Data Science Predictive Application for Country Instability (DSAPCI)

An essay submitted in partial fulfillment of

the requirements for graduation from the

Honors College at the College of Charleston

with a Bachelor of Science in

Applied Mathematics

Courtney Beckham

May 2019

Advisor: Lancie Affonso

Contents

1 Abstract

This project explores the follow-up model of the Analyzing Complex Threats and Operations Readiness (ACTOR) model employed by the United States government. The ACTOR model aggregates data on 12 factors from 159 countries over the period 1975-1999 to predict the level of instability a country will experience. The new modeling system is known as the Integrated Crisis Early Warning System (ICEWS) and incorporates the same background as the previous model, but with a few updates and a public database. This project used that database to explore the importance of data visualization in predictive modeling. Building off of data from the ICEWS database, an interactive website was created to fill the gap between non-technical and technical people alike. By allowing the data visualizations to be customized, users are allowed to explore the data and build a report out of their findings.

2 Introduction

2.1 The Need for Data Visualizations

Storytelling is one of the oldest crafts in human history. In recent years, the concept of storytelling has been extended to data and how to convey truths from a data set to the non-technical user. Some scholars assert that data is the very "antithesis of stories. If stories have people in all their richness, data have points, rich meanings reduced to numbers." [1] When presented with a dataset alone, gleaning wisdom and truth from a spreadsheet proves to be very difficult. This is where the need for data visualization comes into play. Data visuals breathe life into numbers and give color and meaning to an otherwise boring spreadsheet. However, before data visualization can derive a sense of importance in the eyes of the user, the importance of the context and need for understanding of the data must be conveyed by the storyteller.

2.2 Understanding ICEWS Datasets

2.2.1 Provincial Modeling

To provide a context for the need for data visualization for the ICEWS Database, we must first understand what is provided in the data. For this project, the provincial event aggregations dataset was used to generate visuals, reports, and predictions. The provincial dataset includes the countries represented in Figure 1, with their respective provinces as detailed.

This type of dataset was created to "support province-level modeling of select Middle Eastern countries." [2] Within this dataset, event types were gathered for each month from years 2005-2013 within each specific province. This dataset has 14,580 unique entries, with each entry having 44 specific aggregations.

Country	Provinces Represented
Egypt	All 27 governorates as of October 2014.
Iraq	18 of the 19 current governates as of 2014, excluding Halabja.
Jordan	All 12 governates as of October 2014.
Libya	All 22 current districts as of 2014.
Saudi Arabia	All 13 provinces as of 2014.
Syria	All 14 governates as of 2014.
Yemen	20 of the 21 current governates, plus Amanat Al Asimah. Not including the Soqatra Governorate.
Occupied Palestinian Territory	Both territories as of October 2014.
Israel	All districts excluding Judea and Samaria Area.

Figure 1: This figure lists all of the countries and provinces used in the ICEWS dataset used to create this project. [2] [3]

2.2.2 Event Types and Weights

While the dataset used for this project has 44 specific aggregations, only 11 of those were used for the DSAPCI interactive website. These included fight, assault, reject, mass violence, demand, disapprove, coerce, threaten, posture, reduce relations, and arrest counts. The definition of each of these events was based on the CAMEO (Conflict and Mediation Event Observations) event definitions as defined in the CAMEO Manual, specifically in the verb codebook.

Each data entry was also given attributes entitled "ALLtALLhosscaleav" and "ALLtALLcoopscaleav." These two event types correspond to the intensity values, both positive and negative, respectively, of the event set. This number is known as the Goldstein Value, and is scaled from -10 to 10. More positive numbers represent more positive and cooperative

