Daddy, I want to go to Disney ...

Alejandro Avalos Mar and Steven Lin

Northwestern University, Evanston, IL

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.

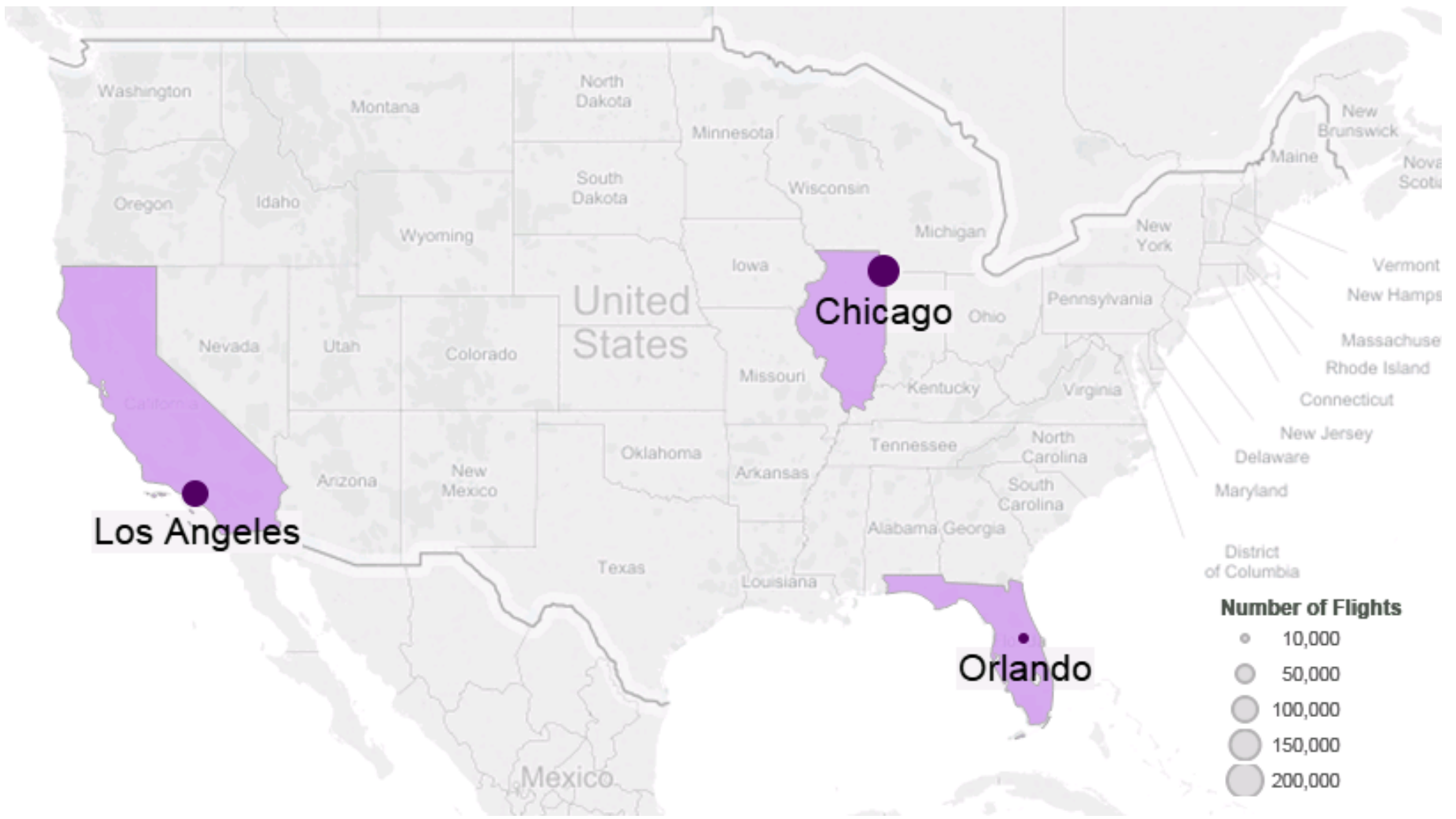

Figure 1. Flights from Chicago, IL (ORD and MDW) to Los Angeles , CALAX) and Orlando, FL (MCO)

Data

The Airline data used for this project is publicly available online from the United States Department of Transportation. The scope of the data is 10 years (1999-2008) and contains over 60 million records, where each record represents a specific flight. There are 29 variables including date time of arrival and departure, delays, cancellations, carrier and airport information.

