Alexander Semaan, Erin Tynan, Clara Mae Wells
PM5 Report

Data Links:

1) https://www.kaggle.com/cdc/foodborne-diseases
2) https://data.boston.gov/dataset/food-establishment-inspections
3) https://www.kaggle.com/jqpeng/boston-weather-data-jan-2013-apr-2018

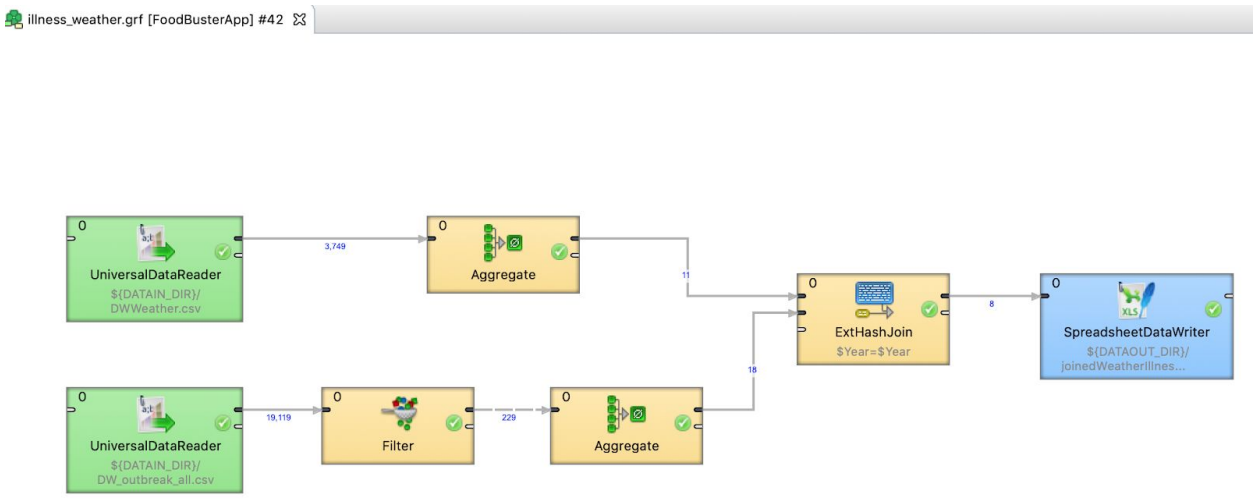
About the Data:

When we had initially started to plan our database and what we would use it for, we had brought in our weather.csv dataset which had multiple data regarding temperature (max, min, avg), humidity, precipitation, weather events, and date. We had cleaned the data in preparation for using it with our database, but didn't get to use it there. Now that we are looking at connecting the data and making hypotheses, we have brought the weather data as an outside data source, in addition to data related to food borne illnesses.

In the Foodborne illness data set, data include year, month, food, location, species, genotype, status, number of illnesses, number of hospitalizations, and number of fatalities for the US. We imported it into Mysql and cleaned the data there. We focused on illnesses rather than fatalities or hospitalizations because the latter two columns had too many blanks for the data to be useful. Illnesses did not have that issue.

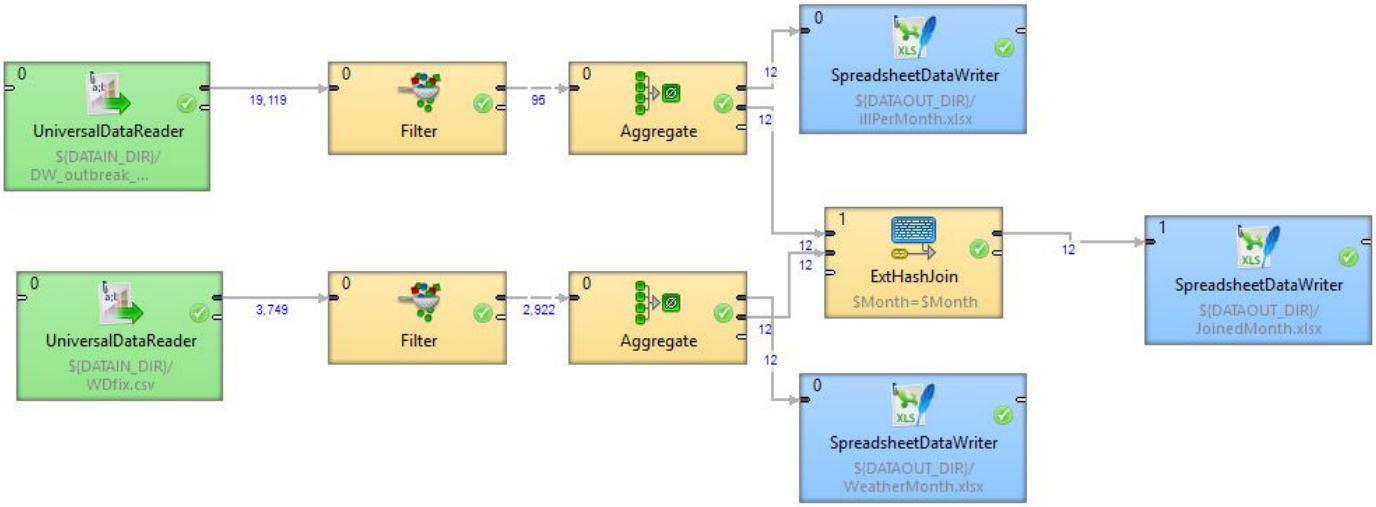
ETLs:

1) **Illnesses_Weather_Per_Year:**
   Two outside data sources are imported, one in each of the UniversalDataReaders, one dataset pertaining to the weather, and the other pertaining to the food illnesses. The outbreak metadata leaves the UDR and is immediately filtered. Since we are looking for data pertaining to one location, we do not want the entire US data. As no city was given in the data, we chose to filter based on state instead. After filtering on Massachusetts, the illness data was aggregated with year and illnesses to give us illnesses per year. For the weather metadata, it was aggregated with a number of columns such as AverageTemperature, Year, AverageHumidity, and Precipitation. This was in anticipation of looking at different items against illnesses. From the aggregators, the metadata are joined in an ExtHashJoin on year in each dataset. Since the data was grouped by year, this left this ETLs data with 8 records. From there, the data were sent to the SpreadSheetDataWriter where it was sent to the data-out folder in an xlsx file. From the file, we created two charts described in the chart section.



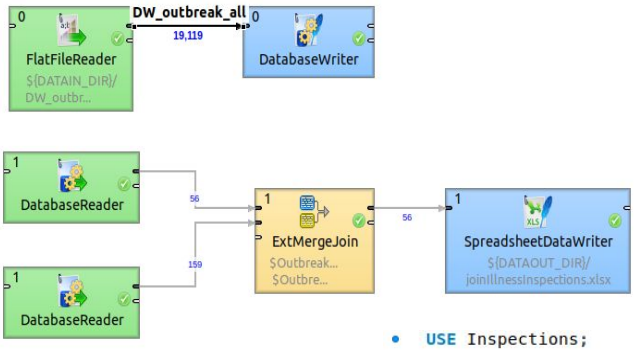
2) Illnesses_Weather_Per_Month

Similarly to the previous ETL, two outside data sources are imported, one in each of the UniversalDataReaders, one dataset pertaining to the weather, and the other pertaining to the food illnesses. The outbreak metadata leaves the UDR and is immediately filtered. Since we are looking for data pertaining to one location, we do not want the entire US data. As no city was given in the data, we chose to filter based on state instead. After filtering on Massachusetts and for years starting 2008 (since the weather data starts in 2008), the illness data was aggregated by month and illnesses to give us illnesses per month, which was saved in a separate spreadsheet in case we need the data separate. For the weather metadata, it was first filtered by year (last year is 2015 for the illnesses data), then it was aggregated with a number of columns such as Months, Temperature, Humidity, Wind and Precipitation. This was in anticipation of looking at different items against illnesses. Same as for the illnesses, we saved the aggregated data using the spreadsheetwriter before joining it in case we needed it separately.
From the aggregators, the metadata is joined in an ExtHashJoin on month in each dataset. Since the data was grouped by month, this left this ETLs data with 12 records. From there, the data was sent to the SpreadSheetDataWriter where it was sent to the data-out folder in an xlsx file. From the file, we created two charts described in the chart section.



Alexander Semaan, Erin Tynan, Clara Mae Wells

3) Inspections_Illness_Monthly



Phase 0:
We used a FlatFileReader (alias UniversalDataReader) to import data from our foodborne disease external dataset to our database. The SQL file that created the table is shown in the bottom right corner of the ETL. The load from the csv to our database yielded 19,119 records, which included disease in various states as well as years beyond the range for which we have food establishment inspection data.

