Assessment the Supplemental Security Income Eligibility for Individuals in the Midwest

By: Hanadi Fahrat, Timothy Lee, Matthew Brown, Diana Paicu, Aya Salka

# Table of Contents

Introduction……………………………………………………………………………………… 1

Methods………………………………………………………………………………………….. 2

Results…………………………………………………………………………………………… 3

Discussion……………………………………………………………………………………….. 4

Conclusion………………………………………………………………………………………. 5

References……………………………………………………………………………………….. 6

# Introduction

Supplemental Security Income or SSI, is a federal program, that provides government assistance to individuals who meet the eligibility requirements. This is done using tax money with the goal of providing simple cash to support those who are eligible. To qualify, individuals must be citizens of the United States of America, either 65 and older, blind, or disabled and have limited resources or low income. This program also provides many benefits such as medical and governmental assistance such as in the form of healthcare access and food stamps. According to the Social Security Administration, in December 2017, alone, 8.2 million people received a minimum monthly income of $542(Social Security Administration). While this grants individuals a better quality of life, the acceptance of individuals into the program itself can seem arbitrary as there are many factors that regulate eligibility. For example, it can be quite difficult to get approved having certain kinds of medical limitations over others and other applicants have difficulty providing proof of their disability. Therefore, the goal of this research paper is to identify certain characteristics or patterns in terms of eligibility for SSI in the Midwest region of the United States of America using data from the National Health Interview Survey.

The data used throughout this paper was the accumulation of information from surveys conducted on people ages 17 and above by the United States Department of Health and Human Services and from January 2004 to May 2005. It looks at the amount and distribution of illnesses and if any of the illnesses caused any further complications, such as disabilities or chronic impairments. In addition, it also looks at how much health services were used by individuals, as well as how much and the type of financial aids these individuals used. The data is made up of 7 subsurveys which ask different types of questions about families and their health status.

# Methods

Due to the relative size and abundance of the overall data set, subserveys 1- 4 were used. A description of each subsurvey is provided below:

DS01: This set contains information on the type of living quarters, number of families in the household responding and not responding, and the month and year of the interview for each sampling unit(National Health Interview Survey, 2004).

DS02: This family level set contains information on the sex, age, race, marital status, Hispanic origin, education, veteran status, family income, family size, major activities, health status, activity limits, and employment status, along with industry and occupation(National Health Interview Survey, 2004).

DS03: This set is referred to as the person level file and contains information on all family members with respect to health status, limitation of daily activities, cognitive impairment, and health conditions. The set also houses data on years at current residence, region variables, height, weight, bed days, doctor visits, hospital stays, and health care access and utilization(National Health Interview Survey, 2004).

DS04: This set holds information on the respiratory conditions, use of nasal spray, renal conditions, AIDS, joint symptoms, health status, limitation of daily activities, and behaviors such as smoking, alcohol consumption, and physical activity(National Health Interview Survey, 2004).

Data Cleaning

Using the above data sets, four variables were created to organize and further evaluate the intrinsic properties of the overall data and formulate a conclusion concerning characteristics involved in the categorization of SSI in the Midwest region. Data\_Household (Data set 1) , Data\_Family (Data set 2), Data\_Person (Data set 3), and Data\_Sample\_Adult (Data set 4).

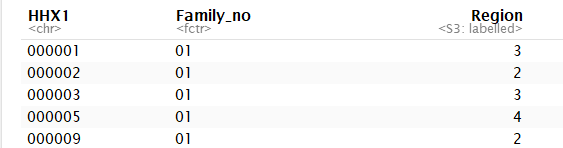
The Data\_Household contains 42089 observations of 17 variables. For the purposes of structuring and identification the only variables used from this dataset included HHX for the identification of households, and Region, the categorization of the United States’ 4 regions. Although the data obtained covered the United States of America, for the purposes of this project, only the midwestern region was taken into account. This included 12 states: Illinois, Iowa, Indiana, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, and Wisconsin. These variables were renamed and included in a data frame.

Data\_Family contains 37466 observations of 100 variables. However, only 2 of the 100 variables will be used from this data set. These include HHX, and Family\_no which noted the serial number of each family. This employs a certain knowledge concerning the eligibilities of SSI that can be further analyzed through the later manipulation of the data. These 2 variables were renamed and placed in a new data frame.

The third set, Data\_Person, houses 94460 observations of 537 variables. Due to its relative size and importance to the overall goal of the outcome, 20 variables were picked from this data set and a new variable was created. The 20 variables chosen can be categorized based on the following properties : Salary, disability, employment status, and citizenship. These variables are especially important due to the fact that the requirement of obtaining SSI relies on low relative income as well as being a U.S citizen. Variables concerning previous application and actual attainment of SSI or Social Security were also included as it was necessary and important to be able to determine characteristics of eligibility of SSI based on factors such as the number of times one has applied to SSI or Social Security. All of the selected variables were renamed, leveled, and included in a separate data frame.

