Big Data Final Project

Use Data to Understand COVID-19

Team name:
Airline with COVID

Team members:
Shantanu Tripathi
Dongzi Qu
Binghan Li

Goal

- Understand the relationship between COVID-19 and air travel
  - How COVID-19 affect the air travel industry
    - Show the relationship between flights and cases by line charts
  - How air travel helped the spread of COVID-19 within the United States
  - Attempt to use data to
    - Trace the origin of the virus for a selected city
    - Numerically present which city contributed to the COVID-19 spread in a selected city

Datasets Used

- COVID-19
  - County level data (date, cases, deathes)
  - Updated daily
  - Provided by New York Times
  - https://github.com/nytimes/covid-19-data/blob/master/us-counties.csv
- Flights
  - Worldwide flight data (date, origin, destination)
  - Updated monthly (January 2020 to April 2020 are currently available)
  - Provided by OpenSky Network
  - https://opensky-network.org/datasets/covid-19/
- Airports
  - Worldwide airports data (geolocation, airport code)
  - Updated daily
  - Provided by OurAirports
  - https://ourairports.com/data/

Data Processing and Challenges

- Flights
  - Consists of to and from **Airport Codes**. (Not city names)
  - Requires merging 4 data files to 1 large data file.
- COVID-19
  - Consists of **city name**. (Not airport codes)
  - Two different counties with same name in different state. (Selecting the city of choice)
  - The mapping between county and city
- Airports
  - Consists of **airport code** and **city name**.
  - Acts as a join between Flights and Covid-19 dataset.
  - City names for the same city vary in Airports and COVID-19. (Needed consistency)
  - One city may have multiple airports. (Mapped different cities to the same city)

Datasets Generation

- covid_flight_count_data.csv
  - For each selected city on each date after COVID is discovered:
    - COVID-19 case number
    - COVID-19 death number
    - Total number of incoming flight on each date
    - Total number of outgoing flight on each date
  - Generated using all three datasets listed in previous section

city	day	cases	deaths	incoming_flight_count	outgoing_flight_count
BOSTON	2020-02-01	1	0	341	310
BOSTON	2020-02-02	1	0	299	293
BOSTON	2020-02-03	1	0	395	405
BOSTON	2020-02-04	1	0	423	424

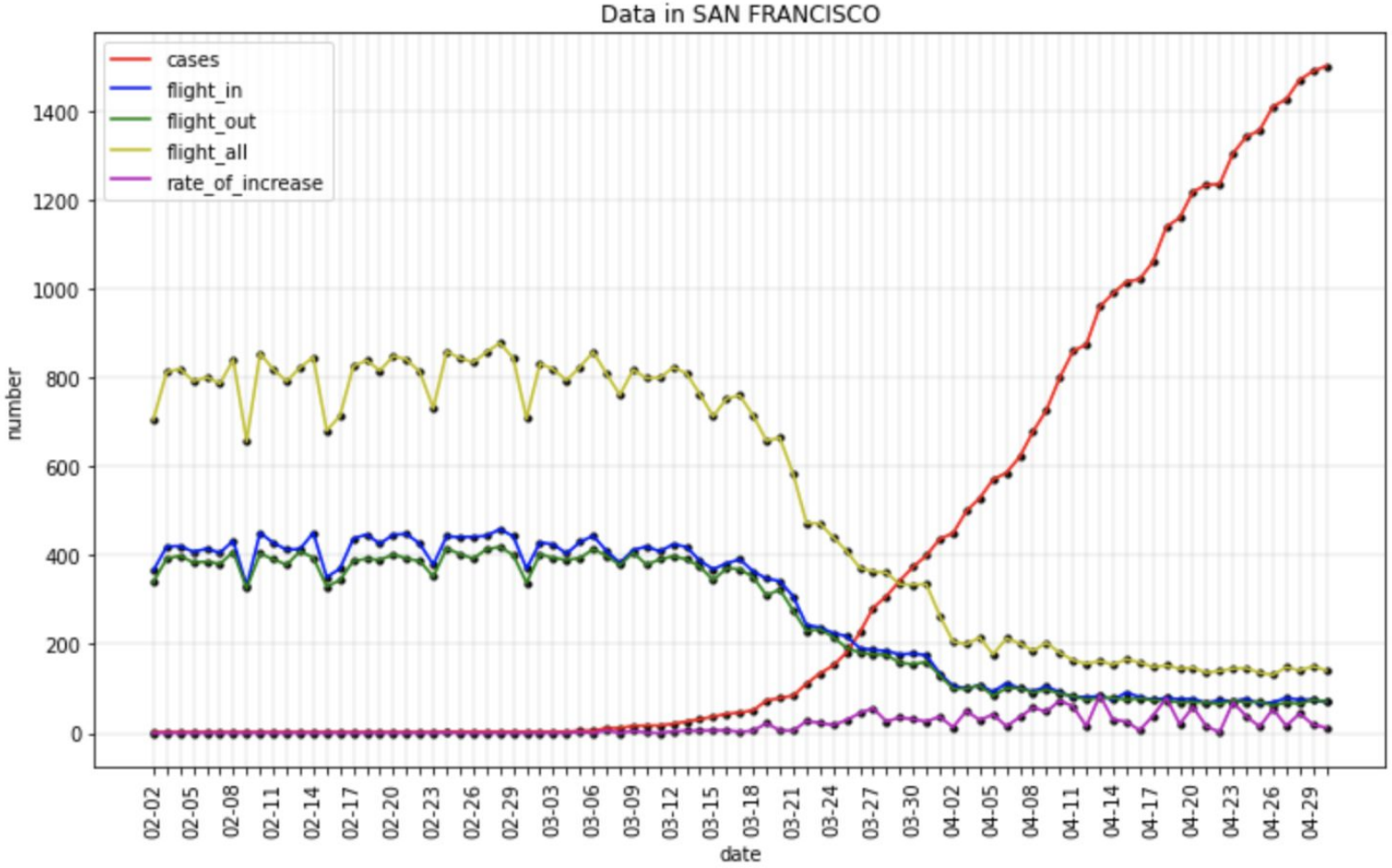
Datasets Generation

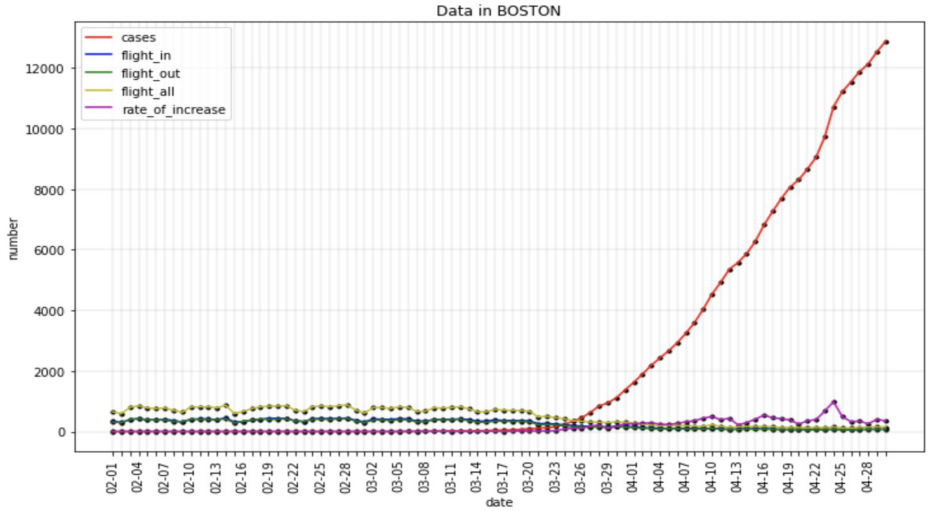
- Inter_city_flight_data.csv
  - Each record gives the details of a flight from one city to another.

