Data Analysis

2025-04-24

R Markdown

```r
# Load required libraries
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: ’dplyr’

## The following objects are masked from ’package:stats’:
##
##     filter, lag

## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```r
library(readr)
library(tidyr)
library(RColorBrewer)
library(countrycode)
```

```
## Warning: package ’countrycode’ was built under R version 4.4.3
```

```r
# Load CSV data
data <- read_csv("osint_data.csv")
```

```
## Rows: 500 Columns: 4

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): IP, Risk Level, Country, Abuse Score
##
## 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.
```

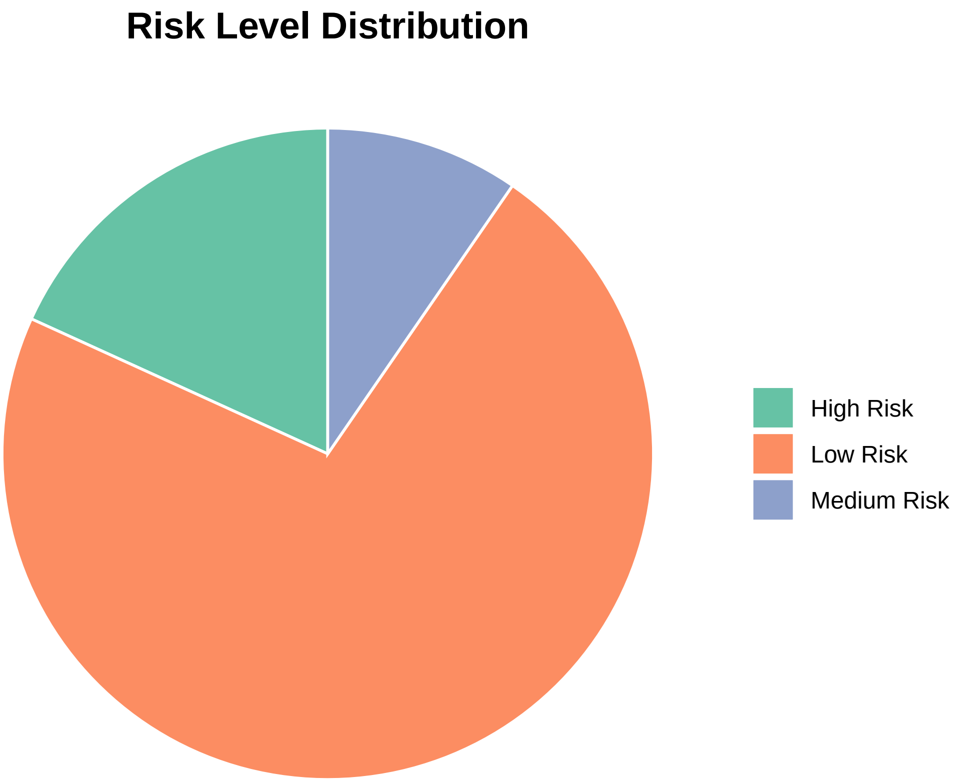
```r
# Clean Abuse Score column
data <- data %>%
  mutate(`Abuse Score` = as.numeric(gsub("Abuse Score: ", "", `Abuse Score`)))
```

```r
# Map country codes to full country names
data <- data %>%
  mutate(Country_Name = countrycode(Country, origin = 'iso2c', destination = 'country.name'))

# Define custom color palette
risk_colors <- brewer.pal(n = 3, name = "Set2")
```

1. Pie Chart of IP Distribution by Risk Level

This chart visualizes the proportion of IP addresses that fall into each risk level category (Low, Medium, High). It's useful to understand the overall distribution of perceived threats.

