Assignment Project

Flights and Airports Data

Ting-Hua, Yeh
2022-07-20

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data Understanding
  - EDA with Visualizing data
- Summary of all results
  - Answer Problems with EDA results
  - Establish monitor dashboard

Introduction

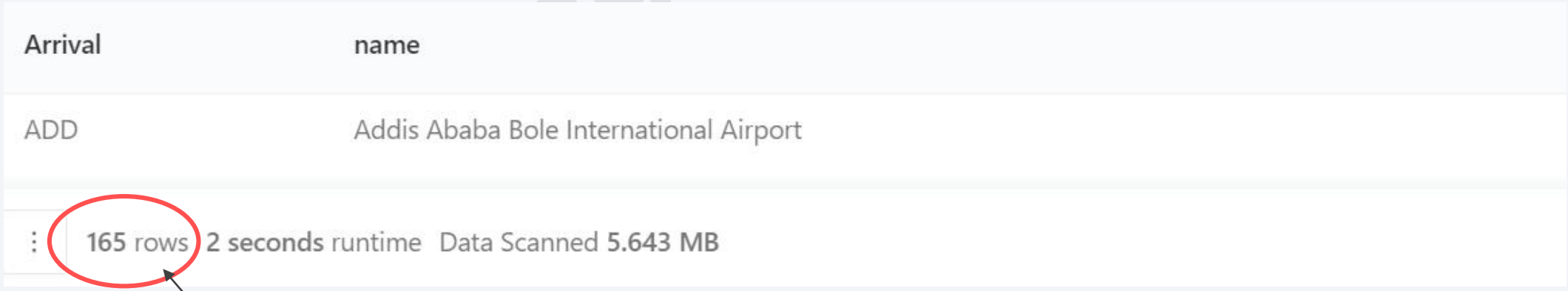
- Project background
  - Traveling by airplane has become a popular choice in the modern world. No matter for the airport, airline, or customer, no one is willing to experience a flight delay. A flight delay would have risk be claimed compensation, and caused inconvenience. The purpose of this project is to build a dashboard to monitor the flight delay information and find out what kind of flight is more likely to be claimed.
- Problems I want to find answers
  1. Which month of a year has more claimed flight?
  2. Which airline has the most claimed flights in the data?
  3. Does more flights means higher claimed rate?
  4. What type of flight looks like the most flight be claimed?

Methodology

Data Understanding -1

- First, I checked the data shape, there are 899,114 flights in the flights table, and there are 39,413 flights are being claimed. In the flights table, there are 123 distinct airlines and 163 distinct arrival airports, but there is only one departure airport, which is Hong Kong International Airport.

  When I try to combine the flights and airports table, I found out there might have duplicate data that have different airport's names but the same iata_code.



There are more than 163 airport names.

Data Understanding -2

- So I located the duplicated data by searching which iata_code appears more than once when I queried for distinct airports and their names.

Arrival	Count
MUC	2
YNT	2

Then, I checked the status of duplicate airports names to decide which to exclude when I queried in the future.

name	iata_code	type	scheduled_service
Flughafen München-Riem	MUC	closed	false
Yantai Penglai International Airport	YNT	large_airport	true
Munich Airport	MUC	large_airport	true
Duplicate--Yantai Penglai International Airport	YNT	medium_airport	true

EDA with Data Visualization

- For different problems, I tried different query and viewed visualization plot if needed.

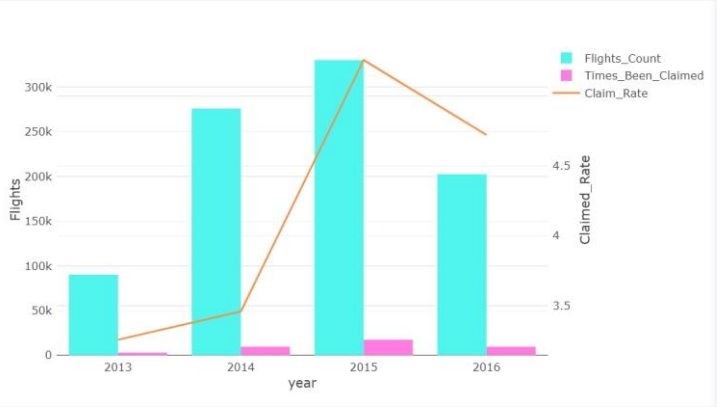
1. Make SQL Query

```sql
SELECT a.Year,
       a.Flights_Count,
       b.Times_Been_Claimed,
       (Times_Been_Claimed/Flights_Count)*100 AS Claim_Rate
FROM
  (SELECT EXTRACT(YEAR
                  FROM flight_date) AS YEAR,
          COUNT(*) AS Flights_Count,
   FROM self_learn.flights
   GROUP BY YEAR) AS a
LEFT JOIN
  (SELECT EXTRACT(YEAR
                  FROM flight_date) AS YEAR,
          count(*) AS Times_Been_Claimed
   FROM self_learn.flights
   WHERE is_claim = 800
   GROUP BY YEAR) AS b ON a.YEAR=b.YEAR
ORDER BY YEAR DESC;
```

2. Examine the Query Result

Year	Flights_Count	Times_Been_Claimed	Claim_Rate
2,016	202,580	9,565	4.72
2,015	330,294	17,363	5.26
2,014	276,084	9,549	3.46
2,013	90,156	2,936	3.26

