Homework 2

df=read.csv("D:/Old_Data/math/Data science toseeh/Files/googleplaystore3-1.csv",header=TRUE, stringsAsFactors = F)

b)

summary(df)

دارای 8 ویژگی زیر است.

X,App,Category,Rating,Reviews,Type,Last. Updates,Installs

دو ویژگی Rating و Reviews دارای مقادیر گمشده هستند.

str(df)

در مورد نوع هر یک از ویژگی ها توضیح داده. به عنوان مثال App کاراکتر است.

c)

دو ویژگی Rating و Reviews به ترتیب دارای 1473 و 200 مقدار گمشده هستند.

d)

df$Category=as.factor(df$Category)

levels(df$Category)

levels(df$Category)=seq(1,33)

levels(df$Category)

در str(df) مشاهده کردیم که category یک کاراکتر است نه فاکتور. پس سطوح آن معنایی ندارد. به همین دلیل ابتدا آن را به یک فاکتور تبدیل کرده و سپس سطوح آن را به 1 تا 33 تغییر داده.

e)

تابع unique برای استخراج یکتای مقادیر در یک بردار است. به این منظور که اگر از یک مقدار به صورت تکراری در بردار داشتیم فقط یک بار در نظر گرفته میشود. (مثل مجموعه ها که عَو تکراری را حذف میکنند و در پایان از هر عَو فقط یکی داریم.)

unique(df$Type)

دارای سه مقدار یکتا است با نام های: "Paid" " " "Free".

به نظر میرسد که " " باید مقدار گمشده باشد. نصب هر نرم افزار دو حالت دارد یا رایگان است و یا باید پول پرداخت شود.

f)

تبدیل " " در ویژگی Type به مقادیر گمشده.

df$Type[df$Type==" "]=NA

is.na(df$Type)

g)

در str(df) دیدیم که Type یک کاراکتر است و فاکتور نیست. برای یافتن سطوح آن، اول به فاکتور تبدیل کرده.

df$Type=as.factor(df$Type)

levels(df$Type)

summary(df)

دارای دو سطح free و paid است و 400 مقدار گمشده دارد.

h)

last.update از نوع کارکتر است. تابعی مانند is.date نداشتیم که بررسی کنم آیا این وِیژگی از نوع زمان هست یا نه. اما تابع weekdays را روی اولین داده در این ستون بررسی کردم و خطا میدهد که از نوع کاراکتر است.

```
is.character(df$Last.Updated)

#weekdays(df$Last.Updated[1])

str(df$Last.Updated)

df$Last.Updated=as.Date(df$Last.Updated, tryFormats =
c("%m/%d/%Y","%Y/%d/%m","%Y-%d-%m","%m-%d-
%Y","%d/%m/%Y","%Y/%m/%d","%Y-%m-%d","%d-%m-%Y"))
```

این دستور نوع ستون مورد نظر را به تاریخ تبدیل میکند. اما چون در سیستم های مختلف فرمت تاریخ متفاوت است، حالت های مختلف را در نظر گرفته.

i)

```
df$Updates=Sys.Date()-df$Last.Updated
```

فاصله زمانی اخرین باری که نرم افزار دانلود شده تا تاریخ سیستم را بدست می آورد و در ستون جدیدی به نام Updates قرار میدهد. البته باید توجه داشت اگر میخواهیم این ستون عددی باشد باید دستور سه خط بعدی را اجرا نکرد.

```
df$Updates

summary(df)


df$Updates=as.numeric(df$Updates)

summary(df)

df$Updates

str(df)
```

j)

```
head(df$Installs,5)
```

ستون Installs تعداد دفعاتی که یک نرم افزار نصب شده است را نمایش میدهد. اما مسئله ای که وجود دارد این است که اسن ستون از نوع کاراکتر است و در صورتی که به عدد تبدیل شود تعداد کاراکتر ها را به عنوان عدد در نظر میگیرد.

