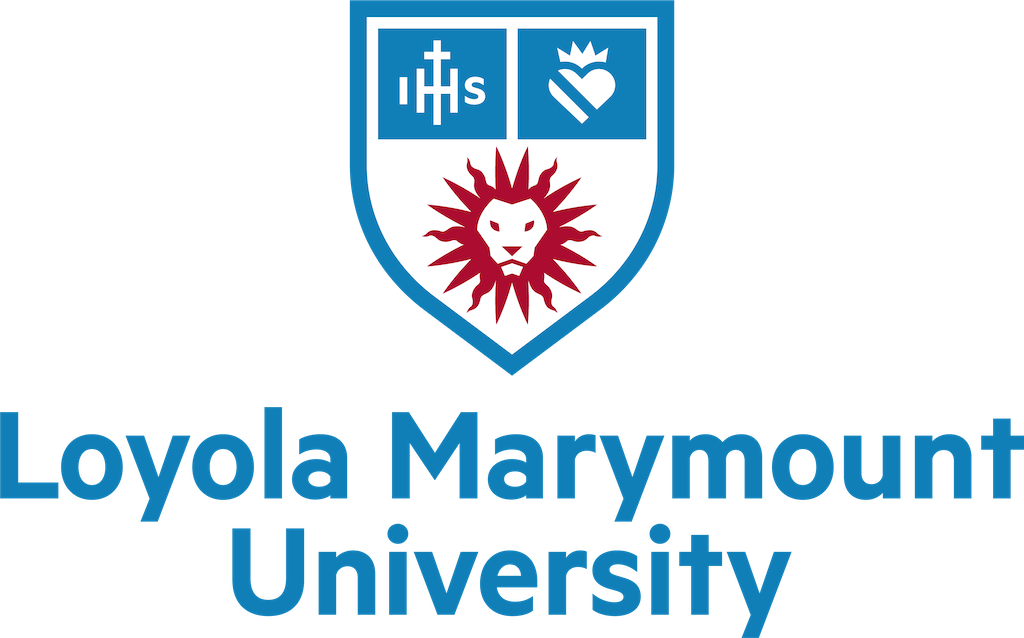
Analytics on   
Consumer Value Dynamics



Carla Fera, Elizabeth Hyde, Vincent Chen

Mentor: Zhen (Richard) Tang

Business Partner: RPA

# Executive Summary

Consumer behaviors are directly impacted by their values, which might be changed a lot under uncertainties like the COVID-19 situation. Marketers and business professionals around the world all hope to understand the possible value changes on consumers so that they can adapt their business strategies to current consumer values. In this study, we examined the dynamics of five consumer values amid the pandemic from Amazon reviews: Social Connectedness, Thrift/Value, Status, Individuality, and Health.

In order to analyze these values, we used the BERT method for preprocessing the review data, transforming them into numerical vectors. After training a review dataset with value labels, we built a neural network to fit the training dataset and make our value predictions on unlabelled reviews. Certain analyses like time series analysis, logit model, and clustering were performed to investigate and understand the value dynamics. We found that the pandemic is positively influencing the social connectedness and health, and companies should revisit those values later to see if the pandemic effect would persist in the future.

# 

# Table of Contents

[**Executive Summary**](#_rkp26yv2oodl) **2**

[**Table of Contents**](#_iv1diu7xznwq) **3**

[**Introduction**](#_ou8df1vavckf) **4**

[RPA](#_bu7jq0dywzkx) 4

[Background & Research Question](#_4rjbkybq58lu) 4

[**Solution Framework**](#_kfr0mairkvw7) **6**

[**Dataset Development**](#_gd9u681ydq9r) **7**

[**Method: BERT + Neural Network**](#_3nix91bqubbi) **9**

[**Findings**](#_8cly8wqbbylh) **12**

[Model-free Findings](#_frb8mvudn2nh) 12

[Model-based Findings](#_7x11pvvdw56k) 16

[**Complimentary Analysis of UGC**](#_927lqmp32oii) **21**

[**Business Implications**](#_veoux35dsfr7) **23**

[**Limitations**](#_8xa43eh7e54k) **23**

[**References**](#_cc75fhpawfuw) **25**

# 

# Introduction

## RPA

RPA is a leading, independent, and full-service marketing agency. They value people-first, building long-lasting relationships, and achieving solid results. Some brands they have partnered with include Honda, Apartments.com, Tik Tok, and Southwest, but what truly sets them apart is their high-quality research on consumers. RPA is full of experts in the consumer decision journey. For Farmers Insurance, they combined qualitative research, Bayesian modeling, and insights on customer segmentation with consumer values like intelligence and diversity, which resulted in increased brand awareness and sales.

## Background & Research Question

For marketers, the most important question to consider right now is how the pandemic is shifting the marketing landscape. The socioeconomic environment is causing a great deal of uncertainty. People across the world have been experiencing the effects of the current coronavirus pandemic. In the United States, cases are still climbing and restrictions are constantly changing. Many businesses have been closed for months. Millions of people do not have income, which is contributing to the uncertainty. The US uncertainty index, developed by McKinsey and Company, is at its highest level in 35 years (Smit, 2020). This uncertainty has already caused a shift in consumer spending patterns. People are spending more on groceries, household supplies, and entertainment streaming while delaying purchases like apparel, and luxury items (Charm, 2020). Likewise, the uncertainty has had an effect on consumer attitudes about physical distance, privacy, health, and financial security (Charm, 2020).

We are looking to determine why behavior has changed. Specifically, we are seeking to understand how uncertainty affects consumer values. It is extremely valuable for companies to understand what people really care about. They can use consumer values to influence their marketing messages, inform media placement, and impact customer segmentation and positioning strategies.

In this study, after consulting with RPA, there are thirteen consumer values considered: Social Connectedness, Environment, Self-Sufficiency, Transparency/Authenticity, Tradition, Individuality, Diversity/Equality, Privacy, Status, Thrift/Value, Innovation, Fun/Adventure, and Health.

