BAIS: 3200

Chicago O'Hare Airport Project

Isaac McSorley, Tim Cook, Josh Orth, Cade Hanrahan, Cade Keenan

Introduction

Flying is something that millions of people do every day. For some people, it is a hectic and nerve-wracking experience to get through the airport while hoping your flight will not be delayed or canceled. We want to make this experience a little easier for people by providing some data showing statistics to help calm people's nerves about flying. After looking at our data, you will be able to choose an airline and day that is statistically better to fly on, which means there is a higher chance that your flight will not be delayed or canceled and that everything will go smoothly. We will focus on the Chicago O'Hare International airport in Chicago, IL because that is one of the largest airports in the country, and the biggest airport in the Midwest. Many people from the University of Iowa use this airport and we are hoping to provide a resource to aid in deciding what the ideal travel plans are when flying through this airport.

Data

https://www.kaggle.com/gauravmehta13/airline-flight-delays?select=flights.csv

This project uses data from 2015 Kaggle data gathered on commercial airflights. Each record represents a single flight, including the airline name, flight number, origin/destination airport and flight distance, as well as scheduled/actual departure and arrival times. The original data set has over five million entries of data. We plan to reduce the data set down that makes the most sense for our questions to be answered.

*We have chosen to reduce the number of data entries by focusing on a single airport, Chicago O' Hare, we will find the optimal days, months, and airlines to use at this airport.

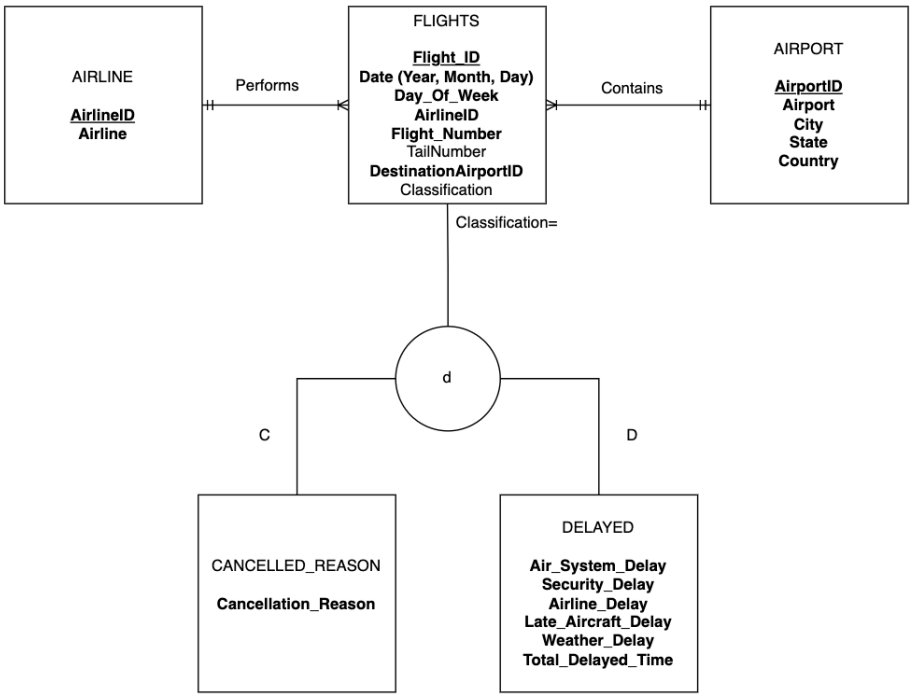
Data Dictionary

ATTRIBUTE	TYPE	DESCRIPTION
Airline ID	Binary	Unique ID for each airline
Airline	Text	Name of each airline
Airport ID	Text	Unique ID for each airport
Airport	Text	Name of each airport
City	Text	City of airport
State	Text	State of airport
Country	Text	Country of airport
FlightID	Numeric	Unique ID for each flight
Cancelation_Reason	Text	Reason why flight was canceled

Air_System_Delay	Numeric	Delay caused by air systems issues
Security_Delay	Numeric	Delay caused by security issues
Airline_Delay	Numeric	Delay caused by specific airline
Late_Aircraft_Delay	Numeric	Delay caused by late aircraft
Weather_Delay	Numeric	Delay caused by weather
TotalDelay	Numeric	Total amount of time flight was delayed
Date_	Date	Date the flight occurred
Day_Of_Week	Numeric	Day of week the flight occurred
Flight_Number	Numeric	The number associated with a given flight (not unique)
Tail_Number	Binary	Unique ID for airplane
Destination_Airport	Text	The airport that is the destination of the flight
Canceled	Numeric	1 if flight was canceled, 0 if not

The supertype entity in this database is FLIGHTS, which is identified by FLIGHT_ID. But FLIGHT is also a child to the AIRPORT table which contains that the full names of each airport that are listed, and the AIRLINE table contains the full names of every airline that is listed in FLIGHTS referencing DestinationAirportID and AirlineID. CANCELED_REASON and DELAYED tables are subtypes of FLIGHTS because only when a flight is canceled or delayed, there will be data for the attributes under CANCELED_REASON and DELAYED. This relationship has the disjointness constraint because a flight cannot be both delayed and canceled, if a flight gets delayed and then canceled, it is only marked as canceled. These subtypes also have partial specialization, because every flight can be either canceled, delayed, or neither (flight was on time). Every flight is fit with its unique flightid, and every FLIGHT must go to one and only one AIRPORT, but each airport can have one or many flights come to it. Similarly with AIRLINE, every flight is performed by one airline at a time, but each AIRLINE can perform one or many flights.

ERD Diagram



Graphical Relational Schema

Based on the ERD diagram above, we have created a Graphical Schema below (Figure 35) with 5 tables. The AIRPORT and AIRLINE are both parent tables to FLIGHTS, and FLIGHTS is the supertype to CANCELED_REASON and DELAYED.

