Assignment Exploratory Data Analysis

Dataset

The dataset contains information about around 17.000 games from the Apple App Store. The data was gathered around the 3rd of August 2019 and contains information about the games such as the price of each game, their overall user rating, their release date and other relevant information about the game on the market. The games in the dataset are all strategy games under which puzzle games and educational games can be found. The data was collected using the iTunes API and the App Store sitemap.

Procedure

I first loaded the dataset and only used a couple of features from the dataset, the ones I deemed to be most interesting to investigate. I then added an extra column in order to capture whether a game is free or not. I also found that some columns had missing values, where Average User Ratings had the most. I therefore, omitted the rows where this column had empty values. Then I calculated some averages to quickly show a small summary of the most important features.

I then looked at the top 5 most expensive games, this showed that the top 3 games were very far away from most of the points and thus removed those points from the following plot that showed the data's density distribution. Before, however, I first calculated and plotted both means of the average user rating of both paid and free games. This showed that there was barely a differences in user rating for both groups. I then looked at which developers had the best ratings, which I summarised in the table. A lot of developers received a 5 star rating, even when adjusting for averages with low User Rating Counts.

Finally, I looked at the frequencies and plotted the overall frequencies of games made throughout the years. I then looked at the frequencies of the previous two groups again and the amount of games developed throughout the years that were free and paid. This was summarised in a table and then plotted. This showed that around 2008 the ratio between the groups was even but as the years progressed, much more free games have been developed. However, overall, but mostly free games, the frequencies are dropping in the latest years. As a last measure, I also plotted the density distribution of the user ratings for both of these groups and found that the free games had higher star ratings that the paid ones.

Loading in the relevant libraries

```
library(tidyverse)
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
## Registered S3 method overwritten by 'rvest':
##   method            from
##   read_xml.response xml2
```

```
## -- Attaching packages ---------------------------------------------------------
```

```
## v ggplot2 3.1.1       v purrr   0.3.2
## v tibble  2.1.1       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0
```

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

```
library(knitr)
```

Loading the dataset

```
games <- read_csv("Data/appstore_games.csv") %>%
          select(Name,
                 `Average User Rating`,
                 `User Rating Count`,
                 Price,
                 Developer,
                 `Original Release Date`)
```

```
## Parsed with column specification:
## cols(
##   URL = col_character(),
##   ID = col_double(),
##   Name = col_character(),
##   Subtitle = col_character(),
##   `Icon URL` = col_character(),
##   `Average User Rating` = col_double(),
##   `User Rating Count` = col_double(),
##   Price = col_double(),
##   `In-app Purchases` = col_character(),
##   Description = col_character(),
##   Developer = col_character(),
##   `Age Rating` = col_character(),
##   Languages = col_character(),
##   Size = col_double(),
```

```
##   `Primary Genre` = col_character(),
##   Genres = col_character(),
##   `Original Release Date` = col_character(),
##   `Current Version Release Date` = col_character()
## )
```

Quick relevant averages

```
#Create an extra column capturing whether the game is free or not.
#Afterwards, remove missing values
games <- games %>% mutate(Paid = Price != 0) %>%
          mutate(Paid = recode(as.character(Paid),
                               "TRUE" = "Yes",
                               "FALSE" = "No")) %>%
          na.omit(`Average User Rating`)

#Quick summary of averages in the data
games %>% summarise(
      "Average User Rating over all games" = mean(`Average User Rating`),
      "Average Rating Count" = mean(`User Rating Count`),
      "Average Price" = mean(Price)) %>% kable()
```

Average User Rating over all games	Average Rating Count	Average Price
4.060905	3306.531	0.5713054

Let's look at the top most expensive games

```
#Show the top 5 most expensive games.
topGames <- games %>% arrange(-Price)
colnames(topGames) <- c("Names",
                        "Rating",
                        "Count",
                        "Price",
                        "Developer",
                        "Date",
                        "Paid" )
kable(topGames[0:10,], format = "latex")
```

Names	Rating	Count	Price	Developer	Date	Paid
Finabase: realtime stocks	4.5	1099	139.99	Astontek Inc	30/09/2013	Yes
GOTO Bridge 19	4.0	50	59.99	GOTO Games	13/10/2018	Yes
Chess Openings Wizard	4.0	9	36.99	Bookup Corp.	22/12/2018	Yes
SmartGo Kifu	4.5	227	19.99	Smart Go, Inc.	1/04/2010	Yes
Panzer Corps	4.5	249	19.99	Slitherine	18/12/2013	Yes
Commander the Great War	3.5	98	19.99	Slitherine	25/07/2014	Yes
Battle Academy 2: Eastern Front	4.0	83	19.99	Slitherine	23/10/2014	Yes
Warhammer 40,000: Armageddon	4.0	72	19.99	Slitherine	19/06/2015	Yes
CrazyStone DeepLearning Pro	4.0	16	16.99	UNBALANCE Corporation	8/10/2016	Yes
Heroes of Normandie	3.5	70	14.99	Slitherine	28/07/2016	Yes

Are paid games rated better than free ones?

