Modeling and profiling people's way of living:
A data mining approach to a health survey.

Luis Gabriel De Alba Rivera
ldealbar@cis.hut.fi

November 26, 2009.

Outline

- Introduction
  - The *Elämä Pelissä* dataset
- Data exploration
  - Variables and queries
- Regression models
  - Methodologies
- Conclusion

Introduction

- *Elämä Pelissä* a Finnish TV show
  - People answered an online survey.
  - 39 questions used to predict the life expectancy.
  - More than half a million Finns answer the survey.
- Duodecim
- Parsing and clean-up
  - 24GB of “raw” data, i.e. 16 million transactions.
  - 65MB database with 457,096 records.

Introduction

- ...
  - Numerical data with informative tags.
  - Three tables in a database to provide fast access by means of SQL.
- Outliers
  - By educated guess.
  - By basic statistics, 3σ from the mean 99.7% of the data.

Data exploration

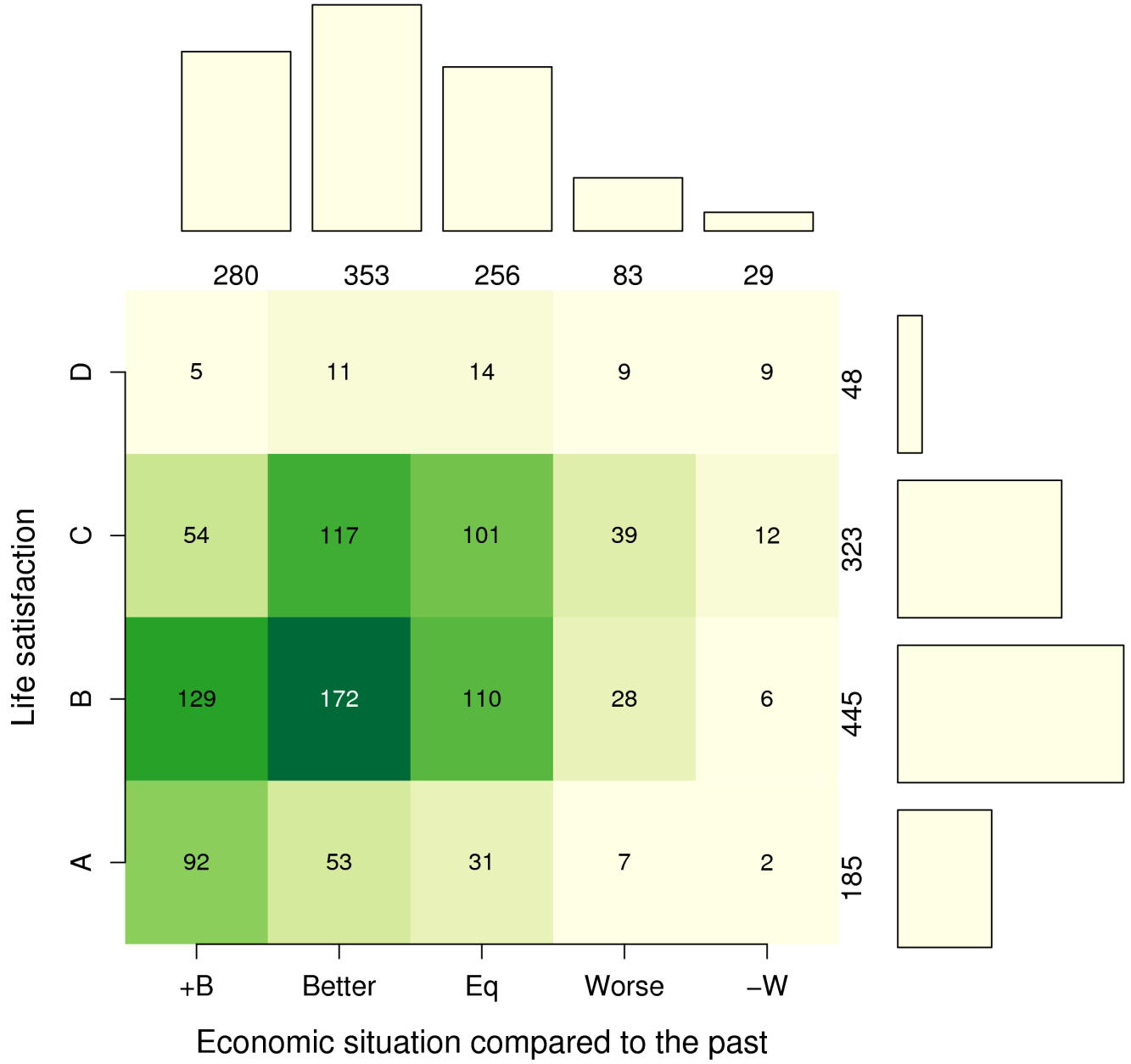
- Data Exploratory Analysis
  - Techniques used to display information.
  - Simple, helpful and easy to interpret.
  - 20 continuous and 29 discrete variables.
  - Different approaches used to display the data depending on the involved variables.
  - Various tools were programed for this purpose.
  - All interesting observations were reported.

