Surafel Demssie, Rakeb Teklehiwot, Anam Yasin, Md Al Amin Iqbal, ChiaWen Chen and Nelson Reyes

INST 490

May 5th, 2020

Professor Mary Francis

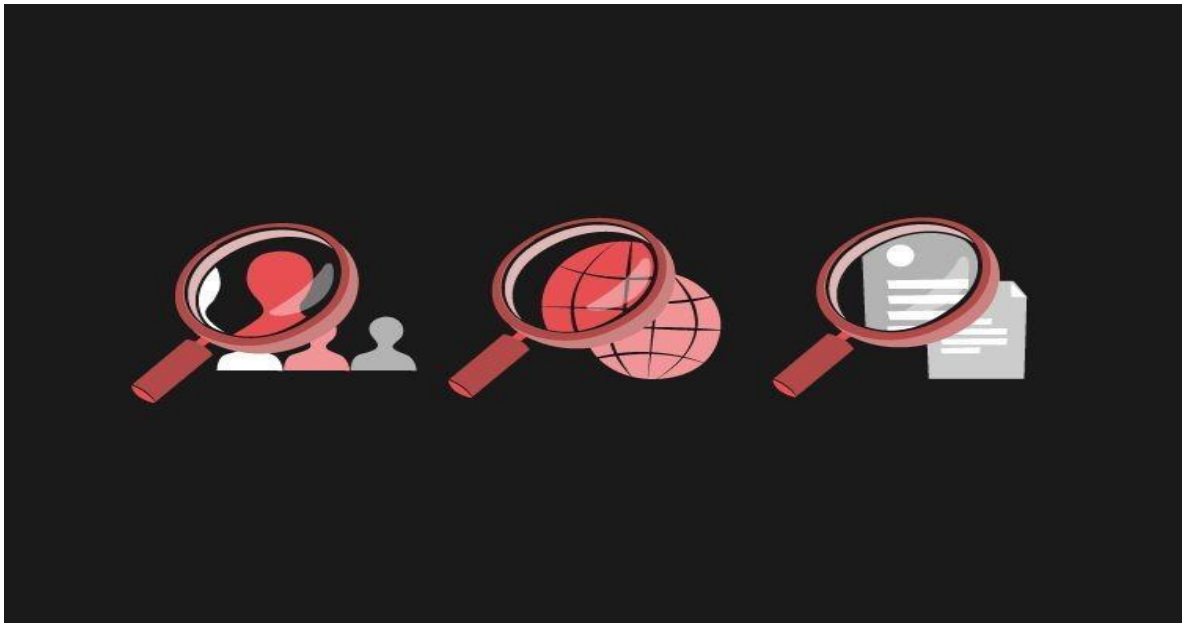
How to use OCCRP

Surafel Demssie, Rakeb Teklehiwot, Anam Yasin, Md Al Amin Iqbal, ChiaWen Chen and Nelson Reyes

Table of contents

How to use OCCRP

Surafel Demssie, Rakeb Teklehiwot, Anam Yasin, Md Al Amin Iqbal, ChiaWen Chen and Nelson Reyes



Datasets

Explore, analyze, and share quality data

Introduction

The Organized Crime and Corruption Reporting Project (OCCRP) is a consortium of investigative centers, media and journalists. OCCRP is the only full time investigative reporting organization that specializes in organized crime and corruption. The organization publishes its stories through local media and on their website, mainly in English and Russian languages.

In this user guide, we will outline and explain stepwise how to go about handling and transforming data in OCCRP website to give the end user insights of what the data is all about.

ACCESS LINK
https://aleph.occrp.org/

How to use OCCRP

Surafel Demssie, Rakeb Teklehiwot, Anam Yasin, Md Al Amin Iqbal, ChiaWen Chen and Nelson Reyes

Types of Datasets

OCCRP supports a variety of dataset publication formats such as CSV files, excel files, PDF files, HTML files, JSON files and many more. Not only are they open, accessible data formats better supported on the platform, they are also easier to work with for more people regardless of their tools.

Supported File Types

- CSV Files
    - The simplest and best-supported file type available on OCCRP is the "Comma-Separated List", or CSV, for tabular data. Each line of the file is a data record and each record consists of one or more fields, separated by commas. A CSV file typically stores tabular data in plain text, each line will have the same number of fields.
- Excel Files
    - Excel is developed by Microsoft. It features calculation, graphing tools and programming language. Spreadsheets is a way to organize data into rows and columns to make it simpler to read and manipulate.
- HTML Files

○ HTML also known as Hypertext Markup Language is designed to be displayed in a web browser. HTML can embed programs written in a scripting language such as JavaScript. HTML elements are the building blocks of HTML pages. HTML provides a means to create structured documents by denoting structural semantics for text such as headings, paragraphs, quotes etc.

How to use OCCRP

Surafel Demssie, Rakeb Teklehiwot, Anam Yasin, Md Al Amin Iqbal, ChiaWen Chen and Nelson Reyes

Filters

Once the user has access to the website and has searched for a topic, the website will provide a page with tabs on the left-hand side with datasets on the right side. The tabs on the left side includes datasets, dates, types, countries, language, emails, phone numbers, names, addresses and file types. The tabs are the filters for the website. In the datasets, there's a list of datasets where a user can click on to view the document. In the Types section, a user can find things like people, companies, contracts, organizations, images, licenses, packages, real estates etc. All of these types will help the user filter into what they are looking for. In the Country's tab, users can find different countries such as Russia, United Kingdom, France, Germany, Slovakia, Italy, Czech Republic etc. Users will be able to find datasets from countries. Having languages as a filter will provide users with different languages such as Spanish, English, Italian, Dutch etc. File types section includes files such as application/pdf, text/html, text/plain, image/jpeg, image/png etc. All of these are lists of filters for users so they can narrow down their search for a dataset.

Searching for Datasets

OCCRP is run by using the Aleph software built by Friedrich Lindenberg. Aleph allows user indexing of large amounts of both text (PDF, Word, HTML) and tabular (CSV, XLS, SQL) data for easy browsing and search.

Aleph has advanced search operators that can be used to find search matches efficiently.

To find exact matches for a given search term in Aleph, e.g. to search for a person or company, try putting the name in quotes.

Proximity search in aleph can be used when users do not want to find a precise string but want to merely specify that two words are supposed to appear close to each other. This will try to find all the requested search terms within a given distance from each other. An example of this search would be: "Bank America"~2

How to use OCCRP

Surafel Demssie, Rakeb Teklehiwot, Anam Yasin, Md Al Amin Iqbal, ChiaWen Chen and Nelson Reyes

In Aleph, users can make sure that a given term must show up in the results or specify terms to never show up. Users can put a plus sign ("+") in front of the term they want to show up and a minus sign, "-", to make sure all documents with the given word are removed.

Aleph allows the user to filter search results by sources, document type, as well as emails, phone numbers, addresses, entity names, countries and more on its left-hand column, after the user runs his/her search.

Users can explore structured data in new ways as OCCRP Data uses entity extraction on documents and emails to find phone numbers, names of people and companies, addresses, ID numbers and other key linkage details of interest. Click on an entity and see the “Tags” option in the preview screen.

OCCRP Data can cross-reference the information on two lists; it also ranks data that closely matches and lets the user compare the information. This can be done by Clicking on a source and then click on the “Cross Reference” option to choose another source with which to do the comparison.

