Analysing Social Determinants of Health – Age, Income, Crime w.r.t. Health Insurance

**MISM 6205 – Data WRANGLING FOR BUSINESS**

Prof.CAROL LEE

# **FINAL PROJECT REPORT**

# Group Members:

Vasatika Ghadiyaram

Hemanshu Dhuria

Siddhant Ambardekar

Mehal Sanghvi

Diagram

Description automatically generated

**Topic:**

Analysing Social Determinants of Health – Age, Income, Crime with respect to Health Insurance.

**Importance:**

“Social determinants of health (SDOH) are the conditions in the environments where people are born, live, learn, work, play, worship, and age that affect a wide range of health, functioning, and quality-of-life outcomes and risks.”As part of your health insurance, Insurance companies determine what tests, drugs, and services they will cover. These choices are based on their understanding of the kinds of medical care that most patients need. These needs are directly related to the social determinants of health. While the exact indices are debatable, not all indices will be dominant in each region. Hence, it is important to understand which indices have more impact in each region.

**Business Questions:**

1. Determine the groups at risk basis social determinants of health (SDOH) that can be targeted for providing insurance.
2. Understand target audience groups and ideate ways to provide health insurance for vulnerable groups and determine the coverage.

**Who would use this?**

Health insurance companies have been moving from paying for volume to paying for value, risk adjustment in current models and measures don’t adequately account for SDOH. The companies now are moving towards quality measurement of health and so out combined dataset is meant for insurance companies in Massachusetts and the result of our analysis would be used by them in designing their health care plans and determining their coverage payments.

**Data:**

The data sets we have used to understand the social determinants of health are the overall US census data based on the SDOH indices this includes all the focused groups we aim to analyse, and Crime Data with types of crimes and some demographic groups.

**SDOH Data:** <https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html#download>

* **Data Description:** Variables in the data are related to five main SDOH domains: social context (e.g., age, race/ethnicity, veteran status), economic context (e.g., income, unemployment rate), education, physical infrastructure (e.g., housing, crime, transportation), and healthcare context (e.g., health insurance).
* **Type of Data:** Secondary Data collected by AHRQ.
* **Data Collection process:** Drawn from multiple sources (original data documents, codebooks, surveys, and environmental scan)

**CRIME Data:** <https://www.kaggle.com/datasets/michaelbryantds/crimedata>

* **Data Description:** Types of crimes, percentages of crime, demographic groups like age, race, and police force. This data has a metrics according to states and counties.
* **Type of Data:** Secondary Data taken from Kaggle.
* **Data Collection process:** Data collection process is unknown and not mentioned in the main Kaggle file, however, we found a data dictionary file for the columns from the following source to understand the columns.
* **Data Dictionary Source:** <http://archive.ics.uci.edu/ml/datasets/communities+and+crime+unnormalized>

**Information quality:**

* The data set is publicly available over government websites making it easily accessible. The excess information provided in the dataset might be difficult for us to find the most relevant columns for the analysis. This can be solved by formatting the data and using most relevant data columns after data profiling.
* Data set collected has missing or null values in the columns and is heterogeneous in nature. Data columns are inconsistent with various types of columns (core, internal and unknown to the study).
* Data sets included are collected from various data sources which do not match. Data wrangling process needs to be carried out to maintain a stable flow of data for analysis.
* Data scope for the analysis is wide and has finely distributed. The Data Collection process for this project is a continuous process. However, the time frame required to collect this data will be longer as the data scope is wide.
* Data sets collected are from reputed sources (government agencies) and has descriptions on how the data was collected and what each column stands for. The government agency website also has accurate citation for the data set.
* Data sets collected from various sources have enough data to address the business questions and comp up with an analysis. However, the data sets would be completely useful once a through data wrangling process is carried out through python and excel (collection, transformation, profiling, processing). These processes will help us cleanse the data and use only the information that is required in the process to come up with an analysis.

**Methods and tools**:

The data wrangling methods and theories that we will be using to transform the data as per the needs of the analysis done to answer the business questions are:

* Data Discovery - collecting and analysing data from various sources to understand overall patterns in the data.
* Data Formatting - organizing the data based on predetermined specifications.
* Data Profiling - creating a report to analyse the summaries of the data.
* Data Pre-processing/Cleaning - preparing the data to be used in the proper form and ensuring any irrelevant, distorted, or incomplete data is cleaned.
* Data Enrichment - filling the gaps by adding additional data to ensure consistency in the dataset.
* Data Integration - combining the datasets and the information from the resources.

