College: Engineering and Information Technology
Department: Information Technology

Busi. Intell.&Data Warehousing course Project
INS402

Analyzing Flight Delays and Cancellations to Identify Airlines with Optimal On-Time Performance

Prepared by:
202011025 – Maryam Ali,
202011604 – Joti Rani Biswas
202011669 – Marwa Fadl
201810326 – Fai Musaad

Supervised by:
Dr. Shahnorbanun Binti Sahran,
Eng. Mahganj Aslam Durra

Academic Year 2023- 2024
Date: November 19, 2023

List of Contents

List of Figures

List of Tables

Abstract

In this report, the complex analysis of flight delays, and cancellations using Kaggle's 2015 dataset will be discussed. This examination is undertaken through a meticulous ETL (Extract, Transform, Load) process, as well as the strategic implementation of advanced analytical models, which includes XGBoost, LightGBM, and CatBoost. The ETL process involves essential steps such as loading data, feature selection, data exploration, handling the missing data, and shaping the dataset for subsequent analysis. The selection of each model is based on its specific strengths: XGBoost's versatility, LightGBM's emphasis on speed and efficiency, and CatBoost's excellence in handling categorical features. The paper also includes a thorough literature review on how these models can be applied in various ways with close attention paid to their strengths and weaknesses. Next, the analytical focus shifts towards the design of a Power BI dashboard, strategically incorporating visual elements such as card visuals, bar charts, date slicers, and multi-slicers. This dashboard aims to visually present and dissect flight cancellation data, emphasizing key insights such as cancellation percentages, the performance of different airlines, and periods of peak delays. This integrated approach seamlessly combines data refinement, advanced modelling, and visualization, offering a holistic solution to comprehend and address issues related to flight delays and cancellations comprehensively.

Overview of the Business Scenario

In our modern, rapidly moving society, air travel plays a crucial role in our daily lives. Despite its significance, the inconvenience of flight delays and cancellations can significantly disrupt our schedules and evoke a sense of frustration. The economic consequences of flight delays reverberate across passengers, airlines, and airports. Faced with the unpredictable nature of these interruptions, travelers frequently allocate extra hours to their journeys, resulting in escalated travel costs to guarantee on-time arrival. The disparity between the initially scheduled, and the real departure or arrival times of a plane serves as a metric for measuring the impact of flight delays. On the other hand, upon Analyzing the root causes of flight cancellations provides valuable information. Therefore, the Weather and Natural Disasters stands out as the predominant factor, accounting for nearly 54% of instances where flights are called off. (Saedi, 2023) There are multiple reasons of flight delays and cancellations such as Air Traffic Congestion, Technical Failures, Aircraft

Turnaround Time, and Mechanical problems, crew scheduling issues, or other operational challenges within the airline can contribute to flight disruptions. This project going to Analyzing and discuss Kaggle's 2015 Flight Delays and Cancellations datasets (2015 Flight delays and cancellations, 2017). Therefore, in this instance, business intelligence (BI) will be applied by leveraging data visualization models to gain insights, analyzing the data related to the problem, identify patterns, and make informed decisions to overcome the problems (Sternberg, 2017).

Business Objective

- Identify the problem by Collect flight, weather, and airport data, then integrate from diverse sources to form a comprehensive dataset on flight delays and cancellations.
- Create drill-down visualizations to investigate the root causes of delays.
- Use BI tools to explore the dataset visually, and Identify trends, patterns, anomalies that may indicate the underlying factors contributing to delays and cancellations.
- Design dashboards that provide a real-time overview by Utilize visuals like card visuals, bar charts, date slicers, multi-slicers, and gauges to illustrate metrics such as on-time performance, cancellation percentages, the performance of different airlines, and periods of peak delays.
- Implement predictive models to forecast potential delays.
- Identify the possible solution for preventing flight delays and cancellations.
- Implementation of Preventive Measures such as improved scheduling, resource allocation, or operational changes.

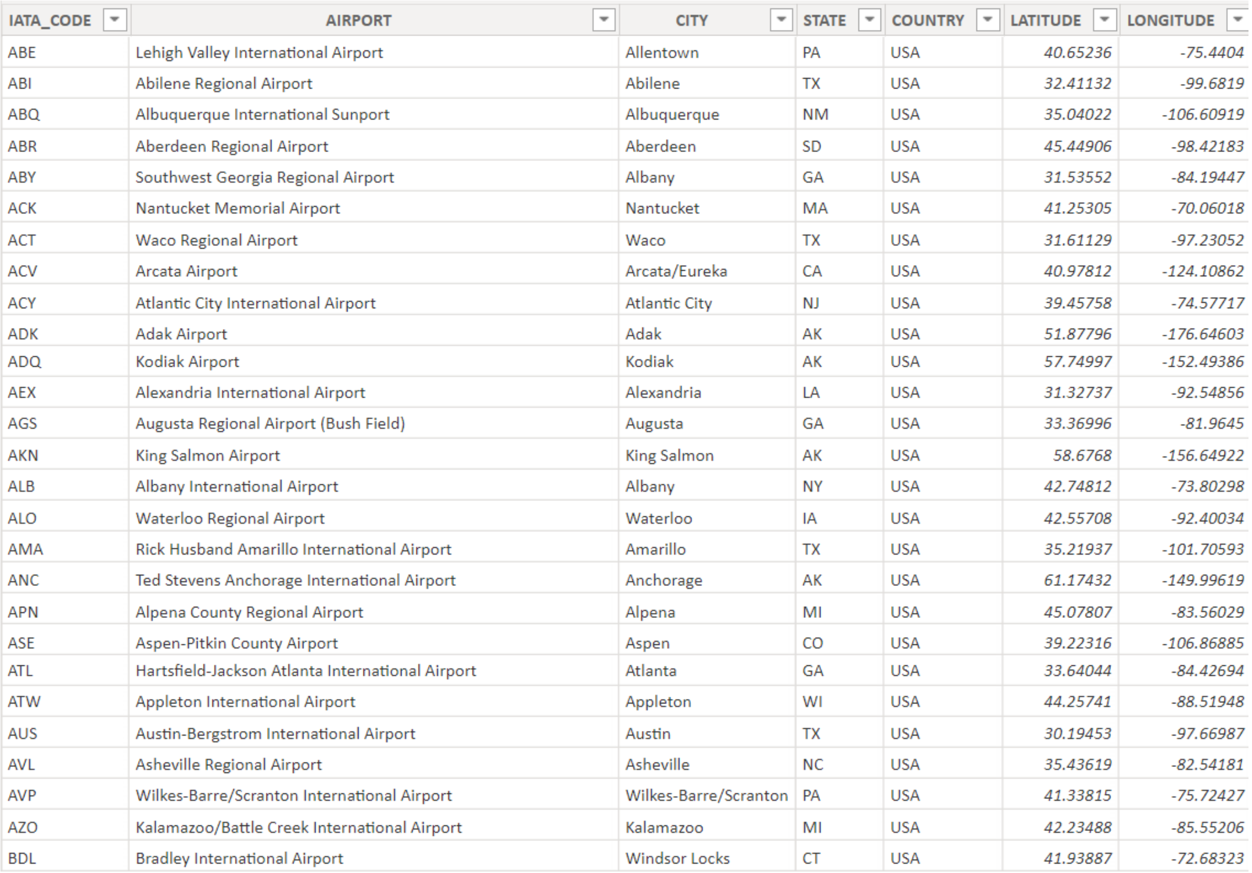
List of Analytical Model Questions

- What can we say about planes cancellations and delay based on the airline associated?
- What are the reasons for such delays and cancellations?
- Is there a specific period where flights cancellations and delays were noticeable at a high peak?
- What airlines need to be more investigated because of poor performance?
- How can we fix these issues and improve airlines' performance?

Data Understanding

i. Snapshot of the Used Data

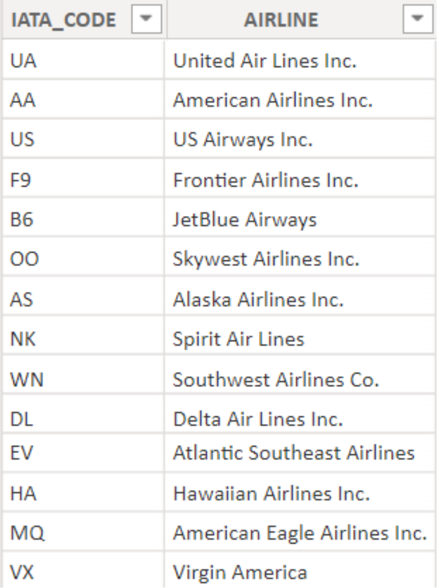
The Airports dataset has seven (7) columns and includes 323 rows of Airports.



