Patterns in Flight Delays and Cancellations

David Ingraham Robby Kendl Jordan Peters Julio Lopez

PROBLEM STATEMENT

(Note for Part 3 – After reviewing our proposal, motivation, and proposed work we've decided to keep everything the same except for adding a sixth question to answer to increase our scope)

Every year thousands of flights get delayed or cancelled leaving frustrated passengers pondering why they were stuck with such poor luck. The Bureau of Transportation Statistics demonstrates that on average 1-2% of flights are cancelled and 16-24% delayed each year. [1] This is an enormous number of wasted minutes and impacted people and because of that our goal for this project is to search millions of flights records and determine if there are patterns on which airlines have the worst number of occurrences and why. Our problem can be broken down into six questions below.

1. Which airlines has the most cancellations and delay?
2. Do all airlines follow similar delay and cancellation patterns throughout the year?
3. Which airlines are most susceptible to weather delay and cancellations and why?
4. Which region locations across America have the most delays and cancellations and is this airline specific?
5. Does the distance of the flight matter and do airlines treat cancellations and delays the same because of it?
6. What is the most common cancellation reason and why?

By finding answers and patterns to these questions, we will be able to confidently present which airlines are more susceptible to and handle delays and cancellations the "best" and "worst" depending on each situation. This information can then be provided to upcoming travelers in helping them determine the likelihood the airline they are traveling on will have a delay or cancellation based on certain situations.

1. LITERATURE SURVEY

In May of 2016, the self-described 'data-driven, travel guide' website, WanderBat released a ranking of the worst to best airlines based off of cancellation rate. Their data consisted of flights ranging from March of 2015 to May of 2016 and they concluded that Spirit airlines had the worst rate of cancellations. While it is unknown how exactly they calculated these percentages, what they've done differently is they combine delays and cancellations into one overall rating while we intend on splitting them up for some of our data-mining. Further, they don't provide the 'why' some airlines struggle over others and our problem surrounds not only identifying which airlines are the worst, but also where, when, and what reasons. Eventually we'd like to provide several lists similar to their ranking but with more details depending on which questions we are answering from our problem statement. [2]

A large study that was done in 2014 by Michael Seelhorst and Mark Hansen explored flight cancellation behavior and aviation system performance. Their study is similar to ours except they wanted to provide feedback to the Federal Aviation Administration in order to mitigate and predict flight delays. They break down their research into two parts, understanding the factors that lead to cancellations and queueing simulation techniques to determine the effect flight cancellations have on delay flights. Their research is thorough and quite impressive but their dataset only covers a span of 160 flight days so it's quite limited. The 71-page report is lengthy and breaks down techniques and factors FAA specific but irrelevant to our research guided toward customer satisfaction. One interesting conclusion they provided though was that American Airlines had the worst cancellation rate over all other airlines they researched (Spirit wasn't included in their dataset though). Their research heavily considers the impact of weather which is something we might reference while looking at our own patterns with weather. [3]

2. PROPOSED WORK

We already have our datasets, which will be described later, so using the collected the data we need to begin searching for patterns. The first step we will do is clean up our data. We've noticed that our dataset contains quite a bit of empty attribute values so we need to reduce the amount of empty space. The tool Data Cleaner should help us clean up these missing values. After that we will be transforming the datasets into a SQL database. This will allow us to process the data specifically for each question we have (time,

weather, distance, etc) and experiment with different patterns. While we are putting the data into database, we will also be running python methods to gather some initial information such as the z score to help guide us moving forward. After our database is set up, we will be writing queries based on each of our problem questions which will help us reduce the attributes that don't matter for each question. Simultaneously, we will be using Qubole to assist us in finding patterns in our processed data. Finally, the patterns we find among the data will be evaluated and converted into visual graphs and tables to help answer the questions we originated with.

While I'm sure we are following a similar workflow to some of the other known works, the previous works before us had limited focus on what they were trying to solve. We are trying to find behaviors among airlines but spread among many different factors that could impact flight delay and cancellation. We'll be breaking down the same dataset multiple times and in numerous ways depending on the question and direction at the time. This is a different approach than the previous works before us as we are trying to provide multiple patterns from the same dataset. In the end, we would like to find patterns between our answers for each question if any are available.

3. DATASET

We have a few primary datasets that will be used for most the data processing. These three datasets are all from Kaggle.com with all the flight data from the USA in the year 2015. [4]

Airline database: 2 Attributes (IATA_CODE, AIRLINE)
14 Rows
Airports dataset: 7 Attributes (IATA_CODE, AIRPORT, CITY, STATE, COUNTRY, LATITUDE, LONGITUDE)
323 Rows
Flights dataset: 31 Attributes, such as (AIRLINE, DISTANCE, ARRIVAL_DELAY, CANCELLATION REASON)
5,800,000+ Rows

Once we examine the dataset from Kaggle, based off the results, if needed the Bureau of Transportation Statistics has an enormous amount of flight delay and cancellation data for all US flights from 1987-2016.
Attributes (Same 31 attributes as above, Rows 100,000,000+ for that entire date range.) [5]

4. EVALUATION METHODS

As we process the data we will explore more and more evaluation methods. For the start, we'll be looking at the most basic analytic practices. For instance, we'll be gathering a lot of counts, how many trips were cancelled, how many were delayed per airlines, etc. We'll then be looking at percentages such as the percentage of cancelled and/or delayed trips per airline, the percentages of cancelled trips per cancellation reason. After that we'll be looking at the average distance of delayed and cancelled flight and then finally using more complex evaluation methods such as the correlation coefficient. We'll need to know how related cancellations between airlines are throughout the year so this will help answer that question. As we take advantage of some of the tools we will find more data processing methods to use then what is initially proposed here.

5. TOOLS

Data Cleaner: Find the clear up the missing values in our data. Help preprocess before moving into a database. [6]
Qubole: Tools used to analyze big data. The free product version might not provide exactly what we need but we'll give it a go first. **[7]**
Weka: Might help us manipulate and display data later once we have broken down datasets. [8]
SQL Management Studio: Will host our database and allow us to do the majority or our data querying and simplification
Python: Used to find patterns and process data **Excel Spreadsheet:** Table and graphical data display
Slack: Online communication.

6. MILESTONES

The Milestones for our project have been broken down into the sections of our workflow. We don't have exact dates for each section but this breakdown will allow us to keep track of what work remains.
Setup:
1. All members have the downloaded datasets
2. All members have downloaded the appropriate tools to clean and process the data

Data Cleaning:
1. Run the datasets through the Data Cleaner.
2. Correct any mixed values in SQL

Pre-Processing:
1. Create a SQL database and tables
2. Move the data into the database

3. Write queries that reduce the number of attributes per questions. (Ex: Have a query that depicts only attributes relevant to distance and cancellation and delay times)

Process:

1. Write python methods that will process the data and depict patterns

Evaluate and Depict the Data:

1. Use Excel to create tables and graphs as necessary.
2. Use WEKA to manipulate and display processed data as necessary.

What we've achieved so far:

At this point we've completed the setup, data cleaning, and pre-processing steps in our milestone. Our setup milestone step was easy to complete as they only required the distribution of the dataset and tools needed to work on the project. Within the first week of working on the project every member had downloaded the appropriated 2015 flight dataset from Kaggle. After that we verified that everyone could download Data Cleaner, Quora, and WEKA. Only one member in our group had easy access to SQL Management Studio so instead of going through the pain in setting everyone up with their own local version, we only had one setup of SQL between the team to save time.

The Data Cleaning is still an ongoing process but we've touched on both steps in that section as start. As it turns out the tool Data Cleaner is a lot less useful to use then we had originally hoped. We spent quite a bit of time struggling to use the tool the way we had envisioned and instead did a lot of the cleaning in SQL instead. With SQL it's easy to run a query to check if a column has NULL values. Once those values are detected we can either fill them in with a 'dummy' value or remove the row completely. One problem we ran into earlier on that was unexpected was a difference in airport codes. For whatever reason in our flight table, for one month of the year, the outgoing and destination airport codes were changed from a three-digit letter code to a five-digit number code. This mixed data was messing up some of our processing so we had to create two additional tables with all the letter and number codes and then join them onto the flight table, changing all number codes to their equivalent letter code. We didn't anticipate this cleaning step but it was good practice in making sure all the data was consistent throughout.

The pre-processing step focused on filtering the data into smaller subsets of the five million row dataset. As of right now we have five tables (flight, airlines, cancellation_reasons, letter_airport_codes, number_airport_codes) that are all contained in one database. We've begun writing simple queries to generate basic subsets of data and answer some simple numerical questions. This allows us to answer the 'what' in some of our questions but it doesn't answer the 'why'. A few of these results are logged in the next section.

We've also briefly begun writing python scripts to help us process the data but no results have been generated from them.

What we need to complete:

With SQL, as we continue working our current focus is to complete the list of 'To Do' items in the results section. A large amount of this work will require us to take some of the basic filtering we've already done and convert the values into percentages of the whole (Ex: We have the number of cancelled flights per month but not how large of a percentage this is). By converting our values to percentages, we will be able to make a greater sense of the impact of these cancellations and delays and better provide results to the reader. On top of the SQL work we also need to get some of our python code to successfully work and run methods in python to handle some evaluation methods we wanted to use has proven to be a bit more challenging than we anticipated and a pain point for the team. While we've had success using SQL, we'd like to continue writing in python to help us analyze the data where SQL can't.

After we are happy with the results in our data we want to use tools like WEKA and Excel to help convert our patterns and results into visual and meaningful depictions. Some of these depictions will be simple line graphs or tables but other visual representations will be a bit harder. For instance, our group wants to cluster cancellation and delay data by location for one of our questions. Once we do this we want to graph our numeric values onto a map to visually show the 'likelihood' of a location in having a delay or a cancellation over another.

Even though we've broken our project proposal apart into six questions we wanted to try to get quantitative data for all questions before visually depicting any of the data. This is why our results section lacks any graphs or meaningful answers to our questions at this time but includes some basic subsets and filtering for the heavier processing to come.

7. RESULTS

As stated above there aren't any visual representation of our data just yet. However, we do have some basic pre-processed data that might be interesting as a start. Here is some of the information we've gathered in SQL to start answering some of our questions. The results are put in bullet points under each question they pertain to.

1. Which airlines has the most cancellations and delay?

- Our dataset has 5,819,079 trips total.
- 89,884 (1.5%) cancelled trips in 2015 out of a total 5,819,079 trips.
- The Bureau of Transportation Statistics predicted 1-2% of trips get cancelled each year, so this values matches up.
- Top five most frequent airline cancellations are

16,043 - Southwest Airlines Co.
15,231 - Atlantic Southeast Airlines
15,025 - American Eagle Airlines
10,919 - American Airlines
99,60- Skywest Airline

- 2,125,618 (36.5%) trips had a Departure Delay > 0 min.
- 1,018,558 (17.5%) trips had a Departure Delay > 15min.

Using trips with a Departure Delay > 15min, top five most frequent airline delays are

254,138 - Southwest Airlines Co.
119,456 - American Airlines Inc.
118,136 - Delta Air Lines
116,153 - United Air Lines
94,024 - Atlantic Southeast Airlines

2. Do all airlines follow similar delay and cancellation patterns throughout the year?

One airline example:
For SouthWest Airlines

Cancellation #	MONTH
1767	1
3454	2
2148	3
621	4
976	5
1884	6
839	7
1056	8
384	9
322	10
970	11
1622	12

Delayed #	MONTH
19935	1
17950	2
21574	3
19468	4
23443	5
29558	6
30562	7
22296	8
12586	9
13665	10
17550	11
25551	12

3. Which airlines are most susceptible to weather delay and cancellations and why?

Top five airlines that cancel due to weather
9,164 – American Eagle Airlines
8,843 – Southwest Airlines
7,306 – American Airlines
5,539 – SkyWest Airlines
5,082- Atlantic Southeast Airlines

4. Which region locations across America have the most delays and cancellations and is this airline specific?

- Top five most frequent airport cancellation locations are

8,598 - Chicago O'Hare International: Chicago, IL
6,667 - Dallas/Fort Worth International: Dallas, TX
4,656 – LaGuardia Airport: New York, NY
3,204 – Newark Liberty International: Newark, NJ
2,703 - Gen. Edward Lawrence Logan International: Boston, MA

- Using trips with a Departure Delay > 15min, top five most frequent airport delay locations are

69,138 - Chicago O'Hare International: Chicago, IL
62,696 – Hartsfield-Jackson Atlanta International: Atlanta GA
52,629 – Dallas/Fort Worth International: Dallas, TX
45,256 Denver International: Denver, CO
41,480 Los Angeles International: Los Angeles, CA

5. Does the distance of the flight matter and do airlines treat cancellations and delays the same because of it?

Number of trips cancelled in terms of flight distance

Distance(Miles)	# Cancellations
500 > D > 0	42,632
1000 > D > 500	30,085
1500 > D > 1000	10,736
2000 > D > 1500	3,269

D > 2000	2,951

Number of trips delayed in terms of flight distance

Distance(Miles)	# Delays
500 > D > 0	346,066
1000 > D > 500	361,157
1500 > D > 1000	165,407
2000 > D > 1500	77,474
D > 2000	66,226

6. What is the most common cancellation reason and why?

- Only four cancellation reasons logged.

48,851 (54.4%) -Carrier
25,262 (28.4%) - Weather
15,749 (17.4%) – National Air System
22 (0.02%) -Security

To do with this SQL data:

- Find the percentage of all airline delay/cancellation values in each question. Graph the percentages if necessary.
- Find a correlation coefficient values between airlines that cancel and delay flights
- Sort the percentage of cancelled/delayed flights into groups and graph them on a map (Does this relate to airline specific? May remove the second half of this in question four)
- In some cases I'm only displaying data for one airline, calculate data for all airlines and compare.
- Research what 'Carrier' means and why this accounts for half of the cancellations caused across the states

As reiterated before we don't have any concrete patterns or results but we have basic filters and values to get us started. These results depict the subsets of data we are looking at with each question in the hopes of getting a better answer as we continue to work.

(Sorry for the weird formatting behavior at the end. Unsure how to fix it ☹)

8. REFERENCES

[1] https://www.transtats.bts.gov/HomeDrillChart.asp
[2] http://airlines.wanderbat.com/stories/13316/airlines-with-the-mostdelayed-flights
[3 http://www.nextor.org/pubs/NEXTOR-II-Flight-Cancellation] 2014.pdf.
[4] **https://www.kaggle.com/usdot/flight-delays**
[5] https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID-236&DB_Short_Name-On-Time
[6] https://datacleaner.org/
[7] https://www.qubole.com/
[8] http://www.cs.waikato.ac.nz/ml/weka/