

GR5206 Midterm Exam

Yuhao Wang and yw3204

10/19/2018

The STAT GR5206 Fall 2018 Midterm is open notes, open book(s), open computer and online resources are allowed. Students are **not** allowed to communicate with any other people regarding the exam. This includes emailing fellow students, using WeChat and other similar forms of communication. Before the exam, the students should **turn off** their cellphone and pass it to the left side of each row. At the same time, please **close** the mailbox and **log out** WeChat and all the other apps for messaging and chatting. If there is any suspicion of one or more students cheating, further investigation will take place. If students do not follow the guidelines, they will receive a zero on the exam and potentially face more severe consequences. The exam will be posted on Canvas at **2:50PM**. Students are required to submit both the .pdf and .Rmd files on Canvas (or .html if you must) by **4:30PM**. Late exams will not be accepted.

Part 1 (Google Play Store Apps Data - Split/Apply/Combine and R plot, 11 + 2 pts)

We work on the **apps** dataset which contains approximately 7,700 Google Play Store apps. There are 13 features that describe a given app. They are:

- **App** – Application name.
- **Category** – Category the app belongs to.
- **Rating** – Overall user rating of the app (between 0 and 5).
- **Reviews** – Number of user reviews for the app.
- **Size** – Size of the app.
- **Installs** – Number of user downloads/installs for the app.
- **Type** – Paid or Free
- **Price** – Price of the app
- **Content Rating** Age group the app is targeted at
- **Genres** An app can belong to multiple genres (apart from its main category). For example, a musical family game will belong to Music, Game, Family genres.
- **Last Updated** Date when the app was last updated on Play Store
- **Current Ver** Current version of the app available on Play Store
- **Android Ver** Minimum required Android version

Read in the dataset using the following code:

```
apps<-read.csv("apps.csv", header = T)
apps$Reviews<-as.numeric(apps$Reviews)
apps$Installs<-factor(apps$Installs, level= c("1+", "5+", "10+", "50+", "100+", "500+", "1,000+", "5,000+"))
head(apps)
```

##		App	Category	Rating
## 1	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	
## 2	Coloring book moana	ART_AND_DESIGN	3.9	

```
## 3 U Launcher Lite - FREE Live Cool Themes, Hide Apps ART_AND_DESIGN 4.7
## 4 Sketch - Draw & Paint ART_AND_DESIGN 4.5
## 5 Pixel Draw - Number Art Coloring Book ART_AND_DESIGN 4.3
## 6 Paper flowers instructions ART_AND_DESIGN 4.4
## Reviews Size Installs Type Price Content.Rating
## 1 159 19.0 10,000+ Free 0 Everyone
## 2 967 14.0 500,000+ Free 0 Everyone
## 3 87510 8.7 5,000,000+ Free 0 Everyone
## 4 215644 25.0 50,000,000+ Free 0 Teen
## 5 967 2.8 100,000+ Free 0 Everyone
## 6 167 5.6 50,000+ Free 0 Everyone
## Genres Last.Updated Current.Ver Android.Ver
## 1 Art & Design 7-Jan-18 1.0.0 4.0.3 and up
## 2 Art & Design;Pretend Play 15-Jan-18 2.0.0 4.0.3 and up
## 3 Art & Design 1-Aug-18 1.2.4 4.0.3 and up
## 4 Art & Design 8-Jun-18 Varies with device 4.2 and up
## 5 Art & Design;Creativity 20-Jun-18 1.1 4.4 and up
## 6 Art & Design 26-Mar-17 1 2.3 and up
```

Problem 1.0

Check the dimension of `apps`, make sure that there are 7,726 lines and 13 variables (features). [1 pt]

```
# code goes here
dim(apps)
```

```
## [1] 7726 13
```

Problem 1.1

In order to get an overview of the dataset, we want to check some summary statistics of each variable, and this can be done by calling the R function `summary()`. Compute the summary statistics of all 13 variables and display the results in a **list**. To receive full credit, you must use a vectorized function from the `apply` family or `plyr` family. [2 pts]

```
# code goes here
lapply(apps, summary)
```

```
## $App
## ROBLOX
## 9
## 8 Ball Pool
## 7
## Candy Crush Saga
## 7
## Bubble Shooter
## 6
## Helix Jump
## 6
## Nick
## 6
## Subway Surfers
## 6
## Temple Run 2
```

##		6
##	Zombie Catchers	
##		6
##	Angry Birds Classic	
##		5
##	BeautyPlus - Easy Photo Editor & Selfie Camera	
##		5
##	Farm Heroes Saga	
##		5
##	Flow Free	
##		5
##	Granny	
##		5
##	MeetMe: Chat & Meet New People	
##		5
##	Plants vs. Zombies FREE	
##		5
##	Skyscanner	
##		5
##	theScore: Live Sports Scores, News, Stats & Videos	
##		5
##	Wish - Shopping Made Fun	
##		5
##	A&E - Watch Full Episodes of TV Shows	
##		4
##	Angry Birds Rio	
##		4
##	Babbel - Learn Languages	
##		4
##	Block Puzzle	
##		4
##	busuu: Learn Languages - Spanish, English & More	
##		4
##	Clash of Clans	
##		4
##	Clash Royale	
##		4
##	DC Super Hero Girls	
##		4
##	DRAGON BALL LEGENDS	
##		4
##	ESPN Fantasy Sports	
##		4
##	Expedia Hotels, Flights & Car Rental Travel Deals	
##		4
##	Garena Free Fire	
##		4
##	Google News	
##		4
##	Hill Climb Racing	
##		4
##	imo free video calls and chat	
##		4
##	Meetup	

##		4
##	My Talking Angela	
##		4
##	Pou	
##		4
##	PUBG MOBILE	
##		4
##	Rosetta Stone: Learn to Speak & Read New Languages	
##		4
##	Shutterfly: Free Prints, Photo Books, Cards, Gifts	
##		4
##	Talking Tom Gold Run	
##		4
##	TED	
##		4
##	The CW	
##		4
##	WatchESPN	
##		4
##	Wordscapes	
##		4
##	Zombie Hunter King	
##		4
##	Adult Dirty Emojis	
##		3
##	Amazon Shopping	
##		3
##	BBW Dating & Plus Size Chat	
##		3
##	BET NOW - Watch Shows	
##		3
##	Block Craft 3D: Building Simulator Games For Free	
##		3
##	Camera for Android	
##		3
##	Camera360: Selfie Photo Editor with Funny Sticker	
##		3
##	Candy Crush Soda Saga	
##		3
##	Cardiac diagnosis (heart rate, arrhythmia)	
##		3
##	Chick-fil-A	
##		3
##	ClassDojo	
##		3
##	CNN Breaking US & World News	
##		3
##	Cut the Rope FULL FREE	
##		3
##	Daily Yoga - Yoga Fitness Plans	
##		3
##	Dog Run - Pet Dog Simulator	
##		3
##	Dr. Panda & Toto's Treehouse	

```

##                                     3
##                               Dream League Soccer 2018
##                                     3
##                               Dropbox
##                                     3
##                               Elmo Calls by Sesame Street
##                                     3
##                               Equestria Girls
##                                     3
##                               ES File Explorer File Manager
##                                     3
##                               Facetune - For Free
##                                     3
##                               Fancy
##                                     3
##                               Fashion in Vogue
##                                     3
##                               Firefox Focus: The privacy browser
##                                     3
##                               Frozen Free Fall
##                                     3
##                               Google Ads
##                                     3
##                               Gyft - Mobile Gift Card Wallet
##                                     3
##                               Hily: Dating, Chat, Match, Meet & Hook up
##                                     3
##                               HISTORY: Watch TV Show Full Episodes & Specials
##                                     3
##                               Human Anatomy Atlas 2018: Complete 3D Human Body
##                                     3
##                               Hungry Shark Evolution
##                                     3
##                               IMDb Movies & TV
##                                     3
##                               JackThreads: Men's Shopping
##                                     3
##                               Just She - Top Lesbian Dating
##                                     3
##                               KAYAK Flights, Hotels & Cars
##                                     3
##                               Khan Academy
##                                     3
##                               Lifetime - Watch Full Episodes & Original Movies
##                                     3
##                               LivingSocial - Local Deals
##                                     3
##                               MARVEL Strike Force
##                                     3
##                               Miraculous Ladybug & Cat Noir - The Official Game
##                                     3
##                               Movies by Flixster, with Rotten Tomatoes
##                                     3
##                               muzmatch: Muslim & Arab Singles, Marriage & Dating

```

```

##                                     3
## mySugr: the blood sugar tracker made just for you
##                                     3
##     Nike Training Club - Workouts & Fitness Plans
##                                     3
##             NYTimes - Latest News
##                                     3
##             0-Star
##                                     3
##             OkCupid Dating
##                                     3
##             Open Camera
##                                     3
##     Papumba Academy - Fun Learning For Kids
##                                     3
##             Periscope - Live Video
##                                     3
##             Pixel Art: Color by Number Game
##                                     3
##             PJ Masks: Moonlight Heroes
##                                     3
##             (Other)
##             7350
##
## $Category
##     ART_AND_DESIGN    AUTO_AND_VEHICLES    BEAUTY
##           59           63           37
## BOOKS_AND_REFERENCE    BUSINESS    COMICS
##       144           246           48
##     COMMUNICATION    DATING    EDUCATION
##       211           173           110
##     ENTERTAINMENT    EVENTS    FAMILY
##       90           38           1617
##           FINANCE    FOOD_AND_DRINK    GAME
##       266           84           974
## HEALTH_AND_FITNESS    HOUSE_AND_HOME    LIBRARIES_AND_DEMO
##       223           56           62
##           LIFESTYLE    MAPS_AND_NAVIGATION    MEDICAL
##       280           95           324
## NEWS_AND_MAGAZINES    PARENTING    PERSONALIZATION
##       169           44           280
##           PHOTOGRAPHY    PRODUCTIVITY    SHOPPING
##       236           235           179
##           SOCIAL    SPORTS    TOOLS
##       177           246           633
## TRAVEL_AND_LOCAL    VIDEO_PLAYERS    WEATHER
##       160           116           51
##
## $Rating
##   Min. 1st Qu.  Median    Mean 3rd Qu.  Max.
## 1.000  4.000  4.300  4.174  4.500  5.000
##
## $Reviews
##   Min. 1st Qu.  Median    Mean 3rd Qu.  Max.

```

```

##          1          107          2324          294777          38959 44893888
##
## $Size
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##    0.008   5.300   14.000   22.959   33.000  100.000
##
## $Installs
##           1+           5+           10+           50+           100+
##             3             9             67             56             303
##          500+        1,000+        5,000+       10,000+       50,000+
##           197           690           420           969           436
##        100,000+      500,000+    1,000,000+    5,000,000+   10,000,000+
##          1037           490           1301           535           825
##    50,000,000+  100,000,000+  500,000,000+ 1,000,000,000+
##           147            201             30             10
##
## $Type
## Free Paid
## 7147  579
##
## $Price
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##    0.000   0.000   0.000   1.128   0.000  400.000
##
## $Content.Rating
##      Everyone Everyone 10+   Mature 17+      Teen
##           6172           318           368           868
##
## $Genres
##
##              Tools              Entertainment
##              633              448
##           Education              Medical
##              417              324
##              Action              Personalization
##              322              280
##           Lifestyle              Finance
##              279              266
##              Sports              Business
##              260              246
##           Photography              Productivity
##              236              235
##           Health & Fitness              Communication
##              223              211
##              Arcade              Simulation
##              186              182
##           Shopping              Social
##              179              177
##              Dating              News & Magazines
##              173              169
##              Casual              Travel & Local
##              160              159
##           Books & Reference              Video Players & Editors
##              144              115
##              Puzzle              Role Playing

```

##	108	103
##	Strategy	Maps & Navigation
##	96	95
##	Food & Drink	Racing
##	84	83
##	Adventure	Auto & Vehicles
##	68	63
##	Libraries & Demo	House & Home
##	62	56
##	Art & Design	Weather
##	53	51
##	Comics	Education;Education
##	47	41
##	Card	Board
##	39	38
##	Events	Beauty
##	38	37
##	Parenting	Educational;Education
##	35	33
##	Casino	Educational
##	32	31
##	Casual;Pretend Play	Trivia
##	30	27
##	Word	Education;Pretend Play
##	24	22
##	Educational;Pretend Play	Puzzle;Brain Games
##	18	18
##	Action;Action & Adventure	Entertainment;Music & Video
##	16	16
##	Casual;Action & Adventure	Music
##	15	15
##	Board;Brain Games	Racing;Action & Adventure
##	14	14
##	Adventure;Action & Adventure	Arcade;Action & Adventure
##	13	13
##	Casual;Brain Games	Simulation;Action & Adventure
##	13	11
##	Casual;Creativity	Art & Design;Creativity
##	7	6
##	Education;Action & Adventure	Educational;Brain Games
##	6	6
##	Entertainment;Brain Games	Education;Creativity
##	6	5
##	Educational;Creativity	Parenting;Music & Video
##	5	5
##	Puzzle;Action & Adventure	Role Playing;Action & Adventure
##	5	5
##	Role Playing;Pretend Play	Educational;Action & Adventure
##	5	4
##	Board;Action & Adventure	Casual;Education
##	3	3
##	Education;Music & Video	Entertainment;Action & Adventure
##	3	3
##	Music;Music & Video	Parenting;Education

##		3		3
##	Simulation;Education		Simulation;Pretend Play	
##		3		3
##	Adventure;Education		Art & Design;Pretend Play	
##		2		2
##	Books & Reference;Education		Card;Action & Adventure	
##		2		2
##	Casual;Music & Video		Entertainment;Creativity	
##		2		2
##	Entertainment;Pretend Play		Puzzle;Creativity	
##		2		2
##	Sports;Action & Adventure		Strategy;Action & Adventure	
##		2		2
##	Video Players & Editors;Creativity		Adventure;Brain Games	
##		2		1
##	Arcade;Pretend Play		Board;Pretend Play	
##		1		1
##	Card;Brain Games		Comics;Creativity	
##		1		1
##	Education;Brain Games		(Other)	
##		1		13
##				
##	\$Last.Updated			
##	3-Aug-18	31-Jul-18	1-Aug-18	2-Aug-18
##	205	189	178	173
##				130
##				124
##				114
##	27-Jul-18	24-Jul-18	16-Jul-18	18-Jul-18
##	101	98	93	84
##				84
##				81
##				80
##	17-Jul-18	12-Jul-18	3-Jul-18	19-Jul-18
##	73	71	71	68
##				66
##				63
##				63
##	24-May-18	9-Jul-18	6-Jul-18	13-Jul-18
##	58	58	56	55
##				48
##				46
##				46
##	13-Jun-18	6-Jun-18	19-Jun-18	2-Jul-18
##	41	41	38	37
##				37
##				36
##				36
##	20-Jun-18	28-Jul-18	4-Jul-18	10-Jul-18
##	35	35	34	33
##				33
##				33
##				32
##	29-Jul-18	7-Aug-18	23-May-18	8-Jun-18
##	30	30	29	29
##				28
##				28
##				27
##	18-Jun-18	28-Jun-18	30-May-18	7-Jun-18
##	26	26	26	26
##				25
##				25
##				25
##	11-Jun-18	15-Jul-18	18-May-18	8-Jul-18
##	24	24	24	24
##				24
##				23
##				22
##				22
##	5-Feb-17	1-Jul-18	6-Mar-18	22-Jun-18
##	22	21	20	19
##				19
##				18
##				18
##	10-May-18	16-Mar-18	17-May-18	20-Mar-18
##	17	17	17	17
##				17
##				17
##				17
##	5-Mar-18	11-Apr-18	21-Jul-18	22-Jul-18
##	17	16	16	16
##				15
##				15
##				15
##	5-Apr-18	9-Apr-18	9-May-18	13-Mar-18
##	15	15	15	14
##				14
##				14
##				14
##	27-Mar-18	9-Jun-18	12-Oct-17	13-Apr-18
##	14	14	13	13
##				13
##				13
##				13
##	2-Jan-18	(Other)		
##	13	3533		
##				