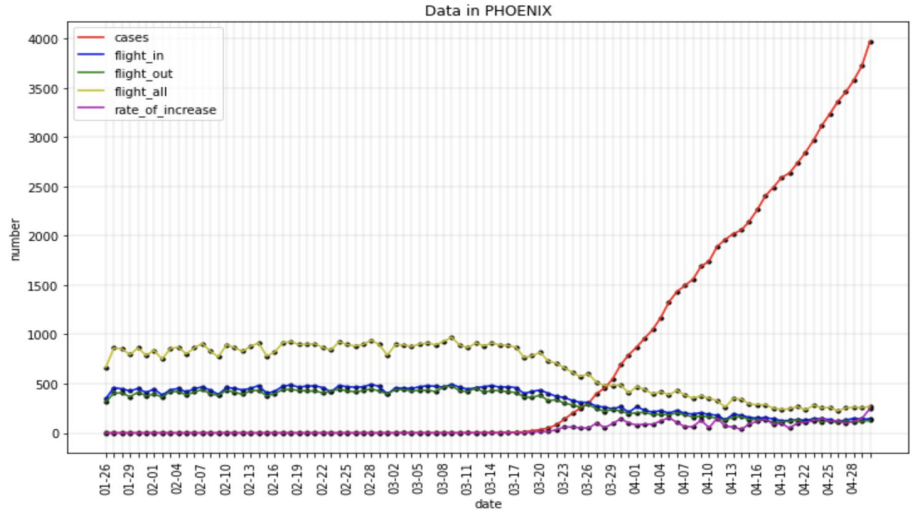
```
+--------------------+-----------+---------+----------+----------+
|           from_city|from_airport|  to_city|to_airport|       day|
+--------------------+-----------+---------+----------+----------+
|WILKES-BARRE/SCRA...|        KAVP|CHARLOTTE|      KCLT|2020-02-15|
|WILKES-BARRE/SCRA...|        KAVP|CHARLOTTE|      KCLT|2020-02-17|
|WILKES-BARRE/SCRA...|        KAVP|CHARLOTTE|      KCLT|2020-02-21|
|WILKES-BARRE/SCRA...|        KAVP|CHARLOTTE|      KCLT|2020-02-21|
+--------------------+-----------+---------+----------+----------+
```

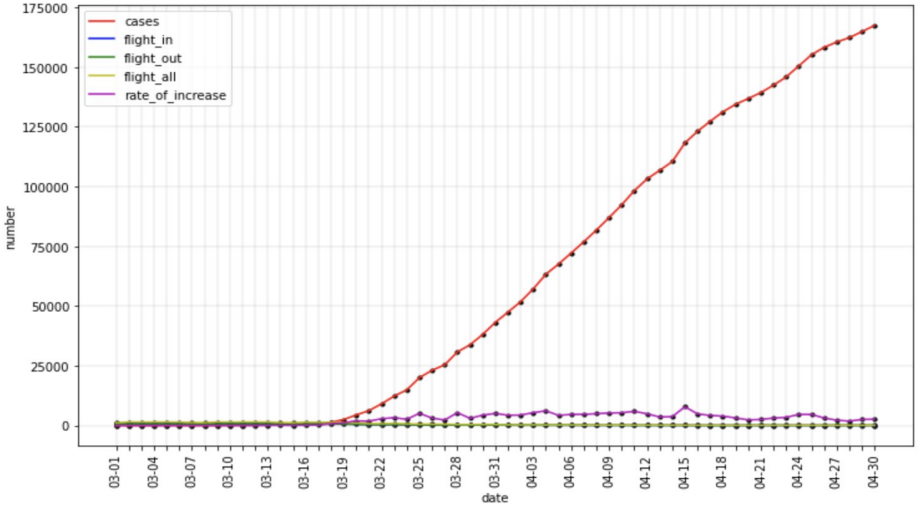
COVID-19 Impacts Air Travel Industry

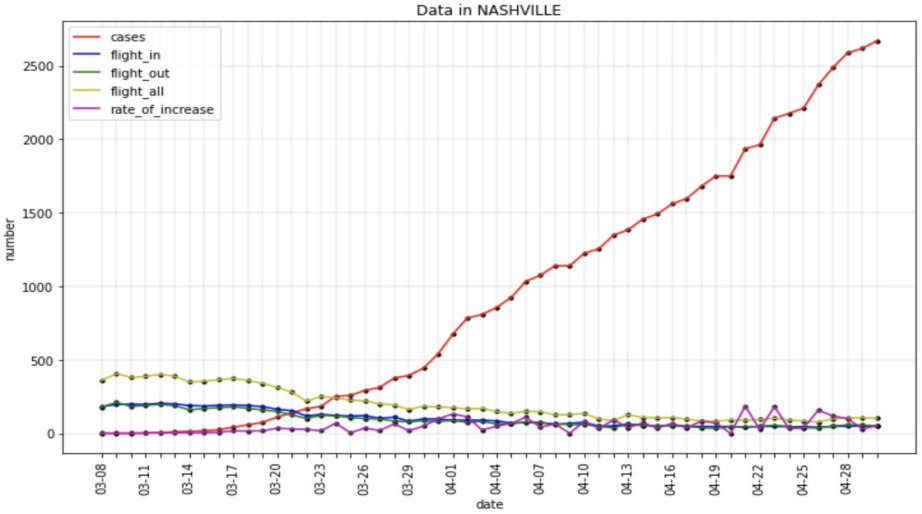
- Aim: Quantify the impact of Covid on Air Industry
- Process:
  - For each city, make a time series plot for:
    - Daily Covid Cases
    - Daily flight count
    - Daily rate of increases
  - Analyse the plots to observe:
    - Rise of Covid Curve
    - Plummet of Flight Curve
  - Decide a date ‘d’ which marks the fall of flight counts. (For us, d = 15th March)
  - Calculate average flights before and after ‘d’
  - For each city, calculate the fall in percentage of flight using the following:
    (flight_before - flight_after) / flight_before

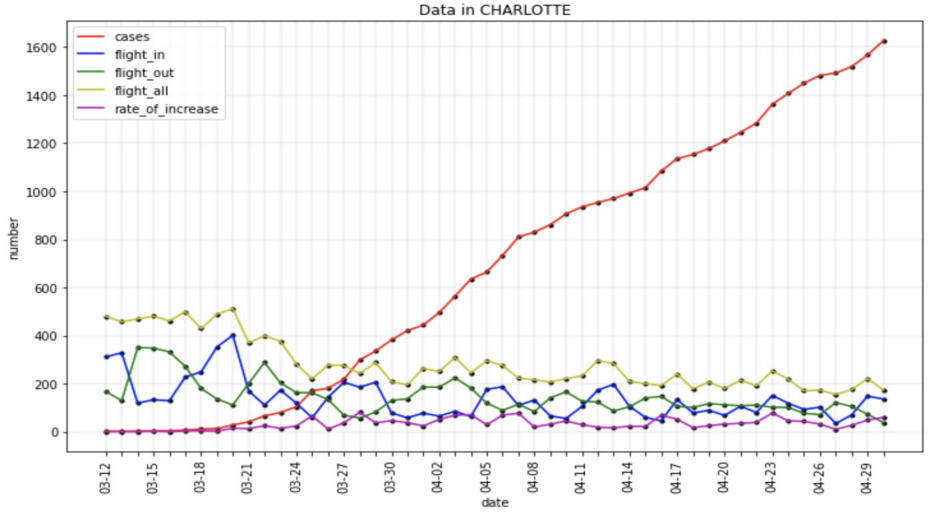

Data in SAN FRANCISCO
cases
flight_in
flight_out
flight_all
rate_of_increase
number
0
200
400
600
800
1000
1200
1400
02-02
02-05
02-08
02-11
02-14
02-17
02-20
02-23
02-26
02-29
03-03
03-06
03-09
03-12
03-15
03-18
03-21
03-24
03-27
03-30
04-02
04-05
04-08
04-11
04-14
04-17
04-20
04-23
04-26
04-29
date

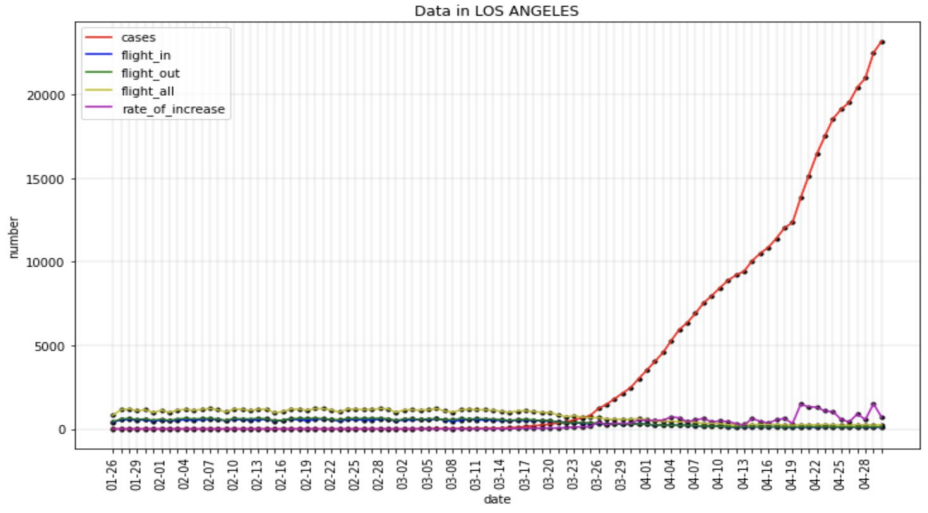

Data in BOSTON
cases
flight_in
flight_out
flight_all
rate_of_increase
number
date