Source: http://stat-computing.org/dataexpo/2009/the-data.html

Method

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.

Flight route 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%).

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. 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.

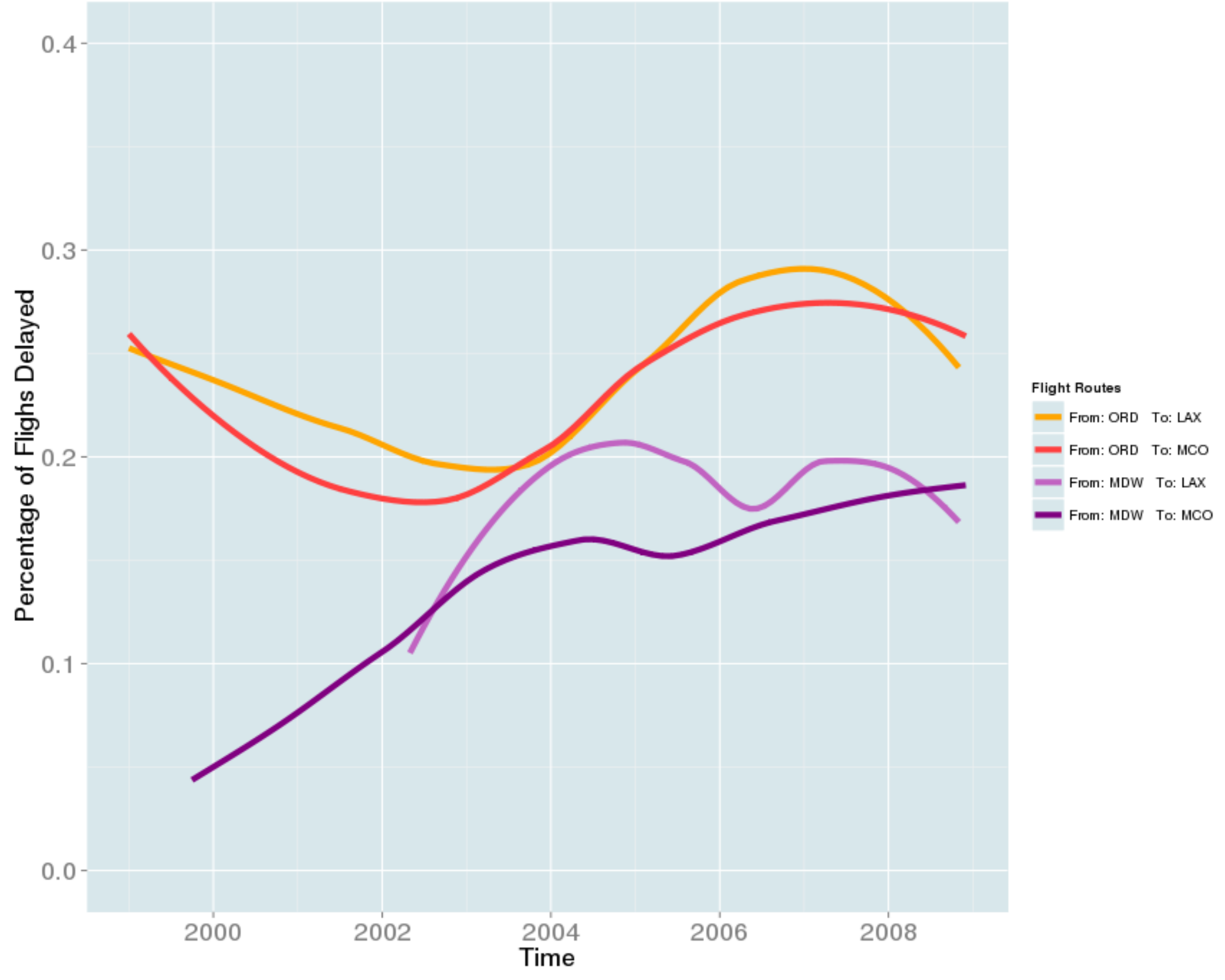

Figure 2. Time series of percentage of flights delayed by flight route

Deeper dive by airline and day of week

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).

