The following activities examine data from Turtle Games. All data were checked for missing values and duplicates.

Activity 1

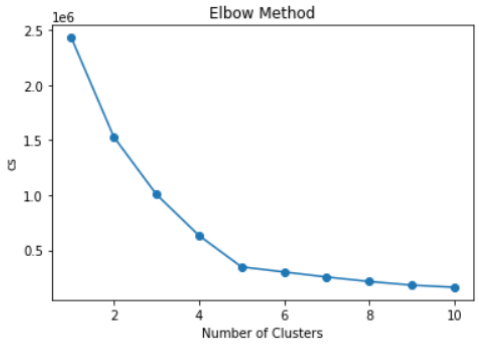
I assessed whether age, income, and spending could predict loyalty points by running a multiple regression in Python with Loyalty Points as the outcome variable and Age, Income, and Spending as predictors.

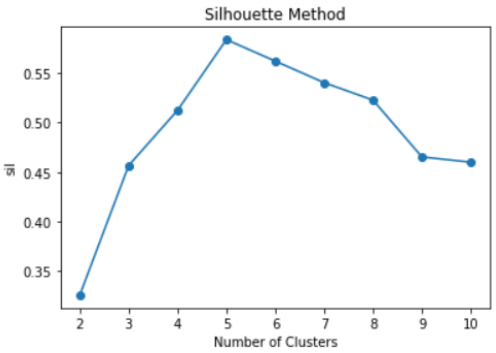
The model explained 84% of loyalty point variance. Income and spending had modest magnitudes of predictive power - higher income/spending suggest higher loyalty points. Loyalty points increase slightly, though reliably, with age.

Activity 2

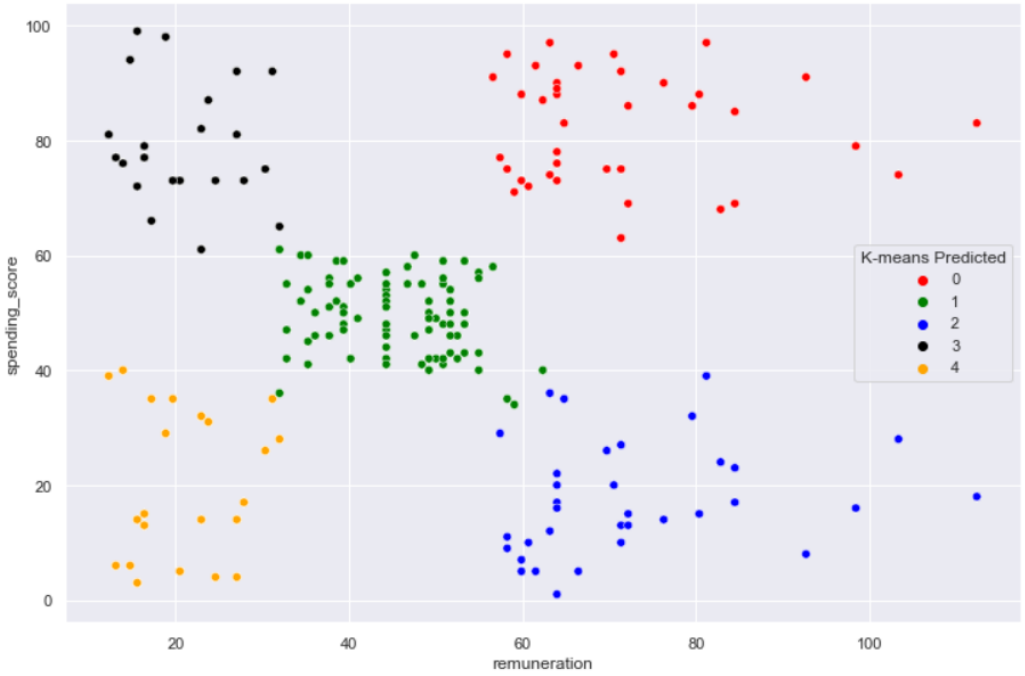
I wanted to investigate whether customers can be categorised by income and spending patterns. This could be useful in targeting loyalty club adverts. I ran a k-means cluster analysis, which explores pockets of association between two variables when little is known about the relationship beforehand.

I used the elbow and silhouette methods to choose cluster number by identifying when additional clusters cease to add explanatory value. These both indicate 5 clusters, seen by the plateau in both plots.





