DATA ANALYSIS AND VISUALIZATION

Final Project v.2

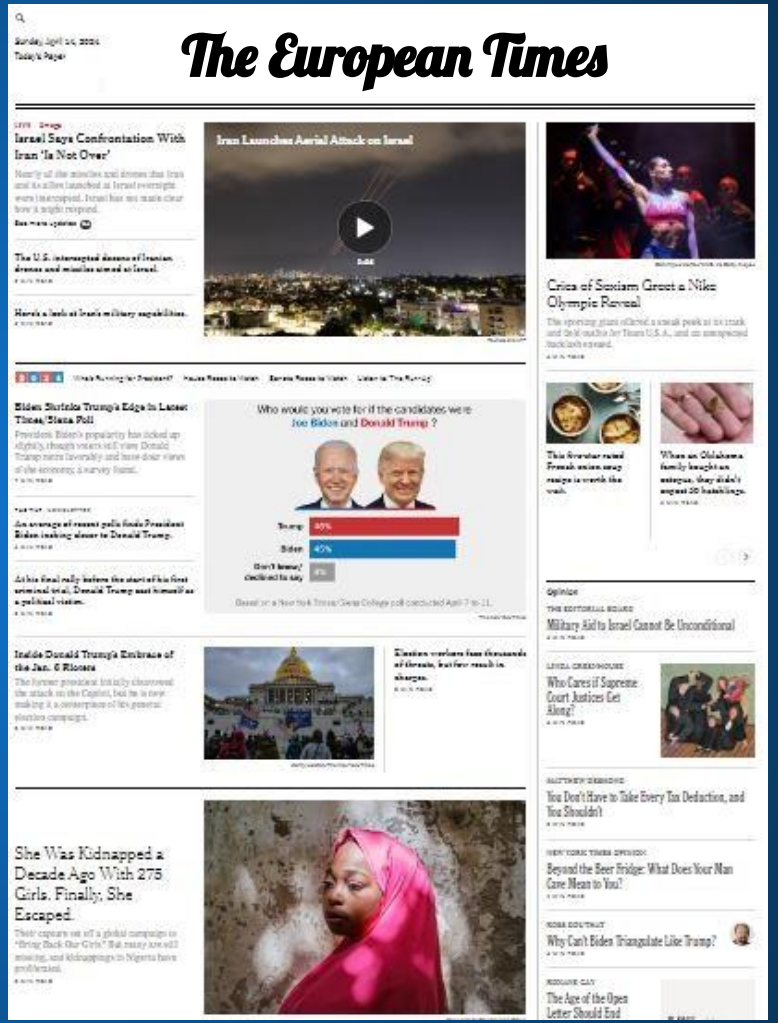
Information

Mirko Rossi

start2impact UNIVERSITY

Introduction

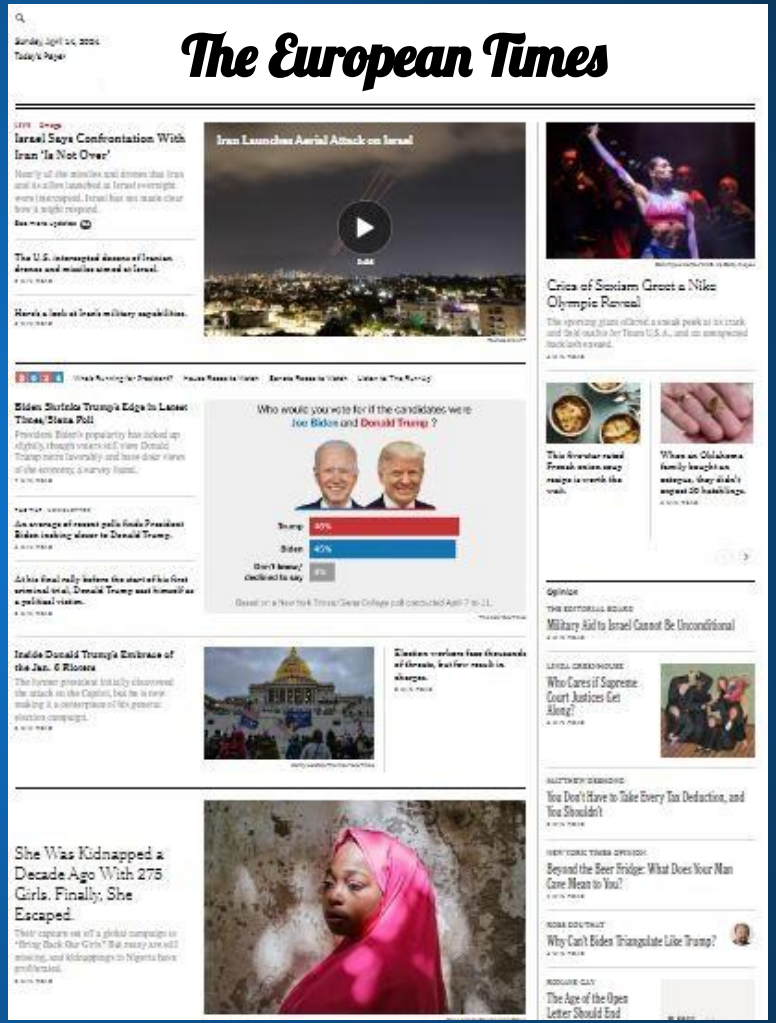
The final project consists of creating a dashboard that includes a strategy proposal and the creation of a user persona for a fictitious online general information newspaper: **The European Times.**



