Labor Market Analysis EDA

Rylee Iacolucci, Sophie McDevitt, Caitlin Howansky

January 20, 2019

#### Overview

For our labor marketing analysis we’ve chosen to analyze Texas. Below are the research questions for this project:

* In Texas, how does earnings vary by household language?
* Does the household language earnings vary by household family type?

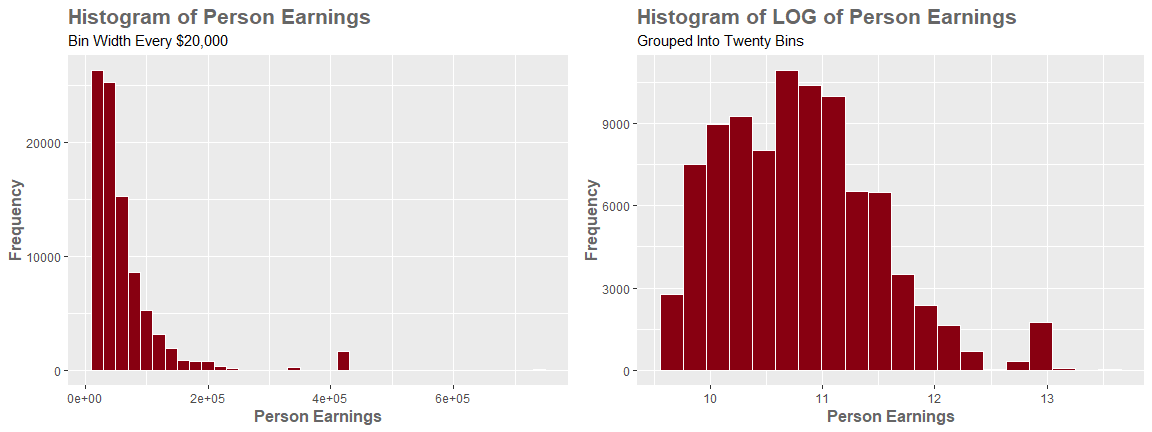
To answer these questions we’ll be analyzing a people and household data file. The people data file contains over 200,000 observations and 280 columns. To answer our research questions we were able to reduce the data file to ten columns. In the rest of this paper we’ll examine this data and describe the filtering out and why it’s necessary for our economic analysis.

## RECORD\_TYPE HOUSING\_UNIT PEOPLE\_NUMBER AREA\_CODE   
## Length:261896 Min. : 1 Min. : 1.000 Min. : 100   
## Class :character 1st Qu.: 381887 1st Qu.: 1.000 1st Qu.:2310   
## Mode :character Median : 763513 Median : 2.000 Median :3801   
## Mean : 762449 Mean : 2.167 Mean :3700   
## 3rd Qu.:1143472 3rd Qu.: 3.000 3rd Qu.:5100   
## Max. :1521433 Max. :20.000 Max. :6900   
##   
## AGEP CITIZEN\_STATUS WORKER\_CLASS SEX   
## Min. : 0.00 Min. :1.000 Min. :1.00 Min. :1.000   
## 1st Qu.:19.00 1st Qu.:1.000 1st Qu.:1.00 1st Qu.:1.000   
## Median :39.00 Median :1.000 Median :1.00 Median :2.000   
## Mean :39.06 Mean :1.556 Mean :2.11 Mean :1.508   
## 3rd Qu.:58.00 3rd Qu.:1.000 3rd Qu.:3.00 3rd Qu.:2.000   
## Max. :92.00 Max. :5.000 Max. :9.00 Max. :2.000   
## NA's :111447   
## EMPLOYMENT\_STATUS PEOPLE\_EARNINGS   
## Min. :1.00 Min. : -5500   
## 1st Qu.:1.00 1st Qu.: 0   
## Median :1.00 Median : 12000   
## Mean :3.06 Mean : 31284   
## 3rd Qu.:6.00 3rd Qu.: 42000   
## Max. :6.00 Max. :749000   
## NA's :54041 NA's :54041

