SIADS 591 – The Relationship Between Terrorism and Tourism
Jim Creighton and Brooke Hawkins

Motivation

Our overarching project question is: How does terrorism in a country relate to tourism? Understanding the relationship between terrorism and tourism is consequential for countries and businesses depending on tourism income. For example, after a series of local terrorist incidents, a hotel company may want to know whether to reduce room capacity and a government may want to anticipate a change in tourism revenue.

We narrowed the scope of our terrorism and tourism question in two key ways.

- We focused on data from 2012 to 2018. We planned to use a decade of data, but the global terrorism dataset changed its data collection methodology mid-2011.
- We compare two groups of countries: High Tourism and Low Tourism countries. We analyzed data from ten countries in each group. We chose ten countries so that we would have enough data to analyze trends in each group (ten countries across seven years for a sample size of 70), but not so much data that it would overwhelm visualizations.

What specific question or goal did you try to address?
We developed three specific questions to investigate how terrorism relates to tourism.

1. DISTRIBUTION: How does the distribution of incident counts and incident severity vary between High and Low Tourism countries?
2. RELATIONSHIP: What is the relationship between the number of terrorist incidents and the percent change in international tourist arrivals in the following year for high and low tourism countries?
3. GROWTH: Are the number of terrorist incidents changing at a rate faster or slower in low tourism countries than high tourism countries?

Data Sources

For a detailed explanation of the important variables in each derived dataset, please refer to the file **Data Dictionary.xlsx.** For step-by-step instructions on how to download each dataset, please refer to *“Appendix 1: Data Download Instructions”* at the end of the report.

Data Source A: World Bank International Tourism Arrivals Dataset

Location: https://data.worldbank.org/indicator/ST.INT.ARVL (CSV file, 4MB)
Important variables: Country Name, Country Code, Year, Arrivals
How many records used or retrieved: 270 records x 58 columns
Time period used: 2012-2018

Data Source B: World Bank Population Dataset

Location: https://data.worldbank.org/indicator/SP.POP.TOTL (XML file, 4MB)
Important variables: Country Name, Country Code, Year, Population
How many records used or retrieved: 270 records x 58 columns

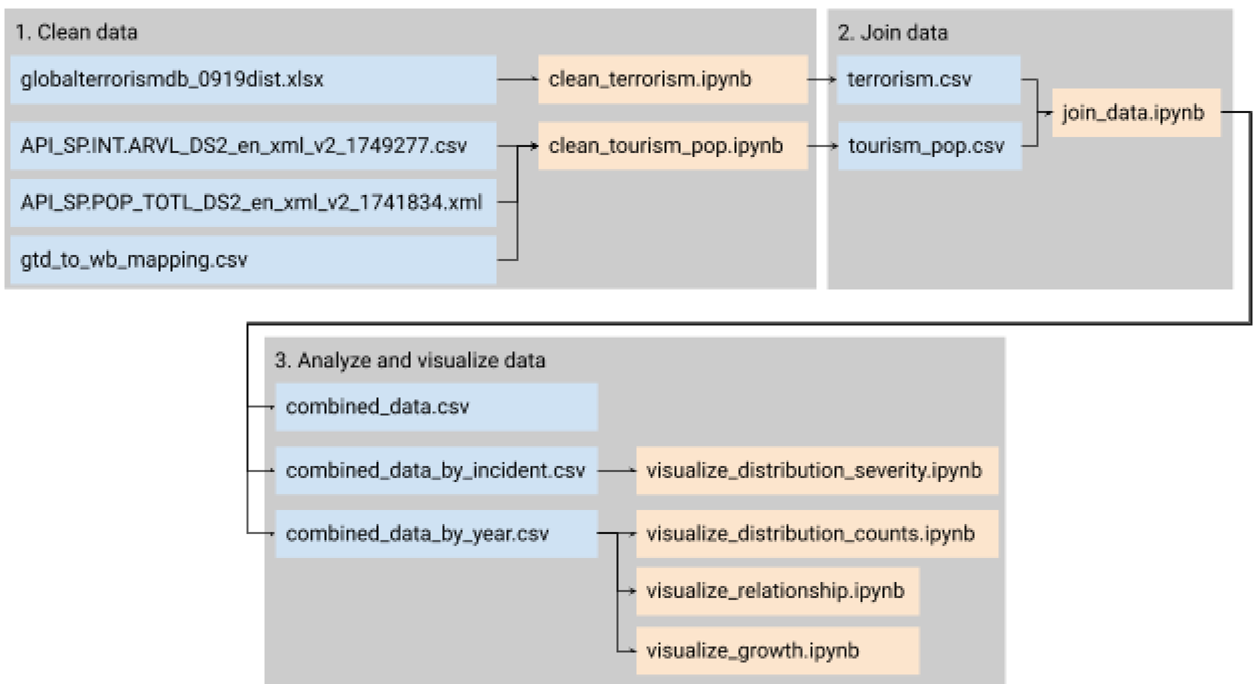
Time period used: 2012-2018

Data Source C: Global Terrorism Database (GTD)

Location: https://www.start.umd.edu/gtd/about/ (Excel spreadsheet, 92 MB)
(See Appendix 1: Data Download Instructions for further download details)
Important variables: Country Name, Country Code, Year, Count Wounded, Count Killed
How many records used or retrieved: 191,464 records x 135 columns
Time period used: 2012-2018

Data Manipulation

Workflow Overview



1. Clean data. In a Jupyter notebook, we cleaned the Global Terrorism Database (GTD), and generated a CSV file for the Join step. In another notebook, we cleaned and combined the World Bank Population and Tourism datasets, creating a second CSV for the Join.
2. Join data. A single notebook was used to join the cleaned GTD data to the cleaned World Bank Tourism and Population data. We created two CSV files needed for visualizations - one that grouped terrorism events by *incidents* and one that grouped incidents by *year*.
3. Analyze and visualize data. We created four notebooks for our main study questions; two to visualize distributions, one to visualize a relationship, and one to visualize growth.

Data Dictionary: We created a data dictionary to define derived variable names and types, along with data origins and how they were to be used, in each step of the analysis. For additional details on the variables included in each derived dataset, please refer to the Data Dictionary.

Code: We programmed all of our analyses in Jupyter notebooks with Python 3. We used the numpy and pandas libraries for data manipulation, and matplotlib and seaborn libraries for data visualization.
GitHub: We used a GitHub repository to collaborate and version control all raw data, Jupyter notebooks, and derived CSV files.

1. Data Manipulation: Clean Data

Sources A and B: World Bank Tourism Arrivals & Population

How specifically did you manipulate the data? How did you perform conversion/processing?
There are several steps in the data cleaning notebook clean_terrorism.ipynb.