Introduction

"The European Times" is a small media start-up launched in 2021 that aims to give a voice to emerging and talented journalists. The project consists of a website and various social channels used to share published articles.

"The European Times" was born with generalist ambitions, to cover topics ranging from economics to art, producing articles in multiple languages to reach multiple international markets and aiming at multi-platform uses.

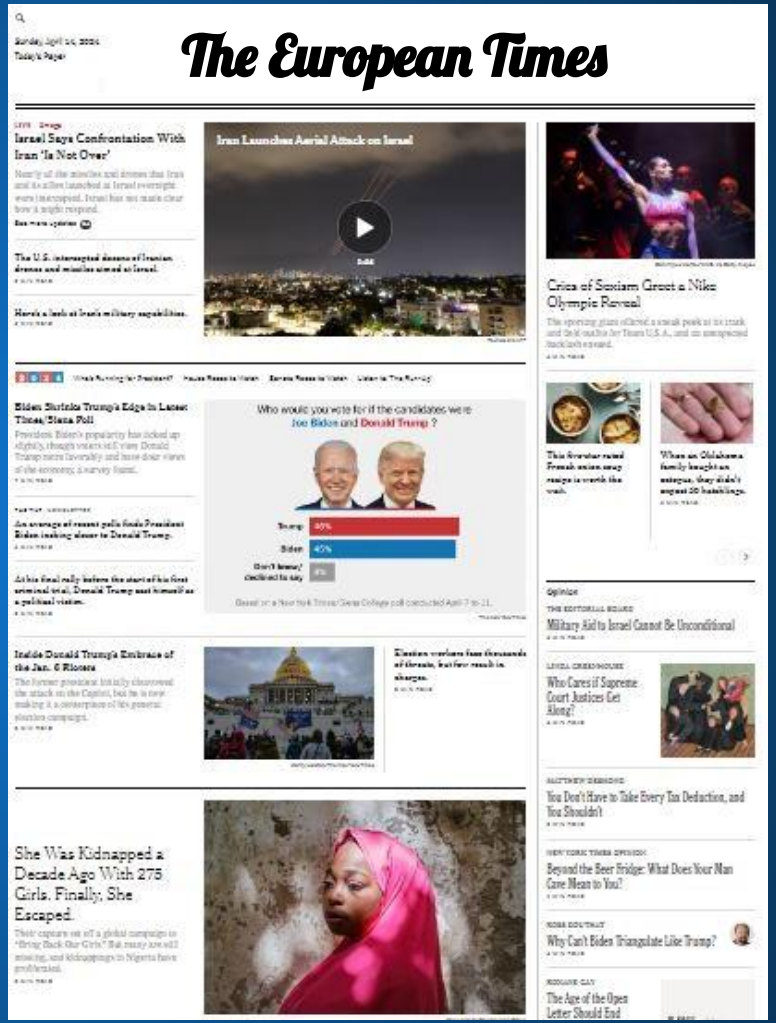


Introduction

The publisher aims to reference itself as a newspaper for a polyglot niche audience and this is the long-term strategy.

After a year, the publisher decides to hire a Data Analyst to analyze the performance of the website and produce, together with marketing, a strategy for the year 2022.

The Data Analyst has access to a database in which accesses to the site by registered users are recorded.



