Patterns in Flight Delays and Cancellations

David Ingraham   Robby Kendl   Jordan Peters   Julio Lopez

1. ABSTRACT

Our research surrounded finding answers for the six following questions.

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?
5. Does the distance of the flight matter for how likely a flight is to be cancelled or delayed?
6. What is the most common cancellation reason and why?

Summary of Results
Q1.
- American Eagle Airlines has the largest percentage of cancelled flights.
- Spirit Air Lines has the largest percentage of delayed flights.
- Both American Eagle and Spirit have the two worst combined percentage of cancelled and delayed flights.
Q2.
- Cancellation and Delay behavior changes from airline to airline over the year.
- The majority of airlines largest percentage of flight cancellations happened in February.
- The majority of airlines lowest percentage of flight cancellations happened between September and October.
- The majority of airlines largest percentage of flight delays happened between June – September
- The majority of airlines lowest percentage of flight delays happened in April
Q3.
-Delta Air Lines and American Airlines were the most suseptible airlines to cancel due to weather.
- Virgin America and Hawaiian Airlines were extreme outliers and rarely cancelled their flights due to weather
Q4.
- More cancellations happen out East then anywhere else in the United States
-Delays are fairly consistent all throughout the United States
Q5.
- The shorter a flight is, the more likely it is for a flight to be cancelled.
- The longer a flight, the more likely the trip is to be delayed.
Q6.
- Weather is the most frequent cancellation reason in our dataset and accounted for 54.4% of all cancelled flights.
- The months between January and March experience the largest spike in flight cancellations across all cancellation reason.

2. INTRODUCTION

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 because of that our goal for this project was to search millions of flights records and determine if there are patterns to why airlines might be cancelled or delayed under different conditions.

Our first question above looks at a basic comparison between airlines, determining which airlines have the worst cancellation and delay rates. The second question breaks down five airlines and compares their cancellation and delay amounts throughout the year. Questions three, four, and five analyze the cancellations rates in relation to weather, location, and distance while attempting to relate them back to specific airlines. Finally, question six explores the most common flight cancellation reason and how it might relate to the other results. These results are then used to together to help classify when a flight is most and least likely to be cancelled or delayed. This in turn, will help reduce the number of frustrated passengers in the future.

3. RELATED WORK

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 both splitting them up and combining them in 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. [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 like ours except they wanted to provide feedback to the Federal Aviation Administration 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 also compares flight cancellation behaviors among airlines which is something we reference later when analyzing our own results. [3]

4. DATASET

We have five primary datasets that will be used for most the data processing. These five 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

Letter Airport Codes dataset: (CODE, DESCRIPTION)
6,429 Rows
Number Airport Codes dataset: 2 Attributes (CODE, DESCRIPTION)
6,414 Rows
Flights dataset: 31 Attributes, such as (AIRLINE, DISTANCE, ARRIVAL_DELAY, CANCELLATION REASON)
5,800,000+ Rows

The Airline dataset is used to reference the origin and destination airline codes in the flight dataset. The airport dataset is used to answer if location matters in flight cancellation and delays. The letter and number airport codes are used to help preprocess and clean the data as some airline codes were inconsistent in the flight dataset. Finally, the flight dataset is the main dataset that will be used to query, cluster, and process against.

5. EVALUATION METHODS

Our first step in evaluating our six questions was to import the .csv dataset files into their own SQL tables. After this, the data needed to be cleaned. The largest issue with the flight dataset was that the month of October had numerical airline codes while the other months had alphabetical airline codes. To maintain consistency in the data, the numerical airline codes were replaced with their corresponding alphabetical codes. After this, the data was cleaned on a question to question basis. Depending on the question, each query run had to take into consideration removing any null value attributes that were important to each question.

Processing the data was relatively the same for each question. Each question was evaluated by first gathering counts, such as "how many trips were cancelled for 'x'" and "how many trips are there total for 'x' when 'y'". After that, the results were placed in an Excel Sheet and converted into percentages. For example, if we wanted to find out how many trips were cancelled for each airport, simply returning the number of cancelled trips at each airport only told us which airports had more trips then others. By taking the percentage of total number of cancelled trips divided by the total number of trips at each airport, our data became more valuable and useful in comparing airports to one another.

After gathering percentages, the data was analyzed and often sorted by percentage to help find patterns among the results. In some cases, such as question two, the data was exported and run through python code to find the correlation coefficient among the data.

6. RESULTS

Each question is broken down individually to explain the findings for each one.

In general:
- Our dataset has 5,819,079 trips total.
- 89,884 (1.5%) were cancelled.
- 1,015,168 (17.44%) were delayed.
- Trips are considered cancelled if they have a cancelled attribute of 1.
- Trips are considered delayed if they have a departure delay time of greater than 15minutes.

Table 1: The total amount of flights per airlines in 2015

Airline	Total Number of Flights
Southwest Airlines	1,261,855
Delta Airlines	875,881
American Airlines	725,984
SkyWest Airlines	588,353
Atlantic Southeast	571,977
United Airlines	515,723
American Eagle Airlines	294,632
JetBlue Airways	267,048
US Airways	198,715
Alaska Airlines	172,521
Spirit Airlines	117,379
Frontier Airlines	90,836
Hawaiian Airlines	76,272
Virgin America	61,903

1. Which airlines has the most cancellations and delays?

First, table 2 and table 3 depict percentages of cancelled and delayed trips separately. Table 4 depicts the total combined percentage of trips impacted by cancellations or delays per airline.

Table 2: Percentage of Flights Cancelled per Airline

Airline	Percentage of Flights Cancelled
American Eagle Airlines	5.10
Atlantic Southeast	2.66
US Airways	2.05
Spirit Airlines	1.71
SkyWest Airlines	1.69
JetBlue Airways	1.60
American Airlines	1.50
United Airlines	1.27
Southwest Airlines	1.27
Virgin America	0.86
Frontier Airlines	0.64
Delta Airlines	0.44
Alaska Airlines	0.39
Hawaiian Airlines	0.22

Table 3: Percentage of Flights Delayed per Airline

Airline	Percentage of Flights Delayed
Spirit Airlines	26.51
United Airlines	22.75
Frontier Airlines	22.30
JetBlue Airways	21.20
Southwest Airlines	20.36
American Eagle Airlines	19.35
Virgin America	17.30
Atlantic Southeast Airlines	16.81
American Airlines	16.66
SkyWest Airlines	16.13
US Airways	14.36
Delta Airlines	13.54
Alaska Airlines	10.39
Hawaiian Airlines	6.87

Table 4: Total Percentage of Flights Cancelled and Delayed per Airline

Airline	Total Percentage of Flights Cancelled and Delayed
Spirit Airlines	28.22
American Eagle Airlines	24.45
United Airlines	24.02
Frontier Airlines	22.94
JetBlue Airways	22.80
Southwest Airlines	21.63
Atlantic Southeast Airlines	19.47
Virgin America	18.17
American Airlines	18.16
SkyWest Airlines	17.82
US Airways	16.41
Delta Airlines	13.97
Alaska Airlines	10.77
Hawaiian Airlines	7.10

-Top three airlines with the largest cancellation percentage: American Eagle Airlines, Atlantic Southeast Airlines, and US Airways.
- Top three airlines with the largest delay percentage: Spirit Airlines, United Airlines and Frontier Airlines.
- For the total percentage, it tells us that the top three flights that are most likely to experience a delay or cancellation are: Spirit Air Lines, American Eagle Airlines, and United Airlines.

Interesting enough, in the related work that was researched, the travel guide' website, WanderBat, also concluded that Spirit Airlines had the largest percentage of airline cancellations and delays. Considering their research overlapped the dates of our dataset, it is comforting knowing that our results are consistent. This comes to no surprise though after a little research. According to thepointsguy.com, the CEO of Spirit holds a "money first" mentality. Spirit Airlines is unique because of their stripped-down airline business model. A part of this model strips down on-ground operations, meaning that the amount of time allowed for each flight and resources are limited for each flight. This results in less effective and reliable flights and the slightest issue can result in a flight delay or cancellation. The company mentality doesn't mind these flight delays or cancellations which would explain why they have four percent more cancelled and delayed flights then the next airline. [5]

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

Five airlines were used to answer this questions:
- The two most like-active airlines from Table 1
 (Delta, American)
- The two least active airlines from Table 1
(Hawaiian, Virgin)
- The most cancelled and delayed airline from the previous question from Table 4
(Spirit)

Each airline had the percentage of cancellations and delayed flights grouped by month. An example of this is demonstrated in Table 5. This data was then compared between the airlines and the correlation coefficient was calculated between them to determine how like the data is to each other.

Table 5: Delta Airlines Cancellations Percentage by Month

Month	Cancellation Percentage
1	17.73
2	44.35
3	13.70
4	3.84
5	3.73
6	6.2
7	2.77
8	1.80
9	0.15
10	0.44
11	0.36
12	4.83

Table 6: Correlation Coeffiecient for Cancelled Flights

Airlines Compared	Correlation Coefficient
Delta & American (Most Active Airlines)	0.89
Hawaiian & Virgin (Least Active Airlines)	-0.28
Delta & Virgin (Most Active & Least Active)	0.94
Delta & Spirit (Most Active & Most Cancelled\Delayed)	0.27
Virgin & Spirit (Least Active & Most Cancelled\Delayed)	0.16

Table 7: Correlation Coeffiecient for Delayed Flights

Airlines Compared	Correlation Coefficient
Delta & American (Most Active Airlines)	0.43
Hawaiian & Virgin (Least Active Airlines)	-0.13
Delta & Virgin (Most Active & Least Active)	0.52
Delata & Spirit (Most Active & Most Cancelled\Delayed)	0.71
Virgin & Spirit (Least Active & Most Cancelled\Delayed)	0.48

Table 8: Correlation Coeffiecient between Cancelled and Delayed Flights

Airline	Correlation Coefficient
Delta Airlines	0.27
American Airlines	-0.31
Hawaiian Airlines	0.81
Virgin America	-0.11
Spirit Airlines	0.83

Based on the results above we can see some inconsistent behavior in that some airlines cancel trips throughout the year similar to others while other airlines have no similiaries in their cancellation and delay behaviors. This might come to no surprise as each airline has to deal with their own issues (weather, technical, etc) independent of other airlines. The study done by Michael Seelhorst and Mark Hansen found similar results when they conclude "We found large differences in the cancellation behavior across airlines". [6] Yet, despite the overall differences, looking at each of largest, smallest, and average cacnellation and delay rates per month shows some commonalities.

Table 9: Largest and Smallest Cancellation and Delay Months per Airline

Airline	LCM	LDM	SCM	SDM
Delta	2	6	9	10
American	2	7	9	2
Hawaiian	8	8	2	4
Virgin	2	12	9	9
Spirit	6	6	10	4

*LCM – Largest Cancelletation Month
*LDM – Largest Delay Month
*SCM – Smallest Cancelletation Month
*SDM – Smallest Delay Month

- 3/5 airlines had February as the largest percentage of their flights cancelled
- 4/5 airlines had September and October as the smallest percentage of their flights cancelled

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

The percentage of flight cancellations and delays due to weather are demonstrated in Tables 10 and 11 respectively.

Table 10: Percentage of Flight Cancellations due to Weather

Airline	Percentage of Weather Cancellations
Delta Airlines	77.75
American Airlines	66.91
US Airways	61.22
American Eagle Airlines	60.99
JetBlue Airways	57.62
SkyWest Airlines	55.61
Southwest Airlines	55.12
Spirit Airlines	53.29
United Ailines	50.39
Frontier Airlines	47.62
Alaska Airlines	47.38
Atlantic Southeast Airlines	33.37
Virgin America	2.25
Hawaiian Airlines	0.58

Table 11: Percentage of Flight Delays due to Weather

Airline	Percentage of Weather Delays
American Eagle Airlines	15.32
Delta Air Lines	9.22
Virgin America	9.14
American Airlines	7.66
United Air Lines	5.98
Hawaiian Airlines	5.58
US Airways	5.09
SkyWest Airlines	4.04
Alaska Airlines	3.85
Southwest Airlines	3.78
JetBlue Airways	3.65
Spirit Air Lines	2.69
Atlantic Southeast Airlines	2.63
Frontier Airlines	2.36

The most surprising results from these tables is the bottom two airlines for the flight cancellations, Virgin Airlines and Hawaiian Airlines. Both airlines are extreme outliers considering the next closest Airline is 31% away from their percentage of weather cancellations. Looking deeper into these results, it might make sense that Hawaiian Airlines suffers the least weather trouble considering the destination is always Hawaii, a place known for its favorable weather.
However Virgin airlines travels across America, everywhere from Lost Angeles to New York City. Considering that they cancel so few flights due to weather is alarming compared to Delta Airlines who also travels throughout America but cancels 77.5% of their flights due to weather. Interesting enough, Virgin America's cancelled 68% of its flights due to the National Air System (air traffic control, heavy air traffic, operations). On Virgin

America's website, they have a traveler alert page. Interesting enough the last five traveler alerts were related to relocation, construction and disease. While these alerts took place in 2016 and 2017, if the same concerns occurred in 2015, our data would reflect some of these issues in that Virgin America cancelled and delayed flights due to the large amount of operation changes that have taken place the few years. This might explain why so few cancellations are relevant to weather. [7]

4. Which region locations across America have the most delays and cancellations?

Table 12: Top Five Airports with the Largest Percentage of Cancellations

Airport: City, State	Percentage of Flight Cancellations
Ithaca Tompkins Regional Airport: Ithaca, NY	11.76
Mammoth Yosemite Airport: Mammoth Lakes, CA	10.26
Friedman Memorial Airport: Hailey, ID	8.84
Del Norte County Airport: Crescent City, CA	8.42
Devils Lake Regional Airport: Devil's Lake, ND	8.01

Table 13: Top Five Airports with the Largest Percentage of Delays

Airport: City, State	Percentage of Flight Delays
Adak Airport: Adak, AK	39.05
Gustavus Airport: Gustavus, AK	37.66
Wilmington Airport: Wilmington, DE	37.00
St. Cloud Regional Airport: ST Cloud, MN	30.12
Martha's Vineyard Airport: Martha's Vineyard, MA	30.09

- Two Californian airports are in the top five largest cancellation percentage of all their flights.
- Two Alaskan airports are the top two airports with the largest percentage of flights delayed.

Next the USA was broken down by longitudes, and the number of airports in each of those groups was compared to how many cancellations and delays occurred in each group as well. These results are show in Tables 14 -18.

The longitudes were split up in degrees like so
-180 to -110 (Hawaii, Alaska, West Coast)
-110 to -95 (Midwest)
-95 to -80 (Mideast)
-80 to -60 (East Coast)

Table 14: Flight Cancellation Amount by Longitude Groups

Longitude Group	Amount of Cancelled Flights
-180 to -110	13,192
-110 to -95	18,729
-95 to -80	35,200
-80 >-60	22,763

Table 15: Flight Delay Amount by Longitude Groups

Longitude Group	Amount of Delayed Flights
-180 to -110	245,503
-110 to -95	199,008
-95 to -80	392,739
-80 >to -60	177,917

Table 16: Number of Airports in Longitude Groups

Longitude Group	Number of Airports
-180 to -110	81
-110 to -95	66
-95 to -80	120
-80 >to -60	55

Table 17: Average Amount of Cancelled Flights per Airport

Longitude Group	Average Amount of Cancelled Flights Per Airport
-180 to -110	162.86
-110 to -95	83.77
-95 to -80	293.33
-80 >to -60	413.87

Table 18: Average Amount of Delayed Flights per Airport

Longitude Group	Average Amount of Delayed Flights Per Airport
-180 to -110	3,030.90
-110 to -95	3,015.27
-95 to -80	3,272.82
-80 >to -60	3,234.85

From the data, we can see that on average each section across the United States has the same average amount of delays per airport then other locations. However, the amount of flight cancellation per location heavily favors the East Coast. This can most likely be explained by the large airports out east. For instance, the airport that cancels the largest percentage of its trips was in Ithaca New York. Further, looking at only the number of cancellations at

each Airport, four of the top five airports with the most number of flight cancellations occurred out East (Chicago, New York City, Newark, Boston). The sheer number of cancelled trips most likely skewed the data up for the average of cancelled flights per airport out East. Still, this heavily suggests that airports out East are more likely to cancel a trip than anywhere else in the United States.

5. Does the distance of the flight matter for how likely a flight is to be cancelled or delayed?

Similar to how question four was handled, the distance of each flight was split up into groups.
The groups in miles are:
0 to 300
300 to 600
600 to 1000
1000 to 20000
2000 to 5000

Table 19: Flight Cancellation Amount by Distance

Distance Group	Flight Cancellation Amount
0 to 300	23,234
300 to 600	27,096
600 to 1000	22,594
1000 - 2000	14,009
2000 - 5000	2,951

Table 20: Flight Delay Amount by Distance

Distance Group	Flight Delay Amount
0 to 300	154,867
300 to 600	275,583
600 to 1000	278,693
1000 - 2000	243,189
2000 - 5000	66,226

Table 21: Number of Flights per Distance Groups

Distance Group	Flight Amount
0 to 300	996,263
300 to 600	1,652,725
600 to 1000	1,523,464
1000 - 2000	1,272,645
2000 - 5000	373,982

Table 22: Percentage of Flights Cancelled

Distance Group	Percentage of Cancelled Flights
0 to 300	2.33
300 to 600	1.64
600 to 1000	1.48
1000 - 2000	1.10
2000 - 5000	0.79

Table 23: Percentage of Flights Delayed

Distance Group	Percentage of Delayed Flights
0 to 300	15.54
300 to 600	16.67
600 to 1000	18.29
1000 - 2000	19.12
2000 - 5000	17.71

Based off the results in Table 22, we can see that the shorter the flight, the more likely it is for a flight to be cancelled. In contrast, Table 23 suggests that the longer the flight, the more likely the trip is to be delayed.

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

Only four cancellation reasons in the dataset.
Weather: 48,851 (54.4%)
Carrier: 25,262 (28.4%)
National Air System: 15,749 (17.4%)
Security: 22 (0.02%)

What does each one mean?
Weather: Meteorological conditions that prevent flight.
Carrier: Due to circumstances within the airline's control (fueling, baggage loading, crew problems)
National Air System: NAS preventing operation such as air traffic control, heavy air traffic, operations
Security: Evacuation of terminal because of security breach

It's clear that majority of cancelled trips are due to Weather. In comparison, the Bureau of Transportation Statistics only keep track of Extreme Weather as a cancellation reason (such as blizzards, tornados, etc) while our cancellation reason accounts for any weather. They claim that only about 4% of flights are cancelled due to weather per year. Out of all the flights, weather cancellations account for 8% of all flights so our numbers are only a bit higher than the Bureau's data but still noticeably different. This is most likely due to the cancellation weather accounting for any weather condition extreme or not. [8]

Looking at each cancellation reason more closely, patterns can be examined for each cancellation reason across the year. By month cancellation results are broken down in Tables 24, 25, and 26 for all cancellation reasons except Security (too few data values).

Table 24: Cancellation Percentages by Month for Carrier

Month	Cancellation Percentage
1	11.38
2	11.14
3	9.81
4	7.10
5	7.94
6	14.32
7	10.24
8	9.36
9	4.31
10	3.77
11	4.30
12	6.25

Table 25: Cancellation Percentages by Month for Weather

Month	Cancellation Percentage
1	14.37
2	31.62
3	14.05
4	3.66
5	5.69
6	6.81
7	1.81
8	2.68
9	1.03
10	1.20
11	4.79
12	11.49

Table 26: Cancellation Percentages by Month for National Air System

Month	Cancellation Percentage
1	13.25
2	14.31
3	10.40
4	5.94
5	5.75
6	13.82
7	8.48
8	8.74
9	3.06
10	3.33
11	7.39
12	5.53

Table 27: Total Sum of All Percentages per Month

Month	Total Cancellation Percentage
1	38.99
2	57.08
3	34.33
4	16.71
5	19.39
6	34.95
7	20.53
8	20.78
9	8.40
10	9.10
11	16.47
12	23.27

In general, adding up all the percentages of each flight cancellation reason, we can see that there's a larger amount of cancellations overall between January – March than any other time of the year. Similarly, there's a dip in flight cancellations around September and October across all cancellation flights.

7. APPLICATIONS

Wrap all these results together and they can be applied toward classifying the worst and best cancellation and delay conditions below.

Table 28: Best and Worst Cancellation Conditions

Airline	Month	Location	Distance
American Eagle	2	East	Short
Hawaiian	8-10	Midwest	Long

Table 29: Best and Worst Delay Conditions

Airline	Month	Location	Distance
Spirit	6-8	Mideast	Long
Hawaiian	2-4	Midwest	Short

The top row in each table represents the worst conditions for a flight to be cancelled or delayed. The bottom row represents the best conditions in catching a flight that won't be delayed or cancelled. Of course, there are many variables that will change for each person's flight, but using Tables 28 and 29 and every other table as a guide, can help users predict the possibility that their flight will be cancelled. This research will hopefully reduce the amount of wasted and frustrated minutes by travelers. While it's easy to keep in mind that only 1-2% of all trips will be cancelled and ~15% of trips will be delayed, it's easy to reduce the chances of falling into either of those categories with a little careful planning. Each condition,

airline, month, location, and distance, can change drastically how likely someone is to experience unexpected frustrations. By applying and considering even one of the factors can help make all future flights and travels easy and stress-free.

Disclaimer:
While we've come up with some useful results, we do understand that the results are based only off the data of one year. The trends and patterns we've discovered could potentially carry over and represent general flight cancellation and delay behavior, but the research isn't promised to be all inclusive for each year.

8. REFERENCES

[1] https://www.tanstats.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://thepointsguy.com/2016/07/spirit-airlines-responds-to-complaints-with-explanation/
[6] http://www.nextor.org/pubs/NEXTOR-II-Flight-Cancellation-2014.pdf
[7] https://www.virginamerica.com/cms/traveler-alerts
[8]https://www.rita.dot.gov/bts/help/aviation/html/understanding.html#q3