Data Analytics

Lecture 3

Delivering Business Value

- Communicating with your tools

Keep AGILE

- (Specification, Explainer)
- Slack bonus homework?

Lecture 03 content

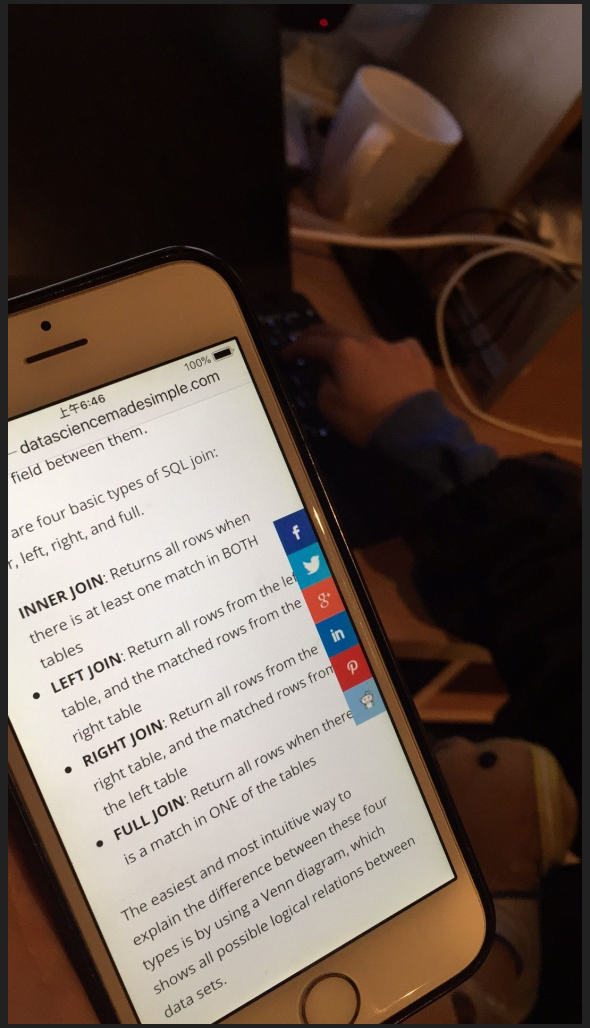
Review/learn key DB concept: Normal Form;

Identify the right business question;

Choose the right chart style

SQL data processing

Saw on train 18th



Cartesian join-调包

Cross Join

Inner join without any condition, full join

```
SELECT COUNT(1)
FROM
(select *
from
Flight_data f1 limit 100)
tab1
CROSS JOIN
(select *
from
Flight_data f1 limit 10)
tab2
```

Normal Form

1 NF:应该是原子属性，每列不可以再分（不要用逗号分隔多个东西存一列）

2 NF每个列都依赖于主键(不要把师生信息都存在成绩表里，不然学生毕业，成绩删除，可能教授信息全部没了）

3 NF不包含其他表中的非关键字信息（只存ID就好，避免冗余）

Normal Form, ERD

Product ID	Color	Price
1	red, green	15.99
2	yellow	23.99
3	green	17.50
4	yellow, blue	9.99
5	red	29.99

TABLE_PRODUCT_PRICE

Product ID	Price
1	15.99
2	23.99
3	17.50
4	9.99
5	29.99

TABLE_PRODUCT_COLOR

Product ID	Color
1	red
1	green
2	yellow
3	green
4	yellow
4	blue
5	red

TABLE_PURCHASE_DETAIL

Customer ID	Store ID	Purchase Location
1	1	Los Angeles
1	3	San Francisco
2	1	Los Angeles
3	2	New York
4	3	San Francisco

TABLE_BOOK_DETAIL

Book ID	Genre ID	Genre Type	Price
1	1	Gardening	25.99
2	2	Sports	14.99
3	1	Gardening	10.00
4	3	Travel	12.99
5	2	Sports	17.99

