GIS 3 Lab 2

Ryan (Yigong) Wang

This lab will practice skills learned in Chapters 3 and 4. In this lab, I will use a spatial dataset to explore the statistical and visualization features for attributes in the dataset available in R.

The dataset I will assemble today is from the ACS public use data, curated by IPUMS from the University of Minnesota. (See "References" section at the end of this document for more information.)

1. Creating the Spatial Dataset

The Non-spatial Part

I assemble this dataset to prepare for some exploratory data analysis for aspects of work and earnings across different areas in the Chicago Metropolitan Area. I first customize a non-spatial dataset from IPUMS, selecting relevant earnings and work variables such as earnings from salary and wages, travel time to work, job types etc. See codebook for more details. I also select geographic variables (PUMA - our area unit, Metropolitan Area, and State) to help joining to a shapefile to make a spatial dataset. This dataset contains data from both 2017 and 2018.

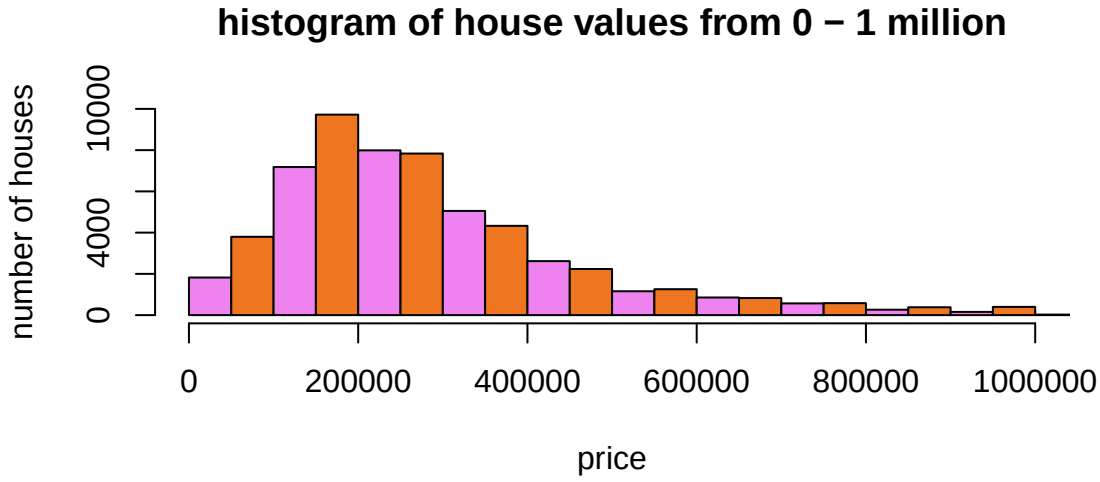
```
ddi <- read_ipums_ddi("/Users/ryan/Desktop/GIS 3/2017 2018 Work Geo/usa_00002.xml")
data <- read_ipums_micro(ddi)
```

a. Data Cleaning

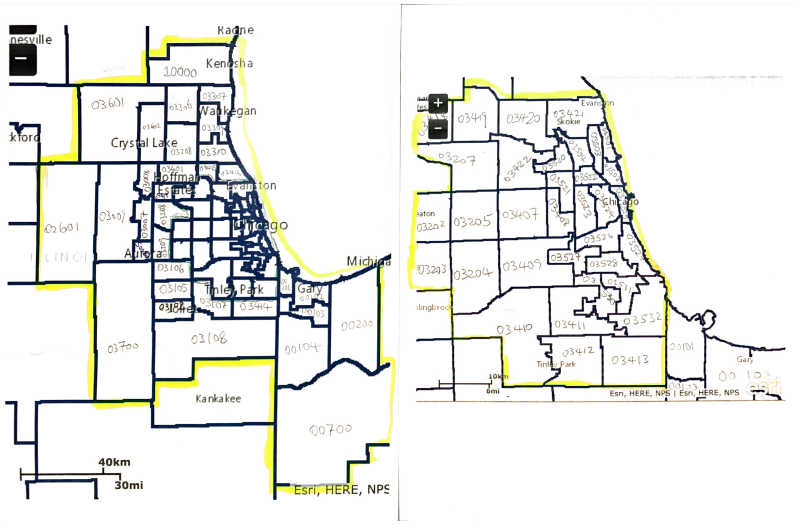
I clean the data by selecting out some irrelavant basic attributes, and apply a filter by selecting out people and households that have missing data for the attributes I am interested in (in this case earnings, household income, and house value), and limit the data to Chicago Metropolitan Area.

```
data$PUMA <- sprintf("%05d", data$PUMA)
data_cleaned <- data %>% select(-SAMPLE , -GQ ) %>%
  filter(INCWAGE != 0 & INCWAGE != 999999 & INCWAGE != 999998 & MET2013 == 16980 &
           VALUEH != 0 & VALUEH != 9999999)
```

I am quite interested to see what the housing price distribution is in Chicago without a detailed area look (non-spatial). Below is a histogram of the housing prices under $1m.