Introduction

I divided the project into four phases:

- EDA (Exploratory Data Analysis) with Python
- Data Visualization
- Creation of the User Persona
- Strategy proposal

In this project, I used the Python language for EDA and Tableau software for Data Visualization.

Exploratory Data Analysis

Dataset Study - head method

	read_date	user_uuid	category	journalist_id	language	length	country	subscription_date	platform	article_id	stars	personas
0	2021-02-25	243	art	117	it	short	it	2020-08-24	tablet	5128	3	P3
1	2021-07-08	157	weather	111	it	long	it	2020-12-02	tablet	732766	5	P1
2	2021-04-17	181	sport	114	en	short	uk	2020-09-12	pc	313130	1	P1
3	2021-11-17	138	finance	111	it	short	it	2020-06-04	pc	612403	3	P2
4	2021-10-04	94	news	103	it	long	it	2020-03-24	pc	632117	5	P3

Dataset study - head and info methods

The dataset is already clean and easy to understand.

It is made up of 9 fields that describe each reading that took place on the site.

There are 90 records and there are no null values.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 90 entries, 0 to 89
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   read_date          90 non-null     object
 1   user_uuid          90 non-null     int64 
 2   category           90 non-null     object
 3   journalist_id      90 non-null     int64 
 4   language           90 non-null     object
 5   length             90 non-null     object
 6   country            90 non-null     object
 7   subscription_date  90 non-null     object
 8   platform           90 non-null     object
 9   article_id         90 non-null     int64 
 10  stars              90 non-null     int64 
 11  personas           90 non-null     object
dtypes: int64(4), object(8)
memory usage: 8.6+ KB
```

Dataset study - dataset fields

Column	Description
read_date	The date the user read the article
user_uuid	User identifier
category	News category
journalist_id	Article author identifier
language	Article language
length	Item length
country	User nationality
subscription_date	Day the user signed up
platformer	Platform from which the article was read
article_id	Item identifier
stars	Average stars assigned to the article (from 1 to 5)

Statistics

Based on the data provided, the site was used for only 80 days with an average of 1.12 articles read per day.

Just 53 users registered on the site with an average of 1.70 articles read per user.

There are 22 journalists working for the site with an average of 4 articles written per journalist.

```
Number of users: 53
Number of journalists: 22
Number of articles read: 90
Number of access days: 80
```

Statistics

The most productive journalist is number 117 and, by sorting the articles by category, each article was read only once.

February and October are the months with more reads.

Journalists who have written the most articles:

journalist_id	Number of articles written
117	8
111	7
105	6

Most read articles by category:

Article ID	Category	Number of Readings
2129	economy	1
711601	weather	1
633861	finance	1

Top three months with the highest number of articles read:

Month	Number of articles read
February	11
October	11
November	10

Statistics

The most loyal user is number 94, who has read 5 articles.

Sorting the individual articles by number of readings also shows that **each article has been read once.**

Users who have read the most articles:

user_uuid	Number of articles read
94	5
209	4
34	3

Most read articles:

	article_id	language	length	platform	Read Count
0	2129	it	short	mobile	1
1	711601	fr	short	tablet	1
2	633861	en	long	mobile	1

Statistics

The most accessed month in 2021 is February, in which 11 articles were read, the highest number.

The month with the least number of accesses in 2021 is May, with 2 accesses and 2 articles read.

The chart could be simplified to a two-column chart since only subscribed users are recorded and every access to the website is related only to the opening of an article page.

Month	Unique Users	Articles Read	Subscriptions	Total Accesses
January	6	7	7	7
February	9	11	11	11
March	8	8	8	8
April	3	3	3	3
May	2	2	2	2
June	5	5	5	5
July	8	8	8	8
August	10	10	10	10
September	9	9	9	9
October	11	11	11	11
November	9	10	10	10
December	6	6	6	6

Distribution of values

The values of the analysed fields describe a prevalence of interest in articles in the weather category in Italian and read by Italian users.

The most-read articles are long and viewed on tablets and PCs.

	Category Count	Percentage
weather	24	26.7%
sport	15	16.7%
finance	15	16.7%
lifestyle	13	14.4%
news	10	11.1%
economy	9	10.0%
art	4	4.4%

	Language Count	Percentage
it	47	52.2%
en	28	31.1%
fr	15	16.7%

	Country Count	Percentage
it	47	52.2%
uk	28	31.1%
fr	15	16.7%

	Platform Count	Percentage
tablet	33	36.7%
pc	32	35.6%
mobile	25	27.8%

	Length Count	Percentage
long	40	44.4%
short	32	35.6%
medium	18	20.0%

Distribution of values

The average rating is 2.85 stars out of 5 but almost 50% of the reviews are below average.

