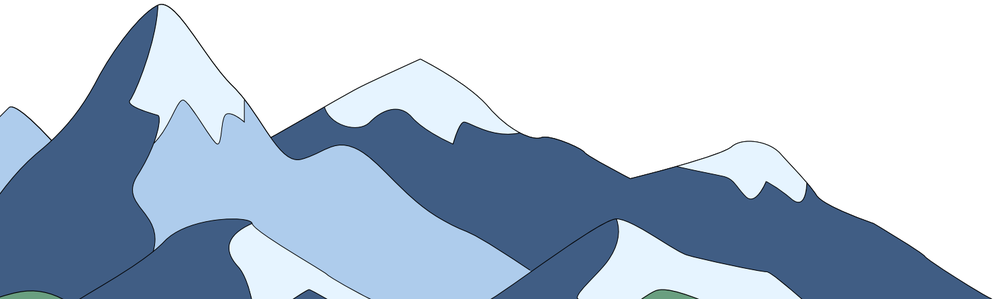
PREDICTING WEATHER FOR RISK MANAGEMENT AND ACCESSIBILITY

GROUP 7

Mufaddal Abizar Ezzi,
Areej Ahmed,
Syed Ammar,
Ziyaan Mir,
Imaad Muhammad Ismail



UNDERSTANDING THE UNPREDICTABILITY OF WEATHER CONDITIONS

Weather patterns are increasingly **unpredictable**, impacting critical areas like risk management, energy efficiency, and accessibility.

Image source : https://climatecrisis247.com/news/from-separate-incidents-to-simultaneous-disasters-the-new-normal-in-extreme-weather-patterns/

OUR OBJECTIVES

Create a regression model to **predict sunshine** in hours per day to assist with optimal use of solar panels.

Create a classification model where we also cluster information and attributes related to **prediction of next day rainfall** for disaster preparedness.

Classify images of different weather phenomenons in order to assist visually impaired people with scene understanding.

Image source : https://climatecrisis247.com/news/from-separate-incidents-to-simultaneous-disasters-the-new-normal-in-extreme-weather-patterns/

SUNSHINE PREDICTION

WHY IS IT NEEDED?

"The **amount of sunlight** that hits a solar panel is one of the **biggest factors in how much electricity** it will generate. The **more sunlight** available to the panel, the **more electricity** it can produce" (Lane, Cappuccio and SolarReviews, 2024).

Image source : https://www.mahindrasolarize.com/blog/harnessing-the-power-of-the-sun-a-guide-to-solar-panel-systems-for-homes/

SUNSHINE PREDICTION

REQUIREMENTS

The **requirements** we intend to satisfy with the work done with this dataset as per the rubrics of our course are:

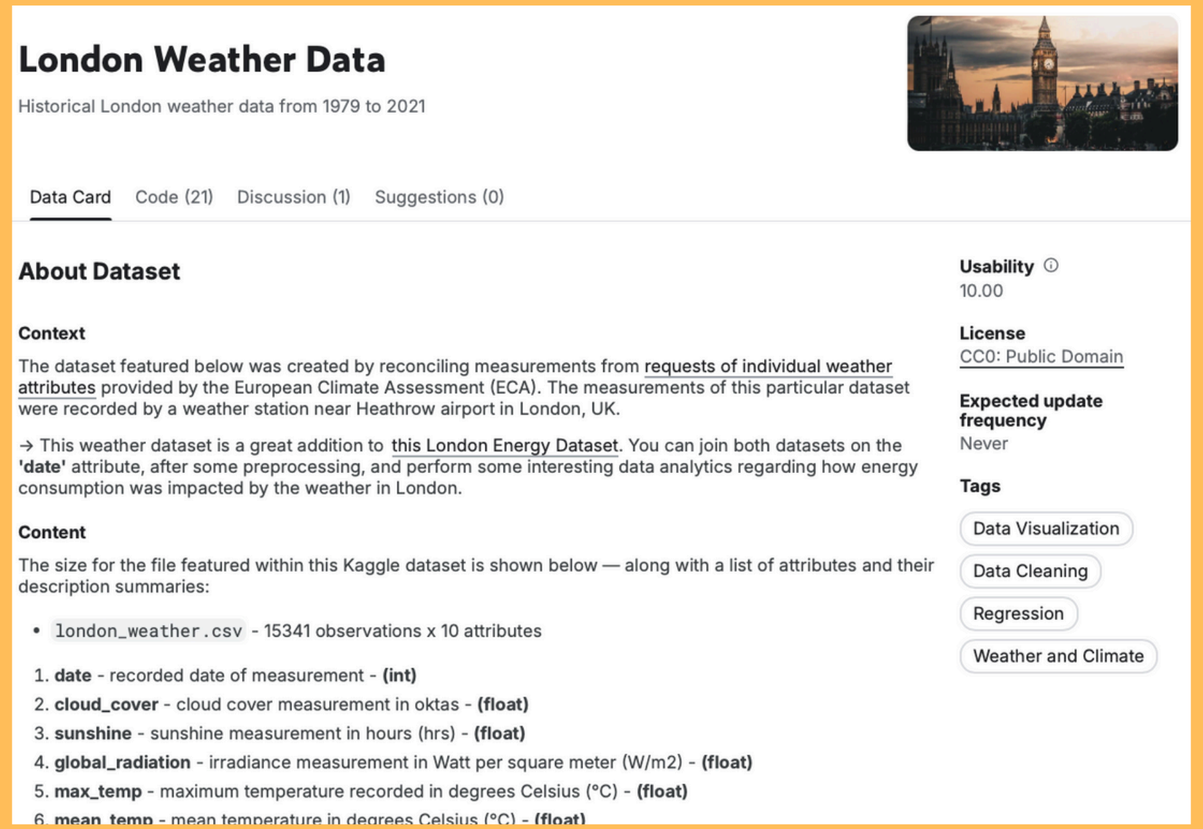
R1: Project topic and directions
R2: Data Analysis
R4: Decision Trees and Regression models

Image source : https://www.mahindrasolarize.com/blog/harnessing-the-power-of-the-sun-a-guide-to-solar-panel-systems-for-homes/

SUNSHINE PREDICTION

DATASET 1

- This dataset is published on **Kaggle**.
- It is sourced from **European Climate Assessment** and licensed **Creative Commons CC0: Public Domain** license.
- It is **numerical** and **tabular** in format.
- We plan to use **Random Forest, Decision Tree** and regression models like **Linear Regression and Perceptron** to predict amount of sunshine received



London Weather Data

Historical London weather data from 1979 to 2021

Data Card | Code (21) | Discussion (1) | Suggestions (0)

About Dataset

Context

The dataset featured below was created by reconciling measurements from requests of individual weather attributes provided by the European Climate Assessment (ECA). The measurements of this particular dataset were recorded by a weather station near Heathrow airport in London, UK.

→ This weather dataset is a great addition to this London Energy Dataset. You can join both datasets on the **'date'** attribute, after some preprocessing, and perform some interesting data analytics regarding how energy consumption was impacted by the weather in London.

Content

The size for the file featured within this Kaggle dataset is shown below — along with a list of attributes and their description summaries:

- `london_weather.csv` - 15341 observations x 10 attributes

1. **date** - recorded date of measurement - **(int)**
2. **cloud_cover** - cloud cover measurement in oktas - **(float)**
3. **sunshine** - sunshine measurement in hours (hrs) - **(float)**
4. **global_radiation** - irradiance measurement in Watt per square meter (W/m2) - **(float)**
5. **max_temp** - maximum temperature recorded in degrees Celsius (°C) - **(float)**
6. **mean_temp** - mean temperature in degrees Celsius (°C) - **(float)**

Usability
10.00

License
CC0: Public Domain

Expected update frequency
Never

Tags

Data Visualization
Data Cleaning
Regression
Weather and Climate

Source : https://www.kaggle.com/datasets/emmanuelfwerr/london-weather-data

WHY DID WE PICK THIS DATASET? (R1)

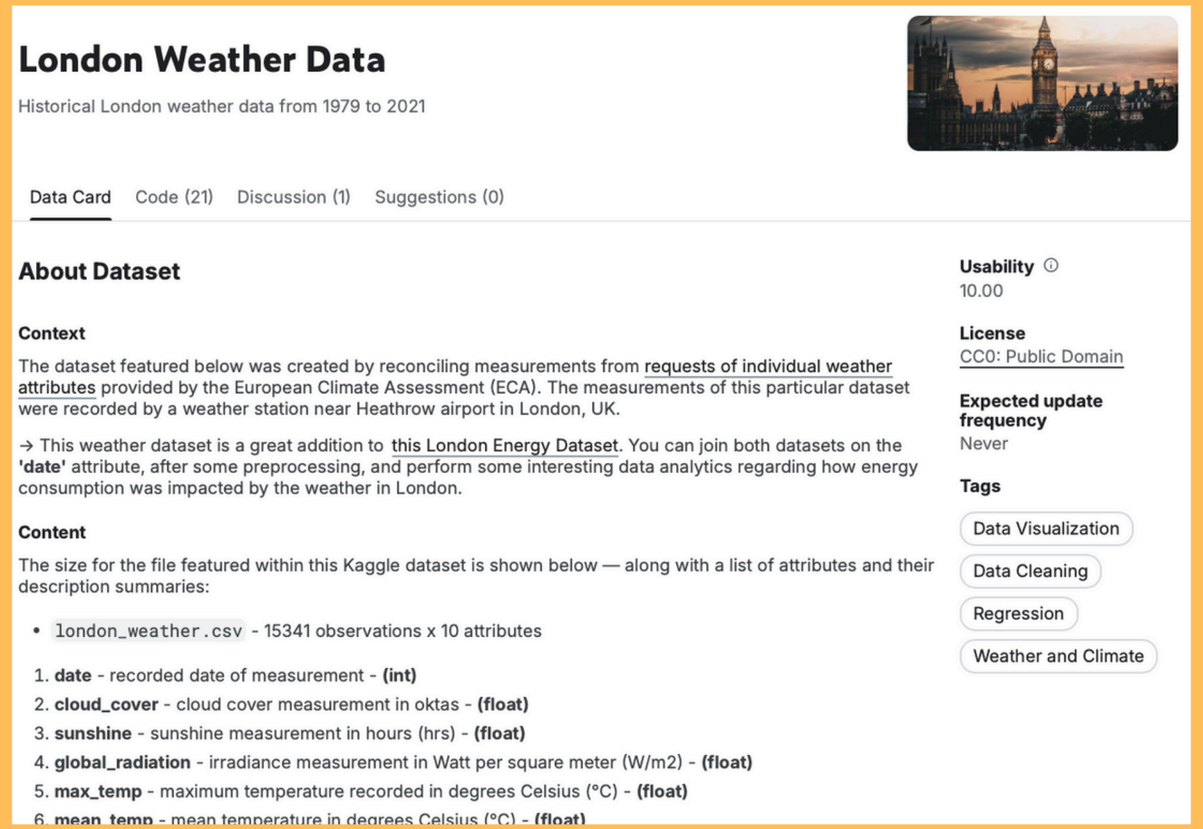
DATASET 1

- The dataset contains **42 years of daily weather observations**, tracking 10 diverse meteorological attributes. Which shows a very **thorough tracking of weather attributes** making it an **ideal dataset** for regression.

