

مقدمه

در دنیایی که به سرعت در حال تغییر و تحول است، درک وقایع باید بلادرنگ صورت گیرد، در غیر این صورت ممکن است زیان‌های جبران ناپذیری بوجود آید. ذهن و تفکر انسانی در درک و شناخت اجسام و اشکال قوی‌تر از اعداد و ارقام عمل می‌کند. البته این موضوع به این علت است که انتزاع اعداد برای بیان پدیده‌ها، پردازش بیشتری در ذهن انسان احتیاج دارد. برای مثال درک بلندی و کوتاهی قد دو نفر با چشم بسیار بهتر و سریع‌تر درک می‌شود تا دیدن رقم مربوط به ارتفاع قد آن‌ها. حتی ممکن است با توجه به موقعیت‌های جغرافیایی متفاوت و در کشورهای مختلف، مقیاس‌های اندازه‌گیری طول با یکدیگر متفاوت بوده و امکان مقایسه صحیح داده‌های حاصل از اندازه‌گیری قد آن‌ها نیز به راحتی میسر نباشد. برای مثال ارتفاع قد یک نفر برابر با ۵ فوت و قد نفر دوم ۱۶۷ سانتی‌متر باشد. واضح است که برای مقایسه قد این دو نفر فقط براساس ارقام دچار مشکل خواهیم شد. در این نوشتار با تکنیک و شیوه‌های مصورسازی داده آشنا شده و روش‌هایی برای توصیف اعداد و ارقام با سریع‌ترین و البته دقیق‌ترین میزان انتقال اطلاعات، ارائه خواهیم کرد. در این بین نیمی‌نگاهی هم به انواع نمودارهای آماری و اطلاعاتی داشته و خصوصیات هر یک را مرور می‌نماییم.

مصورسازی داده

نمایش تصویری اطلاعات را «مصورسازی داده (Data Visualization)» می‌نامند. در این شیوه، اطلاعات و داده‌ها، به واسطه تصاویر و شکل‌ها، قابل نمایش شده و بیننده قادر به درک سریع‌تر و بهتر اطلاعات نهفته در داده‌ها خواهد شد.

ارتباط تصویری، نگاشتنی سیستمی، بین تصاویر و مقادیر متغیرها (کمی یا کیفی) است که در مصورسازی داده به بهترین نحو صورت می‌گیرد. ویژگی‌های متنوع در نمودارها (Graphs)، آن‌ها را به ابزاری مهم برای مصورسازی داده تبدیل کرده است. وجود رنگ، اندازه‌های مختلف برای هر دنباله از داده‌ها و همچنین نمایش روند تغییرات، از مواردی است که نمودارها را برای ارتباط تصویری بهتر، نسبت به جداول و گزارشات متنی، متمایز می‌کند.

نمودارهای آماری، اینفوگرافیک، شکل‌ها و نمادها (Icons) ابزارهایی مهمی در مصورسازی داده محسوب می‌شوند. به این ترتیب، داده‌های عددی (کمی) یا اطلاعات کیفی، به صورت تصاویر، خطوط، میله یا قطاع‌هایی از دایره، تبدیل شده و اطلاعات مربوطه را منتقل می‌کنند.

ارائه یک روش موثر برای مصورسازی داده کاربران و محققین حوزه تحلیل داده را در شناخت و کشف روابط بین پدیده‌ها و داده‌های حاصل از اندازه‌گیری آن‌ها، یاری می‌رساند. به این ترتیب مجموعه داده‌های پیچیده، به شکلی ساده و موثر، نمایش داده می‌شوند و امر استخراج قواعد یا تحلیل آن‌ها آسان می‌گردد. البته اغلب برای مقایسه یا خلاصه‌کردن داده‌ها از جدول نیز استفاده می‌شود ولی زمانی که تعداد گروه‌ها برای مقایسه زیاد باشند، جدول‌ها نیز کارایی خود را از دست می‌دهند.

از دیدگاه بسیاری از صاحب‌نظران، مصورسازی داده یک هنر و در عین حال یک علم است و به عقیده بعضی، تکنیک‌های مصورسازی داده در شاخه آمار توصیفی (Descriptive Statistics) قرار می‌گیرد.

مصورسازی داده چیست؟

مصورسازی داده به روش و تکنیک‌های گفته می‌شود که به منظور برقراری ارتباط داده یا اطلاعات با بیننده با سرعت و دقت، مورد استفاده قرار می‌گیرد. تبدیل مقادیر عددی به صورت اشیاء بصری (به عنوان مثال، نقاط، خطوط یا میله‌ها) مصورسازی داده نامیده می‌شود. در مصورسازی داده هدف این است که اطلاعات را به روشنی و با میزان اثرگذاری چشم‌گیر، به کاربران ارائه کنیم.