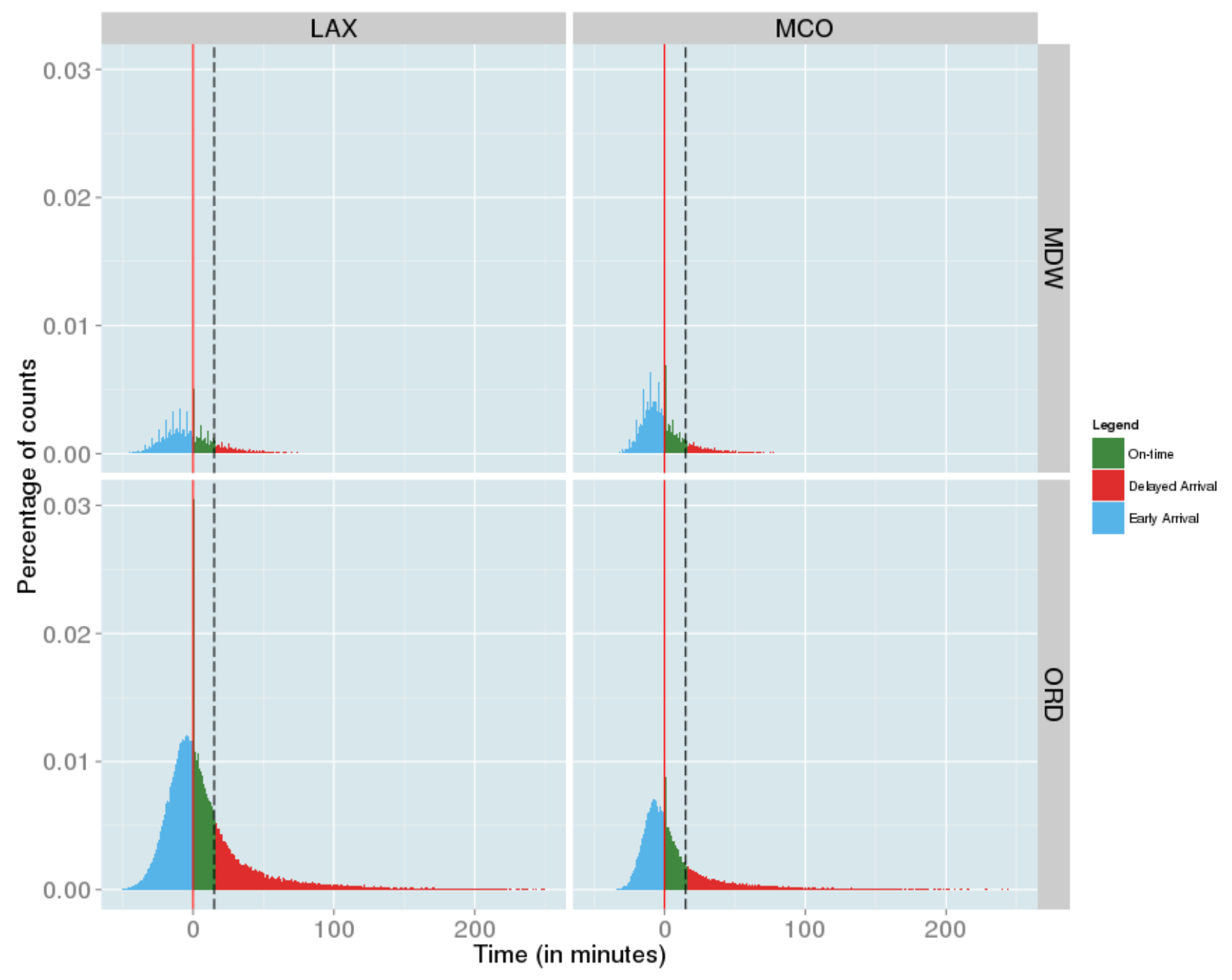

Figure 3. Distribution of flights delayed

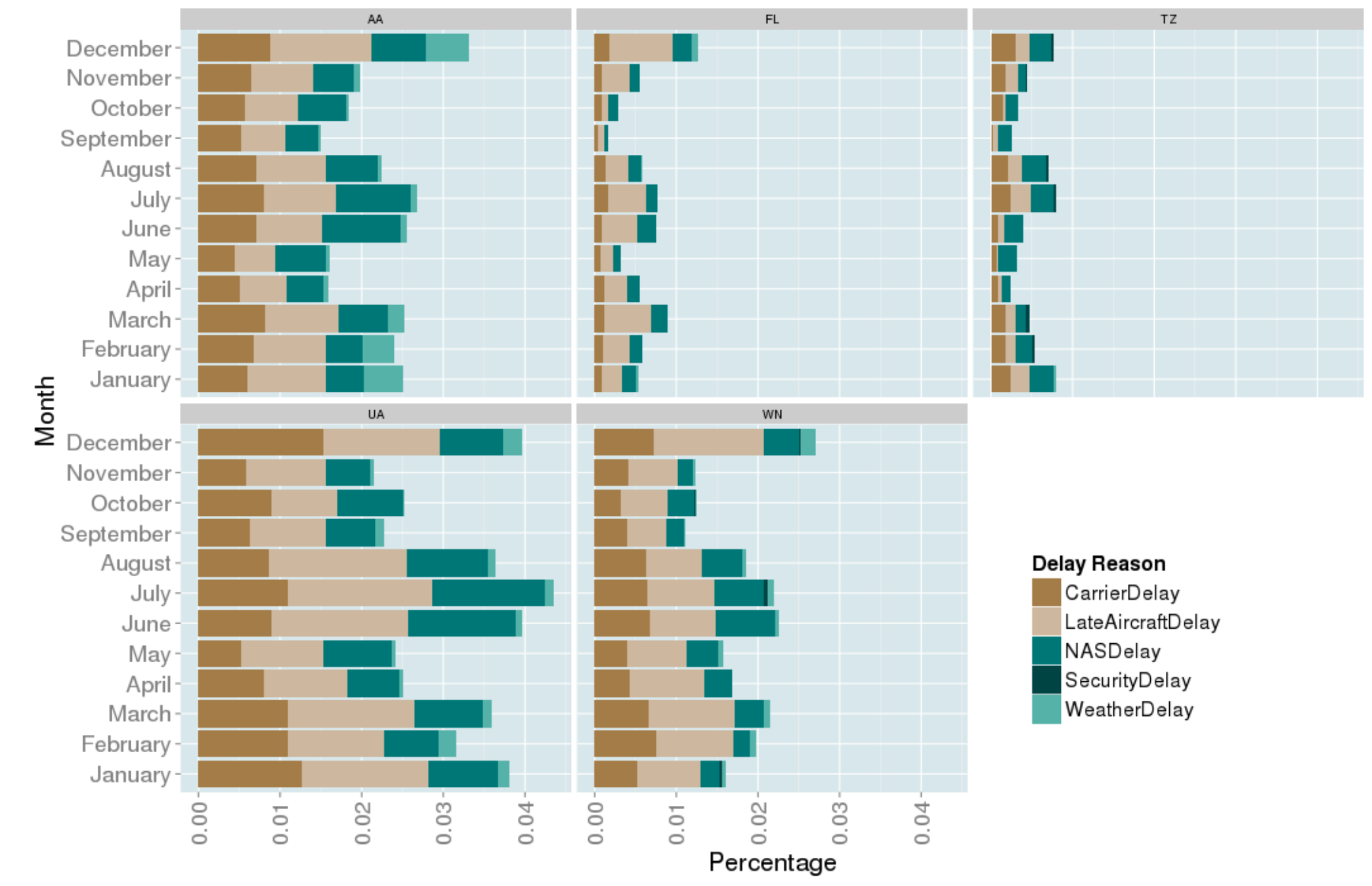

Figure 4. Distribution of delay reasons by airline and month for flights to MCO

