Homework 3

2023-10-30

Task 1

```r
# Load necessary libraries
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyr)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

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

```r
library(viridis)
```

```
## Loading required package: viridisLite
```

```r
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```r
library(RColorBrewer)
library(grid)

# Read the data
vgsales <- read.csv("vgsales.csv")

# Extract games that are available on the PC platform
pc_games <- vgsales %>% filter(Platform == "PC")

# Identify the top 5 publishers with the highest number of PC games
pc_top_5_count <- pc_games %>%
  group_by(Publisher) %>%
  summarize(Count = n()) %>%
  arrange(desc(Count)) %>%
  head(5)

# Print the top 5 publishers
print(pc_top_5_count)
```

```
## # A tibble: 5 x 2
##   Publisher          Count
##   <chr>              <int>
## 1 Electronic Arts      148
## 2 Ubisoft               76
## 3 Activision            56
## 4 Take-Two Interactive  45
## 5 Sega                  33
```

```r
# Create a summary for the entire data
all_games <- vgsales %>%
  group_by(Genre) %>%
  summarize(Count = n(), .groups = "drop")

# Filter out games from the top 5 publishers
pc_top_5 <- pc_games %>% filter(Publisher %in% pc_top_5_count$Publisher)

# Create histograms
all_plot <- ggplot(all_games, aes(x = Genre, y = Count)) +
  geom_bar(stat = "identity") +
  labs(title = "All Platforms", x = NULL, y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())

pc_plot <- ggplot(pc_top_5, aes(x = Genre)) +
  geom_bar(stat = "count") +
  labs(title = "Top 5 PC Platforms", x = NULL, y = NULL) + theme(element_blank()) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())
```
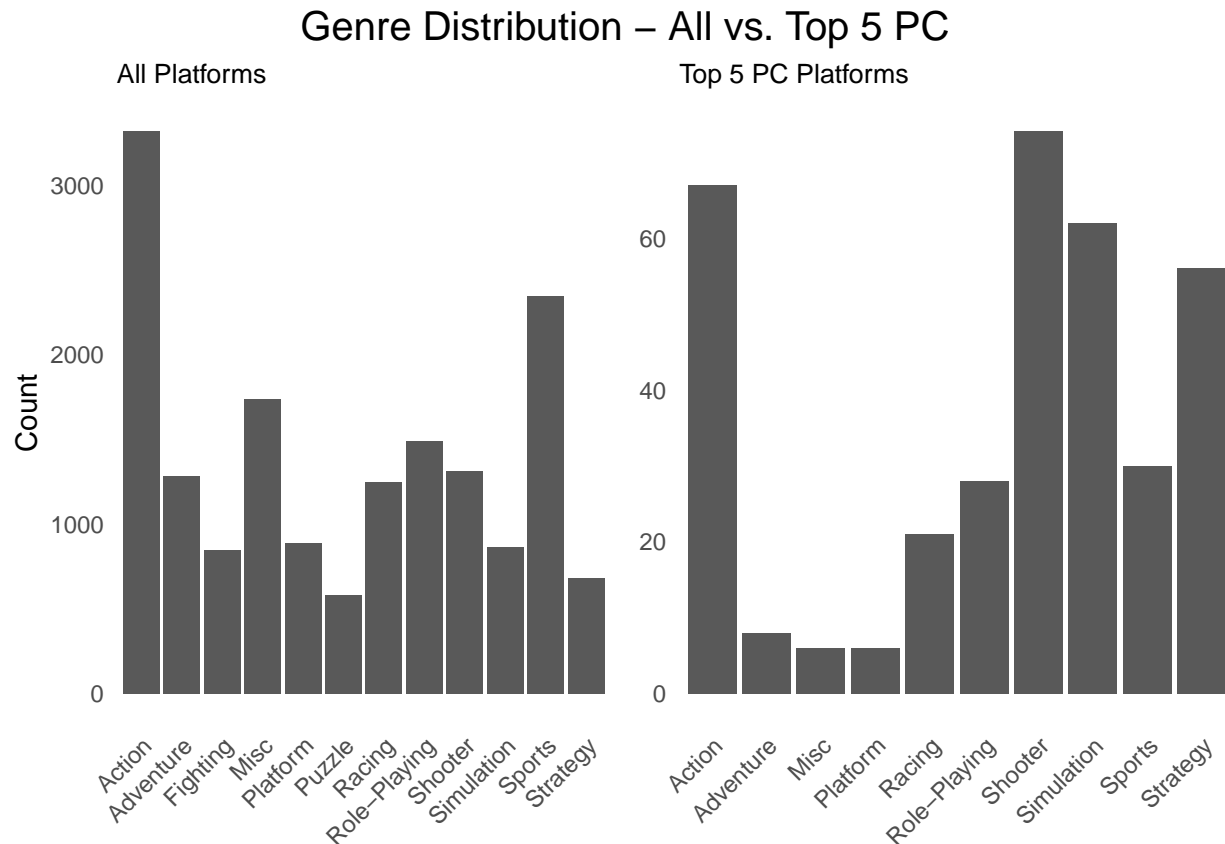
```
all_plot <- all_plot + theme(plot.title = element_text(size = 10))
pc_plot <- pc_plot + theme(plot.title = element_text(size = 10))