TABLE_BOOK

Book ID	Genre ID	Price
1	1	25.99
2	2	14.99
3	1	10.00
4	3	12.99
5	2	17.99

TABLE_GENRE

Genre ID	Genre Type
1	Gardening
2	Sports
3	Travel

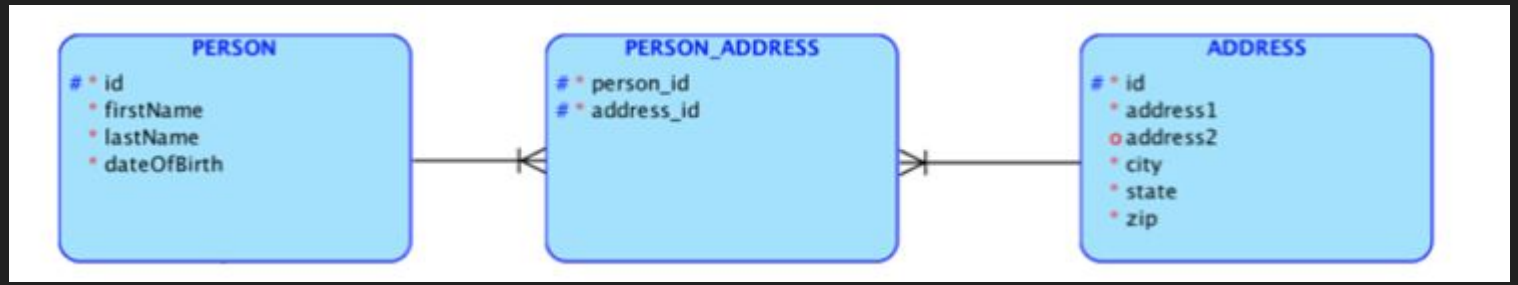
4 Basic concepts of Relational DB

many-to-many relationship exists between customers and products: customers can purchase various products, and products can be purchased by many customer

Cardinality of a given table in relation to another

https://www.agiletrailblazers.com/blog/modernized-technology/4-core-concepts-you-need-to-understand-sql-databases

https://docs.microsoft.com/en-us/power-bi/desktop-create-and-manage-relationships



How to make the elephant dance, with your data

手套型号	价格	采购分公司
A	$ 5	NY
A	$ 7	PALO ALTO
A(similar)	$ 3	SEATTLE
B	$ 3.22	DC
B model2	$ 10.55	NY factory 2

Save 2% purchase cost,
$1,000,000,000

Hans Rosling - the man who can show history in 4 minutes

British Petroleum (BP) ’s problem

- 1991, 10% drill success rate
- Industry leader of best rate already
- Trial drill gives possibility, <20% then unlikely to produce

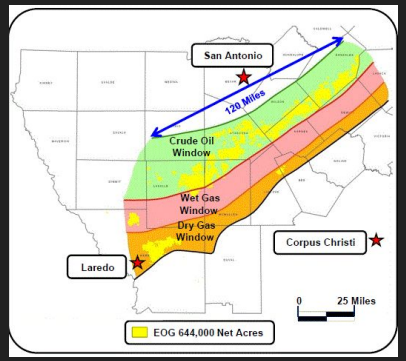
Set the right goal, then use the right type of chart

- Price of each trial:$4,000,000 – $40,000,000

- Revised target: no dry well

- Colour-coded layer, all green=go

- After 9 years : 67% success, 600% lift



Requirement behind user statement

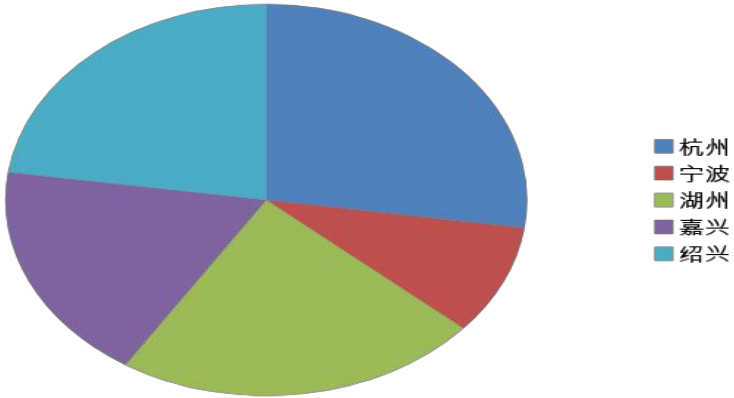
I need a better (horse, CSV file...)