OCCRP Data also has alerts feature that allows the user to monitor a search term so when new information is added to the database the user will receive a notification. All the user needs to do is switch on the bell icon right next to his/her search query.

OCCRP Data also supports content in foreign languages. The interface is translated and supports Russian and Bosnian-Serbo-Croatian. Search results on the database can also be filtered by language.

Navigating the Dataset Interface

When a user is looking for a dataset to use, he/she has the options to filter datasets based on Categories, Countries, Language, Email address, Name, Addresses, and File type. Although these filtering tools are extremely useful in navigating the website, due to the vast number of files, the website holds these filters aren’t enough in providing a smooth navigation. With some links, the

website provides the user with previews of the files and a way for the content to be downloaded; however, some links may lead to a blank white page.

Exporting Data

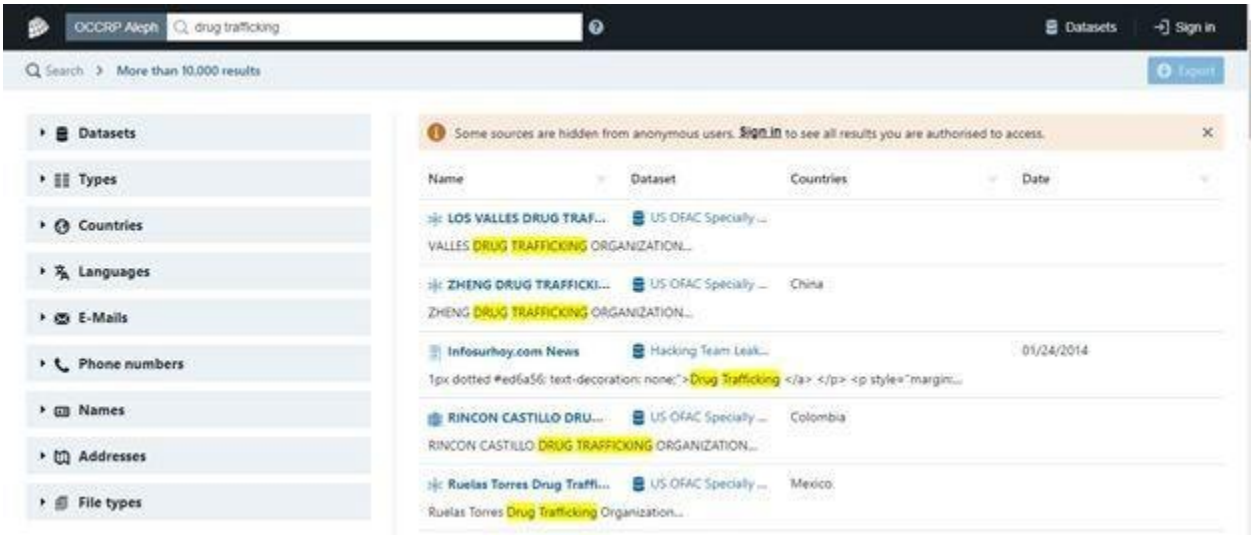
One of the great ways to export data in OCCRP is to categorize your preference regarding file type, language, country and then pressing the export button at the top of the screen to be able to download a zip file containing information related to your search. Follow the image for better understanding.

Step 1: Search for a term in this case I am searching for “drug trafficking”:

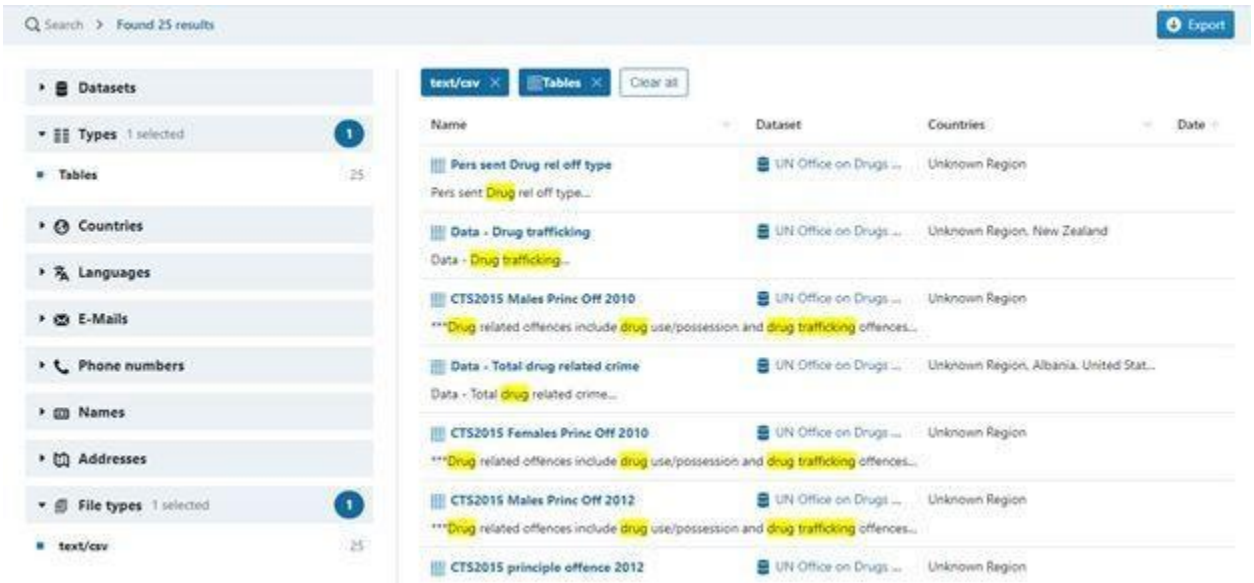
Step 2: After pressing enter you will be taken into the result page related to your search term____ similar to the image below. On the left side of the page there are 10 filter options to categorize your data from. You can select the file types, language etc. by clicking on the filter you want to categorize from.

How to use OCCRP

Surafel Demssie, Rakeb Teklehiwot, Anam Yasin, Md Al Amin Iqbal, ChiaWen Chen and Nelson Reyes



