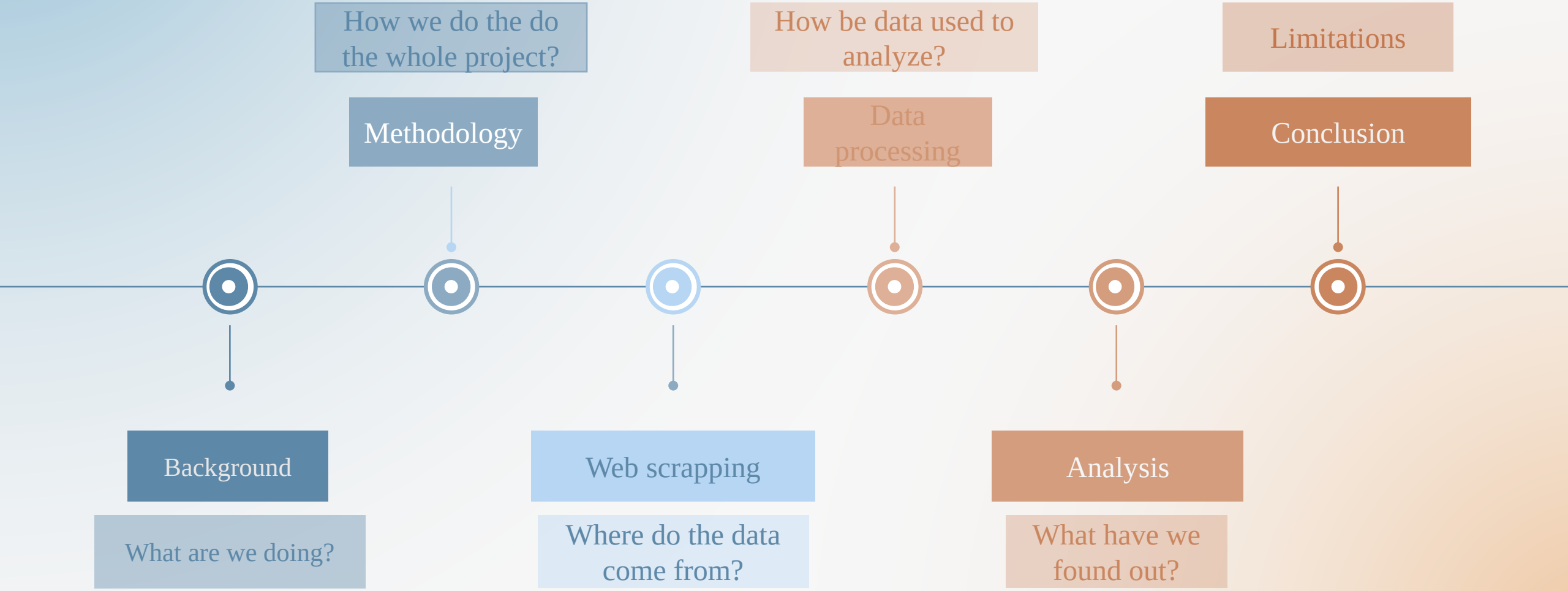
Web Scrapping Project

Group List:

Ricardo Wu
Ka Ho LI
Anmol Thadani


How we do the do the whole project?
How be data used to analyze?
Limitations
Methodology
Data processing
Conclusion
Background
What are we doing?
Web scrapping
Where do the data come from?
Analysis
What have we found out?

