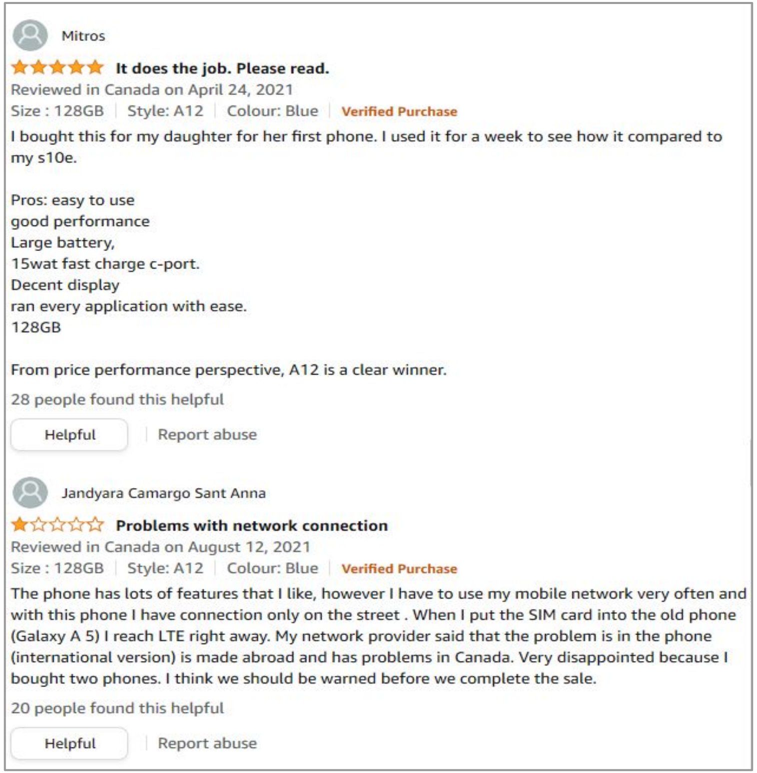
ML1010 Group Project Milestone 1

Al, Mike, Oana

Voluminous Raw Data



Mitros

★★★★★ **It does the job. Please read.**

Reviewed in Canada on April 24, 2021

Size : 128GB | Style: A12 | Colour: Blue | Verified Purchase

I bought this for my daughter for her first phone. I used it for a week to see how it compared to my s10e.

Pros: easy to use
good performance
Large battery,
15wat fast charge c-port.
Decent display
ran every application with ease.
128GB

From price performance perspective, A12 is a clear winner.

28 people found this helpful

Jandyara Camargo Sant Anna

★☆☆☆☆ **Problems with network connection**

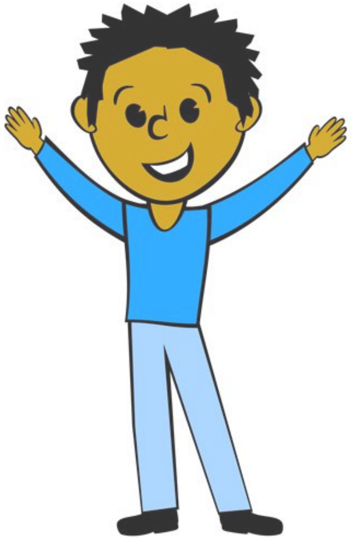
Reviewed in Canada on August 12, 2021

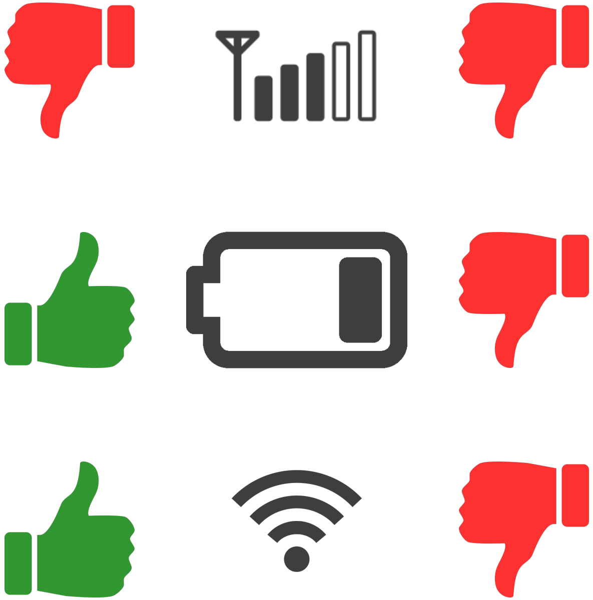
Size : 128GB | Style: A12 | Colour: Blue | Verified Purchase

