We analysed Turtle Game's sales and customers' review data to gain insight and improve overall sales performance. Company provided us with the following data:

- Details of customer reviews.
- Details of video games sold globally over time.

Customer Reviews:

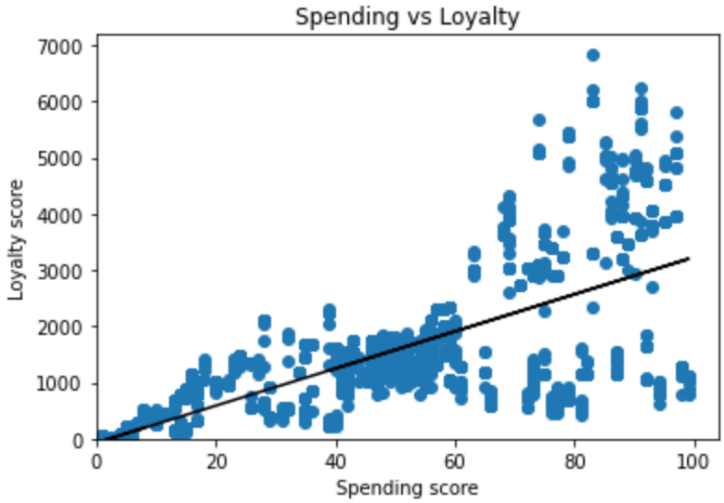
Initial analysis of data:

All the customer reviews data has been analysed for properties and data quality. Data has been reviewed and checked for missing values. Unnecessary columns were removed, and customer loyalty points regressed against three variables: age, renumeration and customer spending.

Regression analysis:

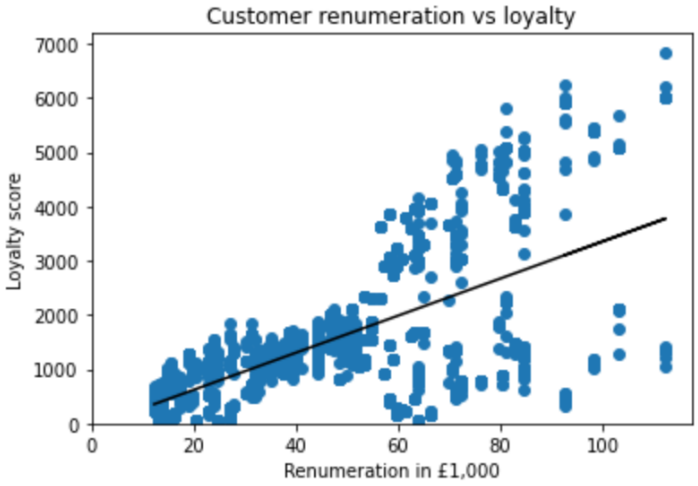
1) Spending vs Loyalty

There is a strong, statistically significant (p value of slope = 0 to 3 decimals points) relationship between the spending and loyalty points as shown on the scatterplot below:



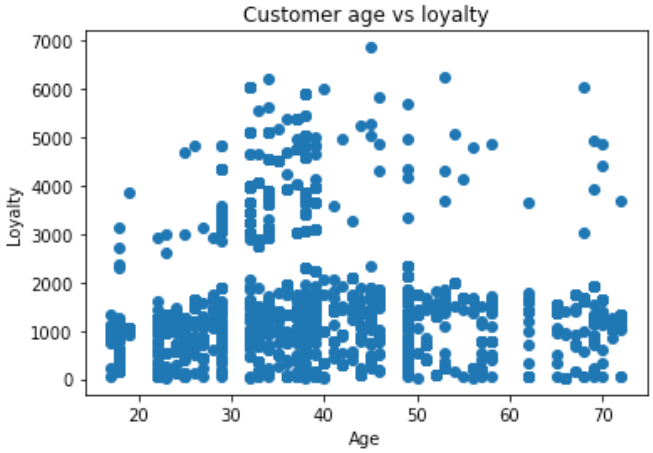
According to metadata, loyalty points are based on customer spendings. Regression seems to confirm that, although it is worth noting that relationship becomes weaker as spending reach high levels. Possible reasons could be due to bonus points or other incentives awarded to high earners.

2) Customer renumeration vs loyalty



Looking at renumeration (total income per customer), I found a strong and significant positive relation (p value of the slope coefficient is zero) between the variable and loyalty points. Given loyalty points and spendings are correlated, and capacity to spend depends on each individual wealth, that makes sense. The more customer earns, the more he is likely to spend. For every 10k of extra income, loyalty points improve on average by 341.

3) Customer age vs loyalty



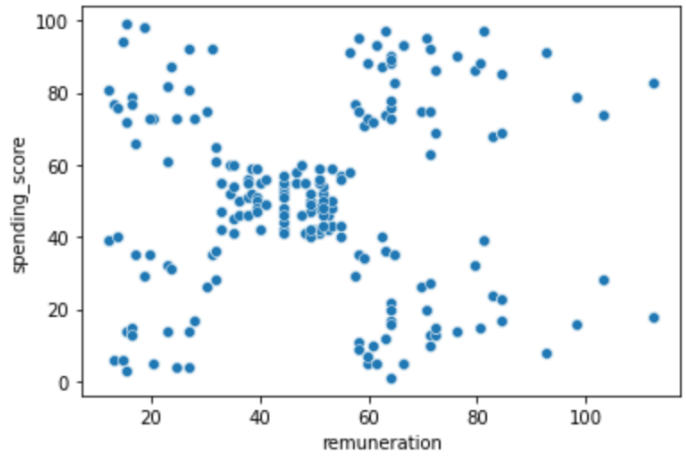
Relationship between age and loyalty is not statistically significant (at 5%). Age does not seem to be a good predictor of loyalty.