# Solution Framework

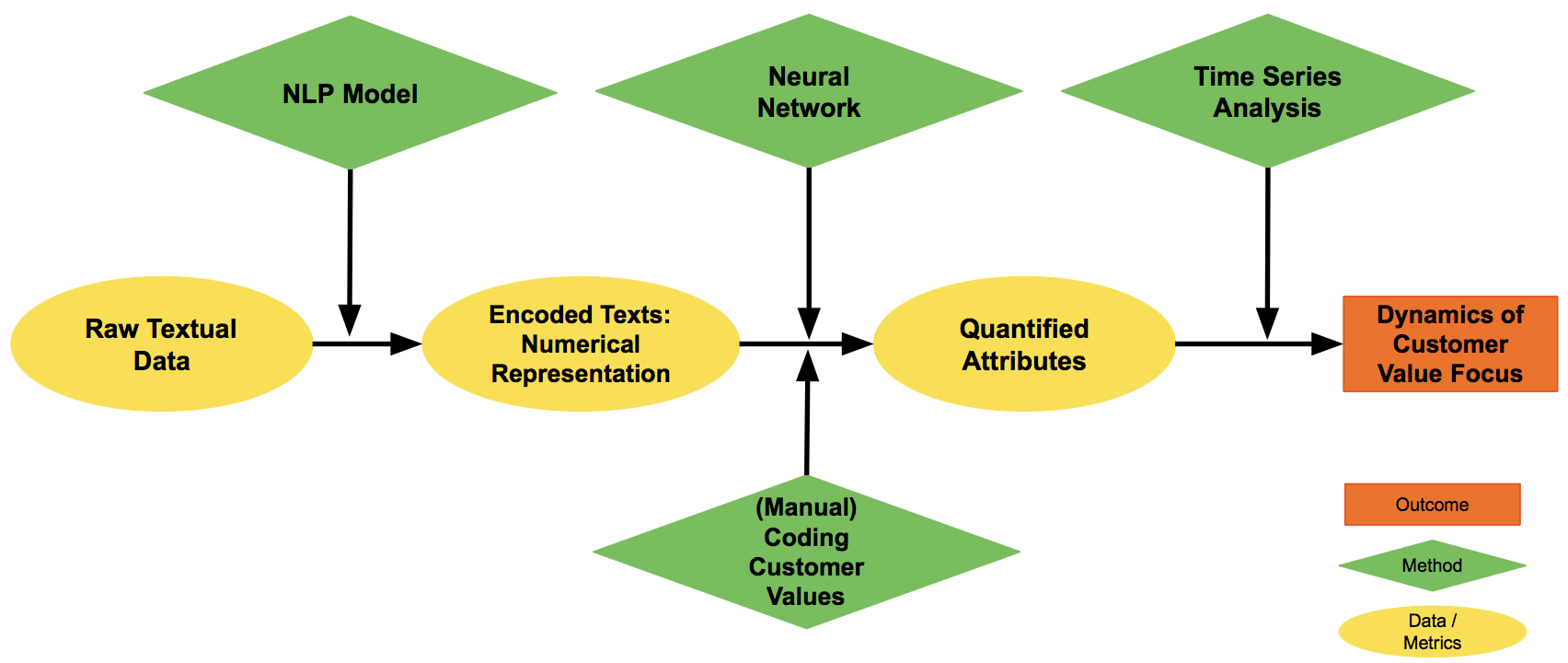


Chart 1, Solution Framework

Chart 1 above visualizes the process we followed to study consumer values. We started with the raw text data, which is the User Generated Content. The text data was put through a Natural Language Processing (NLP) model, BERT, which outputs the encoded text. We had some of the reviews manually coded 1,600 reviews for the thirteen different customer values beforehand as a training set and fed it to a neural network to train on the quantified attributes. Based on the trained neural network model, we were able to predict the values for more than 130,000 unlabeled reviews. Finally, we performed some analysis techniques like time series analysis to understand the dynamics of customer values over time.

# Dataset Development

We chose Amazon Reviews as the data source for our research. In summary, Amazon Reviews have the following main characteristics, which would coincide with our ideal data attributes:

* User-Generated Content
* Consumption Related
* Longitudinal
* Time + Cost Efficiency
* Scalability
* Reflect Values
* Product Diversity

Chart 2 below describes all the characteristics of the different possible data sources:

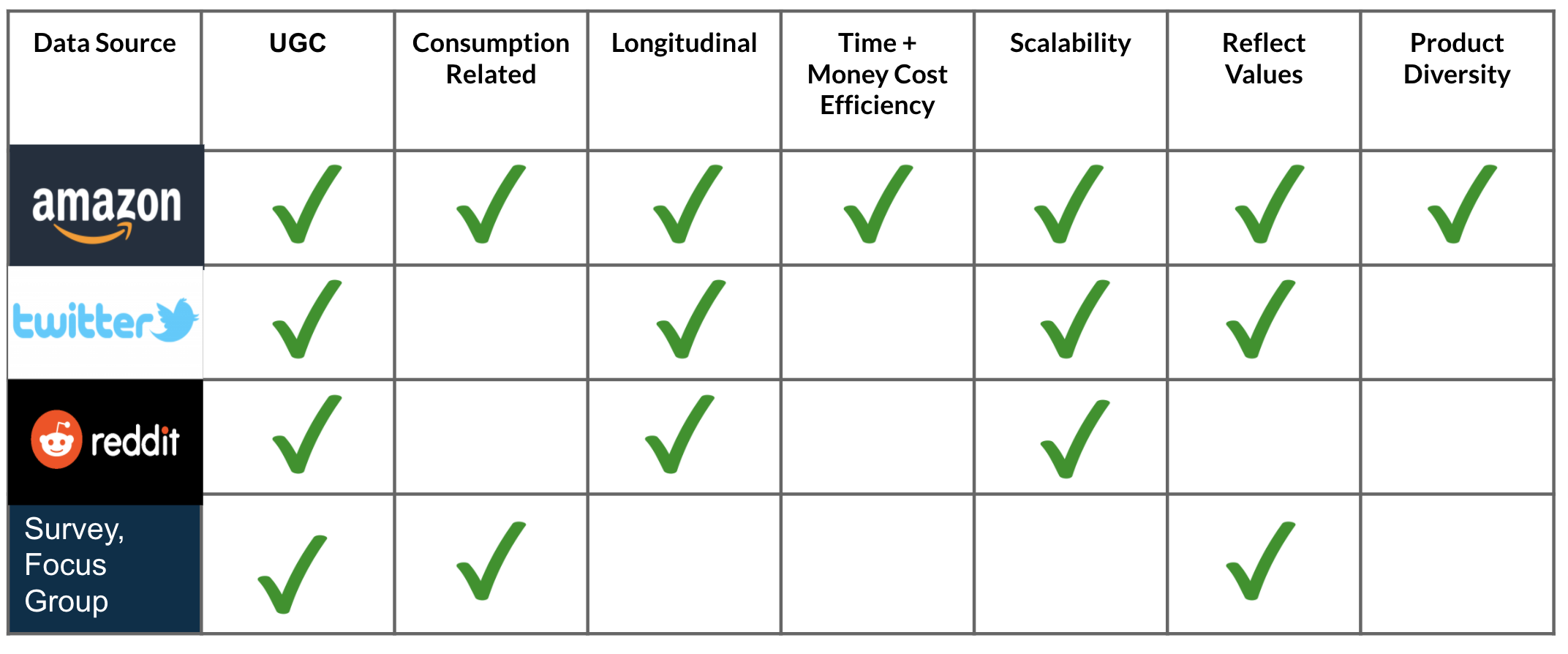


Chart 2, Comparison of the data sources

As our data source was defined, we used a publicly available dataset of Amazon reviews from 2000 to 2015 to train our model. A total of 1,600 reviews were manually labeled on the values that consumers are pursuing. For example, if a review is implying a specific value like status, we would label “1” on the “Status” attribute for this review, otherwise we would label it as “0”. To ensure the coding quality, we labeled some of the reviews with our clients. During the meeting, we set up some coding rules which helped in the consistency of the coding. Cross-validation was also performed within our team to achieve the highest quality of the training dataset.

