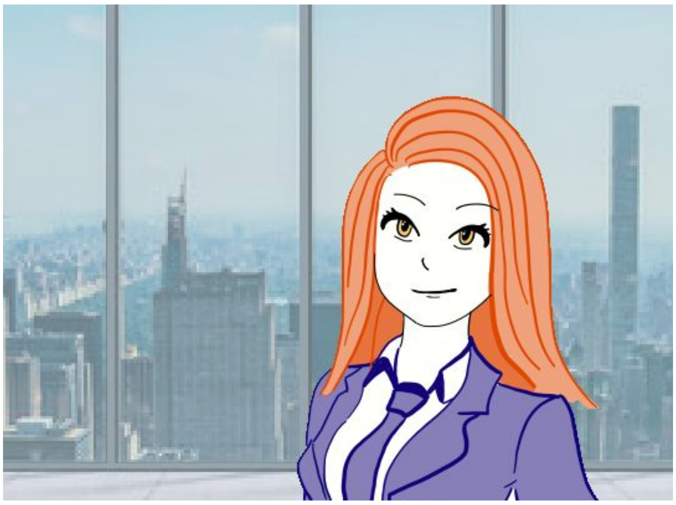
Math & Mystery Chronicles

2016 US Election Analysis • Bayesian Hospital Insights • Flying Through Borealia

The Enigmatic Anomaly

Where Clues Lead to Questions

Meet Jenny, the brilliant lead investigator on a quest to unravel the unexplained and peculiar arrival and departure delays that unfolded at US airports in 2019. Today, her relentless pursuit stands on the brink of a chilling revelation. Brace yourself for an electrifying journey as they navigate the labyrinth of this unsolved mystery.



An Investigation into Mysterious Flight Patterns by Jenny Lewis

Unravel the method behind her madness, but be warned – once you're in, there's no turning back.

Testimonies Reveal the Unusual

By Jenny Lewis

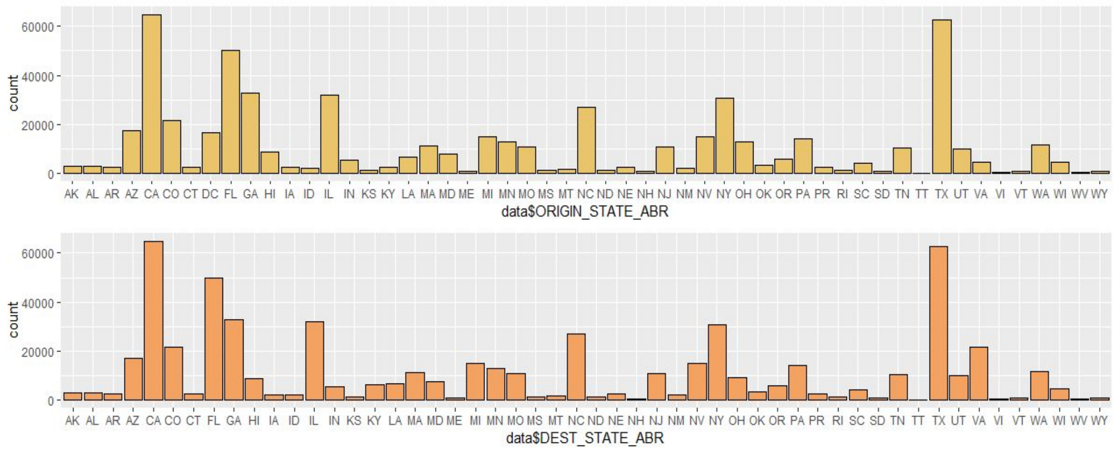
As I was handed a large amount of data on flights in the US from January of 2019, something initially seemed off to me. I began to notice some flights that had very long delays, unlike the others at the same airport. For many years, I was stumped. I could not figure out what was happening that might cause such irregular flights. I knew I had to dig deeper and get to the bottom of this sticky situation.

I started by doing a general exploration of the data by visualizing the flights to see what the patterns were (if any) and if I could find an explanation. Once I found some potential exceptions to those patterns, I wanted to take a closer look at a random subset of flights and investigate further into the causes. I talked to many witnesses and interrogated many suspects as I was putting all of the pieces of this puzzle together. Read along further if you’re curious to see how I embarked on this adventure into the world of flying in the United States and proved that not every flight will go as expected.

Witness 1 Speaks

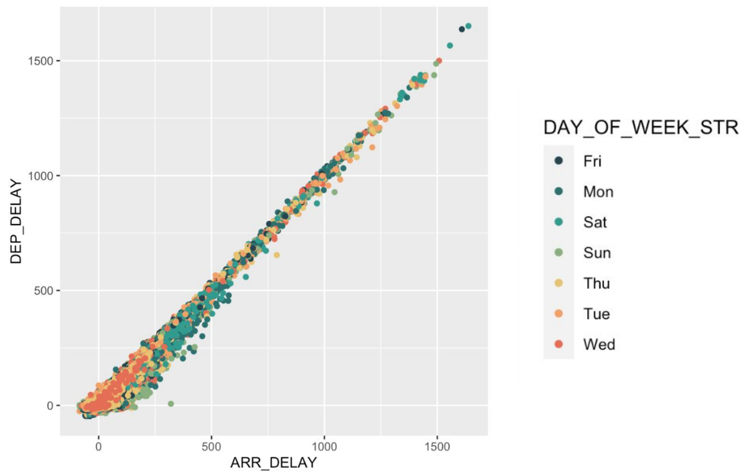
Our First Witness Recalls Their Initial Encounter With the Data

