

**FIT5147 Data Exploration and Visualisation**

**Data Exploration Project**



Student Name : Sriram Vasan

Student ID : 33514569

Applied : 05

Tutor’s Name : Megan Power

TABLE OF CONTENTS

[1.INTRODUCTION 3](#_Toc145625041)

[1.1 PROBLEM DESCRIPTION 3](#_Toc145625042)

[1.2 MOTIVATION 3](#_Toc145625043)

[2.DATA CLEANING AND DATA WRANDLING 4](#_Toc145625044)

[2.1 DATA DESCRIPTION 4](#_Toc145625045)

[2.2 DATA CLEANING 6](#_Toc145625046)

[2.3 DATA WRANGLING 7](#_Toc145625047)

[3.DATA EXPLORATION 8](#_Toc145625048)

[4.CONCLUSION 11](#_Toc145625049)

[5.REFLECTION 12](#_Toc145625050)

[6.BIBLIOGRAPHY 13](#_Toc145625051)

# 1.INTRODUCTION

Global technological advancement and the evergrowing landscape of technology has seen an immense growth in the past few decades. Evloving from the introduction of World wide web to the rapid development of using smartphones and further more evloving technologies like artificial intelligence and blockchain is becoming part of our daily life in an alarming rate. This has only increased the way we communicate and the number of interactions we have with fellow users over the internet leveraging this technology.

However the rise of these technologies has given rise to a darker side of the internet and digital realm leading to technological scams. With every new advancement in technological growth, scammers and cybercriminals have taken advantage of the trusting individuals and unprotected industries.

## 1.1 PROBLEM DESCRIPTION

For this project report, we thrive to analyse and answer the questions below:

1. Is there any geographical pattern to the occurrence of tech support scams (e.g., higher concentration in certain countries or regions), and does this align with regions experiencing higher unemployment rates?

2. Are there any particular domains, URLs, or email addresses (from the Tech Support Scams Dataset) that are consistently associated with higher numbers of reported scams in countries with higher unemployment rates?

## 1.2 MOTIVATION

The correlation between rising unemployment and the surge in tech support scams is a stark reminder of the real-world impact of data analysis. It's a chance to understand the societal challenges we face and how data science can be a powerful tool for addressing them.

What's truly motivating is the potential for our findings to make a difference. By digging deep into this dataset, we are contributing to the collective effort to combat tech scams, protect individuals, and support policy initiatives. This project is a reminder that as students, we have the opportunity to drive change through data-driven insights. So, let's embrace this challenge, keep our curiosity alive, and use our skills to create a positive impact on our society and beyond.

# 2.DATA CLEANING AND DATA WRANDLING

## 2.1 DATA DESCRIPTION

In this report, I am making use of two datasets

1. Tech Scams dataset

**Source** : [Choozn](https://github.com/choozn/PopupDB-Data/) – Popup.org

**Context** : The dataset contains the data from PopupDB project which has been collecting tech support scams and their popups since 2018 with urls, domains, hosts and IPs different tech scams links and the website phone number and their location from the patterns.

**Scope** : The data includes 11375 rows and 14 columns. The data is tabluear containing spatial , temporal and categorical data. Loaded into tableau the dataset looks like below.

* **UID :** Unique ID of the popup report
* **Domain :** Domain of the url
* **Url :** Url of the scam popup
* **Number :** Tech scam Phone number
* **Host :** The host of the website or service
* **Country :** The country the website or service is hosted in
* **City :** The city data
* **IP :** Ip of the popup website
* **ASN :** ASN provider of the IP
* **Latitude :** Latitude of the location
* **Longitude :** Longitude of the location
* **Hash :** Hashed value the URL
* **Date :** Date the popup was reported
* **Mail :** Mail of the website

