

Thunchanok Nakpasom
6009680106

June 10 , 2021

Rain in Australia

o o o o

Predict next-day
rain by training
classification
models on the
target variable
RainTomorrow.

กำหนดการ ของวันนี้

○ ○ ○ ○

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

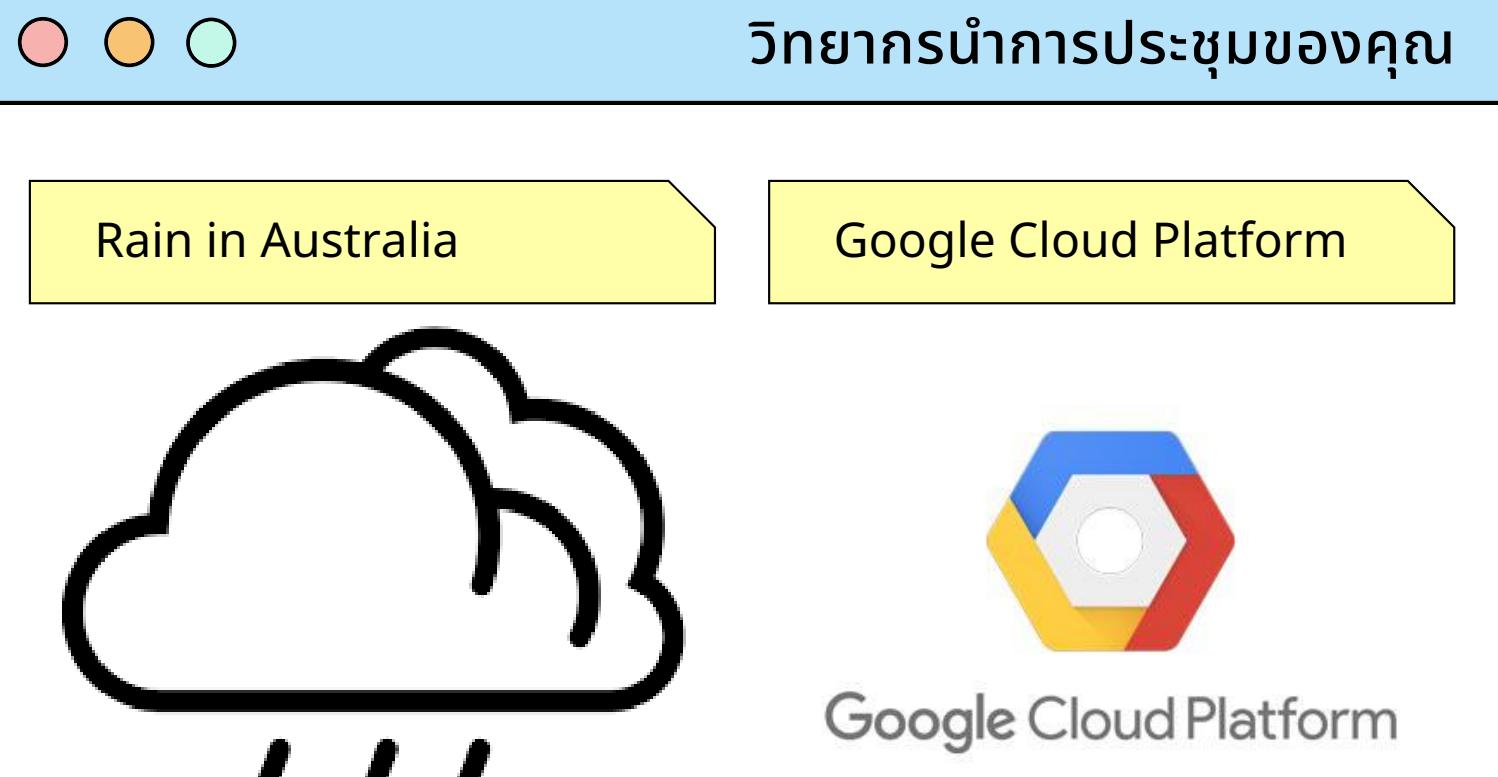
- กี่มาและความสำคัญ
- เป้าหมาย
- รายละเอียดข้อมูล
- สถาปัตยกรรมระบบเบื้องต้น
- แผนการดำเนินงาน
- วิธีการดำเนินงานตามหลัก CRISP-DM
- ผลลัพธ์ที่ได้ และสรุปผล
- สิ่งที่ได้เรียนรู้และแนวทางพัฒนาต่อไป

ที่มาและความสำคัญ

○ ○ ○ ○

ปัจจุบันหลายประเทศทั่วโลกต่างกำลังเผชิญกับปัญหาภัยพิบัติที่เกิดจากการเปลี่ยนแปลงของสภาพภูมิอากาศ

ทำให้มีผู้เดินรับความเดือดร้อนเป็นจำนวนมาก ถ้าหากเราสามารถนำ Big Data มาวิเคราะห์และนำนายล่วงหน้าได้อย่างถูกต้องแม่นยำว่าฝนจะตกในวันพรุ่งนี้หรือไม่ จะส่งผลให้ผู้คนสามารถใช้ชีวิตได้ดีและสะดวกยิ่งขึ้น



ເປົ້າໝາຍ

○ ○ ○ ○

Rain in Australia

classification
models

Predict next-
day rain



ເປົ້າໝາຍ (Goal)

Should I be prepared to handle the rain?

ຈາກຂໍ້ມູນເກີ່ວກັບຝນໃນປະເທດອອສເຕຣເລີຍ ຜູ້ຈັດກຳນີ້ເປົ້າໝາຍໃນສ້າງຄວາມເຂົ້າໃຈເກີ່ວກັບຂໍ້ມູນ
ສາມາຮັບອຸກໄດ້ວ່າປ່ອຈັຍໄດ້ມີຜລຕ່ອກເກີດຝນຕົກ ແລະສ້າງໂນເດລທີ່ສາມາຮັກມຳນາຍວ່າ ໃນວັນພຣຸ່ງນີ້ຝນຈະ
ຕົກໃນປະເທດອອສເຕຣເລີຍຫຼືອ໌ໄມ້ ໄດ້ອ່າຍ່າງຄຸກຕ້ອງແມ່ນຢໍາ

ໂດຍນີ້ເກັນກົດການຕັດສິບໃຈຄົວ

"ເຮົາຈະກຳການເຕຣີຍນຕົວຮັບນີ້ກັບຝນຕົກ ລ້າທາກໂອກາສທີ່ຝນຈະຕກອງຢູ່ກ່ຽວມານ **70%**"



- Factors
- Model
- Decision

รายละเอียดข้อมูล

○ ○ ○ ○

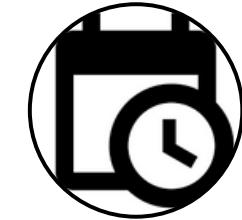
1

Kaggle --> Rain in Australia

2

"weatherAUS.csv"
145460 rows 23 columns

● ● ● Date



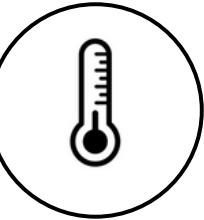
วันที่สังเกต
String

● ● ● Location



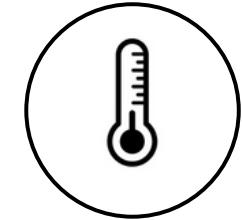
ชื่อที่ตั้งของสถานีตรวจน้ำอากาศ
String

● ● ● MinTemp



อุณหภูมิต่ำสุด (องศาเซลเซียส)
Float

● ● ● MaxTemp



อุณหภูมิสูงสุด (องศาเซลเซียส)
Float

รายละเอียดข้อมูล

○ ○ ○ ○

1

Kaggle --> Rain in Australia

2

"weatherAUS.csv"
145460 rows 23 columns

● ● ● Rainfall



ปริมาณน้ำฝนที่บันทึกไว้
ในแต่ละวัน
(มิลลิเมตร)

Float

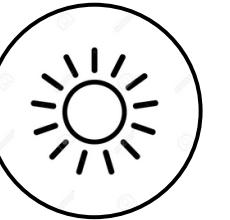
● ● ● Evaporation



การระเหย ณ เวลา 9.00 น.
โดยใช้เครื่องมือวัดแบบ
Class A pan
(มิลลิเมตร)

Float

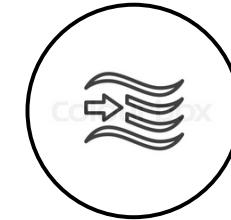
● ● ● Sunshine



จำนวนชั่วโมงที่มีแสงแดด
ในแต่ละวัน

Float

● ● ● WindGustDir



ทิศทางลมกระโชกแรงที่สุด

String

รายละเอียดข้อมูล

○ ○ ○ ○

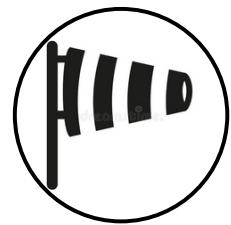
1

Kaggle --> Rain in Australia

2

"weatherAUS.csv"
145460 rows 23 columns

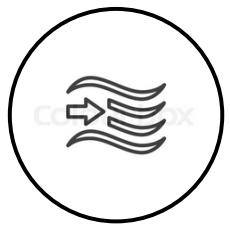
● ● ● WindGust Speed



ความเร็วของลม
กระซอกแรงที่สุด
(กม./ชม.)

Float

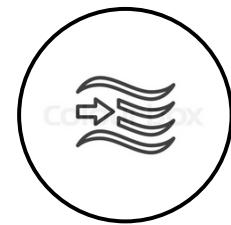
● ● ● WindDir9am



ทิศทางของลมเวลา 9.00 น.

String

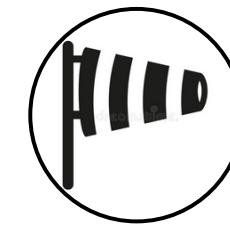
● ● ● WindDir3pm



ทิศทางของลมเวลา 15.00 น.

String

● ● ● WindSpeed 9am



ความเร็วลม (กม./ชม.)
เวลา 9.00 น.

Float

รายละเอียดข้อมูล

○ ○ ○ ○

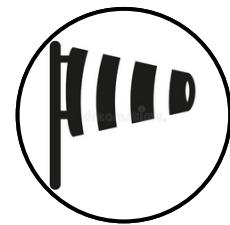
1

Kaggle --> Rain in Australia

2

"weatherAUS.csv"
145460 rows 23 columns

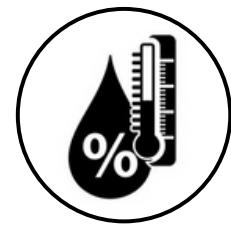
● ● ● WindSpeed
3pm



ความเร็วลม (กม. / ชม.)
เวลา 15.00 น.

Float

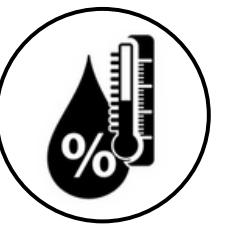
● ● ● Humidity
9am



ความชื้น (เปอร์เซ็นต์)
เวลา 9.00 น.

Float

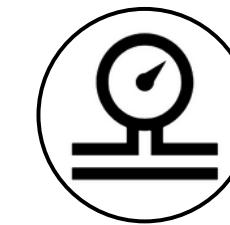
● ● ● Humidity
3pm



ความชื้น (เปอร์เซ็นต์)
เวลา 15.00 น.

Float

● ● ● Pressure
9am



ความดันบรรยากาศ (hpa)
เวลา 9.00 น

Float

รายละเอียดข้อมูล

○ ○ ○ ○

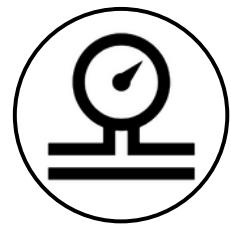
1

Kaggle --> Rain in Australia

2

"weatherAUS.csv"
145460 rows 23 columns

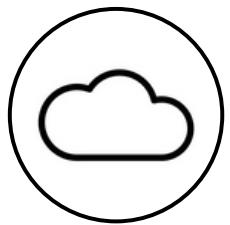
● ● ● Pressure
3pm



ความดันบรรยากาศ (hpa)
เวลา 15.00 น

Float

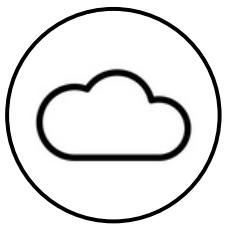
● ● ● Cloud
9am



มาตราส่วนกำหนดเมฆปกคลุม
เวลา 9.00 น.

Float

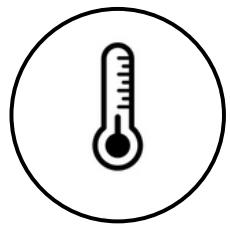
● ● ● Cloud
3pm



มาตราส่วนกำหนดเมฆปกคลุม
เวลา 15.00 น.

Float

● ● ● Temp
9am



อุณหภูมิ (องศาเซลเซียส)
เวลา 9.00 น.

Float

รายละเอียดข้อมูล

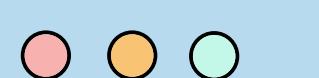
○ ○ ○ ○

1

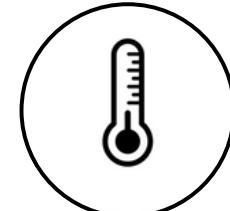
Kaggle --> Rain in Australia

2

"weatherAUS.csv"
145460 rows 23 columns



Temp
9am

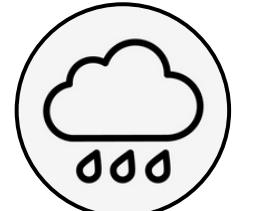


อุณหภูมิ (องศาเซลเซียส)
เวลา 15.00 น.