"I knew something was wrong with the data"



Intriguingly, the witness disclosed that the most popular origin and destination states were California, Texas, Florida, Illinois, Georgia, New York, and North Carolina. These states seemed to be the primary hubs of air travel, attracting a significant amount of flights. On the other hand, the witness also revealed the least popular origin and destination states. Surprisingly, New Hampshire, Trust Territories, Virgin Islands, Vermont, West Virginia, Wyoming, and Maine appeared to have fewer connections, making them less frequented by flights. The first witness proved to be a valuable source of information, shedding light on crucial insights about the US airports. They introduced two important terms: ORIGIN_STATE_ABR, representing the two-letter State abbreviations for the origin airports, and DEST_STATE_ABR, denoting the two-letter State abbreviations for the destination airports.

The ARR_DELAY and DEP_DELAY data revealed delays or advancements in either the arrival or departure flights, respectively. Where negative values indicate delays and positive values indicate early arrivals. The witness also shared the DAY_OF_WEEK_STR, a shortened representation of the days of the week when the flights occurred.

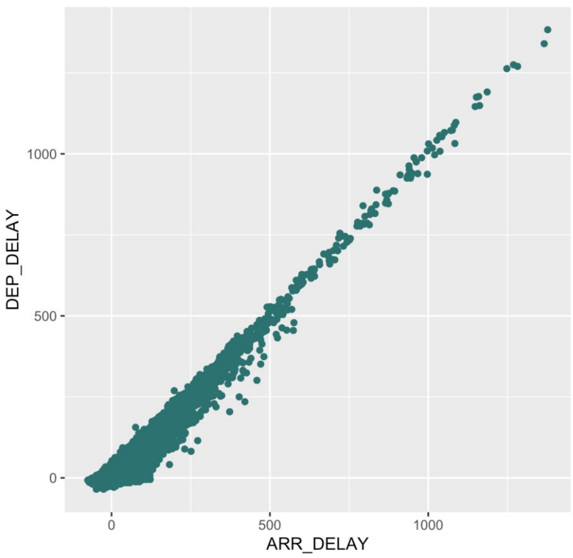


Here's where things got interesting - anomalous activities seemed to be concentrated mostly on Sundays, Mondays, and Thursdays. These particular days appeared to have a higher incidence of delays or advancements in both arrivals and departures.

Witness 2 Exposes

Anomaly Analysis Leads to Justice

The revelations from the second witness added a new layer of intrigue to the investigation, focusing on the arrival and departure delays on specific days of the week - Monday, Sunday, and Thursday, which had been the three days previously identified as the most anomalous by witness 1. While all three days exhibited similar cluster patterns, it was their anomalies that set them apart. In general, the arrival and departure times were typically delayed by the same amount.

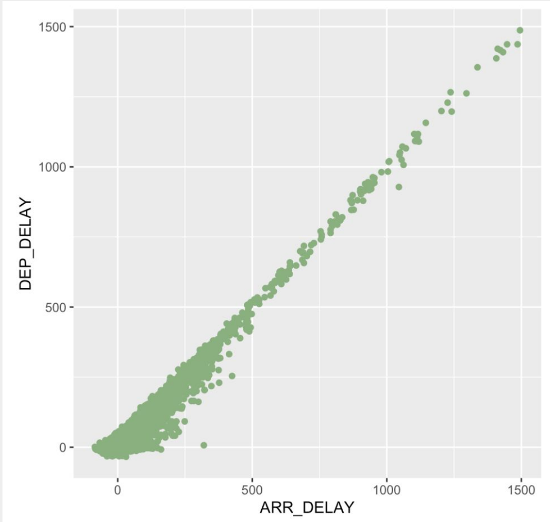


Delays of flights when travelling on Mondays.

I was informed by the witness to start with Mondays, as they believed there might be something peculiar happening. At a first glance, I noticed there were several data points that fell slightly below and out of the general pattern of the delays. In particular, there were quite a few flights with arrival delays between 200 and 600 minutes that didn't have quite as high of a departure delay as the rest of the flights. I took note of this strange behaviour that might be able to help me piece this story together further.

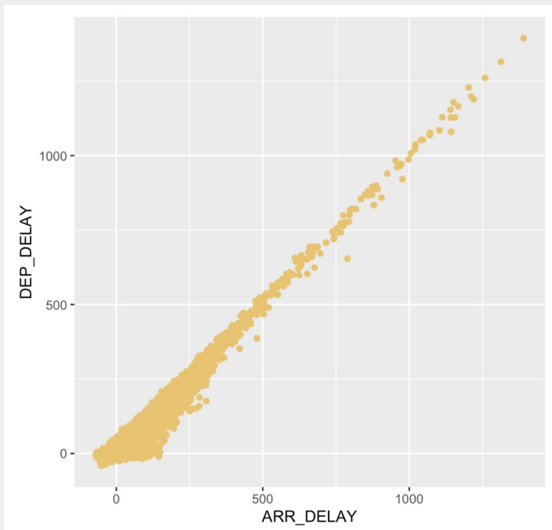
What's Happening on Sundays?

