CPE213 Data Models

Sales Games Prediction

สมาชิกกลุ่ม

1.นาย ปวริศ ร้านชิตวงศ์ 61070501034

2.นาย พีรภัทร เขมะชิต 61070501039

3.นาย ยศกร นุ่นปาน 61070501043

นักศึกษาชั้นปีที่ 2 คณะวิศวกรรมศาสตร์ ภาควิชาวิศวกรรมคอมพิวเตอร์

นำเสนอ

ผศ.ดร.สันติธรรม พรหมอ่อน

รายงานนี้เป็นส่วนหนึ่งของวิชา CPE 213 Data Models

มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

ปีการศึกษา 2562

Introduction to the problem

ตั้งแต่อดีตจนถึงปัจจุบันได้มีเกมหลากหลายประเภทถูกผลิตและจัดจำหน่ายเป็นจำนวนมาก และในแต่ละยุคก็จะมีความนิยมในการเล่นเกมในแต่ละประเภทที่แตกต่างกัน การที่เราสามารถทำนายได้ ว่าเกมประเภทใดที่จะได้รับความนิยมในช่วงนั้นๆ ทำให้ผู้ผลิตหรือผู้จัดจำหน่ายสามารถตัดสินใจผลิตเกมออกมาได้ตรงต่อความต้องการของผู้เล่นในช่วงเวลานั้นและเพื่อความคุ้มค่าของเม็ดเงินที่ต้อใช้จ่ายไปในการลงทุนของบริษัท ทางกลุ่มของพวกเราจึงมีความสนใจที่จะศึกษาเกี่ยวกับแนวโน้มของความนิยมในเกมประเภทต่างๆตั้งแต่ในอดีตจนถึงปี ค.ศ.2020 เพื่อเปรียบเทียบและทำนายความคุ้มมค่าในการผลิตเกม

Analytic objective

เพื่อสร้างโมเดลศึกษาความคุ้มค่าในการลงทุนผลิตเกมประเภทต่างๆของแต่ละบริษัท

Data descriptive

ข้อมูลแสดงรายชื่อเกมที่วางจำหน่ายมียอดขายมากกว่า 10,000 ชุดโดยผู้จัดจำหน่าย ตั้งแต่ปี ค.ศ.1980 ถึง ค.ศ. 2020 ซึ่งมีข้อมูลเกี่ยวกับชื่อ แพลตฟอร์ม ปีที่จัดจำหน่าย ประเภทเกม ผู้จัดจำหน่าย ยอดขายในอเมริกาเหนือ ยุโรป ญี่ปุ่น ในภูมิภาคอื่นๆ และยอดทั่วขายรวมทั่วโลกโดยยอดขายทั้งหมดแสดงในหน่วยล้านชุดโดยข้อมูลทั้งหมดถูกรวบรวมโดยเว็บไซต์ vgchartz (https://www.vgchartz.com/) ซึ่งเป็นเว็บไซต์เกี่ยวกับข้อมูลต่างของเกมต่างๆที่ถูกจัดจำหน่ายในแต่ละปี

Ref: https://www.kaggle.com/gregorut/videogamesales?select=vgsales.csv

Data dictionary

Column	Description
Rank	Ranking of overall sales
Name	The games name
Platform	Platform of the games release (i.e. PC,PS4, etc.)
Year	Year of the game's release
Genre	Genre of the game
Publisher	Publisher of the game
NA_Sales	Sales in North America (in millions)
EU_Sales	Sales in Europe (in millions)
JP_Sales	Sales in Japan (in millions)
Other_Sales	Sales in the rest of the world (in millions)
Global_Sales	Total worldwide sales.

Data preparation

Data Cleaning

ข้อมูลใน คอลลัมน์ Year ซึ่งเป็นข้อมูลประเภทตัวเลข (numerical) ที่แสดงปีที่จัดจำหน่ายของแต่ละเกมโดยบางเกมไม่ปรากฏข้อมูลของปีที่จัดจำหน่าย (แสดงด้วย “N/A”) เราจึงกรองข้อมูลส่วนในนี้ออก เพื่อง่ายต่อการนำข้อมูลไปใช้งานและกรองข้อมูลที่มีปีจัดจำหน่าย 2020 และ 2017 ออกเนื่องจากปริมาณข้อมูลมีเพียง 1 และ 3 แถวตามลำดับและทำการเปลี่ยนจาก Factor เป็น numeric เพื่อให้ง่ายต่อการเปรียบเทียบยอดขายเกมในแต่ละปี

ข้อมูลใน column Rank ซึ่งเป็นข้อมูลแสดงลำดับเกมที่ไม่ถูกนำไปใช้เราจึงทำการลบคอลลัมน์ ในส่วนนี้ออก

ข้อมูลใน คอลลัมน์ Publisher ซึ่งเป็นข้อมูลประเภท categorical ที่แสดงผู้จัดจำหน่ายของแต่ละเกมโดยบางเกมไม่ปรากฏข้อมูลของผู้จัดจำหน่าย (แสดงด้วย “N/A”) เราจึงกรองข้อมูลข้อมูลส่วนนี้ออกเพื่อให้ง่ายต่อการใช้งานและ กรองข้อมูลให้เหลือเฉพาะ Publisher ที่ผลิตเกมตั้งแต่ 100 เกมขึ้นไปเพื่อลดจำนวน outliner

```
game %>% select(-Rank) %>% filter(Year != "N/A" ) %>% filter(Year != "2020") %>% filter(Year != "2017") %>%
  filter(Publisher!= "N/A") -> FilterGame

FilterGame$Year <- as.numeric(as.character(FilterGame$Year))

FilterGame %>% group_by(Publisher) %>% summarise(n=n()) %>% filter(n>100) -> nopublish

FilterGame %>% filter(Publisher %in% nopublish$Publisher) ->FilterGame
```

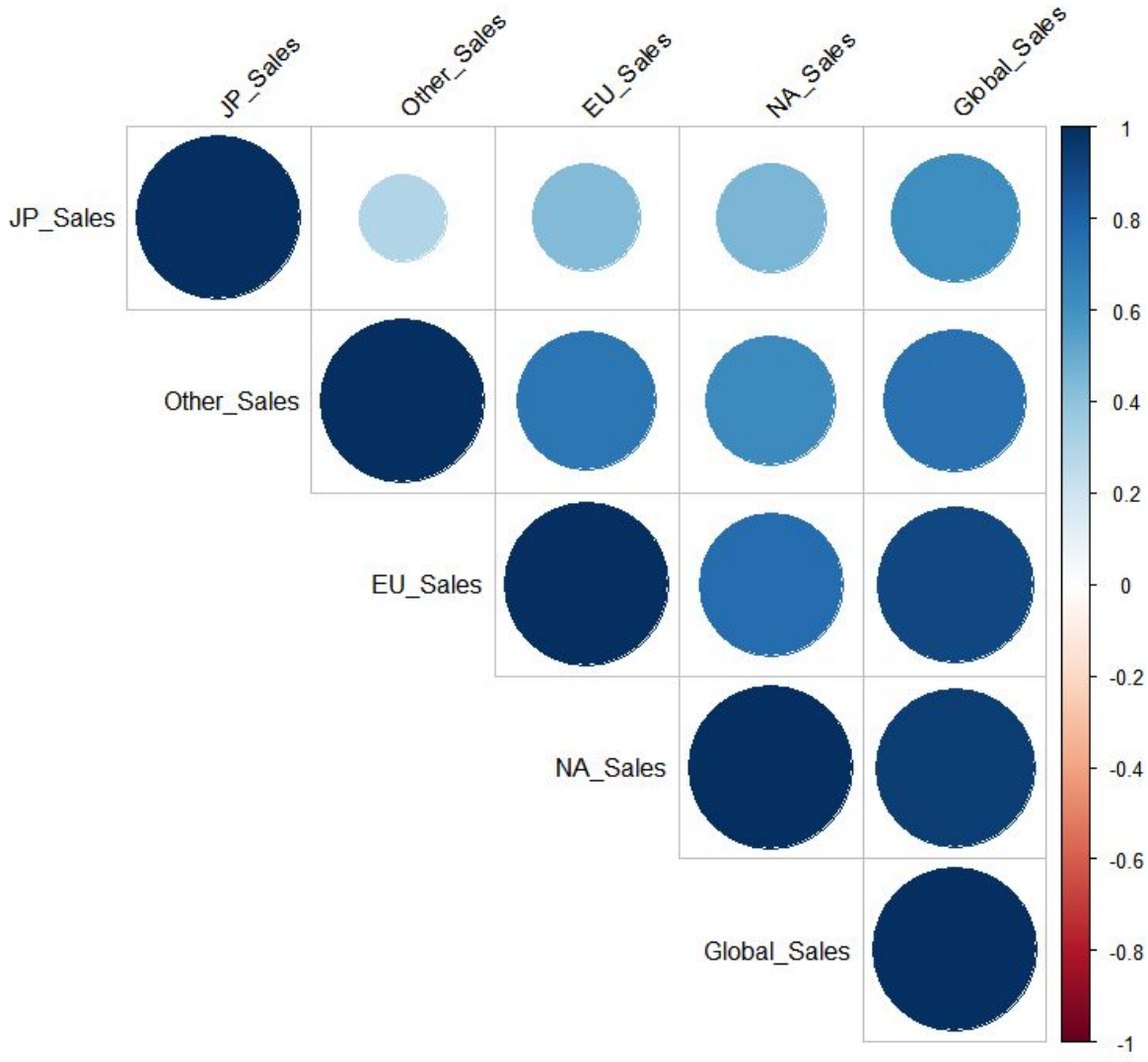
	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
1	Wii Sports	Wii	2006	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00
5	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37
6	Tetris	GB	1989	Puzzle	Nintendo	23.20	2.26	4.22	0.58	30.26
7	New Super Mario Bros.	DS	2006	Platform	Nintendo	11.38	9.23	6.50	2.90	30.01
8	Wii Play	Wii	2006	Misc	Nintendo	14.03	9.20	2.93	2.85	29.02
9	New Super Mario Bros. Wii	Wii	2009	Platform	Nintendo	14.59	7.06	4.70	2.26	28.62
10	Duck Hunt	NES	1984	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.31
11	Nintendogs	DS	2005	Simulation	Nintendo	9.07	11.00	1.93	2.75	24.76
12	Mario Kart DS	DS	2005	Racing	Nintendo	9.81	7.57	4.13	1.92	23.42
13	Pokemon Gold/Pokemon Silver	GB	1999	Role-Playing	Nintendo	9.00	6.18	7.20	0.71	23.10
14	Wii Fit	Wii	2007	Sports	Nintendo	8.94	8.03	3.60	2.15	22.72

