### **Project Details**

### What should my final project submission include?

Other than the presentation, submit your final Proof of Concept, Written Brief, and Supplementary materials in a Google Drive by June 1, 11:59PM!

### Proof of Concept

This is what you were able to build. Make sure you have a way of presenting and sharing this

* If you are making an **analysis and visualization report, the graphs and reports are enough**. Make sure it is something they can easily share with people in their company.
* If you are making a **dashboard or data application, make sure you can send people a working version of the dashboard** that they can interact with!
* If you are making a **model, make sure you are able to show how the model was able to generate predictions** / include something on how they can integrate the predictions to what they are doing (e.g. generate a list of emails your model has recommended marketing to, generate a tagged list of customers based on clusters, etc.)

Some sample pegs:

* <https://stories.thinkingmachin.es/ulssi-doc-ai/>
* <https://stories.thinkingmachin.es/nighttime-lights-covid-19-southeast-asia/>
* <https://pudding.cool/2017/03/film-dialogue/>
* <https://tjpalanca.com/posts/2015-03-19-airline-agitation/>
* <https://medium.com/airbnb-engineering/beyond-how-may-i-help-you-fd6a0d385d02>
* <https://research.fb.com/blog/2021/04/introducing-causal-network-motifs-a-new-approach-to-identifying-heterogeneous-spillover-effects/>

### Supplementary Materials

* Include any code, files, and links in the same folder or repository as your final submission!

### Client Presentation

During the last week of class, we will have presentations with our clients on what you were able to build!

15 minutes max, and go through the high-level flow of what is in your written brief:

* Context: What problem are you trying to solve?
* Solution: How did you solve the problem?
* Impact: What value does your project create?
* Sustainability: How can your project potentially be implemented moving forward?

The tone of the presentation should generally revolve more around selling what you built and talking about the impact you were able to create over talking too deeply about how you technically coded it. We can leave that for the Q&A portion if the panelists want to ask about it.

### **Grading**

for detailed explanation of each criteria.

* Content (30%)
  + 10% Context and Impact
  + 10% Analytical Methodology
  + 10% Communication
* Presentation (15%)
  + A % of this will come from the panel, and they will be using the same syllabus criteria to grade you.
* Peer Review (5%)

### Timelines

* **May 27: Practice Presentations**
* **June 1: Submit Requirements**
* **June 3-5: Final Presentations (Dependent on Client Availability)**

### 

### 05/12/21 Brainstorming

Tasks:

* Pre-processing and cleaning data
  + Combine datasets (done)
  + Lemmas, tokens (done)
* EDA
  + Lengths of reviews
    - \* Understand the average lengths of each rating (ish)
  + Distribution of ratings (ish)
    - Maraming 1 and 5 (we can focus on them)
  + Topic modelling (gensim)
    - (2) run a topic model around certain ratings
  + Word frequency (wordcloud)(nltk)
    - Word Cloud depending on rating
* Predicting rating according to review (classification)
  + You can (1) run an exploratory regression/tree model to see what features influence a good rating
  + (3) run a k-means (will teach this next time) cluster for insight on themes

### 05/14/21 Consultation

* **Ask questions** you wanna answer and focus on how to get the answers
  + Will help us know if we need to analyze further or we’ve answered the question already
* How in depth should the solution be? Will the solution we did be enough
  + Yes. The goal of the project is to come up with insights
* Histogram of score is useful → **shows most of the reviews are either 1 or 5**
* Verified, explore certain words (over time)
* APP Store vs Play Store → more people are android users (back it up with research)
* Correlation with dates, imputing? What version existed in a certain date -> see if logic
* Gcash -> New update forces users to update to use it
* Do a lot of the exploratory stuff → continue to do what Jasper is doing
  + Maraming 1 and 5, 2-4 may be negligible already
  + Frequency → 1 gravitates toward specific words like verify, email, fix → look for the reviews where the words are mentioned
* Be curious! Fix what?? → see all reviews with “fix” and check if there’s a trend
* Topic modelling → u can look at it at a broad perspective then go into the more specific (you can focus on topic modelling reviews with rating 1 lang for example)
  + These are the trends of that specific factor
* Double check codes and data if there are deleted rows
* Tagalog dictionary
* Insight backed up by data would be super useful na
* Cash -> might be cash-in
* If groupby,
* Before group, filter data first -> make base as flat as possible first

