INSEAD
The Business School for the World®

Data Science for Business: Final Project

Group 3 - EB:

Cherrie Liu, Curtis Graham, Daniel Hew

Joe Sawma, Kian Jer Koh, Marco Stoppini

The Business Problem

Rotterdam, Netherlands
One of the biggest polluters in the world per capita

Rotterdam Government Health Department
Survey in 2012

14K+ participants
890+ variables:

- *Physical information,*
- *Medical history (both mental and physical)*
- *Residential information*

Key Questions

How does public health vary across the city?

What are the important drivers of health?

How to implement policy measures to improve the city’s environment and health?

The Business Solution Process

1. Dataset cleaning;
   - Reduced variables from 898 to 397 by removing variables that would not add any value to the analysis
   - Combined rare categories
   - Fixed missing values
2. Apply Dimensionality Reduction methods to reduce the number of input variables in the dataset, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. E.g., by using Principal Component Analysis.
3. Use the results of this analysis to derive insights and make business decisions, e.g., determine how public health varies across the city of Rotterdam, determine which areas are the most heavily polluted

The Dataset

AGOJB201	Surveynummer	koppelsleutel	MIREB201	GemCode	POSCODN
2012	1	1220021347	6	599	3011
2012	1	1220021354	6	599	3013
2012	1	1220021355	6	599	3011
2012	1	1220021358	6	599	3011
2012	1	1220021367	6	599	3011
2012	1	1220021368	6	599	3011
2012	1	1220021369	6	599	3011
2012	1	1220021370	6	599	3011
2012	1	1220021374	6	599	3011
2012	1	1220021377	6	599	3011
2012	1	1220021379	6	599	3011
2012	1	1220021391	6	599	3011
2012	1	1220021393	6	599	3011
2012	1	1220021400	6	599	3011
2012	1	1220021402	6	599	3011
2012	1	1220021405	6	599	3011
2012	1	1220021406	6	599	3011
2012	1	1220021408	6	599	3011
2012	1	1220021420	6	599	3011
2012	1	1220021423	6	599	3011
2012	1	1220021429	6	599	3011
2012	1	1220021435	6	599	3011

Variable	Label	Measurement	Missings
AGOJB201	Year of research	Nominal	9999
Survey number	Survey number	Scale	
koppelsleutel	Security number	Scale	
MIREB201	GGD region	Nominal	99
GemCode	Municipality code [415]	Scale	
POSCODN	Post code (four numbers)	Scale	

KAPITAAL5	KAPITAAL6	MMSCB203	KAPITAAL7	KAPITAAL8	MMVWB201	SUNINB201	SUNINB202	SUNINB203	GROEN1
3	2	6	4	4	2	1	2	1	6
2	1	5	4	1	2	2	1	3	7
2	2	4	4	1	2	1	2	1	7
3	2	4	3	0	1	2	2	2	8
3	3	1	4	3	2	2	1	2	7
2	2	6	4	1	2	2	2	1	7
2	1	3	2	2	1	1	2	2	5
2	3	2	4	3	2	1	2	2	4
3	2	3	2	2	1	1	2	1	5
2	2	4	4	6	2	1	2	1	7
3	4	1	2	1	2	2	1	1	6
3	3	5	4	2	2	2	2	1	5
2	2	1	4	28	1	1	2	1	9
2	3	3	4	3	1	2	2	1	7
2	2	3	4	2	2	1	2	3	8

Variable	Label	Measurement	Missings
KAPITAAL5	How many adults from the neighborhood do you know by sight	Nominal	9
KAPITAAL6	How many of your friends live in your neighborhood	Nominal	9
MMSCB203	Contacts with neighbors	Nominal	9
KAPITAAL7	Participation in block parties or meetings in the neighborhood	Nominal	9
KAPITAAL8	Number of years living in the neighborhood	Nominal	99
MMVWB201	Voluntary work	Nominal	9
SUNINB201	I give money to charities	Nominal	9
SUNINB202	I sometimes do something for my neighbors	Nominal	9
SUNINB203	I bring glass to the bottle bank	Nominal	9

