The Importance of Outlier Detection With Applications in R

Rajesh Mahato - 400878749

About the Analysis

Packages Used:

- `dplyr` - Data manipulation
- `ggplot2` - Data visualization

Air Quality Dataset:

- 153 observations
- 6 variables
- Target: `Ozone Levels` +
- Context: Daily air quality in New York (May-Sept 1973)

What Are Outliers?

Definition: Data points that differ significantly from other observations in a dataset.

Types of Outliers:

1. **Point Outliers** - Single extreme values
2. **Contextual Outliers** - Unusual in specific contexts
3. **Collective Outliers** - Groups of unusual points

Causes:

- Measurement errors
- Data entry mistakes
- Natural variation
- Fraudulent activities
- Equipment malfunction

Why Outlier Detection Matters

Environmental Importance:

- **Protects Public Health**: Early warnings save lives during pollution spikes
- **Ensures Regulatory Compliance:** Keeps companies accountable to environmental laws
- **Urban Management:** Guide traffic controls and industrial activity during high pollution
- **Climate Research:** Identify extreme weather patterns and environmental shifts

Business & Economic Importance:

- **Financial**: Billions saved annually by catching fraud early
- **Quality Control**: Reduces waste and customer complaints
- **Enables Better Decisions**: Clean data leads to accurate predictions
- **Risk Management**: Identifies problems before they escalate

Statistical Impact:

- **Ensures Data Integrity:** Prevents distorted averages and misleading results
- **Improves Model Accuracy:** Leads to better predictions and insights

Key Detection Methods

1. IQR Method (Box Plot Rule)

- Uses Interquartile Range (Q3 - Q1)
- Outliers: < Q1 - 1.5×IQR or > Q3 + 1.5×IQR
- Robust to extreme values

2. Z-Score Method

- Measures standard deviations from mean
- Typically |Z-score| > 3 indicates outlier
- Assumes normal distribution

3. Local Outlier Factor (LOF)

- Density-based method
- Compares local density to neighbors
- Good for multivariate data

Key Concepts

Outliers: Data points that differ significantly from other observations. They can indicate measurement errors, rare events, or interesting phenomena that require special attention.

IQR Method: Uses the Interquartile Range (Q3 - Q1) to identify outliers. Points outside Q1 - 1.5×IQR or Q3 + 1.5×IQR are considered outliers. This method is robust to extreme values.

Z-score Method: Measures how many standard deviations a point is from the mean. Typically, |Z-score| > 3 indicates an outlier. Useful for normally distributed data.

Boxplots: Visual tool that displays data distribution through quartiles and automatically shows outliers as individual points beyond the whiskers

Multivariate Outliers: Observations that are unusual when considering multiple variables simultaneously, not just individually. These can reveal complex patterns.

Step 1: Install & Load Packages

```
# Load required packages
library(dplyr)
library(ggplot2)

# Load and prepare air quality data
data("airquality")
air_clean <- airquality[complete.cases(airquality), ]

cat("Air Quality Dataset Loaded Successfully\n")
cat("Clean dataset dimensions:", dim(air_clean), "\n")
cat("Variables:", names(air_clean), "\n")
```

Purpose: Load the dataset and prepare it for analysis by removing missing values

Step 2: Load & Prepare Data

```
# Basic statistics and data overview
cat("--- Air Quality Dataset Summary ---\n")
```

```
--- Air Quality Dataset Summary ---
```

```
summary(air_clean[, c("Ozone", "Solar.R", "Wind", "Temp")])
```

```
     Ozone           Solar.R           Wind             Temp      
 Min.   :  1.0   Min.   :  7.0   Min.   : 2.30   Min.   :57.00  
 1st Qu.: 18.0   1st Qu.:113.5   1st Qu.: 7.40   1st Qu.:71.00  
 Median : 31.0   Median :207.0   Median : 9.70   Median :79.00  
 Mean   : 42.1   Mean   :184.8   Mean   : 9.94   Mean   :77.79  
 3rd Qu.: 62.0   3rd Qu.:255.5   3rd Qu.:11.50   3rd Qu.:84.50  
 Max.   :168.0   Max.   :334.0   Max.   :20.70   Max.   :97.00  
```

```
cat("\n--- Environmental Context ---\n")
```

```

--- Environmental Context ---
```

```
cat("Ozone Health Guidelines:\n")
```

```
Ozone Health Guidelines:
```

```
cat("- Good: < 50 ppb\n")
```

```
- Good: < 50 ppb
```

```
cat("- Moderate Risk: 50-70 ppb\n")
```

```
- Moderate Risk: 50-70 ppb
```

```
cat("- High Risk: 70-100 ppb\n")
```

```
- High Risk: 70-100 ppb
```

```
cat("- Very High Risk: > 100 ppb\n")
```

```
- Very High Risk: > 100 ppb
```

