**Global Terrorism Data**

**Introduction**

Terrorism is an act that affects the world. It’s global and detrimental threat and effect on people around the world and the sovereignty of many countries was magnified as a result of 9/11. Since then, The U.S has developed a Department of HomeLand Security, who aim is specifically to defend and protect the U.S from terrorist attacks locally and internationally. There is a whole debate on whether the U.S uses legal methods to protect and defend itself, however the point is that terrorism as a global threat is now more prevalent than ever, mainly due to political and ideological differences between the groups and the countries they attack.

This project then uses the “Global Terrorism Database” a database that contains most terrorist acts perpetrated since 1970. Using this database I would like to find common attributes of terrorist acts among regions, and also the likelihood of a specific attack in a region.

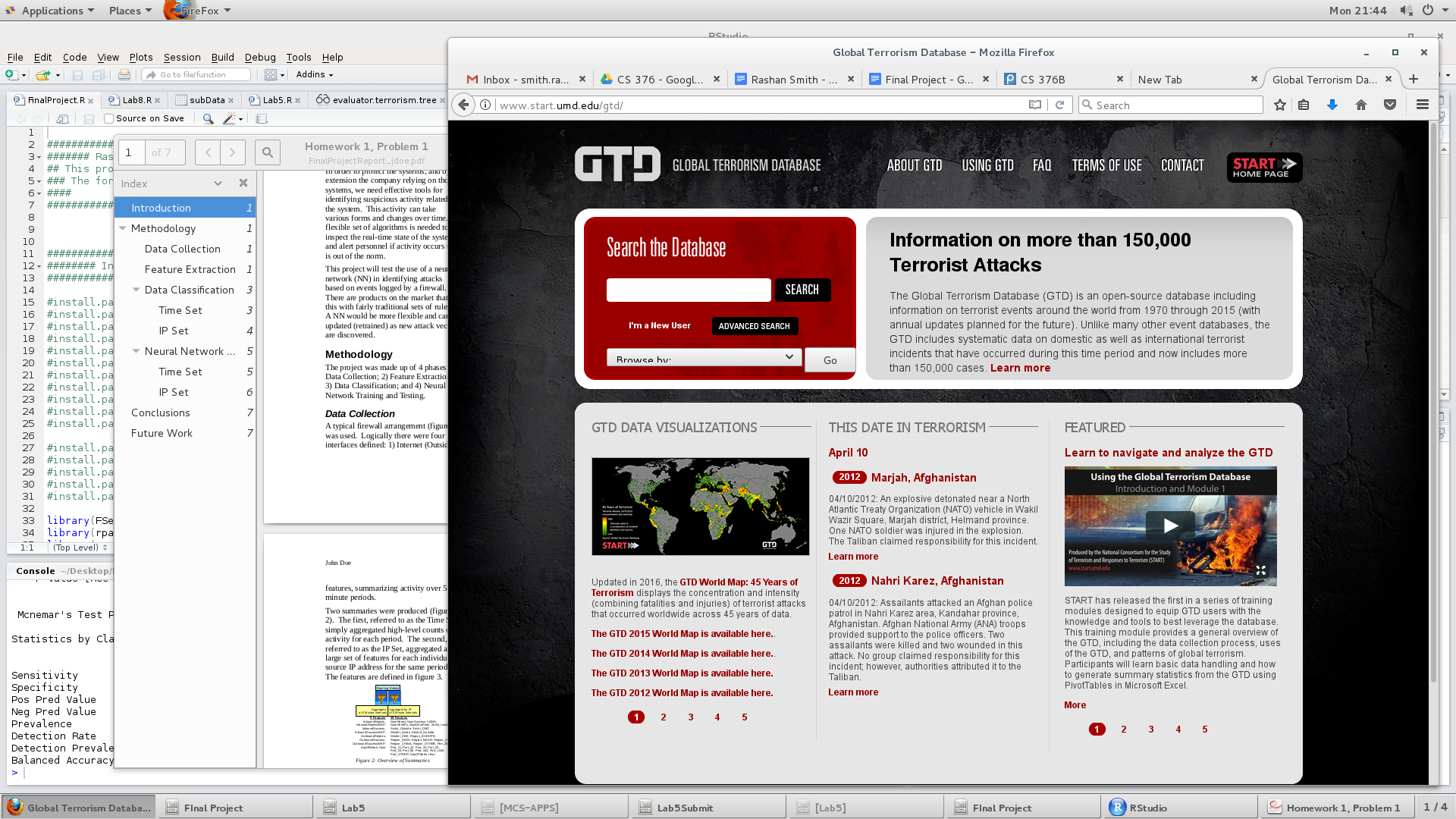
My interest in this topic stems from an International Affairs course I am currently taking called “Terrorism & International Law”. In addition, as someone from The Bahamas, a small island state, I would like to know what threat and possibility of terrorist act we would be most likely to have if we did , using data.

