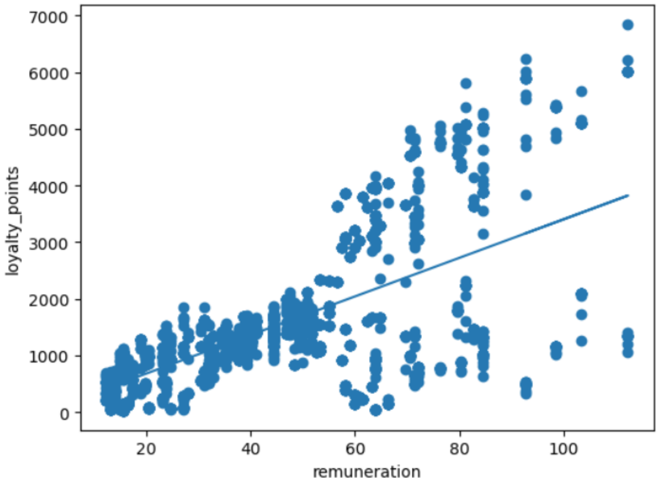
Predicting future outcomes: Report

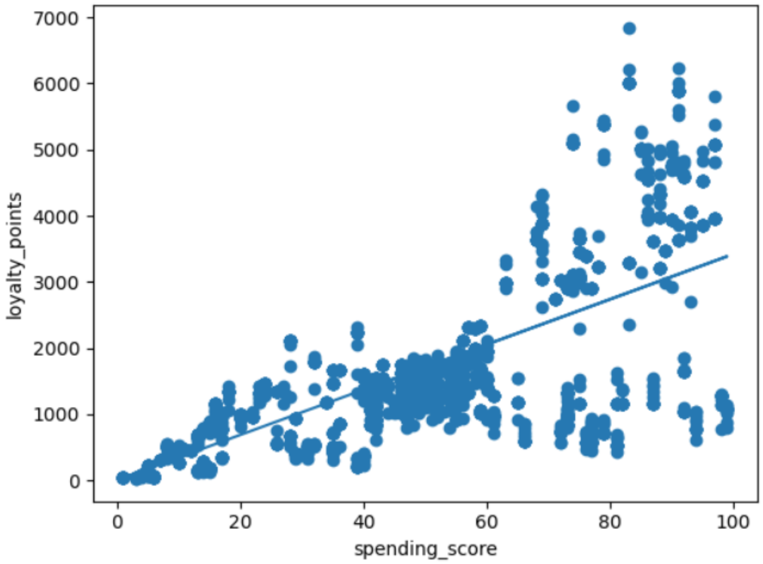
How Do Customers Accumulate Loyalty Points?

I create a number of plots looking at possible connections between loyalty points and various customer metrics. For example, there is a trend where the higher a customer's yearly income (remuneration) the more loyalty points they accumulate. The very highest loyalty points are accrued by customer groups making over 100K per year:



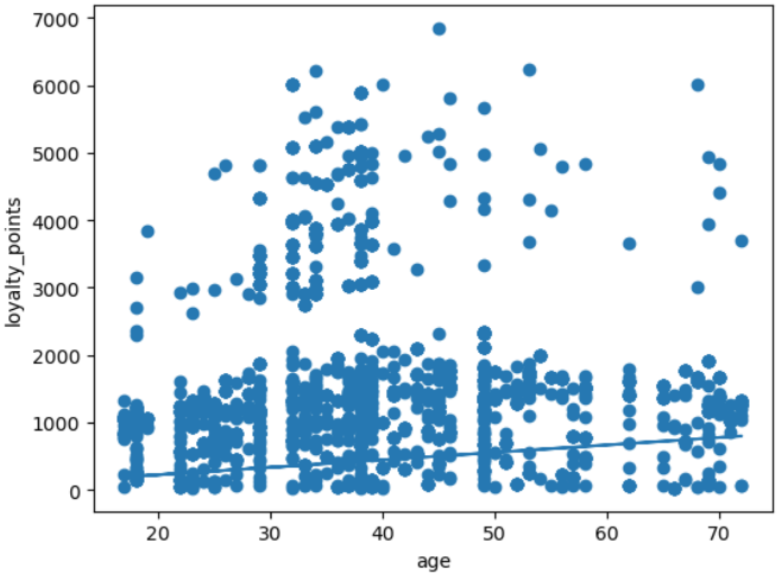
In terms of spending score, the largest group clusters around a score between 40 and 60. After 60 we see the most loyal customers, with a strong correlation between loyalty and spending score.

