

Fashion Marketplace

Consumer Analysis

Master in BI & Big Data Analytics (2022/2023)

Business Intelligence & Big Data Management – Team project work

Team:

Michele Iannotta

Luca Izzo

Fabio Jr Lorenzini

Milano, September 2023

Agenda

1. Business Scenario
2. Process
3. BI Architecture
4. Data Ingestion
5. Data Profiling & ETL
6. Data Quality
7. Data Visualization
8. Final Results
9. Conclusions and Next Steps



1. Business Scenario

Nowadays most **fashion** firms sell products and services on their **proprietary ecommerce**, but in parallel they can leverage on **wholesale** web platforms (e.g. Amazon, JD, Zalando, etc.).

However, it is often difficult to capture customer's "voice" and needs.

Sentiment analysis represents a powerful tool to collect useful insights from customers, also from direct competitors, that can boost business' growth.



Business perspective

Allow stakeholders to:

- Have an overview about consumer's product perception
- See how different platforms present and sell the same product
- Take action in order to improve product features (e.g. materials, comfort, etc.)
- Take action in order to enhance proprietary ecommerce UX and/or marketing strategies



Customer perspective

Get information regarding:

- Price, rating, reviews by product and platform
- Other clients' perception about product features

2. Process

1

Web scraping of products' ratings, reviews from different platforms to create 3 main datasets (Accessory, Apparel, Footwear).



2

Data profiling and data **quality** analysis.



3

Datasets integration and connection to data **visualization** tools – ETL.



4

Qualitative analysis (UI/UX) by product among platforms.

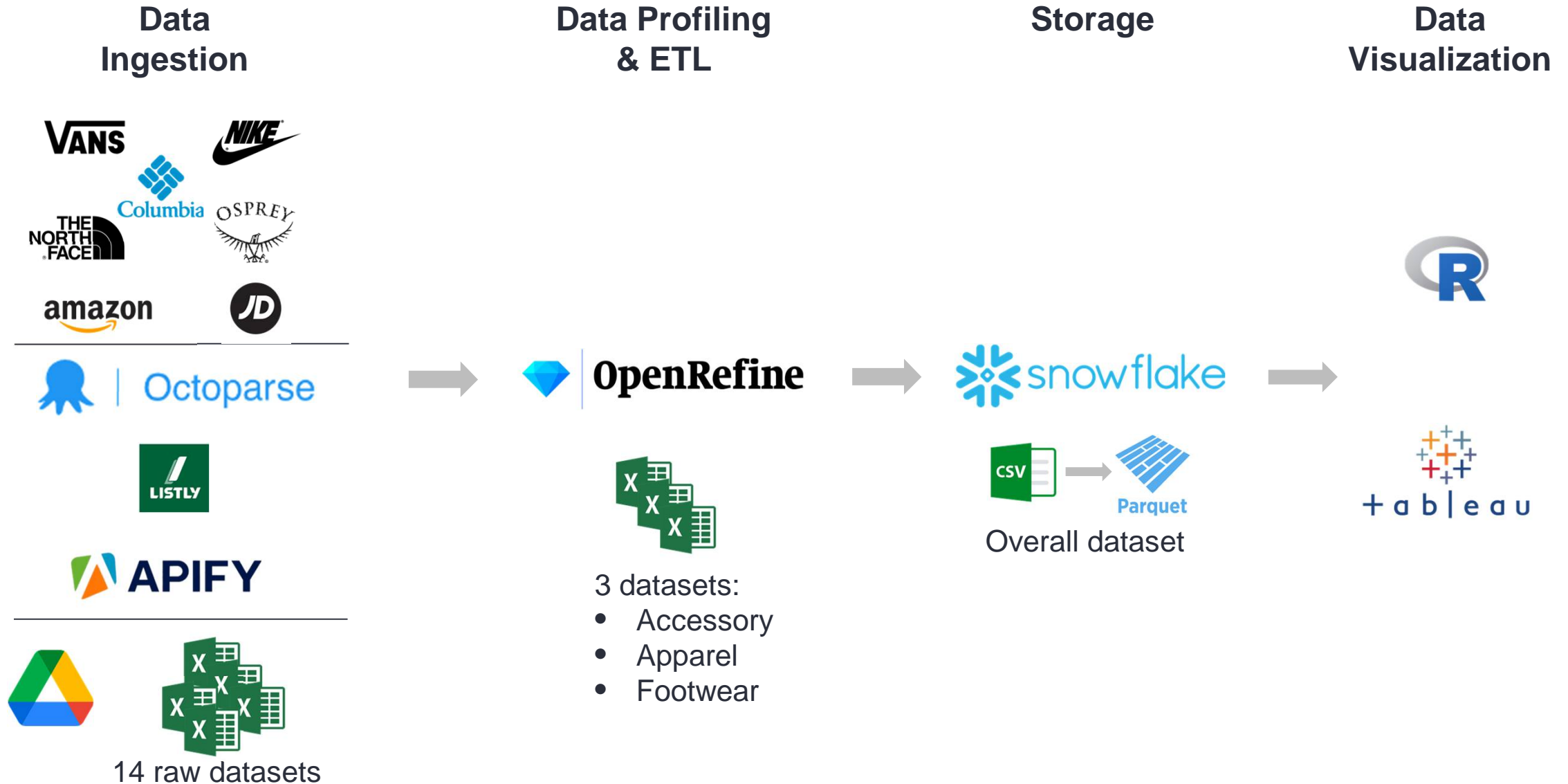


5

Sentiment Analysis by product category.



3. BI Architecture



4. Data Ingestion 1/2

Three product categories: **Accessory • Apparel • Footwear**

For each category, two products have been compared to analyze customer's rating and reviews.

Primary raw dataset for each product collected by scraping **proprietary ecommerce**.

Accessory (backpack)



Borealis Classic



Talon 22



Footwear



Old Skool
VANS



Air Force 1
NIKE

Apparel



Puffect Jacket



Nuptse 1996







amazon



In addition to proprietary ecommerce platform, other main wholesale web marketplaces chosen were **Amazon** (general) and **JD** (fashion focused) to capture different consumer behavior and target.

4. Data Ingestion 2/2

Platform	Scraper
Proprietary Ecommerce	  Octoparse
Amazon	 APIFY
JD	 Octoparse

⚠ For Data Ingestion phase, “*Python-Selenium*” (first attempt tool used) resulted not the best option due to:

- large amount of dataset (→ time-consuming)
- errors encountered (change page and captcha) on platforms (especially on Amazon)



- “**Octoparse**”: quick and reliable web scraper
- “**Apify**”: resulted the best tool for scraping Amazon
- “**Listly**”: user-friendly tool, even if with limitations (e.g. few records scraped by single launch)

Dataset elements

Category	Product	Ecommerce	Amazon	JD
Accessory	Borealis Classic	✓	✓	✗
	Talon 22	✓	✓	✗
Apparel	Nuptse 1996	✓	✓	✓
	Puffect Jacket	✓	✓	✓
Footwear	Old Skool	✓	✓	✓
	Air Force 1	✓	✓	✓

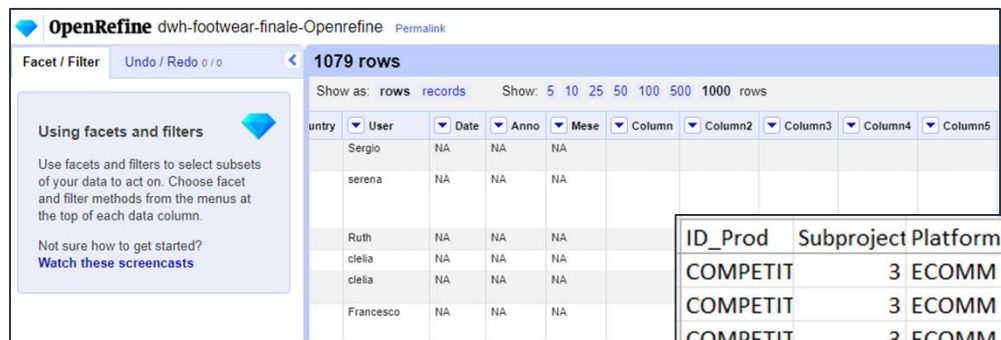
Each product category dataset include information coming from at least two platforms.

Original **14 raw datasets** have been unified in **3 datasets organized by category**.

5. Data Profiling & ETL 1/2

For each product category, once scraped raw data, put them together in a main dataset and having re-organized attributes, here follow the main cleaning steps executed with “**OpenRefine**”:

1. Removed undesired *blank* columns.
2. Added useful attributes (e.g. “*ID_Prod*”, “*Subproject*”, “*Platform*” and “*Category*”) and renaming.
3. Checked and adjusted data type for each column, e.g.:
 - “*Product_name*”, “*Review_title*” → text
 - “*Rating*”, “*Year*”, “*Month*” → number
 - “*Date*” → date.
4. Trimmed out any leading, trailing and consecutive whitespace.
5. Replaced any “*NA*” value with *null*.



