



Module 5 – The Endgame, or Putting it All Together



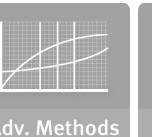
Introduction



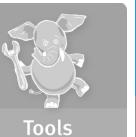
Analytics Lifecycle



Basic Methods



Adv. Methods



Tools



Lab

Module 5 – The Endgame, or Putting it All Together

Upon completion of this module, you should be able to:

- Articulate three tasks needed to operationalize an analytics project
- Explain how the four common deliverables of an analytics lifecycle project meet the needs of key stakeholders
- Use a framework for creating final presentations for sponsors and analysts
- Evaluate a data visualization and identify ways to improve it
- Apply these concepts to a big data analytics problem



Introduction



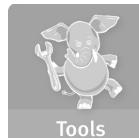
Analytics Lifecycle



Basic Methods



Adv. Methods



Tools



Lab

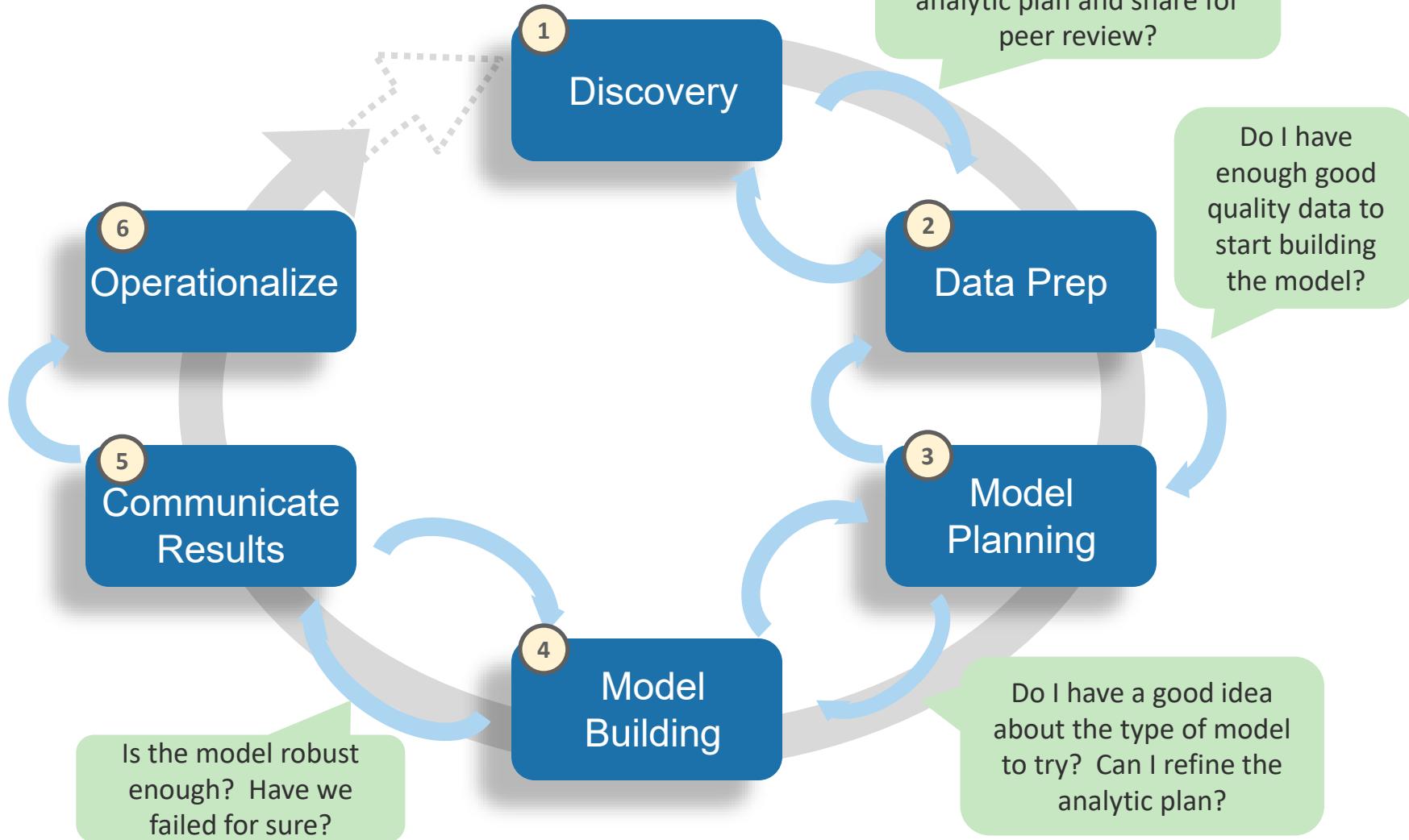
Module 5 – The Endgame, or Putting it All Together

Lesson 1: Operationalizing an Analytics Project

During this lesson the following topics are covered:

- Operationalizing a data analytics lifecycle project
- Key outputs needed for a successful analytic project, by stakeholder role
- 4 core deliverables to meet most stakeholder needs

Data Analytics Lifecycle



Data Analytics Lifecycle

Final Deliverables



6

Operationalize

Communicate Results

Is the model robust enough? Have we failed for sure?

Discovery

- Run a pilot
- Assess the benefits
- Provide final deliverables
- Implement the model in the production environment
- Define process to update, retrain, and retire the model, as needed

Model Building

Do I have enough information to draft an analytic plan and share for peer review?

Do I have

Do I have a good idea about the type of model to try? Can I refine the analytic plan?



Key Outputs from a Successful Analytic Project, by Role

Role	Description	What the Role Needs in the Final Deliverables
Business User	Someone who benefits from the end results and can consult and advise project team on value of end results and how these will be operationalized	<ul style="list-style-type: none">• Sponsor Presentation addressing:<ul style="list-style-type: none">• Are the results good for me?• What are the benefits of the findings?• What are the implications of this for me?
Project Sponsor	Person responsible for the genesis of the project, providing the impetus for the project and core business problem, generally provides the funding and will gauge the degree of value from the final outputs of the working team	<ul style="list-style-type: none">• Sponsor Presentation addressing:<ul style="list-style-type: none">• What's the business impact of doing this?• What are the risks? ROI?• How can this be evangelized within the organization (and beyond)?
Project Manager	Ensure key milestones and objectives are met on time and at expected quality.	
Business Intelligence Analyst	Business domain expertise with deep understanding of the data, KPIs, key metrics and business intelligence from a reporting perspective	<ul style="list-style-type: none">• Show the analyst presentation• Determine if the reports will change
Data Engineer	Deep technical skills to assist with tuning SQL queries for data management, extraction and support data ingest to analytic sandbox	<ul style="list-style-type: none">• Share the code from the analytical project• Create technical document on how to implement it.
Database Administrator (DBA)	Database Administrator who provisions and configures database environment to support the analytical needs of the working team	<ul style="list-style-type: none">• Share the code from the analytical project• Create technical document on how to implement it.
Data Scientist	Provide subject matter expertise for analytical techniques, data modeling, applying valid analytical techniques to given business problems and ensuring overall analytical objectives are met	<ul style="list-style-type: none">• Show the analyst presentation• Share the code

4 Core Deliverables to Meet Most Stakeholder Needs



1. Presentation for Project Sponsors

- “Big picture” takeaways for executive level stakeholders.
- Determine key messages to aid their decision-making process.
- Focus on clean, easy visuals for the presenter to explain and for the viewer to grasp.

2. Presentation for Analysts

- Business process changes.
- Reporting changes.
- Fellow data scientists will want the details and are comfortable with technical graphs (such as ROC curves, density plots, histograms).

3. Code for technical people

4. Technical specs of implementing the code



Introduction



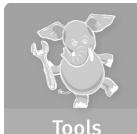
Analytics Lifecycle



Basic Methods



Adv. Methods



Tools



Lab

Module 5: The Endgame, or Putting it All Together

Lesson 1: Summary

During this lesson the following topics were covered:

- Operationalizing a data analytics lifecycle project
- Key outputs needed for a successful analytic project, by stakeholder role
- 4 core deliverables to meet most stakeholder needs



Introduction



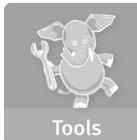
Analytics Lifecycle



Basic Methods



Adv. Methods



Tools



Lab

Module 5: The Endgame, or Putting it All Together

Lesson 2: Creating the Final Deliverables

During this lesson the following topics are covered:

- Brief review of YoyoDyne case study
- Using a core set of materials to deliver presentations for two different audiences
- Comparing the main focus areas for sponsors and analyst audiences
- Using a framework to organize the main pieces of your final presentations
- Tips for sharing your code and technical documentation



YoyoDyne Churn Prediction Case Study

Mini Case Study: Churn Prediction for Yoyodyne Bank

Situation Synopsis

- Retail Bank, Yoyodyne Bank wants to improve the Net Present Value (NPV) and retention rate of the customers .
- They want to establish an effective marketing campaign targeting customers to reduce the churn rate by at least five percent.
- The bank wants to determine whether those customers are worth retaining. In addition, the bank also wants to analyze reasons for customer attrition and what they can do to keep them.
- The bank wants to build a data warehouse to support marketing and other related customer care groups.

