Project for "Introduction To Data Analysis"

# Analyzing Historical Airlines Data with Destination to Chicago O'Hare Airport

Min-Huey Wang, Tomoaki Tanaka

*UCSC extention Santa Clara USA, Aug. 25, 2016)*

# Outline

- Motivation
- State of problem
- Dataset
- Assumption
- Finding and modeling
- Discussion and conclusion

# Motivation

- Flight delays impact the US economics.
- It's worthy to study the delay pattern to better understand the causes such that we can help either the air line companies or passengers to minimize the impacts.

National Center of Excellence for Aviation Operations Research

**Total Delay Impact Study**

*A Comprehensive Assessment of the Costs and Impacts of Flight Delay in the United States*

Final Report — October, 2010

*Prepared by:*

Michael Ball, Cynthia Barnhart, Martin Dresner, Mark Hansen, Kevin Neels, Amedeo Odoni, Everett Peterson, Lance Sherry, Antonio Trani, Bo Zou

*With Assistance from:*

Rodrigo Britto, Doug Fearing, Prem Swaroop, Nitish Uman, Vikrant Vaze, Augusto Voltes

# Annual U.S. Impact of Flight Delays (NEXTOR report)

On December 16, 2010, NEXTOR published its revised final report, entitled Total Delay Impact Study: A Comprehensive Assessment of the Costs and Impacts of Flight Delay in the United States.

In 2007, domestic flight delays were found to cost the U.S. economy **$31.2 billion** in 2007, including $8.3 billion in direct costs to airlines, **$16.7 billion in direct costs to passengers**, $2.2 billion from lost demand and $4.0 billion in forgone GDP.

| Line Item Cost Component | Category | $ Billions |
|---|---|---|
| Flight Delay Against Schedule | Airlines | 4.6 |
| Intrinsic Flight Delay due to Schedule Buffer | Airlines | 3.7 |
| Excess Travel Time due to Schedule Buffer | Passengers | 6.0 |
| Passenger Delay Against Schedule: Delayed Flights | Passengers | 4.7 |
| Passenger Delay Against Schedule: Canceled Flights | Passengers | 3.2 |
| Passenger Delay Against Schedule: Missed Connections | Passengers | 1.5 |
| Capacity-Induced Schedule Delay | Passengers | 0.7 |
| Voluntary Early-Departure-Time Adjustment | Passengers | 0.6 |
| Welfare loss due to switch from air to automobile | Shared | 2.0 |
| Externality cost from increased road traffic | Shared | 0.2 |
| Forgone GDP | Shared | 4.0 |
| Total U.S. Cost | All | 31.2 |