- This dataset possessed all the ideal atributes we looked for in terms of a **sunshine regression** model such as global radiation and various temperature measurements.

- The high and **perfect usability score** of 10 out of 10 on Kaggle made it a **feasible dataset** for this purpose.

- This dataset allows us to do something **meaningful** and **stimulating** in terms of training models, preprocessing and more.



London Weather Data

Historical London weather data from 1979 to 2021

Data Card | Code (21) | Discussion (1) | Suggestions (0)

About Dataset

Context

The dataset featured below was created by reconciling measurements from requests of individual weather attributes provided by the European Climate Assessment (ECA). The measurements of this particular dataset were recorded by a weather station near Heathrow airport in London, UK.

→ This weather dataset is a great addition to this London Energy Dataset. You can join both datasets on the **'date'** attribute, after some preprocessing, and perform some interesting data analytics regarding how energy consumption was impacted by the weather in London.

Content

The size for the file featured within this Kaggle dataset is shown below — along with a list of attributes and their description summaries:

- `london_weather.csv` - 15341 observations x 10 attributes

1. **date** - recorded date of measurement - **(int)**
2. **cloud_cover** - cloud cover measurement in oktas - **(float)**
3. **sunshine** - sunshine measurement in hours (hrs) - **(float)**
4. **global_radiation** - irradiance measurement in Watt per square meter (W/m2) - **(float)**
5. **max_temp** - maximum temperature recorded in degrees Celsius (°C) - **(float)**
6. **mean_temp** - mean temperature in degrees Celsius (°C) - **(float)**

Usability ⓘ
10.00

License
CC0: Public Domain

Expected update frequency
Never

Tags

Data Visualization
Data Cleaning
Regression
Weather and Climate

Source : https://www.kaggle.com/datasets/emmanuelfwerr/london-weather-data

ABOUT THE DATASET (R2)

DATASET 1

This dataset consists of :

- **date**: stores date as int
- **cloud_cover**: stores float measured in oktas
- **sunshine**: stores float measured in hours **global_radiation**: stores float measured in (W/m²) max_temp: stores float measured in (°C)
- **mean_temp**: stores float measured in (°C)
- **min_temp**: stores float measured in (°C)
- **precipitation**: stores float measured in(mm)
- **pressure**: stores float measured in (Pa)
- **snow_depth**: stores float measured in (cm)

First 5 rows of the dataset:

	date	cloud_cover	sunshine	global_radiation	max_temp	mean_temp	min_temp	precipitation	pressure	snow_depth
0	19790101	2.0	7.0	52.0	2.3	-4.1	-7.5	0.4	101900.0	9.0
1	19790102	6.0	1.7	27.0	1.6	-2.6	-7.5	0.0	102530.0	8.0
2	19790103	5.0	0.0	13.0	1.3	-2.8	-7.2	0.0	102050.0	4.0
3	19790104	8.0	0.0	13.0	-0.3	-2.6	-6.5	0.0	100840.0	2.0
4	19790105	6.0	2.0	29.0	5.6	-0.8	-1.4	0.0	102250.0	1.0

Shape before dropping missing values: (15341, 10)

ABOUT THE DATASET (R2)

DATASET 1

Here is our **data statistical analysis** on the colums we plan to use for our model.

cloud_cover:

- **Count**: 15,322 valid observations
- **Mean**: 5.27 (scale: 0-9)
- **Std Dev**: 2.07 (most values within ±2 of the mean)
- **Min**: 0 (no cloud cover)
- **Median**: 6 (half the days had cloud cover ≤ 6)
- **Max**: 9 (complete cloud cover)
- Mostly moderate to high, with many overcast days.

sunshine:

- **Count**: 15,341 valid observations
- **Mean**: 4.35 hours/day
- **Std Dev**: 4.03 (large variation in sunshine duration)
- **Min**: 0 (no sunshine)
- **Median**: 3.5 hours (half the days had sunshine ≤ 3.5 hours)
- **Max**: 16 hours (maximum daylight)
- Highly variable, with many days having minimal sunshine and some full-day sunshine.

Dataset Statistics (excluding date):

	date	cloud_cover	sunshine	global_radiation	max_temp	mean_temp	min_temp	precipitation	pressure	snow_depth
count	1.534100e+04	15322.000000	15341.000000	15322.000000	15335.000000	15305.000000	15339.000000	15335.000000	15337.000000	13900.000000
mean	1.999567e+07	5.268242	4.350238	118.756951	15.388777	11.475511	7.559867	1.668634	101536.605594	0.037986
std	1.212176e+05	2.070072	4.028339	88.898272	6.554754	5.729709	5.326756	3.738540	1049.722604	0.545633
min	1.979010e+07	0.000000	0.000000	8.000000	-6.200000	-7.600000	-11.800000	0.000000	95960.000000	0.000000
25%	1.989070e+07	4.000000	0.500000	41.000000	10.500000	7.000000	3.500000	0.000000	100920.000000	0.000000
50%	2.000010e+07	6.000000	3.500000	95.000000	15.000000	11.400000	7.800000	0.000000	101620.000000	0.000000
75%	2.010070e+07	7.000000	7.200000	186.000000	20.300000	16.000000	11.800000	1.600000	102240.000000	0.000000
max	2.020123e+07	9.000000	16.000000	402.000000	37.900000	29.000000	22.300000	61.800000	104820.000000	22.000000

ABOUT THE DATASET (R2)

DATASET 1

The displayed statistics on missing values indicate that several columns in the dataset contain gaps, which highlights the need for preprocessing. For instance:

- Columns like **cloud_cover** and **global_radiation** have 19 missing values each, while **mean_temp** is missing 36 values.

- The **snow_depth** column has the highest number of missing values, with **1,441 entries absent**, representing **9.39% of the data**.

These **gaps are not evenly distributed** across the dataset, which makes **preprocessing necessary** to ensure data consistency.

```
General Information:

Missing Values per Column (excluding date):
date                0
cloud_cover        19
sunshine            0
global_radiation   19
max_temp            6
mean_temp          36
min_temp            2
precipitation       6
pressure            4
snow_depth       1441
dtype: int64
```

```
Percentage of Missing Data per Column (excluding date):
date               0.00%
cloud_cover        0.12%
sunshine           0.00%
global_radiation   0.12%
max_temp           0.04%
mean_temp          0.23%
min_temp           0.01%
precipitation      0.04%
pressure           0.03%
snow_depth         9.39%
dtype: object
```

DATA CLEANING (R2)
DATASET 1

Reduced dataset from 15,341 to 13,843 rows after **removing missing (NaN) values**.

- Crucial step due to many missing values in the original data.
- Ensures a more **reliable dataset** for better predictions.
- **Improves model accuracy** and quality.
- Example of **undersampling** by removing incomplete rows.
- **Handling missing values** prevents bias and incorrect results.
- Methods like **dropping**, **imputing**, or **interpolation** can be used based on data significance.

```
Handling Missing Values (Dropping Rows with Missing Values, excluding date):
Shape after dropping missing values: (13843, 10)
First 5 rows of the cleaned dataset (excluding date):

       date  cloud_cover  sunshine  global_radiation  max_temp  mean_temp  \
0  19790101          2.0       7.0              52.0       2.3       -4.1
1  19790102          6.0       1.7              27.0       1.6       -2.6
2  19790103          5.0       0.0              13.0       1.3       -2.8
3  19790104          8.0       0.0              13.0      -0.3       -2.6
4  19790105          6.0       2.0              29.0       5.6       -0.8

   min_temp  precipitation  pressure  snow_depth
0      -7.5            0.4  101900.0         9.0
1      -7.5            0.0  102530.0         8.0
2      -7.2            0.0  102050.0         4.0
3      -6.5            0.0  100840.0         2.0
4      -1.4            0.0  102250.0         1.0
```