I ran into a problem of selecting Chicago Metropolitan Area out of this large shape file of 2500+ PUMAs. In order to find all the PUMA area codes in the Chicago Metropolitan Areas, I had to print out the shape file and manually match the codes one by one to record. I did not know at the time that there is a crosswalk file on the IPUMS website listing out the specific PUMA codes for each metropolitan area. Below is a photo of the hand drawn and hand matched PUMA codes atlas (I had fun doing it though despite wasting time):



With the crosswalk file, it became a lot easier to create a list of the PUMA codes that is in the Chicago Metropolitan Area. List created and imported into R. The output shows some of the codes and names of the PUMAs. (The codes are all five digits in original form, but for easier processing I converted them into numerical forms, hence some codes are not five digits due to first digits being zeroes.)

```
## [1] 2600 2700  101
```

```
## [1] West Central Texas COG (Outside Taylor County) PUMA
## [2] West Central Texas COG--Taylor County--Abilene City PUMA
## [3] San Sebastián, Aguada, Moca, Añasco & Rincón Municipios--Carr 2-Carr 111 PUMA
## 2168 Levels: Acadiana Regional Development District 2--Acadia & Vermilion Parishes PUMA ...
```

b. Aggregation - Weighted Average

In order to join this non-spatial part of the data with the spatial part, we need to aggregate, and in this case, the "aggregation" is done by a weighted average to reflect the average value for each of the attributes in each PUMA.

```
data_aggregated <- data_cleaned %>% group_by(PUMA) %>% summarize(VALUEH = weighted.mean(VALUEH, HHWT),
                                      INCWAGE = weighted.mean(INCWAGE, PERWT),
                                      HHINCOME = weighted.mean(HHINCOME, HHWT))
```

The Spatial Part

a. The Shape File

I download a shape file of all US PUMA areas from the IPUMS website. A reference map could also be found from the Census Bureau. https://www.census.gov/programs-surveys/geography/guidance/geo-areas/pumas.html

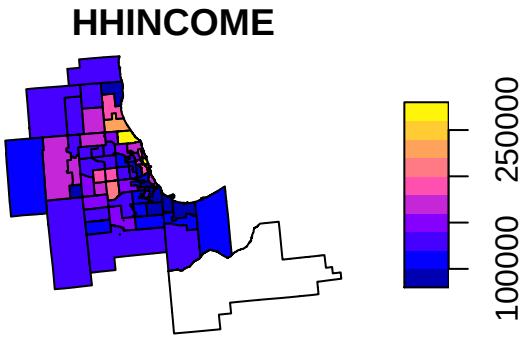
For the sake of time and convenience, I then created the shape file needed for this dataset by selecting the above PUMAs from the original whole US shape file in QGIS. Then I import this new shape file to be joined with the non-spatial part of the dataset.

b. Spatial Join

I `left_join` the spatial and the non-spatial parts, and plot it (using household income) to see the result.

```
## Joining, by = "PUMA"
```

```
## Warning: Column `PUMA` joining factor and character vector, coercing
## into character vector
```



I find that one of the PUMAs is showing nothing in the plot. We run a little bit of code to determine whether it is the criteria in this dataset excludes this PUMA area from the metro area, or whether it is that there is just no data in this PUMA.

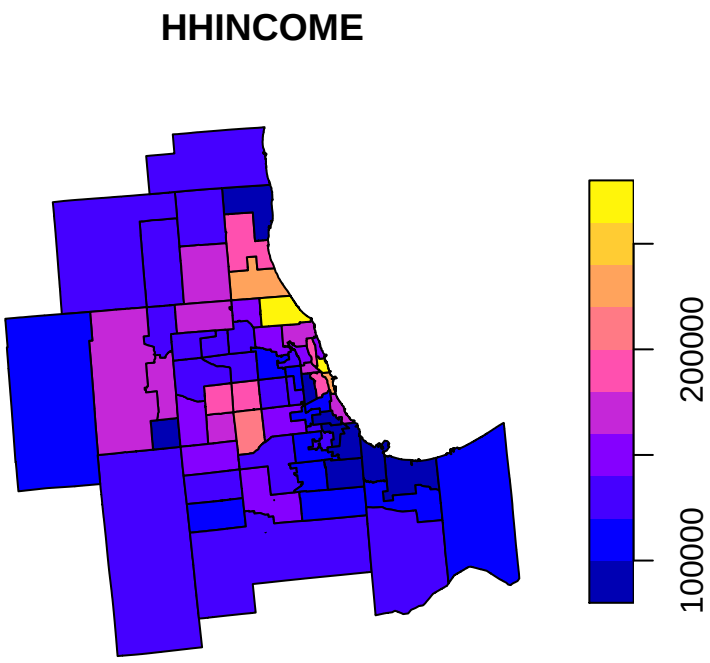
```
filter(data_cleaned, PUMA == 00700)
filter(data, PUMA == 00700 & STATEFIP == 18)
```

