Youko Fujino – LSE | Data Analytics Career Accelerator
DA301_Assignment
22.04.2024

Predicting Outcomes for Turtle Games

BACKGROUND & OBJECTIVE

Turtle Games, a global game manufacturer and retailer, has hired your data team to analyse customer trends. The goal is to improve sales performance by mapping their customer profile and enhancing the efficiency of their loyalty programme.

MODULE 1

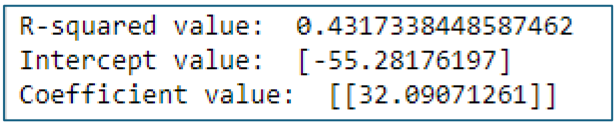
1. **Import files, packages and libraries**
2. **Pre-processing**
3. **Descriptive statistics**

TG customers are relatively young, with an average age of 39 and an average remuneration of £48k. 25% are 29, while the majority are between 29 and 49. On average, their spending score is 50 out of 100, and they have an average of 1578 loyalty points, ranging from 25 to 6847, indicating a wide variation in customer engagement.

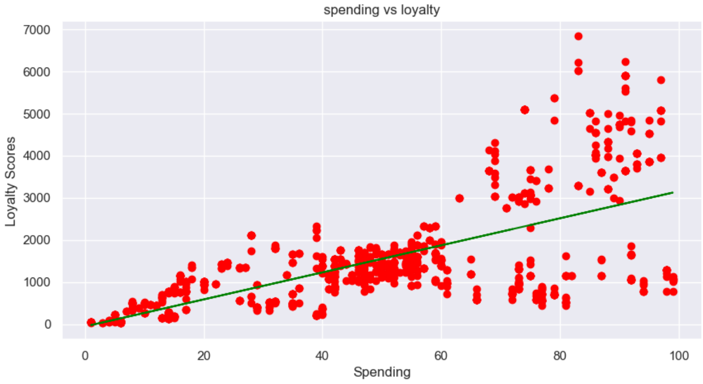
4. Linear regression to predict loyalty points accumulation

Steps in the appendix[A]

5. **LM: Spending vs Loyalty**

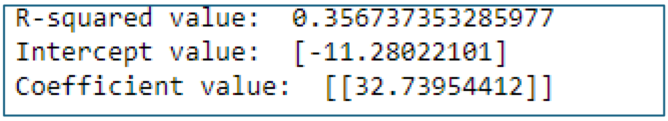


```
R-squared value:  0.4317338448587462
Intercept value:  [-55.28176197]
Coefficient value:  [[32.09071261]]
```

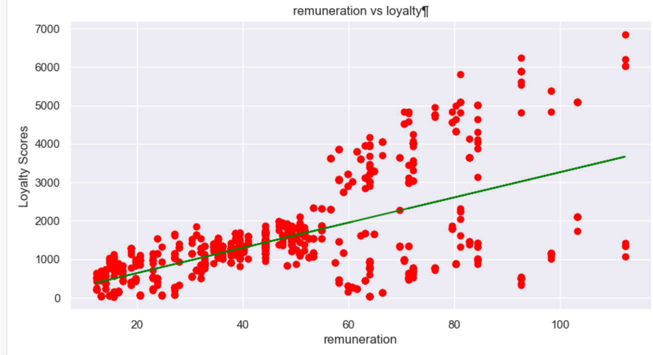


- R-squared : changes in spending can explain 43.17% of the variability in loyalty points.
- Coeff (32.09): for each unit increase in spending_score, the loyalty points increase by 32.09 units.
- Intercept(55.28): implies loyalty points could be harmful when the spending score is zero, which doesn’t have a practical meaning.
- A significant positive relationship between spending score and loyalty points

6. **LM: Remuneration vs Loyalty**

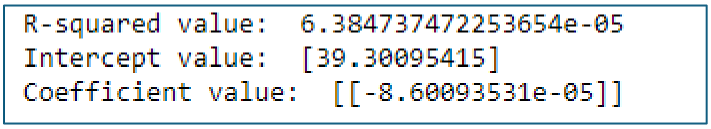


```
R-squared value:  0.356737353285977
Intercept value:  [-11.28022101]
Coefficient value:  [[32.73954412]]
```

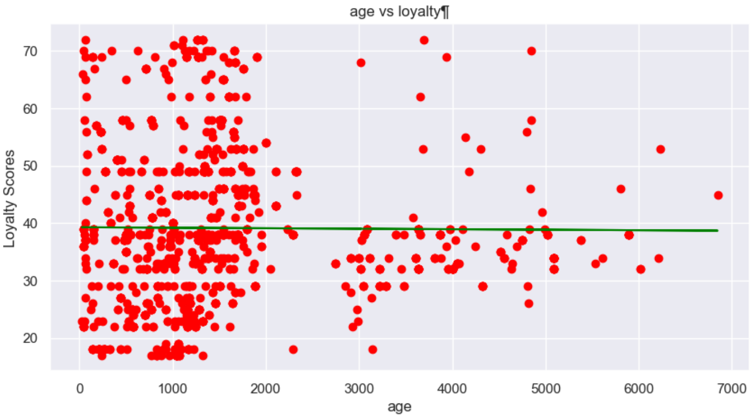


- R-squared: changes in remuneration can explain around 35.67% of the variability in loyalty points.
- Coefficient 32.74: for every unit increase in remuneration, the loyalty points increase by 32.74.
- The intercepts suggest that the estimated loyalty points are around -11.28 when remuneration is zero, which is not practically meaningful.
- Remuneration and loyalty points have a positive but weak relationship.

7. LM: Age vs Loyalty

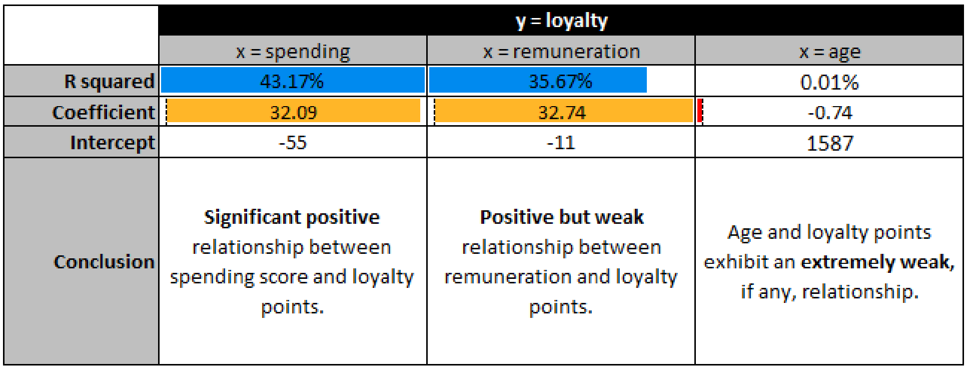


```
R-squared value:  6.384737472253654e-05
Intercept value:  [39.30095415]
Coefficient value:  [[-8.60093531e-05]]
```



- R-squared is very low (0.000064): age explains minimal variability in loyalty points.
- Coefficient: also very low (-0.742): tiny decrease of -0.742 in loyalty points for each year increase in age.
- Age and loyalty points display an extremely weak, if any, relationship.

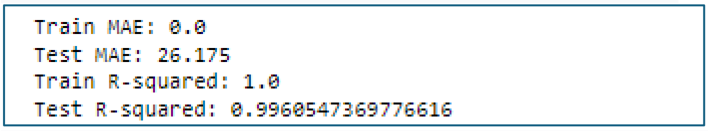
8. Conclusions



	y = loyalty		
	x = spending	x = remuneration	x = age
R squared	43.17%	35.67%	0.01%
Coefficient	32.09	32.74	-0.74
Intercept	-55	-11	1587
Conclusion	**Significant positive** relationship between spending score and loyalty points.	**Positive but weak** relationship between remuneration and loyalty points.	Age and loyalty points exhibit an **extremely weak**, if any, relationship.

