**MIT 5032 – 100**

**Analyzing Cyber Security Breaches**

*Pablo Gutiérrez Félix, Chukwuma Ilodibe, Khalil Tafni, Vishal Reddy Musku, Matej Mestrovic*

**Introduction**

Analyzing Cyber Security Breaches was chosen to address the rising issues of information security and the costs that security breaches pose for institutions and individuals affected. In this project, we will analyze about 1000 incidents that happened in the United States in 1997 and between 2002 and 2014. We will show what happened, who was targeted, how, and where it happened. This will help us better understand how security breaches happen and how we can protect ourselves and how institutions can mitigate the risk of losing information, money, and customers. This project is designed to test basic python skills as well as an understanding of the “pandas” library by analyzing a data set about Cyber Security Breaches in the United States. In this project, the team had to clean up data in order to make it easier for us to analyze and gather useful information from it.

**Problem statement and significance/importance of topic**

We have collectively decided to focus on this issue because data security breaches in healthcare environments have become an increasingly significant problem in recent years. The healthcare industry is a prime target for cybercriminals due to the sensitive information that is stored in electronic health records and other healthcare databases. This information includes personal identification information, medical history, and insurance information, making it highly valuable on the black market. To help raise awareness on cyber security concerns in the US, we formulated different questions that investigate the nature of this flagrant issue in the healthcare industry. These questions will be addressed later in the document, with an explanation of the conclusions we retrieved from our data analysis.

**Description of dataset and how we got it**

The dataset we have chosen to work with is called Cyber Security Breaches and it is a collection of data from the late 90’s to the late 2010’s about cybersecurity breaches in the healthcare sector. It was collected from Kaggle and a link to the source of the dataset has been provided in the bibliography of this report. The dataset consists of 14 attributes of which the most essential ones are: Entity (where a breach incident occurred), State, Entity Type (“healthcare provider” or “health plan”), Individuals Affected, Breach Submission Date, Hacking, Unauthorized Access/Disclosure, Improper Disposal, Loss, Theft, Other Types or Unknown, Location of Breached Information.

**Description of python libraries used**

To conduct our analysis for this project, we used Pandas and Matplotlib libraries. Pandas is an open-source Python package that is most widely used for various data analysis and machine learning tasks as it has functions for reading, analyzing, cleaning, exploring, manipulating, and loading data sets in many formats very easily (Tutorialspoint, 2021). It is built on top of a package named NumPy, which provides support for multi-dimensional arrays, and it provides two popular data structures for manipulating data. They are the Series, DataFrame. Series is a one-dimensional array holding data of any type and a DataFrame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns. We used Pandas DataFrame data structure to work on our analysis and leveraged many functions of it to achieve the results. Pandas is extremely fast, flexible, and easy to use for data analysis. (Geeksforgeeks, 2023).

Matplotlib is a python library used to create 2D graphs and plots using python scripts. It has a module named ‘pyplot’ which makes things easy for plotting by providing features to control line styles, font properties, formatting axes etc. It supports a very wide variety of graphs. It is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002. Matplotlib consists of several plots like line, bar, scatter, histogram etc. Matplotlib comes with a wide variety of plots that help us to understand trends and patterns, and to make correlations. (Geeksforgeeks, 2018). Seaborn is a data visualization library based on Matplotlib that provides a high-level interface for creating informative and attractive statistical graphics. It is built on top of Matplotlib and is designed to work with Panda’s data structures, making it easy to visualize data from Panda’s data frames and series. It provides visualization functions for matrix plots, time series plots, regression plots, etc. (Simplilearn, 2023)

**Problem analysis**

***Question 1*** *- What location of breached Information is most frequent overall? Additionally, do public or private institutions have had more cybersecurity breaches?*

From our data, some common sources of breaches includes laptop, desktop computers, network servers, paper, emails, electronic medical records, portable electronic devices and others. To figure out which source of breach was the most frequent, we used a for loop to iterate through each location in and for each location, check whether it already exists as a key in the ‘locationDictionary’. If it does not, we creates a new key for that location and set its value to 1, indicating that this is the first time that location has been mentioned in a breach. If the location already exists as a key in the dictionary, it simply increments the value associated with that key by 1. With this code, we deduced that Laptop is the leading source of cybersecurity breaches with it occurring 257 times, then Paper, others, desktop computers, portable electronic devices, network servers, emails, electronic medical record, in that order (Appendix 1A). This result shows that the healthcare industry needs to implement access control to restrict access to sensitive information only to authorized individuals, as well as providing training and resources to employees on protecting sensitive information. Another method for increased security is to use data encryption and password protected computers when storing customer data to prevent data breaches.

