University of Texas at Arlington

Data Science Visualization & Exploratory Data Analysis

Group 12

Anil Thapa - 1002190895

Jair Rea - 1001887311

Yash Prakash Joshi - 1002207784

Rama Jyothi Swaroop Janapareddy – 1002199950

Data Visualization 5305-001

Roza Zaruba

September 27, 2024

Group 12

Data Visualization

September 24, 2024

HomeWork 1 – Exploratory Data Analysis

Introduction

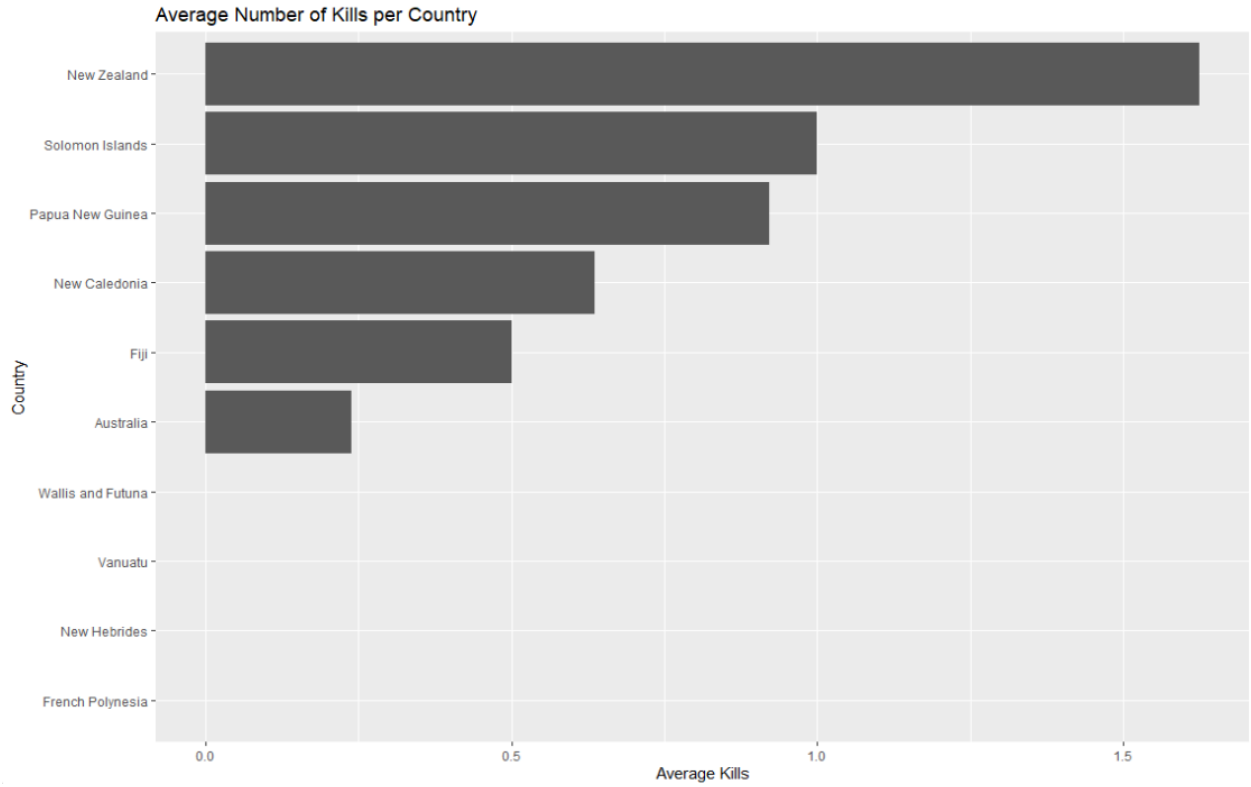
: The dataset our team received was a list of terrorist attacks and the details of what, when, where, and how it happened. The details also include what type the attack consisted of and the main target to be attacked. Our objective in analyzing this attack is to find any type of correlation, check out distributions and highlight key statistics that may provide some valuable insight. Our team decided to focus on a specific part of the dataset that centralized on attacks and whether specific types or methods were more successful and which weapon was more commonly used within the attack. By creating a more focused objective, the team preprocessed the dataset. Doing so, the dataset has considerably decreased in size in terms of the number of variables. However, we want to preface that this does not mean the data has become unreliable and shallow. We chose features that closely corresponded to what we were searching for, making the dataset more centralized and more understandable. Cutting redundant or features that held no use were first to be cut.

Tool Comparison

: The team used two different types of tools, Python and R. In terms of usability, we found that R was more useful and helpful when trying to process the data. In other words, selecting the variables we wanted and outlining key statistics was simpler in R compared to Python. A key factor in why we believe in that is because R is more geared towards data science and overall statistical analysis while Python is an amalgamation of libraries that are geared towards other topics and studies. That is not to say that it is not useful in any way. Although R is geared more towards analysis and data science, we found that the plots created by Python were more visually understandable. There are ways to form and change the graphs using libraries that are solely in Python. To summarize, we found that in terms of usability, R was the prime winner but in terms of visualization quality, Python was the prime winner. Each software/language is a valuable tool and that is one thing the team advocates and does not deny.

