Athens University of Economics and Business

Department of Informatics

M.Sc. in Data Science

Data Visualization and Communication

Report for Final Project

Students:

Alexandros Chasapis (F 335 19 14)
Konstantinos Papilas (F 335 19 10)
Apostolos Tamvakis (F 335 19 13)

The Dataset

The dataset consists of airplane passengers filing claims over issues such as property damage/loss and physical injuries.

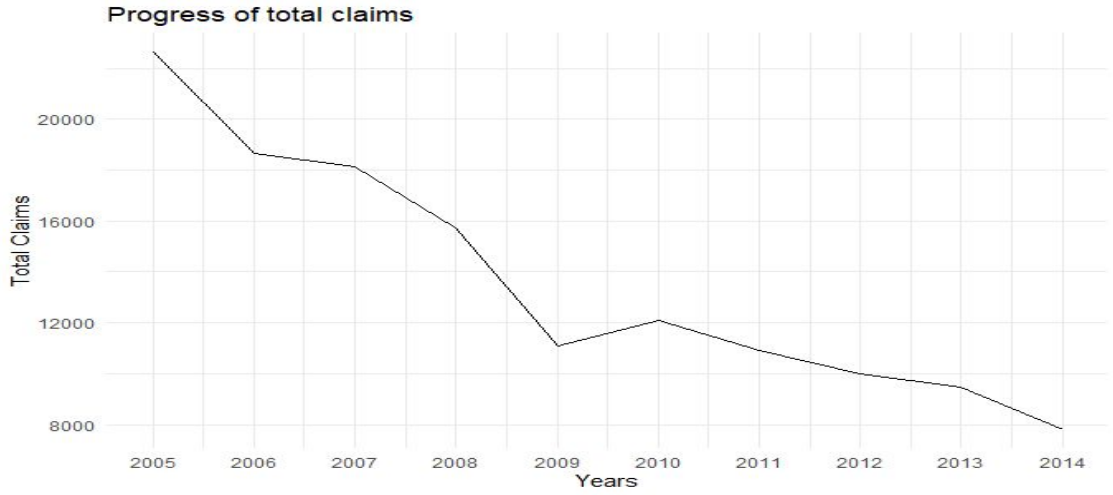
Content

The dataset includes claims filed between 2002 through 2015.

- Claim Number
- Date Received
- Incident Date
- Airport Code
- Airport Name
- Airline Name
- Claim Type
- Claim Site
- Item
- Claim Amount
- Status
- Close Amount
- Disposition

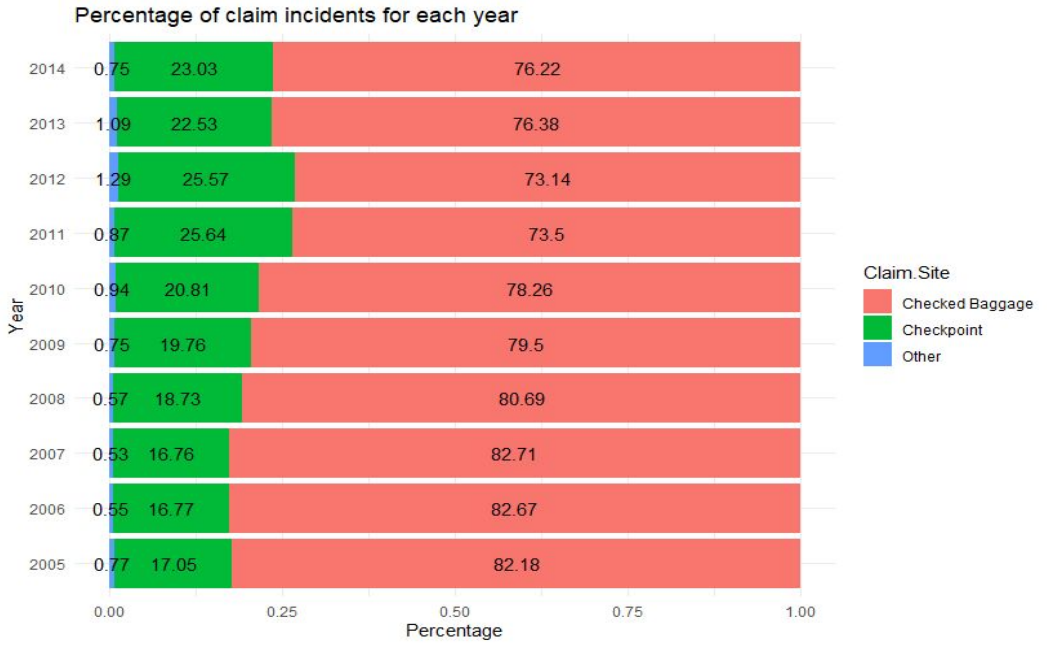
Initial approach on aggregate level.

First of all, in order to get an intuitive view of our data, we are focusing on how the claims are filed in the flow of time.