# Arrange plots side by side with a bigger title
grid.arrange(all_plot, pc_plot, ncol = 2, top = textGrob("Genre Distribution – All vs. Top 5 PC", gp = g
```

Genre Distribution – All vs. Top 5 PC

All Platforms: The count for genres spans from around 1000 to over 3000. Top 5 PC Platforms: The count for genres lies between 0 and 60 because we did filtering.

In All Platforms the "Action" genre as the most dominant. On the other hand "Action" is in the second place after "Shooter" genre on Top 5 PC Platforms. For the top 5 PC platforms, genres like "Shooter," "Action," "Simulation," and "Strategy" are quite popular. This could suggest that PC gamers have a preference for these genres, possibly because PCs can offer more complex controls and richer graphics. We do not see genres like "Fighting" and "Puzzle" on the top 5 PC platforms compared to all platforms. They seem to be not popular for top 5 PC platforms.

```
# Filter data
top_5_data <- vgsales %>%
  filter(Publisher %in% pc_top_5_count$Publisher) %>%
  group_by(Year, Publisher) %>%
  summarize(Total_Global_Sales = sum(Global_Sales, na.rm = TRUE), .groups = "drop")

# Convert the Year column to numeric and remove NAs
suppressWarnings({
  top_5_data$Year <- as.numeric(as.character(top_5_data$Year))
})
```

```
top_5_data <- top_5_data %>% filter(!is.na(Year))

# Plot the line chart
line_chart <- ggplot(top_5_data, aes(x = Year, y = Total_Global_Sales, color = Publisher)) +
  geom_line(aes(group = Publisher), size = 1) +
  labs(title="Global Sales over Years for Top 5 Publishers",
       x = NULL,
       y = NULL) +
  scale_x_continuous(breaks = seq(1980, 2020, by = 5)) +
  theme_minimal() +
  scale_color_brewer(palette ="Set1", name ="Publisher") +
  theme(legend.position ="top", legend.title=element_blank())
```
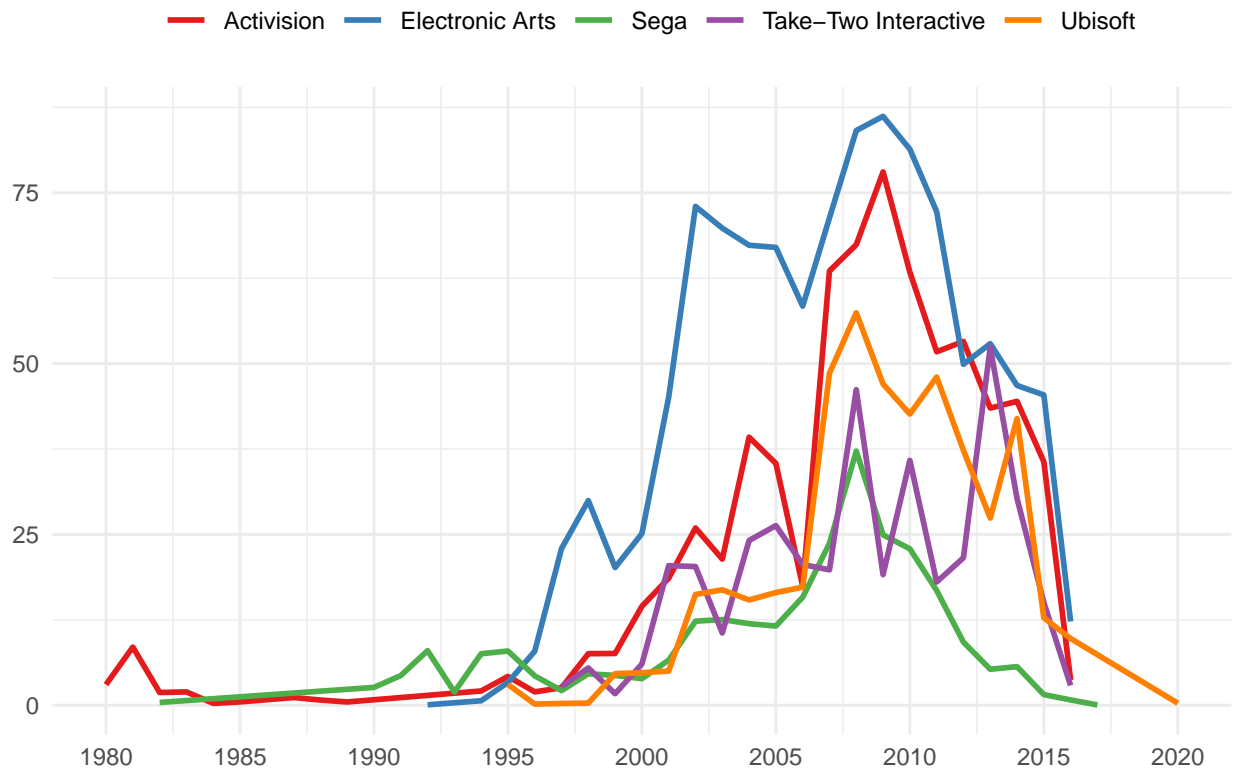
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
# Print the line chart
print(line_chart)
```


