## Background & Business Context

How can *Turtle Games* (*TG*) improve sales performance? There are several recognised metrics[1]:

1. Gain new customers.
2. Avoid losing customers (i.e., churn).
3. Increase the frequency of transactions per customer.
4. Increase the size of customer transactions (e.g., upselling).

*TG* runs a loyalty program, the typical purpose is to reward repeat customers (discounts, multi-buy, bundles, pre-ordering) in return for valuable personal data[2]. What relationships can we find between this data and spending habits?

Which customers do Turtle Games need to keep satisfied? Which customers offer the most potential for growth in sales?

Is there any way to make future predictions of video games sales?

## 1. Analytical Approach

The process of preparing the data begin with importing and sense-checking the raw data using Python. Initial analysis revealed the following:

i. Turtle_sales.csv
   a. Missing values in the 'Year' column.
   b. Duplicate values in the 'Product' column.
   c. 'Global Sales' range from 0.01 to 67.85, average 5.33 (millions).
   d. Unique values for 'Platform' = 22, 'Genre' = 12, 'Publisher' = 24.

ii. Turtle_reviews.csv
   a. 'Age' range is 17 to 72, average 39.45.
   b. 'Remuneration' range is 12.30 to 112.34, average 48.08 (suggests outliers at the top end of the range).
   c. 'Spending Score' range 1 to 99, average 50.
   d. 'Loyalty Points' range 24 to 6847, average 1578.03.
   e. Duplicate values in the 'Product' column.

Data wrangling involved (i) editing column names, (ii) dropping unused columns, (iii) filling missing values with the mean value (year column), (iv) formatting text in the education column, (v) multiply remuneration by 1000 to show true value.

In attempting to join the two datasets by product_id, duplicate product_id's in the turtle_reviews.csv were found to correspond incorrectly to the review text (Fig. 1), which meant it was not viable to investigate patterns or trends between the sales data with the review data.

My first question was how is the spending_score calculated? According to the meta data the score is 'based on the customer's spending nature and behaviour'. A similar score used in marketing is RFM which includes 3 measures: Recency (how long ago a purchase was made), Frequency (how often purchases are made), Monetary Value (how much is spent) [3].

1.1 Predicting Loyalty Points

How should the spending_score relate to loyalty_points? Remember, loyalty_points are 'based on the point value of the purchase'. It can be assumed that there will be a strong positive relationship between these variables. This can be tested using a correlation, and indeed performing an OLS linear regression test on the data shows that 45.20% of the variance in loyalty_points can be explained by the spending_score (Fig. 2). The age and remuneration data were added to create an MLR model which improved explanatory variance of loyalty_points to 84.60% and all p-values were significant (Fig. 3). VIF was calculated which showed no multicollinearity between variables.

1.2 Group Classification

There appeared to be clusters or groupings within the scatterplots which could not be explained by the MLR model, hence a k-means clustering classification algorithm was deployed to gain insights into the relationship between spending_score and remuneration. The data was standardised, then by implementing the elbow (Fig. 4) and silhouette methods, 5 unique groups were identified (Fig. 5).

1.3 Sentiment Analysis

In order to gain predictive accuracy, several NLP models (Vader[4], TextBlob[5], BERT[6,7]) were tested to better understand customer sentiment towards the products purchased. The results will help *TG* to identify marketing opportunities that can increase sales and reduce churn rates. Vader and TextBlob required pre-processing the text to create a 'bag of words', while BERT models do not require this step, there are complications with the maximum length of the text.

1.4 Sales Data Analysis

The distribution of the data was positively skewed and Leptokurtic. Transformation using Log and Cube Root proved to be effective at normalising the distribution, although this was avoided due to complications with describing the results of an MLR model. The data was grouped by the product_id, then used to create an MLR model which explained 96.64% of the variance in global_sales (Fig. 25). This would be expected as 78.01% of global_sales consist of na_sales and eu_sales (Fig. 14).

## 2. <u>Visualisation and Insights</u>

2.1 Predicting Loyalty Points

The MLR model describes the relationship between the 4 variables (Fig. 3):

- For every increase of 1 in the age variable, there will be an increase of 10.84 in loyalty_points.

- For every increase of 1000 (£) in the remuneration variable, there will be an increase of 34.90 in loyalty_points.
- For every increase of 1 in the spending_score variable, there will be an increase of 34.69 in loyalty_points.

In plain terms, as the customers get older and earn a higher yearly income, spending power increases.

### 2.2 Group Classification

The table below shows the group classification predicted by the k-means algorithm, with relevant insights and the value of each group to *TG's* marketing and customer service departments. Groups are ranked (1 = most important) according to importance to increasing sales.

| Group Classification (Rank) | Avg. Age | Characteristics |
|---|---|---|
| high spending_score, high remuneration (Group 1) | 35.59 | Frequent and large buyers of the more expensive products. Possible promoters, bringing new customers with positive reviews/comments. 2nd largest group (Fig. 6). Average accumulated loyalty points are 3,988 (Fig. 7). |
| low spending_score, high remuneration (Group 2) | 40.67 | Perhaps new customers, not interested in the products, or casual buyers. Requires investigation. This group has potential to improve sales significantly. |
| medium spending_score, medium remuneration (Group 3) | 42.13 | 'bread-and-butter' customers, largest group (Fig. 6), 2nd highest avg. age. |
| high spending_score, low remuneration (Group 4) | 31.60 | Low average age, spending high relative to income, impulsive. Lowest avg. age. Possible opportunities to engage with promotions, multi-buy offers, pre-ordering of popular products. |
| low spending_score, low remuneration (Group 5) | 43.51 | Low risk, spend relative to income. Highest avg. age. Unlikely to buy impulsively. |

Further insights include:

- Loyalty points were not evenly distributed between groups (Fig. 7).
- There is no clear imbalance between groups when comparing by gender (Fig. 8) or education (Fig. 9).

### 2.3 Sentiment Analysis

Before implementation, it is important to verify the accuracy of a predictive model. The BERT models gave similar results (between 83.90% and 86.67% positive sentiment). In this analysis, the NLP Town[7] star rating model was chosen because the rating from 1 to 5 adds sensitivity to the model and it will be easy to understand for *TG* employees. According to this model most reviews contained positive sentiment (Fig. 10). The TextBlob model performed well, but tended to overrate sentiment as too positive (Fig. 11). The models were combined to retrieve the Top 20 Positive (Fig. 12) and Negative (Fig. 13) reviews.

### 2.4 Sales Data Analysis

47.18% of all sales are in North America, while 30.83% were from Europe (Fig. 14). The sales data is missing some regions (e.g., Asia). Average sales by product are £10.73 Million, and the top product represented 3.62% of total sales (Fig. 15). Popular video game genres include; Shooter, Platform, Action, Role-Playing and Sports (Fig. 16). There are 24 publishers, the largest being Nintendo which makes all the games for its consoles (Fig. 17), and this is reflected in the platform sales data which shows Nintendo making up 5 of the Top 10 (Fig. 18). Regional variations are evident across all segments, for example, a significant portion of sales for the Role-Playing genre are outside North America and Europe.

## 3. <u>Patterns and Predictions</u>

### 3.1 Predicting Loyalty Points

By implementing a 'train-test-split' methodology, predictions for loyalty_points can be forecast based on age, remuneration and spending_score (X variables). The model performed well while loyalty_points were close to the mean value of the data, with most predictions being accurate to within 500 loyalty_points (RMSE = 522). However, the model predictions can be wildly wrong, including negative scores (Fig. 19 & Fig. 20). Outside extreme input values of the X variables, the model performed well and can help *TG* predict which customers are likely to accumulate loyalty_points by spending more.

### 3.2 Group Classification

Classification of customers into groups allows for closer investigation of the different variables. For example, the importance of spending_score and remuneration in the accumulation of loyalty_points (Fig. 21 & Fig. 22). It helps in identifying the 'low spending_score, high remuneration' group to the marketing department as potential for increasing sales through promotional advertising.

### 3.3 Sentiment Analysis

The sentiment star rating scores allowed us to identify:

- Unsatisfied customers in target groups (Fig. 23 & Fig. 24).
- Products that customers are very satisfied with.

Keeping top customers (high spending_score, high remuneration) satisfied, can help turn them into brand promoters and bring in new customers to increase sales.

### 3.4 Sales Data Analysis

The MLR model predicted Global sales based on the sales in North America and Europe (Fig. 26). However, the dataset was too small to generate meaningful results. Ideally, any scenario involving predicting or forecasting sales would require time-based analysis. In this case, the data was limited to a snapshot in time, if the data was monthly then a time-series model could be built to any show trends and/or seasonality, and forecast future sales.

### 3.5 Key Recommendations

- Survey of ‘high spending_score, high remuneration’ customers to find Net Promoter Score (NPS)[9]. These customers should be kept satisfied through special marketing promotions.
- Survey of ‘low spending_score, high remuneration’ customers to understand more about spending behaviour.
- Use the loyalty program to collect more meaningful data that can be analysed and added to the predictive models. The primary goal is to find key metrics that explain the accumulation of loyalty points.
- Implement BERT model for predicting sentiment on any customer contact such as product reviews, chat and emails. Focus on (i) keeping customers satisfied (e.g., fast-track complaints), (ii) products trending negative or positive, (iii) and identifying and rewarding happy customers who actively promote the brand.

Further research should be focused on:

- Develop a logistic regression model using the target groups to identify more variables that may affect spending e.g., education, gender.

## References


1. What Are The 4 Methods to Increase Revenue'?
   https://personalmba.com/4-methods-to-increase-revenue
2. Loyalty Program: Definition, Purposes, How It Works, Example
   https://www.investopedia.com/terms/l/loyalty-program.asp
3. What Is Recency, Frequency, Monetary Value (RFM) in Marketing?
   https://www.investopedia.com/terms/r/rfm-recency-frequency-monetary-value.asp
4. VADER (Valence Aware Dictionary and sEntiment Reasoner)
   https://pypi.org/project/vaderSentiment/
5. TextBlob: Simplified Text Processing
   https://textblob.readthedocs.io/en/dev/
6. SiEBERT: sentiment-roberta-large-english
   https://huggingface.co/siebert/sentiment-roberta-large-english
7. NLP Town: bert-base-multilingual-uncased-sentiment
   https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment
8. Net Promoter Score
   https://en.wikipedia.org/wiki/Net_promoter_score

## Appendix

Fig. 1: Duplicate product_id values in turtle_reviews.csv

| | product_id | review |
|---|---|---|
| **1220** | 107 | Great addition to my game. Hopefully we'll get more of these for the new eddition. |
| **1602** | 107 | Great doll to go with the book & animals! Can't wait to read book with the doll to the grandkids! |
| **39** | 107 | I gave this as a Christmas present and it seems it went over quite well. |
| **1026** | 107 | Brutal. You have to really plan out and strategize to win. Just the first adventure alone is alr... |
| **831** | 107 | Buying and selling hotels sounds kind of boring on the surface. But if you went for that knee j... |
| **1811** | 107 | If you're lucky enough to get the right letters in your hand, you'll probably have a good time. ... |
| **635** | 107 | I really wanted to love this, I like it ok. The eggs themselves are fine. A nice product. Alth... |
| **438** | 107 | Easy-to-use great for anger management groups |
| **1413** | 107 | This is a very well designed and balanced set of expansions to the original Lords Of Waterdeep b... |
| **238** | 107 | brings very little wool |

> The duplicated product_id's do not appear to match with the reviews. For example, the reviews for product_id=107 are:
>
> - [1602] 'Great doll to go with the book & animals! Can't wait to read book with the doll to the grandkids!'
> - [1026] 'Brutal. You have to really plan out and strategize to win.'
> - [831] 'Buying and selling hotels sounds kind of boring on the surface.'
> - [1413] 'This is a very well designed and balanced set of expansions to the original Lords Of Waterdeep...'

