SI 618 Project Part I Report

By: Shivangi Kumar

Motivation

Nearly 20% of U.S. adults experience some form of disability. People with disabilities often self-report reduced access to preventive health services and poorer health than people without disability. Risk factors for chronic disease are more prevalent in people with disabilities, increasing risk for secondary conditions including cardiovascular diseases. It has been demonstrated that even modest levels of physical activity can offer benefits in ameliorating cardiovascular diseases risk in adults of all ages which is what motivated me to take up such datasets that can give statistics for persons with disabilities who are at risk of cardiovascular diseases. If proper analysis is done with multivariate data, I believe that the causes of the link between cardiovascular disease and persons with disability can be examined. On a high level, through this project I try to address the risk percentages for cardiovascular diseases associated with each disability type.

Data Sources

Disability and Health Data System (DHDS)
Source: https://data.cdc.gov/Disability-Health/Disability-and-Health-Data-System-DHDS-/k62p-6esq

The first dataset is the Disability and Health Data System (DHDS) which is an online source of state-level data on adults with disabilities. This dataset has been provided by Centers for Disease Control and Prevention, National Center on Birth Defects and Developmental Disabilities. This data can be exported in .csv format, although it will require concatenation to develop a longitudinal dataset. Through this dataset, we can access information on six functional disability types: cognitive (serious difficulty concentrating, remembering, or making decisions), hearing (serious difficulty hearing or deaf), mobility (serious difficulty walking or climbing stairs), vision (serious difficulty seeing), self-care (difficulty dressing or bathing) and independent living (difficulty doing errands alone). This dataset contains 546K rows and 32 columns most important being Year, location_description containing names of states across US, data_value containing prevalence percentages (prevalence is the proportion of a population who have a specific characteristic in each time period [2]), Stratification1 containing disability type and stratification2 containing demographic variables. Within the year column in this dataset, the values range from 2016 to 2020 which has been filtered to use data from 2016 to 2018 so that both datasets can be concatenated.

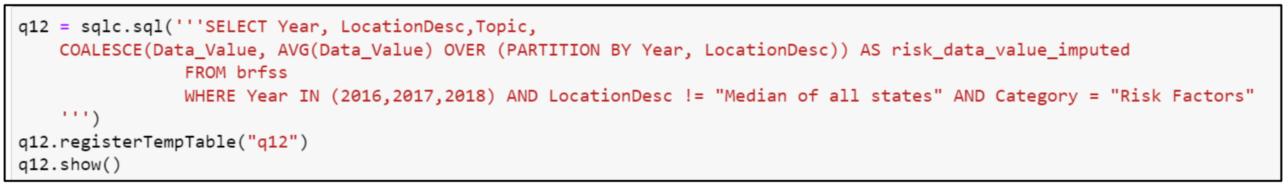
Behavioral Risk Factor Surveillance System (BRFSS) - National Cardiovascular Disease Surveillance Data
Source: https://chronicdata.cdc.gov/Heart-Disease-Stroke-Prevention/Behavioral-Risk-Factor-Surveillance-System-BRFSS-N/ikwk-8git

The other dataset is based on cardiovascular diseases and risk factors associated with it. The Behavioral Risk Factor Surveillance System is a continuous, state-based surveillance system that collects information about modifiable risk factors for chronic diseases and other leading causes of death. This is one of the datasets provided by the National Cardiovascular Disease Surveillance System. This data can be exported in .csv format, although it will require concatenation to develop a longitudinal dataset. The data are organized by location (national, regional, state, and selected sites) and indicator, and they include CVDs (e.g., heart failure) and risk factors (e.g., hypertension).

The data can be plotted as trends and stratified by age group, sex, and race/ethnicity. This dataset contains 126K rows and 30 columns most important ones being Year, location description, break_out_category containing the demographic variables, topic specifying the type of risk and data_value containing the risk percentages. Within the year column in this dataset, the values range from 2011 to 2018 which has been filtered to use data from 2016 to 2018 so that both datasets can be concatenated.