Global Sales over Years for Top 5 Publishers

The graph provides a comprehensive overview of the global sales trends for five major game publishers over four decades. It captures the rise and fall in sales, indicating both individual publisher strategies and industry-wide influences that have shaped the gaming market. For example, Activision and Electronic Arts publishers show significant growth, especially after 2005. And Ubisoft's growth is gradual.

```r
# Create a cumulative sum of Total_Global_Sales for each publisher
top_5_data <- top_5_data %>%
  arrange(Year) %>%
  group_by(Publisher) %>%
  mutate(Cumulative_Sales = cumsum(Total_Global_Sales))

# Plot the cumulative line chart
cumulative_chart <- ggplot(top_5_data, aes(x = Year, y = Cumulative_Sales, color = Publisher)) +
  geom_line(aes(group = Publisher), size = 1) +
  labs(title="Cumulative Global Sales over Years for Top 5 Publishers",
       x = NULL,
       y = NULL) +
  scale_x_continuous(breaks = seq(1980, 2020, by=5)) +
  theme_minimal() +
  scale_color_brewer(palette ="Set1", name ="Publisher") +
  theme(legend.position ="top", legend.title = element_blank())

# Print the cumulative line chart
print(cumulative_chart)
```

Cumulative Global Sales over Years for Top 5 Publishers
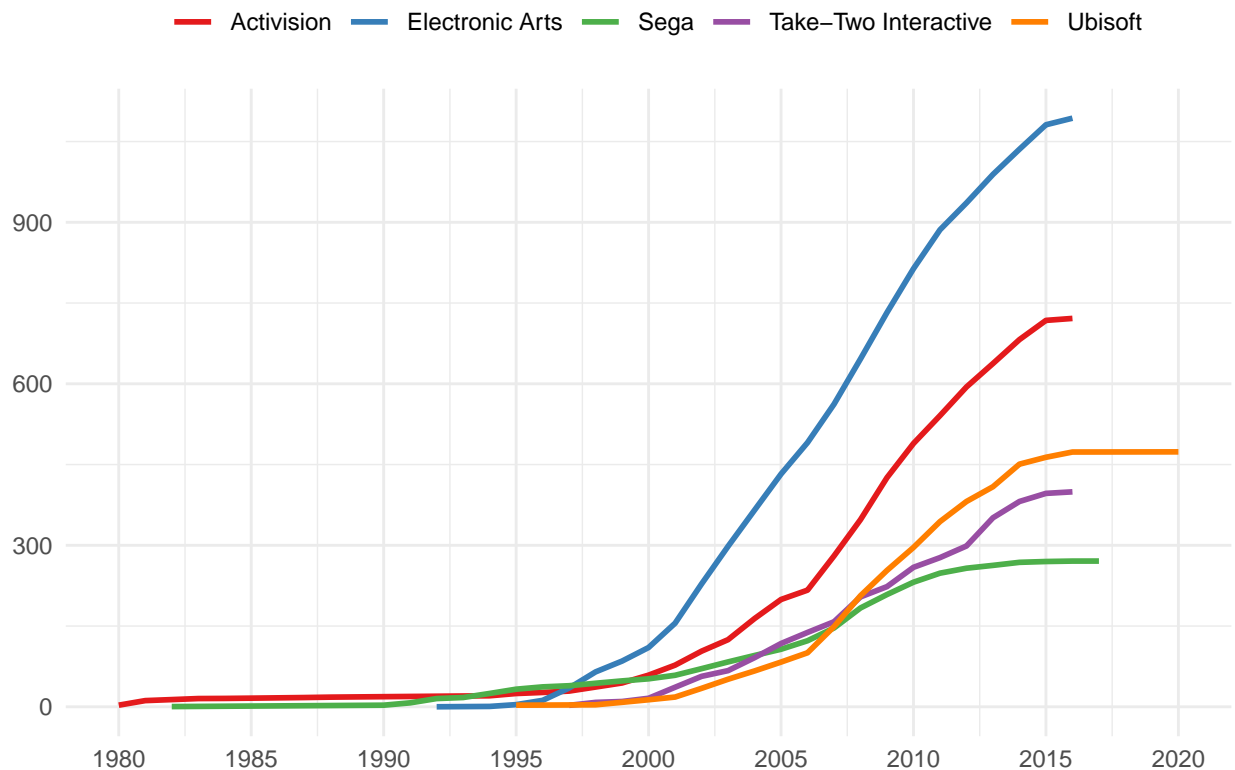
Electronic Arts has the highest cumulative sales, approaching 900 by 2020.

Activision follows closely, reaching just above 600 by 2020.

