Mason Obegi
12/9/22
Data 115
Final Project Report

I selected the video games sales data set because I, like many other kids, grew up playing video games, many of which are in the top 25 in this data set. This data set does not have a lot of information but its size of it combined with the information it does have makes it very interesting. Having sales for all regions, platforms, genres, and publishers makes it a data set that could be very fun to work with in order to find a big question. Some things were pretty obvious in the video game sales: Nintendo is the most popular publisher, North America almost always buys somewhere between ⅓ and ¾ of all copies sold, and the Wii was extremely popular. A few things that stood out to me though were that the genre almost had no correlation to what games were the highest ranked, the year also seemed somewhat irrelevant to what games were ranked the highest, and the proportions between copies sold between countries varied drastically. This means different regions like different games at other times. I wanted to do a quick check to see the correlation between regions and their sales so I made a correlation matrix between all the regions in the data set.

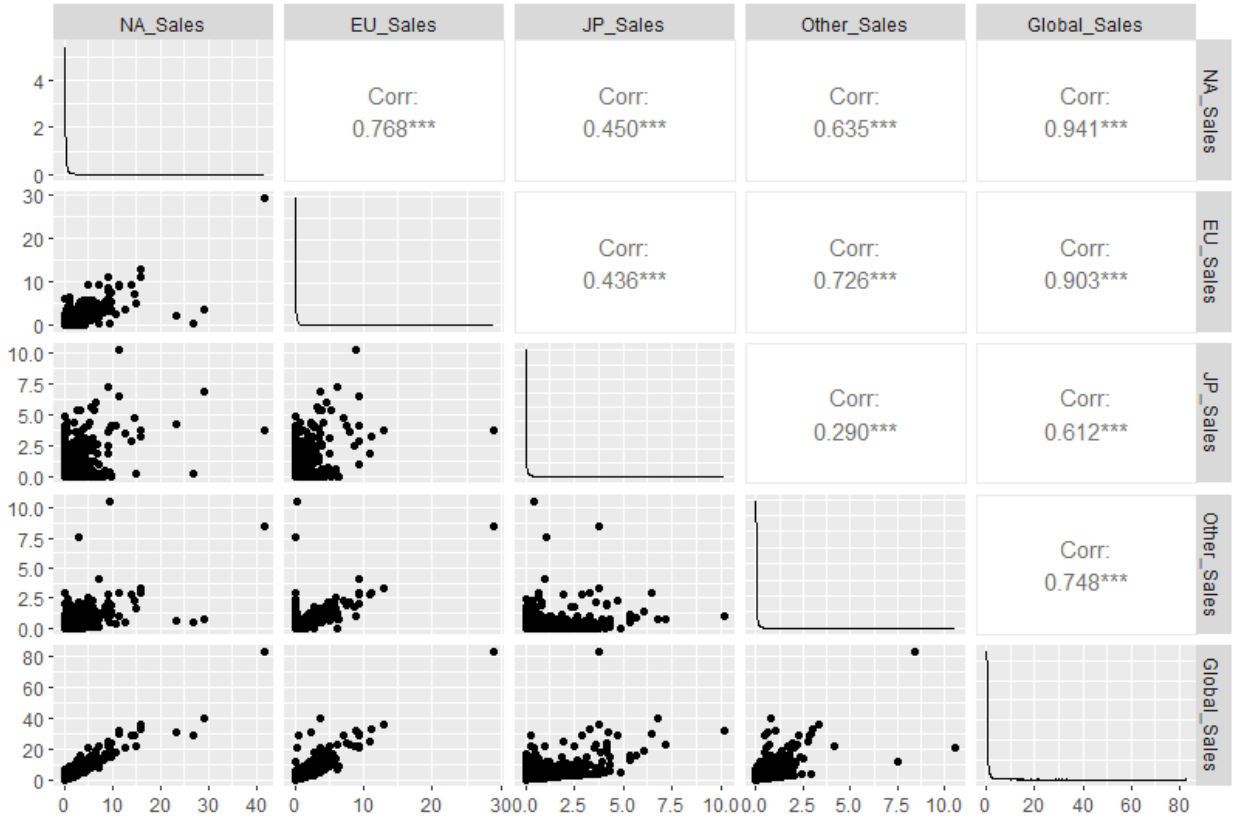
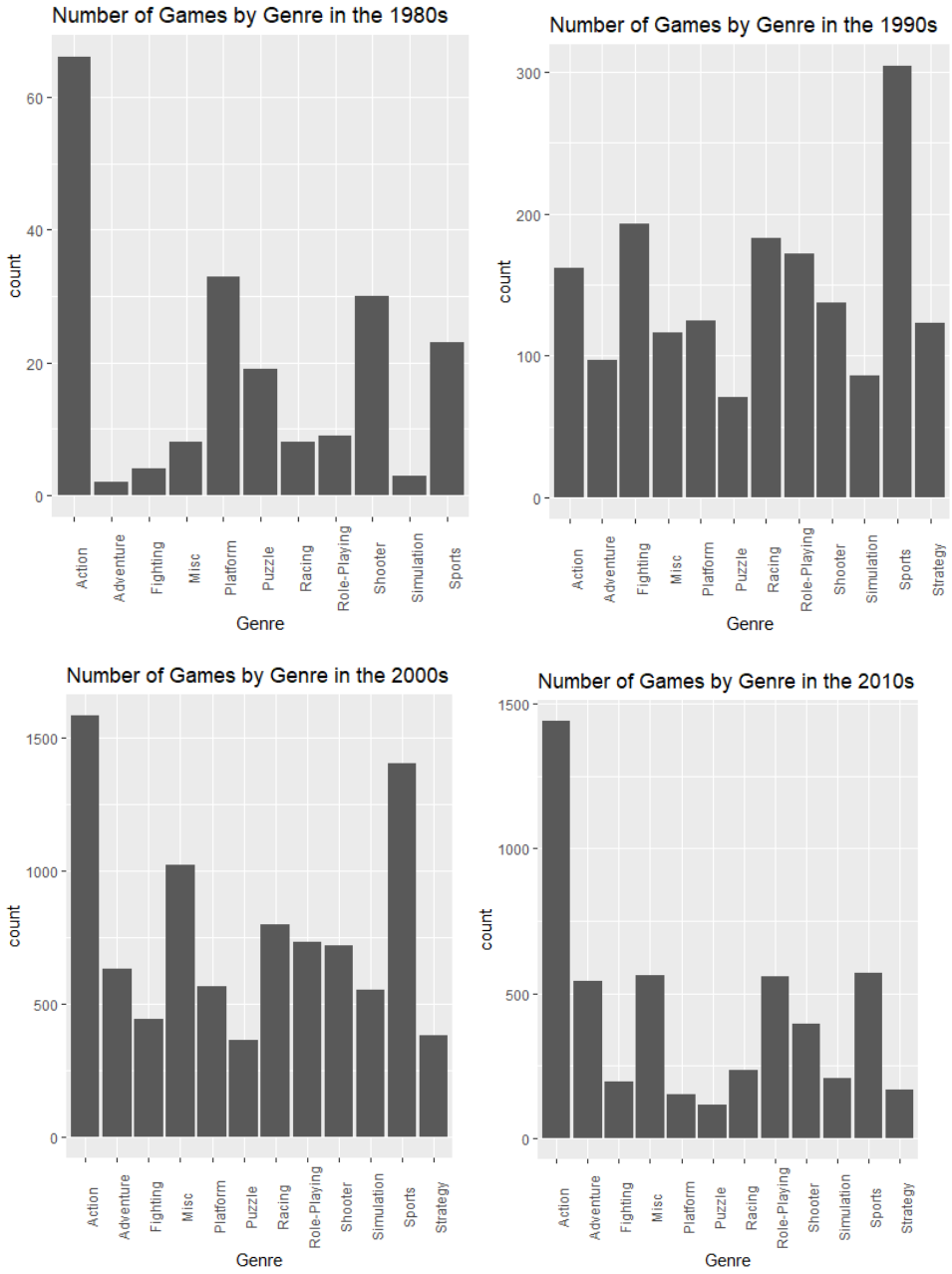