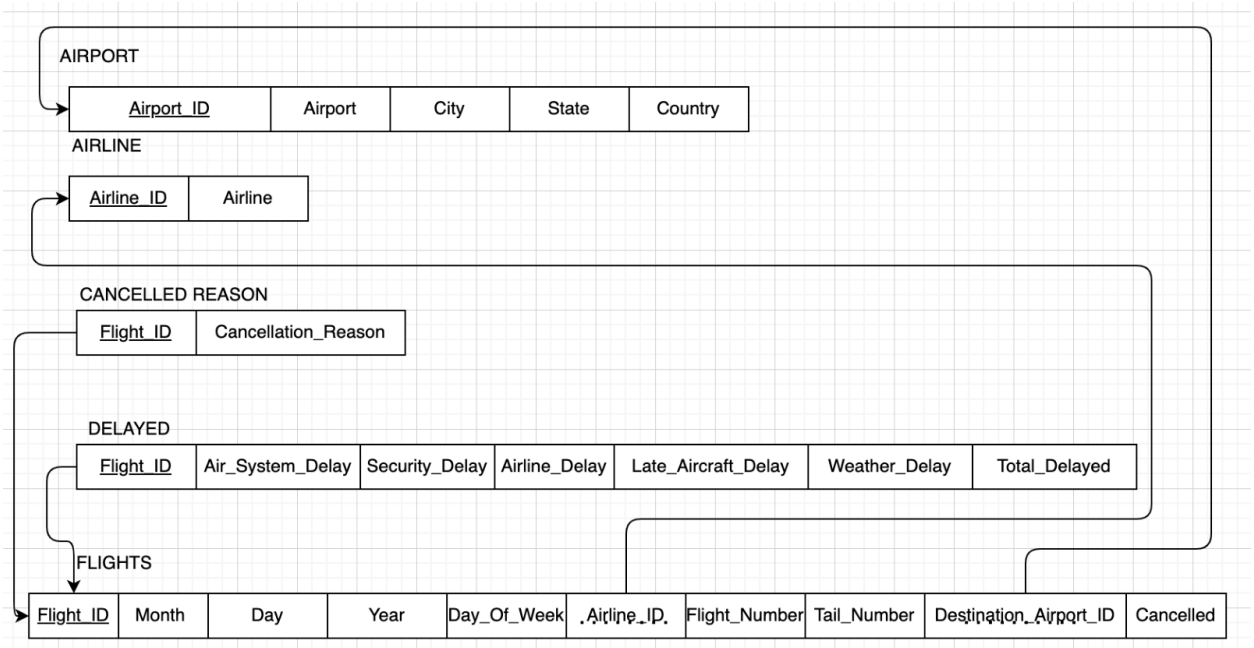


Figure 35: Graphical Relational Schema

Database Implementation

To create the database in Oracle Apex, we wrote the following CREATE TABLE commands

AIRLINE: Because AIRLINE is the parent of FLIGHTS (AirlineID), we will create this table first

```
CREATE TABLE AIRLINE (
AIRLINEID CHAR(2) NOT NULL,
AIRLINE VARCHAR(30),
CONSTRAINT AIRLINE_PK PRIMARY KEY (AIRLINEID));
INSERT INTO AIRLINE VALUES ('UA', 'United Airlines Inc.');
INSERT INTO AIRLINE VALUES ('AA', 'American Airlines Inc.');
INSERT INTO AIRLINE VALUES ('US', 'US Airways Inc.');
INSERT INTO AIRLINE VALUES ('F9', 'Frontier Airlines Inc.');
```

AIRPORT: AirportID is also a FK in FLIGHTS to the AIRPORT table, so we will create that table before FLIGHTS as well.

```
CREATE TABLE AIRPORT (
AIRPORTID CHAR(3) NOT NULL,
AIRPORT VARCHAR(50),
CITY VARCHAR(25),
STATE CHAR(2),
COUNTRY CHAR(3),
CONSTRAINT AIRPORT_PK PRIMARY KEY (AIRPORTID));
INSERT INTO AIRPORT VALUES ('ABE', 'Lehigh Valley International Airport', 'Allentown', 'PA', 'USA');
INSERT INTO AIRPORT VALUES ('ABI', 'Abilene Regional Airport', 'Abilene', 'TX', 'USA');
INSERT INTO AIRPORT VALUES ('ABQ', Albuquerque International Support', 'Albuquerque', 'NM', 'USA');
INSERT INTO AIRPORT VALUES ('ABR', 'Aberdeen Regional Support', 'Aberdeen', 'SD', 'USA');

CREATE TABLE FLIGHTS (
FLIGHT_ID CHAR(4) NOT NULL,
DATE DATE,
DAY_OF_WEEK NUMBER(1,0),
AIRLINEID CHAR(2),
FLIGHT NUMBER NUMBER(6,0),
TAIL_NUMBER CHAR(6) NOT NULL,
DESTINATIONAIRPORTID CHAR(3),
```

```
CANCELED NUMBER(1,0),
CONSTRAINT FLIGHTS_PK PRIMARY KEY (FLIGHT_ID),
CONSTRAINT FLIGHTS_FK FOREIGN KEY (AIRLINEID) REFERENCES AIRLINE (AIRLINEID),
CONSTRAINT FLIGHTS_FK2 FOREIGN KEY (DESTINATIONAIRPORTID) REFERENCES AIRPORT (AIRPORTID));
INSERT INTO FLIGHTS ('FLIGHT_ID', 'DATE', 'DAY_OF_WEEK', 'AIRLINEID', 'FLIGHT_NUMBER', 'DESTINATIONAIRPORTID',
'CANCELED') VALUES ('F156', '02/02/2015', 'AA', '40', '', 'DCA', '1');
INSERT INTO FLIGHTS VALUES ('F157', '03/05/2015', '4', 'AA', '40', 'N3APAA', 'DCA', '1');
INSERT INTO FLIGHTS VALUES ('F158', '01/02/2015', '5, 'AA', '40', 'N3AUAA', 'DCA', '0');
INSERT INTO FLIGHTS VALUES ('F159', '01/03/2015', '6', 'AA', '40', 'N3KMAA', 'DCA', '0');

CREATE TABLE CANCELED_REASON (
FLIGHT_ID CHAR(4) NOT NULL,
CANCELATION_REASON VARCHAR(25),
CONSTRAINT CANCELED_REASON_PK PRIMARY KEY (FLIGHT_ID),
CONSTRAINT CANCELED_REASON_FK FOREIGN KEY (FLIGHT_ID) REFERENCES FLIGHTS (FLIGHT_ID),
);
INSERT INTO CANCELED_REASON VALUES ('F155', 'Weather');
INSERT INTO CANCELED_REASON VALUES ('F156', 'Weather');
INSERT INTO CANCELED_REASON VALUES ('F157', 'Weather');
INSERT INTO CANCELED_REASON VALUES ('F158', 'Weather');

CREATE TABLE DELAYED (
FLIGHT_ID CHAR(4) NOT NULL,
AIR_SYSTEM_DELAY NUMBER(3,0),
SECURITY_DELAY NUMBER(3,0),
AIRLINE_DELAY NUMBER(3,0),
LATE_AIRCRAFT_DELAY NUMBER(3,0),
WEATHER_DELAY NUMBER(3,0),
TOTAL_DELAY NUMBER(4,0),
CONSTRAINT DELAYED_PK PRIMARY KEY (FLIGHT_ID),
CONSTRAINT DELAYED_FK FOREIGN KEY (FLIGHT_ID) REFERENCES FLIGHTS (FLIGHT_ID)
);
INSERT INTO DELAYED VALUES ('F2453', '0', '0', '6', '55', '0', '61');
INSERT INTO DELAYED VALUES ('F2457', '1', '0', '6', '56', '0', '63');
INSERT INTO DELAYED VALUES ('F2462', '13', '0', '0', '75', '143', '231');
INSERT INTO DELAYED VALUES ('F2463', '0', '0', '0', '32', '39', '71');
```

Analysis

Our Analysis is intended to help future Chicago O'Hare fliers make better choices on when to fly, what airline to fly through, and ultimately choose the best possible options that satisfy their flying needs.