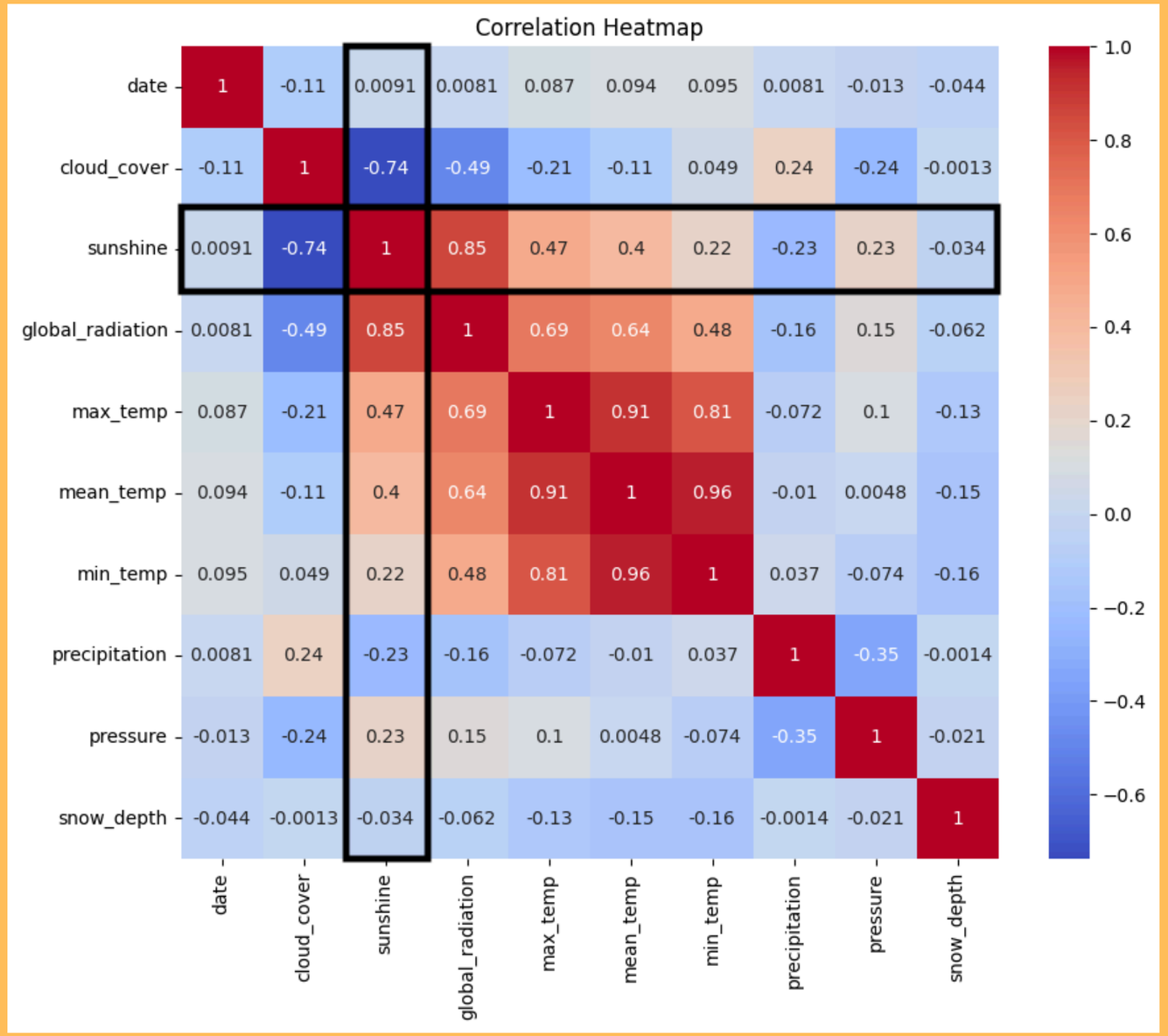
ABOUT THE DATASET (R2)

DATASET 1

The correlation heatmap shows the relationship between **sunshine** with **other attributes** in the datasets.

According to (Wayne W., 2021)
+/-0.2 to 0.4 Weak association,
+/-0.4 to 0.6 is Moderate association ,
+/-0.6 to 0.8 is Strong association,
+/-0.8 to 1.0 is Very strong association

We can see the dataset has **good correlations** which makes this dataset a g**ood candidate** for regression models.



RAINFALL PREDICTION

WHY IS IT NEEDED?

- Helps manage **flood risks** and **damage**.
- Could be important for regions with **low-lying areas**, which are more vulnerable.
- Climate change worsens extreme rainfall; creating greater need for more accurate prediction models.
- Rising temperatures increase **flood exposure**.
- Accurate predictions improve disaster preparedness.
- A two-degree Celsius increase in global temperatures could raise the flood-exposed population by **30%**. (Salas, 2024)

Image source : https://ameyawdebrah.com/record-rainfall-in-uae-dubai-in-chaos-as-streets-flood/

RAINFALL PREDICTION

REQUIREMENTS

The **requirements** we intend to satisfy with the work done with this dataset as per the rubrics of our course are:

- **R1**: Project topic and directions
- **R2**: Data Analysis
- **R3**: Clustering
- R4: Basic classifiers and Decision Trees

Image source : https://ameyawdebrah.com/record-rainfall-in-uae-dubai-in-chaos-as-streets-flood/

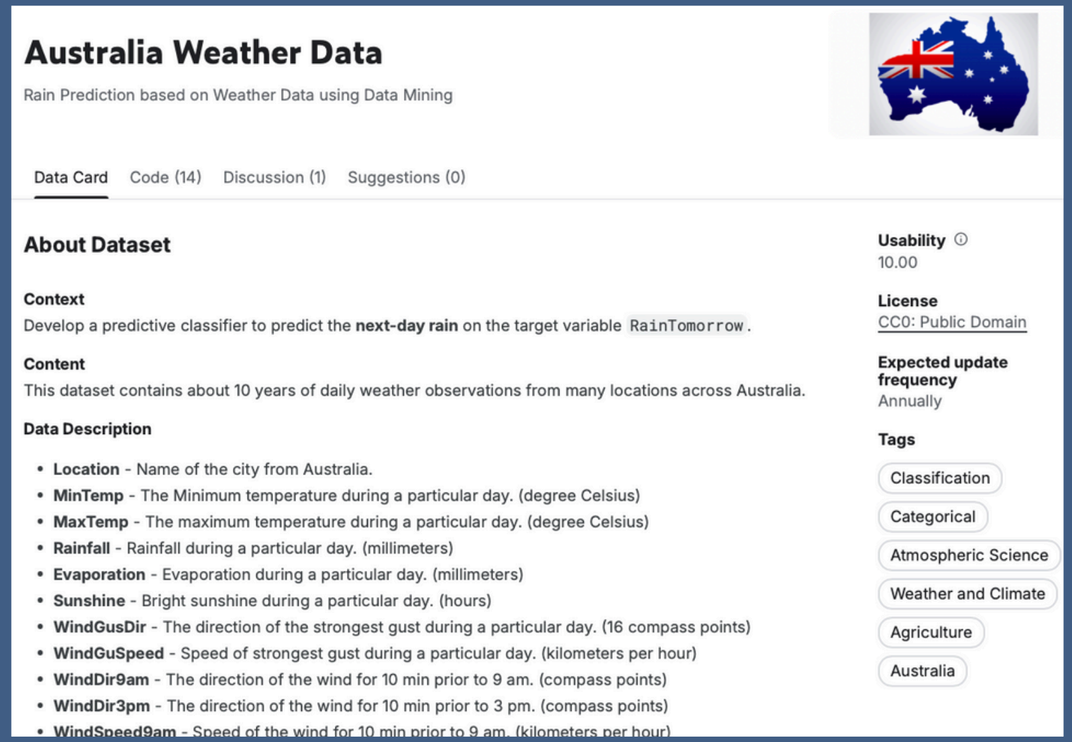
AUSTRALIAN WEATHER DATA

OUR GOAL: RAINFALL PREDICTION (R1)

- This dataset is published on **Kaggle**, sourced from the **Australian Bureau of Meteorology** and it is licensed under the **Creative Commons CC0: Public Domain license**.

- It is numerical and tabular in format.

- We plan to use clustering techniques **(R3)** such as **K-means**, **Random Forest**, **Decision Trees** and classification **(R4)** models like **Naïve Bayes** and **k-Nearest Neighbours** to predict whether it will rain or not.

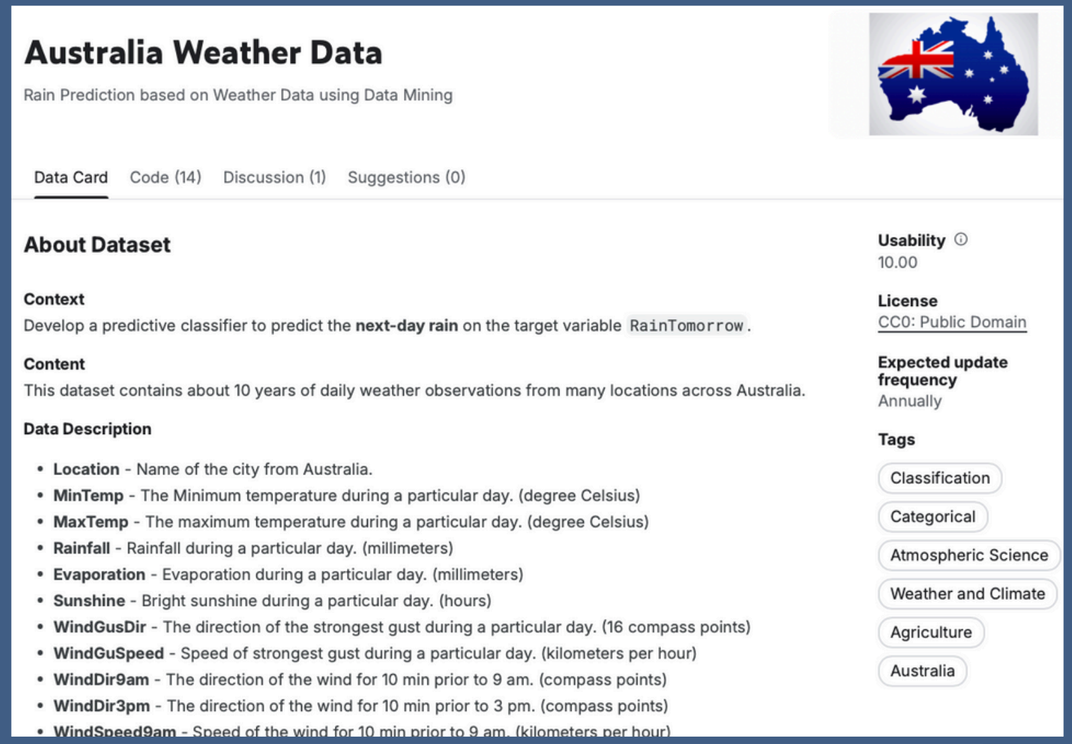


Source : https://www.kaggle.com/datasets/arunavakrchakraborty/australia-weather-data

AUSTRALIAN WEATHER DATA

WHY DID WE PICK THIS DATASET? (R1)

- The dataset contains **10 years of daily weather observations**, tracking **22 diverse meteorological attributes**. Which shows a very thorough tracking of weather attributes making it an **ideal dataset** for the classification.
- The **large dataset reduces** the impact of **weaker correlations and noise**, ensuring that even with lower correlations in some features, the model's overall predictive power remains strong by highlighting more significant patterns.
- This dataset allows us to do something **meaningful and stimulating** in terms of training models, preprocessing and more.



