探索式数据分析

覃雄派

探索式数据分析

提纲

- 什么是探索式数据分析
- 探索式数据分析的目的
- 探索式数据分析的具体任务
    - 单变量分析
    - 双变量分析
    - 多变量分析
    - 降维
- 数据采样
- 探索式分析的“坑”

探索式数据分析

- 探索式数据分析

探索式数据分析

- 什么是探索式数据分析
  - – Data Analysis is basically where you use statistics and probability to figure out trends in the data set
    - It helps you to sort out the “real” trends from the statistical noise
  - – Exploratory Data Analysis (EDA) in Python is the first step in your data analysis process developed by “John Tukey” in the 1970s
    - In statistics, exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods

探索式数据分析

- 探索式数据分析 - Exploratory Data Analysis
  - 不断探索数据，进行描述统计分析的过程
  - 目标是发现数据中的模式，从而更好地理解数据

Exploratory data analysis is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those that we believe to be there

From John Turkey

π@RUC

探索式数据分析

- 什么是探索是数据分析
  - Exploratory data analysis means studying the data to its depth to extract actionable insight from it
  - It includes analyzing and summarizing massive datasets, often in the form of charts and graphs
    - Hence, it’ s unarguably the most crucial step in a data science project, which is why it takes almost 70-80% of time spent in the whole project
    - The better you know your dataset, the better you can make use of it!

探索式数据分析

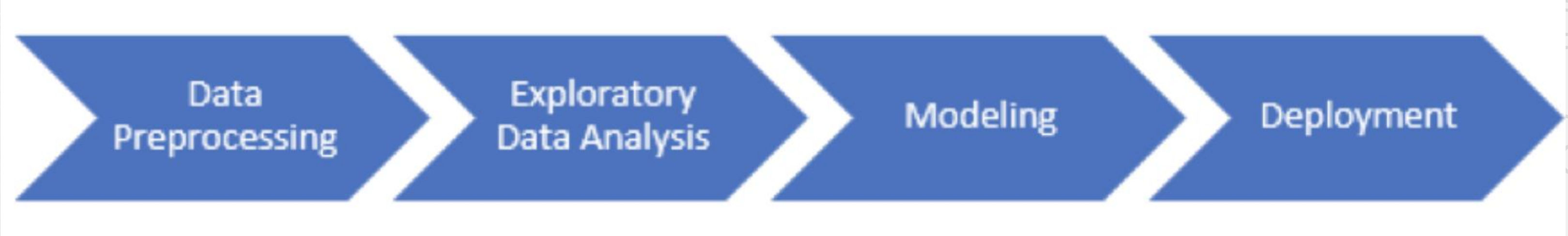
- 探索式数据分析

数据预处理

探索式分析

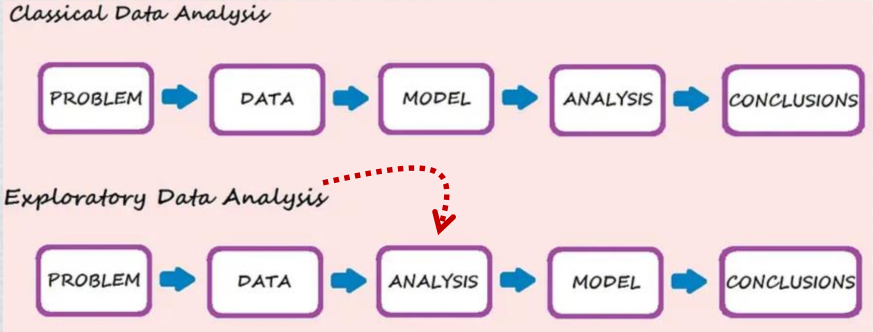
分析与建模

结果呈现与解读



探索式数据分析

- 什么是探索式数据分析
  - **The Need For Exploratory Data Analysis**
  - Exploratory Data Analysis is a crucial step before you jump to machine learning or modeling of your data, By doing this you can get to know
    - whether the selected features are good enough to model
    - are all the features required
    - are there any correlations
    - based on which we can either go back to the Data Pre-processing step or move on to modeling



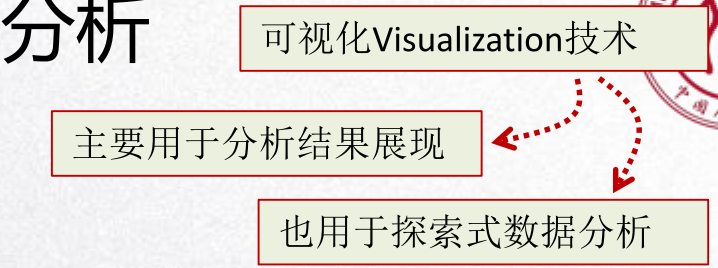
探索式数据分析

探索式数据分析

- 探索式数据分析：目的(primary purpose of EDA)
  - Discover Patterns, maximizing insight into the underlying structure of the data set
  - Pinpointing the important variables and factors
  - Identifying outliers and anomalies
  - Verifying **assumptions** and achieving confident conclusions
  - Uncovering simple efficient models with great explanatory power
    - i.e. models which can explain the data with minimum parameters