هدف نوشتن تابعی است که هر یک از درایه های این ستون را به عدد تبدیل کند به این صورت که دو کاراکتر "+" و "," را حذف کند.

```
split=function(x){
  for (i in 1:length(df$Installs)){
    spinst=strsplit(df$Installs[i],"")
    spinst=unlist(spinst)
    spinst=spinst[spinst!="+" & spinst!=","]
    df$Installs[i]=paste(spinst,collapse = "")
  }
return(df$Installs)
}
df$Installs=split(df$Installs)
df$Installs
df$Installs=as.numeric(df$Installs)
df$Installs
```

k)

```
library(mice)
library(VIM)
```

این دو کتابخانه برای بررسی مقادیر گمشده فراخوانی شده اند.

```
df=df[,c(-1,-2,-7,-8)]
```

حذف ستون های اول، دوم، هفتم و هشتم.

```
summary(df)
impute=mice(df,2,meth=c("polyreg","sample","pmm","logreg","pmm"))
```

پس از حذف ستون ها، 5 ویژگی باقی مانده است که سه تای آنها دارای مقادیر گمشده هستند. حال با استفاده از این دستور مقادیر گمشده هر ستون را با استفاده از روش های مختلف برآورد میکنیم. فقط باید توجه داشت که logreg برای برآورد مقادیر گمشده ستون هایی به کار می رود که دارای دو سطح باشند و poly reg برای برآورد مقادیر گمشده ستون هایی که دارای حداقل دو سطح باشند و یا کارکتر ها که قابل مقایسه نباشند.

برآورد با pmm به این صورت است که یک ستون با مقادیر گمشده را به عنوان متغیر پاسخ در نظر میگیرد و ستون های که مقدار گمشده ندارند را به عنوان متغیر های مستقل. از روش رگرسیون برای براورد استفاده میکند و از بین دو مرحله آن برآوردی را انتخاب میکند که توزیع آن به توزیع داده ها نزدیک تر است.

برآورد با sample به این صورت است که نمونه ای تصادفی از داده ها در نظر میگیرد و جایگزین میکند. چون دو بار براورد کردن را انجام میدهیم آن براوردی را در نظر میگیرد که مقادیر مشاهده شده با باروردشان نزدیک تر است.

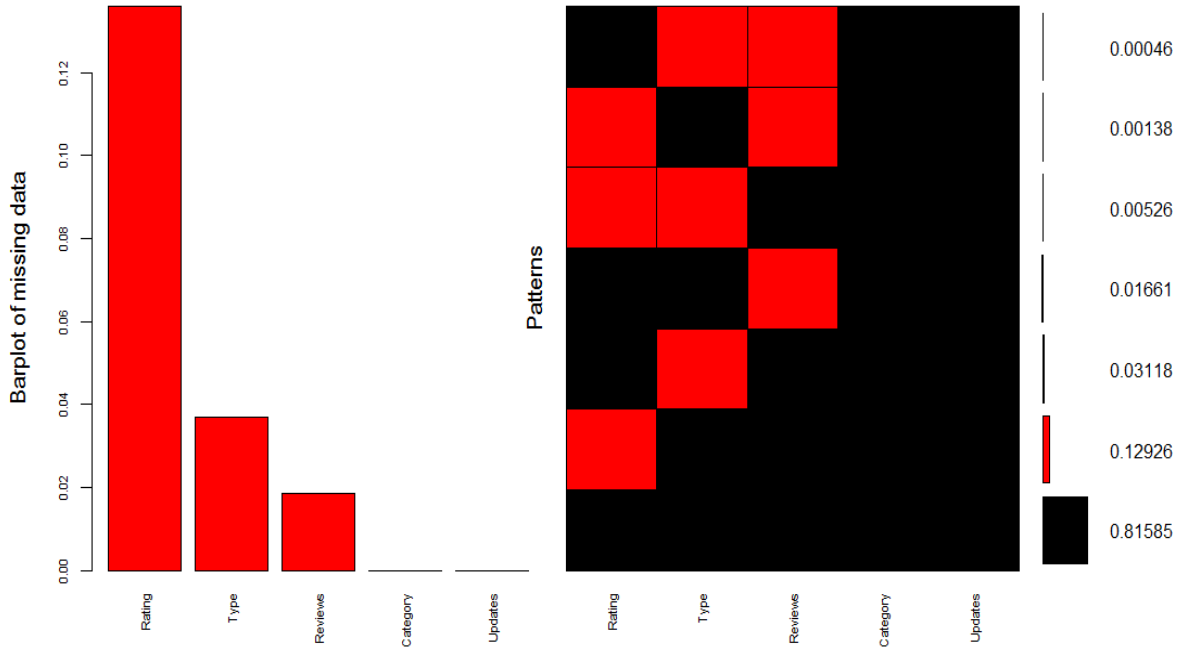
البته روش های دیگری نیز وجود دارد و ساده ترین آن این است که یا سطر هایی که مقادیر گمشده دارند حذف شود (که توصیه نمشود چون داده از دست میدهیم) و یا اینکه میانگین هر ستون را به جای مقادیر گمشده گذاشت.

