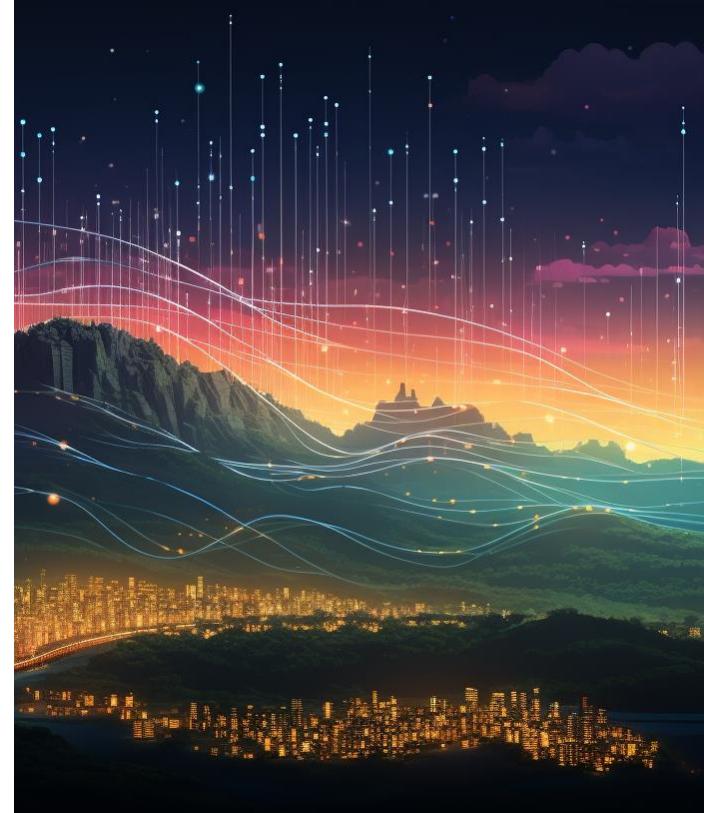




# کاریست زیان برنامهنویس R در شهرسازی و علوم اجتماعی؛ نقشه راهی برای پادگیری تحلیل داده

---

ارائه‌دهنده: مهدی سلیمانی  
با همکاری انجمن علمی مدیریت شهری دانشگاه تهران



بنام حضرت راوند جان و

«IN GOD WE TRUST.  
ALL OTHERS  
**MUST**  
**BRING**  
**DATA»**

W. EDWARDS  
DEMING



## مهندی سلیمانی

- ❖ ورودی استعداد درخشان و فارغ التحصیل مدیریت شهری دانشگاه تهران
- ❖ فارغ التحصیل تحلیل داده دانشگاه صنعتی شریف (ترم تخصصی)
- ❖ مدرس و تحلیل گر داده تائید شده از سوی کمپانی IBM
- ❖ برنده جایزه پژوهشی DAAD آلمان (سال ۱۴۰۰)



# محتوای جلسه

- ۱ مبانی تحلیل داده، اهمیت، و فرایند کلی آن
- ۲ معرفی R و سایر ابزار تحلیل داده
- ۳ معرفی کتابخانه‌های توصیفی و تحلیلی در R
- ۴ معرفی مدل‌های آماری پرکاربرد در حوزه تحلیل داده (آمار توصیفی، استنتاجی، مدل‌های خوشبندی، رگرسیون، یادگیری ماشین)
- ۵ اجرای گام به گام یک پروژه تحلیل داده
- ۶ منابع پیشنهادی برای یادگیری خودآموز

# پژوهش اول مبانی تحلیل داده، اهمیت، و فرایند کلی آن



# مبانی تحلیل داده، اهمیت، و فرایند کلی آن

## داده چیست و چه انواعی دارد؟

"مجموعه‌ای از حقایق (facts) یا آماره (Statistics) که با روش‌هایی چون مشاهده، برداشت و اندازه‌گیری برای ارجاع یا تحقیق جمع‌آوری می‌شود (Oxford Dictionary, 2010)." مهم‌ترین نوع داده عبارتست از:

داده عددی (Numeric) | -1.26, 2.37, 86, ...

داده منطقی (Logical)/باینری (Binary) | True (1), False (0)

داده گسته‌عددی (Integer) | 0, 1, 2, ...

داده رسته‌ای (Character) | "Hello", ...

داده‌های مرکب (Complex) | مانند فهرست، ماتریس، جدول و ...

سایر داده‌ها (داده رقومی، چندرسانه و...)



# مبانی تحلیل داده، اهمیت، و فرایند کلی آن

علم تحلیل داده چیست؟

"علم داده (Data Science) یک حوزه میان رشته‌ای است که از تکنیک‌های آماری و کامپیوتری برای استخراج نظاممند دانش از داده‌ها استفاده می‌کند (MIT Press, 2011)".

علم داده شامل سه مرحله کلان:

پردازش (Process)

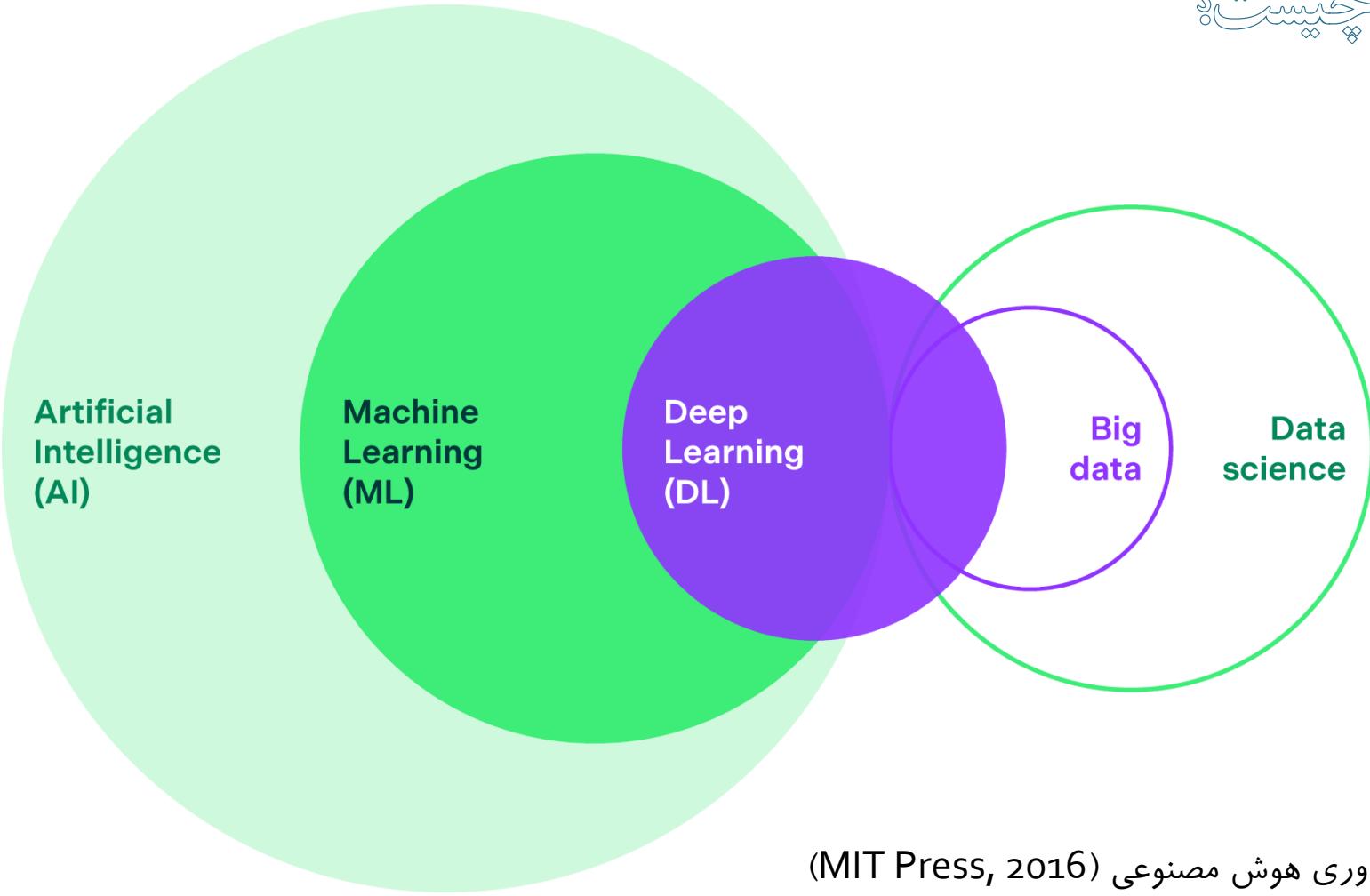
تحلیل (Analysis)

گزارش‌دهی (Report)



# مبانی تحلیل داده، اهمیت، و فرایند کلی آن

علم تحلیل داده چیست؟



ارتباط حوزه تحلیل داده با فناوری هوش مصنوعی (MIT Press, 2016)



# مبانی تحلیل داده، اهمیت، و فرایند کلی آن

## چرا پادگیری تحلیل داده اهمیت دارد؟

چون برای ما افزایش حقوق در پی دارد؟

چون باعث افزایش دانش می شود؟  
چون یک مهارت جدید می گیریم؟

# چون مجبوریم!

چون باعث تقویت رزومه و افزایش فرصت شغلی می شود؟

چون به انتشار مقالات بیشتر و بهتر کمک می کند؟



# مبانی تحلیل داده، اهمیت، و فرایند کلی آن

## فرایند تحلیل داده



# مبانی تحلیل داده، اهمیت، و فرایند کلی آن

## فرایند تحلیل داده: ۱- طرح سوال

هر نوعی از تحلیل داده، با یک یا چند سوال مشخص آغاز می‌شود که به فرایند تحلیل داده **جهت** می‌دهد. این سوال یا سوالات باید دارای ویژگی‌های زیر باشد:

- عقلانی **بوده** و از یک منطق اثبات شده در ادبیات نظری یا گزارشات عملی پیروی نماید
- به شکل مشخص بیان شده باشد و به نوعی **ورودی** و **خروجی** تحلیل را مشخص کند
- قابل سنجش **بوده** و دارای معیار، شاخص و سنجه مشخصی برای تحلیل باشد



# مبانی تحلیل داده، اهمیت، و فرایند کلی آن

فرایند تحلیل داده: ۲- سُنساپای پایگاه و ۳- استخراج داده



## دقت

میزان صحت داده  
انتخابی در بازتاب  
یک پدیده یا کلیت  
مورد بررسی



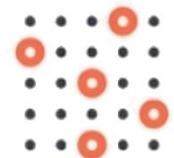
## جامعیت

میزان عدم وجود  
داده گمشده در  
مجموعه  
جمع آوری شده و  
نمایندگی از تمام  
خصوصیات جامعه



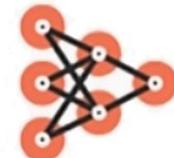
## پیوستگی

میزان عدم وجود  
تناقض یا تعارض  
اطلاعاتی در  
مجموعه  
جمع آوری شده



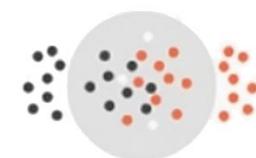
## معتبر بودن

میزان اعتبار  
اطلاعات بر حسب  
مرجع، روش، یا  
کارشناس مسئول



## تمامیت

میزان همبستگی و  
همخوانی اطلاعات  
جمع آوری شده با  
یکدیگر



## یکتایی

میزان یکتایی داده و  
عدم وجود داده  
تکراری یا همپوشان



# مبانی تحلیل داده، اهمیت، و فرایند کلی آن

## فرایند تحلیل داده: ۴- مرتب‌سازی داده (به اصطلاح *Tidy*)

به طور کلی کیفیت داده ورودی، بر زمان (سرعت)، دقت و هزینه پردازش اثرگذار است و رابطه مستقیمی با کیفیت خروجی دارد. یکی از وظایف مهندس/تحلیل‌گر اطمینان از کیفیت داده ورودی با بررسی موارد زیر است؛

