统计制图课堂作业

中央财经大学
王思雨

一、“大数据里的情人节”案例可视化分析

曾经在情人节时在“今日头条”app 中看到过这么一个根据搜索引擎的上搜索数据来描述统计情人节时期中国的南北方相关的信息的，可视化分析报告名为《大数据里的情人节》[1]，直观地展现有关情人节方面的相关统计分析，通过根据搜索引擎内搜索词条频率来形象的展现人们在情人节中所关注的事情和事物，并通过色彩鲜明，冲击力强，方法创新的可视化作图方法来展示一些令人好奇而又新颖的问题。本文截取报告的几张有特色的可视化统计视图简单分析一下。

➢ 情人节与七夕两个节日关注度对比

《大数据里的情人节》对中西方两个关于爱情的节日，情人节以及七夕节进行关注度的分析对比，如图所示；

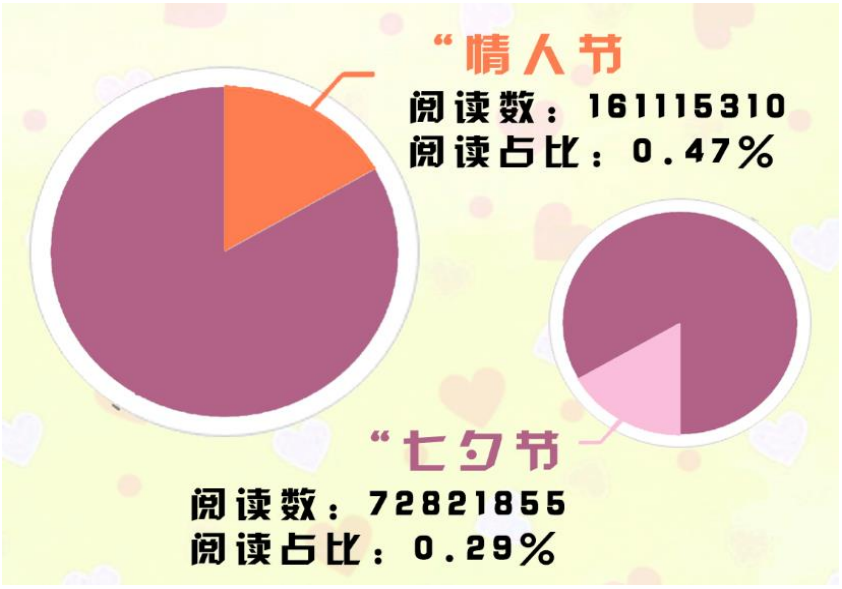

图 1.1 情人节、七夕节浏览器阅读占比

在对比两个节日的关注度的可视化方法中《大数据里的情人节》报告使用的是饼图，使用的变量是“七夕节前两周浏览器中的阅读次数”，“情人节前两周浏览器中的阅读次数”，“全网该时段的总阅读次数”，皆为定量变量，从图中我们可以看出相对于两个节日前两周的总阅读次数来说，情人节的阅读占比较高，说明在中国人们在提到情侣间的节日，更倾向于过情人节而非传统节日七夕节，人们对情人节更为看重和关注。

[1] 报告来源于“今日头条 app”中的“算数中心”栏目

- **情人节人群画像**

《大数据里的情人节》还描述了关注情人节的人群年龄状况，统计描述了在浏览器中阅读关于情人节相关内容的人群年龄段结构，并绘制成统计图如图所示；

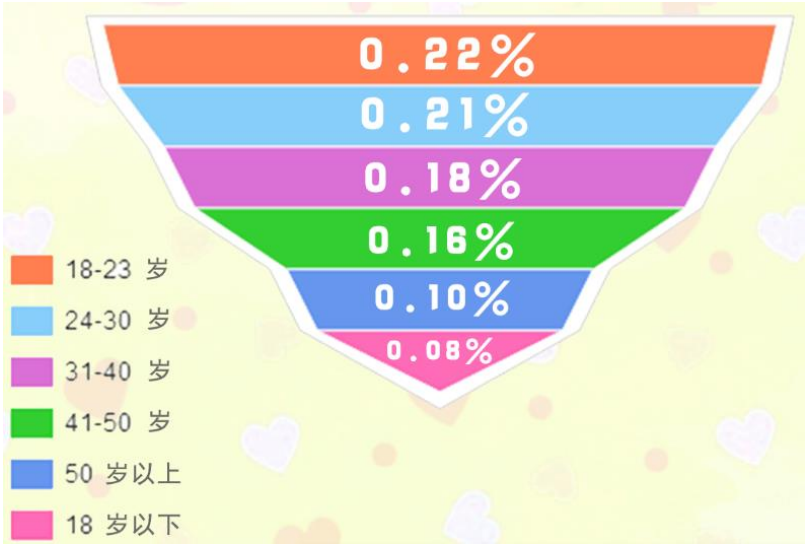

图 1-2 情人节人群年龄结构

在描述情人节关注人群年龄结构分布的可视化方法中《大数据里的情人节》报告使用的是堆栈图，使用的变量是“各年龄段阅读关于情人节信息次数占总阅读次数的比例”，皆为定量变量，从图中我们可以看出关注情人节最多的年龄段是 18-23 岁以及 24-30 岁这两个年龄段，而且成年人随着年龄的增长，对情人节的关注降低。