I)

```
aggr<- aggr(df, col=c('black','red'), numbers=TRUE,
sortVars=TRUE,labels=names(df), cex.axis=.7, gap=1, ylab=c("Barplot of missing data","Patterns"))
```

اهداف رسم این نمودار:

1) یک نمودار میله ای داریم که تعداد مقادیر گمشده هر یک از ستون ها را نمایش میدهد و متوجه میشویم که بیشترین مقدار گمشده در کدام ستون بوده است.
2) یک الگو از مقادیر گمشده میدهد. چند درصد از داده ها هیچ مقدار گمشده ای در این 5 ستون نداشته اند. چند درصد در ستون اول و سوم مقدار گمشده داشته اند....

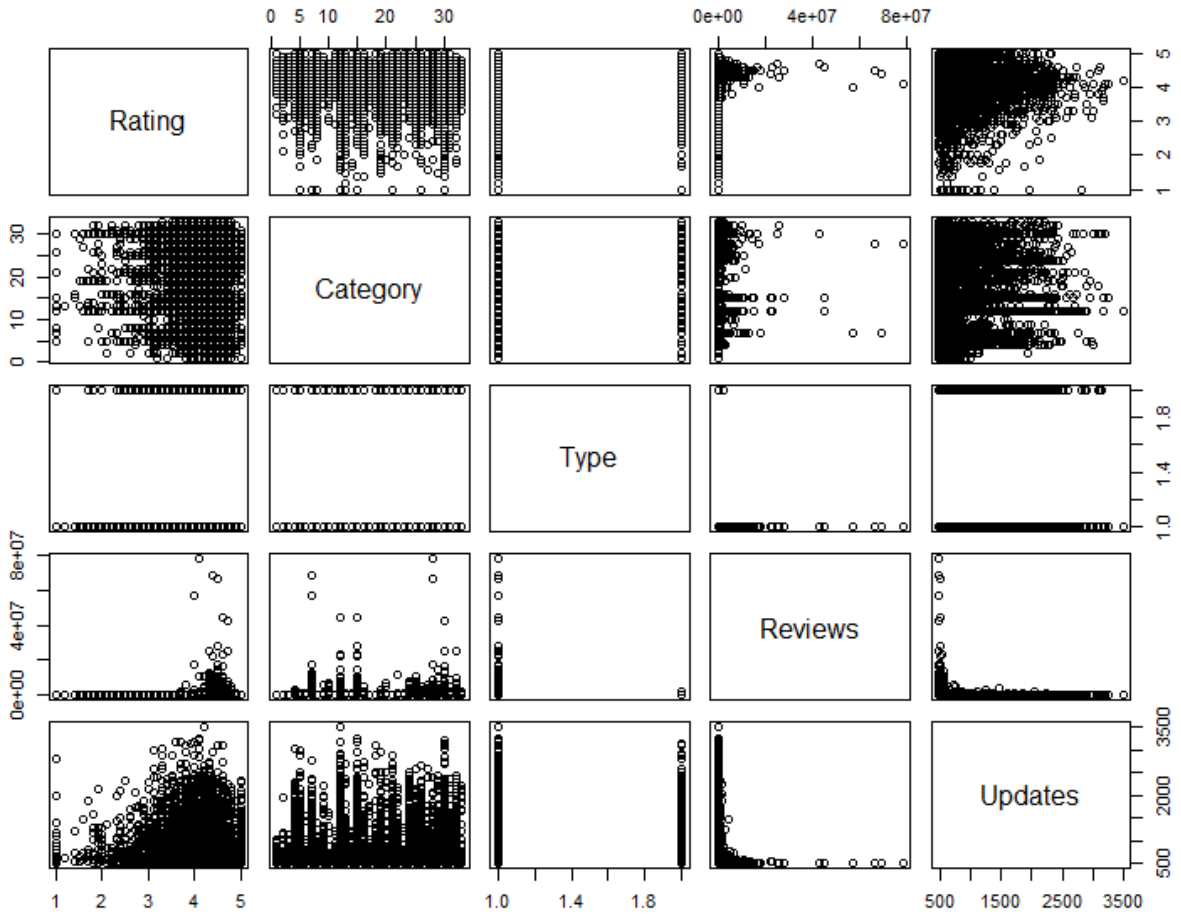


در اینجا مشاهده میککنیم که بیشترین مقدار گمشده مربوط به rating است و دو ستون category و updates مقدار گمشده ندارند.

در الگوی داده شده میبینیم که حدود 81 درصد سطرها هیچ مقدار گمشده ای در 5 ستون ندارند. سطرهایی با مقادیر گمشده در ستون rating و type داریم و...