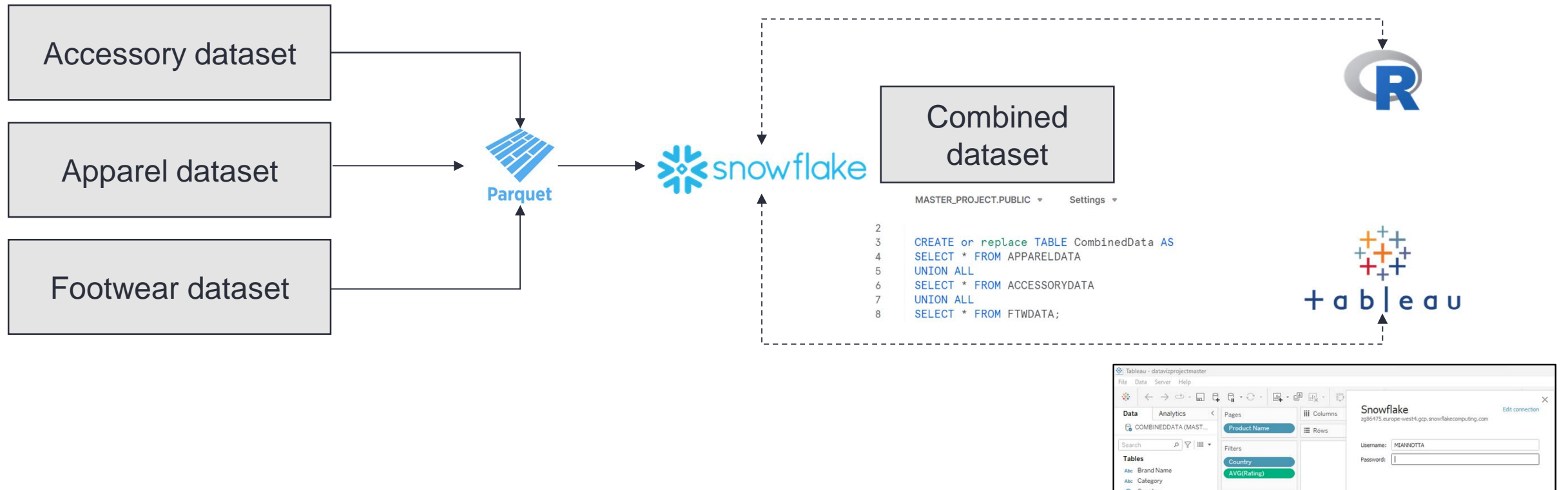
ID_Prod	Subproject	Platform	Category	Product_n	Brand_nar	Rating	Review_title	Review_cc	Country	User	Date	Year	Month
COMPETIT	3	ECOMM	Footwear	Nike Air Fc Nike		4	Une paire emblématique			FRED71	2022-11-19	2022	11
COMPETIT	3	ECOMM	Footwear	Nike Air Fc Nike		5	The Forces is always a go for me...		United Sta	Allfred	2022-11-17	2022	11
COMPETIT	3	ECOMM	Footwear	Nike Air Fc Nike		5	My all time favorite shoes! They go		United Sta	Raven12	2022-11-17	2022	11
COMPETIT	3	ECOMM	Footwear	Nike Air Fc Nike		5	I like the shoes because they are cc		United Sta	Malachi31	2022-11-16	2022	11
COMPETIT	3	ECOMM	Footwear	Nike Air Fc Nike		5	Best shoes ever.		United Sta	thomas28	2022-11-16	2022	11
COMPETIT	3	ECOMM	Footwear	Nike Air Fc Nike		5	I like hibbit sport bc of tl			Ratelbush	2022-11-16	2022	11
COMPETIT	3	ECOMM	Footwear	Nike Air Fc Nike		5	Ordered online for picku		United Sta	Jose55	2022-11-16	2022	11

(Illustrative extraction from cleaned “*Footwear*” dataset.)

5. Data Profiling & ETL 2/2

Once obtained the 3 category datasets, “**Snowflake**” was introduced to:

1. Create the overall dataset and act as storage.
2. Integrate with “**R**” and “**Tableau**” for data visualization purposes.



6. Data Quality

Column	Data Type	Accuracy	Completeness	Consistency
Rating	Number	NA	100%	100%
Review_title	Text	NA	95%	87%
Review_content	Text	NA	100%	100%
Country	Text	100%	84%	100%
User	Text	99%	21%	99%
Date	Date	NA	100%	100%

Accessory dataset
819 records

Column	Data Type	Accuracy	Completeness	Consistency
Rating	Number	NA	100%	100%
Review_title	Text	NA	86%	100%
Review_content	Text	NA	100%	99%
Country	Text	100%	64%	100%
User	Text	94%	46%	94%
Date	Date	NA	71%	100%

Apparel dataset
465 records

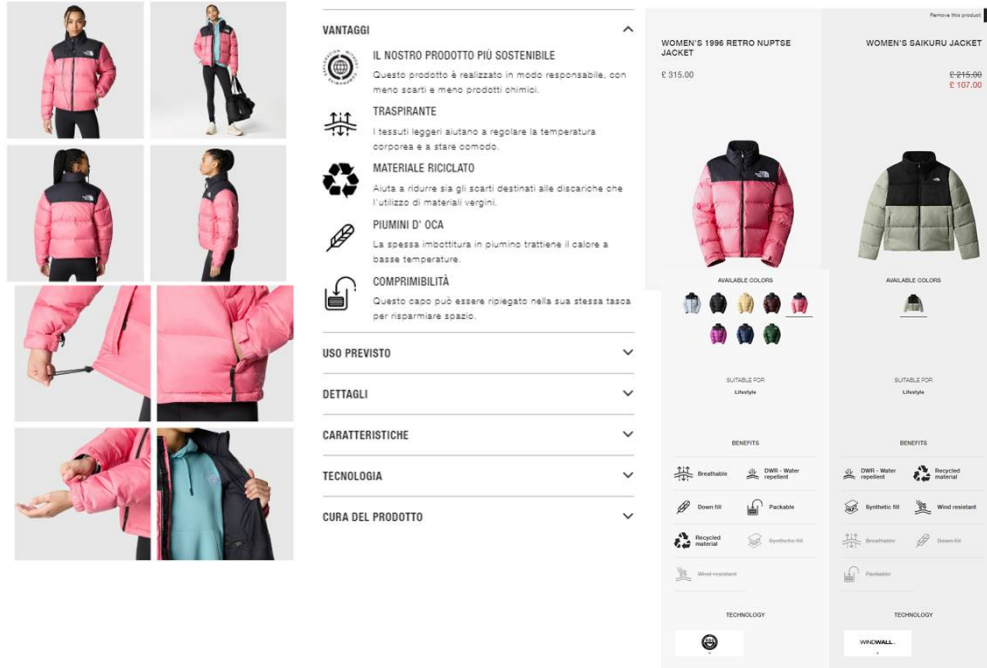
Column	Data Type	Accuracy	Completeness	Consistency
Rating	Number	NA	100%	100%
Review_title	Text	NA	100%	99%
Review_content	Text	NA	61%	100%
Country	Text	100%	82%	100%
User	Text	100%	44%	100%
Date	Date	NA	81%	100%

Footwear dataset
1078 records

- Apparel dataset presented less records compared to Accessory and Footwear, but completeness and consistency for ratings and reviews were reliable as well.
- Completeness for “User” column resulted poor, but not significant for sentiment analysis.
- Volatility, currency and timeliness dimensions were not considered.

7. Data Visualization - UI/UX Analysis

TNF Ecommerce (.it)



Images

- TNF ecomm images are more focused on the **product**, showing a total of 13 images, emphasizing product details and features.
- JD is more **fashion oriented**, 8 images (+ 1 video) capturing model on different non-natural poses, often full body, to show matching outfits.