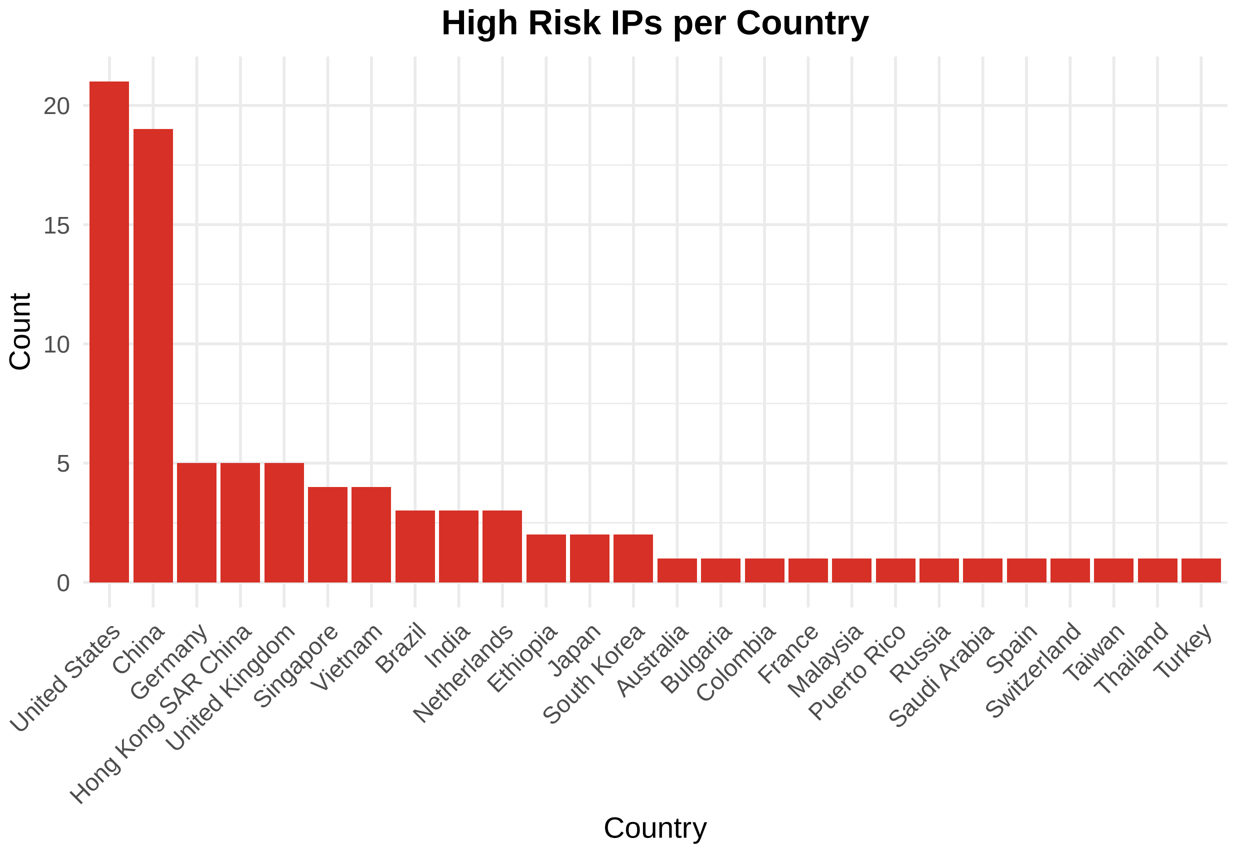
```r
# 1. Pie Chart (no inner text)
risk_counts <- data %>%
  count(`Risk Level`) %>%
  mutate(prop = n / sum(n))

ggplot(risk_counts, aes(x = "", y = prop, fill = `Risk Level`)) +
  geom_col(width = 1, color = "white") +
  coord_polar(theta = "y") +
  scale_fill_manual(values = risk_colors) +
  labs(title = "Risk Level Distribution") +
  theme_void() +
  theme(legend.title = element_blank(),
        plot.title = element_text(hjust = 0.5, face = "bold", size = 14))
```



2. Histogram of High-Risk IPs by Country

This bar chart shows which countries have the highest number of high-risk IP addresses. It helps to identify geographical concentrations of potentially malicious activity.

