CAIRO
UNIVERSITY

DATA SCIENCE PROJECT

TEAM MEMBERS

Donia Abdelfattah
Sec: 01 BN: 28

Raghad Khaled
Sec: 01 BN: 30

Menna Alla Ahmed
Sec: 02 BN: 29

Nada El-Sayed
Sec: 02 BN: 32

CONTRIBUTION

Donia Abdelfattah

Crawled 'lockbit'

Q4,Q6

Raghad Khaled

Crawled 'BianLian', 'Play News' websites

Q2,Q5

Menna Alla Ahmed

Crawled 'Royal', 'Vice Society' websites

Q1,Q7

Nada El-Sayed

Crawled 'black basta' websites

Cleaned the data

Q3

Our Customer

Our client, Saluslab, is a security company that has developed a tool to protect networks and companies from potential attackers.

SALUSLAB

What We Do

As part of our marketing strategy, we are considering various approaches, including gathering data on past attacks to identify patterns and predict the likelihood of future attacks. By leveraging this information, we can target companies that may benefit from Saluslab's security tool and offer them a solution to mitigate potential security risks.

Data analysis cycle and epicycle

- Stating the question:

The initial research question posed was how to enhance sales for the customer in the security services industry. To effectively address this question, it was deemed essential to identify the specific companies that are at a high risk of being attacked. Three key pieces of information pertaining to such companies were identified in the preliminary research - industry, region, and revenue. These were accordingly designated as the primary targets for data collection.

- Cleaning and Exploring data:

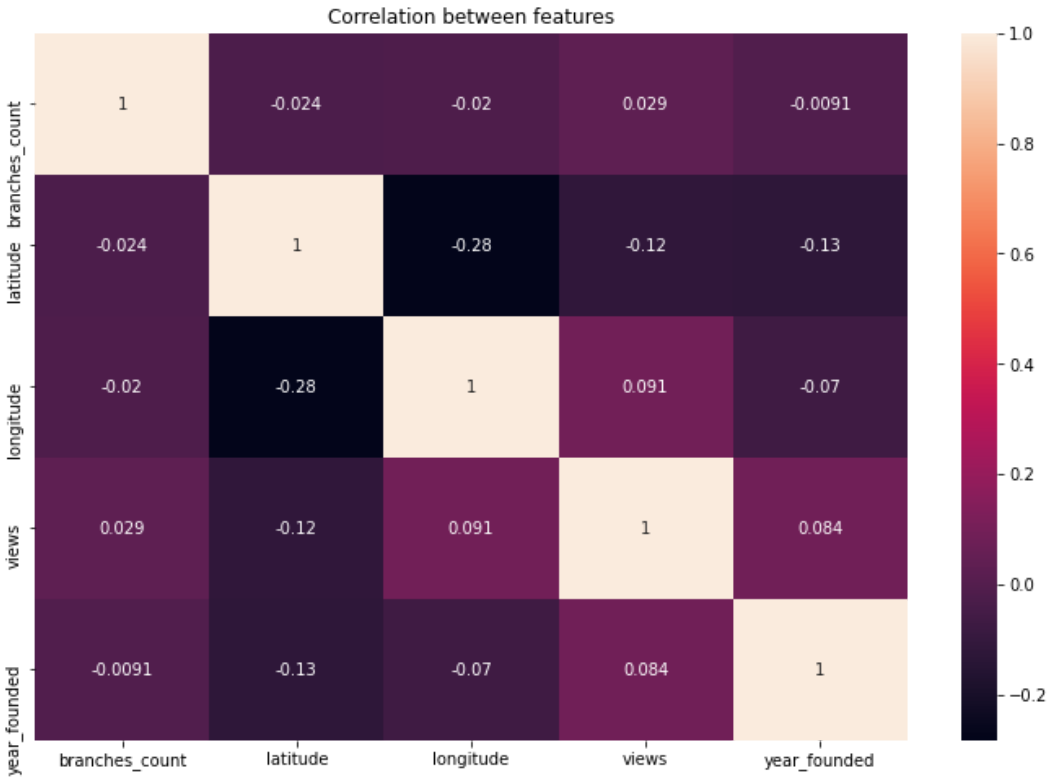
To obtain the names of the attacked companies, an exploration of gangs' websites on the deep web was conducted using Tor and Selenium. Data regarding these companies was subsequently gathered from a range of APIs (Yahoo Finance, ChatGPT, and Bing), with the ORB API eventually proving most effective. This API was utilized to generate a list of companies based on their names, along with associated information such as country of origin, year of foundation, industry, revenue range, and employee range.
The collected data was then subjected to a series of data cleaning processes, with employee ranges and revenue ranges obtained from both ORB and certain gang websites being formalized into columns. Chi-square tests were used to establish relationships between categorical columns, while correlation was utilized for numerical columns. In addition, information regarding the publishing of data (whether a company had refused to pay a ransom and hence the data was made public) was gathered and formalized accordingly.

- Building models:

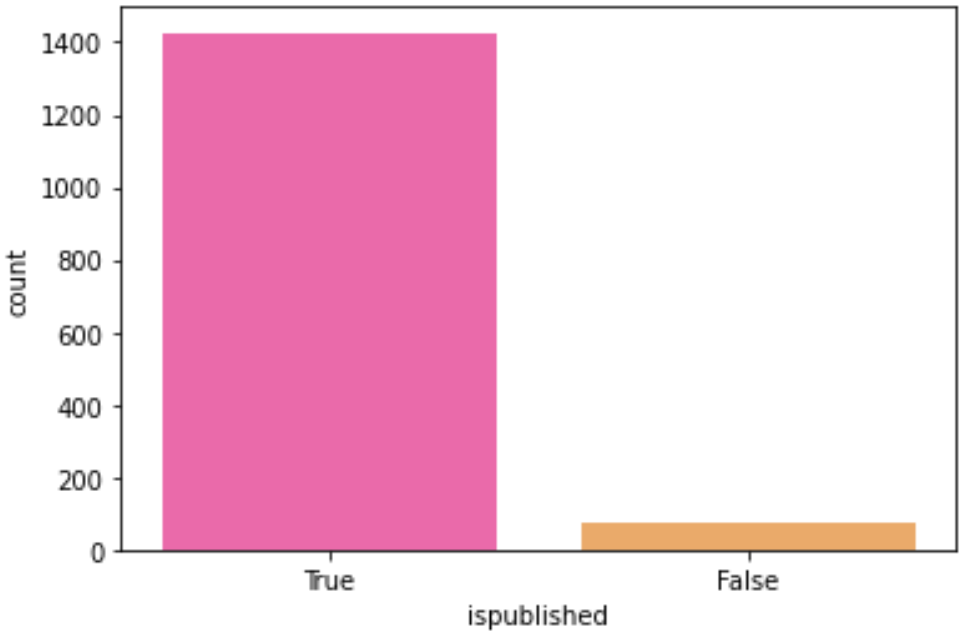
To address the initial research question of identifying the companies at high risk of being attacked, machine learning models were developed. The ORB API was utilized once again to identify unattacked companies which data distribution is similar to attacked companies, and the resultant data was used to train two models - a logistic regression model and a random forest model. The former used industry, region, revenue range, employee range, and year of foundation as predictors to estimate the probability of a given company being attacked. The latter was trained to predict the most likely gang to attack a company based on the source data, in addition to the aforementioned predictors.