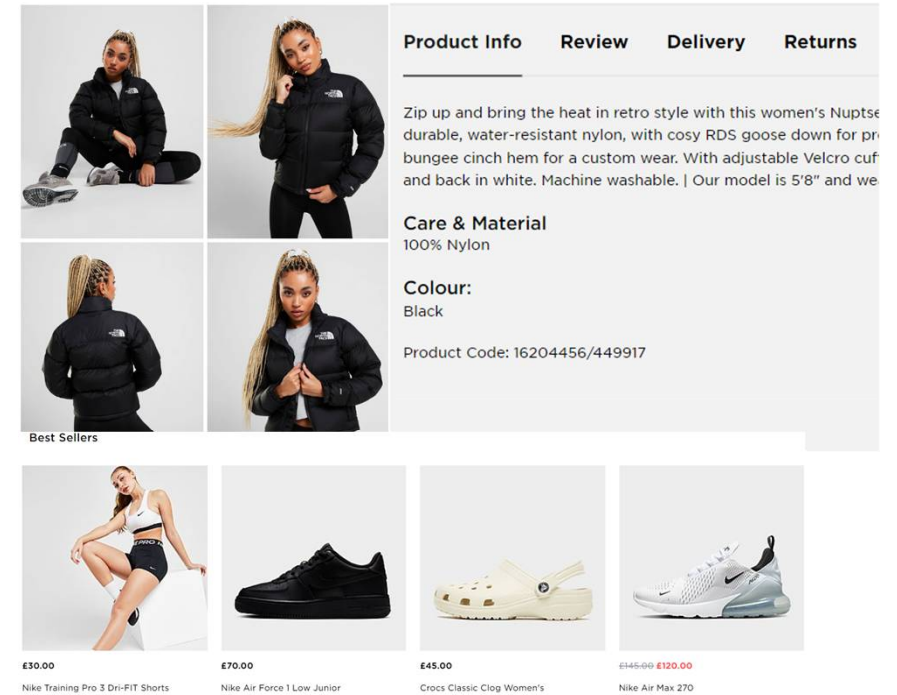
Product Description

- TNF shows very **exhaustive product description** (intended use, advantages, technology, caring instructions, details, feature).
- JD highlights client services (delivery and returns).

Consumer Experience

- TNF offers possibility to **compare** its products + other products recommendation.
- JD does not really create such a comprehensive consumer experience.

JD (.uk)



7. Data Visualization | Sentiment Analysis



- **Sentiment analysis** is a technique that extracts and evaluates the emotional tone expressed in a text to gain valuable insights.
- The emotions associated with a given text can be positive, neutral or negative.

How does it work?

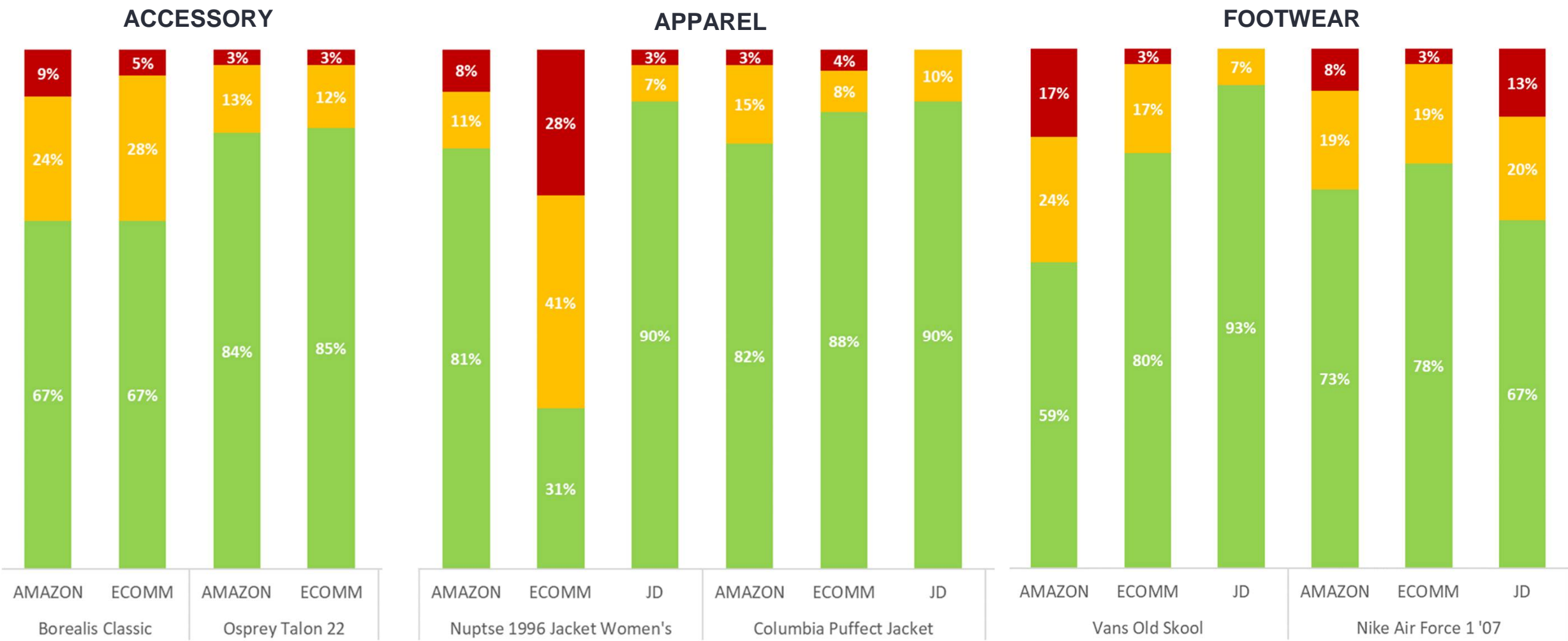
- Text preprocessing (drops all the stop words from the input sequences and other information irrelevant).
- Applying “*get_sentiment*” R-function that iterates over a vector of strings and returns sentiment values based on the default method.
- Tokenization of reviews (process of taking text and breaking it into individual terms) and grouping word for similar meaning.

We want to capture consumer voice from 3 different digital platforms:

- Amazon (generic marketplace)
- JD (fashion marketplace)
- Brand Ecommerce

7. Data Visualization | Sentiment Analysis

- Backpacks' perception is likely to be aligned across platforms due to the nature of the product (less variables in place)
- In general, competitors are more aligned cross platforms compared to TNF and Vans
- There is not a common pattern across different categories when it comes to product appreciation *



* For TNF ecomm, we are considering also the unpublished reviews (subjected to a worse rate)

7. Data Visualization | TNF Borealis

Avg Reviews Rating

Amazon 3,6

Ecomm 4

Lifestyle Backpack, but also used for travel.
Some students/workers who find it **small, heavy and expensive** → negative impact on the perceived quality.

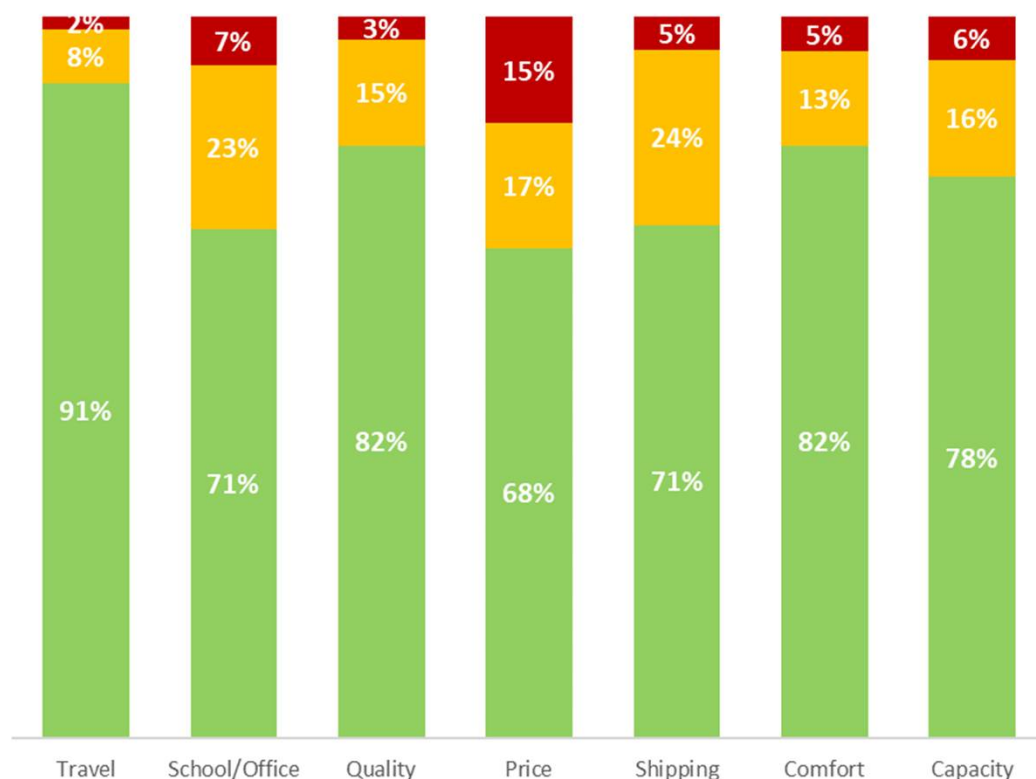
Amazon customer is more sensitive to price and shipping.