MODULE 2

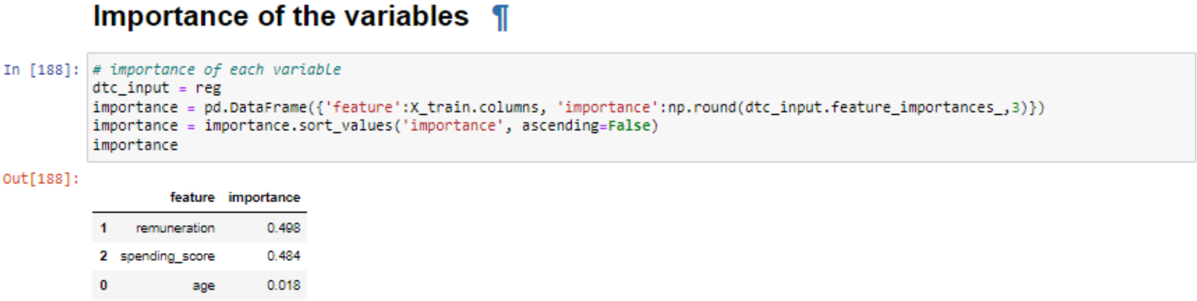
I used only the numeric columns (spending_score, age and remuneration) to build the decision tree to predict loyalty points. Then, I split the data in training.



```
Train MAE: 0.0
Test MAE: 26.175
Train R-squared: 1.0
Test R-squared: 0.9960547369776616
```

The model performed well on the training data, with an MAE of 0 and a perfect R-squared value of 1. The MAE was low on test data, at 26.175, but the R-squared value was slightly lower, at approximately 0.996. Overall, the model showed high predictive accuracy.

I ranked the importance of the variables, with remuneration and spending_score showing roughly the same values.



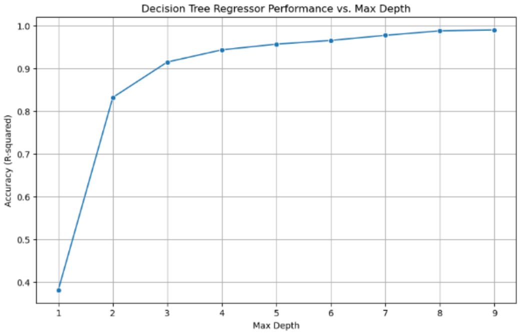
Importance of the variables ¶

```
In [188]: # importance of each variable
          dtc_input = reg
          importance = pd.DataFrame({'feature':X_train.columns, 'importance':np.round(dtc_input.feature_importances_,3)})
          importance = importance.sort_values('importance', ascending=False)
          importance
```

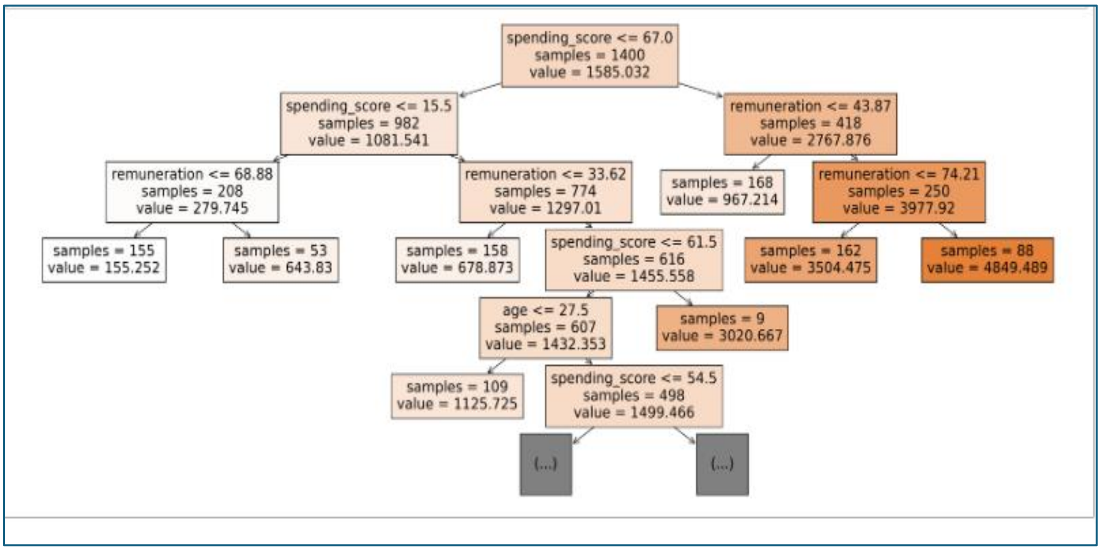
Out[188]:

	feature	importance
1	remuneration	0.498
2	spending_score	0.484
0	age	0.018

I test the ideal length of the Tree before pruning it:



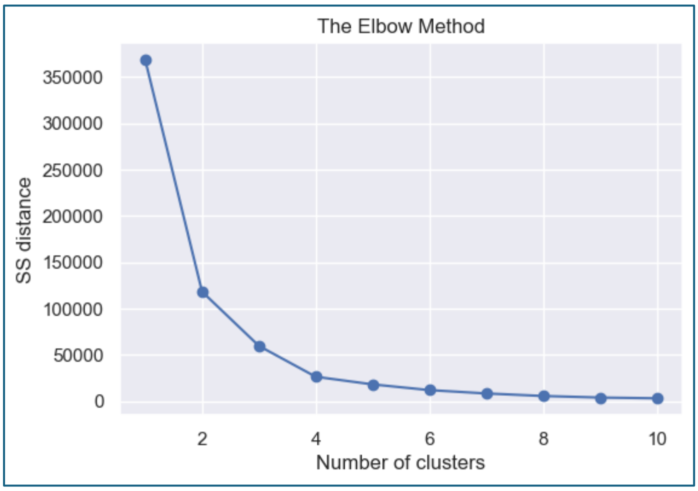
Obtaining the final pruned tree:



MODULE 3

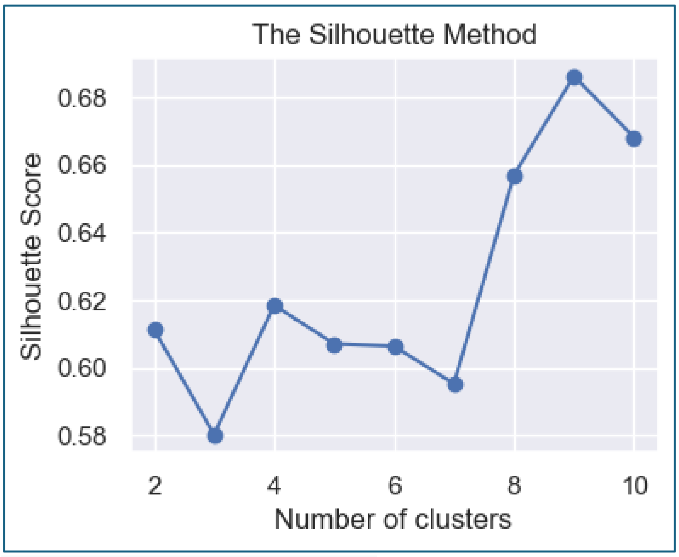
K-means clustering groups data based on similarity using centroids and a chosen distance metric. The algorithm splits data into clusters based on distances from centroids. To determine the optimal number of clusters, the Elbow and Silhouette Method are used to reduce the dataset while maintaining group homogeneity.

Elbow Method
4 to 6 is the ideal number of clusters to maintain a significant level of homogeneity.

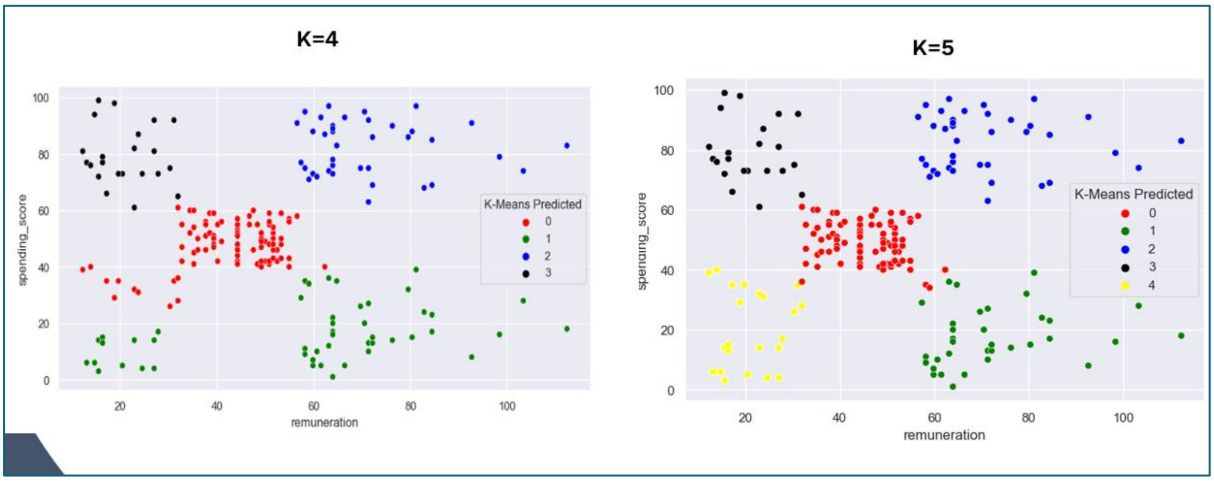


The Silhouette
The silhouette score measures an object's similarity to its assigned cluster compared to others.
In the chart below, 4 clusters are preferable to 5,6,7.