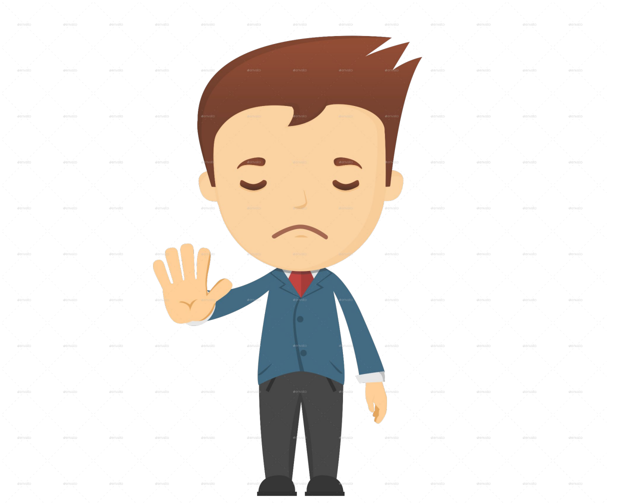
The phone has lots of features that I like, however I have to use my mobile network very often and with this phone I have connection only on the street . When I put the SIM card into the old phone (Galaxy A 5) I reach LTE right away. My network provider said that the problem is in the phone (international version) is made abroad and has problems in Canada. Very disappointed because I bought two phones. I think we should be warned before we complete the sale.

20 people found this helpful

Problem: Roadmap Planning







Narrowing down the data to clarify the problem definition

	Raw data source	In scope
Documents	1,128,437	63,461
Main Categories	29	1
Categories	133	1

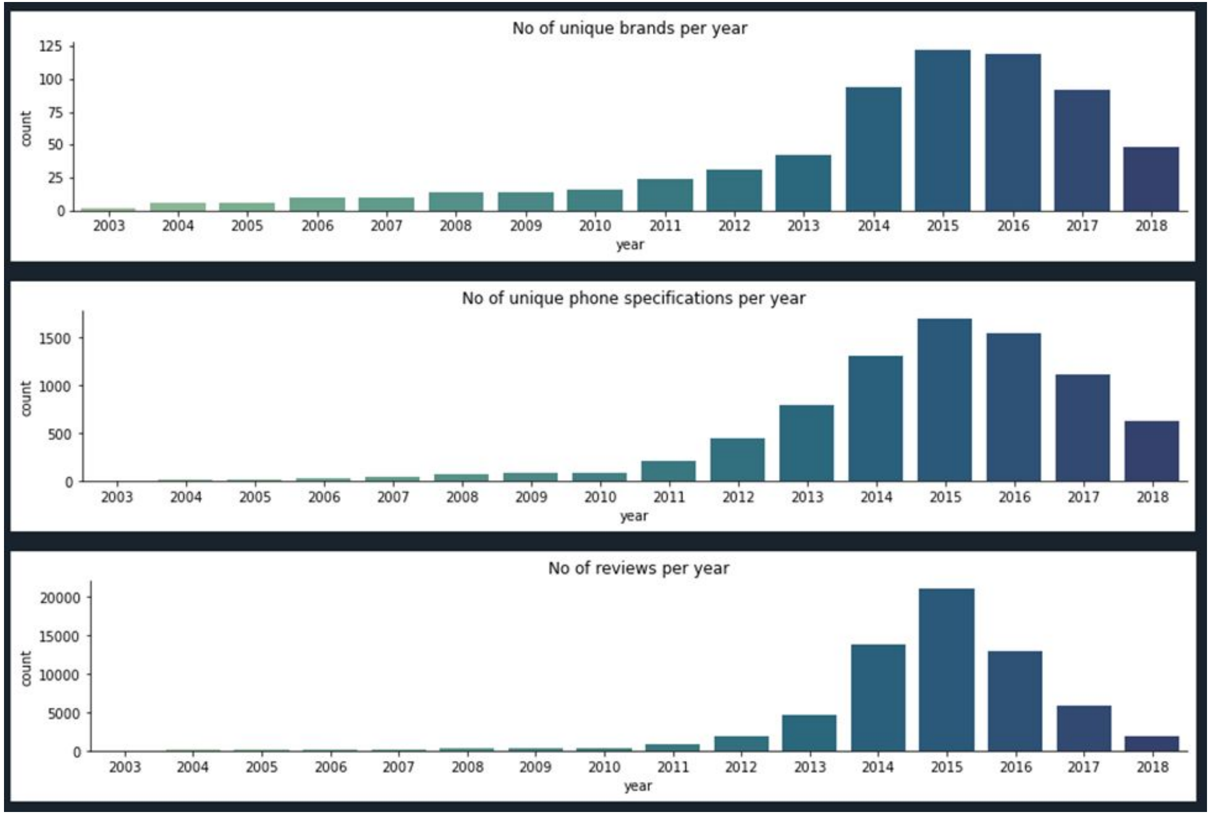
Problem definition

- The **primary** use case is to assist manufacturers in determining which features are important and well done, and which features are not viewed positively so they may be improved in the next iteration. As such, based on user reviews text we can explore what features/issues/topics are making it a high rating and/or low rating product.