Insights for business:

- Create a product with a similar design, but specifically for students (lighter, bigger, less expensive)
- Develop a new sales channel large-scale distribution (Hypermarket, Supermarket) with special discounts
- Investigate shipping issues with Amazon

AMAZON



Average Reviews Rating scale varies from 1 to 5.

ECOMM



7. Data Visualization | Osprey Talon 22

Avg Reviews Rating

Amazon 4,5

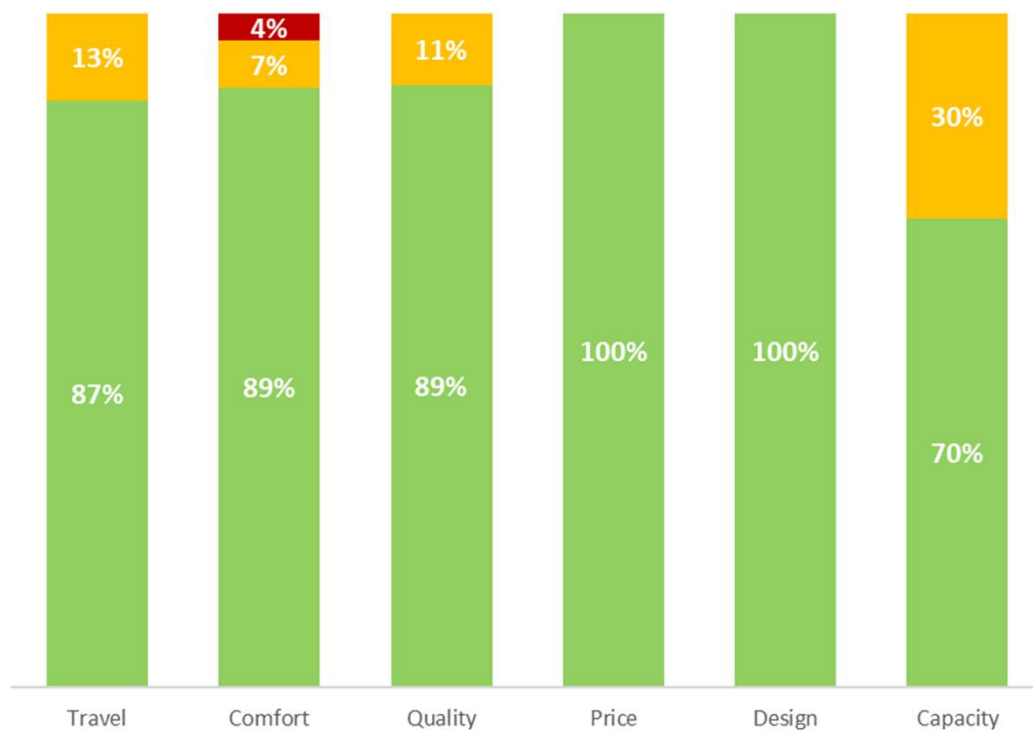
Ecomm 4,6

Compared to Borealis, Osprey Talon is used exclusively for travel by satisfied customers.

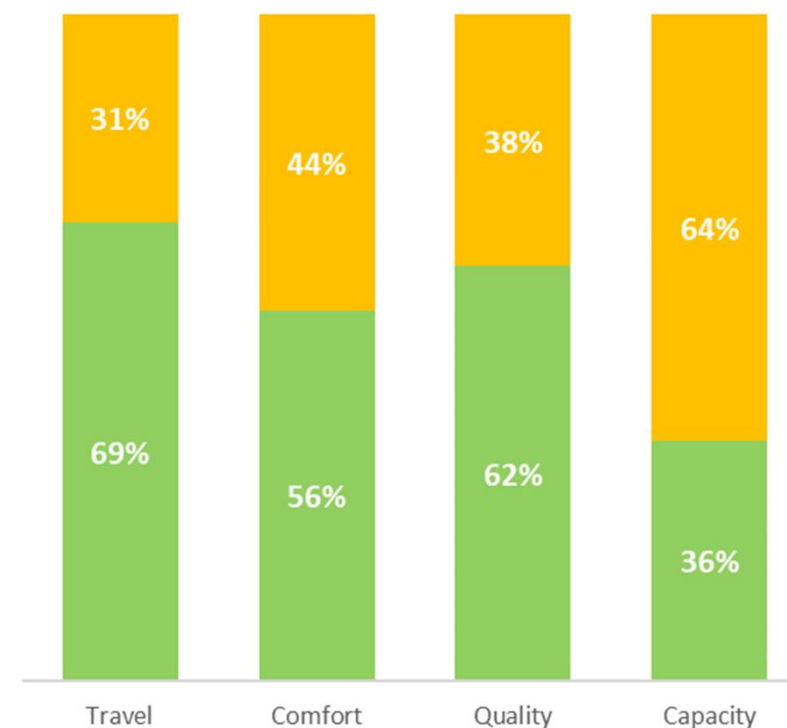
Insights for business:

- The target of this backpack is well defined (hiking people)
- The price, size and capacity of Talon should be analyzed by TNF product development to improve performance

AMAZON



ECOMM



Average Reviews Rating scale varies from 1 to 5.

7. Data Visualization | Vans Old Skool

Avg Reviews Rating

Amazon 3,4

Ecomm 4,7

JD 5

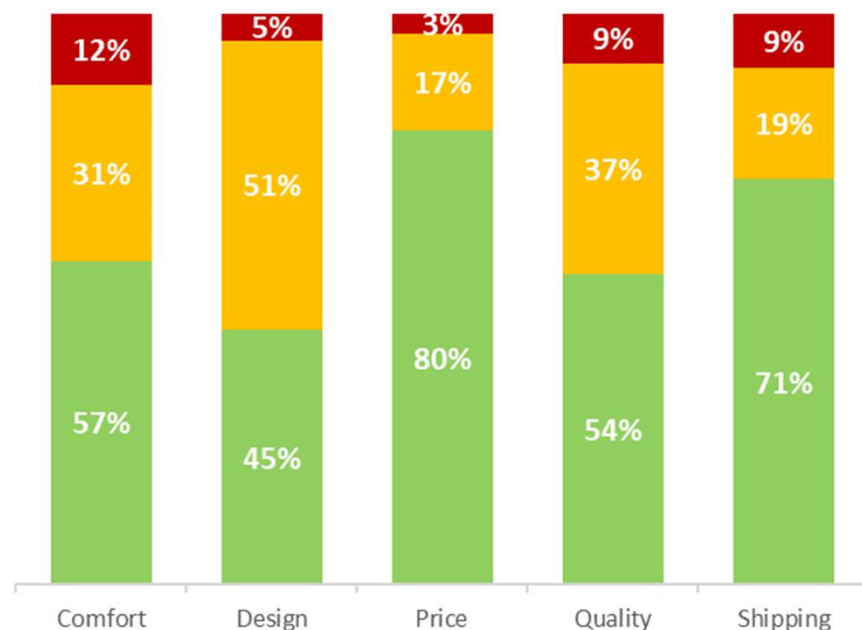
- On Vans.com, shoes appreciated for their design and price, but issue with comfort. Customers complain heel problems.
- Amazon customer looks more critic, it is clearly not a Vans customer, who prefers style rather than comfort.



Insights for business:

- Work with product development team to improve the quality and comfort of the shoes
- Enhance collaboration with iconic brands to ramp-up sales

AMAZON



ECOMM



JD



ALL POSITIVE*

**REVIEWS FOCUSED ON
COMFORT FOR
EVERYDAY USAGE
AND GOOD SHIPPING**

7. Data Visualization | Nike Air Force 1

Avg Reviews Rating

Amazon 4,1

Ecomm 4,5

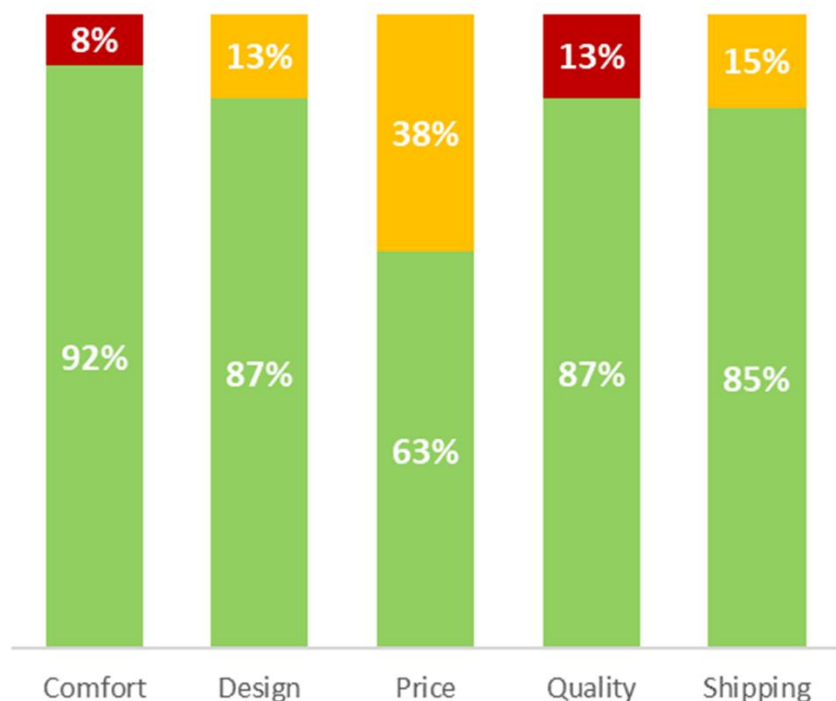
JD 4,5

Nike Air Force are really appreciated for their design and better comfort compared to Vans Old Skool. The price is perceived high and on Amazon sometimes is not original.