Figure 1.

We can see from this visualization that North America and the EU have a strong correlation with global sales, but most other correlations vary.



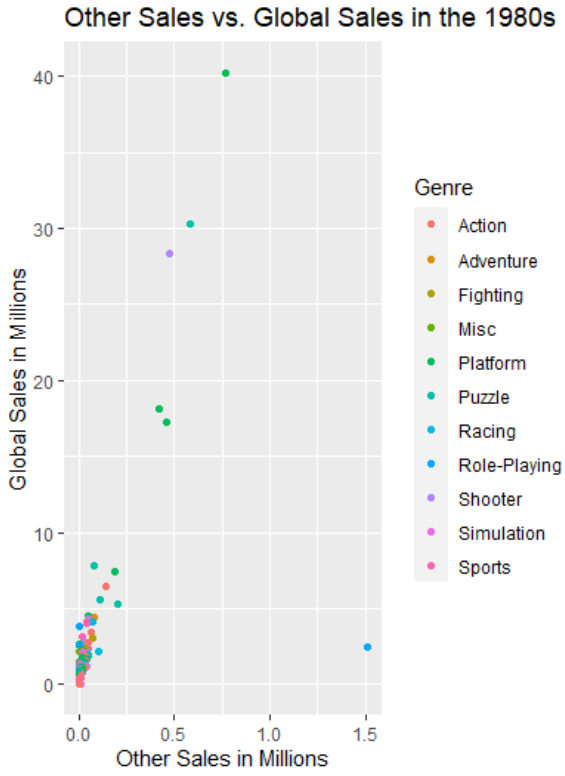
Figures 2-5, left to right, top to bottom.

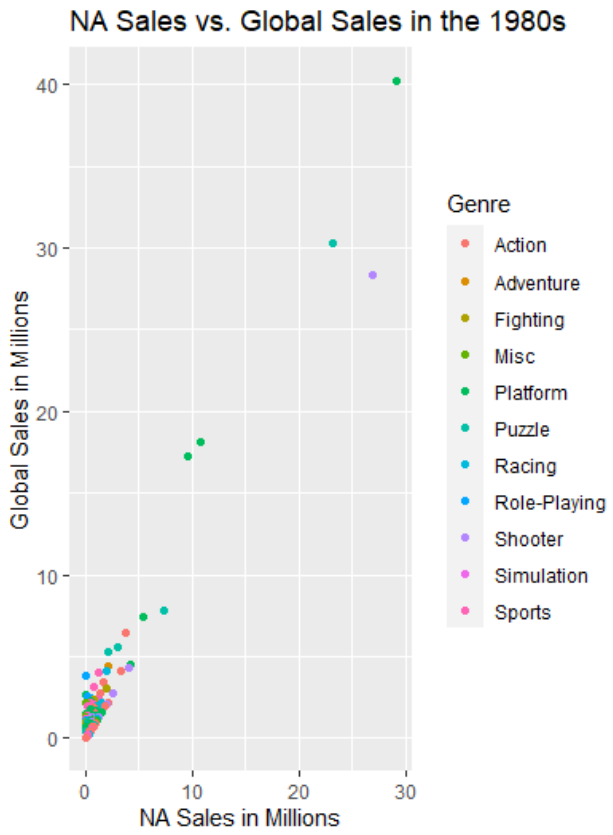
Some more visualizations show an interesting pattern of what kinds of games came out in different years. I took each decade and graphed how many of each type of genre came out every decade. What we can see is that action dominated early, then sports came into the mix then back to action. One thing I was wondering was what kind of games did each of the regions like? Looking at the raw data, it appears that North America liked Platform games while Japan liked Role-Playing. What was very odd was that even though action looked to dominate the gaming world for most decades, the 17th ranked game is the highest-ranked in terms of global sales for that genre. That lead me to ask; if the different regions do not all buy the same games and the action genre, which was the most popular genre 3 in the 4 decades, is not as high

