Python for Data Analysis

FINAL PROJECT

DE VINCI HIGHER EDUCATION

ESILV ENGINEERING SCHOOL DE VINCI PARIS

OUR TEAM

Emrys MEZIANI

👤 DIA5 Student

🌐 https://github.com/EmrysMz/Python-project/tree/main

📍 Courbevoie, 92

Sébastien MOINE

👤 DIA5 Student

🌐 https://github.com/EmrysMz/Python-project/tree/main

📍 Courbevoie, 92

SUMMARY

1 PROJECT'S PURPOSE

2 PROBLEMATIC

3 DATASET

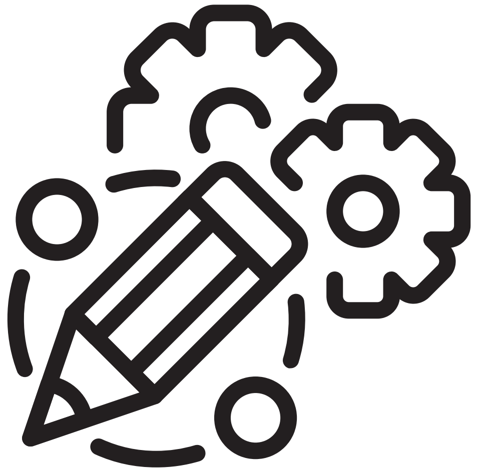
4 PREPROCESSING

5 PREDICTIVE MODELS

6 MAGIC RECIPE

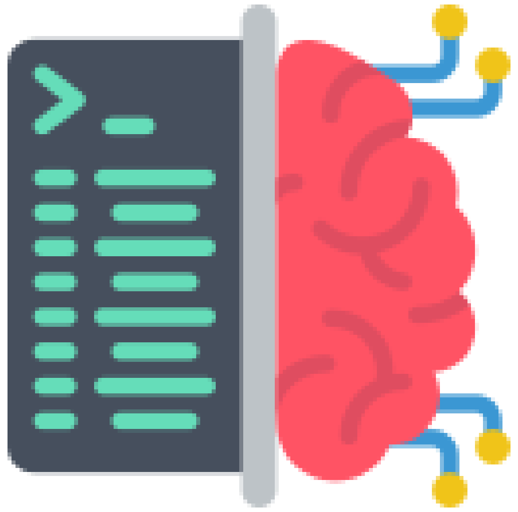
1. Project’s Purpose

Projects works application, and more !
First step into Data Science



Preprocessing

Data Visualisation



Modeling & Optimization

API

2. Problematic

Measuring and predicting the popularity of an article

Which features and how they impact the number of shares

The recipe for the success

3. Dataset

Online News Popularity

Features

61 columns
- 58 predictive
- 1 target (shares)

Input Features' Groups

● Tokens
● Keywords
● References
● Weekdays
● Polarity

● Videos & Images
● Pos/Neg Words Rate
● LDAs & Data Channels
● Subjectivity

Sources

UCI Machine Learning Repository
Released on 05/30/2015
Articles from Mashable (2013-2014)

Characteristics

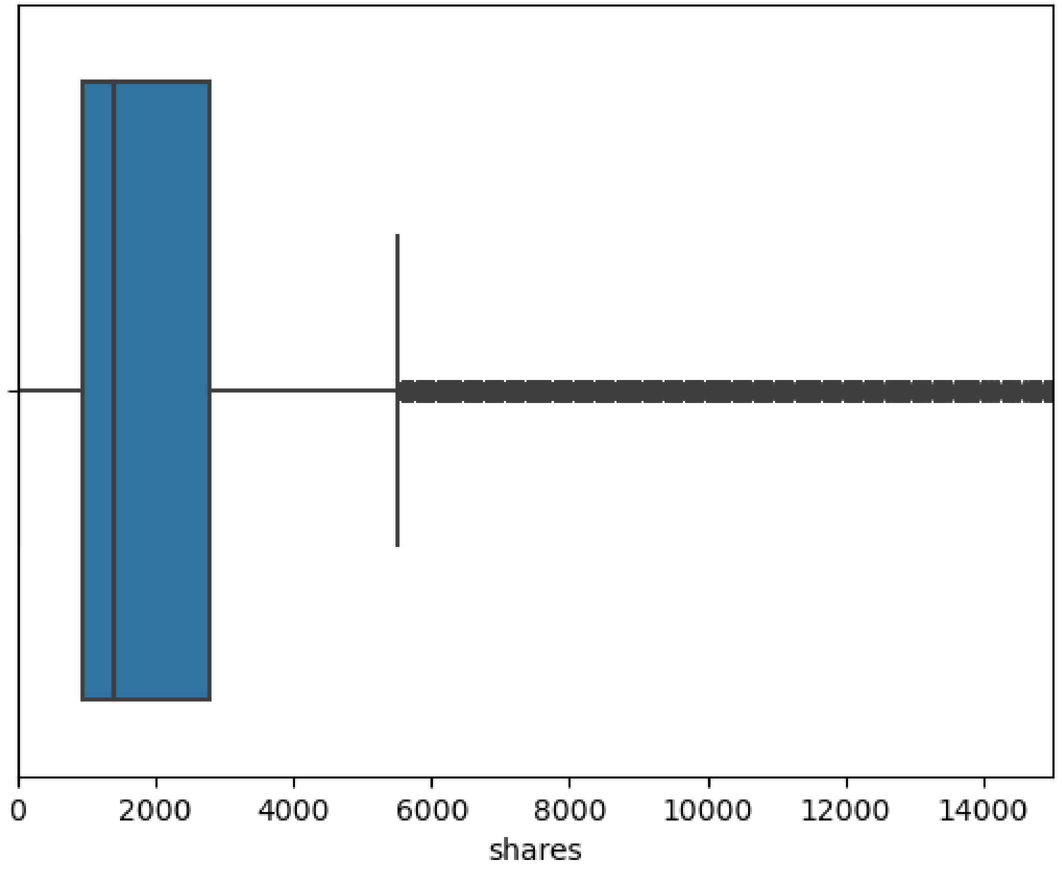
0 NaN values
39 644 rows
float64(59), int64(1), object(1)

4. Preprocessing

Values distributions

- There was a lot of outliers in the dataset, with some extremely high values.

Boxplot of the Shares' distribution



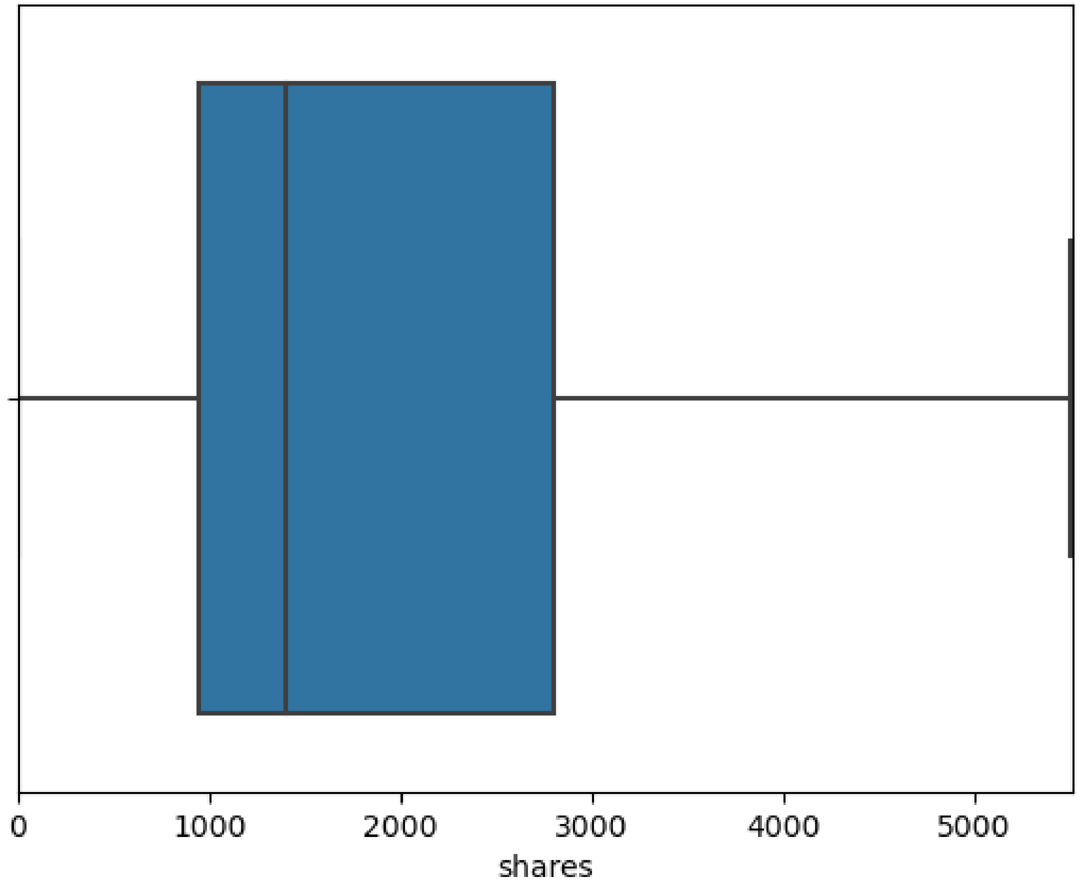
IQR Rule

- Statistic method to find outliers
- IQR = 3rd quartile - 1st quartile
- Upper limit = 3rd quartile + 1,5*IQR
- Lower limit = 1st quartile - 1,5*IQR
- Remove values above upper limit

Target feature : Shares

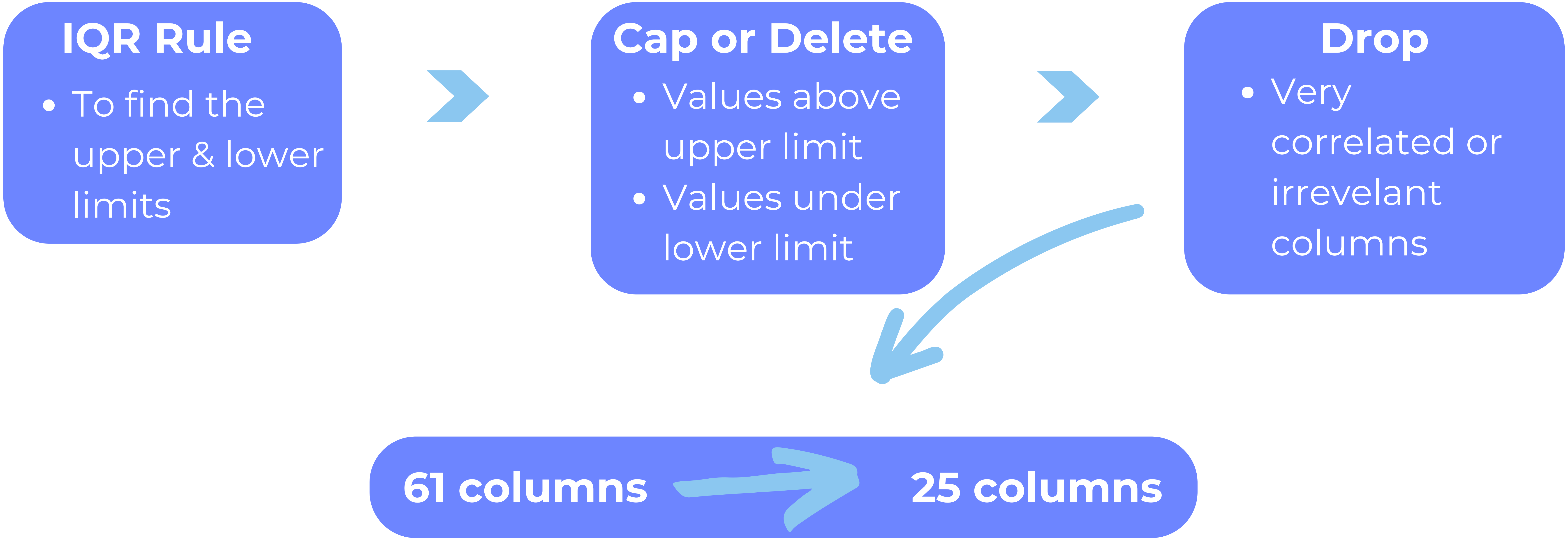
- The column ‘shares’ had many outliers, since it’s the target we needed to fix that.

