بسمه تعالی

موضوع : مبانی آمار و احتمال برای ورود به علم داده و تحلیل داده

بخش نهم:

- **نمودارها:**
  - هیستوگرام
  - نمودار دایره‌ای
  - Box Plot
  - نمودار روند
  - نمودار پارتو

در این پارت از اموزش به بررسی **نمودار های رایج در امار توصیفی** می پردازیم .

در اموزش های قبلی درباره **معیار های تمرکز و پراکندگی و شکل توزیع داده ها** صحبت کردیم .

تا اینجا با مفاهیم مختلف آمار توصیفی آشنا شدیم؛ از جمله میانگین، میانه، نما، واریانس، انحراف معیار، چولگی و چارک‌ها.

در حوزه آمار توصیفی بررسی شاخص های گفته شده هم به صورت جدول امکان پذیر است هم به صورت نمودار (مخصوصا در بررسی شکل توزیع ها)

بنابراین در این اموزش به بررسی نمودار های رایج و کاربردی می پردازیم :

این نمودار ها عبارتند از :

۱– نمودار box-plot :

این نمودار با مباحثی که در جلسه قبل درباره **چارک‌ها** گفتیم ارتباط مستقیم دارد.

Box Plotبه طور خلاصه اطلاعات مربوط به توزیع داده‌ها را در قالب یک نمودار فشرده نمایش می‌دهد.

در شکل کلاسیک، پنج مقدار مهم را در نظر می‌گیریم:

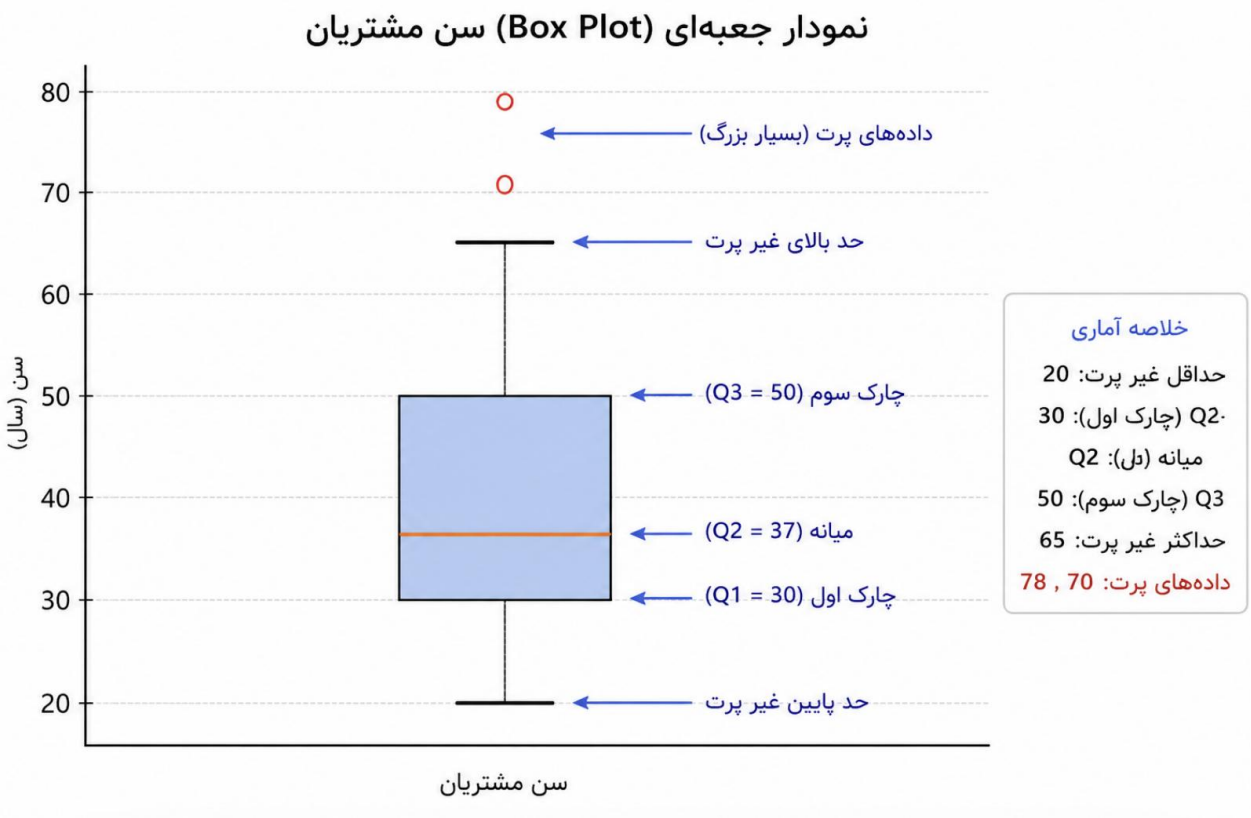
۱. حد پایین یا Minimum

۲. چارک اول Q1

۳. میانه Q2

۴. چارک سوم Q3

۵. حد بالا یاMaximum



که در این نمودار حدود پایین و بالای نمودار به کمک :

$$\text{حد پایین} = Q1 - 1.5 \times IQR$$

$$\text{حد بالا} = Q3 + 1.5 \times IQR$$

که به کمک این حدود می توان داده های پرت را نیز شناسایی کرد.

۲– نمودار هیستوگرام :

ابتدا با هیستوگرام شروع می‌کنیم.

هیستوگرام یکی از مهم‌ترین نمودارها برای نمایش توزیع داده‌های کمی است.

در هیستوگرام، داده‌های عددی را به چند بازه یا طبقه تقسیم می‌کنیم و تعداد داده‌هایی که در هر بازه قرار گرفته‌اند را نمایش می‌دهیم.

مثال : فرض کنید سن ۵۰۰ مشتری یک بانک را در اختیار داریم.

سن مشتریان می‌تواند از ۱۸ تا ۸۰ سال باشد.

اگر بخواهیم همه ۵۰۰ عدد را به صورت جدول نمایش دهیم، بررسی آن دشوار است.

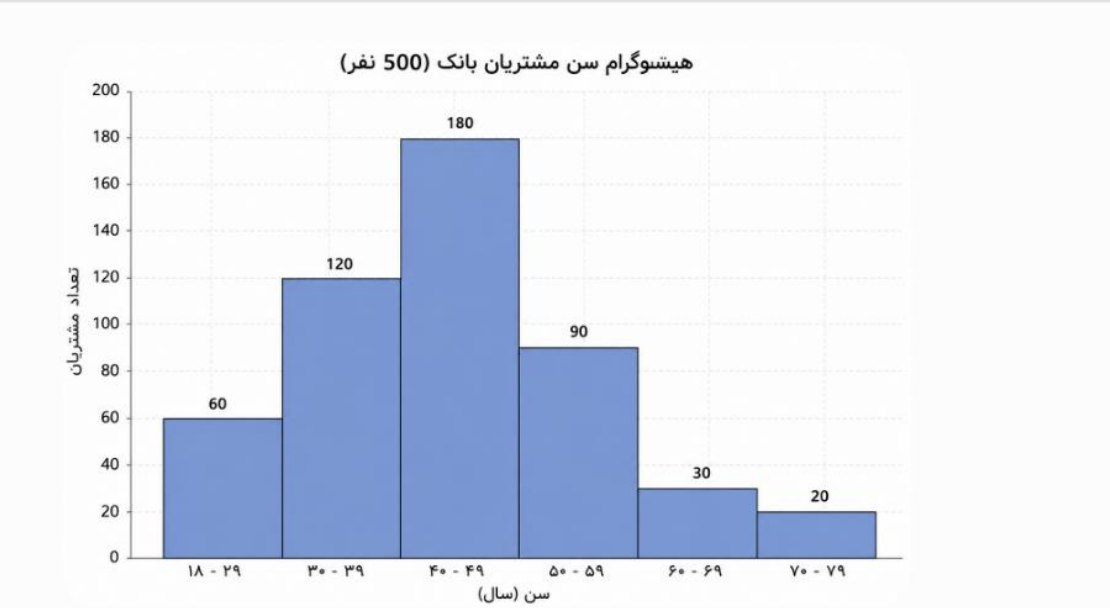
بنابراین سن‌ها را به بازه‌هایی مانند:

- ۱۸ تا ۲۹ سال
- ۳۰ تا ۳۹ سال
- ۴۰ تا ۴۹ سال
- ۵۰ تا ۵۹ سال
- ۶۰ تا ۶۹ سال
- ۷۰ تا ۷۹ سال

تقسیم می‌کنیم.