Aggregation	Definition
Fight	Use of conventional military force, imposing blockades and restricting movement, occupying territory using armed forces, fight with small arms and light weapons, fight with artillery and tanks, employ aerial weapons, employ precision-guided aerial munitions, employ remotely piloted aerial munitions, violate ceasefire
Assault	Use unconventional violence, take hostages, physically assault, sexually assault, torture, kill by physical assault, conduct suicide bombing, carry out suicide bombing, carry out vehicular bombing, carry out roadside and location bombing, use as human shield, attempt to assassinate, assassinate
Reject	Reject material and economic cooperation, reject military and judicial cooperation, reject intelligence cooperation, reject request or demand for material aid, reject request for economic and military aid, reject request for humanitarian aid and military protection, reject request for political reform and leadership change, reject request to change policy, reject request for rights, reject request for change in regime, refuse to yield, refuse to ease administrative sanctions, refuse to ease popular dissent and release persons, refuse to ease economic sanctions, refuse to allow international involvement, refuse to de-escalate military engagement, reject proposal to negotiate, reject mediation and agreement, defy norms, veto
Mass Violence	Use massive unconventional force, engage in mass expulsion of peoples, engage in mass killings, engage in ethnic cleansing, use weapons of mass destruction, use chemical weapons, use of radiological weapons, use of biological weapons, detonate nuclear weapons
Demand	Demand material cooperation, demand economic cooperation, demand military and judicial cooperation, demand intelligence cooperation, demand policy support, demand material and economic aid, demand military and humanitarian aid, demand military protection, demand political reform and leadership change, demand policy change, demand rights, demand change in institutions, demand easing of administrative sanctions and political dissent, demand release of persons or property, demand easing of economic sanctions, demand to allow international involvement, demand negotiation and de-escalation of military engagement, demand settling of dispute and mediation
Disapprove	Criticize or denounce actions, accuse, accuse of crime, accuse of human rights abuses, accuse of aggression, accuse of war crimes, accuse of espionage, accuse of treason, rally opposition against, complain officially, bring lawsuit against, find guilty or legally liable
Coerce	Seize or damage property, confiscate property by force, destroy property, impose administrative sanctions, impose restrictions on political freedoms, ban political parties, impose curfew, impose martial law, arrest or expel individuals, use repression, attack cybernetically
Threaten	Threaten to reduce or stop aid, threaten to boycott, threaten to break relations, threaten with administrative sanctions, threaten with restrictions on political freedoms, threaten to ban political parties, threaten to impose curfew, threaten to impose martial law, threaten political dissent, threaten to halt negotiations, threaten to halt mediation, threaten to halt international involvement, threaten with repression, threaten with military force, threaten blockade
Posture	Exhibit military or police power, increase police alert status, increase military alert status, mobilize or increase power of police and armed forces, mobilize or increase cyber-forces
Reduce Relations	Reduce or break diplomatic relations, reduce or stop material aid, reduce or stop economic assistance, reduce or stop military assistance, reduce or stop humanitarian assistance, impose embargo, halt negotiations, halt mediation, expel or withdraw peacekeepers, expel or withdraw inspectors, expel or withdraw aid agencies.
Arrest	Legal or extrajudicial arrests, detentions, or imprisonments

Figure 2: This table describes each of the 11 event types focused on in the DSAPCI website and data visualization renderings. Descriptions are used as in alignment with the CAMEO Manual and verb codebook.[3]

actions, whereas negative numbers represent more negative and hostile actions. This number was derived from the Goldstein scale for WEIS (World Event/Interaction Survey) event coding.

1. It is important to note that this scale was developed from work done by a professional political scientist.

2. This scale represents a *neutral* point of view. For example, "the US would consider a military agreement between China and North Korea to be a hostile action, [but] as military agreements in general are defined as being cooperative, it would be classified as a cooperative event." [3]

3. This scale also doesn't take into account the intensity of a singular event. For example, "the killing of 1000 persons is classified as no more hostile than the killing of a single person." [3] This is, in part, the reasoning behind providing the user with individual data visualization options for event type counts, rather than just considering the hostility and cooperation scales. The individual graphics can offer insight into the intensity of the event, whereas the scales lose some of that significance.