Without outliers > Quantile 88,55%



4. Preprocessing

For all columns of the dataset :



IQR Rule

- To find the upper & lower limits

Cap or Delete

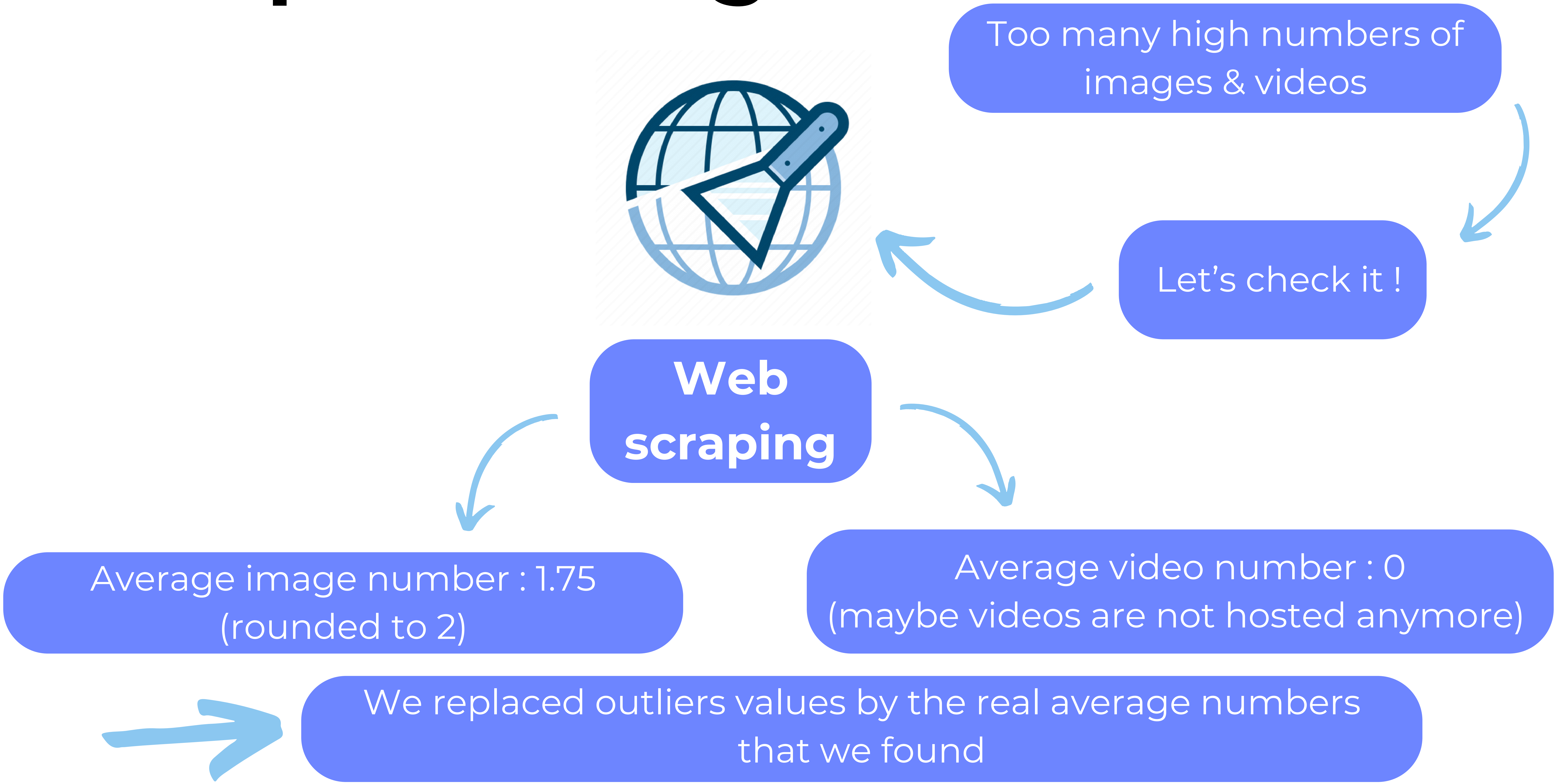
- Values above upper limit
- Values under lower limit

Drop

- Very correlated or irrevelant columns

61 columns → **25 columns**

4. Preprocessing



Too many high numbers of images & videos

Let's check it !

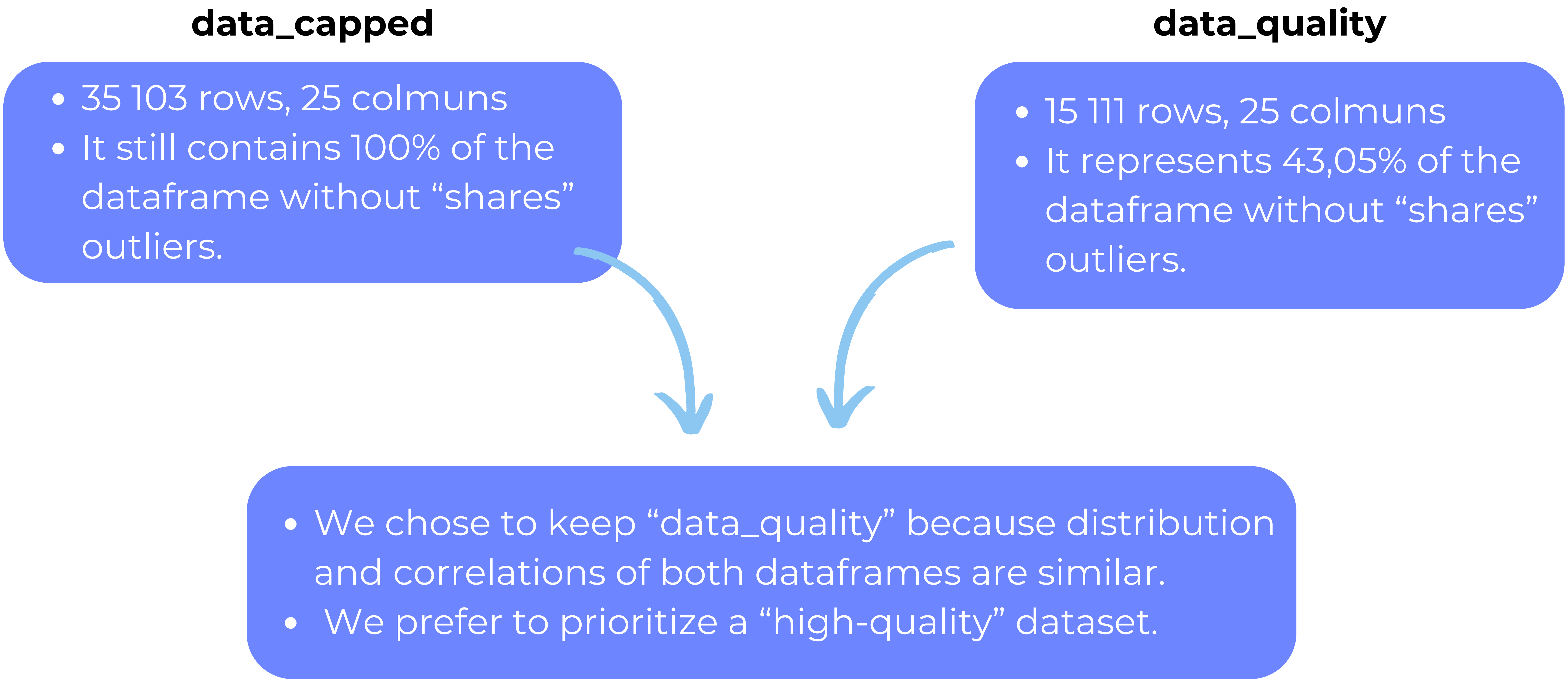
Web scraping

Average image number : 1.75 (rounded to 2)

Average video number : 0 (maybe videos are not hosted anymore)