## \$Current.Ver		
##	1	1.1 1.2
##	476	206 133
##	2	1.3 1.4
##	128	119 82
##	1.0.1	1.5 Varies with device
##	80	75 73
##	1.0.0	1.6 2.1
##	67	57 56
##	1.0.2	1.0.4 1.7
##	51	47 46
##	1.0.3	3 1.0.6
##	44	44 43
##	1.2.1	2.0.0 1.8
##	43	41 39
##	1.0.5	4 1.1.0
##	38	37 36
##	1.2.0	1.9 1.0.9
##	36	33 32
##	2.3.2	2.4 1.1.1
##	32	32 31
##	2.2	3.1 5
##	31	30 29
##	1.4.0	2.0.1 2.5
##	27	26 26
##	1.0.7	1.0.8 1.1.3
##	25	25 24
##	1.1.2	1.3.0 3.0.0
##	23	22 22
##	1.2.2	1.2.3 2.1.1
##	21	21 21
##	3.3	5.1 6
##	21	21 21
##	2.3	3.1.0 2.6
##	20	20 19
##	4.1	1.1.4 8.2
##	19	18 18
##	1.1.6	1.5.0 2.0.5
##	17	17 17
##	7	1.3.1 2.1.0
##	17	16 16
##	2.1.2	2.7 2.9
##	16	16 16
##	1.01	1.5.1 1.6.1
##	15	15 15
##	2.4.0	2.5.1 3.0.1
##	15	15 15
##	2.0.7	2.4.1 3.1.4
##	14	14 14
##	1.03	1.1.5 1.2.7
##	13	13 13
##	2.8	3.2 5.2
##	13	13 13
##	1.2.6	1.5.2 1.6.2

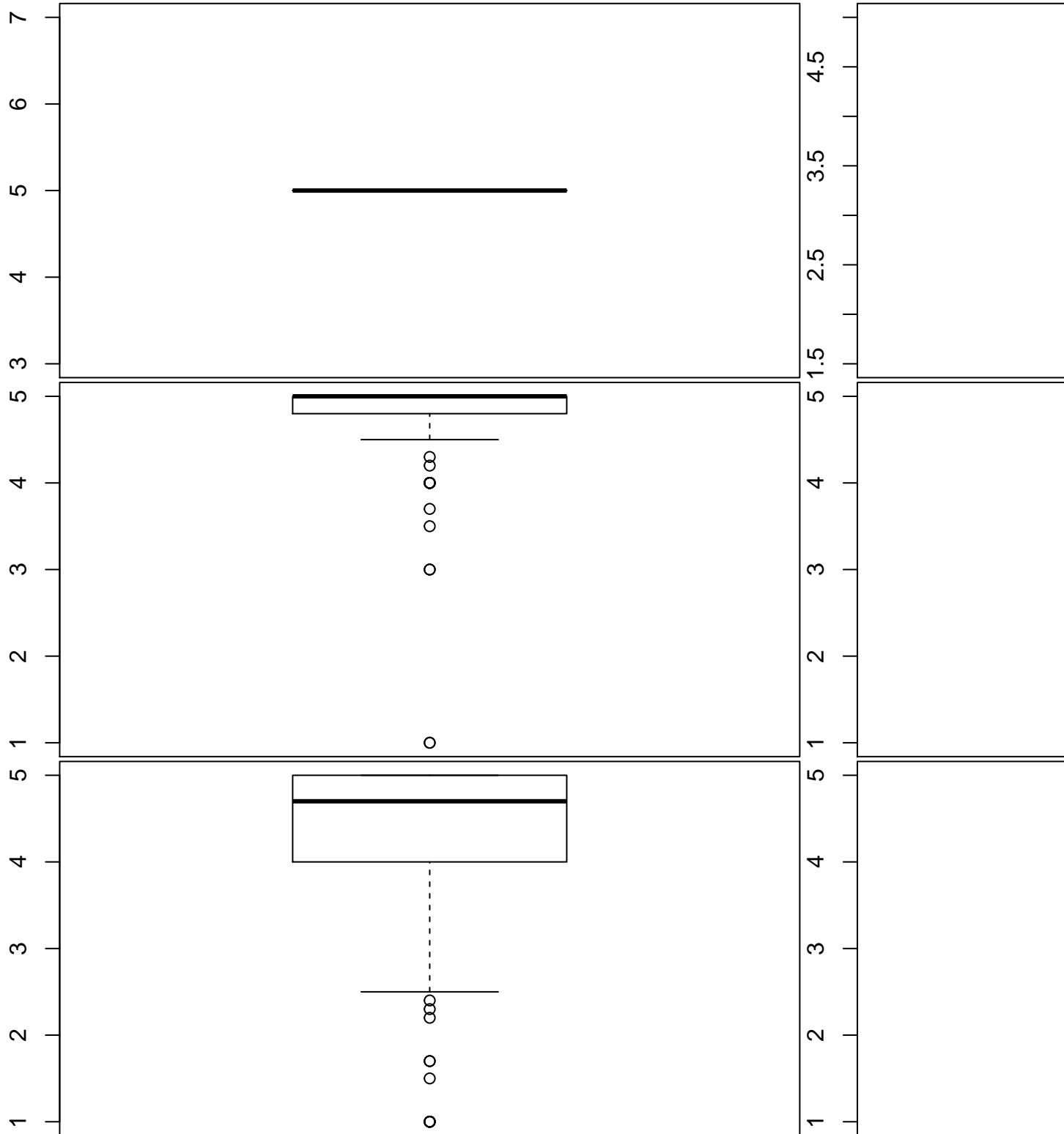
```
##          12          12          12
##        3.0.5        3.6.1        3.8.0
##          12          12          12
##          0.1        1.0.11        1.1.7
##          11          11          11
##          1.11        1.3.3        2.0.3
##          11          11          11
##          2.0.4        2.1.3        2.3.1
##          11          11          11
##          4.1.0        5.9.1.0        1.2.5
##          11          11          10
##          1.2.9        1.6.0        1.7.0
##          10          10          10
##        (Other)
##        4328
##
## $Android.Ver
##    1.0 and up    1.5 and up    1.6 and up
##          2          15          87
##    2.0 and up    2.0.1 and up    2.1 and up
##          27          7          113
##    2.2 and up    2.3 and up    2.3.3 and up
##          206          566          234
##    3.0 and up    3.1 and up    3.2 and up
##          211          8          31
##    4.0 and up    4.0.3 - 7.1.1    4.0.3 and up
##          1109          2          1194
##    4.1 - 7.1.1    4.1 and up    4.2 and up
##          1          1929          318
##    4.3 and up    4.4 and up    4.4W and up
##          195          805          6
##    5.0 - 6.0    5.0 - 8.0    5.0 and up
##          1          2          490
##    5.1 and up    6.0 and up    7.0 - 7.1.1
##          17          45          1
##    7.0 and up    7.1 and up    8.0 and up
##          39          2          5
##          NaN Varies with device
##          2          56
```

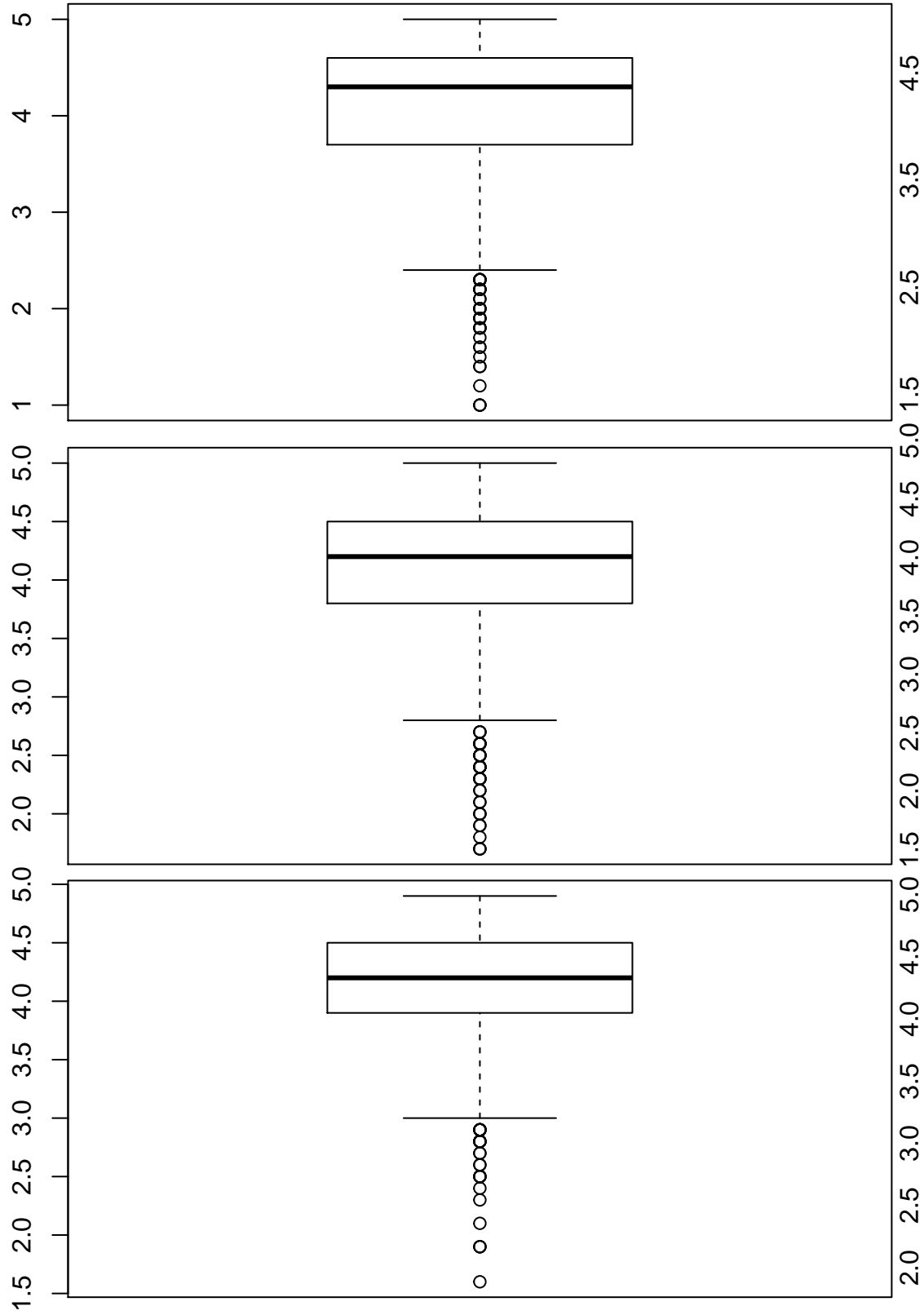
Problem 1.2

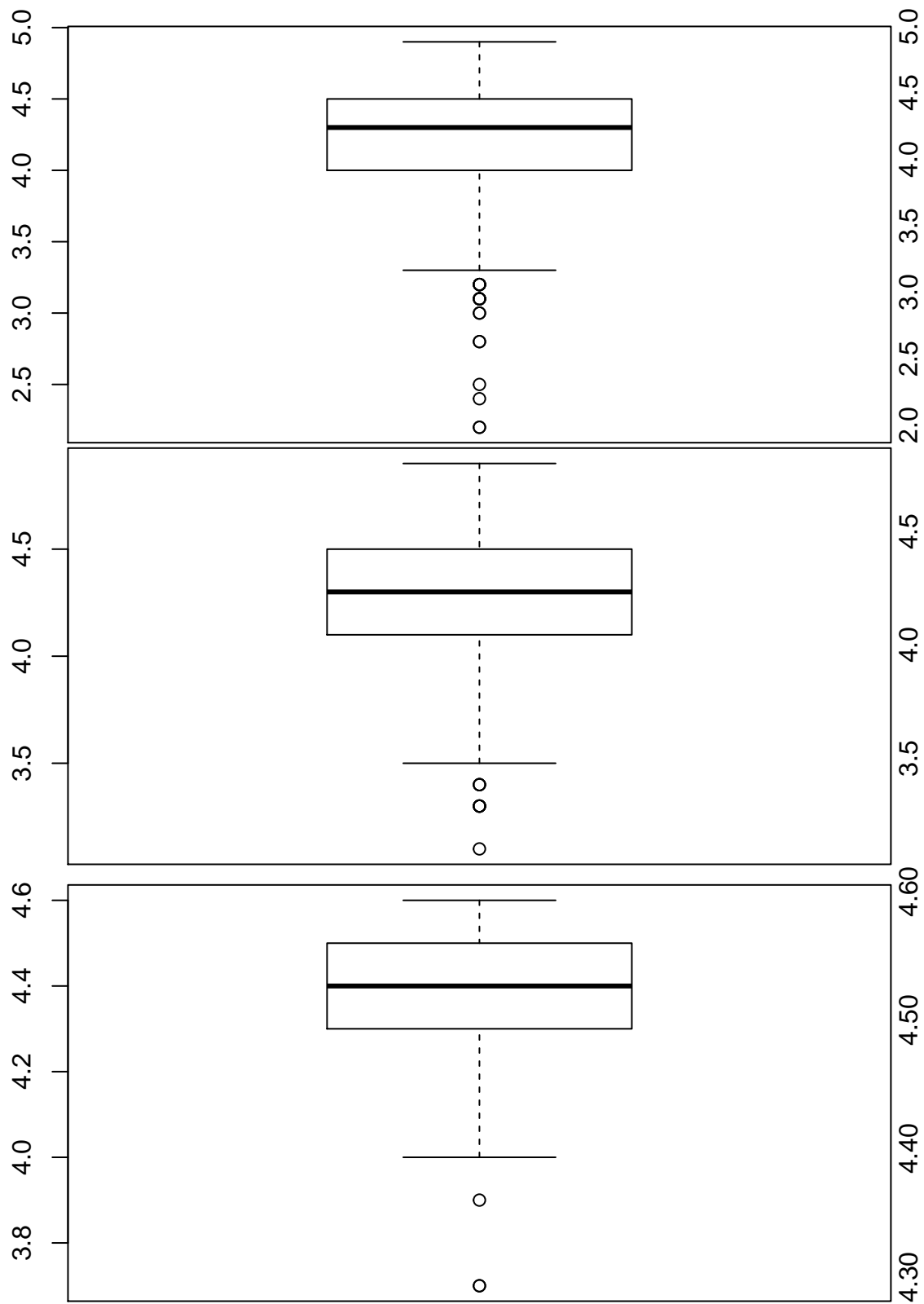
We want to investigate the association between the user's overall rating (`Rating`) of an app and the total number of installs (`Installs`). Use function `plot()` to construct a multiple boxplot of the overall rating of an app split by number of installs (`Installs`). Can you see any relationship in the plot? [3 pts]