Insights extracted throughout the project

- Interpreting the results:

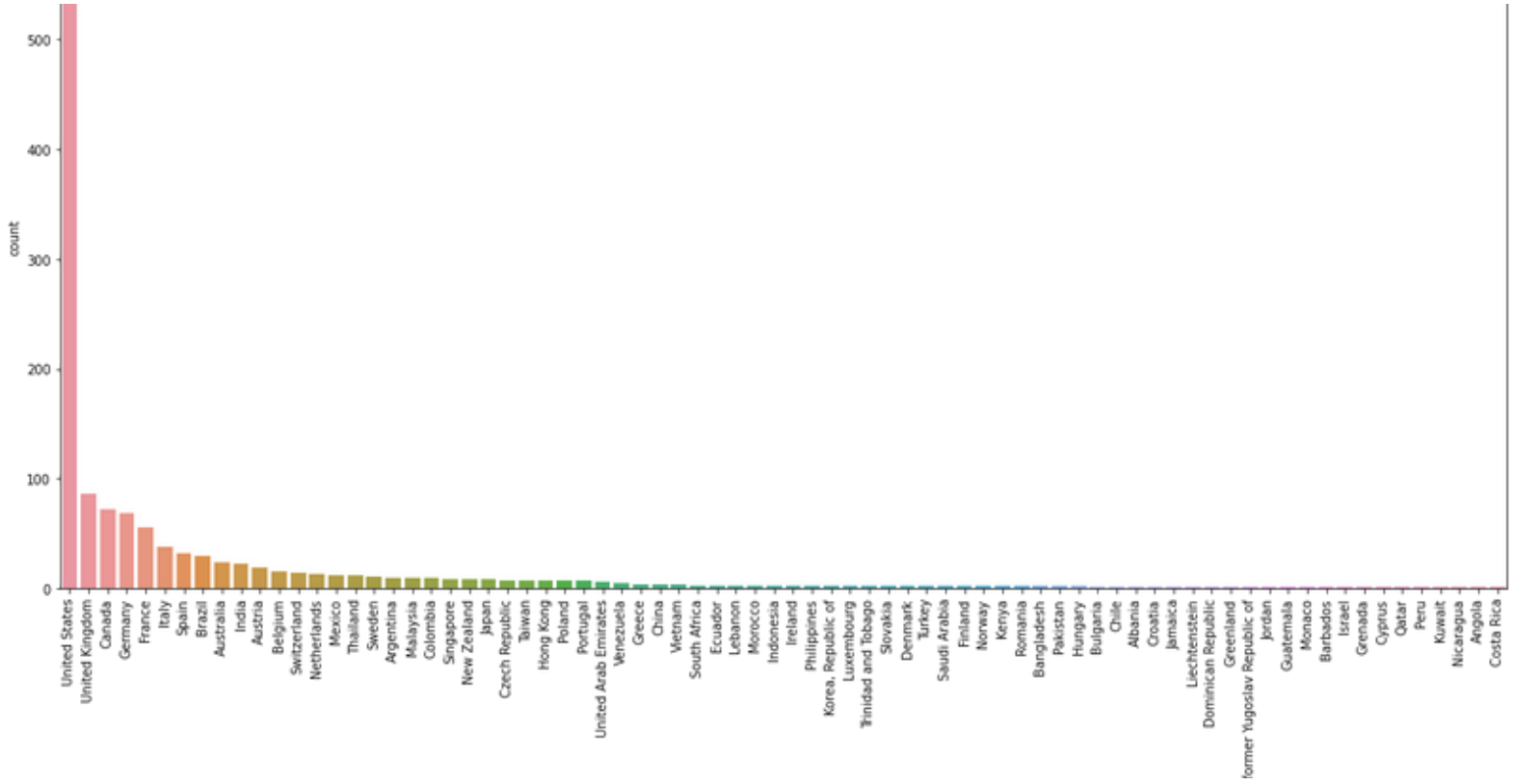


We are unable to identify a strong correlation between the numerical features.

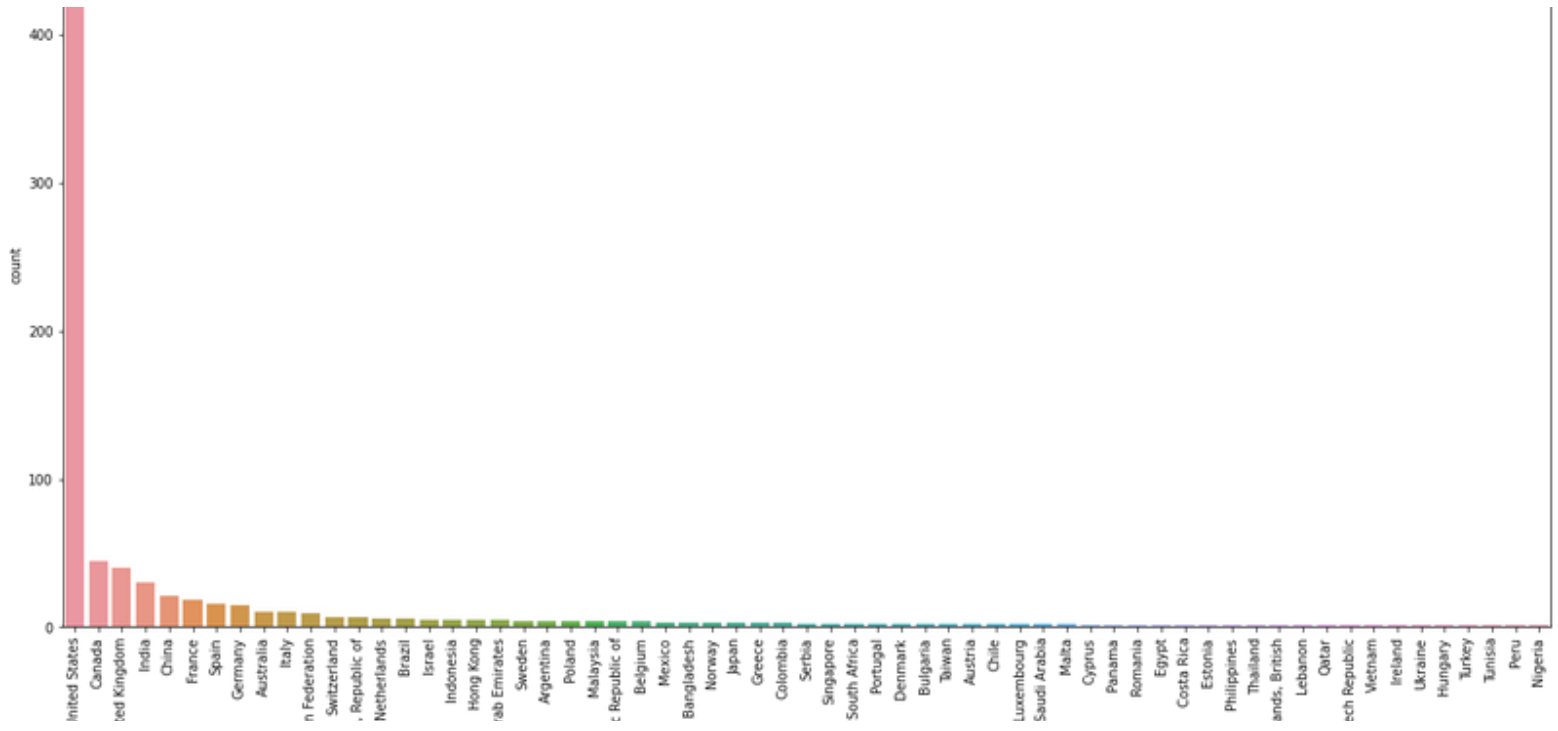


Companies that do not pay the ransom have their data published on the website. Therefore, we have more published data than unpublished data since the data is removed from the website when the company pays the ransom.