IATA_CODE	AIRPORT	CITY	STATE	COUNTRY	LATITUDE	LONGITUDE
ABE	Lehigh Valley International Airport	Allentown	PA	USA	40.65236	-75.4404
ABI	Abilene Regional Airport	Abilene	TX	USA	32.41132	-99.6819
ABQ	Albuquerque International Sunport	Albuquerque	NM	USA	35.04022	-106.60919
ABR	Aberdeen Regional Airport	Aberdeen	SD	USA	45.44906	-98.42183
ABY	Southwest Georgia Regional Airport	Albany	GA	USA	31.53552	-84.19447
ACK	Nantucket Memorial Airport	Nantucket	MA	USA	41.25305	-70.06018
ACT	Waco Regional Airport	Waco	TX	USA	31.61129	-97.23052
ACV	Arcata Airport	Arcata/Eureka	CA	USA	40.97812	-124.10862
ACY	Atlantic City International Airport	Atlantic City	NJ	USA	39.45758	-74.57717
ADK	Adak Airport	Adak	AK	USA	51.87796	-176.64603
ADQ	Kodiak Airport	Kodiak	AK	USA	57.74997	-152.49386
AEX	Alexandria International Airport	Alexandria	LA	USA	31.32737	-92.54856
AGS	Augusta Regional Airport (Bush Field)	Augusta	GA	USA	33.36996	-81.9645
AKN	King Salmon Airport	King Salmon	AK	USA	58.6768	-156.64922
ALB	Albany International Airport	Albany	NY	USA	42.74812	-73.80298
ALO	Waterloo Regional Airport	Waterloo	IA	USA	42.55708	-92.40034
AMA	Rick Husband Amarillo International Airport	Amarillo	TX	USA	35.21937	-101.70593
ANC	Ted Stevens Anchorage International Airport	Anchorage	AK	USA	61.17432	-149.99619
APN	Alpena County Regional Airport	Alpena	MI	USA	45.07807	-83.56029
ASE	Aspen-Pitkin County Airport	Aspen	CO	USA	39.22316	-106.86885
ATL	Hartsfield-Jackson Atlanta International Airport	Atlanta	GA	USA	33.64044	-84.42694
ATW	Appleton International Airport	Appleton	WI	USA	44.25741	-88.51948
AUS	Austin-Bergstrom International Airport	Austin	TX	USA	30.19453	-97.66987
AVL	Asheville Regional Airport	Asheville	NC	USA	35.43619	-82.54181
AVP	Wilkes-Barre/Scranton International Airport	Wilkes-Barre/Scranton	PA	USA	41.33815	-75.72427
AZO	Kalamazoo/Battle Creek International Airport	Kalamazoo	MI	USA	42.23488	-85.55206
BDL	Bradley International Airport	Windsor Locks	CT	USA	41.93887	-72.68323

Figure I: Airports Dataset

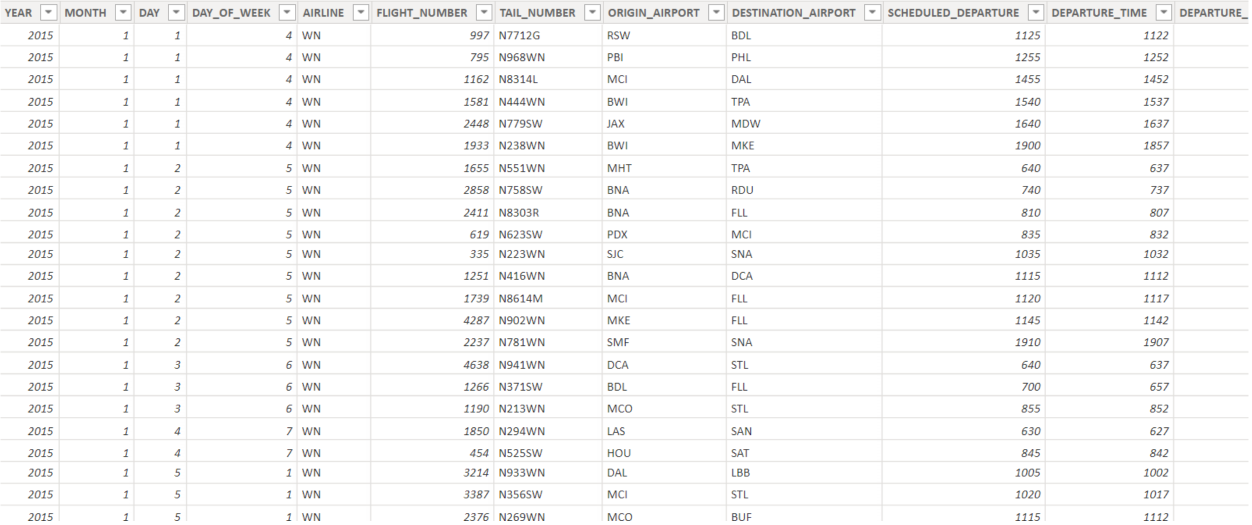
Airlines dataset has 2 columns and includes 14 rows of recorded airlines.



IATA_CODE	AIRLINE
UA	United Air Lines Inc.
AA	American Airlines Inc.
US	US Airways Inc.
F9	Frontier Airlines Inc.
B6	JetBlue Airways
OO	Skywest Airlines Inc.
AS	Alaska Airlines Inc.
NK	Spirit Air Lines
WN	Southwest Airlines Co.
DL	Delta Air Lines Inc.
EV	Atlantic Southeast Airlines
HA	Hawaiian Airlines Inc.
MQ	American Eagle Airlines Inc.
VX	Virgin America

Figure II: Airlines dataset

The Flights dataset has 31 columns and includes 1,048,575 rows of recorded flights.