1. Import the raw World Bank Tourism data from API_ST.INT.ARVL_DS2_en_csv_v2_1740745.csv with **read_csv()** and Population data from API_SP.POP.TOTL_DS2_en_xml_v2_1741834.xml with an **ElementTree** script
2. Filter both datasets for the 2012-2018 timeframe with **boolean indexing**
3. Join the Tourism and Population data into one dataframe with **merge()**
4. Match variable names and data types from data dictionary with **rename()** and **astype()**
5. Import gtd_to_wb_country_mapping.csv and use **merge()** to map Global Terrorism Database country codes to the World Bank data
6. Add a tourism rank column based on Tourism Arrivals to enable future High and Low Tourism country groupings with **rank( )**
7. Export derived dataset tourism_pop.csv with **to_csv()**

Important variables for joins

- The two World Bank datasets (of Tourism and Population) were joined based upon the standardized country names that the World Bank uses.
- The combined World Bank data was then joined to a manually created file gtd_to_wb_country_mapping. that contained a mapping of the World Bank country codes to the corresponding GTD country codes.

How did you handle missing, incomplete, or incorrect data?

- The only missing data encountered in the World Bank datasets was when there was no value reported for specific years. Because our analysis was predicated on studying the trends in each year, we dropped those countries from the data.

What challenges did you encounter and how did you solve them?

- As mentioned, missing World Bank data led us to drop approximately 30 less-developed countries from our analysis. This created an information loss because those included a few hotspots such as Afghanistan, Iraq, and Somalia. Despite this, we stuck with the World Bank dataset since it was by far the most comprehensive set found.
- The GTD and World Bank datasets had different country names and country codes. To enable a future join between the datasets, we manually created a CSV file that mapped GTD country codes to World Bank country codes. In multiple cases the GTD data had legacy names of countries which merged or split apart over time. Through manual web research, we made a best effort to match countries across the two datasets, but still had to drop some obscure countries (<10) for which we could not create a mapping.
- Importing the XML file of the World Bank Population dataset was a challenge. First, there is no standard pandas ability to import XML, so we employed the ElementTree module. Second, the World Bank uses a proprietary XML format that nests the data one level

lower than standard XML files. Through online research and external advice, we were eventually able to import the file by modifying an ElementTree script.

Source C: Global Terrorism Database

How specifically did you manipulate the data? How did you perform conversion/processing?
There are several steps in the data cleaning notebook clean_terrorism.ipynb.

1. Import data from Excel spreadsheet globalterrorismdb_0919dist.xlsx with **read_xlsx()**
2. Filter dataset with **boolean indexing** of dataframe
    - Filter for records that occurred since 2012 based on year column
    - Filter for records that are certainly terrorism based on doubt terrorism column
    - Filter for records that are successful terrorism based on success column
3. Add missing data
    - Convert blank cells to 0's in number of killed and wounded with **fillna()**
    - Create number of victims column by summing number killed and wounded
4. Match variable names and data types to data dictionary
    - Recast data type of text columns from objects to strings with **astype()**
    - Rename columns to match data dictionary with **rename()**
5. Export derived dataset terrorism.csv with **to_csv()**

How did you handle missing, incomplete, or incorrect data?

- The columns with the number of people killed and the number of people non-fatally injured were incomplete; with 3000 and 5216 blank values, respectively. Blank values indicated that there was not enough information in the news reports used to generate records to determine these values. We decided to fill these cells with 0 values. This data transformation could result in an underestimate of these values, but our decision matched an existing undercounting bias in the dataset, which would report the lowest counts from conflicting recent reports. We also preferred an underestimating incident severity to excluding incidents altogether.
- The dataset was missing a column with the number of victims killed or injured in an incident, which we used to visualize incident severity. We created this column by summing the number of people killed and non-fatally injured in an incident.

What challenges did you encounter and how did you solve them?

- The biggest challenge was deciding how to filter records in the dataset. We decided to include those that were certainly terrorism (rather than other types of criminal activity) and that were successful (had a tangible effect). We included flags in our notebook to change these filters, and to additionally filter based on attack type and target type. We did not have the time to explore all reasonable avenues of data transformation, but we think it would be a good avenue for future work.
- The size of the GTD was also a challenge. We needed to track the file with Git large file storage in order to add the data to our repository on GitHub. It also took over five minutes to load the dataset as a pandas dataframe, which was a bottleneck to rerunning the notebook while developing it. It would have been reasonable to consider different computation techniques designed for bigger datasets, such as Spark.

2. Data Manipulation: Join Data

How specifically did you manipulate the data? How did you perform conversion/processing?
There are five main steps in the data manipulation notebook join_data.ipynb.

1. Import clean data from CSV file tourism_pop.csv and terrorism.csv with **read_csv()**
2. Add tourism group column based on tourism rank column with **map()**
3. Join datasets
    - Join datasets based on year and country code columns with **join()**
    - Write dataset as CSV file combined_data.csv with **to_csv()**
4. Group related incidents into one record
    - Group records that are related and occur in same city on same day based on date columns, location columns, and related column with **groupby()**
    - Aggregate data across records with **agg()**
    - Write dataset as CSV file combined_data_by_incident.csv with **to_csv()**
5. Group records by country and year into one record
    - Group records by year and country columns with **groupby()**
    - Aggregate data across records with **agg()**
    - Write dataset as CSV file combined_data_by_year.csv with **to_csv()**

Important variables for joins

- The two cleaned datasets (of tourism/population and terrorism) were joined based on the year column and the standardized country code column.

How did you handle missing, incomplete, or incorrect data?

- Missing tourism/population data: If a country had terrorism data but no tourism data, we dropped its terrorism records since our analysis is focused on tourism groups.
- Missing terrorism data: The GTD tracks terrorist incidents. Therefore, it has no records corresponding to a country without terrorism. If a country had tourism data but no terrorism data for a year, we kept the tourism records, and assumed the number of terrorist incidents and victims were 0.
- Incomplete incident data: We collapsed related terrorist records into one event. The GTD separates incidents conducted by the same group for the same target if they were discontinuous in time or space. For example, three bombings in one city at different times on one day would have three separate records. The GTD includes a related column, which indicates any related records for each incident. For our analysis, we decided related events should be considered one incident, if they occurred in the same city and same day. We chose this approach because we saw cases where one event had several records, yet the general public (importantly, tourists) would view it as one event.

What challenges did you encounter and how did you solve them?

- The biggest challenge was matching our dataset format to our desired visualizations. In particular, one visualization required each record to represent an *incident*, and the remaining visualizations required each record to represent *all incidents in a country in a year*. To solve this problem, we repeated the grouping and aggregation steps with slightly different group bys, and generated two datasets: combined_data_by_incident.csv and combined_data_by_year.csv.

3. Analysis and Visualization

Question 1: DISTRIBUTION: How does the distribution of incident counts and incident severity vary between high and low tourism countries?

Before exploring the relationship and growth rate of terrorism and tourism, we started with an exploratory plot to understand how incidents compare betweeh High and Low Tourism countries.