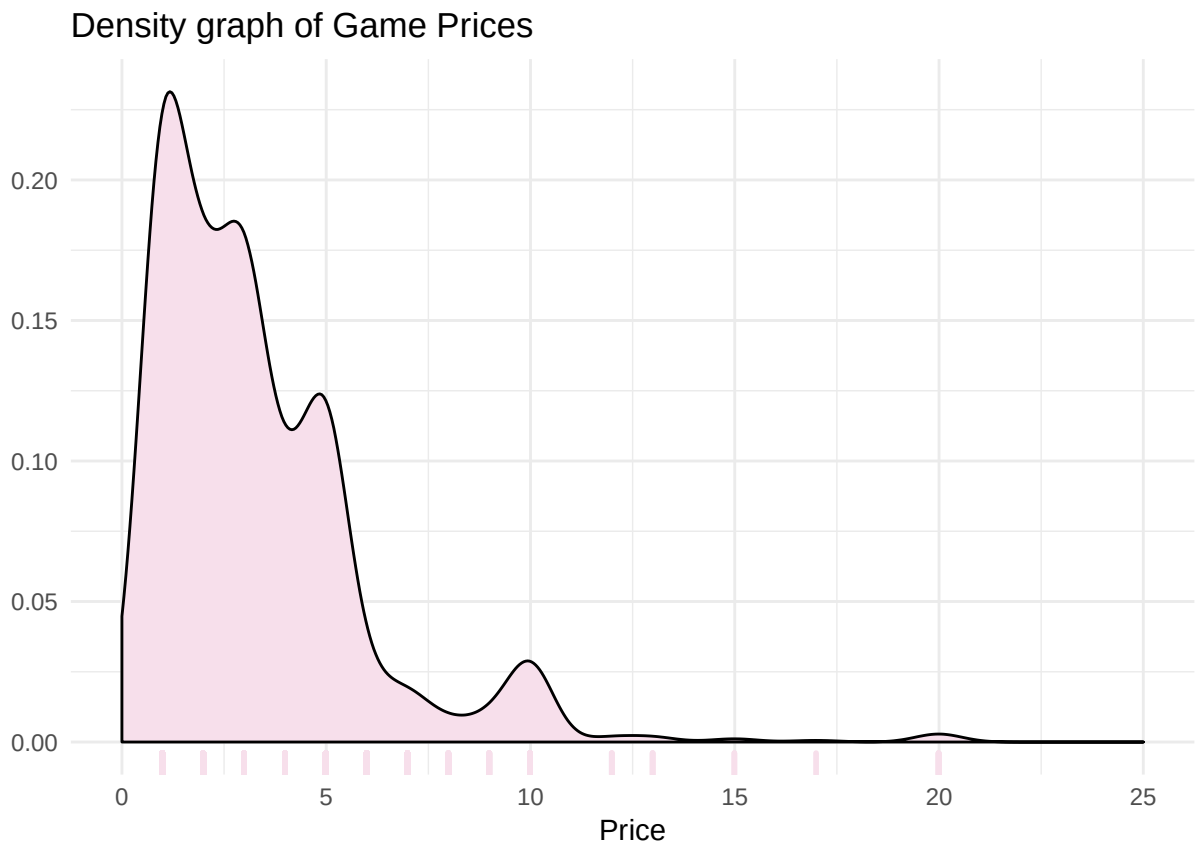
```
plotData <- topGames %>%
            group_by(Paid) %>%
            summarise(mean = mean(`Rating`),
                      variance = var(`Rating`))
kable(plotData)
```

Paid	mean	variance
No	4.071225	0.5563983
Yes	4.008091	0.6039831

Plotting the data

```
#Plot of distribution of the price of games under 25
#(to account for three outliers as can be seen in the previous table)
games %>% filter(Paid == "Yes") %>%
          ggplot(aes(x = Price)) +
          geom_density(fill = "#f7dfeb") +
          theme_minimal() +
          geom_rug(size = 1, colour = "#f7dfeb")+
          labs(y = "") +
          xlim(0,25) +
          ggtitle("Density graph of Game Prices")
```

```
## Warning: Removed 3 rows containing non-finite values (stat_density).
```