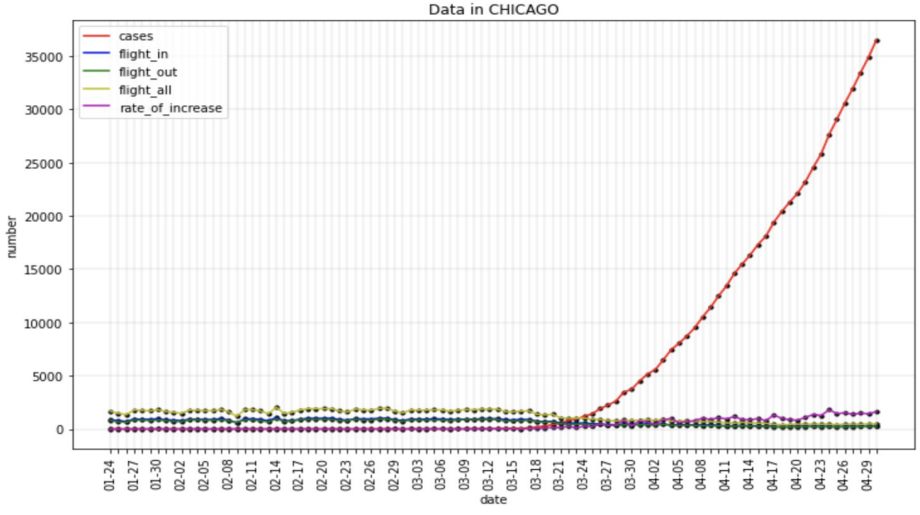

Data in PHOENIX
cases
flight_in
flight_out
flight_all
rate_of_increase
number
date

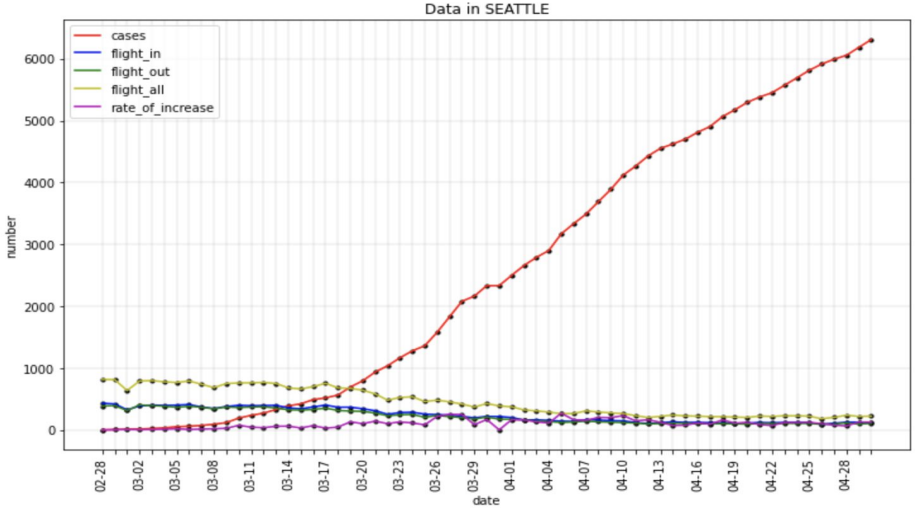

Data in NEW YORK CITY
cases
flight_in
flight_out
flight_all
rate_of_increase
number
date

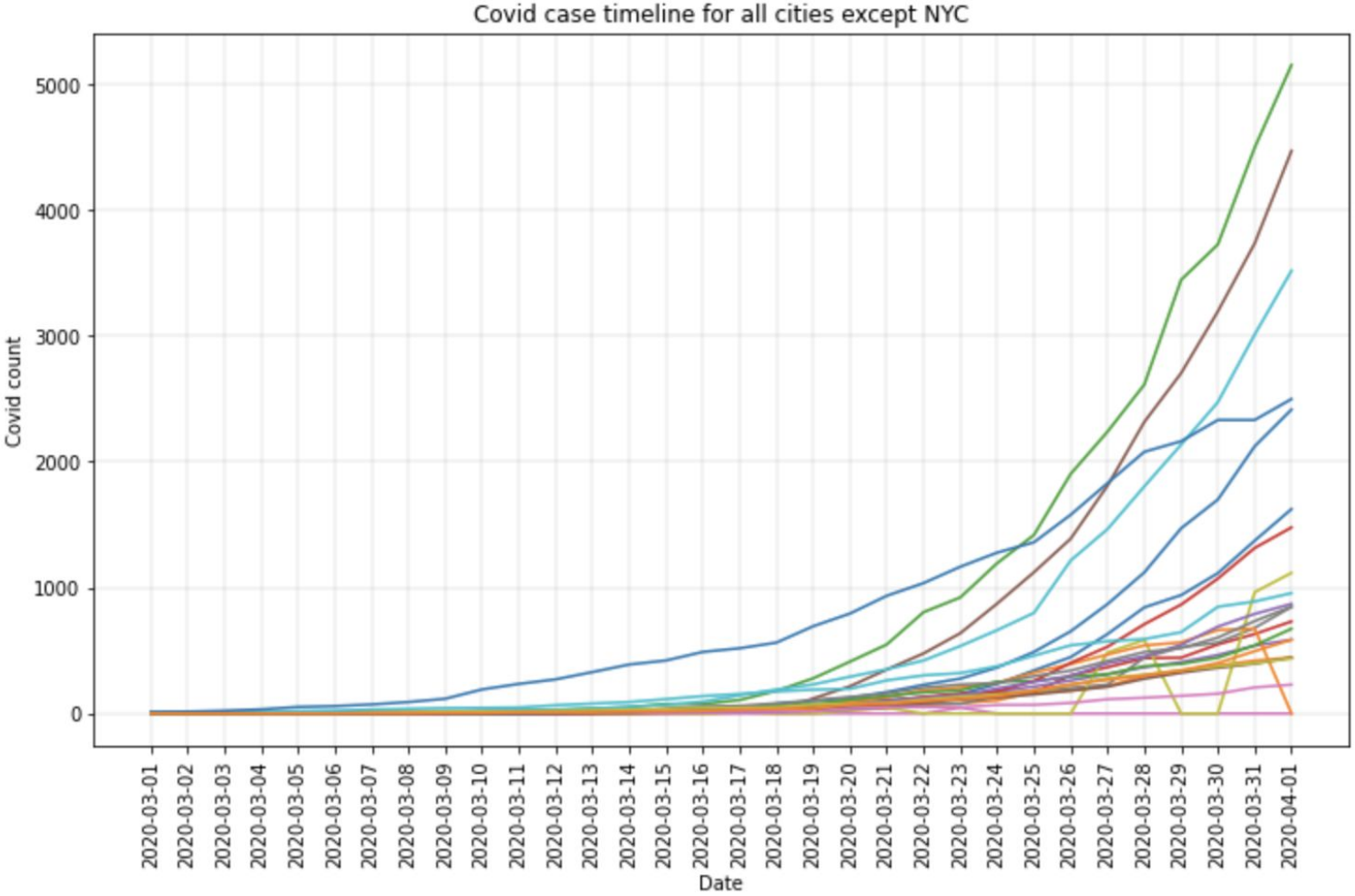

Data in NASHVILLE
cases
flight_in
flight_out
flight_all
rate_of_increase
number
date