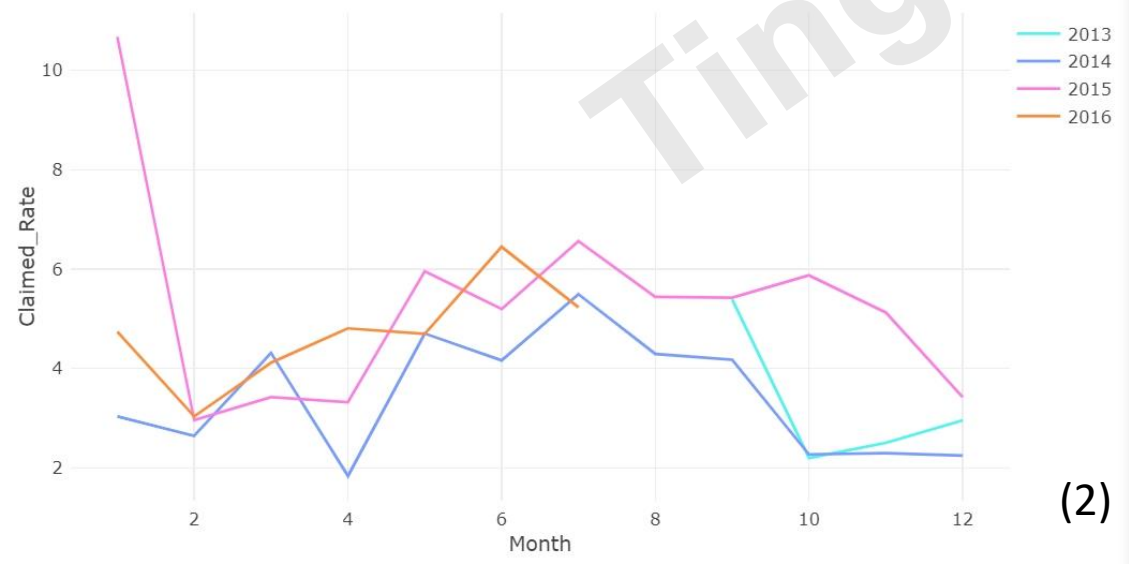
3. Data Visualize



Results

Claimed Flight by Month

(1)



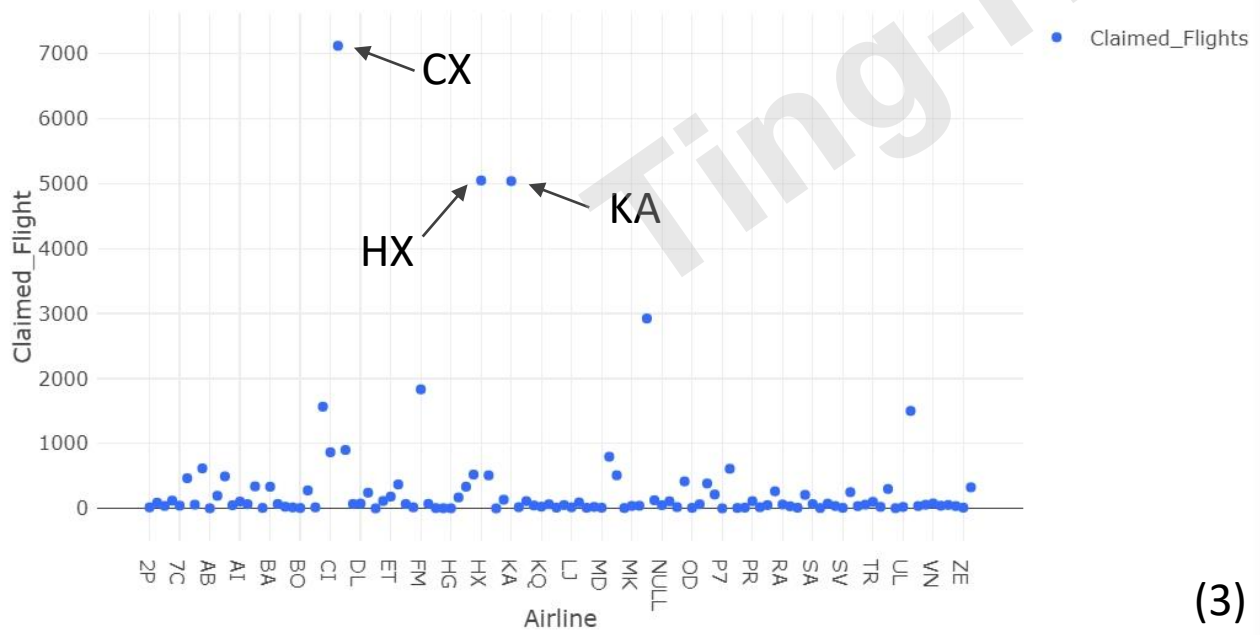
(2)

- For problem 1, I tried to find which month of a year has more claimed flights. From plot (1), it showed that there are more likely to have higher claimed flights in January, May, and July.

- From the claimed rate in each month grouped by year in plot (2), it showed a similar pattern to plot(1).