Country Groupings

High Tourism (top 10)

Rank	Country	Tourist Arrivals (000s)	Incidents
1	France	81,980	16
2	United States	66,967	16
3	China	57,725	3
4	Spain	57,464	0
5	Italy	46,360	7
6	Turkey	35,698	100
7	Germany	30,411	2
8	United Kingdom	29,282	23
9	Russian Federation	28,177	90
10	Malaysia	25,033	2

Low Tourism (bottom 10)

Rank	Country	Tourist Arrivals (000s)	Incidents
145	Moldova	89	0
146	Dominica	79	0
147	Liechtenstein	62	0
148	Sierra Leone	60	0
149	Timor-Leste	58	0
150	Tonga	47	0
151	Mali	32	17
152	Solomon Islands	24	0
153	Comoros	23	0
154	Tuvalu	1	0

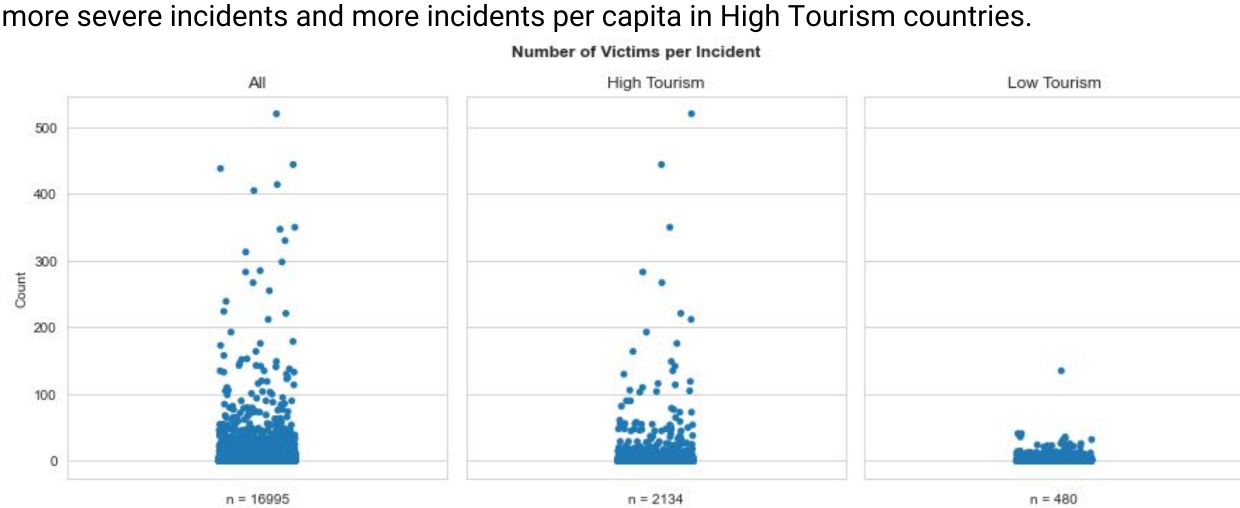
** 2012 Data*

What analysis steps did you perform?

- We chose to visualize the distribution of incidents with two swarm plots. We used swarm plots to represent each data point discretely and clearly, and to implicitly represent the density of data points in an intuitive way. To understand the distribution of incident severity, we plotted the number of victims per terrorist incident (y-axis) for each tourism group (x-axis), where each data point represents one incident.
- To understand the distribution of incident counts, we plotted the number of terrorist incidents per capita (y-axis) for each tourism group (x-axis), where each data point represents one country and one year. Because each country has varying populations (and therefore cities and numbers of terrorist targets), we normalized the incident counts. We therefore chose terrorist incidents *per capita* rather than incident counts, allowing a fairer cross-country comparison.
- We wrote a function called **plot_strip_plot()** to visualize data that uses seaborn's **stripplot()** function for each tourism group.

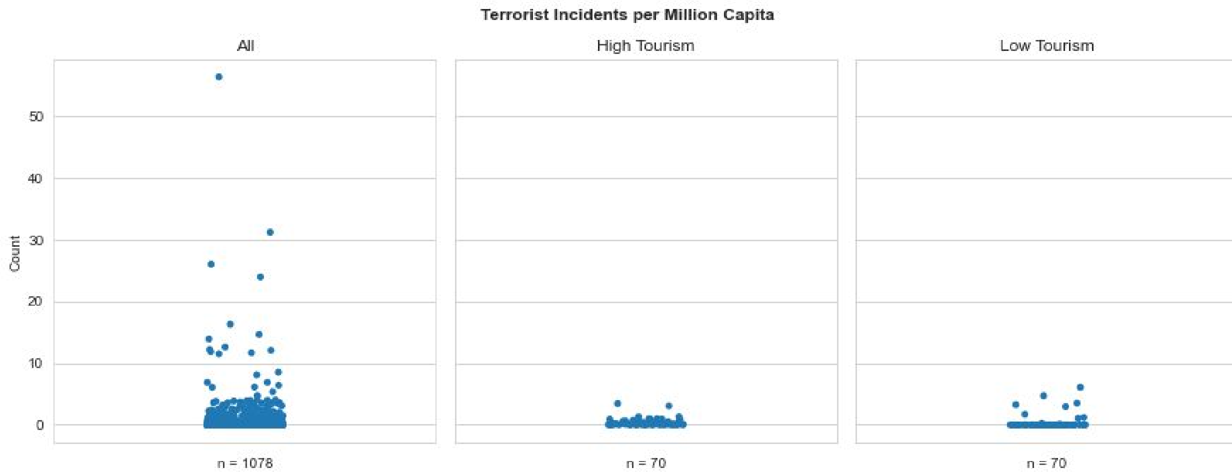
What interesting relationships or insights did you get from your analysis?

We expected to see incidents and victims across most countries and most years. We expected more severe incidents and more incidents per capita in High Tourism countries.



We found that:

- Most incidents have no victims. In High and Low Tourism countries, nearly half of incidents have 0 victims and fall along the x-axis (1016 incidents in High, 199 in Low).
- On average, High and Low Tourism countries have equally severe incidents (median of 1 victim per incident). However, there are exceptions in the high tourism countries that have more victims per incident than any Low Tourism countries.
- Higher Tourism countries have more incidents, which could be related to higher populations or number of cities in these larger countries.



We found that:

- In All countries, there are few incidents overall (most values are zero). Out of the 1078 records per year and country, over half (598) had 0 incidents.
- On average, High Tourism countries have more incidents per million capita (0.13 median) than Low Tourism countries (0.00 median). However, there are exceptions in the Low Tourism countries (namely, Mali) that have more incidents per million capita than any High Tourism countries.

What didn't work, and why?

