Advanced Analytics Assignment

(Turtle Games using Python /R)

Eunmi K Rhee

July 2022

1. Background of analyzing Turtle games

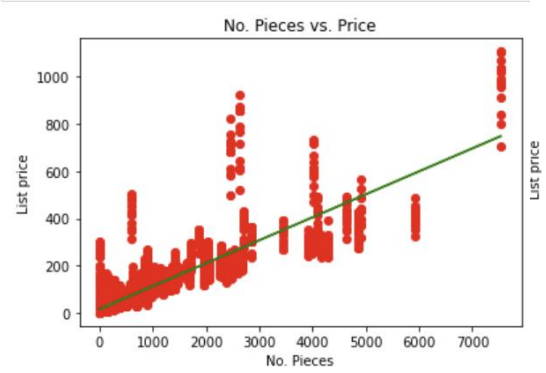
- Turtle games is a game manufacturer and retailer. Their products include Lego, board games, video games and toys. They have a global customer base and wish to improve the overall sales performance.
- The objective of the analysis is to determine:
  - the optimal price based on the number of Lego pieces and customer age
  - customer groups that most likely leave a review of the products they have purchased
  - most popular and most expensive products purchased by a particular group of customers
  - general sentiment of customers across various products
  - sales they can expect to achieve for particular products they offer

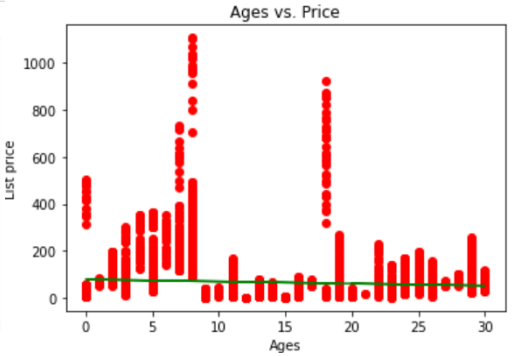
[#Q1] Lego sets:

- The first question is to determine the optimal price of certain number of pieces of Lego and customer group including:
  - What price should be set for the 8000 Lego pieces
  - What price should be set for the Lego sets that have 8000 Lego pieces and are most likely to be purchased by customers who are 30 years old

1) Analysis (using Python)

- The data (lego.csv file) has 7 columns and 12,261 records
- No missing values
- Run regression model using number of pieces as x variable and list price as y.
  - Using number of pieces and price, the R-squared is 75.6%. For 8000 pieces, the predicted price is $792.6.
- Running another regression using ages as x variable and list price as y.
  - The R score is less than 1% (0.6%) which indicates ages is a poor predictor of price
- Running a multivariate regression model with both ages and pieces as independent variables
  - The multivariate model moves the R squared barely as expected, given the low correlation with the age variable and R squared remains largely the same level (75.6%).



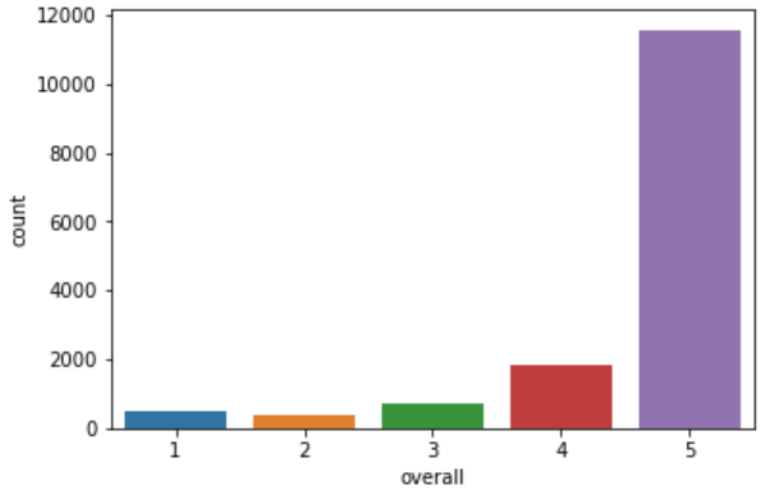


2) Comments/ Highlights

- For 8000 pieces, the predicted price is $792.6. The predicted price of $793 is primarily the same for customers who are 30 years old as well in purchasing the 8000 pieces.

