Course 3: Final Assignment Python/R Presentation

Javier Conde Pascual | 12th September 2022

1. Scenario and system set up

Figure 1.1: Background, project goal and initial set of questions to answer

ROLE: assuming the role of data analyst working with game manufacturer and retailer Turtle Games. Its product range includes books, board games, video games, and toys.

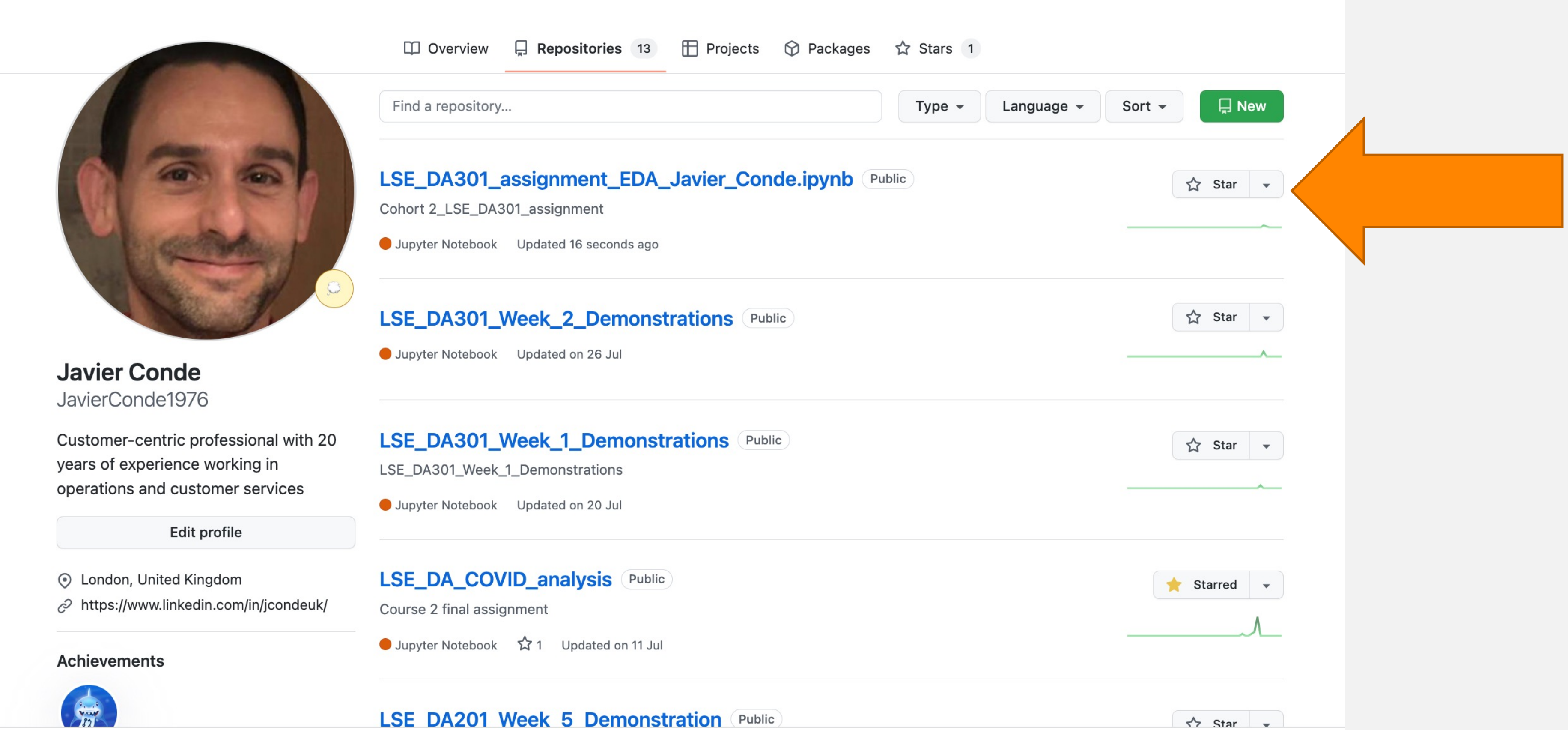
PROJECT GOAL: analyse available data from sales and customer reviews to extract and share insights with stakeholders**.** The ultimate target is improving overall sales performance by utilising customer trends.

INITIAL SET OF QUESTIONS:

- 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.