The reviews are not related to the product_id, so I am assuming that the product_id is incorrect for this dataset.

Fig. 2: Linear Regression: spending_score vs. loyalty_points

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | y | **R-squared:** | 0.452 |
| **Model:** | OLS | **Adj. R-squared:** | 0.452 |
| **Method:** | Least Squares | **F-statistic:** | 1648. |
| **Date:** | Mon, 03 Jul 2023 | **Prob (F-statistic):** | 2.92e-263 |
| **Time:** | 17:12:01 | **Log-Likelihood:** | -16550. |
| **No. Observations:** | 2000 | **AIC:** | 3.310e+04 |
| **Df Residuals:** | 1998 | **BIC:** | 3.312e+04 |
| **Df Model:** | 1 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | -75.0527 | 45.931 | -1.634 | 0.102 | -165.129 | 15.024 |
| **X** | 33.0617 | 0.814 | 40.595 | 0.000 | 31.464 | 34.659 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 126.554 | **Durbin-Watson:** | 1.191 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 260.528 |
| **Skew:** | 0.422 | **Prob(JB):** | 2.67e-57 |
| **Kurtosis:** | 4.554 | **Cond. No.** | 122. |

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

> The summary indicates:
>
> - $R^2$: 45.20% of the total variability of `y` (how many loyalty_points the customer has collected), is explained by the variability of `X` (the customer spending_score).
> - `X`: The coefficient of `X` describes the slope of the regression line, in other words, how much the response variable `y` change when `X` changes by 1 point. In this model, if the customer spending_score ( `X` ) changes by 1 point, then the loyalty_points collected by the customer ( `y` ) will change by 33.06 points.
> - The `P>[t]` for the `X` coefficient value tests the hypothesis that the slope is significant or not. If the corresponding probability is small (typically smaller than 0.05) the slope is significant. In this case, the probability of the t-value is zero, thus the estimated slope is significant.
> - The last two numbers describe the 95% confidence interval of the true `X` coefficient, i.e. the true slope. For instance, if you take a different sample, the estimated slope will be slightly different. If you take 100 random samples each of 500 observations of `X` and `y` , then 95 out of the 100 samples will derive a slope that is within the interval (31.464 , 34.659).

Fig. 3: MLR Model: age, remuneration, spending_score vs. loyalty_points

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         loyalty_points   R-squared:                       0.846
Model:                            OLS   Adj. R-squared:                  0.846
Method:                 Least Squares   F-statistic:                     2555.
Date:                Sat, 15 Jul 2023   Prob (F-statistic):               0.00
Time:                        16:09:56   Log-Likelihood:                -10716.
No. Observations:                1400   AIC:                         2.144e+04
Df Residuals:                    1396   BIC:                         2.146e+04
Df Model:                           3
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const          -2268.7469     62.611    -36.235      0.000   -2391.569   -2145.924
age               10.8414      1.041     10.418      0.000       8.800      12.883
remuneration       0.0349      0.001     59.581      0.000       0.034       0.036
spending_score    34.6944      0.540     64.290      0.000      33.636      35.753
==============================================================================
Omnibus:                       12.107   Durbin-Watson:                   2.010
Prob(Omnibus):                  0.002   Jarque-Bera (JB):               12.786
Skew:                           0.186   Prob(JB):                      0.00167
Kurtosis:                       3.284   Cond. No.                     2.47e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.47e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

> Adjusted $R^2$ = 0.846, which means that 84.6% of the variance in the data can be explained by this model. This is a high value, so this model can be used to make predictions. `P>[t]` values are all less than 0.05 so all `x` variables are significant in explaining the variance of `y` .
>
> - For every increase of 1 in the 'age' variable, there will be an increase of 10.84 in 'loyalty_points'.
> - For every increase of 1000 in the 'remuneration' variable, there will be an increase of 34.90 in 'loyalty_points'.
> - For every increase of 1 in the 'spending_score' variable, there will be an increase of 34.69 in 'loyalty_points'.

