美国数据科学家职位数据分析

1 数据与方法

1.1 数据说明与问题提出

本报告以美国数据科学家工作职位数据为研究对象，数据来源于领英网站上发布的职位信息。共收集到样本数据 6964 条，包含了 position、company、description、reviews 和 location 五个字段的内容，position 即具体的职位名称，company 即职位对应的公司名称，description 即职位描述，reviews 即对公司职位的评论数量，location 即职位所在地。

通过对字段的初步解读，提出以下拟分析的问题：

（1）不同地区对数据科学家职位的需求情况；

（2）不同公司对数据科学家职位的需求情况；

（3）数据科学家职位的热度值，或受关注程度；

（4）数据科学家职位描述的核心内容，以及主要关键词；

（5）数据科学家职位对应的工具型技能的需求。

1.2 数据预处理

本报告使用 R 语言对美国数据科学家工作职位数据进行分析，需要对数据进行预处理，得到格式规范的数据用于后续分析，具体的方法流程如下：

（1）加载原始数据

即将本地的数据加载到 R 语言环境中，用于后续分析。

```
# 读取数据
usdata_scientist <- read.csv("E:/mine/R/usdata_scientist858/usdata_scientist.csv")
```

（2）查看是否有缺失值

用 VIM 包的 aggr 函数来识别数据集是否存在缺失值。

```
library(VIM)
# 识别缺失值
aggr(usdata_scientist,prop=T,numbers=T)
```

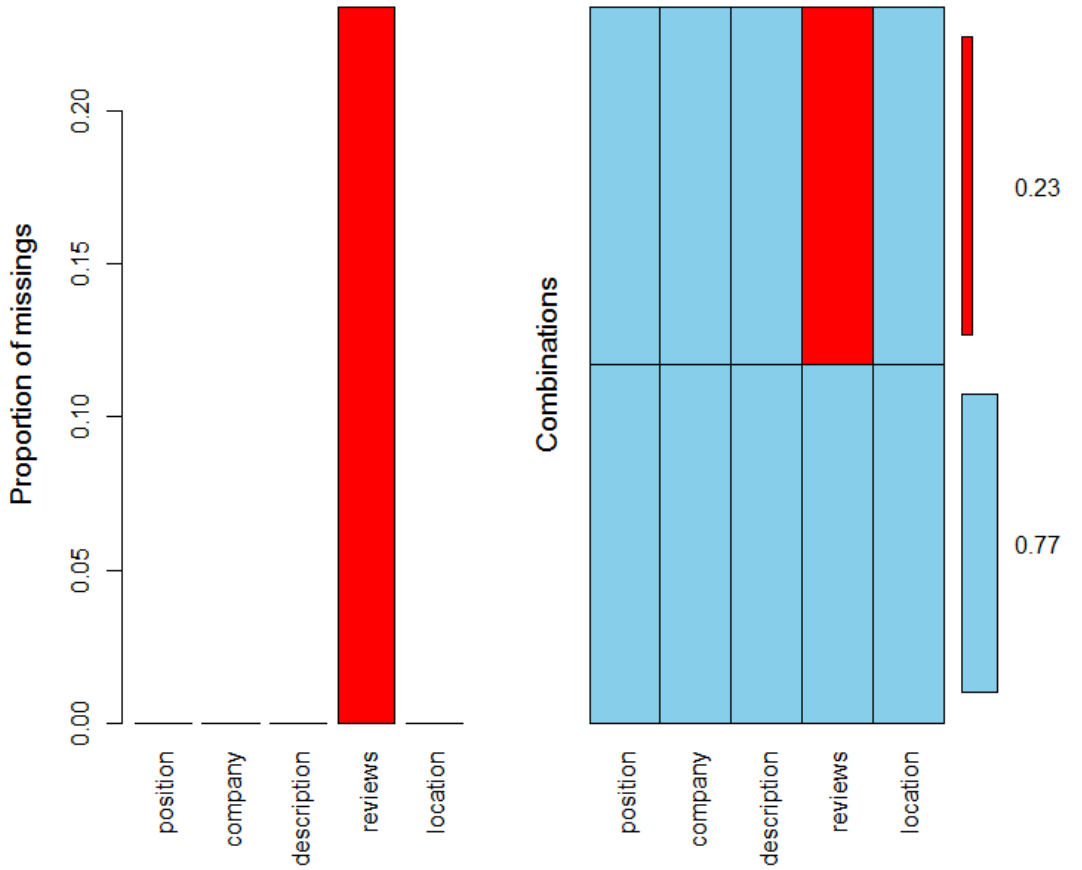

图 1 数据缺失值情况

由上图可以看出，reviews 列中存在缺失值，且缺失值占了总数的 23%，在涉及该列数据的分析中，需要对这部分数据作进一步处理。

（3）删除空白行

在数据分析过程中，发现数据集中存在完全空白的行，需要将这些空白行逐一删除。

```
Warning message:
Expected 2 pieces. Missing pieces filled with `NA` in 11 rows [303, 332, 467, 483, 3338
, 5015, 5060, 5104, 5115, 6094, 6169].
```

图 2 空白行情况

将系统提示的 11 条空白行删除，得到数据集中包含 6953 条数据。

	position	company	description	reviews	location
1	Development Director	ALS TDI	Development Director ALS Therapy Devel...	NA	Atlanta, GA 30301
2	An Ostentatiously-Ex...	The Hexag...	Job Description "The road that leads to a...	NA	Atlanta, GA
3	Data Scientist	Xpert Staffi...	Growing company located in the Atlanta,...	NA	Atlanta, GA
4	Data Analyst	Operation ...	DEPARTMENT: Program OperationsPOSITI...	44	Atlanta, GA 30303
5	Assistant Professor -...	Emory Uni...	DESCRIPTION The Emory University Depar...	550	Atlanta, GA
6	Manager of Data En...	McKinsey ...	Qualifications Bachelor's degree in Com...	385	Atlanta, GA 30318
7	Product Specialist - P...	McKinsey ...	Qualifications Bachelor's degree 5-7 year...	385	Atlanta, GA 30318
8	Junior to Mid-level E...	Wood	Overview / Responsibilities Wood Enviro...	899	Atlanta, GA
9	Analyst - CIB Credit R...	SunTrust	Works closely with senior CIB profession...	3343	Atlanta, GA
10	Senior Associate - Co...	KPMG	Known for being a great place to work a...	4494	Atlanta, GA 30338

Showing 1 to 11 of 6,953 entries

图 3 删除空白行后的数据集

（4）处理数据

通过对数据字段内容的分析以及本文所研究问题的需要，对数据进行处理。

通过 tidyverse 包下的 separate()函数将 location 字段下的（城市，州）的数据分隔，分别存储为两列 city 和 state，并将结果赋值给新的数据框 usdata_scientist_new：

```
# 分隔列
library(tidyverse)
usdata_scientist_new <- separate(data = usdata_scientist, col = location, into = c("city", "state"), sep = ", ", remove = FALSE)
```

通过 gsub()函数，利用正则表达式将 state 列中的邮编数字去除：

```
# 去掉 state 列的邮编数字
usdata_scientist_new$state <- gsub('[0-9]*|\\s',",usdata_scientist_new$state)
```

将 position、state 和 city 变量转换为因子类型，并将 position 转化为小写，消除大小写差异，便于后续分析：

```
# 将数据类型转换为因子类型
usdata_scientist_new$position <- factor(tolower(usdata_scientist_new$position))
usdata_scientist_new$state <- factor(usdata_scientist_new$state)
usdata_scientist_new$city <- factor(usdata_scientist_new$city)
```

在对 description 列进行的文本挖掘操作中，需要对文本进行分词和去停用词的处理，该步骤的处理需要将 description 字段中的文本由因子型转换为字符型：

```
# 用于文本挖掘
usdata_scientist_new$description <- as.character(usdata_scientist_new$description
```

1.3 美国数据科学家职位数据分析

本报告使用 R 语言对美国数据科学家工作职位数据进行分析，具体内容如下：

（1）总体描述统计信息查看

summary()函数提供了最小值、最大值、四分位数和数值型变量的均值，以及因子向量和逻辑型向量的频数统计。通过该函数对数据集总体描述统计信息进行查看：

```
#观察数据的总体描述统计信息
summary(usdata_scientist_new)
```

（2）美国各个州拥有的职位数情况

目的：以州为分类单位，对各个州的公司发布的数据科学家的职位的数量进行统计，可看出该职位在各个州的需求情况。

方法：通过对 state 列的计数，可实现职位数量的统计。采用 ggplot2 包的条形图对该统计信息进行呈现，具体代码如下：

```
# 对 state 计数
library(plyr)
state_freq <- count(usdata_scientist_new$state)
library(ggplot2)
# 各个州拥有的职位数情况
names(state_freq) <- c("State","Freq") # 重命名列名
ggplot(state_freq,aes(x = reorder(State,-Freq), y= Freq, fill = State))+
  geom_bar(stat = "identity")+
  geom_text(aes(label=Freq),vjust=-0.2,size=3)+ # 添加数值标签
  xlab("State") # 修改 x 轴坐标文本
```

（3）职位需求在美国各个州的地理分布情况

目的：将拥有数据科学家职位需求的州在美国地图上标示出来，可反映该职位在地理上的分布情况以及集中程度。

方法：以每个州的首府作为该州的代表，根据首府的经纬度将其位置显示在地图上，位置点的分布即代表了职位需求在各个州的分布情况。采用 maps 和 mapdata 包进行作图：