As we continued the investigation with witness 2 we were lead in the direction of the flights on Sundays, as it was one of the most frequent travel days, and there may be some potential outliers hiding there.



Delays of flights when travelling on Sundays

As the witness continued to provide more information, there was a slight indication that something abnormal might be happening on Mondays. As I looked at its delays, I noticed there were a few Monday flights that weren't quite following the pattern of the others, noting them down to potentially look into.

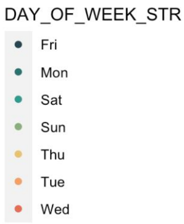


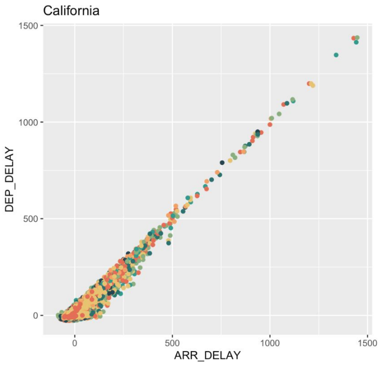
Delays of flights when travelling on Thursdays

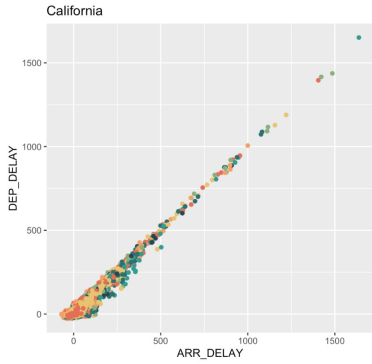
Witness 3 Breaks Silence

With our next witness, they wanted to bring light to some possible abnormalities that may have arisen from the states that have the most travellers passing through. Using ARR_DELAY, DEP_DELAY, DEST_STATE_ABR, and ORIGIN_STATE_ABR, the witness helped identify a few states that had many travellers with some flights looking strange.

Legend:



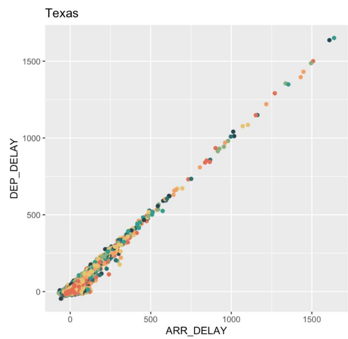




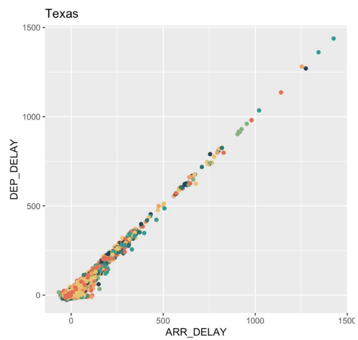
California flights by destination (left) and origin (right)

With California being the most popular state to fly into and out of, the witness recalled seeing some passengers waiting for abnormal amounts of time while travelling there.

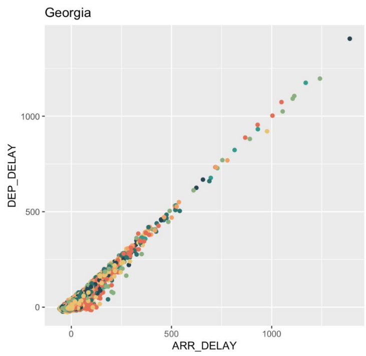
Flights destined for Texas:

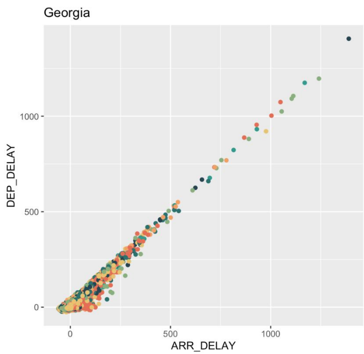


Flights originating in Texas:



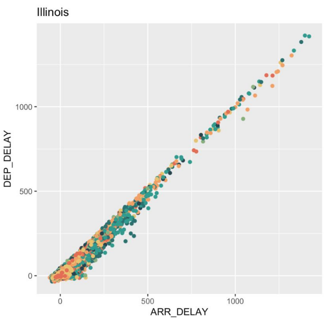
Since Texas was the second highest travelled through state, the witness identified some points that fell out of the general pattern of the delays indicating we should take note.

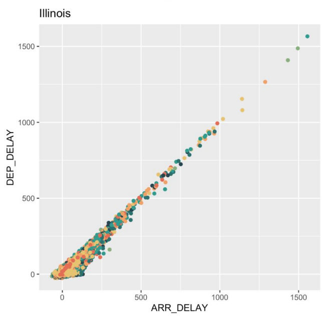




Georgia flights by destination (top) and origin (bottom)

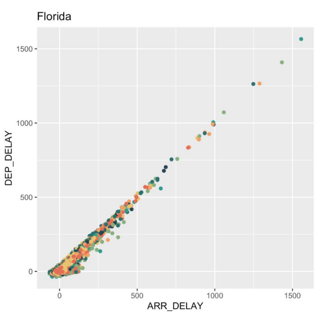
The witness had indicated that Georgia might have a few flights that had some irregular delay times, as indicated above.

