Data Analysis Report for Lab1

Part 1

After analyzing New York's Air Quality, we can say that up to a certain level, an increase in ozone is accompanied by rise in temperature, as indicated by below graph.

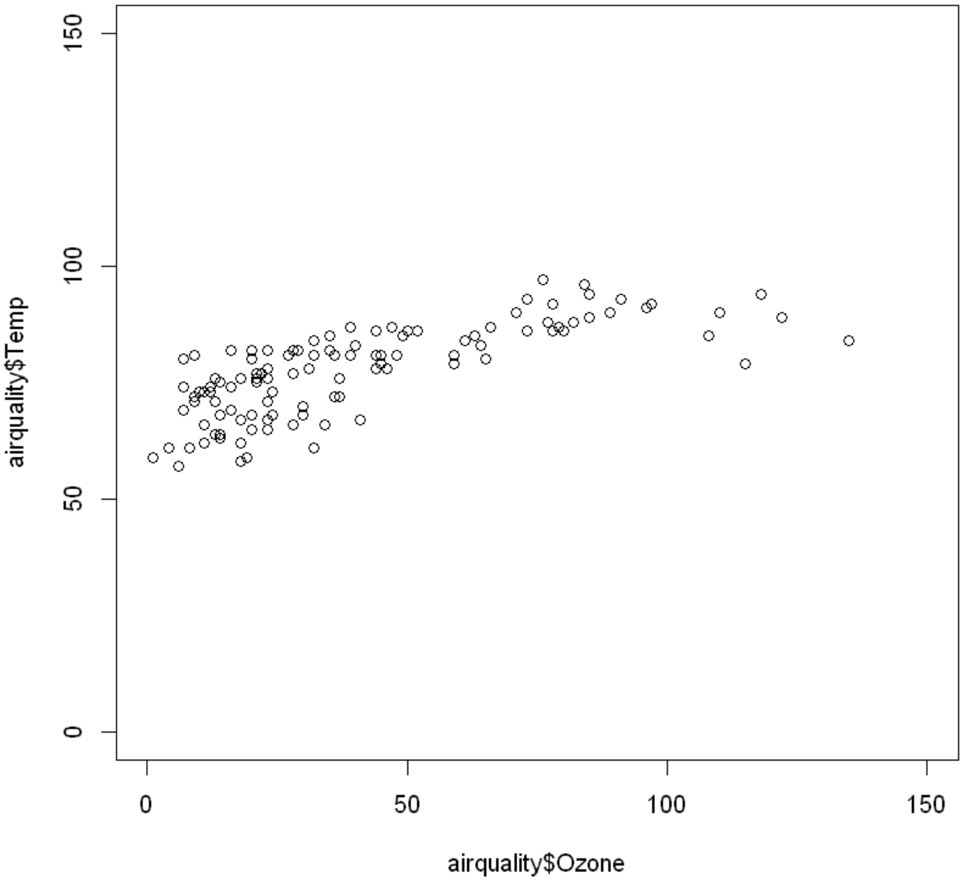
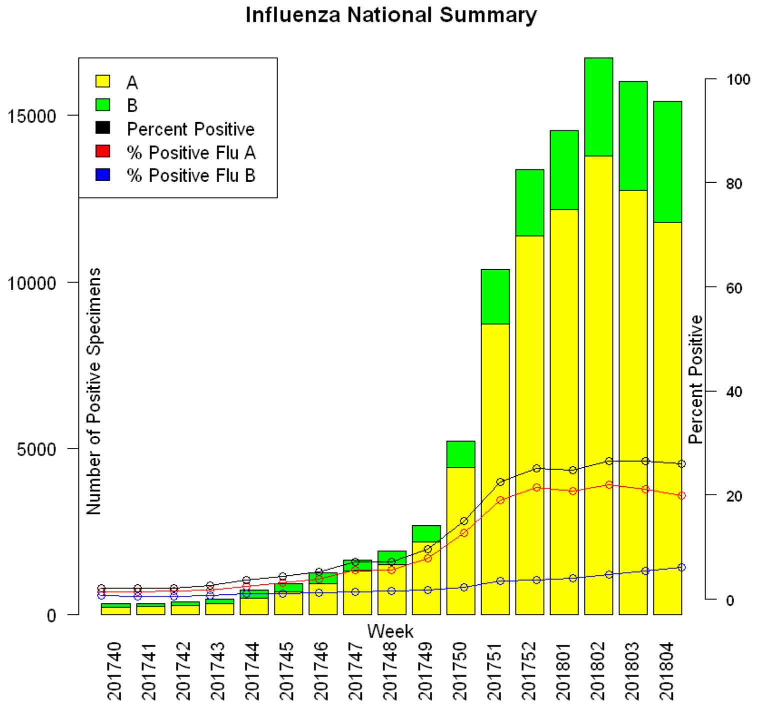
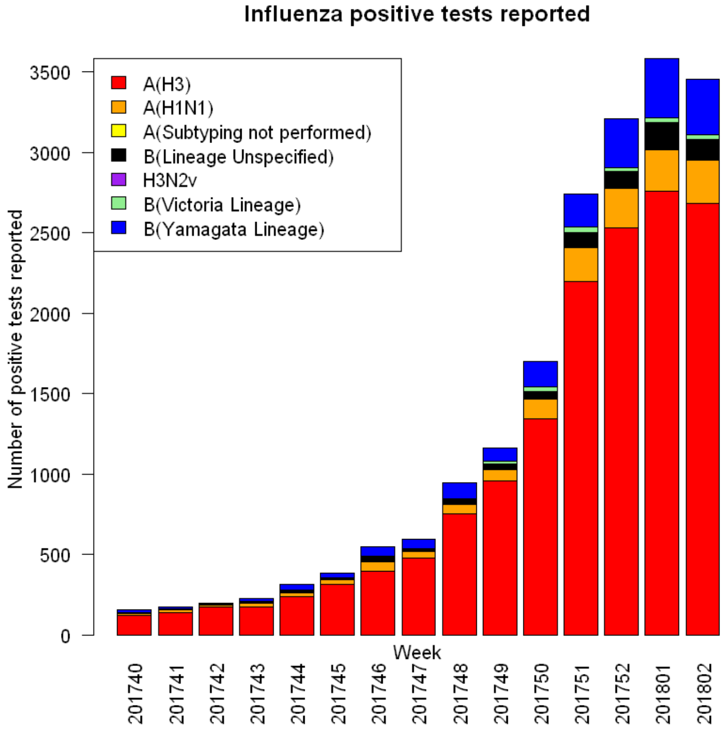


