第三章：数据可视化

数据可视化

- **数据可视化**– 用有意义的方式展示数据以增强商业洞见的过程
  - 数据可视化工具为经理人提供了简易掌握的分析能力，以减少对专业IT人员的依赖。因为信息变得更直观，在组织中增加了信息共享与合作

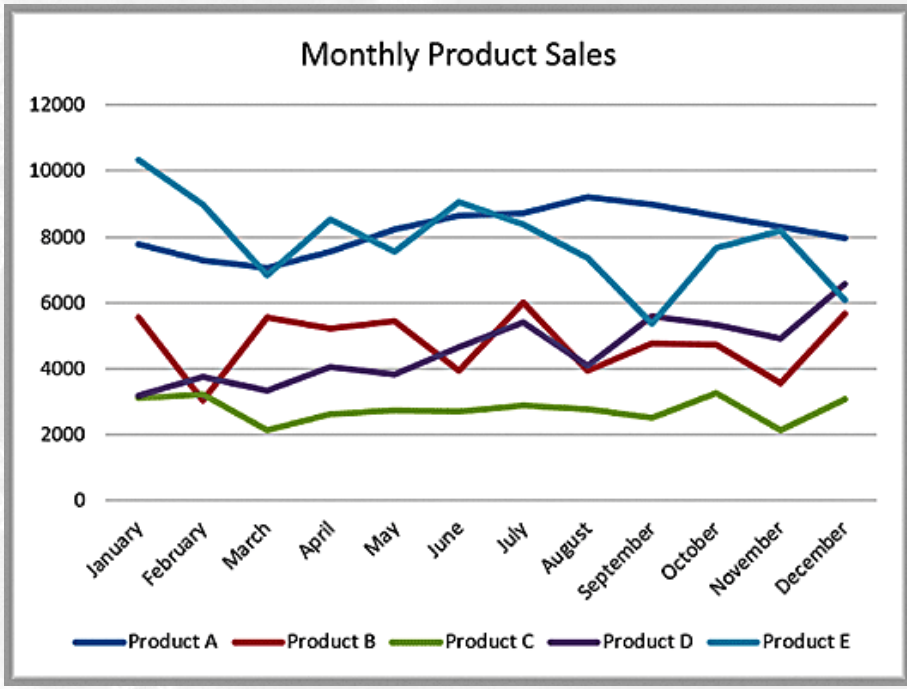
例 3.1：表格数据与可视化数据

- 表格数据提供了精确的数值信息，例如各产品每月销量，可以对此进行比较或计算
  - 例如，通过表格中数据的计算，发现A产品在二月份销量降了6.7%（1 - B3/B2）。但表格数据难以提供更宏观的结论（big picture）。

	A	B	C	D	E	F
1	Month	Product A	Product B	Product C	Product D	Product E
2	January	7792	5554	3105	3168	10350
3	February	7268	3024	3228	3751	8965
4	March	7049	5543	2147	3319	6827
5	April	7560	5232	2636	4057	8544
6	May	8233	5450	2726	3837	7535
7	June	8629	3943	2705	4664	9070
8	July	8702	5991	2891	5418	8389
9	August	9215	3920	2782	4085	7367
10	September	8986	4753	2524	5575	5377
11	October	8654	4746	3258	5333	7645
12	November	8315	3566	2144	4924	8173
13	December	7978	5670	3071	6563	6088

例3.1：表格数据与可视化数据

- 可视化数据提供了更宏观的直观信息：
  - 比较四种产品的总体销量（产品C销量最低）
  - 识别趋势性（产品D销量在增长，产品C销量较稳定，产品B销量波动较大，产品E销量在九月份有大幅下降）