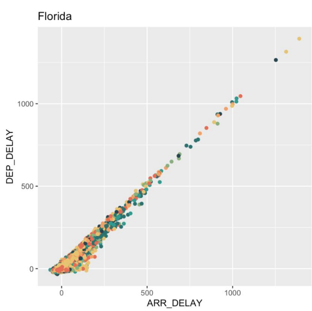




Illinois flights by destination (top) and origin (bottom)

Illinois was one of the states the witness provided insight on potential outliers that didn't follow the pattern.



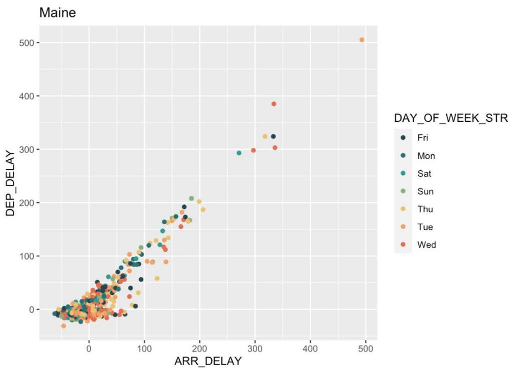


Florida flights by destination (left) and origin (right)

Although the witness provided Florida as a state with potential outliers, when I looked into the dataset I noticed most of the points for flights originating in Florida seemed to follow the general pattern of delays. That being said, when I looked into the flights destined for Florida, there were quite a few flights that fell below the others.

Witness 4’s Breakthrough

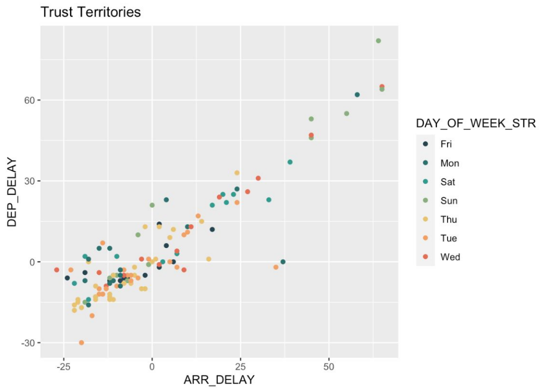
The disclosure from witness 4 led me to investigate states with the lowest number of flights. Among them were Maine, Trust Territories,

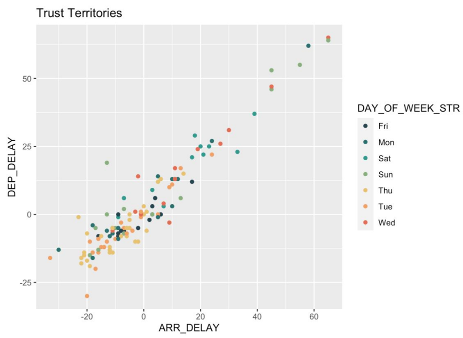


Flights originating in Maine

and New Hampshire, each exhibiting intriguing anomalies. Maine's data showed sparsity, but the general shape of arrival and departure delays remained consistent.

For Trust Territories, both originating and arrival flights displayed erratic behavior, deviating from the patterns seen before. Sparse data points and a lack of sufficient evidence made it challenging to establish





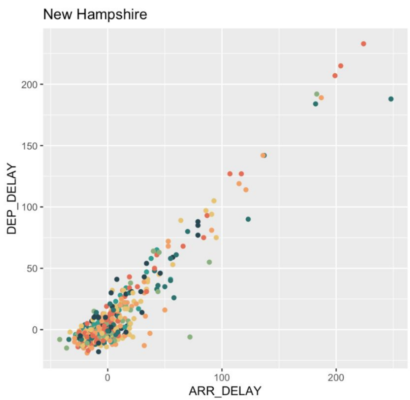
Flights originating in the Trust Territories

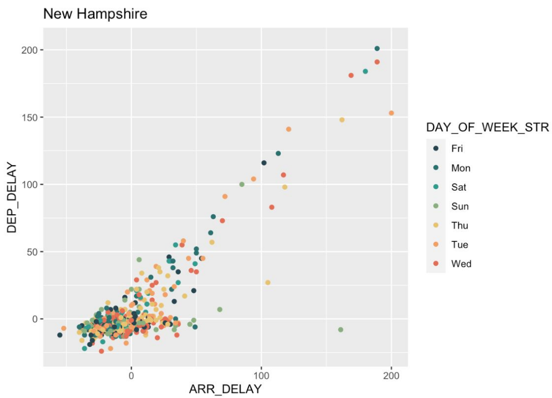
a clear relationship between arrival and departure delays. However, interestingly, Trust Territories consistently showed Thursdays with flights typically delayed by a smaller amount, observed at the bottom left of the scatter plot.

Flights destined for the Trust Territories (left)

New Hampshire's Unusual Patterns

New Hampshire, although having more data points than Trust Territories but fewer than Maine, had presented significant issues with numerous outliers and gaping disparities. Destined flights to New Hampshire had a concentration of outliers on Sunday, Tuesday, and Monday. On the other hand, origin flights from New Hampshire exhibited outliers on Sunday, Thursday, Wednesday, and Tuesday.