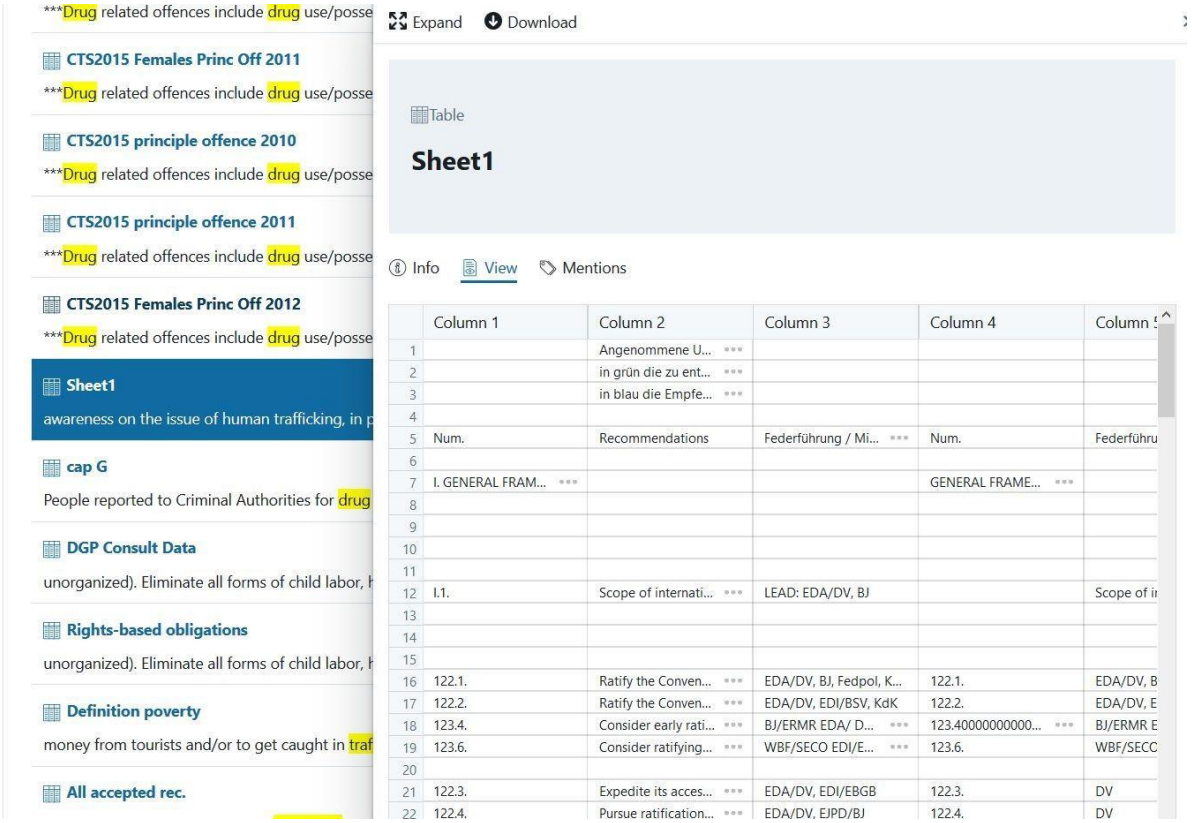
Step 3: After categorizing your preference, in my case I chose table as types and CSV as file types. And then by pressing the export button at the right side of the page will allow you to download all the files that match the preference.



Step 4: When you click on the small table figure, there will be a popup that will show a preview of the content.

How to use OCCRP

Surafel Demssie, Rakeb Teklehiwot, Anam Yasin, Md Al Amin Iqbal, ChiaWen Chen and Nelson Reyes

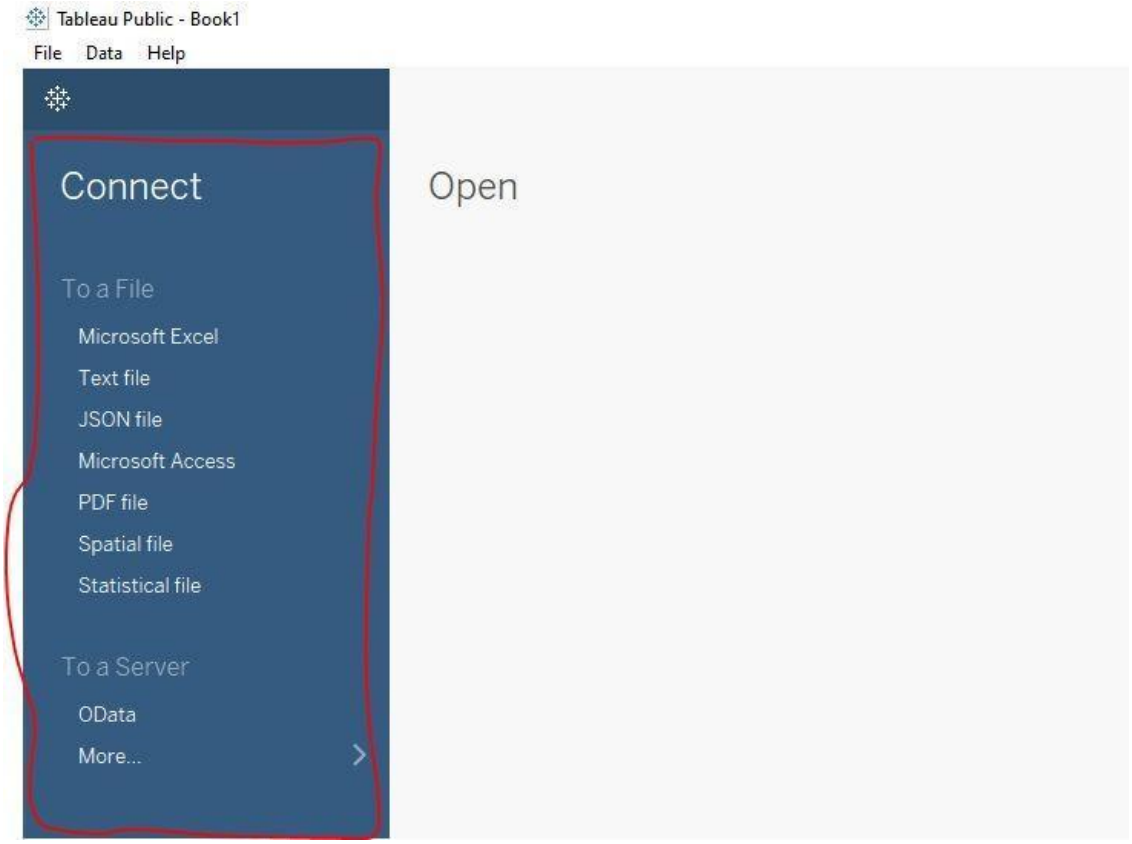


Creating A Visualization Using The Data

OCCRP data can be visualized in the form of graphs and charts to allow for easy interpretation of data. Users can use various visualization tools such as tableau, R and excel. Tableau graphs or visualizations could help users to understand what it is the content of the website as a whole and as deep as a CSV file.

To create visualizations, you need to export data from the OCCRP website by first categorizing preference regarding file type, language, country and then press the export button at the top of the screen to be able to download a zip file containing information related to your search. Once you download your data and put it into file types such as csv or excel(as explained above on how

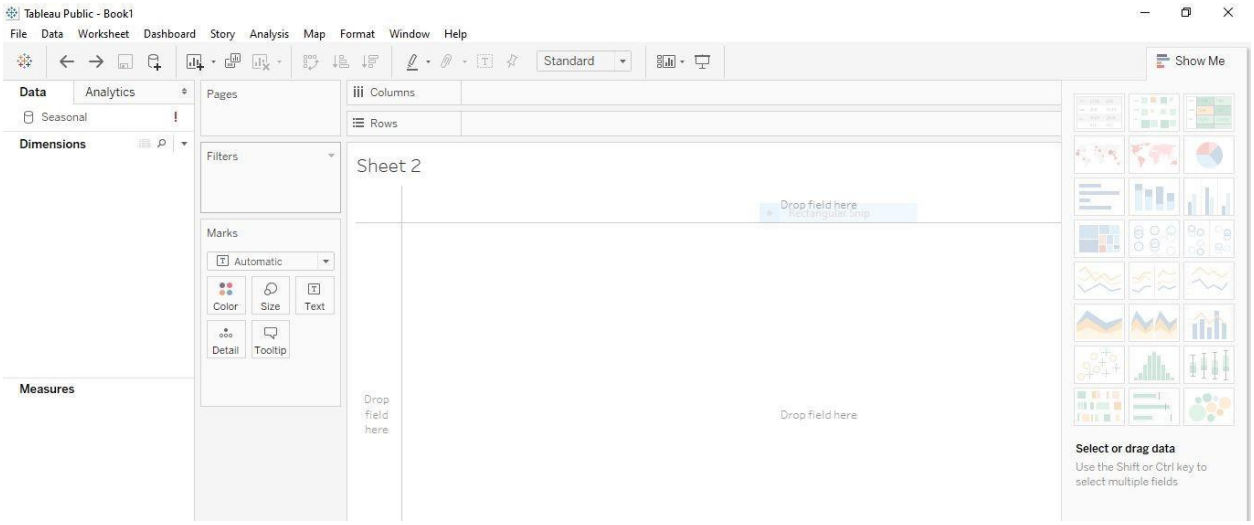
to export data), you open the visualization tool such as Tableau, and R. In Tableau for instance, when you open it, you need to choose the file type of the data you want to import and connect it with the Tableau as shown below.