**Objective:** improve services and entice more users to use the app

**Decisions:**

* What features do we need to improve and focus on for the next app update?
  + What are the **strong points** of GCash that we can use to market to potential customers? (marketing)
  + What are the **weak points** of GCash in general that must be addressed to reduce negative ratings and gain more customers? (product)

**Information:**

* As an Apple/Android user, what **problems** do I experience while using the app?
  + Comparative topic modelling:
    - Split data into groups by os and time?
    - Topic model
* What are the responses to the ratings and corresponding content/reviews of users? (look at the data on GCash responses) which ratings have the most response?
  + Filter out content w response then EDA
* What problems are commonly presented by the users among the low-rating reviews? (more specific: what common words make up the low-rating reviews?) (done -Japs)
* What features are commended by the users among the high-rating reviews? (done -Japs)
* For updates: does gcash address the concerns of the users per update?
  + Explore certain words over time

### 05/20/21 Meeting

| **Things to do** | **Deadline** | **Roles** |
| --- | --- | --- |
| Outline and PPT for writeup | May 23 (Sunday) | Yssa, David, Pau, Lanz |
| Code the pie chart for topic analysis | May 24 (Monday) | Yssa, David, Pau, Lanz |
| Update dash app by incorporating pie chart | May 24 (Monday) | Jasper |
| Consult sir | May 25 (Tuesday) | Everyone |
| ~~Style dashboard~~ | ~~May 25 (Tuesday)~~ | ~~Jasper~~ |
| Practice presentation | May 27 (Thursday 6pm) | Everyone |
| Finish writeup and PPT | May 31 (Thursday) | Yssa, David, Pau, Lanz |
| Pass | June 1 (?) | Everyone |
| Present to Client | June 3-5 | Everyone |

**Pie Chart**

Input: slider, dates, rating (from dropdown)

Shows: Two pie charts based on one-rating and five-rating topic models with colors representing each topic. The pie chart will show the contributions of each topic within the time period specified by the slider. For instance, the five-star ratings have three topics. Topic 1 was shown 30% of the time, Topic 2 was shown 50% of the time, and Topic 3 was shown 20% of the time from December 2020 to February 2021 (date range indicated by slider).

### Outline

**The written brief should be able to concisely answer the following questions:**

* Context: What problem are you trying to solve?
  + What is the background of the company? **Info about GCash**
  + What is the background around the dataset you have? **Reviews and ratings by the users**
* Solution: How did you solve the problem?
  + What methodologies were employed to solve the problem? **Text Analytics (specifically, word frequency and LDA topic modelling) and Dash (?).**
  + Why were these methodologies chosen?
  + How did you contextualize this for your problem? **Insert special considerations in pre-processing/cleaning the dataset**
* Impact: What value does your project create?
  + What key insights or recommendations was your project able to generate?
  + How does this tool help improve the decisions they can make?
  + If applicable, what does it optimize (speed, cost, revenue, etc.)?
* Sustainability: How can your project potentially be implemented moving forward?
  + What can the company do to continue to extract value from your project?

### PPT LINK:

### <https://www.canva.com/design/DAEfLmkjLTg/share/preview?token=A6rDPFt1bLZX6m9U9gM_ew&role=EDITOR&utm_content=DAEfLmkjLTg&utm_campaign=designshare&utm_medium=link&utm_source=sharebutton>

\*NOTE: **The tone of the presentation should generally revolve more around selling what you built and talking about the impact you were able to create over talking too deeply about how you technically coded it. We can leave that for the Q&A portion if the panelists want to ask about it.**

### 

### Write-up

1. **CONTEXT**
2. **Company Background**

GCash is one of the companies under Mynt, a FinTech startup partnership between Globe Telecom, Ayala Corporation, and Ant Financial. Mynt’s five key services revolve around payments, remittances, loans, business solutions, and platforms ([Mynt](https://www.mynt.xyz/about-us/#:~:text=GCash,GCash%20American%20Express%20Virtual%20Pay)). GCash’s services also revolve around these solutions. Since its establishment in 2004, it has effectively been providing its mobile money services to its customers. The company is able to give Globe’s mobile phone subscribers, both local and international, access to a cashless and cardless transaction. Customers are entitled to online money remittance, donations, disbursement of salaries, and payment of bills, products, and services ([GCash Background](https://www.mobilepaymentsworld.com/case-study-gcash/)). Nowadays, especially in the context of the pandemic, people have been using GCash to facilitate online transactions so they would no longer have to go out to pay for their needs and wants.

