选取2016年9月的数据进行分析。该数据包括纽约公共自行车2016年9月1日至9月20日使用流水的记录。

一、纽约公共自行车使用人群分析

1.性别

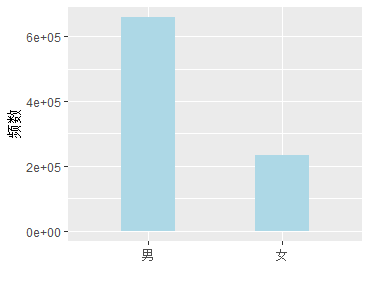


图1 公共自行车使用的性别构成

男性使用占73.77%，女性使用占26.23%。可以看出，男性是纽约公共自行车的主要使用群体。

2.年龄

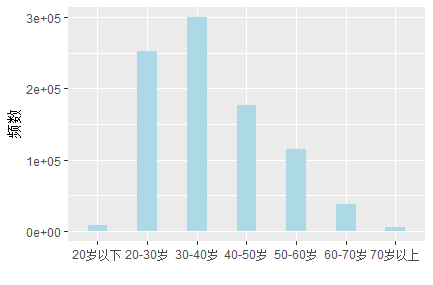


图2 公共自行车使用的年龄构成

20-30岁及30-40岁人群对公共自行车的使用最多，分别占28.29%和33.49%。40-60岁的人群也有较多的使用。20岁以下、70岁以上人群使用公共自行车最少，分别仅占0.95%和0.60%。

