Patterns in Flight Delays and Cancellations

David Ingraham Robby Kendl Jordan Peters Julio Lopez

PROBLEM STATEMENT

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 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 five 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?

By finding answers and patterns to these questions, we will be able to confidently present which airlines 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 looks into 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 groups, 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 many 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.

3. DATASET

We have a few primary datasets that will be used for the majority of 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)
5,000,000+ Rows

Once we examine the dataset from Kaggle, based off of 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, though, 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 we can think of at the moment.

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 on 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.
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.

7. SUMMARY OF PEER REVIEW SESSION

We went into the peer review session with a broad idea that we would explore flight cancellation and delay reasons. After presenting the biggest feedback we got was that we didn't exactly have any specific questions to answer. Taking this feedback, it was decided that the focus would be on the airlines themselves and so we formed specific questions to help break this focus down even further. The feedback also helped us specialize our work as another group is also researching flight delay and cancellation reasons. By focusing specifically on the airline comparisons, we can look and present new information the other group won' have.

8. REFERENCES

[1] https://www.transtats.bts.gov/HomeDrillChart.asp
[2] http://airlines.wanderbat.com/stories/13316/airlines-with-the-most-delayed-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/