Question 1: In the Chicago O'Hare airport in 2015, are certain days of the week better/worse to fly on? What days had the highest percentage of cancelations?

To answer this question we counted the number of total flights on each day, divided by the number of flights that were canceled grouped by the day of the week. After finishing this problem, we rank the days by the highest percentage of cancelations (worst day to fly).

```
SELECT Day, ROUND(COUNT(FLIGHTS.FLIGHT_ID)/Number_of_Flights,3)*100 || '%' AS PercentageCanceled

FROM FLIGHTS JOIN (SELECT Day_of_week,

CASE

WHEN DAY_OF_WEEK = '1' THEN 'Monday'

WHEN DAY_OF_WEEK = '2' THEN 'Tuesday'

WHEN DAY_OF_WEEK = '3' THEN 'Wednesday'

WHEN DAY_OF_WEEK = '4' THEN 'Thursday'

WHEN DAY_OF_WEEK = '5' THEN 'Friday'

WHEN DAY_OF_WEEK = '6' THEN 'Saturday'

WHEN DAY_OF_WEEK = '7' THEN 'Sunday'

END AS Day,

COUNT(FLIGHT_ID) AS Number_of_Flights

FROM FLIGHTS GROUP BY DAY_OF_WEEK, CASE

WHEN DAY_OF_WEEK = '1' THEN 'Monday'

WHEN DAY_OF_WEEK = '2' THEN 'Tuesday'

WHEN DAY_OF_WEEK = '3' THEN 'Wednesday'

WHEN DAY_OF_WEEK = '4' THEN 'Thursday'

WHEN DAY_OF_WEEK = '5' THEN 'Friday'

WHEN DAY_OF_WEEK = '6' THEN 'Saturday'

WHEN DAY_OF_WEEK = '7' THEN 'Sunday' END) T ON T.Day_of_week = FLIGHTS.day_of_week

WHERE CANCELED = '1'

GROUP BY Day, Number_of_flights

ORDER BY COUNT(FLIGHTS.FLIGHT_ID)/Number_of_Flights DESC
```

The results of the query are shown below in (Figure 1). This indicates that the *worst* day to fly on if you are trying to avoid a canceled flight in Chicago O'Hare in 2015 was Sunday, followed by Monday. The day with the smallest chance of getting canceled was Friday. We also showed this via bar graph visualization (Figure 13 below). We put it in descending order to match the query results.

Day	Percentage Cancelled
Sunday	12.7%
Monday	9.6%
Tuesday	6.1%
Thursday	5.4%
Wednesday	3.5%
Saturday	2.6%
Friday	2.4%

Fig 1: % cancelations by day of week

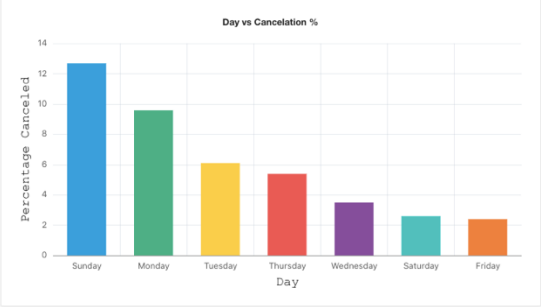


Fig 13: Bar chart of % cancelations by day of week

Question 2: In 2015 Chicago O'Hare airport, what airline has the highest percentage of cancelations?

We used another table subquery similar to question 1, where we divide the total flights grouped by each airline by the number of flights canceled grouped by each airline. We then ordered it by the highest percentage (worst airline for canceled flights) first.

```
SELECT AIRLINE.Airline, FLIGHTS.AIRLINEID, ROUND(COUNT(CANCELED)/TotalFlights,3)*100 ||'%' AS Percentage_Canceled
FROM FLIGHTS JOIN (SELECT AIRLINEID, COUNT(AIRLINEID) AS TotalFlights FROM FLIGHTS GROUP BY AIRLINEID) T ON
T.AirlineID = FLIGHTS.AirlineID

JOIN AIRLINE ON AIRLINE.AIRLINEID = FLIGHTS.AIRLINEID

WHERE CANCELED = '1' GROUP BY FLIGHTS.AIRLINEID, TotalFlights, AIRLINE.Airline

ORDER BY ROUND(COUNT(CANCELED)/TotalFlights,3)*100 DESC
```

In the results of this query below (Figure 2), you can see that JetBlue Airways had the highest percentage of canceled flights, and Alaska Airlines had the smallest percentage of cancelations. We included a bar chart to represent the data in a more user-friendly way (see below figure 18)

Airline ID	Airline	Canceled %
B6	JetBlue Airways	7.2%
OO	Skywest Airlines Inc.	6%
EV	Atlantic Southeast Airlines	5.3%
US	US Airways Inc.	5.2%
AA	American Airlines Inc.	4.2%
UA	United Air Lines Inc.	3.7%
NK	Spirit Air Lines	2.6%
F9	Frontier Airlines Inc.	2%
MQ	American Eagle Airlines Inc.	11.6%
VX	Virgin America	1.2%
DL	Delta Air Lines Inc.	.9%
AS	Alaska Airlines Inc.	.5%

Fig. 2: % cancelations by airline

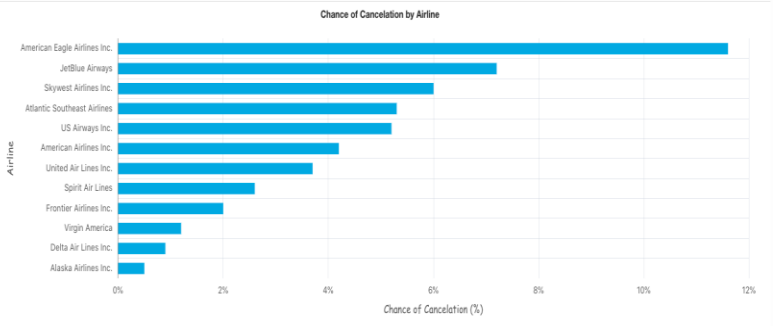


Fig. 18: Bar chart % cancelations by airline

Question 3: What are the different reasons for flight delays? If there is a delay, how likely are each reason to be the cause of the delay? In order to answer this question, we wrote a query that calculates the percentage of delayed flights