（注：使用的经纬度来源：“在线地图经纬度查询：http://www.gpsspg.com/maps.htm”）

```
# 职位在各个州的地理分布情况
library(maps)
library(mapdata)
par(mar=rep(0,4))
# 经纬度数据
dat = read.csv(text = "state-capital,jd,wd
                       California-Sacramento,-121.4943996,38.5815719
                       Colorado-Denver,-104.990251,39.7392358
                       Washington DC,-77.0368707,38.9071923
                       Georgia-Atlanta,-84.3879824,33.7489954
                       Illinois-Springfield,-72.5898110,42.1014831
                       Massachusetts-Boston,-71.0588801,42.3600825
                       New Jersey-Trenton,-74.7597170,40.2205824
                       New York-Albany,-73.7562317,42.6525793
                       Texas-Austin,-97.7430608,30.2671530
                       Washington-Olympia,-122.9006951,47.0378741")
# 地图
map("state", col = "lightblue", fill = TRUE, ylim = c(18, 54), panel.first = grid())
# 地图上显示的点
points(dat$jd, dat$wd, pch = 19, cex = 1.5, col = rgb(0,0, 0, 0.5))
# 地图上显示的文本
text(dat$jd, dat$wd, dat[, 1], cex = 0.7, col = rgb(0,0, 0, 0.7),
     pos = c(1,2,1,1,3,2,2,1,1,1))
# 四周的坐标轴
axis(1, lwd = 0); axis(2, lwd = 0); axis(3, lwd = 0); axis(4, lwd = 0)
```

（4）每个州中各城市拥有的职位数情况

目的：以每个州下的城市为分类单位，对每个城市的公司发布的数据科学家的职位的数量进行统计，可看出该职位在各个城市的需求情况。

方法：选取发布职位数大于 500 的州，分别抽取每个州的子集进行分析，通过对 city 列的计数，可实现各城市职位数量的统计。采用 ggplot2 包的条形图对该统计信息进行呈现，以 CA 州为例，具体代码如下：

```
# 选取 CA 的子集
CA_df <- usdata_scientist_new[which(usdata_scientist_new$state=="CA"),]
CA_city_freq <- count(CA_df$city)
# CA 州各城市的职位数
ggplot(CA_city_freq,aes(x=x,y=freq,fill=x))+
  geom_bar(stat = "identity")+
  geom_text(aes(label=freq),vjust=-0.2,size=2.5)+ #  添加数值标签
  theme(axis.text.x = element_text(angle = 90))+ # x 轴文本旋转 90 度
```

```
  xlab("CA-city")+ #  修改 x 轴坐标文本
  guides(fill=FALSE) #  删除图例
```

其他发布职位数大于 500 的州为 MA、WA 和 NY，分别抽取各自的子集进行统计分析，代码实现与 CA 州类似。

（5）公司对数据科学家职位的需求情况

目的：以公司为分类单位进行统计，可以分析各个公司对数据科学家职位的需求。

方法：通过对 company 列的计数，可实现对每个公司发布的职位数量的统计。取排名前 15 的公司子集，采用 ggplot2 包的条形图对该统计信息进行呈现，具体代码如下：

```
# 分析各公司对数据科学家岗位的需求
company_pos <- count(usdata_scientist_new$company)
# 各公司发布的职位数降序
library(dplyr)
#排序
company_pos_desc <- arrange(company_pos, desc(freq))
# 取前 15 名
company_pos_top <- company_pos_desc[1:15,]
company_pos_top
ggplot(company_pos_top,aes(x=reorder(x,-freq),y=freq,fill=x))+
  geom_bar(stat = "identity")+
  theme(axis.text.x = element_text(angle = 90))+ #x 轴文本旋转 90 度
  geom_text(aes(label=freq),vjust=-0.2,size=2.5)+
  xlab("company-top15")+
  guides(fill=FALSE)
```

（6）美国各个州中数据科学家职位的热度情况

目的：以州为分类单位，分析数据科学家职位在各个州的热度的总体情况。

方法：以评论数来描述职位热度，通过箱型图的形式对职位热度的总体情况进行呈现。由于 reviews 列存在缺失值，所以仅抽取 reviews 值未缺失的数据进行分析，具体代码如下：

```
# 过滤掉 reviews 为 NA 的行
reviews_com <- usdata_scientist_new[complete.cases(usdata_scientist_new[,5]),]
# 各个州的数据科学家热度的总体情况
ggplot(reviews_com,aes(x = state,y = reviews))+geom_boxplot()+
  ylim(0,max(reviews_com$reviews)) #  设置 y 轴精度
```

由于 reviews 的值范围跨度较大，使得箱型图的特征不能较好的显示，所以考虑将 y 轴的精度进行调整：

```
ggplot(reviews_com,aes(x = state,y = reviews))+geom_boxplot()+
  ylim(0,10000) #  设置 y 轴精度
```

（7）数据科学家职位描述关键词

目的：分析数据科学家职位描述中的关键词汇，了解该职位对应的工作内容、主要要求以及能力需求等信息。

方法：通过对 description 列的内容进行文本挖掘处理，统计词频获得高频数词汇，并通过词频统计图和词汇的词云图进行直观展示：

```
# 数据科学家职位描述的关键词
# 文本挖掘
library(tidytext)
```

```
tidy_words <- usdata_scientist_new%>%
  unnest_tokens(word,description)%>% # 分词
  anti_join(stop_words) # 去停用词
head(tidy_words)
word_freq <- tidy_words %>% count(word,sort=TRUE) # 计算词频并排序
head(word_freq)
word_freq[which(word_freq$word=="excel"),]
summary(word_freq)
# 词频统计图
library(ggplot2)
word_freq %>% # 统计词频
  filter(n > 5000) %>%
  mutate(word=reorder(word,n)) %>%
  ggplot(aes(word,n,fill=word))+
  geom_col()+
  xlab(NULL)+
  coord_flip()+
  guides(fill=FALSE) # 删除图例
# 词云图
library(wordcloud2)
wc_freq <- word_freq %>% filter(n > 3000) # 取词频大于 3000 的词画云图
nrow(wc_freq)
wordcloud2(wc_freq,size=1,shape = 'star')
```

（8）数据科学家职位对工具型技能的要求

目的：了解当前数据科学家职位对数据分析人员应该具有的能力和素质以及需掌握的技能的要求。

方法：通过对 description 列进行分词、去提用词等文本挖掘操作，将代表工具型技能的词汇抽取出来，对这些词汇进行统计分析，计算不同技能需求占总职位需求的比例，从中获得不同技能的重要程度。

```
# 数据科学家职位对工具型技能的要求
seg <- usdata_scientist_new %>%
  unnest_tokens(word,description) # 分词
# 转换为大写
seg$word <- toupper(seg$word)
# 只保留工具型技能的词
tools_df <- seg[which(seg$word %in% c("SQL","R","PYTHON",
                                      "EXCEL","SAS","SPSS","HIVE",
                                      "PPT","HADOOP","MYSQL",
                                      "SPARK","ORACLE","BI",
                                      "KPI","JAVA")),]
# 转换为因子类型
tools_df$word <- factor(tools_df$word)
# 求解频数
tools_freq <- count(tools_df$word)
```

```
# 降序输出
library(dplyr)
arrange(tools_freq, desc(freq))
# 计算占比
tools_freq$freq <- tools_freq$freq/nrow(usdata_scientist_new)
ggplot(tools_freq) +
  geom_bar(aes(x=reorder(x,-freq),y=freq),fill="lightblue",stat = "identity") +
  labs(x="工具型技能",y="不同技能需求占总职位需求量的比率") +
  theme(axis.text.x = element_text(angle = 30,hjust = 1))+
  geom_text(aes(x=x,y=freq,label=paste(round(tools_freq$freq,3)*100,'%',sep = "")),vjust=-
0.2,size=3.5)+
  scale_y_continuous(labels = scales::percent) +
  guides(fill=FALSE)
```

（9）不同公司对数据科学家职位工具型技能的要求

目的：分析不同公司对数据科学家需掌握的工具型技能有哪些共同的要求或有不同的偏向。

方法：选取本数据集当中发布职位数排名前十的公司作为研究对象，分析这 10 家公司对数据科学家工具型技能的要求有何异同：

```
# 分析不同公司对数据科学家职位工具型技能的要求
# 工具型技能词汇集合
tools <- c("SQL","R","PYTHON",
           "EXCEL","SAS","SPSS","HIVE",
           "PPT","HADOOP","MYSQL",
           "SPARK","ORACLE","BI",
           "KPI","JAVA")
# 发布职位数前 10 的公司
company_10 <- c("Amazon.com","Ball Aerospace","Microsoft","Google",
                "NYU Langone Health","Fred Hutchinson Cancer Research Center",
                "KPMG","Lab126","Broad Institute","Facebook")
company_df <- usdata_scientist_new[which(usdata_scientist_new$company %in% company_10),]
seg_company <- company_df %>%
  unnest_tokens(word,description) # 分词
head(seg_company)
# 转换为大写
seg_company$word <- toupper(seg_company$word)
# 只保留工具型技能的词
tools_company <- seg_company[which(seg_company$word %in% tools),]
# 转换为因子类型
tools_company$word <- factor(tools_company$word)
head(tools_company)
# 求解频数并按 company 分面可视化
ggplot(tools_company,aes(x=word,fill=word)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90,hjust = 1)) +
  facet_wrap(~company) + #分面
```

guides(fill=FALSE)

