图片包含 徽标

描述已自动生成

**DATA1002 Group Assignment Stage 1**

**Group**

**470000946 Xiaobing Wang**

**470021332 Shiyu HU**

图形用户界面, 文本, 应用程序, 信件

描述已自动生成

**Section1 Description of data sets**

The aim of this report is to indirectly understand climate change by studying the weather conditions and changes in Sydney in recent years.

We selected data from three different sources, mainly BOM, world bank and xxx.

The BOM data is relatively extensive, and for Australia, there are a lot of world bank data, but not as detailed as BOM. Among the websites we were looking for, only the Australian Bureau of Meteorology data can be downloaded for free and comprehensive.

We found data on Sydney’s own temperature, wind speed, humidity, for everyday and the last 20 years from BOM, which shows the temperature and climate trends in Australia over the last 20 years.

Using the comprehensiveness of the world bank data, we downloaded information about Australian Carbon dioxide/forecast

The database is about the in Australia. It shows the.

We can further analyze the correlation between weather and climate, as

However, the database is too complicated, we should clean it and only get meaningful data as well as use it.

In the end of the report a conclusion will be drawn that a question proposed. The dataset used in this report comes from the Source: Australian Bureau of Meteorology, World Bank and NSW government data which are reliable resource.

We use Python to put them together, delete unnecessary data, draw some graph and use that data to draw conclusions.

我已经原地去世了

我们要研究啥？？？

这数据怎么比房子还不好找？？？

1. **Dataset**

必备要求

Data is sufficiently large that automated processing shows genuine benefits

It is produced by combining data from at least two different sources.

Pass是100个；Distinction是500个，2 source ；

满分要求（1分）

at least 3 sources where there are at least 1000 values from each of these sources

澳洲统计局收集的3个dataset 会被当作同一个source，NASA收集的会被当成另一个

1. **Transformation**

1分：at least one aspect of data quality in a simple dataset has been automatically checked

2分：handles in a sensible way, several different kinds of data quality and format difficulties, automatically, integrate the data from a complex dataset

Transforming and cleaning

写3个程序分别处理不同source来clean data，或1个程序处理3个。

* 去掉corrupted or missing values 或填上missing values，
* 修正拼写错误，修正data formats in one source to match another
* 整合不同来源的数据，去掉重复

写程序检查data is clean ( no missing data, every entry has an appropriate value)

描述Metadata：写清来源，changes made，变量的意义。

1. **Analysis**

1分：*at least four useful summaries* and furthermore, at least one of which the summaries must be a grouped- aggregate

2分：at least four sensible grouped-aggregate statistics, where these don’t all use the same attribute to define the groups, and they also don’t all give the same kind of aggregate (sum, maximum, count…)

写一个程序处理integrated file

Produce a few summaries (aggregates) of some attributes

1. **Write a report and submit it**

Four deliverables：Raw data, clean data (with metadata), code, report

### Report 结构：

**Section 1 Description of data sets**

描述the data source(s), the format/contents of the data, the rights associated with the data, and some comment on any strengths or limitations of the dataset;

1.1 data source(s)

Raw Dataset 1， Population estimates by Local Government Area, 2018 to 2019

<https://www.abs.gov.au/statistics/people/population/regional-population/latest-release#data-downloads-geopackages>

Raw Dataset 2， Population estimates by age and sex - summary statistics, by LGA, 2019

<https://www.abs.gov.au/statistics/people/population/regional-population-age-and-sex/latest-release#data-download>

Raw Dataset 3, NSW COVID-19 tests by location and result

COVID-19 cases by notification date and postcode, local health district, and local government area. The dataset is updated daily, except on weekends. The data is for confirmed...

<https://data.nsw.gov.au/data/dataset/covid-19-cases-by-location>

Raw Dataset 4,NSW Hospitals

Hospitals: A full list of health services, including name, address, contact details and Local Health District is available for download in CSV format.

<https://data.gov.au/dataset/ds-nsw-b4573657-81dc-46e7-9677-65b607f734d4/distribution/dist-nsw-e17840df-ecfc-4e38-b51b-9f49af5dc21a/details?q=>

The dataset is obtained from\*\*\*

1.2 the rights associated with the data

The data is authorized by \*\*\*, a famous official \*\*\* organization of \*\*\*. It gives us \*\*\* based on the comprehensive strength among \*\*\* as well as some \*\*\* elements such as \*\*\*.

1.3 the format/contents of the data

可以用code显示，比如data.

也可以如下描述

|  |  |  |
| --- | --- | --- |
| Variable | Description | Type |

1.4 some comments on any strengths or limitations of the dataset;

**Section 2 Data transformation**

描述the initial transformation and cleaning that you did (include here the parts of Python code that you used, or a description that is detailed enough to be followed); this section should end with a brief explanation of where (in the submitted material) is found the metadata for the resulting dataset(s) which you will use in the rest of your project;

2.1 Check and remove null values, fill in missing values

2.2 Transform data formats

2.3 Merge dataset and remove duplicates

2.4 Automated checking for quality verification

2.5 Metadata of the merged dataset (这个应该和clean data 在一起)

**Section 3 Analysis of the data**

描述并解释some simple analysis that you have done (again, show the code and also the output of the analysis).

3.1 Aggregation of \*\*\* by processing the integrated file

Code (as appendix or used as screenshots)

Output (printed values or screenshots of summaries)

3.2 Explanation/Interpretation