Most Claimed Airline -1

Airline	Claimed_Flights
CX	7,122
HX	5,049
KA	5,039
MU	2,924
FM	1,833

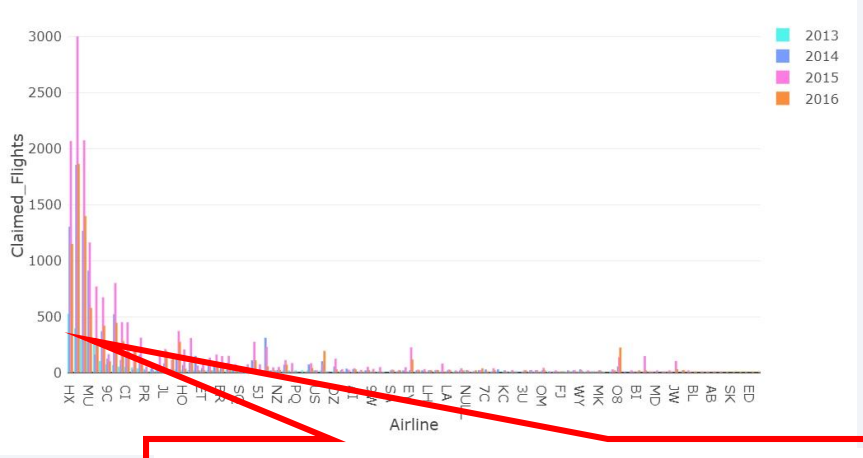


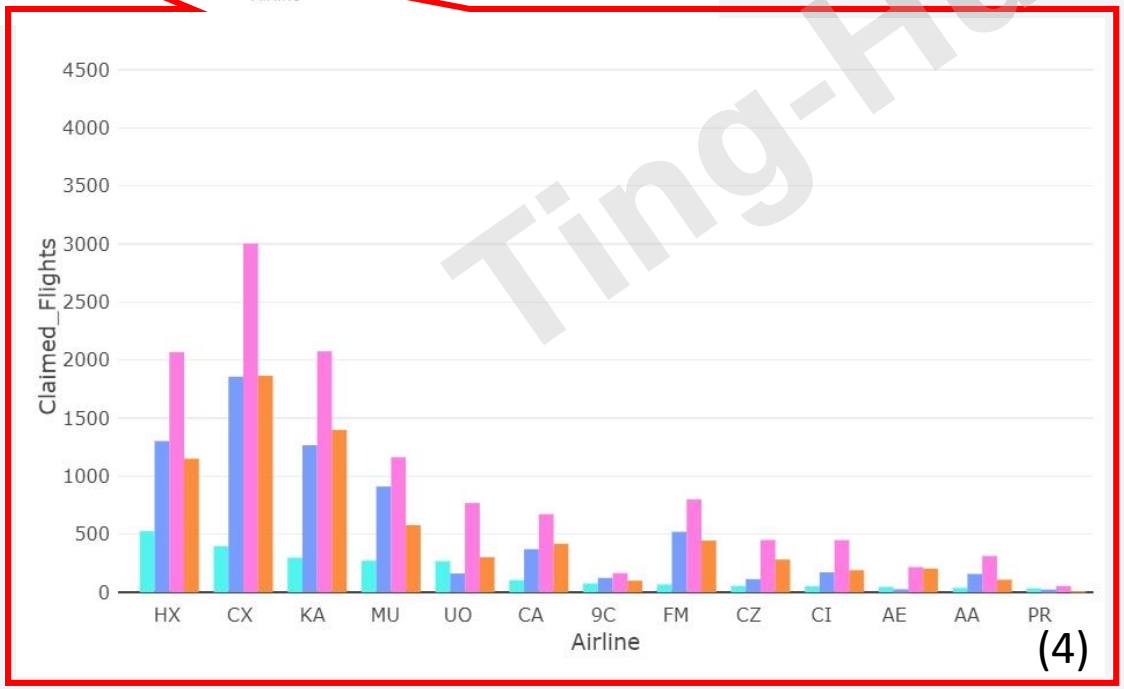
(3)

- For problem 2, the CX, HX, KA airline are the top 3 airlines that have the most claimed flights from 2013 to 2016.

- From the scatter plot, we can see that these 3 airlines' claimed flights number are way larger than other airlines.

Most Claimed Airline -2

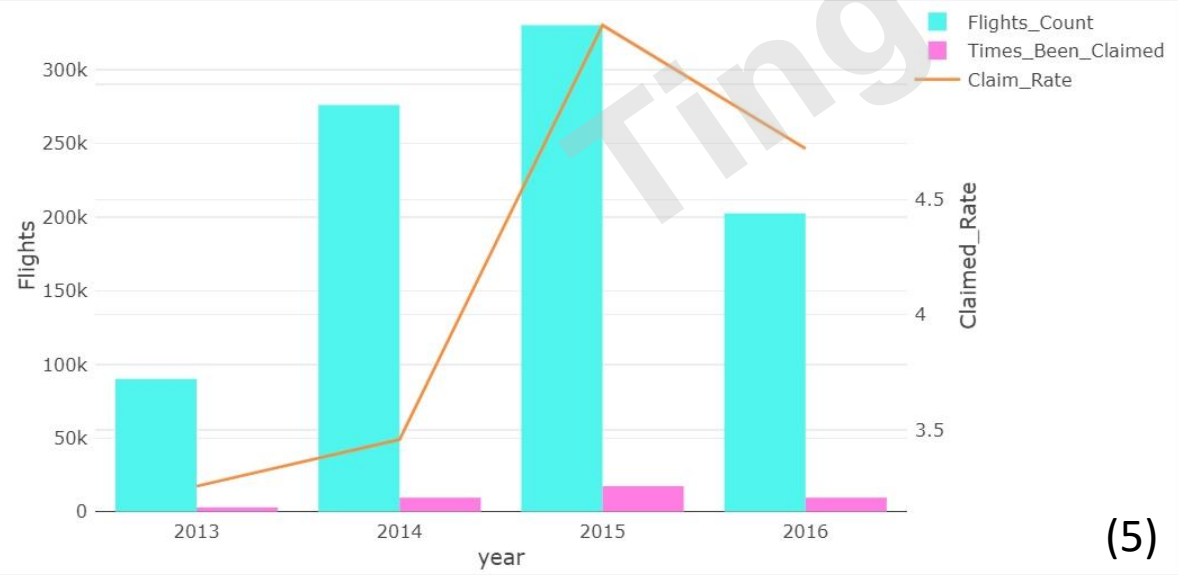




- Then I grouped the claimed fights by different years, it also showed that these 3 airlines always had the top claimed flights in the past different years.