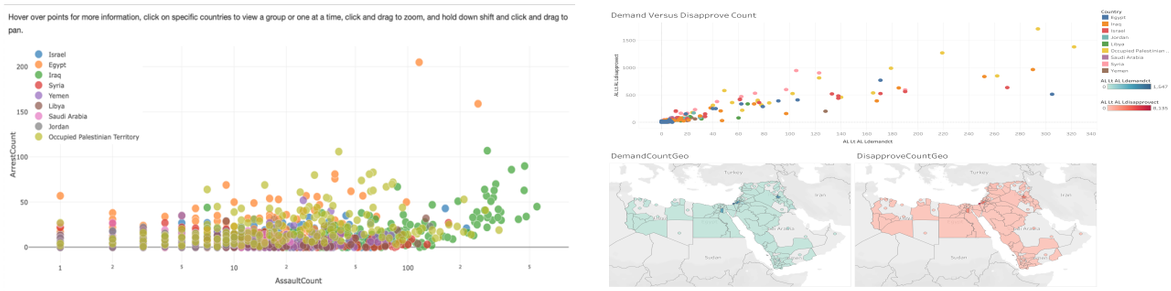
3 The Process of Choosing Software

Throughout the year of working on the DSAPCI website, there was a long process of continually deciding which software was best to use, and which ones had less documentation or proved to be less intuitive.

3.1 Eclipse Versus XCode

At the start of this project, work was began with a Java backend through the Eclipse integrated development environment. While Eclipse is a great resource for Java, it ended up proving to have little documentation and online community to support common errors. For the professional software developer that is a seasoned veteran in web development with Java backends, this would likely not be an issue, but Eclipse is not very beginnger-friendly for the average computer scientist.

After failing to use Eclipse and the Apache Tomcat Server, other avenues were explored. Eventually, after comparing various alternatives, XCode was determined to be the best fit since development was run on a MacBook Pro. XCode offered the flexibility of working with Django, an open-source web framework with extensive documentation and a Python backend. Django offers a wide array of plugins and libraries, and works seamlessly with XCode. It provides an easy start to developing an entire web project, and allows the software developer to focus solely on the development of the website without the distraction of several incompatabilities or outdated documentation.



Scheme 1: The figure on the left displays the Dash application created to display correlation between arrest and assault counts. The figure on the right is a dashboard created with Tableau that displays the correlation between the demand and disapprove counts as well as the demand and disapprove counts by geographic location.

3.2 Dash by Plotly Versus Tableau

When first generating data visualizations, Dash by Plotly provided an excellent framework to quickly and easily generate dashboards that could then be saved as individual images or PDF reports. Dash is a Python framework that is useful for building web applications and has a lot of online documentation. However, the biggest caviat of Dash applications is that they tend to be less deployable. After having a functional Dash application that runs on your local machine, steps must be taken to deploy the app. There are only two options for this deployment: deploying as a Flask app and deploying to Heroku. However, the documentation and commands for this deployment are somewhat outdated and error messages are not documented well by the online community of Dash users. Ultimately, Dash deployment proved to be a very frustrating and under-documented chore, even though the data visualizations were nice.

After failing on multiple accounts to get the Dash app to production, Tableau was considered as an alternative. Tableau is a software company that seeks to change the way people think about data. Tableau offers the ability to customize data visualizations and create predictions for future data. This ability allows the consumer to create workbooks (sets of data visualizations) as well as dashboards (displaying multiple visualizations at once on one screen). Both of these abilities were utilized for the DSAPCI website. In terms of deployment, Tableau provides a service called Tableau Public where workbooks and dashboards can be saved and accessed through an online link. This takes away the hassle of deployment and allows visualizations to be embedded into websites as interactive iframes. Tableau also allows the user to download the data visual in a myriad of types: image, data, Crosstab, PDF, PowerPoint, and Tableau Workbook. While this ability is a great feature, it can also be viewed as a danger, because not all developers may want people to be able to download the dataset. However, since the ICEWS Database is a public database and the datasets are available for download, that was not a concern during this project.

3.3 Python Versus Java

In terms of web development, Python and Java both have various applications that allow for backend development with either language. However, there are some factors to consider when choosing between these popular coding languages for web development.

1. Python is dynamically typed whereas Java is statically typed. [4] For the developer, this means that more can be done in Python in fewer lines of code. This allows for the development process to quickly take off. On the other hand, Java is statically typed which means that it has a stricter set of coding rules which is a guardrail for the developer as it makes code less prone to bugs.

2. Python and Java are both open-source languages with extensive communities to back them. However, in recent years Python has gained more traction than Java and hence has some more advantageous plug-ins and documentation. Depending on the development framework being used, Java can have less documentation with some frameworks than the extensive documentation Python currently has.

3. Speed of development is something to consider when choosing between these languages. For this project, since it was a single individual developing the website, Python was

more advantageous because it was faster to code large web page files than coding in Java.