Data Manipulation Methods

The first and foremost manipulation that I did for both datasets was to filter them such that they contain rows which belong to years 2016 to 2018 as these were the common years in both the datasets. This step would also be incorporated at times when join is being performed based on the Year column. Additionally, I performed a join on both datasets for each of the three tasks separately depending upon the required results and select columns for use. Another manipulation method used was to replace null values of the risk percentages column by the aggregate mean of the risk percentages based on year and location (state of the US) using the COALESCE function.



```
q12 = sqlc.sql('''SELECT Year, LocationDesc,Topic,
    COALESCE(Data_Value, AVG(Data_Value) OVER (PARTITION BY Year, LocationDesc)) AS risk_data_value_imputed
                FROM brfss
                WHERE Year IN (2016,2017,2018) AND LocationDesc != "Median of all states" AND Category = "Risk Factors"
    ''')
q12.registerTempTable("q12")
q12.show()
```

Fig 1: Use of COALESCE function for imputation of null values

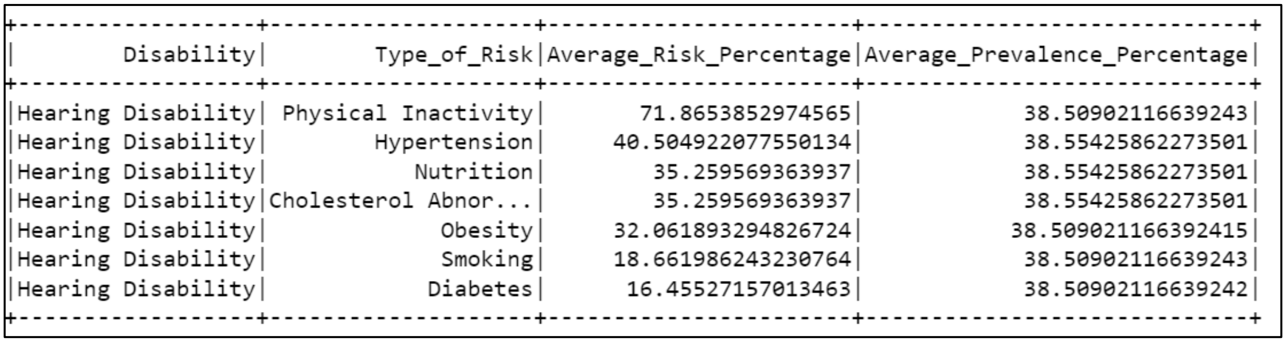
In task 1, I performed a JOIN based on the Year and LocationDesc which is location description column. This was ideal in the case of wanting to explore the trends of risk and prevalence percentages experienced by individuals with hearing disability who are at some kind of risk (hypertension, diabetes). In task 2, I performed a JOIN based on year, location, demographic_value columns as we needed the male to female ratio as a part of the question. In task 3, I again performed a JOIN between manipulated datasets based on the Year and LocationDesc columns which allowed me to take a look at the individual counts in the final dataset per state.

Analysis and Visualization

Task 1: Determine the trend of risk and prevalence percentages experienced by individuals with hearing disability who are at some kind of risk (hypertension, diabetes, etc.).

First, for the disability dataset, I grouped by year, location, and stratification1 (disability type) columns which would give the average prevalence percentages for each year, state, and disability type combinations. Then I performed manipulation on the other cardiovascular dataset by null imputation through COALESCE function, excluding "median of all states" from list of states, and restricting the analysis to include only risk factors associated with cardiovascular diseases for the years 2016 to 2018. After manipulating this data frame, I fetched the average risk percentages by grouping on year, location (state), and type of risk (topic column). After joining the two tables based on year and location, I filtered to fetch results for only the hearing disability type and found the respective averages of prevalence and risk percentages for the various risk factors associated with

the hearing disability. One thing that didn't work here was considering all the disability types and restricting to only the hearing disability as there was not much variation in the risk percentages for each of the disability type to analyze it further.