Insights extracted throughout the project

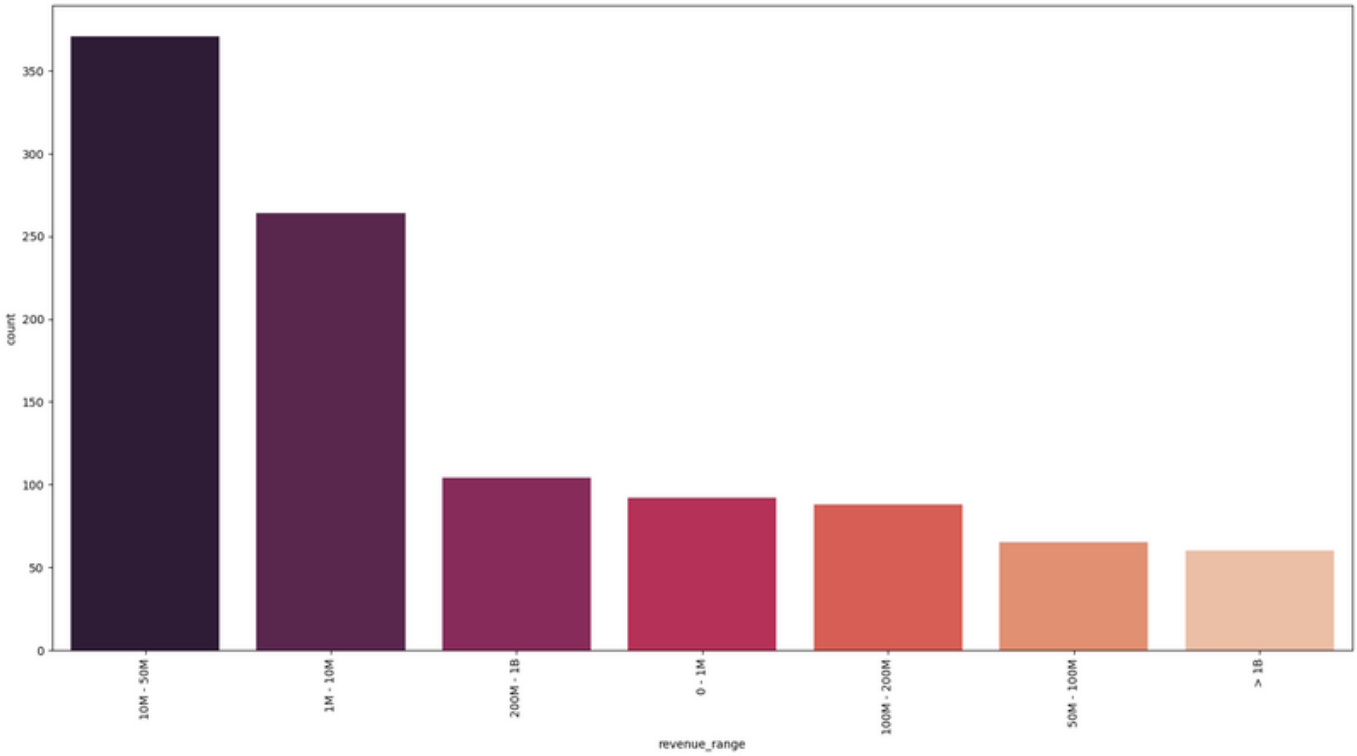


United states expriencies the largest number of attacked companies

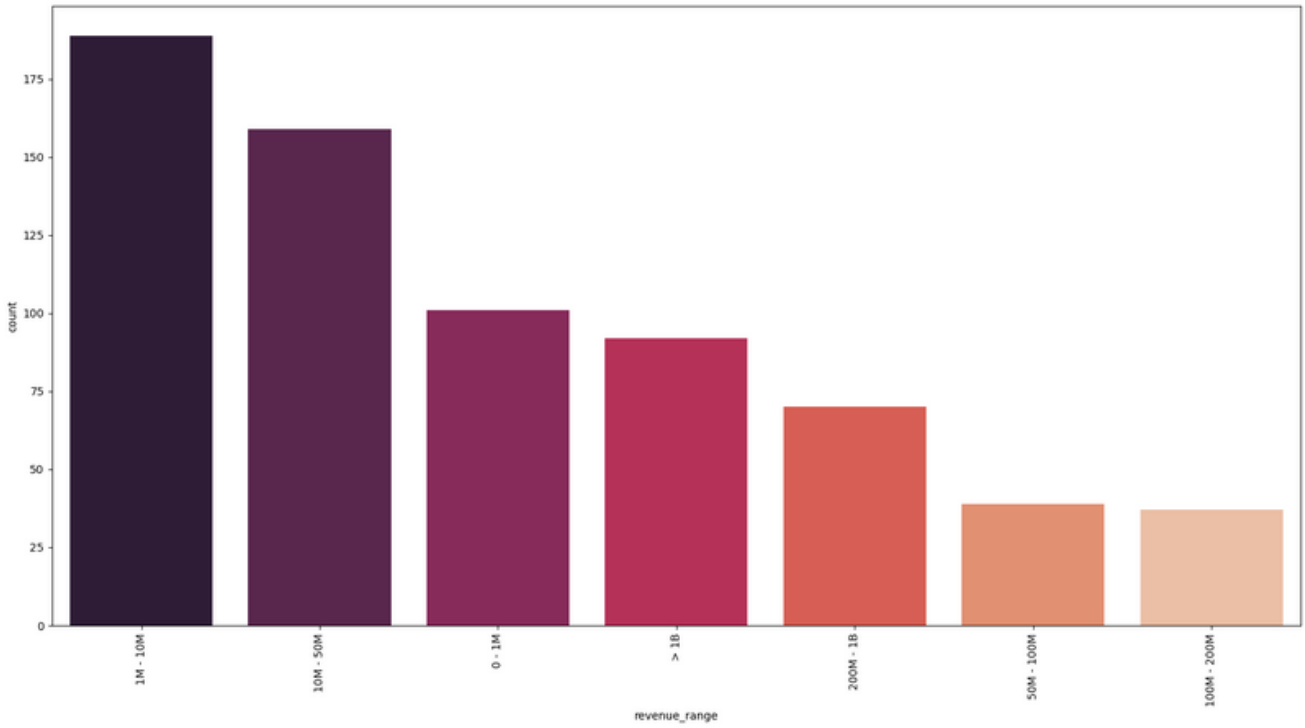


United states expriencies the largest number of un-attacked companies

Insights extracted throughout the project

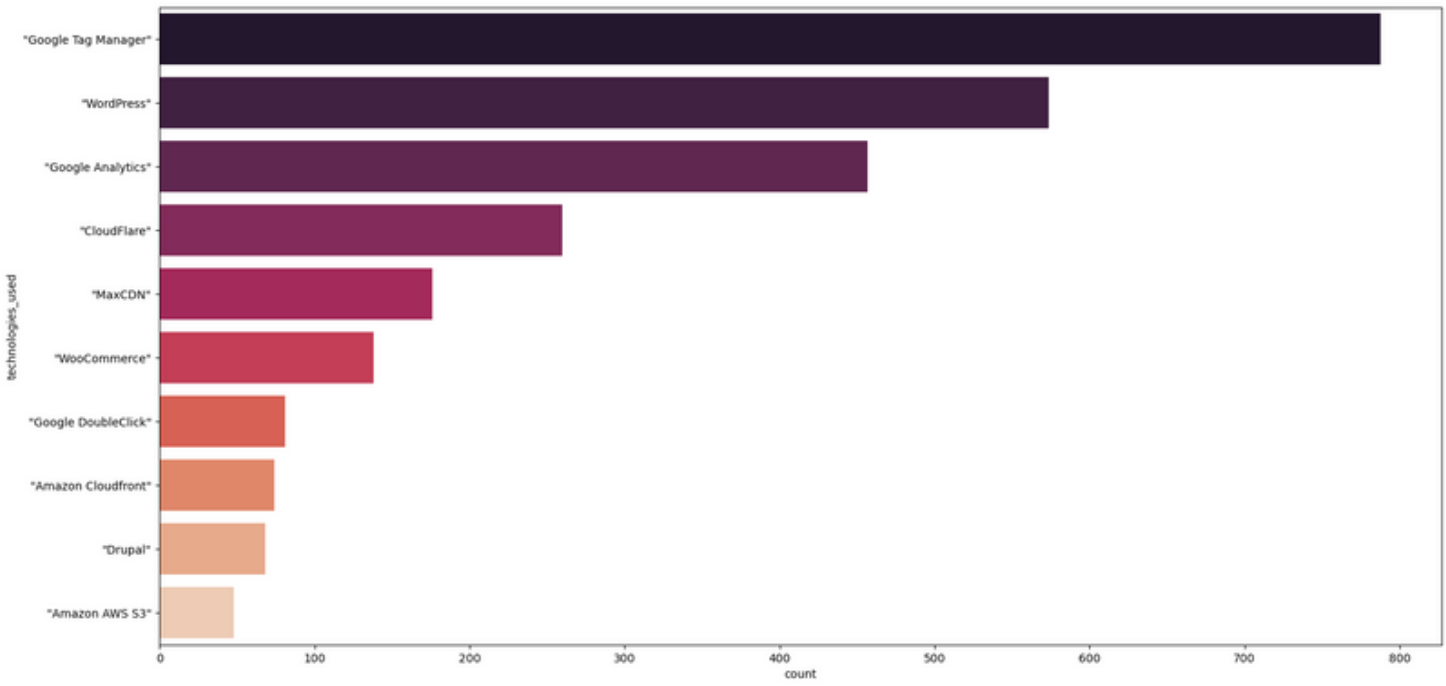


Cybercriminal gangs tend to target companies with a mean revenue between 10-50 million dollars our insight is that startups may be more vulnerable and unable to pay a ransom. Conversely, larger companies with a revenue of 1 billion dollars or more typically have stronger security measures in place.