4. In terms of machine learning and data science, Python has the cutting edge against Java and has very powerful specialized libraries. Python is also a lot more flexible than Java which makes Python better for complex data science projects. [4]

Based on weighing the pros and cons of each of these factors, Python was ultimately chosen over Java because of the flexibility and speed of development.

4 DSAPCI Website

4.1 Website Design

Before a user can utilize a website, they must first decide whether or not they trust the site. This snap judgment happens in a second and is centered on the website's aesthetics. As the usage of internet has increased, the importance of website designed has drastically increased. With so many websites to choose between, a user has the freedom to choose to use which one they like the best, usually based on aesthetic and familiarity of design. One scholar asserts that this snap judgement is based on the perception of five qualities: unity, complexity, intensity, novelty, and interactivity of design. Each of these facets were considered and implemented during the creation of the DSAPCI website.

1. Unity "refers to the congruity among the elements of a design such that they look as though they belong together." [5] When thinking about unity for website design, streamlined fonts, colors, and webpage layout are three crucial aspects. Luckily for the software developer, a tool called Bootstrap makes this streamlining easy. Bootstrap is a front-end component library that provides streamlined HTML and CSS design templates for fonts, buttons, navigation, etc. The main uses of Bootstrap in this website was for the interactive navigation bar and the streamlining of fonts and buttons. Figure 3 shows the navigation bar for the DSAPCI website which was created using Bootstrap.

2. Complexity describes the "amount of information and the differences between different pieces of information that can be found within an aesthetic object." [5] In terms of a website, the number of pages, elements on a webpage, and variety of tools is considered. This was the motivation behind including not only individual visualizations in five types: scatter, line, pie, geographic, and scatter correlations, but it was also the motivation behind combining these elements in various formats to build user reports. This provides a lot of interactivity for the customer and a variety of products to choose from.

3. Intensity of design is defined as the "pitch, hue, or birghtness" that evoke a certain emotional response in the viewer. [5] In website design, intensity of design is a guardrail against having a bland website–one with no excitement to draw the eye to. Using vibrant colors that don't distract from the content of the website, but enhance the user experience is an important aspect of accurately incorparating this aspect. The intensity of a website is in "stirring the emotions, attitudes, and moods of website visitors." [5] This is why vibrant colors like blue and yellow were used for the website framework, and why Tableau visuals were generated with a rich colorscheme.

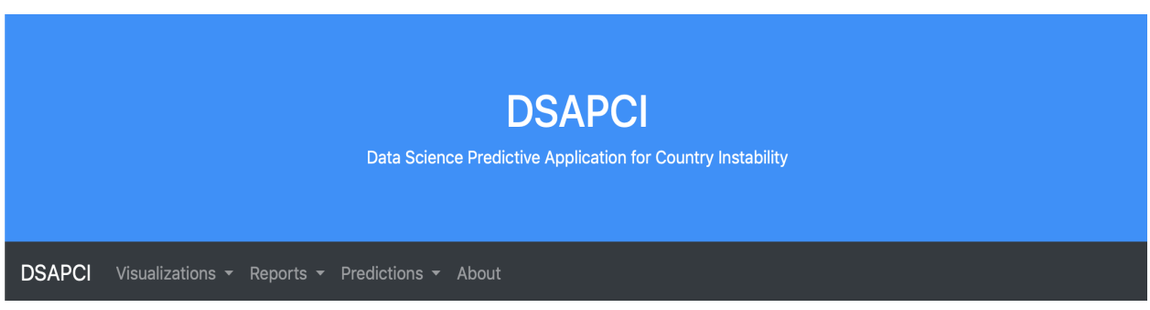


Figure 3: This figure shows the navigation bar for the DSAPCI website that was created using Bootstrap.

4. Novelty of design is the quality of being "new and unusual, different from anything in prior existence." [5] The presentation of a customized interface in website development contributes to this facet. This is, in part, why the DSAPCI website allows for varying products to be customized by the user. The ability that Tableau offers to download in multiple data types contributes to the feeling of novelty to the application.

