12

The Endgame, or Putting It All Together

Key Concepts

Communicating and operationalizing an analytics project
Creating the final deliverables
Using a core set of material for different audiences
Comparing main focus areas for sponsors and analysts
Understanding simple data visualization principles
Cleaning up a chart or visualization

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.

12.1 Communicating and Operationalizing an Analytics Project

As shown in Figure 12-1, 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.

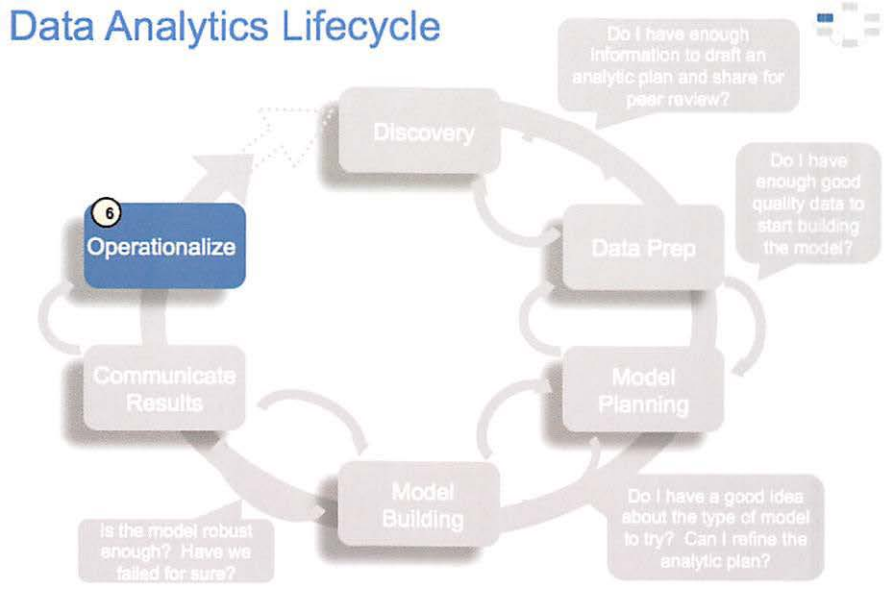

FIGURE 12-1 *Data Analytics Lifecycle, Phase 6: operationalize*

The team's ability to quantify the benefits and share them in a compelling way with the stakeholders will determine if the work will move forward into a pilot project and ultimately be run in a production environment. Therefore, it is critical to identify the benefits and state them in a clear way in the final presentations.

As the team scopes the effort involved to deploy the analytical model as a pilot project, it also needs to consider running the model in a production environment for a discrete set of products or a single line of business, which tests the model in a live setting. This allows the team to learn from the deployment and make adjustments before deploying the application or code more broadly across the enterprise. This phase can bring in a new set of team members—namely, those engineers responsible for the production environment who have a new set of issues and concerns. This group is interested in ensuring that running the model fits smoothly into the production environment and the model can be integrated into downstream processes. While executing the model in the production environment, the team should aim to detect input anomalies before they are fed to the model, assess run times, and gauge competition for resources with other processes in the production environment.

Chapter 2 included an in-depth discussion of the Data Analytics Lifecycle, including an overview of the deliverables provided in its final phase, at which time it is advisable for the team to consider the needs of each of its main stakeholders and the deliverables, illustrated in Figure 12-2, to satisfy these needs.

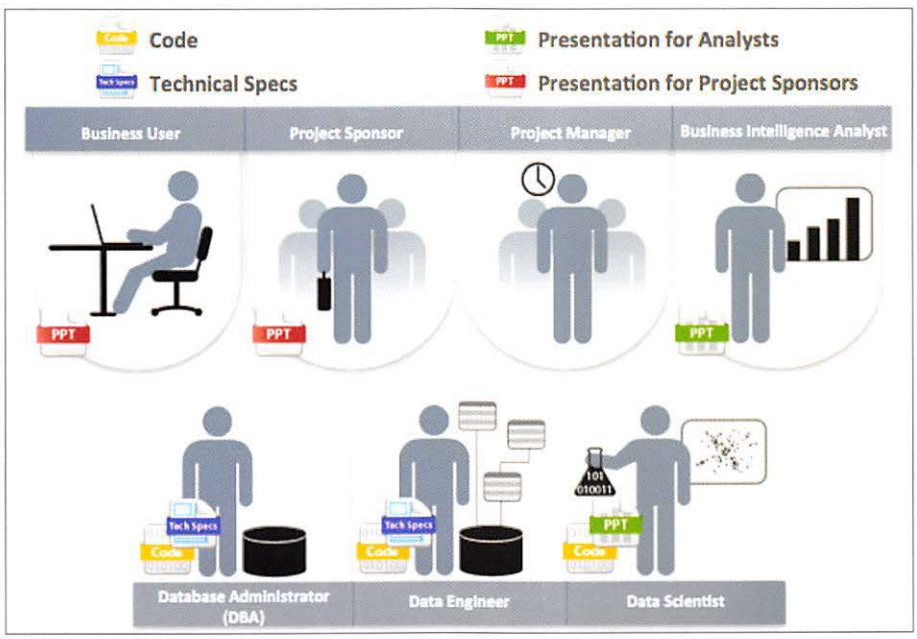

FIGURE 12-2 *Key outputs from a successful analytic project*

Following is a brief review of the key outputs for each of the main stakeholders of an analytics project and what they usually expect at the conclusion of a 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.

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

As a rule, the more executive the audience, the more succinct the presentation needs to be for project sponsors. Ensure that the presentation gets to the point quickly and frames the results in terms of value to the sponsor's organization. When presenting to other audiences with more quantitative backgrounds, focus more time on the methodology and findings. In these instances, the team can be more expansive in describing the outcomes, methodology, and analytical experiments with a peer group. This audience will be more interested in the techniques, especially if the team developed a new way of processing or analyzing data that can be reused in the future or applied to similar problems. In addition, use imagery or data visualization when possible. Although it may take more time to develop imagery, pictures are more appealing, easier to remember, and more effective to deliver key messages than long lists of bullets.

12.2 Creating the Final Deliverables

After reviewing the list of key stakeholders for data science projects and main deliverables, this section focuses on describing the deliverables in detail. To illustrate this approach, a fictional case study is used to make the examples more specific. Figure 12-3 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 a competing bank.

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.

FIGURE 12-3 *Synopsis of YoyoDyne Bank case study example*

Based on this information, the data science team may create an analytics plan similar to Figure 12-4 during the project.

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.

FIGURE 12-4 *Analytics plan for YoyoDyne Bank case study*

In addition to guiding the model planning and methodology, the analytic plan contains components that can be used as inputs for writing about the scope, underlying assumptions, modeling techniques, initial hypotheses, and key findings in the final presentations. After spending substantial amounts of time in the modeling and performing in-depth data analysis, it is critical to reflect on the project work and consider

the context of the problems the team set out to solve. Review the work that was completed during the project, and identify observations about the model outputs, scoring, and results. Based on these observations, begin to identify the key messages and any unexpected insights.

In addition, it is important to tailor the project outputs to the audience. For a project sponsor, show that the team met the project goals. Focus on what was done, what the team accomplished, what ROI can be anticipated, and what business value can be realized. Give the project sponsor talking points to evangelize the work. Remember that the sponsor needs to relay the story to others, so make this person's job easy, and help ensure the message is accurate by providing a few talking points. Find ways to emphasize ROI and business value, and mention whether the models can be deployed within performance constraints of the sponsor's production environment.

In some organizations, the data science team may not be expected to make a full business case for future projects and implementation of the models. Instead, it needs to be able to provide guidance about the impact of the models to enable the project sponsor, or someone designated by that person, to create a business case to advocate for the pilot and subsequent deployment of this functionality. In other words, the data science team can assist in this effort by putting the results of the modeling and data science work into context to help assess the actual value and cost of implementing this work more broadly.

When presenting to a technical audience such as data scientists and analysts, focus on how the work was done. Discuss how the team accomplished the goals and the choices it made in selecting models or analyzing the data. Share analytical methods and decision-making processes so other analysts can learn from them for future projects. Describe methods, techniques, and technologies used, as this technical audience will be interested in learning about these details and considering whether the approach makes sense in this case and whether it can be extended to other, similar projects. Plan to provide specifics related to model accuracy and speed, such as how well the model will perform in a production environment.

Ideally, the team should consider starting the development of the final presentation during the project rather than at the end of the project as commonly occurs. This approach ensures that the team always has a version of the presentation with working hypotheses to show stakeholders, in case there is a need to show a work-in-process version of the project progress on short notice. In fact, many analysts write the executive summary at the outset of a project and then continually refine it over time so that at the end of the project, portions of the final presentation are already completed. This approach also reduces the chance that the team members will forget key points or insights discovered during the project. Finally, it reduces the amount of work to be done on the presentation at the conclusion of the project.

12.2.1 Developing Core Material for Multiple Audiences

Because some of the components of the projects can be used for different audiences, it can be helpful to create a core set of materials regarding the project, which can be used to create presentations for either a technical audience or an executive sponsor.

Table 12-1 depicts the main components of the final presentations for the project sponsor and an analyst audience. Notice that teams can create a core set of materials in these seven areas, which can be used for the two presentation audiences. Three areas (Project Goals, Main Findings, and Model Description), can be used as is for both presentations. Other areas need additional elaboration, such as the Approach. Still other areas, such as the Key Points, require different levels of detail for the analysts and data scientists than for the project sponsor. Each of these main components of the final presentation is discussed in subsequent sections.

TABLE 12-1 *Comparison of Materials for Sponsor and Analyst Presentations*

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.

12.2.2 Project Goals

The Project Goals portion of the final presentation is generally the same, or similar, for sponsors and for analysts. For each audience, the team needs to reiterate the goals of the project to lay the groundwork for

the solution and recommendations that are shared later in the presentation. In addition, the Goals slide serves to ensure there is a shared understanding between the project team and the sponsors and confirm they are aligned in moving forward in the project. Generally, the goals are agreed on early in the project. It is good practice to write them down and share them to ensure the goals and objectives are clearly understood by both the project team and the sponsors.

Figures 12-5 and 12-6 show two examples of slides for Project Goals. Figure 12-5 shows three goals for creating a predictive model to anticipate customer churn. The points on this version of the Goals slide emphasize what needs to be done, but not why, which will be included in the alternative.