- **情人节关键词词云图**

《大数据里的情人节》还描述了情人节的关键词统计，统计描述了与情人节相关的各关键词出现的频率，并绘制成统计词云图如图所示；

图 1-3 关键词词云图

在描述情人节关注关键词的可视化方法中《大数据里的情人节》报告使用的是词云图，使用的变量是“各关键词与情人节同时出现并被阅读的次数”，皆为定量变量，从图中我们可以看出在情人节期间，人们关注的最多的事物和事项是“结婚“”，其次与情人节联系紧密的是“电影”，然后就是“吃饭”和“分手”。很符合我们对情人节的认知。

- **不同城市对情人节的关注**

《大数据里的情人节》不仅描述了情人节的关键词统计，还统计描述了与情人节相关关键词在不同的城市中出现的频率，并绘制成统计饼图如图所示：

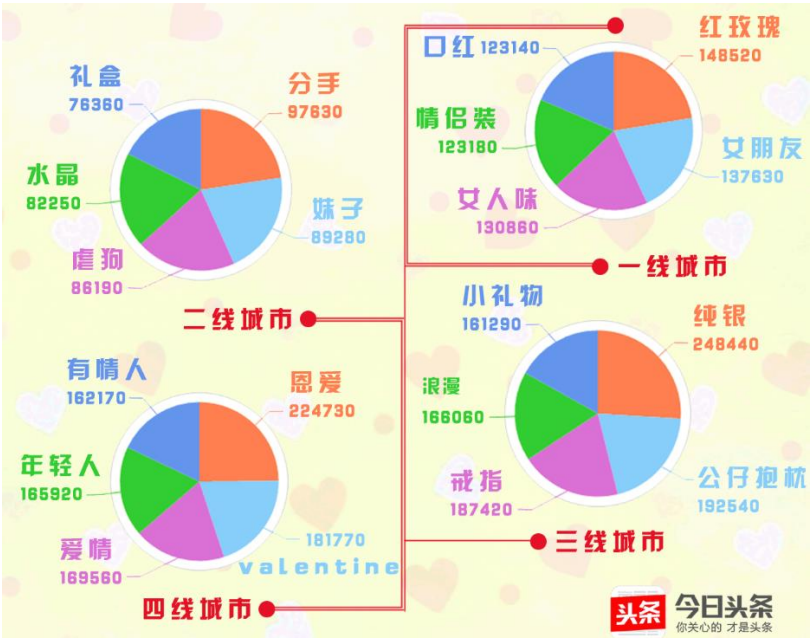

图 1-4 不同城市对情人节的关注

在描述不同城市对情人节关注方面可视化方法中《大数据里的情人节》报告使用的是分类饼图，使用的变量是“不同城市各关键词与情人节同时出现并被阅读的次数”，皆为定量变量，从图中我们可以看出在情人节期间，一线城市人们关注的最多的事物和事项是“红玫瑰”和“女朋友”，二线城市人们关注最多的是“妹子”和“分手”，看来一二线城市的单身的男性应该比较多，很缺爱。三线城市与情人节联系紧密的分别是“纯银”和“公仔抱枕”。而四线城市才是最有爱的，情人节关注的方面是与感情方面联系很紧密的“恩爱”和“爱情”，看来，大城市套路深，还是小城市更有“爱情”啊！

- **关注情人节的南北方差异**

《大数据里的情人节》不还统计描述了南北方对于情人节的关注程度，从一个侧面对比了南北方的浪漫程度，绘制成统计地图如图所示;