- نبود داده خالی یا **Null** در مجموعه داده ورودی
- یکسان بودن نوع داده ورودی با نوع داده مورد انتظار
- همخوانی فرمت داده با استانداردهای نرم‌افزاری (مانند Unicode-08) و در نتیجه عدم استفاده از علائم پارسی

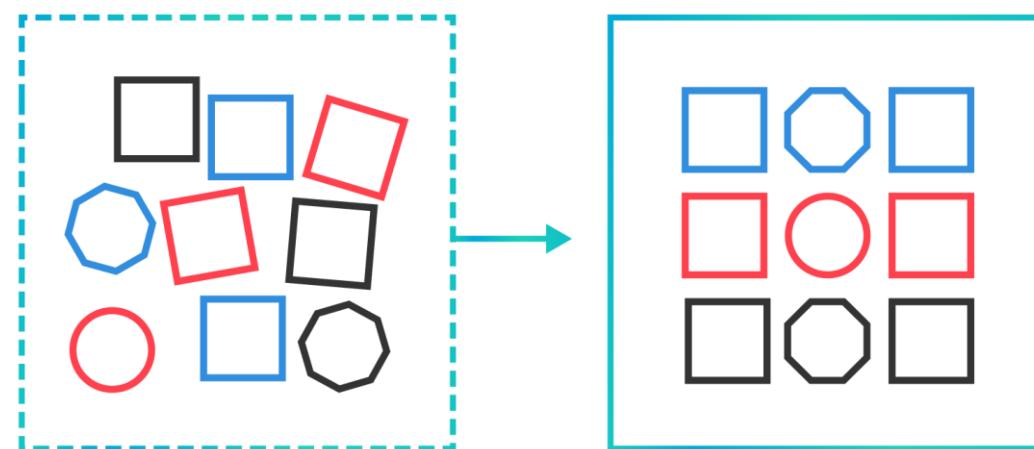


# مبانی تحلیل داده، اهمیت، و فرایند کلی آن

## فرایند تحلیل داده: ♡-قیدیل داده

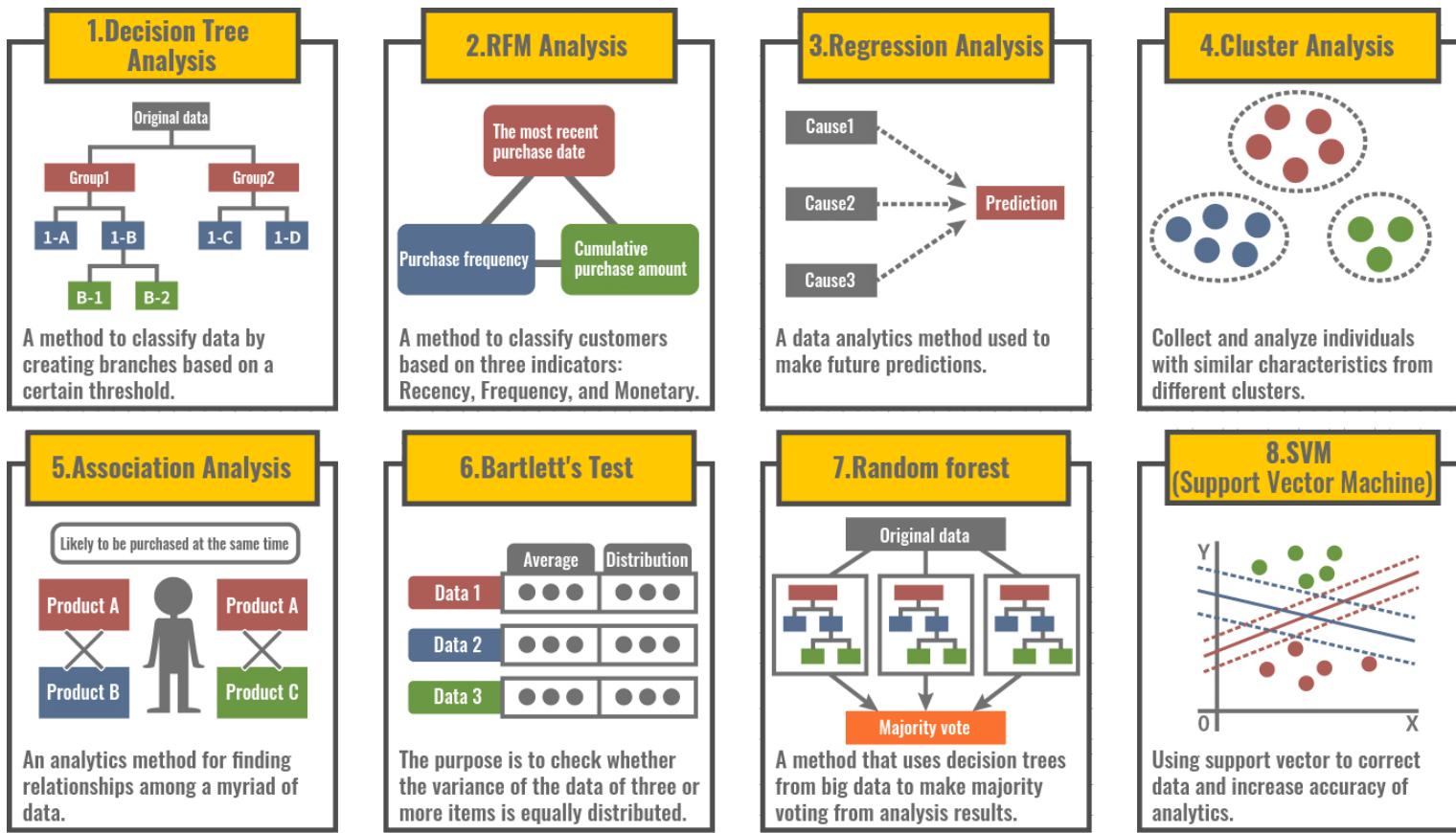
مرتب‌سازی داده برای تحلیل به شکل ساختار عرضی یا طولی:

- ◀ ساختار عرضی: برای هر مشاهده تنها یک ردیف وجود دارد و تمام شاخص‌ها به صورت ستونی ارائه شده است
- ◀ ساختار افقی: برای هر مشاهده بیشتر از یک ردیف وجود دارد



# مبانی تحلیل داده، اهمیت، و فرایند کلی آن

## فرایند تحلیل داده: ۶- تحلیل و اجرای مدل



مدل‌های متداول تحلیل کمی  
داده (MIT Press, 2016)



# مبانی تحلیل داده، اهمیت، و فرایند کلی آن

## فرایند تحلیل داده: ۷- ارائه شواهد و گزارش

با یک مقدمه شروع کنید و کلیت فرایند کار (از سوال تا روش و یافته‌ها) را به اختصار شرح دهید

شاخص‌های مورد بررسی خود را به وضوح تعریف نمایید

از نمودارهای مناسب برای ارائه نتایج خود استفاده نمایید

زبان شیوا و قابل فهم در گزارش داشته باشد و کمتر از اصطلاحات پیچیده تخصصی استفاده نمایید

نتایج و یافته‌ها را جمع‌بندی کنید

پیشنهادات خود را با دقت و بر اساس شواهد ارائه کنید

همه چیز را دوباره چک نمایید!

بخش‌ها و ویژگی‌های  
یک گزارش  
(Databox, 2023)

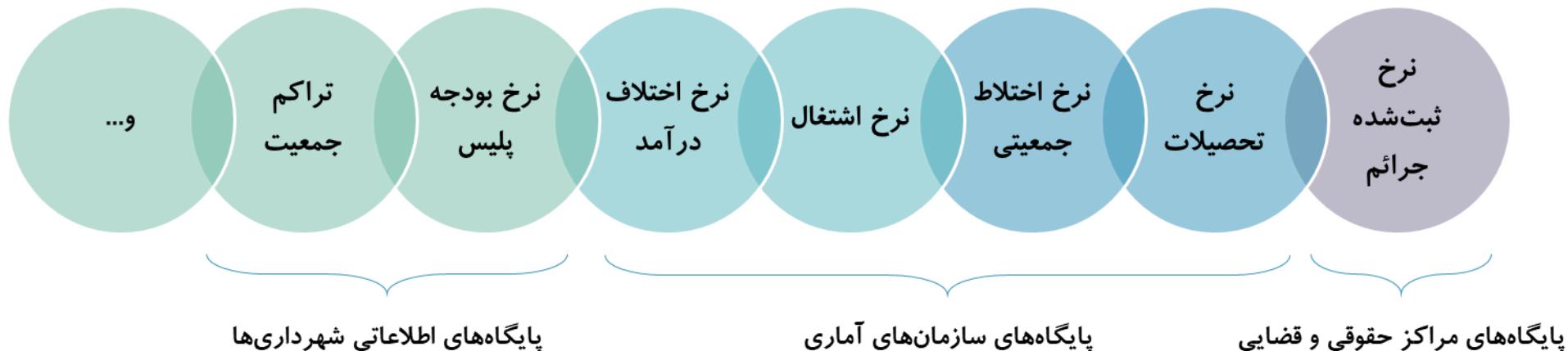


# اجرای گام به گام یک پروژه تحلیل داد

## تعریف پروژه: ۱- طرح سوال / ۲- شناسایی پایگاه داد

عنوان پژوهشی: تحلیل عوامل موثر بر رخداد جرائم خشن در ایالات متحده آمریکا

- سوال مورد تحلیل: چه عوامل اجتماعی-اقتصادی بر نرخ رخداد جرائم خشن در ایالت‌های آمریکا اثر می‌گذارد؟
- پایگاه داده و اطلاعات مورد بررسی:





# پژوهش ۲۹م معرفی R و سایر ابزار تحلیل داده



# معرفی R و سایر ابزار تحلیل داده

## تاریخچه پیدائش زبان برنامه نویسی / تحلیل داده

نرم افزار و زبان برنامه نویسی R در سال ۱۹۹۷ توسط دو آماردان سیاسی در دانشگاه اکلند نیوزلند توسعه داده شد و امروزه به مجموعه‌ای گسترده از کتابخانه‌ها و ابزار آلات تحلیل داده تبدیل شده است؛

□ دانلود نسخه پایه R: <https://cran.r-project.org/bin/windows/base/>

□ دانلود پوسته R-Studio: <https://posit.co/download/rstudio-desktop/>



# معرفی R و سایر ابزار تحلیل داده

## مزیت ها در تحلیل داده



# معرفی R و سایر ابزار تحلیل داده

## سایر ابزار تحلیل داده

Which Software Should I Choose?	Python	R	SAS	SQL
<b>Best for:</b>	General programming; Data analysis; Deep learning; Repeated tasks	Statistical analysis; Data analysis; Single passes of data	Statistical analysis; Data analysis	Database manipulating, updating, querying; Extracting, wrangling data
<b>Availability</b>	Free, open source	Free, open source	Paid (free for university edition); Closed source	Open and closed source versions available (free and paid)
<b>Easy to learn?</b>	Yes, especially for software engineers	Steep learning curve; Relatively easier if no prior coding experience	Yes, especially if you already know SQL	Relatively easy for basic level; Learning curve for more complex tasks
<b>Advantages</b>	Easy to deploy; General purpose language; Widely used by corporations	Minimal coding required for statistical models	Highly reliable, secure and stable	Very readable
<b>Disadvantages</b>	Requires rigorous testing	Very statistics oriented; Not a general-purpose program	Relatively expensive	Not general purpose: very specific, limited capability