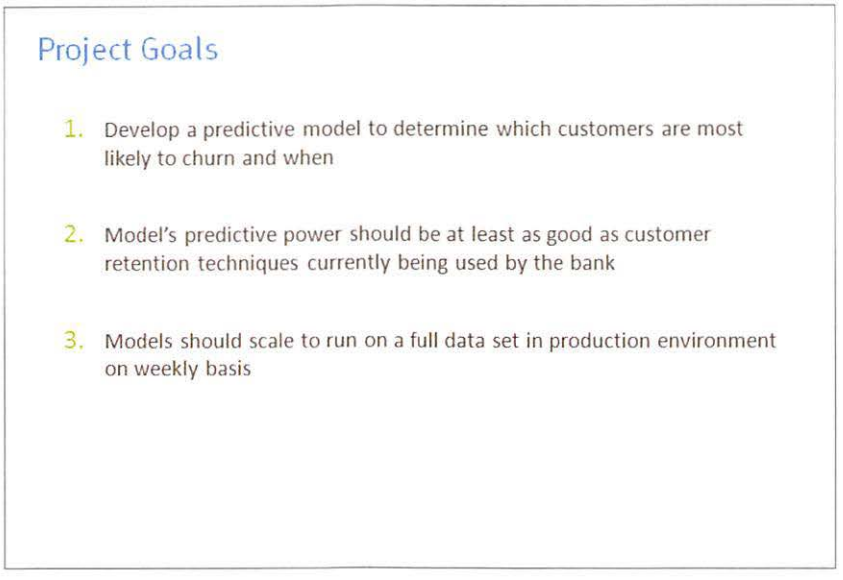

FIGURE 12-5 *Example of Project Goals slide for YoyoDyne case study*

Figure 12-6 shows a variation of the previous Project Goals slide in Figure 12-5. It is a summary of the situation prior to listing the goals. Keep in mind that when delivering final presentations, these deliverables are shared within organizations, and the original context can be lost, especially if the original sponsor leaves the group or changes roles. It is good practice to briefly recap the situation prior to showing the project goals. Keep in mind that adding a situation overview to the Goals slide does make it appear busier. The team needs to determine whether to split this into a separate slide or keep it together, depending on the audience and the team's style for delivering the final presentation.

One method for writing the situational overview in a succinct way is to summarize it in three bullets, as follows:

- **Situation:** Give a one-sentence overview of the situation that has led to the analytics project.
- **Complication:** Give a one-sentence overview of the need for addressing this now. Something has triggered the organization to decide to take action at this time. For instance, perhaps it lost 100

customers in the past two weeks and now has an executive mandate to address an issue, or perhaps it has lost five points of market share to its biggest competitor in the past three months. Usually, this sentence represents the driver for why a particular project is being initiated at this time, rather than in some vague time in the future.

- **Implication:** Give a one-sentence overview of the impact of the complication. For instance, if the bank fails to address its customer attrition problem, it stands to lose its dominant market position in three key markets. Focus on the business impact to illustrate the urgency of doing the project.

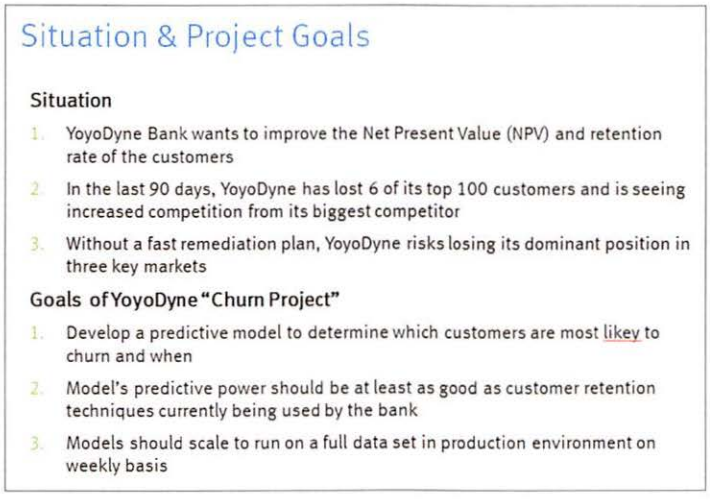

FIGURE 12-6 *Example of Situation & Project Goals slide for YoyoDyne case study*

12.2.3 Main Findings

Write a solid executive summary to portray the main findings of a project. In many cases, the summary may be the only portion of the presentation that hurried managers will read. For this reason, it is imperative to make the language clear, concise, and complete. Those reading the executive summary should be able to grasp the full story of the project and the key insights in a single slide. In addition, this is an opportunity to provide key talking points for the executive sponsor to use to evangelize the project work with others in the customer's organization. Be sure to frame the outcomes of the project in terms of both quantitative and qualitative business value. This is especially important if the presentation is for the project sponsor. The executive summary slide containing the main findings is generally the same for both sponsor and analyst audiences.

Figure 12-7 shows an example of an executive summary slide for the YoyoDyne case study. It is useful to take a closer look at the parts of the slide to make sure it is clear. Keep in mind this is not the only format for conveying the Executive Summary; it varies based on the author's style, although many of the key components are common themes in Executive Summaries.

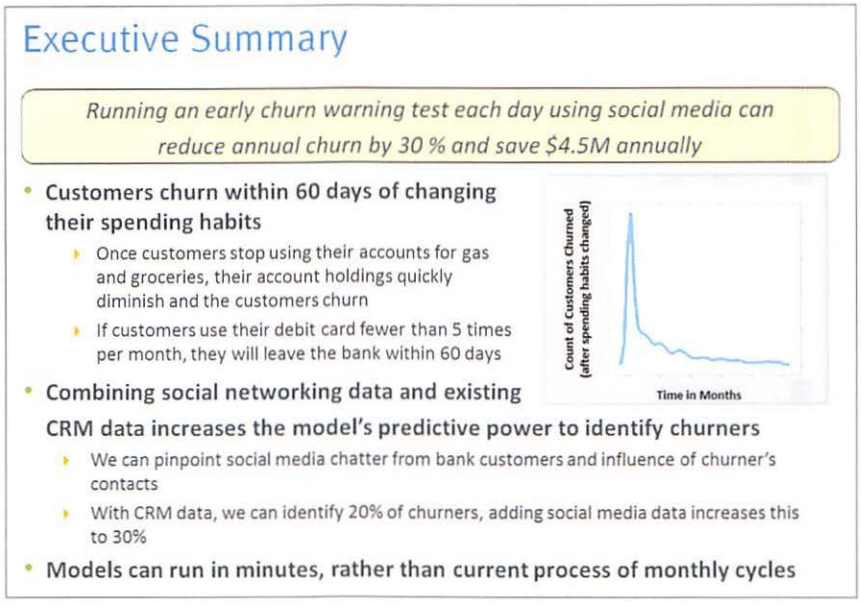

FIGURE 12-7 *Example of Executive Summary slide for YoyoDyne case study*

The key message should be clear and conspicuous at the front of the slide. It can be set apart with color or shading, as shown in Figure 12-8; other techniques can also be used to draw attention to it. The key message may become the single talking point that executives or the project sponsor take away from the project and use to support the team's recommendation for a pilot project, so it needs to be succinct and compelling. To make this message as strong as possible, measure the value of the work and quantify the cost savings, revenue, time savings, or other benefits to make the business impact concrete.

Follow the key message with three major supporting points. Although Executive Summary slides can have more than three major points, going beyond three ideas makes it difficult for people to recall the main points, so it is important to ensure that the ideas remain clear and limited to the few most impactful ideas the team wants the audience to take away from the work that was done. If the author lists ten key points, messages become diluted, and the audience may remember only one or two main points.

In addition, because this is an analytics project, be sure to make one of the key points related to if, and how well, the work will meet the sponsor's service level agreement (SLA) or expectations. Traditionally, the SLA refers to an arrangement between someone providing services, such as an information technology (IT) department or a consulting firm, and an end user or customer. In this case, the SLA refers to system performance, expected uptime of a system, and other constraints that govern an agreement. This term has become less formal and many times conveys system performance or expectations more generally related to performance or timeliness. It is in this sense that SLA is being used here. Namely, in this context, SLA

refers to the expected performance of a system and the intent that the models developed will not adversely impact the expected performance of the system into which they are integrated.

Finally, although it's not required, it is often a good idea to support the main points with a visual or graph. Visual imagery serves to make a visceral connection and helps retain the main message with the reader.

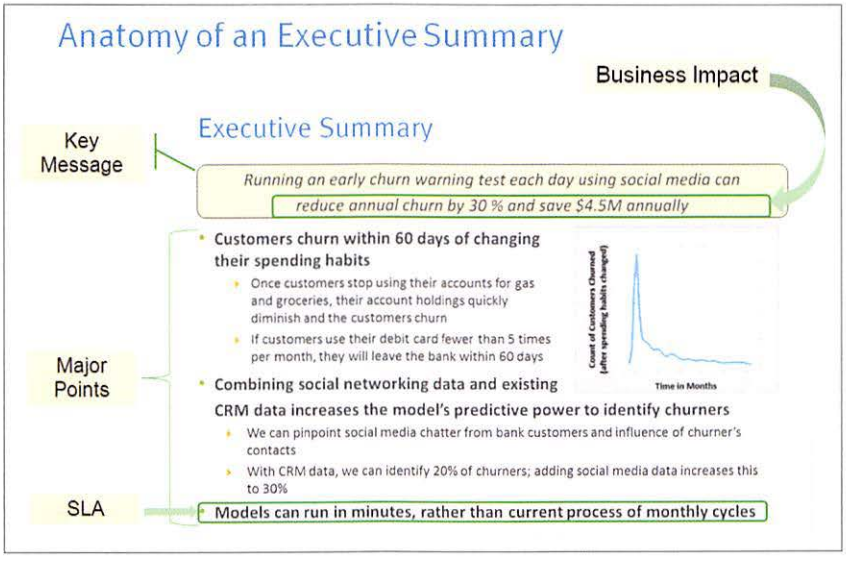

FIGURE 12-8 *Anatomy of an Executive Summary slide*

12.2.4 Approach

In the Approach portion of the presentation, the team needs to explain the methodology pursued on the project. This can include interviews with domain experts, the groups collaborating within the organization, and a few statements about the solution developed. The objective of this slide is to ensure the audience understands the course of action that was pursued well enough to explain it to others within the organization. The team should also include any additional comments related to working assumptions the team followed as it performed the work, because this can be critical in defending why they followed a specific course of action.