Fig: Graph of ozone and temperature

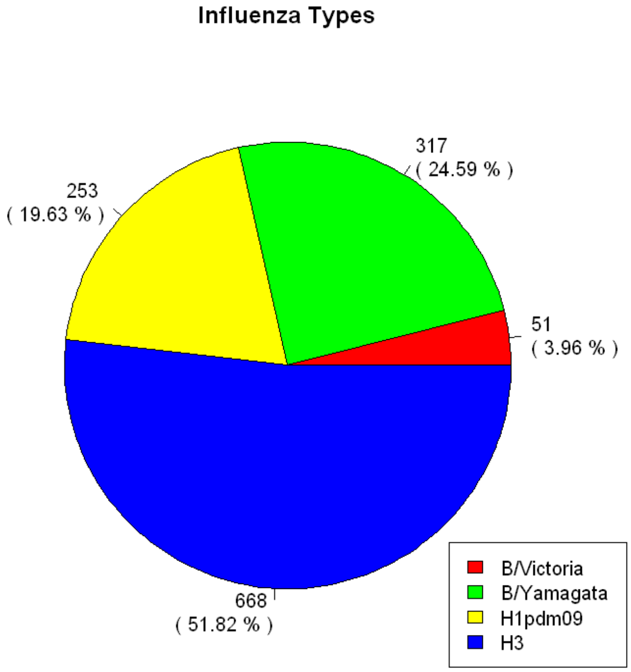
Part 2

By Influenza National Summary and positive test cases reported, we can say that more specimens were reported till week 2, 2018 after which a decrease was observed in that number of specimens indicating that maybe weather got better or people got vaccinations, as indicated by below graphs.