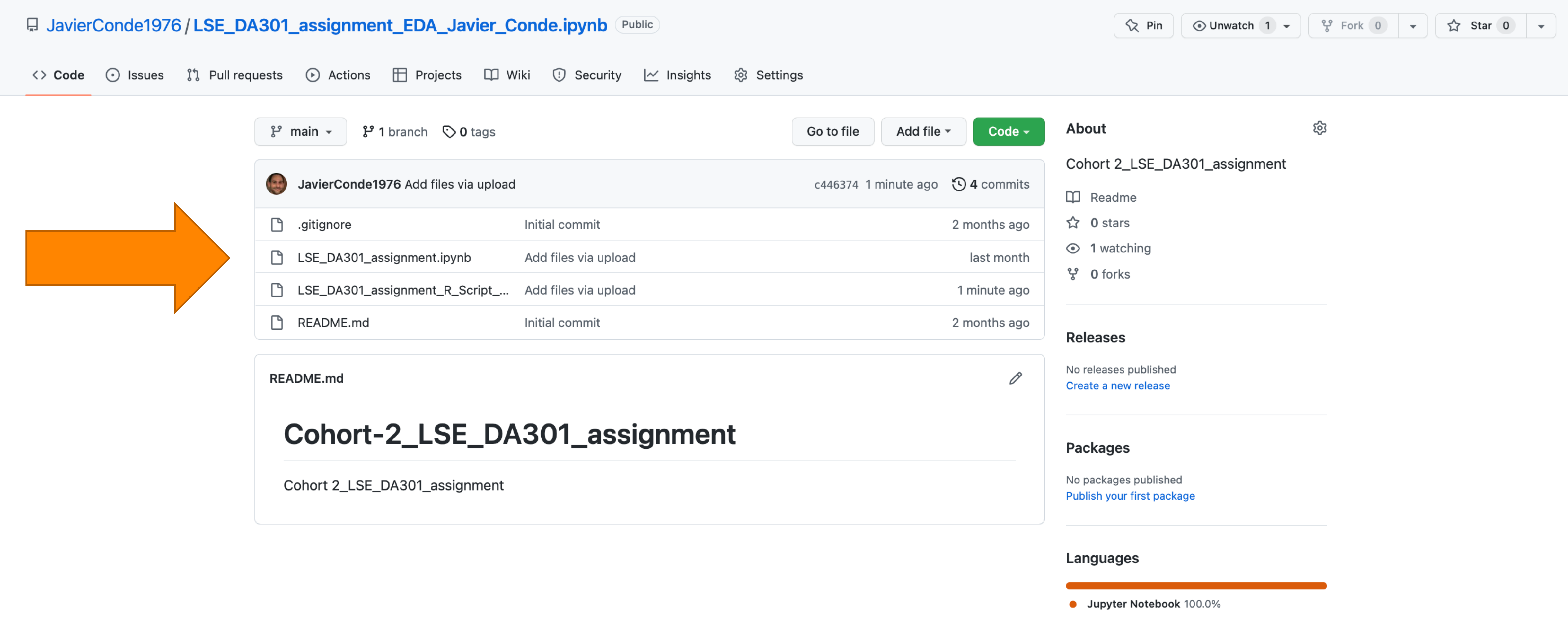
Source: LSE (2022)

<u>Figure 1.2:</u> Project Github repository (I)



Source: Javier Conde (2022)

Figure 1.3: Project Github repository (II)



Source: Javier Conde (2022)

Figure 3.1: Insights on customer reviews

Customer reviews

- Loyalty has a slight correlation with spending and remunerations score. This is advised to be studied further.

- Loyalty has almost no correlation with customer’s age

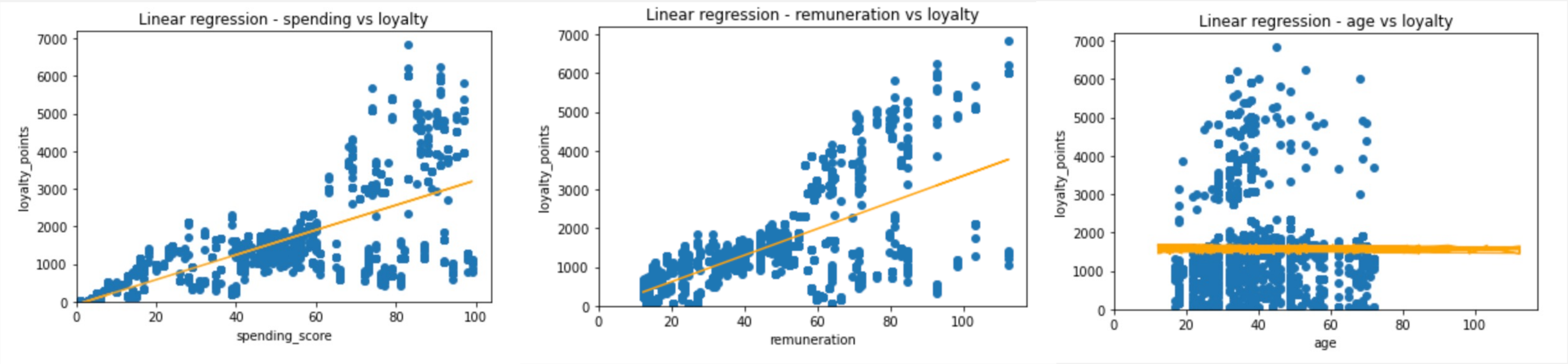
- K-means clustering analysis reveals 5 potential groups marketing department could consider:

    - low income/low spending
    - low income/higher spending
    - average income/average spending
    - higher income/low spending
    - higher income/higher spending

- The 5 most used words in the reviews are ‘stars’ (427) ‘five’ (342), ‘game’(319), ‘great’ (295) and ‘fun’(218), suggesting a positive sentiment confirmed by the polarity analysis (Mean +0.2 to +0.25)

Source: Javier Conde (2022)

<u>Figure 2.3:</u> Loyalty vs spending, remuneration, age



OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.452
Model:	OLS	**Adj. R-squared:**	0.452
Method:	Least Squares	**F-statistic:**	1648.