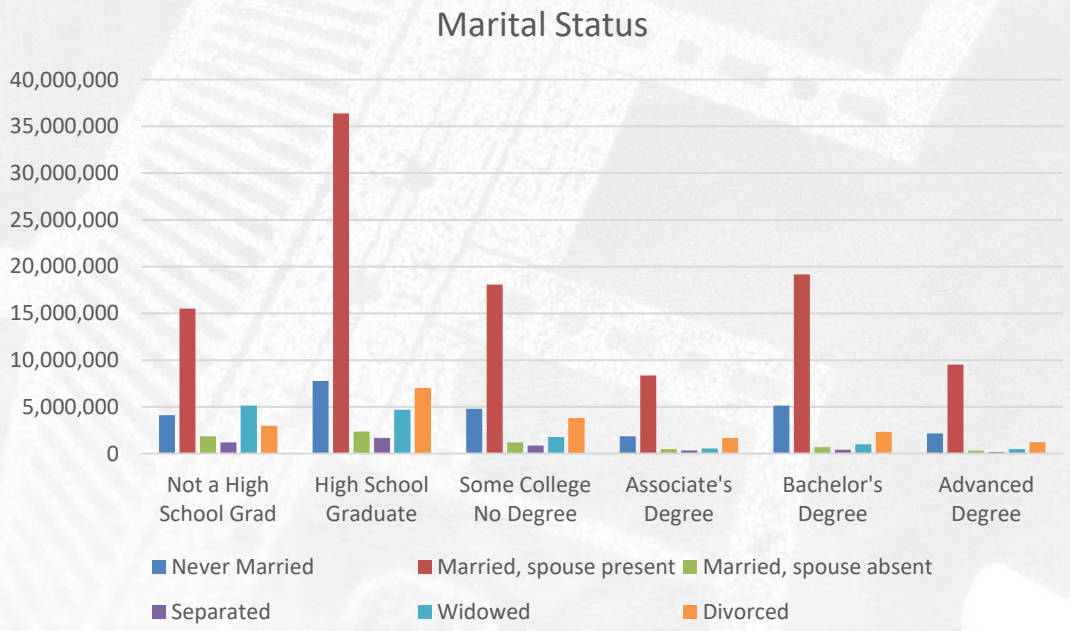


分类型数据（名义数据）

- 柱状图和条形图：比较不同类别的数量
- 柱状图：纵向；条形图：横向
  - 簇状柱形图：比较不同类型的数量
  - 堆积柱形图：比较不同类型的数量并查看其对总和的贡献
  - 百分比堆积柱形图：比较不同类型的数量并显示其在总体中的百分比