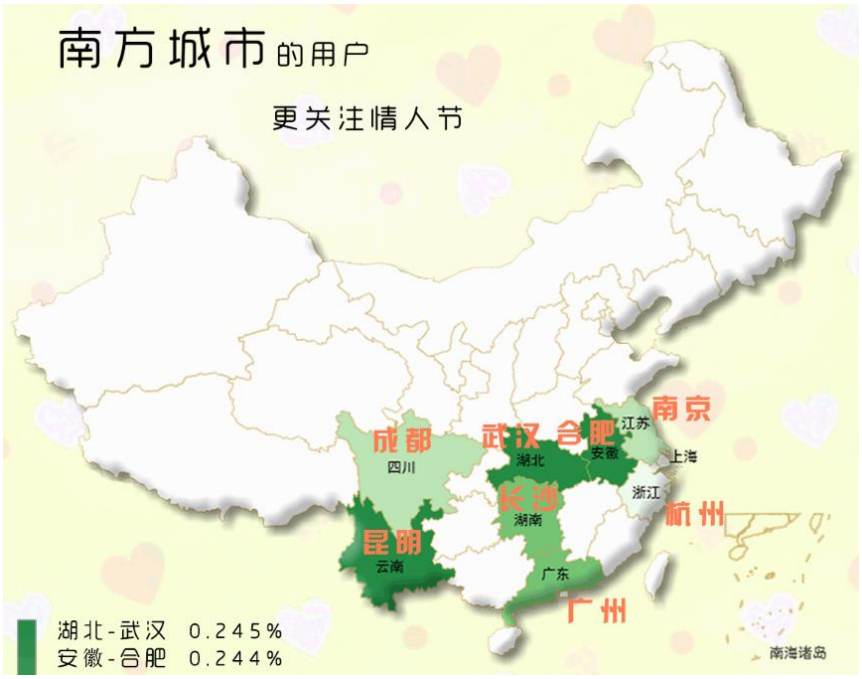

图 1-5 情人节南北方关注程度

在描述南北方对情人节关注方面可视化方法中《大数据里的情人节》报告使用的是统计地图，使用的变量是“关注情人节的各城市用户数占各城市用户总数的比例”，皆为定量变量，从图中我们可以看出在关注情人节方面，南方的省市明显颜色要更加深一些，代表关注的人数比例高，从一个小侧面也可以不严谨的说南方人更加浪漫！

总结：

《大数据里的情人节》给我的印象是一篇极为成功的可视化数据分析报告。首先它能把握数据的使用者，也就是关注情人节数据报告的用户群体必定大多数是年轻人，所以他们所需要的数据必定带有潮流性，《大数据里的情人节》正是把握了这一点所以在进行描述分析时可以加入了年轻化的词汇和时尚简约的配色，引人入胜，可观赏性很强。对于数据的把握上，这篇报告使用了搜索引擎上的一手数据，数据来源可靠。

在图的使用上，《大数据里的情人节》尝试者使用多种统计图，多种变量相会对比的方式，充分利用色块大小，不同的位置表示关系，颜色来直观传达信息，从各个侧面反映了在中国人们在情人节中的行为和表现。非常直观的呈现一些原本不易理解或表达的数据，比如南北方对情人节的关注程度等，改用区域和颜色这种更容易被人理解的方式来呈现，将数据变得更通俗易懂。如果能将数据可视化还要适时适当融入交互性元素，将会更有利于帮助普通用户或商业用户快速理解数据的含义或变化，发现数据的潜在关联。

二、人口学背景单变量描述分析

2.1 变量的选择

对于变量的选择，选取了两个定性变量，和一个定量变量，变量说明如下：

表 2-1 人口学变量说明表

变量类型	变量名	详细说明	取值范围	备注
定量变量	bd006	结束学业年龄 单位：岁	[1-120]	1-120 岁整数值
定性变量	bd001	学历	文盲，专科 本科，硕士等	包含 11 个学历
	bb006	16 岁之前居住地	城市/乡村	

2.2 单变量描述分析

- **数据与处理**

```
setwd("D:/大数据实验/可视化/北大国家发展中心数据/demographic_background/")
mydata=read.dta("demographic_background.dta")
  #添加工作目录，并导入数据 dta 数据，作为 mydata 数据框的值。
 summary(mydata)
  #查看 mydata 数据框
which(is.na(mydata))
  #检查里面有无空值
```

- **绘制柱状图并对比**

简单柱状图：

```
data1<-mydata$bd001
data1<-na.omit(data1)
summary(data1)
 #数据框 mydata 提取 bd001 学历字符型数据向量并赋值给 data1,删除所有 NA 样本。
data1<-table(data1)
#将 data1 转化为 table 型做完所有的柱状图数据准备
label<-c("文盲","能够读、写","私塾","小学毕业","初中毕业","高中毕业","中专","大专","本
        科","硕士","博士")
barplot(data1,main="受访者学历柱状图",xlab = "学历",ylab="人数"
        ,width=0.7,col="lightblue",names=label)
  #画简单柱状图并将图 x 轴的标签换为中文，作图效果如下：
```