After analyzing the location of the breached information, we now wanted to take a closer look at the number of individuals affected depending on whether the institution where the breach took place is public or private. We noticed that there are specific keywords that all the private entities had as value for the column “Name\_of\_Covered\_Entity”. We therefore created a list with every keyword, named “private\_keywords”. After that, we used a for loop that runs through every object (data breach) inside the “allDataBreaches” dictionary. By using the “intersection()” inside of an if statement to check whether one word inside the “Name\_of\_Covered\_Entity” instance is the same as one of the keywords of the “private\_keywords” dictionary and assign the consequent number of individuals affected to the the “public” or “private” key respectively. Our findings show that public entities had a substantial higher amount of cyber security breaches (19 million) compared to private entities (13 million) (Appendix 1B). This could potentially indicate that public entities may have weaker cybersecurity measures in place, making them more vulnerable to attacks. This conclusion is crucial for customers, particularly to those enrolled in public clinics and hospitals, since data breaches in these entities account for 60% of the total data breaches of our dataset. This means that if customers want to have their data better protected, it may be in their interest to switch to private healthcare institutions, even though it may be at a higher cost.

***Question 2*** *– What type of breach (Type\_of\_Breach column) was the most common one in every single year (1997, 2002-2014) and what are the top 3 locations of every single type of breach that happened (Location\_of\_Breached\_Information)?*

After importing the .csv file we had to count and select the most frequent type of breach every year and output it. This shows us that the most common breach was theft. It happened 8 out of 14 times (Appendix 2A). This information tells us that the most significant security issue that institutions, businesses, and individuals face is stolen information which is one of the most lucrative goods on the black market. The theft of information can compromise not only an individual, but also a whole business and it can cost millions of dollars to fix.

The second part of our problem is finding out what are the top 3 locations where every single type of breach happened. This means finding where the institutions are most vulnerable. By using the Location\_of\_Breached\_Information and Type\_of\_Breach columnwe were able to find and count the locations where theft, loss of information, hacking, and other types of breaches have occurred (Appendix 2B). This helps us figure out where the institutions have lacked security and what others need to pay attention to in order to prevent any security breaches. The data shows that institutions/organizations are most vulnerable to a lack of security on laptops, desktop computers, emails, and other electronic devices (Appendix 2B). This is a reasonable finding considering that most of today’s interactions are conducted through electronic devices and email. Despite having millions of dollars invested in the security of devices and systems there is always a risk of breaches happening. The second interesting finding is that a lot of breaches happen through paper, meaning that information is stolen or lost through theft or damage of physical records within institutions/businesses or organizations. This surprising revelation allows for better understanding of where the threats can come from and how to anticipate them.

***Question 3*** *– What year had the highest number of individuals affected and what type of breach was the most common for that year?*

For this question the panda and matplot libraries were used. Pandas were used to create a data frame of the columns from the csv file we used. The columns used for this question were Individuals\_Affected, Date\_of\_Breach, year, and Type\_of\_Breach. First step in this question was determining what year had the highest number of individual affected. From the code we wrote we were able to deduce that 2011 had the highest number of individuals affected from cyber security breaches with more than 11 million victims (Appendix 3A). Then to find out what kind of breach was the most common for that year we had to set the row index to year == 2011 and then we grouped it by type of breach to determine which breach was the most common for that year. The most common breach for that year was theft with 116 incidents of it occurring (Appendix 3B). This raised another question of which breach had the highest value for individuals affected. For this we had to make sure the value of individuals affected was summarized by type of breach for the year 2011. The output of that code showed that loss of data was the breach in 2011 that had affected most people with a bit over 6 million victims (Appendix 3C). Theft only accounted for 2.85 million victims although it was the most common breach for 2011.