探索式数据分析

探索式数据分析



可视化Visualization技术

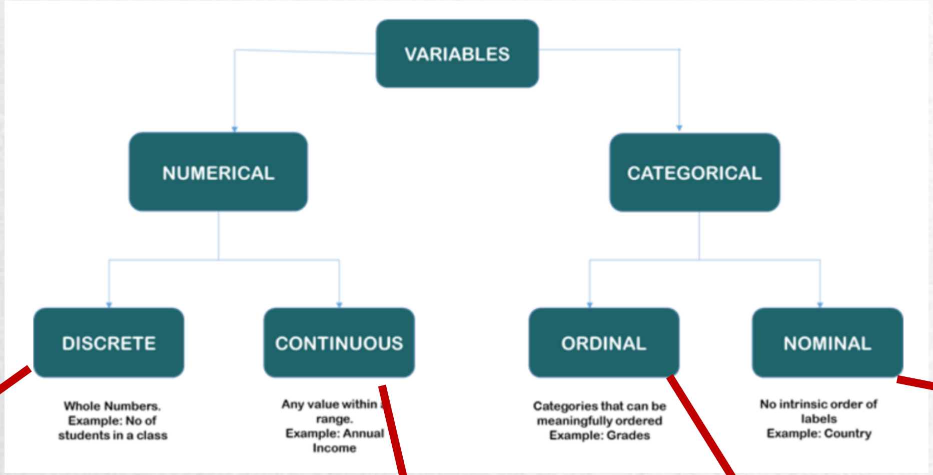
主要用于分析结果展现

也用于探索式数据分析

- 探索式数据分析：做些什么工作
  - 装载数据
  - 描述性分析Description analysis
    - Univariate analysis: considering one variable at a time, learning each variable's distribution and summary statistics
  - 清洗数据Cleaning data（必要的时候）
  - 双变量分析Pair exploration
    - identify relationships between pairs of variables using two-dimensional graphs
  - 多变量分析Multivariate analysis
    - the relationships between larger groups of variables can be analyzed to investigate and identify more complex relationships
  - 假设检验Hypothesis testing and estimation
    - The assumptions made regarding the data set can be tested and estimations are made regarding the variability of variables.
  - 数据编码和转换Transform data to the required format （必要的时候）
  - 降维Dimensional reduction

探索式数据分析

- 变量类型


VARIABLES
NUMERICAL
CATEGORICAL
DISCRETE
CONTINUOUS
ORDINAL
NOMINAL
Whole Numbers. Example: No of students in a class
Any value within range. Example: Annual Income
Categories that can be meaningfully ordered Example: Grades
No intrinsic order of labels Example: Country

.班级人数
.所属区间(Intervals)，区间可以编号，每个区间是一个数据值的范围

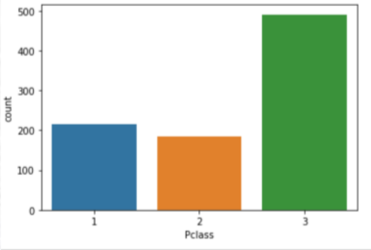
.销售额
.薪水
.身高
.年龄
.百分比(ratio)

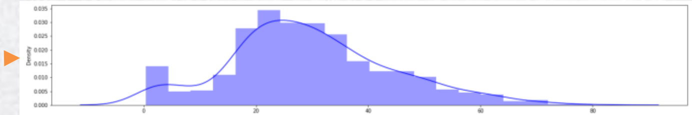
.高中、大专、大学、研究生、博士
.优、良、中、及格、不及格
.分数的A、B、C

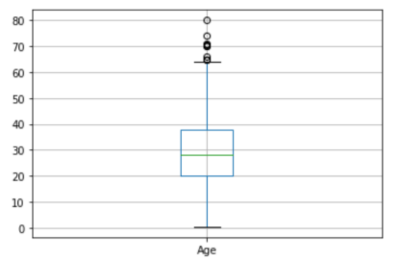
.ID、姓名、电话等，和记录行数相当
.国家
.地区
.邮编
.性别
.型号

探索式数据分析

- 探索式数据分析：做些什么工作
- Univariate Analysis：数据的描述性分析
  - 类别型变量
    - count/distinct value count/ distinct values/top values and frequency
    - histogram直方图
  - 数值型变量
    - count/ max/min/mean/median/各个4分位点
    - Line chart线条图
    - density分布图(Central Tendency / Dispersion)
    - box plot箱线图



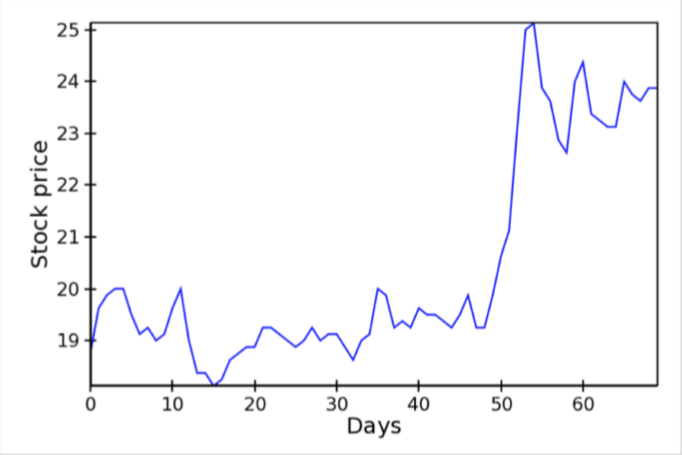




探索式数据分析

- 线条图
  - 一个变量：数值型
  - 看趋势

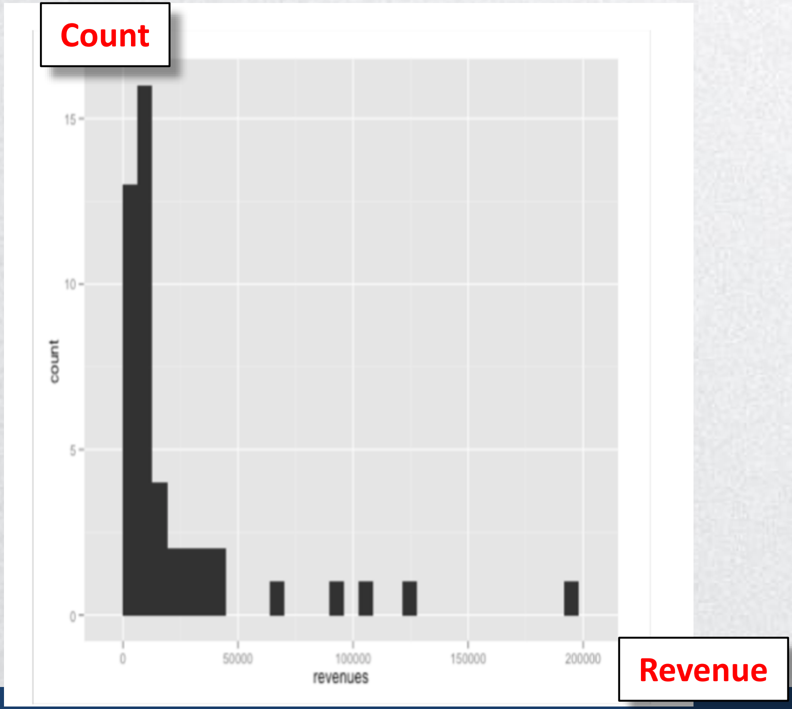
Days	Price
1	15.34
2	17.12
3	18.56
4	19.21
…	…