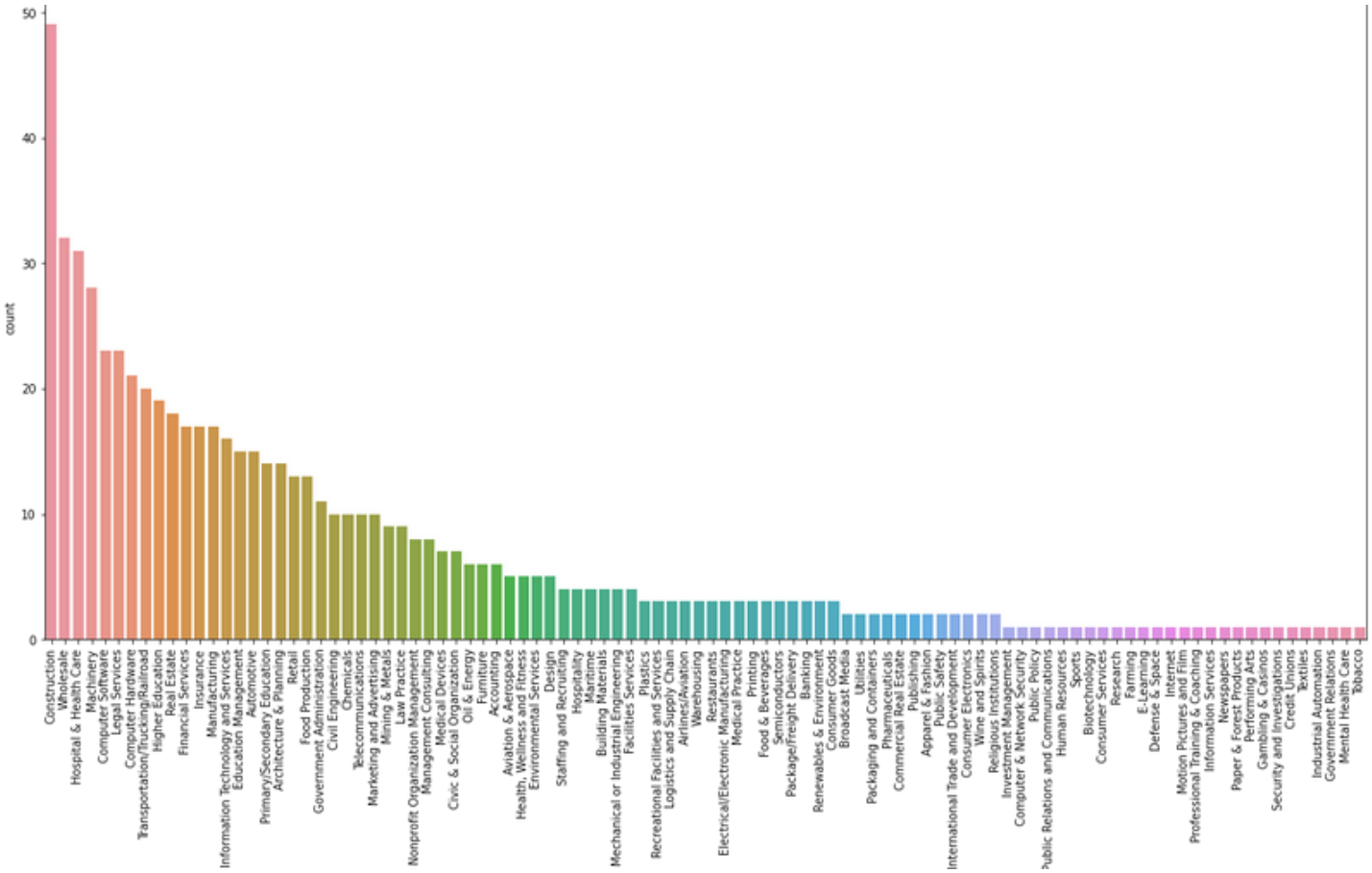


The distribution of the revenu of un-attacked companies.

Insights extracted throughout the project

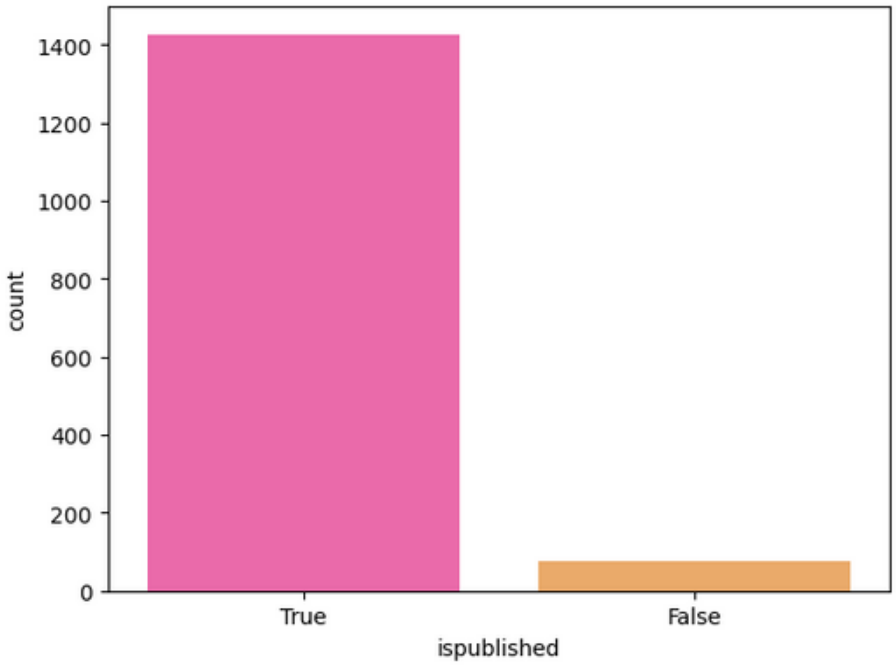


The types of technologies used attacked companies, it does not give any indication as most of the companies in the market use these technologies. and this is a web application and cloud technologies

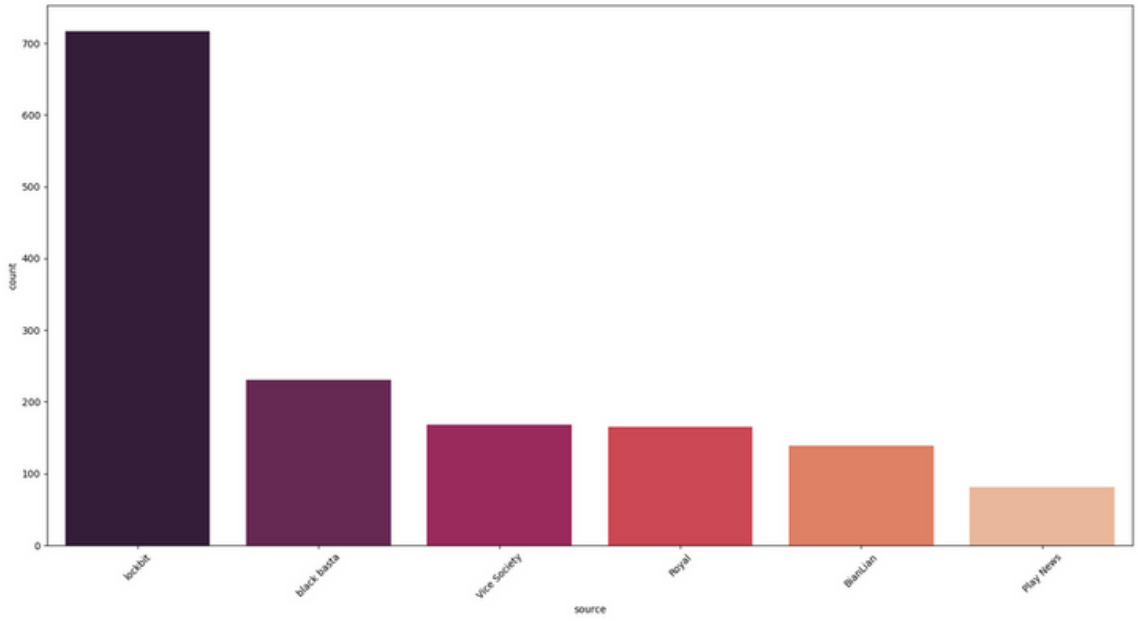


The construction sector has experienced the highest number of attacks. However, experts are surprised by this observation because construction typically does not deal with critical data that would companies have been afraid of the data being published.