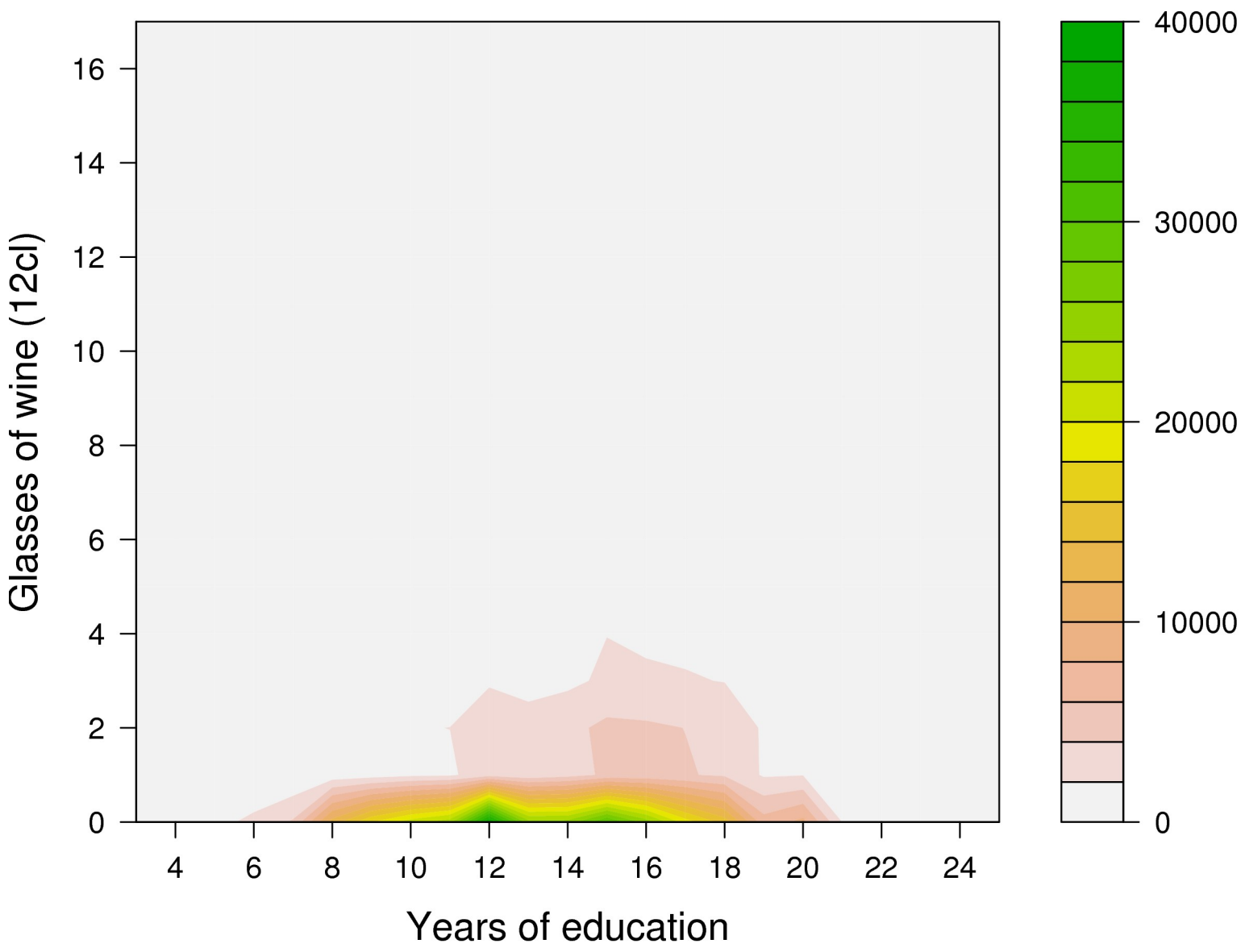
Data exploration

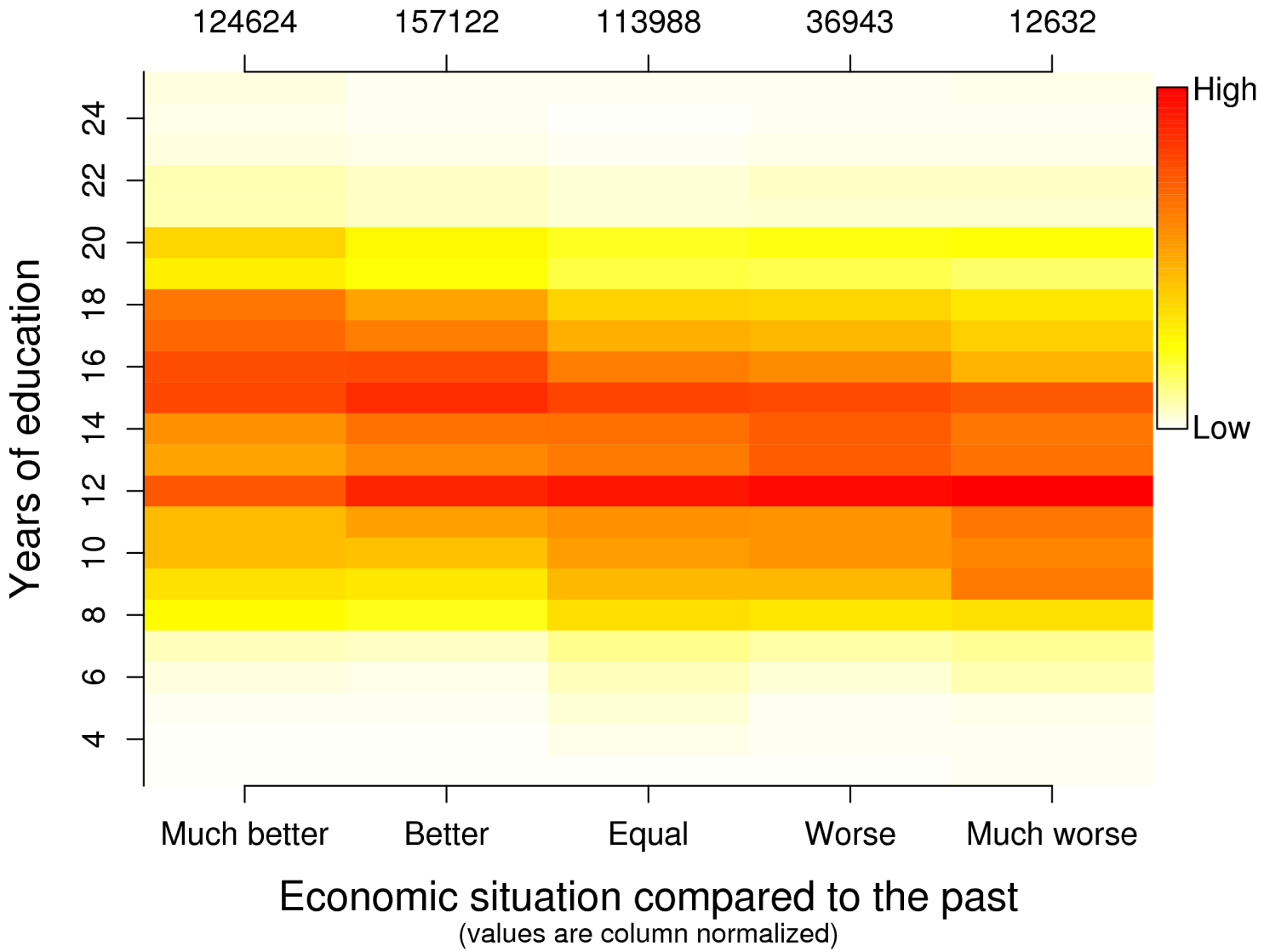
- Single variables
  - Each variable visualized independently.
- Discrete variables
  - Visualized against each other in a matrix-like figure.
  - 406 plots were generated from which 31 interesting facts were found.