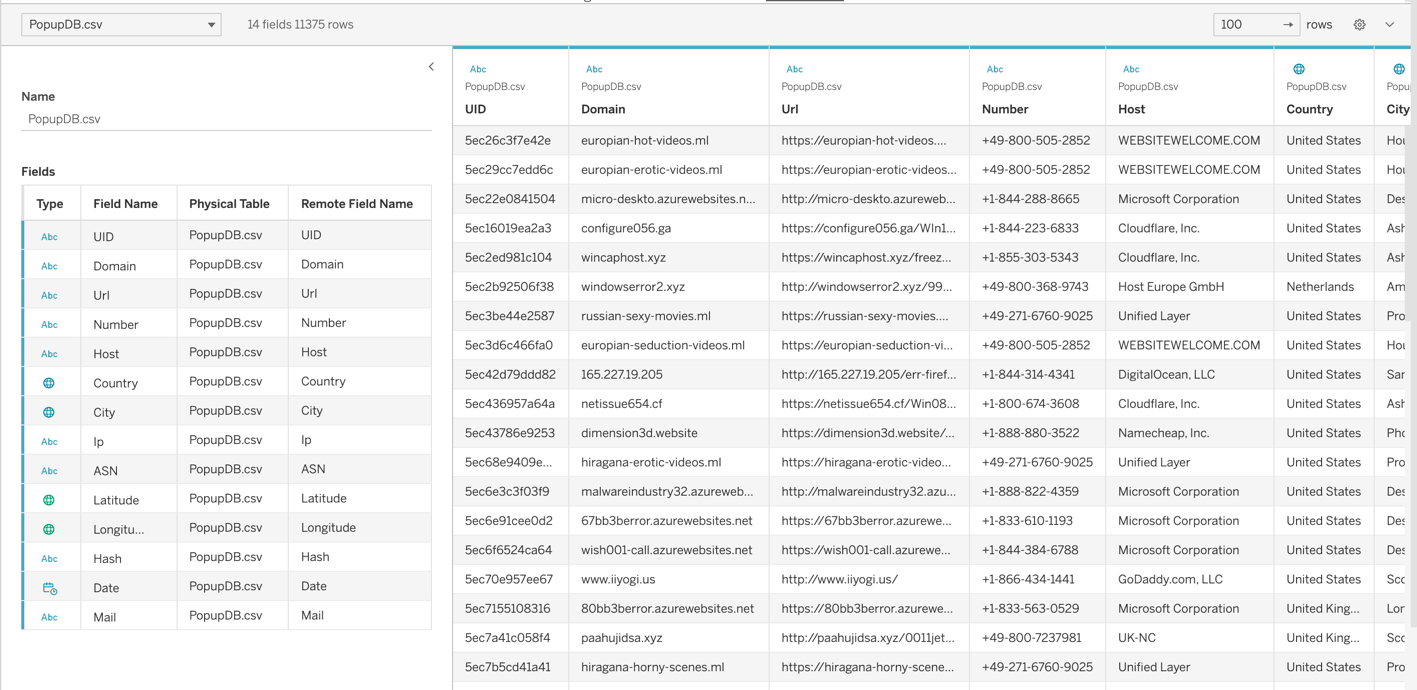


Figure 2.1: Tech scams data view

1. Unemployment Dataset

**Source** : [The World Bank](https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS)

**Context** : The dataset contains the data from the world bank indicators of the unemployment data as a percentage of the total labour force of the respective country in different years.

**Scope** : The data includes 270 rows and 68 columns. The data is tabluear containing temporal and numerical data. Loaded into tableau the dataset looks like below.

**Country Name** : The name of the country

**Indicator Name** : The indicator is all the same of the unemployment data

**Indicator Code** : The code for the indicator

**Country Code** : The code given for each country

**Years as Columns** : 1960 – 2022 with data for each country of unemployment as the total percentage of the available labour force

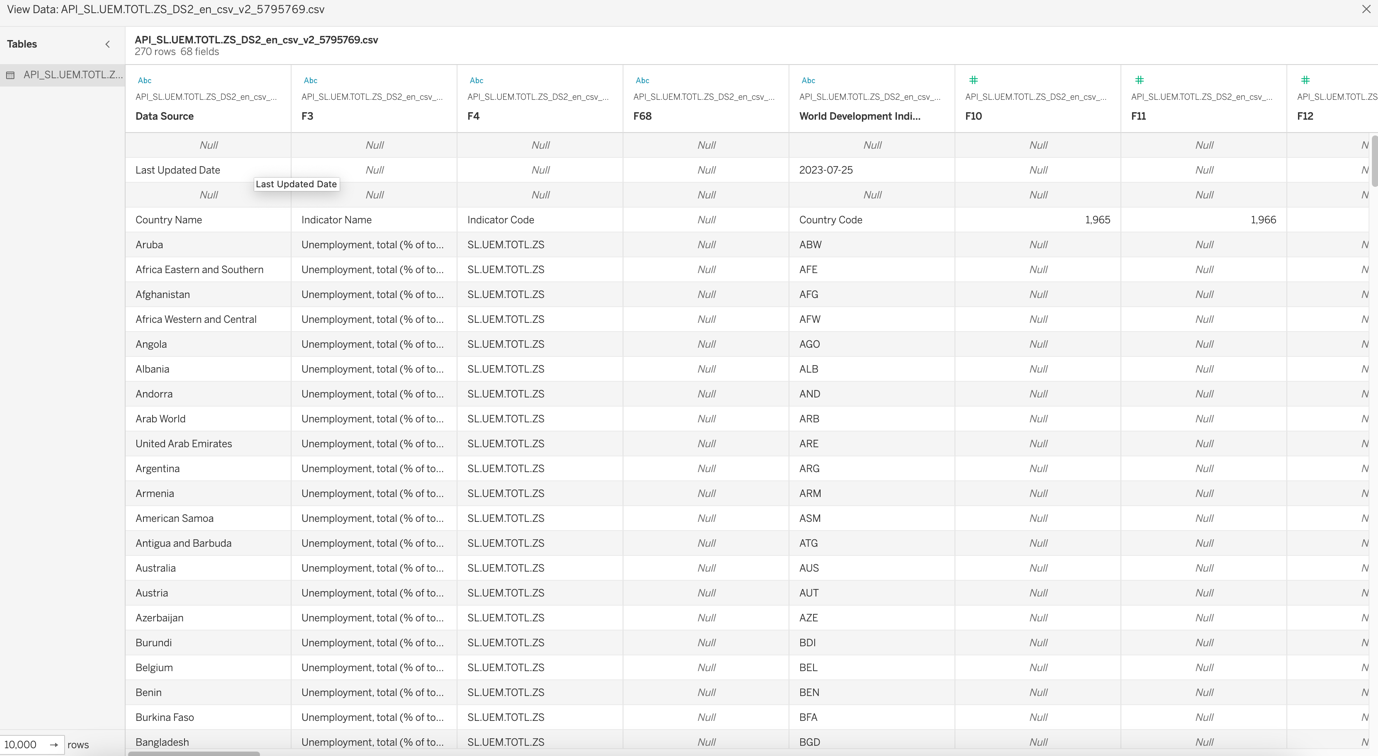


Figure 2.2: Unemployment Data view

## 2.2 DATA CLEANING

Exploring the data misses using vis\_dat library in R , we can see the below for both the datasets.