Insights extracted throughout the project

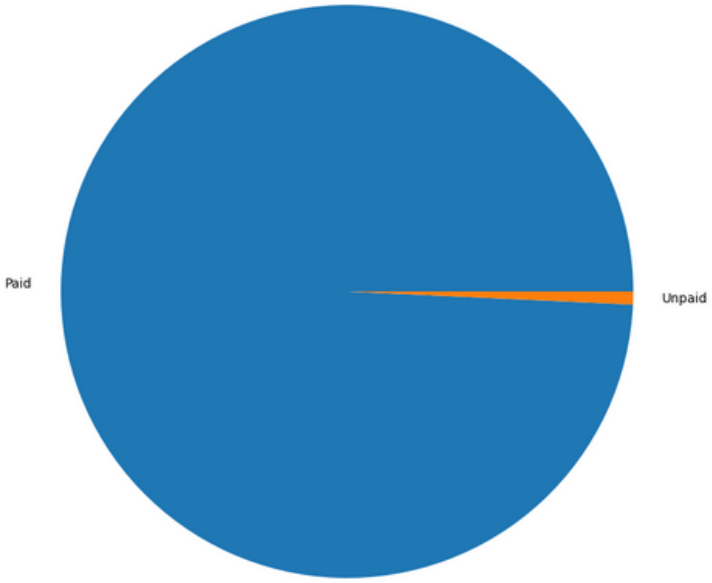


Upon revisiting the data, it became evident that the amount of information obtained by the crawler for published companies is significantly greater than that for unpublished companies. This can be explained by the fact that historical data for published companies is available on the website, whereas data for companies that have paid is removed. This could potentially account for the higher number of attacks targeting construction companies, as other critical sectors may be more likely to pay for their data and therefore not appear on the website.

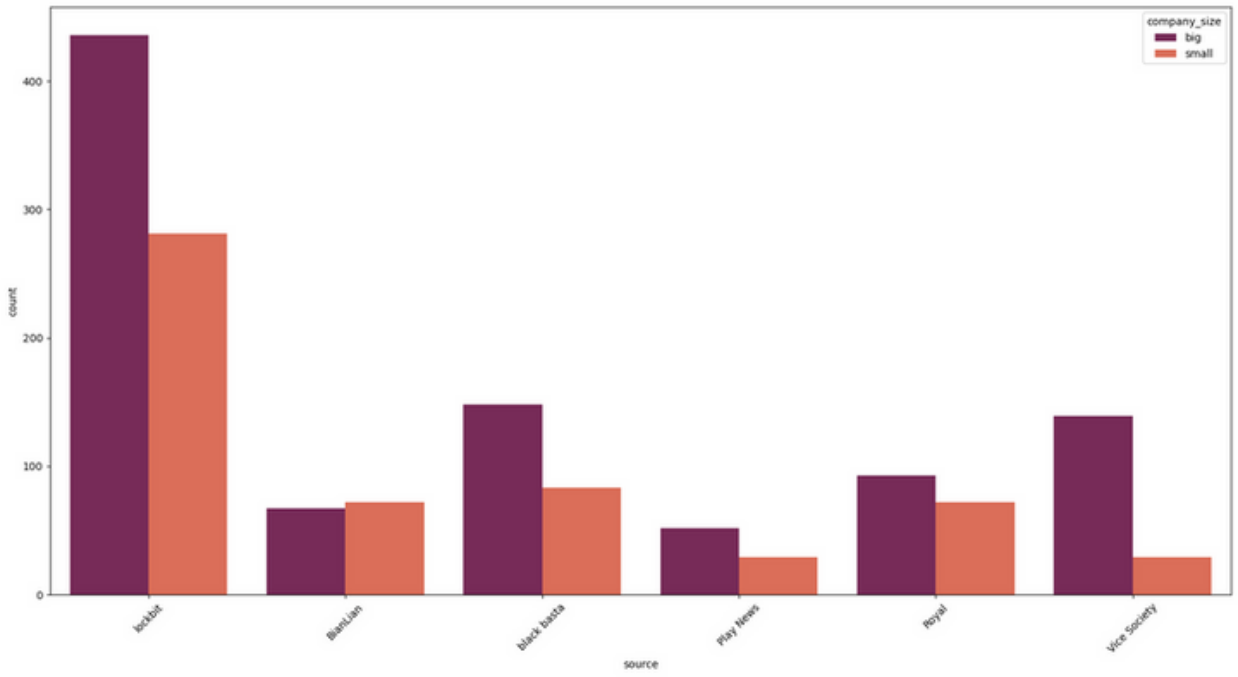


The highest numbers of attacks get from the Lockbit gang and the reason of their high number of attacks is because it is a ransomware-as-a-service (RaaS) model, meaning that it can be rented out to other cybercriminals who may not have the technical expertise to carry out attacks themselves.

Insights extracted throughout the project

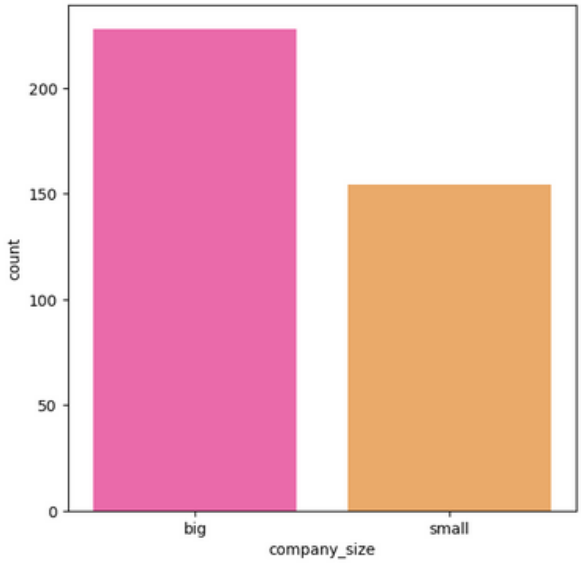


Due to the limited duration of our crawling period, it appears that the majority of companies were not paid for their data. Additionally, the number of companies that paid and subsequently had their data removed from the website during our crawling period was relatively small compared to the total number of companies in the gang's history.

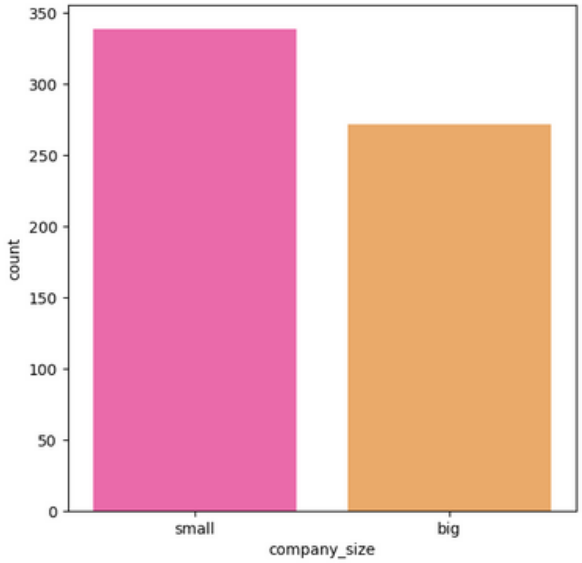


The ratio of company sizes for each gang appears to be consistent, suggesting that there is homogeneity in the data across different gangs, This finding may aid in resolving inferential questions.