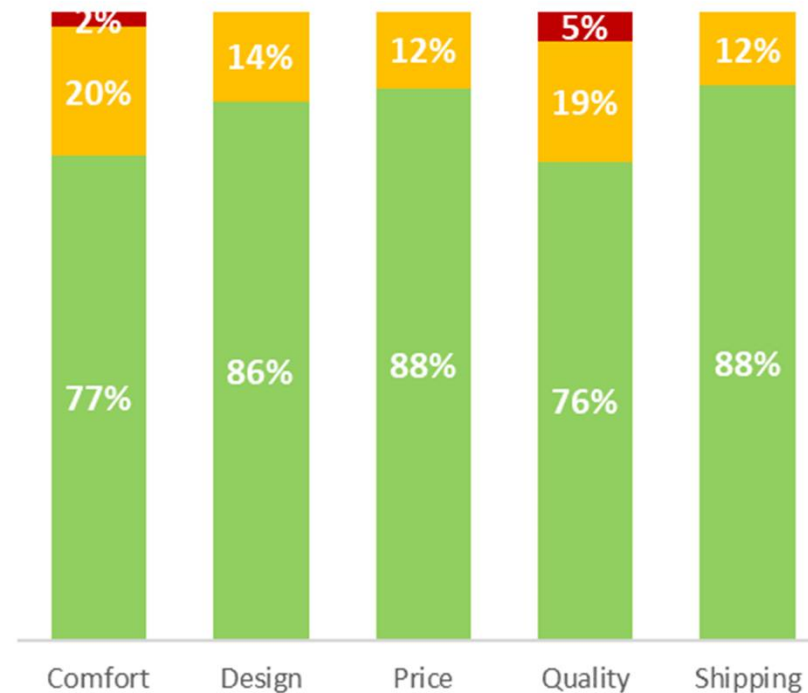
Insights for business:

- Vans could improve and communicate the quality of the product specifying suggested use (e.g. in some reviews there are clients who use Old Skool also to play tennis!)

AMAZON



ECOMM



JD



ALL POSITIVE*

**REVIEWS FOCUSED ON
SPORT USAGE AND
QUALITY**

7. Data Visualization | TNF Nuptse 1996

Avg Reviews Rating

Amazon 4,5

Ecomm 3,1

JD 4,8

TNF Nuptse is really appreciated by customers for its comfort, design and price. Compared to Amazon, on the official website there is an important issue with product availability.

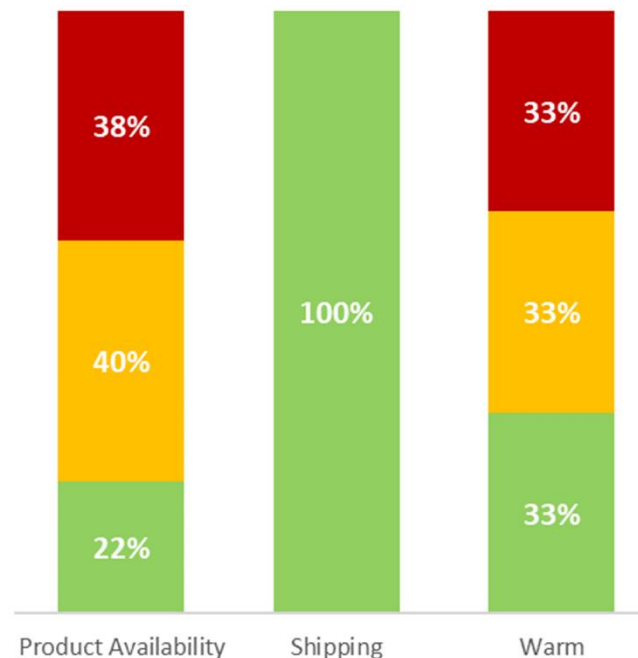
Insights for business:

- Analyze root cause of product unavailability on the official web site
- Customers complain about product quality (issue with feather). Improve jacket seams.
- Customers buy the product as a gift for children/relatives/friends. Create special offers on birthday and Christmas.

AMAZON



ECOMM



JD



ALL POSITIVE*

**REVIEWS FOCUSED ON
WARM CAPACITY AND
SHIPPING**

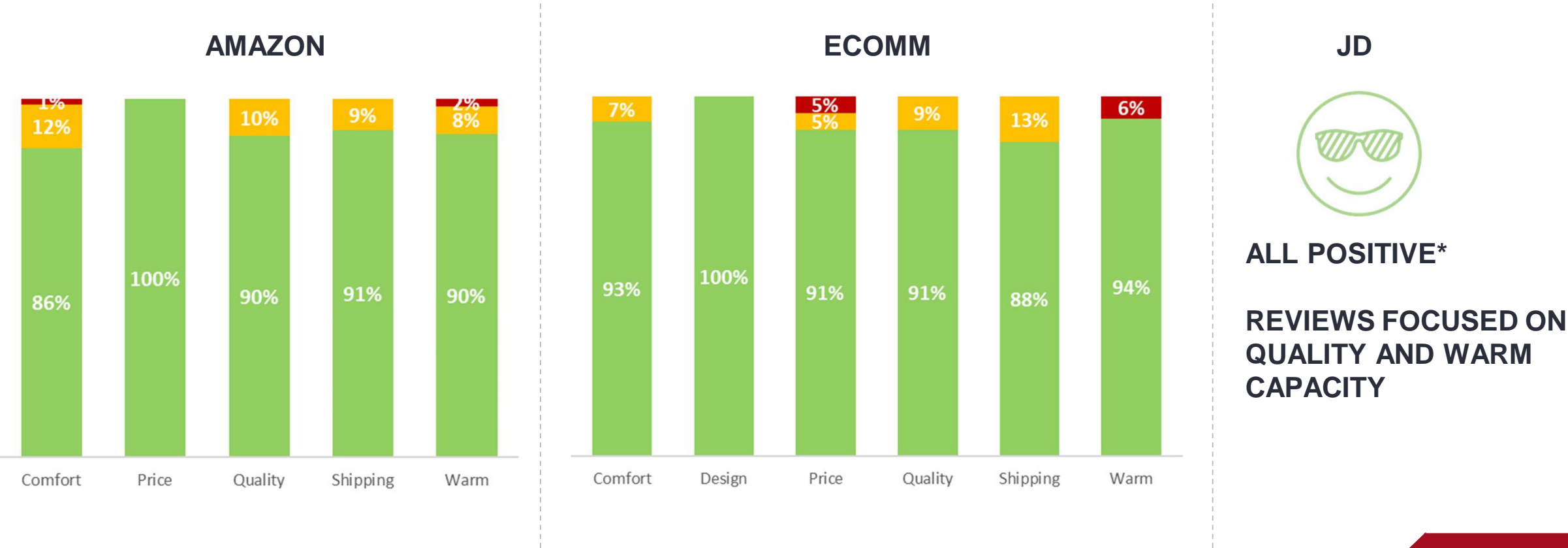
7. Data Visualization | Columbia Puffect

Avg Reviews Rating		
Amazon 4,4	Ecomm 4,8	JD 4,5

Columbia Puffect Jacket is really appreciated by customers for its comfort and design. Some customers complain about the warm power and the price as well (similar too Nuptse).

Insights for business

- Analyze Columbia Puffect material that keep it that warm.
- Improve warm power of the product.