مصورسازی داده اولین گام در مراحل تجزیه و تحلیل داده‌ها در «علم داده (Data Science)» محسوب می‌شود. «ویتالی فریدمن (Vitaly Friedman)»، مصورسازی داده و ویژگی‌ها آن اینطور توصیف می‌کند:

«هدف اصلی مصورسازی داده یا تجسم آن‌ها، برقراری ارتباط واضح و مؤثر از طریق ابزارهای گرافیکی است. این بدان معنا نیست که لزوماً مصورسازی داده باعث ایجاد یک تصویر زیبا شود بلکه درک اطلاعات به شیوه ساده و راحت منظور این روش توصیفی محسوب می‌شود. به طور مؤثر، هم فرم زیبا شناختی و هم عملکرد باید دست به دست هم دهند و با برقراری ارتباط با جنبه‌های اصلی آن به روشی بصری، اطلاعات نهفته در داده‌های نسبتاً پراکنده و پیچیده را ارائه دهند. مصورسازی داده که بدون هدف تولید شده و فقط جاذبه‌های بصری داشته باشند، منظور نظر مصورسازی داده نخواهد بود.»

در واقع، «فرناندا ویگاس (Fernanda Viejas)» و «مارتین واتنبرگ (Martin M. Wattenberg)» اظهار می‌دارند که مصورسازی داده در حالت ایده‌آل نه تنها باید به طور واضح با مخاطب و کاربر ارتباط برقرار کند، بلکه باید توجه و حساسیت بیننده را هم بر بیانگیزند.

مصورسازی داده در ارتباط نزدیک با «گرافیک اطلاعات (Information Graphics)»، «تجسم اطلاعات (Information Visualization)»، «تجسم علمی (Scientific Visualization)»، «تجزیه و تحلیل داده‌های اکتشافی (Exploratory Data Analysis)» و «نمودارهای آماری (Statistical Graphics)» است.

مصورسازی داده با نمایش اطلاعات عددی

«استفان فیو (Stephen Few)» به هشت رویکرد یا نوع اطلاعات عددی اشاره می‌کند که ممکن است کاربران را با مصورسازی داده مواجه کند. این موضوع نشانگر اهمیت ارتباط و تبادل اطلاعات بین بیننده و تصویری است که مصورسازی داده ایجاد کرده است.

- 1. سری زمانی:** اصطلاحاً سری زمانی (Time Series) برای مجموعه اطلاعاتی به کار می‌رود که وابسته به زمان بوده و بر اساس پدیده‌هایی متغیر برحسب زمان ایجاد می‌شوند. برای مثال می‌توان میزان یا نرخ بیکاری در طول یک دوره ۱۰ ساله در یک کشور خاص اشاره کرد. نمایش چنین اطلاعاتی توسط نمودار خطی (Line Chart) بسیار مؤثر است. با این کار، روند (Trend) تغییرات چنین پدیده‌ای در طول زمان مشخص می‌شود.
- 2. رتبه‌بندی:** می‌توان مقادیر عددی را برحسب ترتیب صعودی (Ascending-) یا نزولی (Descending) مرتب کرد. برچسبی که به این مقادیر مرتب شده، به صورت شماره سریال، داده می‌شود، رتبه (Rank) گفته می‌شود. البته دسته‌ها یا گروه‌هایی که از طبقه‌بندی داده‌های کیفی (ترتیبی) ساخته می‌شوند نیز دارای رتبه خواهند بود. برای مثال برحسب شعبه‌های مختلف یک فروشگاه زنجیره‌ای (متغیر طبقه‌ای) می‌توان رتبه فروش را محاسبه کرد و در یک نمودار ستونی (Column Chart) یا نمودار میله‌ای (Bar Chart) نمایش داد.
- 3. جزء نسبت به کل:** مقادیر نسبت یا نسبی از حاصل تقسیم یک جزء به کل (Part-to-Whole) ساخته می‌شوند. به این ترتیب داده‌های دو بعدی را می‌توان به صورت نسبت در آورده و در یک نمودار دایره‌ای (Pie Chart) نمایش داد. برای مثال نمایش سهم از بازار در بین چند فروشگاه زنجیره‌ای بواسطه چنین نمودارهایی، به خوبی امکان مقایسه را به بیننده می‌دهد.
- 4. نمایش اختلاف‌ها:** گاهی لازم است که یک مقدار کمی را در بین چند گروه با یکدیگر مقایسه کنیم. بخصوص زمانی که هدف مقایسه گروه‌ها با یک استاندارد یا گروه مرجع مطرح باشد. برای مثال نمودار میله‌ای (Bar Chart) برای مقایسه و نمایش اختلافات بودجه محقق شده و برنامه‌ریزی شده در حوزه‌های مختلف صنعت، خدمات و بانکداری، بسیار کارآمد عمل می‌کند.
- 5. توزیع فراوانی:** یک نمودار فراوانی (Frequency Chart) که اغلب به هیستوگرام (Histogram) معروف است، مقادیر مختلف عددی را به صورت دسته یا گروه‌هایی که شامل فاصله‌های عددی مشخص و طبقه‌بندی کرده و تعداد اعضای گروه‌ها یا فراوانی آن‌ها را شمارش می‌کند. شکل این نمودار به صورت مستطیل‌هایی با ضخامت فاصله‌های عددی و ارتفاع برحسب فراوانی ترسیم می‌شود. معمولاً این نمودارهای براساس جدول فراوانی (Frequency Table) ایجاد می‌شوند. به این ترتیب گروه‌ها از لحاظ اعضایشان قابل مقایسه می‌شوند.

