Introduction to Data Science:
DATA7001, S1, 2021
Project Report of Group 8

Weather Data Analysis in Australia

Group member:

46281740 Zichuan Huang

46130035 Runqi Shi

45749663 Zenghui Liu

4661597 Savit Anawatmongkol

Statement

We DO Not give consent for this to be used as a teaching resource

Contents

1. Introduction

Weather affects our daily lives for every single event we going to do it. Predicting the weather is one of the most outstanding topics that has influenced people's lives for a long time. Moreover, knowing the weather that will rain tomorrow can be deciding factor between a soak muddy climbing activity or movie night with relaxing raining sound. Having an accurate weather forecast can also save people lives in some situation like a flood.

This project aims to use a multiple logistic regression model to predict rain in three different parts of Australia: middle, west, and east. Our objective is to predict whether or not rain will fall the next day in Australia. This prediction can help people for several reasons. For example, it helps people determine what to wear or what activity they will do on a given day, whether the weekend will permit outdoor events. It also helps people to decide which date they should carry an umbrella or put on a coat or not.

Compared to previous project, some changes were made below:

- The title of our project was changed, in order to make a more general conclusion.
- Since the lack of the hypothesis in the presentation, some explanation of the reason that multiple model regression was used in this project.
- In the project pitch, the aim of our project is not so clear, so in the following analysis, we make it more specific.

2. Problem solving with data

Weather data is a geographic research data. When we are analyzing this type of data set, usually the first step is to understand the content of the data. This data set contains data values of different weather change detection types, as well as a summary of weather conditions. By analyzing weather data, we can understand the changes in weather conditions in different geographic environments. We mainly have two parts: the first part is to understand the data set; the second part is to design the data analysis method. Our design idea is to compare different local data from point to surface, from part to whole. By comparing and understanding the weather characteristics of different regions, the analysis results can be obtained.

3. Getting the data I need

Our data set is a ready-made data set. We did not carry out additional crawling and collection work, but selected the data set that others have collected for data analysis. Data set link: https://www.kaggle.com/jsphyg/weather-dataset-rattle-package. What we pointed out was the issue of data privacy. We believe that the content of this data set does not involve data privacy issues, because the data content is about geography and weather, it is scientific and statistical data, not private data.

4. Is my data fit for use

In this part, we mainly carried out the following operations.

- Due to the large amount of data, we deleted data rows with missing values when filtering the data.
- When choosing a special location, we choose location according to the completeness of data of the target feature.
- When analyzing data, we adopt partial data analysis according to different locations, and try to select the data of the location with the best data quality for analysis. The figure below is about

the data of different selected cities, such as Adelaide, Albury, Alice Springs, Badgerys Creek, Ballarat, Bendigo.

We mainly deal with missing values in the data. When analyzing part of the data, we adopted the idea of data imputation. The figure below shows the statistics of missing values in the data set.

