Unidad 4 - Análisis exploratorio de datos: Visualizaciones

Fundamentos de ciencia de datos

Introducción

Recomendamos la siguiente lectura: Articulo del Gorila

Dicho artículo trata sobre el costo oculto de tener una hipótesis en la investigación científica. Allí se argumenta que si bien tener una hipótesis puede ser útil para guiar el diseño de pruebas y analizar los resultados de manera efectiva, también puede limitar la creatividad y la exploración de nuevos descubrimientos. Se utiliza la metáfora de la ciencia nocturna y diurna para describir las dos fases de la investigación científica. La fase diurna es cuando se realizan pruebas y se analizan los resultados para probar una hipótesis, mientras que la fase nocturna es cuando se exploran nuevos descubrimientos sin tener una hipótesis específica en mente. El artículo argumenta que la limitación de la creatividad que surge de tener una hipótesis puede ser particularmente problemática en el contexto de conjuntos de datos biológicos modernos que son grandes y contienen múltiples descubrimientos potencialmente emocionantes.

Para ilustrar este punto, el artículo describe un experimento en el que los estudiantes analizan un conjunto de datos. A un grupo de estudiantes se les pide que consideren tres hipótesis específicas al analizar los datos, mientras que a otro grupo se les pide simplemente que analicen los datos sin tener una hipótesis específica en mente. Los resultados del experimento muestran que los estudiantes que no tienen una hipótesis específica tienden a identificar más descubrimientos y patrones en los datos que los estudiantes que tienen una hipótesis específica. El artículo destaca la importancia de equilibrar la investigación diurna y nocturna para maximizar el potencial de descubrimientos emocionantes y útiles en la ciencia.

Visualización de datos

La visualización de datos, también conocida como data visualization, es una herramienta fundamental en la ciencia de datos que permite representar información compleja de manera visual y fácil de entender. En general, la ciencia de datos implica el análisis de grandes cantidades de datos para extraer información útil y obtener conocimiento sobre fenómenos complejos. Sin embargo, la simple acumulación de datos no es suficiente para tomar decisiones informadas. Es necesario que los datos sean presentados de una manera que permita identificar patrones, tendencias y relaciones.

La visualización de datos, por lo tanto, se refiere al uso de gráficos, diagramas, mapas, tablas y otros recursos visuales para representar datos de una manera que sea fácil de entender y que permita identificar patrones y tendencias que podrían no ser evidentes de otra manera. El objetivo principal de la visualización de datos es comunicar información de manera efectiva y ayudar a los usuarios a tomar decisiones informadas.

En la ciencia de datos, la visualización de datos se utiliza **en todas las etapas del proceso**, desde la exploración y limpieza de los datos, hasta la presentación de los resultados finales. Es una herramienta clave para explorar y comprender los datos, identificar patrones, relaciones y outliers, y comunicar los resultados a audiencias no técnicas. Además, la visualización de datos también se utiliza para comunicar los resultados de modelos complejos y hacer que los datos sean más accesibles y comprensibles para personas con diferentes niveles de habilidad y conocimiento técnico.

¿Por qué visualizar datos?

(Extraído del libro “Data Visualization: Exploring and Explaining with Data” - Jeffrey D. Camm)

Creamos visualizaciones de datos por dos razones: para explorar los datos y para comunicar/explicar un mensaje. Discutamos estos usos de la visualización de datos con más detalle, examinemos las diferencias entre los dos usos y consideremos cómo se relacionan con los tipos de análisis descritos anteriormente.