**Methodology**

The project was made up of 5 major phases comprised from the CRISP-DM process: Business Understanding, Data Understanding, Data Preparation, Modeling, and Evaluation.

**Data**

The data used for my project was retrieved from the Global Terrorism Database website, a project under the “National Consortium for the Study of Terrorism and Responses to Terrorism” at the University of Maryland.

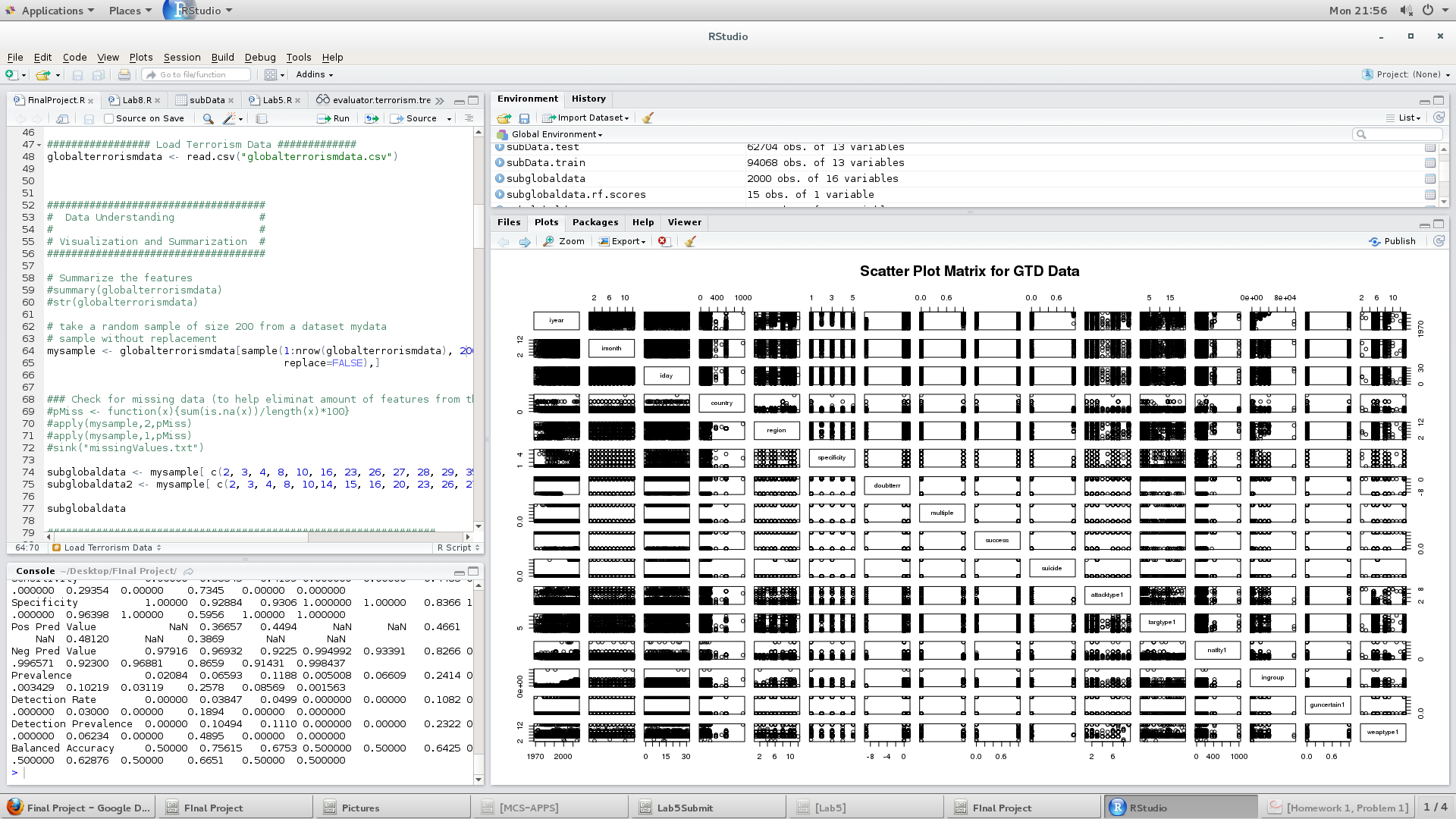


**Part 1: Business Understanding**

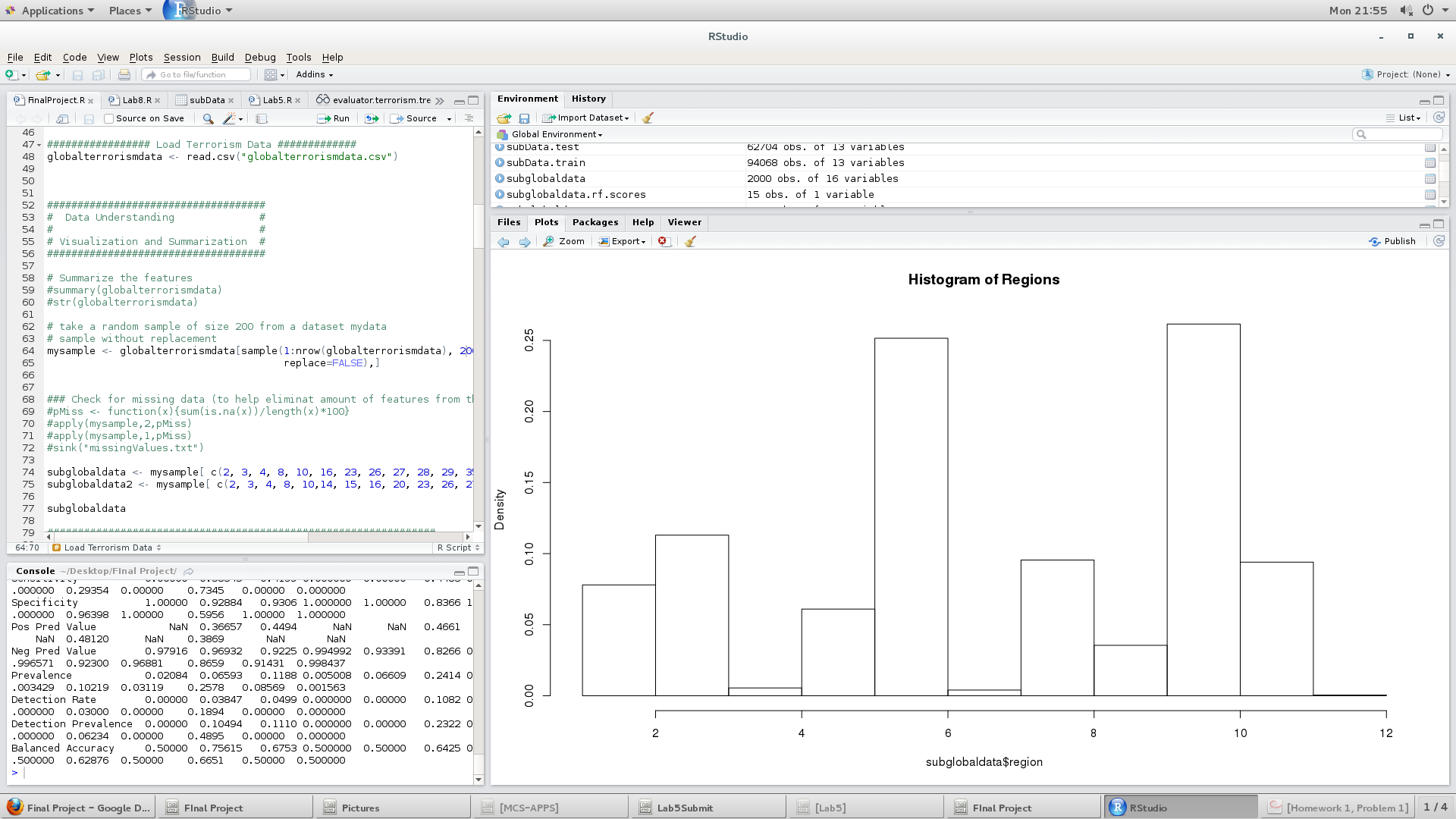
The end product of this project is to be able to predict the type of terrorist attack a region is likely to face.

