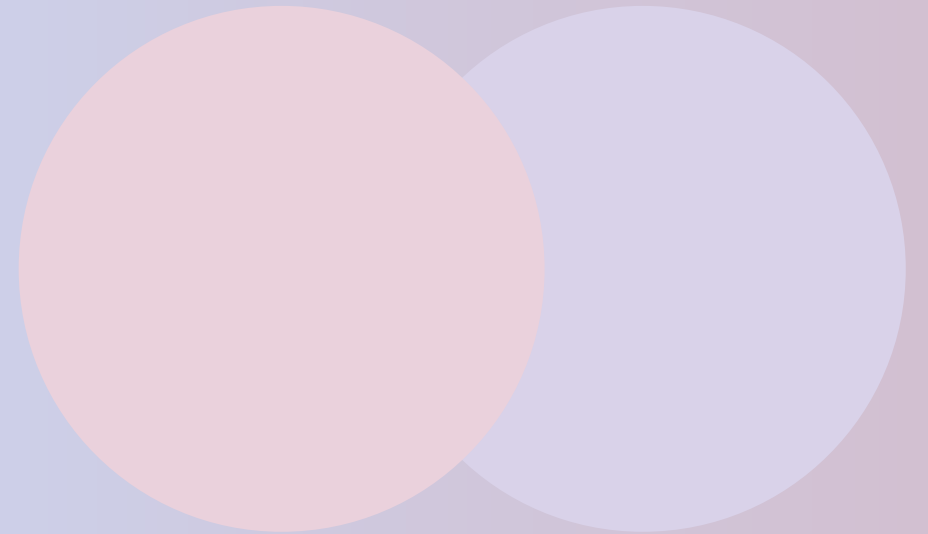


Project in Data Mining

Using association rule mining to identify combinations of factors contributing to high and low suicide rates