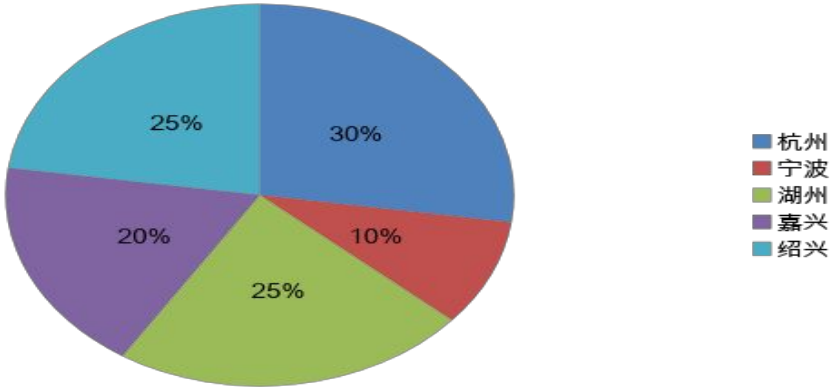
I want to reach more customers

Tableau Example

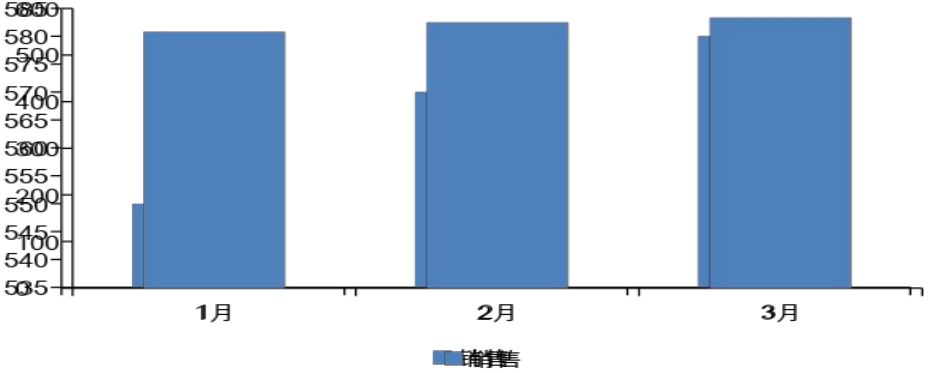
https://public.tableau.com/views/TotalInjuriesEachMonth/InjuriesofAirportAccidentsReport?:embed=y&:display_count=yes&:origin=viz_share_link