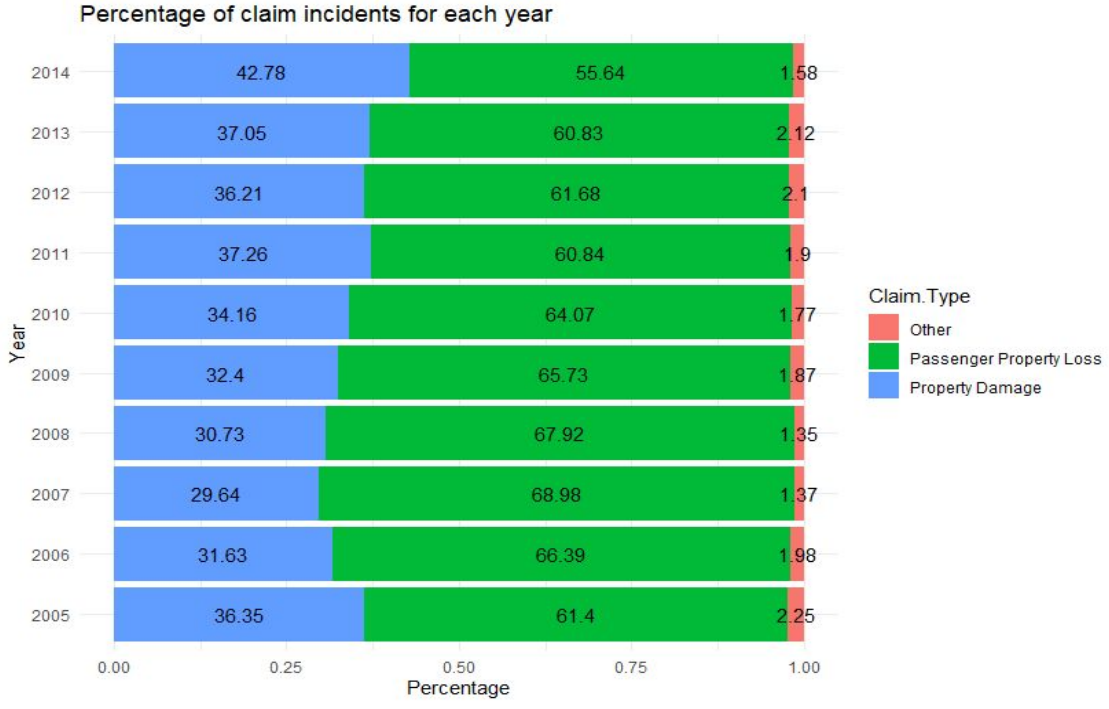


We observe that claims over time tend to decrease which is an indication that the airports or the airlines improved their security policy.

A follow up question occurs….Where do the incidents happen;

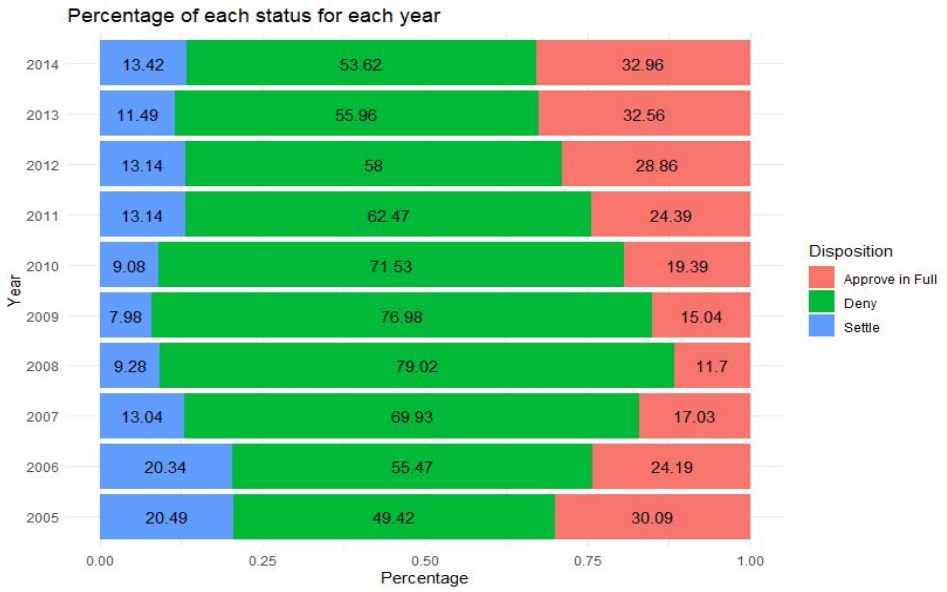


While there are several airport locations, almost in their entirety,the incidents took place in the Checked Baggage area and in the Checkpoint.
How about the type of the claims that are reported. The graph below presents the percentage of these types