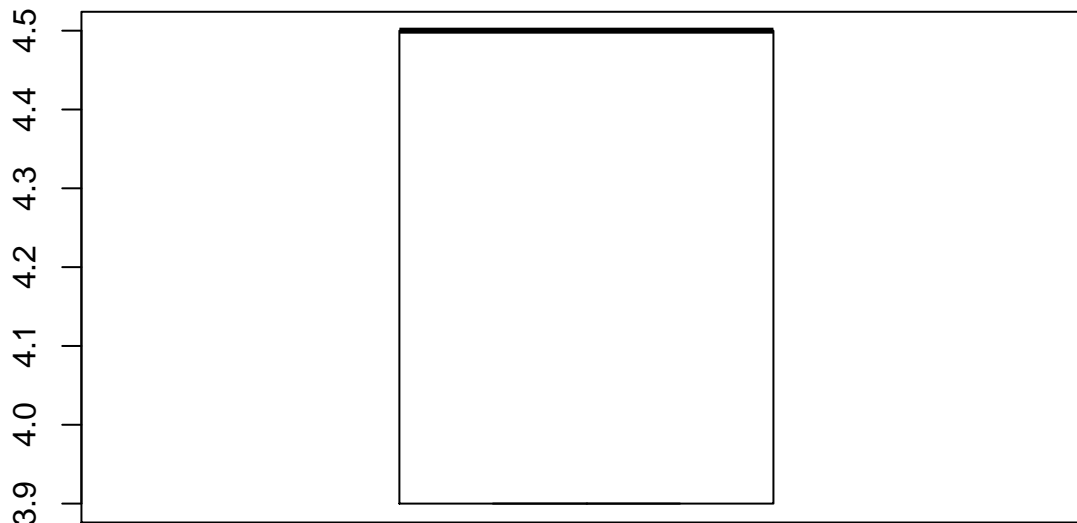
Adjust the labels of *x*-axis. Make sure that all levels of the variable `Installs` show in the plot. [2 extra pts]

```
# code goes here
inst_split <- split(apps, apps$Installs)
box_plt <- function(df) {
  boxplot(df$Rating)
}
sapply(inst_split, box_plt)
```









```
##      1+      5+      10+      50+      100+      500+
## stats Numeric,5 Numeric,5 Numeric,5 Numeric,5 Numeric,5 Numeric,5
## n      3       9      67      56      303      197
## conf  Numeric,2 Numeric,2 Numeric,2 Numeric,2 Numeric,2 Numeric,2
## out   Numeric,0 1.5      Numeric,14 Numeric,6 Numeric,11 Numeric,11
## group Numeric,0 1       Numeric,14 Numeric,6 Numeric,11 Numeric,11
## names "1"      "1"      "1"      "1"      "1"      "1"
##      1,000+    5,000+    10,000+    50,000+    100,000+    500,000+
## stats Numeric,5 Numeric,5 Numeric,5 Numeric,5 Numeric,5 Numeric,5
## n      690     420     969     436     1037     490
## conf  Numeric,2 Numeric,2 Numeric,2 Numeric,2 Numeric,2 Numeric,2
## out   Numeric,38 Numeric,15 Numeric,40 Numeric,17 Numeric,31 Numeric,11
## group Numeric,38 Numeric,15 Numeric,40 Numeric,17 Numeric,31 Numeric,11
## names "1"      "1"      "1"      "1"      "1"      "1"
##      1,000,000+ 5,000,000+ 10,000,000+ 50,000,000+ 100,000,000+
## stats Numeric,5 Numeric,5 Numeric,5 Numeric,5 Numeric,5
## n      1301    535     825     147     201
## conf  Numeric,2 Numeric,2 Numeric,2 Numeric,2 Numeric,2
## out   Numeric,18 Numeric,16 Numeric,9  Numeric,3 Numeric,3
## group Numeric,18 Numeric,16 Numeric,9  Numeric,3 Numeric,3
## names "1"      "1"      "1"      "1"      "1"
##      500,000,000+ 1,000,000,000+
## stats Numeric,5 Numeric,5
## n      30       10
## conf  Numeric,2 Numeric,2
## out   Numeric,2 Numeric,0
## group Numeric,2 Numeric,0
## names "1"      "1"
```

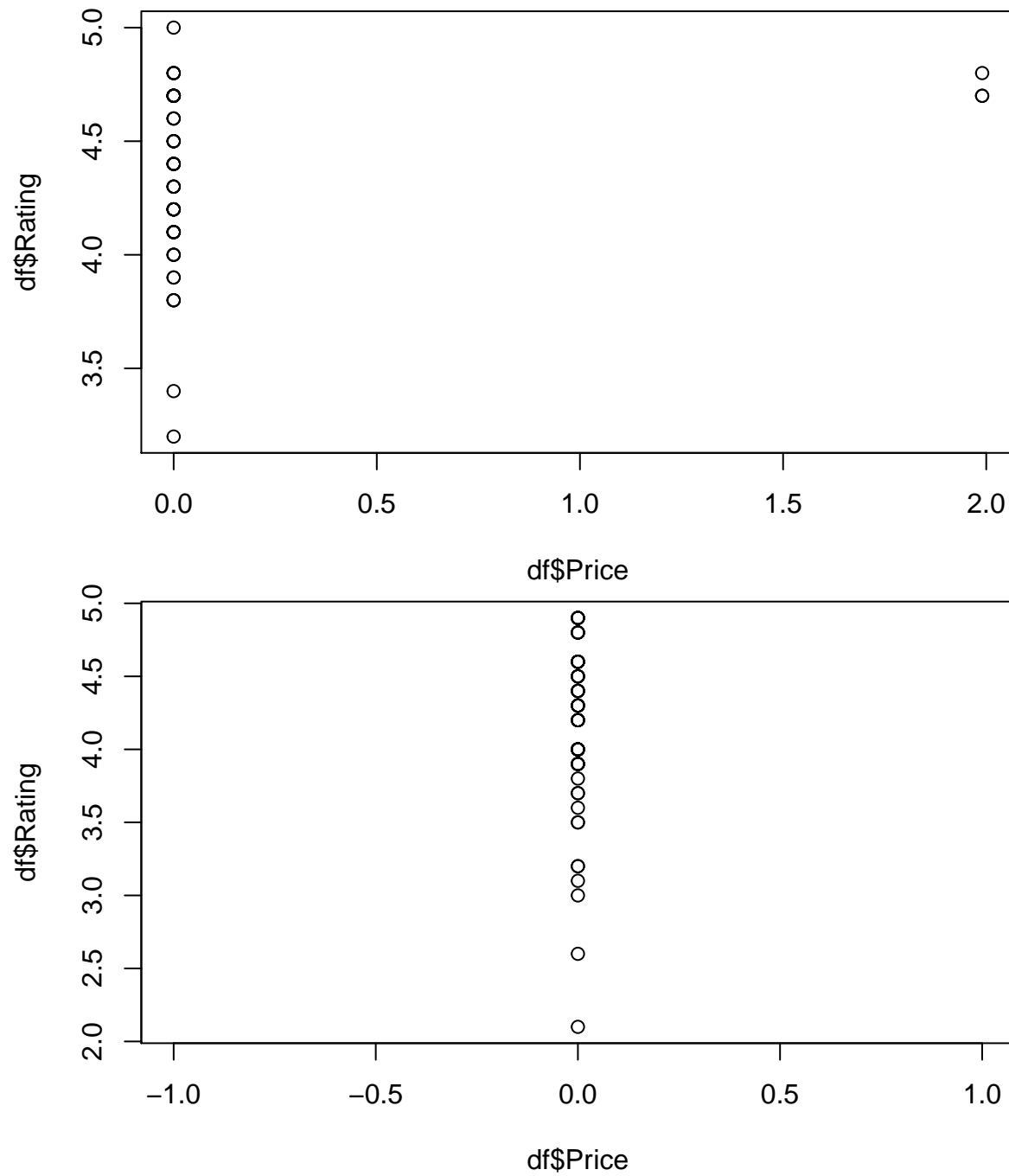
Problem 1.3

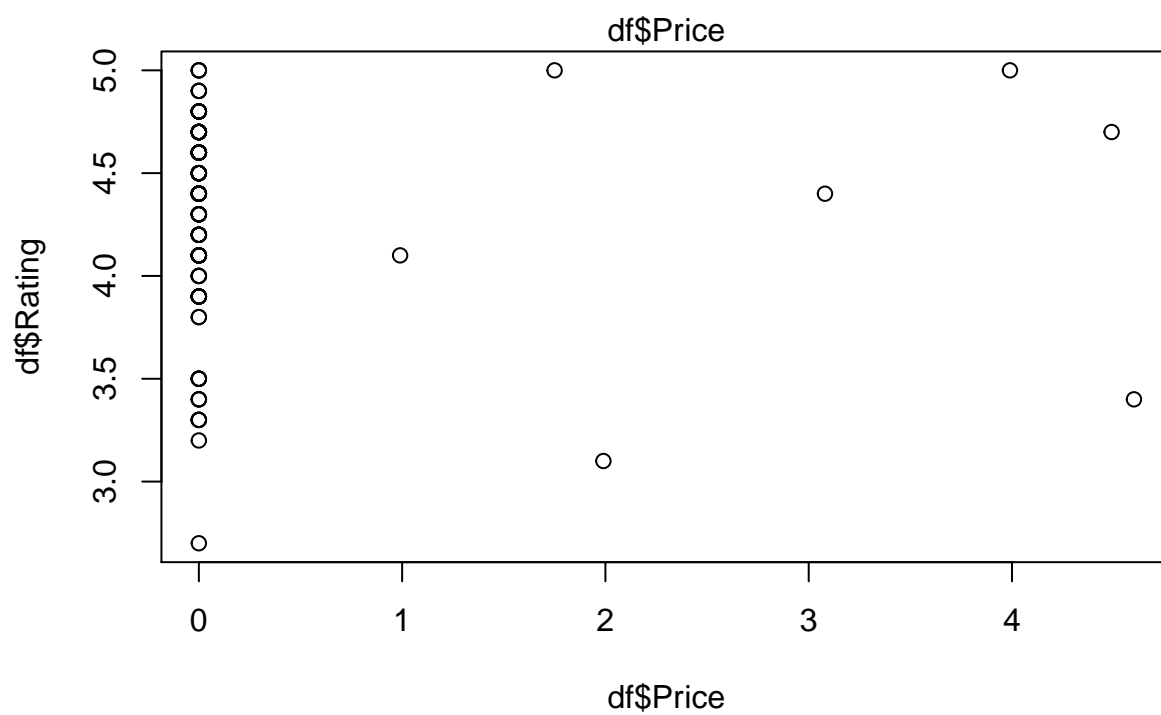
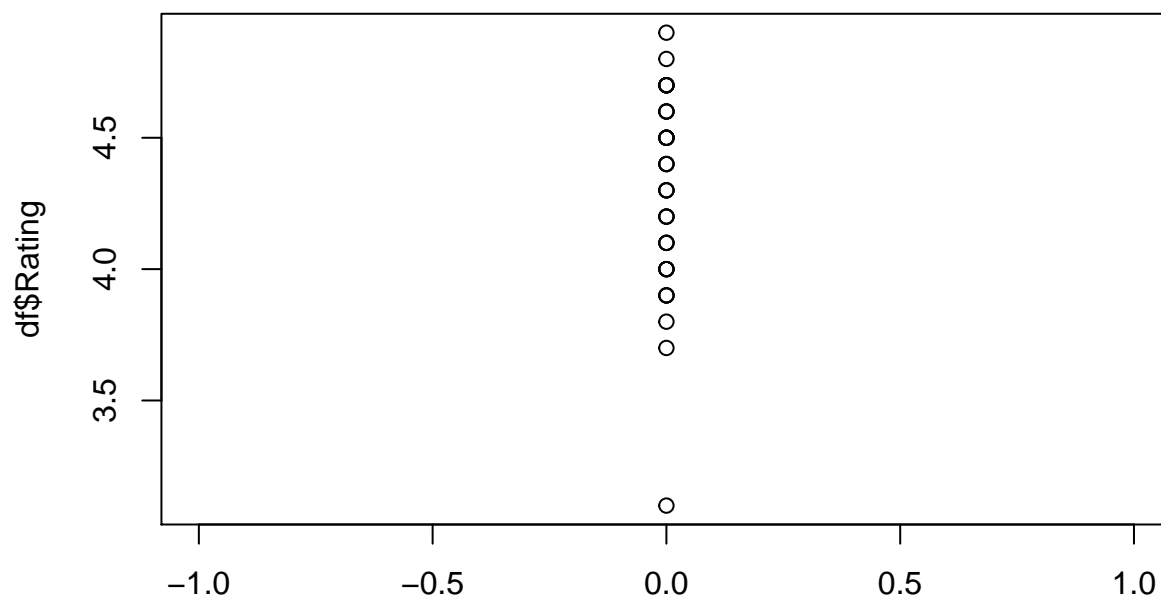
We now investigate how the overall rating (**Rating**) is associated with the category of the app (**Category**) and its price (**Price**). Use **Split/Apply/Combine** strategy to split the dataset by **Category**. For each category, generate a plot of user's rating (**Rating**) against the app's price (**Price**). Display all plots in one figure. To receive full credit, you must use a vectorized function from the **plyr** family. Make sure your figure contains **33** subplots, with each of them corresponding to one category. [5 pts]