سپس تعداد مشتریان هر بازه را محاسبه می‌کنیم.

حالا این فراوانی‌ها را به صورت نمودار نمایش می‌دهیم.



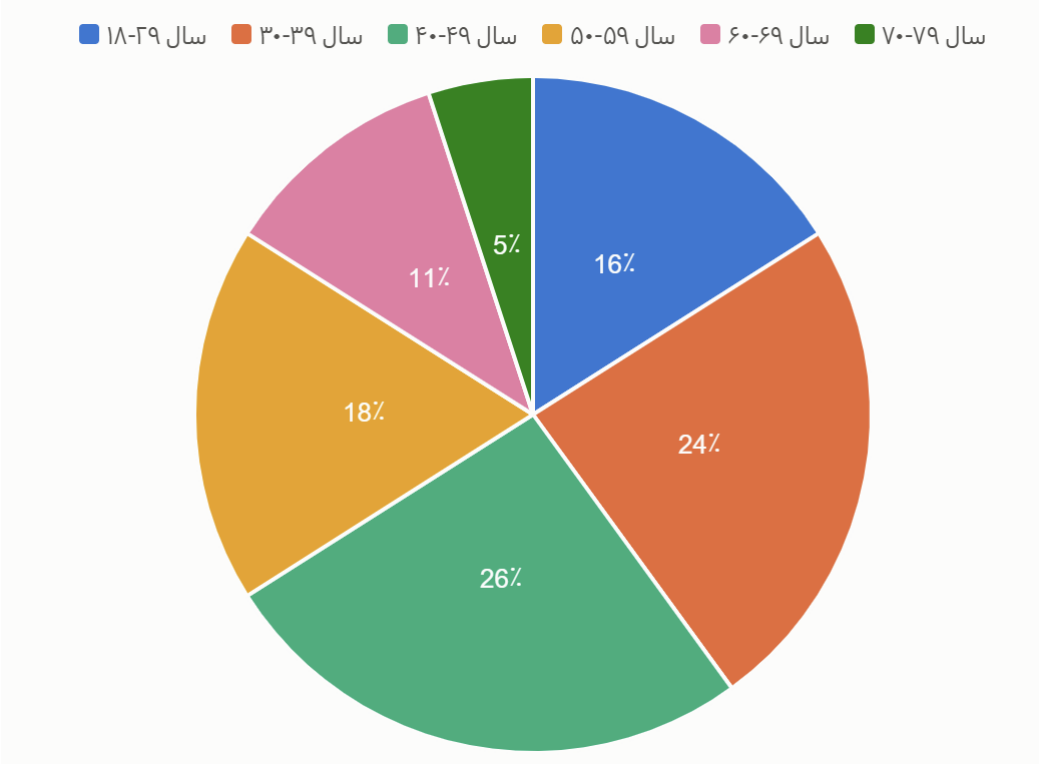
۳– نمودار دایره ای pie chart:

اگر نمودار طبقه بندی کمی داده ها را به صورت درصدی مورد بررسی قرار دهیم نمودار خروجی نمودار دایره ای خواهد بود.

نمودار دایره‌ای یکی از روش‌های نمایش گرافیکی داده‌های **کیفی (طبقه‌بندی‌شده)** است که در آن یک دایره کامل، معادل کل داده‌ها (۱۰۰٪) در نظر گرفته می‌شود و هر بخش (قطاع/برش) از دایره، نسبت یا درصد یک طبقه از داده را نشان می‌دهد.

نکته : برای داده‌های کمی پیوسته یا داده‌هایی با طبقات زیاد، pie chartگزینه مناسبی نیست (نمودار میله‌ای یا هیستوگرام بهتر است)

مثال :



