Describing & Forecasting Flight delays

At Chicago O'Hare International Airport

TEAM #14
David Forteguerre
Vijet Muley
Advait Ramesh Iyer
Anisha Kumar
Kshitij Sankesara

MBC638 Final Project Presentation

PROJECT OBJECTIVE

→ **Analyze *2015 US flights data* to uncover insights and attempt to explain/predict flight delays**

Presentation overview:

About the data | I. Descriptive analysis | II. Forecasting

Data source

Data: **2015 Flight Delays and Cancellations**

- available on Kaggle at https://www.kaggle.com/usdot/flight-delays
- originally came from the Department of Transportation Statistics website
  (available at https://www.transtats.bts.gov/ > Data Finder[By Mode [Aviation]] > Airline On-Time Performance Data > On-Time Performance > [all months from 2015 can be downloaded from this page])

The folder downloaded (592 MB) contained 3 datasets : **airlines.csv** (2 columns), **airports.csv** (7 columns), **flights.csv** (31 columns) (main dataset, about 6 millions rows; lots of missing values)

- 2015 only (all 12 months)
- US airports only
- US airlines only

The dataset was **cleaned** extensively, only **Chicago Airport (ORD)** was kept as the origin

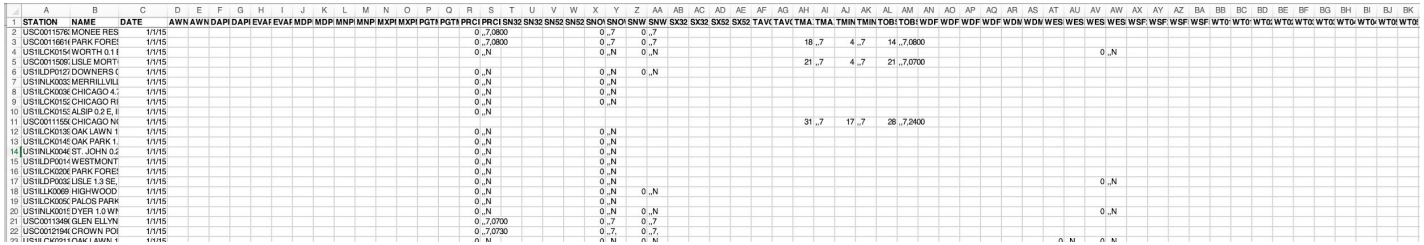
- Note that the month of October (11) was omitted due to the absence of airport codes (departure and destination) for the entire month of October in the original dataset
- Some data preprocessing was done (e.g. times were discretized: morning, afternoon, evening, etc.)

MONTI	DAY	DAY_C	AIRLIN	ORIGI	DESTI	FLIGH	SCHEDUL	SCHEDUL	STATUS	DEPAF	ARRIV
1	1	4	B6	ORD	BOS	867	night	morning	on time	-11	-33
1	1	4	NK	ORD	LGA	733	night	morning	delayed	-9	27
1	1	4	DL	ORD	ATL	606	night	morning	on time	2	-21
1	1	4	UA	ORD	MCO	1005	morning	morning	on time	-5	-21
1	1	4	UA	ORD	SFO	1846	morning	morning	on time	8	-12
1	1	4	UA	ORD	LAX	1744	morning	morning	on time	0	0

Weather data

In the city of Chicago, we expected **weather** to be a very good predictor of flight delays.

- Retrieved weather data from the National Oceanic and Atmospheric Administration
  https://www.ncdc.noaa.gov/cdo-web/
- The data consisted of many weather stations' records at different points in time



- It was very **noisy** and and many **values were missing** → a pivot table was created with averages aggregated by day

Date	AWND	PRCP	SNOW	TAVG	Fog	Heavy_	Thund	Ice_pe	Hail	Glaze	Haze	Blowir	Tornac	Damagi
1/1/15	13.6433333	0	0	20	0	0	0	0	0	0	0	0	0	0
1/2/15	4.02666667	0.00208333	0	26	0	0	0	0	0	0	0	0	0	0
1/3/15	4.69666667	0.21463158	0.30140845	31	1	0	0	1	0	1	0	0	0	0
1/4/15	14.24	0.58297872	0.76835443	29	1	0	0	0	0	1	1	1	0	0
1/5/15	9.91666667	0.07940476	0.93636364	2	1	0	0	0	0	0	1	0	0	0
1/6/15	12.38	0.1616129	2.40714286	5	1	0	0	0	0	0	1	1	0	0
1/7/15	16.2566667	0.00197802	0.02682927	3	0	0	0	0	0	0	1	0	0	0
1/8/15	16.4033333	0.00336957	0.08915663	-2	0	0	0	0	0	0	1	1	0	0
1/9/15	15.2833333	0.10822222	1.83382353	8	0	0	0	0	0	0	1	1	0	0

- Finally, **key variables** (both qualitative and quantitative) that were expected to impact delays were selected and **merged** to our main dataset (VLOOKUP)

e.g. =VLOOKUP(A:A,'Final dataset before merging'!A:O,2)

About the variables in the final dataset

ORIGINAL VARIABLES

WEATHER VARIABLES WE ADDED

DATE	MONT	DAY	DAY_(	AIRLIN	ORIGI	DESTI	FLIGH	SCHEDUL	SCHEDUL	STATUS	DEPAI	ARRIV	AWND	PRCP	SNOW	TAVG	Fog	Heavy_	Thund	Ice_pe	Hail	Glaze	Haze	Blowir	Torna	Damag
1/1/15	1	1	4	B6	ORD	BOS	867	night	morning	on time	-11	-33	13.6	0.0	0.0	20.0	0	0	0	0	0	0	0	0	0	0
1/1/15	1	1	4	NK	ORD	LGA	733	night	morning	delayed	-9	27	13.6	0.0	0.0	20.0	0	0	0	0	0	0	0	0	0	0
1/1/15	1	1	4	DL	ORD	ATL	606	night	morning	on time	2	-21	13.6	0.0	0.0	20.0	0	0	0	0	0	0	0	0	0	0
1/1/15	1	1	4	UA	ORD	MCO	1005	morning	morning	on time	-5	-21	13.6	0.0	0.0	20.0	0	0	0	0	0	0	0	0	0	0
1/1/15	1	1	4	UA	ORD	SFO	1846	morning	morning	on time	8	-12	13.6	0.0	0.0	20.0	0	0	0	0	0	0	0	0	0	0
1/1/15	1	1	4	UA	ORD	LAX	1744	morning	morning	on time	0	0	13.6	0.0	0.0	20.0	0	0	0	0	0	0	0	0	0	0
1/1/15	1	1	4	NK	ORD	LAX	1744	morning	morning	on time	-6	-12	13.6	0.0	0.0	20.0	0	0	0	0	0	0	0	0	0	0
1/1/15	1	1	4	F9	ORD	ATL	606	morning	morning	delayed	65	61	13.6	0.0	0.0	20.0	0	0	0	0	0	0	0	0	0	0
1/1/15	1	1	4	UA	ORD	MCO	1005	morning	morning	on time	14	3	13.6	0.0	0.0	20.0	0	0	0	0	0	0	0	0	0	0

CANDIDATE DEPENDENT VARIABLES FOR MODELING

VARIABLE DICTIONARY

DATE → we *created* based on MONTH and DAY variables
MONTH → (categorical) month of year (1 = Jan, 2 = Feb, etc.)
DAY → (categorical) day of the month (1, 2, 3, ..., 29, 30, 31)
DAY_OF_WEEK → (categorical) weekday (1 = Mon, 2 = Tue, etc.)
AIRLINE → (categorical) airline IATA code
ORIGIN → (categorical) origin airport IATA code (ORD only)
DESTINATION → (categorical) destination airport IATA code
FLIGHT_DISTANCE → (quantitative) flight distance range for given flight (miles)
SCHEDULED_DEPARTURE → (categorical) scheduled departure time of day
SCHEDULED_ARRIVAL → (categorical) scheduled arrival time of day
STATUS (categorical) → we *created* using delay variable & official FAA definition
DEPARTURE_DELAY → (quantitative) departure delay (minutes)
ARRIVAL_DELAY → (quantitative) arrival delay (minutes)

VARIABLE DICTIONARY

AWND → (quantitative) average daily wind speed (mph)
PRCP → (quantitative) precipitation average for the day (inches)
SNOW → (quantitative) snowfall average for the day (inches)
TAVG → (quantitative) temperature average for the day (Fahrenheit)
Fog, Heavy_fog, Thunder, Ice_pellets, Hail, Glaze, Haze, Blowing_snow, Tornado, Damaging_winds → (dummy) all these weather dummy variables indicate whether the corresponding condition occurred that day. By including all these dummies in a regression, the baseline would be "normal weather conditions".

> *The Federal Aviation Administration (FAA) considers a flight to be delayed when it takes off and/or lands **15 minutes** later than its scheduled time.*
>
> *Wikipedia*

About the data

- Our dataset is **unique**
- **Population of interest:** all domestic flights departing from the Chicago O'Hare International Airport (ORD)
- The data collected is a **sample** (non random) which should be representative of of the population of interest (***not*** accounting for international flights)

I. Main dataset *(our focus)*: **(# of rows: 59,644)**

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	DATE	MONT	DAY	DAY_(	AIRLIN	ORIGI	DESTI	FLIGH	SCHEDUL	SCHEDUL	STATUS	DEPAI	ARRIV	AWND	PRCP	SNOW	TAVG	Fog	Heavy	Thund	Ice_pe	Hail	Glaze	Haze	Blowir	Torna	Damag
2	1/1/15	1	1	4	B6	ORD	BOS	867	night	morning	on time	-11	-33	13.6	0.0	0.0	20.0	0	0	0	0	0	0	0	0	0	0
3	1/1/15	1	1	4	NK	ORD	LGA	733	night	morning	delayed	-9	27	13.6	0.0	0.0	20.0	0	0	0	0	0	0	0	0	0	0
4	1/1/15	1	1	4	DL	ORD	ATL	606	night	morning	on time	2	-21	13.6	0.0	0.0	20.0	0	0	0	0	0	0	0	0	0	0
5	1/1/15	1	1	4	UA	ORD	MCO	1005	morning	morning	on time	-5	-21	13.6	0.0	0.0	20.0	0	0	0	0	0	0	0	0	0	0
6	1/1/15	1	1	4	UA	ORD	SFO	1846	morning	morning	on time	8	-12	13.6	0.0	0.0	20.0	0	0	0	0	0	0	0	0	0	0
7	1/1/15	1	1	4	UA	ORD	LAX	1744	morning	morning	on time	0	0	13.6	0.0	0.0	20.0	0	0	0	0	0	0	0	0	0	0
8	1/1/15	1	1	4	NK	ORD	LAX	1744	morning	morning	on time	-6	-12	13.6	0.0	0.0	20.0	0	0	0	0	0	0	0	0	0	0
9	1/1/15	1	1	4	F9	ORD	ATL	606	morning	morning	delayed	65	61	13.6	0.0	0.0	20.0	0	0	0	0	0	0	0	0	0	0
10	1/1/15	1	1	4	UA	ORD	MCO	1005	morning	morning	on time	14	3	13.6	0.0	0.0	20.0	0	0	0	0	0	0	0	0	0	0