Use Analytic Plan to Guide Final Presentation

Components of Analytic Plan	Retail Banking: Yoyodyne Bank
Discovery Business Problem Framed	How do we identify churn/no churn for a customer?
Initial Hypotheses	Transaction volume and type are key predictors of churn rates
Data & Scope	5 months of customer account history
Model Planning - Analytic Technique	Logistic regression to identify most influential factors predicting churn
Result & Key Findings	<p>Once customers stop using their accounts for gas and groceries, they will soon erode their accounts and churn.</p> <p>If customers use their debit card fewer than 5 times per month, they will leave the bank within 60 days.</p>
Business Impact	<p>If we can target customers who are high-risk for churn, we can reduce customer attrition by 25%. This would save \$3 million in lost customer revenue and avoid \$1.5 million in new customer acquisition costs each year.</p>

Key Aspects of Final Presentation Material

- **Reflect on the project:**
 - ▶ Consider the context of the problems you set out to solve.
 - ▶ Identify observations about the model outputs, scoring, results.
 - ▶ Identify Key Messages, and any unexpected insights.
- **Tailor outputs to the audience**

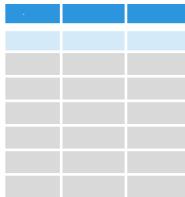
	Project Sponsor Presentation	Analyst Presentation
Focus	What	How
Objectives	<ul style="list-style-type: none">• Show <u>that</u> you met the project goals• Give your sponsor talking points to evangelize the work<ul style="list-style-type: none">• Emphasize ROI and business value• Mention if the models can be deployed within sponsor's SLA	<ul style="list-style-type: none">• Show <u>how</u> you met the project goals• Share your methods so analysts can learn from it for future projects<ul style="list-style-type: none">• Discuss methods, techniques, and technologies used.• Provide specific model accuracy and speed (example: how well will it meet SLAs).

Develop Core Material you can use to Deliver Presentations to 2 Main Audiences

= Same components for both presentations

= Different components for Sponsor vs. Analyst presentation

Presentation Component	Project Sponsor Presentation	Analyst Presentation
Project Goals	<ul style="list-style-type: none">• List top 3 agreed upon goals	<ul style="list-style-type: none">• List top 3 agreed upon goals
Main Findings	<ul style="list-style-type: none">• Emphasize key message	<ul style="list-style-type: none">• Emphasize key message
Approach	<ul style="list-style-type: none">• High Level Methodology	<ul style="list-style-type: none">• High Level Methodology• Relevant details on modeling techniques and technology
Model Description	<ul style="list-style-type: none">• Overview of the modeling technique	<ul style="list-style-type: none">• Overview of the modeling technique
3 key points supported with data	<ul style="list-style-type: none">• Support key points with simple charts and graphics (example: bar charts)	<ul style="list-style-type: none">• Show details to support the key points• Analyst-oriented charts and graphs (ROC curves, histograms)• Visuals of key variables and significance of each
Model Details	<ul style="list-style-type: none">• Omit this section, or discuss only at a very high level	<ul style="list-style-type: none">• Show the code or main logic of the model, Include the model type, variables, technology used to execute it and score data.• Identify key variables and impact of each• Describe expected model performance and any caveats• Detailed description of the modeling technique• Discuss variables, scope, predictive power
Recommendations	<ul style="list-style-type: none">• Focus on business impact of doing this, including risks and ROI• Give the sponsor salient points to help him or her evangelize the work within the organization	<ul style="list-style-type: none">• Supplement recommendations with any implications for the modeling, or for deploying in a production environment.



Project Goals

Presentation Component	Project Sponsor Presentation	Analyst Presentation
Project Goals	<ul style="list-style-type: none">• List top 3 agreed upon goals	<ul style="list-style-type: none">• List top 3 agreed upon goals

Example 1 of Goals slide

Project Goals

1. Develop a predictive model to determine which customers are most likely to churn and when
2. Model's predictive power should be at least as good as customer retention techniques currently being used by the bank
3. Models should scale to run on a full data set in production environment on weekly basis

Example 2 of Goals slide, with Situation overview

Situation & Project Goals

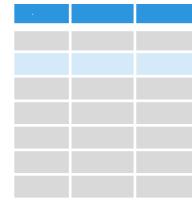
Situation

1. Yoyodyne Bank wants to improve the Net Present Value (NPV) and retention rate of the customers .
2. In last 90 days, Yoyodyne has lost 6 of its top 100 customers, and is seeing increased competition from their biggest competitor
3. Without a fast remediation plan, Yoyodyne risks losing its dominant position in three key markets

Goals of Yoyodyne "Churn Project"

1. Develop a predictive model to determine which customers are most likely to churn and when
2. Model's predictive power should be at least as good as customer retention techniques currently being used by the bank
3. Models should scale to run on a full data set in production environment on weekly basis

Main Findings (Executive Summary)



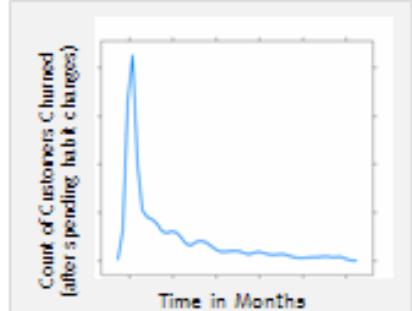
Presentation Component	Project Sponsor Presentation	Analyst Presentation
Main Findings (executive summary)	<ul style="list-style-type: none">Emphasize key message	<ul style="list-style-type: none">Emphasize key message

- Enable reader to grasp full synopsis in 1 slide
- Frame outcomes in terms of business value
- Generally same, or very similar, for both types of audiences

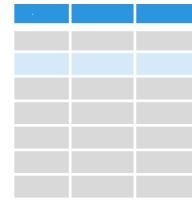
Executive Summary

Running an early churn warning test each day using social media data can reduce annual churn by 30% and save \$4.5M annually

- Customers churn within 60 days of changing their spending habits
 - Most often after customers stop using bank cards for gas and grocery
 - If customers use their debit card fewer than 5 times per month, they will leave the bank within 60 days.
- Combining social networking data and existing CRM data increases the model's predictive power to identify churners
 - We can pinpoint social media chatter from bank customers and influence of chunner's contacts
 - With CRM data we can identify 20% of churners, adding social media data increases this to 30%
- Models can run in minutes, rather than current process of monthly cycles



Anatomy of an Executive Summary



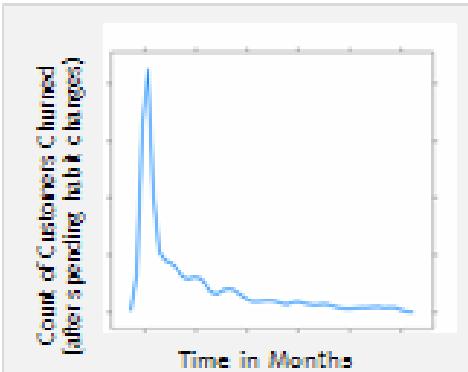
Business Impact

Key Message

Executive Summary

Running an early churn warning test each day using social media data can reduce annual churn by 30% and save \$4.5M annually

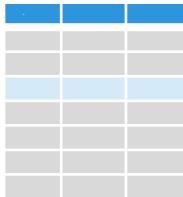
- Customers churn within 60 days of changing their spending habits
 - ▶ Most often after customers stop using bank cards for gas and grocery
 - ▶ If customers use their debit card fewer than 5 times per month, they will leave the bank within 60 days.
- Combining social networking data and existing CRM data increases the model's predictive power to identify churners
 - ▶ We can pinpoint social media chatter from bank customers and influence of chunner's contacts
 - ▶ With CRM data we can identify 20% of chunners, adding social media data increases this to 30%



Major Points

SLA

- Models can run in minutes, rather than current process of monthly cycles



Approach

Presentation Component	Project Sponsor Presentation	Analyst Presentation
Approach	<ul style="list-style-type: none">• High Level Methodology	<ul style="list-style-type: none">• High Level Methodology• Relevant details on modeling techniques and technology