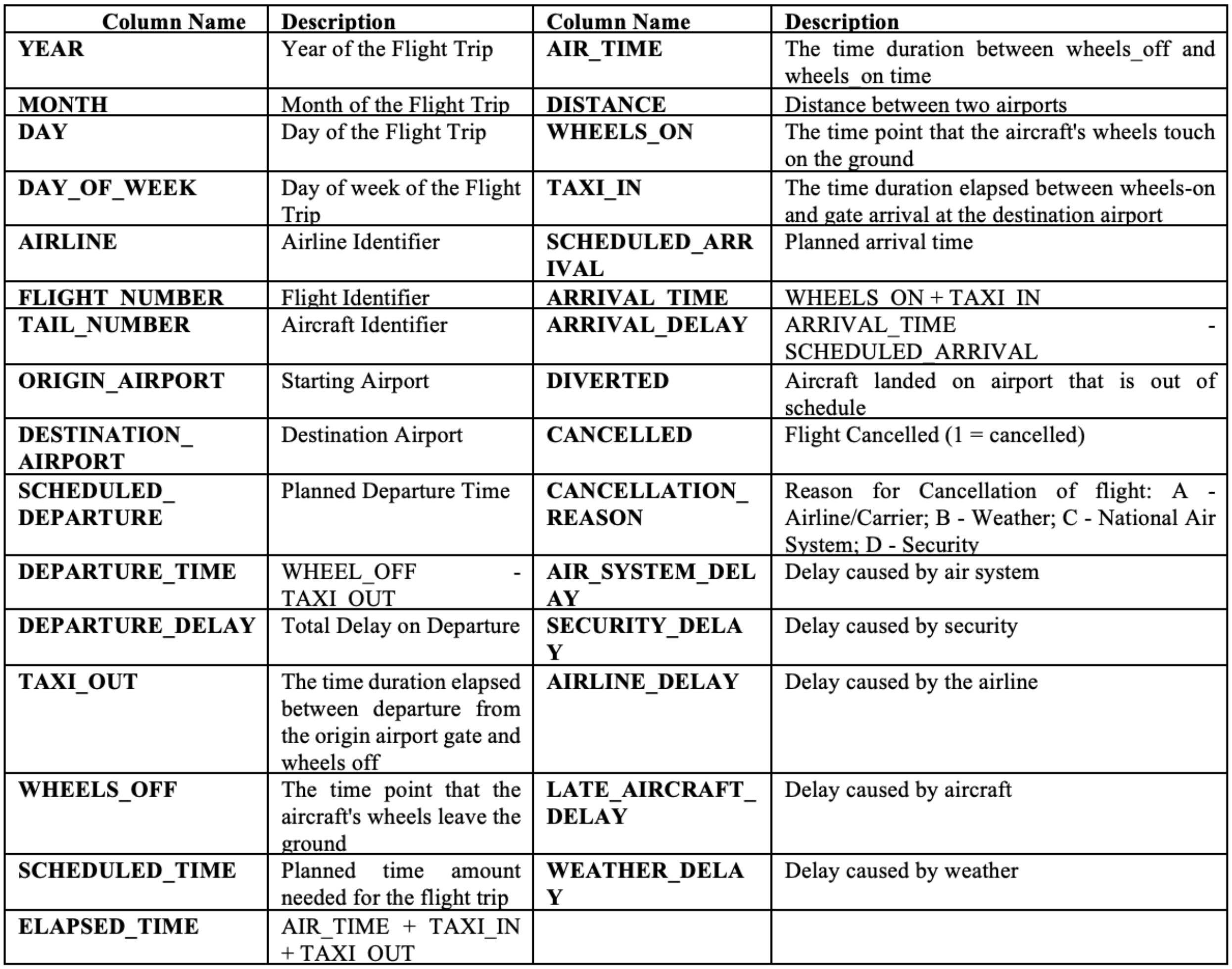


YEAR	MONTH	DAY	DAY_OF_WEEK	AIRLINE	FLIGHT_NUMBER	TAIL_NUMBER	ORIGIN_AIRPORT	DESTINATION_AIRPORT	SCHEDULED_DEPARTURE	DEPARTURE_TIME	DEPARTURE_
2015	1	1	4	WN	997	N7712G	RSW	BDL	1125	1122	
2015	1	1	4	WN	795	N968WN	PBI	PHL	1255	1252	
2015	1	1	4	WN	1162	N8314L	MCI	DAL	1455	1452	
2015	1	1	4	WN	1581	N444WN	BWI	TPA	1540	1537	
2015	1	1	4	WN	2448	N779SW	JAX	MDW	1640	1637	
2015	1	1	4	WN	1933	N238WN	BWI	MKE	1900	1857	
2015	1	2	5	WN	1655	N551WN	MHT	TPA	640	637	
2015	1	2	5	WN	2858	N758SW	BNA	RDU	740	737	
2015	1	2	5	WN	2411	N8303R	BNA	FLL	810	807	
2015	1	2	5	WN	619	N623SW	PDX	MCI	835	832	
2015	1	2	5	WN	335	N223WN	SJC	SNA	1035	1032	
2015	1	2	5	WN	1251	N416WN	BNA	DCA	1115	1112	
2015	1	2	5	WN	1739	N8614M	MCI	FLL	1120	1117	
2015	1	2	5	WN	4287	N902WN	MKE	FLL	1145	1142	
2015	1	2	5	WN	2237	N781WN	SMF	SNA	1910	1907	
2015	1	3	6	WN	4638	N941WN	DCA	STL	640	637	
2015	1	3	6	WN	1266	N371SW	BDL	FLL	700	657	
2015	1	3	6	WN	1190	N213WN	MCO	STL	855	852	
2015	1	4	7	WN	1850	N294WN	LAS	SAN	630	627	
2015	1	4	7	WN	454	N525SW	HOU	SAT	845	842	
2015	1	5	1	WN	3214	N933WN	DAL	LBB	1005	1002	
2015	1	5	1	WN	3387	N356SW	MCI	STL	1020	1017	
2015	1	5	1	WN	2376	N269WN	MCO	BUF	1115	1112	

Figure III: Flights dataset

ii. Data Attributes

Table I: Data Attributes Description



Column Name	Description	Column Name	Description
YEAR	Year of the Flight Trip	**AIR_TIME**	The time duration between wheels_off and wheels_on time
MONTH	Month of the Flight Trip	**DISTANCE**	Distance between two airports
DAY	Day of the Flight Trip	**WHEELS_ON**	The time point that the aircraft's wheels touch on the ground
DAY_OF_WEEK	Day of week of the Flight Trip	**TAXI_IN**	The time duration elapsed between wheels-on and gate arrival at the destination airport
AIRLINE	Airline Identifier	**SCHEDULED_ARRIVAL**	Planned arrival time
FLIGHT_NUMBER	Flight Identifier	**ARRIVAL_TIME**	WHEELS_ON + TAXI_IN
TAIL_NUMBER	Aircraft Identifier	**ARRIVAL_DELAY**	ARRIVAL_TIME - SCHEDULED_ARRIVAL
ORIGIN_AIRPORT	Starting Airport	**DIVERTED**	Aircraft landed on airport that is out of schedule
DESTINATION_ AIRPORT	Destination Airport	**CANCELLED**	Flight Cancelled (1 = cancelled)
SCHEDULED_ DEPARTURE	Planned Departure Time	**CANCELLATION_ REASON**	Reason for Cancellation of flight: A - Airline/Carrier; B - Weather; C - National Air System; D - Security
DEPARTURE_TIME	WHEEL_OFF - TAXI_OUT	**AIR_SYSTEM_DELAY**	Delay caused by air system
DEPARTURE_DELAY	Total Delay on Departure	**SECURITY_DELAY**	Delay caused by security
TAXI_OUT	The time duration elapsed between departure from the origin airport gate and wheels off	**AIRLINE_DELAY**	Delay caused by the airline
WHEELS_OFF	The time point that the aircraft's wheels leave the ground	**LATE_AIRCRAFT_ DELAY**	Delay caused by aircraft
SCHEDULED_TIME	Planned time amount needed for the flight trip	**WEATHER_DELAY**	Delay caused by weather
ELAPSED_TIME	AIR_TIME + TAXI_IN + TAXI_OUT		

Chosen Data Mining/Analytical Model

i. ETL

a. Loading Data:

A large dataset comprised of over 5.8 million entries and 31 attributes was read using the Pandas library in the first step. A comprehensive understanding of the dataset's structure and content was necessary to fully comprehend it. We then randomly selected 10,000 rows from the full dataset to facilitate more focused analyses and modelling.

b. Selecting Features:

To identify the most relevant attributes for subsequent analyses, a meticulous feature selection process was conducted. Features such as temporal details (MONTH, DAY, DAY_OF_WEEK), airline-specific data (AIRLINE, FLIGHT_NUMBER), and operational specifics (DESTINATION_AIRPORT, ORIGIN_AIRPORT, SCHEDULED_DEPARTURE, DEPARTURE_TIME, DEPARTURE_DELAY, SCHEDULED_ARRIVAL, ARRIVAL_TIME, ARRIVAL_DELAY, AIR_TIME, DISTANCE) were strategically chosen. Pandas Profiling was also used to provide more insight, and certain features showed high correlations.

c. Data Exploration:

Identifying potential areas for further research was accomplished by exploring patterns and relationships between selected features. As a result of this phase, a high correlation within the dataset was detected, which provided valuable insight into interdependencies. To make informed decisions during future analyses, it was crucial to have a thorough understanding of these correlations.

d. Handling Missing Data:

The preparation of data included dealing with missing values. In terms of departure and arrival times, departure delays, arrival times, and flight times, different degrees of missing data were observed. Approximately 2% of the data set contained missing values, which were carefully evaluated. The instances have been removed from the dataset after being reviewed for their impact on overall dataset integrity. Imputing values to critical temporal features would lead to potential inaccuracies.

e. Final Dataset:

A dataset containing 9,013 rows and 15 features was developed through the cleaning of the data and the selection of features. Using this curated dataset, exploratory data analysis (EDA) and subsequent modelling could be carried out with a refined and structured foundation.

f. Outcome:

A dataset containing 9,013 rows and 15 features was developed through the cleaning of the data and the selection of features. By analyzing this prepared dataset, a deeper exploration of factors affecting flight delays and cancellations was possible.