The plot below shows the 5 groups of income / spending behaviour. Surprisingly, lower incomes are clustered with low (cluster 4) and high (cluster 3) spending. Similarly, higher incomes appear linked to high (cluster 0) and low (cluster 2) spending. Why is income distributed in this manner - what distinguishes a high income/high spending customer from a high income/low spending customer? Answering this question may help convert low spenders to higher spenders.
Note, however, that data are clustered in regions of different shape/density; a generalised method of k means would be preferred.



Activity 3

I conducted a sentiment analysis on a sample of product reviews/summaries.This analysis estimates the positivity/negativity of unstructured text. Each review/summary was assigned a polarity score (-1=most negative; 1=most positive). The histograms show that reviews/summaries were mostly neutral to positive.

I next examined the top 20 negative and top 20 positive reviews/summaries. Several negative reviews/summaries indicate dissatisfaction instructions or assembly. This may be an area to investigate further; is this a general issue across products or specific to certain products?

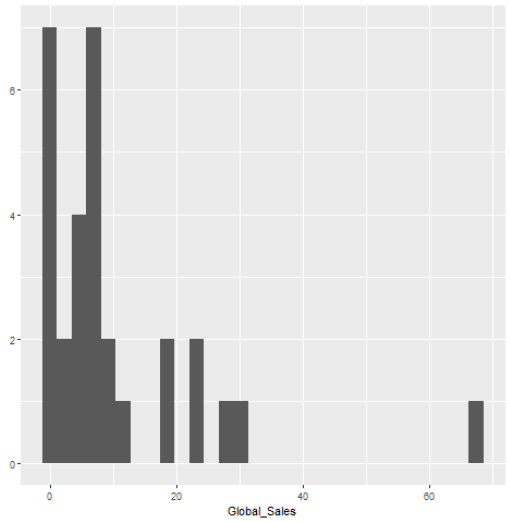
The positive reviews/summaries are less informative at face value. Perhaps their distribution across products would shed additional light? The top 20 positive reviews/summaries have all max positive ratings (+1); the negative reviews/summaries have max ratings that quickly fall towards neutral, reinforcing the more positive tone of these reviews/summaries.

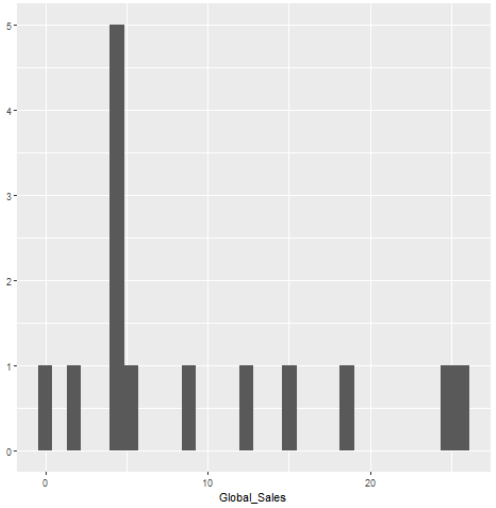
Activity 4

I explored sales by market (EU, NA) and by most popular platform (total sales across products per platform) to identify trends. North American sales totals are greater than EU totals, accounting for £884.64 million in sales compared to £578.67 million. The four most profitable platforms are Wii (£312.56M), Xbox360 (£253.81M), PS3 (£211.61M), and DS (£205.02M).

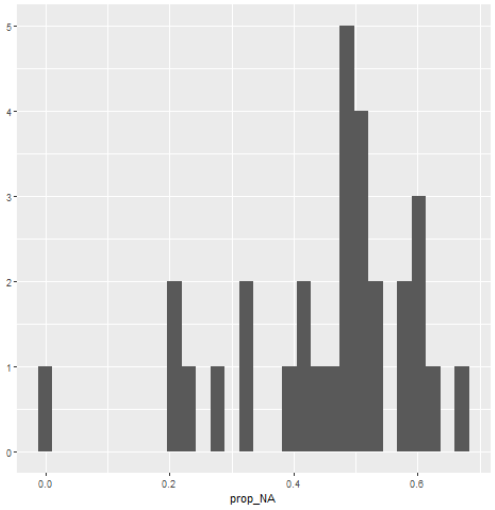
Next, I visualised global sales for each platform. The histograms below identify two features. FIrst, they show whether most products on each platform are equally profitable, or whether there are some very good/poor performers. Second, they show whether sales are composed of a small number of very profitable products or a large number of less profitable ones. Does