Currently, the company, specifically its marketing team, drives its decisions based on qualitative insights generated from reading user reviews about their app. Most of their evidence is anecdotal and they are in need of a structured way of looking at their data to create more informed decisions when improving their service.

1. **Dataset Background**

For this project, the two datasets provided contain both quantitative and qualitative information about the GCash app reviews of Apple and Android phone users from March 2012 to April 2021. Data from both sets involved ratings, reviews, usernames, and date posted. To give more context, users have reviewed the app based on its features and usability. They would give a rating of 1 to 5 depending on how satisfied they are with the app and provide comments (reviews) to justify their rating.

1. **The Problem Being Tackled**

Given that the datasets contain information about users’ reviews and ratings, the group could use the insights generated from these to drive different strategies in improving the app’s services to keep existing users and even entice more users to use the app. They would have to look at the app’s strong points which the company can use in targeting potential customers. Furthermore, they would also need to analyze the reviews to pinpoint the app’s weak points and identify what features the company should improve and focus on for their succeeding updates on the app.

1. **SOLUTION**
2. **Brief Description of the Solution**

**Brief Summary:** A dashboard that shows the frequency of relevant words and proportions of topics from GCash customer reviews over a period of time.

**Visualization:** The frequency of words is visually represented by a line graph (x = time and y = count) and the distribution of topics is represented by a pie chart (colors = topic). Two pie charts are provided to represent topics from one-star ratings and five-star ratings.

**Interactive Elements:** The user may opt to adjust the slider to change the date range, which filters the data being visualized in the dashboard. There is also an interactive table containing the most frequently used words per rating within the period specified by the slider. To check the usage trend of a specific word, one may either select a word from the table or type in the search bar. Two search bars are provided to allow for comparing trends between two different words.

1. **Methodology**
   1. **Process**

Setting up the data for use required1) combining the two datasets, 2) setting all words to lowercase, 3) removing English and Filipino stop words and non-word characters, and 4) tokenization to set the appropriate data format for the analytical tools to be used.

Afterwards, a text analytics approach utilizing word frequency and topic analysis was applied. Frequency distribution from the Natural Language Toolkit (NLTK) was used for the former, while Latent Dirichlet Allocation (LDA) was used for the latter. In addition to this, part of speech (POS) tagging was used to generate the data to be used by the interactive table. Shaping this data required ranking the word frequencies per rating, filtering out words that were not nouns, verbs, or adjectives, and removing the top 100 common words across all ratings.

For the topic modelling, special cleaning was made to remove meaningless words that impeded on the coherence of the topic models. Next, the combined dataset had to be split into two: one dataset containing data with rating = 1, and the other dataset containing data with rating = 5. Afterwards, coherence scores were generated from a list of LDA models across a range of two to ten topics (interval = 2). This was done to find the optimal number of topics for both datasets. In the end, the group decided from the coherence score graphs that four topics for both topic models would suffice. Then, looking at the top 10 words for each topic, the team had to create topic labels to make sense of them. To create the pie charts, the probability that each review belonged to a certain topic was summed up.

Finally, tying everything together is Dash, which required HTML and some CSS elements to ground the dashboard. The link between interactive elements (ie. table, slider, and search bars) and visualization elements (ie. line graph and pie charts) were also made possible through Dash, coupled with custom-made functions that made the process more efficient.

* 1. **Rationale**

The group merged the two datasets provided into a single dataset to generate meaningful insights from all reviews. Since GCash is targeting both Apple and Android gadget users, it would be beneficial for them to focus on the whole picture and analyze it in one go instead of just looking at one or the other as the features and app updates available for both users will generally be the same. The merged dataset may provide more insights for the company as it widens their perspective and this may help them form better strategic decisions that will address the concerns of both users ([Combining Datasets](https://www.netquest.com/blog/en/5-reasons-to-combine-different-data-sources-in-your-online-research)).

The text analytics approach was deemed appropriate because the datasets involve user reviews. Since reviews vary from person to person, language processing techniques such as **stopword removal** and **tokenization** were used to structure the data. **Frequency distribution** was used to explore the words that appeared most frequently and to identify patterns among the reviews made by users across the different ratings. Lastly, the construction of reviews may vary from one user to another but may actually point out the same idea so the group used **topic modelling** to understand similarities among the words found in one-star and five-star reviews, while classifying them into relevant topics that the company can better interpret than singular words. This is in line with the aim of this project to generate insights that would help the company in marketing the strong points of the app to potential users as well as in addressing their weak points to improve their service.