اما رابطه خاصی بین این مقادیر گمشده نیست مثلا نمیتوان گفت اگر داده ای در ستون اول دارای مقدار گمشده بوده در ستون دوم هم مقدار گمشده دارد. در حقیقت ستون ها رابطه خاصی با هم ندارند.

در نمودار نقطه ای زیر هم میتوان عدم همبستگی بین ستون ها را دید.



Margin Plot

marginplot(df[,c(2,3)])

از این نمودار برای مشاهده توزیع داده ها در ستون دوم زمانی که درایه نظیر در ستون سوم مشاهده شده یا گمشده است استفاده میکنیم.

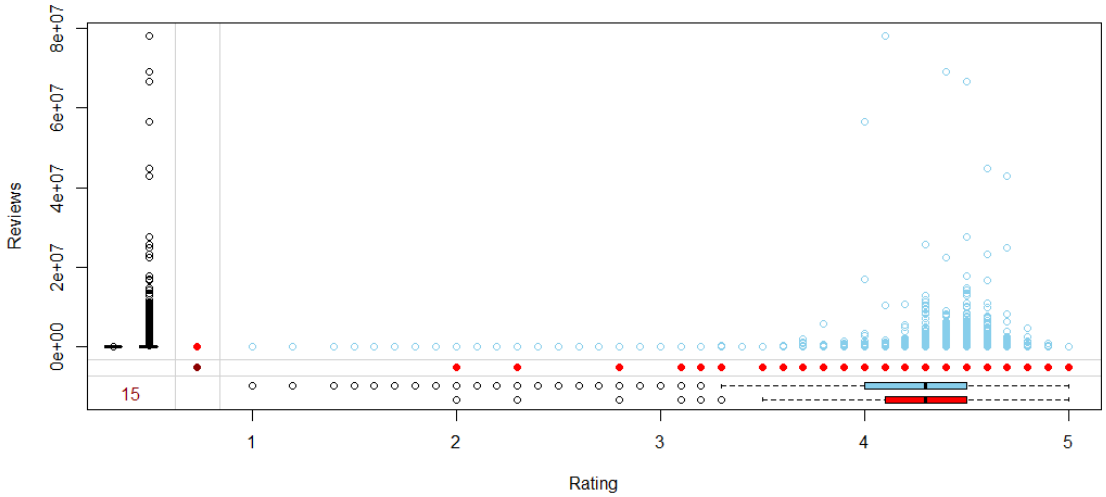
نمودار زیر مربوط به دو ویژگی دوم و سوم است. در این نمودار مشاهده میکنیم که 15 داده هستند که در این دو ستون مقدار گمشده دارند. دو مقدار دیگر که نشان دهنده تعداد کل مقادیر گمشده هر یک از این دو ستون هستند نمایش داده نشده است.