After having our training dataset ready, we collected more than 130,000 reviews ( mainly from 2019 to 2020) from the amazon website, which was used to predict our values in our analysis. They were from the following categories: Apparel, Electronics, Health & Personal Care, Home Improvement, and Software. Based on RPA’s expertise and experience, these categories could provide insights with some level of generalisability and they are implicitly relevant to some of the values. For example, it would be interesting to see how home improvement purchases are discussed during the pandemic.

It is important to mention that due to the sparsity of our data, we set a threshold of 5% of the total reviews. It means that only consumer values equal to or above this threshold are qualified for training, since very sparse training values will probably hurt the performance of our model. Based on the statistics from Chart 3 below, the values chosen by us are Social Connectedness, Individuality, Status, Thrift/Value, and Health.

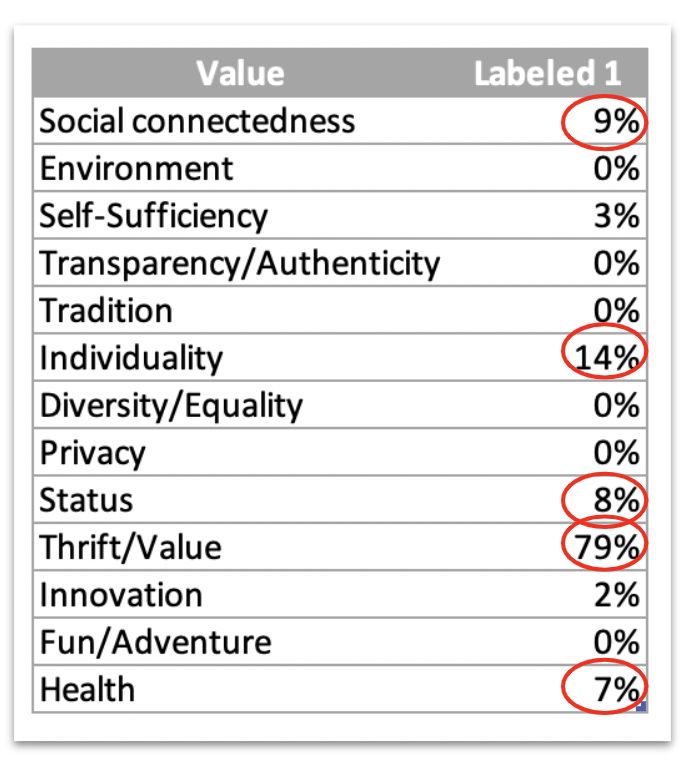


Chart 3, Selection of consumer values

# Method: BERT + Neural Network

The main method used in our analysis is BERT and Neural Network. Compared to the Neural Network, BERT might be more unfamiliar for researchers. BERT is the encoding system of the transformer which was published in 2018 by Jacob Devlin and his colleagues from Google. Being trained on a large corpus of words from Wikipedia and many books, BERT is powerful at taking raw inputs of words and converting them into contextualized numerical vectors.

The most important advantage of BERT is that it can take a good grasp on the semantic meaning among the words. Moreover, BERT has many heads so it can read the words with different perspectives to find out multiple meanings of words. Finally, it is convenient to transfer our textual data to be used for other machine learning tasks like classification and prediction.

The process to implement our model is the following:

1. Create the Dataset Instance and the DataLoader. We have to preprocess our data to fit BERT’s requirements. Then BERT transforms the textual data into the numerical representations. The DataLoader’s main function is to feed the data into the neural network with smaller batches. Additionally, the data needs to be split into the training dataset and validation dataset, which in our case was 0.9 and 0.1 respectively.
2. Create the Neural Network using the nn.Module in PyTorch. Here we need to specify the type of layers we want in our network and the type of activation function.
3. Train the model with a loop. We have to train our model using a loop to feed the data to the neural network into batches, incorporating the calculation of the loss and backpropagation to update weights and improve performance.
4. Evaluate the model with performance metrics. At this stage we would calculate out some performance metrics such as Accuracy, F1-Score, and AUC to assess the model.

|  |  |
| --- | --- |
| Typical Model Specification | NN (Neural Network) Structure |
| Epochs: 5  Training Batch size: 32  Validation batch size: 8  Learning strategy: adamW  Initial learning rate: 10e-5  Max sentence length: 128 | * BERT encoder layer * One fully-connected linear layer (768, 1) * Dropout rate: (0.3 - 0.7) * Activation function: ReLU |

Chart 4, Parameters of a Typical BERT/Neural Network implementation

In Chart 5 below, we can see the performance of our model on the value of Thrift. Each model has its training process, which means the performance will vary.

