HW\_2\_Template

Lina Cao , Thursdays 1:15PM

SDGB 7844; Prof. Nagaraja; Fall 2017

## Question 1

What is a census tract? How many census tracts are in New York County? (Provide the citations for references used.)

The "Census Tract" is an area roughly equivalent to a neighborhood established by the Bureau of Census for analyzing populations.A census tract is smaller than a City but larger than a Block Group or Census Block. The Bureau of Census generally encompass a population between 2,500 to 8,000 people in a "Census Tract".

There are 288 census tracts in New York County in total. Bureau of Census describes census tracts as "relatively permanent", but they do change over time. I chose the data from 2010. Source:United States Census Bureau <https://www.census.gov/geo/maps-data/maps/2010ref/st36_tract.html> <https://parks.ny.gov/shpo/tax-credit-programs/documents/ListofEligibleCensusTracts.pdf> <https://www2.census.gov/geo/maps/dc10map/tract/st36_ny/c36061_new_york/DC10CT_C36061_CT2MS.txt>

## Question 2

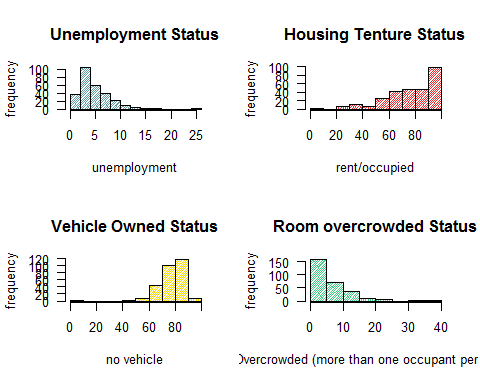
Describe one advantage and one disadvantage of computing estimates after combining 5-years of data.

ADVANTAGE Small deviation of difference years' census can be balanced due to the combination. DISADVANTAGE We cannot find the consecutive year change in a certain census tract, only five-year-change can be found.

## Question 3

Download the ACS data. Clean the data and merge the tables into one data frame, each row representing a census tract, each column representing one of the Townsend variables (keep the geography columns). For each variable, construct a histogram and compute the following summary statistics: mean, median, standard deviation, maximum, and minimum. Describe the shape of each histogram.

## GEO.id2 unemp rent car oc  
## 283 36061030700 6.1 57.6 63.8 95.5  
## 284 36061030900 8.3 92.1 65.4 88.9  
## 285 36061031100 - - - -  
## 286 36061031703 2.2 94.8 77.0 93.3  
## 287 36061031704 3.3 62.1 62.9 98.3  
## 288 36061031900 - - - -



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | mean | median | standard\_deviation | max | min |
| unemp | 4.926855 | 4.00 | 3.250082 | 25.0 | 0 |
| rent | 77.531429 | 80.75 | 19.173203 | 100.0 | 0 |
| car | 77.413571 | 78.60 | 9.783584 | 100.0 | 0 |
| oc | 5.883214 | 4.10 | 5.164891 | 36.4 | 0 |

From the histogarm above, 1. the unemployment status and room overcrowded status both show a right skewed curve; 2. the housing tenture status and vehicle owned status both illustrate a left skewed curve.

## Question 4

How many observations are missing for each variable? What percentage of census tracts do not have complete data? Is this a problem for our analysis? Justify your answer. (Note: do not delete tracts with missing data.)

## [1] 5

## [1] 8

## [1] 8

## [1] 8

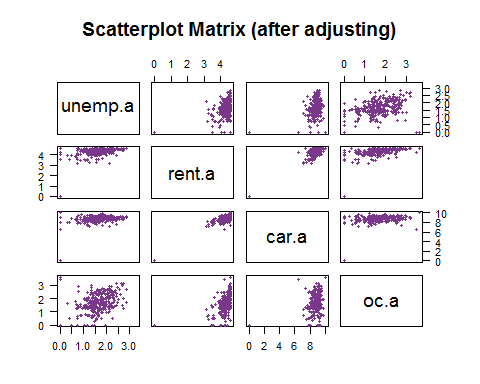
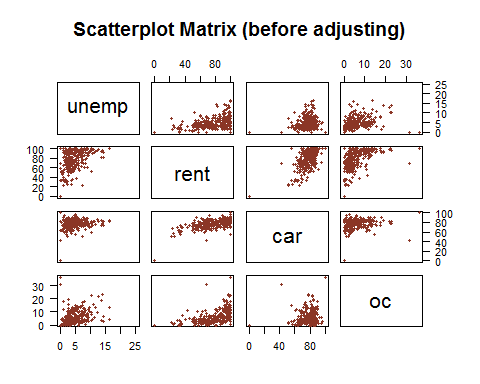
## [1] 0.02777778

## [1] 0.02777778

There are 5 missing data in unemployment, while 8 missing data in house tenture, car occupancy and room overcrowded variable. The groups with incomplete data takes about 2.78% of total data set. The percentage of the incomplete data is a small amount, I don't think it will effect our analysis.

## Question 5

Construct a scatterplot matrix of the four variables. Are they linearly related? Now, transform the variables as given in step (a), adding the transformed variables to your data frame. Make another scatter plot matrix with the transformed variables. Are they linearly related? Construct a correlation matrix of the transformed variables and describe your results.



## unemp.a rent.a car.a oc.a  
## unemp.a 1.0000000 0.4311138 0.2788261 0.3725944  
## rent.a 0.4311138 1.0000000 0.7842310 0.4730620  
## car.a 0.2788261 0.7842310 1.0000000 0.2141658  
## oc.a 0.3725944 0.4730620 0.2141658 1.0000000

The variables are not linearly related, because the dots in the scatterplot are randomly distributed from the first scatterpolt matrix.

