Data Exploration of Video Game Sales

Author: Colin Beran

Dataset source: <https://www.kaggle.com/gregorut/videogamesales>

Kaggle notebook: <https://www.kaggle.com/colinberan/data-exploration-of-video-game-sales>

This is an exploratory data analysis of video game titles with over 100,000 total sales that I retrieved from the above listed source.

I began by loading a few useful libraries and importing the dataset. As a preprocessing step, I removed all rows with N/A values to avoid working with any incomplete data.

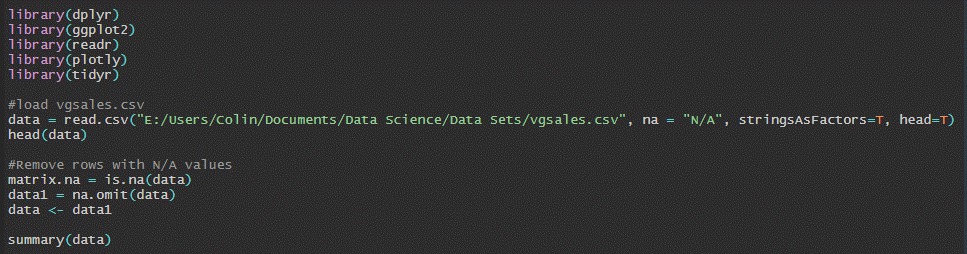


Fig. 1: R code for importing data and slight preprocessing.

I grouped all video games by the release year and noticed incomplete data for 2017 and 2020. Additionally, I wanted to focus on games released in the 21st century, so I created a subset of the data, filtering “Year” for 2000-2016.

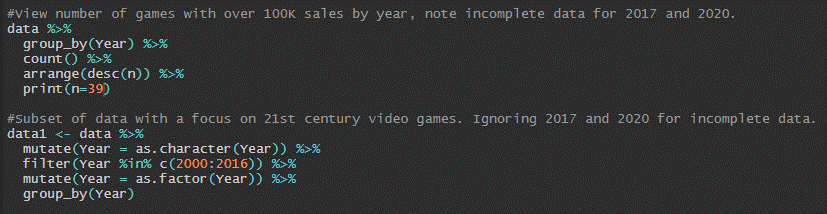
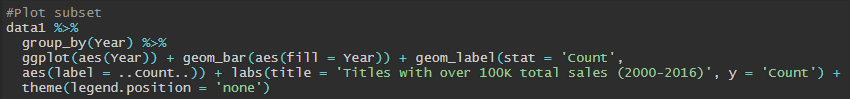


Fig. 2: R code for subset of data filtered to 2000-2016.

I then plotted this subset visually, highlighting 2009 as it was the year with the highest count of video games sold.

Fig. 3: R code plotting 2000-2016 subset by year.

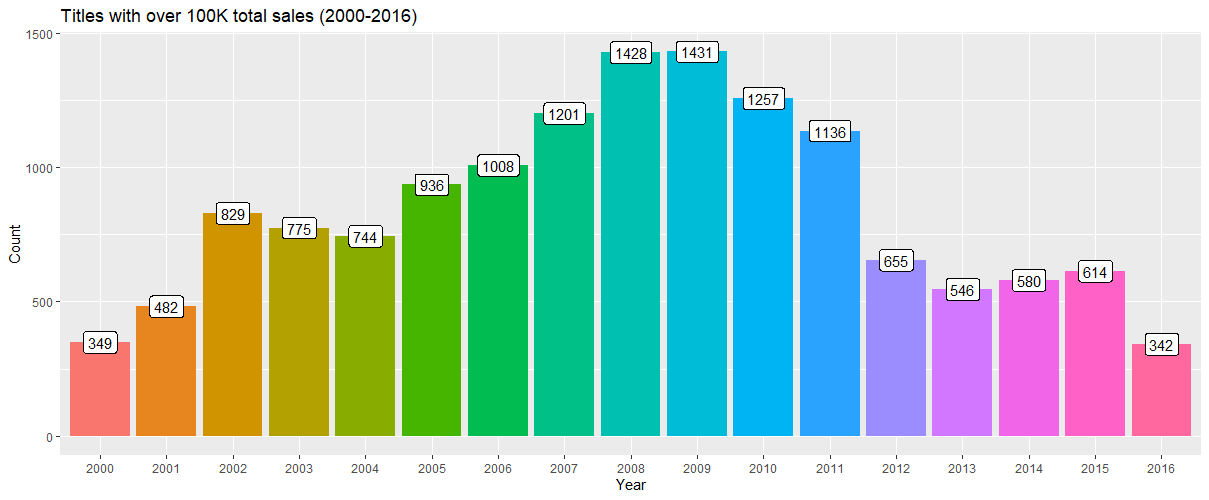


Fig. 4: Plot of video games with over 100,000 total sales by year.

