CSCI446/946 Big Data Analytics

Week 13 Advanced Analytical Theory and Methods: The Endgame

School of Computing and Information Technology
University of Wollongong Australia

Advanced Analytical Theory and Methods: The Endgame

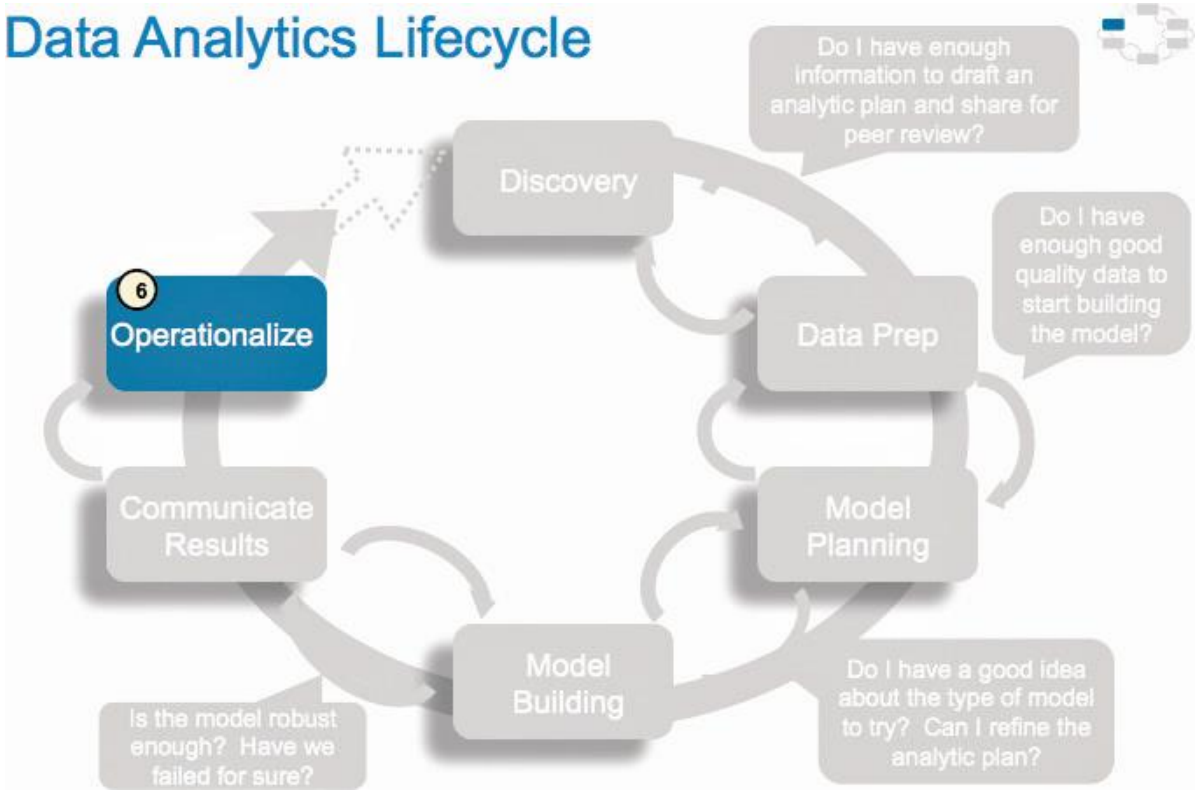
- Overview
- Communicating and Operationalizing an Analytics Project
- Creating the Final Deliverables
- Data Visualization Basics
- Summary

Overview

- This chapter focuses on the final phase of the Data Analytics Lifecycle: operationalize. In this phase, the project team delivers final reports, code, and technical documentation.
- At the conclusion of this phase, the team generally attempts to set up a pilot project and implement the developed models from Phase 4 in a production environment. As stated in Chapter 2, “Data Analytics Lifecycle,” teams can perform a technically accurate analysis, but if they cannot translate the results into a language that resonates with their audience, others will not see the value, and significant effort and resources will have been wasted. This chapter focuses on showing how to construct a clear narrative summary of the work and a framework for conveying the narrative to key stakeholders.

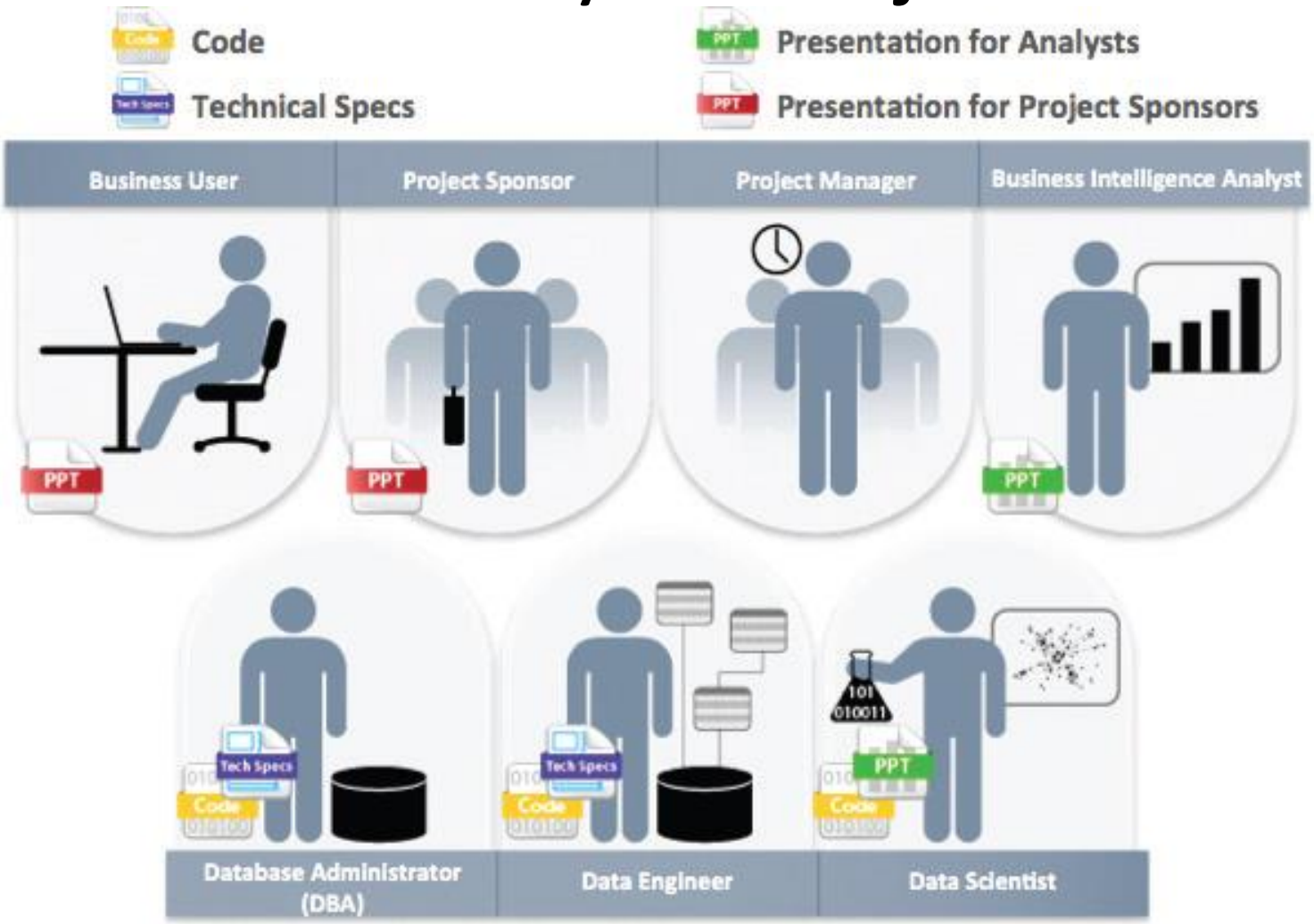
All the figures, tables and codes are from the book “Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data” unless indicated otherwise.

Communicating and Operationalizing an Analytics Project



- As shown in Figure, the final phase in the Data Analytics Lifecycle focuses on operationalizing the project. In this phase, teams need to assess the benefits of the project work and set up a pilot to deploy the models in a controlled way before broadening the work and sharing it with a full enterprise or ecosystem of users. In this context, a pilot project can refer to a project prior to a full-scale rollout of the new algorithms or functionality. This pilot can be a project with a more limited scope and rollout to the lines of business, products, or services affected by these new models.

Communicating and Operationalizing an Analytics Project



- *Key outputs from a successful analytic project*

Communicating and Operationalizing an Analytics Project

- **Business User** typically tries to determine the benefits and implications of the findings to the business.
- **Project Sponsor** typically asks questions related to the business impact of the project, the risks and return on investment (ROI), and how the project can be evangelized within the organization and beyond.
- **Project Manager** needs to determine if the project was completed on time and within budget.
- **Business Intelligence Analyst** needs to know if the reports and dashboards he manages will be impacted and need to change.
- **Data Engineer *and* Database Administrator (DBA)** typically need to share the code from the analytical project and create technical documents that describe how to implement the code.
- **Data Scientists** need to share the code and explain the model to their peers, managers, and other stakeholders.