- Originally, we proposed creating violin plots to visualize these distributions, though the many zero values (years without incidents) made our original mockup of a violin plot somewhat irrelevant. All violin plots depicted a high density of points around zero, so we switched to the discrete encoding of a strip plot. This enabled easier comparison of non-zero values, while still representing that most values were zero.
- We also tried overlaying box-and-whisker plots to show the distribution, but once again, the zero values made this unhelpful. We opted to leave the strip plots as is.

Question 2: RELATIONSHIP: What is the relationship between the number of terrorist incidents and the percent change in international tourist arrivals in the following year for High and Low Tourism countries?

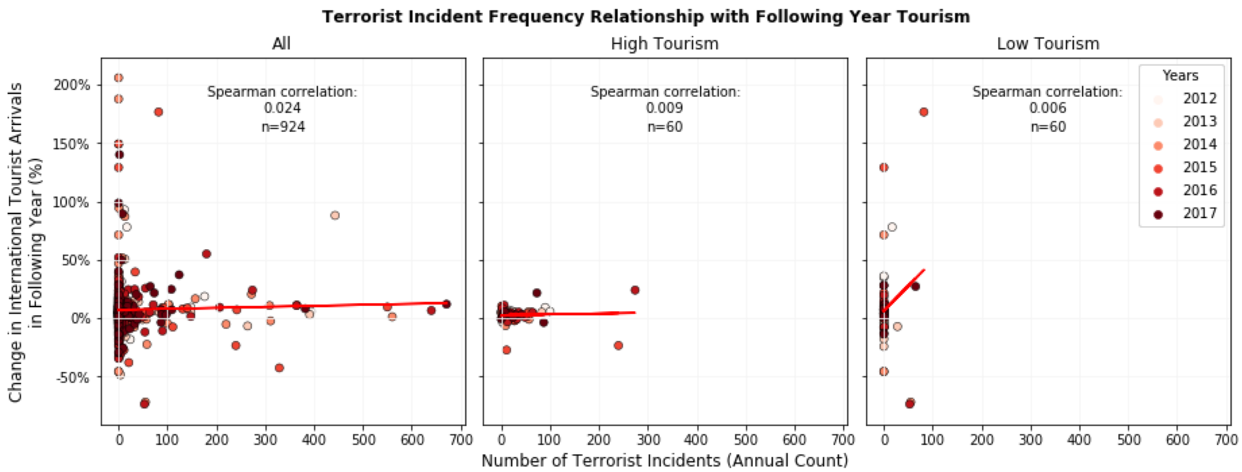
What analysis steps did you perform?

- We chose to explore this relationship with scatter plots, plotting the annual terrorist incident count of each country (x-axis) against the year-to-year percent change in tourist volume (y-axis) the following year, where each data point represents one country and one year. We then found a best-fit line and correlation.

- Because each country has very different tourism volumes, we normalized the data to plot *percent* change rather than tourism *count* change.
- We used a **groupby()** of each country and created a new column using **df.pct_change()** to calculate the tourism percent change.
- We calculated a best-fit line with **np.polyfit()** and a correlation with **.spearmanr()** .

What interesting relationships or insights did you get from your analysis?

Our initial hypothesis had been that we would see a decline in a specific country's tourist arrivals in the year following a large number of terrorism incidents and that this effect would be more pronounced in countries with greater tourism.



We found that:

- There was no meaningful relationship between the number of terrorist incidents and change in tourist volume the following year. Spearman correlation coefficients were almost zero. In most years, the number of incidents is zero (or very low), causing the slope of most best-fit lines to swing heavily based upon which outliers were included.
- Low Tourism countries had very few incidents in most years. In instances where the incident counts were above 10, it was equally likely to see an increase in tourism as a decrease. The range of annual tourism percent change was larger (both positive and negative) for these countries in part because they had a small tourism base.
- Even when zooming in on a smaller range of the High Tourism percent change (than shown in the visualization), there is no discernible trend for High Tourism countries.

What didn't work, and why?

- We intended to use **.pearsonr()** to calculate correlation, but the incident counts were weighted toward zero (not normally distributed). We therefore switched to **.spearmanr()**.

Question 3: GROWTH: Are the number of terrorist incidents changing at a rate faster or slower in Low Tourism countries than High Tourism countries?

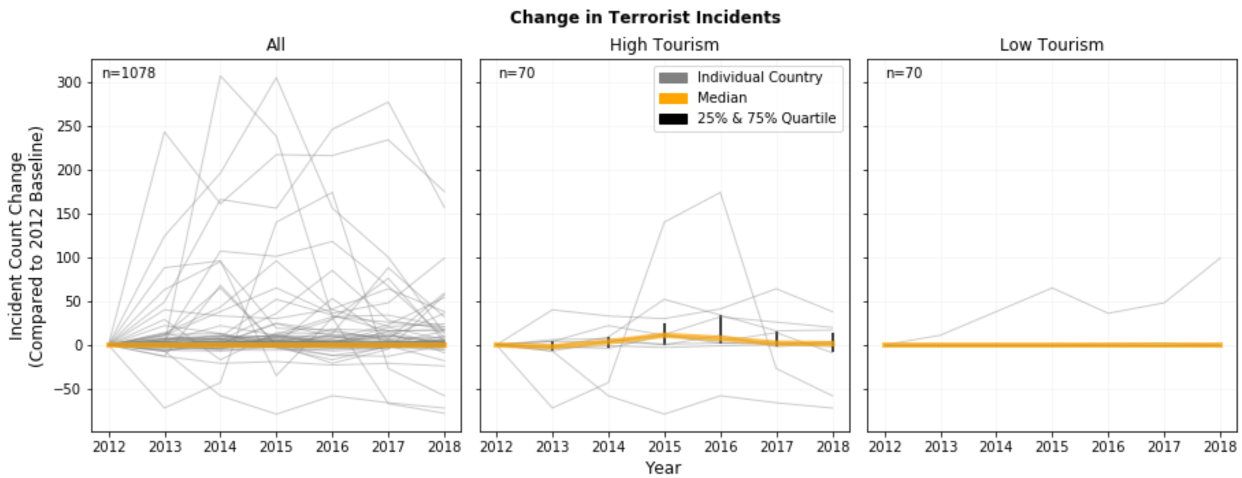
What analysis steps did you perform?

- We chose to explore this trend with time series line graphs plotting year (x-axis) against the terrorism count change since 2012 (y-axis), where each data point represents one country and one year.

- We normalized the incident data against a baseline, since there's a range of starting points of terrorism. For example: If Aruba had 20 incidents in 2012 and 30 incidents in 2014, 2014 would be plotted as 10.
- Annual incident counts were available, but we needed to calculate change since 2012.
- The dataframe was pivoted from long to wide format, with count differences as values, to enable plotting and calculation of medians and quartiles (of country groupings).
- Because the data was not normally distributed, we chose to represent the medians and 25% / 75% quartile ranges (rather than the originally-planned averages and standard deviations) to show variation within the country groupings.