Note that there is likely incomplete data for 2016 as well.

I created a second subset using the data from 2009 and created a plot to highlight the top performing platforms at the time.

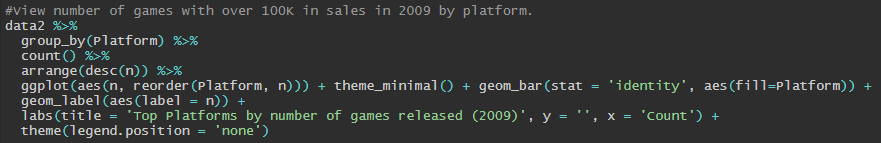


Fig. 5: R code plotting top platforms in 2009.

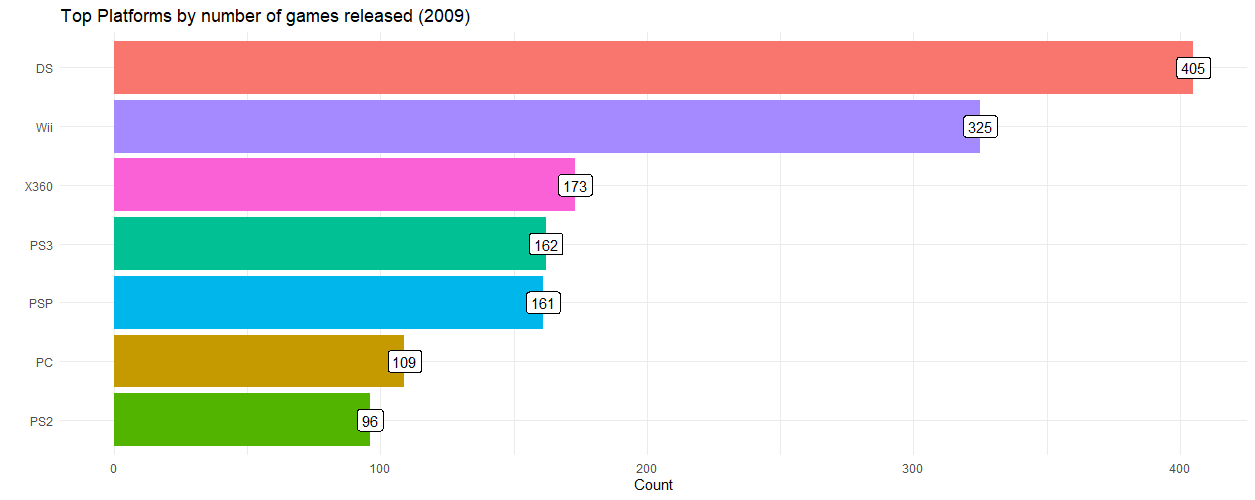


Fig. 6: Plot of top platforms in 2009 by number of games released with over 100,000 total sales.

To explore the data further, I used these two subsets to also create a plot to highlight the top 20 video game publishers during 2000-2016 and the highlighted year 2009. Note that the “head” function can be changed to alter the number of top publishers highlighted.

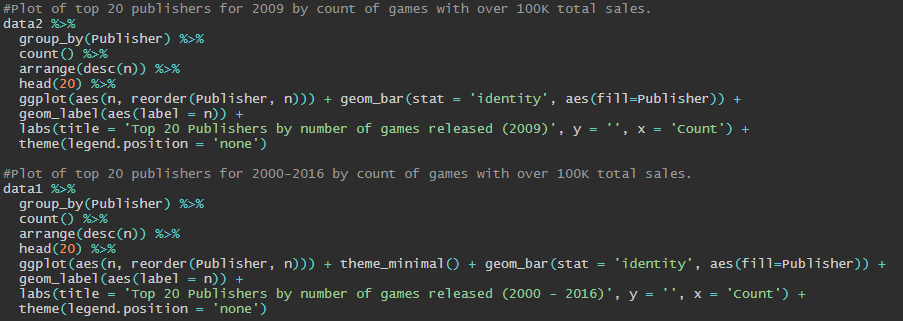


Fig. 7: R code plotting top 20 publishers for 2009 as well as the 2000-2016 range.