Dashboard was chosen as the data product and is intended for the interactive use of the company’s marketing and product development teams. It was designed to display the appropriate visualizations that contain the relevant information they would need to make decisions regarding what features to improve on and what angle they can use to market the app. The clients may search for specific words to determine usage trends, hopefully providing them with insights on whether they are able to address previous issues or if there are recurring issues that need to be solved. They may also opt to choose certain periods they want to explore either on the line graph or pie chart and it will display the most frequently used words or common topics within the specified period which could be helpful in their preparations for succeeding app updates or marketing campaigns.

* 1. **Contextualization (How is the model you made relevant in the business context?)**

By applying word frequency and topic modelling on reviews, the group was able to create a dashboard that can be used to determine what specific areas GCash must improve on to satisfy their customers. We identify the dashboard to be relevant in two different areas: product development and marketing.

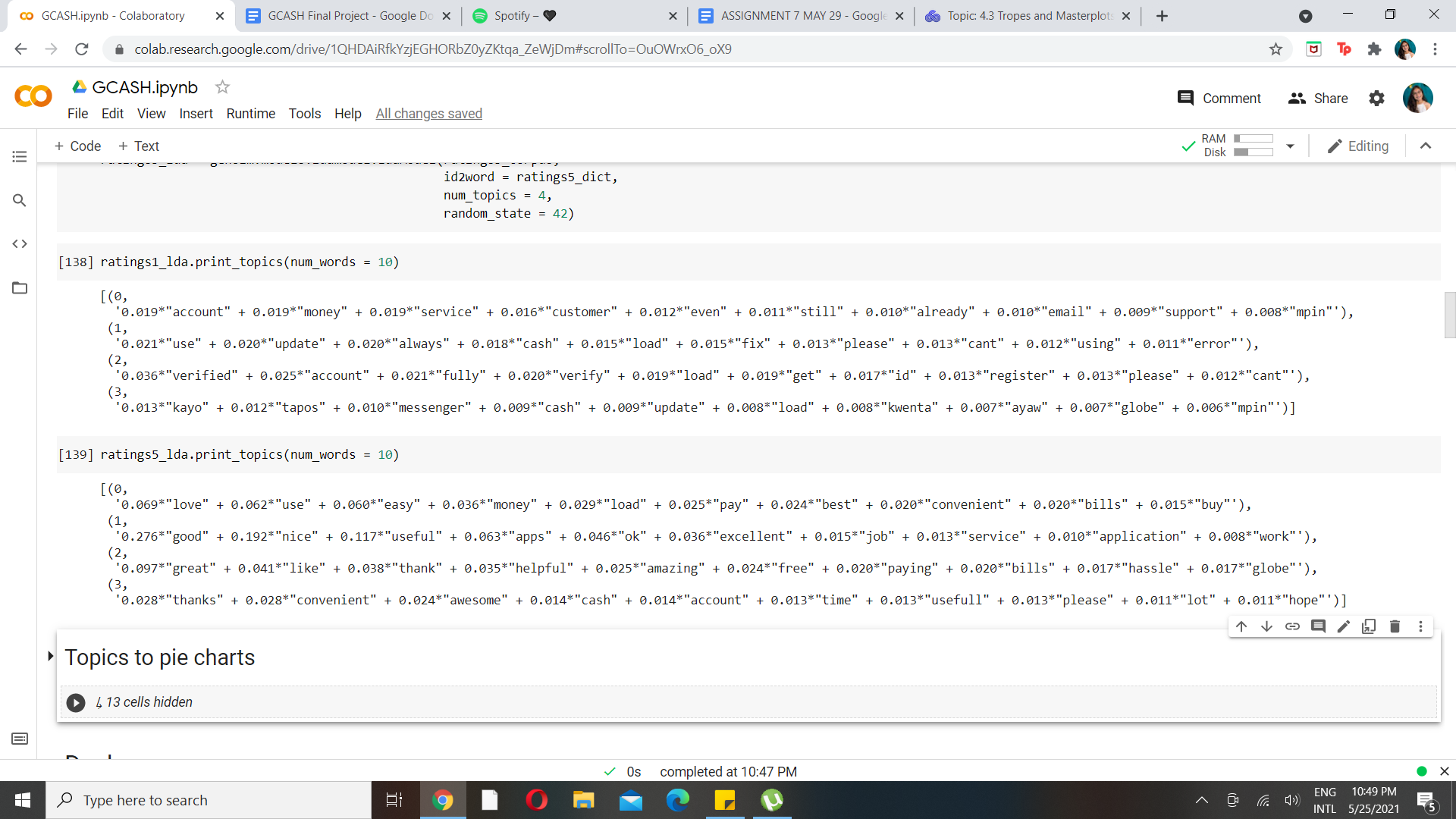
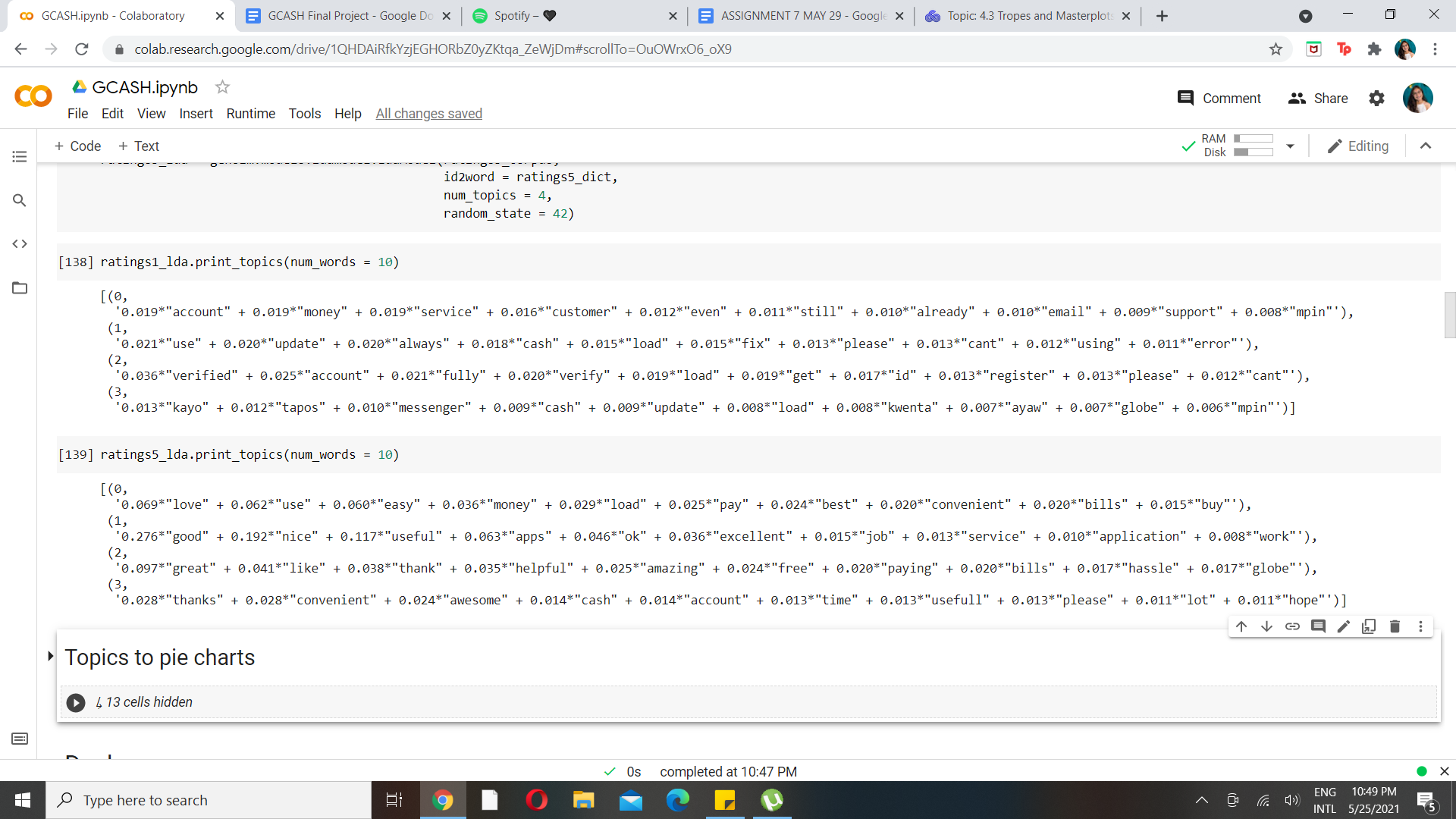
Product development is a crucial component for GCash because it provides numerous services to its customers. It is even more crucial today due to the pandemic, where more and more people are making digital transactions to avoid close contact ([Endo, 2020](https://asia.nikkei.com/Business/Companies/Digital-payment-grows-in-Philippines-amid-COVID-19-fears)). GCash needs to effectively update their app so users will continue to have a seamless and convenient experience. With this, the dashboard can give the product team a direction with regards to which app features they must improve on and what bugs they have to fix by looking at the word and topic trends.

