北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

Data Analytics with R
语言数据分析

观数以形

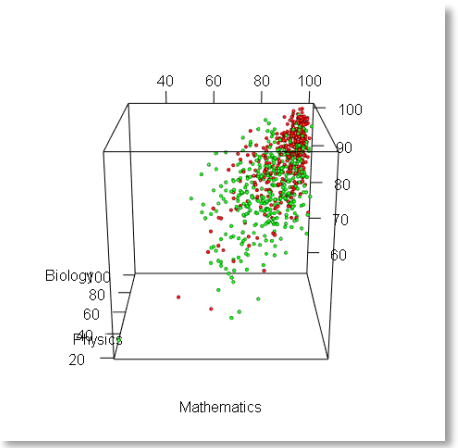
艾新波 / 2018 • 北京

课程体系

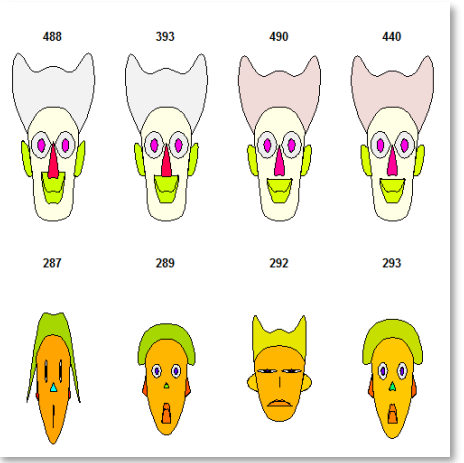
- R语言数据分析
  - 上部：论道 
    - 第1章 气象万千、数以等观
    - 第2章 所谓学习、归类而已
    - 第3章 格言联璧话学习
    - 第4章 源于数学、归于工程
  - 中部：执具 
    - 第5章 工欲善其事必先利其器
    - 第6章 基础编程
    - 第7章 数据对象
  -
    - 第8章 人人都爱tidyverse
    - 第9章 最美不过数据框
  - 下部 博术 
    - 第10章 观数以形
    - 第11章 相随相伴、谓之关联
    - 第12章 既是世间法、自当有分别
    - 第13章 方以类聚、物以群分
    - 第14章 庐山烟雨浙江潮