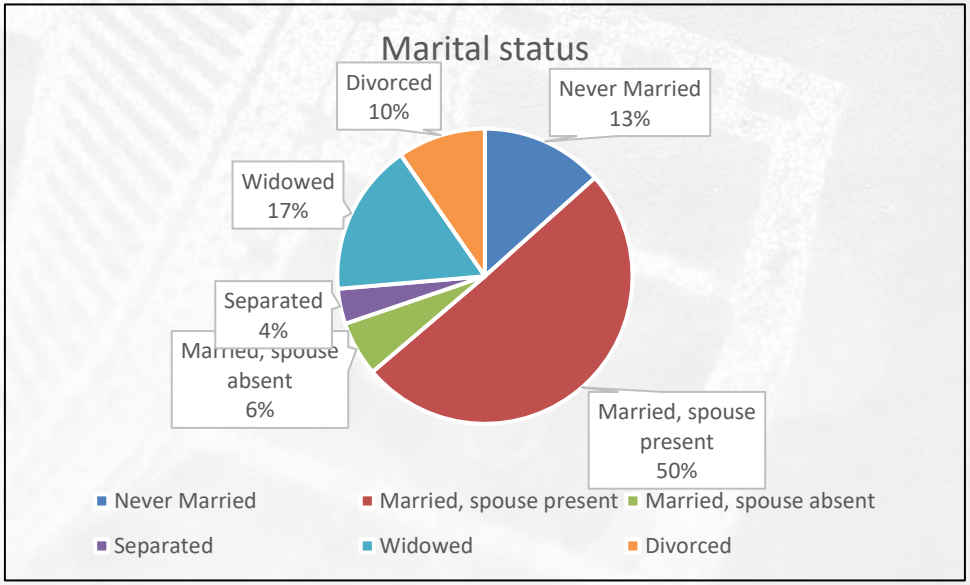
例3.2：绘制柱状图

例3.2 绘制柱状图：“Census Education Data”记录了人口普查数据中的教育情况。查看不同学历中各婚姻状况分布情况。选中A19:C24。



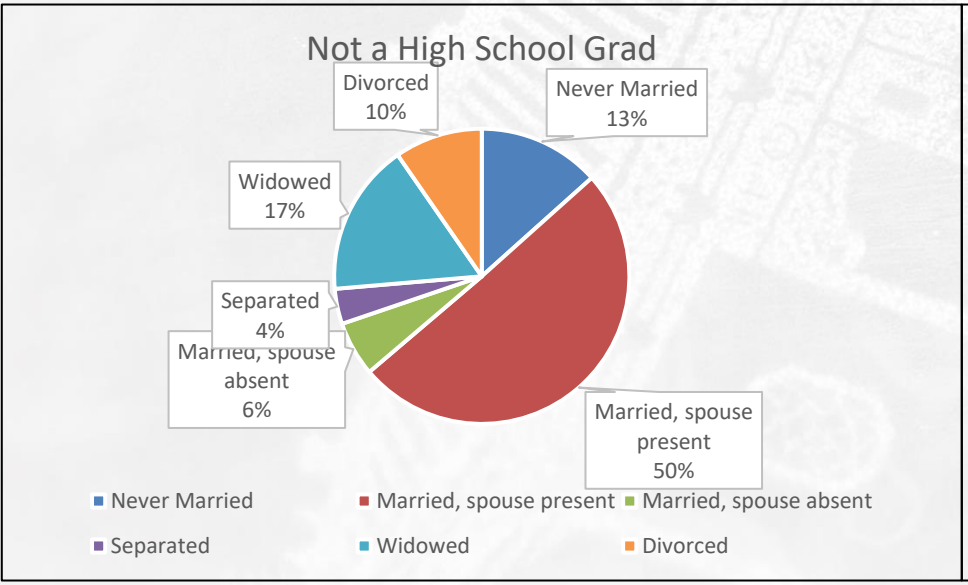
分类型数据

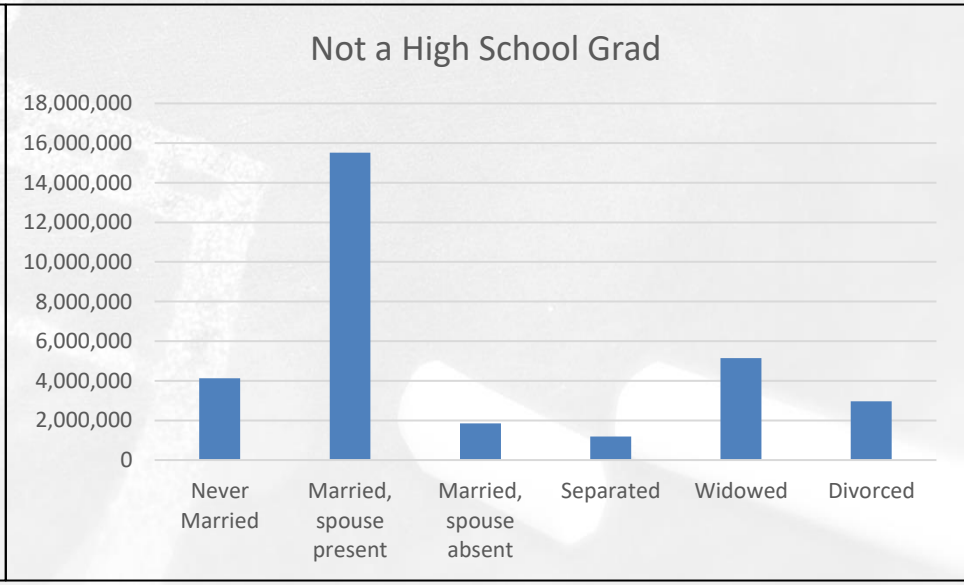
- 饼图
- 饼图展示了各类别所占百分比。
  - 例3.3 绘制饼图：”Census Education Data”记录了人口普查数据中的教育情况。查看高中以下学历中各婚姻状况分布情况。选中A19:B24。



饼图与柱状图

- 柱状图中更易显示出各类别的多少
- 饼图对比例的显示更直观
- 饼图不适用于类别太多的情况





上海财经大学 商學院 COLLEGE OF BUSINESS SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

思考题

- 利用柱状图和饼图对2020年全国第七次人口普查数据中感兴趣的数据进行分析。
- http://www.stats.gov.cn/sj/pcsj/rkpc/7rp/indexch.htm

汇总数据OR原始数据

- “Census Education Data”展示的是汇总数据
- “Purchase Orders”数据集展示的是原始数据

	A	B
100	Item Description	Frequency
101	Airframe fasteners	14
102	Bolt-nut package	11
103	Control Panel	4
104	Door Decal	2
105	Electrical Connector	8
106	Gasket	10
107	Hatch Decal	2
108	Machined Valve	4
109	O-Ring	12
110	Panel Decal	1
111	Pressure Gauge	7
112	Shielded Cable/ft.	11
113	Side Panel	8

