LARA SULZBACH
28/11/24

Data Report

The intention of my project work is exploring if the amount of physical activity has an impact on mental health. As it is common knowledge that physical activity increases the dopamine level and causes a better feeling, I want to explore if such a correlation can be identified in statistical data of the American society. To be more precise, my goal is to show that in US states where people are more physically active problems with mental health are less common.

Data Sources

For answering my question, I will compare a study about the time people spend on physical activities with data of a different survey about mental health.

Datasource 1: Physical Activities

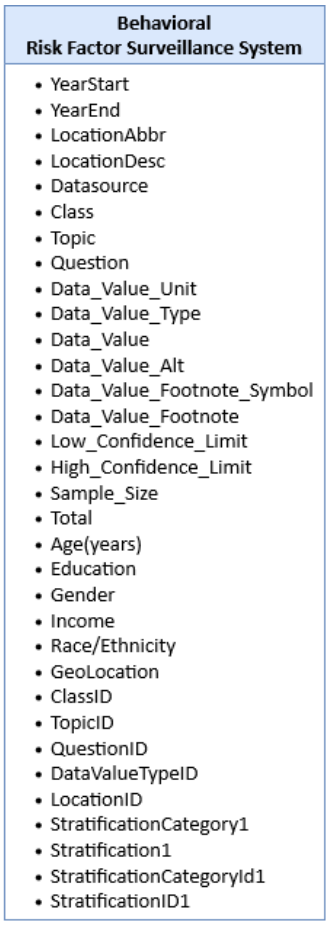

Figure 1: Database schema Physical Activities

My first datasource includes data on adult's diet, physical activity, and weight status from the Behavioral Risk Factor Surveillance System. The file is labelled as *public use data*[1] which allows a free use of the dataset without restrictions. It can be downloaded as csv file[2]. The survey is published by the Centers for Disease Control and Prevention which is an agency of the US Department of Health and Human Services which ensures the credibility and reliability of the survey.

The survey answers nine different questions about the participants lifestyle, which were asked between 2011 to 2023 every one to two years whereas each record is supplemented by information about state, income and race. The dataset comprises 104 273 rows. The fixed structure of the tabular data can be seen in the diagram on the left side displaying the column names.

I have chosen this datasource because it includes four different gradations describing the amount of time people are spending on physical activities in each US state:

- no time at all
- muscle-strengthening activities on 2 or more days a week
- at least 150 minutes a week of moderate-intensity aerobic physical activity or 75 minutes a week of vigorous-intensity aerobic physical activity
- at least 300 minutes a week of moderate-intensity aerobic physical activity or 150 minutes a week of vigorous-intensity aerobic activity (or an equivalent combination).

The gradation is important for answering the base question of my work because it enables a comparison with the survey results regarding mental health in each state. The data in the rows containing the total population in a state without other restrictions will be used for data analysis. They are consistent and, apart from data of the state New Jersey, complete. The survey

[1] https://catalog.data.gov/dataset/nutrition-physical-activity-and-obesity-behavioral-risk-factor-surveillance-system
[2] https://data.cdc.gov/api/views/hn4x-zwk7/rows.csv?accessType=DOWNLOAD

was completed until 2023 which makes it up to date. To avoid impact of the Covid 19 crisis I decided to focus on data collected in 2019. Since the analysed survey system is advertised as the *largest continuously conducted health survey system in the world*[3], the datasource can be considered as accurate. However, it must be mentioned that the survey sample size of 400 000 adults per year is only a small fraction of the total population.

Datasource 2: Mental Health

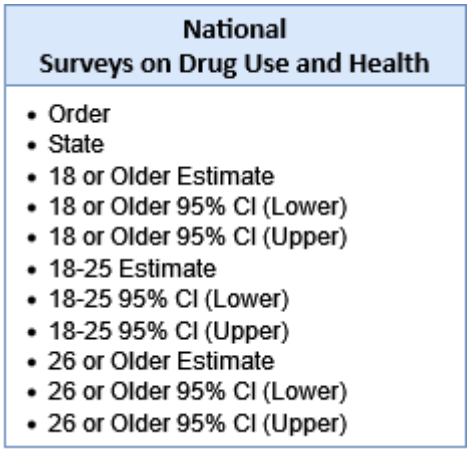

Figure 2: Database schema Mental Health

My second datasource is a package of datasets from the National Surveys on Drug Use and Health (NSDUH) covering the years from 2018 to 2019. The publishing website also labels their data as public files which are shared for public use[4]. The Zip-package[5] can be downloaded and includes 33 csv files, of which only the dataset *Any Mental Illness in the Past Year* is of interest for my work. Since the survey is published by the Substance Abuse and Mental Health Services Administration, which is another agency of the US Department of Health and Human Services, the dataset can be classified as trustworthy. The dataset *Any Mental Illness in the Past Year* is a csv file starting with some unstructured comments followed by tabular data with a fixed scheme showing the percentage of people suffering under mental health problems. The data is separated by state and age as can be seen from the column names in the diagram on the left-hand side. The accuracy of the dataset is documented by including confidence intervals for each data point. Nevertheless, as survey it reflects only information collected from a part of the total population. The given data is complete and consistent. As stated before, the year 2019 will be analysed because the latest published data is from the years 2020 to 2021 which is most probably strongly influenced by the COVID 19 crisis. The dataset is relevant because the mental health situation in each state can be derived from the data and it can be compared to the amount of physical activities from datasource 1.