Source : https://www.kaggle.com/datasets/arunavakrchakraborty/australia-weather-data

AUSTRALIAN WEATHER DATA

OUR GOAL: RAINFALL PREDICTION (R1)

Our dataset includes the following key attributes:

- **Location & Temperature**: Cities in Australia, with min/max temperatures and readings at 9 am and 3 pm (°C).
- **Rain & Evaporation**: Daily rainfall and evaporation (mm), as well as indicators for whether it rained today and if it will rain tomorrow.
- **Sunshine**: Daily hours of sunshine.
- **Wind**: Wind gust direction (compass points) and speed (km/h), along with wind data at 9 am and 3 pm.
- **Humidity & Pressure**: Humidity (%) and atmospheric pressure (hPa) recorded at 9 am and 3 pm.
- **Cloud Cover**: Cloud cover at 9 am and 3 pm (measured in eighths).

Our goal is to build a classification model to predict whether it will rain the next day.
The primary target variable is the **“RainTomorrow”** attribute, which indicates if it **will rain ('Yes' or 'No')**.
We'll use this as our output variable and **analyze the correlations** with **other features** to determine which attributes **most effectively predict** the **likelihood of rain**.

First 5 rows of the dataset (before cleaning):

	row ID	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	\
0	Row0	Albury	13.4	22.9	0.6	NaN	NaN	
1	Row1	Albury	7.4	25.1	0.0	NaN	NaN	
2	Row2	Albury	17.5	32.3	1.0	NaN	NaN	
3	Row3	Albury	14.6	29.7	0.2	NaN	NaN	
4	Row4	Albury	7.7	26.7	0.0	NaN	NaN	

	WindGustDir	WindGustSpeed	WindDir9am	...	Humidity9am	Humidity3pm	\
0	W	44.0	W	...	71.0	22.0	
1	WNW	44.0	NNW	...	44.0	25.0	
2	W	41.0	ENE	...	82.0	33.0	
3	WNW	56.0	W	...	55.0	23.0	
4	W	35.0	SSE	...	48.0	19.0	

	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	\
0	1007.7	1007.1	8.0	NaN	16.9	21.8	No	
1	1010.6	1007.8	NaN	NaN	17.2	24.3	No	
2	1010.8	1006.0	7.0	8.0	17.8	29.7	No	
3	1009.2	1005.4	NaN	NaN	20.6	28.9	No	
4	1013.4	1010.1	NaN	NaN	16.3	25.5	No	

	RainTomorrow
0	0
1	0
2	0
3	0
4	0

[5 rows x 23 columns]

AUSTRALIAN WEATHER DATA

OUR GOAL: RAINFALL PREDICTION (R1)

This Dataset Statistical Analysis :

- Missing Data:
  - **Evaporation** (43%) and **Sunshine** (48%) have significant missing values, while **MinTemp**, **MaxTemp**, and **Rainfall** are mostly complete.
- **Temperature**:
  - **MinTemp** ranges from -8.5°C to 33.9°C,
  - **MaxTemp** from -4.1°C to 48.1°C.
- **Rainfall**: **Mean** is 2.35mm, **max** at 371mm indicates occasional heavy rainfall.
- **Humidity**: Higher at 9am (avg. 68.9%) than 3pm (51.4%).
- **Pressure** : Standard range around sea-level (978.2 hPa to 1041.1 hPa).
- **Cloud Cover**: Mean is ~4.5 out of 9, indicating mostly clear to partly cloudy skies.
- **Location**: 49 unique locations, with Canberra being most frequent. Potential location bias.

	row ID	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am
count	99516	99516	99073.000000	99286.000000	98537.000000	56985.00000	52199.000000	92995	93036.000000	92510
unique	99516	49	NaN	NaN	NaN	NaN	NaN	16	NaN	16
top	Row0	Canberra	NaN	NaN	NaN	NaN	NaN	W	NaN	N
freq	1	2393	NaN	NaN	NaN	NaN	NaN	6843	NaN	8052
mean	NaN	NaN	12.176266	23.218513	2.353024	5.46132	7.615090	NaN	39.976966	NaN
std	NaN	NaN	6.390882	7.115072	8.487866	4.16249	3.783008	NaN	13.581524	NaN
min	NaN	NaN	-8.500000	-4.100000	0.000000	0.00000	0.000000	NaN	6.000000	NaN
25%	NaN	NaN	7.600000	17.900000	0.000000	2.60000	4.800000	NaN	31.000000	NaN
50%	NaN	NaN	12.000000	22.600000	0.000000	4.80000	8.400000	NaN	39.000000	NaN
75%	NaN	NaN	16.800000	28.200000	0.800000	7.40000	10.600000	NaN	48.000000	NaN
max	NaN	NaN	33.900000	48.100000	371.000000	86.20000	14.500000	NaN	135.000000	NaN

	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow
count	98283.000000	97010.000000	89768.000000	89780.000000	61944.000000	59514.000000	98902.000000	97612.000000	98537	99516.000000
unique	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2	NaN
top	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	No	NaN
freq	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	76481	NaN
mean	68.866376	51.433296	1017.684638	1015.286204	4.447985	4.519122	16.970041	21.681340	NaN	0.224677
std	19.074951	20.777616	7.110166	7.045189	2.886580	2.716618	6.488961	6.931681	NaN	0.417372
min	0.000000	0.000000	980.500000	978.200000	0.000000	0.000000	-7.000000	-5.100000	NaN	0.000000
25%	57.000000	37.000000	1013.000000	1010.500000	1.000000	2.000000	12.300000	16.600000	NaN	0.000000
50%	70.000000	52.000000	1017.700000	1015.300000	5.000000	5.000000	16.700000	21.100000	NaN	0.000000
75%	83.000000	65.000000	1022.400000	1020.000000	7.000000	7.000000	21.500000	26.400000	NaN	0.000000
max	100.000000	100.000000	1041.000000	1039.600000	9.000000	9.000000	40.200000	46.700000	NaN	1.000000

AUSTRALIAN WEATHER DATA

DATA CLEANING (R2)

As we saw, this dataset contains many **missing values**, making it crucial to clean and properly utilize it. Here are some of the **preprocessing steps** we have applied:

- **Removed** all rows with **missing values**, leaving 39,574 rows and 23 columns of complete data.
- Ensured that important weather variables like temperature, rainfall, wind direction, humidity, and pressure are retained without gaps.
- Applied **undersampling** to address **potential imbalances** in the dataset, particularly in overrepresented categories, **ensuring a balanced and manageable dataset for further analysis**.