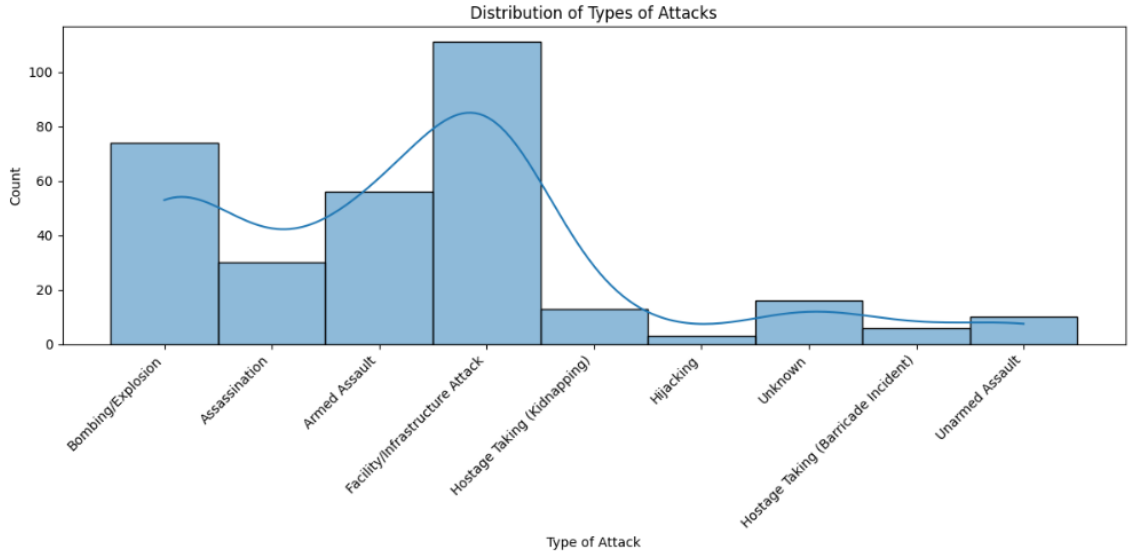
Findings (Visualizations will be specified whether it was made in R or Python)

: 1st Plot (R)



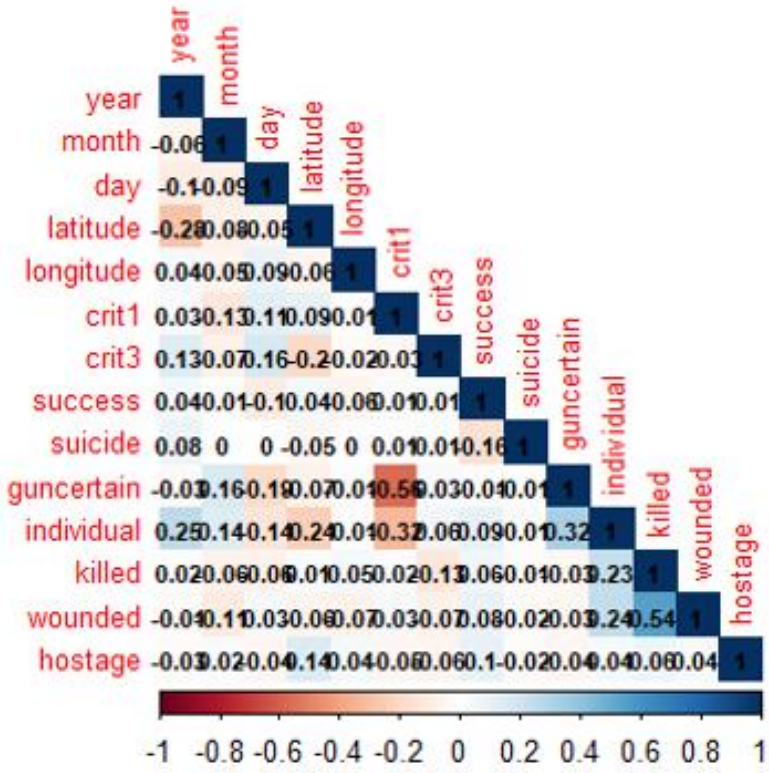
Using a bar plot, the visualization shows/summarizes the average kills per country. It is sorted by average kills. Understanding this, the plot shows that New Zealand is the top country in terms of people being killed within the terrorist attacks. (This does not count wounded.)

: 2nd Plot (Python)



The plot above shows the distribution of the type of terrorist attacks. By doing this, we can show that terrorist opt to use Facility/Infrastructure and bombing attacks more than any other type. The reason being as to why they might choose one over the other is possibly due to risk and what may cause the most amount of damage to a country. Overall, it is clear that the most damaging attack is what terrorists clearly opt for when it comes to an attack.

