
بسم الله الرحمن الرحيم

پیش پردازش داده‌ها

- پاکسازی داده ها
- یکپارچه سازی داده ها
- تقلیل داده ها
- تغییر شکل داده و گسسته کردن داده

مسائل و مشکلات داده‌ها

- **کامل نبودن: فقدان مقادیر صفت، فقدان برخی صفات مطلوب**
  - مثلاً : شغل = " " (داده مفقود)
- **نویزی بودن: داده شامل نویز، خطا یا مقادیرپرت**
  - مثلاً : حقوق = "-10" (خطا)
- **متناقض بودن: وجود اختلاف در کدها یا اسامی برای مثال:**
  - سن = "42" و تاریخ تولد = "2010/07/03"
  - رتبه بندی قبلی "1,2,3" ، رتبه بندی جدید "A,B,C"
  - اختلاف بین رکورد های تکراری
- **گاهی داده مفقود به طور عمدی به شکل دیگری تبدیل شده**
  - مثلاً : اول فروردین به عنوان روز تولد همه

مسائل و مشکلات داده‌ها

- دلیل وجود داده مفقود
  - خرابی دستگاه
  - ناسازگاری با دیگر داده های ثبت شده و در نتیجه حذف آنها
  - وارد نکردن داده به دلیل نامفهوم بودن آن
  - مهم نبودن برخی داده ها در زمان ورود آنها

مسائل و مشکلات داده‌ها

با داده‌های مفقود در یک مجموعه داده چه باید کرد؟

- نادیده گرفتن تاپل: معمولاً زمانی انجام میشود که عنوان یا برچسب کلاس مفقود باشد (زمان انجام رده بندی) - اگر درصد مقادیر مفقوده در هر صفت به طور قابل ملاحظه ای اختلاف داشته باشند عملکرد ضعیف خواهد بود.
- پرکردن دستی داده مفقوده: زمان بر و غیرکاربردی
- پرکردن اتوماتیک با
  - یک ثابت سراسری: مثلاً برچسب "unknown" - یک کلاس جدید به وجود میآورد که ممکن است منجر به نتایج داده کاوی اشتباه شود
  - میانگین یا میانه صفت
  - میانگین یا میانه برای تمام نمونه های متعلق به کلاس مشابه
  - محتمل ترین مقدار: با استفاده از ابزارهای استنتاج مثل فرمول بیز یا درخت تصمیم

مسائل و مشکلات داده‌ها

داده پرت یا داده دورافتاده (Outlier) به آن دسته از داده‌ها گفته می‌شود، که فاصله زیادی با دیگر داده‌های تحقیق داشته باشد، در اصل Outlier مقادیری است که نسبت به میانگین کل داده‌ها فاصله زیادی دارد.

مشکلات داده‌های پرت

- حساس بودن روش های پارامتریک به داده‌های پرت: عدم وجود داده‌های دور افتاده برای بسیاری از آزمون های آماری پارامتریک، یک پیش فرض مهم است. شما برای اجرای این دسته از آزمون ها باید داده‌های پرت خود را حذف یا اصلاح نمایید، چرا که این داده‌ها می تواند روش اجرای شما را زیر سوال ببرد.

- ایجاد خطا در نتایج : داده‌های دور افتاده می تواند نتایج به دست آمده را به کلی دچار اشکال کند. به این مثال توجه کنید : شما می خواهید میانگین درآمدی کارکنان یک شرکت را بررسی نمایید. از 50 نفر از این کارکنان درآمدشان را می پرسید. اکثر کارکنان با درآمد 5 الی 6 میلیون در ماه هستند، ولی 4 نفر از آنان با درآمد ماهانه 40 میلیون کار می کنند. آنان مربوط به سطوح بالای سازمان هستند. حال اگر بخواهید میانگین درآمدی را بگیرید و بگویید به طور میانگین درآمد افراد چقدر است، باید بدانید که اضافه کردن آن 4 نفر به تحلیل، به شما میانگین اشتباه و غیر واقع بینانه ای را می دهد. پس راهکار این است که این 4 نفر را از تحلیل کنار بگذارید و یا مثلا برای بخش مدیریت تحلیل ها و بررسی های جداگانه ای را ارائه دهید.