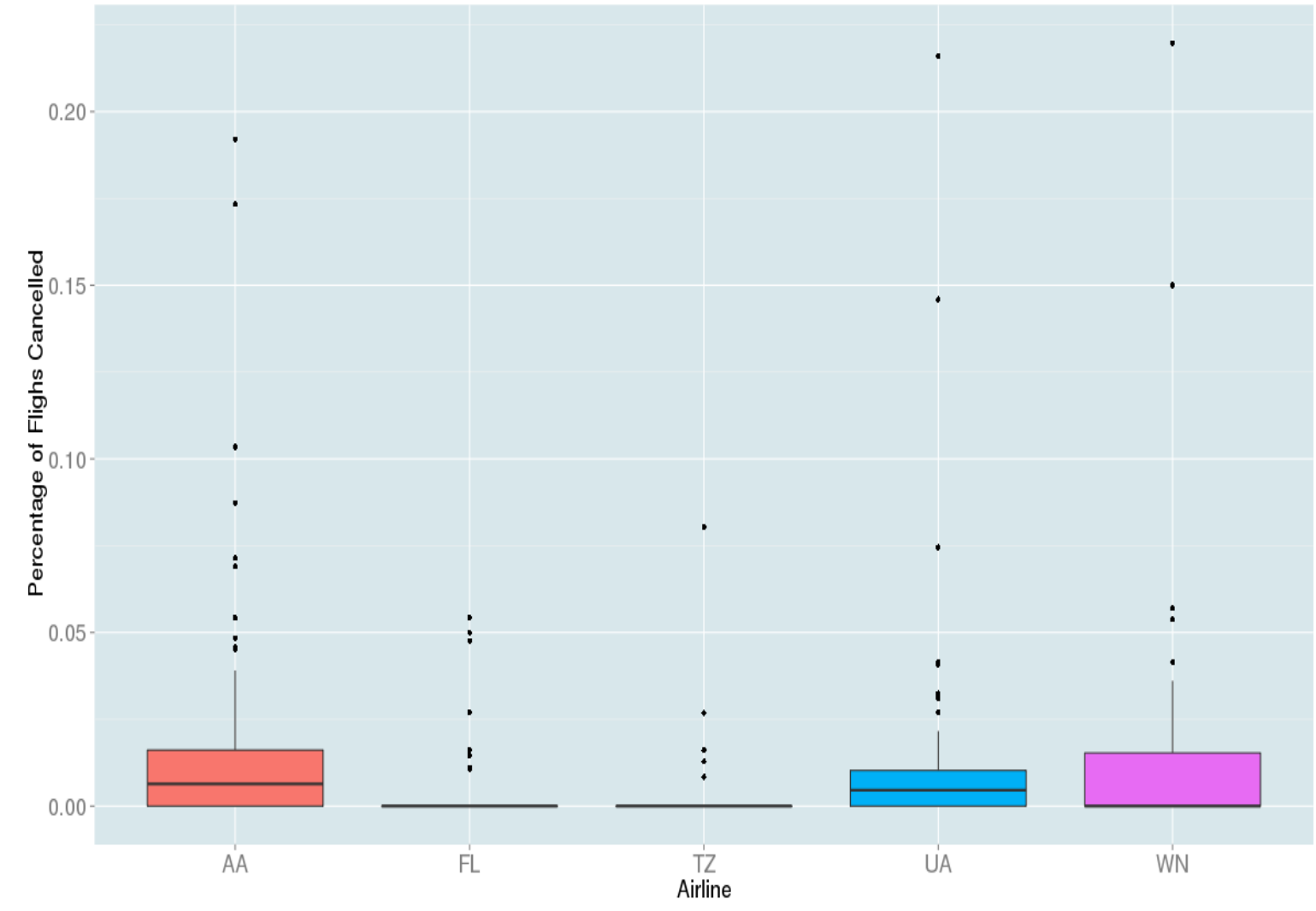

Figure 5 Box plot Cancellations and Flight delays by airline for flights to MCO

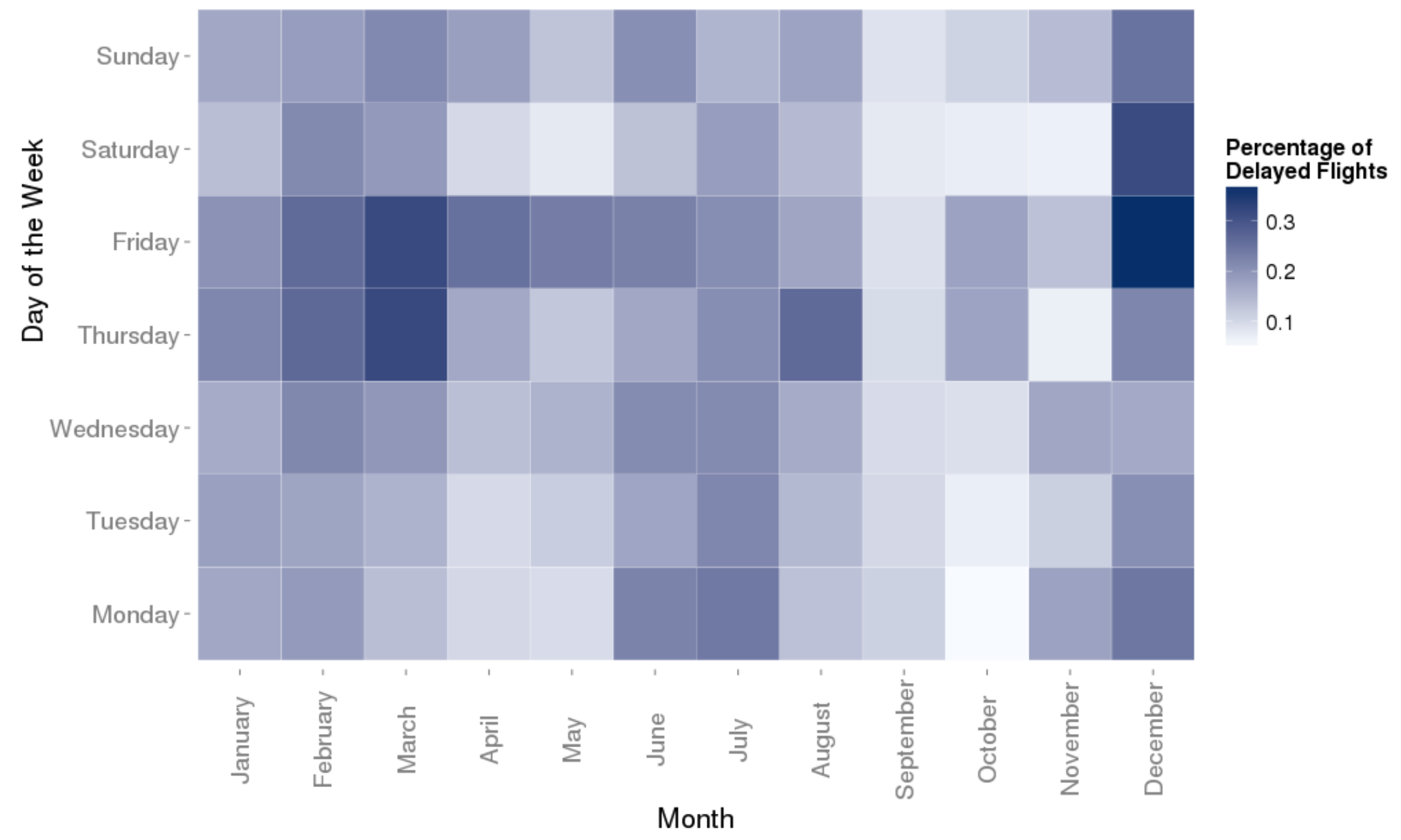

Figure 6 . Delayed flights by day of week and month for MDW-MCO route

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) airlines**
- **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.

Contact information

Northwestern University
Robert R. McCormick School of Engineering and Applied Science

2145 Sheridan Road C210
Evanston, IL 60208

T: (847) 467-4520.
E: analytics@northwestern.edu