CSC1109 Group Assignment - Final Report

Ronan Kelly – 21374881
Eoin Quinn – 21356173

Gitlab Repo Location:
https://github.com/Kellyronan03/CSC1109_assignment2

1: Introduction

Air Travel has become a crucial part of modern life for a lot of people. Whether it's corporate travel for individuals who work across different countries, or people who are lucky enough to go on holiday abroad, efficient travel systems are vital for economic growth, personal convenience, and customer satisfaction. As we have gained strong interests in travelling through our own experiences, we were intrigued by the idea of analysing a large dataset of US flight information from 2008, uncovering insights around travel patterns and operational performance.

With the use of Apache Pig, Hive, Spark as well as limited use of python, we hoped to show how these big data technologies can be extremely useful in analysing flight data, retrieving useful information regarding travel efficiency through big data analytics.

2: What Data

We found transportation data from the Harvard Dataverse at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HG7NV7 we were able to find data regarding flights from numerous airlines across a vast amount of airports in the USA. The data revolves around the timeliness of the flight arrival, providing details like expected departure and arrival times, actual departure and arrival times, reasons for delay, whether a flight was cancelled or not, flight origin and destination among others. With each row representing a flight, the dataset has 2389217 rows and 29 columns. Some of the crucial information such as actual airport and airline names, were encoded in the main dataset and had to be joined with other tables to retrieve the actual names. Some of the raw data can be seen below:

Month	DayofMon	DayOfWee	DepTime	CRSDepTi	ArrTime	CRSArrTim	UniqueCa	FlightNum	Actual Ela	CRSElapse	AirTime	ArrDelay	DepDelay	Origin	Dest	Distance	TaxiIn	TaxiOut	Cancelled
1	1	2	2152	2115	2316	2256	FL	66	84	101	66	20	37	ATL	IAD	533	5	13	0
1	1	2	2048	2050	12	2315	CO	1571	204	145	135	57	-2	CLE	MCO	895	5	64	0
1	1	2	1655	1612	1900	1759	OH	5269	125	107	82	61	43	CMH	LGA	478	10	33	0
1	1	2	1033	1035	1158	1220	AA	319	145	165	116	-22	-2	LGA	ORD	733	16	13	0

5: Related Work

We have seen the importance of analysing and predicting flight delays through numerous pieces of literature relevant to the analysis we want to convey. Understanding that flight delays heavily contribute to operational efficiency and flyer satisfaction. A study by Thiagarajan et al. utilized two machine learning models, one to predict whether a flight would be delayed or not, and then another to predict how much the flight would be delayed by[1]. Another study by Sai Durga et al. tried binary classification models and a long short term memory regression model to predict flight delays[2].

These related studies leveraged additional features using the likes of weather data and aircraft types. While we did not have access to these types of features, we hoped to achieve similar goals with the flight data we did have.

3: What Analytics

1. **Dataset Integration**
   Our analysis initially required the combination of three datasets: flights, carriers, and airports.
   - **Flights dataset:** Served as the primary base for integration.
   - **Carriers dataset:** Matched using the airline's unique carrier code to add a new column containing the full name of each airline.
   - **Airports dataset:** Matched based on the IATA codes for origin and destination airports, adding columns for the name, city, and state of both airports.

   Initially, we attempted to complete this integration in Apache Pig; however, due to its complexity, we decided to switch to Python, where pandas provided an efficient solution for merging and adding columns.
2. **Data Cleaning**
   After integrating the datasets, we performed cleaning tasks primarily in Apache Pig:
   - **Null Values:** Replaced empty cells with appropriate values to ensure consistent analysis and avoid issues caused by missing data.
   - **Duplicate Flights:** Removed duplicate records to maintain data integrity and prevent biases in our analysis.
   - **New Columns:**
     - Created a Status column to classify flights as *early*, *delayed*, *on-time*, or *cancelled*. This was based on whether the ArrDelay column value was less than, equal to, or greater than 0.
     - Added a Date column to consolidate date-related information and simplify the dataset. Columns indicating specific date components were subsequently dropped.
3. **Outcome**
   These steps were essential to create a clean, organised dataset, enabling us to conduct analytics experiments effectively. By resolving integration and cleaning challenges early, we ensured smoother downstream analysis and minimised potential errors.