频数

数据透视表

- 快速生成感兴趣的变量的汇总信息
  - “Purchase orders”中，“插入”-“数据透视表”，拖动感兴趣的变量至行、列中。“值”为需要汇总的量，可以是计数、加总、求平均值等。

行标签	计数项:Supplier
Airframe fasteners	14
Bolt-nut package	11
Control Panel	4
Door Decal	2
Electrical Connector	8
Gasket	10
Hatch Decal	2
Machined Valve	4
O-Ring	12
Panel Decal	1
Pressure Gauge	7
Shielded Cable/ft.	11
Side Panel	8
总计	94

数据透视图

- 数据透视图将数据透视表的结果可视化
- 在“数据透视表分析”下选择数据透视图
- 行标签处可以进行筛选

频率分布

- 频率是各类别数目占总数的比例
  - 如果一组数据有n个观测值，第i类的频率为

$$\frac{i\text{类的频数}}{n}$$

  - 频率一般用百分比表示
  - 频率分布用表格展示了各类别的频率
  - 值—值汇总方式—值显示方式

	A	B	C
100	Item Description	Frequency	Relative Frequency
101	Airframe fasteners	14	0.1489
102	Bolt-nut package	11	0.1170
103	Control Panel	4	0.0426
104	Door Decal	2	0.0213
105	Electrical Connector	8	0.0851
106	Gasket	10	0.1064
107	Hatch Decal	2	0.0213
108	Machined Valve	4	0.0426
109	O-Ring	12	0.1277
110	Panel Decal	1	0.0106
111	Pressure Gauge	7	0.0745
112	Shielded Cable/ft.	11	0.1170
113	Side Panel	8	0.0851
114	Total	94	1.0000

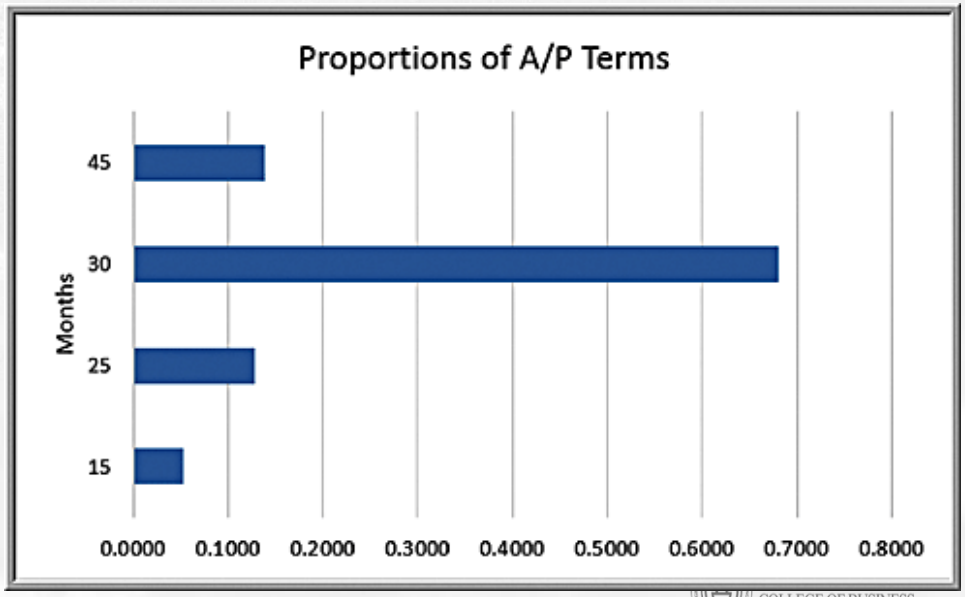
数量型数据（比率数据）

- 直方图：用柱状图展示的数量型数据频数分布图
  - 数据-数据分析-直方图
  - （文件-选项-加载项-分析工具库-转到-分析工具库-确定）
  - 定义“接收区域”：例如A/P terms 接收区域为15，25，30，45

数量型数据的频数分布

- 只有少量几个数值，与名义数据频数分布一样
  - 例3.4， A/P terms取值只有15，25，30，45

	A	B	C
117	A/P Terms	Frequency	Relative Frequency
118	15	5	0.0532
119	25	12	0.1277
120	30	64	0.6809
121	45	13	0.1383
122	Total	94	1.0000