- تغییر شکل توزیع متغیرها: داده‌های دور افتاده می تواند شکل توزیع نرمال را تغییر دهد. نرمال بودن توزیع یکی از پیش فرض های بسیاری از تحلیل های آماری است و داده‌های پرت می تواند این توزیع را به هم بریزد .

مسائل و مشکلات داده‌ها

دلایل ایجاد داده‌های نویزی در یک مجموعه داده:

- دستگاه های خطادار جمع آوری داده
- مشکلات ورود داده
- مشکلات انتقال داده
- محدودیت تکنولوژی
- ناسازگاری در قوانین نام گذاری

مسائل و مشکلات داده‌ها

با داده‌های نویزی در یک مجموعه داده چه باید کرد؟

- Binning
  - ابتدا داده ها را مرتب کرده و درون bin هایی با تعداد برابر افراز می کنند. بعد از آن می توان با میانگین ، میانه یا کران آن را هموار سازی کرد.
- Regression
  - هموارسازی به وسیله قرار دادن داده ها در توابع رگرسیون
- Clustering
  - شناسایی و حذف داده های پرت
- بازرسی ترکیبی انسان و کامپیوتر
  - تشخیص مقادیر مشکوک و چک کردن آن توسط کاربر

Binning:

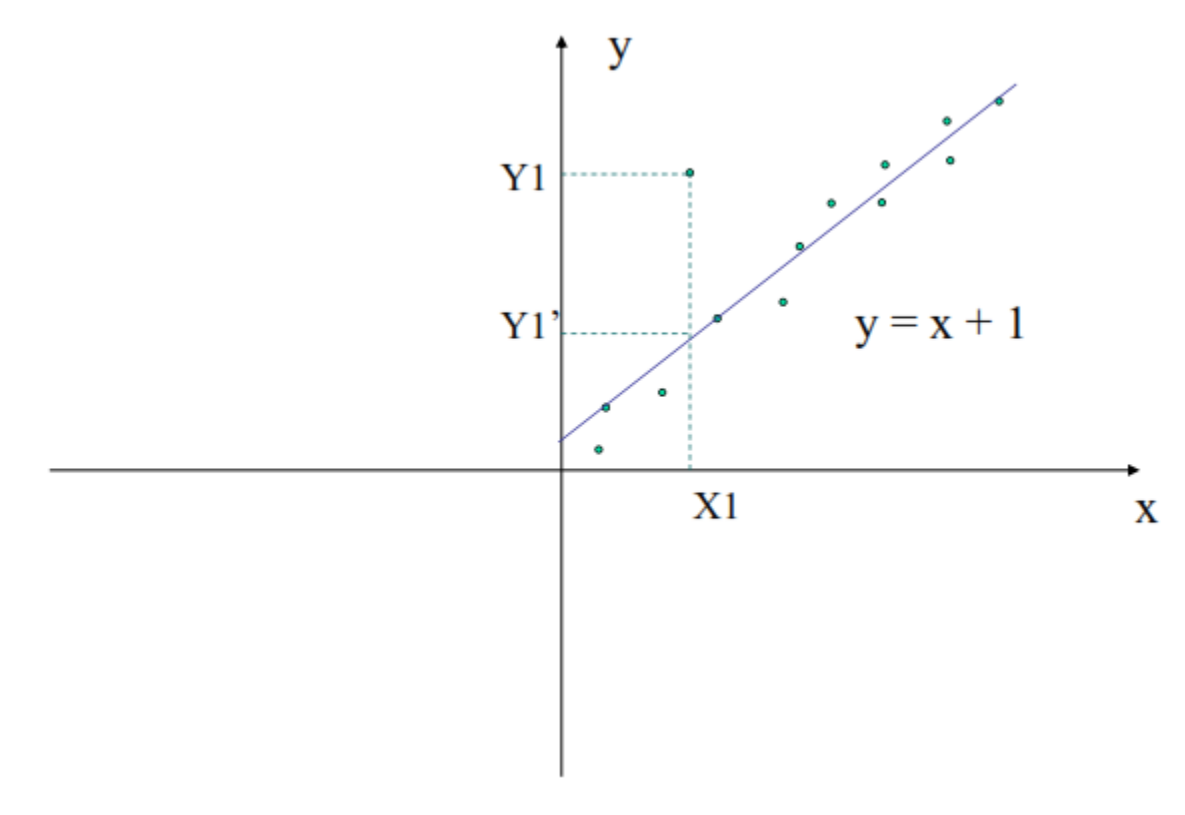
- افراز Equal-width (فاصله):
  - تقسیم دامنه به N بازه با سایز برابر
  - اگر A و B پایین ترین و بالاترین مقادیر صفت باشند، عرض بازه ها برابرخواهد بود با
    $W = (B-A)/N$
  - *ساده ترین روش است اما ممکن است داده های پرت نمایش را برهم بزنند.*