Visualización de datos para la exploración

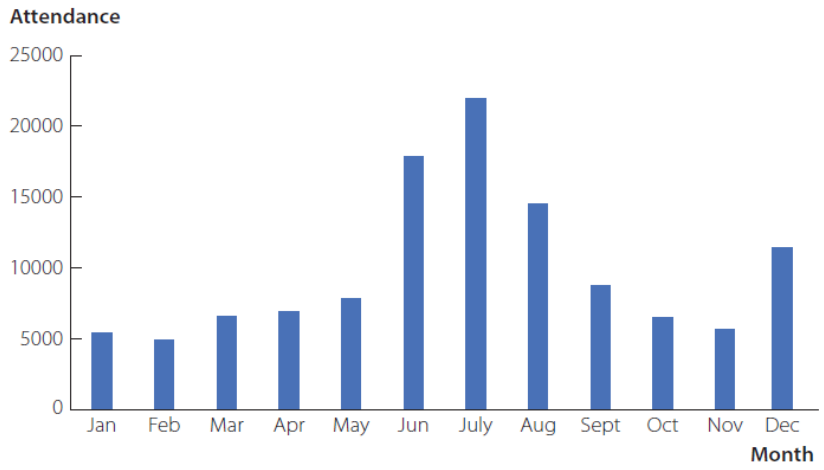
La visualización de datos es una herramienta poderosa para explorar datos y poder identificar patrones, reconocer anomalías o irregularidades en los datos y comprender mejor las relaciones entre variables. Nuestra capacidad para detectar este tipo de características de los datos es mucho más fuerte y rápida cuando vemos una visualización de los datos en lugar de una simple lista.

Como ejemplo de visualización de datos para la exploración, consideremos los datos de asistencia al zoológico mostrados en la Tabla 1.1 y la Figura 1.1. Comparando la Tabla 1.1 y la Figura 1.1, observe que el patrón en los datos es más detectable en el gráfico de columnas de la Figura 1.1 que en una tabla de números. Un gráfico de columnas muestra datos numéricos mediante la altura de la columna para una variedad de categorías o períodos de tiempo. En el caso de la Figura 1.1, los períodos de tiempo son los diferentes meses del año.

TABLE 1.1 Zoo Attendance Data

Month	Jan	Feb	Mar	Apr	May	Jun
Attendance	5422	4878	6586	6943	7876	17843
Month	July	Aug	Sept	Oct	Nov	Dec
Attendance	21967	14542	8751	6454	5677	11422