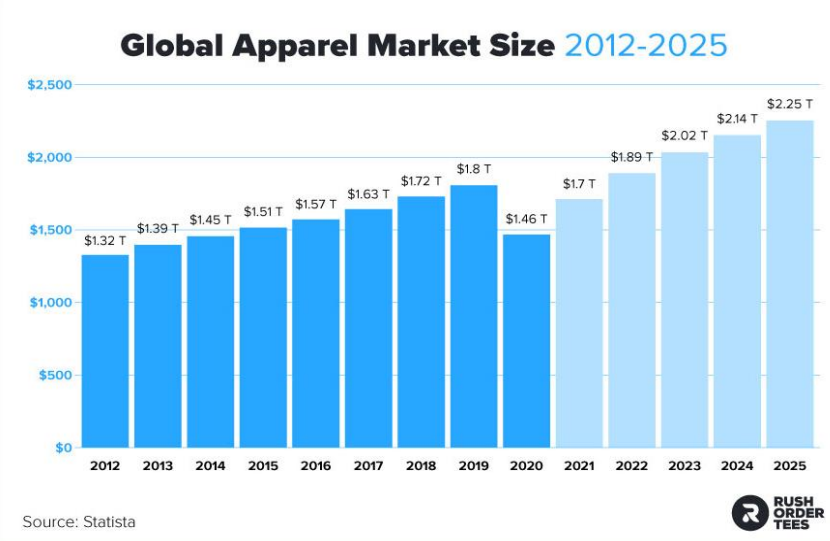
Background

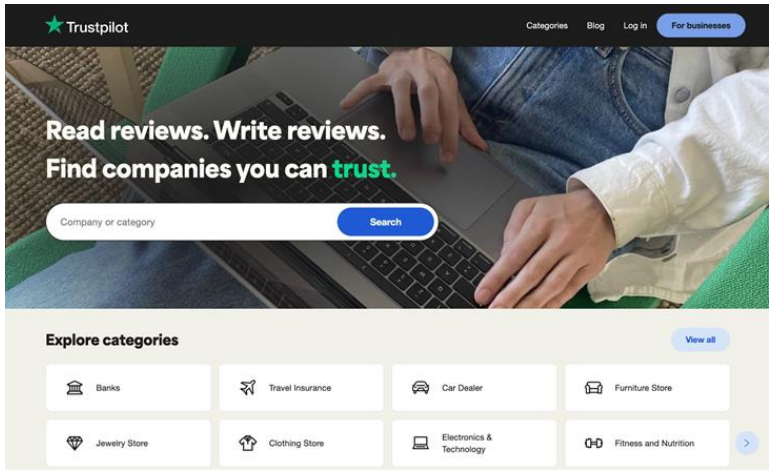
What are we doing?

Background

- Retail is always a significant part in a local economy, apparel market can be regarded as the most symbolic representative

- Situation: We are employees of a US clothing retail group and are trying to randomly collect reviews from 20 retailers with highest rating and 20 with lowest rating from an integrated platform





https://www.trustpilot.com/

Aims of the Project

1. To have a better understanding of customer feedback on the shopping experiences.
2. To improve customer services through insights from web scraping, data mining on text analysis and Statistics through below analysis

- find out best rated 20 & worst rated 20 Clothing stores
- find correlation between reviews & trust score
- do sentiment analysis through the comments
- find out what are the most mentioned keywords in general

what customers concern most?

Methodology

How do we collect the data?

Methodology - Data Preprocessing

- We scrape the data through BeautifulSoup
- By Sampling, we only scrape the best 20 and worst 20 rated companies and only if they are verified
- We classify the keywords into 3 groups based on their nature

Union 22

Reviews 505 • Excellent

4.9

VERIFIED COMPANY

Hastamuerte

Reviews 974 • Excellent

4.8

VERIFIED COMPANY

Methodology - Analysis

- The scrapped data is plain text and cannot conduct any research on it. The first step is data mining to quantify the data. TF-IDF scoring algorithm is considered too complicated; creating Chorus is technically impossible now. **Afinn** scoring algorithm is taken for simplicity

- **Afinn** is the simplest, yet most popular lexicon used for sentiment analysis developed by *Finn Årup Nielsen*. It contains 3300+ words with a polarity score associated with each word. In python, there is an in-built function for this lexicon.

- **Afinn** scoring algorithm acts similar with TD-IDF algorithm but has sentiment words preloaded, for **Afinn score > 0,** the sentence can be treated as positive comments, and vice versa. At the end, as we need to take total length into consideration for fair comparison, the **Afinn** scores were standardized through being divided by the sentence length.

Data processing

How are data used to analyze?

Data processing

The data collected is stored in 4 separated CSV files, named ***cus_good20***, ***cus_bad20***, ***best_comments and worst_comments***

- ***cus_good20*** and ***cus_bad20*** stored best 20 and worst 20 rated retailers' information
- ***best_comments*** and ***worst_comments*** stored reviews from sampled best 20 and worst 20 rated retailers

Next, the data in ***cus_good20*** and ***cus_bad20*** are used to count both types of retailers' reviews to compare the average reviews per retailer

The data in ***best_comments*** and ***worst_comments*** then are used to conduct sentiment analysis using Afinn scoring algorithm

Libraries used in data processing are: ***Afinn***, ***Re, Numpy, Matplotlib***

Analysis

What have we found out?