that each reason is responsible for. The code below divides the number of flights with each type of delay by the number of delayed flights in total.

```
SELECT ROUND (
(SELECT COUNT(Flight_ID) FROM Delayed
WHERE Air_System_Delay <> '0') /
(SELECT COUNT(Flight_ID) FROM Delayed),4) * 100 || '%' AS Air_System_Delay,
ROUND (
(SELECT COUNT(Flight_ID) FROM Delayed
WHERE Security_Delay <> '0') /
(SELECT COUNT(Flight_ID) FROM Delayed),4) * 100 || '%' AS Security_Delay,
ROUND (
(SELECT COUNT(Flight_ID) FROM Delayed
WHERE Airline_Delay <> '0') /
(SELECT COUNT(Flight_ID) FROM Delayed),4) * 100 || '%' AS Airline_Delay,
ROUND (
(SELECT COUNT(Flight_ID) FROM Delayed
WHERE Late_Aircraft_Delay <> '0') /
(SELECT COUNT(Flight_ID) FROM Delayed),4) * 100 || '%' AS Late_Aircraft_Delay,
ROUND (
(SELECT COUNT(Flight_ID) FROM Delayed
WHERE Weather_Delay <> '0') /
(SELECT COUNT(Flight_ID) FROM Delayed),4) * 100 || '%' AS Weather_Delay
FROM Delayed
FETCH FIRST ROW ONLY
```

The results of this query are shown in the table below (Figure 3). Each number represents a percentage of delayed flights that each reason is responsible for. You can see that the reason with the highest percentage of delays was air system delay, and the reason with the lowest percentage of delays was security delay.

Air System Delay	Security Delay	Airline Delay	Late Aircraft Delay	Weather Delay
57.48%	.09%	49.1%	49.06%	26.67%

Figure 3: % of delayed flights that each reason is responsible for

Question 4: What airlines had the most flights out of O'Hare in 2015? Return the top 5 airlines with the most flights.

For this question, we used a join query to join AIRINE and FLIGHTS, and counted the FLIGHT_ID grouped by airlines. We then ordered by the count in descending order, and fetched the top 5 so we can focus on the airlines with the most flights.

SELECT AIRLINE.AIRLINE, COUNT (FLIGHTS.FLIGHT_ID) AS TotalFlights FROM FLIGHTS

JOIN AIRLINE ON AIRLINE.AIRLINEID = FLIGHTS.AIRLINEID

GROUP BY AIRLINE.AIRLINE

ORDER BY COUNT (FLIGHTS.FLIGHT_ID) DESC

FETCH FIRST 5 ROWS ONLY

The results of this query are shown below (Figure 4). As you can see, American Eagle Airlines is had the most flights out of Chicago O' Hare in 2015.

Airline	Total Flights
American Eagle Airlines Inc.	12353
United Air Lines Inc.	10192
Atlantic Southeast Airlines	8863
American Airlines Inc.	8745
Skywest Airlines Inc.	7012

Download

Figure 4: Total flights by airline

Going off of question 4, to represent the percentage of flights out of the total each airline had out of Chicago O'Hare, we also conducted this query below and represented it as a pie chart in Apex.

SELECT AIRLINE, ROUND(COUNT(FLIGHT_ID)/(SELECT COUNT(FLIGHT_ID) AS total FROM FLIGHTS),3) AS percentage_of_total FROM FLIGHTS JOIN AIRLINE ON AIRLINE.AIRLINEID= FLIGHTS.AIRLINEID GROUP BY AIRLINE

As you can see below (Figure 5), this is what the results we got back, we did not include this on our web page, but we did include the pie chart below (figure 8) from this query to better demonstrate how much of the pie each airline takes up in terms of total flights.

AIRLINE	PERCENTAGE_OF_TOTAL
United Air Lines Inc.	0.192
JetBlue Airways	0.007
Virgin America	0.003
US Airways Inc.	0.026
Skywest Airlines Inc.	0.132
Alaska Airlines Inc.	0.004
Spirit Air Lines	0.033
Delta Air Lines Inc.	0.024
Frontier Airlines Inc.	0.012
American Eagle Airlines Inc.	0.233
American Airlines Inc.	0.165
Atlantic Southeast Airlines	0.167

Fig. 5: % of total flights grouped by airline.

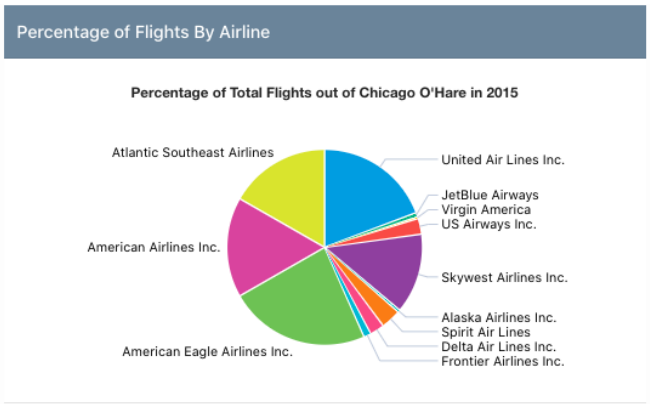


Fig. 8: % of total flights by airline

Question 5: In 2015 Chicago O'Hare airport, what is the percentage chance that your flight got canceled, and what is the percentage chance that your flight got delayed?

We used a subquery to find the percentage of delayed flights divided by the total flights as well as the percentage of canceled flights divided by the total flights

SELECT (SELECT ROUND((SELECT COUNT(flight_id) * 100 FROM Flights WHERE Canceled = '1') / (SELECT COUNT(Flight_ID) FROM Flights),3) || '%' FROM Flights FETCH FIRST ROW ONLY) AS PercentageCanceled,

(SELECT ROUND((SELECT COUNT(Delayed.Flight_ID) * 100

FROM Flights JOIN Delayed ON Flights.Flight_ID = Delayed.Flight_ID) /

(SELECT COUNT(Flight_ID)

```
FROM Flights),2) || '%'

FROM Flights

FETCH FIRST ROW ONLY  ) AS PercentageDelayed

FROM FLIGHTS

FETCH FIRST ROW ONLY
```

The results are shown below (Figure 6). At O'Hare airport in 2015, 6.12% of flights got canceled, and 32.38% of flights got delayed. This does not include how long the delay averages were, but our question #6 helps answer that question.