高维数据空间形态

三维散点图

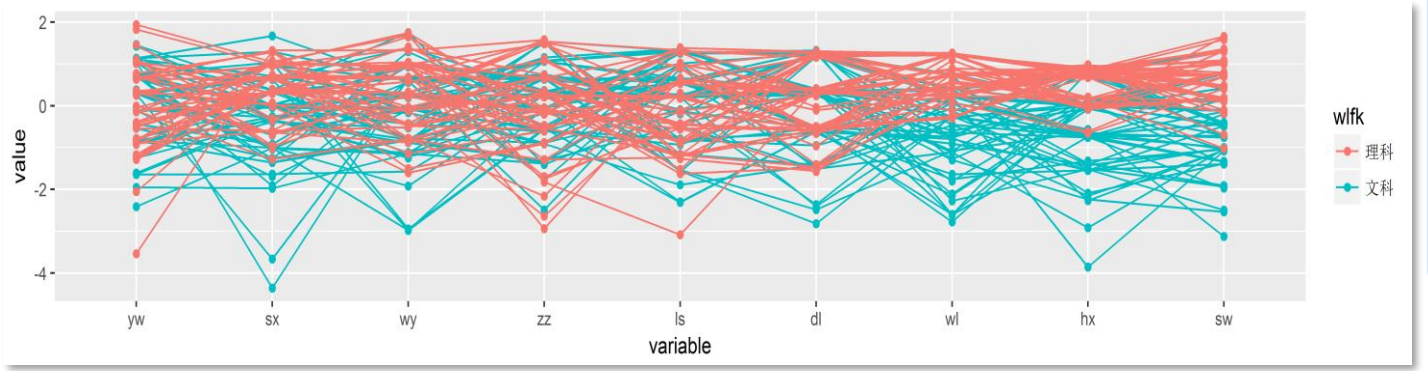


脸谱图



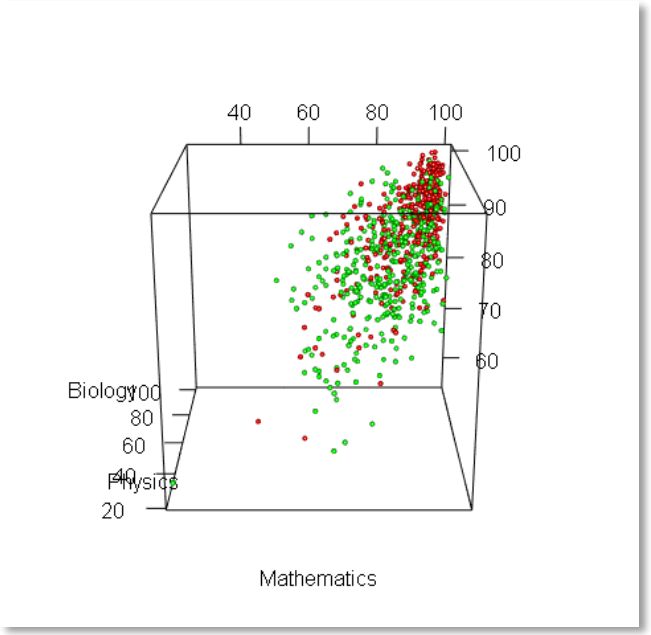
高维数据空间形态

平行坐标图



三维散点图

```
#绘制三维散点图
library(rgl)
plot3d(
  x = cjb$sx,
  y = cjb$wl,
  z = cjb$sw,
  xlab = "Mathematics",
  ylab = "Physics",
  zlab = "Biology",
  type = "s",
  size = 0.6,
  col = c("red", "green")[cjb$wlfk])
```

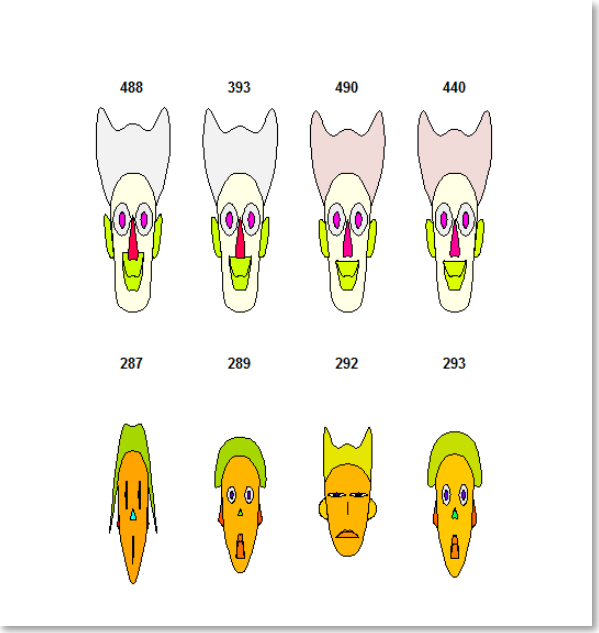


脸谱图

```
#绘制脸谱图
library(aplpack)
selected_cols <- c("wl", "hx", "sw")
selected_rows <-
  c(488, 393, 490,  440,
    287, 289,  292, 293)
faces(cjb[selected_rows,
          selected_cols],
      ncol.plot = 4,
      nrow.plot = 2,
      face.type = 1)
```