Project goals



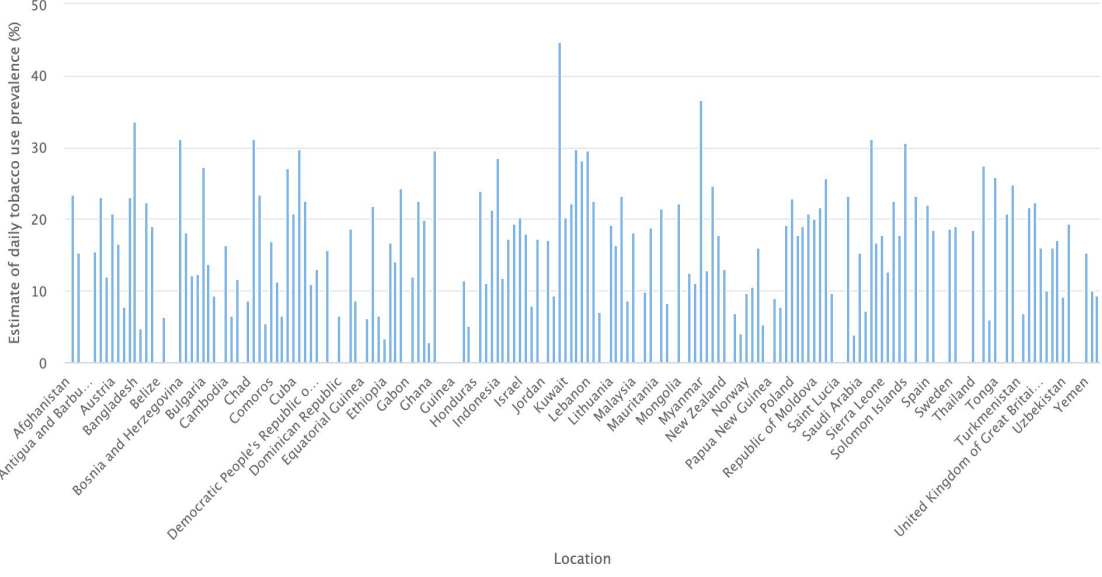
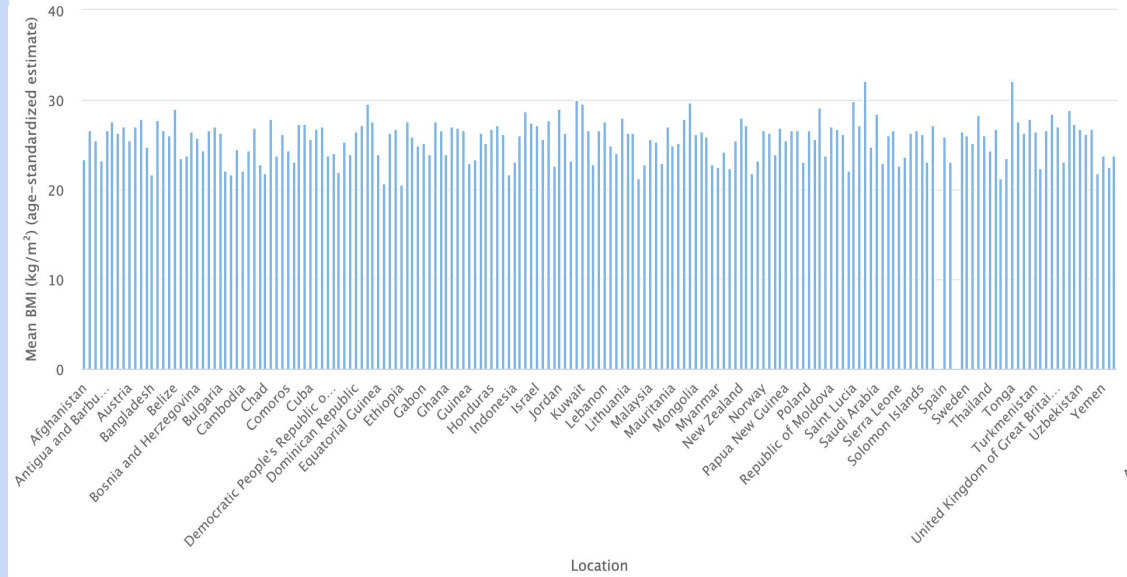
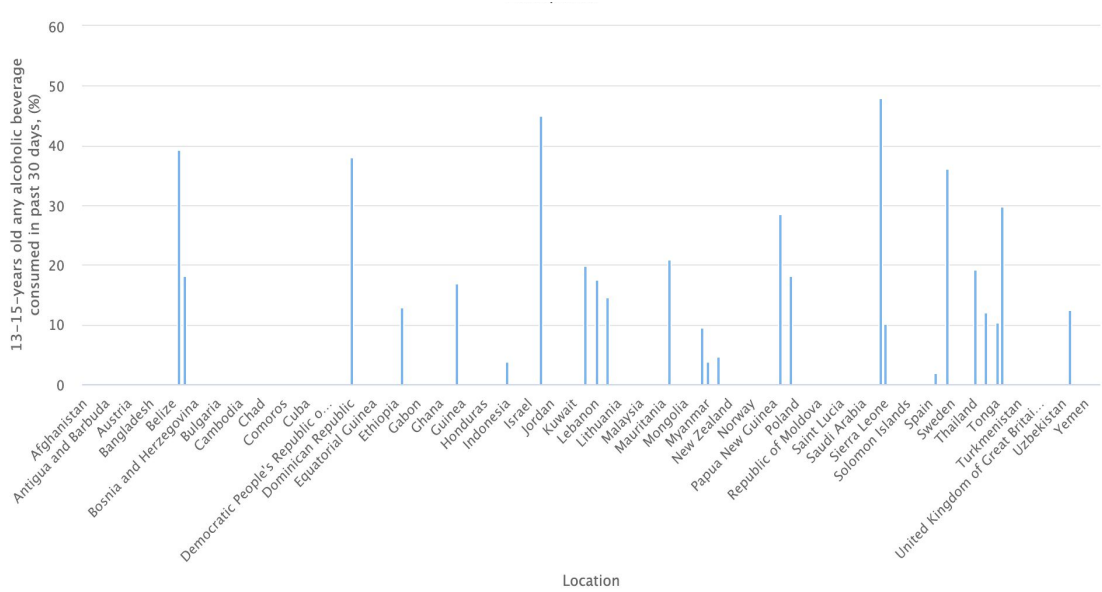
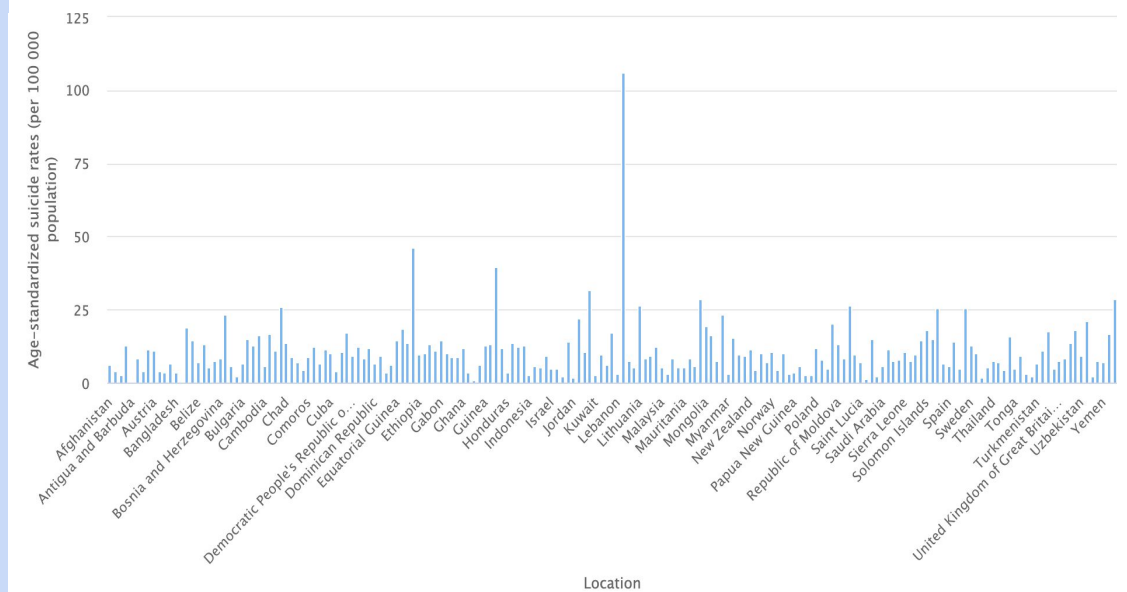
Can we identify any
combination of factors
contributing to
high/low suicide rates?