探索式数据分析

- 直方图（Histogram）
  - 直方图表示了数据在各组的**频数**分布情况

Revenue
2000
11000
5400
204944
32244
1232
…



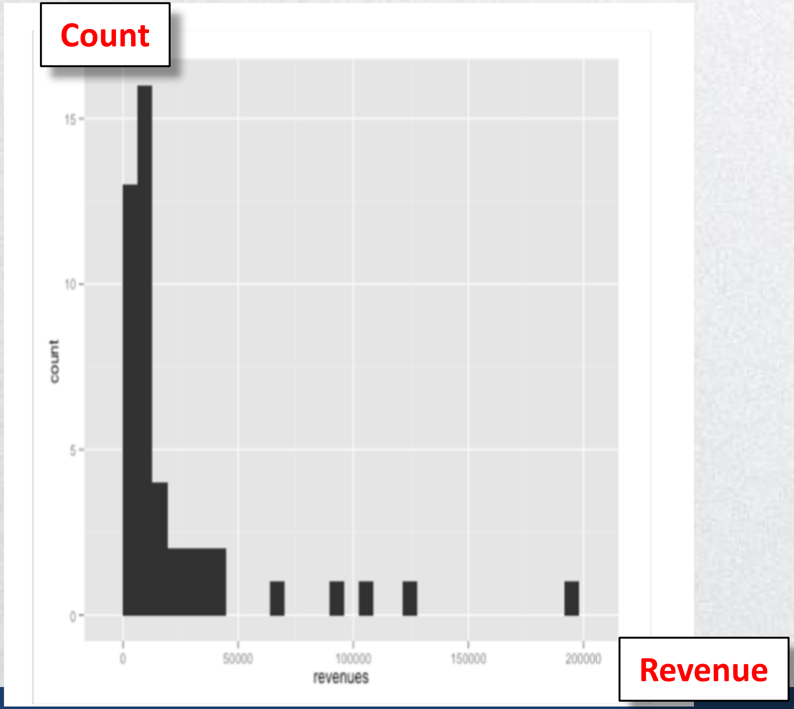
在绘制40个企业营收的直方图时，这些企业的营收落在了4个桶中。如果第一、二、三组的频数分别是5、8、15，则第四组的频数是多少？

A. 12
B. 40
C. 15

探索式数据分析

- 直方图（Histogram）
  - 直方图表示了数据在各组的**频数**分布情况

Revenue
2000
11000
5400
204944
32244
1232
…

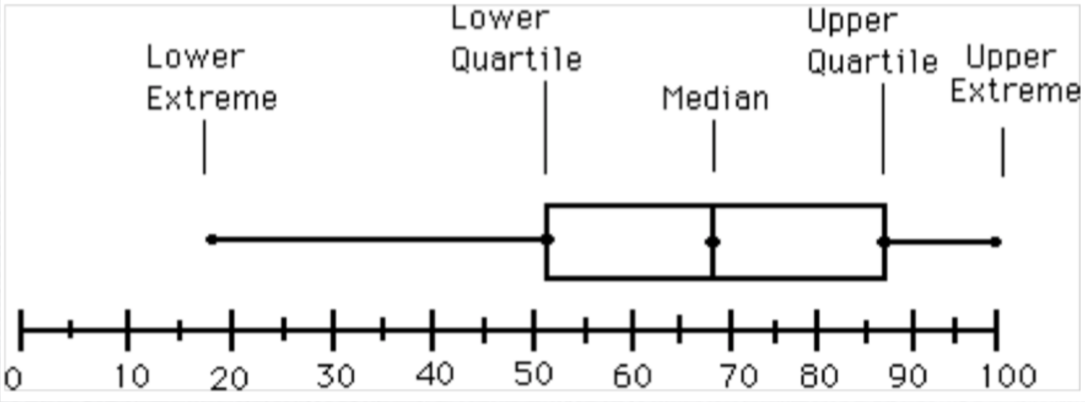


在绘制40个企业营收的直方图时，这些企业的营收落在了4个桶中。如果第一、二、三组的频数分别是5、8、15，则第四组的频数是多少？

[✓ A. 12]
B. 40
C. 15

探索式数据分析

- 箱线图：通过图形的方式表示数据的
  - Min, 25% Quartile, Median, 75% Quartile, Max



探索式数据分析

- 箱线图
  - Find the median, lower quartile and upper quartile of the following numbers.
    - The **median** divides the data into a lower half and an upper half
    - The **lower quartile** is the middle value of the lower half
    - The **upper quartile** is the middle value of the upper half

请给出右侧一组数据的
lower quartile

12, 5, 22, 30, 7, 36, 14, 42, 15, 53, 25

A. 22

B. 12

探索式数据分析

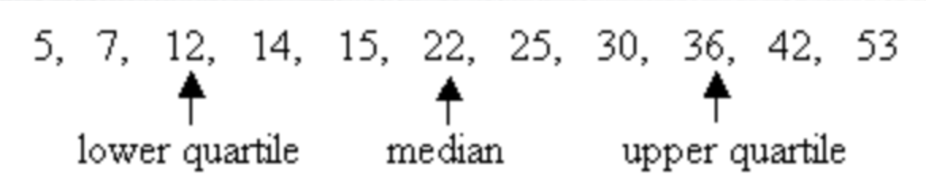
- 箱线图
  - Find the median, lower quartile and upper quartile of the following numbers.
    - The **median** divides the data into a lower half and an upper half
    - The **lower quartile** is the middle value of the lower half
    - The **upper quartile** is the middle value of the upper half

请给出右侧一组数据的
lower quartile

A. 22

B. 12

12, 5, 22, 30, 7, 36, 14, 42, 15, 53, 25



5, 7, 12, 14, 15, 22, 25, 30, 36, 42, 53

lower quartile

median

upper quartile

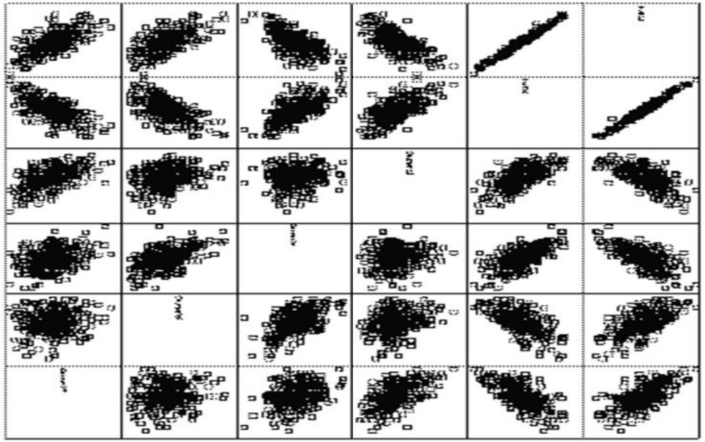
探索式数据分析

- 探索式数据分析：做些什么工作
  - Bivariate analysis
  - refers to studying the relationship between any two variables in the dataset
    - Understanding relationships and new insights through plots
    - between any two predictor variables or with the target variable
      - if strong relationships exist among predictor, the relationship could cause problems during the model development
    - Some of the techniques
      - Scatter Plots
      - Regression Analysis
      - Correlation Coefficients: heat map

