



# Airline On-Time Performance

Arzu Yesilova

Brad Peters

Irem Ataibis

Kanokporn Laochunsuwan

Sergio Iovanovich

# Agenda

- ✦ Problem/Issue
- ✦ Data for Analysis
- ✦ Analysis of Data
  - ✦ Top three models
  - ✦ Model comparison
  - ✦ What does it really mean?
- ✦ Recommendations for Future Studies
- ✦ Questions

# Problem/Issue

- ✦ Predict delay in domestic flights departing from Washington DC Metro Area before the actual flight day
- ✦ Why? - More information is available on flight performance once the flight takes off but not before-hand
- ✦ Customer Point-of-View - Airlines are among the companies that process significant amount of information for operational efficiency

# Data for Analysis

- ✦ Bureau of Transportation Statistics
  - ✦ Extensive resources in every form of transportation
  - ✦ Databases and Consumer Reports
  - ✦ Mission - To lead in developing transportation data and information of high quality, and to advance their effective use in both public and private transportation decision making.
  - ✦ [www.transtats.bts.gov](http://www.transtats.bts.gov)

# BTS Data

- ☀ Who uses the data from BTS?
  - ☀ Congress and Federal Agencies
  - ☀ State and Local Governments
  - ☀ Metropolitan Planning Organizations
  - ☀ Universities
  - ☀ Private Sector
  - ☀ General Public – Us!!

# BTS Data



## Bureau of Transportation Statistics

### Numbers to Move People

#### Search Entire Site

[Advanced Search](#)

#### Data

#### National Transportation Library

#### Products

#### Programs

#### Laws and Regulations

#### Press Room

#### About BTS

#### Events

#### Jobs

#### External Links

#### :: Quick Vote

What do you think of  
our redesigned

[Home](#) > [NTDA](#) > [OAI](#)

## Airline On-Time Statistics

[Summary Statistics](#) [Detailed Statistics](#) [Special Reports](#)

### Detailed Statistics

Detailed statistics are collected and reported by the Bureau of Transportation Statistics on flight delays, cancellations and diversions for the nation's major airlines and the airports they serve since 1995:

- Departure Statistics** with scheduled departure time, actual departure time, scheduled flight time, actual flight time, departure delay, wheels-off time and taxi-out time
- Arrival Statistics** with scheduled arrival time, actual arrival time, scheduled flight time, actual flight time, arrival delay, wheels-on time, taxi-in time
- Airborne Time** by airport and airline for individual flights
- Cancellation** by airport and airline for individual flights
- Diversion** by airport and airline for individual flights

# BTS Data

## ☀ Departure Statistics Variables

- ☀ Scheduled Departure Time
- ☀ Actual Departure Time
- ☀ Scheduled Elapsed Time
- ☀ Actual Elapsed Time
- ☀ Departure Delay
- ☀ Wheels-off Time
- ☀ Taxi-out Time

# BTS Data - Query

## ✦ Airport

- ✦ DCA – Reagan National
- ✦ BWI – Baltimore International
- ✦ IAD – Dulles International

## ✦ Airline

- ✦ Continental, Delta, Northwest, Southwest, US Airways

## ✦ Date (one week)

- ✦ May 10 – May 16 2003

# BTS Data

Airport: Baltimore, MD - Baltimore/Washington International (BWI)