Communicating and Operationalizing an Analytics Project

Although these seven roles represent many interests within a project, these interests usually overlap, and most of them can be met with four main deliverables:

- **Presentation for Project Sponsors** contains high-level takeaways for executive-level stakeholders, with a few key messages to aid their decision-making process. Focus on clean, easy visuals for the presenter to explain and for the viewer to grasp.
- **Presentation for Analysts**, which describes changes to business processes and reports. Data scientists reading this presentation are comfortable with technical graphs (such as Receiver Operating Characteristic [ROC] curves, density plots, and histograms) and will be interested in the details.
- **Code** for technical people, such as engineers and others managing the production environment
- **Technical specifications** for implementing the code

Communicating and Operationalizing an Analytics Project

- **Creating the Final Deliverables**
  - **Example:** Figure describes a scenario of a fictional bank, YoyoDyne Bank, which would like to embark on a project to do churn prediction models of its customers. *Churn rate* in this context refers to the frequency with which customers sever their relationship as customers of YoyoDyne Bank or switch to

Synopsis of YoyoDyne Bank Case Study

- YoyoDyne Bank is a retail bank that wants to improve its Net Present Value (NPV) and its customer retention rate.
- It wants 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 wants to analyze reasons for customer attrition and what it can do to keep customers from leaving.
- The bank wants to build a data warehouse to support marketing and other related customer care groups.

Creating the Final Deliverables

- Based on this information, the data science team may create an analytics plan a competing bank.

Components of Analytic Plan	Retail Banking: YoyoDyne Bank
Discovery Business Problem Framed	How can the bank identify customers with the highest likelihood for churn?
Initial Hypotheses	Transaction volume and type are key predictors of churn rates
Data and Scope	5 months of customer account history
Model Planning - Analytic Technique	Logistic regression to identify most influential factors predicting churn
Result and Key Findings	Key predictors of churn are: 1. Once customers stop using their accounts for gas and groceries, their account holdings quickly diminish and the customers churn. 2. If the customers use their debit card fewer than 5 times per month, they will leave the bank within 60 days.
Business Impact	By targeting customers who are at high risk for churn, customer attrition can be reduced by 23%. This would save $3 million in lost customer revenue and avoid $1.5 million in new customer acquisition costs each year for the bank.

Creating the Final Deliverables

- **Developing Core Material for Multiple Audiences**

Presentation Component	Project Sponsor Presentation	Analyst Presentation
Project Goals	List top 3–5 agreed-upon goals.	
Main Findings	Emphasize key messages.	
Approach	High-level methodology	High-level methodology Relevant details on modeling techniques and technology
Model Description	Overview of the modeling technique	
Key Points Supported with Data	Support key points with simple charts and graphics (example: bar charts).	Show details to support the key points. Analyst-oriented charts and graphs, such as ROC curves and histograms Visuals of key variables and significance of each
Model Details	Omit this section, or discuss only at a high level.	Show the code or main logic of the model, and include model type, variables, and technology used to execute the model 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, and predictive power.
Recommendations	Focus on business impact, including risks and ROI. Give the sponsor salient points to help her evangelize work within the organization.	Supplement recommendations with implications for the modeling or for deploying in a production environment.

Creating the Final Deliverables

- **Project Goals**

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

Creating the Final Deliverables

Situation & Project Goals

Situation

1. YoyoDyne Bank wants to improve the Net Present Value (NPV) and retention rate of the customers
2. In the last 90 days, YoyoDyne has lost 6 of its top 100 customers and is seeing increased competition from its 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 likey 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

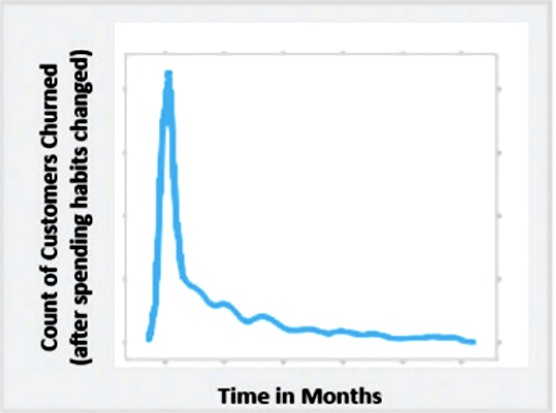
Creating the Final Deliverables

- **Main Findings**

Executive Summary

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

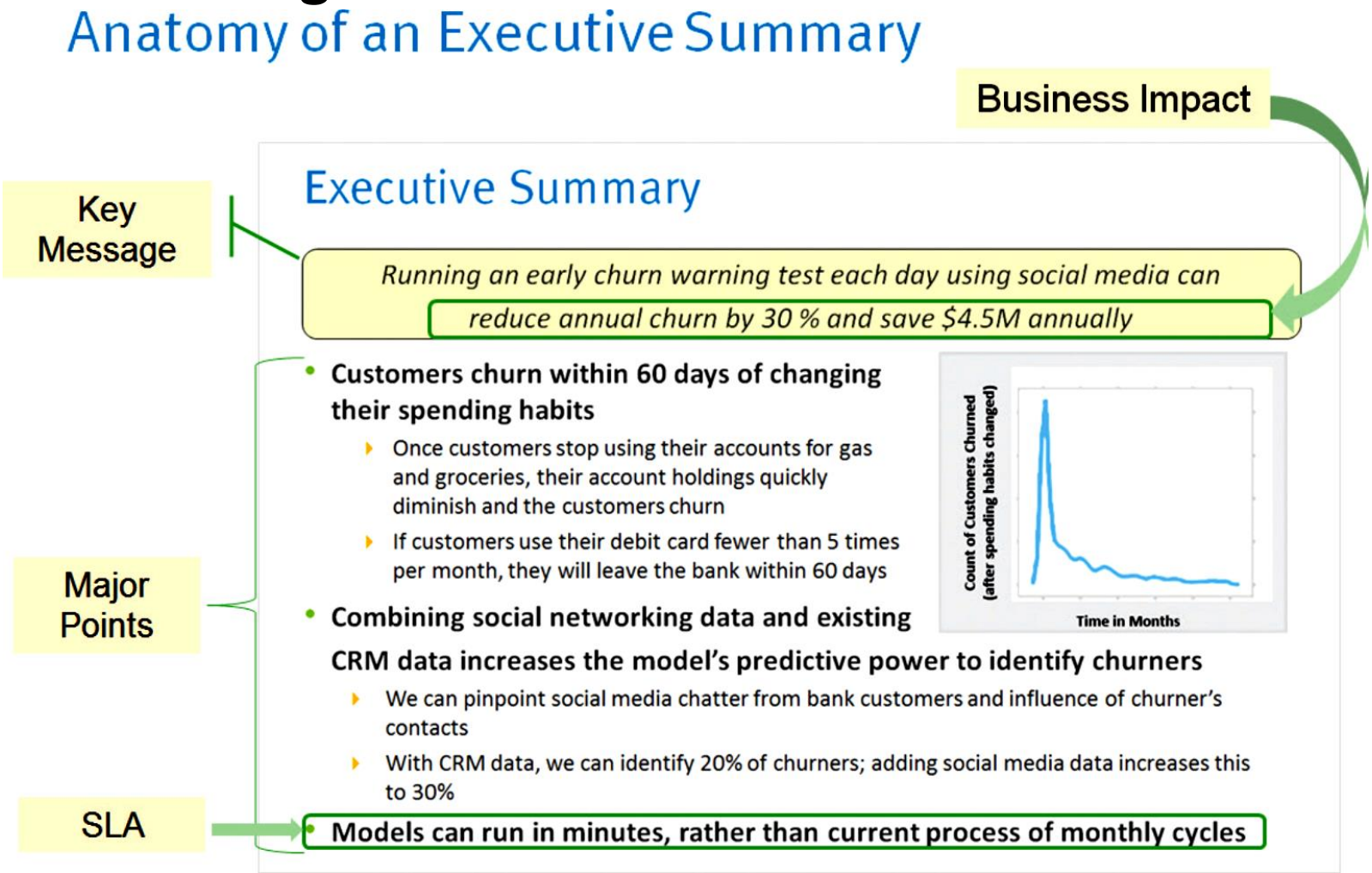
- **Customers churn within 60 days of changing their spending habits**
  - Once customers stop using their accounts for gas and groceries, their account holdings quickly diminish and the customers churn
  - If customers use their debit card fewer than 5 times per month, they will leave the bank within 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**

Creating the Final Deliverables

- **Main Findings**



Creating the Final Deliverables

- **Approach**

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 datasets and assess data quality and availability
- Developed churn model to identify customers most likely to leave the bank
  - Identify most influential factors
  - Provide 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

Creating the Final Deliverables

- **Approach**

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 datasets and assess data quality and availability
- Developed churn model in R using a Generalized Addictive Modeling technique
  - Minimizes variable transformations and binning
  - Provide greater explanatory power for analyzing impact of different factors on churn
- Examined impact of social network variables and found that it helped identify more potential churners
- Work 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

Creating the Final Deliverables

- **Model Description**
  - After describing the project approach, teams generally include a description of the model that was used. Figure provides the model description for the Yoyodyne Bank example. Although the model Description slide can be the same for both audiences, the interests and objectives differ for each. For the sponsor, the general methodology needs to be articulated without getting into excessive detail. Conveythe basic methodology followed in the team's work to allow the sponsor to communicate this to others within the organization and provide talking points.

