



# Explaining Navy Reserve Training Expense Obligations

*Emily Franklin*

*Roxana Garcia*

*Mike Hulsey*

*Raj Kanniyappan*

*Daniel Lee*



# Agenda

- Defining The Problem
- Data Analysis
- Data Cleaning
- Exploration
- Models & Methods
- Model Performance
- Recommendations

# Defining The Problem

Explanation or Prediction?

- Explain the outstanding travel obligations within the US Navy Reserve.

What is the analysis going to be used for?

- Determine whether travel policy changes are needed.

Who will be the users?

- Navy Reserve Headquarters staff

What is the currently implemented?

- Access tool implemented by contractors
- Travel Responsibility Manual

# Data Analysis

## Data Source

- Navy Reserve Order Writing System (NROWS) database

## Data Quality

- Directly entered by reservist in NROWS and approved by appropriate official.
- Pay disbursements fed from Navy Reserve financial system

## Size of the Data

- Training and travel records for fiscal year 2009
- 86,000 records in total (liquidated and unliquidated costs)
- 10,000 sample dataset used for modeling and visualizations

## Security and Privacy

- Social security numbers and other personal information were removed prior to obtaining dataset

# Data Cleaning

## Dataset Generation

- Expense report generated from three separate reports from NROWS
- Report generated on August 28, 2009
- 86,000 total records
- Created random sample of 10,000 records as final data set

## Incomplete Records

- Removed records with missing data elements

## Dummy Variables: created dummy variables for the following categorical variables

- Two Order Type Ref Variables: ADT as reference value
- Two ACRN Ref Variables: AA as reference value
- Five Region Ref Variables: RCC MA as reference value
- One Travel System Ref Variables: DTS as reference value

