Personal Project - Final Report

CS686 - Data Processing in Cloud
Lacee Xu

Motivation

Due to recent events of infectious diseases, I would like to learn about characteristics of other diseases and epidemics. Specifically conditions relating to influenza, pneumonia and whether the efficiency of vaccines.

About the data

The data is retrieved from Project Tyco through their web API interface. The data is csv format and split into a specific condition name and location.

The data has been gathered into two locations, USA and rest of world (ROW) and the condition names have been aggregated.

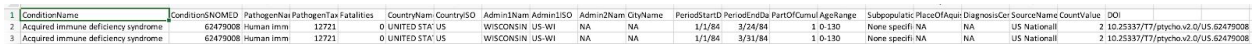
Data files

- tycho_ALL_ROW.csv
    - 23.38 MB
    - 101,042 rows
- tycho_USA_all_data.csv
    - 1.14 GB
    - 5575383 rows

Data description

- 92 different types of conditions
    - In the USA, condition cases ranges from 47 count (Syphilis) to 436932 count (Measles).
    - Each entry is counted ('CountValue') for the specified time frame ('PeriodStartDate' and 'PeriodEndDate').
    - SNOMED format is used for the condition, pathogen and id naming.

Example



	ConditionName	ConditionSNOMED	PathogenNa	PathogenTax	Fatalities	CountryNam	CountryISO	Admin1Nam	Admin1ISO	Admin2Nam	CityName	PeriodStartD	PeriodEndDa	PartOfCumul	AgeRange	Subpopulatic	PlaceOfAqui	DiagnosisCer	SourceName	CountValue	DOI
2	Acquired immune deficiency syndrome	62479008	Human imm	12721	0	UNITED STA	US	WISCONSIN	US-WI	NA	NA	1/1/84	3/24/84	1	0-130	None specifi	NA	NA	US Nationall	2	10.25337/T7/ptycho.v2.0/US.62479008
3	Acquired immune deficiency syndrome	62479008	Human imm	12721	0	UNITED STA	US	WISCONSIN	US-WI	NA	NA	1/1/84	3/31/84	1	0-130	None specifi	NA	NA	US Nationall	2	10.25337/T7/ptycho.v2.0/US.62479008

Data preprocessing

A python script was created to pull the data using Tyco's API and combine all condition names into one csv. The data limit for each call is 20,000 entries - a loop is needed to retrieve the complete dataset for each condition. The script also filters out any corrupted data if the entry has extra columns otherwise it will produce an error during conversion in BigQuery.

```
max_count = 20000

outfile = open("tyco_USA_all_data.csv","w")
skip_header = False

for condition in condition_map:

    count = condition[1]
    offset = 0
    while count > 0:
        condition_url = condition[0].replace(" ", "%20")
        url =
            "https://www.tycho.pitt
            .edu/api/query?apikey=ca9c5d59101a526ddcc7&ConditionName=" +
            condition_url + "&CountryISO=US&offset=" + str(offset) + "&limit=20000"

        print("Getting from url: " + url)
        req = urllib.request.urlopen(url)
        if skip_header:
            line = req.readline().decode("utf8")
#            print(line)
            col_num = len(line.split(','))
            if col_num is not 20:
                print("Column number is not 20")
                break
#            print(col_num)
        else:
            skip_header = True
        content = req.read()
        outfile.write(content.decode("utf8"))
        count -= max_count
        offset += max_count

        print(count)

outfile.close()
```

BigQuery

The csv was uploaded to the GCP bucket and the table was converted to a BigQuery table.

The schema between the tycho_USA_all_data.csv and tycho_ALL_ROW.csv data sets are similar with the only difference is in the Type and lack of the DOI field. The reason why the tycho_USA_all_data.csv dataset has a STRING field is that some of the Integer and Float fields have unexpected data which would cause BigQuery to throw an error when importing.

Load gs://tyco_bucket/tyco_USA_all_data.csv to just-experience-270022:cs686.tyco_usa_all
12:23 PM

Error while reading data, error message: CSV table encountered too many errors, giving up. Rows: 1; errors: 1. Please look into the errors[] collection for more details.

Error while reading data, error message: Could not parse 'CountValue' as double for field CountValue (position 19) starting at location 0

However, when using DataStudio, the correct types will be applied.