Ubisoft and Take-Two Interactive (Purple Line) have a similar growth trajectory, with sales around 500 and slightly below 400 respectively by 2020.

Sega has the lowest cumulative sales among the group, with a relatively flat growth after the early 2000s.

Task 2

```r
# Filter and aggregate data for PC by Genre
pc_genre_counts <- vgsales %>%
  filter(Platform == "PC") %>%
  group_by(Genre) %>%
  summarize(Count = n()) %>%
  arrange(desc(Count))

pc_genre_counts
```

```
## # A tibble: 12 x 2
##    Genre        Count
##    <chr>        <int>
##  1 Strategy       188
##  2 Action         165
##  3 Shooter        148
##  4 Simulation     115
##  5 Role-Playing   104
##  6 Adventure       65
##  7 Racing          60
##  8 Sports          49
##  9 Puzzle          25
## 10 Misc            24
## 11 Platform        11
## 12 Fighting         6
```
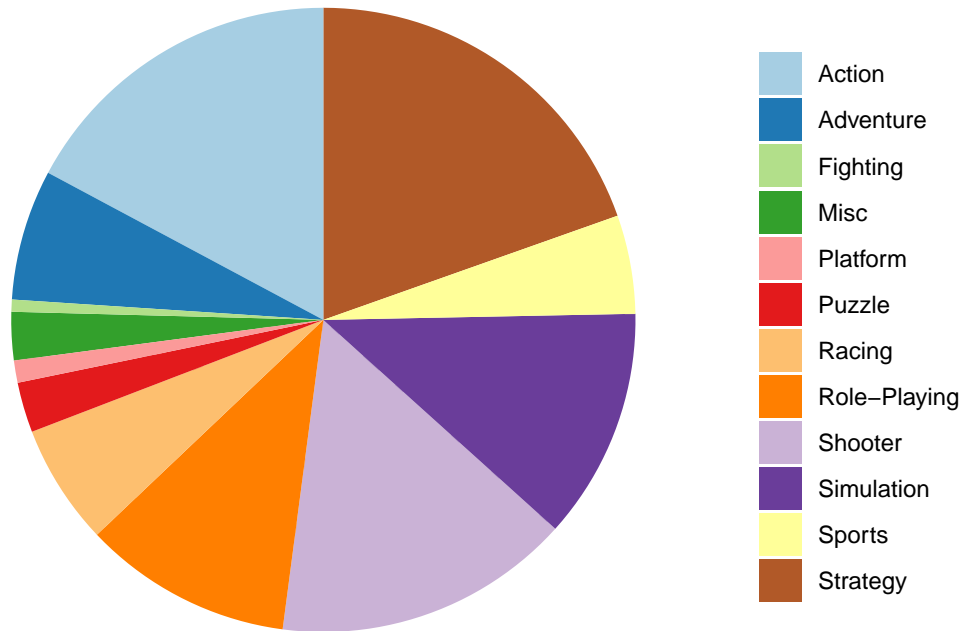
```r
# Define a custom color palette with 12 colors
colors <- brewer.pal(n = 12, name = "Paired")

# Create pie chart using the custom color palette
pie_chart <- ggplot(pc_genre_counts, aes(x = "", y = Count, fill = Genre)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  scale_fill_manual(values = colors) +
  labs(title = "Distribution of PC Games by Genres", fill = "Genre") +
  theme_minimal() +
  theme(axis.text.x = element_blank(),
        axis.ticks = element_blank(),
        axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        plot.title = element_text(hjust = 0.5),
        panel.grid = element_blank(),
        legend.title = element_blank())

print(pie_chart)
```