Grouping / clustering of customers

1) Overview

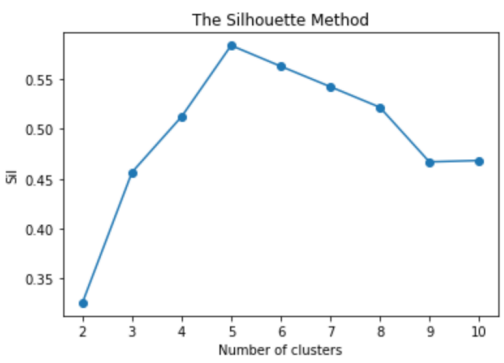
Customer data (Spending Score and Renumeration) was analysed to find commonality in customer groups that can be targeted for advertising. We used k-means clustering algorithm from sklearn library.

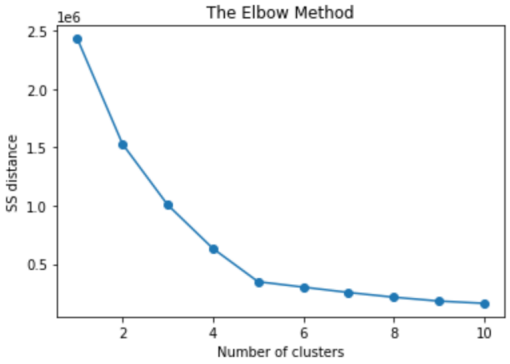
Scatterplot of renumeration and spending score:



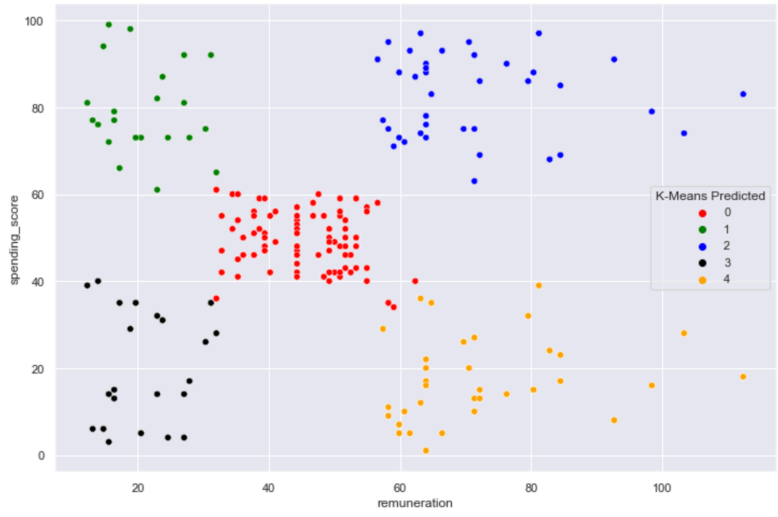
As the k-means clustering model requires k (number of clusters) to be pre-determined, we have decided to use k=5. This is due to:

- Visual inspection of data (data looks clearly to be concentrated around 5 different centres)
- Diagnostics tests. Both Elbow method and Silhouette Method strongly indicated suitability of 5 as the best clustering algorithm:





The results of clustering are shown below:



Customers' reviews and sentiment

As part of the dataset provided, we have received 2000 reviews (in two columns: "review" and "summary") from customers who purchased and used a product from Turtle Games. The review data was messy, so the following data cleaning processes have been implemented.

- Moving all words to lower cases

```
apply(lambda x: " ".join(x.lower() for x in x.split()))
```

- Removing punctuations

```
df3['review'] = df3['review'].str.replace('[^\w\s]','')
```

- Dropping duplicated reviews (through analysing both 'review' and 'summary' columns and double checking that removal algorithm had intended consequences)

After data cleansing, the reviews have been tokenised to create a first set of word clouds. This has been found to be of little value, due to presence of stopwords ('and', 'to', 'for' etc).

We have therefore removed the stopwords and reran the wordcloud with following results:

1) 'Review' column

Word	game	great	fun	one	play	like	love	really	get	cards	tiles	good	time	would	book
Frequency	1679	586	553	530	502	414	326	319	319	301	297	292	291	280	273

Looking at results, sentiment across "review" column is not very clear. There are some positive words sticking out (like "great"), however many of them are of neutral or unknown polarity ("game" or "play")

2) ‘Summary’ column

Word	stars	five	game	great	fun	love	good	four	like	expansion	kids	cute	book	one	awesome
Frequency	448	363	319	295	218	93	92	58	54	52	50	45	43	38	36

“Summary” column looks more positive with words/ phrases like “five stars”, “great”, “fun” dominating the picture. We notice some duplications in the wordcloud (in words like “stars”, “game”, “five” etc). This is due to the WordCloud algorithm settings. As the default, WordCloud uses collocation, which set frequently used two words (bigrams) as one word. We can turn off that option (by setting collocation algorithm to false as per below)