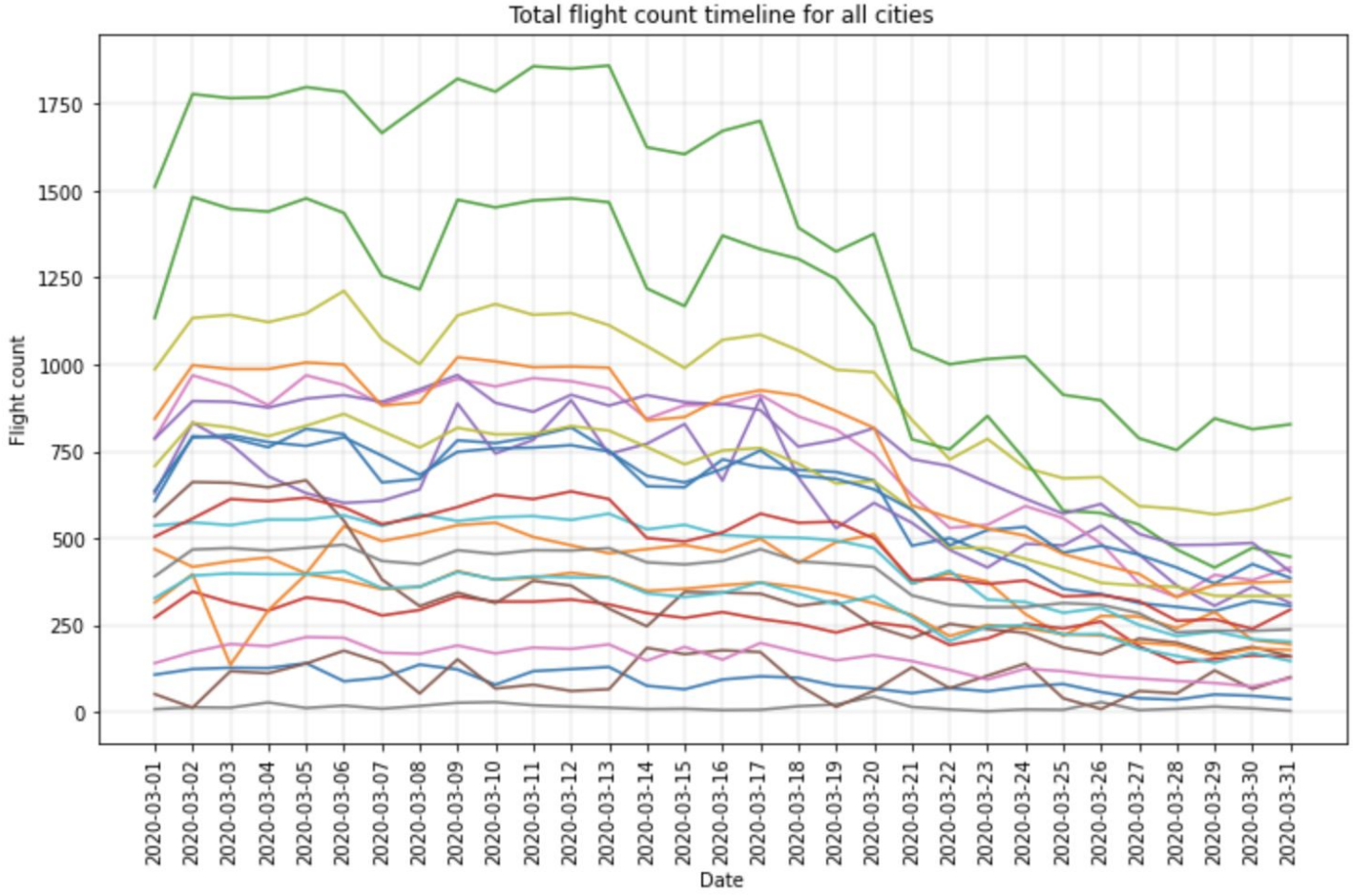

Data in CHARLOTTE
cases
flight_in
flight_out
flight_all
rate_of_increase
number
date

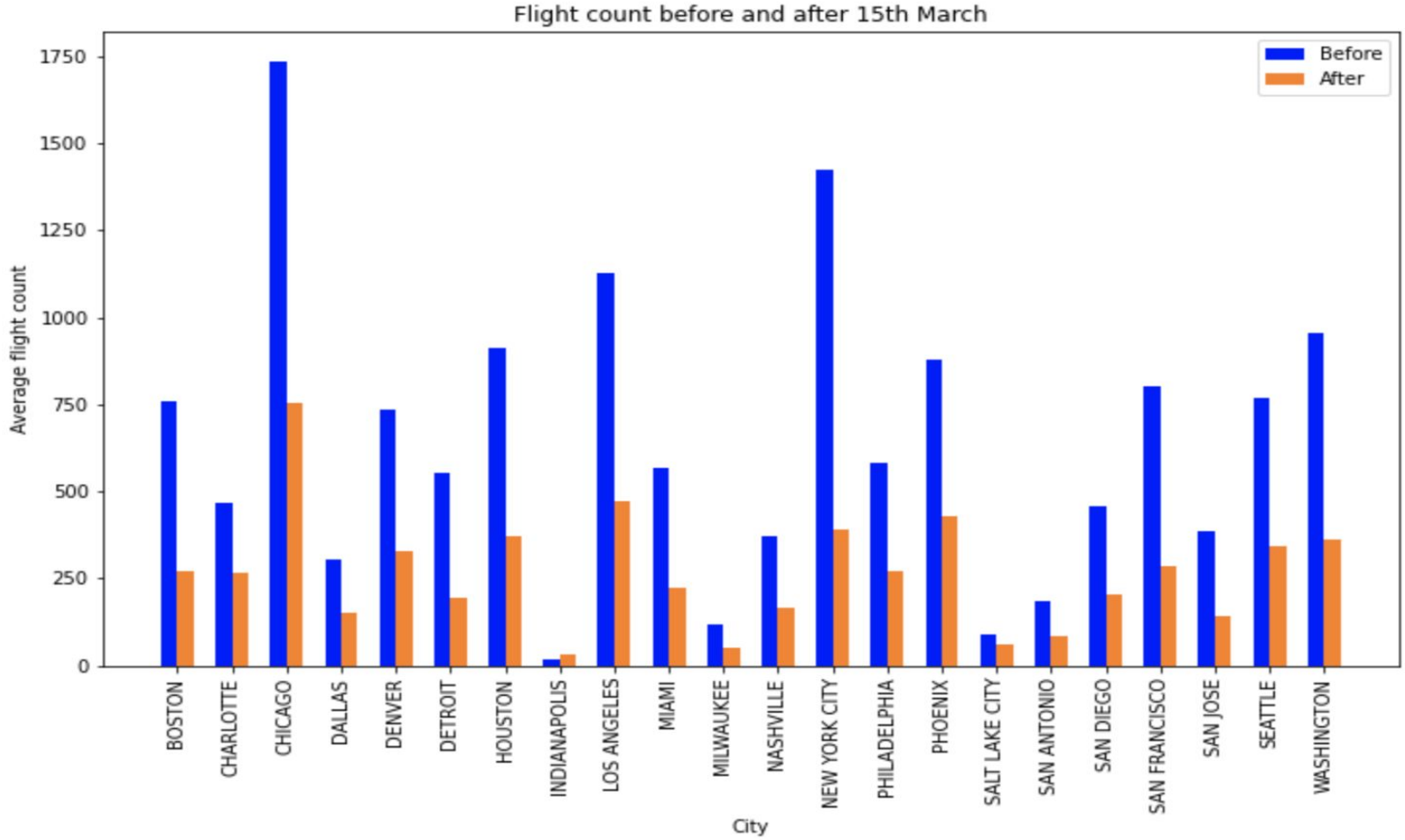

Data in LOS ANGELES
cases
flight_in
flight_out
flight_all
rate_of_increase
number
date


Data in CHICAGO
cases
flight_in
flight_out
flight_all
rate_of_increase
number
date


Data in SEATTLE
cases
flight_in
flight_out
flight_all
rate_of_increase
number
date


Covid case timeline for all cities except NYC
Covid count
5000
4000
3000
2000
1000
0
2020-03-01
2020-03-02
2020-03-03
2020-03-04
2020-03-05
2020-03-06
2020-03-07
2020-03-08
2020-03-09
2020-03-10
2020-03-11
2020-03-12
2020-03-13
2020-03-14
2020-03-15
2020-03-16
2020-03-17
2020-03-18
2020-03-19
2020-03-20
2020-03-21
2020-03-22
2020-03-23
2020-03-24
2020-03-25
2020-03-26
2020-03-27
2020-03-28
2020-03-29
2020-03-30
2020-03-31
2020-04-01
Date


Total flight count timeline for all cities
Flight count
1750
1500
1250
1000
750
500
250
0
2020-03-01
2020-03-02
2020-03-03
2020-03-04
2020-03-05
2020-03-06
2020-03-07
2020-03-08
2020-03-09
2020-03-10
2020-03-11
2020-03-12
2020-03-13
2020-03-14
2020-03-15
2020-03-16
2020-03-17
2020-03-18
2020-03-19
2020-03-20
2020-03-21
2020-03-22
2020-03-23
2020-03-24
2020-03-25
2020-03-26
2020-03-27
2020-03-28
2020-03-29
2020-03-30
2020-03-31
Date