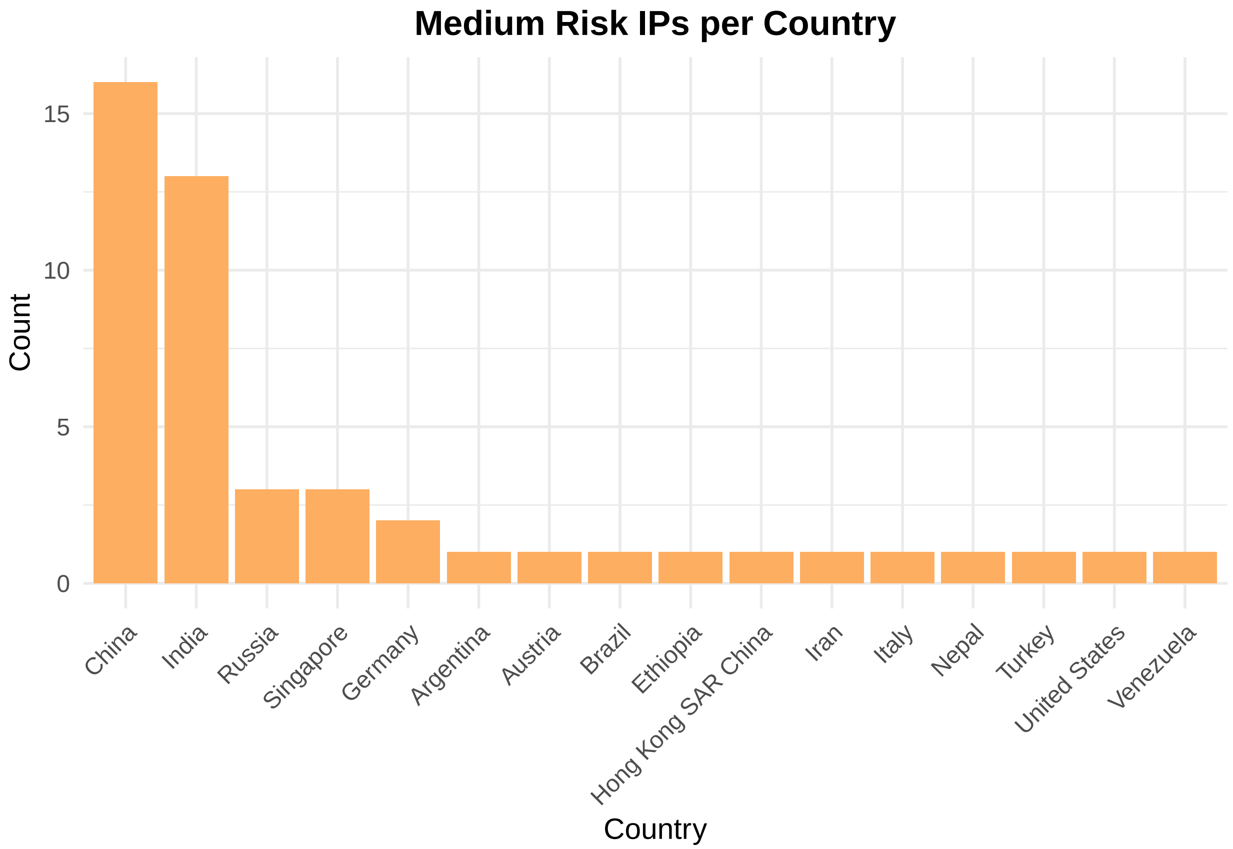
```r
# 2. Histogram: High Risk per Country (ordered, full names)
high_risk <- data %>%
  filter(`Risk Level` == "High Risk") %>%
  count(Country_Name) %>%
  arrange(desc(n))

ggplot(high_risk, aes(x = reorder(Country_Name, -n), y = n)) +
  geom_bar(stat = "identity", fill = "#D73027") +
  labs(title = "High Risk IPs per Country", x = "Country", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(face = "bold", size = 13, hjust = 0.5))
```



3. Histogram of Medium-Risk IPs by Country

This chart focuses on medium-risk IP addresses, helping to highlight different regions that may be contributing to moderate-level cyber activity or abuse patterns.