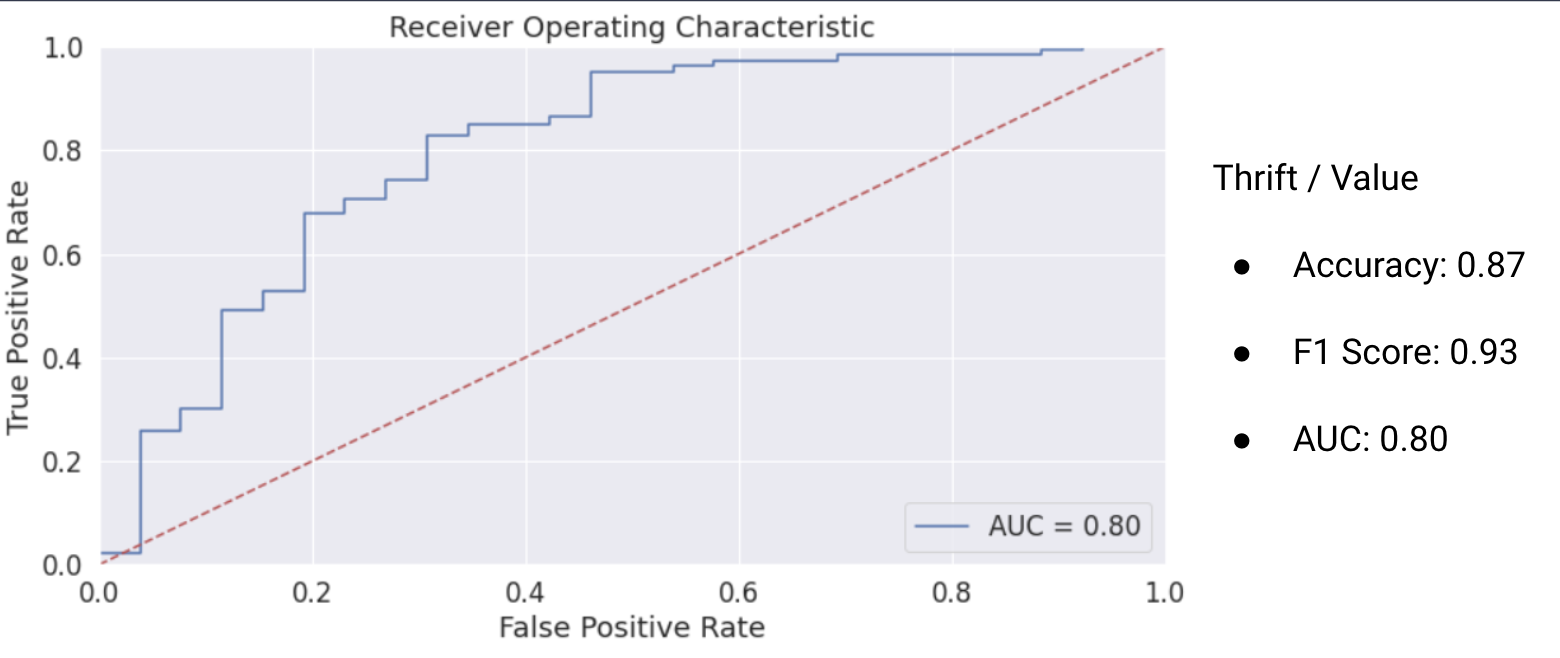


Chart 5, Model performance on Thrift

# Findings

In summary, we have run three types of analysis. The first one is on the time series analysis on different values. We would see how they have changed since last year and what are the current trends about these values right now. The second one is the logit model, with which we could see which factors are significant in affecting those value dynamics. The last one is the cluster analysis, which could be used to analyze how customers are different by grouping them together.

## Model-free Findings

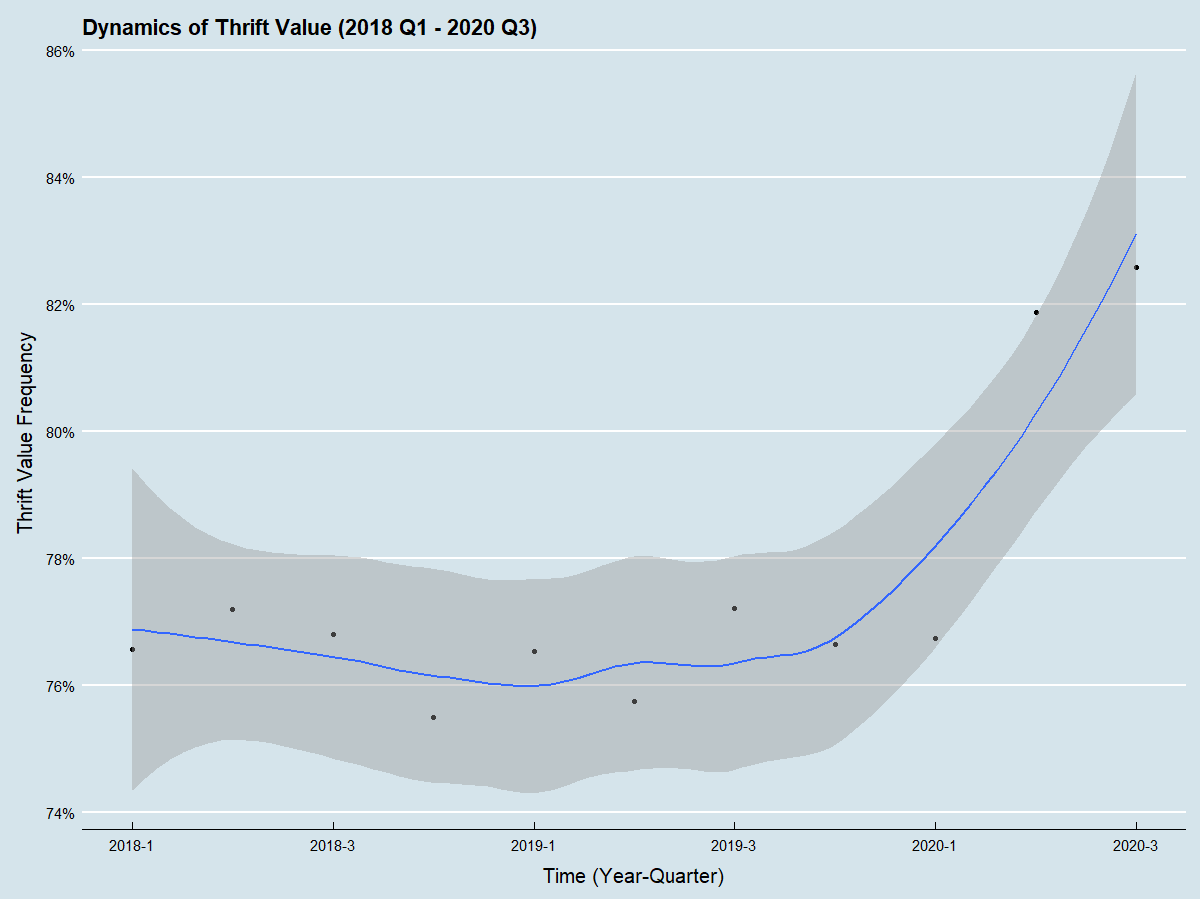


Chart 6, Dynamics of Thrift from 2018

Chart 6 indicates that the thrift value has been increasing from the fourth quarter of 2019. Based on the recent high levels of thrift, consumers now may care about thrift more than ever.