Companies Rating & Correlation between Trust Scores & Reviews

Insights

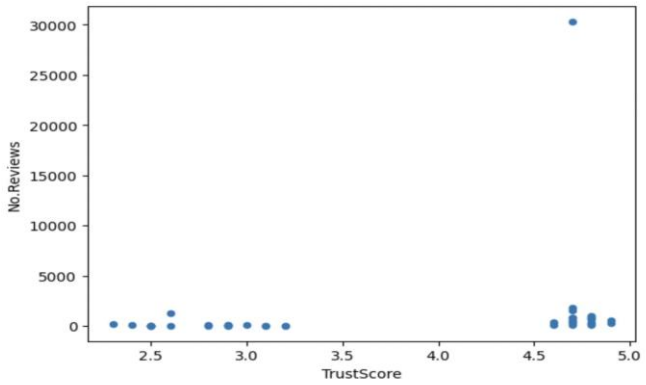
- Companies with higher Trust Scores obtain more reviews
- However, the relations are not in straight linear structure
- Correlation score is 0.199

	Company_Name	Trust_Score	No_of_Reviews	TrustScore	No.Reviews
0	Union 22	TrustScore 4.9	505 reviews	4.9	505
1	A.K. RIKK's	TrustScore 4.9	282 reviews	4.9	282
2	Hastamuerte	TrustScore 4.8	974 reviews	4.8	974
3	Otrium	TrustScore 4.8	929 reviews	4.8	929
4	Angeljackets	TrustScore 4.8	750 reviews	4.8	750
5	Lebo's	TrustScore 4.8	746 reviews	4.8	746
6	CelticClothing.com	TrustScore 4.8	370 reviews	4.8	370
7	USA Kilts	TrustScore 4.8	205 reviews	4.8	205
8	New Horizons Trading	TrustScore 4.8	60 reviews	4.8	60
9	JAXXON	TrustScore 4.7	30,316 reviews	4.7	30316
10	Tailor Store	TrustScore 4.7	1,788 reviews	4.7	1788

4 Correlation between TrustScore & Reviews

```
In [46]: print(result['TrustScore'].corr(result['No.Reviews']))
0.19934361816888282
```

```
In [48]: result.plot.scatter(x = 'TrustScore', y = 'No.Reviews')
Out[48]: <AxesSubplot:xlabel='TrustScore', ylabel='No.Reviews'>
```



Analysis – Keyword analysis

From the output we classify keywords into 3 groups

1. **Delivery**
2. **Product**
3. **Service**

By comparing keywords from B20 & W20 retailers

- **There are numbers of intersection on the keywords i.e., order, quality, received**

Out[13]:

	B20_Keywords	B20_Times	W20_Keywords	W20_Times
0	Order	111	Order	112
1	Quality	85	Received	62
2	Service	60	Get	59
3	Received	49	Service	50
4	Fit	48	Quality	41
5	Time	43	Email	31
6	Size	40	Product	27
7	Experience	38	Return	27
8	Shipping	36	Website	23
9	Fast	31	Size	23

Keywords Analysis (2)

For Best 20 retailers

- **customer admires most for delivery, they expect timely and fast dispatch**
- **for products, quality and fit sizing win customers' compliments**
- **service and experience also contribute to the good comments**

For Worst 20 retailers

- **customer complains most for delayed and wrong delivery, return is also an issue from them**
- **for products, it shares similar situation with the best 20 companies**
- **lastly for service, email reply and website guidelines may affect customers' shopping experience**