Upon connecting, drag the specific data you want to visualize in the rows section and choose the type of visual you need to create i.e graphs, maps or charts by clicking on the top right side of the worksheet. You can build visualizations by adding data elements such as measure, text, or attribute from the ‘marks’ section and work with color to make visualizations more attractive, dynamic, and informative as shown below.

How to use OCCRP

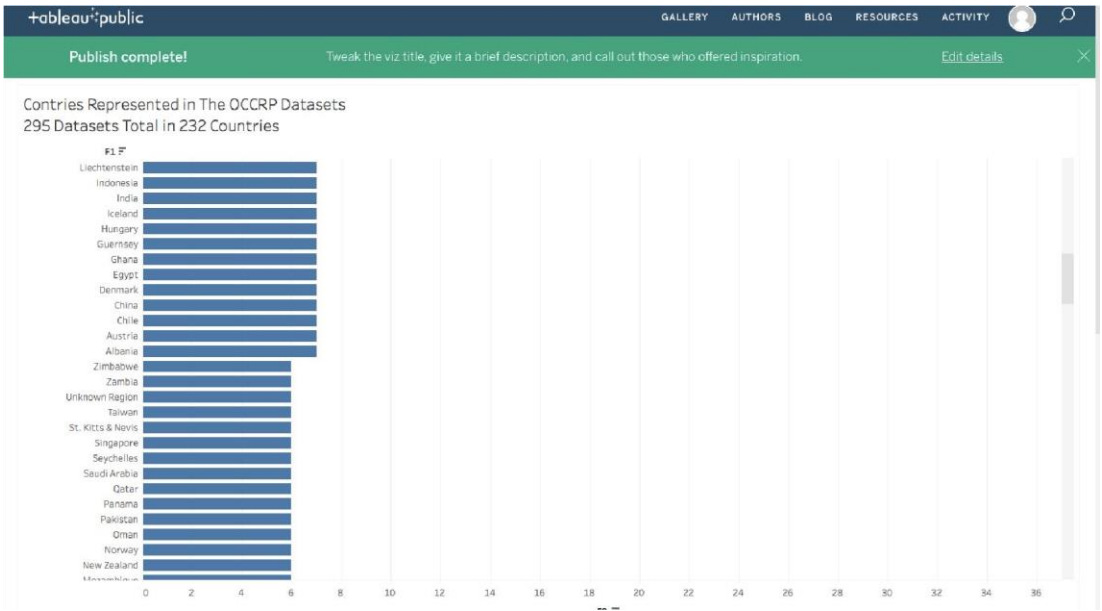
Surafel Demssie, Rakeb Teklehiwot, Anam Yasin, Md Al Amin Iqbal, ChiaWen Chen and Nelson Reyes



You can change visualization types to best suit the data you're exploring. As you build the visualization, you add as many data elements as required and thereafter create a dashboard to explore and analyze your data of your interest in the form of graphs, maps and charts.

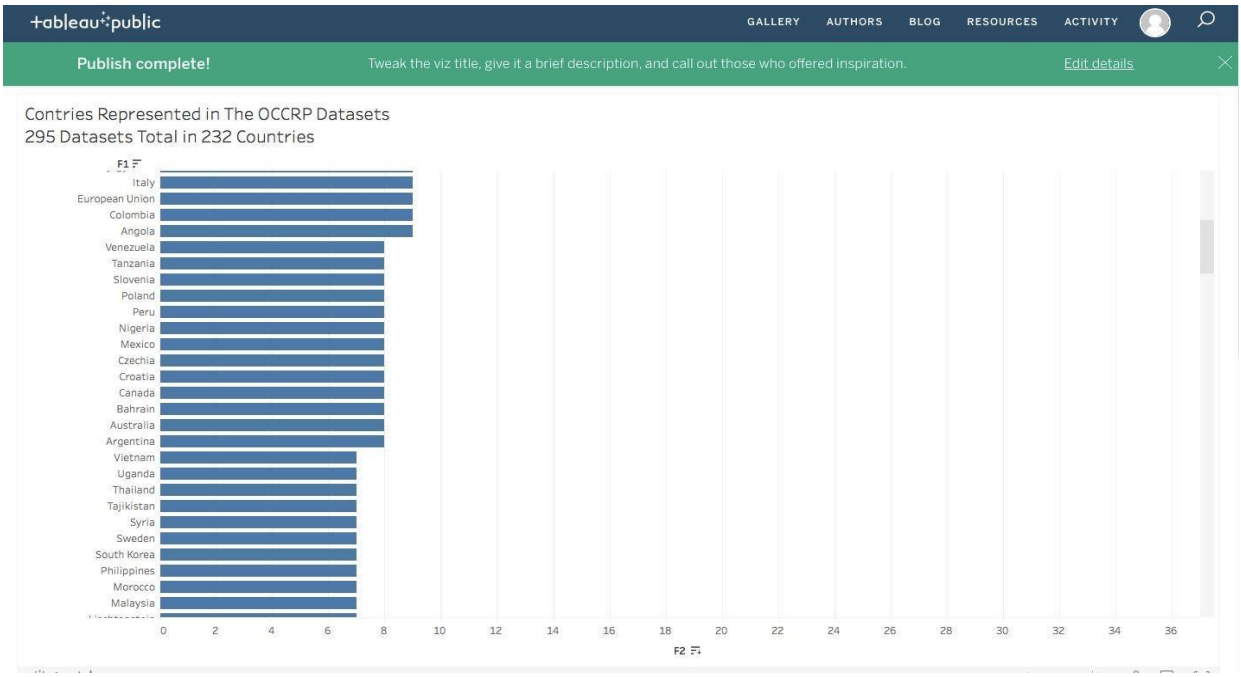
- ❖ Below are some of the visualizations created from data obtained from OCCRP.

Figure 1-An overview of the whole website including 295 datasets in total in 232 countries in



How to use OCCRP

Surafel Demssie, Rakeb Teklehiwot, Anam Yasin, Md Al Amin Iqbal, ChiaWen Chen and Nelson Reyes



total. This is a bar graph that shows an overview as a whole of the OCCRP website.

Figure 2-An overview of the whole website including 295 datasets in total in 232 countries in total.

Figure 3-An overview of the whole website including 295 datasets in total in 232 countries in total. This bar graph is an overview as a whole of the OCCRP website.

Figure 4-An overview of the whole website including 295 datasets in total in 232 countries.