Fig. 4: The Elbow Method

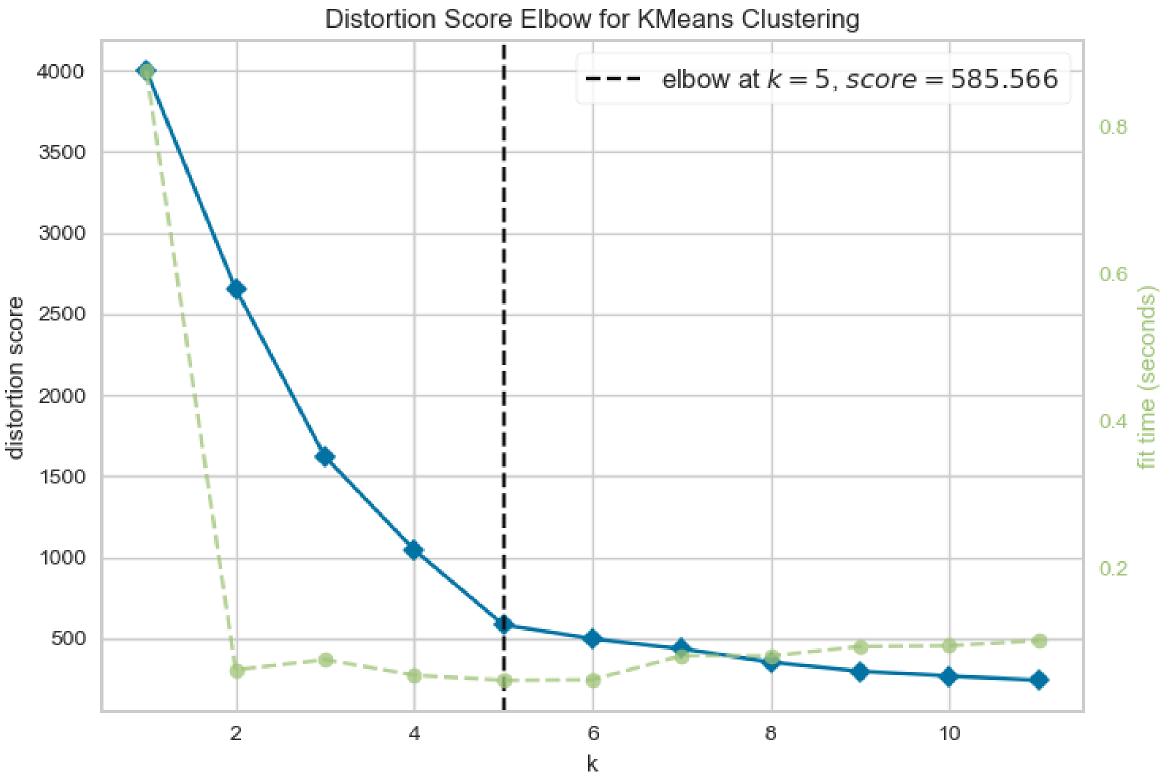


Fig. 5: K-means Classification Model: Identified 5 Clusters

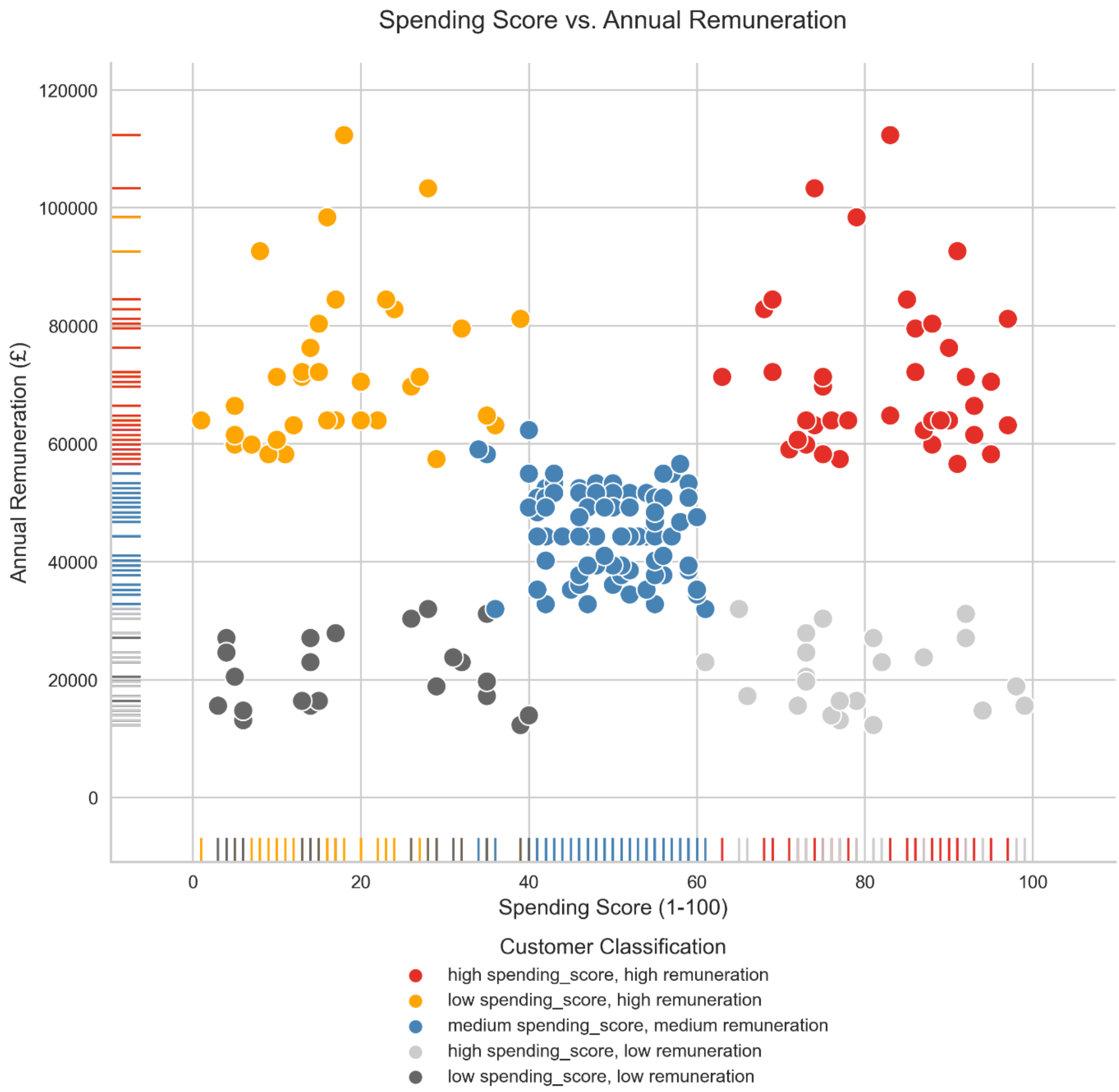


Fig. 6: K-means Classification Model: Percentage of Customers by Classification

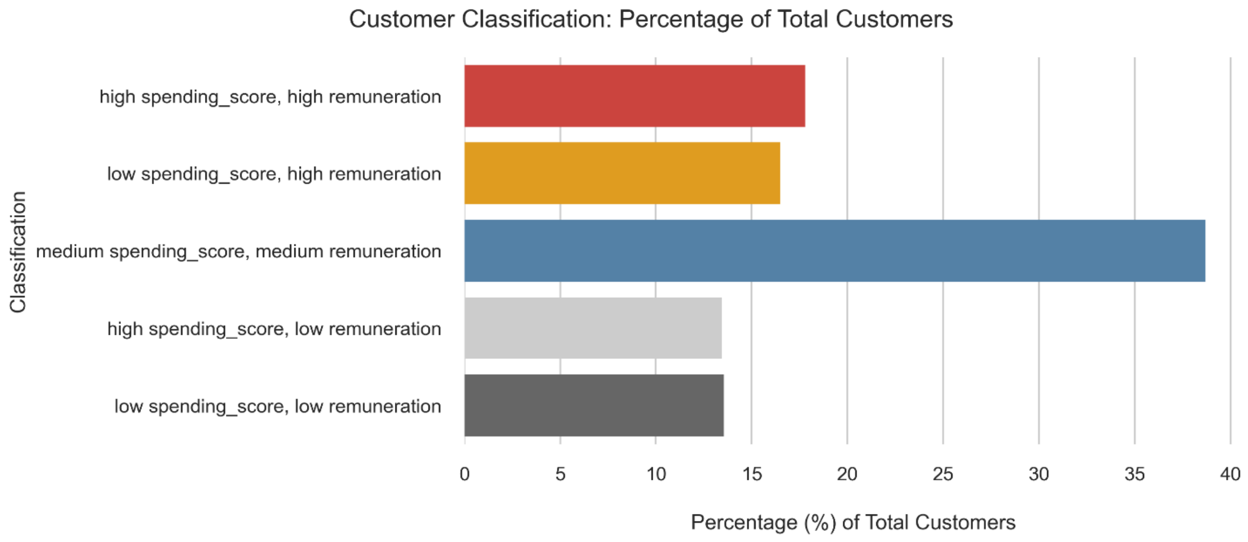



Fig. 7: K-means Classification Model: Average Loyalty Points Accumulated by Customer Classification

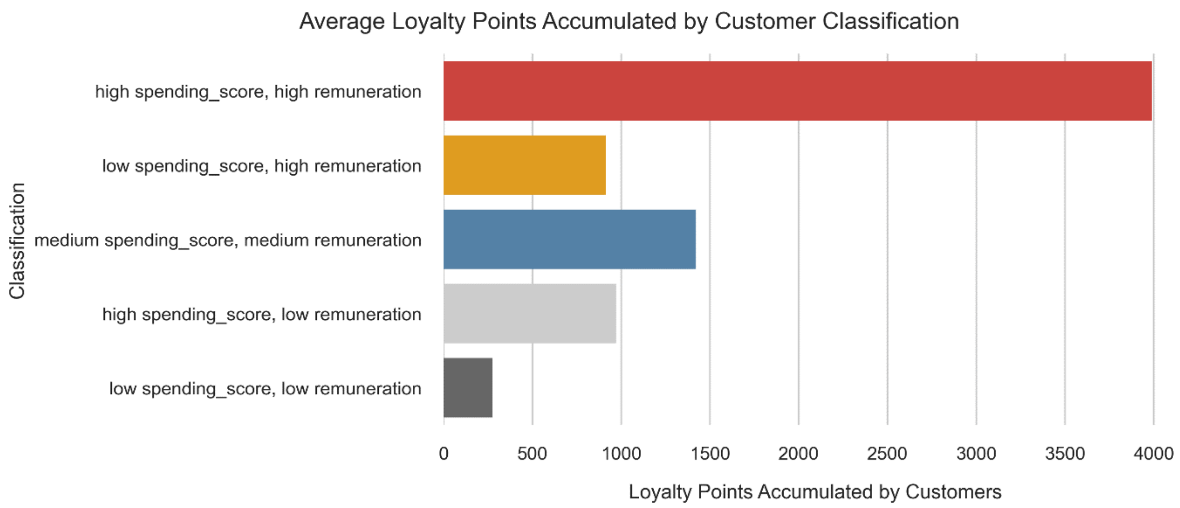


Fig. 8: K-means Classification Model: Average Loyalty Points Accumulated by Customer Classification & Gender

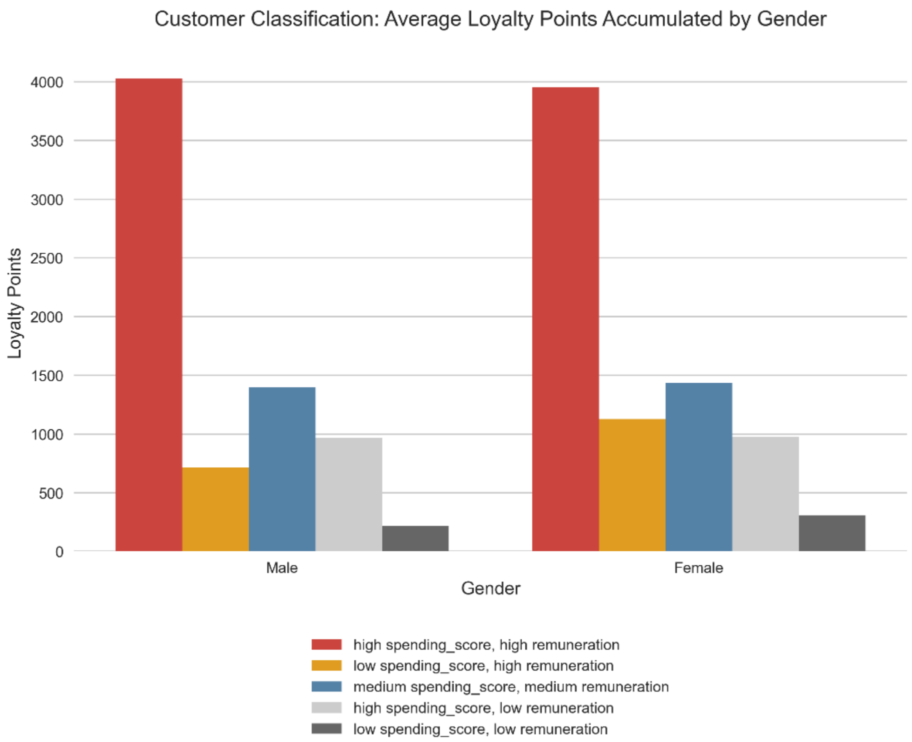



Fig. 9: K-means Classification Model: Average Loyalty Points Accumulated by Customer Classification & Education