Float



RainToday



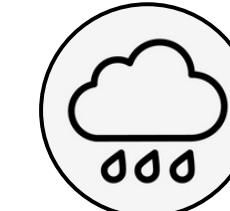
ปริมาณห้าฝน (มิลลิเมตร)

หากเกิน 1 มม. = 1 ถ้าไม่เกิน = 0

String



Rain
Tomorrow



ปริมาณฝนในวันถัดไป
(มิลลิเมตร)

หากเกิน 1 มม. = 1 ถ้าไม่เกิน = 0

String

1

df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 145460 entries, 0 to 145459
Data columns (total 23 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   Date        145460 non-null   object  
 1   Location    145460 non-null   object  
 2   MinTemp     143975 non-null   float64 
 3   MaxTemp     144199 non-null   float64 
 4   Rainfall    142199 non-null   float64 
 5   Evaporation 82670 non-null   float64 
 6   Sunshine    75625 non-null   float64 
 7   WindGustDir 135134 non-null   object  
 8   WindGustSpeed 135197 non-null   float64 
 9   WindDir9am   134894 non-null   object  
 10  WindDir3pm   141232 non-null   object  
 11  WindSpeed9am 143693 non-null   float64 
 12  WindSpeed3pm 142398 non-null   float64 
 13  Humidity9am  142806 non-null   float64 
 14  Humidity3pm  140953 non-null   float64 
 15  Pressure9am  130395 non-null   float64 
 16  Pressure3pm  130432 non-null   float64 
 17  Cloud9am     89572 non-null   float64 
 18  Cloud3pm     86102 non-null   float64 
 19  Temp9am      143693 non-null   float64 
 20  Temp3pm      141851 non-null   float64 
 21  RainToday    142199 non-null   object  
 22  RainTomorrow 142193 non-null   object  
dtypes: float64(16), object(7)
memory usage: 25.5+ MB
```

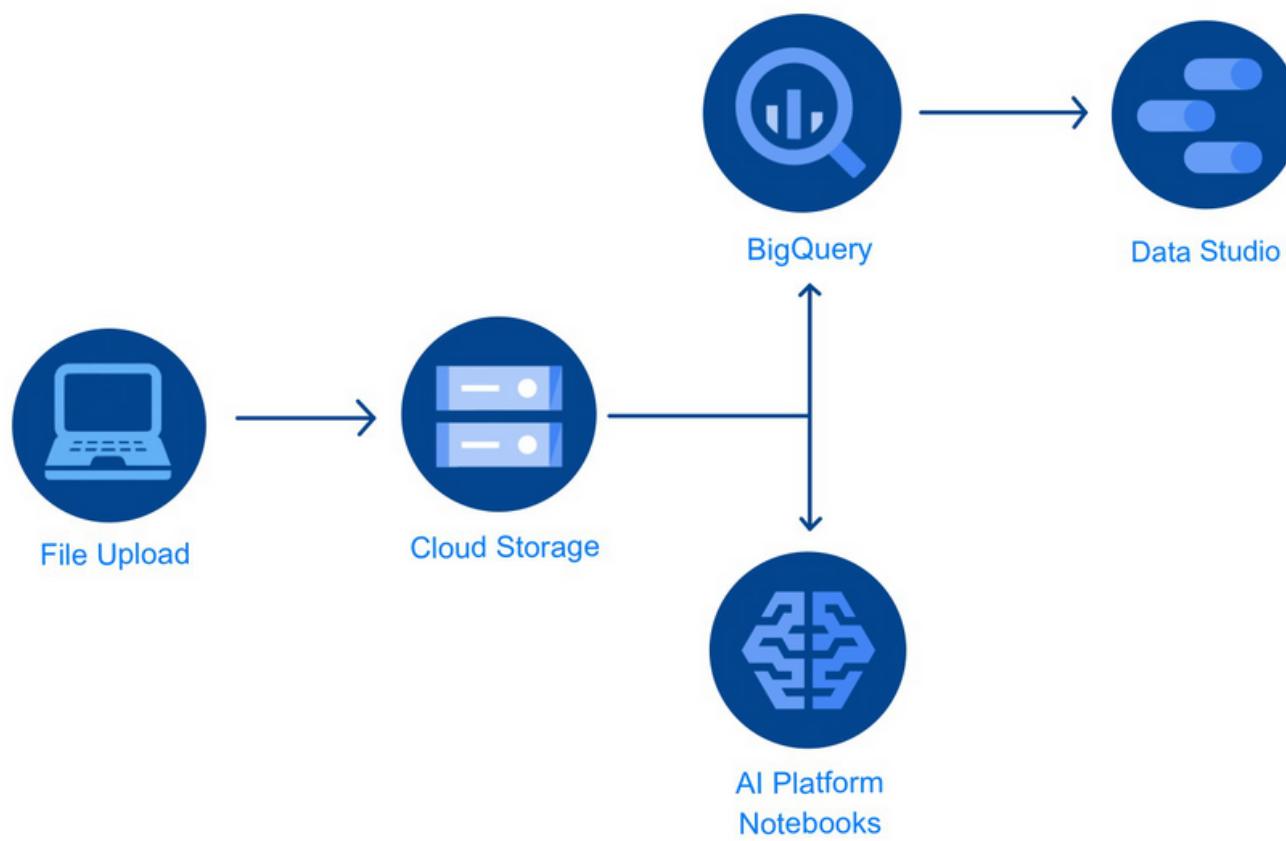
2

df

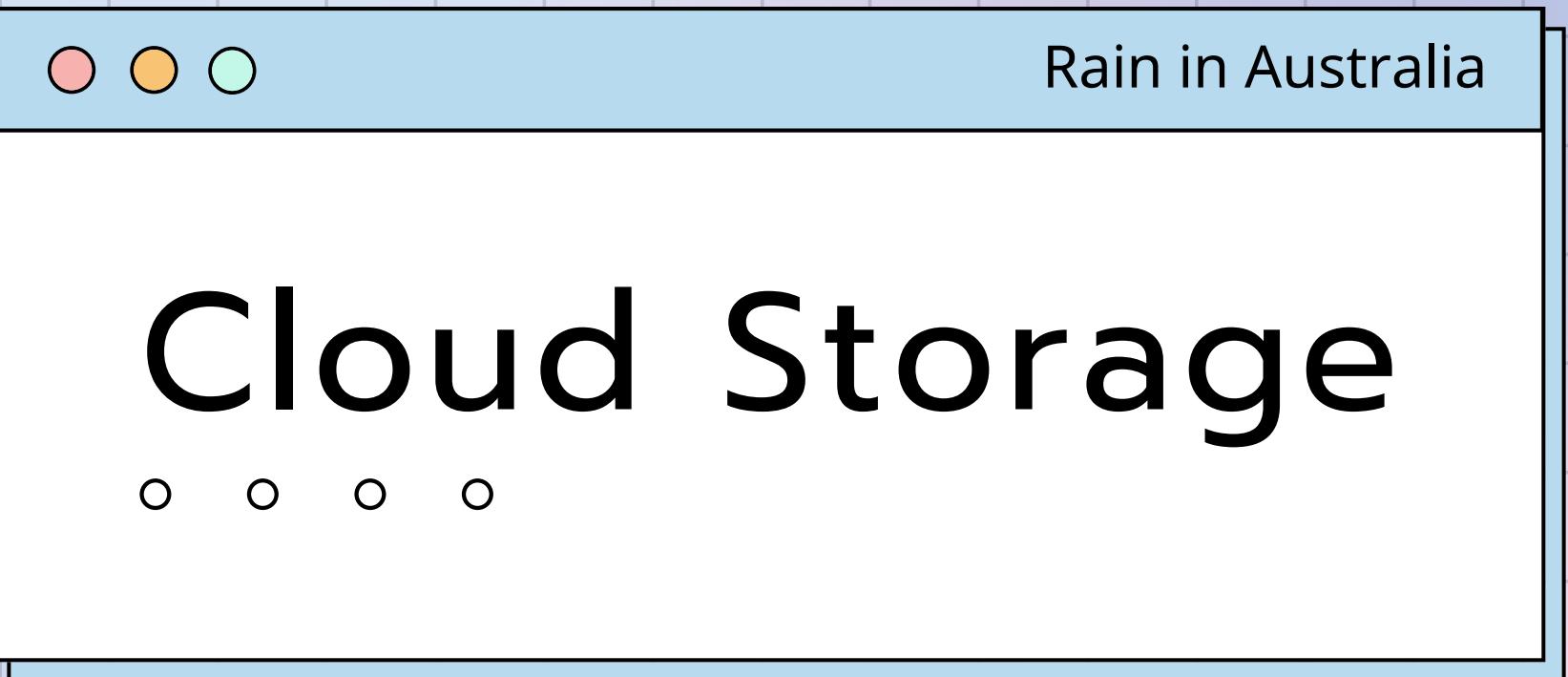
	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am
0	2008-12-01	Albury	13.4	22.9	0.6	NaN	NaN	W	44.0	W
1	2008-12-02	Albury	7.4	25.1	0.0	NaN	NaN	NNW	44.0	NNW
2	2008-12-03	Albury	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W
3	2008-12-04	Albury	9.2	28.0	0.0	NaN	NaN	NE	24.0	SE
4	2008-12-05	Albury	17.5	32.3	1.0	NaN	NaN	W	41.0	ENE
...
145455	2017-06-21	Uluru	2.8	23.4	0.0	NaN	NaN	E	31.0	SE
145456	2017-06-22	Uluru	3.6	25.3	0.0	NaN	NaN	NNW	22.0	SE

ສາມາດ ຮະບົບເປື້ອງຕັນ

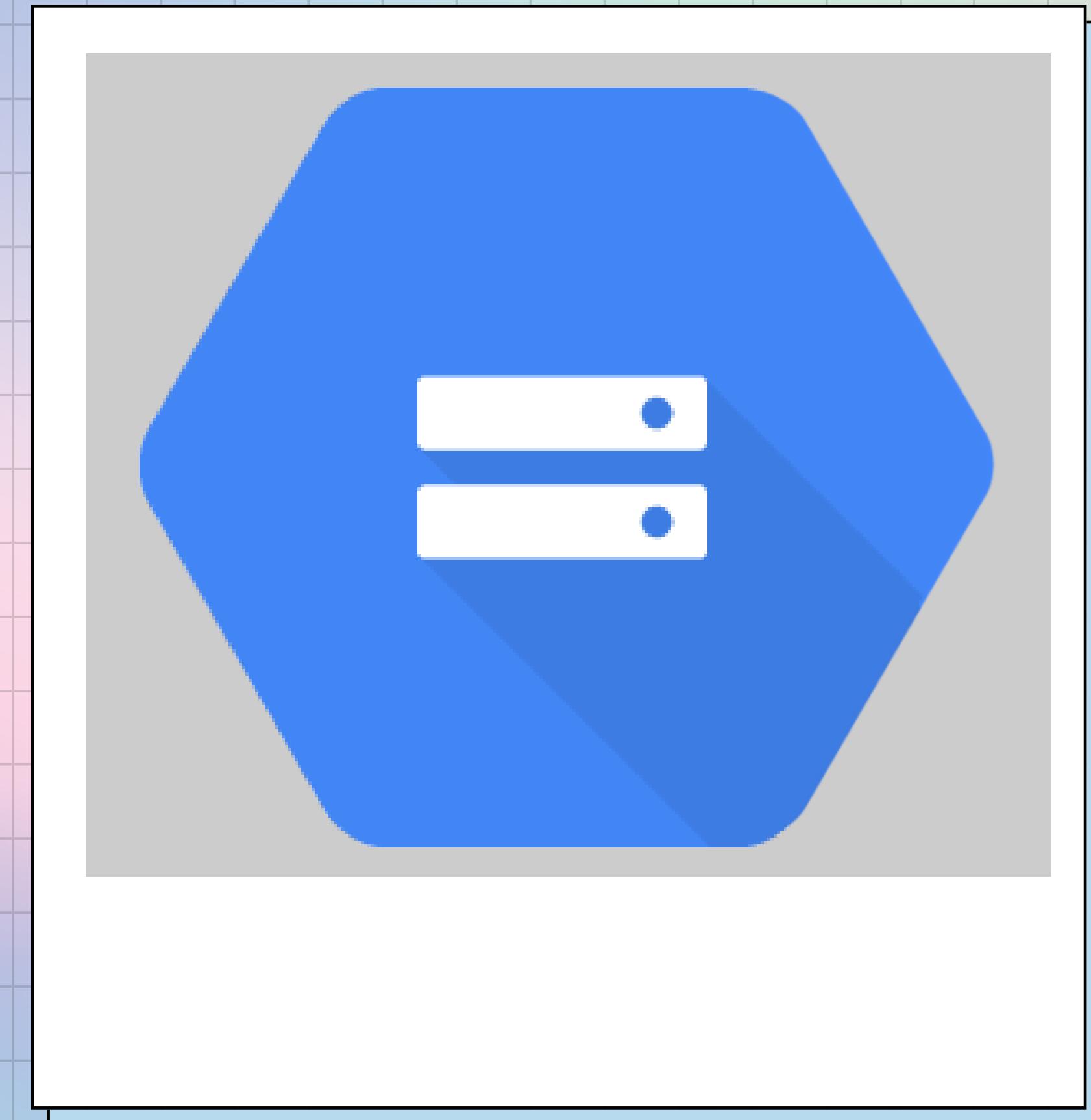
○ ○ ○ ○

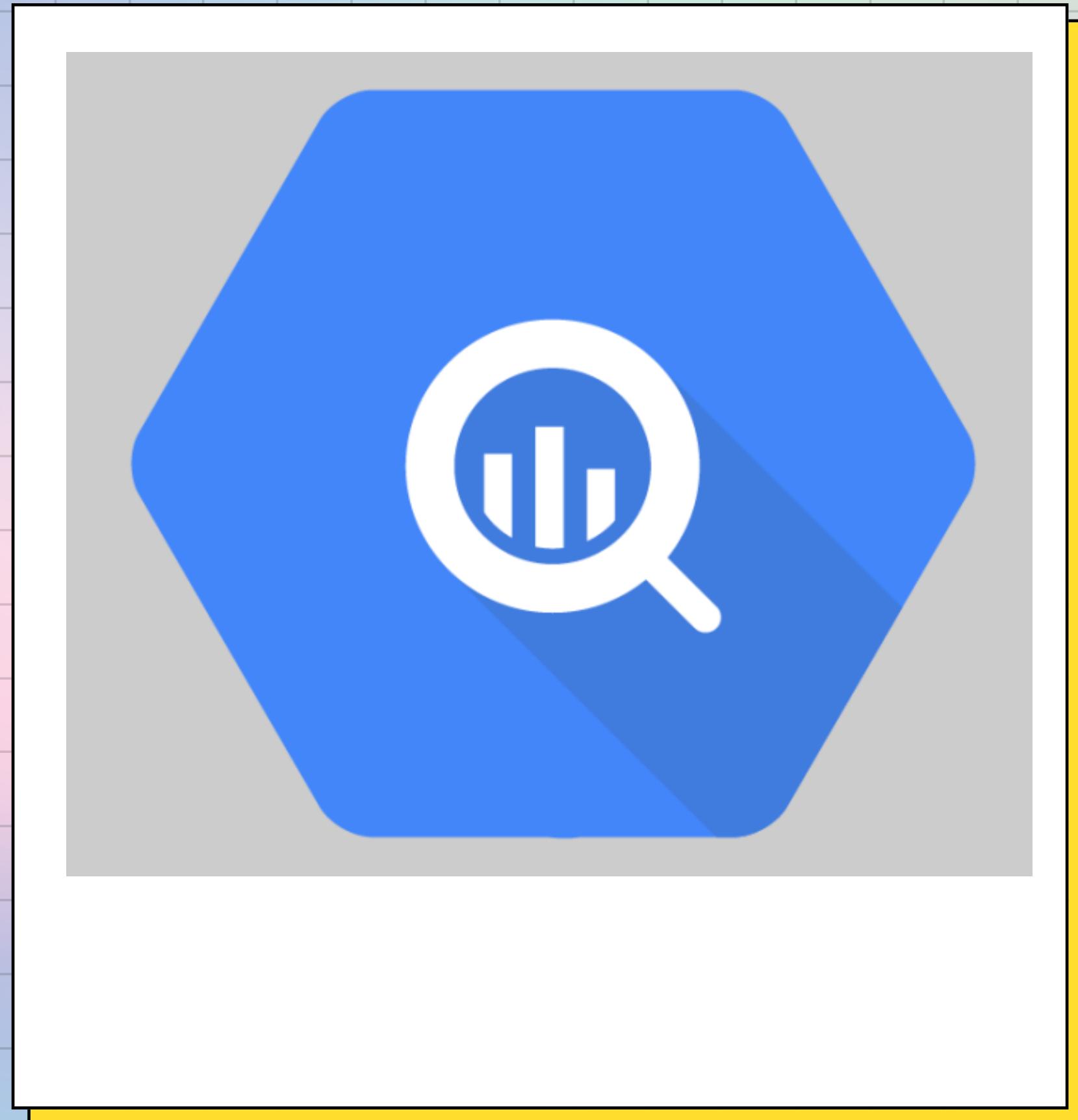
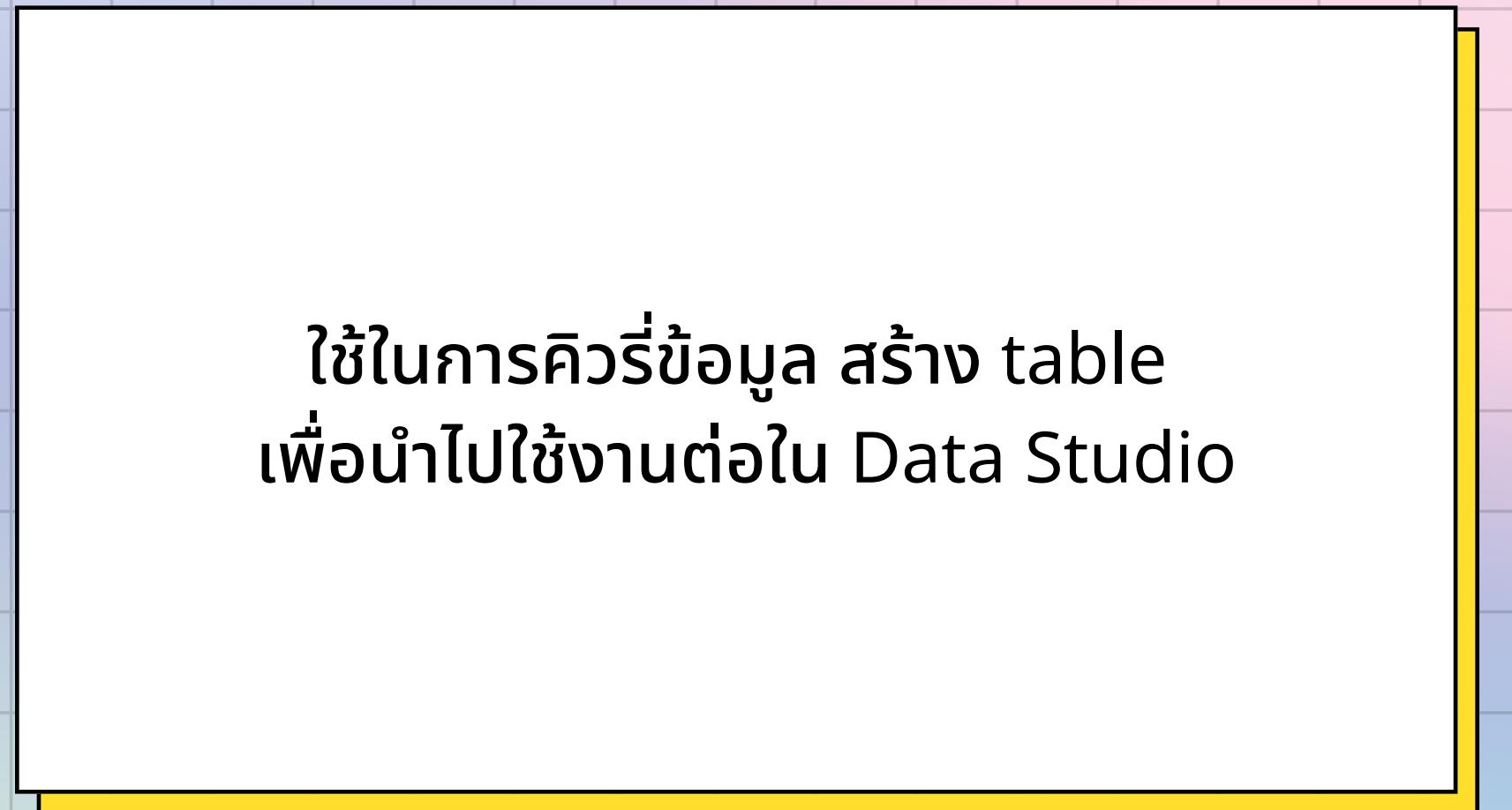
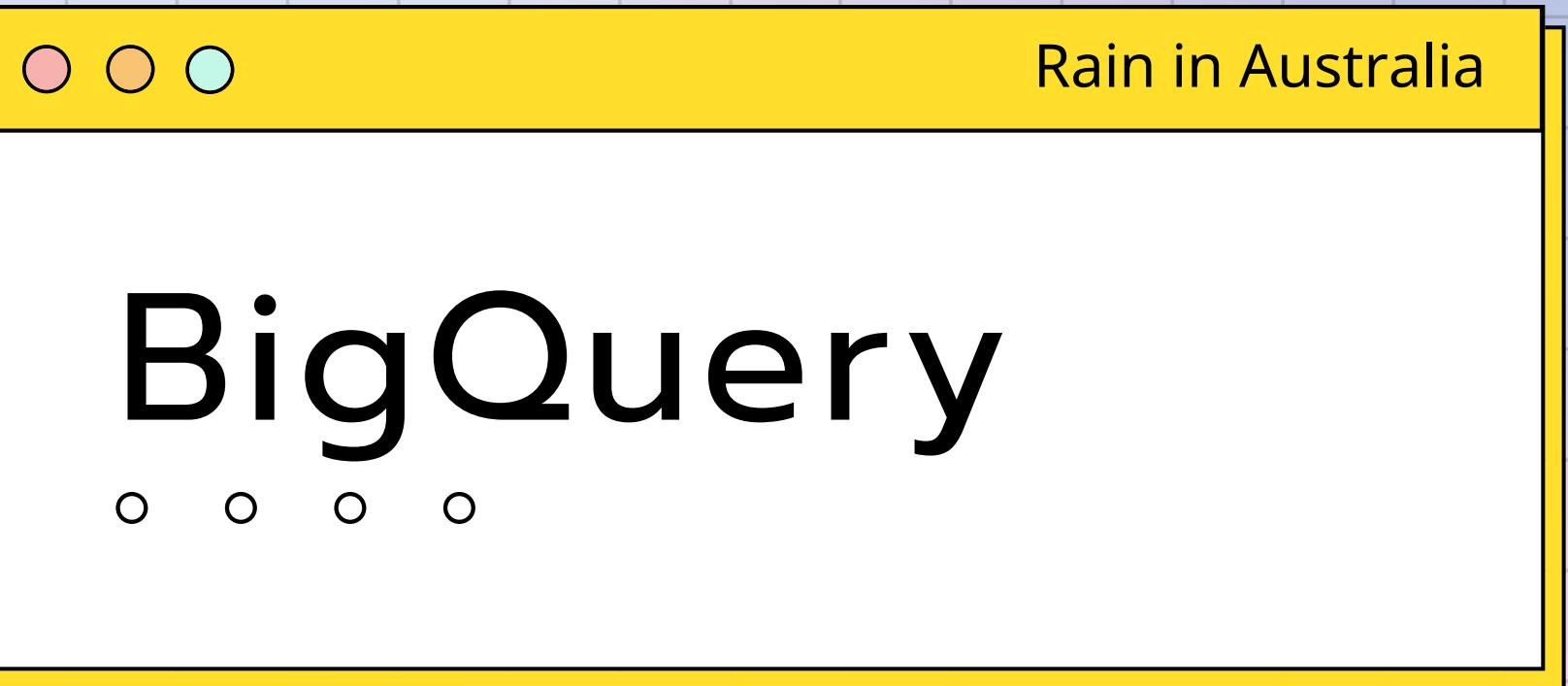


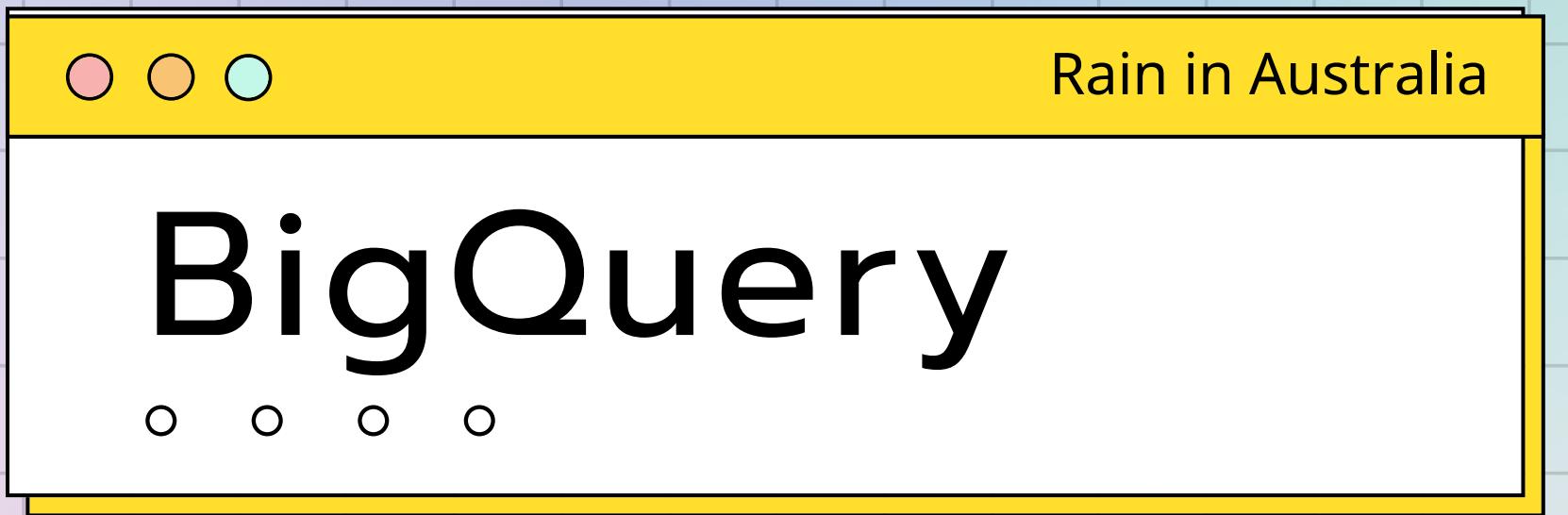
- 1 Cloud Storage
- 2 BigQuery
- 3 Data Studio
- 3 Cloud AI Platform Notebooks



ใช้ในการเก็บไฟล์ข้อมูล
"weatherAUS.csv"
ลงใน Bucket ชื่อว่า "cs358-finalproj"
folder "data/"







Explorer + ADD DATA

Type to search

Viewing pinned projects.

cs358-finalproj

rain

corr

explore

CORR EXPLORE

explore

SCHEMA DETAILS PREVIEW

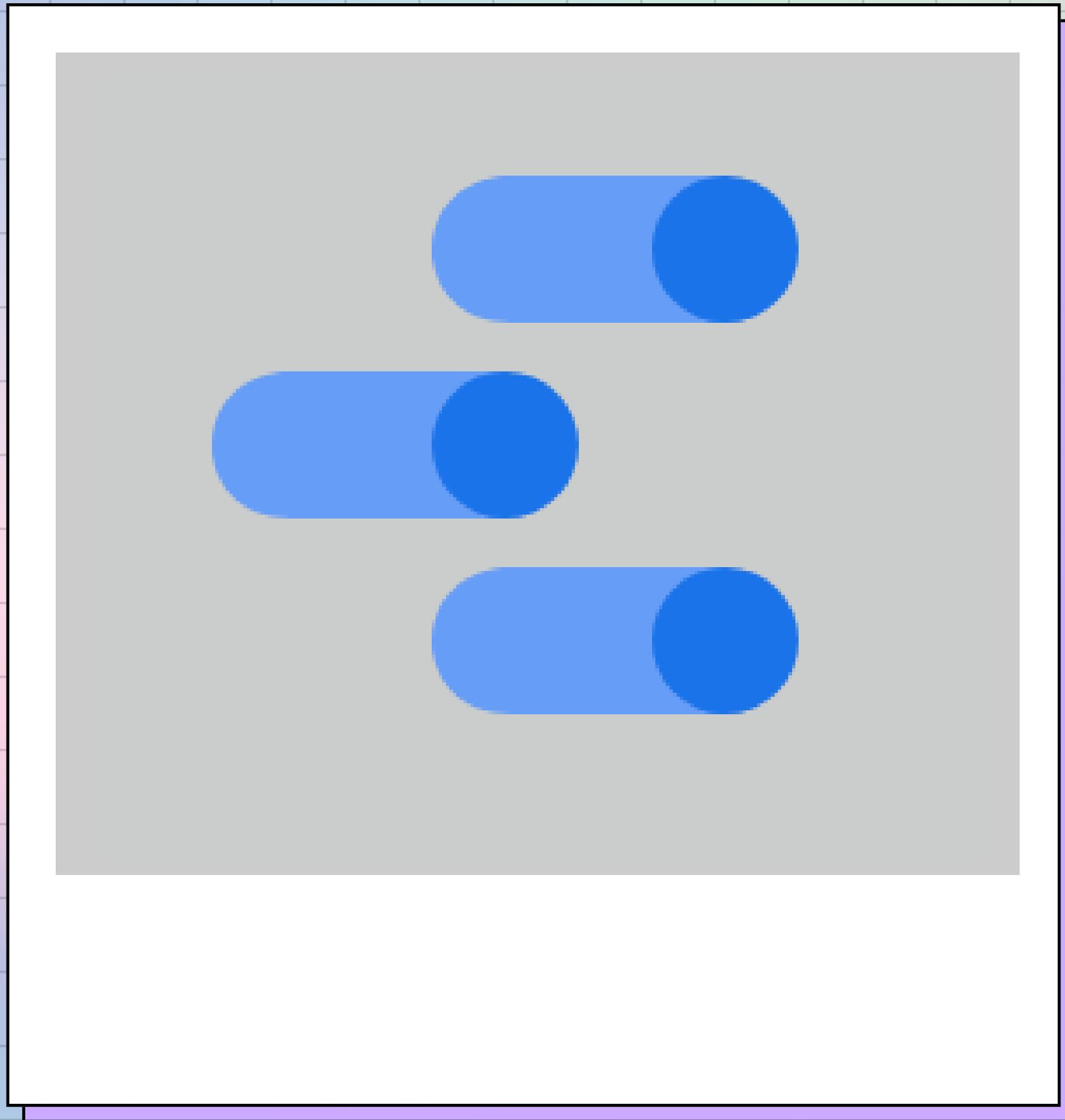
Table schema

Filter Enter property name or value

Field name	Type	Mode	Policy Tags	Description
Location	STRING	NULLABLE		
MinTemp	FLOAT	NULLABLE		
MaxTemp	FLOAT	NULLABLE		

1. Table ชื่อ “explore”
นำไป explore data ที่มีเบื้องต้นได้ โดยการสร้างกราฟใน Data Studio

2. Table ชื่อ “corr”
เป็น table เก็บค่า correlation ระหว่างตัวแปรแต่ละตัว



ใช้ในการสร้าง Dashboard
ที่นำเสนอประกอบการวิเคราะห์ต่างๆ



Rain in Australia

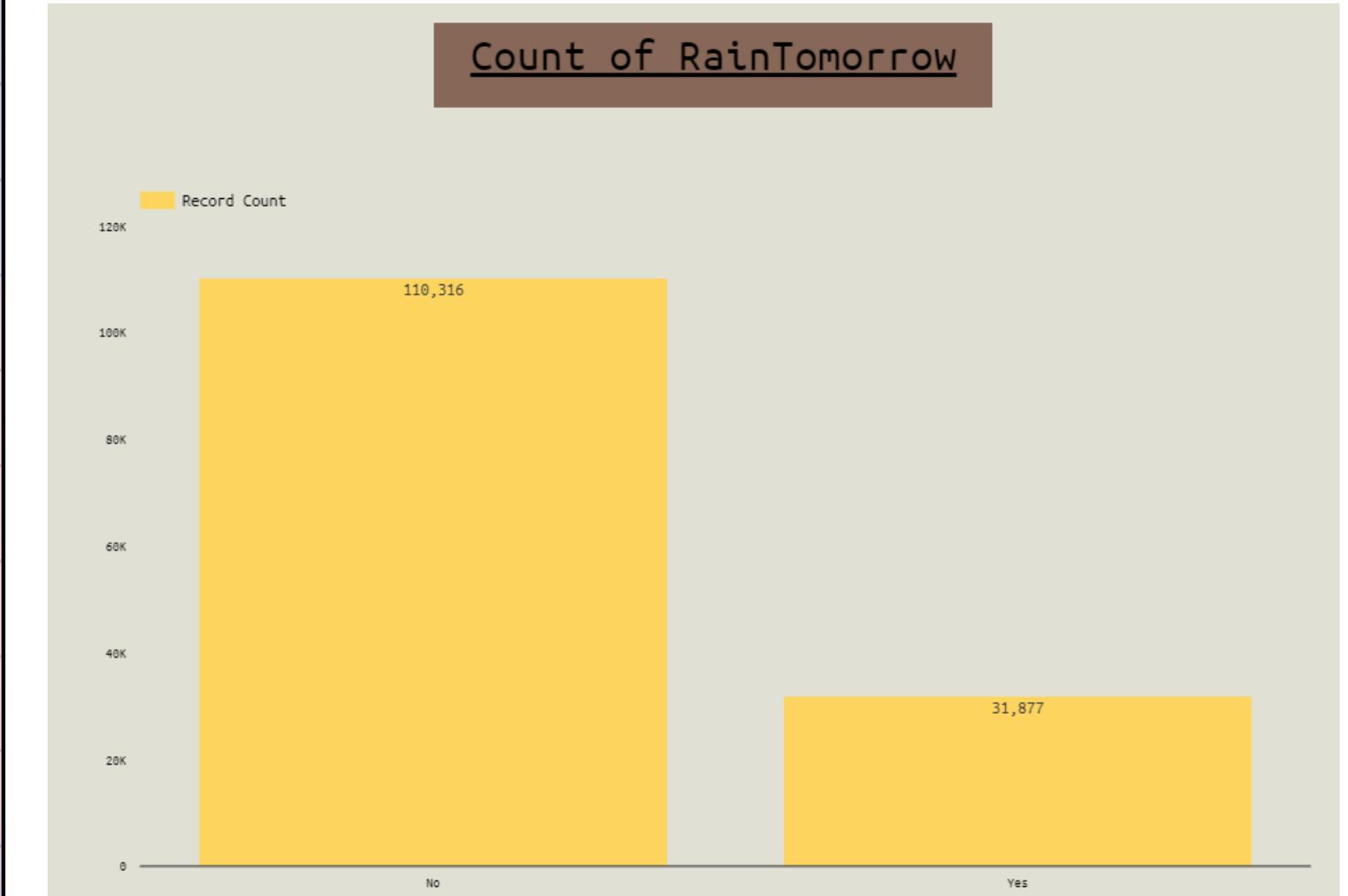
Data Studio

○ ○ ○ ○

1. กราฟ Count of RainTomorrow

เพื่อแสดงจำนวนที่มีและรูปแบบของตัวแปรที่สนใจ
(ตัวแปร Y) ในที่นี้มีค่า Yes เท่ากับ 31,877 ค่า
และมีค่า No เท่ากับ 110,316 ค่า

Count of RainTomorrow





Rain in Australia

Data Studio



2. กราฟ Correlation Heatmap of Rain in Australia Dataset

เพื่อแสดงความสัมพันธ์ระหว่างตัวแปรแต่ละตัว
และทำให้เห็นภาพชัดขึ้นด้วยสีของ heatmap

Correlation Heatmap of Rain in Australia Dataset

Name	Cloud3pm	Cloud9am	Day	Evaporation	Humidity3pm	Humidity9am	MaxTemp	MinTemp	Month
Cloud3pm	1	0.6	-0	-0.18	0.52	0.36	-0.28	0.02	-0
Cloud9am	0.6	1	0.01	-0.18	0.52	0.45	-0.29	0.08	-0.01
Day	-0	0.01	1	-0.01	0.01	0.02	0	0	0.01
Evaporation	-0.18	-0.18	-0.01	1	-0.39	-0.5	0.59	0.47	-0.03
Humidity3pm	0.52	0.52	0.01	-0.39	1	0.67	-0.51	0.01	-0.02
Humidity9am	0.36	0.45	0.02	-0.5	0.67	1	-0.5	-0.23	-0.09
MaxTemp	-0.28	-0.29	0	0.59	-0.51	-0.5	1	0.74	-0.16
MinTemp	0.02	0.08	0	0.47	0.01	-0.23	0.74	1	-0.2
Month	-0	-0.01	0.01	-0.03	-0.02	-0.09	-0.16	-0.2	1
Pressure3pm	-0.08	-0.06	-0.02	-0.29	0.05	0.19	-0.43	-0.46	0.03
Pressure9am	-0.15	-0.13	-0.02	-0.27	-0.03	0.14	-0.33	-0.45	0.03
Rainfall	0.17	0.2	0	-0.06	0.26	0.22	-0.07	0.1	-0.03
Sunshine	-0.7	-0.68	-0	0.37	-0.63	-0.49	0.47	0.07	0.02
Temp3pm	-0.32	-0.3	-0	0.57	-0.56	-0.5	0.98	0.71	-0.18
Temp9am	-0.13	-0.14	0	0.55	-0.22	-0.47	0.89	0.9	-0.14
WindGustSpe...	0.11	0.07	-0.01	0.2	-0.03	-0.22	0.07	0.18	0.06
WindSpeed3p...	0.03	0.05	-0.01	0.13	0.02	-0.15	0.05	0.18	0.06
WindSpeed9a...	0.05	0.03	-0.01	0.19	-0.03	-0.27	0.01	0.18	0.05
Year	0.04	0.07	-0.01	0.08	-0.01	0.01	0.06	0.04	-0.11