Example Approach slide, for Sponsors

Approach (for Sponsors)

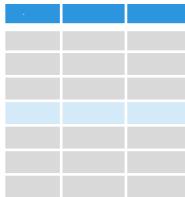
- Interviewed 14 members of retail lending team to understand Yoyodyne's lending policies and marketing practices for customer retention
- Collaborated with IT to identify relevant data sets, assess data quality and availability
- Developed churn model to identify customers most likely to leave the bank
 - ▶ Identify most influential factors
 - ▶ Provides greater explanatory power for analyzing impact of different factors on churn
- Mined and added social media data to the model to improve predictive power
- Worked with IT to simulate model performance within Yoyodyne's production environment

Example Approach slide, for Analysts

Approach (for Analysts)

- Interviewed 14 members of retail lending team to understand Yoyodyne's lending policies and marketing practices for customer retention
- Collaborated with IT to identify relevant data sets, assess data quality and availability
- Developed churn model in R using a Generalized Additive Modeling technique
 - ▶ Minimizes variable transformations and binning
 - ▶ Provides greater explanatory power for analyzing impact of different factors on churn
- Impact of social network variables was examined and found to help identify more potential churners
- Worked with IT to simulate model performance within Yoyodyne's production environment
- The model can be rapidly scored in the database over large datasets using a SQL code generator for the purpose

Note: Green boxes highlight differences between slides



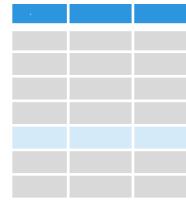
Model Description

Presentation Component	Project Sponsor Presentation	Analyst Presentation
Model Description	<ul style="list-style-type: none">• Overview of the modeling technique	<ul style="list-style-type: none">• Overview of the modeling technique

Model Description

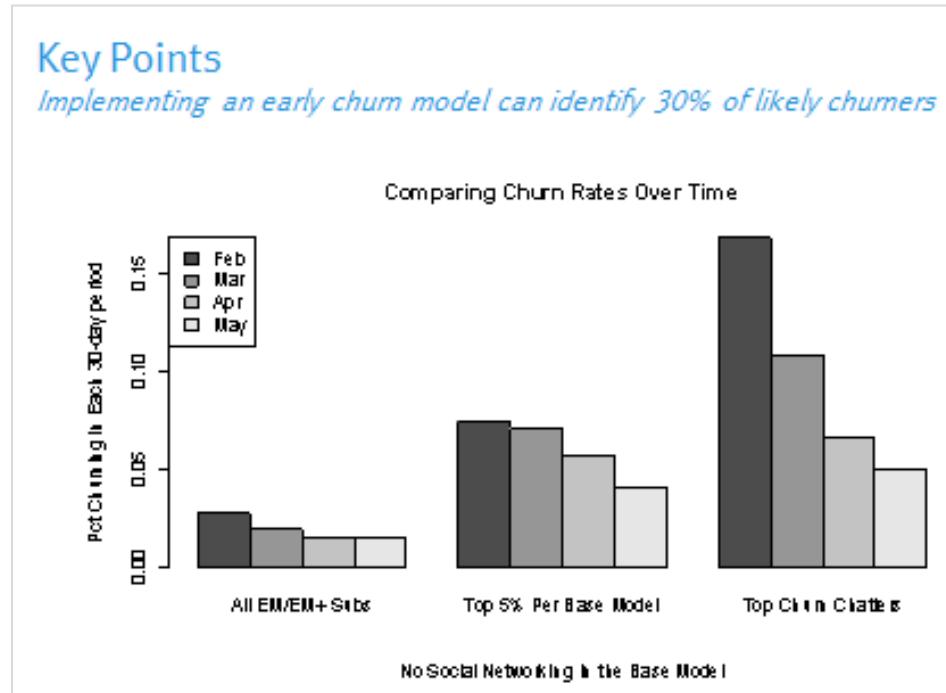
- **Overview of Basic Methodology:** predict the likelihood of churn for each customer. Identify customers with a greater probability for churn then compare with actual churn outcomes to train the algorithm and enable predictions for existing customers.
- **Model:** Logistic regression model
- **Dependent variable:** Binary variable, of churn/no churn
- **Scope:**
 - ▶ 500,000 Yoyodyne bank customers, based on churn within a 150 day period after 1/31/2011
 - ▶ 500,000 Customers with all churners through 6/30/11, plus a random sample of 45,000 accounts
 - ▶ All selected customers were Active, Suspended or Pending as of 2011-01-31
 - ▶ Call History detail data extracted from Call Data Record Warehouse for customers from 1/31/11 to 6/30/11
- **Sampling**
 - ▶ Training sample: 50,000 subscribers
 - ▶ Testing sample: 100,000 subscribers
- **The model developed has predictive power at least as good as the bank's current churn model**
 - ▶ We created a baseline model without social networking variables and the bank's marketing analytics team verified that the predictive power was at least as good as the current model
 - ▶ Social networking variables were added to the model and that further increased its predictive power

Key Points Supported With Data

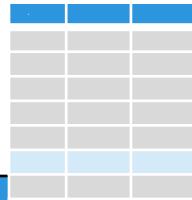


Presentation Component	Project Sponsor Presentation	Analyst Presentation
Key Points Supported With Data	<ul style="list-style-type: none">Support key points with simple charts and graphics (such as bar charts)	<ul style="list-style-type: none">Show details to support the key pointsAnalyst-oriented charts and graphs (ROC curves, histograms)Visuals of key variables and significance of each

- Identify key points based on your insights and observations resulting from the data and model scoring results
- Illustrate your key points with charts and visualizations
- Use simpler charts for Sponsors



Model Details



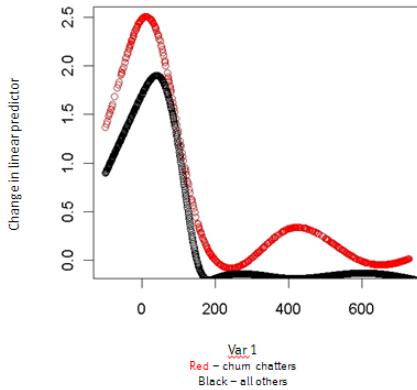
Presentation Component	Project Sponsor Presentation	Analyst Presentation
Model Details	<ul style="list-style-type: none">Omit this section, or discuss only at a very high level	<ul style="list-style-type: none">Show the code or main logic of the model, Include the model type, variables, technology used to execute it and score data.Describe expected model performance and any caveatsDetailed description of the modeling technique, variables, scope, predictive power

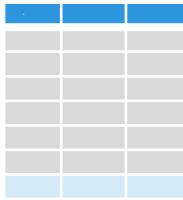
Model Details

- Candidate variables: 22 from CRM, 154 from call history, and 12 social networking variables
- Through PCA and discussion with domain experts, we reduced ~190 variables to the 9 most predictive of customer churn
- General Additive Model (GAM) model built in R :

```
gam.wsn.by2 <- bam(volchurn.120.p~  
  s(var1, bs="cs",by=c30,k=length(custom.knots))  
  +s(var2, bs="cs",by=c30)  
  +s(var3, bs="cs",k=5)  
  +s(var4,bs="cs",k=5,by=c30)  
  +s(tvar5,bs="cs",k=5)  
  +var6  
  +var7  
  +s(var8)  
  +s(var9),  
  knots=list(var1=custom.knots),  
  data=train.df,family=binomial, weight=weight, gamma=1.4)
```

Var 1 has a larger and earlier impact on churn-chatters





Recommendations

Presentation Component	Project Sponsor Presentation	Analyst Presentation
Recommendations	<ul style="list-style-type: none">Focus on business impact of the project, including risks and ROIGive the sponsor salient points to help him or her evangelize the work within the organization	<ul style="list-style-type: none">Supplement recommendations with any implications for the modeling, or for deploying in a production environment.
<h3>Recommendations</h3> <ul style="list-style-type: none">Implement the model as a pilot, before more wide-scale rollout – test and learn from initial pilot on performance and precision.<ul style="list-style-type: none">Addressing these promptly can potentially save more customers from churning over time and also prevent more networking that seems to drive additional churn.An early churn warning trigger can be set up based on this model.Run the predictive model daily or weekly to be proactive on customer churn<ul style="list-style-type: none">In-database scorer can score large datasets in a matter of minutes and can be run dailyEach customer retained via early warning trigger saves 4 hours of account retention efforts & 50k in new account acquisition costsDevelop targeted customer surveys to investigate the causes of churn, which will make the collection of data for investigation into the causes of churn easier.		