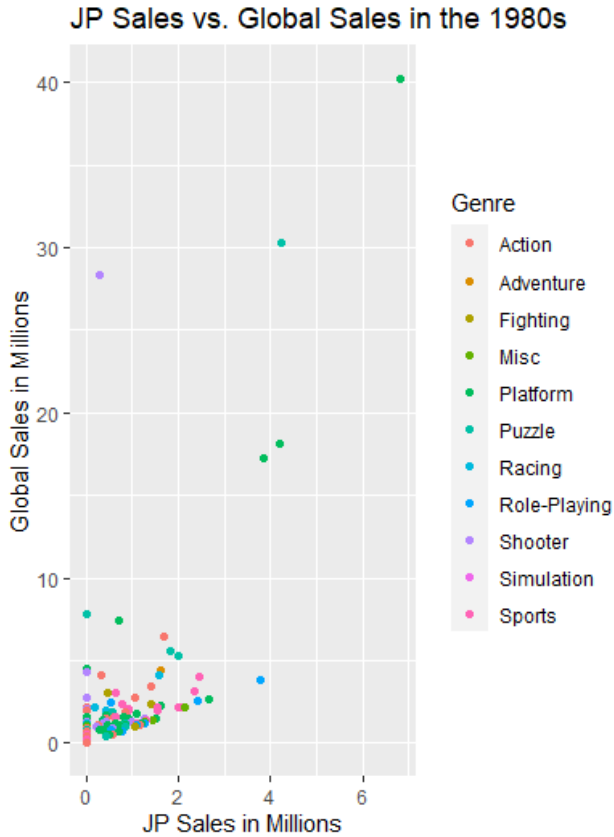
selling as other, less popular genres in terms of the number of releases, then over the years, what kinds of games have been the most popular in each region, and does every region like the most popular genre from the decade? This data set works for my “big question” because there are over 15000 games spanning 40 years in this data set, all with the number of sales for each region, and the year they came out, among a few other categories that may help me narrow down my question a little more, such as platform and publisher. This to me is not a black-and-white question. There are many factors that could be used to determine the most “popular” game, there could be total sales, total games released, and the average number of sales per title release, among many others.

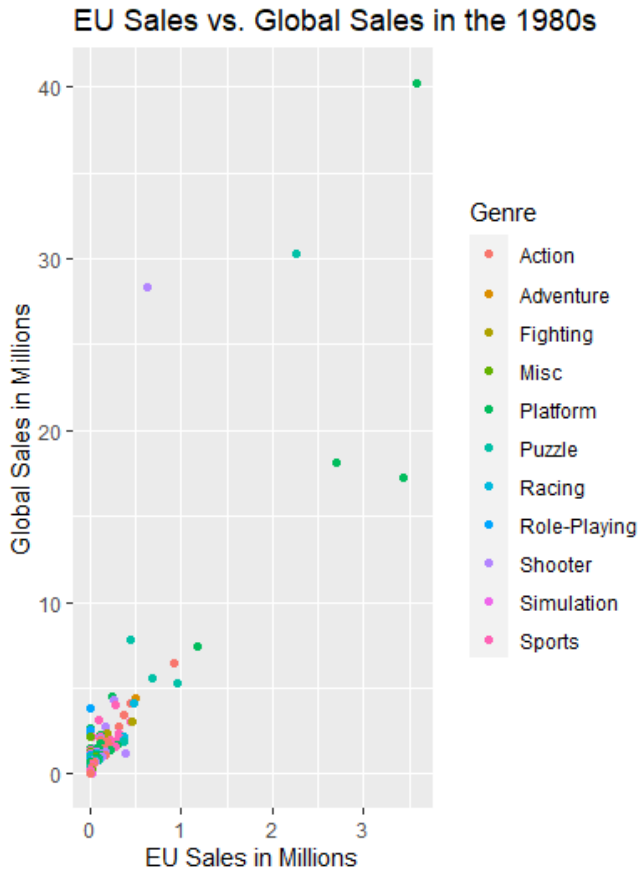
There was only one major issue with this data set, and that was that there were about 250 games that had no year associated with them. Unfortunately due to the question I am trying to answer, I had to omit those games. Other than that, this data set was very straightforward.

So, looking back to find the most popular games through each decade in each region, I made graphs of all region's sales against the global sales for each decade, with every dot colored by the genre.




NA Sales vs. Global Sales in the 1980s
Global Sales in Millions
NA Sales in Millions
0
10
20
30
40
Genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role-Playing
Shooter
Simulation
Sports

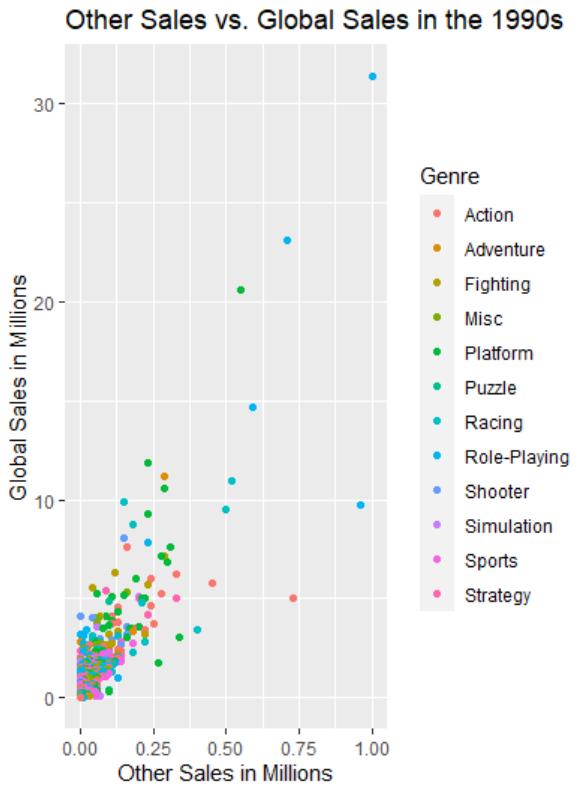

JP Sales vs. Global Sales in the 1980s
Global Sales in Millions
JP Sales in Millions
0
2
4
6
10
20
30
40
Genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role-Playing
Shooter
Simulation
Sports