脸谱图

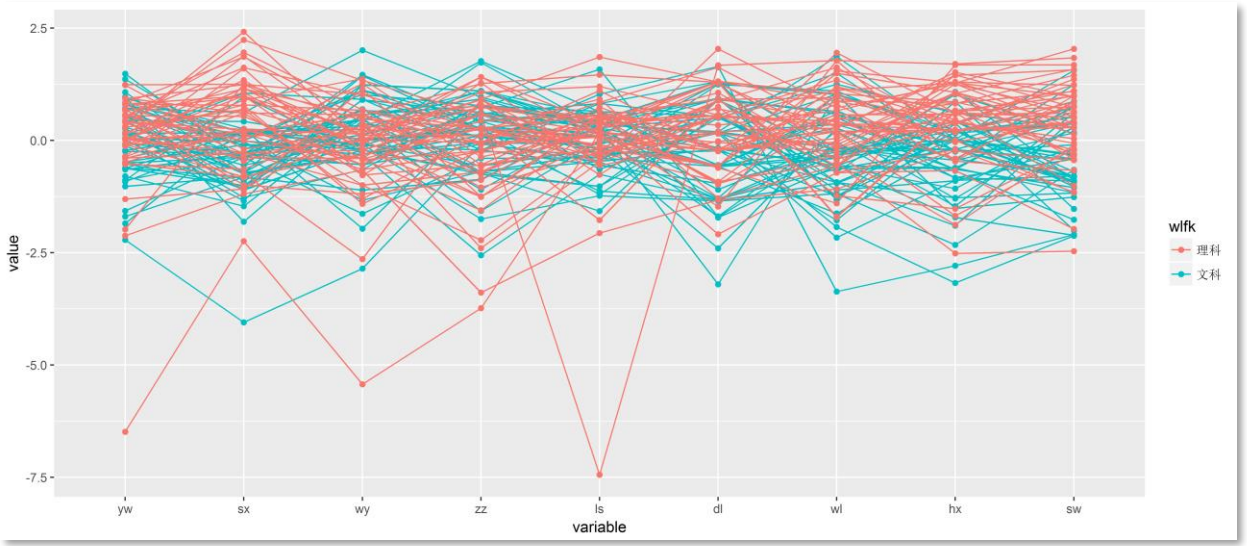
```
#> effect of variables:
#>   modified item       Var
#> "height of face    " "wl"
#> "width of face     " "hx"
#> "structure of face" "sw"
#> "height of mouth   " "wl"
#> "width of mouth    " "hx"
#> "smiling           " "sw"
#> "height of eyes    " "wl"
#> "width of eyes     " "hx"
#> "height of hair    " "sw"
#> "width of hair    "  "wl"
#> "style of hair    "  "hx"
#> "height of nose   "  "sw"
#> "width of nose    "  "wl"
#> "width of ear     "  "hx"
#> "height of ear    "  "sw"
```



平行坐标图

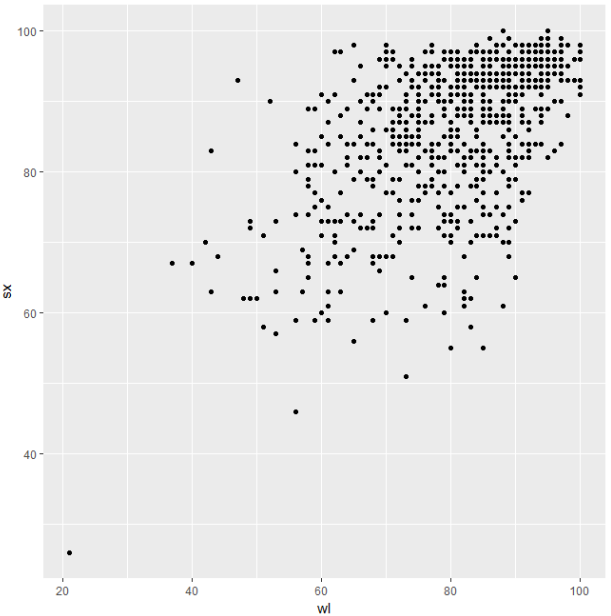
```
cjb_top_w <- cjb %>%
  filter(wlfk == "文科") %>%
  arrange(zcj) %>%
  select(4:13) %>%
  mutate_at(vars(yw:sw), jitter) %>%
  head(n = 50)
# (采用同样的方法得到理科cjb_top_l, 此处略)
cjb_top <- rbind(cjb_top_w, cjb_top_l)
Ggally::ggparcoord(cjb_top,
           columns = 1:9,
           groupColumn = 10) +
  geom_point()
```