On the other hand, marketing is also crucial as this is how the company could gather more customers to grow their business. Marketing requires GCash to know about their selling point from the perceptions of its customers. Therefore, by looking at word usage and topic trends of five-star app reviews, the marketing team can come up with advertisements and campaigns that would highlight the perceived strengths of the app. These marketing campaigns would prove effective because the messages conveyed are based on actual customer experiences.

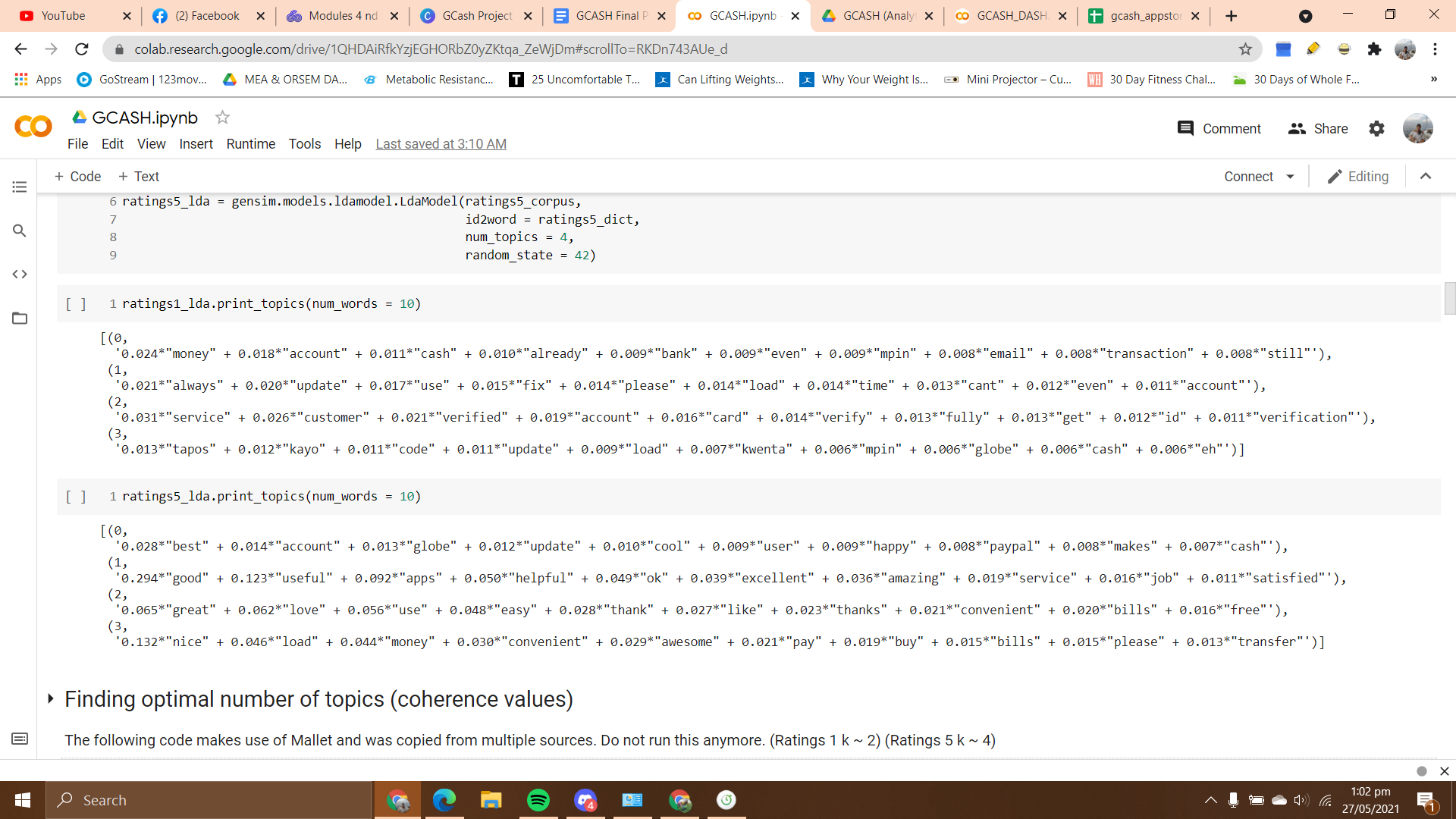
1. **IMPACT**
   1. **Key Insights and Recommendations**

* Looking at the histogram of ratings, the distribution is concentrated at rating = 1 and rating = 5. Calculating the proportions, 58.7% of reviews rated the app a 5, while 21.6% rated the app a 1.