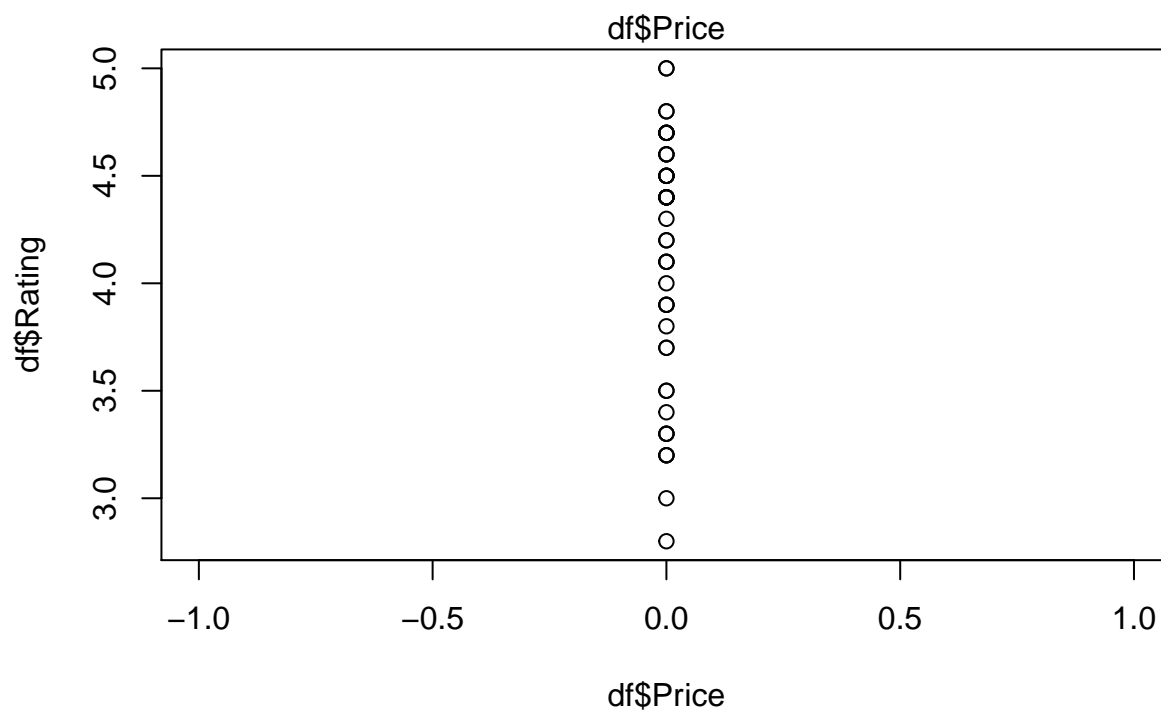
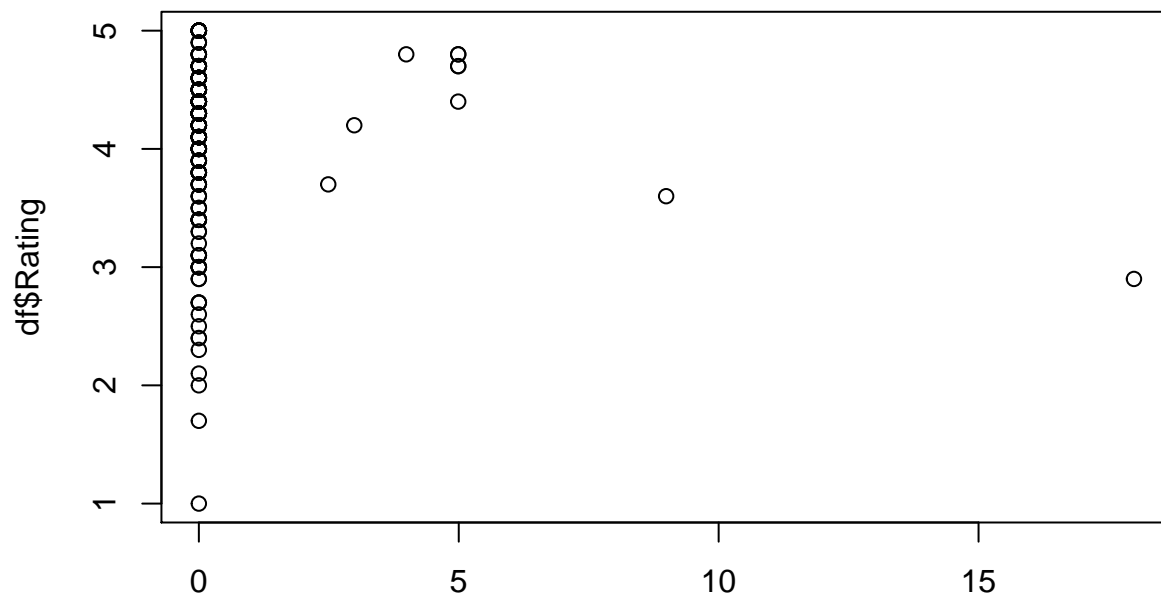
```

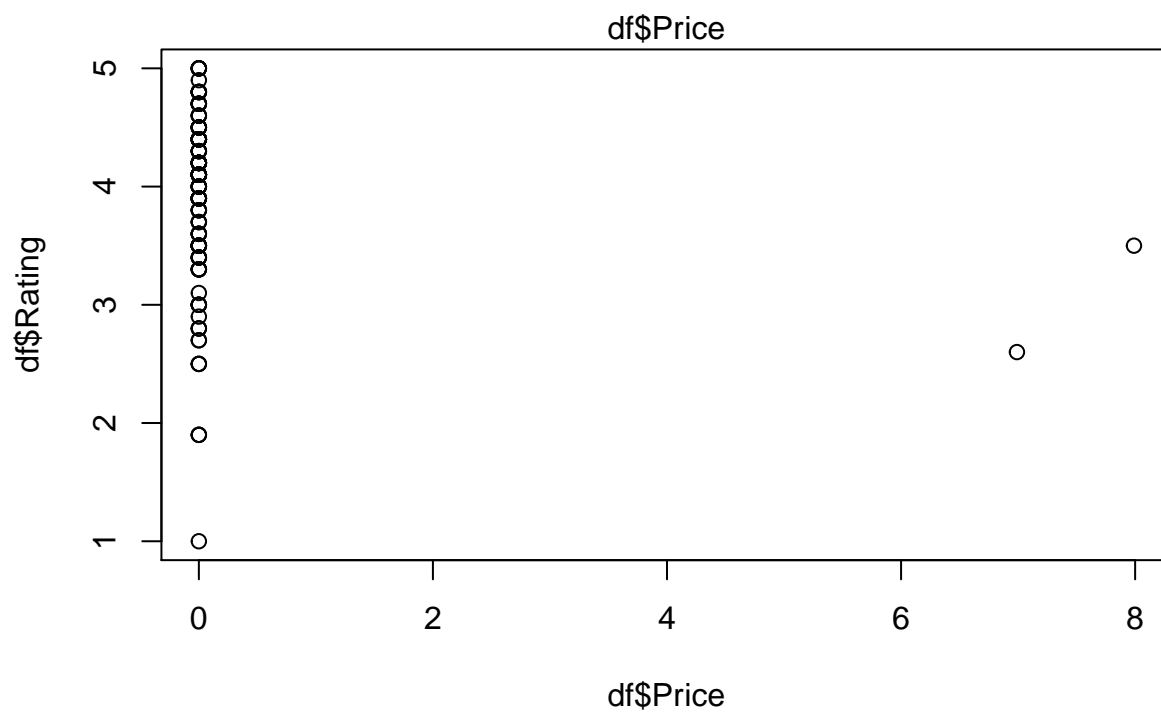
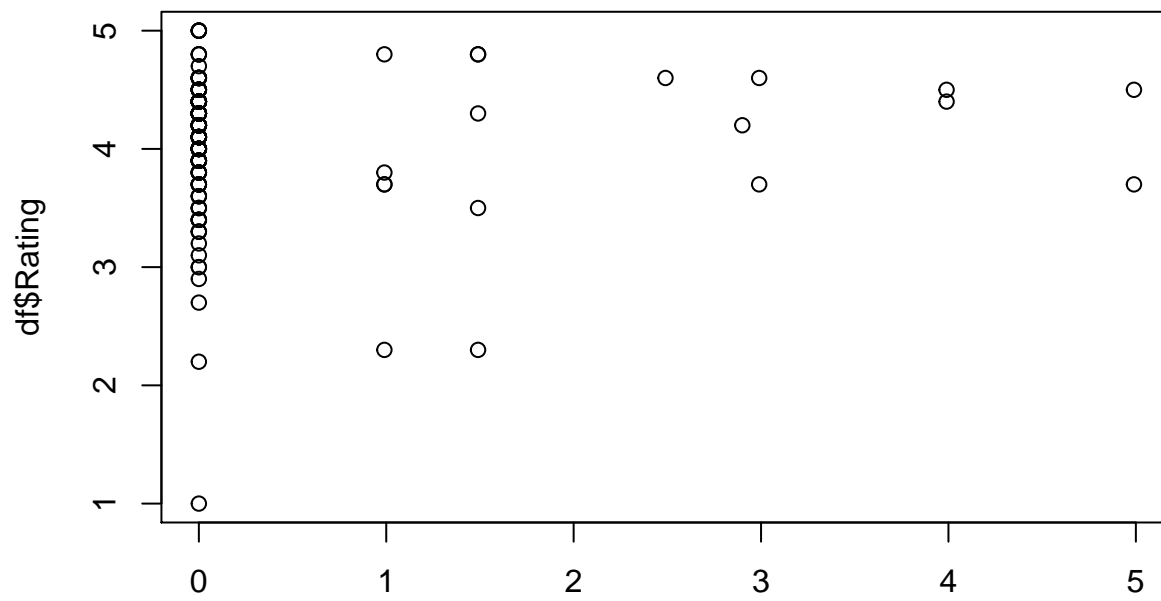
# code goes here
library(plyr)
rat_plot <- function(df) {
  plot(df$Price, df$Rating)
}
#par(mfrow = c(3, 11))
d_ply(apps, .(Category), rat_plot)

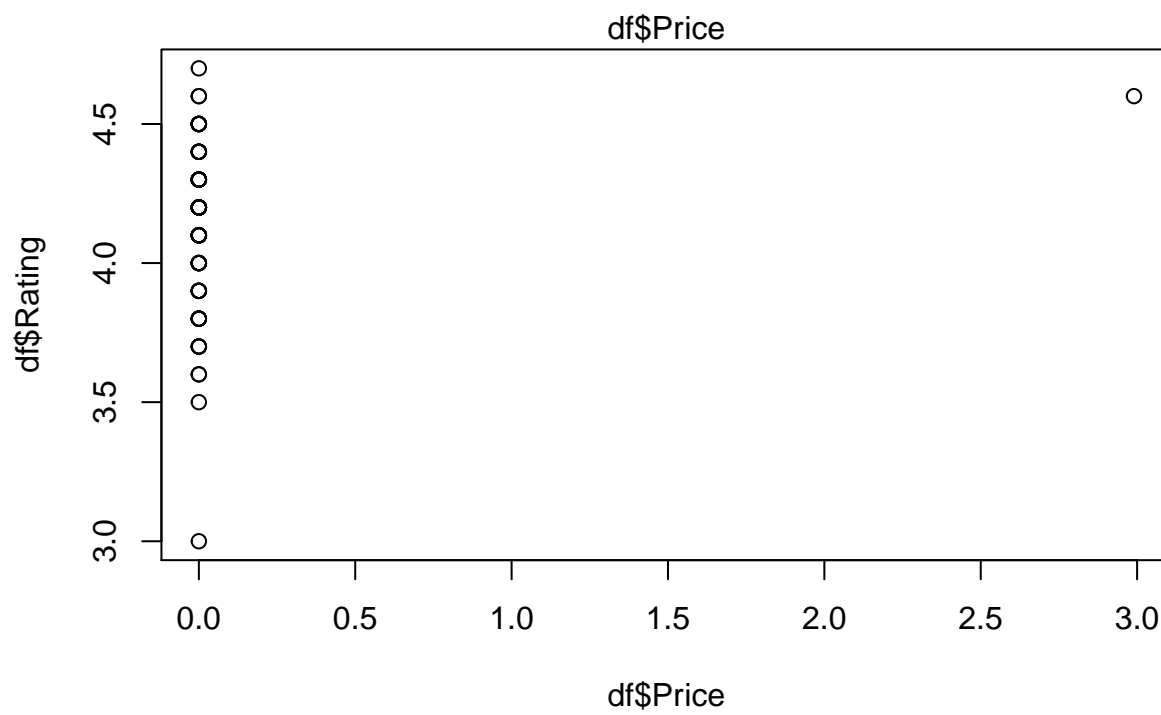
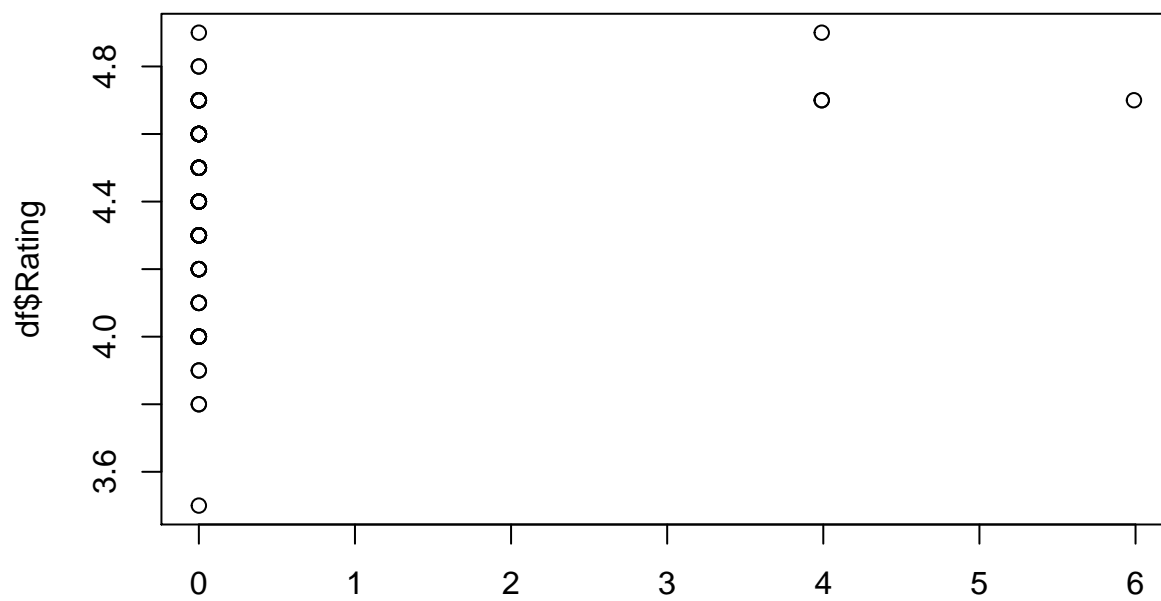
```

