COMP20008 Elements of Data Processing
Phase 4 Presentation

Semester 1 2018

Name: Xulin Yang
Student id: 904904

Table of content

Structure of the talk:

- 1 research questions
- 2 why questions are valuable?
- 3 data sets used & why
- 4 data preprocessing methods used
- 5 findings
- 6 challenges

research questions

- **Questions:**
- (1) Identify the correlation between community citizens' chronic disease risks and citizens' health risks.
- (2) What are features in communities that causes citizens to have high risks?
- outcomes

THE UNIVERSITY OF MELBOURNE

why questions are valuable?

- Purpose
- Who benefits & interests?
- Innovation

data sets used & why

- LGA11 Chronic Disease Modeled Estimate
- LGA11 Health Risk Factors Modeled Estimate
- LGA11 Psychological Distress Modeled Estimate
- Local Government Area LGA profiles data 2011

	Data_set	#Rows	3 example columns	Column data type	Realiable data%
2	LGA11 Chronic Disease Modeled Estimate	80	"Arthritis"	Discrete	70%
3			"Arthritis_RRMSE"	Qualitative	
4			"Local_Government_Area_Code"	Categorical	
5	LGA11 Health Risk Factors Modeled Estimate	80	"Current_Smokers"	Discrete	84%
6			"Current_Smokers_RRMSE"	Qualitative	
7			"Local_Government_Area_Code"	Categorical	
8	LGA11 Psychological Distress Modeled Estimate	80	"People_felt_high_mental_pressure"	Discrete	97.50%
9			"People_felt_high_mental_pressure_RRMSE"	Qualitative	
10			"Local_Government_Area_Code"	Categorical	
11	Local Government Area LGA profiles data 2011	79	"Unemployment_rate"	Discrete	70%
12			"People_lack_physical_activity"	Discrete	
13			"Local_Government_Area_Code"	Categorical	