When explaining the solution, the discussion should remain at a high level for the project sponsors. If presenting to analysts or data scientists, provide additional detail about the type of model used, including the technology and the actual performance of the model during the tests. Finally, as part of the description of the approach, the team may want to mention constraints from systems, tools, or existing processes and any implications for how these things may need to change with this project.

Figure 12-9 shows an example of how to describe the methodology followed during a data science project to a sponsor audience.

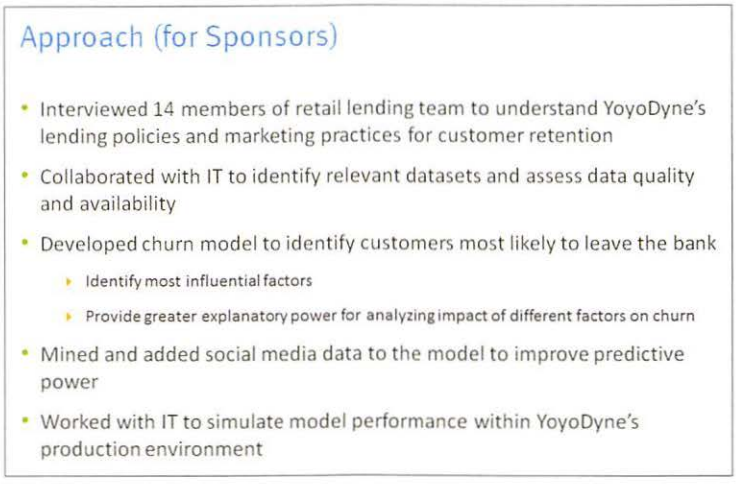

FIGURE 12-9 *Example describing the project methodology for project sponsors*

Note that the third bullet describes the churn model in general terms. Furthermore, the subbullets provide additional details in nontechnical terms. Compare this approach to the variation shown in Figure 12-10.

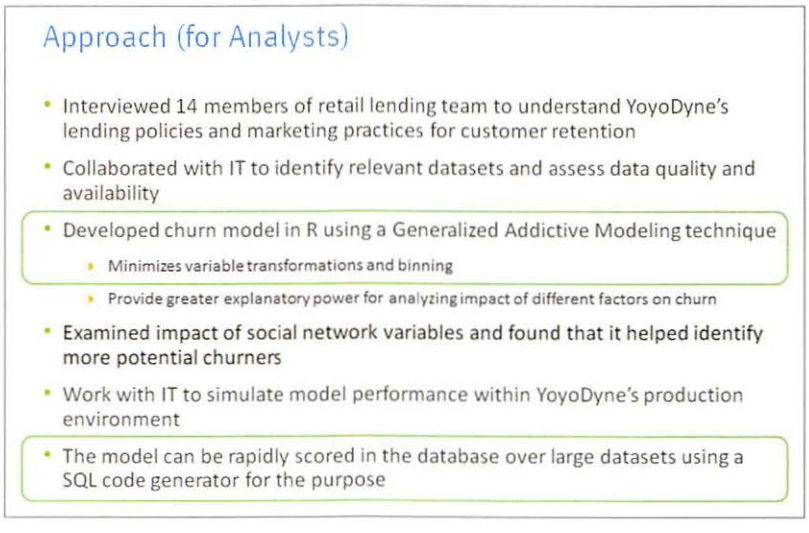

FIGURE 12-10 *Example describing the project methodology for analysts and data scientists*

Figure 12-10 shows a variation on the approach and methodology used in the data science project. In this case, most of the language and description are the same as in the example for project sponsors.

The main difference is that this version contains additional detail regarding the kind of model used and the way the model will score data quickly to meet the SLA. These differences are highlighted in the boxes shown in Figure 12-10.

12.2.5 Model Description

After describing the project approach, teams generally include a description of the model that was used. Figure 12-11 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. Convey the basic methodology followed in the team's work to allow the sponsor to communicate this to others within the organization and provide talking points.

Mentioning the scope of the data used is critical. The purpose is to illustrate thoroughness and exude confidence that the team used an approach that accurately portrays its problem and is as free from bias as possible. A key trait of a good data scientist is the ability to be skeptical of one's own work. This is an opportunity to view the work and the deliverable critically and consider how the audience will receive the work. Try to ensure it is an unbiased view of the project and the results.

Assuming that the model will meet the agreed-upon SLAs, mention that the model will meet the SLAs based on the performance of the model within the testing or staging environment. For instance, one may want to indicate that the model processed 500,000 records in 5 minutes to give stakeholders an idea of the speed of the model during run time. Analysts will want to understand the details of the model, including the decisions made in constructing the model and the scope of the data extracts for testing and training. Be prepared to explain the team's thought process on this, as well as the speed of running the model within the test environment.

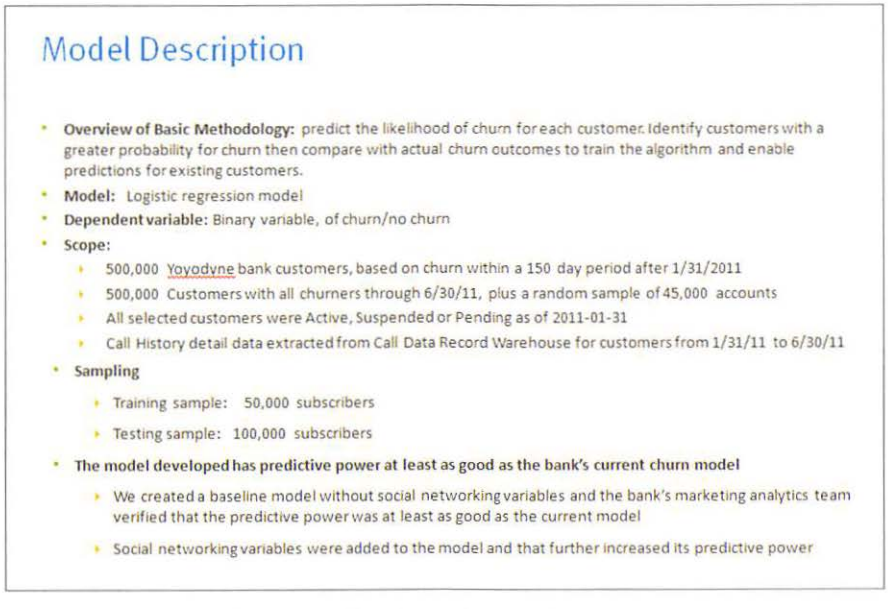

FIGURE 12-11 *Example of a model description for a data science project*

12.2.6 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.

Figure 12-12 shows an example of providing supporting detail regarding the rate of bank customers who would churn in various months. When developing the key points, consider the insights that will drive the biggest business impact and can be defended with data. For project sponsors, use simple charts such as bar charts, which illustrate data clearly and enable the audience to understand the value of the insights. This is also a good point to foreshadow some of the team's recommendations and begin tying together ideas to demonstrate what led to the recommendations and why. In other words, this section supplies the data and foundation for the recommendations that come later in the presentation. Creating clear, compelling slides to show the key points makes the recommendations more credible and more likely to be acted upon by the customer or sponsor.

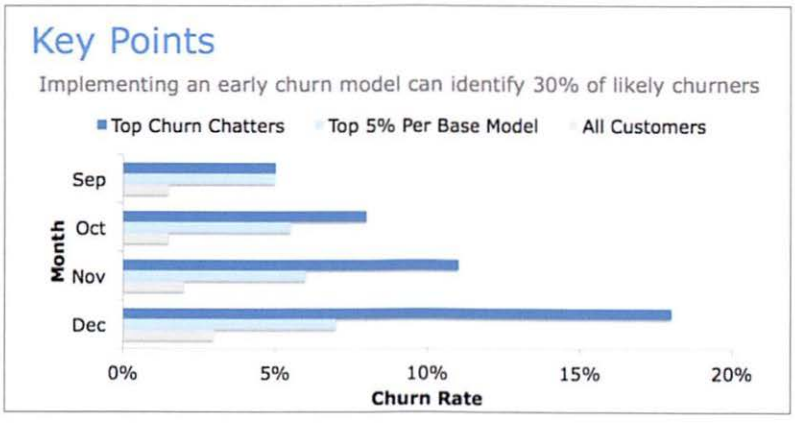

FIGURE 12-12 *Example of a presentation of key points of a data science project shown as a bar chart*

For analyst presentations, use more granular or technical charts and graphs. In this case, appropriate visualization techniques include dot charts, density plots, ROC curves, or histograms of a data distribution to support decisions made in the modeling techniques. Basic concepts of data visualization are discussed later in the chapter.

12.2.7 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.

This is where the team can provide discussion or written details related to the variables used in the model and explain how or why these variables were selected. In addition, the team should share the actual code (or at least an excerpt) developed to explain what was created and how it operates. This also serves to foster discussion related to any additional constraints or implications related to the main logic of the code. In addition, the team can use this section to illustrate details of the key variables and the predictive power of the model, using analyst-oriented charts and graphs, such as histograms, dot charts, density plots, and ROC curves.

Figure 12-13 provides a sample slide describing the data variables, and Figure 12-14 shows a sample slide with a technical graph to support the work.

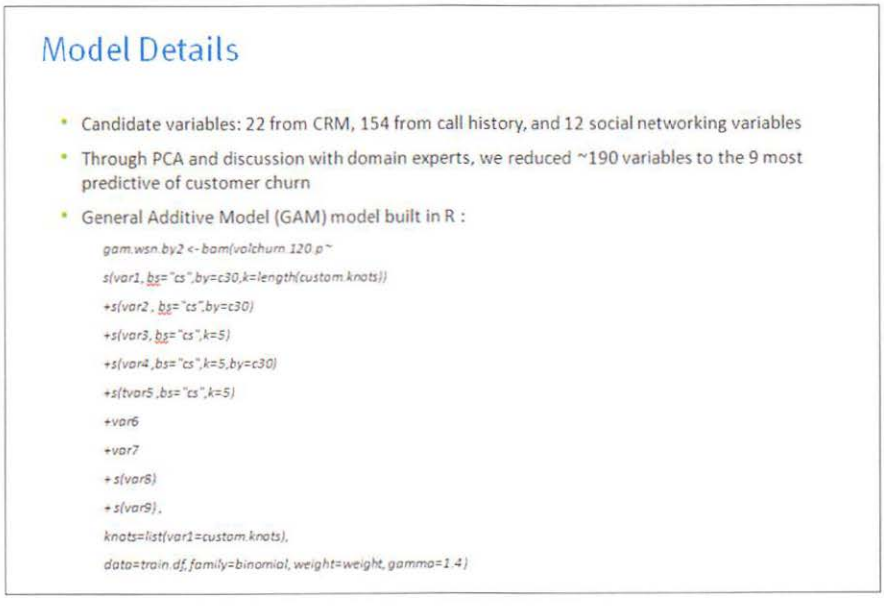

FIGURE 12-13 *Example of model details showing model type and variables*

As part of the model detail description, guidance should be provided regarding the speed with which the model can run in the test environment; the expected performance in a live, production environment; and the technology needed. This kind of discussion addresses how well the model can meet the organization's SLA.