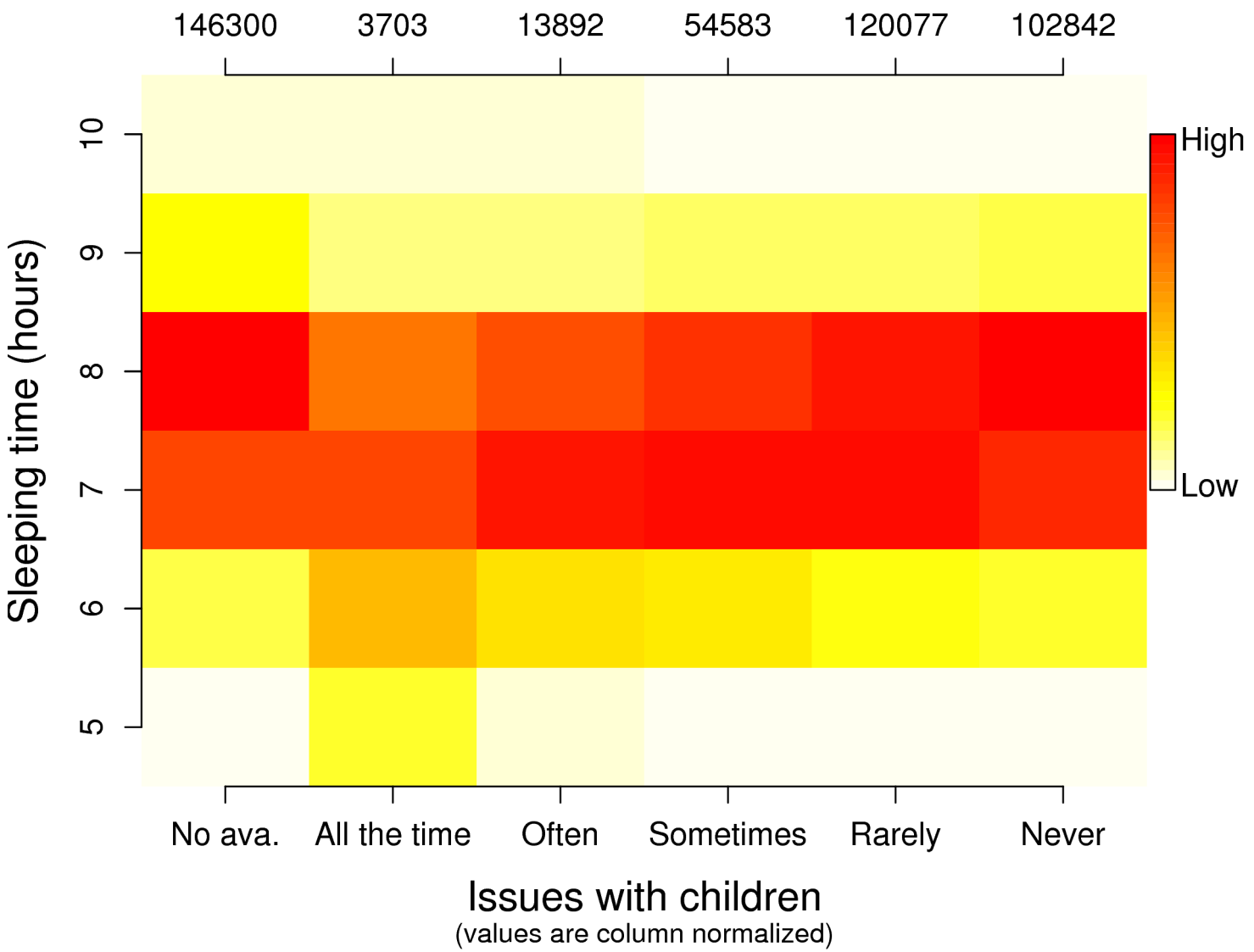
Data exploration

- Continuous variables
  - Visualized against each other in a contour-like figure.
  - 190 plots were generated from which 20 interesting facts were found.
- Mixed variables
  - Continuous and discrete variables in the same plot.
  - Discrete variables on *x-axis* Continuous on *y-axis*.
  - 580 plots with 119 interesting findings.


280
353
256
83
29
D
C
B
A
5
11
14
9
9
54
117
101
39
12
129
172
110
28
6
92
53
31
7
2
48
323
445
185
Life satisfaction
+B
Better
Eq
Worse
–W
Economic situation compared to the past

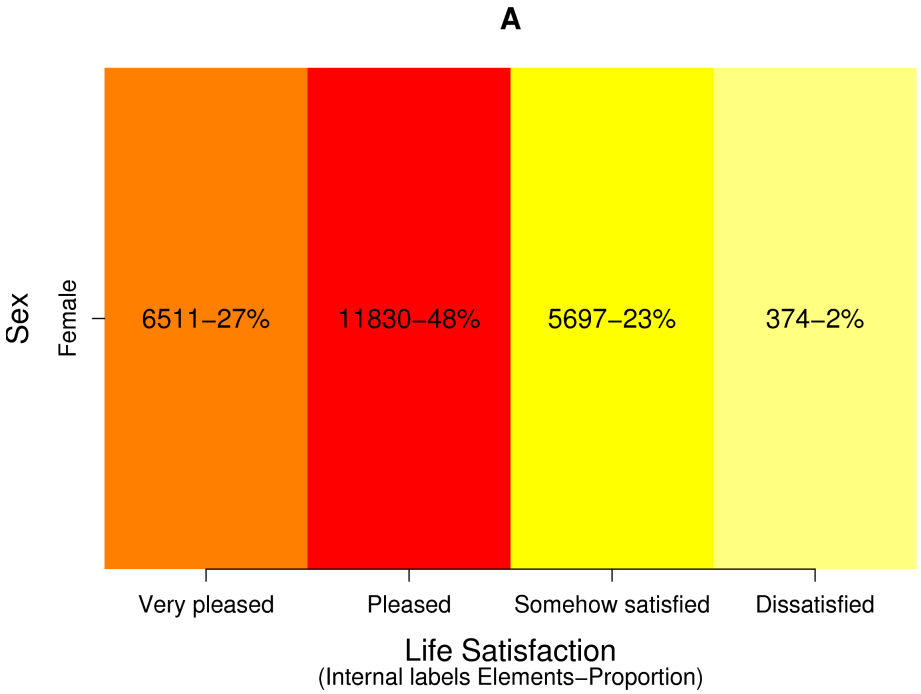

Glasses of wine (12cl)
0
2
4
6
8
10
12
14
16
4
6
8
10
12
14
16
18
20
22
24
Years of education
40000
30000
20000
10000
0