We replaced outliers values by the real average numbers that we found

4. Preprocessing



data_capped

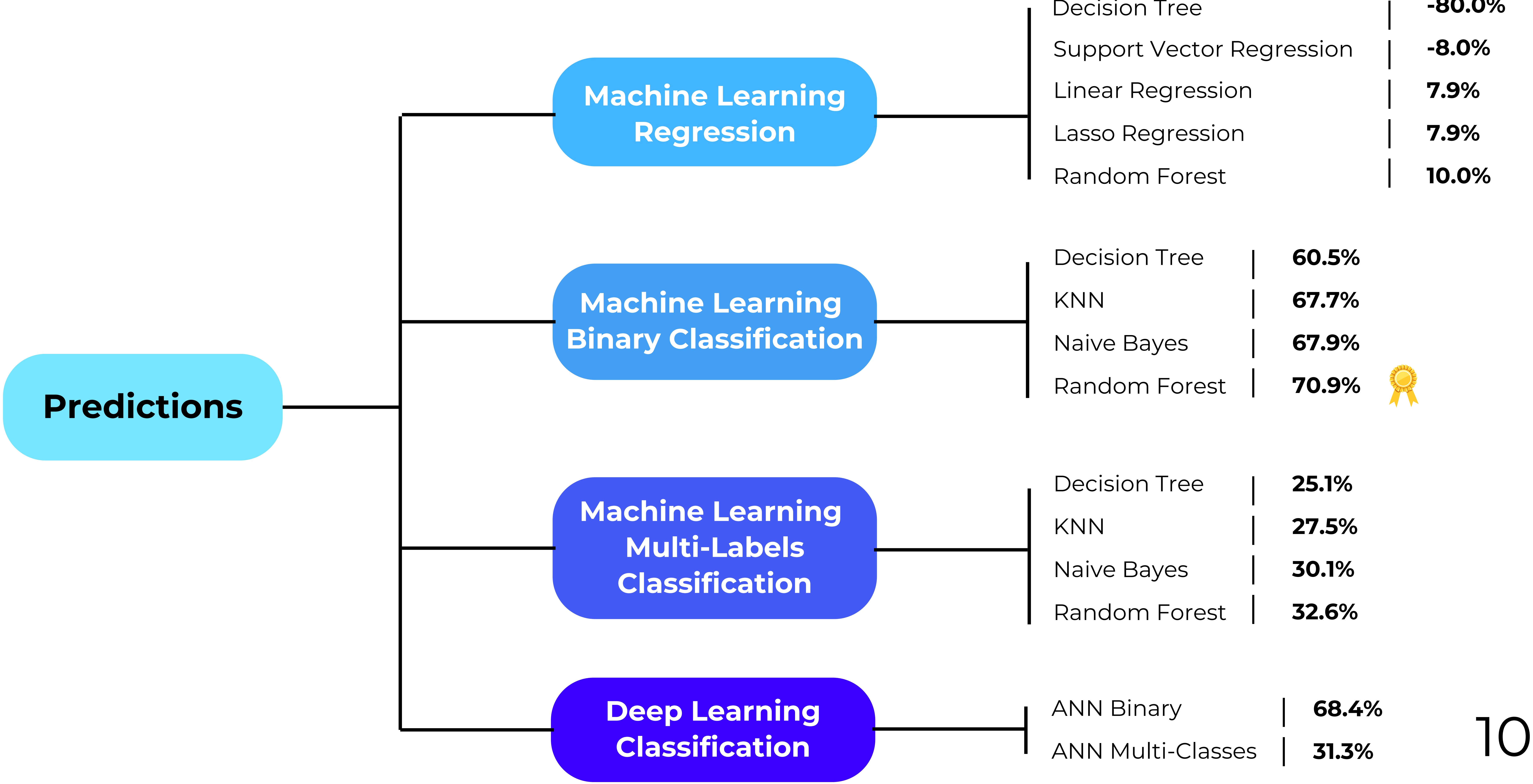
- 35 103 rows, 25 colmuns
- It still contains 100% of the dataframe without “shares” outliers.

data_quality

- 15 111 rows, 25 colmuns
- It represents 43,05% of the dataframe without “shares” outliers.

- We chose to keep “data_quality” because distribution and correlations of both dataframes are similar.
- We prefer to prioritize a “high-quality” dataset.

5. Predictive Models



Predictions

Machine Learning Regression

Decision Tree	-80.0%
Support Vector Regression	-8.0%
Linear Regression	7.9%
Lasso Regression	7.9%
Random Forest	10.0%