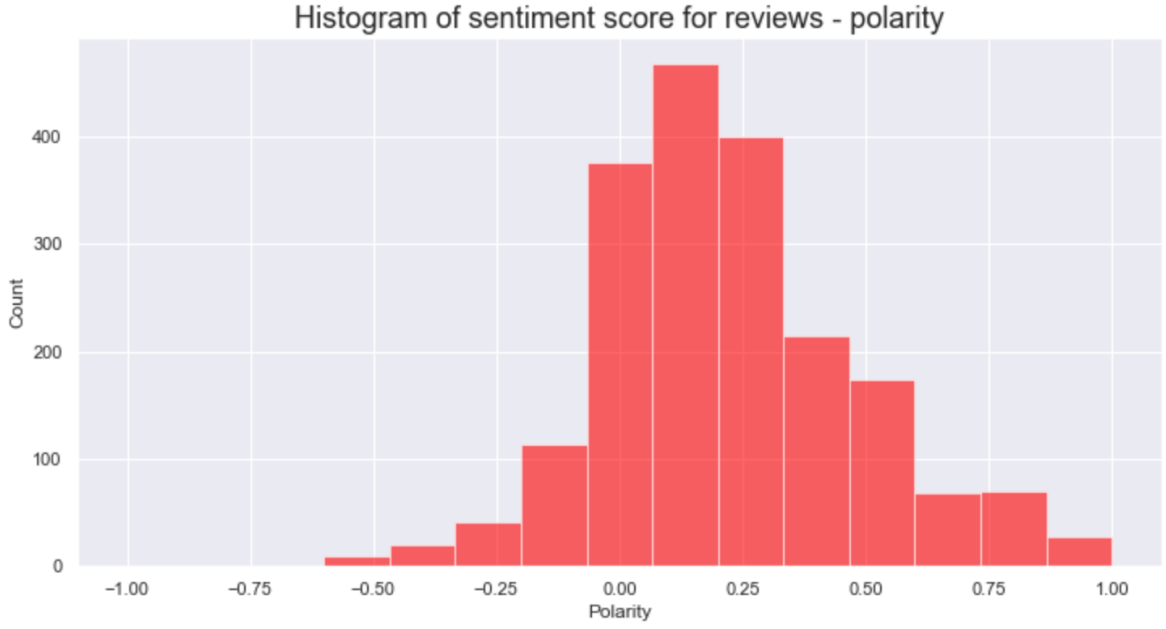
```
word_cloud5 = WordCloud(width = 1600, height = 900,
                background_color ='white',
                colormap = 'plasma',
                stopwords = 'none',
                collocations=False,
                min_font_size = 10).generate(all_summary2)
```

This gives following (visually clearer) results:

Our visual analysis has so far indicated that the sentiment of customers to Turtle Games is generally positive. Let's confirm our results with formal sentiment analysis.

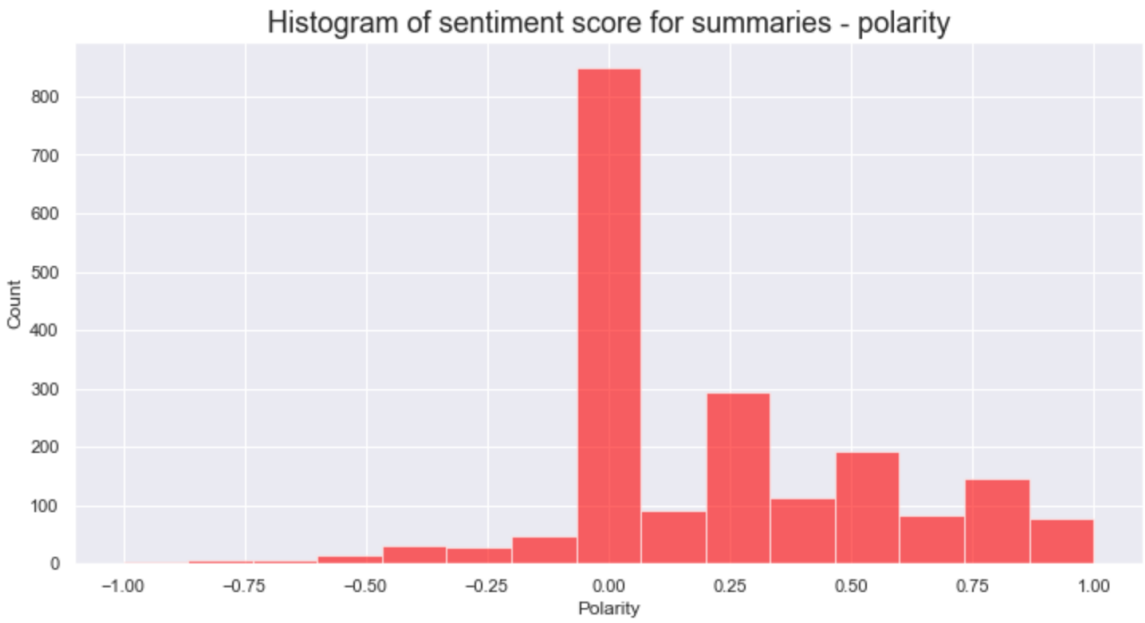
To perform the analysis, we have used sentiment analysis from NLTK library which classifies each statement by polarity from -1 to 1 (where zero is neutral, +1 max positive sentiment, -1 max negative sentiment)

1) "Review" column



Review column seems to have rather neutral to slightly positive statements, with distribution skewed to the right (positive) and mean/ median around 0.2.