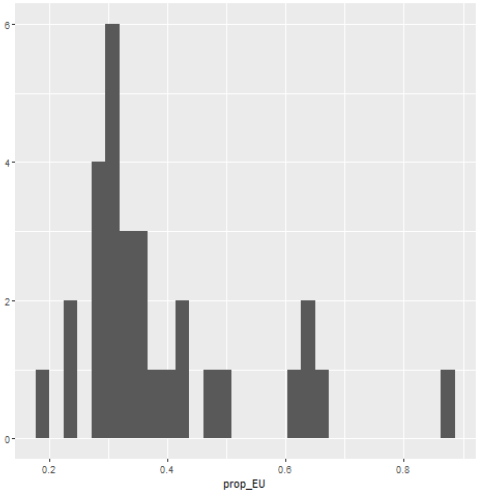
a particular platform have a hit title or a lot of smaller, steady performers? We can see that most of the Wii's titles sell in the £10M or less range (below left). In contrast, the DS has titles spread more evenly across profitability bands (below right).



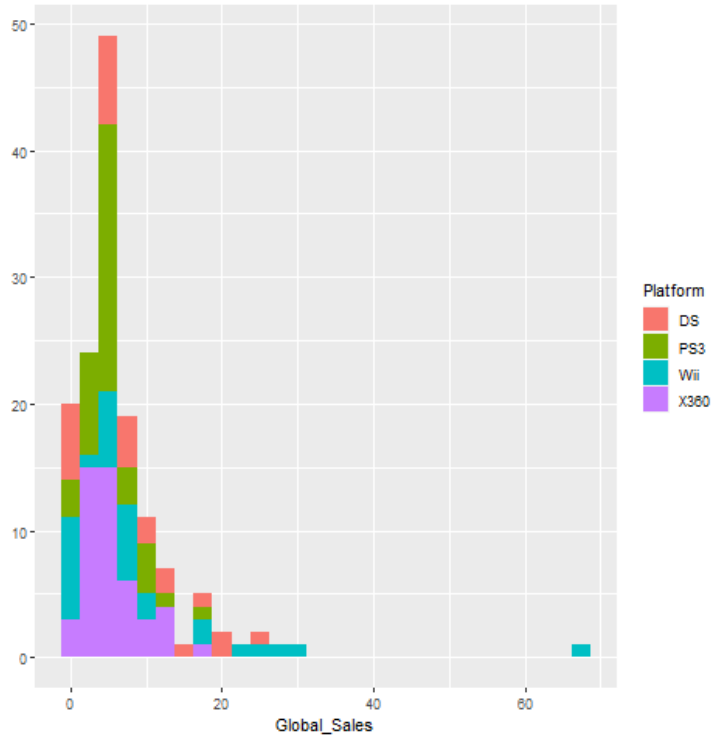


Next, I computed variables (prop_NA and prop_EU) to represent regional sales as a function of total sales (e.g, NA_Sales / Global_Sales). Most Wii titles do 40% + of overall sales in North America (below left), with only a few titles selling a greater share in the EU (below right).





The figure below shows a snapshot of the relative performance of the top platforms globally. The PS3 sells a wide range of titles in the lower total sales region. Wii sales are less concentrated, but have a few titles in the very high total sales region.



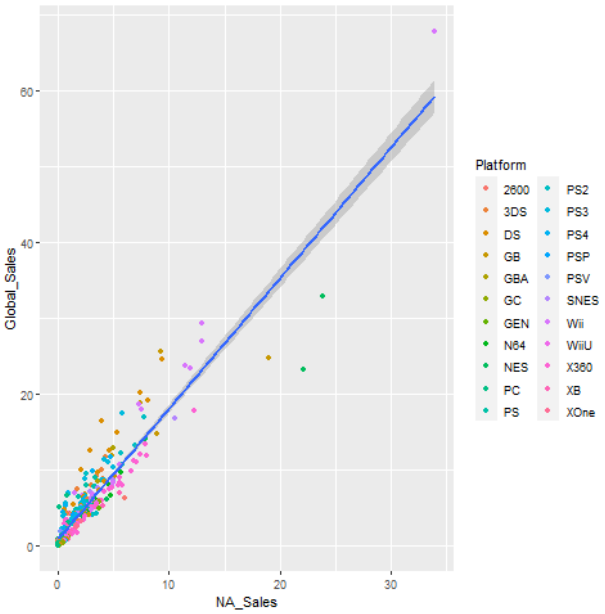
This is a starting point for understanding trends across platforms and regions. A good next step would be to look to identify any particularly weak/strong products for further analysis.