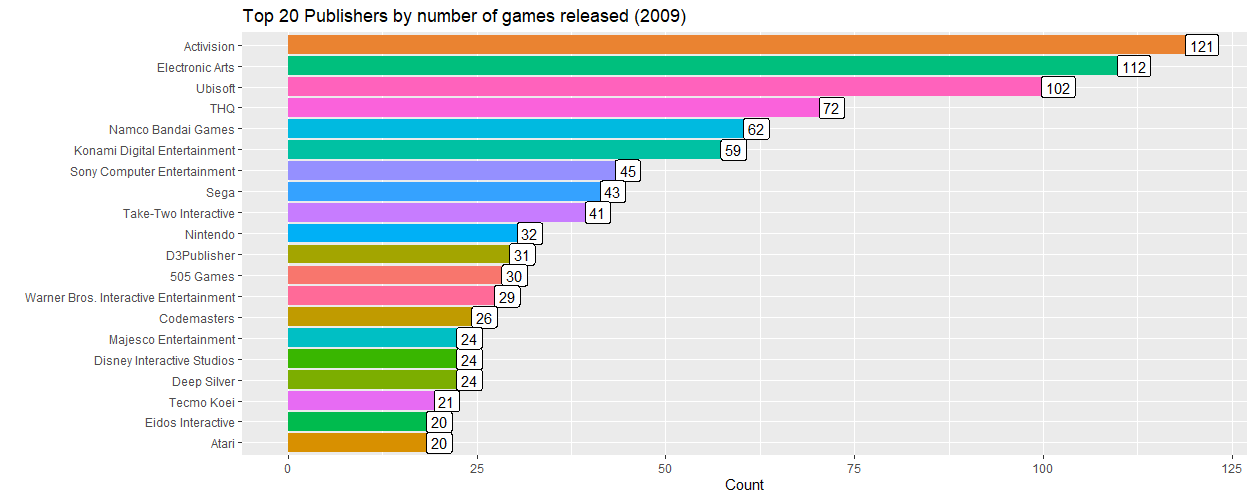


Fig. 8: Plot of top 20 publishers for 2009 by number of games with over 100,000 total sales.

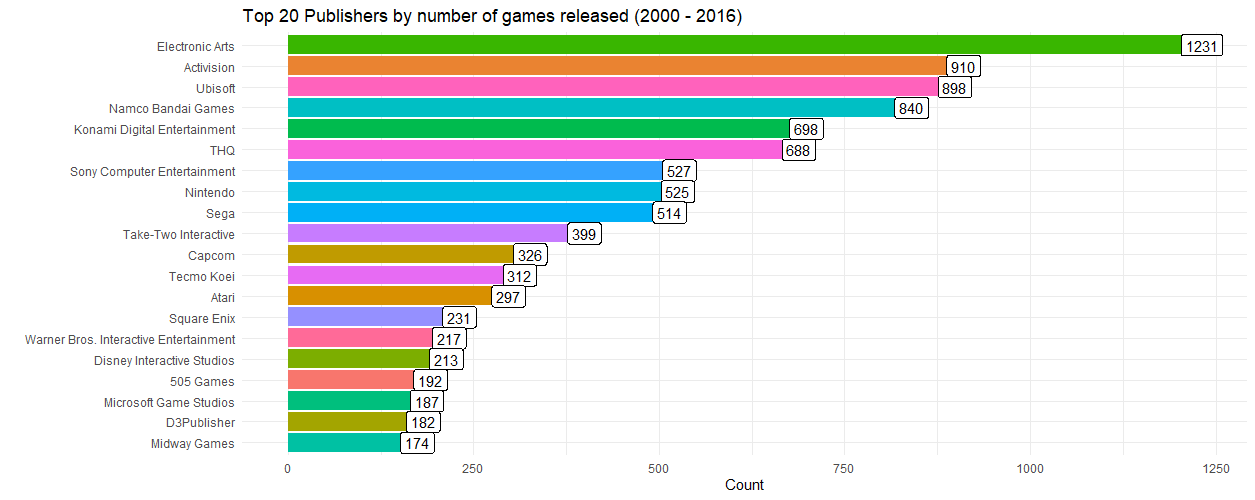


Fig. 9: Plot of top 20 publishers for 2000-2016 by number of games with over 100,000 total sales.