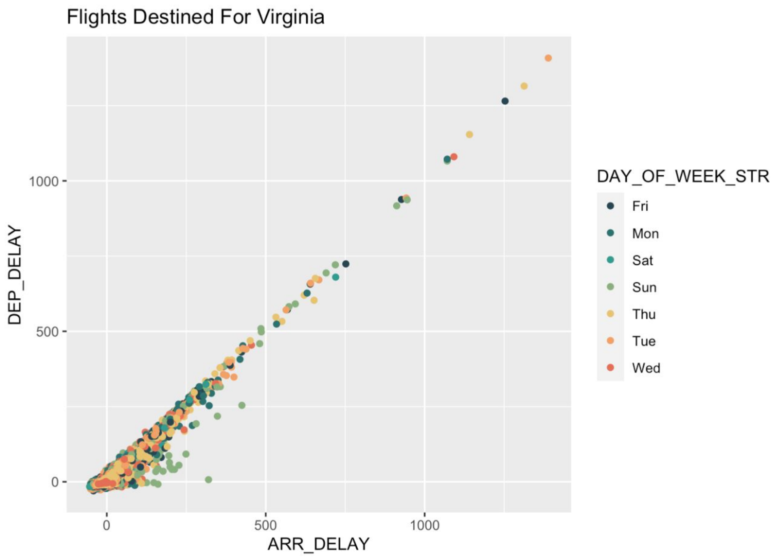
New Hampshire flights by destination (left) and origin (right)

Decoding the States, Seeking the Missing Piece

Each state's unique behavior provided valuable clues that I knew I must unravel to decipher the puzzling anomalies hidden within the flight data.

A Crucial Lead in the Investigation

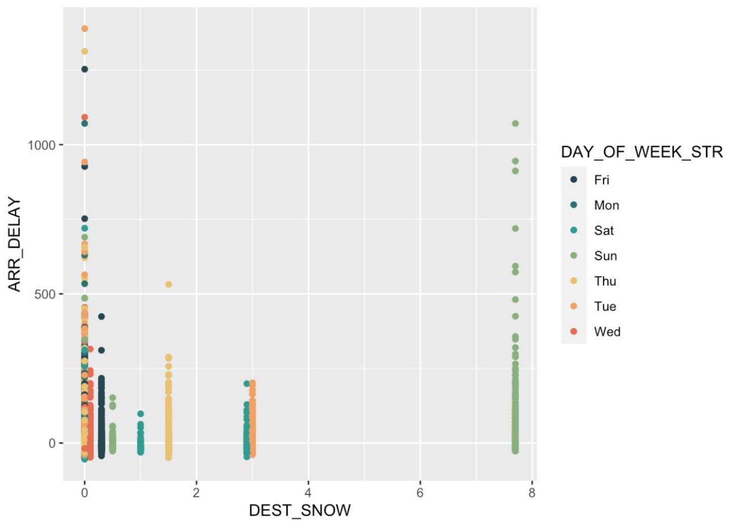
Unraveling Virginia's Flight Anomalies



After consulting the four witnesses, I set out to investigate the erratic arrival and departure delays. A vital clue emerged when I received evidence from an airport, redirecting my focus to Virginia. Discussions with Virginia airport attendees revealed that Sundays showed the most anomalies in flights destined for the state. Taking a closer look, I noticed many gaping issues possible indicating removed, lost, or unknown recorded destination flights.

An intriguing pattern emerged as I analyzed the destined flights for Virginia. Up to a certain value, both arrival and departure delays showed a positively correlated path. However, beyond the value of "1000," I noticed a significant drop in the frequency of flights, indicating that those with larger delays were a rare occurrence. This finding added a layer of fascination to the investigation, leaving me eager to explore the reasons behind this distinctive behavior.

The airport attendees also introduced the term "DEST_SNOW," representing the amount of snowfall in inches (in) that each state received. Interestingly, this behavior was confined to January, suggesting the typical assumption of weather delays. Snowfall in the region was predominantly below 1 inch. However, what caught my attention were the gaps where there seemed to be no recorded snow of 4-6 inches.



Indeed, further investigation revealed that Sundays experienced the highest snowfall. This correlation between snowfall and irregular flight delays provided a compelling lead to follow.

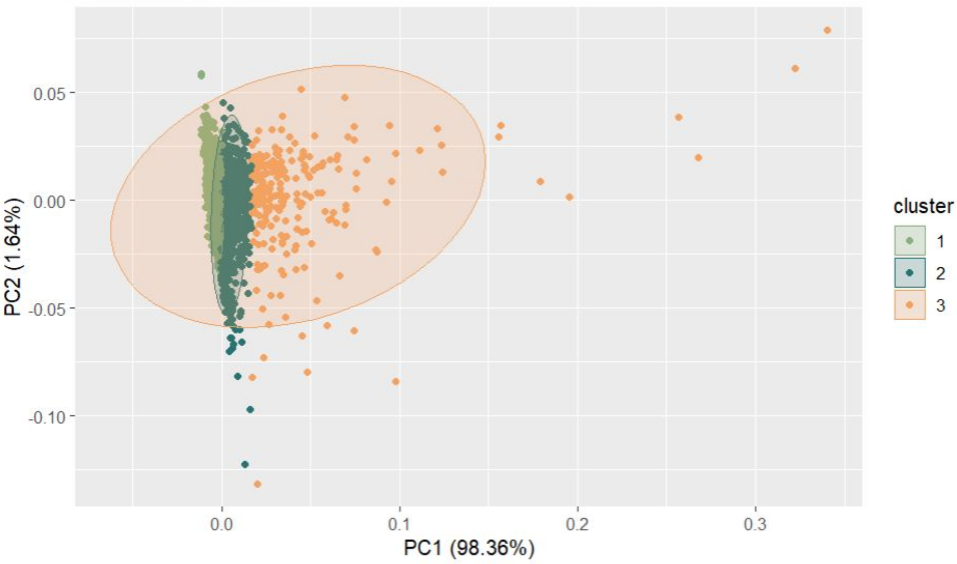
Coming Up Next: *How I used a smaller subset to pinpoint the outliers.*