Carrier Code	Date (MM/DD/YYYY)	Flight Number	Tail Number	Destination Airport	Scheduled Departure Time	Actual Departure Time	Scheduled Elapsed Time (Minutes)	Actual Elapsed Time (Minutes)	Wheels-off Time	Taxi-out Time (Minutes)
DL	05/13/2003	0343	N621DL	CVG	6:15	6:08	0095	0088	6:18	0010
DL	05/13/2003	0618	N662DN	ATL	10:10	10:05	0111	0104	10:19	0014
DL	05/13/2003	1003	N636DL	ATL	11:45	11:51	0108	0104	12:02	0011
DL	05/13/2003	1019	N699DL	CVG	9:40	9:32	0093	0088	9:45	0013
DL	05/13/2003	1135	N110DL	ATL	13:05	13:03	0115	0115	13:21	0018
DL	05/13/2003	1169	N960DL	ATL	17:35	17:28	0117	0112	17:46	0018
DL	05/13/2003	1211	N659DL	ATL	7:00	6:53	0119	0120	7:17	0024
DL	05/13/2003	1445	N651DL	ATL	16:15	16:18	0114	0106	16:31	0013
DL	05/13/2003	1448	N626DL	ATL	19:05	18:59	0117	0102	19:10	0011
DL	05/13/2003	1490	N995DL	ATL	5:45	5:35	0107	0104	5:50	0015
DL	05/13/2003	1563	N657DL	CVG	17:55	17:49	0103	0092	18:05	0016
DL	05/13/2003	1816	N661DM	ATL	14:35	14:30	0121	0120	14:49	0019

- ☀ Database Tables
  - ☀ Carrier Codes
  - ☀ Destination Airport

# Additional Data

## ☀ Modified Variables

### ☀ Destination Airport → Regions

- 5 geographic regions (Northeast, North-Midwest, Southeast, South-Midwest, West)

### ☀ Date → Day of the Week

### ☀ Scheduled Departure Time → Minutes

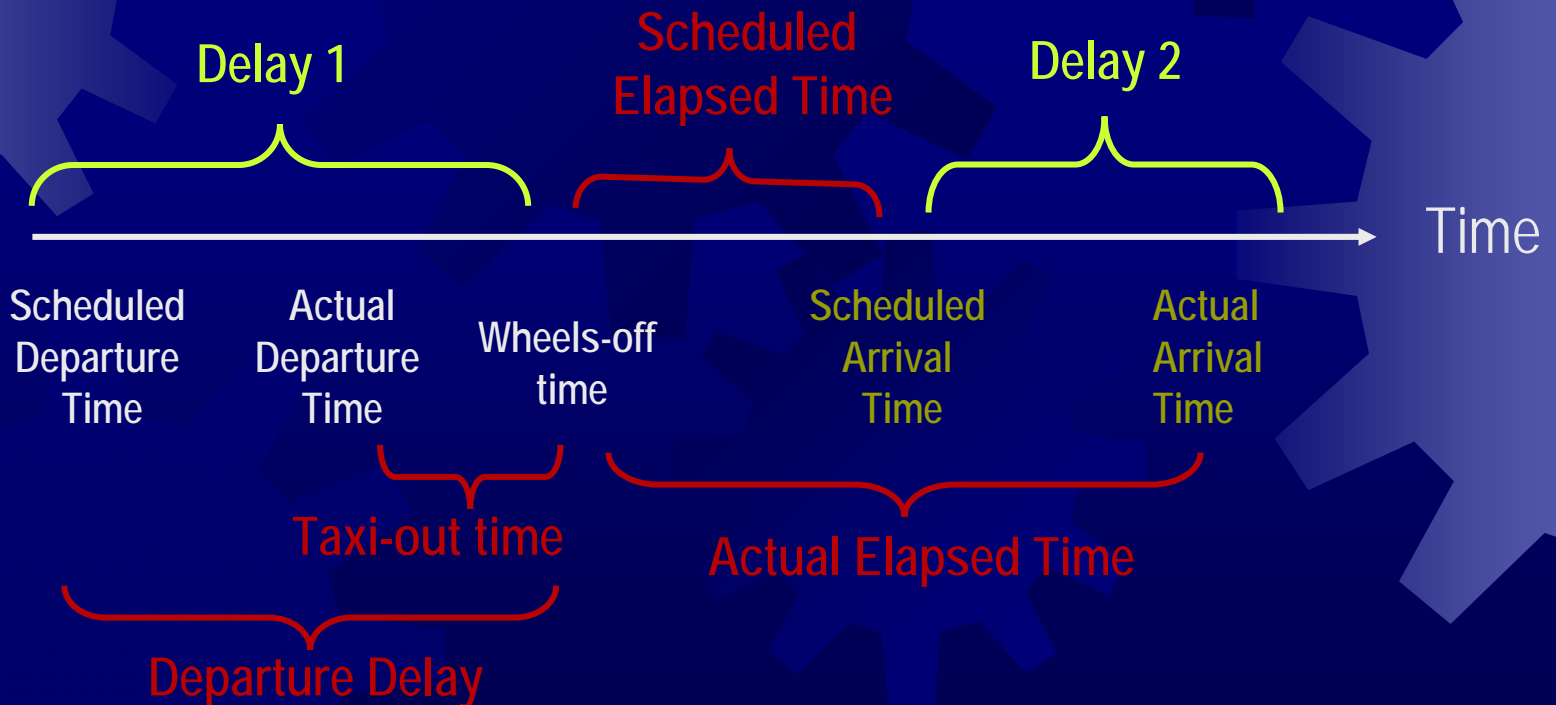
- Transformed into minutes through a formula
- $\{(hh,mm/60)/24\} * 1440$
- Ex: 09:45 am →  $(9,75/24) * 1440 = 585$  mn.