Machine Learning Binary Classification

Decision Tree	60.5%	
KNN	67.7%	
Naive Bayes	67.9%	
Random Forest	70.9%	🎖

Machine Learning Multi-Labels Classification

Decision Tree	25.1%
KNN	27.5%
Naive Bayes	30.1%
Random Forest	32.6%

Deep Learning Classification

ANN Binary	68.4%
ANN Multi-Classes	31.3%

5. Predictive Models

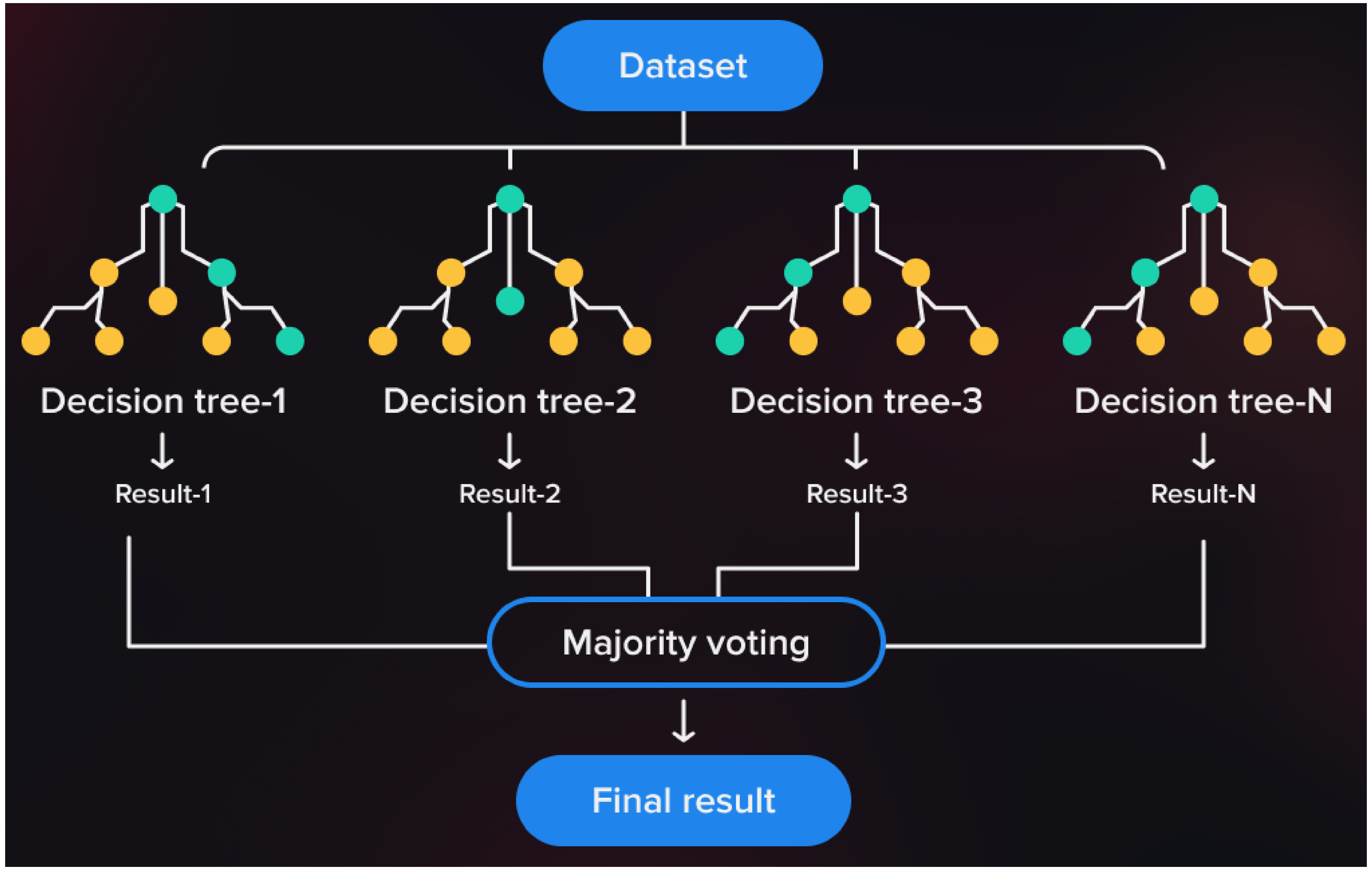
24 inputs

2 classes

Popular

Non-Popular

Best model : Random Forest