OLD TOPICS (?):



CURRENT TOPICS (LDA V2):



**Suggested topics:**

| **Topic** | **Rating 1** | **Rating 5** |
| --- | --- | --- |
| **1** | Customer Service | Easy and Fast Transactions |
| **2** | Service Reliability | App Usability |
| **3** | Log-in/Verification | Services Offered |
| **4** | --STILL THINKING--  UI/UX Errors (?)  Customer Discontent/Dissatisfaction (?) | Customer Satisfaction |

**Topic Title Explanations:**

Rating 1

* Customer Service → how fast or often does GCash respond to customers’ messages?
* Service Reliability → transaction errors/bugs
* Log-in/Verification → bugs when logging in, hassle verification process, mpin errors
* Customer Dissatisfaction (for now) → general discontent/adjectives

Rating 5

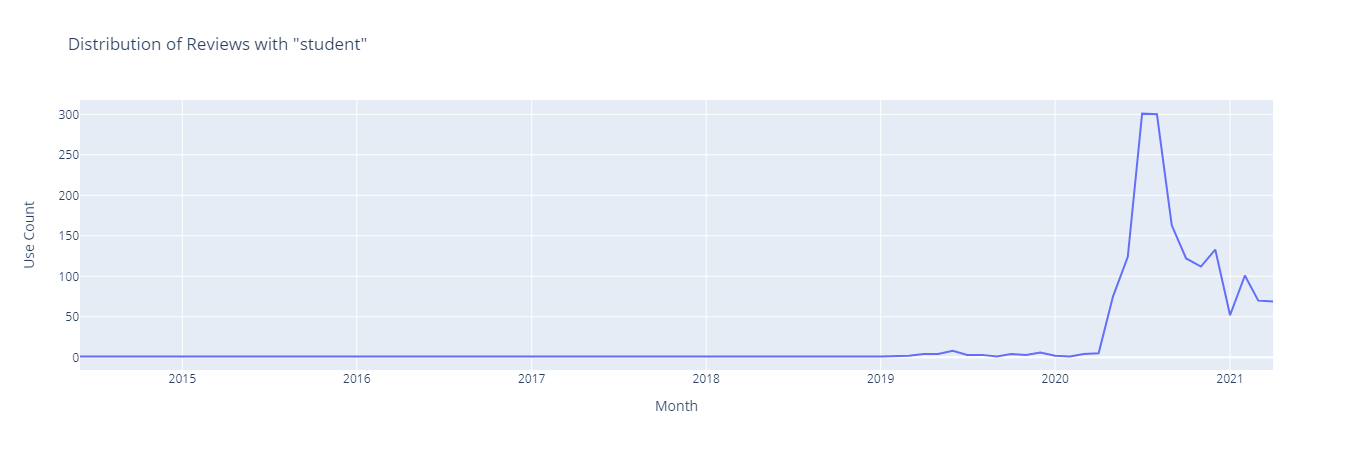
* Easy & Fast Transactions → no errors encountered during transactions **SMOOTH TRANSACTIONS**
* App Usability → refers to overall user experience,
* Services Offered → refers to diverse services customers are able to use **(the customers are able to use a lot of other services such as loans, investing, saving money)**
* Customer Satisfaction → praises from customers **(adjectives)**

**Notes:**

* **Last Gcash app update:** April 25 2021 (based on play store) version 5.40.0
  + Users are required to update the app to its latest version
  + haven't found info on how often Gcash updates their app
* **mpin** = Mobile Personal Identification Number (a 4-digit passcode that you can use to authenticate all your GCash transactions)
* **ticket** = Gcash app users send a ticket regarding their concern then the Gcash support team will respond to them
  + “If you have already submitted a ticket using your GCash-registered email address, you should be able to see and track the status of your ticket inside your GCash app. There is no need to create a new ticket to follow up on your concern.”