How to use OCCRP

Surafel Demssie, Rakeb Teklehiwot, Anam Yasin, Md Al Amin Iqbal, ChiaWen Chen and Nelson Reyes

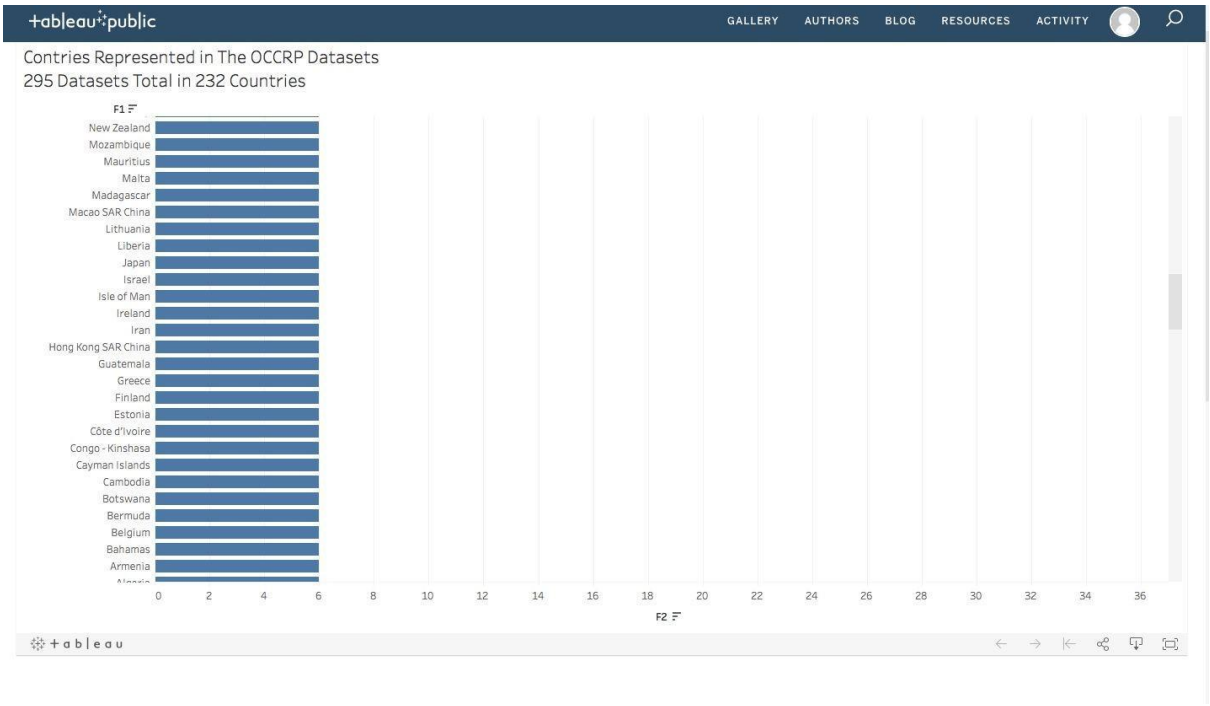


Figure 5 - An overview of the whole website including 295 datasets in total in 232 countries in total.

How to use OCCRP

Surafel Demssie, Rakeb Teklehiwot, Anam Yasin, Md Al Amin Iqbal, ChiaWen Chen and Nelson Reyes

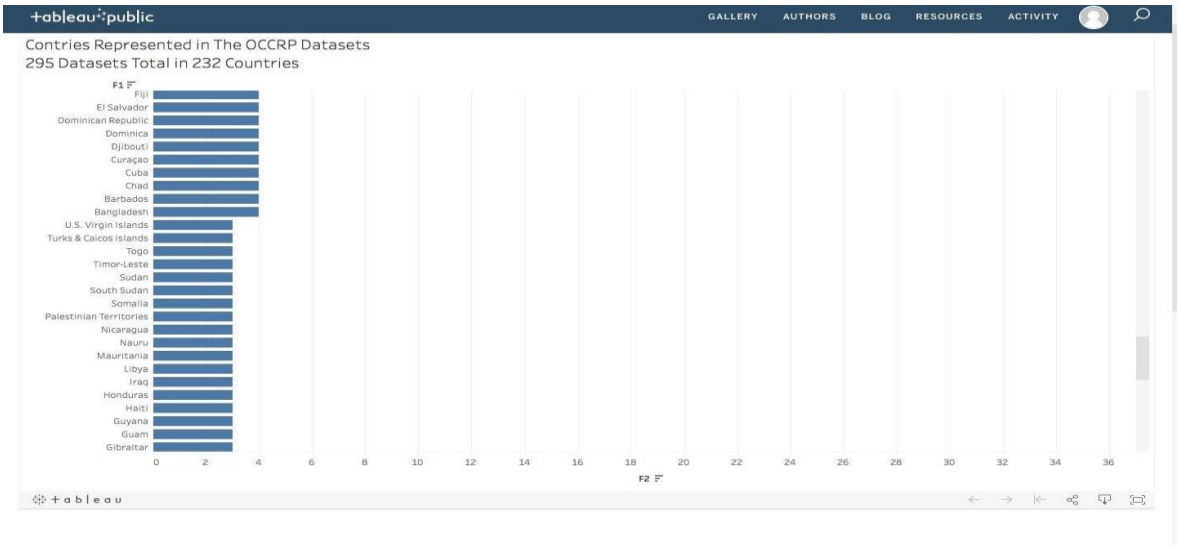


Figure 6- An overview of the whole website including 295 datasets in total in 232 countries in total.

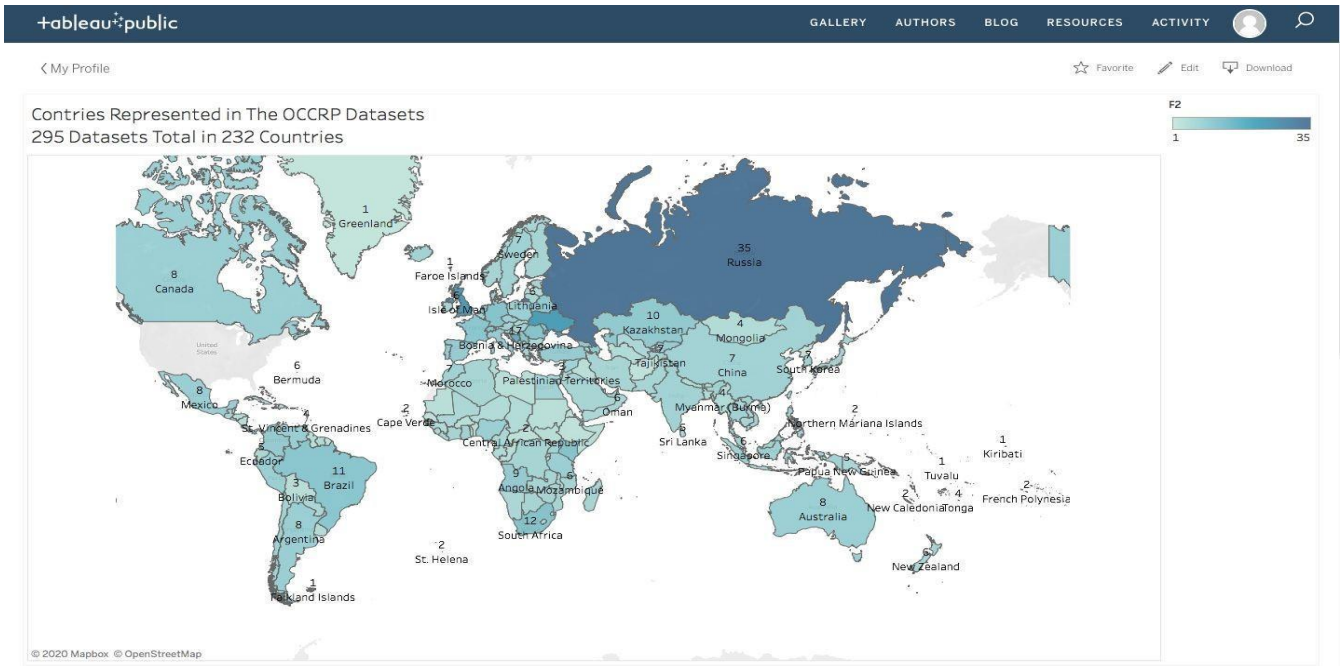


Figure 7-This graph shows the number of dataset by categories for the OCCPR Website.

