Airline Dataset Exploratory Analysis

Alejandro Avalos Mar
Steven Lin

MSiA - Data Visualization
Spring 2015

Abstract

The goal of this exploratory data analysis was to generate and investigate hypotheses through visual analysis. Visualization tools were used to help formulate hypotheses that should then be analyzed with rigorous statistical tests.

The exploratory analysis explores outliers, trends and relationships between flights from Chicago's International Airports, MDW and ORD, to Los Angeles and Orlando, where Disneyland or Disney World are located. The analysis indicates that the best route to travel from Chicago to Disney is from MDW to Orlando's International Airport (MCO) with AirTran (FL) or Scoot (TZ) airlines based on on-time arrival performance and cancellations. In addition, any day in September is a great option to fly, but more specifically, Mondays in October is the best option in terms of lowest percentage of flights with delays in the past.

Introduction

The Walt Disney Company is a world leader in high-quality family entertainment. The company operates the most visited theme parks in the world, attracting millions of visitors every year and generating airline traffic. In the United States, Disney operates Disneyland and Disney World, which are located in Los Angeles, CA and Orlando, FL.

This study used visualization tools to explore and analyze relationships in flight patterns from Chicago, IL to Los Angels, CA and Orlando, FL. The objective is to determine the best month, day of week and airline to travel to Disneyland or Disney World. The number of flights in the time period 1999-2008 from Chicago's Midway International Airport (MDW) and O'Hare International Airport (ORD) to Los Angeles International Airport (LAX) and Orlando International Airport (MCO) are shown in Figure 1. All figures can be found in the Appendix.

Method and Tools

Due to the large size of the dataset, the data was directly transferred from the source website to the Social Sciences Computing Cluster (SSCC) server at Northwestern University. The data was decompressed to csv files in linux and then merged and loaded into R. The dataset was then filtered using SQL-like queries, keeping only the flights from ORD and MDW to either MCO or LAX, resulting in 19 variables and 168,455 records. The selected visualization tools were R (package ggplot2) and Tableau 8.0. Features of the visualized data that looked interesting were highlighted and further investigated by isolating the subset of the data containing the interesting feature.

RStudio was used to clean, parse, query data and do the visualizations. The main R libraries used were:

- data.table - To load the data
- sqldf - To query and subset data from the main data frame
- ggplot2 - For visualization: Histograms, bar plots, scatter plots and time series

Data.data was extremely useful to load the data, as the server was not functioning and the default function read.csv() would have taken extremely long to load all the data. Sqldf was great to build sub datasets for each specific plot, by either reducing the number of fields and rows to improve efficiency, or to build new datasets (percentages, aggregating variables, etc). Ggplot2 was our main tool for visualization; the best feature ggplot2 has is the ability to add layers, which made it extremely flexible in terms of the output and the aesthetics of the plots.

In addition, Tableau was used to plot a map locating the areas of interest. Dragging the variables in tableau, and creating professional graphs with no code are some of the great attributes Tableau has. The best functionality, in our opinion, was the map feature; it was extremely easy to create maps, simply by providing the name of the city and state.

These two tools had all of the functionalities we required for this project. The only problem we had was using the geom_map function in ggplot2. Although it followed the same rubric as all of the other ggplot functions, there was lack of flexibility in terms of highlighting specific points we wanted to make, mainly highlighting specific states. Although there may be a way to do this task with geom_map, searching for the answer or help was extremely time consuming. To solve this, Tableau was used as mentioned above.

Flight Analysis

All possible direct routes from Chicago to Orland and L.A. were considered, and a time series of the flights delayed across years aggregated at the monthly level was generated (Figure 2). Note that The Federal Aviation Administration (FAA) considers a flight to be delayed when it is 15 minutes past its scheduled time.

The plot shows that the smooth average percentage of delayed flights ranged from about 5% to 30% across all years. However, the range is lower for flights from MCO (5% to 20%) compared to flights from ORD (20% to 30%). These results might be expected since ORD is one of the busiest airports in the world. The plot also indicates that the overall percentage of flights delayed has been increasing for flights from MDW, while for flights from ORD, it decreased up to 2003, but then increased. All curves seem to have a constant or decreasing trend for years greater than 2008. These fluctuating pattern with big swings was surprising. A key observation is that curves for flights departing from MDW are below those of ORD across all years.Additionally, compared to ORD, MDW tends to have less traffic and smaller carriers, which tend to have lower flight fares. Therefore the flight route MDW-MCO was further investigated.

