1. Clear contribution of each member.

NotImplemented()

1. your effort in applying the data analysis cycle and epicycle in detail.

First question was how to improve sells for our costumer (we should sell security services).

To answer this question, we should know which company get attacked to market our project to it. In the beginning research we found that there are three important information about any company that gat attacked: sector (industry), region (country), revenue. Then these three info was our first target to collect.

To collect names of the companies that gat attacked we must access the gangs’ websites on the deep web; we do this using tor and selenium. (momken nsii7 hena ala kol mohawalatna)

After we knew the attacked companies we must to collect data about this company we try lot of APIs most: Yahoo finance, ChatGPT, Bing, ORB (the successful trail).

ORB API take company name and it return more than one company with probabilities we take company with the higher probability; it returns information (e.g. country, year founded, revenue rang, employee rang, …).

Now we need to clean data: we get employee and revenue range from ORB or from websites (same gangs provide an information company revenue and employee count) we formalize the columns values. And using chi square test we get the relation between some categorical columns, for numerical columns we used correlation. We also crawl the info about publishing data (the data will be publishing if the company didn’t pay the ransom) there is gangs tell is publish or not and some gangs give a percentage; we also formalized this column values.

Now to answer first question, which companies threat to be attacked; we must provide model with attacked companies and non-attacked companies, we do that using ORB API again we provide company name to different end-point and it search for similar companies.

Then using machine learning models (logistic regression and random forest) we build two models:

First we logistic regression using (country, branches count, revenue range, industry, employees range, year founded) - all of them are correlated - to predict the probability of this company to be attacked given its information.

Second random forest using the same data and (source (gang name)) we train the model to predict which gang more probable to attack this company.

* Stating and refining the question.
* Exploring the data.
* Building formal statistical models.

1. Knowledge and insights extracted throughout the project ordered chronologically and aided visually with graphs, charts, diagrams ... etc.

* Interpreting the results.
* Communicating the results.

1. Final findings and results.
2. Future work and enhancements

* With more data we should be able to collect big data about companies that pay the ransom and number of attacks for the same company; and with this information we can know more insights about which companies more targeted from gangs.
* We will try to get the information about ransomware amount with this information we can detect future ransom; than we can convince customer to buy our product.
* Also we need to collect data over a year or more to detect each gang active time and off time; this will able us to predict when to do offers or free trails and when we don’t, also we can build time series forecasting model to predict attack to the company on range of time

**Effort in applying the data analysis cycle and epicycle in detail.**

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.

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 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.

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 similar 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.

**Future work**

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.