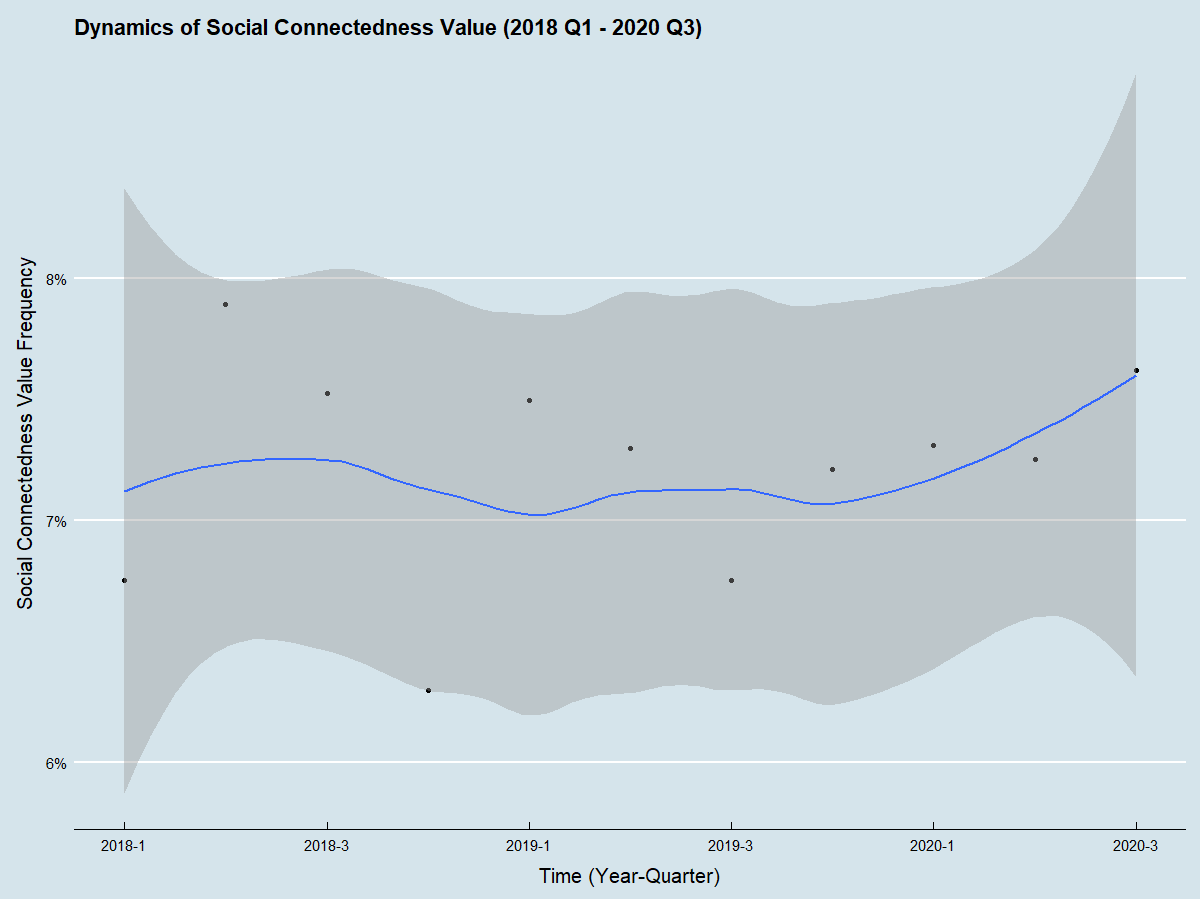


Chart 7, Dynamics of Social Connectedness from 2018

Chart 7 has not shown a very significant trend for social connectedness since 2018, just a little increasing during the pandemic. It seems like the consumers would still value the social connectedness in the future.

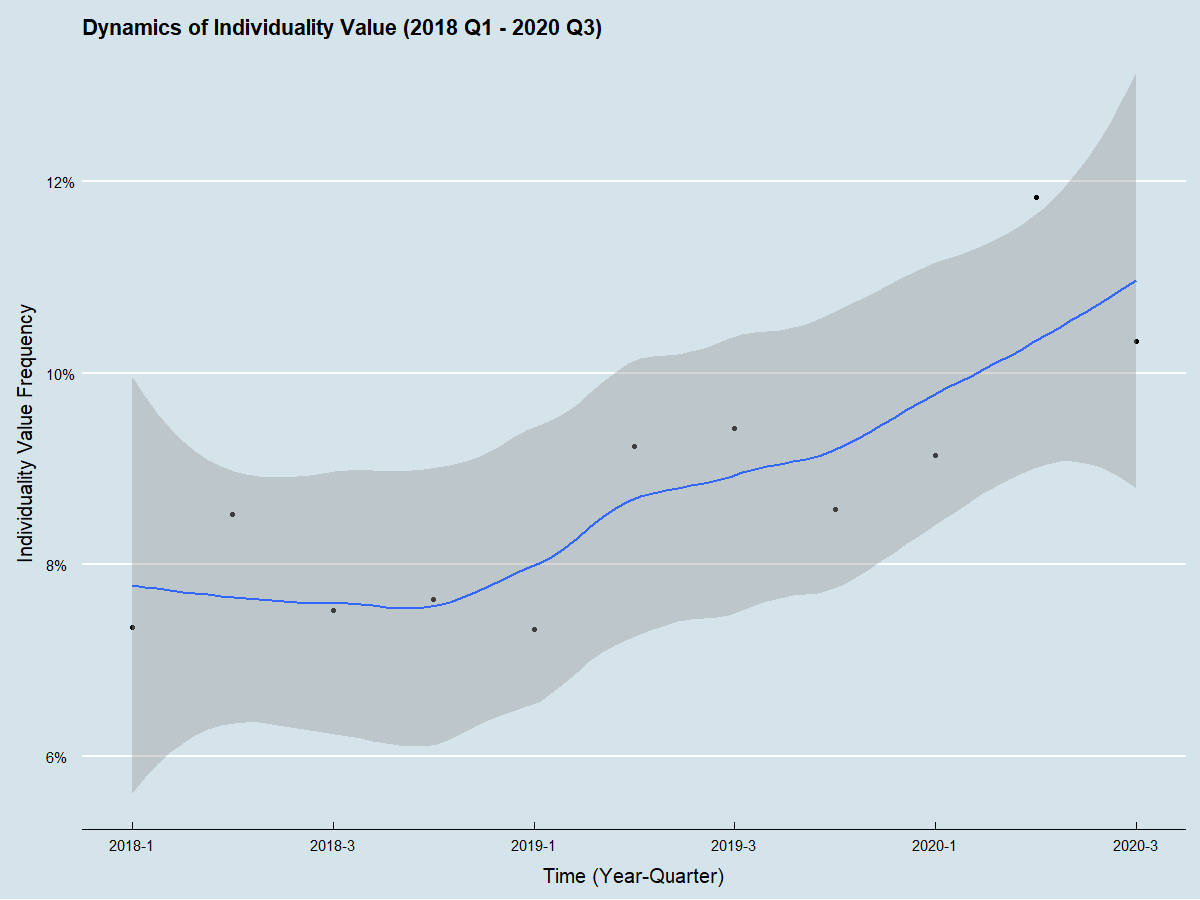


Chart 8, Dynamics of Individuality from 2018

In Chart 8, the curve for individuality has been increasing since the first quarter of 2019. It is clear in the chart that the consumers may want more value in individuality in the long run.