2) “Summary” column



Histogram of polarity scores shows unexpected results. Although comments are generally positive (distribution skewed to the right with very few negative comments), it is surprising to see so many neutral comments. After investigation this was found to be due to certain positive statements (like “five stars”, “four stars”) being interpreted as neutral.

See below for example of polarity score for frequently used phrases: “Five stars”, “Four stars”

```
# Check polarity score for statement 'five stars'
print(TextBlob('Five stars').sentiment)

# Check polarity score for statement 'four stars'
print(TextBlob('Four stars').sentiment)
```

```
Sentiment(polarity=0.0, subjectivity=0.0)
Sentiment(polarity=0.0, subjectivity=0.0)
```

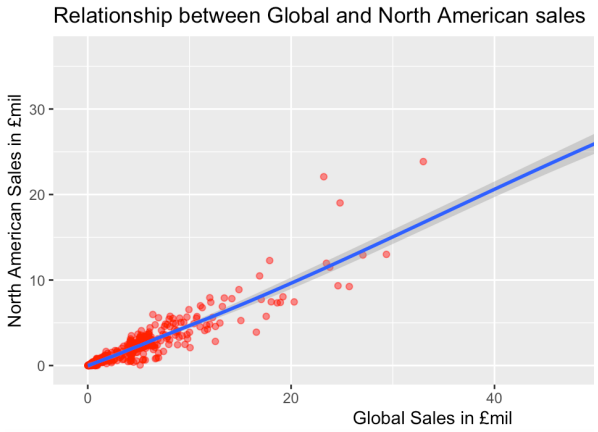
Notwithstanding the above anomaly, comments are positive with mean of 0.22.

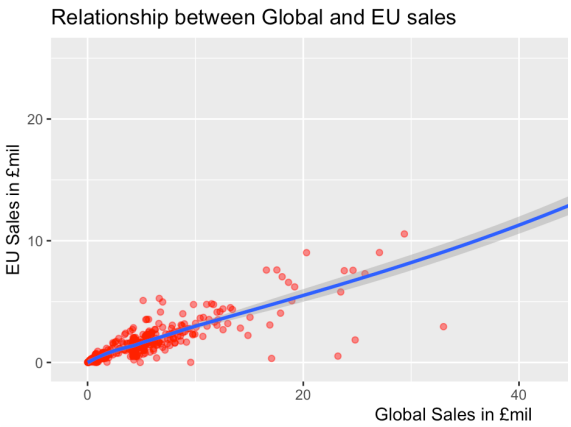
Customer Sales:

Subsequent analysis has been concentrated on the data showing video games sold globally. The analysis has been performed using R.

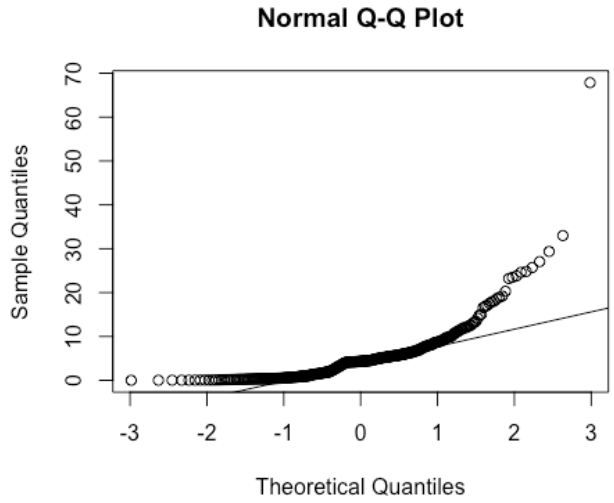
Current data and trends:

- The data shows strong correlation between sales in North America, Europe and Globally as shown on the below graphs.





The sales data is not normally distributed (see Q-Q plot below and Shapiro test). On the example of global sales, the data shows right skewness (skewness=4) and excess kurtosis (kurtosis=32.6).