Quick Summary of Final Presentation Components

Situation & Project Goals

Situation

1. Yoyodyne Bank wants to improve the Net Present Value (NPV) and retention rate of the customers .
2. In last 90 days, Yoyodyne has lost 6 of its top 100 customers, and is seeing increased competition from their biggest competitor
3. Without a fast remediation plan, Yoyodyne risks losing its dominant position in three key markets

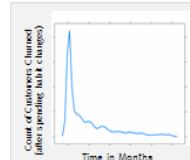
Goals of Yoyodyne "Churn Project"

1. Develop a predictive model to determine which customers are most likely to churn and when
2. Model's predictive power should be at least as good as customer retention techniques currently being used by the bank
3. Models should scale to run on a full data set in production environment on weekly basis

Executive Summary

Running an early churn warning test each day using social media data can reduce annual churn by 30% and save \$4.5M annually

- Customers churn within 60 days of changing their spending habits
 - Most often after customers stop using bank cards for gas and grocery
 - If customers use their debit card fewer than 5 times per month, they will leave the bank within 60 days.
- Combining social networking data and existing CRM data increases the model's predictive power to identify churners
 - We can pinpoint social media chatter from bank customers and influence of churner's contacts
 - With CRM data we can identify 20% of churners, adding social media data increases this to 30%
- Models can run in minutes, rather than current process of monthly cycles



Approach (for Sponsors)

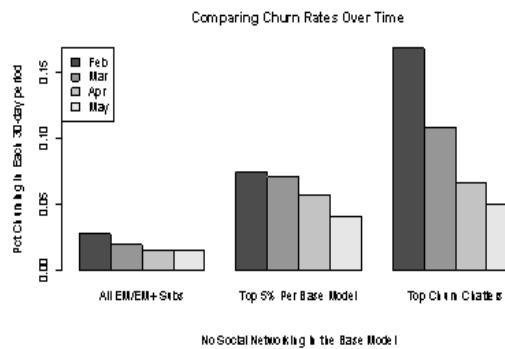
- Interviewed 14 members of retail lending team to understand Yoyodyne's lending policies and marketing practices for customer retention
- Collaborated with IT to identify relevant data sets, assess data quality and availability
- Developed churn model to identify customers most likely to leave the bank
 - Identify most influential factors
 - Provides greater explanatory power for analyzing impact of different factors on churn
- Mined and added social media data to the model to improve predictive power
- Worked with IT to simulate model performance within Yoyodyne's production environment

Model Description

- Overview of Basic Methodology: predict the likelihood of churn for each customer. Identify customers with a greater probability for churn then compare with actual churn outcomes to train the algorithm and enable predictions for existing customers.
- Model: Logistic regression model
- Dependent variable: Binary variable, of churn/no churn
- Scope:
 - 500,000 Yoyodyne bank customers, based on churn within a 150 day period after 1/31/2011
 - 500,000 Customers with all churners through 6/30/11, plus a random sample of 45,000 accounts
 - All selected customers were Active, Suspended or Pending as of 2011-01-31
 - Call History detail data extracted from Call Data Record Warehouse for customers from 1/31/11 to 6/30/11
- Sampling
 - Training sample: 50,000 subscribers
 - Testing sample: 100,000 subscribers
- The model developed has predictive power at least as good as the bank's current churn model
 - We created a baseline model without social networking variables and the bank's marketing analytics team verified that the predictive power was at least as good as the current model
 - Social networking variables were added to the model and that further increased its predictive power

Key Points

Implementing an early churn model can identify 30% of likely churners



Model Details

- Candidate variables: 22 from CRM, 154 from call history, and 12 social networking variables
- Through PCA and discussion with domain experts, we reduced ~190 variables to the 9 most predictive

Recommendations

- Implement the model as a pilot, before more wide-scale rollout – test and learn from initial pilot on performance and precision.
 - Addressing these promptly can potentially save more customers from churning over time and also prevent more networking that seems to drive additional churn.
 - An early churn warning trigger can be set up based on this model.
- Run the predictive model daily or weekly to be proactive on customer churn
 - In-database scorer can score large datasets in a matter of minutes and can be run daily
 - Each customer retained via early warning trigger saves 4 hours of account retention efforts & 50k in new account acquisition costs
- Develop targeted customer surveys to investigate the causes of churn, which will make the collection of data for investigation into the causes of churn easier.

Top 10 Tips, Tricks, & Pitfalls to Avoid for the Final Presentation

1. Be visual. Generally, the more visual the better. Up to a point.
2. Be MECE (Mutually Exclusive and Collectively Exhaustive).
3. Tie your ideas together....don't force people to tie your ideas together, guide people and help them draw logical connections.
4. Don't forget that not everyone has gone through the Discovery phase like you have.
5. Context is key. Orient people to the project itself, as well as the graphics you use, the terminology and jargon (spell out acronyms).
6. Don't assume people see the obvious benefits.
7. Measure and quantify the benefits. Be specific. “\$8.5M in annual cost savings” is much stronger than “Great Value”.
8. Be patient. You may have to tell your story more than once...consider these sessions opportunities to refine your message and share good work that was done.
9. Let the intended audience guide you in shaping the right message and level of detail.
10. Avoid long bulleted lists ☺

Overview of Code & Technical Documentation

- **Consider the interests of your technical audience:**
 - ▶ How will the project affect them?
 - ▶ In what ways will it change their day-to-day roles, or existing processes?
 - ▶ Be aware of the implications of your work on their roles as you create these technical deliverables.
- **2 Technical deliverables:**
 - ▶ Code
 - ▶ Technical specifications and documentation.

Considerations for Technical Specifications & Documentation

Approach the documentation as if it's for an API (application programming interface)

- **Inputs & Pre-processing:**

- ▶ Discuss the expected pre-processing steps before data goes to the model code.
- ▶ Document expected input, data format, source tables, and units.
- ▶ Describe the processing script are you using.
- ▶ Explain how the outputs are created.

- **Exception handling :**

- ▶ Explain how to deal with exceptions to the model.
- ▶ Provide guidance for making decisions on the exceptions.

- **Post-processing:**

- ▶ After you create the output, discuss any post-processing before going to the next step.
- ▶ Interpreting a threshold as opposed to a simple yes/no.

Providing Your Code

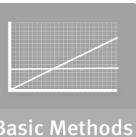
- Test for accuracy in the production environment
- Ensure the code will run quickly and meet SLAs
- Include comment lines in the code
- Hold a briefing with the engineers who will implement the code



Introduction



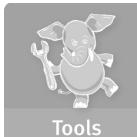
Analytics Lifecycle



Basic Methods



Adv. Methods



Tools



Lab

Module 5: The Endgame, or Putting it All Together

Lesson 2: Summary

During this lesson the following topics were covered:

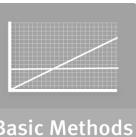
- Brief review of YoyoDyne case study
- Using a core set of materials to deliver presentations for two different audiences
- Comparing the main focus areas for sponsors and analyst audiences
- Using a framework to organize the main pieces of your final presentations
- Tips for sharing your code and technical documentation



Introduction



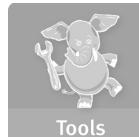
Analytics Lifecycle



Basic Methods



Adv. Methods



Tools



Lab

Module 5: “The Endgame” or Putting it All Together

Lesson 3: Data Visualization Techniques

During this lesson the following topics are covered:

- Survey of data visualization tools
- Creating different visualizations for sponsors and analysts
- Developing visuals to support your key points
- How to clean up a chart or visualization
- Tips and tricks

Key Points Supported With Data

Overview of Visualization Tools

Open Source

- R
 - ▶ Base package
 - ▶ ggplot
 - ▶ Lattice
- Ggobi/Rggobi
- Inkscape
- Processing
- Modest Maps
- GnuPlot

Commercial Tools

- Tableau
- Spotfire
- Qlikview
- Adobe Illustrator

Key Points Supported With Data

Tables of Information

44 years of BigBox stores data

Year	1962	1964	1965	1967	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	Grand Total
SuperBox	1	1	1	1	5	4	4	14	13	14	20	14	17	29	24	37	33	117	42	65	79	81	90	92	82	86	106	72	62	40	49	22	26	33	47	78	71	67	64	91	91	33	1980	
BigBox					1	1	1	4	5	5	10	10	6	21	33	21	22	20	29	31	50	52	43	45	72	91	76	94	67	80	31	34	33	33	27	35	47	32	39	27	4	1196		
Grand Total	1	1	1	2	5	5	5	15	17	19	25	19	27	39	34	43	54	150	63	87	99	110	121	142	125	131	178	163	138	156	107	129	53	60	66	80	105	106	114	96	130	118	37	3176