3. กราฟ Correlation Heatmap of Rain in Australia Dataset (ต่อ)

เพื่อแสดงความสัมพันธ์ระหว่างตัวแปรแต่ละตัว
และทำให้เห็นภาพชัดขึ้นด้วยสีของ heatmap

Correlation Heatmap of Rain in Australia Dataset (ต่อ)											
Name	Pressure3pm	Pressure9am	Rainfall	Sunshine	Temp3pm	Temp9am	WindGustS...	WindSpeed...	WindSpeed...	Year	
Cloud3pm	-0.08	-0.15	0.17	-0.7	-0.32	-0.13	0.11	0.03	0.05	0.04	
Cloud9am	-0.06	-0.13	0.2	-0.68	-0.3	-0.14	0.07	0.05	0.03	0.07	
Day	-0.02	-0.02	0	-0	-0	0	-0.01	-0.01	-0.01	-0.01	
Evaporation	-0.29	-0.27	-0.06	0.37	0.57	0.55	0.2	0.13	0.19	0.08	
Humidity3pm	0.05	-0.03	0.26	-0.63	-0.56	-0.22	-0.03	0.02	-0.03	-0.01	
Humidity9am	0.19	0.14	0.22	-0.49	-0.5	-0.47	-0.22	-0.15	-0.27	0.01	
MaxTemp	-0.43	-0.33	-0.07	0.47	0.98	0.89	0.07	0.05	0.01	0.06	
MinTemp	-0.46	-0.45	0.1	0.07	0.71	0.9	0.18	0.18	0.18	0.04	
Month	0.03	0.03	-0.03	0.02	-0.18	-0.14	0.06	0.06	0.05	-0.11	
Pressure3pm	1	0.96	-0.13	-0.02	-0.39	-0.47	-0.41	-0.26	-0.18	0.02	
Pressure9am	0.96	1	-0.17	0.04	-0.29	-0.42	-0.46	-0.3	-0.23	0.03	
Rainfall	-0.13	-0.17	1	-0.23	-0.08	0.01	0.13	0.06	0.09	-0.01	
Sunshine	-0.02	0.04	-0.23	1	0.49	0.29	-0.03	0.05	0.01	0.01	
Temp3pm	-0.39	-0.29	-0.08	0.49	1	0.86	0.03	0.03	0	0.05	
Temp9am	-0.47	-0.42	0.01	0.29	0.86	1	0.15	0.16	0.13	0.05	
WindGustSpe...	-0.41	-0.46	0.13	-0.03	0.03	0.15	1	0.69	0.61	-0.03	
WindSpeed3p...	-0.26	-0.3	0.06	0.05	0.03	0.16	0.69	1	0.52	-0.03	
WindSpeed9a...	-0.18	-0.23	0.09	0.01	0	0.13	0.61	0.52	1	-0.02	
Year	0.02	0.03	-0.01	0.01	0.05	0.05	-0.03	-0.03	-0.02	1	



ใช้ในการสร้างโมเดล หรือการทำ Evaluation และการ
ปรับปรุงแก้ไขต่างๆ

ในที่นี่ทำการสร้าง Instance Name
ชื่อว่า cs358-finalproj โดยภายใน instance นี้
จะมีไฟล์ cs358-project.ipynb
ซึ่งมีกั้งหมด 15 ส่วน



- 1. Installing dependencies**
- 2. Import Library**
- 3. Import Data from Bucket**
- 4. Exploratory data analysis**
- 5. Check Seasonal of Data**
- 6. Univariate Analysis**
- 7. Bivariate Analysis**
- 8. Export Dataframe to Bigquery**
- 9. Multivariate Analysis**
- 10. Declare feature vector and target variable**
- 11. Split data into separate training and test set**
- 12. Feature Engineering**
- 13. Feature Scaling**
- 14. Model training**
- 15. Evaluation**

แผนการดำเนินงาน

○ ○ ○ ○

April 01, 2021 - June 10, 2021

ลำดับ	กิจกรรม	เม.ย.				พ.ค.				มิ.ย.	
		1	2	3	4	1	2	3	4	1	2
1	วางแผน หาข้อมูล และเป้าหมาย ที่สนใจ	↔									
2	ดำเนินการเพื่อให้ได้ผลเบื้องต้น เช่น การนำข้อมูลเข้าไปเก็บใน Storage	↔									
3	นำเสนอเค้าโครงโปรเจค	↔									
4	ปรับปรุงแก้ไข	↔	→								
5	Data Understanding			↔	→						
6	Data Preparing		↔	→							
7	Modeling เบื้องต้น			↔	→						
8	รายงานความก้าวหน้า			↔	→						
9	ปรับปรุงแก้ไข			↔	→						
10	Modeling (ต่อ)			↔	→						
11	Evaluation				↔	→					
12	ปรับปรุงแก้ไข					↔	→				
13	นำเสนอผลงานโปรเจคทั้งหมด						↔	→			

វិវេការណ៍បែងចាយ តាមអត្ថក **CRISP-DM**

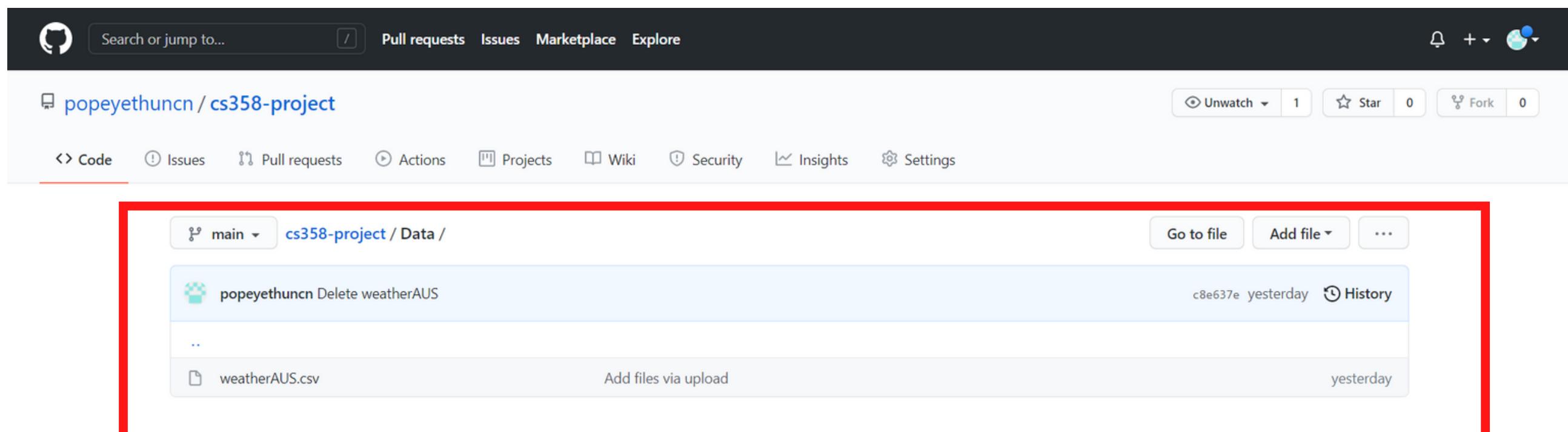
○ ○ ○ ○

1. Business Understanding

- ศึกษา Machine Learning ประเภท Supervised Learning
- Classification Model เพื่อกำหนดว่าฝนจะตกหรือไม่ในวันต่อไป
- ส่งผลให้ประชาชนสามารถวางแผนป้องกันและลดความเสียหายอุบัติเหตุ และความไม่สงบสบายนในการใช้ชีวิตได้อย่างมีประสิทธิภาพ

2. Data Understanding

- ทำการสร้าง Github เพื่อเก็บข้อมูลที่จำเป็นต้องใช้ในการทำโปรเจคนี้ทำการนำเข้าไฟล์ข้อมูล "weatherAUS.csv" ไปยัง Cloud Storage



- สร้าง Project ชื่อว่า “cs358-proj-6009680106” และได้ทำการสร้าง Bucket ที่มีชื่อว่า “cs358-finalproj” เพื่อเก็บไฟล์ข้อมูล weatherAUS.csv ลงใน Folder “data”

The screenshot shows the Google Cloud Platform Storage interface. The top navigation bar includes the project name "cs358-proj-6009680106". The main title is "Bucket details" for "cs358-finalproj". The left sidebar has options for Buckets, Objects, Configuration, Permissions, Retention, and Lifecycle. The "OBJECTS" tab is selected. The breadcrumb navigation shows "Buckets > cs358-finalproj > data". Below are buttons for UPLOAD FILES, UPLOAD FOLDER, CREATE FOLDER, MANAGE HOLDS, DOWNLOAD, and DELETE. A filter bar allows filtering by name prefix. The main table lists the file "weatherAUS.csv" with details: Name (weatherAUS.csv), Size (13.4 MB), Type (text/csv), Created time (Apr 4, 2021, 10:58...), Storage class (Standard), Last modified (Apr 4, 2021, 10:5...), Public access (Not public), Encryption (Google-managed key), Retention expiration date (None), and Holds (None). The entire table row is highlighted with a red border.

Name	Size	Type	Created time	Storage class	Last modified	Public access	Encryption	Retention expiration date	Holds
weatherAUS.csv	13.4 MB	text/csv	Apr 4, 2021, 10:58...	Standard	Apr 4, 2021, 10:5...	Not public	Google-managed key	–	None

- ວິທີການໄພລໍຂ້ອມ໌ weatherAUS.csv ໄປ່ນ Cloud Storage