Step 3: Create Outlier Detection Function

```r
# Create comprehensive outlier detection function
detect_outliers_iqr <- function(data, column_name) {
  values <- data[[column_name]]

  Q1 <- quantile(values, 0.25, na.rm = TRUE)
  Q3 <- quantile(values, 0.75, na.rm = TRUE)
  IQR_value <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR_value
  upper_bound <- Q3 + 1.5 * IQR_value

  outliers <- which(values < lower_bound | values > upper_bound)

  return(list(
    outliers = outliers,
    bounds = c(lower_bound, upper_bound),
    values = values[outliers],
    indices = outliers,
    count = length(outliers)
  ))
}

cat("Robust outlier detection function created successfully\n")
```

```
Robust outlier detection function created successfully
```

Step 4: Detect Ozone Outliers

```
# Detect outliers in Ozone (primary air quality indicator)
ozone_outliers <- detect_outliers_iqr(air_clean, "Ozone")

cat("--- Ozone Outlier Analysis ---\n")
```

```
--- Ozone Outlier Analysis ---
```

```
cat("IQR Bounds:", round(ozone_outliers$bounds, 2), "ppb\n")
```

```
IQR Bounds: -48 128 ppb
```

```
cat("Number of high ozone days:", ozone_outliers$count, "\n")
```

```
Number of high ozone days: 2
```

```
cat("Outlier ozone levels:", sort(round(ozone_outliers$values, 2)), "ppb\n")
```

```
Outlier ozone levels: 135 168 ppb
```

```
# Health risk assessment
high_risk_days <- sum(ozone_outliers$values > 100)
moderate_risk_days <- sum(ozone_outliers$values > 70 & ozone_outliers$values <= 100)

cat("\n--- Public Health Impact ---\n")
```

```
--- Public Health Impact ---
```

```
cat("Days with high health risk (Ozone > 100 ppb):", high_risk_days, "\n")
```

```
Days with high health risk (Ozone > 100 ppb): 2
```

```
cat("Days with moderate health risk (Ozone > 70 ppb):", moderate_risk_days, "\n")
```

```
Days with moderate health risk (Ozone > 70 ppb): 0
```

Key Point: Identified days with dangerously high ozone levels that exceed health safety guidelines

Step 5: Detect Solar Radiation Outliers

```
# Detect outliers in Solar Radiation
solar_outliers <- detect_outliers_iqr(air_clean, "Solar.R")

cat("--- Solar Radiation Outlier Analysis ---\n")
```

```
--- Solar Radiation Outlier Analysis ---
```

```
cat("IQR Bounds:", round(solar_outliers$bounds, 2), "Langleys\n")
```

```
IQR Bounds: -99.5 468.5 Langleys
```

```
cat("Number of unusual solar days:", solar_outliers$count, "\n")
```

```
Number of unusual solar days: 0
```

```
cat("Outlier solar radiation values:", sort(round(solar_outliers$values, 2)), "Langleys\n")
```

```
Outlier solar radiation values:  Langleys
```

```
# Relationship with ozone
cat("\n--- Solar-Ozone Relationship ---\n")
```

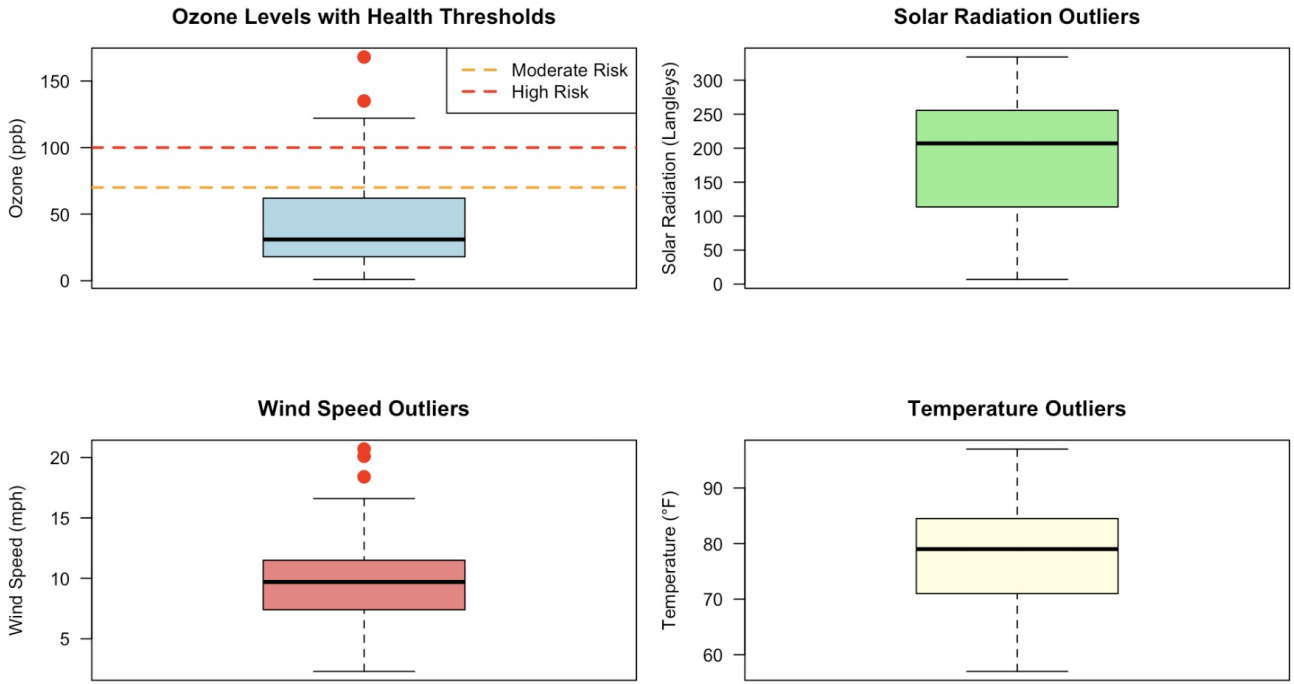
```

--- Solar-Ozone Relationship ---
```

```
high_solar_ozone <- mean(air_clean$Ozone[solar_outliers$indices])
cat("Average ozone on high solar days:", round(high_solar_ozone, 2), "ppb\n")
```

```
Average ozone on high solar days: NaN ppb
```