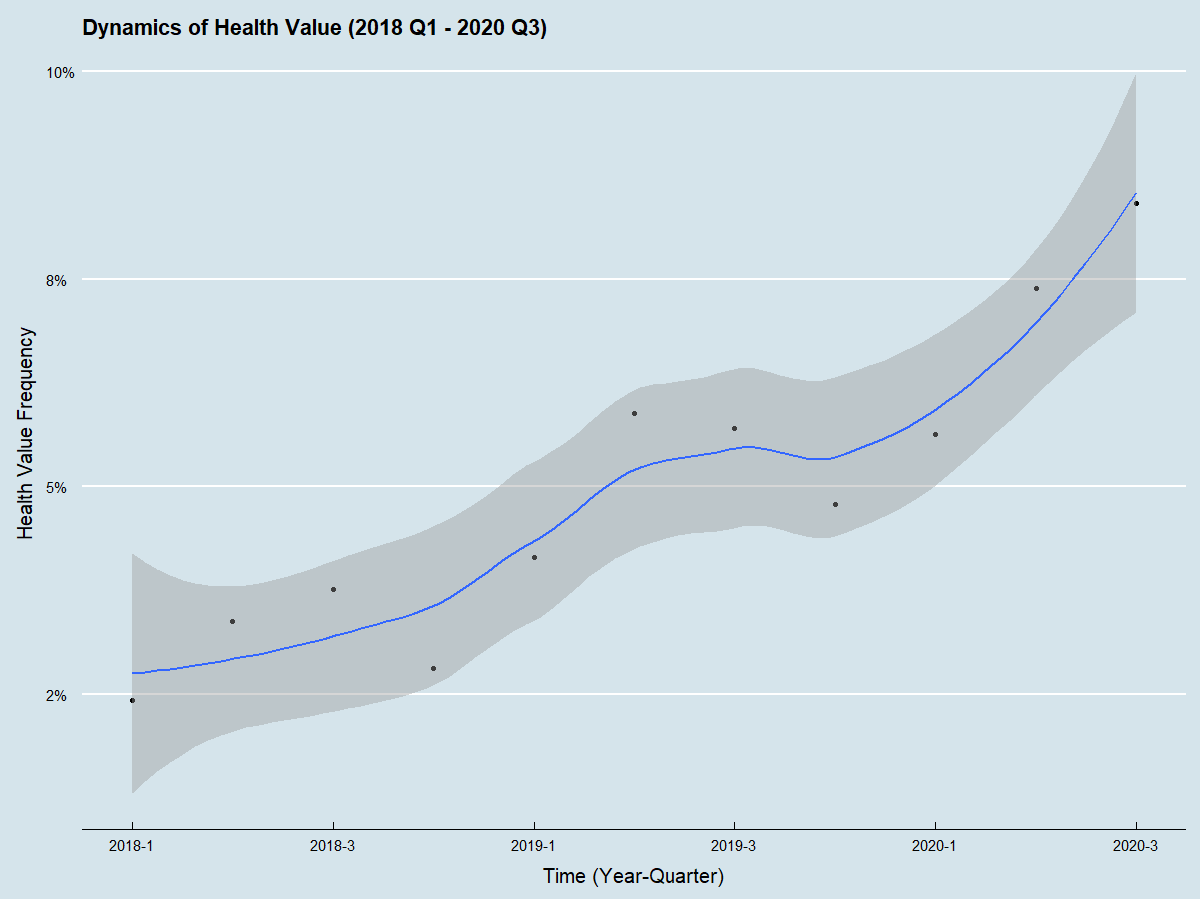


Chart 9, Dynamics of Health from 2018

The curve in Chart 9 representing health is increasing since the first quarter of 2018. We think the value of health is getting more and more important especially under the COVID situation.