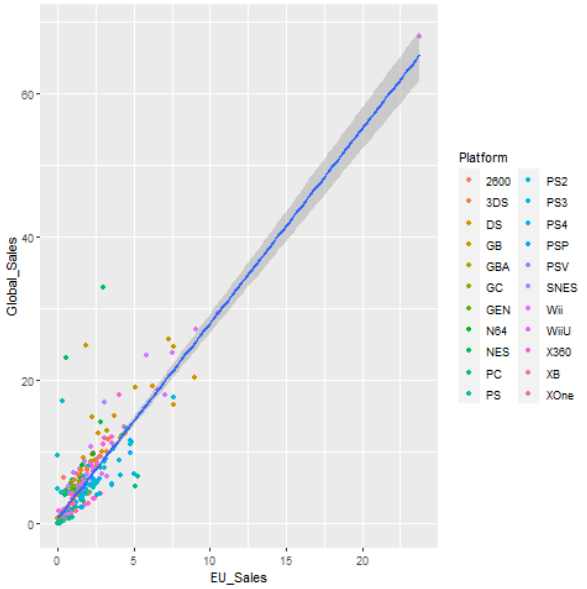
Activity 5

The goal of this analysis was to dig deeper in to the distribution of sales data, particularly with respect to individual products

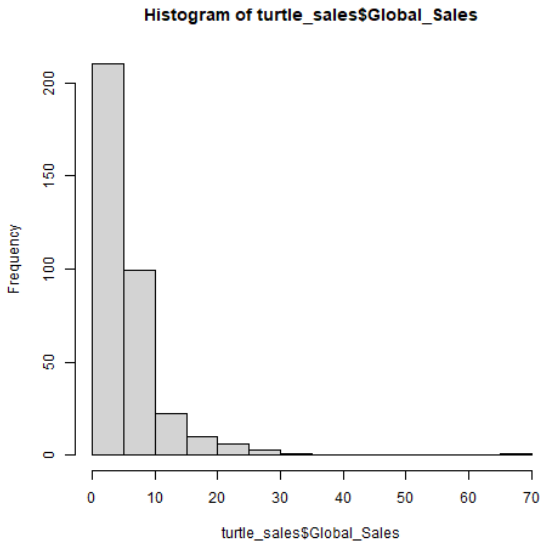
Summary statistics of Global Sales showed the average product sales were £5.33M globally, with the most profitable product achieving £67.85M and the least profitable product pulling only £10k. Median global sales per product was well below average (£4.32M), indicating that the average value is inflated by a few profitable products.

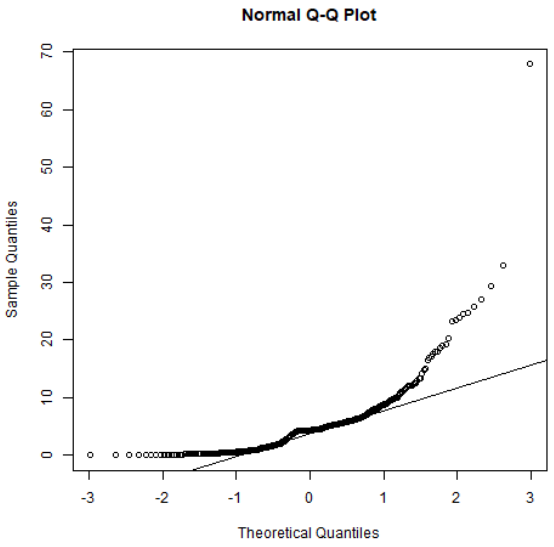
Next, I explored the association between regional sales and global sales by plotting NA Sales by Global Sales and EU Sales by Global Sales. These both showed a strong positive relationship. This visual trend was confirmed by Pearson's r (EUxGlobal r=.88; NAxGlobal r=.93).