: 3rd Plot (R)



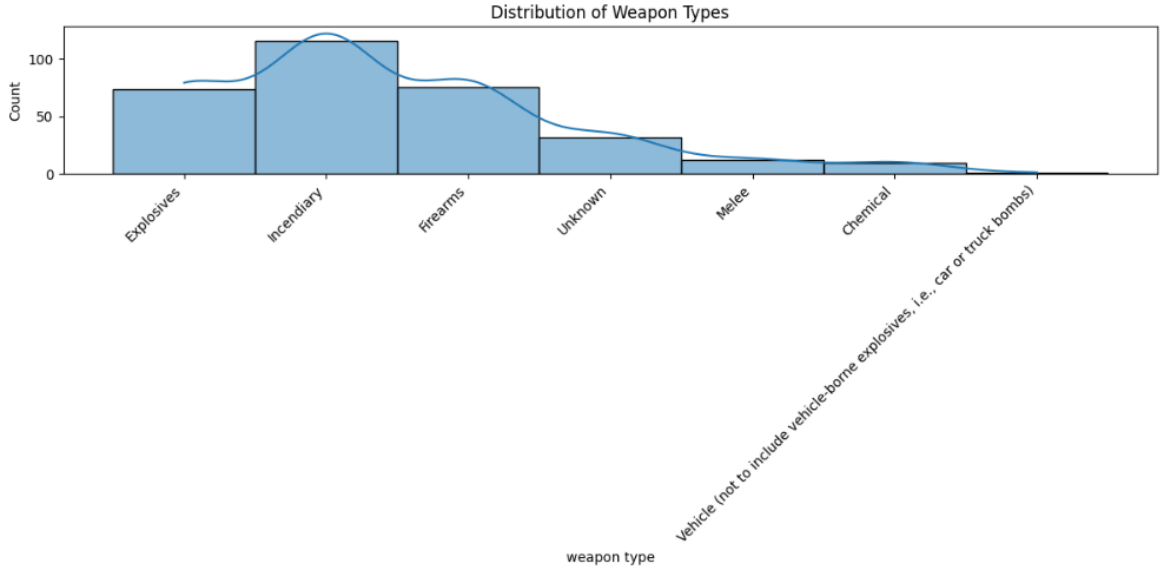
The plot above represents a detailed correlation map in which type of variable is more correlated with another variable. Going more into depth, it is shown that most of the variables show no type of correlation to one another. This is possibly due to the variables we selected but overall, when it comes to attacks and deaths, there seems to be no type of trend in between with one exception.

```
# Correlation between killed and wounded
correlation_value <- cor(data_cleaned$killed, data_cleaned$wounded, use =
"complete.obs")
print(paste("Correlation between Number of Kills and Number of Wounded: ",
correlation_value))

## [1] "Correlation between Number of Kills and Number of Wounded:
0.540566046614495"
```

The correlation coefficient of 0.54 suggests a moderate positive correlation between the two variables, meaning that as the number of kills increases, the number of wounded also tends to

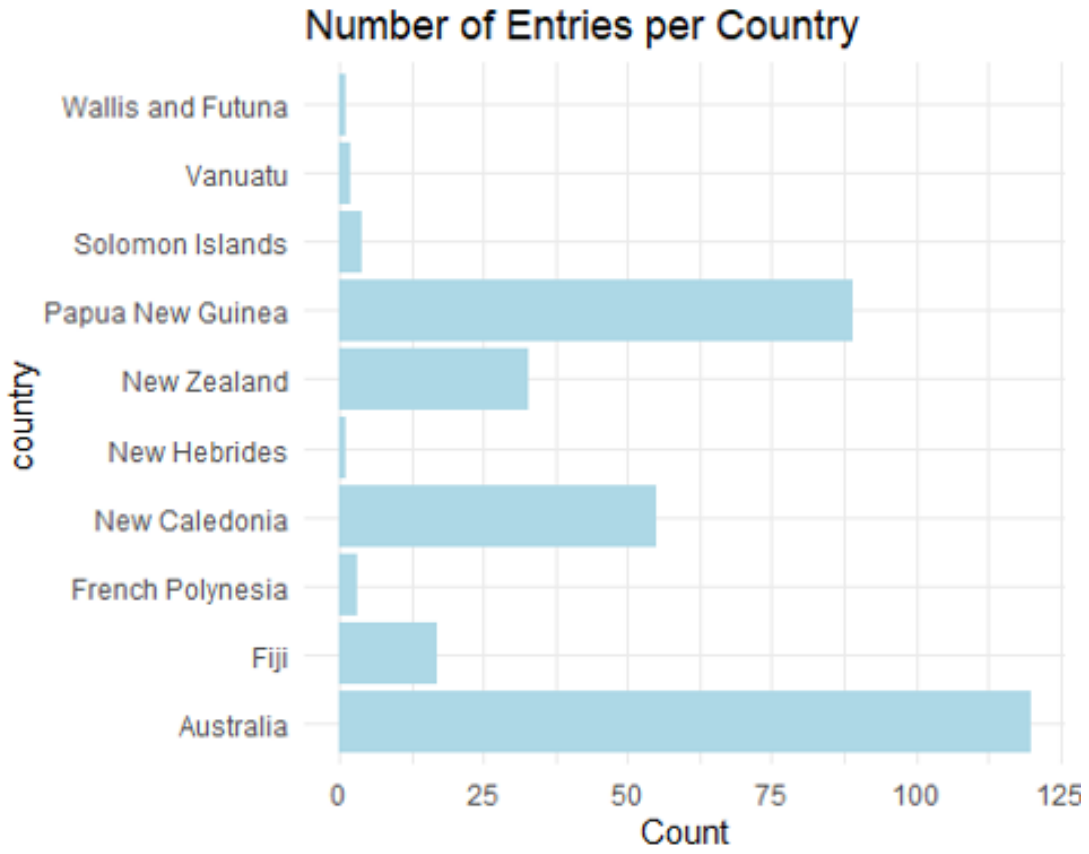
increase, though the relationship is not very strong. Overall, while higher numbers of kills generally correspond with higher numbers of wounded, the majority of incidents involve relatively low casualties.