เมื่อทำการกรองข้อมูลเสร็จจะได้ข้อมูลจาก 16,598 แถวเหลือ 11,735 แถว

```
FilterGame          11735 obs.
```

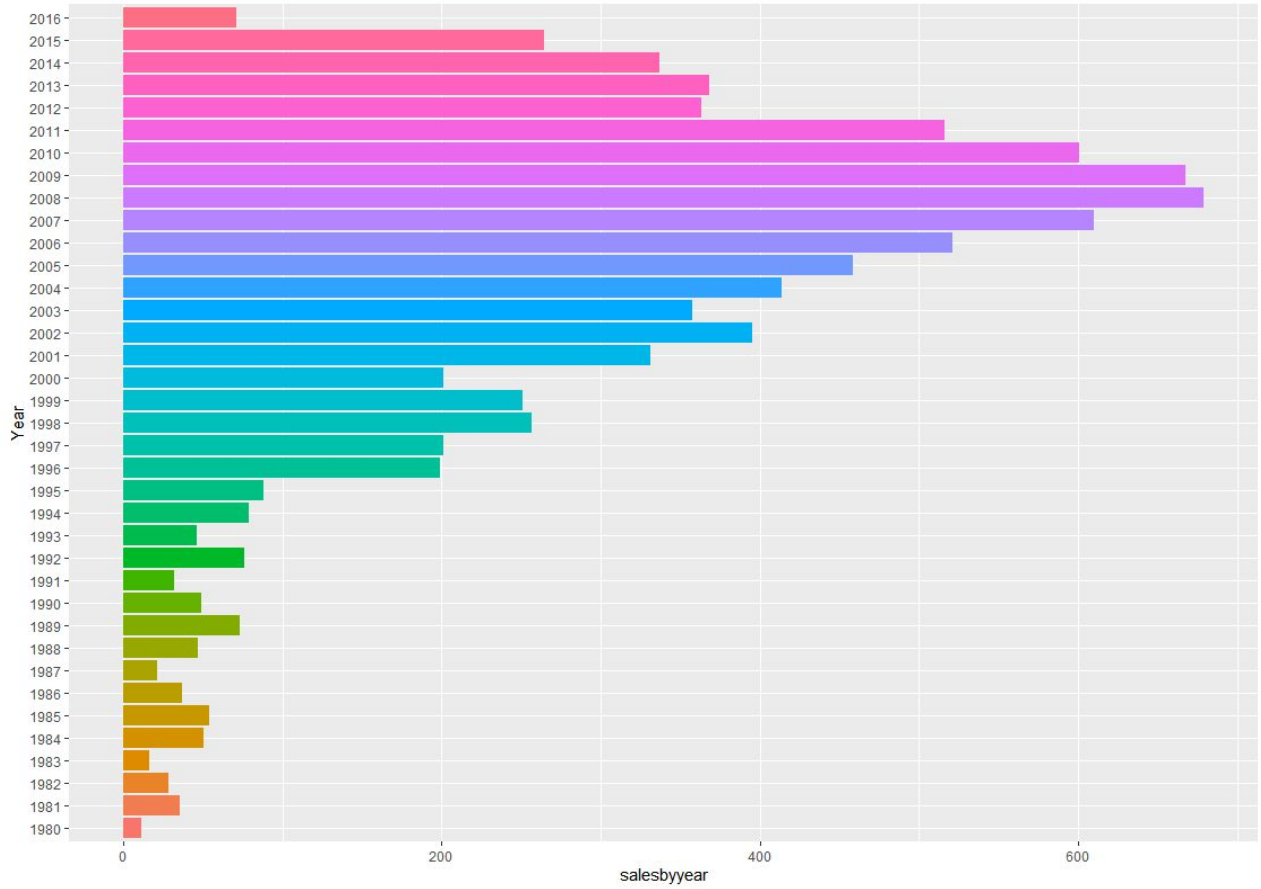
Data exploration and visualization

Correlation



แสดงความสัมพันธ์ของตัวแปร numberic ระหว่างยอดขายในแต่ละส่วนจากกราฟเราสามารถบอกได้ว่าข้อมูลมีความสัมพันธ์กันโดยเฉพาะ NA_Sales กับ Global_Sales มีความสัมพันธ์กันมากที่สุดจึงสามารถบอกได้ว่ายอดว่าของ Global_sales ส่วนใหญ่นั้นมาจาก NA_sales

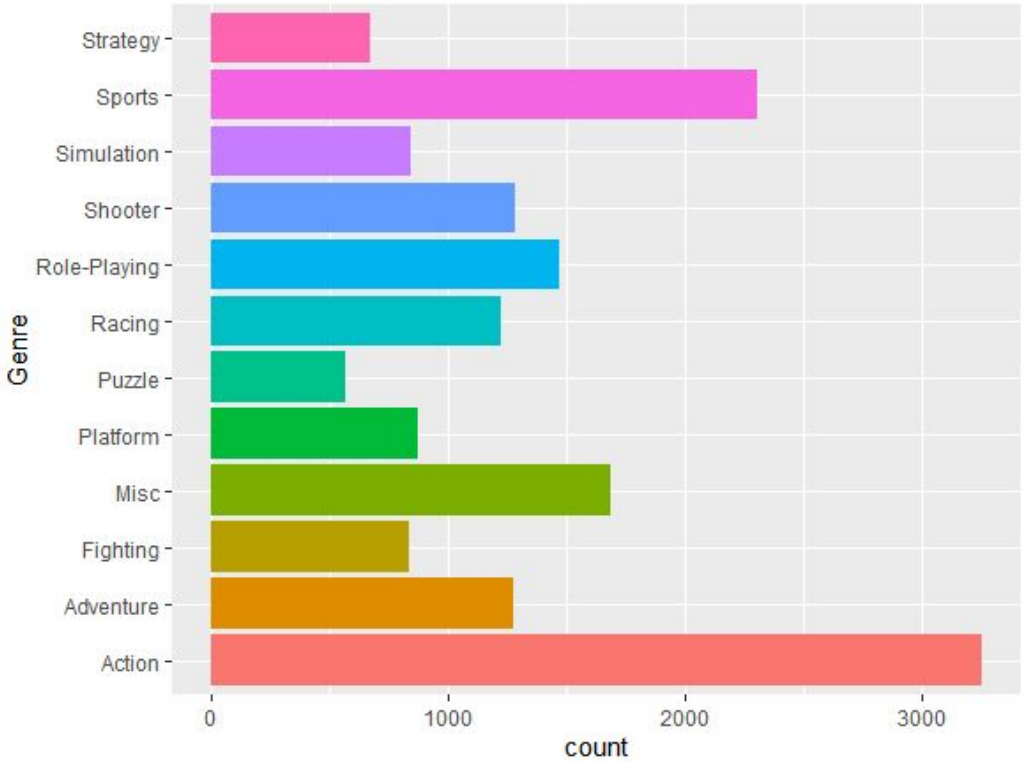
กราฟแสดงยอดขายรวมในแต่ละปี




```
FilterGame %>% group_by(Year) %>% summarise(salesbyyear = sum(Global_Sales)) %>% ggplot() +
  geom_col(mapping = aes(x = Year, y = salesbyyear,fill=Year))+coord_flip()
```