Percentage Cancelled	Percentage Delayed
6.12%	32.38%

Figure 6: Chance of cancelation/delay

Question 6: **What is the average delay time for each airline?**

For this query we used a function to average the total delay time and grouped it by airlines in descending order.

```
SELECT AIRLINE.AIRLINE, ROUND(AVG(TOTALDELAY)) AS AVERAGEDELAY

FROM DELAYED JOIN FLIGHTS ON FLIGHTS.FLIGHT_ID = DELAYED.FLIGHT_ID JOIN  AIRLINE

ON AIRLINE.AIRLINEID = FLIGHTS.AIRLINEID GROUP BY AIRLINE.AIRLINE

ORDER BY AVERAGEDELAY DESC
```

As you can see below (Figure 7), *When there is a delay* the averages times are as followed: The longest average delay is from **Frontier Airlines** with 17 minutes longer on average than any other airline.

Airline	Average Delay Time (minutes)
Frontier Airlines Inc.	87
Delta Air Lines Inc.	70
American Eagle Airlines Inc.	67
JetBlue Airways	67
Spirit Air Lines	67
Alaska Airlines Inc.	64
Skywest Airlines Inc.	63
United Air Lines Inc.	63
American Airlines Inc.	58
Atlantic Southeast Airlines	58
US Airways Inc.	55
Virgin America	54

Figure 7: Average delay time per airline

We also showed this query as a bar chart below (Figure 9). This makes it easier for users to see that Frontier Airlines was quite a bit higher than other airlines, and that could factor in to the users decision making when choosing which airline to fly from.

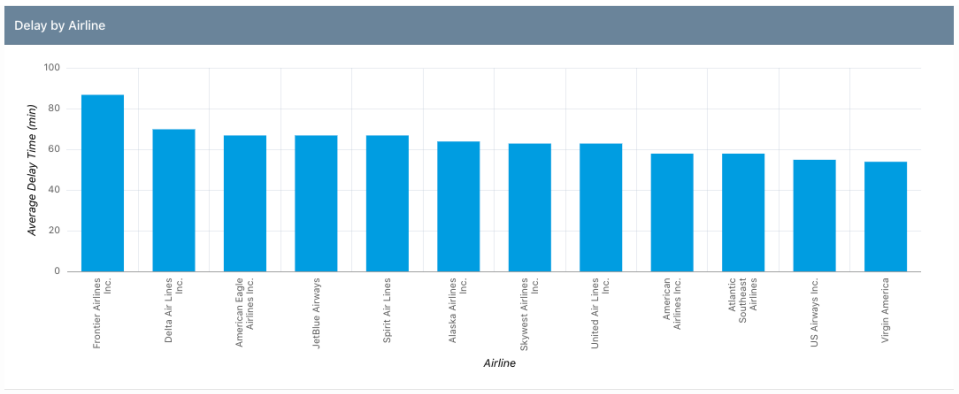


Fig. 9: Bar chart of Average delay in minutes

Web Design

Link: [Chicago O'Hare Project Link]

Home Page:

Our homepage features an overview of our project's mission (with a hyperlink and description of the original data source), the questions we answer throughout the project, a creative common license's image of O'Hare International Airport (link), and a navigation menu that features custom icons for each page(fa-paper-plane). We decided that using a consistent color scheme (accent 8) throughout the project, along with graphs and images would improve the readability of the project and peak the user's attention. Figures 10 and 11 below show screenshots of the homepage.

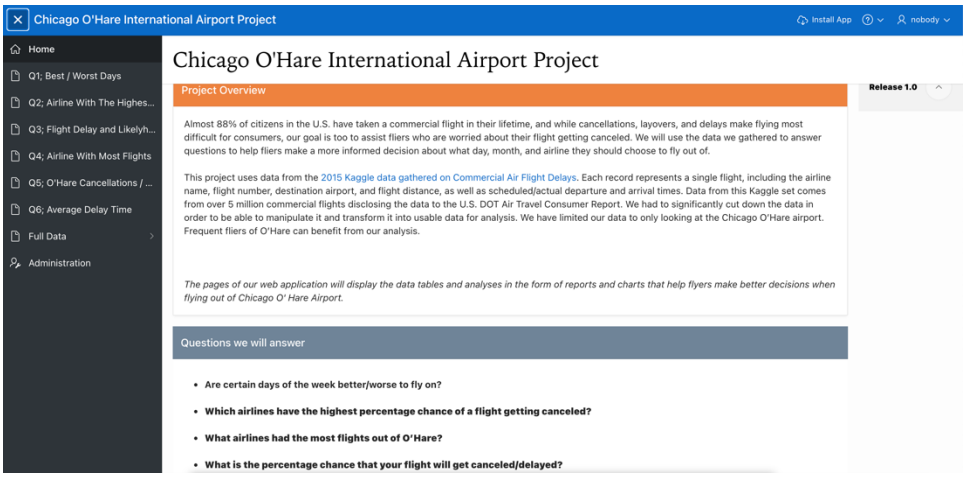


Figure 10: Homepage

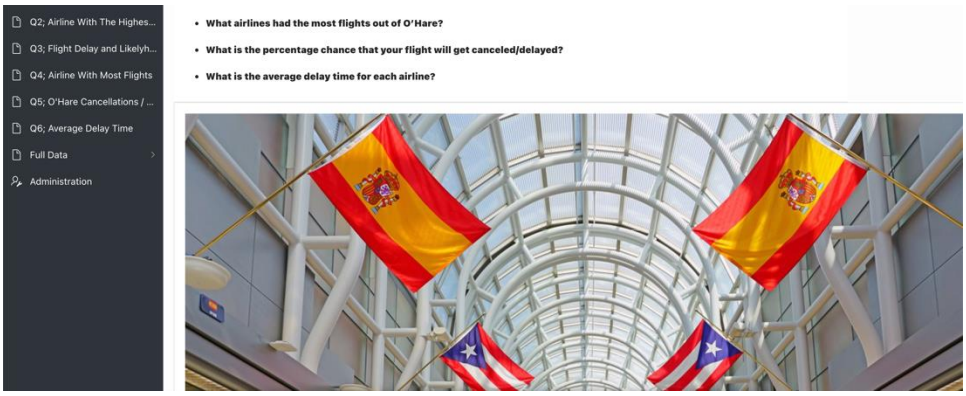


Figure 11: Homepage

Tables

We created a report for each of the database's tables that have blue column headings to increase the user's readability. Each table has the name of the table in a title heading and a description of the table below it in bold. The data also contains a list of actions that allow you to search, filter, format, graph, and group the data. The tables are shown below in figures 12-16.

