

<b>Synopsis of YoyoDyne Bank Case Study</b>	
<ul style="list-style-type: none"> <li>▪ YoyoDyne Bank is a retail bank that wants to improve its Net Present Value (NPV) and its customer retention rate.</li> <li>▪ It wants to establish an effective marketing campaign targeting customers to reduce the churn rate by at least five percent.</li> <li>▪ 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.</li> </ul> <p>The bank wants to build a data warehouse to support marketing and other related customer care groups.</p>	

FIGURE 12-3 Sinopsis del ejemplo de estudio de caso de YoyoDyne Bank

Residencia en  
durante el p

similar a la Figura 12-4

<b>Retail Banking: YoyoDyne Bank</b>	
<b>Components of Analytic Plan</b>	
<b>Discovery</b>	How can the bank identify customers with the highest likelihood for churn?
<b>Business Problem Framed</b>	
<b>Initial Hypotheses</b>	Transaction volume and type are key predictors of churn rates
<b>Data and Scope</b>	5 months of customer account history
<b>Model Planning - Analytic Technique</b>	Logistic regression to identify most influential factors predicting churn
<b>Result and Key Findings</b>	<p>Key predictors of churn are:</p> <ol style="list-style-type: none"> <li>Once customers stop using their accounts for gas and groceries, their account holdings quickly diminish and the customers churn.</li> <li>If the customers use their debit card fewer than 5 times per month, they will leave the bank within 60 days.</li> </ol>
<b>Business Impact</b>	<p>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.</p>

FIGURE 12-4 Plan de análisis para el caso de estudio de YoyoDyne Bank

Además de guiar la planificación y la metodología del modelo, el plan analítico contiene componentes que pueden usarse como entradas para escribir sobre el alcance, los supuestos subyacentes, las técnicas de modelado, las hipótesis iniciales y los hallazgos clave en las presentaciones finales. Después de pasar una cantidad considerable de tiempo en el modelado y realizar un análisis de datos en profundidad, es fundamental reflexionar sobre el trabajo del proyecto y considerar

el contexto de los problemas que el equipo se propuso resolver. Revise el trabajo que se completó durante el proyecto e identifique las observaciones sobre los productos, la puntuación y los resultados del modelo. Con base en estas observaciones, comience a identificar los mensajes clave y cualquier conocimiento inesperado.

Además, es importante adaptar los resultados del proyecto a la audiencia. Para un patrocinador de proyecto, demuestre que el equipo cumplió con los objetivos del proyecto. Concéntrese en lo que se hizo, lo que logró el equipo, qué ROI se puede anticipar y qué valor comercial se puede lograr. Dé puntos de conversación al patrocinador del proyecto para evangelizar el trabajo. Recuerde que el patrocinador debe transmitir la historia a los demás, así que facilite el trabajo de esta persona y ayude a garantizar que el mensaje sea preciso proporcionando algunos puntos de conversación. Encuentre formas de enfatizar el ROI y el valor comercial, y mencione si los modelos se pueden implementar dentro de las limitaciones de desempeño del entorno de producción del patrocinador.

En algunas organizaciones, es posible que no se espere que el equipo de ciencia de datos presente un caso comercial completo para proyectos futuros e implementación de los modelos. En su lugar, debe poder brindar orientación sobre el impacto de los modelos para permitir que el patrocinador del proyecto, o alguien designado por esa persona, cree un caso de negocios para abogar por el piloto y la implementación posterior de esta funcionalidad. En otras palabras, el equipo de ciencia de datos puede ayudar en este esfuerzo poniendo los resultados del trabajo de modelado y ciencia de datos en contexto para ayudar a evaluar el valor real y el costo de implementar este trabajo de manera más amplia.

Cuando se presente a una audiencia técnica, como científicos de datos y analistas, concéntrese en cómo se hizo el trabajo. Discuta cómo el equipo logró los objetivos y las decisiones que tomó al seleccionar modelos o analizar los datos. Comparta métodos analíticos y procesos de toma de decisiones para que otros analistas puedan aprender de ellos para proyectos futuros. Describa los métodos, técnicas y tecnologías utilizadas, ya que esta audiencia técnica estará interesada en conocer estos detalles y considerar si el enfoque tiene sentido en este caso y si puede extenderse a otros proyectos similares. Planifique proporcionar datos específicos relacionados con la precisión y la velocidad del modelo, como el rendimiento del modelo en un entorno de producción.

Idealmente, el equipo debería considerar comenzar el desarrollo de la presentación final durante el proyecto y no al final del proyecto, como ocurre comúnmente. Este enfoque asegura que el equipo siempre tenga una versión de la presentación con hipótesis de trabajo para mostrar a las partes interesadas, en caso de que sea necesario mostrar una versión del trabajo en proceso del progreso del proyecto con poca antelación. De hecho, muchos analistas escriben el resumen ejecutivo al comienzo de un proyecto y luego lo van perfeccionando continuamente a lo largo del tiempo, de modo que al final del proyecto, partes de la presentación final ya estén completadas. Este enfoque también reduce la posibilidad de que los miembros del equipo olviden puntos clave o conocimientos descubiertos durante el proyecto. Por último, reduce la cantidad de trabajo a realizar en la presentación al finalizar el proyecto.

### 12.2.1 Desarrollo de CoreMaterial para múltiples audiencias

Debido a que algunos de los componentes de los proyectos se pueden usar para diferentes audiencias, puede ser útil crear un conjunto básico de materiales relacionados con el proyecto, que se puede utilizar para crear presentaciones para una audiencia técnica o un patrocinador ejecutivo.

La Tabla 12-1 describe los componentes principales de las presentaciones finales para el patrocinador del proyecto y una audiencia analista. Tenga en cuenta que los equipos pueden crear un conjunto básico de materiales en estas siete áreas, que se pueden utilizar para las dos audiencias de presentación. Se pueden utilizar tres áreas (Objetivos del proyecto, Hallazgos principales y Descripción del modelo) tal cual para ambas presentaciones. Otras áreas necesitan una elaboración adicional, como el Enfoque. Otras áreas, como los puntos clave, requieren diferentes niveles de detalle para los analistas y científicos de datos que para el patrocinador del proyecto. Cada uno de estos componentes principales de la presentación final se analiza en las secciones siguientes.

**T PODER 12-1 Comparación de materiales para presentaciones de patrocinadores y analistas**

Objetivos del proyecto	Enumere los 3-5 objetivos principales acordados.	
Hallazgos principales	Enfatice los mensajes clave.	
Acercarse	Nivel alto metodología	Metodología de alto nivel Detalles relevantes sobre técnicas y tecnología de modelado
descripción del modelo	Descripción general de la técnica de modelado	
Puntos clave	Clave de soporte	Muestre detalles para respaldar los puntos clave.
Apoyado con	puntos con gráficos simples	Cuadros y gráficos orientados al analista, como curvas e histogramas
Datos	y gráficos (ejemplo: bar gráficos).	ROC Imágenes de variables clave y significado de cada una
Detalles del modelo	Omita esta sección ción, o discutir solo en un alto nivel.	Muestre el código o la lógica principal del modelo e incluya el tipo de modelo, las variables y la tecnología utilizada para ejecutar el modelo y puntuar los datos.  Identifique las variables clave y el impacto de cada una.  Describa el rendimiento esperado del modelo y cualquier advertencia.  Descripción detallada de la técnica de modelado Discutir las variables, el alcance y el poder predictivo.
Recomendaciones	Centrarse en el negocio impacto de ness, incluyendo riesgos	Complemente las recomendaciones con implicaciones para el modelado o para la implementación en un entorno de producción.  y ROI.  Dar el patrocinio sor saliente puntos para ayudar su evangelizar trabajar dentro del organización.

**12.2.2 Objetivos del proyecto**

La parte de las Metas del Proyecto de la presentación final es generalmente la misma, o similar, para los patrocinadores y para los analistas. Para cada audiencia, el equipo debe reiterar los objetivos del proyecto para sentar las bases para

la solución y las recomendaciones que se comparten más adelante en la presentación. Además, la diapositiva de Objetivos sirve para garantizar que haya un entendimiento compartido entre el equipo del proyecto y los patrocinadores y para confirmar que están alineados para avanzar en el proyecto. Generalmente, los objetivos se acuerdan al principio del proyecto. Es una buena práctica escribirlos y compartirlos para asegurarse de que tanto el equipo del proyecto como los patrocinadores entiendan claramente las metas y los objetivos.

Las figuras 12-5 y 12-6 muestran dos ejemplos de diapositivas para los objetivos del proyecto. La figura 12-5 muestra tres objetivos para crear un pred enfatizar lo que n

## 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

FIGURE 12-5 Ejemplo de diapositiva de objetivos del proyecto para el estudio de caso de YoyoDyne

La Figura 12-6 muestra una variación de la diapositiva de Objetivos del Proyecto anterior en la Figura 12-5. Es un resumen de la situación antes de enumerar los objetivos. Tenga en cuenta que al realizar presentaciones finales, estos entregables se comparten dentro de las organizaciones y el contexto original se puede perder, especialmente si el patrocinador original abandona el grupo o cambia de roles. Es una buena práctica recapitular brevemente la situación antes de mostrar los objetivos del proyecto. Tenga en cuenta que agregar una descripción general de la situación a la diapositiva Objetivos hace que parezca más ocupada. El equipo debe determinar si dividir esto en una diapositiva separada o mantenerlo unido, según la audiencia y el estilo del equipo para realizar la presentación final.

Un método para escribir la descripción general de la situación de manera sucinta es resumirla en tres viñetas, de la siguiente manera:

- **Situación:** Brinde una descripción general de una oración de la situación que ha llevado al proyecto de análisis.
- **Complicación:** Dé una descripción general de una oración sobre la necesidad de abordar esto ahora. Algo ha provocado que la organización decida actuar en este momento. Por ejemplo, tal vez perdió 100

clientes en las últimas dos semanas y ahora tiene un mandato ejecutivo para abordar un problema, o tal vez ha perdido cinco puntos de participación de mercado frente a su mayor competidor en los últimos tres meses. Por lo general, esta oración representa el impulsor de por qué se inicia un proyecto en particular en este momento, en lugar de en algún momento vago en el futuro.

- **Implicación:** Brinde una descripción general de una oración del impacto de la complicación. Por ejemplo, si el banco no resuelve su problema de deserción de clientes, puede perder su posición dominante en el mercado en tres de hacer el proyecto.

## 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 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

**FIGURE 12-6** Ejemplo de diapositiva de situación y objetivos de proyecto. *Muchos de los datos de muestra fallan al ser ejecutados en el entorno de producción*

### 12.2.3 Principales hallazgos

Escriba un resumen ejecutivo sólido para describir los principales hallazgos de un proyecto. En muchos casos, el resumen puede ser la única parte de la presentación que leerán los gerentes apresurados. Por esta razón, es imperativo que el lenguaje sea claro, conciso y completo. Quienes lean el resumen ejecutivo deben poder captar la historia completa del proyecto y las ideas clave en una sola diapositiva. Además, esta es una oportunidad para proporcionar puntos de conversación clave para que el patrocinador ejecutivo los utilice para evangelizar el trabajo del proyecto con otros en la organización del cliente.

Asegúrese de enmarcar los resultados del proyecto en términos de valor comercial tanto cuantitativo como cualitativo. Esto es especialmente importante si la presentación es para el patrocinador del proyecto.

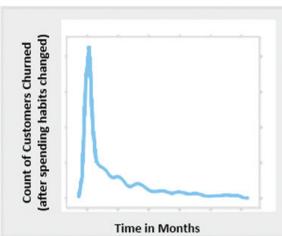
La Figura 12-7 muestra un ejemplo de una diapositiva de resumen ejecutivo para el estudio de caso de YoyoDyne. Es útil observar más de cerca las partes de la diapositiva para asegurarse de que esté clara. Tenga en cuenta que este no es el único formato para transmitir el Resumen Ejecutivo; varía según el estilo del autor, aunque muchos de los componentes clave son temas comunes en los resúmenes ejecutivos.

## 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 churker's contacts
- ▶ With CRM data, we can identify 20% of churners, adding social media data increases this

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

El mensaje clave debe ser claro y visible en la parte delantera de la diapositiva. Puede diferenciarse con color o sombreado, como se muestra en la Figura 12-8; También se pueden utilizar otras técnicas para llamar la atención. El mensaje clave puede convertirse en el único tema de conversación que los ejecutivos o el patrocinador del proyecto extraen del proyecto y lo utilizan para respaldar la recomendación del equipo para un proyecto piloto, por lo que debe ser conciso y convincente. Para que este mensaje sea lo más fuerte posible, mida el valor del trabajo y cuantifique los ahorros de costos, ingresos, ahorro de tiempo u otros beneficios para que el impacto comercial sea concreto.

Siga el mensaje clave con tres puntos de apoyo principales. Aunque las diapositivas del Resumen ejecutivo pueden tener más de tres puntos principales, ir más allá de las tres ideas dificulta que las personas recuerden los puntos principales, por lo que es importante asegurarse de que las ideas permanezcan claras y limitadas a las pocas ideas más impactantes que el equipo desea que la audiencia se lleve de trabajo que se hizo. Si el autor enumera diez puntos clave, los mensajes se diluyen y la audiencia puede recordar solo uno o dos puntos principales.

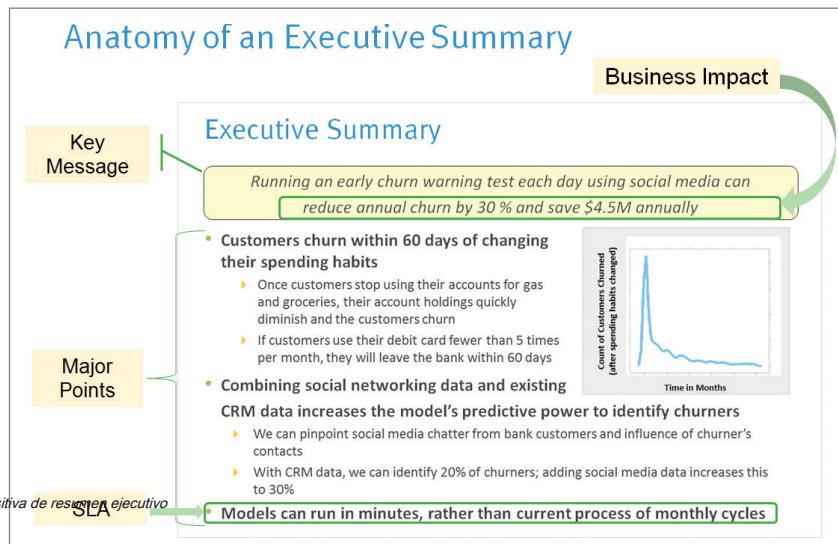
Además, debido a que se trata de un proyecto de análisis, asegúrese de tomar uno de los puntos clave relacionados con si el trabajo cumplirá con las expectativas o el acuerdo de nivel de servicio (SLA) del patrocinador. Tradicionalmente, el SLA se refiere a un acuerdo entre alguien que brinda servicios, como un departamento de tecnología de la información (TI) o una empresa de consultoría, y un usuario final o cliente. En este caso, el SLA se refiere al rendimiento del sistema, el tiempo de actividad esperado de un sistema y otras restricciones que rigen un acuerdo. Este término se ha vuelto menos formal y muchas veces transmite el desempeño del sistema o las expectativas más generalmente relacionadas con el desempeño o la puntualidad. Es en este sentido que aquí se utiliza SLA. Es decir, en este contexto, SLA

FIGURE 12-7 Ejemplo de diapositiva de resumen ejecutivo para el estudio de caso de YoyoDyne

se refiere al desempeño esperado de un sistema y la intención de que los modelos desarrollados no impacten adversamente el desempeño esperado del sistema en el que están integrados.

Finalmente, aunque

Servicio de imágenes visuales



## 12.2.4 Enfoque

En la parte de Enfoque de la presentación, el equipo debe explicar la metodología seguida en el proyecto. Esto puede incluir entrevistas con expertos en el dominio, los grupos que colaboran dentro de la organización y algunas declaraciones sobre la solución desarrollada. El objetivo de esta diapositiva es asegurar que la audiencia comprenda el curso de acción que se siguió lo suficientemente bien como para explicarlo a otros dentro de la organización. El equipo también debe incluir cualquier comentario adicional relacionado con los supuestos de trabajo que el equipo siguió mientras realizaba el trabajo, ya que esto puede ser crítico para defender por qué siguió un curso de acción específico.