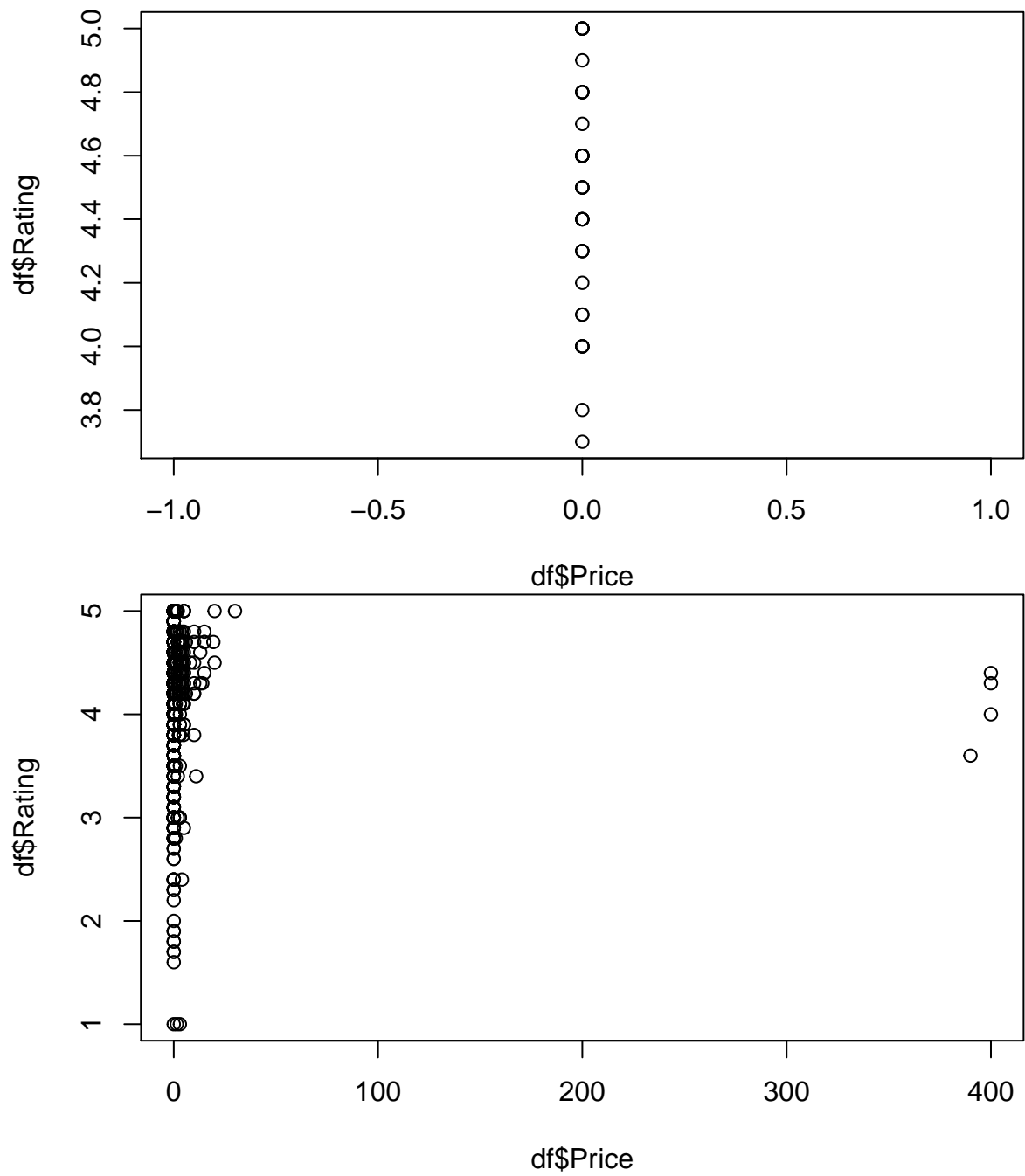


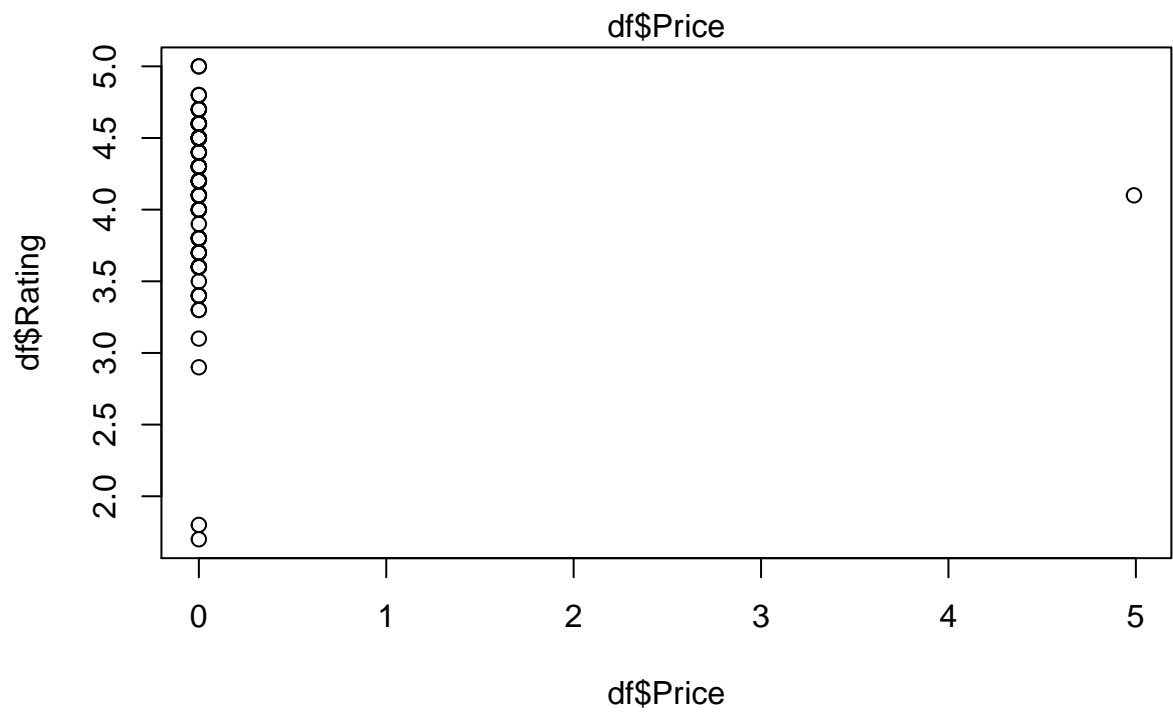
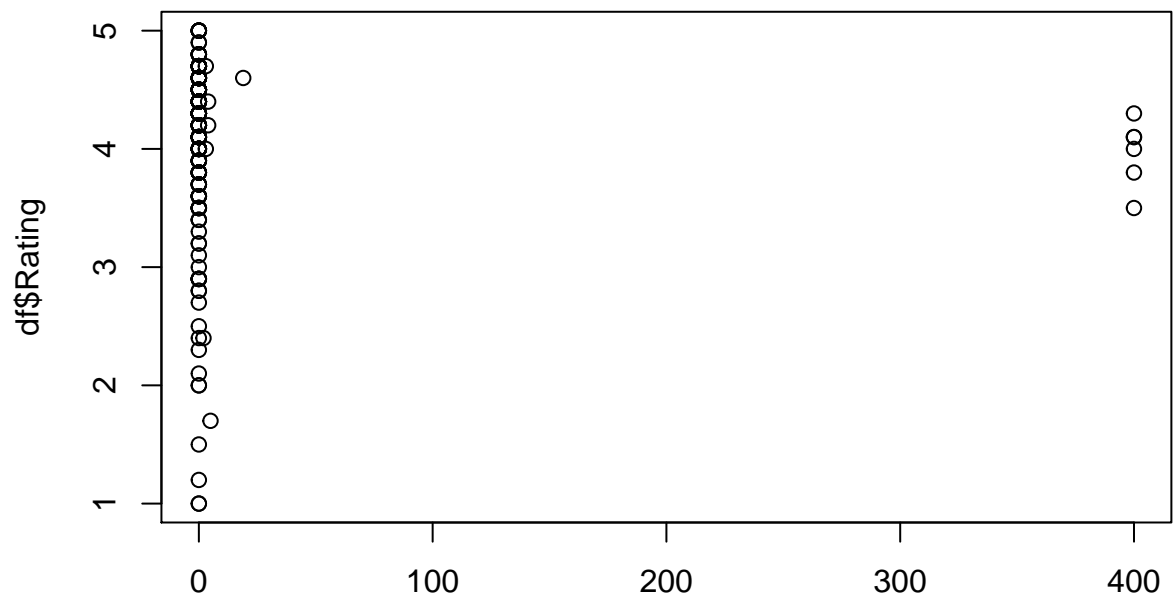


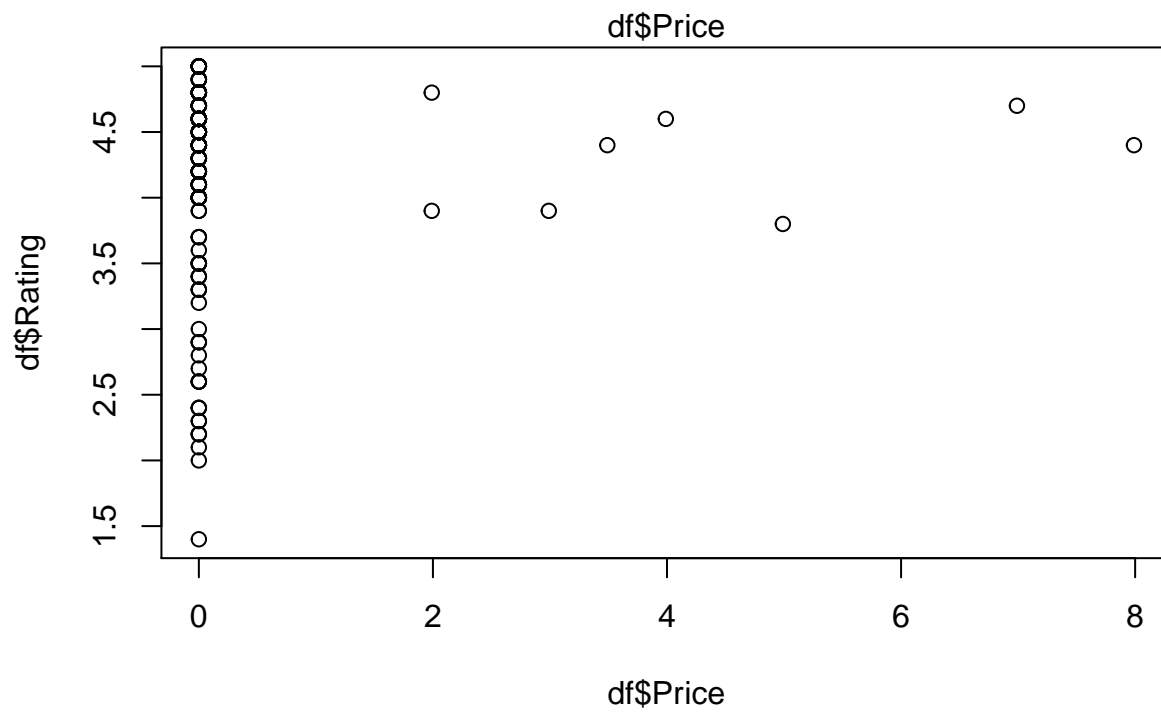
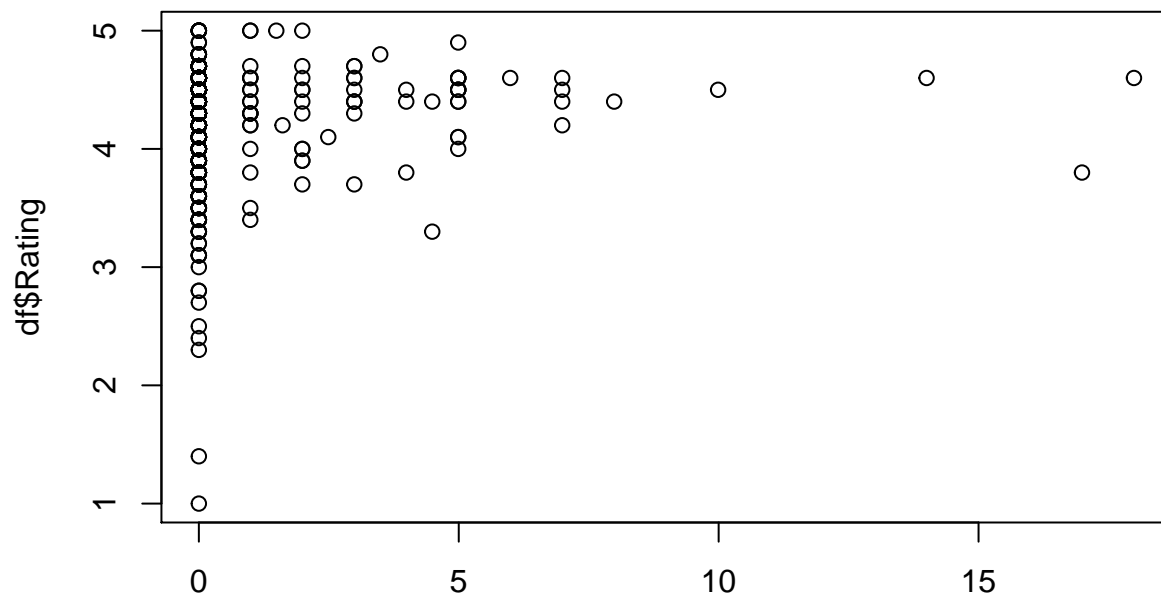


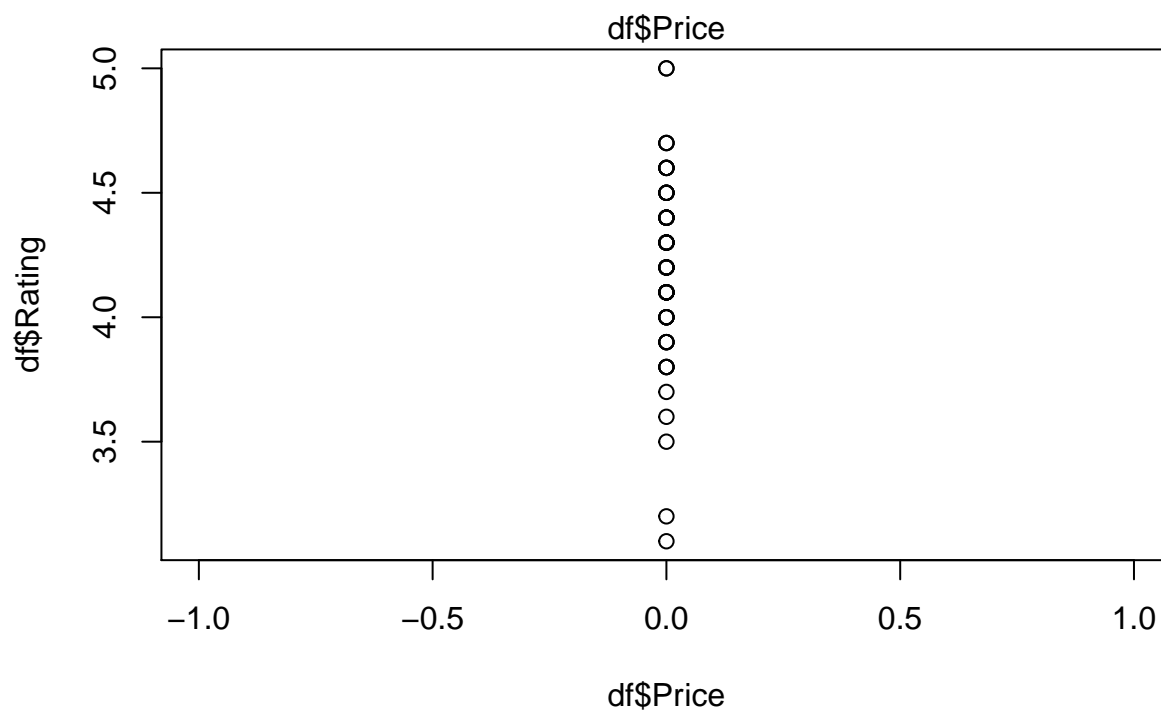
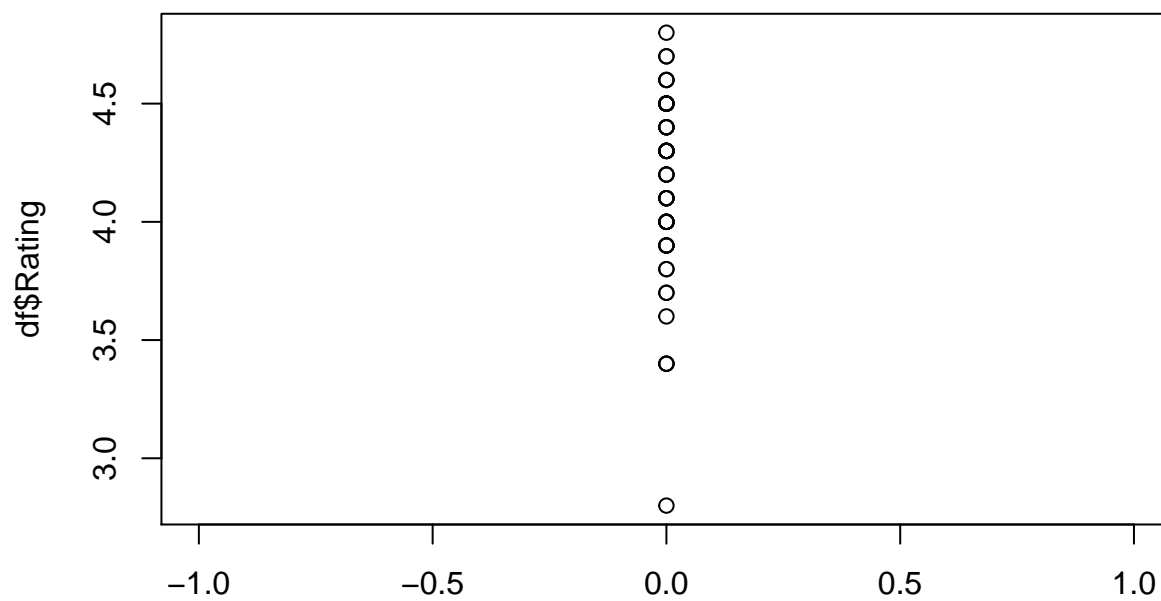


