Stat 605 Final Project: First Draft

Project Group 4

November 2020

Authors

Ji, Qingchuan, qji5
Kou, Huitong, hkou2
Rui, Peibin, prui
Shi, Daiyi, dshi42
Su, Haohao, hsu69

1 Abstract

This project aim at making an exploratory data analysis on Total Person's Income (PINCP) based on the ACS PUMS 2015-2019 dataset.

In the first draft of our study, we focused on the 2019 data. After variable selection and data cleaning for missing values, we get a dataset containing 74 variables and 2452515 units. With this data, it is clear to show that there exists the huge difference of average personal income among states and heat maps are drawn to help reflect this difference intuitively. Furthermore, classification and regression trees are used to find out the characters of people with different levels of personal income, which implies the considerable income gap among all groups of people.

2 Data Introduction

Our project will be based on personal record data of the 5 one-year ACS PUMS datasets of the whole US, from 2015 to 2019. The Public Use Microdata Sample (PUMS) data contains a data sample of actual response to the American Community Survey (ACS) made by U.S. Census Bureau, covering approximately 1% of the population in the United States. Every one-year dataset can be divided into two parts of records, household-level and person-level. In the first draft, we will focus on PUMS 2019 data first.

This dataset covers the most fields of daylife of nearly 1% of American people, which can help find out the potential factors affecting personal income.

With these factors, we are able to prove the existence of income gap and track its source.

3 Data Pre-processing

As many other social surveys, missing data problem is unavoidable and occurs in this data set. Besides, not all about 300 numeric and categorical variables are closely related to our aim. Thus, data pre-processing is an essential prerequisite for further statistical analysis.

Totally, 73 variables, including 14 numeric variables and 59 categorical variables, are subjectively considered to be correlated to total person's income (PINCP) and selected together with PINCP. These 73 variables will work as explanatory variables and divided into 7 groups including **region, race, education, employment, insurance and other personal info** by their meaning.

For convenience, our study objects will be limited to adults (no less than 18 years old) living in 50 states and DC of the United States. This can help cut down the missing rate, since many missing values occur due to low age. Besides, some allocations have already been made by staff of U.S. Census Bureau and we will accept these allocations considering the reliability. As for the rest missing values, they all occur in categorical variables and most of them contain some information but not pure vacancy. These "NA" would be better to be assigned as an extra "class" of the corresponding factor.

Now, only the variable OC (own child) still suffers from missing data problem with about 70000 missing values. We just simply remove these units and after the steps above, we have a dataset containing 74 variables and 2452515 units, without any missing value.

4 Regional Difference in Personal Income

Based on near 2.5 million effective observations in ACS PUMS 2019 data, the first we hope to present is the regional difference of average personal income (PINCP). From the aspect of mean PINCP, the top 5 states in 2019 are shown as follows, with their corresponding mean and median values of PINCP.

State	Personal Income Mean	Personal Income Median
District of Columbia	85658.58	62000
Connecticut	66652.14	40000
New Jersey	63871.01	40000
Massachusetts	63495.54	41000
Maryland	62168.96	42000

Table 1: Top 5 States in mean personal income

And also the similar table for the last 5 states.

State	Personal Income Mean	Personal Income Median
Mississippi	35749.88	23000
West Virginia	36410.11	25000
Arkansas	36816.44	24700
Alabama	38960.41	25000
Oklahoma	39057.17	25700

Table 2: Last 5 States in mean personal income

To make the results more intuitive, heat maps are drawn. Here is the heat map based on mean PINCP values of each states.

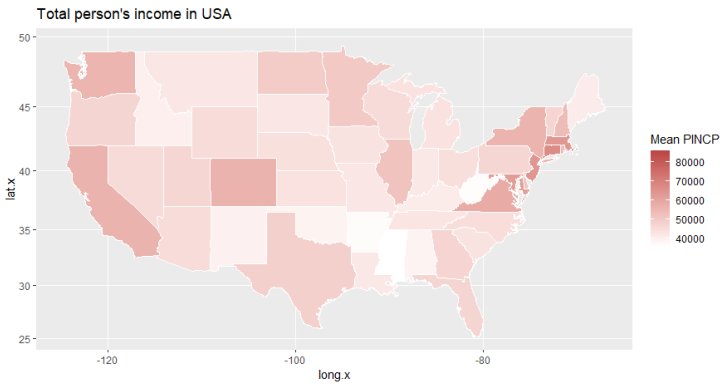


Figure 1: Heat map on mean PINCP values

Honestly speaking, it may still be difficult to search the difference between states. To make comparison more clear, another map is created based on the rank interval.