How to use OCCRP

Surafel Demssie, Rakeb Teklehiwot, Anam Yasin, Md Al Amin Iqbal, ChiaWen Chen and Nelson Reyes

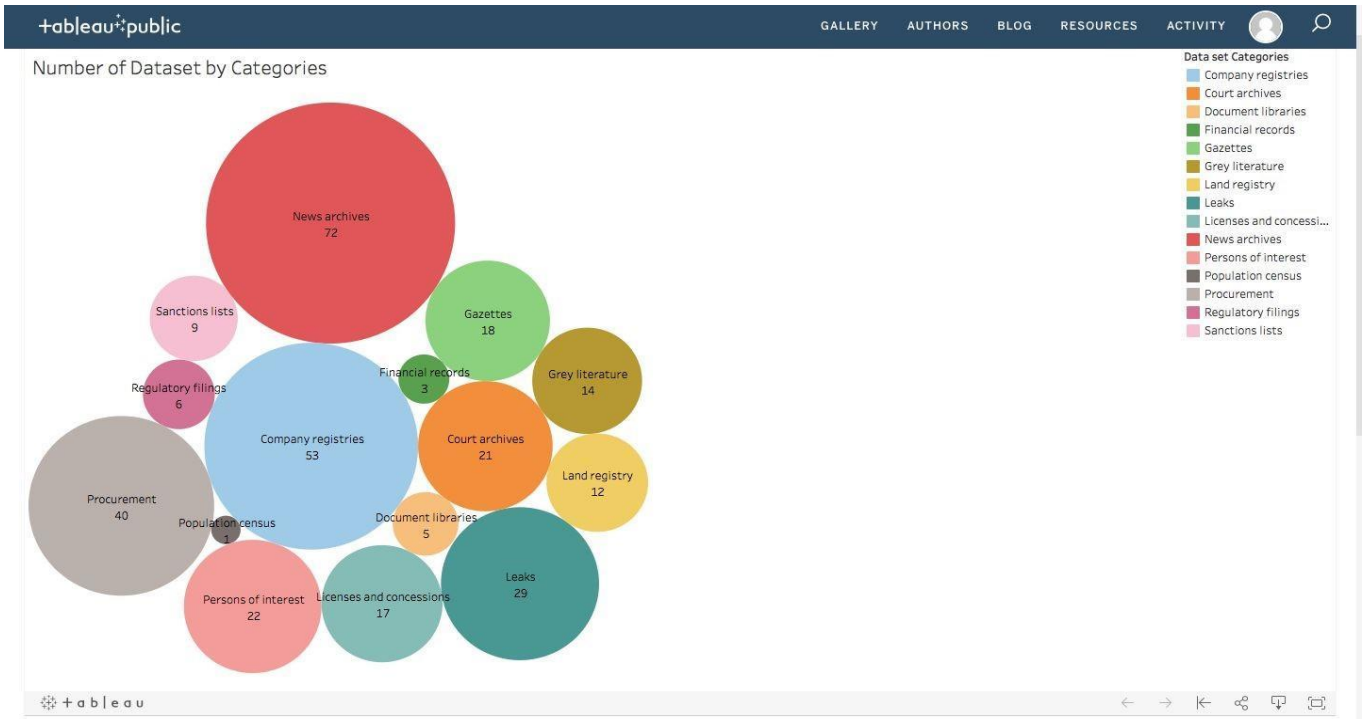


Figure 8- An overview of the whole website including 295 datasets in total in 232 countries in total using a MAP on Tableau.

How to use OCCRP

Surafel Demssie, Rakeb Teklehiwot, Anam Yasin, Md Al Amin Iqbal, ChiaWen Chen and Nelson Reyes

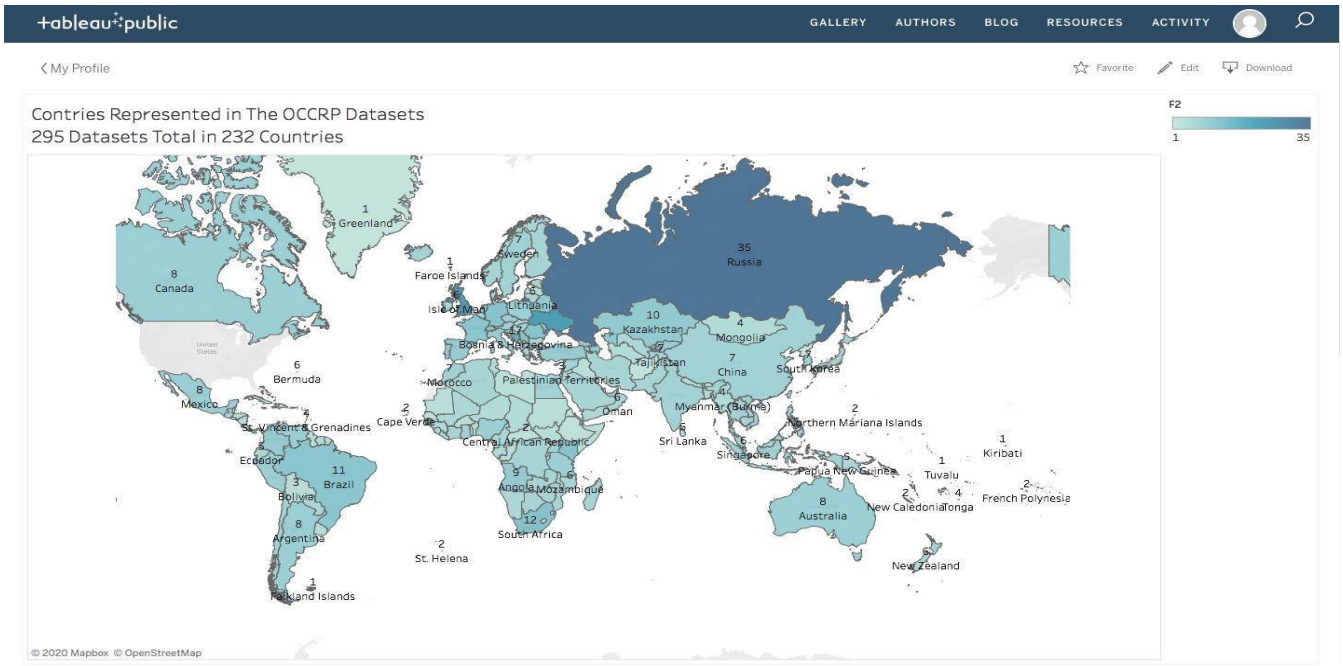


Figure 9 - An overall view of the whole website of the number of dataset by the categories.