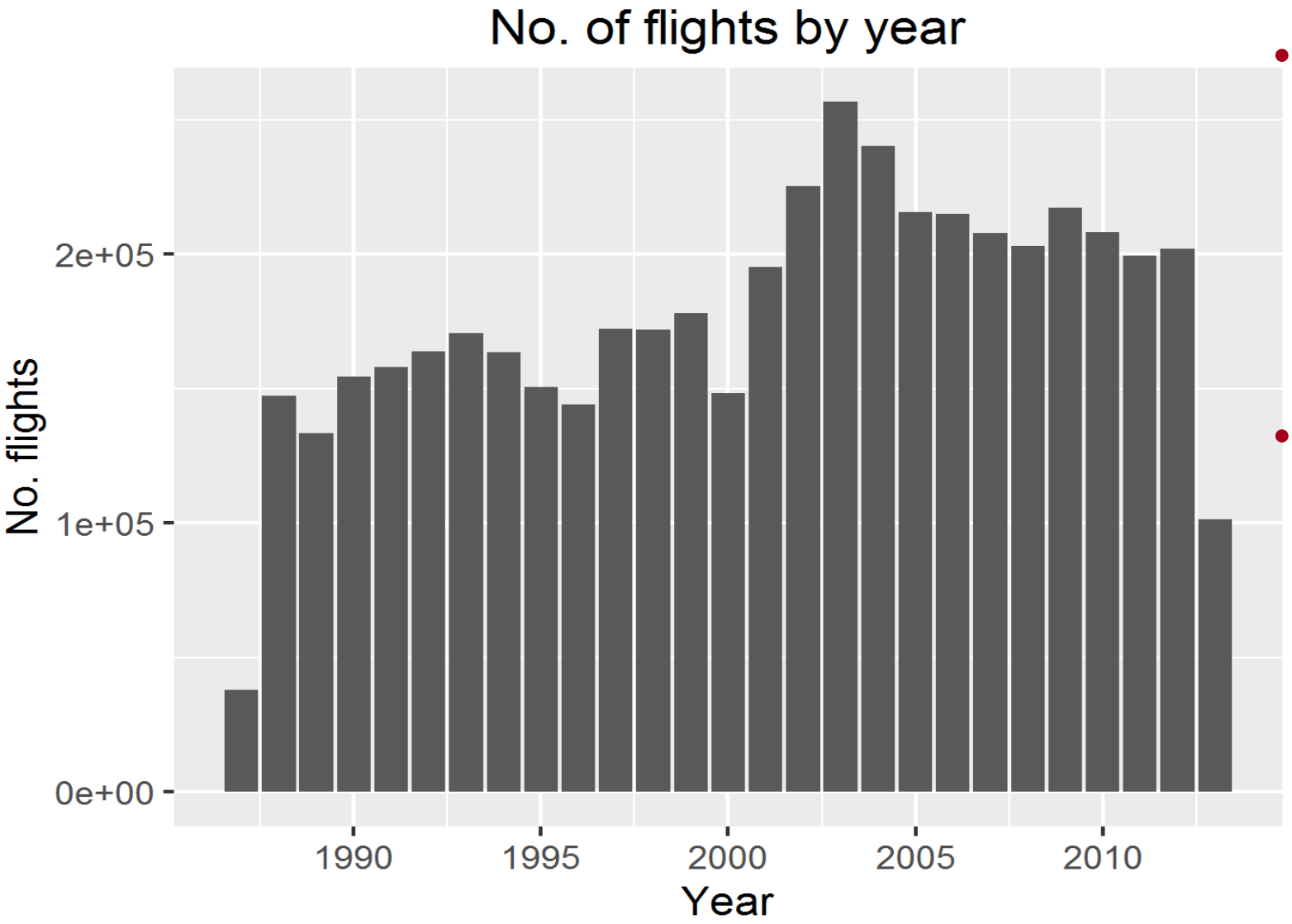
# The problems we want to solve

- Predict the delay of the flight
  - How to establish the right model among 31 features?
  - Which airline company is more reliable?
  - Could we improve the cause of delay? i.e. reduce the delay rate.
  - If delay is inevitable, could we estimate the delay time in advance such that either airline company or the customer can prepare for it.