All downloaded from aurin

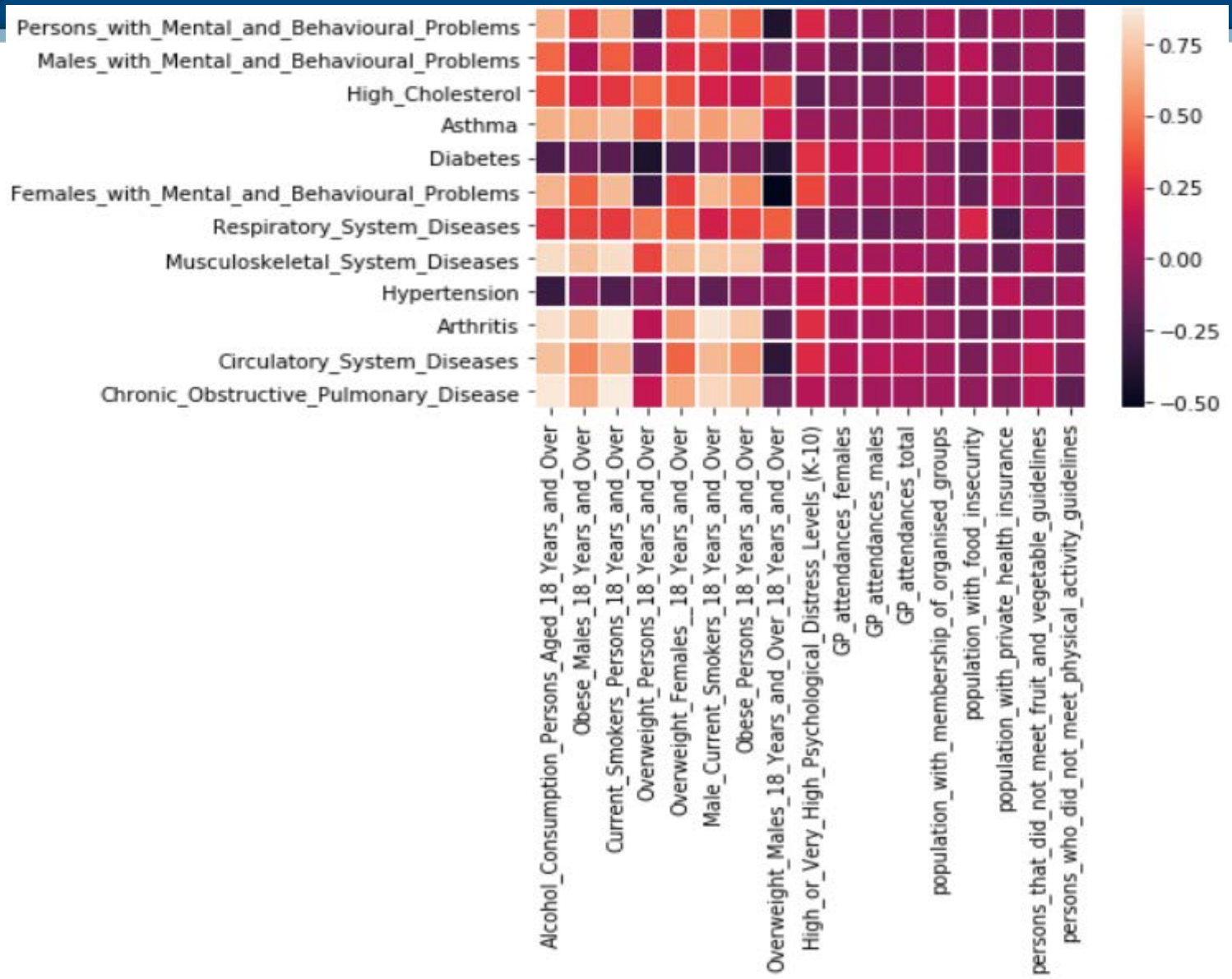
data preprocessing methods used

- Wrong data corrected
- Unreliable data dropped
- Missing data
- Columns replaced by schema
- Data normalization
- Deriving data from raw data