We can conclude from this to perform in-depth analyses and modelling, the 2015 Flight Delay & Cancellation dataset was meticulously prepared using Data Extraction, Transformation, and Load (ETL). Initially, Pandas DataFrame was loaded with over 5.8 million rows and 31 features to analyze the dataset. Among the many details in this dataset were temporal information, airline-specific data, and operational details. Randomly selecting 10,000 rows from the dataset provided a manageable yet still robust subset for further processing while maintaining the dataset's representativeness. A segment of essential features was selected for analysis during the data exploration phase by using its structure and characteristics to understand factors associated with flight cancellations and delays. It was discovered that certain features had a high correlation, highlighting potential areas for improvement, because of implementing Pandas Profiling. In addition to departure time, departure delay, arrival time, arrival delay, and airtime, this exploration identified certain features with missing values. Since missing values accounted for less than 2% of the dataset, it was decided to drop instances with missing values. To ensure that the dataset was consistent and of high quality for further analysis, this approach was used. As a result of the refinement and focus of this dataset, the exploration and modelling of the data can be more meaningful. In the aviation industry, flight delays and cancellations are influenced by a variety of complex factors. A meticulous ETL process lays the foundation for extracting valuable information about these factors.

ii. Model Description

XGBoost (Extreme Gradient Boosting):

- It is renowned for its speed and performance as a scalable and optimized gradient boosting library. By combining multiple low-confidence models (typically decision trees) into one, boosting ensemble techniques can make accurate and robust predictions. With XGBoost, the objective function is regularized, which reduces the chance of overfitting. The program excels at handling large datasets and can be used for both classification and regression tasks.

LightGBM (Light Gradient Boosting Machine):

- With LightGBM, Microsoft focused on speed and efficiency when developing gradient boosting frameworks. It is an excellent tool for handling large data sets and high-dimensional data. LightGBM provides distributed training, which makes it suitable for implementation in parallel or distributed computing systems. A multitude of classification and regression tasks are performed using it, especially when sparse features are present.

CatBoost:

- With CatBoost, categorical features are handled seamlessly using gradient boosting. There is no need to manually encode categorical variables with this model since it supports categorical variables out-of-the-box. CatBoost performs better in tasks involving ordinal categories due to its use of an ordered boosting technique. Using categorical features to predict flight delays, for example, is a good use of the intuitive and robust system.

iii. Literature Review on the Selected Model

XGBoost:

- XGBoost has been studied extensively since its introduction and gained widespread popularity. Through regularization techniques, it prevents overfitting when dealing with complex datasets. Finance, healthcare, and natural language processing are some of the fields in which XGBoost has been tested. XGBoost models are investigated for their interpretability in some studies, while hyperparameters are refined to optimize performance.

LightGBM:

- When dealing with large datasets, LightGBM's speed and memory efficiency are often emphasized. The performance of this algorithm compared to other gradient boosting algorithms has been studied in scenario scenarios with sparse features. It has also been analyzed for big data analytics due to its distributed computing capabilities.

CatBoost:

- The CatBoost approach manages categorical features without the need for extensive preprocessing, as opposed to other approaches. According to literature, CatBoost excels at categorical data analysis. Compared to traditional gradient boosting models, it is more effective in dealing with categorical data

iv. Model Functionality

XGBoost:

- With XGBoost, you can use gradient boosting to boost your performance. In ensemble learning, multiple weak learners, typically decision trees, combine their predictions to produce a final prediction. There are numerous machine learning problems you can solve with XGBoost, including classification, regression, and ranking. Support for parallel processing, regularization to prevent overfitting, and handling of missing data are some of the key functionalities. Because of its flexibility and speed, XGBoost is popular for predictive modeling, and it can be customized for different datasets and objectives.

LightGBM:

- This framework is designed to optimize performance and efficiency through gradient boosting. The tree-based approach provides faster training compared to XGBoost but utilizes histogram-based bins for continuous features. A lightweight, memory-efficient, scalable, and distributed training framework, LightGBM can handle large datasets well. When dealing with sparse and high-dimensional data, it is particularly effective.

CatBoost:

- Yandex has developed a gradient boosting library called CatBoost, which stands for Categorical Boosting. Categorical features can be handled without extensive preprocessing, which is its primary strength. A technique called ordered boosting is used in CatBoost to handle categorical variables naturally. As well as handling overfitting, it incorporates advanced strategies that make it robust and easy to use. A dataset that includes categorical features is particularly well suited for CatBoost classification tasks.

v. Model Evaluation

XGBoost:

- The ROC AUC score of XGBoost has been used to evaluate the performance of the model. The ROC AUC score of XGBoost has been used to evaluate the performance of the model. By comparing the positive and negative classes, it assesses the ability of the model to make distinctions between them. Performance is better when ROC AUC is higher. As shown in the results for our dataset, XGBoost showed ROC AUC Train of 0.884 and ROC AUC Test of 0.844.

LightGBM:

- Similar metrics can be used to evaluate LightGBM, known for its efficiency and speed. ROC AUC, precision, recall, and F1 score are all commonly used metrics for evaluating LightGBM. Model performance can be evaluated based on these metrics across a variety of classification aspects.

CatBoost:

- ROC AUC, precision, recall, and F1 score are a few of the metrics CatBoost has to offer due to its specialization in handling categorical features. Modelling becomes more efficient by simplifying categorical variables' preprocessing with CatBoost. The system can be tested and trained using both training and testing datasets, which will help ensure robust classification capabilities.

General Considerations:

Each model must be considered not only for its performance but for its generalization to unknown data as well. Underfitting and overfitting should be avoided at all costs. Additionally, confusion matrices, precision-recall requested the importance of features can be used to gain a deeper understanding of model behavior. Also, the accuracy merely shows if the model used was able to predict whether a trip is going to be delayed or not based on specific features.

vi. Advantages and Disadvantages of the Chosen Model

XGBoost:

- Advantages:
  - Versatility: There are many types of datasets that can be worked with by XGBoost, and it performs well on all of them.
  - Regularization: Overfitting can be prevented with regularization techniques.
  - Ensemble Learning: In XGBoost, weak learners are combined to enhance prediction accuracy.
  - Speed and Efficiency: The speed and efficiency of this program are attributed to parallel computing and tree pruning.
- Disadvantages:
  - Sensitivity to Imbalanced Data: If highly imbalanced datasets are not adjusted or adjusted with additional techniques, XGBoost may not perform optimally.
  - Complexity: Hyperparameter tuning is required for optimal results due to the extensive set of parameters.

LightGBM:

- Advantages:
  - Speed and Efficiency: Large datasets are efficiently handled by LightGBM due to its low memory usage.
  - High-Dimensional Data: Using this method is especially suitable for datasets that have many features or dimensions.
  - Distributed Training: Easily scales to large datasets with distributed training.
  - Categorical Feature Handling: The algorithm is fast and does not require one-hot encoding when handling categorical features.
- Disadvantages:
  - Sensitivity to Noisy Data: When complex models are fitted, data that are noisy may be more sensitive to them.
  - Potential Overfitting Risks: When using noisy or small datasets, LightGBM may overfit, despite its speed and efficiency.

CatBoost:

- Advantages:
  - Categorical Feature Handling: CatBoost excels in handling categorical features, eliminating the need for extensive preprocessing.
  - Reduction of Overfitting: Built-in regularization techniques reduce the risk of overfitting.
  - Ease of Implementation: User-friendly and requires less parameter tuning.
  - Performance on Diverse Datasets: Performs well on various types of datasets.
- Disadvantages:
  - Encoding Overhead: Encoding might introduce some computational overhead even though it is efficient for handling categorical features.
  - Limited Interpretability: As CatBoost models are complex, it can be difficult to understand how they make decisions, hindering transparency.

At the end we can conclude that Our machine learning models include XGBoost, LightGBM, and CatBoost because of their unique strengths. It has also been renowned for its success in various competitions, in addition to employing ensemble learning and regularization techniques. For large data sets and scenarios with sparse features, LightGBM excels in speed and memory efficiency. The algorithm has been optimized to use low amounts of memory, so it can handle datasets that might otherwise be impossible to fit into its memory. Its user-friendly implementation is well-suited to tasks that involve categorical variables, as CatBoost specializes in effortless handling of categorical features. As a result of the combination of the two models, machine learning challenges can be handled efficiently and accurately. LightGBM optimizes memory usage, and XGBoost scales efficiently to large datasets, enabling CatBoost to simplify preprocessing by handling categorical features without encoding one-hot. LightGBM's complex models may be affected by noise due to its categorical feature encoding. CatBoost's encoding of categorical features may introduce some computational overhead. A highly imbalanced dataset may present a challenge for XGBoost. Even though each model has limitations, they provide a robust framework for handling a wide range of machine learning scenarios. Here in this table, it shows the result that we get in our machine learning. (Team, 2023)

Table II: Machine Learning Models Summarization

	XGBoost	**LightGBM**	**CatBoost**
Advantages	Scalability	Low Memory Usage	Handling Categorical Features
Disadvantages	Imbalanced Data	Increased Sensitivity to Noisy Data	Categorical Feature Encoding Overhead
Accuracy	91	90	90
Evaluation Matric	metrics.roc_auc_score	ROC AUC	ROC AUC
Mean Squared Error	0.082	0.079	0.063

Dashboard

i. Objective

This section's objective is to prepare and analyze the data of interest, flights delay and cancellation data. This includes having the data in the desired format and using all the informative columns that would aid with understanding the behavior of different airports and how it is connected to customer churn and unsatisfactory.