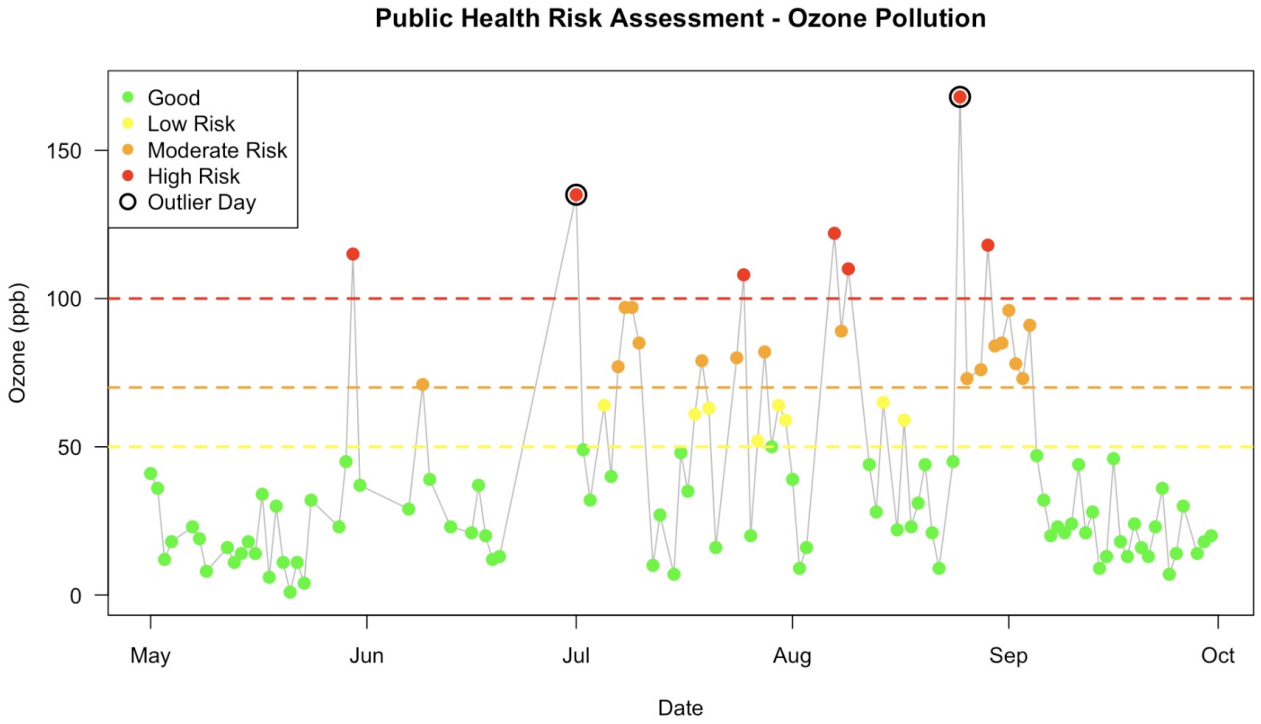
Key Point: Extreme solar radiation days often correlate with higher ozone levels due to photochemical reactions

Step 6: Time Series Analysis of Pollution Events & Public Health Risk Assessment



Observation: Boxplots clearly show environmental outliers with health risk thresholds for immediate interpretation

Step 7: Time Series Analysis of Pollution Events & Public Health Risk Assessment



Application: Real-time environmental monitoring system that automatically flags days requiring public health advisories.