[#Q2] Analyzing Customer Sentiment

- The second question is to determine the general sentiment of customers across all products.
- Based on polarity of the sentiment, identify the top 20 positive and top 20 negative reviews



1) Analysis (Using Python)

- The data (game_reviews.csv file) has 9 columns and 15,000 records
- No missing values
- Using the overall scores 89% of the respondents gave 4 or 5s and if we add '3's, 94% gave positive responses.
- To analyze the review text itself, natural language processing was applied where:
  - Data was pre-processed with lower casing and removing punctuations
  - Removal of stop words
  - Applied lemmatization where in natural language processing involves working with words according to their root lexical components.
  - Using Sentiment Intensity Analyzer, data is classified to negative, neutral, and positive reviews.
  - Using a threshold of 0, converted the scores into positive and negative reviews where 14,402 (96%) reviews were positive & 598 (4%) were negative

2) Comments

- Using the **overall score data**, 89% of the respondents gave 4 or 5s and if we add '3's, **94% gave positive responses**.
- Applying **NLP**(Natural Language Processing), **96% of the respondents gave positive reviews** and 4% negative reviews, which is largely in line with the overall comments.
- Top 20 positive/ negative comments include:

[Top 20 positive]

5659	Awesome
11500	Excellent way to start conversations
5433	Awesome
9272	Just as picture. Perfect gift !
13623	It was perfect since she picked it out herself...
9269	Awesome!!!
5420	excellent trade
13607	Awesome!!!
3222	Perfect give
9514	Perfect condition when it arrived!
13605	Came in excellent shape. Haven't opened-as is ...
8902	Awesome
14734	Perfect to go with book.
524	Perfect, just what I ordered!!
8928	Had the best season with my elf.
9505	He is perfect we can't wait to get started.
14761	it was perfect
14763	The Best!!!
496	Excellent activity for teaching self-managemen...
8961	Perfect

[Top 20 negative]

3359	some of the suggestions are disgusting
2043	Kids did not like it. Thought it was boring.
8319	Awful. We did not receive what was advertised....
208	BOOO UNLES YOU ARE PATIENT KNOW HOW TO MEASUR...
8758	I hate the holidays bcuz of the Elf, he was di...
13181	I do not under stand how you keep score or rea...
9143	Cliche and stupid. I should not drink and Amaz...
9260	Just stupid.
8398	I only recieved the book from this place when ...
7988	Was the elf on the shelf but it didn't have th...
12510	These stickers were not the same as the ones I...
9511	I haven't even taken it out of the box yet but...
526	Keeps clients engaged while helping them devel...
14285	I like this product for my daughter. She is in...
8860	She arrived today. I'm so disappointed in her ...
9058	This toy is designed to desensitize youngsters...
11971	Horrible and incomplete flash cards....DO NOT ...
2147	This was a bit disappointing. My students find...
174	I sent this product to my granddaughter. The p...
13982	had no idea the extent you have to go through ...