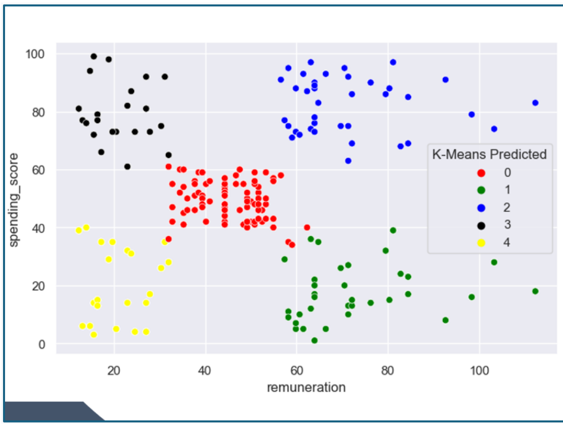
Youko Fujino – LSE | Data Analytics Career Accelerator
DA301_Assignment
22.04.2024



Tested and plotted k=4 and k=5



Choosing the right K is a subjective choice. The plotted data suggests that K=5 brings a higher level of homogeneity. Therefore, I continued with naming the cluster as presented below.
Prioritise high-spending consumers but push incentives for the 'potential' cluster.

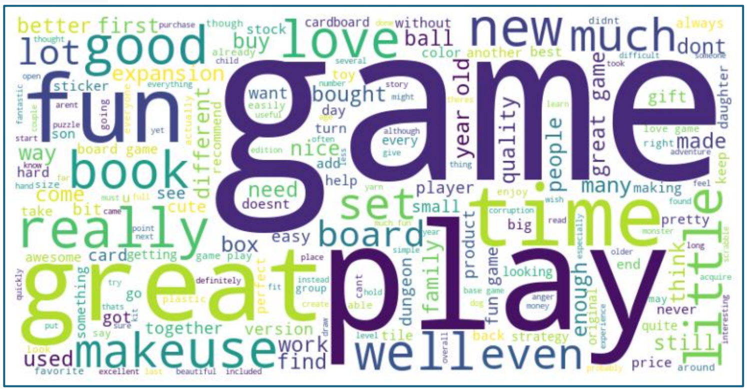


1. **Potential: (Green)** – High income **but** low spending score
2. **Average (Red)** – Avg remuneration **and** avg spending score
3. **VIP (Blue)**– High-income **and** high spending
4. **Irresponsible (Black)** – Low income **but** high spending score
5. **Irrelevant (Yellow)** – low income **and** low spending score

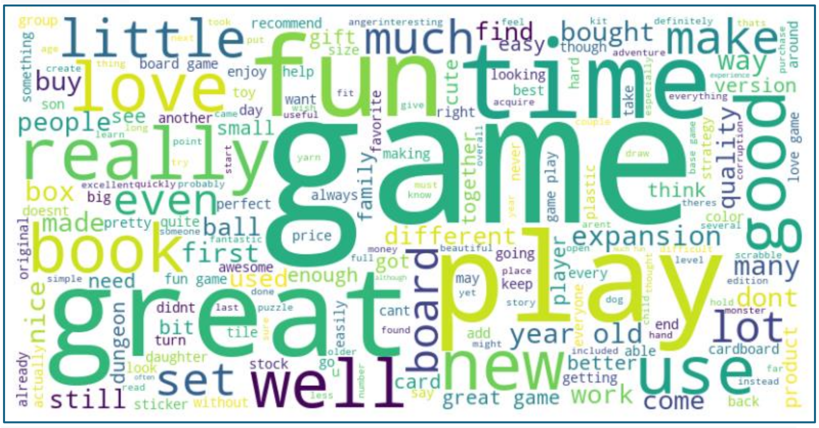
MODULE 4

To perform sentiment analysis, I preprocessed the data, tokenised comments, and removed stopwords and alphanumeric words. Then, I generated the wordclouds.

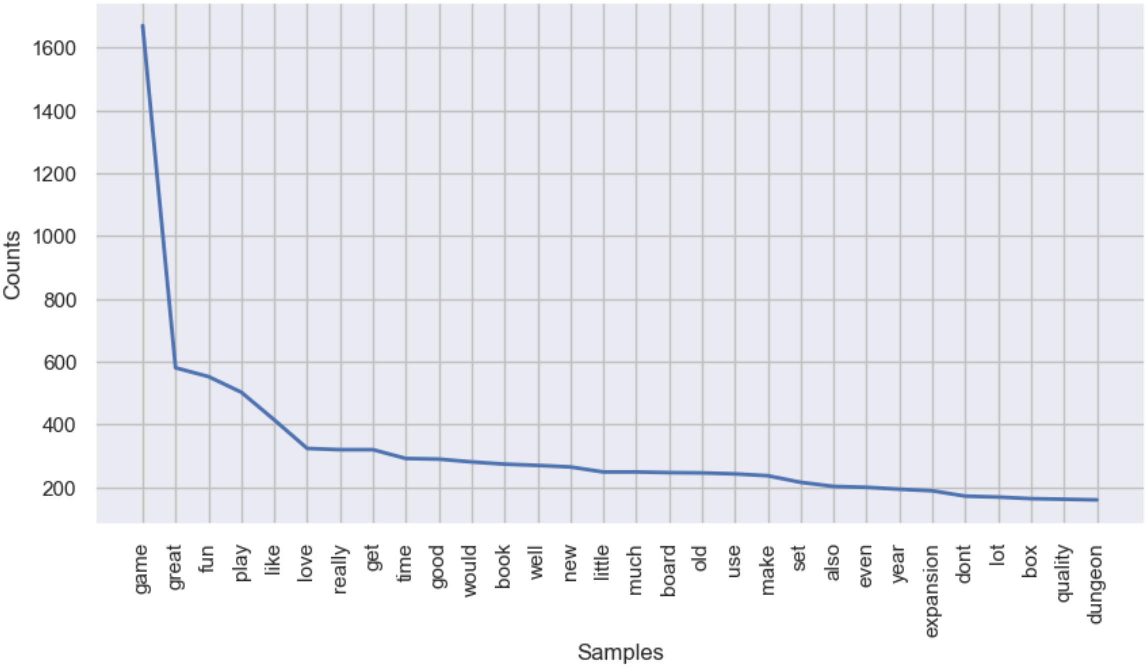
- Reviews



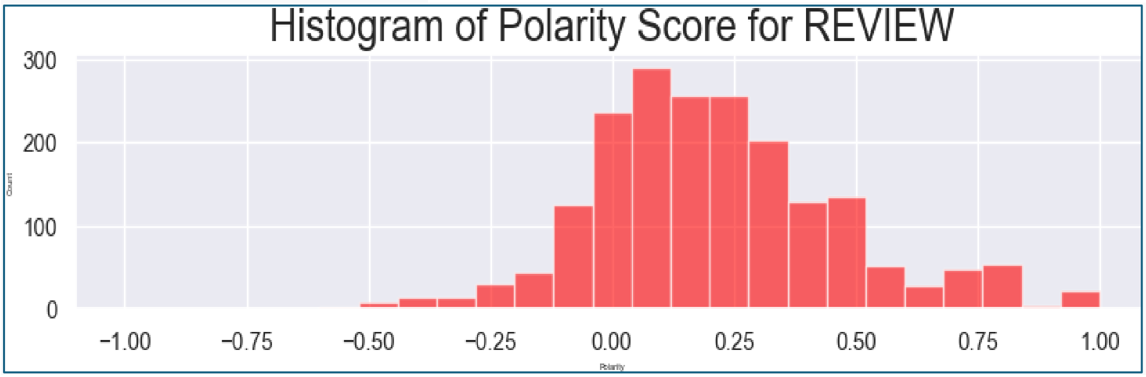
Summary:

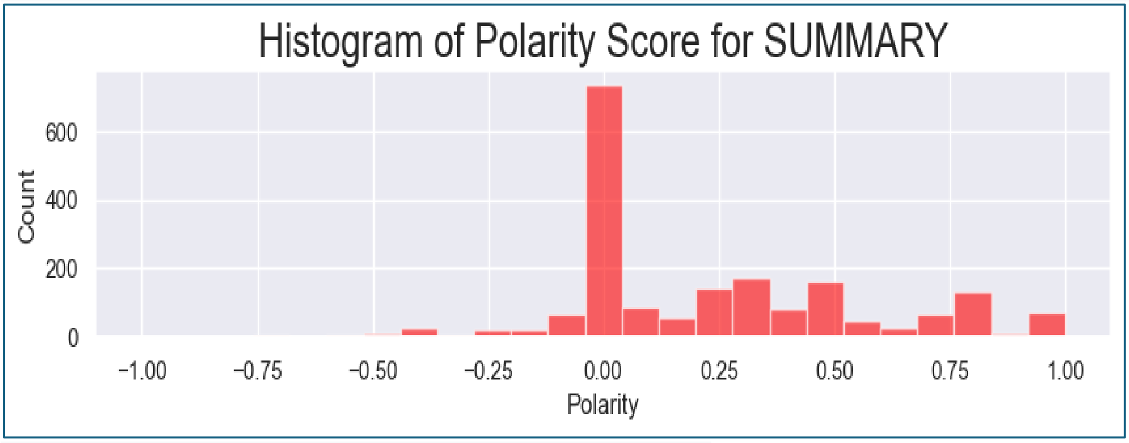


- Game, Great and Fun are the most common words in the reviews

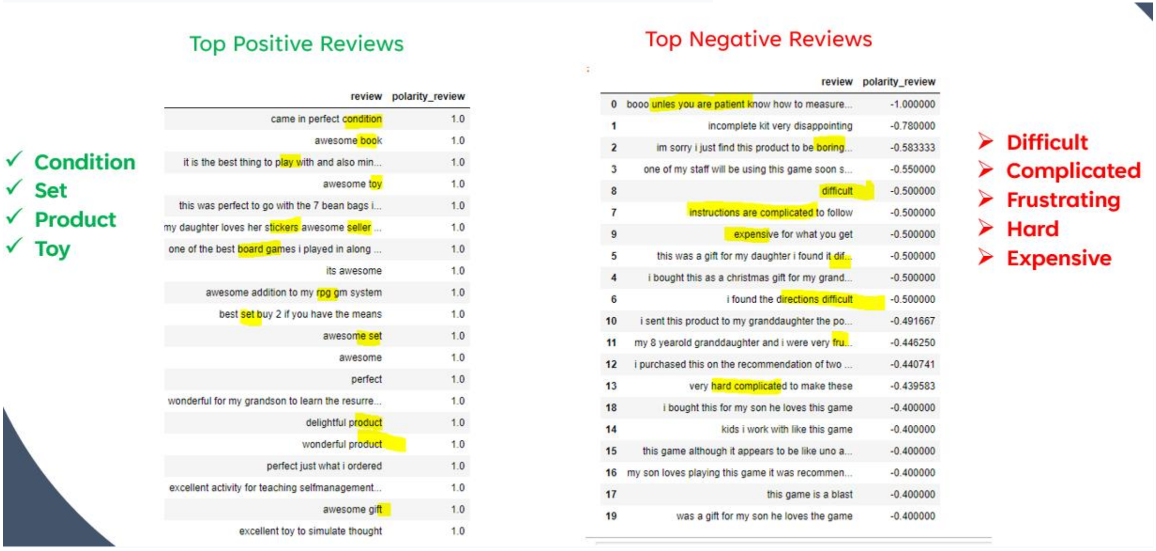


The polarity histograms of reviews and summaries show a right-tailed distribution, indicating an overall positive sentiment among reviews. When analysing the Subjectivity Scores, higher levels were found for reviews than summaries.[B]





Top Positive/ Top Negative Reviews using the polarity score



Top Positive Reviews

review	polarity_review
came in perfect condition	1.0
awesome book	1.0
it is the best thing to play with and also min...	1.0
awesome toy	1.0
this was perfect to go with the 7 bean bags i...	1.0
my daughter loves her stickers awesome seller ...	1.0
one of the best board games i played in along ...	1.0
its awesome	1.0
awesome addition to my rpg gm system	1.0
best set buy 2 if you have the means	1.0
awesome set	1.0
awesome	1.0
perfect	1.0
wonderful for my grandson to learn the resurre...	1.0
delightful product	1.0
wonderful product	1.0
perfect just what i ordered	1.0
excellent activity for teaching selfmanagement...	1.0
awesome gift	1.0
excellent toy to simulate thought	1.0

- ✓ Condition
- ✓ Set
- ✓ Product
- ✓ Toy

Top Negative Reviews

	review	polarity_review
0	booo unles you are patient know how to measure...	-1.000000
1	incomplete kit very disappointing	-0.780000
2	im sorry i just find this product to be boring...	-0.583333
3	one of my staff will be using this game soon s...	-0.550000
8	difficult	-0.500000
7	instructions are complicated to follow	-0.500000
9	expensive for what you get	-0.500000
5	this was a gift for my daughter i found it dif...	-0.500000
4	i bought this as a christmas gift for my grand...	-0.500000
6	i found the directions difficult	-0.500000
10	i sent this product to my granddaughter the po...	-0.491667
11	my 8 yearold granddaughter and i were very fru...	-0.446250
12	i purchased this on the recommendation of two ...	-0.440741
13	very hard complicated to make these	-0.439583
18	i bought this for my son he loves this game	-0.400000
14	kids i work with like this game	-0.400000
15	this game although it appears to be like uno a...	-0.400000
16	my son loves playing this game it was recommen...	-0.400000
17	this game is a blast	-0.400000
19	was a gift for my son he loves the game	-0.400000

- ➢ Difficult
- ➢ Complicated
- ➢ Frustrating
- ➢ Hard
- ➢ Expensive

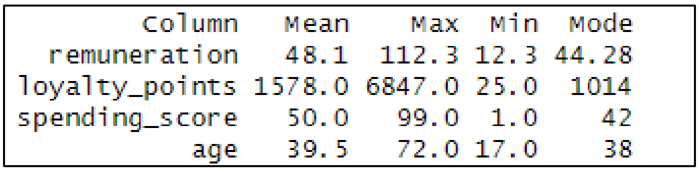
Overall, consumers show a positive sentiment towards TG, with conditions, set and the product mentioned in top positive reviews. However, there are complaints about the games' difficulty and complexity and instructions.

MODULE 5

Steps:

Youko Fujino – LSE | Data Analytics Career Accelerator
DA301_Assignment
22.04.2024

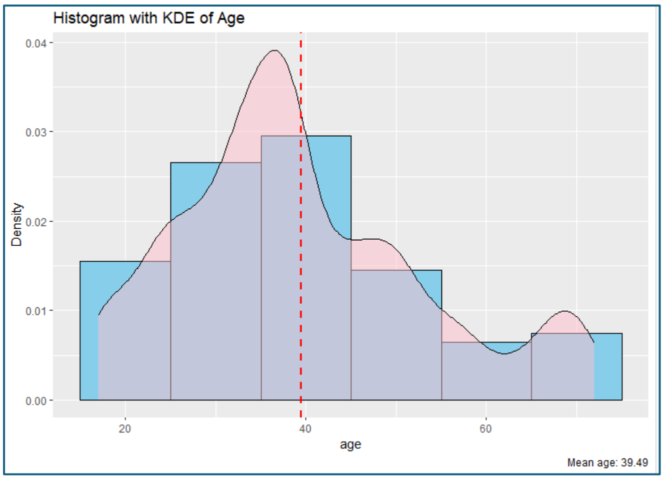
- Install Packages & Import Libraries
- Pre-processing Data
- Exploratory Data Analysis



```
         Column   Mean    Max  Min  Mode
   remuneration   48.1  112.3 12.3 44.28
 loyalty_points 1578.0 6847.0 25.0  1014
 spending_score   50.0   99.0  1.0    42
            age   39.5   72.0 17.0    38
```

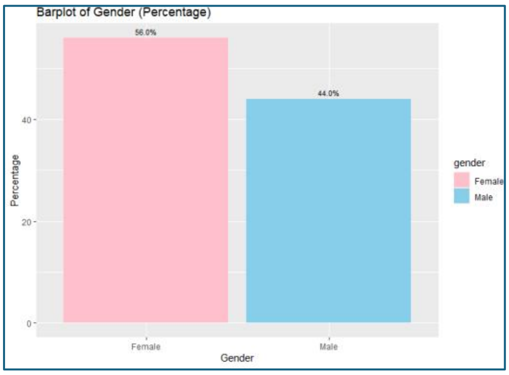
---

1) AGE



Customers range in age from 17 to 72, with a mode of 38. The KDE exhibits a left-tailed distribution, with a higher density of younger customers.