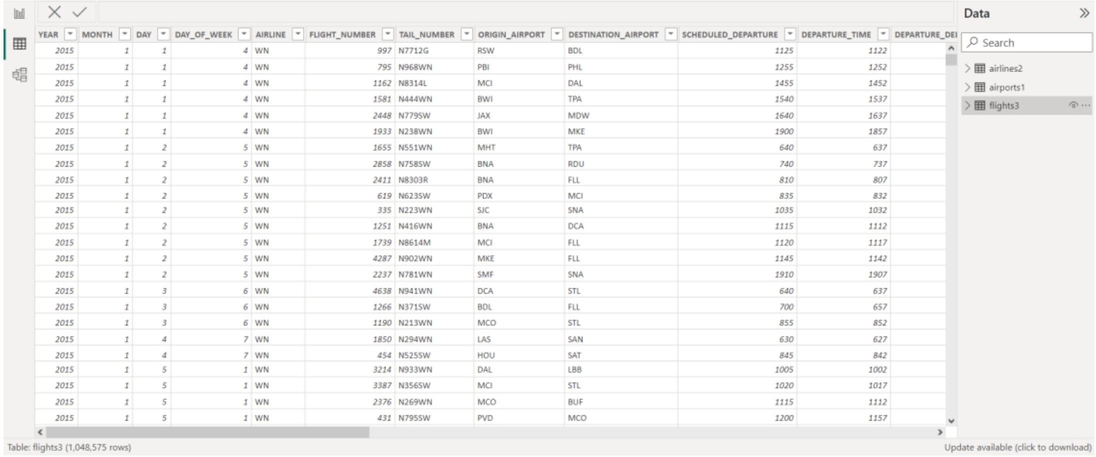
ii. Requirements

To achieve the desired output, this project has accessed Kaggle platform to download the data necessary to visualize the flights cancellation and delay problem. It also utilized Microsoft Power BI software to transform and load the data for visualizations of all the tables. It also needs to understand the relationship between those tables, familiarizing oneself with the different concepts. In addition to that feature engineering, command line knowledge, data analysis, visualization design skills, and performing quality check.

iii. Steps

Step 1: Download the data set needed for the project objective.

Step 2: Upload the data to the Power BI, this ensures that the data is placed in a unified location for visualizations.



YEAR	MONTH	DAY	DAY_OF_WEEK	AIRLINE	FLIGHT_NUMBER	TAIL_NUMBER	ORIGIN_AIRPORT	DESTINATION_AIRPORT	SCHEDULED_DEPARTURE	DEPARTURE_TIME	DEPARTURE_DEI
2015	1	1	4	WN	997	N7712G	RSW	BDL	1125	1122	
2015	1	1	4	WN	795	N968WN	PBI	PHL	1255	1252	
2015	1	1	4	WN	1162	N8314L	MCI	DAL	1455	1452	
2015	1	1	4	WN	1581	N444WN	BWI	TPA	1540	1537	
2015	1	1	4	WN	2448	N779SW	JAX	MDW	1640	1637	
2015	1	1	4	WN	1933	N238WN	BWI	MKE	1900	1857	
2015	1	2	5	WN	1655	N551WN	MHT	TPA	640	637	
2015	1	2	5	WN	2858	N758SW	BNA	RDU	740	737	
2015	1	2	5	WN	2411	N8303R	BNA	FLL	810	807	
2015	1	2	5	WN	619	N623SW	PDX	MCI	835	832	
2015	1	2	5	WN	335	N223WN	SJC	SNA	1035	1032	
2015	1	2	5	WN	1251	N416WN	BNA	DCA	1115	1112	
2015	1	2	5	WN	1739	N8614M	MCI	FLL	1120	1117	
2015	1	2	5	WN	4287	N902WN	MKE	FLL	1145	1142	
2015	1	2	5	WN	2237	N781WN	SMF	SNA	1910	1907	
2015	1	3	6	WN	4638	N941WN	DCA	STL	640	637	
2015	1	3	6	WN	1266	N371SW	BDL	FLL	700	657	
2015	1	3	6	WN	1190	N213WN	MCO	STL	855	852	
2015	1	4	7	WN	1850	N294WN	LAS	SAN	630	627	
2015	1	4	7	WN	454	N525SW	HOU	SAT	845	842	
2015	1	5	1	WN	3214	N933WN	DAL	LBB	1005	1002	
2015	1	5	1	WN	3387	N356SW	MCI	STL	1020	1017	
2015	1	5	1	WN	2376	N269WN	MCO	BUF	1115	1112	
2015	1	5	1	WN	431	N795SW	PVD	MCO	1200	1157	

Table: flights3 (1,048,575 rows)

Update available (click to download)

Figure IV: Uploading the data.

Step 3: Create the relationships between the different tables.

Initially the three tables were related like follows:

For this case, the following relationships are created:

- airlines(IATA_CODE) with flights(AIRLINE), 1 → N (one to many)
- airports(IATA_CODE) with flights(ORIGIN_AIRPORT) 1 → N (one to many)

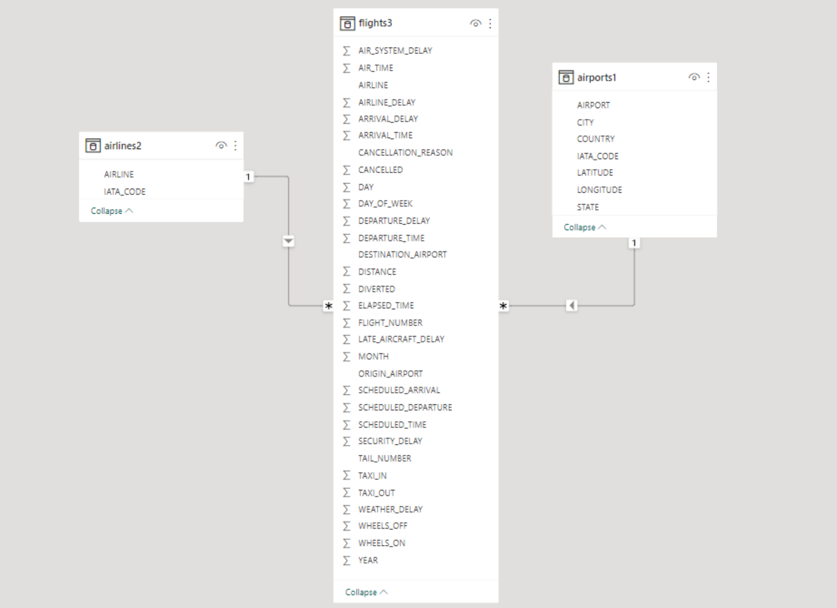
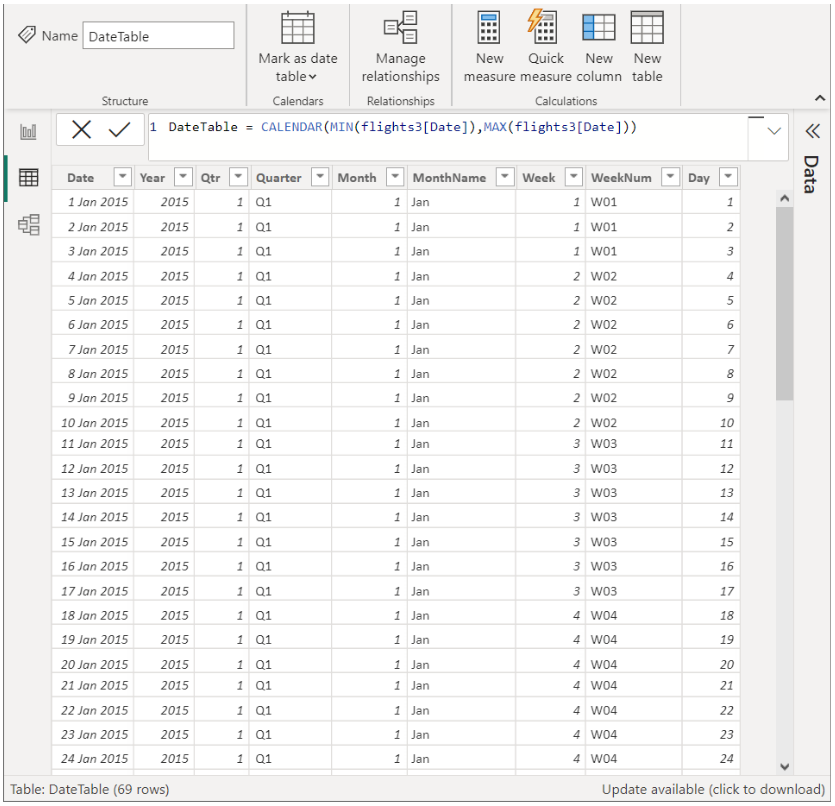