PCA

The first suspect identified

Patrick C. Adams (PCA), a prime suspect in the case, has been linked to the irregular behavior observed at the US airports in 2019. The pressing question that lingers is whether Patrick C. Adams acted alone, or was there someone else involved in his stealthy activities?

Patrick exposed three distinct clusters in the flight data. Cluster 1 formed a scattered line slightly pointing left, while Cluster 2 displayed points in a straight line. Cluster 3, on the other hand, had scattered points spread across the plot. Additionally, it was revealed that PC1 (Principal Component 1)



Principal Component Analysis (PCA) clustered scatter plot

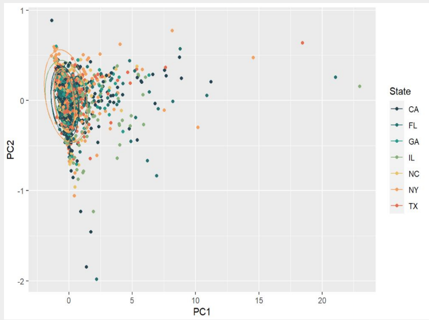
represented the overall departure delay, while PC2 (Principal Component 2) symbolized the arrival delay of flights at U.S. airports. Remarkably, PC1 alone explained 98.36% of the variation in the flight data. During the investigation, Patrick confessed to deliberately crafting certain points within the clusters identified by their X values.

Patrick Confessed to Crafting Outliers

Confessed Outliers:
93182, 477922, 259286, 467791, 42767, 479941, 312513, 381994, 451093, 259423, 270720, 352242, 382028, 46425, 93004.

Algorithm Triumph!

As I dug into the flight data, I noticed striking outliers from prominent destination states like California, Florida, Texas, and Georgia - precisely the

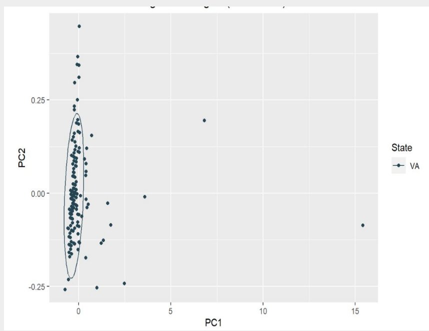


PCA - Largest Origin States

states that the third witness had mentioned were the most frequent on the list.

Chasing the Clues

My curiosity led me to focus on Virginia, one of the largest destination states. As suspected, there were indeed numerous outliers among the destination flights for Virginia.

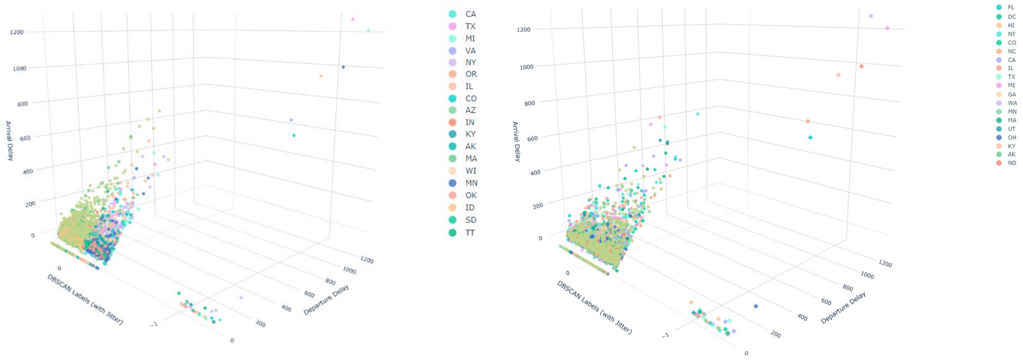


PCA - Virgina Destination

DBSCAN

The second suspects brought in for questioning

The suspects I brought in were Debs & Candie, also known as DBSCAN, a dynamic duo known for detecting outliers. DBSCAN (density based spatial clustering of applications of noise) looks at clusters within the dataset to potentially identified possible outliers. Debs and Candie provided me with the DEST_STATE_ABR (destination state abbreviation), ORIGIN_STATE_ABR (origin state abbreviation), DEP_DELAY (departure delay), and ARR_DELAY (arrival delay).