Al explicar la solución, la discusión debe permanecer a un alto nivel para los patrocinadores del proyecto. Si se presenta a analistas o científicos de datos, proporcione detalles adicionales sobre el tipo de modelo utilizado, incluida la tecnología y el rendimiento real del modelo durante las pruebas. Finalmente, como parte de la descripción del enfoque, el equipo puede querer mencionar las restricciones de los sistemas, herramientas o procesos existentes y cualquier implicación sobre cómo estas cosas pueden necesitar cambiar con este proyecto.

Figura 12-9 mostrar  
proyecto a un patrocinador

una ciencia

## 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

**FIGURE 12-9** Ejemplo que describe la metodología del proyecto para patrocinadores de datos

para analistas y científicos de datos. La descripción muestra el desarrollo de un modelo de rendimiento dentro de YoyoDyne's

producción ambiente

Tenga en cuenta que la  
tercera vide detalles adicionales

O-  
0.

## 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

**FIGURE 12-10** Ejemplo que describe la metodología del proyecto para analistas y científicos de datos

para analistas y científicos de datos. La descripción muestra el desarrollo de un modelo de rendimiento dentro de YoyoDyne's

producción ambiente

La figura 12-10 muestra una variación del enfoque y la metodología utilizados en el proyecto de ciencia de datos. En este caso, la mayor parte del lenguaje y la descripción son los mismos que en el ejemplo para los patrocinadores del proyecto.

La principal diferencia es que esta versión contiene detalles adicionales sobre el tipo de modelo utilizado y la forma en que el modelo puntuará los datos rápidamente para cumplir con el SLA. Estas diferencias se destacan en los cuadros que se muestran en la Figura 12-10.

### 12.2.5 Descripción del modelo

Después de describir el enfoque del proyecto, los equipos generalmente incluyen una descripción del modelo que se utilizó. La figura 12-11 proporciona la descripción del modelo para el ejemplo de Yoyodyne Bank. Aunque la diapositiva de Descripción del modelo puede ser la misma para ambas audiencias, los intereses y objetivos difieren para cada una. Para el patrocinador, la metodología general debe articularse sin entrar en detalles excesivos. Transmitir la metodología básica seguida en el trabajo del equipo para permitir que el patrocinador comunique esto a otros dentro de la organización y proporcione puntos de conversación.

Es fundamental mencionar el alcance de los datos utilizados. El propósito es ilustrar la minuciosidad y transmitir confianza en que el equipo utilizó un enfoque que retrata con precisión su problema y está lo más libre de sesgos posible. Un rasgo clave de un buen científico de datos es la capacidad de ser escéptico del propio trabajo. Esta es una oportunidad para ver el trabajo y el entregable de manera crítica y considerar cómo la audiencia recibirá el trabajo. Trate de asegurarse de que sea una visión imparcial del proyecto y los resultados.

Suponiendo que el modelo cumplirá con los SLA acordados, mencione que el modelo cumplirá con los SLA según el rendimiento del modelo dentro del entorno de prueba o ensayo. Por ejemplo, uno puede querer indicar que el modelo procesó 500,000 registros en 5 minutos para dar a los interesados una idea de la velocidad del modelo durante el tiempo de ejecución. Los analistas querrán comprender los detalles del modelo, incluidas las decisiones tomadas al construir el modelo y el alcance de las extracciones de datos para pruebas y capacitación. Estar preparado

el entorno de prueba

### 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

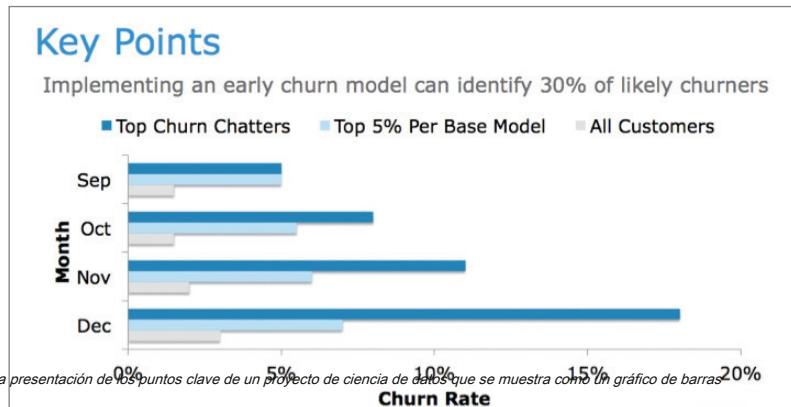
FIGURE 12-11 Ejemplo de descripción de modelo para un proyecto de ciencia de datos

## 12.2.6 Puntos clave compatibles con los datos

El siguiente paso es identificar puntos clave basados en conocimientos y observaciones resultantes de los datos y los resultados de la puntuación del modelo. Encuentra formas de ilustrar los puntos clave con gráficos y técnicas de visualización, utilizando gráficos más simples para los patrocinadores y una visualización de datos más técnica para los analistas y científicos de datos.

La Figura 12-12 muestra un ejemplo de proporcionar detalles de respaldo con respecto a la tasa de clientes del banco que se perderían en varios meses. Al desarrollar los puntos clave, tenga en cuenta los conocimientos que generarán el mayor impacto comercial y que se pueden defender con datos. Para los patrocinadores de proyectos, utilice gráficos simples como gráficos de barras, que ilustran los datos con claridad y permiten que la audiencia comprenda el valor de los conocimientos. Este también es un buen punto para presagiar algunas de las recomendaciones del equipo y comenzar a unir ideas para demostrar qué condujo a las recomendaciones y por qué. En otras palabras, esta sección proporciona los datos y la base para las recomendaciones que vienen más adelante en la presentación. Crear diapositivas claras y atractivas para

e actuó  
sobre por el



Para presentaciones de analistas, utilice cuadros y gráficos más detallados o técnicos. En este caso, las técnicas de visualización apropiadas incluyen gráficos de puntos, diagramas de densidad, curvas ROC o histogramas de una distribución de datos para respaldar las decisiones tomadas en las técnicas de modelado. Los conceptos básicos de visualización de datos se tratan más adelante en este capítulo.

## 12.2.7 Detalles del modelo

Los detalles del modelo suelen ser necesarios para las personas que tienen una comprensión más técnica que los patrocinadores, como los que implementarán el código o los colegas del equipo de análisis. Los patrocinadores del proyecto suelen estar menos interesados en los detalles del modelo; por lo general, se centran más en las implicaciones comerciales del trabajo que en los detalles del modelo. Esta parte de la presentación debe mostrar el código o la lógica principal del modelo, incluido el tipo de modelo, las variables y la tecnología utilizada para ejecutar

el modelo y puntúe los datos. El segmento de detalles del modelo de la presentación debe centrarse en describir el desempeño esperado del modelo y cualquier advertencia relacionada con el desempeño del modelo. Además, esta parte de la presentación debe proporcionar una descripción detallada de la técnica de modelado, las variables, el alcance y la efectividad esperada del modelo.

Aquí es donde el equipo puede proporcionar una discusión o detalles escritos relacionados con las variables utilizadas en el modelo y explicar cómo o por qué se seleccionaron estas variables. Además, el equipo debe compartir el código real (o al menos un extracto) desarrollado para explicar qué se creó y cómo funciona. Esto también sirve para fomentar la discusión relacionada con cualquier restricción o implicación adicional relacionada con la lógica principal del código. Además, el equipo puede usar esta sección para ilustrar los detalles de las variables clave y el poder predictivo del modelo, utilizando tablas y gráficos orientados al analista, como histogramas, gráficos de puntos, diagramas de densidad y curvas ROC.