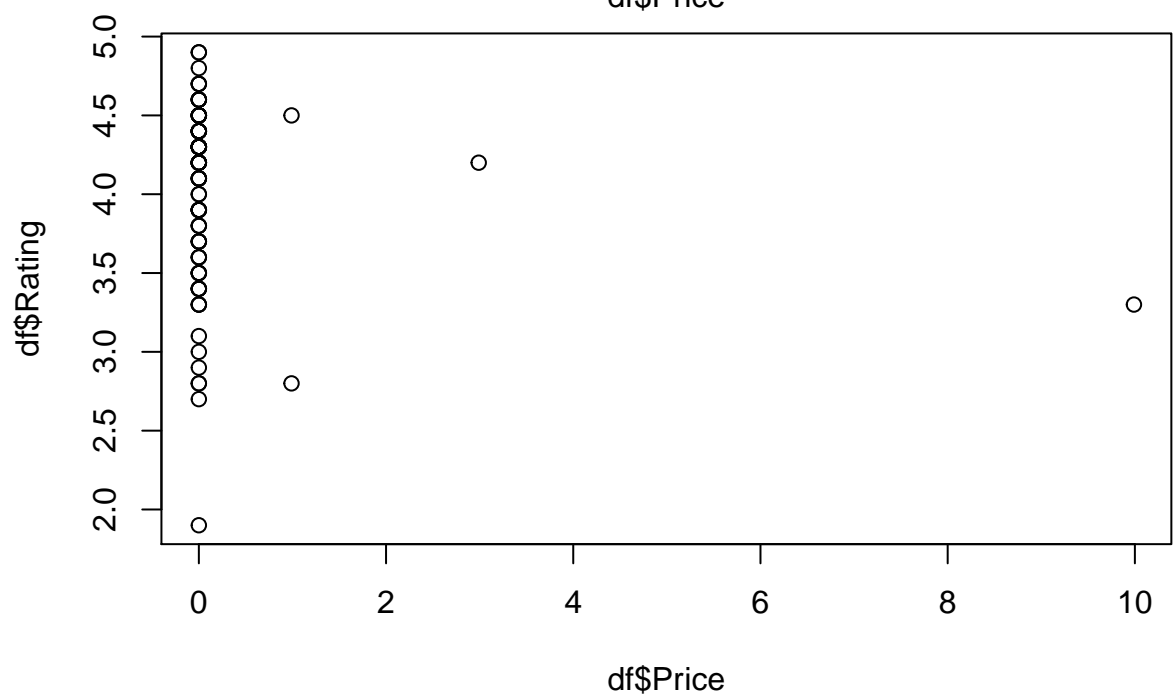
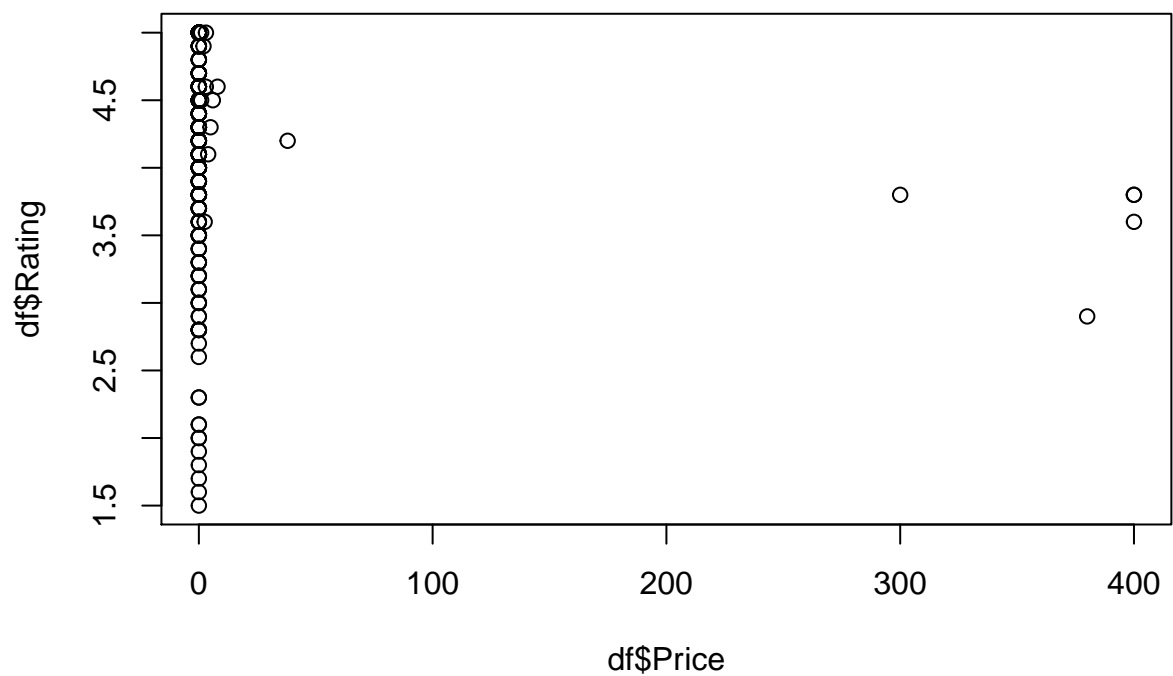


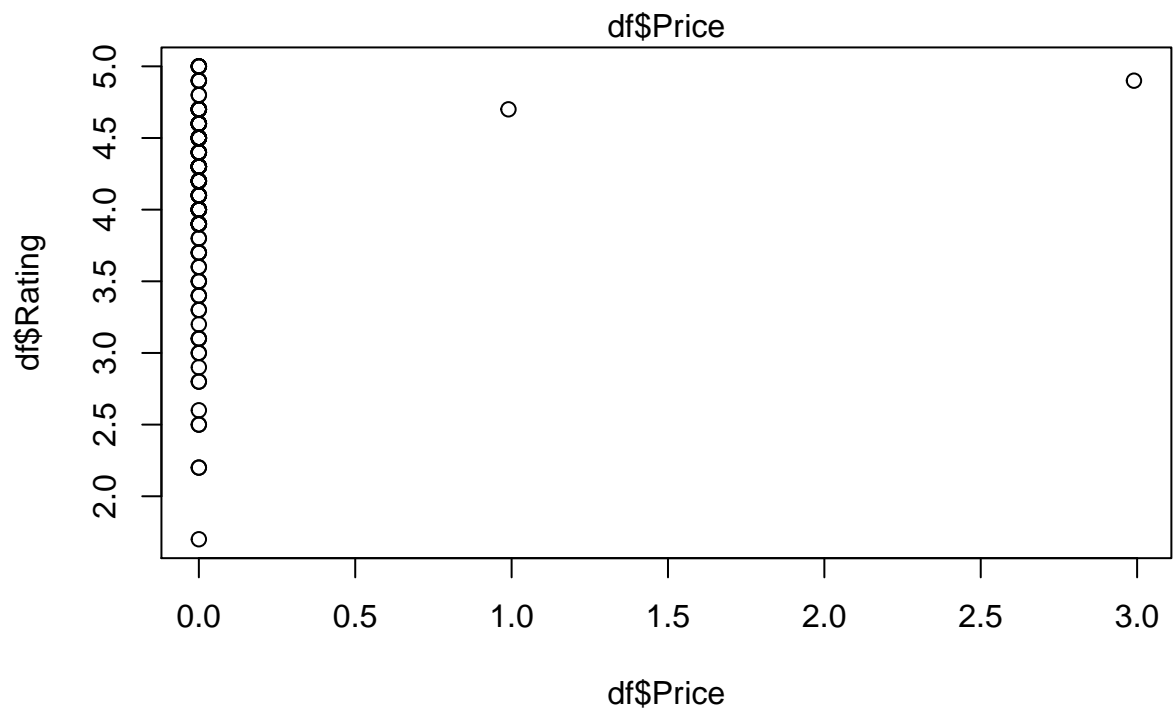
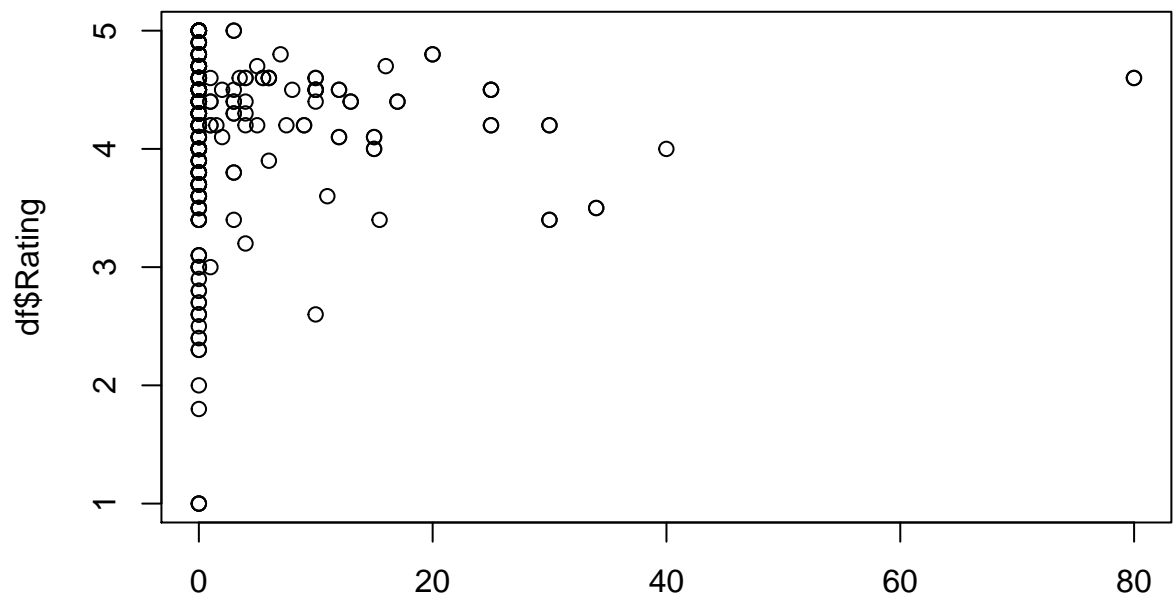


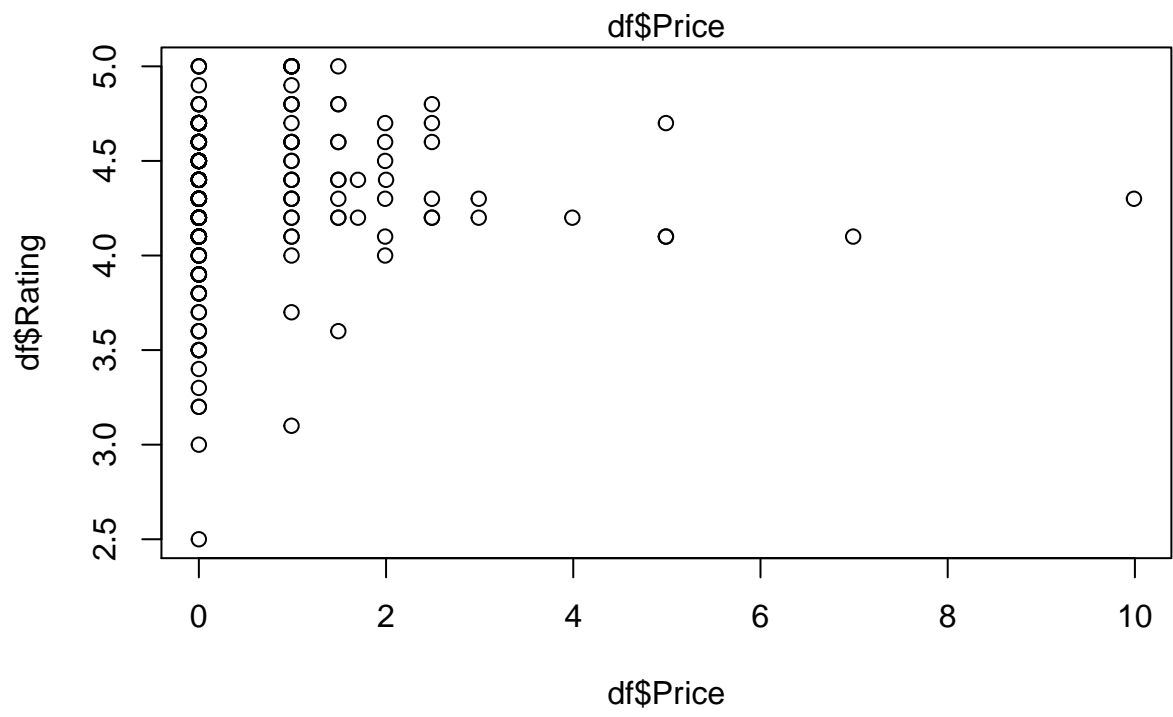
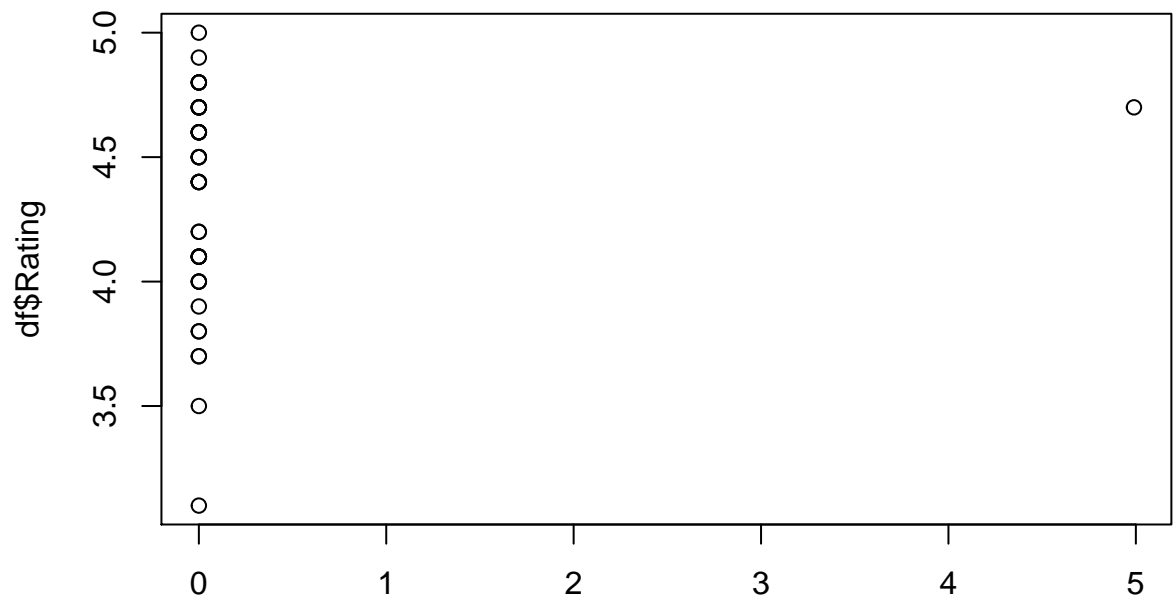