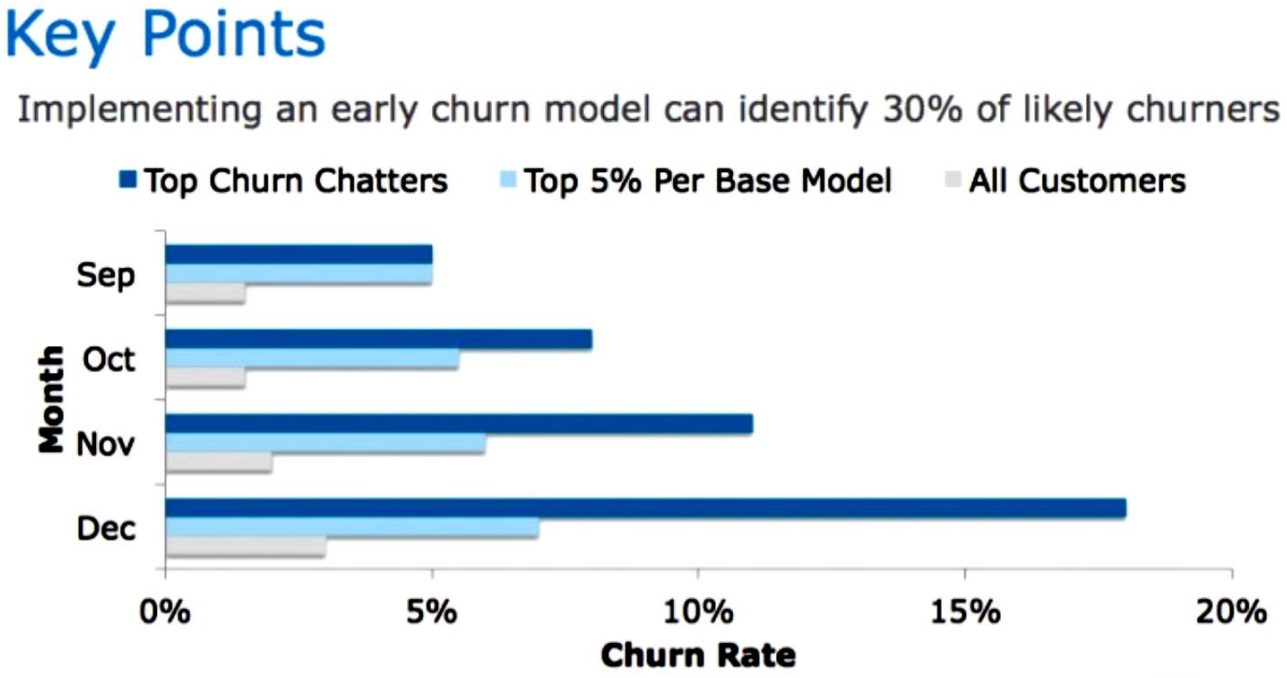
Creating the Final Deliverables

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

Creating the Final Deliverables

- **Key Points Supported with Data**
  - The next step is to identify key points based on insights and observations resulting from the data and model scoring results. Find ways to illustrate the key points with charts and visualization techniques, using simpler charts for sponsors and more technical data visualization for analysts and data scientists.



Creating the Final Deliverables

- **Model Details**
  - Model details are typically needed by people who have a more technical understanding than the sponsors, such as those who will implement the code, or colleagues on the analytics team. Project sponsors are typically less interested in the model details; they are usually more focused on the business implications of the work rather than the details of the model. This portion of the presentation needs to show the code or main logic of the model, including the model type, variables, and technology used to execute the model and score the data. The model details segment of the presentation should focus on describing expected model performance and any caveats related to the model performance. In addition, this portion of the presentation should provide a detailed description of the modeling technique, variables, scope, and expected effectiveness of the model.

Creating the Final Deliverables

- **Model Details**

Model Details

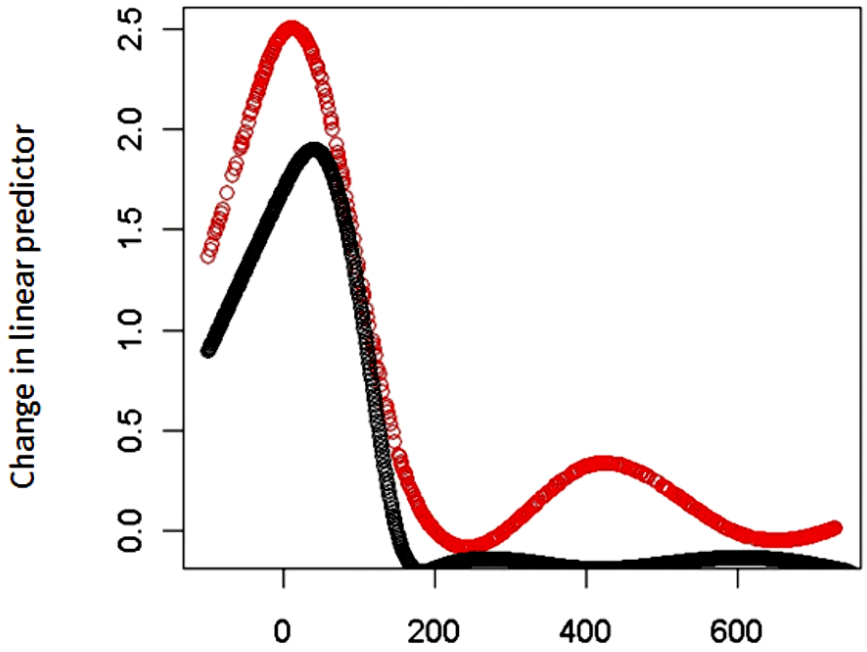
- Candidate variables: 22 from CRM, 154 from call history, and 12 social networking variables
- Through PCA and discussion with domain experts, we reduced ~190 variables to the 9 most predictive of customer churn
- General Additive Model (GAM) model built in R :

```
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 )
```

Creating the Final Deliverables

- **Model Details**

Var 1 has a larger and earlier impact on churn chatters



Creating the Final Deliverables

- **Recommendations**
  - The final main component of the presentation involves creating a set of recommendations that include how to deploy the model from a business perspective within the organization and any other suggestions on the rollout of the model's logic.

Creating the Final Deliverables

- **Recommendations**

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

Creating the Final Deliverables

- **Additional Tips on the Final Presentation**
  - Use imagery and visual representations.
  - Make sure the text is mutually exclusive and collectively exhaustive (MECE)
  - Measure and quantify the benefits of the project
  - Make the benefits of the project clear and conspicuous

Creating the Final Deliverables

- **Providing Technical Specifications and Code**
  - In addition to authoring the final presentations, the team needs to deliver the actual code that was developed and the technical documentation needed to support it.
  - The team should consider how the project will affect the end users and the technical people who will need to implement the code.
  - the team may need to consider a compromise of nightly batch jobs to process the data.

Data Visualization Basics

- As the volume of data continues to increase, more vendors and communities are developing tools to create clear and impactful graphics for use in presentations and applications. Although not exhaustive, Table lists some popular tools.

Open Source	Commercial Tools
R (Base package, `lattice`, `ggplot2`)	Tableau
GGobi/Rggobi	Spotfire (TIBCO)
Gnuplot	QlikView
Inkscape	Adobe Illustrator
Modest Maps	
OpenLayers	
Processing	
D3.js	
Weave	

Data Visualization Basics

- **Key Points Supported with Data**
  - Depending on the point trying to be made, the analyst must take care to organize the information in a way that intuitively enables the viewer to take away the same main point that the author intended. If the analyst fails to do this effectively, the person consuming the data must guess at the main point and may interpret something different from what was intended.

Data Visualization Basics

- **Key Points Supported with Data**

Year	1962	1963	1964	1965	1966	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	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	62	40	49	22	26	33	47	78	71	67	64	91	91	33	1980
BigBox						1		1	1	1	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	33	27	35	47	32	39	27	4	1196
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

FIGURE Forty-five years of store opening data

Data Visualization Basics

- **Key Points Supported with Data**
  - This map is a more powerful way to depict data than a small table would be. The approach is well suited to a sponsor audience.