平行坐标图



数据空间的其它形态：密度

- 密度是数据空间的形态之一
- 物理上的密度是指物质每单位体积内的质量：$\rho = \frac{m}{V}$
- **数据空间当然没有质量，密度只能是指密集程度而已**
- 一个简单的计算方法是：单位面积/体积内数据点的个数
- 换言之，这里的密度，指的是单位体积内的数量，而非质量



数据空间的其它形态：密度

#汇总统计

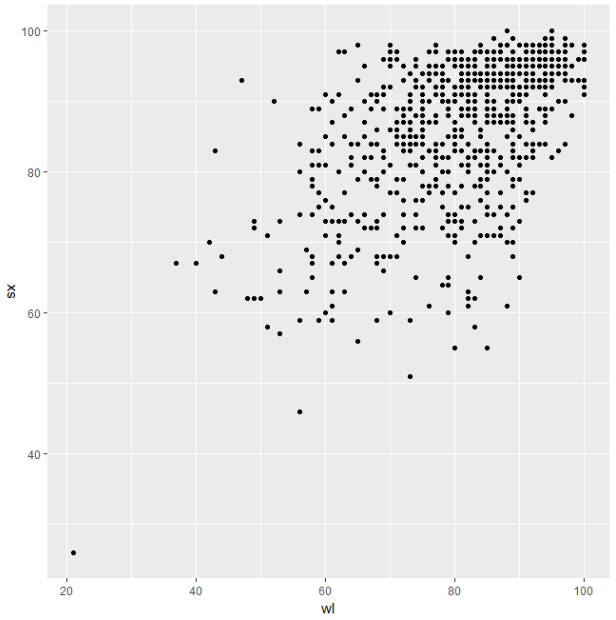
```
breaks <-  c(0,seq(50, 100, len=11))
wl_sx_freq <- cjb %>%
  select(wl, sx) %>%
  mutate_at(
    vars(wl, sx),
    function(x) {
      cut(x, breaks = breaks)
    }) %>%
  group_by(wl, sx) %>%
  summarise(freq = n()) %>%
  complete(wl, sx, fill = list(freq = 0))
```

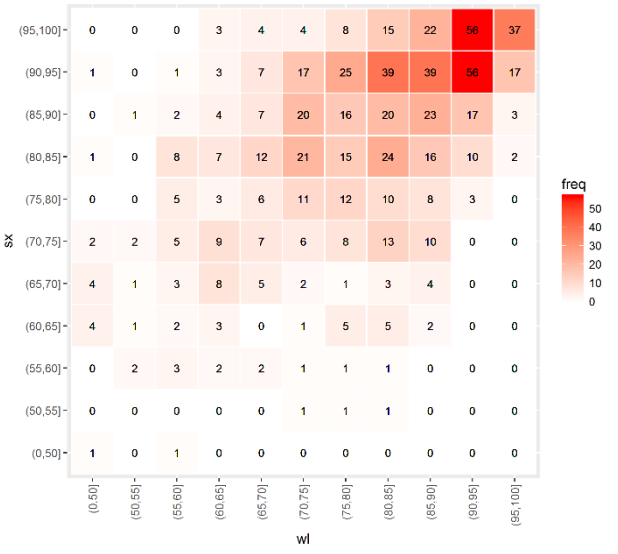
```
# A tibble: 1,331 x 3
# Groups:   wl [11]
   wl      sx       freq
   <fct>   <fct>   <dbl>
 1 (0,50] (0,50]      1
 2 (0,50] (50,55]     0
 3 (0,50] (55,60]     0
 4 (0,50] (60,65]     4
 5 (0,50] (65,70]     4
 6 (0,50] (70,75]     2
 7 (0,50] (75,80]     0
 8 (0,50] (80,85]     1
 9 (0,50] (85,90]     0
10 (0,50] (90,95]     1
# ... with 1,321 more rows
```