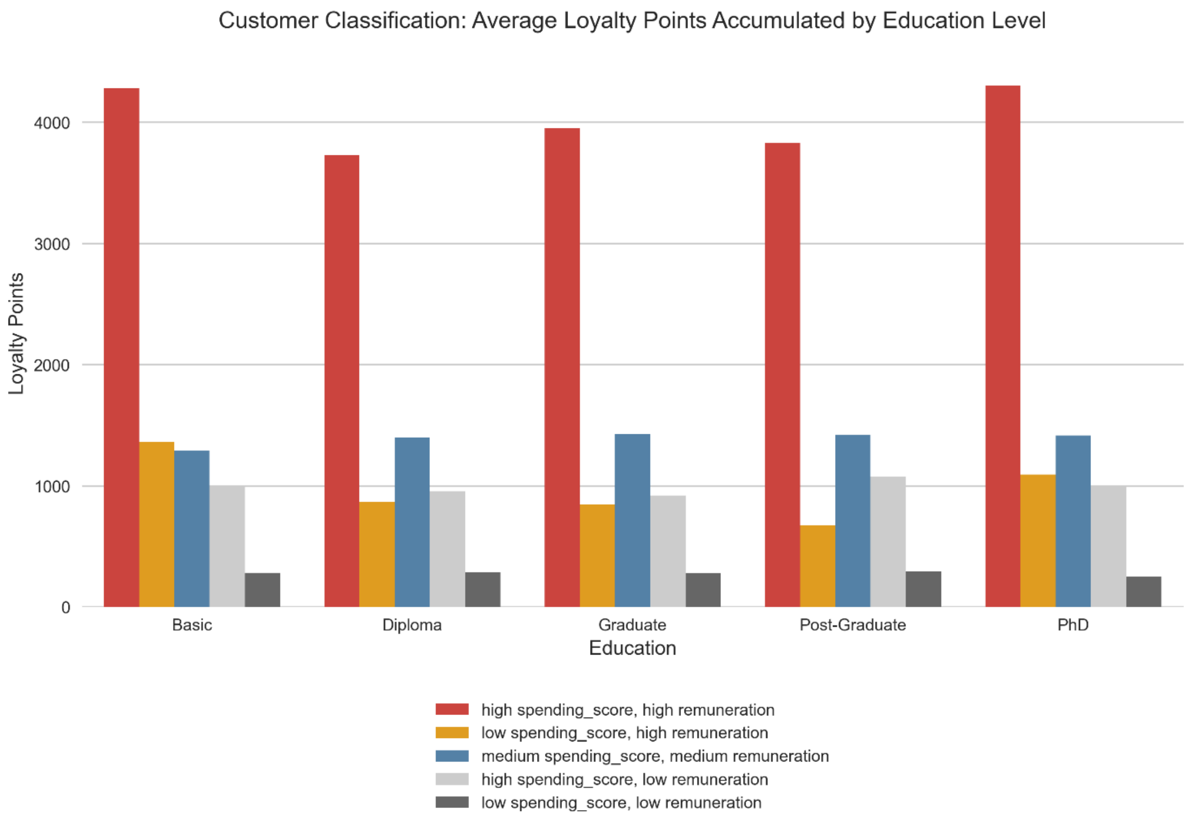


Fig. 10. Sentiment Analysis: NLP Town Model Star Ratings of Reviews

Fig. 11. Sentiment Analysis: Star Ratings vs. TextBlob

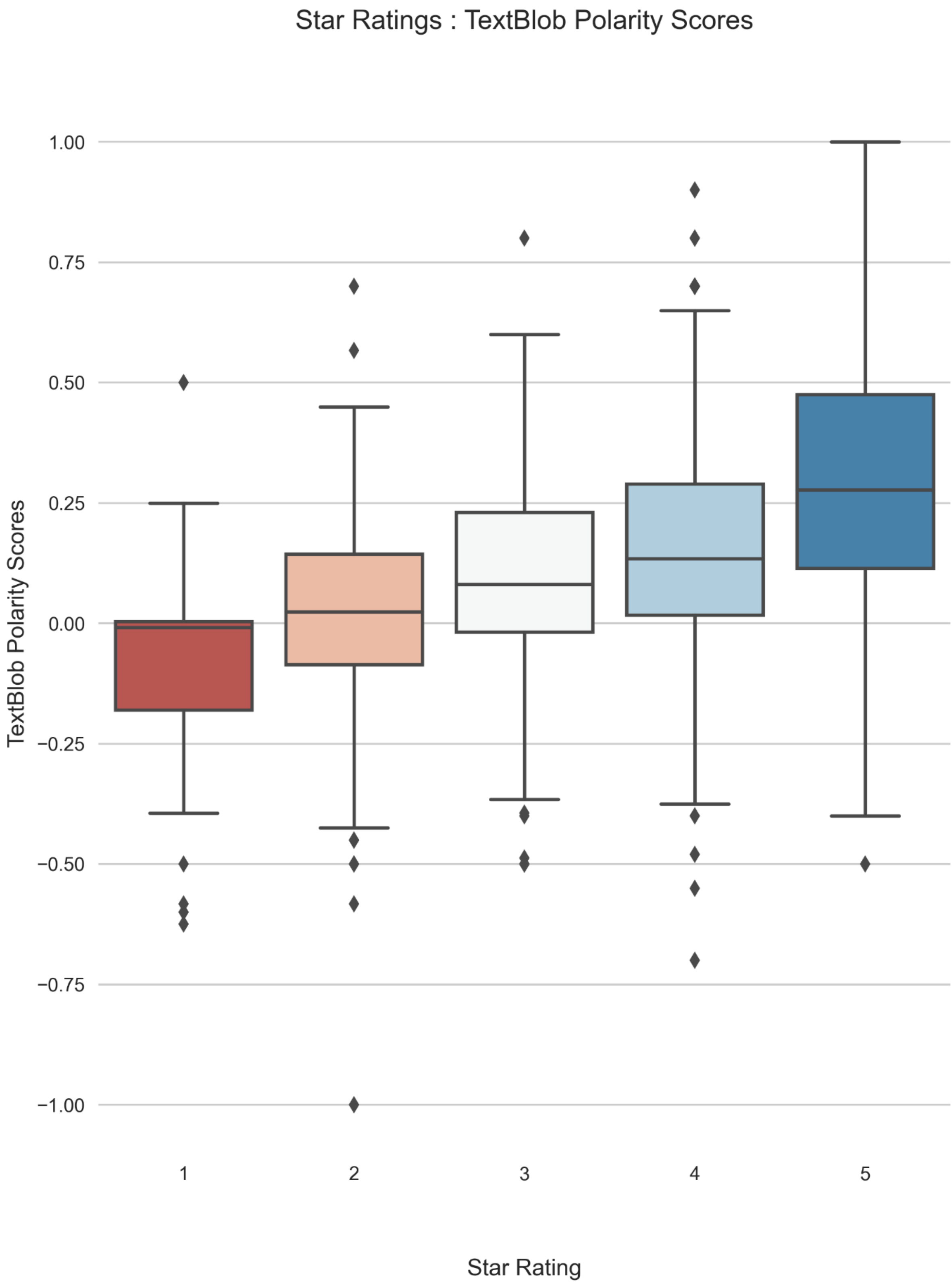


Fig. 12. Sentiment Analysis: Top 20 Positive Reviews

| | review | siebert_score |
|---|---|---|
| 36 | The pictures are great , I've done one and gave it to a friend of mine who likes dragons. | 0.999 |
| 1153 | Great product for my 5th edition adventure | 0.999 |
| 1931 | we play quiddler all the time at our home and the book is great | 0.999 |
| 1866 | I bought this as a family gift for Christmas. We've played it a couple times; it's great for my kids to create words and then add up their own points. | 0.999 |
| 1720 | It is the best thing to play with and also mind -blowing in some ways | 0.999 |
| 1719 | Great toy for puzzle lovers! | 0.999 |
| 1705 | A great creation tool. It helps me concentrate. | 0.999 |
| 1657 | These were perfect for Joey and the price was very good. | 0.999 |
| 1609 | this was perfect to go with the 7 bean bags , I just wish they were not separate orders. | 0.999 |
| 1518 | Great! Just like the photo! ;) | 0.999 |
| 1249 | This is a great accessory to the starter set. I would recommend this to anyone who owns the starter set. | 0.999 |
| 1245 | The best part I see is the box! What a wonderfully diverse and rounded set for the cost. I am so happy, and, as the DM, you know that if I am happy, my players are happy! | 0.999 |
| 1005 | Great! Thank you. | 0.999 |
| 194 | Awesome gift | 0.999 |
| 956 | Excellent quality. Beautifully manufactured. | 0.999 |
| 936 | This is a GREAT way to get minis and an awesome dungeon crawl system to boot. It has inspired me to get into the world of DnD more thoroughly and I have been having a blast painting the minis as well. The Dungeon cards are great for assembling into just about any sort of underground encounter you could imagine. | 0.999 |
| 703 | Great quality, very cute and perfect for my toddler! | 0.999 |
| 655 | Again, a great price for a puzzle. The best part is the carrying case. You can take it with you without worrying about losing pieces. | 0.999 |
| 648 | LOVE the Easter Story Egg! Such a great way to celebrate and remember Easter and what Jesus did for us! The book and eggs are beautiful and are a great way to teach children about the days leading up to Jesus' death and resurrection. Also, the customer service fro Star From Afar Kids is EXCELLENT! We had an issue and contacted them by email. They were quick to respond and wonderful in helping us! Will recommend The Easter Story Egg to everyone I know! | 0.999 |
| 620 | Great Easter gift for kids! | 0.999 |

Fig. 13. Sentiment Analysis: Top 20 Negative Reviews

| | review | siebert_score |
|---|---|---|
| **4** | As my review of GF9's previous screens these were completely unnecessary and nearly useless. Skip them, this is the definition of a waste of money. | 1.000 |
| **406** | This game is extremely boring and takes far too long. | 1.000 |
| **1123** | The product description specifies that this product "contains 50 miniatures" BUT THIS IS FALSE.--The package actually only contains FIVE miniatures. VERY DISAPPOINTING, especially considering a COMPLETE STARTER GAME SET-UP costs less than $25includingpostageonAmazon(andthis5-pieceboostercostme13$ including postage). | 1.000 |
| **847** | Do not waste your money on this. There are no plastic parts whatsoever. Its all cardboard and paper. the tiles slide around because they don't lock into place like the "real" versions of this game. If it wasn't a gift I would return it immediately. | 1.000 |
| **131** | This is advertised as stickers to build your own robot, what I received was the cover of the book with one sheet of stickers folded inside (not even stapled just stuck inside) and they were not even robots, they were firetrucks. | 1.000 |
| **602** | Received defective product. I could see that the largest egg was not closed inside the box. After further inspection, its defective & won't close. Very frustrating as Im now dealing with hassle of return. | 1.000 |
| **593** | If I could give this egg zero stars I would. It is poorly made and rudiculously hard to open. What should be a tender moment spent with your children is a huge headache. I had to use a knife to open it and the knife literally broke off. That is how difficult it is. Horrible product. Dont buy. | 1.000 |
| **623** | Was so dissappointed in this. I could not open it, no matter how hard I tried. | 1.000 |
| **359** | This is horrible! The directions are very hard for a child to read and comprehend themselves. The yarn made a mess. My daughter was so excited to get this and cried when she couldn't understand how to make them. I would not recommend this to anyone! | 1.000 |
| **337** | Only buy this for an adult who is super patient and likes to be covered in fuzz. It was a frustrating project to do with an 8 year old. There will be crying when the glue doesn't hold and the puppies head falls off. It makes a terrible mess and there is not even that much yarn included. I would never be able to make the puppies look anywhere close to the ones in the book. Ours turned out crazy looking. Book went straight to Goodwill. | 1.000 |
| **273** | Cute idea - horrible execution. If you want your child in tears then this is your book. My seven year old got very frustrated with this whole thing. | 1.000 |
| **182** | Incomplete kit! Very disappointing! | 1.000 |
| **174** | I sent this product to my granddaughter. The pom-pom maker comes in two parts and is supposed to snap together to create the pom-poms. However, both parts were the same making it unusable. If you can't make the pom-poms the kit is useless. Since this was sent as a gift, I do not have it to return. Very disappointed. | 1.000 |
| **1804** | I'm sorry. I just find this product to be boring and, to be frank, juvenile. | 1.000 |
| **882** | A crappy cardboard ghost of the original. Hard to believe they did this, but they did. Shame on Hasbro. Disgusting. | 0.999 |
| **890** | The game tiles, board, and tile stands are all made of paper. After using few times it will not sustain. Paper board tiles will move on the board making the game messy and inconvenient to manage. Its a shame what they have done to such a brilliant game. | 0.999 |
| **1003** | If you play Dungeons and Dragons. Then you will find this board game to be dumb and boring. Stick with the real thing. | 0.999 |
| **1119** | Here is my review, cross-posted from boardgamegeek.com: I have fond memories of D&D from my youth, that I occasionally attempt to recapture. I remember the sense of vague foreboding conjured by RPG-like board games such as Runequest and Spacehulk. I had read some favorable reviews of the new "RPG-as-boardgame" series promoted by Wizards on the Coast, and I was excited to try this. I splunked down my hard-earned cash and amazon duly delivered a hefty box of dreams to my doorstep. Wrath of Ash... | 0.999 |
| **173** | Horrible! Nothing more to say Would give zero stars if possible | 0.999 |
| **1315** | Got the product in damaged condition | 0.999 |

