# Git Set Go

Siddhi Jadhav Sanch Anand Thi Lan Anh Tran

2025-05-27

## Table of contents



## 1 Cleaning the data

```
# Load the dataset
data <- read.csv("data/weatherAUS.csv")
```

```
# 1. Drop high-missing columns
data <- data %>%
  select(-Evaporation, -Sunshine, -Cloud9am, -Cloud3pm)
```

```
# 2. Convert 'Date' to Date type
data <- data %>%
  mutate(Date = ymd(Date))
```

```
# 3. Drop rows where 'RainTomorrow' is missing
data <- data %>%
  filter(!is.na(RainTomorrow))
```

```
# 4. Drop wind direction columns
data <- data %>%
  select(-WindGustDir, -WindDir9am, -WindDir3pm)
```

```r
# 5. Convert RainToday and RainTomorrow to binary 0/1
data <- data %>%
  mutate(
    RainToday = if_else(RainToday == "Yes", 1, 0),
    RainTomorrow = if_else(RainTomorrow == "Yes", 1, 0)
  )

# 6. Drop any remaining NAs
data <- data %>% drop_na()
```

```r
# 7. Save cleaned data
write_csv(data, "data/data_cleaned.csv")
```

## 2 Methodology

The dataset used for this analysis, weatherAUS.csv, contains over 120,000 daily weather observations recorded across various Australian cities. To ensure data quality and usability, we cleaned the dataset by removing columns with a high number of missing values, such as Evaporation and Sunshine, and imputing the remaining missing values using the median for numerical variables. Categorical variables were imputed using the mode, where necessary. Wind direction columns were excluded due to their limited relevance to the city-level comparative focus. The resulting cleaned dataset, consisting of 16 variables, enabled us to conduct meaningful comparisons of weather characteristics across regions.

The analysis focused on four primary variables: **Rainfall**, **Temperature at 3pm**, **Humidity at 3pm**, and **Pressure at 9am** chosen to represent precipitation, thermal patterns, atmospheric moisture, and pressure systems respectively. The data was grouped by `Location`, enabling city-wise summary statistics and comparisons. To ensure consistency and reduce noise, we focused on cities with the highest number of observations.

To visualize precipitation differences across cities, we generated a bar plot of the ten cities with the highest average daily rainfall (Figure 1). This figure highlights distinct tropical patterns, especially in Darwin and Cairns. To complement the visual insight, a detailed summary of weather metrics including temperature, humidity, and pressure for these top 10 wettest cities is presented in Table 1.

All analysis was conducted using the `tidyverse` and `ggplot2` packages in R, structured within a reproducible Quarto project. GitHub was used to manage version control and collaboration, with each contributor working in individual branches.

```r
top_rainfall <- data %>%
  group_by(Location) %>%
  summarise(avg_rain = mean(Rainfall, na.rm = TRUE)) %>%
  arrange(desc(avg_rain)) %>%
  slice_head(n = 10)

ggplot(top_rainfall, aes(x = reorder(Location, avg_rain), y = avg_rain)) +
  geom_col(fill = "burlywood3") +
  coord_flip() +
  labs(x = "City", y = "Average Rainfall (mm)")
```

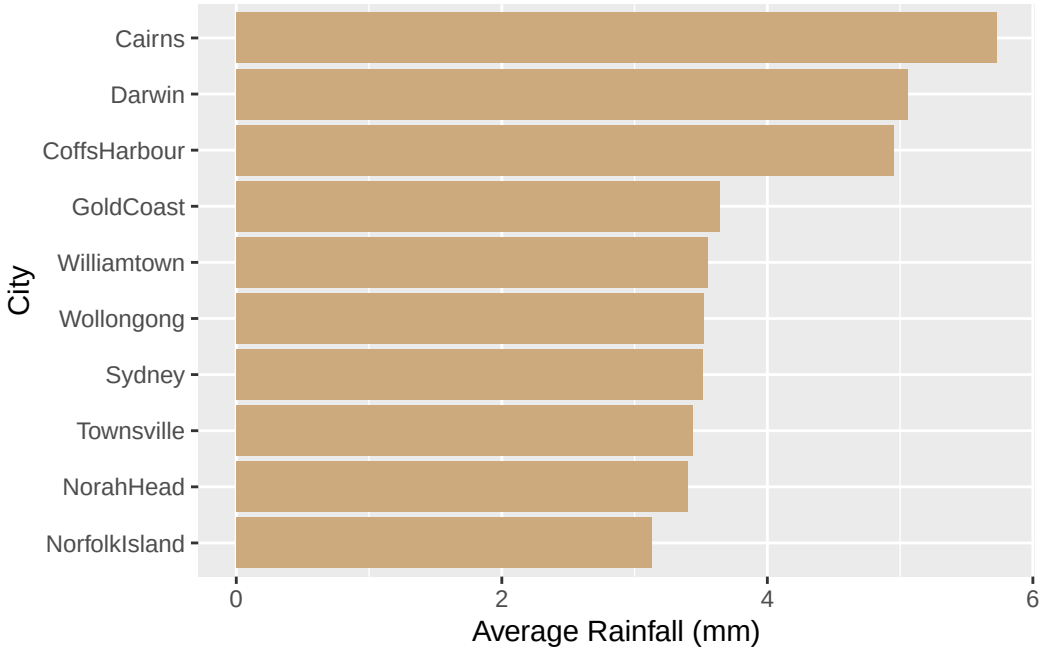


Figure 1: Top 10 cities by average daily rainfall (mm)

```r
top_cities <- top_rainfall$Location

data %>%
  filter(Location %in% top_cities) %>%
  group_by(Location) %>%
  summarise(
    `Avg Rainfall` = round(mean(Rainfall), 1),
    `Avg Temp (3pm)` = round(mean(Temp3pm), 1),
    `Avg Humidity (3pm)` = round(mean(Humidity3pm), 1),
```

```
    `Avg Pressure (9am)` = round(mean(Pressure9am), 1)
  ) %>%
  knitr::kable()
```

Table 1: Summary of key weather variables in top 10 wettest cities

| Location | Avg Rainfall | Avg Temp (3pm) | Avg Humidity (3pm) | Avg Pressure (9am) |
|---|---|---|---|---|
| Cairns | 5.7 | 27.9 | 61.7 | 1014.2 |
| CoffsHarbour | 5.0 | 22.3 | 62.3 | 1018.3 |
| Darwin | 5.1 | 31.1 | 51.7 | 1011.9 |
| GoldCoast | 3.6 | 23.7 | 62.8 | 1018.0 |
| NorahHead | 3.4 | 20.8 | 67.5 | 1018.3 |
| NorfolkIsland | 3.1 | 20.4 | 67.8 | 1017.7 |
| Sydney | 3.5 | 21.8 | 54.3 | 1018.3 |
| Townsville | 3.4 | 27.8 | 57.3 | 1015.2 |
| Williamtown | 3.6 | 22.7 | 53.2 | 1018.4 |
| Wollongong | 3.5 | 19.9 | 65.1 | 1018.1 |