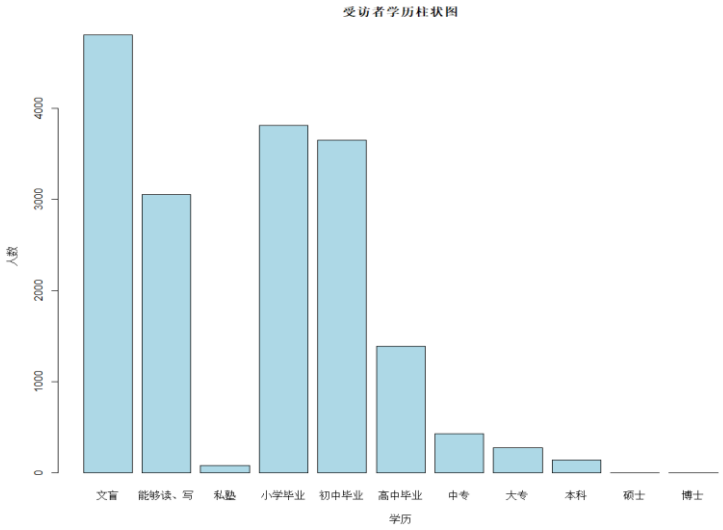

图 2-1　受访者学历分布

ggplot2 柱状图：

```
mydata<-mydata[-which(is.na(mydata$bd001)),]
```

在 mydata 中删除在 bb001 中的为空值的所有行

```
mydata<-mydata[-which(is.na(mydata$bd006)),]
ggplot(data=mydata,aes(x=bd006))+geom_histogram(fill='orchid1',colour = 'black')
                +xlab("年龄") + ylab("频数")
                +ggtitle('读完书年龄直方图')
                 +theme(plot.title = element_text(hjust = 0.5,
                        family="myFont",size=18,color="red"),
                        panel.background=element_rect(fill='skyblue',color='black'))
```

利用 ggplot 开始绘制柱状图将参数调节完毕，绘制出的图形如下：

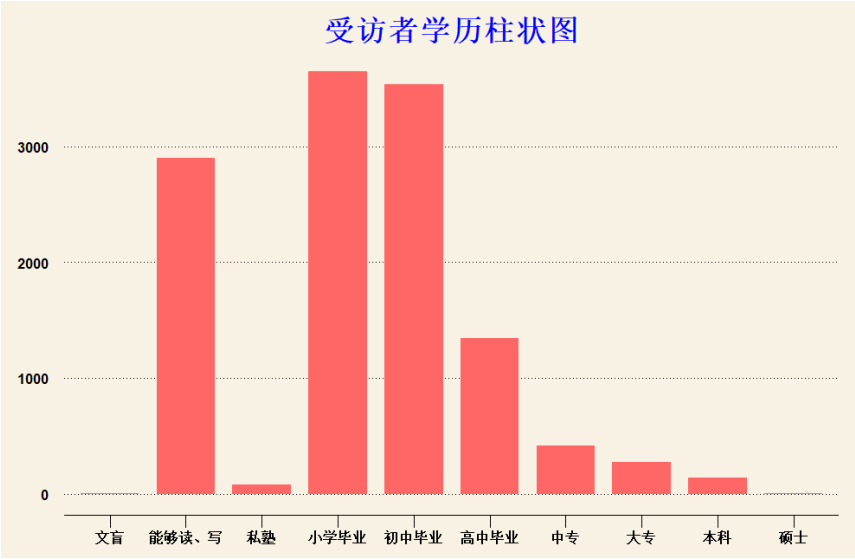

图 2-2 受访者学历分布柱状图

➢ 饼图对比

简单饼图：

```
data<-mydata$bb006
summary(data)
which(is.na(data))
data<-na.omit(data)
```

#数据框 mydata 提取 bb006 城乡字符型数据向量并赋值给 data，删除所有 NA 样本。

```
par(mfrow = c(1, 2))
slices <- c(15759,1912)
lbls <- c("城镇", "乡村")
pie(slices,labels= lbls,col=brewer.pal(5,"Set2"),
    border="white",font=2,labelcex=1,explode=0.1,
    radius=0.95,main = "16 岁之前受访者居住地")
```

#开始绘制简单饼图，以出现的频数为准，并调节参数，使图变美观
#利用 par(mfrow = c(1, 2))进行图层的分割，再绘制百分比饼图

```
pct <- round(slices/sum(slices) * 100)
lbls2 <- paste(lbls, " ", pct, "%", sep = "")
pie(slices,labels=lbls2, col=brewer.pal(5,"Set2"),
    border="white",font=2,labelcex=1,explode=0.1,
    radius=0.95, main = "16 岁之前受访者居住地")
```