Which developers are the most popular?

```
table <- games %>% filter(`User Rating Count` > 5) %>%
    group_by(Developer) %>%
    summarise(mean = mean(`Average User Rating`)) %>%
    arrange(-mean)

kable(table[0:10,])
```

Developer	mean
“Don’t Blink Studios”	5
“It’s All A Game LLC”	5
6df1573354d454214e925a3179d16280670996505l6c53f8	5
70ed95e86e38620f	5
10K BULBS LLC	5
1791 Entertainment LLC	5
1C Mobile Ltd	5
99bosses	5
A Dark Matter Creation LLC	5
A Sharp, LLC	5

Frequency of games over the years

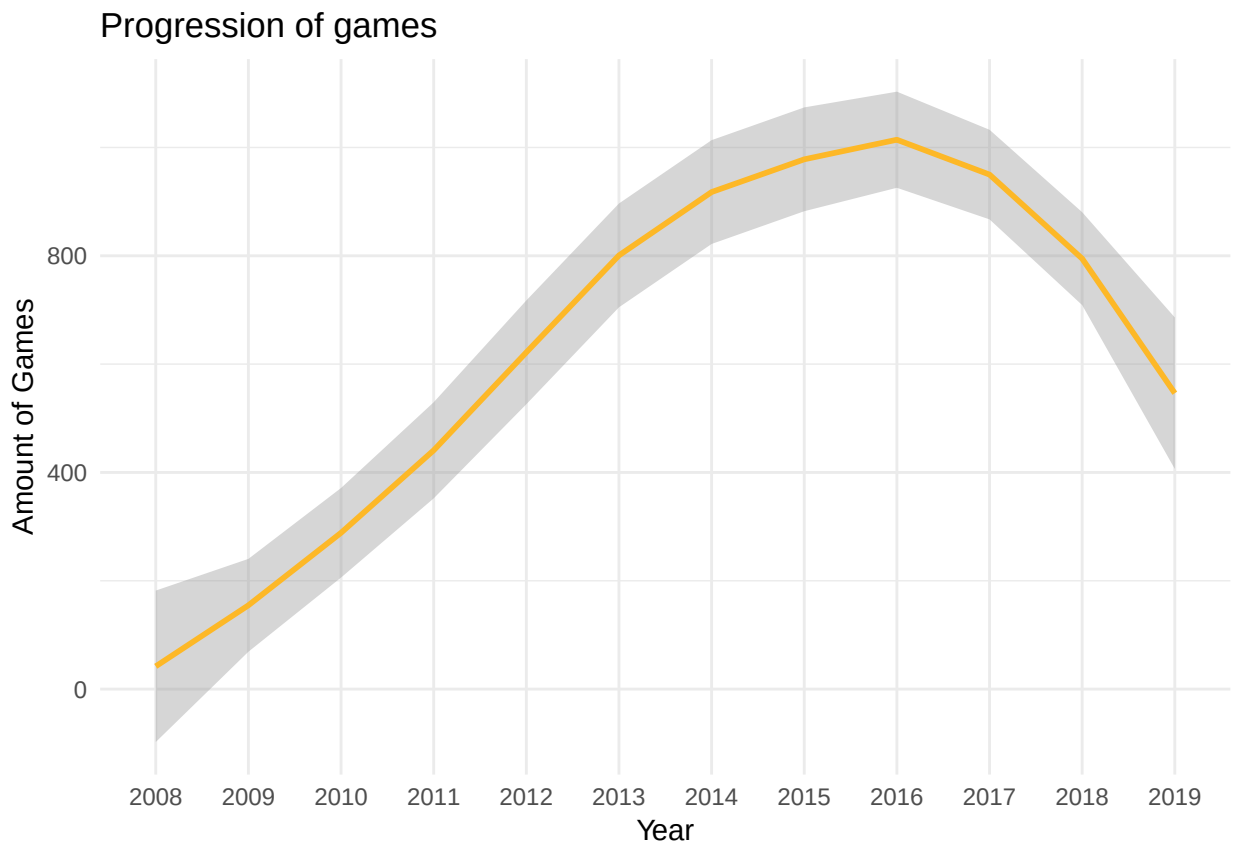
```
#Convert the release date column to date type and extract only
#the years from the date
games$`Original Release Date` <- as.Date(games$`Original Release Date`,
                                         "%d/%m/%Y")
releaseDates <- games %>%
  mutate("Year Released" = as.numeric(format(games$`Original Release Date`,
                                             "%Y")))

#Create a frequency table
frequencies <- table(releaseDates$`Year Released`) %>% data.frame()
colnames(frequencies) = c("Year", "Frequency")

kable(frequencies)
```