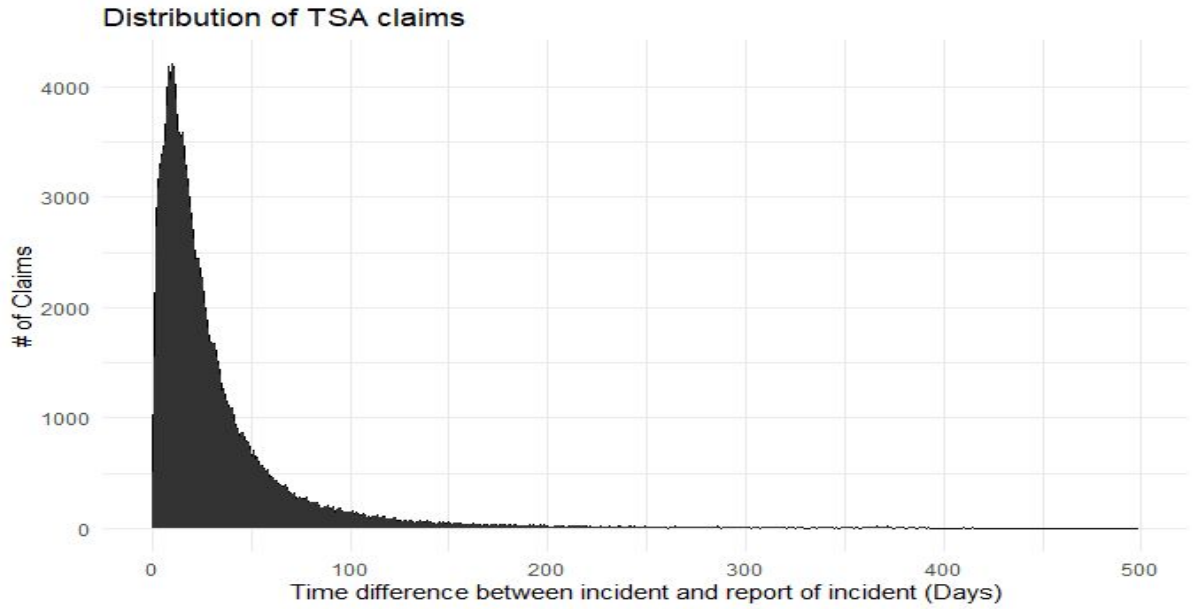
Again, we see that the majority of types consist of two categories: Passenger Property loss & Property Damage. Types as Personal Injury or Complaints and other hold up as about 2% of the total percentage.

Moving on, we have to review how the claims we processed in terms of disposition.

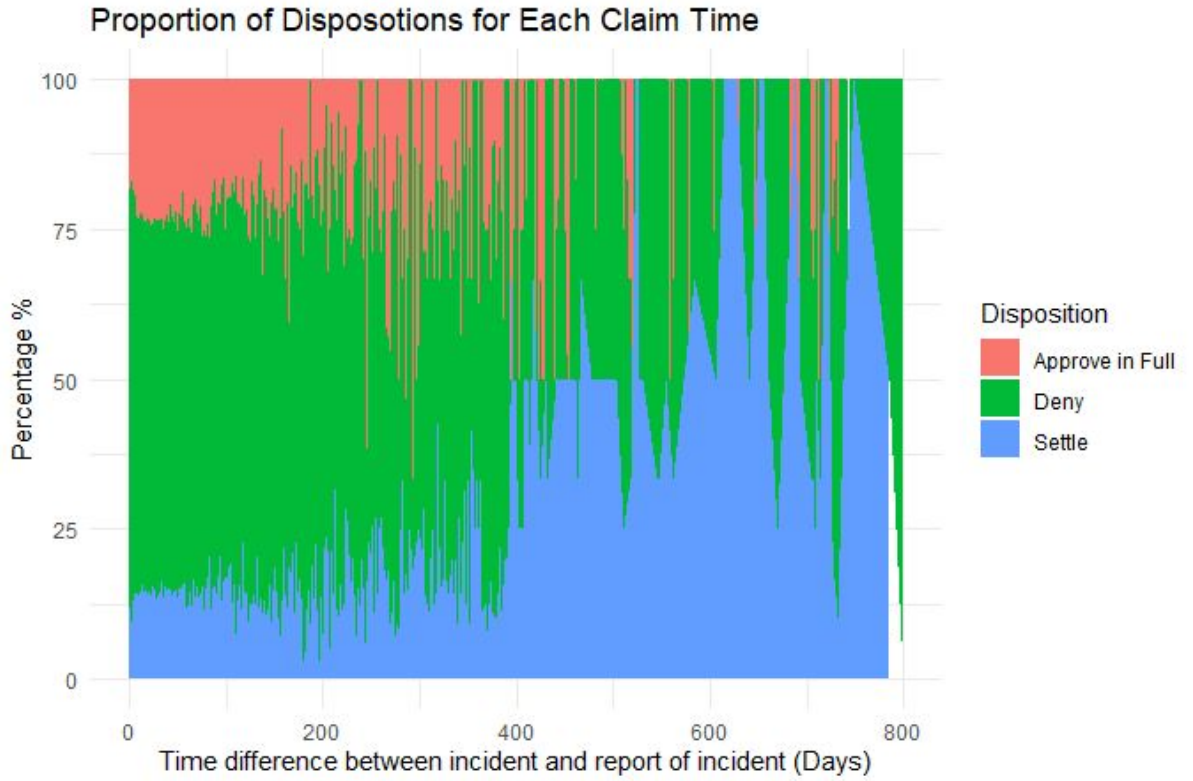


It is apparent that most of the time the claim was denied. An interesting observation is that from 2008 the denial percentage steadily dropped.