Flight count before and after 15th March
Before
After
Average flight count
1750
1500
1250
1000
750
500
250
0
BOSTON
CHARLOTTE
CHICAGO
DALLAS
DENVER
DETROIT
HOUSTON
INDIANAPOLIS
LOS ANGELES
MIAMI
MILWAUKEE
NASHVILLE
NEW YORK CITY
PHILADELPHIA
PHOENIX
SALT LAKE CITY
SAN ANTONIO
SAN DIEGO
SAN FRANCISCO
SAN JOSE
SEATTLE
WASHINGTON
City

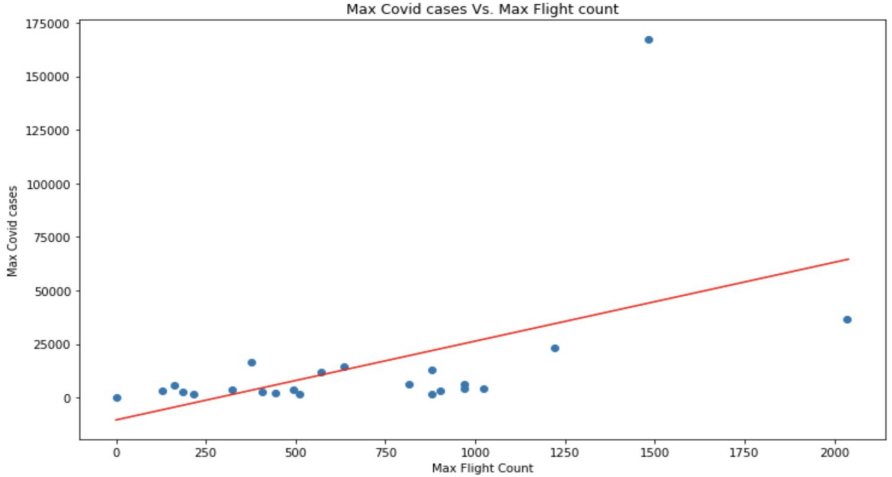
Percentage drop in air traffic before and after 15th March

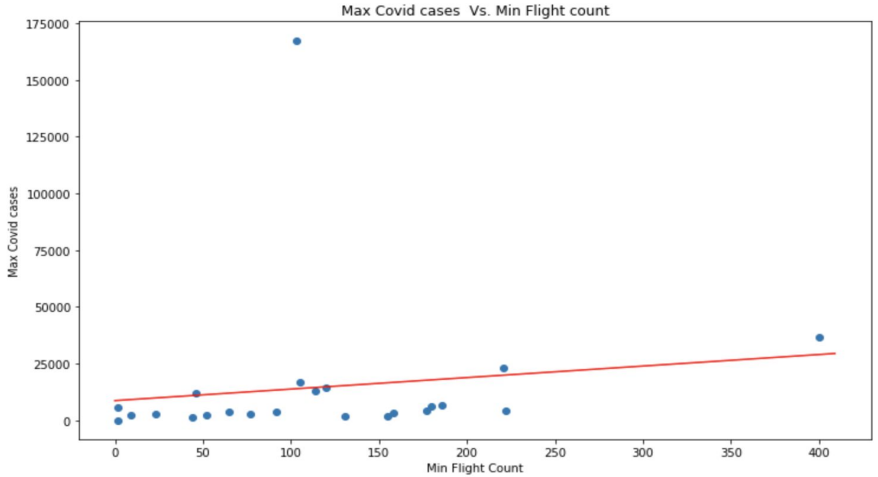
City	Percentage Drop
BOSTON	64.03436967210123
CHARLOTTE	43.11551776804807
CHICAGO	56.505997503012864
DALLAS	50.338988011042716
DENVER	54.9450223763374
DETROIT	64.85845388780345
HOUSTON	59.147483381948206
INDIANAPOLIS	-115.80912600468326
LOS ANGELES	58.198958205447504
MIAMI	60.60229655656081
MILWAUKEE	56.32749771850065
NASHVILLE	55.87767603679751
NEW YORK CITY	72.61226798921905
PHILADELPHIA	53.21531460914357
PHOENIX	50.857507661454505
SALT LAKE CITY	32.595765732339416
SAN ANTONIO	53.19578719956792
SAN DIEGO	55.02529416092192
SAN FRANCISCO	64.67235906153029
SAN JOSE	63.68037133585418
SEATTLE	55.2172628604457
WASHINGTON	62.07914197429196

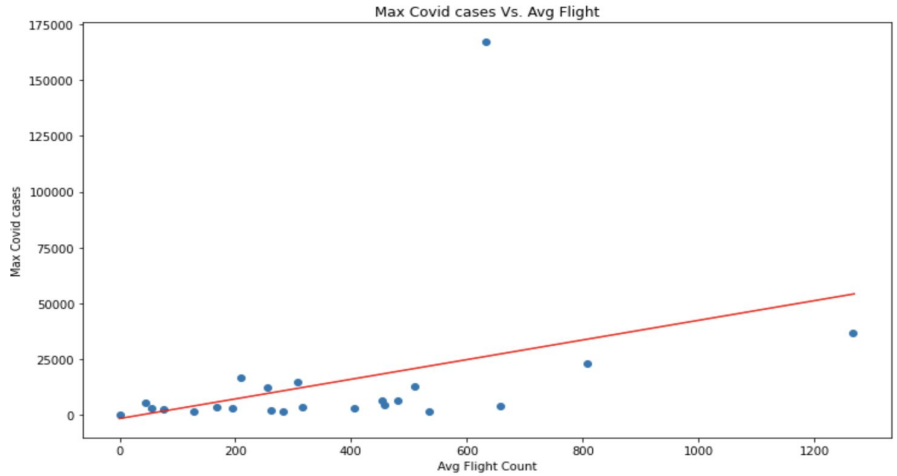
Air Travel impacts COVID-19 Spread - Analysis 1

- Aim:
  - Analyse if the flight count shares some relationship with:
    - Count of Covid cases in a city.
    - Average rate of Increase of Covid cases in a city.
    - Time taken by the city to reach a threshold Covid count.
- Process:
  - For each city, collect the following data:
    - Max Covid Count Reached
    - Max / Min / Average flight count
    - Average Rate of increase of Covid Count
    - Starting from 1st Jan, number of days taken to reach Covid count of 2000.
  - Using the data:
    - Plot graphs
    - Obtain best fit lines
    - Calculate Pearson’s Coefficient