Both tests return a zero result! Turns out the surveyers of the Census Bureau just decided not to survey this area, hence it is not showing any data, even in the source file. We hence remove this area from the shape file to better visualize and represent the data.

Plotting results again, and this time it looks normal.

```
## Joining, by = "PUMA"
```

```
## Warning: Column `PUMA` joining factor and character vector, coercing
## into character vector
```



2. Data Analysis

INCWAGE

As metioned before, INCWAGE is the variable for salary/wage earnings for individual persons.

Descriptive Statistics

I first look at the median annual salary income in the Chicago metropolitan area.

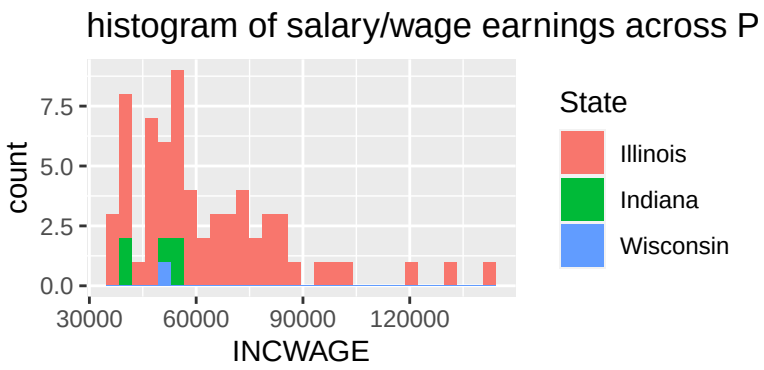
```
## [1] 55997.4
```

The median number is 55997.4 (dollars/year), which is not bad for a metropolis.

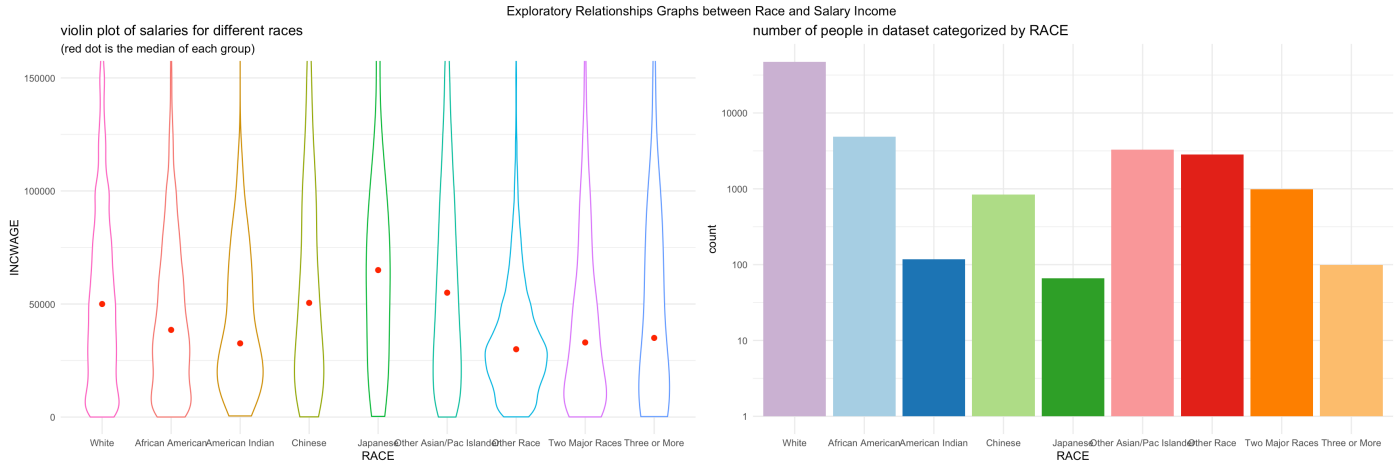
Distribution

I want to look at how income level differes in different communities (PUMAs), and this histogram gives some general information. We find that incomes are in a left-skewed distribution, with most communities averaging between 40K and 60K a year. There are a couple of outliers on the right, which we will visualize in the mapping part.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Before heading to the mapping part, I am also curious in how salary earnings differs across races. I use the non-spatial dataset to draw a violin plot to visualize how incomes are distributed for different race group. I also draw a bar-chart to see how many people are in each ethnicity group (the y-axis of this chart is using a log scale because white respondents are of overwhelming majority.)