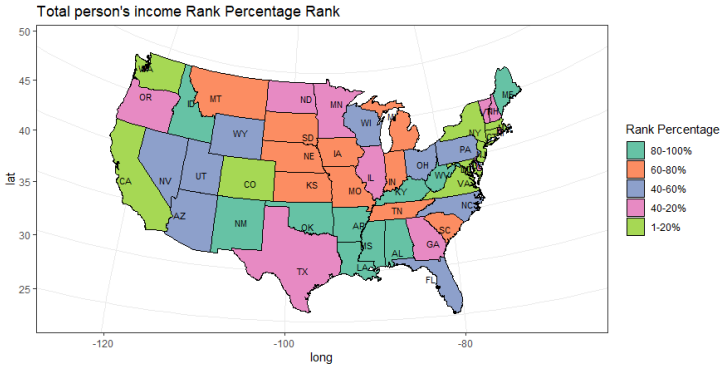


Figure 2: Heat map on rank intervals of mean PINCP

Based on the tables and heat maps, we can straightly know the mean Personal Income in different states. It is clear that there exist huge regional difference in personal income among states. In general, the East and West Coast are regions with high average personal income, while personal income are low in East and West South Central.

5 Classification

In this section, we hope to draw the characters of people with different levels of personal income. Since we already have divided the variables into several groups by their meaning, we are going to select variables from each group and find their relationship with person's income.

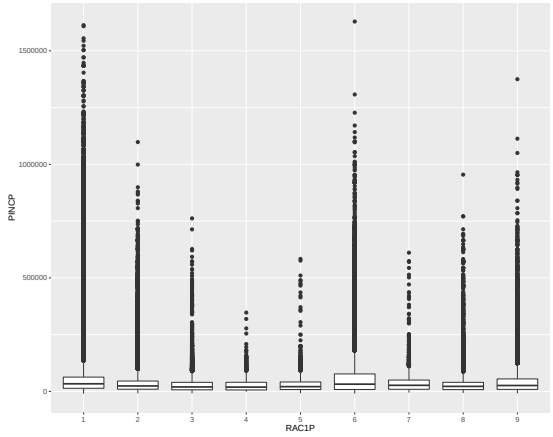
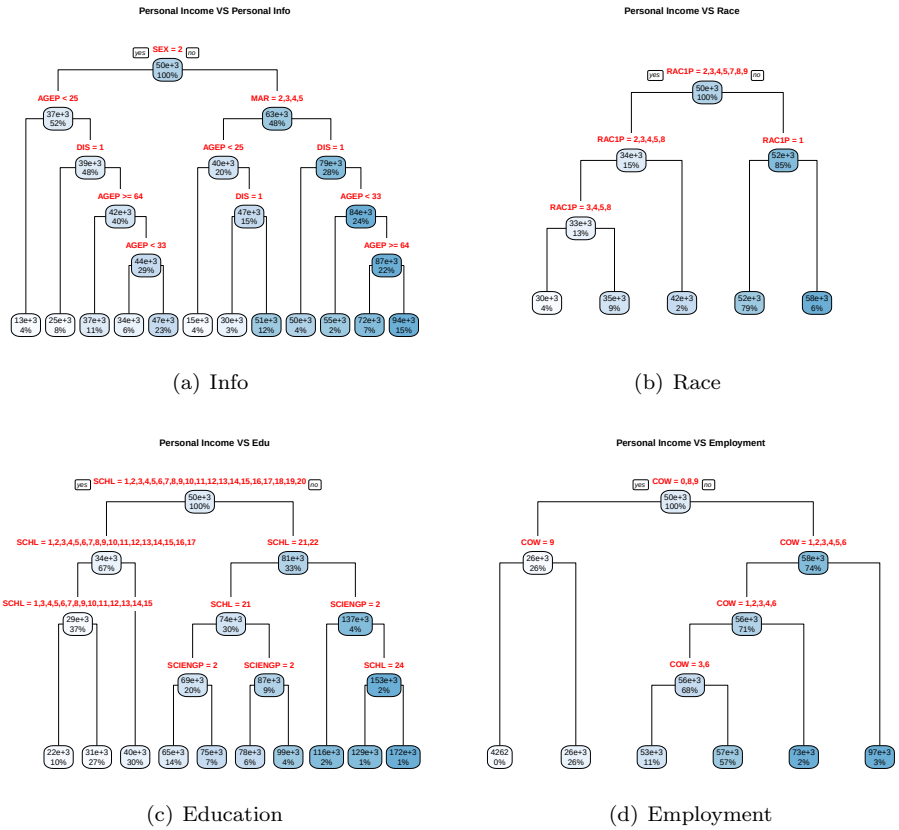