How to use OCCRP

Surafel Demssie, Rakeb Teklehiwot, Anam Yasin, Md Al Amin Iqbal, ChiaWen Chen and Nelson Reyes

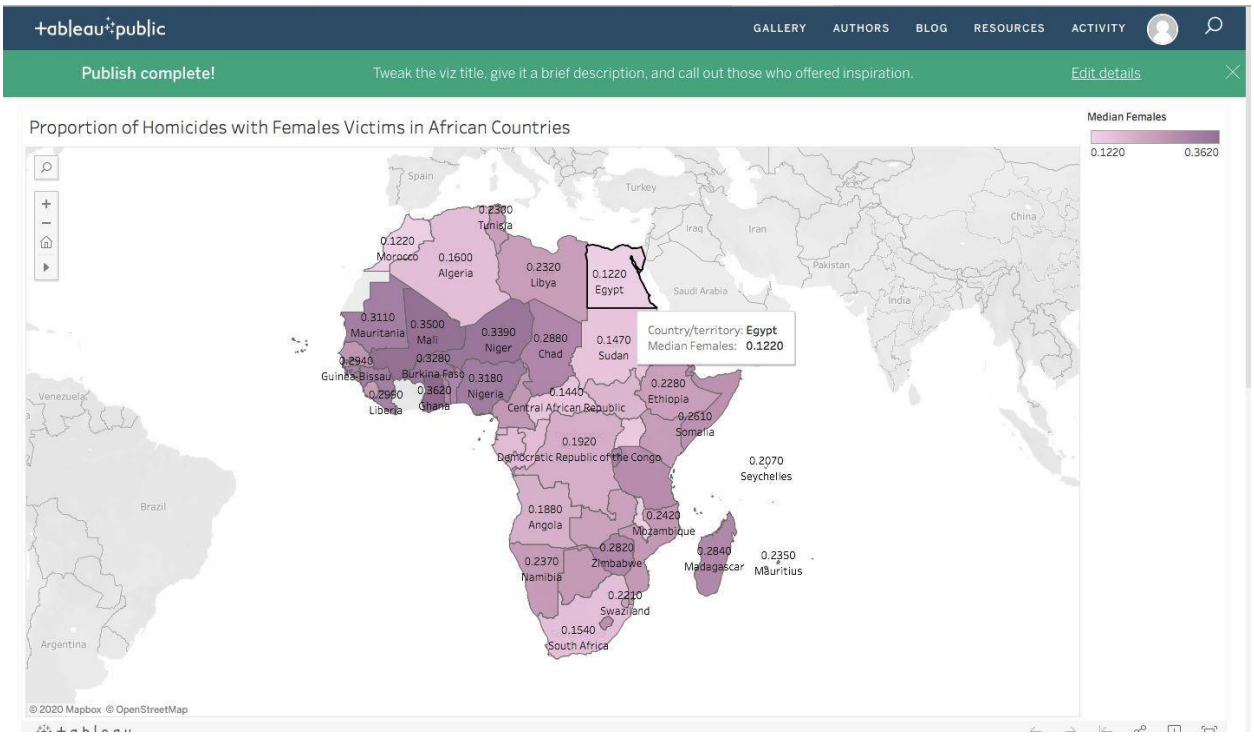


Figure 10-This is the proportion of homicides with males victims in African Countries

How to use OCCRP

Surafel Demssie, Rakeb Teklehiwot, Anam Yasin, Md Al Amin Iqbal, ChiaWen Chen and Nelson Reyes

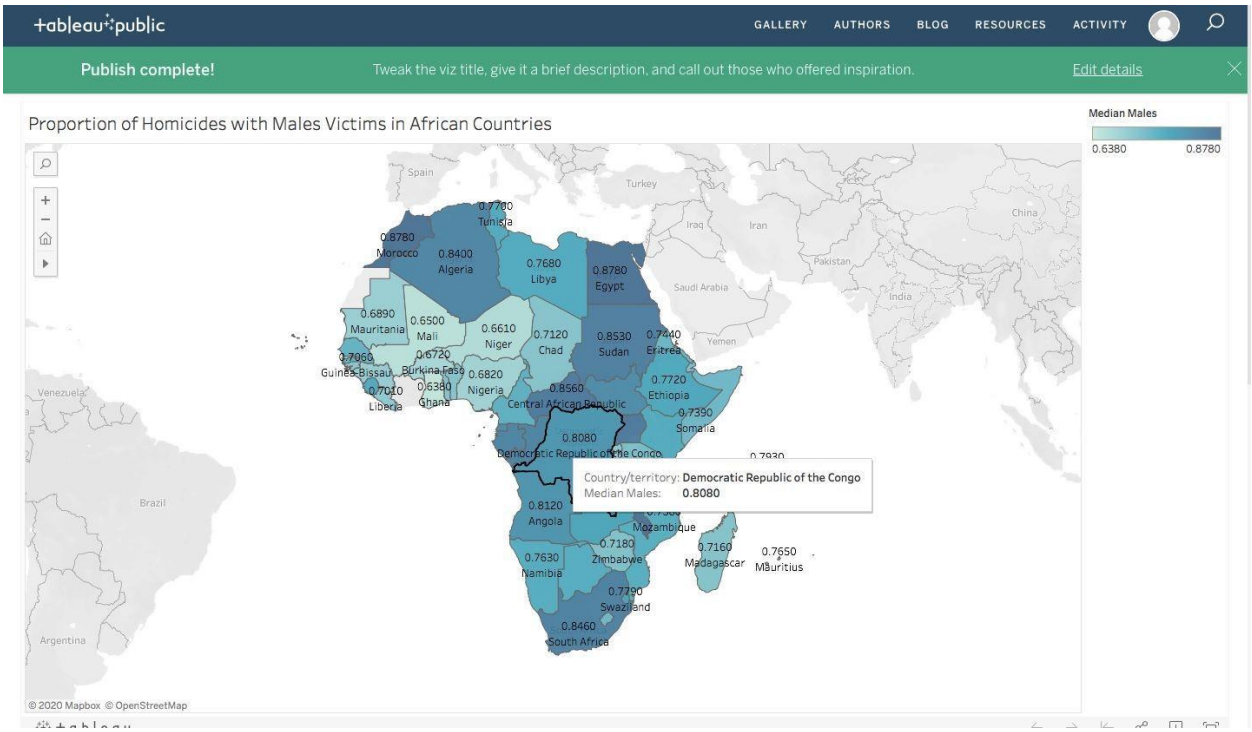
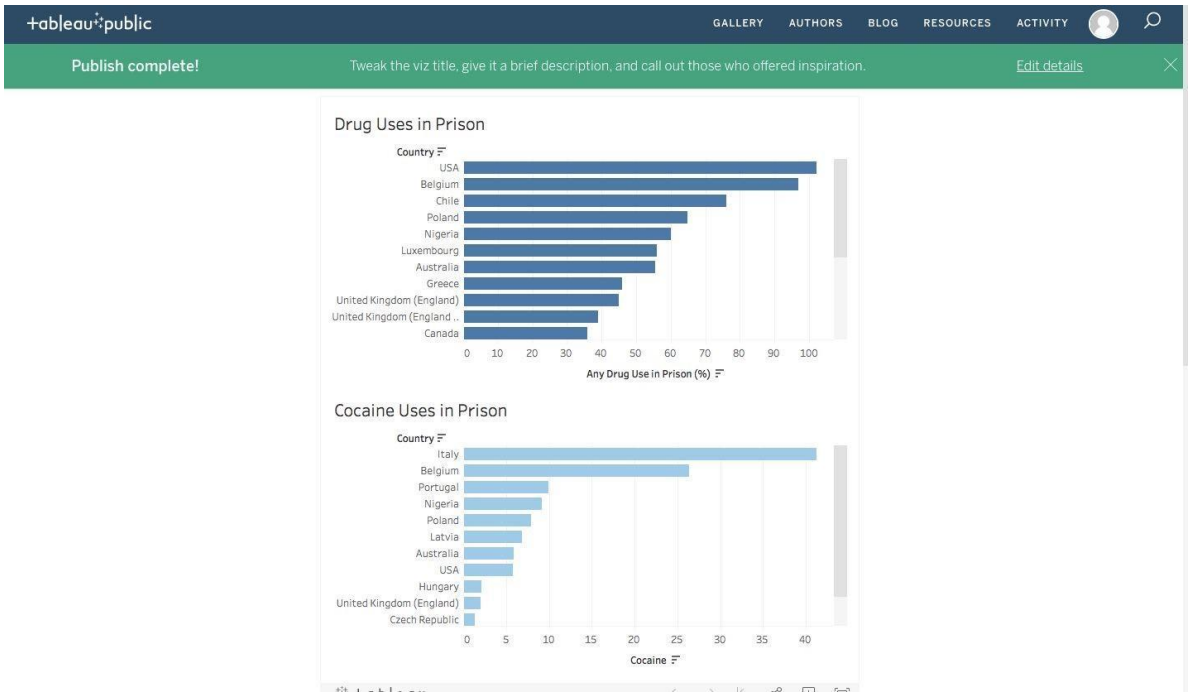


