PS1

Rui

10/11/2019

1. Data set for the Google Play Store is obtained through Kaggle https://www.kaggle.com/lava18/google-play-store-apps

2. A bar plot is used to compare the mean of rating for game and non game apps. And a histogram is used to look at the the distribution of number of reviews received by

```
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts -------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(skimr)
```

```
## 
## Attaching package: 'skimr'
```

```
## The following object is masked from 'package:stats':
## 
##     filter
```

```
# I want to know if gaming apps are popular in the GOOGLE play store and
# review their number of installs

#Load the google play store data
appdata <- read.csv(file="/Users/ruihe/Documents/UML/PS1/googleplaystore.csv")
summary(appdata)
```

```
##                                                  App       
##  ROBLOX                                            :    9  
##  CBS Sports App - Scores, News, Stats & Watch Live:    8  
##  8 Ball Pool                                       :    7  
##  Candy Crush Saga                                  :    7  
##  Duolingo: Learn Languages Free                    :    7  
##  ESPN                                              :    7  
##  (Other)                                           :10796  
##          Category        Rating          Reviews     
##  FAMILY      :1972   Min.   : 1.000   0      : 596  
##  GAME        :1144   1st Qu.: 4.000   1      : 272
##  TOOLS       : 843   Median : 4.300   2      : 214
##  MEDICAL     : 463   Mean   : 4.193   3      : 175
##  BUSINESS    : 460   3rd Qu.: 4.500   4      : 137
##  PRODUCTIVITY: 424   Max.   :19.000   5      : 108
##  (Other)     :5535   NA's   :1474     (Other):9339
##                 Size             Installs      Type           Price
##  Varies with device:1695   1,000,000+ :1579   0    :    1   0      :10040
##  11M               : 198   10,000,000+:1252   Free:10039   $0.99  :  148
##  12M               : 196   100,000+   :1169   NaN :    1   $2.99  :  129
##  14M               : 194   10,000+    :1054   Paid:  800   $1.99  :   73
##  13M               : 191   1,000+     : 907                $4.99  :   72
##  15M               : 184   5,000,000+ : 752                $3.99  :   63
##  (Other)           :8183   (Other)    :4128                (Other):  316
##         Content.Rating           Genres        Last.Updated
##                 :   1   Tools        : 842   3-Aug-18 : 326
##  Adults only 18+:   3   Entertainment: 623   2-Aug-18 : 304
##  Everyone       :8714   Education    : 549   31-Jul-18: 294
##  Everyone 10+   : 414   Medical      : 463   1-Aug-18 : 285
##  Mature 17+     : 499   Business     : 460   30-Jul-18: 211
##  Teen           :1208   Productivity : 424   25-Jul-18: 164
##  Unrated        :   2   (Other)      :7480   (Other)  :9257
##             Current.Ver                   Android.Ver
##  Varies with device:1459   4.1 and up        :2451
##  1                 : 842   4.0.3 and up      :1501
##  1.1               : 276   4.0 and up        :1375
##  1.2               : 185   Varies with device:1362
##  2                 : 165   4.4 and up        : 980
##  1.3               : 145   2.3 and up        : 652
##  (Other)           :7769   (Other)           :2520
```