Further Analysis by airline, day of week and month

Figures 2 shows differences in terms of average percentage of delayed flights among the flight routes. Thus, the actual distribution of the arrival delay in minutes was plotted to verify this finding. As Figure 3 shows, all distributions are right skewed towards more on-time and early arrivals. However, flights from ORD tend to have a larger mass on the late tail (the reference line are for 0 minutes and 15 minutes to indicate when a flight is considered early, on-time or late).

Furthermore, the plot also shows that flights from ORD to LAX have the highest frequency of late flights. Based on these conclusions, the data was further analyzed by month, delay reasons, day of week and airline for flights from Chicago to MCO. Note that American Airlines (AA) and United Airlines (UA) fly from ORD to MCO, while Scoot (TZ), AirTran (FL) and Southwest Airlines (WN) fly from MDW to MCO. Combining the results from Figures 4, 5 & 6, we conclude that the best times to fly, in terms of delays and cancellations, are any day in September or a Monday in October (Figures 6) with airlines TZ or FL (Figures 4 & 5). The best performer airlines are expected because the airlines are tied to the airport. The fact that September overall is the best month to travel also makes sense since it is the month after summer. The fact that Monday in October was the best day of the week is unexpected.

Summary

The exploratory analysis provided insight on the dataset and helped formulate hypotheses regarding the initial question. The following is a summary of the most important conclusions:

- Regardless of destination, MDW seems to be the best departure airport performer
- The flight route MDW-MCO in particular appears to be the best route for our case study
- The best airline to travel seems to be with AirTran (FL) or Scoot (TZ) airline
- September is the best month to fly overall, but more specifically, Mondays in October

Future Work

With the aid of visualization tools, the exploratory analysis helped formulate interesting hypotheses. The next steps would involve conducting more rigorous statistical analysis to test the validity of these hypothesis. For example, a correlation analysis and ANOVA should be done to determine if there is a significant difference in delays among airlines. A predictive model can be also built to forecast the time series and see if the trend of delays across years of ORD vs. MCO holds.

Appendix

	Name	Description
1	Year	1987-2008
2	Month	1-12
3	DayofMonth	1-31
4	DayOfWeek	1 (Monday) - 7 (Sunday)
5	DepTime	actual departure time (local, hhmm)
6	CRSDepTime	scheduled departure time (local, hhmm)
7	ArrTime	actual arrival time (local, hhmm)
8	CRSArrTime	scheduled arrival time (local, hhmm)
9	UniqueCarrier	unique carrier code
10	FlightNum	flight number
11	TailNum	plane tail number
12	ActualElapsedTime	in minutes
13	CRSElapsedTime	in minutes
14	AirTime	in minutes
15	ArrDelay	arrival delay, in minutes
16	DepDelay	departure delay, in minutes
17	Origin	origin IATA airport code
18	Dest	destination IATA airport code
19	Distance	in miles
20	TaxiIn	taxi in time, in minutes
21	TaxiOut	taxi out time in minutes
22	Cancelled	was the flight cancelled?
23	CancellationCode	reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
24	Diverted	1 = yes, 0 = no
25	CarrierDelay	in minutes
26	WeatherDelay	in minutes
27	NASDelay	in minutes
28	SecurityDelay	in minutes
29	LateAircraftDelay	in minutes