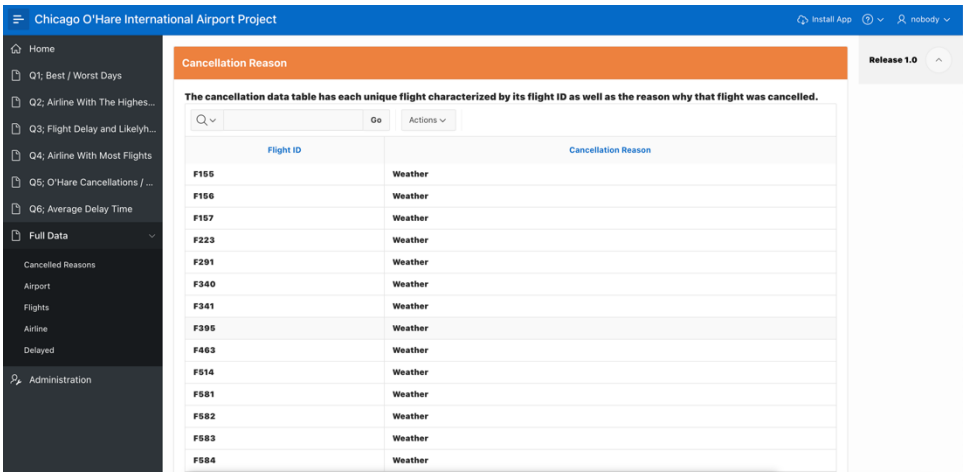


Figure 12: Canceled Reasons

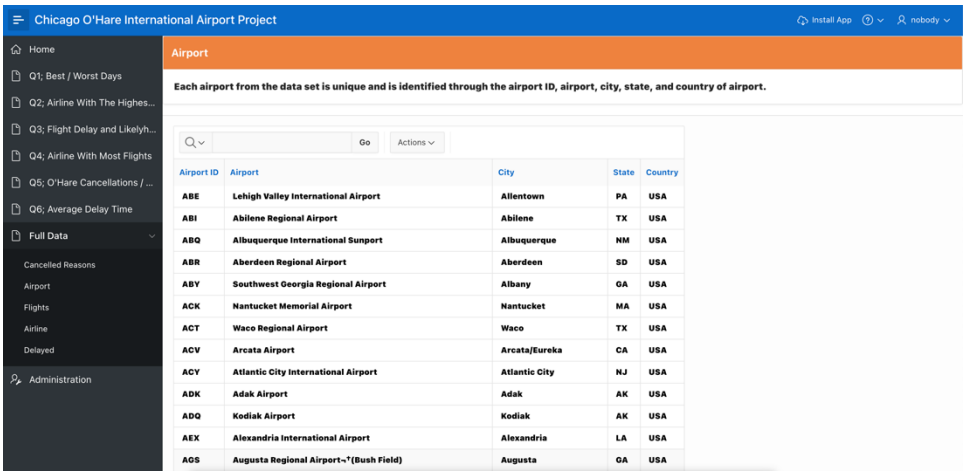


Figure 13: Airport

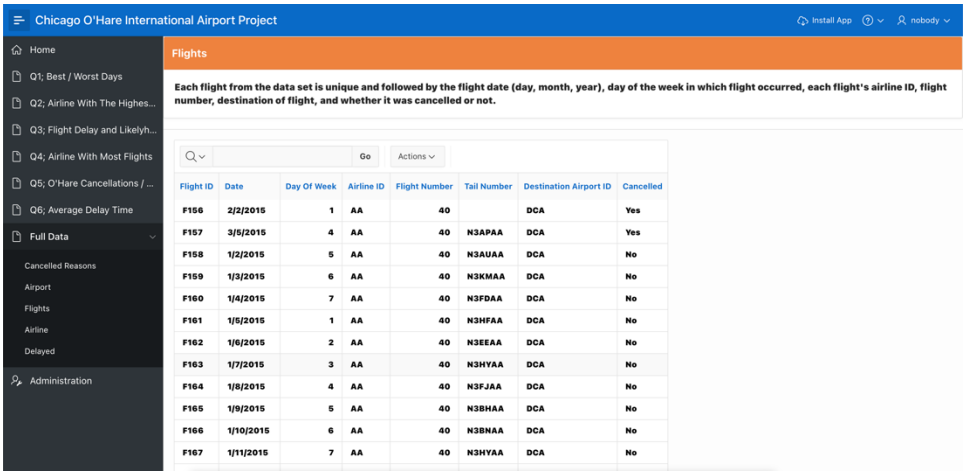


Figure 14: Flights

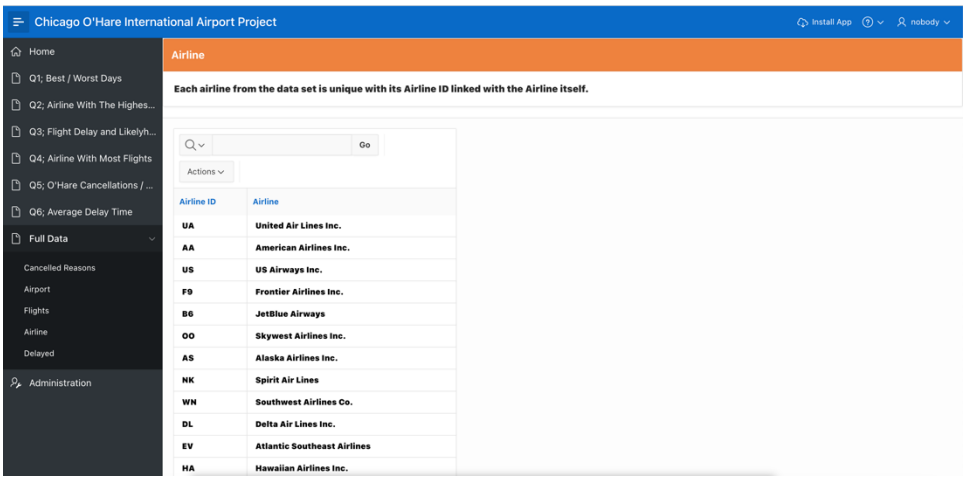


Figure 15: Airline

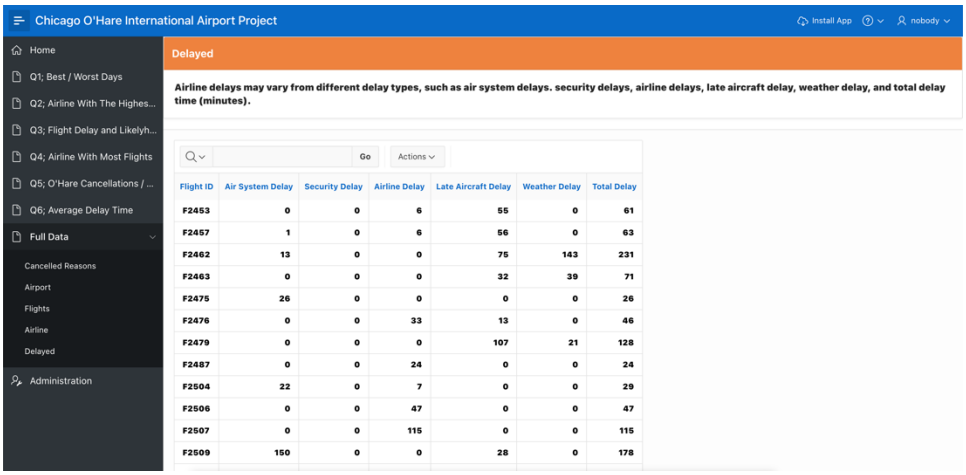


Figure 16: Delayed

Queries

We presented our first research question that focuses on what days of the week have the most cancelations in the form of a classic report in figure 17. In all of our queries we provide the question in large white text and an orange background with our answer directly below it in its own text box. We put the word "**Answer"** in bold and added a break (</br>) so that it is easily distinguishable that it is the answer to the question listed above. For this query, we presented the information in a bar chart to help visualize the data with the day of the week on the x-axis and the percentage of cancelations on the y-axis. We grouped the bar chart in descending order so that it is consistent with the query results. We also made every bar a different color so that it is the easiest to tell the difference between.

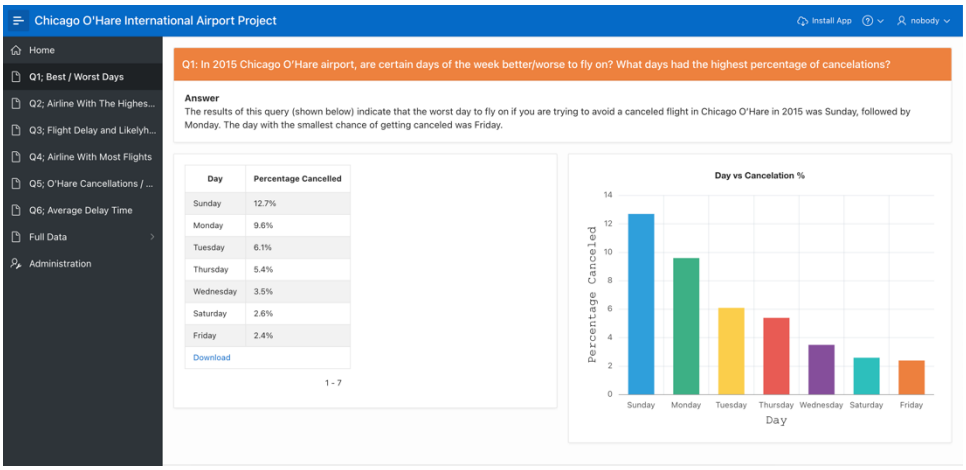


Figure 17: Best / Worst Days