**Conclusion**

After analyzing the data breaches in the healthcare industry, it is clear that there is a need for improved cybersecurity measures. From our analysis, laptops were found to be the leading source of cybersecurity breaches, followed by paper, desktop computers, portable electronic devices, network servers, emails, and electronic medical records. It is important for the healthcare industry to implement access control to restrict access to sensitive information to only authorized individuals, as well as provide training and resources to employees on protecting sensitive information. The use of data encryption and password-protected computers when storing customer data is also crucial to prevent data breaches.

We also found that public entities had a substantially higher number of cyber security breaches compared to private entities, with public entities accounting for 60% of the total data breaches in our dataset. This suggests that public entities may have weaker cybersecurity measures in place, making them more vulnerable to attacks. Customers, particularly those enrolled in public clinics and hospitals, should be aware of the potential risks and may want to consider switching to private healthcare institutions for better data protection, even though it may come at a higher cost.

Additionally, our analysis of the dataset shows that theft is the most frequent type of breach, occurring 8 out of 14 times. This highlights the significant security issue that institutions, businesses, and individuals face in terms of stolen information, which can compromise not only an individual, but also a whole business and cost millions of dollars to fix.

Furthermore, we found that the year 2011 had the highest number of individuals affected by cybersecurity breaches, with over 11 million victims. The most common type of breach for that year was theft, with 116 incidents of it occurring. However, the breach that affected the most people in 2011 was loss of data, with over 6 million victims, even though it was not the most common type of breach. These findings highlight the importance of regularly reviewing and improving cybersecurity measures to protect sensitive data from breaches.

In conclusion, our analysis of the healthcare industry's cybersecurity breaches underscores the need for better data protection measures. The value of our project is such that it provides evidence that the healthcare industry must prioritize cybersecurity by implementing access control, providing employee training and resources, and using data encryption and password-protected computers to prevent data breaches. Public entities should take particular care to improve their cybersecurity measures to prevent further breaches, and customers should be informed of the risks and consider switching to private healthcare institutions for better data protection. Finally, regular reviews and improvements of cybersecurity measures are necessary to stay ahead of potential threats and to protect sensitive data from breaches.

**Works cited**

Aluko-Sayo, E. (2021). Cyber Security Breaches Data. Kaggle. Retrieved from <https://www.kaggle.com/datasets/alukosayoenoch/cyber-security-breaches-data>

Devara, G. (2021, November 18). *Why do we use Pandas in python*. Tutorials Point. Retrieved March 7, 2023, from <https://www.tutorialspoint.com/why-do-we-use-pandas-in-python>

Aggarwal, N. (2023, February 9). *Introduction to pandas in Python*. GeeksforGeeks. Retrieved March 7, 2023, from <https://www.geeksforgeeks.org/introduction-to-pandas-in-python/>

Kattamuri, M. (2018, May 14). *Python: Introduction to matplotlib*. GeeksforGeeks. Retrieved March 7, 2023, from <https://www.geeksforgeeks.org/python-introduction-matplotlib/>

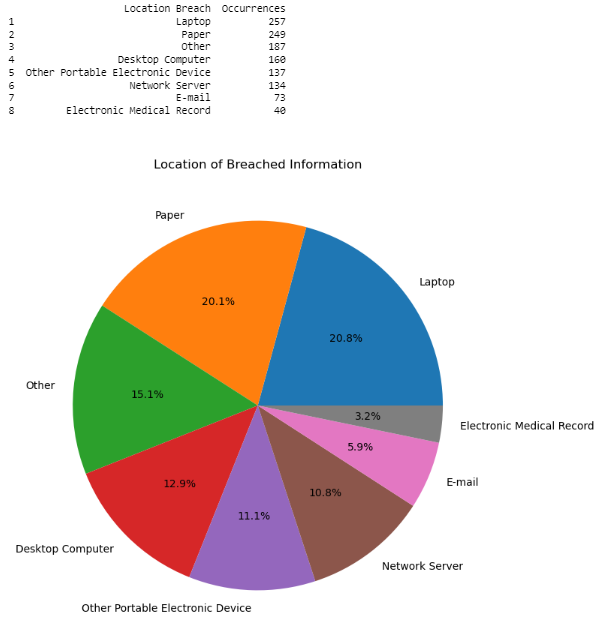
S, R. A. (2023, January 27). *What is python seaborn: Multiple plots & examples: Simplilearn*. Simplilearn.com. Retrieved March 9, 2023, from <https://www.simplilearn.com/tutorials/python-tutorial/python-seaborn>