همچنین «نمودار جعبه‌ای (Boxplot)» که توسط آمار شناس بزرگ «جان توکی (John Tukey)» ابداع شده، روشی دیگری برای نمایش توزیع فراوانی داده‌ها است. این نمودار، شاخص‌های دیگری مانند میانه و چارک‌ها را هم نشان می‌دهد. نمایش نقاط پرت (Outlier) از کاربردهای مهم این نمودار محسوب می‌شود. او اعتقاد دارد یک نمودار باید آن چیزی که شما با نگاه کردن به داده‌ها متوجه نمی‌شوید را در دیدتان قرار داده و شما را متعجب سازد.

6. **همبستگی:** یکی از روش‌های نمایش وابستگی بین دو متغیر، رسم نمودارهای پراکندگی (Scatter/Dot Plot) است. معمولاً در چنین نموداری، همبستگی (Correlation) بین دو متغیر به خوبی دیده می‌شود. مشخص است که در صورت همبستگی شدید بین دو متغیر، با دانستن مقدار یکی از آن‌ها، دیگری را می‌توان پیش‌بینی کرد. معمولاً متغیر مستقل در محور افقی (X) و متغیر وابسته در محور عمودی (Y) مشخص می‌شود. توابع ریاضی (Functions) یا رابطه رگرسیونی (Regression Model) در چنین نمودارهایی به خوبی، ظاهر خواهند شد.

7. **مقایسه‌های متغیرهای اسمی:** گاهی مقایسه بین دسته‌های مختلف که بوسیله داده‌هایی با مقیاس اسمی (Nominal Measure) ایجاد شده‌اند، هدف محقق است. از آنجایی که این گونه مقادیر دارای ترتیبی نیستند، نمی‌توان برای آن‌ها نمودارهای یک شکلی در نظر گرفت. ولی اغلب براساس بیشترین فراوانی می‌توان گروه‌ها را مرتب و در یک نمودار ستونی یا میله‌ای نمایش داد. برای مثال «نمودار پارتو (Pareto Chart)» یکی از این گونه روش‌ها در مصورسازی داده محسوب می‌شود.

8. **نمودارهای جغرافیایی یا نقشه‌ای:** امروزه نمایش گزارشات مرتبط با حوزه‌های جغرافیایی مختلف در نمودارهایی که شامل نقشه یک ناحیه است، تکمیل می‌شوند. به این ترتیب علاوه بر نمایش شاخص‌هایی مورد نظر، نواحی و مختصات طول و عرض جغرافیایی نیز به همراه نمودار دیده می‌شود. نمودار نقشه آماری (Cartogram) یک نمونه از این گونه نمودارها است.

تحلیلگرانی که مجموعه‌ای از داده‌ها را بررسی می‌کنند، ممکن است بیش از یک نمودار را به منظور مصورسازی داده به کار گیرند. به این ترتیب انتخاب بهترین شیوه یا روش برای نمایش داده‌ها، یک هنر محسوب شده که البته تکنیک‌های فنی نیز در آن نقش خواهد داشت. به همین علت مصورسازی داده را یک هنر و علم قلمداد کرده‌اند.