3.临时用户与订阅用户

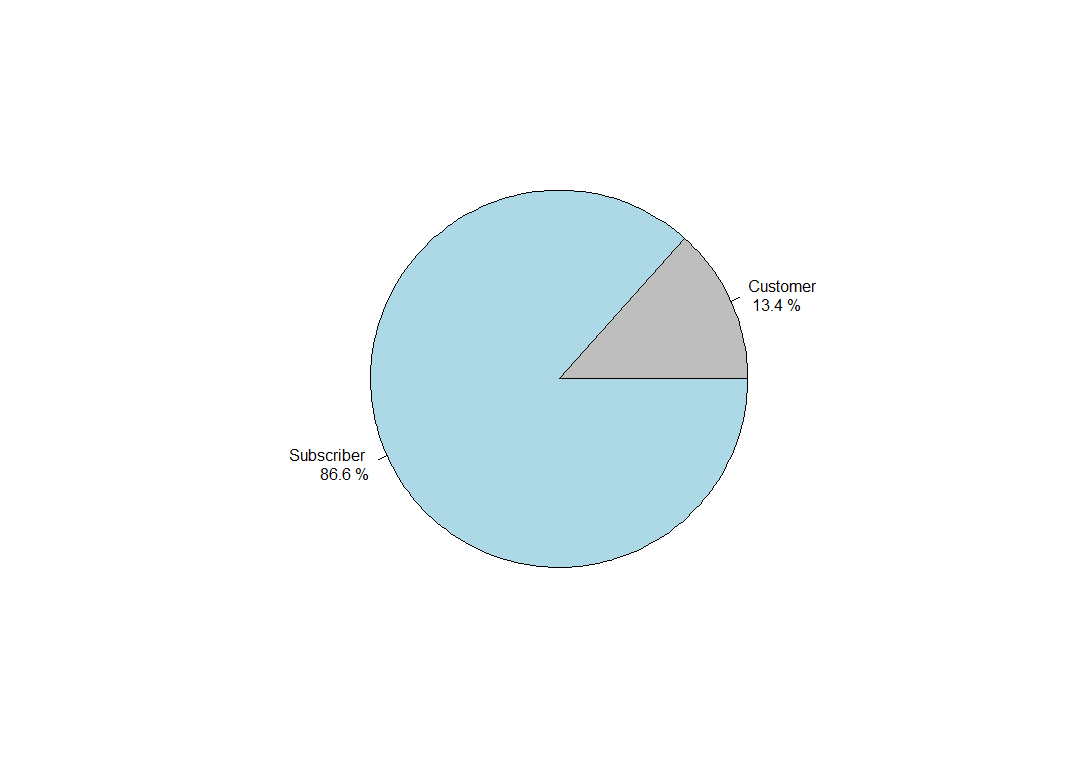


图3 公共自行车使用者的类别构成

纽约公共自行车9月的使用流水中，86.6%的自行车使用是由订阅用户产生，仅有13.4%由临时用户产生。可见订阅用户是纽约公共自行车使用的主力军。

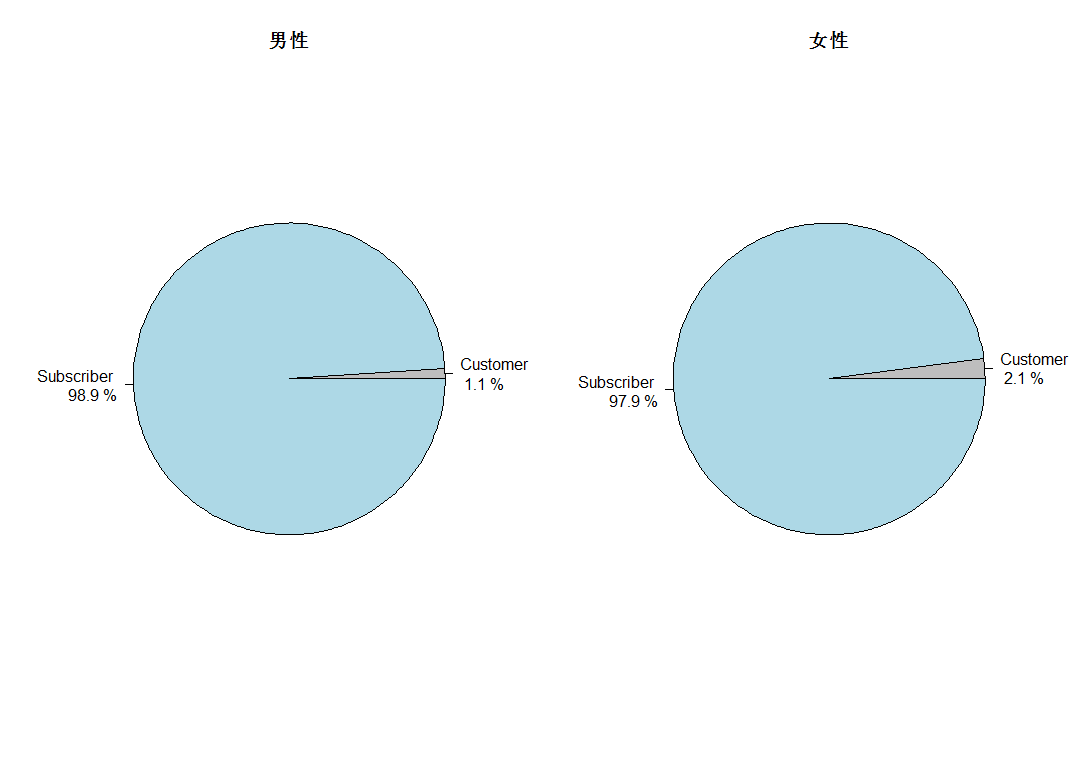


图4 不同性别公共自行车使用者的类别构成

从上图可以看出，女性用户中的“临时用户”比例略高于男性用户中的“临时用户”比例。

二、用户使用情况的描述性分析

1.时长

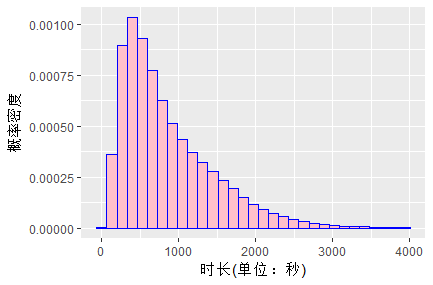


图5 公共自行车使用时长直方图

9月，纽约公共自行车的使用时长平均为11.48分钟，中位数为14.33分钟。可见公共自行车主要用于短时间使用。从下图还可以看出，女性用户借出的平均时长高于男性。

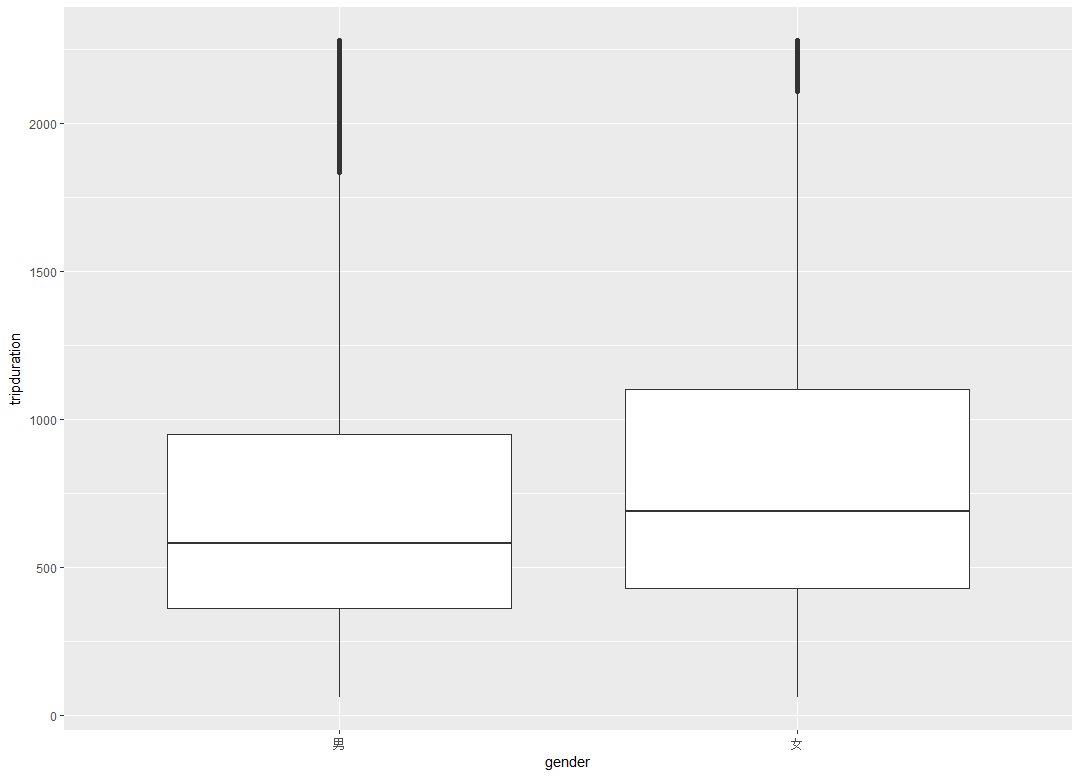


图6 不同性别人群使用时长箱线图

2.星期

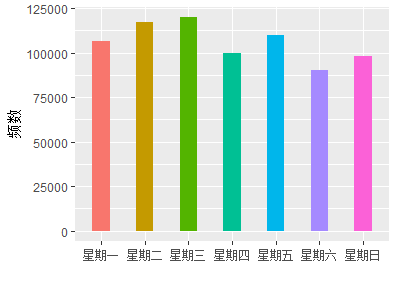


图7 自行车使用的星期分类