Flights VS. Claimed Rate -1

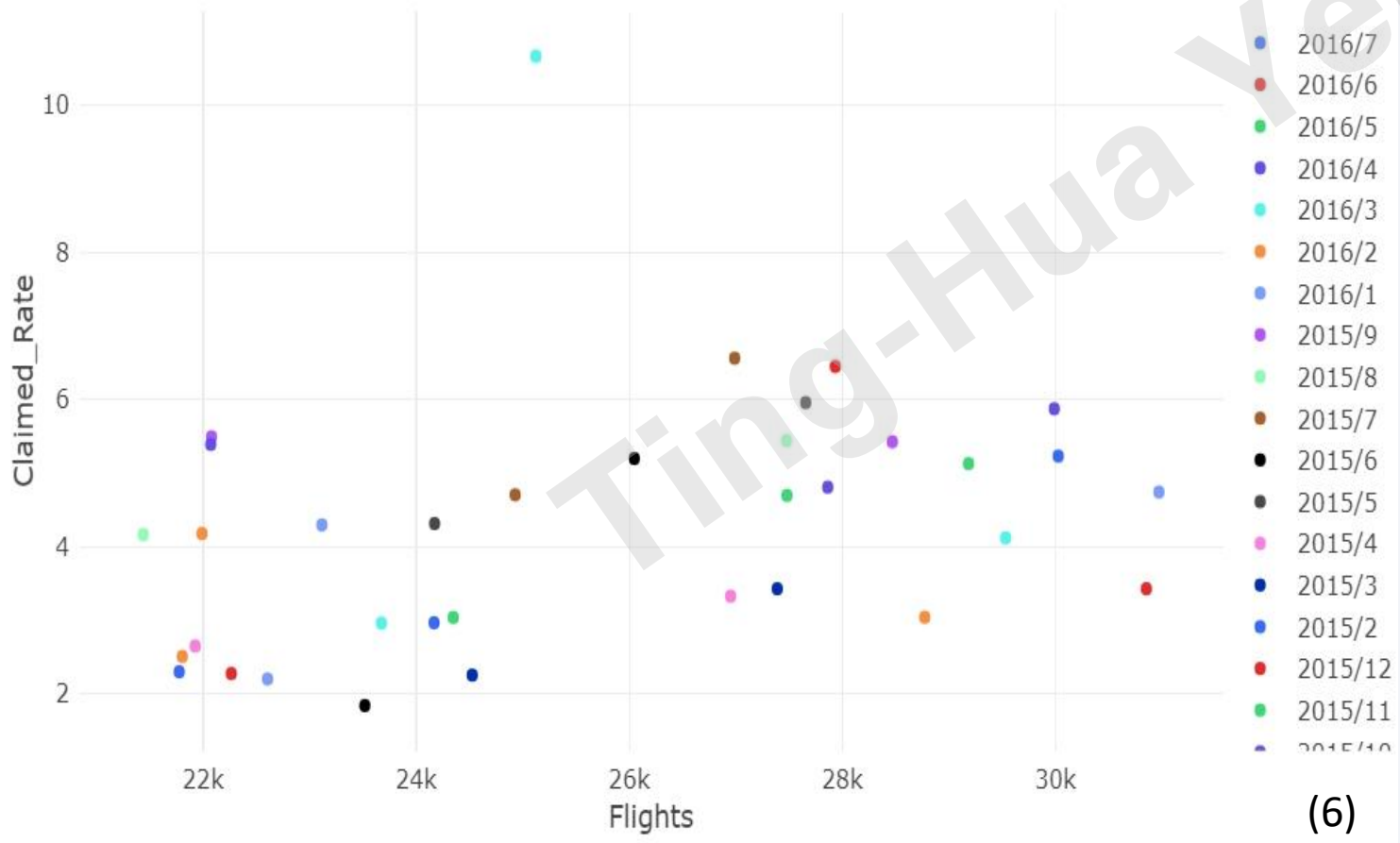
Year	Flights_Count	Times_Been_Claimed	Claim_Rate
2,016	202,580	9,565	4.72
2,015	330,294	17,363	5.26
2,014	276,084	9,549	3.46
2,013	90,156	2,936	3.26



(5)

- For problem 3, it seems reasonable to say that more flights lead to a more claimed rate from 2013-2015. But if we take a look at 2016's data and compared it with 2014's, it doesn't quite followed this pattern. The data of 2016 are only recorded till July, but the claimed flights are already surpassed the whole year's number of 2014. 2016 had higher claimed rate with fewer flights than 2014.

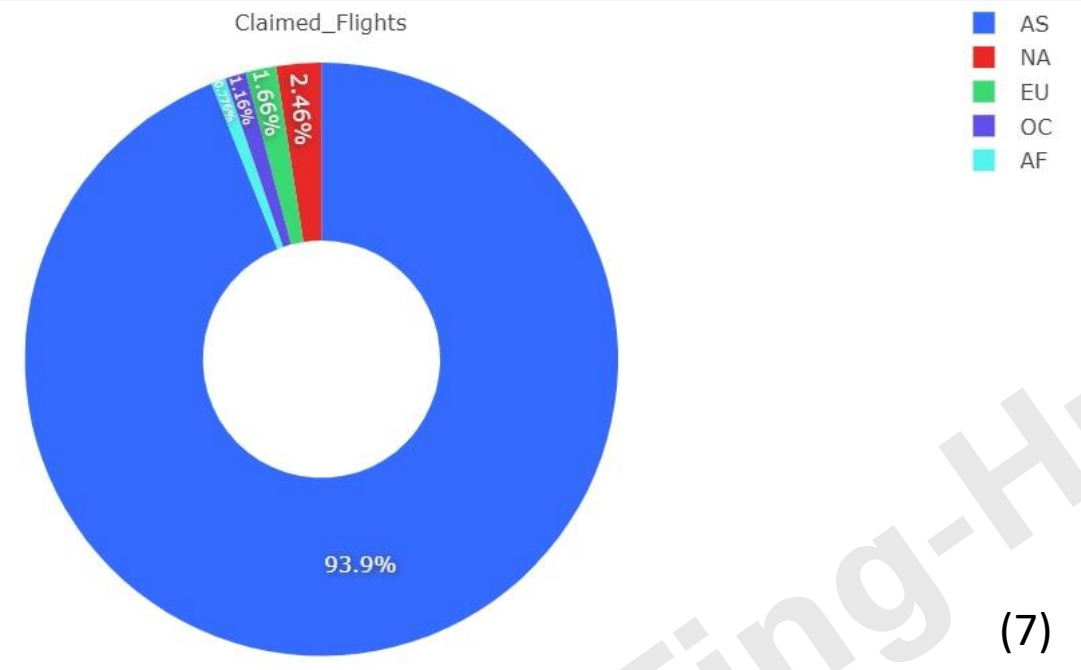
Flights VS. Claimed Rate -2



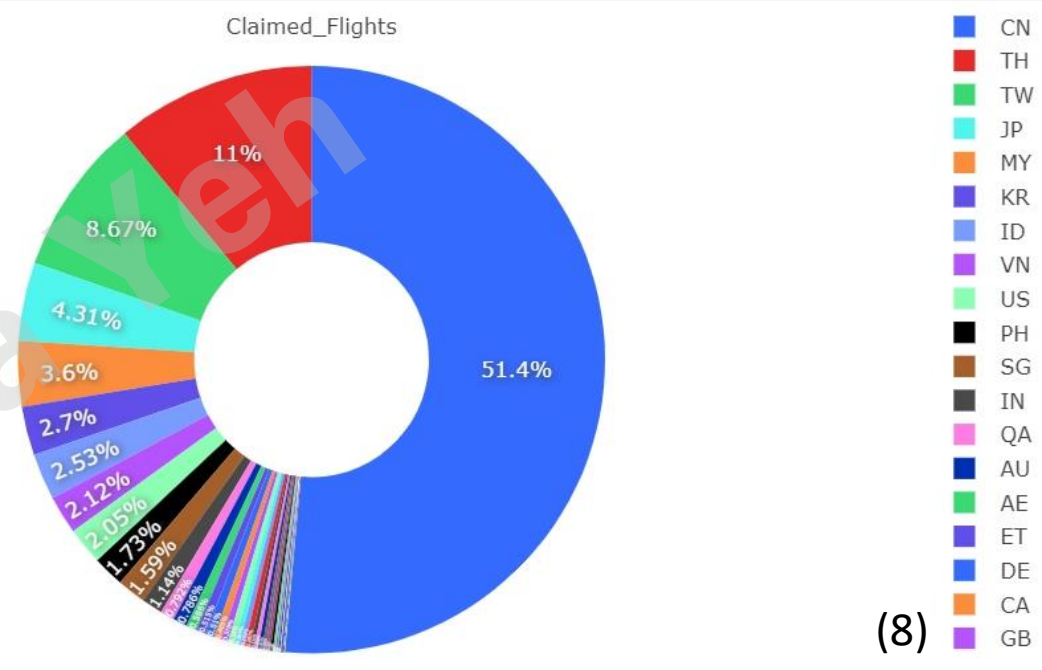
(6)