: 4th Plot (Python)



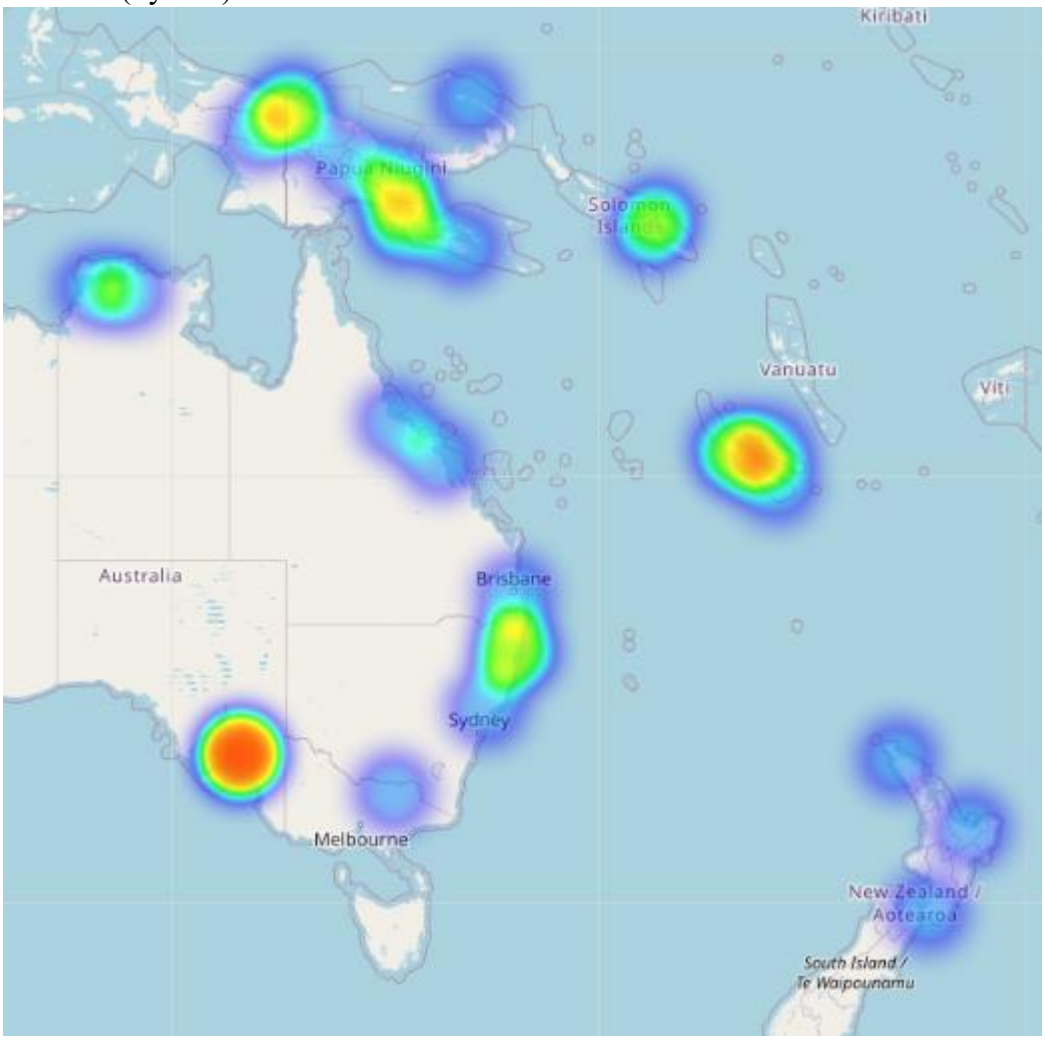
In the plot above, the findings that we found was that “Incendiary” weapons were the likely choice when it came to the type of weapon used within terrorist attacks.

: 5th Plot (R)