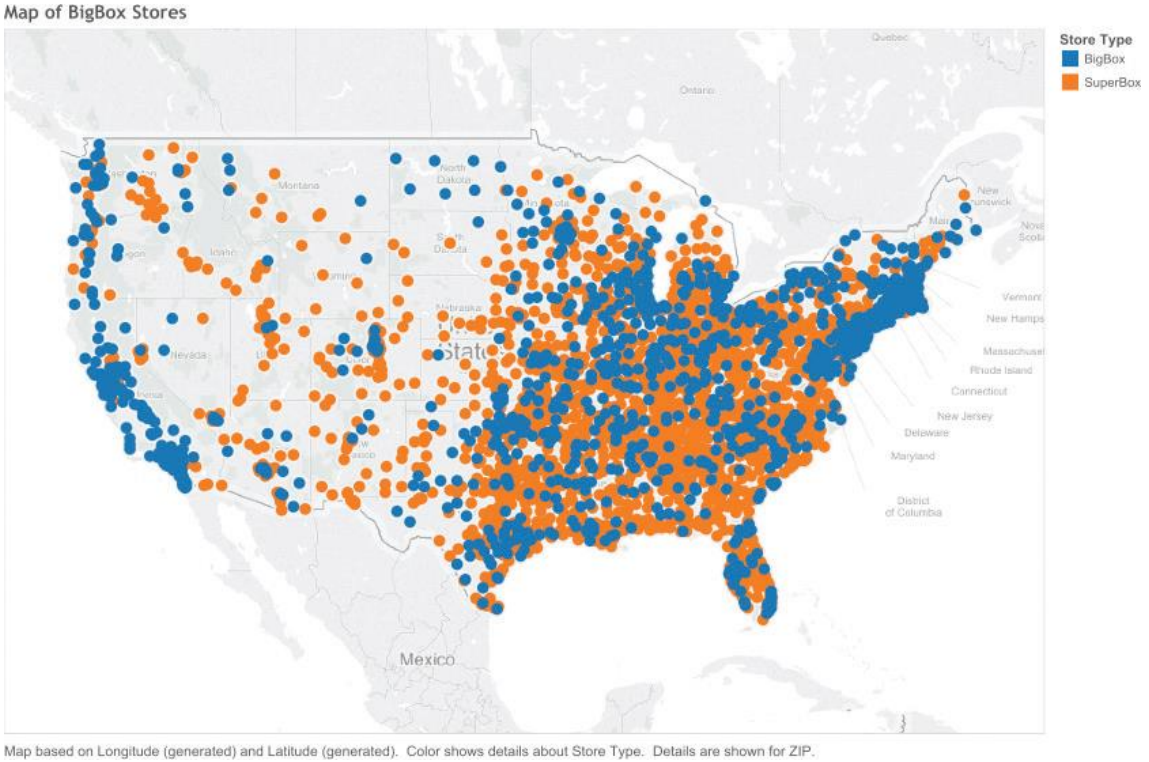


FIGURE Forty-five years of store opening data, shown as map

Data Visualization Basics

- **Evolution of a Graph**
  - Visualization allows people to portray data in a more compelling way than tables of data and in a way that can be understood on an intuitive, precognitive level. In addition, analysts and data scientists can use visualization to interact with and explore data.

Data Visualization Basics

- **Evolution of a Graph**

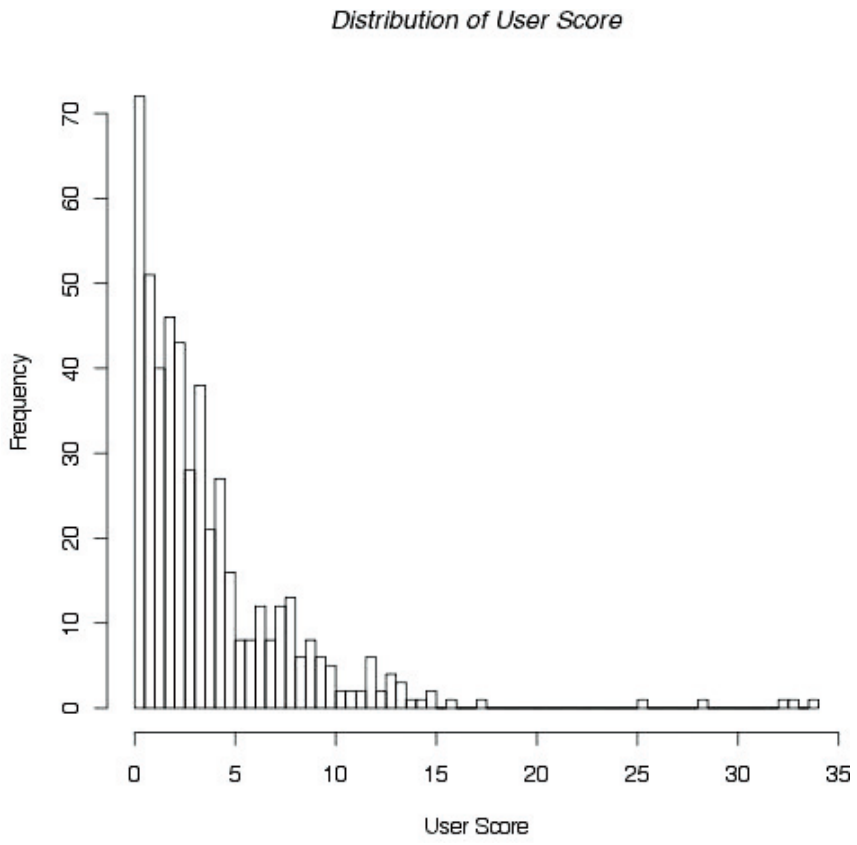


FIGURE Frequency distribution of user scores

Data Visualization Basics

- **Evolution of a Graph**

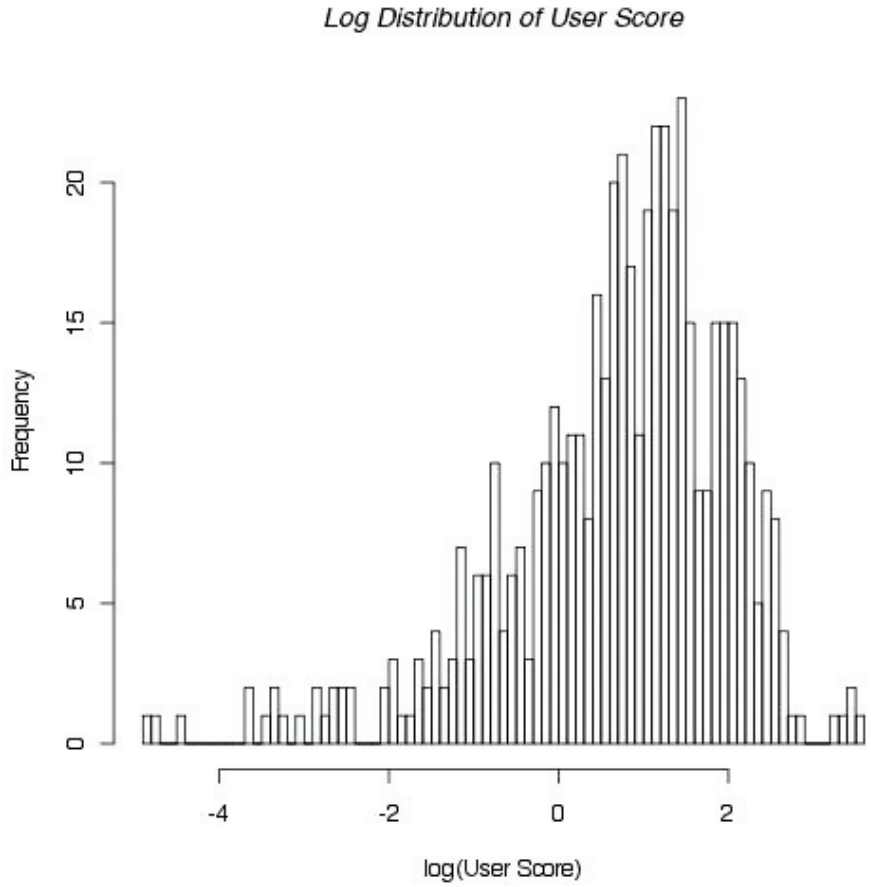


FIGURE Frequency distribution with log of user scores

Data Visualization Basics

- **Evolution of a Graph**

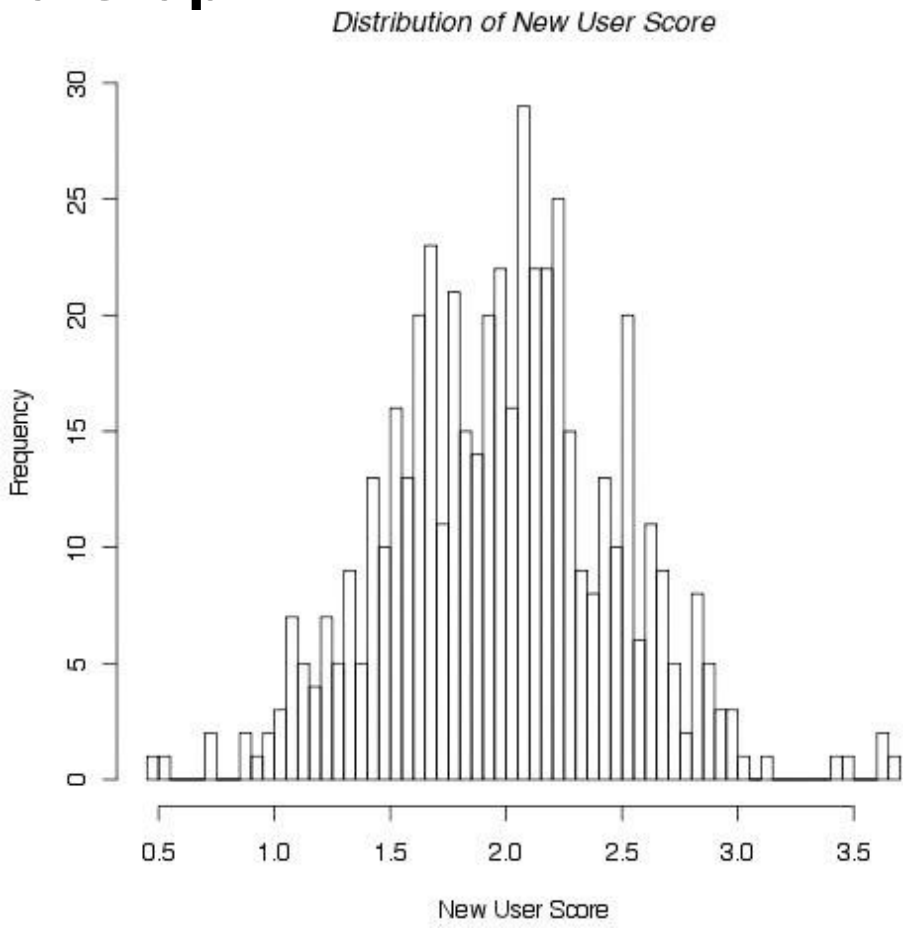


FIGURE Frequency distribution of new user scores

Data Visualization Basics

- **Evolution of a Graph**
  - Data scientists typically iterate and view data in many different ways, framing hypotheses, testing them, and exploring the implications of a given model. This case explores visual examples of pricing distributions, fluctuations in pricing, and the differences in price tiers before and after implementing a new model to optimize price. The visualization work illustrates how the data may look as the result of the model, and helps a data scientist understand the relationships within the data at a glance.