2 研究结果

通过运行 R 语言代码，对领英社交平台上美国数据科学家工作职位的数据进行分析和挖掘，得到该职位相关的统计计量、地理分布、需求分析和技能要求等信息。

（1）总体描述统计信息查看

summary()函数可以获取数据集的描述性统计量，提供最大值、最小值、四分位数和数值型变量的均值，以及因子向量和逻辑型向量的频数统计。

通过以下 summary()函数的运行结果可以看到美国数据科学家工作职位数据集的总体描述统计信息。如 company 变量为因子型，在所有的数据科学家职位中，亚马逊（Amazon.com）发布的数量最多，为 358 条，然后依次是 Ball Aerospace、Microsoft、Google、NYU Langone Health 和 Fred Hutchinson Cancer Research Center；reviews 变量为数值型，所以从结果中可以看到评论数最少为 2 条，最多达到了 148114 条，并且该变量包含了 1627 条缺失值，这与通过 aggr()函数识别的缺失值比例（0.23）相符合。

```
> summary(usdata_scientist_new)
                     position
 data scientist           : 355
 senior data scientist    :  98
 research analyst         :  66
 data engineer            :  63
 machine learning engineer:  56
 research scientist       :  32
 (Other)                  :6283
                                company     description
 Amazon.com                        : 358   Length:6953
 Ball Aerospace                    : 187   Class :character
 Microsoft                         : 137   Mode  :character
 Google                            : 134
 NYU Langone Health                :  77
 Fred Hutchinson Cancer Research Center:  70
 (Other)                           :5990
    reviews                  location                city
 Min.    :     2   Seattle, WA       : 563   New York      : 848
 1st Qu.:    27   New York, NY      : 508   Seattle       : 777
 Median :   230   Cambridge, MA     : 487   Cambridge     : 694
 Mean   :  3179   Boston, MA        : 454   Boston        : 629
 3rd Qu.:  1578   San Francisco, CA: 425   San Francisco: 564
 Max.   :148114   San Diego, CA     : 294   Chicago       : 471
 NA's   :1627     (Other)           :4222   (Other)       :2970
     state
 CA      :2152
 MA      :1323
 WA      : 935
 NY      : 926
 IL      : 471
 DC      : 340
 (Other): 806
```

图 4 数据集总体描述统计信息

（2）美国各个州拥有的职位数情况

以州为工作地点的分析单位，采用条形图对该数据集中美国各个州拥有的职位数量进行呈现，可以看出加利福尼亚州（CA）对数据科学家职位的需求最大，达到 2152 个。其次是马萨诸塞州（MA），为 1323 个，但仅为加利福尼亚州的一半左右。再者就是华盛顿州（WA）和纽约州（NY），分别是 935 个和 926 个，其余包含该职位的 6 个州的需求量均小于 500 个，另外还有 40 个州未呈现对数据科学家岗位的需求。以上反映了美国各个州对于数据科学家职位的需求量

仅集中于几个州，并且这几个州的需求量差异较大。

加利福尼亚是美国经济最发达的州之一，洛杉矶、旧金山、硅谷等大城市都位于该州，并且硅谷作为世界著名的高科技产业区，拥有谷歌、Facebook、苹果等一大批大大小小的高新技术公司，对数据科学家职位的需求是必不可少的。另外，马萨诸萨州的波士顿、华盛顿州的西雅图和纽约州的纽约市等，都是美国金融、商业、教育较为发达的地区，依托于这些经济发达的城市，对数据科学家职位的需求量也得到了保障。通过上述的分析，认为本报告数据集所反映的职位需求分布较为可靠。

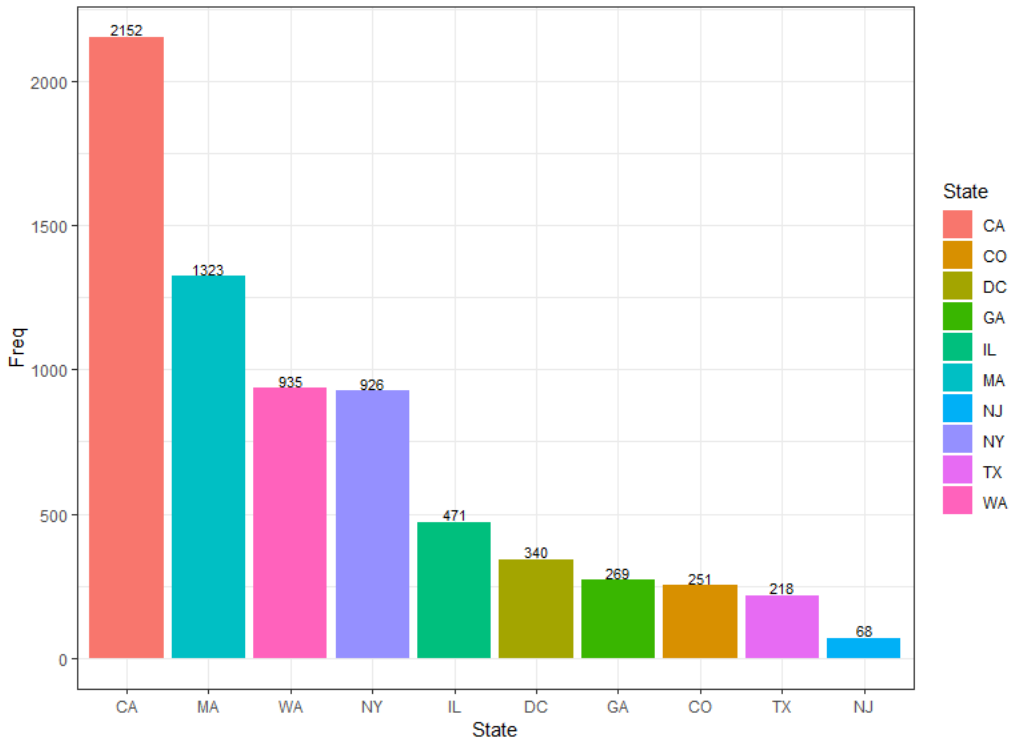

图 5 美国每个州对数据科学家职位需求情况

（3）职位需求在美国各个州的地理分布情况

将包含数据科学家职位的州在美国地图上进行标识，可以看出该岗位需求在地理上的分布以及集中情况。地图上点的位置分布反映了对数据科学家职位的需求主要是在美国沿海的地区。这些地区通常也是经济发达、高新技术产业集中的地区，对数据科学家职位有丰富的需求。

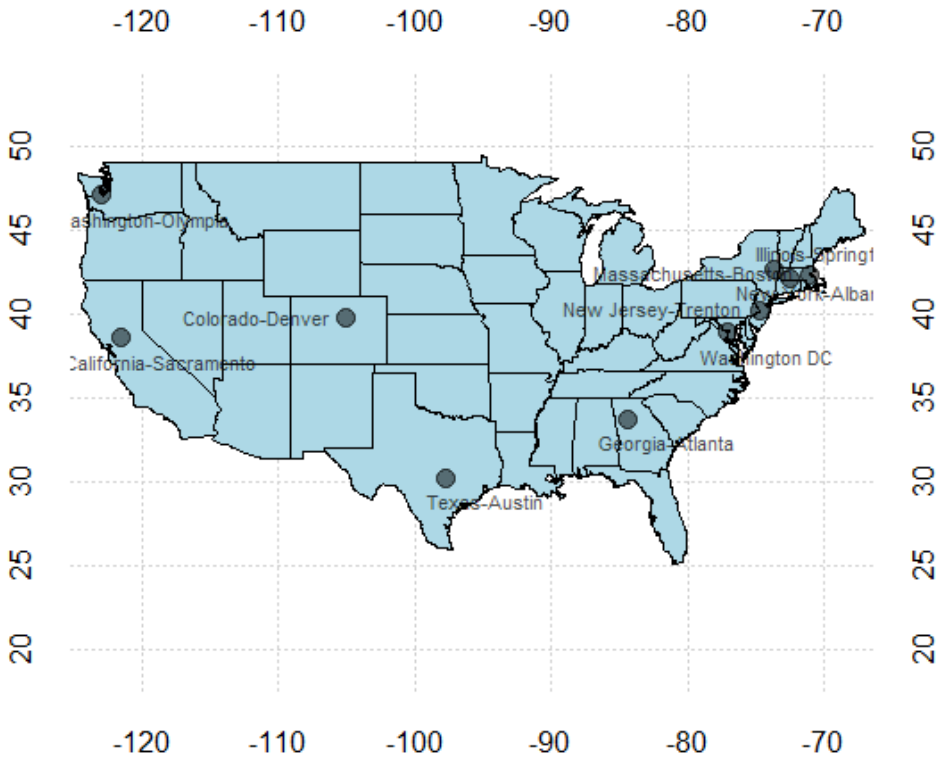

图 6 数据科学家职位需求的地理分布情况

（4）每个州中各城市拥有的职位数情况

根据上述（2）和（3）的分析结果，对职位需求量排名前 4 的州（即 CA、MA、WA、NY）作进一步的细化研究，分析其城市职位需求量的统计情况。

①加利福尼亚州（CA）各大城市数据科学家岗位的分布情况

加利福尼亚州对数据科学家职位的需求较为广泛，分布于各个城市，其中以 San Francisco（旧金山）、San Diego（圣地亚哥）、Mountain View（山景城）、Sunnyvale（森尼维尔）和 Los Angeles（洛杉矶）为主，其余城市的需求量均小于 100。另外，从以下城市分布可以看到，硅谷并未包含其中，即本报告数据集未涉及对硅谷数据科学家职位需求的内容。