Upon processing our data, we observed that time period that the claims where received vary in relation to the date of the claimed incident. The following displays the claim distribution over time.

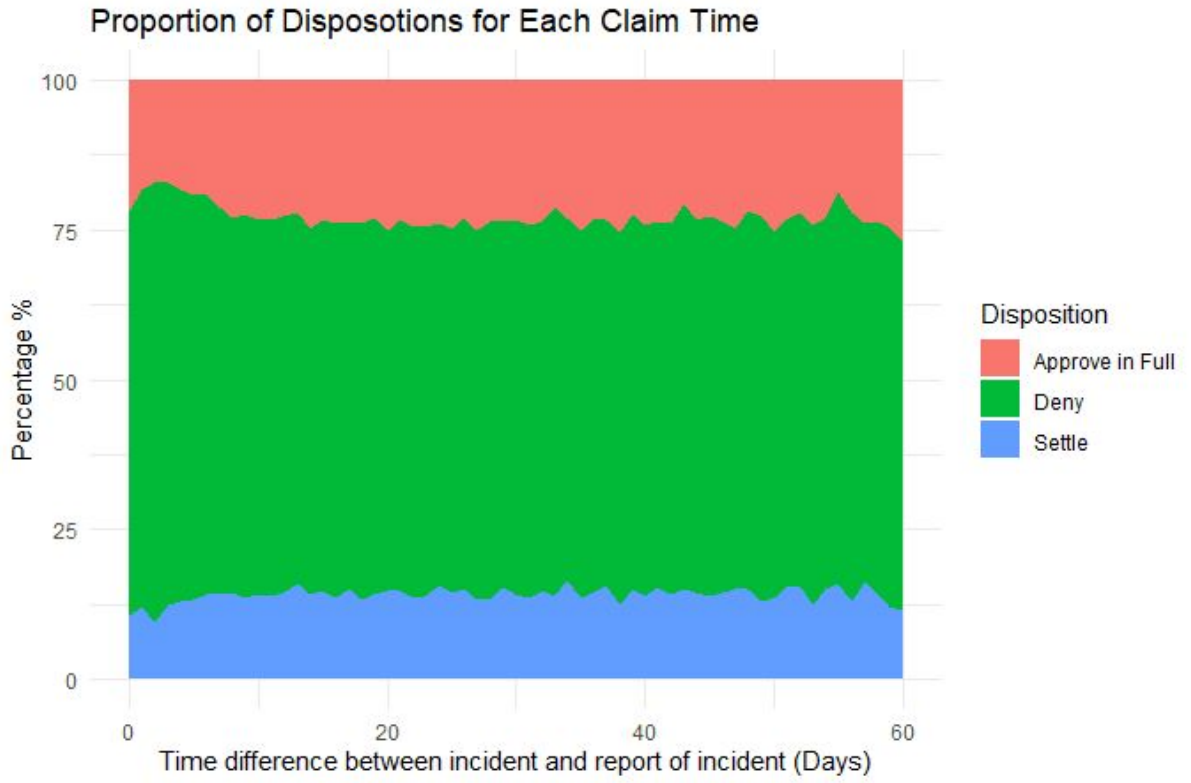


It's interesting why someone would report the incident after such a long period of time. Based on that premise, we calculated the proportion of cases according to their disposition status and plotted them in correlation with the time difference between the report of the claim and the incident.



The data suggest that claims which are filled more than 200 days after the reported incident have an extremely unpredictable chance of being approved or denied.

Since the majority of claims were reported within 50-60 days, we should focus on this subset as well.