As we can see, Chinese, Japanese and other Asian racies score very high in their salaries, and Whites are also in a leading position.

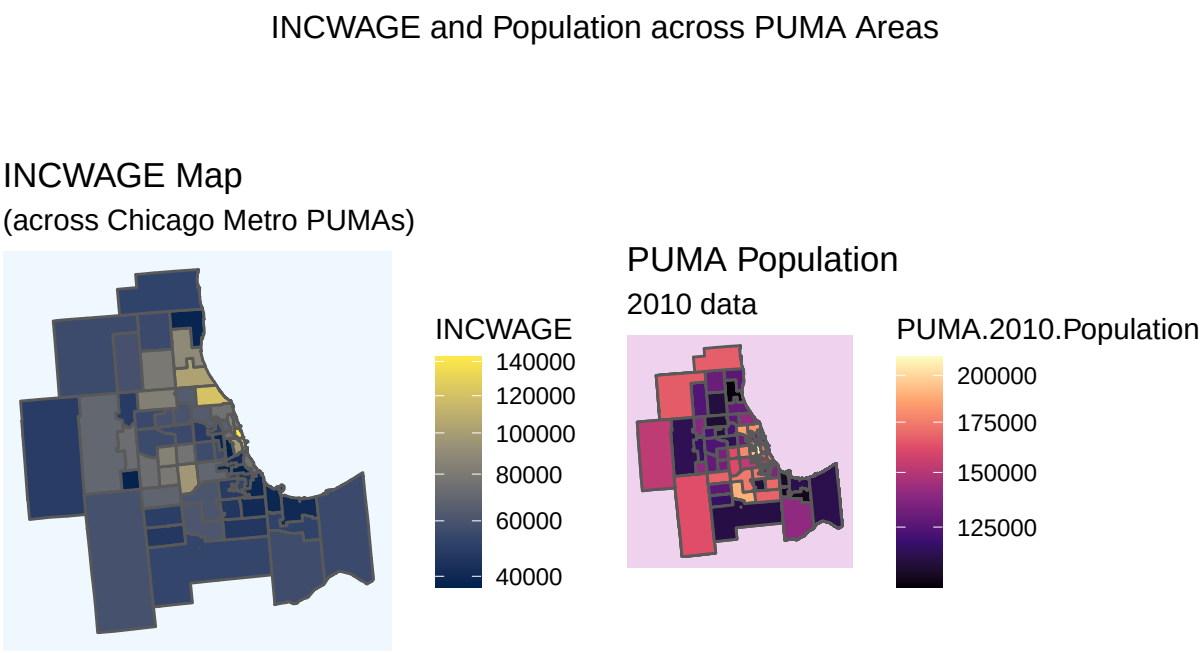
Mapping

We then plot a chloropleth map showing the distrubution of INCWAGE across all PUMA areas.

Population

The PUMA area is a specially designated area for the Census to provide as detailed geographical information as possible without compromising privacy of people who are surveyed. Each PUMA has at least 100K population. Out of curiosity, I want to see how many people are located in each PUMA in the Chicago Metropolitan Area.

```
## Joining, by = "PUMA"
```



The first plot shows average weighted salary income of each PUMA area. Turns out, near north side of Chicago and the northern suburbs (*Evanston*, *Glencoe* etc.) has very high salaryearnings, which exceeds $100K annulaly measured by the weighted average.

The second plot shows that urban and downtown areas generally have higher population for each PUMA, but some suburban areas also have relatively high population for the PUMA, and this might be due to the PUMA boundary following the guidlines of counties, and they try not to separate counties, rusulting in suburban areas having larger areas for individual PUMA, hence containing more pupolation.

For the following labs, I will try to explore more variables and potential relationships.

I am still figuring out how to get rid of some of the code outputs of ggplot graphics, and will try to make it look nicer for following labs assignments.

References

The MET2013 variable in the IPUMS dataset for ACS contains information for the metropolitan areas. https://usa.ipums.org/usa-action/variables/MET2013#description_section

IPUMS USA, University of Minnesota, www.ipums.org.

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0

-END OF LAB 2-