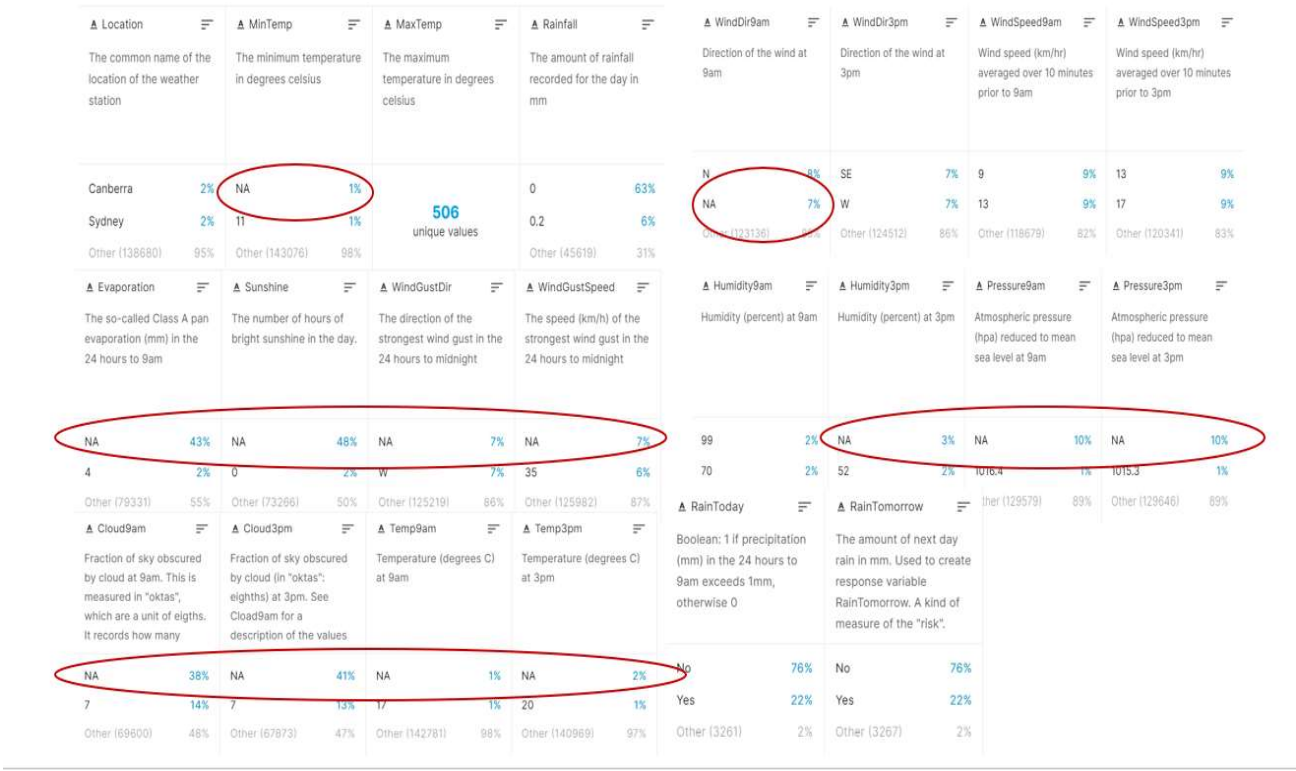


Figure 1 Overview of missing data

When we deal with missing data values, the method we use is based on the specific content of data set and the purpose of analysis.

- Some locations without observation values of the two attributes of Evaporation or Sunshine, so there is not an obvious relationship between Evaporation and Sunshine in some locations. Therefore, we chose data having values for the attribute of Evaporation, and only the attribute of Evaporation is considered. For the non-empty value of Evaporation data, which is extracted.
- In addition, the non-empty value distributions of Cloud9am and Cloud3pm are irregular and account for a large proportion, so these two variables are not taken into consideration in this analysis.
- we deleted the rows with 'NA 'values in the remaining attributes.

In the original data, the data type of the attributes Rain Today and Rain Tomorrow is character data type.

- To facilitate subsequent analysis, we assigned the data of 'Yes' to 1 (numeric type) and assigned the data of 'No' to 0 (numeric type).
- About the attributes wind direction, the data type is also character data, so we did not take it into consideration.

Moreover, to facilitate the analysis of weather conditions throughout Australia, 44 locations have been added with longitude and latitude coordinates.

5. Making data confess

The main target of this part is to predict rainy day by analyze weather factors. Since the dataset is large, we chose 2 to 3 different locations in 3 area of Australia, which is mid-land, west side and east side.

5.1 Hypothesis

The main target in this part is to analyze these factors to predict if it is going to rain tomorrow. The output should be yes or no, and the input are multi-value, so a multiple logistic regression model should be suitable for this scenario.

5.2 Sampling

Because Australia has a large longitude span, we decided to select 2 to 3 neighboring cities in each region of the east, west, and central of AUS for analysis. In the middle-land, we chose Alice Springs and Uluru. In the west side, we chose Perth and Perth Airport. In the east-side, we chose Sydney, Sydney Airport and Wagga Wagga. By dividing them into three groups, it should be better to recognize a more specific weather pattern in each area. Another reason that these locations were selected is that some locations don't have records of some attributes like 'Evaporation' or 'Sunshine', so those locations that have these attributes were chosen.

In each area, random sampling with replacement was used to divide the dataset into training set and testing set, which ratio is 7 : 3. The reason that random sampling was applied is that the size of the dataset is large enough, which will not result in incorrect representation.

```
set.seed(33)
sampleidx <- sample(x=1:nrow(east), size=nrow(east)*0.7)
east_train <- east[sampleidx, ]
east_test <- west[-sampleidx, ]
```

Figure 2 Sampling of training set and testing set

5.3 Regression model

In the process of fitting, the logistic regression model was built twice. The first fitting is to remove those variables that are not relevant. It can be implied from the p value of each attributes.