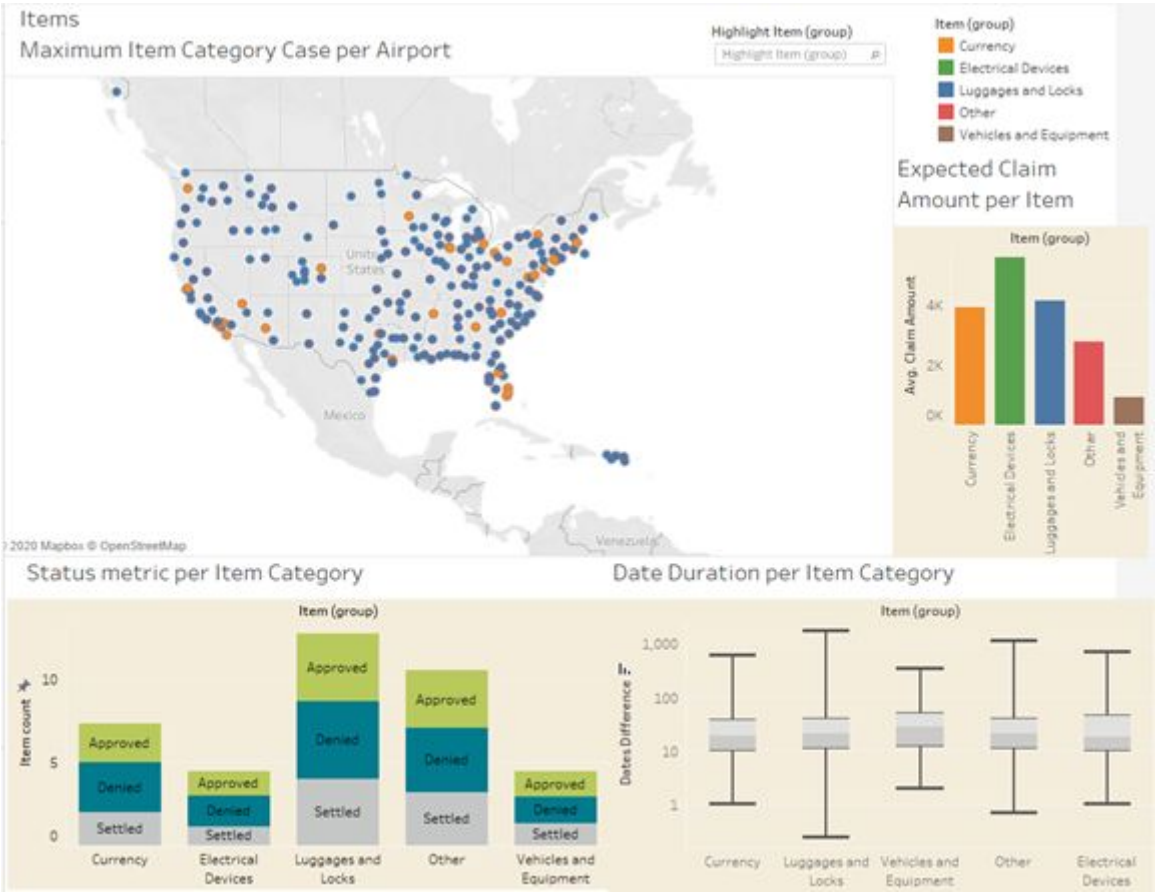


In this situation the disposition percentage appears to remain stable for all categories. From that, we infer that the disposition is not affected by the number of days passed since the incident.

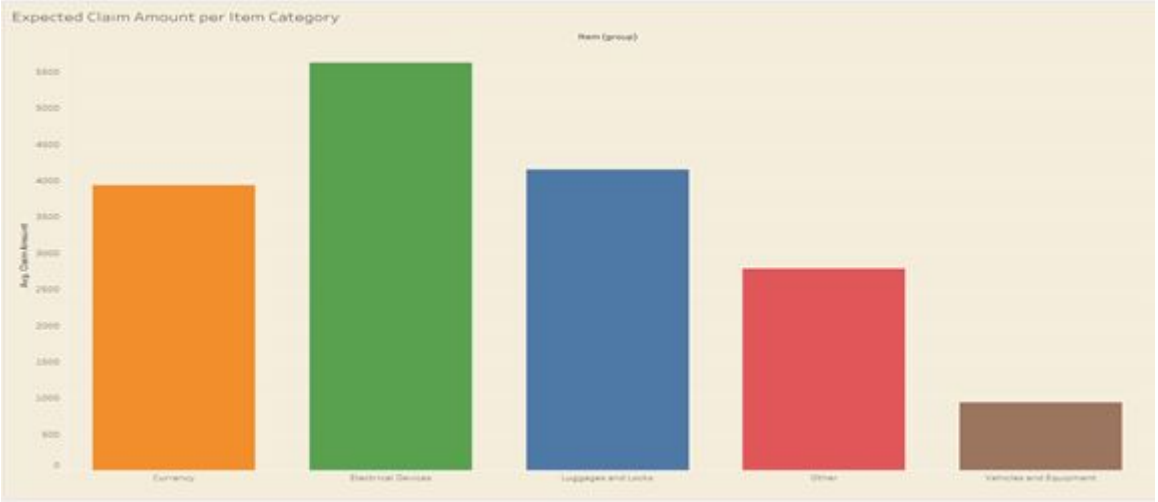
Item-centric approach.