مقایسه زبان های برنامه نویسی متداول  
تحلیل داده (IBM, 2019)





# پژوهش سوم

# معرفی کتابخانه‌های توصیفی و تحلیلی



# معرفی کتابخانه‌های توصیفی و تحلیلی در R



# معرفی کتابخانه‌های توصیفی و تحلیلی در **محیط Base.R**

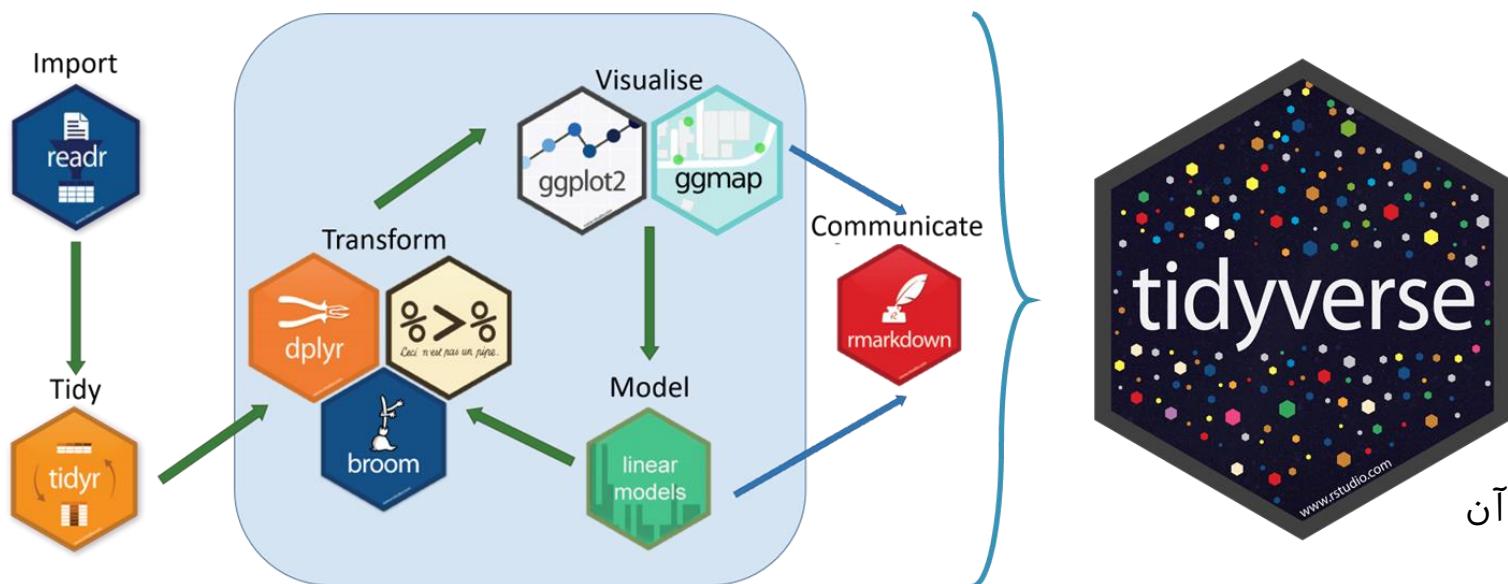
به محیط پایه در نرم افزار ، به اصطلاح Base.R گفته می‌شود و برای انجام موارد زیر مناسب است؛

- انجام عملگرهای اصلی (شامل عملگرهای محاسباتی، مقایسه‌ای، و منطقی)
- انجام پردازش‌های پایه روی داده و اطلاعات مورد نیاز
- تعریف و تخصیص متغیرها و توابع کاربردی
- فراخوانی اعضا و متغیرهای ورودی



# معرفی کتابخانه‌های توصیفی و تحلیلی در *tidyVerse* محیط

یک محیط بهینه‌شده برای پردازش و آماده‌سازی داده جهت تحلیل در R است که فرایند مراحل استخراج، مرتب‌سازی و تبدیل داده را تسهیل می‌کند و خود از چند کتابخانه (پکیج) تشکیل می‌شود



محیط *tidyVerse* و پکیج‌های آن  
(MIT Press, 2016)



# معرفی کتابخانه‌های توصیفی و تحلیلی در پکیج *readR* و کتابخانه *readr*

پکیج *readR* یکی از پکیج‌های محیط tidyVerse است که برای موارد زیر کاربرد دارد؛

- ❑ فراخوانی داده‌های جدول‌مُبنا مانند داده‌های *.csv* یا *.tsv*.
- ❑ خروجی گرفتن از نتایج و یافته‌های تحلیل به شکل جدول
- ❑ تبدیل سایر داده‌ها به داده‌های جدولی

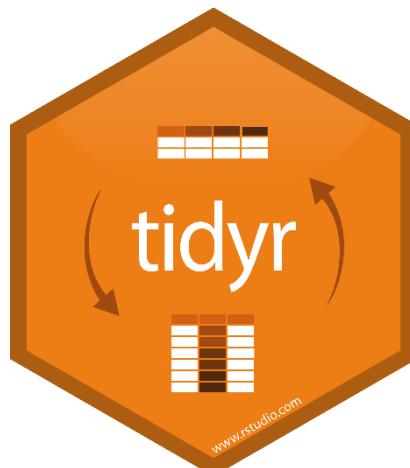


# معرفی کتابخانه‌های توصیفی و تحلیلی در R

## پکیج و کتابخانه tidyR

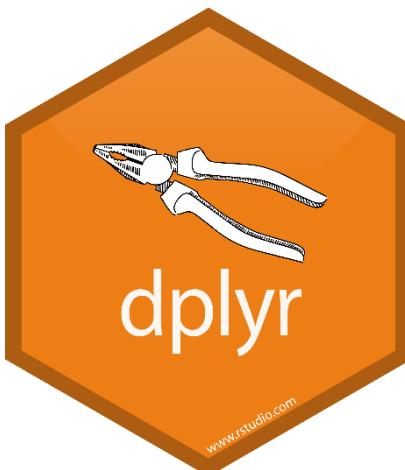
پکیج tidyR یکی از پکیج‌های محیط tidyVerse است که برای مرتب کردن یا به اصطلاح "تایدی" کردن داده از آن استفاده می‌شود. داده مطلوب یا "تایدی" شده شامل مشخصات زیر است:

- هر ستون معرف یک شاخص یا متغیر است
- مشاهدات در سطرها قرار دارند
- هر سلول تنها یک مقدار مشخص دارد



# معرفی کتابخانه‌های توصیفی و تحلیلی در پکیج و کتابخانه *dplyR*

پکیج *dplyR* یکی از پکیج‌های محیط tidyVerse است که از آن برای تغییر و پردازش داده ورودی استفاده می‌شود. مهم‌ترین کاربرد کتابخانه مذکور در تغییر و پردازش داده عبارتست از:



- ❑ فیلتر کردن مشاهدات دارای ویژگی مشخص (*filter*)
- ❑ تغییر چیدمان مشاهدات و شاخص‌ها (*arrange*)
- ❑ ساخت شاخص جدید با تلفیق شاخص‌های موجود (*mutate*)
- ❑ خلاصه کردن مشاهدات بر اساس شاخص‌ها (*summarise*)





# معرفی کتابخانه‌های توصیفی و تحلیلی در پکیج **ggplot2** و کتابخانه **tidyverse**

پکیج **ggplot2** یکی از پکیج‌های محیط tidyVerse است که از آن برای بصری‌سازی (Visualization) داده و یافته‌ها استفاده می‌شود. این کتابخانه بر اساس لایه‌های ورودی زیر کار می‌کند

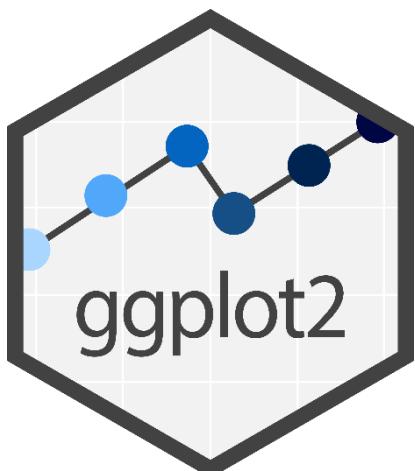
: مجموعه داده پایه برای بصری‌سازی **Data** □