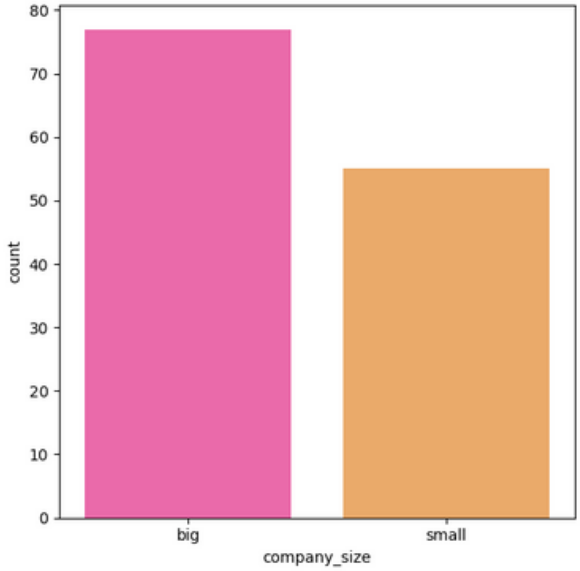
Insights extracted throughout the project



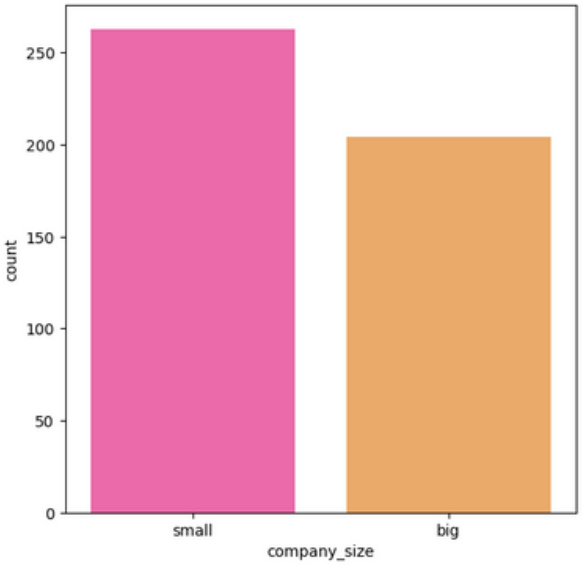
number of attacks for small and big companies in European



number of attacks for small and big companies in United States



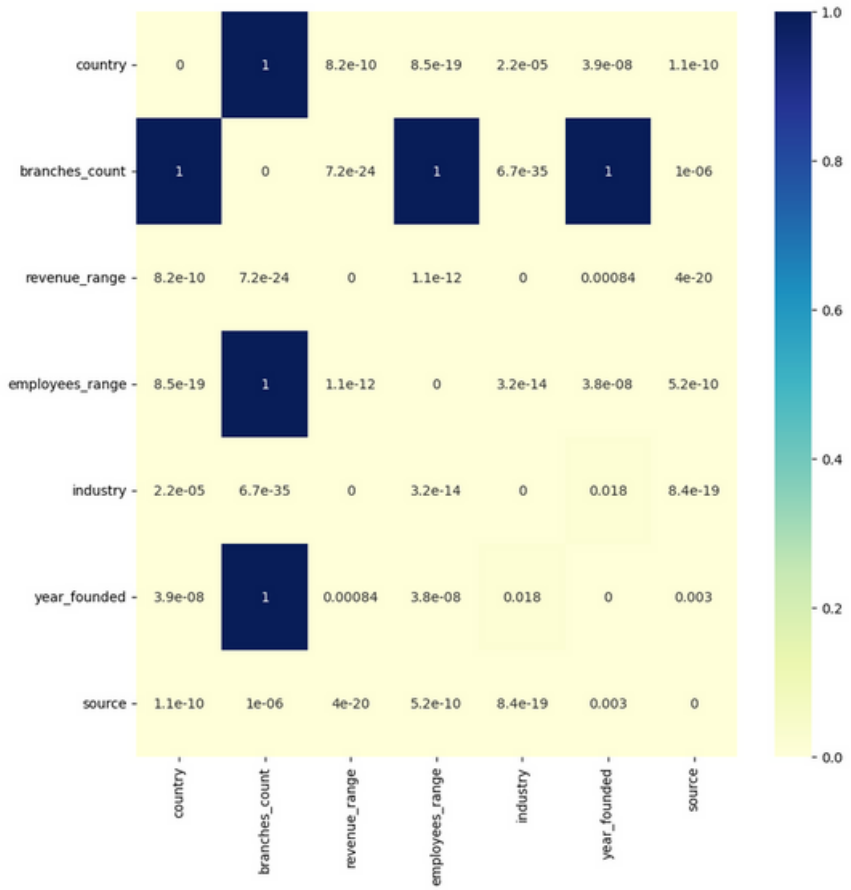
number of unattacks for small and big companies in European



number of unattacks for small and big companies in United States

we use this data to infer the distribution of Europe company size given USA company size distribution.

Insights extracted throughout the project



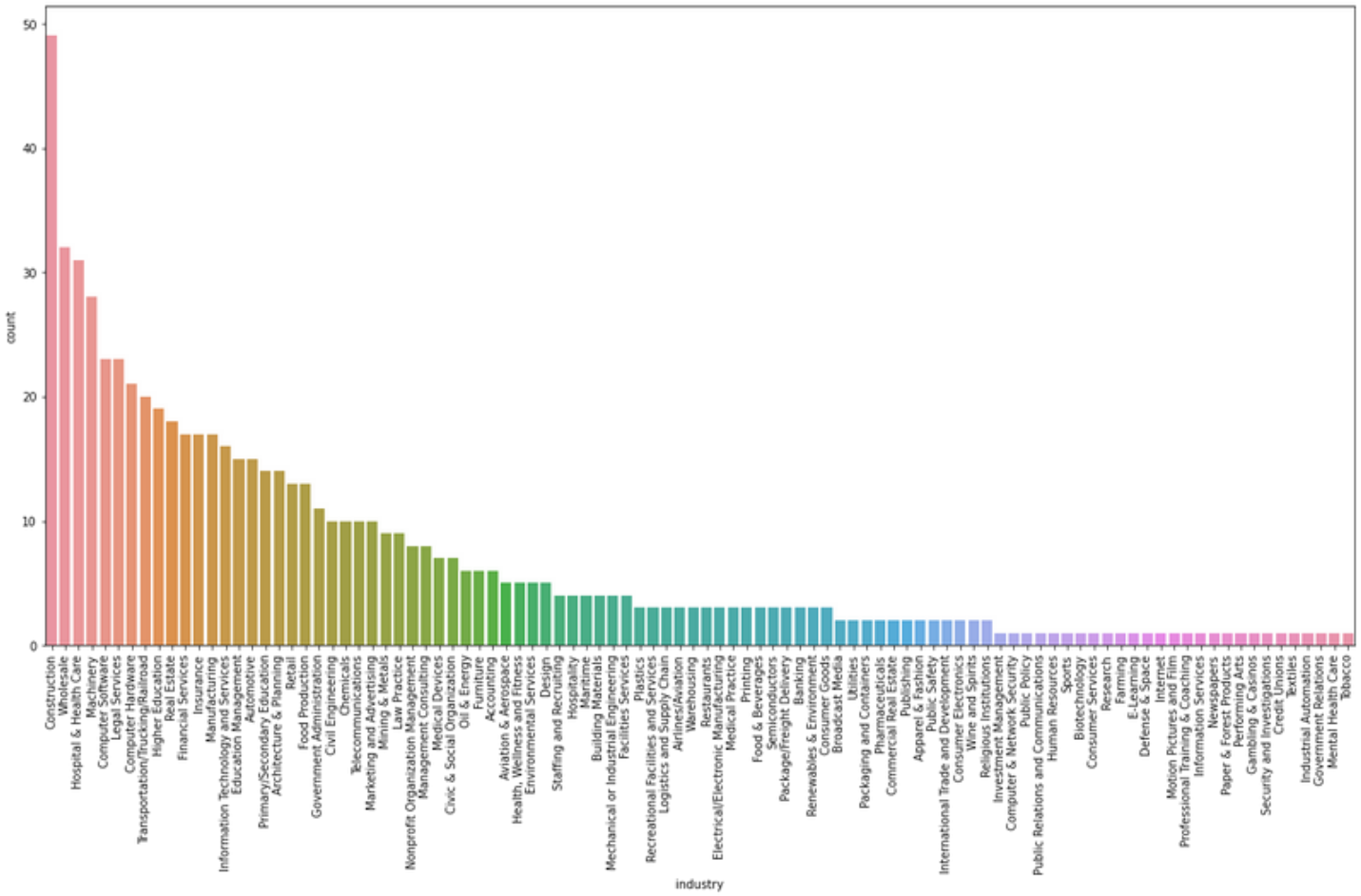
first, we select the data columns that are important for the attacked and unattached company then we fill in the missing data and convert it to numerical.
after that, we calculate the chi-square test to test their correlation.
heat map between features that shows a small correlation between selected features.

