Team Project: Analysis of Airports' TSA Complaints

UC Berkeley MIDS 2024: Data Science 200, Professor Ysis Tarter

Members: Sooneui Kim, Carrie Li, Emily Lopez

April 17, 2024

Team GitHub: UC-Berkeley-I-School/Project2_Kim_Li_Lopez

Team Google Colab: Airport_Complaints.ipynb

Introduction

We set out to explore the data containing TSA complaint reports from Jan 2015 through Jan 2024. We anticipated seeing some uptick in complaint counts as more people are traveling now than ever before and also from the often publicized bad behaviors by travelers. We also wanted to see if the data can allude to any change in people's behavior and expectations over time.

Data & Source

TSA Complaint Counts: This is our primary dataset and contains the monthly numbers of traveler complaints by airport, category, and subcategory. This data is sourced from the Data Liberation Project who converted the data into csv format for people to analyze more easily. The underlying data was collected and published by the US Transportation Security Administration (TSA) in PDF format.

Airports: This was our initial supplemental dataset which sourced from Data.Gov. We were interested in using its information on Airport codes and their corresponding United States (US) state. After data cleaning and sanity checks, we saw the data was incorrect (states and airport codes were mismatched) and incomplete (not all states were represented), so we did not use it.

Airport Codes by State 2024: This was our new supplemental dataset sourced from the World Population Review which we used in our analysis. It lists the number of major airports in each US State (50 total states) and its columns include: airport code, airport state, and number of airports in each state.

Data Cleaning & Sanity Check Questions

We used these questions as guidelines for our sanity checks after we looked at overall dataset statistics and information, like size, non-null values count per variable, and data types. Insights from this section were used for our assumptions, data cleaning, and main question answers.

TSA Complaint Counts

1. Why are the null values in the airport variable only?
    a. ANSWER: From the data documentation: Null/blank values appear to represent complaints not associated with any airport in particular. They are not grand totals.
2. What years did the nulls in the airport variable correspond to?
    a. ANSWER: All years 2015 - 2023 have values for each month in nan airports and 2024 only has values for January.
3. How many unique airports do we have in our dataset?
    a. ANSWER: There are 448 unique airports.
4. How many complaints are there in total per airport in our data?
    a. ANSWER: There are 7 airports with complaint counts over 1600 and 8 airports with complaint counts under 10, for example. We can't look at individual airport complaint counts since there are so many airports.

Airports

After our sanity c[1]hecks, we still had questions our dataset couldn't answer since it lacked data:

1. Why are there 59 states if there are only 50 states in the US?
    a. THEORY: It may be because non-USA airports are included, like US territories and international airports, but because of null state values, we can't confirm this.
2. Why don't the number of unique airport names and abbreviations/codes match?
    a. THEORY: Airport names may have more than one code or there is missing data.
3. Why don't the state and city listed most frequently match the airport that is listed most frequently? Most frequent state was Texas (2036 times), city was Houston (107 times), and airport was Fort Worth (911).
    a. THEORY: Since Fort Worth is also a city in Texas, logically, the most frequent city should have been Fort Worth not Houston.

[1]

Airport Codes by State 2024

1. Is there complete data for all of the states?
    a. ANSWER: Yes, all states have at least one airport and the total number of codes per state matches the “number of airports” variable per state.
2. Do the airport codes in this dataset match the airport codes in our TSA dataset?
    a. ANSWER: Yes, they do. We identified the unique airport codes before and after we merged our datasets, and all airport codes in our subset matched.

After our preliminary data studies, we made the following modifications to our final datasets:

TSA Complaint Counts

- Converted the date column (“year_month”) to datetime data type and added two columns of data (“year” and “month”) based on that date column.
- Complaint dates ranged from Jan 2015 through Jan 2024. We decided to exclude 2024 data since the data for that year was incomplete.
- Dropped "pdf_report_date" and "category" after making our assumptions listed below.
- Dropped "clean_cat_status" because it was not relevant for our analysis.
- Dropped the rows with null values from the “airport” column since the null values were only 1.67% of our total dataset: 4042 null values/241588 total values

Airport Codes by State 2024:

- Renamed the “airport_code” variable to “airport” in order to merge datasets.

Airports:

- Dropped this dataset from our analysis.

Assumptions

After our data cleaning, we made the following assumptions for the **TSA Complaint Counts** dataset before continuing our analysis:

1. Used the “cleaned_by_category.csv” file provided by the data owner for our analysis:
    a. To minimize analyst bias
    b. Per recommendation on data repo README file.
2. Out of the two date fields in the initial data “pdf_report_date” & “year_month”, we only used the “year_month” column as the date of reference.