We also see that at the upper levels of remuneration, the data loosely splits into two groups: those with high income and low loyalty and those with high levels in both. There is a gap, where there are not so many costumes with mid-low loyalty and high income. Our demographics could be said to split here. . This may be a good group to market to, given their high level of potential spend on Turtle Games products with successful marketing.



In simple terms, the more a customer spends *and* has available to spend, the more loyal they tend to be. The meta-data doesn't contain information on the methodology of the 'spending score', so it may be that it includes 'frequency of purchasing' or a similar metric. This would probably explain the correlation with loyalty points. However, something that is interesting to note here is that we do *not* see a big trend of highly-scored but not-very-loyal customers, suggesting that people who spend 'well' at Turtle Games are also repeat customers, and not one-offs.

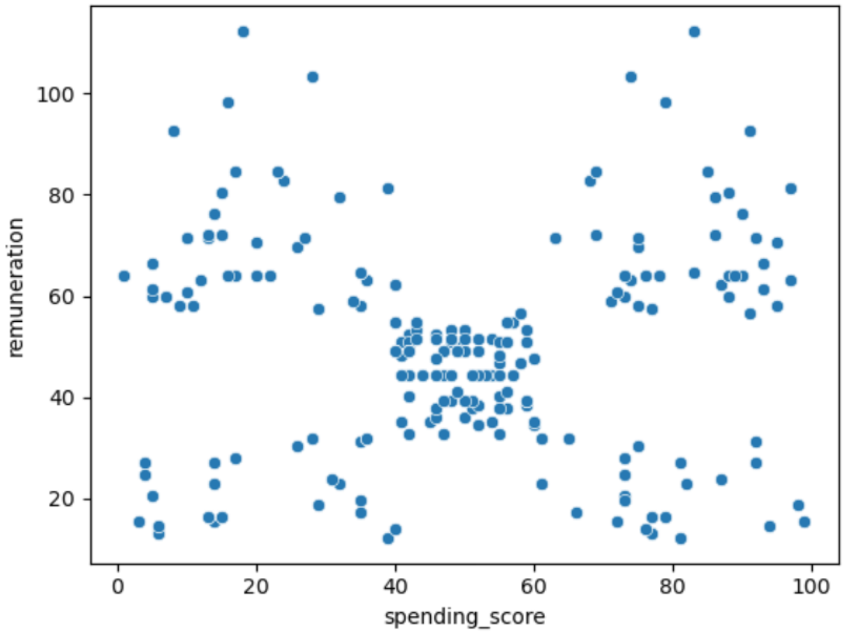
Finally I wanted to understand more about the customer groups so I plotted age against loyalty points. We see the largest cluster around the 30-40 range. This is to be expected from a demographic of video gamers, with the average fan being 35 years old, according to research by GameSparks (https://dataprot.net/statistics/gamer-demographics/#:~:text=The%20majority%20of%20gamers%20in,gamers%20are%20older%20than%2055).



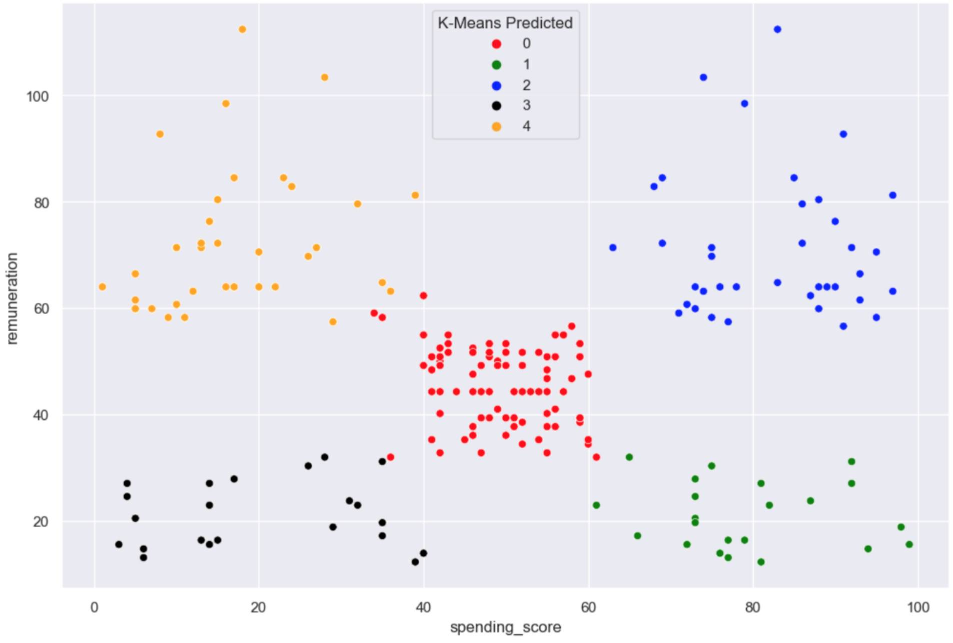
How can groups within the customer base be used to target specific market segments?