Questions

1. What are the sectors which attackers target regularly?

It is descriptive question get answer directly from the data

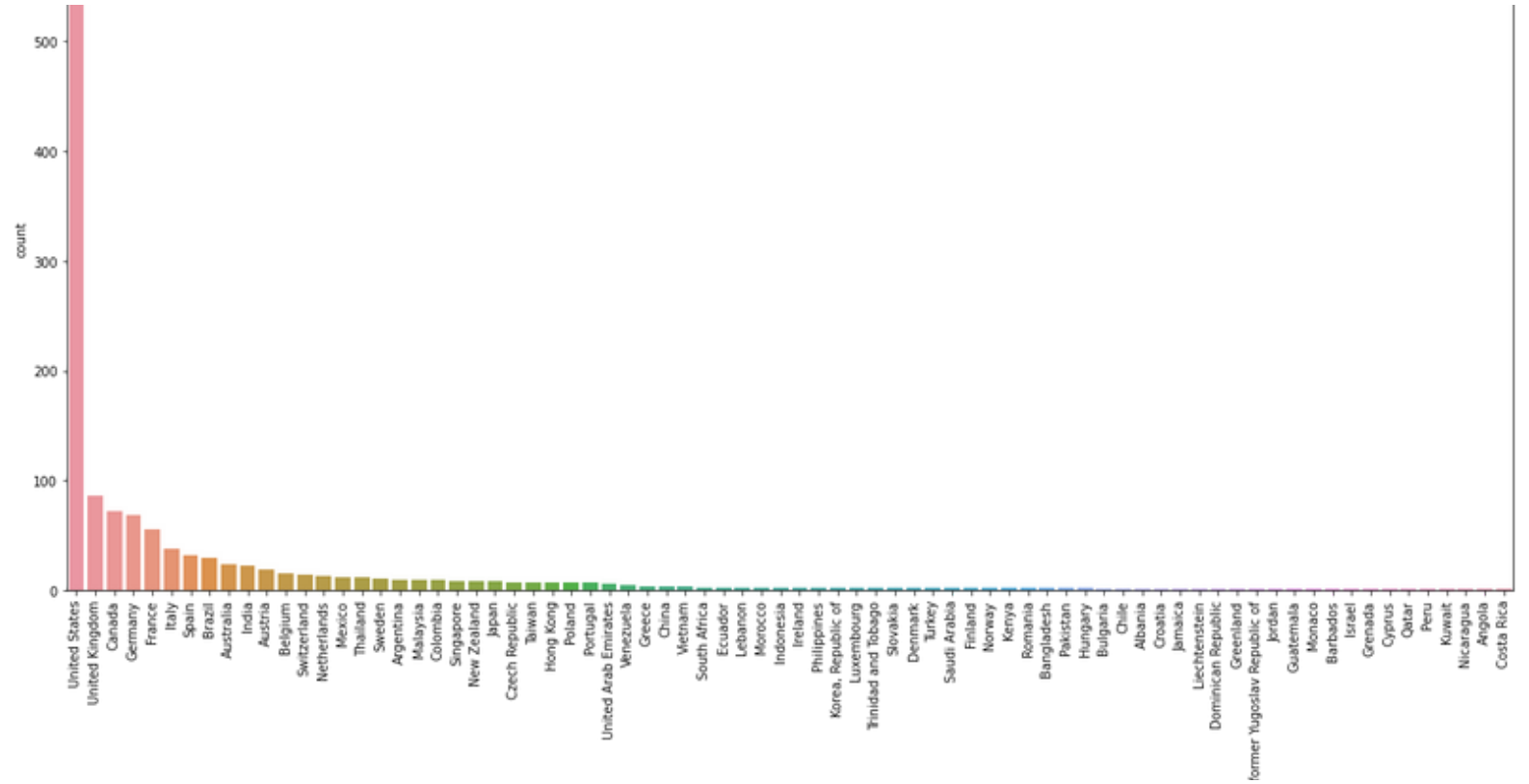
(Construction, Wholesale, Hospital & Health Care,.....)



Questions

2. Do attackers target companies that lie in specific geographical regions?

According to the data, it appears that attackers are focusing a significant portion of their attacks on the United States.



Questions

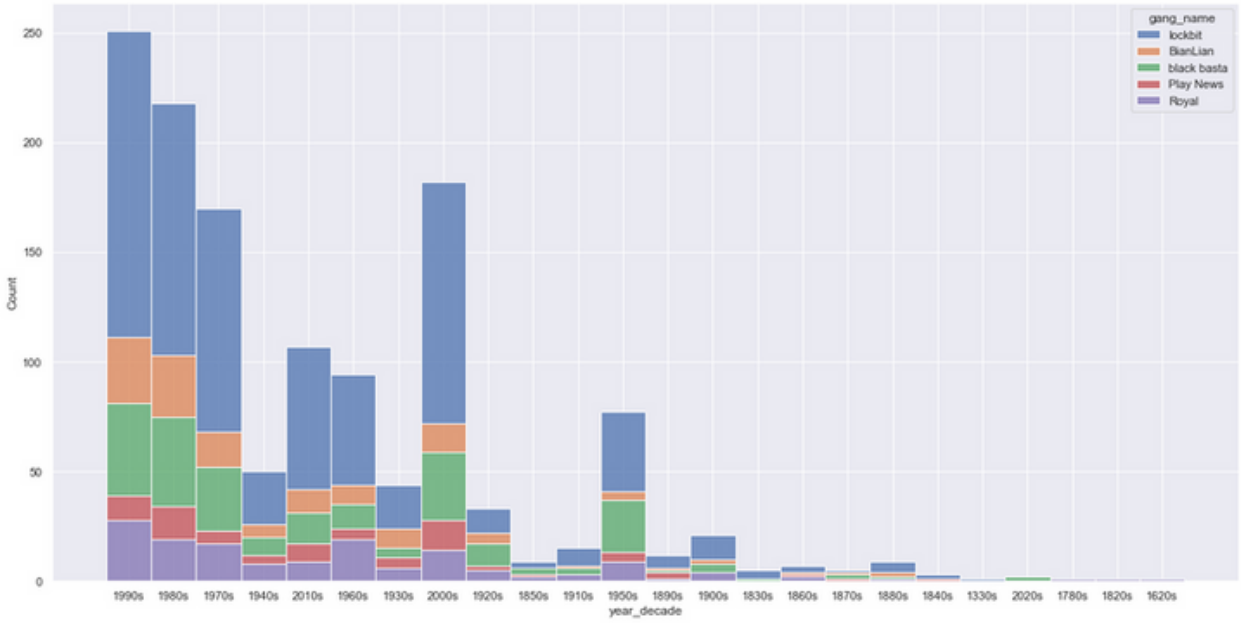
3. Are small companies in Europe vulnerable to attacks, given that companies in the USA rank among the countries most vulnerable to attacks?

This Inferential question answered we detect that the distribution in the USA can be generalized to Europe. That may able us to predict the distribution in Arab Areas where our customer market the security tool.