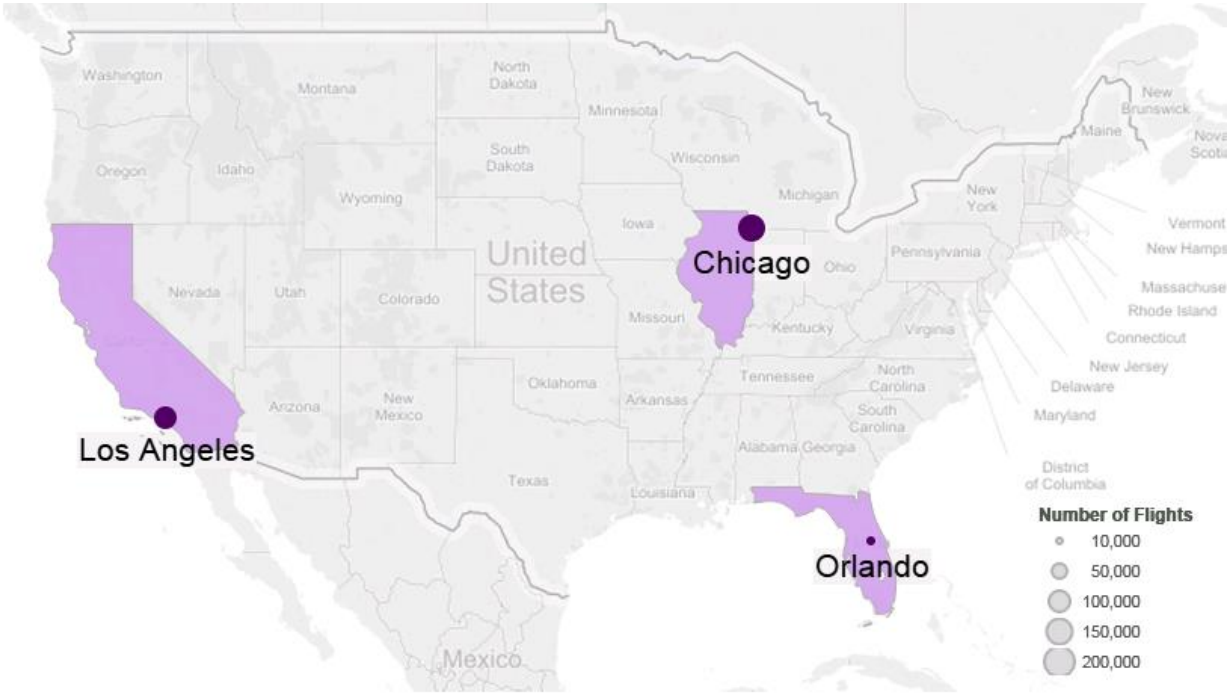

Figure 1. Flights from Chicago, IL (ORD and MDW) to Los Angeles , CALAX) and Orlando, FL (MCO)

