

CIS1006-Secure data acquisition

Report



March 16, 2022

Teesside university

Middlesbrough TS1 3BX

|  |  |
| --- | --- |
| Submitted by: | Md Monajir Hussain (B1393756) |
| Submitted to: | Dr. Zia Ush Shamszaman  Mr. Lewis Golightly |

Table of contents

Introduction…………………………………………………………………………………………………………………………

Major findings………………………………………………………………………………………………………………………

Challenges……………………………………………………………………………………………………………………………

Presentation and analysis of the results?.................................................

Were you surprised at any of the results?...............................................................................

References.................................................................................................................................

Introduction

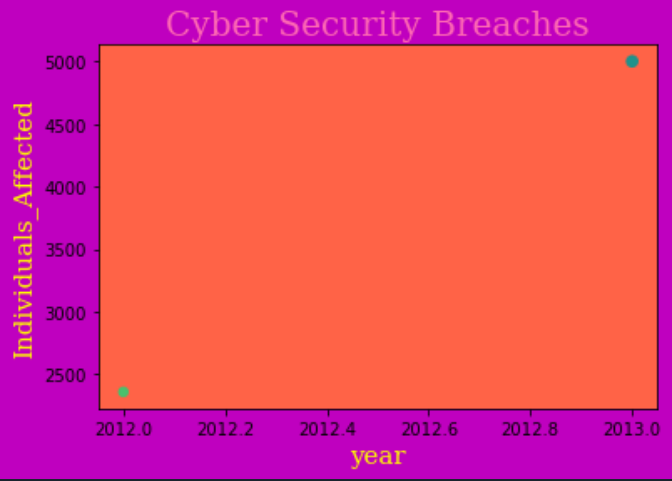
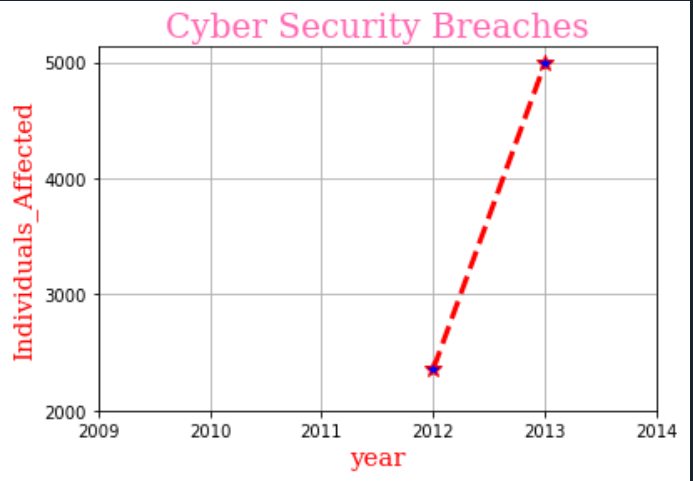
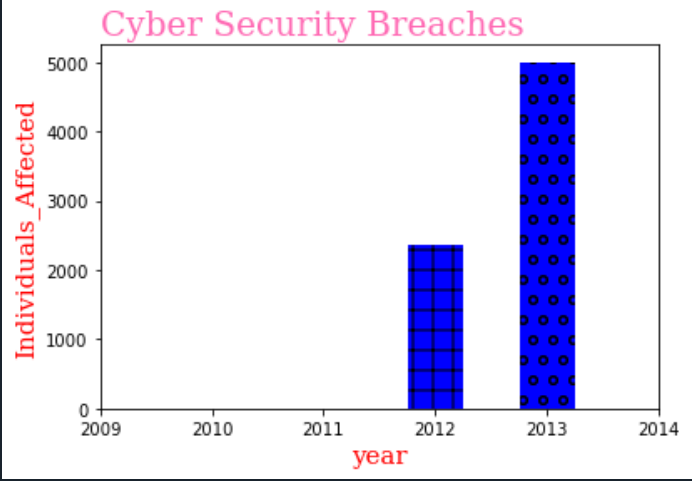
The portion of dataset from the KDD Cup Data from 1999, (International Knowledge Discovery and Data Mining Tools Competition) is selected for the data analysis. The dataset is cleaned properly to make it ready for the analysis using Python and pandas.

The dataset from the KDD Cup Data is commonly used for educational and research purposes, despite its age. This is the data set used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99 The Fifth International Conference on Knowledge Discovery and Data Mining (KDD Cup 1999 Data, 1999).

For the data analysis the small portion of the dataset is taken and cleaned for visualization. Python was used and the Spyder interpreter was applied.