Data

from the World Health Organization

- **ID:** Countries (nominal)
- Age-standardized suicide rates (per 100 population) (ratio)
- Mean BMI (age standardized estimate) (interval)
- UV radiation (ratio)
- Mental hospital admissions (per 100,000) (ratio)
- Government expenditures on mental health as a percentage of total government expenditures on health (%) (ratio)
- Non-age-standardized estimates of daily tobacco use (tobacco smoking and cigarette smoking) (ratio)
- Raised blood pressure (SBP \geq 140 or DPR \geq 90) (crude estimate) (ratio)
- Alcohol, total per capita (15+ years) consumption (in liters of pure alcohol) (ratio)
- 13-15-years old any alcoholic beverage consumed in past 30 days (%) (ratio)
- Proportion of women aged 20-24 who were married or in a union at by age 15 (%) (ratio)
- Out-of-pocket expenditure as percentage of current health expenditure (CHE) (%) (ratio)
- Adolescent birth rate (per 1000 women aged 15-19 years) (ratio)
- Psychiatrists working in mental health sector (per 100,000) (ratio)
- Social workers working in mental health sector (per 100,000) (ratio)
- Nurses working in mental health sector (per 100,000) (ratio)

Data



Preprocessing

What decisions had to be made?

- How do we handle the variety of different ranges of years present in the WHO data?
- How do we convert the numerical data to binomial, so that we can do association rule mining?

Chose to focus on years 2015-2017 since many countries had data present for those years. A mean of the attribute values for those years were taken.

In the cases where data for *some* of these years were missing, the mean of the existing years in that range was taken.

When *no* data existed for those years, the existing year was used.

Divide all attributes into four ranges: Low, Medium-Low, Medium-High, High.

The ranges were obtained by splitting the range for each attribute values into quartiles.

Preprocessing

Creating a CSV file using pandas in Python

Step 1: Sorting out what data we wanted and took the average in the range of years 2015-2017

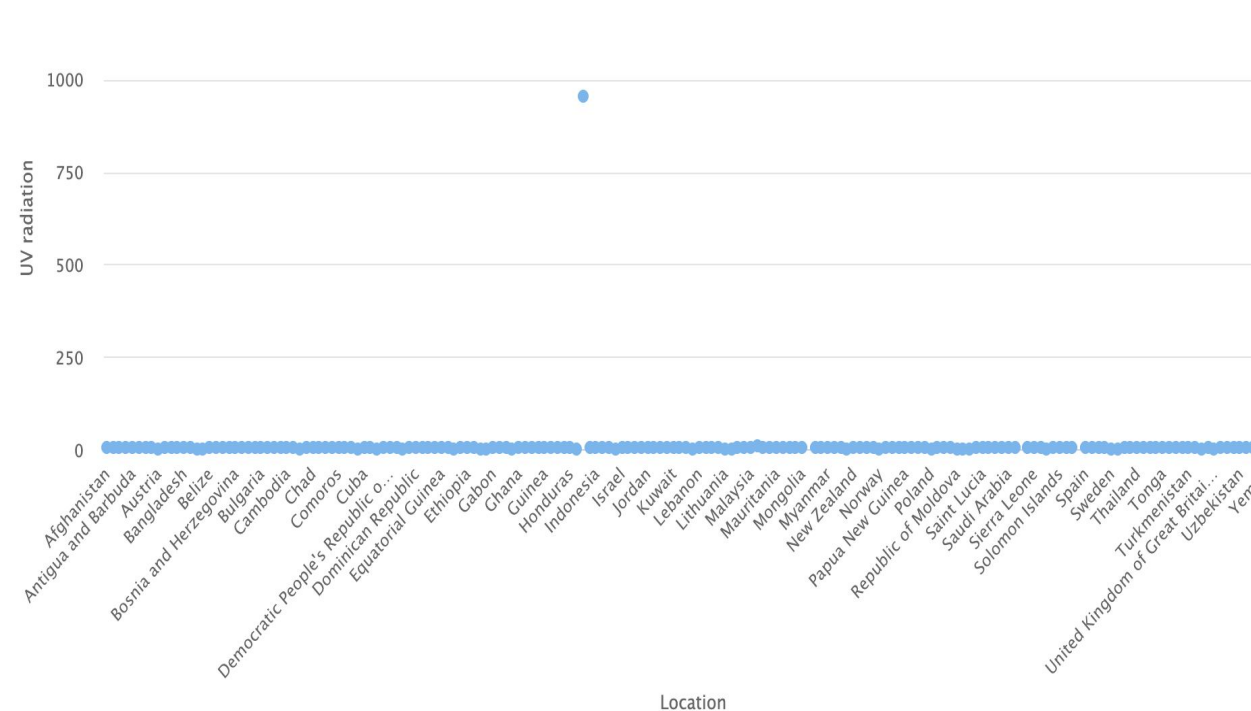
Attribute	Years present in WHO's database	Years chosen to use	Percentage of missing values [%]
Suicide Rates	2000-2019	2015-2017	0.00
UV radiation	2004	2004	2.18
Raised Blood Pressure	1975-2015	2015	1.09
Social Workers	2015-2017	2015-2017	47.54
Psychiatrists	2015-2017	2015-2017	24.04
Women Married	2010-2017 (one value for those years)	2010-2017	33.88
Out of Pocket	2000-2018	2015-2017	2.73
Nurses	2013-2017	2015-2017	33.88
Tobacco	2007-2018 (every other year)	2016	21.31
Mental Hospital	2014	2014	37.16
BMI	1975-2016	2015-2016	1.09
Government Expenditures	2011	2011	60.66
Alcohol Total	2016-2018 (one value for those years)	2016-2018	0.55
Young Birth	2000-2019	2015-2017	18.57
Alcohol Youth	2003-2017	2015-2017	86.34