กราฟนี้คือกราฟ Bar Chart ที่แสดงข้อมูลจำนวนเกมทุกประเภทที่ขายได้ในหน่วยล้านชุดตั้งแต่ปี ค.ศ.1980 ถึงปี ค.ศ.2016 โดยจากกราฟนั้นสามารถบอกได้ว่าปีที่ขายเกมได้ปริมาณมากที่สุดคือปี ค.ศ. 2008 และปีที่น้อยที่สุดคือปี ค.ศ.1980

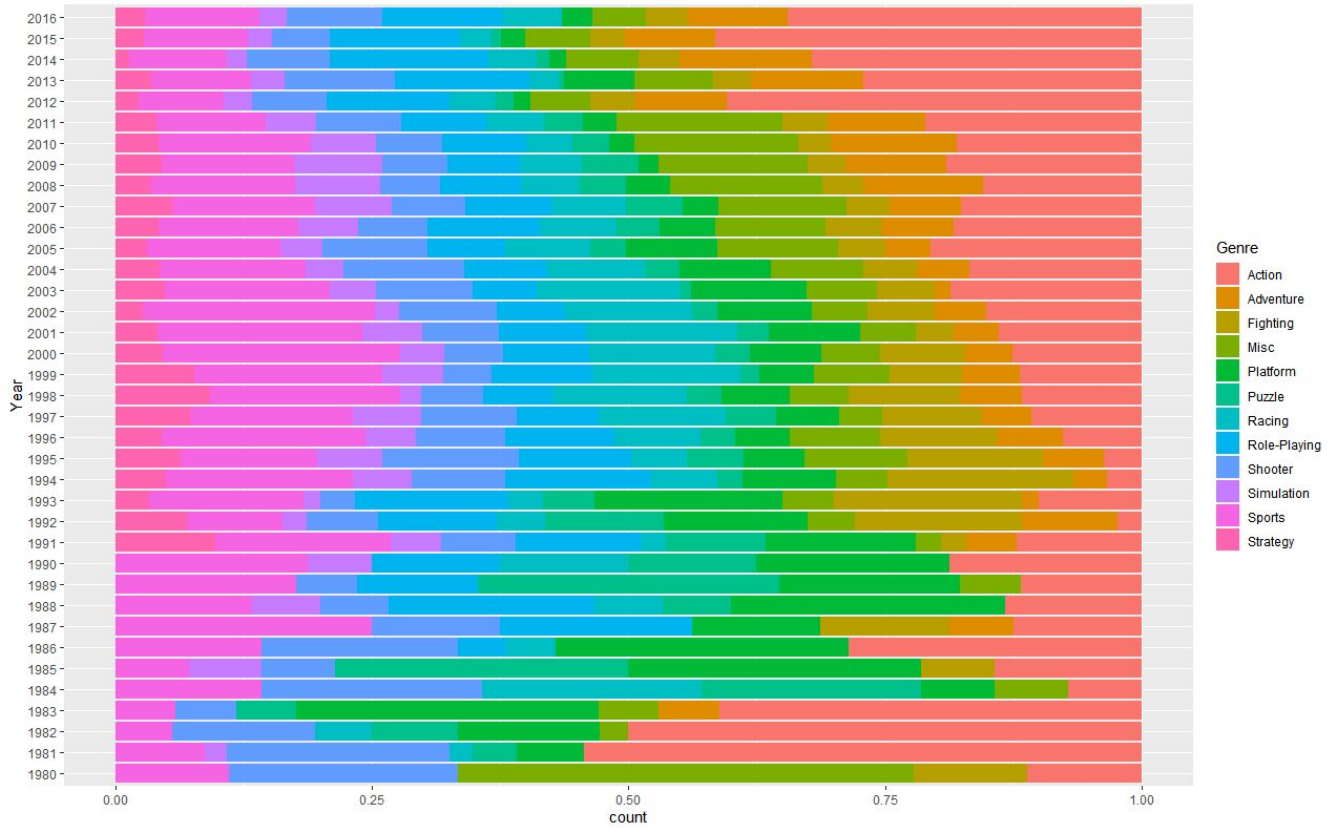
กราฟแสดงประเภทเกมทั้งหมด




```
FilterGame %>% group_by(Genre) %>% summarise(count=n()) %>%
  ggplot(aes(x=Genre,y=count,fill=Genre))+geom_col()+coord_flip()
```

กราฟนี้คือกราฟ Bar Chart ชนิด Value ใช้คำสั่ง geom_col ที่แสดงข้อมูลว่าในแต่ละประเภทของเกมมีจำนวนเกมวางจำหน่ายได้จำนวนกี่ล้านชุด โดยประเภทเกมที่จำหน่ายได้มากที่สุดคือเกมประเภท Action และรองลงมาคือเกมประเภท Sport และที่น้อยที่สุดคือเกมประเภท Puzzle

กราฟแสดงประเภทเกมขายได้ในแต่ละปี

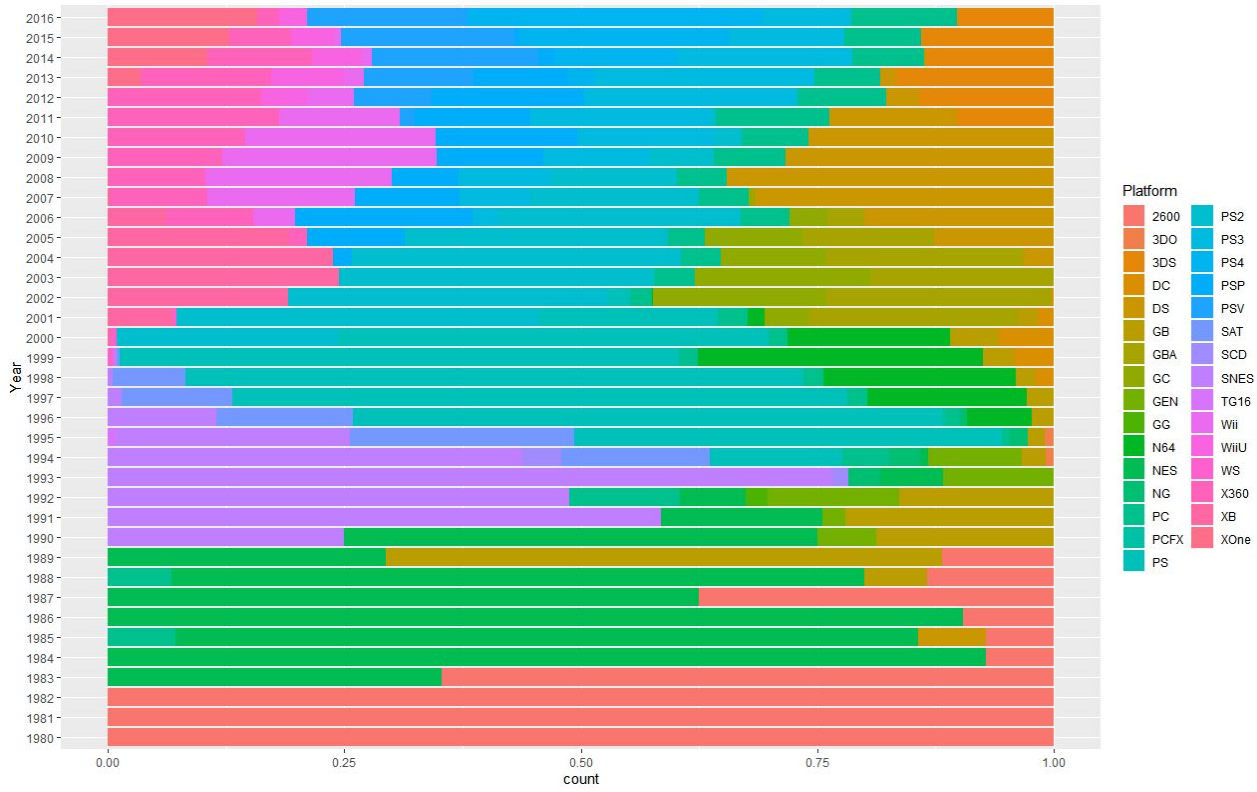



```
FilterGame %>%  ggplot(mapping = aes(x=Year,fill=Genre))+geom_bar(position = 'fill')+coord_flip()
```