Environmental Applications:

- **Public Health Monitoring:** Automated alert systems for high pollution days to protect vulnerable populations.
- **Regulatory Compliance:** Flagging industrial violations of Clean Air Act standards
- **Climate Research:** Identifying extreme weather and long-term climate shifts

Business Applications:

- **Fraud Detection:** Spotting unusual financial transactions
- **Quality Control:** Identifying defective products on production lines
- **Sales Forecasting:** Using clean data to predict future revenue

Technical & Research Applications:

- **Data Cleaning:** Removing errors before analysis
- **Model Development:** Preparing data for machine learning

Conclusion

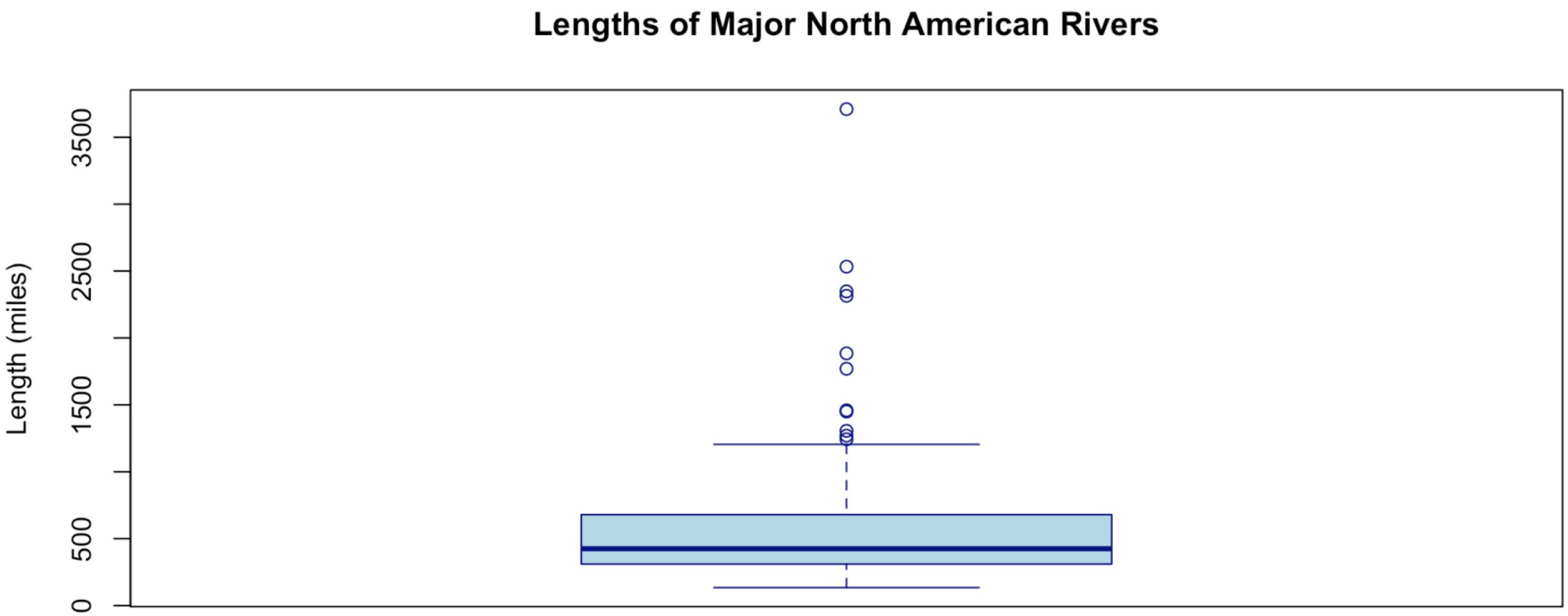
- **Outlier Detection**: is crucial for environmental monitoring and public health protection.
- **IQR Method**: provides robust identification of extreme environmental conditions.
- **Contextual Thresholds**: transform statistical outliers into actionable environmental insights.
- **Time Series Analysis**: reveals patterns in pollution events and weather extremes.
- **Automated Monitoring**: enables real-time public health alerts and regulatory responses.

🌊🐟 EXERCISE !! 🌊🌊🌊

- Look at the built-in rivers dataset in R (lengths of major North American rivers in miles).
- Create a box plot of the river lengths How many rivers appear as outliers in the box plot?

Which river(s) are the longest outliers?

```
Number of outlier rivers: 11
Longest outlier(s): 3710 miles
All outliers: 1243 1270 1306 1450 1459 1770 1885 2315 2348 2533 3710 miles
```



Thank You!

Questions?