Preprocessing

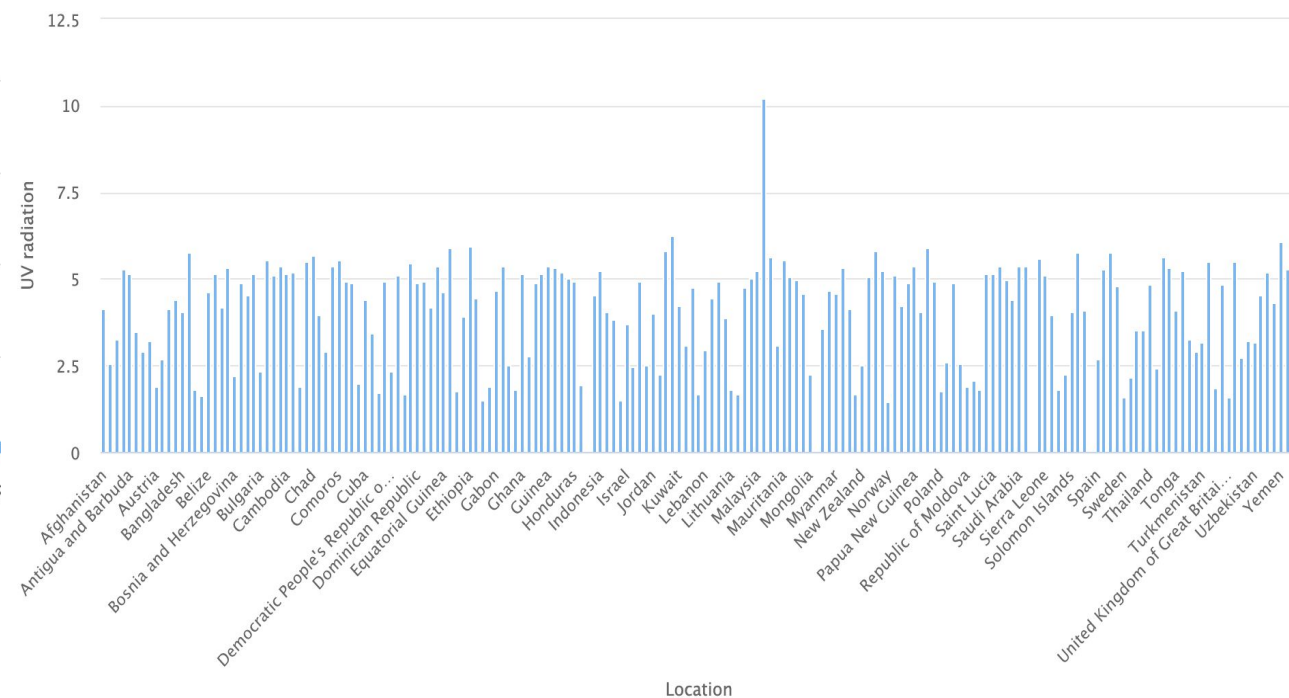
Detect and remove outliers

Step 2: Visualized the data using scatterplots to identify and remove outliers.

