Weather Pattern Analysis Across 10 U.S. Cities

Amoni Haynes

2025-04-05

Introduction

This case study analyzes weather patterns across 10 major U.S. cities. The goal is to identify trends in temperature, humidity, precipitation, and wind speed, and how these factors vary over time and location.

Dataset Overview

```
library(dplyr)
library(ggplot2)
library(readr)
```

Loading the Data

```
weather_data <- read_csv("weather_data.csv")
```

```
## Rows: 1000000 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): Location
## dbl  (4): Temperature_C, Humidity_pct, Precipitation_mm, Wind_Speed_kmh
## dttm (1): Date_Time
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(weather_data)
```

```
## # A tibble: 6 x 6
##   Location     Date_Time           Temperature_C Humidity_pct Precipitation_mm
##   <chr>        <dttm>                      <dbl>        <dbl>             <dbl>
## 1 San Diego    2024-01-14 21:12:46         10.7          41.2              4.02
## 2 San Diego    2024-05-17 15:22:10          8.73         58.3              9.11
## 3 San Diego    2024-05-11 09:30:59         11.6          38.8              4.61
## 4 Philadelphia 2024-02-26 17:32:39         -8.63         54.1              3.18
## 5 San Antonio  2024-04-29 13:23:51         39.8          72.9              9.60
## 6 San Diego    2024-01-21 08:54:56         27.3          49.0              9.17
## # i 1 more variable: Wind_Speed_kmh <dbl>
```

```r
summary(weather_data)
```

```
##    Location           Date_Time                      Temperature_C
##  Length:1000000     Min.   :2024-01-01 00:00:06.00   Min.   :-19.97
##  Class :character   1st Qu.:2024-02-04 16:28:23.75   1st Qu.:  2.27
##  Mode  :character   Median :2024-03-10 11:43:28.00   Median : 14.78
##                     Mean   :2024-03-10 10:40:58.90   Mean   : 14.78
##                     3rd Qu.:2024-04-14 03:51:32.50   3rd Qu.: 27.27
##                     Max.   :2024-05-18 19:44:10.00   Max.   : 40.00
##   Humidity_pct   Precipitation_mm    Wind_Speed_kmh
##  Min.   :30.00   Min.   : 0.000009   Min.   : 0.000051
##  1st Qu.:45.01   1st Qu.: 2.580694   1st Qu.: 7.490101
##  Median :60.02   Median : 5.109917   Median :14.993777
##  Mean   :60.02   Mean   : 5.109639   Mean   :14.997598
##  3rd Qu.:75.04   3rd Qu.: 7.613750   3rd Qu.:22.514110
##  Max.   :90.00   Max.   :14.971583   Max.   :29.999973
```

Data Cleaning

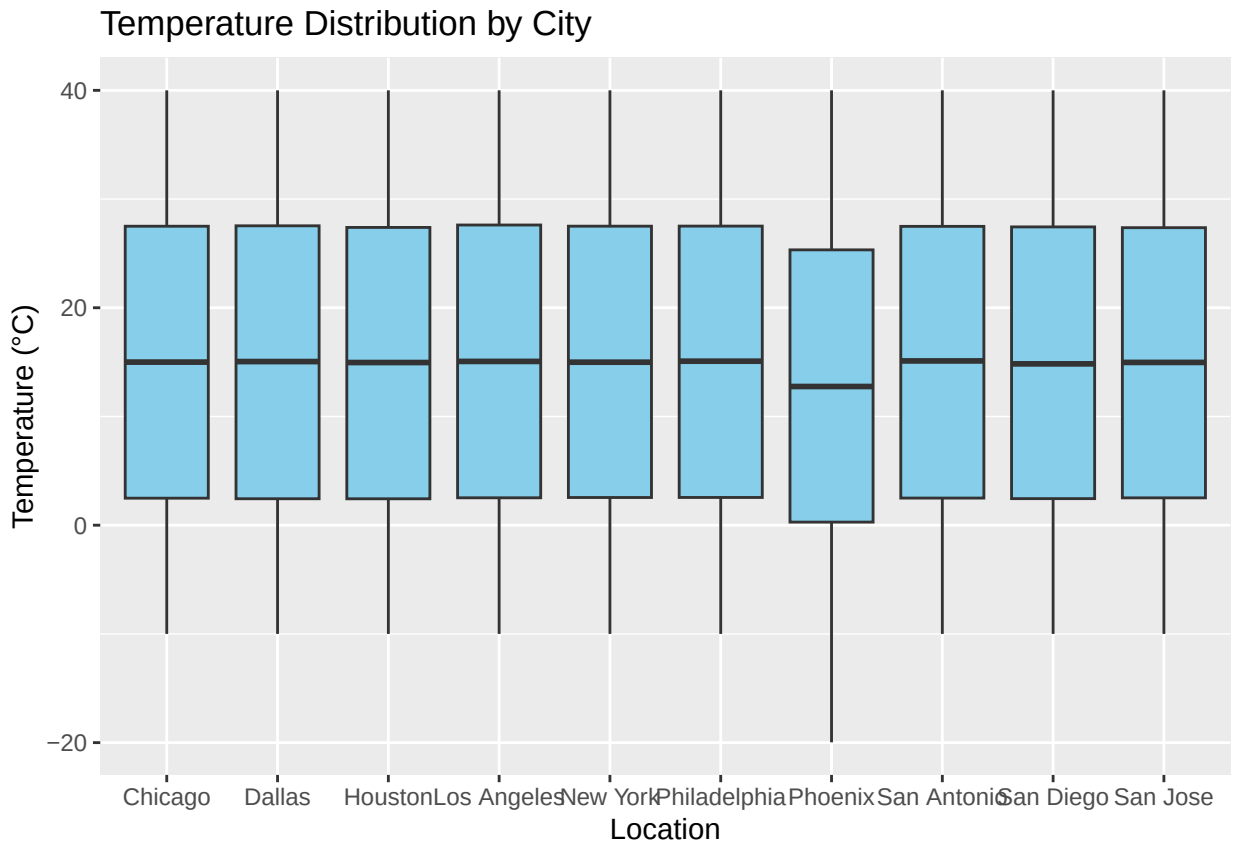
- Converting date-time format
- Checking for missing values
- Removing duplicates

```r
weather_data$Date_Time <- as.POSIXct(weather_data$Date_Time)
weather_data <- distinct(weather_data)
weather_data <- na.omit(weather_data)  # Removes rows with missing values
```