探索式数据分析

- 数据冗余
  - 属性上的相关性会导致一些属性间存在冗余信息
  - 皮尔森相关系数

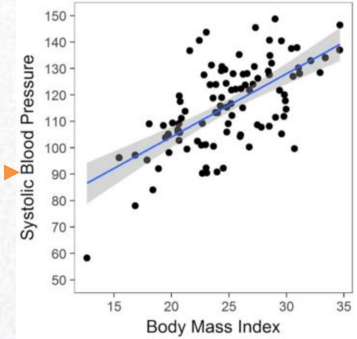
$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

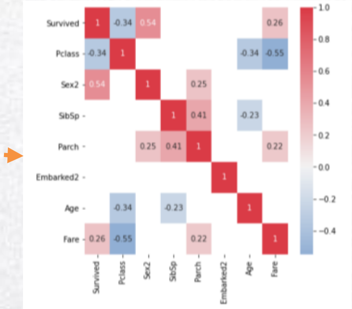


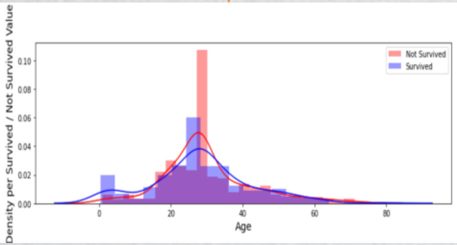
π@RUC

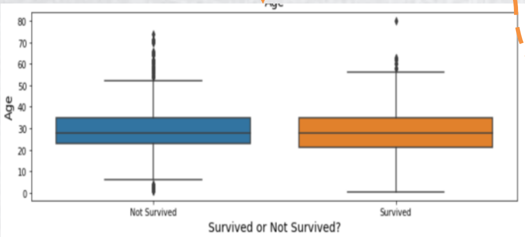
探索式数据分析

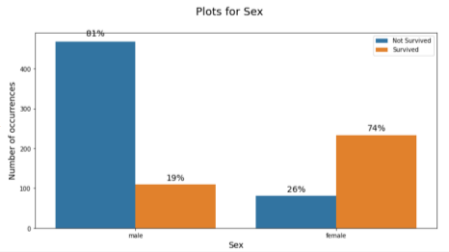
- Pair variables：可视化
  - Numerical vs. Numerical
    - Scatterplot/ Heat map for correlation
  - Categorical vs. Numerical
    - Histogram / categorical box plot
  - Two Categorical Variables
    - Bar chart/Grouped bar chart





