Comparing Airline Delays & Cancellations by Season

Katie Kotler, Kristina Urberg, Angelina Sala,
Ethan Shang, Steven Peng
Section 8, Team 5

Introduction

- Status of flights in the U.S.
- CORGIS Data Set Project
  - Created by Austin Cory Bart
  - March 27, 2015
- Observations based on:
  - Airport code (e.g. LAX, JFK, etc...)
  - Time (year / month)
  - Status (delayed / cancelled / rescheduled / on time)
  - Reasoning for delays

	Airport.Code	Airport.Name	Time.Label	Time.Month	Time.Month Name	Time.Year
1	ATL	Atlanta, GA: Hartsfield-Jackson Atlanta International	2003/06	6	June	2003
2	BOS	Boston, MA: Logan International	2003/06	6	June	2003
3	BWI	Baltimore, MD: Baltimore/Washington International Thurgo...	2003/06	6	June	2003
4	CLT	Charlotte, NC: Charlotte Douglas International	2003/06	6	June	2003
5	DCA	Washington, DC: Ronald Reagan Washington National	2003/06	6	June	2003
6	DEN	Denver, CO: Denver International	2003/06	6	June	2003
7	DFW	Dallas/Fort Worth, TX: Dallas/Fort Worth International	2003/06	6	June	2003
8	DTW	Detroit, MI: Detroit Metro Wayne County	2003/06	6	June	2003
9	EWR	Newark, NJ: Newark Liberty International	2003/06	6	June	2003
10	FLL	Fort Lauderdale, FL: Fort Lauderdale-Hollywood International	2003/06	6	June	2003
11	IAD	Washington, DC: Washington Dulles International	2003/06	6	June	2003
12	IAH	Houston, TX: George Bush Intercontinental/Houston	2003/06	6	June	2003
13	JFK	New York, NY: John F. Kennedy International	2003/06	6	June	2003
14	LAS	Las Vegas, NV: McCarran International	2003/06	6	June	2003
15	LAX	Los Angeles, CA: Los Angeles International	2003/06	6	June	2003
16	LGA	New York, NY: LaGuardia	2003/06	6	June	2003

Research question: Of the top five busiest airports in the U.S., when and where are there the most flight delays and/or cancellations?

Literature Review

Understanding the Summer Air Travel Mess, NYT

- Summer is the busiest travel season
- Constant demand leads to greater impact when scheduling conflicts occur
- Major reasons: weather, staffing shortages, equipment malfunctions
- Least reliable airports: Newark, LaGuardia, and Orlando

- **To minimize the impact of delays:**
  - Book direct flights
  - Early morning departures
  - Avoid flying during weekends
  - Never book the last flight of the day

Methods

Data Manipulation

- Variables of interest variables include:
  - Number of flights delayed
  - Number of flights canceled
- Created a new variable **airlines_season**
- Filtered data to only include 5 busiest airports: **ATL, ORD, DEN, DFW, LAX**

EDA

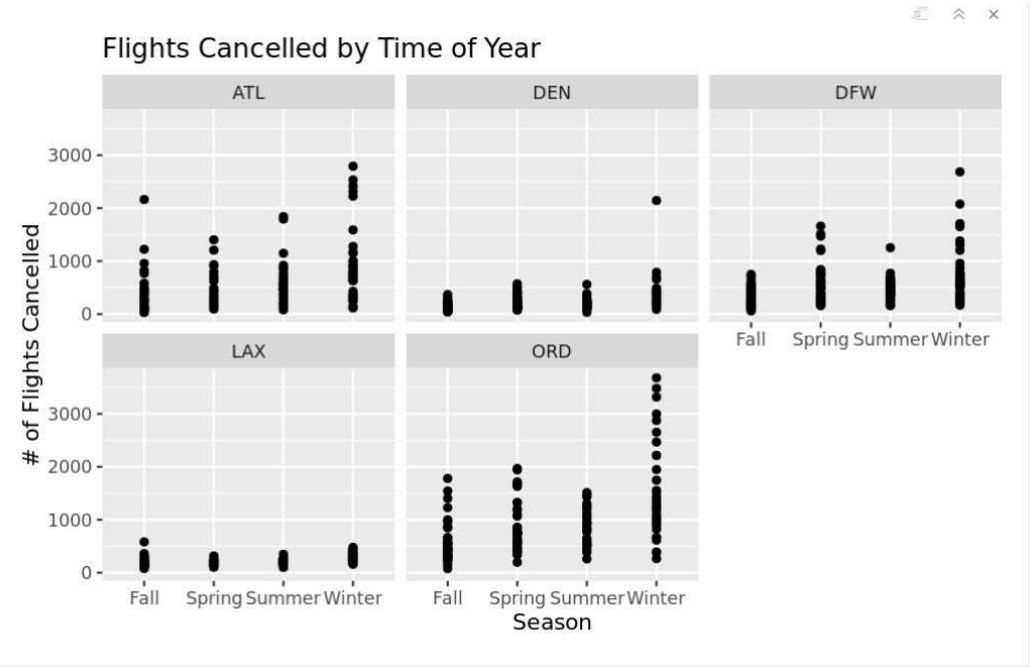
- Visualized flights delayed and canceled by time of year
- Calculated the mean number of flights canceled and delayed for each airport
- Highest cancellation rate: **ORD**
- Highest delay rate: **ATL**