กราฟนี้คือกราฟ Barchart แบบ fill ที่แสดงให้เห็นสัดส่วนให้เห็นว่าในแต่ละปี มีเกมแต่ละประเภทออกมาขายเป็นสัดส่วนเท่าไรบ้าง

- ยกตัวอย่างเช่น ในปี 1982 มีเกมประเภท Action ถึง 50% ของเกมทั้งหมดในปีนั้น ปี 1991 เป็นปีแรกที่มีการผลิตเกมแนว Strategy

กราฟแสดงเกมในแต่ละแพลตฟอร์มที่ขายได้ในแต่ละปี

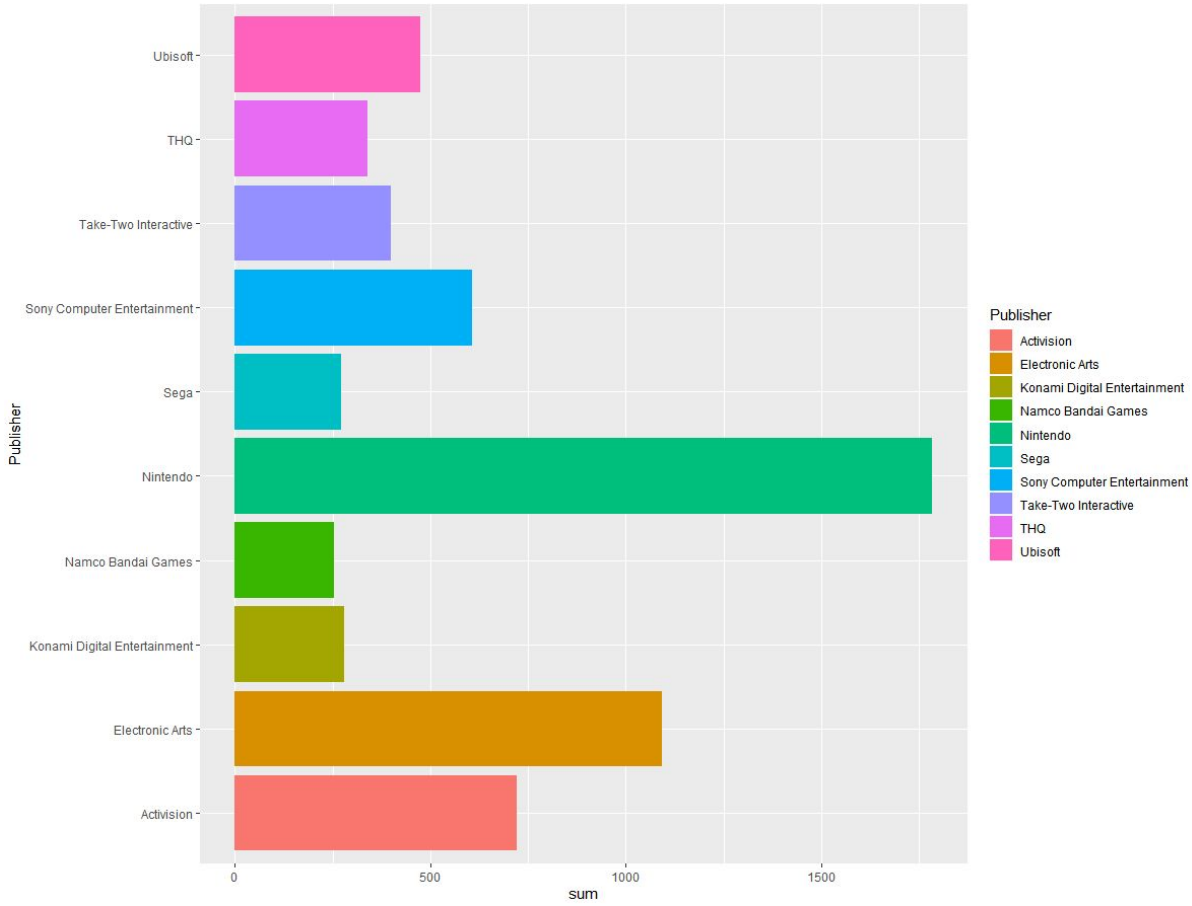



```
FilterGame %>%  ggplot(mapping = aes(x=Year,fill=Platform))+geom_bar(position = 'fill')+coord_flip()
```

กราฟนี้คือกราฟ Barchart แบบ fill ที่แสดงให้เห็นสัดส่วนว่าในแต่ละปีตั้งแต่ ค.ศ.1980 ถึง ค.ศ.2016 มีเกมของแต่ละแพลตฟอร์มขายได้เป็นสัดส่วนเท่าไรบ้าง

ยกตัวอย่างเช่น ในปี 1980 - 1982 มีเป็นเกมของเครื่อง Atari 2600 เกือบทั้งหมด หลังจากปี 1982 ก็เริ่มมีเกม Platform อื่นออกมา จนกระทั่งในปี 1990 ตลาดก็เลิกผลิตเกมของเครื่อง Atari 2600

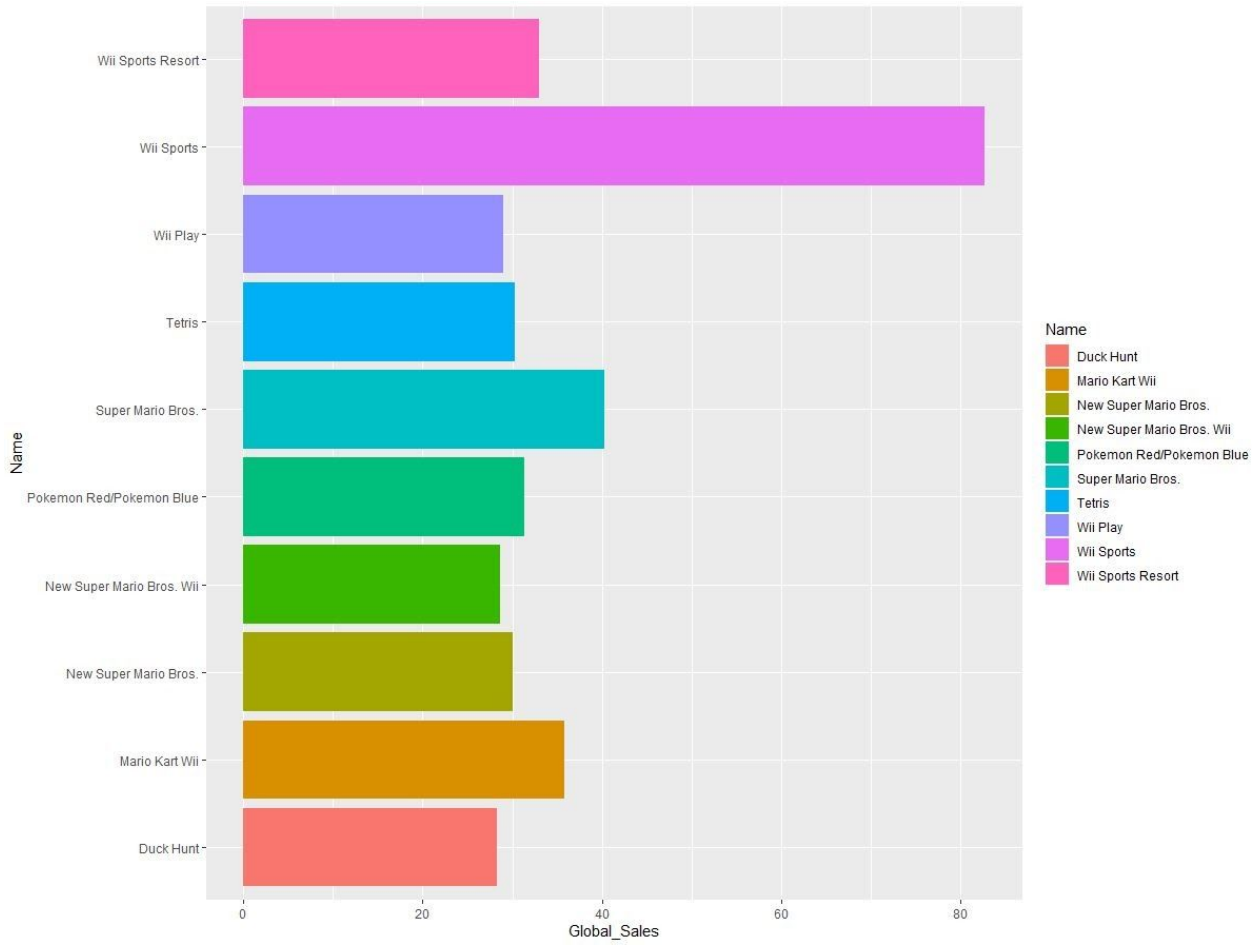
กราฟแสดง 10 อันดับผู้จัดจำหน่ายที่ขายเกมได้มากที่สุด




```
FilterGame %>% select(Publisher,Global_Sales) %>%  group_by(Publisher) %>% summarise(sum=sum(Global_Sales)) %>%
  arrange(desc(sum)) %>% head(10) %>%
  ggplot()+geom_col(mapping=aes(fill=Publisher,x=Publisher,y=sum)) + coord_flip()
```