# Dataset

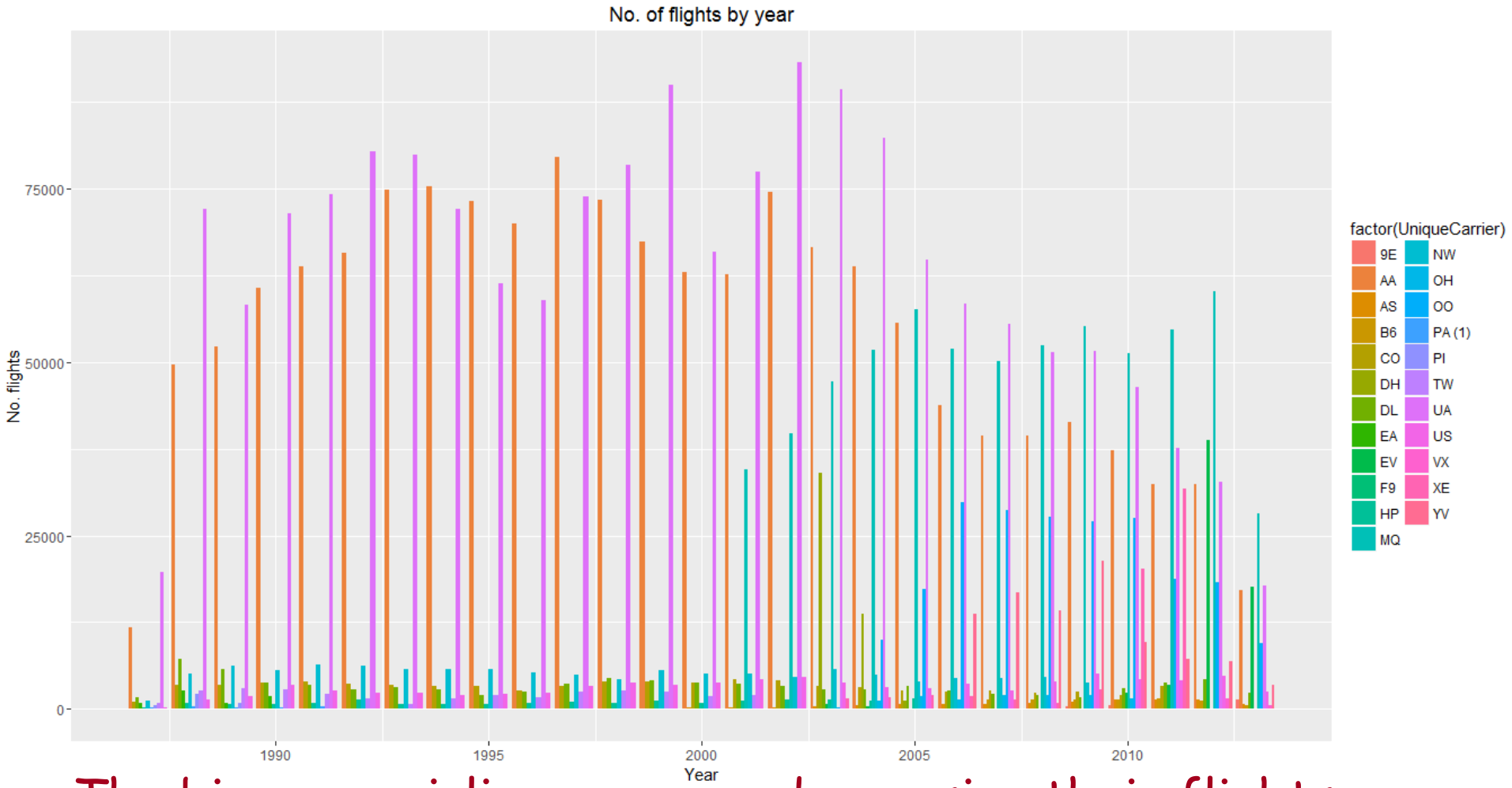
- The data is download from the website of United States Department of Transportation.
- The data contain 4,078,094 recorders of arriving flight to Chicago O'Hare airport.
- The time span is from 1987 to 2013.
- Including 23 carriers from 180 cities.
  - AA, UA, DL, .. ; SFO, JFK….
- Every recorder has 31 features.
  - UniqueCarrier, ArrTime, ArrDelay (+ delay, - arrive early), Distance, IsArrDelayed ……..