    a. Per the data description document provided by the data owner
    b. EDS confirmed that “pdf_report_date” to be the data/document import date rather than the complaint or the incident date: only 7 unique values for the period between 2019 - 2024.
3. Decided to focus analysis on the complaint “cleaned category” rather than including both category and subcategory:
    a. EDS revealed that the subcategory information appears more granular so it may be more suitable for deep dive study into a specific complaint category
    b. The cleaned/standardized category allows us to make cleaner complaint groups
4. Some complaint categories appeared similar so decided to bucket them into three groups for generalized analysis.
    a. Property related complaints (mishandling, special handling, locks etc)
    b. Screening related complaints (screening, patdown, expedited passenger program)
    c. Others (customer service, civil rights, lost and found)
5. We will have sufficient data to answer our main question if we only focus on airports with over 1,000 complaints. This reduction will help us narrow our scope to get meaningful insights.
    a. There are 69 airports with complaint counts over 1000 and 89,193 rows

Data Stories: Answering Exploratory Questions

Main Question:

What are the various types of TSA Complaints in US Airports over the years and what complaints are the most prevalent? (See conclusions for the response)

Supporting Questions:

We will answer this overall question by analyzing the complaint count and types of complaints in our data. The supporting questions that guide our analysis are:

Question 1: What are the most common complaints topics reported and do we see a growth pattern in these complaints over the years?

Subquestion 1.1: What are the most common complaints?

The most common types of complaints are 'mishandling of passenger property' with over 30,000 complaints. Note that the complaints listed are buckets with multiple types of subjects in each. For example, the complaint of "Mishandling of Passenger Property' can consist of loss, damage, stolen, misrouting, tampering, delayed, improper handling and / or incorrect tagging of an individual's property. This can result in an individual's financial loss, inconvenience or distress for travelers. (see figure 1 below)

Figure 1

clean_cat	count
Mishandling of Passenger Property	30947
Screening	28817
Customer Service	28094
Property - Special Handling	21847
Patdown	21155
Civil Rights	18389
Persons w/ Disabilities (PWD)	18288
Locks	16564
Expedited Passenger Screening Program	14227
Lost and Found	7285

Figure 2

Subquestion 1.2: How did the complaints grow over the years?

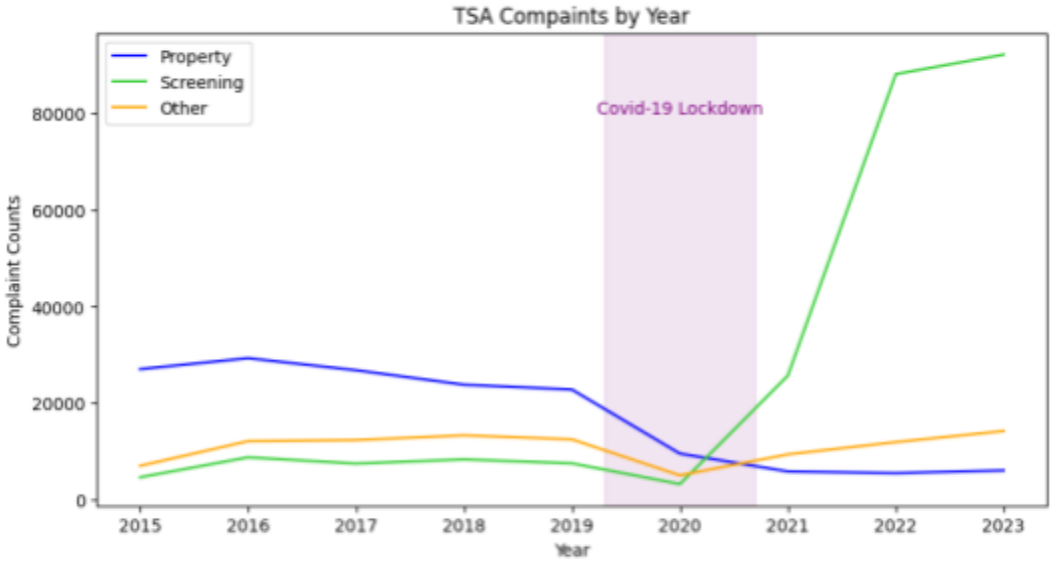
From 2015 to 2019, the number of complaints remained consistently between 40,000 and 60,000. However, there was a significant decrease in complaints from 2019 to 2020, followed by a gradual increase from 2020 to 2021. This deviation can be attributed to the onset of the pandemic and the implementation of stay-at-home orders. During this period, travel was heavily restricted, and advisories urged against it. As restrictions began to ease, people resumed traveling. Despite the relaxation of stay-at-home orders, travel regulations became stricter, including mandates for masks, social distancing

of six feet, separate seating arrangements, and more. Consequently, the uptick in complaints during this time aligns with the imposition of these stringent regulations. (see figure 2 above)

Subquestion 1.3: How have the top complaints grown over the years?

The primary grievances regarding 'property' and 'screening' remained steady from 2015 to 2020. However, there was a notable surge in 'screening' complaints in 2020. This escalation may be attributed to the resumption of travel following the pandemic. With individuals exercising greater caution, screening procedures may necessitate more thorough searches of both individuals and their belongings.

Figure 3



(Lives saved and lost in the first six month of the US COVID-19 pandemic: A retrospective cost-benefit analysis - PMC: US lockdowns were around march 2019 to july 2020)

Question 2: What are the patterns in the amount or type of complaints related to the disruption caused by COVID-19?

