

DATA MINING FINAL PROJECT

GROUP 2: CLAUDE SHANNON

AUTHORS

- Garri Romzova, Antonio
- Mayol Matos, Sergi
- Medina Perelló, Alejandro
- Palmer Perez, Ruben
- Rodríguez Arguimbau, Alejandro

INDEX

1. Dataset
2. Objective
3. Considerations left behind
4. Conclusions

DATASET

An exploratory introduction

CONTEXT

Medical data around the US regarding health metrics

OBSERVATIONS

Over 400k

counting duplicated data

VARIABLES

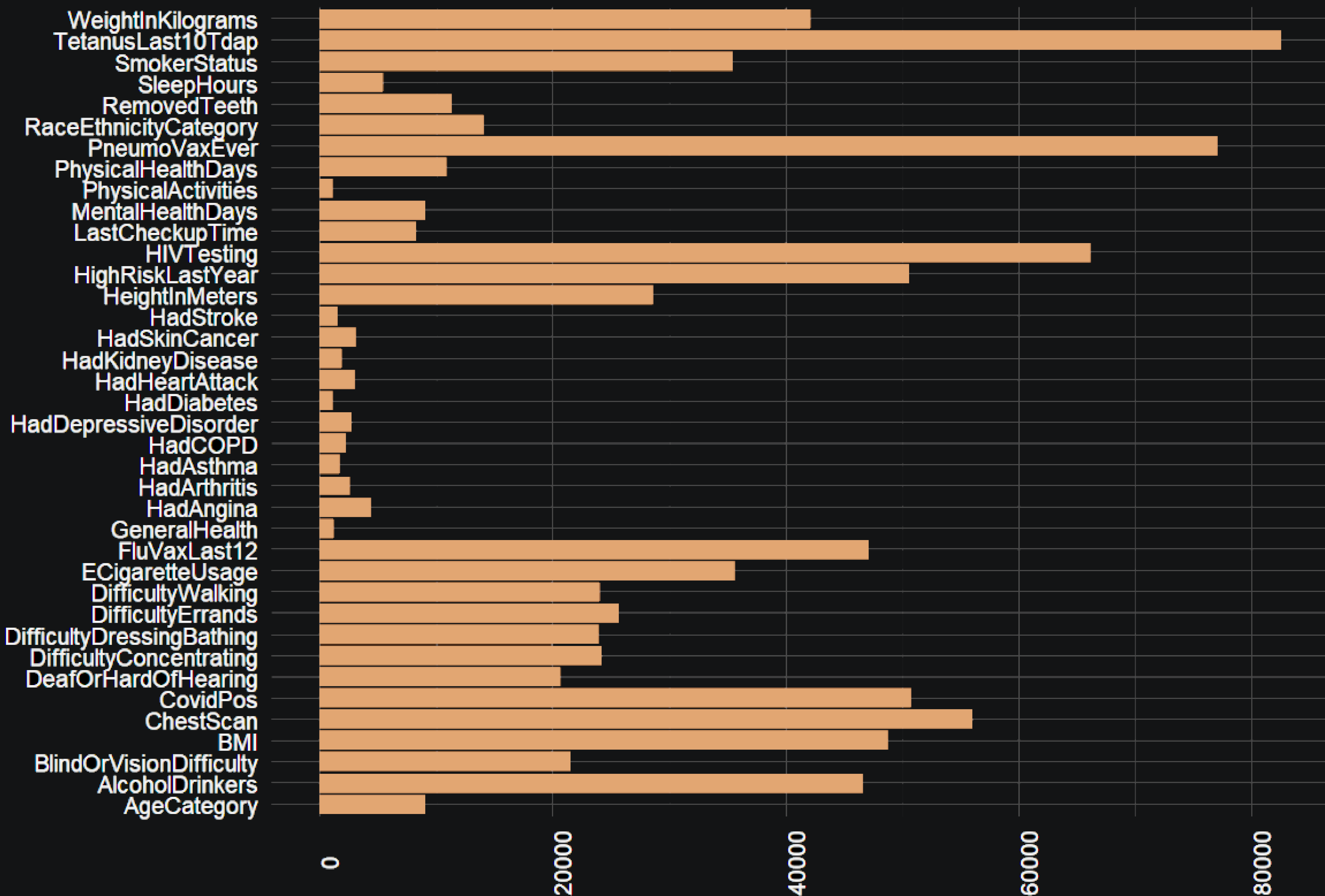
Most of them are categorical or *Boolean*, e.g.