These preprocessing steps ensure the dataset is ready for reliable and accurate analysis.

```
Handling Missing Values (Dropping Rows with Missing Values, excluding date):
Shape after dropping missing values: (39574, 23)
First 5 rows of the cleaned dataset (excluding date):
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

	row ID	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	\
4183	Row4262	Cobar	17.9	35.2	0.0	12.0	12.3	
4185	Row4264	Cobar	27.1	36.1	0.0	13.0	0.0	
4186	Row4265	Cobar	23.3	34.0	0.0	9.8	12.6	
4187	Row4266	Cobar	16.1	34.2	0.0	14.6	13.2	
4188	Row4267	Cobar	19.0	35.5	0.0	12.0	12.3	

	WindGustDir	WindGustSpeed	WindDir9am	...	Humidity9am	Humidity3pm	\
4183	SSW	48.0	ENE	...	20.0	13.0	
4185	N	43.0	N	...	26.0	19.0	
4186	SSW	41.0	S	...	33.0	15.0	
4187	SE	37.0	SE	...	25.0	9.0	
4188	ENE	48.0	ENE	...	46.0	28.0	

	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	\
4183	1006.3	1004.4	2.0	5.0	26.6	33.4	
4185	1007.7	1007.4	8.0	8.0	30.7	34.3	
4186	1011.3	1009.9	3.0	1.0	25.0	31.5	
4187	1013.3	1009.2	1.0	1.0	20.7	32.8	
4188	1008.3	1004.0	1.0	5.0	23.4	33.3	

	RainToday	RainTomorrow
4183	No	0
4185	No	0
4186	No	0
4187	No	0
4188	No	0

[5 rows x 23 columns]

AUSTRALIAN WEATHER DATA

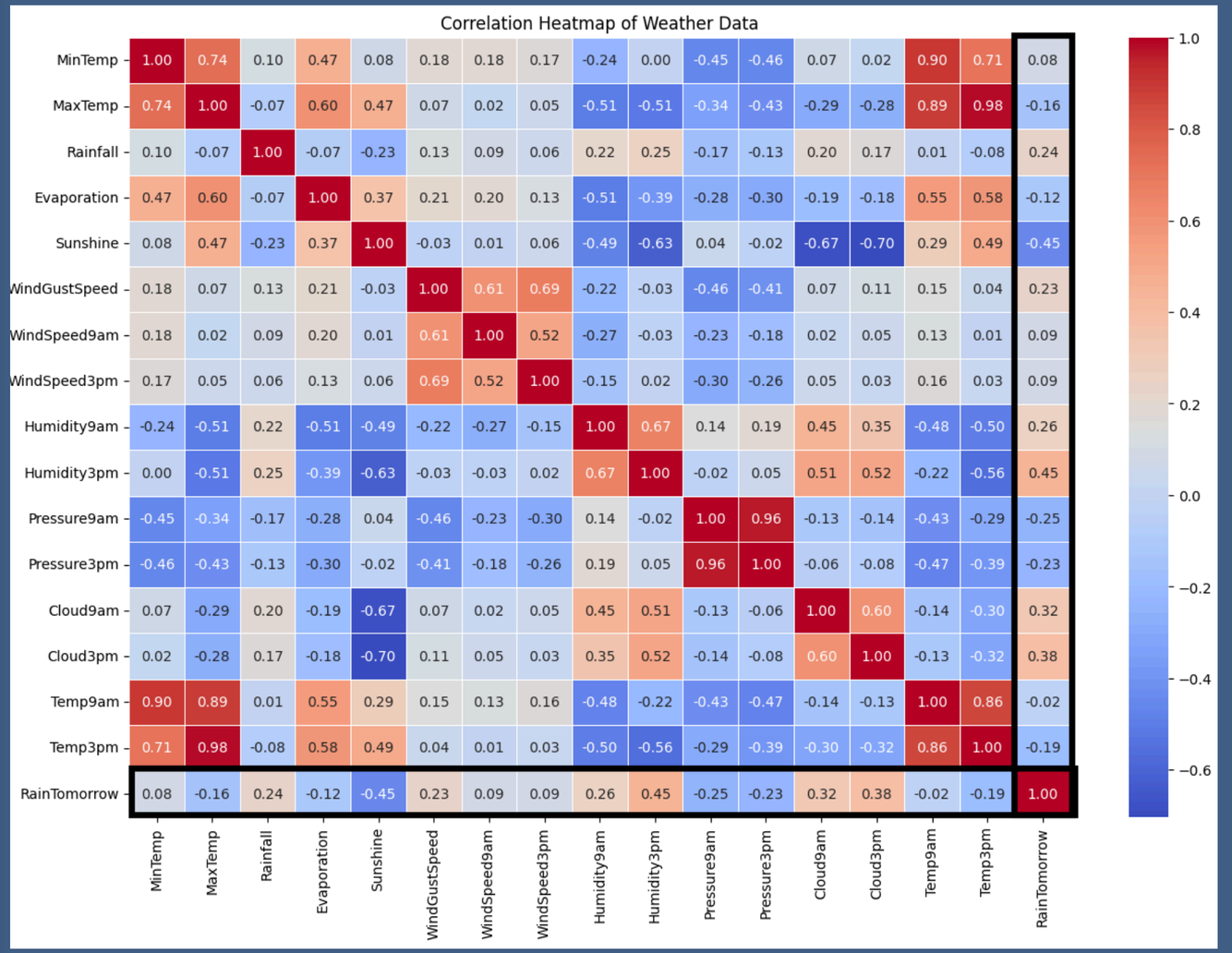
CORRELATION BETWEEN ATTRIBUTES (R2)

The **correlation heatmap** shows the relationship between the **RainTomorrow** attribute with **other attributes** in the dataset.

We intend on using the **strength of correlation** between attributes to **identify** the strongest **links** to make use of within the classification **model**.

Correlations with **RainTomorrow** identified that are worth exploring are:

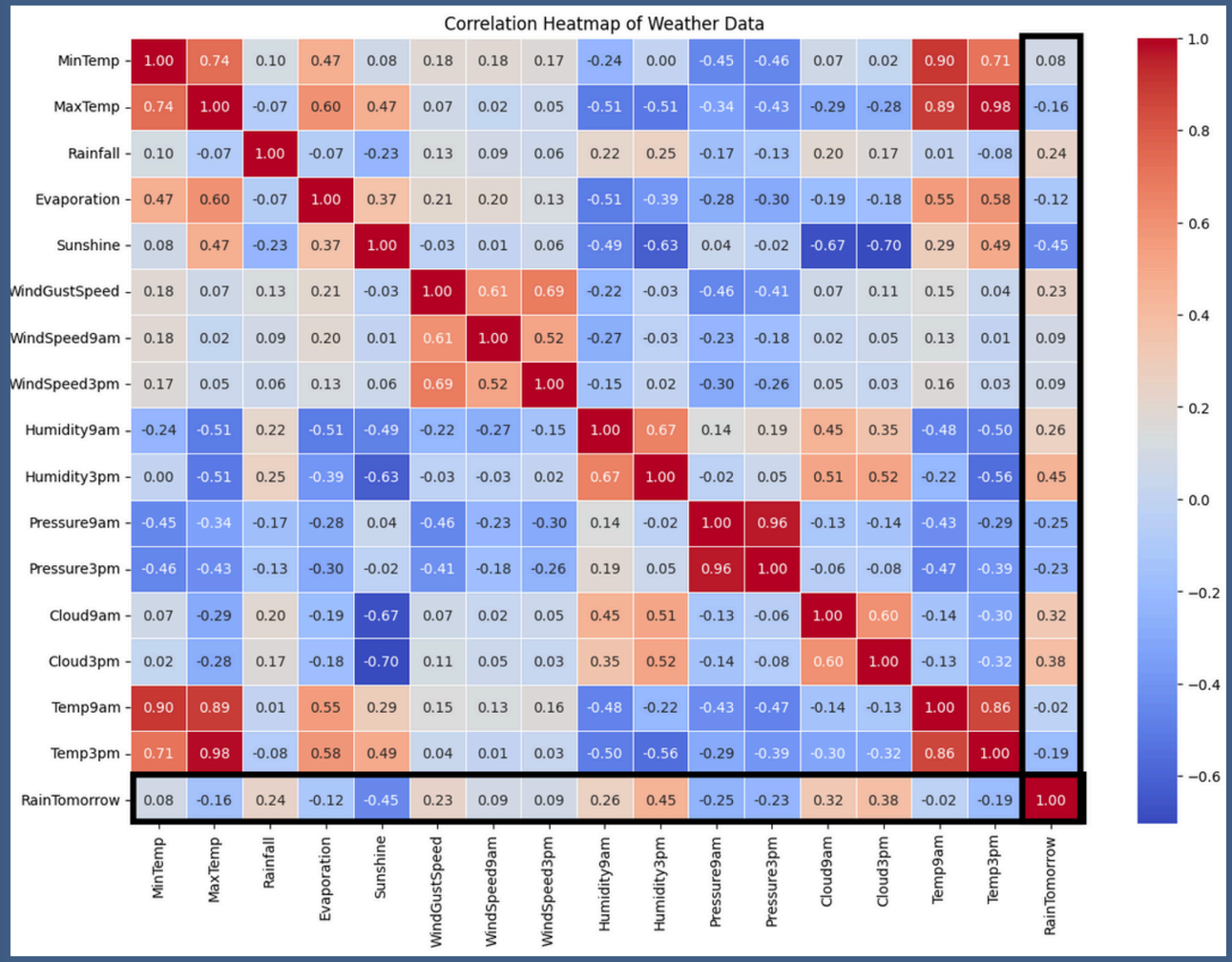
- **Sunshine** with a **negative correlation** of **-0.45**.
- **Humidity** at 3pm with a correlation of **0.45**.
- **Cloud at 3pm** with a correlation of **0.38**.
- **Cloud at 9am** with a correlation of **0.32**.



AUSTRALIAN WEATHER DATA

LOTS OF WEAK CORRELATIONS A BAD SIGN? (R2)

Due to the size of our dataset, our analysis is **less affected** by lower-valued correlations, as the larger sample size helps to **minimize the impact of noise** or random variations. This means that even though some features may show weaker correlations with the target variable, the overall predictive power of the model can still be strong, as more **significant patterns** are likely to emerge from the data.



AUSTRALIAN WEATHER DATA