**Appendices:**

**Appendix 1A**

**Code, Data Frame, and Pie Chart showing the percentages of individuals affected by the location of breach**

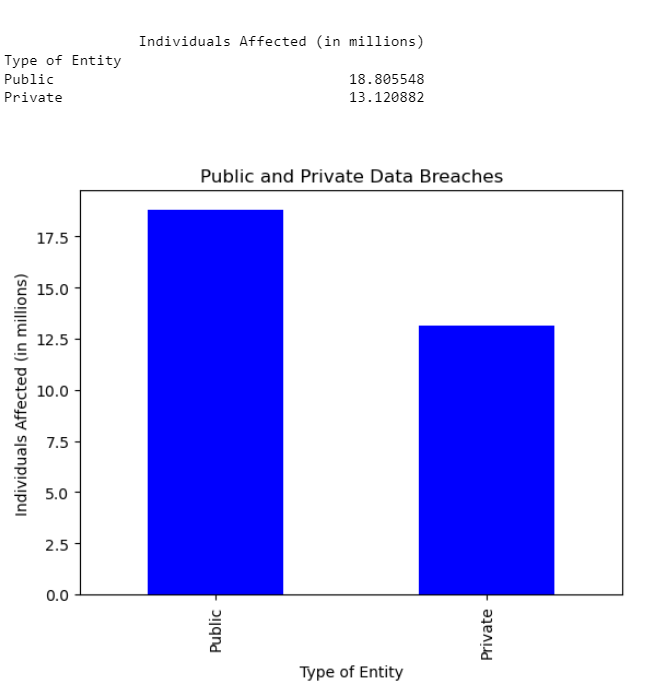




**Appendix 1B**

**Code, Data Frame, and Bar Graph for individuals affected sorted by private and public entities**

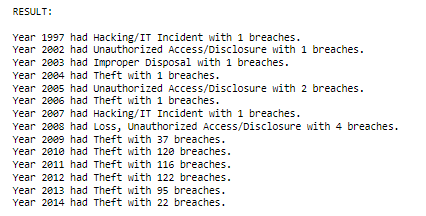


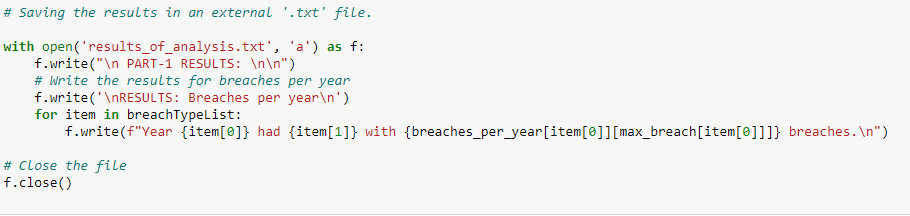


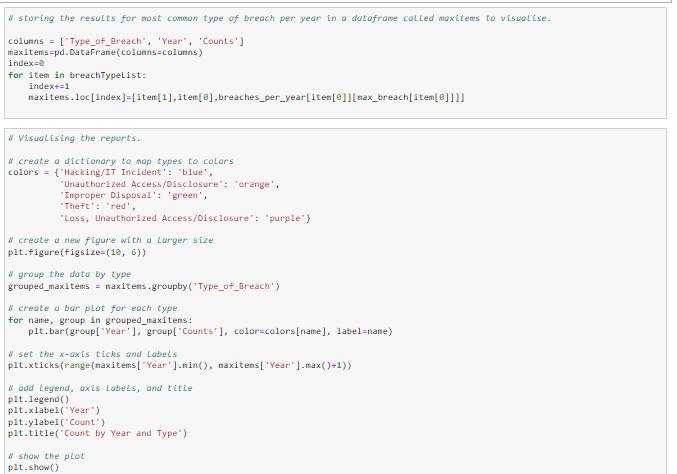
**Appendix 2A**

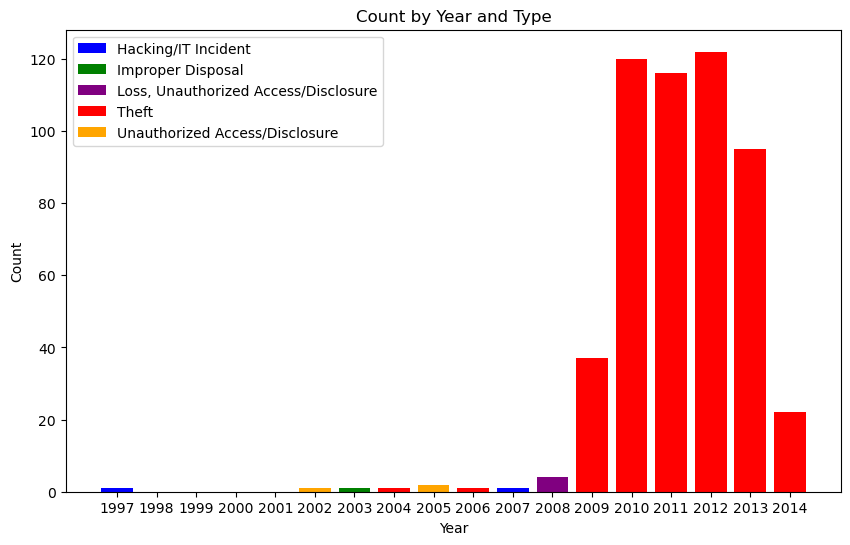
**Code and Bar Graph for most common type of breach in every single year**