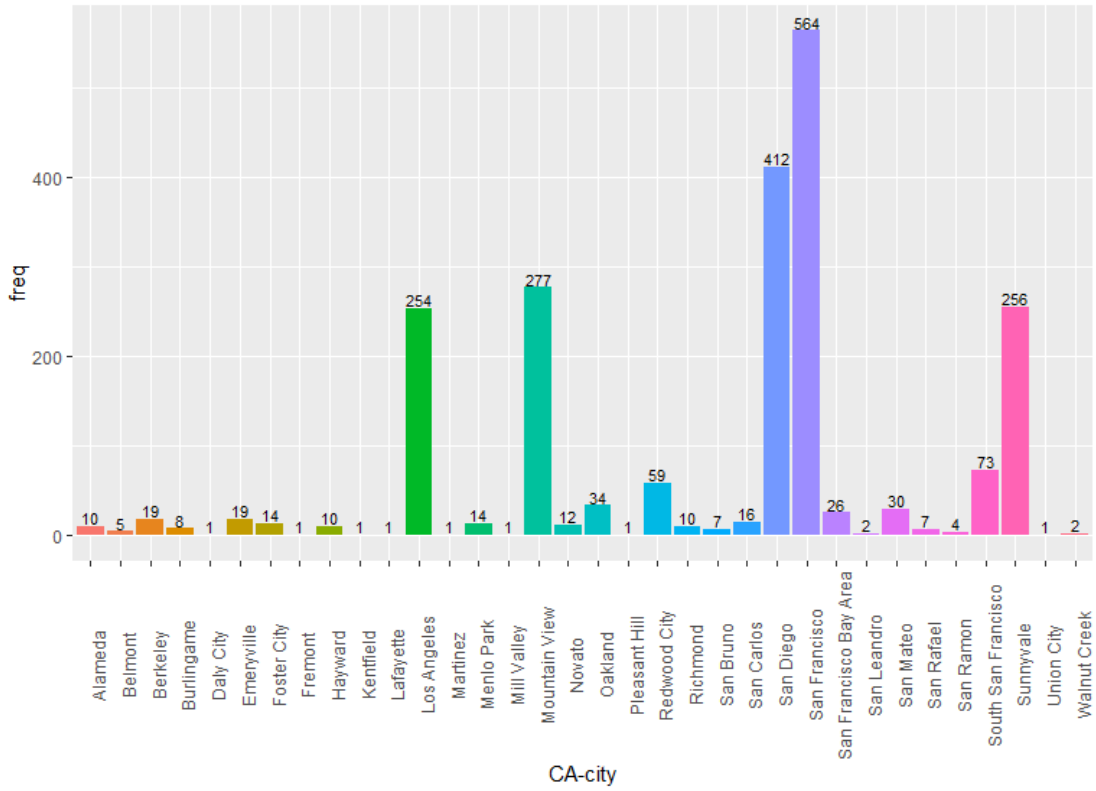

图 7 加利福尼亚州各城市对职位的需求情况

②马萨诸塞州（MA）各大城市数据科学家岗位的分布情况

马萨诸塞州的职位需求主要集中于 Boston（波斯顿）和 Cambridge（剑桥镇），并且这两个城市的需求相近。

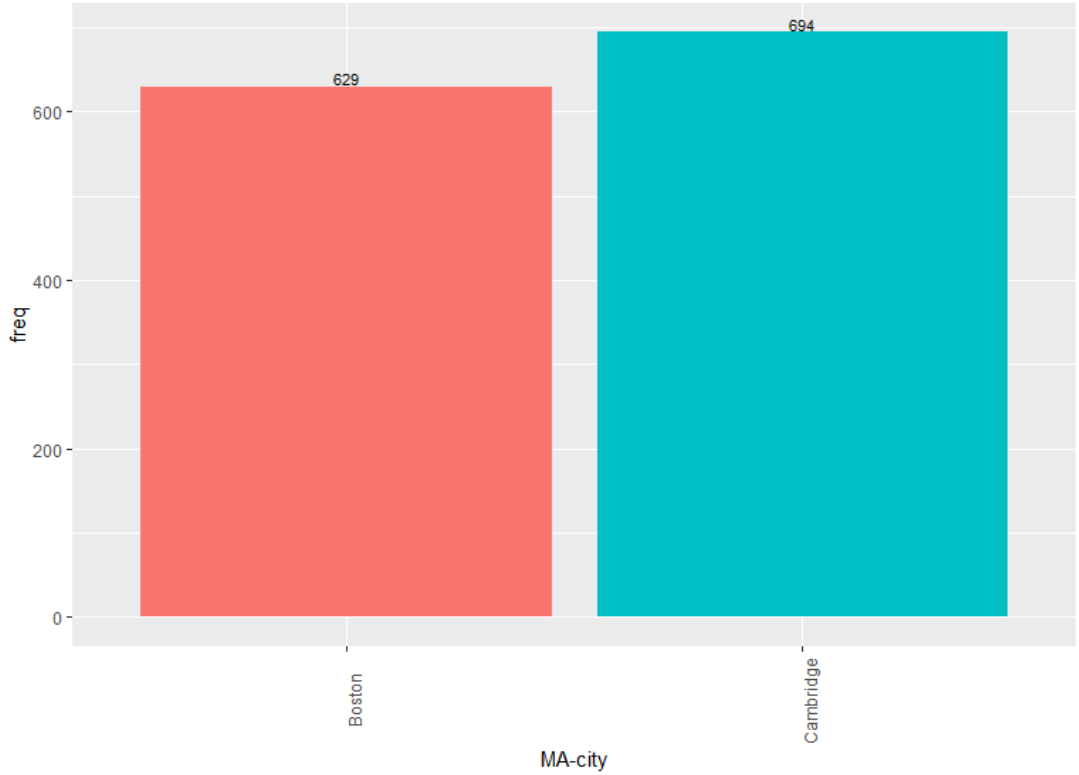

图 8 马萨诸塞州各城市对职位的需求情况

③华盛顿州（WA）各大城市数据科学家岗位的分布情况

华盛顿州的职位需求分布于 Redmond（雷德蒙德）和 Seattle（西雅图）两个城市，以 Seattle 为主，为 777 个，而 Redmond 较少，为 158 个。