- So, I grouped the data by year and month. From the scatter plot, it showed that there is no obvious relationship between the number of flights and the claimed rate.

Typical Claimed Flight -1



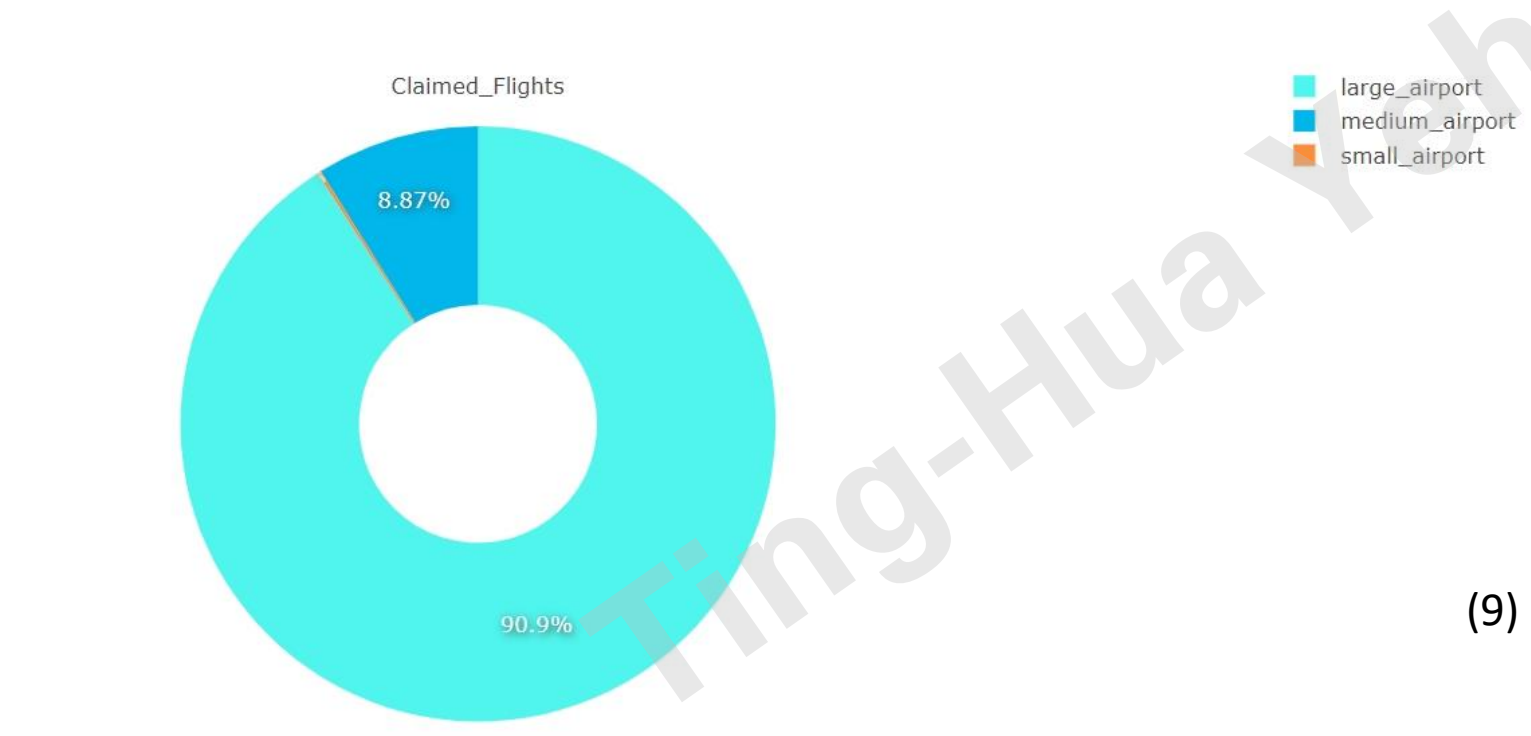
(7)



(8)

- For the final problem, what is usually a claimed flight's characteristic? From plot(7), we can see over 90% of claimed flights were flown to Asia. From plot(8), these claimed fights to Asia, over half of them were heading to China, then Taiwan, Thailand, etc.

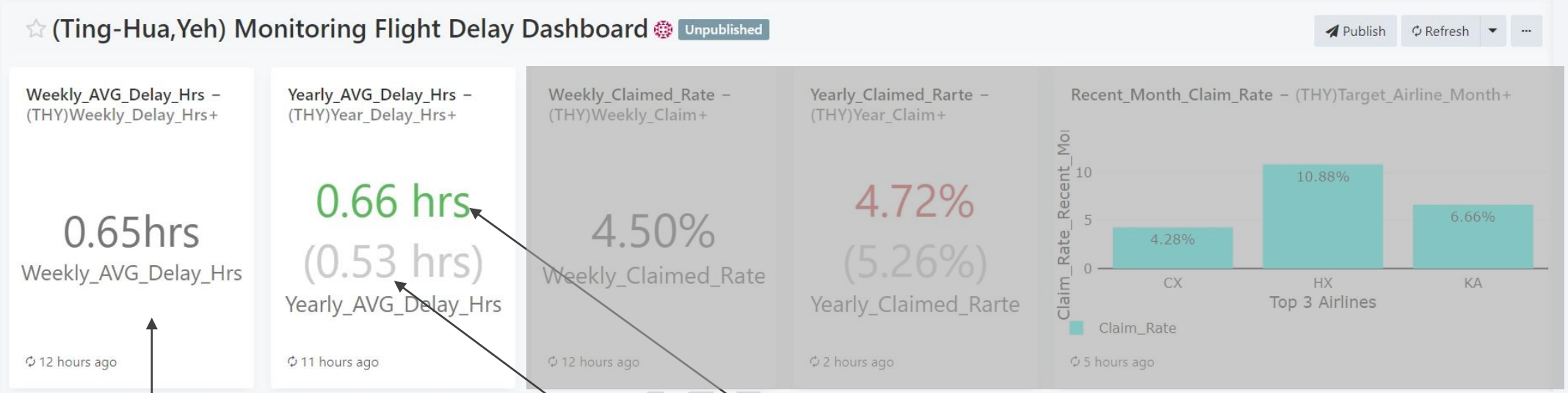
Typical Claimed Flight -2



(9)

- From plot(9), we can see most claimed flights heading to China, Taiwan, and Thailand, were mostly flown to large airports, and rarely went to small airports.