OLS Regression Results

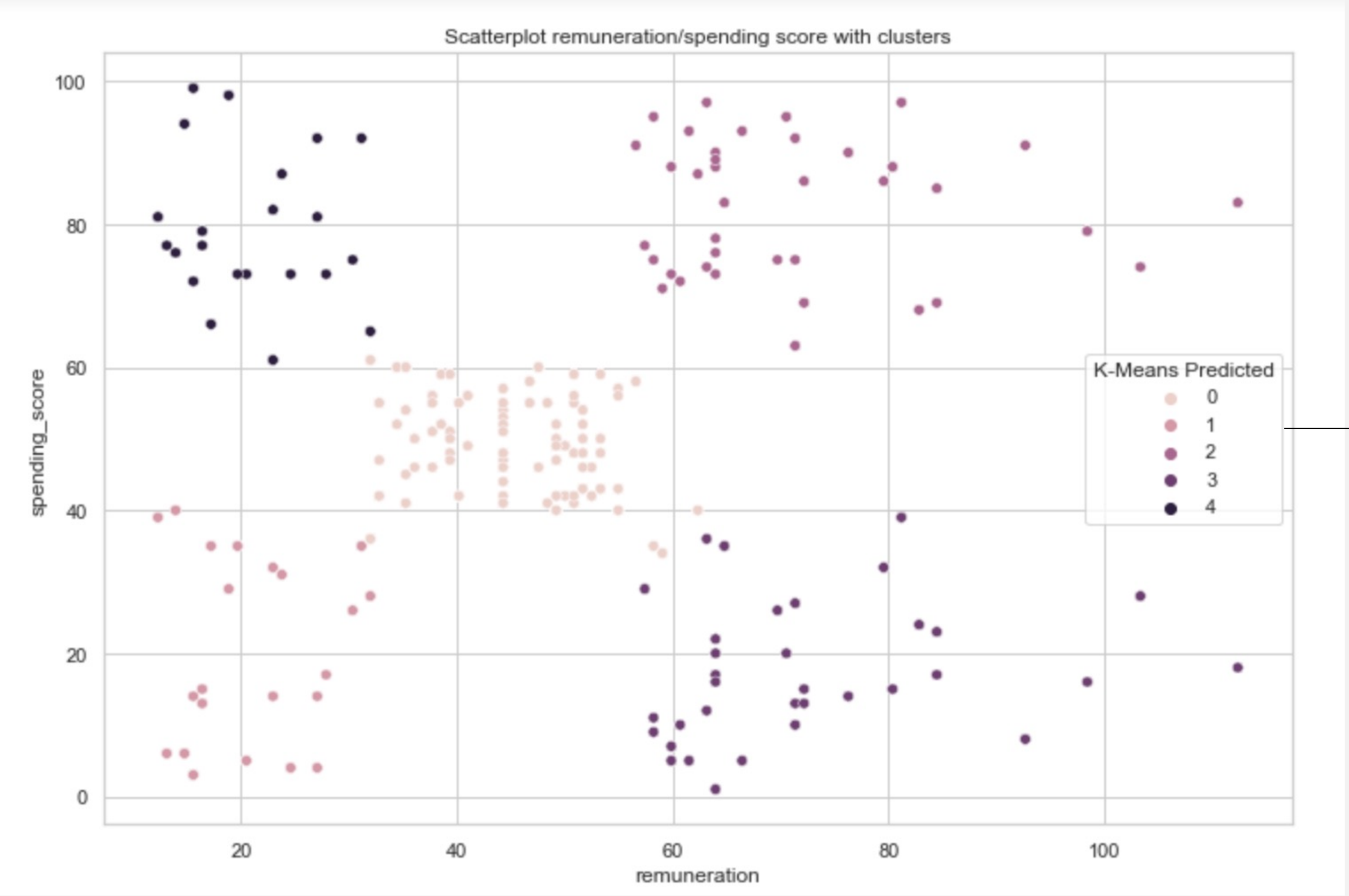
Dep. Variable:	y1	**R-squared:**	0.380
Model:	OLS	**Adj. R-squared:**	0.379
Method:	Least Squares	**F-statistic:**	1222.

OLS Regression Results

Dep. Variable:	y2	**R-squared:**	0.002
Model:	OLS	**Adj. R-squared:**	0.001
Method:	Least Squares	**F-statistic:**	3.606

Source: Javier Conde (2022)