What interesting relationships or insights did you get from your analysis?

We expected High Tourism countries to have a higher growth rate due to terrorist groups selecting tourist destinations with the aim of gaining greater publicity.



We found that:

- There is no evidence that the number of incidents is increasing globally. Incidents vary by year, and some countries have more variability than others (Egypt, India, Philippines, Turkey, and Ukraine), but the median changes in terrorism incidents in all tourism groups remains at 2012-like levels.
- Among High Tourism countries, median incident count change peaked at 11 in 2015 and decreased to 1.5 by 2018. The interquartile range peaked at 35 in 2016, but was often below 15. This suggests a slight uptick in terrorism during 2015 and 2016 and that some countries (namely, Turkey) in the High Tourism group experience greater variability.
- Among Low Tourism countries, there are so many years with zero incidents (exception of Mali) that the median and interquartile range remain at zero throughout the time span. This suggests that Low countries are experiencing little growth or variability in terrorism.

What didn't work, and why?

- We had originally normalized the data by creating an *index* to 2012: we calculated current year incidents divided by 2012 incidents. We abandoned this approach because we found that many countries (including 9/10 Low Tourism countries) had zero 2012 incidents. We therefore instead used a count difference compared to the 2012 baseline.

Summary and Future Exploration

Our analysis did not support many of the initial hypotheses that we held about terrorist incident occurrences, their growth, and their relation to tourism. We found:

- Terrorist incidents are often outliers, thankfully. Many countries have no incidents yearly.
- High and Low Tourism countries have equally severe incidents.
- High Tourism countries have more incidents, and incidents per capita, than Low.
- Indications of only marginal growth of incidents in High Tourism countries since 2012.
- No relationship between terrorism and tourism (in the following years).

It's possible that our findings could change by taking a different view of the data, such as:

- Our temporal resolution may be too coarse to capture near-term effects or too fine to capture long-term effects.
- Perhaps not all terrorist incidents are equal in terms of their relationship with future tourism. Severity of incident (casualties), type of target (tourists vs infrastructure), and weapons are a few factors to consider.
- High vs Low Tourism countries could be an uninformative grouping. Grouping by region or by size of economy (GDP) may be more instructive.
- Looking at tourist arrivals for a country in aggregate may be less informative than considering where tourists originated. For example, US citizens may be more likely to reduce travel to the UK after several incidents than a French citizen.
- There are many reasonable approaches to this dataset. Instead of selecting one potential path, it could be useful to explore all paths with a multiverse analysis.

A literature review suggests identifiable relationships emerge only in very specific instances.

Final Thoughts

- Through this analysis we grew to appreciate the challenge in just collecting terrorist incident data. Despite the GTD being a very good source, we question whether it suffers from an availability bias. Specifically, we suspect that terrorist incidents are unreported or underreported in lower tourism and less developed countries, or are reported by news outlets or in languages that are not monitored by the Global Terrorism Database.
- Our general expectation was that terrorism worldwide was ocurring in most countries, was growing, and was impacting tourism. While we don't know the root of this probable overestimation, we suspect that terrorist incidents with many victims are more newsworthy and receive more coverage in the US, even if they are not representative of incidents overall. Media tendency to overreport severe events likely shaped expectations.

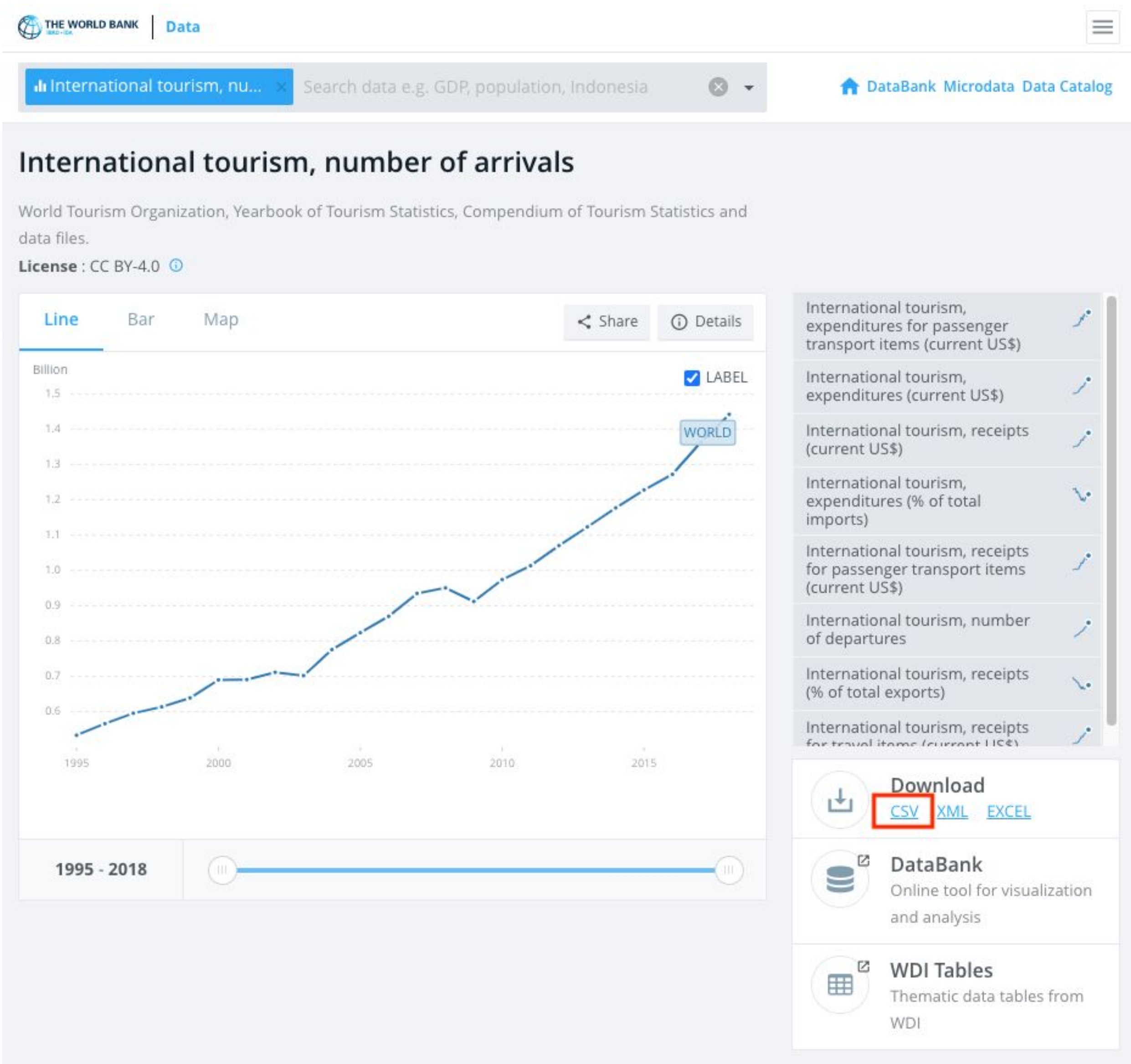
Statement of Work

We both sought additional experience across the data manipulation, visualization, and writing phases. We therefore approached this collaboratively, attempting to segment each phase equally. Initially, Jim defined a data dictionary while Brooke developed a workflow, including GitHub usage. Jim took the lead on World Bank tourism and population datasets while Brooke worked with the Global Terrorism Database and constructed the larger joins. Brooke conducted the Distribution visualizations while Jim conducted the Growth and Relationship visualizations. Both wrote the final report. Throughout the project, we paused to review each other's work and give feedback. Overall, we were pleased with how we worked in tandem.