#开始绘制简单百分比饼图，以出现占比为准，并调节参数，使图变美观。

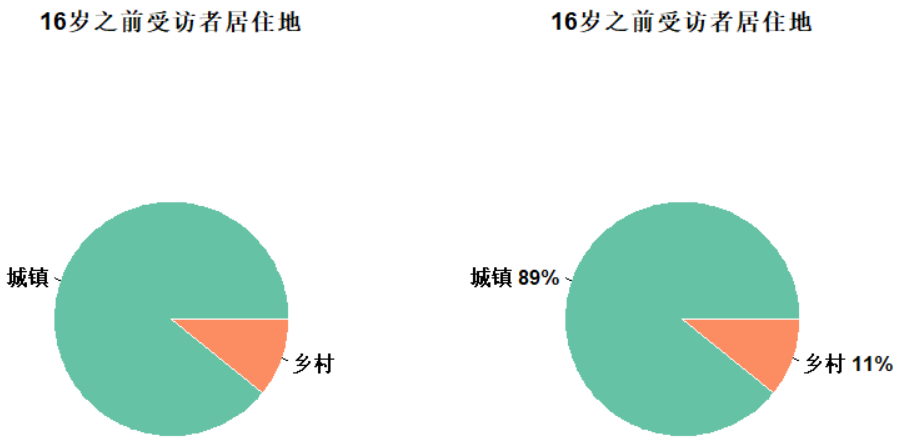

图 2-3　居民身份分布饼图

ggplot2 饼图：

```
mydata<-mydata[-which(is.na(mydata$bb006)),]
```

#在 mydata 中删除在 bb006 中的为空值的所有行

```
ggplot(mydata,aes(x=factor(1),fill=bb006))
        +geom_bar()+coord_polar(theta="y")
        +ggtitle('受访者 16 岁前居住状况')
      +theme(plot.title = element_text(hjust = 0.5,family="myFont",size=18,color="red"),
             panel.background=element_rect(fill='aliceblue',color='black'))
```

#利用 ggplot 开始绘制饼图，无法在 ggplot 直接作图，采用 coord_polar(theta="y")在条形图中进行极坐标变换得出饼图，将参数调节完毕，绘制出的图形如下：

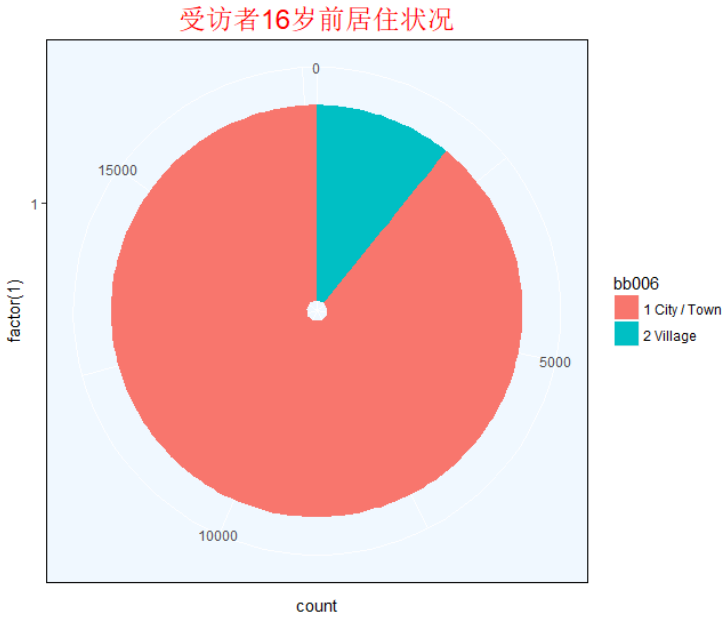

图 2-4　居民身份分布饼图

- **直方图对比**

简单直方图：

```
data2<-mydata$bd006
summary(data2)
which(is.na(data2))
data2<-na.omit(data2)
```

#数据框 mydata 提取 bd006 读书年龄数据向量并赋值给 data 2，删除所有 NA 样本。

```
summary(data2)
hist(data2,xlab="年龄",ylab="频数",main="读完书年龄",col="lightblue")
```

#开始绘制简单直方图，以出现的频数为准，并调节参数，使图变美观，详细。

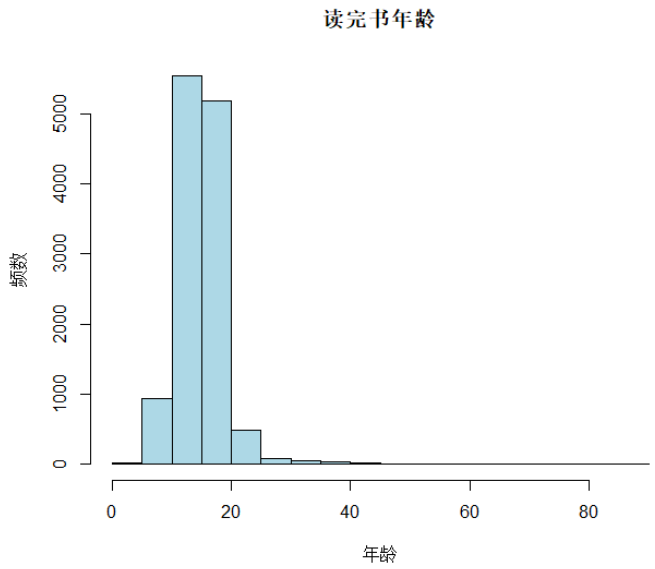

图 2-5 毕业时间直方图

ggplot2 直方图：

```
mydata<-mydata[-which(is.na(mydata$bd006)),]
```

#在 mydata 中删除在 bd006 中的为空值的所有行

```
ggplot(data=mydata,aes(x=bd006))
      +geom_histogram( fill = 'orchid1', colour = 'black')
      +xlab("年龄") + ylab("频数")
      +ggtitle('读完书年龄直方图')
      +theme(plot.title = element_text(hjust = 0.5,
            family="myFont",size=18,color="red"),
            panel.background=element_rect(fill='skyblue',color='black'))
```