好看，还是信息？

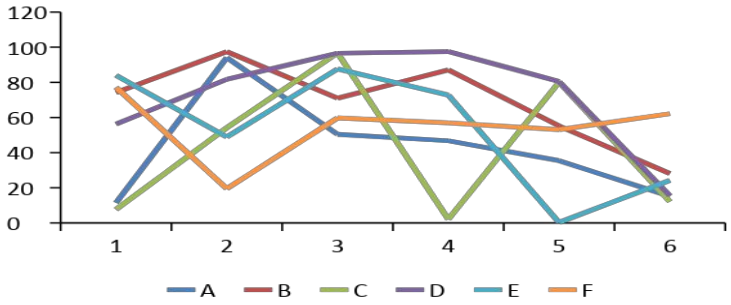




数据一定反应真相吗？

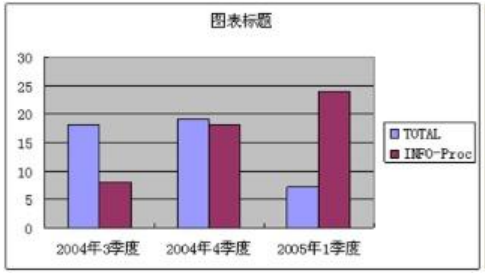


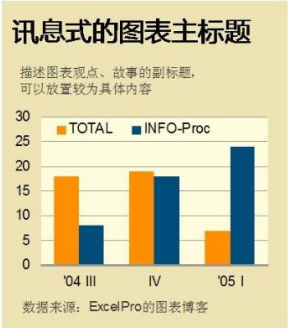
化繁为简



指标说明	WK10本周数据	周趋势
流入量		
线流入量		
预估波动率	10.3%	
估波动率	16.3%	
呼损率	5.6%	
线呼损率	10.8%	

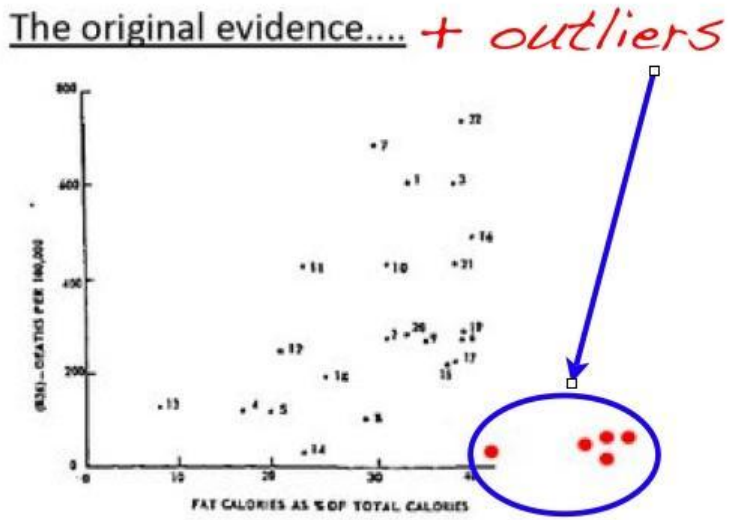
脱颖而出-色系 布局





来自ExcelPro博客 RGB255 150 24, 82 101 115

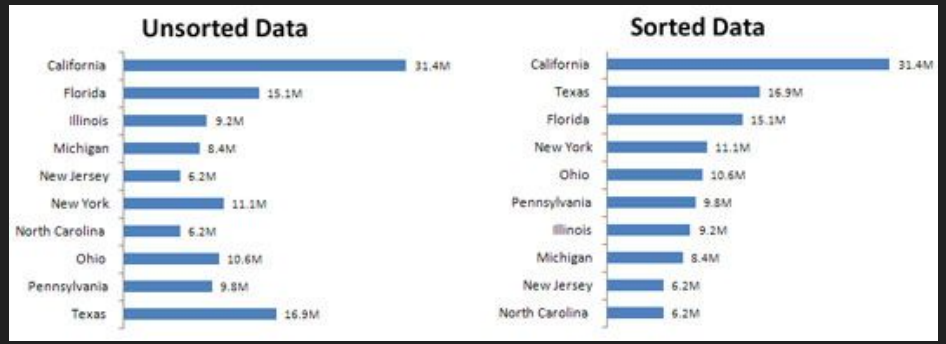
最坑人的分析师，“饱和脂肪酸=〉心血管疾病”



简单，还是复杂好？

1,2,3 of chart style

1. Don’t add fancy elements (background colour, 3D); Remove (default) gridline, border etc
2. Format numbers consistently (to $M, thousands
3. Have a meaningful title and sort order, legend