Dataset

Decision tree-1

Result-1

Decision tree-2

Result-2

Decision tree-3

Result-3

Decision tree-N

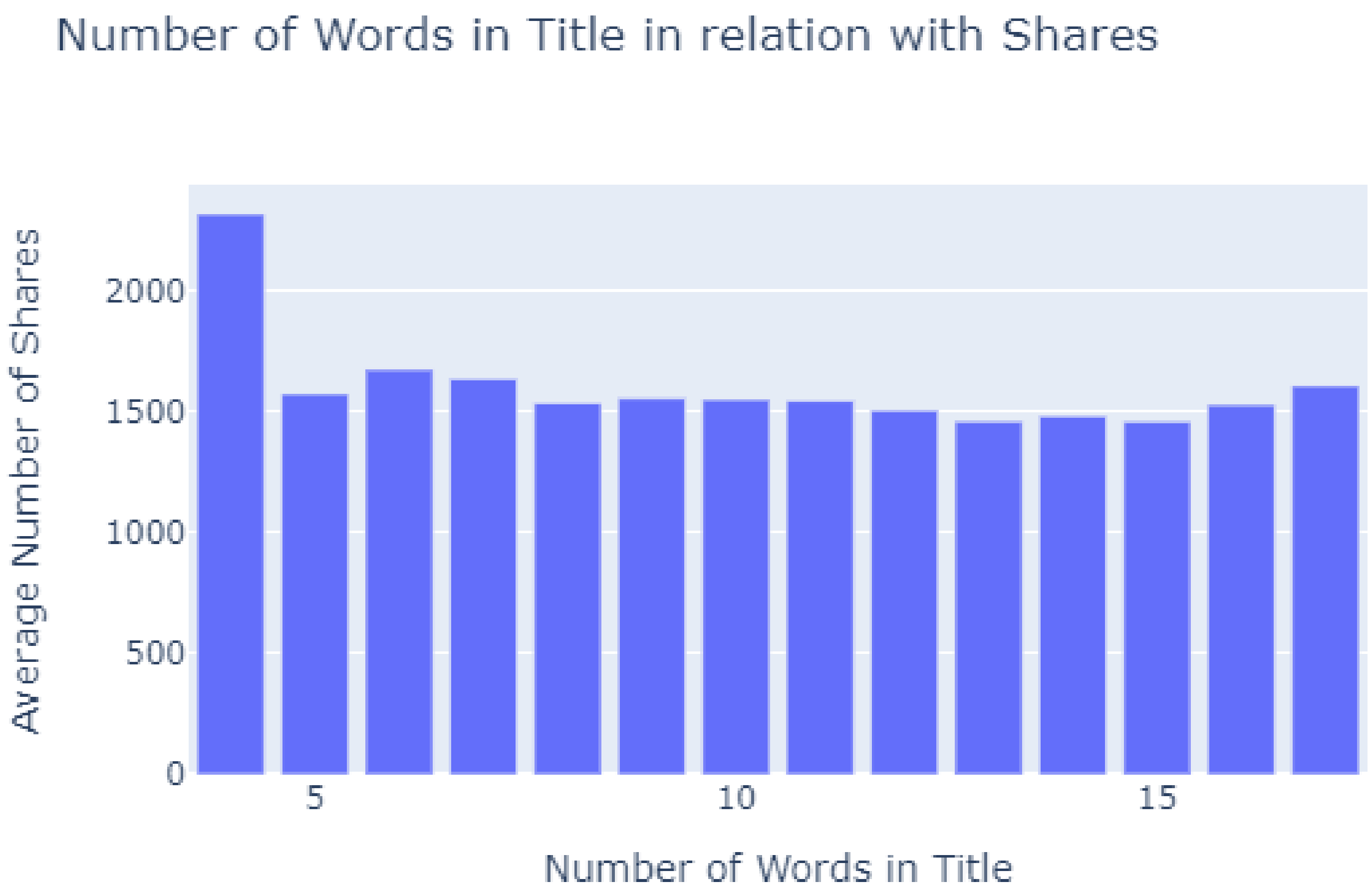
Result-N

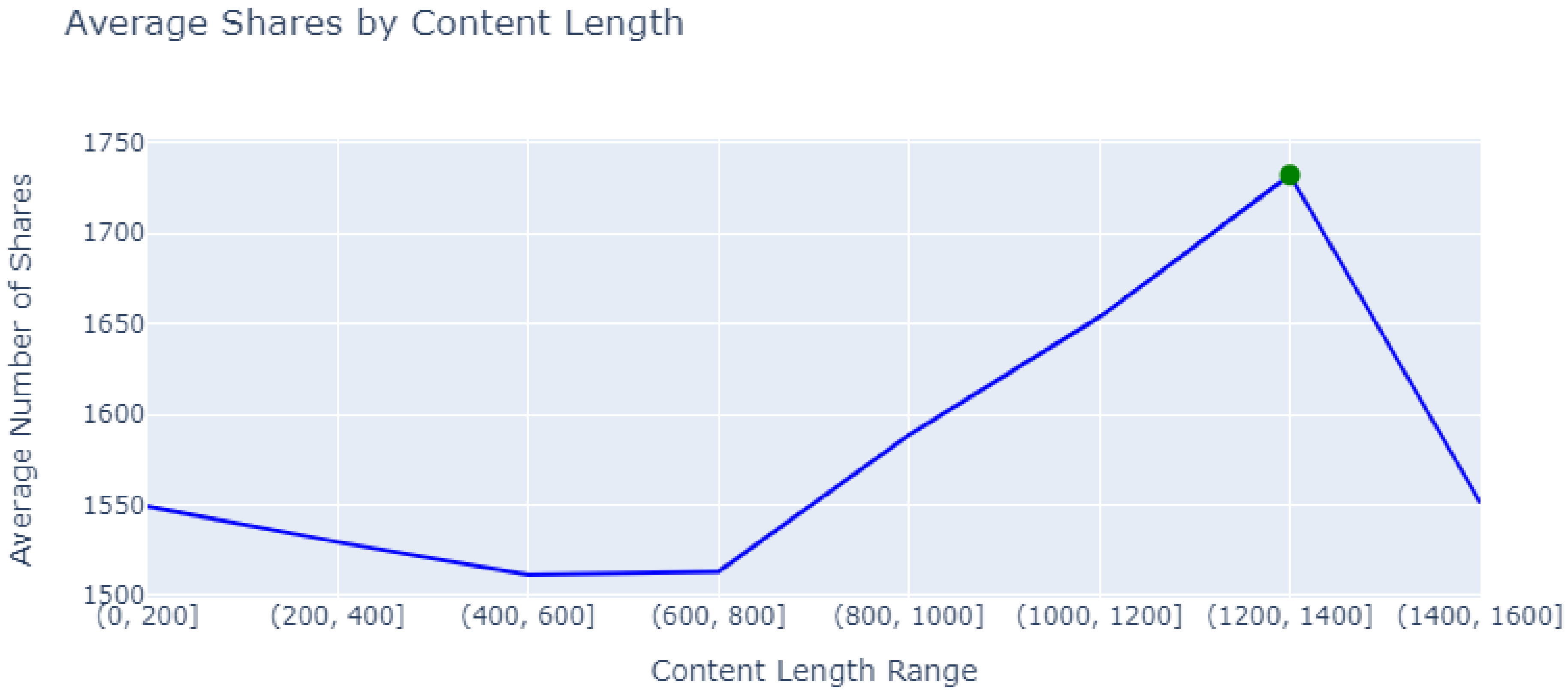
Majority voting

Final result

Serokell.io : Random Forest Classifier: Basic Principles and Applications

6. Magic Recipe

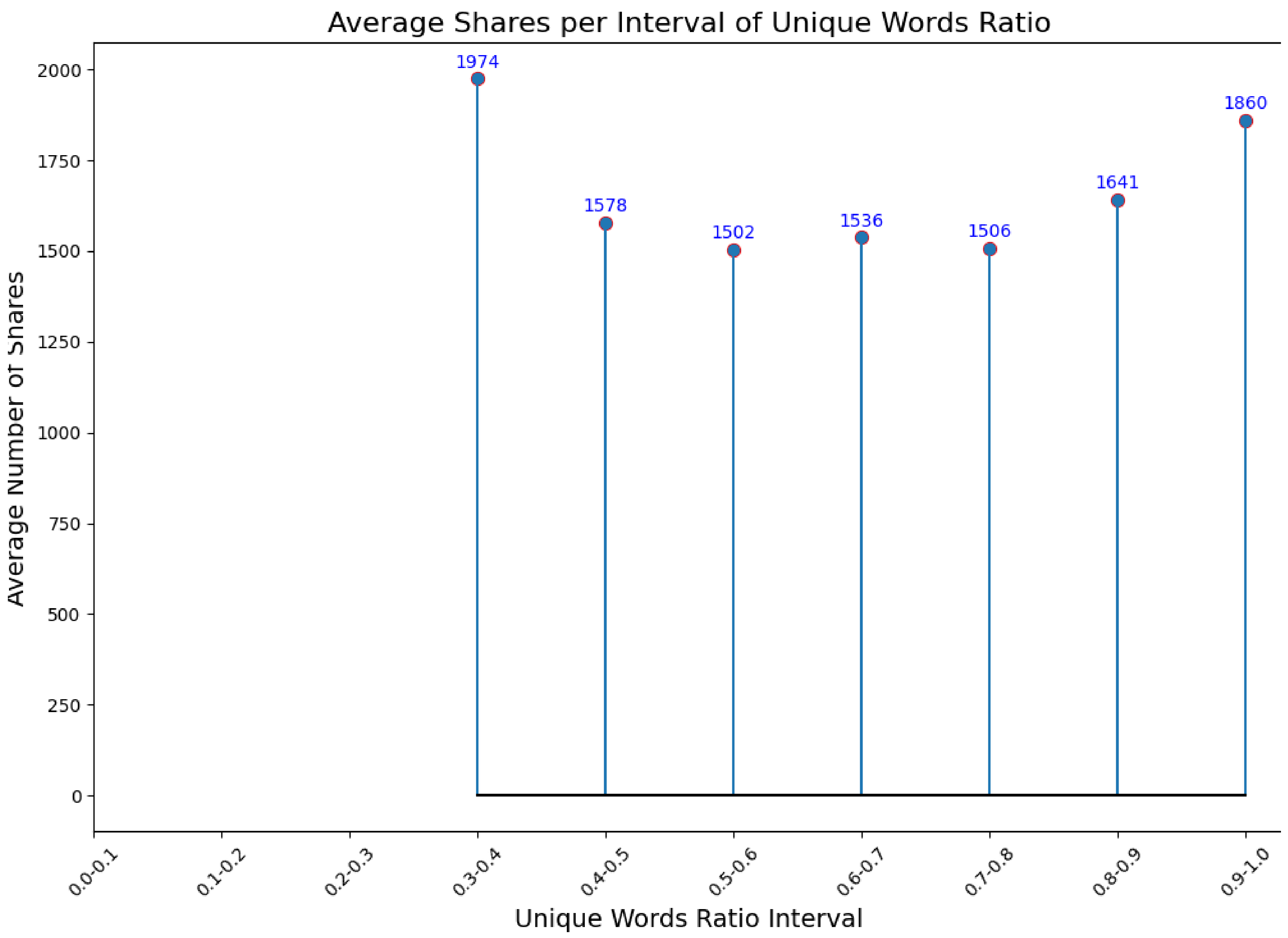


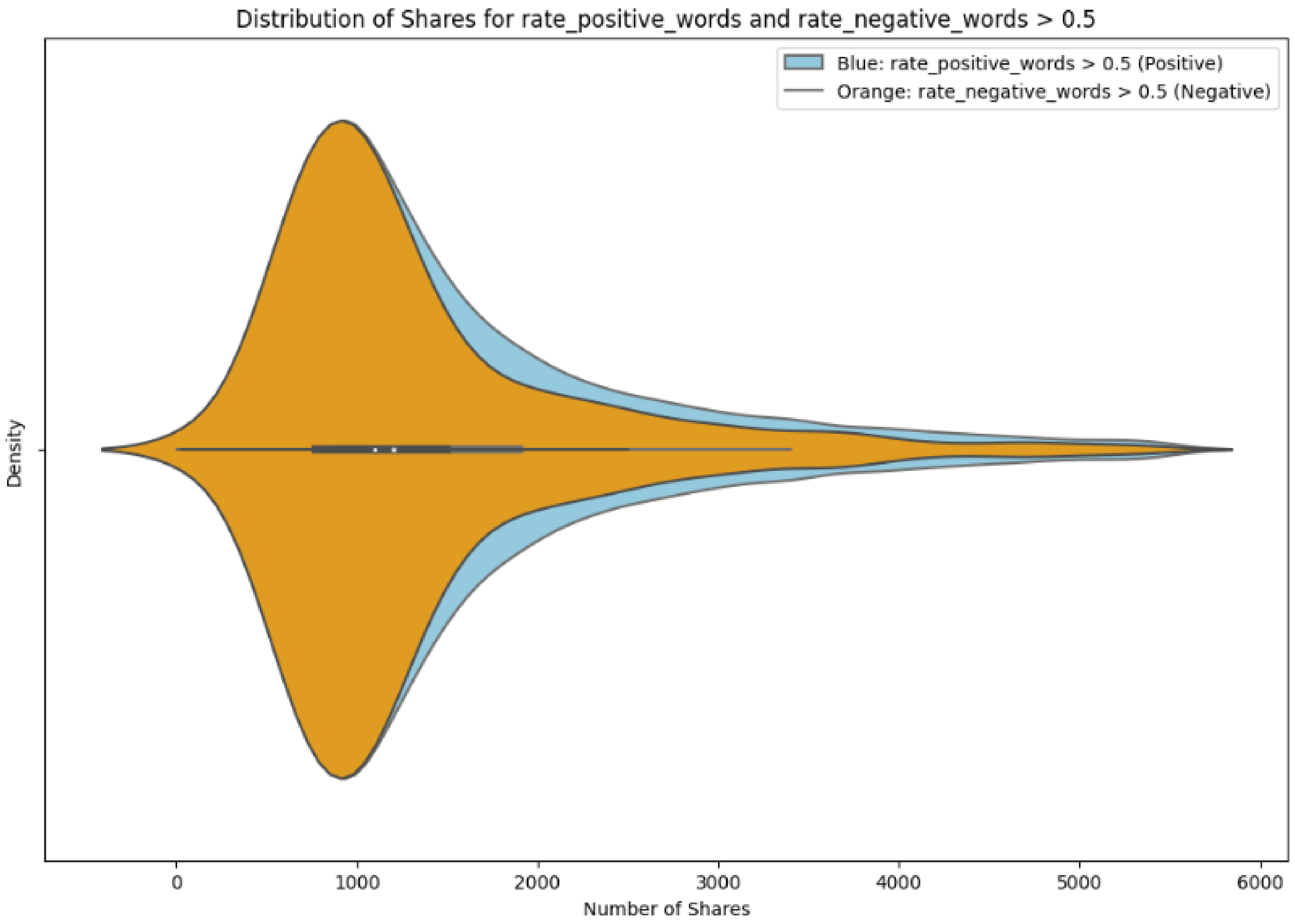


Aim for a concise title with precisely 4 words

