Video Game Sales from 2000 to 2017

Jonathan Marvin

2/23/2022

Asking the important questions

Scenario

You are a junior data analyst working for a business intelligence consultant. You have been at your job for six months, and your boss feels you are ready for more responsibility. He has asked you to lead a project for a brand new client — this will involve everything from defining the business task all the way through presenting your data-driven recommendations. You will choose the topic, ask the right questions, identify a fresh dataset and ensure its integrity, conduct analysis, create compelling data visualizations, and prepare a presentation.

Business task

Research the sales of video games from 2000 to 2017 in the US, Europe and Japan and discover what genres of video games are most popular and the revenue from each region comparatively.

Stakeholders

My company was hired by a video game marketing firm who want to know what the best marketing strategy is and what region would be the best for advertising. They are mostly interested in action, adventure and role-playing so I

The Data

The dataset used in this case study, Video Game Sales, was found on Kaggle and was originally posted on Data.World. The base dataset covers video game sales from 1977 to 2017.

Data limitations

- The data is only up to 2017, effectively excluding anything released afterwards.
- There are NULL values that have to be removed so as to keep the data fair and will impact the revenue from each region.
- The data was collected by a third-party source.

Data cleaning

Tools I used

I decided to clean the white spaces using Google sheets and filter the data using both SQL and R Studio. I prefer the R route being that I can make the visualizations in the same script and create the markdown in the same software. The SQL script is a show of the work done before moving into R Studio.

Google Sheets:

I wanted to use only clean data so I started in Google Sheets to make sure there were no blank rows or spaces out of place. The data was already fairly cleaned so I didn't have too much to do here. I decided to add total sales from each region to compare the revenue from across the globe.

SQL:

The dataset I worked with starts at 1977 so I had to cut it down to fit my needs. I removed any NULL values to keep the data unbiased. For example, there were 296 NULL years and I recognized some of the consoles, such as 2600, to be older than the data I needed. Next, I decided to group the data into distinct genres based on the global sales to see what the highest selling genre was and I took the top fifty from each category so as not to make the visualizations seem too cluttered. The SQL code can be found here

R Studio

When I knew exactly what I needed, I came to R Studio to filter the data down further. I first had to load the tidyverse package and my data into R studio.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

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

```
vg_clean <- read.csv('Game sales/Video_Games_clean.csv')
```

I started with the year filter to take out any data before 2000, followed by the genre filter to only show action, adventure and role-playing.

```
vg_filter = filter(vg_clean, between(Year, '2000','2017'))
```

```
## Warning in between(Year, "2000", "2017"): NAs introduced by coercion
```

```
genres <- c('Action', 'Adventure', 'Role-Playing')
vg_filter1 = filter(vg_filter, Genre %in% genres)
```

I then filtered the new data to show only the top fifty games from each genre.

```
action = filter(vg_filter1, Genre == 'Action')
action_slice = slice(action, 1:50)

adv = filter(vg_filter1, Genre == 'Adventure')
avd_slice = slice(adv, 1:50)

rp = filter(vg_filter1, Genre == 'Role-Playing')
rp_slice = slice(rp, 1:50)
```

With the tables limited to fifty values, I was able to bind them.

```
temp <- rbind(action_slice, avd_slice)

genres <- rbind(temp, rp_slice)
```

I also had to convert all of the data into the right format using the retype function.

```
library(hablar)
```

```
##
## Attaching package: 'hablar'

## The following object is masked from 'package:dplyr':
##
##     na_if
```

```
## The following object is masked from 'package:tibble':
##
##     num
```

```
genres %>%
  retype()
```

```
## # A tibble: 150 x 15
##    Name    Platform  Year Genre Publisher NA_Sales EU_Sales JP_Sales Other_Sales
##    <chr>   <chr>    <int> <chr> <chr>        <dbl>    <dbl>    <dbl>       <dbl>
##  1 Grand ~ PS3       2013 Acti~ Take-Two~     7.02     9.09     0.98        3.96
##  2 Grand ~ PS2       2004 Acti~ Take-Two~     9.43     0.4      0.41       10.6
##  3 Grand ~ X360      2013 Acti~ Take-Two~     9.66     5.14     0.06        1.41
##  4 Grand ~ PS2       2002 Acti~ Take-Two~     8.41     5.49     0.47        1.78
##  5 Grand ~ PS2       2001 Acti~ Take-Two~     6.99     4.51     0.3         1.3
##  6 Grand ~ PS4       2014 Acti~ Take-Two~     3.96     6.31     0.38        1.97
##  7 Pokemo~ DS        2009 Acti~ Nintendo      4.34     2.71     3.96        0.76
##  8 Grand ~ X360      2008 Acti~ Take-Two~     6.76     3.07     0.14        1.03
##  9 Grand ~ PS3       2008 Acti~ Take-Two~     4.76     3.69     0.44        1.61
## 10 FIFA S~ PS3       2012 Acti~ Electron~     1.06     5.01     0.13        1.97
## # ... with 140 more rows, and 6 more variables: Global_Sales <dbl>, X <dbl>,
## #   X.1 <dbl>, X.2 <dbl>, X.3 <dbl>, X.4 <dbl>
```