Figure V: Initial Relations Between the 3 tables

Step 4: Feature Engineering, in addition to creating useful tables.

Since the flights dataset contains a date attribute, a calendar table must be added. This calendar table will contain all information needed for a date attribute. This step utilized DAX commands in Power Bi to create the appropriate attributes. Please note that the data is recorded for first three months of 2015.



```
DateTable = CALENDAR(MIN(flights3[Date]),MAX(flights3[Date]))
```

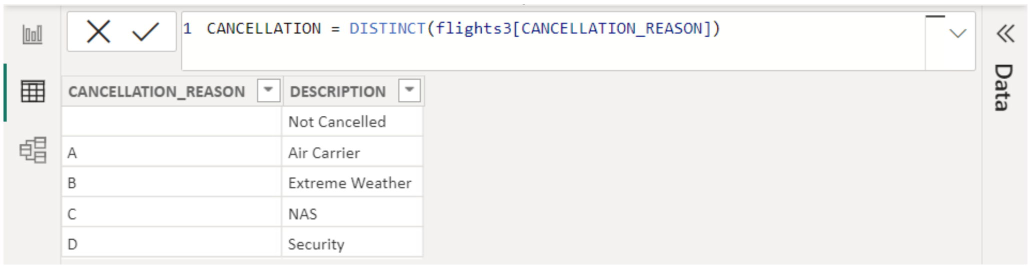
Date	Year	Qtr	Quarter	Month	MonthName	Week	WeekNum	Day
1 Jan 2015	2015	1	Q1	1	Jan	1	W01	1
2 Jan 2015	2015	1	Q1	1	Jan	1	W01	2
3 Jan 2015	2015	1	Q1	1	Jan	1	W01	3
4 Jan 2015	2015	1	Q1	1	Jan	2	W02	4
5 Jan 2015	2015	1	Q1	1	Jan	2	W02	5
6 Jan 2015	2015	1	Q1	1	Jan	2	W02	6
7 Jan 2015	2015	1	Q1	1	Jan	2	W02	7
8 Jan 2015	2015	1	Q1	1	Jan	2	W02	8
9 Jan 2015	2015	1	Q1	1	Jan	2	W02	9
10 Jan 2015	2015	1	Q1	1	Jan	2	W02	10
11 Jan 2015	2015	1	Q1	1	Jan	3	W03	11
12 Jan 2015	2015	1	Q1	1	Jan	3	W03	12
13 Jan 2015	2015	1	Q1	1	Jan	3	W03	13
14 Jan 2015	2015	1	Q1	1	Jan	3	W03	14
15 Jan 2015	2015	1	Q1	1	Jan	3	W03	15
16 Jan 2015	2015	1	Q1	1	Jan	3	W03	16
17 Jan 2015	2015	1	Q1	1	Jan	3	W03	17
18 Jan 2015	2015	1	Q1	1	Jan	4	W04	18
19 Jan 2015	2015	1	Q1	1	Jan	4	W04	19
20 Jan 2015	2015	1	Q1	1	Jan	4	W04	20
21 Jan 2015	2015	1	Q1	1	Jan	4	W04	21
22 Jan 2015	2015	1	Q1	1	Jan	4	W04	22
23 Jan 2015	2015	1	Q1	1	Jan	4	W04	23
24 Jan 2015	2015	1	Q1	1	Jan	4	W04	24

Table: DateTable (69 rows)

Update available (click to download)

Figure VI: DateTable Dataset

Another table that was added is the Cancellation table, which contains the information about why an airport would cancel a trip. This information has been obtained from the source discussion on Kaggle.

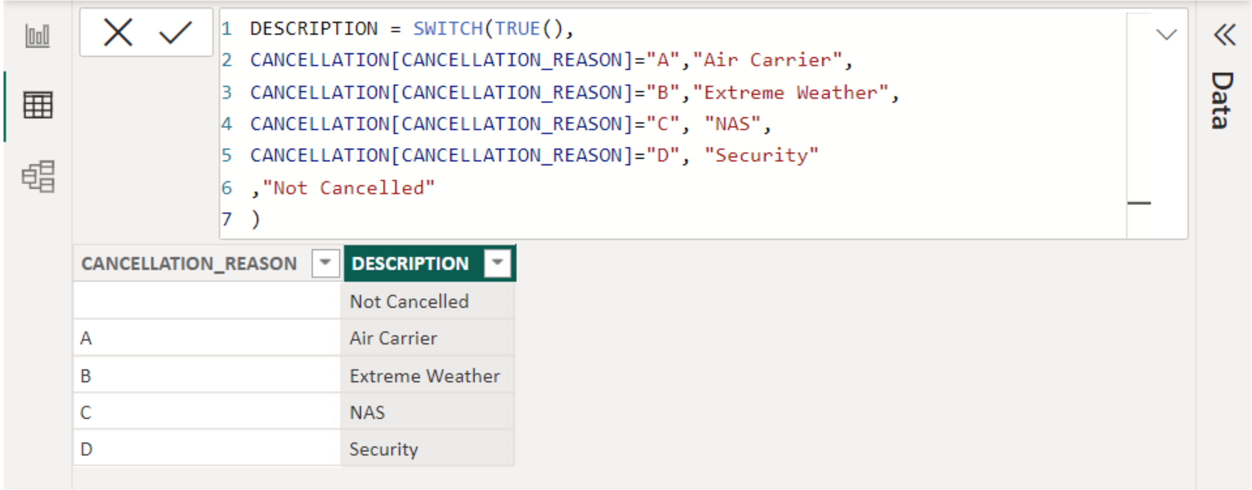


```
CANCELLATION = DISTINCT(flights3[CANCELLATION_REASON])
```

CANCELLATION_REASON	DESCRIPTION
	Not Cancelled
A	Air Carrier
B	Extreme Weather
C	NAS
D	Security

Figure VII: Cancellation dataset

The descriptions were achieved using the following command respectfully for each distinct value.



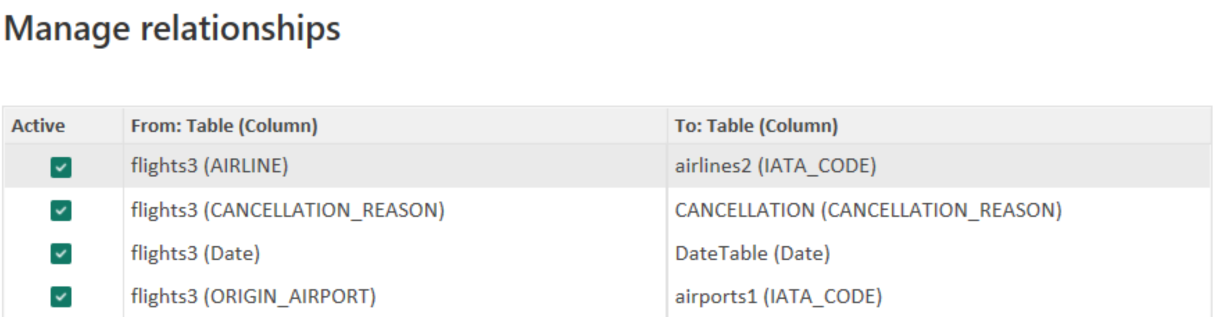
```
DESCRIPTION = SWITCH(TRUE(),
CANCELLATION[CANCELLATION_REASON]="A","Air Carrier",
CANCELLATION[CANCELLATION_REASON]="B","Extreme Weather",
CANCELLATION[CANCELLATION_REASON]="C", "NAS",
CANCELLATION[CANCELLATION_REASON]="D", "Security"
,"Not Cancelled"
)
```

CANCELLATION_REASON	DESCRIPTION
	Not Cancelled
A	Air Carrier
B	Extreme Weather
C	NAS
D	Security

Figure VIII: Example of Commands Used.

