Analysis of Outbreaks in Toronto Healthcare Institutions

dataset provided by the City of Toronto

Written by:
Layne Moran Atencio
Cameron Milligan
Atef Alqashqish

Abstract

Group 9 has chosen to study the Outbreaks in Toronto Healthcare Institutions dataset provided by the City of Toronto. The dataset is available publicly at City of Toronto website (URL). The dataset covers incidents of diseases outbreak from January 3rd, 2016 up to March 27, 2019.

Objectives

The objective of the analysis is to conduct a data analysis, uncover key statistical features and insights about the data set. The source of data was an open data set provided by the City of Toronto. The data was available in multiple batches based on its year beginning in 2016. The 2019 data is updating weekly.

Specifically, our goal was to answer the questions: What is the most prevalent disease in the dataset? What are the most prevalent diseases based on the outbreak setting? Which disease occurs most frequently? Which disease had the longest duration? Is there a seasonality trend in the data? If so, what is the seasonality? Lastly, we wanted to conduct a simple forecast of 2019's potential outbreaks.

Data Preparation

The dataset was available in an excel format and accessible by a URL on the City of Toronto website. We brought the data into our model by utilizing the pd.read_excel() function on the URL of each data set and were able to stream the data directly into pandas dataframes.

The quality of the data was strong, in total we had over 1200 rows spanning multiple years. This meant we had sufficient quantity for analysis. Upon examination of the data we noticed several ways in which the data could be cleaned and presented in a more useful manner.

As expected with a dataset on disease outbreaks there were three columns containing information on the Causative Agents. One of the columns, Causative Agent-2 only had 121 non null values compared to over 1200 rows. Similarly, the Causative Agent column only contained 268 non null values. The third column, Causative Agent-1, contained 976 non null values which indicated to us that if we consolidated all three columns we would have sufficient data for analysis. They were consolidated using a simple join and by dropping NaN values.

We also noticed the Etiological Agent2 column only contained 33 non-null values, so this column was dropped as it could not be used in a statistically significant way.

To make further use of the newly consolidated disease column, we wrote a function called extract_dease_name which takes the consolidated disease column as an input and outputs the disease name, type, and subtype. For example the disease Influenza A would become disease

name influenza and type A. This enabled us to make a broader analysis on the grouping of the disease without being overly granular on the specific type and subtype.

Included in the dataset were columns containing institution names and addresses of the locations where outbreaks occurred. This data needed to be cleaned up because the Institution name also contained floor numbers or hospital wings. This additional information made it less easy to perform grouping analysis. By striping all information after the dashes we were able to clean this column.

The last area for data preparation was to create a disease duration column. The dataset included Date Outbreak Began and Date Declared Over columns. These columns were datetime values but the date declared over column also contained NaT values. We cleaned these NaT values by replacing with the current system date. The reason these cells were NaT was because the disease outbreak was still active. We still wanted to include active outbreaks in our analysis. By subtracting the begin date from the date declared over we created a Duration column.

Analysis and Conclusions

What is the most prevalent disease in the dataset?

The most prevalent diseases across the Data Set are: **Influenza** at about **43.0%** followed by **norovirus-like** at **10.0%**.

Disease	Count
influenza	530
unable to identify	225
norovirus-like	121
respiratory	103
rhinovirus	74

Which disease occurs most frequently?

- The top 3 diseases occurred most frequently assuming 'daily' frequency are:
    - influenza, occurred 288 times out of 1180 at about 24.0%

    - norovirus-like, occurred 104 times out of 1180 at about 9.0%
    - respiratory, occurred 86 times out of 1180 at about 7.0%
- The top 3 diseases occurred most frequently assuming 'weekly' frequency are:
    - influenza, occurred 91 times out of 170 at about 54.0%
    - norovirus-like, occurred 62 times out of 170 at about 36.0%
    - respiratory, occurred 49 times out of 170 at about 29.0%
- The top 3 diseases occurred most frequently assuming 'monthly' frequency are:
    - influenza, occurred 27 times out of 39 at about 69.0%
    - parainfluenza, occurred 26 times out of 39 at about 67.0%
    - norovirus-like, occurred 25 times out of 39 at about 64.0%