34 years of BigBox stores data

Year	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	Grand Total		
SuperBox	13	14	20	14	17	29	24	37	33	117	42	65	79	81	90	92	82	86	106	72	62	62	40	49	22	33	47	78	71	67	64	91	91	33	1980			
BigBox				4	5	5	5	10	10	10	6	21	33	21	22	20	29	31	50	43	45	72	91	76	94	67	80	31	34	33	27	35	47	32	39	27	4	1196
Grand Total	17	19	25	19	27	39	34	43	54	150	63	87	99	110	121	142	125	131	178	163	138	156	107	129	53	60	66	80	105	106	114	96	130	118	37	3176		

- What do you observe from this data?
- What's the main message?
- What is the author trying to emphasize with the data?
- Tailor outputs to the audience

Key Points Supported With Data

Using Visuals to Illustrate Key Points

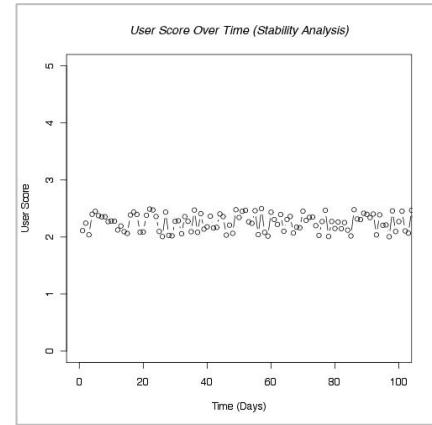
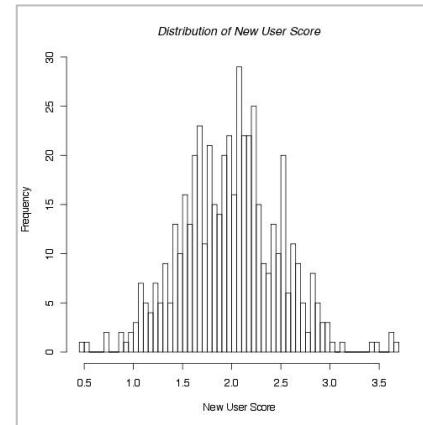
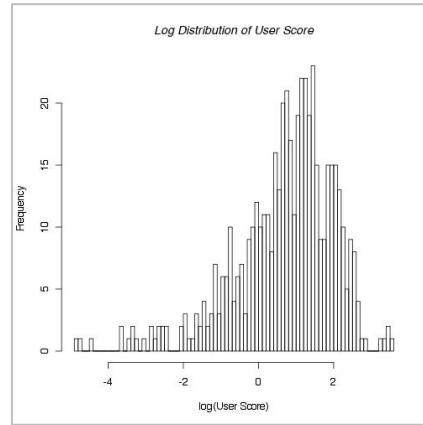
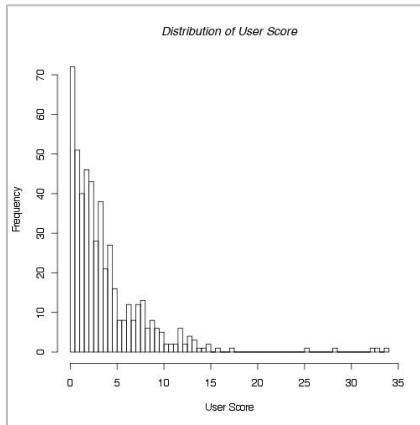
Example of a Visual to help tell a story to a Sponsor



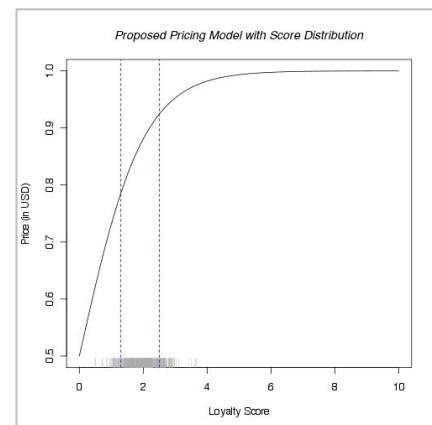
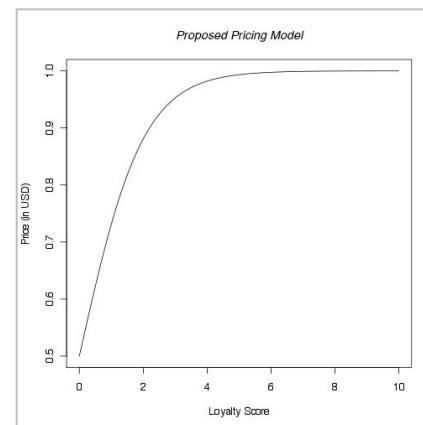
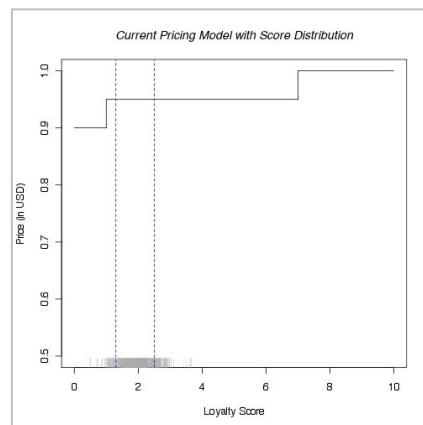
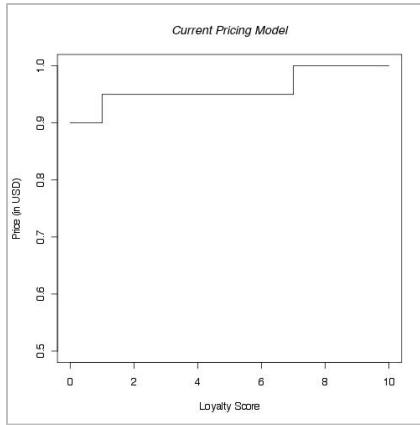
Evolution of a Graph

Hypothetical Example: Exploring Pricing Data

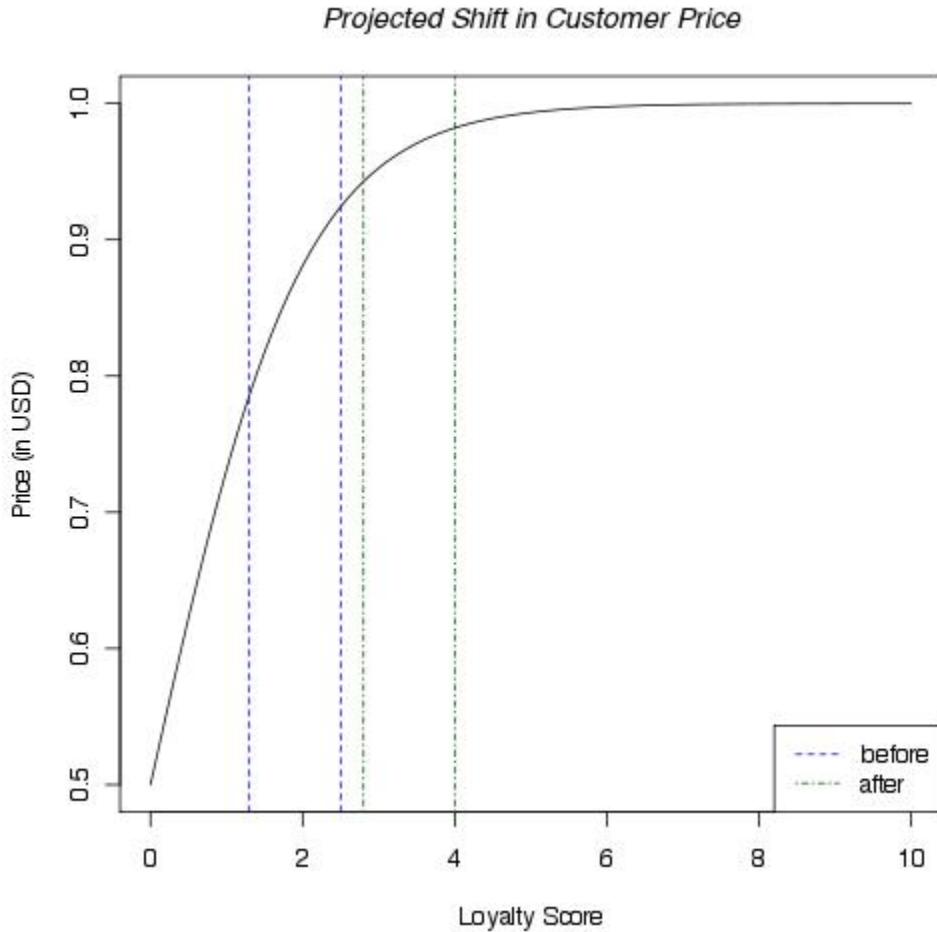
Example of exploring customer price data, price distributions and stability over time