: لایه ورودی برای بصری‌سازی مانند محورهای  $x$  و  $y$  **Aesthetics** □

: نوع نمودار یا شی مانند نمودار نقطه‌ای، میله‌ای و... **Geometry** □

: طیف‌های عددی و رنگی برای نمایش **Scales** □

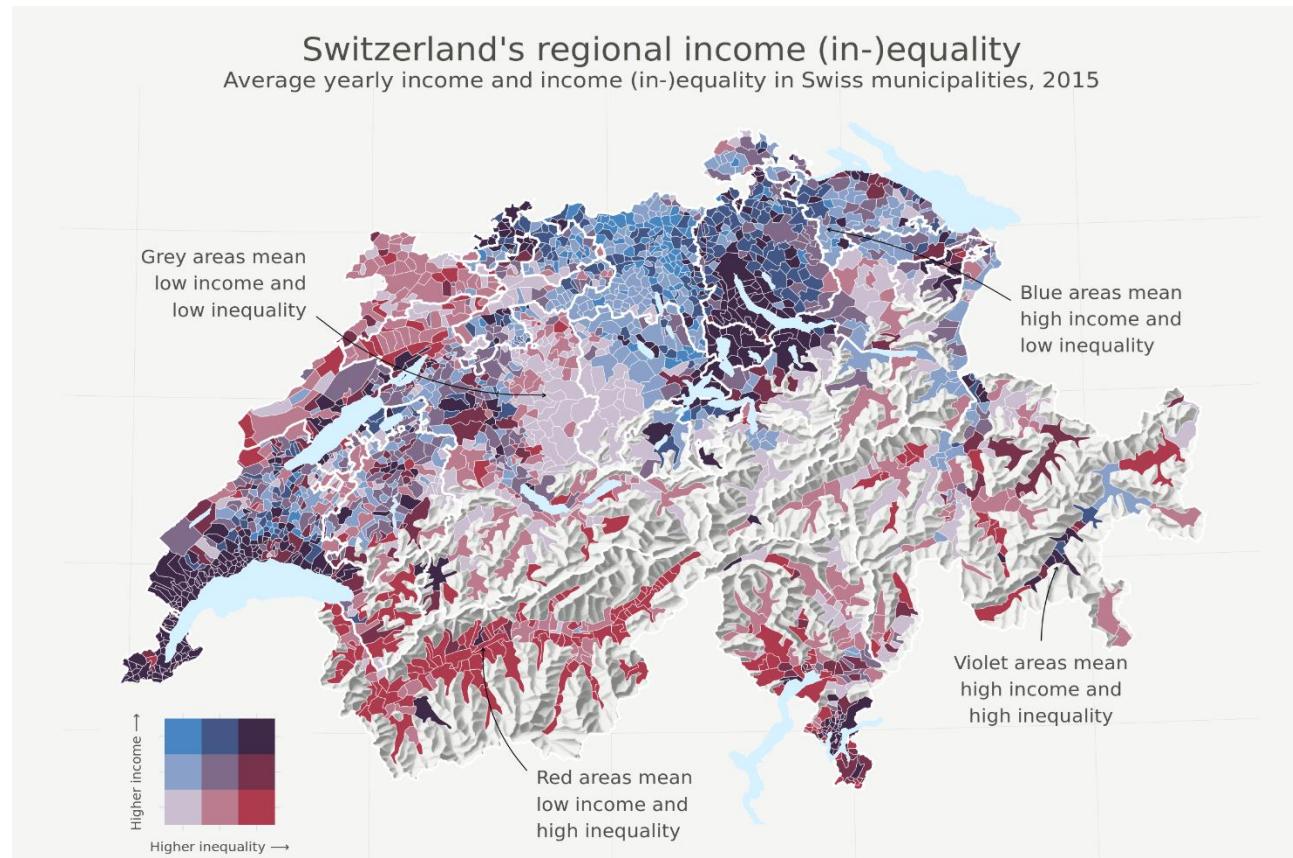
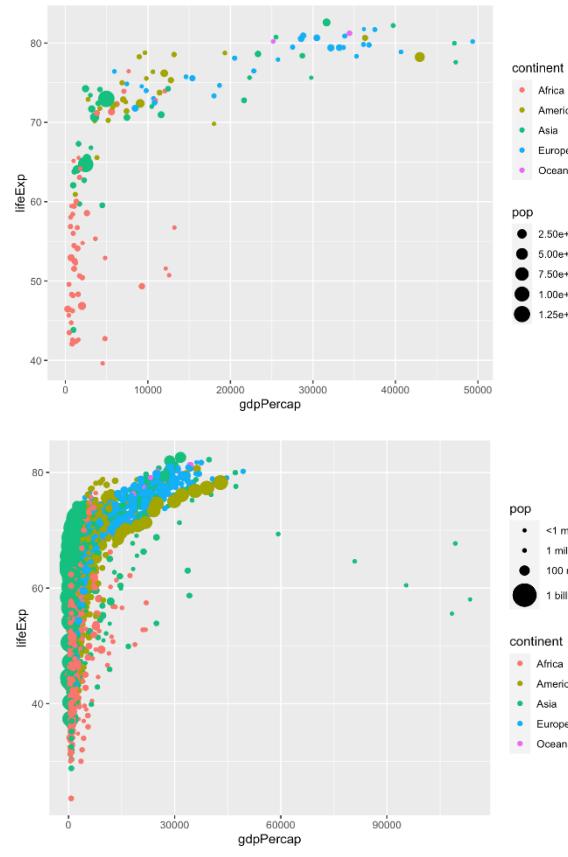
و... □





# معرفی کتابخانه‌های توصیفی و تحلیلی در

## پکیج *ggplot2* و کتابخانه



# معرفی کتابخانه‌های توصیفی و تحلیلی در پکیج و کتابخانه R-ArcGIS Bridge



پکیج R-ArcGIS Bridge یک پکیج ارتباطی جدید است که در سال ۲۰۲۳ توسط کمپانی ESRI توسعه داده شده و نیازمند دسترسی به نسخه اصلی نرمافزار ArcGIS Pro برای موارد زیر است؛

- برقراری ارتباط میان محیط نرمافزارهای R و ArcGIS Pro
- تسهیل فراخوان داده‌های فضایی (رسترن و وکتوری) در R
- ایجاد امکان تحلیل فضایی (مانند رگرسیون فضایی) در R

# اجرای گام به گام یک پروژه تحلیل داده

## نصب و فراخوان کتابخانه‌های مورد نیاز

نصب پکیج *tidyverse*

یا

نصب کتابخانه *ReadR*: برای وارد کردن اطلاعات

نصب کتابخانه *GGPlot2*: برای مصورسازی اطلاعات

محیط پایه R: برای مدل‌سازی و گزارش‌گیری



# اجرای گام به گام یک پروژه تحلیل داده

## بارگیری و فرآخوان داده (۲- استخراج، ۳- مرتبسازی، ۴- قبیل)

متغیر وابسته	متغیرهای مستقل	سنجه	داده / اطلاعات
		تعداد رخداد جرائم خشن در هر ایالت به ازای هر ۱ میلیون نفر	نرخ رخداد جرائم خشن
		تعداد مردان ۱۴ تا ۲۴ سال به ازای هر ۱۰۰۰ نفر	نوع توزیع سنی
		ده برابر متوسط تعداد سالهای تحصیل کرد جمعیت بالا ۲۵ سال	نرخ تحصیلات
		میزان هزینه کرد پلیس (دلار) به ازای هر نفر در سال ۲۰۱۴	سرانه هزینه کرد پلیس در سال ۲۰۱۴
		میزان هزینه کرد پلیس (دلار) به ازای هر نفر در سال ۱۹۵۹	سرانه هزینه کرد پلیس در سال ۱۹۵۹
		تعداد نیروی کار شاغل به ازای هر ۱۰۰۰ نفر مرد بین ۱۴ تا ۲۴	نرخ نیروی کار
		تعداد مردان به ازای هر ۱۰۰۰ زن ساکن	نسبت جنسیتی
		جمعیت ایالت در واحد ۱۰۰ هزار نفر	اندازه ایالت
		تعداد ساکنان غیرسفیدپوست به ازای هر ۱۰۰۰ نفر	نرخ مهاجران و اقوام غیرسفیدپوست
		تعداد بیکاران به ازای هر ۱۰۰۰ نفر	نرخ بیکاری
		متوسط دارایی قبل انتقال خانوار بر حسب واحد ۱۰ دلار	اندازه ثروت
		تعداد خانوار با درآمد پائین‌تر از نصف میانگین درآمدی به ازای هر ۱۰۰۰ خانوار	نابرابری درآمدی



# پنجشی چهارم

## معرفی مدل‌های آماری پرکاربرد در حوزه تحلیل داده



# کاربردها کلی مدل‌های آماری در تحلیل داده

به طور کلی مدل‌های آماری در تحلیل داده سه کاربرد اساسی دارد؛

- |                                                               |                                                                       |                           |
|---------------------------------------------------------------|-----------------------------------------------------------------------|---------------------------|
| آمار توصیفی (Descriptive): مشاهده و توصیف کلیت یک جامعه آماری | آمار استنباطی (Inferential): شامل نمونه‌گیری، استنتاج/استنباط و تعمیم | تحلیل وضع موجود           |
| {                                                             |                                                                       | □ شناخت عوامل اثرگذار     |
| }                                                             |                                                                       | □ پیش‌بینی آینده‌های ممکن |



# کاربردها کلی مدل‌های آماری در تحلیل داده

استخراج فرضیه از داده موجود، شامل اکتشاف الگوریتم‌ها، تولید فرضیه و مدل‌سازی

تحلیل اکتشافی

بررسی و تائید/رد فرضیه‌ها و مدل‌های ساخته شده

تحلیل تاییدی



# آمار توصیفی

شاخص‌های آماری که هدف آن‌ها نمایش گرایش کلی یک مجموعه داده و روند کلی آن است. مهم‌ترین شاخص‌های گرایش به مرکز عبارتست از؛ میانگین، میانه، و مود

شاخص‌های حول مرکز

این شاخص‌ها نشان می‌دهد که داده موجود تا چه اندازه متنوع است یا از مرکزیت فاصله دارد. مهم‌ترین شاخص‌های پراکندگی عبارتست از؛ واریانس و دامنه بین چارکی

شاخص‌های پراکندگی



# آمار توصیفی

## شاخص‌های گرایش به مرکز

□ میانگین ( $\bar{x}$ ): حاصل تقسیم مجموع داده بر تعداد آنها که می‌تواند به شکل میانگین ساده (حسابی)، وزن‌دار (هندسی) و متحرک باشد. میانگین به وجود مقادیر پرت (extreme) در مجموعه حساس است.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

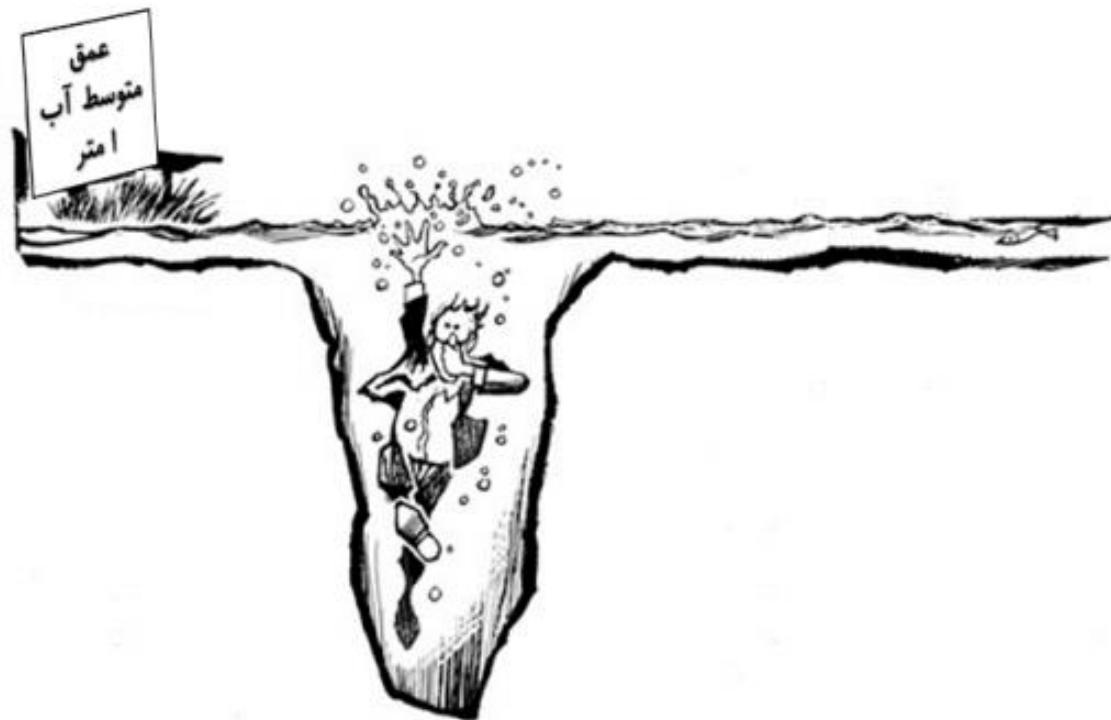
□ میانه: زمانی که نقاط داده به ترتیب صعودی یا نزولی مرتب شوند، میانه مقدار وسطی در یک مجموعه داده است. میانه بر خلاف میانگین به وجود داده پرت حساس نیست. میانه داده را به دو قسمت مساوی تقسیم می‌کند

□ مد: مقداری است که بیشتر از سایر مقادیر در یک مجموعه داده رخ می‌دهد.