I used K-means to generate potential customers segments that could be targeted for marketing. After trying 3 and 4 clusters, I decided to use 5 due to the higher silhouette score, as well as the clearly clustered shapes of the original data (*see below).*

Original data:



Clustered data:



From a marketing perspective, it could be valuable to target the group in red with marketing around accessible, but not overly cheap products. This is also a fairly large group and so would be valuable to convert into sales. Their spending scores are also high. Given the previously mentioned correlation between high spending scores and loyalty, it is likely that many in this group can or would be loyal customers going forward.

One could also create a different marketing campaign that targets the blue and green groups with campaigns around extremely high quality products (for the high earning blues) and affordable options for the greens. Given the high spending score of the green group, and the previously-discussed correlation between loyalty and spending score, it may be advantageous to target this group with discounts for repeat-customers, given their comparably low income but high score.

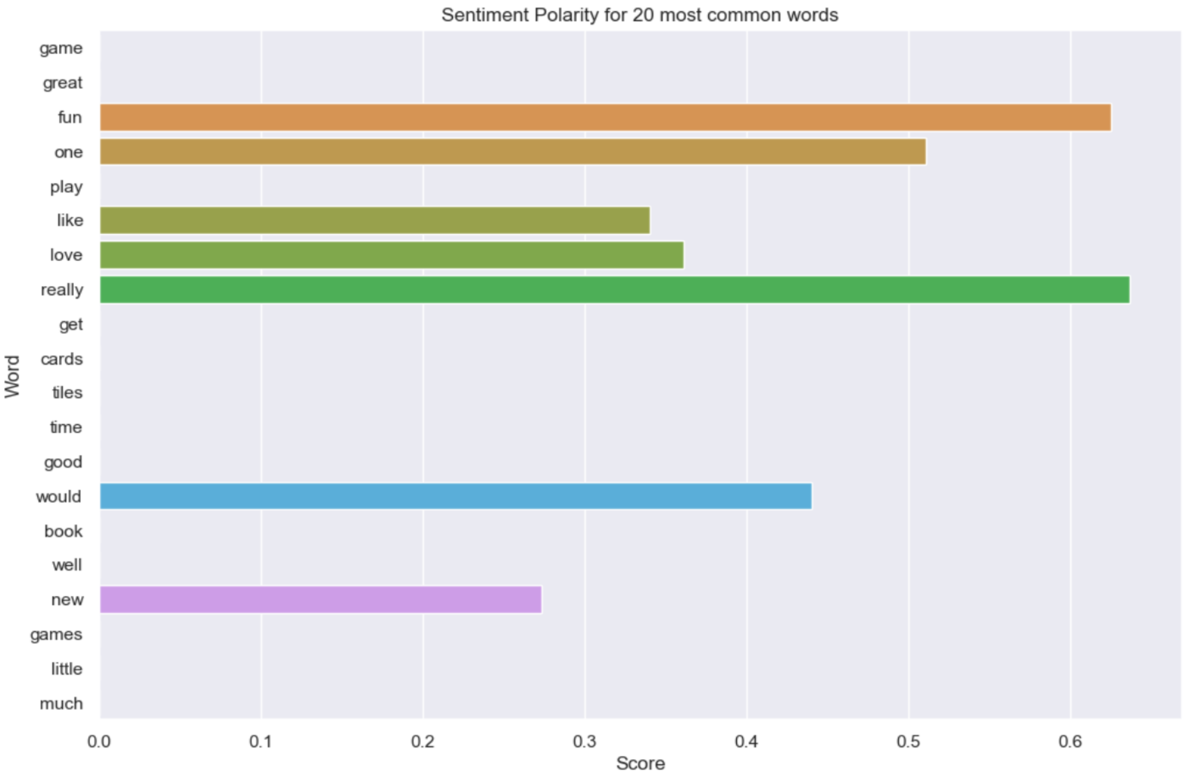
But most importantly marketing should be focused on the orange group. With their high income and low spending score, they have the most potential money to spend at Turtle Games. With successful marketing they could become highly profitable segment of their customer base.