Fig. 14. Total Sales by Region

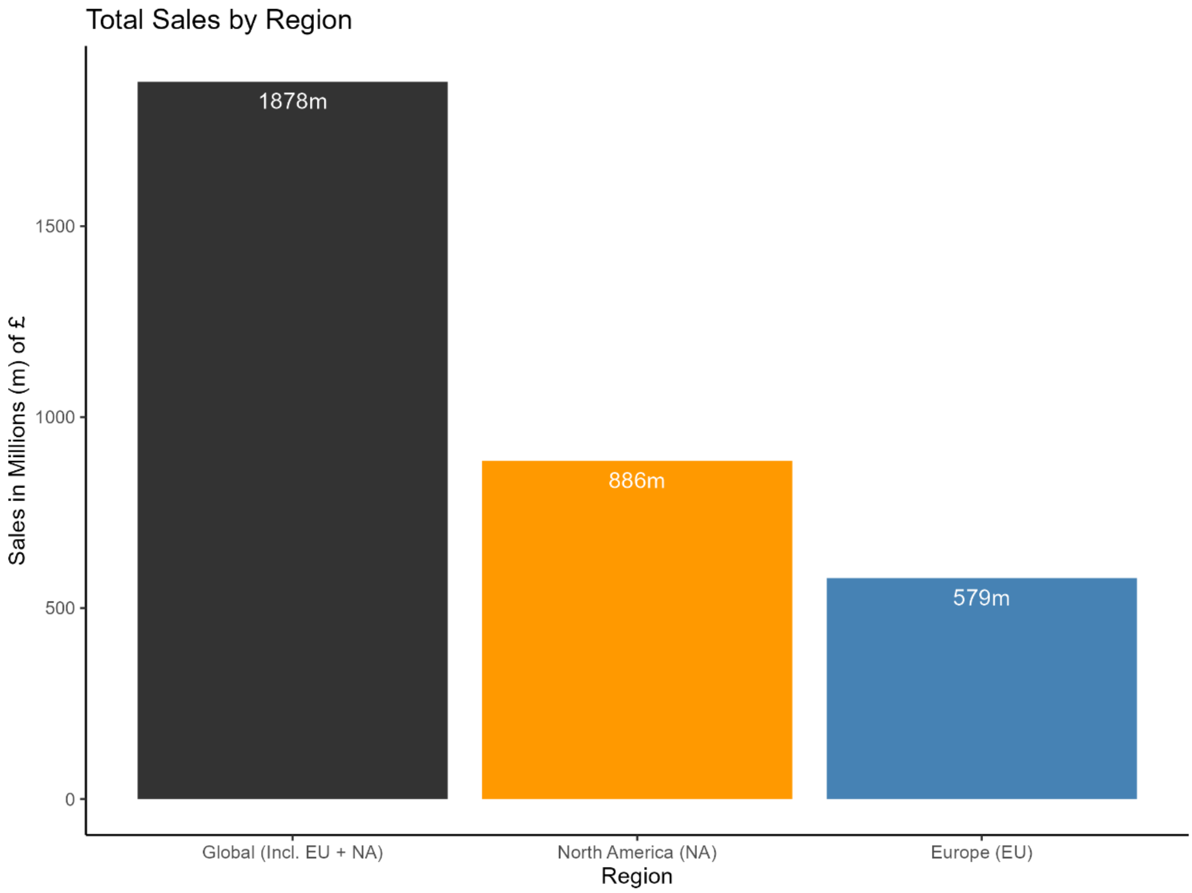


Fig. 15. Total Sales by Product ID

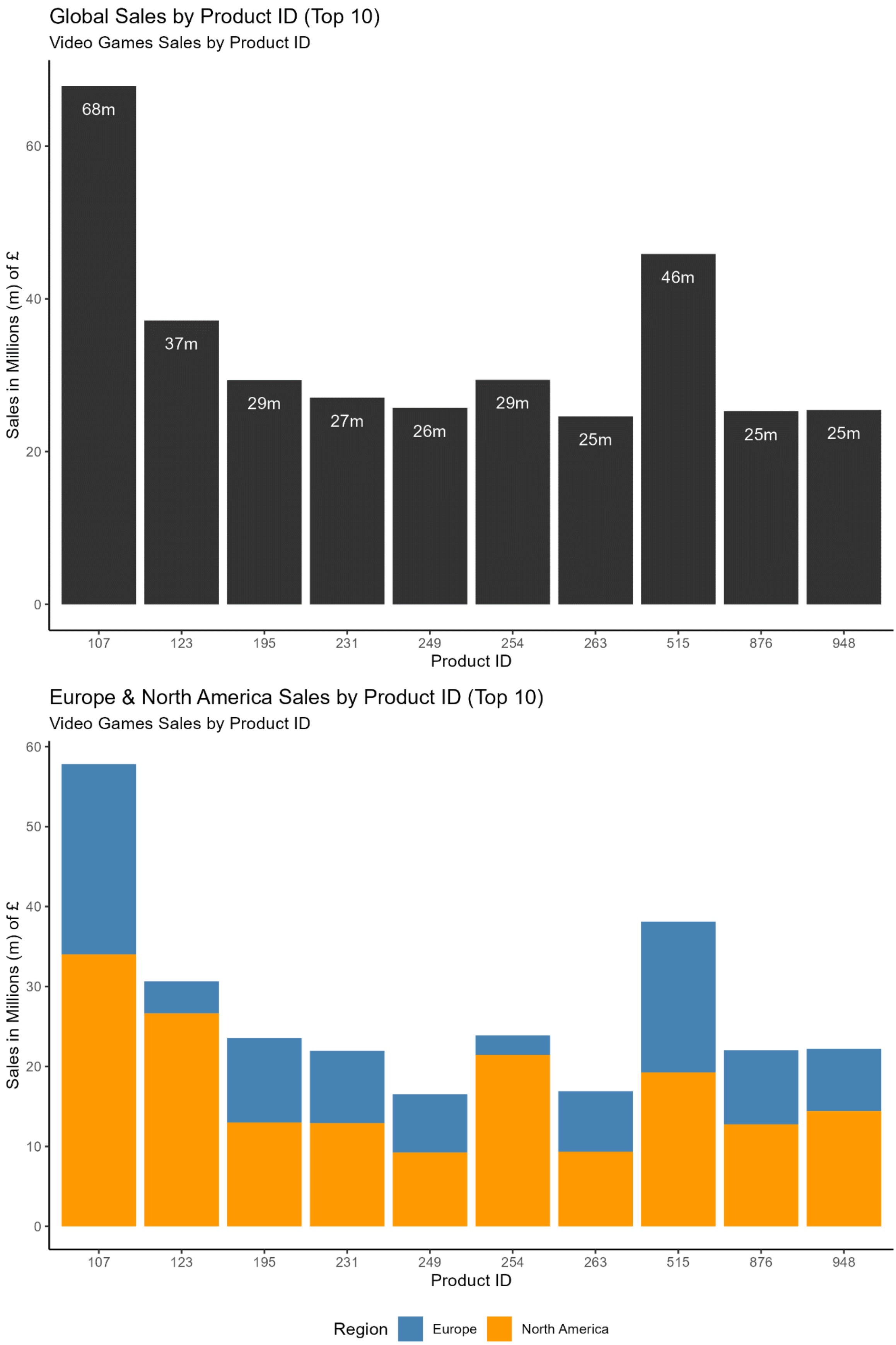
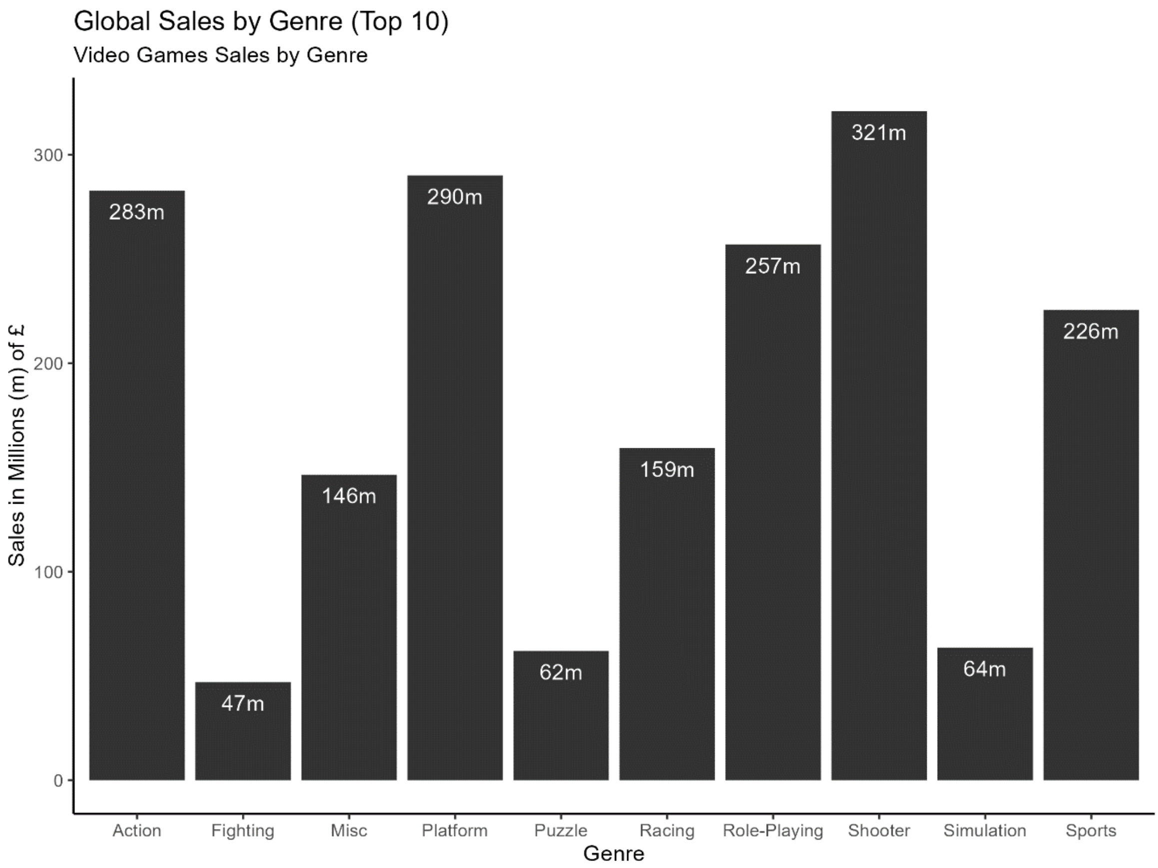