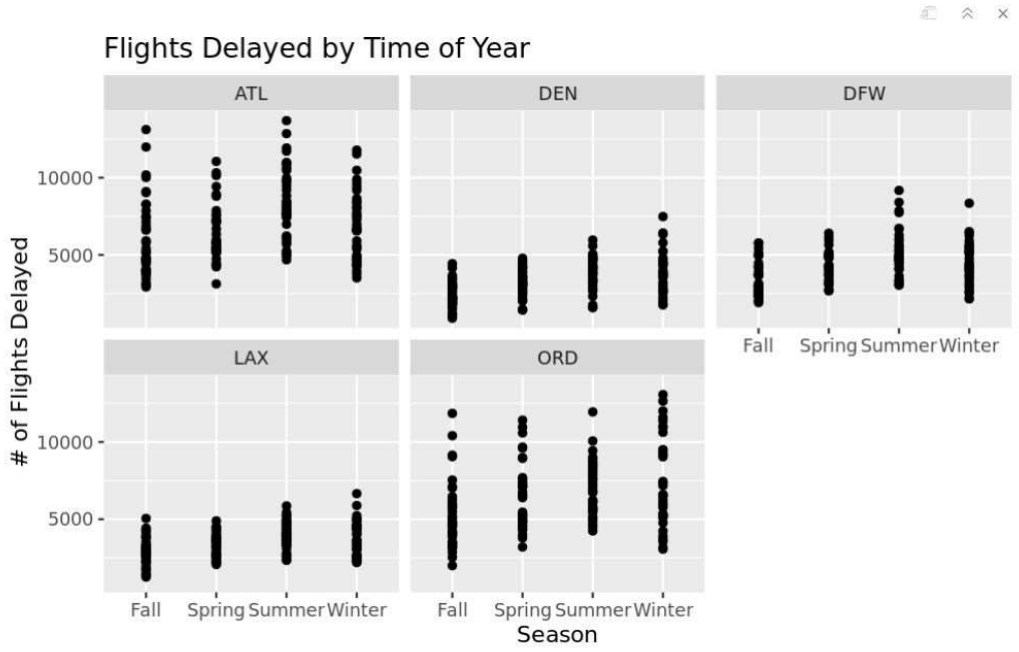
Further Analysis

Linear Regression Modeling

- Simple: relationship between season & delays/cancellations
- Multiple: relationship between season, airport & delays/cancellations (interactive)

Methods Visualizations





Results

- **Delays:**
  - Greatest coefficient in front of Summer

- **Cancellations:**
  - Greatest coefficient in front of Winter

A tibble: 4 × 5

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	3988.6821	160.1178	24.910929	1.787278e-100
seasonSpring	767.6568	231.1101	3.321607	9.380054e-04
seasonSummer	1688.0205	226.4407	7.454581	2.472648e-13
seasonWinter	1251.3495	227.9256	5.490167	5.485703e-08

4 rows

A tibble: 4 × 5

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	316.4462	33.89535	9.335975	1.090616e-19
seasonSpring	150.8594	48.92372	3.083563	2.119768e-03
seasonSummer	149.9641	47.93526	3.128471	1.824623e-03
seasonWinter	445.2486	48.24960	9.228027	2.708504e-19

4 rows

Results cont.

Interactive Delay Model:

Not all airports have the highest number of delays in the same season

- Greatest seasonal coefficient for ATL is Summer
- Greatest seasonal coefficient for Denver is Winter

```r
{r}
#| label: lin-reg-airport-code

model2_delayed <- linear_reg() |>
  set_engine("lm") |>
  fit(Statistics.Flights.Delayed ~ season * Airport.Code, data =
airlines_season)

model2_delayed |>
  tidy()
```

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>
(Intercept)	6021.92308	263.7509	22.83186109
seasonSpring	440.88248	380.6916	1.15810935
seasonSummer	2336.51282	373.0000	6.26410863
seasonWinter	791.62955	375.4460	2.10850456
Airport.CodeDEN	-3592.89744	373.0000	-9.63243156
Airport.CodeDFW	-2596.89744	373.0000	-6.96219062
Airport.CodeLAX	-3100.28205	373.0000	-8.31174705
Airport.CodeORD	-876.12821	373.0000	-2.34886888
seasonSpring:Airp…	301.98077	538.3792	0.56090720
seasonSummer:Air…	-884.30769	527.5017	-1.67640720

```
[1] 0.07468645
[1] 0.508484
```

Discussion

Major Findings

- Season w/ highest # delays: Summer
- Season w/ highest # cancellations: Winter
- Airport also seemed to have a significant impact on the number of delays or cancellations

Limitations

- Only examined five most busy airports → small portion of the airlines data set
- There are likely more variables than the season that impact the number of delays and cancellations

Further Research

- Examining how the airport impacts flight delays or cancellations
- Exploring what are the root causes of delays and cancellations across the U.S.

Thank You