Year	Frequency
2008	44
2009	172
2010	255
2011	412
2012	598
2013	835
2014	955
2015	898
2016	1047
2017	964
2018	904
2019	477

Plot of the previous data over time

```
frequencies %>%
      ggplot(aes(x = Year, y = Frequency, group = 1)) +
      geom_smooth(colour = "#fdb827") +
      theme_minimal() +
      ggtitle("Progression of games") +
      labs(x = "Year", y = "Amount of Games")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Group by paid or non paid

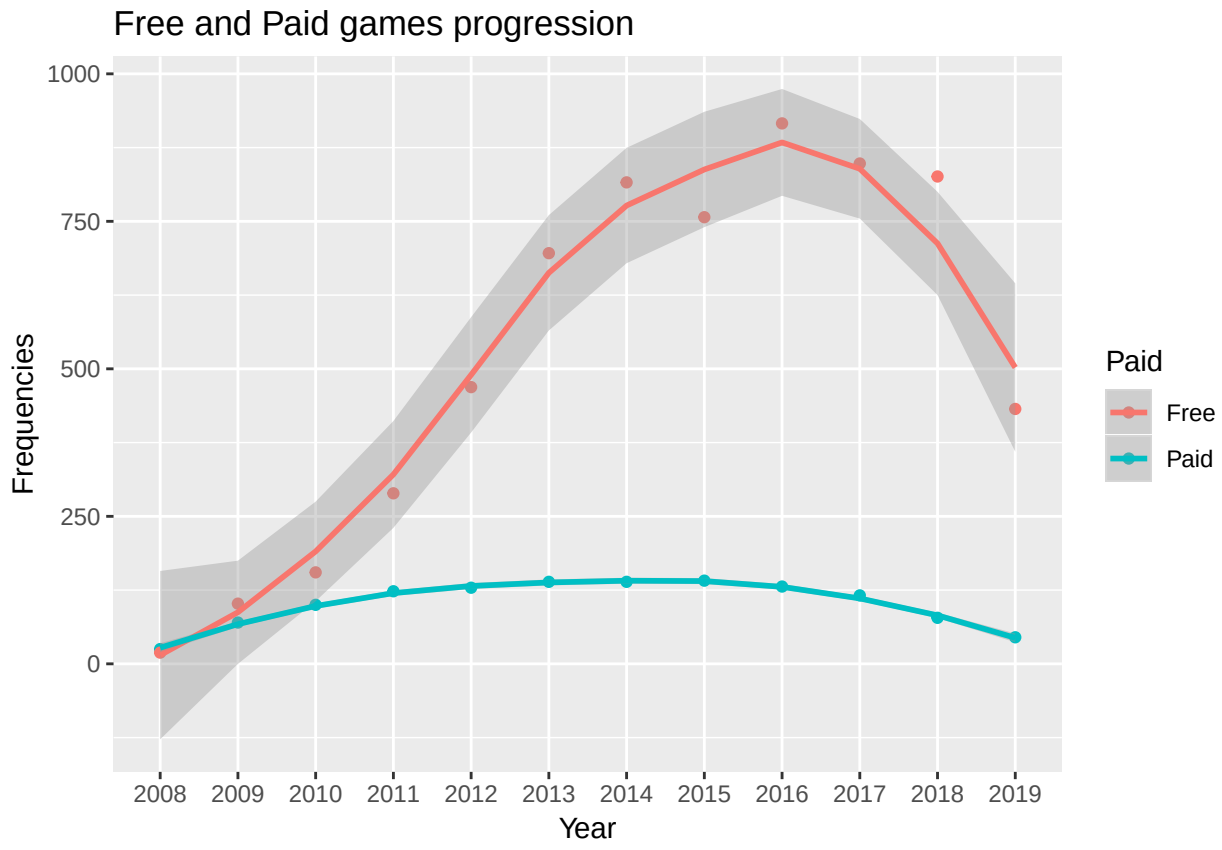
```
#Separate paid and free games
paidReleaseDates <- releaseDates %>% filter(Paid == "Yes")
paidReleaseDates <- table(paidReleaseDates$`Year Released`) %>% data.frame()

freeReleaseDates <- releaseDates %>% filter(Paid == "No")
freeReleaseDates <- table(freeReleaseDates$`Year Released`) %>% data.frame()