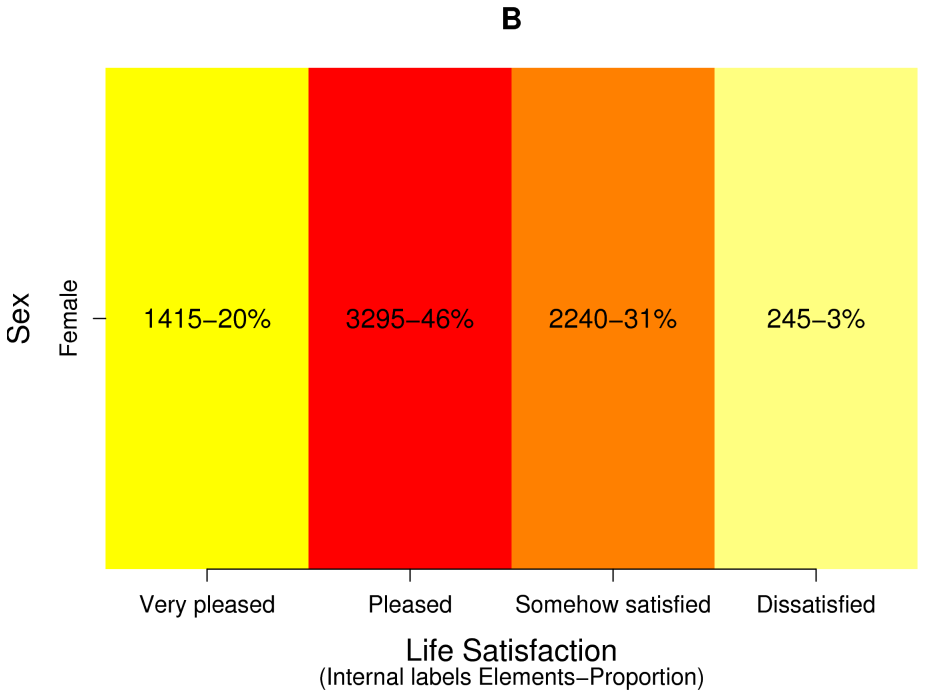

124624
157122
113988
36943
12632
Years of education
4
6
8
10
12
14
16
18
20
22
24
High
Low
Much better
Better
Equal
Worse
Much worse
Economic situation compared to the past
(values are column normalized)