Graph | Source

Problems | Properties | Console | Data Inspector | Regex Tester

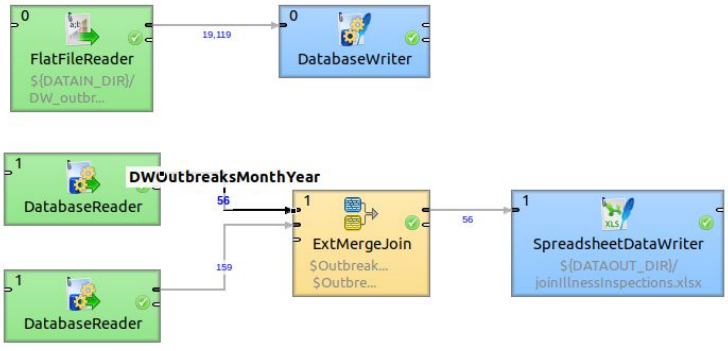
Edge [FlatFileReader → DatabaseWriter] Run ID: 24

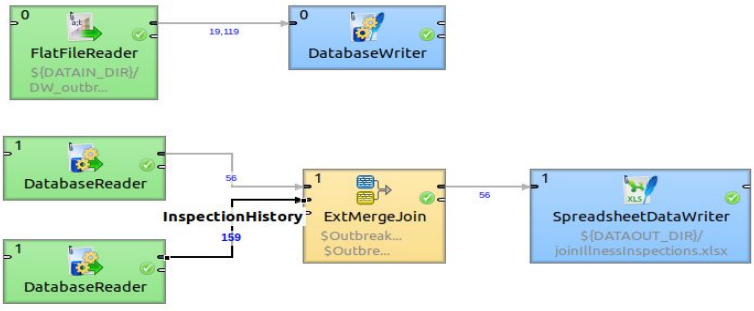
#	Id	Year	Month	State	Illnesses
1	1	1998	January	California	20
2	2	1998	January	California	112
3	3	1998	January	California	35
4	4	1998	January	California	4

```
USE Inspections;

DROP TABLE IF EXISTS DWOutbreaksMonth;

CREATE TABLE IF NOT EXISTS DWOutbreaksMonth (
    IdKey Int AUTO_INCREMENT NOT NULL,
    OutbreakYear Int,
    OutbreakMonth VARCHAR(45),
    Location VARCHAR(45),
    Illnesses Int,
    CONSTRAINT pk_IdKey
        PRIMARY KEY (IdKey)
)
```

Phase 1:

For the next phase of the ETL, we used the DatabaseReader twice. The first reader extracted the formerly external foodborne illness data, and the metadata SQL query selected only years that align with our inspection data records (> 2007 and < 2019). The SQL query also specified where records where location == Massachusetts. While our establishment data is limited to Boston, it is reasonable to examine disease data by state - partly because that is as specific as the data gets but also because movement between cities leads to a reasonable assumption that some illness did not originate in the city in which a person dined. Furthermore, illness in a nearby city can influence actions in another city (e.g. food establishment inspections). This metadata resulted in 56 records.





Graph | Source

Problems | Properties | Console | Data Inspector | Regex Tester

Edge [DatabaseReader → ExtMergeJoin] Run ID: 25

#	Outbr	OutbreakMc	Illnes	Location
1	2008	April	21	Massachusetts
2	2008	August	4	Massachusetts
3	2008	December	40	Massachusetts
4	2008	July	0	Massachusetts