Example of exploring price tiering for most and least loyal customers

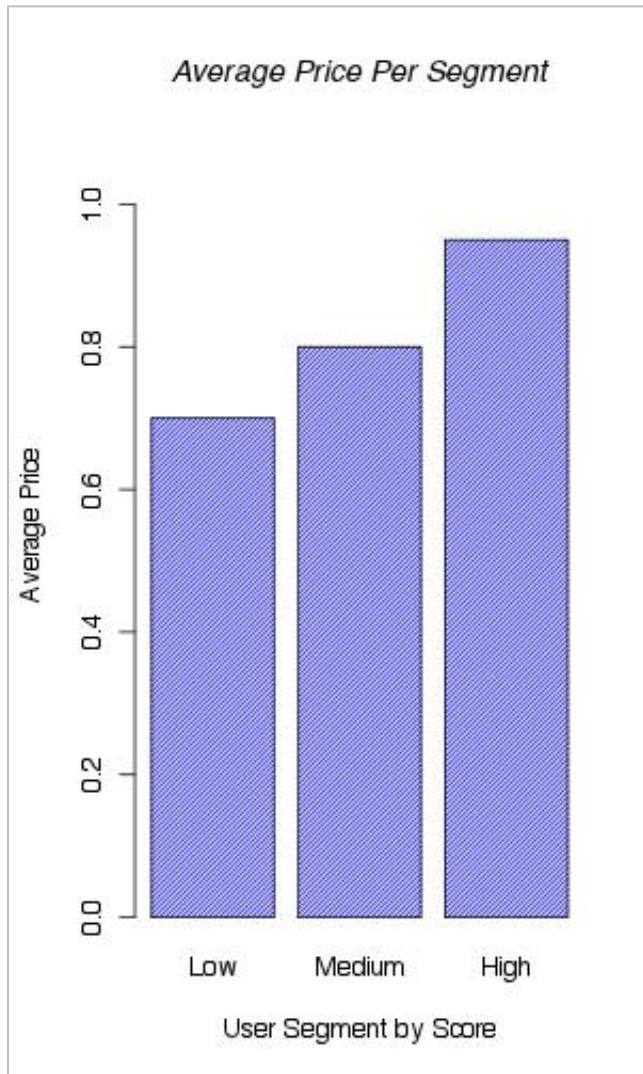


Evolution of a Graph, Analyst Example



- Implementing new price tiering approach increases the precision of price promotions by 23%
- Price optimization model explains 92% of customer behavior
- Model can be run in production environment on daily basis, if needed, to tailor changes to direct mail campaigns and web promotional offers

Evolution of a Graph, Sponsor Example



- Before the project, pricing promotions were offered to all customers equally
- With the new approach:
 - ▶ Highly loyal customers do not receive as many price promotions, since their loyalty is not strongly influenced by price
 - ▶ Customers with low loyalty are influenced by price, and we can now target them for this purpose better
- We project multiple cost savings with this approach
 - ▶ \$2M in lost customers
 - ▶ \$1.5M in new customer acquisition costs
 - ▶ \$1M in reductions for pricing promotions

Key Points Supported With Data

Common Representation Methods

If you want to compare this kind of information....	...consider this kind of chart
Components	Pie chart
Item	Bar chart
Time Series	Line chart
Frequency	Line charts, histograms
Correlation	Scatterplot, side-by-side bar charts

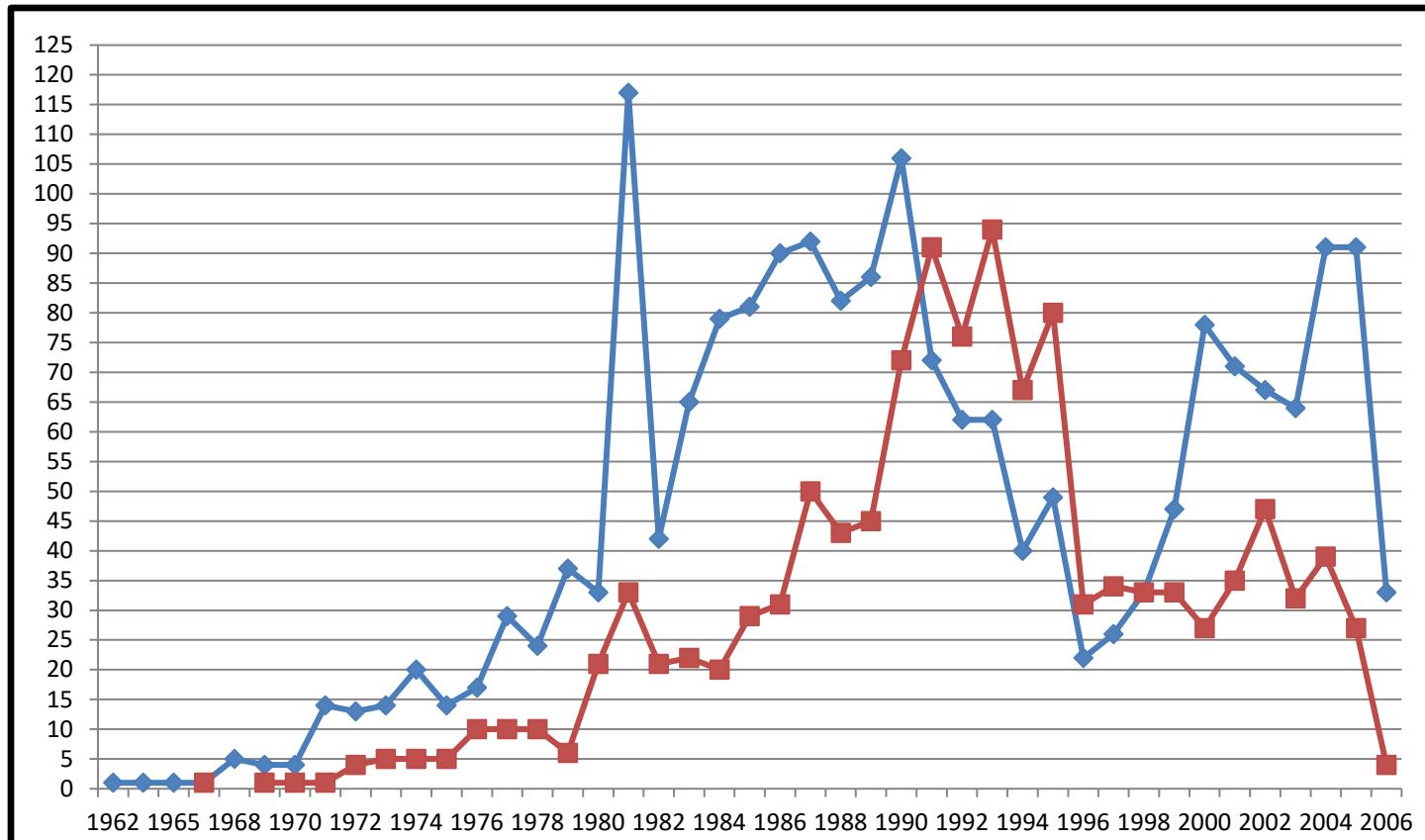
How to Clean Up a Graphic, Example 1

The Before Picture

- What are the main messages here? What is the author trying to emphasize?
- What's wrong with this picture?