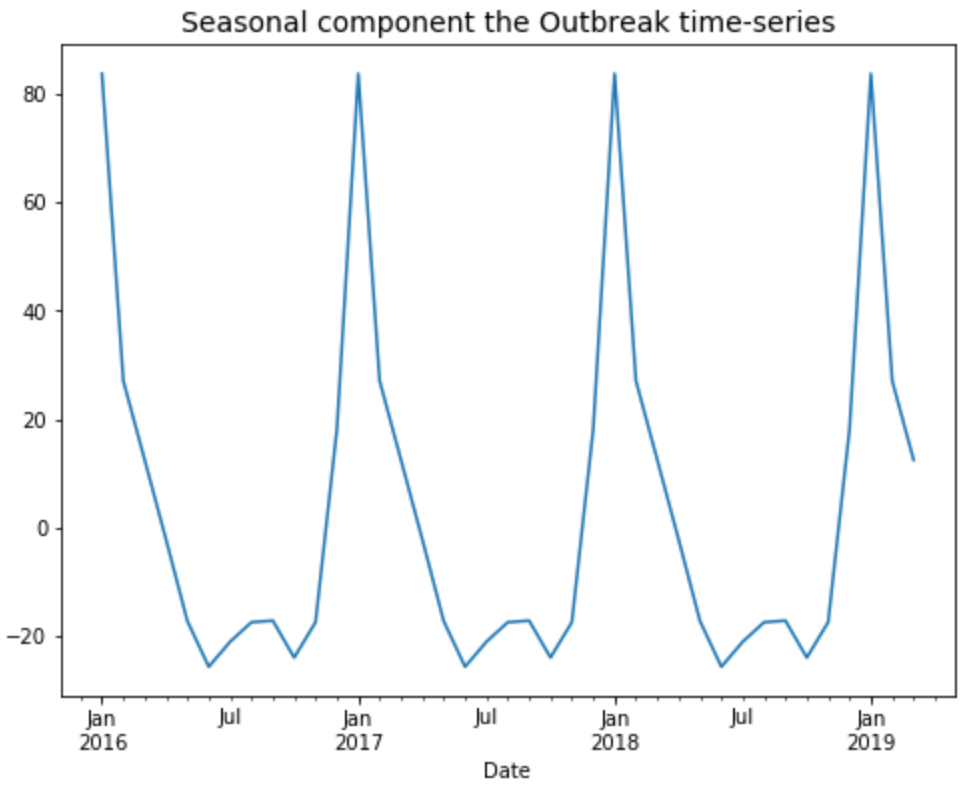
Which disease had the longest duration?

The average outbreak is 14 days with 75% of the data falling under 16 days. The longest outbreak was a streptococcus outbreak that lasted 578 days in a Shelter. This indicates that the 578 day outbreak was also a significant outlier in the dataset.

Is there a seasonality trend in the data? If so, what is the seasonality?

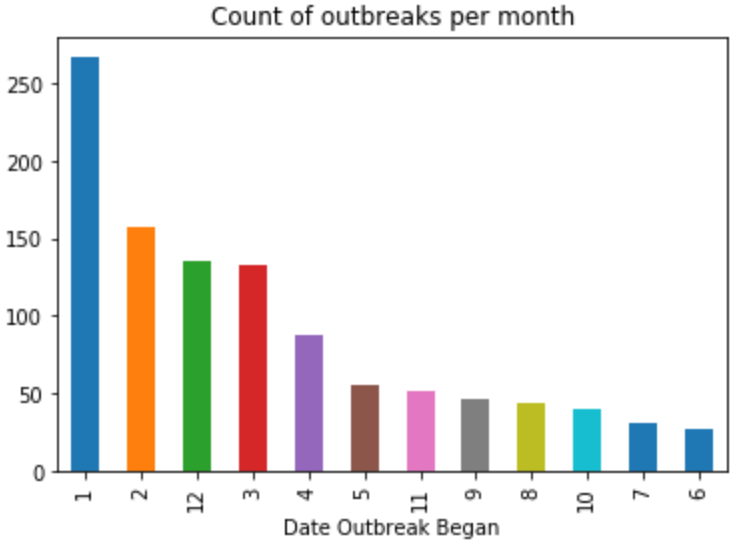
Another important question regarding our dataset was whether or not there was a seasonality component. Since our data is time series in nature we have the perfect opportunity to detect a seasonal trend. Our initial hypothesis is that there likely is a seasonality component since there is a so called “Flu Season”. We wanted to see if the data actually supported such a thing.

By using the statsmodel package and the decompose function we were able to decompose the dataset and uncover a seasonal trend. This trend is plotted below. As expected, there is a seasonality peak in outbreaks for the winter months, specifically January.



Which months have the most outbreaks?

After analyzing the seasonality we wanted to verify which months actually had the most outbreaks. The seasonality indicates it should be the winter months. To do this we summed the counts of outbreaks grouped by month for 2016 through 2018. Since 2019 is a partial year, we excluded those data points. As anticipated, the winter months January, February, and December experienced the highest number of outbreaks.



What would a simple forecast of 2019's potential outbreaks look like?

One method of forecasting is to do a naive forecast by shifting 2018's values forward 12 months and plotting them. Since we already have actual data for January through March, we only need to forecast April through December. The below chart displays this forecast. Interestingly, it appears the January spike for 2019 was not nearly as pronounced as it was in 2017 and 2018. We do not have enough data to determine the cause, but an interesting hypothesis could be that 2019's flu shot was more effective than in previous years. This would have allowed for fewer outbreaks to occur because immunization rates would be higher.