Covid Count Vs. Flight Count - Graphs with the outlier NYC



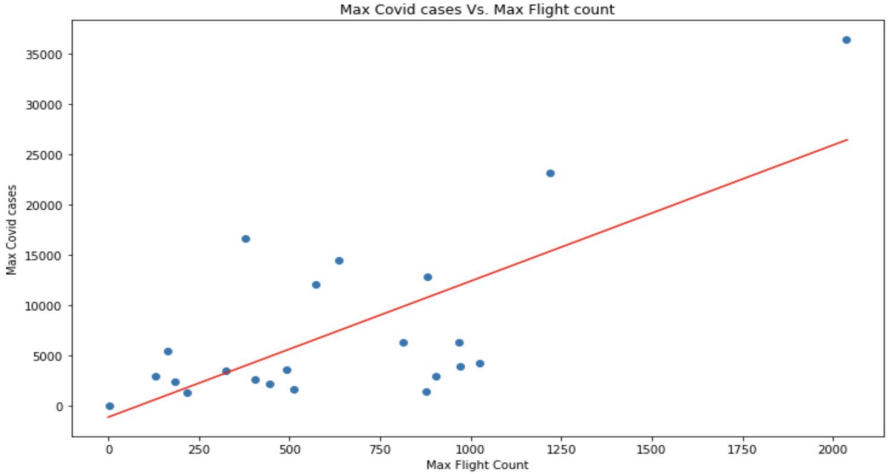


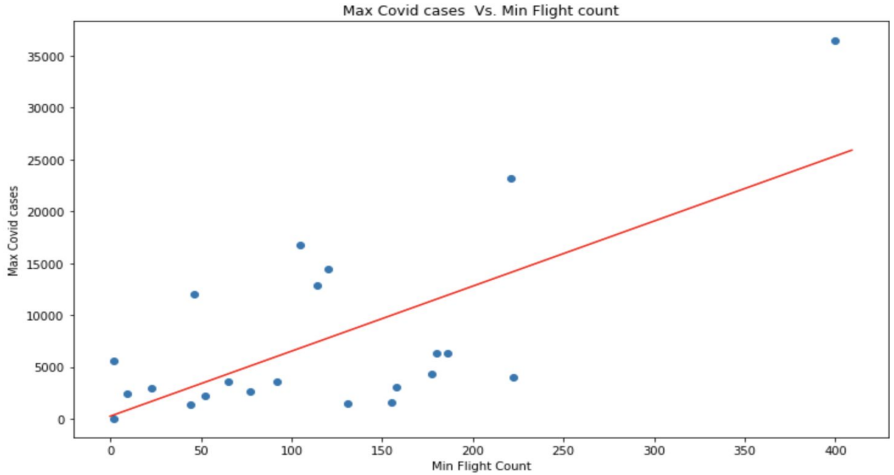


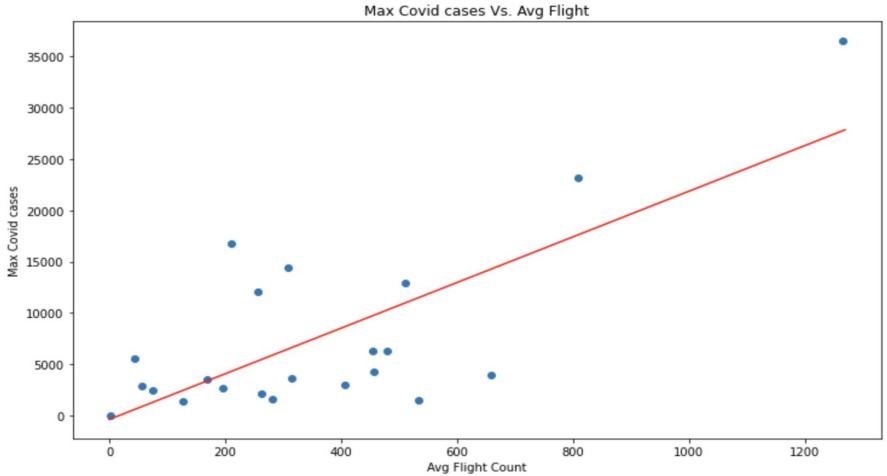
Graph	Pearson's Correlation
Max Flight	0.52
Avg Flight	0.37
Min Flight	0.14