<u>Figure 2.8:</u> *K-means chosen model for k=5 with cluster interpretation*

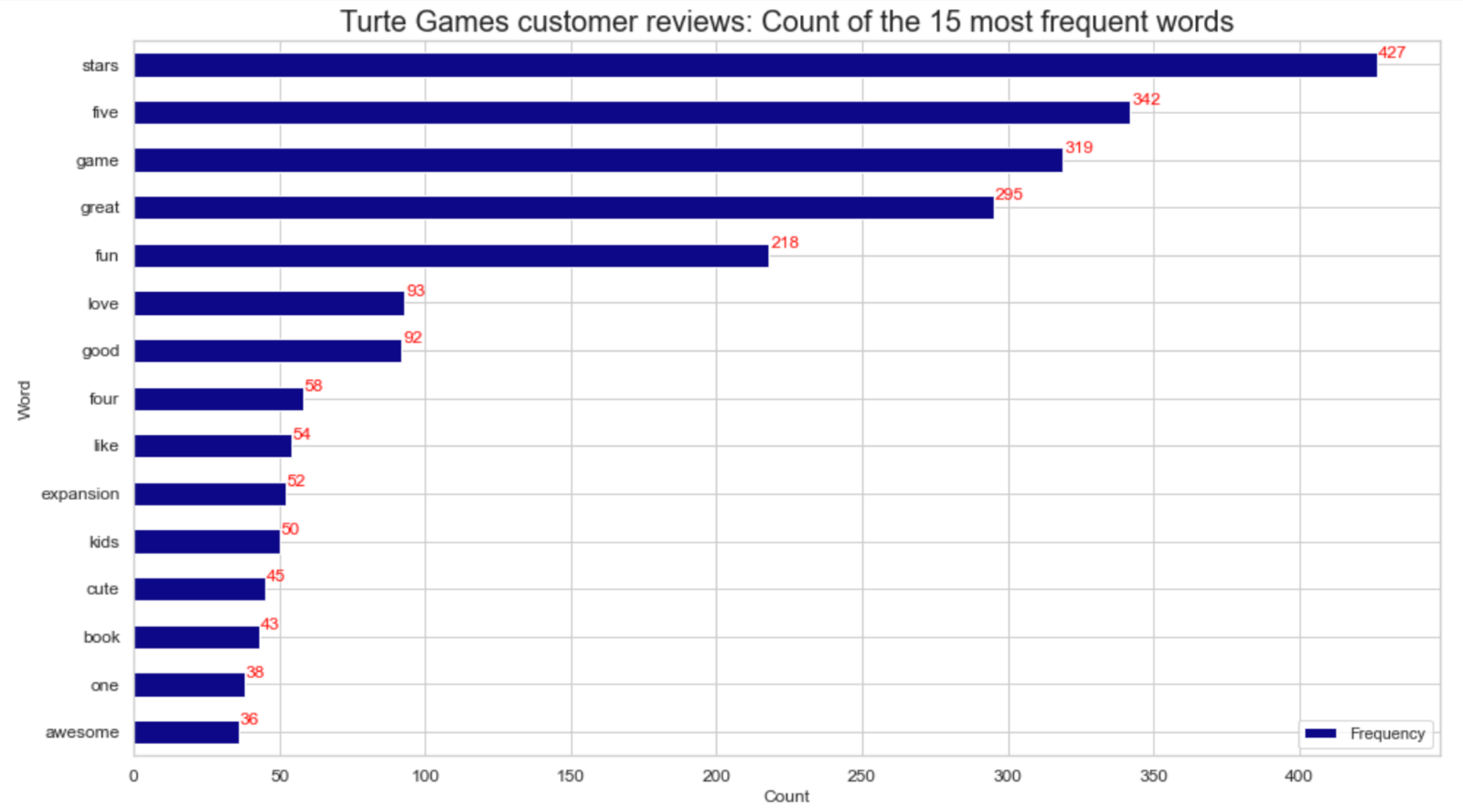


Cluster interpretation:

- group 1, low income and low spending
- group 4, low income and higher spending
- group 0, average income and average spending
- group 3, higuer income and low spending
- group 2, higher income and higher spending

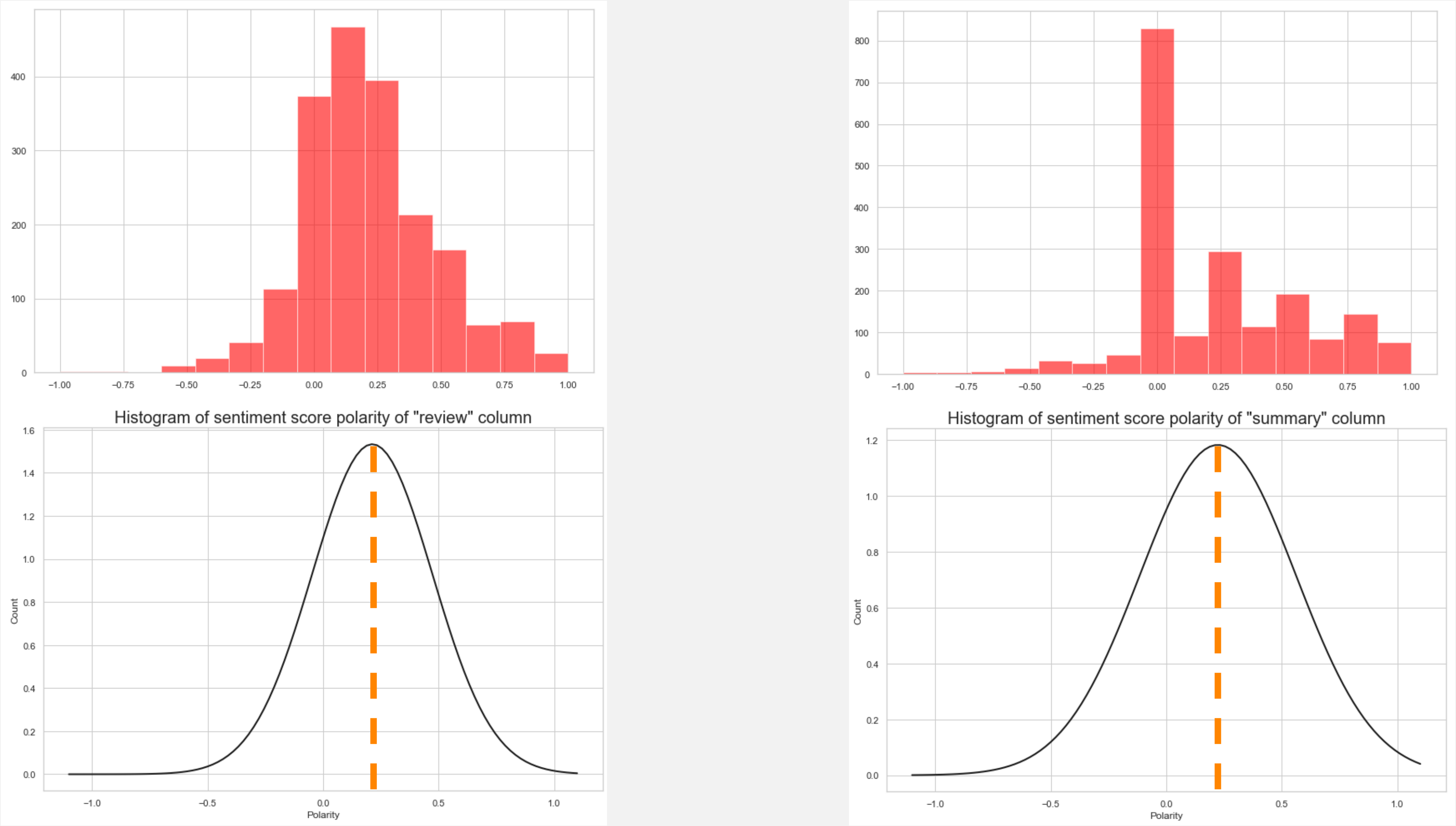
Source: Javier Conde (2022)