```
df['stars'].describe()
```

```
count    90.000000
mean      2.855556
std       1.583348
min       1.000000
25%       1.000000
50%       3.000000
75%       4.000000
max       5.000000
Name: stars, dtype: float64
```

	Stars Count	Percentage
1	27	30.0%
5	21	23.3%
2	17	18.9%
4	16	17.8%
3	9	10.0%

Summary by category

Summary by category

Category	Language	Length	Platform	Average Ratings	Percentage Articles	Percentage Reads
weather	it	long	tablet	3.00	29%	29%
finance	it	short	mobile	3.00	18%	18%
sport	en	long	pc	2.33	18%	18%
lifestyle	it	long	[pc, tablet]	2.85	16%	16%
news	it	long	tablet	3.20	12%	12%
economy	it	short	pc	2.33	11%	11%
art	it	[long, short]	tablet	3.75	4%	4%

Since each article has been read only once, “Percentage Articles” and “Percentage Reads” show the same percentages.

Data Visualization

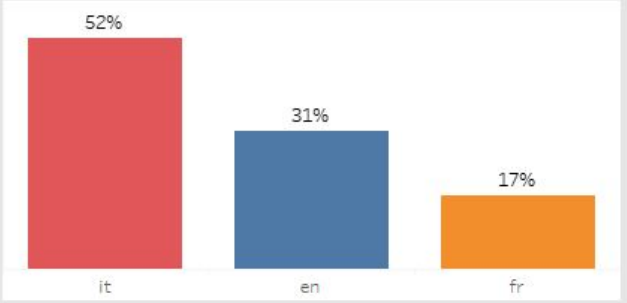
KPI - Key Performance Indicator

The data visualization in Tableau dashboards summarises in graphical form the results already obtained with Python.

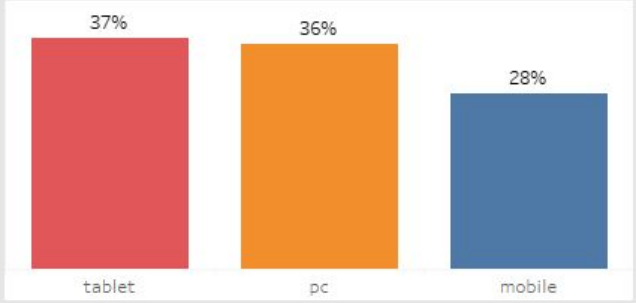
The KPIs (Key Performance Indicators) used are the number of users, the number of articles read and the average score.

WEBSITE ANALYTICS 2021

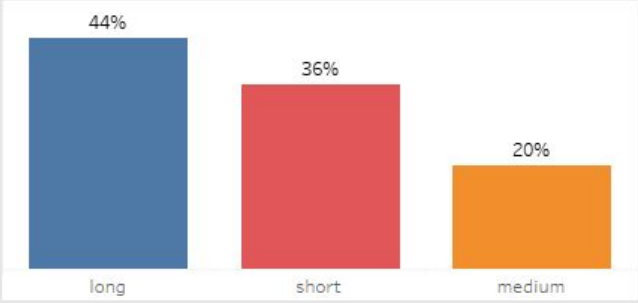
Language



Platform



Length



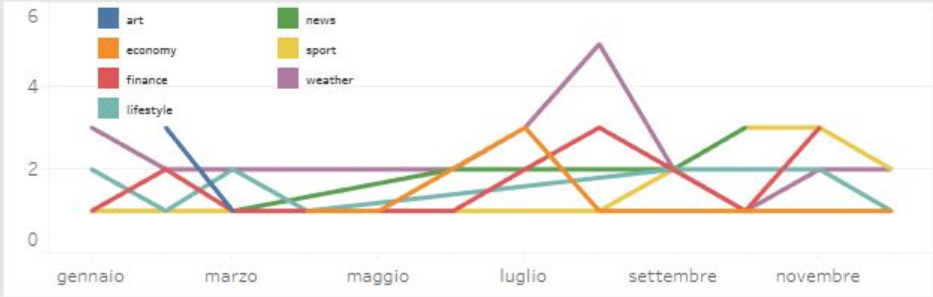
53
USERS

90
ARTICLE READS

2,86/5
ARTICLE AVERAGE SCORE

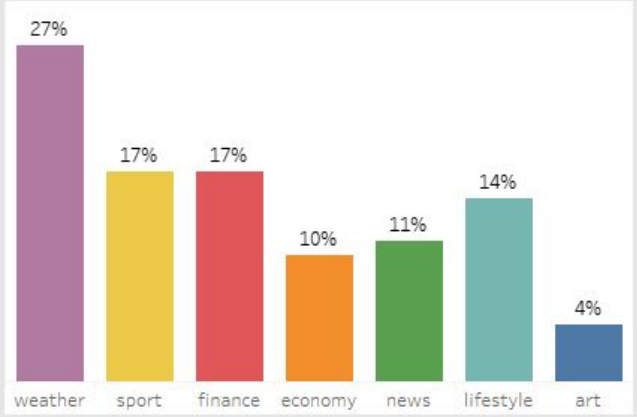
Unique users and articles read

