Flight Delay Analysis – Project Report

Context

Tens of millions of flight disruptions occur each year around the world causing travel chaos for passengers and significant financial losses for airlines and airports. By analyzing a dataset of flight delays and cancelations from 2015 in the United States I will gain insight into the patterns and causes of these delays and seek to understand how their negative impact can be reduced in the future.

This dataset was chosen as it not only fulfills the briefing specifications but also appeals to my interest in working with travel data. Having traveled quite extensively myself I have experienced a number of disruptions over the years and to analyze the cause of these delays and try to mitigate them as best as possible is a goal of every airline around the world.

Data Sourcing

The flight delay and cancellation data were collected and published by the Department of Transportation's Bureau of Transportation Statistics.

Data Collection

The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled, and diverted flights is published in DOT's monthly Air Travel Consumer Report and in this dataset of 2015 flight delays and cancellations.

Data Contents

The dataset contains many details about each of the 5.8 million flights that suffered from delays or cancelations in the US in 2015. The full list of variables is listed below in the data profile however among them are the location of their departure and arrival, expected and actual take-off and landing times as well as the attributed reason for the cancelation or delay.

Data Cleaning

Action	Columns	Notes

Check for duplicates	All	No duplicates found
Check for missing values	All	A number of values were empty; however these missing values represented meaning of some sort and thus were left empty
Merging flights, airlines, and airports databases	On airline code and airport code	Combined databases to create a master database that held all relevant information
Dropped columns	'AIRLINE_x', 'FLIGHT_NUMBER', 'TAIL_NUMBER', 'IATA_CODE_x' ,'ORIGIN_AIRPORT', 'DESTINATION_AIRPORT'	Some of these columns were duplicates of other columns after the merge, others were not needed for the analysis
Renamed columns	'IATA_CODE_y':'ORIGIN_AIRPORT_CODE', 'AIRPORT_x':'ORIGIN_AIRPORT', 'CITY_x':'ORIGIN_CITY', 'STATE_x' : 'ORIGIN_STATE', 'COUNTRY_x': 'ORIGIN_COUNTRY', 'LATITUDE_x': 'ORIGIN_LATITUDE', 'LONGITUDE_x': 'ORIGIN_LONGITUDE', 'IATA_CODE':'DESTINATION_AIRPORT_CODE', 'AIRPORT_y':'DESTINATION_AIRPORT', 'CITY_y':'DESTINATION_CITY', 'STATE_y' : 'DESTINATION_STATE', 'COUNTRY_y': 'DESTINATION_COUNTRY', 'LATITUDE_y': 'DESTINATION_LATITUDE', 'LONGITUDE_y': 'DESTINATION_LONGITUDE'	After the merge a number of columns were renamed for ease of understanding

Data Profile

Variables	Description	Time	Structure	Data Type	
YEAR	Year of flight	invariant	structured	quantitative	discrete
MONTH	Month of flight (1 -12)	invariant	structured	quantitative	discrete
DAY	Day of month (1 – 31)	invariant	structured	quantitative	discrete
DAY_OF_WEEK	Day of the week of flight (1-7)	invariant	structured	quantitative	discrete

SCHEDULED_DEPARTURE	Scheduled departure time (HHMM)	invariant	structured	quantitative	continuous
DEPARTURE_TIME	Actual departure time (HHMM)	invariant	structured	quantitative	continuous
DEPARTURE_DELAY	Time delayed (mins)	invariant	structured	quantitative	continuous
TAXI_OUT	Time for taxi out (mins)	invariant	structured	quantitative	continuous
WHEELS_OFF	Time for wheels off (mins)	invariant	structured	quantitative	continuous
SCHEDULED_TIME	Scheduled duration (mins)	invariant	structured	quantitative	continuous
ELAPSED_TIME	Actual duration (mins)	invariant	structured	quantitative	continuous
AIR_TIME	Time in air (mins)	invariant	structured	quantitative	continuous
DISTANCE	Distance covered (miles)	invariant	structured	quantitative	continuous
WHEELS_ON	Time for wheels on (mins)	invariant	structured	quantitative	continuous
TAXI_IN	Time for taxi in (mins)	invariant	structured	quantitative	continuous
SCHEDULED_ARRIVAL	Scheduled arrival time (HHMM)	invariant	structured	quantitative	continuous
ARRIVAL_TIME	Actual arrival time (HHMM)	invariant	structured	quantitative	continuous
ARRIVAL_DELAY	Delay time (mins)	invariant	structured	quantitative	continuous
DIVERTED	Was the flight diverted (1 = yes, 0 = no)	invariant	structured	quantitative	nominal
CANCELLED	Was the flight cancelled (1 = yes, 0 = no)	invariant	structured	quantitative	nominal
CANCELLATION_REASON	Why was the plan cancelled? A - Airline/Carrier B - Weather C - National Air System D - Security	invariant	structured	quantitative	nominal
AIR_SYSTEM_DELAY	Length of air system delay (minutes)	invariant	structured	quantitative	continuous
SECURITY_DELAY	Length of security delay (minutes)	invariant	structured	quantitative	continuous
AIRLINE_DELAY	Length of airline delay (minutes)	invariant	structured	quantitative	continuous

LATE_AIRCRAFT_DELAY	Length of aircraft delay (minutes)	invariant	structured	quantitative	continuous
WEATHER_DELAY	Length of weather delay (minutes)	invariant	structured	quantitative	continuous
AIRLINE	Flight airline	invariant	structured	qualitative	nominal
ORIGIN_AIRPORT_CODE	Origin airport code	invariant	structured	qualitative	nominal
ORIGIN_AIRPORT	Origin airport name	invariant	structured	qualitative	nominal
ORIGIN_CITY	Origin city	invariant	structured	qualitative	nominal
ORIGIN_STATE	Origin state	invariant	structured	qualitative	nominal
ORIGIN_COUNTRY	Origin country	invariant	structured	qualitative	nominal
ORIGIN_LATITUDE	Origin latitude	invariant	structured	quantitative	continuous
ORIGIN_LONGITUDE	Origin longitude	invariant	structured	quantitative	continuous
DESTINATION_AIRPORT_CODE	Destination airport code	invariant	structured	qualitative	nominal
DESTINATION_AIRPORT	Destination airport name	invariant	structured	qualitative	nominal
DESTINATION_CITY	Destination city	invariant	structured	qualitative	nominal
DESTINATION_STATE	Destination state	invariant	structured	qualitative	nominal
DESTINATION_COUNTRY	Destination country	invariant	structured	qualitative	nominal
DESTINATION_LATITUDE	Destination latitude	invariant	structured	quantitative	continuous
DESTINATION_LONGITUDE	Destination longitude	invariant	structured	quantitative	continuous

Database size – 5,332,914 x 41

Limitations and Ethics

- This data is taken from 2015 and thus is quite outdated at this point. However, many of the patterns and causes of delays and cancelations remain the same today and thus are still relevant.
- While the dataset was taken from Kaggle which could put into question its accuracy the source is the US Department of Transportation's Bureau of Transportation Statistics and thus is to be considered reliable.
- It would have been good to have data on the exact type of aircraft to if there was a relationship between different aircraft and manufacturers in terms of reliability and punctuality, however, the dataset does not include this data.

Exploration Questions

1) What is the leading cause of delays and cancelations for flights?
2) Which airlines have the highest rate of delays and cancelations?
3) Are there some airports or states that have a higher rate of delays and cancelations in flights departing and or arriving at their airports?

4) How successful are airlines and airports at making up lost time if they depart late?
5) Do some causes of delays result in longer delays than others?
6) How does the time of the year and day of the week impact the rate of delays and cancelations?