In the first fitting, we include almost every weather factor, which is MinTemp, MaxTemp, Evaporation, Humidity9am, Humidity3pm, Pressure9am, Pressure3pm, WindGustSpeed, WindSpeed9am and WindSpeed3pm. The graph below shows the result of first fitting in midland. Some irrelevant attributes whose p-values are bigger than 0.05 can be seen on the graph, so these factors should be moved in the second fitting.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6713  -0.2722  -0.1699  -0.1120   3.2101

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   60.711983  35.425181   1.714  0.08656 .
MinTemp        0.069382   0.044156   1.571  0.11611
MaxTemp       -0.018811   0.049154  -0.383  0.70194
Evaporation   -0.022527   0.032615  -0.691  0.48975
Humidity9am   -0.007228   0.009371  -0.771  0.44050
Humidity3pm    0.087888   0.012268   7.164 7.84e-13 ***
Pressure9am    0.268966   0.121331   2.217  0.02664 *
Pressure3pm   -0.337183   0.122789  -2.746  0.00603 **
WindGustSpeed  0.069682   0.011446   6.088 1.14e-09 ***
WindSpeed9am  -0.016084   0.017246  -0.933  0.35102
WindSpeed3pm  -0.048355   0.016338  -2.960  0.00308 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3 The first fitting of midland area's data

In the second fitting, there was only Humidity3pm, Pressure9am, Pressure3pm, WindGustSpeed and WindSpeed3pm left. These factors have high correlation with the output.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8235  -0.2705  -0.1719  -0.1141   3.1913

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   100.348382  16.347202   6.139 8.33e-10 ***
Humidity3pm     0.089183   0.005245  17.003  < 2e-16 ***
Pressure9am     0.246554   0.085918   2.870 0.004109 **
Pressure3pm    -0.354608   0.086249  -4.111 3.93e-05 ***
WindGustSpeed   0.078886   0.008642   9.129  < 2e-16 ***
WindSpeed3pm   -0.047472   0.012590  -3.770 0.000163 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

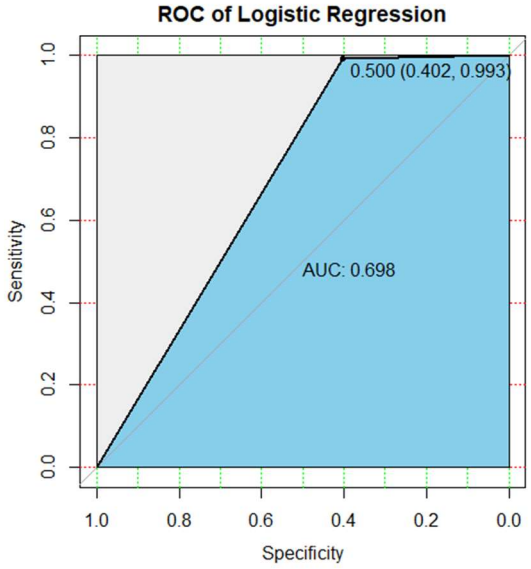
Figure 4 The second fitting of midland area's data

The other two areas were fitted by the same method, and the results can be seen in the code.

5.4 Model evaluation

In order to evaluate the validity of the model, ROC (Receiver Operating Characteristic) curve and confusion matrix for binary classification were used. In figure 8, the curve locates at the top-left of the graph and the AUC (Arear Under Curve) is 0.698, which implies that the precision of the model is good. From the confusion matrix, the Accuracy Rate can be calculated, which is

$$(TP+TN)/(TP+TN+FN+FP) = 94\%$$



	Predict Value	
Actual Value	0	1
0	1149	8
1	61	41

Figure 5 ROC curve and confusion matrix of mid-land area

For the other two groups, the evaluate results can be calculate as above, which are also good for prediction.

6. Storytelling with data

According to the dataset, which contains daily weather observations from 2009 to 2016 in 44 locations across the whole of Australia, it is found that there are a few different meteorological features in different regions.

Figure 6 shows the distribution of rainfall in 44 locations. As we can see, in general, the sum of rainfall in northern Australia is higher than that in the south.