Appendix 1: Data Download Instructions

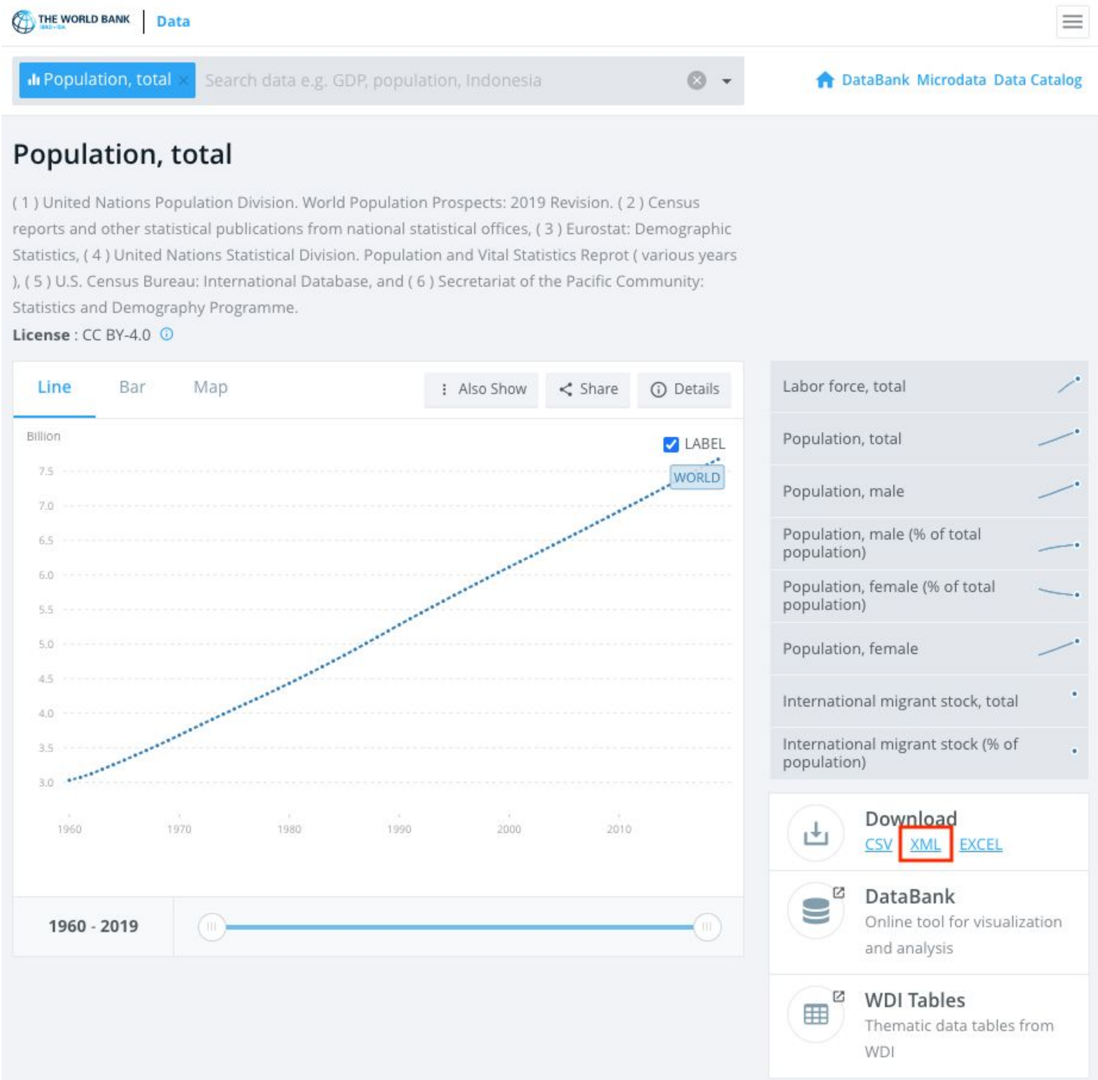
Data Source A: World Bank Tourism Arrivals Dataset

1. Navigate to https://data.worldbank.org/indicator/ST.INT.ARVL
2. Select **CSV** in Download sidebar



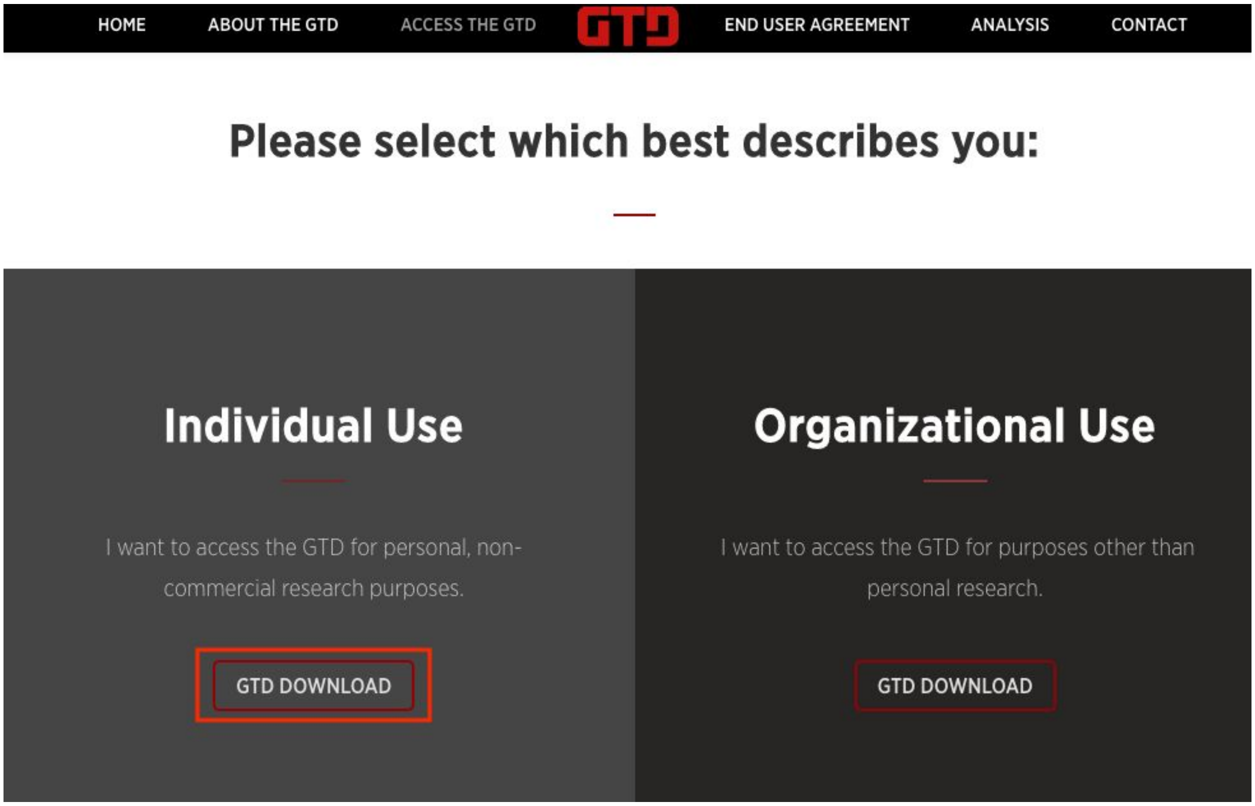
Data Source B: World Bank Population Dataset