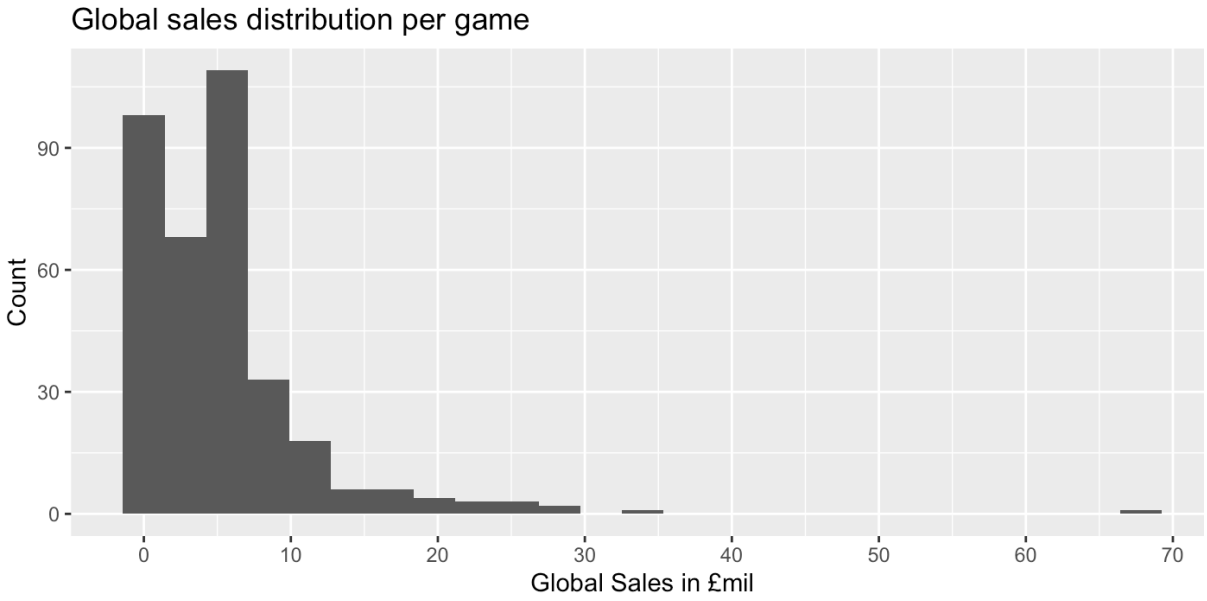
Shapiro test confirms that the data is not normally distributed (p value very small so we can reject the null hypothesis of normality at 1% or even 0.1% significance level)

```
        Shapiro-Wilk normality test

data:  sales_clean$Global_Sales
W = 0.6818, p-value < 2.2e-16
```

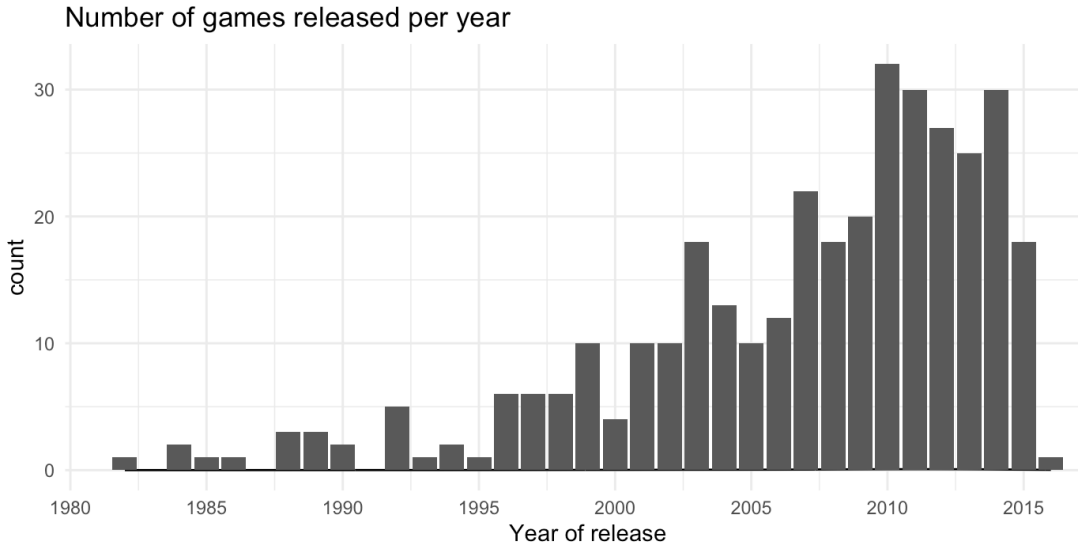
There is strong correlation between regions' sales data as shown in the table below. As the data is not normally distributed, we used Spearman correlation.

	NA_Sales	EU_Sales	Global_Sales
NA_Sales	1.0000000	0.7329771	0.9225854
EU_Sales	0.7329771	1.0000000	0.8661192
Global_Sales	0.9225854	0.8661192	1.0000000

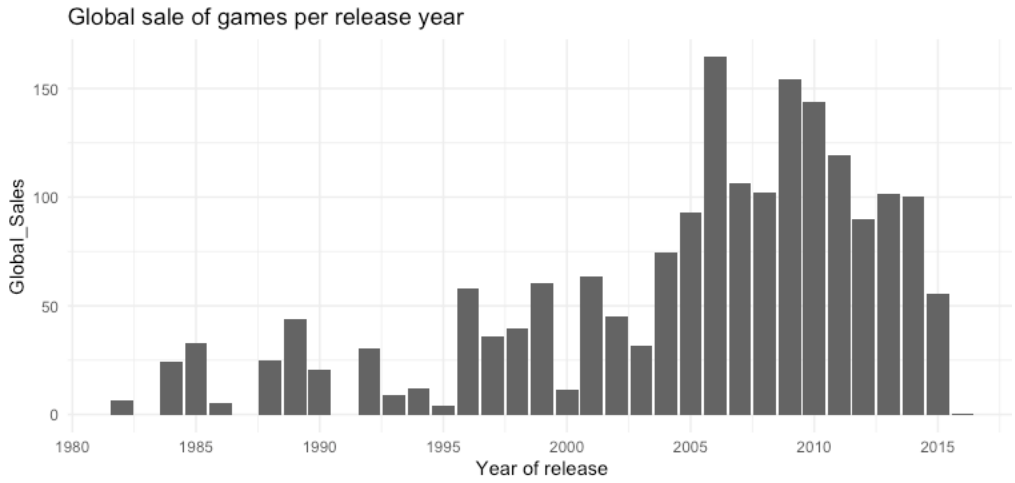
- The sales are concentrated among lower sales titles with couple of game hits as outliers.