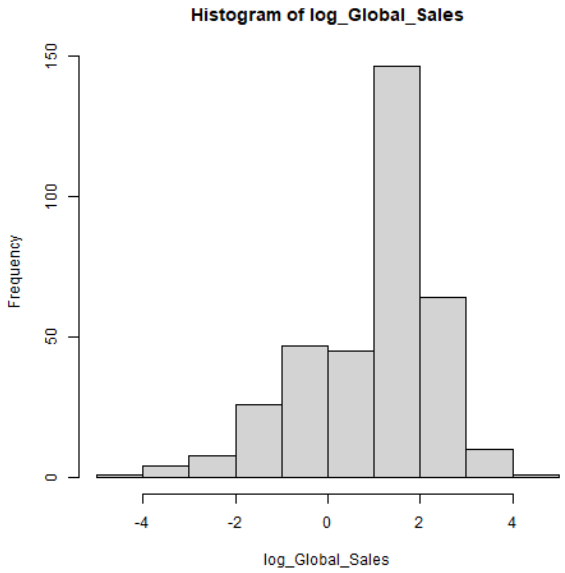


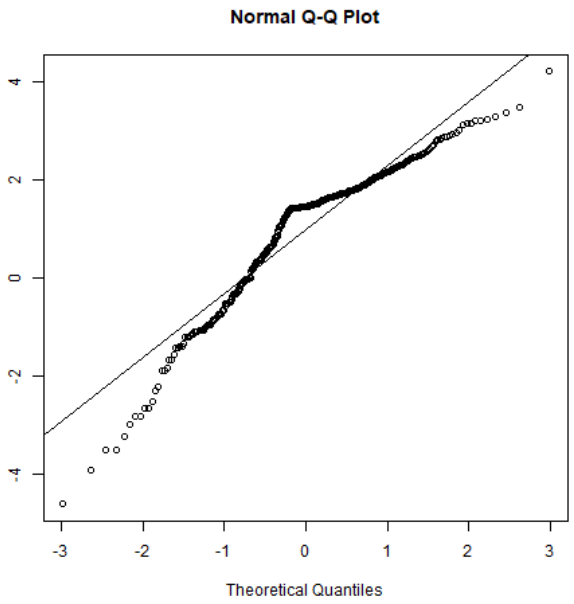
I assessed the distributional properties of Global Sales to determine its appropriateness for parametric analysis. The histogram, boxplot, and qqplot indicated that Global Sales data were not normally distributed. Most sales figures were concentrated in the lower range of the scale, creating a heavily positively skewed distribution. Global Sales data have a few utliers at the high end. NA and EU data displayed a similar pattern of skew.



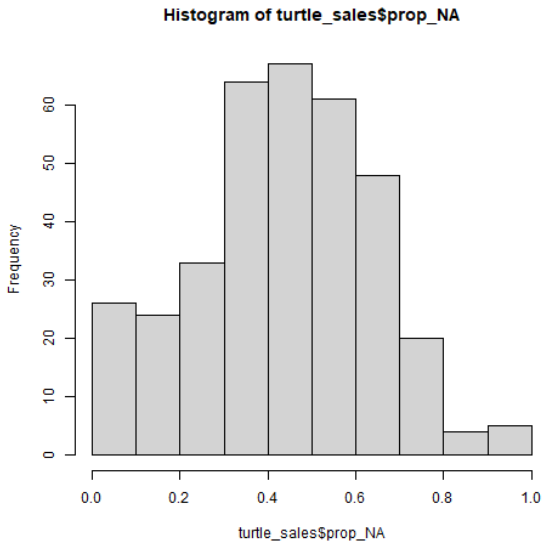


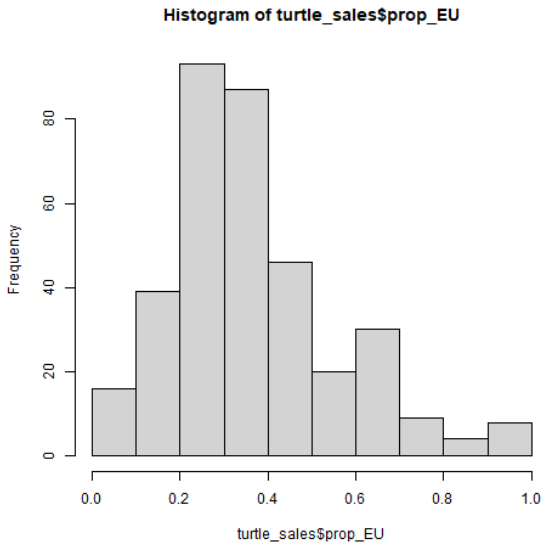
I explored ways to achieve normality in sales data while retaining all data points.A log transformation of Global Sales (see histogram and qqplot below) improved the distribution, but not sufficiently to pass for normal ($w = .9$, $p < .0001$).