Figure 2.11: Plot for the data frame 15 most frequent words



Source: Javier Conde (2022)

Figure 2.12: Sentiment score polarity analysis for columns ‘review’ and ‘summary’



Source: Javier Conde (2022)

Figure 3.2: Insights on Turtle Games sales

Turtle Games sales

- The ‘turtle_sales’ data set provided is of great data quality (no duplicates, just two NA in ‘Year’ column). Outliers are not eliminated and they need to be closely monitored in further calculations

- From the initial scatterplot is visible a possible relationship between the variables ‘Global_Sales’ and ‘NA/EU_Sales’ robust enough to allow predictive studies on future global sales through a Multiple Linear Regression model with EU and NA sales as independent variables (modelA)

- Data sets provided on sales are far from normality (according to tests run on Q-Q plots, Shapiro-Wilk, skewness, and kurtosis), very relevant for further study

- Building a second model (modelB) including the variable “Product’ for further exploration could be considered, as it adds some robustness (but also complexity) to modelA.

- Tested on 5 aleatory observations, predictive accuracy of modelA results average to good

Source: Javier Conde (2022)

Figure 2.14: Data cleaning and subset creation: NA, duplicated observations, unnecessary columns

```r
# Explore NA and duplicates.

is.na(turtle_sales)
apply(is.na(turtle_sales), 2, which)

# Only two NA in 'Year' column, rows 180 and 258. Remove these columns in next
# step so no action taken.

duplicated(turtle_sales)

# No duplicated observations.

# Create a new data frame from a subset of the sales data frame.
# Remove unnecessary columns (Ranking, Year, Genre, Publisher).

turtle_sales2 <- subset(turtle_sales, select=-c(Ranking, Year, Genre, Publisher))

# View the data frame and structure.

turtle_sales2
str(turtle_sales2)

# View the descriptive statistics.

summary(turtle_sales2)
```

```
> apply(is.na(turtle_sales), 2, which)
$Ranking
integer(0)

$Product
integer(0)

$Platform
integer(0)

$Year
[1] 180 258

$Genre
integer(0)

$Publisher
integer(0)

$NA_Sales
integer(0)

$EU_Sales
integer(0)

$Global_Sales
integer(0)
```

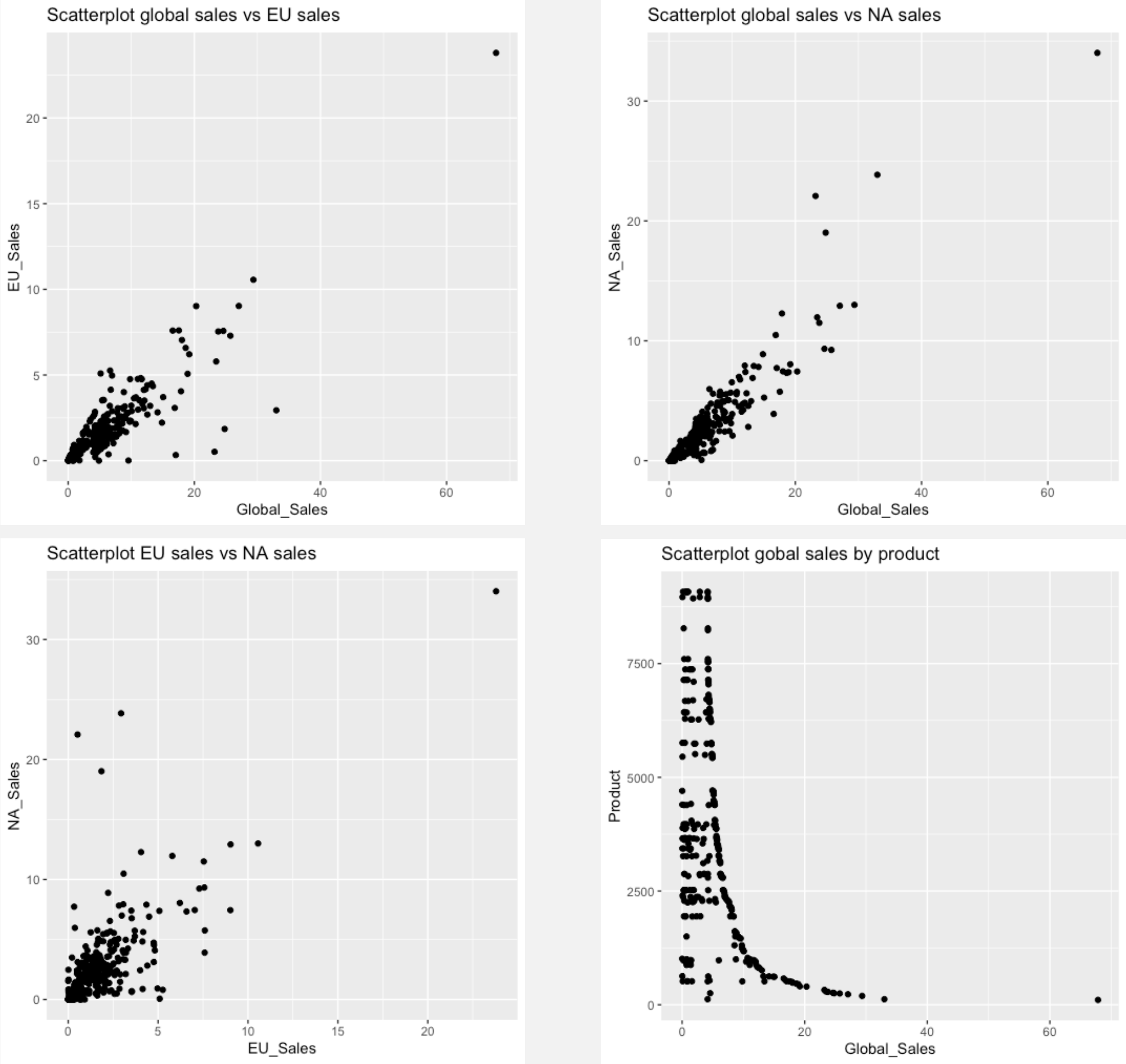
```
> duplicated(turtle_sales)
  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [34] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [45] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [56] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [67] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [78] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [89] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[100] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[111] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[122] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[144] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[155] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[166] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[177] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[188] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[199] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[210] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[221] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[232] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[243] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[254] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[265] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[276] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[287] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[298] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[309] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[320] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[331] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[342] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
> summary(turtle_sales2)
    Product        Platform            NA_Sales          EU_Sales        Global_Sales
 Min.   : 107   Length:352         Min.   : 0.0000   Min.   : 0.000   Min.   : 0.010
 1st Qu.:1945   Class :character   1st Qu.: 0.4775   1st Qu.: 0.390   1st Qu.: 1.115
 Median :3340   Mode  :character   Median : 1.8200   Median : 1.170   Median : 4.320
 Mean   :3607                      Mean   : 2.5160   Mean   : 1.644   Mean   : 5.335
 3rd Qu.:5436                      3rd Qu.: 3.1250   3rd Qu.: 2.160   3rd Qu.: 6.435
 Max.   :9080                      Max.   :34.0200   Max.   :23.800   Max.   :67.850
>
```