	USA	Europe
attacked	339	154
Non attacked	263	55

The p-value is equal to 1.3632153665534083e-05 <.05

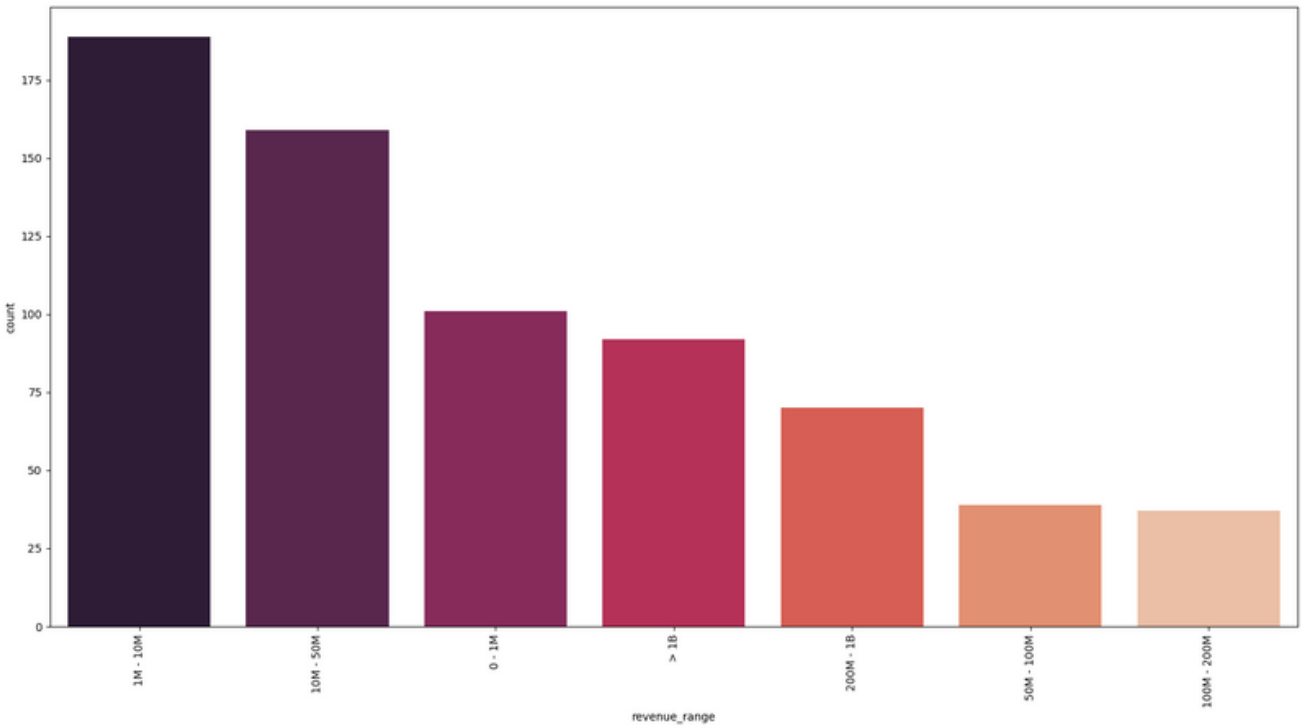
4. What is the relation between the gang and the company's year of establishment?



Get no clear relation between gang and company's year of establishment, as there is no gang focus on one type of companies.

Questions

5. Do the attackers target the companies with higher revenue more than those with lower revenue?



The data suggests that this hypothesis is false, as it appears that medium-sized companies are more likely to be attacked compared to small or large companies.

6. Can we predict the probability of a specific company being attacked ?

used logistic regression and SVM to predict if the company will be attacked or not based on companies information(country, branches_count, revenue_range, employees_range, industry, year_founded source)

	source	precision	recall	f1_score	support
0	BianLian	0.969925	0.984733	0.977273	131
1	Play News	0.962500	0.962500	0.962500	80
2	Royal	0.956790	1.000000	0.977918	155
3	Vice Society	0.985816	0.939189	0.961938	148
4	black basta	0.981818	0.977376	0.979592	221
5	lockbit	1.000000	0.934820	0.966312	583

Questions

7. Can we predict the attack on the specific company, which gang is more probable to do this attack?

use random forest to predict what is the gang attacked the company if it will be attacked

Classification report:

	precision	recall	f1-score	support
0.0	0.00	0.00	0.00	26
1.0	0.00	0.00	0.00	16
2.0	0.12	0.10	0.11	31
3.0	0.41	0.30	0.35	30
4.0	0.29	0.20	0.24	44
5.0	0.52	0.76	0.62	117
accuracy			0.42	264
macro avg	0.22	0.23	0.22	264
weighted avg	0.34	0.42	0.37	264

The accuracy of the findings may be limited, as the data is not well-balanced and there does not appear to be a clear pattern that can be used for classification.

Final findings and results

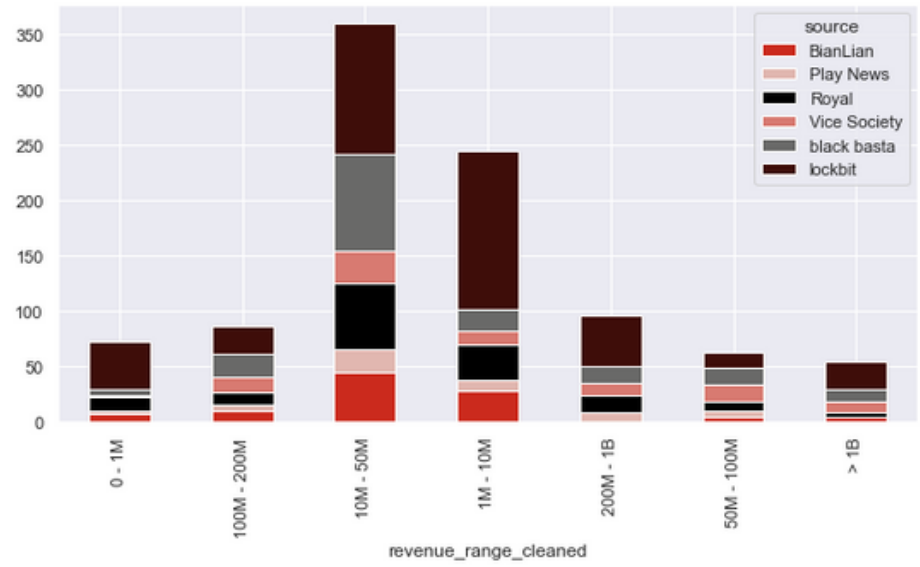
1. Based on our analysis, it appears that the United States is a good market for our company's products, as companies in this region experience a high number of cyber attacks.
2. We recommend targeting construction companies specifically and promoting the importance of security tools within this sector.
3. When determining the price of our subscription-based security tool, we should consider the revenue of the targeted companies and customize the product for those with revenue between 1 million and 10 million, as they tend to experience more attacks.
4. Our focus should be on developing robust security tools that are able to combat the encryption techniques used by Lockbit, as this gang is currently one of the most active.
5. We suggest utilizing a predictive model to assess the likelihood of a given company being targeted for an attack.

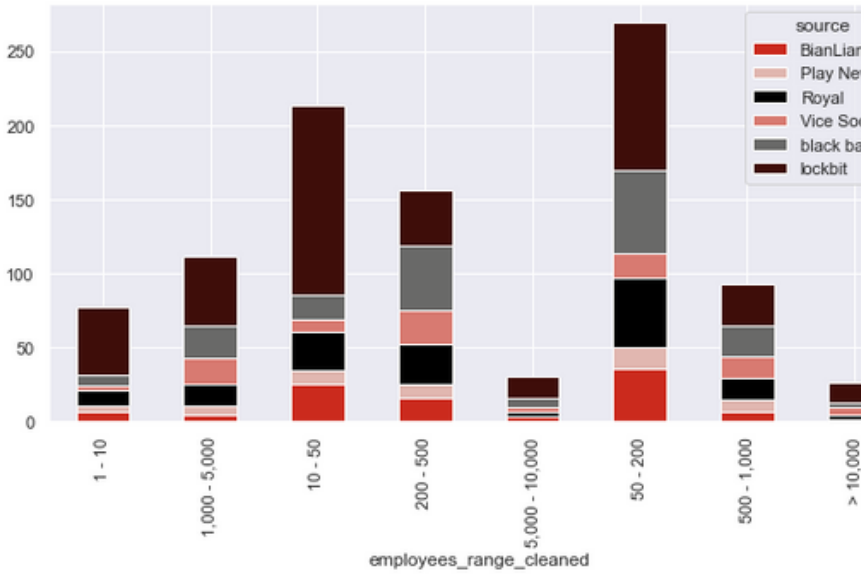
Future work and enhancements

Moving forward, we aim to expand our data collection efforts to include information on companies that have paid ransom and the number of attacks they have experienced. This will allow us to gain deeper insights into which companies are being targeted most frequently by cyber gangs. Additionally, we plan to gather data on the amounts of ransomware paid in order to better predict future ransom demands and use this information to persuade customers to purchase our security services.

Furthermore, we recognize the importance of understanding the timing of cyber-attacks and the activity patterns of different gangs. To achieve this, we will extend our data collection period to cover at least one year and use this data to build time series forecasting models. These models will help us predict when to offer discounts, free trials, and other promotions to customers to maximize sales and mitigate the risk of attacks.

	source	country	counts
171	lockbit	United States	221
106	black basta	United States	139
58	Royal	United States	103
15	BianLian	United States	79
89	Vice Society	United States	48
40	Play News	United States	21





Proportion of each gang in the data