- Number of games released per year is increasing.

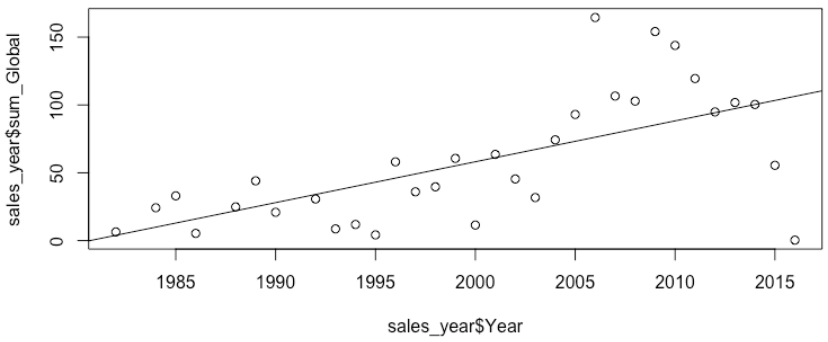


- Global sale of games is linearly dependent with year of release (newer games sell better)



Despite recent dip (after 2010), the linear relation between release year and global sales is strongly significant (at 0.1% or better) as shown on the below regression:

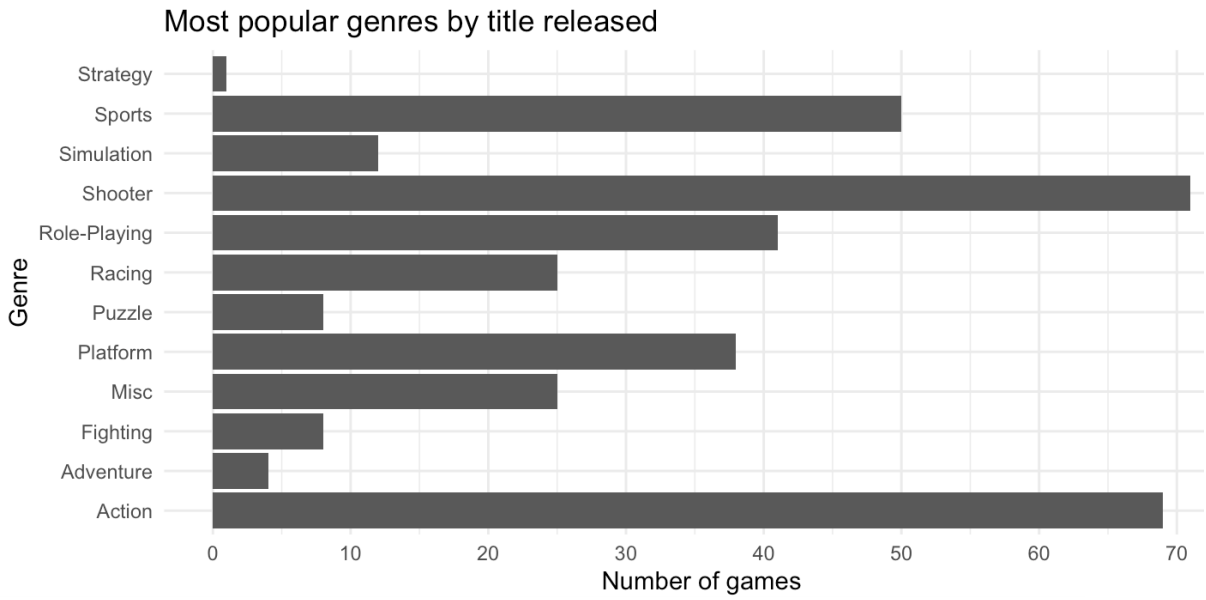
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5967.2420  1320.7367  -4.518 9.05e-05 ***
Year            3.0127     0.6603   4.562 7.99e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



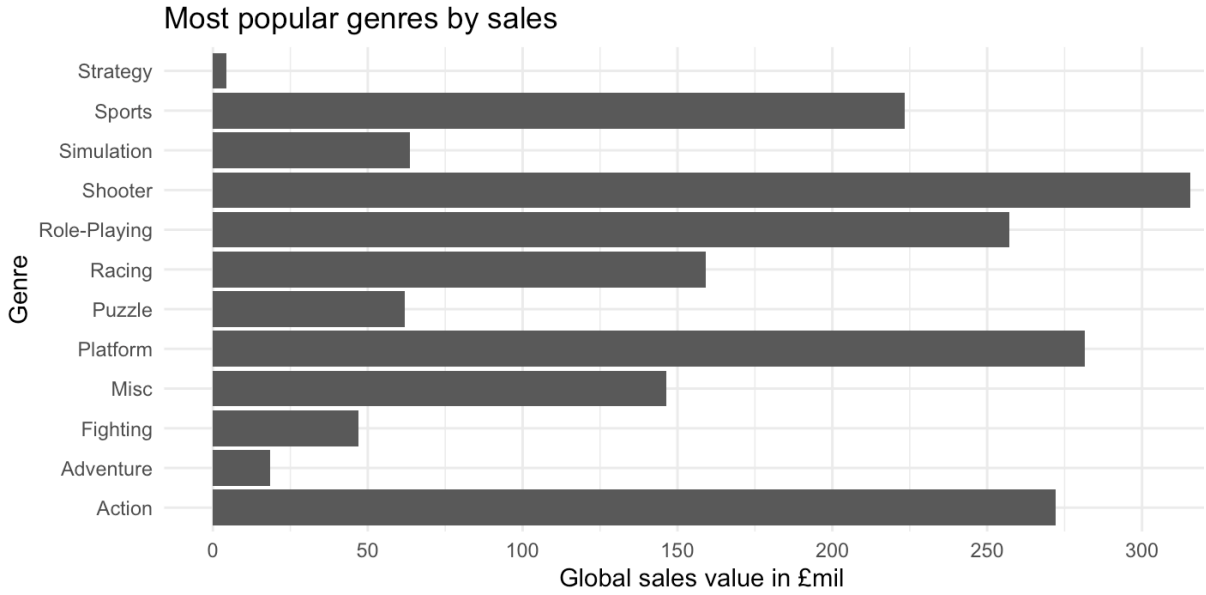
Games released in the future are predicted to have even higher sales, as the global market for games seems to be increasing.