```
popeyethunchanok@cloudshell:~ (cs358-proj-6009680106)$ git clone https://github.com/popeyethuncn/cs358-project
Cloning into 'cs358-project'...
remote: Enumerating objects: 14, done.
remote: Counting objects: 100% (14/14), done.
remote: Compressing objects: 100% (10/10), done.
remote: Total 14 (delta 0), reused 0 (delta 0), pack-reused 0
Unpacking objects: 100% (14/14), done.
popeyethunchanok@cloudshell:~ (cs358-proj-6009680106)$ ls
10 cs358-project data-science-on-gcp README-cloudshell.txt
popeyethunchanok@cloudshell:~ (cs358-proj-6009680106)$ cd cs358-project/
popeyethunchanok@cloudshell:~/cs358-project (cs358-proj-6009680106)$ cd Data/
popeyethunchanok@cloudshell:~/cs358-project/Data (cs358-proj-6009680106)$ ls
weatherAUS.csv
popeyethunchanok@cloudshell:~/cs358-project/Data (cs358-proj-6009680106)$ gsutil cp weatherAUS.csv gs://cs358-finalproj/data/
Copying file://weatherAUS.csv [Content-Type=text/csv]...
\ [1 files][ 13.4 MiB/ 13.4 MiB]
Operation completed over 1 objects/13.4 MiB.
```

- ทำการสร้าง Instance ที่มีชื่อว่า cs358-finalproj โดยภายใน instance นี้ เมื่อเปิดผ่าน JupyterLab แล้ว จะมีไฟล์ cs358-project.ipynb

The screenshot shows a Jupyter Notebook interface. At the top, there is a toolbar with buttons for NEW INSTANCE, REFRESH, START, STOP, RESET, and DELETE. Below the toolbar, a message states: "Create and use Jupyter Notebooks with a notebook instance. Notebook instances have JupyterLab pre-installed and are configured with GPU-enabled machine learning frameworks. [Learn more](#)". A "Filter" input field is present. The main area displays a table of notebook instances:

Instance name	Zone	Environment Version	Auto-upgrade
cs358-finalproj	asia-southeast1-a	M69	-

The "cs358-finalproj" row is highlighted with a red box. Below the table, a file browser shows a directory structure with files: src, tutorials, cs358-project.ipynb, and weatherAUS.csv. The "cs358-project.ipynb" file is selected and highlighted with a red box. In the bottom right corner, a code editor window is open, showing the following code:

```
1. Installing dependencies
[1]: %pip freeze
...
[2]: %pip install google-cloud
%pip install google-cloud-storage
%pip install pandas
%pip install git+https://github.com/pydata/pandas-gbq.git
##%pip install git+git://github.com/scikit-learn/scikit-learn.git
%pip install --upgrade category_encoders
...
```

- ការ Installing dependencies ឬទិន្នន័យ dependencies កែចាប់

1. Installing dependencies

```
[1]: %pip freeze
...
[2]: %pip install google-cloud
%pip install google-cloud-storage
%pip install pandas
%pip install git+https://github.com/pydata/pandas-gbq.git
##%pip install git+git://github.com/scikit-learn/scikit-learn.git
%pip install --upgrade category_encoders
...
[3]: %%bash
sudo apt-get update
sudo apt-get -y install python-mpltoolkits.basemap
...
...
```

- Import Library : นำเข้าไลบรารีที่จำเป็น

2. Import Library

[4]:

```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import numpy as np
from pandas.io import gbq
```

- Import Data from Bucket : นำเข้าไฟล์ข้อมูล "weatherAUS.csv" จาก Bucket ชื่อว่า "cs358-finalproj" folder "data/"

```
[31]: from google.cloud import storage
import pandas as pd

bucket_name = "cs358-finalproj"

storage_client = storage.Client()
bucket = storage_client.get_bucket(bucket_name)

# When you have your files in a subfolder of the bucket.
my_prefix = "data/" # the name of the subfolder
blobs = bucket.list_blobs(prefix = my_prefix, delimiter = '/')

for blob in blobs:
    if(blob.name != my_prefix): # ignoring the subfolder itself
        file_name = blob.name.replace(my_prefix, "")
        blob.download_to_filename(file_name) # download the file to the machine
        df = pd.read_csv(file_name) # load the data
        print(df)
```

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	\
0	2008-12-01	Albury	13.4	22.9	0.6	NaN	
1	2008-12-02	Albury	7.4	25.1	0.0	NaN	
2	2008-12-03	Albury	12.9	25.7	0.0	NaN	
3	2008-12-04	Albury	9.2	28.0	0.0	NaN	
4	2008-12-05	Albury	17.5	32.3	1.0	NaN	
...	
145455	2017-06-21	Uluru	2.8	23.4	0.0	NaN	
145456	2017-06-22	Uluru	3.6	25.3	0.0	NaN	
145457	2017-06-23	Uluru	5.4	26.9	0.0	NaN	
145458	2017-06-24	Uluru	7.8	27.0	0.0	NaN	
145459	2017-06-25	Uluru	14.9	NaN	0.0	NaN	

- Exploratory data analysis : ทำการ explore ข้อมูลเบื้องต้น
- โดยมีตัวแปรที่เป็น object ทั้งหมด 7 ตัว และตัวแปร Float 16 ตัว

4. Exploratory data analysis

```
[6]: df
```

```
[6]:
```

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	...
0	2008-12-01	Albury	13.4	22.9	0.6	NaN	NaN	W	44.0	W	...
1	2008-12-02	Albury	7.4	25.1	0.0	NaN	NaN	WNW	44.0	NNW	...
2	2008-12-03	Albury	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	...
3	2008-12-04	Albury	9.2	28.0	0.0	NaN	NaN	NE	24.0	SE	...
4	2008-12-05	Albury	17.5	32.3	1.0	NaN	NaN	W	41.0	ENE	...
...

```
[7]: df.describe(include='all')
```

```
[7]:
```

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	...
count	145460	145460	143975.000000	144199.000000	142199.000000	82670.000000	75625.000000	135134	135197.000000	134894	...
unique	3436	49	NaN	NaN	NaN	NaN	NaN	16	NaN	16	...
top	2016-08-16	Canberra	NaN	NaN	NaN	NaN	NaN	W	NaN	N	...
freq	49	3436	NaN	NaN	NaN	NaN	NaN	9915	NaN	11758	...
mean	NaN	NaN	12.194034	23.221348	2.360918	5.468232	7.611178	NaN	40.035230	NaN	...
std	NaN	NaN	6.398495	7.119049	8.478060	4.193704	3.785483	NaN	13.607062	NaN	...
min	NaN	NaN	-8.500000	-4.800000	0.000000	0.000000	0.000000	NaN	6.000000	NaN	...
25%	NaN	NaN	7.600000	17.900000	0.000000	2.600000	4.800000	NaN	31.000000	NaN	...
50%	NaN	NaN	12.000000	22.600000	0.000000	4.800000	8.400000	NaN	39.000000	NaN	...
75%	NaN	NaN	16.900000	28.200000	0.800000	7.400000	10.600000	NaN	48.000000	NaN	...
max	NaN	NaN	33.900000	48.100000	371.000000	145.000000	14.500000	NaN	135.000000	NaN	...

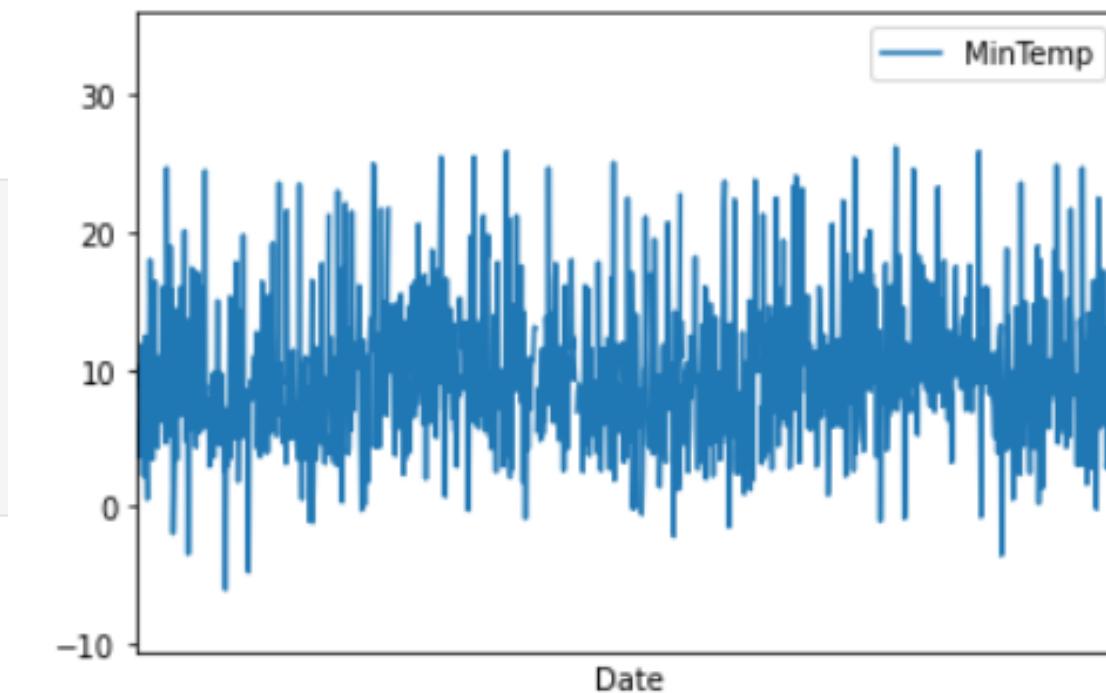
```
[8]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 145460 entries, 0 to 145459
Data columns (total 23 columns):
 #   Column        Non-Null Count  Dtype  
--- 
 0   Date          145460 non-null   object 
 1   Location      145460 non-null   object 
 2   MinTemp       143975 non-null   float64
 3   MaxTemp       144199 non-null   float64
 4   Rainfall      142199 non-null   float64
 5   Evaporation   82670 non-null   float64
 6   Sunshine      75625 non-null   float64
 7   WindGustDir   135134 non-null   object 
 8   WindGustSpeed 135197 non-null   float64
 9   WindDir9am    134894 non-null   object 
 10  WindDir3pm    141232 non-null   object 
 11  WindSpeed9am  143693 non-null   float64
 12  WindSpeed3pm  142398 non-null   float64
 13  Humidity9am   142806 non-null   float64
 14  Humidity3pm   140953 non-null   float64
 15  Pressure9am   130395 non-null   float64
 16  Pressure3pm   130432 non-null   float64
 17  Cloud9am      89572 non-null   float64
 18  Cloud3pm      86102 non-null   float64
 19  Temp9am       143693 non-null   float64
 20  Temp3pm       141851 non-null   float64
 21  RainToday     142199 non-null   object 
 22  RainTomorrow  142193 non-null   object 
dtypes: float64(16), object(7)
memory usage: 25.5+ MB
```

- Check Seasonal of Data : ทำการเช็คตัวแปรแต่ละตัวว่ามี seasonal หรือไม่ เพื่อใช้ในการตัดสินใจในการเลือกวิธีจัดการกับข้อมูลในขั้นตอนการทำ Feature Engineering

5. Check Seasonal of Data

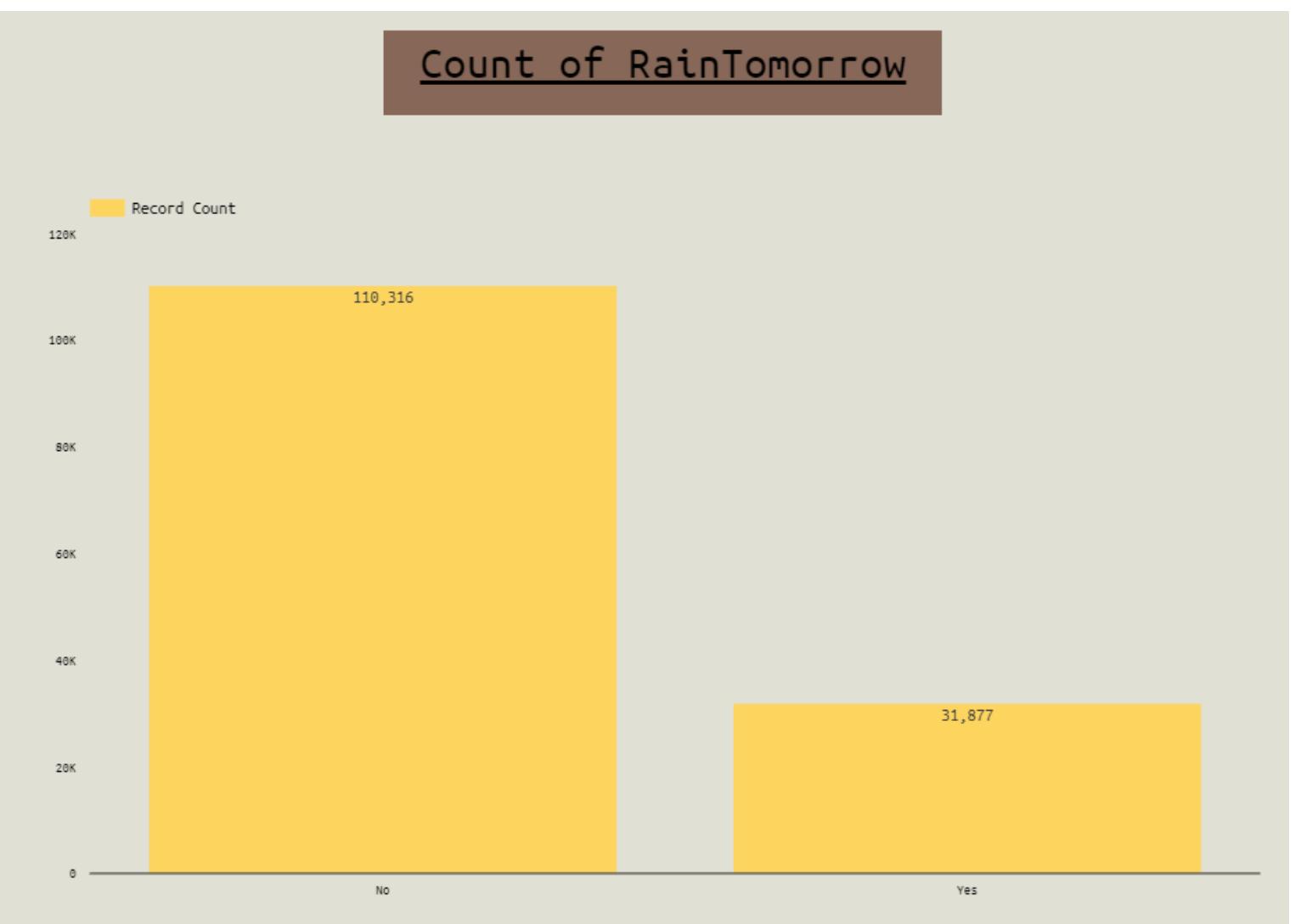
```
[9]: import datetime
numerical = [var for var in df.columns if df[var].dtype != 'O']
print(numerical)
dfsort = df.sort_values('Date')
dfsort.plot('Date', 'MinTemp')
plt.xticks(rotation=60)
plt.xlim([datetime.date(2010, 1, 20), datetime.date(2013, 1, 28)])
```



- Univariate Analysis : ทำการวิเคราะห์ตัวแปร RainTomorrow ซึ่งเป็นตัวแปรที่เราสนใจ โดยมีค่า Missing Value เท่ากับ 3267 ค่า Yes 31877 ค่า และ No 110316 ค่า

```
[38]: df['RainTomorrow'].isnull().sum()  
[38]: 3267  
  
[39]: df['RainTomorrow'].unique()  
[39]: array(['No', 'Yes', nan], dtype=object)  
  
[40]: df['RainTomorrow'].value_counts()  
[40]: No      110316  
      Yes     31877  
      Name: RainTomorrow, dtype: int64  
  
[41]: df['RainTomorrow'].value_counts()/len(df)  
[41]: No      0.758394  
      Yes     0.219146  
      Name: RainTomorrow, dtype: float64
```

- ทำการ plot graph ใน Data Studio เพื่อให้เห็นภาพของตัวแปร RainTomorrow ชัดขึ้น



- Bivariate Analysis : ทำการวิเคราะห์ตัวแปรต่างๆ โดยแบ่งเป็นตัวแปร Categorical และตัวแปร Numerical

```
[17]: # check missing values in categorical variables
```

```
df[categorical].isnull().sum()
```

```
[17]: Date          0  
Location        0  
WindGustDir    10326  
WindDir9am     10566  
WindDir3pm      4228  
RainToday       3261  
RainTomorrow    3267  
dtype: int64
```

```
[24]: # check missing values in numerical variables
```

```
df[numerical].isnull().sum()
```

```
[24]: MinTemp        1485  
MaxTemp         1261  
Rainfall        3261  
Evaporation    62790  
Sunshine        69835  
WindGustSpeed   10263  
WindSpeed9am    1767  
WindSpeed3pm    3062  
Humidity9am     2654  
Humidity3pm     4507  
Pressure9am     15065  
Pressure3pm     15028  
Cloud9am        55888  
Cloud3pm        59358  
Temp9am         1767  
Temp3pm         3609  
Year             0  
Month            0  
Day              0  
dtype: int64
```

- การแปลง field “Date” ให้แยกเป็น Date Month Year

```
[18]: # check for cardinality in categorical variables

for var in categorical:

    print(var, ' contains ', len(df[var].unique()), ' labels')

Date contains 3436 labels
Location contains 49 labels
WindGustDir contains 17 labels
WindDir9am contains 17 labels
WindDir3pm contains 17 labels
RainToday contains 3 labels
RainTomorrow contains 3 labels
```

```
[19]: #Feature Engineering of Date Variable

# parse the dates, currently coded as strings, into datetime format
df['Date'] = pd.to_datetime(df['Date'])

# extract year from date
df['Year'] = df['Date'].dt.year

# extract month from date
df['Month'] = df['Date'].dt.month

# extract day from date
df['Day'] = df['Date'].dt.day
```

```
[21]: # drop the original Date variable

df.drop('Date', axis=1, inplace = True)

df.head()
```

```
[21]:   e WindGustDir WindGustSpeed WindDir9am WindDir3pm ... Pressure3pm Cloud9am Cloud3pm Temp9am Temp3pm RainToday RainTomorrow Year Month Day
0   N        W       44.0        W      WNW   ...     1007.1      8.0      NaN    16.9    21.8      No      No  2008     12      1
1   N       WNW       44.0       NNW      WSW   ...     1007.8      NaN      NaN    17.2    24.3      No      No  2008     12      2
2   N       WSW       46.0        W      WSW   ...     1008.7      NaN      2.0    21.0    23.2      No      No  2008     12      3
3   N        NE       24.0        SE        E   ...     1012.8      NaN      NaN    18.1    26.5      No      No  2008     12      4
4   N        W       41.0       ENE       NW   ...     1006.0      7.0      8.0    17.8    29.7      No      No  2008     12      5
```

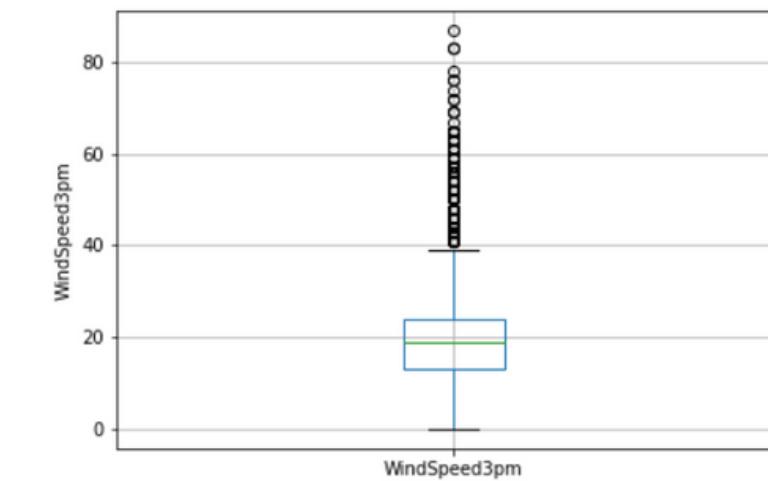
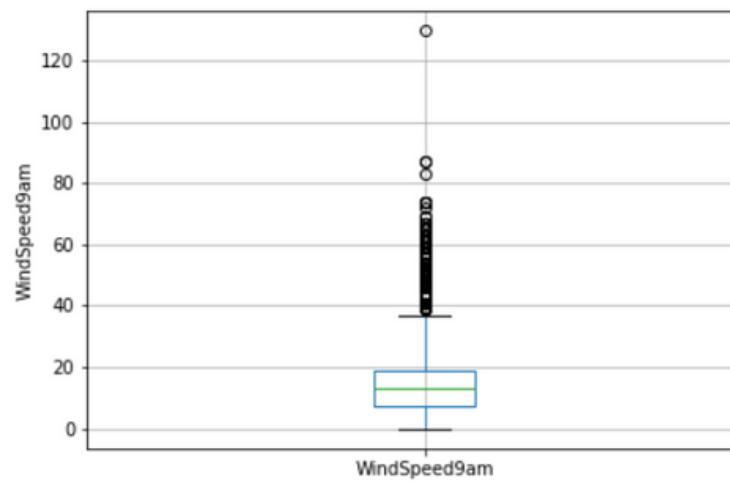
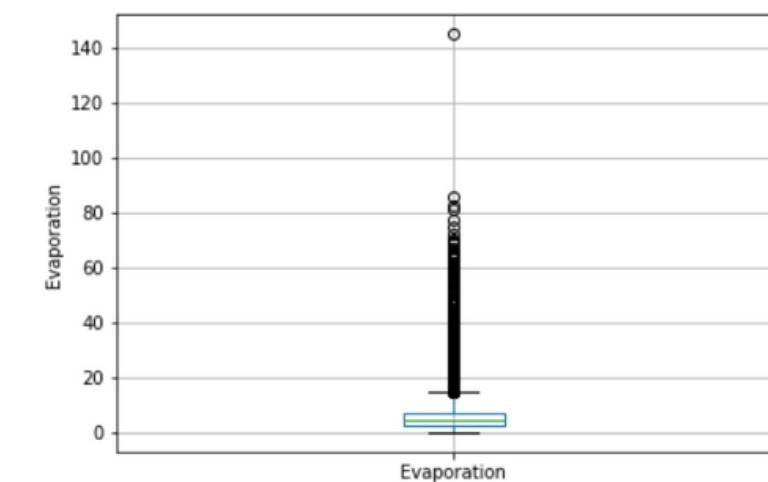
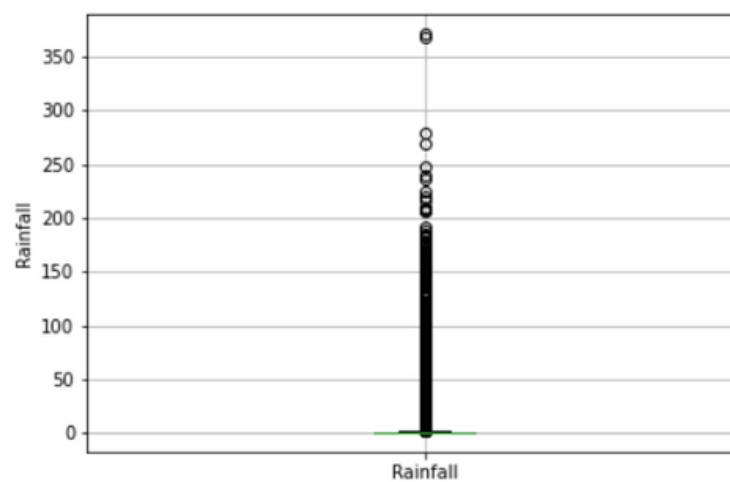
- ทำการตรวจสอบ Outliner ในแต่ละตัวตัวแปรโดยเริ่มดูจากตาราง Descriptive Statistics

```
[25]: #Outliers in numerical variables  
#Rainfall, Evaporation, WindSpeed9am, WindSpeed3pm  
  
print(round(df[numerical].describe()),2)
```

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed
count	143975.0	144199.0	142199.0	82670.0	75625.0	135197.0
mean	12.0	23.0	2.0	5.0	8.0	40.0
std	6.0	7.0	8.0	4.0	4.0	14.0
min	-8.0	-5.0	0.0	0.0	0.0	6.0
25%	8.0	18.0	0.0	3.0	5.0	31.0
50%	12.0	23.0	0.0	5.0	8.0	39.0
75%	17.0	28.0	1.0	7.0	11.0	48.0
max	34.0	48.0	371.0	145.0	14.0	135.0

	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am
count	143693.0	142398.0	142806.0	140953.0	130395.0
mean	14.0	19.0	69.0	52.0	1018.0
std	9.0	9.0	19.0	21.0	7.0
min	0.0	0.0	0.0	0.0	980.0
25%	7.0	13.0	57.0	37.0	1013.0
50%	13.0	19.0	70.0	52.0	1018.0
75%	19.0	24.0	83.0	66.0	1022.0
max	130.0	87.0	100.0	100.0	1041.0

- จึงทำการตรวจสอบ Box Plot เพื่อให้แน่ใจว่าตัวแปรดังกล่าวมีค่า Outliner



- ทำการหาขอบเขตของ Outliner เพื่อกำหนดการจัดการกับค่าเหล่านั้นต่อไป

```
[28]: # find outliers for Rainfall variable
IQR = df.Rainfall.quantile(0.75) - df.Rainfall.quantile(0.25)
Lower_fence = df.Rainfall.quantile(0.25) - (IQR * 3)
Upper_fence = df.Rainfall.quantile(0.75) + (IQR * 3)
print('Rainfall outliers are values < {lowerboundary} or > {upperboundary}'.format(lowerboundary=Lower_fence, upperboundary=Upper_fence))

# find outliers for Evaporation variable
IQR = df.Evaporation.quantile(0.75) - df.Evaporation.quantile(0.25)
Lower_fence = df.Evaporation.quantile(0.25) - (IQR * 3)
Upper_fence = df.Evaporation.quantile(0.75) + (IQR * 3)
print('Evaporation outliers are values < {lowerboundary} or > {upperboundary}'.format(lowerboundary=Lower_fence, upperboundary=Upper_fence))

# find outliers for WindSpeed9am variable
IQR = df.WindSpeed9am.quantile(0.75) - df.WindSpeed9am.quantile(0.25)
Lower_fence = df.WindSpeed9am.quantile(0.25) - (IQR * 3)
Upper_fence = df.WindSpeed9am.quantile(0.75) + (IQR * 3)
print('WindSpeed9am outliers are values < {lowerboundary} or > {upperboundary}'.format(lowerboundary=Lower_fence, upperboundary=Upper_fence))