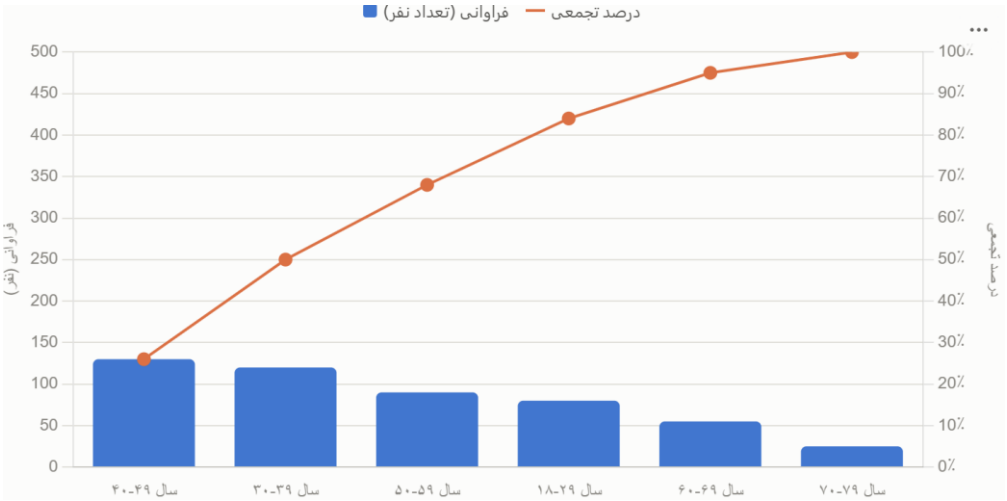
۳– نمودار پارتو :

برای توضیح این نمودار باید به اصل پارتو اشاره کرد . پارتو اقتصاددان و جامعه‌شناس ایتالیایی در اواخر قرن نوزدهم این اصل را کشف کرد. که هشتاد درصد ثروت دست ۲۰ درصد از افراد جامعه است .

به این اصل ، اصل ۲۰/۸۰ گفته می شود که بعد ها با تعمیم آن به حوزه های دیگر به این نتیجه رسیدن ۸۰ درصد از نتایج و پیامدها، از تنها ۲۰ درصد علت‌ها و تلاش‌ها به دست می‌آیند.

نموداری که این اصل را نشان می دهد که نموداری ترکیبی از **نمودار میله‌ای** (fraweb ferequency)و **نمودار خطی تجمعی (cumulative)** است. بر پایه **اصل پارتو (قانون ۲۰/۸۰)** ساخته شده: معمولاً بخش کوچکی از طبقات، بخش بزرگی از کل داده‌ها رو تشکیل می‌دن را نمودار پارتو می گویند .

که اگر بخواهیم مثال قبلی را پیاده سازی کنیم به همچین نموداری می رسیم :



- **بازه ۴۹-۴۰** پرفراوانی‌ترین گروهه (۱۳۰ نفر، ٪۲۶) و اول قرار گرفته
- خط تجمعی نشون می‌ده که فقط با **۳ بازه اول** (۴۹-۴۰، ۳۹-۳۰، ۵۹-۵۰) به ٪**۶۸** از کل مشتریان می‌رسیم — این دقیقاً همون ایده‌ی قانون ۲۰/۸۰ هست که در دنیای واقعی (مثل شکایات مشتریان یا علل نقص محصول) پرکاربردتره
- محور عمودی سمت چپ = فراوانی، محور سمت راست = درصد تجمعی (همیشه تا ٪۱۰۰)

۴- نمودار روند داده ها :

اگر بخواهیم تصادفی بودن نمونه و داده ها را بررسی کنیم و این بررسی را در یک بازه معین انجام دهیم (که معمولا این بازه براساس زمان است) از این نمودار استفاده می کنیم.

نمودار روند، تغییرات یک متغیر را در طول زمان (روز، ماه، سال، فصل و...) به‌صورت خط پیوسته نمایش می‌دهد. برخلاف pie chart و پارتو که برای مقایسه طبقات هستند، نمودار روند برای دیدن الگوی تغییر در طول زمان به‌کار می رود.

به عنوان **مثال** : نمودار روند فروش محصول به مشتریان در طول یکسال :



- **روند کلی صعودی**: از فروردین (۴۲ هزار) تا اسفند (۸۵ هزار)، تقریباً دو برابر شده
- **افت موقت** در شهریور و مهر — نمونه‌ی خوبی برای توضیح این نکته که روند همیشه یکنواخت نیست و می‌تونه **نوسان کوتاه‌مدت** داشته باشه، حتی وقتی روند بلندمدت صعودیه
- **جهش اسفند**: می‌شه با دانشجویان بحث کرد که آیا این یک **الگوی فصلی** (مثلاً افزایش خرید و تراکنش پایان سال) هست یا یک رویداد استثنایی — این تمایز بین **روند** (**trend**)، **نوسان فصلی** (**seasonality**) و **نوسان تصادفی** یکی از مباحث مهم بعدی در تحلیل سری‌های زمانیه