Now let's talk about the data view through Items (we have chosen the top 5 categories Currency(orange) ,Electrical Devices(green), Luggage's and Locks(blue) , Vehicles and

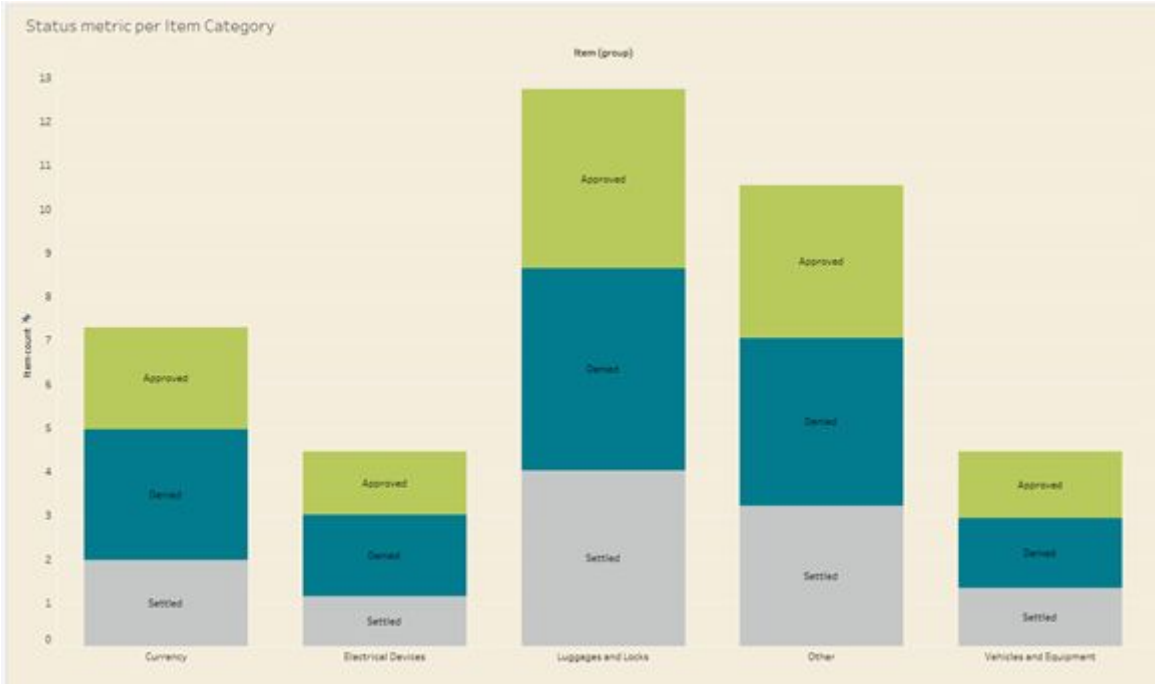
Equipment(brown) and Other(red)) .



Firstly we start with Expected Claim Amount per Item Category . Every bar is giving the average claim amount for every item.

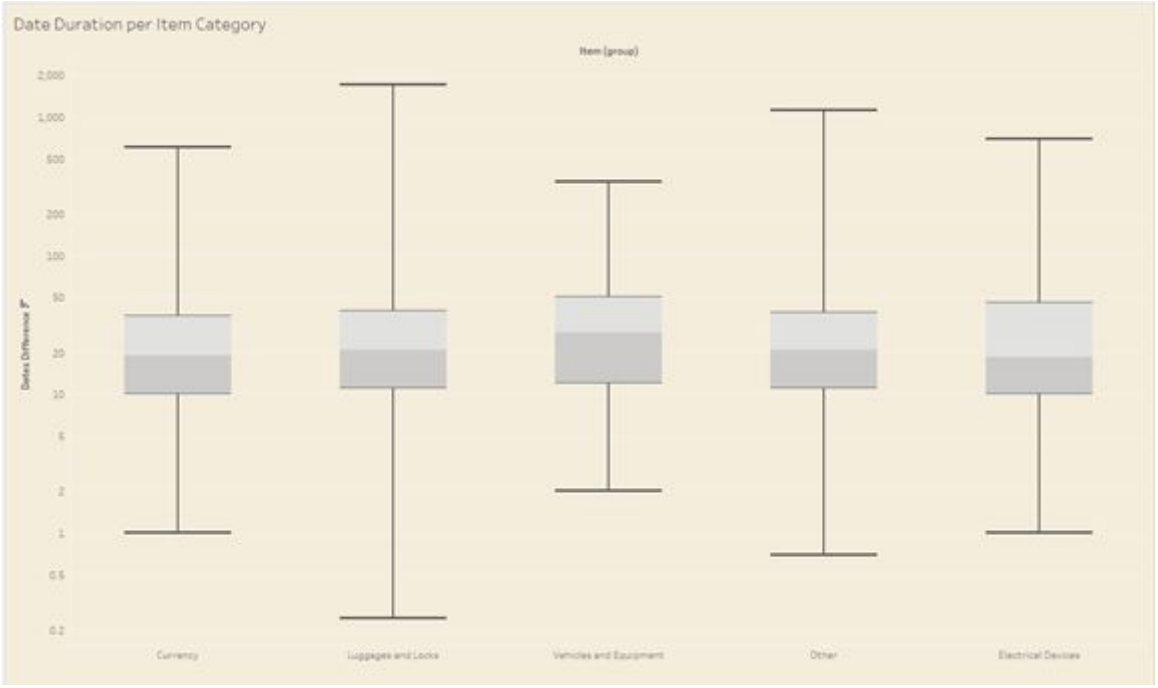


Secondly,



We came across the Status per Category . As bigger is the part of each bar as is the Item count (Logarithmic scaling used at Item count).(Assumption: consider in top 3 Status Category by Item count)