نمودار های جعبه ای :

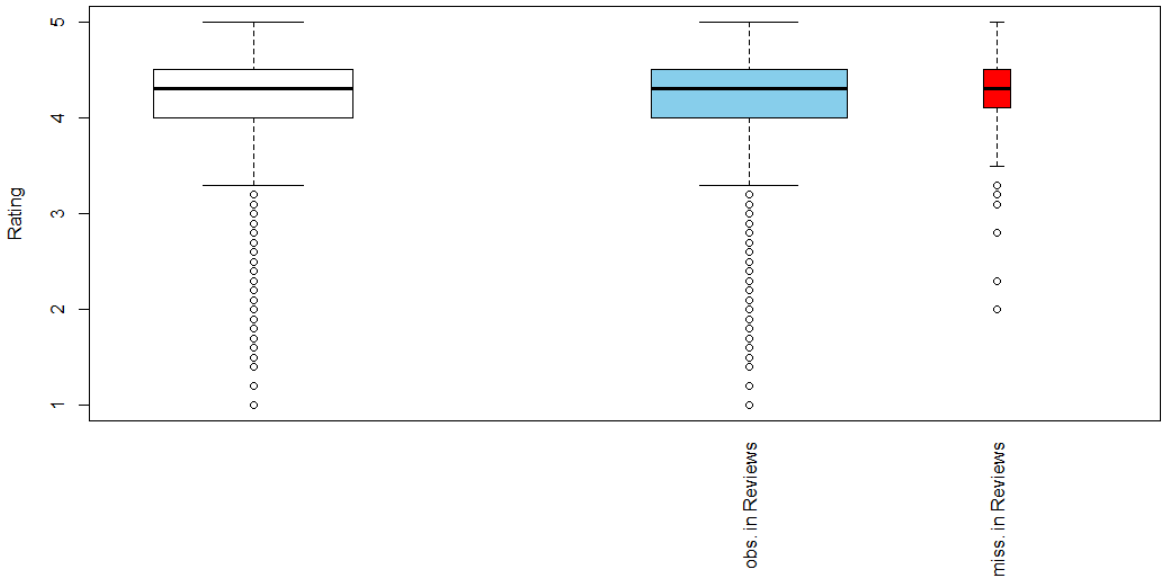
نمودار جعبه ای ابی رنگ در محور افقی نشان دهنده توزیع rating است زمانی که reviews مشاهده شده است.

و نمودار قرمز رنگ در محور افقی نشان دهنده توزیع rating است زمانی که reviews آن ها مقدار گمشده است.

مینیمم rating برای داده هایی که reviews مشاهده شده دارند از مینیمم rating برای داده هایی که reviews مشاهده شده ندارند، کمتر است. میانه در هر دو نمودار جعبه ای برابر است و باید به همین گونه باشد. چون توزیع rating برای مقادیر مشاهده شده و گمشده reviews یکسان است دو نمودار جعبه ای باید همانند هم باشند. تعداد داده های پرت زمانی که reviews مشاهده شده است بیشتر است.از نمودار نقطه ای میتوان دریافت که این دو ویژگی رابطه مستقیمی ندارند. (در کل رابطه ای ندارند) و هر چه reviews کمتر بوده پراکندگی rating کمتر است و همچنین اکثر داده هایی که review کمتری دارند دارای rating بیشتری هستند.

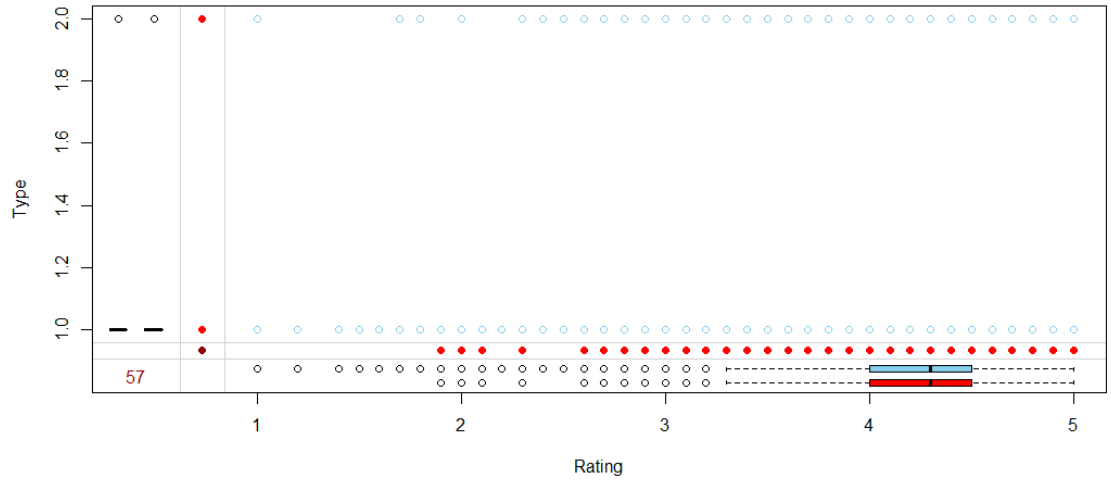


pbox(df[,c(2,3)])