# آمار توصیفی

## شاخص‌های گرایش به مرکز



تلهی میانگین و شاخص‌های گرایش به مرکز (MIT Press, 2016)

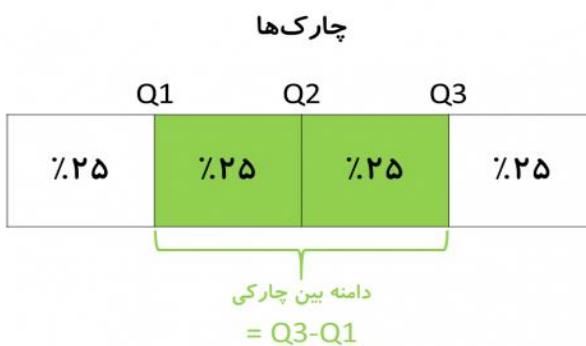


# آمار توصیفی

## شاخص‌های پراکنده

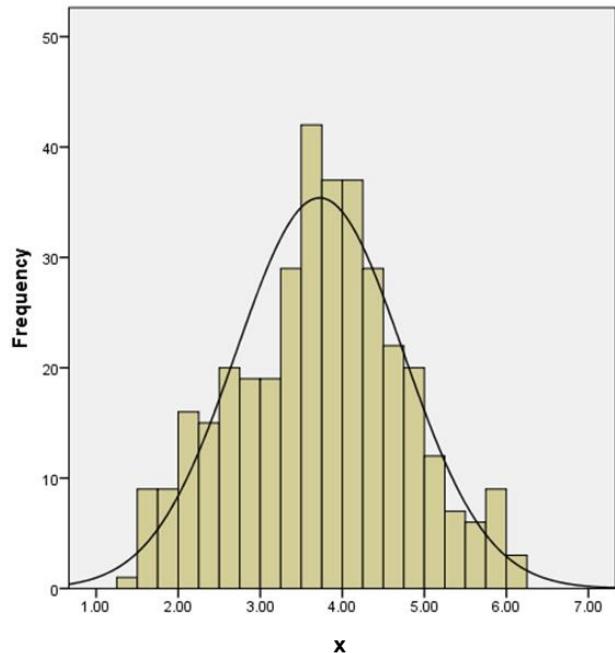
□ واریانس: واریانس به زبان ساده میانگین مجذور فاصله داده‌ها از مرکز آن‌هاست.

□ چارک و دامنه بین چارکی (IQR): چارک داده‌ها را به چهار قسمت مساوی تقسیم می‌کند؛ هر بخش شامل ۲۵ درصد مشاهدات است. میانه چارک دوم محسوب می‌شود. دامنه بین چارکی اختلاف سقف چارک سوم و کف چارک دوم است. دامنه بین چارکی به شناسایی و حذف داده پرت کمک می‌کند



# آمار توصیفی

## هیستوگرام و نمایش فراوانی



نوعی نمودار ستونی برای نمایش فراوانی هر دسته از داده است که توزیع داده بر اساس فراوانی را نشان می‌دهد

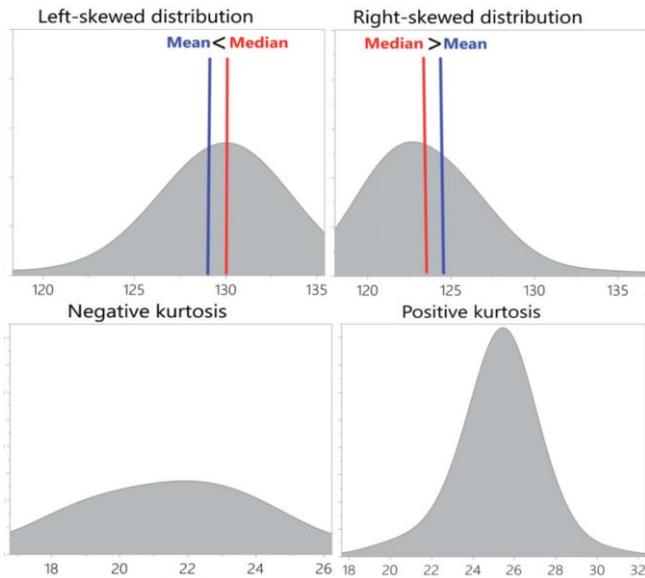
□ محور افقی در هیستوگرام شامل دسته‌هایی با طول یکسان است که به آنها سبد (Bin) گفته می‌شود

□ محور عمودی در هیستوگرام معرفی میزان فراوانی هر دسته است که می‌توان به شکل اسمی یا نسبی ارائه شود

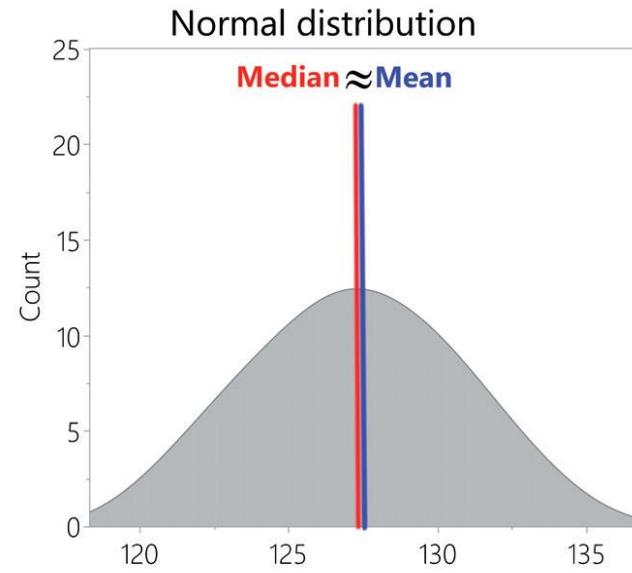


# آمار توصیفی

## توزیع نرمال و غیرنرمال



توزيع غیرنرمال:



توزيع نرمال:

- اختلاف اندک میانگین و میانه
- تقارن نمودار توزیع فراوانی
- متداول‌ترین الگوی توزیع
- چولگی (Skewness)
- کشیدگی (Kurtosis)
- اگر  $-5 < \text{skewness} < +2$  و  $-2 < \text{kurtosis} < +5$  باشد توزیع نرمال است!



# آمار استنتاجی

نوعی از مدل‌های آماری/یادگیری ماشین است که در آن، داده از پیش‌ تعیین‌ شده‌ای برای نظارت و مقایسه خروجی وجود ندارد. مانند مدل‌های خوشبندی

استنتاج بدون نظارت

در این مدل‌های آماری/یادگیری ماشین، مجموعه داده برچسب‌ گذاری شده وجود دارد که بر صحت و دقت خروجی نظارت می‌کند. مانند مدل‌های رگرسیون، مدل KNN و جنگل تصمیم

استنتاج با نظارت

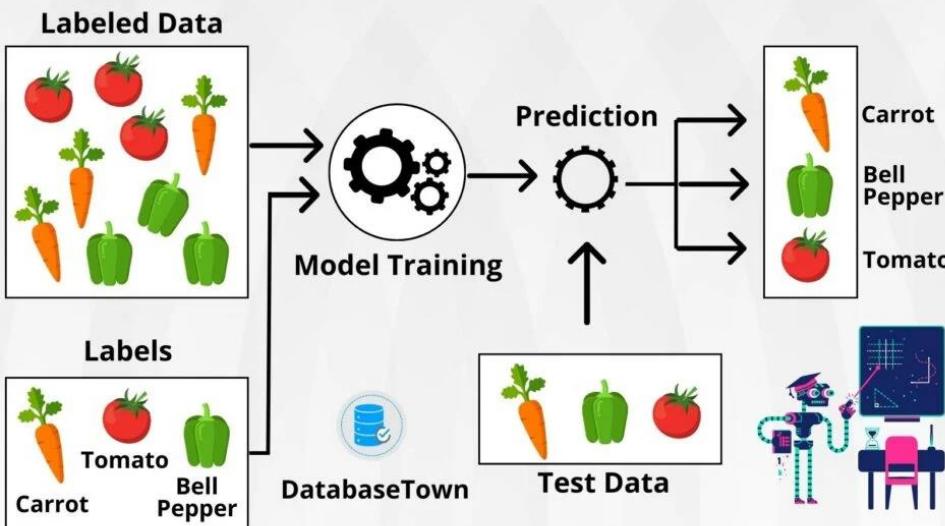


# آمار استنتاجی

## یادگیری بدون نظارت و با نظارت

### SUPERVISED LEARNING

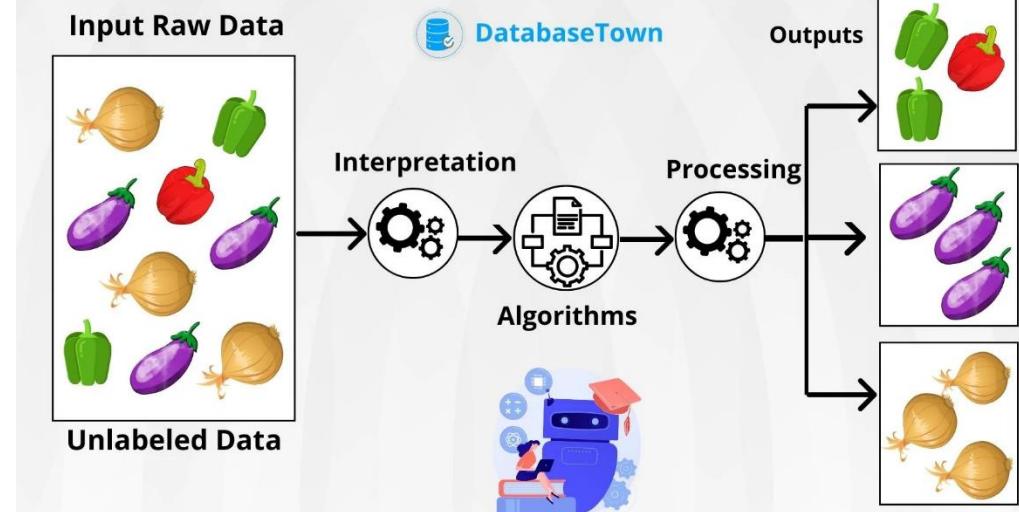
Supervised machine learning is a branch of artificial intelligence that focuses on training models to make predictions or decisions based on labeled training data.



یادگیری با نظارت (Supervised)

### UNSUPERVISED LEARNING

Unsupervised learning is a type of machine learning where the algorithm learns from unlabeled data without any predefined outputs or target variables.



یادگیری بدون نظارت (Unsupervised)