The languages, and scripts we used are:

* Python (Jupyter Notebook)
  + Pandas - machine learning library with various tools and structures suitable for data wrangling.
  + NumPy - numerical python allows multidimensional data analysis and use of mathematical functions.
  + Scikit-learn - can be used for supervised and unsupervised learning algorithms. Correlation, regression, cluster analysis, etc.
  + Matplotlib - used for data visualization and plotting.
  + Seaborn – used for regression analysis and plotting.
* Excel - Storing the data and performing basic calculations using the functions available. Making Pivot tables to understand the data in a more concise way and charts.

**Data Wrangling Process**:

Our data wrangling process was entirely performed in Python, it gave us a good platform to keep our steps neat and concise and introduced us to new unique libraries, and functions. Following is the description of the Data Wrangling steps and in-depth description of the process in each of the steps.

1. **Data Discovery:**
   * The project required thorough research to find census data with SDOH indices and the source found included a data dictionary sheet with clear description of all the columns, and their data types as well.
   * Finding recent crime data was not a hard process, because it was available in Kaggle, however, it did not have a dictionary with descriptions, so it required further research to find a relevant source that included column descriptions and relevance.
2. **Data Formatting:**

* Imported the Pandas, NumPy, Seaborn, Pandas Profiling libraries.
* Split the formatting process into two for customization according to the different data sets.

**SDOH Census Data**

* Imported the excel file into the console using pd.read\_excel, and selecting sheet\_name as 1 because Sheet 0 includes the data dictionary.
* Listed down all the columns to understand which columns will be needed.
* After selecting the columns. Subsetting the data according to the unit of analysis selected which Massachusetts, and County wise.
* Created a new DataFrame to assign the selected columns which reduced columns from 682, to 81.
* Converted the datatype of “CountyFIPS” column to for splitting into two, one of county, and other for state. We used as.type(str) and str.split for these steps.
* After splitting, changing the datatype back to int.
* Next, we deleted the state code column and changed the data type for the county code to integer.
* Added the county code’s column to the census dataset and removed other columns that could be dropped.
* Created a subset DataFrame from the Mass\_Census data create charts and analyze relationships between variables. Health\_Insurance was created using percentage of uninsured people, and county codes. For this we created a pie chart to understand the proportion dynamics.

Chart, pie chart

Description automatically generated

Fig.1- Pie chart showing % of people who are uninsured in each county.

Health insurance companies could look upon the chart to focus on people who are currently uninsured.

* Created a subset DataFrame for people with income less than 10000 according to county code and created a pie chart.

Chart, pie chart

Description automatically generated

Fig.2. – Pie Chart showing % income less than 10000 with Counties

* Created a subset DataFrame Medicaid availability according to county codes and created another pie chart.

Chart, pie chart

Description automatically generated

Fig.3- Pie chart shows % of people taking Medicaid in each county

From Fig. 2 and Fig. 3 we can infer that proportion of people with income less than 10000 in counties 13 and 25 is aligning with the proportion of people who are provided Medicaid in the same counties.

* Performed correlation between Medicaid and Income less than 10000

Graphical user interface, text, application

Description automatically generated

This confirmed that there is a strong positive correlation between people with less than $10000 income and people with Medicaid insurance provided.

* In the same way we performed correlation between private insurance and income greater than $100,000 and created a double-bar graph to show the comparison between both.

Chart, bar chart

Description automatically generated

Fig.4. – Bar Graph shows that % people buying private insurance and % who has income greater than 10000 in each county.

Fig.4 shows health insurance companies need to focus on counties who have income greater than 100000 USD but still not buying private health insurance.

* We also created a regression plot using the seaborn library which was unique to our process. We used Age groups and uninsured.

Chart, scatter chart

Description automatically generated

Fig.5.- Regression chart shows that age 15-17 are the maximum uninsured.

* We concluded by creating a profile report for the final Mass\_Census data.

**Crime Data**

* Imported the Crime Data to the variable using pd.read\_csv.
* Listed down all the columns to understand which columns would align well with the SDOH census data. Limited the columns to age group and types of crimes keeping unit of analysis as county code.
* Subsetted the data for Massachusetts and created a new DataFrame with the columns chosen for analysis.
* Changed the required datatypes of the relevant columns for further formatting.
* Formatted the Community name column removing the suffix for each and sorted the Mass level crime data by county code number for further merging process.
* Row wise aggregation was the most unique process encountered in the formatting of the crime data. Ensured that no values were 0 as it would cause a problem for aggregation.
* Selected a type of aggregation function for each of the columns and performed row-wise merging by aggregation.
* New data became concise with 23 columns and 12 rows according to county code compared to the original data with 146 columns.
* Performed correlation between the age groups and types of crimes according to county and got the following table and analysis.