- A **secondary** potential use case is determining comparable phones of same or different brands for a better brand positioning in a certain phone type class.

- A **third** potential use case is to see if a pre-trained sentiment analysis tool performs well in estimating the review text sentiment to predict the overall rating and explore if it is possible to train our own, cell phones specific, sentiment model.

Descriptive stats

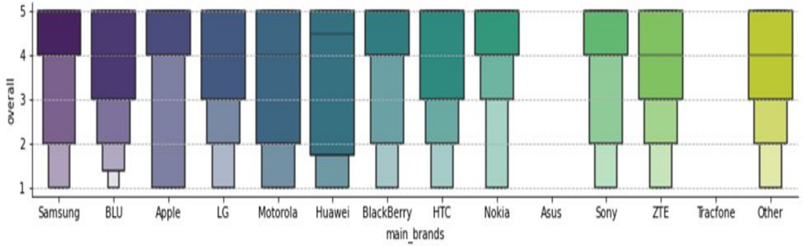


The data has a time series component, both in terms of observations available as well as the phone specifications and available reviews, reflecting the cell phone industry technological and brand name evolution

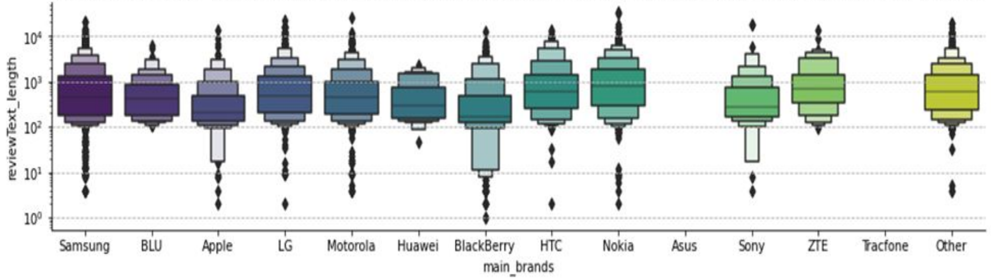
Descriptive stats

Across time the overall ratings as well as the review length distributions have changed

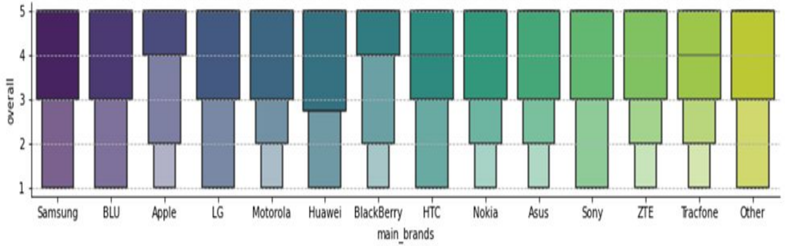
Overall ratings distribution, 2013 & before (optimized number of quantiles, see https://vita.had.co.nz/papers/letter-value-plot.pdf)



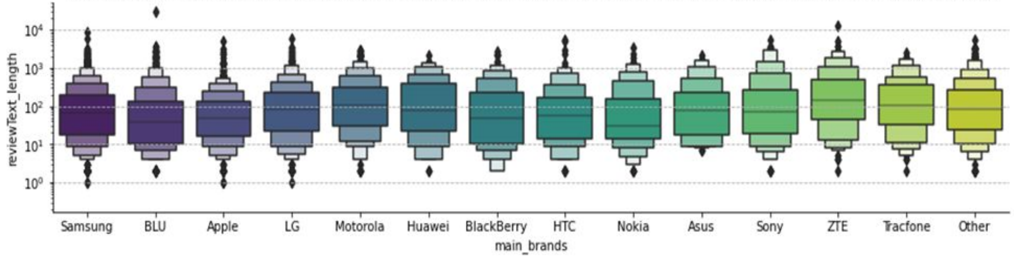
Overall review length distribution, 2013 & before (optimized number of quantiles, see https://vita.had.co.nz/papers/letter-value-plot.pdf)



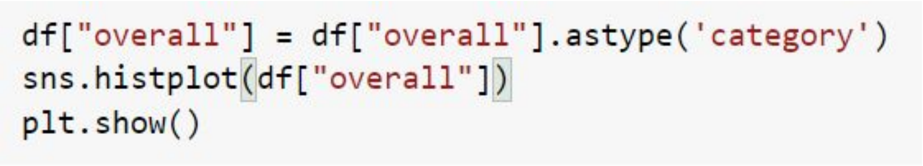
Overall ratings distribution, 2017 & after (optimized number of quantiles, see https://vita.had.co.nz/papers/letter-value-plot.pdf)