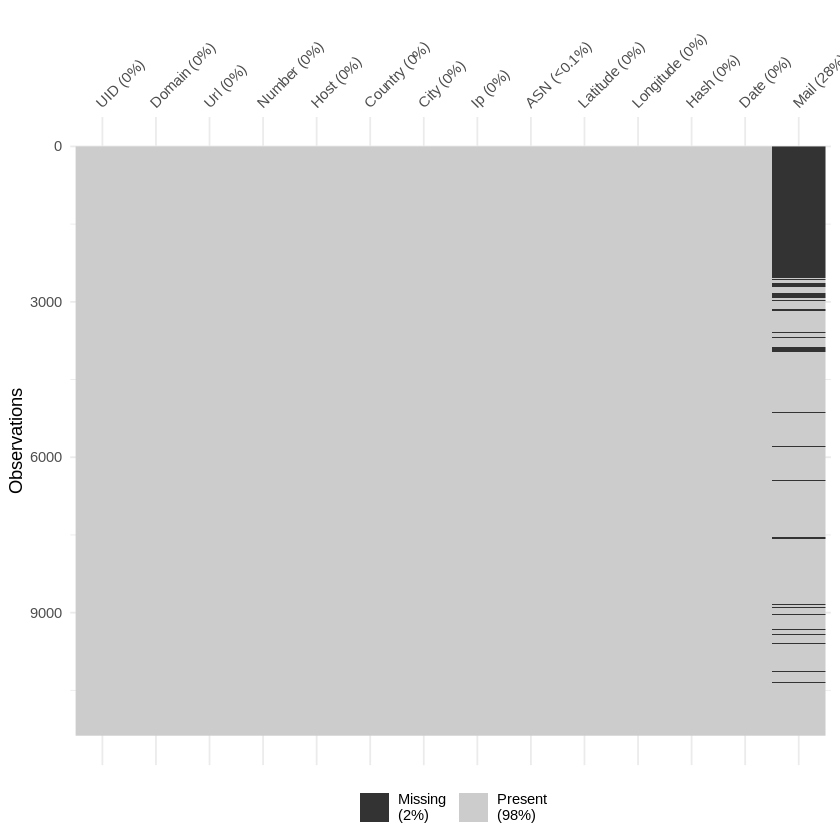


Figure 2.3: Missing values of tech scams data

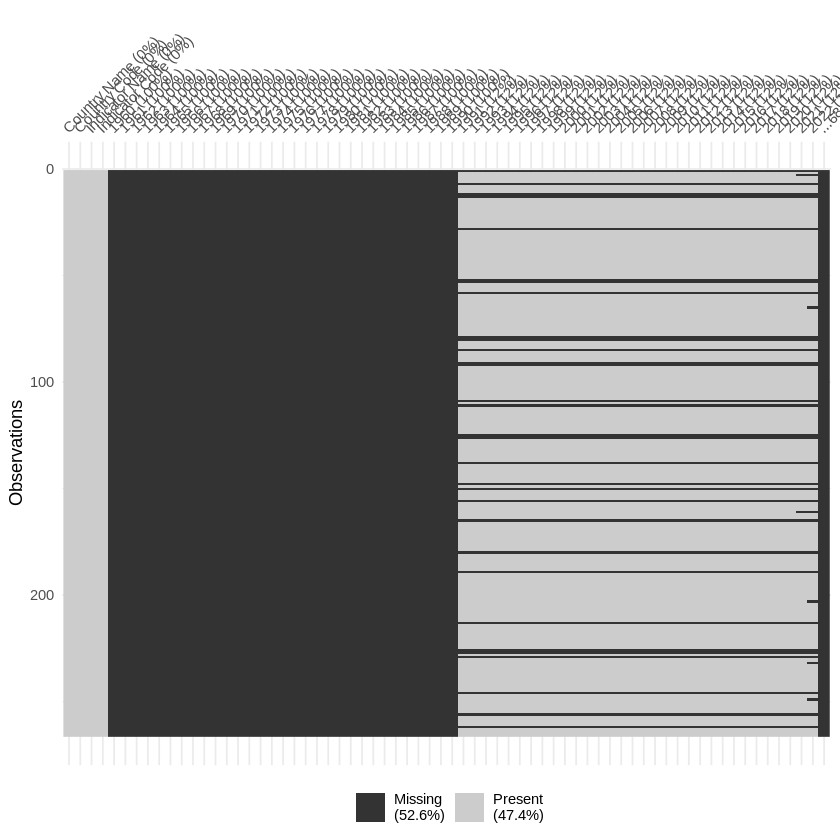


Figure 2.4: Missing values of Unemployment Data

The dataset1 – tech scams data, has less than 2% of the data to be missing, which seems to be mostly from the Mail column, so we decide to drop the column. And on further exploration, the data column named Hash, has the hashed values of the urls; which is not of much use to us without the hash key to decode them, so we decide to drop the hash column as well.