$\underbrace{\hspace{2em}}$   $\underbrace{\hspace{2em}}$   
n-th percentile    # of minutes in a day

# Additional Data

## ☀ Variables Added

- ☀ Weather Condition ([www.weather.com](http://www.weather.com))
- ☀ Flight Status (Delayed/Not Delayed)



# Additional Data

- ✦ Flight Status – Delayed/Not Delayed
  - ✦ Delay = Departure Delay + Actual Elapsed Time – Scheduled Elapsed Time
  - ✦ If Delay < 10min → Not-Delayed
  - ✦ If Delay > 10 min → Delayed

# Issues with Data Collection

- ✦ Availability of the data
  - ✦ Query on database vs. all data
  - ✦ Departure Statistics vs. Arrival Statistics
- ✦ Format of the data
  - ✦ Text with spaces
  - ✦ Time format (text vs. time)
- ✦ Technical definitions of the variables
  - ✦ Scheduled Elapsed Time etc.

# Issues with Data Collection

## ☀ Weather variable

- ☀ Past weather statistics – average temperature vs. weather condition (sunny/cloudy etc.)

## ☀ Irrelevant observations

- ☀ International Flights (Mexico)
- ☀ Flights with Actual Elapsed Time = 0
  - Cancelled Flights

# Putting It All Together - Data

- ★ Size = 2,388 (n) \* 7 (k)
- ★ Observation = Flight
- ★ Flight Status (Delayed/Not Delayed) as a function of...
  - ★ Carrier
  - ★ Departure Airport
  - ★ Day of the Week
  - ★ Time of the Day
  - ★ Weather
  - ★ Destination Airport

# Sample Data

Row ID	Departure Airport	Carrier Code	Day of the Week	Weather	Destination Airport_Region	Time_in_Minutes	Flight Status
5	DCA	Delta	Wednesday	Clear	Northeast	1110	Not Delayed
7	BWI	Southwest	Friday	Cloudy/Rain	West	650	Not Delayed
8	BWI	Southwest	Friday	Cloudy/Rain	West	1175	Not Delayed
9	DCA	Delta	Thursday	Clear	Northeast	1050	Not Delayed
10	BWI	Southwest	Friday	Cloudy/Rain	West	1060	Not Delayed
11	BWI	Southwest	Wednesday	Clear	West	420	Not Delayed
12	BWI	Southwest	Thursday	Clear	West	580	Not Delayed
13	BWI	Southwest	Wednesday	Clear	West	1175	Not Delayed
14	BWI	Southwest	Thursday	Clear	West	1060	Not Delayed
15	BWI	Southwest	Thursday	Clear	West	1175	Not Delayed
16	BWI	Southwest	Friday	Cloudy/Rain	West	1190	Not Delayed
17	BWI	Southwest	Friday	Cloudy/Rain	West	485	Not Delayed
18	BWI	Southwest	Friday	Cloudy/Rain	West	1065	Not Delayed