Our second research question focuses on which airlines have the highest chance of having a flight get canceled (Figure 30). We found our answer by taking the number of cancelations each airline had divided by the total number of flights ran through that specific airline. We again went with the same consistent theme as the page before. We chose to represent this query as a bar chart. We made this bar chart horizontal to better fit the page, we

also added labels and sorted it in descending order, so it is easily visible to see the airline with the highest cancelation percentage (American Eagle), and lowest percentage(Alaska Airlines).

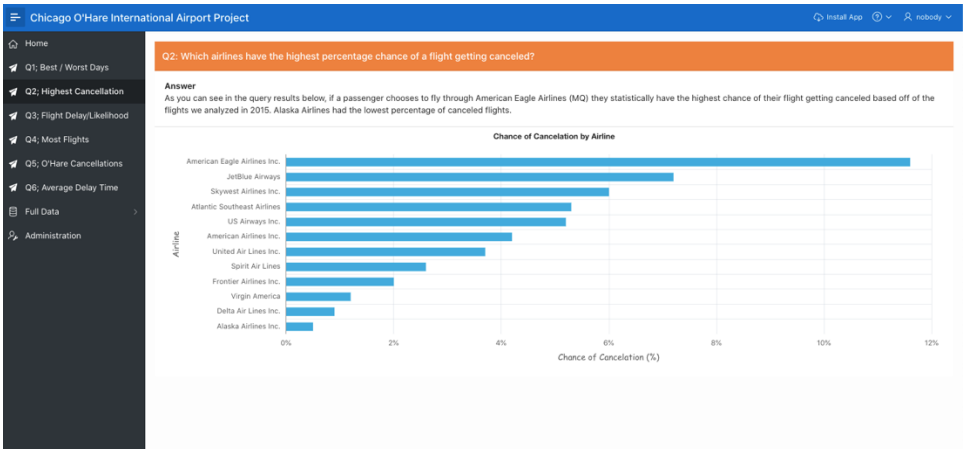


Fig. 30: Q2 page

Our third research question goes into the details of several types of cancelations. We present our data in a classic report and provide a text box with the answer to our questions and why we believe the data supports our answer. We show all of this below in figure 19.

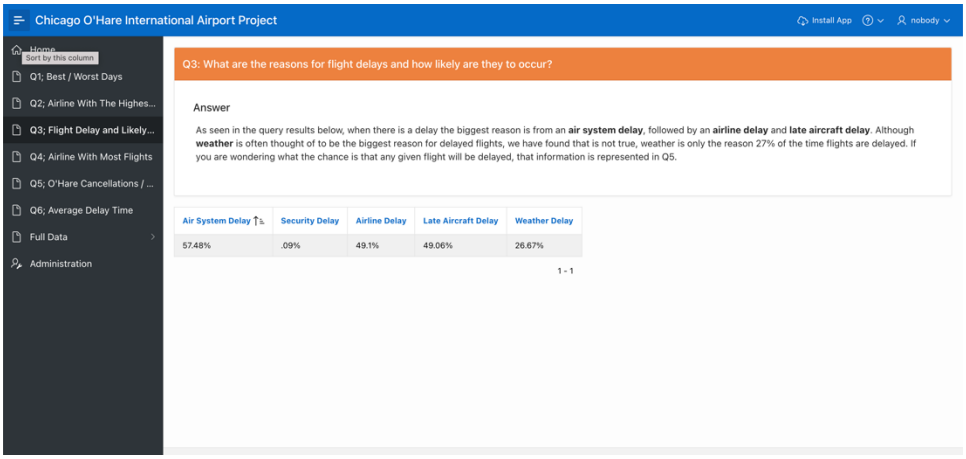


Figure 19: Flight Delay and Likelihood

Our fourth question presents a report showing which airlines have the most flights throughout the year. To visualize this data, we showed every airline's total number of flights in a pie chart with a picture of the airline that had the most yearly flights (American Airlines). All this information is shown in figures 20 and 21. We listed only the top 5 airlines in our query results because the top 5 airlines made up over 80% of all of the flights out of O'Hare. You can see this visually through the pie chart (Figure 20 on the right) where we did not limit it to the top 5.