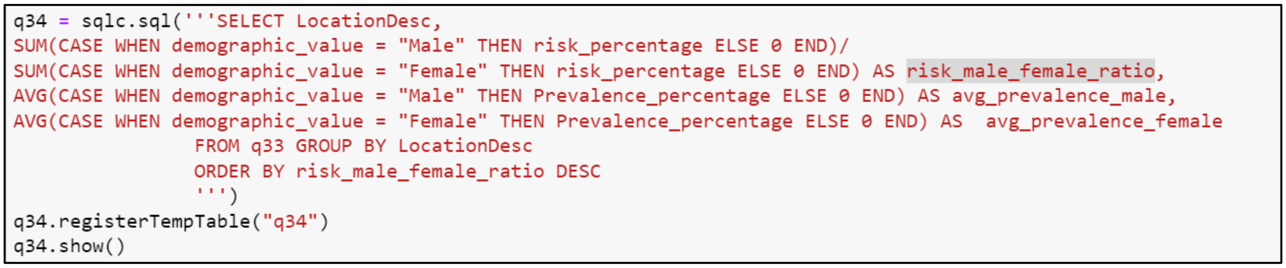
```
+------------------+--------------------+-----------------------+-----------------------------+
|        Disability|        Type_of_Risk|Average_Risk_Percentage|Average_Prevalence_Percentage|
+------------------+--------------------+-----------------------+-----------------------------+
|Hearing Disability| Physical Inactivity|       71.8653852974565|            38.50902116639243|
|Hearing Disability|        Hypertension|     40.504922077550134|            38.55425862273501|
|Hearing Disability|           Nutrition|        35.259569363937|            38.55425862273501|
|Hearing Disability|Cholesterol Abnor...|        35.259569363937|            38.55425862273501|
|Hearing Disability|             Obesity|     32.061893294826724|           38.509021166392415|
|Hearing Disability|             Smoking|     18.661986243230764|            38.50902116639243|
|Hearing Disability|            Diabetes|      16.45527157013463|            38.50902116639242|
+------------------+--------------------+-----------------------+-----------------------------+
```

Figure 2: Hearing disability and risk percentage

Interpretation: It can be seen from this task that 38.5% of the population (prevalence percentage) who have hearing disability are at a 71.86% risk due to physical inactivity which is the highest in case of hearing disability. People with hearing disability are at least risk of diabetes with a risk percentage of 16.4% for 38.5% of the population with hearing disability.

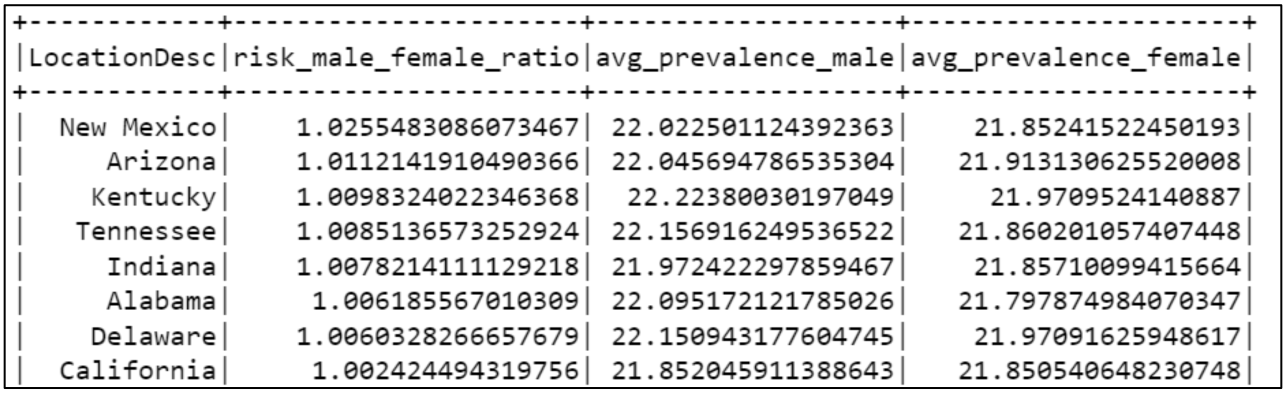
Task 2: Calculate the male to female ratio based on aggregated risk percentages for each state in the United States such that it shows the gender distribution of the state whose individuals are at a comparatively higher risk (based on prevalence percentage).