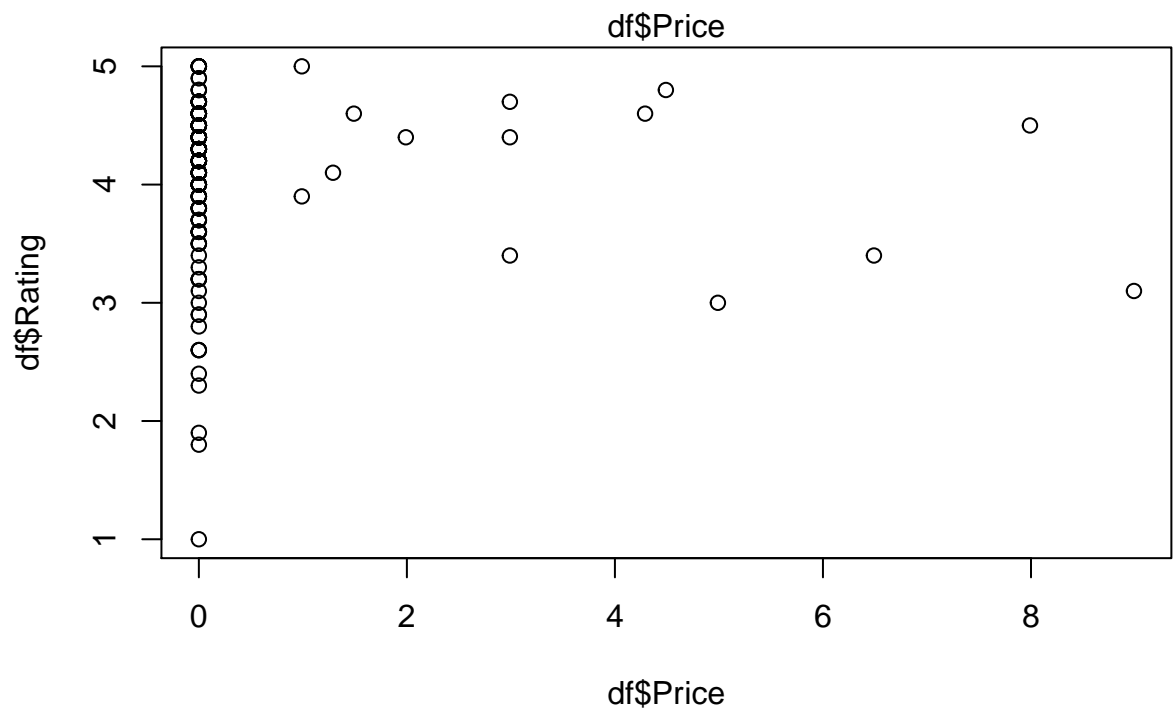
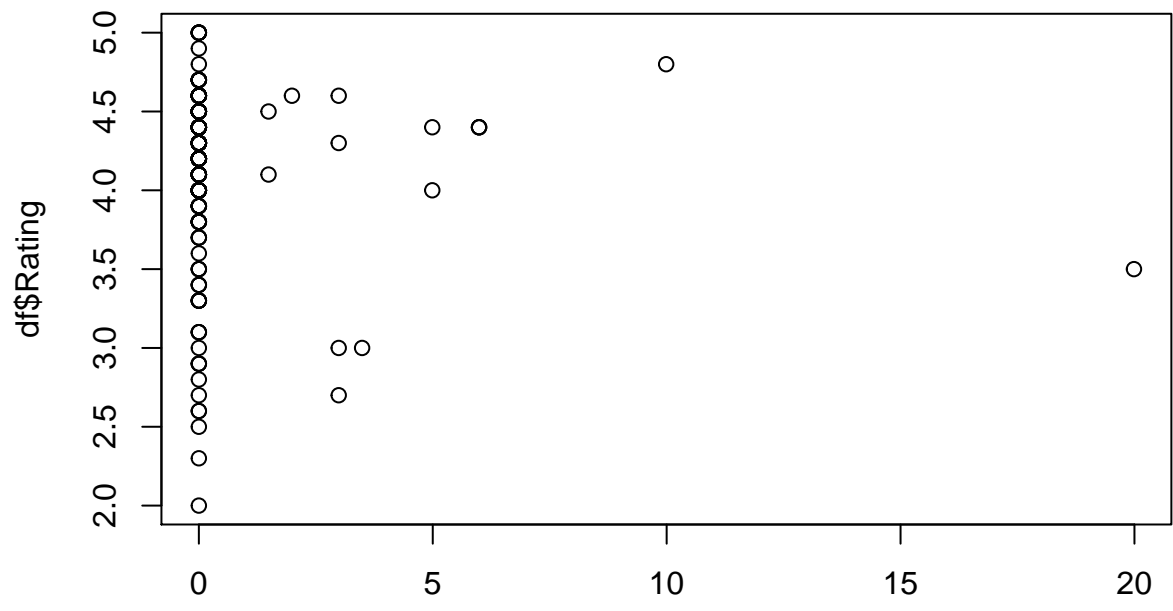


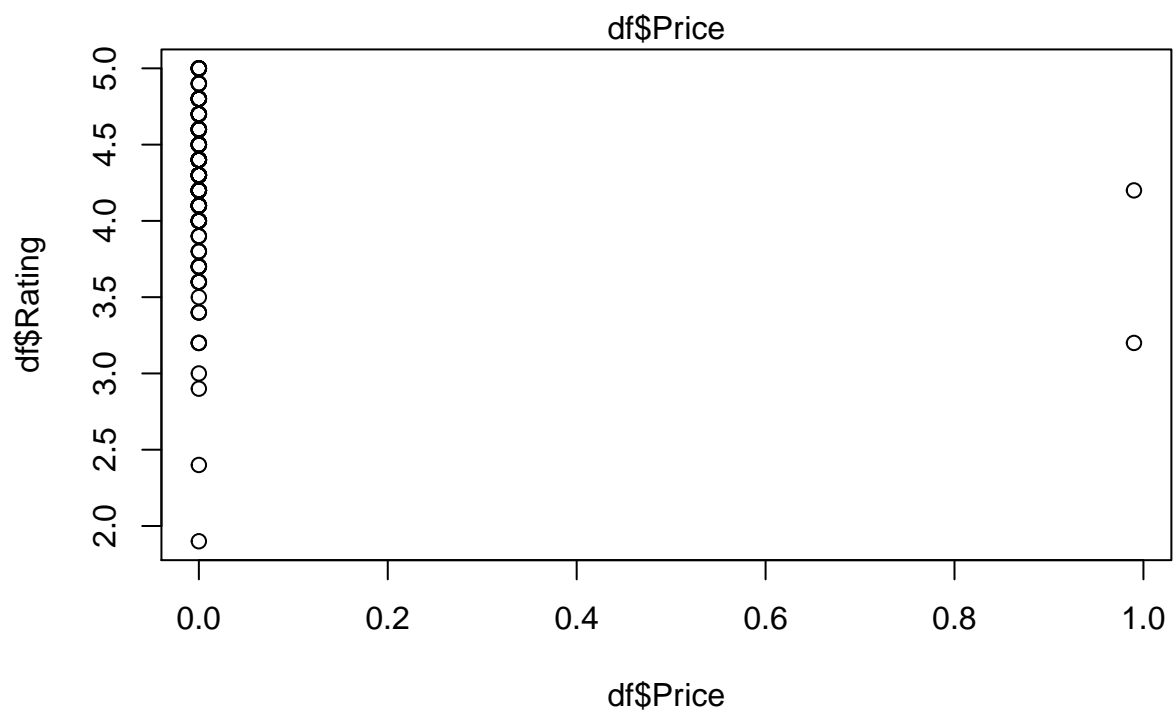
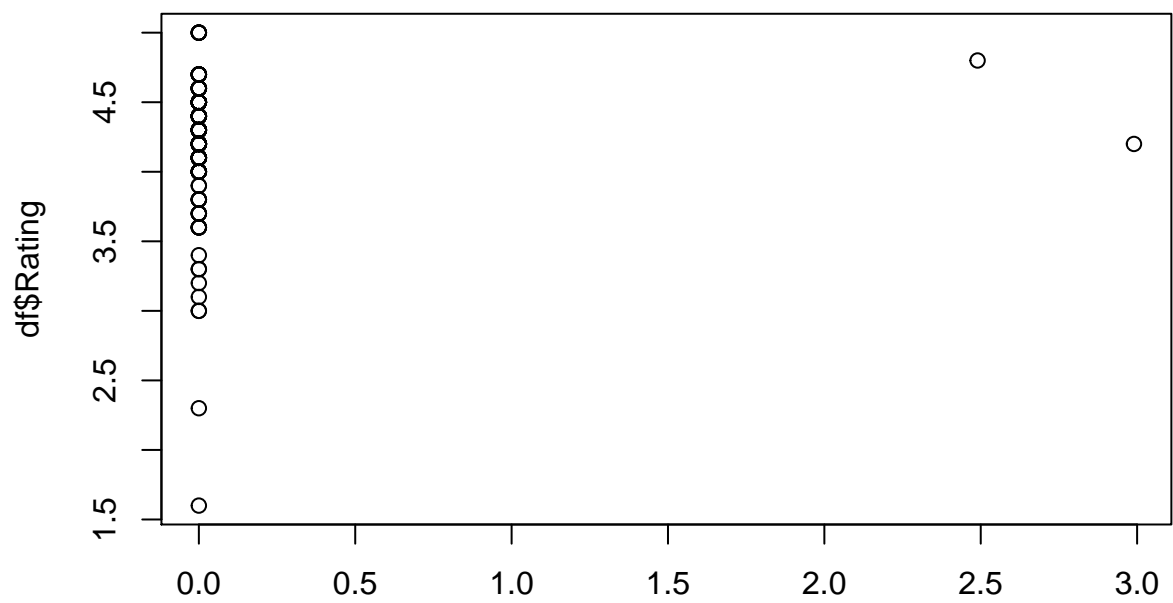


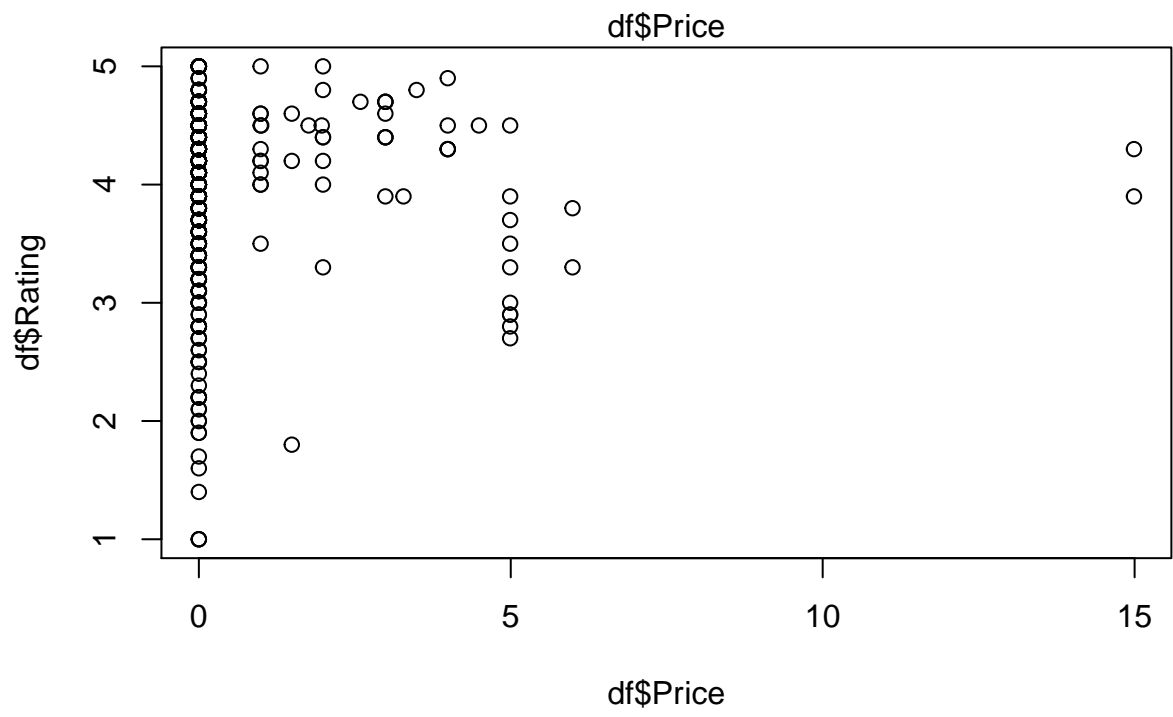
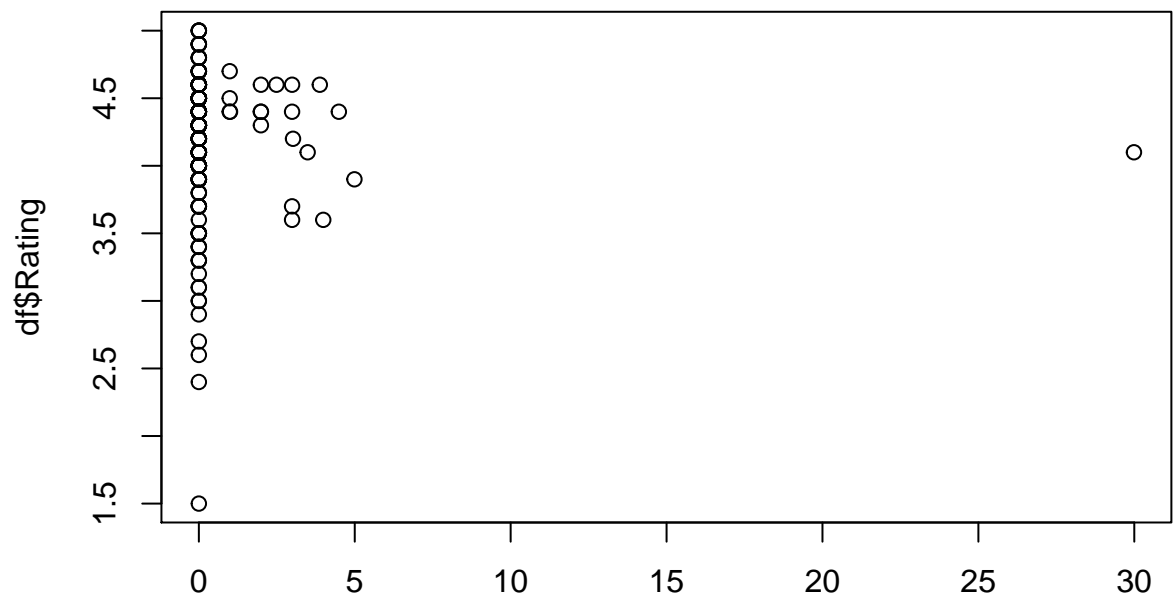