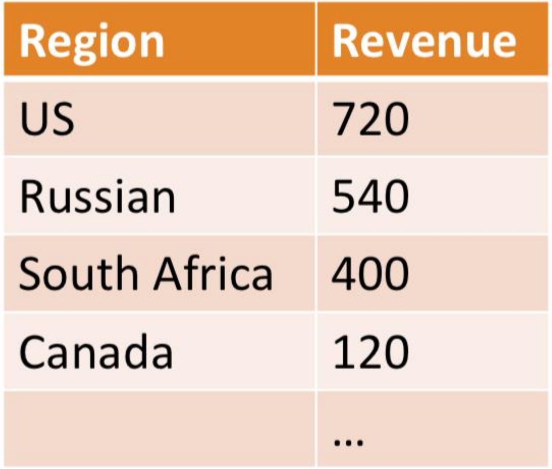




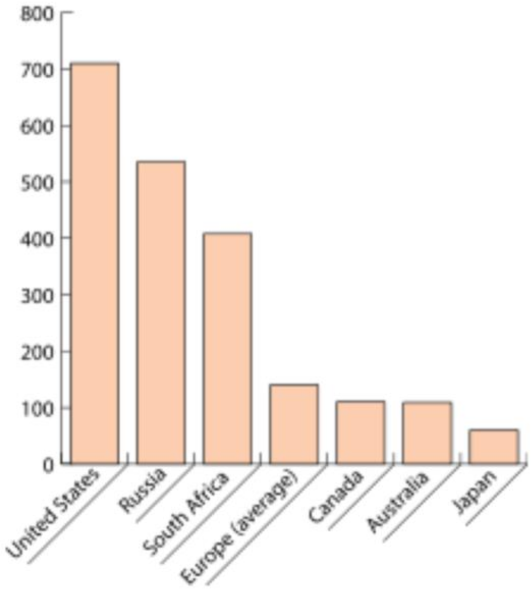


探索式数据分析

- 柱状图
  - 一个变量为离散型（Categorical）、一个变量为数值型

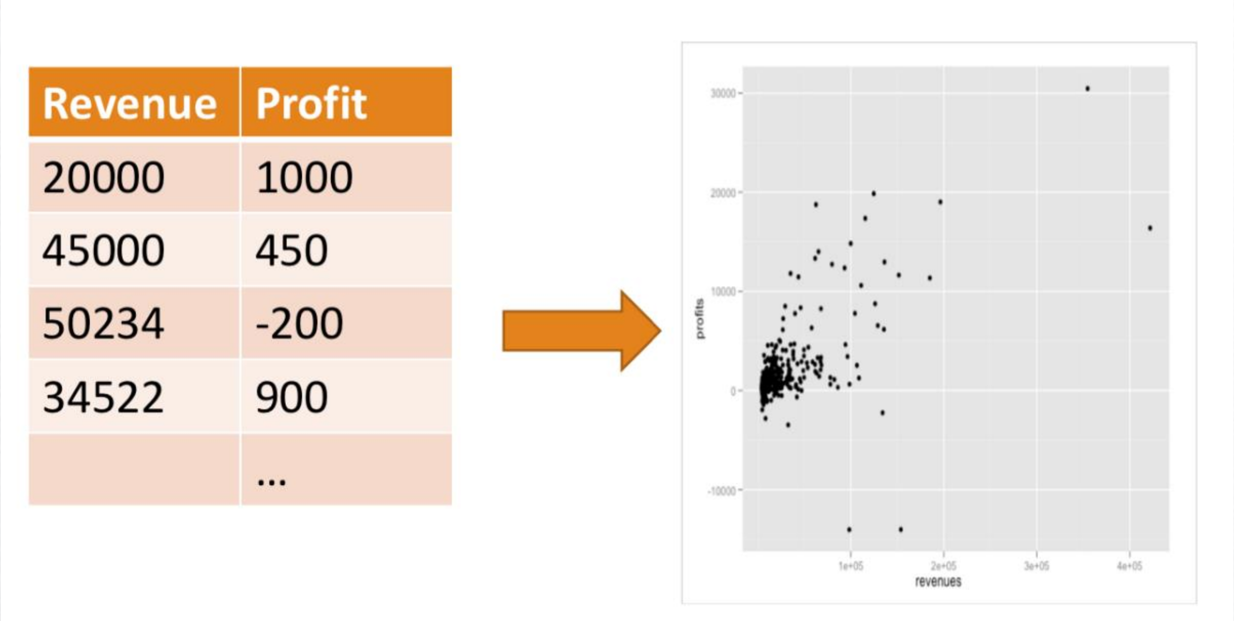


Region	Revenue
US	720
Russian	540
South Africa	400
Canada	120
	…



探索式数据分析

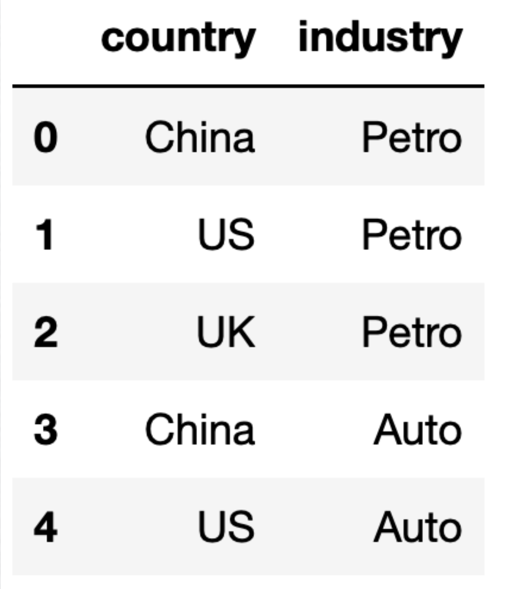
- 散点图
  - 两个变量均为数值型



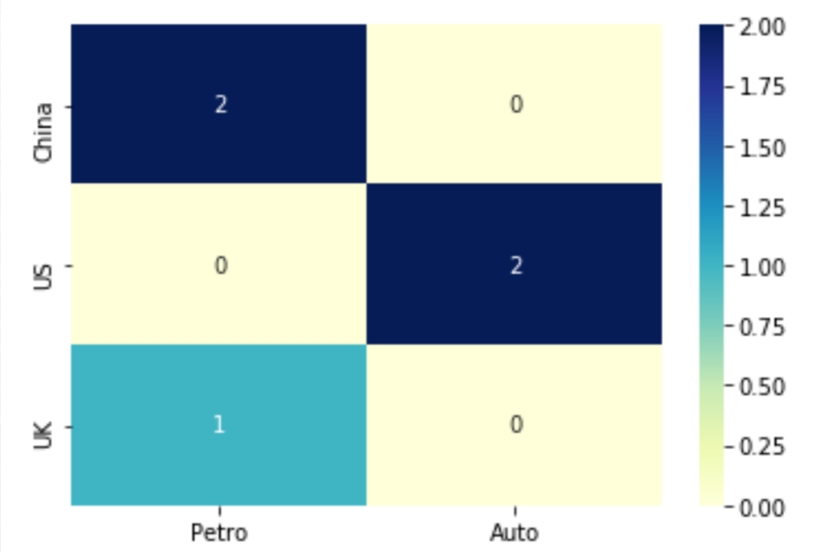
Revenue	Profit
20000	1000
45000	450
50234	-200
34522	900
	…

探索式数据分析

- 热力图示例
  - 两个变量均为离散型，计算相关性



	country	industry
0	China	Petro
1	US	Petro
2	UK	Petro
3	China	Auto
4	US	Auto



探索式数据分析

探索式数据分析

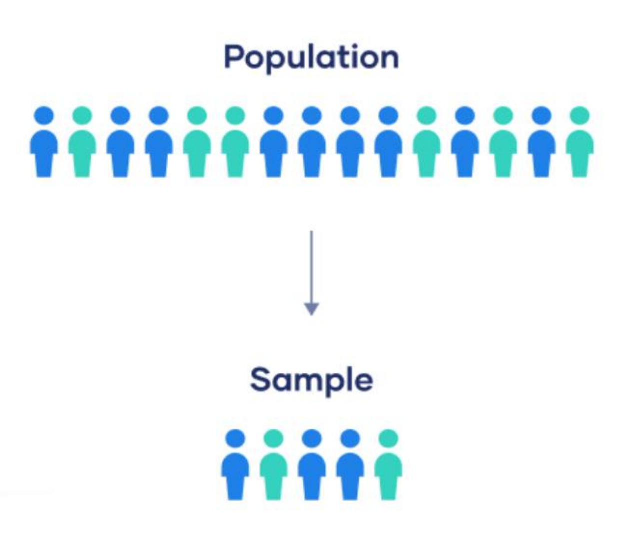
- 多变量分析Multivariate analysis and visualization
  - performed to understand interactions between different fields in the dataset

- 降维Dimensionality reduction
  - Deal with high dimensional data
  - Reduce the dimension first
    - understand the fields in the data that account for the most variance between observations
    - Reduce data volume
  - The visualize in 2d or 3d space to understand the data