In short

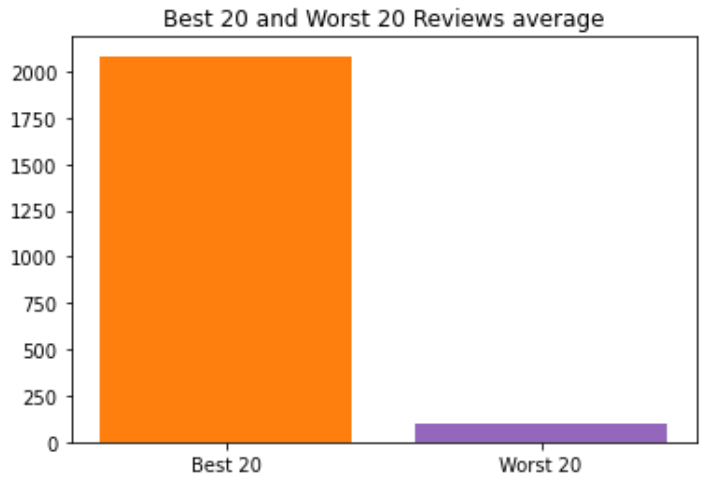
❖ By tackling the issues on orders delivery, service and product quality, we can already cater what customers concern most on their spendings

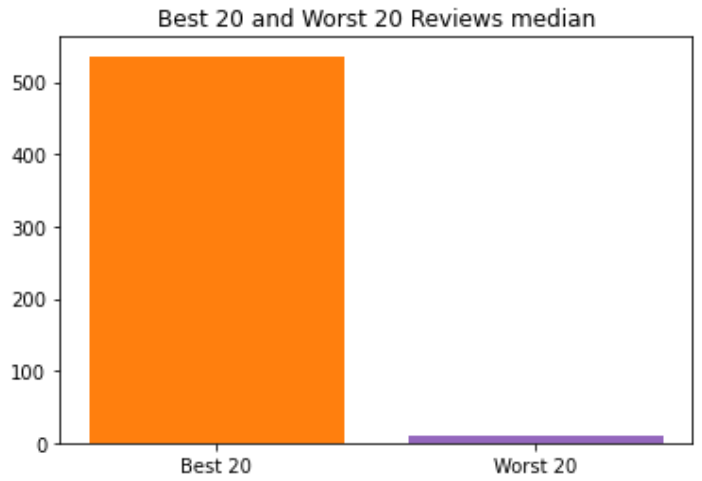
Analysis – Population data overview

There are 41708 reviews for the best 20 rated retailers in total and 1925 reviews for the worst 20 in total

For best 20 rated retailers, each retailer has 2085.4 reviews on average
While each worst rated retailers has 96.25 reviews on average

However, there are extreme outliers that have reviews more than 30000 reviews
Thus, the median of reviews number for the best 20 rated retailers is 536, 11 for the worst 20



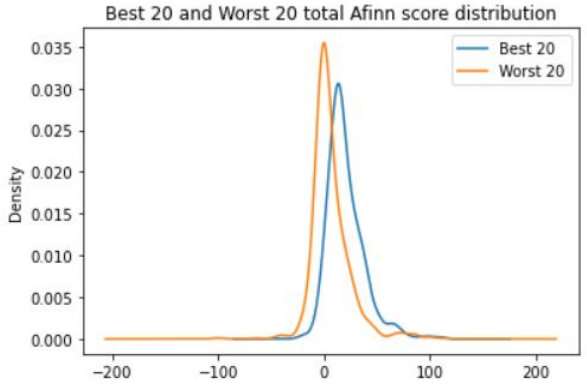


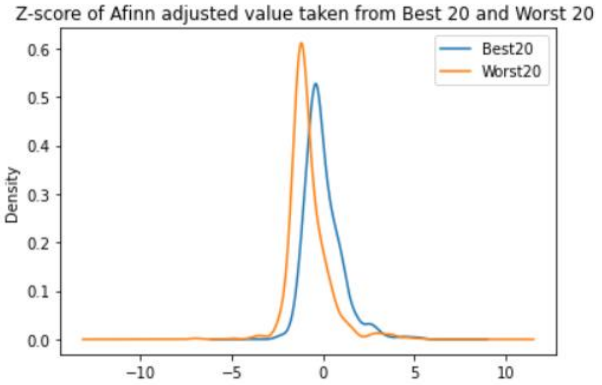
Analysis – Sentiment analysis

Through trivial sampling, we obtained 400 reviews from the best 20 retailers and 764 reviews from the worst 20 retailers, accounting for 43.2% and 7.8% in their section. Through checking the probability density, We assume they follow normal distribution:

	Comments	score	wordcount	afinn_adjusted
0	All this place does is double the price of Ama...	1.0	107	0.934579
1	The jacket purchased for my husband fits perfe...	5.0	47	10.638298
2	I received my order from Angel Jackets and hav...	12.0	46	26.086957
3	I'm a new customer but am so impressed with th...	7.0	41	17.073171
4	It is greatl ordered a leather jacked. I order...	6.0	48	12.500000
...	...	...	...	...
395	Great website and customer service. Very speed...	11.0	43	25.581395
396	Luxire has once again been the best service l'...	19.0	105	18.095238
397	My first order with Luxire-a pair of trousers ...	13.0	105	12.380952
398	Fantastic made-to-measure service. Ordered a c...	16.0	98	16.326531
399	The whole process was great and the communicat...	0.0	47	0.000000

	Comments	score	wordcount	afinn_adjusted	Z_score
0	All this place does is double the price of Ama...	1.0	107	0.934579	-1.121361
1	The jacket purchased for my husband fits perfe...	5.0	47	10.638298	-0.557934
2	I received my order from Angel Jackets and hav...	12.0	46	26.086957	0.339062
3	I'm a new customer but am so impressed with th...	7.0	41	17.073171	-0.184306
4	It is greatl ordered a leather jacked. I order...	6.0	48	12.500000	-0.449838
...	...	...	...	...	...
395	Great website and customer service. Very speed...	11.0	43	25.581395	0.309707
396	Luxire has once again been the best service l'...	19.0	105	18.095238	-0.124961
397	My first order with Luxire-a pair of trousers ...	13.0	105	12.380952	-0.456750
398	Fantastic made-to-measure service. Ordered a c...	16.0	98	16.326531	-0.227658
399	The whole process was great and the communicat...	0.0	47	0.000000	-1.175625





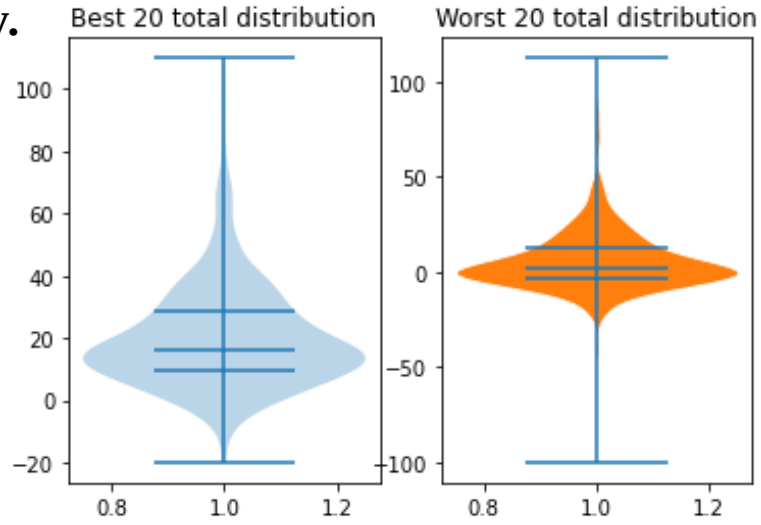
Analysis – Sentiment analysis

Best 20 retailers only have 16 negative comments while worst 20 retailers get 267 negative comments

The ratio of negative reviews to the whole data are 4% and 35% respectively

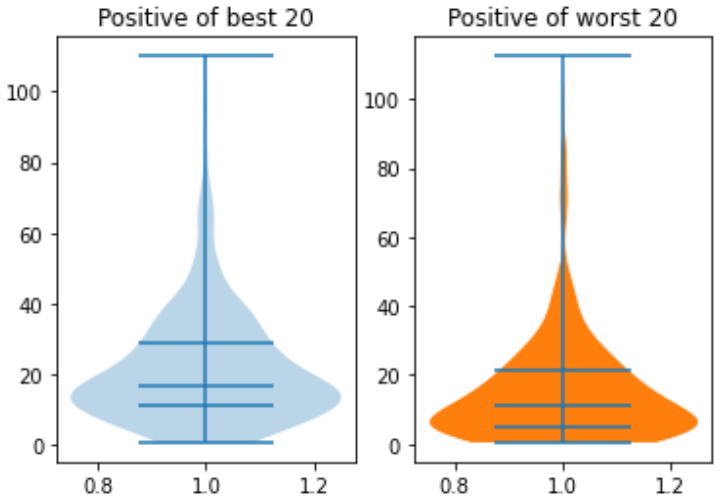
Variance of best 20 is 297, for the worst 20 is 338