Figura 12  
deslice con en

## 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)
```

FIGURE 12-13 Ejemplo de detalles del modelo que muestran el tipo de modelo y las variables

Como parte de la descripción detallada del modelo, se debe proporcionar orientación con respecto a la velocidad con la que el modelo puede ejecutarse en el entorno de prueba; el rendimiento esperado en un entorno de producción en vivo; y la tecnología necesaria. Este tipo de discusión aborda qué tan bien el modelo puede cumplir con el SLA de la organización.

Esta sección de la presentación debe incluir advertencias, suposiciones o restricciones adicionales del modelo y el rendimiento del modelo, como sistemas o datos con los que el modelo necesita interactuar, rendimiento

problemas y formas de incorporar los resultados del modelo a los procesos comerciales existentes. El autor de esta sección debe describir las relaciones de las principales variables en los objetivos del proyecto, como los efectos de las variables clave en la predicción de la deserción y la relación de las variables clave con otras variables. El equipo puede incluso querer hacer sugerencias para mejorar el modelo, resaltar cualquier riesgo de introducir sesgos en el modelo técnico.

del metodo

poder

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

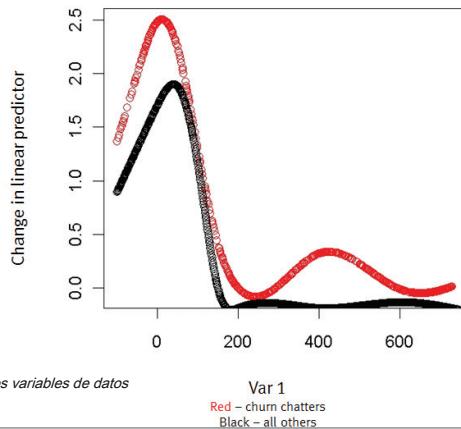


FIGURE 12-14 Detalles del modelo comparando dos variables de datos

#### 12.2.8 Recomendaciones

El componente principal final de la presentación implica la creación de un conjunto de recomendaciones que incluyen cómo implementar el modelo desde una perspectiva empresarial dentro de la organización y cualquier otra sugerencia sobre la implementación de la lógica del modelo. Para el ejemplo de Yoyodyne Bank, la Figura 12-15 proporciona posibles recomendaciones del proyecto. En esta sección de la presentación, medir el impacto de las mejoras y establecer cómo aprovechar ese impacto dentro de las recomendaciones es clave. Por ejemplo, la presentación podría mencionar que cada cliente retenido representa un ahorro de tiempo de seis horas para uno de los gerentes de cuenta del banco o \$ 50,000 en ahorros de adquisiciones de nuevas cuentas, debido a costos de marketing, ventas y costos relacionados con el sistema.

Para una presentación a una audiencia patrocinadora del proyecto, enfóquese en el impacto comercial del proyecto, incluidos los riesgos y el ROI. Debido a que los patrocinadores del proyecto estarán más interesados en el impacto comercial del proyecto, la presentación también debe proporcionar al patrocinador puntos destacados para ayudar a evangelizar el trabajo dentro de la organización. Al preparar una presentación para analistas, complemente el conjunto principal de recomendaciones con las implicaciones para el modelado o para la implementación en un entorno de producción. En cualquier caso, el

el equipo debe fo  
recibirá será

e cliente

## 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, must have in mind the collection of data for investigation into the causes of churn easier**

FIGURE 12-15 Recomendaciones de modelo para el manejo de la rotación de clientes

### 12.2.9 Consejos adicionales sobre la presentación final

A medida que un equipo completa un proyecto y se esfuerza por pasar al siguiente, debe recordar invertir el tiempo adecuado en el desarrollo de las presentaciones finales. Es importante orientar a la audiencia hacia el proyecto y proporcionar contexto. En ocasiones, un equipo está tan inmerso en el proyecto que no proporciona un contexto suficiente para sus recomendaciones y los resultados de los modelos. Un equipo debe recordar deletrear la terminología y los acrónimos y evitar el uso excesivo de jerga. También se debe tener en cuenta que las presentaciones pueden compartirse ampliamente; por lo tanto, es posible que los destinatarios no estén familiarizados con el contexto y el viaje que ha atravesado el equipo a lo largo del proyecto.

Es posible que sea necesario contar la historia varias veces a diferentes audiencias, por lo que el equipo debe ser paciente al repetir algunos de los mensajes clave. Estas presentaciones deben verse como oportunidades para refinar los mensajes clave y evangelizar el buen trabajo que se hizo. En este punto del proceso, el equipo ha invertido muchas horas de trabajo y ha descubierto conocimientos para el negocio. Estas presentaciones son una oportunidad para comunicar estos proyectos y generar apoyo para proyectos futuros. Como ocurre con la mayoría de las presentaciones, es importante evaluar la audiencia para orientar la configuración del mensaje y el nivel de detalle. Aquí hay varios consejos más sobre cómo desarrollar las presentaciones.

- **Utilice imágenes y representaciones visuales:** Las imágenes tienden a hacer la presentación más atractiva. Además, las personas recuerdan las imágenes mejor que las palabras, porque las imágenes pueden tener un impacto más visceral. Estas representaciones visuales pueden ser datos estáticos e interactivos.

- **Asegúrese de que el texto sea mutuamente exclusivo y colectivamente exhaustivo (MECE):** Esto significa tener economía de palabras en la presentación y asegurarse de que los puntos clave estén cubiertos pero no repetidos innecesariamente.
- **Medir y cuantificar los beneficios del proyecto:** Esto puede ser un desafío y requiere tiempo y esfuerzo para ajustarlo. Este tipo de medición debe intentar cuantificar los beneficios que tienen beneficios financieros y de otro tipo de una manera específica. Hacer la afirmación de que un proyecto proporcionó "\$ 8.5 millones en ahorros de costos anuales" es mucho más convincente que decir que tiene un "gran valor".
- **Haga que los beneficios del proyecto sean claros y visibles:** Después de calcular los beneficios del proyecto, asegúrese de articularlos claramente en la presentación.

### 12.2.10 Proporcionar especificaciones técnicas y código

Además de crear las presentaciones finales, el equipo debe entregar el código real que se desarrolló y la documentación técnica necesaria para respaldarlo. El equipo debe considerar cómo el proyecto afectará a los usuarios finales y el personal técnico deberá implementar el código. Se recomienda que el equipo piense en las implicaciones de su trabajo sobre los destinatarios del código, el tipo de preguntas que tendrán y sus intereses. Por ejemplo, indicar que el modelo necesitará realizar un monitoreo en tiempo real puede requerir grandes cambios en un entorno de ejecución de TI, por lo que el equipo puede necesitar considerar un compromiso de trabajos por lotes nocturnos para procesar los datos. Adicionalmente,

El equipo debe anticipar preguntas de TI relacionadas con lo costoso computacionalmente que será ejecutar el modelo en el entorno de producción. Si es posible, indique qué tan bien se ejecutó el modelo en los escenarios de prueba y si hay oportunidades para ajustar el modelo o el entorno para optimizar el rendimiento en el entorno de producción.

Los equipos deben abordar la escritura de documentación técnica para su código como si fuera una interfaz de programación de aplicaciones (API). Muchas veces, los modelos se encapsulan como funciones que leen un conjunto de entradas en el entorno de producción, posiblemente realizan un preprocesamiento de datos y crean una salida, incluido un conjunto de resultados de posprocesamiento.

Considere las entradas, salidas y otras restricciones del sistema para permitir que un técnico implemente el modelo analítico, incluso si esta persona no ha tenido una conexión con el proyecto de ciencia de datos hasta este momento. Piense en la documentación como una forma de presentar los datos que necesita el modelo, la lógica que está utilizando y cómo otros sistemas relacionados necesitan interactuar con ella en un entorno de producción para que funcione correctamente. Las especificaciones detallan las entradas que necesita el código y el formato y las estructuras de los datos. Por ejemplo, puede ser útil especificar si se necesitan datos estructurados o si los datos esperados deben ser formatos numéricos o de cadena. Describa cualquier transformación que deba realizarse en los datos de entrada antes de que el código pueda usarlo y si se creó un script para realizar estas tareas.

Con respecto al manejo de excepciones, el equipo debe considerar cómo el código debe manejar los datos que están fuera de los rangos de datos esperados de los parámetros del modelo y cómo manejará los valores de datos faltantes (Capítulo 3, "Revisión de los métodos analíticos de datos básicos usando R"), nulo valores, ceros, NA o datos en un formato o tipo inesperado. La documentación técnica describe cómo tratar estas excepciones y qué implicaciones pueden surgir en los procesos posteriores. Para los resultados del modelo, el equipo debe explicar hasta qué punto se posprocesa el resultado. Por ejemplo, si el modelo devuelve un valor que representa

la probabilidad de abandono de clientes, puede ser necesaria una lógica adicional para identificar el umbral de puntuación para determinar qué cuentas de clientes marcar como en riesgo de abandono. Además, se deben tomar algunas disposiciones para ajustar este umbral y entrenar el algoritmo, ya sea en forma de aprendizaje automatizado o con intervención humana.

Aunque el equipo debe crear la documentación técnica, muchas veces los ingenieros y otro personal técnico reciben el código y pueden intentar usarlo sin leer toda la documentación. Por lo tanto, es importante agregar comentarios extensos en el código. Esto dirige a las personas que implementan el código sobre cómo usarlo, explica qué partes de la lógica se supone que deben hacer y guía a otras personas a través del código hasta que se familiarizan con él. Si el equipo puede hacer un trabajo completo agregando comentarios en el código, es mucho más fácil para otra persona mantener el código y ajustarlo en el entorno de ejecución. Además, ayuda a los ingenieros a editar el código cuando cambia su entorno o cuando necesitan modificar procesos que pueden proporcionar entradas al código o recibir sus salidas.

## 12.3 Conceptos básicos de visualización de datos

A medida que el volumen de datos continúa aumentando, más proveedores y comunidades están desarrollando herramientas para crear gráficos claros e impactantes para usar en presentaciones y aplicaciones. Aunque no es exhaustivo, la Tabla 12-2 enumera algunas herramientas populares.

**T PODER 12-2** *Herramientas comunes para la visualización de datos*

R (paquete básico, celosía, ggplot2)	Cuadro
GGobi / Rgobi	Fusión paralela (TIBCO)
Gnuplot	QlikView
Inkscape	Ilustrador
ModestMaps	Adobe
OpenLayers	
Procesando	
D3.js	
Tejido	

A medida que ha aumentado el volumen y la complejidad de los datos, los usuarios se han vuelto más dependientes del uso de imágenes nítidas para ilustrar ideas clave y representar datos ricos de una manera sencilla. Con el tiempo, la comunidad de código abierto ha desarrollado muchas bibliotecas para ofrecer más opciones para representar visualmente los datos gráficos. Aunque este libro mostró ejemplos que utilizan principalmente el paquete base de R, ggplot2 proporciona opciones adicionales para crear una visualización de datos de aspecto profesional, al igual que enrejado biblioteca para R.

Gnuplot y GGobi tienen un enfoque basado en la línea de comandos para generar visualización de datos. La génesis de estas herramientas surgió principalmente de la computación científica y la necesidad de expresar visualmente datos complejos. GGobi

también tiene una variante llamada Rggobi que permite a los usuarios acceder a la funcionalidad de GGobi con el software R y el lenguaje de programación. Hay muchas herramientas de creación de mapas de código abierto disponibles, incluidos Modest Maps y OpenLayers, ambos diseñados para desarrolladores que deseen crear mapas interactivos e incrustarlos en sus propios proyectos de desarrollo o en la web. El entorno de desarrollo del lenguaje de programación de software, Processing, emplea un lenguaje similar a Java para que los desarrolladores creen una visualización de datos de aspecto profesional. Debido a que se basa en un lenguaje de programación en lugar de una GUI, Processing permite a los desarrolladores crear una visualización sólida y tener un control preciso sobre la salida. D3.js es una biblioteca de JavaScript para manipular datos y crear visualizaciones basadas en web con estándares, como Hypertext Markup Language (HTML), Gráficos vectoriales escalables (SVG) y hojas de estilo en cascada (CSS). Para obtener más ejemplos del uso de herramientas de visualización de código abierto, consulte el sitio web de Nathan Yau, [flowingdata.com](http://flowingdata.com) [1], o su libro *Visualiza esto* [2], que analiza métodos adicionales para crear representaciones de datos con herramientas de código abierto.

Con respecto a las herramientas comerciales que se muestran en la Tabla 12-2, Tableau, Spotfire (de TIBCO) y QlikView funcionan como herramientas de visualización de datos y como herramientas interactivas de inteligencia empresarial (BI). Debido al crecimiento de los datos en los últimos años, las organizaciones por primera vez están comenzando a poner más énfasis en la facilidad de uso y visualización en BI sobre más herramientas y bases de datos de BI tradicionales. Estas herramientas facilitan la visualización y tienen interfaces de usuario que son más limpias y sencillas de navegar que sus predecesoras. Aunque tradicionalmente no se considera una herramienta de visualización de datos, Adobe Illustrator se enumera en la Tabla 12-2 porque algunos profesionales lo utilizan para mejorar la visualización realizada en otras herramientas. Por ejemplo, algunos usuarios desarrollan una visualización de datos simple en R, guardan la imagen como PDF o JPEG, y luego use una herramienta como Illustrator para mejorar la calidad del gráfico o unir el trabajo de visualización múltiple en una infografía. Inkscape es una herramienta de código abierto que se utiliza para casos de uso similares, con gran parte de la funcionalidad de Illustrator.