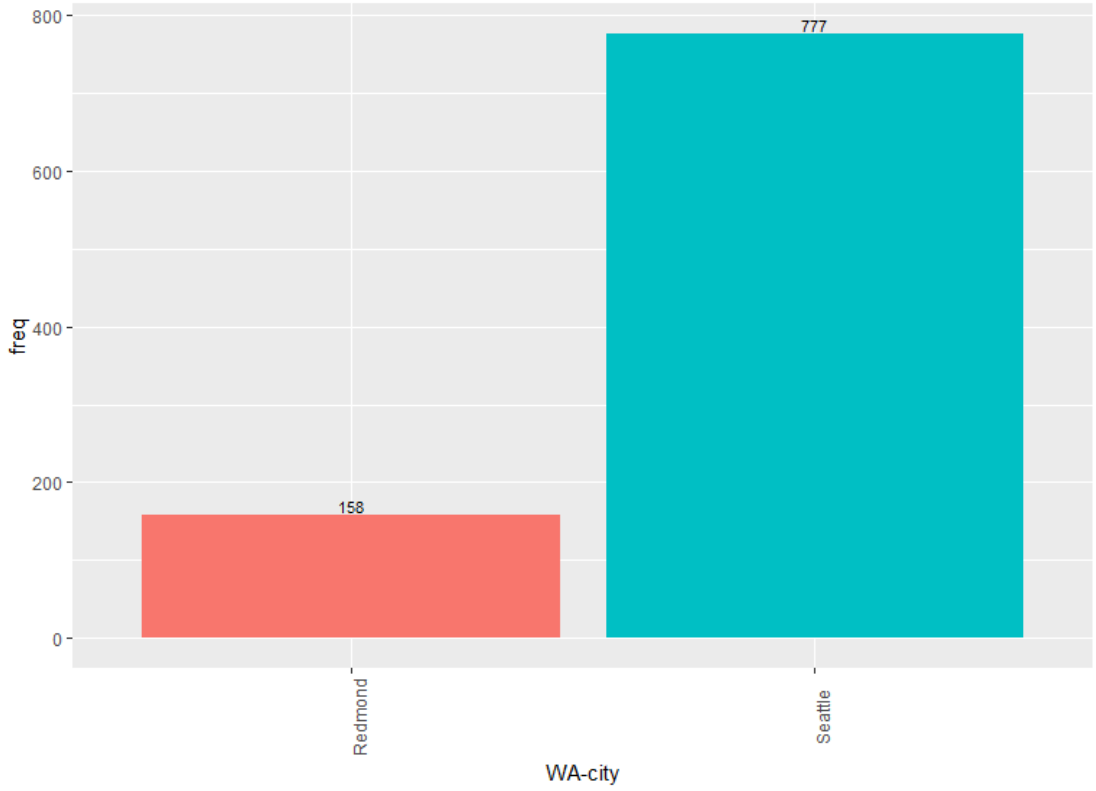

图 9 华盛顿州各城市对职位的需求情况

④纽约州（NY）各大城市数据科学家岗位的分布情况

纽约州的职位需求分布特征较为明显，以纽约市的需求量最为突出，而其他城市的需求量较少。

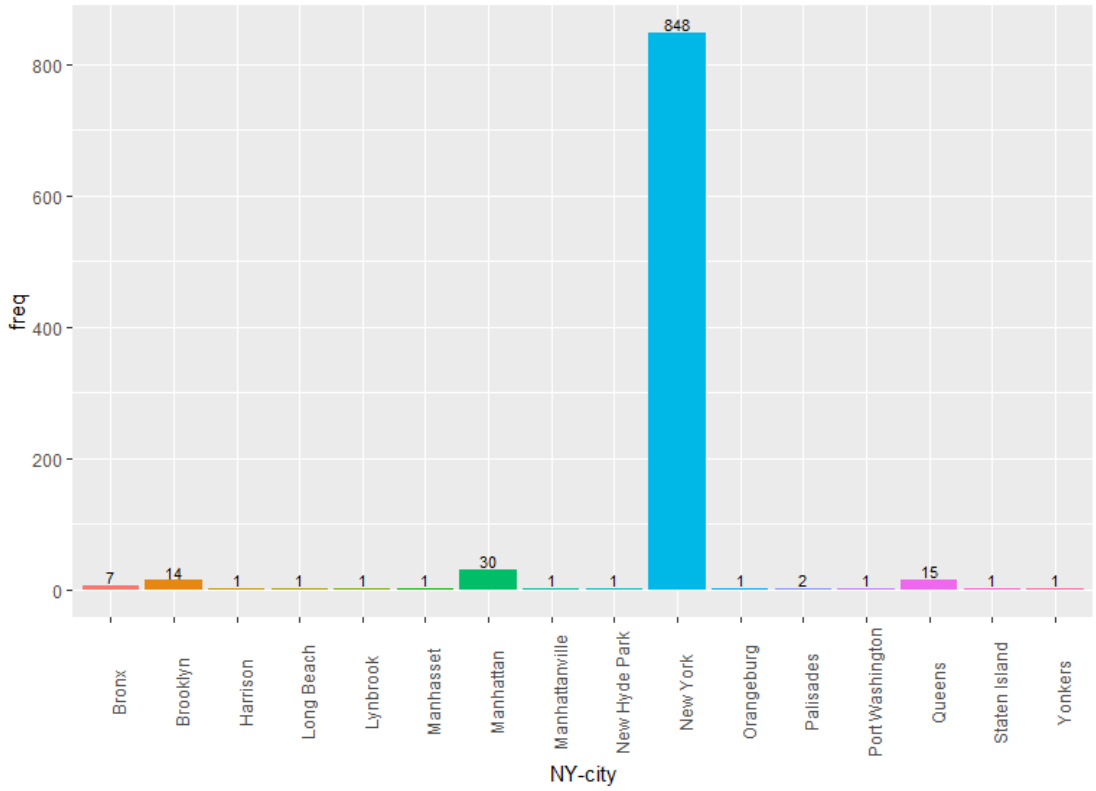

图 10 纽约州各城市对职位的需求情况

（5）公司对数据科学家职位的需求情况

统计各个公司发布的数据科学家职位数量，并选取前 15 家进行图形呈现，可以了解公司对该职位需求的分布情况以及需求程度情况。从以下条形图得到 Amazon、Ball Aerospace、Microsoft 和 Google 对数据科学家职位的需求较大，且以 Amazon 最为突出，而其他公司最职位的需求量稍低且较为相近。

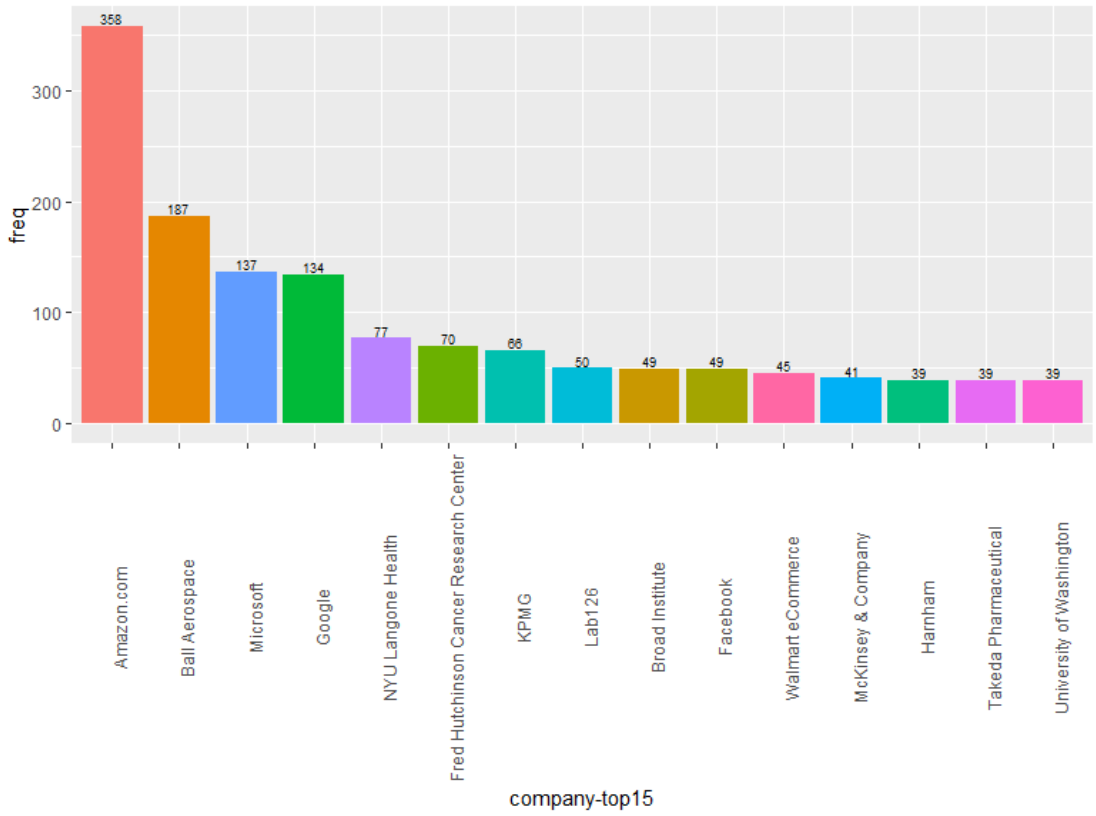

图 11 对数据科学家需求排名前 15 的公司及其需求情况

（6）美国各个州中数据科学家职位的热度情况

通过数据集所给的对每个职位的评论数来表示该职位的热度。将职位按州进行划分，通过箱型图描述不同州中数据科学家职位的热度情况。

由以下箱型图可以看出，加利福尼亚州（CA）和得克萨斯州（TX）存在极大的异常点，通过查看原始数据集，发现这些异常点是来自于对同一公司 Walmart 不同地区分公司发布的职位的评价，说明该公司的数据科学家职位热度值高，较受欢迎。而总体的评论数则较为集中在低于 10000 条的区间内。因此，为了更加清楚的描绘出职位评价数的分布情况，将箱型图纵轴的显示精度修改为（0,10000），得到以下第二张箱型图。

由第二张箱型图可以看出，新泽西州（NJ）的总体评论数均值最高，且新泽西州和华盛顿州的评论数分布较平稳，不存在评论数异常高的点。而其他地区的职位评论数分布总体差异较大，存在多个数值异常高的点。另外，佐治亚州（GA）和得克萨斯州（TX）的职位热度值平均值也处于较前位置。

职位评论数的多少也间接反映了该职位的热度以及求职人员或在职人员对职位的关注程度。通过以上对各个州中数据科学家职位的评论数的分析，并将其与第（2）步中的结果进行比较分析，发现对数据科学家职位需求量较大的州与该职位热度值较高的州并不完全对应，认为新泽西州、佐治亚州和得克萨斯州未来有可能成为数据科学工作人员较为关注的去向和集中的地区。

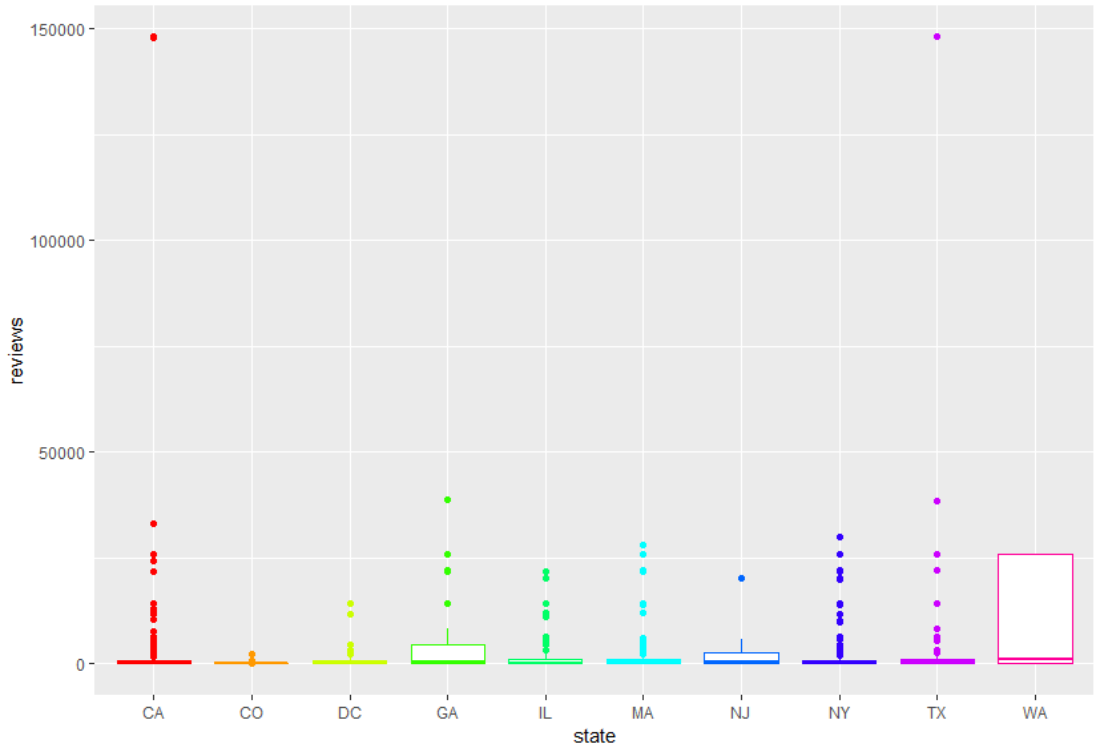

图 12 美国每个州中数据科学家职位热度的总体情况

1	Staff Software Enginee	Walmart	148114	Sunnyvale, CA 94086
2	Information Security F	Walmart	148114	Sunnyvale, CA 94086
3	Senior Data Scientist	Walmart	148114	Sunnyvale, CA 94086
4	Data Scientist-ISD	Walmart	148085	Austin, TX 78716
5	Staff Data Scientist	Walmart	148051	San Bruno, CA 94066