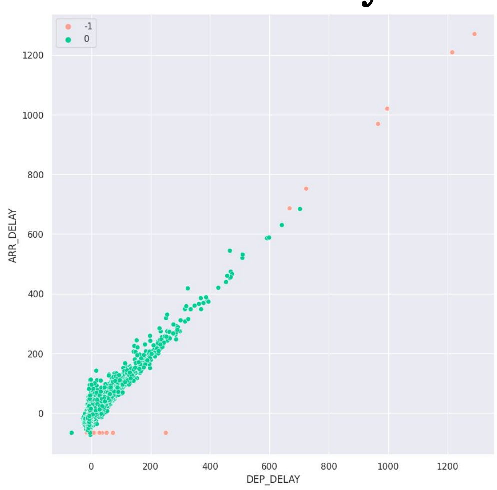


DBSCAN with flight delays by destination (left) and origin (right) states.

Where Are These Potential Outliers Coming From?

I asked Debs and Candie where all the these flights were coming from and going to and asked them to provide us with graphs that show potential outliers. If Debs or Candie provided us with a flight labelled as "-1", we had a strong indication that the flight was a potential outlier. All other flights were labelled as "0" and were not believe to have any anomalous behaviours. Many of the potential outliers strayed far from the other flights and had significantly higher delays.

Outliers Debs & Candie Confessed To and Their Delays



Confessed outliers (pink) and non-outliers (green) from Debs & Candie

After interrogating Debs and Candie, we looked at our flights and their delays. I investigated the behaviour of all of the flights in the data and separated those that Debs and Candie has confessed to from the rest. From their confession, I were able to identify the possible outliers, with many of them not following the general pattern of the flight delays.

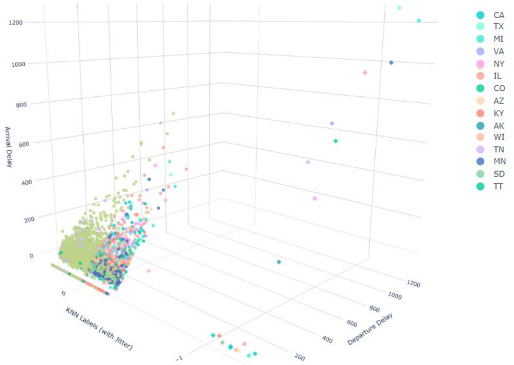
They Confessed to Many Outliers!

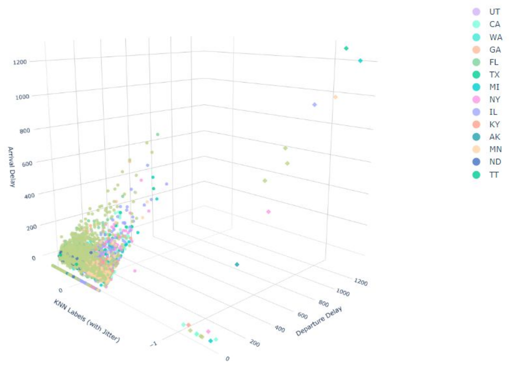
490158, 259286, 282050, 113891, 473726, 250208, 42767, 219580, 239692, 477922, 93182, 259372, 417185, 449318, 210546, 469875, 54463, 354955, 529616, 479941, 467791, 544652, 469356, 171352, 256452, 228490, 526012, 82692, 11902, 501614, 492652, 121347, 77196, 240

KNN

Our Third Suspected Individual

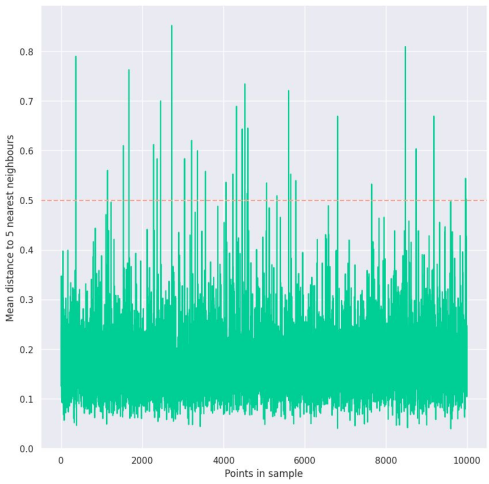
The next suspect I interviewed was Kevin Noah Newman, an expert in the K-Nearest Neighbours algorithm (KNN), who grouped flights together based on different characteristics and classifications. Kevin was the second suspect I interviewed who had a background in using clusters to analyse data and find potential outliers. The two of us used ARR_DELAY (arrival delay) and DEP_DELAY (departure delay) in our clusters.





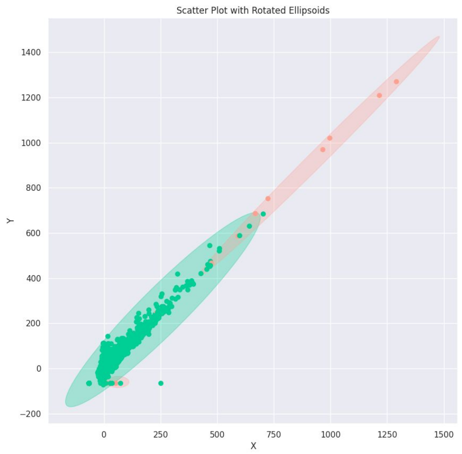
KNN by destination (top) and origin (bottom) state

Kevin produced two graphs with his initial findings, one based on DEST_STATE_ABR (destination state abbreviation) and one based on ORIGIN_STATE_ABR (origin state abbreviation), where those labelled as “-1” were found to be potential outliers, and those labelled as “0” were not.