### **12.3.1 Puntos clave compatibles con los datos**

Es más difícil observar información clave cuando los datos están en tablas en lugar de en gráficos. Para subrayar este punto, en *Dígalos con gráficos*, Gene Zelazny [3] menciona que para resaltar datos, es mejor crear una representación visual a partir de ellos, como un cuadro, gráfico u otra visualización de datos. Lo opuesto también es cierto. Suponga que un analista opta por restar importancia a los datos. Compartirlo en una mesa atrae menos atención y hace que sea más difícil de digerir para las personas.

La forma en que uno elige organizar lo visual en términos de combinación de colores, etiquetas y secuencia de información también influye en cómo el espectador procesa la información y lo que percibe como el mensaje clave del gráfico. La tabla que se muestra en la Figura 12-16 contiene muchos puntos de datos. Dado el diseño de

FIGURE	ITEM	Quarterly and annual data of exports of goods, services and factor income											
		1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
F	SuperBox	1	1	1	1	5	4	14	15	20	14	17	29
F	BigBox	1	1	1	1	4	5	5	10	10	6	21	33
F	Total,	1	1	1	2	5	5	15	12	19	25	19	27
F	Services	1	1	1	1	4	4	14	15	20	14	17	29
F	Goods	1	1	1	1	5	5	5	10	10	6	21	33
F	Factor Income	1	1	1	1	5	5	5	10	10	6	21	33

Incluso mostrar algo menos de datos sigue siendo difícil de leer para la mayoría de las personas. La figura 12-17 oculta los primeros 10 años, dejando 35 años de datos en la tabla.

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
BigBox	4	5	5	5	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 Siete y cinco años de datos de apertura de tiendas.

Como observarán la mayoría de los lectores, es un desafío dar sentido a los datos, incluso a escalas relativamente pequeñas. Hay varias observaciones en los datos que uno puede notar, si mira de cerca las tablas de datos:

- BigBox experimentó un fuerte crecimiento en las décadas de 1980 y 1990.
- En la década de 1980, BigBox comenzó a agregar más tiendas SuperBox a su mezcla de cadenas de tiendas.
- Las tiendas SuperBox superan en número a las tiendas BigBox en casi 2 a 1 en total.

Dependiendo del punto que se intenta hacer, el analista debe tener cuidado de organizar la información de una manera que permita intuitivamente al espectador extraer el mismo punto principal que pretendía el autor. Si el analista falla en hacer esto de manera efectiva, la persona que consume los datos debe adivinar el punto principal y puede interpretar algo diferente de lo que se pretendía.

La Figura 12-18 muestra un mapa de los Estados Unidos, con los puntos que representan las ubicaciones geográficas de las tiendas. Este mapa es una forma más poderosa de representar datos que lo que sería una tabla pequeña. El enfoque se adapta bien a una audiencia patrocinadora. Este mapa muestra dónde la tienda BigBox tiene saturación de mercado, dónde ha crecido la empresa y dónde tiene tiendas SuperBox y otras tiendas BigBox, según el color y el sombreado. La visualización en la Figura 12-18 se comunica claramente de manera más efectiva que las tablas densas en la Figura 12-

ization tec

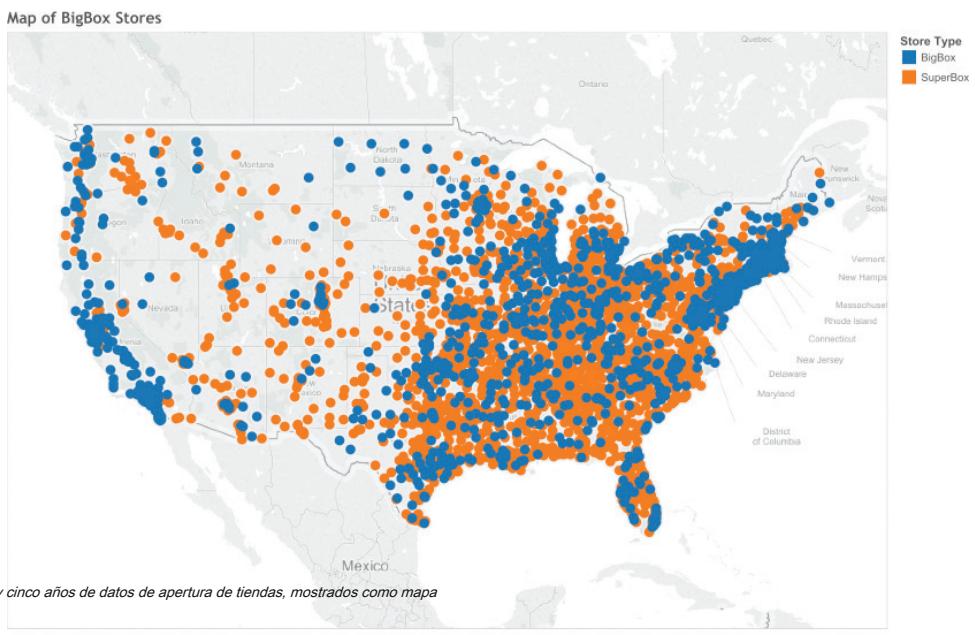


FIGURE 12-18 Cuarenta y cinco años de datos de apertura de tiendas, mostrados como mapa

Map based on Longitude (generated) and Latitude (generated). Color shows details about Store Type. Details are shown for ZIP.

### 12.3.2 Evolución de un gráfico

La visualización permite a las personas representar los datos de una manera más convincente que las tablas de datos y de una manera que se puede entender en un nivel intuitivo y precognitivo. Además, los analistas y científicos de datos pueden utilizar la visualización para interactuar con los datos y explorarlos. A continuación se muestra un ejemplo de los pasos que puede seguir un científico de datos al explorar los datos de precios para comprender mejor los datos, modelarlos y evaluar si un precio actual

ción de los datos de precios como una puntuación de usuario

reflejando

*Distribution of User Score*

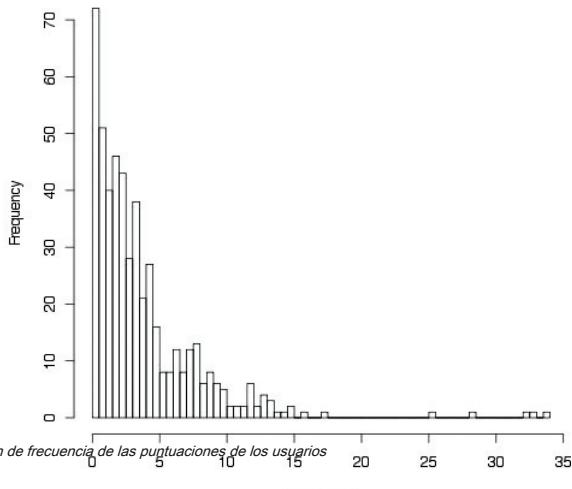


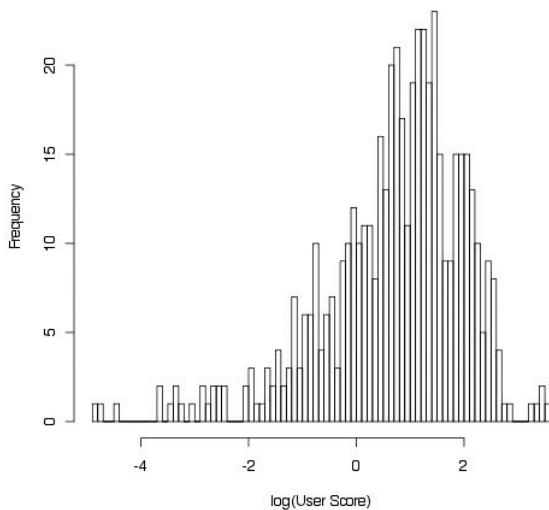
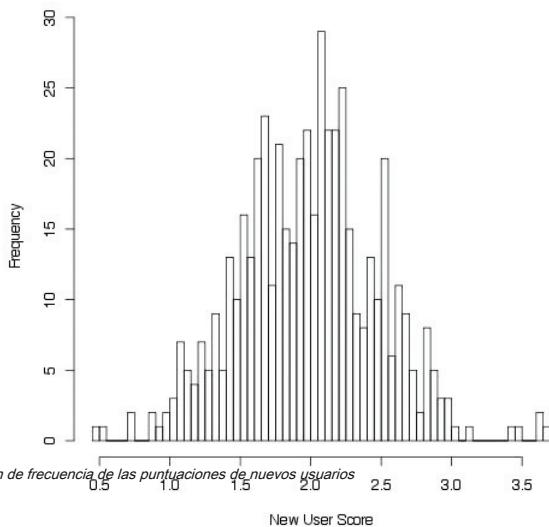
FIGURE 12-19 Distribución de frecuencia de las puntuaciones de los usuarios

El primer paso de un científico de datos puede ser ver los datos como una distribución sin procesar de los niveles de precios de los usuarios.

Debido a que los valores tienen una cola larga a la derecha, en la Figura 12-19, puede ser difícil tener una idea de cuán estrechamente agrupados están los datos entre las puntuaciones de usuario de cero y cinco.

Para comprender esto mejor, un científico de datos puede volver a ejecutar esta distribución mostrando una distribución de registro (Capítulo 3) de la puntuación del usuario, como se muestra en la Figura 12-20.

Esto muestra una distribución menos sesgada que puede ser más fácil de entender para un científico de datos. La Figura 12-21 ilustra una vista reescalada de la Figura 12-20, con la mediana de la distribución alrededor de 2.0. Este gráfico proporciona la distribución de una nueva puntuación de usuario, o índice, que puede medir el nivel de sensibilidad al precio de un usuario cuando se expresa en forma de registro.

*Log Distribution of User Score*FIGURE 12-20 *Fre**Distribution of New User Score*FIGURE 12-21 *Distribución de frecuencia de las puntuaciones de nuevos usuarios*

Otra idea puede ser analizar la estabilidad de las distribuciones de precios a lo largo del tiempo para ver si los precios ofrecidos a los clientes son estables o volátiles. Como se muestra en un gráfico como el de la Figura 12-22, los precios parecen estables. En este ejemplo, la puntuación de precio del usuario permanece dentro de una banda estrecha entre dos y tres independientemente del tiempo

persigue un producto dado no

significativo

La puntuación del usuario, que se muestra en el eje y.

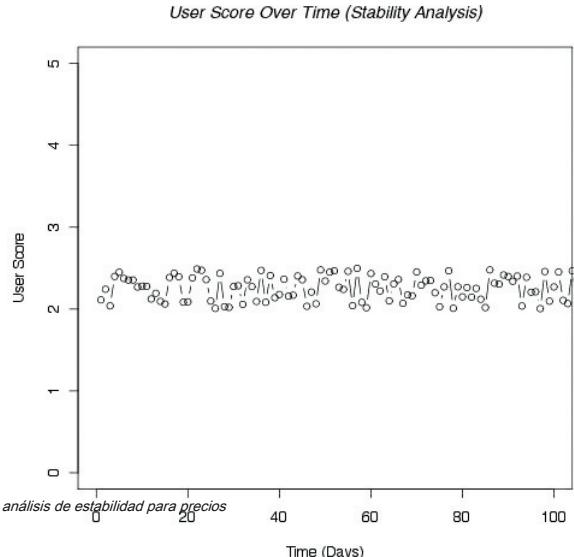


FIGURE 12-22 Gráfico de análisis de estabilidad para precios

En este punto, el científico de datos ha aprendido lo siguiente sobre este ejemplo y ha realizado varias observaciones sobre los datos:

- La mayoría de las puntuaciones de los usuarios están entre dos y tres en términos de sensibilidad al precio.
- Después de tomar el valor de registro de las puntuaciones de los usuarios, se creó un nuevo índice de puntuación de usuarios, que volvió a centrar los valores de datos alrededor del centro de la distribución.
- Las puntuaciones de precios parecen ser estables en el tiempo, ya que la duración del cliente no parece tener una influencia significativa en la puntuación de precios del usuario. En cambio, parece ser relativamente constante en el tiempo, dentro de una pequeña banda de puntajes de usuario.

En este punto, los analistas pueden querer explorar la gama de niveles de precios que se ofrecen a los clientes. Las figuras 12-22 y 12-23 muestran ejemplos de la clasificación de precios actualmente en vigor dentro de la base de clientes.

La figura 12-23 muestra la distribución de precios para una base de clientes. En este ejemplo, el puntaje de lealtad y el precio están correlacionados positivamente; a medida que aumenta el puntaje de lealtad, también aumentan los precios que los clientes están dispuestos a pagar. Puede parecer un fenómeno extraño que los clientes más leales en este ejemplo estén dispuestos a pagar precios más altos, pero la realidad es que los clientes que son muy leales tienden a ser menos sensibles a las fluctuaciones o aumentos de precios. La clave, sin embargo, es comprender qué clientes son muy leales para poder cobrar los precios adecuados a los grupos de personas adecuados.

La figura 12-24 muestra una variación de 12-23. En este caso, el nuevo gráfico muestra los mismos niveles de precios del cliente, pero la parte inferior para reflejar la distribución ciación de la

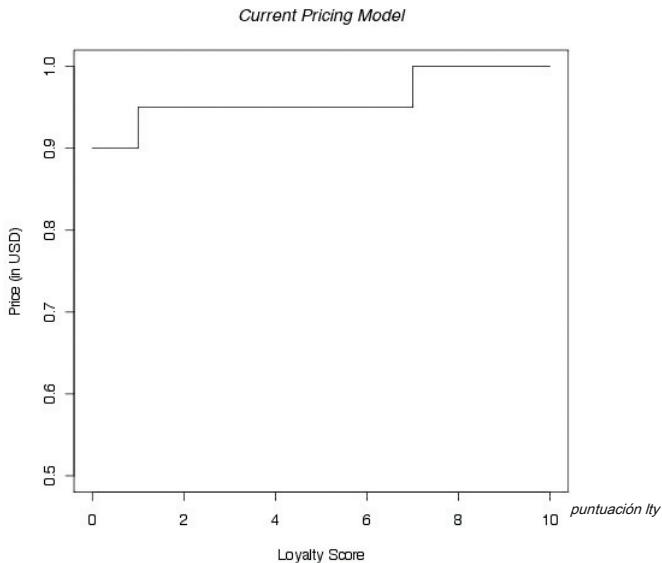


FIGURE 12-23 Gr

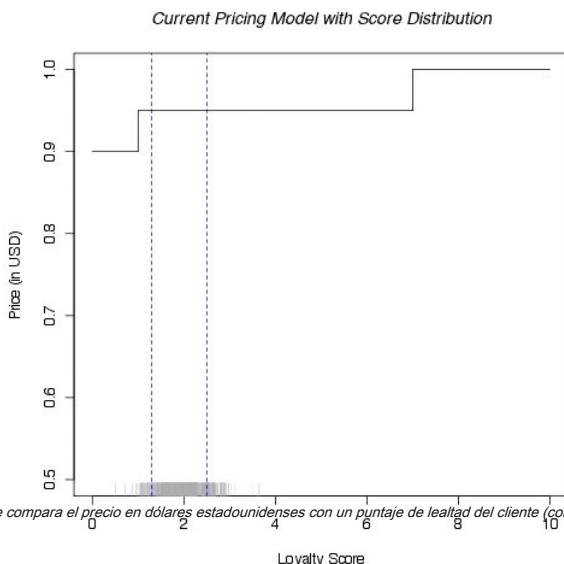


FIGURE 12-24 Gráfico que compara el precio en dólares estadounidenses con un puntaje de lealtad del cliente (con representación de alfombra)

Esta alfombra indica que la mayoría de los clientes en este ejemplo se encuentran en una banda estrecha de puntuaciones de lealtad, entre aproximadamente 1 y 3 en el eje x, todos los cuales ofrecen el mismo conjunto de precios, que son altos (entre 0.9 y 1.0 en el eje y). El eje y en este ejemplo puede representar una puntuación de precio o el valor bruto de un cliente en millones de dólares. El aspecto importante es reconocer que el precio es alto y se ofrece de manera consistente a la mayoría de los clientes en este ejemplo.

### *Proposed Pricing Model with Score Distribution*

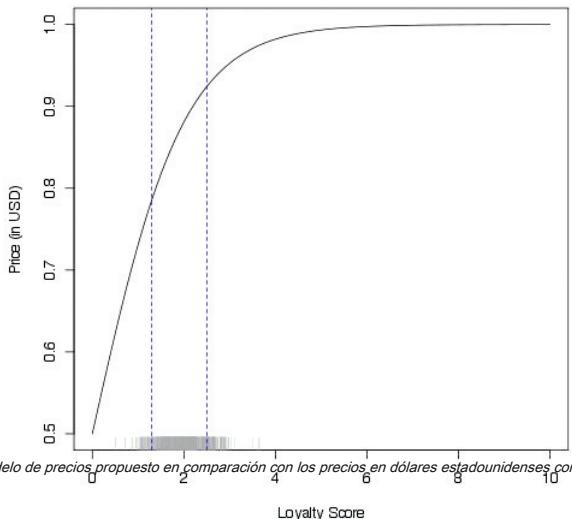
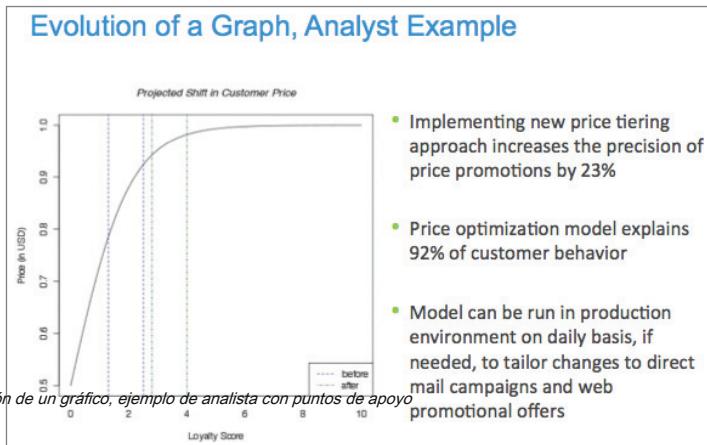


FIGURE 12-25 Nuevo modelo de precios propuesto en comparación con los precios en dólares estadounidenses con alfombra.

Los científicos de datos normalmente iteran y ven los datos de muchas formas diferentes, enmarcando hipótesis, probándolas y explorando las implicaciones de un modelo dado. Este caso explora ejemplos visuales de distribuciones de precios, fluctuaciones en los precios y las diferencias en los niveles de precios antes y después de implementar un nuevo modelo para optimizar el precio. El trabajo de visualización ilustra cómo pueden verse los datos como resultado del modelo y ayuda a un científico de datos a comprender las relaciones dentro de los datos de un vistazo.

El gráfico resultante en el escenario de precios parece ser técnico con respecto a la distribución de precios en una base de clientes y sería adecuado para una audiencia técnica compuesta por otros científicos de datos. La figura 12-26 muestra un ejemplo de cómo se puede presentar este gráfico a una audiencia de otros científicos de datos o analistas de datos. Esto demuestra una relación curvilinear entre los niveles de precios y la lealtad del cliente cuando se expresa como un índice. Tenga en cuenta que los comentarios a la derecha del gráfico se relacionan con la precisión de la focalización de precios, la cantidad de variabilidad en la robustez del modelo y las expectativas del modelo específico.



La Figura 12-27 muestra otro ejemplo del resultado del escenario del proyecto de optimización de precios, mostrando cómo se puede presentar esto a una audiencia de patrocinadores del proyecto. Esto demuestra un gráfico de barras simple que representa el precio promedio por segmento de cliente o usuario. La figura 12-27 muestra una imagen de aspecto mucho más simple que la figura 12-26. Muestra claramente que los clientes con puntajes de lealtad más bajos tienden a obtener precios más bajos debido a la orientación de las promociones de precios. Tenga en cuenta que el lado derecho de la imagen se centra en el impacto comercial y el ahorro de costos en lugar de las características detalladas del modelo.

## Evolution of a Graph, Sponsor Example

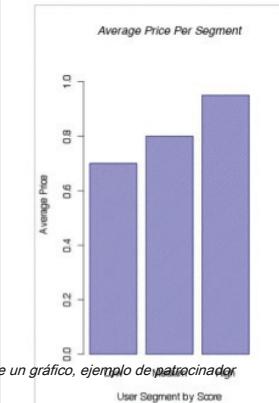


FIGURE 12-27 Evolución de un gráfico, ejemplo de patrocinador

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

Los comentarios al lado derecho del gráfico en la Figura 12-27 explican el impacto del modelo a un alto nivel y el ahorro de costos de implementar este enfoque para la optimización de precios.

### 12.3.3 Métodos de representación común

Aunque existen muchos tipos de visualizaciones de datos, varios tipos fundamentales de gráficos representan datos e información. Es importante saber cuándo utilizar un tipo particular de cuadro o gráfico para expresar un tipo de datos determinado. La tabla 12-3 muestra algunos tipos de gráficos básicos para guiar al lector a comprender que los diferentes tipos de gráficos se adaptan mejor a una situación según los tipos específicos de datos y el mensaje que el equipo está intentando transmitir. El uso de un tipo de gráfico para los datos para el que no está diseñado puede parecer interesante o inusual, pero generalmente confunde al espectador. El objetivo del autor es encontrar el mejor gráfico para expresar los datos con claridad, de modo que lo visual no obstaculice el mensaje, sino que ayude al lector a eliminar el mensaje deseado.

#### T PODER 12-3 Métodos comunes de representación para datos y gráficos

Componentes (partes de un todo)	Gráfico circular
Artículo	Gráfico de barras
Serie de tiempo	Gráfico de líneas
Frecuencia	Gráfico de líneas o histograma
Correlación	Diagrama de dispersión, gráficos de barras uno al lado del otro
La tabla 12-3 muestra las representaciones de datos más fundamentales y comunes, que se pueden combinar, embellecer y hacer más sofisticadas según la situación y la audiencia. Es recomendado	

que el equipo considere el mensaje que está tratando de comunicar y luego seleccione el tipo de visual apropiado para apoyar el punto. El uso indebido de gráficos tiende a confundir a la audiencia, por lo que es importante tener en cuenta el tipo de datos y el mensaje deseado al elegir un gráfico.