5. Interactivity of design is defined as the "ability of an artifact to allow users' participation in modifying its form and content." [5] This facet was considered the most important at the onset of this project. The ability for data visualizations to effectively communicate to the user is partly made up of the user's ability to interact with the visualizations and glean truths from their customization. That is why Tableau is such a powerful tool in data visualization, it allows for users to customize visuals by selecting a subset of the data, or changing which type of graph they would like to see, and explore correlations between different variables. Tableau puts the power in the users' hands which allows for maximum interactivity.

4.2 Website Structure

4.2.1 Visualizations

The visualizations tab offers 5 different webpages to choose from, with each webpage containing a workbook of Tableau data visuals specific to that visual type. Every webpage includes the 11 event counts as outlined in Figure 2.

Scatter Comparisons This page is the most unique webpage under the visualizations tab, because it doesn't generate visualizations for singular event types, but rather combines two event types to display a correlation. 6 pairs were created: fight versus assault, reject versus mass violence, demand versus disapprove, coerce versus threaten, posture versus reduce relations, and assault versus arrest. An example of these correlations in displayed in Figure 6 (a).

Scatter Plots This page is similar to the scatter comparisons webpage, but allows the breakout of individual event types and plots over the years from 2005-2013. The advantage of scatter plots is that they offer the ability to easily pick out outliers. Figure 6 (b) shows the visualizations offered by this webpage.

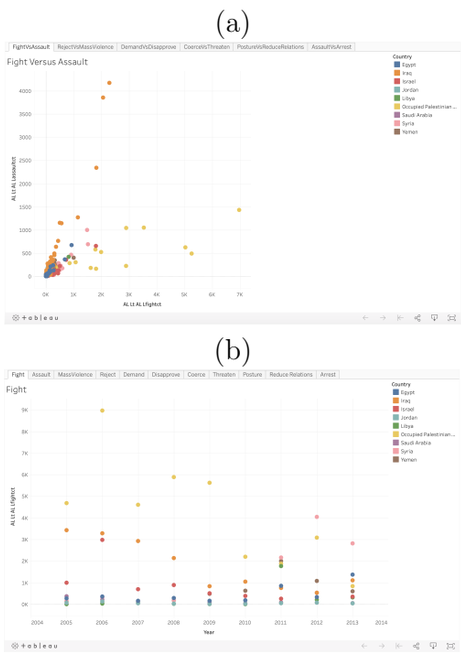
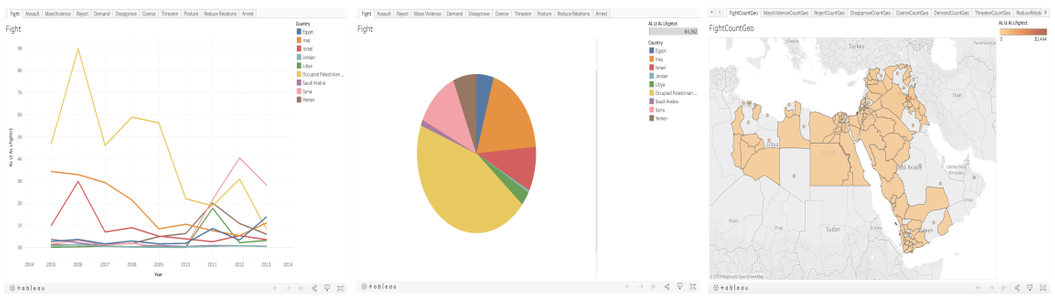


Figure 6: (a) This image is an example of the Scatter Comparisons webpage, with this particular tab displaying the correlation between fight and assault counts. (b) This is an example of the Scatter Plots webpage, with this particular tab displaying the fight count for each country over the years of 2005-2013.

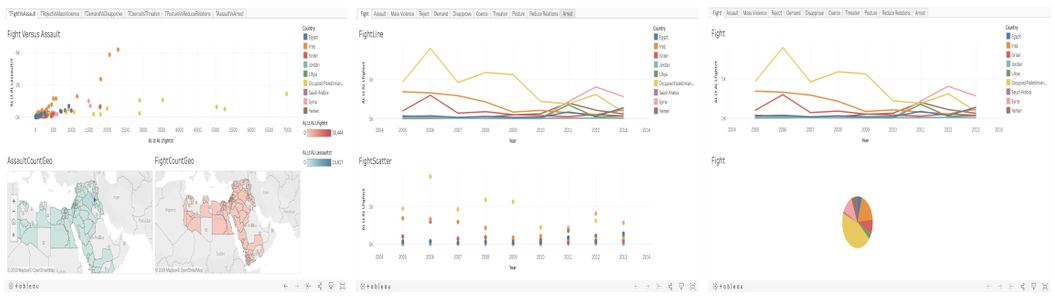
Line Graphs Where scatter plots allow the user to easily view outliers, line graphs allow the user to see the pattern of different data counts over the years. This makes it easier to pick out trends, and to see during which years counts spiked or dramatically fell. Scheme 2 shows the visualizations offered by this webpage.