Before



After



Preprocessing

Creating a CSV file using pandas in Python

Step 3: Merged the 15 different data sets into one csv-file

Countries	Attribute 1	Attribute 2	...	Attribute 15

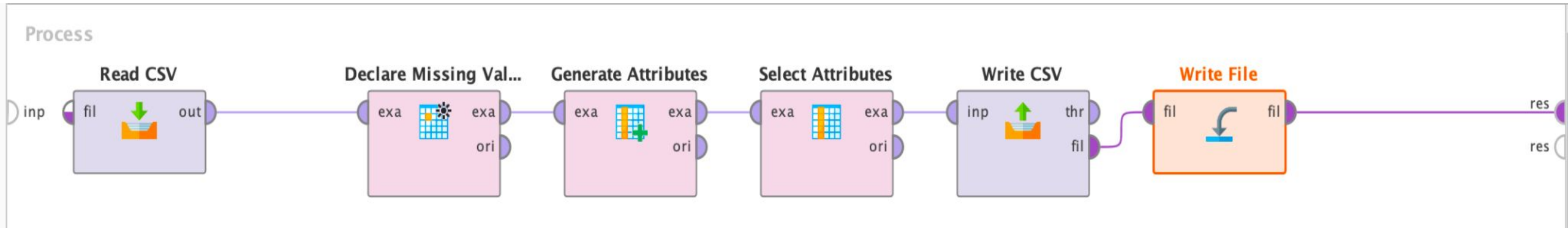
Preprocessing

Converting numerical data to binomial using RapidMiner

Step 4: Declare missing values

Step 5: Generate the new binomial (true or false) attributes:
Low \leq Medium-Low \leq Medium-High \leq High

Step 6: Convert data into a CSV-file (to do Association rule mining in Python, due to a bug in RapidMiner)



Preprocessing

Converting numerical data to binomial using RapidMiner

Attributes	Low	Medium-Low	Medium-High	High
Suicide Rates	≤ 5.372	(5.372, 8.72]	(8.72, 13.143]	(13.143, 105.963]
BMI	≤ 23.8	(23.8, 26.15]	(26.15, 27.0]	(27.0, 32.15]
UV Radiation	≤ 2.826	(2.826, 4.552]	(4.552, 5.176]	(5.176, 10.224]
Nurses working in mental health sector	≤ 0.87	(0.87, 4.37]	(4.37, 14.83]	(14.83, 150.3]
Adolescent birth rate	≤ 10.85	(10.85, 32.9]	(32.9, 67.5]	(67.5, 180.0]
Alcohol - young age	≤ 10.4	(10.4, 17.5]	(17.5, 28.5]	(28.5, 47.9]
Alcohol	≤ 2.3	(2.3, 5.7]	(5.7, 9.2]	(9.2, 20.5]
Tobacco	≤ 10.4	(10.4, 17.25]	(17.25, 22.25]	(22.25, 44.8]
Government expenditures on mental health	≤ 1.28	(1.28, 2.87]	(2.87, 5.02]	(5.02, 12.91]
Psychiatrists working in mental health sector	≤ 0.225	(0.225, 1.39]	(1.39, 6.82]	(6.82, 48.04]
Mental hospital admissions	≤ 3.495	(3.495, 32.65]	(32.65, 141.5]	(141.5, 985.4]
Out of pocket health-expenditures	≤ 16.947	(16.947, 29.595]	(29.595, 44.435]	(44.435, 82.21]
Raised Blood Pressure	≤ 18.6	(18.6, 20.9]	(20.9, 23.8]	(23.8, 37.6]
Social workers working in mental health sector	≤ 0.081	(0.081, 0.45]	(0.45, 1.41]	(1.41, 145.4]
Women married at a young age	≤ 1.0	(1.0, 4.0]	(4.0, 9.0]	(9.0, 30.0]

Preprocessing

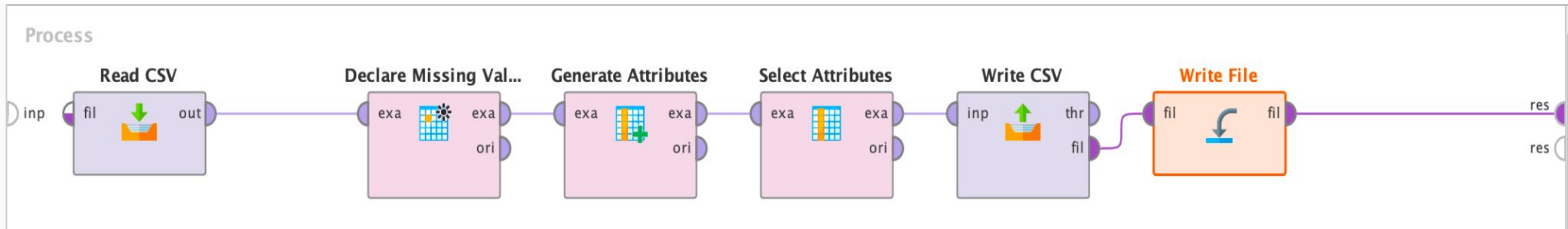
Converting numerical data to binomial using RapidMiner

Step 4: Declare missing values

Step 5: Generate the new binomial (true or false) attributes:

Low \leq Medium-Low \leq Medium-High \leq High

Step 6: Convert data into a CSV-file (to do Association rule mining in Python, due to a bug in RapidMiner)



Preprocessing

Converting numerical data to binomial using RapidMiner

- **What about correlation issues?**

Since every country only can exist in max one of the four attributes for each original attribute, no correlation issues will occur.