Dataset Cleaning

	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1														
2	WEEGF	WEEGF	WEEGF	WEEGFACTOR4_OUD	geboor	geslach	MIDGV	AGLFS2	LFT010	AGLFS2	Lftkl_er	Herkom	Genera	inkkwir	MIE
3	13,228	13,194	13.23	1,281,401,187	1989	1	1	23	22	1	1	0	0	1	
4	6,053	6,039	6.02	1,015,850,597	1993	2	1	19	19	1	1	0	0	5	
5	9,592	9,572	9,599	1,281,401,187	1993	1	1	19	19	1	1	0	0	5	
6	15,624	15,692	15,686	1,015,850,597	1992	2	1	20	20	1	1	5	2	1	
7	9,688	9,661	9,651	1,015,850,597	1990	2	1	22	22	1	1	0	0	1	
8	13,025	13,019	13,066	1,281,401,187	1989	1	1	23	23	1	1	0	0	1	
9	12,326	12,499	12,446	1,015,850,597	1989	2	1	23	23	1	1	6	1	1	
10	16,338	16,598	16,652	1,281,401,187	1989	1	1	23	22	1	1	6	2	1	
11	13,228	13,194	13.23	1,281,401,187	1990	1	1	22	22	1	1	0	0	1	
12	6,861	6.82	6,832	1,015,850,597	1988	2	1	24	23	1	1	0	0	3	
13	7,462	7,488	7,518	1,015,850,597	1991	2	1	21	21	1	1	0	0	4	
14	8,426	8,433	8,321	1,015,850,597	1991	2	1	21	20	1	1	0	0	5	
15	14,948	14,952	14,896	1,015,850,597	1989	2	1	23	22	1	1	5	1	-1	
16	9,688	9,661	9,651	1,015,850,597	1990	2	1	22	22	1	1	0	0	1	
17	7,739	7,734	7,787	1,015,850,597	1989	2	1	23	22	1	1	0	0	3	
18	9,592	9,572	9,599	1,281,401,187	1990	1	1	22	22	1	1	0	0	5	
19	6,931	6,953	6,975	1,015,850,597	1988	2	1	24	23	1	1	0	0	5	
20	9,891	9,835	9,815	1,015,850,597	1992	2	1	20	20	1	1	0	0	1	
21	9,891	9,835	9,815	1,015,850,597	1990	2	1	22	21	1	1	0	0	1	
22	18,285	18,311	18,311	1,281,401,187	1989	1	1	23	22	1	1	5	1	1	
23	6,861	6.82	6,832	1,015,850,597	1989	2	1	23	23	1	1	0	0	3	
24	7,064	6,994	6,996	1,015,850,597	1991	2	1	21	21	1	1	0	0	3	

Removal of 450+ variables:

Irrelevant Data (e.g. leisure times to do jobs assigned MET values)

Descriptive values (e.g. open ended questions)

Repetition

Unclear description about variables (e.g. Other Neighbourhoods)

Delete variables with more than 5% data missing

Deleted 676 rows with more than 40 blank values

Dimensionality Reduction

1. Confirm the data is metric
2. Scale the data: applied standardization
3. Check correlations: find correlations of the factors
4. Choose number of factors: used Principal Component Analysis
- Reduced variables from 890+ to 34 factors: including alcohol, residential living, mental health etc.
5. Interpret the factors: use only a few, non-overlapping original attributes
6. Save factor scores