# استنتاج/یادگیری بدون ناظارت

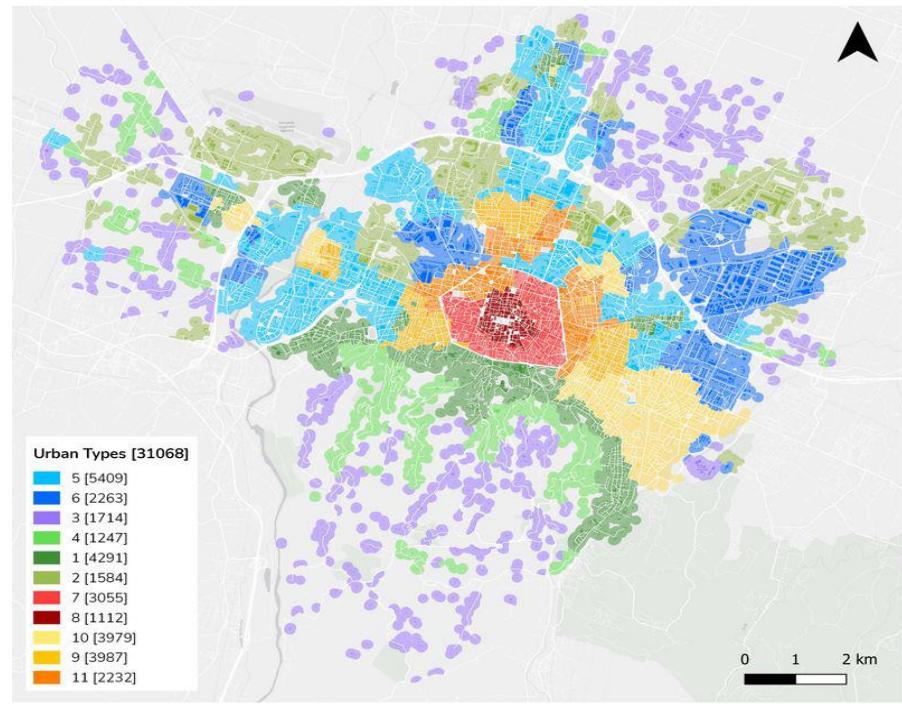
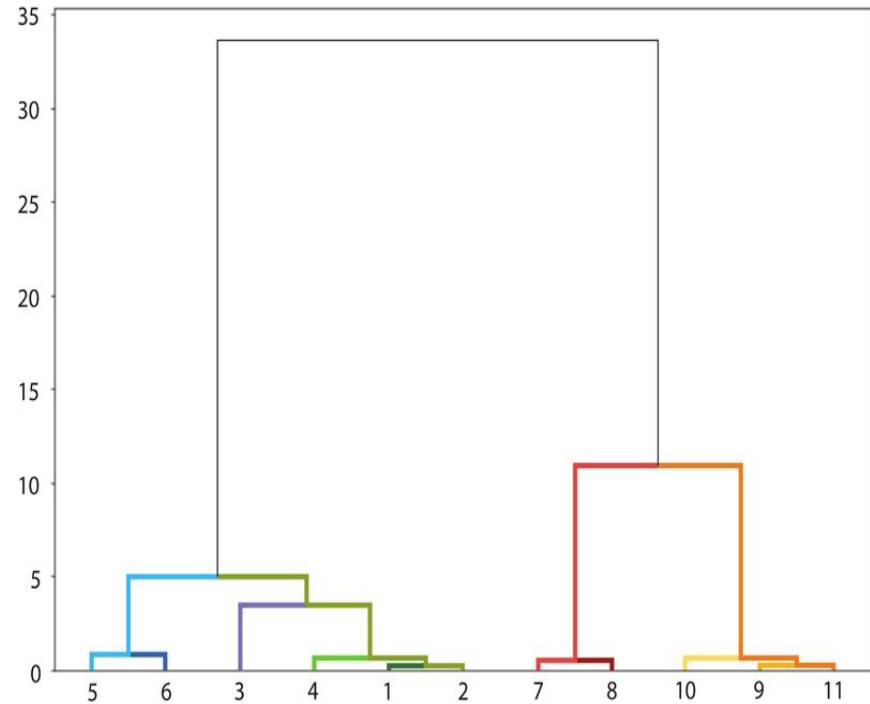
## (*Cluster Analysis*) تحلیل خوشه‌ای

- ❑ نوعی تحلیل اکتشافی بدون ناظارت است که هدف اصلی آن دسته‌بندی داده موجود و نمایش ساختار کلی مجموعه داده ورودی می‌باشد
- ❑ از تحلیل خوشه‌ای برای **خلاصه‌سازی مشاهدات** بر حسب شاخص‌ها استفاده می‌شود. به عبارت دیگر مشاهداتی که در یک خوشه قرار می‌گیرند، خصوصیات یکسانی بر حسب شاخص‌ها دارند
- ❑ کاربرد اصلی تحلیل خوشه‌ای در شهرسازی **تحلیل مورفومتری** (ریخت‌شناسی شهر) و در علوم اجتماعی، شناسایی پایگاه‌های اقتصادی-اجتماعی است



# استنتاج/یادگیری بدون ناظرت (*Cluster Analysis*)

خوشبندی بلوک‌های  
شهر بولونیا بر اساس  
شاخص فرمی، شامل  
ارتفاع، سطح اشغال، عرض  
معبر، کاربری و...  
(Porta et al., 2022)



# استنتاج/یادگیری با ناظارت

## مدل‌های رگرسیون

- مدل‌هایی از استنتاج با ناظارت هستند که ارتباط/میزان اثرگذاری یک یا چند متغیر مستقل ( $X$ ) را بر یک متغیر وابسته یا پاسخ ( $y$ ) تحلیل می‌کنند
- کاربرد اصلی مدل‌های رگرسیون در بررسی آن است که متغیر مستقل  $X$  چقدر و چگونه بر متغیر وابسته  $y$  اثر می‌گذارد تا بتوانند با کمک  $X$ ، شدت و احتمال رخداد  $y$  را پیش‌بینی کنند
- در شهرسازی و علوم اجتماعی از مدل‌های رگرسیون برای شناسایی تاثیر عوامل مختلف محیطی و غیرمحیطی بر پدیده‌ها (مانند تاثیر فرم بر شدت گرما) و پیش‌بینی ناهنجاری‌ها (مانند جداسدگی اجتماعی) استفاده می‌شود
- مدل‌های رگرسیون متعددی تاکنون توسعه داده شده‌اند که در دو دسته رگرسیون خطی و غیرخطی قرار می‌گیرد. انتخاب مدل رگرسیون باید با توجه به هدف تحقیق، تعداد مشاهدات، و نوع متغیر وابسته ( $y$ ) انجام شود



# استنتاج/یادگیری با ناظارت

## مدل‌های رگرسیون

Method	Dependent variable	Independent variable	Test statistics
Linear Regression	Continuous	Continuous + Categorical	R2
Logistic Regression	Binary	Continuous + Categorical	R2
Poisson/Negative binomial Regression	Count	Continuous + Categorical	R2
Multinomial logistic Regression	Categorical	Continuous + Categorical	R2

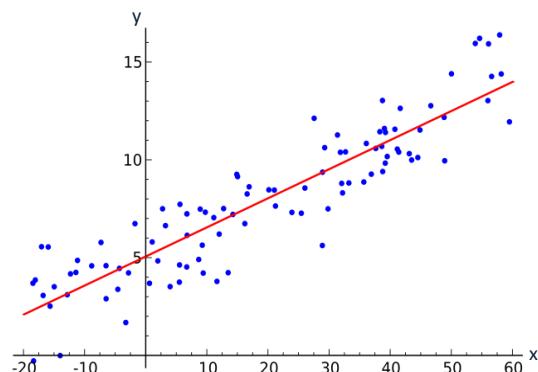
مدل‌های رگرسیون  
مناسب با توجه به نوع  
متغیر پاسخ



# استنتاج/یادگیری با نظرات

## مدل‌های رگرسیون - رگرسیون خطی

- از مدل رگرسیون خطی زمانی ( $lm$ ) استفاده می‌شود که متغیر وابسته یا پاسخ ( $y$ ) از جنس عددی پیوسته باشد مانند دما، جمعیت، مساحت و...
- خروجی مدل رگرسیون خطی، ضرایب رگرسیون ( $\beta$ ) هست که شدت تاثیر هر یک از متغیرهای مستقل ( $X$ ) بر متغیر وابسته ( $y$ ) را نشان می‌دهد.



$$y = mx + b$$
$$Y = \beta_0 + \beta_1 X + \epsilon$$

single value of dependent variable  
slope  
single value of independent variable  
y-intercept  
all observed values for dependent variable  
y-intercept aka "bias"  
slope aka "coefficient"  
all observed values of independent variable  
error\*  
 $\alpha$



# استنتاج/یادگیری با ناظارت

## مدل‌های رگرسیون - رگرسیون خطی

مثال: خروجی زیر حاصل از اجرای مدل رگرسیون خطی بر یک مجموعه داده به جهت شناسایی ارتباط قد کودکان (متغیر وابسته) با سن آنها، بر حسب ماه، و تعداد فرزندان خانواده (متغیرهای مستقل) است.

```
call:  
lm(formula = height ~ age + no_siblings, data = ageandheight)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-0.28029 -0.22490 -0.02219  0.14418  0.48350  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 64.95872   0.55752 116.515 1.28e-15 ***  
age          0.63516   0.02254  28.180 4.34e-10 ***  
no_siblings -0.01137   0.05893  -0.193   0.851  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.2693 on 9 degrees of freedom  
Multiple R-squared:  0.9888,    Adjusted R-squared:  0.9863  
F-statistic: 397.7 on 2 and 9 DF,  p-value: 1.658e-09
```

❖ مطابق این ضرایب معادله زیر برای قد کودک بر حسب سن و تعداد فرزندان خانواده برقرار است:

$$64.95 + (\text{تعداد فرزندان خانواده}) * 0.63 - (\text{سن کودک}) * 0.01 = \text{قد کودک}$$

❖ بر اساس آماره *p-value* چون ارتباط معناداری بین تعداد فرزندان خانواده با قد کودک وجود ندارد از آن صرف نظر می‌شود

❖ در نتیجه، پیش‌بینی می‌شود که قد یک کودک ۲ ساله (۲۴ ماهه) برابر باشد با:

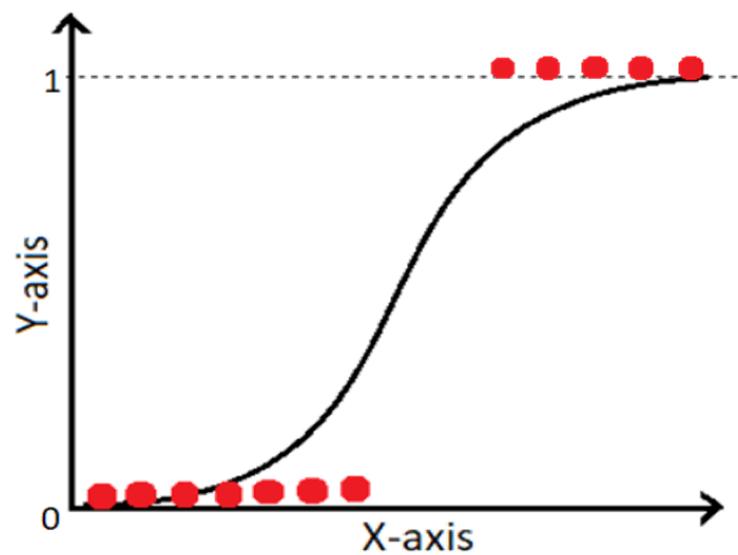
$$64.95 + 0.63 * 24 = 80.07 \text{ سانتی‌متر} = \text{قد کودک}$$



# استنتاج/یادگیری با ناظارت

## مدل‌های رگرسیون - رگرسیون غیرخطی (رگرسیون لجستیک ساده)

- از مدل رگرسیون غیرخطی زمانی (*glm*) استفاده می‌شود که متغیر وابسته یا پاسخ (y) از جنس غیرپیوسته باشد مانند متغیرهای باینری (صفر و یک)، ترتیبی (مثال: رتبه مالی شهرها)، دسته‌بندی (مثال: پایگاه‌های اجتماعی-اقتصادی)
- رگرسیون لجستیک ساده، یکی از مدل‌های رگرسیون غیرخطی است که از آن زمانی استفاده می‌شود که متغیر وابسته (y) از جنس کمیت‌های باینری (صفر و یک) باشد.



- خروجی رگرسیون لجستیک، ضریب احتمال است که نشان می‌دهد متغیرهای مستقل چه تاثیری بر احتمال رخداد متغیر وابسته دارد!



