Dan Thorneloe 10/09/2022

Couse 3 Assignment – Predicting future outcomes

Background

Turtle Games is a global game manufacturer and retailer with a business objective of improving overall sales performance.

Turtle Games wants to understand:

- how customers accumulate loyalty points
- how groups within the customer base can be used to target specific market segments
- how social data (e.g. customer reviews) can be used to inform marketing campaigns
- the impact that each product has on sales
- how reliable the data is (e.g., normal distribution, skewness, or kurtosis)
- what the relationship(s) is/are (if any) between North American, European, and global sales.

Analytical approach

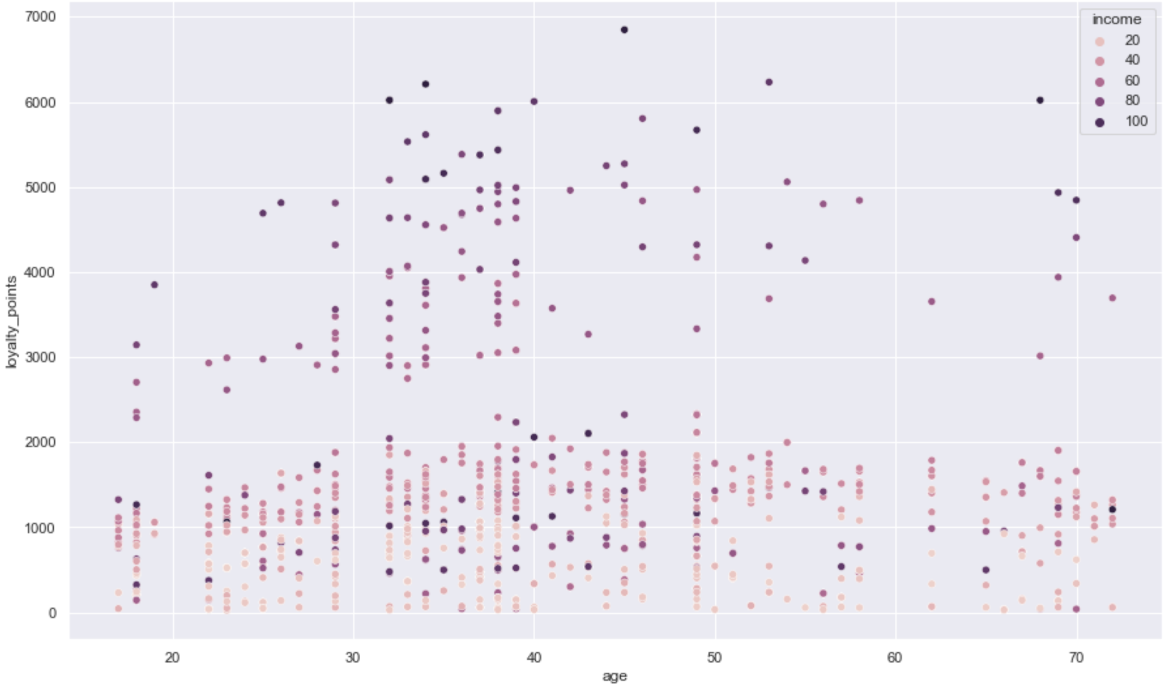
- Reviewed the data and established no missing or anomalous values
- Unnecessary columns dropped to simplify analysis
- OLS used for trend-line but note R and Seaborn have simpler tools for line-of best-fit
- k-means clustering used to identify clusters. Elbow and Silhouette method used to determine cluster numbers but experimentation yields best results
- Added hue to plots for third dimension to improve efficiency of analysis
- Preparation for NLP involved removing punctuation and duplicates, tokenising, removing stop words
- TextBlob and Vader (lexicon) sentiment analysis both used and produced directionally similar result. Vader is better suited to social media data i.e., short stings. Anomalous results seen in output e.g., “toy helped manage anger issues” was classed as negative and “5 star” was classes as) polarity
- R was used for quick insights, correlation, and prediction. I’d argue excel pivots are quicker for small data sets but for large data sets R is very useful as well as supporting more complex visualisation
- Correlation and predict functions in R were used to model relationships between variables

Insights

How customers accumulate loyalty points

- Loyalty points increases with spend_score/ income but bifurcate into two clusters: low loyalty point accumulators and high loyalty point accumulators
- 30s-40s age band has higher concentration of high loyalty points (more likely to have children?). 20-30s and 40-50s have significant, low loyalty points populations
- Clear division between low and high loyalty point populations (see below)

Dan Thorneloe 10/09/2022



Opportunities:

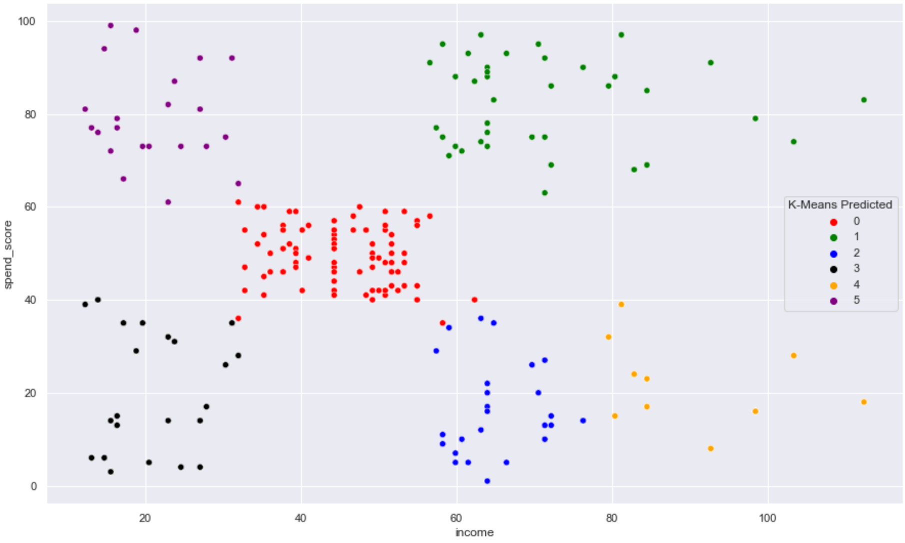
- Targeting high spend_score customers with low points e.g., shift product mix
- Target low loyalty points populations in the 20s to 30s and 40s to 50s groups

Further analysis/ considerations:

- Identify demographics/ behaviours that account for stark clusters (2) in loyalty points vs age. Income is a factor but something discrete driving he observation
- No data on % of customers not in loyalty programme (if opt in)

Identify groups to target specific market segments

- 6 clusters deemed to be optimal for spend vs. income
- Cluster 0 (mid spend/ mid income) is the largest with 767 customers, cluster 4 the smallest 123 customer



Opportunities:

- Clusters 2 & 4, high income/ low spend are a priority segment to look at re. sales volume/ product mix to move to cluster 1.
- Cluster 0 with the bulk of customers is also an attractive target segment

Further analysis

- Separation of clusters e.g., why are there no middle-income/ high spend customers is of interest as suggests discrete behaviour difference which could inform marketing

How social media can inform marketing campaigns?

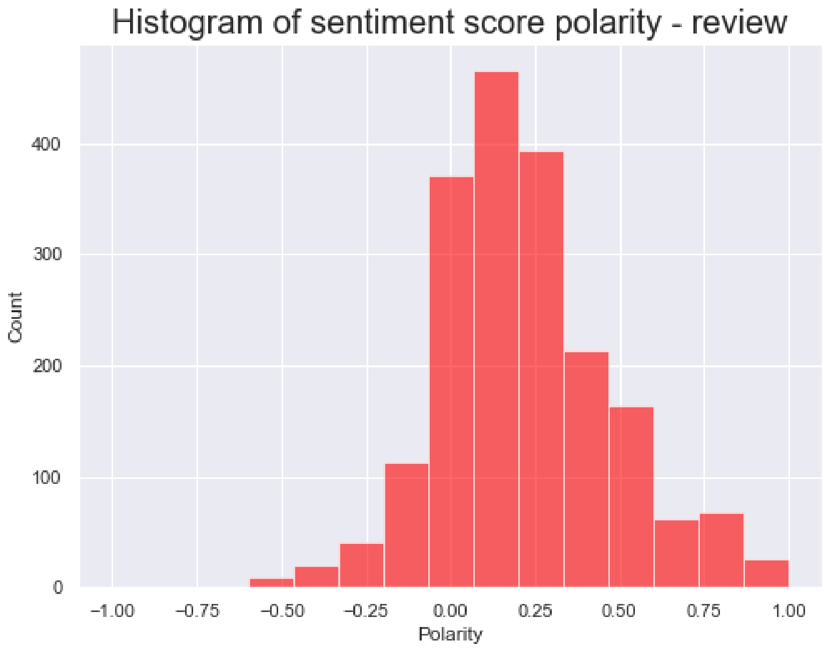
- Reviews, overall, are favourable (polarity >0 is positive). This is supported by an altenative method (VADER)
- Top 20 positive reviews are largely short expresions of delight but a few mention (grand) children
- Bottom 20 are more discrptive in their disatisfaction mentioning product quality, game quality etc.

Opportunity

- Focus campaigns on the joy gifts bring to receivers
- Align products to age categories more effectively to ensure suitability
- Target issues around product quality surfaced in reviews

Further analysis

- Refine analysis of positive reviews to identify more situations that are delighting customers to play back in marketing approaches.
- Align reviews to demographics to see what delight/ frustrates segments to inform campaigns
- Enhance NLP tools (train bespoke) to improve accuracy on summary and review analysis