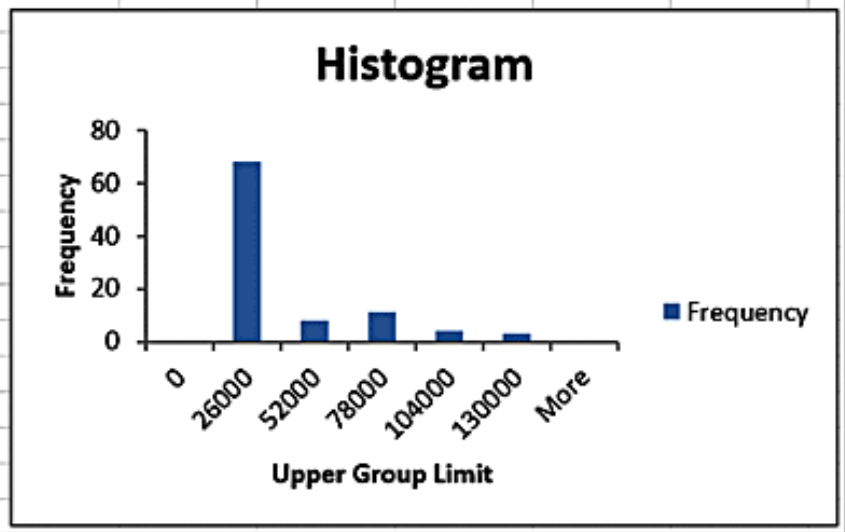
直方图

- 数据取值较多时如何定义接收区域？
  - 先确定分为几组（5-15组较合适）
  - 根据组数确定每组宽度
    - 每组宽度=（UL-LL）/组数
      - UL：大于等于数据中最大值的某一整数
      - LL:小于等于数据最小值的某一整数
  - 确定每组的上限和下限

例3.5 绘制Cost per Order直方图

- 数据范围为68.75 至127,500（排序可得）；
- LL设为0 ， UL设为130,000
- 选择5组，则每组宽度为 (130,000 – 0) / 5 = 26,000
- 接收区域为：0， 26000，52000，78000，104000，13000
  - 注意，以第二组为例，统计的是cost per order大于0小于等于26000的订数数量

	A	B
1	*Upper Group Limit*	*Frequency*
2	0	0
3	26000	68
4	52000	8
5	78000	11
6	104000	4
7	130000	3
8	More	0
9		
10		
11		
12		
13		
14		
15		



练习

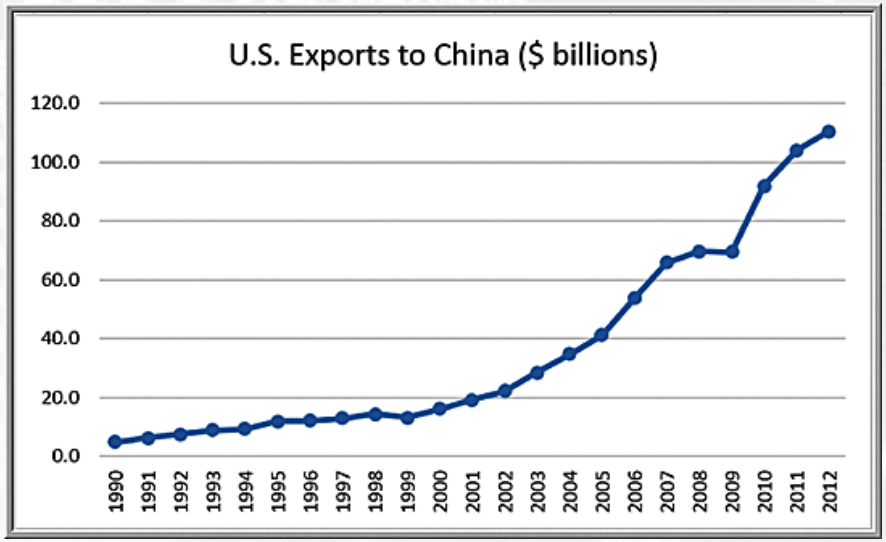
- 上题中组数选择10组，绘制cost per order直方图
- 累积百分率