4: Analytics Tasks and Results

1. **Time-Based Analysis**
   - **Worst Delays:** American Airlines had the worst delays, averaging 15.7 minutes on departure and 16 minutes on arrival.
   - **Best Delays:** Aloha Airlines had the best performance, arriving early on average with -1.5 minutes on departure and -2.8 minutes on arrival.

```
AirTran Airways Corporation,9.697029448826248,10.921710925467867
Alaska Airlines Inc.,7.0851685990634525,5.349467312843792
Aloha Airlines Inc.,-1.4824561403508771,-2.888673890608875
American Airlines Inc.,15.715686071439668,16.07303975889889
American Eagle Airlines Inc.,13.573224520149244,13.84414626953701
Atlantic Southeast Airlines,14.498481497186472,12.046070802342506
Comair Inc.,11.252396444547726,11.378347760413641
Continental Air Lines Inc.,13.087049755422699,11.828697070013341
Delta Air Lines Inc.,7.76639164250914,7.54839276422929
Expressjet Airlines Inc.,12.146476571694139,11.862788309459024
Frontier Airlines Inc.,6.682126141459104,7.183811736981639
Hawaiian Airlines Inc.,-1.2663454167122987,-1.1391961450005477
JetBlue Airways,11.970368026995676,9.817523086425332
Mesa Airlines Inc.,13.394302867249582,13.20407044814787
Northwest Airlines Inc.,8.106498060589018,11.289960550147628
Pinnacle Airlines Inc.,11.930286413649943,10.124908508417226
Skywest Airlines Inc.,9.362174311877391,10.056010042218086
Southwest Airlines Co.,11.864636351630043,7.603865065577601
US Airways Inc.,6.675109847505816,2.7307659171189798
United Air Lines Inc.,16.04111802408217,14.346067204805706
```

Fig 1. Average flight delays by airline

- **Delays and Cancellations:** Most delays occurred on Wednesdays (35,000+), while cancellations peaked on Tuesdays (11,701). March was the worst month for delays, and February had the most cancellations. Interestingly, delays were higher midweek rather than on weekends, contrary to expectations.

```
1,17203
2,20523
3,16064
4,10309
```

```
1,337141
2,345707
3,353765
4,339632
5,337655
6,281025
7,320783
```

Fig 2. Total Monthly Delays, Fig 3. Total Delays by day

```
1,586165
2,546456
3,597615
4,585472
```

```
1,9440
2,11763
3,10315
4,8787
5,11501
6,6155
7,6138
```

Fig 4. Total Monthly Cancellations, Fig 5. Total Cancellations by day

2. Geographical Analysis

- **Delay Totals and Rates:** California had the most delays (>15,000), reflecting its high flight volume. New Jersey had the highest delay rate, with 64% of flights delayed.

```
OriginState,TotalFlights,DelayedFlights,CancelledFlights,DelayRate
NJ,49054,31474,1935,64.16194398010356
IL,156513,98234,9863,62.76411544088989
MI,69663,43124,2201,61.90373656029743
MD,34575,20797,413,60.150397686189436
MN,47791,28599,1054,59.84181121968572
```

Fig 6. Delay Rate by State

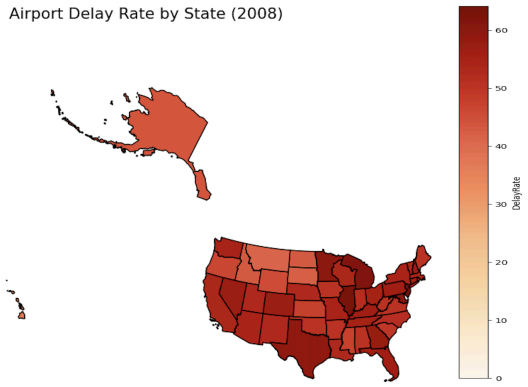

Fig 7. Map of Average state delays

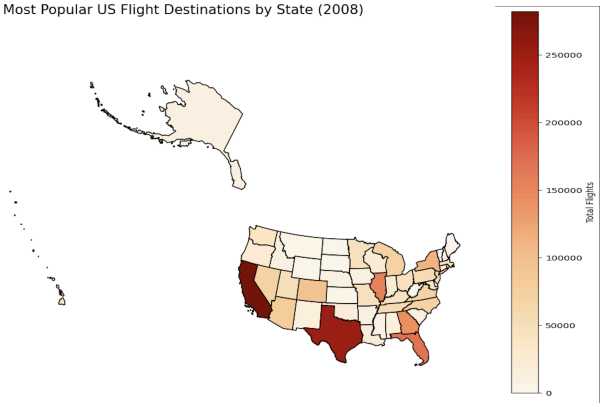

Fig 8. Map of total state delays

- **Tourist Hotspots:** The top destinations were California, Texas, and Florida, likely due to their tourist appeal. Surprisingly, New York didn't rank in the top five.

```
DestState,TotalFlights
CA,282434
TX,249999
FL,169545
IL,156511
GA,142504
```

Fig 9. Most popular flight destinations

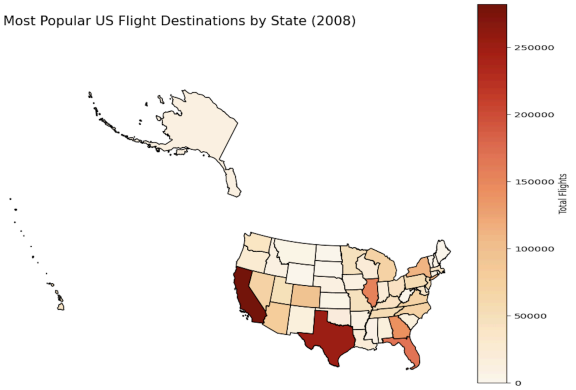

Fig 10. Map of most popular flight destinations

3. **Categorical Analysis**
    - **Delay Reasons:** Carrier delays were the most common cause of flight disruptions in the first four months of 2008, often due to late aircraft or airline-related issues.

```
DelayType,Count
No Delay,1797514
Carrier Delay,257621
NAS Delay,209787
Late Aircraft Delay,78649
Weather Delay,34941
Security Delay,1295
```