For the dataset2 – Unemploument data, a lot of columns have null values and around 52.6% of the data is missing which is a huge chunk of data. This could point to many countries not existing in the 20th century and also with the unavailablity of data as well. Going through these raw data, we can notice that the dates necessary for us is only among the years, 2021 and 2022.

After dropping the columns the Mail and Hash, the data seems only to have less than 0.1% of the data to be missing as shown in *Figure 2.5*.

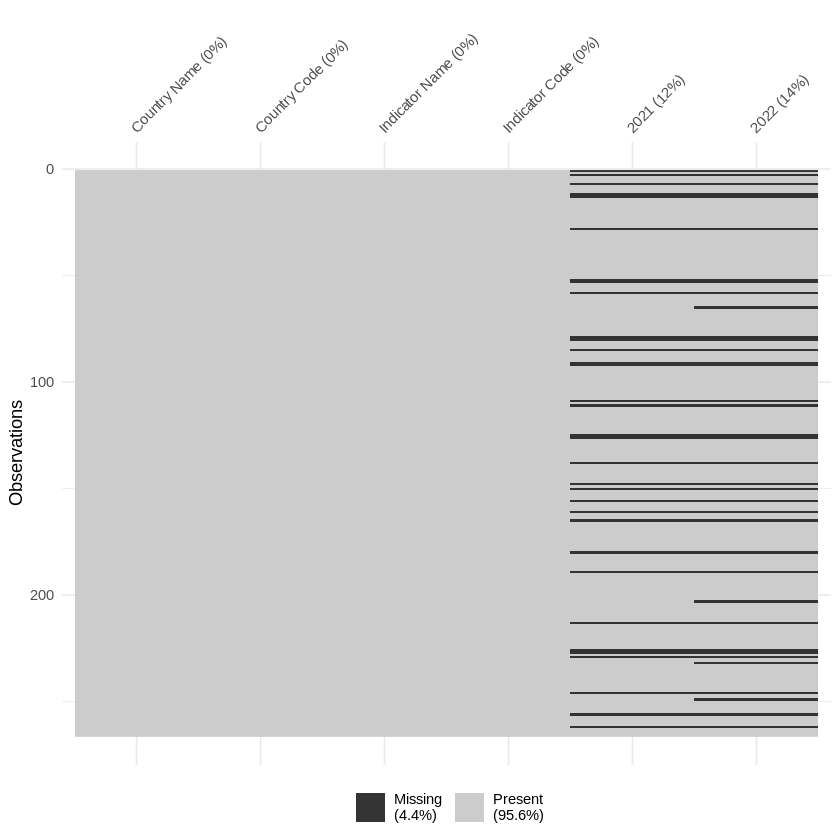


Figure 2.5: Unemployment Data columns Dropped

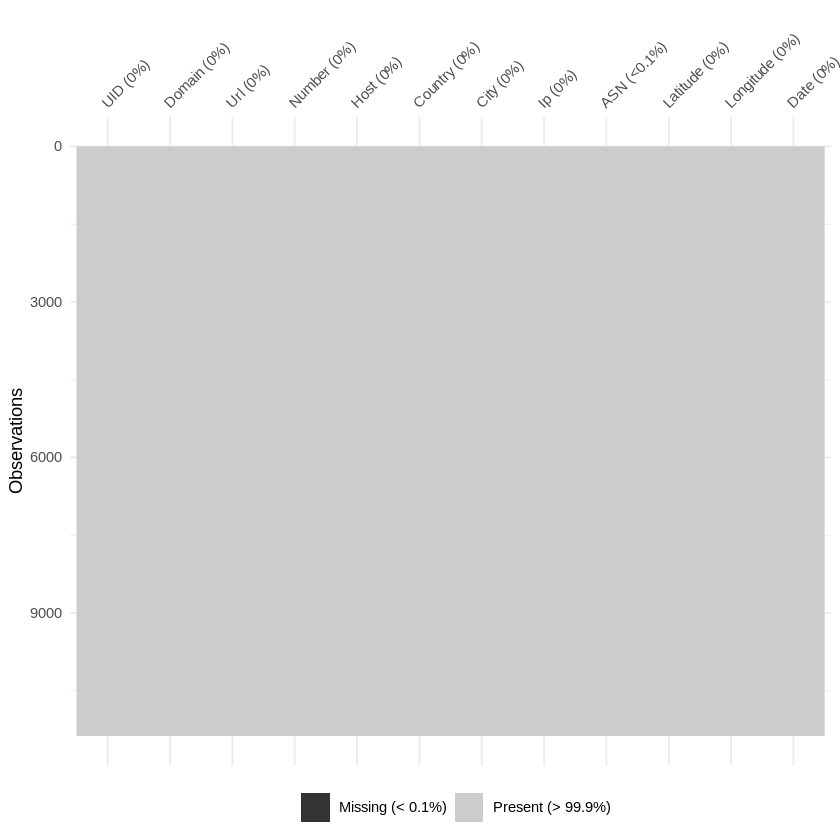


Figure 2.6: Scams Data post Columns dropped

The observations for the unemployment data has been cleaned to only include the years 2021 and 2022 as shown in *Figure2.6*. Even though the data has around 4.5% of missing data, we are not gonna impute or replace it anyway.

## 2.3 DATA WRANGLING

Data wrangling refers to a variety of processes designed to transform raw data into more readily used formats[3]. To get the data and convert it into the most useful some aspects of the data needs to be analysed in detail.