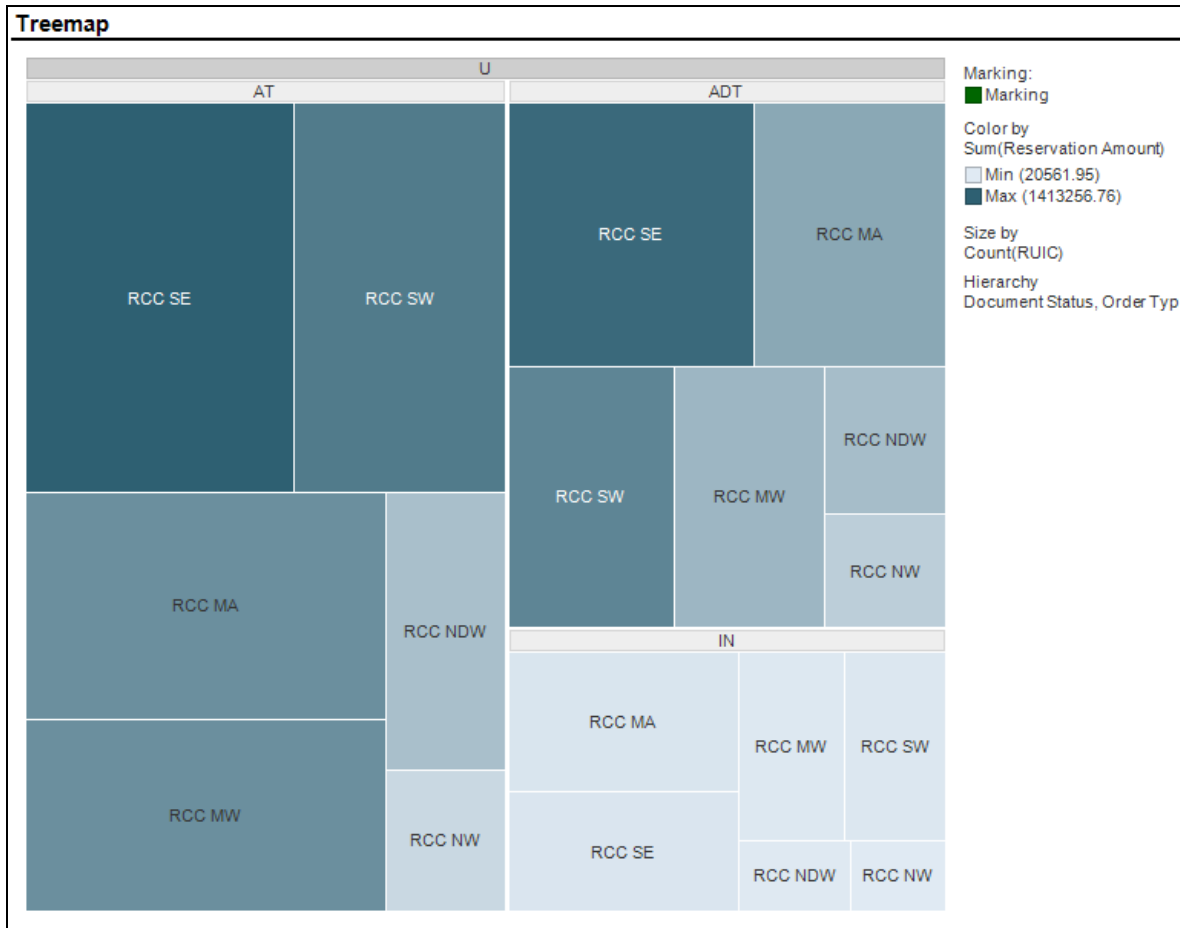
## Data Record Adjustment

- Created new variables (i.e.  $\text{Log}[\text{Reservation Amount}]$ )
- Removed insignificant variables

# Exploration

## Treemap Chart: Number of Unliquidated Records

- *Unliquidated Records Only*
- *Hierarchy*: Document Status, Order Type
- *Interpretation*: Of the unliquidated data records, the majority of the outstanding expense records on Annual Training and then Active Duty Training



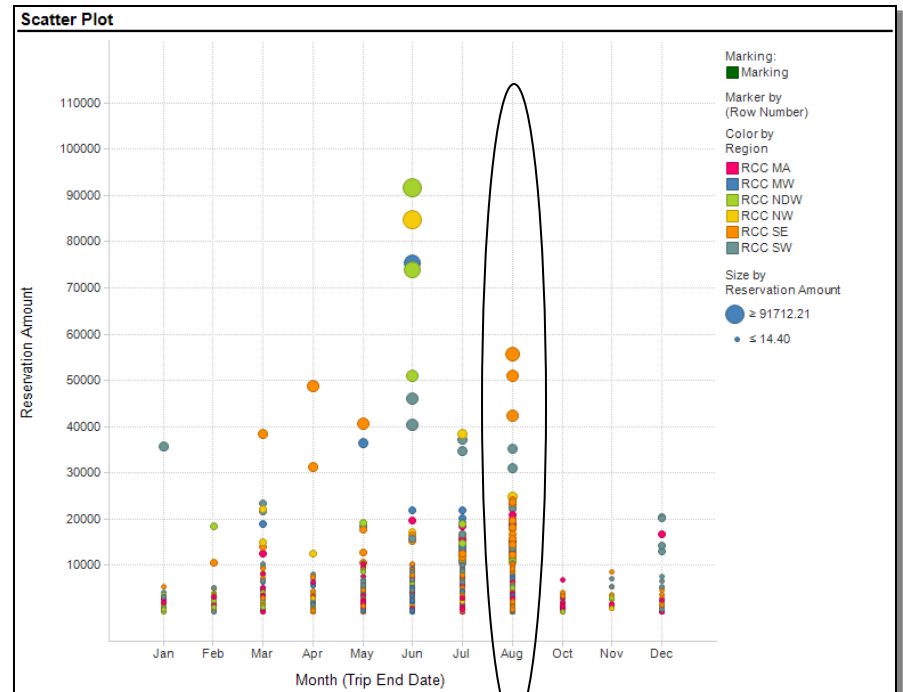
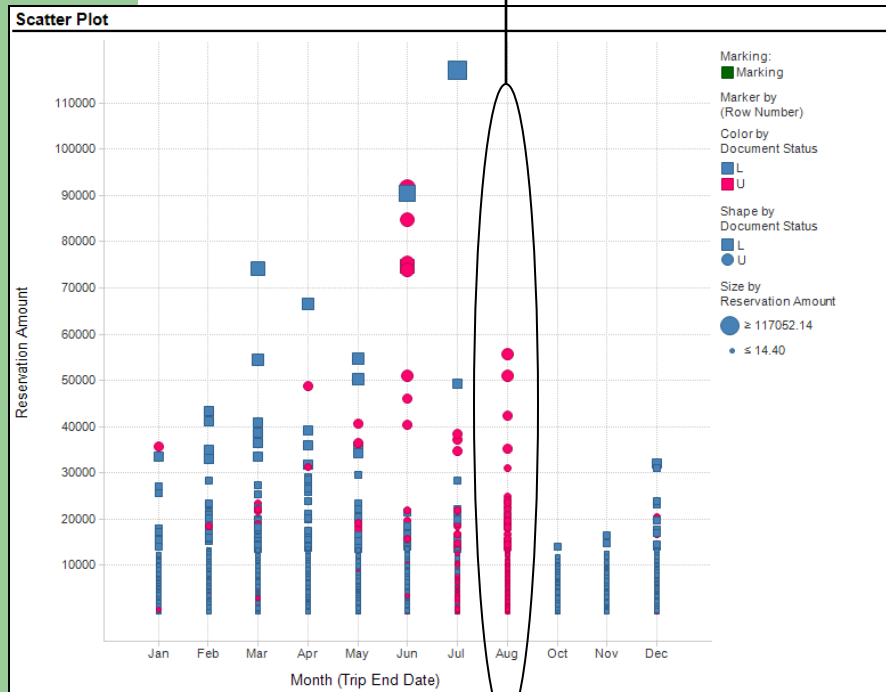
## Order Type