图 13 Walmart 公司的评论数

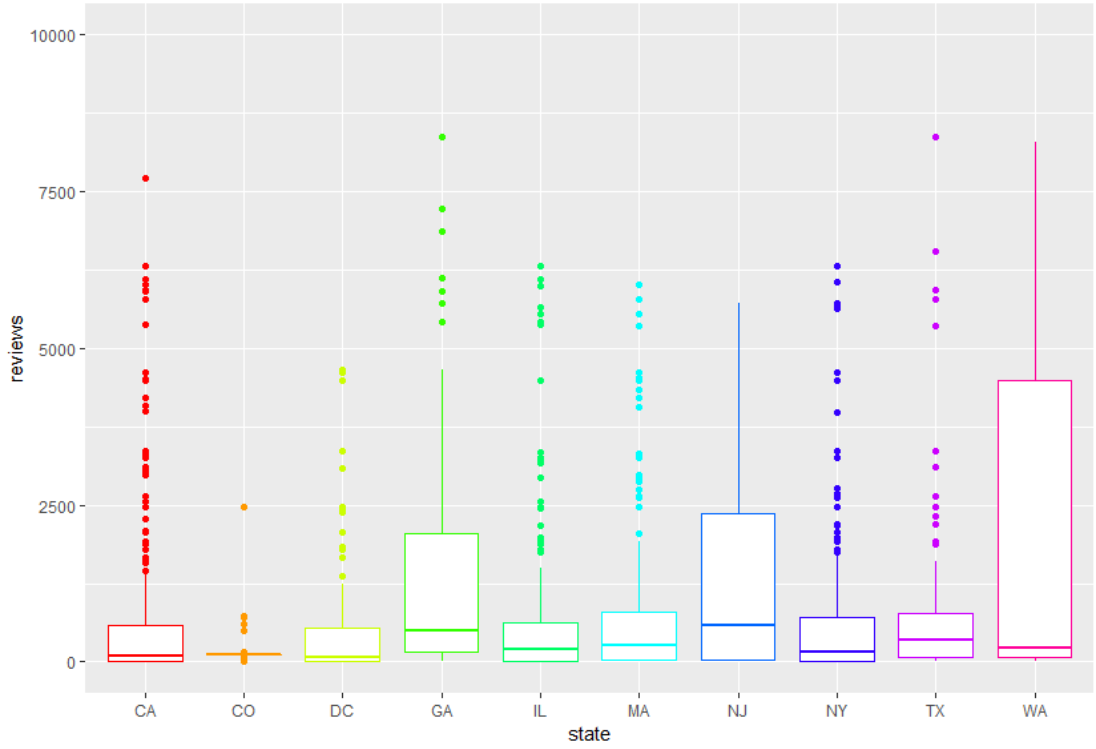

图 14 调整 y 轴精度后的职位热度总体情况

（7）数据科学家职位描述关键词

通过对职位描述内容 description 列进行分词、去停用词、计算词频等文本处理，可以对数据科学家这一职位的总体工作内容，关键职位特征以及主要岗位要求等信息进行表征。

①高频词

抽取出现频次大于 5000 的词作为高频词，以条形图的形式进行有序呈现。

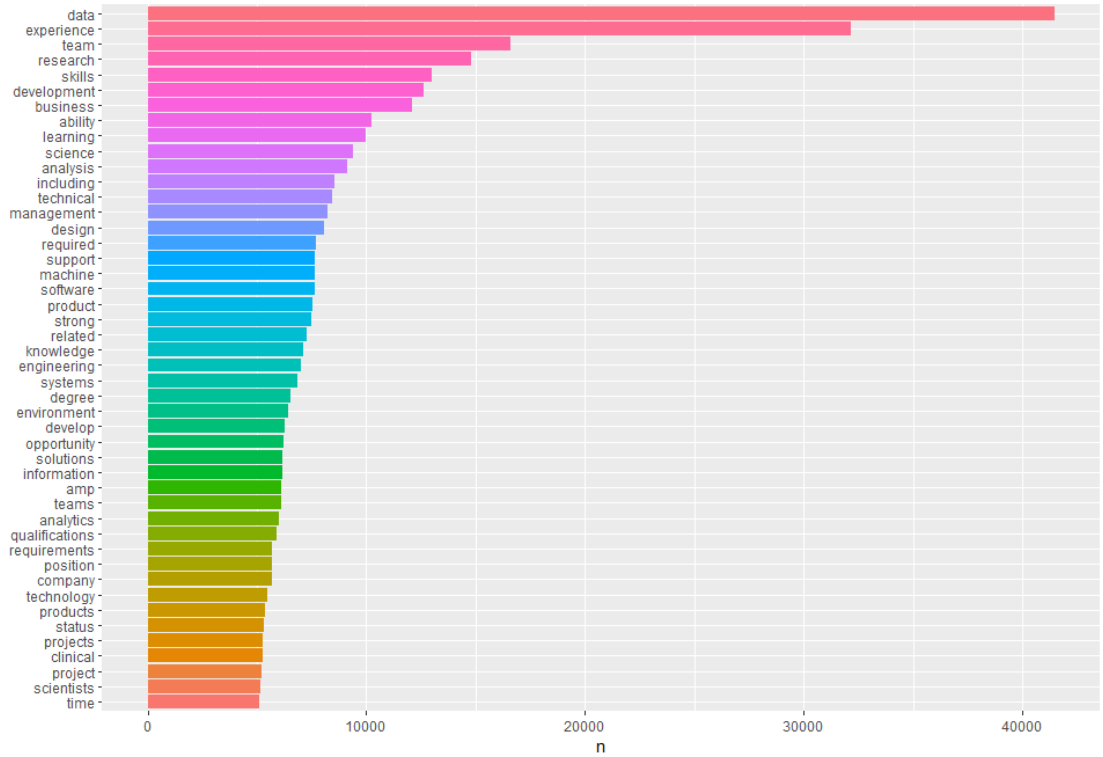

图 15 频次大于 5000 的词频统计图

②词云

另一种对词汇的展示形式为词云图。在词云图中，词汇的显示大小表示该词汇词频的高低。通过词云图可以直观的看出词汇的重要程度。为避免词汇显示过于密集，选择词频大于 3000 的词汇进行词云图的绘制。

在下图中，data 出现频次最高，由于本报告分析的是数据科学家职位的描述，所以不管是工作内容或工作要求，都离不开对数据的描述；

第二个关键词 experience 表明了数据科学家职位对于求职人员工作经验的要求十分看重。具有丰富工作经验的人在求职过程中都受到更多的青睐，而在本次分析中，experience 一词作为仅次于 data 出现的高频词，可以看出工作经验是数据科学家求职中的一个关键要素。在开展工作的过程中，工作经验丰富的数据科学家往往能够更好地对复杂的数据进行处理和解读，更加高效地完成对数据的挖掘和利用。

数据科学家职位也重视团队（team）的协作与管理。在具体开展工作的过程中，每一项任务都需要团队人员之间的相互合作，而并非一个特定职位的人员就能够完成。

同时，数据科学家职位还注重技能（skills）的掌握和能力（ability）的培养。数据分析和挖掘的技术和工具都在不断改进和更新，如机器学习（machine learning）等，而这些新技术、新工具也在逐步应用于企业的工作中。数据科学家需要通过不断学习与研究，来提升自身的能力，更好的适应于职位的要求。

图 16 词频大于 3000 的词云图

（8）数据科学家职位对工具型技能的要求

对职位描述 description 中表示工具型技能的词汇进行统计分析，可以得到这些技能在数据科学家求职中的重要程度。

从对技能词汇的统计排序以及所占比例中可以得到：

R 和 Python 是两个最热门的开源数据分析工具，分别有 51.7%和 41.8%的公司对其提出了要求，因此核心掌握两门语言的其中一门都会让数据科学家具备有力的竞争优势；

SQL 和 Excel 的需求也较高，这两种是数据分析过程中经常用到的基础技能。大多数企业都有自己的数据库体系或系统平台，因此企业工作人员读取和处理数据还是以数据库和 Excel 为主；

在数据库的使用过程中，往往需要结合 Java 语言进行操作，因此对 Java 语言的掌握也对求职有不少帮助；

数据科学家还需掌握 Spark、Hadoop 和 Hive 等主流大数据工具，SAS 等数据分析软件，以及 BI 可视化分析工具。

该分析结果为数据科学家职位的求职人员对自身技能的培养和能力的提升提供了一定的参考价值。根据所分析的结果，对自身所缺乏的技能进行及时的学习，能够提高自己的竞争优势。

```
> arrange(tools_freq, desc(freq))
         x freq
1   PYTHON 3592
2        R 2906
3      SQL 2235
4     JAVA 1488
5    SPARK 1183
6    EXCEL 1142
7   HADOOP 1105
8      SAS 1040
9       BI  557
10    HIVE  547
11 ORACLE  259
12    SPSS  223
13   MYSQL  177
14     KPI   31
15     PPT   12
```