This section of the presentation needs to include additional caveats, assumptions, or constraints of the model and model performance, such as systems or data the model needs to interact with, performance

issues, and ways to feed the outputs of the model into existing business processes. The author of this section needs to describe the relationships of the main variables on the project objectives, such as the effects of key variables on predicting churn, and the relationship of key variables to other variables. The team may even want to make suggestions to improve the model, highlight any risks to introducing bias into the modeling technique, or describe certain segments of the data that may skew the overall predictive power of the methodology.

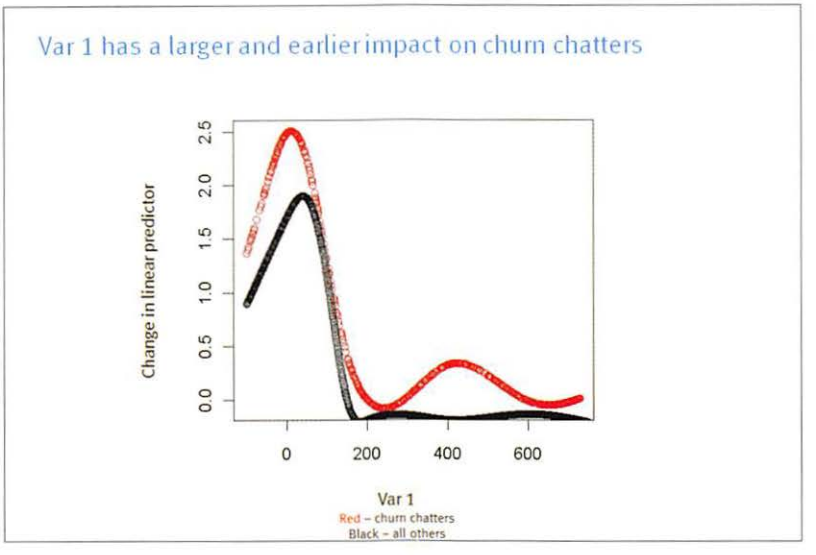

FIGURE 12-14 *Model details comparing two data variables*

12.2.8 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. For the Yoyodyne Bank example, Figure 12-15 provides possible recommendations from the project. In this section of the presentation, measuring the impact of the improvements and stating how to leverage that impact within the recommendations are key. For instance, the presentation might mention that every customer retained represents a time savings of six hours for one of the bank's account managers or $50,000 in savings of new account acquisitions, due to marketing costs, sales, and system-related costs.

For a presentation to a project sponsor audience, focus on the business impact of the project, including risks and ROI. Because project sponsors will be most interested in the business impact of the project, the presentation should also provide the sponsor with salient points to help evangelize the work within the organization. When preparing a presentation for analysts, supplement the main set of recommendations with any implications for the modeling or for deployment in a production environment. In either case, the

team should focus on recommending actions to operationalize the work and the benefits the customer will receive because of implementing these recommendations.

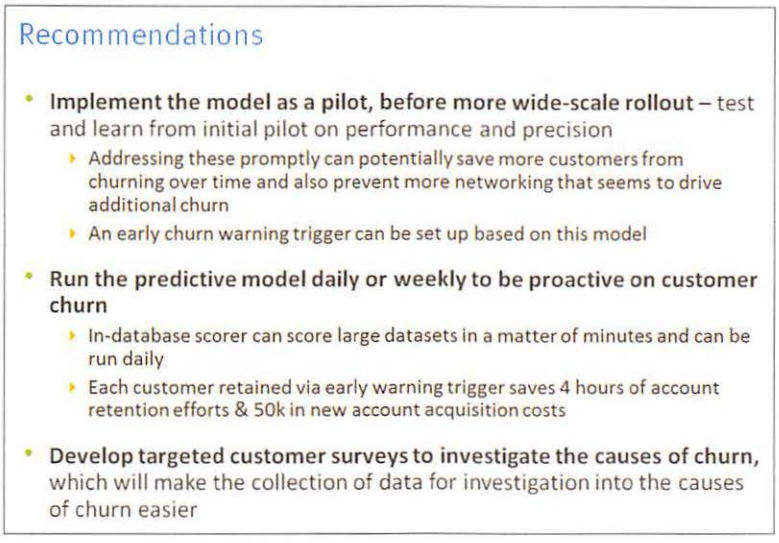

FIGURE 12-15 *Sample recommendations for a data science project*

12.2.9 Additional Tips on the Final Presentation

As a team completes a project and strives to move on to the next one, it must remember to invest adequate time in developing the final presentations. Orienting the audience to the project and providing context is important. On occasion, a team is so immersed in the project that it fails to provide sufficient context for its recommendations and the outputs of the models. A team needs to remember to spell out terminology and acronyms and avoid excessive use of jargon. It should also keep in mind that presentations may be shared extensively; therefore, recipients may not be familiar with the context and the journey the team has gone through over the course of the project.

The story may need to be told multiple times to different audiences, so the team must remain patient in repeating some of the key messages. These presentations should be viewed as opportunities to refine the key messages and evangelize the good work that was done. By this point in the process, the team has invested many hours of work and uncovered insights for the business. These presentations are an opportunity to communicate these projects and build support for future projects. As with most presentations, it is important to gauge the audience to guide shaping the message and the level of detail. Here are several more tips on developing the presentations.

- **Use imagery and visual representations:** Visuals tend to make the presentation more compelling. Also, people recall imagery better than words, because images can have a more visceral impact. These visual representations can be static and interactive data.

- **Make sure the text is mutually exclusive and collectively exhaustive (MECE):** This means having an economy of words in the presentation and making sure the key points are covered but not repeated unnecessarily.
- **Measure and quantify the benefits of the project:** This can be challenging and requires time and effort to do well. This kind of measurement should attempt to quantify benefits that have financial and other benefits in a specific way. Making the statement that a project provided "$8.5M in annual cost savings" is much more compelling than saying it has "great value."
- **Make the benefits of the project clear and conspicuous:** After calculating the benefits of the project, make sure to articulate them clearly in the presentation.

12.2.10 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. It is recommended that the team think through the implications of its work on the recipients of the code, the kinds of questions they will have, and their interests. For instance, indicating that the model will need to perform real-time monitoring may require extensive changes to an IT runtime environment, so the team may need to consider a compromise of nightly batch jobs to process the data. In addition, the team may need to get the technical team talking with the project sponsor to ensure the implementation and SLA will meet the business needs during the technical deployment.

The team should anticipate questions from IT related to how computationally expensive it will be to run the model in the production environment. If possible, indicate how well the model ran in the test scenarios and whether there are opportunities to tune the model or environment to optimize performance in the production environment.

Teams should approach writing technical documentation for their code as if it were an application programming interface (API). Many times, the models become encapsulated as functions that read a set of inputs in the production environment, possibly perform preprocessing on data, and create an output, including a set of post-processing results.

Consider the inputs, outputs, and other system constraints to enable a technical person to implement the analytical model, even if this person has not had a connection to the data science project up to this point. Think about the documentation as a way to introduce the data that the model needs, the logic it is using, and how other related systems need to interact with it in a production environment for it to operate well. The specifications detail the inputs the code needs and the data format and structures. For instance, it may be useful to specify whether structured data is needed or whether the expected data needs to be numeric or string formats. Describe any transformations that need to be made on the input data before the code can use it, and if scripting was created to perform these tasks. These kinds of details are important when other engineers must modify the code or utilize a different dataset or table, if and when the environment changes.

Regarding exception handling, the team must consider how the code should handle data that is outside the expected data ranges of the model parameters and how it will handle missing data values (Chapter 3, "Review of Basic Data Analytic Methods Using R"), null values, zeros, NAs, or data that is in an unexpected format or type. The technical documentation describes how to treat these exceptions and what implications may emerge on downstream processes. For the model outputs, the team must explain to what extent to post-process the output. For example, if the model returns a value representing

the probability of customer churn, additional logic may be needed to identify the scoring threshold to determine which customer accounts to flag as being at risk of churn. In addition, some provision should be made for adjusting this threshold and training the algorithm, either in an automated learning fashion or with human intervention.

Although the team must create technical documentation, many times engineers and other technical staff receive the code and may try to use it without reading through all the documentation. Therefore, it is important to add extensive comments in the code. This directs the people implementing the code on how to use it, explains what pieces of the logic are supposed to do, and guides other people through the code until they're familiar with it. If the team can do a thorough job adding comments in the code, it is much easier for someone else to maintain the code and tune it in the runtime environment. In addition, it helps the engineers edit the code when their environment changes or they need to modify processes that may be providing inputs to the code or receiving its outputs.

12.3 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 12-2 lists some popular tools.

TABLE 12-2 *Common Tools for Data Visualization*

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	

As the volume and complexity of data has grown, users have become more reliant on using crisp visuals to illustrate key ideas and portray rich data in a simple way. Over time, the open source community has developed many libraries to offer more options for portraying graphics data visually. Although this book showed examples primarily using the base package of R, `ggplot2` provides additional options for creating professional-looking data visualization, as does the `lattice` library for R.

Gnuplot and GGobi have a command-line-driven approach to generating data visualization. The genesis of these tools mainly grew out of scientific computing and the need to express complex data visually. GGobi

also has a variant called Rggobi that enables users to access the GGobi functionality with the R software and programming language. There are many open source mapping tools available, including Modest Maps and OpenLayers, both designed for developers who would like to create interactive maps and embed them within their own development projects or on the web. The software programming language development environment, Processing, employs a Java-like language for developers to create professional-looking data visualization. Because it is based on a programming language rather than a GUI, Processing enables developers to create robust visualization and have precise control over the output. D3.js is a JavaScript library for manipulating data and creating web-based visualization with standards, such as Hypertext Markup Language (HTML), Scalable Vector Graphics (SVG), and Cascading Style Sheets (CSS). For more examples of using open source visualization tools, refer to Nathan Yau's website, flowingdata.com [1], or his book ***Visualize This*** [2], which discusses additional methods for creating data representations with open source tools.