How can social data (e.g. customer reviews) be used to inform marketing campaigns?

To answer this I performed sentiment analysis on a dataset of reviews and summaries of those reviews. To a degree I believe this can be used to inform future marketing campaigns by Turtle. However I also believe there are limitations and problems using this analysis framework.

Many of our most positive reviews mention children or childhood, for example (in summaries) “absolutely adorable, I was excited to teach my daughter”. Or another positive review that mentions buying the game for their son and him still playing it years later.

The sentiment analysis also indicates positive views of Turtle Games generally. From the analysis of the 20 most common words we get no negative results, only positive and neutral.



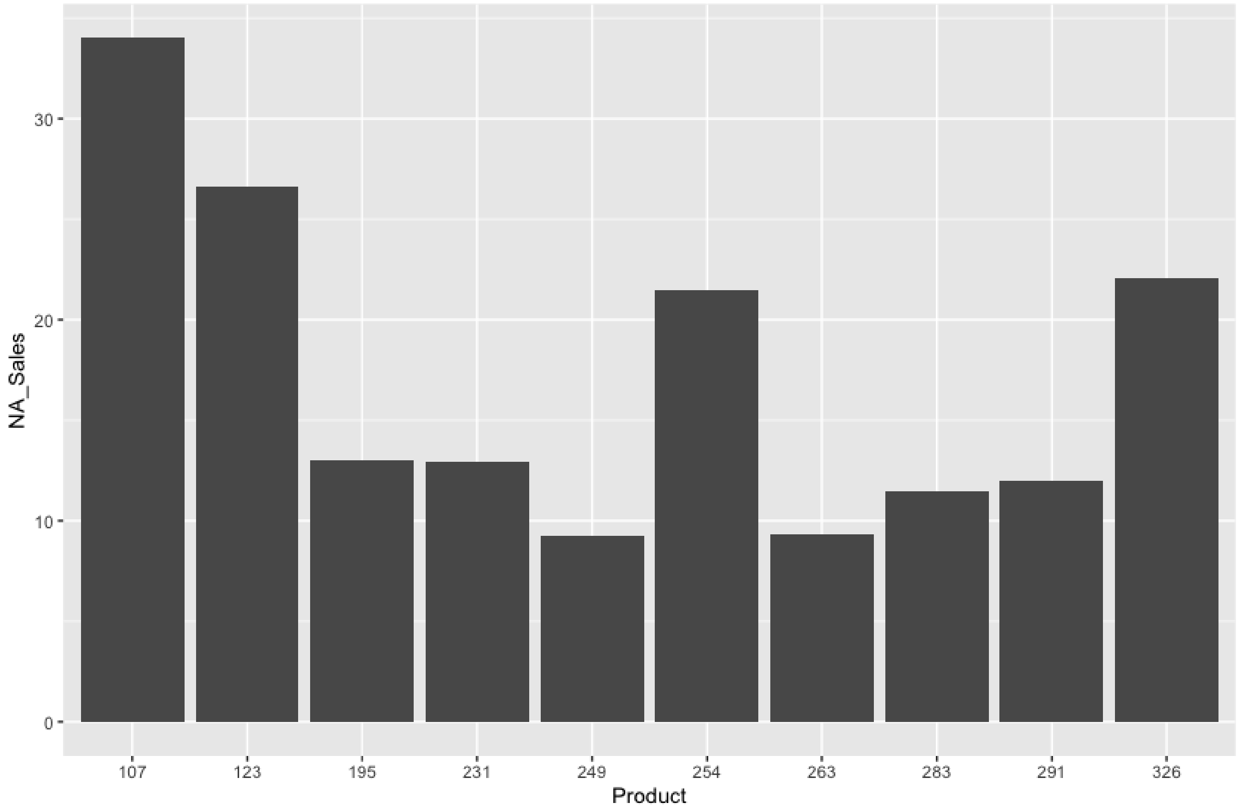
There are, however, issues with the sentiment analysis library. For example our worst rated review is actually a positive, yet nuanced, review. Some of the other ‘negative’ reviews merely mention that the game takes time to get into, but is then rewarding.