Schema

tyco_usa_all

Schema Details Preview

Field name	Type	Mode
ConditionName	STRING	NULLABLE
ConditionSNOMED	STRING	NULLABLE
PathogenName	STRING	NULLABLE
PathogenTaxonID	STRING	NULLABLE
Fatalities	STRING	NULLABLE
CountryName	STRING	NULLABLE
CountryISO	STRING	NULLABLE
Admin1Name	STRING	NULLABLE
Admin1ISO	STRING	NULLABLE
Admin2Name	STRING	NULLABLE
CityName	STRING	NULLABLE
PeriodStartDate	STRING	NULLABLE
PeriodEndDate	STRING	NULLABLE
PartOfCumulativeCountSeries	STRING	NULLABLE
AgeRange	STRING	NULLABLE
Subpopulation	STRING	NULLABLE
PlaceOfAquisition	STRING	NULLABLE
DiagnosisCertainty	STRING	NULLABLE
SourceName	STRING	NULLABLE
CountValue	STRING	NULLABLE

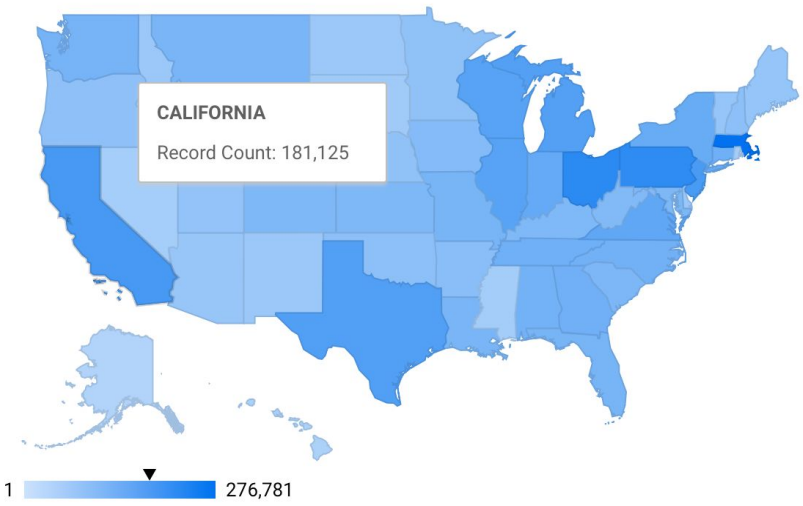
tycho_all_row

Schema Details Preview

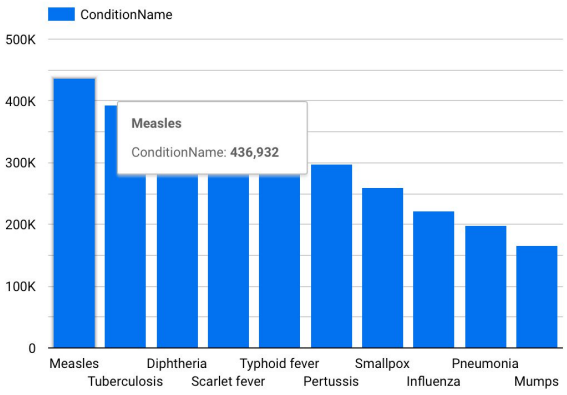
Field name	Type	Mode
ConditionName	STRING	NULLABLE
ConditionSNOMED	INTEGER	NULLABLE
PathogenName	STRING	NULLABLE
PathogenTaxonID	INTEGER	NULLABLE
Fatalities	INTEGER	NULLABLE
CountryName	STRING	NULLABLE
CountryISO	STRING	NULLABLE
Admin1Name	STRING	NULLABLE
Admin1ISO	STRING	NULLABLE
Admin2Name	STRING	NULLABLE
CityName	STRING	NULLABLE
PeriodStartDate	DATE	NULLABLE
PeriodEndDate	DATE	NULLABLE
PartOfCumulativeCountSeries	INTEGER	NULLABLE
AgeRange	STRING	NULLABLE
Subpopulation	STRING	NULLABLE
PlaceOfAquisition	STRING	NULLABLE
DiagnosisCertainty	STRING	NULLABLE
SourceName	STRING	NULLABLE
CountValue	FLOAT	NULLABLE
DOI	STRING	NULLABLE