Data like the phone numbers and IP addresses contain a lot more info in them like geographic locations where they come from and the organizations they belong to. To accommodate and include complex information that entails and covers these , we take necessary steps to include certain fact table to understand them better. Two fact tables for both IP addresses and Phone number are obtained to get more insights.

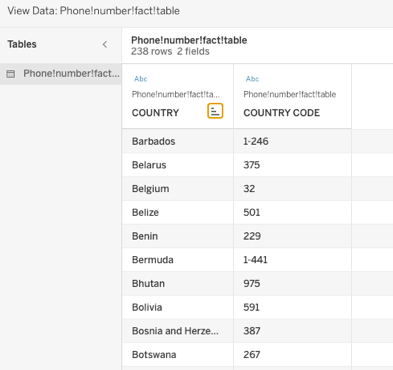


Figure 2.7: Phone Fact table view

Phone number fact table, this dataset is from the [countrycode.org](https://countrycode.org/) where the fact table only for the phone number country code and the respective country is taken to join with the scams data to get the country of the phone number it belongs to. The dataset for the fact table is as shown in *Figure2.7.* The phone number fact table is then inner joined with the scams data on the country code column.

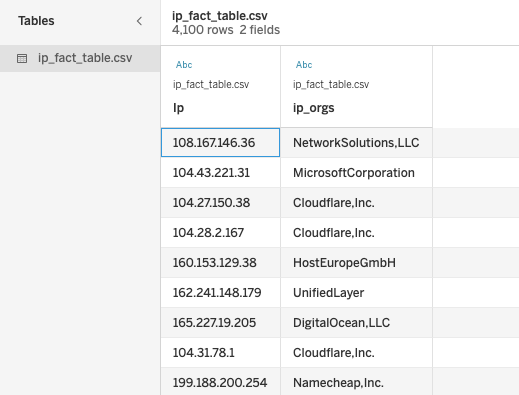


Figure 2.8:IP Fact table View

In the IP fact table, the dataset is obtained by using the [ipinfo.io](https://ipinfo.io/) API to get the organizations where the website or the scams have been hosted. This gives a rich understanding of the IP data and also helps us extract other data from this for hidden and complex data type. The data from the API is extracted using python and added to tableau and joined based on the IP addresses in the scams dataset. The dataset looks as shown in *Figure2.8*. The organizations extracted from the different IPs along with the Phone number fact table are joined as shown below in tableau.

A calculated column is created to extract the phone number country from the table and then the phone number code is joined with the Phone fact table and the existing IPs are joined based on the IPs. The join logic and the joined data columns are as shown below.

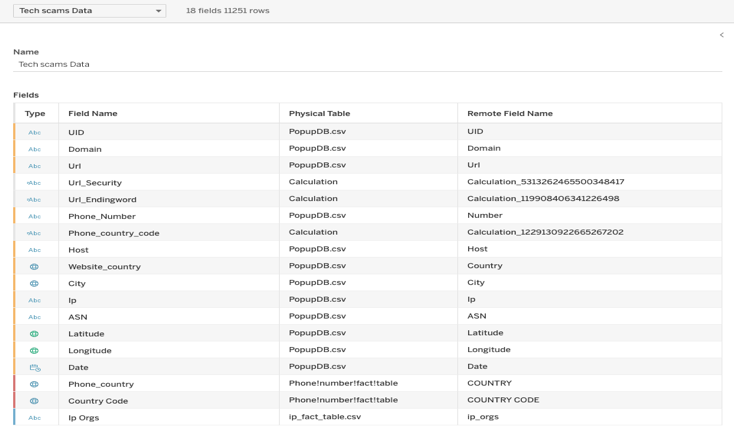


Figure 2.10: Cleaned Tech Scams

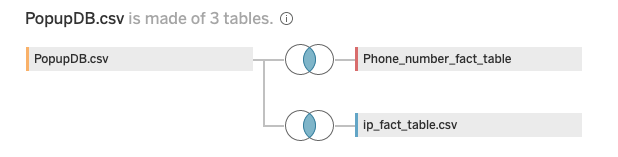


Figure 2.9: Join Logic of the fact tables

Further more the dataset of employment is transposed into usable format with year as a column and unemployment percentage as the other column and the unnecessary columns are dropped. The data looks like as it is shown in *Figure2.11*. Which is then futher joined with the tech scams data based on the country as in *Figure2.12*.