Data Visualization Basics

- **Evolution of a Graph**

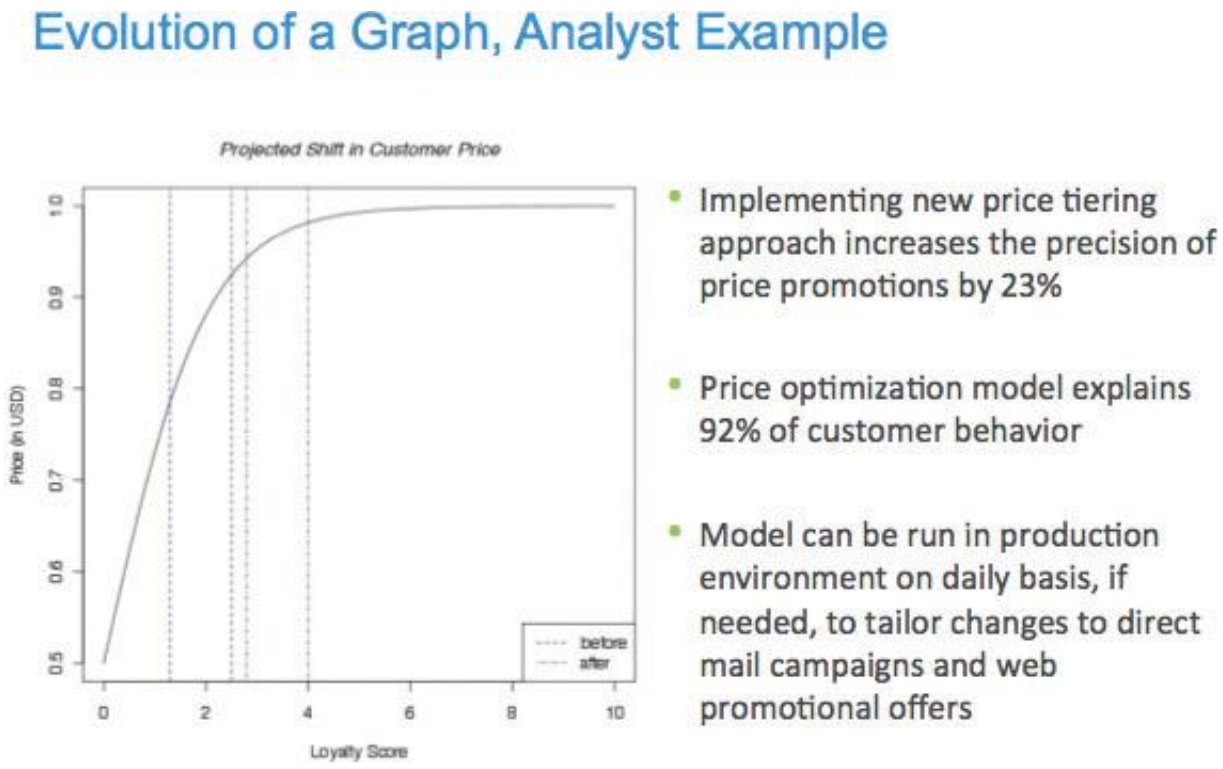


FIGURE Evolution of a graph, analyst example with supporting points

Data Visualization Basics

- **Evolution of a Graph**

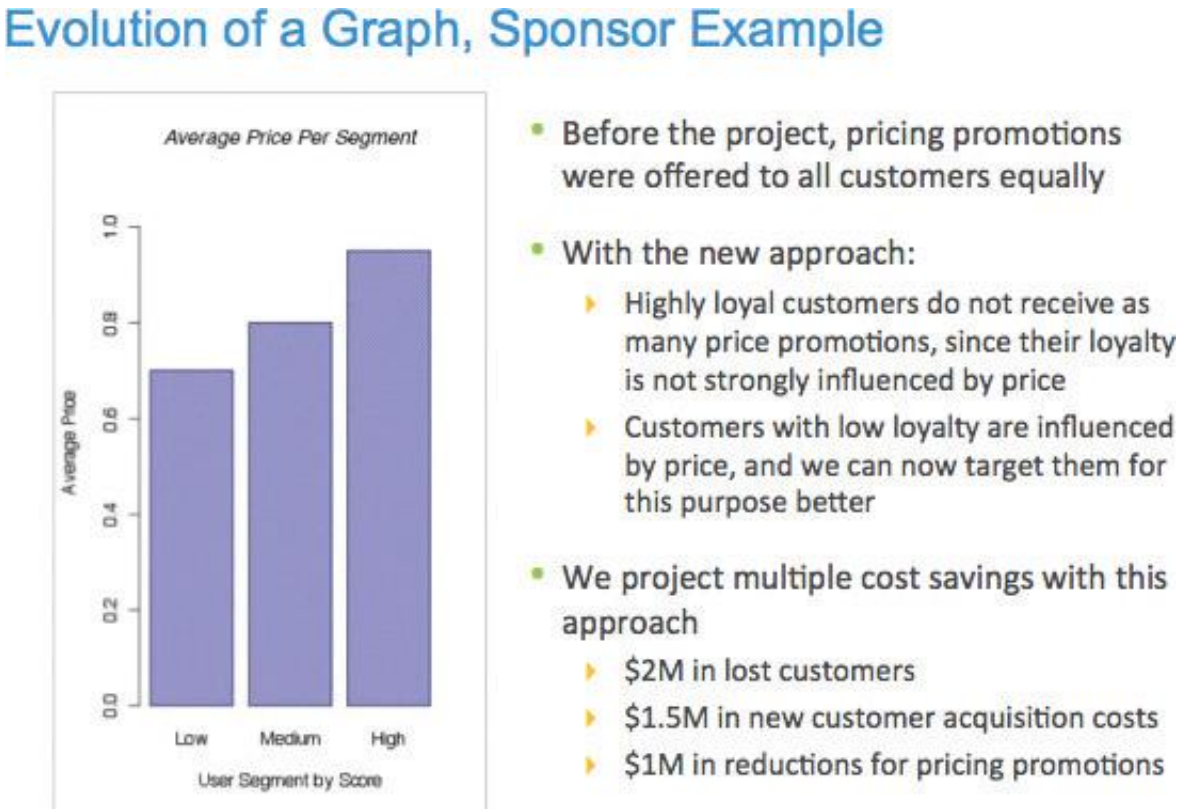


FIGURE Evolution of a graph, sponsor example

Data Visualization Basics

- **Common Representation Methods**
  - Although there are many types of data visualizations, several fundamental types of charts portray data and information. It is important to know when to use a particular type of chart or graph to express a given kind of data.

Data Visualization Basics

- **Common Representation Methods**

TABLE Common Representation Methods for Data and Charts

Data for Visualization	Type of Chart
Components (parts of whole)	Pie chart
Item	Bar chart
Time series	Line chart
Frequency	Line chart or histogram
Correlation	Scatterplot, side-by-side bar charts

Data Visualization Basics

- **How to Clean Up a Graphic**
  - Many times software packages generate a graphic for a dataset, but the software adds too many things to the graphic. These added visual distractions can make the visual appear busy or otherwise obscure the main points that are to be made with the graphic. In general, it is a best practice to strive for simplicity when creating graphics and data visualization graphs. Knowing how to simplify graphics or clean up a messy chart is helpful for conveying the key message as clearly as possible.