FROM THE SCATTER PLOT After transform variables, from the second matrix, variables are more clustered, except for the relationship of room overcrowded and unemployment. Although the linear relation is not relatively obvious, the result after transforming variables can be reported to be more linear related to some extent.

FROM CORRELATION MATRIX Consist to the intuitive impression from the plot, the correlation between each variable is not high. Unemployment rate and housing rent rate, car owned rate and overcrowded rate, as well as room overcrowded rate with house rent and car owned rate are all weakly correlated, while car owned rate and house rented rate is the most correlated.

## Question 6

Compute the Townsend index value for each census tract. Add the index to your data frame. For how many census tracts are you able to compute the Townsend index? Why does this number not equal the total number of census tracts?

## [1] 280

## [1] 0.9722222

Totally 280 census tract can derive Townsend Index, which is 97.22% of total census tracts. As long as there is data missing in one census tracts, the "NA" will prevent the calculation of Townsend index. Therefore, the Townsend will also disappear in the census tracts of which any variable value is missing. Thus the number of total Townsend index will be less than the number of census tracts.

## Question 7

Identify which census tract is the most deprived and which is the least deprived (give the census tract number and deprivation index level). Based on your results, would you like to live in the least deprived census tract? Justify your answer.

## [1] "36061028500"

## [1] "36061021703"

## GEO.id2 unemp rent car oc Townsend  
## 224 36061021703 0.0 0.0 0.0 100.0 -28.58582

The geography ID of the most deprived census tract is 36061028500. The geography ID of the least deprived census tract is 36061021703  
The data in the least deprived tract shows zero in unemployment rate, housing tenture rate and car owned rate, while the overcrowded status is considerably high, which means the house will be too crowded to live in. I don't think I want to live in such a place in spite of the low unemployment rate.

## Question 8

The ACS data includes not only estimates but their margins of error which we ignored in our calculations. What are the implications?

Margin of error (MOE) is an estimate of the sampling error for each value, sum, or ratio reported for the sample population in ACS (American Community Survey). Usually it can be the result of the difference between the upper and lower limits of the sample estimate's confidence interval. MOE is used to describe the accuracy of estimates calculated by ACS. Although MOE increases for smaller areas, in this task, the numbers of each variable are only used to compare the relationship between each census tract, not for accurate value. Therefore, ignoration of MOE can simplify the calculation process. (source: <http://www.esri.com/news/arcuser/0708/demoarticle.html> )

## Question 9

Construct a map color-coded by the deprivation index value. Include a legend and plot title. Describe the patterns you see, especially in relation to what you know about neighborhoods in New York City. What does the large rectangle in the middle of the map represent?

## Warning: package 'rgdal' was built under R version 3.4.2

## Loading required package: sp

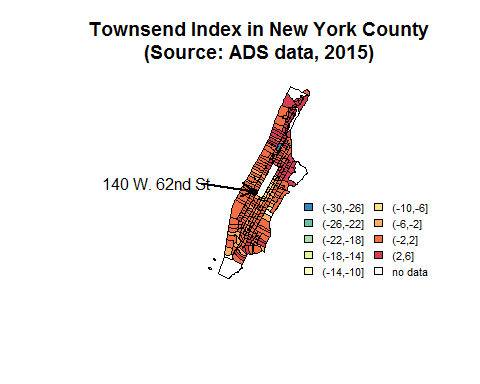
## Warning: package 'sp' was built under R version 3.4.2

## rgdal: version: 1.2-13, (SVN revision 686)  
## Geospatial Data Abstraction Library extensions to R successfully loaded  
## Loaded GDAL runtime: GDAL 2.2.0, released 2017/04/28  
## Path to GDAL shared files: C:/Program Files/R/R-3.4.1/library/rgdal/gdal  
## Loaded PROJ.4 runtime: Rel. 4.9.3, 15 August 2016, [PJ\_VERSION: 493]  
## Path to PROJ.4 shared files: C:/Program Files/R/R-3.4.1/library/rgdal/proj  
## Linking to sp version: 1.2-5

## OGR data source with driver: ESRI Shapefile   
## Source: "tl\_2015\_36\_tract", layer: "tl\_2015\_36\_tract"  
## with 4918 features  
## It has 12 fields  
## Integer64 fields read as strings: ALAND AWATER

## [1] -28.585820 5.044114

## [1] 218



In the map I derived, it seems most of the New York County is more deprived, except for an census tract in upper east part. The large rectangle in the middle of the map represents the Central Park, where no residents live, therefore no census data can be collected. Hence the heat map is white, with no data in it.

## Question 10

In which census tract is 140 W. 62nd St. (where we have class)? What is the deprivation level rank (where a rank of 1 is the most deprived)? Mark it on the map (use the computer, not by hand) and add it to your legend. (Provide the citations for references used.)

The census tract number of our classroom is 145. The source of the tract number is: <https://www2.census.gov/geo/maps/dc10map/tract/st36_ny/c36061_new_york/DC10CT_C36061_002.pdf>

The deprived level rank of 140W 62nd St is the 63rd among 280 tracts (which has full data in each variable).

## Question 11

New York County is an urban county, however New York state has roughly 22 counties classified as rural (e.g., Allegany, Essex, Otsego, Sullivan). Would it make sense to compute the Townsend index values for all census tracts within New York state combined? Why or why not?

I don't think it make sense. Since urban and rural county will have many differences from many aspect, for example GDP, average income per capita and even etc. All of them can separate urban and rural county's unemployment rate, car owned rate and so forth. Therefore if we mix them and calculate mean and standard deviation for the total New York state, some extreme value might change mean and derive a larger standard deviation. Both will affect when standardize the data before final calculation of Townsend Index. And generate more factors we should take into consideration. Hence I think we should calculate by different counties( urban and rural) separately.