# Summary Statistics

Departure Airport	Count	Percentage
BWI	1352	56.62%
DCA	836	35.01%
IAD	200	8.38%
Grand Total	2388	100.00%

Carrier Code	Count	Percentage
Continental	152	6.37%
Delta	384	16.08%
Northwest	265	11.10%
Southwest	1031	43.17%
US Airways	556	23.28%
Grand Total	2388	100.00%

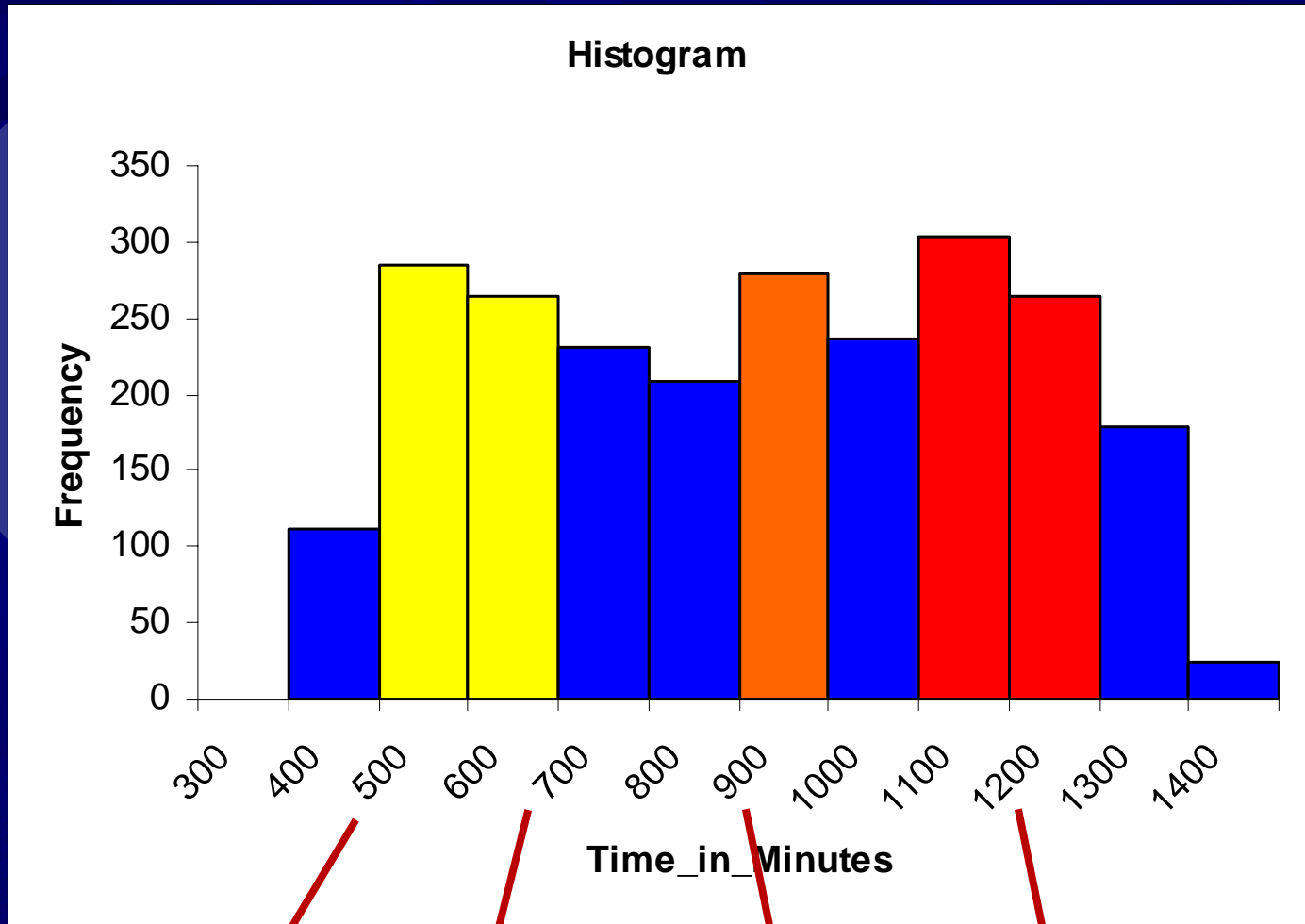
Day of the Week	Count	Percentage
Sunday	327	13.69%
Monday	359	15.03%
Tuesday	355	14.87%
Wednesday	344	14.41%
Thursday	361	15.12%
Friday	362	15.16%
Saturday	280	11.73%
Grand Total	2388	100.00%