探索式数据分析

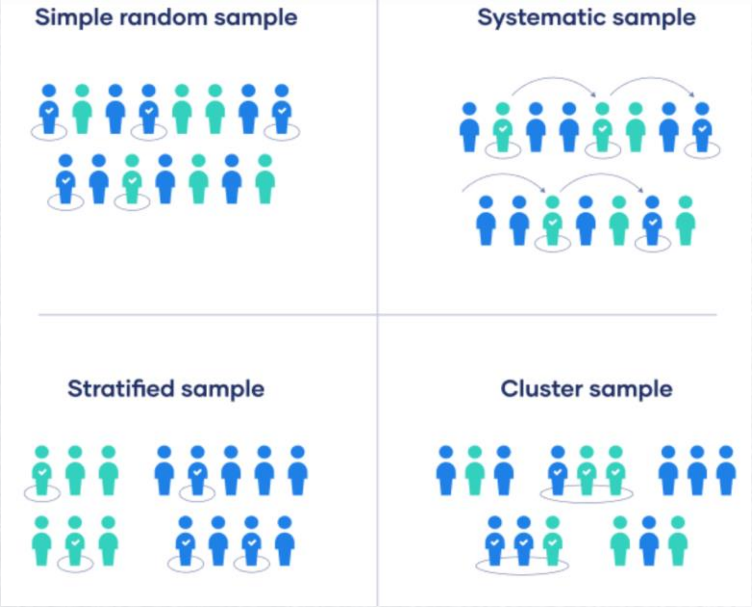
探索式数据分析

- 数据采样：为什么要数据采样
  - 当数据量很大的时候，为了在探索式分析中提高效率
  - 需要对数据进行采样：选择有代表性的数据子集，减小数据量
    - 总体和采样



探索式数据分析

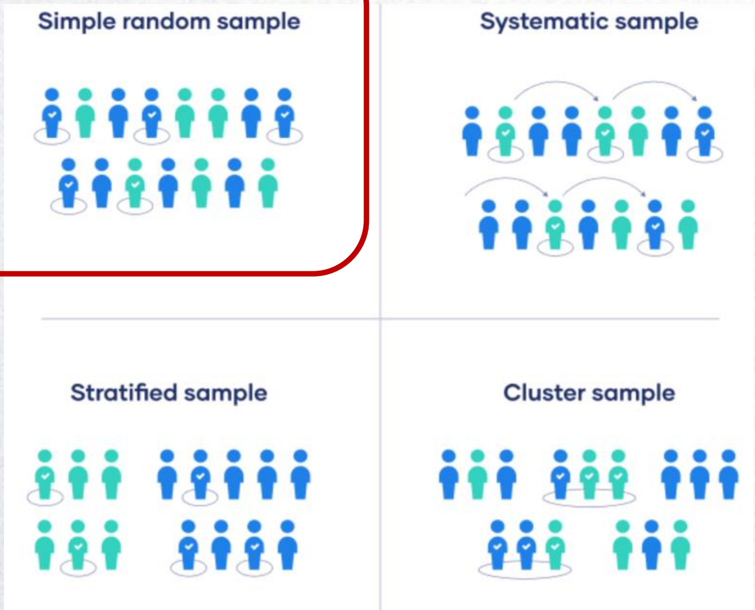
- 主流采样方法
  - 随机采样
  - 系统采样
  - 分层采样
  - 整群采样



探索式数据分析

- 主流采样方法
  - 随机采样
  - 系统采样
  - 分层采样
  - 整群采样

- **Simple random sampling**
- In a simple random sample, every member of the population has an equal chance of being selected
- Your sampling frame should include the whole population.



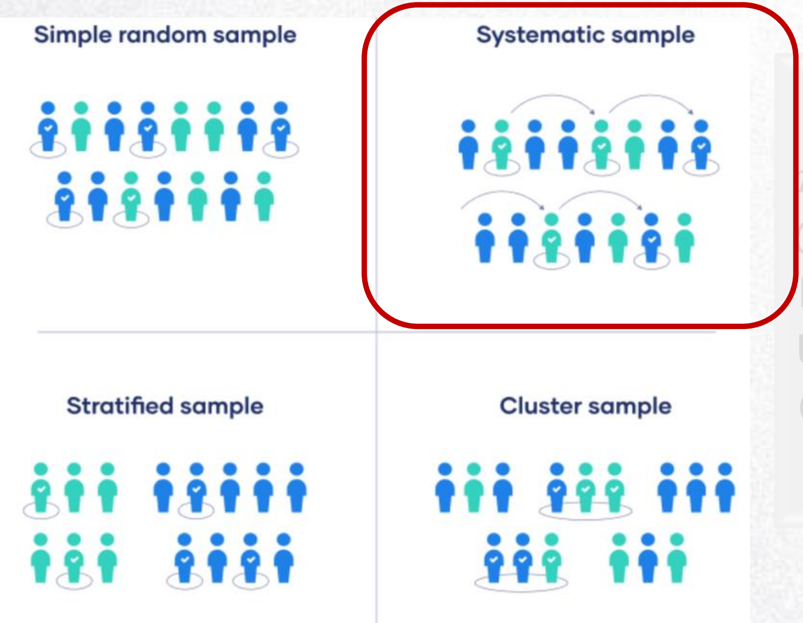
π@RUC

探索式数据分析

- 主流采样方法
  - 随机采样
  - 系统采样
  - 分层采样
  - 整群采样

Systematic sampling
Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals

Example
All employees of the company are listed in alphabetical order. From the first 10 numbers, you randomly select a starting point: number 6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you end up with a sample of 100 people

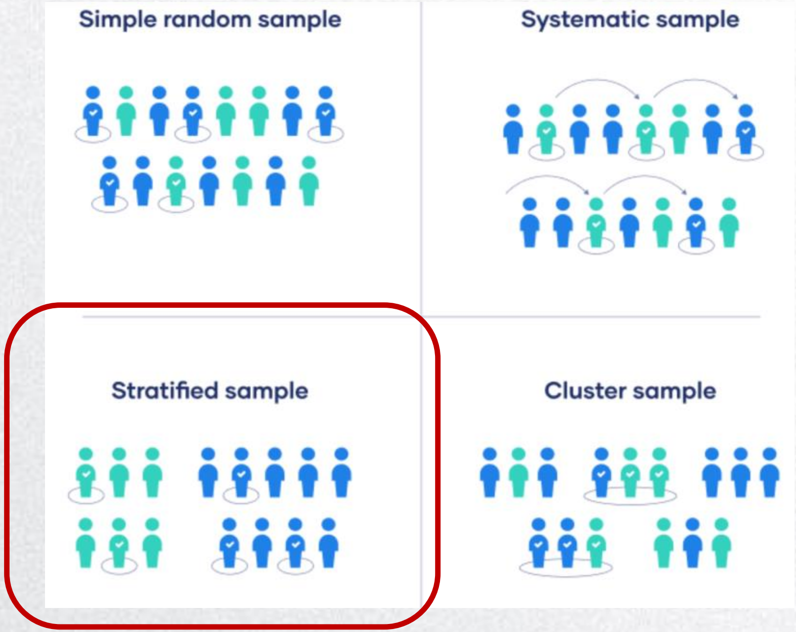


探索式数据分析

- 主流采样方法
  - 随机采样
  - 系统采样
  - 分层采样
  - 整群采样

Stratified sampling
It involves dividing the population into subpopulations that may differ in important ways.
It allows you draw more precise conclusions by ensuring that every subgroup is properly represented in the sample