Source: Javier Conde (2022)

Figure 2.15: Exploratory scatterplots with qplot

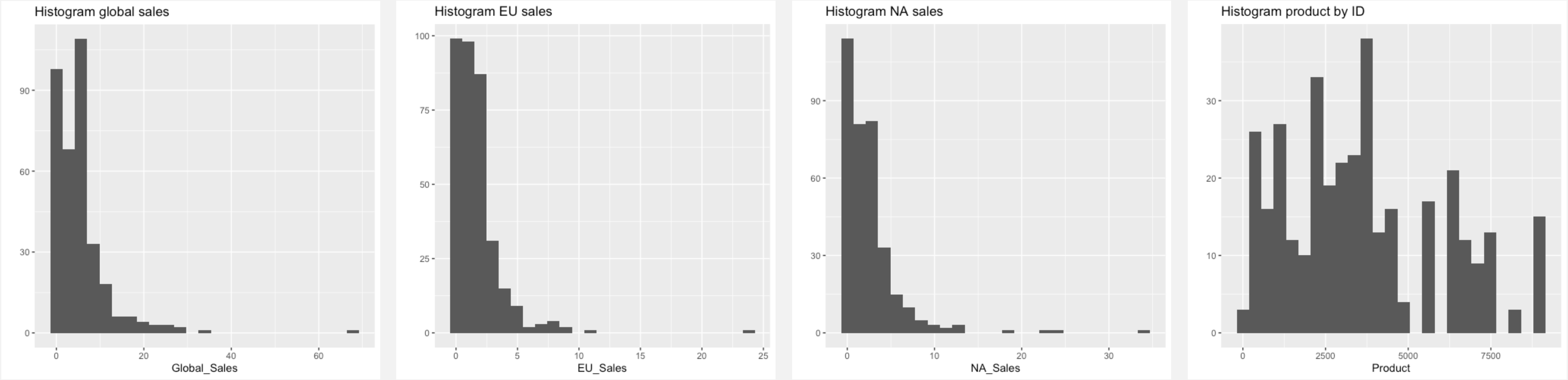
```r
qplot(Global_Sales, EU_Sales, data=turtle_sales2,
      main='Scatterplot global sales vs EU sales')

qplot(Global_Sales, NA_Sales, data=turtle_sales2,
      main='Scatterplot global sales vs NA sales')

qplot(EU_Sales, NA_Sales, data=turtle_sales2,
      main='Scatterplot EU sales vs NA sales')

qplot(Global_Sales, Product, data=turtle_sales2,
      main='Scatterplot gobal sales by product')
```



Source: Javier Conde (2022)

Figure 2.16: *Exploratory histograms with qplot*

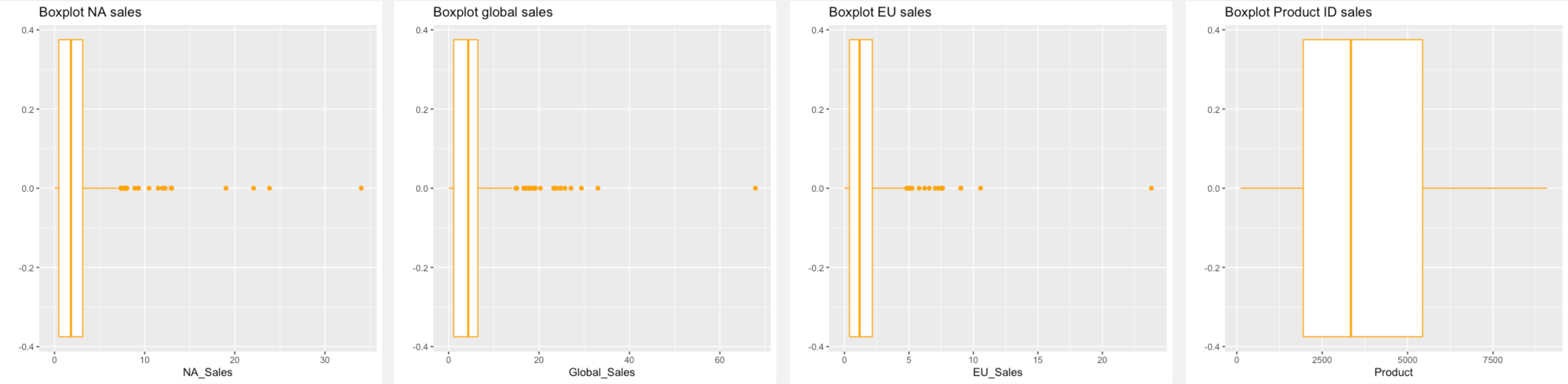
```r
qplot(Global_Sales, bins=25, data=turtle_sales2, main='Histogram global sales')
qplot(EU_Sales, bins=25, data=turtle_sales2, main='Histogram EU sales')
qplot(NA_Sales, bins=25, data=turtle_sales2, main='Histogram NA sales')
qplot(Product, bins=25, data=turtle_sales2, main='Histogram product by ID')
```