数据空间的其它形态：密度

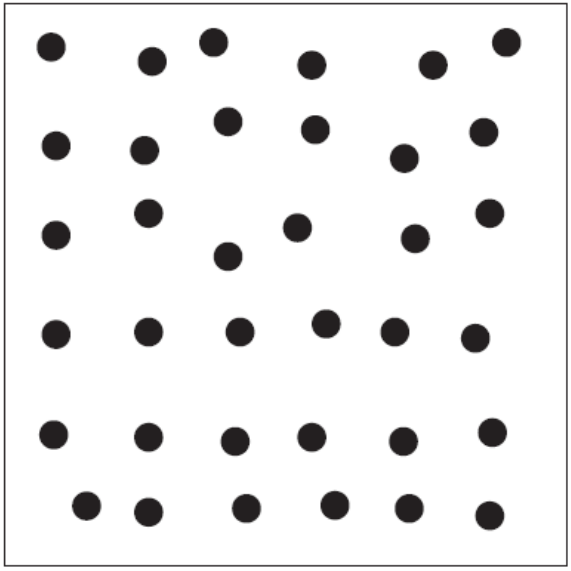
```
ggplot(wl_sx_freq, aes(x = wl, y = sx, fill = freq)) +
  geom_tile(colour="white", size = 0.5) +
  geom_text(aes(label = freq), size = 3) +
  scale_fill_gradient(
    low = "white",
    high = "red")+
  theme(axis.text.x =
          element_text(
            angle = 90,
            hjust = 1,
            vjust = 0.5)) +
  coord_fixed()
```

数据空间的其它形态：密度





数据空间的其它形态：均匀程度



A data set that is uniformly distributed in the data space

数据空间的其它形态：均匀程度

Hopkins统计量告诉我们所拿到的数据多大程度上接近于均匀散布的形态

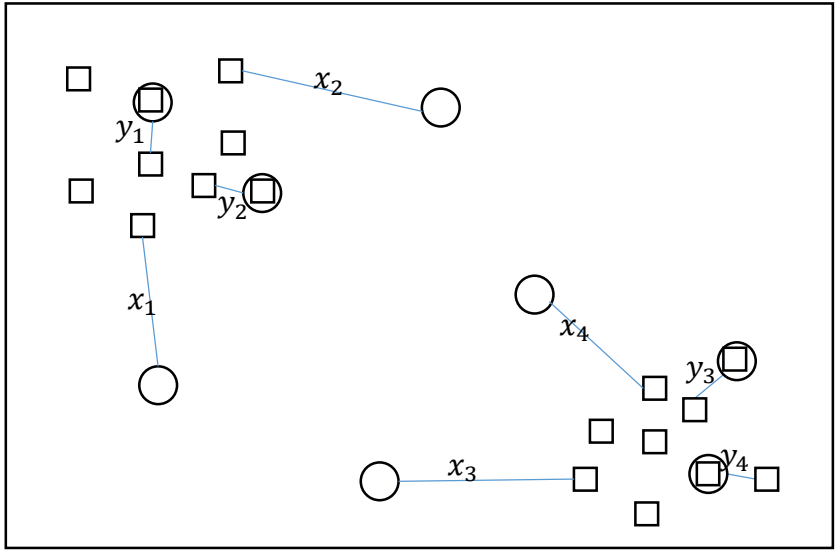
对于给定数据集 D：

(1) 均匀地从D的空间中抽取 n 个点 $p_1, p_2, \ldots, p_n$，对于每个点 p_i 令x_i为p_i与它在D中的最近邻的距离：$x_i = \min_{v \in D}\{dist(p_i, v)\}$

(2) 从D中抽取 n 个点$q_1, q_2, \ldots, q_n$，对于每个点 q_i，令y_i为q_i与它在D-$\{q_i\}$中最近邻之间的距离：$y_i = \min_{v \in D-\{q_i\}}\{dist(q_i, v)\}$