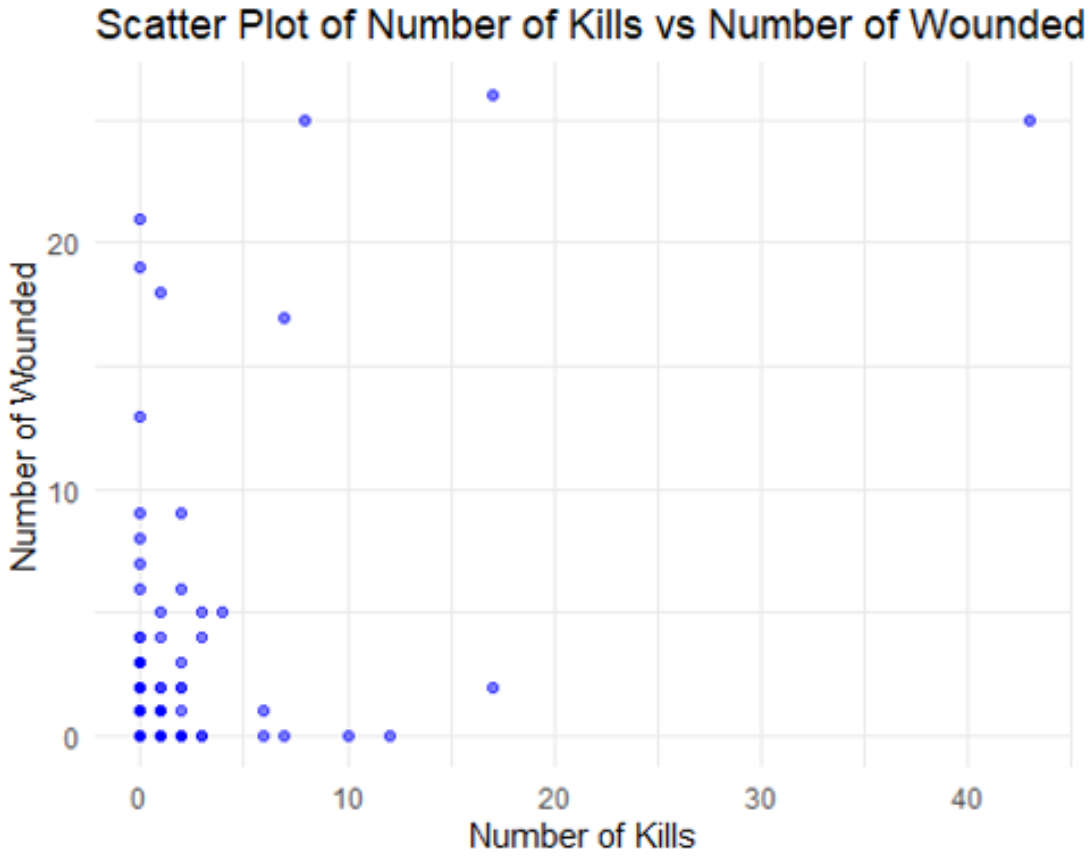
The bar plot shows a clear imbalance in the number of entries per country, with Australia and Papua New Guinea dominating the dataset. Australia has the highest representation, with around 125 entries, followed by Papua New Guinea with about 100. New Caledonia and New Zealand also have a moderate presence, but countries like Wallis and Futuna, Vanuatu, New Hebrides, and French Polynesia are minimally represented, with very few entries. Fiji and Solomon Islands fall in between, with slightly more occurrences than the least represented countries. The distribution is highly skewed, indicating that the dataset focuses primarily on a few countries, particularly Australia and Papua New Guinea, while providing significantly less data for other regions. This concentration suggests that any analysis of trends or patterns in the dataset may heavily reflect the conditions of these dominant countries, with less reliable insights for the underrepresented ones.

: 6th Plot (Python)



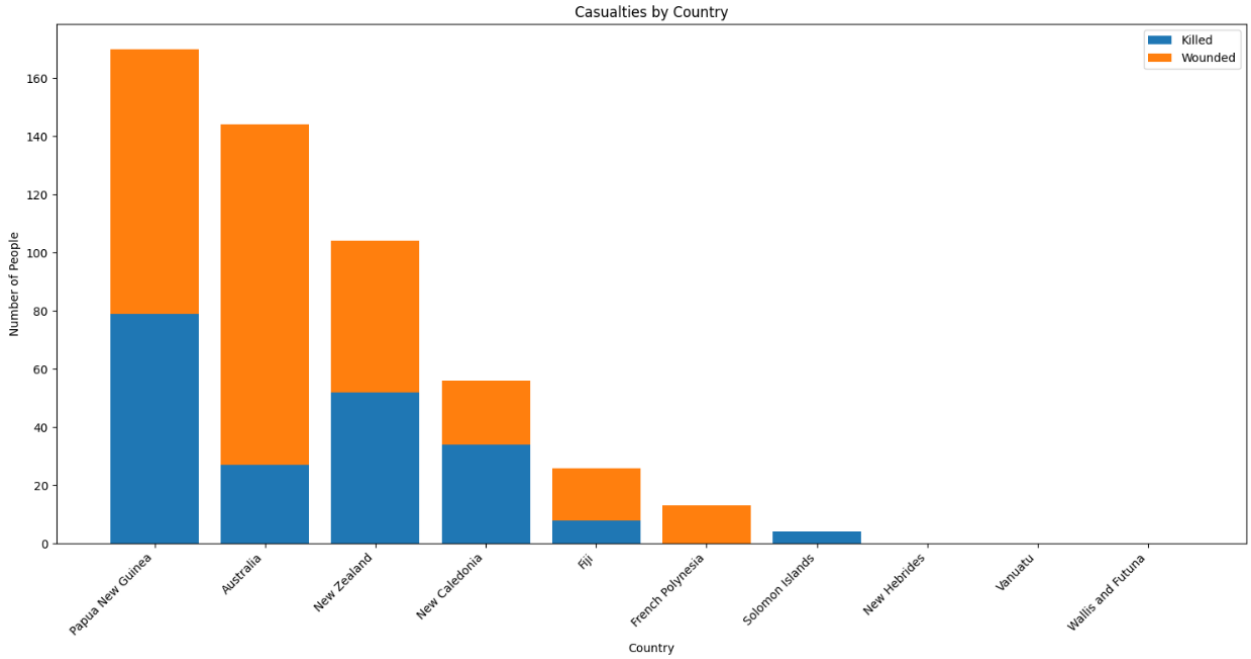
The plot above is a heatmap that highlights the areas where casualties are killed within the country. This is based on longitude and latitude. Just by looking at the plot, it is obvious to notice that there are four main key areas that are prominent. This lines up with a previous plot in which it showed the top countries where there were casualties, but this plot shows a map that could illustrate to viewers what main areas to avoid. Another to thing to highlight is that, although New Zealand was on top, you can see that the area in which is affected is spread out, not in one sole area. This is important to understand as attacks do not happen in a centralized spot but scattered.