- افراز Equal-depth (فراوانی):
  - تقسیم دامنه به N بازه که هریک شامل تعداد تقریباً برابر از نمونه ها است.
  - مقیاس گذاری خوب داده ها

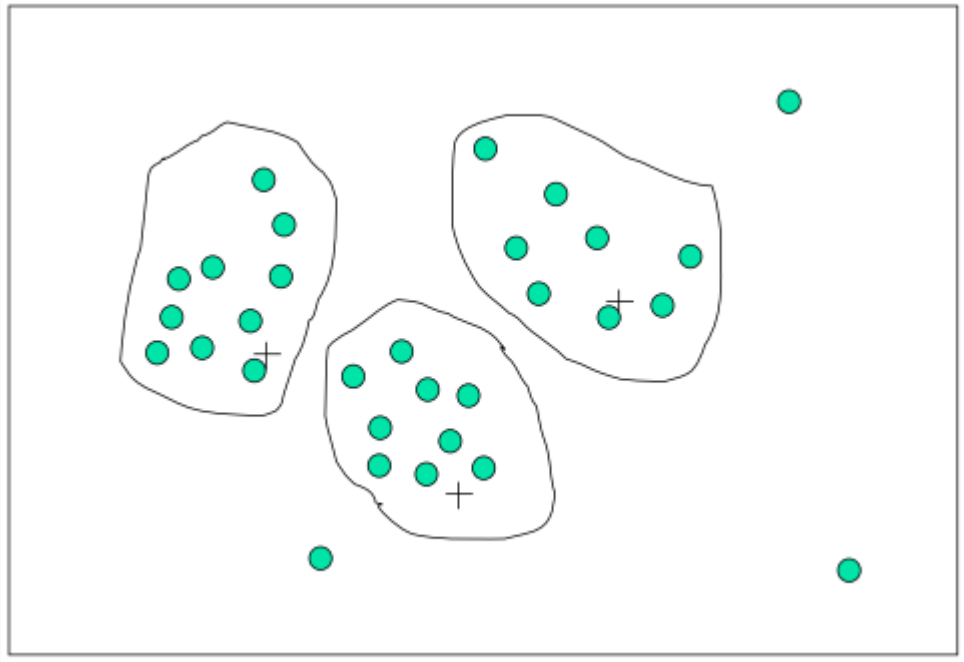
مثال از Binning:

- داده ذخیره شده برای قیمت: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- ✓ افراز به bin هایی با فراوانی برابر (equal-depth):
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
- ✓ هموارسازی با میانگین bin:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
- ✓ هموارسازی با کران bin: داده در هر دسته به هرکران نزدیک تر است آن کران قرار می گیرد.
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Regression:



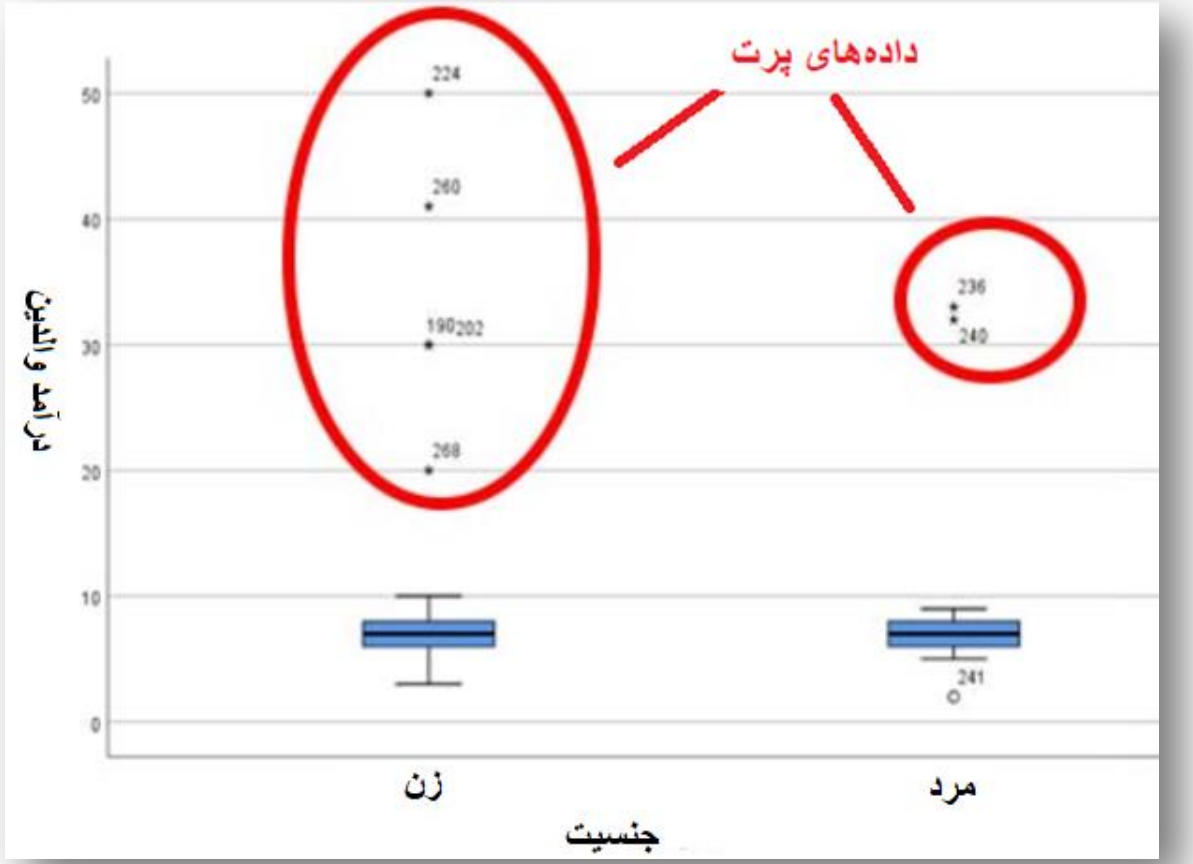
Cluster Analysis :



سایر روش‌های شناسایی داده‌های پرت:

- شناسایی داده های پرت با نمودار جعبه‌ای (Box Plot)
- شناسایی داده های پرت با نمودار میله‌ای (Bar Chart)
- شناسایی داده های پرت با نمودار پراکندگی (Scatter Plot)