Graph | Source

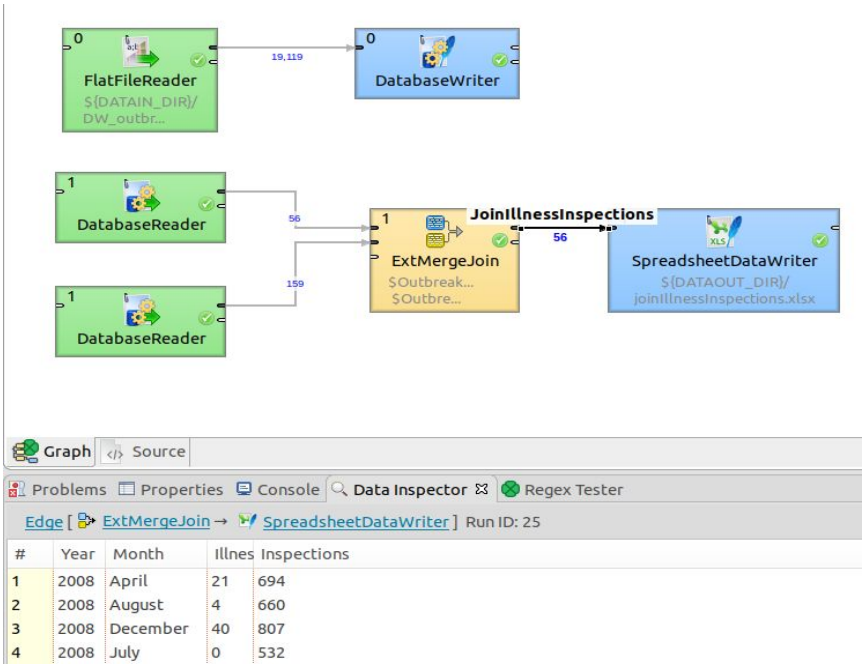
Problems | Properties | Console | Data Inspector | Regex Tester

Edge [DatabaseReader → ExtMergeJoin] Run ID: 24

#	Inspe	InspectionM	InspectionCount
1	null	null	2345
2	2006	April	1
3	2006	August	1
4	2006	December	3

The other DataBaseReader resulted in 159 records. The metadata selected data from an existing relation in our database: InspectionHistory. We used SELECTYEAR(InspectionData), MONTHNAME(InspectionDate), and the count of InspectionResults.

For both DatabaseReader SQL queries, we grouped and ordered by year and month to get sorted monthly inspections. We then used the ExtMergeJoin to produce only the overlapping months and dates of foodborne diseases and food establishment inspections. This resulted in 56 records that were transferred to a SpreadsheetDataWriter.



Graph | Source

Problems | Properties | Console | Data Inspector | Regex Tester

Edge [ExtMergeJoin → SpreadsheetDataWriter] Run ID: 25

#	Year	Month	Illnes	Inspections
1	2008	April	21	694
2	2008	August	4	660
3	2008	December	40	807
4	2008	July	0	532

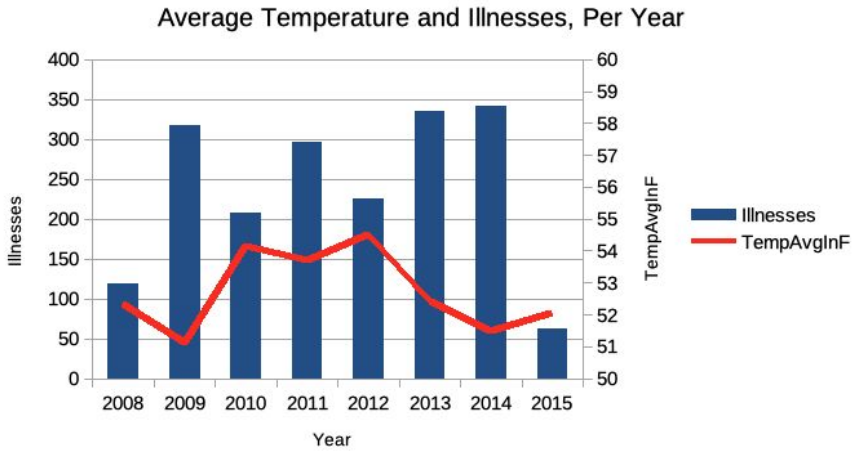
Alexander Semaan, Erin Tynan, Clara Mae Wells

Hypotheses:

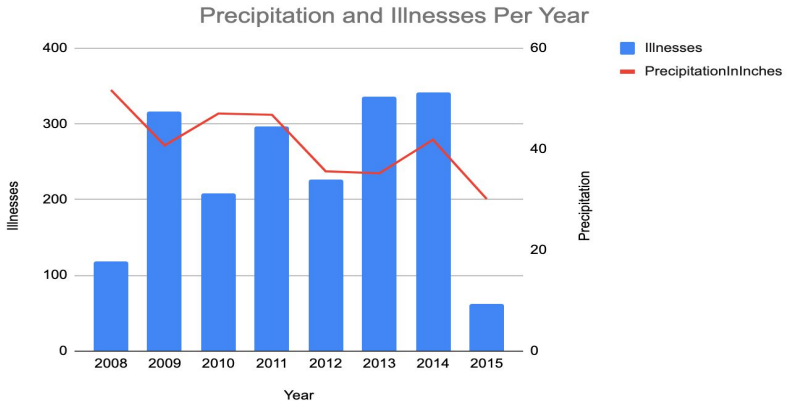
1) Illnesses/Temperature Per Year
    - Illnesses increases when Average Temperature increases.
2) Illnesses/Precipitation Per Year
    - Illnesses increases when Precipitation Per Year increases..
3) On a monthly basis, food-born illnesses increase because of increased air temperature averages, based on the fact that heat reduces produce shelf-life and makes it harder to store.
4) On a monthly basis, food-born illnesses increase because of increased humidity averages, based on the fact that moisture is damaging to the useful life of food and produce.
5) An increase in foodborne illness leads to an increase in food establishment inspections.