# Flight Counts



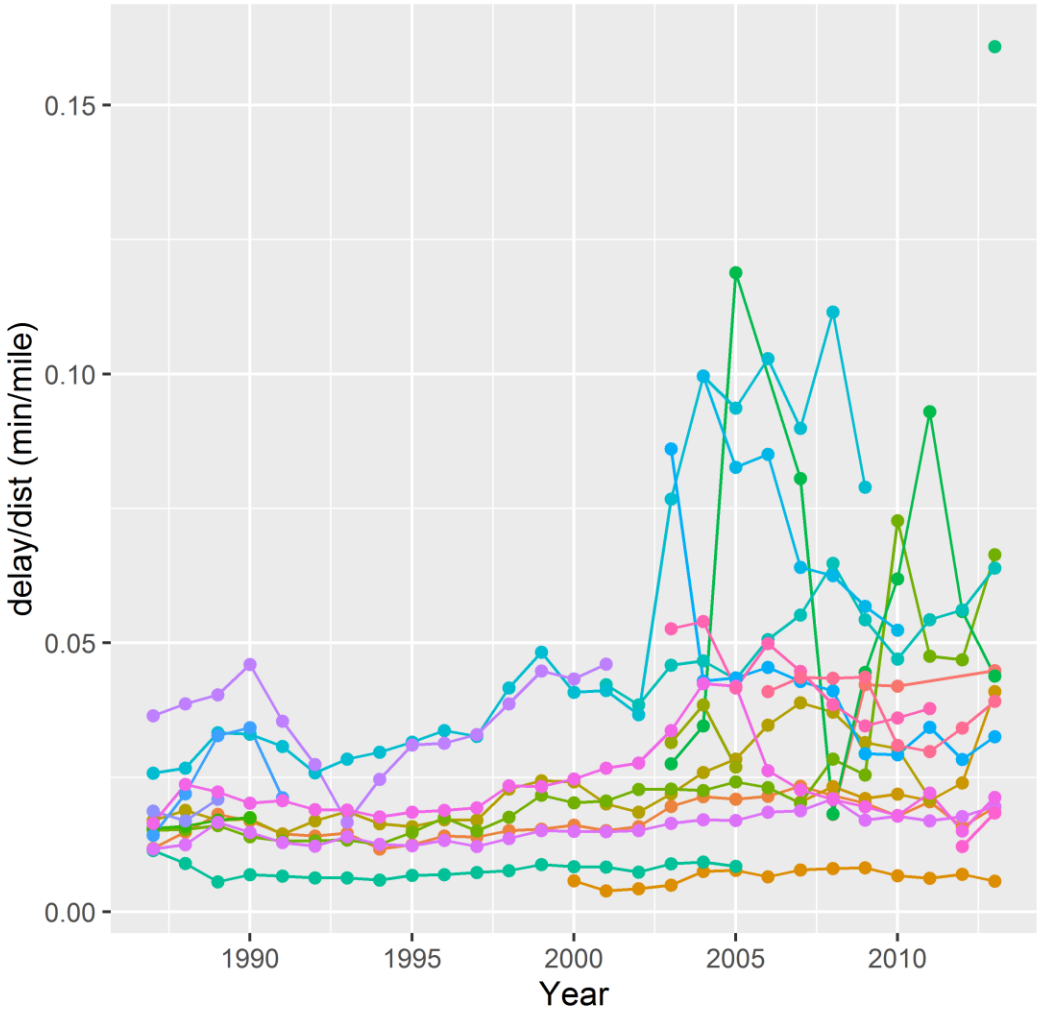

- There is a step increase after year 2000.
- Year 1987 and 2013 are not counted whole year.