Mean distance to 5 nearest neighbours

Kevin chose K = 5 for his clustering, and provided me with the mean distance to 5 nearest neighbours. The points above the pink line were found to be outliers, and those below were not. I found that most of the data fell below the 0.5 line, with only about 30 points rising above it. Those 30 points were classified as not following the pattern of flight delays as the rest.



Scatter plot with rotated ellipsoids

Kevin provided me with some very insight data by showing me different clusters, or neighbourhoods, of datapoints. The points that are in pink have been identified as potential outliers, whereas the points that are green were not. There were two clusters that contained potential outliers, those with high delays and those with very low delays.

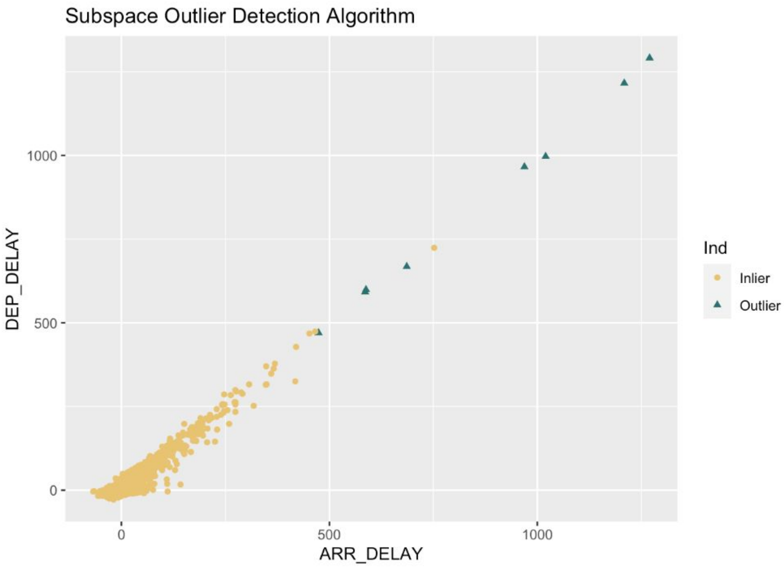
Outliers Kevin Confessed To!

259286, 282050, 113891, 473726, 42767, 239692, 477922, 93182, 259372, 417185, 449318, 381994, 337032, 354955, 529616, 479941, 467791, 268199, 544652, 171352, 256452, 228490, 266925, 262613, 82692, 502920, 501614, 121347, 77196, 454429, 92801

SOD

The Fourth and Final Suspect

The final suspect I interviewed was Sarah Oliver-Davids, also known as SOD. She was linked to many irregularities in the US flights in January of 2019. Now that the three other identified suspects have provided me with incredibly useful information, I want to ask the final suspect what she may know about these irregular flights. When talking with Sarah about them, I asked her about the DEP_DELAY (departure delay), ARR_DELAY (arrival delay), and X values that were uniquely associated to each flight in the dataset. SOD (subspace outlier detection) help use my subset to look at potential outliers in the delayed flights.

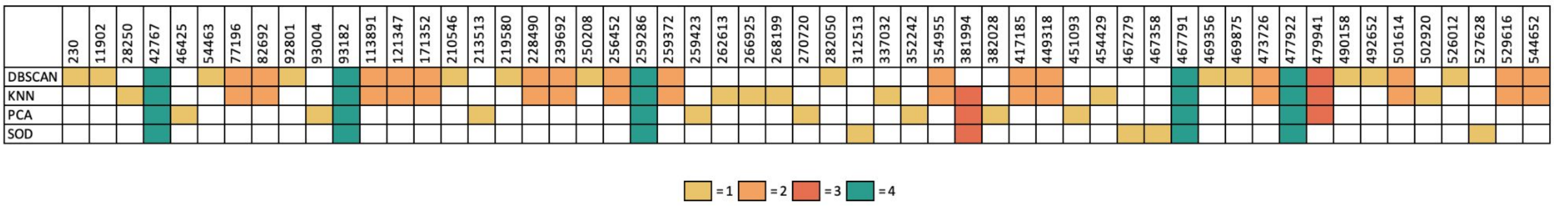


Throughout Sarah's interrogation, we were able to pinpoint 10 anomalous flights and identified them by their unique X values. Sarah confessed to the following outliers:

259286, 467279, 527628, 381994, 42767, 477922, 467791, 312513, 93182, 467358

The Verdict: Guilty

Case Closed - Joint Efforts to Individual Actions



It seems that a few outliers, specifically #42767, #93182, #259286, #467791, and #477922, stood out as instances where all four suspects collaborated, highlighted in green. These are the outliers that I can confirm since they were consistently identified by multiple accomplices. There were additional outliers that were associated with the involvement of three suspects.

For example, #381994 was identified by PCA, SOD, and KNN, while #479941 was identified by DBSCAN, KNN, and PCA. Furthermore, there were other potential outliers where two suspects collaborated, or where a suspect acted alone, which makes these outliers remain as *potential* anomalies.

M & MC