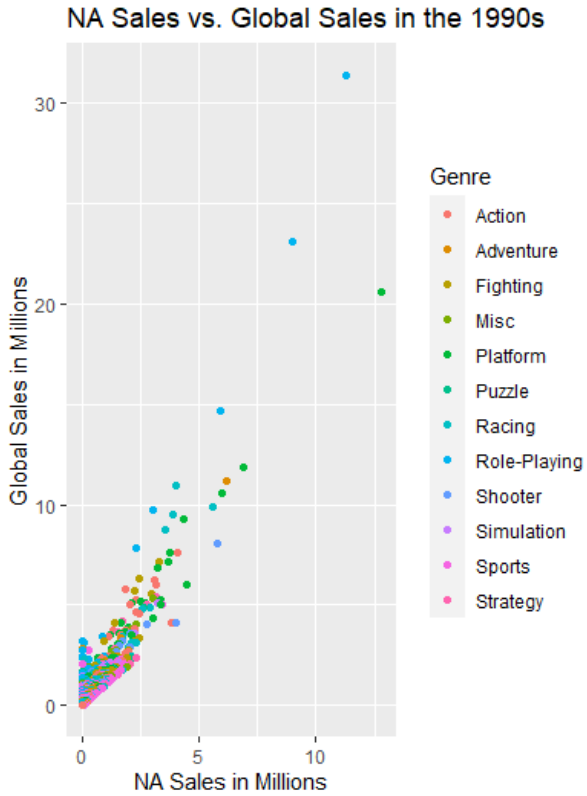


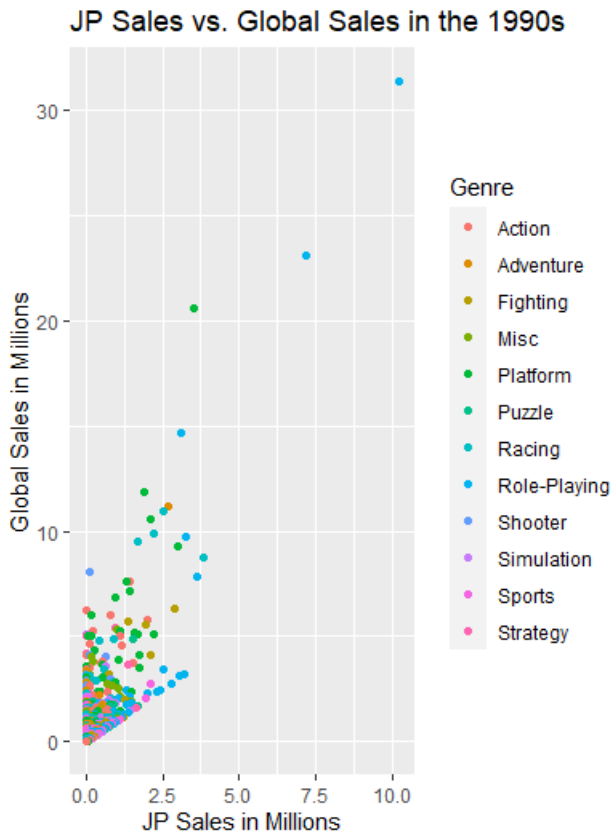
Figures 6-9 top to bottom.

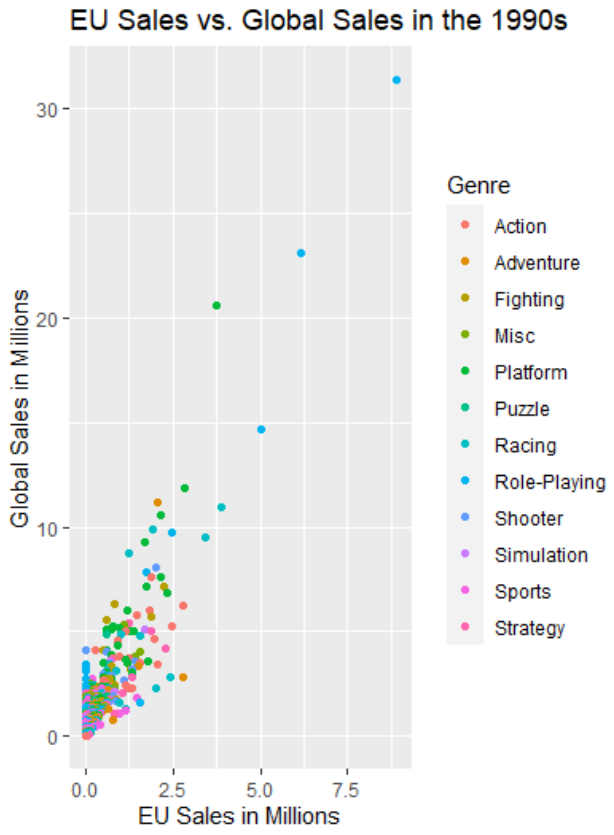
The first thing that jumps out on all these graphs, there are a lot of unpopular games in every country that are clustered towards the very bottom left of each graph, meaning no global sales and no regional sales. Another thing we can see, even though action games were the most released type of genre, platformer was the most sold game in every region except for others, where it was role-playing. What is even crazier is that, if you look at the outliers in each graph, none of them is action, meaning even though a lot of action games were released, they were not popular, perhaps explaining the sports genre overtaking it as the most released genre in the 1990s. Going through the data, the most popular game this decade in every region was Super Mario Bros, except for other sales where it was Dragon Warrior II. For the most part, it looks like all regions had platformers as their genre of the decade, while there were still unique genres towards the outlier range for each region, platformers were consistently in all of the region's outliers multiple times.

Next, looking at the 1990s:


Other Sales vs. Global Sales in the 1990s
Global Sales in Millions
Other Sales in Millions
0
10
20
30
0.00
0.25
0.50
0.75
1.00
Genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role-Playing
Shooter
Simulation
Sports
Strategy


NA Sales vs. Global Sales in the 1990s
Global Sales in Millions
NA Sales in Millions
0
10
20
30
0
5
10
Genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role-Playing
Shooter
Simulation
Sports
Strategy