<u>Step 5:</u> Adjust the relationships between the tables:

To relate the Date table with the flights table, feature engineering has been performed. This merging the different date related attributes in flights table to one date attribute that contained all information needed (DD/MM/YYYY).



Manage relationships

Active	From: Table (Column)	To: Table (Column)
☑	flights3 (AIRLINE)	airlines2 (IATA_CODE)
☑	flights3 (CANCELLATION_REASON)	CANCELLATION (CANCELLATION_REASON)
☑	flights3 (Date)	DateTable (Date)
☑	flights3 (ORIGIN_AIRPORT)	airports1 (IATA_CODE)

Figure IX: Managing Relationships

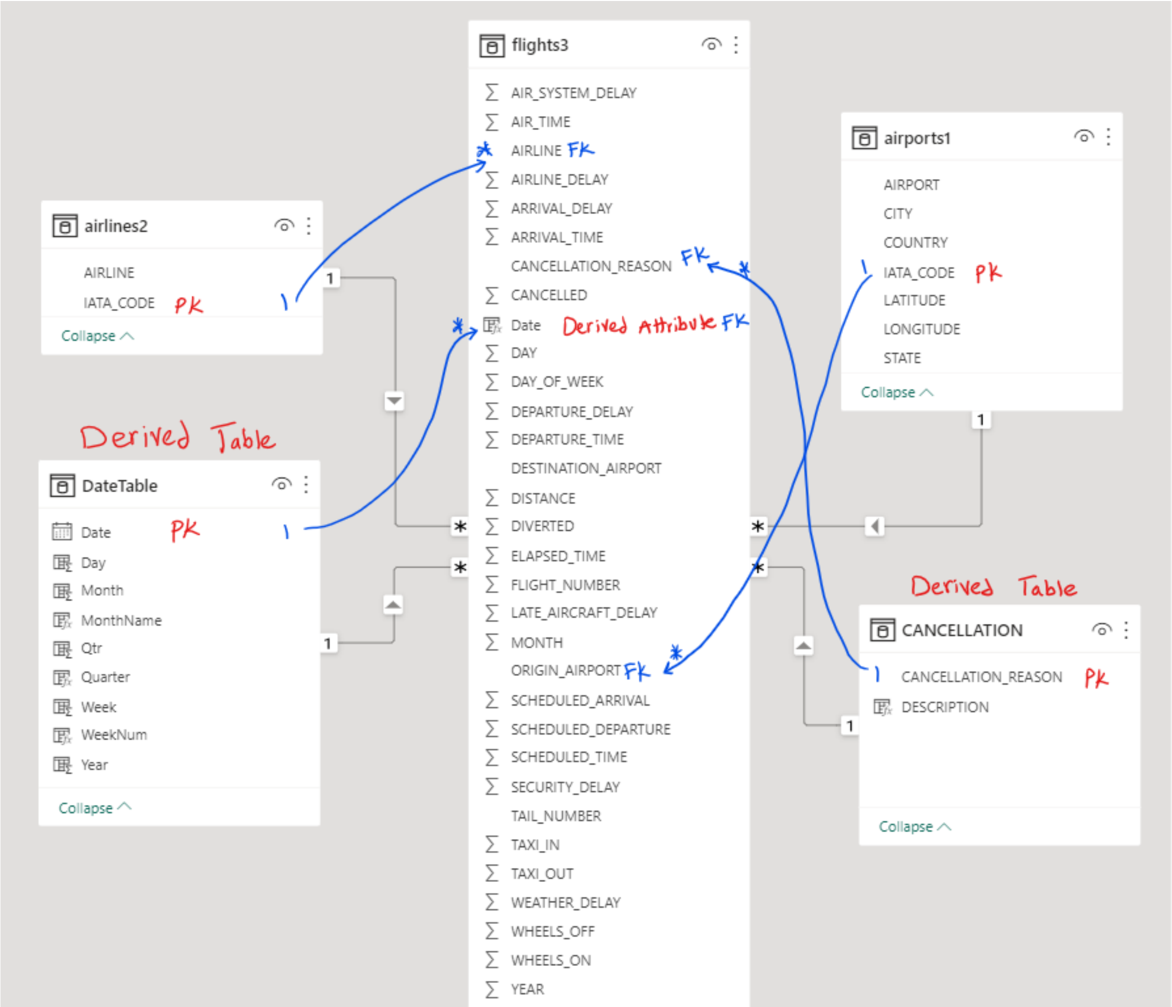
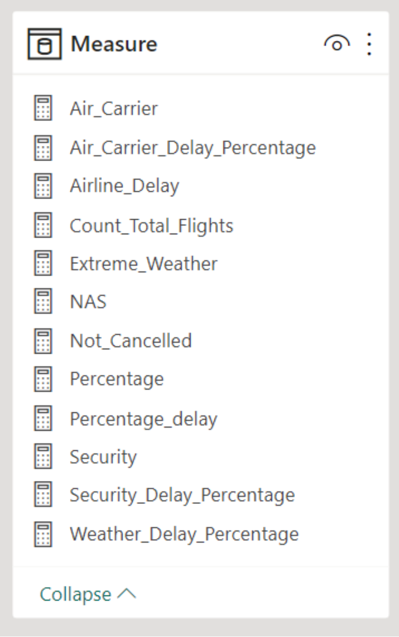


Figure X: Connections Between All Tables Achieved

Step 6: Creating the Measures.

As a professional step, a measurement table has been created after much thought of what needs to be presented in the dashboard. It acts as a centralized location for defining and organizing the key measures required to make informative visualizations based on the flight delay and cancellation data.



Measure

Air_Carrier
Air_Carrier_Delay_Percentage
Airline_Delay
Count_Total_Flights
Extreme_Weather
NAS
Not_Cancelled
Percentage
Percentage_delay
Security
Security_Delay_Percentage
Weather_Delay_Percentage

Figure XI: Measurements

iv. Dashboard design

The first dashboard is designed to visualize and analyze the cancellation and delay across different airlines, along with the reasons behind such performance. It focused on weather, NAS, air carriers, and security as potential reasons. That was achieved by employing card visuals, where each represents a value of the measures derived from the cancellation table in conjunction with the flights table.

A line and clustered bar chart have also been created, illustrating different airlines with the percentage of canceled flights as the bars and the line showing the percentage of delay in respect to the total recorded flights for each airline. The chart has been sorted from highest to lowest cancellation percentage to show which airline made the most cancellations in their trips.

To aid with investigating targeted durations, a date visual – slicer – has been incorporated, allowing users to explore the part of time they are interested in. This helps when trying to understand the data during a specific event or a global crisis. However, a line plot showing the x-axis as the date and y-axis as the arrival delay helped view the trend in delays made by the airlines over time. From that one can view the peaks and understand the data in a better manner.