: 7th Plot (R)



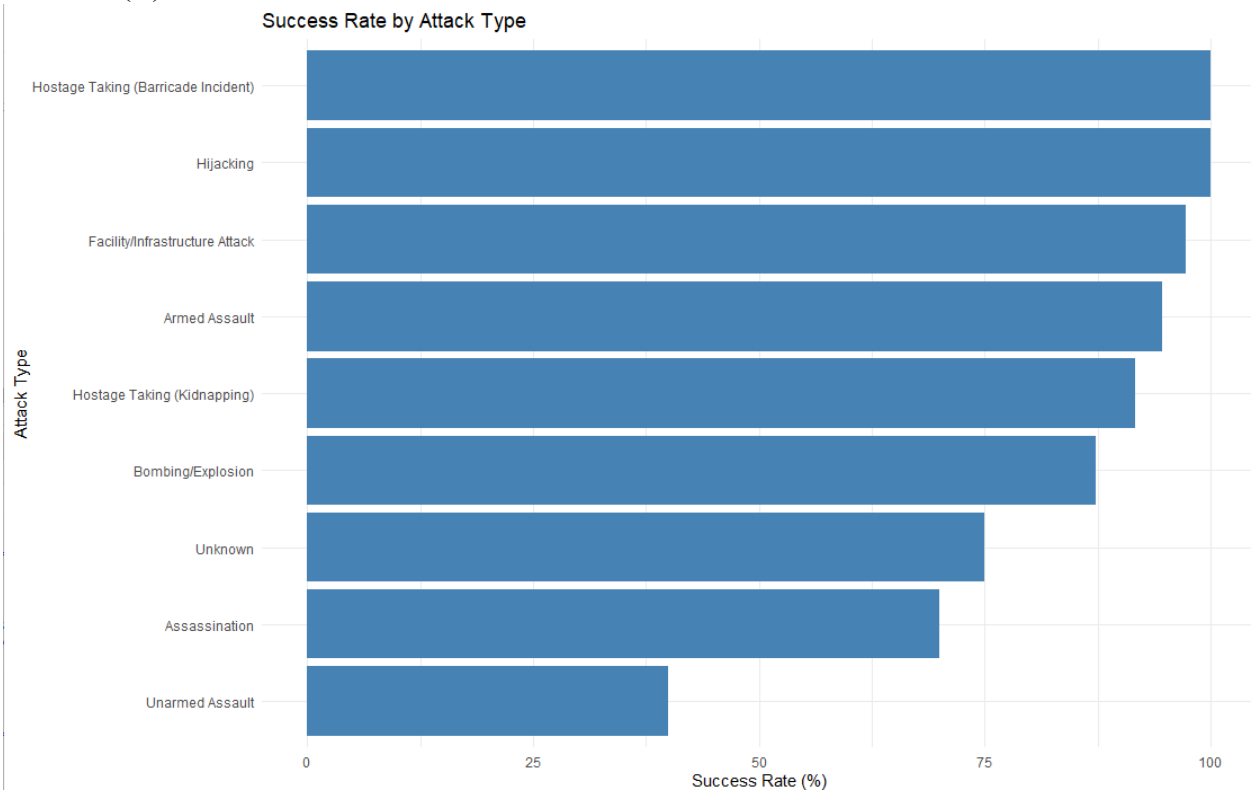
Looking at the scatter plot above it is clear that most of the points are concentrated near zero. This indicates that most of the terrorist attacks had few casualties. There are some outliers shown but the lack of points show that attacks with high amount of casualties are rare. Overall there is a trend that suggests that incidents with more people killed often also involve more wounded, but there is not a perfectly linear relationship.

: 8th Plot (Python)



In the stacked bar plot above, we can see that when we involve the wounded and a different type of dataset where we also include missing values, Papua New Guinea has the highest number of casualties, with over 160 people affected. It is worthy to note that the plot is split with Blue representing those who were killed while orange represents those were wounded. Some key insight that may be worthy to note is the plot suggests that most casualties in countries like Australia and New Zealand are from wounded individuals, while in Papua New Guinea, both killed and wounded numbers are relatively close. This could indicate differences in attacks or the response capabilities of these countries. Another thing that is worthy to mention is in most countries, the number of wounded is greater than those killed. This could suggest that attacks are either becoming less lethal or that medical services are capable of saving lives in high-casualty situations, which is a great thing.