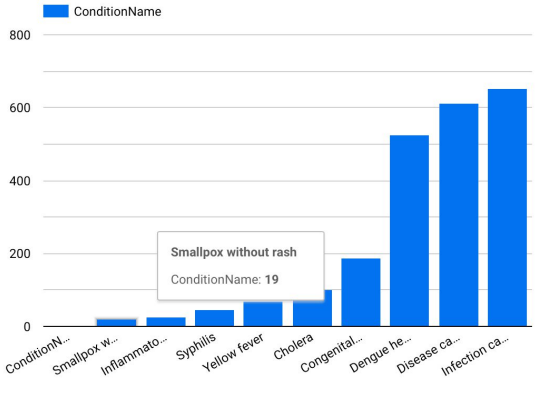
Shape of the data

Below is the heat map of the number of all conditions for each state in the US over time.



Since there are different 92 conditions in the dataset, it'll be more meaningful to analyze specific conditions of time and location, specifically the United states. From the data, we can top and bottom 10 conditions in terms of count in the US. In the analysis, we'll focus on the most prominent conditions such as Measles, Influenza and Pneumonia.



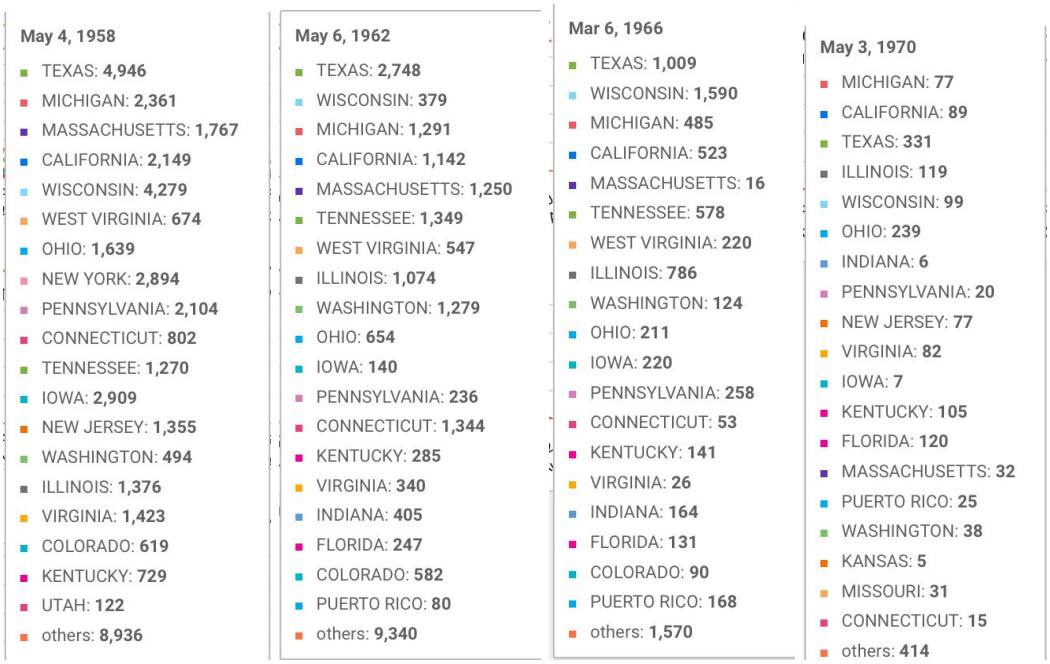