Weather	Count	Percentage
Clear	1419	59.42%
Cloudy/Rain	969	40.58%
Grand Total	2388	100.00%

Destination Airport	Count	Percentage
Northeast	789	33.04%
North-Midwest	444	18.59%
Southeast	841	35.22%
South-Midwest	195	8.17%
West	119	4.98%
Grand Total	2388	100.00%

Flight Status	Count	Percentage
Delayed	419	17.55%
Not Delayed	1969	82.45%
Grand Total	2388	100.00%

# Summary Statistics



~8:20 am

~11:40 am

~3:00 pm

~8:00 pm

# Delayed vs. Not-Delayed

Carrier Code	Delayed	Not Delayed	Grand Total
Continental	19.74%	80.26%	100.00%
Delta	23.44%	76.56%	100.00%
Northwest	20.75%	79.25%	100.00%
Southwest	14.65%	85.35%	100.00%
US Airways	16.73%	83.27%	100.00%
Grand Total	17.55%	82.45%	100.00%

Departure Airport	Delayed	Not Delayed	Grand Total
BWI	15.16%	84.84%	100.00%
DCA	19.02%	80.98%	100.00%
IAD	27.50%	72.50%	100.00%
Grand Total	17.55%	82.45%	100.00%

Weather	Delayed	Not Delayed	Grand Total
Clear	11.98%	88.02%	100.00%
Cloudy/Rain	25.70%	74.30%	100.00%
Grand Total	17.55%	82.45%	100.00%

# Delayed vs. Not-Delayed

Destination Airport	Delayed	Not Delayed	Grand Total
Northeast	15.59%	84.41%	100.00%
North-Midwest	18.69%	81.31%	100.00%
Southeast	19.74%	80.26%	100.00%
South-Midwest	16.41%	83.59%	100.00%
West	12.61%	87.39%	100.00%
Grand Total	17.55%	82.45%	100.00%

Day of the Week	Delayed	Not Delayed	Grand Total
Sunday	24.16%	75.84%	100.00%
Monday	10.58%	89.42%	100.00%
Tuesday	4.79%	95.21%	100.00%
Wednesday	9.88%	90.12%	100.00%
Thursday	22.44%	77.56%	100.00%
Friday	23.48%	76.52%	100.00%
Saturday	30.36%	69.64%	100.00%
Grand Total	17.55%	82.45%	100.00%

# Summary Statistics

## ☀ Correlations (Chi-Square)

P-values	Departure Airport	Carrier Code	Day_Grouped	Weather	Destination Airport_Region	Flight Status
Departure Airport	0					
Carrier Code	0	0				
Day_Grouped	0.92	0.964	0			
Weather	0.908	0.969	0	0		
Destination Airport_Region	0	0	0.942	0.803	0	
Flight Status	0	0.001	0	0	0.113	0

\* Low p-value says that variables are not independent

\* Carrier Code, Departure Airport and Destination Airport are correlated due to the routing decisions airlines make

# How did summary stats help?

- ✦ Overall understanding of the data
- ✦ How to handle dummy variables
  - ✦ Change in reference categories
    - DCA → IAD
  - ✦ Day of the week
    - Weekend vs. Weekday
    - Mon-Wed vs. Thu-Sun



# Analysis of Data

# Analysis Technique

## ★ Logistics Regression

- ★ The qualitative nature of the response variable
- ★ The qualitative nature of the predictor variables (other than departure time)
- ★ The binary distribution of predictor variables