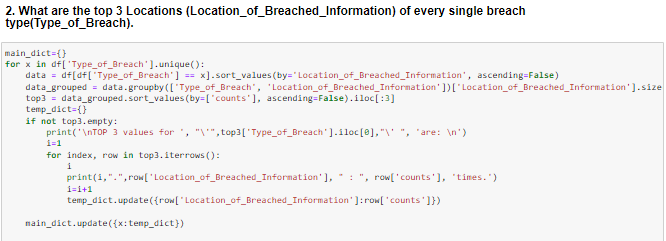


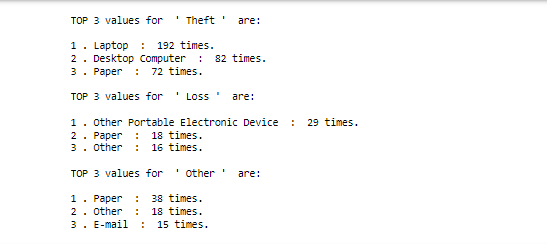


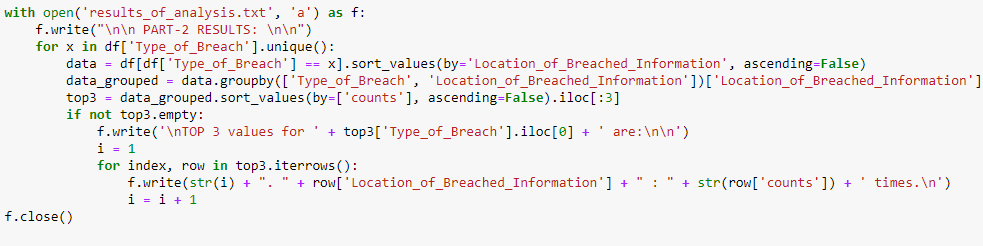


**Appendix 2B**

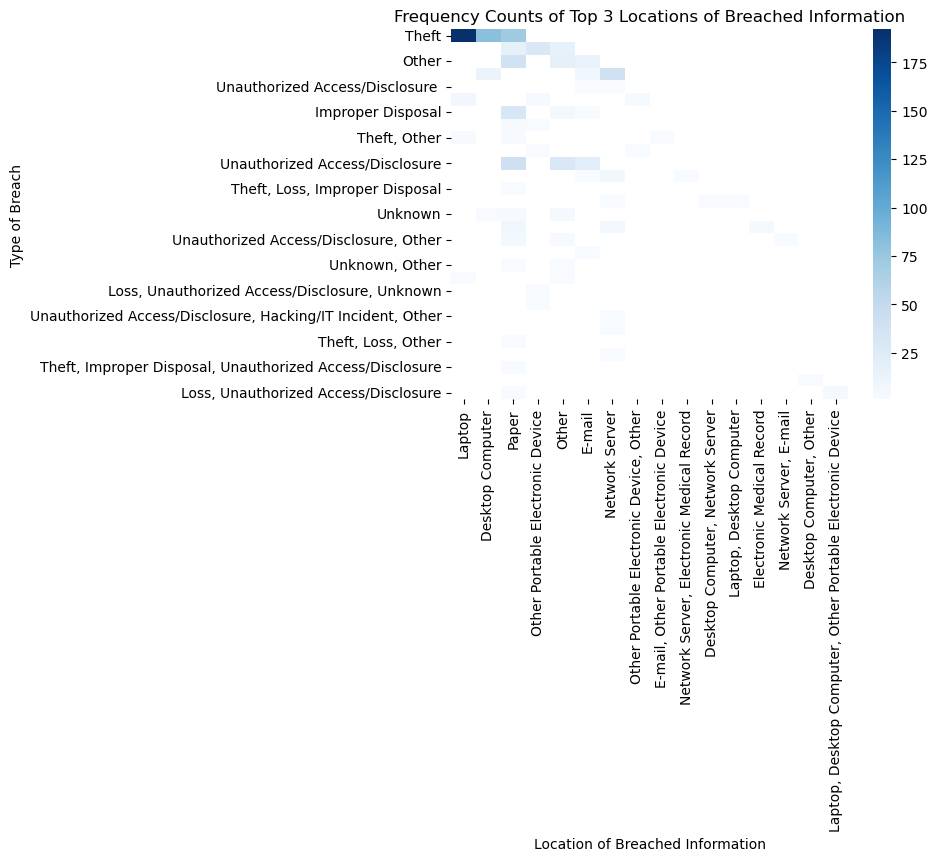
**Code and Heat map for top 3 locations of breach for every single type of breach**