Regarding the commercial tools shown in Table 12-2, Tableau, Spotfire (by TIBCO), and QlikView function as data visualization tools and as interactive business intelligence (BI) tools. Due to the growth of data in the past few years, organizations for the first time are beginning to place more emphasis on ease of use and visualization in BI over more traditional BI tools and databases. These tools make visualization easy and have user interfaces that are cleaner and simpler to navigate than their predecessors. Although not traditionally considered a data visualization tool, Adobe Illustrator is listed in Table 12-2 because some professionals use it to enhance visualization made in other tools. For example, some users develop a simple data visualization in R, save the image as a PDF or JPEG, and then use a tool such as Illustrator to enhance the quality of the graphic or stitch multiple visualization work into an infographic. Inkscape is an open source tool used for similar use cases, with much of Illustrator's functionality.

12.3.1 Key Points Supported with Data

It is more difficult to observe key insights when data is in tables instead of in charts. To underscore this point, in ***Say it with Charts,*** Gene Zelazny [3] mentions that to highlight data, it is best to create a visual representation out of it, such as a chart, graph, or other data visualization. The opposite is also true. Suppose an analyst chooses to downplay the data. Sharing it in a table draws less attention to it and makes it more difficult for people to digest.

The way one chooses to organize the visual in terms of the color scheme, labels, and sequence of information also influences how the viewer processes the information and what he perceives as the key message from the chart. The table shown in Figure 12-16 contains many data points. Given the layout of the information, it is difficult to identify the key points at a glance. Looking at 45 years of store opening data can be challenging, as shown in Figure 12-16.

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 12-16 *Forty-five years of store opening data*

Even showing somewhat less data is still difficult to read through for most people. Figure 12-17 hides the first 10 years, leaving 35 years of data in the table.

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	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	26	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	33	27	35	47	32	39	27	4	1196
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

FIGURE 12-17 *Thirty-five years of store opening data*

As most readers will observe, it is challenging to make sense of data, even at relatively small scales. There are several observations in the data that one may notice, if one looks closely at the data tables:

- BigBox experienced strong growth in the 1980s and 1990s.
- By the 1980s, BigBox began adding more SuperBox stores to its mix of chain stores.
- SuperBox stores outnumber BigBox stores nearly 2 to 1 in aggregate.

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.

Figure 12-18 shows a map of the United States, with the points representing the geographic locations of the stores. 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. This map shows where the BigBox store has market saturation, where the company has grown, and where it has SuperBox stores and other BigBox stores, based on the color and shading. The visualization in Figure 12-18 clearly communicates more effectively than the dense tables in Figure 12-16 and Figure 12-17. For a sponsor audience, the analytics team can also use other simple visualization techniques to portray data, such as bar charts or line charts.

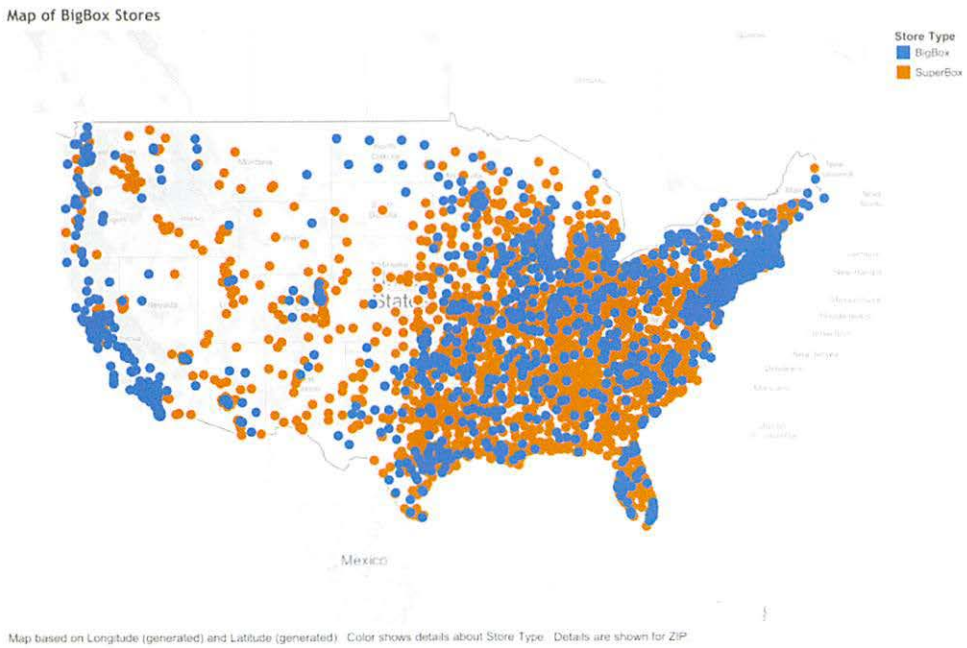

FIGURE 12-18 *Forty-five years of store opening data, shown as map*

12.3.2 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. Following is an example of the steps a data scientist may go through in exploring pricing data to understand the data better, model it, and assess whether a current pricing model is working effectively. Figure 12-19 shows a distribution of pricing data as a user score reflecting price sensitivity.

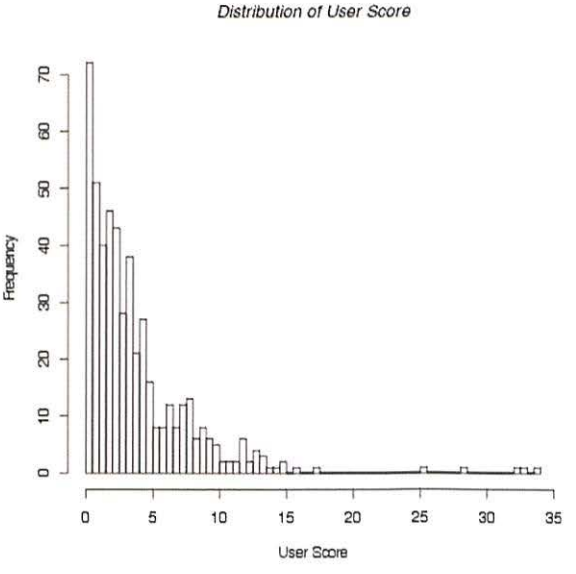

FIGURE 12-19 *Frequency distribution of user scores*

A data scientist's first step may be to view the data as a raw distribution of the pricing levels of users. Because the values have a long tail to the right, in Figure 12-19, it may be difficult to get a sense of how tightly clustered the data is between user scores of zero and five.

To understand this better, a data scientist may rerun this distribution showing a log distribution (Chapter 3) of the user score, as demonstrated in Figure 12-20.

This shows a less skewed distribution that may be easier for a data scientist to understand. Figure 12-21 illustrates a rescaled view of Figure 12-20, with the median of the distribution around 2.0. This plot provides the distribution of a new user score, or index, that may gauge the level of price sensitivity of a user when expressed in log form.

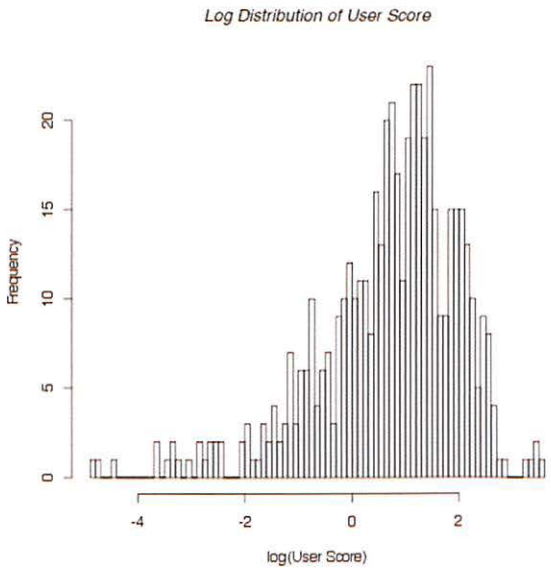

FIGURE 12-20 *Frequency distribution with log of user score*

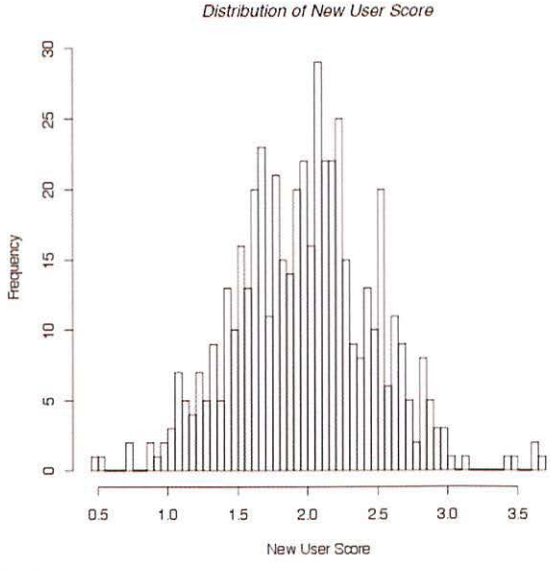

FIGURE 12-21 *Frequency distribution of new user scores*

Another idea may be to analyze the stability of price distributions over time to see if the prices offered to customers are stable or volatile. As shown in a graphic such as Figure 12-22, the prices appear to be stable. In this example, the user score of pricing remains within a tight band between two and three regardless of the time in days. In other words, the time in which a customer purchases a given product does not significantly influence the price she is willing to pay, as expressed by the user score, shown on the y-axis.