## ★ Discriminant Analysis

- ★ To validate/double-check the findings from the logistic regression

# Modeling Process

- ✦ Getting the data ready for analysis
  - ✦ Creating Dummy Variables
  - ✦ Partitioning Data (60%-40%)
    - Main objective is to predict
  - ✦ Various combination of variables put into the model
    - Common Sense
    - Chi-Squared
    - Summary Statistics

# Modeling Process

## ☀ Cut-off value calculation

- ☀ Success Class = Delayed
- ☀ Cut-off value = the probability of delay from 3 departure airports in May 2003

Departure Airport	Count	%	% of Delayed Flights*
BWI	1352	56.62%	13.60%
DCA	836	35.01%	10%
IAD	200	8.38%	17.40%
Grand Total	2388	100.00%	12.66%

\* Bureau of Transportation Statistics -  
Aviation Consumer Report July 2003

# Modeling Process

## ★ Cost for various errors

Classification Confusion Matrix		
	Predicted Class	
Actual Class	Delayed	Not Delayed
Delayed	a	b
Not Delayed	c	d

- ★ The cost of 'b' is not equal to the cost of 'c'
- ★ The relative costs change from customer-to-customer (business vs. leisure)
- ★ No fact on the relative importance of costs of 'b' and 'c'

# Models

## ☀ Model A

Cut off Prob.Val. for Success (Updatable)

**0.126**

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-2.41725111	0.14583632	0	*
Delta 2	0.48739767	0.17616045	0.00566125	1.62807393
Th-Su	1.06250262	0.21579513	0.00000085	2.89360332
Cloudy/Rain	0.28823787	0.18842809	0.12609175	1.33407462

### Training Data scoring - Summary Report

Error Report			
Class	# Cases	# Errors	% Error
Delayed	264	50	18.94
Not Delayed	1169	696	59.54
<b>Overall</b>	<b>1433</b>	<b>746</b>	<b>52.06</b>

Null deviance	1897.7553
Model deviance	1820.5935
Improvement	77.1618
p-value	0.0000

### Validation Data scoring - Summary Report

Error Report			
Class	# Cases	# Errors	% Error
Delayed	155	33	21.29
Not Delayed	800	402	50.25
<b>Overall</b>	<b>955</b>	<b>435</b>	<b>45.55</b>

# Models

## ☀ Model B

Cut off Prob.Val. for Success (Updatable)

**0.126**

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-1.98874629	0.25156367	0	*
Delta 2	0.35831341	0.18807726	0.05676194	1.43091404
Th-Su	1.2787348	0.16228618	0	3.59209204
BVM_2	-0.50969231	0.23630914	0.0310144	0.60068035
DCA_2	-0.37287655	0.23657468	0.11499183	0.68875027

### Training Data scoring - Summary Report

Error Report			
Class	# Cases	# Errors	% Error
Delayed	264	54	20.45
Not Delayed	1169	622	53.21
<b>Overall</b>	<b>1433</b>	<b>676</b>	<b>47.17</b>

### Validation Data scoring - Summary Report

Error Report			
Class	# Cases	# Errors	% Error
Delayed	155	33	21.29
Not Delayed	800	402	50.25
<b>Overall</b>	<b>955</b>	<b>435</b>	<b>45.55</b>

Null deviance	1900.6523
Model deviance	1823.0526
Improvement	77.5997
p-value	0.0000

# Models

## ☀ Model C

Cut off Prob.Val. for Success (Updatable)

**0.126**

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-1.7849009	0.23806627	0	*
Delta 2	0.37025481	0.1883163	0.0492833	1.44810355
Time*Day	0.00122882	0.00015436	0	1.00122952
BVM_2	-0.53791696	0.23616017	0.02274068	0.58396339
DCA_2	-0.41011131	0.23597959	0.08222665	0.66357636