**Part 2: Data Understanding**

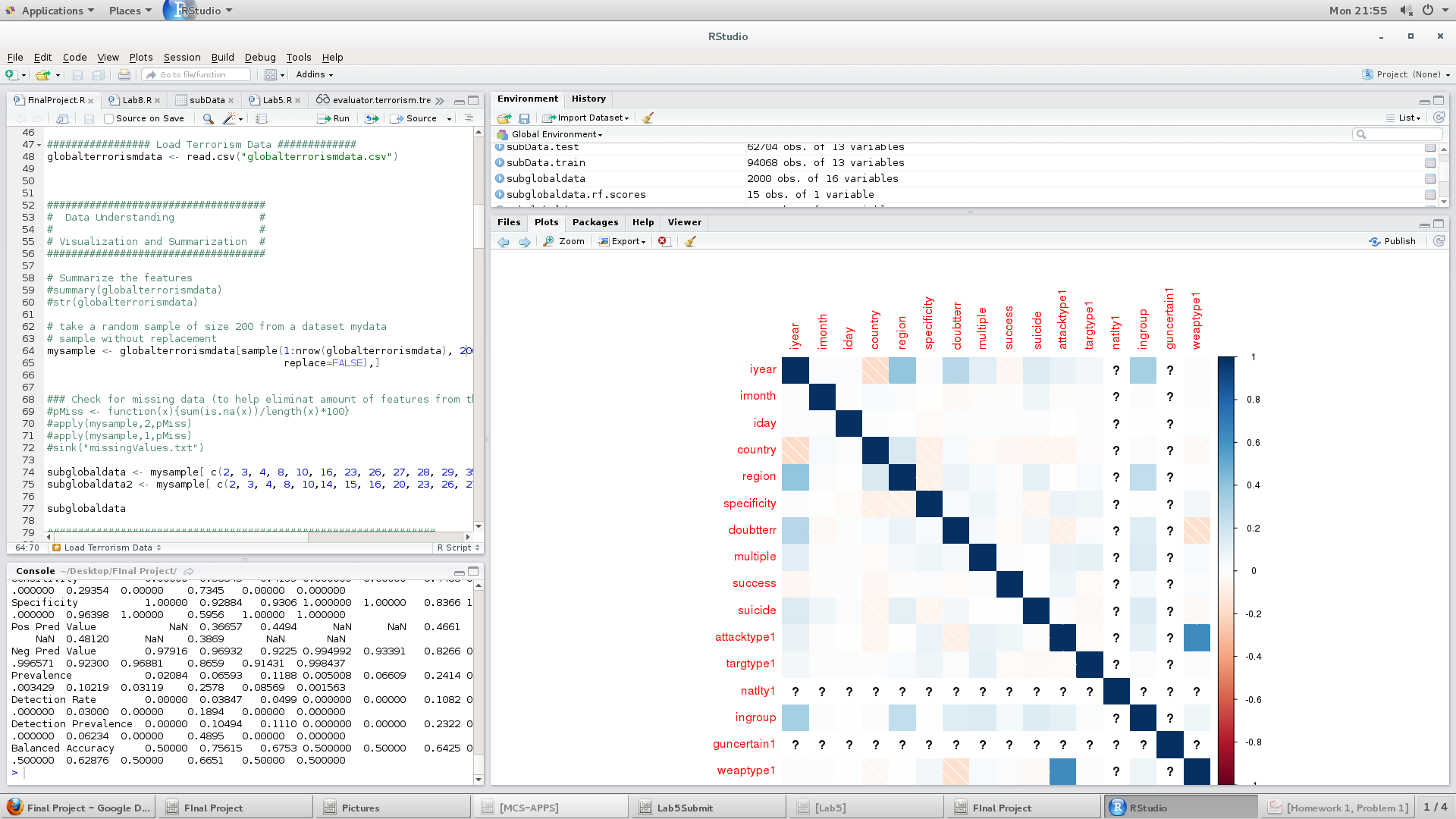
This GTD dataset has over 150,000 rows of data, therefore a random subset of 2000 was used to gain some basic visualization of it.