[#Q3] Visualize data to analyze customers

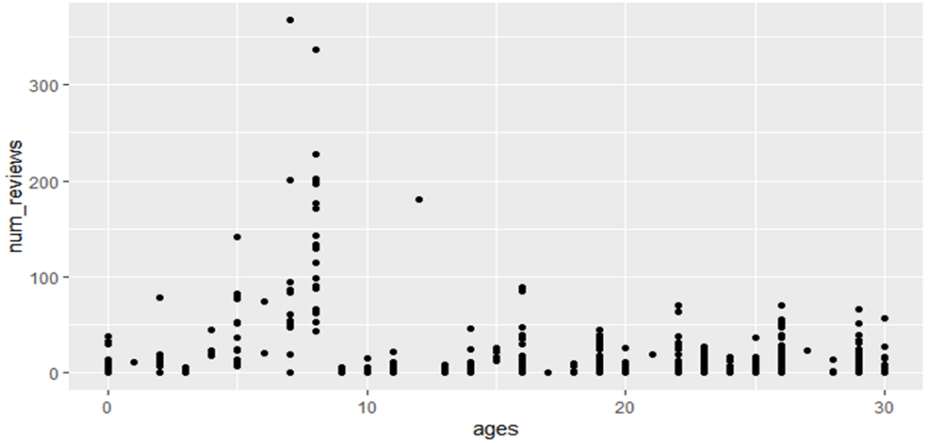
- The third question to address is to determine the customer group most likely to leave a review on products
- Identify the most popular, expensive product purchased by a particular group of customers

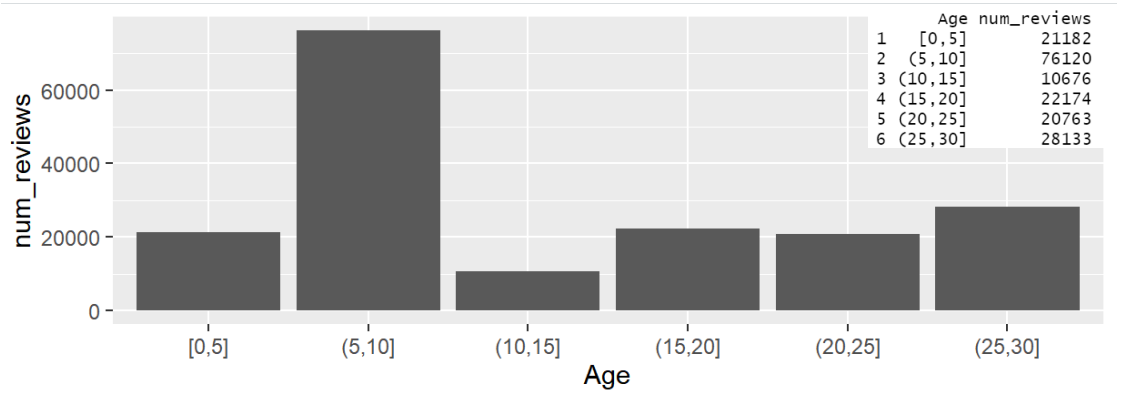
1) Analysis (Using R)

- The data (lego.csv file) has 7 columns and 12,261 records
- Using the summary function in R:
  - Ages range from 0~30
  - List price from $2.3~$1104
  - Number of pieces from 1~7541
  - Rating and difficulty from 0~5

```
      ages          list_price        num_reviews
 Min.   : 0.00   Min.   :   2.272   Min.   :  0.0
 1st Qu.:11.00   1st Qu.:  19.990   1st Qu.:  1.0
 Median :19.00   Median :  36.588   Median :  4.0
 Mean   :16.69   Mean   :  65.142   Mean   : 14.6
 3rd Qu.:23.00   3rd Qu.:  70.192   3rd Qu.: 11.0
 Max.   :30.00   Max.   :1104.870   Max.   :367.0
  piece_count     play_star_rating review_difficulty
 Min.   :   1.0   Min.   :0.00     Min.   :0.000
 1st Qu.:  97.0   1st Qu.:3.60     1st Qu.:0.000
 Median : 216.0   Median :4.40     Median :2.000
 Mean   : 493.4   Mean   :3.71     Mean   :1.989
 3rd Qu.: 544.0   3rd Qu.:4.70     3rd Qu.:4.000
 Max.   :7541.0   Max.   :5.00     Max.   :5.000
```

- With the qplot, we can see that reviews are not even across ages and for better clarity , ages have been grouped into age group with 5 intervals, as below.






```
  Age num_reviews
1 [0,5]       21182
2 (5,10]      76120
3 (10,15]     10676
4 (15,20]     22174
5 (20,25]     20763
6 (25,30]     28133
```