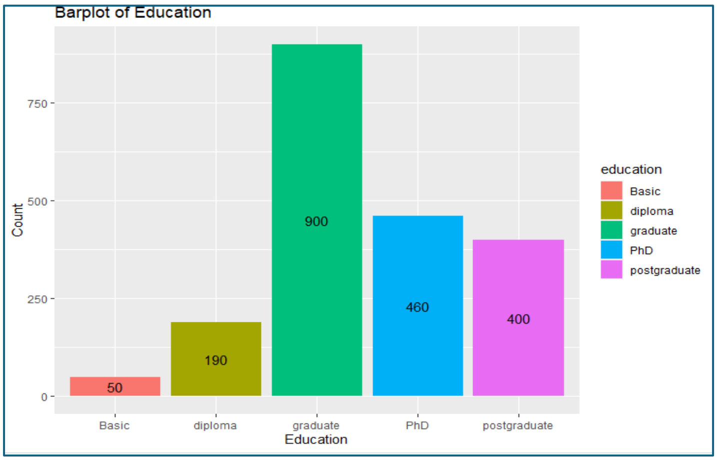
2) GENDER



Majority women (56%), men (44%)

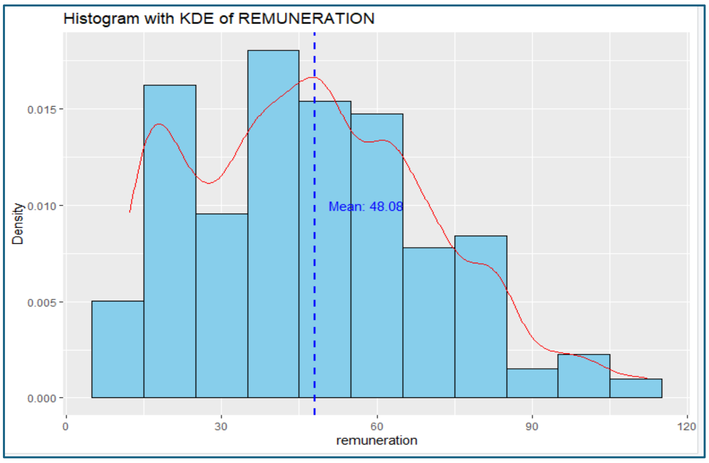
3) EDUCATION

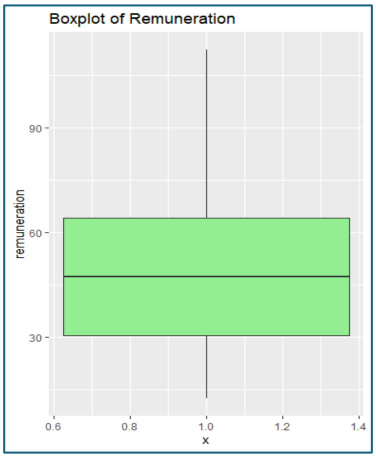
Most consumers are highly educated, with bigger groups: 1) graduate, 2)PHD, 3) Postgraduate.



4) REMUNERATION

Customers receive a remuneration of $48.1k annually, with a concentration of values between $15k and $60k. Beyond $60k/year, the distribution falls, showing a right-tailed pattern as evidenced by the KDE line.

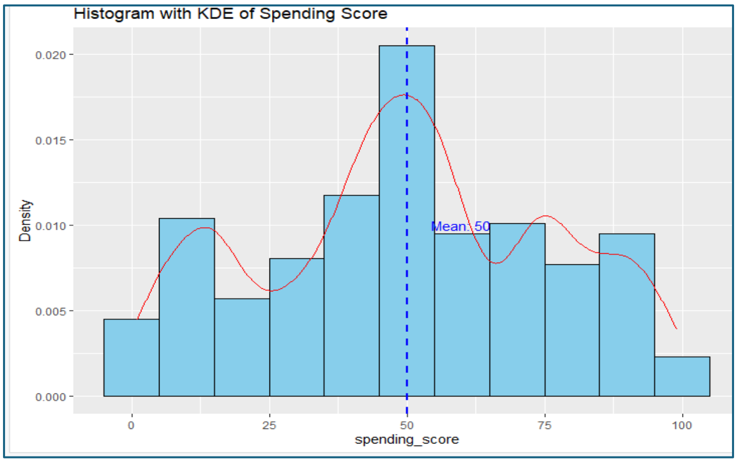


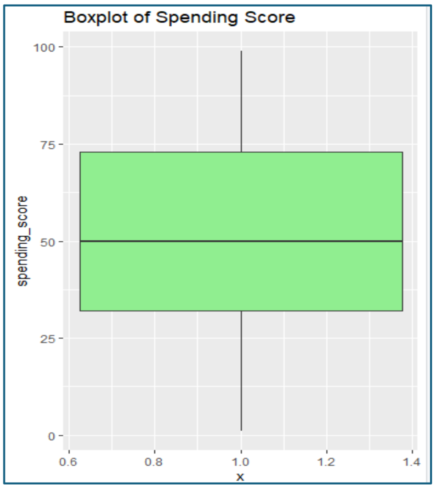


.

5) SPENDING SCORE

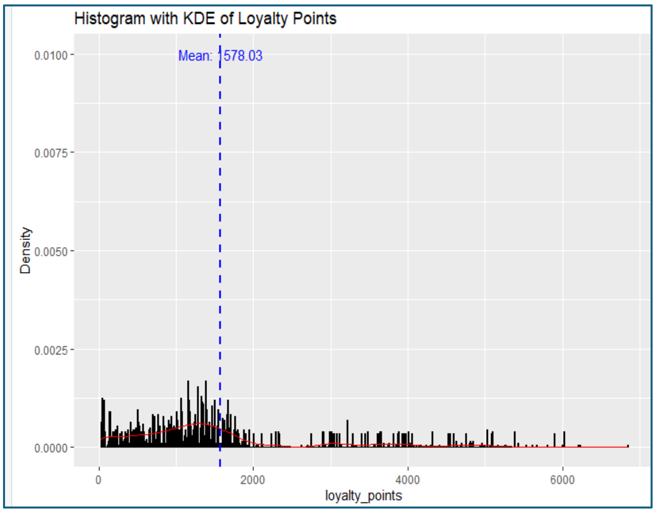
The spending score exhibits a reasonably symmetrical distribution, centred around an average of 50.0. A prominent peak around the 50k range resembles a bell-shaped curve indicative of a normal distribution.

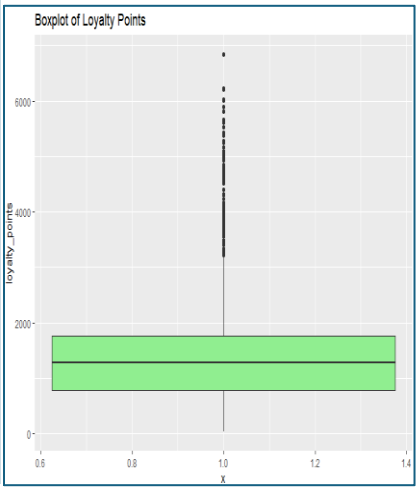




6) LOYALTY POINTS

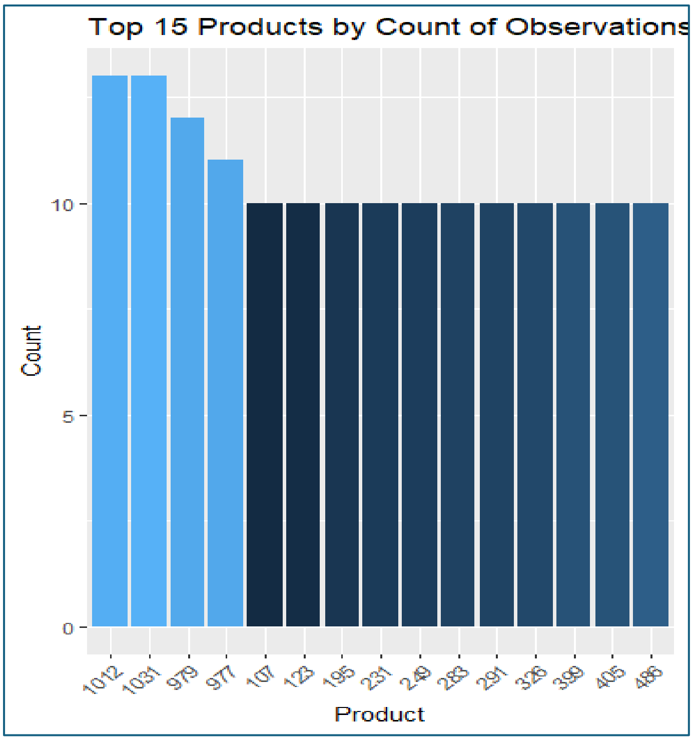
The loyalty points boxplot indicates a wide presence of outliers in the high-end loyalty system. Most consumers have a range between 1k and 2k loyalty points, with outliers superior to 3k.