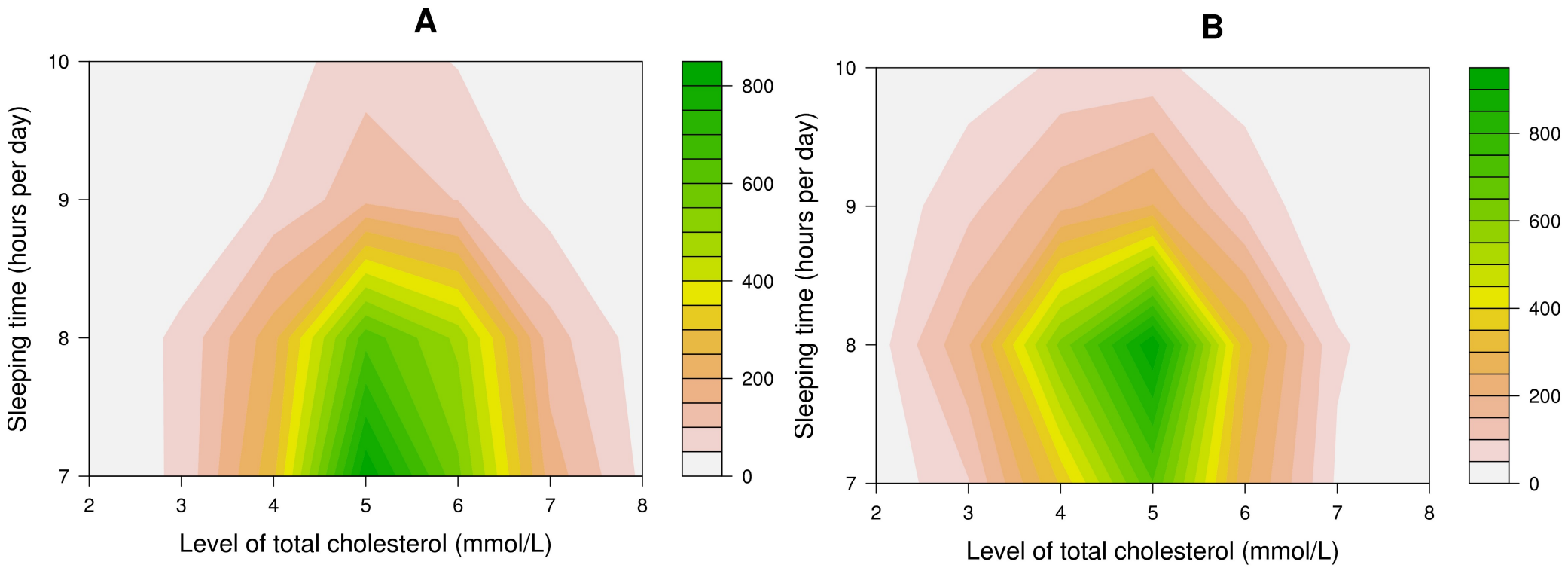

146300
3703
13892
54583
120077
102842
Sleeping time (hours)
10
9
8
7
6
5
High
Low
No ava.
All the time
Often
Sometimes
Rarely
Never
Issues with children
(values are column normalized)

Data exploration

- Specific consults
  - Grouping and filtering data according to different variables and characteristics.
  - It is possible to select ranges for Continuous variables and categories for Discrete variables.
  - Interesting results,


A
Sex
Female
6511–27%
11830–48%
5697–23%
374–2%
Very pleased
Pleased
Somehow satisfied
Dissatisfied
Life Satisfaction
(Internal labels Elements–Proportion)


B
Sex
Female