Ensure the article has between 1200 and 1400 words

6. Magic Recipe

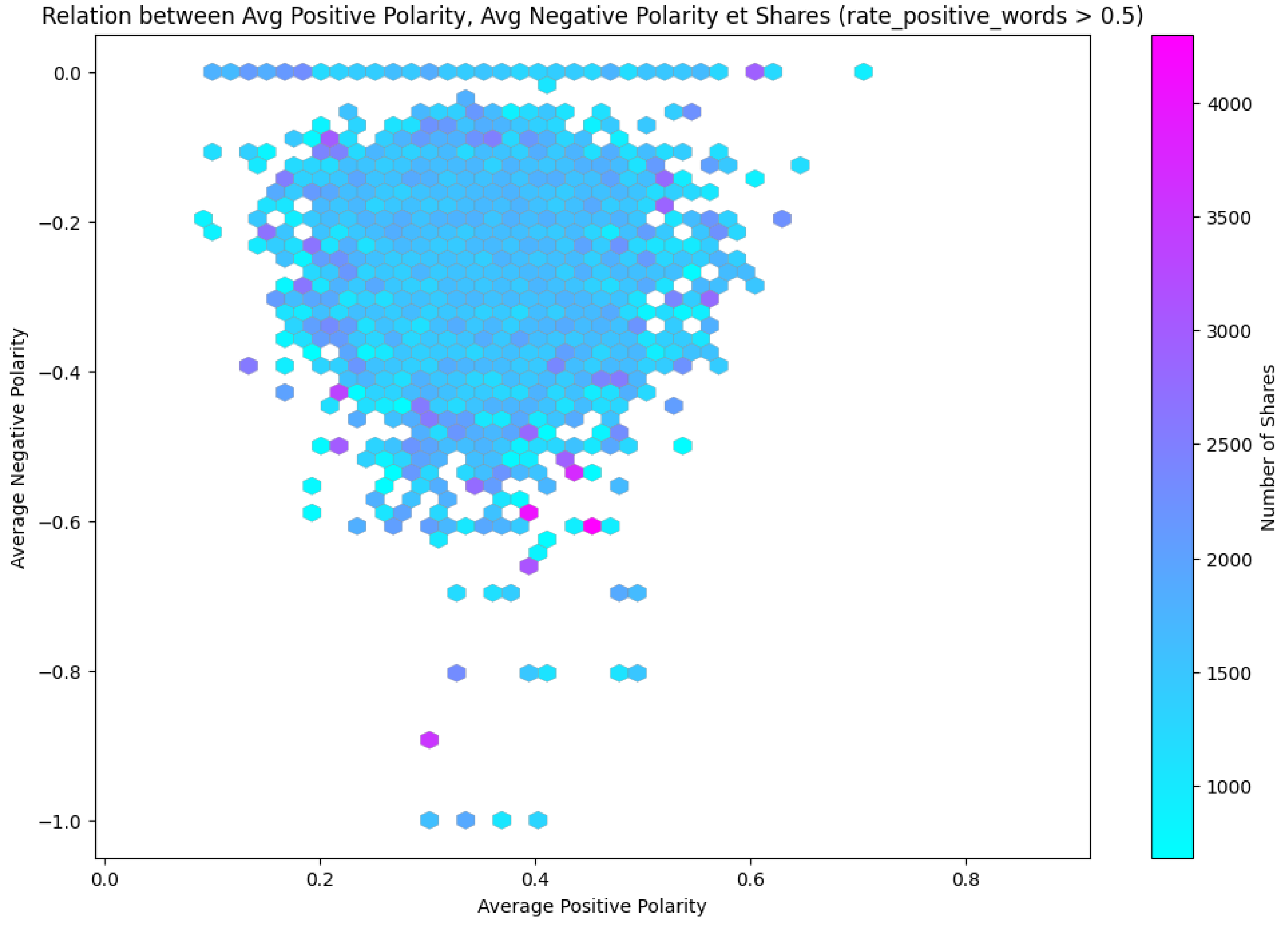


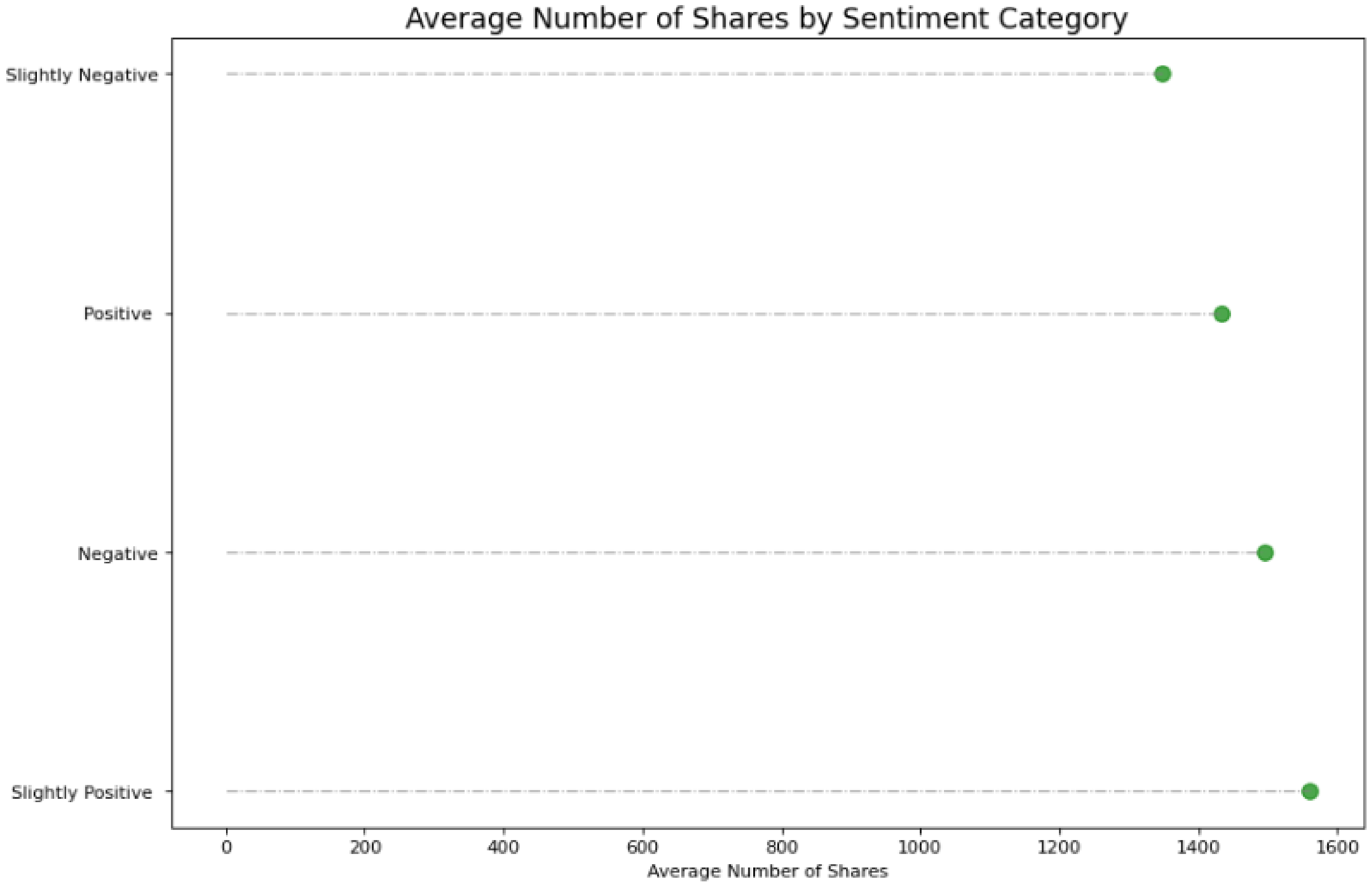


Maintain a unique words ratio between 30% to 40%

Strive for a higher global rate of positive words than negative words

6. Magic Recipe



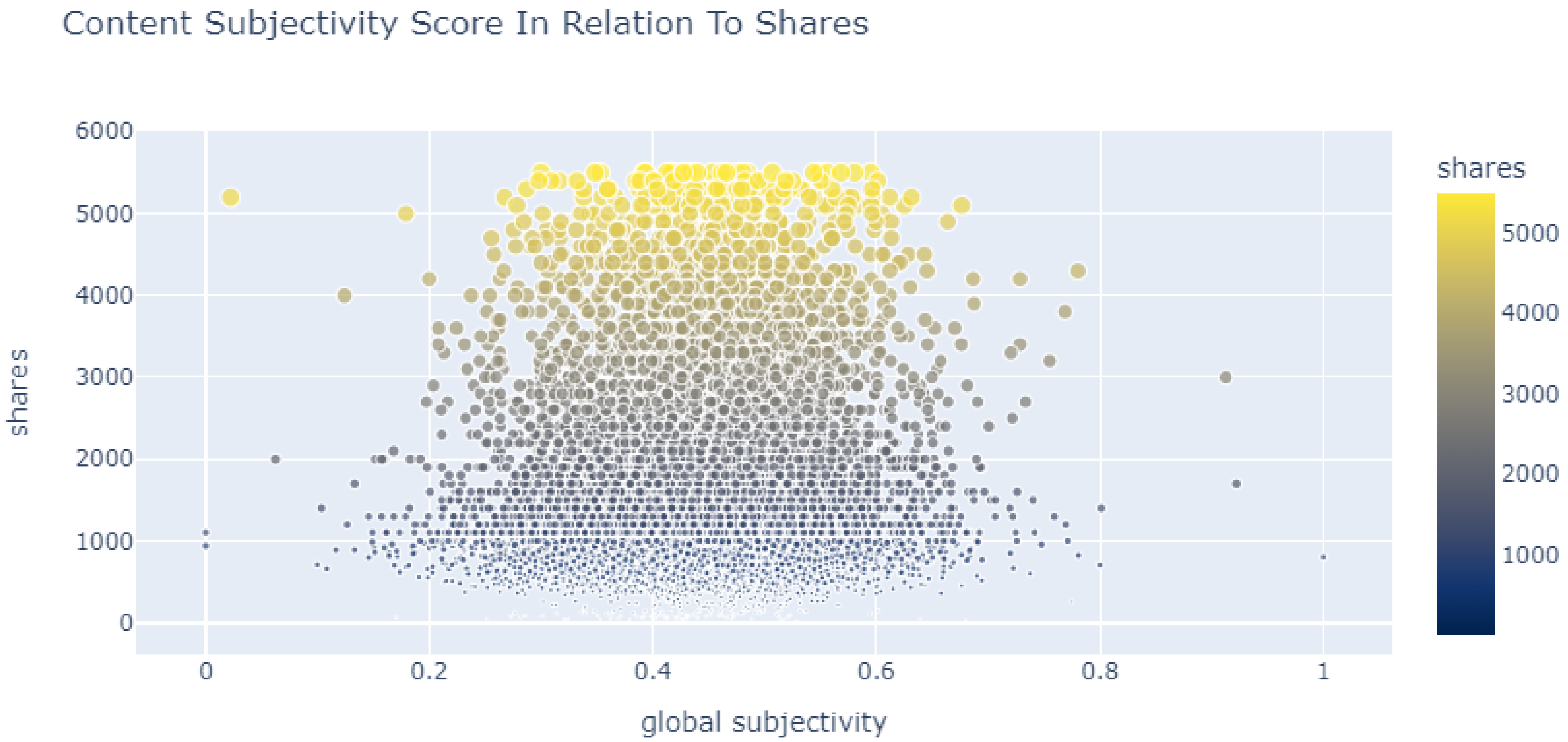


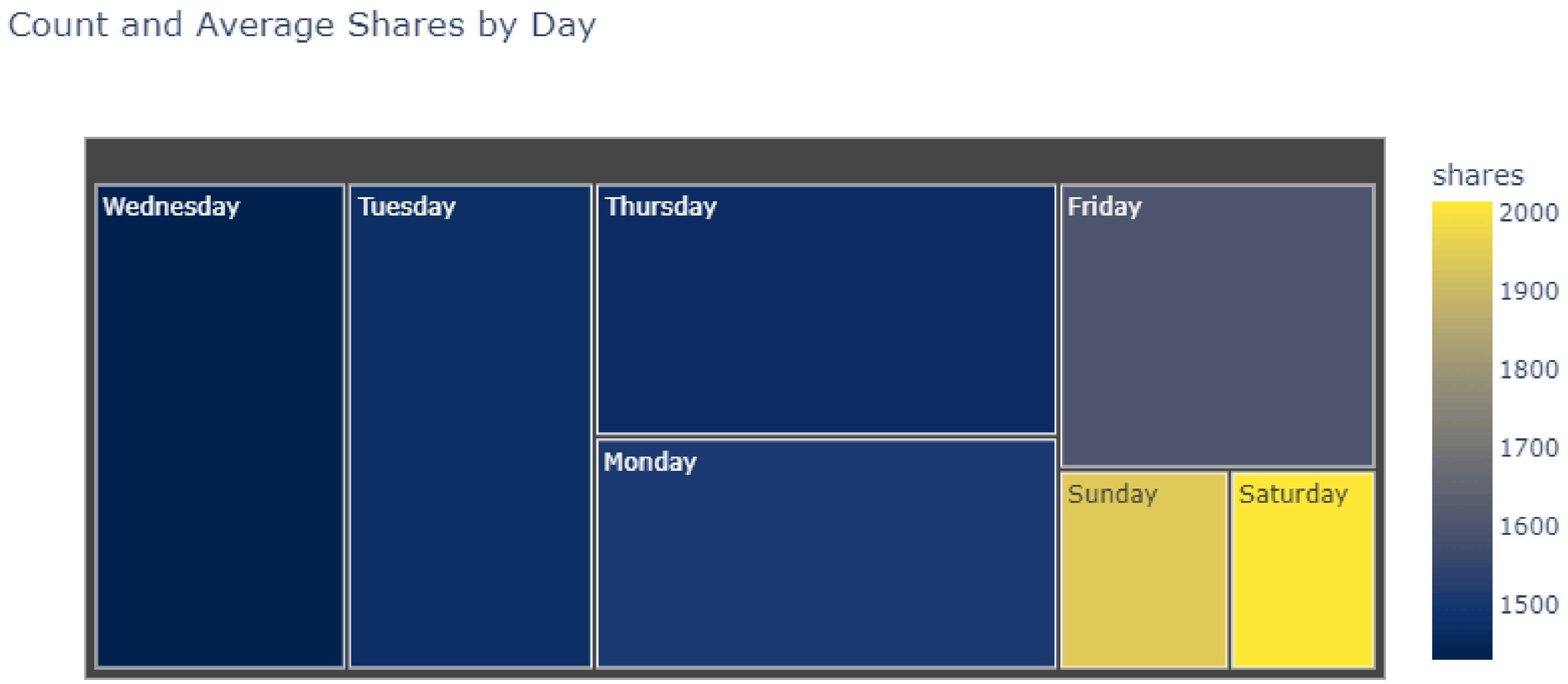
Positive words polarity should range from 20% to 60%