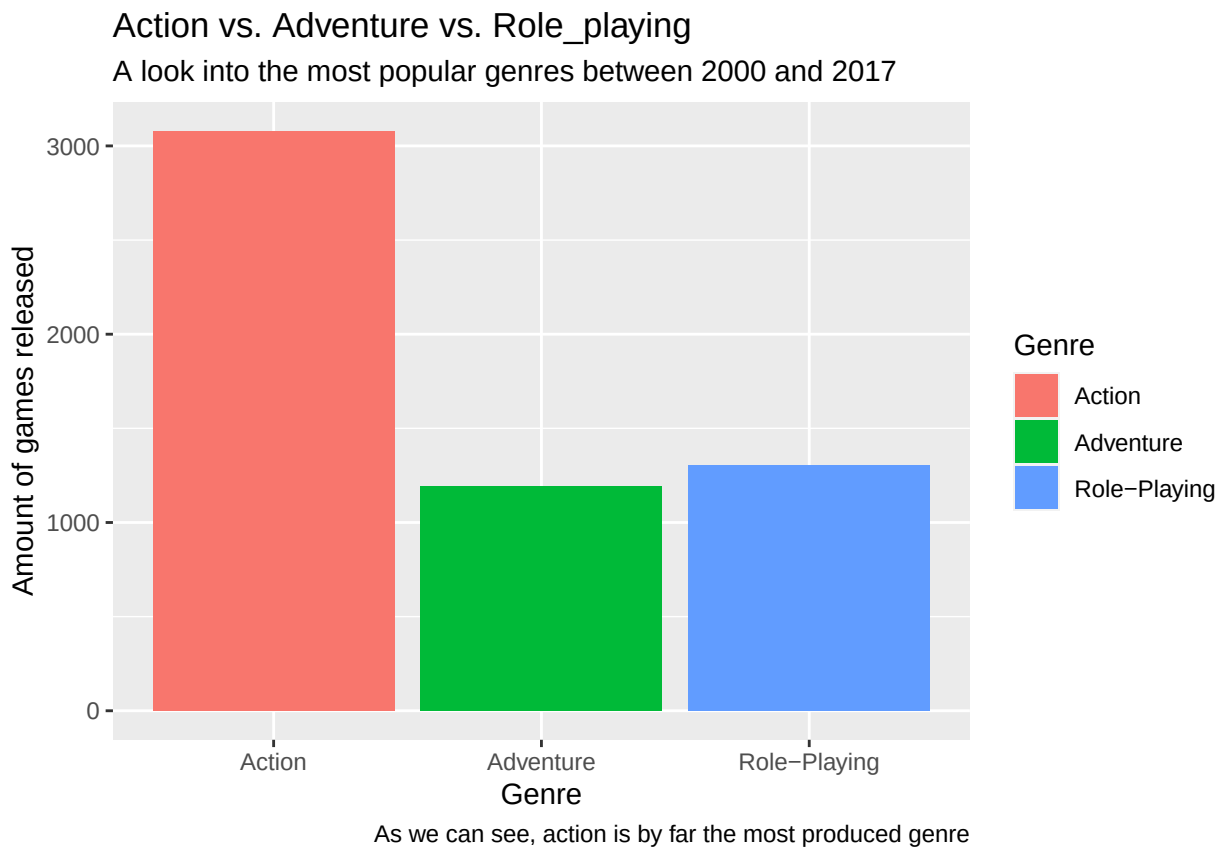
With the data broken down to the right size, we're ready to make the visualizations.

Charting and visualization

Action, adventure and role-playing games tend to intersect but the base genre is what we want to see. Not all action games have adventures just as not all adventure games are role-playing. As stated earlier, I took the top fifty games from each genre and compared them based on sales in each region in the dataset. I'll specifically be looking at the amount of games created for each genre from 2000 to 2017, and the sales in North America, Europe and Japan.