Example
The company has 800 female employees and 200 male employees. You want to ensure that the sample reflects the gender balance of the company, so you sort the population into two strata based on gender. Then you use random sampling on each group, selecting 80 women and 20 men, which gives you a representative sample of 100 people



探索式数据分析

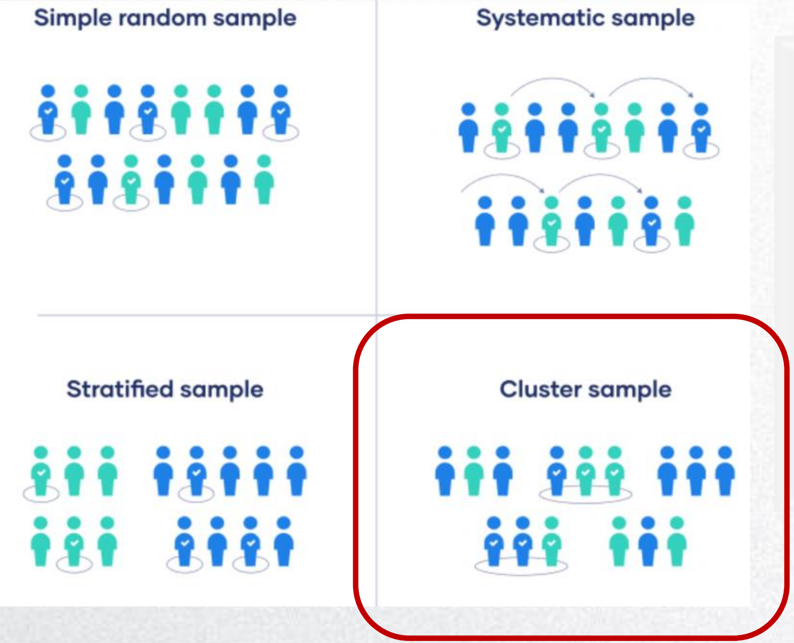
- 主流采样方法
  - 随机采样
  - 系统采样
  - 分层采样
  - 整群采样

Cluster sampling

Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups

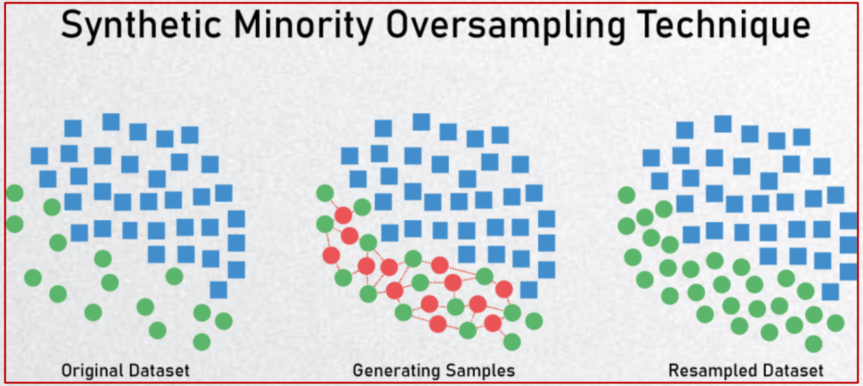
Example

The company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You don’ t have the capacity to travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters



探索式数据分析

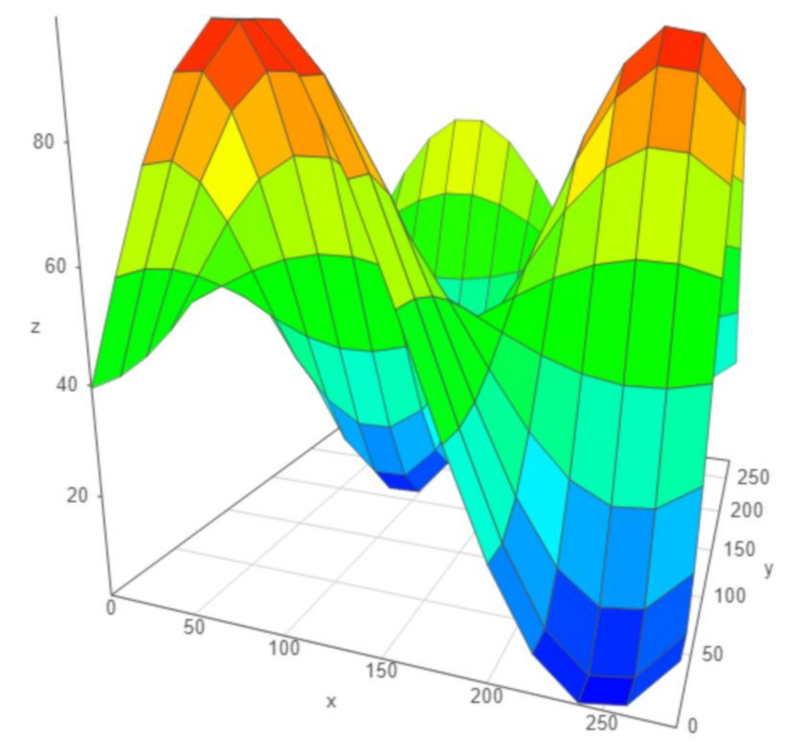
- 数据倾斜问题
  - 90%的人没有感染
  - 99%不买商品
  - 99.99%不是恐怖分析
- 正负样本的极度不均衡，给数据分析带来很多挑战
  - 少采多数样本
  - 重复采少数样本
  - 制造一些假的少数样本SMOTE



探索式数据分析

探索式数据分析

- 数据探索与数据预处理
  - 工具集
    - Pandas
    - Numpy
    - Matplotlib
    - Seaborn
    - plotly
    - Bokeh