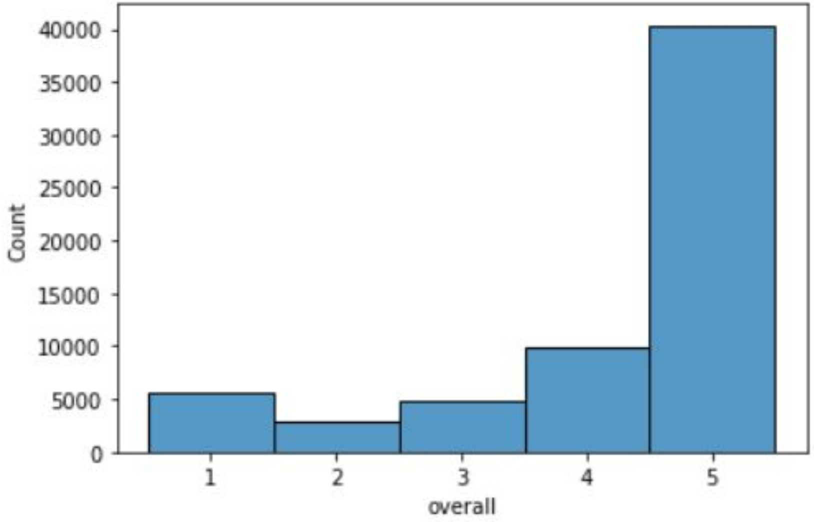
Overall review length distribution, 2017 & after (optimized number of quantiles, see https://vita.had.co.nz/papers/letter-value-plot.pdf)



What did we find?



```
df["overall"] = df["overall"].astype('category')
sns.histplot(df["overall"])
plt.show()
```

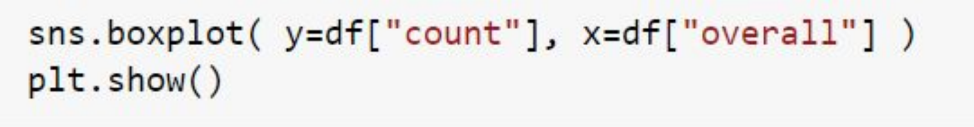


Reviews are highly unbalanced

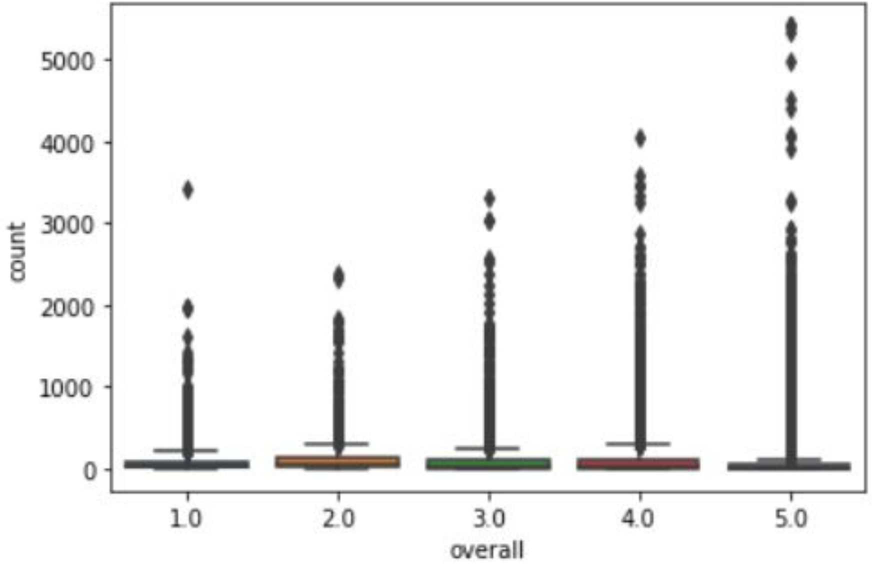
- Almost 50% are 5-star
- About 8% 1-star

Issues for training, and our intention to mine both positive and negative reviews

What did we find?



```
sns.boxplot( y=df["count"], x=df["overall"] )
plt.show()
```



Size of Reviews varies “astronomically”

- Median Review 18 words long
- Some reviews up to 5000 words in length!

This will be an issue for model training

Where do we go from here?

- Explore text variables in depth:
  - Extract topics from review text
  - Break down “parts of speech” for product features (topics)
  - Evaluate short text column “Summary”
  - How to categorize reviews that have less than 30 letters(they are usually non informative, usually)
- Explore non-text variables further
  - Evaluate if reviewer vote/rating is useful
  - Explore if price is useful(missing data issue)
- Make some filtering decisions given:
  - too large time span (retain 2017 & 2018)
  - Retain best represented brands
- Strategy for dealing with sample imbalance

Thank You