# استنتاج/یادگیری با ناظارت

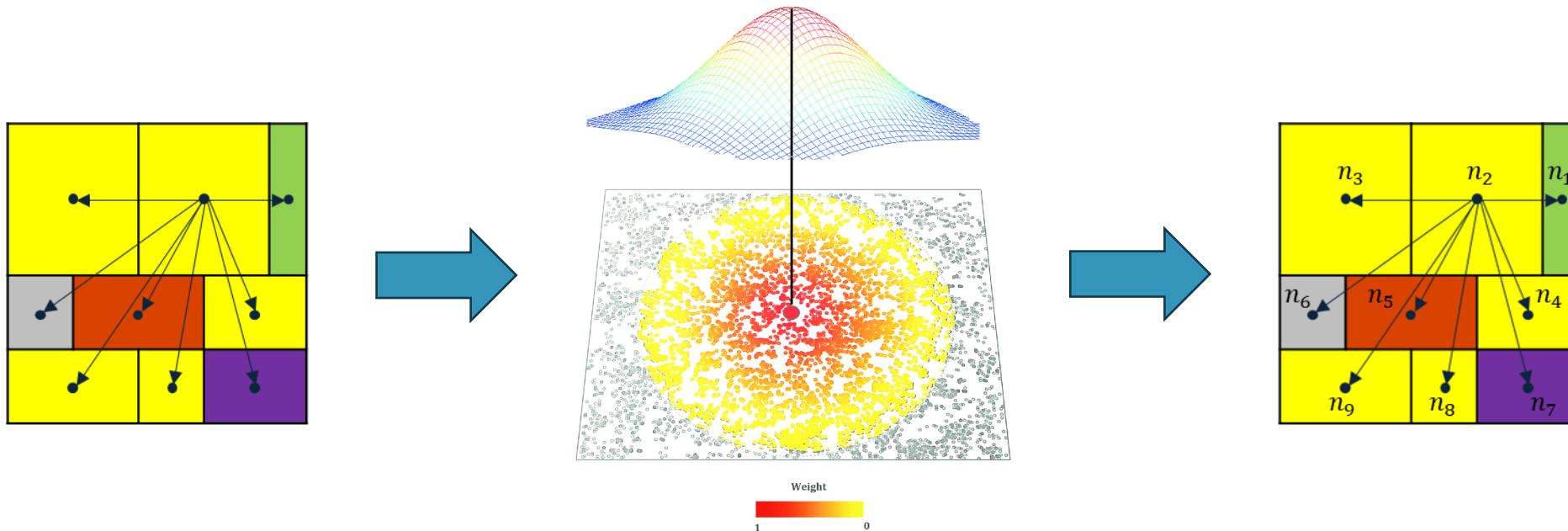
## مدل‌های رگرسیون - رگرسیون فضایی

- در مدل‌های رگرسیون فضایی، عامل مکان (یا فاصله) متغیرها به عنوان یک عامل سوم علاوه بر متغیر وابسته ( $y$ ) و متغیرهای مستقل ( $X$ ) به مدل اضافه خواهد شد
- در رگرسیون فضایی، تحلیل روابط بر اساس یک ماتریس ضرایب فضایی ( $W$ ) انجام می‌شود، به این صورت که رابطه متغیرها در مشاهداتی که به یکدیگر نزدیک‌تر هستند، شدیدتر است
- مدل‌های رگرسیون فضایی (مانند Spatial Lag یا MGWR) کاربرد بسیار گسترده‌ای در حوزه شهرسازی دارد و از آن می‌توان در تحلیل عوامل موثر بر قیمت مسکن، تاثیر فرم شهری بر حرارت و... استفاده کرد



# استنتاج/یادگیری با ناظارت

## مدل‌های رگرسیون - رگرسیون فضایی



۵۴

۱۴۰۳/۰۲/۰۵

انجمن علمی مدیریت شهری دانشگاه تهران | کاریست زبان برنامه‌نویسی R در شهرسازی و علوم اجتماعی؛ نقشه راهی برای یادگیری تحلیل داده | ارائه‌دهنده: مهدی سلیمانی



# استنتاج/یادگیری با نظرات

## p-Value و R2 - آماره مدل‌های رگرسیون

```
call:  
lm(formula = height ~ age + no_siblings, data = ageandheight)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-0.28029 -0.22490 -0.02219  0.14418  0.48350  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 64.95872  0.55752 116.515 1.28e-15 ***  
age          0.63516  0.02254  28.180 4.34e-10 ***  
no_siblings -0.01137  0.05893  -0.193   0.851  
---  
signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 0.2693 on 9 degrees of freedom  
Multiple R-squared:  0.9888, Adjusted R-squared:  0.9863  
F-statistic: 397.7 on 2 and 9 DF, p-value: 1.658e-09
```

در تفسیر نتایج حاصل از اجرای مدل‌های رگرسیون به ترتیب  
باید به سه آماره توجه کنیم!

آماره **p-value** یا مقدار احتمال: به دو شکل کلی و به تفکیک متغیرهای  
مستقل ارائه می‌شود. اگر کمتر از ۰.۰۵ باشد حاکی از ارتباط معنادار و  
وجود نتایج قابل تفسیر است

آماره **R2** و **Adjusted R2**: ضریبی بین ۰ تا ۱ است که میزان تعریف  
کنندگی مدل را نشان می‌دهد. هر چه این عدد بالاتر باشد، به این معناست  
که متغیرهای مستقل (X) بهتر می‌توانند متغیر پاسخ (Y) را تعریف کنند

ضریب رگرسیون: این ضریب در رگرسیون‌های خطی یک عدد حقیقی است  
که بزرگی آن شدت اثرگذاری و علامت آن (+ یا -) جهت اثرگذاری  
متغیر مستقل بر متغیر پاسخ را نشان می‌دهد



# استنتاج/یادگیری با ناظارت

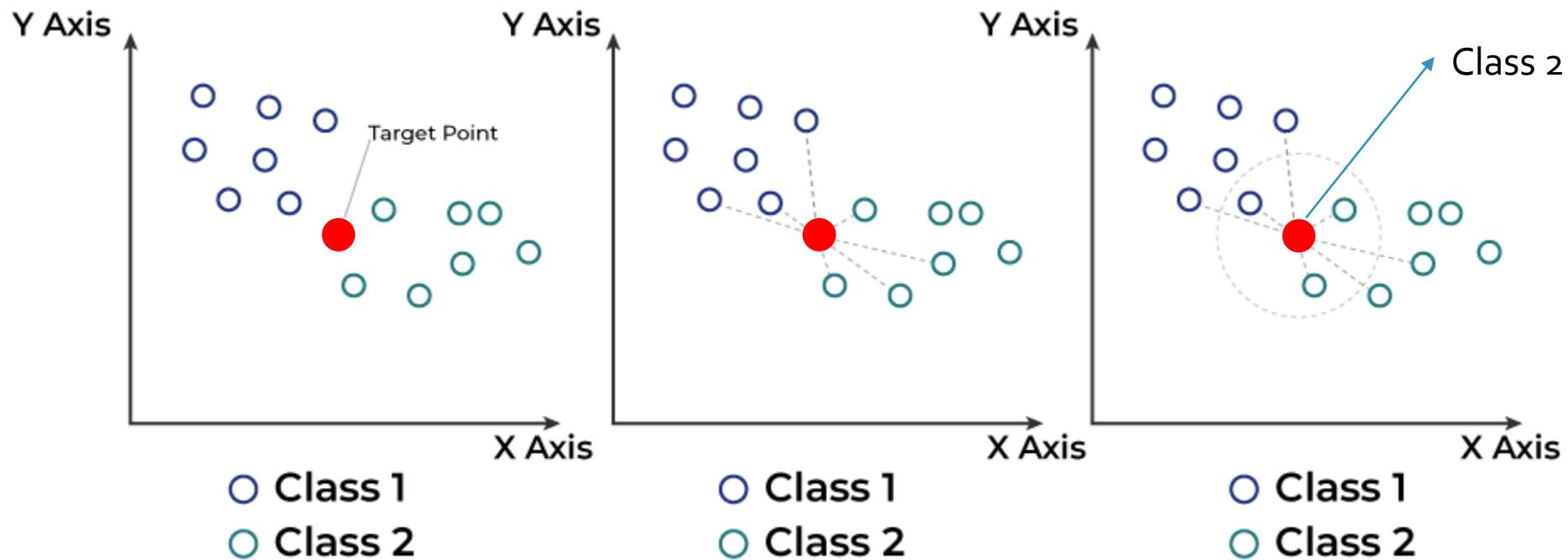
## تحلیل **KNN** (یادگیری ماشین)

- یکی از الگوریتم‌های یادگیری با ناظارت، الگوریتم **K Nearest Neighborhood** یا **KNN** است که از آن برای پیش‌بینی داده‌های مجهول بر اساس داده‌های معلوم مجاور استفاده می‌شود
- الگوریتم **KNN** خود شامل مجموعه‌ای از مدل‌های آماری می‌شود که کاربرد اصلی آنها شناسایی **K** همسایه نزدیک یک نقطه و تعیین وضعیت آن نقطه بر اساس وضعیت همسایگانش است
- کاربرد الگوریتم **KNN** در شهرسازی بیشتر در حوزه **پردازش تصاویر، درون‌یابی (مانند تولید خطوط کانتور از روی نقاط ارتفاعی)** و پیش‌بینی الگوی توسعه **شهرها** بر اساس هسته‌های جمعیتی اطراف است



# استنتاج/یادگیری با ناظارت

## تحلیل KNN (یادگیری ماشین)



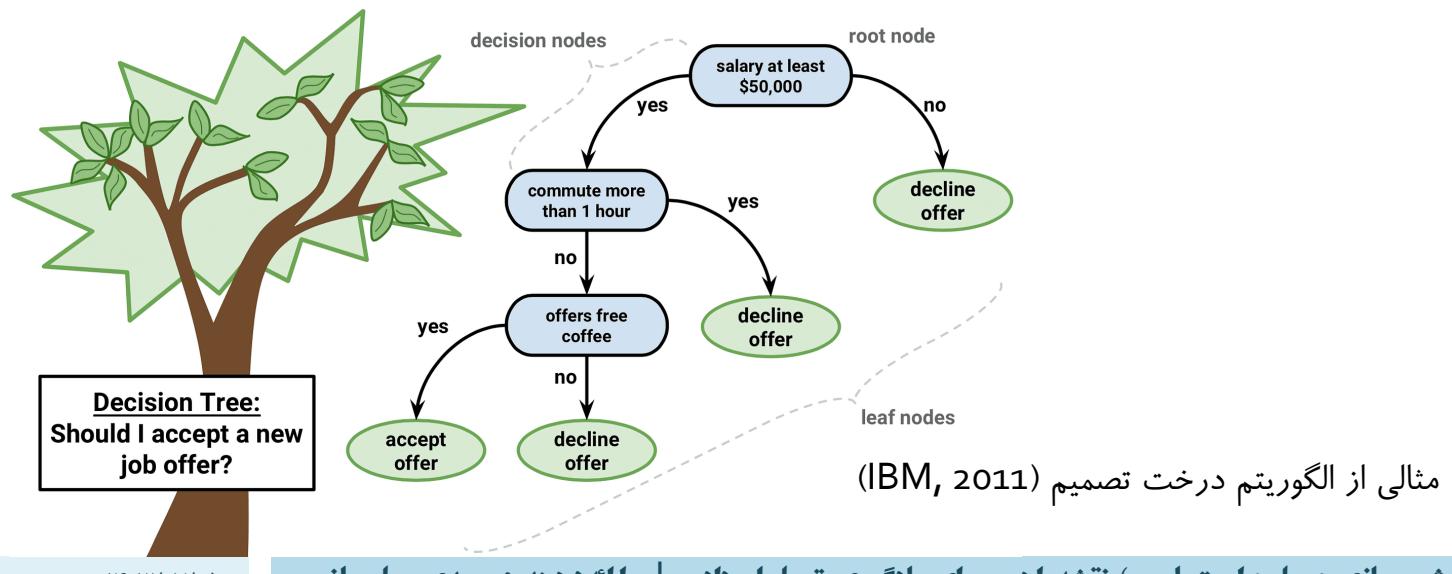
فرایند کلی عملکرد الگوریتم (MIT Press, 2010) KNN