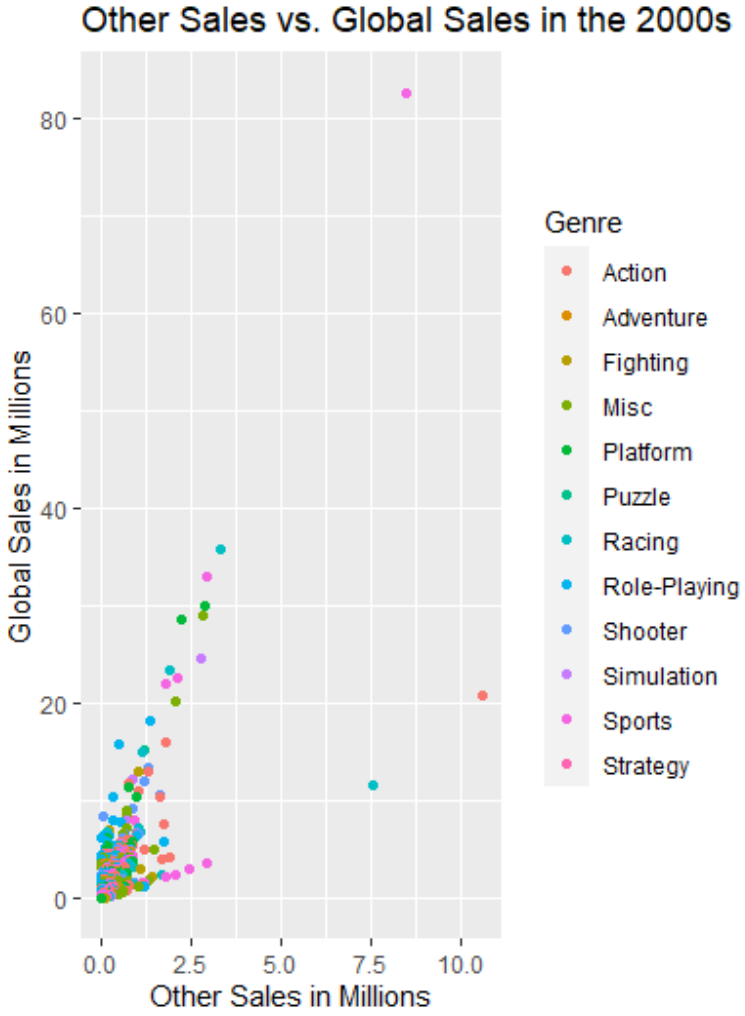

JP Sales vs. Global Sales in the 1990s
Global Sales in Millions
JP Sales in Millions
0
10
20
30
0.0
2.5
5.0
7.5
10.0
Genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role-Playing
Shooter
Simulation
Sports
Strategy

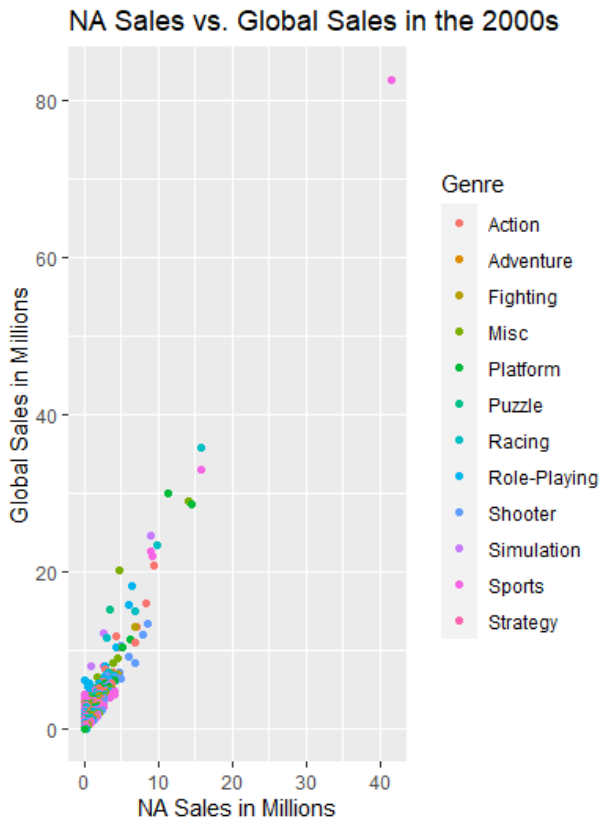

EU Sales vs. Global Sales in the 1990s
Global Sales in Millions
EU Sales in Millions
0
10
20
30
0.0
2.5
5.0
7.5
Genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role-Playing
Shooter
Simulation
Sports
Strategy

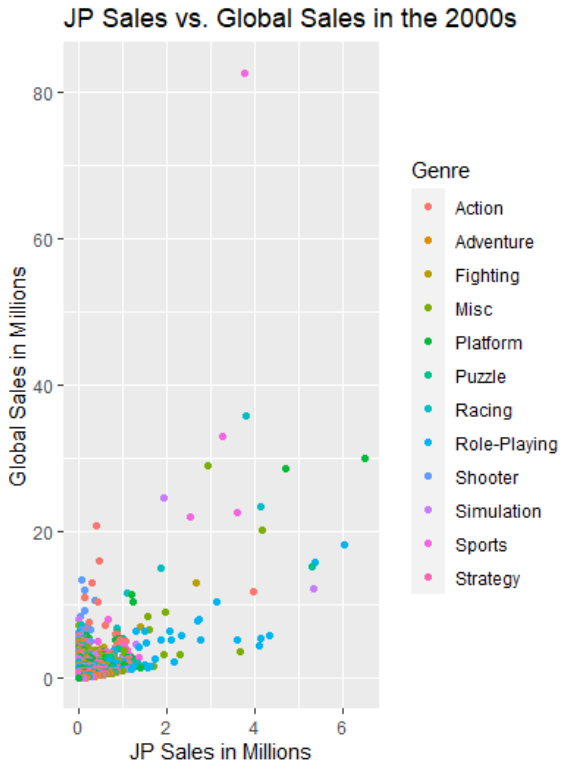
Figures 10-13 top to bottom.

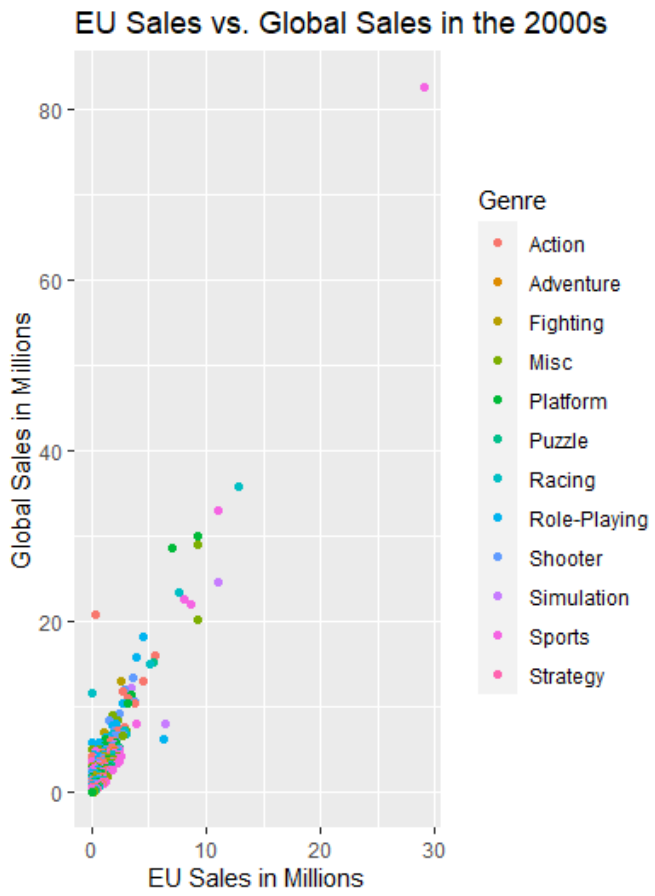
We can see that every region had a role-playing game as their most popular game, except for NA which had it as a close second. This seems odd considering in figure 3, we can see that in the 1990s, role-playing games are only the 4th most released genre. It looks like a through-line from the 1980s to 1990s is that the most sold game genre was the same in every region except one, and the genre most released that year was not even an outlier when it came to sales.