7) PRODUCTS CONSUMPTION

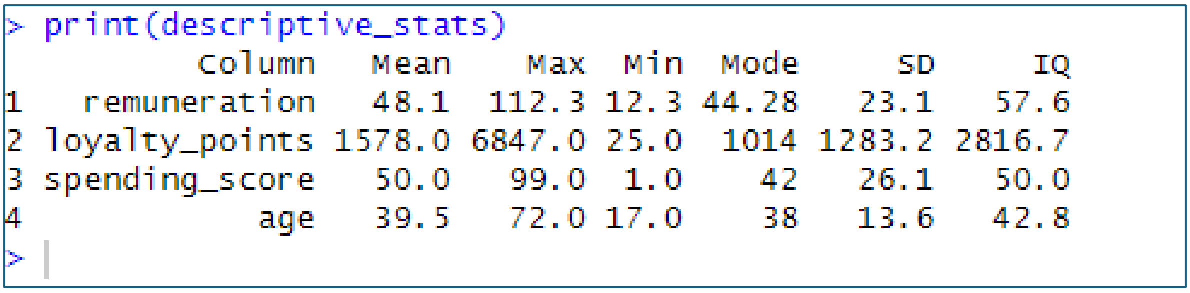
According to Metafile, products are assigned a unique code based on their description. There are 200 unique products registered.



Youko Fujino – LSE | Data Analytics Career Accelerator
DA301_Assignment
22.04.2024

MODULE 6

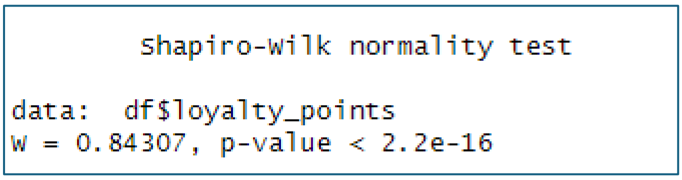
DESCRIPTIVE STATISTICS WITH R: Details in Appendix ^{C}



```
> print(descriptive_stats)
          Column   Mean    Max  Min  Mode     SD     IQ
1   remuneration   48.1  112.3 12.3 44.28   23.1   57.6
2 loyalty_points 1578.0 6847.0 25.0  1014 1283.2 2816.7
3 spending_score   50.0   99.0  1.0    42   26.1   50.0
4            age   39.5   72.0 17.0    38   13.6   42.8
> 
```

LOYALTY POINTS

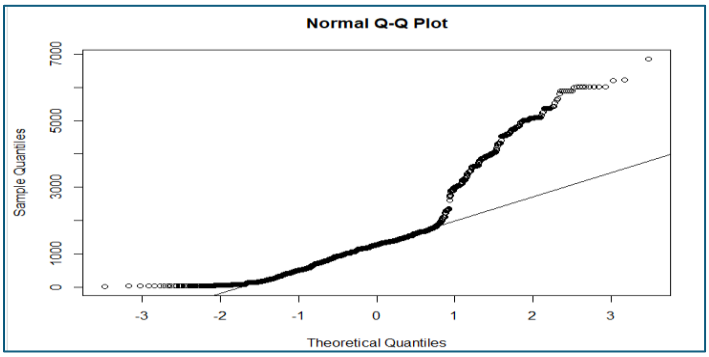
- The Shapiro Test



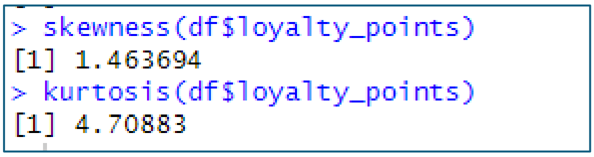
```
        Shapiro-Wilk normality test

data:  df$loyalty_points
W = 0.84307, p-value < 2.2e-16
```

The very small p-value < 2.2e-16 invalidates the null hypothesis of normality.

- Normal QQ Plot for Loyalty Points



- Skewness & Kurtosis

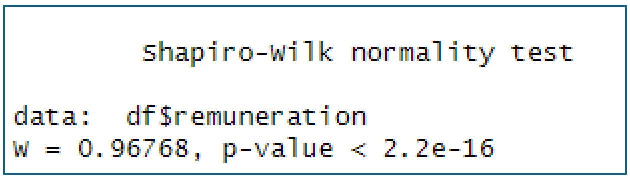


```
> skewness(df$loyalty_points)
[1] 1.463694
> kurtosis(df$loyalty_points)
[1] 4.70883
```

a. The strong positive skewness of 1.46 indicates a significant right skewness, suggesting that the distribution is stretched more towards the higher values.
b. The kurtosis of 4.71 indicates a high degree of peakedness and heavy tails, meaning there are more extreme values (outliers) than a normal distribution.

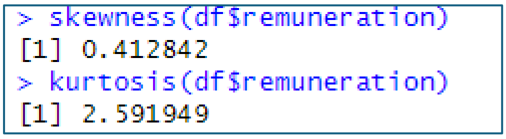
REMUNERATION

The Shapiro Test: The p-value < 2.2e-16 invalidates the null hypothesis of normality



```
        Shapiro-Wilk normality test

data:  df$remuneration
W = 0.96768, p-value < 2.2e-16
```

- Skewness & Kurtosis

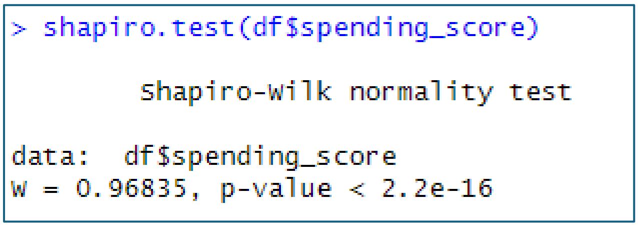


```
> skewness(df$remuneration)
[1] 0.412842
> kurtosis(df$remuneration)
[1] 2.591949
```

- Positive skewness: distribution has a longer right tail, implying that some individuals have relatively high incomes, pulling the distribution towards the right.
- The kurtosis (2.59) indicates that the distribution is leptokurtic: heavier tails and a sharper peak than normal distribution.

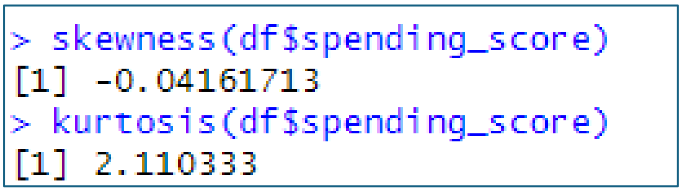
SPENDING SCORE

- The Shapiro Test : low p-level, invaliding the null hypothesis of normality



```
> shapiro.test(df$spending_score)

        Shapiro-Wilk normality test

data:  df$spending_score
W = 0.96835, p-value < 2.2e-16
```

- Skewness & Kurtosis



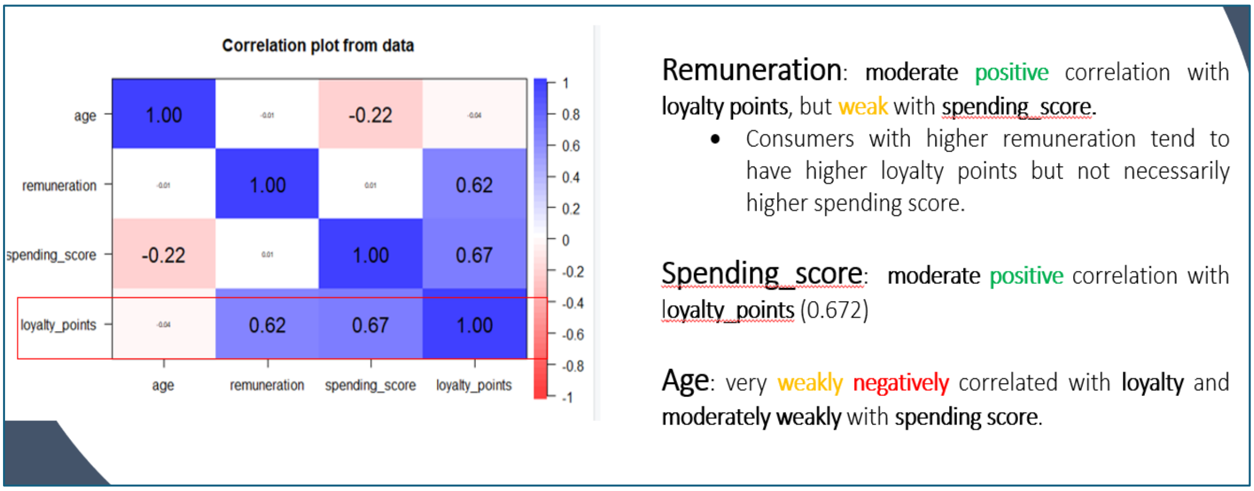
```
> skewness(df$spending_score)
[1] -0.04161713
> kurtosis(df$spending_score)
[1] 2.110333
```

The spending_score data has a slightly left-skewed distribution with a moderate degree of peakedness and heavier tails, but less pronounced than in a leptokurtic distribution. Despite a slight deviation from normality, the distribution is relatively close with minor deviations.