(3) 计算霍普金斯统计量：$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$

数据空间的其它形态：均匀程度



霍普金斯统计量：

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

若D分布均匀，则H接近于0.5

若D是高度倾斜的，则H接近于0

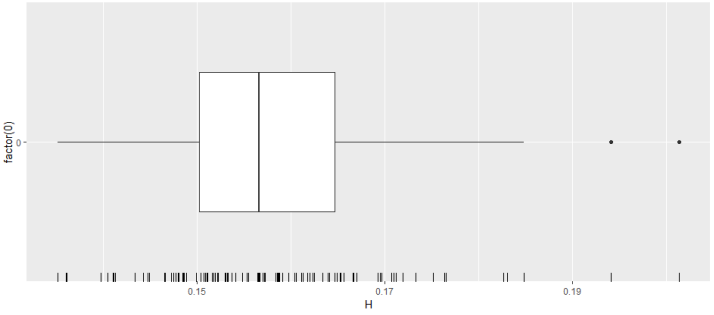
一般而言， $n \ll |D|$

推荐的做法：

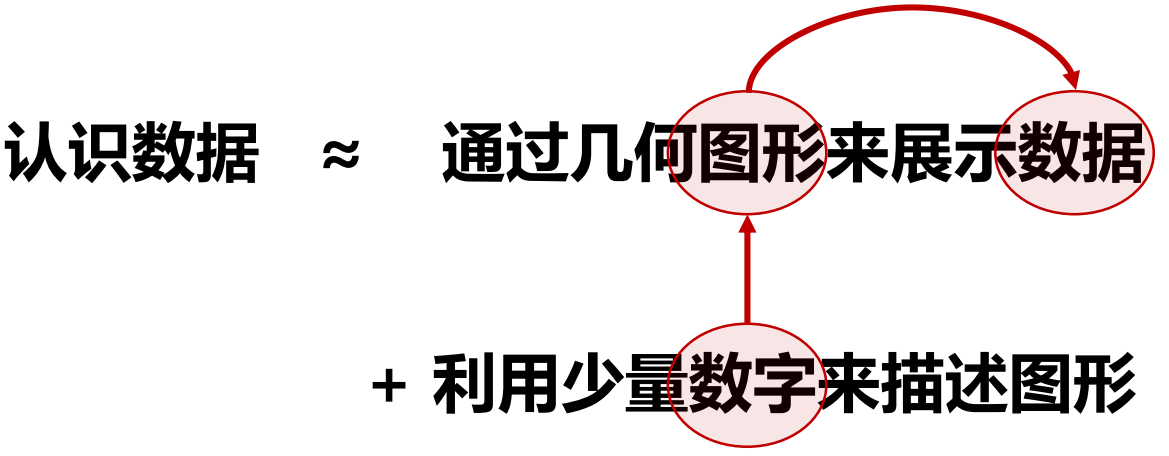
$n = 0.05 \times |D|$ *或是* $n = 0.1 \times |D|$

数据空间的其它形态：均匀程度

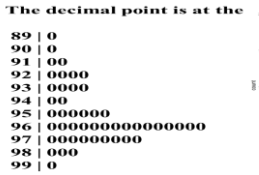
```
library(clustertend)
set.seed(2012)
scores <- cjb %>%
  select(yw:sw)
n <- floor(nrow(cjb) * 0.05)
hopkins_stat <- unlist(replicate(100, hopkins(scores, n)))
mean(hopkins_stat)
#> [1] 0.1577968
```