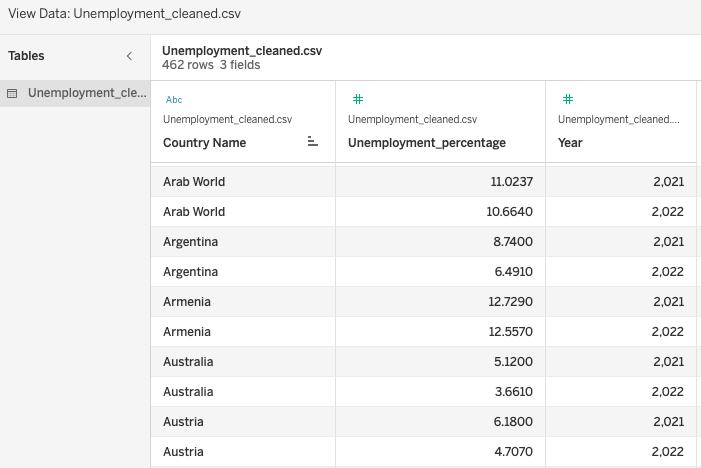


Figure 2.11: Unemployment Dataset



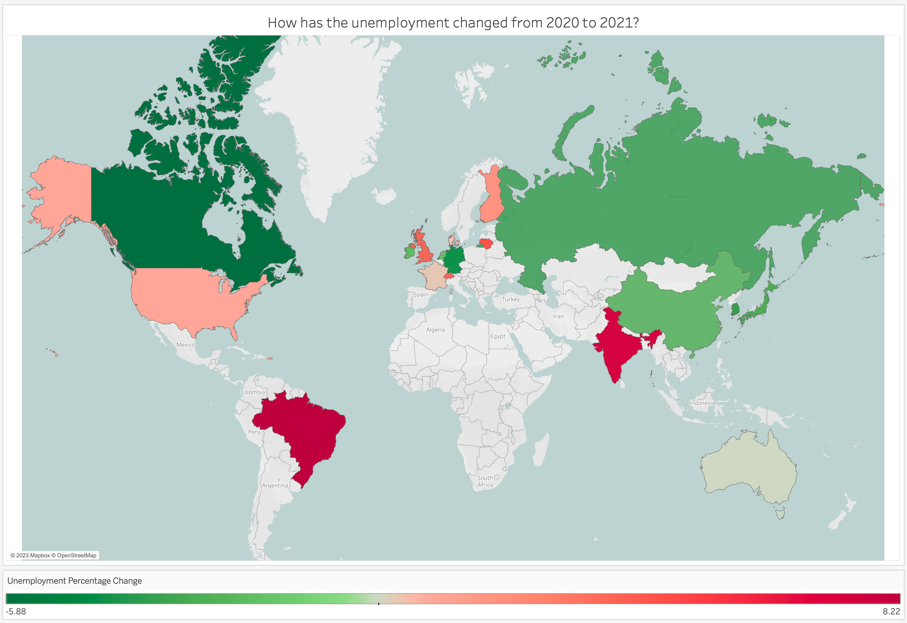
Figure 2.12: Data join of Tech scams and Unemployment

# 3.DATA EXPLORATION

Now that the dataset is cleaned and transformed into the way we would like, we dive into the exciting part of the project - data exploration. From the dataset we are trying to answer the questions outlined in the Problem Statement Section.

**Is there any geographical pattern to the occurrence of tech support scams (e.g., higher concentration in certain countries or regions), and does this align with regions experiencing higher unemployment rates?**

Figure 3.1: Unemployment increase around the world.



From this visual the countries unemployment shown has increased from the year 2020 to 2021. The choropleth map shows the countries with increase in unemployment percentage in red and the countries in green with a decrease in the percentage of unemployment.

Infering the *Figure3.2* below we can see that there is a huge increase in the percentage of scams coming out of India changing from 2020 to 2021. The *Figure3.3* supports the above deductions with a general idea on how the scams have a linear increase with the increase in unemployment rate.

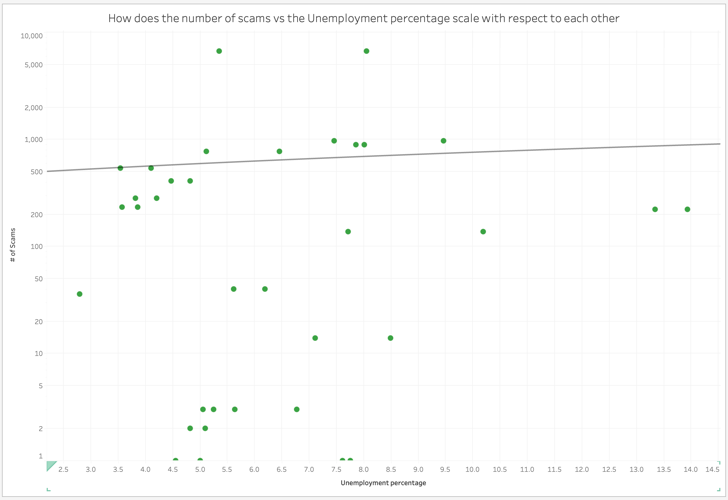


Figure 3.3: Logarithmic Scatter plot of Scams and Unemployment

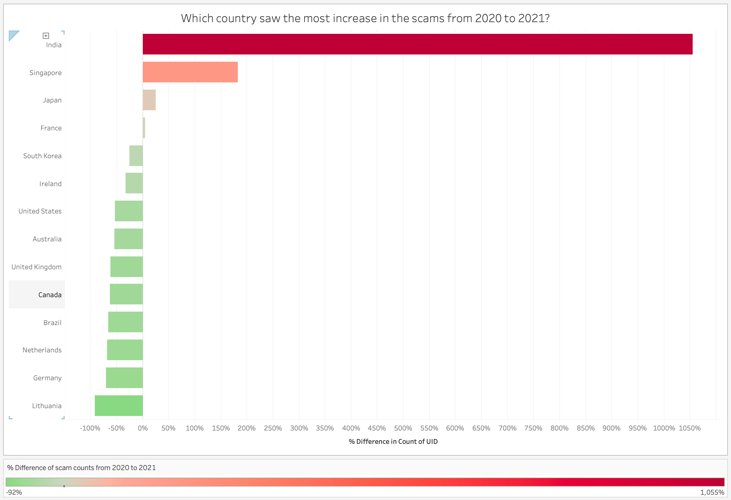


Figure 3.2: Scam increase by countries.