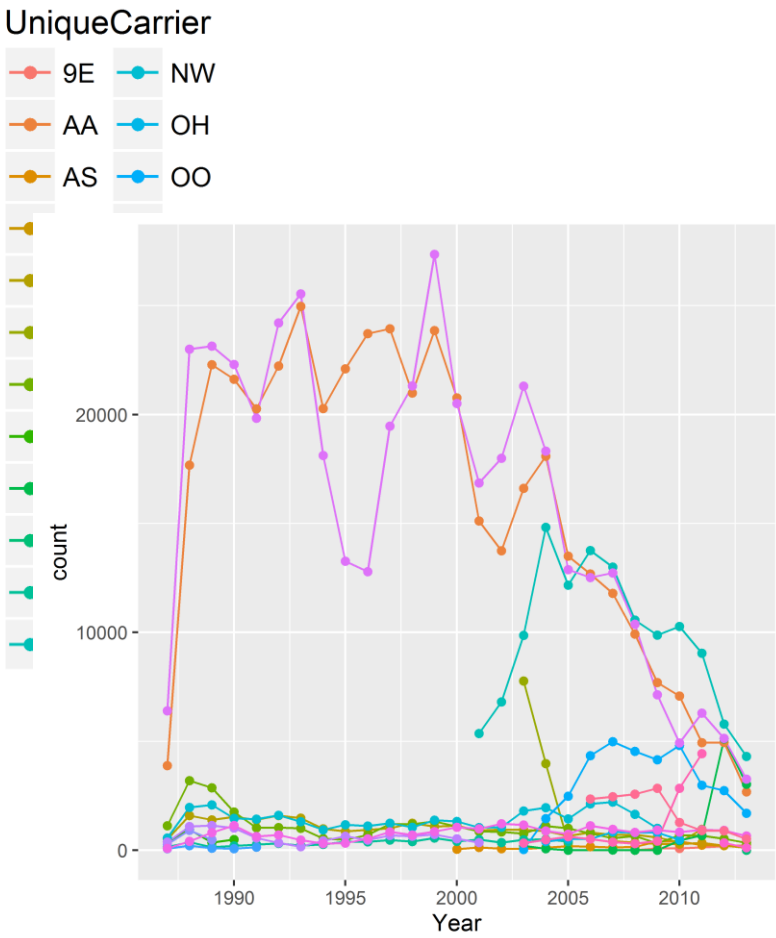
# Flight Counts




The big name airline company decreasing their flights to Chicago, local or small airline company takes the loads.

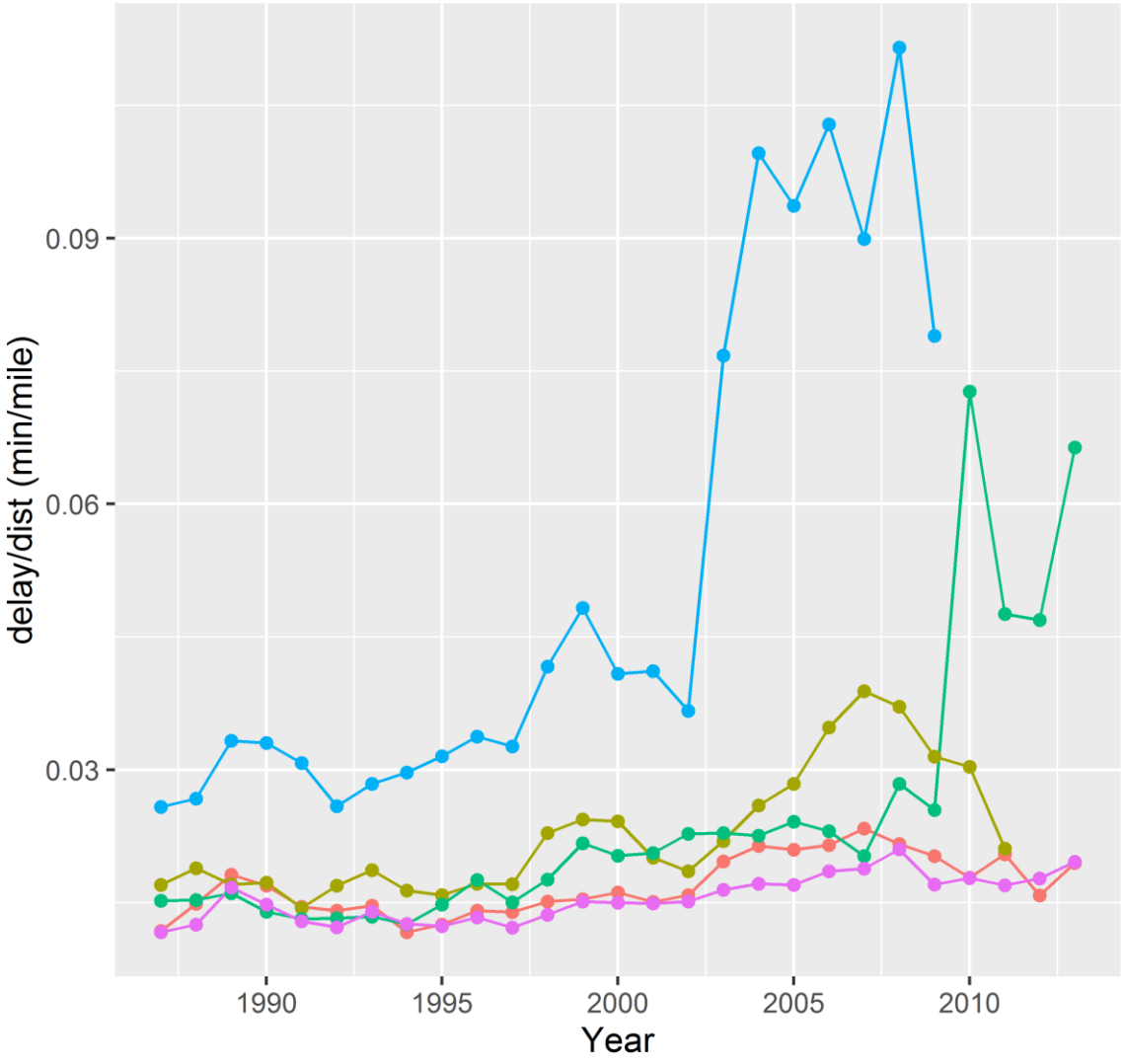
# Flight Delay (I)