Monitor DashBoard -1



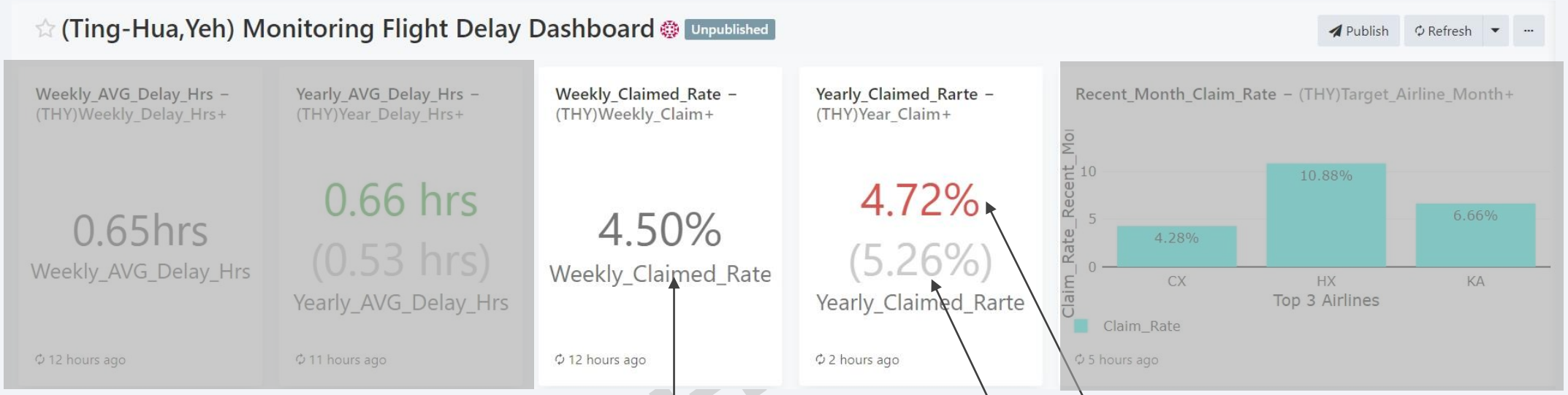
AVG Delay Hours of recent week

AVG Delay Hours of recent year

AVG Delay Hours of last year

- I establish a dashboard with three parts to monitor the flight delay. The first part shows the most recent information. The average delay hours of a recent week, and I can compare the average delay hours of a recent year, and I also can see the average delay hours of last year.