Additionally, I plotted the top-performing video game genres for 2000-2016.

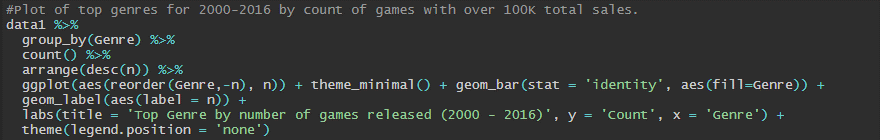


Fig. 10: R code for plotting top genres.

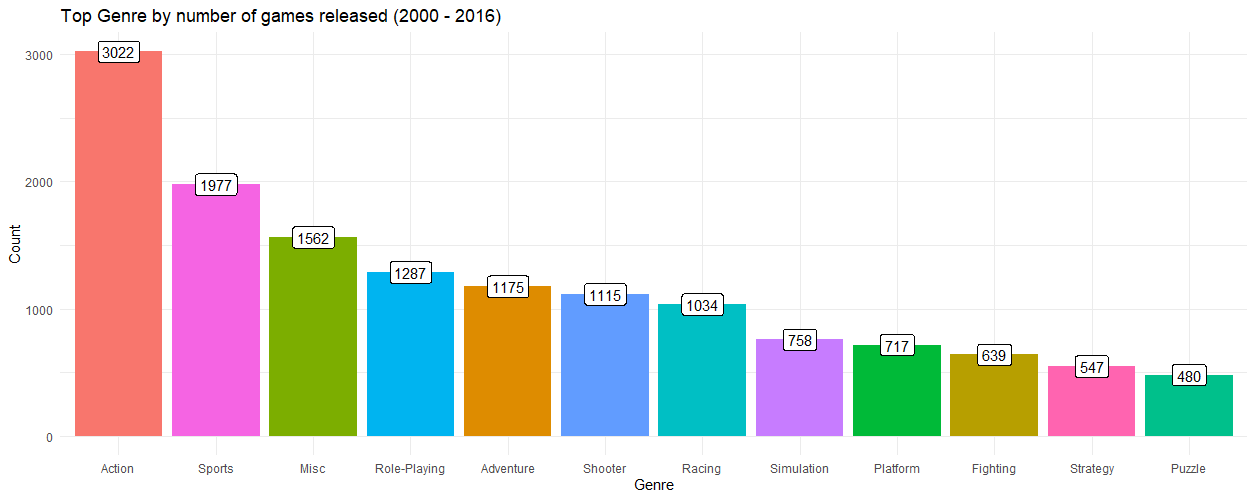


Fig. 11: Plot of top genres for 2000-2016.

Next, I wanted to take a look at the top performing games for each region (North America, Japan, and Europe) as well as globally. To do this, we first must aggregate the sales columns to combine all games with the same name that were released on different platforms.

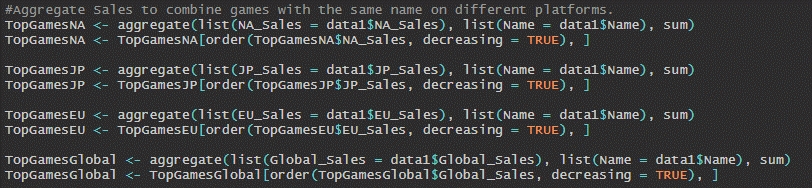


Fig. 12: R code aggregating games with the same name on different platforms.

With our data aggregated, we are able to plot it using ggplot to view our plots for regional and global sales.

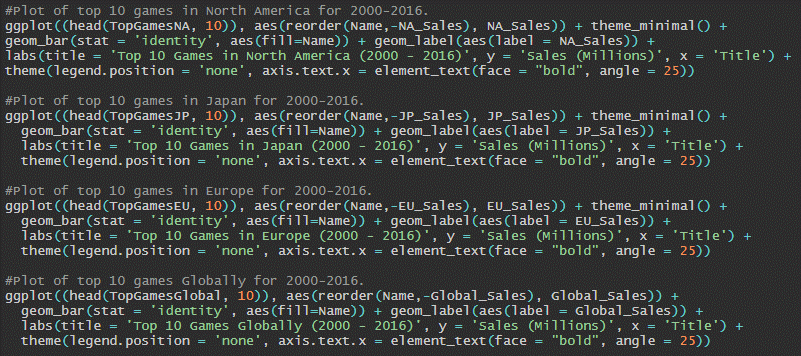


Fig. 13: R code plotting top 10 games for each region and globally.