So this answers out first question that yes, the scams frequency have a significant positive correlation between the increase in the unemployment and they do have an effect on each other.

**Are there any particular domains, URLs, or email addresses (from the Tech Support Scams Dataset) that are consistently associated with higher numbers of reported scams in countries with higher unemployment rates?**

From the dataset since we dropped the mail column on the data cleaning we dive into the analysis of the URLs and the domains of the websites.

Understanding the distinction between HTTP and HTTPS is crucial in grasping web security fundamentals. HTTPS encrypts standard HTTP requests and responses, significantly enhancing the security of data exchanges[5]. So to analyse the URLs , using tableau calculated columns and splitting the urls based on protocol used. I get the following visual.

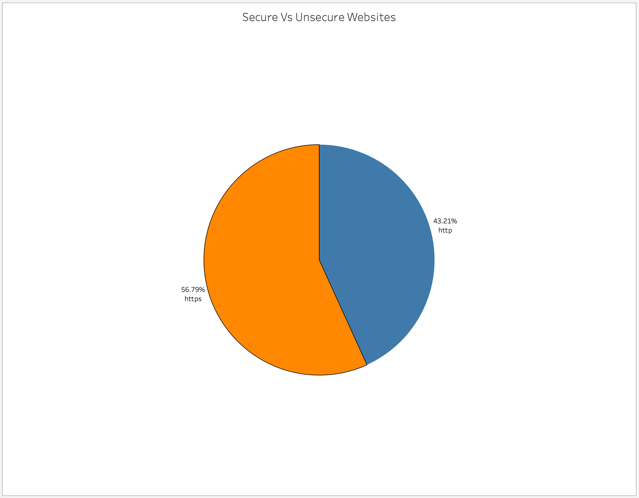


Figure 3.4: Secure vs Unsecure Urls

On contrary to popular belief one would think that scam links would have limited to no security, these scam wesbites or links have almost half of them to be secure furthermore increasing their penetration into the tech scam world and being hidden in the browser alert radar.

Clustering the organizations that the websites are hosted based on the number of scams we can find that there is a major change in the way the organization Microsoft, which has significantly reduced the number of scam websites it had hosted, but company like cloudflare failed to not only reduce but infact went the opposite direction in facilitating the number of scam websites it has hosted.

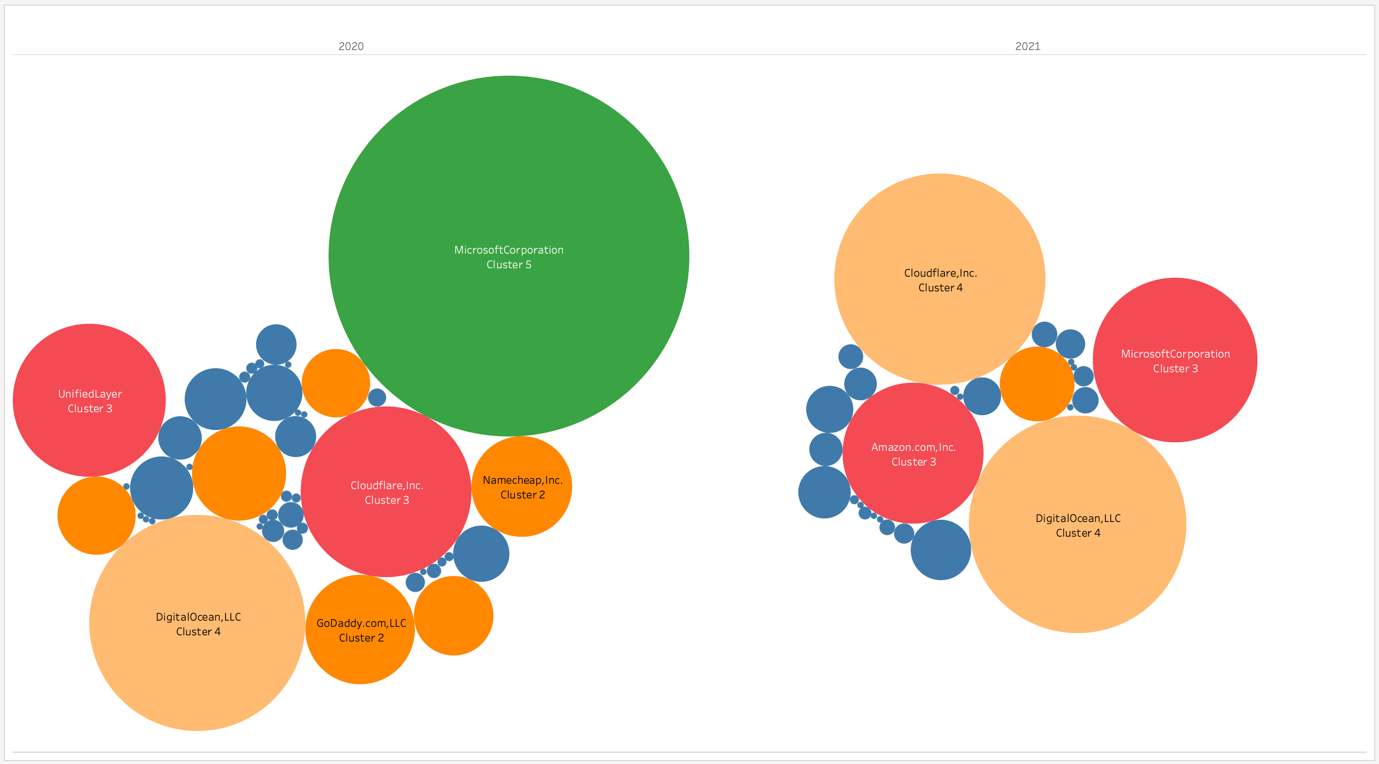


Figure 3.5: Clustering to find any changes over the years.