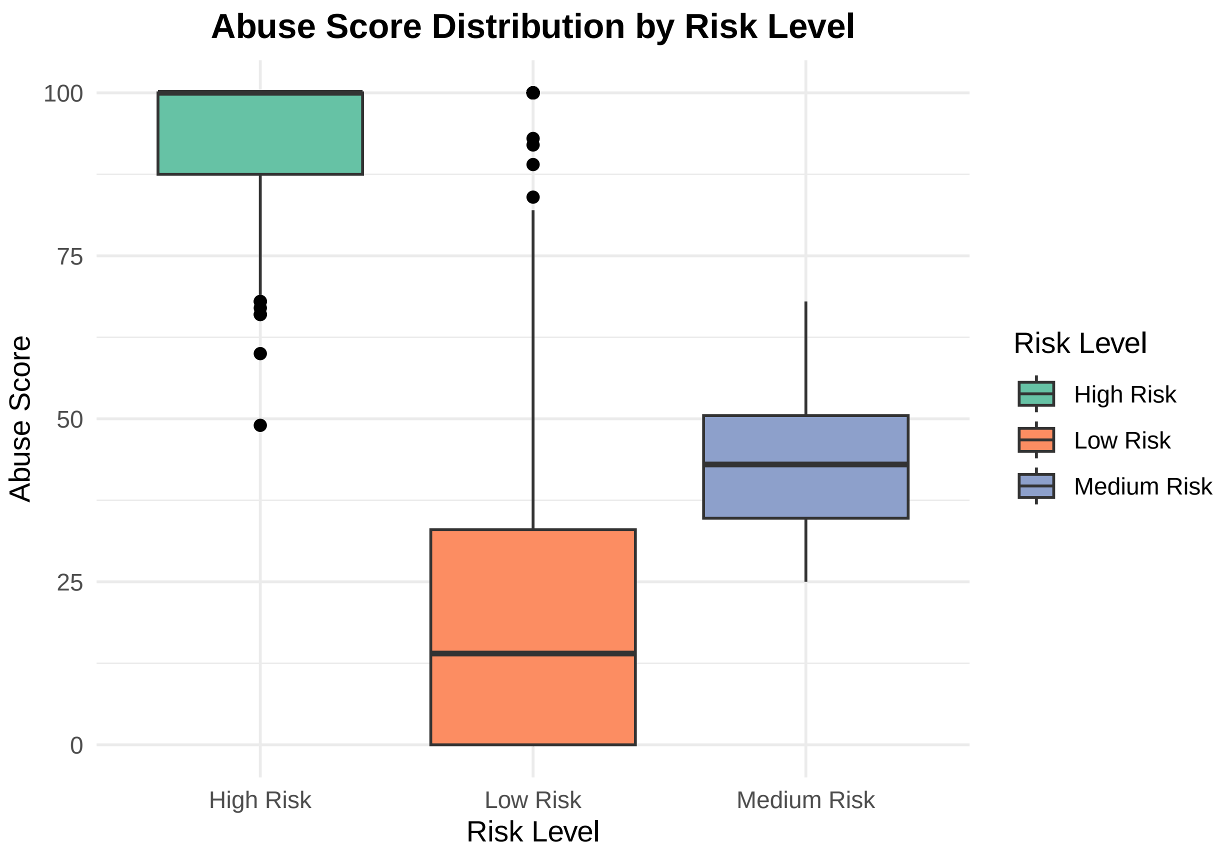
```r
# 3. Histogram: Medium Risk per Country (ordered, full names)
medium_risk <- data %>%
  filter(`Risk Level` == "Medium Risk") %>%
  count(Country_Name) %>%
  arrange(desc(n))

ggplot(medium_risk, aes(x = reorder(Country_Name, -n), y = n)) +
  geom_bar(stat = "identity", fill = "#FDAE61") +
  labs(title = "Medium Risk IPs per Country", x = "Country", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(face = "bold", size = 13, hjust = 0.5))
```



4. Boxplot of Abuse Score by Risk Level

This boxplot shows the distribution of Abuse Scores within each risk category. It helps assess whether risk labels align with quantitative indicators of threat level and identify outliers or misclassifications.

```
# 4. Boxplot: Abuse Score by Risk Level
ggplot(data, aes(x = `Risk Level`, y = `Abuse Score`, fill = `Risk Level`)) +
  geom_boxplot(outlier.color = "black", outlier.shape = 16, outlier.size = 2) +
  scale_fill_manual(values = risk_colors) +
  labs(title = "Abuse Score Distribution by Risk Level", x = "Risk Level", y = "Abuse Score") +
  theme_minimal() +
  theme(plot.title = element_text(face = "bold", size = 13, hjust = 0.5))
```


Abuse Score Distribution by Risk Level
Abuse Score
100
75
50
25
0
High Risk
Low Risk
Medium Risk
Risk Level
Risk Level
High Risk
Low Risk
Medium Risk