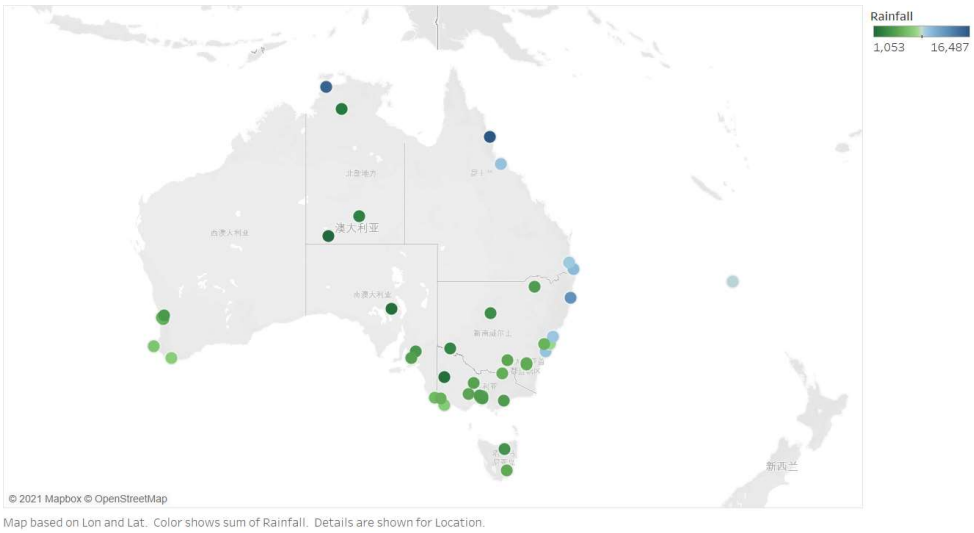


Figure 6 The distribution of rainfall in 44 location

Figure 7 shows the distribution of 44 locations' average sunshine. It can be seen that the value of average sunshine in the northern area is higher than that in the south and that in the inland area is higher than the coastal area.

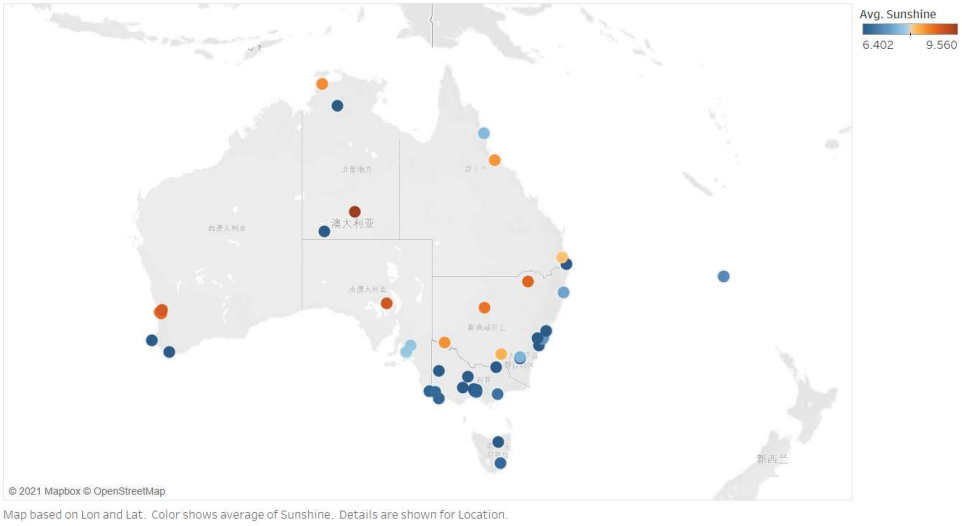


Figure 7 Average sunshine distribution

The followed figure (Figure 8) illustrates the distribution of the average maximum temperature in 44 observation stations. It can be seen that the value of the average maximum temperature in northern Australia is higher than that in the south, and the inland area has a higher temperature value than coastal cities.