- HadSkinCancer
- HadDepressiveDisorder
- SmokerStatus
- Etc

Only six variables are numeric:

- PhysicalHealthDays
- MentalHealthDays
- SleepHours
- HeightInMeters
- WeightInKilograms
- BMI

MISSING DATA



OBJECTIVE

A common goal

PREDICTIVE MODEL

with HadHeartAttack as our target

PROBLEM

Large amount of variables

SOLUTION

**DATA
MODIFICATIONS**

To use the least amount of variables

OBSERVATION REMOVAL LIMIT

15% of the original dataset

~ 66k

BMI

Remove Height and Weight variables and fill non-numeric values on BMI applying

$$BMI = \frac{Weight}{Height^2}$$

Classification	BMI Score
Underweight	< 18.5
Normal	18.5 - 24.9
Overweight	25.0 - 29.0
Obese	30.0 - 40.0
Extreme Obese	> 40.0

HOW TO FILL MISSING DATA

- Predictive models
- Median
- Mean
- Remove

OUTLIERS

IQR

$$X < (Q1 - 1.5 \times IQR)$$

Or

$$X > (Q3 + 1.5 \times IQR)$$

best for unknown or non-normal distributed data

Z-score

$$\left| \frac{X - \text{mean}(X)}{\text{sd}(X)} \right| > 3$$

best for normal distributed data

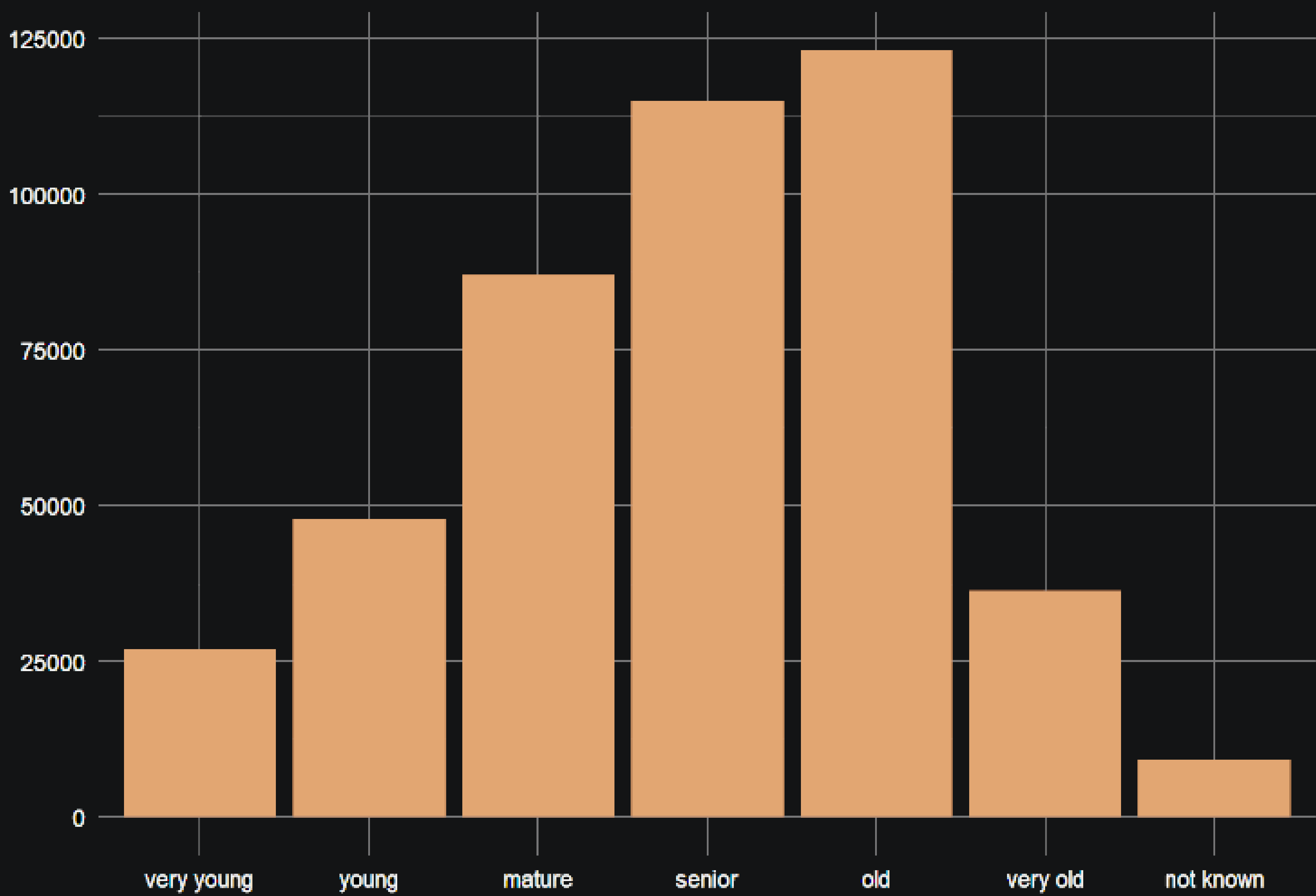
VARIABLE REDUCTION

- Random Forest
- XGBoost

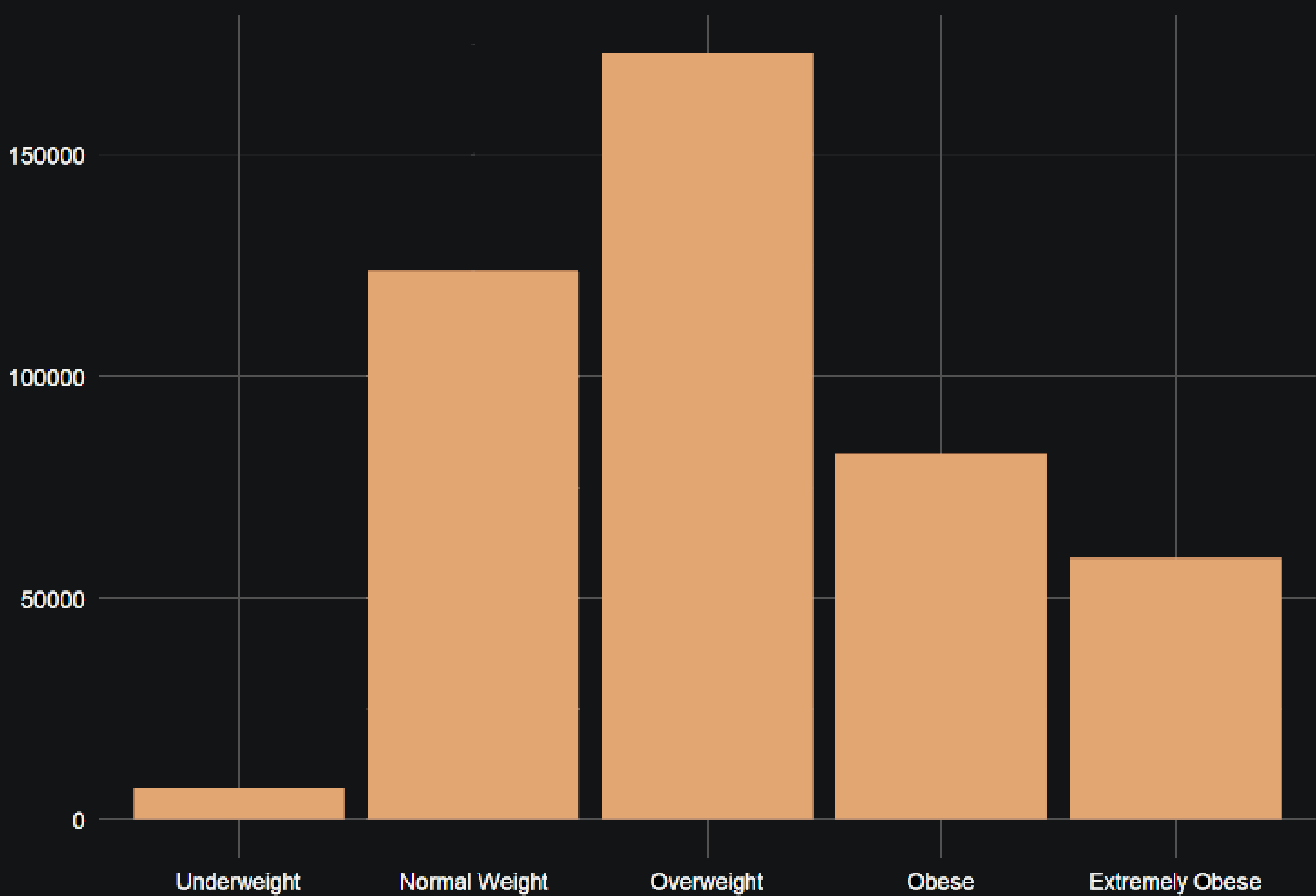
at least explain 95% of the data

SIMPLE QUESTIONS

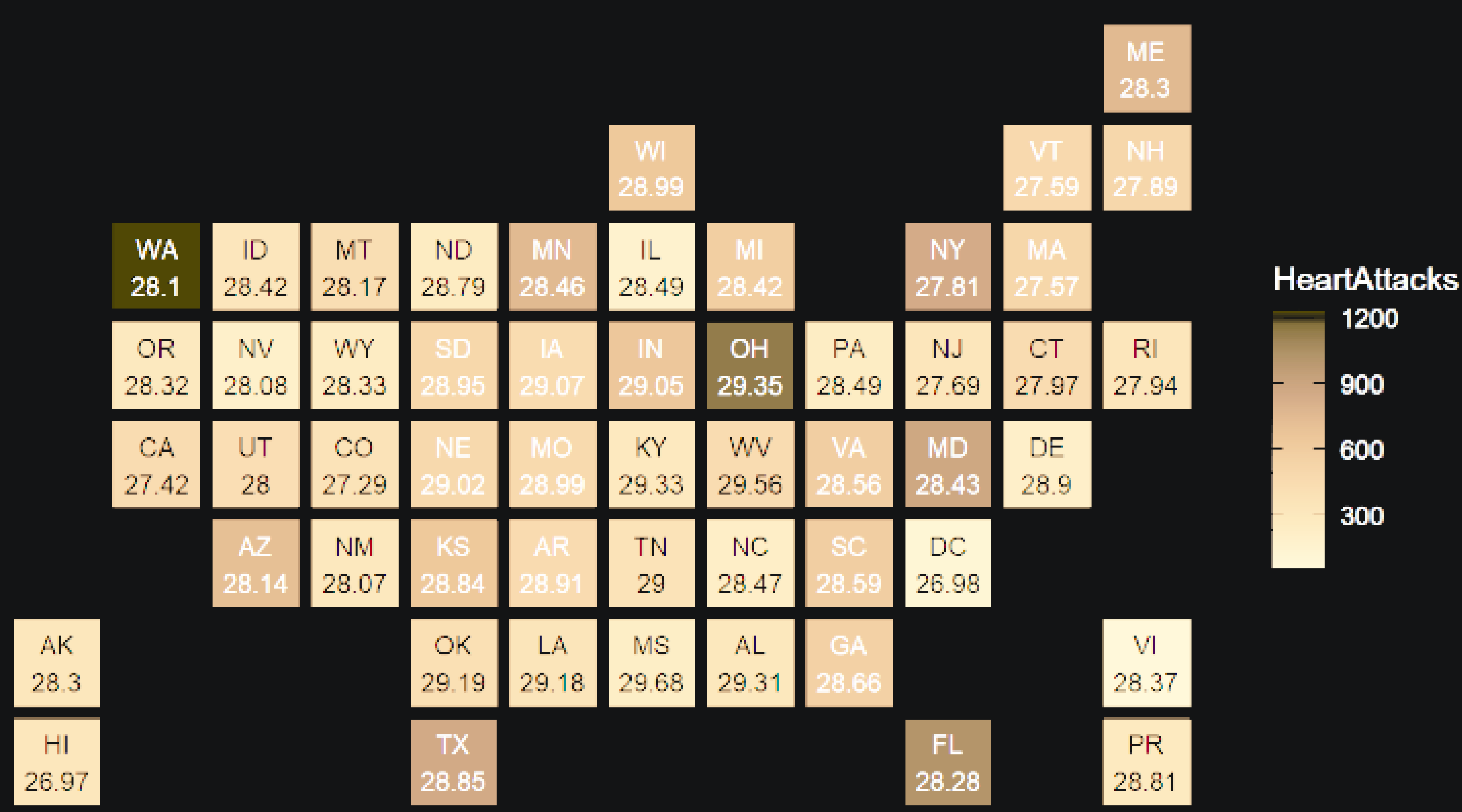
AGE DISTRIBUTION



BMI



HEART ATTACK PER STATE



CONCLUSIONS