Articles read by category

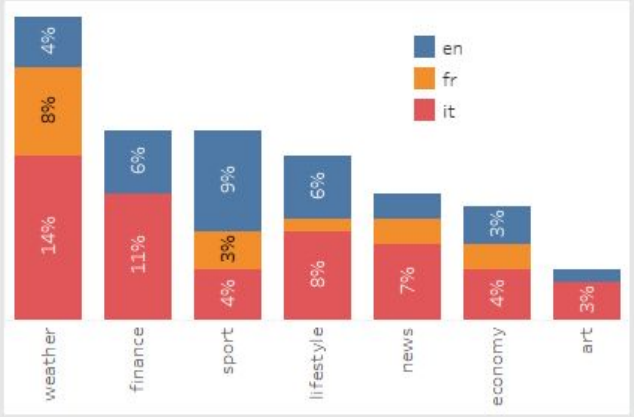


WEBSITE ANALYTICS 2021

Category

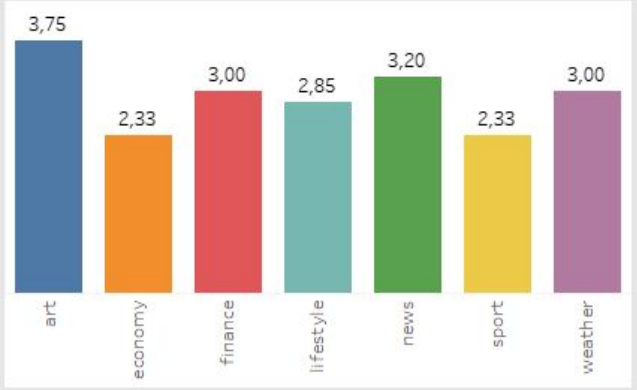


Category and Language

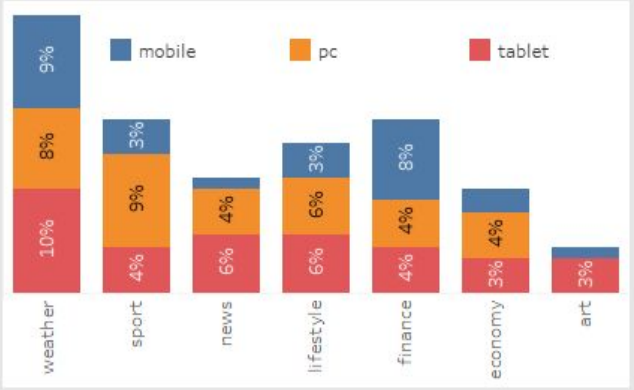


Category and Country

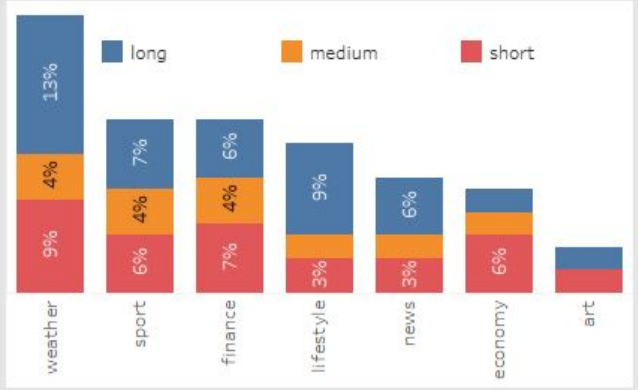
Category and Stars



Category and Platform



Category and Length



Creation of the User Persona

User clustering

To create the User Personas, I chose to use a **data-driven approach** with Python.

I calculated the percentages of users who have a preference for each category and, subsequently, their preference for all the others.

	Preferred Cat.	% Users	weather	finance	sport	lifestyle	news	economy	art
0	weather	24.53%	59%	7%	10%	7%	7%	7%	2%
1	finance	13.21%	9%	44%	15%	15%	6%	9%	3%
2	sport	22.64%	14%	21%	52%	10%	3%	0%	0%
3	lifestyle	11.32%	8%	16%	12%	52%	4%	8%	0%
4	news	13.21%	11%	11%	5%	5%	53%	11%	5%
5	economy	11.32%	18%	14%	0%	9%	14%	41%	5%
6	art	3.77%	11%	11%	0%	0%	22%	11%	44%

User clustering

Below I have selected some reasonably coherent clusters:

1. weather and sports
2. economy and finance
3. lifestyle, art and news

	Preferred Cat.	% Users	weather	finance	sport	lifestyle	news	economy	art
0	weather	24.53%	59%	7%	10%	7%	7%	7%	2%
1	finance	13.21%	9%	44%	15%	15%	6%	9%	3%
2	sport	22.64%	14%	21%	52%	10%	3%	0%	0%
3	lifestyle	11.32%	8%	16%	12%	52%	4%	8%	0%
4	news	13.21%	11%	11%	5%	5%	53%	11%	5%
5	economy	11.32%	18%	14%	0%	9%	14%	41%	5%