帕累托分析

- 意大利经济学家Vilfredo Pareto在1906年发现意大利大部分财富由很少数的人拥有。
- 对“Bicycle Inventory”数据集中每个产品的价值做帕累托分析

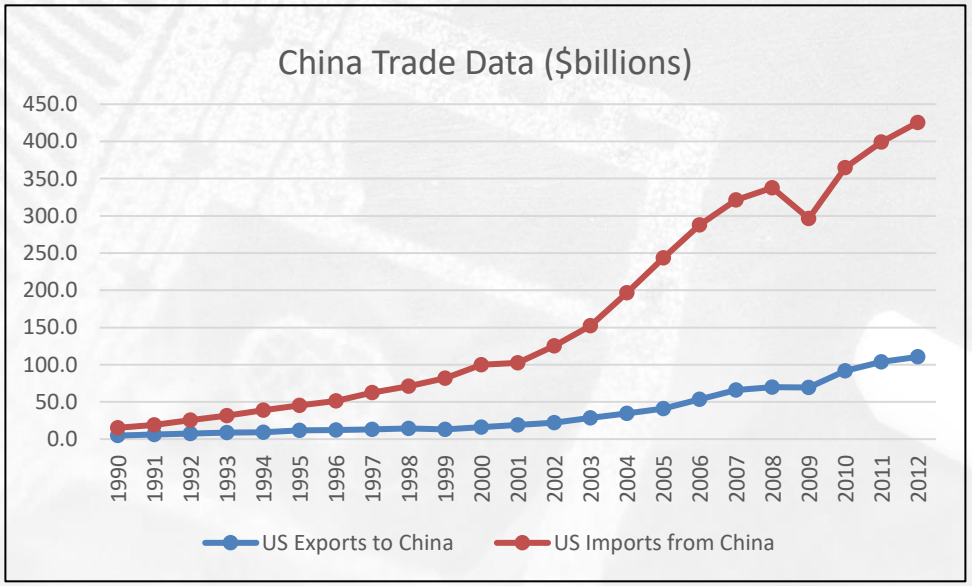
变量相关性可视化

- 折线图展示了数据的趋势
  - 注意：可以在一张折线图中展现多个数据系列，但当数量级差别较大时，显示效果较差，影响对数据的解读。因此，数量级差异大时，建议用多张图展现多个数据系列。



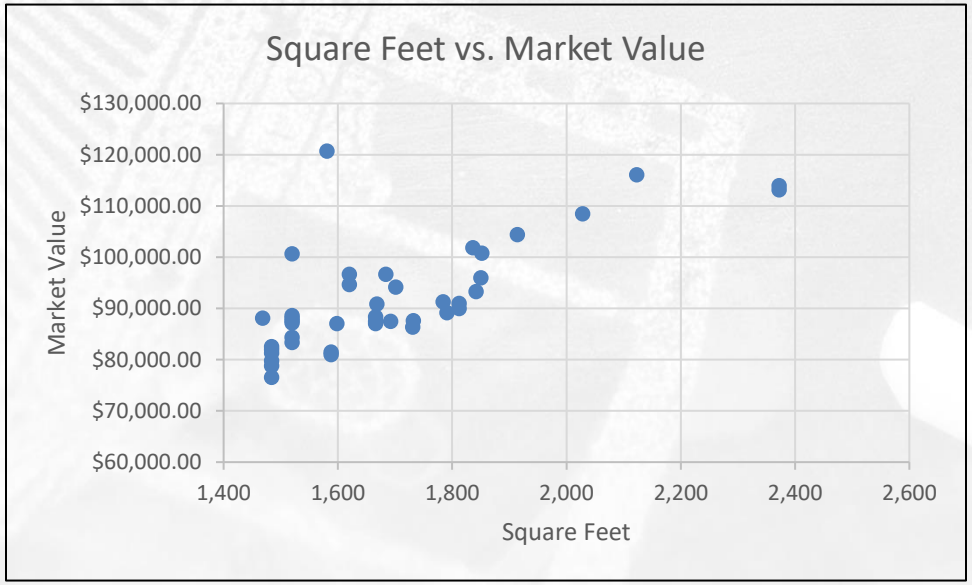
例3.6：绘制折线图

- “China Trade Data”展示了中美之间历年进出口数据。选择数据B3:C26.
- 面积图



散点图

- 散点图直观地展现了两个变量之间的关系，是分析两个变量关系的初步探索。
  - 例3.7：”Home Market Value”记录了房龄、面积、价值信息。研究面积与价值的关系。