Fig 11. Most common causes for delay

4. **Machine Learning Experiment**
    - **Objective:** Predict flight outcomes as on-time, early, delayed, or cancelled using features like date, times, flight distance, taxi time, origin, destination, airline, and flight number.
    - **Models and Results:** Logistic Regression performed best (64.8% accuracy), followed by Random Forest and Naïve Bayes (59.3%). All models excelled at predicting delays, with Random Forest achieving 100% accuracy for this category.

```
Logistic Regression Accuracy: 0.6477896695160643
Logistic Regression Confusion Matrix:
Predicted\Actual  0    1    2    3
       0       170009   58895   12   0
       1       96214   124557   19   0
       2       1   13   13938   0
       3       7424   5159   0   0
```

```
Random Forest Accuracy: 0.6343321973538607
Random Forest Confusion Matrix:
Predicted\Actual  0    1    2    3
       0       193698   35206   12   0
       1       126326   94445   19   0
       2       0   0   13952   0
       3       9344   3239   0   0
```

```
Naive Bayes Accuracy: 0.5928384998351675
Naive Bayes Confusion Matrix:
Predicted\Actual  0    1    2    3
       0       160105   68811   0   0
       1       101499   119291   0   0
       2       10706   308   2938   0
       3       7544   5039   0   0
```

Fig 12,13,14. Classification Model results

6: Challenges

As mentioned, one major challenge we found was right at the beginning when attempting to join the 3 tables together. We initially tried to do this in PIG spending multiple hours trying to no avail. At this point we transitioned to doing the table joins through python before conducting our proper cleaning through PIG.

Another issue we had was when we had to create the heatmaps for our popular destinations and location delays queries. We conducted this through GeoPandas which was very useful, however, it required a lot of time creating the dataframe and mapping the state variables and then finding the exact map shapefile that worked for our analysis was a task of trial and error.

Finally, and what was properly the most difficult task was creating the classification experiment. Although, on the surface the task was fairly simple, this was an experiment we had very little experience in, so had very little understanding of how to create the functions and the preprocessing that would be necessary to run the models in spark. This was an experiment we spent a lot of time working on before we got it right.

That being said, it was a great experience to learn how to conduct these experiments in a new environment and language, and getting the opportunity to deal with problems of this nature.

7: Responsibility statement

Despite the fact we largely worked on the majority of this project together, we felt it was important to separate some of the tasks so that we had the ability to work simultaneously, or also when we were not together.

Eoin was predominantly the leader of the group, and so he led much of the initial direction of the project, including the decision on the project idea and dataset. With his experience working on cleaning and processing large datasets throughout his summer internship, we felt it was best to let him conduct the cleaning and processing through PIG and Python. He also led the machine learning experiment. Since it was his idea to conduct this experiment from the start, we agreed it would be fitting for him to carry out this part of the work.

Ronan spent considerable time conducting various Hive and PIG experiments during the first assignment, so we thought it best for him to handle both the simple and advanced queries through Hive and Spark. He also felt the most confident working on Spark queries during the semester, so we believed he would enjoy this role the most.

However, as this was a group project, we emphasized the importance of communicating with each other at every stage. We also worked closely to ensure any issues we encountered were addressed with multiple viewpoints, allowing us to solve them efficiently.

8: Response to peer feedback

The feedback we received on the mid-way report was very useful for us. When we wrote that report the project was still in the very early stages and we didn't have a full idea of exactly the route we were going to go down, so the feedback provided a good bit of guidance for us as we were conducting the experiments.

The discussion of what models we would be using was incredibly useful, as it encouraged us to use a variety of models to test out their accuracies. They also emphasised the importance of clean data for those classification models which was very important as we encountered these issues when we began modelling.

The idea of using geographical visualisations was also not an idea we had considered, so this was very useful for us as we conducted some of our experiments.

Bibliography

1. B. Thiagarajan, L. Srinivasan, A. V. Sharma, D. Sreekanthan, and V. Vijayaraghavan, "A Machine Learning Approach for Prediction of On-time Performance of Flights," in *Proceedings of the IEEE Conference on Emerging Trends in Engineering and Technology*, 2023. [Online]. Available: balasubramanian.in.1995@ieee.org.

2. R. S. Durga and K. S. Sri, "Predicting Flight Delays Using Machine Learning Techniques and Aviation Big Data," *International Journal of Research in Electronics and Computer Engineering (IJRECE)*, vol. 11, no. 2 Apr.–June 2023. ISSN: 2393-9028 (Print), ISSN: 2348-2281 (Online).