Source: Javier Conde (2022)

Figure 2.17: Exploratory boxplots with qplot

```
qplot(Global_Sales, data=turtle_sales2, colour=I('orange'),
      main='Boxplot global sales', geom='boxplot')

qplot(EU_Sales, data=turtle_sales2, colour=I('orange'),
      main='Boxplot EU sales', geom='boxplot')

qplot(NA_Sales, data=turtle_sales2, colour=I('orange'),
      main='Boxplot NA sales', geom='boxplot')

qplot(Product, data=turtle_sales2, colour=I('orange'),
      main='Boxplot NA sales', geom='boxplot')
```



Source: Javier Conde (2022)

<u>Figure 2.23:</u> Variables correlation

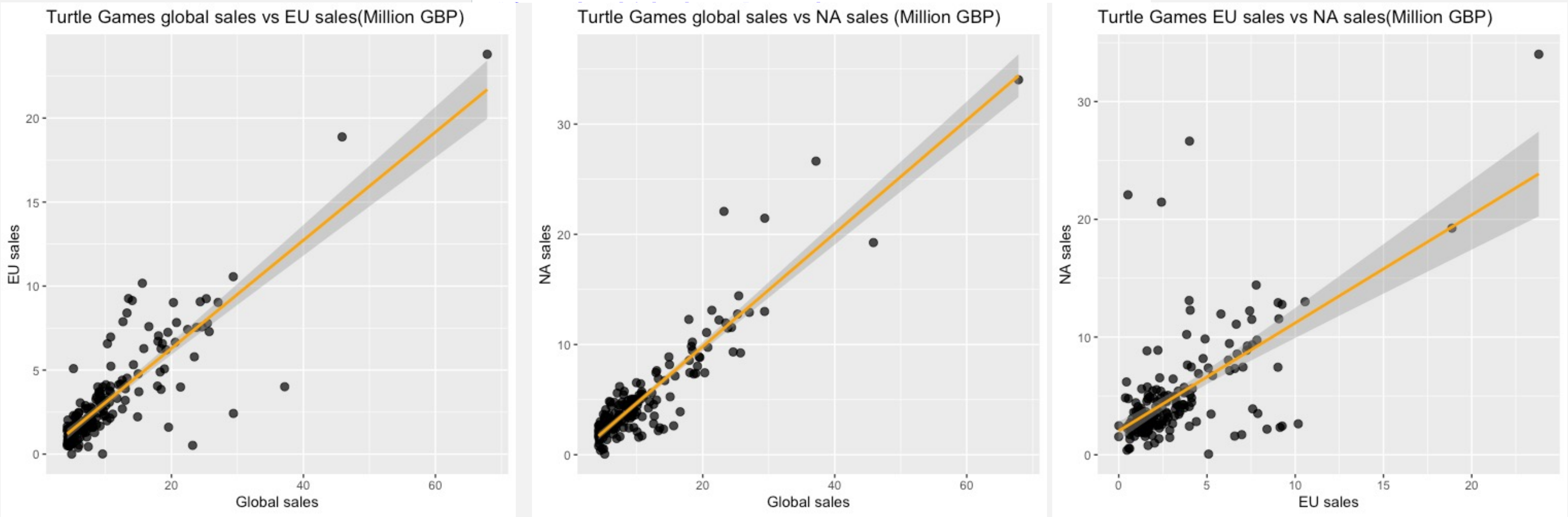
```r
# Correlation between the sales data columns.

round(cor(turtle_sales_product), digits=2)
```

```
> round(cor(turtle_sales_product), digits=2)
             Product NA_Sales EU_Sales Global_Sales
Product         1.00    -0.54    -0.45        -0.61
NA_Sales       -0.54     1.00     0.62         0.92
EU_Sales       -0.45     0.62     1.00         0.85
Global_Sales   -0.61     0.92     0.85         1.00
```

Source: Javier Conde (2022)

Figure 2.24: *Advanced plotting (ggplot with linear approach)*

```
> ggplot(data=turtle_sales_product,mapping=aes(x=Global_Sales, y=NA_Sales)) +
+   geom_point(color='black',
+              alpha=0.75,
+              size=2.5) +
+   geom_smooth(method='lm', color='orange') +
+   scale_x_continuous("Global sales") +
+   scale_y_continuous("North America sales") +
+   labs(title="Turtle Games global sales vs North America sales (Million GBP)")
`geom_smooth()` using formula 'y ~ x'
```



Source: Javier Conde (2022)

<u>Figure 2.31:</u> Multiple linear regression model (sales)

```
> modelA = lm(Global_Sales~NA_Sales+EU_Sales, data=turtle_sales_noproduct)
> summary(modelA)

Call:
lm(formula = Global_Sales ~ NA_Sales + EU_Sales, data = turtle_sales_noproduct)

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
```

Source: Javier Conde (2022)