#利用 ggplot 开始绘制直方图，将参数调节完毕，绘制出的图形如下：

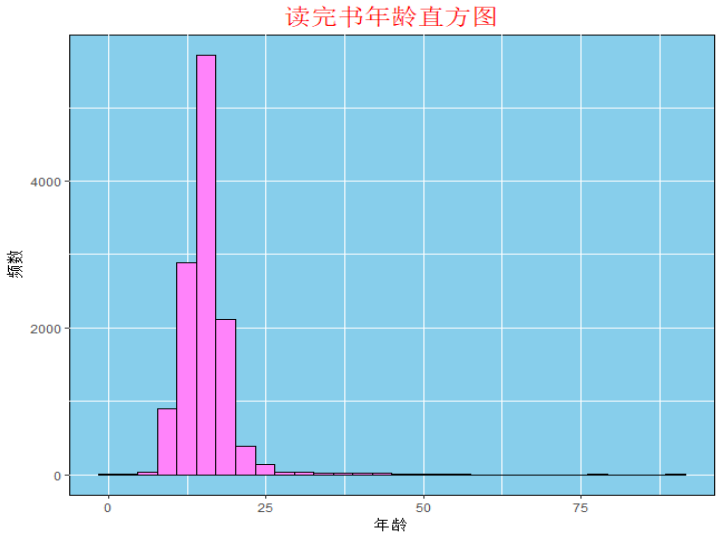

图 2-6 毕业时间直方图

➢ 绘制分面统计图

绘制分面统计图 R 代码如下所示：

```
mydata=read.dta("demographic_background.dta")
mydata<-mydata[-which(is.na(mydata$bb006)),]
p<-ggplot(mydata,aes(x=bd006,fill=bb006))
p+geom_histogram(position="identity",alpha=0.6)+facet_grid(bd001~bb006,scales="free")
```

将数据导入后，将空值列进行删除后，进行 ggplot 分面统计图绘制，利用 bd001 和 bb006 即，居民身份和学历作为分类变量绘制出如下 2*10 的分面统计图。从图中可以清晰的看出某个居民身份的学历对应的受访者人数。

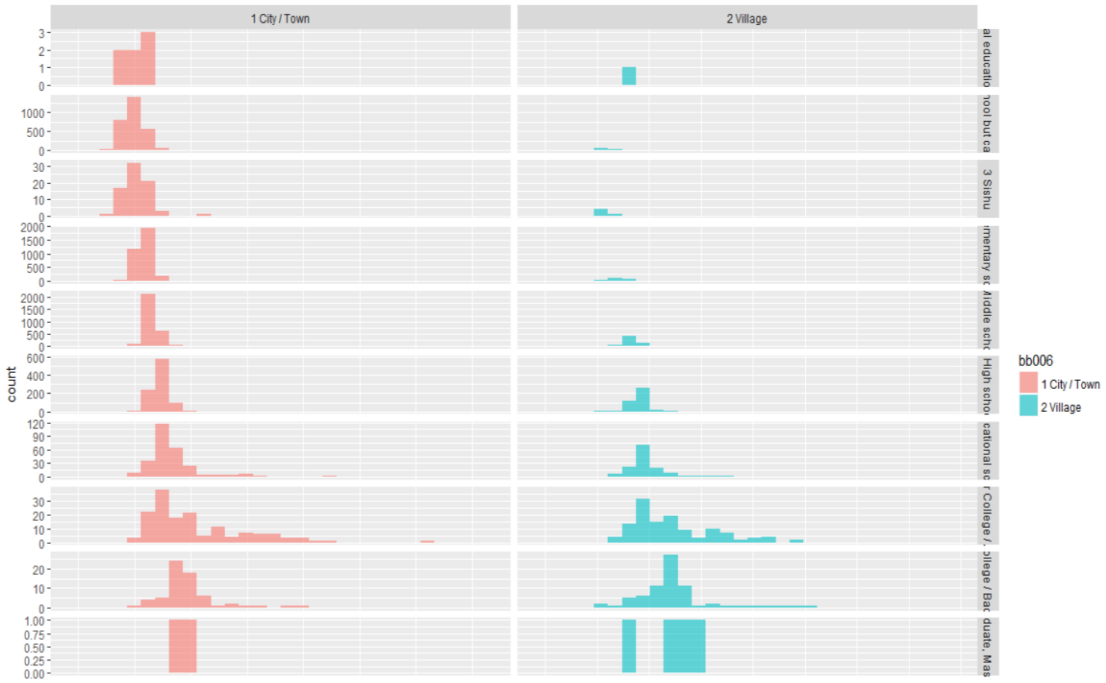

图 2-7 分面统计图

➢ 原始制图与 ggplot 作图区别与联系

首先 ggplot2 作图包是依据 R 原始作图基础上进行编译的基本的图形与 R 原始作图相似。可以说 ggplot 将 plot 的图层思想发扬光大，比如 plot 在增加 lines（）和 legend（）时也用了图层思想，所以说 ggplot 把原本在 plot 里面的参数，摘出来形成图层，更贴合人类思维。