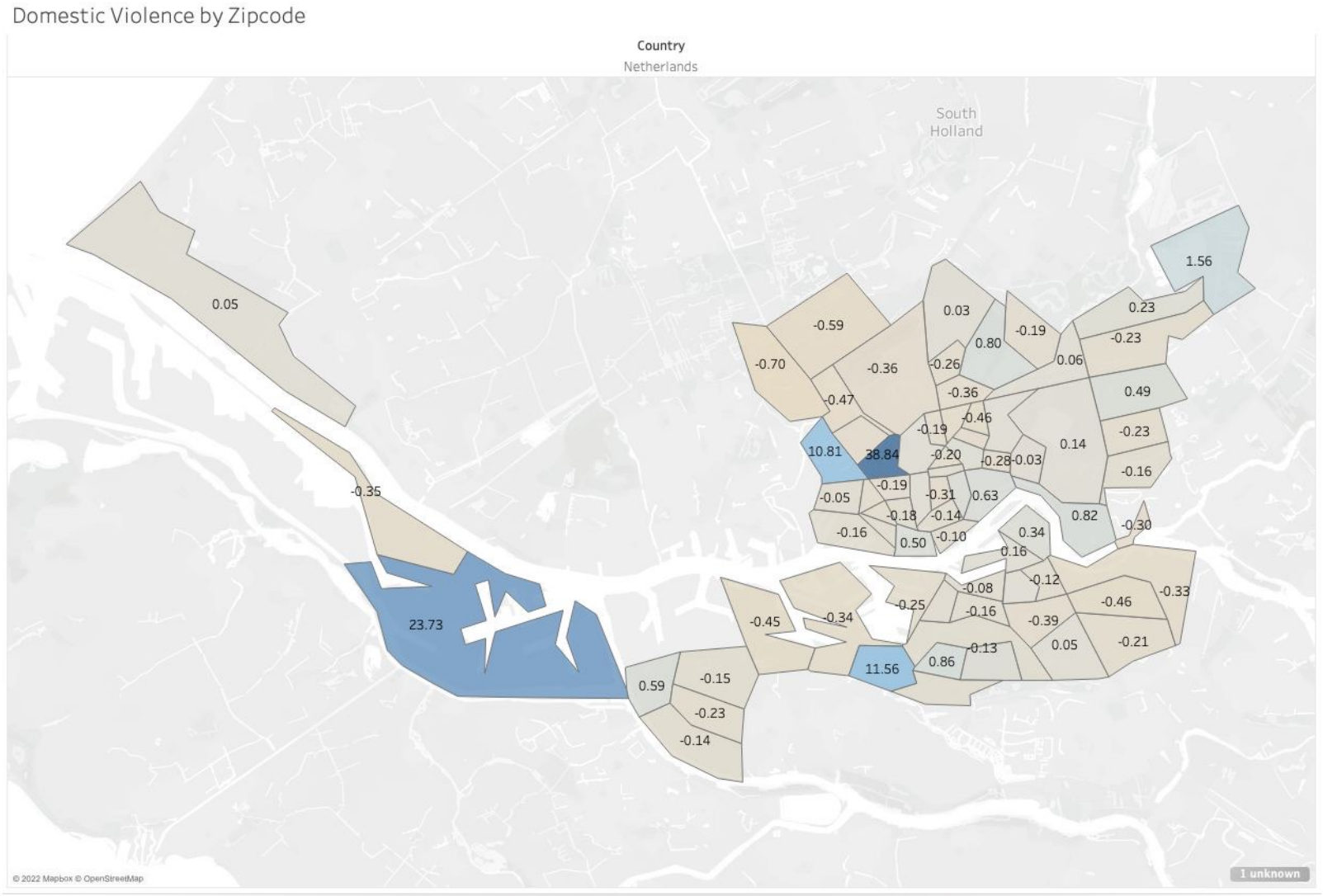
Factor Interpretation

Applied factor rotations to transform the estimated factors into new ones that satisfy that, while capturing the same information