Games released

```
ggplot(data = vg_filter1)+
  geom_bar(aes(x = Genre, fill = Genre))+
labs(title = 'Action vs. Adventure vs. Role_playing',
     subtitle = 'A look into the most popular genres between 2000 and 2017',
     caption = 'As we can see, action is by far the most produced genre')+
  ylab('Amount of games released')
```



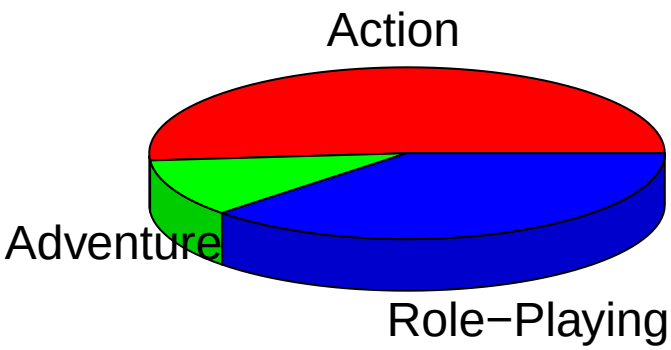
North America sales

```
library(plotrix)


NAslices <- c(159.46, 33.94, 117.20)

NAlbls <- c('Action', 'Adventure', 'Role-Playing')

pie3D(NAslices, labels = NAlbls, main = 'Sales in North America based on genre')
```



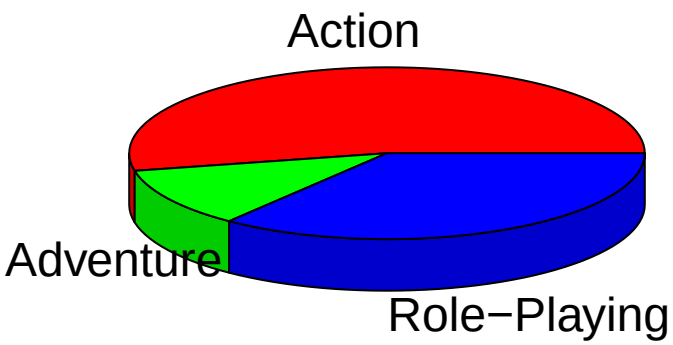
We can see that action games sold the best in North America with 159 million dollars in revenue. Role-playing made 117 million and adventure made almost 34 million.

Europe Sales

```
EUslices <- c(116.50, 24.53, 77.59)

EUlbls <- c('Action', 'Adventure', 'Role-Playing')

pie3D(EUslices, labels = EUlbls, main = 'Sales in Europe based on genre')
```



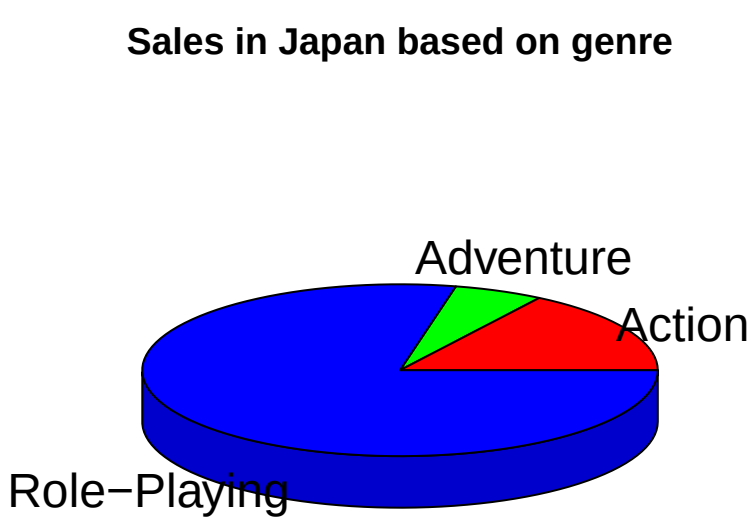
Similarly to North America, the action genre dominates the revenue at 116 million.

Japan Sales

```
JPslices <- c(18.03, 6.35, 88.90)

JPlbls <- c('Action', 'Adventure', 'Role-Playing')

pie3D(JPslices, labels = JPlbls, main = 'Sales in Japan based on genre')
```



Japan, however, prefers role-playing games to action games with a revenue of almost 89 million.

Recommendations

As the data shows, marketing action games in North America and Europe would be the best option whereas in Japan, role-playing games would be the most marketable. All of the genres sell well but each region has their own niche.