First, I manipulated the disability dataset by grouping it by year, location (state), and the gender type to get the respective prevalence percentages. This means that each state will appear 6 times in the dataset for each gender (male or female) for each year (2016 to 2018). I performed a similar manipulation based on year, location, and gender type for the cardiovascular dataset to fetch the average risk percentages associated with each gender type for each state in 2016 to 2018. For the cardiovascular dataset, there were missing values for gender type (demographics) column which I decided to drop before merging it with the other dataset. One thing that didn't work here was that I could not group on race/ethnicity and age groups as the format and range of age groups in each dataset was different which made it impossible to merge the datasets based on race/ethnicity and age groups. After performing the manipulation tasks, I proceeded to join the two tables based on year, location (state), and gender type (demographics). The result dataset gives the output for prevalence and risk percentages for each gender type for each state in the US from 2016 to 2018. Now aggregating just by the states (and removing years), I used CASE WHEN to get the sum of average risk percentages for each gender type and using it to calculate the ratio of male to female based on the sums.



```
q34 = sqlc.sql('''SELECT LocationDesc,
SUM(CASE WHEN demographic_value = "Male" THEN risk_percentage ELSE 0 END)/
SUM(CASE WHEN demographic_value = "Female" THEN risk_percentage ELSE 0 END) AS risk_male_female_ratio,
AVG(CASE WHEN demographic_value = "Male" THEN Prevalence_percentage ELSE 0 END) AS avg_prevalence_male,
AVG(CASE WHEN demographic_value = "Female" THEN Prevalence_percentage ELSE 0 END) AS  avg_prevalence_female
                FROM q33 GROUP BY LocationDesc
                ORDER BY risk_male_female_ratio DESC
                ''')
q34.registerTempTable("q34")
q34.show()
```

Figure 3: Calculate male-to-female ratios based on sum of risk percentages for all three years



```
+------------+----------------------+-------------------+---------------------+
|LocationDesc|risk_male_female_ratio|avg_prevalence_male|avg_prevalence_female|
+------------+----------------------+-------------------+---------------------+
|  New Mexico|    1.0255483086073467| 22.022501124392363|    21.85241522450193|
|     Arizona|    1.0112141910490366| 22.045694786535304|   21.913130625520008|
|    Kentucky|    1.0098324022346368|  22.22380030197049|     21.9709524140887|
|   Tennessee|    1.0085136573252924| 22.156916249536522|   21.860201057407448|
|     Indiana|    1.0078214111129218| 21.972422297859467|    21.85710099415664|
|     Alabama|     1.006185567010309| 22.095172121785026|   21.797874984070347|
|    Delaware|    1.0060328266657679| 22.150943177604745|    21.97091625948617|
|  California|     1.002424494319756| 21.852045911388643|   21.850540648230748|
```

Figure 4: Result of Task 2

Interpretation: It can be interpreted from this table that 22% of males and 21% of females in New Mexico of the total population considered are at a high risk and form a part of the male to female risk ratio. Almost all states have the risk male to female ration close to 1 indicating that approximately more or less both males and females are equally at risk of cardiovascular diseases in 2016 to 2018. Cardiovascular disease (CVD) is the leading cause of death worldwide, yet important differences exist between men and women as there has been research on this. From past research, men generally develop CVD at a younger age and have a higher risk of coronary heart disease (CHD) than women and women, in contrast, are at a higher risk of stroke, which often occurs at older age [1]. Thus, in our analysis, we don't see a contrasting difference between men and women as in this dataset we considered the overall risk to cardiovascular diseases instead of the bifurcation between coronary hear disease (CHD) and stroke for our analysis.

Visualization:

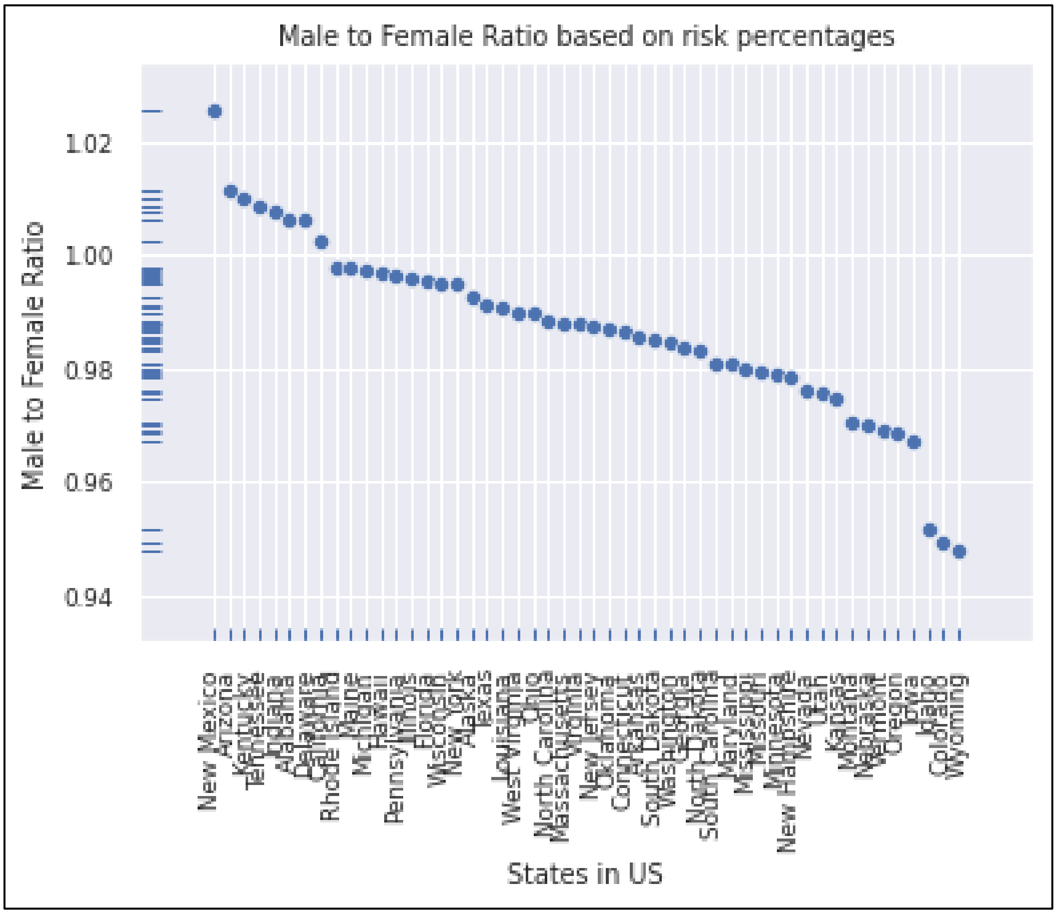
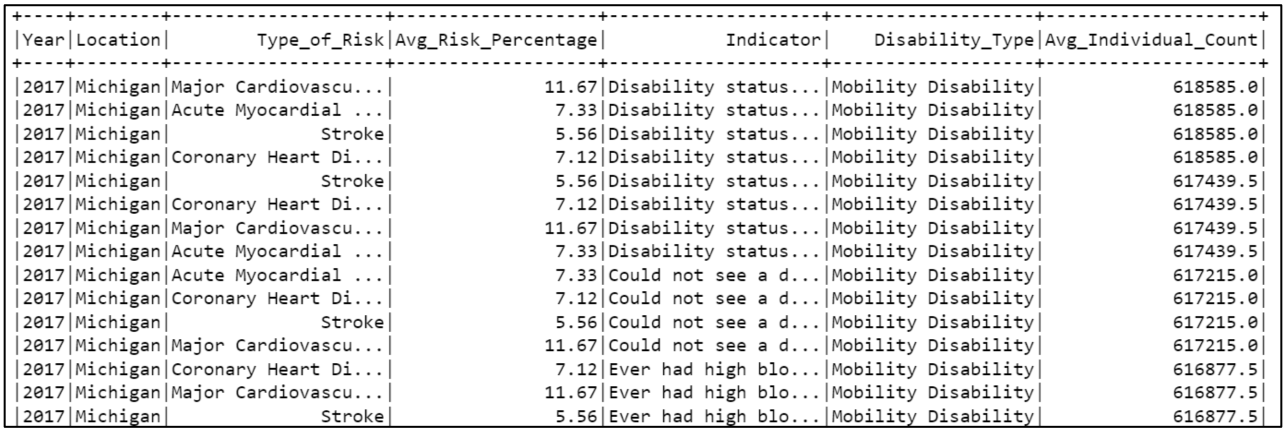


Figure 5: Scatterplot and rug plot indicating the male to female ratio based on risk percentages for each state in the US from 2016 to 2018

I plotted a scatterplot and rug plot to fit the male to female ratios calculated for each state. To do this, I used seaborn library in python to generate this visualization. Additionally, as the state names were overlapping with the default display of x-axis labels, I rotated the x-axis labels by 90 degrees to be clearly visible. On the x-axis are the state names and on the y-axis are the male to female ratios. One thing which I learnt was to first convert the spark data frame to a pandas data frame to use the seaborn library functions.

Task 3: Calculate the individual count for each indicator of a cardiovascular disease for people in Michigan between years 2016 to 2018 who have mobility disability.