Finally have three DataFrame: disease_risks, health_risks, lga_features

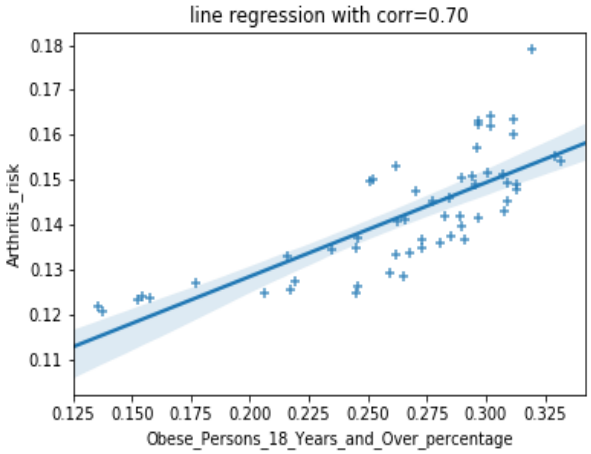
THE UNIVERSITY OF MELBOURNE

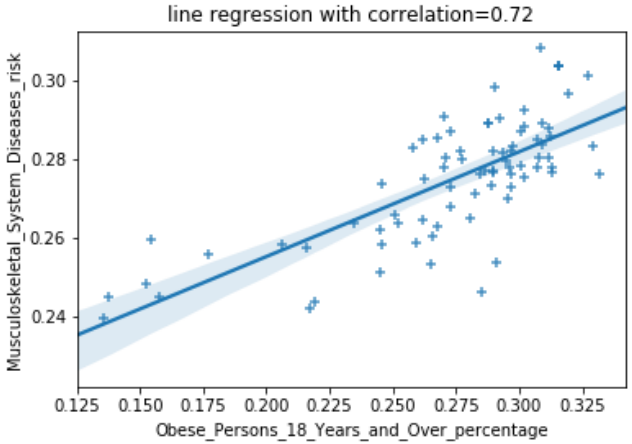
findings

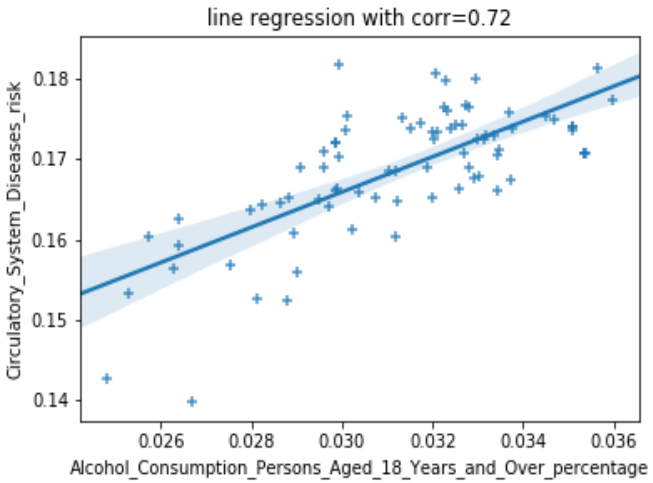


heat map against disease risk and health risk

findings

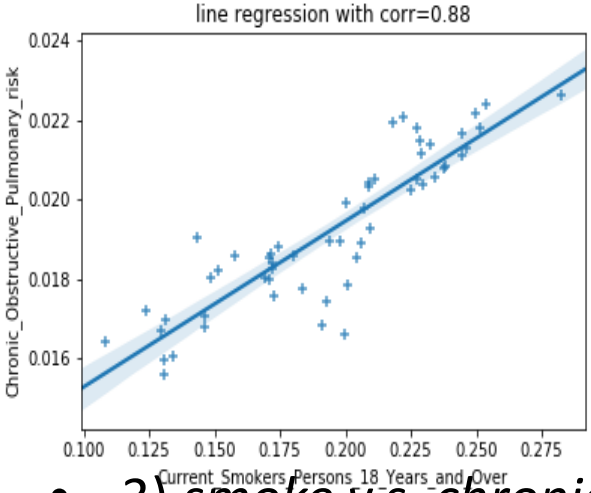


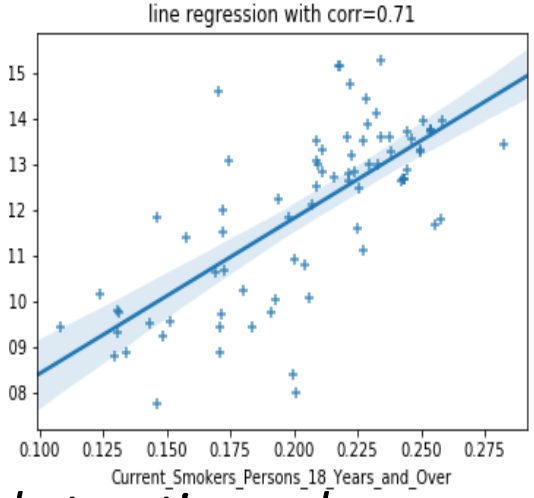




1) *obese v.s. arthritis risk* 2) *obese v.s. musculoskeletal system diseases risk*

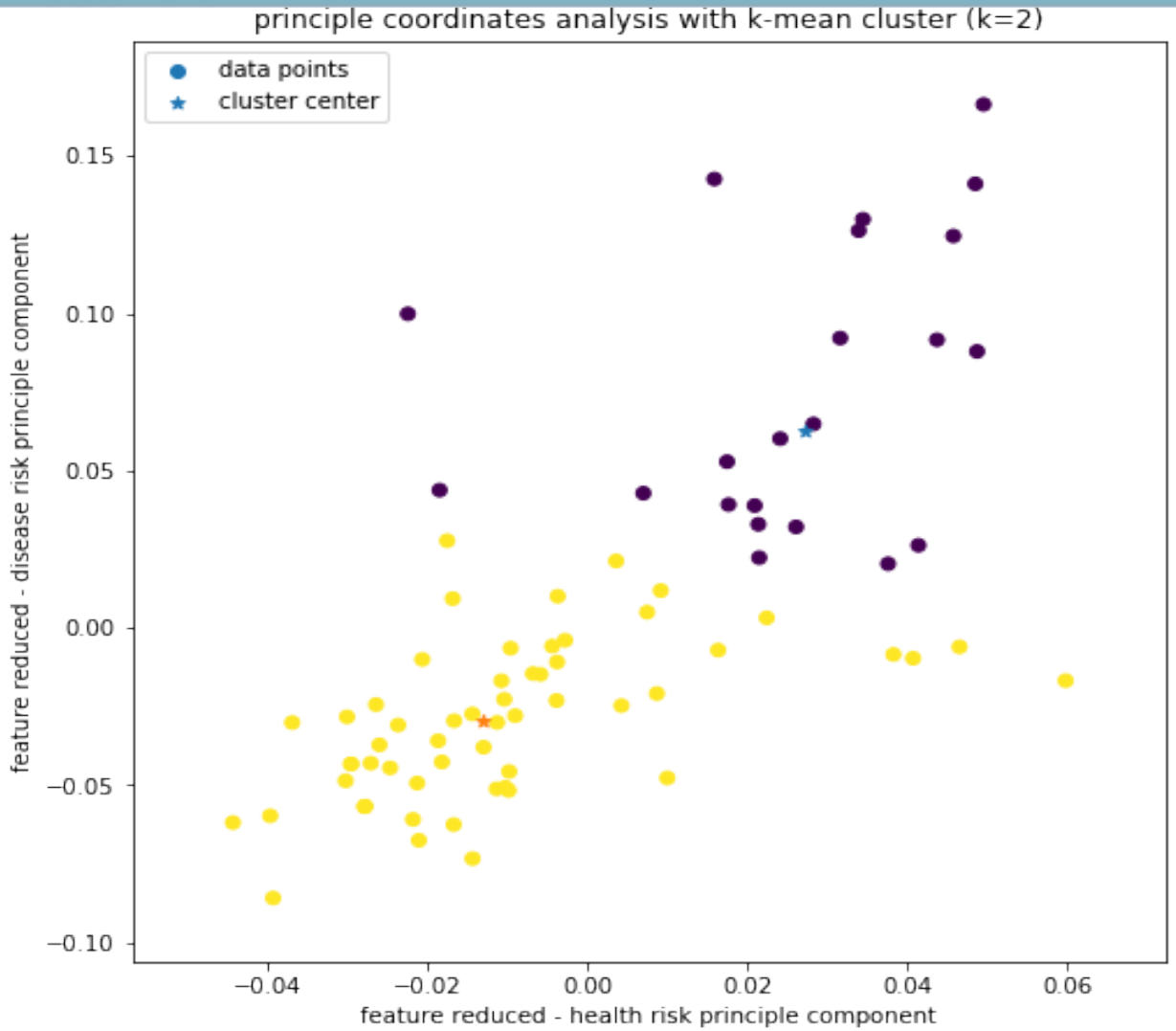
5) *alcohol v.s. circulatory system disease*