Major Findings

Analysing more than thousands of data using programming language can be challenging, however manually doing it can be 10 times more challenging. By cleaning the data followed by visualization, it can be seen that most of the cyber security breaches took place in the year 2012 and 2013. The figures of the affected individuals range from 2500 to 5000 in those respective years. In addition, there was an upper trend of the breach as illustrated in the graphs.



Bar Graph Line graph Scatter Graph

With regards to the data analysis using Python, the fact that stands out is that thousands of data, even if not in format, can be visualised and made easy for the users.

However, the dataset was statically analysed, and the analysis showed that there were some flaws regarding the dataset. There are some critical issues in the dataset which highly affect the performance of evaluated systems, and results in an extremely poor evaluation of the anomaly detection approach. It includes few redundant data in the train set which makes the classifiers biased towards more frequent records (Mahbod, Ebrahim, Lu and Ali, 2009).

Challenges

Purifying thousands of data with various redundancies can be challenging sometimes. In the context of visualizing a portion of data from KDD CUP <http://kdd.ics.uci.edu/databases/kddcup99/kddcup.data_10_percent.gz> , there were few challenges tackled which were solved using some python properties whereas some of them remain unchanged. To use these Python properties for cleaning and analysing, Panda and NumPy packages or libraries can be leveraged so it should be installed and imported to use python properties. The major challenge faced during data analysis process, was data cleaning. In fact, a lot of data scientists argue that the initial steps of obtaining, and cleaning data constitute 80% of the job (Agarwal, 2017).

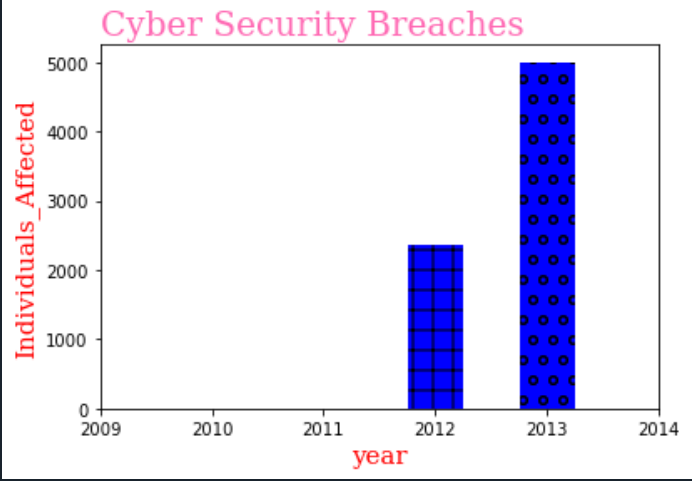
There are a few issues associated with data cleaning, the columns with unknown or no values were processed for data cleaning and then passed for visualization. When retrieving these data, it stood out that there were some faults in the cleaning process which resulted in inaccurate data analysis. There are data from the year 2009 to 2014 where only data of 2012 and 2013 were plotted. In conclusion, it stands out that it went sideways when retrieving a mean for the data and sorting the year column at the same time.

In addition, there were efforts made for passing the two columns namely breach\_start and breach\_end, for visualization along with the other two successfully visualized columns but it was not being plotted instead made the analysing process dysfunctional.

Therefore, some other minor challenges came in the way but solved using different learning platforms and YouTube tutorials.

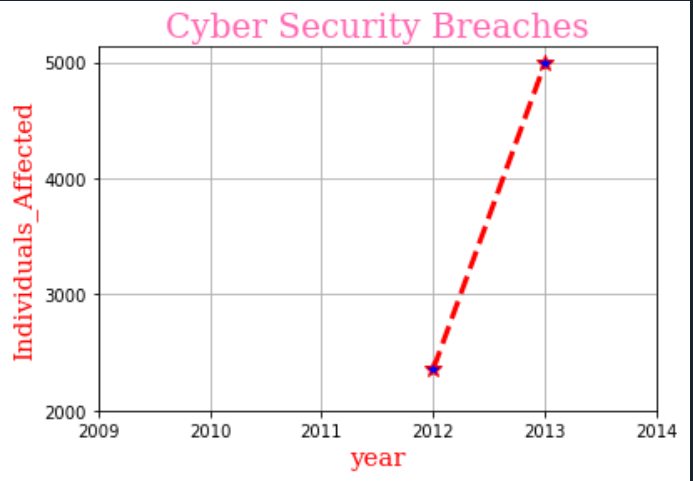
Presentation and analysis of the results

Presentation of the data plays a major role in data visualisation as it shapes the final output of the whole analysis. For visualization, necessary package ‘matplotlib’ should be imported. Three graphs are plotted for better data analysis, along with some general customizations for making it visually appealing. The two major columns were processed for visualisation namely ‘year’ and ‘Individuals\_Affected’. The title on the graphs are added using python function called plt.title and few customizations can be done with it using loc and fontdict.