Los gráficos circulares están diseñados para mostrar los componentes o partes relativas a un conjunto completo de cosas. Un gráfico circular es también el tipo de gráfico que se utiliza con más frecuencia. Si la situación requiere el uso de un gráfico circular, úselo solo cuando muestre solo 2 o 3 elementos en un gráfico, y solo para las audiencias de los patrocinadores.

Los gráficos de barras y los gráficos de líneas se utilizan con mucha más frecuencia y son útiles para mostrar comparaciones y tendencias a lo largo del tiempo. Aunque la gente usa gráficos de barras verticales con más frecuencia, los gráficos de barras horizontales permiten al autor más espacio para ajustarse a las etiquetas de texto. Los gráficos de barras verticales tienden a funcionar bien cuando las etiquetas son pequeñas, como cuando se muestran comparaciones a lo largo del tiempo utilizando años.

Para la frecuencia, los histogramas son útiles para demostrar la distribución de datos a una audiencia de analistas o científicos de datos. Como se muestra en el ejemplo de precios anteriormente en este capítulo, las distribuciones de datos suelen ser uno de los primeros pasos cuando se visualizan datos para preparar la planificación del modelo. Para evaluar cualitativamente las correlaciones, los diagramas de dispersión pueden ser útiles para comparar relaciones entre variables.

Como con cualquier presentación, considere la audiencia y el nivel de sofisticación al seleccionar el gráfico para transmitir el mensaje deseado. Estos gráficos son ejemplos simples, pero pueden volverse más complejos fácilmente cuando se agregan variables de datos, se combinan gráficos o se agrega animación cuando sea apropiado.

### 12.3.4 Cómo limpiar un gráfico

Muchas veces, los paquetes de software generan un gráfico para un conjunto de datos, pero el software agrega demasiadas cosas al gráfico. Estas distracciones visuales agregadas pueden hacer que la imagen parezca ocupada u oscurecer los puntos principales que se deben hacer con el gráfico. En general, es una buena práctica esforzarse por la simplicidad al crear gráficos y gráficos de visualización de datos. Saber cómo simplificar gráficos o limpiar un cha desordenado

mi  
gráfico con

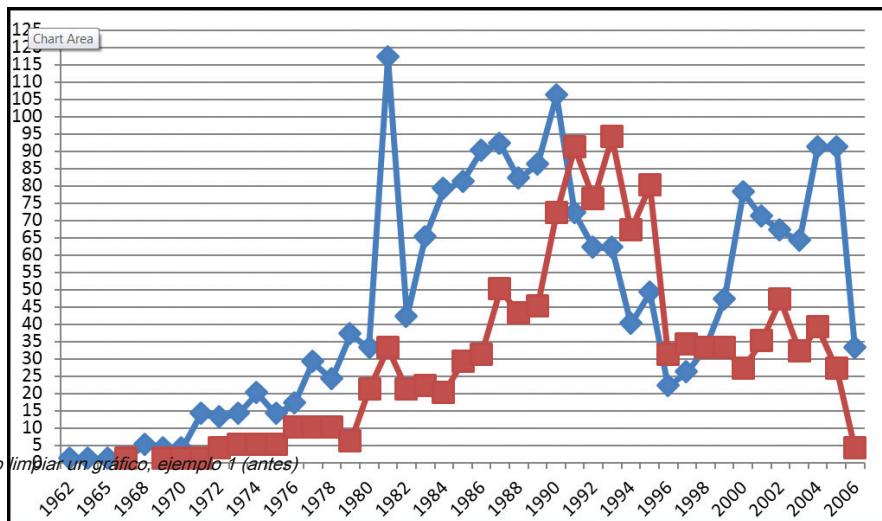


FIGURE 12-28 Cómo limpiar un gráfico, ejemplo 1 (antes)

## Cómo limpiar un gráfico

El gráfico de líneas que se muestra en la Figura 12-28 compara dos tendencias a lo largo del tiempo. El gráfico parece ocupado y contiene una gran cantidad de basura que distrae al espectador del mensaje principal. *Gráfico basura* se refiere a elementos de visualización de datos que proporcionan materiales adicionales pero que no contribuyen a la parte de datos del gráfico. Si se eliminara la basura de los gráficos, el significado y la comprensión del gráfico no disminuirían; en cambio, se aclararía más. Hay cinco tipos principales de "basura de gráficos" en la Figura 12-28:

- **Líneas de cuadrícula horizontales:** Estos no tienen ningún propósito en este gráfico. No proporcionan información adicional para la tabla.
- **Puntos de datos gruesos:** Estos puntos de datos representados como grandes bloques cuadrados atraen la atención del espectador hacia ellos, pero no representan ningún significado específico aparte de los puntos de datos en sí.
- **Uso excesivo de colores de énfasis en las líneas y el borde:** El borde del gráfico es una línea gruesa y en negrita. Esto fuerza la atención del espectador hacia el perímetro del gráfico, que no contiene ningún valor de información. Además, las líneas que muestran las tendencias son relativamente gruesas.
- **Sin contexto ni etiquetas:** El gráfico no contiene ninguna leyenda que proporcione contexto a lo que se muestra. Las líneas también carecen de etiquetas para explicar lo que representan.
- **Etiquetas de eje abarrotadas:** Hay demasiadas etiquetas de eje, por lo que parecen abarrotadas. No es necesario que las etiquetas en el eje y aparezcan cada cinco unidades o que los valores en el eje x aparezcan cada dos unidades. Mostrado de esta manera, las etiquetas de los ejes distraen al espectador de los datos reales que están representados por las líneas de tendencia en el gráfico.

Las cinco formas de basura en los gráficos de la Figura 12-28 se corrigen fácilmente, como se muestra en la Figura 12-29. Tenga en cuenta que no hay en la Figura 1 lo que se muestra

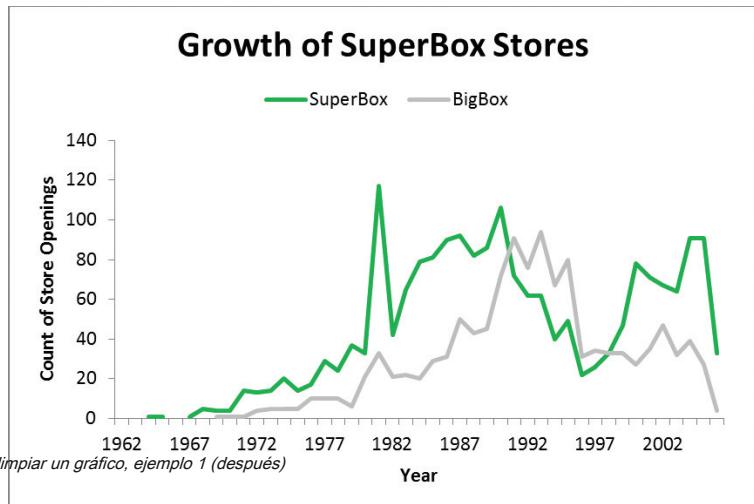
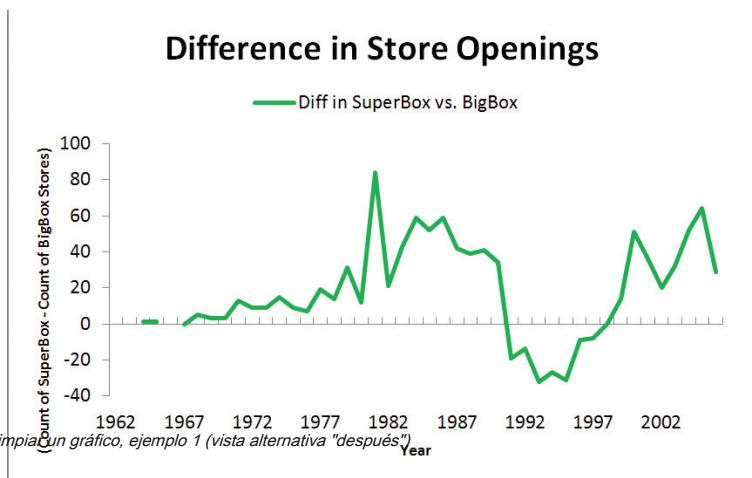


FIGURE 12-29 Cómo limpiar un gráfico, ejemplo 1 (después)



**FIGURE 12-30** Cómo limpian un gráfico, ejemplo 1 (vista alternativa "después")

Las Figuras 12-29 y 12-30 muestran dos ejemplos de versiones limpias del cuadro que se muestra en la Figura 12-28. Tenga en cuenta que se han solucionado los problemas con la basura de gráficos. Hay una etiqueta y un título claros para cada gráfico para reforzar el mensaje, y el color se ha utilizado siempre para resaltar el punto que el autor está tratando de expresar. En la Figura 12-29, se muestra un color verde fuerte para representar el recuento de tiendas SuperBox, porque aquí es donde debe dibujarse el foco del espectador, mientras que el recuento de tiendas BigBox se muestra en un color gris claro.

Además, observe la cantidad de espacio en blanco que se utiliza en cada uno de los dos gráficos que se muestran en las Figuras 12-29 y 12-30. Eliminar líneas de cuadrícula, ejes excesivos y el ruido visual dentro del gráfico permite un contraste claro entre los colores de énfasis (los gráficos de líneas verdes) y los colores estándar (el gris más claro de las tiendas BigBox). Al crear gráficos, es mejor dibujar la mayoría de las imágenes principales en colores estándar, tonos claros o matices de color para que los colores de énfasis más fuertes puedan resaltar los puntos principales. En este caso, la tendencia de las tiendas BigBox en gris claro se desvanece en el fondo pero no desaparece, mientras que la tendencia de las tiendas SuperBox en un gris más oscuro (verde brillante en el gráfico en línea) lo hace prominente para respaldar el mensaje que el autor está haciendo sobre el crecimiento de las tiendas SuperBox.

En la Figura 12-30 se muestra una alternativa a la Figura 12-29. Si el mensaje principal es mostrar la diferencia en el crecimiento de nuevas tiendas, se puede crear la Figura 12-30 para simplificar aún más la Figura 12-28 y graficar solo la diferencia entre las tiendas SuperBox en comparación con las tiendas BigBox normales. Se muestran dos ejemplos para ilustrar diferentes formas de transmitir el mensaje, dependiendo de lo que el autor de estos cuadros quisiera enfatizar.

### Cómo limpiar un gráfico, segundo ejemplo

Otro ejemplo de limpieza de un gráfico se muestra en la Figura 12-31. Este gráfico de barras verticales sufre más de los problemas típicos relacionados con la basura de los gráficos, incluido el uso inadecuado de los esquemas de color y la falta de contexto.

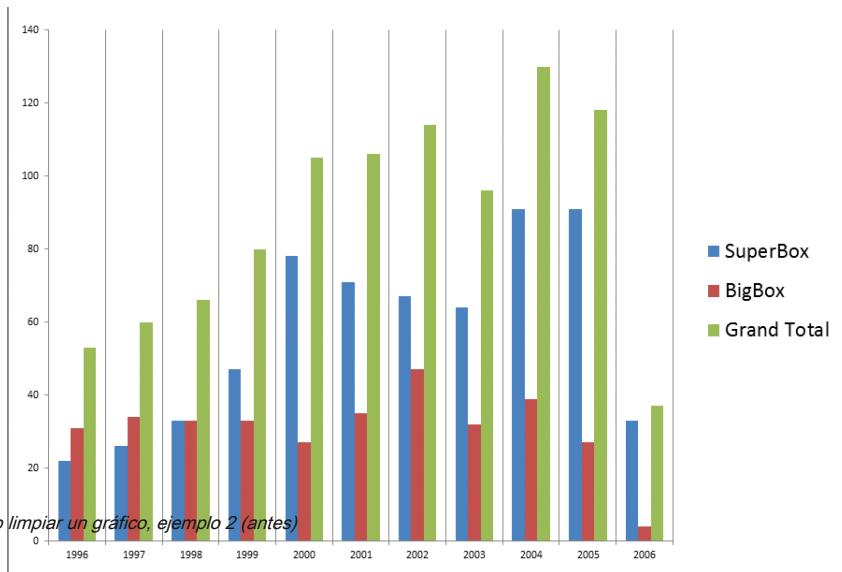


FIGURE 12-31 Cómo limpiar un gráfico, ejemplo 2 (antes)

Hay cinco tipos principales de basura de gráficos en la Figura 12-31:

- **Líneas de cuadrícula verticales:** Estas líneas de cuadricula verticales no son necesarias en este gráfico. No proporcionan información adicional para ayudar al espectador a comprender el mensaje de los datos. En cambio, estas líneas de cuadricula verticales solo distraen al espectador de mirar los datos.
- **Color de énfasis demasiado:** Este gráfico de barras utiliza colores fuertes y una escala de grises oscura de contraste demasiado alto. En general, es mejor utilizar tonos sutiles, con un gris de bajo contraste como color neutro, y luego enfatizar los datos subrayando el mensaje clave en un tono oscuro o un color fuerte.
- **Sin título de gráfico:** Debido a que el gráfico carece de un título de gráfico, el espectador no está orientado hacia lo que está viendo y no tiene el contexto adecuado.
- **Legenda a la derecha que restringe el espacio del gráfico:** Aunque hay una leyenda para el gráfico, se muestra en el lado derecho, lo que hace que el gráfico de barras verticales se comprima horizontalmente. La leyenda tendría más sentido colocada en la parte superior, encima del gráfico, donde no interferiría con los datos que se expresan.
- **Etiquetas pequeñas:** Las etiquetas de los ejes horizontal y vertical tienen el espacio adecuado, pero el tamaño de la fuente es demasiado pequeño para leerlo fácilmente. Deben ser un poco más grandes para que se puedan leer fácilmente, sin que parezcan demasiado prominentes.

Las figuras 12-32 y 12-33 muestran dos ejemplos de versiones limpias del cuadro que se muestra en la figura 12-31. Se han abordado los problemas con la basura de gráficos. Hay una etiqueta y un título claros para cada gráfico para reforzar el mensaje, y se han utilizado colores apropiados para resaltar el punto que el autor está tratando de expresar. Las figuras 12-32 y 12-33 muestran dos opciones para modificar el gráfico, dependiendo del punto principal que el presentador esté tratando de hacer.

La Figura 12-32 muestra un color de fuerte énfasis (azul oscuro) que representa las tiendas SuperBox para respaldar el título del gráfico: Crecimiento de las tiendas SuperBox.

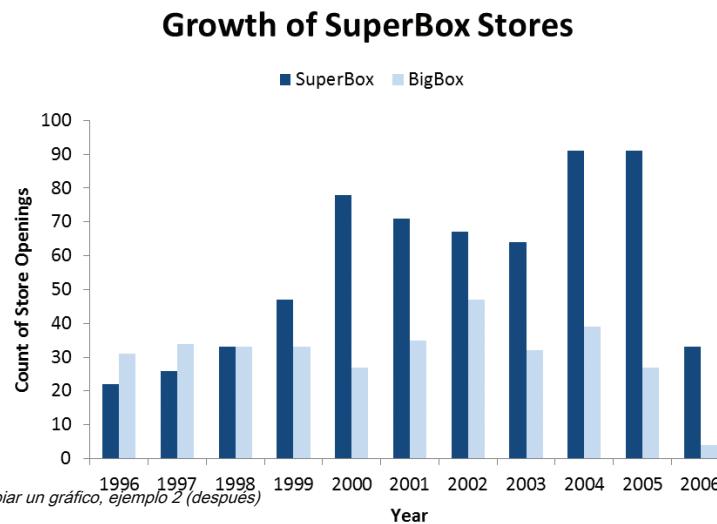


FIGURE 12-32 Cómo limpiar un gráfico, ejemplo 2 (después)

Supongamos  
mostrando th

lugar. Un gráfico de líneas

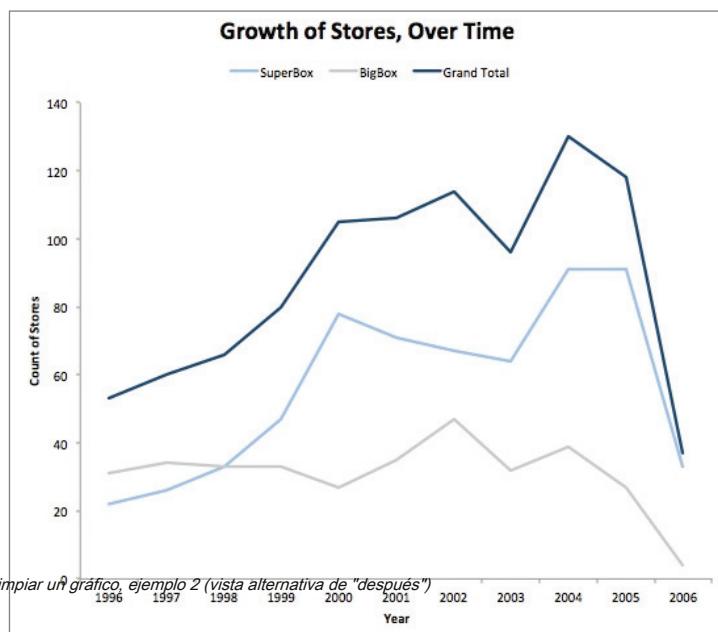


FIGURE 12-33 Cómo limpiar un gráfico, ejemplo 2 (vista alternativa de "después")

En ambos casos, se han eliminado el ruido y las distracciones dentro del gráfico. Como resultado, se ha restado importancia a los datos del gráfico de barras para proporcionar contexto, mientras que otros datos se han hecho más prominentes porque refuerzan el punto clave como se indica en el título del gráfico.

### 12.3.5 Consideraciones adicionales

Como se indicó en los ejemplos anteriores, el énfasis debe estar en la simplicidad al crear cuadros y gráficos. Cree gráficos libres de basura en los gráficos y utilice el método más simple para representar gráficos con claridad. El objetivo de la visualización de datos debe ser respaldar la transmisión de mensajes clave de la manera más clara posible y con pocas distracciones.

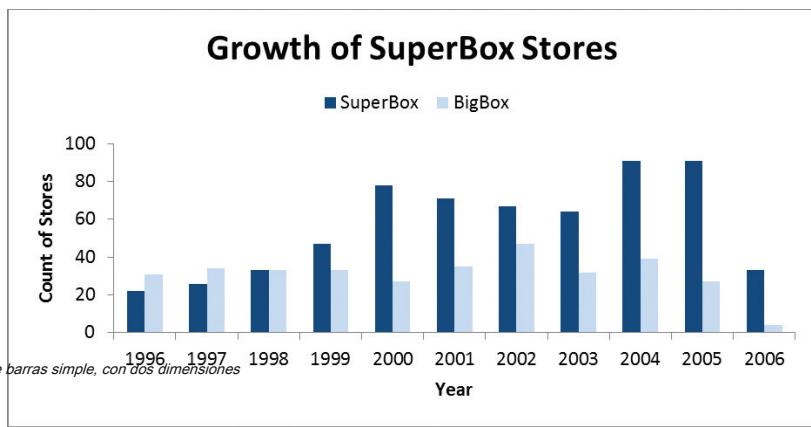
Al igual que la idea de eliminar la basura de los gráficos, es consciente de la relación datos-tinta. *Tinta de datos* se refiere a la parte real de un gráfico que representa los datos, mientras que *tinta sin datos* se refiere a etiquetas, bordes, colores y otra decoración. Si uno imagina la tinta necesaria para imprimir una visualización de datos en papel, la relación tinta-datos podría pensarse como  $(\text{tinta de datos}) / (\text{tinta total utilizada para imprimir el gráfico})$ . En otras palabras, cuanto mayor es la proporción de tinta de datos en lo visual, más ricos son los datos y menos distracciones tiene [4].

#### **Evite el uso de tres dimensiones en la mayoría de los gráficos**

Un ejemplo más en el que las personas suelen equivocarse es agregar sombras, profundidad o dimensiones innecesarias a los gráficos. La figura 12-34 muestra un gráfico de barras verticales con dos dimensiones visibles. Este ejemplo es simple y fácil de entender, y la atención se centra en los datos, no en los gráficos. El autor de la tabla ha optado por resaltar las tiendas SuperBox en un color azul oscuro, mientras que las barras de BigBox en la tabla están en un azul más claro. El título i

k, alto

contraste s



Compare la Figura 12-34 con la Figura 12-35, que muestra un gráfico tridimensional. La figura 12-35 muestra el gráfico de barras original en ángulo, con algún intento de mostrar la profundidad. Este tipo de perspectiva tridimensional hace más difícil para el espectador medir los datos reales y la escala se vuelve engañosa.

Tres tenues  
sión para de

ddimen-

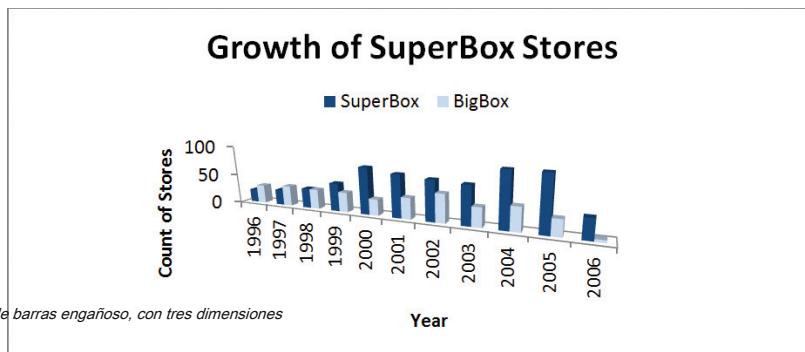


FIGURE 12-35 Gráfico de barras engañoso, con tres dimensiones

Los gráficos de las figuras 12-34 y 12-35 muestran los mismos datos, pero es más difícil juzgar la altura real de las barras de la figura 12-35. Además, el sombreado y la forma del gráfico hacen que la mayoría de los espectadores dediquen tiempo a mirar la perspectiva del gráfico en lugar de la altura de las barras, que es el mensaje clave y el propósito de esta visualización de datos.

## Resumen

Comunicar el valor de los proyectos analíticos es fundamental para mantener el impulso de un proyecto y generar apoyo dentro de las organizaciones. Este apoyo es fundamental para convertir un proyecto exitoso en un sistema o integrarlo correctamente en un entorno de producción existente. Debido a que un proyecto de análisis puede necesitar ser comunicado a audiencias con antecedentes mixtos, este capítulo recomienda la creación de cuatro entregables para satisfacer la mayoría de las necesidades de varias partes interesadas.

- Presentación para patrocinador de un proyecto
- Una presentación para una audiencia analítica
- Documentos de especificación técnica
- Código de producción bien anotado

La creación de estos entregables permite al equipo del proyecto de análisis comunicar y evangelizar el trabajo que realizó, mientras que el código y la documentación técnica ayudan al equipo que desea implementar los modelos dentro del entorno de producción.

Este capítulo ilustra la importancia de seleccionar representaciones visuales claras y simples para apoyar los puntos clave en las presentaciones finales o para retratar los datos. La mayoría de las representaciones de datos y gráficos se pueden mejorar simplemente eliminando las distracciones visuales. Esto significa minimizar o eliminar la basura de los gráficos, que distrae al espectador del propósito principal de un gráfico y no agrega valor de información.