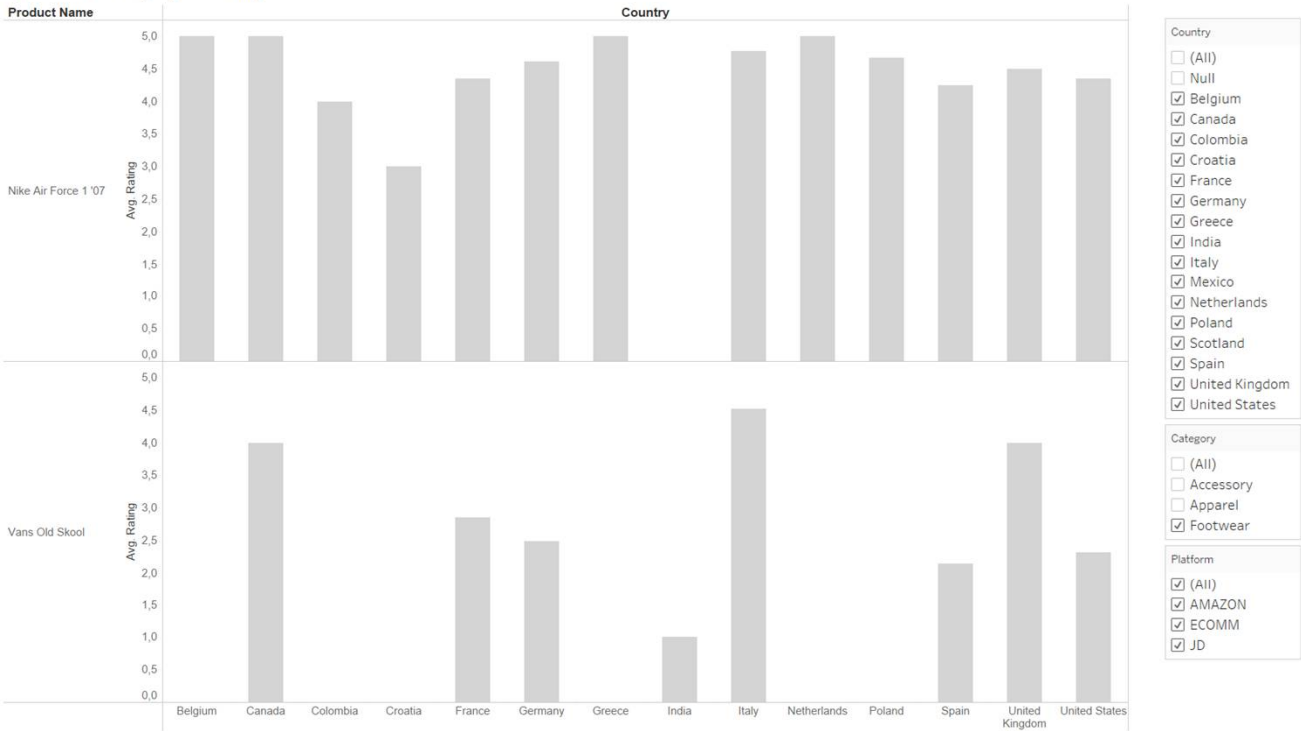
Average Reviews Rating scale varies from 1 to 5.

*Only 30 reviews

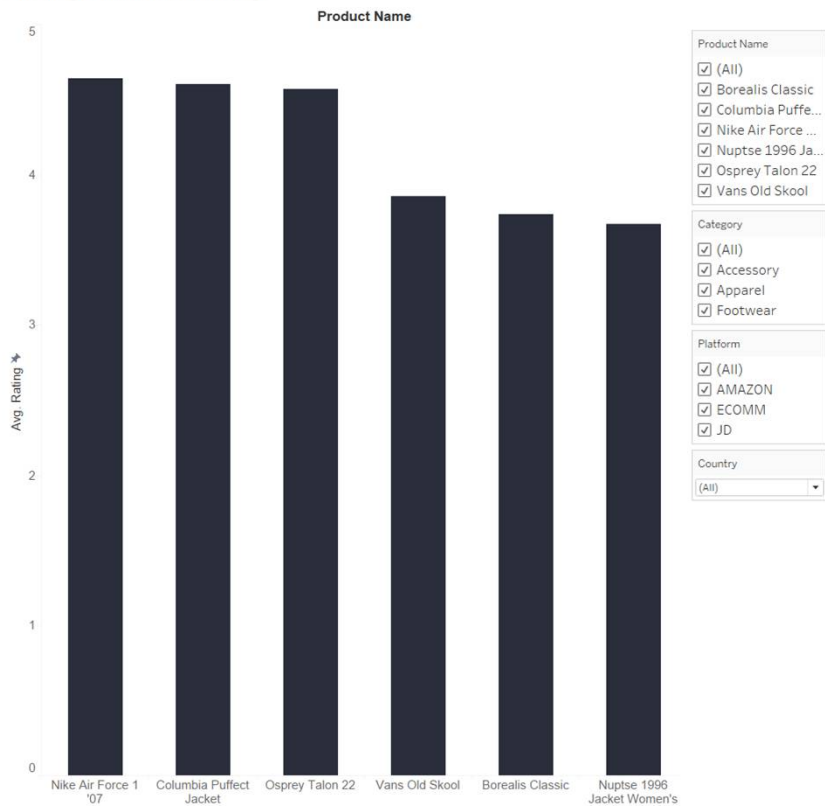
7. Data Visualization | Product Rating



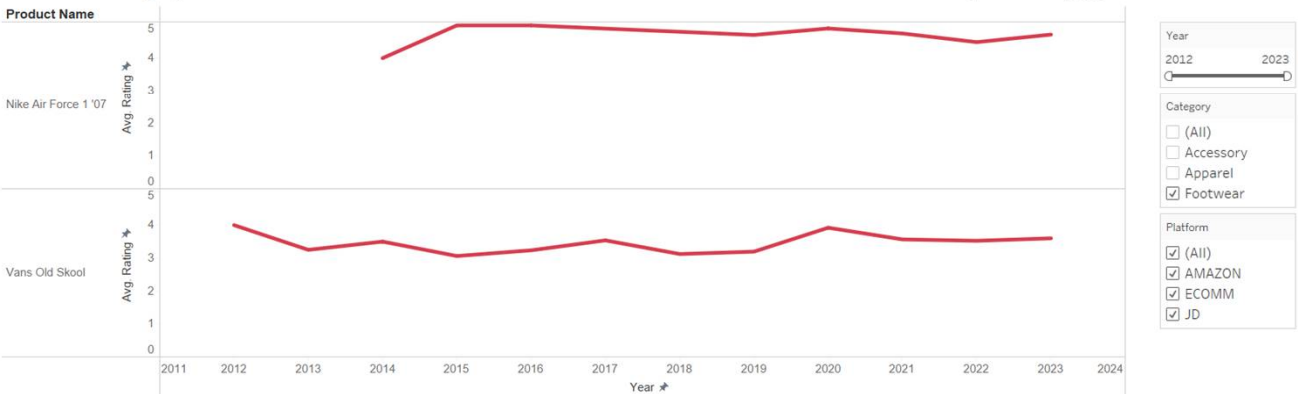
Product Rating by Country



Average Product Rating



Product Rating by Year



8. Final Results

Achieved Objectives



1. Created a **dataset** for each product category.
2. Executed **data profiling** and data **quality** analysis to ensure valuable results.
3. Implemented an **ETL** procedure to integrate datasets and visualize analysis.
4. Performed **sentiment analysis** to capture insights and compare products.
5. Carried out a **qualitative analysis** to compare how platforms sell a same product.

Challenges / Critical factors



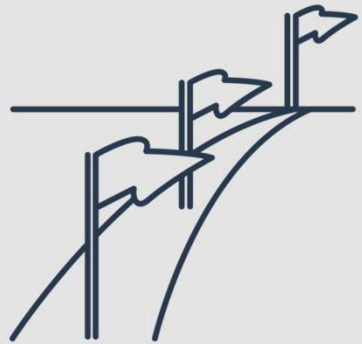
1. Very **time consuming** to find **same products** available on **different platforms** with enough rating and reviews to be analyzed.
2. Zalando (originally ecommerce selected) was replaced by JD, since from mid July '23 online **reviews** were **no more available on Zalando**.
3. The **JD** platform presented only **few review records** compared to official ecommerce or Amazon. This impacted the data analysis.
4. Faced **difficulties** during first **data ingestions** attempts through Python-Selenium. This redirected the exploration of other online web scraping tools (i.e. Octoparse, Listly, Apify).
5. Online scraping tools presented **limitations** on number of **records scraped** (e.g. Listly: only 10 records at a time) or not correct attributes alignment among different launches.

9. Conclusions and Next Steps

Moving along the overall data analysis process performed:

- **Business** is able to have a view about pros&cons of their products' features (and competitors) on proprietary ecommerce and main wholesale platforms. So they can take action in several ways, e.g.:
 - improve R&D on products
 - organize better marketing campaigns or business development strategy
 - implement a more engaging ecommerce UX/UI
- **Customers** are able to compare products' rating and reviews at first sight before proceeding with a purchase on a specific platform.