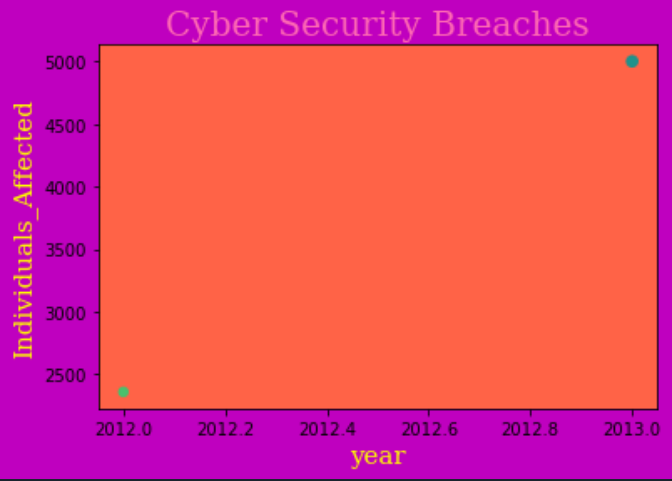
Bar Graph

In the bar graph, the data is plotted with ‘year’ on the x-axis and Individuals\_Affected affected on the y-axis using plt.xlabel and plt.ylabel function respectively. The range of the axis can be defined using plt.xticks function. For adding patterns to the plot, we can use set\_hatch functions.



Line Graph

In the Line graph, the data is plotted with ‘year’ on the x-axis and Individuals\_Affected affected on the y-axis using plt.xlabel and plt.ylabel function respectively. The graph is customized using various fonts and colours available in matplotlib package. The graph is also grid by using plt.grid(True) function. The line in the line graph is made visually appealing by using few personalized functions like linewidth, linestyle, marker and so on using legal values.



Scatter Graph

In the scatter graph, the whole graph is made unique although the data plotted gives the same information. The size of the graph is maintained along with the colour. The data is plotted with ‘year’ on the x-axis and Individuals\_Affected affected on the y-axis using plt.xlabel and plt.ylabel function respectively. The background is coloured with magenta colour using fig.patch.set\_facecolor('m') and the patch where the data are highlighted is coloured using ax.set(facecolor = "tomato").

Were you surprised at any of the results?

Having little to no knowledge about programming, I was really amazed to have the thousands of data which were not in structure being plotted in the various graph graphs. The data were not in format, some of the data were corrupted whereas some had unknown values which were cleaned and passed for visualization.

It was really fascinating to see small function of python doing big tasks, that if done manually would take days. Even slightest of functions can make great changes and having the proper skill to utilize them can help in doing almost every data retrieving tasks possible.

Therefore, it proves that the world of data visualization has exploded. From arc diagrams and bullet charts to violin plots and waterfalls, there are more techniques than ever before (qlik, no date).

Not to miss, I was also surprised to see the other columns not being plotted properly. Also, it was somewhat upsetting to see the data being plotted inaccurately after it was cleaned, which is the problem in the cleaning of data

References

KDD Cup 1999 Data (October 28, 1999) Available at: <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html> (Accessed: 12/03/3033)

CISDA (2009) A Detailed Analysis of the KDD CUP 99 Data Set. Available at: <https://www.ee.ryerson.ca/~bagheri/papers/cisda.pdf> (Accessed: 12/03/3033)

Malay Agarwal (2017) Real Python: Pythonic Data Cleaning With Pandas and NumPy. Available at: <https://realpython.com/python-data-cleaning-numpy-pandas/> (Accessed: 13/03/3033)

Qlik (no date) 10 Ways to Take Your Data Visualizations to the Next Level. Available at: <https://www.qlik.com/us/lp/sem/10-ways-to-take-your-data-visualizations-to-the-next-level?utm_team=DIG&utm_subtype=cpc_nb&ppc_id=f6PAE1gA&kw=visualizing%20data&utm_content=f6PAE1gA_pcrid_547089939001_pmt_p_pkw_visualizing%20data_pdv_c_mslid__pgrid_127139118173_ptaid_kwd-136226542&utm_source=google&utm_medium=cpc&utm_campaign=Qlik_UnitedKingdom_Google_NB_DA_Dashboard_EN&utm_term=visualizing%20data&gclid=EAIaIQobChMIicLwzvzK9gIVQu3tCh3urwCoEAAYASAAEgKkJPD_BwE> (Accessed: 15/03/3033)