بیشترین تعداد داده های پرت زمانی است که reviews مشاهده شده است. نمودار جعبه ای سمت چپ زمانی است که تمامی مقادیر مشاهده شده و گمشده ستون سوم را همزمان در نظر گرفته ایم.

marginplot(df[,c(2,4)])



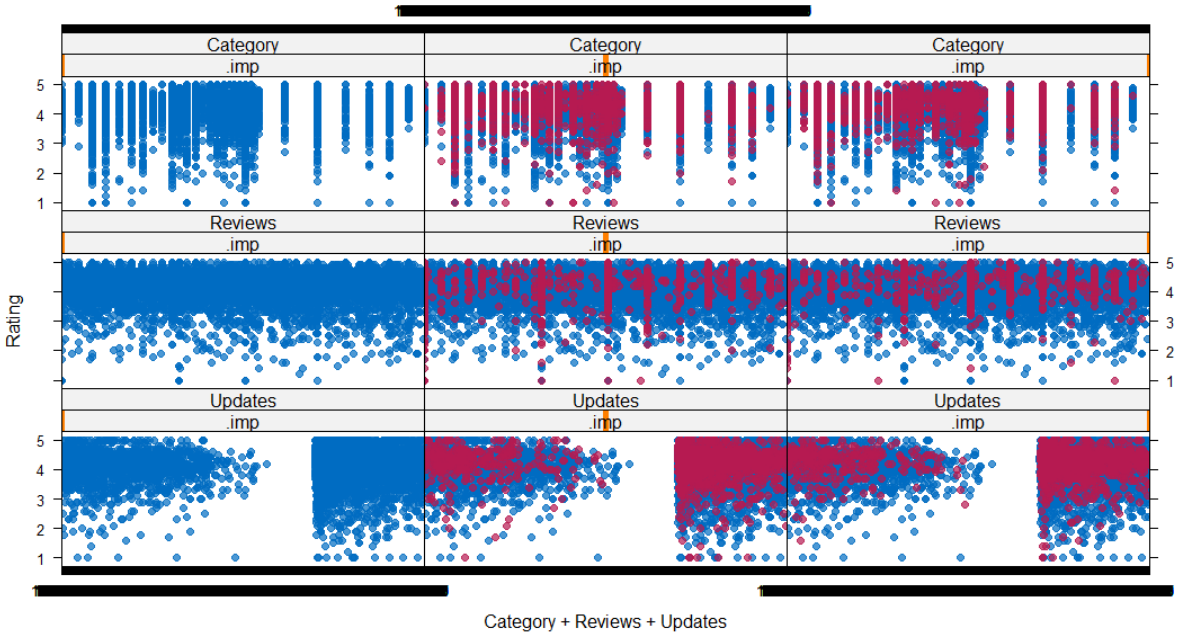
در اینجا مشاهده میکنیم که دو نمودار میله ای یکسان هستند. در حقیقت توزیع داده ها در هر دو حالت ویژگی type یکسان است. 57 سطر داریم که مقادیر ستون دوم و چهارم برای انها گمشده است.

m)

این نمودار برای نمایش براورد نقاط است در دو مرحله ای که براورد را انجام دادیم.