From a marketing standpoint we can draw two conclusions here: one might be to believe the sentiment analysis, and work to find and market games that have more ‘instant appeal’ so as to not gain the ‘negative’ reviews critiquing how long they take to get into. The other would be to do further analysis, perhaps to use a different sentiment analysis algorithm and see how the results compare? Given the results I would argue in favor of the second.

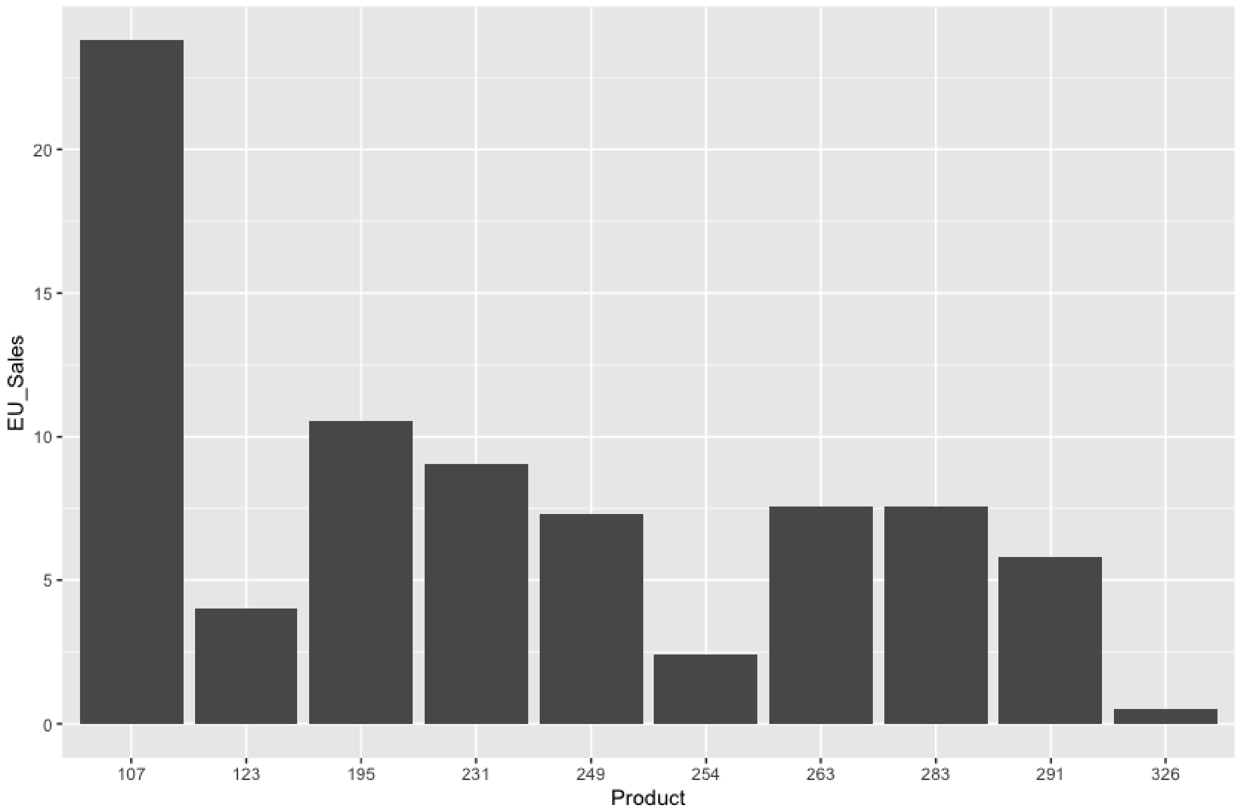
<u>What is the impact each product has on sales?</u>

With 352 unique products, I did not think it would be useful to create visualizations show each of them against sales. However, I have created visualizations to show the top 10 best selling products by region:

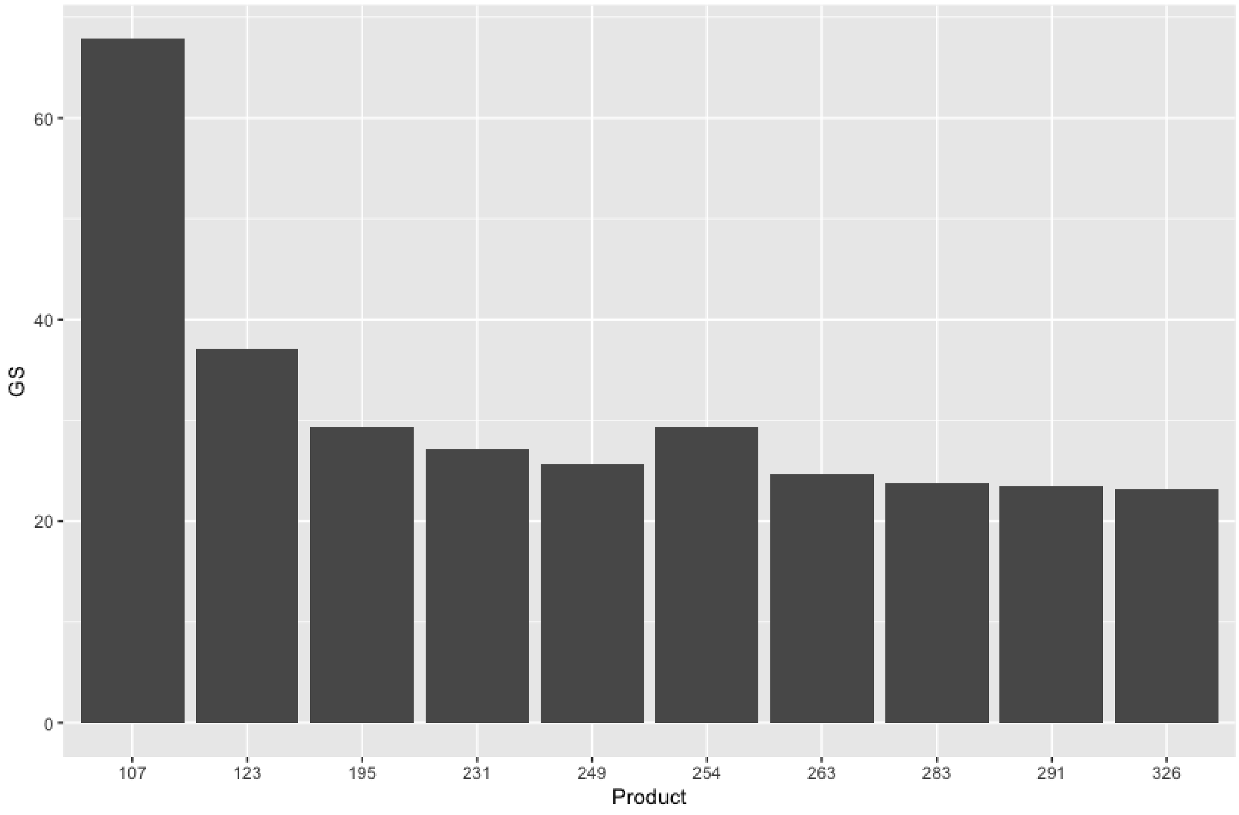
Best-Selling North American Products



Best- Selling European Products



Best-Selling Products Globally



Looking at these, we can see that North American and European top products are the same (if we investigate further we can see a lot of similarity further down the rankings, but I thought it would be too much information to show here). Furthermore, the top 10 Global Products account for 16.6 percent of total sales (out of 352 total products). As we will see subsequently, success between European and North-America markets has a high degree of predictive ability (i.e. high sales in one, correlates with high sales in the other).

Thus, I would argue for further investment in these top 10 products: we know they will perform well in the two largest markets, and we know that they make up a large market share of Turtle Games' total sales.

How reliable is the data in terms of skewness and kurtosis?