Results:

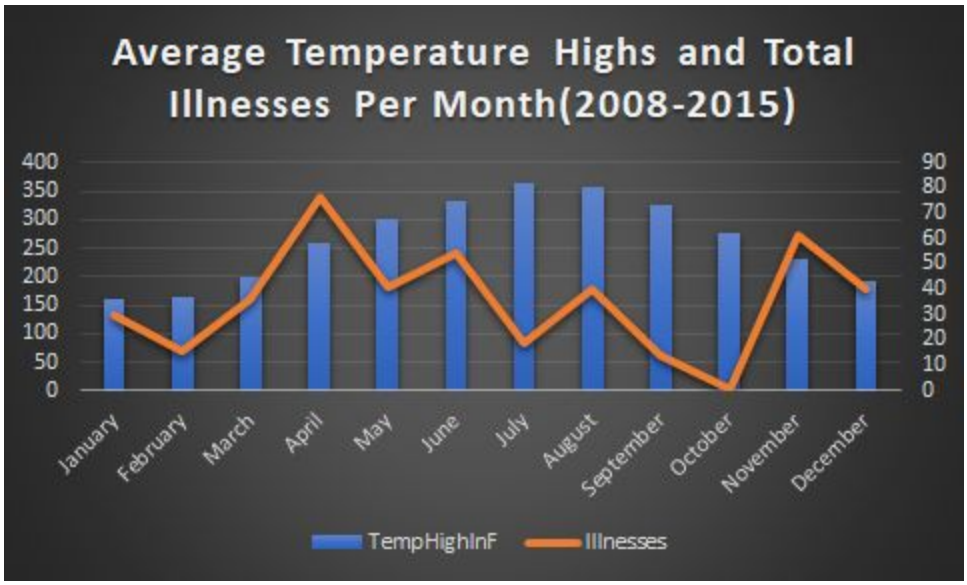
1) Illnesses appeared to be lower when the temperature was higher, and increased when temperature was lower. The opposite of hypothesis 1.



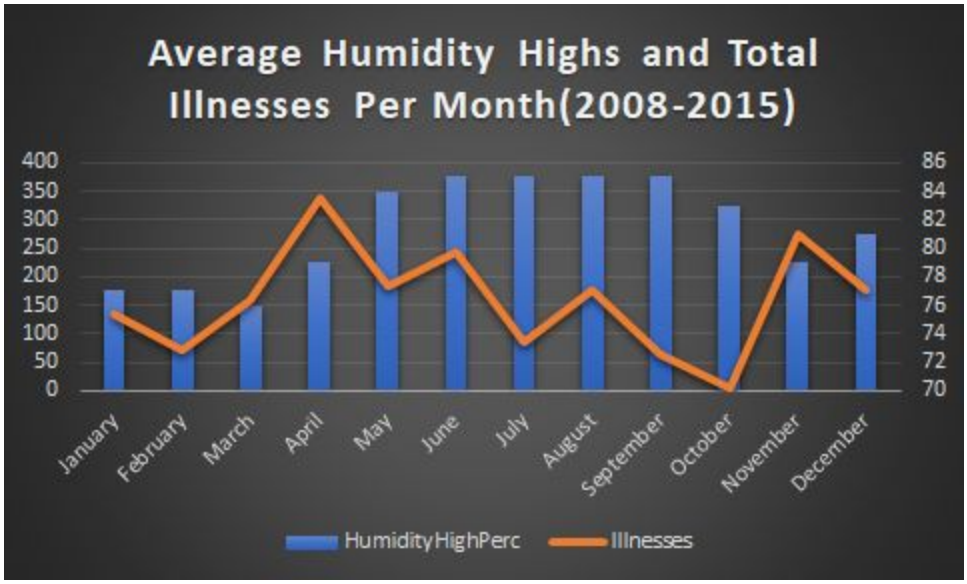
2) Illnesses and precipitation levels do not appear to correlate as much, with the two lowest number of illnesses occurring in both the wettest and least wet years. So Hypothesis 2 did not appear correct.