Figure 2.32: Multiple linear regression model (sales and product ID)

```
> modelB = lm(Global_Sales~NA_Sales+EU_Sales+Product, data=turtle_sales_product)
> summary(modelB)

Call:
lm(formula = Global_Sales ~ NA_Sales + EU_Sales + Product, data = turtle_sales_product)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.3388 -0.9149 -0.2399  0.7364  5.9643 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.451e+00  3.167e-01   7.741 8.24e-13 ***
NA_Sales     1.068e+00  3.179e-02  33.601  < 2e-16 ***
EU_Sales     1.160e+00  4.421e-02  26.233  < 2e-16 ***
Product     -2.753e-04  5.278e-05  -5.215 5.26e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.388 on 171 degrees of freedom
Multiple R-squared:  0.9714,	Adjusted R-squared:  0.9709 
F-statistic:  1933 on 3 and 171 DF,  p-value: < 2.2e-16
```

Source: Javier Conde (2022)

Figure 2.33: *Value prediction (model A, sales) (I)*

```r
# A. NA_Sales_sum of 34.02 and EU_Sales_sum of 23.80

NA_Sales <- c(34.02)
EU_Sales <- c(23.80)

sales1 <- data.frame(NA_Sales, EU_Sales)

# Predicted Global_Sales value
predict(modelA, newdata = sales1)

# Predicted value 68.056 vs observation value 67.85 good
```

```r
# B. NA_Sales_sum of 3.93 and EU_Sales_sum of 1.56.
# Values not on provided data set
# Most similar 3.94/1.28 with observed Global_sales value 8.36

NA_Sales <- c(3.94)
EU_Sales <- c(1.28)

sales2 <- data.frame(NA_Sales, EU_Sales)

# Predicted Global_Sales value
predict(modelA, newdata = sales2)

# Predicted value 7.03 vs observation value 8.36: average
```

Source: Javier Conde (2022)

<u>Figure 2.34:</u> Value prediction (model A, sales)(II)

```r
# C. NA_Sales_sum of 2.73 and EU_Sales_sum of 0.65, observed value 4.32

NA_Sales <- c(2.73)
EU_Sales <- c(0.65)

sales3 <- data.frame(NA_Sales, EU_Sales)

# Predicted Global_Sales value
predict(modelA, newdata = sales3)

# Predicted value 4.90 vs observation value 4.32 good
```

```r
# D. NA_Sales_sum of 2.26 and EU_Sales_sum of 0.97.
# Values not on provided data set
# Most similar 2.27/2.30 with observed Global_sales value 5.60

NA_Sales <- c(2.27)
EU_Sales <- c(2.30)

sales4 <- data.frame(NA_Sales, EU_Sales)

# Predicted Global_Sales value
predict(modelA, newdata = sales4)

# Predicted value 6.36 vs observation value 5.60 average
```

Source: Javier Conde (2022)

<u>Figure 2.35:</u> Value prediction (model A, sales)(III)

```r
# E. NA_Sales_sum of 22.08 and EU_Sales_sum of 0.52, Global sales 23.21

NA_Sales <- c(22.08)
EU_Sales <- c(0.52)

sales <- data.frame(NA_Sales, EU_Sales)

# Predicted Global_Sales value
predict(modelA, newdata = sales)

# Predicted value 26.62 vs observation value 23.21: average
```

Source: Javier Conde (2022)

Figure 3.3: Recommendations to the marketing department

- Consider the 5 groups uncovered during the analysis to investigate further other possibly useful relationships (suggested to start with loyalty/gender, loyalty/education, loyalty/product (any product customers keep coming back for?)

- Study further the products associated to great reviews where these words appear. Are there any products with great reviews that haven't been given the marketing exposure?

- Study further also products associated to more negative reviews, data may be available here to understand lack of sales/interest and how to amend it

- Due to the good quality of the data provided, if budget and resources allow consider expanding the study to a bigger sample for more in-depth insights, maybe considering other social media (Instagram, Twitter)

- Ensure communication with other departments is fluid, this may be key to shift projects' priorities (i.e. understanding best selling products per region, provided by the sales department) (Figure 3.4)

Source: Javier Conde (2022)

Figure 3.4: *Turtle Games best sellers*

Product	NA_Sales	EU_Sales	Global_Sales
107	34.02	23.80	67.85
515	19.25	18.88	45.86
123	26.64	4.01	37.16
254	21.46	2.42	29.39
195	13.00	10.56	29.37
231	12.92	9.03	27.06
249	9.24	7.29	25.72
948	14.42	7.79	25.45

Product	NA_Sales	EU_Sales	Global_Sales
107	34.02	23.80	67.85
515	19.25	18.88	45.86
195	13.00	10.56	29.37
3967	2.63	10.17	15.59
2371	2.44	9.26	13.49
876	12.77	9.25	25.28
3645	2.33	9.14	14.06
979	11.55	9.07	24.36

Product	NA_Sales	EU_Sales	Global_Sales
107	34.02	23.80	67.85
123	26.64	4.01	37.16
326	22.08	0.52	23.21
254	21.46	2.42	29.39
515	19.25	18.88	45.86
948	14.42	7.79	25.45
535	13.11	3.99	21.38
195	13.00	10.56	29.37

Source: Javier Conde (2022)

Figure 3.5: *Recommendations to the sales department*

- Consider investing budget and resources in further study on MLR models, maybe including variables like product or game genre, with an extended data set to improve accuracy. This may also improve normality in the data set

- Consider comparing sales data from various years and stablish a 3/5 year time-series study with the sales evolution of products/game genres/platforms per region

- Communicate with other departments (i.e. marketing) sales figures to develop a joint strategy on how to promote products that could be potential hits in the future. Consider creating an interactive dashboard (i.e. Tableau, interactive visualisations in RStudio) for other departments to have access to sales information in real time

Source: Javier Conde (2022)

Thank You