Report

The aim of the project is to visualise lifestyle disease such as obesity and hypertension along with age and sex in a particular population. The population studied here is adult population from data gathered by National Health and Nutrition Examination Survey (NHANES) for the year 2005-2006.

The datasets we have used in our analysis include Body Measures, Blood Pressure and Demographic Variables & Sample Weights for the year 2005-2006.

Data Visualisation tasks undertaken:

1. Prevalence of Obesity against age
2. Prevalence of Hypertension against age

The following DAG represents the data visualisation process :

Diagram

Description automatically generated

The data is first explored and cleaned as explained in detail in scripts/readme.md file. The cleaned data is filtered, mutated, and stored to intermediates folder. This intermediate results are used to produce the final output plots and store to results folder. This process is pipelined using snakemake. Git is used to version control and track the workflow.

1. **Prevalence of Obesity against age**

Raw data used : BMX\_D.csv, DEMO\_D.csv

1. The raw data is first cleaned by converting column names to lowercase, merging required columns based on unique identifier, sequence number and is stored in a variable, data0. The gender column is centred around 0, with 0 representing male and 1 representing female. Similarly pregnant women is centred around 0 and 1, with 0 representing negative pregnancy. Pregnant women are removed from visualisation as it can bias our results. Assumed that the participant is not pregnant if the pregnancy status is null. The data is then filtered for adult participants, addressed outliers and missing observations.

1. The cleaned data is written to file obesityData.csv and is then transformed and written to files obesityTransformedData.csv, obeseProbData.csv, obeseMaleProbData.csv, obeseFemaleProbData.csv
2. The transformed data is then used to plot probability of obesity in participants in each age, for male participants only, for female participants only, for both male and female participants.

Each of these plots are saved as png files to results/obesity\_results.

1. **Prevalence of Hypertension against age**

Raw data used : BPX\_D.csv, DEMO\_D.csv

1. The raw data is first cleaned by converting column names to lowercase, merging required columns based on unique identifier, sequence number and this data is stored in a variable, data1. The gender column is centred around 0, with 0 representing male and 1 representing female, filtered for adult participants, addressed missing observations and outliers by removing rows with systolic bp > 300 mmHg, systolic and diastolic bp having odd number values, systolic bp < diastolic bp, systolic bp is 0.
2. The cleaned data is written to file hypertensionData.csv and is then transformed and written to files hypertensionTransformedData.csv, hypertensionProbData.csv, hypertensionMaleProbData.csv, hypertensionFemaleProbData.csv.
3. The transformed data is then used to plot probability of obesity in participants in each age, for male participants only, for female participants only, for both male and female participants.

Each of these plots are saved as png files to results/hypertension\_results

Folder Structure:

Text

Description automatically generated with medium confidence

Running the snakemake pipeline will create 3 folders cleanData, intermediates, results with the corresponding contents inside.

Once this pipelining is done, the pipeline rule, shiny\_web\_app is ran to generate our dynamic visualization in the form of a web application.

To run the application execute shell script executable.sh in terminal as follows inside the project folder :

$./executable.sh

**Output plots for Obesity against Age in adults :**

The x-axis represents age and y-axis represents the corresponding obesity probability.

Chart, line chart

Description automatically generatedA picture containing chart

Description automatically generated

Chart

Description automatically generatedChart

Description automatically generated

Legend :

* Blue – male
* Red ­– female

We can observe, probability of developing obesity increases by age. In both males and females, probability obesity is highest after 50 years, gradually decreases by age 70 years.

**Output plots for Hypertension against Age in adults :**

The x-axis represents age and y-axis represents the corresponding hypertension probability.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated with medium confidence

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Legend :

* Blue – male
* Red ­– female

We can observe, probability of developing hypertension increases by age. In females, probability of developing hypertension reached its peak around 50 years and in males, peak after 50 years.

The dynamic visualization is done using Shiny framework with r script. Below is a snapshot of the application in localhost. It takes 2 inputs, Association between and Gender. Renders the corresponding output plot and data table.

Graphical user interface

Description automatically generated

Refer README.md file for further instructions on getting started and running application.