For this, I ran a skewness and kurtosis test on our sales columns, with results indicating the data is positively skewed and has more outliers than normal distribution:

NA Sales

Skewness
4.30921

Kurtosis
31.36852

EU Sales

Skewness
4.818688

Kurtosis
44.68924

Global Sales

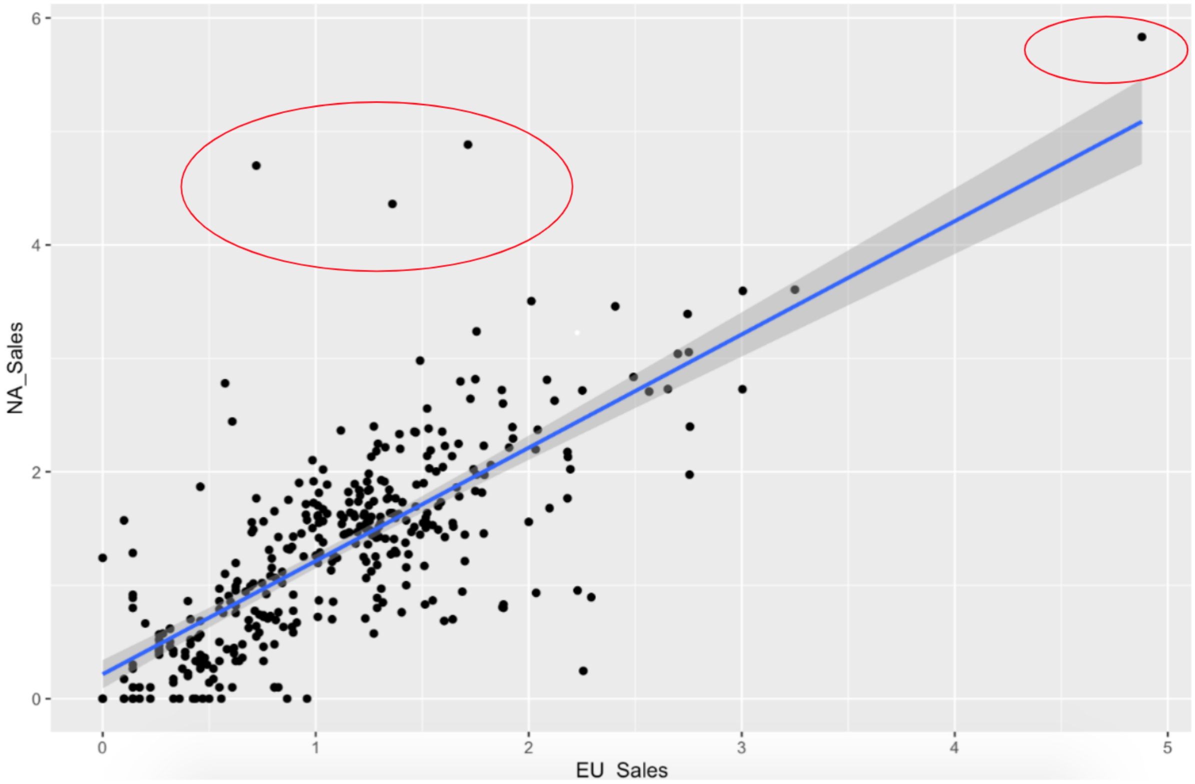
Skewness
4.045582

Kurtosis
32.63966

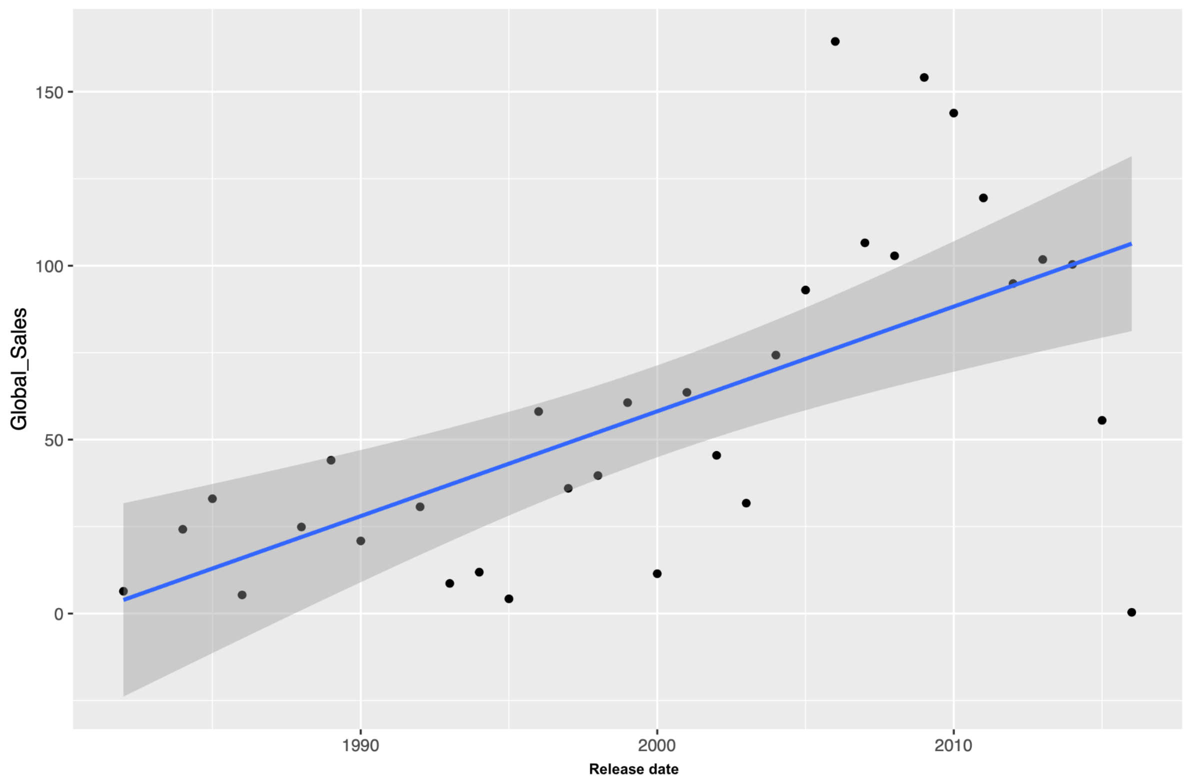
These results make sense given data I have also plotted, and can be (somewhat) easily explained when sales for various products are ranked. We see that a few video-games do exceptionally well and stay performing well in the market for years after their initial release, creating outliers. Some of these can be easily observed when looked at sales by region:

It is also important to note that the 3 outliers in the larger red circle do well in North America but not Europe. Currently we don't know why, but it may be that they were released earlier: as the second chart shows, there is a higher level of sales for earlier released games.

EU & NA Sales



Sales by release date



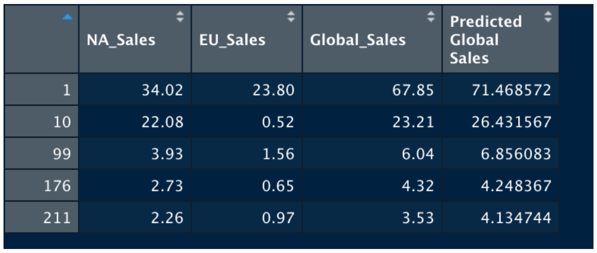
What the relationship(s) is/are (if any) between North American, European, and global sales?