Future investigation ideas



- Extend data gathering to other competitors' ecommerce and/or products.
- Find out any clustering among users who buy a product on a specific platform.
- Adopt a *Word2vec* dedicated model to cluster review words.

THANKS

Team:

Michele Iannotta / micheleiannotta273@gmail.com

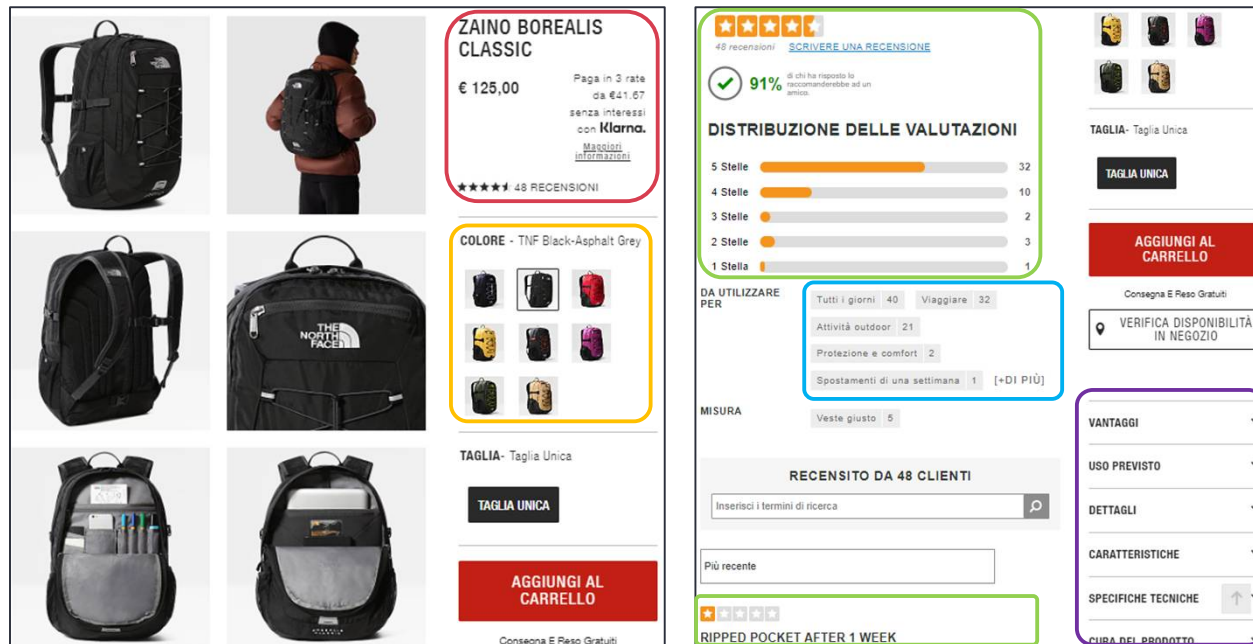
Luca Izzo / l.izzo7@campus.unimib.it

Fabio Jr Lorenzini / fabiojr.lorenzini@gmail.com , <https://github.com/FabioJrLorenzini>

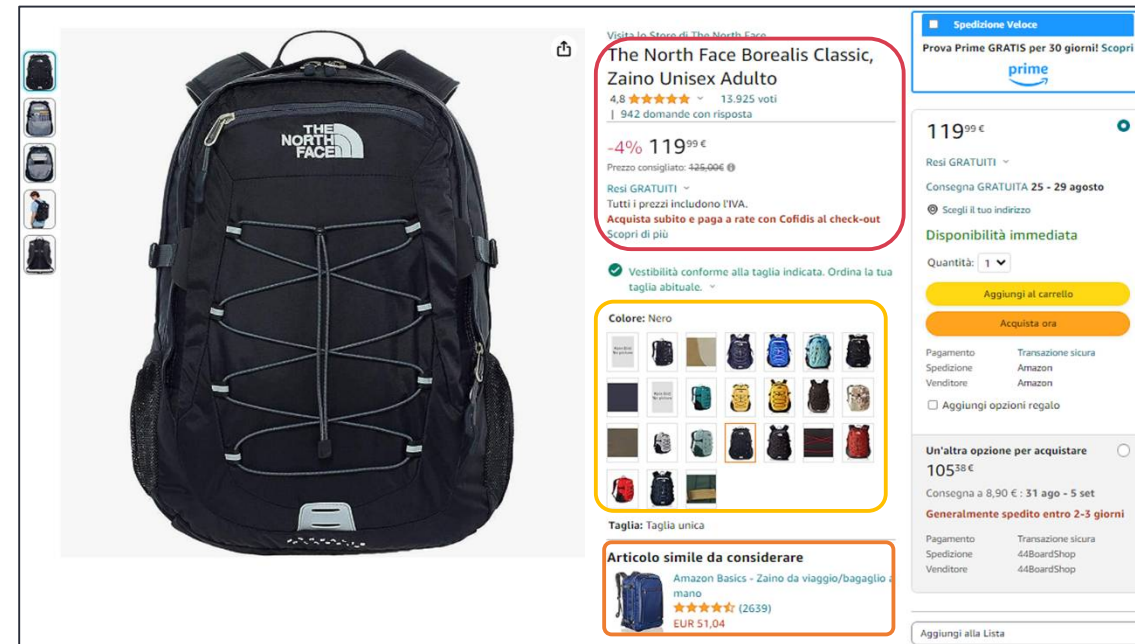
APPENDIX

7. Data Visualization - UI/UX Analysis

TNF Ecommerce (.it)

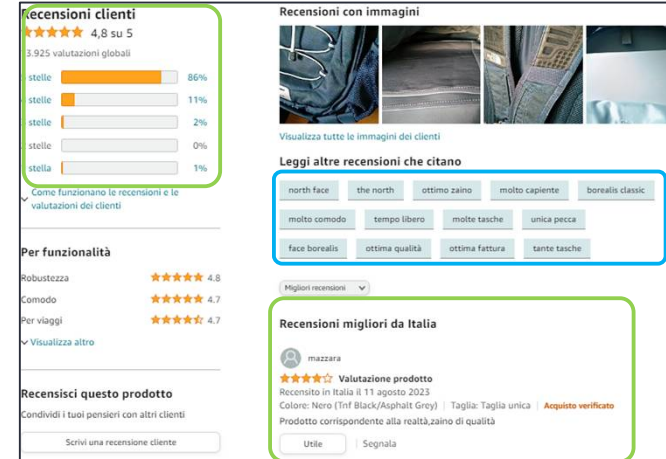


Amazon (.it)



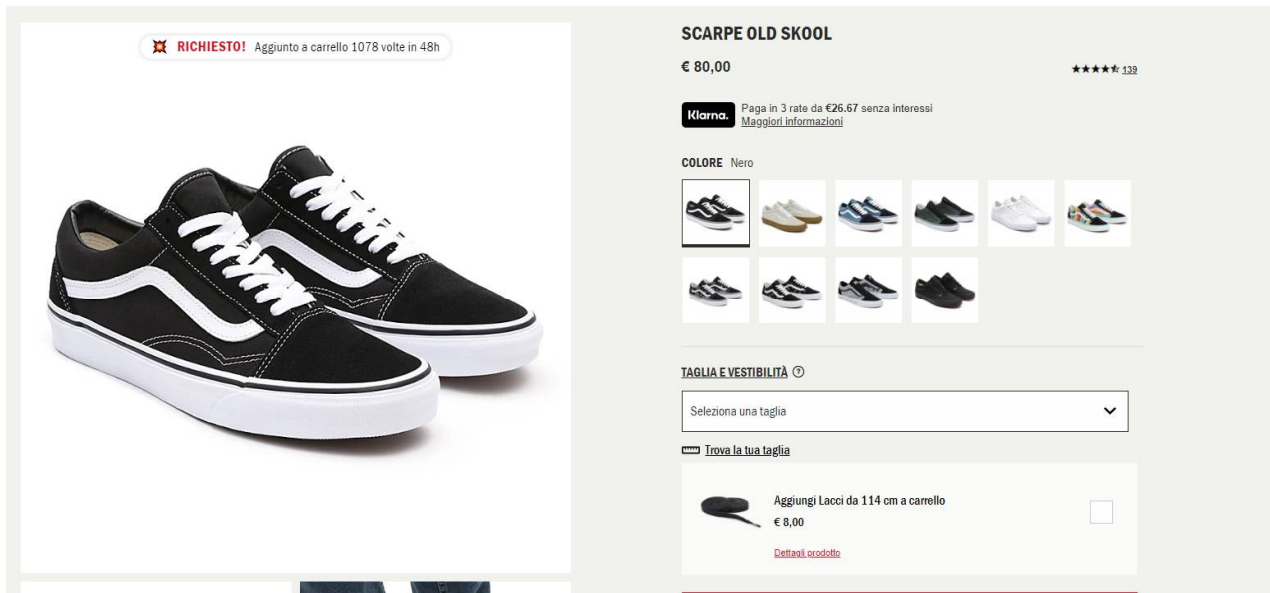
Both platforms have similar **product presentation features**:

- Summary section with product name, price, possibility for deferred payment, total number of reviews and average rating.
→ Amazon has a discount on the proprietary price list.
- Several (zoomable) images and possibility to select backpack's color.
- Rating distribution bar chart and user reviews detail sections.
- Keywords dedicated section.
- Product description section:
→ TNF ecommerce uses several dropdown description menus
→ Amazon provides an overall description.
- Amazon allows customer to consider other product/s as alternative.