1415–20%
3295–46%
2240–31%
245–3%
Very pleased
Pleased
Somehow satisfied
Dissatisfied
Life Satisfaction
(Internal labels Elements–Proportion)


A
B
Sleeping time (hours per day)
Level of total cholesterol (mmol/L)
10
9
8
7
2
3
4
5
6
7
8
0
200
400
600
800

Regression models

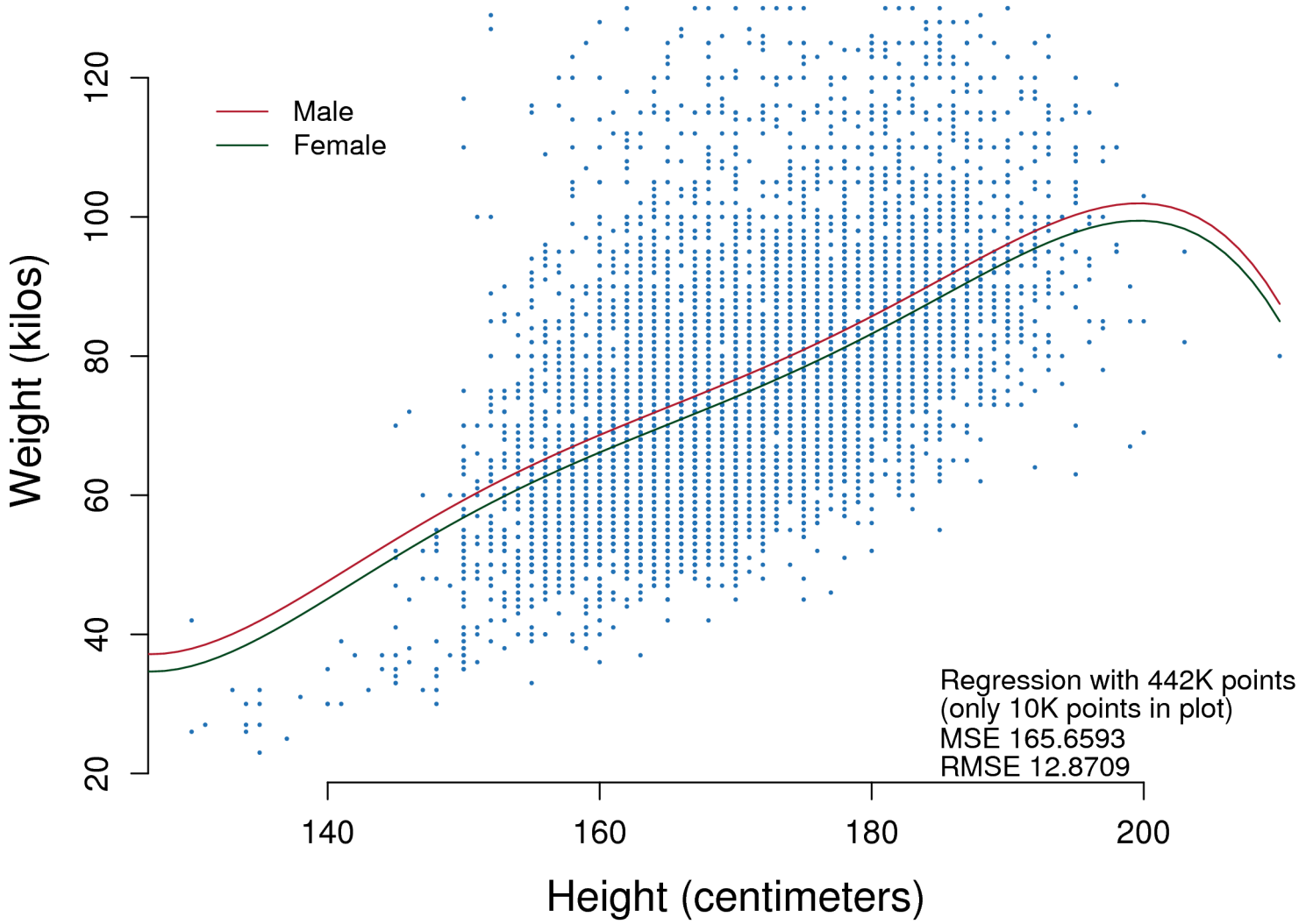
- Missing variables
  - In the dataset there are variables with lots of missing values.
  - Specially those related to *difficult* questions.
  - Thesis only focused on continuous variables, but not all discrete variables were discarded. E.g.
    - Sex: Male(-1) Female(+1)
    - Stress: Yes(3) More than average(2) Somehow(1) No(0)