BUILDING AN MLR TO PREDICT LOYALTY POINT

VARIABLES

- Age: very weakly negatively correlated with loyalty and moderately weakly with spending
- Remuneration: moderate positive correlation with loyalty points, but weak with spending_score, implying that individuals with higher remuneration tend to have higher loyalty points but not necessarily the same uplift in spending_score.
- Spending_score: moderate positive correlation with loyalty_points (0.672).



Correlation plot from data

	age	remuneration	spending_score	loyalty_points
age	1.00	-0.01	-0.22	-0.04
remuneration	-0.01	1.00	0.01	0.62
spending_score	-0.22	0.01	1.00	0.67
loyalty_points	-0.04	0.62	0.67	1.00

Remuneration: moderate positive correlation with loyalty points, but weak with spending_score.

- Consumers with higher remuneration tend to have higher loyalty points but not necessarily higher spending score.

Spending_score: moderate positive correlation with loyalty_points (0.672)

Age: very weakly negatively correlated with loyalty and moderately weakly with spending score.

CREATING A MODEL TO PREDICT LOYALTY POINTS

MODEL A = REMUNERATION + SPENDING_SCORE

Syntax R (appendix)[D]

- Remuneration and spending score are significant predictors of loyalty points ($p < 0.001$).
- For every unit increase in remuneration, loyalty_points increase by 34 units.
- For every unit increase in spending_score, loyalty_points increase by 33 units.
- The model explains 82.7% of the variance in loyalty points.

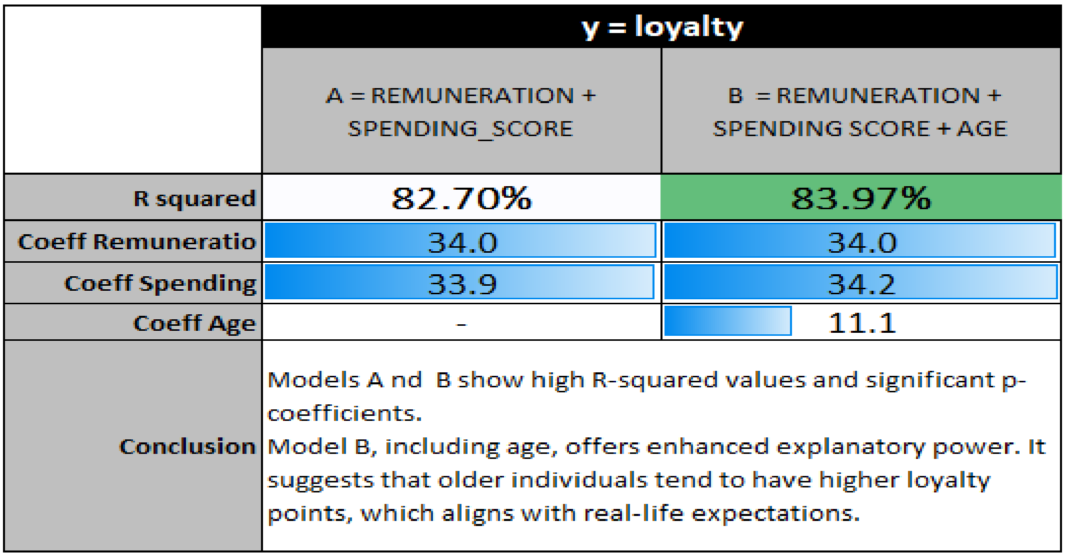
MODEL B = REMUNERATION + SPENDING SCORE + AGE

Syntax R (appendix)[E]

- All predictors are significant ($p < 0.001$) in predicting loyalty_points.
- For every unit increase in remuneration, loyalty_points increase by 34 units.
- For every unit increase in spending_score, loyalty_points increase by 34 units.
- For every year increase in age, loyalty_points increase by 11 units
- The model explains 83.97% of the variance in loyalty_points.

MODEL A OR MODEL B?

Models A and B show high R-squared values and significant p-coefficients.
Model B, including age, offers enhanced explanatory power. It suggests that older individuals tend to have higher loyalty points, which aligns with real-life expectations.



	y = loyalty	
	A = REMUNERATION + SPENDING_SCORE	B = REMUNERATION + SPENDING SCORE + AGE
R squared	82.70%	83.97%
Coeff Remuneratio	34.0	34.0
Coeff Spending	33.9	34.2
Coeff Age	-	11.1
Conclusion	Models A nd B show high R-squared values and significant p-coefficients. Model B, including age, offers enhanced explanatory power. It suggests that older individuals tend to have higher loyalty points, which aligns with real-life expectations.	

DEMO: USING THE MODEL B TO PREDICT LOYALTY POINTS

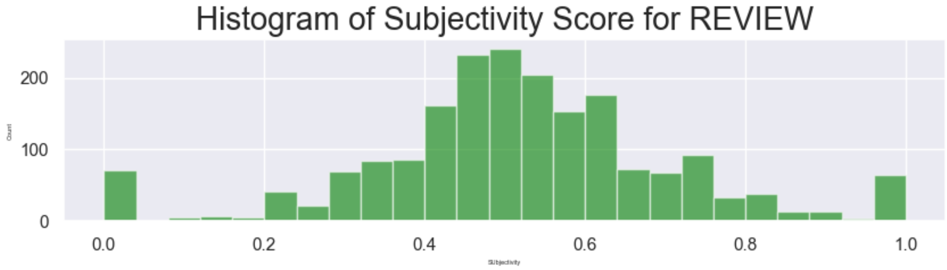
Appendix[F]

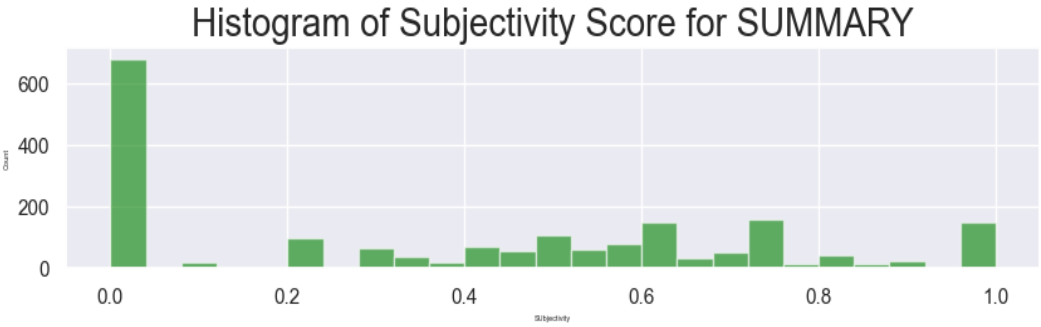
Appendix

[A] STEPS TO BUILD A LINEAR REGRESSION

1) Define the dependent variable and independent variables
2) Specify the linear regression model (lm), test the model and predict the training dataset
3) Evaluate the goodness of fit of the model, obtaining the R-squared, intercept and coefficient value
4) Plot the regression

[B] SUBJECTIVITY HISTOGRAMS





^{C} DESCRIPTIVE STATISTICS WITH R

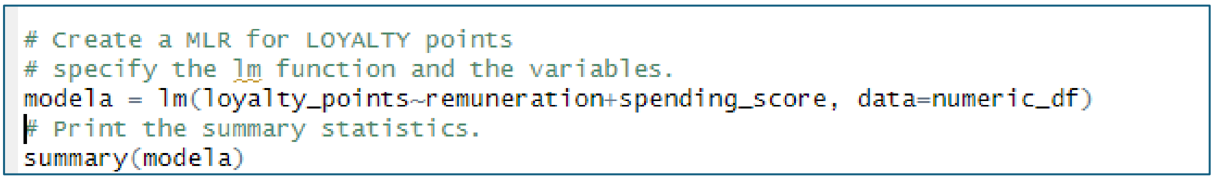
1) Remuneration: the average annual income is $48.1k, with a notable standard deviation of $23.1k, indicating variability. The maximum remuneration is $112.3k, while the minimum is $12.3k, with a wide range of remuneration in the dataset.

2) Loyalty points: Customer loyalty levels vary significantly, with a mean of 1578.0 and a substantial standard deviation of 1283.2.

3) Spending score distribution displays a symmetrical pattern, with a mean of 50.0 and a standard deviation of 26.1, indicating moderate variability in spending behaviour. Notably, the mode of spending score closely aligns with the mean, suggesting a concentration of values around the average spending level.