ggplot 基于图层概念，更接近于人作图的思维，使作图更加顺利，流畅，增加容错率。而且基于它图层作图的优势，使做图灵活，变化丰富，而且易于修改和增强。其次 ggplot 提供了众多的模板和配色板，作图更加美观。Ggplot 在每个函数图层上又增加了众多参数，使图像的可调节性更强。但比较遗憾的是 ggplot 并不提供饼图的作图方式，因此想要做饼图需要进行极坐标变化，在堆积柱形图上进行极坐标变换，但这也是得 ggplot 多出了风玫瑰这样的更加美观的图形。

ggplot 与原始作图的区别还在于所应用的数据，原始作图利用的大多是对

向量的操作，利用的数据是数据框抽出来的向量，而 ggplot 的操作对象是数据框，这样使用起来更加方便，加少了对矩阵的操作。

三、人口学背景双变量描述分析

➢ 分类箱线图

想考察受访者的居民身份和学习年龄之间的关系，想要探究城市居民的身份是否会使受教育年龄和教育程度提高，从图 3－1 中可以看出关于城镇和乡村受教育年龄的分布情况，城市的毕业平均年龄为 1 5 岁，而乡村为 1 7 岁左右，发现总体毕业年龄乡村要大于城镇的毕业年龄，说明居民的身份情况还是对毕业年龄有一定影响。

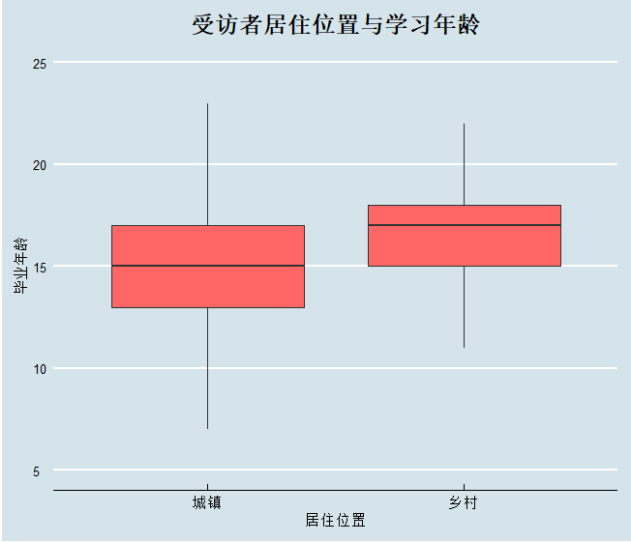

图 3-1　居住位置与学习年龄关系

➢ 堆积柱状图

想考察受访者的居民身份和学习年龄之间的关系，想要探究城市居民的身份是否会使受教育年龄和教育程度提高，还可以从堆积柱状图中找到答案，从图 3－2 堆积柱状图中可以看出关于城镇和乡村受教育年龄的分布情况，发现不论城市还是乡村受教育年龄的分布都是右偏分布，并且两者都是在 1 5－1 7 岁结束教育的人数最多，而乡村人口结束教育的时间相对更加正态分布。

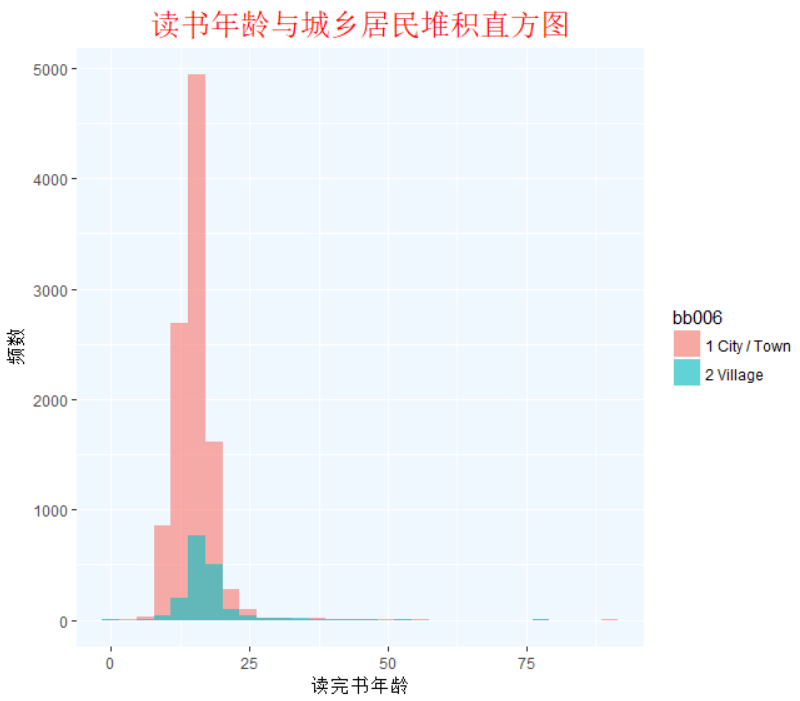

图 3-2 读书年龄与城乡居民身份关系

- **堆积柱状图**

通过堆积柱状图考察绘制居民身份与学历之间两个定性变量之间的关系，观察城市农村人口在各个文化程度之间的分布情况，在图 3－3 中我们可以明显的看到本次的调查对象显然是城市人口较多，城市人口中文盲的比重较大，而在农村人口中可以看到是初中毕业人数较多，占比较大。