กราฟนี้คือกราฟ Barchart ที่แสดงให้เห็นจำนวนเกมที่ขายได้ (ในหน่วยล้านชุด) ของแต่ละผู้จัดหน่ายตั้งแต่ปี ค.ศ.1980 ถึงปี ค.ศ.2016 โดยจากกราฟสามารถบอกได้ว่าผู้จัดจำหน่าย Nintendo ขายเกมได้จำนวนมากที่สุดรองลงมาคือ Electonics Arts

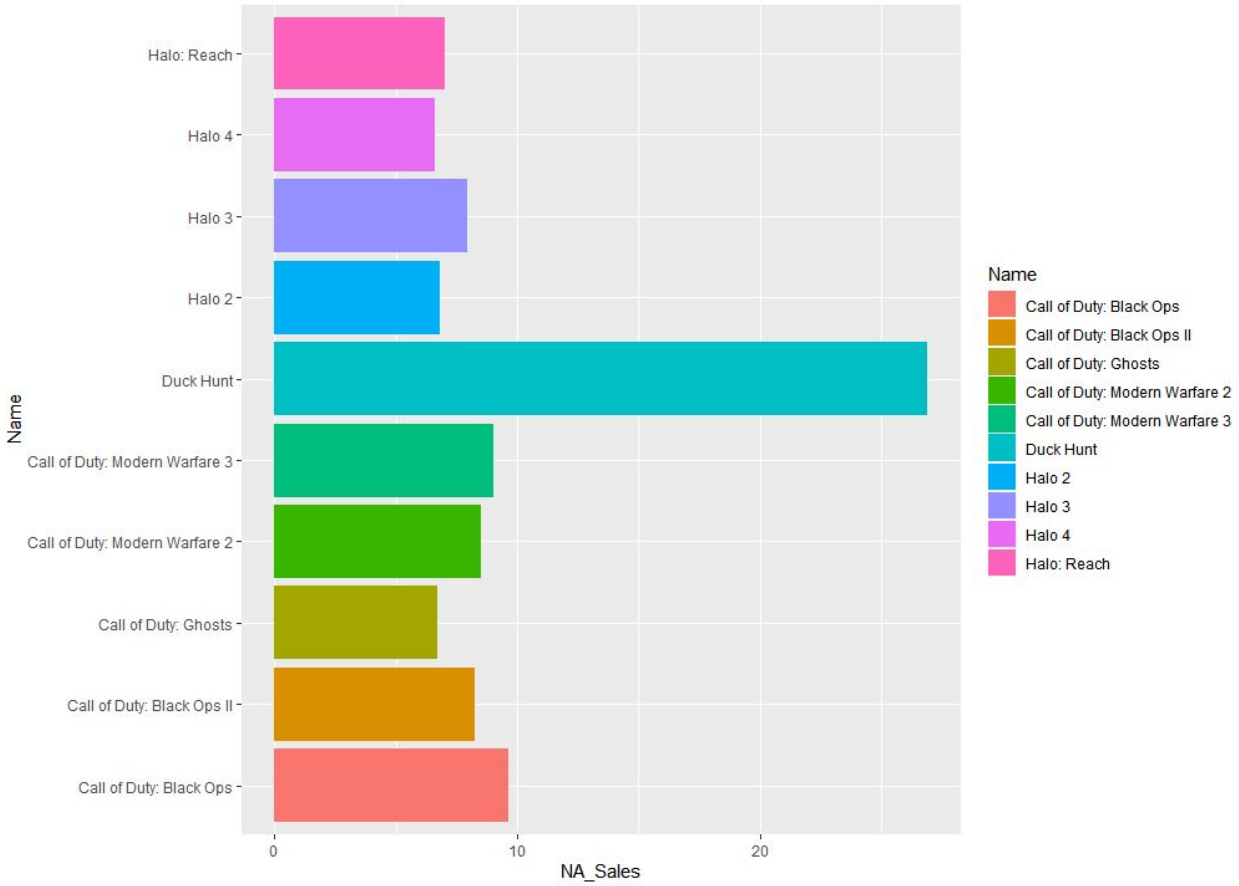
กราฟแสดง 10 อันดับเกมที่ขายดีที่สุด




```
FilterGame %>% select(Name,Global_Sales) %>% arrange(desc(Global_Sales)) %>% head(10) %>%
  ggplot()+geom_col(mapping=aes(fill=Name,x=Name,y=Global_Sales)) + coord_flip()
```

กราฟนี้คือกราฟ Barchart ที่แสดงให้เห็นจำนวนยอดขายเกม(หน่วย : ล้านชุด)สูงสุด 10 อันดับแรกของเกมทุกประเภททั่วโลกบนทุกแพลตฟอร์มตั้งแต่ปี ค.ศ.1980 ถึงปี ค.ศ.2016 โดยจากกราฟสามารถบอกได้ว่าเกม Wii Sports ขายดีที่สุดรองลงมาคือเกม Super Mario Bros

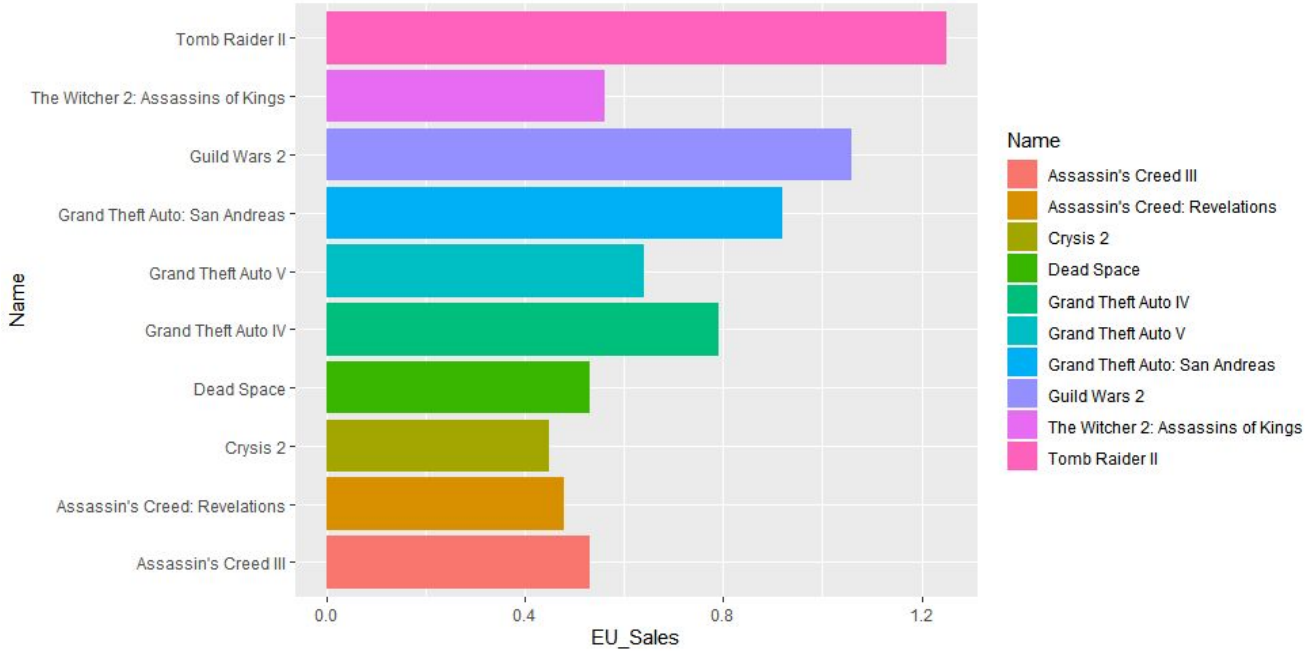
กราฟแสดง 10 อัพดับเกมประเภท Shooter ที่ขายดีที่สุดในอเมริกาเหนือ




```
FilterGame%>%select(Name,NA_Sales,Genre) %>% filter(Genre=="Shooter") %>%
  arrange(desc(NA_Sales)) %>% head(10) %>%
  ggplot()+geom_col(mapping=aes(fill=Name,x=Name,y=NA_Sales)) + coord_flip()
```