Thirdly,



We have the Date Duration between Date Received and Incident Date for every Item Category. No negative Differences are plotted.(Logarithmic scaling used at Dates Difference)

Finally ,

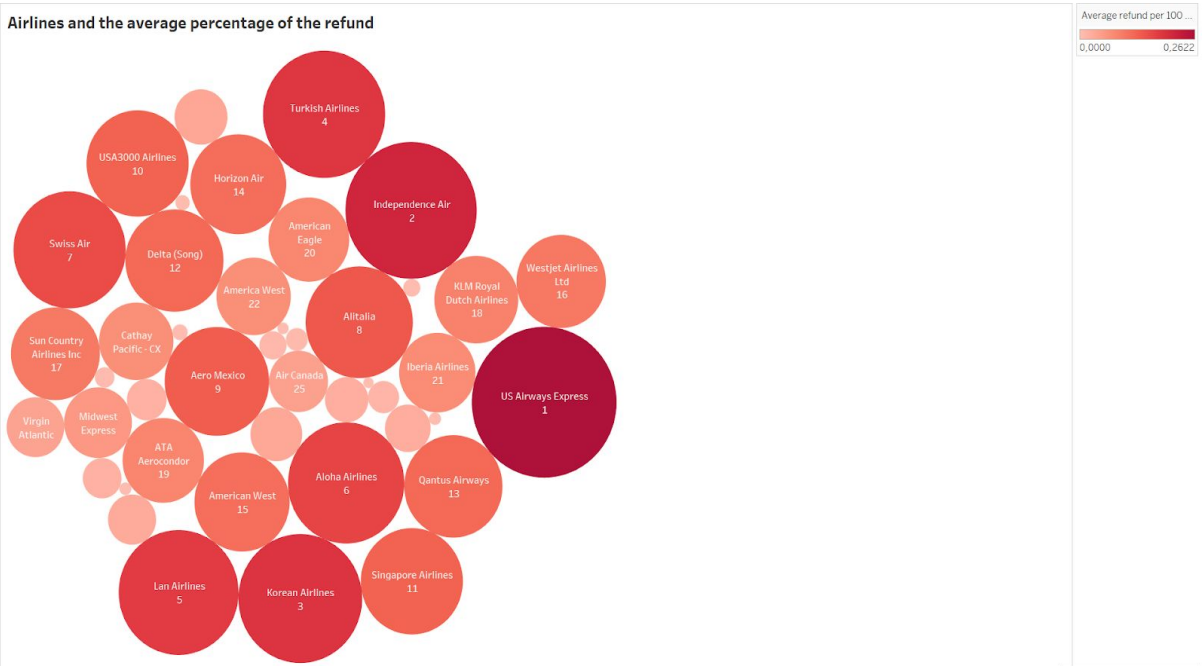
There is a visualization that is giving the most Item Category Case for every airport . Based on our assumption , we took the airports that have at least 10 instances .

View from the airline perspective.

Now, we will focus on the airlines. The plots above show the behavior of each airline on the passengers' claims.

For the first plot, we had to create and apply a new measurement for each airline. That measurement expresses the average percentage of refund of the approved claims for each claim. For example, if passenger X had claimed that the value of his/her/its lost item was $100 and he/she/it finally got $60, the previous explained measurement would have been equal to 0.4.

The next plot shows the airlines with the previous measurement and their ranking based on that measurement per 100 claims. We need to note that we included only the airlines that they have received more than 100 claims in total.



You can see that each airline is represented as a circle. The bigger the diameter of each circle, the higher the value of the explained measurement. Moreover, on the circle you can describe the ranking of each airline based on the average percentage of refund. In details, the best airline in terms of refund is the US Airways Express, while at the bottom of the ranking is Delta Airlines.

Now, we kept the first 10 airlines from the ranking and we splitted them into two categories, the american airlines and the non-american airlines.

The 10 best airlines based on the average percentage of refund

It is very interesting that even almost all the airports are in US, half of the best airlines are non-american. On this plot also, you could see some interactive features (that are not visible on the paper). If you click on the box of each airline, you will browse to the wikipedia page of this airline.

Even if we have found the airlines with the best refund policy, it is very important to see how much time they spend to approve or disapprove a claim. For that purpose, we created a measurement, which is calculated by the difference of the date that each airline took a decision for the claim from the incident date.
For that plot, we have data only for five years (2005-2009).