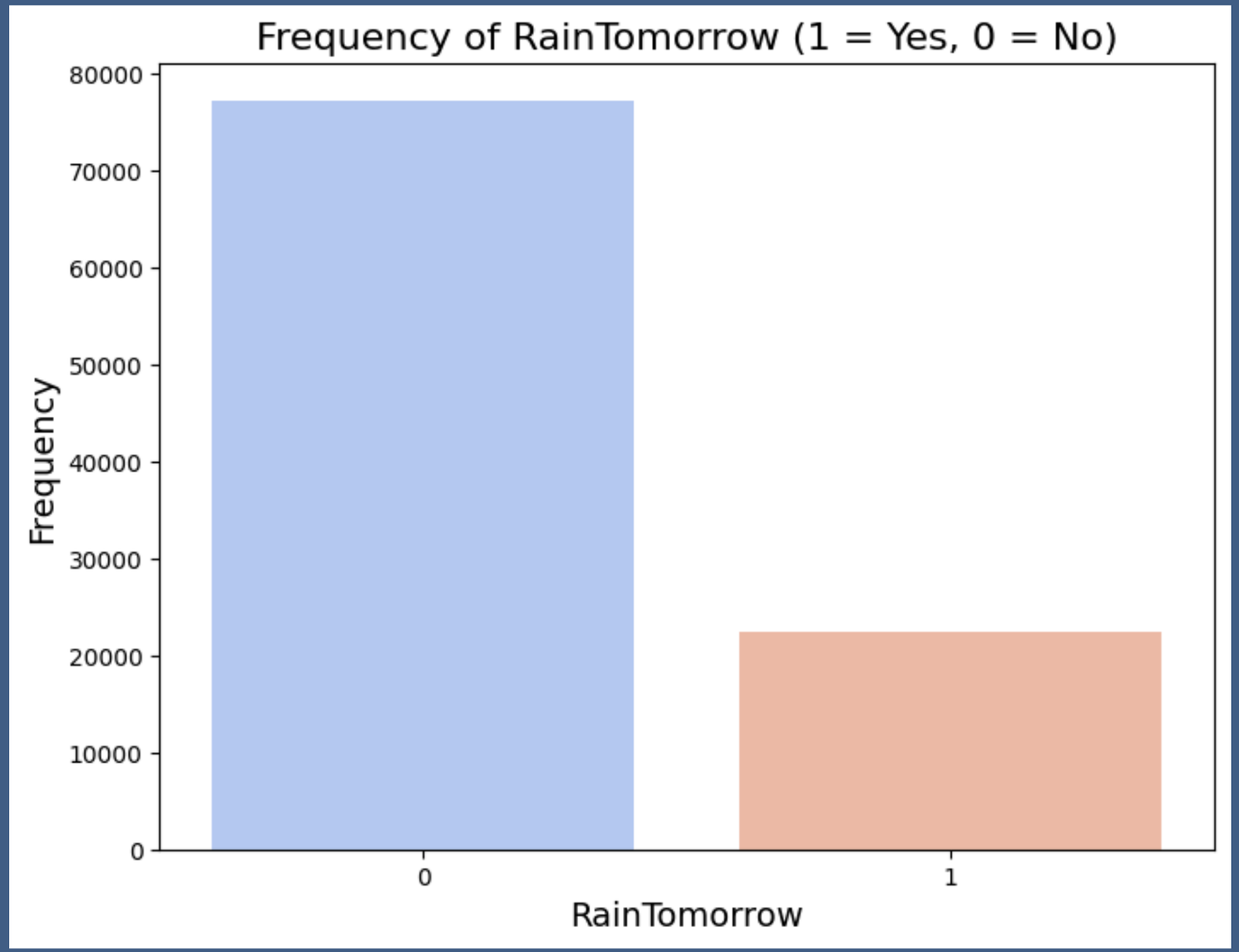
HOW MUCH DATA IS MISSING? (R2)

Data analysis has revealed unbalanced data when it comes to frequency of occurence of rain the next day.

As it stands, the data shows that:

- The occurence of a 0 (no rain tomorrow) is 77157.
- The occurence that it is 1 (it does rain tomorrow) is 22359.

This imbalance is something we have to deal with through methods such as **oversampling** or **undersampling** in order to select a standard size of attributes to fairly train the models.

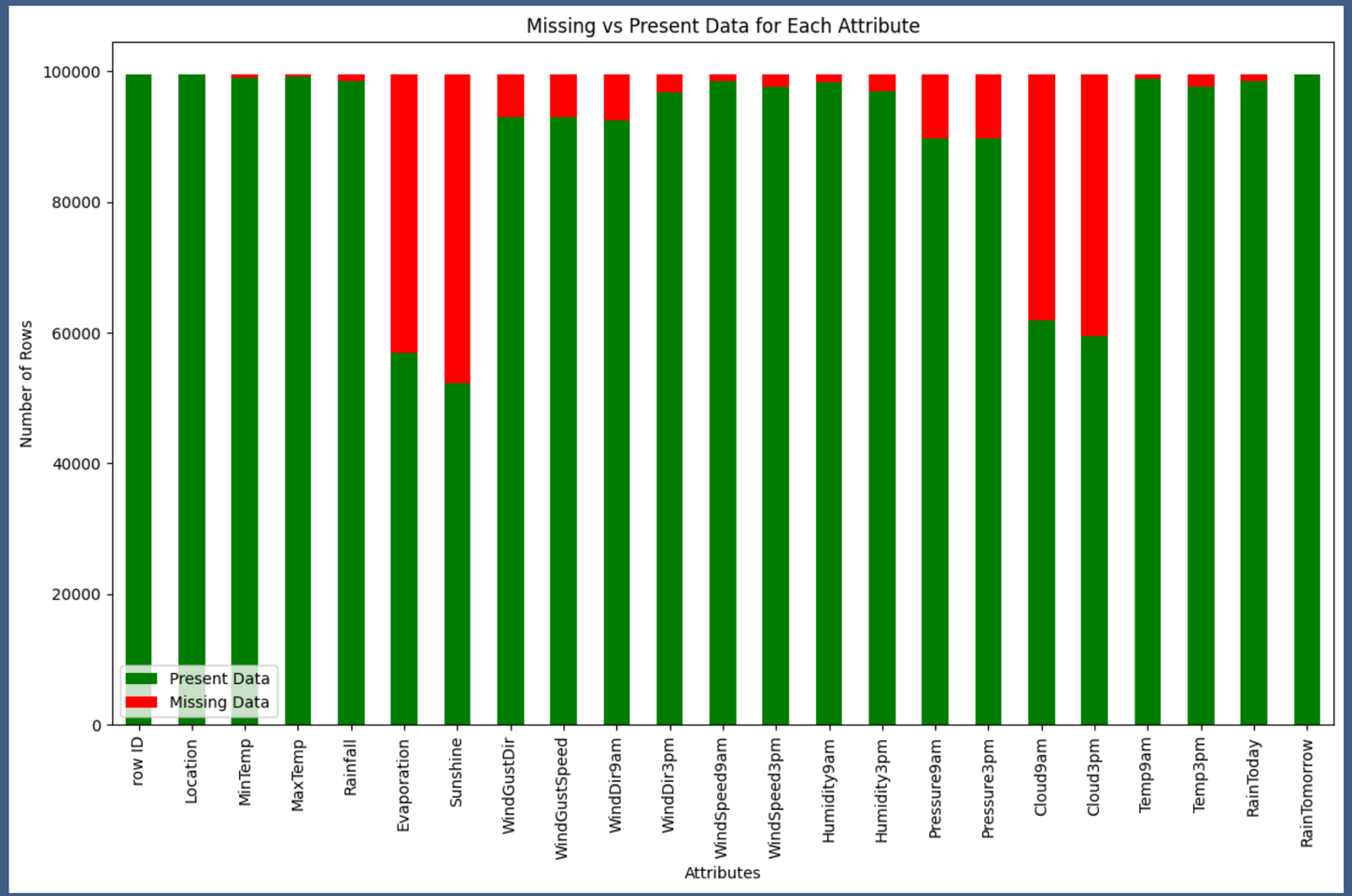


AUSTRALIAN WEATHER DATA

HOW MUCH DATA IS MISSING? (R2)

Most attributes have minimal missing data.

- Significant missing data in key attributes correlated with RainTomorrow:
  - **Sunshine**: 52,199 entries (47.5%)
  - **Cloud9am**: 61,944 entries (37.7%)
  - **Cloud3pm**: 59,514 entries (40.1%)
- Non-significant attributes like **Evaporation** also have large gaps.
- The **large dataset** helps **mitigate the impact** of missing values.
- Techniques like **data imputation** can manage missing data by **estimating values**.
- The dataset size allows for **identifying patterns** despite missing data.



WEATHER IMAGE CLASSIFICATION

WHY IS IT NEEDED?

- WHO: 285 million visually impaired, 39 million blind.
- Deep learning weather image classification offers real-time **auditory/tactile** updates .
- CNN-based system alerts **visually impaired individuals** to weather changes like **storms** or **rain**.
- Enhances **safety** and **independence** by providing critical weather info without **visual cues** (Shidore et al., 2023).

Image source: https://yaleclimateconnections.org/2023/03/with-global-warming-of-just-1-2c-why-has-the-weather-gotten-so-extreme/

WEATHER IMAGE CLASSIFICATION REQUIREMENTS

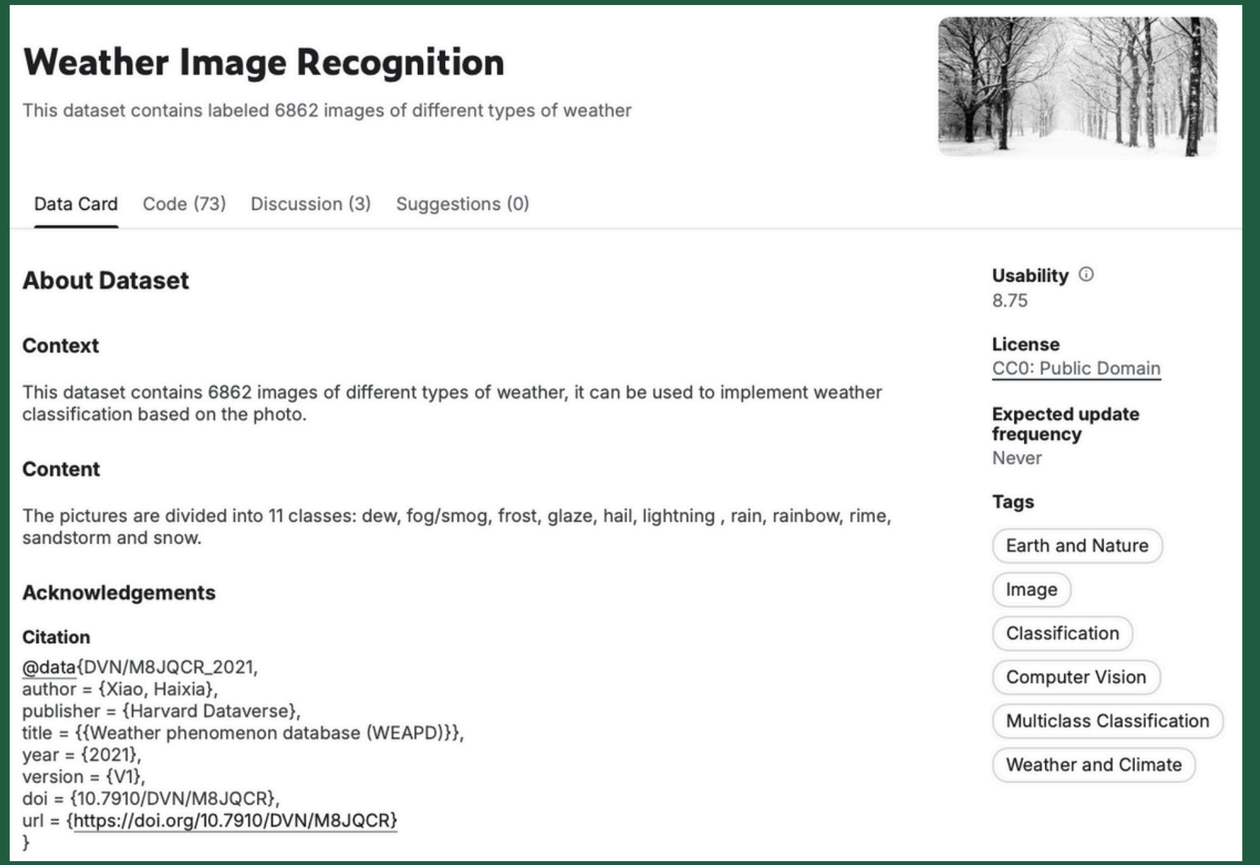
- The **requirements** we intend to satisfy with the work done with this dataset as per the rubrics of our course are:
- **R1**: Project topic and directions
- **R2**: Data Analysis
- **R3**: Clustering
- **R4**: Basic classifiers and Decision Trees
- **R5**: Neural Networks

Image source: https://yaleclimateconnections.org/2023/03/with-global-warming-of-just-1-2c-why-has-the-weather-gotten-so-extreme/

WEATHER IMAGE CLASSIFICATION (R1)

DATASET 3

- The dataset is sourced from **Kaggle.**
- It is licensed under the **Creative Commons CC0: Public Domain license**.