Graphical user interface, application, table

Description automatically generated

We can see a few examples where the correlation between a certain age group and type of crime is high. For example, for the percentage of people in the age group 12 to 29 the correlation between murder and the age group is 0.785905 which is approximately 79%. This is a strong indicator that murders have occurred more in the percentage of population of the age group 12 to 29.

Another example is the correlation between the percentage of people above the age of 65 and all the crimes. We can see negative values for all the crime which indicates a negative correlation meaning that there are less crimes in the population of people who are 65 and above.

* Created a profile report of the polished Crime dataset.

**Merged Data:**

* With the merged data we used correlation to assess relationships between crime, income, age, and overall health insurance.

1. **Data Preprocessing:**

* Found the number of null values in all columns.
* Replaced the 0's with Nan values in the rows to avoid incorrect calculation.
* Created a county code column as a common column in Crime data for merging.
* Brought the CountyCode column to first column using (df.pop(“column”).

1. **Data Enrichment:**

* Created new data frames for sub-level analysis of columns.
* Converted County code column to string for the pie chart.
* Created a pie chart, regression plot and bar charts using seaborn library.
* Performed correlation between various relevant columns using [.corr].
* Merged both datasets together using outer join.
* Dropped the common age columns for repetition.
* Exported the merged dataset to csv.

1. **Data Profiling:**

* Profiled both the datasets to check for outliers, null values, missing values highly correlated, data quality and overall characteristics and quality of data.

1. **Data Visualization in Python:**

* Plotted pie charts, bar graphs and regression charts using pandas’ function and seaborn libraries.
* Function like plot, kind=bar, pie, sns.regplot are used.
* Visualized analysed data with different columns showing relations between age, income, and health insurance by correlation or comparing them.

**Validation Rules:**

* After importing the datasets, we will check for the missing values and Null values in the dataset when first entered in the system by data collector. There can be multiple sources of same data (Data Quality problem). It is a system validation rule type. The specific process step and task is to run the data profiling before importing the dataset and use it.
* While dropping columns which are irrelevant, we need to validate if the columns are irrelevant or not. This helps in data formatting and cleaning. It helps in analysing better. This serves as a business rule as this is checked by a person or process without using information technology. Ex- dropping the columns named people who have broadband, as it is not necessary for specific business problem.
* While splitting and changing cases of the columns, data needs validation if these types of changes are allowed or not. This is a system validation rule type. ex- we can validate the data by checking if the cause is not impacted by splitting the address columns and changing its case.

**Analysis and Outputs**:

* The outcome expected is an integrated dataset which has 14 rows and 100 columns containing attributes related to social determinants of health and mortality rate associated with it with demographic representation in each county
* The dominance and dependence of a determinant could be gauged to calculate the risk percentage, which could be used in deciding the premium and coverage of different health plans.
* As part of our example analysis, we created regression model to check for correlation between murder & income groups and murder & age group to identify the groups where impact is the most and the groups where the correlation is negative.
* The income groups under $10000 are eligible for government provided Medicaid and the Age group above 65 is eligible for government provided Medicare and as per data the groups have been availing these benefits provided, hence these groups could be neglected.
* The T.1 shows the correlation between Income and Murders, the high-risk groups found here are income groups under $10000 and income groups under $14999. Given the Medicaid benefits provided by the government, it is safe to say that the focus could be put more on the less than $14999 income group. At the same time, the groups with <49999 and have a negative correlation, it will be safe to say that the group is at low risk for murders and their coverage could be adjusted accordingly.
* From T.2. The regression is performed between Age groups and murders, and the high-risk group with 75% correlation are 30-44 age. At the same time, people of age group 45-64 have a negative correlation with murders and hence the coverage payments could be minimised.
* T.3. Identifies the counties where presence of both the groups is significant is highlighted below:

|  |  |  |
| --- | --- | --- |
| **COUNTY NAME** | **INC<14999** | **AGE(30-44)** |
| Barnstable County | 2.92 | 13.22 |
| Berkshire County | 4.34 | 15.79 |
| **Bristol County** | **4.81** | **18.61** |
| Dukes County | 1.82 | 16.59 |
| **Essex County** | **4.6** | **18.22** |
| Franklin County | 4.6 | 17.68 |
| Hampden County | 5.94 | 18.03 |
| Hampshire County | 3.97 | 14.78 |
| Middlesex County | 2.97 | 20.88 |
| Nantucket County | 1.48 | 24.5 |
| Norfolk County | 2.9 | 19.2 |
| Plymouth County | 3.19 | 17.07 |
| **Suffolk County** | **6.01** | **23.47** |
| Worcester County | 4.11 | 18.66 |