选取9月自行车使用流水的前两个完整星期（9月1日-9月14日）进行分析。可以看出，纽约公共自行车在工作日的使用多于在周末的使用。工作日中，周一至周三的使用多于周四、周五。这可能与纽约公共自行车使用者多是上班族有关。

3.出发时间

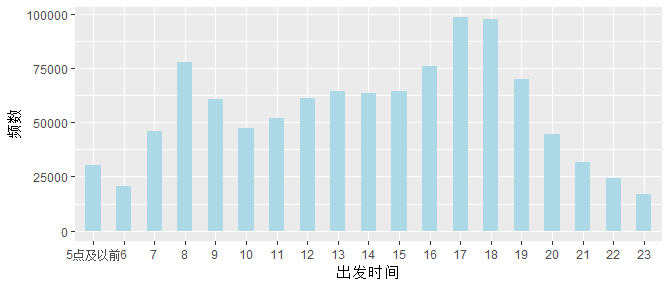


图8 公共自行车使用的出发时间图

纽约公共自行车在一天中的各个时段均有使用，但主要集中在早晚上下班高峰期（早上8、9点和下午4点至6点）。早上6点至8点，自行车使用有明显增加趋势；中午12点至15点，自行车使用率较稳定；经下班高峰期之后，自行车使用率又有明显下降。这再一次印证了纽约公共自行车主要使用人群是为了上班而使用公共自行车。

代码：

（胡练）

setwd("C:/Users/hulian/Desktop/机器学习/吕晓玲")

data=read.csv("201609-citibike-tripdata.csv",header=T,sep=',')

library(ggplot2)

######usertype

#####条形图

ggplot(data,aes(x=usertype))+geom\_bar(width=0.4,fill='lightblue')