- The dataset contains **6862 images** of different types of weather divided into 11 different classes.
- The datasets is published by **Harvard Dataverse**.
- It is a labelled image data set of size **615MB**.



Weather Image Recognition

This dataset contains labeled 6862 images of different types of weather

Data Card | Code (73) | Discussion (3) | Suggestions (0)

About Dataset

Context

This dataset contains 6862 images of different types of weather, it can be used to implement weather classification based on the photo.

Content

The pictures are divided into 11 classes: dew, fog/smog, frost, glaze, hail, lightning , rain, rainbow, rime, sandstorm and snow.

Acknowledgements

Citation

@data{DVN/M8JQCR_2021,
author = {Xiao, Haixia},
publisher = {Harvard Dataverse},
title = {{Weather phenomenon database (WEAPD)}},
year = {2021},
version = {V1},
doi = {10.7910/DVN/M8JQCR},
url = {https://doi.org/10.7910/DVN/M8JQCR}
}

Usability 8.75

License CC0: Public Domain

Expected update frequency Never

Tags

Earth and Nature | Image | Classification | Computer Vision | Multiclass Classification | Weather and Climate

Source : https://www.kaggle.com/datasets/jehanbhathena/weather-dataset

WEATHER IMAGE CLASSIFICATION (R1)

DATASET 3

- We plan to use classification models with CNN to **predict what weather** is shown in the image.
- Given the data distribution, our model may become **biased towards certain classes**, like 'Fog/smog' and 'Rime.'
- We may face challenges in training on **under-represented classes**.
- This dataset contains 6862 images and 11 classes to work with.
- Since it is a **Harvard Dataverse Dataset**, it adds **credibility**.

Similar work has been done on it before, which shows that it is possible to make **classification model** on this.

Source : https://www.kaggle.com/datasets/jehanbhathena/weather-dataset

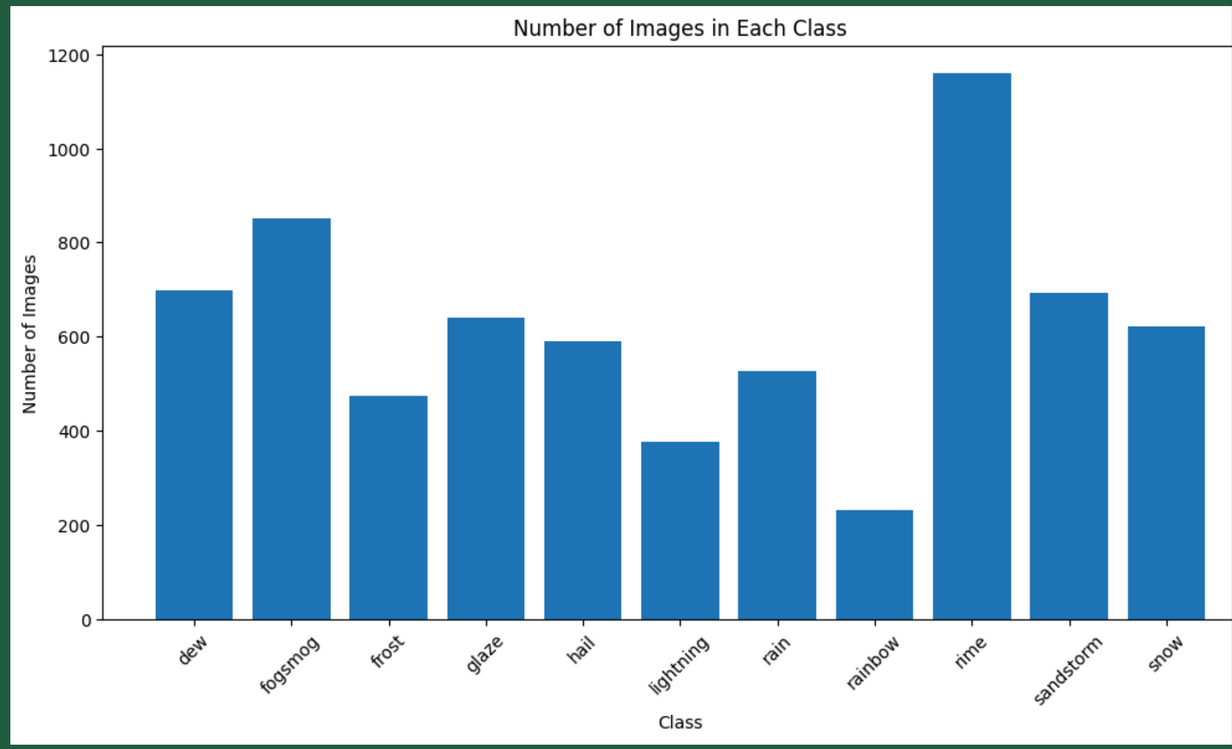
WEATHER IMAGE DATASET

DATA DISTRIBUTION (R2)

This **histogram** shows the number of images in each weather class folder.

- Some classes, like 'Fog/smog' and 'Rime,' have **significantly more images** than others.
- The data distribution may cause our model to become **biased** toward over-represented classes.
- Under-represented classes, like 'Rainbow' and 'Lightning,' may **present challenges** during training.

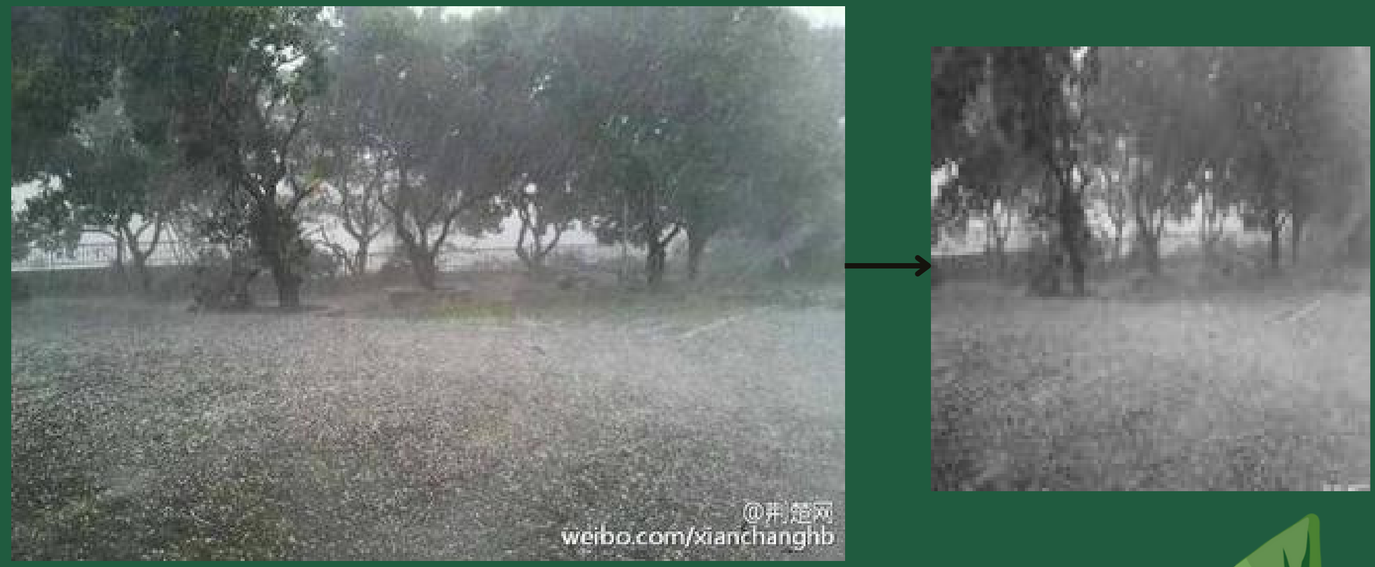
We may need to address this **imbalance** through techniques like **oversampling** or **class weighting** to improve model performance.