- **What about scaling issues?**

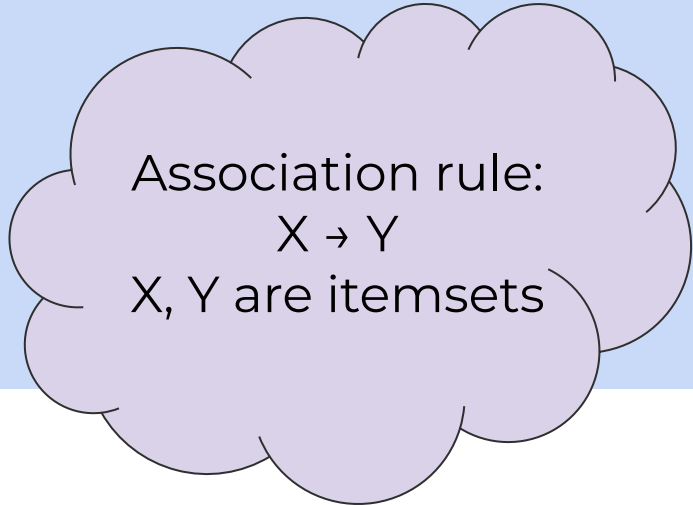
Since we chose to divide the attributes using quartiles, there was no need for scaling since the quartiles assured that every attribute had the same amount of records in them. Scaling or normalizing the data before calculating the quartiles would give the same result.

Association Rule Mining

using mlxtend in Python

- What is mlxtend?
- Generation of association rules:
 - Fp-growth
 - Create association rules
- How to identify the most important relationships: *interest measures for association rules*
 - Support
 - Confidence
 - Lift
 - Conviction

Measures of Interest



Association rule:
 $X \rightarrow Y$
 X, Y are itemsets

- **Support** indicates how frequently an itemset occurs in the dataset. A *higher* value of support indicates that the rule is more common.
- **Confidence** measures how often the consequent (Y) appears in transactions that contain the antecedent (X). A *higher* value of confidence indicates a more reliable association rule.
- **Lift** measures how likely the items in a rule appear together compared to by chance. It is a measure of the importance of the rule.
A lift value >1 is desirable.

Lift > 1 means it is more likely
Lift < 1 means it is less likely
Lift $= 1$ the itemsets in the rule are statistically independent (not desirable!)
- A high **conviction** value implies that the consequent (Y) is highly dependent on the antecedent (X). If conviction $= 1$ the items in the rule are unrelated.



We want to find association rules with high support, high confidence, lift > 1 , and conviction > 1