Next, I plotted histograms for regional sales as proportions of global sales. Both of these variables had a much nicer looking distribution than the untransformed versions. However, they also failed to achieve sufficient normality for the Shapiro Wilk test (both $p < .001$).





It may be necessary to further subdivide the dataset by platform or product type. Otherwise, nonparametric alternatives may be more appropriate.

Activity 6

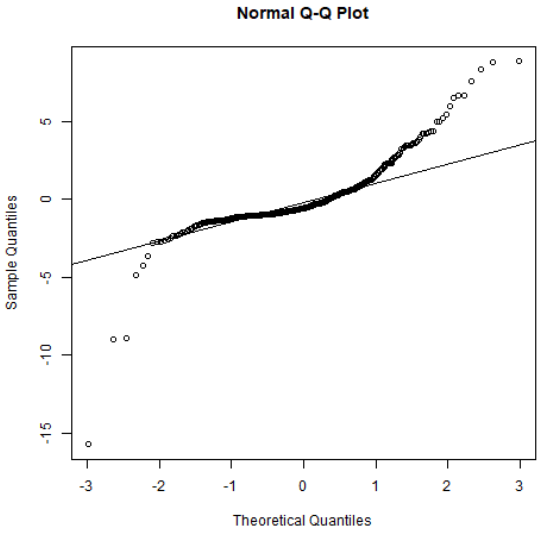
I created a predictive model of global sales using regional data from NA and the EU. NA sales and EU sales are both strongly correlated with Global Sales. The question is whether one region is more strongly predictive than the other of global sales trends.

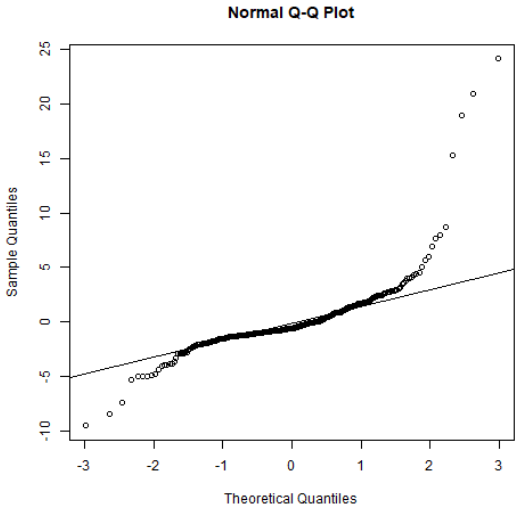
Since regression models are parametric, the issue of Global Sales distribution is relevant. The relationships between regional and global sales may not even be primarily linear (perhaps inverse exponential). Data are highly clustered at the low end of the top right quadrant, so smaller data points may over-influence the model. However, the GLM is fairly robust to violations of normality, so, for the activity, I will use untransformed data.

NA and EU Sales entered into a multiple linear regression with Global Sales as the outcome variable. The model was significant ($F(2,349)=5398$, $p<.0001$) and explained 97% of the variance in Global Sales ($adjR^2=.97$). North American sales ($b=1.155$, $p<.001$) and EU sales ($b=1.34$,$p<.001$) were significant predictors of Global Sales. Regional sales figures are reliable indicators of global sales trends.

The model was assessed on a handful of data points to see whether it hand;ed higher values of x as well as lower values. The error (observed-predicted) does increase at higher values of x (see below), but the proportional error (predicted/observed) does not. However, the residuals in single predictor models (NA and Global; EU and Global) were not normally distributed, as assessed by qqplots (NA sales model left, EU sales model right).

NA Sales	EU Sales	Predicted Global	Observed Global	Difference
2.26	.97	5.02	3.53	1.49
2.73	.65	5.32	4.32	1.00
3.93	1.56	8.40	6.04	2.36
22.08	.52	35.14	23.21	11.93
34.02	23.80	84.84	67.85	16.99





Based on these analyses, both regions provide reliable predictions of global sales. NA sales data is slightly better behaved and NA accounts for a larger share of global sales. Therefore, if a single predictor must be chosen, NA is preferred to EU.