Seguir varios principios de sentido común sobre cómo minimizar las distracciones en las diapositivas y visualizaciones, comunicarse de forma clara y sencilla, usar el color de forma deliberada y tomarse el tiempo para proporcionar contexto aborda la mayoría de los problemas comunes en gráficos y diapositivas. Estas pocas pautas apoyan la creación de imágenes claras y nítidas que transmiten los mensajes clave.

En la mayoría de los casos, las mejores visualizaciones de datos utilizan la imagen más simple y clara para ilustrar el punto clave. Evite los adornos innecesarios y concéntrese en tratar de encontrar el método mejor y más simple para transmitir el mensaje. El contexto es fundamental para orientar al espectador hacia un cuadro o gráfico, porque las personas tienen reacciones inmediatas a las imágenes en un nivel precognitivo. Con este fin, asegúrese de emplear el color con cuidado y oriente al espectador con escalas, leyendas y ejes.

## Ejercicios

1. Describe cuatro entregables comunes para un proyecto de análisis.
2. ¿Cuál es el enfoque de una presentación para un patrocinador de proyecto?
3. Dar ejemplos de gráficos apropiados para crear en una presentación para otros analistas de datos y científicos de datos. tistas como parte de una presentación final. Explique por qué los gráficos son apropiados para mostrar a cada audiencia.
4. Explique qué tipos de gráficos serían apropiados para mostrar datos que cambian con el tiempo y por qué.
5. Como parte de la puesta en funcionamiento de un proyecto de análisis, ¿qué entregable esperaría proporcionar a un Analista de Business Intelligence?

## Referencias y lecturas adicionales

A continuación, encontrará referencias adicionales para obtener más información sobre las mejores prácticas para realizar presentaciones.

- **Diga ItwithCharts, por Gene Zelazny [3]:** Libro de referencia simple sobre cómo seleccionar el enfoque gráfico correcto para representar datos y garantizar que el mensaje se transmita claramente en las presentaciones.
- **Principio de la pirámide, por BarbaraMinto [5]:** Minto fue pionero en el enfoque para construir estructuras lógicas para presentaciones en grupos de tres: tres secciones para las presentaciones, cada una con tres puntos principales. Esto le enseña a la gente cómo tejer una historia a partir de piezas dispares.
- **PresentationZen, por Garr Reynolds [6]:** Enseña cómo transmitir ideas de forma simple y clara y cómo utilizar imágenes en presentaciones. Muestra muchas versiones de gráficos y diapositivas antes y después.
- **Ahora lo ves, por Stephen Few [4]:** Proporciona muchos ejemplos que dan formato al tipo apropiado de visualización de datos para un conjunto de datos determinado.

## Bibliografía

- [1] N. Yau, "flowingdata.com" [en línea]. Disponible: <http://flowingdata.com>.
- [2] N. Yau, *Visualiza esto*, Indianápolis: Wiley, 2011.
- [3] G. Zelazny, *Dígalos con gráficos: la guía ejecutiva de comunicación visual*, McGraw-Hill, 2001.
- [4] S. Pocos, *Now You See It: Técnicas de visualización simples para análisis cuantitativo*, Prensa analítica, 2009.
- [5] B. Minto, *El principio de la pirámide de Minto: lógica en la escritura, el pensamiento y la resolución de problemas*, Aprendiz Hall, 2010.
- [6] G. Reynolds, *Presentation Zen: Ideas simples sobre el diseño y la entrega de presentaciones*, Berkeley: nuevo Jinetes, 2011.



# *Índice*

## Números y símbolos

\ (barra inclinada hacia atrás) como separador, 69  
 / (barra inclinada) como separador, 69 conjuntos de 1 elementos, 147  
 Conjuntos de 2 elementos, 148–149  
 3 Vs (volumen, variedad, velocidad), 2-3 conjuntos de 3 elementos, 149–150  
 Conjuntos de 4 elementos, 150–151

## UN

precisión, 225  
 ACF (función de autocorrelación), 236–237 Ejemplo de análisis de texto ACME, 259–260

colección de texto sin formato, 260–263

agregados (SQL)

pedido, 351–352

de fi nido por el usuario, 347–351

agregadores de datos, 18

AIE (Economía de la información aplicada), 28

algoritmos

agrupación, 134–135

árboles de decisión, 197–200

C4.5, 203–204

CARRITO, 204

ID3, 203

Minero Alphine, 42

hipótesis alternativa, 102–103

proyectos analíticos

Enfoque, 369–371

Analista de BI, 362

usuarios comerciales, 361

código, 362, 376–377

comunicación, 360–361

ingeniero de datos, 362

científicos de datos, 362

DBA (Administrador de base de datos), 362

entregables, 362–364

audiencias, 364–365

corematerial, 364–365

puntos clave, 372

Hallazgos principales, 367–369

descripción del modelo, 371

detalles del modelo, 372–374

operacionalizando, 360–361

salidas, 361

presentaciones, 362

Metas del proyecto, 365–367

director de proyectos, 362

patrocinador del proyecto, 361

recomendaciones, 374–375

partes interesadas, 361–362

especializaciones técnicas, 376–377

cajas de arena analíticas. Ver Arquitectura

analítica de sandboxes, 13–15

analítica

impulsores de negocios, 11

ejemplos, 22–23

nuevos enfoques, 16–19

ANOVA, 110–114

Cuarteto de Anscombe, 82–83

aov () función 78 Apache Hadoop. Ver

Hadoop

API (interfaces de programación de aplicaciones), Hadoop, 304–305

a priori () función, 146, 152–157 Algoritmo a

priori, 139

ejemplo de tienda de comestibles, 143

Comestibles conjunto de datos, 144–146

generación de conjuntos de elementos, 146–151

generación de reglas, 152–157

conjuntos de elementos, 139, 140–141

contando, 158

particionamiento y 158

muestreo y, 158

reducción de transacciones y 158

arquitectura, analítica, 13–15

arima () función, 246

Modelo ARIMA (media móvil integrada autorregresiva),

236

ACF, 236–237

ARMAmodel, 241–244

modelos autorregresivos, 238–239

edificio, 244–252

precauciones, 252–253

varianza constante, 250–251

evaluando, 244–252

modelos de series de tiempo ajustados, pronóstico

249–250, 251–252

modelos de media móvil, 239–241

normalidad, 250–251

PACF, 238–239

razones para elegir, 252–253

media móvil integrada autorregresiva estacional

modelo, 243–244

VARIMA, 253

Modelo ARMA (media móvil autorregresiva), 241–244

matriz () función, 74 matrices

matrices, 74

R, 74–75

reglas de asociación, 138–139

aplicación, 143

reglas candidatas, 141–142

diagnóstico, 158

- pruebas y 157-158  
validación, 157-158
- atributos**  
objetos, k-medias, 130-131  
R, 71-72
- AUC (área bajo la curva), 227 modelos  
autorregresivos, 238-239  
promedios, modelos de promedio móvil, 239-241
- segundo**
- ensacado, 228  
bolsa de palabras en el análisis de texto, 265-266  
**banca, 18**  
gráfico de barras () función, 88 gráficos  
de barras, 93-94  
Teorema de Bayes, 212-214. *Ver también* ingenuo Bayes  
probabilidad condicional, 212
- BI (inteligencia empresarial)**  
herramientas analíticas, 10  
*versus* Ciencia de datos, 12-13 Big
- Data**  
3 Vs, 2-3  
análisis, ejemplos, 22-23  
características, 2  
de fi niciones, 2-3  
conductores, 15-16  
ecosistema, 16-19  
roles clave, 19-22  
McKinsey & Co. en, volumen 3, 2-3  
**impulso, 228-229**  
agregación bootstrap, 228  
diagramas de caja y bigotes, 95-96  
Metodología de Box-Jenkins, 235-236  
ARIMAmódelo, 236  
ramas (árboles de decisión), 193 Brown  
Corpus, 267-268  
impulsores comerciales de la analítica, 11  
Analista de Business Intelligence, fase de operacionalización, 52  
Función de analista de inteligencia empresarial, 27  
Usuario empresarial, fase de puesta en funcionamiento, 52 Función de usuario empresarial, 27  
compradores de datos, 18
- Algoritmo C4.5, 203-204  
proveedores de televisión por cable, 17 reglas candidatas, 141-142  
CART (árboles de clasificación y regresión), 204
- plegado de casos en análisis de texto, 264-265  
algoritmos categóricos, 205  
variables categóricas, 170-171  
cbind () función, 78  
centroídes, 120-122  
posiciones iniciales, 134  
tipos de datos de caracteres, R, 72  
gráficos, 386-387  
tasa de abandono (clientes), 120  
regresión logística, 180-181  
clase( ) función, 72 clasi fi  
**cación**  
ensacado, 228  
impulso, 228-229  
agregación bootstrap, 228  
árboles de decisión, 192-193  
algoritmos, 197-200, 203-204  
decisiones binarias, 206  
sucursales, 193  
atributos categóricos, 205  
árboles de clasificación, 193  
variables correlacionadas, 206  
muñón de decisión, 194  
evaluando, 204-206  
algoritmo codicioso, 204  
nodos internos, 193  
variables irrelevantes, 205  
nodos, 193  
atributos numéricos, 205  
R y, 206-211  
variables redundantes, 206  
regiones, 205  
árboles de regresión, 193  
raíz, 193  
árboles cortos, 194  
divisiones, 193, 194, 197, 200-203  
estructura, 205  
usos, 194  
ingenuo Bayes, 211-212  
Teorema de Bayes, 212-214  
diagnóstico, 217-218  
clasificador de Bayes ingenuo, 214-217 R y, 218-224  
suavizado, 217
- árboles de clasificación, 193  
clasificadores  
precisión, 225  
diagnóstico, 224-228  
recordar, 225  
flujo de círculos, 9  
agrupación, 118  
algoritmos, 134-135  
centroídes, 120-122

**C**

- posiciones iniciales, 134  
diagnósticos, 128-129  
k-medias, 118-119  
algoritmo, 120-122  
segmentación de clientes, 120  
procesamiento de imágenes y, 119  
usos médicos, 119  
razones para elegir, 130-134 cambio de  
escala, 133-134  
unidades de medida, 132-133  
etiquetas, 127  
número de clústeres, 123-127  
código, especifiaciones técnicas en el proyecto, 376-377  
coeficientes, regresión lineal, 169  
combinadores, 302-303  
Comunicar Resultados fase del ciclo de vida, 30, 49-50 componentes,  
árboles cortos como, 194  
entropía condicional, 199  
probabilidad condicional, 212  
clasificador ingenuo de Bayes, 215-216  
confianza, 141-142  
resultado, 172  
parámetros, 171  
intervalo de confianza, 107  
confint () función, 171 matriz de  
confusión, 224, 280 tablas de  
contingencia, 79  
variables continuas, discretización, 211 corpora
- Brown Corpus, 267-268  
corpora in Natural Language Processing, 256 IC (contenido  
de información), 268-269  
análisis de sentimiento y, 278 variables  
correlacionadas, 206  
empresas de tarjetas de crédito, 2  
CRISP-DM, 28  
crowdsourcing, 17  
Archivos CSV (valores separados por comas), 64-65  
importando, 64-65  
segmentación de clientes  
k-medias, 120  
regresión logística, 180-181  
Archivos CVS, 6  
componentes cílicos del análisis de series de tiempo, 235
- re**  
datos  
necesidades de crecimiento, 9-10  
fuentes, 15-16  
datos () función, 84  
agregadores de datos, 17-18
- análisis de datos, exploratorio, 80-82  
visualización y 82-85  
Ciclo de vida de análisis de datos  
Función de analista de inteligencia empresarial, 27  
Función de usuario empresarial, 27  
Fase de comunicación de resultados, 30, 49-50  
Estudio de caso de GINA, 58-59  
Rol de ingeniero de datos, 27-28 Fase  
de preparación de datos, 29,  
36-37  
AlpineMiner, 42 años  
acondicionamiento de datos, 40-41  
visualización de datos, 41-42  
DataWrangler, 42 años  
inventario de conjuntos de datos, 39-40  
ETLT, 38-39  
Estudio de caso de GINA, 55-56  
Hadoop, 42  
OpenRefine, 42  
preparación de la caja de arena, 37-38  
herramientas, 42  
Rol de científico de datos, 28  
Rol DBA (Administrador de base de datos), 27 Fase de  
descubrimiento, 29  
dominio empresarial, 30-31  
identificación de la fuente de datos, 35-36  
encuadre, 32-33  
Estudio de caso de GINA, 54-55  
desarrollo de hipótesis, 35  
recursos, 31-32  
entrevista con el patrocinador, 33-34  
identificación de las partes interesadas, 33  
Estudio de caso de GINA, 53-60  
Fase de construcción de modelos, 30, 46-48  
AlpineMiner, 48 años  
Estudio de caso de GINA, 56-58  
Mathematica, 48  
Matlab, 48 años  
Octava, 48  
PL / R, 48  
Pitón, 48  
R, 48  
SAS EnterpriseMiner, 48 años  
SPSSModeler, 48  
SQL, 48  
ESTADÍSTICA, 48  
WEKA, 48  
Fase de planificación del modelo, 29-30, 42-44  
exploración de datos, 44-45  
Estudio de caso de GINA,  
56 selección de modelos, 45  
R, 45-46

- SAS / ACCESO, 46
- Servicios de análisis SQL, 46
- selección de variables, 44–45
- Fase operativa, 30, 50–53, 360**
  - Analista de Business Intelligence y, 52
  - Business User y, 52
  - Ingeniero de datos y, 52
  - Científico de datos y, 52
  - DBA (Administrador de base de datos) y, 52 estudio de caso de GINA, 59–60
  - ProjectManager y, 52
  - Patrocinador del proyecto y, 52
  - procesos, 28
  - Función de gestor de proyectos, 27
  - funciones de patrocinador del proyecto, 27 funciones, 26–28
  - compradores de datos, 18
  - limpieza de datos, 86
  - recolectores de datos, 17
  - acondicionamiento de datos, 40–41
  - tasa de creación de datos, 3
  - dispositivos de datos, 17
  - Ingeniero de datos, fase de puesta en funcionamiento, 52
  - Rol de ingeniero de datos, 27–28
  - formatos de datos, análisis de texto, 257
  - marcos de datos, 75–76
  - mercados de datos, 10
  - Fase de preparación de datos del ciclo de vida, 29, 36–37**
    - acondicionamiento de datos, 40–41
    - visualización de datos, 41–42
    - inventoryario de conjuntos de datos, 39–40
    - ETLT, 38–39
    - preparación de la caja de arena, 37–38
  - repositorios de datos, 9–11
    - tipos, 10–11
  - Profesionales expertos en datos, 20
  - ciencia de datos *versus* BI, 12–13
  - Científicos de datos, 28
    - actividades, 20–21
    - retos empresariales, 20
    - características, 21–22
  - Fase operativa y, 52**
  - recomendaciones y, 21**
    - modelos estadísticos y fuentes de datos de 20 a 21
      - Fase de descubrimiento, 35–36
      - análisis de texto, 257
    - estructuras de datos, 5–9