#####饼图

usertype.percentage=round(100\*prop.table(table(data$usertype)),digits=1)

mylabs=paste(names(usertype.percentage),"\n",usertype.percentage,"%",sep=" ")

pie(usertype.percentage,labels=mylabs,col=c('grey','lightblue'))

#######tripduration

age=2016-data$birth.year

data1=data.frame(data,age)

summary(data1$tripduration)

data1$tripduration[data1$tripduration>2282]=NA

summary(data1$age)

data1$age[data1$age>71.5]=NA

boxplot(data1$tripduration)

data1$gender[data1$gender%in%c('1')]='男'

data1$gender[data1$gender%in%c('2')]='女'

data1$gender[data1$gender%in%c('0')]=NA

data2=subset(data1,select=c(gender,tripduration,age))

ggplot(na.omit(data2),aes(x=gender,y=tripduration))+geom\_boxplot()

ggplot(na.omit(data2),aes(x=gender,y=age))+geom\_boxplot()

######分性别的usertype

data.male=data1[data1$gender=='男',]

data.female=data1[data1$gender=='女',]

par(mfrow=c(1,2))

usertype.percent1=round(100\*prop.table(table(data.male$usertype)),digits=1)

mylabs1=paste(names(usertype.percentage),"\n",usertype.percent1,"%",sep=" ")

pie(usertype.percent1,labels=mylabs1,main="男性",col=c('grey','lightblue'))

usertype.percent2=round(100\*prop.table(table(data.female$usertype)),digits=1)

mylabs2=paste(names(usertype.percent2),"\n",usertype.percent2,"%",sep=" ")

pie(usertype.percent2,labels=mylabs2,main="女性",col=c('grey','lightblue'))

（肖婧学）

data<-read.csv("201609-citibike-tripdata.csv",header=T,sep=",")

bike<-data

library("ggplot2")

##性别比例

bike<-subset(bike,gender!="0")

bike$gender<-ifelse(bike$gender==1,"男","女")

summary(as.factor(bike$gender))

ggplot(bike,aes(x=gender,fill=gender))+geom\_bar(width=0.4,fill='lightblue')+

labs(x="",y="频数")

##年龄构成

bike$age<-2016-bike$birth.year

bike$age1[bike$age<20]<-"20岁以下"

bike$age1[bike$age>=20&bike$age<30]<-"20-30岁"

bike$age1[bike$age>=30&bike$age<40]<-"30-40岁"

bike$age1[bike$age>=40&bike$age<50]<-"40-50岁"

bike$age1[bike$age>=50&bike$age<60]<-"50-60岁"

bike$age1[bike$age>=60&bike$age<70]<-"60-70岁"

bike$age1[bike$age>=70]<-"70岁以上"

bike$age1 = factor(bike$age1,

levels=c('20岁以下','20-30岁','30-40岁','40-50岁','50-60岁','60-70岁','70岁以上'))

summary(bike$age1)

bike<-subset(bike,age1!="NA")

ggplot(bike,aes(x=age1))+geom\_bar(width=0.4,fill='lightblue')+

labs(x="",y="频数")

##时刻

bike$starttime[bike$starttime<=5]<-"5点及以前"

bike$starttime=factor(bike$starttime,

level=c("5点及以前","6","7","8","9","10","11","12","13","14","15","16","17","18","19","20","21","22","23"))

summary(bike$starttime)

ggplot(bike,aes(x=starttime))+geom\_bar(width=0.5,fill='lightblue')+

labs(x="出发时间",y="频数")

##星期

bike$date<-as.Date(bike$date,"%m/%d/%Y")

bike$week<-format(bike$date,format("%A"))

bike$week<-factor(bike$week,

levels=c('星期一','星期二','星期三','星期四','星期五','星期六','星期日'))

bike1<-subset(bike,bike$date<as.Date("2016-09-15"))

summary(bike1$week)

ggplot(bike1,aes(x=week,fill=factor(week)))+geom\_bar(width=0.4)+

labs(x="",y="频数")+

theme(legend.position="none")

##时长

bike<-subset(bike,tripduration<=4000)

summary(bike$tripduration)

binsize<-diff(range(bike$tripduration))/30

ggplot(bike,aes(x=tripduration))+

geom\_histogram(aes(y=..density..),binwidth=binsize,fill="pink",colour="blue")+

labs(x="时长(单位：秒)",y="概率密度")