Figure 3.6: Top 5 Countries by Unemployment and Website Organizations

Filtering the top 5 Countries based on the unemployment increase in percentage, we can see a strong correlation between the difference in the percentage of scams increase from 2020 to 2021. The organizations in some countries as seen have a drop in the number of , but we can still see a significant unchange in the India which seens to have the podium in the tech scams happenings.

Further more, there are trends we can spot in microsoft as we discussed in the clustering , which has seen a major decrease in its scam hosting websites.

So to answer, the countries with higher unemployment rates, tend to have a lower implementation to stop the scams.

# 4.CONCLUSION

The analysis conducted on tech support scams, unemployment rates, and their geographical patterns reveals significant insights into the relationship between these factors. Here are the key findings:

* There is a noteworthy positive correlation between the occurrence of tech support scams and an increase in unemployment rates. This correlation suggests that economic challenges, such as higher unemployment, may create an environment conducive to the proliferation of tech support scams.
* Geographically, India stands out as a focal point for tech support scams, with a substantial increase in the percentage of reported scams from 2020 to 2021. This indicates a growing issue in the region.
* Analysis of URLs reveals an unexpected trend: a significant portion of scam websites employs HTTPS, implying a level of sophistication in hiding their malicious intent. This challenges the common assumption that scam websites lack security features.
* Clustering hosting organizations by the number of scam websites they facilitate highlights an interesting contrast. While Microsoft has notably reduced its association with scam websites, companies like Cloudflare have seen an increase in hosting such sites. This suggests variations in efforts to combat tech support scams among hosting providers.
* Focusing on the top five countries with the highest increase in unemployment percentages, it becomes evident that the difference in the percentage of scam increase from 2020 to 2021 correlates with these country’s unemployment rates. This suggests that countries with higher unemployment rates may have fewer resources or initiatives to combat tech support scams effectively.

In summary, the analysis underscores the complex relationship between tech support scams, unemployment rates, and online security. The findings suggest that economic factors, such as unemployment, play a role in the prevalence of these scams. Furthermore, the HTTPS adoption among scam websites challenges traditional assumptions about web security.

Understanding these dynamics is crucial for policymakers, cybersecurity experts, and law enforcement agencies in developing strategies to combat tech support scams effectively. Additionally, hosting providers should take note of their role in either facilitating or mitigating these scams. As the threat landscape evolves, it is imperative to adapt and implement measures to protect users and businesses from falling victim to these deceptive practices.

# 5.REFLECTION

Throughout this project, I've gained valuable insights into the intricate interplay between tech support scams, unemployment rates. One key takeaway is the strong positive correlation between scam occurrences and rising unemployment, shedding light on the socio-economic factors influencing the prevalence of these scams. The unexpected finding of scam websites using HTTPS underscores the need for a more nuanced approach to web security. Analyzing hosting organization’s roles in facilitating scams also revealed intriguing patterns that deserve further investigation.

In hindsight, I might have explored the temporal aspect of the data more extensively to identify potential seasonality in scam occurrences. Additionally, delving deeper into the specifics of hosting provider’s involvement in scams could have provided a richer understanding of their role in perpetuating or mitigating this issue. Overall, this project has reinforced the importance of adaptability and a comprehensive approach to data analysis in the ever-evolving landscape of cybersecurity and data science.

‌

‌

# 6.BIBLIOGRAPHY

[1] Lebied, M. (2017). *Data Exploration Vs. Data Presentation – Learn More About The 5 Key Differences*. <https://www.datapine.com/blog/wp-content/uploads/2016/12/Header-image.png>

[2] countrycode.org. (n.d.). *Country Codes, Phone Codes, Dialing Codes, Telephone Codes, ISO Country Codes*. [online] Available at: <https://countrycode.org/>.

[3] Stobierski, T. (2021). *Data Wrangling: What It Is & Why It’s Important*. [online] Business Insights - Blog. Available at: <https://online.hbs.edu/blog/post/data-wrangling>.

[4] Ipinfo.io. (2013). *IP Address API and Data Solutions - geolocation, company, carrier info, type and more - IPinfo IP Address Geolocation API*. [online] Available at: <https://ipinfo.io/>.

[5] Venafi (2023). *Venafi*. [online] Venafi.com. Available at: <https://venafi.com/blog/what-are-differences-between-http-https-0/> [Accessed 14 Sep. 2023].

[6] World Bank (2021). *Unemployment, total (% of total labor force) (modeled ILO estimate) | Data*. [online] Worldbank.org. Available at: <https://data.worldbank.org/indicator/sl.uem.totl.zs>.

[7] choozn (2022). *About*. [online] GitHub. Available at: <https://github.com/choozn/PopupDB-Data/>.