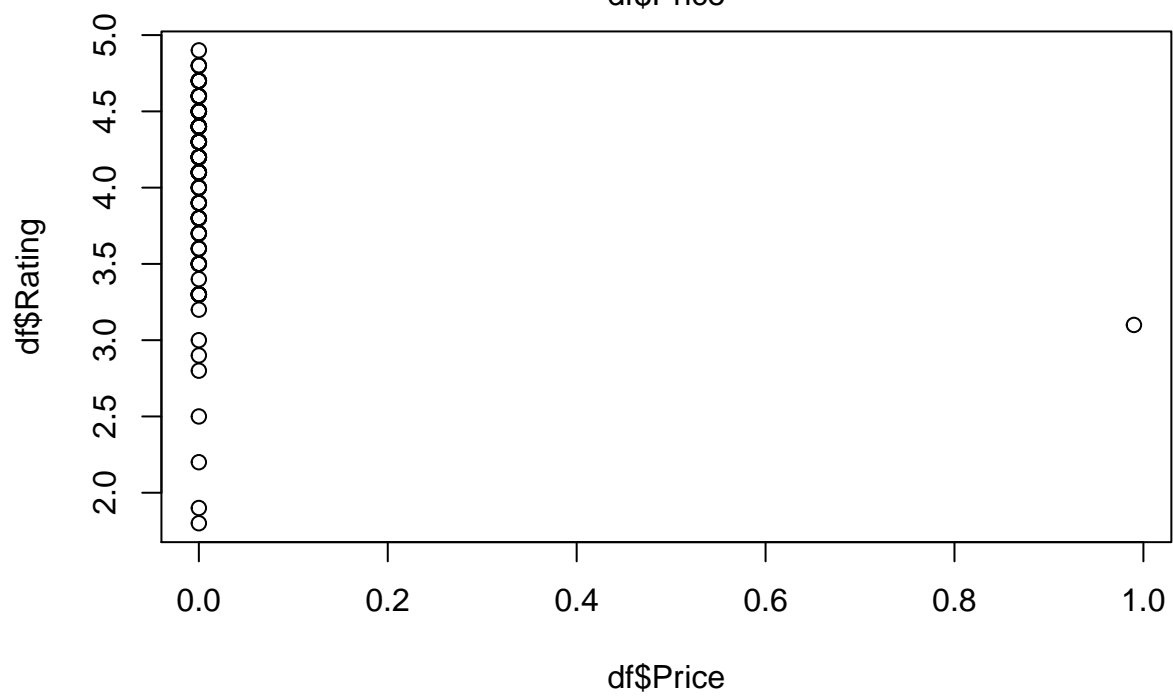
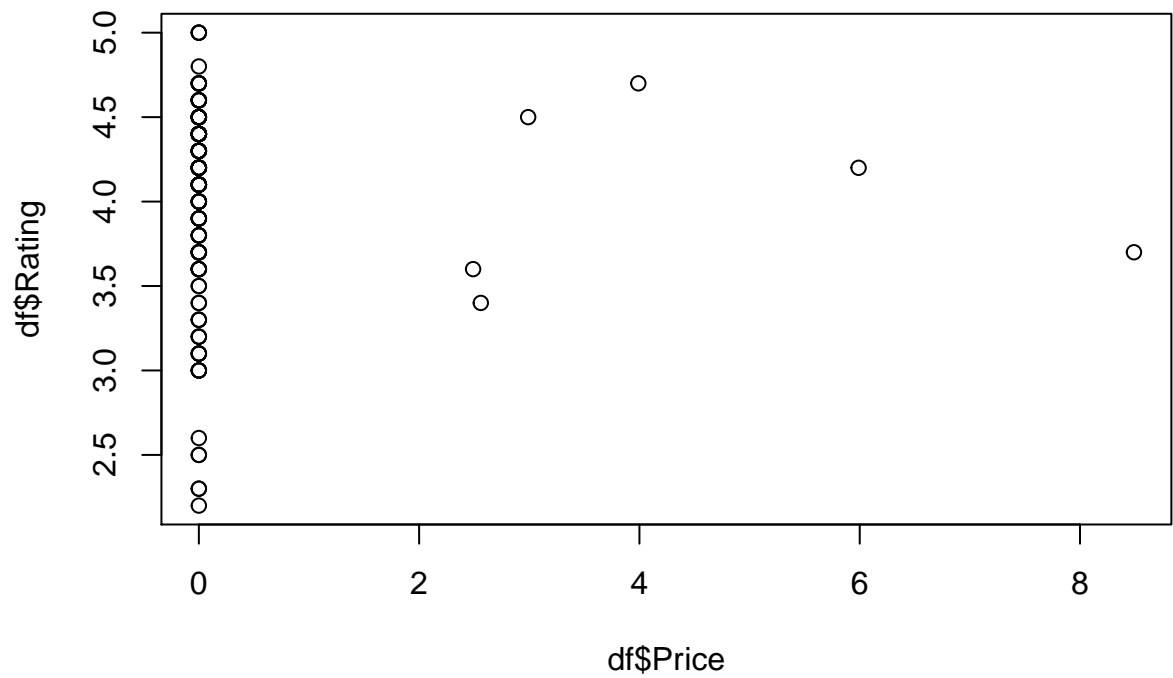


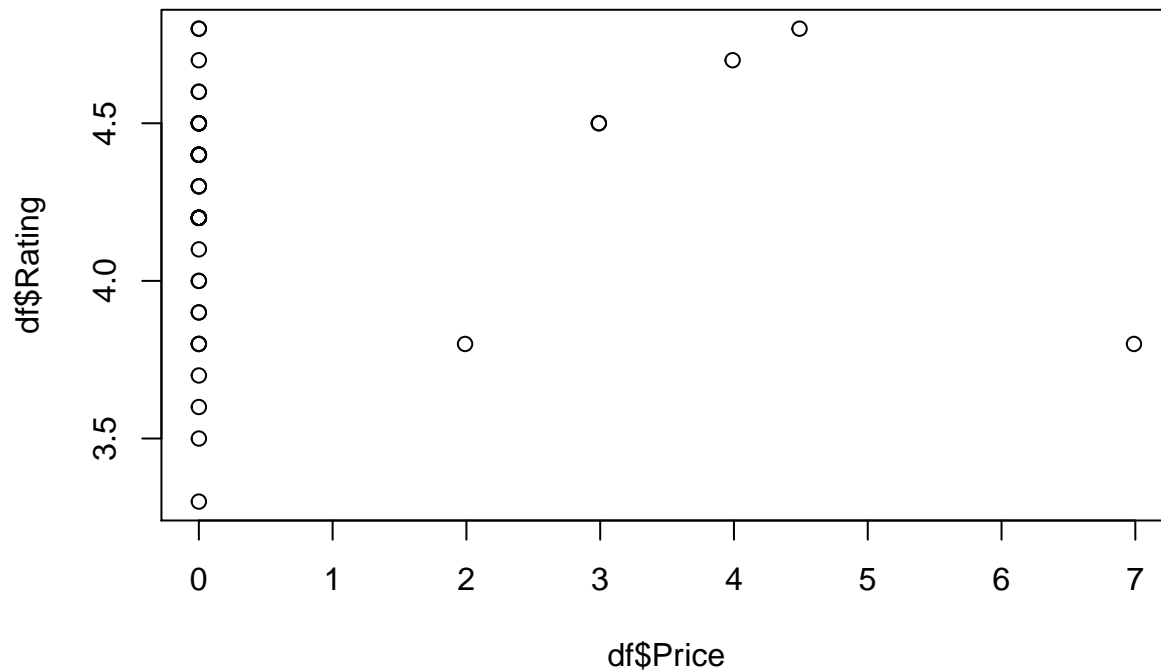












Part 2 (Basic Web Scraping, 15 + 2 pts)

In this part, we look at the voting record of the 2018 US Congress for roll call 274. The votes were compiled from <http://clerk.house.gov>. The raw data have been saved in the file `Roll_Call_274.xml`. We want to extract the voting results for 427 members of the House of Representatives. First, we read in data.

```
rollCall1274<-readLines("Roll_Call_274.xml")
```

Problem 2.0

Check the number of lines contained in the `Roll_call_274.xml` file. There should be 485 lines. [1 pt]

```
# code goes here
length(rollCall1274)
```

```
## [1] 485
```

Problem 2.1

Use the `grep()` function to find the lines in the file that correspond to the votes. Make sure `grep()` finds 427 lines. Hint: such a line starts with `<recorded-vote>`. [2 pts]

```
# code goes here
pat <- '<recorded-vote>'
length(grep(pat, rollCall1274))
```

```
## [1] 427
```

```
lines <- rollCall1274[grep(pat, rollCall1274)]
```


Problem 2.2

Write a regular expression that will capture the ID of a member. Using it extract the ID of each member. Hint: you can use the fact that `name-id=` appears before the ID. The ID is inside a pair of quotes, and it consists of one capital letter and six digits. [2 pts]

```
# code goes here
pat1 <- "name-id=\"[A-Z][0-9]{6}\""
id <- regmatches(lines, regex(pat1, lines))
id <- substring(id, 10, nchar(id)-1)
```

Problem 2.3

Using a regular expression extract the name of each member. Make sure that you can extract all names for 427 members. [2 extra pts]

```
# code goes here
pat2 <- "unaccented-name=\".+\" party"
name <- regmatches(lines, regex(pat2, lines))
name <- substring(name, 18, nchar(name)-7)
```

Problem 2.4

Extract the party of each member by using a regular expression. There should be 193 Democrats and 234 Republicans. [2 pts]

```
# code goes here
pat3 <- "party=\"[A-Z]\""
party <- regmatches(lines, regex(pat3, lines))
party <- substring(party, 8, nchar(party)-1)
```

Problem 2.5

Extract the state for each member by using a regular expression. [2 pts]

```
# code goes here
pat4 <- "state=\"[A-Z]+\""
state <- regmatches(lines, regex(pat4, lines))
state <- substring(state, 8, nchar(state)-1)
```

Problem 2.6

Last, use a regular expression to extract the vote. [2 pts]

```
# code goes here
pat5 <- "<vote>([A-z]+)</vote>"
vote <- regmatches(lines, regex(pat5, lines))
vote <- substring(vote, 7, nchar(vote)-7)
```

Problem 2.7

Make the extracted vote as a factor, and check its levels. Make a new variable called `numeric.vote`, which takes value 1 if the member voted “Yes (Aye)”, 0 if the vote is “No”, and -1 for “Not Voting”. [2 pts]

```
# code goes here
# vote <- factor(vote)
# levels(vote)
# numeric.vote <- ifelse(vote == 'Aye', 0, 1, -1)
```

Problem 2.8

Create a dataframe `rollCall274`, which contains the following five variables: `name`, `state`, `party`, `vote`, `numeric.vote`. Use `id` to name the rows of this dataframe. [2 pts]

```
# code goes here
# rollCall274 <- data.frame(name, state, party, vote, numeric.vote)
# row.names(rollCall274) <- id
```

Part 3 (Bootstrap, 4 + 2 pts)

We consider the `strikes` data which we used in Lecture 6. The data set is about strikes in 18 countries over 35 years (compiled by Bruce Western, in the Sociology Department at Harvard University). The measured variables are:

- **country**, **year** – country and year of data collection
- **strike.volume** – days on strike per 1000 workers
- **unemployment** – unemployment rate
- **inflation** – inflation rate
- **left.parliament** – leftwing share of the government
- **centralization** – centralization of unions
- **density** – density of unions

In this problem, we *only* look at the strikes in Italy. On this subset, we run a simple linear regression using `strike.volume` as the response (Y) and `left.parliament` as the predictor (X). Our model is

$$Y = \beta_0 + \beta_1 X + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

```
strikes<-read.csv("strikes.csv", header = T)
italy.strikes<-strikes[strikes$country == "Italy", ]
dim(italy.strikes)
```

```
## [1] 35  8
```

```
lm.fit<- lm(strike.volume ~ left.parliament, data = italy.strikes)
round(lm.fit$coefficients,3)
```

```
##      (Intercept) left.parliament
##      -738.745      40.291
```

Problem 3.1

Denote our estimate of β_1 as $\hat{\beta}_1$, which is 40.291 according to the analysis above. Use the Bootstrap method to estimate the variance of $\hat{\beta}_1$. Here, you may draw 100 Bootstrap samples. [4 pts]

```
# code goes here
n <- nrow(italy.strikes)
B <- 100
resampled_values <- matrix(NA, nrow = B, ncol = n)
for (b in 1:B) {
  resampled_values[b, ] <- sample(1:n, n, replace = TRUE)
}

resampled_ests <- matrix(NA, nrow = B, ncol = 2)
colnames(resampled_ests) <- c("Intercept_Est", "Slope_Est")

for (b in 1:B) {
  resampled_ests[b, ] <- coef(lm(strike.volume ~ left.parliament, italy.strikes[resampled_values[b, ], ]))
}

var(resampled_ests[, "Slope_Est"])

## [1] 563.8825
```

Problem 3.2

Construct a 95% confidence interval of β_1 based on the result in part 3.1. [2 extra pts]

```
# code goes here
Cl_1 <- quantile(resampled_ests[, "Slope_Est"], 0.025)
Cu_1 <- quantile(resampled_ests[, "Slope_Est"], 0.975)
int_1 <- c(Cl_1, Cu_1)
int_1

##      2.5%      97.5%
## -9.248921 73.009120
```