purpose and outcomes of the talk early enough?

communicate the structure of the talk?

1.What is the research question?

2. Why is it worth tackling (i.e. motivation)?

3. What are the datasets you used and why?

4. What data wrangling methodologies have you used to investigate your research question?

5. What did you find? Why is it interesting? What have you learnt?

6. What have been the challenges and what (if anything) would you have done differently?

show progression and connection between the parts of the presentation

Good morning every one:

My name is Xulin Yang. I am going to give the presentation of Elements of data processing phase four about the report in health and community domain.

The structure of the presentation is the following six major areas: research questions I derived, value of the questions, data sets I used and reasons of choices, data wrangling methods I used, findings of data wrangling and challenges I have faced.

Firstly, I am going to talk about the research questions I have explored and outcomes I discovered. This first question is whether citizens’ chronic disease risks are related to health risks for local government area wide which is lga wide. The second question is what are the features of communities with high risks? After spending several hours on doing data wrangling, I found five chronic disease risks are correlated to three health risks. And there are some features for high risk communities.

Secondly, the values of the questions are demonstrated in these aspects. The purpose of this investigation is to help to show the cause of chronic disease and give suggestions for high risk communities. Citizens in communities can benefit from this investigation as it can help local government to make policies and communities activities for improving health level of the community. In “The Chronic Care Model (Wagner et al.1999)”, it states that the chronic disease risk is correlated to community health service delivery rate and the personal health system. Compared to it, my investigation explores a wider range of community features and look for relation between these features and disease risks. With the above exploration, some characteristic community development advice can be obtained and provided for particular communities.

Thirdly, I have used all listed data sets. The details are showed in the table. The first one is used for disease risks factors. The second and third one are used for citizens’ health risks factors. The fourth data set is chosen because it can provide extra information about each lga area. For example: unemployment rate and people who is lack of enough physical activity.

* Wrong data corrected

In the second data set, {‘obese\_m\_me \_3\_rrmse\_3\_11\_7\_13’ : Obese Females 18 Years and Over - RRMSE } is wrong and corrected to ‘Obese Males 18 Years and Over - RRMSE’.

* Unreliable data dropped

There are columns like sth-PRMSE with 1 means this row’s data is unreliable and I dropped it.

* Missing data

There are some missing caused at data reading stage. As in the original csv file it is ‘<1%’ stores in string format, and can’t be read. So I used 0 to replace it.

* Columns replaced by schema

I write function to read json file which has schema information. And use it to replace the original name in DataFrame. By doing this, it make data with real meaning and easier to plot.

* Data normalization

Some data like “Number of hospital” can’t reflect anything about each lga. Then I use them divided by the population in that lga.

* Deriving data from raw data

Some data needed to be derived from one columns. For example, LGA\_name columns is a string with and type of local government type. I uses regular expression to split these two data to name column and lga\_type column.

After doing all these data preprocessing steps above, the data set is cleaned and ready for investigation compared to the raw data. About 30% data are drooped as they are not reliable.

* The community with higher percentage of adults drinking alcohol has a higher risk of having the circulatory system disease(r[[1]](#footnote-2) 0.72).

·The community with higher percentage of obese adults has a higher risk of having musculoskeletal system diseases(r ) and arthritis(r ).

·The community with higher percentage of adults who smoke results in a higher risk in asthma(r ) and chronic obstructive pulmonary disease(r 0.88).

1. Pearson correlation coefficient [↑](#footnote-ref-2)