	Year	Global_sales
1	2017	109.3447
2	2018	112.3574
3	2019	115.3701
4	2020	118.3827

- Action and shooter games seems to dominate in terms of number of titles released:

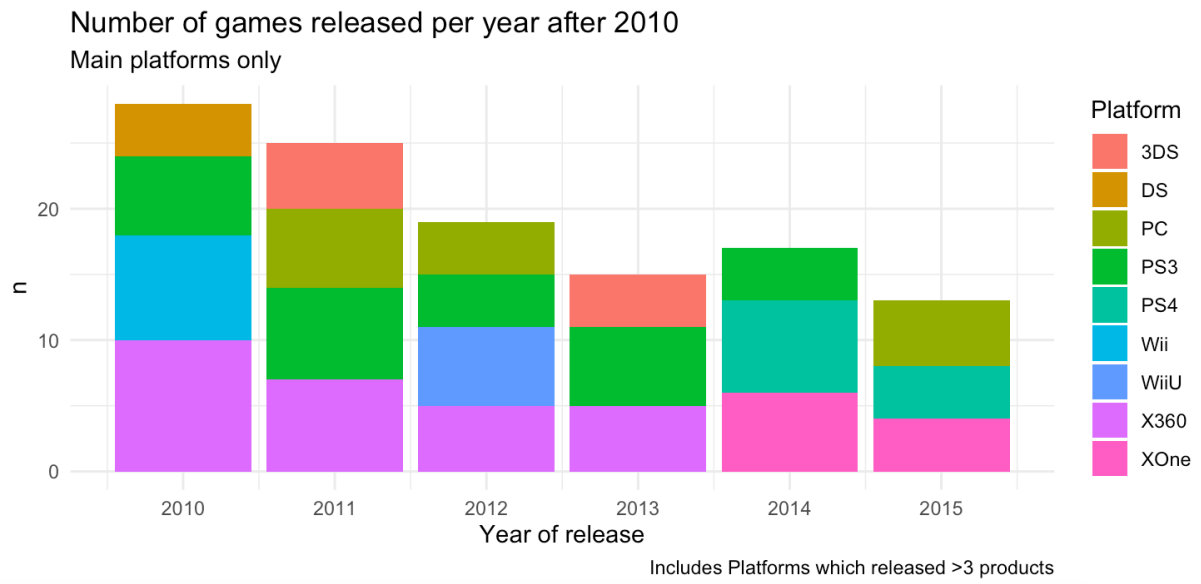


However, they don't dominate to the same extent in terms of global sales:



This might indicate that genres other than 'Shooter' or 'Action' have a higher average global sales per title. Some of the 'Shooter' or 'Action' might be just flops with low sales revenue. Further investigation of distribution of sales for each genre would be necessary to confirm.

- PC, PS4 and XOne platforms seem to dominate latest game releases, with share of other platforms decreasing since 2010