6	art	3.77%	11%	11%	0%	0%	22%	11%	44%

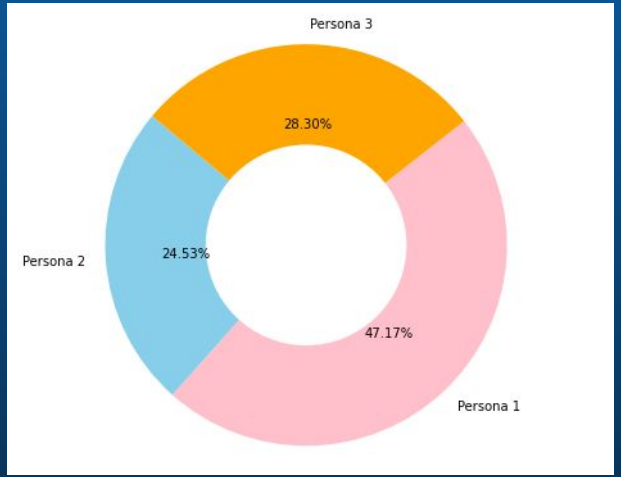
	Name	% Users	weather	finance	sport	lifestyle	news	economy	art
0	P1	47.17%	36%	14%	31%	8%	5%	3%	1%
1	P2	24.53%	13%	29%	7%	12%	10%	25%	4%
2	P3	28.3%	10%	12%	5%	19%	26%	10%	16%

User clustering

Finally, I calculated the average approval percentages after the merger.

I carried out the study of the profiles and the creation of the dashboards with Tableau.

	Name	% Users	% Preference
0	P1	47.17%	67%
1	P2	24.53%	54%
2	P3	28.3%	61%



USER PERSONA 1 - Giulio

Giulio is the user most connected to the platform, as half of the site views come from him.

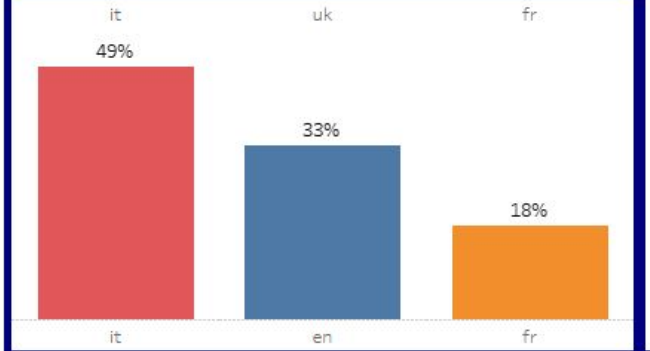
Giulio is passionate about outdoor **sports** such as football, basketball and skiing, which is why he regularly consults the **weather forecast** on our site.

Giulio reads articles mainly in Italian via the PC. He prefers in-depth articles, but doesn't disdain short ones such as the results of the matches that interest him.

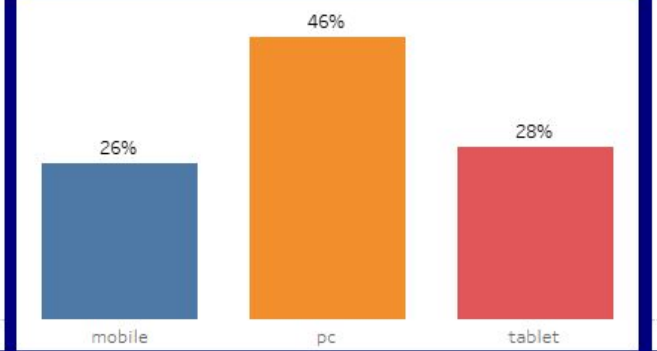
His rating is above average.

INTERESTS

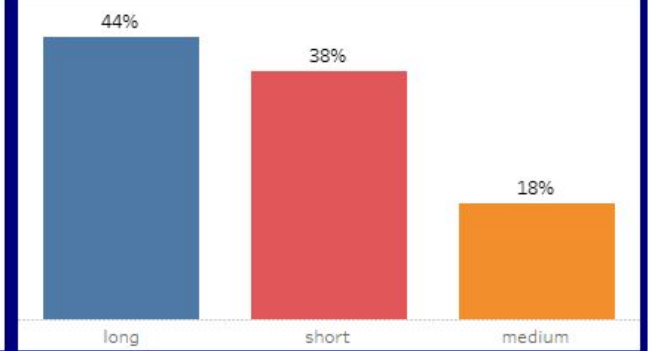
Language and Country



Platform



Length



47%
USERS

43%
ARTICLE READS

3,03/5
ARTICLE AVERAGE SCORE

USER PERSONA 2 - Cristina

Cristina is a user responsible for a quarter of the portal's views.

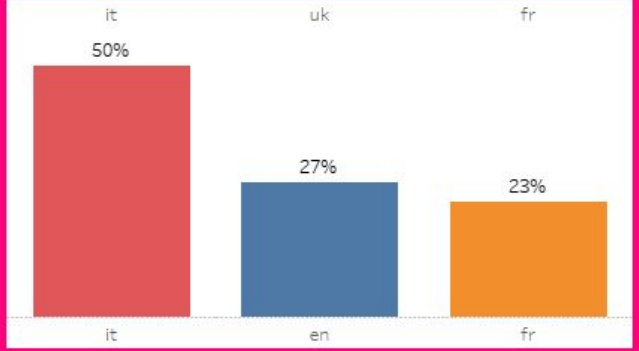