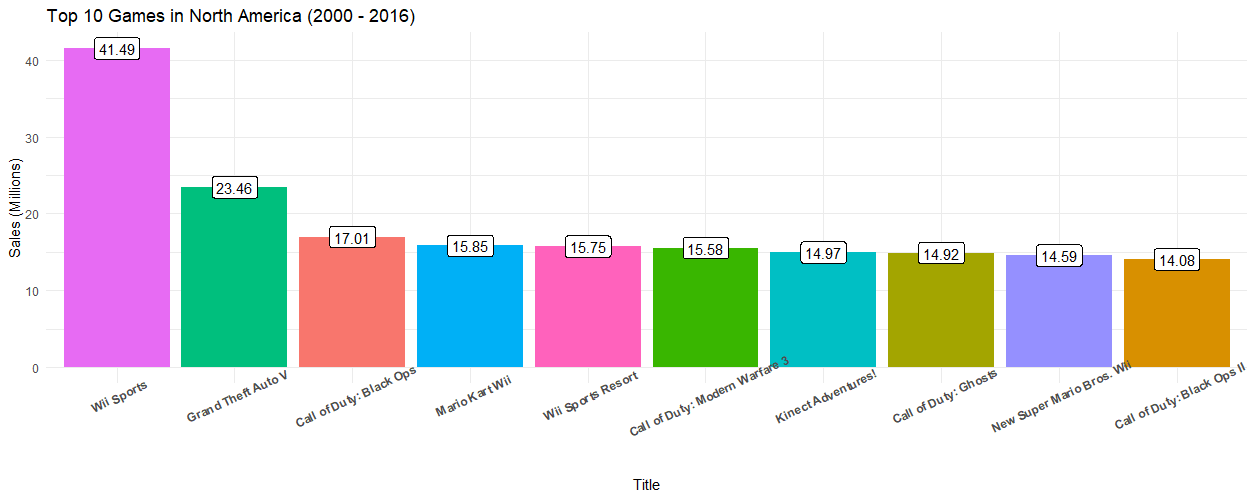


Fig. 14: Plot of top 10 games in North America (2000-2016).

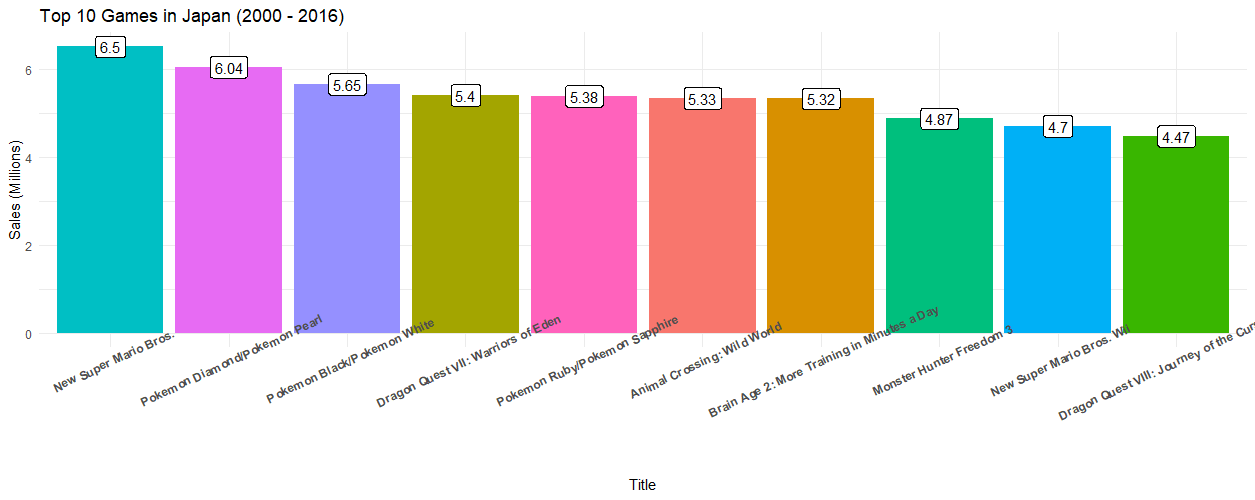


Fig. 15: Plot of top 10 games in Japan (2000-2016).