شناسایی داده های پرت با نمودار جعبه‌ای

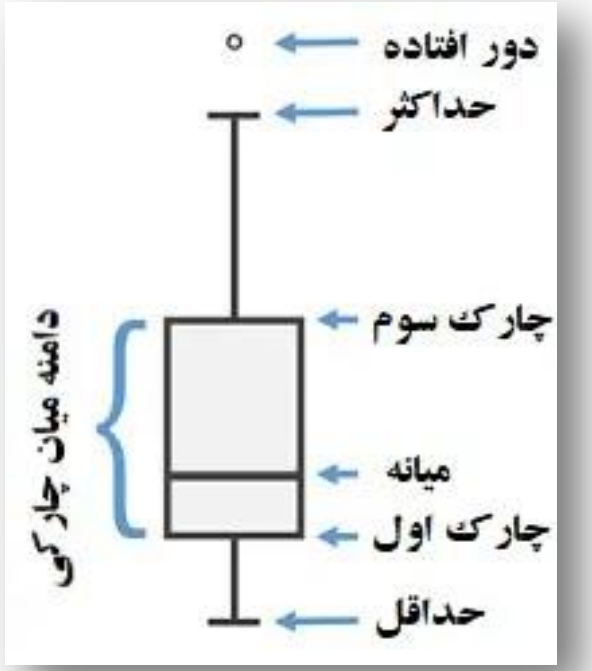


تفسیر خروجی های نمودار جعبه‌ای

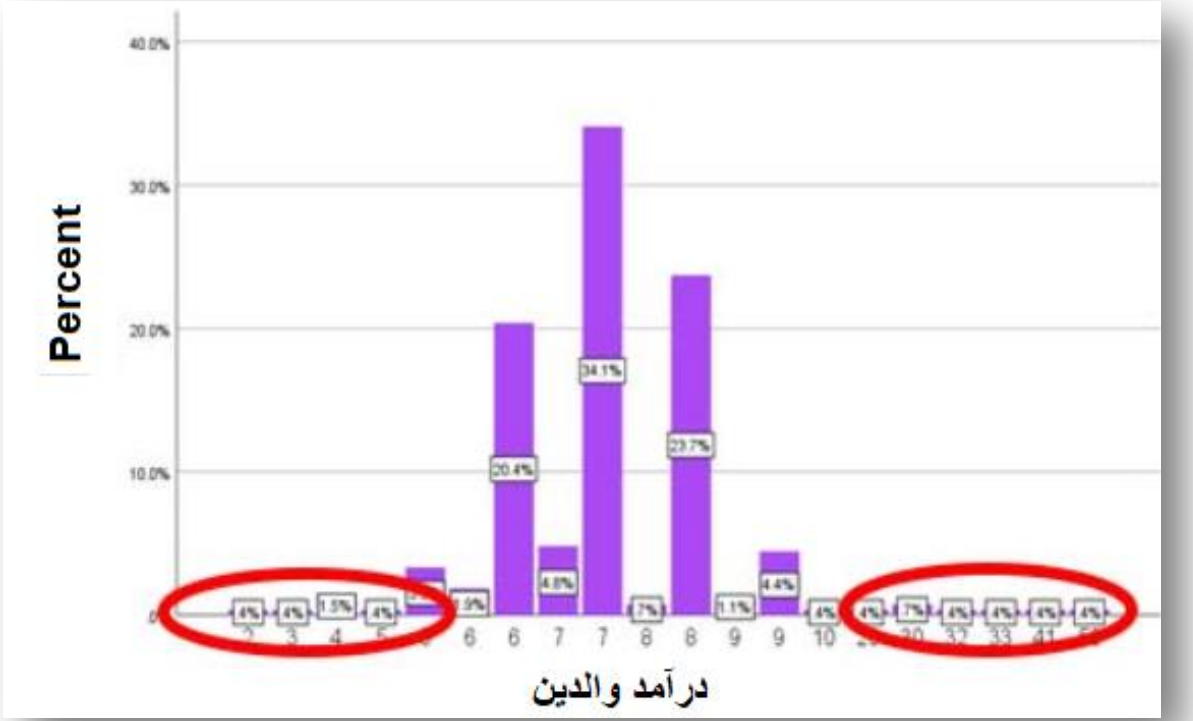
همان طور که از تصویر مشخص است، داده هایی که از توزیع ما بسیار پرت هستند را در قالب یک سری شکل دایره ای در نمودار نشان می‌دهد. بالای هر کدام از این دایره ها عدد کیس مورد نظر را نیز نوشته است. می‌توان به موارد مشخص شده در نمودار مراجعه کرد و دید چرا جواب های آن ها پرت است؟ آزمون دهنده اشتباه کرده یا آزمون گیرنده؟