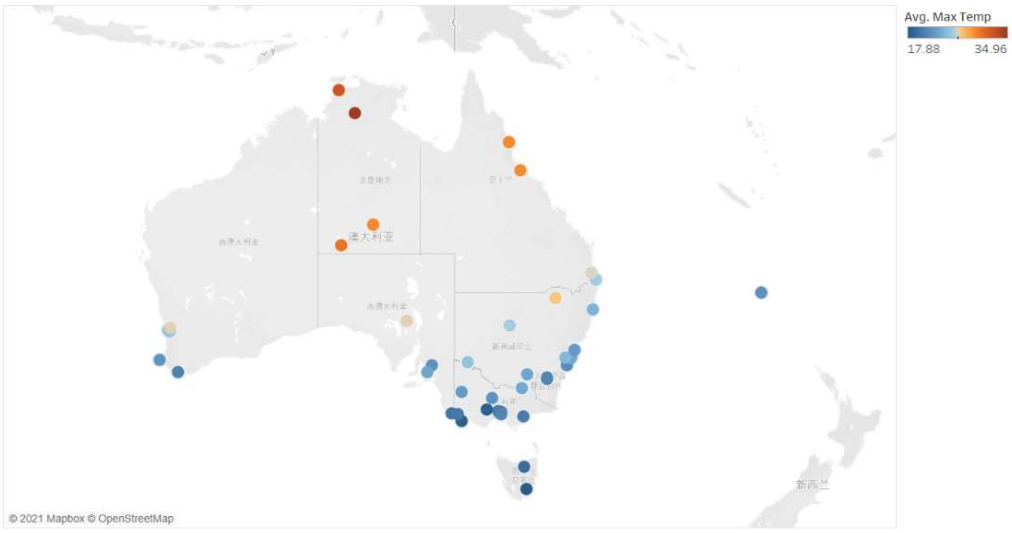


Figure 8 Average maximum temperature

The distribution of the average minimum temperature in 44 locations is shown in Figure 9. It shows that the value of the average minimum temperature in northern Australia is higher than in the south. However, another conclusion can be seen is that there is no obvious difference between inland and coastal area.

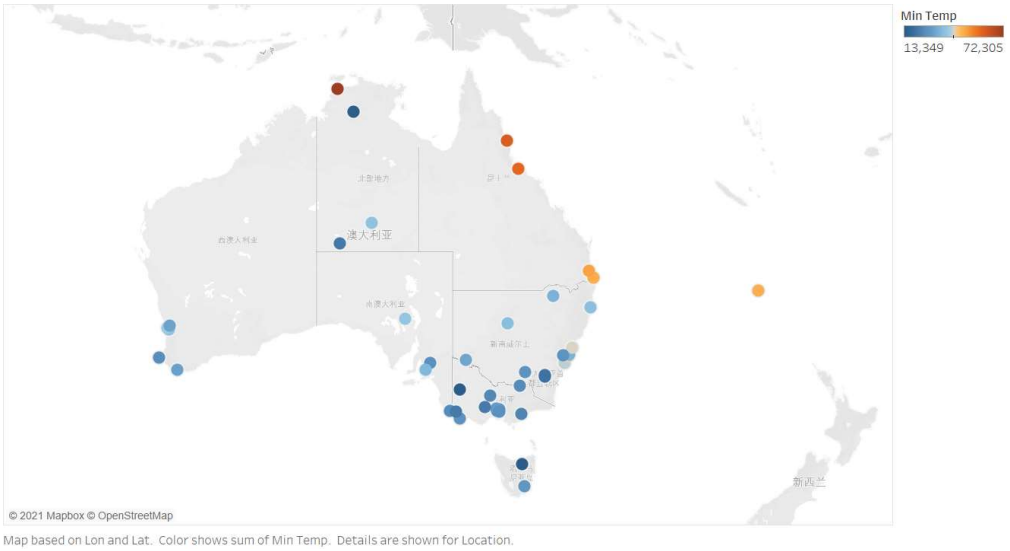


Figure 9 Average minimum temperature

In order to study the difference from various regions of Australia, 5 locations from different regions are selected as representatives.

Firstly, it compares the average humidity at 9am and 3pm from 5 stations. As we can see from Figure 10, Alice Springs has the lowest average humidity. Besides, compare morning and afternoon data from these two lines charts, we can see that the location which has the biggest difference between morning and afternoon is Hobart.

Figure 10 Average humidity from 2007 to 2017

Figure 11 illustrates the average temperature at 9am and 3pm in 5 locations. It can be seen that generally the place with the lowest average temperature is Hobart and the temperature's change is similar in the morning and afternoon in all five locations.