Negative words polarity should range from -10% to -60%

Aim for an overall global sentiment polarity of 0% to 33% positive

6. Magic Recipe

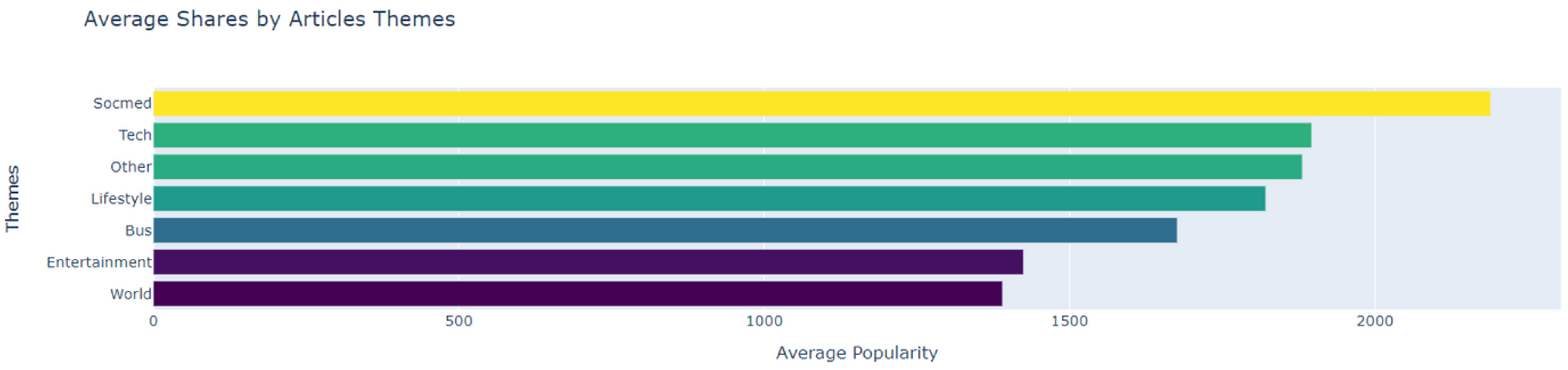

Content Subjectivity Score In Relation To Shares

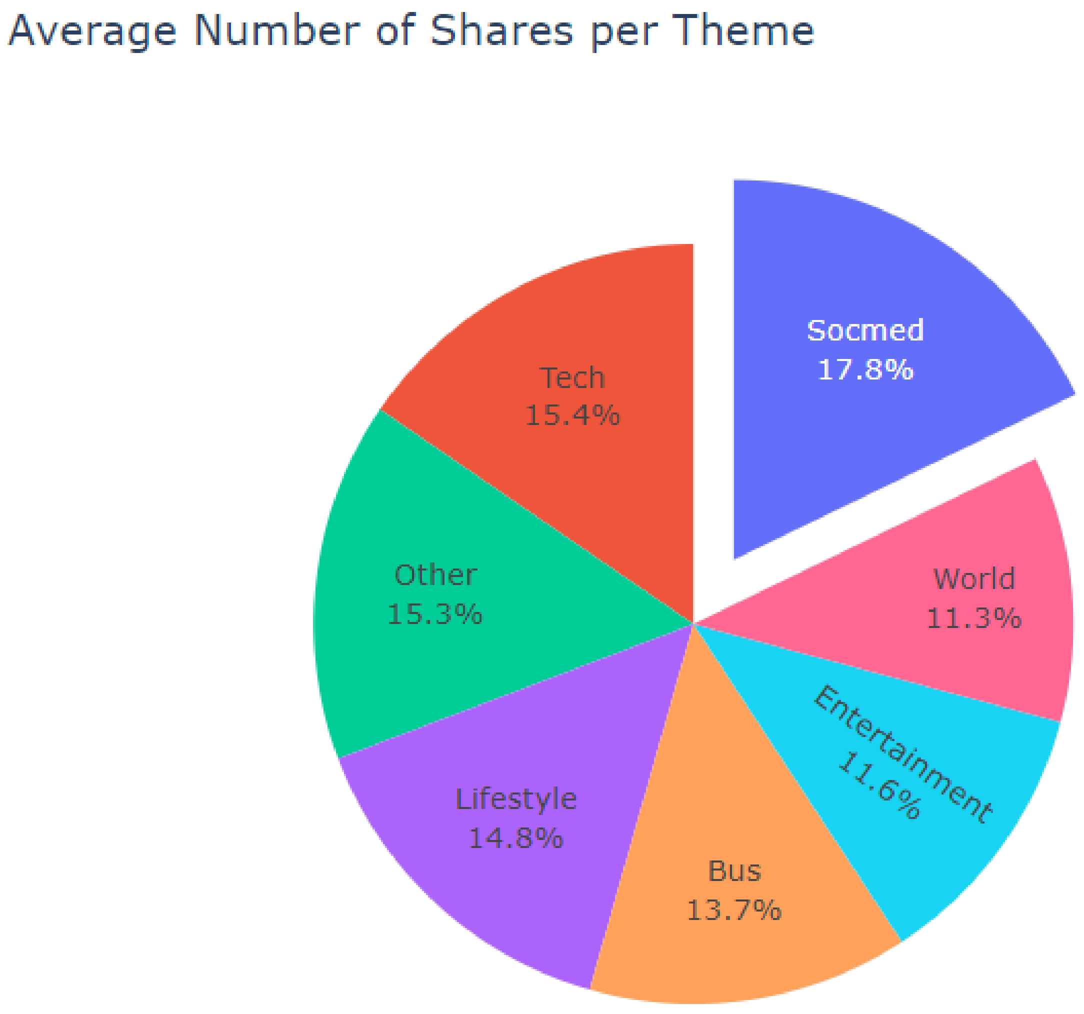

Count and Average Shares by Day

Keep the content subjectivity within the range of 30% to 60%

Optimal publishing day is Saturday

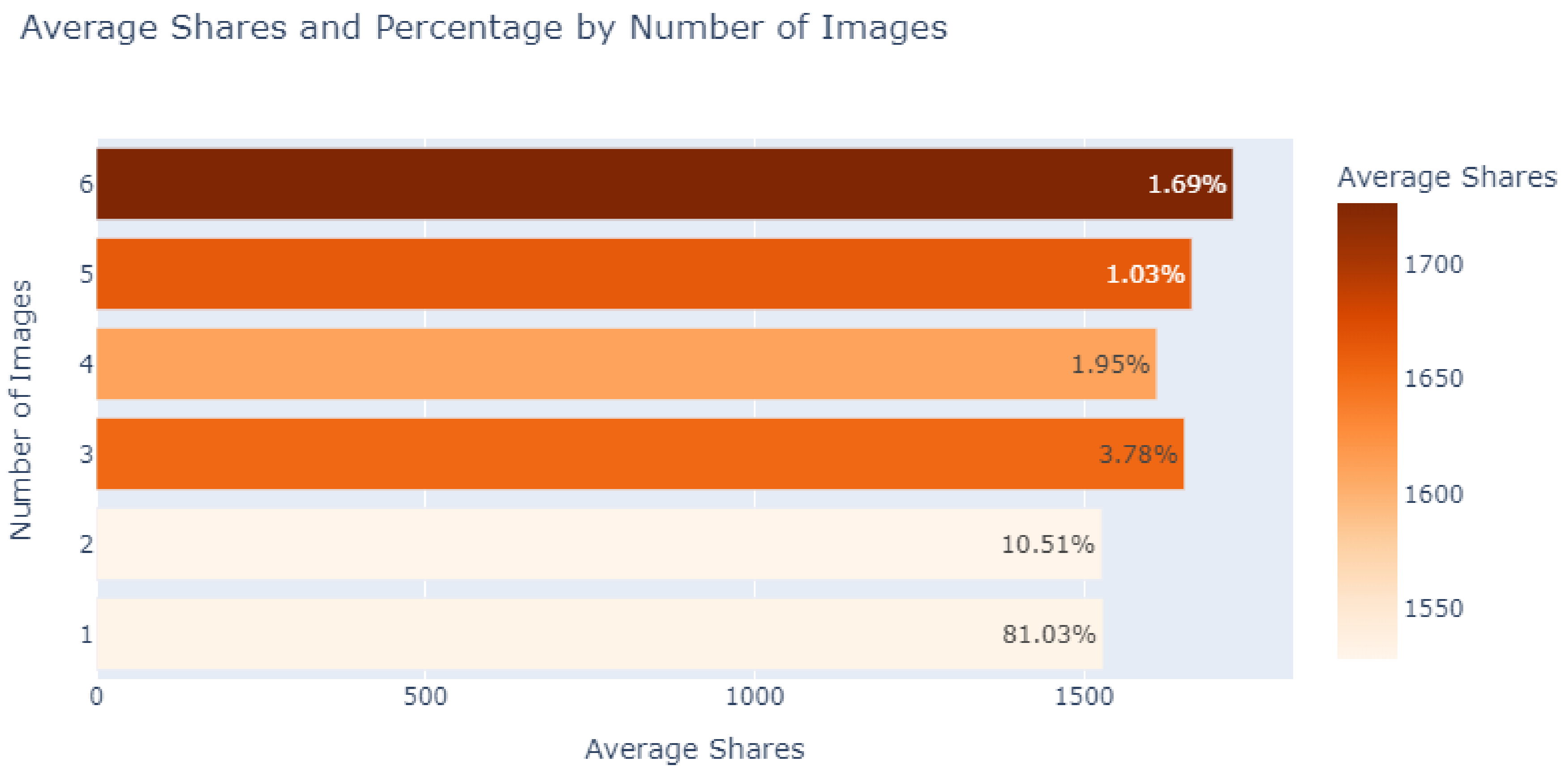
6. Magic Recipe

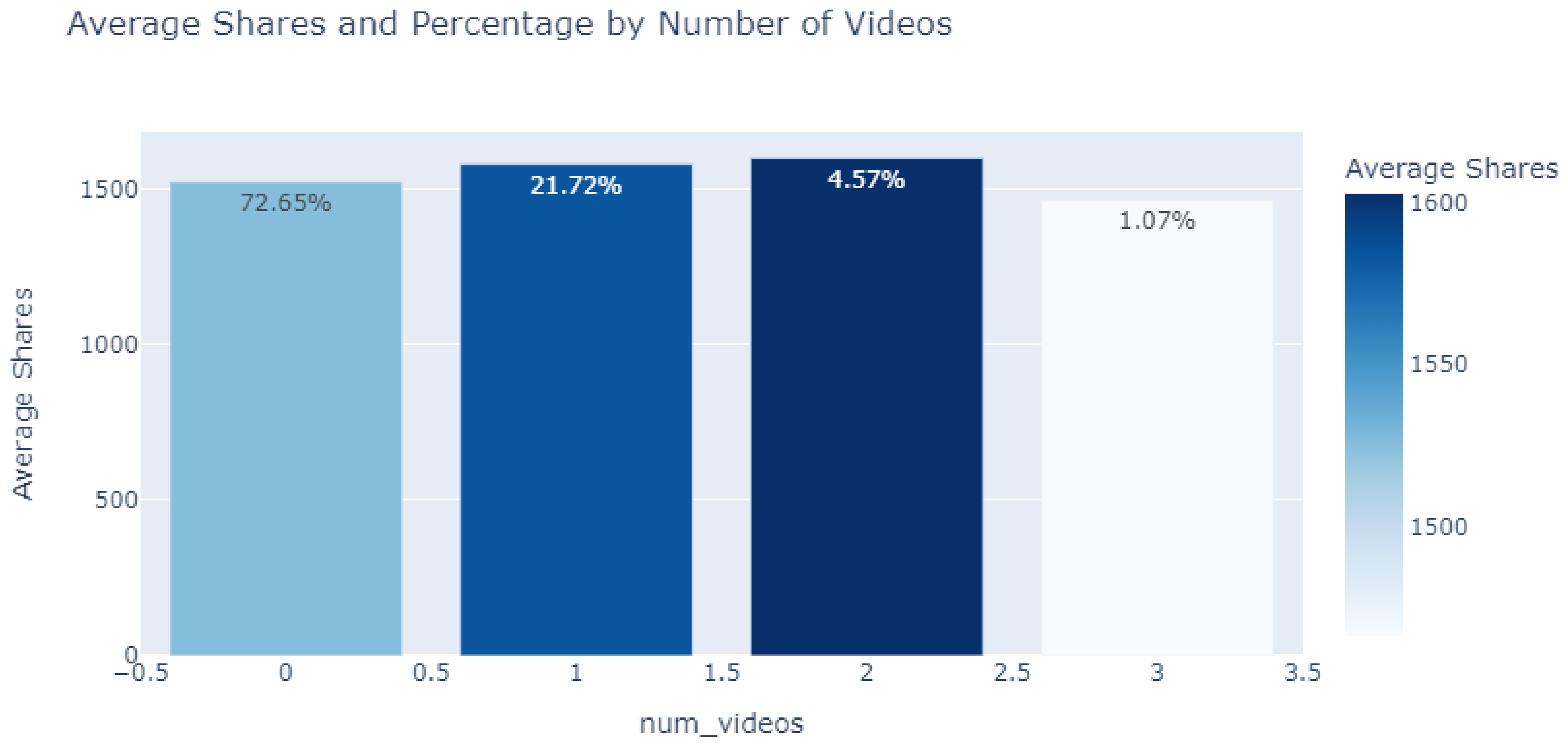




Tailor the article content to revolve around social media.