Chart junk

1. Horizontal Grid Lines
2. Chunky data points
3. Overuse of emphasis colors; lines & border
4. No context or labels
5. Crowded axis labels



How to Clean Up a Graphic, Example 1

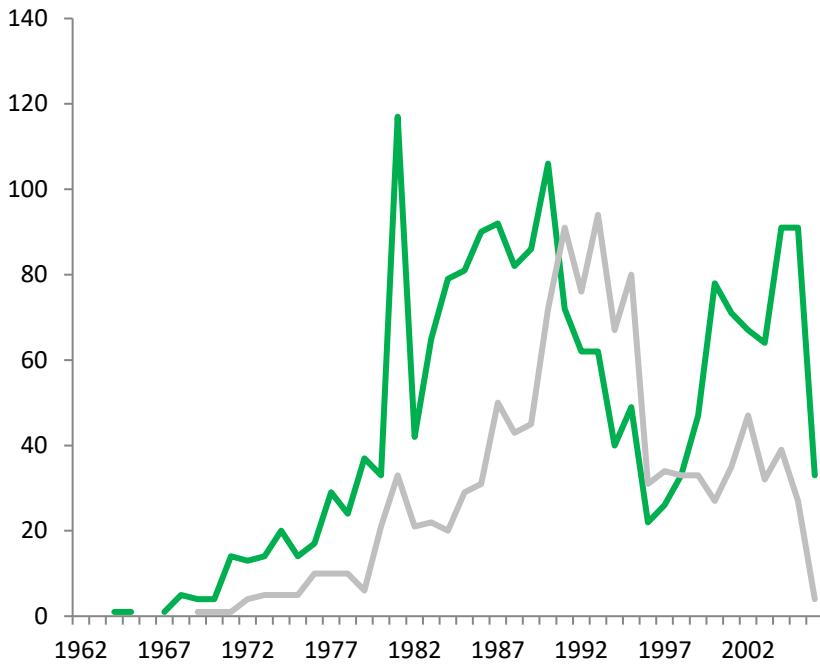
After

- What are the main messages here?
- What is the author trying to emphasize?

Growth of SuperBox Stores

(Count of Stores)

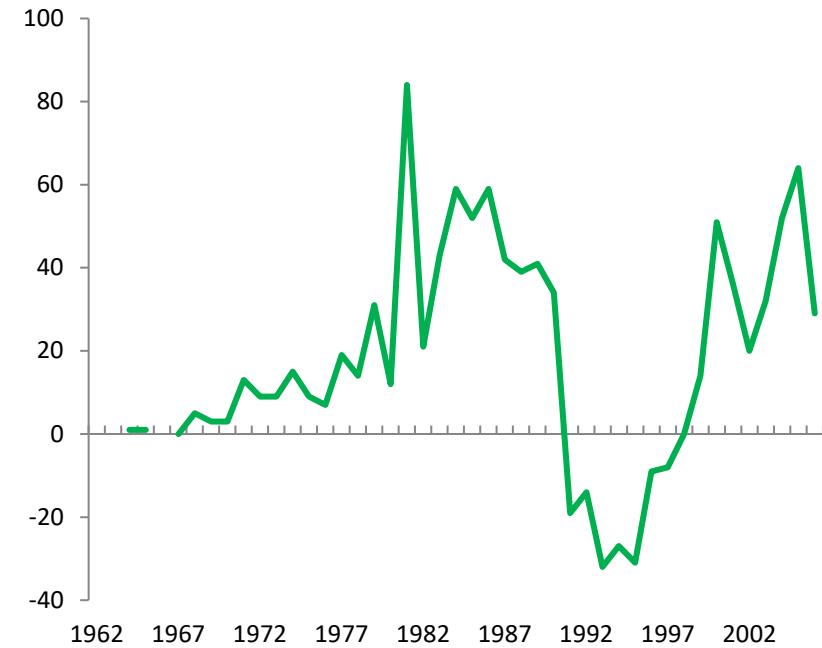
— SuperBox — BigBox



Difference in Store Openings

(Count of SuperBox - Count of BigBox Stores)

— Diff in SuperBox vs. BigBox



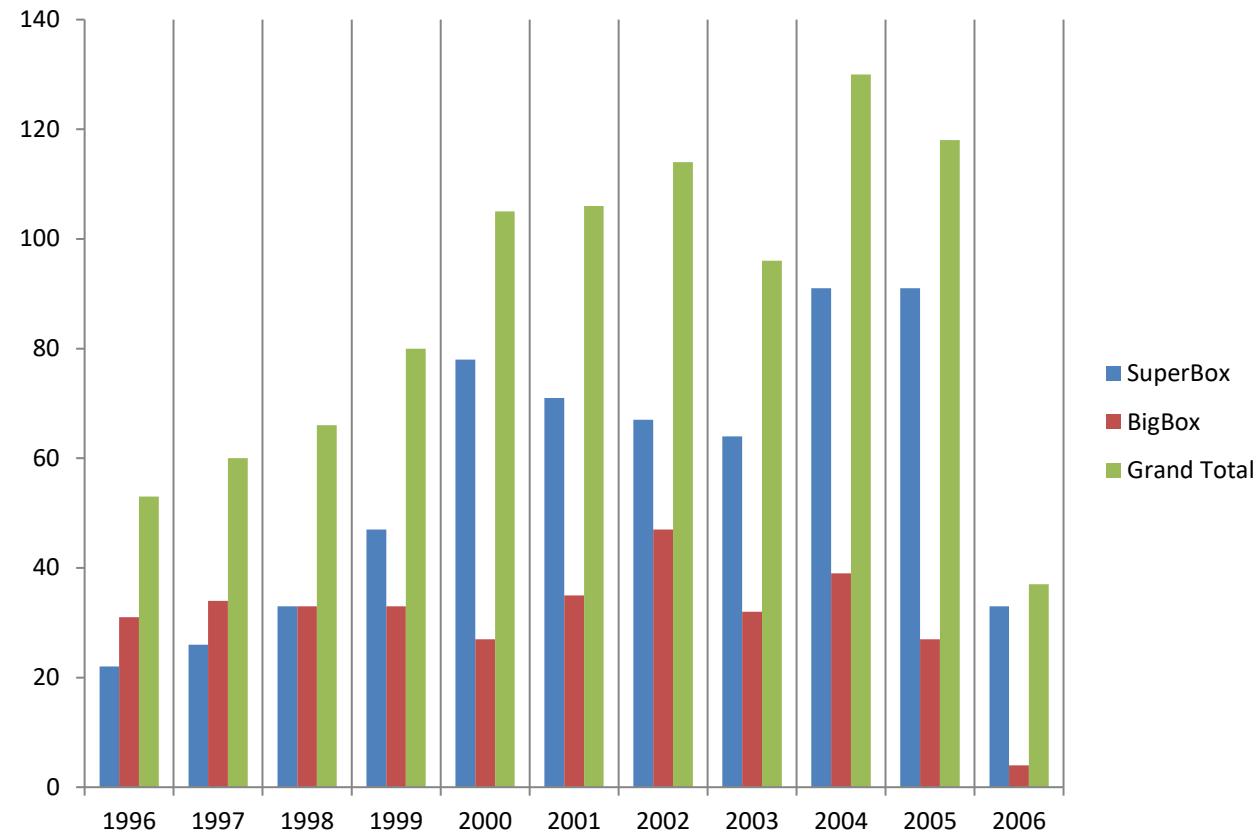
How to Clean Up a Graphic, Example 2

The Before Picture

- What are the main messages here? What is the author trying to emphasize?
- What's wrong with this picture?

Chart junk

1. Vertical Grid Lines
2. Too much emphasis colors
3. No chart title
4. Legend at right restricts chart space
5. Labels are too small



How to Clean Up a Graphic, Example 2

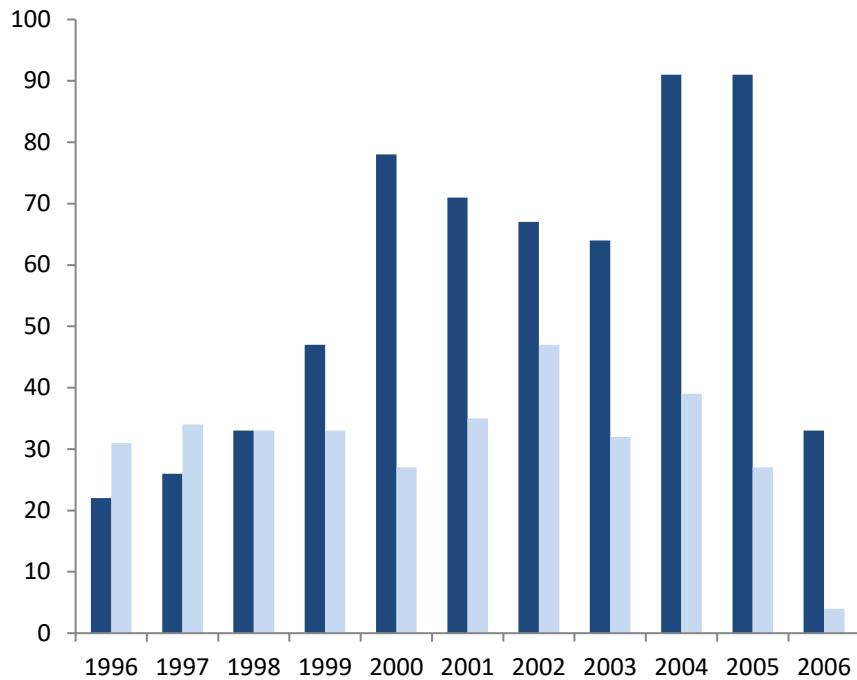
After

- What are the main messages here?
- What is the author trying to emphasize?