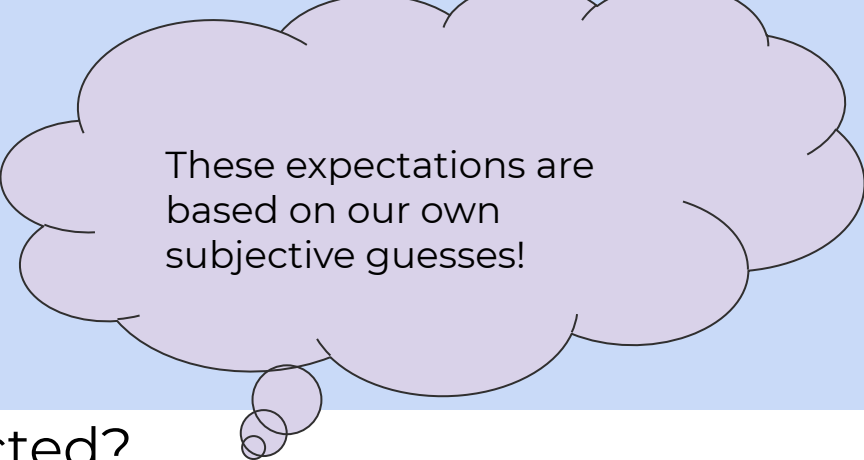
Result

Generated association rules

Minimum support = 0.03
Minimum confidence = 0.75

Antecedents	Consequents	Support	Confidence	Lift	Conviction
{'BMI High', 'Alcohol Low'}	{'Suicide Rates Low'}	0.060	0.846	3.366	4.866
{'Women married at a young age Low', 'Alcohol Low'}	{'Suicide Rates Low'}	0.049	1.000	3.978	inf
{'BMI High', 'Alcohol Low', 'UV Radiation Medium-Low'}	{'Suicide Rates Low'}	0.049	1.000	3.978	inf
{'Raised Blood Pressure Low', 'BMI High', 'UV Radiation Medium-Low'}	{'Suicide Rates Low'}	0.038	0.875	3.481	5.989
{'BMI High', 'Social workers working in mental health sector Medium-Low'}	{'Suicide Rates Low'}	0.038	0.875	3.481	5.989
{'Raised Blood Pressure Low', 'BMI High', 'Alcohol Low'}	{'Suicide Rates Low'}	0.038	0.778	3.094	3.369
{'Women married at a young age Low', 'Alcohol Low', 'UV Radiation Medium-Low'}	{'Suicide Rates Low'}	0.033	1.000	3.978	inf
{'Adolescent birth rate Medium-Low', 'Out of pocket health-expenditures Medium-High', 'UV Radiation Low'}	{'Suicide Rates High'}	0.033	0.750	2.984	2.985
{'Psychiatrists working in mental health sector Medium-High', 'Out of pocket health-expenditures Low'}	{'Suicide Rates Low'}	0.033	0.750	2.984	2.985
{'Adolescent birth rate Low', 'Alcohol Low'}	{'Suicide Rates Low'}	0.033	0.750	2.984	2.985
{'Women married at a young age Low', 'Mental hospital admissions Medium-High'}	{'Suicide Rates Low'}	0.033	0.750	2.984	2.995
{'BMI Medium-High', 'Adolescent birth rate Medium-Low', 'Psychiatrists working in mental health sector High'}	{'Suicide Rates High'}	0.033	0.750	2.984	2.995
{'Psychiatrists working in mental health sector High', 'Adolescent birth rate Medium-Low', 'BMI Medium-High', 'UV Radiation Low'}	{'Suicide Rates High'}	0.033	0.750	2.984	2.995
{'Psychiatrists working in mental health sector High', 'Adolescent birth rate Medium-Low', 'BMI Medium-High', 'Raised Blood Pressure High'}	{'Suicide Rates High'}	0.033	0.750	2.984	2.995
{'Adolescent birth rate Medium-Low', 'UV Radiation Low', 'Raised Blood Pressure High', 'Psychiatrists working in mental health sector High', 'BMI Medium-High'}	{'Suicide Rates High'}	0.033	0.750	2.984	2.995
{'Mental hospital admissions High', 'BMI Medium-High', 'Raised Blood Pressure High', 'UV Radiation Low'}	{'Suicide Rates High'}	0.033	0.750	2.984	2.995

Analysis



These expectations are based on our own subjective guesses!

→ Are the results similar to what was expected?

List of attributes expected to contribute to *Low Suicide Rate*:

- Medium BMI
- High amount of UV radiation
- High amount of nurses/psychiatrists/social workers working in mental health sector
- Low adolescent birth rate
- Low consumption of alcohol (both in total and at a young age)
- Low tobacco use
- High government expenditures on health
- Low mental hospital admissions
- Low out-of-pocket expenditures
- Low raised blood pressure
- Low women married at a young age

...and vice versa for *high suicide rates*

The anticipated factors or combination of factors associated with a low suicide rate was expected to correlate with the welfare of the country and the overall health of the population of a country. We associate a strong welfare system and a healthy population with, for example, Medium BMI, Low consumption of alcohol, Low tobacco use and Low raised blood pressure, Low out-of-pocket expenditures on mental health and so on.

Some of these expected attributes reoccured in several association rules, for example 'low consumption of alcohol' and 'low amount of women married at a young age' exist in several "Low suicide rate" outcomes.

Analysis

→ Are there any discussion-worthy individual cases?

- We were somewhat surprised by the fact that high BMI often contributed to low suicide rates, as we anticipated that a medium value of the BMI would contribute to low suicide rates. This may be explained by the fact that we happen to live in countries where starvation and hunger is most often not an issue.
- We expected that countries that invest much in mental health (such as high number of Psychiatrists working in mental health sector) would result in lower suicide rates, but the results shows that this was the opposite: High number of 'Psychiatrists working in mental health sector' is contributing to a high suicide rate. We concluded that this found can be due to that countries with high suicide rates have more people with mental illness and thus higher demand of psychiatrists working in mental health sector.

Antecedents	Consequents	Support	Confidence	Lift	Conviction
{'BMI High', 'Alcohol Low'}	{'Suicide Rates Low'}	0.060	0.846	3.366	4.866
{'Women married at a young age Low', 'Alcohol Low'}	{'Suicide Rates Low'}	0.049	1.000	3.978	inf
{'BMI High', 'Alcohol Low', 'UV Radiation Medium-Low'}	{'Suicide Rates Low'}	0.049	1.000	3.978	inf

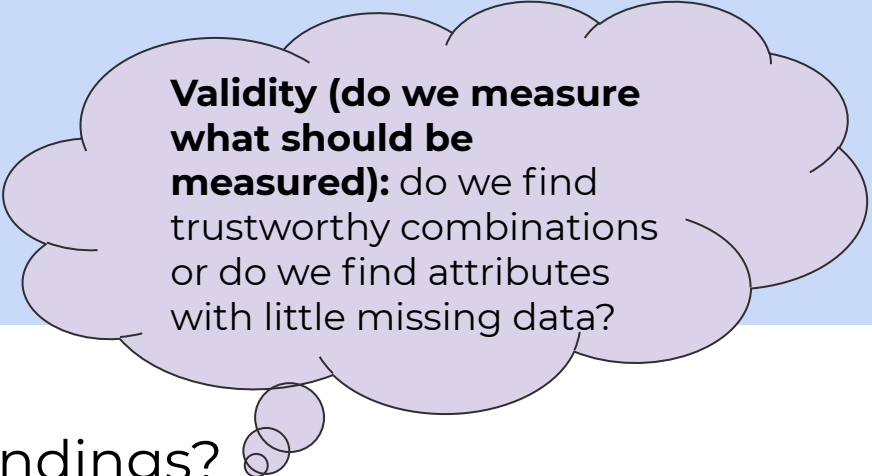
Analysis

→ Are there any discussion-worthy individual cases?