Delay rate : delay time/traveling distance (min/mile)

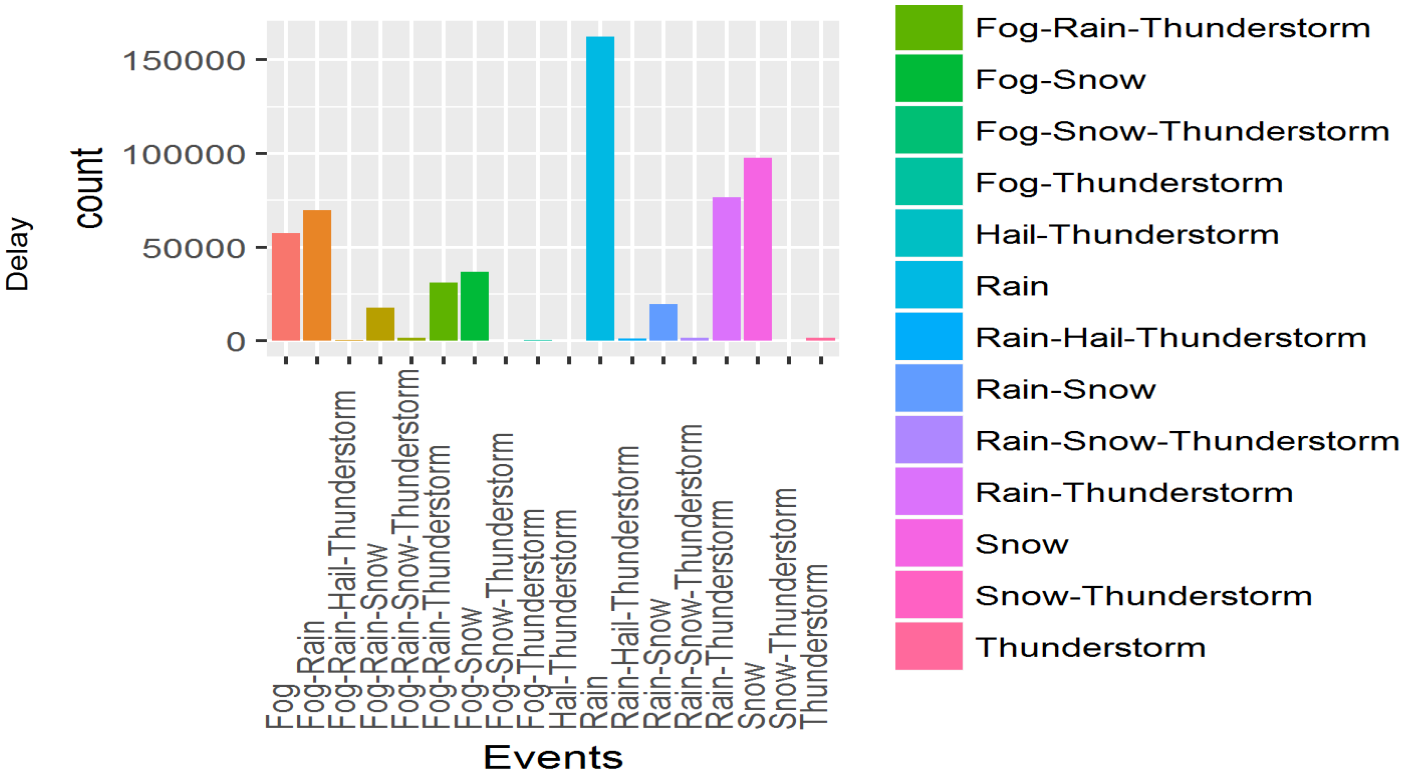
# Flight Delay (II)