Cristina is passionate about **economics** and, given that she follows market trends, also about **finance**. She is responsible for approximately 30% of accesses to the website.

Cristina reads articles mainly in Italian, regardless of whether via smartphone or PC. She prefers in-depth articles, but does not disdain short ones such as international stock market trends.

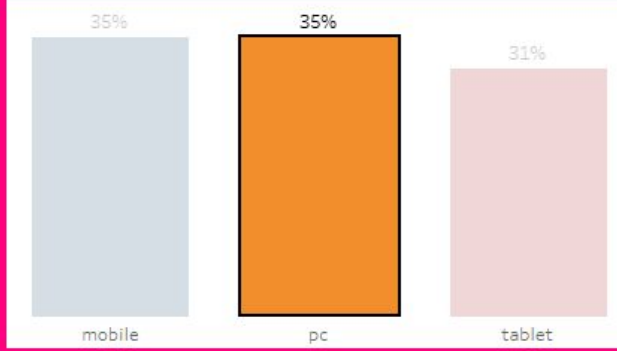
Her rating is below average.

INTERESTS

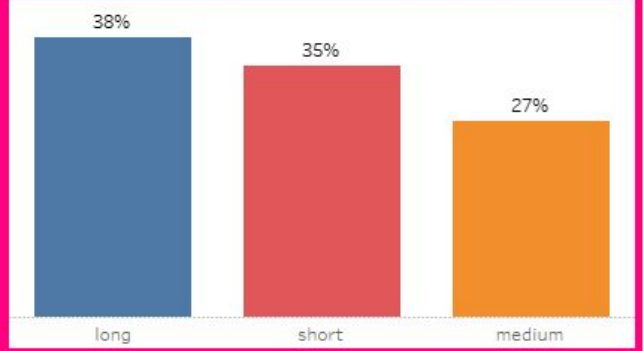
Language and Country



Platform



Length



25%
USERS

29%
ARTICLE READS

2,54/5
ARTICLE AVERAGE SCORE

USER PERSONA 3 - Stefano

Stefano is a reader who recently discovered our site and recommends it to friends.

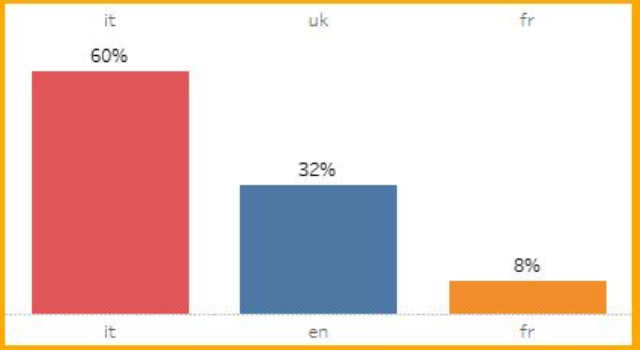
Stefano is passionate about **art**. He is interested in **lifestyle** readings such as fashion and travel, and, for this reason he is always up to date on the latest **news**. He is responsible for approximately 30% of hits to the website.

Stefano reads articles mainly in Italian and from tablets. He prefers in-depth articles, such as travel guides or art criticism, and only sporadically reads shorter articles such as press agencies.

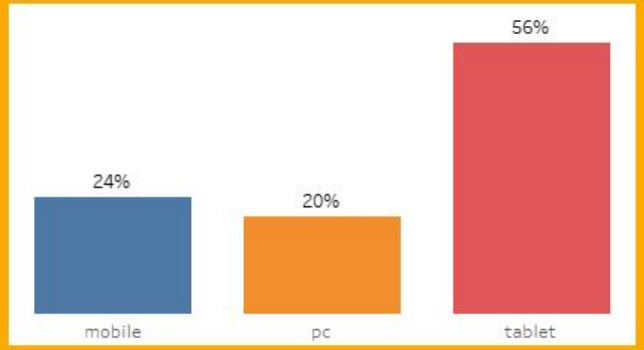
His rating is above average.

INTERESTS

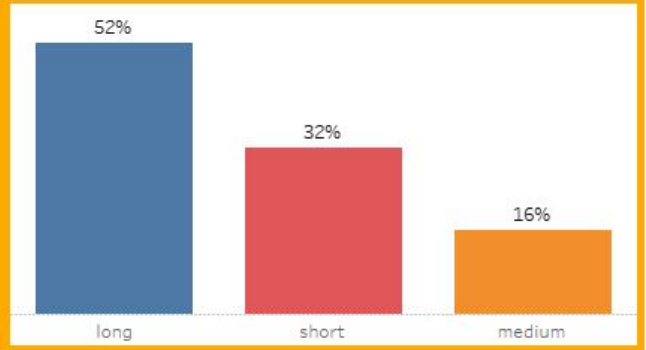
Language and Country



Platform



Length



28%

USERS

28%

ARTICLE READS

2,92/5

ARTICLE AVERAGE SCORE

Strategy proposal

Strategy Proposal

For organic growth of the various sections of the site, it is necessary to invest in art articles, which represent just 4% of the total articles and in general in articles in French, whose percentage fraction is under 20%. This objective is achievable by increasing collaborations with journalists specializing in art and journalists of all categories who write in French.

Strategy Proposal

Regarding Giulio, user type 1, the use of this user on our site must be consolidated by inserting content in line with the current level of quality, which Giulio appreciates.

It is possible to enhance its presence during the sale of advertising spaces for use on PCs.

Strategy Proposal

Regarding Cristina, user type 2, we need to increase the share of this user and invest resources to produce higher quality content.

By increasing the percentage of this user, it will be possible to enhance their presence during the sale of advertising spaces for use on PCs, tablets and smartphones.

Strategy Proposal

Regarding Stefano, user type 3, we need to increase the share of this user and invest resources to produce content in French, which is present in less than 10% of the articles he reads.

By increasing the percentage relating to this user, it will be possible to enhance their presence during the sale of advertising spaces for use, especially on tablets.

Links

The Jupyter notebook and Tableau dashboards from this project are publicly available.