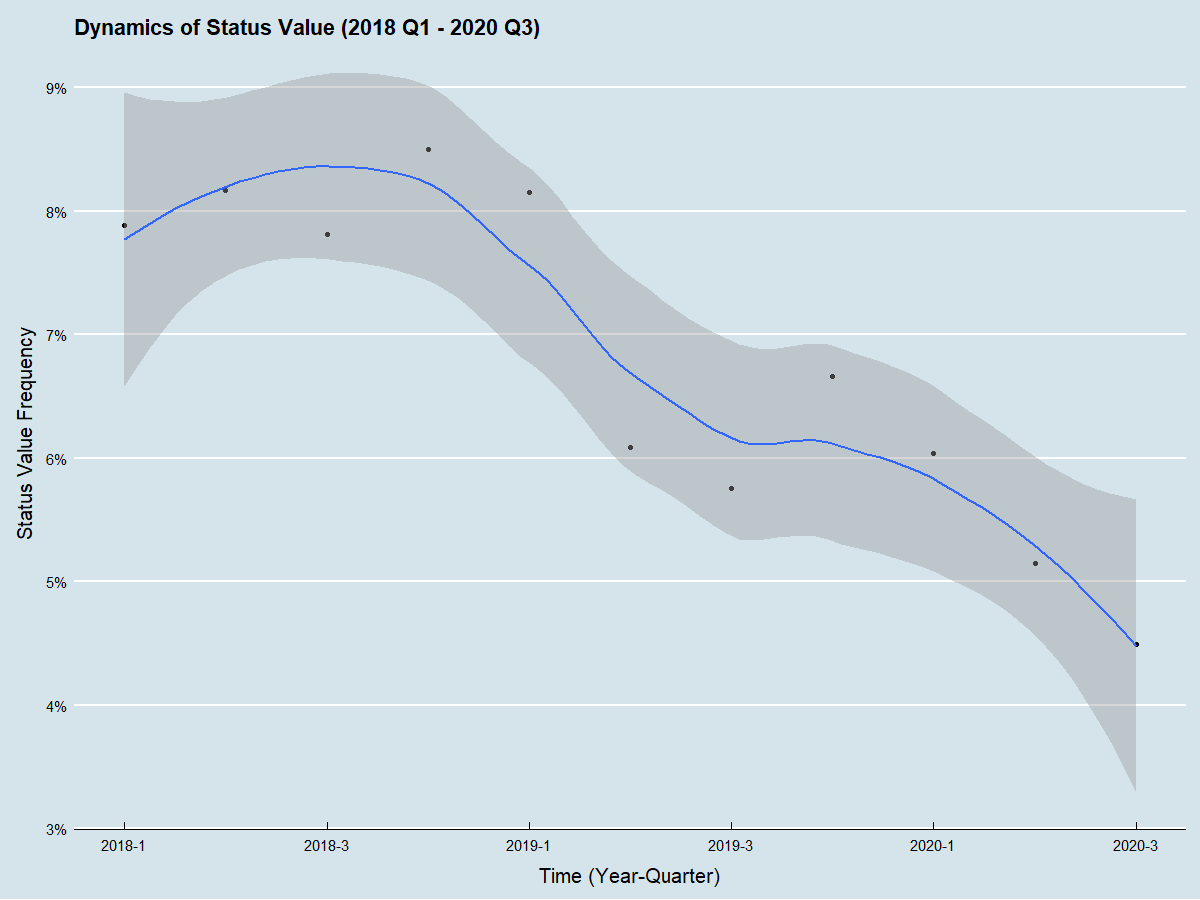


Chart 10, Dynamics of Status from 2018

As Chart 10 shows, the curve for status has decreased since the third quarter of 2018, telling us that the consumers don’t like to mention brands as often as before.

## Model-based Findings

After reviewing the current trends in the five values, let us think about this question: To what extent can we attribute the value dynamics to the pandemic?

To identify the real effect of the pandemic, we need to use a method called DID, Difference in Differences. For example, to define the pandemic effect, it is natural to study the data difference, between the 2020 Q2 and 2020 Q1, let’s call that difference 1. However, difference 1 is a combination of the impact of the pandemic effect and the seasonality effect. To figure out how large the impact of the pandemic is, we have to eliminate the impact of seasonality. One way to do it is to calculate the data difference between the 2019 Q2 and 2019 Q1, let’s call that difference 2. Then the result of the difference 1 minus difference 2, is our impact on the pandemic effect.

In fact, the Pandemic effect can be parameterized, with an interaction item in a Regression Model.

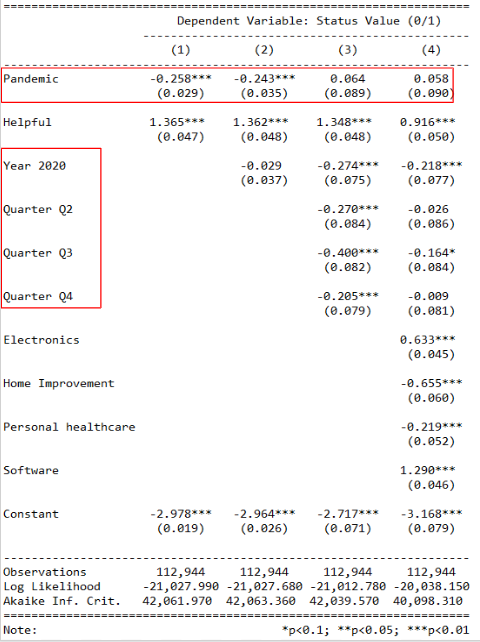


Chart 11, Logit model results on Status

As shown in chart 11, we did a series of logit models on status. From left to right, we gradually added more factors to control the impacts from seasonality and product category. In the first two testings, the pandemic factor is significant. It shows the pandemic has a negative impact on the status, because the coefficients are negative. However, once we have controlled the seasonality effect by adding the time dummies, the pandemic factor becomes irrelevant. Based on the model result, a more complete and prudent conclusion for us is, the status is in a pre-existing declining trend, and the pandemic does not significantly contribute to it, as chart 12 describes.

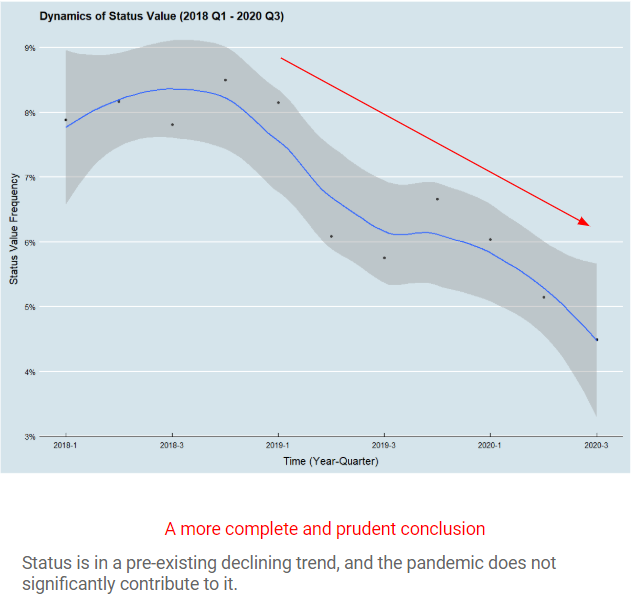


Chart 12, Dynamics of Status from 2018 and how pandemic influenced it

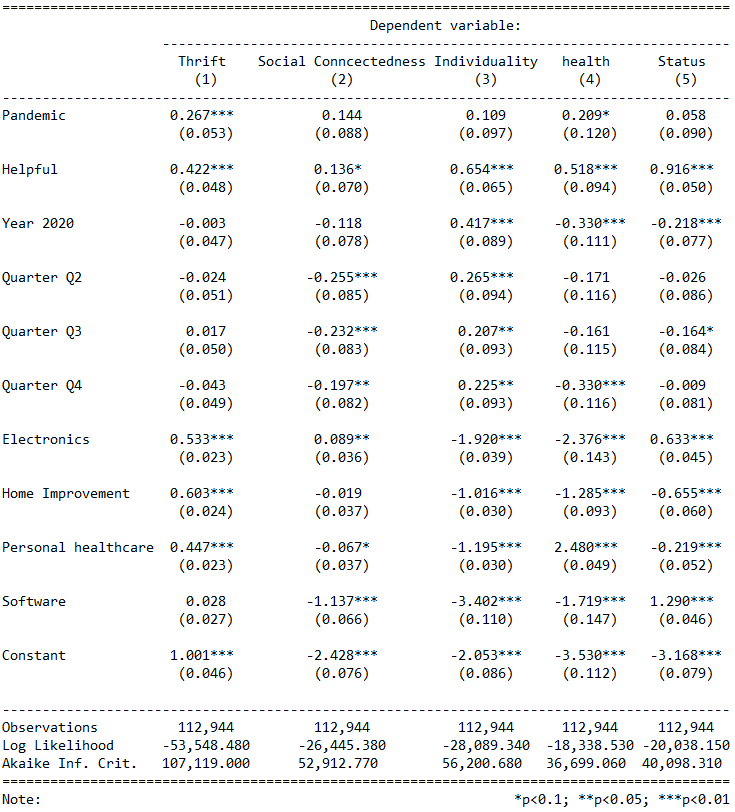


Chart 13, Summary of logit model results on different values

Similarly, if we put all the logit models for the predicted values together while having controls on the seasonality effect, we could see that the pandemic could only have a significant impact on thrift and health, as shown in Chart 13.