Figure 11- These two graphs show Drug and cocaine uses in prison from Two different CSV files.

How to use OCCRP

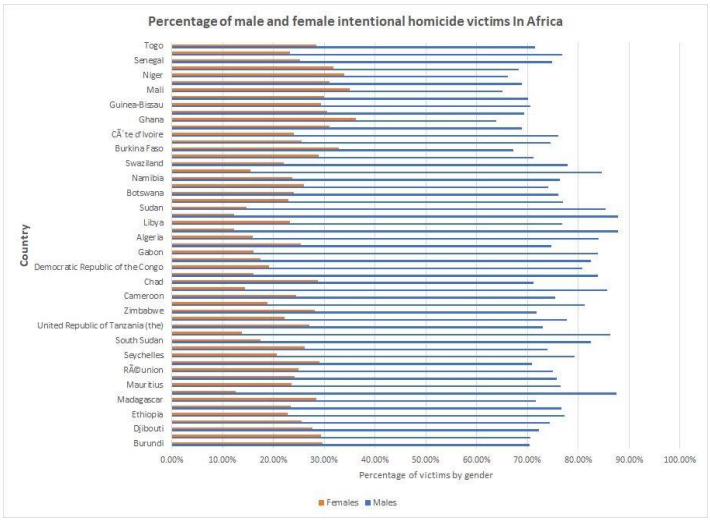
Surafel Demssie, Rakeb Teklehiwot, Anam Yasin, Md Al Amin Iqbal, ChiaWen Chen and Nelson Reyes

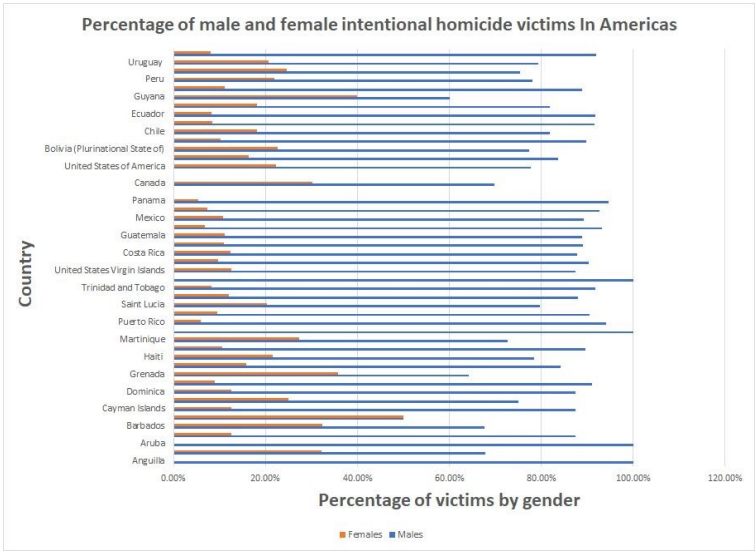


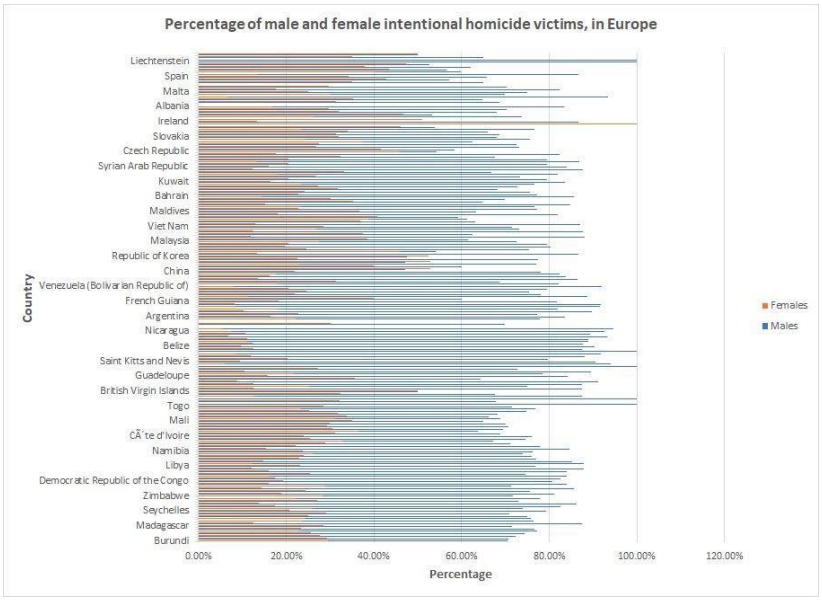
These three charts show murder victim rate by gender for many different countries in 2010 . Blue is for Male victims and orange is for Female victims. The chart on the left contains murder rate from African, the one on the right is from Americans and the one at the bottom is from Europe. Based on the charts some nations have a 100% murder rate for one gender. This could be due to lack of records, or the nation has a very small population. This also tells the user that some of the data might be questionable or may require additional investigation.

How to use OCCRP

Surafel Demssie, Rakeb Teklehiwot, Anam Yasin, Md Al Amin Iqbal, ChiaWen Chen and Nelson Reyes







Technical Specifications

- OCCRP requires users to create an account on their website in order to obtain all of the available data.

- Users must click the link as soon as the page loads otherwise the data will not pull up in a window.
- Data extracted from the OCCRP database is not well structured and therefore needs to be cleaned first in order to perform any statistical analysis.
- The language could be a barrier because files, emails, or the data could be in different Languages.

Limitations

One of the glitches that we found while downloading specific CSV files is that sometimes, the website will display the download button for 1-2 seconds in which time frame you must press the download button to successfully export the file. However, if you fail to click the download button, the whole website crashes and the screen becomes white with no content. Because of that the file or that particular data will not be accessible.

How to use OCCRP

Surafel Demssie, Rakeb Teklehiwot, Anam Yasin, Md Al Amin Iqbal, ChiaWen Chen and Nelson Reyes

Version History

Version	Month and Year
1.0	May 2022
2.0	November 2022
3.0	November 2024