More Flight, more Covid Cases

Covid Count Vs. Flight Count - Graphs without the outlier NYC





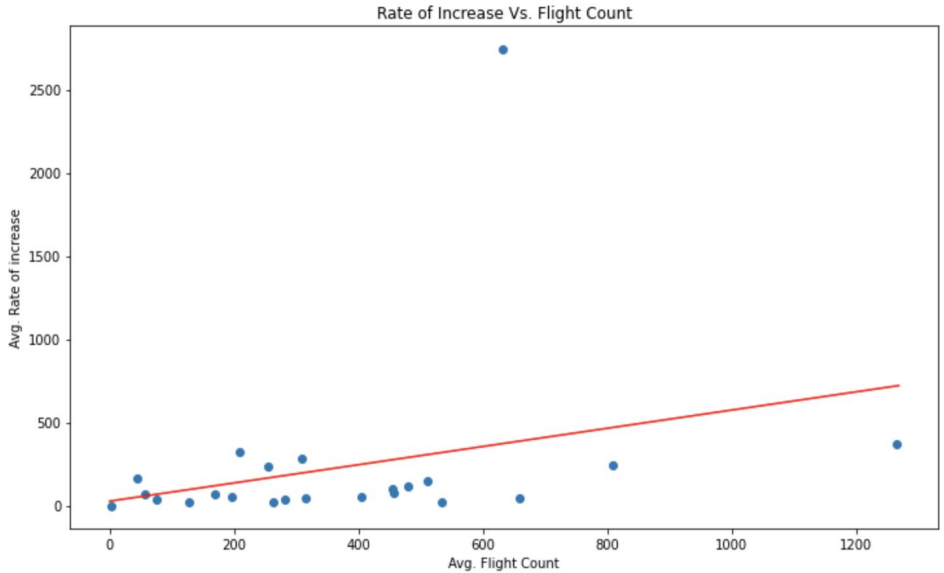


Graph	Pearson's Correlation
Max Flight	0.71
Avg Flight	0.73
Min Flight	0.67

More Flight, more Covid Cases

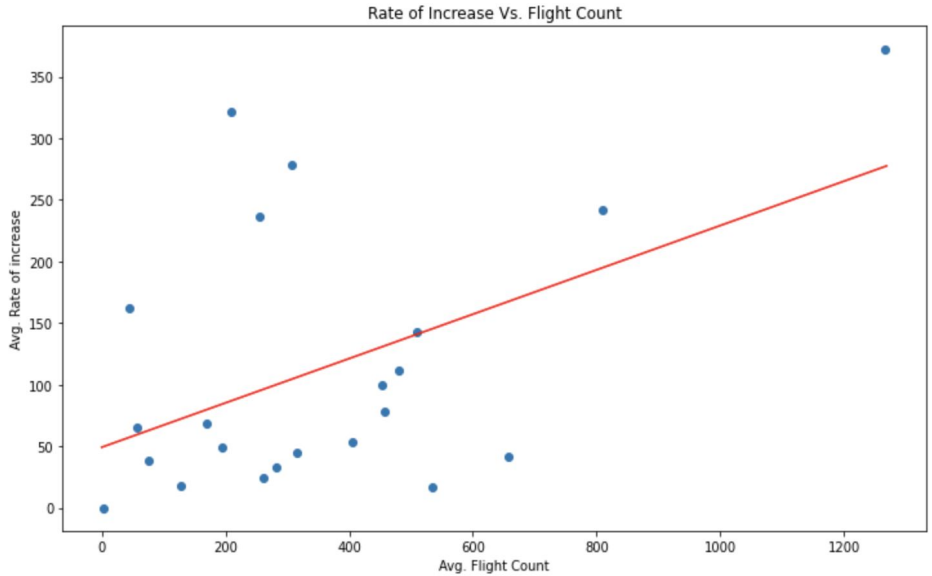
Covid Rate of Increase Vs. Flight Count

With Outlier



Pearson's Correlation : 0.28

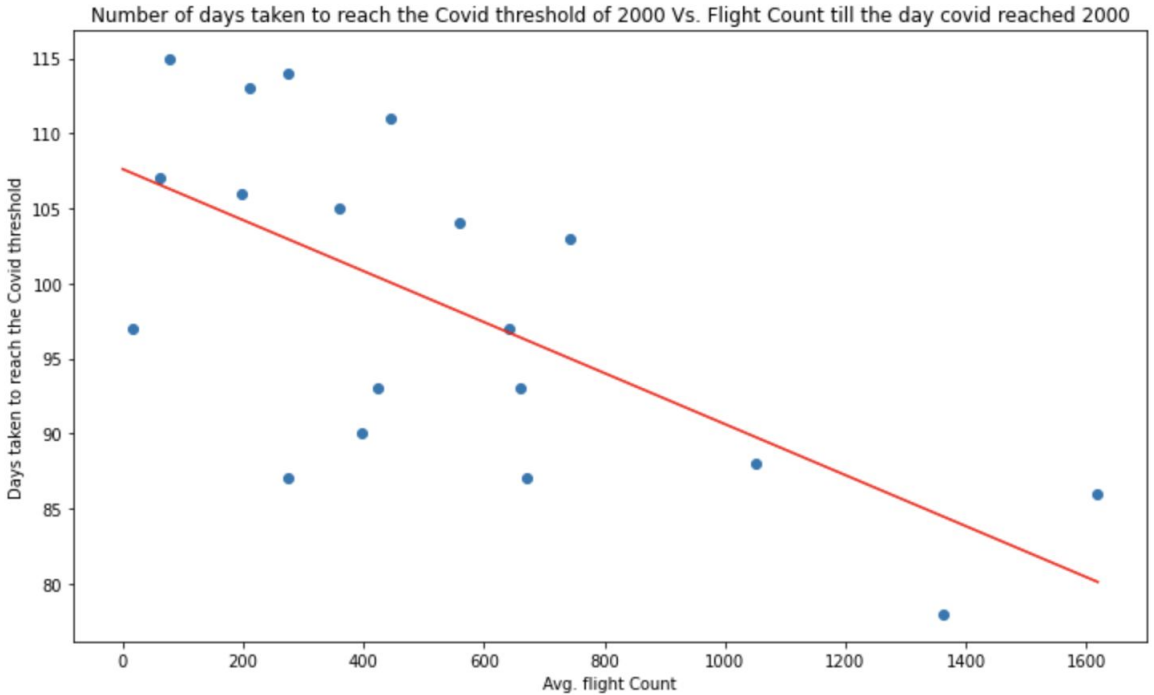
Without Outlier



Pearson's Correlation : 0.48

More Flight, more Rate of Covid Increase

Days to reach covid count of 2000 Vs. Flight Count



Pearson's Correlation:
-0.66

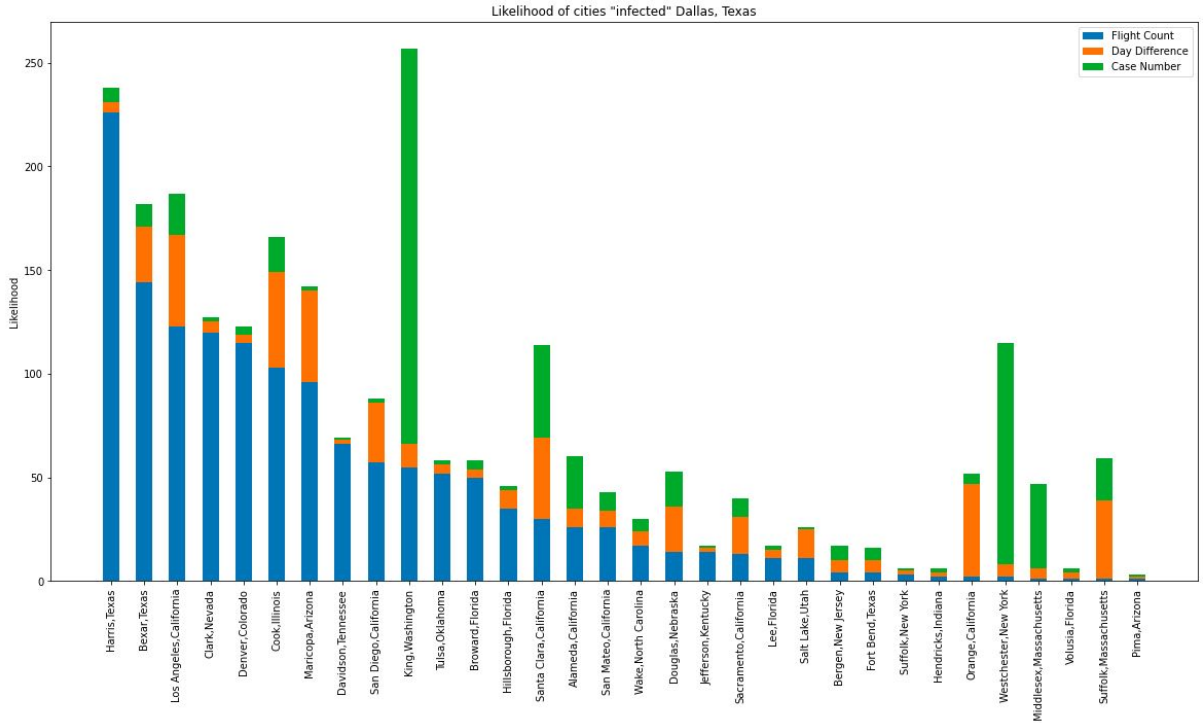
More Flight, less days taken to reach 2000 Covid cases

Air Travel impacts COVID-19 Spread - Analysis 2

- Process
  - Select a target city
  - Find out the date (target date) when COVID-19 first started to spread in target city
  - Filtering out flights that landed in the target city within two week before target date
    - Two week because of incubation period of COVID-19
  - Group these flights by county and state of the origin city and output the count
  - For each county, find out the day difference between it's COVID-19 start date and target date
    - Larger day difference potentially allows more infectant travel to target city
  - For each county, find out its case number on target date
    - Larger case number increase the probability of infectant travel to target city
  - Evaluation:
    - Flight count from each county
    - Day difference of each county
    - Case number of each county