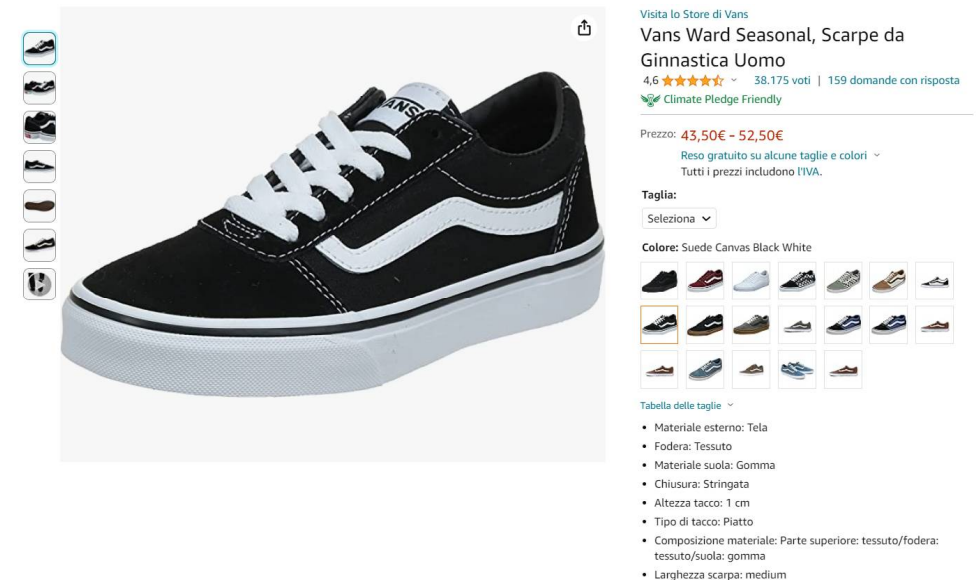


7. Data Visualization - UI/UX Analysis

Taking as example Footwear – Vans Old Skool:
TNF Ecommerce (.it)



Amazon (.it)



Both platforms have similar **product presentation features**:

- Summary section with product name, price, possibility for deferred payment, total number of reviews and average rating.
→ Amazon has a strong discount on the proprietary price list.
- Several (zoomable) images on Vans site and possibility to select different color.
- Vans site describe better the product offering also the possibility to buy customized accessory.



Recensisci questo prodotto
Condividi i tuoi pensieri con altri clienti

Scrivi una recensione cliente



Recensioni con immagini



Migliori recensioni

Recensioni migliori da Italia

Mattia

★★★★★ **Eccezionali**

Recensito in Italia il 19 agosto 2023

Taglia: 42 EU Colore: Suede Canvas Black White **Acquisto verificato**

Comprate un anno fa ed indossate quasi tutti i giorni.

La suola si è consumata veramente poco e la scarpa presenta pochissimi segni di usura dopo un anno di utilizzo. Adesso ovviamente hanno finito il loro ciclo vitale poiché sono diventate molto dure da indossare, ma la loro parte l'hanno fatta e anche molto bene.

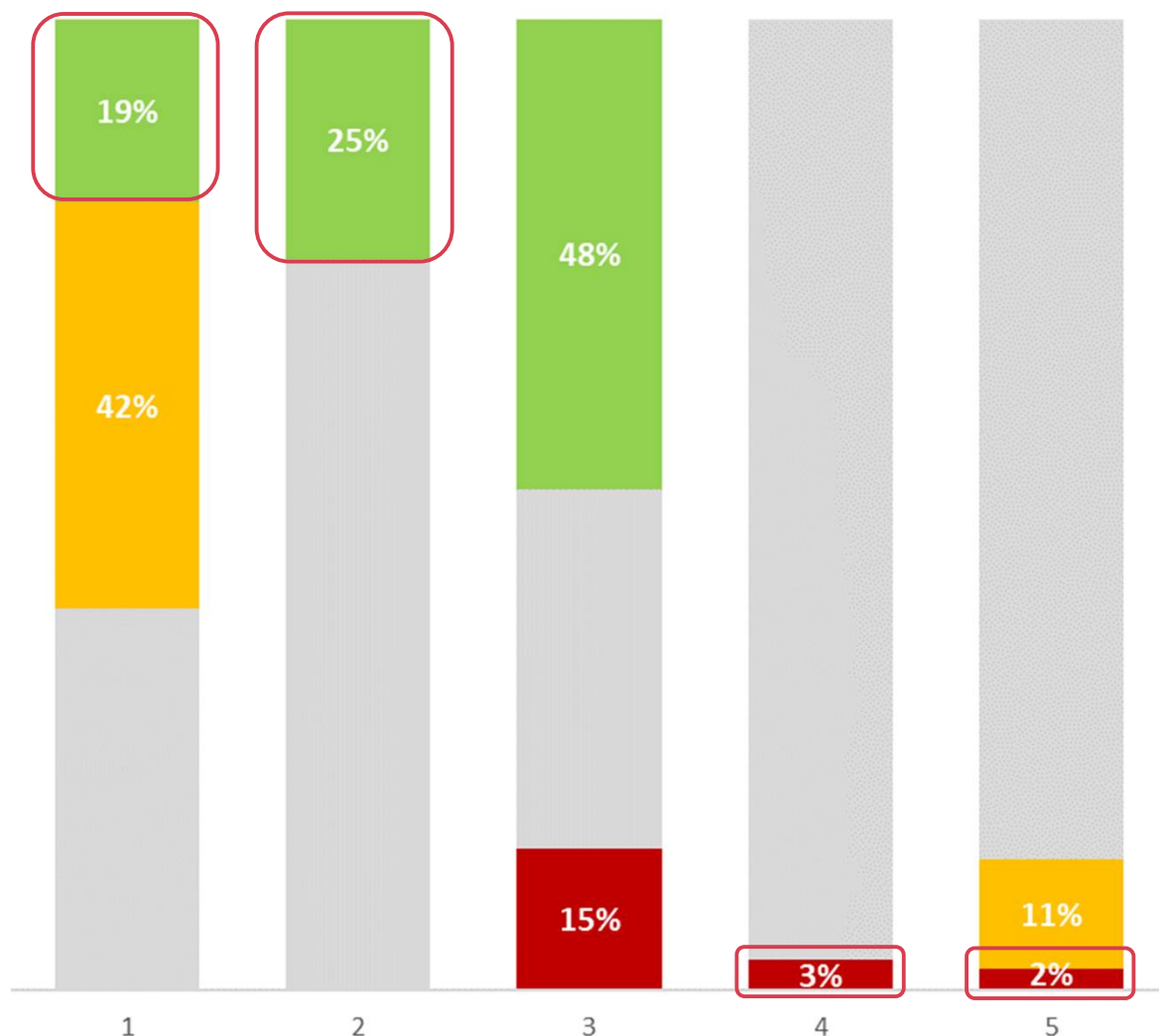
Porto un 42 e ordinandole online pensavo avessero avuto dei difetti di grandezza, invece calzano perfettamente e rispecchiano molto bene il mio numero.

Consigliate sia per ginnastica che per un'occasione un po' più elegante, stanno bene con tutti i vestiti.

Utile | Segnala

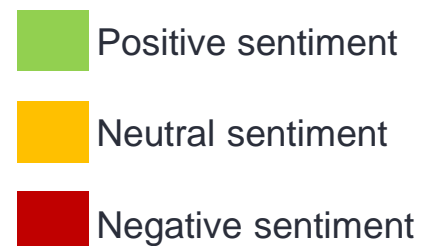
7. Data Visualization – Reviews Score

Total Reviews Sentiment results compared with Product Rating Score



Sentiment analysis methods usually do not identify sarcasm, negation, irony. Thus, it may generate some alert.

- Even if we can expect 100% of negative sentiment on low score, there is on average 20% of positive sentiment on score between 1 and 2. Nuptse product is the main affected by this warning.
- We can see that there are negative sentiment on high score (between 4 and 5). Columbia puffet jacket is the main product affected.



7. Data Visualization | Product Rating



Average Product Review Length