1. Navigate to https://data.worldbank.org/indicator/SP.POP.TOTL
2. Select **XML** in Download sidebar

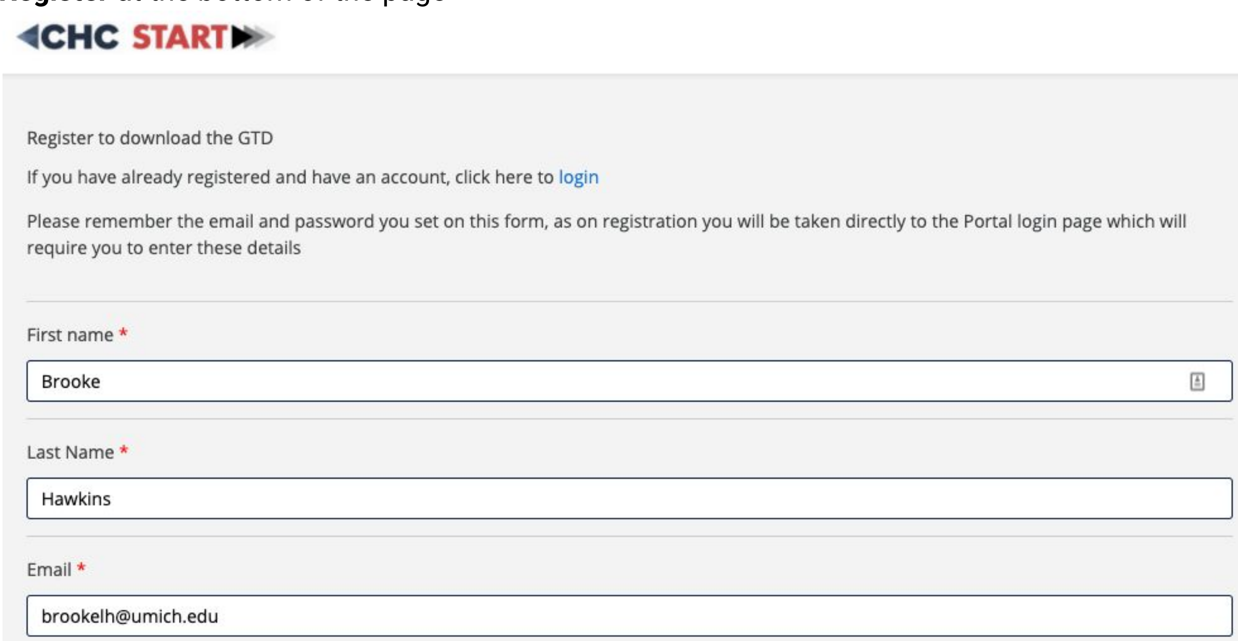


Data Source C: Global Terrorism Dataset

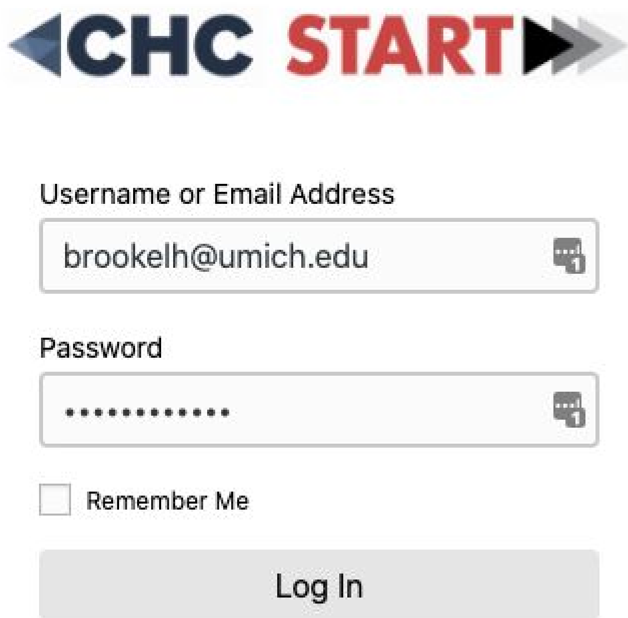
1. Navigate to https://www.start.umd.edu/gtd/access/
2. Scroll down the page, and select **GTD Download** in the Individual Use section