# استنتاج/یادگیری با نظارت

## درخت و جنگل تصمیم (یادگیری ماشین)

- درخت تصمیم (Decision Tree) یکی از الگوریتم‌های یادگیری با نظارت است که از مجموعه تصمیم‌های باینری (با جواب بله و خیر) تشکیل شده و به یک یا چند خروجی مشخص می‌رسد
- همانند یک درخت، درخت تصمیم از یک ریشه شروع شده و با عبور از تنه تصمیم، به تعدادی برگ (خروجی) می‌رسد. از درخت تصمیم هم می‌توان برای دسته‌بندی و هم برای تحلیل رگرسیون استفاده کرد



# استنتاج/یادگیری با ناظارت

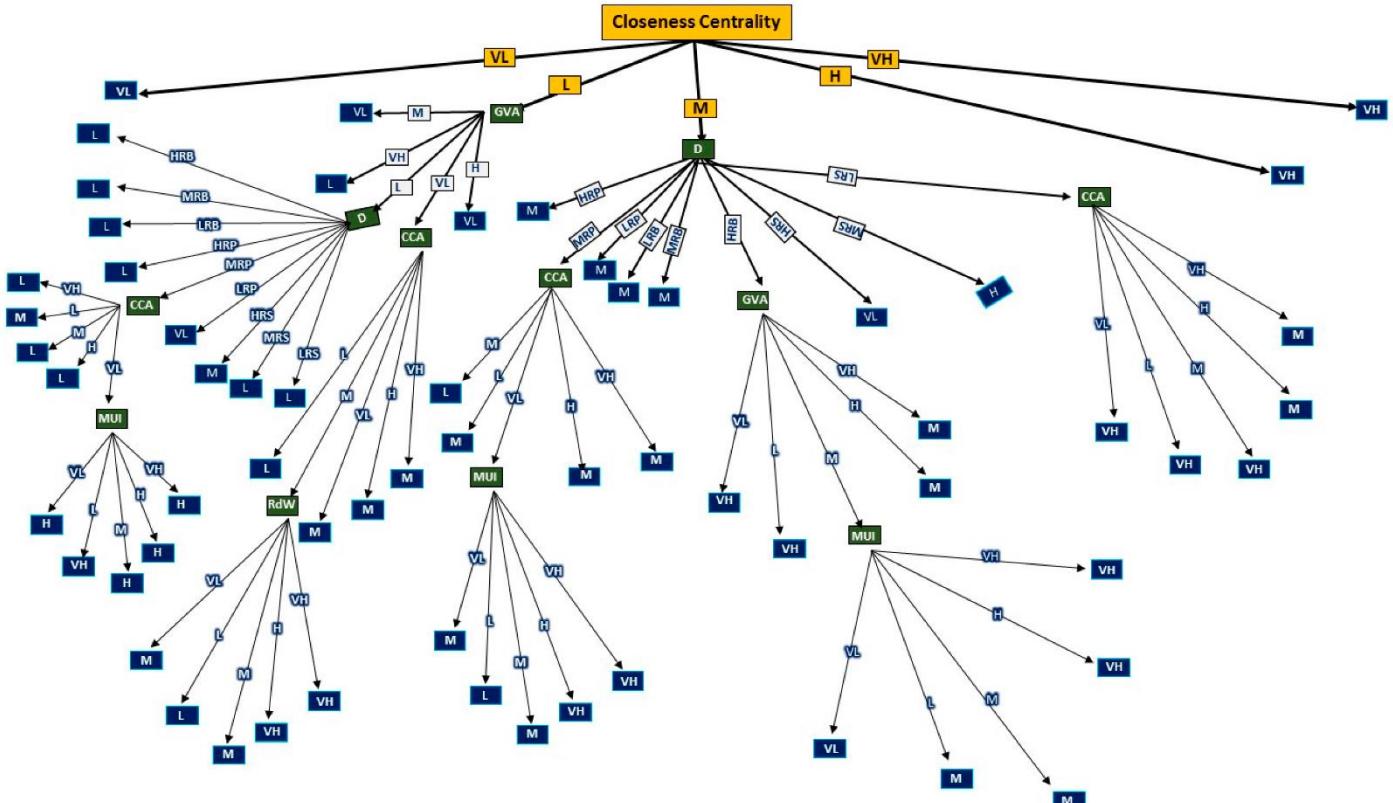
## درخت و جنگل تصمیم (یادگیری ماشین)

- جنگل تصمیم یا جنگل تصادفی (Random Forest) مجموعه‌ای از چندین درخت تصمیم درباره یک موضوع است که هر کدام از این درخت‌ها ساختار منحصر به فردی دارد
- جنگل تصمیم یک الگوریتم پیشرفته یادگیری ماشین است که به دلیل تحلیل چندجانبه روابط کاربرد گسترده‌ای در تحلیل و پیش‌بینی مسائل پیچیده، از جمله مسائل شهری، دارد
- استفاده از الگوریتم جنگل تصمیم، نیازمند مجموعه‌ای گسترده از داده برای آموزش (Train) ماشین است که این داده می‌تواند به صورت کلان داده (Big Data) جمع‌آوری شود یا با تکنیک BootStrap از داده موجود تولید شود



# استنتاج/یادگیری با نظرات

## درخت و جنگل تصمیم (یادگیری ماشین)



LRP – Low Rise Point	LRS – Low Rise Strip	LRB – Low Rise Block	Road Width – RdW	Imperviousness – Imp	Very Low – VL	High – H
MRP – Mid Rise Point	MRS – Mid Rise Strip	MRB – Mid Rise Block	Ground Vegetation % – GVA	Density – D	Low – L	Very High – VH
HRP – High Rise Point	HRS – High Rise Strip	HRB – High Rise Block	Canopy Cover % – CCA	Mixed use Index – MUI	Moderate – M	

کاربست الگوریتم جنگل تصمیم در تحلیل  
شدت جریان روان‌آب‌های سطحی بر اساس  
شاخص‌های فرم شهری، مانند فشردگی شبکه،  
تراکم، پوشش گیاهی و... (Madusanka et al., 2022)



# اجرای گام به گام یک پروژه تحلیل داده انتخاب، اجرا و استنتاج مدل (۶- تحلیل و اجرای مدل)

هدف: تحلیل عوامل اثرگذار (تحلیل ارتباط)



مدل‌های رگرسیون

۹

$y =$  (نرخ رخداد جرائم خشن)

$\beta_0 +$

سرانه هزینه کرد پلیس در سال ۲۰۱۴ \*  $\beta_3$  + نرخ تحصیلات \*  $\beta_2$  + نوع توزیع سنی \*

نسبت جنسیتی \*  $\beta_6$  + نرخ نیروی کار \*  $\beta_5$  + سرانه هزینه کرد پلیس در سال ۱۹۵۹ \*

+ نرخ بیکاری \*  $\beta_9$  + نرخ مهاجران و اقوام غیرسفیدپوست \*  $\beta_8$  + اندازه ایالت \*

$\beta_{10}$  + نابرابری درآمدی \*  $\beta_{11}$  + اندازه ثروت \*

نوع متغیر وابسته؟ عددی پیوسته



رگرسیون خطی

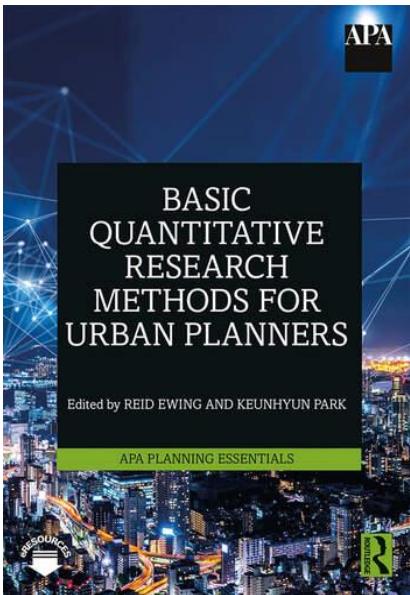


پنجش ششم

# منابع پیشنهادی برای یادگیری خودآموز

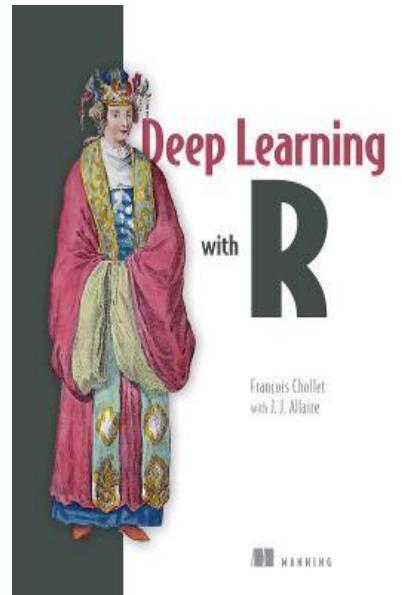


# منابع پیشنهادی برای یادگیری خودآموز



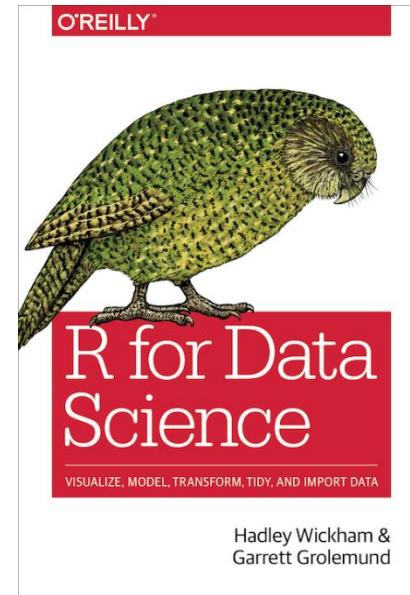
[Quantitative Research Methods...](#) کتاب دو مجلدی

- مناسب برای یادگیری مدل‌های آماری و
- روش‌های تحقیق کمی حوزه شهرسازی



[Deep Learning with R](#) کتاب هوشمند

- مناسب برای کاربرست R در مدل‌سازی عمیق
- مقدمه‌ای برای یادگیری هوش مصنوعی در R



[R for Data Science](#) کتاب هوشمند

- رایگان و در دسترس!
- مناسب برای یادگیری مبانی R
- مناسب برای یادگیری کتابخانه‌های پایه



# منابع پیشنهادی برای یادگیری خودآموز



## IBM Data Analytics with Excel and R Professional Certificate

Prepare for a career in data analytics. Gain the in-demand skills and hands-on experience to get job-ready prior experience required.

-Taught in English | [22 languages available](#) | Some content may not be translated

[IBM Data Analytics](#) دوره

- ارائه گواهینامه معتبر بین المللی
- تمرین و حل مسائل به روز و حرفه‌ای



## R Programming 101

@RProgramming101 · 98.3K subscribers · 83 videos

This channel provides teaching videos on data analysis

[learnmore365.com/pages/membership-r-programming-](#)

Subscribe

[R Programming 101](#) کanal ویدئویی

- رایگان و در دسترس!
- مناسب برای یادگیری مبانی R
- مناسب برای یادگیری کتابخانه‌های پایه



پایان!  
از توجه شما سپاس گزارم  
*[mi.suleimany@ut.ac.ir](mailto:mi.suleimany@ut.ac.ir)*