Figure 2. Time series of percentage of flights delayed by flight route

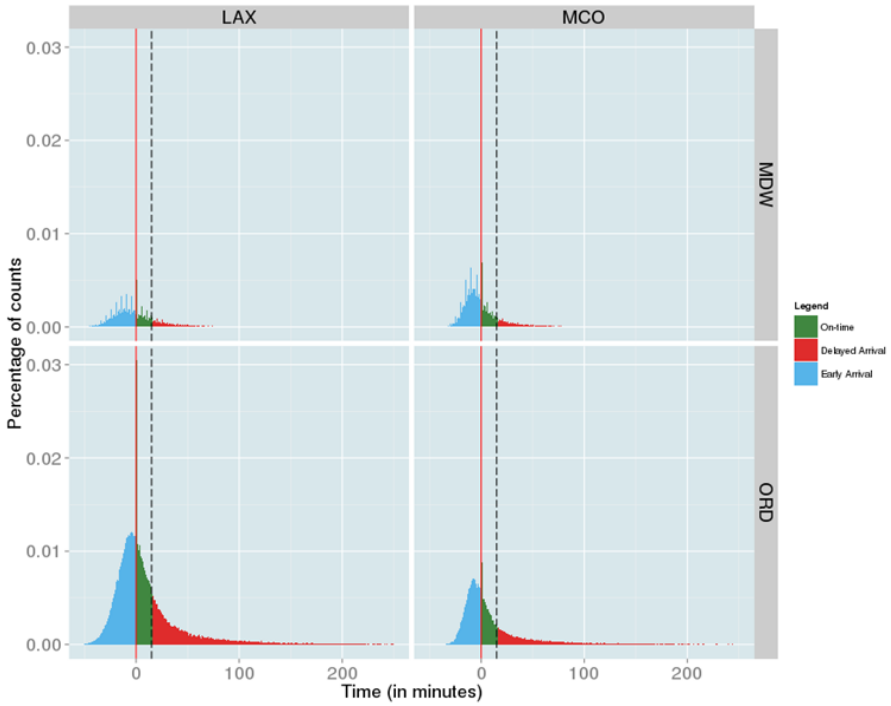

Figure 3. Distribution of flights delayed

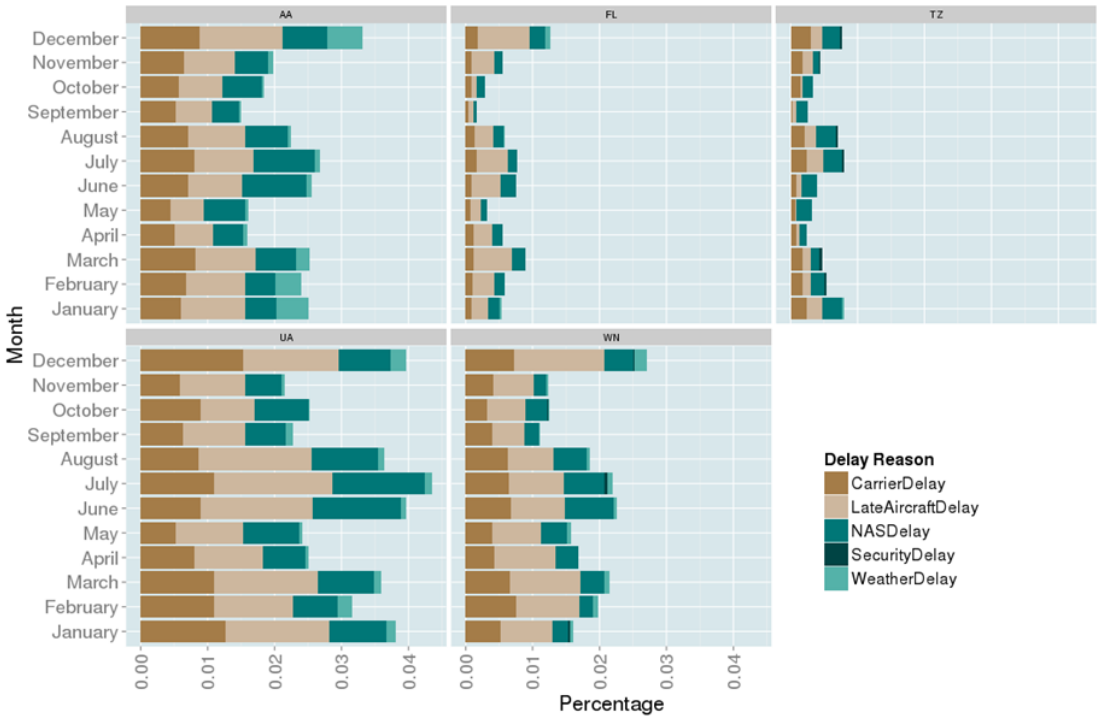

Figure 4. Distribution of delay reasons by airline and month for flights to MCO