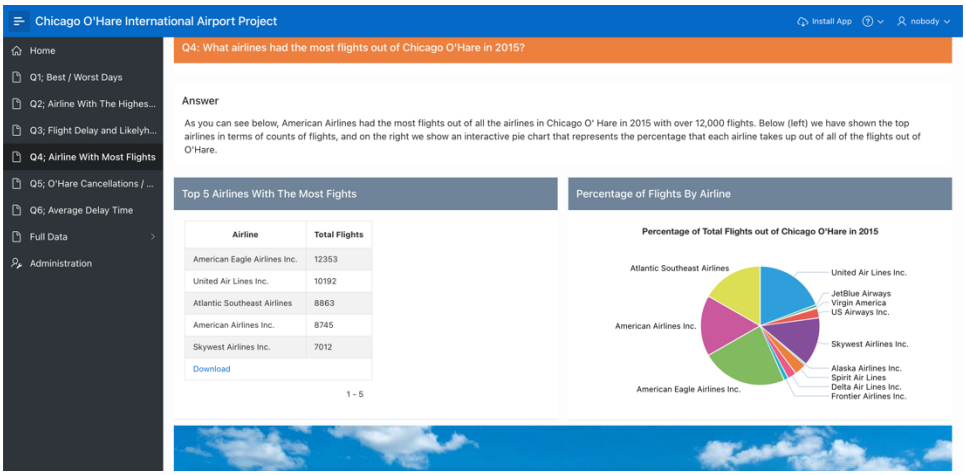


Figure 20: Airline With Most Flights

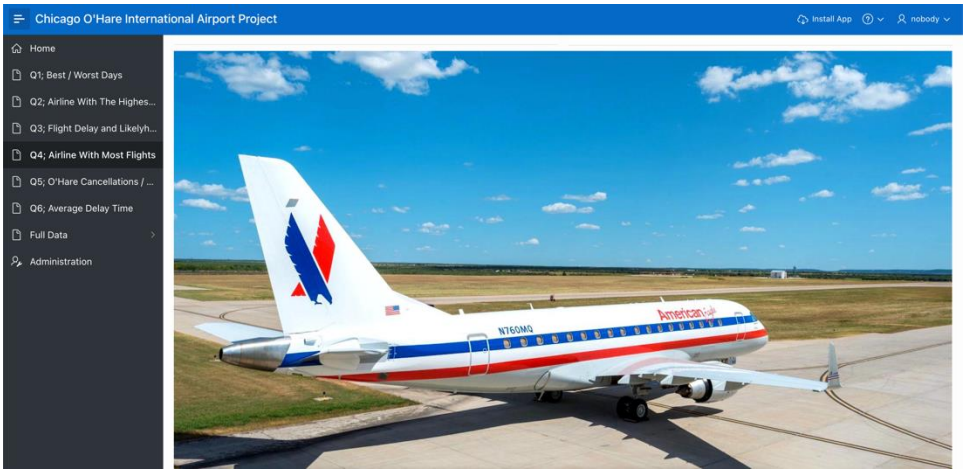


Figure 21: Airline With Most Flights Graphic

Our fifth research question shows a classic report (Figure 22) explaining the percentage of delays and cancelations the entire airport (Chicago O'Hare Airport) experienced throughout the year.

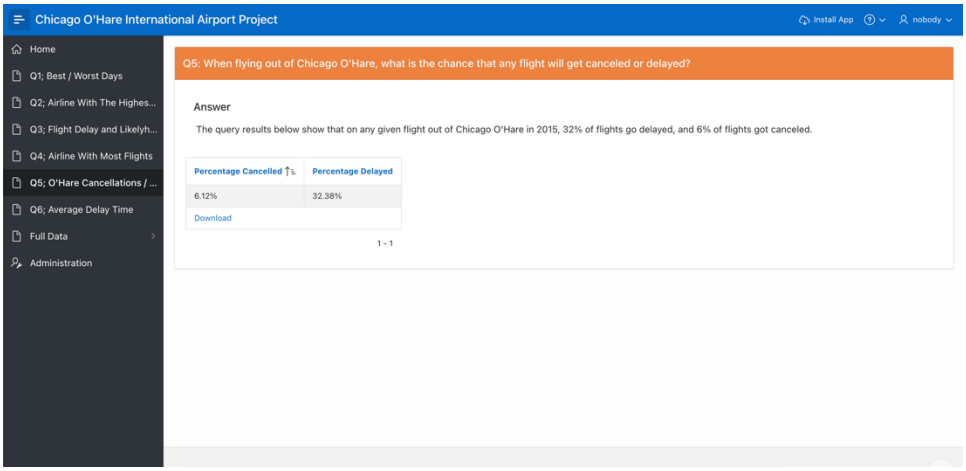


Figure 22: O'Hare Cancelations / Delays

Our last research question contains a classic report of each airline's average delay time in minutes. To better understand the data, we also presented it through a bar chart that has each airline on the x-axis and the average delay time on the y-axis. Lastly, we provided an image of the specific airline that had the highest average delay time (Frontier Airlines) at the bottom of the page. All this information is shown in figures 23 and 24.

Figure 23: Average Delay Time

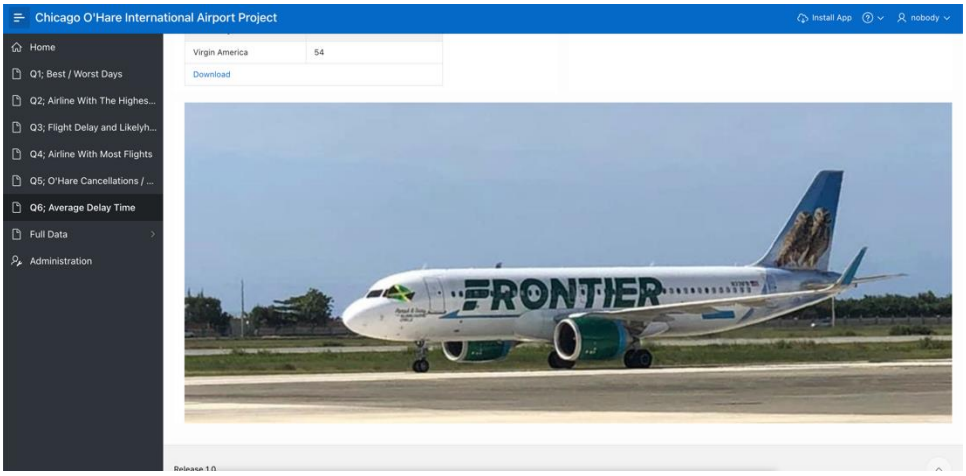


Figure 24: Average Delay Time Graphic