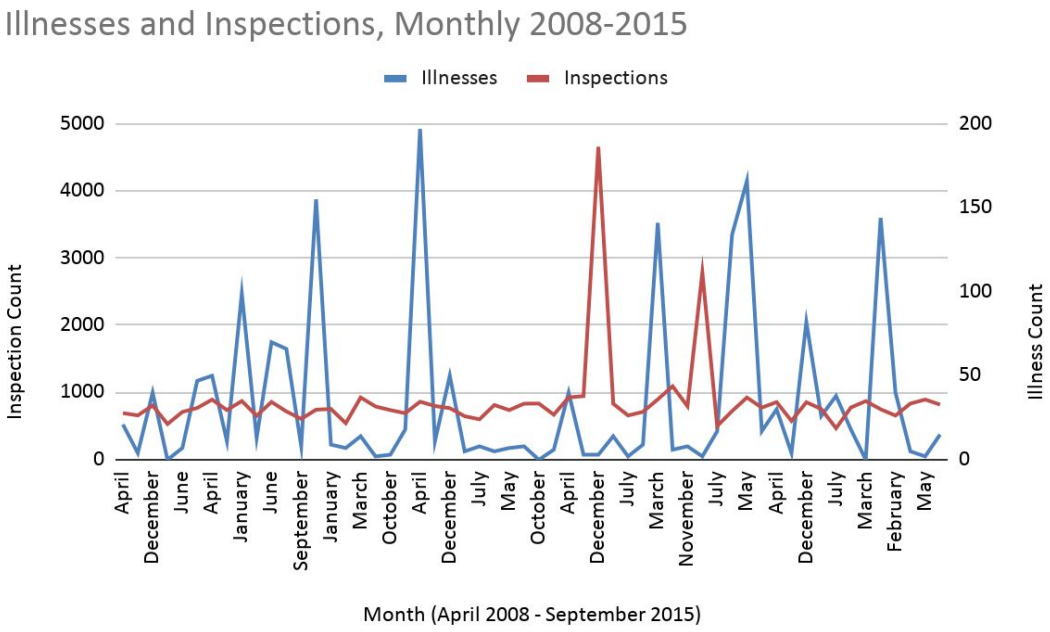
Alexander Semaan, Erin Tynan, Clara Mae Wells



3) By analyzing the chart, we can see that the result is exactly the opposite of what was expected: Food-born illnesses seem to be inversely proportional to increasing average temperatures. At first this might seem odd, but the more you think about it, it would make sense that establishments have a stricter policy and are more careful with the way their products are stored when the weather is warm. The food quality control entity in establishments must be aware of these facts and as a countermeasure they enforce stricter controls to keep products and foods fresh and to keep them from getting spoiled.



4) Very similar to the results of the previous case: we can see that the result is exactly the opposite of what was expected: Food-born illnesses seem to be inversely proportional to increasing average humidity percentages. The food quality control entity in establishments must be aware of the dangers that come with high humidity levels and warm temperatures and as a countermeasure they enforce stricter controls to keep products and foods fresh and to keep them from getting spoiled.



5) We hypothesized that an increase in foodborne illness leads to an increase in food establishment inspections. Some peaks align, which may point to some correlation; however, the data visualization does not show a strong correlation. This leads us to believe that while foodborne illness in the state may play a role in the number of food establishment inspections, something other than foodborne illness primarily drives the number of food establishment inspections. An interesting note is that the most dramatic peak for inspections counts and illnesses counts are over a year apart. The other peaks and dips vary with some seeming fairly correlated while others appearing opposite or unrelated. In light of these results, it may be reassuring for our application users to know that food safety inspectors do not appear to inspect food establishments only in a reactive manner, so they may have a strong internal cadence to promote public health.

Links:

1) Erin Tynan https://vimeo.com/373752430
2) Alexander Semaan https://youtu.be/EWjLlkKInRk
3) Clara Mae Wells https://youtu.be/d26FFiBrliA