Distribution of PC Games by Genres

Legend:
- Action
- Adventure
- Fighting
- Misc
- Platform
- Puzzle
- Racing
- Role–Playing
- Shooter
- Simulation
- Sports
- Strategy

```r
print(dominating_genre_pc <- pc_genre_counts[[1,"Genre"]])
```

```
## [1] "Strategy"
```

```r
colors_pc <- c("#a6cee3", "#ff7f00", "#cab2d6", "#6a3d9a", "#b15928")

# Filter vgsales to include only PC games, group by Genre, and count the number of games
top_5_pc_genres <- vgsales %>%
  filter(Platform == "PC") %>%
  group_by(Genre) %>%
  summarize(Count = n()) %>%
  arrange(desc(Count)) %>%
  head(5)

# Create a pie chart for PC
pc_pie_chart <- ggplot(top_5_pc_genres, aes(x = "", y = Count, fill = Genre)) +
  geom_bar(stat = "identity", width = 1) +
  geom_text(aes(label = Count), position = position_stack(vjust = 0.5)) +
  coord_polar(theta = "y") +
  scale_fill_manual(values = colors_pc) +
  labs(title = "Top 5 Genres - PC Games") +
  theme_minimal() +
  theme(axis.text.x = element_blank(),
        axis.ticks = element_blank(),
        axis.title.x = element_blank(),
```

```
        axis.title.y = element_blank(),
        plot.title = element_text(hjust = 0.5),
        panel.grid = element_blank(),
        legend.title = element_blank())

# Print the pie chart
print(pc_pie_chart)
```

Top 5 Genres – PC Games


```
# Identify the dominating genre for PC
print(dominating_genre_pc <- top_5_pc_genres[[1,"Genre"]])
```

```
## [1] "Strategy"
```