II. Additional datasets *(to answer specific questions in descriptive section)*:

a. Delayed flights with a breakdown of delays **(# of rows: 15,048)**

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	DATE	MONT	DAY	DAY_(	AIRLIN	ORIGI	DESTI	FLIGH	SCHEDUL	SCHEDUL	STATUS	DEPAI	ARRIV	AIR_SYSTEM_D	SECURITY_D	AIRLINE_D	LATE_AIRCRAFT_D	WEAT	AWND	PRCP	SNOW
2	1/1/15	1	1	4	NK	ORD	LGA	733	night	morning	delayed	-9	27	27	0	0	0	0	13.6	0.0	0.0
3	1/1/15	1	1	4	F9	ORD	ATL	606	morning	morning	delayed	65	61	0	0	61	0	0	13.6	0.0	0.0
4	1/1/15	1	1	4	UA	ORD	LAX	1744	morning	morning	delayed	21	16	0	0	16	0	0	13.6	0.0	0.0
5	1/1/15	1	1	4	OO	ORD	ATL	606	morning	morning	delayed	100	88	0	0	63	25	0	13.6	0.0	0.0
6	1/1/15	1	1	4	UA	ORD	SFO	1846	afternoon	afternoon	delayed	46	44	0	0	44	0	0	13.6	0.0	0.0
7	1/1/15	1	1	4	US	ORD	PHL	678	afternoon	afternoon	delayed	25	15	0	0	2	13	0	13.6	0.0	0.0
8	1/1/15	1	1	4	UA	ORD	MCO	1005	afternoon	evening	delayed	133	121	0	0	93	28	0	13.6	0.0	0.0
9	1/1/15	1	1	4	AA	ORD	LGA	733	afternoon	evening	delayed	65	53	0	0	53	0	0	13.6	0.0	0.0
10	1/1/15	1	1	4	OO	ORD	DCA	612	afternoon	evening	delayed	35	15	0	0	15	0	0	13.6	0.0	0.0

b. Canceled flights **(# of rows: 1,303)**

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	DATE	MONTI	DAY	DAY_(	AIRLIN	ORIGI	DESTI	FLIGH	SCHEDUL	SCHEDUL	STATUS	CANCELLATION_F	AWND	PRCP	SNOW	TAVG	Fog	Heavy	Thund	Ice_pe	Hail	Glaze	Haze	Blowin	Torna	Damag
2	1/1/15	1	1	4	UA	ORD	PHL	678	morning	afternoon	cancelled	airline/carrier	13.6	0.0	0.0	20.0	0	0	0	0	0	0	0	0	0	0
3	1/3/15	1	3	6	F9	ORD	ATL	606	morning	morning	cancelled	airline/carrier	4.7	0.2	0.3	31.0	1	0	0	1	0	1	0	0	0	0
4	1/3/15	1	3	6	AA	ORD	SFO	1846	morning	afternoon	cancelled	airline/carrier	4.7	0.2	0.3	31.0	1	0	0	1	0	1	0	0	0	0
5	1/3/15	1	3	6	AA	ORD	MCO	1005	afternoon	afternoon	cancelled	airline/carrier	4.7	0.2	0.3	31.0	1	0	0	1	0	1	0	0	0	0
6	1/3/15	1	3	6	AA	ORD	LGA	733	afternoon	afternoon	cancelled	airline/carrier	4.7	0.2	0.3	31.0	1	0	0	1	0	1	0	0	0	0
7	1/3/15	1	3	6	UA	ORD	LAX	1744	afternoon	evening	cancelled	airline/carrier	4.7	0.2	0.3	31.0	1	0	0	1	0	1	0	0	0	0
8	1/4/15	1	4	7	MQ	ORD	ATL	606	morning	morning	cancelled	airline/carrier	14.2	0.6	0.8	29.0	1	0	0	0	0	1	1	1	0	0
9	1/4/15	1	4	7	AA	ORD	LGA	733	morning	morning	cancelled	airline/carrier	14.2	0.6	0.8	29.0	1	0	0	0	0	1	1	1	0	0
10	1/4/15	1	4	7	AA	ORD	DCA	612	afternoon	afternoon	cancelled	airline/carrier	14.2	0.6	0.8	29.0	1	0	0	0	0	1	1	1	0	0

Airlines in the data

IATA	AIRLINE
UA	United Air Lines Inc.
AA	American Airlines Inc.
US	US Airways Inc.
F9	Frontier Airlines Inc.
B6	JetBlue Airways
OO	Skywest Airlines Inc.
AS	Alaska Airlines Inc.
NK	Spirit Air Lines
DL	Delta Air Lines Inc.
EV	Atlantic Southeast Airlines
MQ	American Eagle Airlines Inc.
VX	Virgin America

Airports in the data

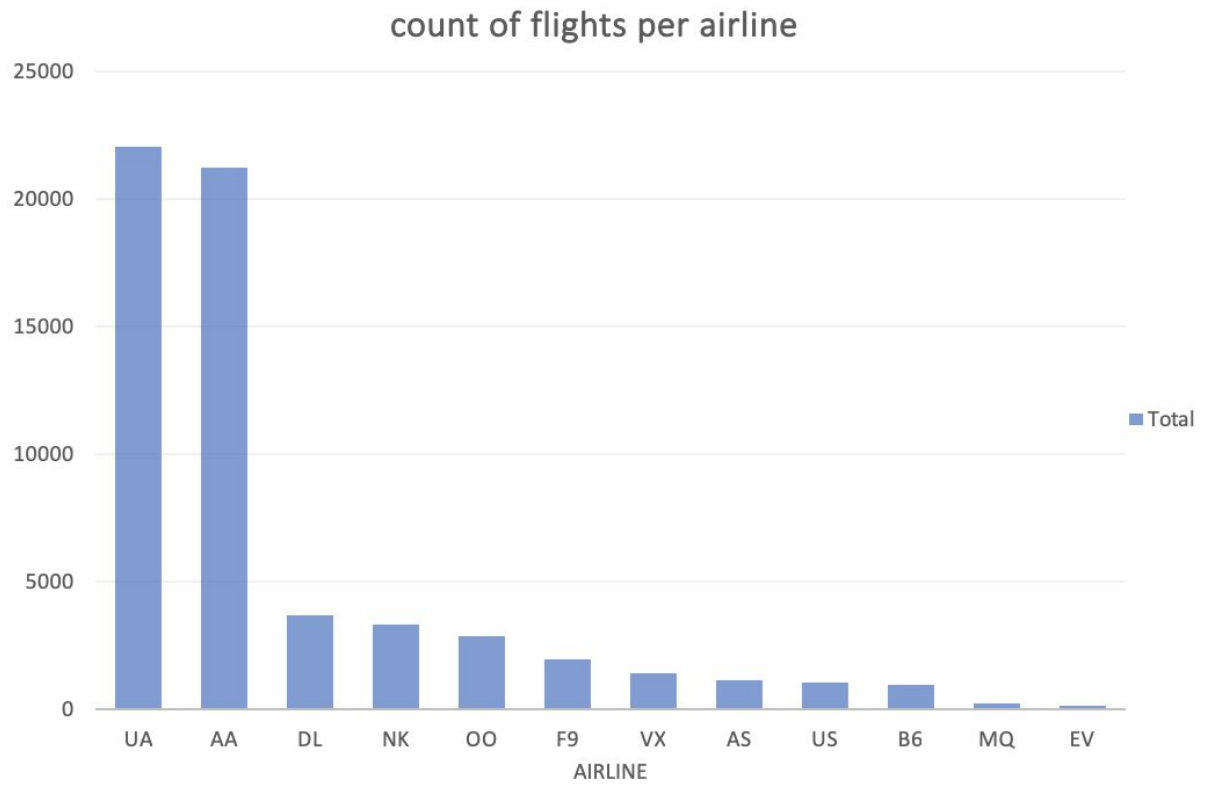
IATA	AIRPORT	CITY	STATE
ATL	Hartsfield-Jackson	Atlanta	GA
BOS	Gen. Edward Lawr	Boston	MA
DCA	Ronald Reagan W	Arlington	VA
LAS	McCarran Internati	Las Vegas	NV
LAX	Los Angeles Intern	Los Angele	CA
LGA	LaGuardia Airport (	New York	NY
MCO	Orlando Internatior	Orlando	FL
ORD	Chicago O'Hare Int	Chicago	IL
PHL	Philadelphia Intern	Philadelph	PA
SEA	Seattle-Tacoma Int	Seattle	WA
SFO	San Francisco Inte	San Franci	CA

A

DESCRIPTIVE ANALYSIS

Objective: use descriptive methods to learn from and understand past data

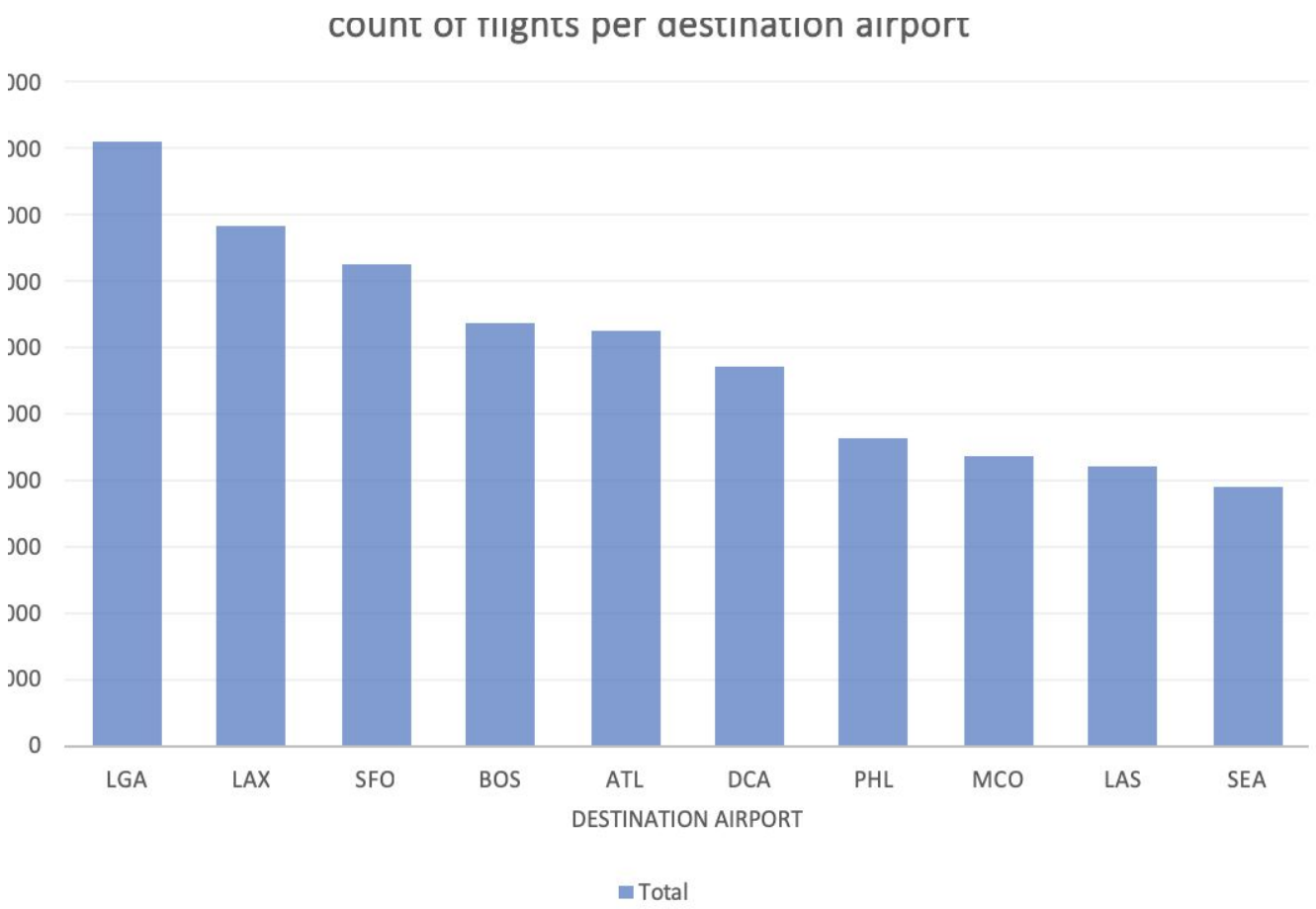
Row Labels	Count of A
UA	22012
AA	21217
DL	3627
NK	3328
OO	2864
F9	1892
VX	1373
AS	1110
US	1021
B6	954
MQ	181
EV	66
Grand Total	**59645**



→ Since Chicago airport (ORD) is known to be the largest hub for United Airlines and American Airlines, this graph clearly shows that the count of flights for these airlines are the highest