# find outliers for WindSpeed3pm variable
IQR = df.WindSpeed3pm.quantile(0.75) - df.WindSpeed3pm.quantile(0.25)
Lower_fence = df.WindSpeed3pm.quantile(0.25) - (IQR * 3)
Upper_fence = df.WindSpeed3pm.quantile(0.75) + (IQR * 3)
print('WindSpeed3pm outliers are values < {lowerboundary} or > {upperboundary}'.format(lowerboundary=Lower_fence, upperboundary=Upper_fence))

Rainfall outliers are values < -2.4000000000000004 or > 3.2
Evaporation outliers are values < -11.800000000000002 or > 21.800000000000004
WindSpeed9am outliers are values < -29.0 or > 55.0
WindSpeed3pm outliers are values < -20.0 or > 57.0
```

- Rainfall มีขอบเขตเก่ากับ (-2.4 , 3.2)
- Evaporation มีขอบเขตเก่ากับ (-11.8 , 21.8)
- WindSpeed9am มีขอบเขตเก่ากับ (-29.0, 55.0)
- WindSpeed3pm มีขอบเขตเก่ากับ (-20.0, 57.0)

- Export Dataframe to Bigquery : เพื่อนำไปใช้สร้างกราฟใน Data Studio

8. Export Dataframe to Bigquery

for plot graph in Data Studio

```
[34]: #table explore
df.to_gbq(destination_table='rain.explore',project_id='cs358-finalproj',if_exists='replace')
```

1it [00:19, 19.68s/it]

```
[35]: #table correlation
name = list(correlation.columns)
correlation['Name'] = name
correlation.to_gbq(destination_table='rain.corr',project_id='cs358-finalproj',if_exists='replace')
```

1it [00:04, 4.73s/it]

- Multivariate Analysis : ทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรแต่ละตัว โดยใช้ Correlation ในการวิเคราะห์

Correlation Heatmap of Rain in Australia Dataset

Name	Cloud3pm	Cloud9am	Day	Evaporation	Humidity3pm	Humidity9am	MaxTemp	MinTemp	Month
Cloud3pm	1	0.6	-0	-0.18	0.52	0.36	-0.28	0.02	-0
Cloud9am	0.6	1	0.01	-0.18	0.52	0.45	-0.29	0.08	-0.01
Day	-0	0.01	1	-0.01	0.01	0.02	0	0	0.01
Evaporation	-0.18	-0.18	-0.01	1	-0.39	-0.5	0.59	0.47	-0.03
Humidity3pm	0.52	0.52	0.01	-0.39	1	0.67	-0.51	0.01	-0.02
Humidity9am	0.36	0.45	0.02	-0.5	0.67	1	-0.5	-0.23	-0.09
MaxTemp	-0.28	-0.29	0	0.59	-0.51	-0.5	1	0.74	-0.16
MinTemp	0.02	0.08	0	0.47	0.01	-0.23	0.74	1	-0.2
Month	-0	-0.01	0.01	-0.03	-0.02	-0.09	-0.16	-0.2	1
Pressure3pm	-0.08	-0.06	-0.02	-0.29	0.05	0.19	-0.43	-0.46	0.03
Pressure9am	-0.15	-0.13	-0.02	-0.27	-0.03	0.14	-0.33	-0.45	0.03
Rainfall	0.17	0.2	0	-0.06	0.26	0.22	-0.07	0.1	-0.03
Sunshine	-0.7	-0.68	-0	0.37	-0.63	-0.49	0.47	0.07	0.02
Temp3pm	-0.32	-0.3	-0	0.57	-0.56	-0.5	0.98	0.71	-0.18
Temp9am	-0.13	-0.14	0	0.55	-0.22	-0.47	0.89	0.9	-0.14
WindGustSpeed	0.11	0.07	-0.01	0.2	-0.03	-0.22	0.07	0.18	0.06
WindSpeed3pm	0.03	0.05	-0.01	0.13	0.02	-0.15	0.05	0.18	0.06
WindSpeed9am	0.05	0.03	-0.01	0.19	-0.03	-0.27	0.01	0.18	0.05
Year	0.04	0.07	-0.01	0.08	-0.01	0.01	0.06	0.04	-0.11

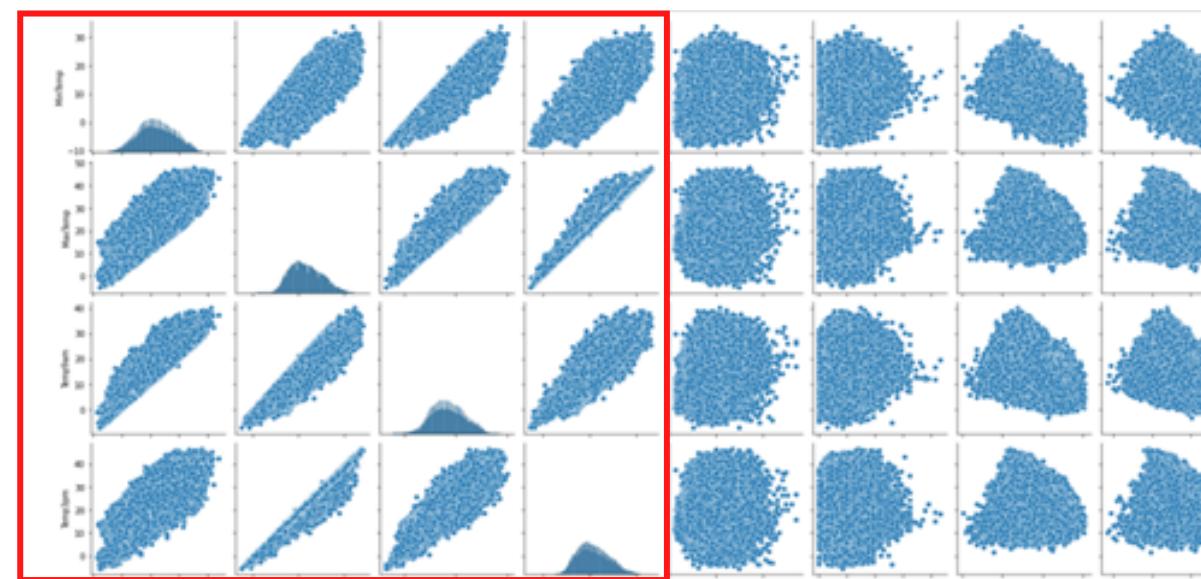
Correlation Heatmap of Rain in Australia Dataset (ต่อ)

Name	Pressure3pm	Pressure9am	Rainfall	Sunshine	Temp3pm	Temp9am	WindGustSpeed	WindSpeed3pm	WindSpeed9am	Year
Cloud3pm	-0.08	-0.15	0.17	-0.7	-0.32	-0.13	0.11	0.03	0.05	0.04
Cloud9am	-0.06	-0.13	0.2	-0.68	-0.3	-0.14	0.07	0.05	0.03	0.07
Day	-0.02	-0.02	0	-0	-0	0	-0.01	-0.01	-0.01	-0.01
Evaporation	-0.29	-0.27	-0.06	0.37	0.57	0.55	0.2	0.13	0.19	0.08
Humidity3pm	0.05	-0.03	0.26	-0.63	-0.56	-0.22	-0.03	0.02	-0.03	-0.01
Humidity9am	0.19	0.14	0.22	-0.49	-0.5	-0.47	-0.22	-0.15	-0.27	0.01
MaxTemp	-0.43	-0.33	-0.07	0.47	0.98	0.89	0.07	0.05	0.01	0.06
MinTemp	-0.46	-0.45	0.1	0.07	0.71	0.9	0.18	0.18	0.04	
Month	0.03	0.03	-0.03	0.02	-0.18	-0.14	0.06	0.06	0.05	-0.11
Pressure3pm	1	0.96	-0.13	-0.02	-0.39	-0.47	-0.41	-0.26	-0.18	0.02
Pressure9am	0.96	1	-0.17	0.04	-0.29	-0.42	-0.46	-0.3	-0.23	0.03
Rainfall	-0.13	-0.17	1	-0.23	-0.08	0.01	0.13	0.06	0.09	-0.01
Sunshine	-0.02	0.04	-0.23	1	0.49	0.29	-0.03	0.05	0.01	0.01
Temp3pm	-0.39	-0.29	-0.08	0.49	1	0.86	0.03	0.03	0	0.05
Temp9am	-0.47	-0.42	0.01	0.29	0.86	1	0.15	0.16	0.13	0.05
WindGustSpeed	-0.41	-0.46	0.13	-0.03	0.03	0.15	1	0.69	0.61	-0.03
WindSpeed3pm	-0.26	-0.3	0.06	0.05	0.03	0.16	0.69	1	0.52	-0.03
WindSpeed9am	-0.18	-0.23	0.09	0.01	0	0.13	0.61	0.52	1	-0.02
Year	0.02	0.03	-0.01	0.01	0.05	0.05	-0.03	-0.03	-0.02	1

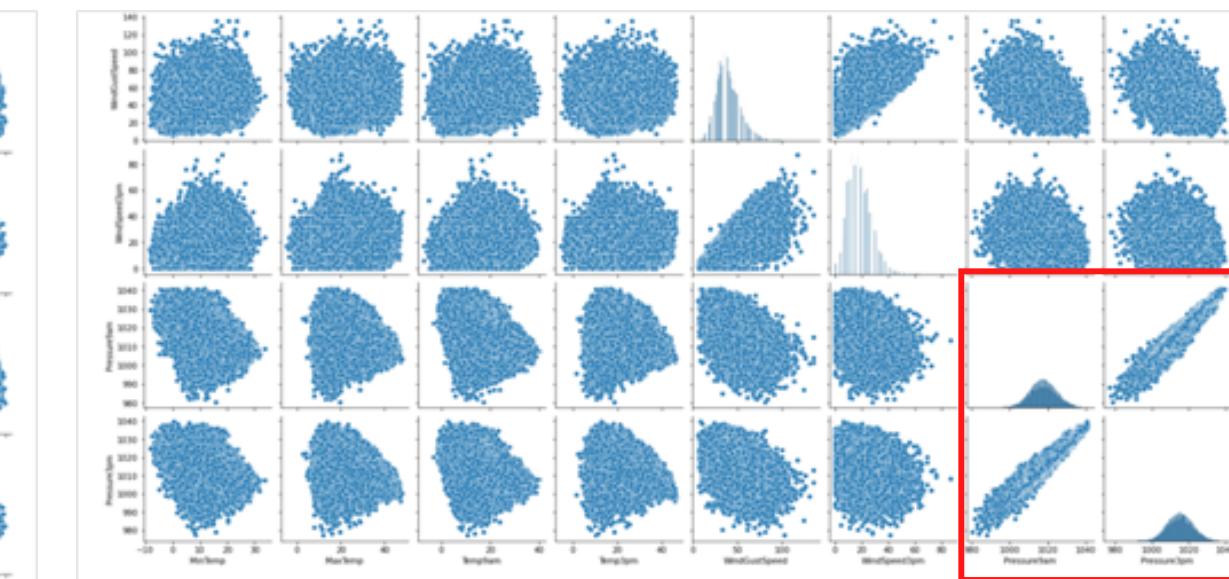
MinTemp and MaxTemp ---> 0.74
 MinTemp and Temp3pm ---> 0.71
 MinTemp and Temp9am ---> 0.90
 MaxTemp and Temp9am ---> 0.89

MaxTemp and Temp3pm ---> 0.98
 WindGustSpeed and WindSpeed3pm ---> 0.69
 Pressure9am and Pressure3pm ---> 0.96
 Temp9am and Temp3pm ---> 0.86

- กราฟ Pair Plot แสดงกราฟ Scatter Plot ระหว่างตัวแปรแต่ละตัวที่มีความสัมพันธ์กันสูงในทิศทางเดียวกัน



- MinTemp และ MaxTemp
- MinTemp และ Temp9am
- MinTemp และ Temp3pm
- MaxTemp และ Temp9am



- MaxTemp และ Temp3pm
- Temp9am และ Temp3pm
- Pressure9am และ Pressure3pm

3. Data Preparation

- Declare feature vector and target variable : จัดข้อมูลให้ X เป็น dataframe ที่ประกอบไปด้วยตัวแปรอิสระ และ y เป็น dataframe ที่ประกอบไปด้วยตัวแปรตามหรือ rainTomorrow

10. Declare feature vector and target variable

```
[31]: x = df.drop(['RainTomorrow'], axis=1)  
y = df['RainTomorrow']
```

- Split data into separate training and test set : แบ่งข้อมูลให้เป็นชุด test และ train ใช้อัตราส่วน 20:80

11. Split data into separate training and test set

```
[32]: #split X and y into training and testing sets

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)

# check the shape of X_train and X_test

X_train.shape, X_test.shape
```

[32]: ((116368, 24), (29092, 24))

- Feature Engineering

- การจัดการกับ Missing Value ของตัวแปร Categorical

```
[40]: # impute missing categorical variables with most frequent value

for df1 in [X_train, X_test]:
    df1['WindGustDir'].fillna(X_train['WindGustDir'].mode()[0], inplace=True)
    df1['WindDir9am'].fillna(X_train['WindDir9am'].mode()[0], inplace=True)
    df1['WindDir3pm'].fillna(X_train['WindDir3pm'].mode()[0], inplace=True)
    df1['RainToday'].fillna(X_train['RainToday'].mode()[0], inplace=True)
```

```
[45]: # impute missing categorical variables with most frequent value

for df2 in [y_train, y_test]:
    df2.fillna(y_train.mode()[0], inplace=True)
```

- Feature Engineering
 - การจัดการกับ Missing Value ของตัวแปร Numerical

```
[51]: # impute missing values in X_train and X_test with respective column median in X_train

for df3 in [X_train, X_test]:
    for col in numerical:
        col_median=X_train[col].median()
        df3[col].fillna(col_median, inplace=True)
```

- Feature Engineering
 - การจัดการกับ outlier ของตัวแปร Numerical

Engineering outliers in numerical variables

```
[58]: def max_value(df3, variable, top):  
    return np.where(df3[variable]>top, top, df3[variable])  
  
for df3 in [X_train, X_test]:  
    df3['Rainfall'] = max_value(df3, 'Rainfall', 3.2)  
    df3['Evaporation'] = max_value(df3, 'Evaporation', 21.8)  
    df3['WindSpeed9am'] = max_value(df3, 'WindSpeed9am', 55)  
    df3['WindSpeed3pm'] = max_value(df3, 'WindSpeed3pm', 57)
```

- Feature Engineering
 - ทำการ encode และสร้างตัวแปร Dummy