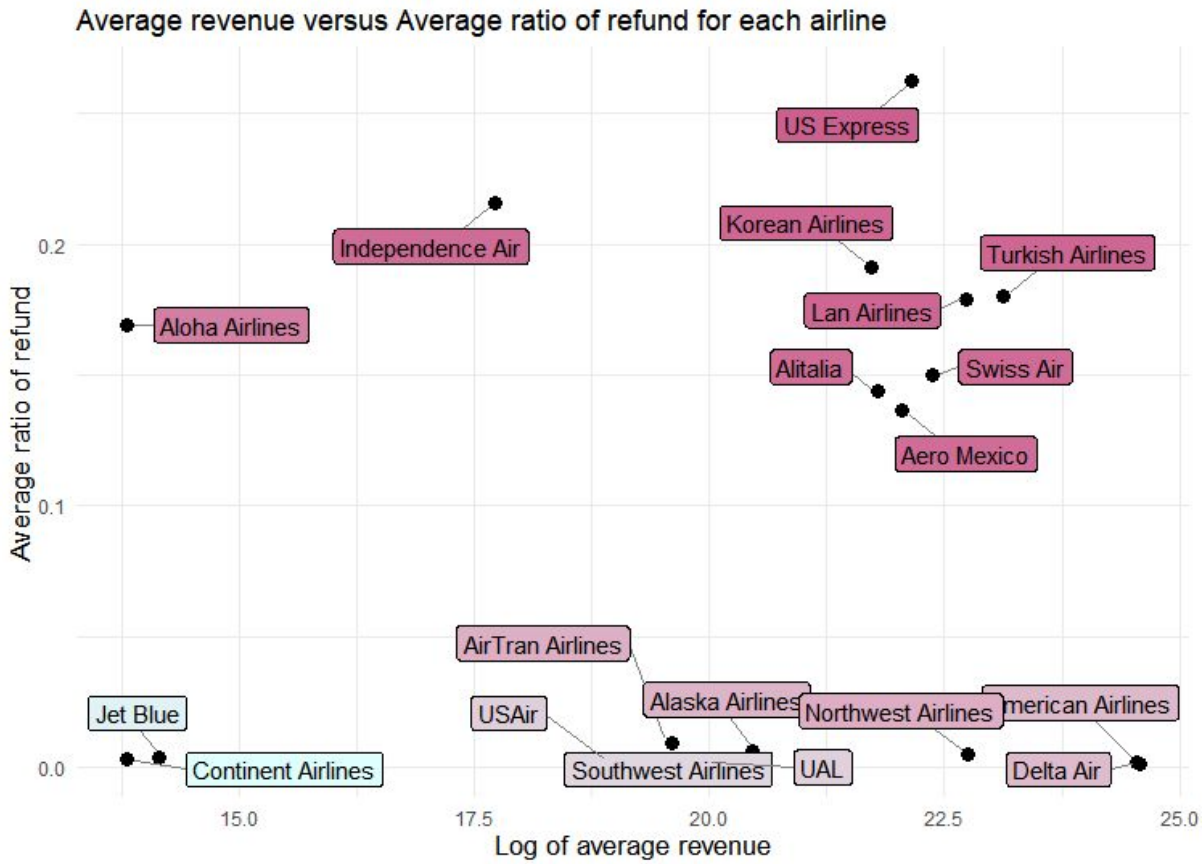
Average rate of difference of the days between the incident day and the day received for each airline

Airline Name	2006	2007	2008	2009
Aero Mexico	-47,2%	225,7%	-50,6%	23,3%
AirTran Airlines	13,3%	-9,4%	1,9%	17,7%
Alaska Airlines	17,8%	-0,8%	-3,4%	27,2%
Alitalia	-19,0%	45,1%	2,0%	10,3%
Aloha Airlines	-5,6%	-3,6%	85,6%	-100,0%
American Airlines	11,5%	1,6%	7,9%	7,3%
Continental Airlines	25,7%	-2,1%	18,3%	5,5%
Delta Air Lines	20,4%	3,6%	-2,6%	13,6%
Independence Air	170,1%	-100,0%	○	○
Jet Blue	25,6%	-15,7%	25,3%	4,3%
Korean Airlines	48,9%	3,0%	40,6%	-33,0%
Lan Airlines	○	216,1%	-77,6%	-55,7%
Northwest Airlines	25,1%	0,8%	2,5%	16,1%
Southwest Airlines	12,9%	-5,2%	3,1%	11,5%
Swiss Air	47,5%	0,2%	-50,7%	-34,0%
Turkish Airlines	-25,0%	-3,8%	79,7%	-24,7%
UAL	21,5%	6,6%	7,3%	13,2%
US Airways Express	14,1%	60,3%	-39,9%	-4,5%
USAir	19,1%	-3,4%	6,2%	1,4%

You can see that we compare that average rate of each year with the previous one. If that ratio for example is smaller during 2007 compared to 2006, means that the airline is improved on the way it handles the claims of each passenger. [1]

[1] The circle on the plot indicates that we do not have data for these years.

For the last one, we selected data for each airline based on their revenues during these years. The main idea is to check if there is a correlation between the revenues and the average percentage of refund for each airline.



As you can see from the above plot, there are some airlines that confirm the previous statement related to the correlation. For example, US Express, Turkish Airlines and Korean Airlines are three of the most rich airlines and at the same time they obtain a high ranking on the average ratio of refund.
Moreover, Jetblue and Continental Airlines have small revenues and their ratios for the refund are being of the bottom of the ranking.
We could not approve or indicate a correlation between these two measurements, but we could find some patterns between our data.

Appendix:

Aggregate level approach: Apostolos Tamvakis

link to related working material: