|  |
| --- |
|  |
| FLIGHTS DELAY PREDICTIONS AND AIRPORT CONNECTIONS IN FEBRUARY 2019 USING SPARK GRAPHFRAMES |
| |  |  |  | | --- | --- | --- | | **Folashikemi Maryam Asani Olaniyan** | **P2586104** | **IMAT5322 Big Data Analytics** | |

abstract

An analysis of flight delay predictions and airport connections in the United States in February 2019.

The analysis identifies the highest number of departure and arrival delays, the total number of delays, the airport with the highest departure and arrival delays, the cause of delays and connections between the airports.

CONTENTS

[Abstract 1](#_Toc41337044)

[Contents 2](#_Toc41337045)

[Data Introduction 3](#_Toc41337046)

[Analysis Flow Chart 4](#_Toc41337047)

[Methodology 5](#_Toc41337048)

[Data Pre-Processing 5](#_Toc41337049)

[Preparing The Graph For Analysis 5](#_Toc41337050)

[Data Analysis 6](#_Toc41337051)

[Departure And Arrival Delays 6](#_Toc41337052)

[Causes Of Delay 7](#_Toc41337053)

[Graph Queries 10](#_Toc41337054)

[Degrees 10](#_Toc41337055)

[Motif Finding 11](#_Toc41337056)

[Connections And Distance Analysis 12](#_Toc41337057)

[Graph Algorithms 13](#_Toc41337058)

[Case Study: Atl As The Busiest Airport 15](#_Toc41337059)

[Conclusion 16](#_Toc41337060)

[Recommendations And Further Research 16](#_Toc41337061)

[References And Bibliography 17](#_Toc41337062)

[Appendix 18](#_Toc41337063)

Data introduction

source and license

The datasets used in this analysis are derived from Openflights.org and The United States Bureau of Transportation Statistics website. Openflights.org is an open-source website derived from airport database containing the detailed description of all airports worldwide. The data from the United States Bureau of Statistics can be access from *transtats.bts.gov* containing information about flight origins, destinations, departure and arrival Delay durations, causes of delay and distance travelled and more.

The data sets are available freely and at no cost to the public. The Bureau of Transportation Statistics published this data voluntarily and updates under the US Depart of Transportation, permitting the copying, publishing, and exploiting of the information.

data format

The complete data from both websites are published to CSV files. The Open flights dataset contains several fields detailing the identity of the airport; this report we will primarily use the IATA (International Air Transport Association) code which a 3-letter code used principally to identify airports globally. In the analysis, the selected data used from the Bureau of Transportation Statistics include flight origin and destination, duration of departure and arrival delay, the distance between two airports and causes of delay.

Data BAckground

The Bureau of Transportation Statistics defines significant airline departure and arrival delay as a flight with delay greater than 15 minutes, the distance between airports in miles and cause of delay by one of the following (Bureau of Transportation Statistics, 2020):

* Carrier: Delay due to airline control issues, e.g. baggage loading, crew problems, etc.
* Weather: Significant weather conditions due to blizzard, tornados, or any extreme weather conditions.
* National Aviation System (NAS): Delays attributed to a broad set of situations, i.e. heavy runway traffic volume, airport operations and traffic control.
* Security: Flight delays due to security breach, inoperative screening equipment and evacuations in the terminals or onboard.
* Late-arriving aircraft: Previous flights with same aircraft arriving late causing present flights to depart late.

Content, scope, range and accuracy

This report investigates flight delay, causes, airport interactions and performance of Hartsfield-Jackson Atlanta International Airport (ATL) as a case study for February 2019; inclusively. The scope limits to analysing flights within the United States of America (USA) using Apache Spark Graphframes written in Python Programming Language.

Spark Graphframes

Graph representations enable new sets of algorithms and functions for evaluating the relationships between flight networks. The Apache Spark Graphframe library is an abstraction of the Spark SQL DataFrame designed to identify graph objects relationships- in this case, flight relationships (Aven, 2017). To create a GraphFrame, a vertex Dataframe should contain a unique column named *id,* and an edges Dataframe containing two unique columns named *src* and *dst* (see *Appendix A*). In this analysis, the column IATA in the data from Openflight represents *id* and the columns origin and dest from Bureau of Transportation Statistics represent *src* and *dst,* respectively. Graph Queries and Graph Algorithms will be used to answer queries about the data.

## analysis flow chart

A picture containing game

Description automatically generated

Figure . Analysis Flow Chart

# methodology

import data

The raw data is uploaded to the HDFS Hue system, and the file directory is imported into Jupyter Notebook using Apache Spark SQL DataFrame configurations. The relevant libraries such as Graphframes, Pixie dust, Pandas and Numpy were imported to the notebook.

Data Pre-ProceSsing

Recall, to create a Graphframe, the vertices data would have a unique column named *id.* To achieve this, the column IATA is renamed to *id.*

A screenshot of a cell phone

Description automatically generated

Figure . Renamed IATA to id

For the edges, the columns src and dst was added to the raw file to ensure the original columns remain in the dataset.

Preparing The Graph for analysis

The vertices and edges were combined to form the Graph DataFrame. At any point, the results returned by from the GraphFrame could easily be displayed using PixieDust library to create visualisations that help summarise the findings and make it readable. Thus, the data is left unchanged within the Graphframe and graph queries are used to produce desired results.

A close up of a map

Description automatically generated

Figure . Circles are vertices (Airport id), and Arrows lines are edges (trips)

total number of airports and trips

The total number of airports and trips in Feb 2019 are 7184 and 533175.

# Data analysis

## Departure AND arrival delays

A simple grouping of the departure and arrival delays shows the maximum departure delay as approximately 2672 minutes and arrival delays as 2649. The minimum delay time for both departure and arrival is 0 minutes which is also recorded as on-time flights. The chosen columns for departure and arrival delays only take account of delays from 0 minutes onwards.

A screenshot of a social media post

Description automatically generated

Figure . Summary of Departure and Arrival Delay

The total time of departure and arrival delays in the US is 197,424 minutes and 204,694 minutes, respectively for Feb 2019.

Flights originating from Detroit Metropolitan Wayne County Airport (DTW) to LaGuardia Airport (LGA) have the most significant flight delays.

A screenshot of a cell phone

Description automatically generated

Figure 5. Airport with the highest delay

The BTS dedicates a separate column for identifying delay factor greater than zero, which signifies delay times greater 15 minutes. The analysis shows that Yuma International Airport (YUM), Arizona has the highest departure and arrival delays.

A screenshot of a social media post

Description automatically generated

Figure . Highest Departure and Arrival Delay

## causes of delay

carrier

The analysis shows that there is a much significant number of delays caused by the airline carrier in Detroit (DTW) and LaGuardia (LGA) airport. An article from The Guardian in Feb 2019 shows the reason for carrier delays was due to staffing issues caused by the government shutdown in New York (see *Appendix A).*

A screenshot of a cell phone

Description automatically generated

Figure . Airport mostly affected by the Airline

weather

Extreme Weather conditions are known to cause minor flight delays with the majority of airlines resulting in cancellations in such cases (Yablonsky et al., 2014). This analysis shows that Flights from Los Angeles International Airport (LAX) to Eagle County Regional Airport, Colorado (EGE) had the most delays due to weather conditions. It may be due to the extreme snowfalls at that time of the year.

A screenshot of a cell phone

Description automatically generated

Figure . Airport affected by Extreme weather

national AVIATION SYSTEM (NAS)

This category of delay frequently happens in most airports with the reason for delay falling under a broad set of reasons including heavy traffic volume (many incoming and outgoing flights), airport operations and air traffic control. Flights from Cincinnati/Northern Kentucky Airport (CVG) to Newark Liberty Airport (EWR) have the most significant NAS delayed flights.

A screenshot of a cell phone

Description automatically generated

Figure . Airport Affected by NAS clearance

security

A major concern of all airport staff and passengers is the safety of aircraft during the flights. Hence, a great deal of time is spent ensuring all passenger are screened before boarding. Security delays can be caused by terminal or concourse evacuations, security breach, inoperative screening equipment or long lines in screening areas. In this analysis, it is observed that flights between Nashville International Airport (BNA) and Orlando International Airport (MCO) have the highest security delays. According to a news article from News Channel5, 90% of guns found at BNA airport were loaded which was a massive increase in the number of guns found between 2017 -2018 (see *Appendix B*).

Observe from the bar chart that all flights departing and arriving at Orlando International Airport (MCO) have security delays. Numerous news articles in Feb 2019 reported gun carry issues, and Orlando ranked in the nation’s top five-gun seizures (see *Appendix C*).

A picture containing drawing

Description automatically generated

Figure . Airport with Security delay

late arriving aircraft

Flights delays from San Luis Obispo County Regional (SBP) to Phoenix Sky Habor Airport have the highest late arrival delay due to the same late-arriving aircraft carrying the delayed passengers.

A screenshot of a cell phone

Description automatically generated

Figure . Airport with Late Arriving Flights causing the delay

GRAPH QUERIES

## Degrees

The airport with the highest number of arriving and departing flights is identified by using the Graphframe degrees method. This method returns information about the degree of each airport (vertex) is inbound and outbound to other airports(vertex). In this analysis, Hartsfield-Jackson Atlanta International Airport has the highest inbound and outbound flights with 56382 trips in total.

A screenshot of a cell phone

Description automatically generated

Figure . Degree Showing the Most Common Airport

find transfer from degree method

Using the degree method, Magic Valley Regional Airport (TWF) has no transfer flights



Figure . Airport without transfer

Yellowstone Regional Airport (COD) and ELM have the highest number of airport transfers.

A screenshot of a cell phone

Description automatically generated

Figure . Airport with the Highest transfer

## Motif Finding

A motif returns unique information about structural patterns in airport connections. The method is implemented using a domain-specific language (DSL) to find arbitrary airport ids and connect them to resultant edges. In this analysis, the motif was used to find the airports with regular connections that are not directly connected and airport with regular connections that are directly connected.

The result shows that John F. Kennedy (JFK) and Newark Liberty airport have no direct connections; this is due to the proximity of both airports- only 33 miles between the two airports.

A screenshot of a social media post

Description automatically generated

Figure . Find Flights that can connect to the airports without Direct Flights

The analysis result for airports with regular connections that are directly connected shows that flight from Honolulu International Airport (HNL) and John F. Kennedy (JFK) have direct links with other airports.

A screenshot of a cell phone

Description automatically generated

## Connections and Distance Analysis

The analysis identifies trips between LaGuardia airport Chicago O’Hare International Airport as the most popular flight routes with almost 2400 journeys on the route.

A screenshot of a cell phone

Description automatically generated

Figure . Most popular route

The trips between Honolulu International (HNL) and John F. Kennedy had the longest distance of about 5000 miles. Observe that the distance between Newark International (EWR) to Honolulu (HNL) is almost the same as trips from JFK; this is because of the proximity of JFK to EWR.

A picture containing drawing

Description automatically generated

Figure . Longest Distance

Graph Algorithms

page ranking

Page Rank is an algorithm created by Google. In this analysis, page rank measures the importance of airports according to the total number of trips by determining the weighting of all trips connected to an airport. The result shows that Hartsfield- Jackson Atlanta International Airport (ATL) has the highest weighting

A screenshot of a cell phone

Description automatically generated

Figure . What is the Busiest Airport?

triangle count

The triangle count identifies three airports that are directly connected provided there is a trip between any two of them and count the number of trips for each triangle. ATL has the highest number of trips.