First, I filtered the cardiovascular diseases dataset to include only the categories which have the type of cardiovascular diseases in it and get rid of the risk factors category. I mutated the risk percentage column by replacing null values with the mean value of risk percentage after which I grouped by year, location (state), and the cardiovascular disease type. In the disability dataset, I manipulated it by performing a group by on year, location, disability type, and the indicator which indicates the reason behind the disability type, and this results in fetching the count of the individuals for this grouped combination. There are 42 unique values for the indicator column in this dataset. After joining the two datasets based on year and location, I filtered the dataset to contain only the Michigan state data for individuals with mobility disability. I do not really see anything that did not work here, but I did face some issues in incorporating type of risk in the graph in addition to all other variables.



```
+----+--------+--------------------+-------------------+--------------------+-------------------+--------------------+
|Year|Location|        Type_of_Risk|Avg_Risk_Percentage|           Indicator|    Disability_Type|Avg_Individual_Count|
+----+--------+--------------------+-------------------+--------------------+-------------------+--------------------+
|2017|Michigan|Major Cardiovascu...|              11.67|Disability status...|Mobility Disability|            618585.0|
|2017|Michigan|Acute Myocardial ...|               7.33|Disability status...|Mobility Disability|            618585.0|
|2017|Michigan|              Stroke|               5.56|Disability status...|Mobility Disability|            618585.0|
|2017|Michigan|Coronary Heart Di...|               7.12|Disability status...|Mobility Disability|            618585.0|
|2017|Michigan|              Stroke|               5.56|Disability status...|Mobility Disability|            617439.5|
|2017|Michigan|Coronary Heart Di...|               7.12|Disability status...|Mobility Disability|            617439.5|
|2017|Michigan|Major Cardiovascu...|              11.67|Disability status...|Mobility Disability|            617439.5|
|2017|Michigan|Acute Myocardial ...|               7.33|Disability status...|Mobility Disability|            617439.5|
|2017|Michigan|Acute Myocardial ...|               7.33|Could not see a d...|Mobility Disability|            617215.0|
|2017|Michigan|Coronary Heart Di...|               7.12|Could not see a d...|Mobility Disability|            617215.0|
|2017|Michigan|              Stroke|               5.56|Could not see a d...|Mobility Disability|            617215.0|
|2017|Michigan|Major Cardiovascu...|              11.67|Could not see a d...|Mobility Disability|            617215.0|
|2017|Michigan|Coronary Heart Di...|               7.12|Ever had high blo...|Mobility Disability|            616877.5|
|2017|Michigan|Major Cardiovascu...|              11.67|Ever had high blo...|Mobility Disability|            616877.5|
|2017|Michigan|              Stroke|               5.56|Ever had high blo...|Mobility Disability|            616877.5|
```

Figure 6: Average individual counts for different types of risk and indicators for people in Michigan with mobility disability

Interpretation: For individuals in Michigan with mobility disability, it can be seen that more number of individuals are at a risk for major cardiovascular disease with a risk percentage of 11.67%. We can also see that more number of individuals in Michigan were at some form of cardiovascular disease risk in the year 2017 as compared to 2016 and 2018.