图 17 工具型技能词汇词频统计

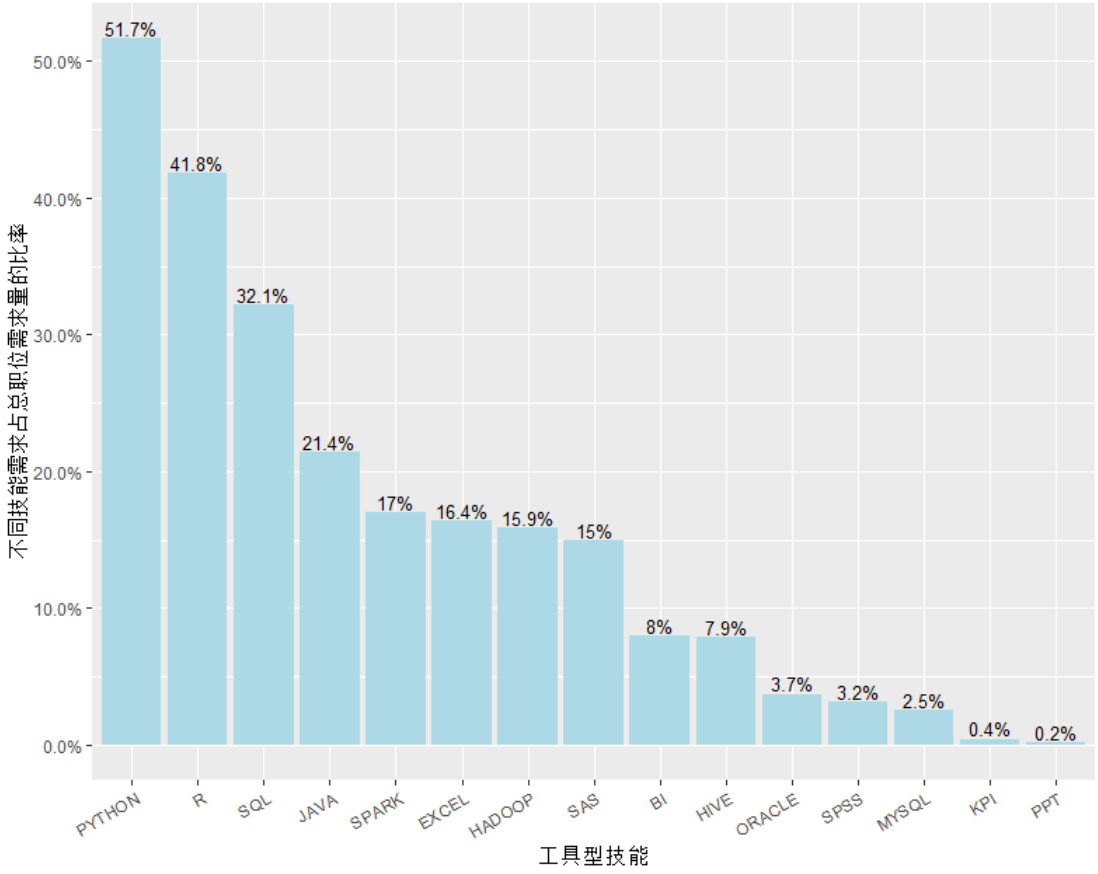

图 18 不同技能需求占总职位需求量的比率

（9）不同公司对数据科学家职位工具型技能的要求

分析不同公司对工具型技能的要求有无异同以及有何异同，可以为求职人员提供有针对性的建议。

选取对数据科学家职位需求排名前 10 的公司进行分析，从以下条形图显示的结果可以得出：

Amazon 公司对工具型技能的要求最高，需要掌握的技能较多，尤其是 Python、Java、SQL 和 R；

Ball Aerospace 公司对 BI 可视化工具的需求尤为突出，所以对于想进入该公司从事数据科学

家工作的求职者来说，掌握 BI 这一工具的使用显然是重中之重；

通过对比这 10 家公司对工具型技能的需求，可以发现，除了 Python 和 R 语言这两种技能是 10 家公司都要求掌握的，还有 SQL 也是所有公司都提出需求的，可见 SQL 这一技能的重要性；

在大部分情况下，公司对 Python 语言的需求要多于对 R 语言的需求。

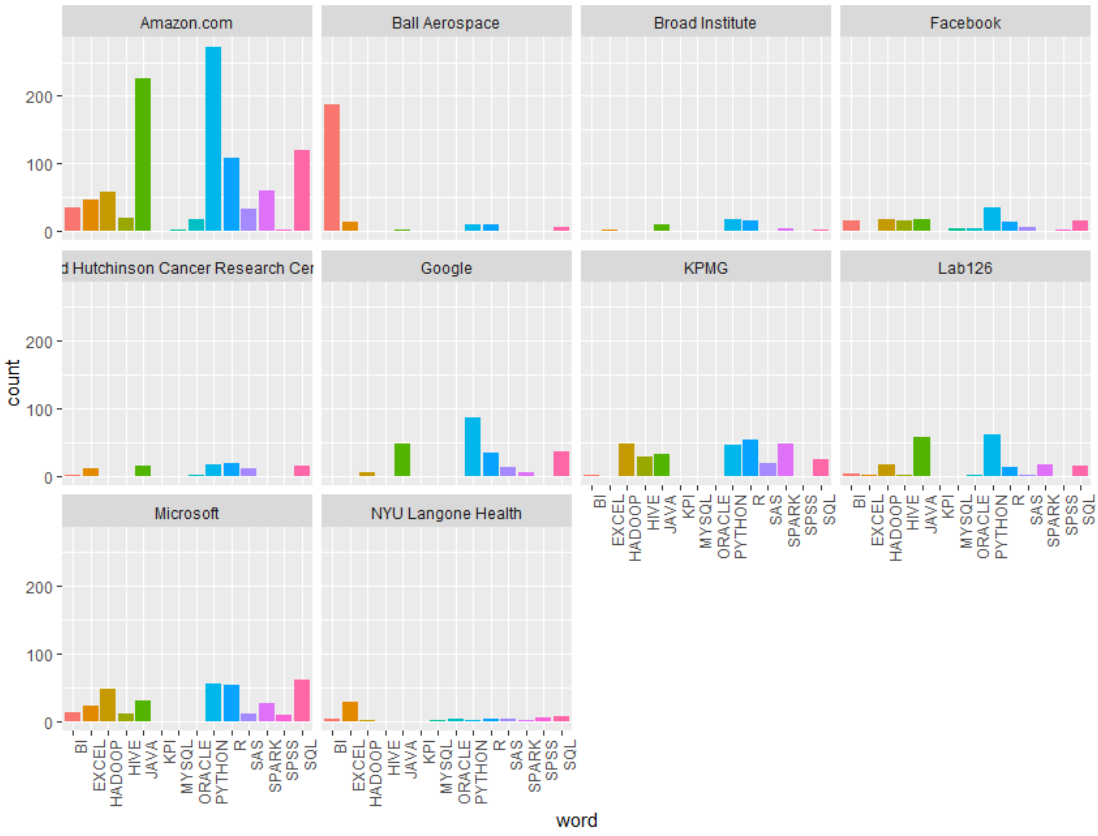

图 19 职位需求量前 10 的公司对技能的要求

3 研究结论

本报告采用 R 语言对美国数据科学家职位数据集进行分析，通过职位所在公司、职位描述、职位评论、工作地点等字段的处理，对数据科学家职位的地区需求、公司需求、职位热度、技能要求等方面的内容进行了不同角度的分析，得出以下结论：

（1）从州的角度来看，加利福尼亚州、马萨诸塞州、华盛顿州和纽约州是对数据科学家职位需求最多的四个州，且以加利福尼亚州最为突出；在地理位置的分布上，对职位的需求主要以沿海的州为主；

（2）从城市的角度来看，旧金山、圣地亚哥、山景城、森尼维尔和洛杉矶这五个城市的需求构成了加利福尼亚州职位需求的主体部分；马萨诸塞州的波斯顿和剑桥镇的需求相当；华盛顿州以西雅图的需求为主；纽约州则是纽约市的职位需求最大；该结果对求职人员工作地点的选择有一定的指导作用；

（3）从公司的角度来看，亚马逊对数据科学家职位的需求最大，Ball Aerospace 次之，接着是微软公司和谷歌公司，而其他公司对数据科学家的需求较低；

（4）从受关注程度来看，新泽西州、佐治亚州和得克萨斯州这三个州虽然不是发布数据科学家职位最多的州，但其发布的职位收到的评论数平均水平高于其他州，具有较高的热度值，受到了求职人员的广泛关注，这也为求职人员在选择职业发展地区时提供了参考。

（5）从总体要求来看，企业更加需要具备多年工作经验，并且业务能力突出，重视团队合

作的分析人才。对于数据科学家个人，应该更加关注自身项目经验的积累，以及新兴技能的学习与关键能力的培养，与时俱进的规划职业的发展路径；

（6）从能力的角度来看，数据科学家需要掌握 R、Python 和 SQL 这三个必备的工具（R 和 Python 可以选择其中一项作为主要工具），若是大数据挖掘方向的工作人员，则需要更加关注 Hadoop、Hive 和 Spark 等工具的使用；