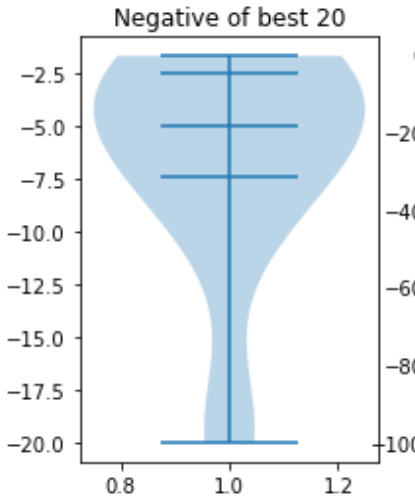
From the following Afinn score distribution, the median of Best 20 retailers is about 17%, while that of the Worst 20 is about 3%. So, customers who bought clothes from Best 20 retailers who have higher Trust Scores are more satisfactory.

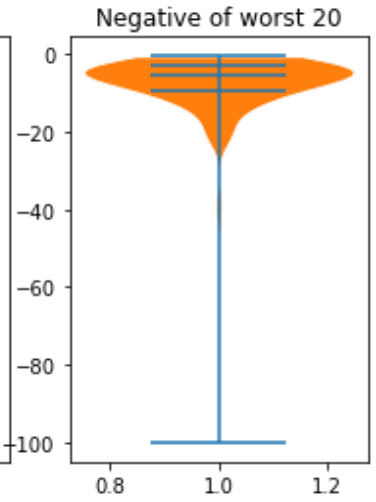


Analysis – Sentiment analysis

The positive reviews of the worst 20 retailers have more spread distribution, while the negative reviews are more centered, this shows the variation in reviews

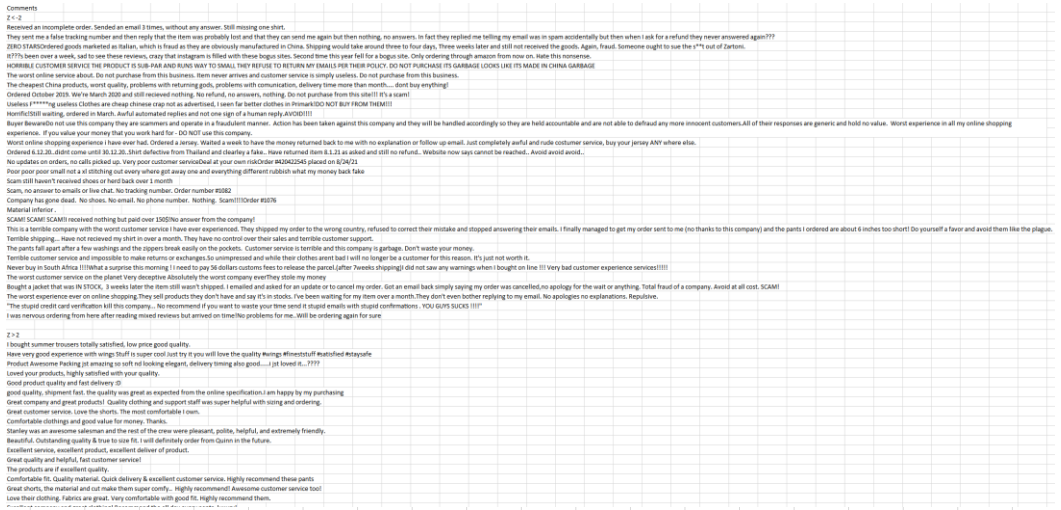






Analysis – Sampling

As we assume the sample data follows normal distribution, so we want to focus more on outliers
We extract outliers which have Z-score of adjusted Afinn score ± 2 S.D.