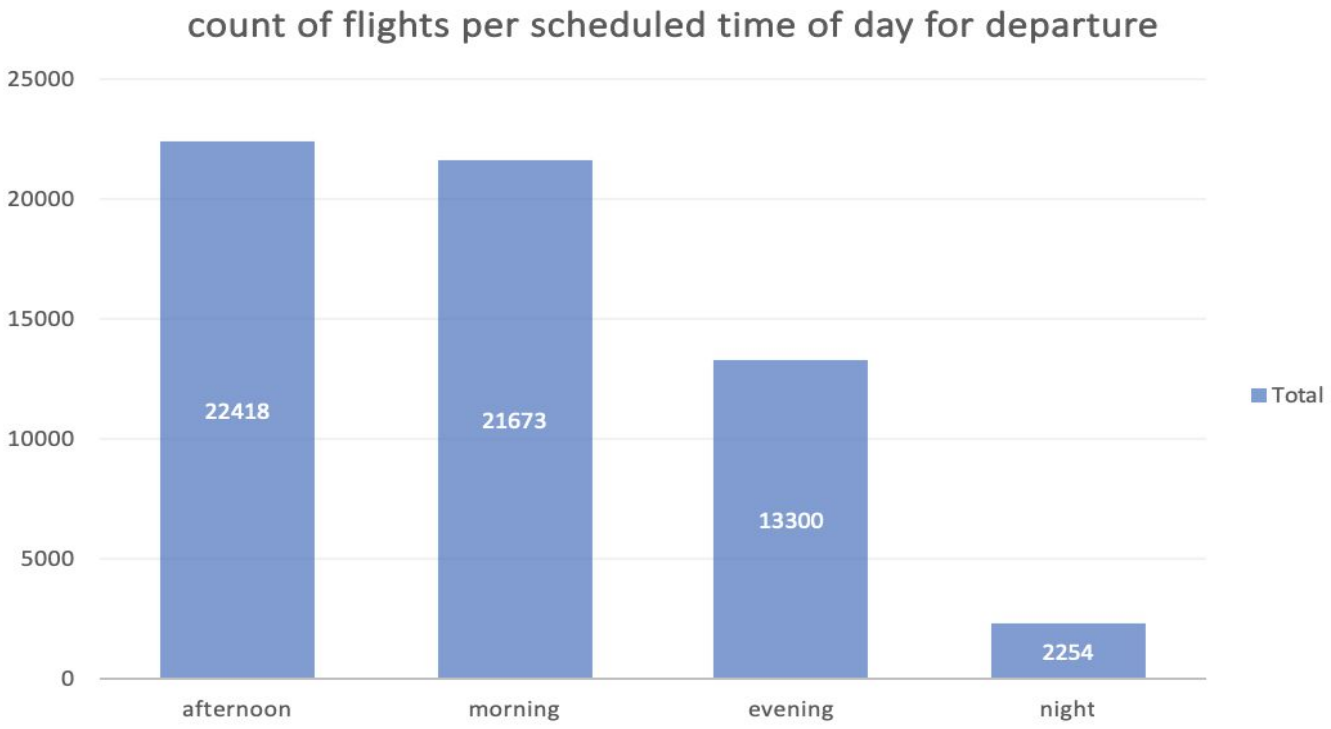
How many flights per airline are there in the data?

Row Labels	Count of
LGA	9101
LAX	7826
SFO	7273
BOS	6378
ATL	6272
DCA	5716
PHL	4617
MCO	4350
LAS	4217
SEA	3895
Grand Total	**59645**



→ The highest number of flights leaving from Chicago airport are the ones departing to LGA, LAX and SFO

How many flights per scheduled destination are there in the data?



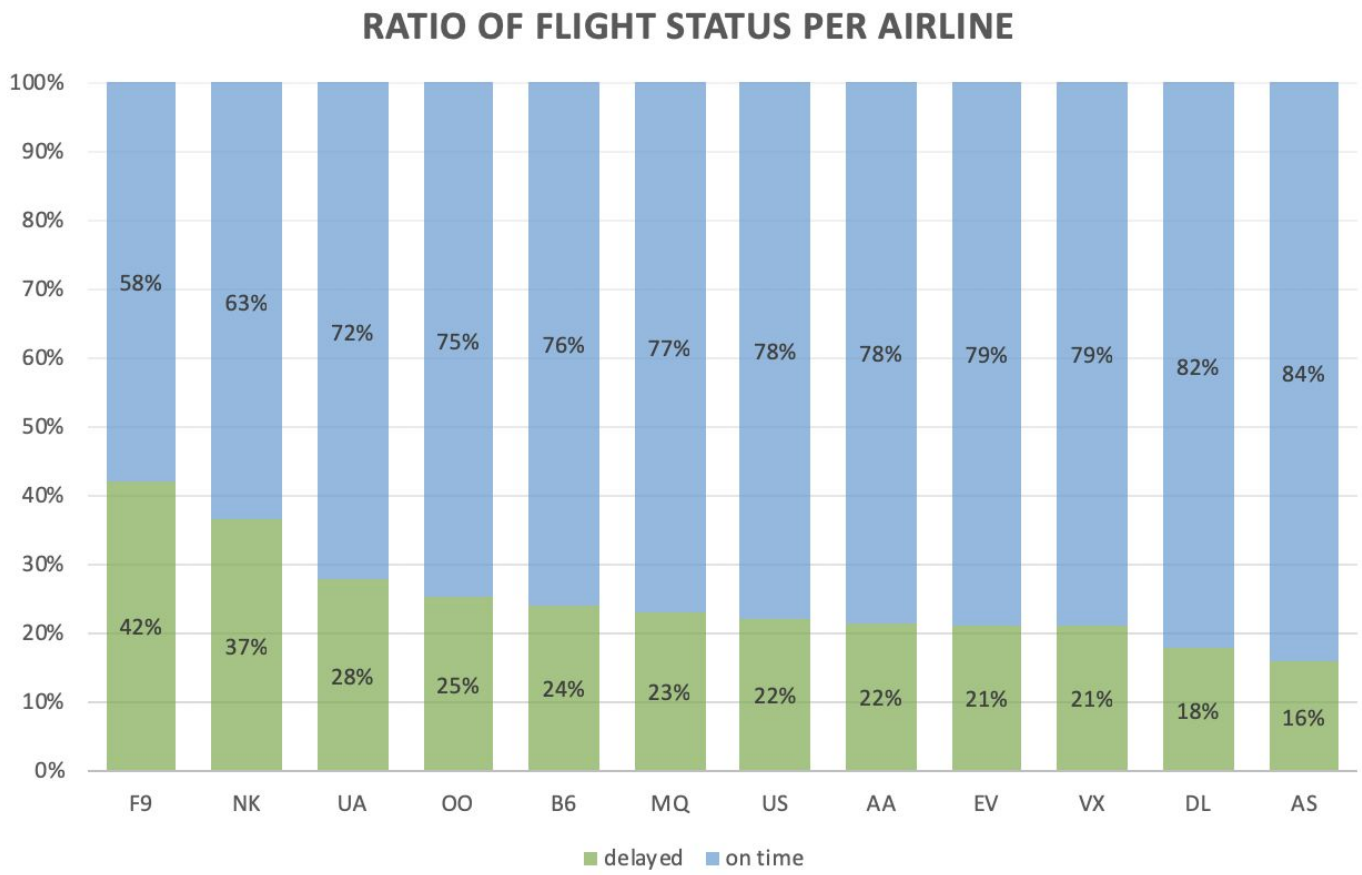
Row Labels	Count of $
afternoon	22418
morning	21673
evening	13300
night	2254
Grand Total	**59645**

→ The highest number of flights leave from Chicago airport are in the afternoon and morning

1. Which airlines typically have significant flight delays? (1) → SUMMARY STATISTICS

AIRLINE	delayed flights count	on time flight counts	delay mean	delay median	delay st dev	delay min	delay max	% delayed flights	% on time flights
AA	4569	16648	3.35	-8.00	41.18	-63	859	22%	78%
AS	176	934	2.30	-6.00	45.42	-51	820	16%	84%
B6	230	724	12.35	-4.00	60.27	-38	1002	24%	76%
DL	646	2981	6.17	-8.00	62.88	-39	1184	18%	82%
EV	14	52	4.24	-9.00	35.28	-30	152	21%	79%
F9	798	1094	27.66	8.00	62.73	-48	762	42%	58%
MQ	42	139	8.53	-5.00	45.76	-32	277	23%	77%
NK	1216	2112	18.57	4.00	46.46	-46	405	37%	63%
OO	725	2139	10.88	-1.00	38.92	-35	396	25%	75%
UA	6115	15897	12.06	-2.00	47.79	-55	623	28%	72%
US	227	794	5.44	-6.00	38.95	-41	289	22%	78%
VX	291	1082	5.38	-3.00	38.71	-48	342	21%	79%

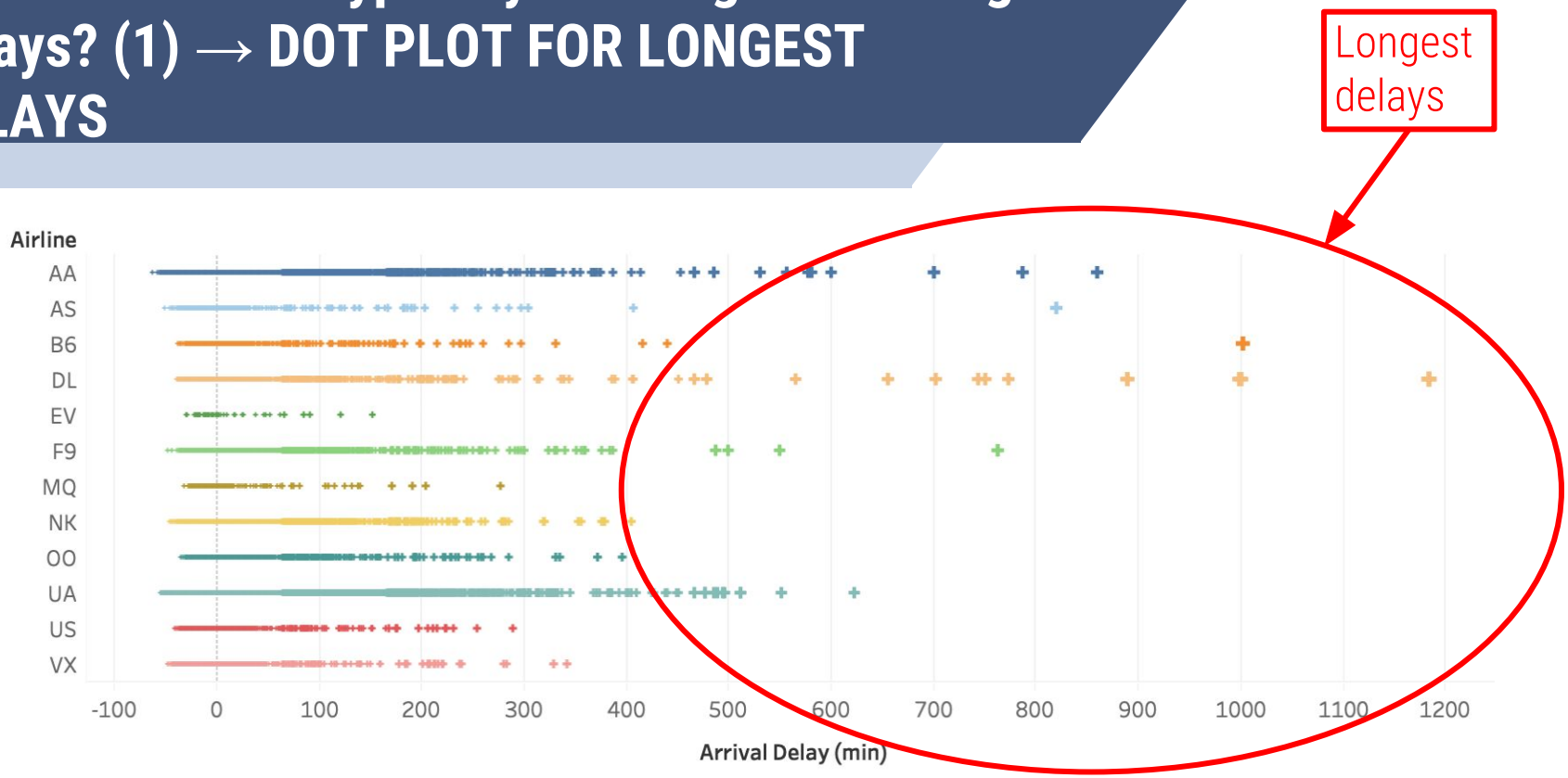
1. Which airlines typically have significant flight delays? (1) → RATIOS OF STATUS PER AIRLINE



Row L	delayed	on time	Grand Total
F9	42%	58%	100%
NK	37%	63%	100%
UA	28%	72%	100%
OO	25%	75%	100%
B6	24%	76%	100%
MQ	23%	77%	100%
US	22%	78%	100%
AA	22%	78%	100%
EV	21%	79%	100%
VX	21%	79%	100%
DL	18%	82%	100%
AS	16%	84%	100%

→ The airlines most commonly delayed are F9 (Frontier), NK (Spirit), and UA (United)