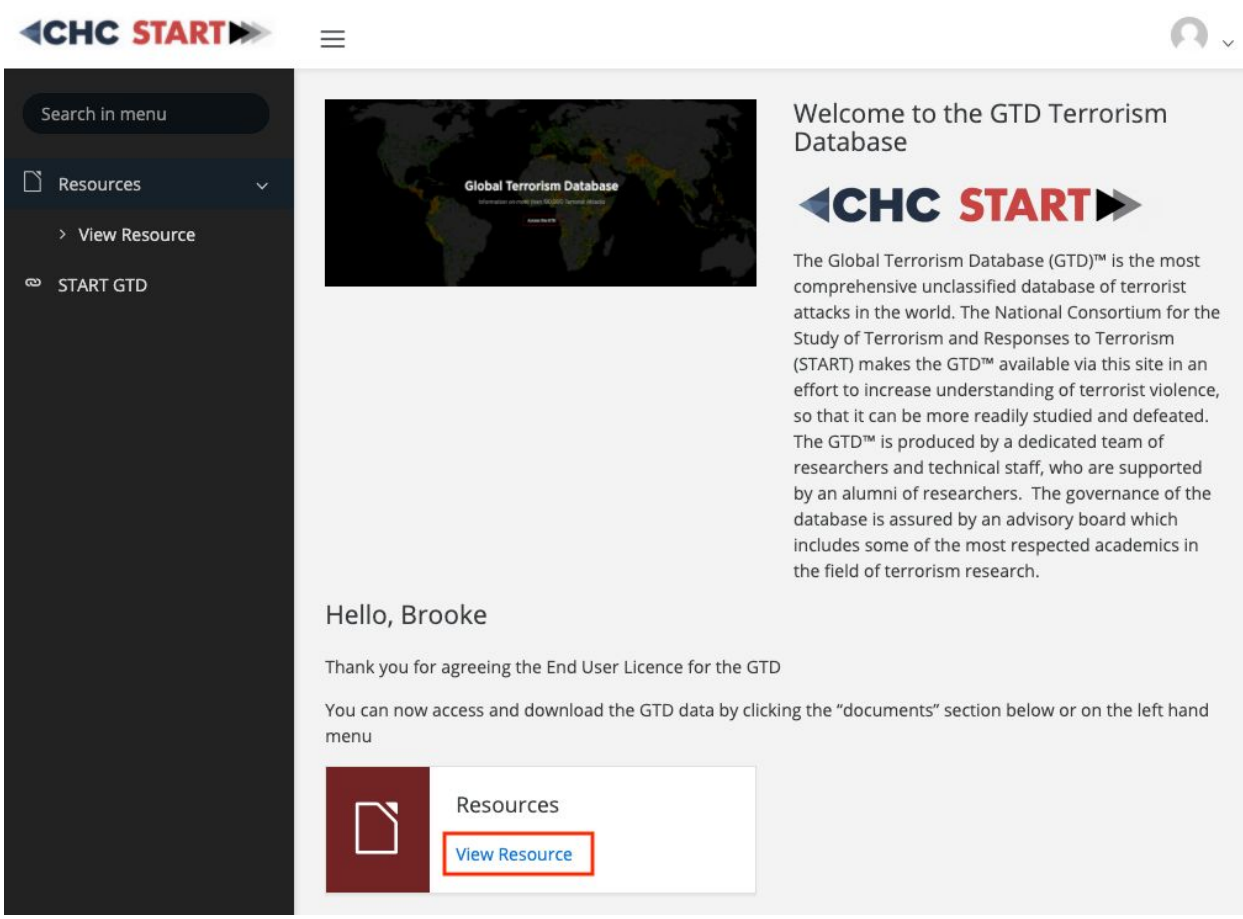
3. On the registration page, fill in your information, agree to EULA agrrement and select **Register** at the bottom of the page



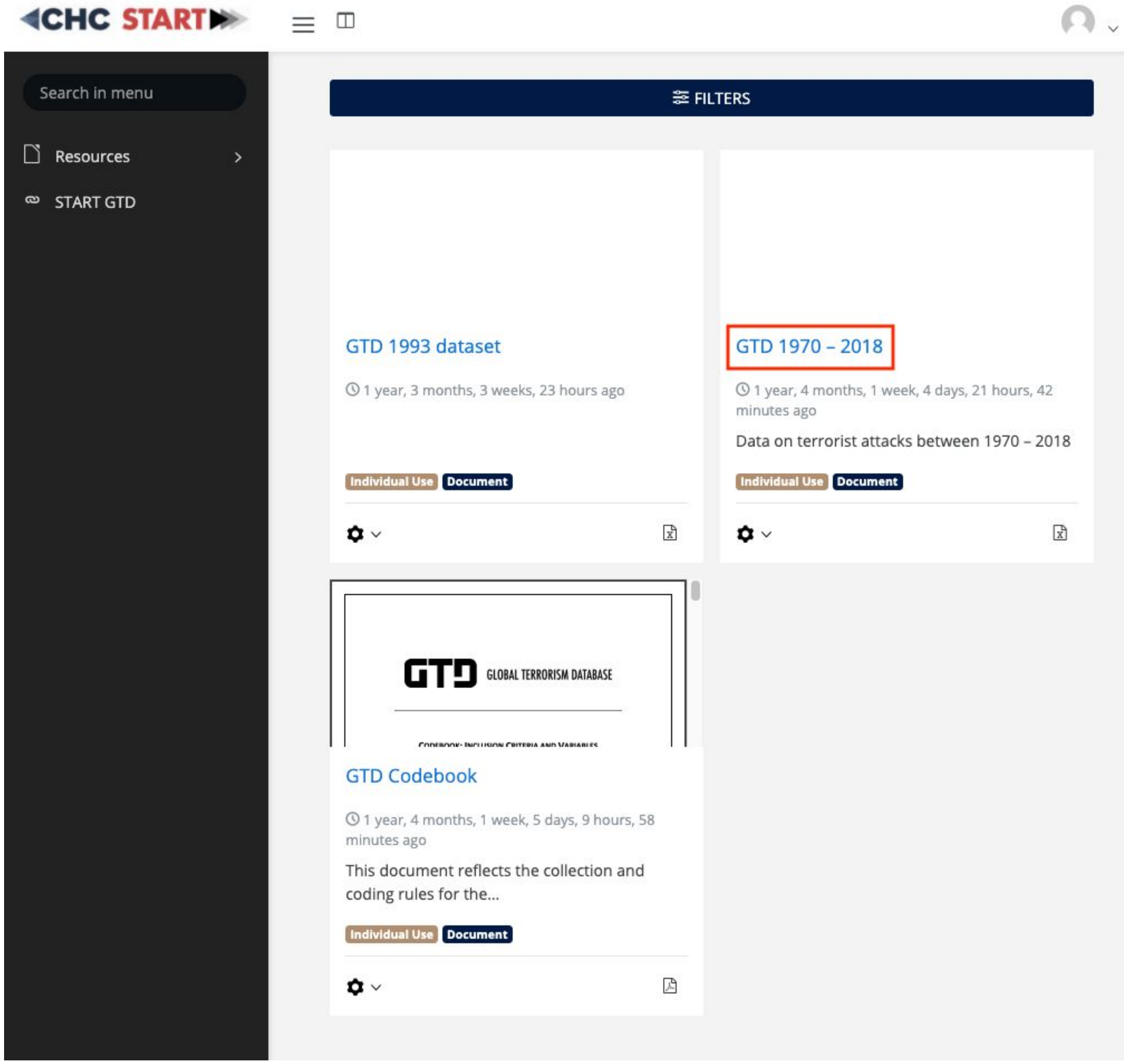
4. On the login page, enter your username and password to and select **Log In**



5. On the GTD Terrorism Database page, select **View Resource**



6. On the resources page, select **GTD 1970 - 2018**



7. On the GTD 1970 - 2018 page, select **DOWNLOAD ATTACHMENT - GLOBALTERRORISMDB_09191DIST**