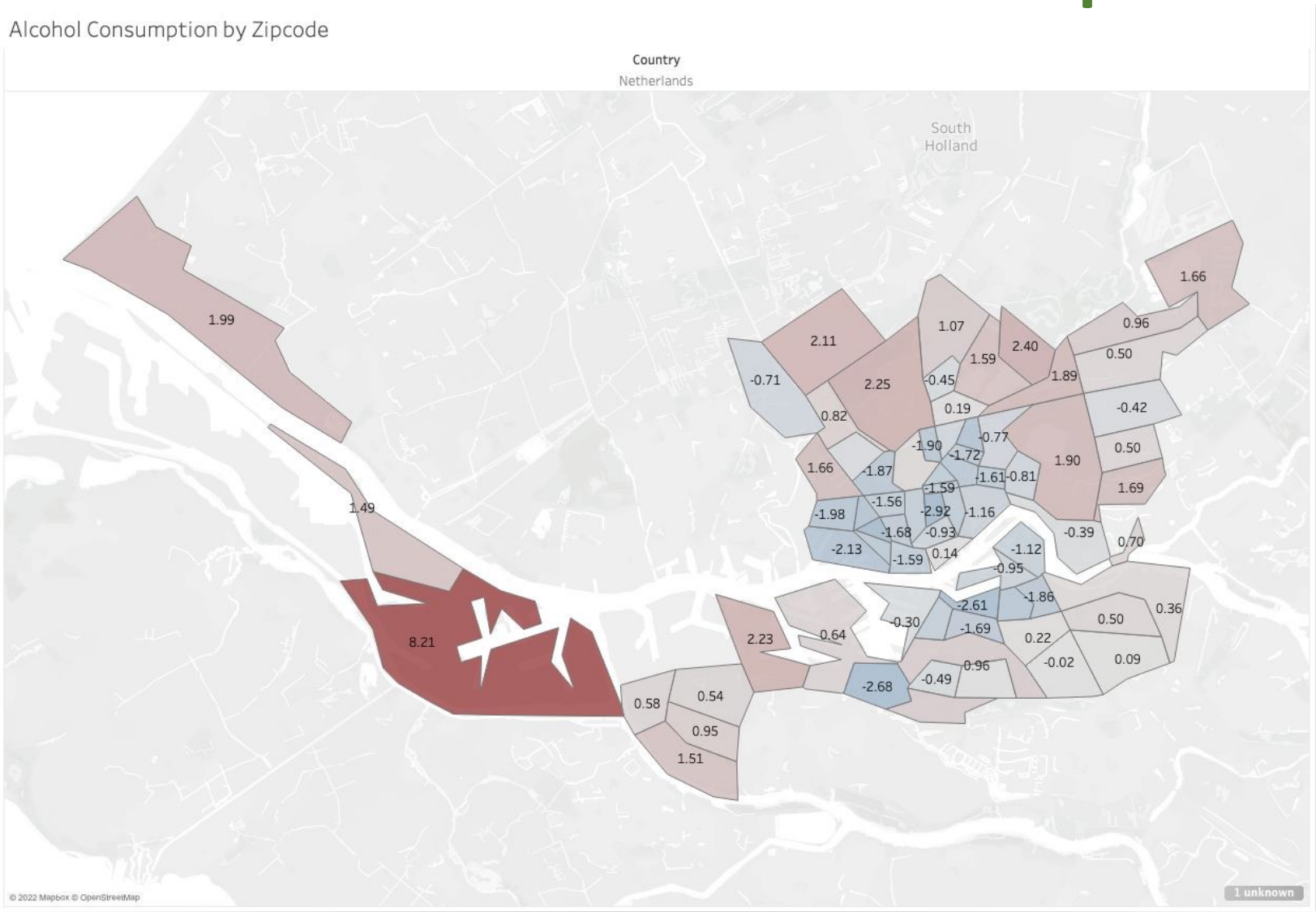
	Alcohol
	Comp.4
KNIBBE2	0.94
KNIBBE3	0.94
KNIBBE4	0.94
KNIBBE5	0.95
KNIBBE6	0.94
prodri	0.94
risicodrinker2	0.94
KNIBBE1	0.94
risicodrinker1	0.91
LFALB207	-0.7
LFALA201	-0.88

VARIABLE NAME	MEANING
KNIBBE2	Tried to stop, did not succeed
KNIBBE3	Skip meals
KNIBBE4	Use alcohol to forget
KNIBBE5	Partner or family memberis worried about alcohol
KNIBBE6	Irritated about remarks about alcohol
PRODRI	Problematic use of alcohol
KNIBBE1	Feel the need to drink less
RISICODRINKER1/2	Problematic use of alcohol and excessive drinking
LFALB207	Never have drunk alcohol