Fig. 16. Total Sales by Genre



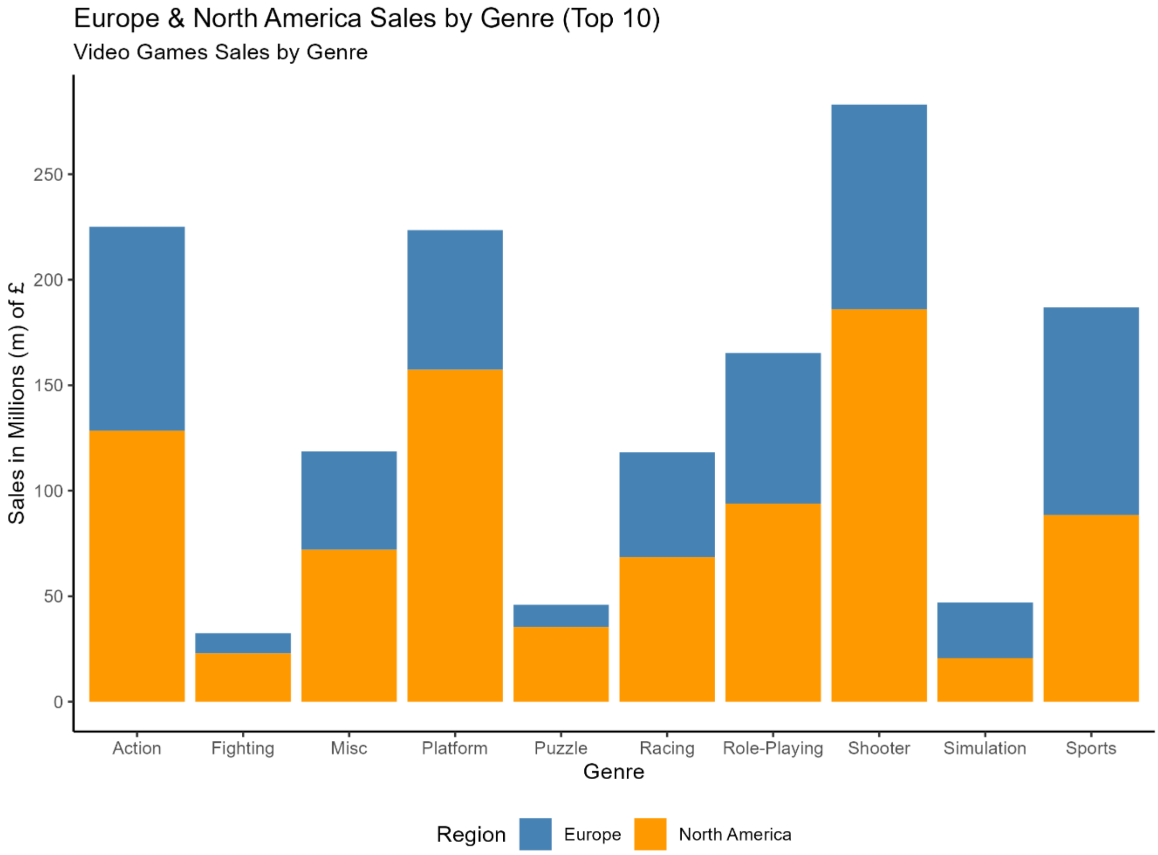
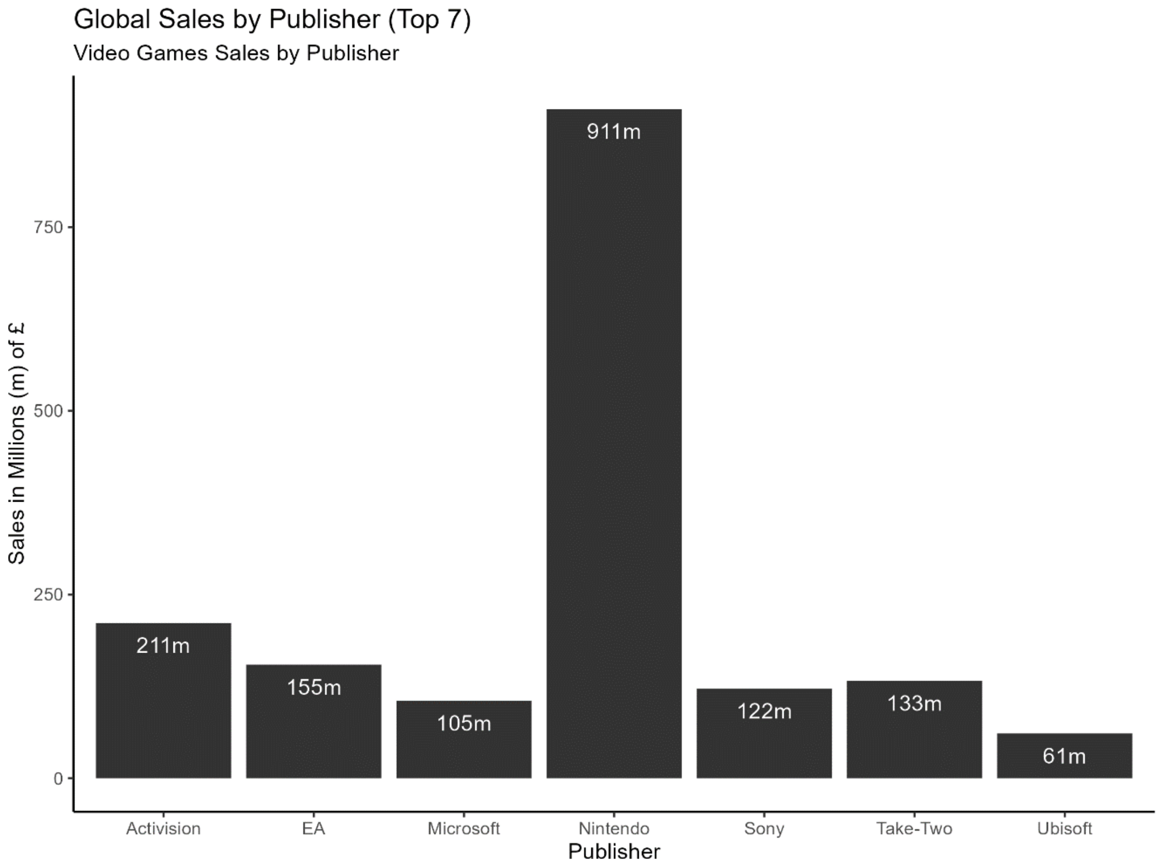



Fig. 17. Total Sales by Publisher



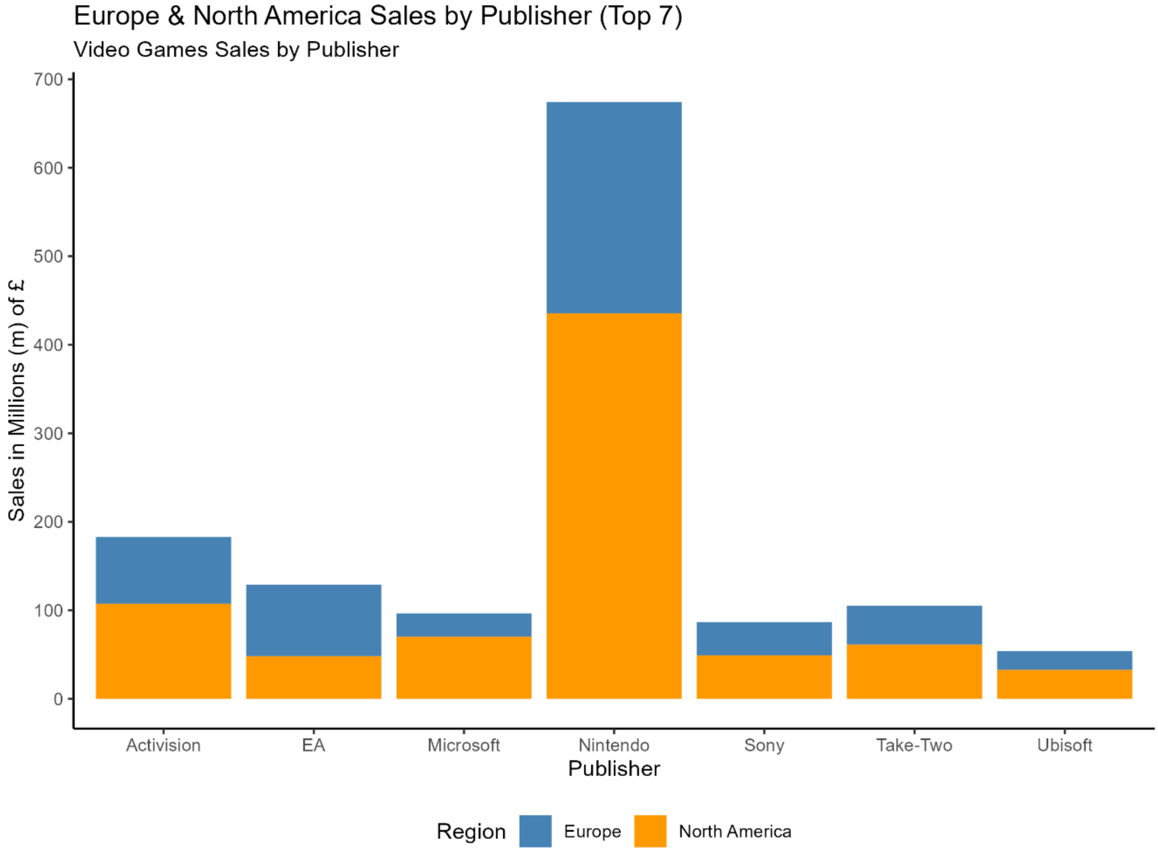
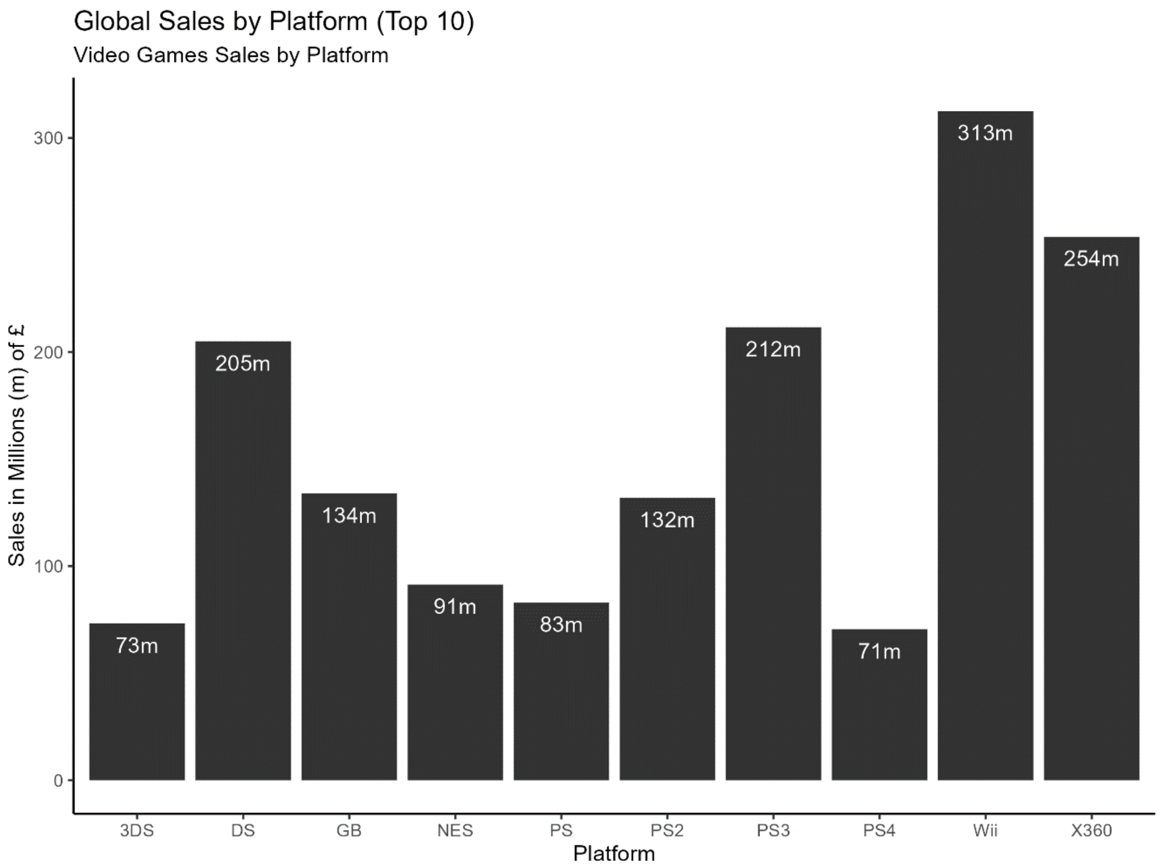