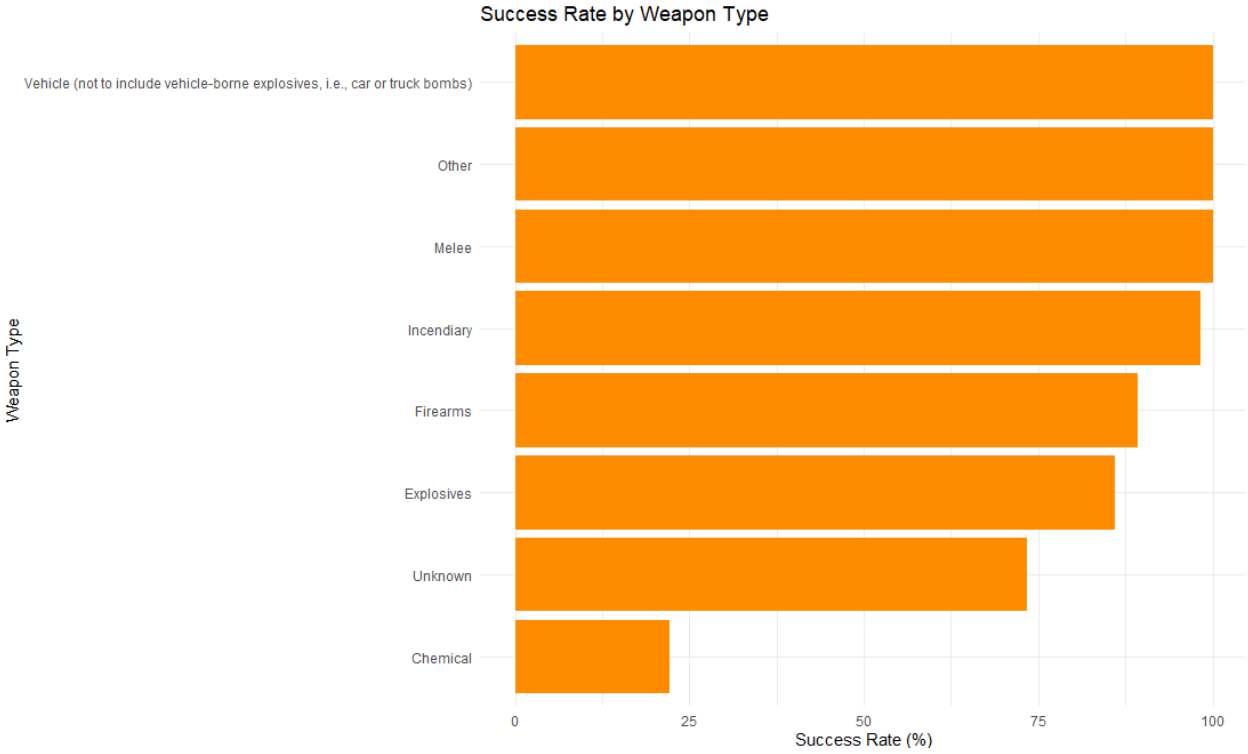
: 9th Plot (R)



This bar plot expresses the success rate of the attack types. The team wanted to see which type of attack was most likely to work. The plot shows that the most successful types of attacks were Hostage taking (specifically barricades) and Hijacking, both reaching a near 100% success rate. I want to also note that there is a specific type that has not been identified which has a good chance of succeeding (Unknown variable.) Overall this plot really highlights that that coordinated and well-planned attacks, especially those involving hostages, tend to be more successful.

```
> summary(attack_type_success)
 attacktype1_txt    total_attacks    successful_attacks  success_rate
 Length:9           Min.   :  3.00   Min.   :  3        Min.   : 40.00
 Class :character   1st Qu.: 10.00   1st Qu.:  6        1st Qu.: 75.00
 Mode  :character   Median : 16.00   Median : 12        Median : 91.67
                    Mean   : 34.89   Mean   : 31        Mean   : 83.99
                    3rd Qu.: 56.00   3rd Qu.: 53        3rd Qu.: 97.27
                    Max.   :110.00   Max.   :107        Max.   :100.00
```

:10th Plot (R)



This plot is the last plot with an explanation as it also highlights the success rate my team is really interested in. This looks really similar to the distribution of weapon types but it looks like other forms of attacks and vehicles lead the cause when it comes to success. It is worthy to note that guns and forms of explosives are always going to be a form of success as damaging property and causing chaos is a form of success when it comes to terrorist attacks.

```
> summary(weapon_type_success)
 weaptype1_txt      total_attacks    successful_attacks  success_rate
 Length:8           Min.   :  1.00   Min.   :  1.00     Min.   : 22.22
 Class :character   1st Qu.:  7.00   1st Qu.:  1.75     1st Qu.: 82.77
 Mode  :character   Median : 21.00   Median : 17.00     Median : 93.80
                    Mean   : 39.25   Mean   : 34.88     Mean   : 83.63
                    3rd Qu.: 72.00   3rd Qu.: 62.50     3rd Qu.:100.00
                    Max.   :115.00   Max.   :113.00     Max.   :100.00
```