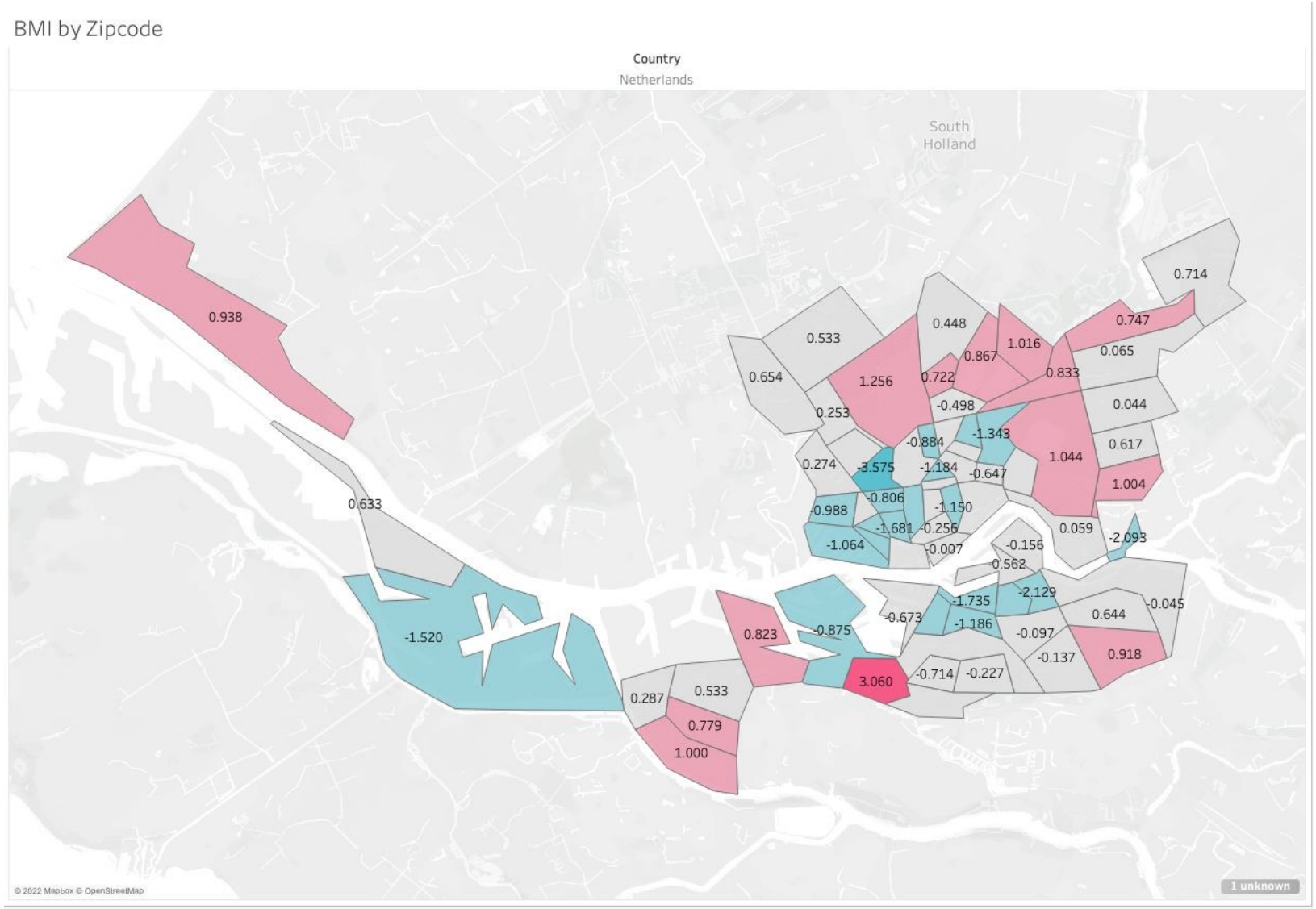
Data Visualization #1 – Domestic Violence



Data Visualization #2 – Alcohol Consumption



Data Visualization #3 -



Key Insights from the Data Analysis

Alcoholic Consumption	Smoking	Cardiovascular Disease	General Health	Domestic Violence	Sports Activity	Green Space in Neighbourhood
3197	**3197**	3065	**3197**	3041	**3197**	3062
3055	3035	3016	**3084**	**3197**	3035	3038
3045	**3084**	3041	3044	3088	**3084**	3065

Based on the analytics result, we identify critical postal areas such as 3197 and 3084, which requires immediate attention from government for more actions to improve the residents' health condition.

Government shall review the data and consider policy improvement & implementation on:

- Programs to reduce alcoholic consumption and smoking
- Provide more sports facilities and green space in neighborhood
- Prevention of domestic violence and precautionary health measures

Challenges

Biggest challenges of the project was the data cleaning!