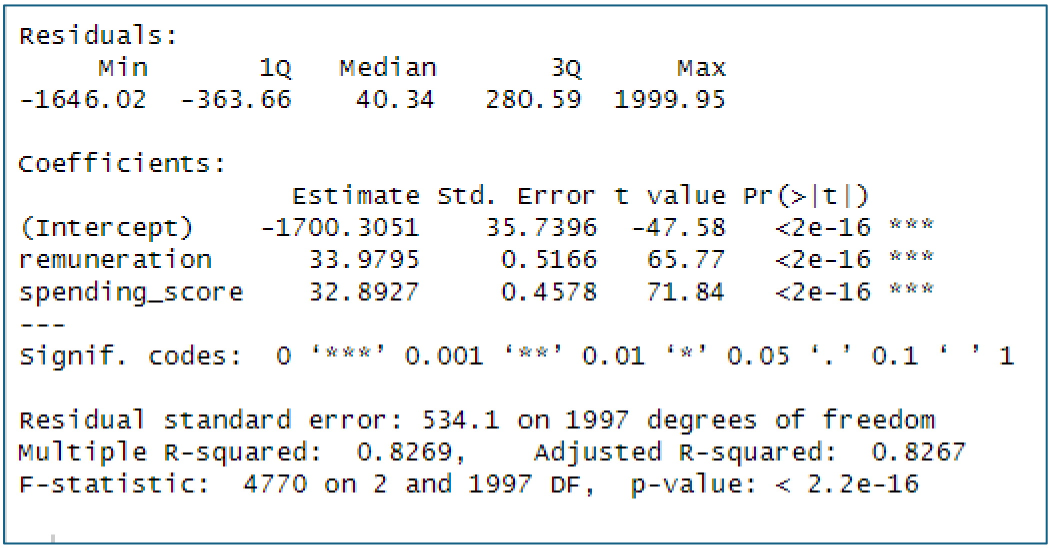
^{D} SYNTEX R – MODEL A

MODEL A = REMUNERATION + SPENDING_SCORE



```
# Create a MLR for LOYALTY points
# specify the lm function and the variables.
modela = lm(loyalty_points~remuneration+spending_score, data=numeric_df)
# Print the summary statistics.
summary(modela)
```

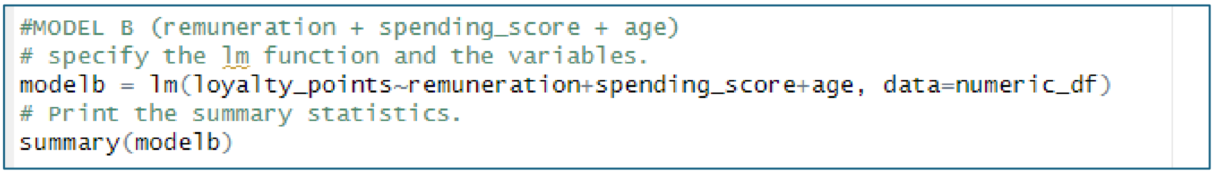
Output



```
Residuals:
     Min       1Q   Median       3Q      Max 
-1646.02  -363.66    40.34   280.59  1999.95 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -1700.3051    35.7396  -47.58   <2e-16 ***
remuneration      33.9795     0.5166   65.77   <2e-16 ***
spending_score    32.8927     0.4578   71.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 534.1 on 1997 degrees of freedom
Multiple R-squared:  0.8269,	Adjusted R-squared:  0.8267 
F-statistic:  4770 on 2 and 1997 DF,  p-value: < 2.2e-16
```

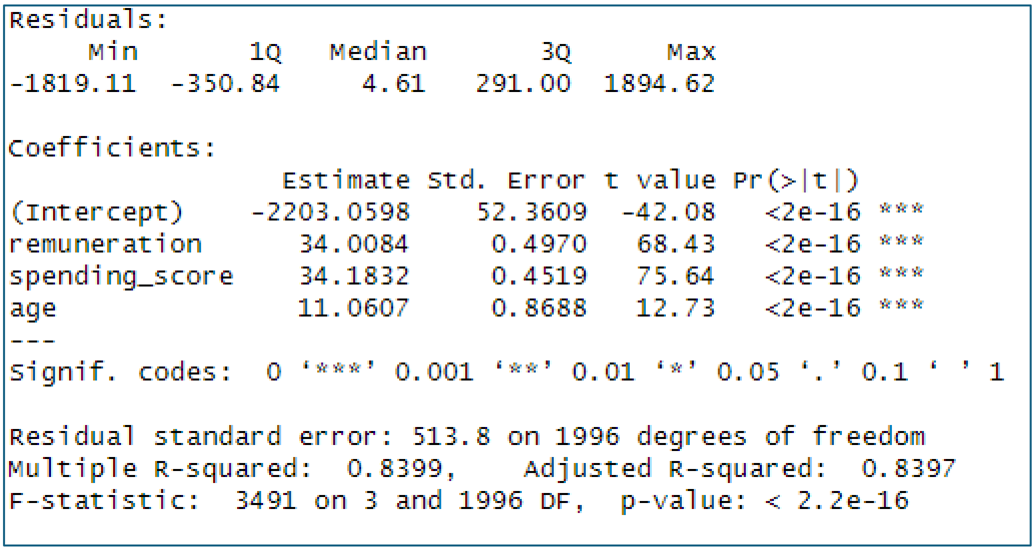
^{E} SYNTEX R – MODEL B

MODEL B = REMUNERATION + SPENDING SCORE + AGE



```
#MODEL B (remuneration + spending_score + age)
# specify the lm function and the variables.
modelb = lm(loyalty_points~remuneration+spending_score+age, data=numeric_df)
# Print the summary statistics.
summary(modelb)
```

Output



```
Residuals:
     Min       1Q   Median       3Q      Max 
-1819.11  -350.84     4.61   291.00  1894.62 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -2203.0598    52.3609  -42.08   <2e-16 ***
remuneration      34.0084     0.4970   68.43   <2e-16 ***
spending_score    34.1832     0.4519   75.64   <2e-16 ***
age               11.0607     0.8688   12.73   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 513.8 on 1996 degrees of freedom
Multiple R-squared:  0.8399,	Adjusted R-squared:  0.8397 
F-statistic:  3491 on 3 and 1996 DF,  p-value: < 2.2e-16
```

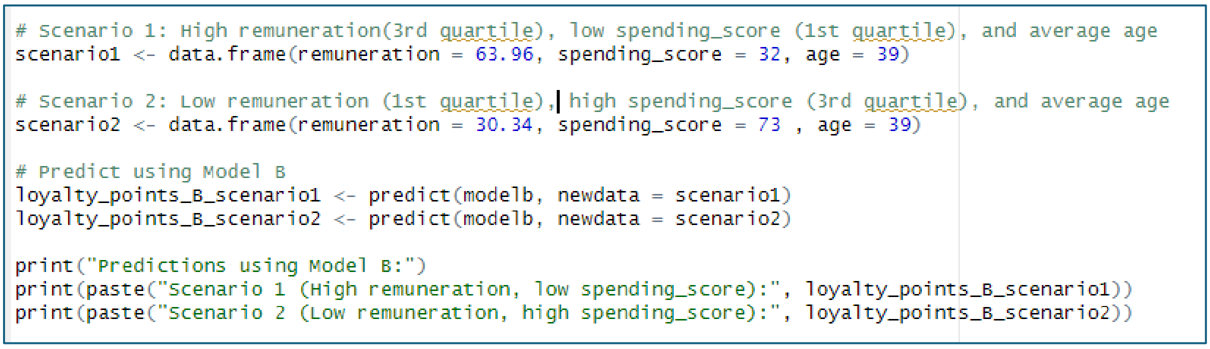
F DEMO: APPLYING MODEL B TO PREDICT LOYALTY POINTS

Scenario 1 (High remuneration, low spending_score, avg age)

The predicted loyalty_points is 1497.34.

Scenario 2 (Low remuneration, high spending_score, avg age)

The predicted loyalty_points is 1755.49.



```
# Scenario 1: High remuneration(3rd quartile), low spending_score (1st quartile), and average age
scenario1 <- data.frame(remuneration = 63.96, spending_score = 32, age = 39)

# Scenario 2: Low remuneration (1st quartile), high spending_score (3rd quartile), and average age
scenario2 <- data.frame(remuneration = 30.34, spending_score = 73 , age = 39)

# Predict using Model B
loyalty_points_B_scenario1 <- predict(modelb, newdata = scenario1)
loyalty_points_B_scenario2 <- predict(modelb, newdata = scenario2)

print("Predictions using Model B:")
print(paste("Scenario 1 (High remuneration, low spending_score):", loyalty_points_B_scenario1))
print(paste("Scenario 2 (Low remuneration, high spending_score):", loyalty_points_B_scenario2))
```