1. Which airlines typically have significant flight delays? (1) → DOT PLOT FOR LONGEST DELAYS



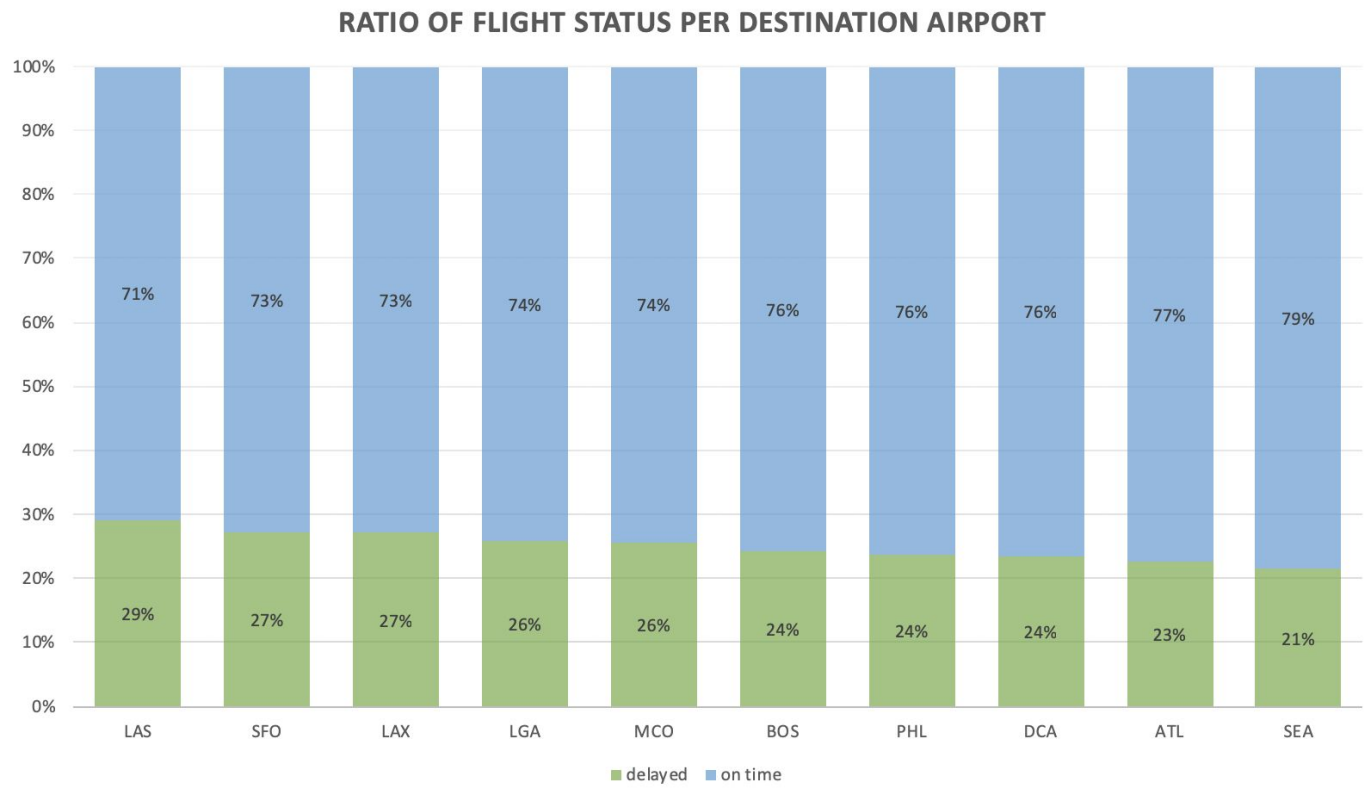
Feel free to use our VLOOKUP query tables to answer questions about delays and airlines:

MIN TO HOUR CONVERSION					
How many minutes?	1180	*(please update the value in the yellow cell -- the value can be positive or negative.)*			
For your information,	*1180*	*minutes is approximately equal to*	*19.7*	*hour(s).*	
Therefore, the flight you selected a delay for arrived approximately			*19.7*	*hour(s)*	***after*** *the original schedule.*
In other words, this flight was delayed.					

AIRLINE CODE TO AIRLINE NAME		
Enter airline code:	DL	*(please update the value in the yellow cell -- enter an airline code.)*
This airline is:	**Delta Air Lines Inc.**	

2. Which scheduled destination airports typically have significant flight delays?

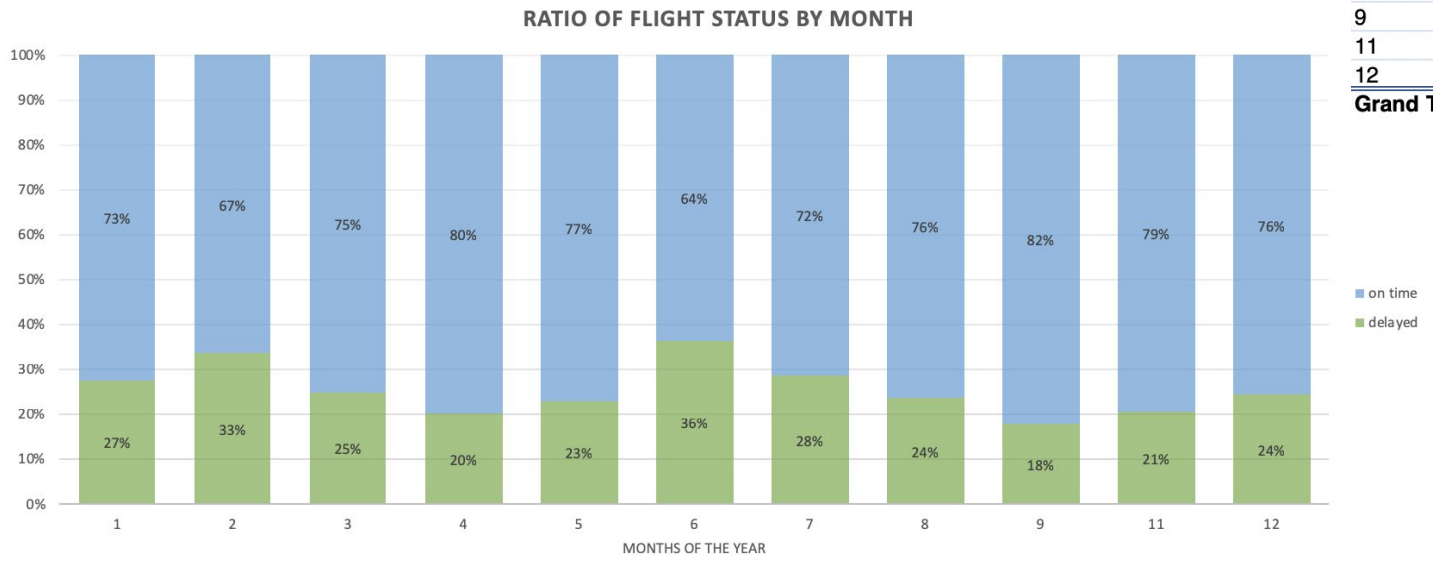
Count of SCHEDULE	Colum		
Row Labels	delayed	on time	Grand Total
LGA	25.85%	74.15%	100.00%
LAX	27.10%	72.90%	100.00%
SFO	27.20%	72.80%	100.00%
BOS	24.22%	75.78%	100.00%
ATL	22.70%	77.30%	100.00%
DCA	23.55%	76.45%	100.00%
PHL	23.76%	76.24%	100.00%
MCO	25.70%	74.30%	100.00%
LAS	29.19%	70.81%	100.00%
SEA	21.46%	78.54%	100.00%
Grand Total	**25.23%**	**74.77%**	**100.00%**



→ The top destination airports for which there tends to be more delays are LAS, SFO, and LAX

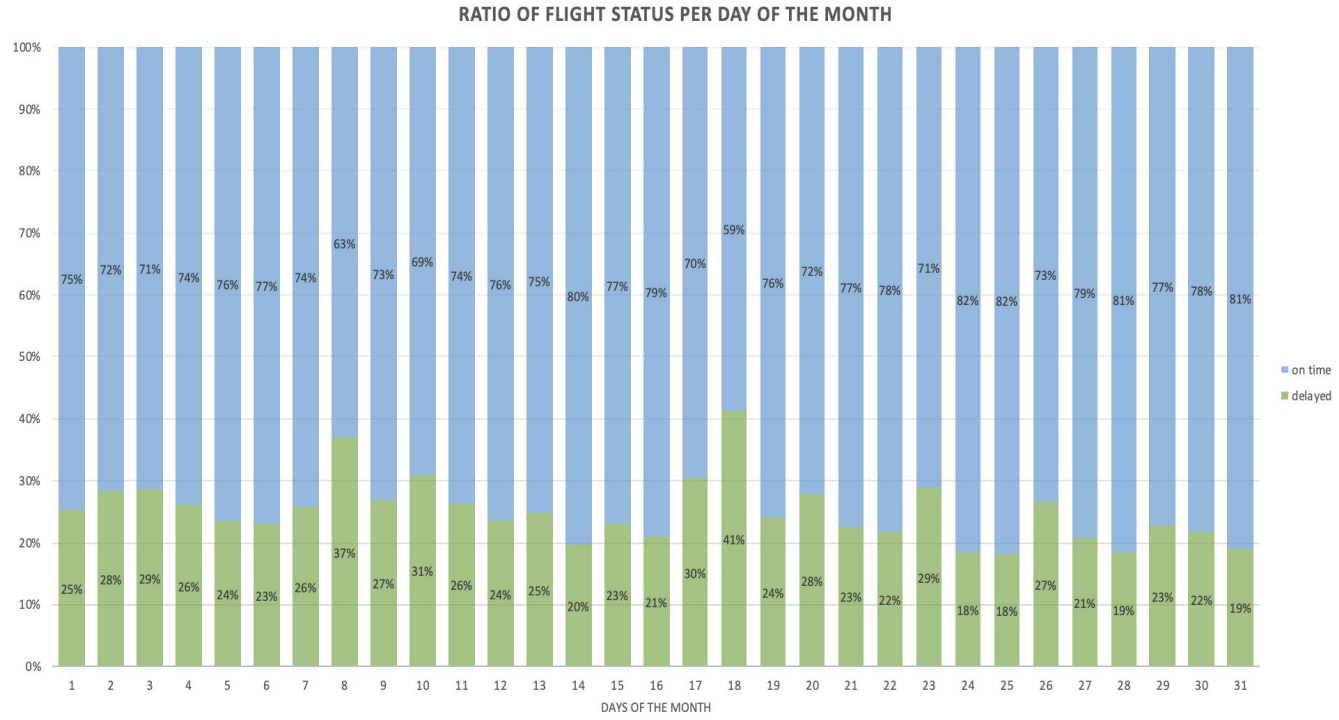
3. Which time of year, month, and day have the most flight delays? (month)

Count of MONTH	Colum		
Row Labels	delayed	on time	Grand Total
1	27%	73%	100%
2	33%	67%	100%
3	25%	75%	100%
4	20%	80%	100%
5	23%	77%	100%
6	36%	64%	100%
7	28%	72%	100%
8	24%	76%	100%
9	18%	82%	100%
11	21%	79%	100%
12	24%	76%	100%
Grand Total	**25%**	**75%**	**100%**



→ The months with most flight delays are December, January, and February (Holiday season), and June and July (summer vacation season)

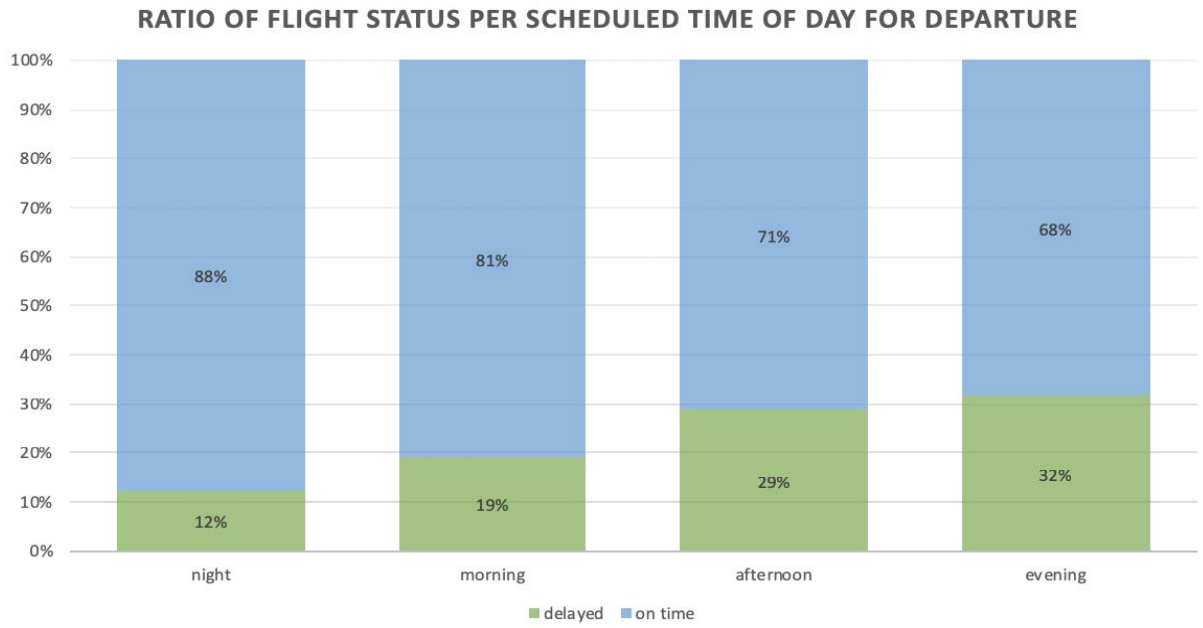
3. Which time of year, month, and day have the most flight delays? (day)