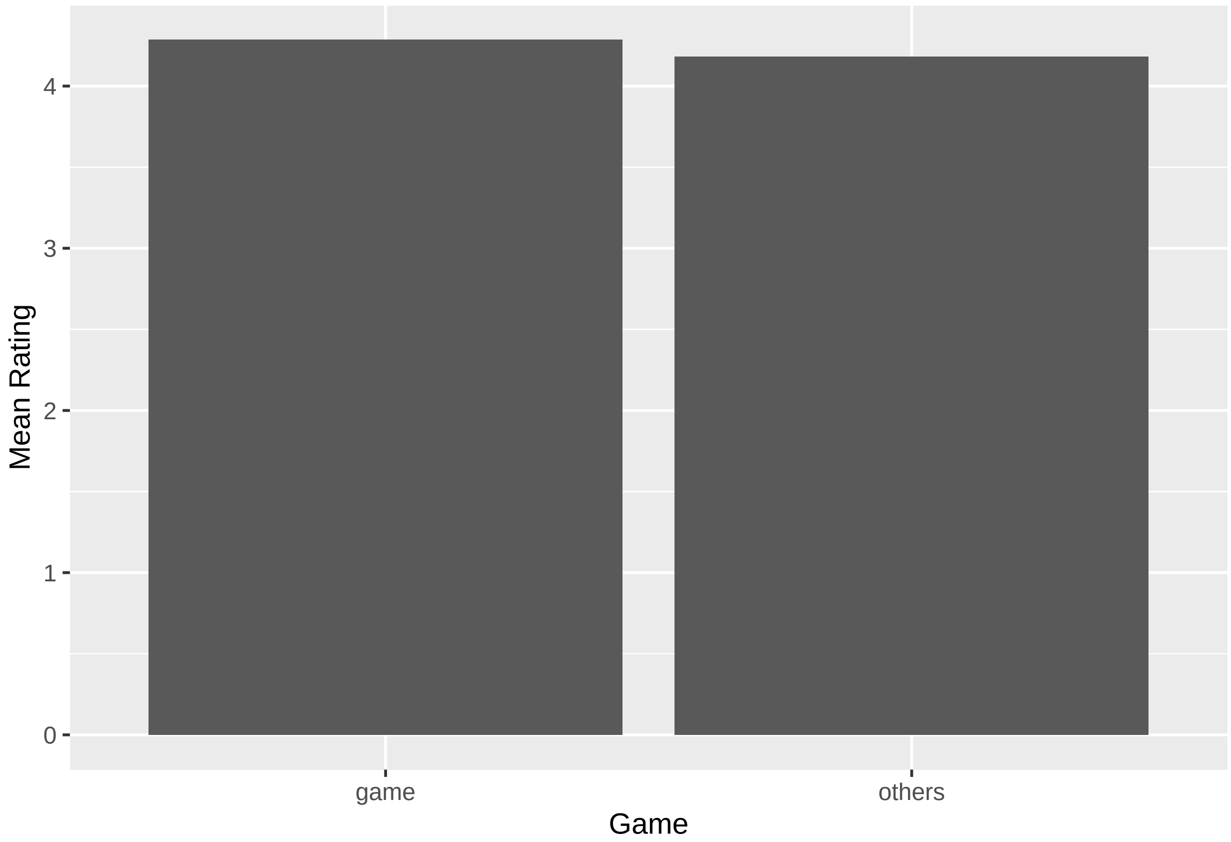
```
appdata_sub <-appdata %>%
  dplyr::select(App, Rating,Installs, Reviews, Category) %>%
  mutate(game = ifelse(Category=="GAME","game","others"),
         Reviews = as.numeric(Reviews)) %>%
  drop_na()
  #mutate(category = case_when(Category == "GAME" ~ "Game",
  #                            Category == "Dating" ~"Dating",
  #                            Category != ("GAME"|"Dating"), ~"Others")) %>%


summary(appdata_sub)
```