6. Magic Recipe

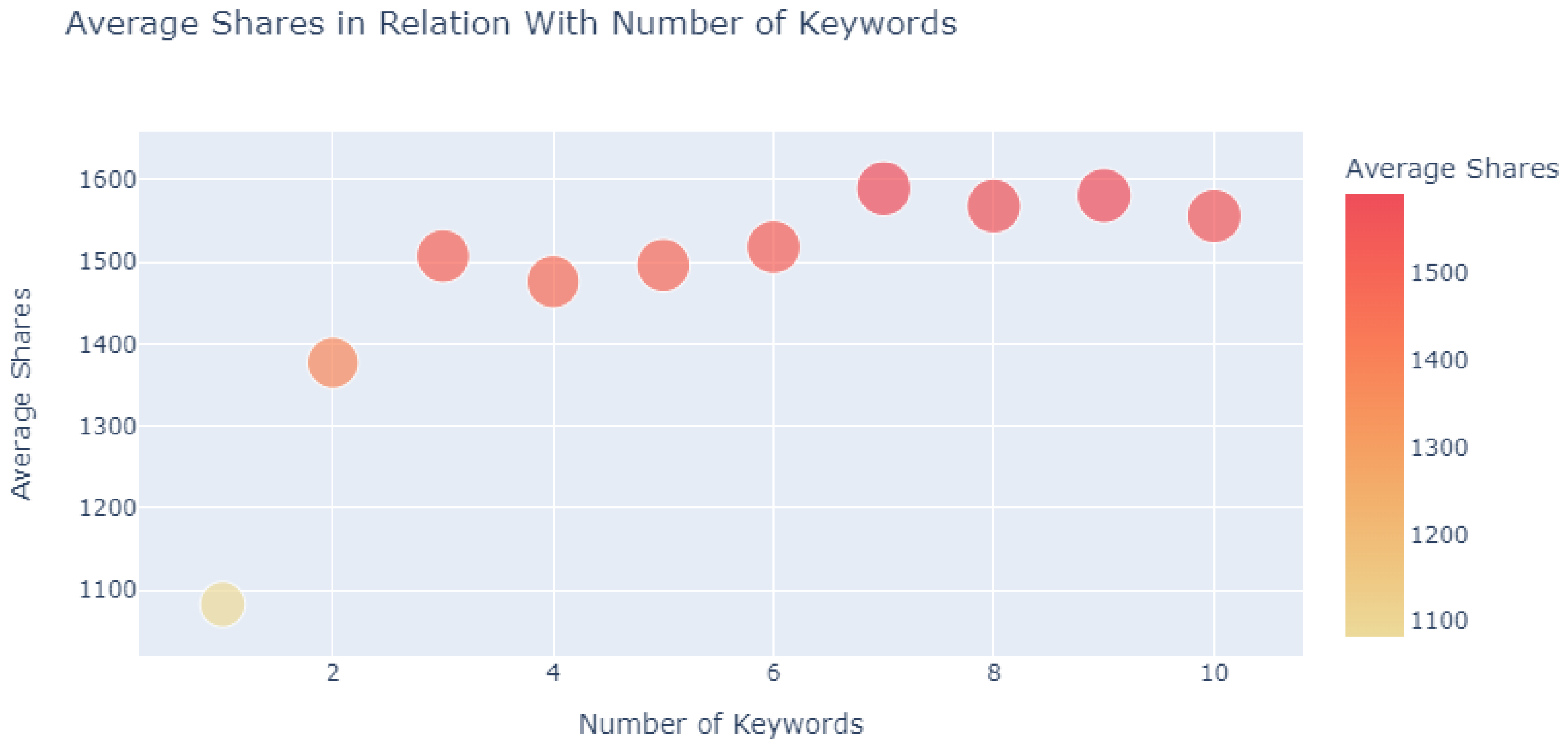


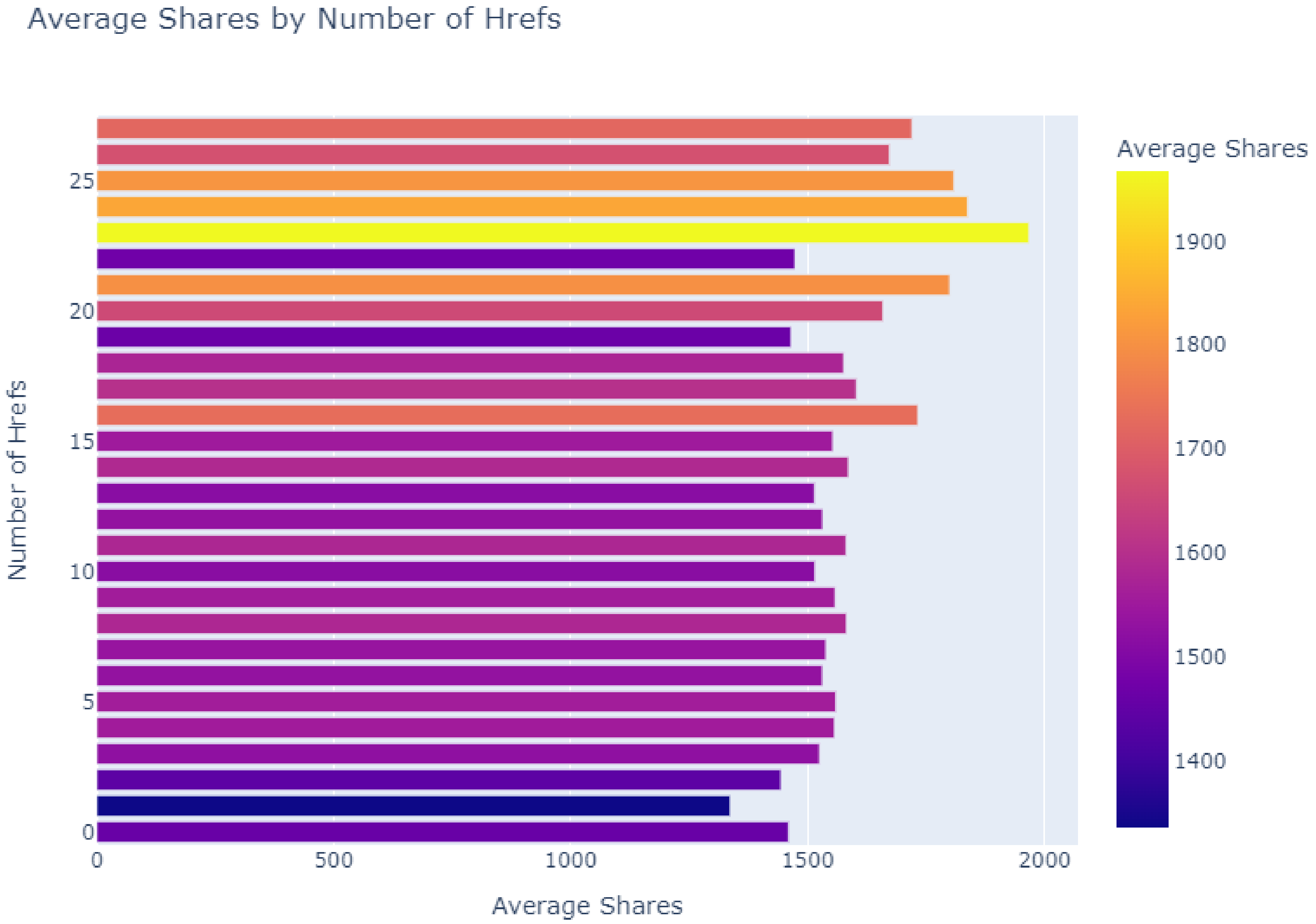


Include 6 images to enhance visual appeal

Embed 2 videos for a dynamic and engaging experience

6. Magic Recipe

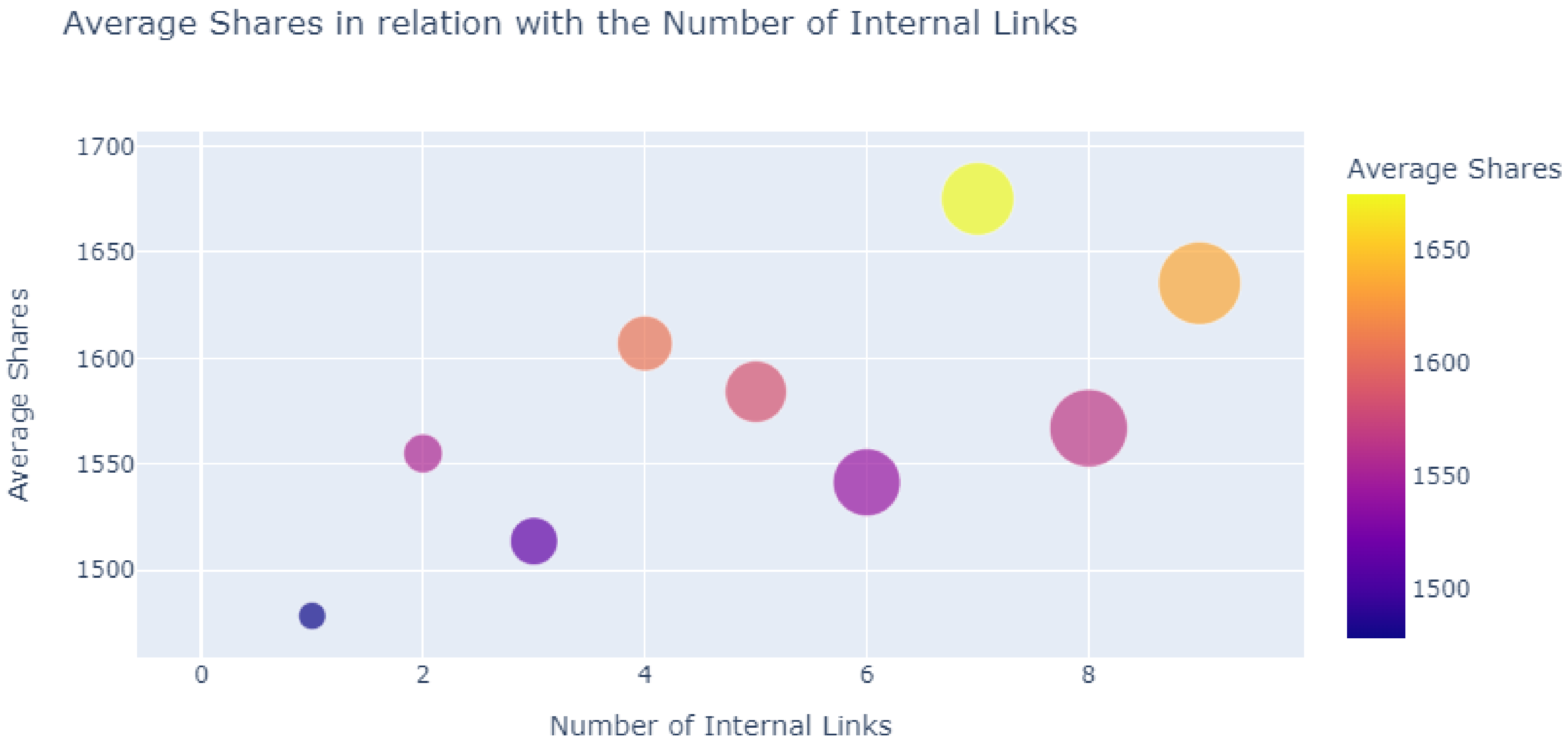


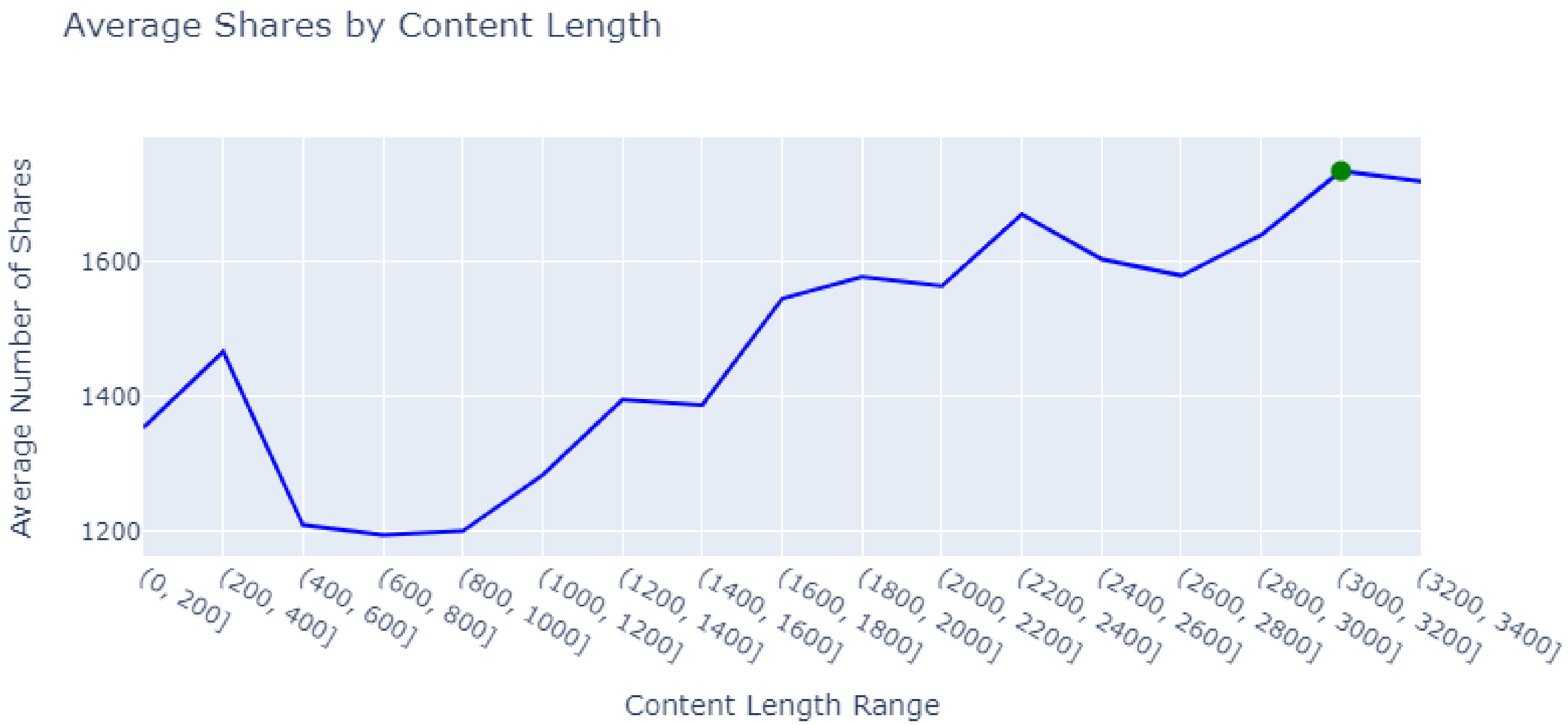


Utilize 7 carefully chosen keywords to enhance search engine visibility.

Include 23 hyperlinks to revelant sources

6. Magic Recipe





Ensure 7 self-referencial hyperlinks within the article

Aim for an average of 3000 to 3200 shares for articles linked within the content

Deliverables

README.txt

Summarizing the task to be accomplished and our conclusions

PDF of the PPT

PowerPoint of the presentation

Jupyter Notebook

Code Jupyter Notebook (.ipynb)

The Flask API

API
Form to the predictive model

TO THE SUCCESS OF YOUR ARTICLES!

ANY QUESTIONS?