*Scatter Plot of Data Set*



*Bar chart of Data Set feature: regions*



*Correlation Chart*

By understanding the data visually it can be noticed that there are too many features, and feature reduction must take place.

**Part 3: Data Preparation**

This step contains a lot of work involving deciding which features to remove, and also changes the types of certain features so that they are able to be used in certain calculations. The main methods used to determine features for the final ABT were Entropy, Random Forest, and Correlation.

Firstly, the majority of features were removed by eliminating features with a missing value percentage greater than 5%. This left the dataset with 15 - 20 features. However, some features that were included actually had no values at all, or were in a format that could be used, so that narrowed the feature set down to 13.

Later, features were changed into factors because of their nature. When looking at the importance of features in relation to “region” (target feature) using random forest importance, it is interesting how a feature “ingroup” seems to be an important attribute, however its true meaning is unknown for this dataset at the moment.

**Part 4: Modeling**

The data was split into 2 sections: train and test.

The models that I used for my project were Decision Tree, Random Forest, SVM, Rule Set and Naive Bayes. For all models I had to make sure that the features were all in numeric factorial form. I analyzed the results by creating a confusion matrix for each. The results of the training data did not produce any great results, and the same happened for the test data as well.

**Part 5: Evaluation**

Below are the models and their levels of accuracy:

|  |  |
| --- | --- |
| Decision Tree | 0.403 |
| Random Forest | 0.358 |
| SVM | 0.357 |
| Naive Bayes | 0.355 |
| Rule Set | 0.295 |

Overall these models did not perform well on the data.

**Part 5: Deployment/Conclusion**

Overall, I do not think this model is fit for deployment. If it were to be deployed, it would be in a website/dashboard form that would allow one to see the biggest attack and target threat that a region would be most vulnerable to at that time.

I think a major drawback during this process was that the data was in factor form, which made it hard to use other models such as Linear Regression & K-Nearest Neighbor that may have produced more accurate results. The data also had a lot of missing values, and this could have set back the quality of the predictions as well. Another aspect of this project that could be more productive is subsetting the data to focus on a particular region, thus conducting in depth analysis on the different cities and their vulnerabilities within the data set. I think doing this would provide more fruitful results because the data would not be as spread out. However, overall I value the opportunity to conduct this research, because it gave me insight into the data side of a global phenomenon.

The biggest question, however, is if this model was perfect, would it be a great indicator for a region on how to prepare themselves against a terrorist attack? The answer is no. Terrorism is not a black and white concept. These acts, although small in number somehow have the biggest impact on our perception of safety and political stability. Terrorism itself does not have a universal definition, and likewise a terrorist cannot be characterized. In recent years the US has seen many “terrorist” like attacks by people who had appeared to be normal. Although I would argue that you can’t really ever predict the exact place and time a terrorist act might occur, a machine learning tool could use trends to help countries safe guard themselves and boost their surveillance in areas where it has in the past.