Figure 3: An example of box plot

We tried to use box plot to find the relationship between PINCP and a categorical factor like what is shown in 3. However, this method does have plenty of drawbacks. For example, the existence of many extreme points in the plot may decrease the intuition of distribution by a lot. And the box plot would ignore the potential interaction of factors to PINCP since it can process only one factor once in a time. Based on this situation, we are going to use classification and regression trees for this analysis.



(a) Info (b) Race

(c) Education (d) Employment

Figure 4: Trees for PINCP

In the group "**personal info**", some variables correlated to PINCP are considered as factors to calssify and explain PINCP as follows.

$AGEP$: age; numeric

MAR : marital status; categorical with 5 classes

SEX : sex; categorical with 2 classes

DIS : disability record; categorical with 2 classes

According to the tree (a) in 4, **SEX** works as the first split factor for PINCP. The average personal income of males in 2019 reaches about \$63000, above the national average personal income \$50000, while the average personal income of females is \$37000, under the overall average. As for people from different age groups, those who age between 33 and 64 (i.e. middle-aged) have highest average personal income, while people aged less than 25 have the lowest. Besides, people with disability usually have lower personal income.

For the average personal income for different **race** groups, Asians rank the first, followed by white people. And American Indians, Alaska natives as well as several other race people have lowest personal income, which is about a half of the highest. Therefore, from the perspective of personal income, the gap among races is still considerable.

Then, take a glance at the relationship between personal income and education. Basically, the annual personal income has positive correlation with education level, which is consistent with our common sense in education and indicates that education does help improve personal income. And for people having at least one bachelor's degree or above, those whose degree is in the field of science or engineering would enjoy higher personal income.

Finally, employment. In this part, we will classify PINCP only by **COW** (i.e. class of work). People self-employed in own incorporated business, professional practice or farm have the most personal income and federal government employees take the second slot. What's more, we can also find some interesting difference among groups in similar job categories like among government employees, people work for federal have highest income, followed by states government employees; and incorporated businesses of farms can bring more income to their owner than businesses not incorporated.

Now, we can take all aspects above into consideration and draw the characters of people with different levels of personal income based on their personal situation, race, education and employment. For example, wealthiest people are married males having at least one bachelor's degree or above and self employed by his own business, professional practice or farm. Their average annual person's income reaches $142000, more than 7 times the lowest.

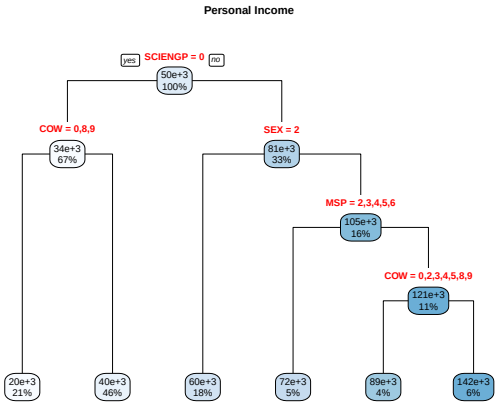


Figure 5: General tree for PINCP

6 Conclusion and Future Plan

From the exploratory analysis on ACS PUMS 2019 data, it is clear to found personal income differences at the scale of region, sex, education background, employment, race and so on. Besides, several factors are found and considered to have impact on person's income, with which the characters of different income levels are drawn by classification.

In following part of project, we will extend our study to the whole 2015-2019 dataset and dig the difference in personal income at the scale of time.