A screenshot of a cell phone

Description automatically generated

Figure . Triangle to know the airport with the Highest Connections

breaDth-first search

The Breadth-first Search finds the quickest route from one airport to another by specifying the airports using Spark Dataframe expression. The distance between HNL and JFK is the longest. Therefore, the breadth-first search is used to find the quickest trip between the route. The result shows direct flights between the airports as the fastest route.

A screenshot of a cell phone

Description automatically generated

Figure . Breath-First to find the Quickest Path

CASE STUDY: ATL AS THE BUSIEST AIRport

So far, we have identified Hartsfield-Jackson Atlanta International Airport (ATL) is the busiest airport in the USA. The visualisations below show that on average, there are an approximately 140 minutes delays for flights arriving from Gulfport-Biloxi International Airport (GPT) and about 150 minutes for flights departing from ATL to Coastal Carolina Airport (EWN).

A screenshot of a cell phone

Description automatically generated

Figure . Average Arrival Delay to Atlanta (ATL)

A screenshot of a cell phone

Description automatically generated

Figure . Average Departure Delay in Atlanta (ATL)

# concLUsion

The Graphframe analysis as given immense details about the airport data, causes of delay and flight connections.

The analysis revealed the total number of arrival and departure delays and the airports affected by the categories of delay causes defined by The Bureau of Transportation Statistics. The busiest airport appears to be Hartsfield-Jackson Airport with 56,382 flights in and out of the airport.

Graph queries using Vertex degrees shows that were no transfers from Magic Valley Regional Airport. The Motif Method identified there were no trips between JFK and EWR.

Further analysis conducted using Graph Algorithms such as Page Rank, Triangle Count and Breadth-First Search identifies ATLas the airport with the highest rank and triangle count. A simple Breadth-first Search shows there are direct flights to the Honolulu Airport (HNL) from ATL.

# Recommendations and further research

Cancelling flights in advance of extreme weather will reduce weather-related delays at Eagle County Regional Airport (EGE) since there is a pattern of extreme weather happening February of each year.

Although expensive, additional airport screening infrastructure coupled with an increased presence of security personnel, would reduce delays in Nashville (BNA) and Orlando (MCO) airports.

More research should be conducted to find the cause of delays in ATL to prevent the airport from having too many passengers at the same time.

refErences and bibLIography

Aven, J., 2017. *Sams Teach Yourself Apache Spark In 24 Hours*. Indianapolis, Indiana: Sams.

Kotheeranurak, D., 2020. *EDA on flight delay prediction with Apache PySpark Graphframes*. [Blog] *medium.com*,. (Kotheeranurak, 2020)

Yandex, 2020. *Big Data Analysis: Hive, Spark SQL, Dataframes And Graphframes*. [video] Available at: <https://www.coursera.org/learn/big-data-analysis/home/welcome> [Accessed 23 May 2020]. (Yandex, 2020)

Yablonsky, G., Steckel, R., Constales, D., Farnan, J., Lercel, D. and Patankar, M., 2014. Flight delay performance at Hartsfield-Jackson Atlanta International Airport. *Journal of Airline and Airport Management*, 4(1).

Appendix

Appendix a: carrier delays

A picture containing indoor, table

Description automatically generated

Source . <https://www.theguardian.com/world/2019/jan/25/flight-delays-laguardia-newark-philadelphia-shutdown>

appendix b: security delay

A screenshot of a cell phone

Description automatically generated

*Source 2.* [*https://www.newschannel5.com/bna-stands-out-as-tsa-warns-of-growing-problem-of-guns-found-at-airport-security*](https://www.newschannel5.com/bna-stands-out-as-tsa-warns-of-growing-problem-of-guns-found-at-airport-security)

appendix c: orlando gun issues

A close up of a sign

Description automatically generated

Source . <https://www.orlandoweekly.com/Blogs/archives/2019/02/07/orlando-international-airport-ranks-in-nations-top-5-for-tsa-gun-seizures>