Monitor DashBoard -2



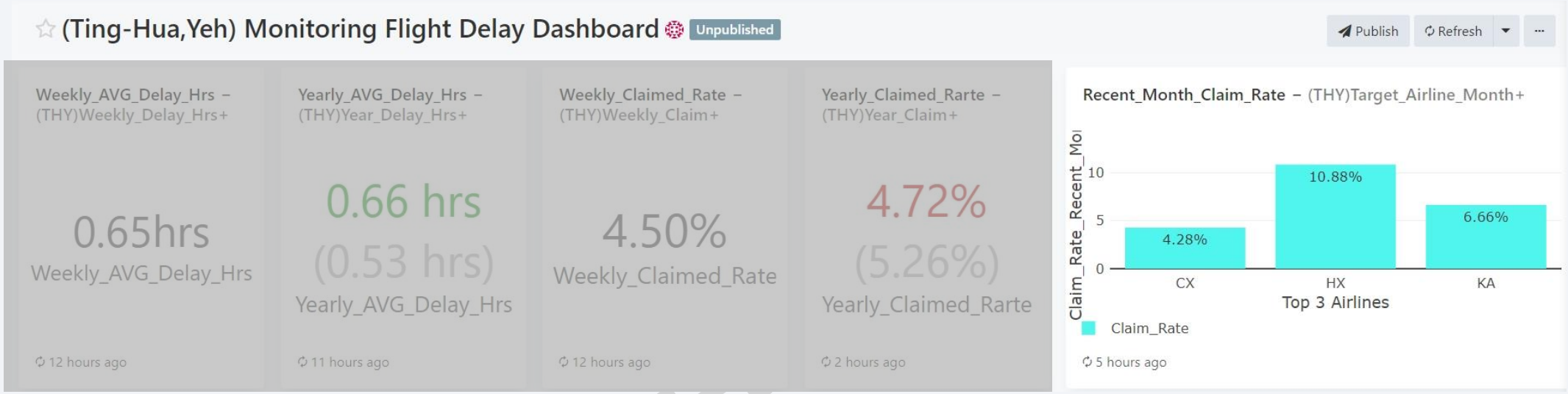
Claimed rate of recent week

Claimed rate of last year

Claimed rate of recent year

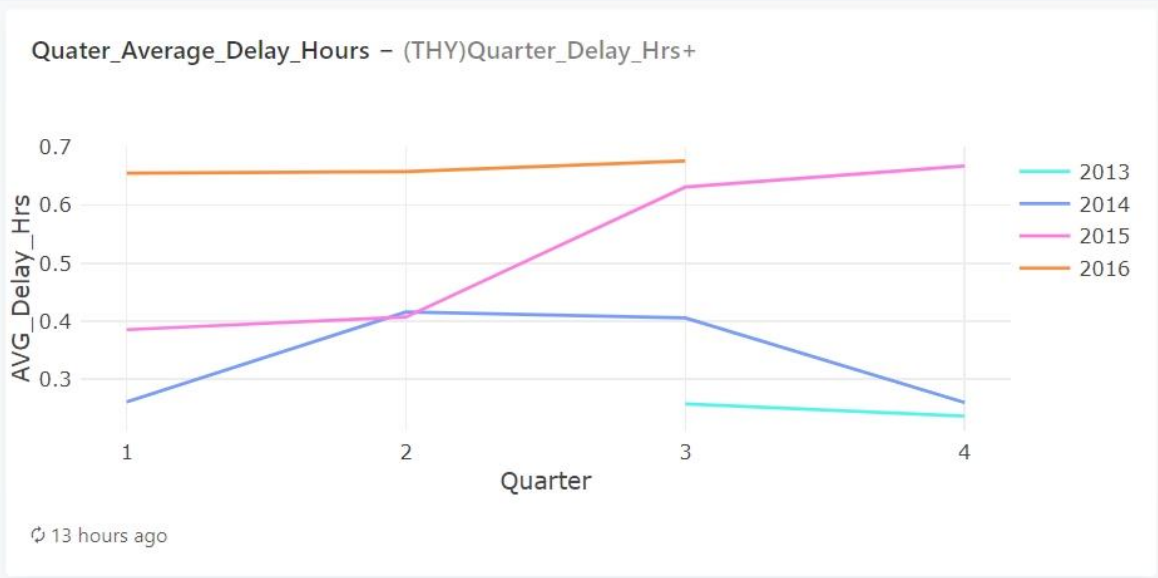
- The middle section of the first part shows the claimed rate of a recent week, and can compare the claimed rate of a recent year, and also can see the claimed rate of last year.

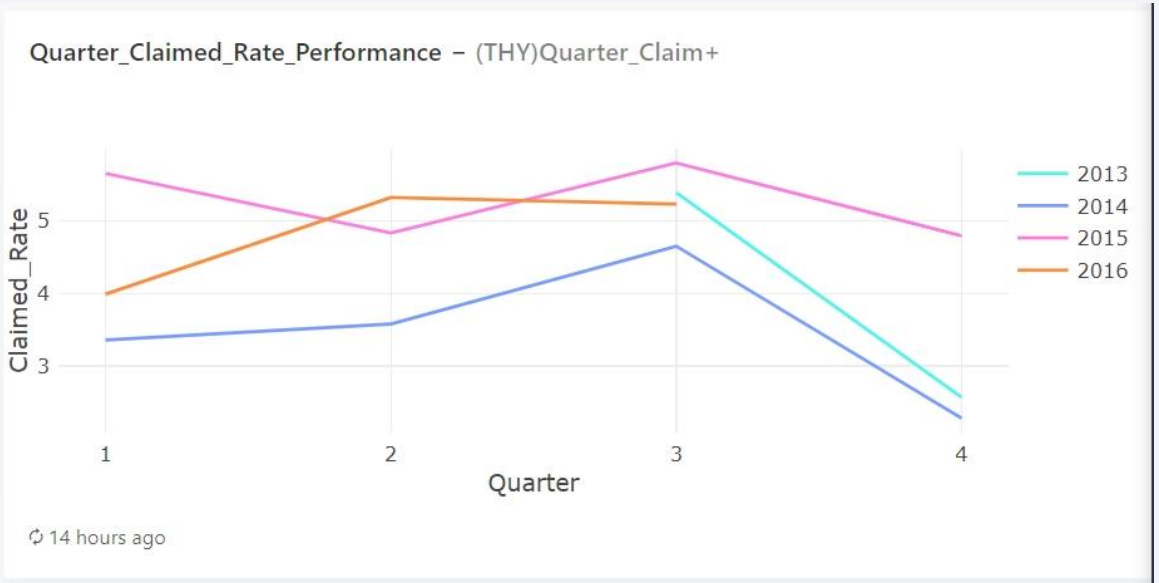
Monitor DashBoard -3



- The final section of the first part shows the major 3 airlines' claim rates in recent month. The purpose of the first section is to monitor the flight delay information by comparing the performance on the different time scale and with the major airlines.