      - datos cuasi estructurados, 6, 7 datos semiestructurados, 6
      - datos estructurados, 6
      - datos no estructurados, 6
  - tipos de datos en R, 71–72
    - personaje, 72
    - lógica, 72
    - numérico, 72
    - vectores, 73–74
  - usuarios de datos, 18
  - visualización de datos, 41–42, 377–378
    - CSS y 378
    - GGobi, 377–378
    - Gnuplot, 377–378
    - gráficos, 380–386
      - limpiar, 387–392
      - tridimensional, 392–393
    - HTML y 378
    - puntos clave con apoyo, 378–379 métodos de representación, 386–387
    - SVG y 378
  - almacenes de datos, 11
  - Data Wrangler, 42 años
  - conjuntos de datos
    - exportando, R y, 69–71
    - importando, R y, 69–71
    - inventoryario, 39–40
  - Davenport, Tom, 28 años
  - DBA (administrador de base de datos), 10, 27
  - Fase operacional y 52 árboles de decisión, 192–193**
    - algoritmos, 197–200
      - C4.5, 203–204
      - CARRITO, 204
      - categórico, 205
      - codicioso, 204
      - ID3, 203
      - numérico, 205
      - decisiones binarias, 206
      - sucursales, 193
      - árboles de clasificación, 193
      - variables correlacionadas, 206
      - evaluando, 204–206
      - algoritmos codiciosos, 204
      - nodos internos, 193
      - variables irrelevantes, 205
      - nodos
        - profundidad, 193
        - hoja, 193
    - R y, 206–211
      - variables redundantes, 206
      - regiones, 205
      - árboles de regresión, 193
      - raíz, 193
      - árboles cortos, 194
      - muñón de decisión, 194

- divisiones, 193, 197  
detección, 200-203  
limitante, 194  
estructura, 205  
usos, 194
- Talento analítico profundo, 19-20
- Marco DELTA, 28
- pronóstico de demanda, regresión lineal y, 162 gráficas de densidad, análisis de datos exploratorios, 88-91 variables dependientes, 162  
estadística descriptiva, 79-80  
desviación, 183-184  
dispositivos, 17  
móvil, 16  
no tradicional, 16  
dispositivos inteligentes, 16
- DF (frecuencia de documentos), 271-272
- diagnóstico por imagen, 16
- diagnóstico
- reglas de asociación, 158
  - clásificadores, 224-228
  - regresión lineal
    - supuesto de linealidad, 173
    - Validación cruzada de N veces, 177-178
    - supuesto de normalidad, 174-177
    - residuos, 173-174
  - Regresión logística
    - desviación, 183-184
    - histograma de probabilidades, 188
    - prueba logarítmica de verosimilitud, 184-185
    - pseudo-R<sup>2</sup>, 183
    - Curva ROC, 185-187
  - ingeniero Bayes, 217-218
- diff () función, 245 diferencia de medias, 104
- intervalo de confianza, 107
  - prueba t de estudiante, 104-106
  - Prueba t de Welch, 106-108
- diferenciando, 241-242
- datos sucios, 85-87
- Fase de descubrimiento del ciclo de vida, 29
- identificación de la fuente de datos, 35-36
  - encuadre, 32-33
- desarrollo de hipótesis, 35
- entrevista con el patrocinador, 33-34
- identificación de las partes interesadas, 33
- discretización de variables continuas, 211
- documentos, categorización, 274-277
- dotchart () función, 88
- mi**
- Eclipse, 304
- ecosistema de Big Data, 16-19
- Profesionales conocedores de datos, 20  
talentos analíticos profundos, 19-20 roles clave, 19-22  
Habilitadores de datos y tecnología, 20 EDW (almacenes de datos empresariales), 10 tamaño de efecto, 110  
Ejemplo de búsqueda de EMC en Google, 7 a 9  
emoticonos, 282  
ingeniería, regresión logística y, 179 métodos de conjunto, árboles de decisión, 194 distribución de errores  
modelo de regresión lineal, 165-166 error estándar residual, 170 ETLT, 38-39
- EXCEPT operator (SQL), 333-3334 análisis de datos exploratorios, 80-82
- gráfico de densidad, 88-91
  - datos sucios, 85-87
  - histogramas, 88-91
  - múltiples variables, 91-92
  - análisis a lo largo del tiempo, 99 gráficos de barras, 93-94
  - diagramas de caja y bigotes, 95-96
  - gráficos de puntos, 93-94
  - hexbinplots, 96-97
  - versus presentación, 99-101
  - matriz de diagramas de dispersión, 97-99
  - visualización y 82-85
  - variable única, 88-91
- exportar conjuntos de datos en R, 69-71  
expresiones, regular, 263
- F**
- Facebook, 2, 3-4
- factores, 77-78
- información financiera, regresión logística y, 179 FNR (tasa de falsos negativos), 225
- previsión
- ARIMA (media móvil integrada autorregresiva)
    - modelo, 251-252
  - regresión lineal y, 162
- FP (falsos positivos), matriz de confusión, 224 FPR (tasa de falsos positivos), 225
- encuadre en la fase de descubrimiento, 32-33
- funciones
- aov (), 78
  - a priori(), 146, 152-157
  - arima (), 246
  - matriz (), 74
  - gráfico de barras (), 88
  - cbind (), 78
  - clase(), 72
  - confint (), 171

datos(), 84  
 diff(), 245  
 dotchart(), 88  
 gl(), 84  
 glm(), 183  
 hclust(), 135  
 cabeza(), sesenta y cinco  
 inspeccionar(), 147, 154-155  
 entero(), 72  
 IQR(), 80  
 es.data.frame(), 75  
 es.na(), 86  
 es.vector(), 73  
 jpeg(), 71  
 kmeans(), 134  
 kmode(), 134-135  
 longitud(), 72  
 biblioteca(), 70  
 lm(), 66  
 cargar imagen(), 68-69  
 matrix.inverse(), 74  
 media(), 86  
 mi\_rango(), 80  
 na.exclude(), 86  
 pamk(), 135  
 Cerdó, 307-308  
 trama(), 65, 153-154, 245  
 predecir(), 172  
 rbind(), 78  
 read.csv(), 64-65, 75  
 read.csv2(), 70  
 read.delim2(), 70  
 rpart, 207  
 SQL, 347-351  
 sqlQuery(), 70  
 str(), 75  
 resumen(), 65, 66-67, 79, 80-82  
 t(), 74  
 ts(), 245  
 tipo de(), 72  
 wilcox.test(), 109  
 funciones de ventana (SQL), 343-347  
 write.csv(), 70  
 write.csv2(), 70  
 write.table(), 70

GINA (Red y análisis de innovación global), Datos  
 Caso de éxito de Analytics Lifecycle, 53-60  
 gl() función, 84  
 glm() función, 183  
 Gnuplot, 377-378  
 Sistemas GPS, 16  
 Búsqueda de gráficos (Facebook), 3 a 4  
 gráficos, 380 a 386  
 limpiar, 387-392  
 tridimensional, 392-393  
 algoritmos codiciosos, 204  
*Huevos Verdes con jamón*, análisis de texto y, 256 ejemplos de  
 tienda de comestibles del algoritmo Apriori, 143  
 Comestibles conjunto de datos, 144-146  
 conjuntos de elementos, generación frecuente, 146-151  
 reglas, generación, 152-157  
 necesidades de crecimiento de datos, 9-10  
 GUI (interfaces gráficas de usuario), R y, 67-69

## H

Hadoop  
 Fase de preparación de datos, 42 Hadoop  
 Streaming API, 304-305 HBase, 311-312  
 arquitectura, 312-317  
 nombres de familias de columnas, 319  
 nombres de calificadores de columnas, 319  
 modelos de datos, 312-317  
 API de Java y, 319  
 filas, 319  
 casos de uso, 317-319  
 control de versiones, 319  
 Cuidador del zoológico, 319  
 HDFS, 300-301  
 Hive, 308-311  
 LinkedIn, 297  
 Mahout, 319-320  
 MapReduce, 22  
 combinadores, 302-303  
 desarrollo, 304-305  
 conductores, 301  
 ejecución, 304-305  
 cartógrafos, 301-302  
 divisores, 304  
 estructuración, 301-304  
 procesamiento del lenguaje natural, 18 Pig, 306-308  
 tubos, 305  
 Watson (IBM), 297  
 Yahoo!, 297-298  
 YARN (Yet Another Resource Negotiator), 305 conjuntos de  
 elementos basados en hash, algoritmo Apriori y 158

## GRAMO

Función de modelo lineal generalizado, 182  
 secuenciación genética, 3, 4  
 genómica, 4, 16  
 genotipado, 4  
 GGobi, 377-378

HAWQ (Hadoop con consulta), 321 HBase,

311-312

arquitectura, 312-317

nombres de familias de columnas, 319

nombres de cali fi cadores de columnas,

319 modelos de datos, 312-317

API de Java y, 319

filas, 319

casos de uso, 317-319

control de versiones, 319

Cuidador del zoológico, 319

**hclust () función, 135**

HDFS (sistema de archivos distribuido Hadoop), 300-301

cabeza () función, 65

hexbinplots, 96-97

histogramas

análisis de datos exploratorios, 88-91

regresión logística, 188

Hive, 308-311

HiveQL (lenguaje de consulta de Hive), 308 Hopper,

Grace, 299

Hubbard, Doug, 28 años

HVE (extensiones de virtualización de Hadoop), 321

hipótesis

hipótesis alternativa, 102-103

Fase de descubrimiento, 35

hipótesis nula, 102

prueba de hipótesis, 102-104

prueba de hipótesis de dos caras, 105 errores de

tipo I, 109-110

errores de tipo II, 109-110

## yo

IBM Watson, 297

Algoritmo ID3, 203

IDE (entorno de desarrollo interactivo), 304 IDF (frecuencia de documentos invertida), 271-272 importando conjuntos de datos en

R, 69-71

análisis en la base de datos

SQL, 328-338

análisis de texto, 338-339

variables independientes, 162

variables de entrada, 192

inspeccionar() función, 147, 154-155

entero () función, 72 nodos internos (árboles

de decisión), 193 Internet de las cosas, 17-18

Operador INTERSECT (SQL), 333

IQR () función, 80

is.data.frame () función, 75

is.na () función, 86

es.vector () función, 73 conjuntos de

elementos, 139

Conjuntos de 1 artículo, 147

Conjuntos de 2 elementos, 148-149

Conjuntos de 3 elementos, 149-150

Conjuntos de 4 elementos, 150-151

Algoritmo a priori, 139

Propiedad Apriori, 139

propiedad de cierre hacia abajo, 139

conteo dinámico, algoritmo Apriori y 158 conjuntos de elementos

frecuentes, 139

generación, frecuente, 146-151

algoritmo Apriori basado en hash y 158 k-itemset,

139, 140-141

## J

uniones (SQL), 330-332

jpeg () función, 71

## K

k grupos

encontrando, 120-122

número de, 123-127

k-itemset, 139, 140-141

k-medias, 118-119

segmentación de clientes, 120

procesamiento de imágenes y grupos

de 119 k

encontrando, 120-122

número de, 123-127

usos médicos, 119

objetos, atributos, 130-131

R y, 123-127

razones para elegir, 130-134 cambio de

escala, 133-134

unidades de medida, 132-133

kmeans () función, 134

kmodes () función, 134-135

## L

retraso, 237

Alisado de Laplace, 217

regresión de lazo, 189

LDA (asignación de Dirichlet latente), 274-275 nodos de

hoja, 192, 193

lematización, análisis de texto y, 258

longitud() función, 72

apalancamiento, 142

biblioteca () función, 70

- ciclo vital. Ver también Incremento del ciclo de vida de análisis de datos, 142
- regresión lineal, 162
- coeficientes, 169
  - diagnóstico
    - supuesto de linealidad, 173
    - Validación cruzada de N veces, 177–178
    - supuesto de normalidad, 174–177
    - residuos, 173–174
  - modelo, 163–165
    - variables categóricas, 170–171
    - errores distribuidos normalmente, 165–166
    - intervalos de confianza de resultados, 172 intervalos de confianza de parámetros, 171 intervalo de predicción sobre el resultado, 172
    - R, 166–170
    - valores p, 169–170
    - casos de uso, 162–163
- LinkedIn, 2, 22–23, 297 listas en R, 76–77
- Im () función, 66
- cargar imagen() función, 68–69 tipos de datos lógicos, R, 72
- regresión logística, 178
- advertencias, 188–189
  - diagnóstico, 181–182
    - desviación, 183–184
    - histograma de probabilidades, 188
    - prueba logarítmica de verosimilitud, 184–185
    - pseudo-R<sup>2</sup>, 183
    - Curva ROC, 185–187
  - Función de modelo lineal generalizado, modelo 182, 179–181
  - multinomial, 190
  - razones para elegir, 188–189 casos de uso, 179
  - prueba logarítmica de verosimilitud, 184–185
  - tarjetas de fidelidad, 17
- ## METRO
- Habilidades MAD (magnéticas / ágiles / profundas), 28, 352–356
- MADlib, 352–356
- Mahout, 319–320
- MapReduce, 22, 298–299
- combinadores, 302–303
  - desarrollo, 304–305
  - controladores, 301–302
  - ejecución, 304–305
  - cartógrafos, 301–302
  - divisores, 304
  - estructuración, 301–304
- análisis de la cesta de la compra, 139
- reglas de asociación, 143
- marketing, regresión logística y, 179 nodos maestros, 301
- matrices
- matriz de confusión, 224
  - R, 74–75
  - matrices de diagramas de dispersión, 97–99
- matrix.inverse () función, 74 MaxEnt (máxima entropía), 278 McKinsey & Co. de iniciación de Big Data, 3
- significa () función, 86
- información médica, 16
- k-medias y, 119
  - regresión lineal y, 162
  - regresión logística y, 179
  - confianza mínima, 141
- datos faltantes, 86
- dispositivos móviles, 16
- empresas de telefonía móvil, 2
- Fase de construcción de modelos del ciclo de vida, 30, 46–48
- Minero alpino, 48
  - Mathematica, 48
  - Matlab, 48 años
  - Octava, 48
  - PL / R, 48
  - Pitón, 48
  - R, 48
  - SAS Enterprise Miner, 48
  - Modelador de SPSS, 48
  - SQL, 48
  - ESTADÍSTICA, 48
  - WEKA, 48
- Fase de planificación del modelo del ciclo de vida, 29–30, 42–44
- exploración de datos, 44–45
  - selección de modelo, 45
  - R, 45–46
  - SAS / ACCESO, 46
  - Servicios de análisis SQL, 46
  - variables, selección, 44–45
- características morfológicas en el análisis de texto, 266–267 modelos de promedio móvil, 239–241
- MPP (procesamiento masivamente paralelo), 5 MTurk (Mechanical Turk), 282
- regresión logística multinomial, 190 análisis de series de tiempo multivariante, 253
- my\_range () función, 80
- ## norte
- na.exclude () función, 86 ingenuo
- Bayes, 211–212
- Teorema de Bayes, 212–214
  - diagnóstico, 217–218

clásificador de Bayes ingenuo, 214–217 R y, 218–224 análisis de sentimiento y, 278 suavizado, 217 procesamiento del lenguaje natural, validación cruzada de 18 N veces, 177–178 NLP (procesamiento de lenguaje natural), 256 nodos maestro, 301 trabajador, 301 nodos (árboles de decisión), 192 profundidad, 193 hoja, 193 nodos de hoja, 192, 193 pruebas no paramétricas, 108–109 dispositivos no tradicionales, 16 normalidad ARIMAmódelo, 250–251 regresión lineal, 174–177 normalización, acondicionamiento de datos, 40–41 NoSQL, 322–323 desviación nula, 183 hipótesis nula, 102 tipos de datos numéricos, R, 72 algoritmos numéricos, 205 subflujo numérico, 216–217

**O**

objetos, k-medias, atributos, 130–131 OLAP (procesamiento analítico en línea), 6 cubos, 10 OpenRefine, 42 Fase operativa del ciclo de vida, 30, 50–53, 360 Analista de Business Intelligence y, 52 Business User y, 52 Ingeniero de datos y, 52 Científico de datos y, 52 DBA (Administrador de base de datos) y, 52 Project Manager y, 52 Patrocinador del proyecto y, 52 operadores, subconjunto, 75 Salir intervalos de confianza, 172 intervalo de predicción, 172