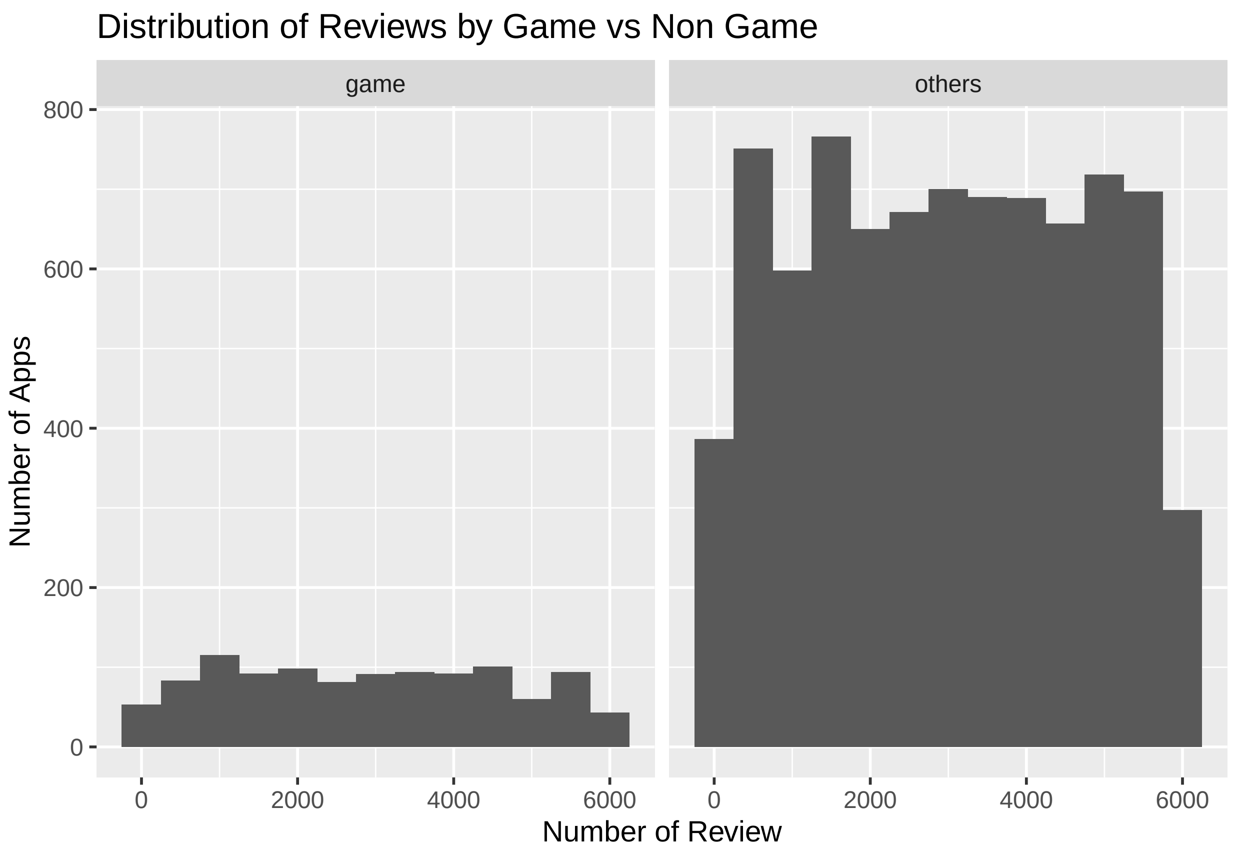
```
##                                                 App           Rating
##  ROBLOX                                        :   9   Min.   : 1.000
##  CBS Sports App - Scores, News, Stats & Watch Live:   8   1st Qu.: 4.000
##  8 Ball Pool                                   :   7   Median : 4.300
##  Candy Crush Saga                              :   7   Mean   : 4.193
##  Duolingo: Learn Languages Free                :   7   3rd Qu.: 4.500
##  ESPN                                          :   7   Max.   :19.000
##  (Other)                                       :9322
##        Installs        Reviews                 Category        game
##  1,000,000+ :1577   Min.   :    2   FAMILY       :1747   Length:9367
##  10,000,000+:1252   1st Qu.:1484   GAME         :1097   Class :character
##  100,000+    :1150   Median :2937   TOOLS         : 734   Mode  :character
##  10,000+     :1010   Mean   :2967   PRODUCTIVITY : 351
##  5,000,000+ : 752   3rd Qu.:4476   MEDICAL       : 350
##  1,000+      : 713   Max.   :6002   COMMUNICATION: 328
##  (Other)     :2913                  (Other)       :4760
```