Growth of SuperBox Stores

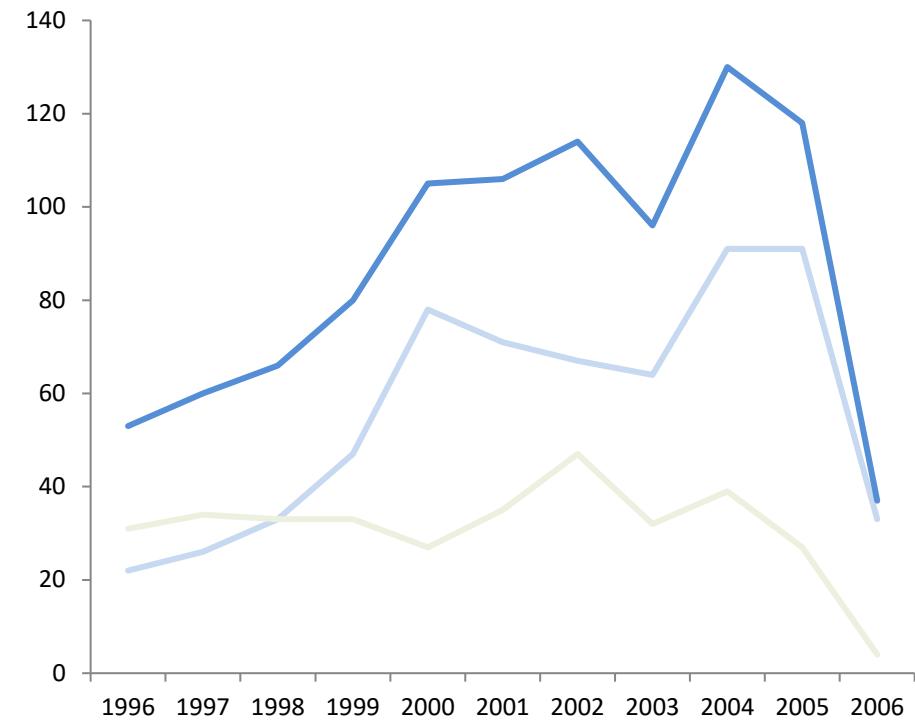
(Count of Stores)

■ SuperBox ■ BigBox



Total Growth of Stores, Over Time

■ SuperBox ■ BigBox ■ Grand Total



A Quick Word About Using 3D Charts: Avoid Them!

2-Dimensional Charts

Growth of SuperBox Stores
(Count of Stores)

■ SuperBox ■ BigBox

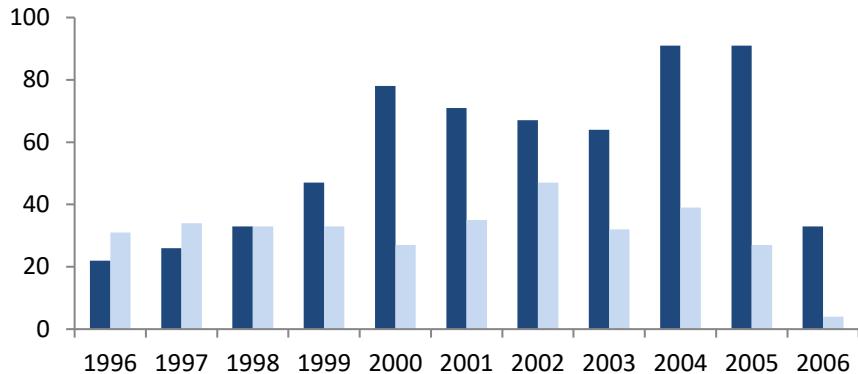


Chart A: 2-Dimensional

- Simple
- Easy to understand
- Focus on the data, not the graphics

3-Dimensional Charts

Growth of SuperBox Stores
(Count of Stores)

■ SuperBox ■ BigBox

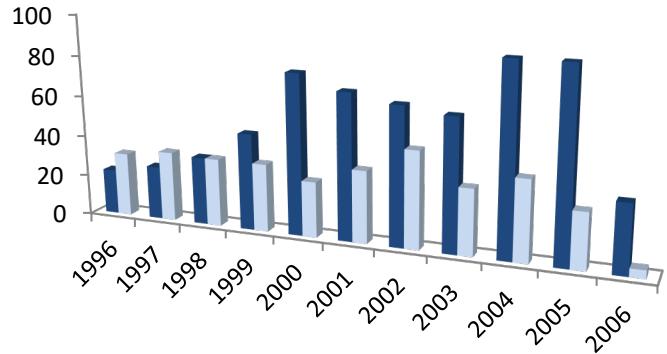


Chart B: 3-Dimensional

- Difficult to gauge actual data
- Scaling becomes deceptive
- Does not make graphic fancier, just harder to understand

Key Points with Data Visualizations

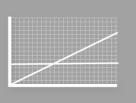
- **Remove distractions**
 - ▶ Minimize “chart junk”
 - ▶ Data-Ink Ratio
- **Choose the simplest, clearest visual for the situation**
 - ▶ Strive to illustrate your points
 - ▶ Charts should serve to reinforce your key points
 - ▶ Charts vs. Data Art
- **Use color deliberately**
 - ▶ Emphasis Colors vs. Standard Colors
 - ▶ In most cases, less is more
 - ▶ Focus on the contrast
- **Context**
 - ▶ Consistent scales, labels, axes
 - ▶ Using logs vs. raw values to show differences



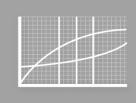
Introduction



Analytics Lifecycle



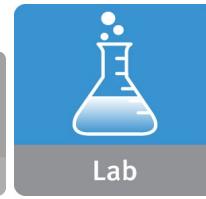
Basic Methods



Adv. Methods



Tools



Lab

Module 5: The Endgame, or Putting it All Together

Lesson 3: Summary

During this lesson the following topics were covered:

- Survey of data visualization tools.
- Creating different visualizations for sponsors and analysts.
- Developing visuals to support your key points.
- How to clean up a chart or visualization.
- Tips and tricks

Check Your Knowledge

- What are the 4 key main deliverables for operationalizing a data analytics project?
- Give an example of an appropriate data visualization for an analyst presentation and one for a sponsor?
- Name 3 considerations for delivering your code or technical documentation
- What is chart junk? What are some ways to address it in a visualization?

Final Lab Exercise on Big Data Analytics



This lab allows students to apply what they have learned from the analytical methods and tools to a big data problem using the Analytics Lab Environment.

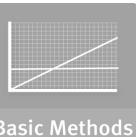
- After completing the tasks in this lab you should be able to:
 - ▶ Explore the big data set provided and prepare the data for analysis.
 - ▶ Assess data quality, outliers and training sets.
 - ▶ Conduct model selection, code, execute and score the model.
 - ▶ Use R and PSQL statements during your analysis of big data.
 - ▶ Create a narrative summary of your findings, using the methods shared earlier in this module.



Introduction



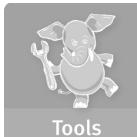
Analytics Lifecycle



Basic Methods



Adv. Methods



Tools



Lab

Module 5: The Endgame, or Putting it All Together

Module 5: Summary

During this module the following topics were covered:

- Three tasks needed to operationalize an analytics project
- Four common deliverables of an analytics lifecycle project meet the needs of key stakeholders
- A framework for creating final presentations for sponsors and analysts
- Evaluation and improvement of data visualizations
- These concepts applied to a big data analytics problem in the final lab

Course Summary

Key points covered in this course:

- Participation as a data science team member on big data and other analytics projects by:
 - ▶ Deploying a structured lifecycle approach to data science and big data analytics projects
 - ▶ Reframing a business challenge as an analytics challenge
 - ▶ Applying analytic techniques and tools to analyze big data, create statistical models, and identify insights that can lead to actionable results
 - ▶ Selecting optimal visualization techniques to clearly communicate analytic insights to business sponsors and others
 - ▶ Using tools such as R and RStudio, MapReduce/Hadoop, in-database analytics, and window and MADlib functions
- How advanced analytics can be leveraged to create competitive advantage and how the Data Scientist role and skills differ from those of a traditional business intelligence analyst

Thank You!