Data Visualization and Analysis

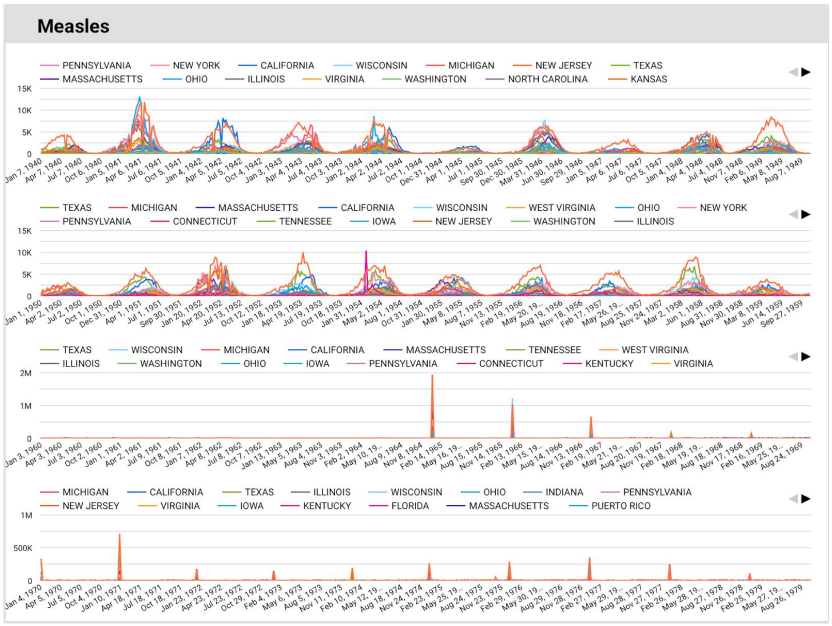
Measles in USA

Looking at measles over the decades, we can see that the measles infection is periodic and that after around 1965, the infection count reduced dramatically. This can be explained by the

measle vaccine being invented in 1963. This observation was inspired after searching online, many other people concluded the same thing.

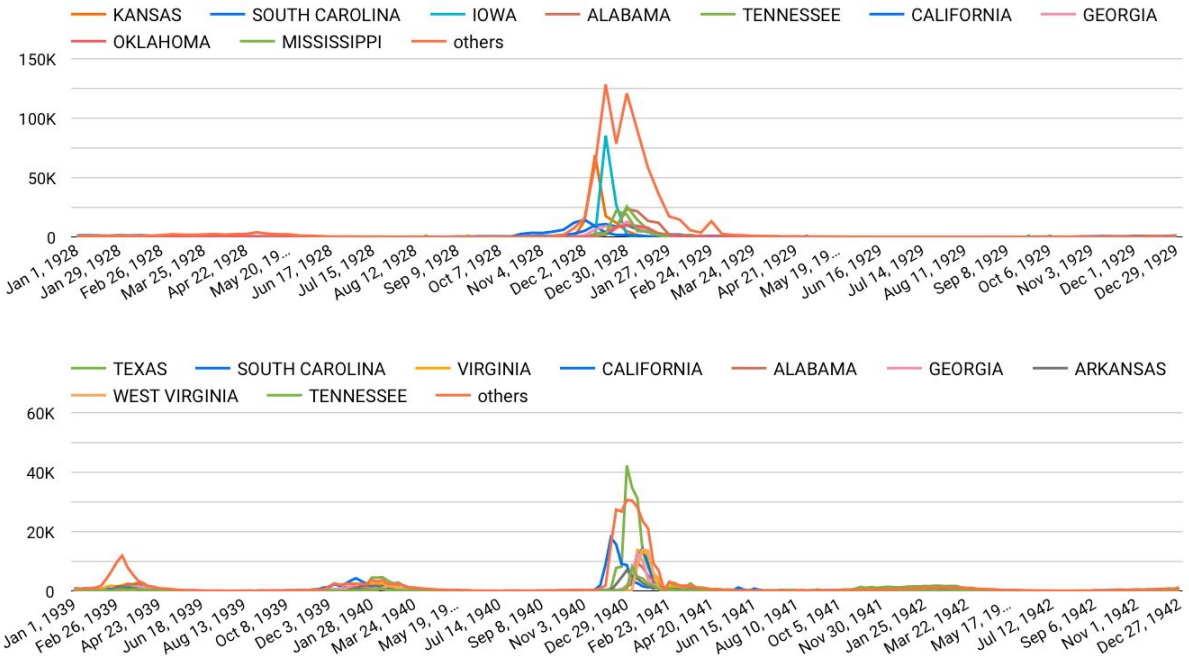


One thing to note is that in the data presented, there seems to be outlier dates that reaches to the millions.