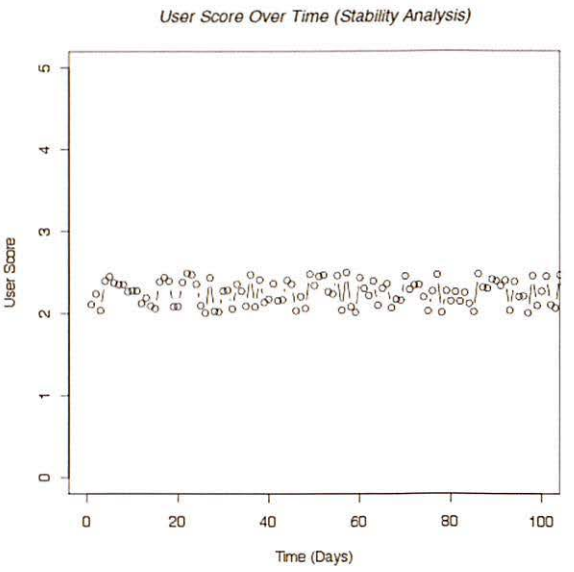

FIGURE 12-22 *Graph of stability analysis for pricing*

By this point the data scientist has learned the following about this example and made several observations about the data:

- Most user scores are between two and three in terms of their price sensitivity.
- After taking the log value of the user scores, a new user scoring index was created, which recentered the data values around the center of the distribution.
- The pricing scores appear to be stable over time, as the duration of the customer does not seem to have significant influence on the user pricing score. Instead, it appears to be relatively constant over time, within a small band of user scores.

At this point, the analysts may want to explore the range of price tiers offered to customers. Figures 12-22 and 12-23 demonstrate examples of the price tiering currently in place within the customer base.

Figure 12-23 shows the price distribution for a customer base. In this example, loyalty score and price are positively correlated; as the loyalty score increases, so do the prices that the customers are willing to pay. It may seem like a strange phenomenon that the most loyal customers in this example are willing to pay higher prices, but the reality is that customers who are very loyal tend to be less sensitive to price fluctuations or increases. The key, however, is to understand which customers are highly loyal so that appropriate pricing can be charged to the right groups of people.

Figure 12-24 shows a variation on 12-23. In this case, the new graphic portrays the same customer price tiers, but this time a rug representation (Chapter 3) has been added at the bottom to reflect the distribution of the data points.

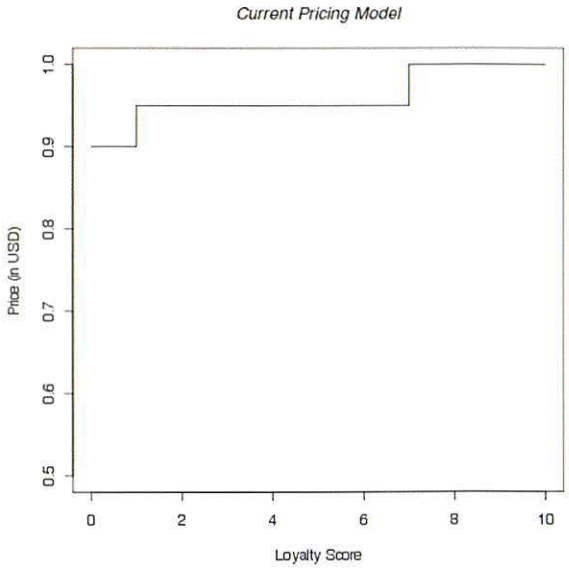

FIGURE 12-23 *Graph comparing the price in U.S. dollars with a customer loyalty score*

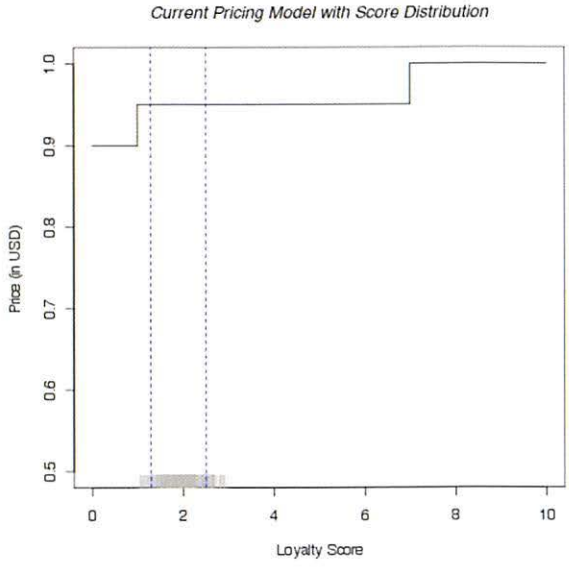

FIGURE 12-24 *Graph comparing the price in U.S. dollars with a customer loyalty score (with rug representation)*

This rug indicates that the majority of customers in this example are in a tight band of loyalty scores, between about 1 and 3 on the x-axis, all of which offered the same set of prices, which are high (between 0.9 and 1.0 on the y-axis). The y-axis in this example may represent a pricing score, or the raw value of a customer in millions of dollars. The important aspect is to recognize that the pricing is high and is offered consistently to most of the customers in this example.

Based on what was shown in Figure 12-25, the team may decide to develop a new pricing model. Rather than offering static prices to customers regardless of their level of loyalty, a new pricing model might offer more dynamic price points to customers. In this visualization, the data shows the price increases as more of a curvilinear slope relative to the customer loyalty score. The rug at the bottom of the graph indicates that most customers remain between 1 and 3 on the x-axis, but now rather than offering all these customers the same price, the proposal suggests offering progressively higher prices as customer loyalty increases. In one sense, this may seem counterintuitive. It could be argued that the best prices should be offered to the most loyal customers. However, in reality, the opposite is often the case, with the most attractive prices being offered to the least loyal customers. The rationale is that loyal customers are less price sensitive and may enjoy the product and stay with it regardless of small fluctuations in price. Conversely, customers who are not very loyal may defect unless they are offered more attractive prices to stay. In other words, less loyal customers are more price sensitive. To address this issue, a new pricing model that accounts for this may enable an organization to maximize revenue and minimize attrition by offering higher prices to more loyal customers and lower prices to less loyal customers. Creating an iterative depicting the data visually allows the viewer to see these changes in a more concrete way than by looking at tables of numbers or raw values.

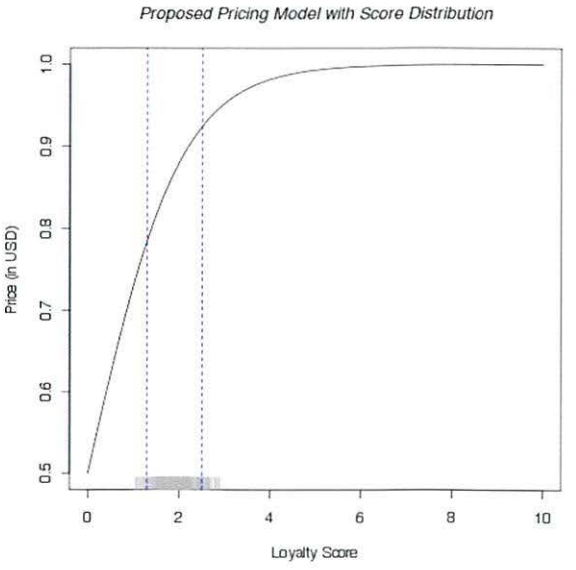

FIGURE 12-25 *New proposed pricing model compared to prices in U.S. dollars with rug*

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.

The resulting graph in the pricing scenario appears to be technical regarding the distribution of prices throughout a customer base and would be suitable for a technical audience composed of other data scientists. Figure 12-26 shows an example of how one may present this graphic to an audience of other data scientists or data analysts. This demonstrates a curvilinear relationship between price tiers and customer loyalty when expressed as an index. Note that the comments to the right of the graph relate to the precision of the price targeting, the amount of variability in robustness of the model, and the expectations of model speed when run in a production environment.

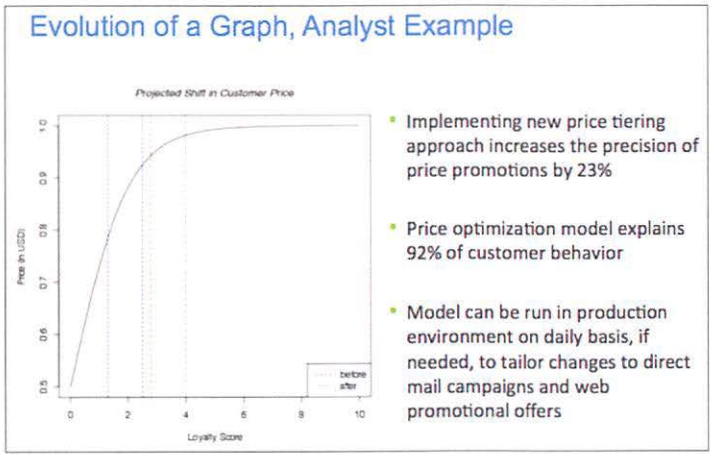

FIGURE 12-26 *Evolution of a graph, analyst example with supporting points*

Figure 12-27 portrays another example of the output from the price optimization project scenario, showing how one may present this to an audience of project sponsors. This demonstrates a simple bar chart depicting the average price per customer or user segment. Figure 12-27 shows a much simpler-looking visual than Figure 12-26. It clearly portrays that customers with lower loyalty scores tend to get lower prices due to targeting from price promotions. Note that the right side of the image focuses on the business impact and cost savings rather than the detailed characteristics of the model.

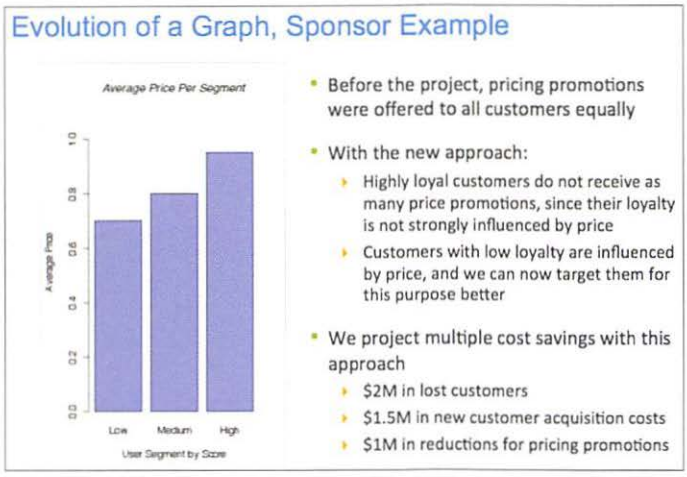

FIGURE 12-27 *Evolution of a graph, sponsor example*

The comments to the right side of the graphic in Figure 12-27 explain the impact of the model at a high level and the cost savings of implementing this approach to price optimization.