Comments
Z < -2
Received an incomplete order. Sended an email 3 times, without any answer. Still missing one shirt.
They sent me a false tracking number and then reply that the item was probably lost and that they can send me again but then nothing, no answers. In fact they replied me telling my email was in spam accidentally but then when I ask for a refund they never answered again???
ZERO STARSOrdered goods marketed as Italian, which is fraud as they are obviously manufactured in China. Shipping would take around three to four days, Three weeks later and still not received the goods. Again, fraud. Someone ought to sue the s**t out of Zartoni.
It???s been over a week, sad to see these reviews, crazy that Instagram is filled with these bogus sites. Second time this year fell for a bogus site. Only ordering through amazon from now on. Hate this nonsense.
HORRIBLE CUSTOMER SERVICE THE PRODUCT IS SUB-PAR AND RUNS WAY TO SMALL THEY REFUSE TO RETURN MY EMAILS PER THEIR POLICY. DO NOT PURCHASE ITS GARBAGE LOOKS LIKE ITS MADE IN CHINA GARBAGE
The worst online service about. Do not purchase from this business. Item never arrives and customer service is simply useless. Do not purchase from this business.
The cheapest China products, worst quality, problems with returning gods, problems with comunication, delivery time more than month.... dont buy anything!
Ordered October 2019. We're March 2020 and still recieved nothing. No refund, no answers, nothing. Do not purchase from this site!!!! It's a scam!
Useless f******g useless Clothes are cheap chinese crap not as advertised, I seen far better clothes in Primark!DO NOT BUY FROM THEM!!!
Horrific!!Still waiting, ordered in March. Awful automated replies and not one sign of a human reply.AVOID!!!!
Buyer BewareDo not use this company they are scammers and operate in a fraudulent manner. Action has been taken against this company and they will be handled accordingly so they are held accountable and are not able to defraud any more innocent customers.All of their responses are generic and hold no value. Worst experience in all my online shopping experience. If you value your money that you work hard for - DO NOT use this company.
Worst online shopping experience I have ever had. Ordered a Jersey. Waited a week to have the money returned back to me with no explanation or follow up email. Just completely awful and rude costumer service, buy your jersey ANY where else.
Ordered 6.12.20, didnt come until 30.12.20..Shirt deffective from Thailand and clearley a fake... Have returned item 8.1.21 as asked and still no refund.. Website now says cannot be reached.. Avoid avoid avoid..
No updates on orders, no calls picked up. Very poor customer serviceDeal at your own riskOrder #420422545 placed on 8/24/21
Poor poor poor small net a xl stitching out every where got away one and everything different rubbish what my money back fake
Scam still haven't received shoes or herd back over 1 month
Scam, no answer to emails or live chat. No tracking number. Order number #1082
Company has gone dead. No shoes. No email. No phone number. Nothing. Scam!!!!Order #1076
Material inferior .
SCAM! SCAM! SCAM!I received nothing but paid over 150$!No answer from the company!
This is a terrible company with the worst customer service I have ever experienced. They shipped my order to the wrong country, refused to correct their mistake and stopped answering their emails. I finally managed to get my order sent to me (no thanks to this company) and the pants I ordered are about 6 inches too short! Do yourself a favor and avoid them like the plague.
Terrible shipping... Have not recieved my shirt in over a month. They have no control over their sales and terrible customer support.
The pants fell apart after a few washings and the zippers break easily on the pockets. Customer service is terrible and this company is garbage. Don't waste your money.
Terrible customer service and impossible to make returns or exchanges.So unimpressed and while their clothes arent bad I will no longer be a customer for this reason. It's just not worth it.
Never buy in South Africa !!!!!What a surprise this morning ! I need to pay 56 dollars customs fees to release the parcel.(after 7weeks shipping)I did not saw any warnings when I bought on line !!! Very bad customer experience services!!!!!
The worst customer service on the planet Very deceptive Absolutely the worst company ever!They stole my money
Bought a jacket that was IN STOCK, 3 weeks later the item still wasn't shipped. I emailed and asked for an update or to cancel my order. Got an email back simply saying my order was cancelled,no apology for the wait or anything. Total fraud of a company. Avoid at all cost. SCAM!
The worst experience ever on online shopping.They sell products they don't have and say it's in stocks. I've been waiting for my item over a month.They don't even bother replying to my email. No apologies no explanations. Repulsive.
"The stupid credit card verification kill this company... No recommend if you want to waste your time send it stupid emails with stupid confirmations . YOU GUYS SUCKS !!!!"
I was nervous ordering from here after reading mixed reviews but arrived on time!No problems for me...Will be ordering again for sure