Count	Colum		
Row	delayed	on time	Grand Total
1	25%	75%	100%
2	28%	72%	100%
3	29%	71%	100%
4	26%	74%	100%
5	24%	76%	100%
6	23%	77%	100%
7	26%	74%	100%
8	37%	63%	100%
9	27%	73%	100%
10	31%	69%	100%
11	26%	74%	100%
12	24%	76%	100%
13	25%	75%	100%
14	20%	80%	100%
15	23%	77%	100%
16	21%	79%	100%
17	30%	70%	100%
18	41%	59%	100%
19	24%	76%	100%
20	28%	72%	100%
21	23%	77%	100%
22	22%	78%	100%
23	29%	71%	100%
24	18%	82%	100%
25	18%	82%	100%
26	27%	73%	100%
27	21%	79%	100%
28	19%	81%	100%
29	23%	77%	100%
30	22%	78%	100%
31	19%	81%	100%
Grand	**25%**	**75%**	**100%**

→ The patterns shown do not reveal any particularly useful insights, as some specific months must have skewed the data

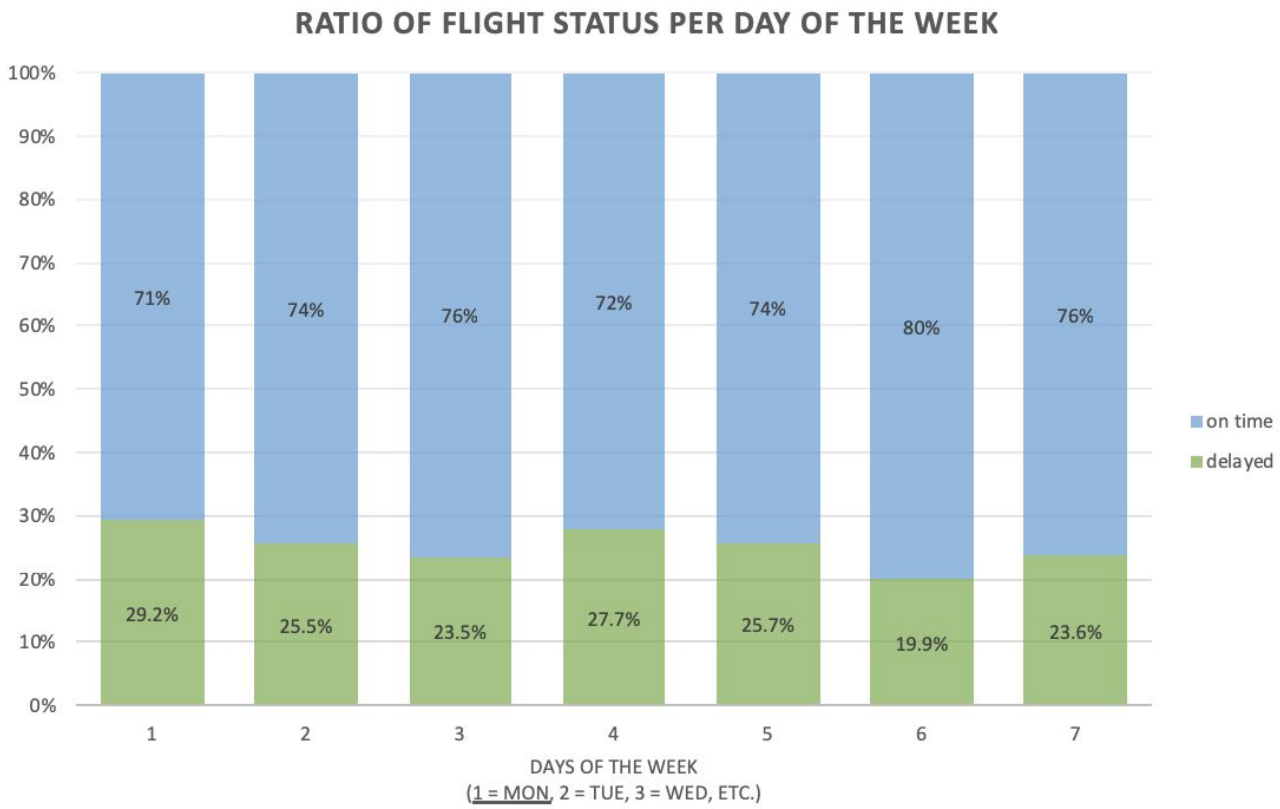
3. Which time of year, month, and day have the most flight delays? (time of day)



Count	Colum		
Row	delayed	on time	Grand Total
night	12.29%	87.71%	100.00%
morning	19.10%	80.90%	100.00%
evening	31.56%	68.44%	100.00%
afterno	28.70%	71.30%	100.00%
Grand	**25.23%**	**74.77%**	**100.00%**

→ Although more flights leave from ORD in the morning and in the afternoon, the percentage of delayed flights is the highest in the evening and in the afternoon (which could be due to the accumulation of late aircraft over the day)

4. Which day(s) of the week typically have the most flight delays?

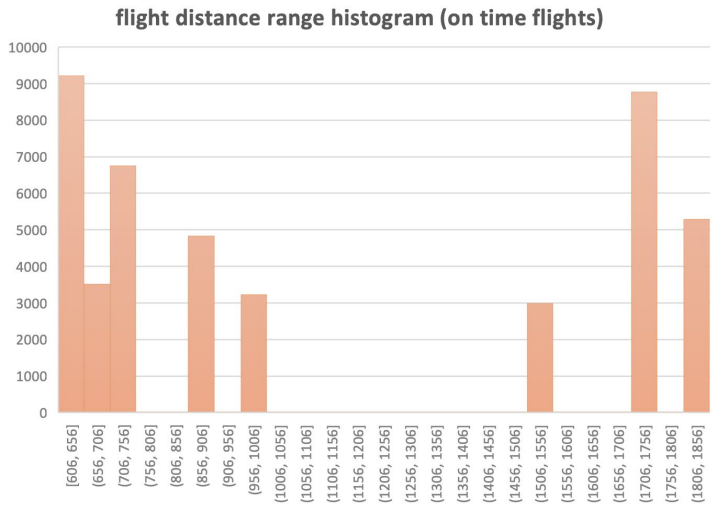


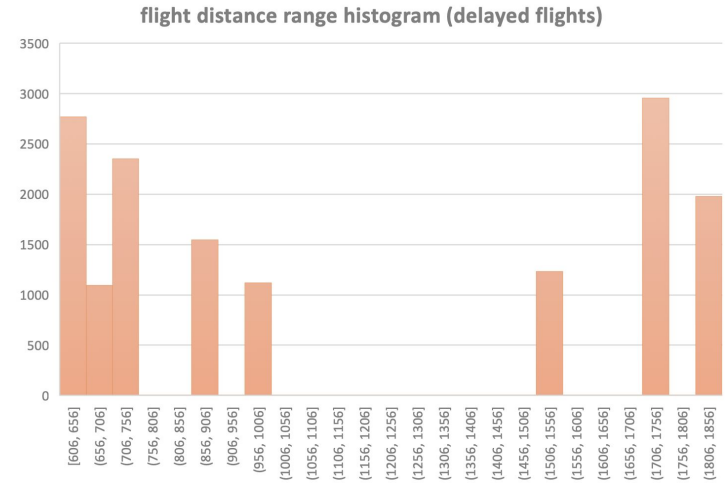
Count	Colum		
Row	delayed	on time	Grand Total
1	29.2%	71%	100%
2	25.5%	74%	100%
3	23.5%	76%	100%
4	27.7%	72%	100%
5	25.7%	74%	100%
6	19.9%	80%	100%
7	23.6%	76%	100%
Grand	**25%**	**75%**	**100%**

→ Flights tend to be delayed more on Mondays, Thursdays, and Fridays in comparison to other days of the week

5. Which flight distance range(s) show(s) the largest number of flight delays?

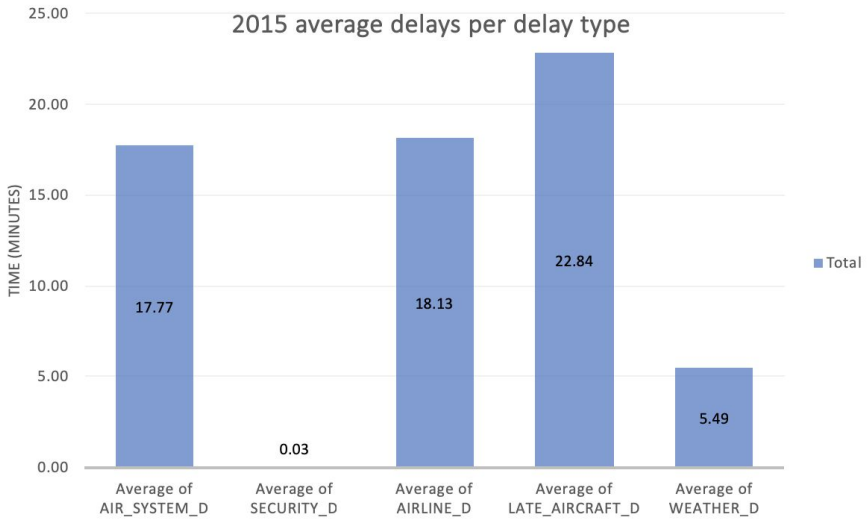
→ These histograms show that the distributions of flight distance ranges for on time flights and delayed flights are similar





6. What are the most common delay reasons? What are the top cancellation reasons? (all data)

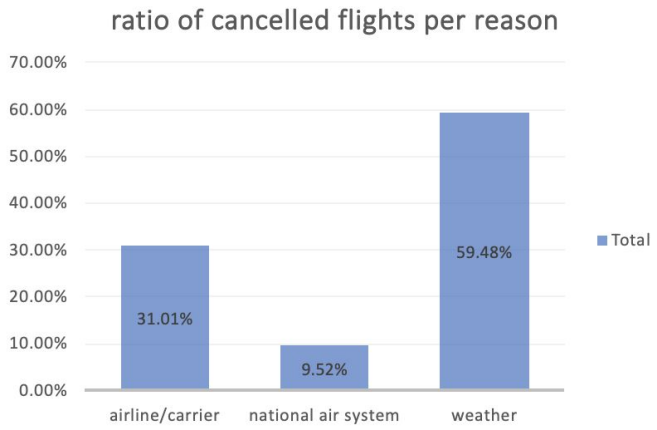
DELAYS



Values	
Average of AIR_SYS	17.77
Average of SECURIT	0.03
Average of AIRLINE_	18.13
Average of LATE_AIF	22.84
Average of WEATHE	5.49

→ In 2015, the average late aircraft delay was 22.84 minutes for flights departing from ORD. The average weather delay was 5.49 minutes.