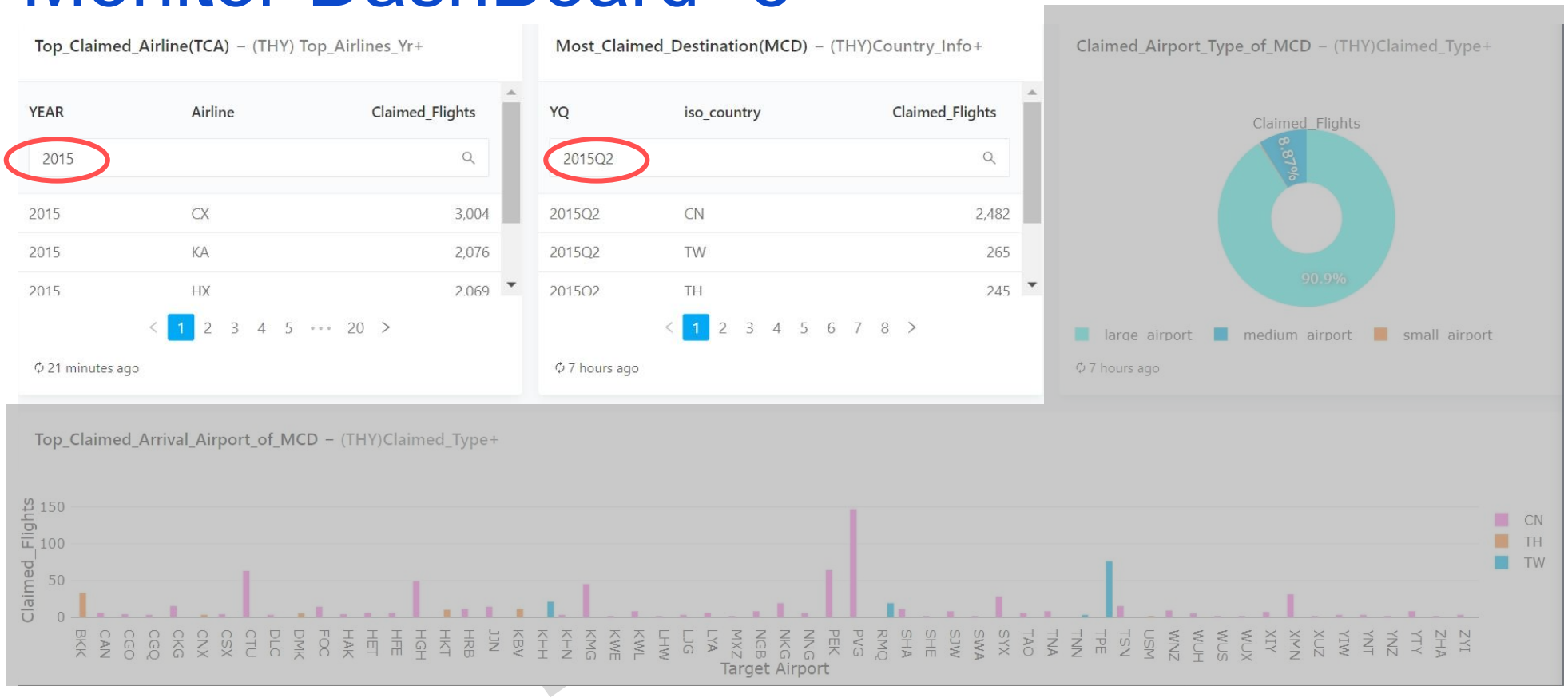
Monitor DashBoard -4





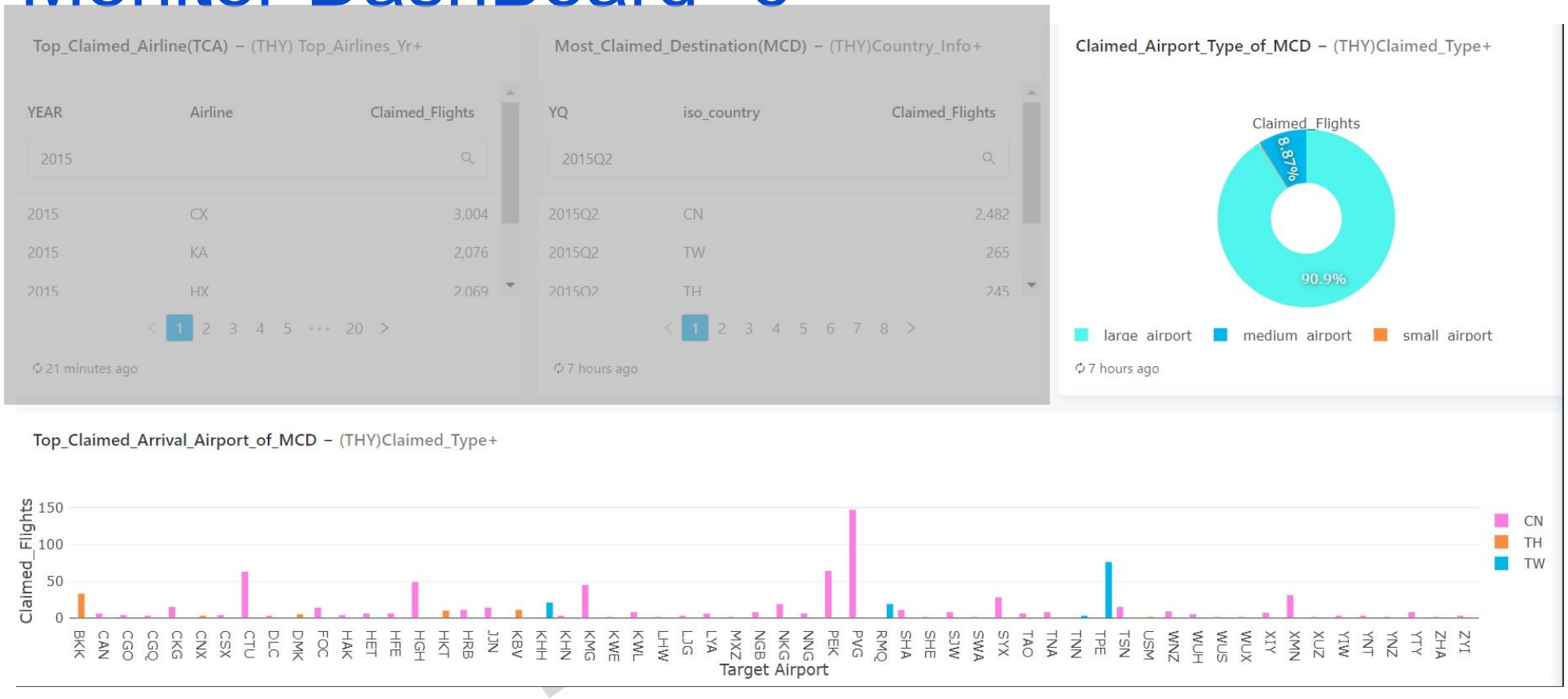
- The second part of the dashboard is to take a midterm trend look of delay hours and claimed rate on a quarterly time scale. From the previous EDA we know there are some months tend to have more claimed flights and claimed rate in a year. I think a quarterly scale will minimize the effect and more precisely than a yearly scale if the year hasn't finished.

Monitor DashBoard -5



- The final part of the dashboard is to monitor the characteristics of the flight most likely to be claimed. The first 2 sections can filter top claimed airlines by year, and the top claimed destination countries by quarter in each year.

Monitor DashBoard -6



- In the previous section, we filtered the most claimed destination countries in a certain period, and the last 2 sections are to show the most claimed airports and airport types in these destination countries.

Conclusions

From this project, I found out there are some regular ups and downs by months in a year about the number of flights. The total number of flights seems to increase over time, but it doesn't have an obvious relationship with claimed rate.

The data is about the flights only departures from HKG airport, there are three major airlines that make a huge impact in this data because they had more flights than other airlines, and their flight also took the main parts of the claimed flights. By analyzing the claimed flights, I found out that most of the claimed flights were heading to Asia, mostly flying to the large airports in China, Taiwan, and Thailand.

A further hypothesis raises that might can be analyzed in the future with other data. Why the majority of records are flying to certain countries? Maybe is because of the distance difference. Why does the large airport tend to have more claimed flights? Maybe is because they have much more loading than other types of airports, or maybe the security issues are more complex in large airports.

Thank You


DEPARTED
GATE CLOSED
GATE CLOSED
GATE CLOSED
GATE CLOSED
GATE CLOSING
GATE CLOSING
NZ3436 03
MH5468 08
LX4154 03
SK8019 03
GA9450 04
VA5872 04
NZ3456 03
03
FY7322 02
S05604 02