- We were somewhat surprised by the fact that high BMI often contributed to low suicide rates, as we anticipated that a medium value of the BMI would contribute to low suicide rates. This may be explained by the fact that we happen to live in countries where starvation and hunger is most often not an issue.
- We expected that countries that invest much in mental health (such as high number of Psychiatrists working in mental health sector) would result in lower suicide rates, but the results shows that this was the opposite: High number of 'Psychiatrists working in mental health sector' is contributing to a high suicide rate. We concluded that this found can be due to that countries with high suicide rates have more people with mental illness and thus higher demand of psychiatrists working in mental health sector.

Antecedents	Consequents	Support	Confidence	Lift	Conviction
{'BMI Medium-High', 'Adolescent birth rate Medium-Low', 'Psychiatrists working in mental health sector High'}	{'Suicide Rates High'}	0.033	0.750	2.984	2.995
{'Psychiatrists working in mental health sector High', 'Adolescent birth rate Medium-Low', 'BMI Medium-High', 'UV Radiation Low'}	{'Suicide Rates High'}	0.033	0.750	2.984	2.995
{'Psychiatrists working in mental health sector High', 'Adolescent birth rate Medium-Low', 'BMI Medium-High', 'Raised Blood Pressure High'}	{'Suicide Rates High'}	0.033	0.750	2.984	2.995
{'Adolescent birth rate Medium-Low', 'UV Radiation Low', 'Raised Blood Pressure High', 'Psychiatrists working in mental health sector High', 'BMI Medium-High'}	{'Suicide Rates High'}	0.033	0.750	2.984	2.995
{'Adolescent birth rate Medium-Low', 'Out of pocket health-expenditures Medium-High', 'UV Radiation Low'}	{'Suicide Rates High'}	0.033	0.750	2.984	2.985

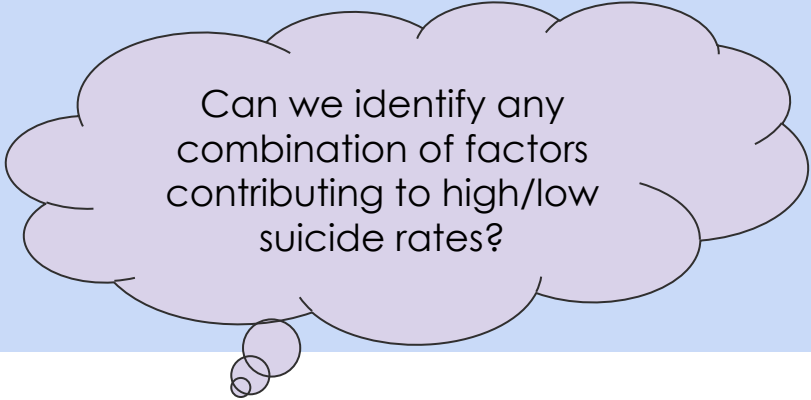
Analysis



Validity (do we measure what should be measured): do we find trustworthy combinations or do we find attributes with little missing data?

- Are there any dataset properties affecting your findings?
- **Impact of missing data:** Most of the attributes present in the final result includes attributes that had small amounts of missing data. This have to be taken into consideration when analyzing the result, as the fact that some attributes do not appear in the final association rules might not necessarily imply that they do not have a great effect on the suicide rates, but can be due to large amount of missing values in those datasets. Hence, the *validity* of this investigation is questionable.

Conclusions



Can we identify any combination of factors contributing to high/low suicide rates?

→ Was the problem solved?

We were able to obtain a result that can be used to answer the research question, however the degree of trustworthiness of the result can be discussed, as the validity is questionable.

→ How can the results be used?

It can be used as a guideline to find areas of focus in where to put effort in the work to prevent suicides. If more data were accessible (and less missing data), the results would be more trustworthy.

→ Have any new hypotheses been generated?

'Psychiatrist working in mental health sector High' is correlating with high suicide rate
→ Action to an already existing societal problem?

Potential Improvements

- Improved initial data exploration: Search for equivalent data from other websites (not only WHO)
 - Manage to do clustering, as initially planned. With our data it was difficult to perform a fair clustering (not desirable to impute missing values obtained by using the values of that attribute from other countries)
- Different handling of the years present in the data
 - Use some other way of handling the different ranges of years present in the data. Instead of taking the average we could have used the median instead. Median is not as sensitive to largely differing values as the mean.
 - Perform the association rule mining on several time periods, which would make it possible to examine the reliability of the model.
- Check for outliers *before* taking the average over the chosen years (this was unfortunately not possible in this project due to lack of time)

**Thanks for
listening!**