Z > 2
I bought summer trousers totally satisfied, low price good quality.
Have very good experience with wings Stuff is super cool Just try it you will love the quality #wings #lineststuff #satisfied #staysafe
Product Awesome Packing jst amazing so soft nd looking elegant, delivery timing also good.....I jst loved it....????
Loved your products, highly satisfied with your quality.
Good product quality and fast delivery :D
good quality, shipment fast, the quality was great as expected from the online specification.I am happy by my purchasing
Great company and great products! Quality clothing and support staff was super helpful with sizing and ordering.
Great customer service. Love the shorts. The most comfortable I own.
Comfortable clothings and good value for money. Thanks.
Stanley was an awesome salesman and the rest of the crew were pleasant, polite, helpful, and extremely friendly.
Beautiful. Outstanding quality & true to size fit. I will definitely order from Quinn in the future.
Excellent service, excellent product, excellent deliver of product.
Great quality and helpful, fast customer service!
The products are if excellent quality.
Comfortable fit. Quality material. Quick delivery & excellent customer service. Highly recommend these pants
Great shorts, the material and cut make them super comfy.. Highly recommend! Awesome customer service too!
Love their clothing. Fabrics are great. Very comfortable with good fit. Highly recommend them.

Comments
>2
Good quality jacket true to size. Kept informed from order to delivery and arrived in good time.Lovely jacket have had lots of compliments. Would recommend.
I love their products, they are beautiful, fast delivery and comfortable to wear
Excellent service great quality clothes fast delivery and packaged with a touch of luxury.
High quality, perfect garments at a very competitive price. Quick, efficient service. Thank you
Amazing brand, beautiful quality, and a price that makes the items feel quality, fire ????
Hasta Muerte hands down the best qualityproducts on the market. Products fit wonderfully and perfect for any occasion. I am 14 shirts in and still get compliments in public. Great brand not only for the amazing quality, but the excellent customer service!
Awesome prices on awesome products; super fast shipping; I highly recommend this online store!
Easy, quick, fast loved my item and loved the quick shipping
Exactly as described. Very fast shipping. Good, quality clothing. Thanks
I am highly recommend this store, very good quality and excellent attitude to the customers
Got my shoes quickly and they were good quality
The service is outstanding and the clothing is wonderful. You can trust these people!
always on time, great quality!!
A great shopping experience, as usual. Your sales are amazing. Thanks.
Nice clothes, nice store, nice customer service! Will be coming back!
Superb quality and fantastic service (e.g. quick shipment turnaround).
<-2
Quality shorts. Easy exchange
I initally gave a 5 Star review but this rating system is to difficult so I chaged it to worst rating
They advertise the wrong color for their golden brown timberland pro boots. Aka they lie to their customers. Extremely disappointed.

Conclusion

Limitations

Challenges and Limitations

Challenges

- There are values that we have to scrape with different class names due to no content inside
- Some worst-rated companies only have good reviews which makes it quite confusing
- Existence of extreme outliers have much more reviews makes it hard to categorize
- Unrevealed data (No trust score of every user shown on the website)

Limitations

- There are too many reviews, we can only do random sampling for that (Time cost)
- It takes a more advanced skillset to better interpret comments & reviews
  i.e. there are negative comments that get positive Afinn score (Algorithm limitation)
- Knowledge limitation on Inferential Statistics

Conclusion

It is hard to scrap data from other websites, but if we have our own websites, we can get complete data so that only need to do a little data cleaning for the noises or incomplete data. This time the data follows the normal distribution, but what if next time it's not?

For big data sets, algorithm validity is of paramount importance which can greatly reduce the human resources needed and efficiently categorize data for analysis.

We hope can do better after learning data mining, deep learning and machine learning.

Company_Name	Trust_Score	No_of_Reviews	Reviews topics type/ Keyword	Comments	Individual Trust Score
A	4.9	505	Transaction	.....	...
			Staff Attitude		
			Goods quality		
			Payment		
B	4.8	282	Transaction		...

Conclusion

Considering the analyses we did, we have below insights:

- The higher trust scores & reviews a company obtains , the more satisfied customers are
- Best 20 companies seldom get negative feedback from customers while there are higher portion of negative feedback on worst 20 companies
- Orders delivery, service and product quality are the top 3 concerns from the customers, if we can make sure of these, we get happy and satisfied customer

Q&A

End