```
[62]: # encode RainToday variable

import category_encoders as ce

encoder = ce.BinaryEncoder(cols=['RainToday'])

X_train = encoder.fit_transform(X_train)

X_test = encoder.transform(X_test)
```

```
[65]: X_train = pd.concat([X_train[numerical], X_train[['RainToday_0', 'RainToday_1']],
                        pd.get_dummies(X_train.Location),
                        pd.get_dummies(X_train.WindGustDir),
                        pd.get_dummies(X_train.WindDir9am),
                        pd.get_dummies(X_train.WindDir3pm)], axis=1)
```

- Feature Scaling
 - ทำการ scaling data โดยใช้ MinMaxScaler

```
[74]: from sklearn.preprocessing import MinMaxScaler  
  
scaler = MinMaxScaler()  
  
X_train = scaler.fit_transform(X_train)  
  
X_test = scaler.transform(X_test)  
  
X_train2 = scaler.fit_transform(X_train2)  
  
X_test2 = scaler.transform(X_test2)
```

4. Modeling

- ใช้ Logistic Regression โดยผลที่ได้จากโมเดลจะเป็นการคำนายค่าของตัวแปร RainTomorrow ว่าจะมีค่าเป็น Yes หรือ No

14. Model training

```
[78]: # train a logistic regression model on the training set
      from sklearn.linear_model import LogisticRegression

      # instantiate the model
      logreg = LogisticRegression(solver='liblinear', random_state=0)

      # fit the model
      logreg.fit(X_train, y_train)
```



```
[78]: LogisticRegression(random_state=0, solver='liblinear')
```

5. Evaluation

- ตรวจสอบค่า Accuracy

Check accuracy score

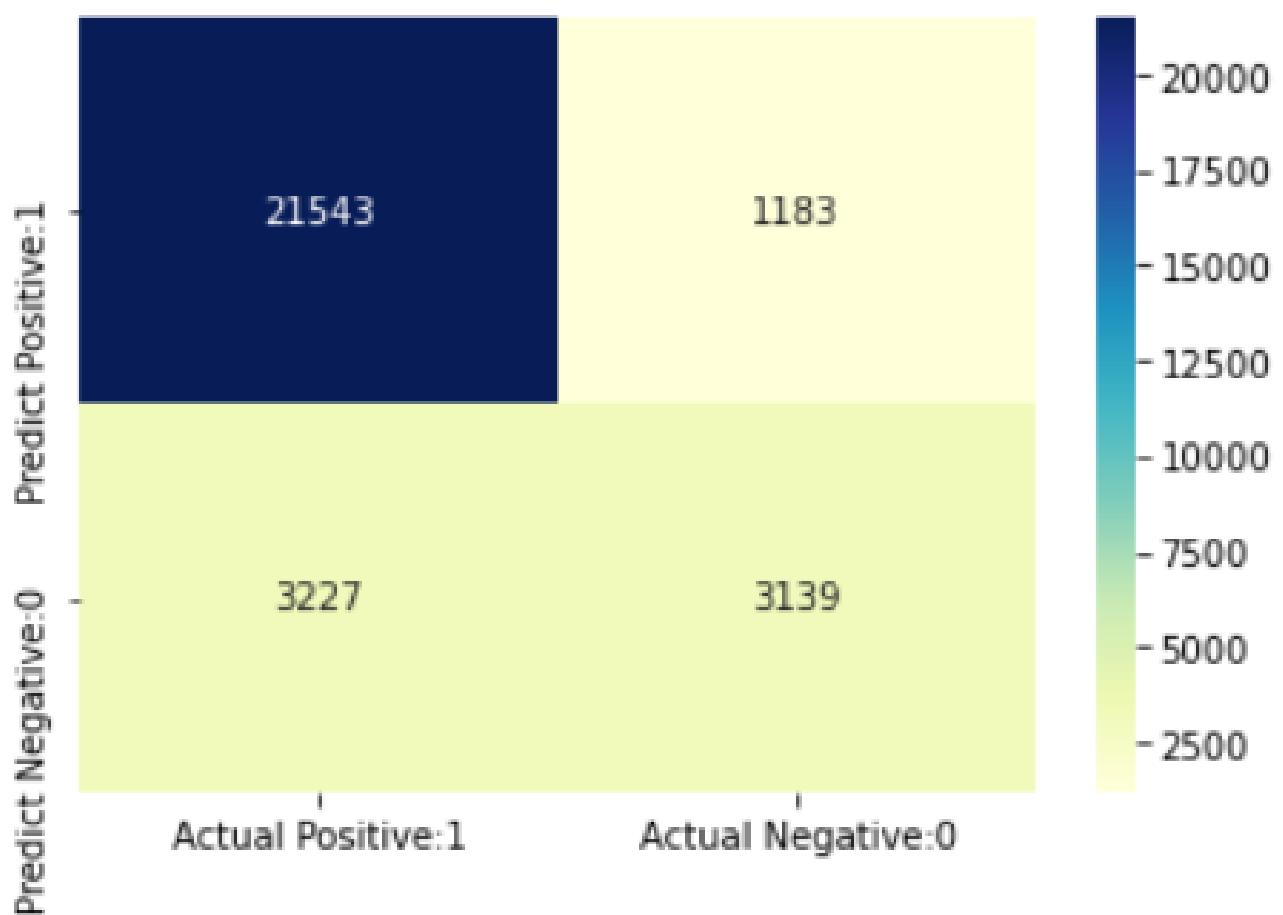
```
[82]: from sklearn.metrics import accuracy_score  
  
print('Model accuracy score: {:.4f}'.format(accuracy_score(y_test, y_pred)))
```

```
Model accuracy score: 0.8484
```

Rain in Australia

June 10 , 2021

• ຕຽວຈສອບ Confusion Metrix



- เมื่อนำมาคิดค่า sensitivity โดยคำนวณได้จาก

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

จะมีค่าเท่ากับ 0.4931 หรือ 49.31%
ซึ่งมีค่าที่น้อย และไม่ตรงกับเป้าหมายที่ต้องการให้มากกว่า 70%

• จึงทำการเพิ่ม Threshold

With 0.1 threshold the Confusion Matrix is

```
[[13291  9435]
 [ 571  5795]]
```

with 19086 correct predictions,

9435 Type I errors(False Positives),

571 Type II errors(False Negatives),

Accuracy score: 0.6560566478757046

Sensitivity: 0.9103047439522463

Specificity: 0.5848367508580481

=====

With 0.3 threshold the Confusion Matrix is

```
[[19744  2982]
 [ 2043  4323]]
```

with 24067 correct predictions,

2982 Type I errors(False Positives),

2043 Type II errors(False Negatives),

Accuracy score: 0.8272721022961639

Sensitivity: 0.679076343072573

Specificity: 0.8687846519405087

=====

With 0.2 threshold the Confusion Matrix is

```
[[17742  4984]
 [ 1365  5001]]
```

with 22743 correct predictions,

4984 Type I errors(False Positives),

1365 Type II errors(False Negatives),

Accuracy score: 0.7817613089509143

Sensitivity: 0.7855796418473139

Specificity: 0.7806917187362492

=====

With 0.4 threshold the Confusion Matrix is

```
[[20840  1886]
 [ 2645  3721]]
```

with 24561 correct predictions,

1886 Type I errors(False Positives),

2645 Type II errors(False Negatives),

Accuracy score: 0.8442527155231678

Sensitivity: 0.5845114671693371

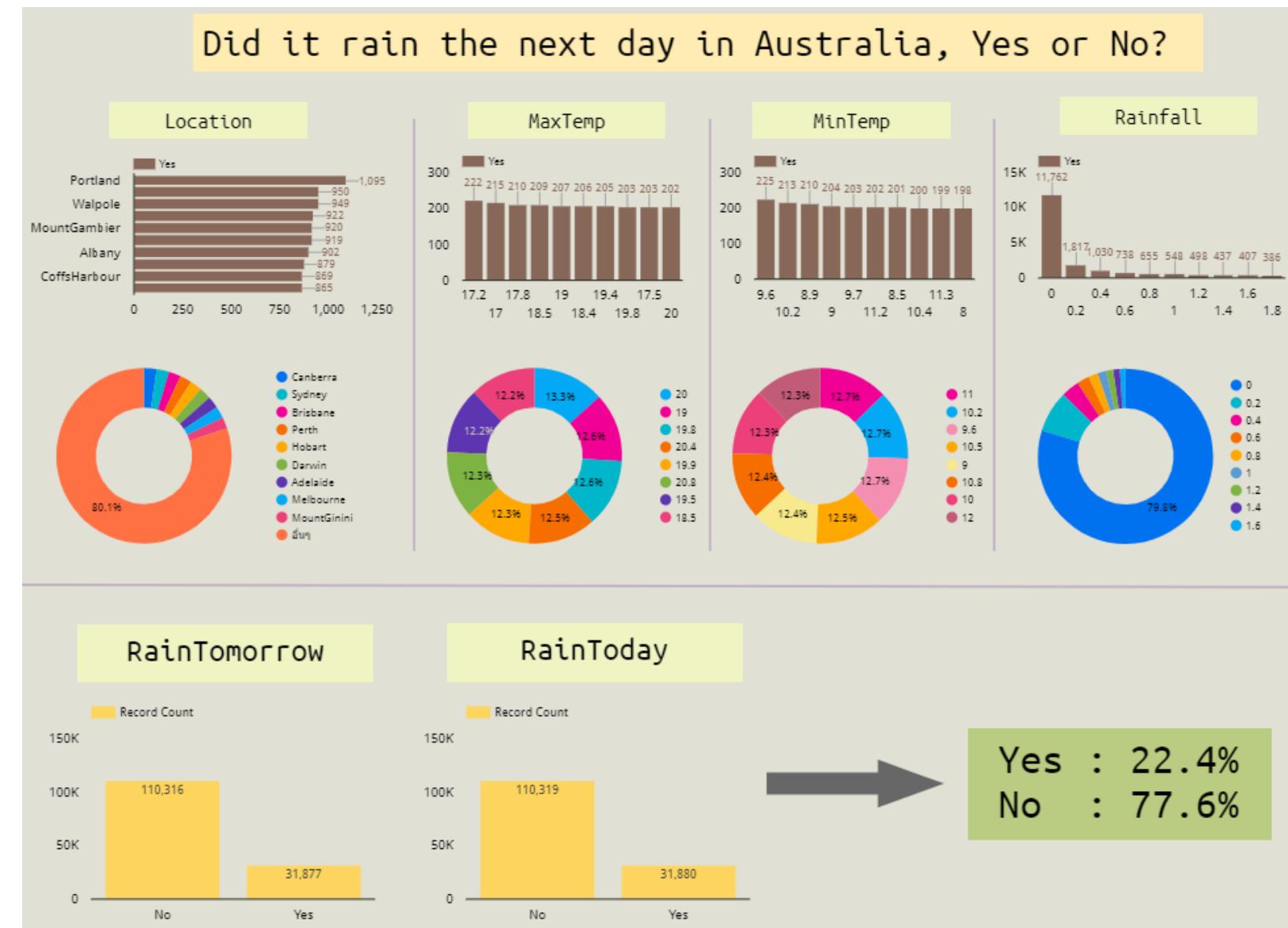
Specificity: 0.9170113526357476

=====

6. Deployment

- อาจนำผลลัพธ์ที่ได้จากการทำโปรเจคนี้ไปเผยแพร่ให้ user ได้ลองใช้งานจริง หรืออาจนำไปเผยแพร่ใน Kaggle เพื่อที่จะได้รับคำแนะนำ หรือเห็นข้อผิดพลาด ในโมเดลเรามากขึ้น เพื่อเป็นแนวทางในการแก้ไขและพัฒนาโมเดลของเราต่อไป ในที่นี่ได้ลองเผยแพร่ใน Google Data Studio เพื่อแสดงให้ users เห็น Dashboard ที่นำเสนอ

• Dashboard



Rain in Australia

June 10 , 2021

ผลลัพธ์ที่ได้ และสรุป

○ ○ ○ ○

เราจะทำการเตรียมตัวรับมือกับฝนตก ถ้าหากโอกาสที่ฝนจะตกอยู่ที่ประมาณ 70%

Threshold 0.1

91.03%

With 0.1 threshold the Confusion Matrix is

```
[[13291  9435]
 [ 571  5795]]
```

with 19086 correct predictions,
9435 Type I errors(False Positives),
571 Type II errors(False Negatives),
Accuracy score: 0.6560566478757046
Sensitivity: 0.9103047439522463
Specificity: 0.5848367508580481

With 0.2 threshold the Confusion Matrix is

```
[[17742  4984]
 [1365  5001]]
```

with 22743 correct predictions,
4984 Type I errors(False Positives),
1365 Type II errors(False Negatives),
Accuracy score: 0.7817613089509143
Sensitivity: 0.7855796418473139
Specificity: 0.7806917187362492

Threshold 0.2

78.56%

Threshold 0.3

67.91%

With 0.3 threshold the Confusion Matrix is

```
[[19744  2982]
 [2043  4323]]
```

with 24067 correct predictions,
2982 Type I errors(False Positives),
2043 Type II errors(False Negatives),
Accuracy score: 0.8272721022961639
Sensitivity: 0.679076343072573
Specificity: 0.8687846519405087

With 0.4 threshold the Confusion Matrix is

```
[[20840  1886]
 [2645  3721]]
```

with 24561 correct predictions,
1886 Type I errors(False Positives),
2645 Type II errors(False Negatives),
Accuracy score: 0.8442527155231678
Sensitivity: 0.5845114671693371
Specificity: 0.9170113526357476

Threshold 0.4

58.45%

ສົ່ງຖິ່ນໄດ້ເຮັດວຽບຮູແລະ ແບວກາງໃນກາර ພັດທະຍອດ

○ ○ ○ ○

สิ่งที่ได้เรียนรู้

- เรียนรู้การใช้ Google Cloud Platform Ex. Cloud Storage, BigQuery
- ได้ฝึกฝน พัฒนาการเขียนภาษา Python
- ได้เรียนรู้ในลักษณะของ CRISP-DM

แนวทางการพัฒนาต่อไป

- นำมาต่อยอดได้เพื่อทำนายหน้าฝนในประเทศไทยได้โดยการประยุกต์ข้อมูล ประยุกต์วิธีการจัดการกับข้อมูลต่างๆ วิธีการสร้างโมเดลให้เหมาะสมกับข้อมูลของประเทศไทยมากขึ้น
- อาจจะมีการใช้ Platform ใน Google Cloud Platform ให้มากกว่านี้ เช่น Google Cloud Dataprep ซึ่งอาจช่วยให้การจัดเตรียมข้อมูล ก่อนนำเข้าโมเดล ง่ายและได้ประสิทธิภาพที่ดีกว่า

Thunchanok Nakpasom
6009680106

June 10, 2021

**Thank
You**

o o o o

Rain in
Australia