### Training Data scoring - Summary Report

Error Report			
Class	# Cases	# Errors	% Error
Delayed	264	46	17.42
Not Delayed	1169	699	59.79
<b>Overall</b>	<b>1433</b>	<b>745</b>	<b>51.99</b>

Null deviance	1899.7046
Model deviance	1823.7226
Improvement	75.9820
p-value	0.0000

### Validation Data scoring - Summary Report

Error Report			
Class	# Cases	# Errors	% Error
Delayed	155	27	17.42
Not Delayed	800	466	58.25
<b>Overall</b>	<b>955</b>	<b>493</b>	<b>51.62</b>

# Model Comparison

- Overall Error as well as Error in Delayed Flights
  - More weight on error in delayed flights

		Model A		Model B		Model C	
	Error	Training	Validation	Training	Validation	Training	Validation
<b>Linear Regression</b>	Delayed	18.94%	21.29%	20.45%	20.65%	17.42%	17.42%
	Overall	52.06%	45.55%	47.17%	47.02%	51.99%	51.62%

		Model A		Model B		Model C	
	Error	Training	Validation	Training	Validation	Training	Validation
<b>Discriminant Analysis</b>	Delayed	18.94%	20.00%	17.42%	17.42%	17.42%	17.42%
	Overall	52.06%	52.36%	51.99%	51.62%	51.99%	51.62%

# Interpretations

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-1.7849009	0.23806627	0	*
Delta 2	0.37025481	0.1883163	0.0492833	1.44810355
Time*Day	0.00122882	0.00015436	0	1.00122952
BW1_2	-0.53791696	0.23618017	0.02274068	0.58396339
DCA_2	-0.41011131	0.23597959	0.08222665	0.66357636

## ☀ Delta\_2

- ☀ The odds of Delta flight getting delayed is 1.45 times the odds of a flight from one the other 4 airlines getting delayed, when all other predictor variables are held constant

# Interpretations

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-1.7849009	0.23806627	0	*
Delta 2	0.37025481	0.1883163	0.0492833	1.44810355
Time*Day	0.00122882	0.00015436	0	1.00122952
BWI_2	-0.53791696	0.23618017	0.02274068	0.58396339
DCA_2	-0.41011131	0.23597959	0.08222665	0.66357636

## ☀ Time\*Day

- ☀ For flight on Monday thru Wednesday, the time of the flight doesn't necessarily contribute to the odds of delay, when all else is constant
- ☀ For flights on Thursday thru Sunday, any additional minute in the time of the flight, increases the odds of delay by a factor of 1.001, when all other variables held constant
- ☀ An additional hour → a factor of 1.08 increase

# Interpretations

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-1.7849009	0.23806627	0	*
Delta 2	0.37025481	0.1883163	0.0492833	1.44810355
Time*Day	0.00122882	0.00015436	0	1.00122952
BWI_2	-0.53791696	0.23618017	0.02274068	0.58396339
DCA_2	-0.41011131	0.23597959	0.08222665	0.66357636

## ★ BWI\_2

- ★ The odds of a flight out of BWI getting delayed is 0.58 times the odds of a flight out of IAD getting delayed when all other predictors are constant.

## ★ DCA\_2

- ★ The odds of a flight out of DCA getting delayed is 0.66 times the odds of a flight out of IAD getting delayed, when all other predictors are constant.

# What does it really mean for you?

- ✦ Don't fly Delta out of Washington DC Metro Area! – if you don't have to!
- ✦ Your airport preference should be BWI, DCA and IAD to minimize the odds of delay
- ✦ Try to fly on Monday, Tuesday or Wednesday
- ✦ If you need to fly on the other days, then fly as early in the day as you can.
- ✦ Your destination doesn't necessarily have too much to do with the delay
- ✦ The weather conditions will be addressed under recommendations

# Recommendations

- ✱ More data covering a larger time frame
  - ✱ A month
  - ✱ A year (to account for seasonality)
- ✱ Weather – Day of the week complication
  - ✱ Mon-Fri → Cloudy
  - ✱ Sat-Sun → Clear

Questions?