- 3) *smoke v.s. chronic obstructive pulmonary risk* 4) smoke *v.s. asthma*

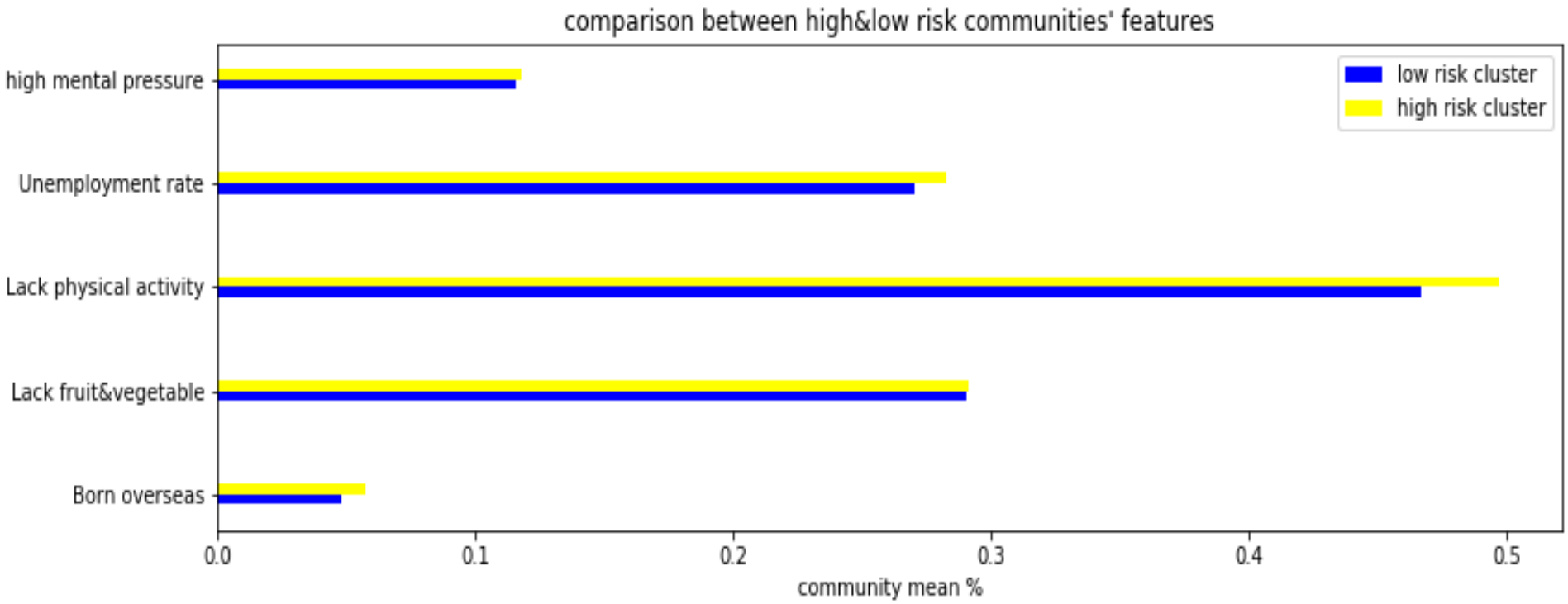
findings



The sixteen communities with high risks：
Moreland，
Kingston，
etc

principle coordinated analysis with K-mean (k=2) clustering

findings



community feature for clustered data

findings from analysis

- circulatory system disease <> drinking alcohol
- musculoskeletal system diseases & arthritis <> obese adults
- Asthma & chronic obstructive pulmonary disease <> smokers
- high people felt high mental pressure,
- high unemployment rate
- High rate of lack of exercises
- High immigrants' population percentages

Features of Community with High risk

<> : correlated

challenges

- Appropriate data sets to use
- Selecting columns(100+ columns in lga census data set) also consumes a lot of time

THE UNIVERSITY OF MELBOURNE

The end

Thanks for your attentions

Q & A

- Any questions and queries?