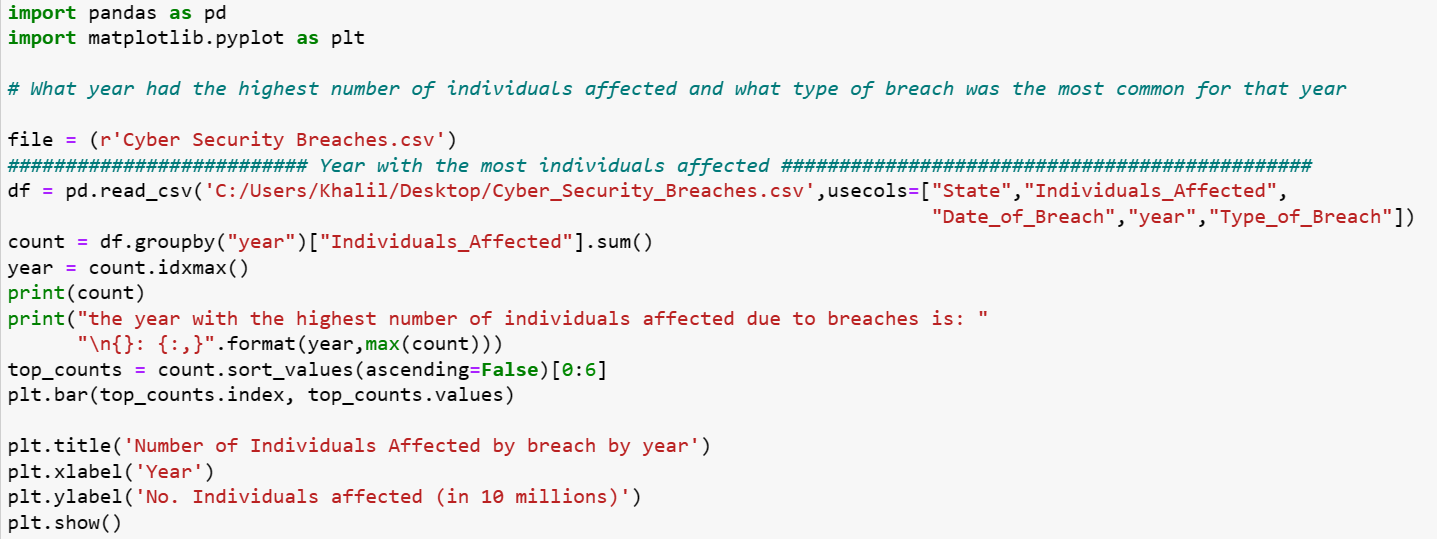


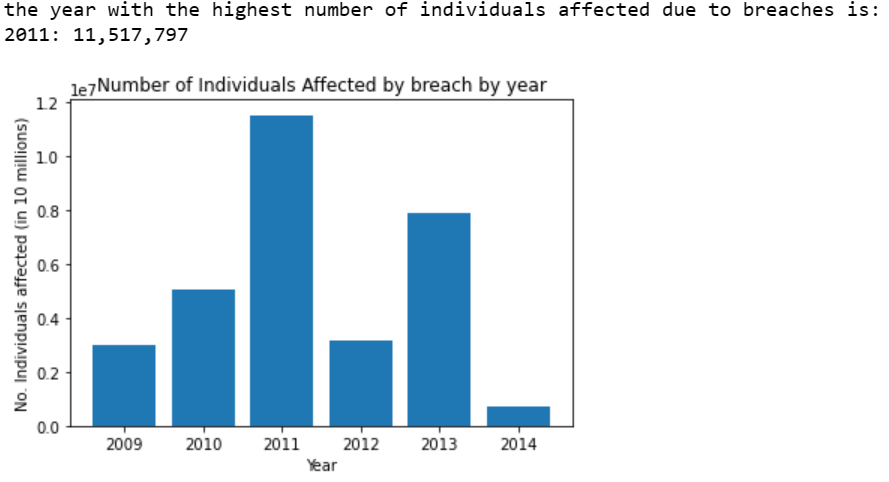




**Appendix 3A**

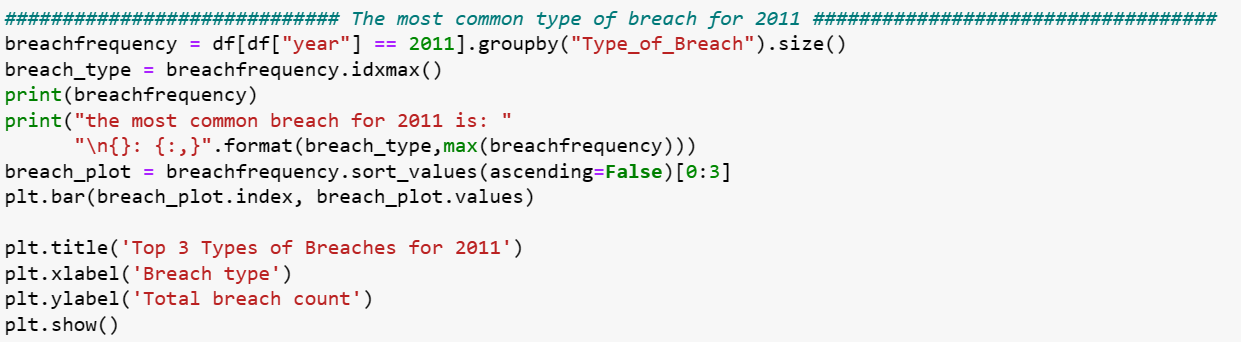
**Code and Histogram For Year With Most Individuals Affected**

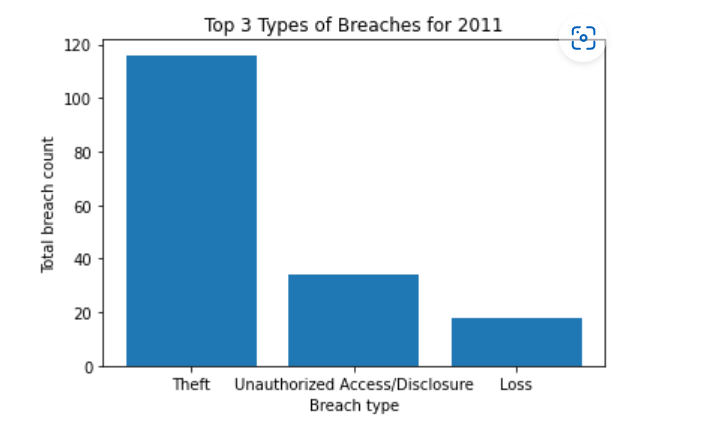




**Appendix 3B**

**Code and Histogram For Most common Breach in 2011**





**Appendix 3C**

**Code and Histogram For breach with most individuals affected in 2011**