Exploratory Data Analysis

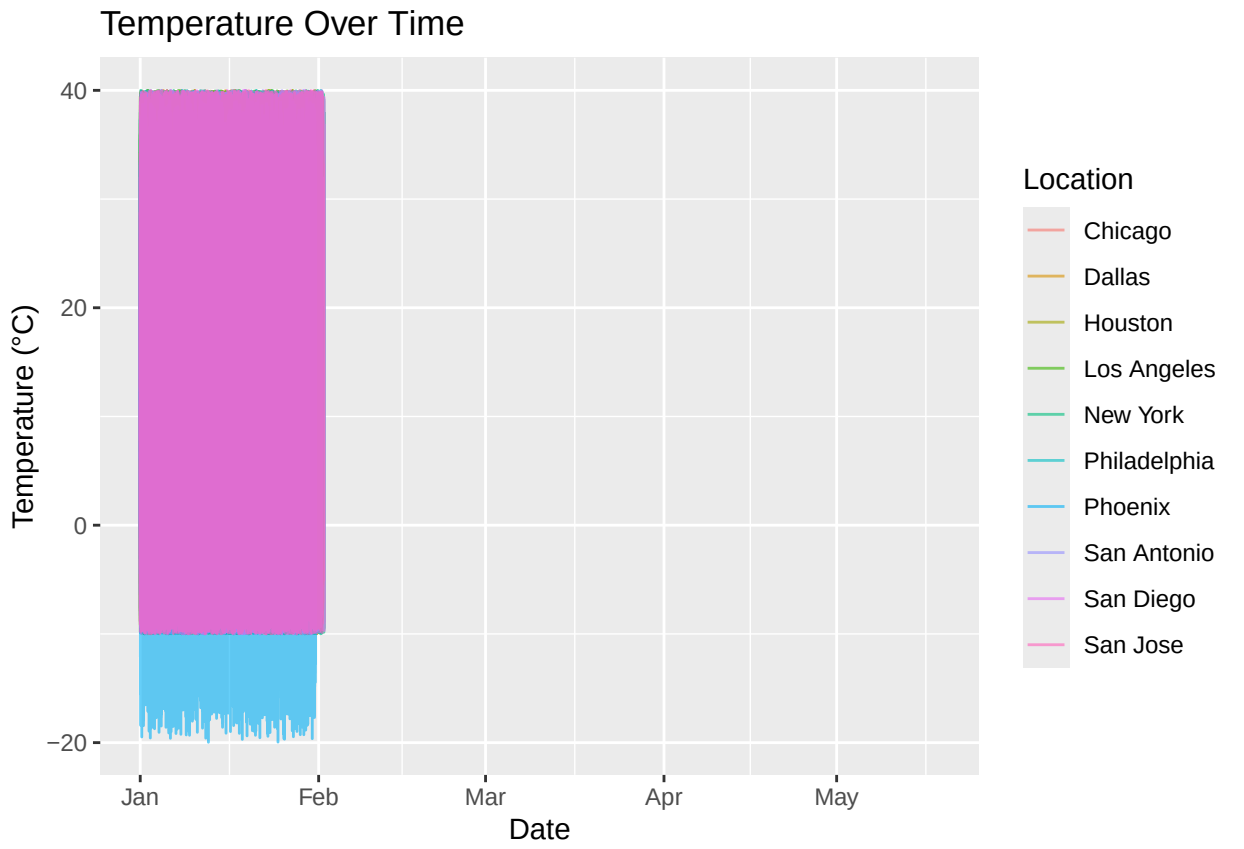
1. Temperature Trends by City

```r
ggplot(weather_data, aes(x = Location, y = Temperature_C)) +
  geom_boxplot(fill = "skyblue") +
  labs(title = "Temperature Distribution by City", y = "Temperature (°C)")
```



2. Temperature Over Time

```
ggplot(weather_data, aes(x = Date_Time, y = Temperature_C, color = Location)) +
  geom_line(alpha = 0.6) +
  labs(title = "Temperature Over Time", x = "Date", y = "Temperature (°C)")
```



Key Findings

- **Hottest city:** Los Angeles, 15.0810627490806
- **Most humid city:** Chicago, 60.0887976453709

Conclusion & Recommendations

- Locations with higher temperatures should prepare for heatwaves.
- Cities with high humidity levels might experience more rainfall.
- Future research can analyze the impact of these weather patterns on agriculture and energy consumption.

Exporting Cleaned Data:

```
write.csv(weather_data, "cleaned_weather_data.csv", row.names = FALSE)
```