❖ نمودار جعبه‌ای

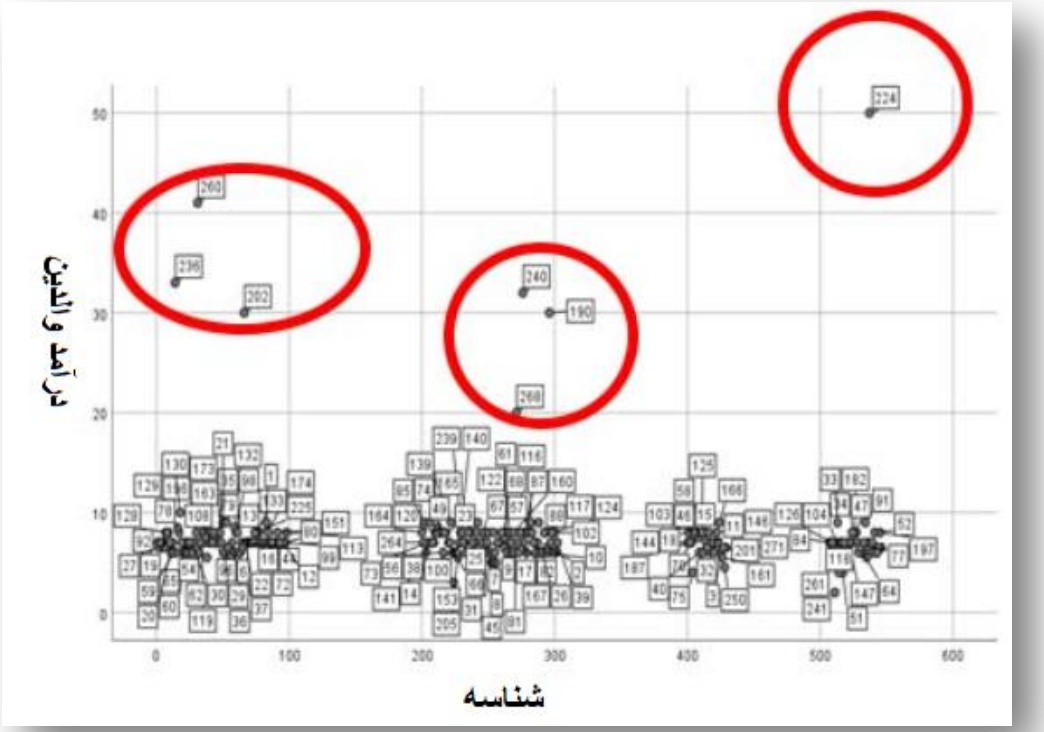
نمودار جعبه‌ای یک روش استاندارد برای نمایش توزیع داده‌ها است که براساس شاخص‌های آماری کوچکترین مقدار (Minimum)، چارک اول (First Quartile -Q1)، میانه (Median)، چارک سوم (Third Quartile- Q3) و بزرگترین مقدار (Maximum) ساخته شده است. همچنین این نمودار می‌تواند در مورد وجود داده‌های دورافتاده (Outlier) یا پرت، اطلاعاتی می‌دهد و مقدار آن‌ها را تعیین می‌کند. همچنین نشان دادن تقارن در داده‌ها از دیگر موارد قابل مشاهده در این نمودار است.