Air Travel impacts COVID-19 Spread

- Example result for *Dallas Texas*

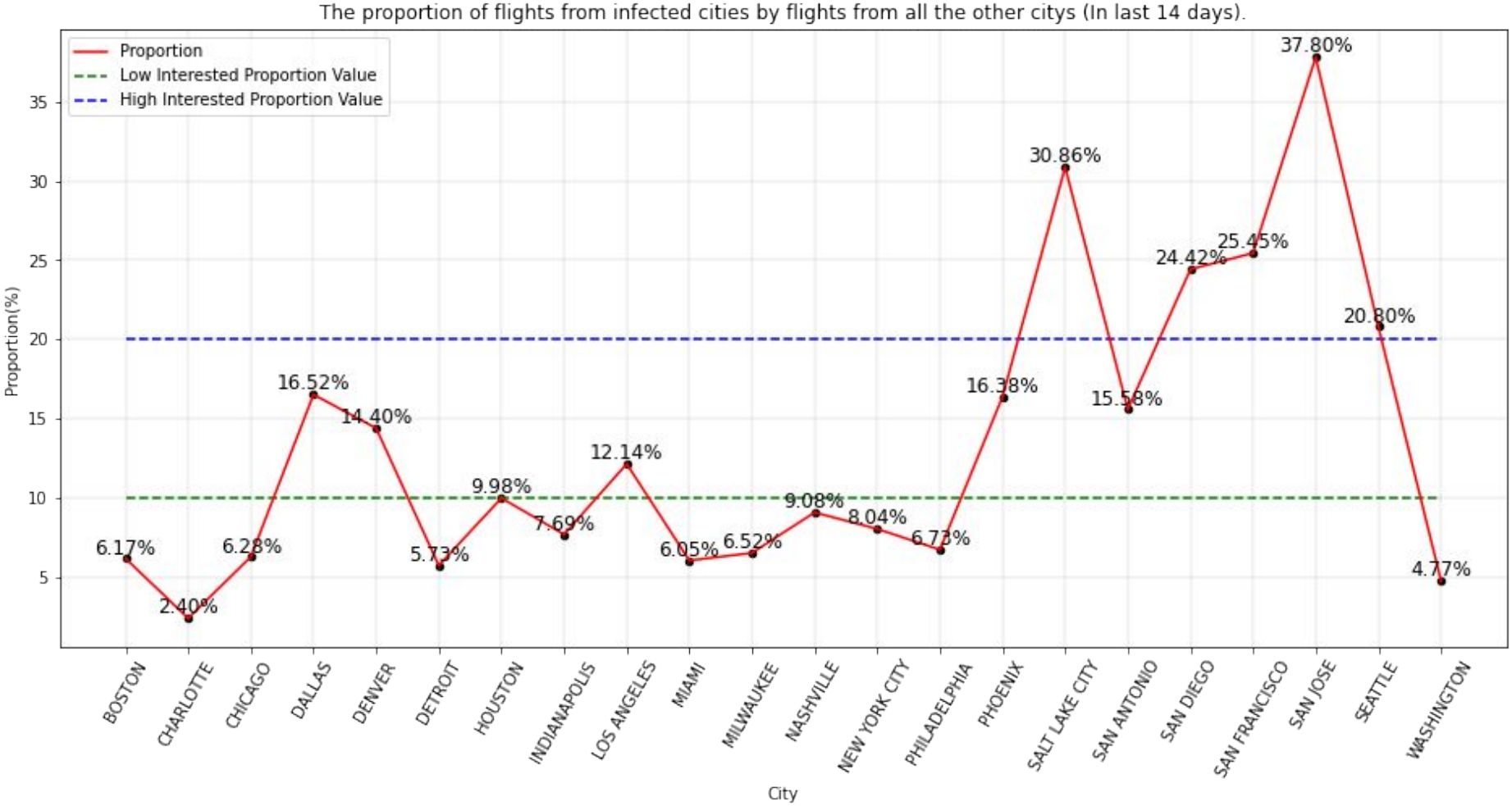


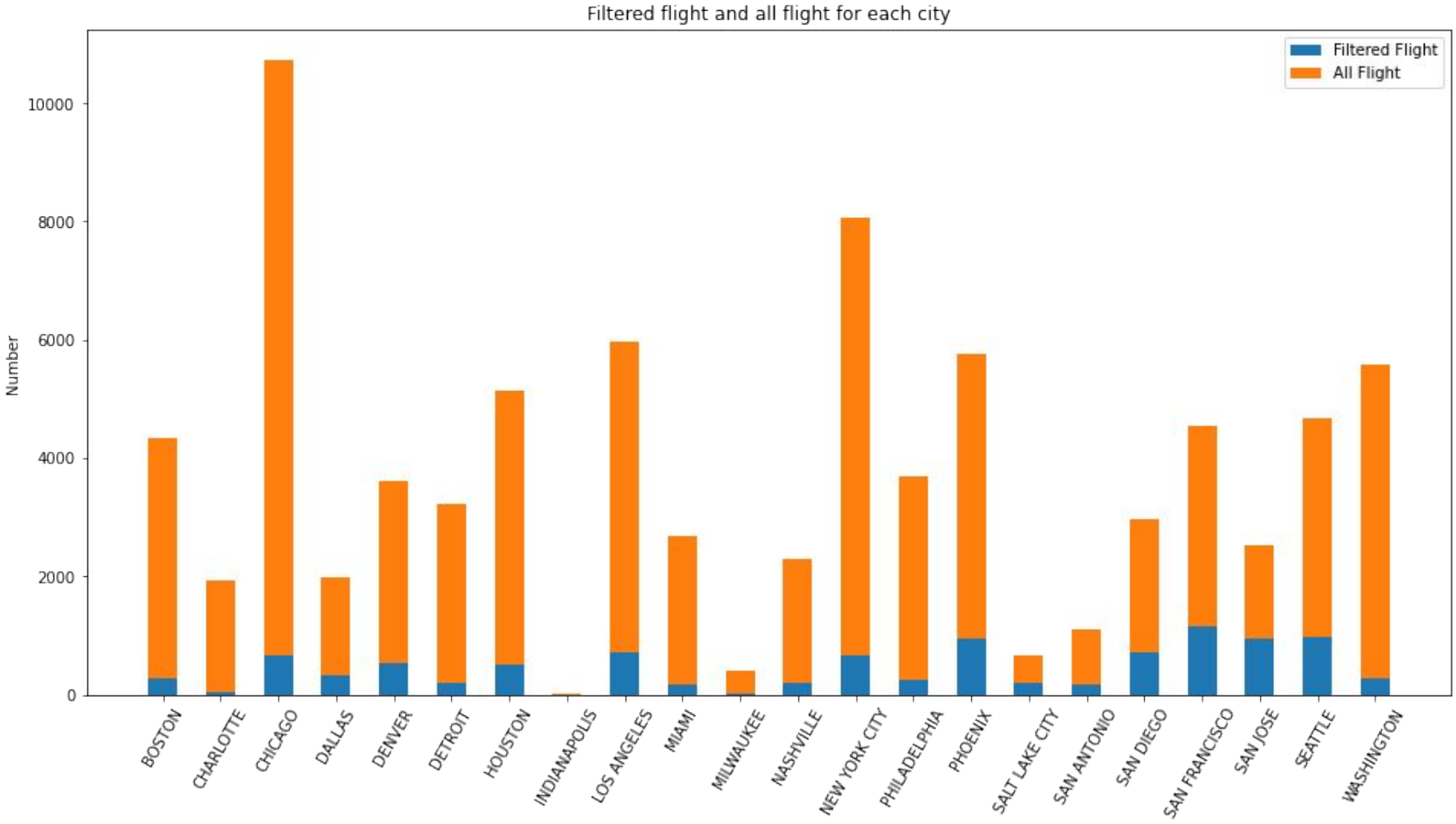
Air Travel impacts COVID-19 Spread - Analysis 3

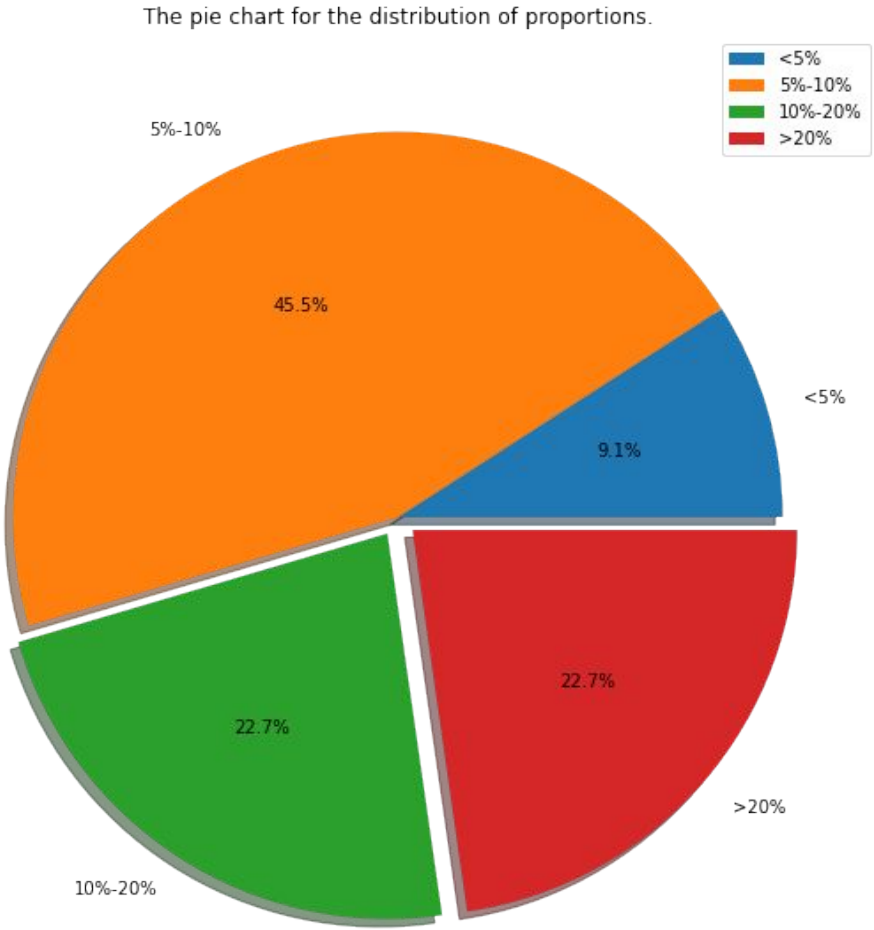
- Process (in more numerical way)
  - Goal: show how the airlines help the spread of Covid
  - Based on the previous process:
    - Select a target city and its target date (first case appear) → filter out the flights to target city within two weeks before the target date
  - For all these flights, figure out where they came from and group them by county/state
    - Remove the flights from the cities outside the US
    - Count the total flights to target city (as denominator **N_all_flight**)
  - Filter out flights using the county/state info where the disease spreaded before the target date
  - For all these county/state and the flights from these place to target city, we compute the average infected case. If the average is higher than a threshold, then this county/state and the corresponding flights are candidates.
  - For the candidates, count the total flights to target city (as numerator **N_infected_flight**)

Air Travel impacts COVID-19 Spread (Cont)

- Numerical formula
  - According to the previous slide:
    - **N_all_flight**: sum of flights landing on the target city in last two weeks
    - **N_infected_flight**: sum of flights from city where the disease has already spreaded out
  - Formula:
    - ***Flight_ratio*** = **N_infected_flight** / **N_all_flight**
  - Explanation: we use this proportion to represent how possible the spread of disease in the target city is resulted from the incoming flights departing from other “infected” cities. In other word, the first case of target city might be infected by the passengers of flights from other cities.
  - Some visualizations:


The proportion of flights from infected cities by flights from all the other citys (In last 14 days).
Proportion
Low Interested Proportion Value
High Interested Proportion Value
Proportion(%)
City
6.17%
2.40%
6.28%
16.52%
14.40%
5.73%
9.98%
7.69%
12.14%
6.05%
6.52%
9.08%
8.04%
6.73%
16.38%
30.86%
15.58%
24.42%
25.45%
37.80%
20.80%
4.77%
BOSTON
CHARLOTTE
CHICAGO
DALLAS
DENVER
DETROIT
HOUSTON
INDIANAPOLIS
LOS ANGELES
MIAMI
MILWAUKEE
NASHVILLE
NEW YORK CITY
PHILADELPHIA
PHOENIX
SALT LAKE CITY
SAN ANTONIO
SAN DIEGO
SAN FRANCISCO
SAN JOSE
SEATTLE
WASHINGTON


Filtered flight and all flight for each city
Filtered Flight
All Flight
Number
0
2000
4000
6000
8000
10000
BOSTON
CHARLOTTE
CHICAGO
DALLAS
DENVER
DETROIT
HOUSTON
INDIANAPOLIS
LOS ANGELES
MIAMI
MILWAUKEE
NASHVILLE
NEW YORK CITY
PHILADELPHIA
PHOENIX
SALT LAKE CITY
SAN ANTONIO
SAN DIEGO
SAN FRANCISCO
SAN JOSE
SEATTLE
WASHINGTON



Some analysis:

- For pie chart: If threshold = 10%
- Green and red parts
- Then, for almost half of our choosing cities, we can say that there is a good chance that the spread of disease in these cities are resulted from the incoming flights.

Conclusion

- Air travel industry is negatively impacted by COVID-19
  - Amount of flight decreases as COVID-19 case number increases
- Air travel helped the spread of COVID-19
  - We can potentially traceback the origin of the virus in selected cities
  - Only the possibilities, more data is needed to confirm our finding