Data Pipeline

For the data pipeline I decided to use Python and Panda as data analysis library. My pipeline starts with downloading the data. For improved robustness the download is repeated up to three times in case of failure. As the mental health dataset is included in a zip package, it must be unzipped by using the library ZipFile and the required dataset must be extracted. Now, the Panda library is used to transform the files into dataframes for further processing. For reducing the data to the columns and rows needed for answering my question, I used filtering methods from Panda with a whitelist. This strategy allows raising exceptions if unexpectedly important columns are missing. The exceptions are not caught because the program should finish in case of changing input data with an error message. Furthermore, I excluded the data of the state New Jersey as there is no data about the amount of physical activities in this state. The unstructured lines at the beginning of the *Mental Health* dataset had been a particular issue which has been solved by deleting these lines before naming the columns and working with them.

[3] https://www.cdc.gov/brfss/about/brfss_faq.htm
[4] https://www.samhsa.gov/data/data-we-collect/n-sumhss-national-substance-use-and-mental-health-services-survey
[5] https://www.samhsa.gov/data/sites/default/files/reports/rpt32805/2019NSDUHsaeExcelPercents/2019NSDUHsaeExcelPercents/2019NSDUHsaeExcelCSVs.zip

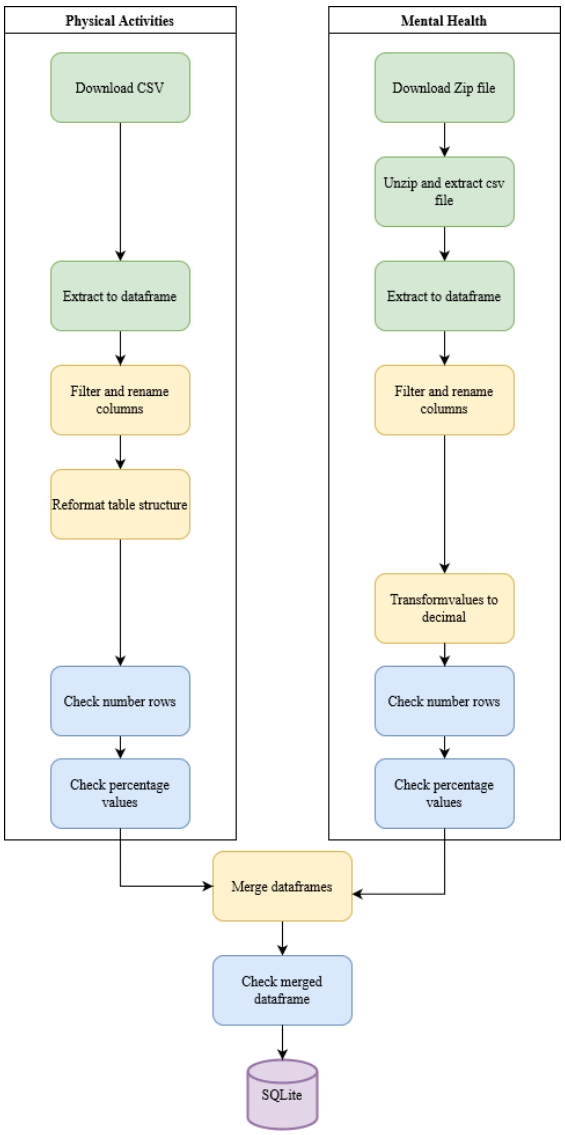

Figure 3: ETL-Pipeline with green extract steps, yellow transform steps, blue quality checks and a purple load step

After filtering the data I had to reformat the dataframe *Physical Activities* as described in the following. At this state, there are four different questions about the amount of physical activity for each state as separate data rows. Therefore, the pipeline merges them to one single row for each state including all four questions. Additionally, the percentage values must be converted from a *String* to a *Float* data type in the dataframe *Mental Health*. Next, the number of rows of both dataframes are checked to be greater or equal the number of states minus one, as the row with *New Jersey* is dropped because of missing data, and if the percentage values are in the right range to secure a good data quality. The two dataframes are merged by an inner join via the common column *State*. The inner join ensures that the state exists in both datasets and removes, for example, the state Guam which is actually just an US foreign territory only included in the datasource about physical activities. After checking the number of rows of the merged dataframe again, the merged dataset is transferred into a SQLite database for further analysing. For further insights every successfully processed step in the pipeline is logged by a log message.

Results and Limitations

The result of my data pipeline is a SQLite database with the joined table *CorrelationPaAndMh*. The resulting dataset provides, for each US state (excluding New Jersey), the percentage distribution of four levels of physical activity and the percentage of people experiencing mental health issues, along with confidence intervals of the mental health data, for the year 2019. Like the datasources the result can be considered as consistent and accurate. I chose SQLite as it is known for its good reading and writing performance combined with Python and to reduce the complexity for accessing the database. Furthermore, compared to other database systems, SQLite offers major advantages of being portable and lightweight[6].

Given the growing importance of mental health in today's society, the resulting data is highly relevant. Identifying a correlation with physical activity could offer a valuable approach for reducing the risk of mental health issues. The most challenging issue is finding and showing the correlation because there are numerous other factors with impact onto metal health that are not covered by the analysed data but could be the particular reason for good or bad mental health. Another limitation of my dataset is its age of five years, which means that the survey results may not reflect the current situation. This is a consequence of deciding against using the COVID 19 influenced years 2020-2021 and the lack of up-to-date data about mental health.

[6] https://www.javatpoint.com/sqlite-advantages-and-disadvantages