variables. [Hint: This [example](https://d2l.depaul.edu/d2l/common/dialogs/quickLink/quickLink.d2l?ou=706272&type=content&rcode=depaul-3294141) of creating simple bar charts using Matplotlib may be useful.



**Analysis:**

From above graph, we can summarize that since majority of people living in the region of ‘INNER\_CITY’, so that whether buy pep (YES) or not buy pep (NO) have more than other regions such as rural, suburban, and town.

Moreover, number of people who living in ‘town’ also significantly higher that other regions; but both people who living in ‘inner\_city’ and ‘town’ are more likely do not buy ‘pep’;

For the level ‘suburban’, number of people who tend to buy pep are more than people who do not tend to buy pep, and this is only region where more people who willing to buy pep.