Lastly, Data\_Sample\_Adult had around 31326 observations and 510 variables. A total of 37 variables were selected based on the fact that they included types of diseases, medical conditions, or limitations. These factors are especially important as they denote the portion of the SSI eligibility concerning being disabled or blind. They can provide further use in their implementation by assessing the relative importance of each type of disease/limitation and their ideal frequency among those with SSI or those without. These variables were also renamed, leveled accordingly, and placed in a new data frame.

The Statistical methods that were implemented in R based on the specified data frames that were created and mentioned above included inner join. This was performed between the Data\_Household and Data\_Family data frames that were created previously connecting HHX between the two sets. This resulted in a Data Join of 37,466 observations of 4 variables. A visualization of the first five outputs can be seen in the table below:



This was then implemented again through the result of this inner join being used against the data frame from Data\_Person. On the other hand this resulted in 97,461 observations. This was not what we were expecting as the inner join should have given us the lowest amount of observations and having over 97,000 signaled alarm. Therefore, we instilled an inner join between Data\_Sample\_Adult\_1 and Data\_Person\_1 by HHX, Person\_no, and Family\_no as the identifier. The outcome of this join was much more viable as it gave us over 31,326 observations to work with.

Inner join was used once again against the previously modified Data\_join from above against the result from the first data join, Data\_join\_1 by HHX and Family\_no. This resulted in 31,326 observations, 59 variables and was labeled as Data\_Join\_4.

Lastly, the inner join between Data\_Join\_4 was joined with Data\_join\_1 by HHX and Family\_no. This resulted in a comprehensive set of observations that were then filtered based on region, sample adult and mutated for the type of disease we wanted to include. This outcome of this filtered join, after some cleaning up, was used in creating a function which was necessary to calculate the percentages of each of the diseases. This is an extremely important implementation as it would aid us in computing the top diseases in terms of their frequencies in the sets based on the previous manipulation, of course.

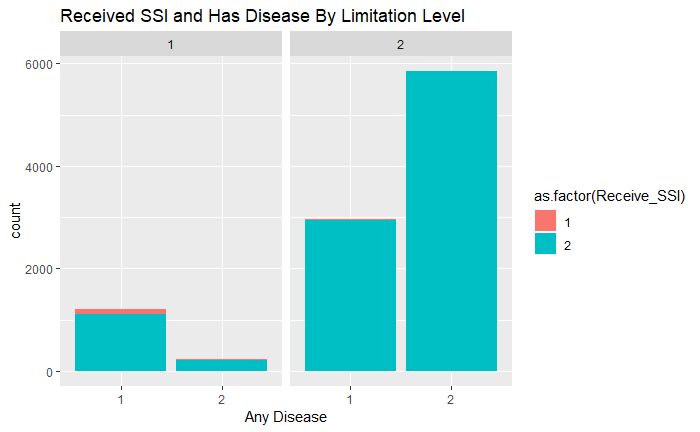
# 

# 

# 

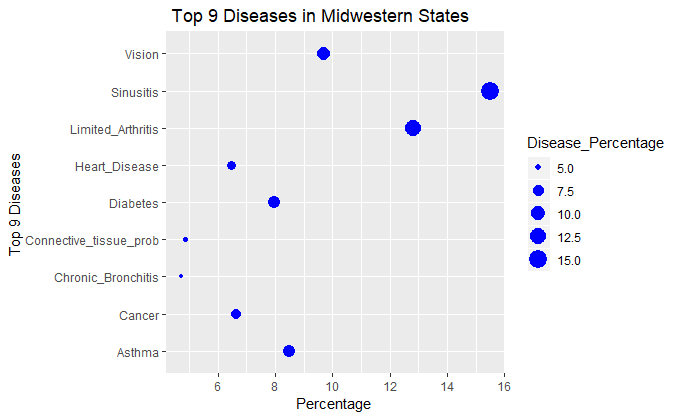
# Results

The manipulation of the data sets were visualized and transformed into the associated figures below.

*Fig 1.*

Explanation: This graph is showing the amount of people who have a disease, received SSI, and are limited in any way. We found that there are a very small amount of people who have a disease and state that they have a limitation received SSI. The majority of people do not have a disease, are not receiving SSI, and are not limited. For any variable, a (1) means “Yes” and a (2) means “No”. The facet labels are those who are limited in any way.

*Fig. 2*



Explanation: In this graph we found the 9 diseases that the most people had in the midwest in order to understand what some of the largest problems were in the region that demand SSI. This helped simplify our data since we had 36 variables for types of disabilities, and since the percentages were so small, it made sense to look only at the top contributors. We found the most common ailment to be sinusitis (15.49%), followed by arthritis (12.83%), vision impairment (9.67%), asthma (8.5%), diabetes (7.96%), cancer (6.66%), heart disease (6.46%), connective tissue disorders (4.86%) and lastly, chronic bronchitis (4.74%).