交叉分组表

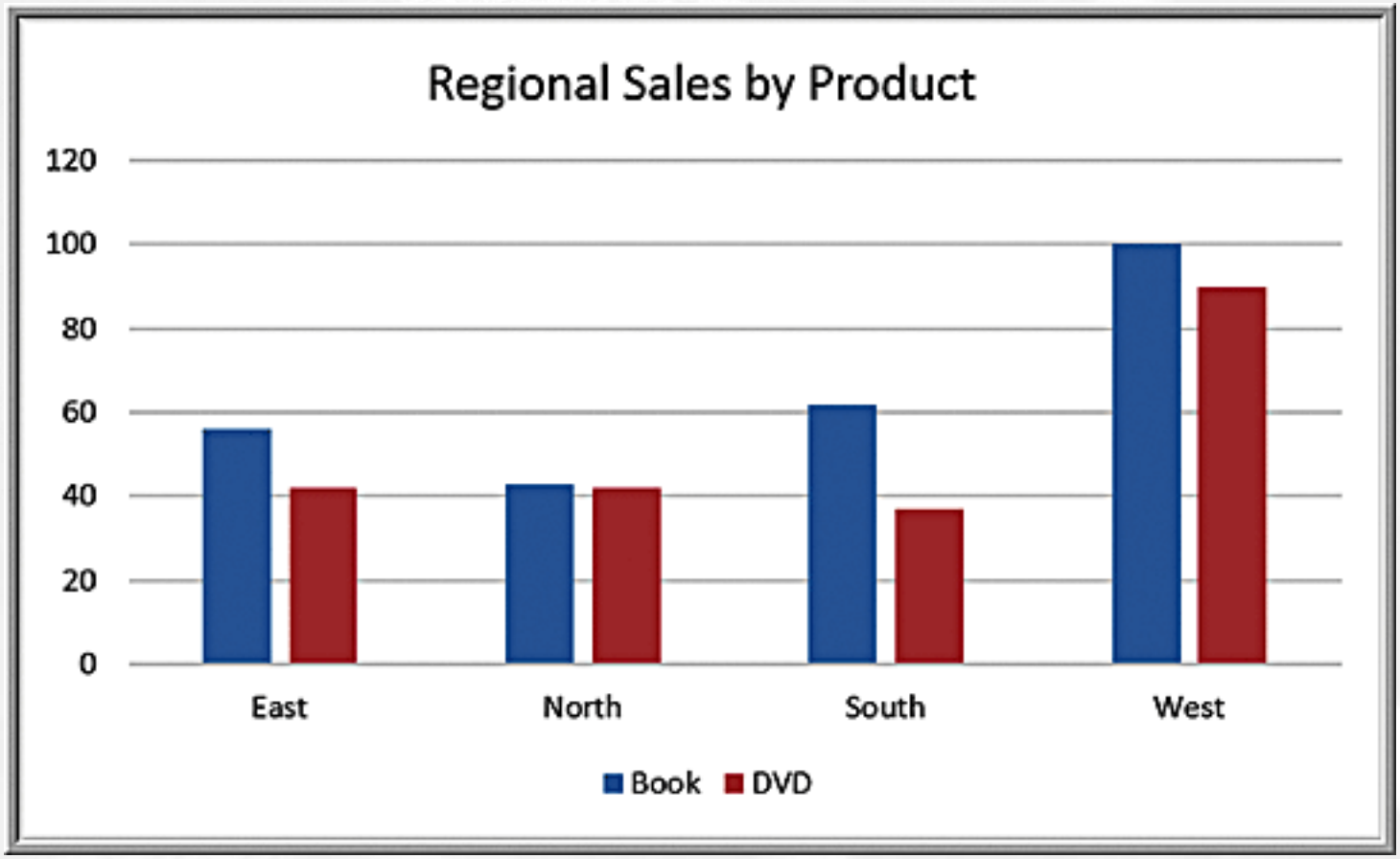
- 交叉分组表：分析两个类别变量之间关系的有效工具。例如在“Sales Transaction”数据集中，探索区域与购买产品的关系。