（7）不同公司对数据科学家需要掌握的技能要求有所不同，求职人员应该以此为基础，有针对性地进行技能的训练和职业的规划。

附录：完整源代码

```
# 美国数据科学家工作职位分析(usdata_scientist)

# 读取数据
usdata_scientist <- read.csv("E:/mine/R/usdata_scientist858/usdata_scientist.csv")
# 查看缺失值
library(VIM)
aggr(usdata_scientist,prop=T,numbers=T)
# 分隔列，存储为 city 和 state，并保留原有列
library(tidyverse)
usdata_scientist_new <- separate(data = usdata_scientist, col = location, into = c("city", "state"), sep =
", ", remove = FALSE)
# 去掉 state 列的邮编数字
usdata_scientist_new$state <- gsub('[0-9]*|\\s','',usdata_scientist_new$state)
# 转换为因子类型
usdata_scientist_new$position <- factor(tolower(usdata_scientist_new$position))
usdata_scientist_new$state <- factor(usdata_scientist_new$state)
usdata_scientist_new$city <- factor(usdata_scientist_new$city)
# 转换 description 为字符型，用于文本挖掘
usdata_scientist_new$description <- as.character(usdata_scientist_new$description)
str(usdata_scientist_new)

# 数据分析
# 1 观察数据的总体描述统计信息
summary(usdata_scientist_new)

# 2 各个州拥有的职位数情况
# 对 state 计数
library(plyr)
state_freq <- count(usdata_scientist_new$state)
state_freq
library(ggplot2)
# 重命名列名
names(state_freq) <- c("State","Freq")
library(ggthemes)
ggplot(state_freq,aes(x = reorder(State,-Freq), y= Freq, fill = State))+
  geom_bar(stat = "identity")+
  geom_text(aes(label=Freq),vjust=-0.2,size=3)+ # 添加数值标签
  xlab("State")+ # 修改 x 轴坐标文本
  theme_bw() #添加主题

# 3 职位在各个州的地理分布情况
library(maps)
```

```
library(mapdata)
par(mar=rep(0,4))
# 经纬度数据
dat = read.csv(text = "state-capital,jd,wd
                       California-Sacramento,-121.4943996,38.5815719
                       Colorado-Denver,-104.990251,39.7392358
                       Washington DC,-77.0368707,38.9071923
                       Georgia-Atlanta,-84.3879824,33.7489954
                       Illinois-Springfield,-72.5898110,42.1014831
                       Massachusetts-Boston,-71.0588801,42.3600825
                       New Jersey-Trenton,-74.7597170,40.2205824
                       New York-Albany,-73.7562317,42.6525793
                       Texas-Austin,-97.7430608,30.2671530
                       Washington-Olympia,-122.9006951,47.0378741")
# 地图
map("state", col = "lightblue", fill = TRUE, ylim = c(18, 54), panel.first = grid())
# 地图上显示的点
points(dat$jd, dat$wd, pch = 19, cex = 1.5, col = rgb(0,0, 0, 0.5))
# points(dat$jd, dat$wd, pch = 19, cex = 1.5, col =
c("blue","red","green","blue","red","blue","blue","red","red","green"))
# 地图上显示的文本；pos：文本显示位置（1,2,3,4-下，左，上，右）
text(dat$jd, dat$wd, dat[, 1], cex = 0.7, col = rgb(0,0, 0, 0.7),
     pos = c(1,2,1,1,3,2,2,1,1,1))
# 四周的坐标轴
axis(1, lwd = 0); axis(2, lwd = 0); axis(3, lwd = 0); axis(4, lwd = 0)

# 4 CA、MA、NY、WA 中各城市的职位数
# 4.1 CA 州各城市的职位数
# 选取 CA 的子集
CA_df <- usdata_scientist_new[which(usdata_scientist_new$state=="CA"),]
CA_city_freq <- count(CA_df$city)
CA_city_freq
ggplot(CA_city_freq,aes(x=x,y=freq,fill=x))+
  geom_bar(stat = "identity")+
  geom_text(aes(label=freq),vjust=-0.2,size=3)+ # 添加数值标签
  theme(axis.text.x = element_text(angle = 90))+ # x 轴文本旋转 90 度
  xlab("CA-city")+ # 修改 x 轴坐标文本
  guides(fill=FALSE) # 删除图例

# 4.2 MA 州各城市的职位数
# 选取 MA 的子集
MA_df <- usdata_scientist_new[which(usdata_scientist_new$state=="MA"),]
MA_city_freq <- count(MA_df$city)
MA_city_freq
```

```
ggplot(MA_city_freq,aes(x=x,y=freq,fill=x))+
  geom_bar(stat = "identity")+
  geom_text(aes(label=freq),vjust=-0.2,size=3)+ # 添加数值标签
  theme(axis.text.x = element_text(angle = 90))+ # x 轴文本旋转 90 度
  xlab("MA-city")+ # 修改 x 轴坐标文本
  guides(fill=FALSE) # 删除图例

# 4.3 WA 州各城市的职位数
# 选取 WA 的子集
WA_df <- usdata_scientist_new[which(usdata_scientist_new$state=="WA"),]
WA_city_freq <- count(WA_df$city)
WA_city_freq
ggplot(WA_city_freq,aes(x=x,y=freq,fill=x))+
  geom_bar(stat = "identity")+
  geom_text(aes(label=freq),vjust=-0.2,size=3)+ # 添加数值标签
  theme(axis.text.x = element_text(angle = 90))+ # x 轴文本旋转 90 度
  xlab("WA-city")+ # 修改 x 轴坐标文本
  guides(fill=FALSE) # 删除图例

# 4.4 NY 州各城市的职位数
# 选取 NY 的子集
NY_df <- usdata_scientist_new[which(usdata_scientist_new$state=="NY"),]
NY_city_freq <- count(NY_df$city)
NY_city_freq
ggplot(NY_city_freq,aes(x=x,y=freq,fill=x))+
  geom_bar(stat = "identity")+
  geom_text(aes(label=freq),vjust=-0.2,size=3)+ # 添加数值标签
  theme(axis.text.x = element_text(angle = 90))+ # x 轴文本旋转 90 度
  xlab("NY-city")+ # 修改 x 轴坐标文本
  guides(fill=FALSE) # 删除图例

# 5 分析各公司对数据科学家岗位的需求
# 统计频数
company_pos <- count(usdata_scientist_new$company)
# 各公司发布的职位数降序
library(dplyr)
#排序
company_pos_desc <- arrange(company_pos, desc(freq))
# 取前 15 名
company_pos_top <- company_pos_desc[1:15,]
ggplot(company_pos_top,aes(x=reorder(x,-freq),y=freq,fill=x))+
  geom_bar(stat = "identity")+
  theme(axis.text.x = element_text(angle = 90))+ #x 轴文本旋转 90 度
  geom_text(aes(label=freq),vjust=-0.2,size=2.5)+
```

```
  xlab("company-top15")+
  guides(fill=FALSE)

# 6 各个州的数据科学家热度的总体情况
# 过滤掉 reviews 为 NA 的行
reviews_com <- usdata_scientist_new[complete.cases(usdata_scientist_new[,5]),]
ggplot(reviews_com,aes(x = state,y = reviews))+
  geom_boxplot(colour = rainbow(10))+ # 调色板
  ylim(0,max(reviews_com$reviews)) # 设置 y 轴精度
ggplot(reviews_com,aes(x = state,y = reviews))+
  geom_boxplot(colour = rainbow(10))+ # 调色板
  ylim(0,10000) # 设置 y 轴精度

# 7 数据科学家职位描述的关键词
# 文本挖掘
library(tidytext)
tidy_words <- usdata_scientist_new %>%
  unnest_tokens(word,description) %>% # 分词
  anti_join(stop_words) # 去停用词
head(tidy_words)
word_freq <- tidy_words%>%
  count(word,sort=TRUE) # 计算词频并排序
head(word_freq)

# 7.1 词频统计图
library(ggplot2)
word_freq %>% # 统计词频
  filter(n > 5000) %>%
  mutate(word=reorder(word,n)) %>%
  ggplot(aes(word,n,fill=word))+
  geom_col()+
  xlab(NULL)+
  coord_flip()+
  guides(fill=FALSE)+ # 删除图例

# 7.2 词云图
library(wordcloud2)
wc_freq <- word_freq %>% filter(n > 3000) # 取词频大于 3000 的词画云图
wordcloud2(wc_freq,size=1,shape = 'star')

# 8 数据科学家职位对工具型技能的要求
seg <- usdata_scientist_new %>%
  unnest_tokens(word,description) # 分词
# 转换为大写
```

```
seg$word <- toupper(seg$word)
# 只保留工具型技能的词
tools_df <- seg[which(seg$word %in% c("SQL","R","PYTHON",
                                      "EXCEL","SAS","SPSS","HIVE",
                                      "PPT","HADOOP","MYSQL",
                                      "SPARK","ORACLE","BI",
                                      "KPI","JAVA")),]
# 转换为因子类型
tools_df$word <- factor(tools_df$word)
# 求解频数
tools_freq <- count(tools_df$word)
# 降序输出
library(dplyr)
arrange(tools_freq, desc(freq))
# 计算占比
tools_freq$freq <- tools_freq$freq/nrow(usdata_scientist_new)
ggplot(tools_freq) +
  geom_bar(aes(x=reorder(x,-freq),y=freq),fill="lightblue",stat = "identity") +
  labs(x="工具型技能",y="不同技能需求占总职位需求量的比率") +
  theme(axis.text.x = element_text(angle = 30,hjust = 1))+
  geom_text(aes(x=x,y=freq,label=paste(round(tools_freq$freq,3)*100,'%',sep = '')),vjust=-
0.2,size=3.5)+
  scale_y_continuous(labels = scales::percent) +
  guides(fill=FALSE)

# 9 分析不同公司对数据科学家职位工具型技能的要求
# 工具型技能词汇集合
tools <- c("SQL","R","PYTHON",
           "EXCEL","SAS","SPSS","HIVE",
           "PPT","HADOOP","MYSQL",
           "SPARK","ORACLE","BI",
           "KPI","JAVA")
# 发布职位数前 10 的公司
company_10 <- c("Amazon.com","Ball Aerospace","Microsoft","Google",
                "NYU Langone Health","Fred Hutchinson Cancer Research Center",
                "KPMG","Lab126","Broad Institute","Facebook")
company_df <- usdata_scientist_new[which(usdata_scientist_new$company %in% company_10),]
seg_company <- company_df %>%
  unnest_tokens(word,description) # 分词
# 转换为大写
seg_company$word <- toupper(seg_company$word)
# 只保留工具型技能的词
tools_company <- seg_company[which(seg_company$word %in% tools),]
# 转换为因子类型
```

```
tools_company$word <- factor(tools_company$word)
head(tools_company)
#  求解频数并按 company 分面可视化
ggplot(tools_company,aes(x=word,fill=word)) +
   geom_bar() +
   theme(axis.text.x = element_text(angle = 90,hjust = 1)) +
   facet_wrap(~company) + #分面
   guides(fill=FALSE)
```