3. Two graphs showed that the rating and distribution reviews for both game and non game apps are very similar, suggesting gaming apps are not more popular than non-game apps.

```
#bar graph of mean ratings for game and non game apps
appdata_sub %>%
  group_by(game) %>%
  summarize(mean_rating = mean(Rating, na.rm = TRUE))%>%
  ggplot()+
  geom_bar(aes(x = game, y = mean_rating), stat = "identity") +
  labs(x = 'Game', y = "Mean Rating")
```



```
#number of reveiws given to game vs. non game apps
  ggplot(data = appdata_sub) +
    geom_histogram(aes(x=Reviews), binwidth = 500) +
    facet_wrap(~game, scales = "fixed", ncol = 2) +
    labs(x="Number of Review", y = "Number of Apps",
         title = "Distribution of Reviews by Game vs Non Game")
```



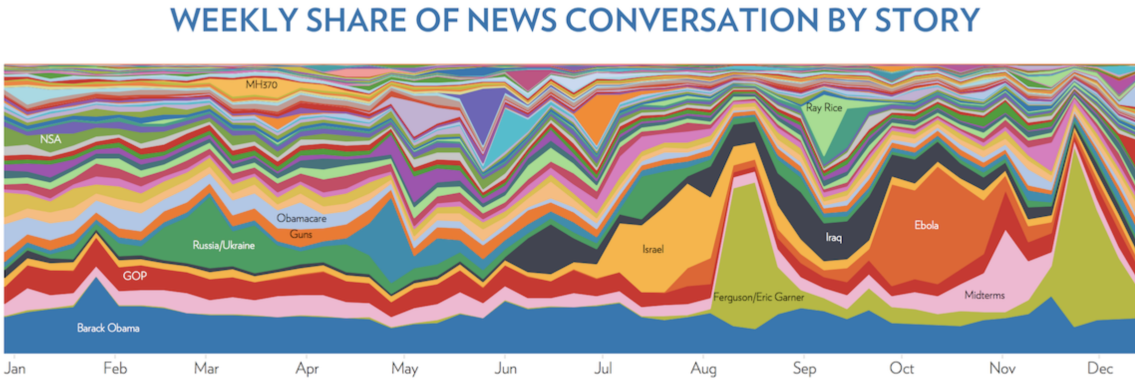
4. Mean, median and std of ratings for game/nongame apps

```
appdata_sub%>%
group_by(game)%>%
summarise(Mean=mean(Rating), Max=max(Rating), Min=min(Rating), Median=median(Rating), Std=sd(Rating))
```

```
## # A tibble: 2 x 6
##   game    Mean   Max   Min Median   Std
##   <chr>  <dbl> <dbl> <dbl>  <dbl> <dbl>
## 1 game    4.29     5     1    4.4 0.365
## 2 others  4.18    19     1    4.3 0.555
```