1. Converting columns from factors to numeric in R to apply correlation.
2. Many variables in the dataset. Lots of planning & iteration involved with cleaning up the data
3. fixNA creating surrogate columns, effectively doubling the numbers of variables (as they had so much missing data). Removal of surrogates column required.

Next Steps

1. Apply further data cleaning
2. Attempt to segment the market using *cluster analysis* techniques described in class
3. Use *classification analysis* to classify people using classification analysis techniques.

Appendix – Principal Component Analysis

	POSCODN	Primaireeenheid	WEEGFACTOR1_OUD	WEEGFACTOR2_OUD
POSCODN	1.00	0.13	0.22-	0.22-
Primaireeenheid	0.13	1.00	0.12-	0.12-
WEEGFACTOR1_OUD	0.22-	0.12-	1.00	0.81
WEEGFACTOR2_OUD	0.22-	0.12-	0.81	1.00
WEEGFACTOR3_OUD	0.22-	0.11-	0.81	0.81
geslacht	0.01-	0.02-	0.08-	0.08-
AGLFS201	0.13	0.91	0.21-	0.21-
Herkomst	0.15-	0.14-	0.24	0.25
Generatie	0.16-	0.25-	0.26	0.27
inkkwin_2010	0.11	0.02	0.12-	0.12-

	Eigenvalue	Pct of explained variance	Cumulative pct of explained variance
Component 1	69.80	17.58	17.58
Component 2	25.59	6.45	24.03
Component 3	16.13	4.06	28.09
Component 4	14.38	3.62	31.71
Component 5	9.76	2.46	34.17
Component 6	9.18	2.31	36.48
Component 7	8.59	2.16	38.64
Component 8	6.80	1.71	40.36
Component 9	6.41	1.62	41.97
Component 10	6.20	1.56	43.53

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	C
lftcat2	0.98	0.02	0.02	0.02	0.02	0.00	0.03	0.02	0.08	0.04-	
MIGHA206_D2	0.97	0.03-	0.02	0.02	0.02	0.00	0.03	0.03	0.07	0.04-	
MIGHA207_D2	0.97	0.03-	0.02	0.02	0.02	0.00	0.03	0.03	0.06	0.03-	
MIGHA208_D2	0.97	0.03-	0.02	0.02	0.02	0.00	0.03	0.04	0.06	0.03-	
MIGHA209_D2	0.97	0.03-	0.02	0.02	0.02	0.01-	0.03	0.03	0.06	0.04-	
MIGHA210_D2	0.97	0.03-	0.02	0.02	0.02	0.01-	0.03	0.04	0.06	0.03-	
LFDGA201	0.97	0.02	0.03	0.05	0.02	0.01-	0.02	0.03	0.06	0.05-	
LFDGA202	0.97	0.02	0.03	0.04	0.02	0.01-	0.02	0.03	0.06	0.04-	
LFDGA203	0.97	0.02	0.02	0.04	0.02	0.01-	0.02	0.03	0.06	0.04-	
LFDGS201	0.97	0.02	0.03	0.04	0.02	0.01-	0.03	0.03	0.05	0.04-	
LFDGS202	0.97	0.02	0.03	0.03	0.02	0.02-	0.03	0.03	0.05	0.03-	
LFDGS203	0.97	0.02	0.03	0.03	0.02	0.01-	0.03	0.03	0.05	0.03-	
LFDGS204	0.97	0.02	0.03	0.03	0.02	0.01-	0.02	0.03	0.05	0.04-	
LFDGS205	0.97	0.02	0.03	0.04	0.02	0.01-	0.03	0.03	0.05	0.04-	
LFDGS206	0.97	0.02	0.03	0.03	0.02	0.02-	0.03	0.03	0.05	0.03-	
LFDGS207	0.97	0.02	0.03	0.03	0.02	0.01-	0.03	0.03	0.05	0.03-	
LFDGS208	0.97	0.02	0.03	0.03	0.02	0.01-	0.02	0.03	0.05	0.03-	
LFDGB202	0.96	0.02	0.03	0.04	0.02	0.01-	0.03	0.03	0.05	0.03-	
LFDGB204	0.96	0.01	0.03	0.03	0.02	0.02-	0.03	0.03	0.04	0.03-