กราฟนี้คือกราฟ Barchart ที่แสดงให้เห็นจำนวนยอดขาย (หน่วย : ล้านชุด)สูงสุดจำนวน 10 อันดับแรกของเกมประเภท Shooter บนทุกแพลตฟอร์ม ในภูมิภาคอเมริกาเหนือ(NA)ตั้งแต่ปี ค.ศ .1980 ถึงปี ค.ศ.2016 โดยจากกราฟสามารถบอกได้ว่าเกม Duck Hunt ขายดีที่สุดรองลงมาคือเกม Call of Duty: Black Ops

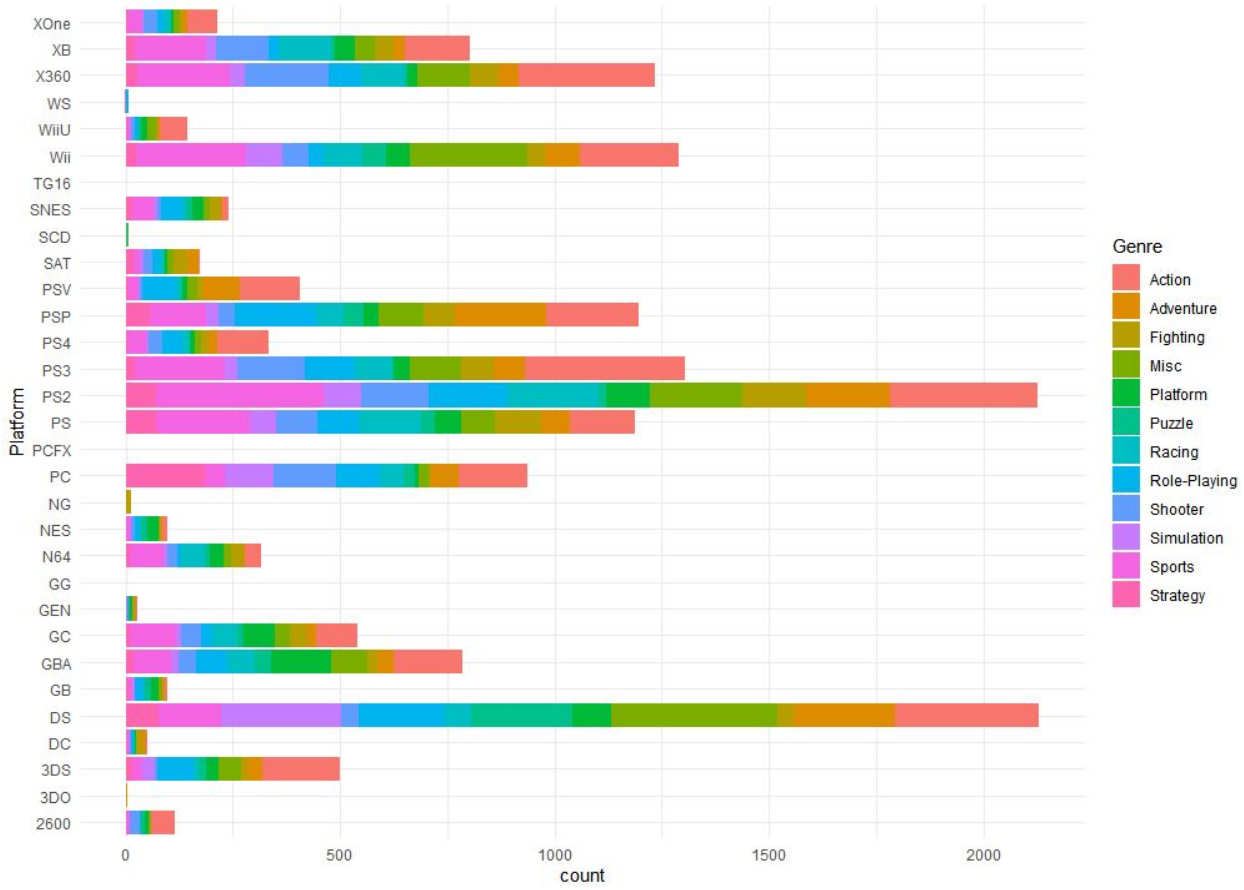
กราฟแสดง 10 อันดับเกมแพลตฟอร์ม PC ประเภท Action ที่ขายดีที่สุดใน EU




```
FilterGame %>% select(Name,EU_Sales,Genre,Platform) %>%  filter(Genre=="Action") %>%
  filter(Platform=="PC") %>% arrange(desc(EU_Sales)) %>% head(10) %>%
  ggplot()+geom_col(mapping = aes(fill=Name,x=Name,y=EU_Sales))+coord_flip()
```

กราฟนี้คือกราฟ Barchart ที่แสดงให้เห็นจำนวนยอดขายสูงสุด (หน่วย : ล้านชุด) 10 อันดับแรกของเกมประเภท Action ในภูมิภาคยุโรป (EU) บนแพลตฟอร์มประเภท PC (Personal Computer) ตั้งแต่ปี ค.ศ.1980 ถึงปี ค.ศ.2016 โดยจากกราฟสามารถบอกได้ว่าเกม Tomb Raider II ขายดีที่สุดรองลงมาคือเกม Guild Wars 2

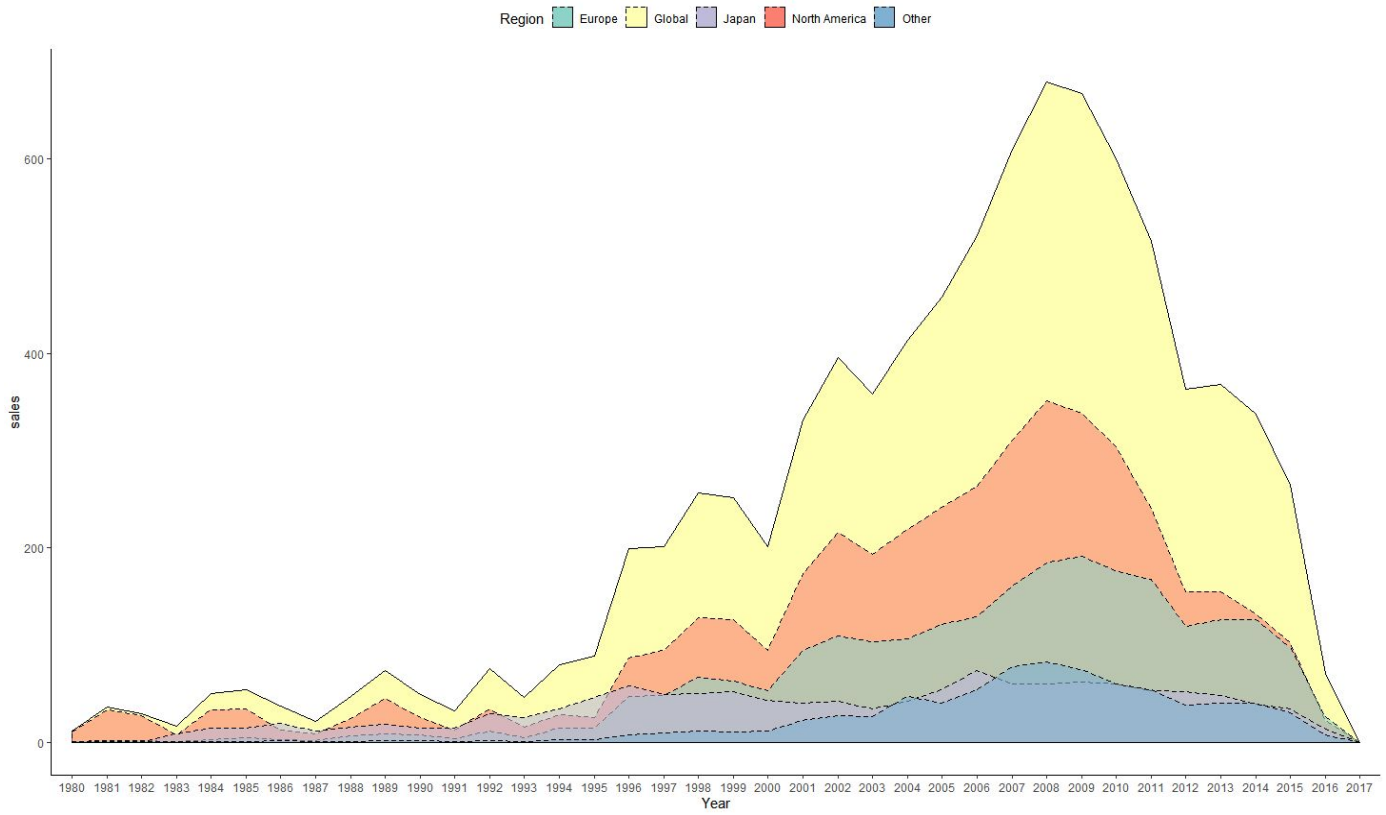
กราฟแสดงเกมแต่ละประเภทในแต่ละ platform




```
FilterGame %>% select(Platform,Genre) %>% group_by(Platform) %>%
  ggplot()+geom_bar(aes(x=Platform,fill=Genre),position = 'stack')+ coord_flip()+theme_minimal()
```