- Delay rate becomes worse as time goes by.
  - NW has worst delay rate among five UA big airlines .
  - After merged with DL it affects DL delay rate.
- In term of delay rate UA and AA are comparable

# Dataset: Weather data

We used weather data at O'Hare airport as well to investigate an effect weather events on aircraft delay.




Weather data includes: Humidity, SealevelPressure, PrecipitationIn, CloudCover, Events….

# Assumption

Can we predict arrival delay might occur?
We presume we could predict delay using logistic regression

```
head(test_arr$IsArrDelayed)
NO NO NO NO YES NO
Levels: NO YES
```

**logistic regression:**

```
arr.glm = glm(IsArrDelayed~CRSElapsedTime
              +Distance
              +UniqueCarrier
              +ts
              +Events
              +PrecipitationIn
              +CloudCover
              +MeanTempF
              ,data=train, family = binomial)
```

Flight data

Weather data

# Logistic Regression model

| > summary(arr.glm) | | | | | |
|---|---|---|---|---|---|
| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
| (Intercept) | 5.22E-02 | 1.61E-01 | 0.323 | 0.746332 | |
| CRSElapsedTime | -1.02E-01 | 3.06E-03 | -33.334 | < 2e-16 | *** |
| AirTime | 1.39E-01 | 3.28E-03 | 42.4 | < 2e-16 | *** |
| Distance | -3.73E-03 | 2.50E-04 | -14.913 | < 2e-16 | *** |
| EventsFog-Rain | 1.24E-02 | 1.55E-01 | 0.08 | 0.936462 | |
| EventsFog-Rain-Hail-Thunderstorm | -4.97E-01 | 7.32E-01 | -0.679 | 0.497338 | |
| EventsFog-Rain-Snow | 6.11E-01 | 2.39E-01 | 2.559 | 0.010503 | * |
| EventsFog-Rain-Snow-Thunderstorm | 2.30E+00 | 8.45E-01 | 2.717 | 0.006588 | ** |
| EventsFog-Rain-Thunderstorm | 2.39E-01 | 1.91E-01 | 1.254 | 0.209802 | |
| EventsFog-Snow | 5.12E-01 | 1.70E-01 | 3.014 | 0.002578 | ** |
| ---------- | ---------- | ---------- | -------- | --------- | --- |
| EventsSnow-Thunderstorm | 2.00E+00 | 8.61E-01 | 2.32 | 0.020354 | * |
| EventsThunderstorm | -2.31E-01 | 3.53E-01 | -0.654 | 0.51313 | |
| PrecipitationIn | 2.15E-01 | 7.65E-02 | 2.805 | 0.005033 | ** |
| CloudCover | 5.91E-02 | 1.63E-02 | 3.627 | 0.000286 | *** |

# Logistic Regression model

Logistic regression will provide probabilities in the form of P(Y=1|X). Our decision boundary will be 0.5.
If P(y=1|X) > 0.5 then y = 1 otherwise y=0.

**fitted.results=P(Y=1|X)=**
**0.2241, 0.4194, 0.0819, 0.3303, 0.9753, 0.12731**
**0, 0 , 0, 0, 1, 0**

Prediction

**tail(test_arr$IsArrDelayed)**
**NO, NO, NO, NO, YES, NO**

Observation

Accuracy on the test set is: 0.8032278"

# Discussion and Conclusion

- Our logistic regression model using flight and weather data at O'Hare airport shows 80% accuracy on the test set.
- Here we only use 1 % of data for training due to memory limits.
- Things we are interested to do but lack of data
  - Flight delay versus the age of the aircraft. It could provide suggestion to the airline company to decide when it's more economic to replace the old aircraft.
  - Delay related to the direction of the flight
- Data analysis checking the common sense
- Power of data analysis exploring the model that beyond one's expectation and imagination