- Active Duty Training
- Annual Training
- Inactive Duty Training

# Exploration

## Scatter Plot: When & Where Unliquidated Records Occur

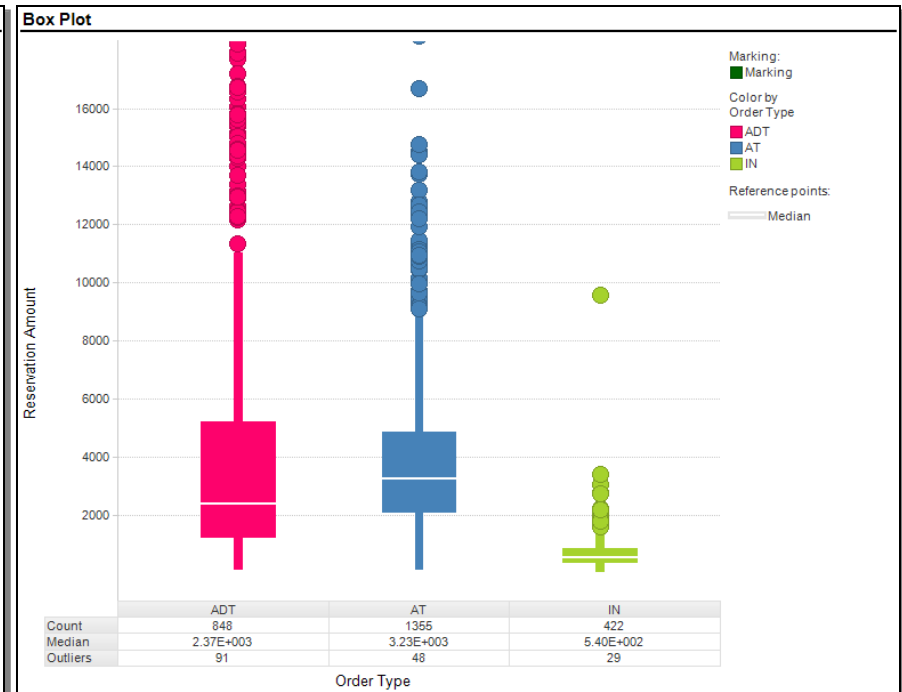
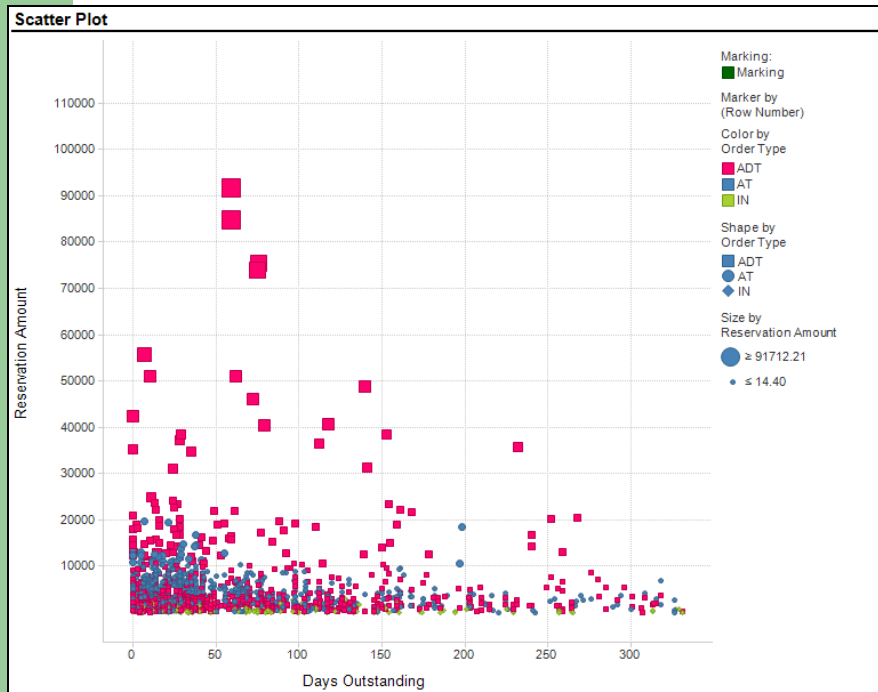
- *Liquidated & Unliquidated Records*
- *Hierarchy: Document Status and Then Region*
- *Interpretation: After determining when the highest amount of unliquidated data records occur, we determined that the majority of the records occur in Region RCC SW*