*Fig. 3*

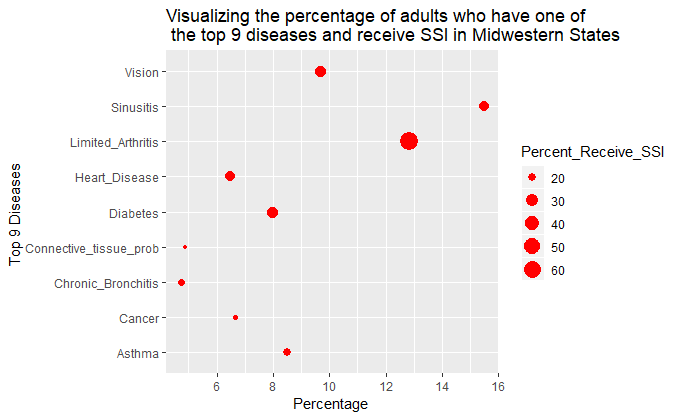


Figure 3 visualize the the top 9 diseases of people live in midwestern states. The size of the points reveals the weight of the percentages of people receiving SSI.

Explanation: We used this graph to explain the number of adults who have one of the top 9 diseases and receive SSI. The largest % receiving SSI is for limited arthritis, which is also the second most common disability. This graph generally follows the trend that the most common disability also is the most likely to receive benefits. One exception is sinusitis, which is the most common condition but does not have the highest reception rate (even though it is close).

*Fig. 4*

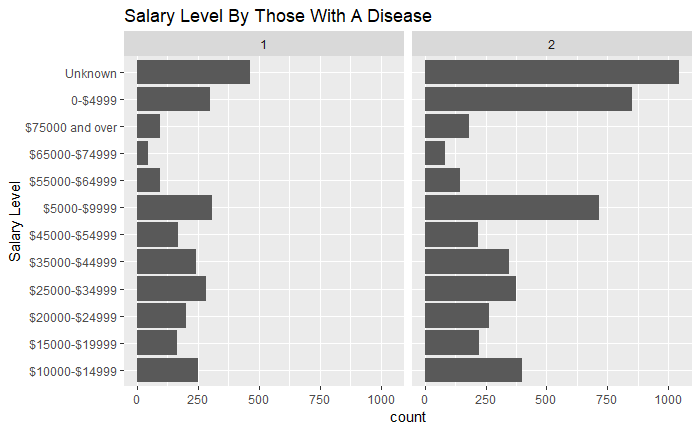


Figure 4 is a visualization of the Salary displayed at the x-axis and facet by whether the person has a disease (1) or not (2). The graph reveals that the health condition of a person has an impact on his salary. This graph is faceted by 1 (people who do not have a disability), and 2 (those who have a disability). Note that people who have a disease are paid less than those who don’t have, especially for people in the salary range of $5000 -$9999 and below which are unlivable wages on their own.