There are two main conclusions we can draw from the analysis. Firstly, the pandemic did spur consumers’ concerns on thrift and health, since their coefficients are positive. Secondly, companies need to revisit the consumer's value focuses on thrift and health after the pandemic, so we could know if the pandemic impact would persist in the future.

# Complimentary Analysis of UGC

Additionally, we have also done some Complementary Analysis of UGC with the clustering method. These analyses are not directly related to our customer value focus, but we just want to illustrate the potential of working with UGC.

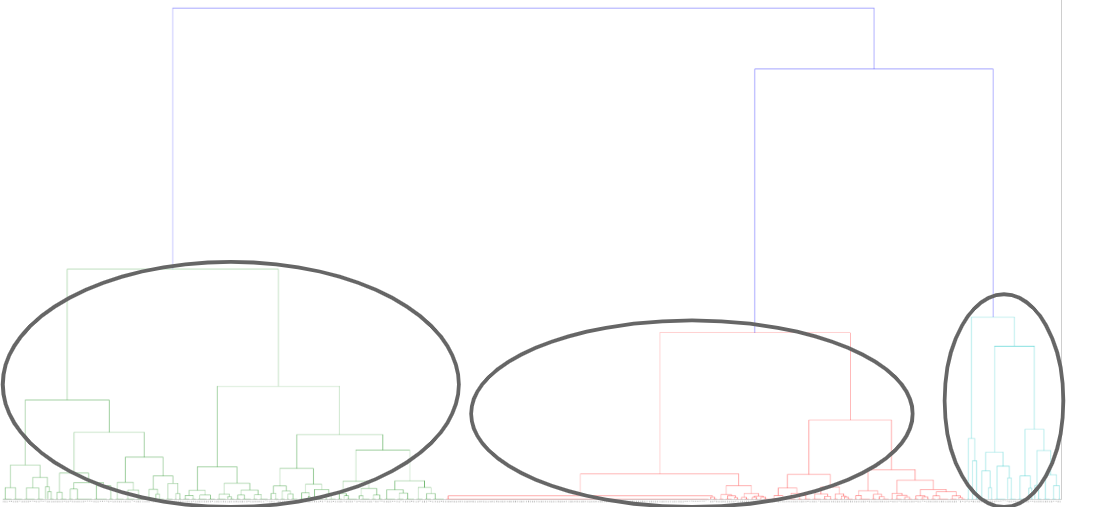


Chart 14, A Dendrogram on useful reviews from Apparel

We have run a cluster analysis on helpful reviews of apparel with Ward’s method and drawn out the dendrogram like Chart 14. After reviewing this dendrogram, we decided to break up these reviews into three groups.

We took some samples from the three clusters to see what they were like, so we could see what the customers were looking for in each of the groups. After our analysis, the first group is focusing on the quality of the products. The second group prefers the specific features of the products, and the third group likes the functions of the products. This might be a clue for companies at the time of developing marketing campaigns. When they have to advertise their products, characteristics like quality, features and functions should be paid more attention than other specifications of the products.

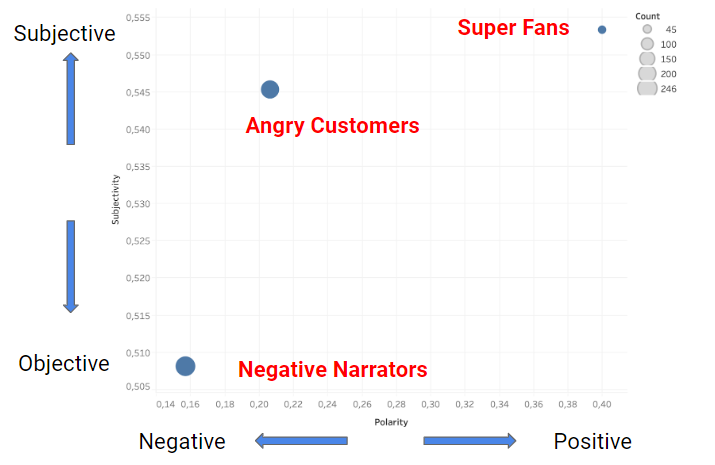


Chart 15, Example of cluster analysis on customer sentiments

We could also apply some sentiment analysis on the clustered groups, and see how different groups performed on the polarity and subjectivity, like chart 15. We can apply similar methods to any specific business to assess their customer segments so that they can develop detailed marketing strategies accordingly.

# Business Implications

Companies can use this approach to develop insights into customer values and needs. With these insights, marketers can reassess customer segmentation and better understand behavior. Further, consumer values can be used to inform the entire marketing strategy, incorporating them into product development, marketing content development, brand communication, and media placement.

# Limitations

There are also some limitations to our study. One of them would be the time difference between the two data sources. The reviews in the training set are from 2000 to 2015, while our data scraped for analysis is mainly in 2019 and 2020. The time difference may affect the accuracy of predictions since the phrases and terms used in the reviews might be different. In addition, we only selected values that were more reflected in the review data and categories we wanted for our analysis. Consumer values would be affected by the chosen categories, which means that our business implications in this study could be only applied to defined categories and defined values. We believe our study already provides a way of conducting similar marketing research that could be used on other product categories and defined consumer values. Maybe there would be more comprehensive research with more advanced data analysis methods in the future.

# 

# References

Charm, T., Bhargava, S., & Das, R. (2020, August 08). Survey: US consumer sentiment during the coronavirus crisis. Retrieved March 20, 2020, from https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/survey-us-consumer-sentiment-during-the-coronavirus-crisis

Hernandez, S., O'Key, S., Watts, A., Manley, B., & Pettersson, H. (2020, August 1). Tracking Covid-19 cases in the US. Retrieved August 1, 2020, from https://www.cnn.com/interactive/2020/health/coronavirus-us-maps-and-cases/

Smit, S., Hirt, M., & Dash, P. (2020, May 29). Crushing coronavirus uncertainty: The big 'unlock' for our economies. Retrieved August 1, 2020, from https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/crushing-coronavirus-uncertainty-the-big-unlock-for-our-economies

Chris McCormick and Nick Ryan. (2019, May 14). BERT Word Embeddings Tutorial. Retrieved from http://www.mccormickml.com

Chris McCormick and Nick Ryan. (2019, July 22). BERT Fine-Tuning Tutorial with PyTorch. Retrieved from http://www.mccormickml.com

Alammar, Jay (2018). The Illustrated Transformer [Blog post]. Retrieved from <https://jalammar.github.io/illustrated-transformer/>