Data Visualization Basics

- **How to Clean Up a Graphic**

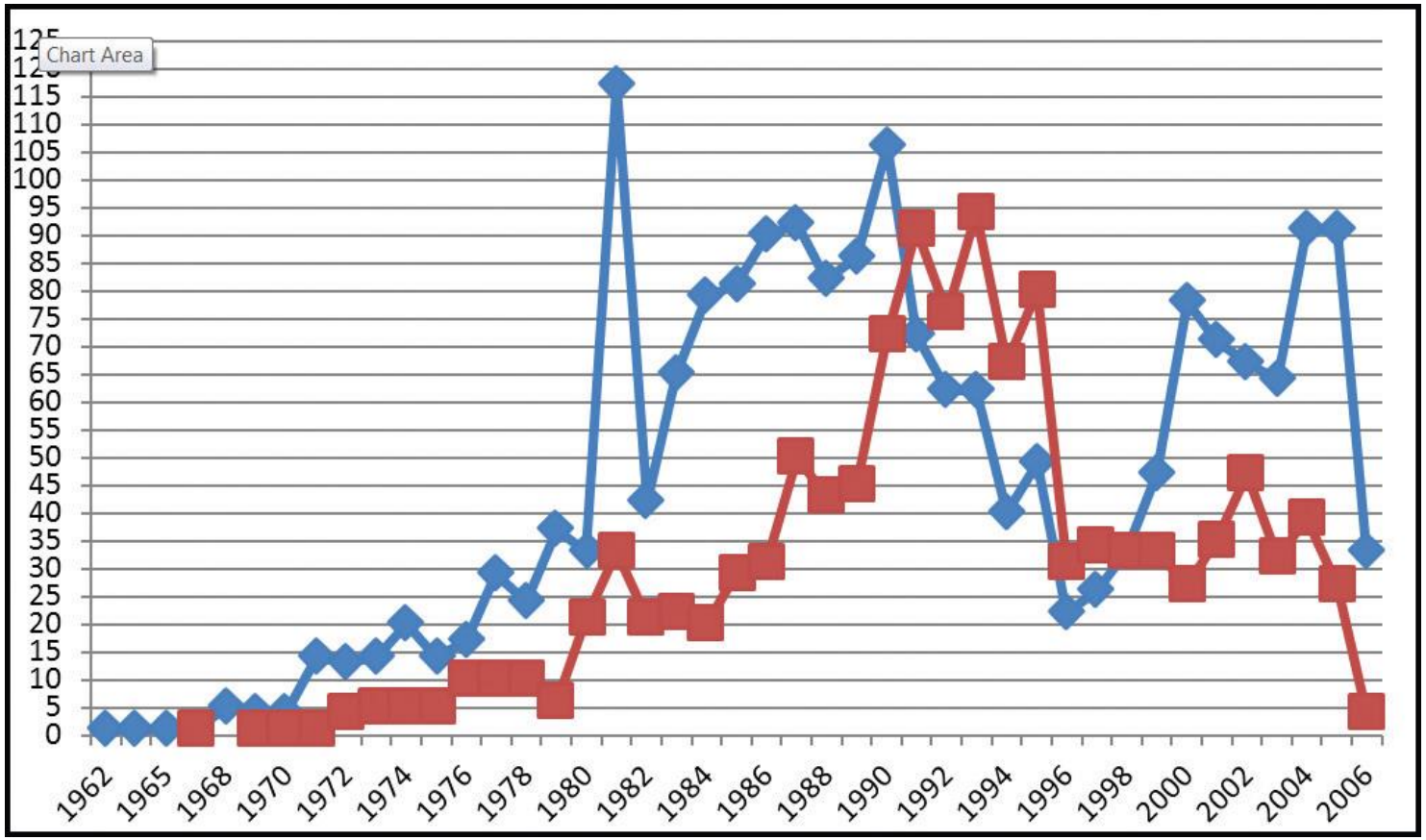


FIGURE How to clean up a graphic, example 1 (before)

Data Visualization Basics

- **How to Clean Up a Graphic**

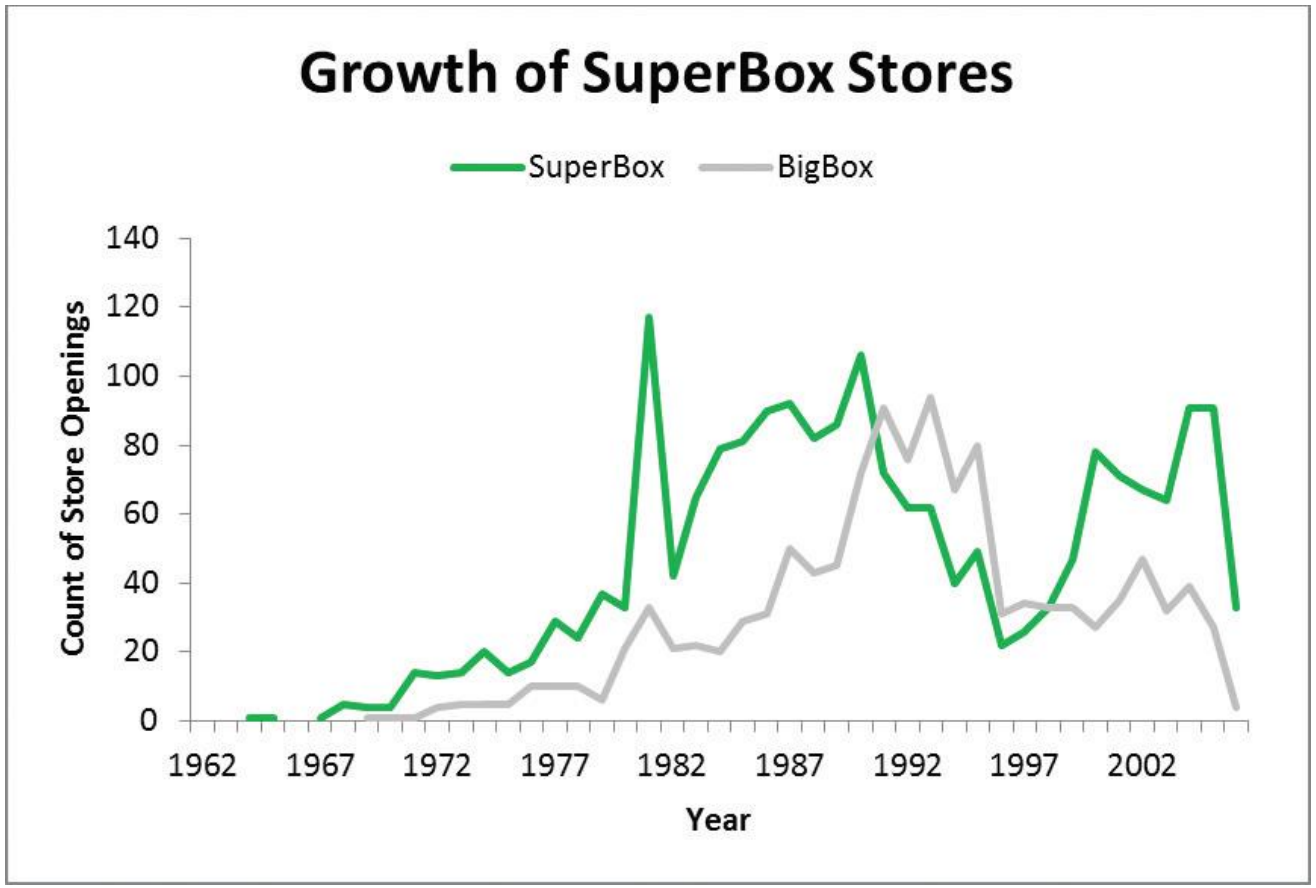


FIGURE How to clean up a graphic, example 1 (after)

Data Visualization Basics

- **How to Clean Up a Graphic**

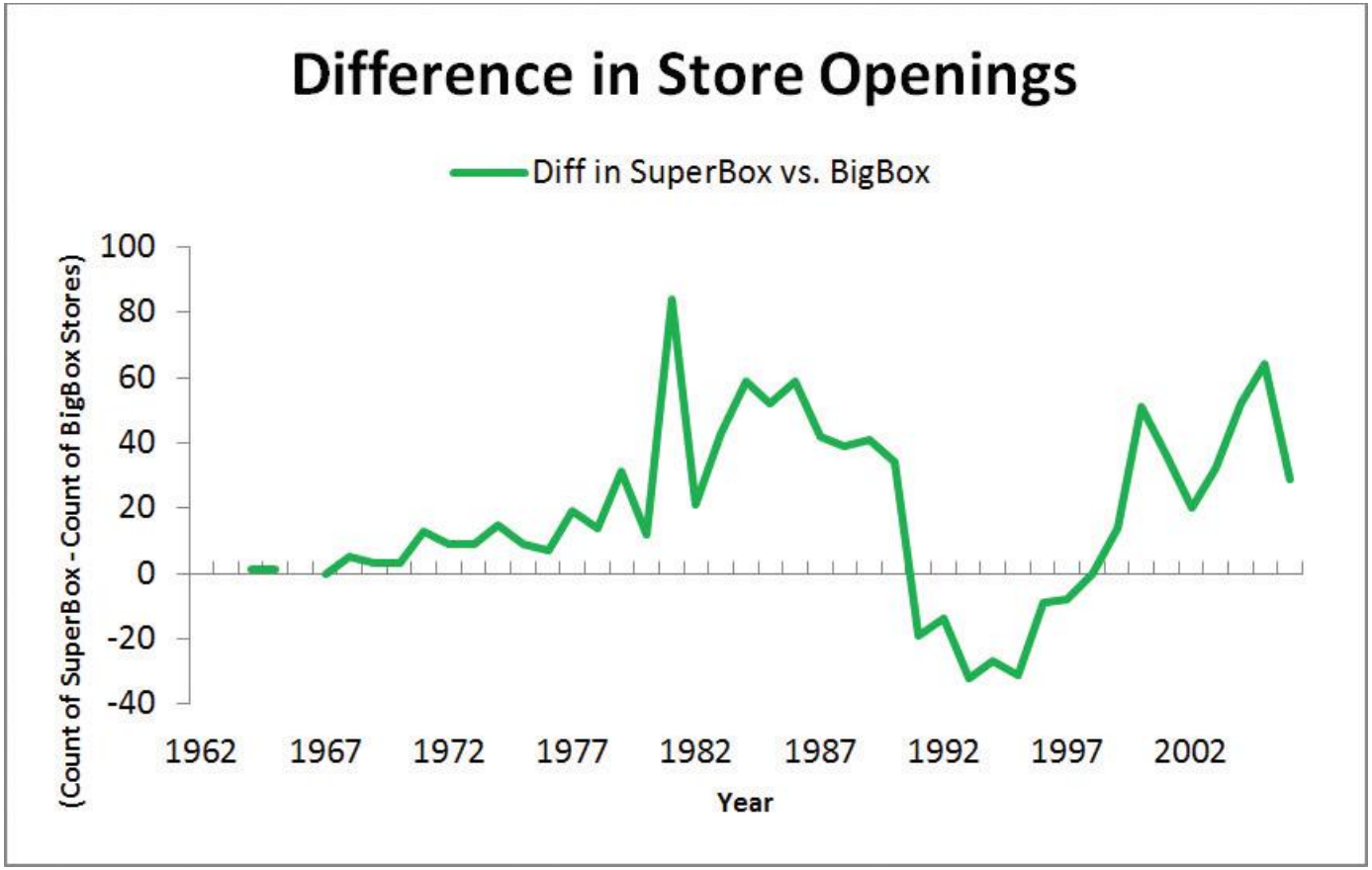


FIGURE How to clean up a graphic, example 1 (alternate “after” view)