In addition to that, a multi-slicer has been leveraged to show the top cities in the USA with the highest cancellations rates, plus delay in arrival. (Team, 2023)

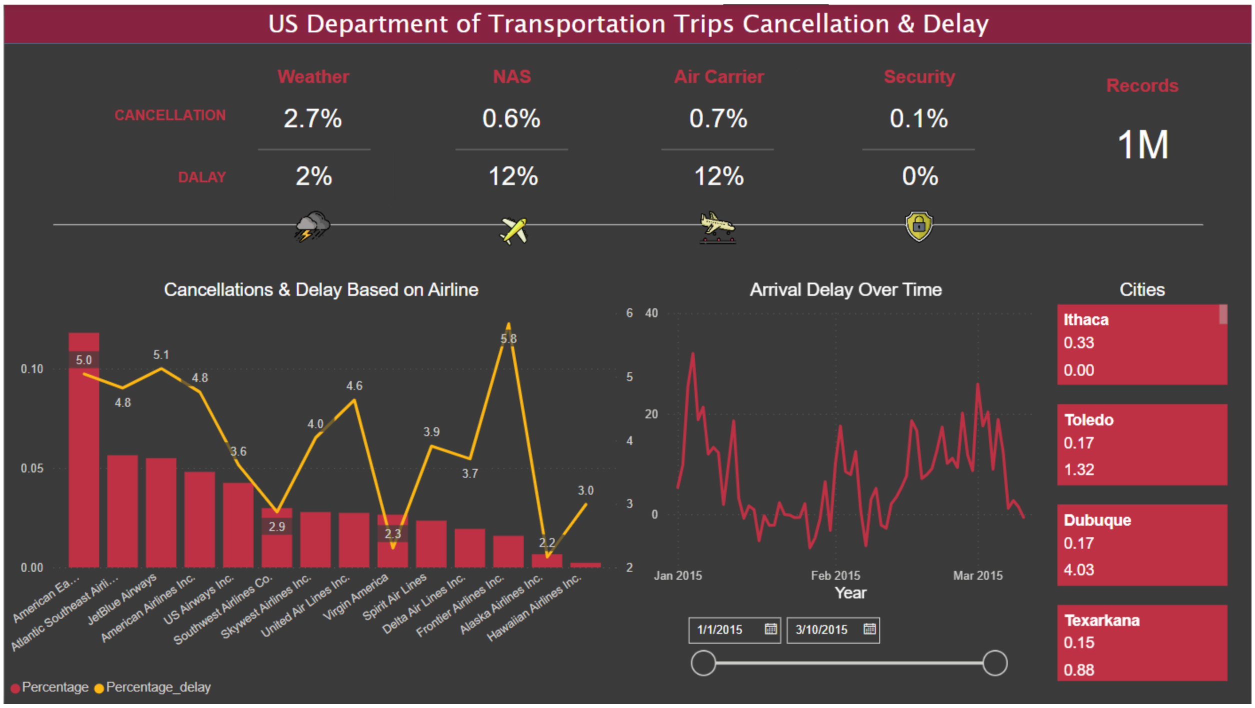


Figure XII: Dashboard

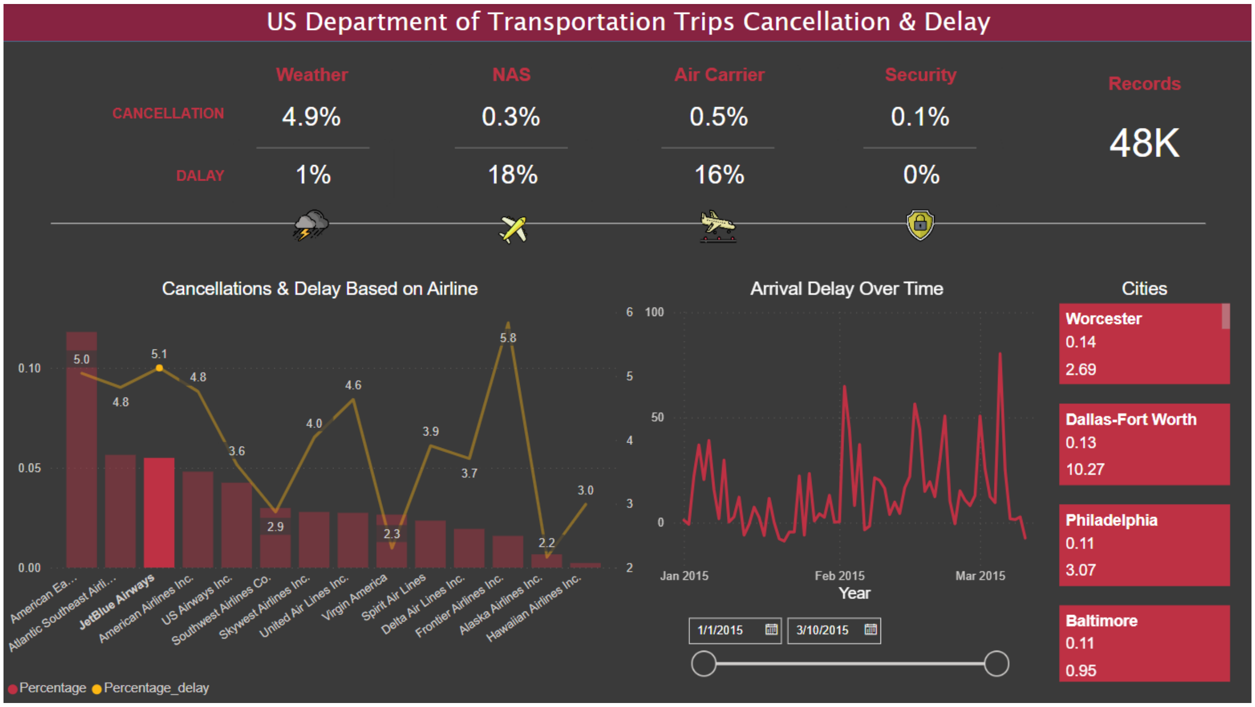


Figure XIII: Interactiveness test

v. Insights

Creating a dashboard means communicating insightful and valuable information to the team. In the above dashboard we can observe some like the following five:

- American Eagle Airlines Inc., recorded 66k instances with the highest cancellation percentage. 5% cancellation rate, mainly due to extreme weather. High delay percentage, primarily caused by air carriers and airline delays. A peak of average 70 minutes delay on January 4, 2015.
    - Solution: invest in advanced weather prediction technologies
- Delta Airlines Inc. was the best performing airline with 147k instances. Low cancellation (2%) and delay (3.7%) percentages. The delay was mostly caused by the airline itself. Peaks in delay on February 2, 2015, and February 24, 2015.
- Peak arrival delay on January 4, 2015, was mostly caused by Frontier Airlines Inc. with a 13.61% delay percentage, United Airlines Inc. with a 12.1% delay percentage, and American Airline Inc. and Atlantic Southeast Airlines with 11.25% delay.
- Most cancellations in Ithaca City – Atlantic Southeast Airlines with a 33.4% cancellation rate mainly caused by air carrier issues. Concentrated cancellation on February 18, decreasing over time but increasing by February 25.
- Most delay in Bangor – Atlantic Southeast Airlines. Delays were mostly caused by the airline itself, suggesting the need to investigate Atlantic Southeast Airlines' performance and employee-related issues.

vi. Evaluation Criteria

Table III: Evaluation Criteria

Evaluation Criteria	Description
Functionality	The dashboard created functions without any error, where all interactive features work smoothly.
Design	The dashboard is designed in a way that is visually appealing, making the information easy to understand by other.
Complexity	The data model obtained in well-structured and has clear and consistent relationships between the different tables.

Originality	The dashboard has been done without any inspiration and constructed valuable insights to help improve the performance of airlines.

Conclusion

In conclusion, and based on the 2015 dataset in hand, a Power BI dashboard, and advanced analytical models, this comprehensive analysis of flight delays and cancellations provides insight into aviation challenges. By loading data, selecting features, exploring, and handling missing data, the ETL procedure prepared a refined dataset. A framework for machine learning was created by using models such as XGBoost, LightGBM, and CatBoost. A literature review emphasized the benefits and drawbacks of the chosen models and demonstrated the researcher's comprehension of them. By using card visuals, a bar chart, date slicers, and multi-slicers, the Power BI dashboard effectively conveyed informative insights into cancellation percentages, airline performance, and peak delay periods.

Reference

2015 Flight delays and cancellations. (2017, February 9). Kaggle. https://www.kaggle.com/datasets/usdot/flight-delays/data

Sternberg, A. (2017, March 1). *(PDF) A Review on Flight Delay Prediction*. ResearchGate. https://www.researchgate.net/publication/315382748_A_Review_on_Flight_Delay_Prediction

Saedi, M. (2023, July 17). *Analyzing Kaggle's 2015 Flight Delays and Cancellations on Tableau: A Visual Exploration*. Medium. https://medium.com/@masoud_saedi/analyzing-kaggles-2015-flight-delays-and-cancellations-on-tableau-a-visual-exploration-ef748daf90df

Fadl, M., Ali, M., Rani, J., Musaad, F. (2023). 2015_flight_delay [Jupyter Notebook]. Project Team.

Ali, M., Fadl, M., Rani, J., Musaad, F. (2023). US Department of Transportation Trips Cancellation & Delay [Power BI File]. Project Team.