Pie Charts Pie charts allow the user to see which country has the largest percentage of a certain event type occurring over the entirety of the years used in the dataset: 2005-2013. This aggregates the data into a view that allows the user to see which country has made up the most of a certain event type over this 8 year period. Scheme 2 shows the visualizations offered by this webpage.

Geo Graphs Geo graphs are extremely useful in conveying to the user the importance and reality of what the numbers and visualizations mean. Through putting a country with a statistic, data truths are more easily communicable to the user. Unfortunately, the ICEWS Database does not include specific latitude and longitudes in their provincial dataset, but



Scheme 2: These figures illustrate the webpage data visualizations for each event from 2005 to 2015 for (left to right) Line Graphs, Pie Charts, and Geo Graphs webpages.[]



Scheme 3: These figures illustrate the webpage reports dashboards for (left to right) Scatter Versus Geo, Scatter Versus Line, and Line Versus Pie webpages.

when uploading a dataset to Tableau that contains country names and provinces, Tableau automatically generates these values and allows for the creation of beautiful geo graphs. Scheme 2 shows the visualizations offered by this webpage.

4.2.2 Reports

The reports tab allows for the user to explore dashboards that combine the different data visuals represented under the visualizations tab. The motivation behind the creation of reports was to allow the user to explore correlations and breakout views at the same time. These dashboards are also easily generated as images or PDFs, making it easy for the user to incorporate it into their data visualization report, project, or presentation. Each combination was created with a specific goal for the user in mind, as explained below.

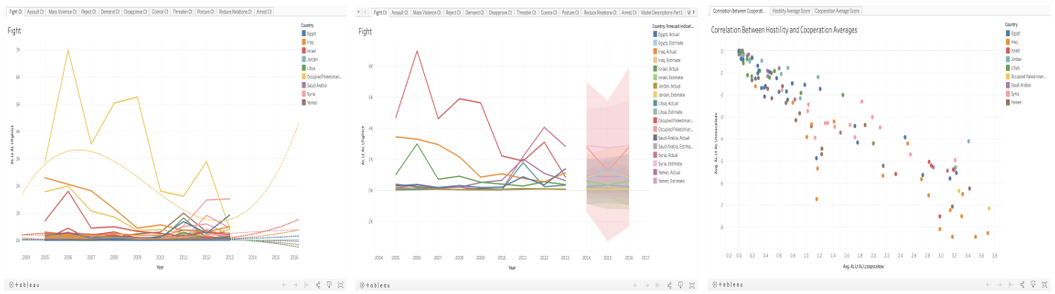
Scatter Versus Geo The Scatter Versus Geo webpage allows for the user to view the scatter correlation plots for two variables while simultaneously viewing the individual geo graphs for both event types used in the scatter correlation. One such example of this is shown in Scheme 3.

Scatter Versus Line While the Scatter Versus Geo webpage uses the scatter correlations, the Scatter Versus Line webpage focuses on single event types rather than correlations. This means that it includes the scatter plot for a singular event as well as its line plot. This allows the user to easily view outliers while simultaneously showing the trend of the country. For example, one could see that Egypt has an outlier point in 2013 for assault counts, but the general trend from 2005-2012 was consistently lower. This allows the user to begin to ask questions about what caused this outlier to occur in the country at the time and make conclusions based on context and the data dashboard. An example of this is shown in Scheme 3.

Line Versus Pie The Line Versus Pie webpage allows the unser to see the trends over the years, and compare the trends with the countries that make up the majority of a specific event count. For example, a country may have a consistently low trend for several years, but because of a spike may make up the majority of the event counts over the years 2005-2013. An example of this is shown in Scheme 3.