Subquestion 2.1: What are the most common complaints topics reported within 2020.

Refer to figure 2 and figure 3. The most common complaints reported related to the disruption caused by COVID-19 is an increase in complaints with screening. Nonetheless, complaints overall shot up as travel began due to the new rules and regulations for traveling.

Question 3: What are the patterns in the amount or type of complaints related to airport sizes or airport locations?

Subquestion 3.1: What are the top states / airports with the most complaints? Investigate any correlation between the states with most complaints and airports with most complaints may indicate geographical link.

The states with the most complaints are Florida, California, Texas, New York and New Jersey, in that order. (Refer to figure 4). Looking at the chart below, it is apparent that the number of major airports in each state has some correlation to the complaint count for the states.

Figure 4 State and Airports With the Most Complaint Counts

By State

	state	num_airports	num_complaints
0	Florida	13	67070
1	California	11	57216
2	Texas	6	42174
3	New York	5	34660
4	New Jersey	3	23945
5	Georgia	3	23198
6	Illinois	4	22877
7	Colorado	5	18753
8	Nevada	2	17516
9	District of Columbia	2	16816

By Airport

state	airport	
California	LAX	24455
New York	JFK	24327
New Jersey	EWR	23945
Georgia	ATL	23198
Florida	MCO	22001
Colorado	DEN	18753
Illinois	ORD	17944
Nevada	LAS	15744
Texas	DFW	15705
Florida	MIA	14707

The airports with the most complaints are LAX, JFK, EWR, ATL and MCO. It is notable that these airports are all well traveled airports in major US cities.

Subquestion 3.2: What are the states / airports with the least complaints?

States with the least complaint counts are Kenturky, Alaska, Nebraska, Wisconsin and New Mexico, in ascending order. These states are not known to be the popular travel destinations.

The airports with the least complaints are MYR, KOA, BUR, SDF & GEG, also in ascending order. Though some of these airports are located in popular destinations, these appear to be the smaller airports in the area so may serve as the secondary airports to another airport. (Refer to Figure 5)

Figure 5 State and Airports with the Least Complaint Counts

By State

	state	num_airports	num_complaints
0	Kentucky	2	1553
1	Alaska	3	1961
2	Nebraska	2	2017
3	Wisconsin	3	2045
4	New Mexico	2	2439
5	Connecticut	2	2555
6	Indiana	4	2654
7	South Carolina	4	3031
8	Louisiana	3	5420
9	Oregon	4	5611

By airport

state	airport	
South Carolina	MYR	1036
Hawaii	KOA	1371
California	BUR	1484
Kentucky	SDF	1553
Washington	GEG	1689
Nevada	RNO	1772
Tennessee	MEM	1791
Virginia	ORF	1821
Alaska	ANC	1961
South Carolina	CHS	1995

Conclusions

We obtained interesting key findings from our analysis of US airports with over 1000 complaints which allowed us to answer our main question: What are the various types of TSA Complaints in US Airports over the years and what complaints are the most prevalent?

One very notable finding is how the amount of TSA complaints increased dramatically after the Covid-19 lockdown. We also found that the primary driver of that increase was the complaints relating to the TSA screening. As seen on figure 3, the other two complaint groups (property and other) either decreased or stayed the same post Covid.

The graph “Growth of Complaints Over the Year” (Refer to Figure 2 above) showed that there was a huge decline from 2019 to 2020 due to stay at home orders caused by the pandemic. Travel was limited and traveling was considered a huge risk during that time period. As stay at home orders start to loosen in 2020 and onward, complaints started to increase. From 2020 to 2023 there has been a constant growth which can be the cause of regulated rules from the pandemic. Travelers were more cautious during that time period and may have had concerns on close spaces and screening procedures may necessitate more thorough searches of both individuals and their belongings.

Our analysis of TSA Complaints data and airports by state helped us understand the complaint trends over the years. We had predictions to see a huge decline during the pandemic which is shown in our

analysis validation that our data and current events go hand in hand. Although we expected some of our outcome, it was interesting to see the types of complaints that were the most common.

After our exploratory data analysis and discussions, we found there are many topics that future studies could expand on. For example, it would be interesting to explore possible causality or dependent relationship between the Covid-19 lockdown and the drastic increase of screening related complaints while the property related complaints have decreased.

Other Notable findings

- The complaints grouped by state data were generally in line with what we expected: states with less airports had less complaints and states with more airports had more complaints. (Refer to figure 4)
- The number of complaints by state can vary dramatically depending on the number of airports in the state and whether the state is considered a popular destination - likely related to the traveler count.
- Busy airports are not necessarily the airports with the most complaints. (Refer to figure 7)

Figure 6

Airports with Most Complaints
(mean val)

airport	num_complaints
LAX	16.912172
JFK	16.492881
EWR	16.233898
ATL	16.132128

Busiest Airports
(mean val)

airport	num_complaints
ATL	16.132128
ORD	13.069192
DEN	13.022917
DFW	10.936630