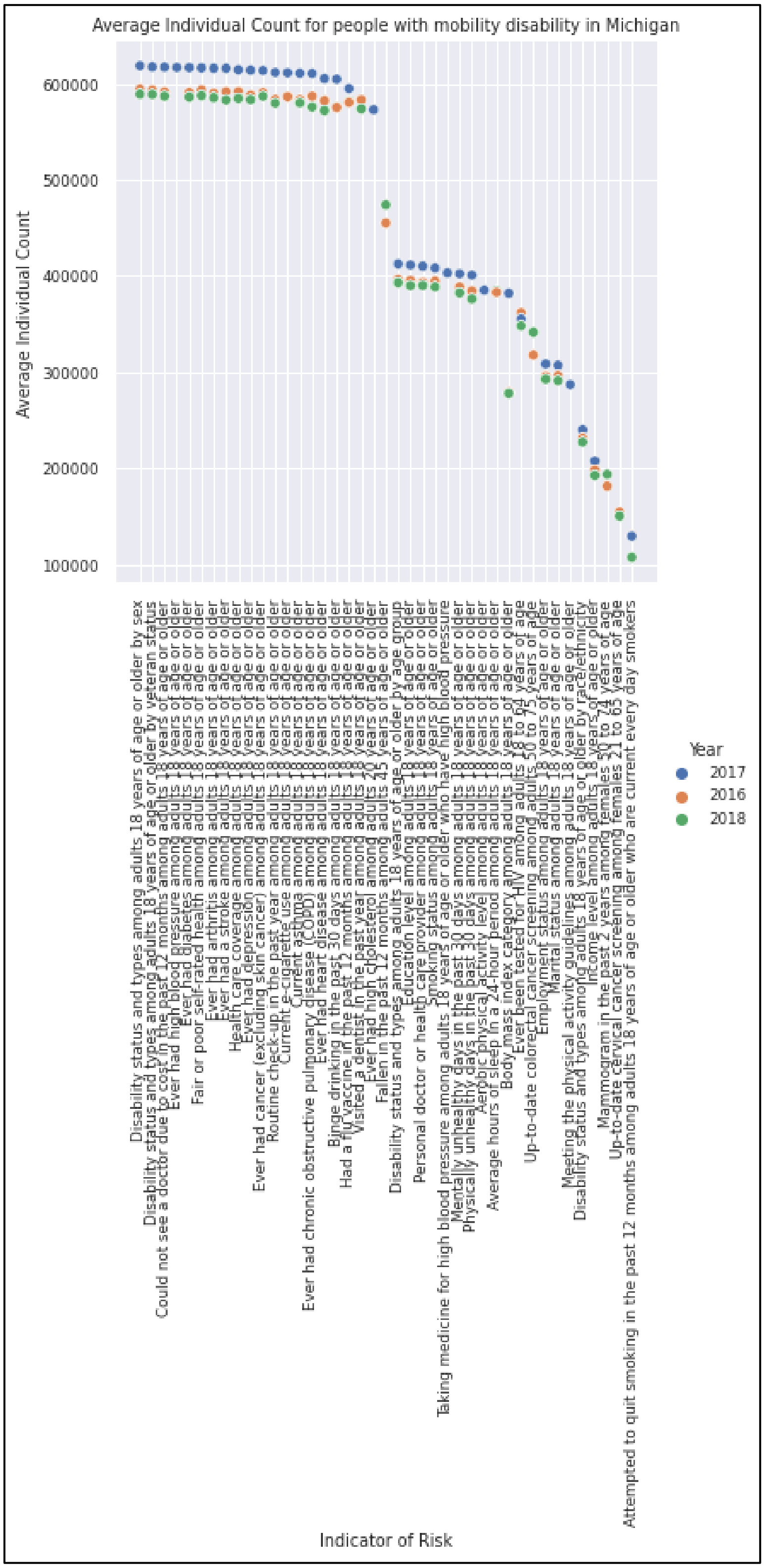


Figure 7: Relational plot displaying the relationship between indicator of risk and average individual count from 2016 to 2018

I plotted a relational plot for this task with indicator being on the x-axis and average individual count being on the y-axis. I used color to indicate the year. Also, for this visualization I used the Seaborn library in python. This graph makes it clear to understand that for most indicator types in people with mobility disability and living in Michigan, the year 2017 has higher average count of individuals who are at risk of some kind of cardiovascular disease (2017 is depicted in blue color).

Challenges

There were three main challenges that I faced throughout this project. One main issue was that I had to cut down on a large chunk of the disability dataset owing to the year constraint from 2016 to 2018 for me to be able to merge it with the other dataset. Another challenge was the issue on dropping missing values that did not have a value for the gender type column in one of the datasets. Also, there were null values for the risk percentage in the dataset which had to be imputed with the aggregated mean of groups. I had to figure out how to take an aggregated mean of each group and replace null values with that mean for that group only. I learnt the use of the COALESCE function throughout this process. Lastly, I had planned to perform analysis for race/ethnicity, gender, and age groups for both datasets, however, due to different notations of race/ethnicity and different ranges of age groups in the datasets, it was not possible to merge them based on these things. Thus, the analysis was restricted to only gender type analysis.

References

[1] Bots SH, Peters SAE, Woodward M
Sex differences in coronary heart disease and stroke mortality: a global assessment of the effect of ageing between 1980 and 2010
BMJ Global Health 2017;2:e000298.

[2] https://www.nimh.nih.gov/health/statistics/what-is-prevalence