CANCELLATIONS

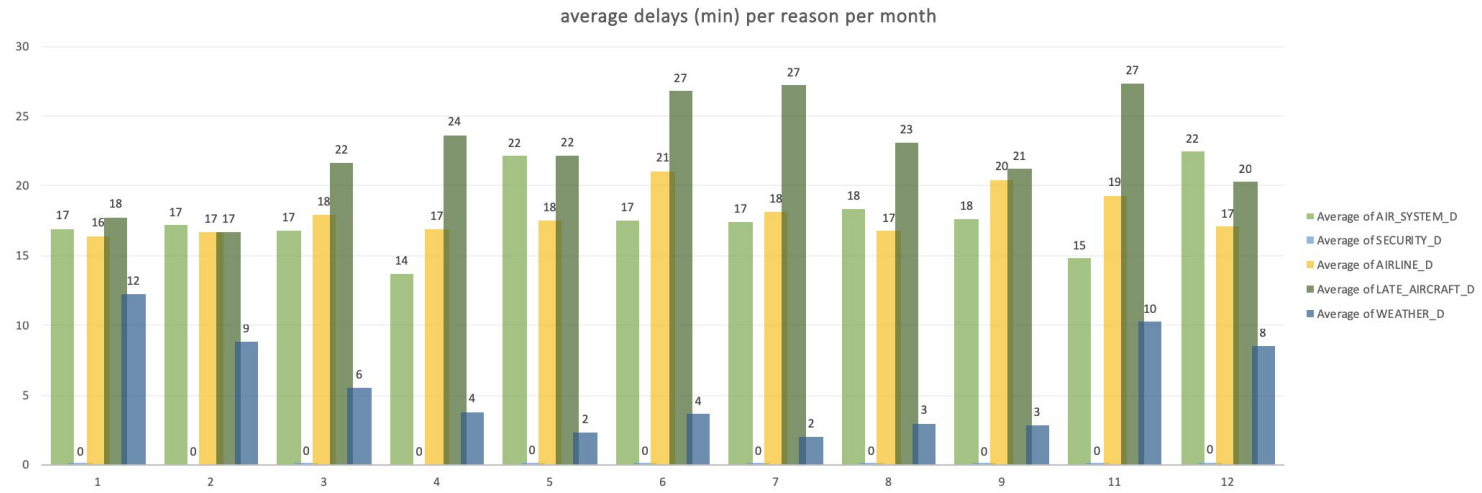


Row Labels	Count of CANCELLA
airline/carrier	31.01%
national air system	9.52%
weather	59.48%
Grand Total	**100.00%**

→ Weather contributes most to aircraft cancellations

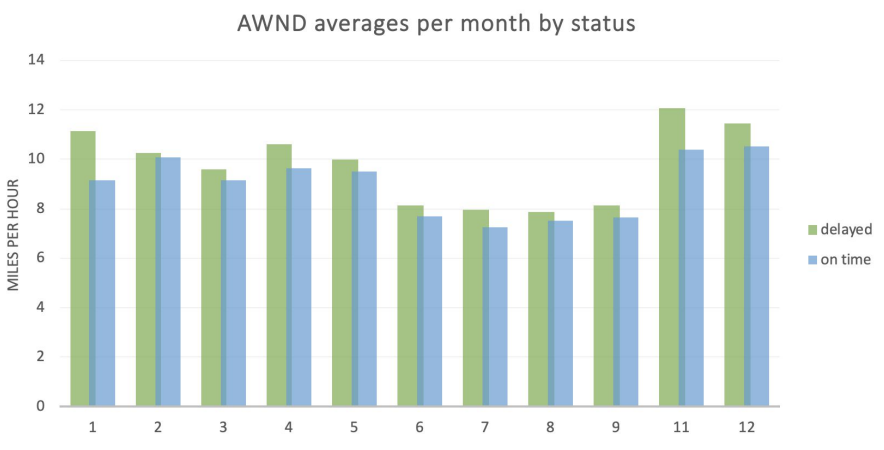
6. What are the most common delay reasons? What are the top cancellation reasons? (by month)

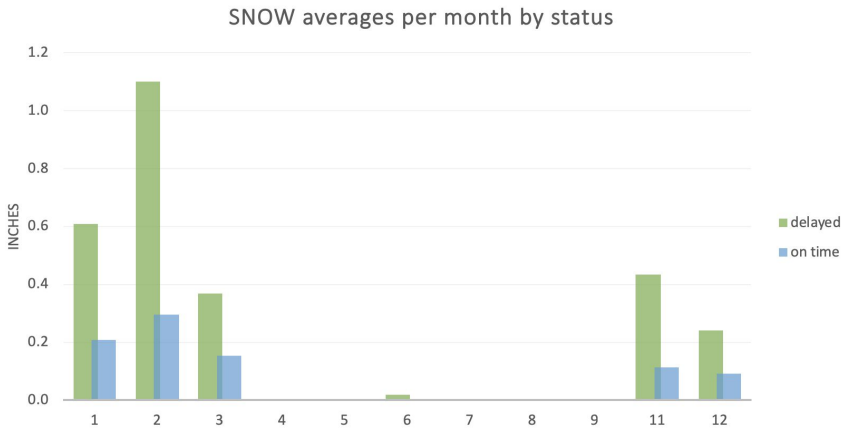
DELAYS

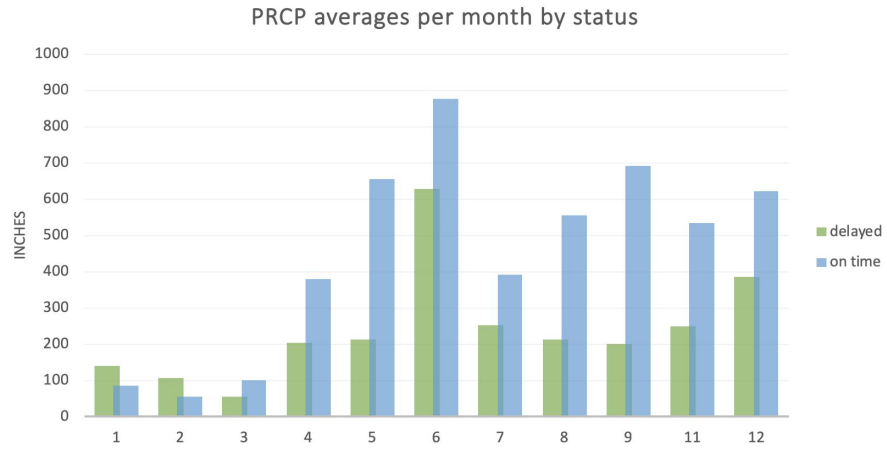


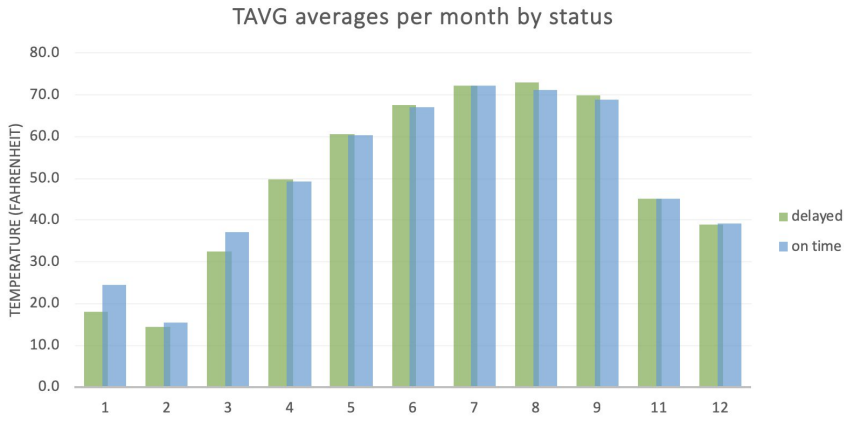
CANCELLATIONS

7. What was the weather data information by month?



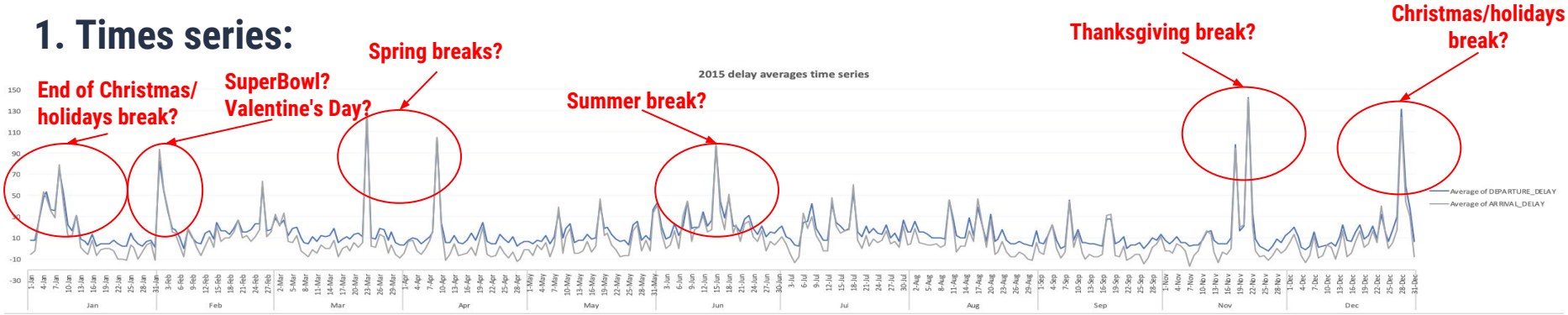






7. Time series and 3D map

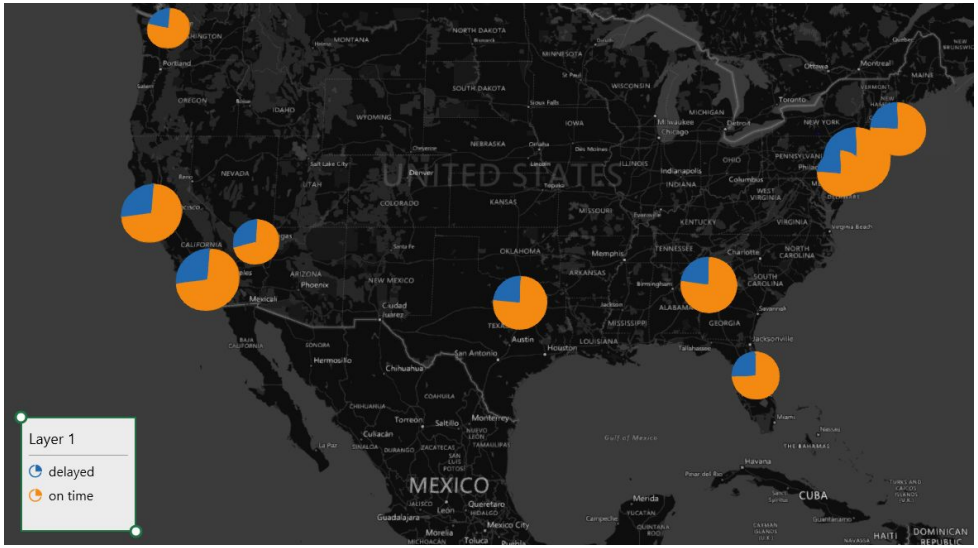
1. Times series:



→ The data is **non-stationary** (it shows no trend but has high volatility, and different cyclicality).

2. 3D map:

Orange = on time
Blue = delayed



B

FORECASTING 

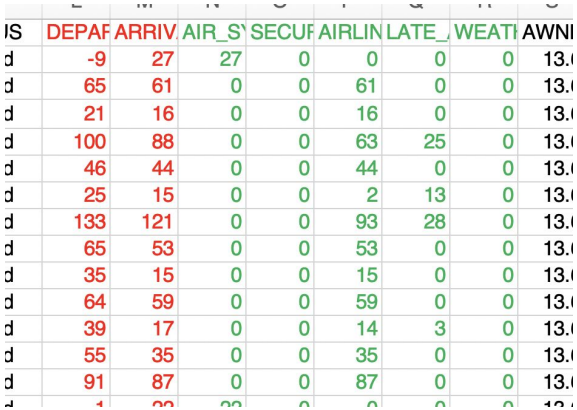
OBJECTIVE: Use regression models to discover good predictors of flight delays

Regression overview

We attempted to understand the **relationship** between **quantitative** variables in the data and conducting a **correlation analysis** of the second dataset (the one with a breakdown of delays)

The difference between **departure delay** vs. **arrival delay** was quite unclear. We conducted several experiments predicting both and the adj R^2 would always be higher when predicting ***arrival delays***

	FLIGHT_	DEPART	ARRIVAL	AIR_SYS	SECURI	AIRLINE_	LATE_AI	WEATHE	AWND	PRCP	SNOW	TAVG
FLIGHT_DISTANCE	1.00											
DEPARTURE_DELAY	-0.04	1.00										
ARRIVAL_DELAY	-0.05	0.95	1.00									
AIR_SYSTEM_D	-0.08	0.14	0.30	1.00								
SECURITY_D	0.02	-0.02	-0.01	-0.01	1.00							
AIRLINE_D	0.02	0.51	0.48	-0.15	-0.02	1.00						
LATE_AIRCRAFT_D	-0.04	0.61	0.58	-0.15	-0.02	-0.05	1.00					
WEATHER_D	0.00	0.34	0.35	-0.04	-0.01	-0.07	0.00	1.00				
AWND	-0.03	0.07	0.09	0.06	0.00	-0.02	0.04	0.10	1.00			
PRCP	-0.02	0.09	0.09	0.00	-0.01	0.02	0.07	0.06	0.17	1.00		
SNOW	-0.02	0.08	0.09	-0.02	-0.01	0.01	0.03	0.17	0.14	0.21	1.00	
TAVG	0.04	0.00	0.01	0.02	0.00	0.02	0.06	-0.13	-0.27	0.13	-0.33	1.00