探索式数据分析

π
@
R
U
C

探索式数据分析

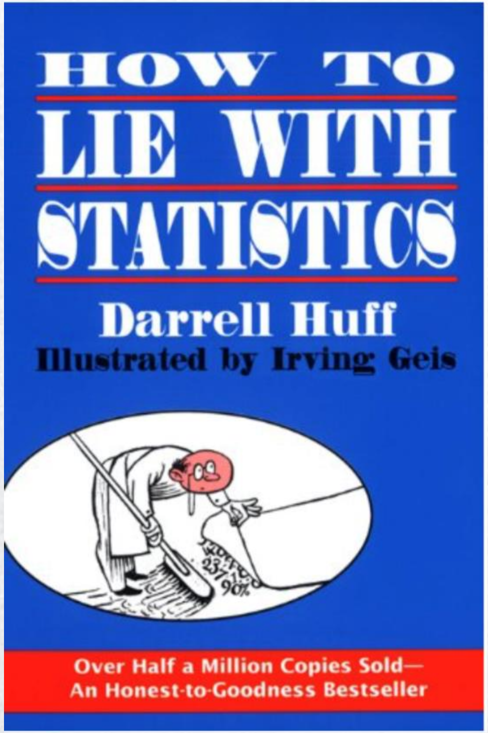
- 警惕EDA的坑
  - 不要迷信“客观”的数字
  - 基于统计，所有人给出的数字都是客观的
  - 只不过有的人可能更客观一些……

There are three kinds of lies:

lies,

damned lies,

and statistics


HOW TO
LIE WITH
STATISTICS
Darrell Huff
Illustrated by Irving Geis
Over Half a Million Copies Sold—
An Honest-to-Goodness Bestseller

探索式数据分析

- EDA需要警惕的“坑”
  - 加州大学伯克利分校（UC Berkeley）的入学申请是否存在性别歧视?

	申请人数	录取率
男生	2691	45%
女生	1835	30%

YES!

✕

探索式数据分析

- EDA需要警惕的“坑”
  - 加州大学伯克利分校（UC Berkeley）的入学申请是否存在性别歧视?

Drill in 院系

院系	男生		女生	
	申请人数	录取率	申请人数	录取率
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%
合计	2691	45%	1835	30%

辛普森悖论

- 在分组比较中都占优势的一方，在总评中有时反而是失势的一方

- 女生录取率低是因为大部分女生申请对男女录取率都低的院系！

探索式数据分析

- EDA需要警惕的“坑”
  - 谁会赢得美国大选，希拉里还是特朗普？
    - 选前民调：希拉里支持率52%、特朗普支持率48%
    - 假设：所有选民在最终投票时不改变主意

WIN

X

探索式数据分析

- 数据分析 ≠ 简单计数（可视化）
  - 假设选前随机找了1000人进行民调
    - 简单计数（可视化）
      - 这1000人就是我要考虑的总体（Population）
      - 统计支持希拉里的人数：520人
      - 统计支持特朗普的人数：480人
      - 结论：希拉里必胜！

探索式数据分析

- 数据分析 ≠ 简单计数（可视化）
  - 假设选前随机找了1000人进行民调
  - 数据分析：应具备统计思维
    - 要考虑的总体是所有在当天投票的人
    - 这1000人是我抽取出的样本（Sample）
    - 通过样本估计出的支持率存在误差：
      - 希拉里：52% ± 3%
      - 特朗普：48% ± 2%

结论：特朗普也有胜算！

探索式数据分析

- 假设你在某家研发移动端APP的公司实习
  - APP用户中有10000人使用Android设备、5000人使用IOS设备，整体的付费转化率是5%。
  - 细分发现其中：
    - IOS设备的总体转化率为4.17%
    - Android设备的总体转化率为5.82%。
  - “英明”的老板得出结论：IOS平台的用户付费转化率低下，建议放弃IOS平台的研发
  - 可聪明的你通过生活经验知道IOS平台的用户更愿意付费，请问你如何用数据说服你的老板?

vs.

另一个坑的实例

探索式数据分析

- 步骤1：提出问题
  - IOS和Android平台的用户，谁付费转化率更高?
- 步骤2：收集数据
  - 请思考你要收集什么数据?

	Android手机	IOS手机	Android平板	IOS平板
转化	50	100	500	100
未转化	1950	3400	7500	1400

- 步骤3：分析数据

转化率	2.56%	2.94%	6.67%	7.14%

请你谈谈，为什么要收集更细粒度的数据?

Drill in
手机 | 平板

探索式数据分析

- 步骤1：提出问题
  - IOS和Android平台的用户，谁付费转化率更高？
- 步骤2：收集数据
  - 请思考你要收集什么数据？

	Android手机	IOS手机	Android平板	IOS平板
转化	50	100	500	100
未转化	1950	3400	7500	1400

- 步骤3：分析数据

转化率	2.56%	2.94%	6.67%	7.14%

请你谈谈，为什么要收集更细粒度的数据？

你觉得那种设备的转化率更高？

A. IOS

B. Android

探索式数据分析

- 步骤4：原因分析
  - 汇总手机和平板的数据时，忽略了二者在“量”的差异——将“值与量”两个维度的数据，归纳成了“值”一个维度的数据

	Android设备	IOS设备
转化	550	200
未转化	9450	4800
转化率	5.82%	4.17%

总量10000　　总量5000

	Android手机	IOS手机	Android平板	IOS平板
转化	50	100	500	100
未转化	1950	3400	7500	1400
转化率	2.56%	2.94%	6.67%	7.14%

- **粗粒度**的数据（如对比IOS和Android总体情况）往往没有多大参考意义
- **要细分**到具体设备、获取渠道等再进行比对才有价值

探索式数据分析

- 步骤5：形成报告
  - 假设你的老板看不懂这么复杂的图表……
  - 对于占总体少数比例的样本加以更高的权重，也就是“逆概加权”(Inverse probability weighting)
    - 依旧是上面的例子，对于汇总的每个子群体加权，权重为该子群体在总群体里出现的概率的倒数

子群

	IOS手机	IOS平板
转化	100	100
未转化	3400	1400

$$\frac{100+100}{3500+1500} \Rightarrow \frac{100*\frac{5000}{3500}+100*\frac{5000}{1500}}{3500*\frac{5000}{3500}+1500*\frac{5000}{1500}} = 4.8\%$$

	Android手机	Android平板
转化	50	500
未转化	1950	7500

$$\frac{50+500}{2000+8000} \Rightarrow \frac{50*\frac{10000}{2000}+500*\frac{10000}{8000}}{2000*\frac{10000}{2000}+8000*\frac{10000}{8000}} = 4.4\%$$

探索式数据分析

π @ R U C