4.2.3 Predictions

Trends In Tableau, the analytic package offers trend lines that interpret the data and predict the future trends of the data. For this dashboard, polynomial trend lines were used of



Scheme 4: These figures illustrate the webpage data visualizations for (left to right) Trend Lines, Forecasts, and Scales webpages.[]

order 3. Polynomial modeling was preferred over linear or logarithmic due to the randomness of the data and the presence of several peaks and valleys amongst the average line plot. Tableau's trend lines technology is typically pretty good, but can gain more accuracy with fine-tuned paramaters, like knowing when to use polynomial over logarithmic. By extending the year axis into the future, we have nice visuals that not only seek to describe our current data, but makes a prediction about the trends of data in the future. When hovering over a trend line, a tooltip pop-up contains three items: event type formula, r-squared, and p-value. The event type formula shows the formula calculated and used to generate the trend line. The two statistical values, r-squared and p-value, allow us to know how well our trend line fits the data. R-squared is a goodness-of-fit measure and is always a value between 0 and 1. Values that are closer to 1 tend to express a better model with higher confidence in future predictions. However, before interpreting on the R-squared alone, the p-value must be taken into consideration. A p-value of less than .05 is means that the trend line model Tableau generated may be significant. Combining these measures allows the users to see which trend lines are good predictors for future events, and which ones are not. Due to the random fluctuation of some data, trend lines may not be able to effectively fit certain countries in a certain event type. An example of this webpage is located in Scheme 4.

Forecasts In addition to trend lines, the analytic package that Tableau offers also has a feature entitled forecasting. Forecasting "uses a technique known as exponential smoothing." Exponential smoothing models "forecast future values of a regular time series of values from weighted averages of past values of the series." [6] These algorithms try to find patterns in the data that can be projected into the future. During this process, Tableau uses eight different models and picks the one that generates the best quality forecast. This means that Tableau optimizes the modeling process for you, instead of having to sift through which models are best for your data. However, there are some ways to improve forecasts through fine-tuning certain parameters. By allowing periods to be automatic, we allow Tableau to find which patterns are most evident in the data apart from "seasons". Also, through creating a custom model with seasons set to "additive" we are able to get a better idea of the pattern going into the future. An additive model is one in which the model components are summed, and just says that the component that effects the trend is present in the data. The descriptions of each model are also present in three dashboard tabs. An example of this webpage is shown in Scheme 4.

Scales The scales tab expresses the correlation between a country's cooperation average score and their hostility average score. This is in reference to the Goldstein value that was created for each country. Please refer back to section 2.2.2 for a more detailed explanation of these weights. An example of this webpage is shown in Scheme 4.

5 Conclusions

Through this project, every aspect of web development was explored. From the need for data visualization and how to craft a pleasing website to writing backend code, I was able to learn about what it takes to build a website from scratch. While this can be a process with its ups and downs, ultimately the product can be a beautiful website that seeks to help others in their data exploration. I hope that the DSAPCI website can be one of many examples for years to come on how we seek to convey data truths to the average user.

Acknowledgements

I would like to express my gratitude to Professor Lancie Affonso for his continued guidance on this project. Thank you to the Honors College at the College of Charleston for their continued support and guidance in seeing this project through to fruition.

References

[1] e. a. Nathalie Henry Riche, "Data-driven storytelling," http://ebookcentral.proquest.com/lib/cofc/detail.action?docID=5331742, 2018.

[2] e. a. Boschee, Elizabeth, "Icews events and aggregations.pdf," https://doi.org/10.7910/DVN/28118/OXJF12, 2015.

[3] P. A. Schrodt, "Cameo conflict and mediation event observations event and actor codebook," data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf, 2012.

[4] C. Addicott, "Python vs java comparison – which language to choose for your app," www.netguru.com/blog/python-vs-java-comparison-which-language-to-choose-for-your-app., 2018.

[5] e. a. Jiang, Zhenhui (Jack), "The determinants and impacts of aesthetics in users' first interaction with websites." *Journal of Management Information Systems*, vol. 33, no. 1, pp. 229–259, 2016.

[6] "How forecasting works in tableau," https://onlinehelp.tableau.com/current/pro/desktop/en-us/forecast_how_it_works.htm.