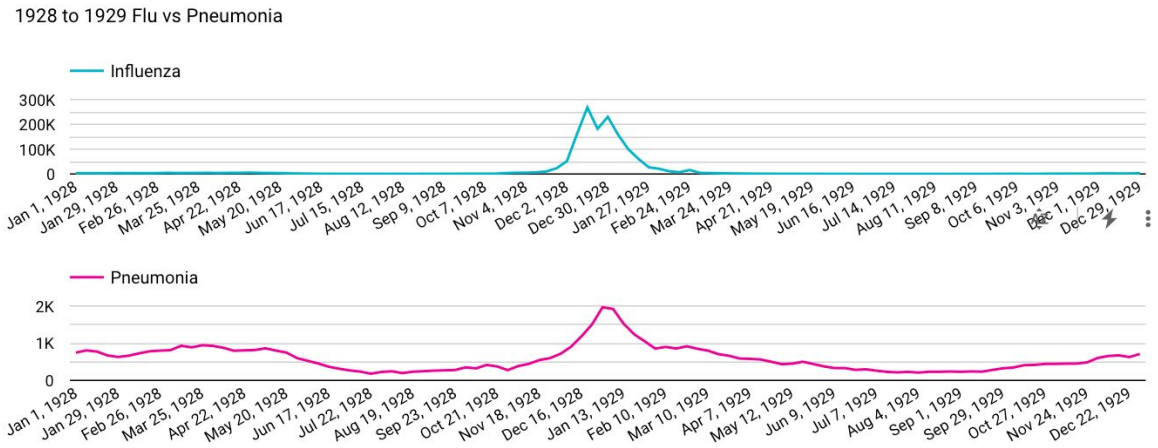
Influenza in USA

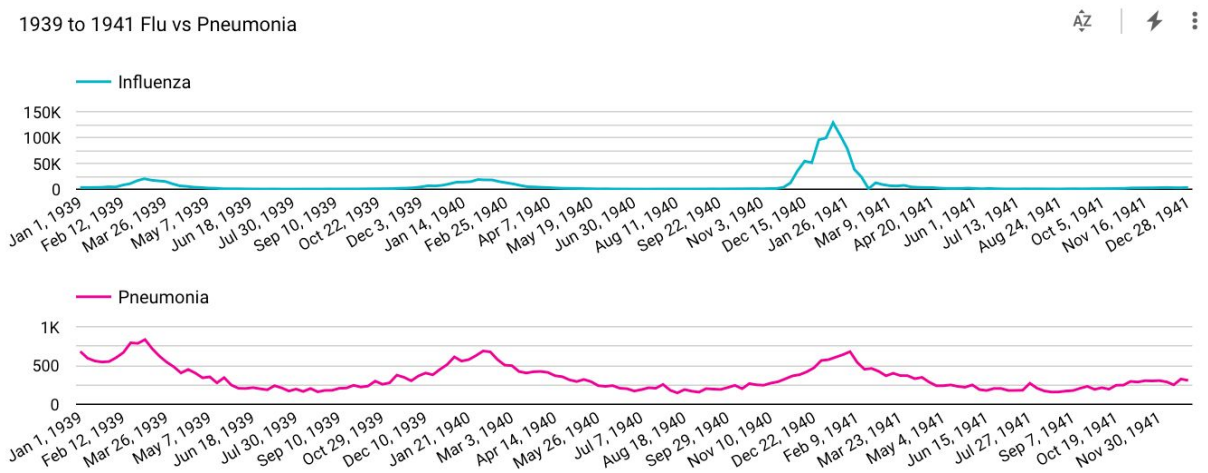
From the dataset, the year with the most amount of influenza in the USA cases is 1928 and 1940. If we take a look at the data over time, it looks like the majority of the cases fall between winter time between December and April.



Pneumonia in USA

To determine if flu and pneumonia have a subtle or significant correlation, we can look at both conditions over the same time range. From the years of 1928 and 1940, we can see pneumonia cases follow that of influenza.





Why BigQuery

BigQuery was used because it's simple to convert the Tycho csv dataset into a BigQuery table.

- BigQuery allows for flexible definition of the schema - whether auto or user defined.
- It's also very quick to process to convert - 1.2GB csv only took around 35.2 seconds.
- Easy to access as one can run query commands on the web browser without installing any external software
- BigQuery can also be integrated with Data Studio which provides an easy and fun way to visualize the data.