Fig. 18. Total Sales by Platform



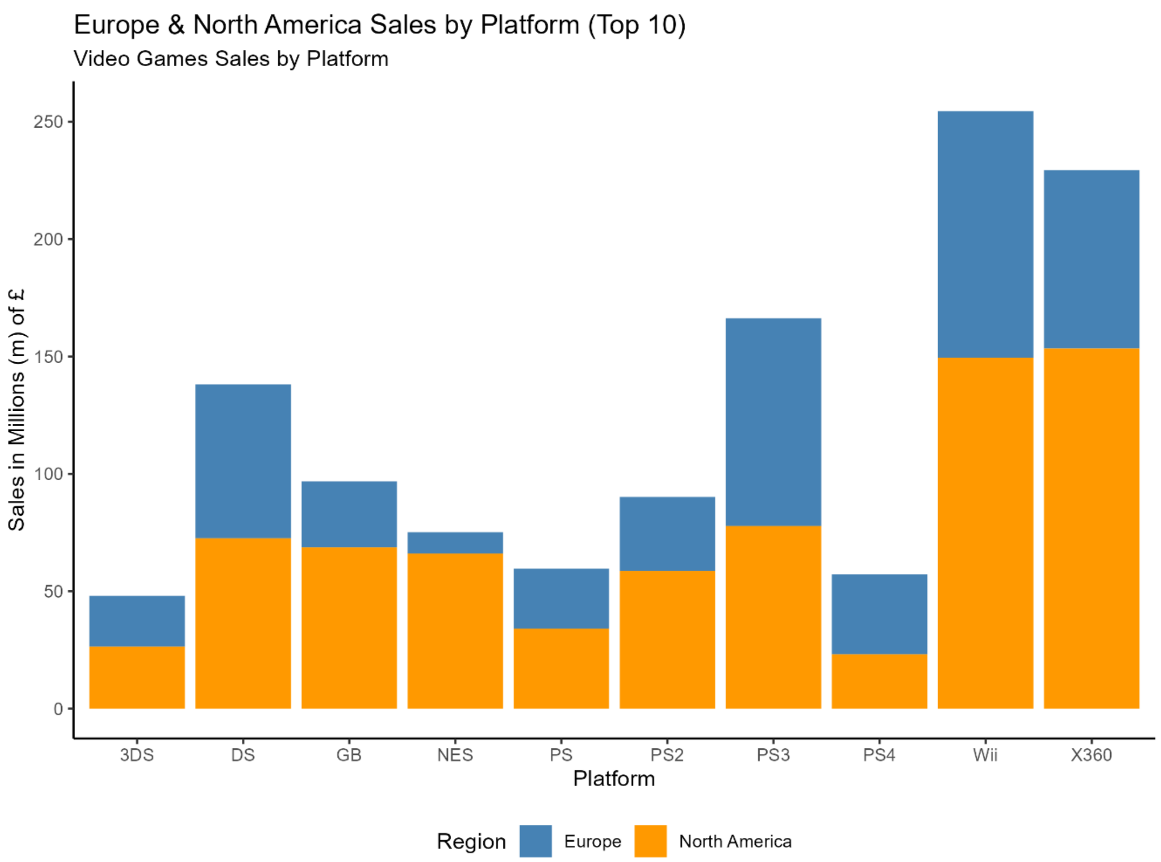



Fig. 19: MLR Model Predictions: Actual Loyalty Points vs. Predicted

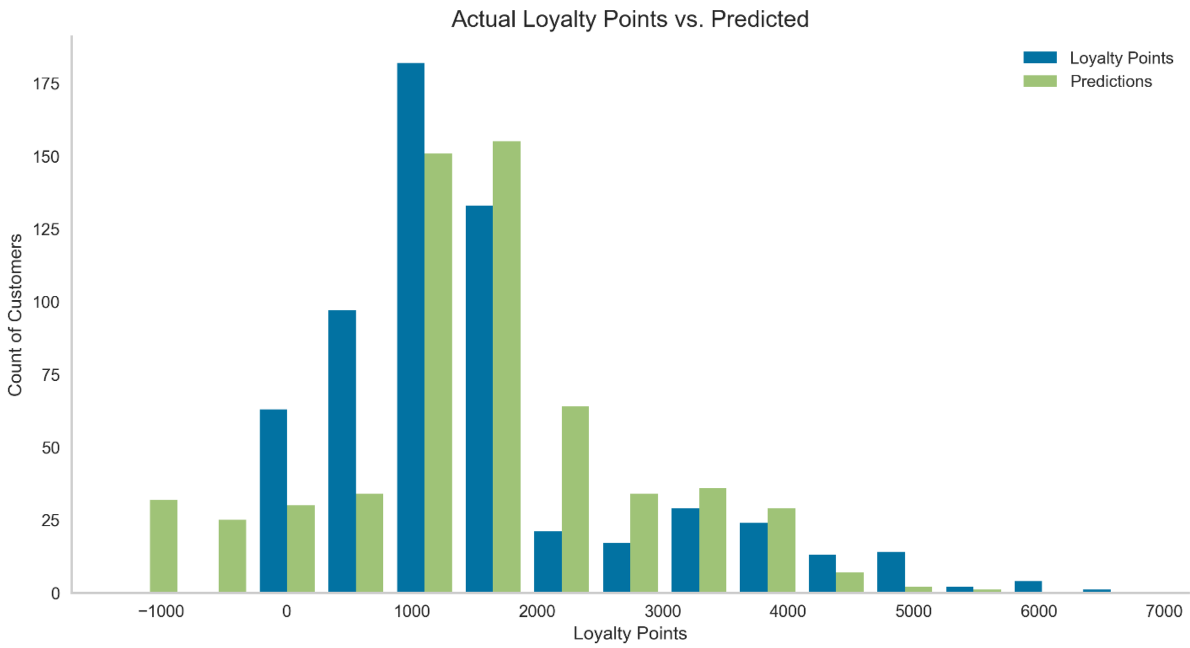



Fig. 20: MLR Model Predictions: Actual Loyalty Points minus Predicted

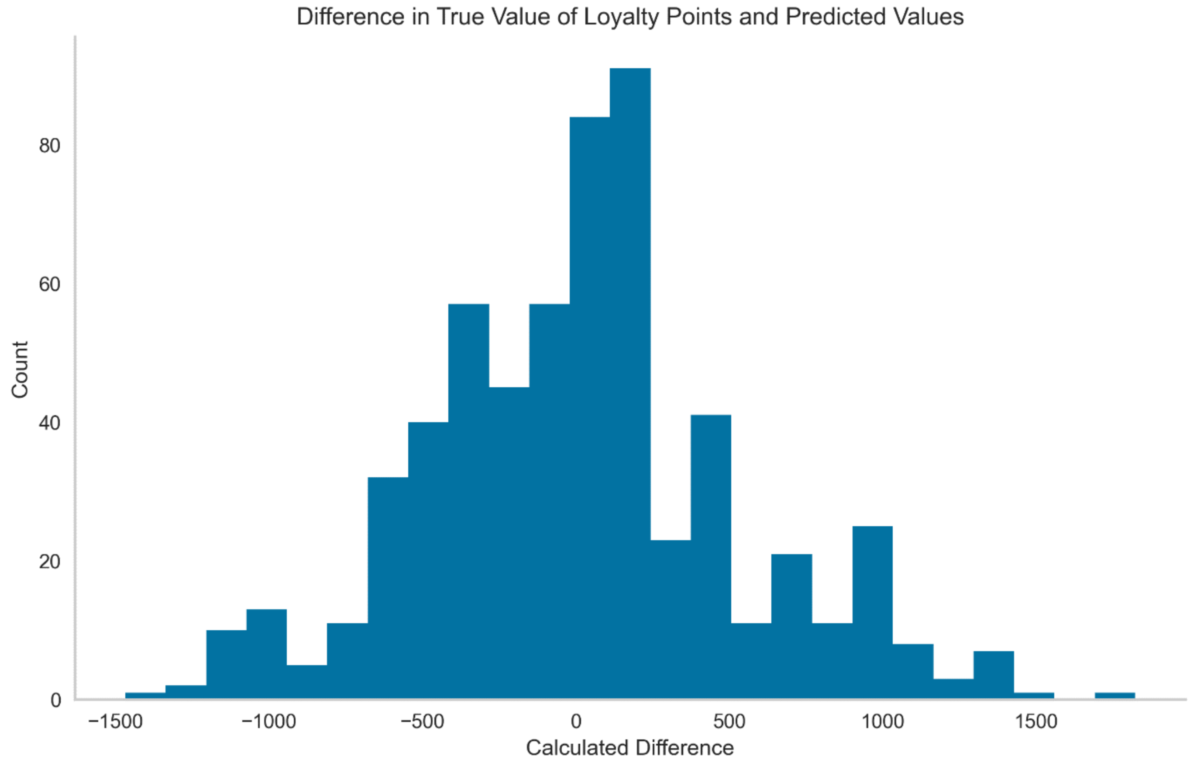


Fig. 21: K-means Classification Model: Remuneration vs. Loyalty Points

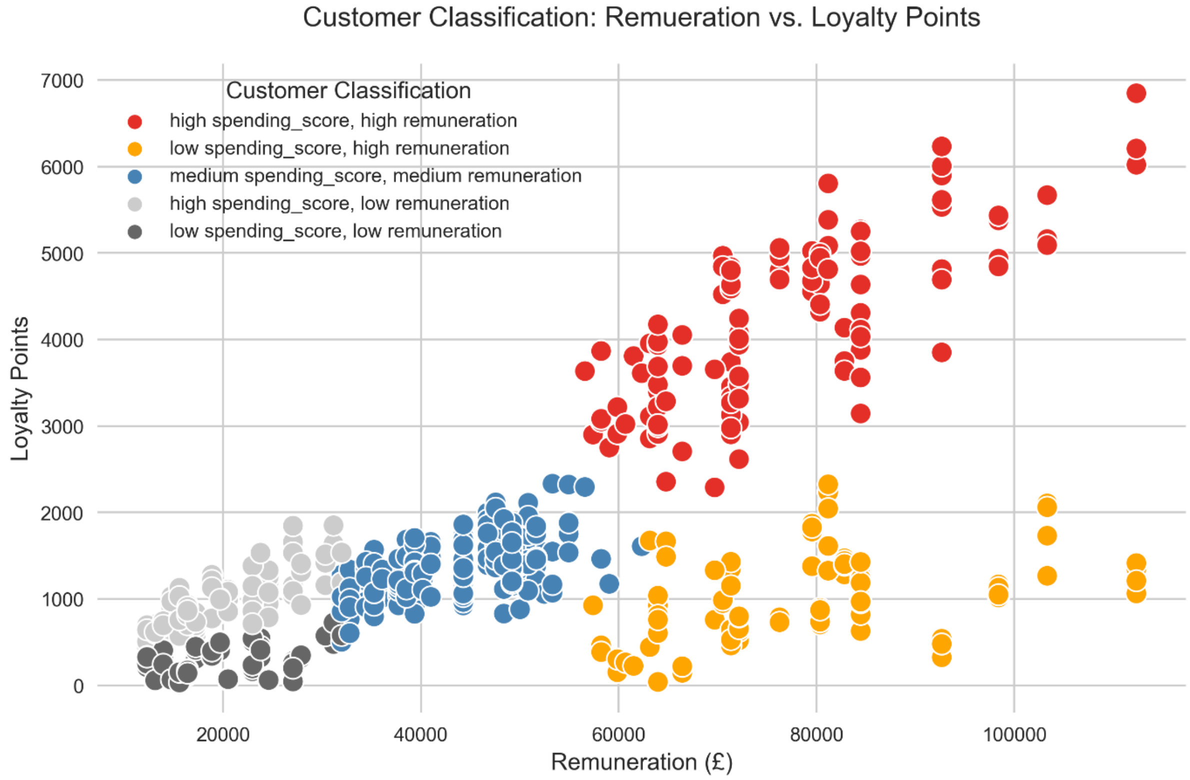



Fig. 22: K-means Classification Model: Spending Score vs. Loyalty Points

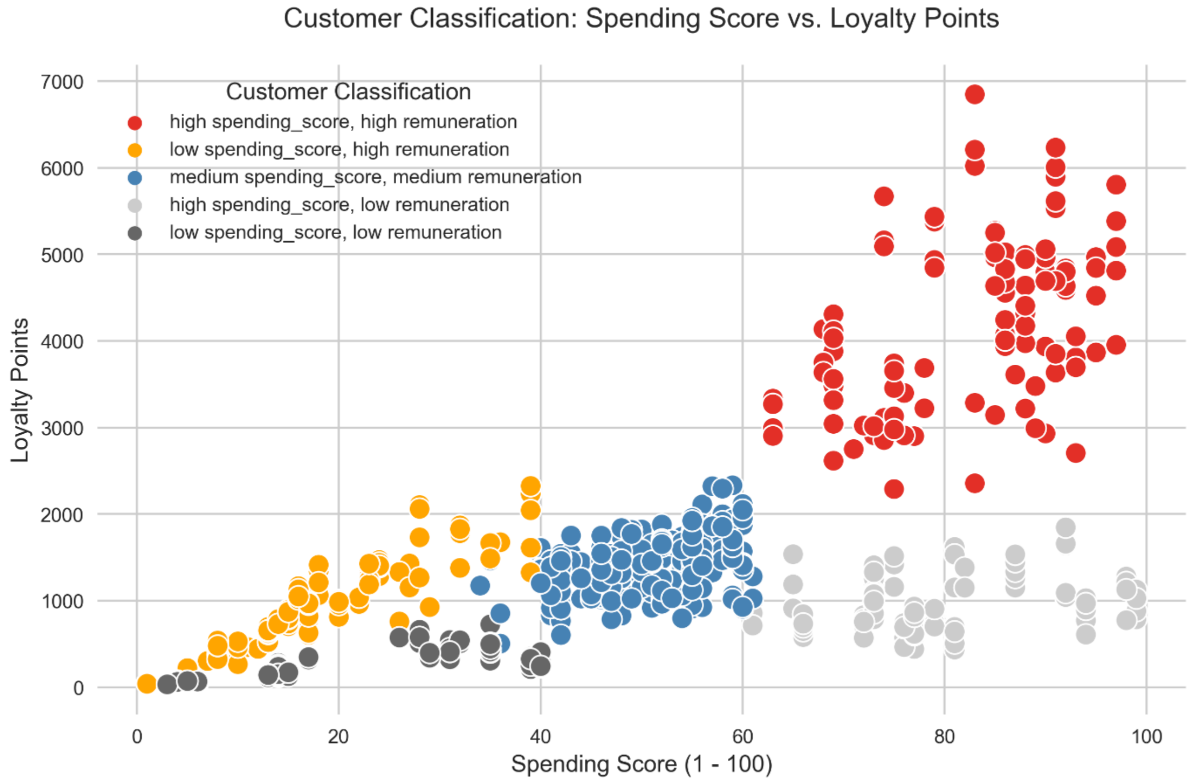


Fig. 23. High Spending Score & High Remuneration: Negative Reviews

| | classification | summary | star_rating |
|---|---|---|---|
| 131 | high spending_score, high remuneration | Not as advertised | 1 |
| 1670 | high spending_score, high remuneration | Not as expected | 1 |
| 1552 | high spending_score, high remuneration | This says it is 5 SHEETS when in fact it ... | 1 |
| 1123 | high spending_score, high remuneration | Not as described | 1 |
| 929 | high spending_score, high remuneration | Wrath of the customer | 1 |
| 784 | low spending_score, high remuneration | One Star | 1 |
| 770 | low spending_score, high remuneration | Bring back the 1999 Version!!!! | 1 |
| 593 | high spending_score, high remuneration | Too hard for children or adults to open | 1 |
| 1694 | high spending_score, high remuneration | No ball to the Ball of Whacks | 1 |
| 359 | high spending_score, high remuneration | Not worth the money! | 1 |
| 337 | high spending_score, high remuneration | Don't try to recreate the photos | 1 |
| 328 | low spending_score, high remuneration | Defective- poor QC | 1 |
| 325 | high spending_score, high remuneration | Cute but not realistic! | 1 |
| 188 | low spending_score, high remuneration | Do not buy! They look cute. But my ... | 1 |
| 183 | high spending_score, high remuneration | Incomplete | 1 |
| 1706 | high spending_score, high remuneration | Fell apart | 1 |
| 145 | high spending_score, high remuneration | Junk | 1 |
| 173 | high spending_score, high remuneration | Horrible! Nothing more to say Would give zero stars ... | 1 |
| 174 | low spending_score, high remuneration | Faulty Product | 1 |
| 182 | low spending_score, high remuneration | INCOMPLETE KIT! | 1 |
| 133 | high spending_score, high remuneration | Smaller than life. | 1 |
| 134 | low spending_score, high remuneration | Perfect for Preschooler | 2 |