## # A tibble: 10 x 10  
## RECORD\_TYPE HOUSING\_UNIT PEOPLE\_NUMBER AREA\_CODE AGEP CITIZEN\_STATUS  
## <chr> <int> <int> <int> <int> <int>  
## 1 P 1 1 100 33 5  
## 2 P 1 2 100 35 5  
## 3 P 1 3 100 12 1  
## 4 P 1 4 100 7 1  
## 5 P 1 5 100 0 1  
## 6 P 14 1 3200 48 1  
## 7 P 14 2 3200 41 1  
## 8 P 14 3 3200 13 1  
## 9 P 20 1 4624 77 1  
## 10 P 46 1 4606 33 5  
## # ... with 4 more variables: WORKER\_CLASS <dbl>, SEX <int>,  
## # EMPLOYMENT\_STATUS <dbl>, PEOPLE\_EARNINGS <dbl>

#### Summary Statistics

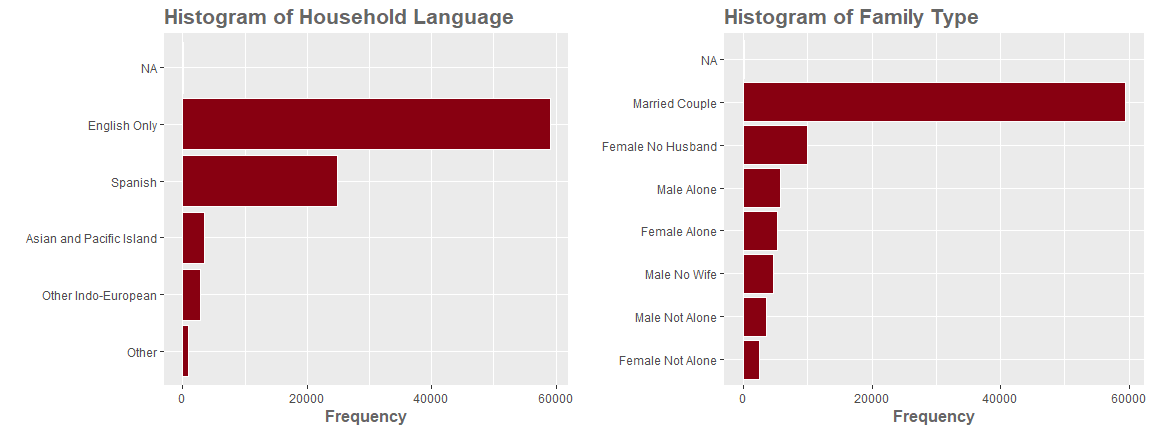
* AGEP
  + From the summary data you’re able to see that age of people surveyed has a very wide range, while no nulls appear to be in the dataset.
  + In our analysis we want individuals who are legally able to be fully employed. In Texas the minimum age for full employment is 14, however, by law that individual is still considered a child.
    - Peoples under 18 years old have regulations limiting the total number of hours they can work in a single week.
    - In this analysis we will want to analyze peoples who are not bound by hour age regulations by the state, so only peoples greater than 18 years old will be analyzed.
* Employment Status
  + In this category 1 stands for “Civilian Employed, at Work”, while other codes suggest the people is currently not working full time. In this analysis we may want to filter for only peoples who are currently employed and working so the results of the analysis are not skewed by invididuals not working and/or not looking for work.
* People Earnings
  + Right away we noticed there are peoples with zero or less earnings.
  + According to the Bureau of Labor Statistics nearly 6.2 million workers in Texas were paid the federal minimum wage of $7.25 per hour.
    - This would make a peoples annual earnings of $15,080, without adjusting for federal and local taxes, if that people worked 40 hours a week for 52 weeks a year.
  + Looking at the first 10 rows of data you see a lot of rounded earnings values like 60,000.
  + In this analysis we may have to keep in mind that when individuals were filling out their income values that they may have rounded up or down.



#### People Earnings

*Note: Inner joining with the household dataset on HOUSING\_UNIT & AREA\_CODE removed 151 observations due to NAs.*

Looking at the left histogram you can see the people earnings data is right skewed. Some people answered their earnings to exceed 500,000, which may be considered an outlier when you compare against the mean of 65,423. In this analysis we wouldn’t exclude these values since they are not top coded and may provide useful insight in explaining our research questions. Instead we can chose to analyze the log of people earnings. As you can see applying the log function makes the data more normally distributed which may help with analyzing the variance in people earnings.



#### Family Type

Both the histgrams above show the answer frequency for the household language and family type of the people surveyed. You can see english and married couple households are the majority of those surveyed. The impact of household and family types will be examined in our technical research paper.