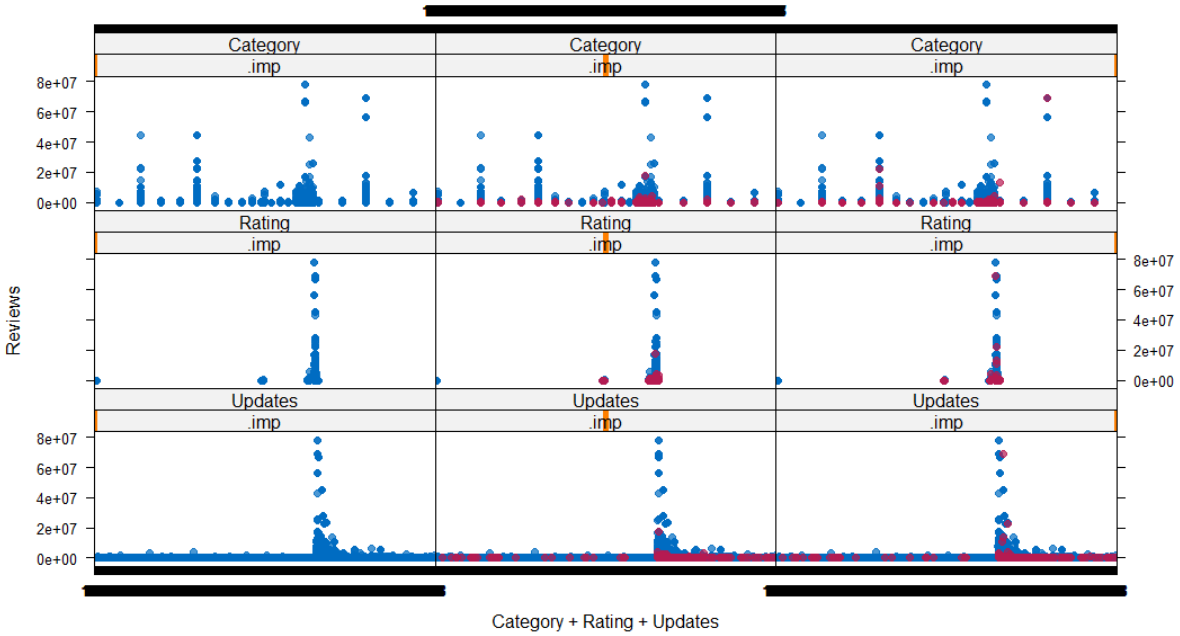
با این نمودار میتوان براورد نقاط در دو مرحله را با هم مقایسه کرد. اگر برآورد ها در دو مرحله خیلی متفاوت بودند یعنی روشی که برای براورد ان ویژگی استفاده کردیم مناسب نبوده و یا دلایل دیگری وجود داشته مثلا تعداد مقادیر گمشده کم بوده.

```
xyplot(impute, Rating ~Category+Reviews +Updates | .imp, pch = 20, cex = 1.4)
```



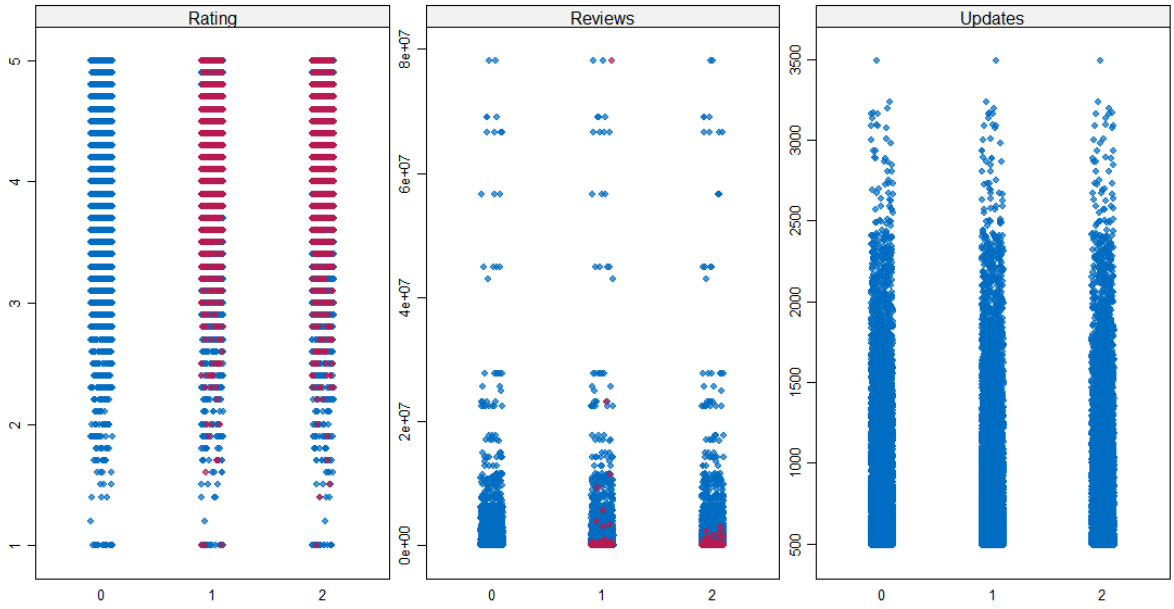
نمودار نقطه ای rating را با دیگر ویژگی ها رسم کرده ایم. میتوان دید براورد ها در دو مرحله تقریا یکسان هستند. پس روشی که برای براورد نقاط در این ستون استفاده کرده ایم مناسب است.