**PAGS**

PACF (función de autocorrelación parcial), 238–239 pamk () función, 135 parámetros, intervalos de confianza, 171 pruebas paramétricas, 108–109 análisis de texto, análisis de texto y partición 257 Algoritmo a priori y, 158 MapReduce, 304 fotografías, 16 Cerdó, 306–308 Pivotal HD Enterprise, 320–321 trama( ) función, 65, 153–154, 245 etiquetado POS (parte del discurso), 258 poder de una prueba, 110 precisión en el análisis de sentimientos, 281 predecir() función, 172 árboles de predicción. Ver presentación de árboles de decisión versus exploración de datos, 99–101 probabilidad, condicional, 212 clasificador ingenuo de Bayes, 215–216 Project Manager, fase de puesta en funcionamiento, 52 Rol de Project Manager, 27 Patrocinador del proyecto, fase de puesta en funcionamiento, 52 Función del patrocinador del proyecto, 27 pseudo-R<sub>2</sub>, 183 valores p, regresión lineal, 169–170

**Q**

datos casi estructurados, 6, 7 consultas, SQL, 329–330 anidado, 3334 subconsultas, 3334

**R**

matrices, 74–75 atributos, tipos, 71–72 tramas de datos, 75–76 tipos de datos, 71–72 personaje, 72 lógica, 72 numérico, 72 vectores, 73–74 árboles de decisión, 206–211 estadística descriptiva, 79–80 análisis de datos exploratorios, 80–82 gráfico de densidad, 88–91 datos sucios, 85–87 histogramas, 88–91 múltiples variables, 91–99 versus presentación, 99–101 visualización y, 82–85, 88–91 factores, 77–78 funciones

**aov ()**, 78  
**matriz ()**, 74  
 gráfico de barras (), 88  
**cbind ()**, 78  
**clase( )**, 72  
**datos ()**, 84  
**dotchart ()**, 88  
**gl ()**, 84  
 cabeza (), sesenta y cinco  
 valores predeterminados de la función de importación, 70  
**entero ()**, 72  
**IQR ()**, 80  
**es.data.frame ()**,  
 75  
**es.na ()**, 86  
**es.vector ()**, 73  
**jpeg ()**, 71  
**longitud( )**, 72  
**biblioteca ()**, 70  
**Im ()**, 66  
 cargar imagen(), 68–69  
**mi\_rango ()**, 80  
**trama( )** función, 65  
**rbind ()**, 78  
**read.csv ()**, 65, 75  
**read.csv2 ()**, 70  
**read.delim ()**, 69  
**read.delim2 ()**, 70  
**read.table ()**, 69  
**str ()**, 75  
**resumen()**, 65, 66–67, 79  
**t ()**, 74  
**tipo de( )**, 72  
 visualización de una sola variable, 88  
**write.csv ()**, 70  
**write.csv2 ()**, 70  
**write.table ()**, 70  
**GUI**, 67–69  
 importación / exportación, 69–71  
 análisis de k-medias, 123–127  
 modelo de regresión lineal, 166–170 listas,  
 76–77  
 matrices, 74–75  
 planificación de modelos y, 45–46 Bayes  
 ingenuos y, 218–224 operadores,  
 subconjuntos, 75  
 descripción general, 64–67  
 técnicas estadísticas, 101–102  
**ANOVA**, 110–114  
 diferencia inmeans, 104–108  
 efect tamaño, 110  
 prueba de hipótesis, 102–104  
 potencia de la prueba,  
 110 tamaño de muestra, 110  
 errores de tipo I, 109–110  
 errores de tipo II, 109–110  
 tablas, tablas de contingencia, 79 R  
**commander GUI**, 67  
 componentes aleatorios del análisis de series de tiempo, 235 Rattle  
**GUI**, 67  
 texto crudo  
     colección, 260–263  
     tokenización, 264  
**rbind ()** función, 78  
**RDBMS**, 6  
**read.csv ()** función, 64–65, 75  
**read.csv2 ()** función, 70  
**read.delim ()** función, 69  
**read.delim2 ()** función, 70  
**read.table ()** función, 69  
 bienes raíces, regresión lineal y, 162 recuerdo en el  
 análisis de sentimiento, 281 variables redundantes,  
 206  
**regresión**  
     lazo, 189  
     lineal, 162  
         coeficientes, 169  
         diagnósticos, 173–178  
         modelo, 163–172  
         valores p, 169–170  
         casos de uso, 162–163  
     logístico, 178  
         advertencias, 188–189  
         diagnósticos, 181–188  
         modelo, 179–181  
         logística multinomial,  
             190  
         razones para elegir, 188–189 casos de  
         uso, 179  
     logística multinomial, 190  
     cresta, 189  
     variables  
         dependiente, 162  
         independiente, 162  
     árboles de regresión, 193  
     expresiones regulares, 263, 339–340  
     relaciones, 141  
     repositorios, 9–11  
         tipos, 10–11  
     métodos de representación, 386–387  
     cambio de escala, k-medias, 133–134  
     desviación residual, 183  
     error estándar residual, 170

residuos, regresión lineal, 173–174

recursos, Fase de descubrimiento del ciclo de vida, 31–32

lectores RFID, 16

regresión de la cresta, 189

Curva ROC (característica operativa del receptor), 185–187, 225 raíces (árboles de decisión), 193

rpart función, 207

RStudio GUI, 67–68

reglas

reglas de asociación, 138–139

  aplicación, 143

  reglas candidatas, 141–142

  diagnóstico, 158

  pruebas y 157–158

  validación, 157–158

generación, ejemplo de tienda de comestibles (Apriori), 152–157

## S

ventas, análisis de series de tiempo y, tamaño de

muestra 234, 110

muestreo, algoritmo Apriori y, 158 cajas de arena, 10,

11. Ver también espacios de trabajo

  Fase de preparación de datos, 37–38 SAS /

ACCESS, planificación del modelo, 46 matriz de

diagramas de dispersión, 97–99

diagramas de dispersión, 81

  Cuarteto de Anscombe, 83

  múltiples variables, 91–92

método científico, 28

búsquedas, análisis de texto y, 257

modelo de media móvil integrado autorregresivo estacional,

243–244

componentes de estacionalidad del análisis de series de tiempo, 235

procesamiento sísmico, 16

datos semiestructurados, 6

SensorNet, 17–18

análisis de sentimientos en el análisis de texto, 277–283

  matriz de confusión, 280

  precisión, 281

  recordar, 281

compras

  tarjetas de fidelidad, 17

  Fichas RFID en carros, 17

árboles cortos, 194

dispositivos inteligentes, 16

teléfonos inteligentes, 17

suavizado, 217

redes sociales, 3–4

fuentes de datos, 15–16

planificación de partes de partes, análisis de series de tiempo y, 234–235

divisiones (árboles de decisión), 193

  detección, 200–203

entrevista con el patrocinador, fase de descubrimiento,

33 spreadmarts, 10

hojas de cálculo, 6, 9, 10

SQL (lenguaje de consulta estructurado), 328–329

agregados

  pedido, 351–352

  de fi nido por el usuario, 347–351

EXCEPTO operador, 333–3334

funciones, de fi nidas por el usuario, 347–351

agrupación, 334–338

Operador INTERSECT, 333

une, 330–332

MADlib, 352–356

consultas, 329–330

  anidado, 3334

  subconsultas, 3334

establecer operaciones, 332–334

Operador UNION ALL, 332–333

funciones de ventana, 343–347

Servicios de análisis SQL, planificación de modelos y 46

sqlQuery () función, 70

partes interesadas, fase de descubrimiento del ciclo de vida, 33

series de tiempo estacionales, 236

técnicas estadísticas, 101–102

ANOVA, 110–114

  diferencia de medias, 104

  prueba t de Student, 104–106

  Prueba t de Welch, 106–108

  efecto tamaño, 110

  prueba de hipótesis, 102–104

  potencia de la prueba,

  110 tamaño de muestra, 110

  errores de tipo I, 109–110

  errores de tipo II, 109–110

  Prueba de suma de rangos de Wilcoxon,

estadísticas 108–109

  Cuarteto de Anscombe, 82–83

  descriptivo, 79–80

derivación, análisis de texto y, 258

negociación de acciones, análisis de series de tiempo y 235

palabras vacías, 270–271

str () función, 75 datos

estructurados, 6

operadores de subconjuntos, 75

resumen() función, 65, 66–67, 79, 80–82 SVM (máquinas de

vectores de soporte), 278

## T

t () función, 74

tablas, tablas de contienda, 79 tiendas

Target, 22

distribución t

- ANOVA, 110-114  
 prueba t de Student, 104-106  
 Prueba t de Welch, 106-108  
 especificaciones técnicas en el proyecto, 376-377  
 Habilidades de datos y tecnología, 20  
 pruebas, reglas de asociación y, 157-158 análisis de texto, 256  
 Ejemplo de ACME, 259-263  
 bolsa de palabras, 265-266  
 corpora, 264-265
  - Brown Corpus, 267-268
  - corpora inNatural Language Processing, 256 IC (corpus de información), 268-269
  - formatos de datos, 257
  - fuentes de datos, 257
  - categorización de documentos, 274-277
  - Huevos Verdes con jamón*, 256 en la base de datos, 338-339
  - lematización, 258
  - características morfológicas, 266-267
  - NLP (procesamiento de lenguaje natural), 256 análisis, 257
  - Etiquetado POS (parte del discurso), 258 texto sin formato, recopilación, 260-263 búsqueda y recuperación, 257
  - análisis de sentimientos, 277-283
  - despalillado, 258
  - palabras vacías, 270-271
  - minería de texto, 257-258
  - TF (frecuencia de término) de palabras, 265-266
    - DF, 271-272
    - FDI, 271-272
    - lematización, 271
    - despalillado, 271
    - palabras vacías, 270-271
    - TFIDF, 269-274
  - tokenización, 264
  - modelado de temas, 267, 274
    - LDA (asignación de Dirichlet latente), raspador de banda 274-275, 262-263
    - nubes de palabras, 284
    - Ley de Zipf, 265-266
  - minería de texto, 257
  - archivos de datos textuales, 6
  - TF (frecuencia de término) de palabras, 265-266
    - DF (frecuencia de documentos), 271-272
    - IDF (frecuencia de documentos invertida), 271-272
    - lematización, 271
    - despalillado, 271
    - palabras vacías, 270-271
    - TFIDF, 269-274
  - TFIDF (frecuencia de término-frecuencia de documento inverso), 269-274, 285-286
  - análisis de series temporales
- ARIMAmodelo, 236  
 ACF, 236-237  
 ARMAmodel, 241-244  
 modelos autorregresivos, 238-239  
 edificio, 244-252  
 precauciones, 252-253  
 varianza constante, 250-251  
 evaluando, 244-252  
 modelos ajustados, 249-250  
 pronóstico, 251-252  
 modelos de media móvil, 239-241  
 normalidad, 250-251  
 PACF, 238-239  
 razones para elegir, 252-253  
**media móvil integrada autogresiva estacional**  
 modelo, 243-244  
**ARMAX (media móvil autorregresiva con**  
 Insumos exógenos), 253  
**Metodología de Box-Jenkins, 235-236**  
 componentes cílicos, 235  
 diferenciando, 241-242  
 modelos equipados, 249-250  
 GARCH (autorregresivo generalizado condicionalmente Heterocedástico), 253  
 Filtrado de Kalman, 253  
 análisis multivariante de series de tiempo, 253  
 componentes aleatorios, 235  
**media móvil integrada autorregresiva estacional**  
 modelo, 243-244  
 estacionalidad, 235  
 análisis espectral, 253  
 series de tiempo estacionarias, 236  
 tendencias, 235  
 casos de uso, 234-235  
 proceso de ruido blanco, 239  
 tokenización en análisis de texto, 264  
 modelado de temas en el análisis de texto, 267, 274
  - LDA (asignación de Dirichlet latente), 274-275 TP (verdaderos positivos), matriz de confusión, 224 TPR (tasa de verdaderos positivos), 225
  - datos de transacciones, 6
  - reducción de transacciones, algoritmo Apriori y, 158 tendencias, análisis de series de tiempo, 235
  - TRP (tasa de verdaderos positivos), 185-187- ts () función, 245
  - prueba de hipótesis de dos caras, 105
  - errores de tipo I, 109-110
  - errores de tipo II, 109-110
  - tipo de() función, 72

**U**

Operador UNION ALL (SQL), 332-333 unidades de medida, k-medias, 132-133 datos no estructurados,

Apache Hadoop, HDFS, 300–301

LinkedIn, 297

MapReduce, 298–299

procesamiento natural del lenguaje,

18

casos de uso, 296–298

Watson (IBM), 297

Yahoo!, 297–298

Técnicas no supervisadas. *Ver* agrupación de usuarios de datos, 18

## V

validación, reglas de asociación y, 157–158 variables

categórico, 170–171

continuo, discretización, 211

correlacionado, 206

árboles de decisión, 205

dependiente, 162

factores, 77–78

independiente, 162

entrada, 192

redundante, 206

VARIMA (Vector ARIMA), 253

vectores, R, 73–74

metraje de video, 16

k-medias y, 119

videovigilancia, 16

visualización, 41–42. *Ver también* Visualización de datos

análisis de datos exploratorios, 82–85

variable única, 88–91

ejemplo de tienda de abarrotes (Apriori), 152–157 volumen, variedad, velocidad. *Ver* 3

Vs (volumen, variedad, velocidad)

## W

Watson (IBM), 297

rascador de banda, 262–263

proceso de ruido blanco, 239

Prueba de suma de rangos de Wilcoxon, 108–109

wilcox.test () función, 109 funciones de

ventana (SQL), 343–347 nubes de palabras,

284

espacios de trabajo, 10, 11. *Ver también* areneros

Fase de preparación de datos, 37 a 38

nodos trabajadores, 301

write.csv () función, 70

write.csv2 () función, 70

write.table () función, 70 WSS (dentro de la suma

de cuadrados), 123–127

## XZ

XML (eXtensible Markup Language), 6 Yahoo!,

297–298

YARN (otro negociador de recursos),

305

Ley de Zipf, 265–266

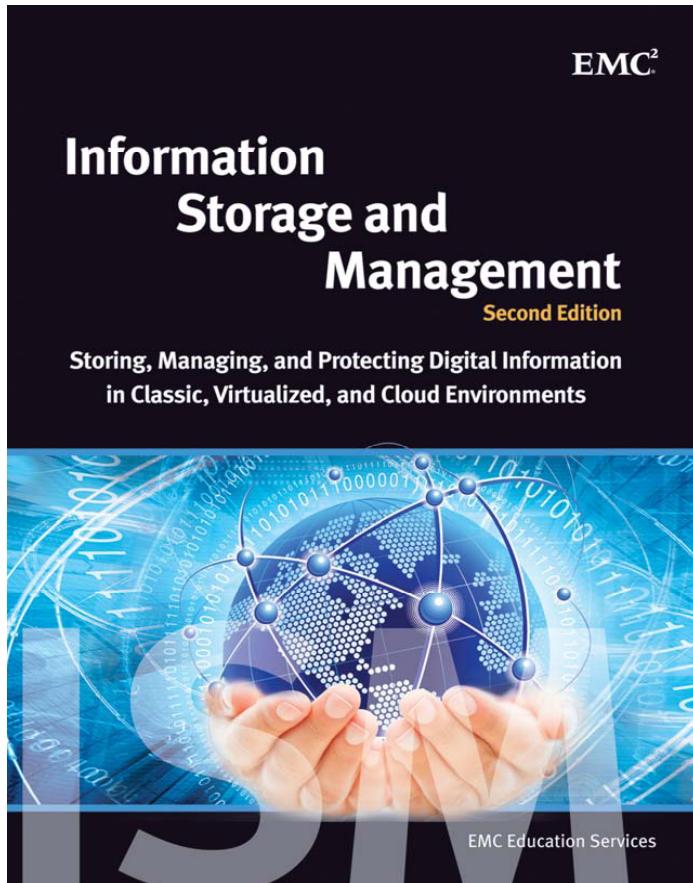






# A comprehensive book on information storage and management by the EMC's Education Services, a global technology leader.

More than ever, the IT industry is challenged with employing and developing highly skilled technical professionals with storage technology expertise across classic, virtualized, and cloud environments. This book covers concepts, principles, and deployment considerations across technologies that are used for storing and managing information.



978-1-118-09483-9 · \$70.00 US · \$77.00 CAN · £47.50 UK

Key Technology Strategies for Classic, Virtualized, and Cloud Environments:

- Challenges and Solutions for Data Storage and Management
- Intelligent Storage, Object-Based Storage, and Unified Storage
- Storage Networking, Federation, and Protocols
- Backup, Recovery, Deduplication, and Archive
- Business Continuity and Disaster Recovery
- Cloud Computing and Converged Infrastructure
- Storage Security and Managing Storage Infrastructure

**WILEY**

## **ACUERDO DE LICENCIA DE USUARIO FINAL DE WILEY**

Vaya a [www.wiley.com/go/eula](http://www.wiley.com/go/eula) para acceder al EULA del libro electrónico de Wiley.