# Exploration

## Scatter Plot, Box Plot: Amount of Unliquidated Expenses

- *Unliquidated Records Only*
- *Hierarchy: Order Type, Size By Reservation Amount*
- *Interpretation: Of the unliquidated records, the highest level of reservation amounts are tied to Active Duty Training*





# Models & Methods

- With the goal of explaining, our team ran the following Models: Logistic Regression, Discriminant Analysis, Classification Tree
- Our team began with more than 86,000 records. Using XLMiner, we took a random sample of 10,000 records so that our dataset was more manageable using the Explanatory Models in XLMiner.
- The "Y" output variable we used is 'Document Status' - Resulting in either Liquidated (L) or Unliquidated (U) data records. The input variables consisted of numerical and non-numerical data, and the non-numerical data, such as ACRN, Region and Order Type were converted to dummy variables.

# Model Performance

Model	Significant Input Variables	Overall Error	Error in Classifying Unliquidated	Multiple R-Squared
Naïve Rule	Majority Rule Predicts Liquidated.	26.25%	100%	
Logistic Regression #1	Days Outstanding, Number of Days, Order Type, Travel System, Reservation Amount, Advance Amount, Region	2.59%	9.83%	0.08751
Logistic Regression #2	Days Outstanding, Number of Days, Order Type, Reservation Amount, Advance Amount, Region	2.59%	9.83%	0.87511
Logistic Regression #3	Days Outstanding, Order Type, Reservation Amount, Advance Amount, Region	2.59%	9.83%	0.87506
Logistic Regression #4	Days Outstanding, Order Type, Reservation Amount, Advance Amount, ACRN	2.52%	9.56%	0.87484
Logistic Regression #5	Days Outstanding, Order Type, Reservation Amount, Advance Amount	2.46%	9.44%	0.87409
<b>Logistic Regression #6</b>	<b>Days Outstanding, Order Type, Log(Reservation Amount)</b>	<b>2.46%</b>	<b>9.49%</b>	<b>0.87344</b>
Discriminant Analysis #1	Days Outstanding, Number of Days, Order Type, Travel System, Reservation Amount, Advance Amount, Region	11.56%	43.85%	
Discriminant Analysis #2	Days Outstanding, Number of Days, Order Type, Reservation Amount, Advance Amount, Region	11.58	43.89%	
<b>Discriminant Analysis #3</b>	<b>Days Outstanding, Order Type, Reservation Amount, Advance Amount</b>	<b>11.53%</b>	<b>43.70%</b>	
Classification Tree	Number of Days, Reservation Amount, Order Type, Advance Amount	25.89%	100%	

# Model Performance

## Logistical Regression Model – Best Model

**Input Variables:** Outstanding, Order Type\_AT, Order Type\_IN

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	3.48984122	0.13432698	0	*
Days Outstanding	-0.64695567	0.02618866	0	0.52363747
Order Type_AT	-0.40805581	0.15914348	0.01034511	0.66494173
Order Type_IN	0.78317082	0.25056556	0.00177435	2.18840027

Residual df	6996
Residual Dev.	1375.511719
% Success in training data	74.11428571
# Iterations used	8
Multiple R-squared	0.87344289

Training:

Error Report			
Class	# Cases	# Errors	% Error
L	5188	0	0.00
U	1812	172	9.49
<b>Overall</b>	<b>7000</b>	<b>172</b>	<b>2.46</b>

Validation:

Error Report			
Class	# Cases	# Errors	% Error
L	2187	0	0.00
U	813	80	9.84
<b>Overall</b>	<b>3000</b>	<b>80</b>	<b>2.67</b>