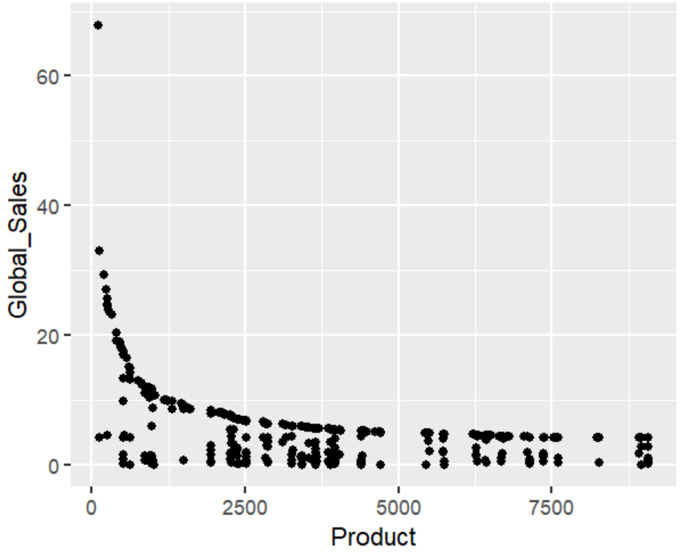
Dan Thorneloe 10/09/2022

Top 20 most negative reviews according to VADER sentiment analysis – items in red seem anomalous

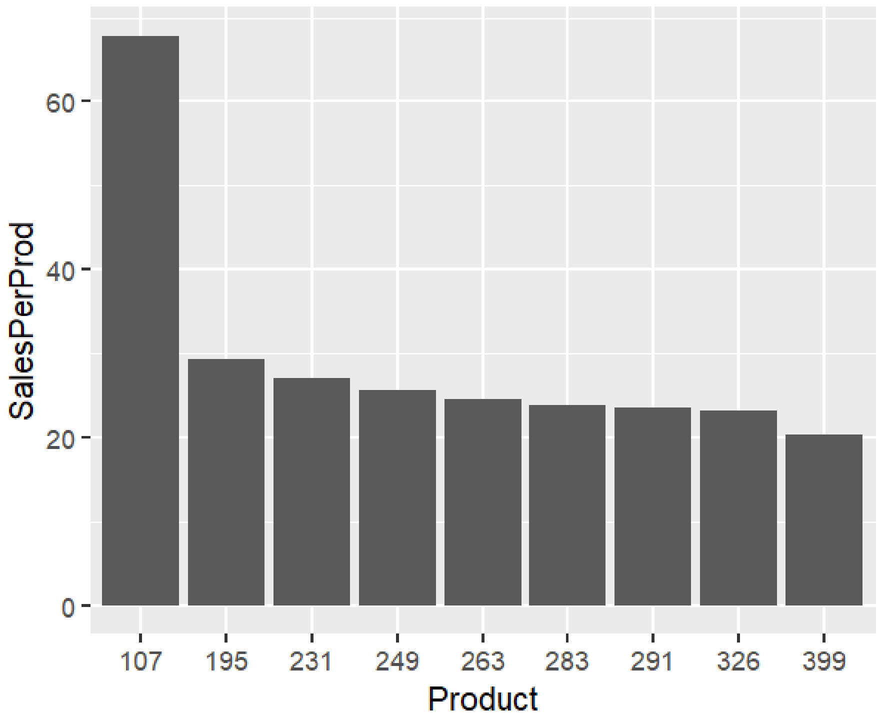
difficult
incomplete kit very disappointing
no more comments
a crappy cardboard ghost of the original hard to believe they did this but they did shame on hasbro disgusting
not a hard game to learn but not easy to win
i found the directions difficult
who doesnt love puppies great instructions pictures fun
different kids had red faces not sure they like
got the product in damaged condition
i bought this thinking it would be really fun but i was disappointed its really messy and it isnt nearly as easy as it seems also the glue is useless for a 9 year old the instructions are very difficult
great game poor quality
we really did not enjoy this game
not as easy as it looks
hard to put together
my 8 yearold granddaughter and i were very frustrated and discouraged attempting this craft it is definitely not for a young child i too had difficulty understanding the directions we were very disappointed
easytouse great for anger management groups
its ok but loses its luster quickly
rather hard for my 11 year old to do alone
smaller than we thought kind of disappointed in it
i really like this game it helps kids recognize anger and talk about difficult emotions

The impact of product on sales

- Products sell across multiple platforms so grouped to establish sale per product
- A few products significantly outperform a long tail of lower performing product; product 107 is the highest seller



Dan Thorneloe 10/09/2022



Opportunity

- Identify traits of higher sales product e.g., genre, platform to stock more high performing products
- Look at long tail of lower performing product to determine if volumes warrant cost of holding stock

Further analysis

- Differences in sales by geography
- Add Rest of World analysis as significant contributor to sales

Data Reliability

- Normality test shows we cannot assume the sample data comes from a population that is normally distributed.

Relationship(s) between North American, European, and global sales.

- NA sales is strongly correlated to global sales which we would expect given volume
- NA and Europe sales are correlated but with a value of 0.7 suggesting regional differences worth exploring
- More useful analysis would be to understand relationship between product, demographics etc. and regional sales to determine underlying driver e.g., % of sports titles launched in sales period as genre contributing most to sales was sports.