In the 2000s:




NA Sales vs. Global Sales in the 2000s
Global Sales in Millions
NA Sales in Millions
0
20
40
60
80
0
10
20
30
40
Genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role-Playing
Shooter
Simulation
Sports
Strategy

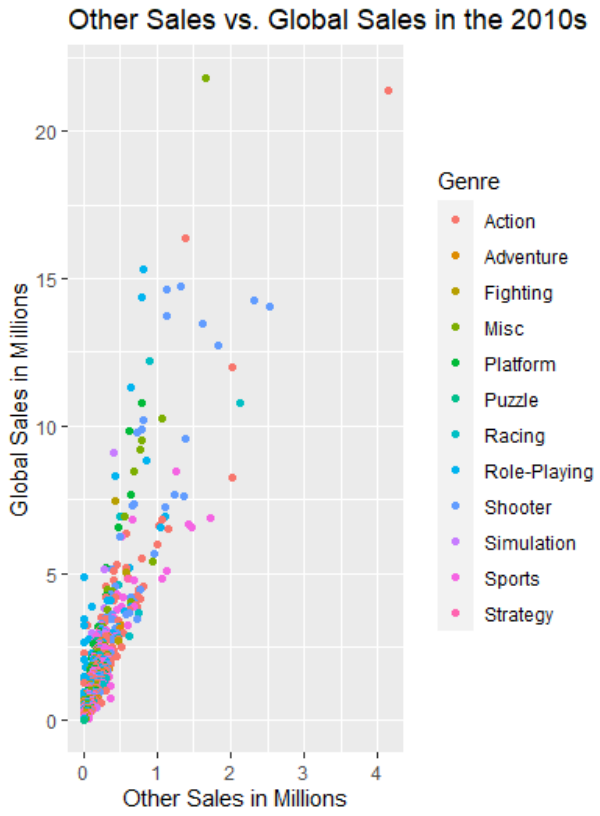

JP Sales vs. Global Sales in the 2000s
Global Sales in Millions
JP Sales in Millions
0
20
40
60
80
0
2
4
6
Genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role-Playing
Shooter
Simulation
Sports
Strategy

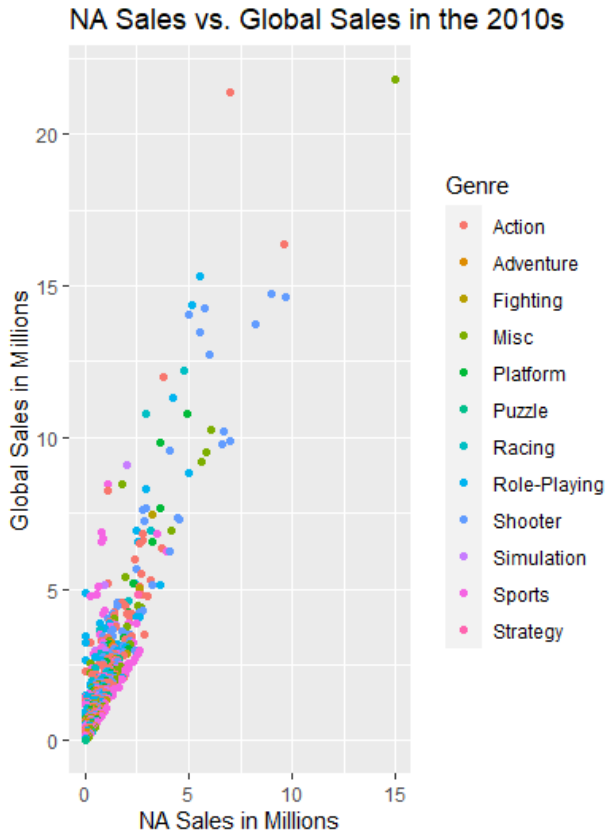


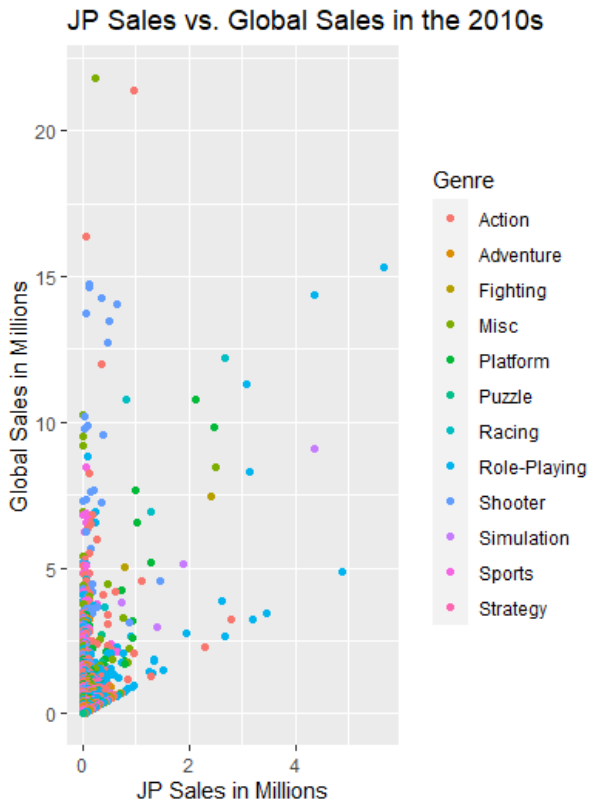
Figures 14-17 top to bottom.