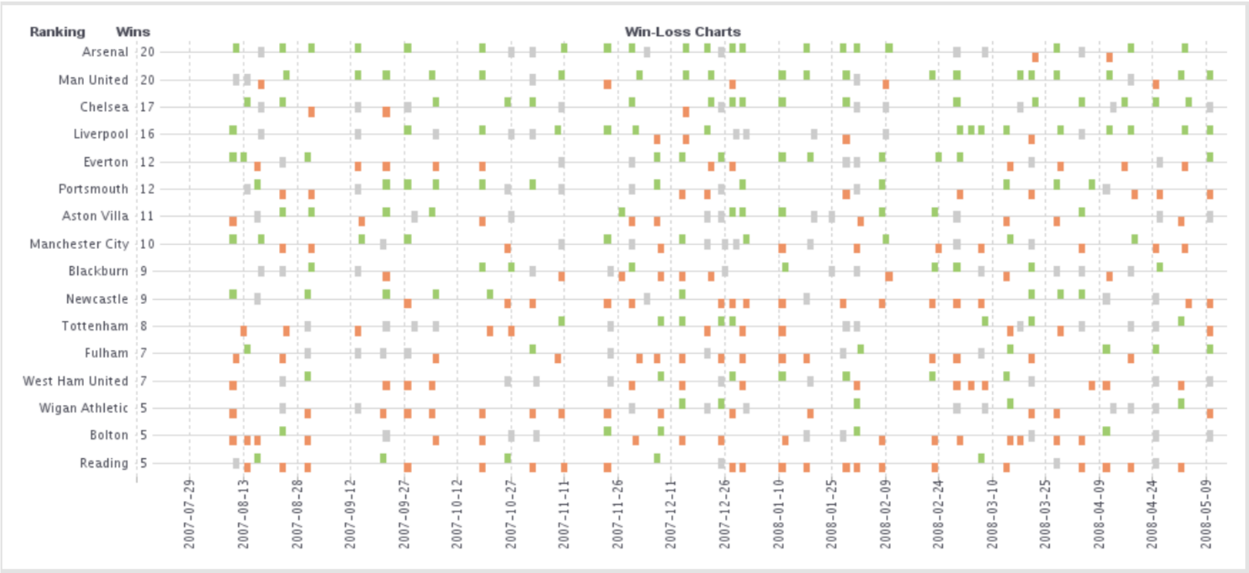
##Critical thinking

1. Visualization of data shows the qualitative trends and patterns in a more straight forward manners whereas numeric EDA shows a quanitative perspective of the data.