FIGURE 1.1 A Column Chart of Zoo Attendance by Month

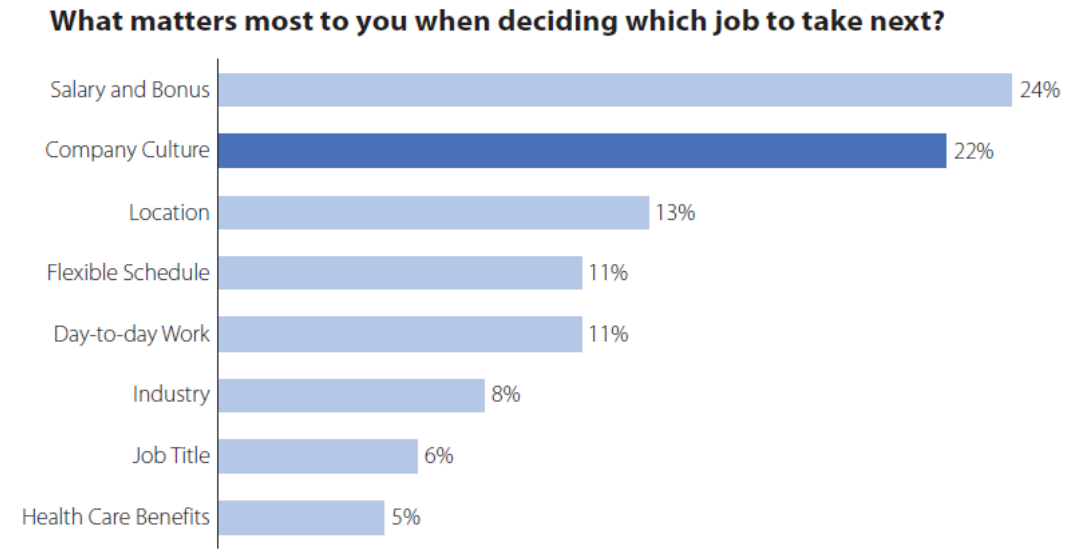


Nuestra intuición y experiencia nos dice que esperaríamos que la asistencia al zoológico sea más alta en los meses de verano, cuando muchos niños en edad escolar están de vacaciones de verano (recordemos que este ejemplo fue tomado de un libro escrito por una persona del hemisferio norte). La Figura 1.1 confirma esto, ya que la asistencia al zoológico es más alta en los meses de verano de junio, julio y agosto (en hemisferio norte). Además, vemos que la asistencia aumenta gradualmente cada mes de febrero a mayo a medida que aumenta la temperatura promedio, y disminuye gradualmente cada mes de septiembre a noviembre a medida que la temperatura promedio disminuye. ¿Pero por qué la asistencia al zoológico en diciembre y enero no sigue estos patrones? Resulta que el zoológico tiene un evento conocido como el "Festival de las Luces" que se lleva a cabo desde finales de noviembre hasta principios de enero. Los niños están de vacaciones escolares durante la última mitad de diciembre y principios de enero para las fiestas de fin de año, y esto lleva a un aumento de la asistencia en las noches en el zoológico a pesar de las bajas temperaturas invernales. La exploración visual de datos es una parte importante del análisis descriptivo. La visualización de datos también se puede usar directamente para monitorear las principales métricas de rendimiento, es decir, medir cómo está funcionando una organización en relación con sus objetivos.

Visualización de datos para explicación

La visualización de datos también es importante para explicar relaciones encontradas en los datos y para explicar los resultados de modelos predictivos y prescriptivos. En general, la visualización de datos es útil para comunicarse con su audiencia y asegurarse de que su audiencia comprenda y se centre en su mensaje previsto.
Consideremos el artículo "Conoce la cultura antes de un nuevo trabajo", que apareció en The Wall Street Journal. El artículo discute la importancia de encontrar un buen ajuste cultural al buscar un nuevo trabajo. La dificultad para entender una cultura empresarial o la falta de alineación con esa cultura puede llevar a la insatisfacción laboral. La figura 1.3 es una recreación de un gráfico de barras que apareció en este artículo. Un gráfico de barras muestra un resumen de datos categóricos utilizando la longitud de las barras horizontales para mostrar la magnitud de una variable cuantitativa.
El gráfico mostrado en la figura 1.3 muestra el porcentaje de los 10.002 encuestados que enumeraron un factor como el más importante al buscar trabajo. Observe que nuestra atención se dirige a la barra azul oscuro, que es "Cultura de la empresa" (el enfoque del artículo). Inmediatamente vemos que solo "Salario y bonificación" se cita con más frecuencia que "Cultura de la empresa". Cuando se mira por primera vez el gráfico, el mensaje que se comunica es que la cultura empresarial es el segundo factor más importante citado por los solicitantes de empleo. Y como lector, en función de ese mensaje, decide si el artículo merece la pena leerlo.

FIGURE 1.3 A Bar Chart of Survey Results of Job Seekers



Visualización de distribuciones: histogramas, densidades, boxplots, gráficos de violines, distribuciones acumuladas

En el mundo de la estadística y el análisis de datos, la visualización de distribuciones es una herramienta esencial para comprender y comunicar información compleja de manera clara y efectiva. Los gráficos y diagramas no sólo facilitan la interpretación de los datos, sino que también permiten una exploración más profunda de las características y patrones subyacentes. En esta parte se abordarán algunas técnicas gráficas clave en la visualización de distribuciones, por ejemplo: histogramas, densidades, boxplots, gráficos de violines y distribuciones acumuladas.

Cada una de estas técnicas ofrece una perspectiva única sobre la estructura y comportamiento de los datos, permitiendo a los analistas identificar tendencias, anomalías y relaciones ocultas.

Histogramas

Con frecuencia nos encontramos con la situación en la que nos gustaría comprender cómo se distribuye una variable particular en un conjunto de datos. Para dar un ejemplo concreto, consideraremos los pasajeros del Titanic. Allí había aproximadamente 1300 pasajeros (sin contar la tripulación) y 756 de ellos con edades conocidas. Podríamos querer saber cuántos pasajeros había de determinadas edades en el Titanic, es decir, cuántos niños, adultos jóvenes, personas de mediana