Top 5 Geners PC Games is a subset with Action, Role-Playing, Shooter, Simulation, Strategy from All PC games. Strategy is Dominant.

```
colors_x360 <- c("#a6cee3", "#33a02c", "#fdbf6f", "#cab2d6", "#ffff99")

# Filter and aggregate data for PS4
top_5_x360_genres <- vgsales %>%
  filter(Platform == "X360") %>%
  group_by(Genre) %>%
  summarize(Count = n()) %>%
  arrange(desc(Count)) %>%
```
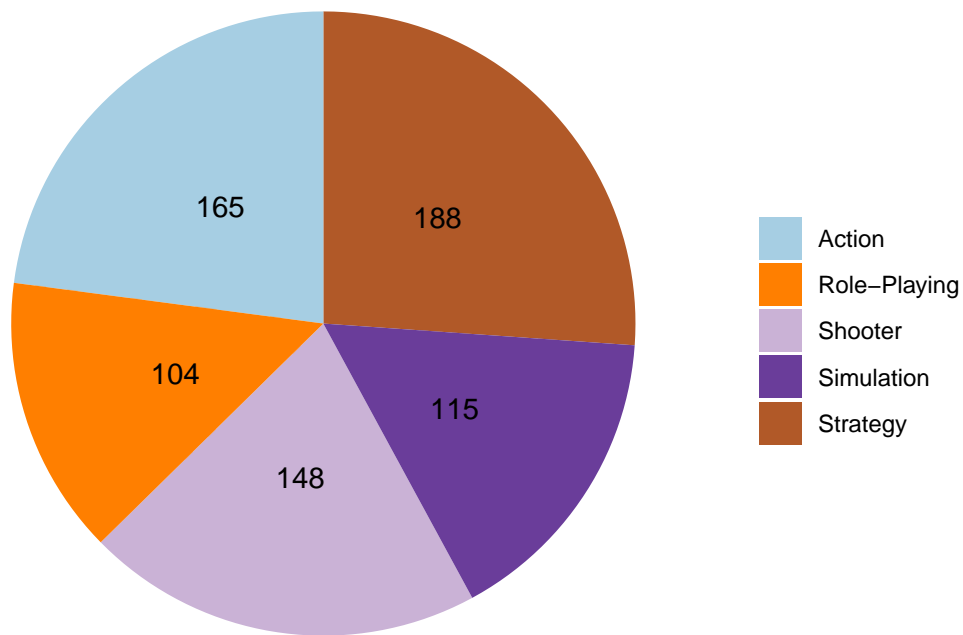
```
  head(5)

# Create a pie chart for PS4
x360_pie_chart <- ggplot(top_5_x360_genres, aes(x = "", y = Count, fill = Genre)) +
  geom_bar(stat = "identity", width = 1) +
  geom_text(aes(label = Count), position = position_stack(vjust = 0.5)) +
  coord_polar(theta = "y") +
  scale_fill_manual(values = colors_x360) +
  labs(title = "Top 5 Genres - X360 Games") +
  theme_minimal() +
  theme(axis.text.x = element_blank(),
        axis.ticks = element_blank(),
        axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        plot.title = element_text(hjust = 0.5),
        panel.grid = element_blank(),
        legend.title = element_blank())

# Print the pie chart
print(x360_pie_chart)
```

Top 5 Genres – X360 Games
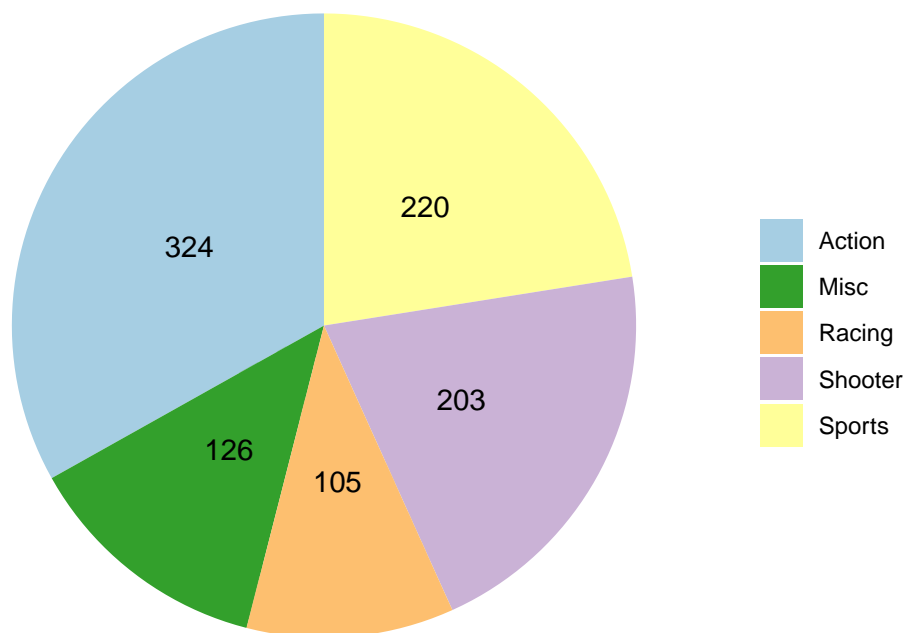

```
# Identify the dominating genre for X360
print(dominating_genre_x360 <- top_5_x360_genres[[1,"Genre"]])
```

```
## [1] "Action"
```