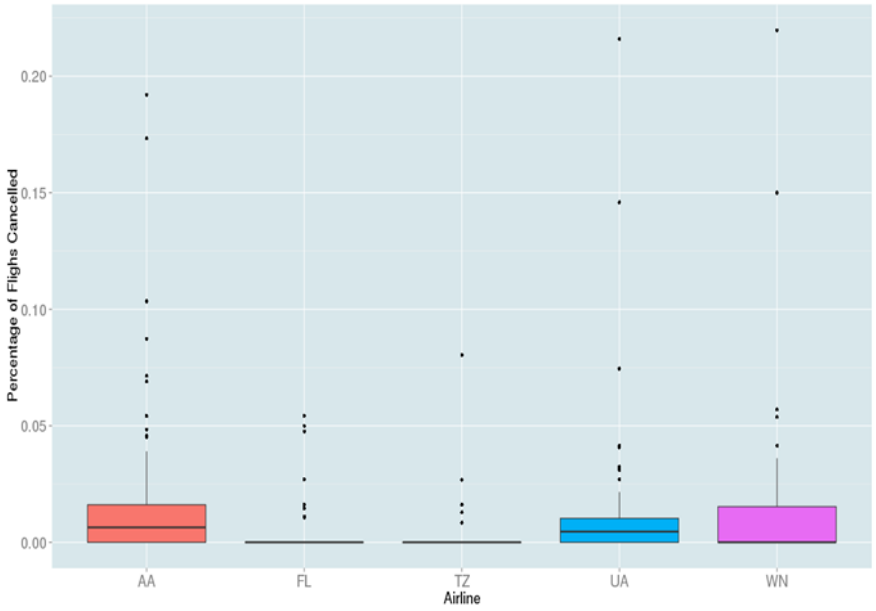

Figure 5 Box plot Cancellations and Flight delays by airline for flights to MCO

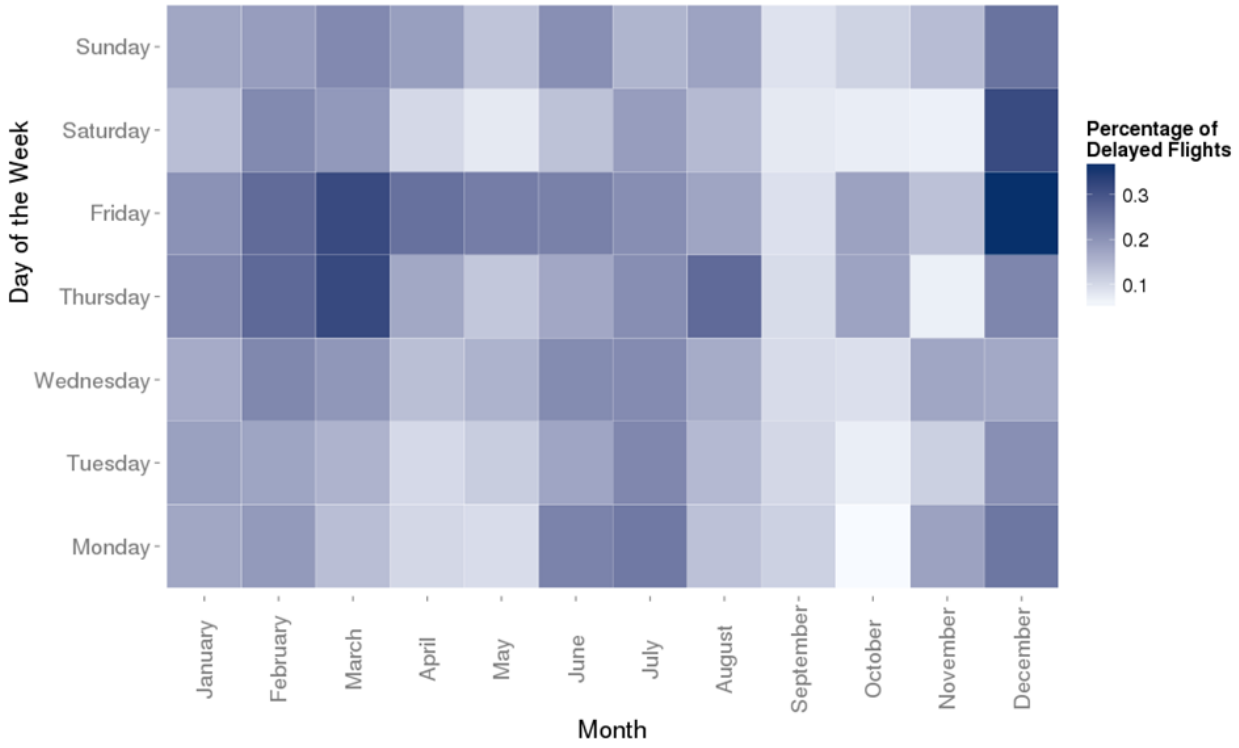

Figure 6 . Delayed flights by day of week and month for MDW-MCO route