#Combines them into one
combined <- data.frame(
  Year = rep(paidReleaseDates$Var1,2),
  Frequencies = c(paidReleaseDates$Freq, freeReleaseDates$Freq),
  Paid = c(rep("Paid", 12), rep("Free", 12))
)
```

Plotting the data

```
combined %>% ggplot(aes(x = Year,
                        y = Frequencies,
                        group = Paid,
```

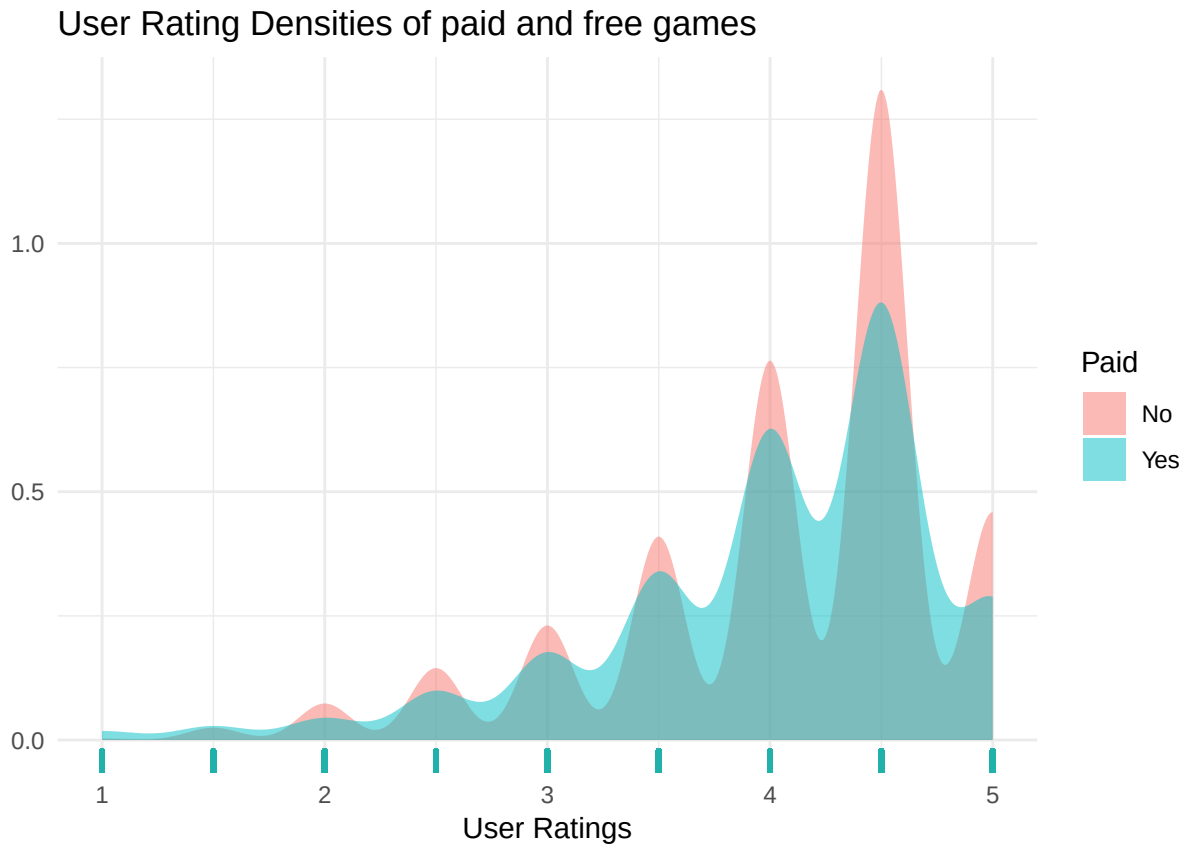
```
                        colour = Paid)) +
  geom_point() +
  geom_smooth() +
  ggtitle("Free and Paid games progression")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Differences in user ratings for paid and free games

```
#Plot the data
games %>%
  ggplot(aes(x = `Average User Rating`, fill = Paid)) +
  geom_density(alpha = .5, colour = NA) +
  geom_rug(size = 1, colour = "light seagreen") +
  theme_minimal() +
  ggtitle("User Rating Densities of paid and free games") +
  labs(y = "", fill = "Paid", x = "User Ratings")
```


User Rating Densities of paid and free games
1.0
0.5
0.0
1
2
3
4
5
User Ratings
Paid
No
Yes