Data Visualization Basics

- **How to Clean Up a Graphic**

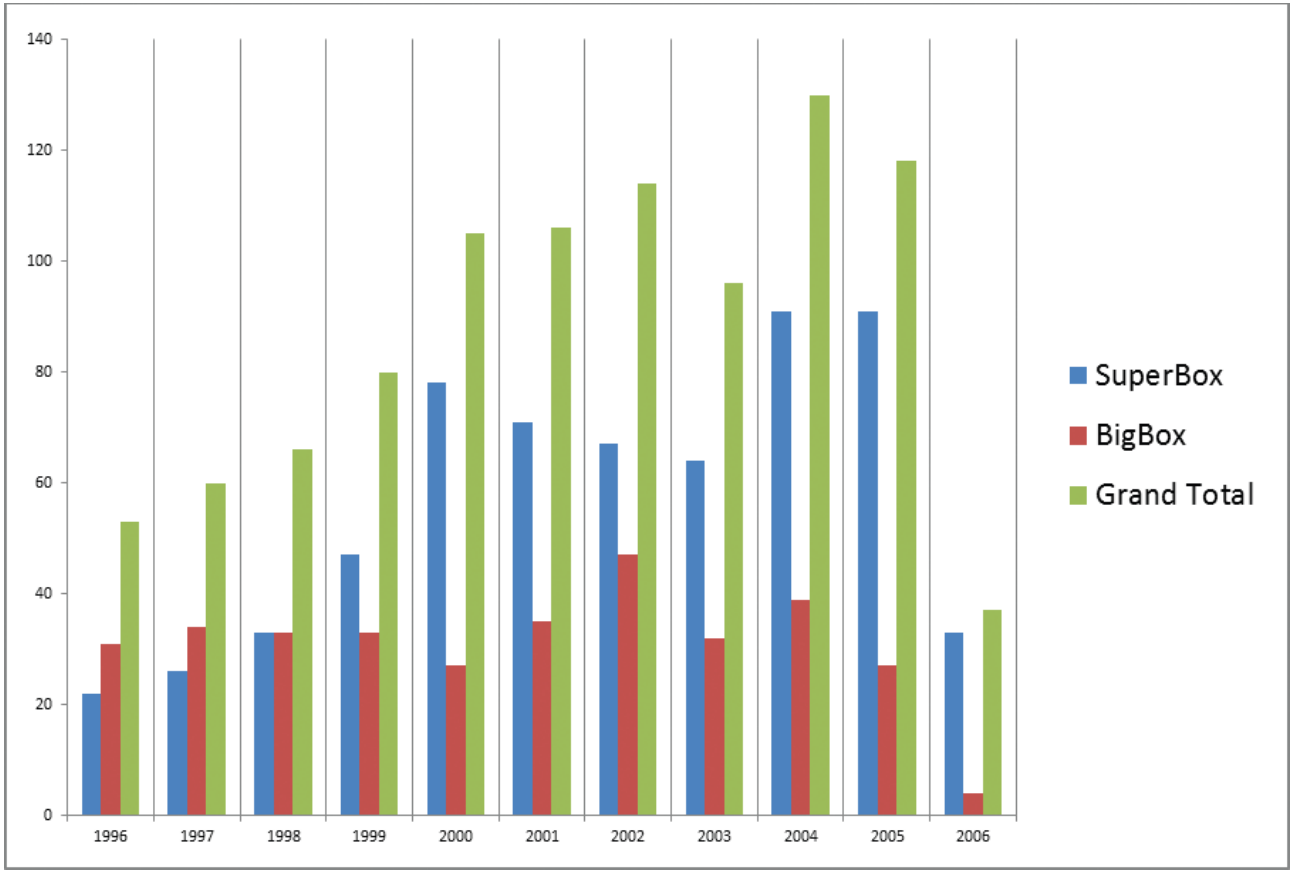


FIGURE How to clean up a graphic, example 2 (before)

Data Visualization Basics

- **How to Clean Up a Graphic**

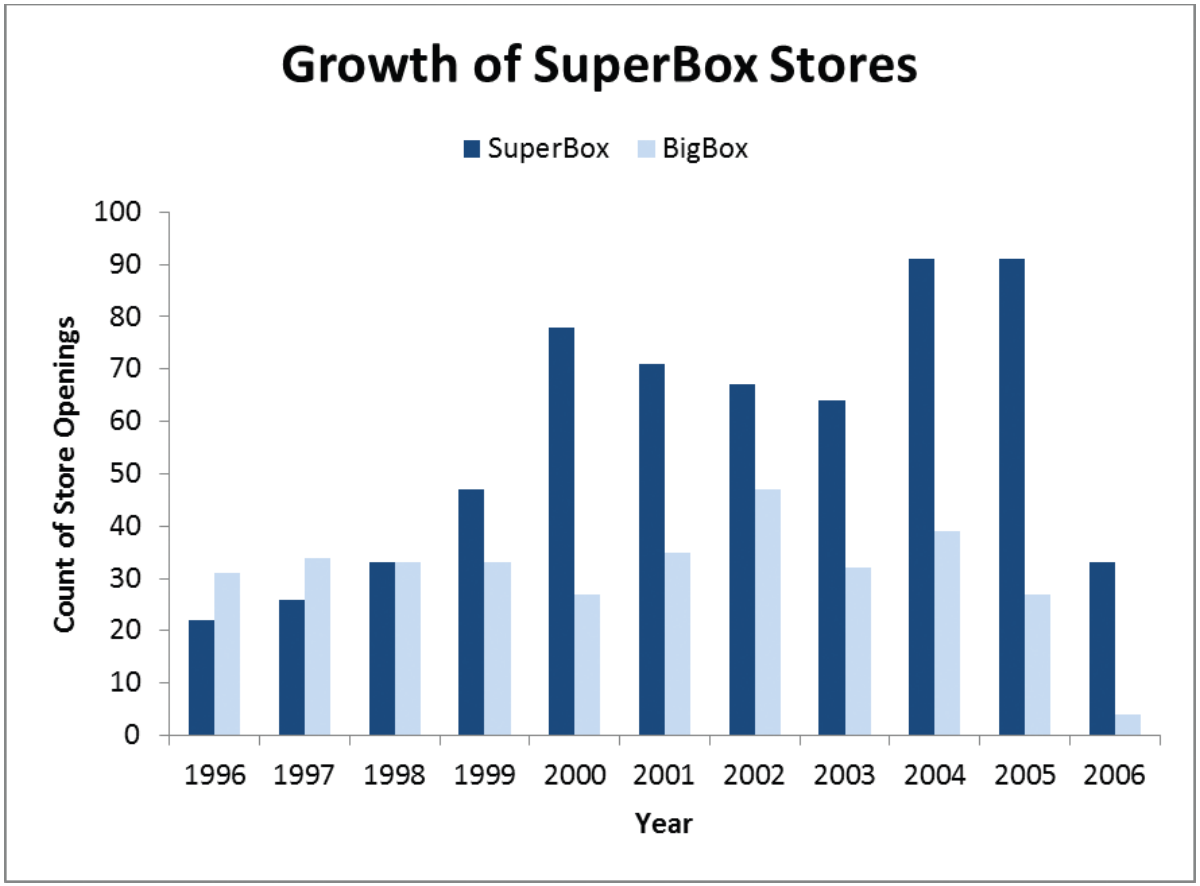


FIGURE How to clean up a graphic, example 2 (after)