	A	B	C	D	E	F	G	H
1	Sales Transactions: July 14							
2								
3	Cust ID	Region	Payment	Transaction Code	Source	Amount	Product	Time Of Day
4	10001	East	Paypal	93816545	Web	$20.19	DVD	22:19
5	10002	West	Credit	74083490	Web	$17.85	DVD	13:27
6	10003	North	Credit	64942368	Web	$23.98	DVD	14:27
7	10004	West	Paypal	70560957	Email	$23.51	Book	15:38
8	10005	South	Credit	35208817	Web	$15.33	Book	15:21
9	10006	West	Paypal	20978903	Email	$17.30	DVD	13:11
10	10007	East	Credit	80103311	Web	$177.72	Book	21:59
11	10008	West	Credit	14132683	Web	$21.76	Book	4:04
12	10009	West	Paypal	40128225	Web	$15.92	DVD	19:35
13	10010	South	Paypal	49073721	Web	$23.39	DVD	13:26

Region	Book	DVD	Total
East	56	42	98
North	43	42	85
South	62	37	99
West	100	90	190
Total	261	211	472

Region	Book	DVD	Total
East	57.1%	42.9%	100.0%
North	50.6%	49.4%	100.0%
South	62.6%	37.4%	100.0%
West	52.6%	47.4%	100.0%

交叉分组表—柱状图


Regional Sales by Product
120
100
80
60
40
20
0
East
North
South
West
Book
DVD

数据透视表

- 快速制作交叉分组表并提供相关分析
  - 例3.8：“Sales Transactions”中，“插入”-“数据透视表”，拖动感兴趣的变量至行、列中。“值”为需要汇总的量，可以是计数、加总、求平均值等。

计数项:Cust ID	列标签		
行标签	Book	DVD	总计
East	56	42	98
North	43	42	85
South	62	37	99
West	100	90	190
总计	261	211	472

数据透视表

- 在“行”中增加“Payment”
- 汇总值为订单金额的平均值
- 只显示“East”和“North”的情况
- 按“Payment”方式进行筛选
- 插入“切片器”，选择“Source”，可进一步进行筛选性展示

辛普森悖论

- 根据汇总交叉数据表和未汇总交叉数据表得出的结论可能相反（统计学家辛普森提出）
  - 例3.9：法官勒基特和肯德尔在民事庭和市政庭主持审理案件，他们判决的部分案件被提出上诉。上诉法庭对大多数上诉案件维持原来的判决，但也有部分判决被推翻。以两个变量—判决（维持或推翻）和法庭类型（民事庭或市政庭）为依据，对每位法官构建交叉分组表。

辛普森悖论

判决	法官		总计
	勒基特	肯德尔	
维持	129（86%）	110（88%）	239
推翻	21（14%）	15（12%）	36
总计（%）	150（100%）	125（100%）	275

辛普森悖论

判决	法官勒基特		总计
	民事庭	市政庭	
维持	29（91%）	100（85%）	129
推翻	3（9%）	18（15%）	21
总计（%）	32（100%）	118（100%）	150

判决	法官肯德尔		总计
	民事庭	市政庭	
维持	90（90%）	20（80%）	110
推翻	10（10%）	5（20%）	15
总计（%）	100（100%）	25（100%）	125

辛普森悖论

- 市政庭被推翻的案件比例较高
- 法官勒基特审理的案件大多数在市政庭
- 在得出结论前应该审查交叉分组表是综合还是未综合形式
- 当交叉分组表包括综合数据时，应该审查是否存在可能影响结论的隐藏变量（法庭类型）

词云（word cloud）

- 文本数据可视化方法
- 对文本中出现频率较高的关键词进行视觉上的突出处理，出现频率越高的词显示越大

十四五规划的词云图

建设
发展
推进
完善
善
提升
体系
制度
推动
健全
机制
国家
提高
加快
服务
深化
实施
创新
强化
保障
改革
水平
战略
管理
企业
产业
能力
保护
构建
支持
经济
优化
重点
领域
政策
规划
治理
投资
统筹
人民
基础设施
城市
教育
资源
文化
农业
区域
开放
布局
政府
落实
基础
生态
市场
科技
技术
功能
网络
数字
社会
组织
风险
健康
合作
结构
第一节
第二节
第三节
统一
环境
消费
建立
现代化
工程
乡村
打造
鼓励
增强
国际
就业
高水平
全球
平台
金融
供给
监管
规范
优势
质量
中心
农村
综合
持续
作用
人才
绿色
公共服务
主体

十九大报告全文的词云图

二十大报告全文的词云图