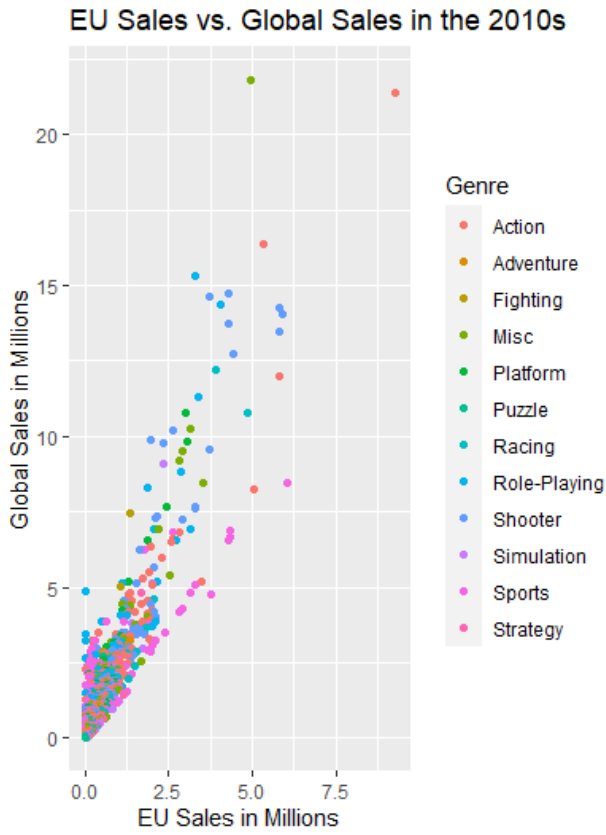
This is the decade in which I believe my question begins to take shape. In the 2000s the internet started coming into more prominence, so instead of just having one major game each decade that dominates all markets, people were able to find games that were not just the most sold each year but were games they wanted to play. As we can tell, in the EU, a sports game was the huge front-runner in terms of sales, along with that, we can see that two other outliers are also in the sports genre. From this, I think we can conclude that in the EU the genre of the decade was sports. Looking at other sales and NA sales we can conclude that there is no main genre that looks to be dominating, however, a sports game is still the top selling in NA and second in selling in other sales. Finally, looking at Japan, we can tell that this is the decade in which it separates itself. The sports game that has been at the top for the other regions, is not top 10 for japan. What we can see is that role-playing games dominate in Japan unlike the rest of the world. This decade action games were the most released, yet we still only have one region that has that genre as a best-selling game.

Next in the 2010s:


Other Sales vs. Global Sales in the 2010s
Global Sales in Millions
Other Sales in Millions
0
5
10
15
20
0
1
2
3
4
Genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role-Playing
Shooter
Simulation
Sports
Strategy


NA Sales vs. Global Sales in the 2010s
Global Sales in Millions
NA Sales in Millions
0
5
10
15
20
0
5
10
15
Genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role-Playing
Shooter
Simulation
Sports
Strategy





Figures 18-21 top to bottom.

The 2010s seem to be when each region starts to separate and understand what kinds of games they enjoy. As I mentioned in the 2000s, the internet started to emerge, allowing people to research games for themselves rather than just buying what is popular, and I think the effects of this show up more in the 2010s. Japan is still heavily dominated in the genre of role-playing. We can finally see action having a top-selling game in the genre making an impact, by having the top-selling game in both other and EU. In NA we can see that a miscellaneous game had the best-selling title. In other, NA and EU, we can see that action, shooter, and role-playing are generally the top 3 genres. In other, sports make a case for being in the top few genres as well. Still, the actions proportion of games released does not seem to be making as big of an impact as one would think, except for in other sales, where it had the best-selling game.

I selected this methodology because to me I felt like it made the most sense to find out popular games in each decade in each region, it would make sense to graph the sales in each region for each decade. The pros to using this method are that it is very easy to tell the outliers for popular games in each decade in each region. The cons are that it is only easy to see the outliers, and there are a lot of points on the scatter plot tattoo close together or even on top of each other to tell them apart. However, for my question, I did not need to know about the bad-selling games. Some alternative methods I considered were making bar graphs about sales from each genre, but I decided against this because I felt like it would not have been a good indicator of popular games just to do total sales or the average sales of each game of each genre. I felt like total sales would have been a poor choice as in many years there were one or two genres that dominated the market it releases and they would have been towards the top by default. I also did not want to do average games sales per game in each genre because a lot of people who play video games are not buying a lot of games, but the few that come out that they think they can enjoy and get a lot of fo time out of, therefore, I thought that looking at the genre of game that was in the outliers made the most sense, showed the games people bought because they enjoyed it, not showing that one genre has one amazing game and a bunch of duds, or too small of a sample size to get an average and skewing the data in that genres favor heavily.

My conclusions are that up until the 2000s, there was almost a negligible difference in what regions like what games. It appeared that there was a top game that was the best seller, or at least close to it in every region, then each of the next best-selling games seemed to be random in the genre, with no strong correlation about what was the best-selling genre, or even best-selling couple genres any given year. Starting in the 2000s, we can see that this is why in figure 1, Japan was not as strongly correlated with other regions, Japan started to fall in love with role-playing games way more than any other region. What we can tell is that the EU leans more towards sports games starting this decade, while others and NA still seem to not have a strong genre for that region. Finally, in the 2010s, we can see that EU, NA, and other generally prefer action, shooter, and role-playing, while other slightly prefers sports as well. While Japan is dominated entirely by role-playing. In the 2000s and 2010s, we can see that there are 3 different genres at the top of each region, unlike in previous decades. We can make all of the conclusions by looking at each of the graphs for each decade. So the answer to my “big

question”, until the 2000s, every region had practically the same taste in video games, having the same trends in genres and usually having the same top-selling genre lacrosse 3 of the regions and being top 3 in the other region. However, starting in the 2000s, Japan was the first region to separate itself from the genres they enjoy, showing that they are into role-playing games while the other regions still do not have a strong correlation with genres. But in the 2010s all regions found what game they enjoy, Japan still with role-playing, and EU and NA liked shooters, action, and role-playing, while others liked role-playing, action, shooters, and sports. Along with this, the most released genre did not seem to affect what was the best-selling game in general.