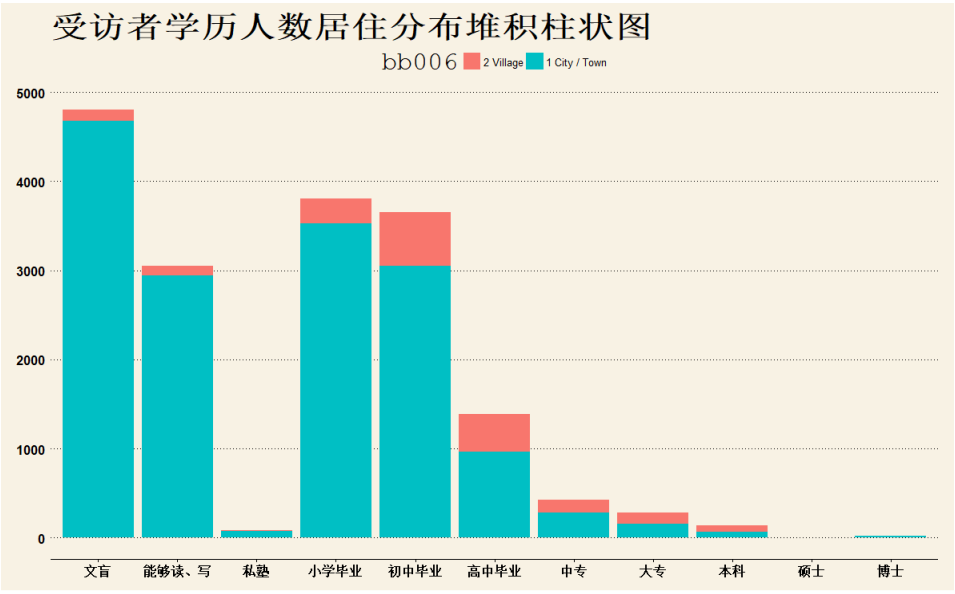

图 3-3 受访者学历与居民身份堆积柱状图

➢ 风玫瑰图

风玫瑰图只是对堆积柱状图进行了极坐标变化，使图形更加美观可视，风玫瑰图如下图所示，分析结果与堆积柱状图相似。在图中可以形象直观的看出每个学历的人数，和每个学历人数在城镇和乡村的占比。

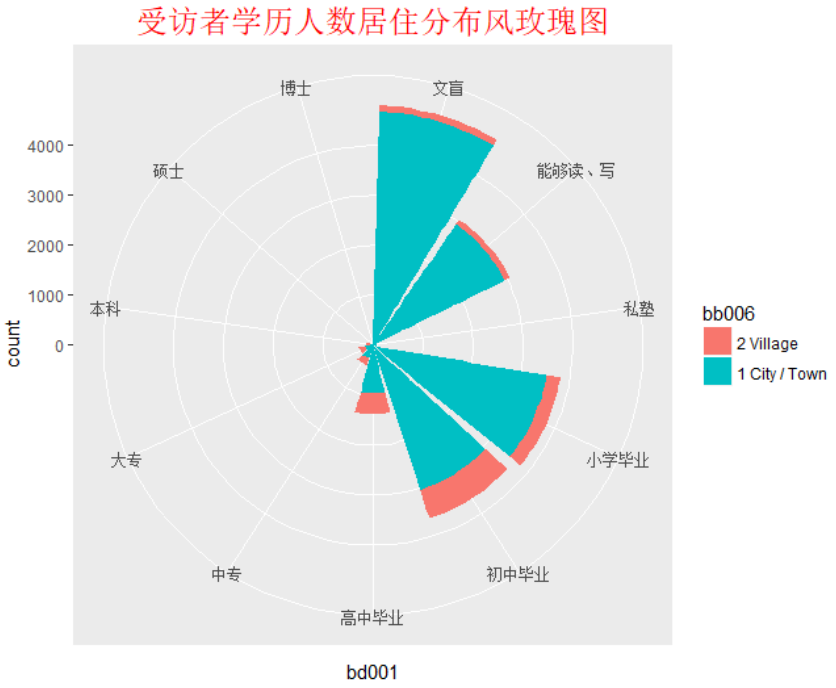

图 3-4　学历人数与城乡居民身份关系