| 793 | high spending_score, high remuneration | Disappointed | 2 |
| 799 | high spending_score, high remuneration | I didn't like how the pieces are made of paper | 2 |
| 143 | high spending_score, high remuneration | Doesn't hold up well | 2 |
| 983 | high spending_score, high remuneration | Disk #5 was corrupted; requires too much ram | 2 |
| 996 | low spending_score, high remuneration | We had hours of fun. Took a bit to figure out the rules | 2 |
| 998 | low spending_score, high remuneration | Four Stars | 2 |

Fig. 24. Reviews by Customers Classification and Star Ratings

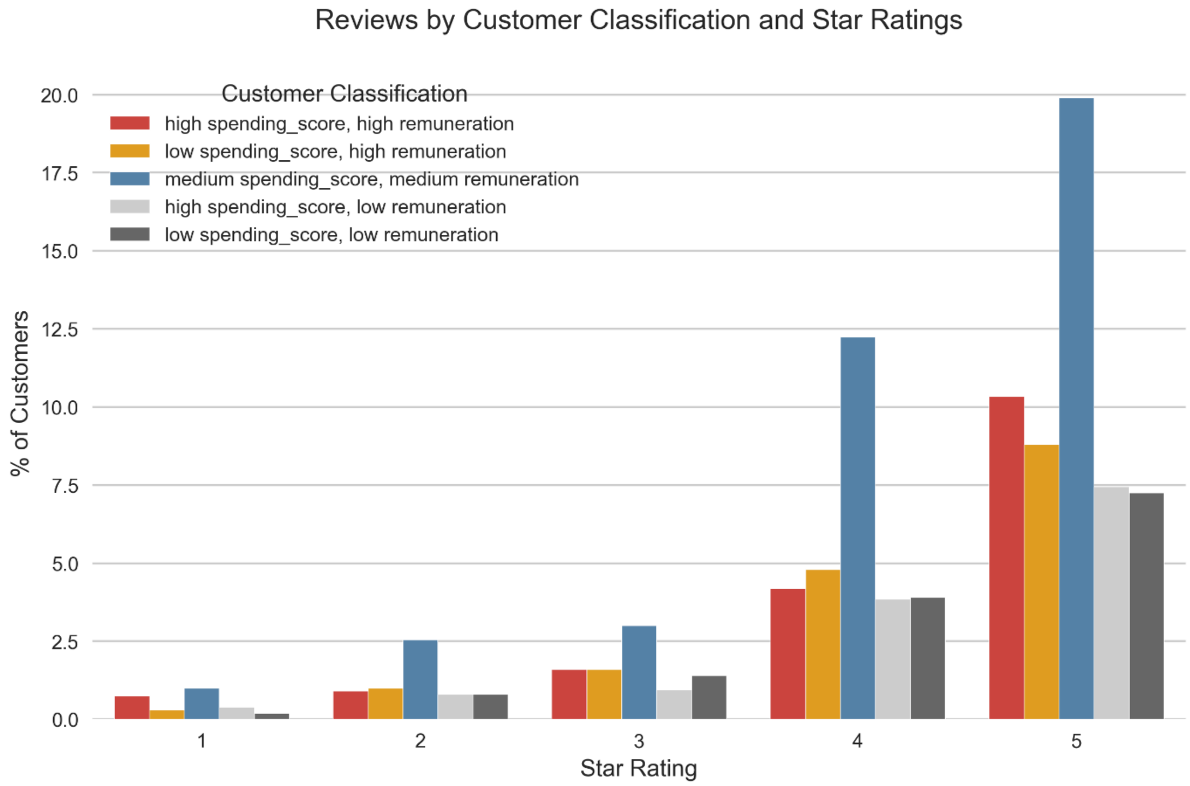


Fig. 25: MLR Model: na_sales and eu_sales vs. global_sales

```
Call:
lm(formula = global_sales ~ na_sales + eu_sales, data = ts_num_grouped)

Coefficients:
(Intercept)     na_sales     eu_sales
      1.042        1.130        1.200


Call:
lm(formula = global_sales ~ na_sales + eu_sales, data = ts_num_grouped)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4156 -1.0112 -0.3344  0.6516  6.6163

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.04242    0.17736   5.877 2.11e-08 ***
na_sales     1.13040    0.03162  35.745  < 2e-16 ***
eu_sales     1.19992    0.04672  25.682  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.49 on 172 degrees of freedom
Multiple R-squared:  0.9668,	Adjusted R-squared:  0.9664
F-statistic:  2504 on 2 and 172 DF,  p-value: < 2.2e-16
```

> The summary indicates:
>
> - Ajusted $R^2$: 96.64% of the total variability of `y` (Global Sales), is explained by the variability of `X` (Europe & North America Sales).
> - `X` : The coefficient of `X` describes the slope of the regression line, in other words, how much the response variable `y` change when `X` changes by 1 point (in this case £1,000,000). In this model, (i) if the North America sales increases by £1M, then the Global sales ( `y` ) will increase by £1,130,400, and (ii) if the Europe sales increases by £1M, then the Global sales ( `y` ) will increase by £1,199,920, and (ii)
> - The `P>[t]` for the `X` coefficient value tests the hypothesis that the slope is significant or not. If the corresponding probability is small (typically smaller than 0.05) the slope is significant. In this case, it has a 3 star rating which means the slope is significant and the null hypothesis can be rejected.

Fig. 26. MLR Model Predicting Global Sales