Data Visualization Basics

- **How to Clean Up a Graphic**

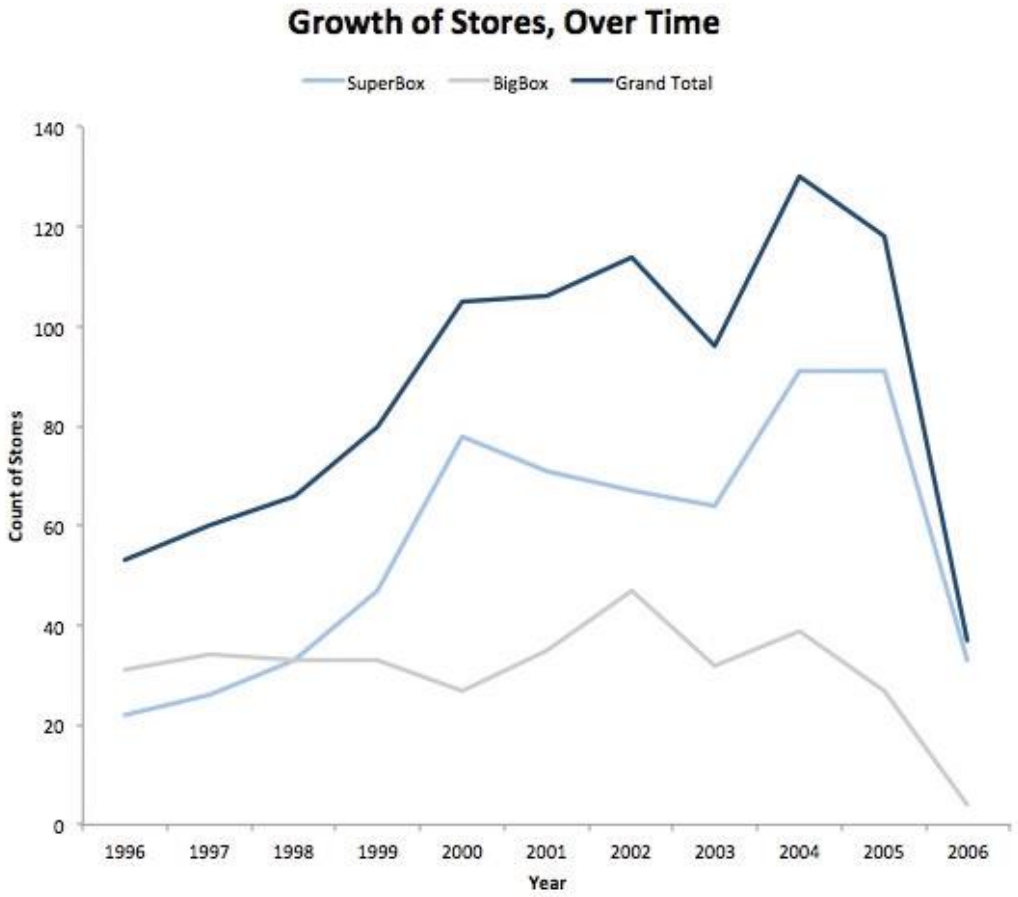


FIGURE How to clean up a graphic, example 2 (alternate “after” view)

Data Visualization Basics

- **Additional Considerations**
  - Avoid Using Three-Dimensions in Most Graphics

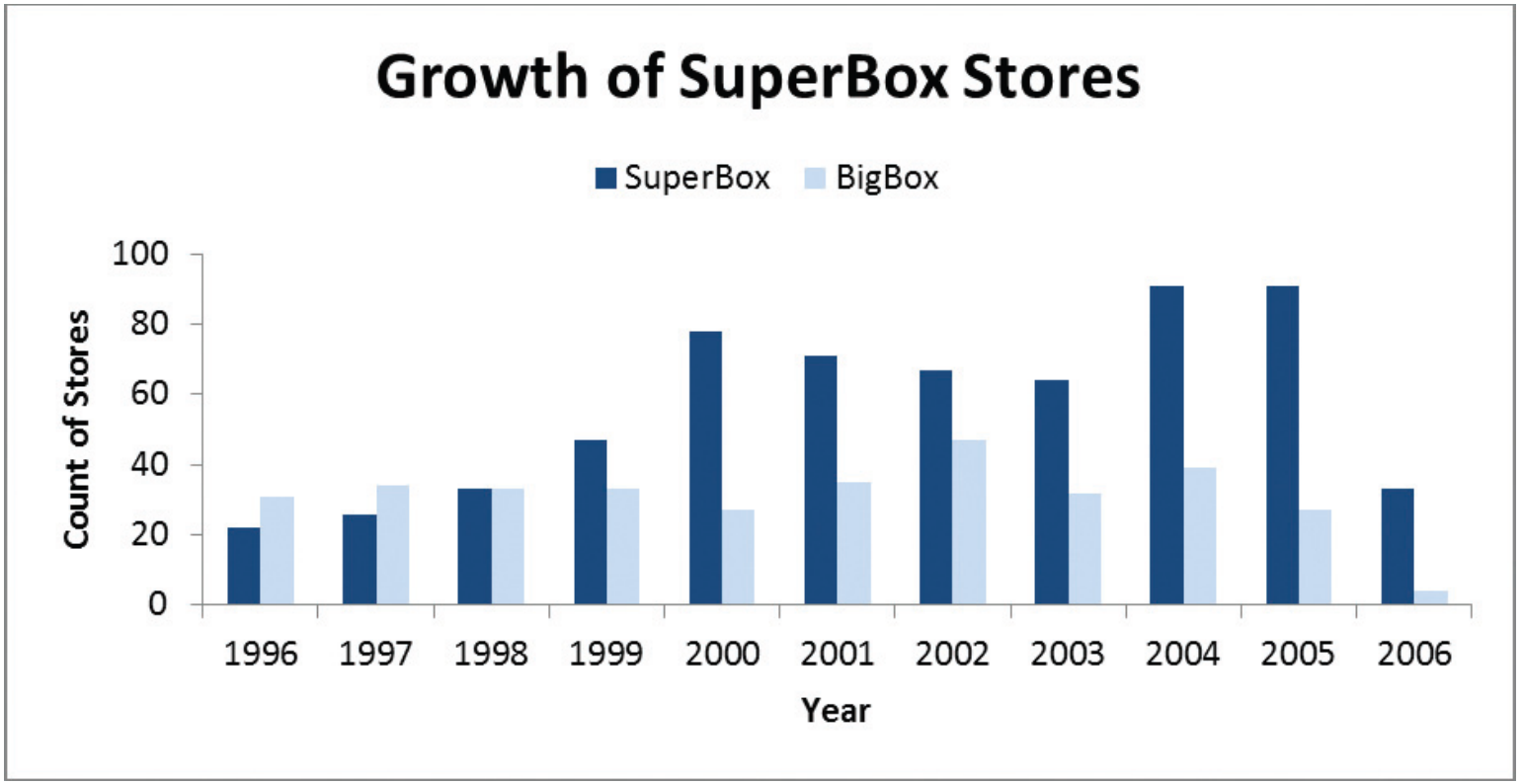


FIGURE Simple bar chart, with two dimensions

Data Visualization Basics

- **Additional Considerations**
  - Avoid Using Three-Dimensions in Most Graphics

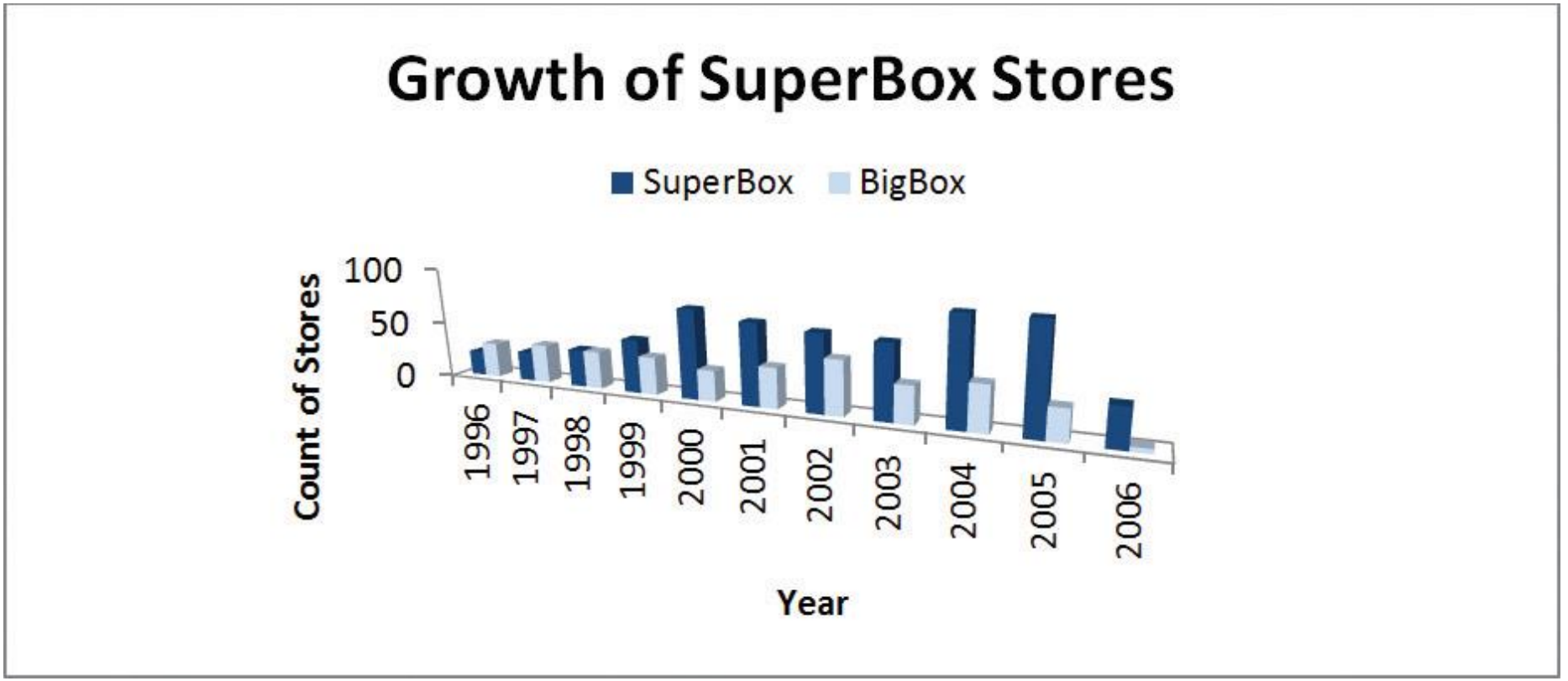


FIGURE Misleading bar chart, with three dimensions

Summary

- Communicating and Operationalizing an Analytics Project
- Creating the Final Deliverables
- Data Visualization Basics

Q&A

Images Courtesy of Google Image