กราฟนี้คือกราฟ Bar Chart ที่แสดงข้อมูลจำนวนของเกมแต่ละประเภทที่ขายได้ในแต่ละ Platform โดยจากกราฟนั้นสามารถบอกได้ว่า Platform Ds (Nintendo DS) ขายเกมได้เป็นจำนวนมากที่สุดรองลงมาคือ Platform PS2(PlayStation 2) และ Platform ที่ขายได้น้อยที่สุดคือ TG16(TurboGrafx-16) และ PCFX

กราฟแสดงยอดขายแต่ละภูมิภาค




```
FilterGame %>% group_by(Year) %>%
  summarise(sales = sum(Global_Sales),nasales = sum(NA_Sales),eusales = sum(EU_Sales),jpsales = sum(JP_Sales),othsales = sum(Other_Sales))%>%
  ggplot(mapping = aes(x = Year,group = 1)) +
  geom_area(aes(y = sales,fill = "Global"),color = "black") +
  geom_area(aes(y = nasales,fill = "North America"),color = "black", linetype="dashed", alpha=0.6) + theme_classic()+
  geom_area(aes(y = eusales,fill = "Europe"),color = "black", linetype="dashed", alpha=0.6) + theme_classic()+
  geom_area(aes(y = jpsales,fill = "Japan"),color = "black", linetype ="dashed", alpha=0.6) + theme_classic()+
  geom_area(aes(y = othsales,fill = "Other"),color = "black", alpha=0.6,linetype ="dashed") + theme_classic()+
  labs(x = "Year",y = "sales",fill = "Region") +
  scale_fill_brewer(palette="Set3") + theme(legend.position="top")
```

กราฟ Area นี้เป็นกราฟที่แสดง ยอดขายเกมในแต่ละปีตั้งแต่ปี ค.ศ. 1980 ถึงปี ค.ศ. 2017 โดยแบ่งตามภูมิภาค กราฟสามารถบอกได้ว่า ในแต่ละภูมิภาค ยอดขายมีแนวโน้มไปทางไหน และสามารถเปรียบเทียบระหว่างภูมิภาคได้ว่า ปีนั้น ภูมิภาคไหนขายเกมได้มากกว่ากันบ้าง

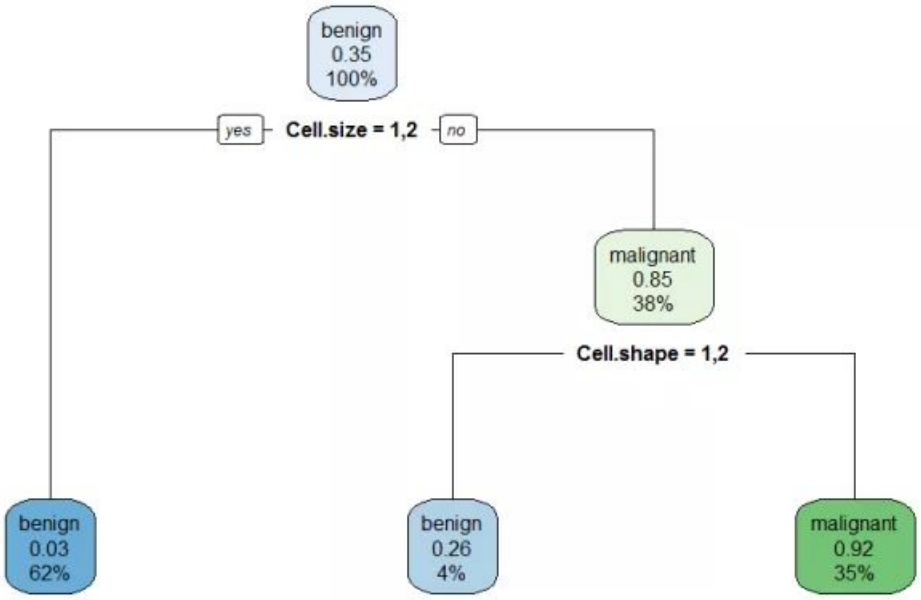
Model Explanation

ในการสร้างโมเดลสำหรับการทำนายเราเลือกใช้แบบ Decision Tree เนื่องจากเราจะใช้ข้อมูลในส่วนของ Platform,Genre,Publisher ซึ่งเป็นข้อมูลประเภท category ทั้งหมด โดยเรานำมาใช้ในการทำนายยอดขายของเกมว่าถ้าเกมที่ออกมาใหม่จะสามารถทำยอดขายได้เกิน 2 แสนชุดหรือไม่

Decision tree model

Machine Learning Model Classification ตัวหนึ่งที่สามารถอธิบายได้ว่าทำไมถึงแบ่งเป็นคลาสนี้ ทำไมต้องเป็นคลาสนี้ สามารถอธิบายได้ด้วยรูปแบบของ “TREE”
นั้นคือมี Node พ่อเป็นตัวตั้งคำถามว่าใช่หรือไม่ Node ลูกตัวแรกจะเป็นใช่ อีกตัวจะเป็นไม่ใช่ โดยปัจจัยสำคัญในการสร้างโมเดลนี้คือ “ความลึกของต้นไม้” ยิ่งต้นไม้ลึก (มีจำนวนชั้นมาก) ก็จะตั้งคำถามได้ละเอียดมากขึ้น แต่ก็จะมี overfit มากขึ้นทำให้เกิดความซับซ้อนมากขึ้น แต่ถ้าจำนวนชั้นน้อยเกินไป ก็จะทำนายได้ไม่แม่นยำพอที่จะใช้งาน

ตัวอย่าง



Modeling Implementation

Decision Tree model

ขั้นตอนเตรียมข้อมูล

- mutate ยอดขายที่มากกว่า 200,000 ชุด เป็นคลาส success = y

```
FilterGame %>% mutate(success = cut(Global_Sales,
                                    breaks = c(0,0.2,Inf),
                                    labels = c("n","y"))) -> FilterGame
```

Global_Sales	success
0.17	n
0.23	y
0.16	n
0.17	n
1.27	y
0.14	n
0.68	y
0.46	y
0.18	n
0.03	n
0.80	y

สร้าง decision tree ด้วยวิธี Hold-Out Method โดยแบ่งข้อเป็น 2 ส่วนคือ ข้อมูลส่วน Test 20 % และข้อมูลส่วน Train 80%