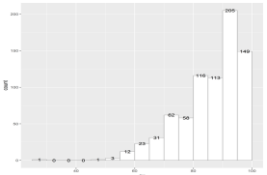


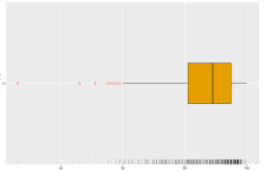
内容小结

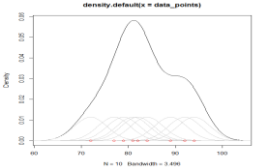


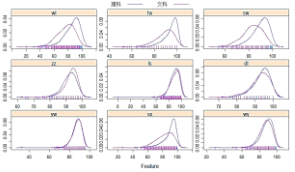
内容小结

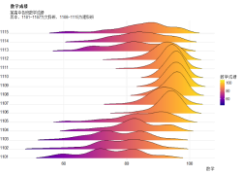


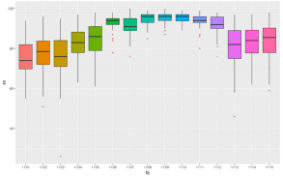


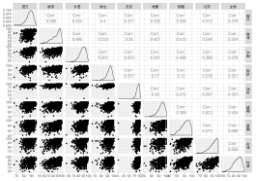


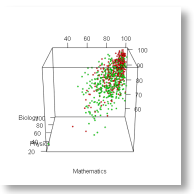












通过几何图形来展示数据

内容小结

1．集中的趋势：均值、中位数、众数、……

2．分散的程度：标准差、方差、极差、……

3．峰度、偏度

4．均匀程度

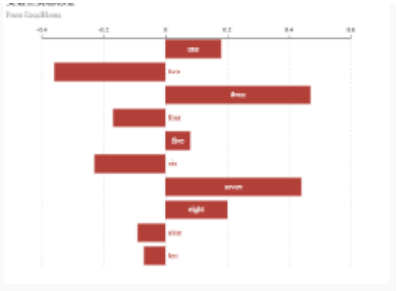
5．密度

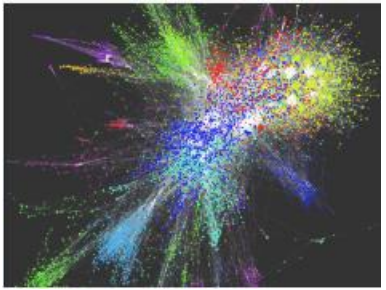
6．相关系数

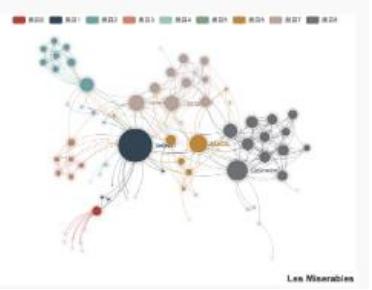
7．……

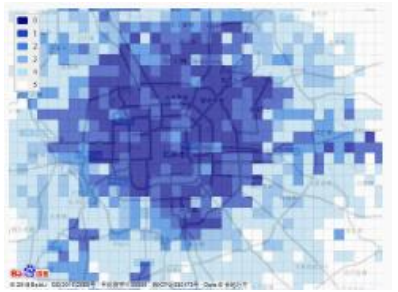
利用少量数字来描述图形

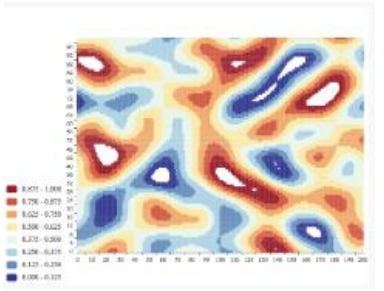
更多内容……

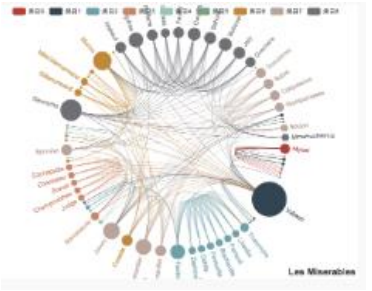












谢谢聆听
Thank you

教师个人联系方式

艾新波

手机：13641159546

QQ： 23127789

微信：13641159546

E-mail: 13641159546@126.com

axb@bupt.edu.cn

地址：北京邮电大学科研楼917室

课程网址：https://github.com/byaxb/RDataAnalytics