شناسایی داده های پرت با نمودار میله‌ای



شناسایی داده های پرت با نمودار پراکندگی



راهبردهای تقلیل داده

- **تقلیل داده:** داده کاوی بر روی یک مجموعه داده کاهش یابد تا به قدر کافی در تولید نتایج آماری نزدیک به داده اصلی باشد.
- چرا تقلیل داده؟ یک پایگاه داده یا انبارداده حجم عظیمی از داده ها را ذخیره میکند. تحلیل های پیچیده بر روی چنین داده هایی میتواند بسیار زمان بر بوده و چنین تحلیلی هایی را غیرکاربردی یا نشدنی کند.
- راهبردهای تقلیل داده:

1. تقلیل بُعد (dimensionality reduction)
   - تبدیل موجک
   - تحلیل مولفه های اصلی
   - انتخاب زیرمجموعه صفت
2. تقلیل تکثر(numerosity reduction)
   - مدل های رگرسیون و log-linear
   - هیستوگرام ها، خوشه بندی، نمونه برداری
   - مکعب تجمیع داده
3. متراکم سازی داده(data compression)

راهبردهای تقلیل داده

- **تقلیل داده:** داده کاوی بر روی یک مجموعه داده کاهش یابد تا به قدر کافی در تولید نتایج آماری نزدیک به داده اصلی باشد.
- چرا تقلیل داده؟ یک پایگاه داده یا انبارداده حجم عظیمی از داده ها را ذخیره میکند. تحلیل های پیچیده بر روی چنین داده هایی میتواند بسیار زمان بر بوده و چنین تحلیلی هایی را غیرکاربردی یا نشدنی کند.
- راهبردهای تقلیل داده:

1. تقلیل بُعد (dimensionality reduction)
   - تبدیل موجک
   - تحلیل مولفه های اصلی
   - انتخاب زیرمجموعه صفت
2. تقلیل تکثر(numerosity reduction)
   - مدل های رگرسیون و log-linear
   - هیستوگرام ها، خوشه بندی، نمونه برداری
   - مکعب تجمیع داده
3. متراکم سازی داده(data compression)

- **چرا تقلیل بعد؟**
  - وقتی ابعاد داده زیاد می شود، داده ها به شدت پراکنده می شوند.
  - چگالی و فاصله بین نقاط که برای خوشه بندی و تحلیل پرت ها بسیار حیاتی است، مفهوم خود را از دست میدهد.
  - به حذف صفات نامربوط و کاهش نویز کمک می کند.
  - فضا و زمان مورد نیاز برای داده کاوی را کاهش می دهد.
  - ویژوال سازی را راحت تر می کند.
- **تکنیک های تقلیل بعد**
  - تبدیل موجک (wavelet transform)
  - تحلیل مولفه های اصلی (principal component analysis - PCA)
  - انتخاب زیرمجموعه صفت

انتخاب زیرمجموعه صفت

- راهی دیگر برای کاهش ابعاد داده ها
- صفات زائد
  - همه یا بیشتر اطلاعات یک یا چند صفت تکراری باشد
  - مثلاً قیمت خرید یک محصول و میزان مالیات فروش آن
- صفات نامربوط
  - اطلاعات مفیدی برای داده کاوی در صفت موجود نباشد.
  - مثلاً شماره دانشجویی ربطی به پیش بینی معدل دانشجو ندارد

تغییر شکل داده

- داده تغییر شکل یافته یا یکپارچه می شود به طوری که نتیجه فرآیند داده کاوی کارآمدتر شده و درک الگوهای پدیدار شده آسانتر گردد.
- راهبردها:
  - هموارسازی: حذف نویز از داده. رگرسیون، خوشه بندی و binning
  - ساخت صفت
    - صفات جدیدی ساخته شده و به مجموعه صفات افزوده می شوند تا به فرایند داده کاوی کمک کنند.
  - تجمیع: عملیات خلاصه سازی یا تجمیع، ساخت مکعب داده
    - مثل محاسبه کل فروش ماهانه و سالانه از طریق تجمیع داده فروش روزانه
  - نرمال سازی: داده صفت تغییر مقیاس می یابد تا در محدوده کوچکتری جای بگیرد.
  - گسسته سازی