12.3.3 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. Table 12-3 shows some basic chart types to guide the reader in understanding that different types of charts are more suited to a situation depending on specific kinds of data and the message the team is attempting to portray. Using a type of chart for data it is not designed for may look interesting or unusual, but it generally confuses the viewer. The objective for the author is to find the best chart for expressing the data clearly so the visual does not impede the message, but rather supports the reader in taking away the intended message.

TABLE 12-3 *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

Table 12-3 shows the most fundamental and common data representations, which can be combined, embellished, and made more sophisticated depending on the situation and the audience. It is recommended

that the team consider the message it is trying to communicate and then select the appropriate type of visual to support the point. Misusing charts tends to confuse an audience, so it is important to take into account the data type and desired message when choosing a chart.

Pie charts are designed to show the components, or parts relative to a whole set of things. A pie chart is also the most commonly misused kind of chart. If the situation calls for using a pie chart, employ it only when showing only 2–3 items in a chart, and only for sponsor audiences.

Bar charts and line charts are used much more often and are useful for showing comparisons and trends over time. Even though people use vertical bar charts more often, horizontal bar charts allow an author more room to fit the text labels. Vertical bar charts tend to work well when the labels are small, such as when showing comparisons over time using years.

For frequency, histograms are useful for demonstrating the distribution of data to an analyst audience or to data scientists. As shown in the pricing example earlier in this chapter, data distributions are typically one of the first steps when visualizing data to prepare for model planning. To qualitatively evaluate correlations, scatterplots can be useful to compare relationships among variables.

As with any presentation, consider the audience and level of sophistication when selecting the chart to convey the intended message. These charts are simple examples but can easily become more complex when adding data variables, combining charts, or adding animation where appropriate.

12.3.4 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. Figure 12-28 portrays a line chart with several design problems.

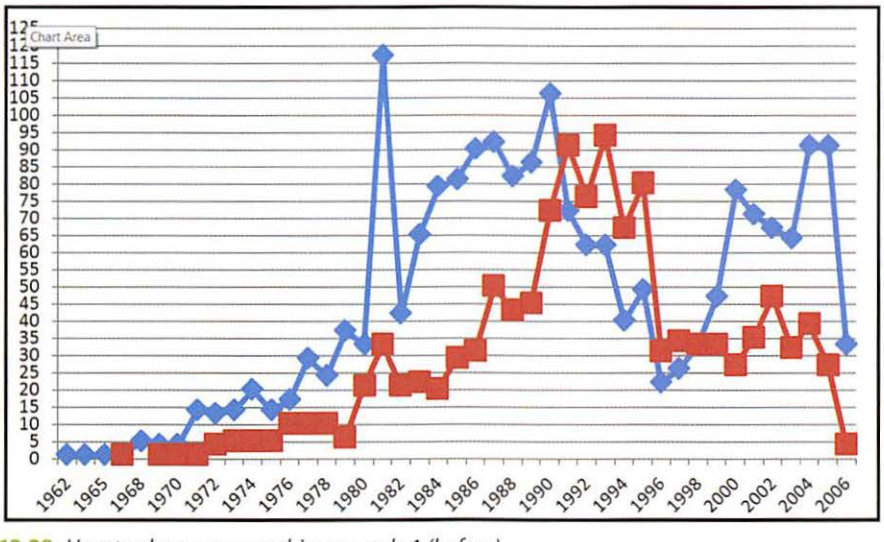

FIGURE 12-28 *How to clean up a graphic, example 1 (before)*

How to Clean Up a Graphic

The line chart shown in Figure 12-28 compares two trends over time. The chart looks busy and contains a lot of chart junk that distracts the viewer from the main message. *Chart junk* refers to elements of data visualization that provide additional materials but do not contribute to the data portion of the graphic. If chart junk were removed, the meaning and understanding of the graphic would not be diminished; it would instead be made clearer. There are five main kinds of "chart junk" in Figure 12-28:

- **Horizontal grid lines:** These serve no purpose in this graphic. They do not provide additional information for the chart.
- **Chunky data points:** These data points represented as large square blocks draw the viewer's attention to them but do not represent any specific meaning aside from the data points themselves.
- **Overuse of emphasis colors in the lines and border:** The border of the graphic is a thick, bold line. This forces the viewer's attention to the perimeter of the graphic, which contains no information value. In addition, the lines showing the trends are relatively thick.
- **No context or labels:** The chart contains no legend to provide context as to what is being shown. The lines also lack labels to explain what they represent.
- **Crowded axis labels:** There are too many axis labels, so they appear crowded. There is no need for labels on the y-axis to appear every five units or for values on the x-axis to appear every two units. Shown in this way, the axis labels distract the viewer from the actual data that is represented by the trend lines in the chart.

The five forms of chart junk in Figure 12-28 are easily corrected, as shown in Figure 12-29. Note that there is no clear message associated with the chart and no legend to provide context for what is shown in Figure 12-28.

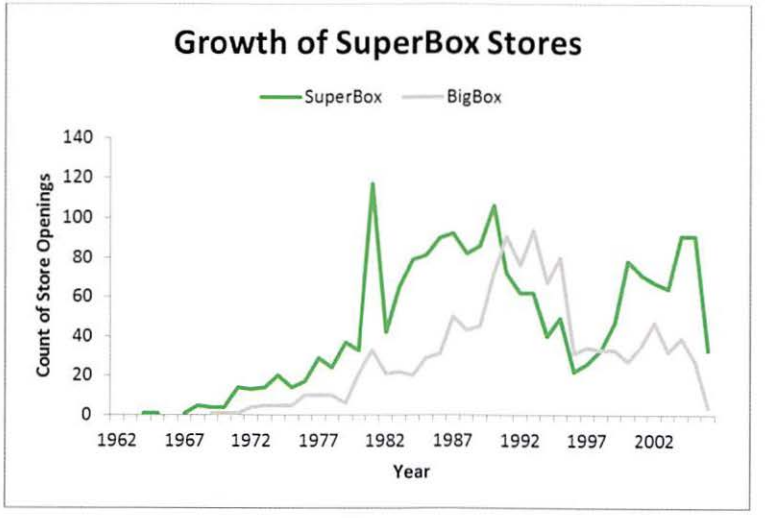

FIGURE 12-29 *How to clean up a graphic, example 1 (after)*

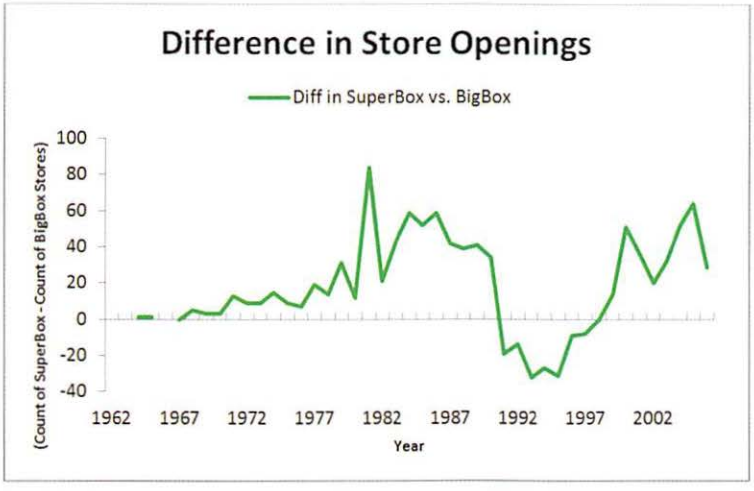

FIGURE 12-30 *How to clean up a graphic, example 1 (alternate "after" view)*

Figures 12-29 and 12-30 portray two examples of cleaned-up versions of the chart shown in Figure 12-28. Note that the problems with chart junk have been addressed. There is a clear label and title for each chart to reinforce the message, and color has been used in ways to highlight the point the author is trying to make. In Figure 12-29, a strong, green color is shown to represent the count of SuperBox stores, because this is where the viewer's focus should be drawn, whereas the count of BigBox stores is shown in a light gray color.

In addition, note the amount of white space being used in each of the two charts shown in Figures 12-29 and 12-30. Removing grid lines, excessive axes, and the visual noise within the chart allows clear contrast between the emphasis colors (the green line charts) and the standard colors (the lighter gray of the BigBox stores). When creating charts, it is best to draw most of the main visuals in standard colors, light tones, or color shades so that stronger emphasis colors can highlight the main points. In this case, the trend of BigBox stores in light gray fades into the background but does not disappear, while making the SuperBox stores trend in a darker gray (bright green in the online chart) makes it prominent to support the message the author is making about the growth of the SuperBox stores.

An alternative to Figure 12-29 is shown in Figure 12-30. If the main message is to show the difference in the growth of new stores, Figure 12-30 can be created to further simplify Figure 12-28 and graph only the difference between SuperBox stores compared to regular BigBox stores. Two examples are shown to illustrate different ways to convey the message, depending on what it is the author of these charts would like to emphasize.

How to Clean Up a Graphic, Second Example

Another example of cleaning up a chart is portrayed in Figure 12-31. This vertical bar chart suffers from more of the typical problems related to chart junk, including misuse of color schemes and lack of context.