# Model Performance

## Discriminant Analysis Model – Best Model

**Input Variables:** Days Outstanding, Order Type\_AT, Order Type\_IN, Reservation Amount, Advance Amount

Variables	Classification Function	
	L	U
Constant	-2.29217935	-3.93947172
Days Outstanding	0.0042434	0.05690328
Order Type_AT	3.3345499	3.82253385
Order Type_IN	4.12011194	3.94162393
Reservation Amount	0.00022529	0.00024252
Advance Amount	-0.00029611	0.00059535

Error Report			
Class	# Cases	# Errors	% Error
L	7375	6	0.08
U	2625	1147	43.70
<b>Overall</b>	<b>10000</b>	<b>1153</b>	<b>11.53</b>

# Model Performance

## Classification And Regression Trees

**Input Variables:** Number of Days, Order Type\_AT, Order Type\_IN, Reservation Amount, Advance Amount

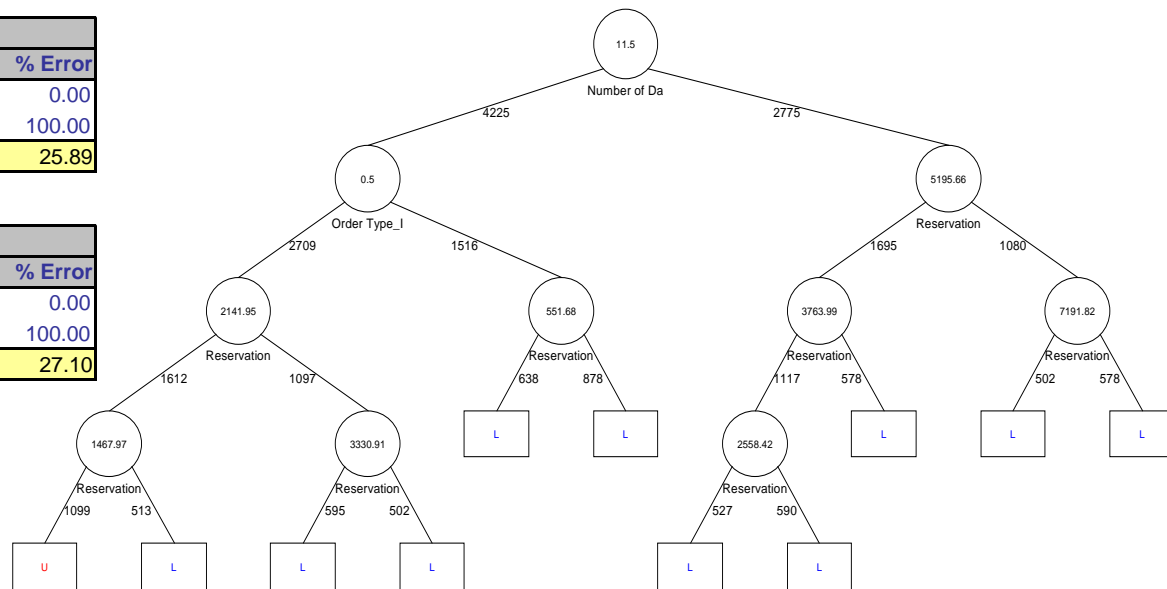
**Pruned Tree = Naïve Rule, predicting all as Liquidated.**

Training:

Error Report			
Class	# Cases	# Errors	% Error
L	5188	0	0.00
U	1812	1812	100.00
<b>Overall</b>	<b>7000</b>	<b>1812</b>	<b>25.89</b>

Validation:

Error Report			
Class	# Cases	# Errors	% Error
L	2187	0	0.00
U	813	813	100.00
<b>Overall</b>	<b>3000</b>	<b>813</b>	<b>27.10</b>



# Recommendations

**Use & Deployment:** Based upon our team's Data Mining Analysis Project, we encourage the Navy Reserve to focus its attention on the following to reduce unliquidated training instances

- Review our team's linear regression model #6 and focus its attention to re-evaluate the training efforts for both Annual Training and Training for Inactive Reservists as these are the most significant variables along with Days Outstanding
- Review the training strategy in Region RCC SE since this region has the largest number of outstanding unliquidated instances
- Review the schedule for when expense training is given to reservists since most of the unliquidated records occurred in August
- Training Emphasis Examples:
  - Trainers who can review the status of orders and work with reservists
  - Trainers who can my be contacted to assist reservists having issues submitting travel claims
  - Training on the Travel Claim System
  - Escalation channels to officers superior to reservists with outstanding travel claims

# Questions / Discussion