WEATHER IMAGE CLASSIFICATION (R2)

DATASET 3

- Images have been **preprocessed** by:
  - Cropped bottom **50px** of all images to deal with overlaying text.
  - Converted to **grayscale**.
  - Rescaled to **128 x 128 pixels**.
- We intend to use this processed data to further **classify** between multiple weather images.

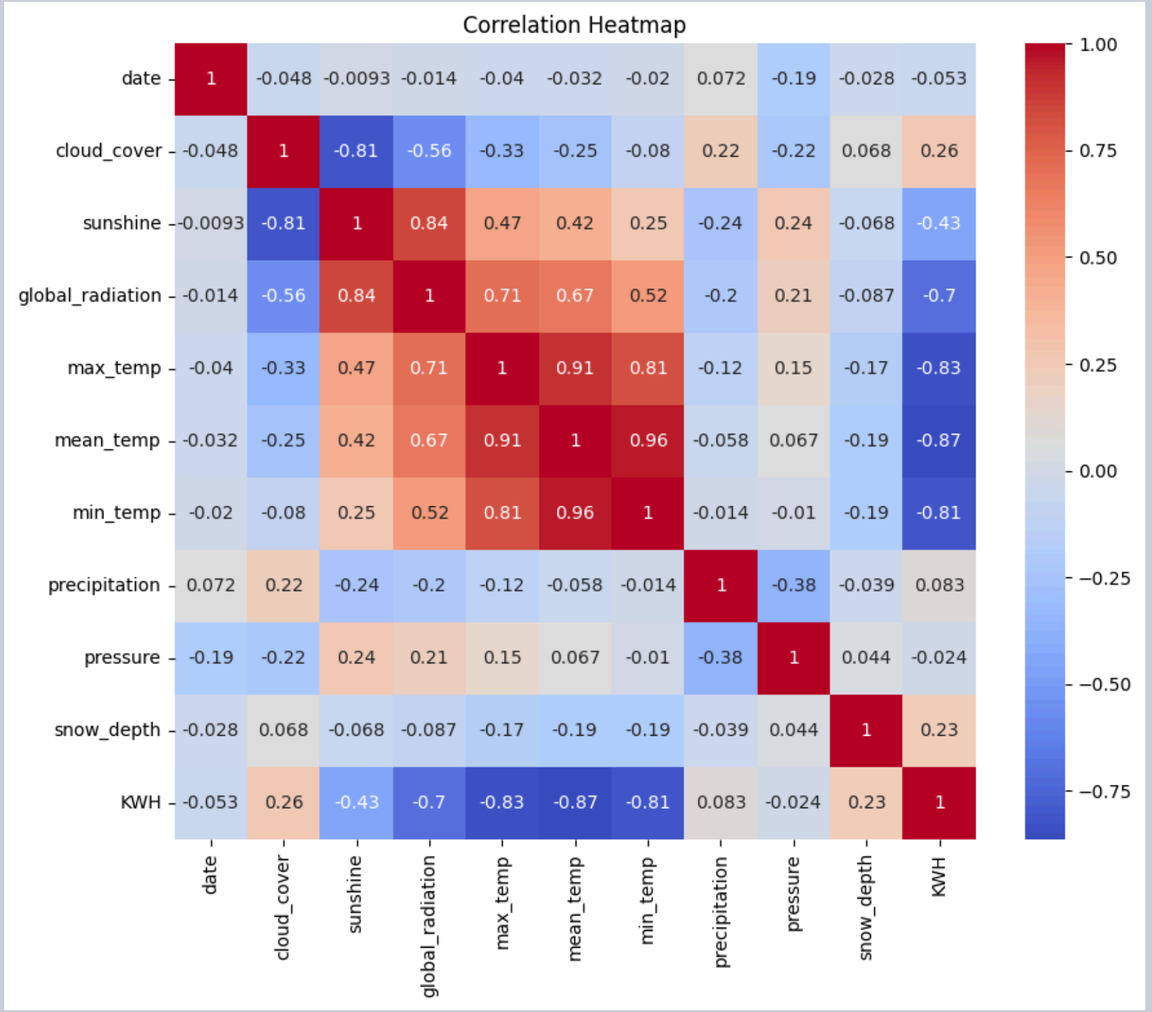


Dataset Exploration (R2)

While selecting topics, we conducted an **analysis** of several **other datasets** and topics. This is our analysis explaining why we chose not to use these datasets.

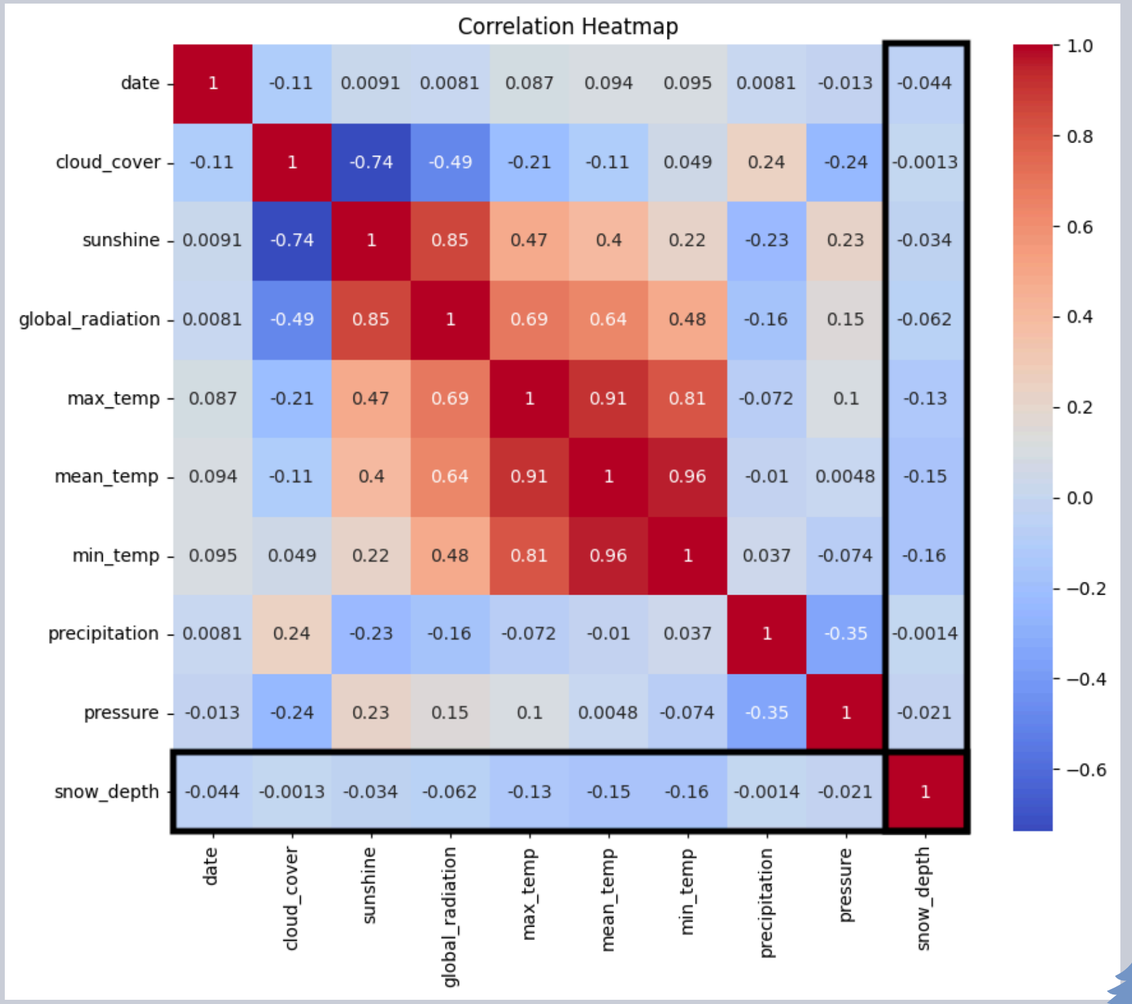
ENERGY CONSUMPTION PREDICTION PER HOUSEHOLD

- Conducted data analysis on the "**London Energy Data**."
- Strong correlation but...
- Limited to **three years.**
- **Less accuracy**.



LONDON WEATHER DATA

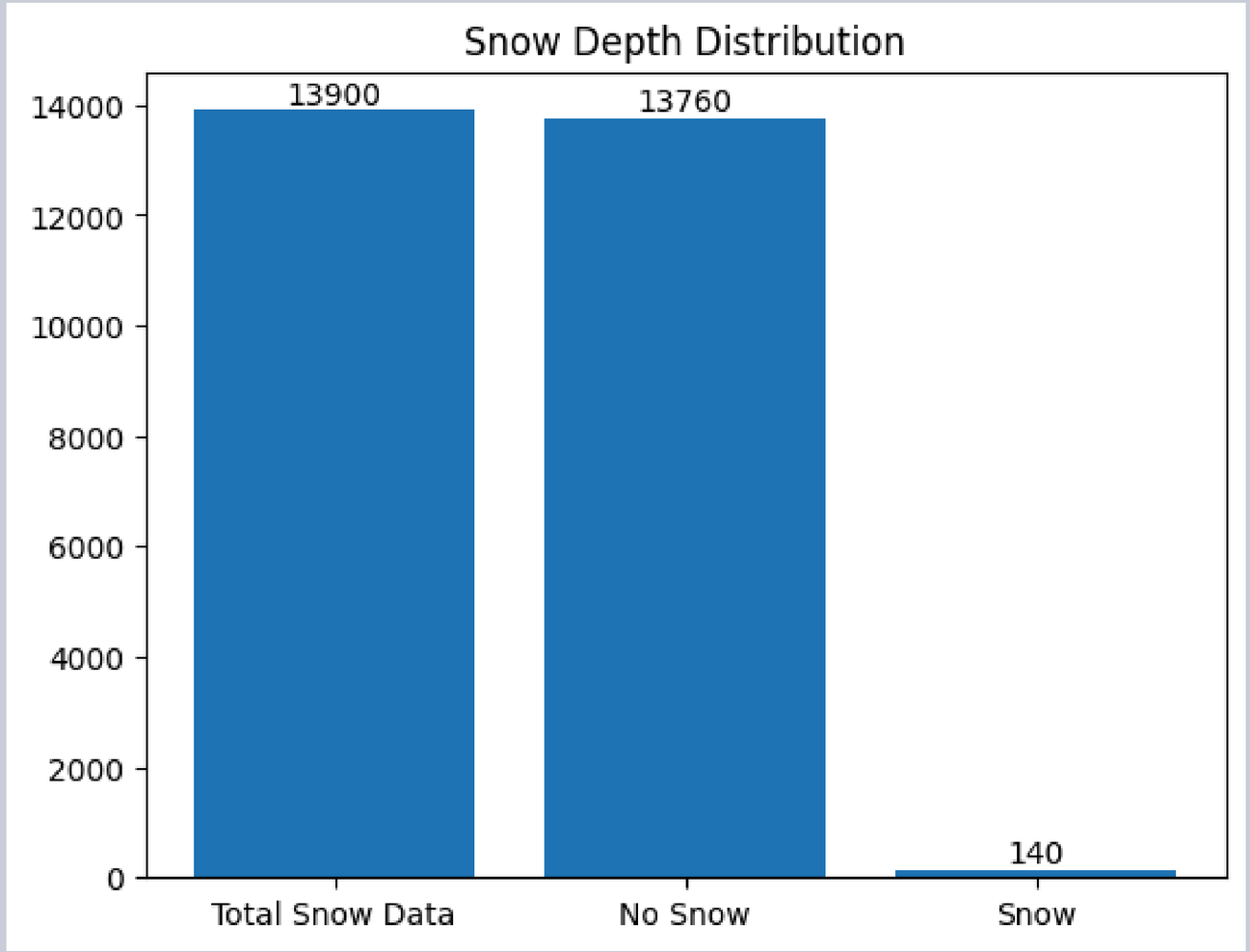
- Conducted data analysis on the "London Weather Data."
- Weak correlations.
- Little Impact on overall weather patterns in this dataset.



LONDON WEATHER DATA

This histogram shows the distribution of snow depth data.

- There is **minimal data** available for the **"Snow" class** (only 140 entries).
- The lack of snow data may lead to a biased model, **heavily favoring "No Snow"** cases.
- The absence of snow in recent data could be due to global warming, explaining the imbalance.



Thank You !

References

Lane, C., Cappuccio, G. and SolarReviews (2024) How Much Energy Does A Solar Panel Produce?, SolarReviews. Available at: https://www.solarreviews.com/blog/how-much-electricity-does-a-solar-panel-produce (Accessed: 3 October 2024).

Salas, E.B. (2024) Countries most exposed to river floods worldwide 2024, Statista. Available at: https://www.statista.com/statistics/1306264/countries-most-exposed-to-floods-by-risk-index-score/ (Accessed: 3 October 2024)

Shidore, M. et al. (2023) 'Eye for Blind: A Deep Learning-Based Sensory Navigation System for the Blind', in S. Fong, N. Dey, and A. Joshi (eds) ICT Analysis and Applications. Singapore: Springer Nature, pp. 229–240. Available at: https://doi.org/10.1007/978-981-99-6568-7_21.

Wayne W., L. and Boston University School of Public Health (2021) The Correlation Coefficient (r). Available at: https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module9-Correlation-Regression/PH717-Module9-Correlation-Regression4.html (Accessed: 3 October 2024).