IS	DEPAI	ARRIV	AIR_S	SECUI	AIRLIN	LATE_	WEATI	AWNI
d	-9	27	27	0	0	0	0	13.(
d	65	61	0	0	61	0	0	13.(
d	21	16	0	0	16	0	0	13.(
d	100	88	0	0	63	25	0	13.(
d	46	44	0	0	44	0	0	13.(
d	25	15	0	0	2	13	0	13.(
d	133	121	0	0	93	28	0	13.(
d	65	53	0	0	53	0	0	13.(
d	35	15	0	0	15	0	0	13.(
d	64	59	0	0	59	0	0	13.(
d	39	17	0	0	14	3	0	13.(
d	55	35	0	0	35	0	0	13.(
d	91	87	0	0	87	0	0	13.(

MODELING OVERVIEW

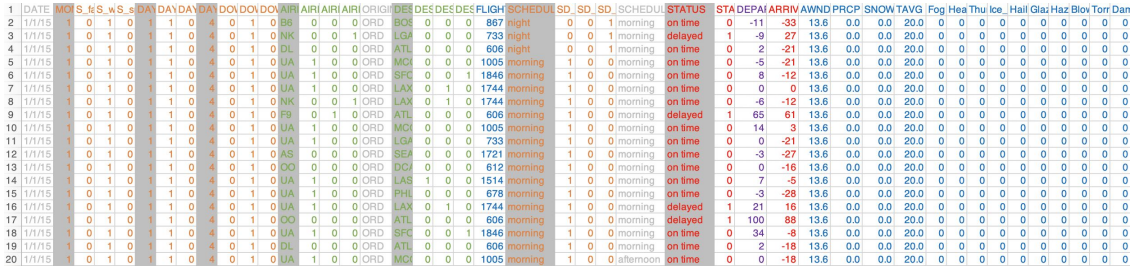
1. **Logistic regression** using all data (**Y = status**)
2. **Simple linear regression** (X=departure delay, **Y=arrival delay**)
3. **Multiple linear regression** (stepwise) (X=weather; weather+airlines(Dummy); weather+airlines(Dummy)+dayofweek(Dummy), etc. **Y=arrival delay**)

Feature engineering and dummy transformation

After discovering which airlines, days of week, destination airports, etc. had **the most significant delays** in the descriptive section and pivot tables, we decided to create dummy variables for the top #3 values with most delays in each category (e.g. AIRLINE: UA, F9, NK)

VARIABLE	DUMMIES CREATED FOR	BASELINE
MONTH	S_fall, S_winter, S_spring	S_summer
DAY	DAY_earlymonth, DAY_latemonth	DAY_middlemonth
DAY_OF_WEEK	DOW_mon, DOW_thu, DOW_fri	All other days of the week
AIRLINE	AIRLINE_UA, AIRLINE_F9, AIRLINE_NK	All other airlines in the data
DESTINATION	DEST_LAS, DEST_LAX, DEST_SFO	All other destination airports in the data
SCHEDULED_DEPARTURE	SD_morning, SD_evening, SD_night	SD_afternoon

Feature engineered dataset overview:



I. Logistic regression: *predicting flight status using all data* (1)

Logistic Regression for STATUS_D(1=delayed)

Summary Measures

Null Deviance	67381.9
Model Deviance	61618.5
Improvement	5763.47
p-Value	< 0.0001

https://stats.idre.ucla.edu/other/mult-pkg/faq/q

0.09 =1-modeldeviance/nulldeviance

Regression Coefficient	Coefficien	Standar Error	Wald Value	p-Value	Lower Limit	Upper Limit	Exp(Coef
Constant	-0.5241	0.1	-6.4	< 0.0001	7.00E-0	-0.4	0.6
S_fall	-0.788	0	-22.5	< 0.0001	9.00E-0	-0.7	0.5
S_winter	-0.863	0	-17.9	< 0.0001	-1	-0.8	0.4
S_spring	-0.647	0	-19	< 0.0001	-0.7	-0.6	0.5
DAY_earlymonth	-0.025	0	-1	0.3378	-0.1	0	1
DAY_latemonth	-0.214	0	-8.5	< 0.0001	-0.3	-0.2	0.8
DOW_mon	0.23	0	7.9	< 0.0001	0.2	0.3	1.3
DOW_thu	0.238	0	8.1	< 0.0001	0.2	0.3	1.3
DOW_fri	0.266	0	9	< 0.0001	0.2	0.3	1.3
AIRLINE_UA	0.355	0	16.4	< 0.0001	0.3	0.4	1.4
AIRLINE_F9	0.935	0.1	18	< 0.0001	0.8	1	2.5
AIRLINE_NK	0.948	0	22.6	< 0.0001	0.9	1	2.6
DEST_LAS	0.052	0	1.1	0.2759	0	0.1	1.1
DEST_LAX	0.16	0	3.4	0.0006	0.1	0.3	1.2
DEST_SFO	0.199	0.1	3.9	< 0.0001	0.1	0.3	1.2
FLIGHT_DISTANCE	0	0	-3.3	0.001	0	0	1
SD_morning	-0.589	0	-24.8	< 0.0001	-0.6	-0.5	0.6
SD_evening	0.098	0	3.9	0.0001	0	0.1	1.1
SD_night	-1.268	0.1	-18.3	< 0.0001	-1.4	-1.1	0.3
AWND	0.046	0	13.9	< 0.0001	0	0.1	1
PRCP	0.097	0	2.4	0.0177	0	0.2	1.1
SNOW	0.231	0	9.9	< 0.0001	0.2	0.3	1.3
TAVG	-0.016	0	-17.2	< 0.0001	0	0	1
Fog	0.25	0	10.7	< 0.0001	0.2	0.3	1.3
Heavy_fog	0.53	0	11.4	< 0.0001	0.4	0.6	1.7
Thunder	0.62	0	20.2	< 0.0001	0.6	0.7	1.9
Ice_pellets	0.568	0.1	7	< 0.0001	0.4	0.7	1.8
Hail	0.032	0.1	0.4	0.6943	-0.1	0.2	1
Glaze	0.218	0.1	3.5	0.0005	0.1	0.3	1.2
Haze	-0.003	0	-0.1	0.9129	-0.1	0	1
Blowing_snow	1.002	0.1	11.8	< 0.0001	0.8	1.2	2.7
Tornado	1.555	0.2	7.8	< 0.0001	1.2	1.9	4.7
Damaging_winds	0.526	0.1	7.1	< 0.0001	0.4	0.7	1.7

GOAL: *Understand how the data would affect the probability of getting delayed through the coefficients, and if the explanatory variables are good predictors for the classification task*

Interpretation:

- Variables that increase P(Y=1(delay)) are in **green**.
- Variables that decrease P(Y=1(delay)) are in **red**.
- The brighter the color, the higher the increase or decrease in probability of P(Y=1(delay)).

Examples:

- The variables that increase the probability of delay the most are: Tornado (dummy), blowing_snow (dummy), airline_NK (dummy), etc.
- The variables that decrease the probability of delay the most are: SD_night, S_winter, S_fall, S_spring, SD_morning, etc.

(always in comparison to the ***baselines***, and ***holding all other variables constant***.)

I. Logistic regression: *predicting flight status using all data* (2)

	Median	25th Percenti	75th Percenti	Logit (25th)	Logit (75th)	Prob (25th)	Prob (75th)	Difference
AWND	8.503	6.637	11.183	-1.090	-0.883	0.252	0.293	0.041
PRCP	0.008	0.000	0.133	-1.006	-0.993	0.268	0.270	0.003
SNOW	0.000	0.000	0.000					
TAVG	55.000	37.000	69.000	-0.720	-1.226	0.327	0.227	-0.100

Equation: **ln(p/(1-p)) = -0.52414074 + 0.046*AWND + 0.097*PRCP + 0.231*SNOW - 0.016*TAVG**

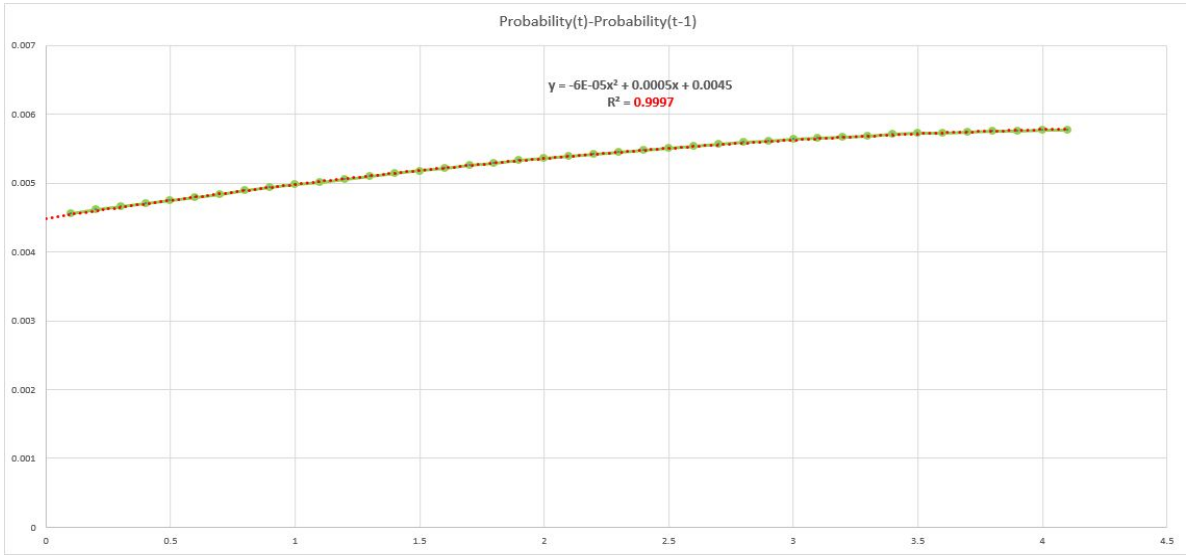
- In the equation above, to study the effect of one variable on the probability of delay, all the dummy variables are held constant (at baseline, i.e. all dummies =0) and median values were used for the remaining qualitative variables to keep them constant
  e.g. while studying the effect of 'AWND' on probability of delay, we used median values of 'PRCP, 'SNOW' and 'TAVG'

Interpretation:

1. While keeping everything else constant, an **increase of 1 mile per hour** in AWND will **increase the probability** of flight being delayed by **0.041**.
2. While keeping everything else constant, **an increase of 1 inch** in PRCP will **increase the probability** of flight being delayed by **0.003**.
3. While keeping everything else constant, **an increase of 1℉** in TAVG will **decrease the probability** of flight being delayed by **0.100**.

I. Logistic regression: *predicting flight status using all data* (3)

Snow(mm)	logits	Probability	Probability(t)-Probability(t-1)
0	-1.0049431	0.26797066	
0.1	-0.9818219	0.27253043	0.004559771
0.2	-0.9587007	0.27713841	0.004607985
0.3	-0.9355795	0.28179413	0.004655717
0.4	-0.9124583	0.28649706	0.004702931
0.5	-0.8893371	0.29124665	0.004749591
0.6	-0.8662158	0.29604231	0.00479566
0.7	-0.8430946	0.30088342	0.004841103
0.8	-0.8199734	0.3057693	0.004885882
0.9	-0.7968522	0.31069926	0.004929962
1	-0.773731	0.31567257	0.004973305
1.1	-0.7506098	0.32068844	0.005015876
1.2	-0.7274886	0.32574608	0.005057638
1.3	-0.7043674	0.33084463	0.005098555
1.4	-0.6812462	0.33598323	0.005138591
1.5	-0.658125	0.34116094	0.005177712
1.6	-0.6350038	0.34637682	0.005215881
1.7	-0.6118825	0.35162988	0.005253064
1.8	-0.5887613	0.35691911	0.005289228
1.9	-0.5656401	0.36224345	0.005324338
2	-0.5425189	0.36760181	0.005358362
2.1	-0.5193977	0.37299308	0.005391268
2.2	-0.4962765	0.3784161	0.005423023
2.3	-0.4731553	0.3838697	0.005453598
2.4	-0.4500341	0.38935266	0.005482962
2.5	-0.4269129	0.39486375	0.005511087
2.6	-0.4037917	0.40040169	0.005537944
2.7	-0.3806705	0.4059652	0.005563508
2.8	-0.3575492	0.41155295	0.005587752
2.9	-0.334428	0.4171636	0.005610652
3	-0.3113068	0.42279579	0.005632184
3.1	-0.2881856	0.42844811	0.005652326
3.2	-0.2650644	0.43411917	0.005671059
3.3	-0.2419432	0.43980753	0.005688361
3.4	-0.218822	0.44551175	0.005704216
3.5	-0.1957008	0.45123036	0.005718606
3.6	-0.1725796	0.45696187	0.005731517
3.7	-0.1494584	0.46270481	0.005742934
3.8	-0.1263372	0.46845765	0.005752847
3.9	-0.1032159	0.4742189	0.005761244
4	-0.0800947	0.47998701	0.005768116
4.1	-0.0569735	0.48576047	0.005773456



1. The data was sparse, most of it was 0. So we monitored the change in probability of delay for 0.1mm increment in snow.
2. The snow level (in mm) quadratically explained the change in probability.
3. For e.g., putting 0.3 as x in equation will give the probability change in delay when snow increases from 0.2mm to 0.3mm

Equation: $y = -6E\text{-}05x^2 + 0.0005x + 0.0045$

$R^2 = 0.9997$

I. Logistic regression: *predicting flight status using all data* (4)

Scenario #1

Vijet's uncle wants to go to Los Angeles on Friday to spend Christmas with his family on Sunday. He has a choice between traveling with Frontier Airlines or Delta airlines on Friday evening. The weather forecast indicates that there will only be wind on that night (10 mph) and that the the temperature average for the day will be 32 F.
Which airline should Vijet's uncle travel with?