xyplot(impute, Reviews~Category+Rating +Updates | .imp, pch = 20, cex = 1.4)



stripplot(impute, pch = 20, cex = 1.2)

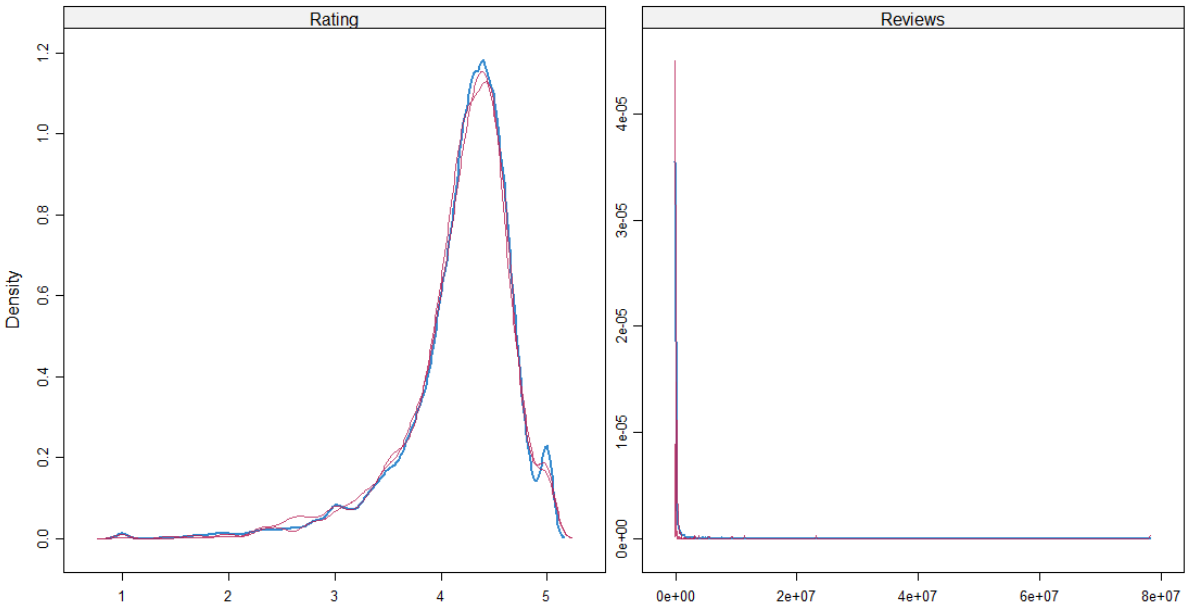
این نمودار براورد هر ویژگی را روی کل داده ها ر دو مرحله نمایش میدهد.



در هر ویژگی تقریبا برراورد دو مرحله شبیه بهم هستندد. ویژگی type چون باینری است رسم نشده.

densityplot(impute)

این نمودار توزیع داده ها برای هر ویژگی قبل از براورد کردن و توزیع داده ها در هر دو مرحله براورد است. توزیع داده ها در هر دو مرحله تقریبا با توزیع ابی رنگ یکسان است پس میتوان هر یک از دو مرحله را انتخاب نمود و داده های براورد شده را جایگزین مقادیر گمشده کرد.



n)

com=complete(impute,1)

summary(com)

داده های براورد شده جایگزین شده اند.

پس از انجام این مرحله دیگر هیچ مقدار گمشده ای نداریم.