درک تصویری و مصورسازی داده

ذهن انسان می‌تواند تفاوت در طول خط، شکل، جهت‌گیری و رنگ (شدت رنگ) را به راحتی تشخیص دهد، بدون آنکه تلاش قابل توجهی برای پردازش آن‌ها صورت گیرد. این ویژگی‌ها به «خصوصیات پیش از توجه (Pre-attentive)» (Attributes) معروف هستند. به عنوان مثال، برای شناسایی تعداد دفعات تکرار رقم «5» در یک سری اعداد، به زمان و تلاش یا «پردازش توجه» زیادی نیاز است، اما اگر این رقم در اندازه، جهت یا رنگ با بقیه اعداد متفاوت باشد، تشخیص آن با سرعت توسط ذهن صورت خواهد گرفت.

نمونه‌ای از نمودارهای مورد استفاده در مصورسازی داده

در ادامه شکل ظاهری بعضی از نمودارهایی که بخصوص در مصورسازی داده به کار می‌روند، اشاره شده است. واضح است که هر یک از آن‌ها، برای بیان ویژگی خاصی به کار می‌روند.

- **نمودار میله‌ای:** مقایسه تعداد نامه‌های الکترونیکی ارسال شده (محور عمودی - داده کمی) برحسب روزهای هفته (محور افقی - داده طبقه‌ای)

نمودار فراوانی (هیستوگرام): (Histogram) - این نمودار، فراوانی گروه‌های مختلف را نمایش می‌دهد. برای مثال نمودار فراوانی قیمت خانه‌ها، شامل فراوانی (محور عمودی - شمارشی) و محور افقی دسته یا گروه‌های قیمتی (کیفی - ترتیبی) است.

نمودار پراکندگی (Scatter Plot): نمایش ارتباط بین دو متغیر کمی (عددی) در این نمودار صورت می‌گیرد. هر دو محور افقی و عمودی عددی بوده و مدل رابطه (همبستگی) بین دو متغیر نمایش داده می‌شود.

نمودار پراکندگی سه بُعدی (3D Scatter Plot): هر سه محور به صورت عددی بوده و برای نمایش توابع دو متغیره مناسب است. مثلاً $z=f(x,y)$ توسط این گونه نمودارها ترسیم می‌شود.

نمودار شبکه‌ای (Network): در این گونه نمودارها، ارتباط بین نقطه‌ها مد نظر بوده و برای نمایش گراف‌ها و نمایش راس و اضلاع یک گراف یا جنگل، که از مفاهیم ریاضیات گسسته هستند، به کار می‌رود. با رشد شبکه‌های اجتماعی، استفاده از این گونه نمودارها گسترش یافته و به این ترتیب نحوه ارتباط بین افراد و اجتماعات انسانی و شبکه‌ها، به خوبی نمایش داده می‌شود.

نمودار جریان (Streamgraph): در این نمودار روند تغییرات برای چندین متغیر (هم واحد) روی یک نمودار دو بُعدی ترسیم می‌شود. به همین دلیل استفاده از این نوع نمودار امکان مقایسه بین مقادیر در طول زمان را میسر خواهد ساخت.

نمودار نقشه درختی (Treemap): این نمودارها برای نمایش دو ویژگی همزمان در بین چندین گروه مناسب است. البته تفاوت آن با نمودار دایره‌ای در تعداد زیاد گروه‌ها است که این نمودار را بخصوص در نمایش نحوه عملکرد شرکت‌ها در بورس مورد پسند کاربران کرده است.

نمودار گانت (Gantt Chart): از نمودار گانت برای نمایش انجام فعالیت‌ها برحسب زمان صورت می‌گیرد. ارتباط بین فعالیت‌ها نیز در چنین نموداری ظاهر شده و اطلاعات زیادی در رابطه با تاریخ شروع، خاتمه و پیش‌نیازهای یک فعالیت در آن دیده می‌شود.

نقشه گرمایی (Heat Map): برای نمایش همزمان چند متغیر و رابطه بین گروه‌ها، این نمودار به خوبی عمل می‌کند. استفاده از این نمودار درست به مانند یک جدول توافقی است که می‌تواند بین دو یا چند گروه، یک یا چند شاخص محاسباتی را مقایسه کند.

گراف نواری (Strip Graphic): این نمودار دارای یک محور افقی است و مقادیر براساس اندیسی که روی محور افقی دارند، مرتب شده‌اند. رنگ‌ها نیز بیانگر شدت یا ضعف (کوچکی یا بزرگی) یک ویژگی عددی هستند.

نمودار مارپیچی (Spiral Graph): نمایش تغییرات یک پدیده در طول زمان در این نمودار توسط دایره‌های که مرکز آن مبدا یا مجموعه مرجع را نشان می‌دهد، توسط نمودار مارپیچی صورت می‌گیرد.