Some additional analysis that I would have liked to use was comparing the publisher with the types of games released. Publishers like Nintendo dominate platformers like Mario and role-playing games like Pokemon. It would be cool to see what would happen to those genres if the big fish in the water were taken out because even though Nintendo does make great games, there is a good amount of people that buy those games because of the brand name. Some more data that would have been nice to have is the release date of the game. I know a lot of publishers release their games in the fall for black Friday and Christmas and it would have been nice to see how the time of year a game is released correlates to its sales.

Code for Figures:

Figure 1: ggpairs(df[,7:11])

Figure 2: df_1980s <- df[df$Year >= 1980 & df$Year <= 1989,]
ggplot(df_1980s, aes(x = Genre))+geom_bar() +ggtitle("Number of Games by Genre in 1980s")
+theme(axis.text.x = element_text(angle = 90))

Figure 3: df_1990s <- df[df$Year >= 1990 & df$Year <= 1999,]
ggplot(df_1990s, aes(x = Genre))+geom_bar() +ggtitle("Number of Games by Genre in 1990s")
+theme(axis.text.x = element_text(angle = 90))

Figure 4: df_2000s <- df[df$Year >= 2000& df$Year <= 2009,]
ggplot(df_2000s, aes(x = Genre))+geom_bar() +ggtitle("Number of Games by Genre in 2000s")
+theme(axis.text.x = element_text(angle = 90))

Figure 5: df_2010s <- df[df$Year >= 2010 & df$Year <= 2019,]
ggplot(df_2010s, aes(x = Genre))+geom_bar() +ggtitle("Number of Games by Genre in
2010s")+theme(axis.text.x = element_text(angle = 90))

Figure 6: ggplot(df_1980s, aes(df_1980s$Other_Sales, df_1980s$Global_Sales))+
geom_point(aes(color= Genre)) + ggtitle("Other Sales vs. Global Sales in the 1980s") + labs(y=
"Global Sales in Millions", x = "Other Sales in Millions")

Figure 7: ggplot(df_1980s, aes(df_1980sJP_Sales, df_1980sGlobal_Sales))+
geom_point(aes(color= Genre)) + ggtitle("JP Sales vs. Global Sales in the 1980s") + labs(y=
"Global Sales in Millions", x = "JP Sales in Millions")

Figure 8: ggplot(df_1980s, aes(df_1980sNA_Sales, df_1980sGlobal_Sales))+
geom_point(aes(color= Genre)) + ggtitle("NA Sales vs. Global Sales in the 1980s") + labs(y=
"Global Sales in Millions", x = "NA Sales in Millions")

Figure 9: ggplot(df_1980s, aes(df_1980sEU_Sales, df_1980sGlobal_Sales))+
geom_point(aes(color= Genre)) + ggtitle("EU Sales vs. Global Sales in the 1980s") + labs(y=
"Global Sales in Millions", x = "JP Sales in Millions")

Figure 10: ggplot(df_1990s, aes(df_1990s$Other_Sales, df_1990s$Global_Sales))+
geom_point(aes(color= Genre)) + ggtitle("Other Sales vs. Global Sales in the 1990s") + labs(y=
"Global Sales in Millions", x = "Other Sales in Millions")

Figure 11: ggplot(df_1990s, aes(df_1990sNA_Sales, df_1990sGlobal_Sales))+
geom_point(aes(color= Genre)) + ggtitle("NA Sales vs. Global Sales in the 1990s") + labs(y=
"Global Sales in Millions", x = "NA Sales in Millions")

Figure 12:ggplot(df_1990s, aes(df_1990sJP_Sales, df_1990sGlobal_Sales))+
geom_point(aes(color= Genre)) + ggtitle("JP Sales vs. Global Sales in the 1990s") + labs(y=
"Global Sales in Millions", x = "JP Sales in Millions")

Figure 13:ggplot(df_1990s, aes(df_1990sEU_Sales, df_1990sGlobal_Sales))+
geom_point(aes(color= Genre)) + ggtitle("EU Sales vs. Global Sales in the 1990s") + labs(y=
"Global Sales in Millions", x = "EU Sales in Millions")

Figure 14: ggplot(df_2000s, aes(df_2000s$Other_Sales, df_2000s$Global_Sales))+
geom_point(aes(color= Genre)) + ggtitle("Other Sales vs. Global Sales in the 2000s") + labs(y=
"Global Sales in Millions", x = "Other Sales in Millions")

Figure 15: ggplot(df_2000s, aes(df_2000sNA_Sales, df_2000sGlobal_Sales))+
geom_point(aes(color= Genre)) + ggtitle("NA Sales vs. Global Sales in the 2000s") + labs(y=
"Global Sales in Millions", x = "NA Sales in Millions")

Figure 16: ggplot(df_2000s, aes(df_2000sJP_Sales, df_2000sGlobal_Sales))+ geom_point(aes(color= Genre)) + ggtitle("JP Sales vs. Global Sales in the 2000s") + labs(y= "Global Sales in Millions", x = "JP Sales in Millions")
Figure 17: ggplot(df_2000s, aes(df_2000sEU_Sales, df_2000sGlobal_Sales))+ geom_point(aes(color= Genre)) + ggtitle("EU Sales vs. Global Sales in the 2000s") + labs(y= "Global Sales in Millions", x = "EU Sales in Millions")
Figure 18: ggplot(df_2010s, aes(df_2010s$Other_Sales, df_2010s$Global_Sales))+ geom_point(aes(color= Genre)) + ggtitle("Other Sales vs. Global Sales in the 2010s") + labs(y= "Global Sales in Millions", x = "Other Sales in Millions")
Figure 19: ggplot(df_2010s, aes(df_2010sNA_Sales, df_2010sGlobal_Sales))+ geom_point(aes(color= Genre)) + ggtitle("NA Sales vs. Global Sales in the 2010s") + labs(y= "Global Sales in Millions", x = "NA Sales in Millions")
Figure 20: ggplot(df_2010s, aes(df_2010sJP_Sales, df_2010sGlobal_Sales))+ geom_point(aes(color= Genre)) + ggtitle("JP Sales vs. Global Sales in the 2010s") + labs(y= "Global Sales in Millions", x = "JP Sales in Millions")
Figure 21: ggplot(df_2010s, aes(df_2010sEU_Sales, df_2010sGlobal_Sales))+ geom_point(aes(color= Genre)) + ggtitle("EU Sales vs. Global Sales in the 2010s") + labs(y= "Global Sales in Millions", x = "EU Sales in Millions")