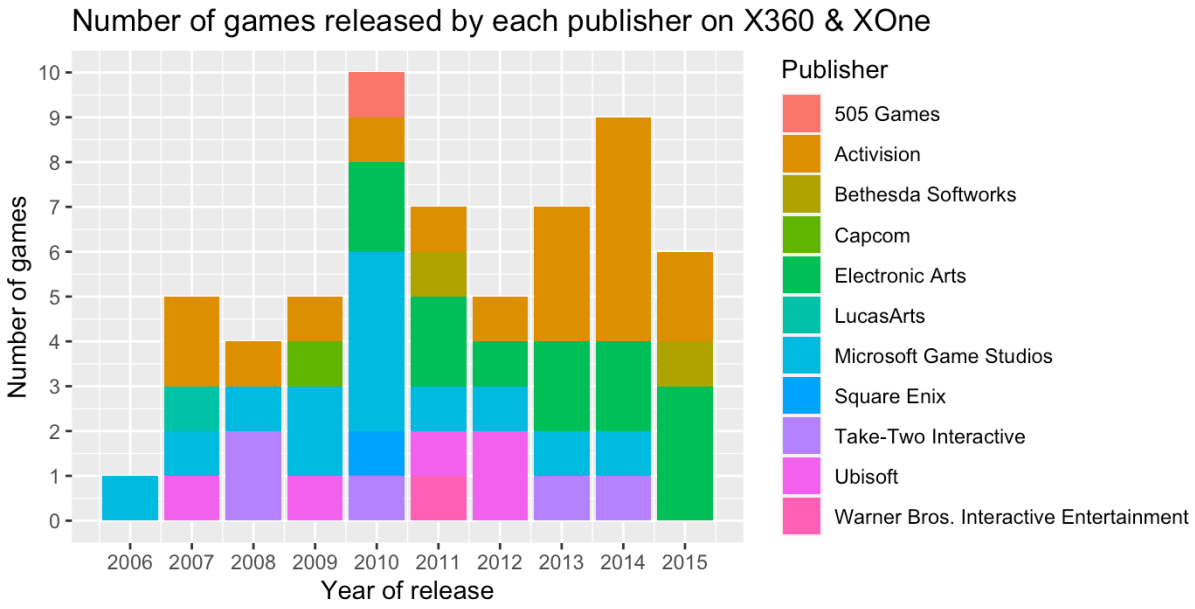


The above graph required some extra preparation in R, to filter data appropriately (for year >= 2010 and filter for platforms with only more than 3 games released – for graph clarity). The following code snippet was used to first count the titles per Platform & Year Released (using tally function) and then subsequently filter the data and pass to ggplot.

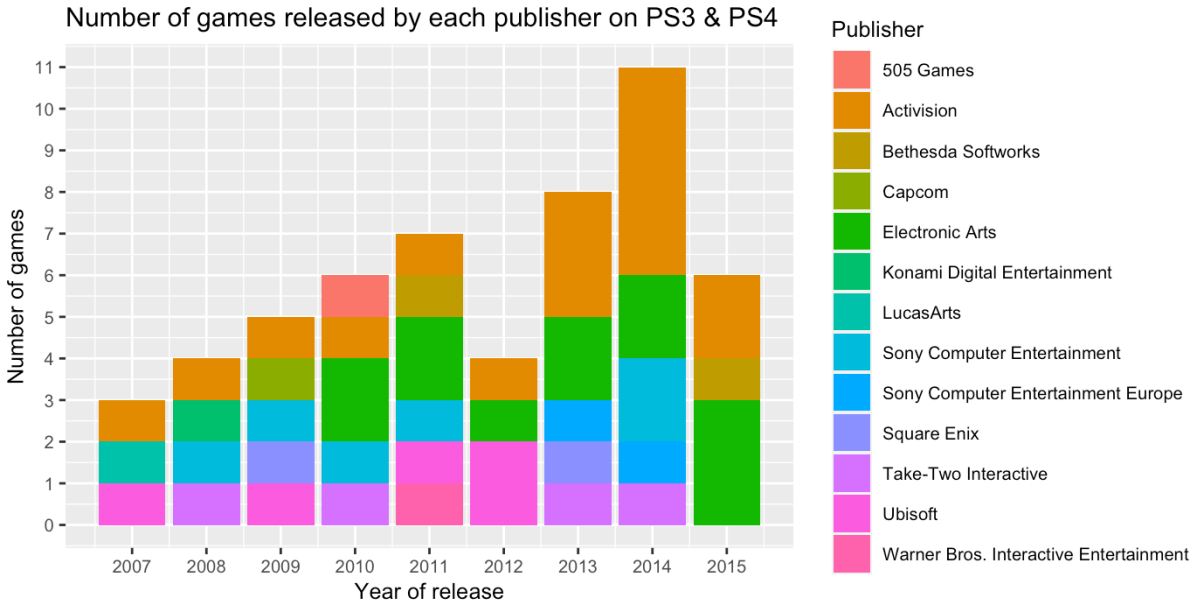
```
# Create new dataframe grouping by Platform and Year and counting
sales_tally <- sales %>% group_by(Platform, Year) %>% tally(sort=TRUE)

# Pass the new dataframe and create graph for year above 2010 and n>3
sales_tally[sales_tally$Year>=2010 & sales_tally$n>3,] %>%
  ggplot(aes(x = Year, y=n, fill= Platform)) +
  geom_col()+
  scale_x_continuous(breaks = seq(2010, 2016, 1), "Year of release")+
  theme_minimal()+
  labs(title="Number of games released per year after 2010",
       subtitle="Main platforms only",
       caption="Includes Platforms which released >3 products")
```

- No single Publisher dominates the current market in the highest selling platforms (PS4/PS3 and XOne/X360). The market is fragmented.

    - X360 & XOne



    - PS3/PS4



The following code was used to create the above graphs (on example of PS3/PS4)- please note work related to adjusting the font size to prevent overlapping of legend and the tile:

```r
# Number of distributors releasing on PS3 & PS4

# Filter dataframe only on platform PS3 & PS4
sales[sales$Platform %in% c("PS3","PS4"),] %>%
  # Pass it to the graph
  ggplot(aes(x = Year, fill= Publisher)) +
  geom_bar()+
  # Define the axis breaks
  scale_x_continuous(breaks = seq(2005, 2015, 1), "Year of release")+
  scale_y_continuous(breaks = seq(0, 12, 1), "Number of games")+
  labs(title='Number of games released by each publisher on PS3 & PS4')+
  theme(text = element_text(size = 9.5))
```