For X360 Games Action, Sports and Shooter have more proportions. Both X360 and PC game distributions highlight the popularity of the 'Action' and 'Shooter' genres. the "Action" and "Shooter" genres have higher proportions for X360 games than in the Top 5 PC games. For example, this can indicate a few things: Xbox 360 players may have a stronger preference for fast-paced and direct engagement games, typical of the "Action" and "Shooter" genres. The "Sports" genre is more popular in the X360 distribution but doesn't appear in the Top 5 PC games. This could be because sports games are often more enjoyable on consoles due to their multiplayer and local co-op features.

```r
# Get the top 15 publishers
top_15_publishers <- vgsales %>%
  filter(Platform == "PC") %>%
  group_by(Publisher) %>%
  summarise(Count = n()) %>%
  arrange(-Count) %>%
  head(15) %>%
  pull(Publisher)

# Filter the data to only include those top publishers
filtered_data <- vgsales %>%
  filter(Publisher %in% top_15_publishers)

# Create a table to count games by publisher and genre
count_table <- filtered_data %>%
  group_by(Publisher, Genre) %>%
  summarise(Count = n(), .groups = 'drop') %>%
  pivot_wider(names_from = Genre, values_from = Count, values_fill = 0)

# Convert wide format back to long format for plotting
count_table_long <- count_table %>%
  pivot_longer(cols = -Publisher, names_to = "Genre", values_to = "Count")
```
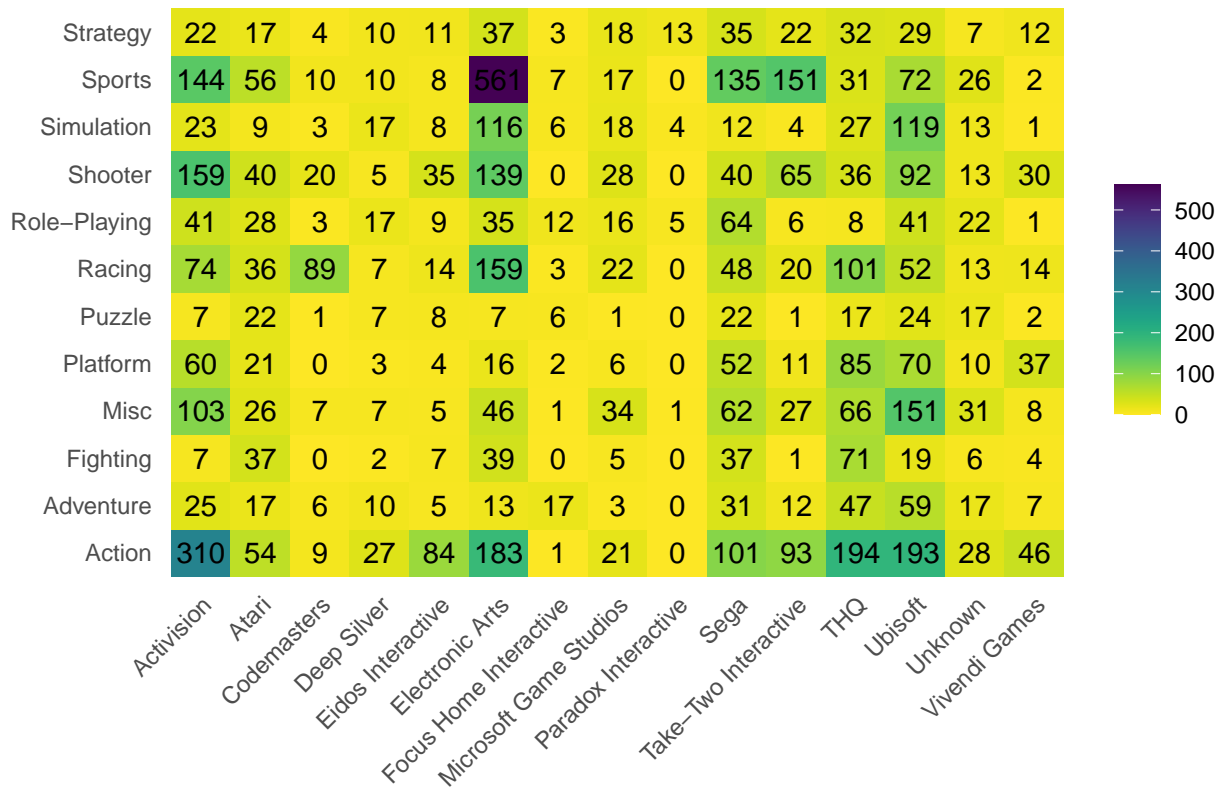
```r
# Plot the Heatmap
ggplot(count_table_long, aes(x = Publisher, y = Genre, fill = Count)) +
  geom_tile() +
  geom_text(aes(label=Count)) +
  scale_fill_viridis(direction = -1) +
  labs(title = "Genres by Top 15 PC Game Publishers",
       x ="Publisher",
       y = "Genre",
       fill = "Number of Games") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.ticks = element_blank(),
        axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        plot.title = element_text(hjust = 0.5),
        panel.grid = element_blank(),
        legend.title = element_blank())
```