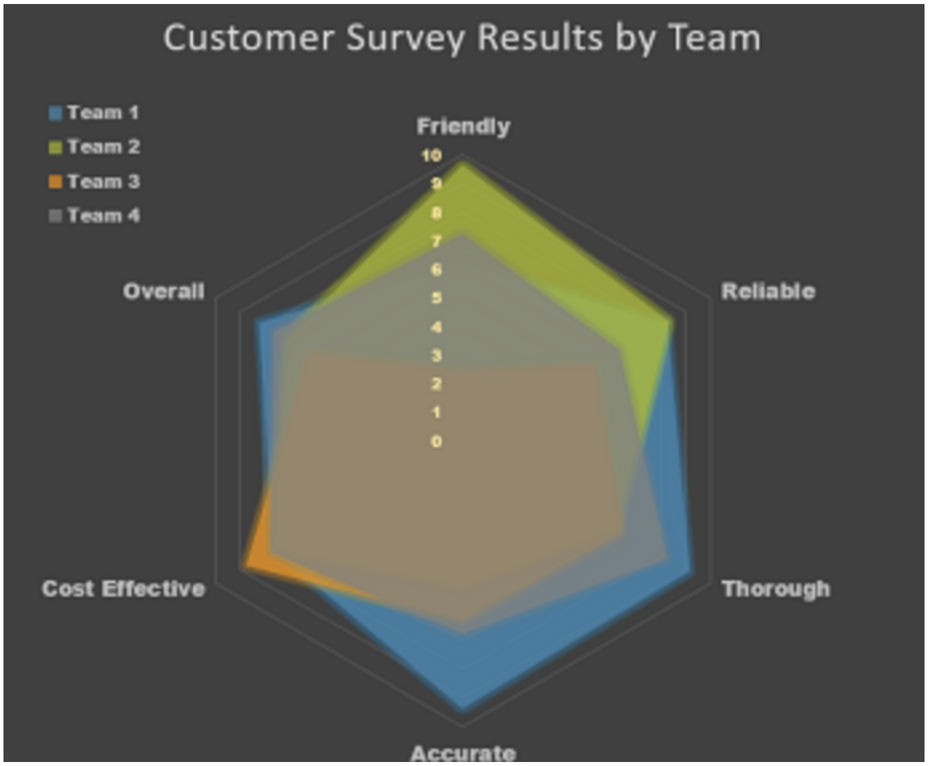


2. Bad visualizations

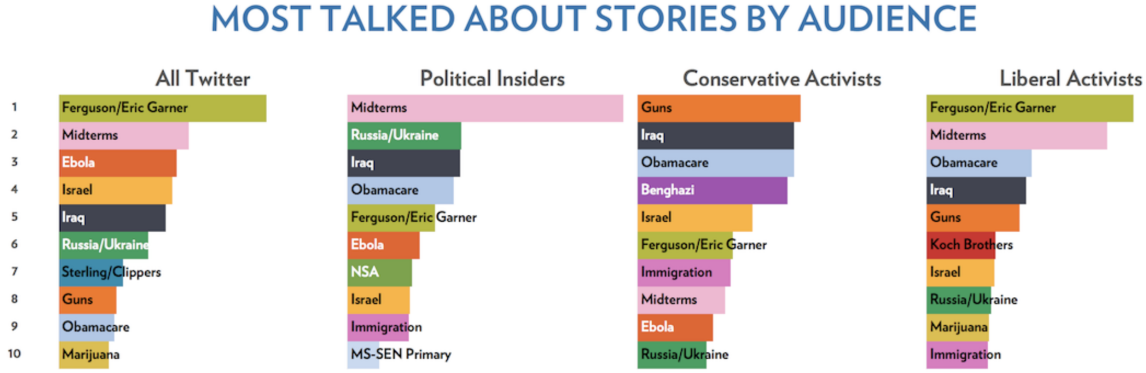
   Example 1: The graph showed the trending topics in the news - I think it is a bad example of visualization because 1. a lot of excessive informaiton is showed with the number of color presented in the graphs. 2. Although the author tried to emphasize the topics by putting labels at the color blocks with large area, some of the color blocks (e.g. the purple one between May-Jun) are unlabeled.



Example2: This example is actually not very bad. Something that could be confusing is to have the color blocks that are too compact and the relative position with relation to the time is not straight forward.The ticks for the timeline is weridly chosen as 15 days, and therefore the date of game requires calculations based on the timeline, which missed the purpose of data visualization.



3.Good visualization Example 1: The radar plot has for overlapping, half-transparent graphs for 4 teams. The scale is clearly labeled and it directly compares the advantages and disadvantages of each team on the 6 aspects listed. Would probably be better if the shared space is more clear on which team is worse, but overall it is informative.



Example 2: This graph showed the most talked abot topics grouped by political identity, which showed nicely what the most talked about topics are and how they varied depending on political affiliation. Some concern about funnel plot is that it can be heavily biased by the large number and unequal sample sizes between each groups. However, the caption of this graphs (not included) showed a similar sample sizes and make these charts more comparable.

4. EDA is good for setting up the baselines of the observations we intend to measure and look for possible previously unpredicted trends and patterns. It helps at the early stage of research by providing a better understanding of the population or the data samples, for example to range of the data or the distribution of the data, so that we know the intended measurements are reasonable to used for the samples. Also, it may establish a baseline for the measurement.EDA is also helpful when there was no certain directional prediction about variables in the data, and can be used to stimulate questions and discussions about the unforseeable trends and patterns.

5.In John Tukey's "We Need Both Exploratory and Confirmatory" (1980) paper, he proposed that exploratory data aalysis is more of "an attitude, a flexibility and a reliance on display, not a bundle of techniques" whereas the confirmatory data analysis is more a set of statistical techniques that can computerize answers for some circumscried questions. For example, if we understand the effect of some drug on an deadly illness, we may first want to do EDA to understand the average life-expectancy of the patients with the traditional treatements, and examine the causes of dead to establish some basic parameters people can measure within a patients. A confirmatory data analyasis would be having a more refined and practical question, such as whether a drug would extend the life-expectancy of a patient by comparing measurments from the control group and the treatment group with some statistical test.