图表类型列表-Ant Design

比较类

分布类

流程类

地图类

占比类

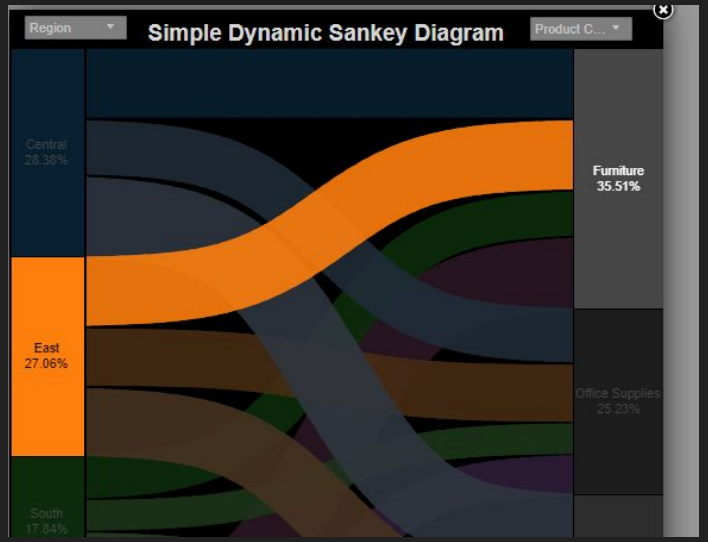
区间类

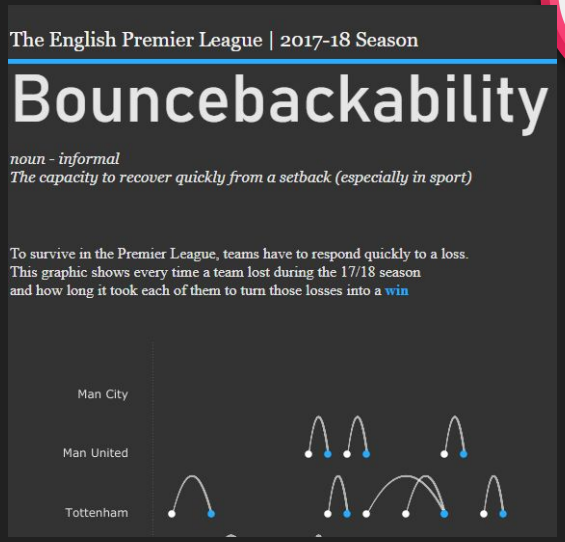
关联类

时间类

趋势类

Example charts





Small Multiples

Tableau functions

Attr()

Relation

SQL clause: Window functions and Partition By

https://www.brentozar.com/sql-syntax-examples/window-function-examples-sql-server/

rn	COUNTRY	EVENT_DATE
1	NULL	2007-06-21 00:00:00
2	NULL	2007-05-13 00:00:00
1	Afghanistan	2014-11-07 00:00:00
2	Afghanistan	2014-06-20 00:00:00
1	Algeria	2017-08-10 00:00:00
2	Algeria	2011-09-22 00:00:00
1	American Samoa	2016-04-23 00:00:00
2	American Samoa	2014-07-22 00:00:00
1	Angola	2011-11-26 00:00:00
2	Angola	2008-01-19 00:00:00
1	Antarctica	2011-12-19 00:00:00
2	Antarctica	2007-12-20 00:00:00
1	Argentina	2018-07-12 00:00:00
2	Argentina	2018-04-10 00:00:00
1	Australia	2018-11-24 00:00:00
2	Australia	2018-09-07 00:00:00

SQL WITH clause-(a.k.a CTE Common Table Expression) -

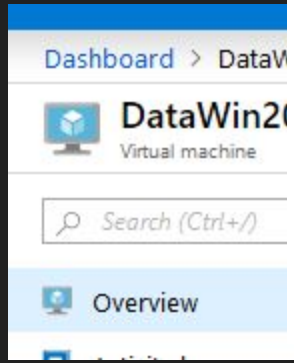
```sql
WITH

Ranked AS --THIS IS A TEMP RESULT SET
(SELECT
row_number() OVER (PARTITION BY COUNTRY ORDER BY EVENT_DATE DESC) as rn
, COUNTRY
,EVENT_DATE
FROM
Flight_data
)

SELECT

RN
,COUNTRY
,EVENT_DATE
FROM
Ranked
WHERE rn<2
```

Add Alteryx



RegEx Tool

The RegEx tool is able to leverage the powerful pattern matching abilities of regular expression syntax for the sake of parsing, matching, or replacing string data.

1) Run the workflow (Ctrl+R).

2) Select a tool to view its output in the Results window.

Introduction

Regular expressions are matching patterns with the versatile ability to extract useful pieces of information from strings.

Expand the 'More Info' boxes to the right of each example to learn about the specific RegEx syntax used.

Users new to RegEx may want to consult an online RegEx resource for additional information to harness the robust capabilities of this tool.

RegEx Match

More Info

The "Match" output method of the RegEx tool returns a boolean value of true if the specified string matches the regular expression, and false if it does not. This method looks for full string matches rather than partial matches.

RegEx Parse (first name / last name)

More Info

The "Parse" output method of the RegEx tool returns each of the groups defined in the regular expression. The output field, type, and length can all be specified in the configuration tool once the expression has been entered.

RegEx Parse (zip4)

More Info

In "Parse" mode, partial matches will be considered, however only the first match will be returned.

RegEx Replace

More Info

Resource group (change) : MyGRoup1

Good security design

Wrong direction - when not to re-invent wheels

Don’t do it or re-invent wheels - 对账时我自己算calculation-问题在于business rule will change

Specification

报表是服务 we delivery information products, we need to collect user data (Tableau, not CSV)
逻辑将死 行为已死 只有数据流传

出错 frequent catches

半可加，命名不一致

Sum(NUMBER_OF_SAIL_DAYS) day 首先不该group因为是半可加。然后两个Branch的名字不一样day, NUM_OF_SAIL_DAYS另一头，又有fill with 0,所以查不出

Data - What level of value is your work @？

Top

Mid

Basic