** Detailed Statistics on the dataset

```
# Descriptive statistics
summary(data_cleaned)

##       year          month             day          country
##  Min.   :1970   Min.   : 1.000   Min.   : 0.00   Length:325
##  1st Qu.:1989   1st Qu.: 4.000   1st Qu.: 8.00   Class :character
##  Median :1995   Median : 7.000   Median :15.00   Mode  :character
##  Mean   :1998   Mean   : 7.003   Mean   :15.11
##  3rd Qu.:2015   3rd Qu.:10.000   3rd Qu.:23.00
##  Max.   :2020   Max.   :12.000   Max.   :31.00
##
##      city              latitude         longitude          crit1
##  Length:325         Min.   :-43.533   Min.   :-176.2   Min.   :0.0000
##  Class :character   1st Qu.:-34.289   1st Qu.: 149.1   1st Qu.:1.0000
##  Mode  :character   Median :-22.276   Median : 155.4   Median :1.0000
##                     Mean   :-23.141   Mean   : 152.3   Mean   :0.9754
##                     3rd Qu.: -7.539   3rd Qu.: 165.8   3rd Qu.:1.0000
##                     Max.   : -2.578   Max.   : 179.4   Max.   :1.0000
##                     NA's   :6         NA's   :6
##      crit2       crit3           success          suicide
##  Min.   :1   Min.   :0.0000   Min.   :0.0000   Min.   :0.000000
##  1st Qu.:1   1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:0.000000
##  Median :1   Median :1.0000   Median :1.0000   Median :0.000000
##  Mean   :1   Mean   :0.9538   Mean   :0.8738   Mean   :0.003077
##  3rd Qu.:1   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.000000
##  Max.   :1   Max.   :1.0000   Max.   :1.0000   Max.   :1.000000
##
##  Type.of.Attack        target          Group.home.country  Group.name
##  Length:325         Length:325         Length:325         Length:325
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    guncertain        individual      weapon.type            killed
##  Min.   :0.00000   Min.   :0.00000   Length:325         Min.   : 0.0000
##  1st Qu.:0.00000   1st Qu.:0.00000   Class :character   1st Qu.: 0.0000
##  Median :0.00000   Median :0.00000   Mode  :character   Median : 0.0000
##  Mean   :0.02154   Mean   :0.05846                      Mean   : 0.6426
##  3rd Qu.:0.00000   3rd Qu.:0.00000                      3rd Qu.: 0.0000
##  Max.   :1.00000   Max.   :1.00000                      Max.   :43.0000
##                                                         NA's   :6
##     wounded           hostage          dbsource
##  Min.   : 0.0000   Min.   :0.00000   Length:325
##  1st Qu.: 0.0000   1st Qu.:0.00000   Class :character
##  Median : 0.0000   Median :0.00000   Mode  :character
##  Mean   : 0.9905   Mean   :0.06769
##  3rd Qu.: 0.0000   3rd Qu.:0.00000
```

:

ANOVA Test

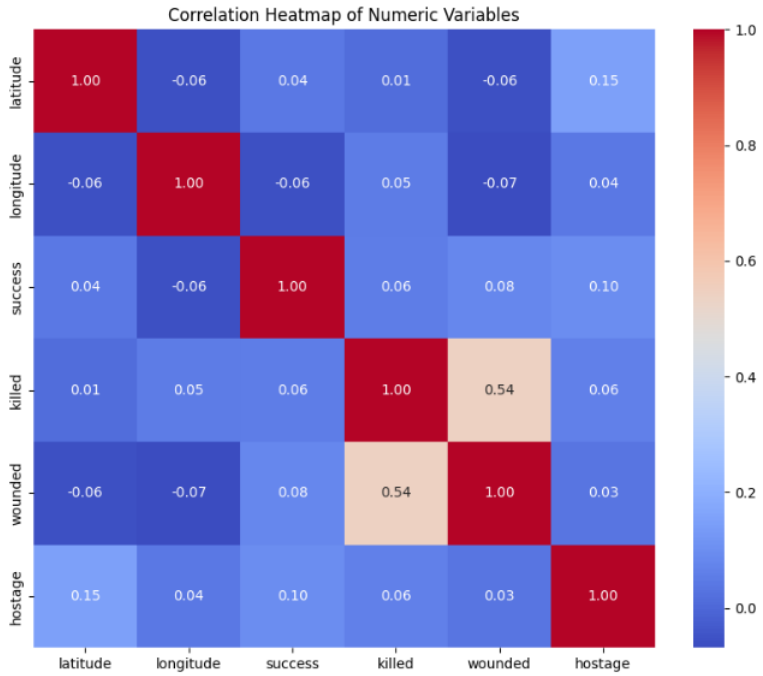
```
# ANOVA to check the relationship between kills and attack type
anova_result <- aov(killed ~ `Type.of.Attack`, data = data_cleaned)
summary(anova_result)

##                 Df Sum Sq Mean Sq F value   Pr(>F)
## Type.of.Attack   8    253   31.63   3.689 0.000395 ***
## Residuals      310   2658    8.57
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 6 observations deleted due to missingness
```

:

F value = 3.689, Pr(>F) = 0.000395: This suggests that the variation between the groups (attack types) is statistically significant, with a p-value well below 0.05, indicating a significant effect of the attack type on the number of kills.

Type.of.Attack has a significant effect on the number of kills, as indicated by the small p-value (0.000395), suggesting that the type of attack has a measurable impact on the outcome (number of kills). The ANOVA results further show that the type of attack significantly affects the number of kills, reinforcing the idea that attack methods have a substantial impact on the lethality of incidents

:Other visualizations that we did some type of analysis on



Conclusion

: To conclude I want to highlight key aspects within the EDA my team has done. We found what types of attacks were most successful and which areas were impacted the most. It is very important to understand that these types of attacks, when analyzed, can be used to understand and more importantly defend against. By highlighting the success rates and the common uses we can ensure that in the future fewer and less attacks will be successful. I want to add that there are many strengths each tool has and as I have already stated, we find that R is best suited for analysis as the libraries it has are very useful, especially for statistics and sadly the weakness has to be the learning curve and the visualizations as when it has lots of data, it gets muddy. As for python the strength comes from its versatility, but that may also come at the expense of focus. It has many wonderful libraries but most are useless when it comes to trying to explain the stats however its visualizations are wonderful when given the proper commands.