AIRLINE	logit	prob(Y=1)
Frontier	-0.192124704	**0.45**
Delta	-1.128	**0.24**

He should travel with Delta Airlines.

Here's how the first probability was computed:
z = constant + S_winter*1 + DAY_latemonth*1 + DOW_fri*1+ AIRLINE_F9*1 + DEST_LAX*1 + SD_evening*1 + AWND*10 + TAVG*32
P(Y=1(delayed)) = $1/(1+e^{-z})$

Scenario #2

David's aunt also wants to go to Los Angeles on the day after, Saturday, and she will be traveling with United Airlines. She has a choice between flying in the morning or in the evening. The weather forecast indicates that there will only be blowing snow during the day and that the the temperature average for the day will be 32 F.
At what time of day should David's aunt travel?

TIME OF DAY	logit	prob(Y=1)
morning	-1.179	**0.24**
evening	-0.492	**0.38**

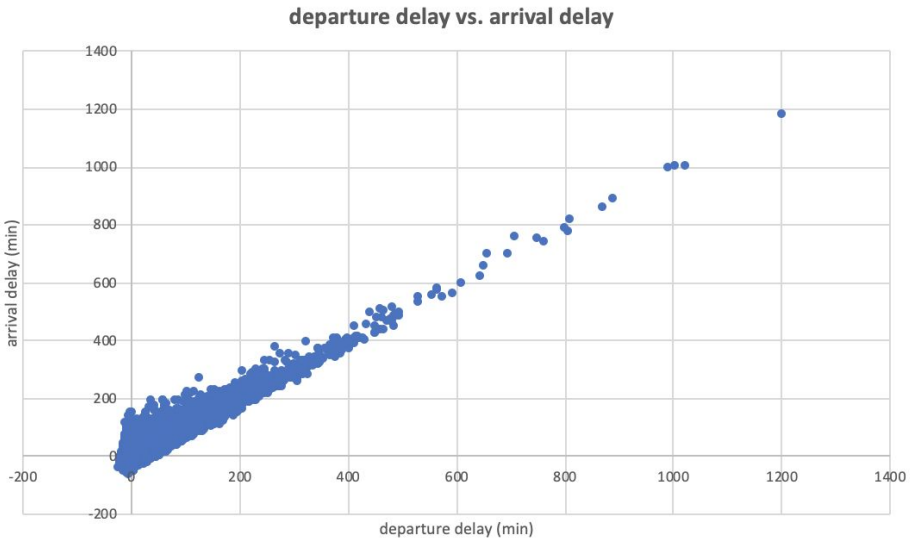
She should travel in the morning.

II. Simple linear regression: *predicting arrival delays using departure delays*

GOAL: *Understand how departure delays affect arrival delays (2 quantitative variables)*

I. Correlation coefficient and scatter plot

r = 0.94 ⟶ strong positive correlation
⟶ strong linear positive relationship



II. Linear regression

Yhat = -7.59 + 1.02*x
R^2 = 0.8923

IV. Interpretation

- A flight with no departure delay (X=0) is predicted to arrive 7.59 minutes early.
- Every extra minute of departure delay is predicted to increase the arrival delay by 1.02 minutes.

Note: *we* attempted to add a quadratic term $(x + x^2)$ to the regression to relax the assumption that the trend was linear (R^2 adj = 0.8924), and a cubic term $(x + x^2 + x^3)$ (R^2 adj = 0.8924). The improvement was neglectable so we stuck with the linear model.

III. Multiple linear regression: *predicting arrival delays using stepwise multiple linear regression* (1)

GOAL: *Understand how weather, airlines, destination airports, scheduled time of day, day of week, etc. all affect arrival delays and if they are good predictors using stepwise regression*

X (predictors) for Y = ARRIVAL_DELAY	adj R-squared
all weather variables	0.0668
FLIGHT_DISTANCE + all weather variables *(+ removed bad weather predictors: Hail (dummy) and Haze (dummy) (they had high p-values))*	0.0673
[...] + AIRLINES	0.0810
[...] + DAY_OF_WEEK	0.0846
[...] + DAY (DAY_earlymonth, etc.)	0.0848
[...] + MONTH (S_winter, S_spring, etc.)	0.0918
[...] + DESTINATION (LAS, LAX, SFO)	0.0919
[...] + SCHEDULED_DEPARTURE (SD_morning, etc.)	0.1033
[...] + INTERACTION VARIABLES	**0.1043**

Interpretation

- Adding more variables consistently helped **improve the predictive power** of the model & **increase adj R^2**
- Although for some chunks of variables (e.g. destination), some categories were **good predictors** (p-value<0.05, e.g. SFO) & some **were not** (p-value>0.05, e.g. LAX, LAS), the adj R^2 still increased
- As a last step, 3 **interactions variables** were added (categorical*quantitative variable): **morning*SNOW, evening*SNOW, night*SNOW**, to allow for the slope of delays to be different when it is snowing during different times of day (morning vs. night)

III. Multiple linear regression: *predicting arrival delays using stepwise multiple linear regression* (2)

Multiple Regression for ARRIVAL_DELAY

Summary

	0.3237	0.1048	**0.1043**	4.4119241	0	0

ANOVA Table	ees of Freeim of Squaran of Squa			F	p-Value
Explained	33	1.00E+07	416927	211.37848	< 0.0001
Unexplained	59611	1.00E+08	1972.4		

Regression Tab	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Upper
Constant	15.005	1.5	10.2	< 0.0001	12.13	17.88
S_fall	-11.229	0.6	-18.8	< 0.0001	-12.4	-10.06
S_winter	-15.311	0.9	-17.5	< 0.0001	-17.02	-13.6
S_spring	-11.873	0.6	-19.4	< 0.0001	-13.07	-10.67
DAY_earlymonth	-1.013	0.5	-2.1	0.0342	-1.95	-0.08
DAY_latemonth	-2.713	0.5	-6	< 0.0001	-3.6	-1.82
DOW_mon	7.679	0.5	14.2	< 0.0001	6.62	8.74
DOW_thu	4.889	0.5	9	< 0.0001	3.83	5.95
DOW_fri	4.645	0.5	8.6	< 0.0001	3.59	5.7
AIRLINE_UA	7.487	0.4	18.9	< 0.0001	6.71	8.26
AIRLINE_F9	21.432	1.1	20	< 0.0001	19.33	23.53
AIRLINE_NK	16.049	0.8	19.3	< 0.0001	14.42	17.68
DEST_LAS	-1.149	0.9	-1.3	0.1922	-2.88	0.58
DEST_LAX	0.856	0.8	1	0.31	-0.8	2.51
DEST_SFO	2.173	0.9	2.4	0.0169	0.39	3.96
FLIGHT_DISTANCE	-0.004	0	-5.9	< 0.0001	-0.01	0
SD_morning	-9.372	0.4	-21.8	< 0.0001	-10.22	-8.53
SD_evening	0.872	0.5	1.8	0.0797	-0.1	1.85
SD_night	-15.639	1	-15.5	< 0.0001	-17.62	-13.66
AWND	1.106	0.1	18.6	< 0.0001	0.99	1.22
PRCP	3.374	0.8	4.1	< 0.0001	1.78	4.97
SNOW	4.637	0.5	9.2	< 0.0001	3.65	5.63
TAVG	-0.223	0	-13.7	< 0.0001	-0.25	-0.19
Fog	3.813	0.4	9.5	< 0.0001	3.03	4.6
Heavy_fog	17.652	0.9	19.2	< 0.0001	15.85	19.46
Thunder	10.082	0.6	17.2	< 0.0001	8.94	11.23
Ice_pellets	15.683	1.6	9.6	< 0.0001	12.48	18.89
Glaze	3.108	1.2	2.7	0.0078	0.82	5.4
Blowing_snow	31.212	1.7	18.3	< 0.0001	27.86	34.56
Tornado	74.911	3.8	19.9	< 0.0001	67.55	82.27
Damaging_winds	9.953	1.5	6.5	< 0.0001	6.95	12.96
SD_morning*SNOW	2.982	0.7	4.3	< 0.0001	1.64	4.33
SD_evening*SNOW	-3.073	0.8	-4	< 0.0001	-4.58	-1.57
SD_night*SNOW	-4.675	1.7	-2.8	0.0058	-8	-1.35

Interpretation of this model *(examples)*

1. Flights that leave from ORD in seasons other than summer (baseline) are predicted to have a shorter arrival delay when holding other variables constant:

e.g. flights that leave in the Fall are predicted to have a delay of 15.01-11.23 = **3.78 min** when all the explanatory variables = 0

2. Flights that leave from ORD on days of the week that are Mondays, Thursdays, and Fridays are predicted to have longer arrival delays than other days (baseline) when holding other variables constant:

e.g. flights that leave on Mondays are predicted to have a delay of 15.00+7.68 = **22.68 min** when all the explanatory variables =0

3. Flights that leave from ORD with severe weather conditions are predicted to have longer delays than with normal weather when holding other variables constant:

e.g. an increase in 1 mph of AWND is predicted to increase the arrival delay by **1.11 min**

4. Flights that leave from ORD while it is snowing are predicted to have different delays based on the time of scheduled departure when holding other variables constant:

e.g. an increase in 1 inch of SNOW fall in the **morning** is predicted to increase the arrival delay by **2.9 min** in comparison to the delay triggered by SNOW in the afternoon

e.g. an increase in 1 inch of SNOW fall in the **evening** is predicted to decrease the arrival delay by **1.57 min** in comparison to the delay triggered by SNOW in the afternoon

Adj R^2: **0.1033**

RESULTS & CONCLUSIONS

SUMMARY

- With **logistic regression**, we were able to determine which factors play a key role in determining P(Y=1(delay))
- With **simple linear regression**, we were able to discover the impact of departure delay on arrival delay
- With **multiple linear regression,** we managed to reach the highest possible adjusted R-squared value by adding variables (quantitative, dummies, interactions) in a stepwise fashion

CONCLUSIONS

Based on the descriptive section, the models implemented, and the findings derived from those models, it can be concluded that:

- Although our task to build a models with a high predictive power was quite **ambitious** considering how little data we had, we found that the majority of variables we added (e.g. airlines, days of week, etc.) were **good predictors** for our linear & logistic regression models
- **Those predictors were found to greatly impact flight delays**

NEXT STEPS

- Collect **more precise weather data** (hourly averages instead of daily averages)
- Retrieve **more data** that could also play a crucial role in impacting flight delays:
  - Airport congestion
  - Fueling,
  - Maintenance Issues
  - Baggage and catering loading
  - Computer Glitches
  - Boarding times
  - Gate changes
  - etc.
- Attempt to implement **new models** for this highly categorical task (e.g. random forests, neural networks)
- Make predictions using a **training**, **validation**, and **testing** splits to evaluate accuracy of the best model

THANK YOU

If you have any questions, feel free to reach out to us.

David Forteguerre (dfortegu@syr.edu)
Vijet Muley (vmuley@syr.edu)
Advait Iyer (aiyer01@syr.edu)
Anisha Kumar (akumar10@syr.edu)
Kshitij Sankesara (kssankes@syr.edu)