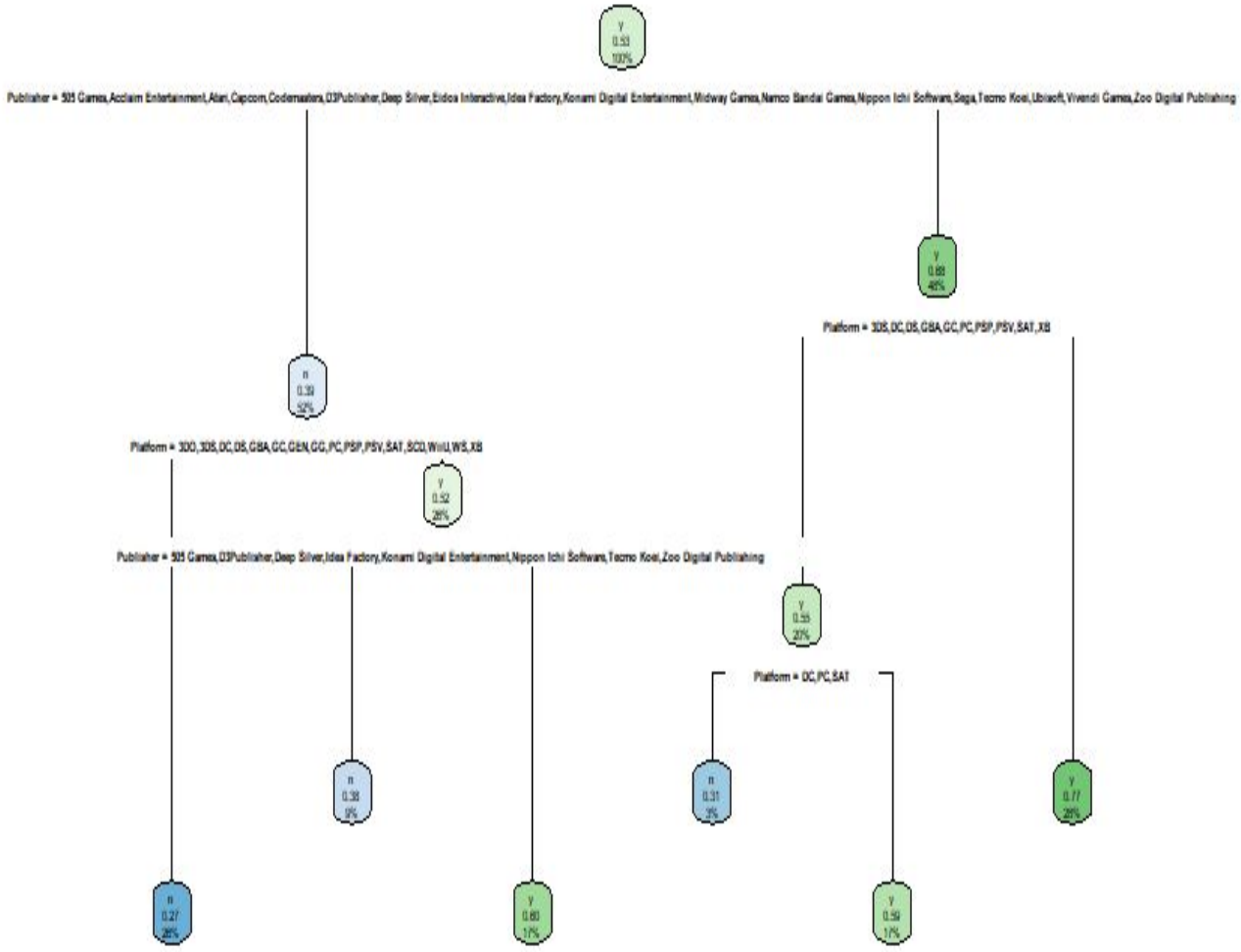
```
FilterGame %>% select(-NA_Sales,-Name,-EU_Sales,-JP_Sales,-Other_Sales,-Global_Sales) -> GameData

set.seed(555)
test_ind <- sample(nrow(GameData),
                   0.2*nrow(GameData))
Data_training <- GameData[-test_ind,]
Data_testing <- GameData[test_ind,]
```

ใช้คำสั่ง rpart ในการสร้าง tree

```
tree <- rpart(success ~ ., data = Data_training)
rpart.plot(tree)
```

รูป Tree plot



Variable Importance

แสดงข้อมูลตัวแปรที่มีผลต่อ Tree มากที่สุดโดยจากข้อมูลสามารถบอกได้ว่า Publisher ส่งผลต่อยอดขายเกมมากที่สุดตัวแปรที่ส่งผลรองลงมาคือ Platform

```
> tree$variable.importance
Publisher  Platform     Genre      Year
462.52501 341.52441  67.69058  19.82499
```

Confusion Matrix

สร้าง confusion Matrix เพื่อทำนายค่าจาก decision tree และดูประสิทธิภาพของ Model

```
res<-predict(tree,Data_testing,type = "class")
```

```
confusionMatrix(res,
                Data_testing$success,
                positive = "y",
                mode = "prec_recall")
```

```
Confusion Matrix and Statistics

          Reference
Prediction   n   y
         n 621 247
         y 493 986

               Accuracy : 0.6847
                 95% CI : (0.6655, 0.7035)
    No Information Rate : 0.5254
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.361

 Mcnemar's Test P-Value : < 2.2e-16

              Precision : 0.6667
                 Recall : 0.7997
                     F1 : 0.7271
             Prevalence : 0.5254
         Detection Rate : 0.4201
   Detection Prevalence : 0.6302
      Balanced Accuracy : 0.6786

       'Positive' Class : y
```

Evaluation

```
Confusion Matrix and Statistics

          Reference
Prediction   n   y
         n 621 247
         y 493 986

               Accuracy : 0.6847
                 95% CI : (0.6655, 0.7035)
    No Information Rate : 0.5254
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.361

 Mcnemar's Test P-Value : < 2.2e-16

              Precision : 0.6667
                 Recall : 0.7997
                     F1 : 0.7271
             Prevalence : 0.5254
         Detection Rate : 0.4201
   Detection Prevalence : 0.6302
      Balanced Accuracy : 0.6786

       'Positive' Class : y
```

จาก Confusion Matrix จะได้

ค่า Accuracy คือความแม่นยำของการทำนายทั้งหมด มีความถูกต้อง 68.47%

ค่า No information Rate 0.5254

Accuracy > No Information Rate แสดงว่า โมเดลนี้สามารถทำนายได้ดีขึ้นอย่างมีนัยสำคัญ

ค่า Kappa 0.361 (fair)

ค่า Precision(TP/(TP+FP)) หมายถึงการทำนายเกมที่จะขายได้เกิน 200,000 ชุดถูกต้องเป็นสัดส่วน 66.67% จากที่ทำนายว่าจะขายเกิน 200,000 ชุดทั้งหมด

ค่า Recall (TP/(TP+FN)) ทำนายเกมที่จะขายได้มากกว่า 200,000 ชุดถูกต้องเป็นสัดส่วน 79.97% จากในความเป็นจริงที่จะขายเกิน 200,000 ชุดทั้งหมด

จากที่เราได้ทำการเลือกให้ที่จำนวน 200,000 ชุด เป็นตัวแบ่งคลาส y กับ n ทำให้คลาสที่ใหญ่ที่สุด เป็นคลาส y ทำให้ Prevalence มีค่าเท่ากับค่า NIR คือ 0.5254 คือสัดส่วนค่าที่จะขายได้มากกว่า 200000 ชุด ในความเป็นจริงจากทั้งหมด (Actual y/ All)

ค่า Detection Prevalence มีค่าเป็น 0.6302 คือสัดส่วนการทำนายว่าจะขายได้มากกว่า 200,000 ชุดจากทั้งหมด (TP+FP/All)

ต่อมา ค่า Detection Rate มีค่า 0.4201 คือการทำนายว่าจะขายได้มากกว่า 200,000 ชุด ได้ถูกต้องจากทั้งหมด

ค่า Balance accuracy มีค่า 0.6786 และจะสังเกตได้ว่า ค่า Balance accuracy ใกล้เคียงกับ Accuracy คือ 0.6847 แสดงว่า คลาสที่เราทำนาย เป็น Balanced Classes ทั้งนี้เกิดจากที่เลือก 200,000 ชุด เป็นตัวแบ่งคลาส

Discussion and Conclusion

จากโมเดลที่กลุ่มพวกเราสร้างขึ้นในการทำนายว่าเกมใดๆจะสามารถจำหน่ายได้ตั้งแต่ 200,000 ชุดขึ้นไป ก่อนที่จะเริ่มสร้างโมเดลเราได้ทำการ clean data หรือก็คือ Data Preparation เพื่อให้ได้ข้อมูลที่ง่ายต่อการวิเคราะห์ และ ตรงตามที่กลุ่มของพวกเราต้องการ ต่อมาได้นำข้อมูลมาทำ visualization เพื่อดูค่าต่างๆที่ปรากฏในข้อมูล หลังจากนั้นทำได้นำข้อมูลไปสร้างโมเดลโดยกลุ่มของพวกเราเลือกใช้ Decision Tree ในการทำโมเดล โดยในตอนแรก ข้อมูลเรา มีแต่ Categorical ของ data และ numerical ที่เป็นยอดขายของแต่ละภูมิภาค เราจึงแปลง Global_sales ให้กลายเป็น Category คือ Success = y หรือ n เพื่อทำนาย

ต่อมาเมื่อทดลองทำตอนแรกที่เรายังไม่ได้ Filter ผู้ผลิตเกมที่ผลิตน้อยกว่า 100 เกมออก ทำให้ Publisher ยังคงมีจำนวนมากเกินไป เราจึงได้ filter เฉพาะผู้ผลิตรายใหญ่ที่ผลิตเกมออกมาคือผลิตมากกว่า 100 เกมขึ้นไปเท่านั้น เพื่อลด Outlier

ผลการทำนายจากการทำ Decosion Tree เมื่อดูจาก Confusion Matrix มีค่า p-value < 2.2e-16 หมายความว่าโมเดลของเราทำให้การทำนายดีขึ้นอย่างมีนัยสำคัญและ
มีค่า Precision, Recall ที่น่าพอใจ