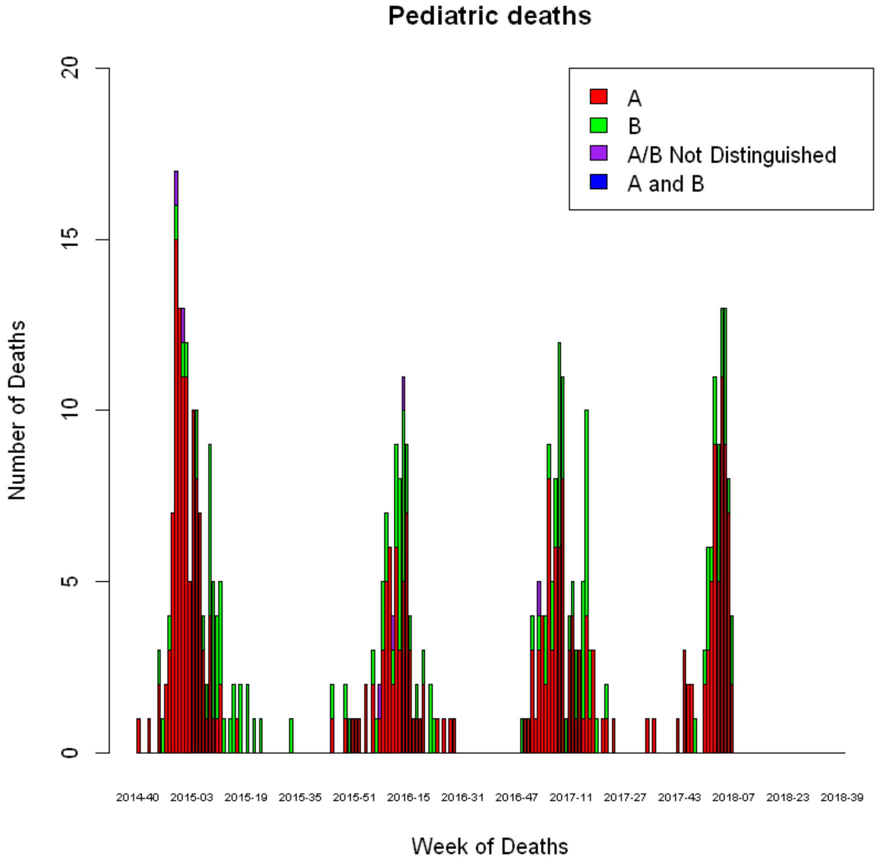

Influenza National Summary


Influenza positive tests reported

By Virus Classification Data, we can conclude that H3 was the most widespread type as indicated by below pie chart.

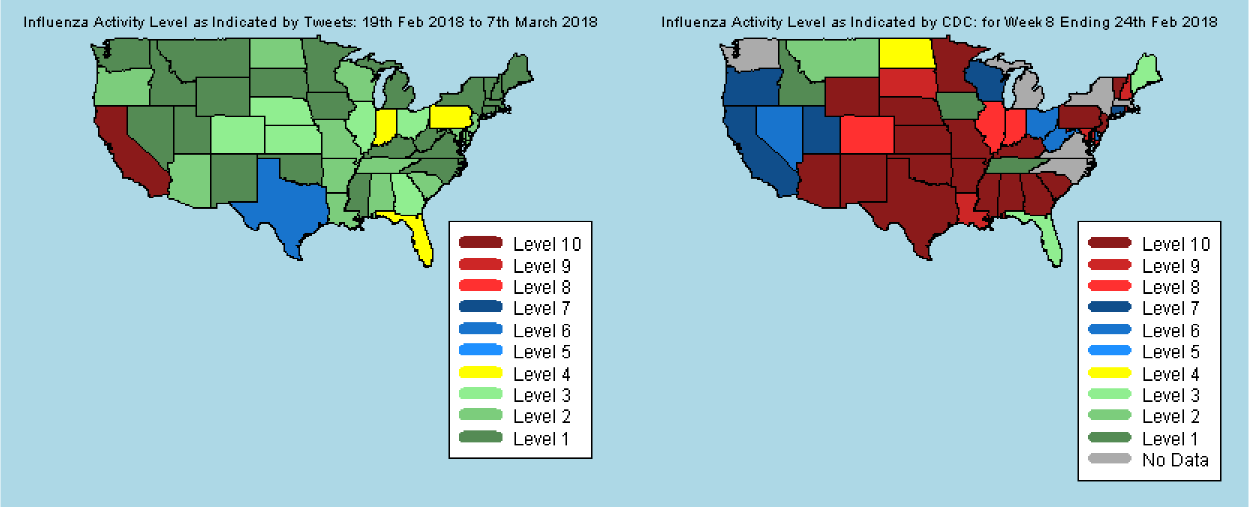


By pediatric deaths observed we can conclude that medical facilities got better as less number of deaths were reported in following weeks, as indicated by below graph.



Part 3

Topic of discussion: Can an epidemic like Influenza be correctly estimated using social media platforms like Twitter?


Influenza Activity Level as Indicated by Tweets: 19th Feb 2018 to 7th March 2018
Level 10
Level 9
Level 8
Level 7
Level 6
Level 5
Level 4
Level 3
Level 2
Level 1
Influenza Activity Level as Indicated by CDC: for Week 8 Ending 24th Feb 2018
Level 10
Level 9
Level 8
Level 7
Level 6
Level 5
Level 4
Level 3
Level 2
Level 1
No Data

The results from above analysis can be divided into 3 categories:

1. **Over Estimations:**
   As reported by CDC, **California had an influence level of 7** but due to more number of tweets from California, the level got over **estimated to level 10**.

2. **Under Estimations:**
   As reported by CDC, **Texas has an influence level of 10** but due to less number of tweets from Texas, the level got under **estimated to level 6.**

3. **Accurate Estimations:**
   Idaho was correctly estimated by Twitter data.

Conclusion:

Reasons for over or under estimations can be as follows:

1. **Less data**
   Due to Twitter API restrictions of providing only a week old data.

2. **Less or absurd locations**
   Most users hide their locations while posting on twitter.
   Also some locations include virtual places like heaven, sunny day, hot place, wet place etc.

3. **Demographics**
   It might be the case that the location with most sever flu, may have people that never use Twitter or never tweet about epidemics, whereas people with less severe areas have very active users of Twitter.
   For all the results to be accurate it is needed that all anomalies are ignored and we have uniformity, which is rarely the case.

Lesser anomalies in actual and virtual scenarios will lead to better results.