Genres by Top 15 PC Game Publishers

Genre	Activision	Atari	Codemasters	Deep Silver	Eidos Interactive	Electronic Arts	Focus Home Interactive	Microsoft Game Studios	Paradox Interactive	Sega	Take–Two Interactive	THQ	Ubisoft	Unknown	Vivendi Games
Strategy	22	17	4	10	11	37	3	18	13	35	22	32	29	7	12
Sports	144	56	10	10	8	561	7	17	0	135	151	31	72	26	2
Simulation	23	9	3	17	8	116	6	18	4	12	4	27	119	13	1
Shooter	159	40	20	5	35	139	0	28	0	40	65	36	92	13	30
Role–Playing	41	28	3	17	9	35	12	16	5	64	6	8	41	22	1
Racing	74	36	89	7	14	159	3	22	0	48	20	101	52	13	14
Puzzle	7	22	1	7	8	7	6	1	0	22	1	17	24	17	2
Platform	60	21	0	3	4	16	2	6	0	52	11	85	70	10	37
Misc	103	26	7	7	5	46	1	34	1	62	27	66	151	31	8
Fighting	7	37	0	2	7	39	0	5	0	37	1	71	19	6	4
Adventure	25	17	6	10	5	13	17	3	0	31	12	47	59	17	7
Action	310	54	9	27	84	183	1	21	0	101	93	194	193	28	46

A publisher is better if it has published more games, then the publishers with more colors tending towards the warmer side of the palette (representing higher game counts) across multiple genres are considered better. In this heatmap, publishers such as Activision, Electronic Arts, and Ubisoft show a diverse portfolio across multiple genres. In this heatmap, publishers such as Activision, Electronic Arts, and Ubisoft show a diverse portfolio across multiple genres. Electronic Arts seems to have a significant role across multiple genres, with the "Sports" genre being especially dominant, evidenced by the dark shade of purple count of 561. We see that "Action" stands out as a popular genre across multiple publishers such as Activision, Electronic Arts, THQ, Ubisoft. Ubisoft have a strong focus on "Action" games, as indicated by the high count of 193. THQ have a significant count of 194 in the "Action" genre. Activision have produced a high number of "Action" games, indicated by the count of 310.

Task 4

```r
df <- read.csv("MobileSmartphones.csv")
selected_brands <- df %>%
                select(Date, Apple, Samsung, Xiaomi, Oppo)
# Convert Date Column
selected_brands$Date <- ym(selected_brands$Date)

# Melt the Data
melted_data <- selected_brands %>%
            gather(key = "Brand", value = "MarketShare", -Date)

# Determine the start and end dates
start_date <- min(selected_brands$Date)
end_date <- max(selected_brands$Date)
```

```
# Plot
ggplot(melted_data, aes(x = Date, y = MarketShare, color = Brand, group = Brand)) +
  geom_line(size = 3) +
  labs(title ="Top 4 Smartphones with Highest Market Share over time",
       y = NULL,
       x = NULL,
       color="Smartphone") +
  geom_segment(aes(x = start_date, xend = end_date, y = 18.63, yend = 17.41), linetype = "dashed", colo
  geom_text(aes(x = start_date, y = 18.63, label = "18.63%"), size = 3.5, hjust = 0, vjust = 1.5, color
    geom_text(aes(x = end_date, y = 17.41, label = "17.41%"), size = 3.5, hjust = 1.7, vjust = 1.5, col
  geom_text(aes(x = median(selected_brands$Date), y = 20, label = "Tendency of market shares of Top 4")
  theme_minimal() +
  scale_x_date(date_breaks = "3 month", date_labels = "%b %Y") + # Format the x-axis for dates
  scale_y_continuous(labels = function(x) paste0(x, "%")) + # Format the y-axis
  theme(plot.title = element_text(hjust = 0.5),
        panel.grid.minor = element_blank())  # This line removes the minor grid lines
```


Top 4 Smartphones with Highest Market Share over time