  - Between the age group of 5~10, number of reviews were 76,120 followed by a distant second of 25~30 (28,133) and 15~20 (22,174).
- With the qplot, we can see that reviews are not even across ages and for better clarity , ages have been grouped (by 5) as below.
- Identifying the most popular Lego sets (i.e. most number of reviews) purchased by customers below 25 years old is:
  - No of customer reviews of **367**, age: 7, list price of 117~181.9 (median of 153.7), number of pieces: 1969, play star-rating: 4.6, and review difficulty of 1
  - The most expensive Lego set purchased by customers who are older than 25 years old is as follows:

ages	list_price	num_reviews	piece_count	play_star_rating	review_difficulty	country
29	259.87	6	1413	4.3	0	16

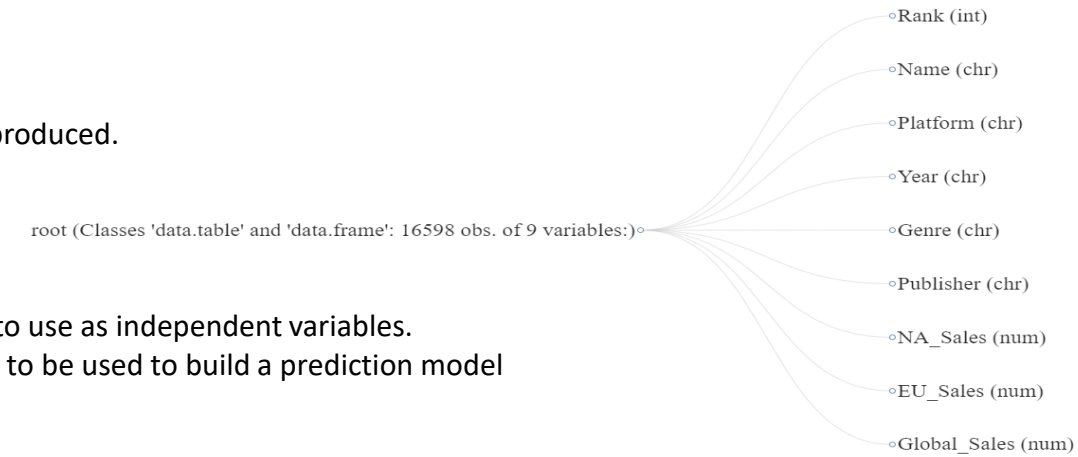
Turtle games (Sales Prediction)

[#Q4] EDA & Sales prediction

- The last question is to predict the sales of North American and European stores for the upcoming financial year using regression analysis.

1) Analysis (Using R)

- The data (game_sales.csv file) has 9 columns and 16,598 records
- Use the summary and 'DataExplorer' to do a sense check of the data where a Data Profiling Report is produced.
- No missing values identified
- Reviewing the data, sales prediction would need to be based on categorical independent variables (e.g. name, platform, year, genre, publisher) to build a sales prediction model
- To assess what could be considered as independent variables, count the unique number of cols
- Name and Publisher have unique number of cols of 11,493 and 579 respectively and look rather large to use as independent variables.
- Platform (31), Year (40), and Genre(12), on the other hand, could be considered as categorical variable to be used to build a prediction model
- Given that the 3 independent variables are characters, we would need to convert them to 'factors' for modelling. This is the equivalent of One Hot Encoding in Python.



- Using the linear regression model with the 3 independent variables above, the model summary in predicting the North America sales and the European sales were below.
  - North America sales: Adjusted R squared of 6.3% and p-value of 2.2 e-16
  - European sales: Adjusted R squared of 3.4% and p-value of 2.2 e-16. Reviewing the summary, we expect that Year has little explanatory power in predicting the EU_Sales hence better off to drop the variable.

2) Comments

- While the p-value of the 3 variables (Year, Platform, and Genre) are low of 2.2 e-16, the R square in both North-America and European sales are 6.3% and 3.4%, respectively hence using a different model other than linear regression could be considered.

GitHub

- Link to Github : https://github.com/krisrhee/Advanced-Analytics