*Fig. 5*

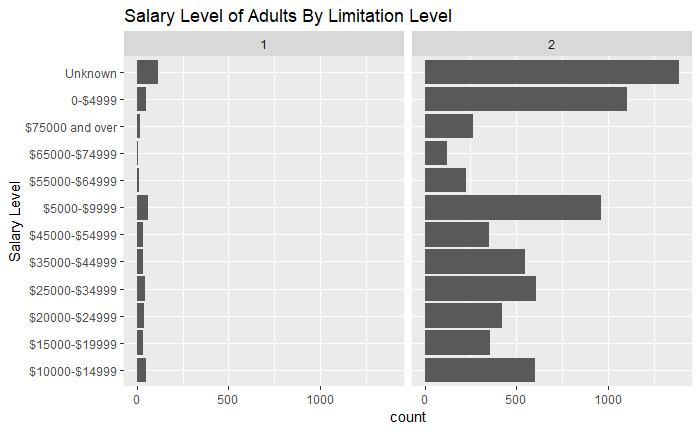


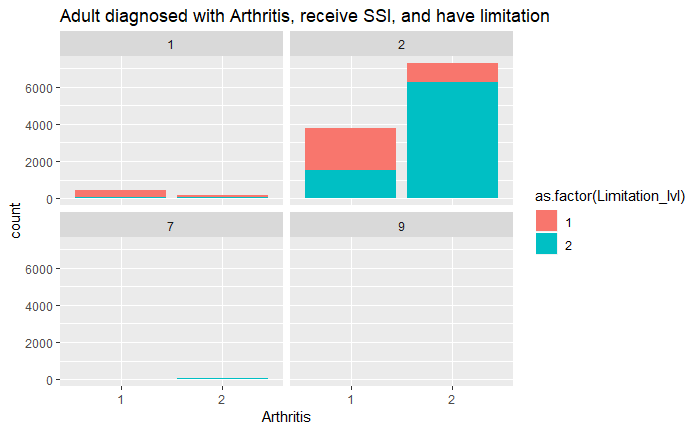
Figure 5 is a visualization of Salary of adults as the x-axis facet by limitation. Explanation: We found that the salary level of people stating that they have a limitation is a lot lower than those without a limitation. There are still a lot of people that did not know their salary and do not have a limitation. The result of this graph matches with the result of figure 4. More proves that the health condition of a person plays an important role on the ones earning. As stated in previous graphs, a (1) means “yes” and a (2) means “no”.

*Fig. 6*



In figure 6, we limited the data for people over 18, the x-axis is the variable Employed, which indicates whether the person is employed or not, the fill is the gender of the person, here is for male and 2 for female. Explanation: The distribution of the working adult is found to be about an equal amount of both genders in the workforce. Though there are more females that are not working, at the moment. In the graph, a (7) means “Refused”, (8) means “Not Associated”, and (9) means “Do not Know”.

*Fig. 7*



In figure 7, we are plotting Arthritis as x-axis (1 being they do have arthritis, 2 being no), whether they consider this or have other conditions considered a limitation as fill proportion (1 as yes, 2 as no) and receive SSI as facet. For the limitation fill, 1 represents Yes and 2 is No. 1, 2, 7, and 9 represent whether a person received SSI respectively for the labels of the facet.

Explanation: This graph is explaining the population that is accepted into SSI and diagnosed with Arthritis that consider it a limitation or live with another condition. However, there are more people who are diagnosed with arthritis and have limitation but do not receive SSI. We conclude that there are other factors that could play a major role in deciding people eligibility other than limitation, we could refer to the other requirements for SSI such as age and low income for these.

# Limitations

We were unable to provide the true estimates for the entire US population based off of the data collected from the NHSI. Due to the fact that there was a significant amount of missing values or respondents refusing to provide any answers, we had to eliminate these values from our analysis. By doing this, our actual data was limited and therefore extremely difficult to manipulate according to actual information/conclusions we wanted to reach concerning the breadth of SSI among the Midwest population.

Another limitation we encountered concerned the scope of the data itself and the method in which it was collected. The data was mainly categorical and as such it was much more difficult than anticipated to manipulate it. For example, the data had combined variables making new rows causing the data to be manipulated already and very difficult to work with. There were variables that were placing more than one row of observations as it was including the scope of two individuals in a household/family. Therefore, when we had to call functions such as inner join, the data was hard to gain some sort of viable outcome and we were left with a huge amount of data still. In the end, we had to use other methods to analyze our data instead of modeling our data linearly.

One limitation specific to the dataframes was that Data\_Households was larger than Data\_Family, therefore, presenting a contradiction within the overall method of data collection. This made the method of data cleaning difficult and limited as it was presenting troubles over the number of rows and columns to aggregate in inner join.

# Conclusion

Although the ongoing battle over correctly cleaning, organizing, and manipulating the data presented an outstanding issue, we were able to successfully visualize what certain aspects such as citizenship, education level, and salary had over the data set as a whole. Overall, we were able to ascertain the top ten diseases across midwest region according to family number (serial number) as well as basic eligibilities for SSI. Based on these findings, we were able to look at the frequencies of each of the diseases and pinpoint the most frequent one as being the top disease that people had who were eligible for SSI. These findings present a relatively important idea concerning the SSI eligibility process and what sort of diseases, at least the ones presented in this set, are more or less eligible for SSI.

In terms of disabilities, you are most likely to be eligible for SSI in the midwest if you have arthritis, or vision impairment/blindness. This just about followed the trend that the disability with the highest percent were also the most likely to receive SSI, apart from sinusitis which was the most common but was in the median range of number of people receiving income for it next to heart disease and diabetes. Those who have a limitation that was found not to be qualifiable may also be receiving it due to low income or old age--essentially other factors that could affect the amount they can work, since most people receiving SSI marked that they do not have a disability/ limitation. All this being said, those that have disabilities that didn't qualify for SSI may be receiving supplemental income from other programs where they were either more qualified for it, or they simply applied to those programs first and got approved.

# References

https://www.healthcare.gov/glossary/federal-poverty-level-fpl/   
  
U.S. Census Bureau. "Census Regions and Divisions of the United States" (PDF). U.S. Census Bureau. U.S. Census Bureau. Retrieved 6 December 2018. https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us\_regdiv.pdf   
  
https://news.umich.edu/new-map-offers-snapshot-of-poverty-well-being-in-michigan/   
  
https://aspe.hhs.gov/2003-poverty-guidelines-federal-register-notice