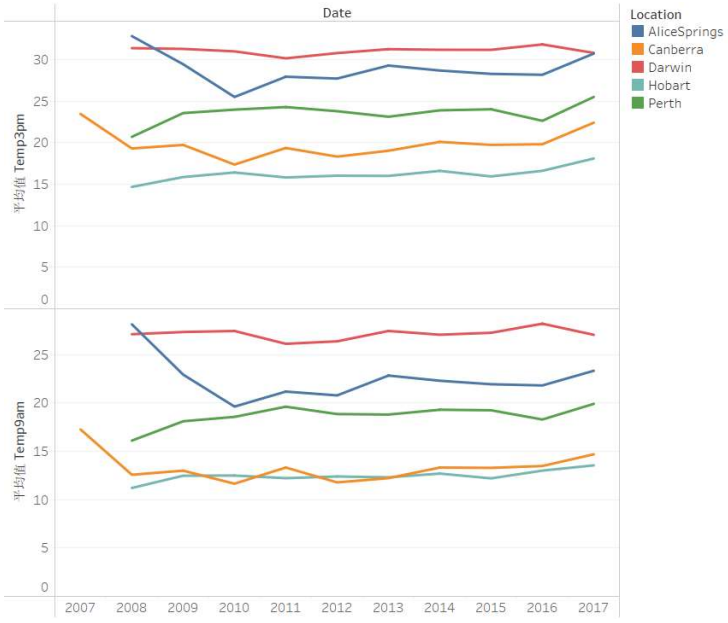


Figure 11 Average temperature from 2007 to 2017

Figure 12 is data of average wind speed at 9am and 3pm. It shows that the wind speed in the afternoon is higher than that in the morning, except for Uluru which has the opposite observation.

Figure 12 Average wind speed from 2009 to 2017

Figure 13 shows the average data of sunshine, evaporation, and rainfall. It firstly can be seen that Darwin has the largest value of rainfall among those five locations. Meanwhile, it is clear that the trend of evaporation and sunshine is the same. It also shows that after evaporation remains high for a period of time, rainfall will increase significantly.

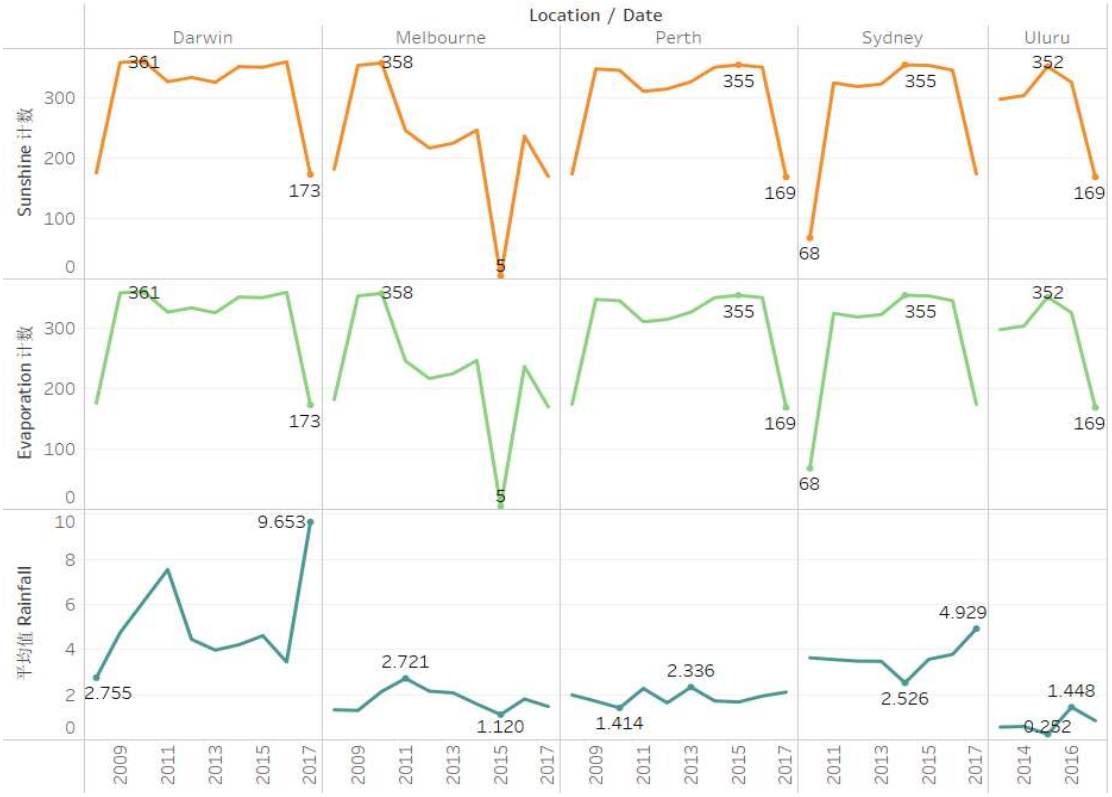


Figure 13 Largest value of rainfall in 5 locations

7. Reference

【1】 Data set link: https://www.kaggle.com/jsphyg/weather-dataset-rattle-package.

【2】 Tools and resources: R studio, Tableau, content of lecture, Multiple logistic regression.

【3】 Multiple logistic regression resources: https://www.cnblogs.com/Hyacinth-Yuan/p/7905855.html

【4】 Multiple logistic regression resources: https://blog.csdn.net/weixin_43216017/article/details/87904663