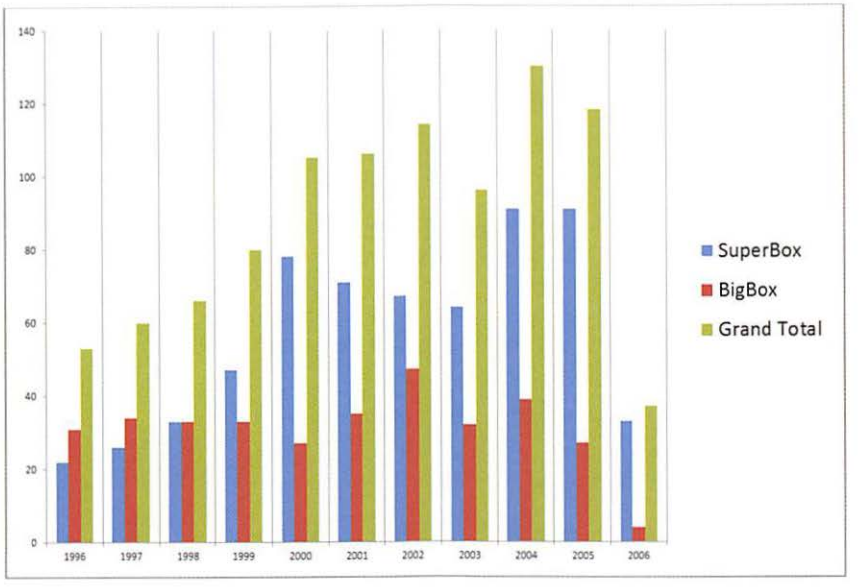

FIGURE 12-31 *How to clean up a graphic, example 2 (before)*

There are five main kinds of chart junk in Figure 12-31:

- **Vertical grid lines:** These vertical grid lines are not needed in this graphic. They provide no additional information to help the viewer understand the message in the data. Instead, these vertical grid lines only distract the viewer from looking at the data.
- **Too much emphasis color:** This bar chart uses strong colors and too much high-contrast dark gray-scale. In general, it is best to use subtle tones, with a low contrast gray as neutral color, and then emphasize the data underscoring the key message in a dark tone or strong color.
- **No chart title:** Because the graphic lacks a chart title, the viewer is not oriented to what he is viewing and does not have proper context.
- **Legend at right restricting chart space:** Although there is a legend for the chart, it is shown on the right side, which causes the vertical bar chart to be compressed horizontally. The legend would make more sense placed across the top, above the chart, where it would not interfere with the data being expressed.
- **Small labels:** The horizontal and vertical axis labels have appropriate spacing, but the font size is too small to be easily read. These should be slightly larger to be easily read, while not appearing too prominent.

Figures 12-32 and 12-33 portray two examples of cleaned-up versions of the chart shown in Figure 12-31. The problems with chart junk have been addressed. There is a clear label and title for each chart to reinforce the message, and appropriate colors have been used in ways to highlight the point the author is trying to make. Figures 12-32 and 12-33 show two options for modifying the graphic, depending on the main point the presenter is trying to make.

Figure 12-32 shows strong emphasis color (dark blue) representing the SuperBox stores to support the chart title: Growth of SuperBox Stores.

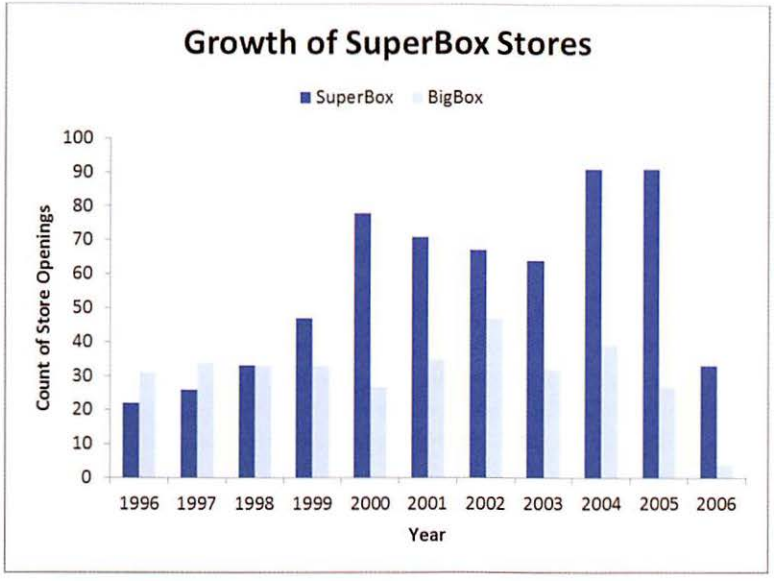

FIGURE 12-32 *How to clean up a graphic, example 2 (after)*

Suppose the presenter wanted to talk about the total growth of BigBox stores instead. A line chart showing the trends over time would be a better choice, as shown in Figure 12-33.

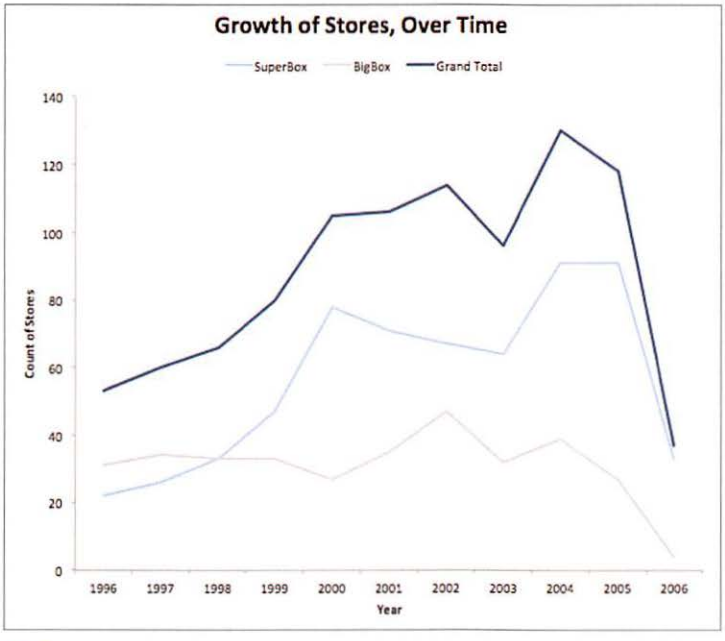

FIGURE 12-33 *How to clean up a graphic, example 2 (alternate view of "after")*

In both cases, the noise and distractions within the chart have been removed. As a result, the data in the bar chart for providing context has been deemphasized, while other data has been made more prominent because it reinforces the key point as stated in the chart's title.

12.3.5 Additional Considerations

As stated in the previous examples, the emphasis should be on simplicity when creating charts and graphs. Create graphics that are free of chart junk and utilize the simplest method for portraying graphics clearly. The goal of data visualization should be to support the key messages being made as clearly as possible and with few distractions.

Similar to the idea of removing chart junk is being cognizant of the data-ink ratio. *Data-ink* refers to the actual portion of a graphic that portrays the data, while *non-data ink* refers to labels, edges, colors, and other decoration. If one imagined the ink required to print a data visualization on paper, the data-ink ratio could be thought of as (data-ink)/(total ink used to print the graphic). In other words, the greater the ratio of data-ink in the visual, the more data rich it is and the fewer distractions it has [4].

Avoid Using Three-Dimensions in Most Graphics

One more example where people typically err is in adding unnecessary shading, depth, or dimensions to graphics. Figure 12-34 shows a vertical bar chart with two visible dimensions. This example is simple and easy to understand, and the focus is on the data, not the graphics. The author of the chart has chosen to highlight the SuperBox stores in a dark blue color, while the BigBox bars in the chart are in a lighter blue. The title is about the growth of SuperBox stores, and the SuperBox bars in the chart are in a dark, high-contrast shade that draws the viewer's attention to them.

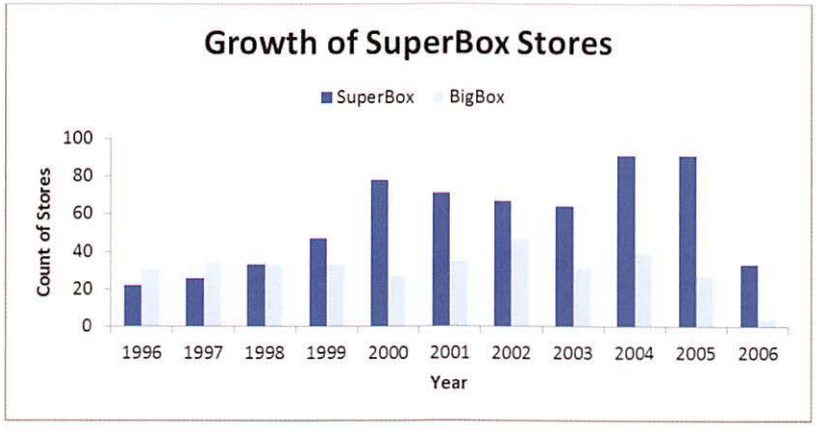

FIGURE 12-34 *Simple bar chart, with two dimensions*

Compare Figure 12-34 to Figure 12-35, which shows a three-dimensional chart. Figure 12-35 shows the original bar chart at an angle, with some attempt at showing depth. This kind of three-dimensional perspective makes it more difficult for the viewer to gauge the actual data and the scaling becomes deceptive.

Three-dimensional charts often distort scales and axes, and impede viewer cognition. Adding a third dimension for depth in Figure 12-35, does not make it fancier, just more difficult to understand.

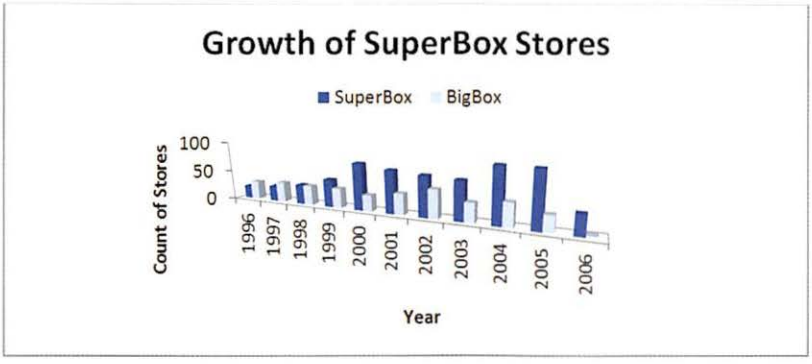

FIGURE 12-35 *Misleading bar chart, with three dimensions*

The charts in Figures 12-34 and 12-35 portray the same data, but it is more difficult to judge the actual height of the bars in Figure 12-35. Moreover, the shadowing and shape of the chart cause most viewers to spend time looking at the perspective of the chart rather than the height of the bars, which is the key message and purpose of this data visualization.

Summary

Communicating the value of analytical projects is critical for sustaining the momentum of a project and building support within organizations. This support is instrumental in turning a successful project into a system or integrating it properly into an existing production environment. Because an analytics project may need to be communicated to audiences with mixed backgrounds, this chapter recommends creating four deliverables to satisfy most of the needs of various stakeholders.

- A presentation for a project sponsor
- A presentation for an analytical audience
- Technical specification documents
- Well-annotated production code

Creating these deliverables enables the analytics project team to communicate and evangelize the work that it did, whereas the code and technical documentation assists the team that wants to implement the models within the production environment.

This chapter illustrates the importance of selecting clear and simple visual representations to support the key points in the final presentations or for portraying data. Most data representations and graphs can be improved by simply removing the visual distractions. This means minimizing or removing chart junk, which distracts the viewer from the main purpose of a chart or graph and does not add information value.