Regression models

- Modeling
  - 90% training set, 10% testing set.
  - 10-fold cross-validation repeated 100 times to estimate the coefficients.
- Single predictor
  - Single variable linear models.
  - Polynomial models with *natural* divisors.
    - Sex and Age were forced to be in the model.

Regression models

- ...
  - Results
    - Height as function of Sex improvement of 51%
    - Weight as function of Height improvement of 31%
  - Results polynomial
    - Sex and Age played important role.
      - Alcohol, Education, Income, etc.
    - Different degrees of polynomials.
      - Weight improved 5 points to 36% using Sex, Age and Height.
      - Number of cigarettes showed an improvement of 71%.


Male
Female
Weight (kilos)
20
40
60
80
100
120
140
160
180
200
Height (centimeters)
Regression with 442K points
(only 10K points in plot)
MSE 165.6593
RMSE 12.8709

Regression models

- Feature selection
  - Forward Stepwise
    - Add the input that decreases the most the RSS.
    - For each variable the best set of features was selected.
  - Results
    - Improvements in some variables up to 67%.
    - On most missing variables:
      - 15% Blood pressure systole, 4% Cholesterol total.
    - Few variables selected, 2-7 of 15 available.

Regression models

- Feature selection
  - SISAL
    - Sequential Input Selection Algorithm (Tikka & Hollmén)
    - Backward elimination using the median and the width; the least significant feature is removed.
  - Results
    - Better results than forward selection.
    - However, more features selected 9-12 of 15.
    - Most missing variables.
      - 19% Blood pressure systole, 7% Cholesterol total.

Regression models

- Regression improvements
  - Artificial Neural Networks
  - Principal Components
    - Using the variables selected by previous algorithms:
      - ALL and TOP
    - Focus on the most missing variables
    - Results
      - Minute or non improvements compared to linear models.
      - Results when using TOP is similar to those when using ALL.
      - The burden of training and testing the ANN does not worth the gain.

Conclusion

- Understanding the data
- Data exploratory analysis
    - Initial set of results and information.
- Modeling of the variables
    - Focus on missing variables.
    - Different approaches.
- Future work
    - Discrete variables.
    - Clusters, etc.

Questions?

Live in peace, plant potatoes and dream.

5% get drunk more than once per week.

Not wearing a seat belt:70% males, 30% females

Heavy smokers, 20+ cigarettes per day, reduce their expectancy of living by 5 – 10 years.

People with not stress at all are between 55 – 65 years old.