* The three main counties where the insurance companies could focus more are Bristol County, Essex County and Suffolk County.

|  |  |
| --- | --- |
| **Income Groups** | **Murders** |
| Under 10000 | 78% |
| Under 14999 | 53.30% |
| Under 49999 | -34.20% |
| Under 99999 | -56.60% |

|  |  |
| --- | --- |
| **Age Group** | **Murders** |
| 18-29 | 52% |
| 30-44 | 75.00% |
| 45-64 | -76.96% |
| Above 65 | -56.78% |

**What more could the companies do?**

The companies could perform the same regression for all the possible social determinants of health and develop a risk score which takes into consideration all the determinants significant in different geographical areas.

**External material**:

We have researched and found out some external websites and articles that support our objective.

1. CHIR. “Insurance Companies Are Investing in the Social Determinants of Health, but Widespread Changes in Benefits Remain to Be Seen.” Centre on Health Insurance Reforms, 7 May 2020,  <http://chirblog.org/insurance-company-investments-social-determinants-of-health/>
   * This article is about the way major health insurance companies are investing in improving SDOH. They are not directly changing their already existing policies especially for SDOH, rather they are just putting straight money through their donations and finds organization. It will really help in analysing as to why people are finding difficulty to reach the insurance companies.
2. Posted on January 17, 2019, by Carli Friedman. “Introducing the Social Determinants of Health Index.” The Council on Quality and Leadership, 17 June 2021,<https://www.c-q-l.org/resources/newsletters/introducing-the-social-determinants-of-health-index/>
   * This article will help us knowing about social determinants of health faced by some people and their importance. It also talks about the social factors that impact health.
3. Social Determinants of Health Database. Content last reviewed July 2022. Agency for Healthcare Research and Quality, Rockville, MD. <https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html>
   * This website has provided us with lots of SDOH focused data. The data focuses on five key SDOH domains and it is linked to other data by geography. The data is distributed year-wise.
4. Townsend, Robin. “Average Cost of Health Insurance (2022).” Value Penguin, Value Penguin, 12 Oct. 2022,<https://www.valuepenguin.com/average-cost-of-health-insurance>
   * This article will help us to know about what an average American pays for health insurance per month. Price of health insurance is one of the major factors in determining social determinants of health.
5. Stasha, Smiljanic. “Health Insurance Statistics and Facts (2021): Policy Advice.” Health Insurance Statistics and Facts (2021) | Policy Advice | Policy Advice,<https://policyadvice.net/insurance/insights/health-insurance-statistics/>
   * This article provides us with all the statistics related to health insurance and facts like number of American uninsured, expenditure on healthcare from household budget etc. These facts will make us understand the rising trend of insurance costs and understand people’s purchase power in terms of insurance.
6. Himber, Vaughn. 9 Reasons to Offer Small Business Health Insurance. <https://www.ehealthinsurance.com/resources/small-business/9-reasons-to-offer-small-business-health-insurance>
   * This article will tell us about the benefits of employee health insurance. This will help us in understanding the difference insurance companies create between people when it comes to minority and majority class.
7. <https://core.ac.uk/download/pdf/131215565.pdf>
8. <https://www.universalclass.com/articles/business/the-impact-of-crime-on-community-development.htm>
9. <https://www.kff.org/racial-equity-and-health-policy/issue-brief/beyond-health-care-the-role-of-social-determinants-in-promoting-health-and-health-equity/>
10. https://bettermedicarealliance.org/wp-content/uploads/2021/08/Innovative-Approaches-to-Addressing-SDOH-for-MA-Beneficiaries-FINAL.pdf

**Additional data and analysis:**

* Develop regression model factoring all SDOH indices to generate score calculating overall risk level.
* Repeat the process for other SDOH factors: While this process has been done for crime (specific to murder) we can do this for other SDOH factors such as education, economic status etc. This will create a holistic view for the health insurance companies to formulate their health care plans.
* Iterate this process for country level or state level data: This project focuses on counties only in Massachusetts. To expand the scope and coverage of the business outcome this process can be conducted for different states in a country or on a country level.