نرمال سازی

- واحد سنجش به کار رفته میتواند در تحلیل داده تأثیرگذار باشد. مثلاً تغییر واحدهای سنجش از متر به اینچ برای ارتفاع میتواند منجر به نتایجی کاملاً متفاوت از هم شود.
- به طورکلی بیان یک صفت در واحدهای کوچکتر منجر به یک بازه بزرگتر برای آن صفت میشود و ازاینرو تأثیر یا وزن بیشتری به آن صفت میدهد.
- برای کمک به اجتناب از وابستگی مرتبط با انتخاب واحدهای سنجش و اندازه گیری، داده باید نرمال یا استاندارد شود.
- داده تغییرشکل میدهد تا در بازه کوچکتر یا عمومی تری همانند **[-1,1]** یا **[0.0,1.0]** قرار گیرد.
- مناسب الگوریتم رده بندی درگیر در شبکه های عصبی یا سنجش فاصله و خوشه بندی است.
- روشهای نرمال سازی
  - نرمال سازی **min-max**
  - نرمال سازی **z-score**
  - نرمال سازی از طریق مقیاس دسیمال

نرمال سازی min- max

- یک تبدیل به شکل خطی برروی داده اصلی انجام میدهد.
- فرض کنید min_A و max_A نشان دهنده مقادیر حداقلی و حداکثری صفت A هستند. نرمال سازی از بازه $[min_A, max_A]$ به بازه $[new\_min_A, new\_max_A]$ با نگاشت یک مقدار v از A به v':

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- مثال: حداقل و حداکثر صفت درآمد۱۲۰۰۰ و ۹۸۰۰۰ است. میخواهیم درآمد را در بازه [0.0,1.0] نگاشت کنیم. مقدار ۷۳۶۰۰ تبدیل به ۰.۷۱۶ میشود.

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$$

- این روش رابطه بین مقادیر داده اصلی را حفظ میکند و اگر یک ورودی برای نرمال سازی خارج از بازه داده اصلی A قرار بگیرد با خطای خارج از محدوده مواجه میشود.

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

σ = population standard deviation

N = the size of the population

x_i = each value from the population

μ = the population mean

نرمال سازی

نرمال سازی z-score (یا نرمال سازی میانه صفر)

- مقادیر برای صفت A براساس میانگین و انحراف معیار A نرمال سازی می شوند.

$$v' = \frac{v - \mu_A}{\sigma_A}$$

- مثال: میانگین و انحراف معیار صفت درآمد ۵۴۰۰۰ و ۱۶۰۰۰ است. مقدار ۷۳۶۰۰ تبدیل به ۱.۲۲۵ می شود.

$$\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$$

- این شیوه زمانیکه حداقل و حداکثر صفت A شناخته نشده باشند یا زمانیکه داده های پرت چیره باشند، عملکرد مناسب تری خواهد داشت. min-max بر نرمال سازی

نرمال سازی

- **نرمال سازی از طریق مقیاس دسیمال:**
- از طریق جابجا کردن نقطه دسیمال مقادیر صفت A
- تعداد نقاط جابجا شده بستگی به حداکثر مقدار مطلق A دارد.

$$v'_i = \frac{v_i}{10^j}$$

- در اینجا j کوچکتری مقدار صحیح است به طوریکه $\max(|v'_i|) < 1$ است.
  - مثال: مقادیر ثبت شده از A در بازه 986- تا 917 است. حداکثر مقدار مطلق A برابر 986 است. برای نرمال سازی از طریق مقیاس دسیمال هر مقدار را به ۱۰۰۰ تقسیم میکنیم (یعنی j=3) .
  - 986- به 0.986- و 917 به 0.917 نرمال میشود.

گسسته سازی

- جایگزین کردن مقادیر خام یک صفت عددی با برچسب های بازه یا برچسب های مفهومی
  - مانند سن که به صورت بازه(۰-۱۰ ، ۱۱-۲۰ و ...) یا به صورت (جوان، بزرگسال، سالمند) بیان میشود
- باعث کاهش سایز داده میشود
- میتواند به صورت بازگشتی بر روی یک صفت انجام شود.