Running a correlation test between the sales columns, we can see a strong correlation between all 3 columns. For Global Sales this intuitively makes sense: EU and North American markets make up the majority of Global Sales, and so changes in these markets will make the largest impact on the final Global Sales figures. However, we can also see a correlation between EU and North American sales, which is perhaps not so obvious. Games that sell well in one of these regions can be used as a metric to predict high sales in the other.

I ran a multiple-linear-regression test, using these figures, and tested it against real sales data. We can see that this test has a fairly high degree of success, predicting similar numbers to the real results.

Going forward I would like to run further tests, removing the outliers I mentioned above and seeing if I can create a more accurate prediction model.

It would also be interesting and useful to understand which factors drive success in one region, so that Turtle Games could potentially invest into these factors (for example, a specific publishing company). If the success of one region can be used to predict success in the other, we could invest in Europe (a smaller market, and thus cheaper) and see if a product does well. If it *does* do well, we could then invest in the North American market with a higher confidence it will succeed.



	NA_Sales	EU_Sales	Global_Sales	Predicted Global Sales
1	34.02	23.80	67.85	71.468572
10	22.08	0.52	23.21	26.431567
99	3.93	1.56	6.04	6.856083
176	2.73	0.65	4.32	4.248367
211	2.26	0.97	3.53	4.134744