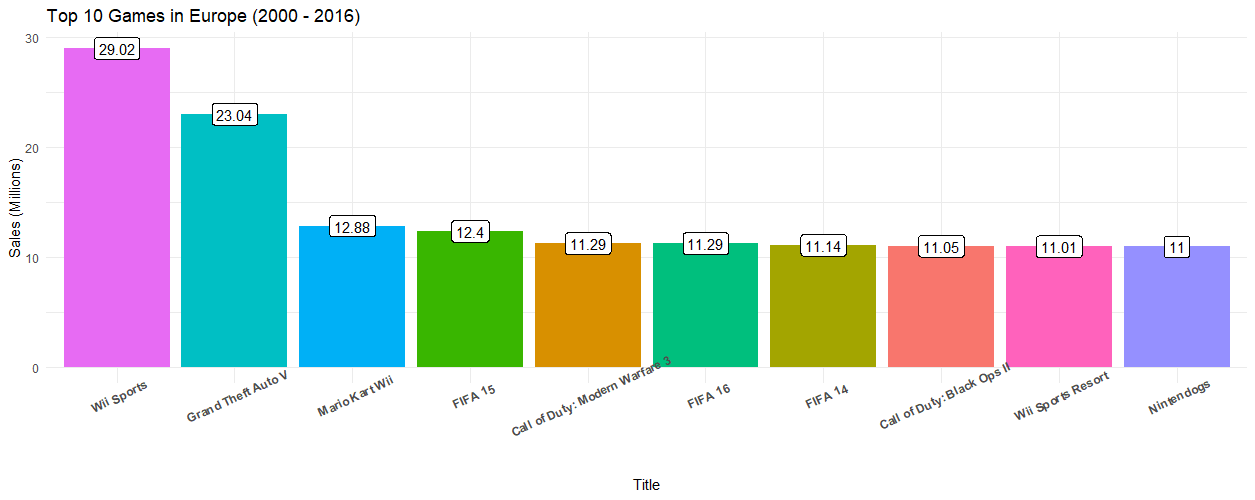


Fig. 16: Plot of top 10 games in Europe (2000-2016).

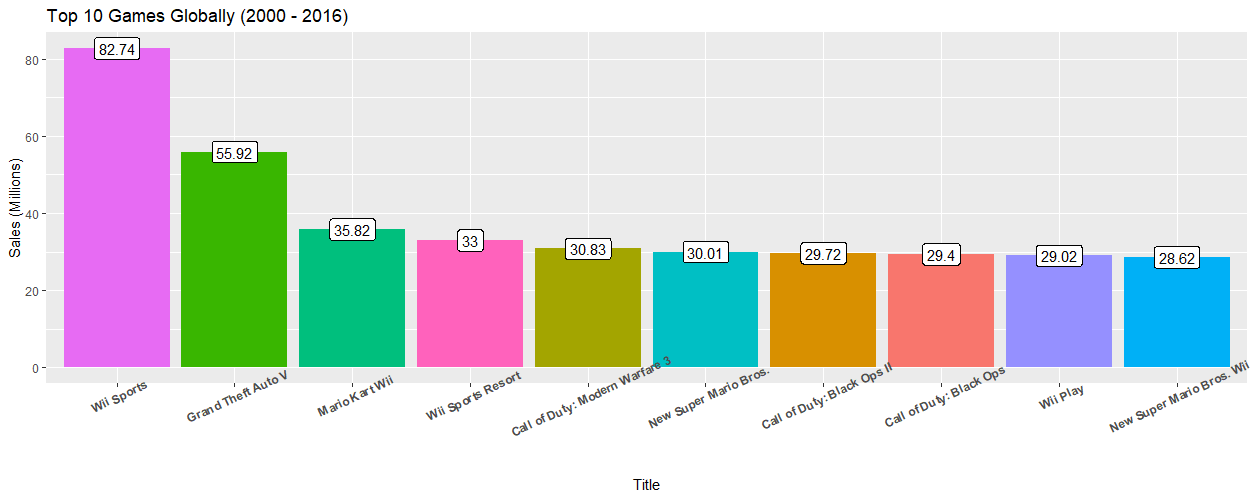


Fig. 17: Plot of top 10 games globally (2000-2016).