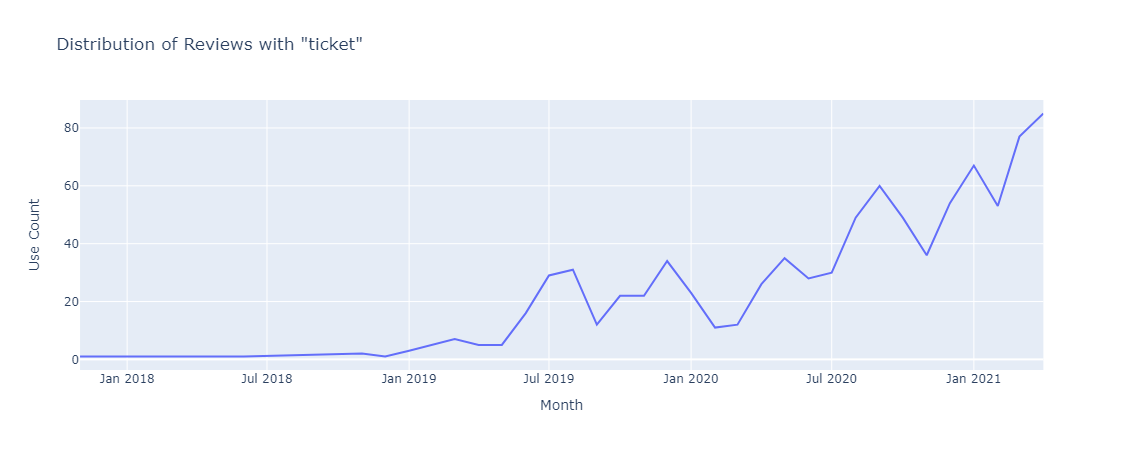
**Random insights:**

* Times that **“student”** was mentioned in rating 1 reviews suddenly rose from 5 to 301 (April vs. July 2020 → **onset of pandemic**) however it can be implied that Gcash was able to resolve it because as of April 2021 it dropped to only 65 mentions among the rating 1 reviews
  + “Not student-friendly”; Hard for students to get verified because they don’t have many valid IDs; doesn’t accept school IDs; school IDs might be most helpful for students during the pandemic because it’s hard to go out to get valid IDs

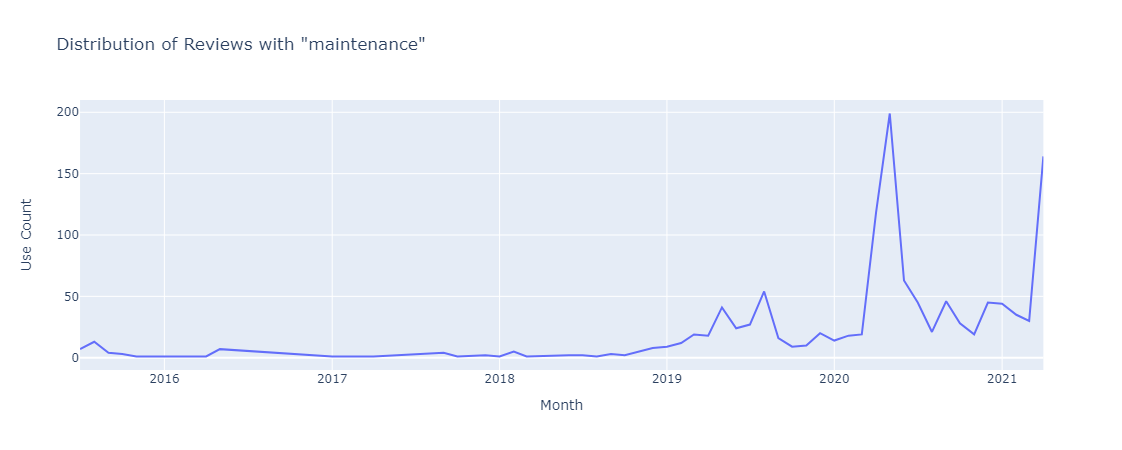


* In the graph below, a comparison in the usage trends of the words “**customer”** and “**service”** in the one-star reviews reveal that the two share similar behaviors which may refer to the overall customer support provided by Gcash for its users. Both started to fluctuate around 2019 and continued thereafter. This could generally mean that the Gcash Customer Support team has not been that consistent and effective in handling and addressing user concerns. An underlying issue might be revealed by how the usage trend of the word “**ticket**” also fluctuates over the periods and started around the same time as well. Gcash users submit a ticket in the app to report their concerns which are handled by the Gcash support team.

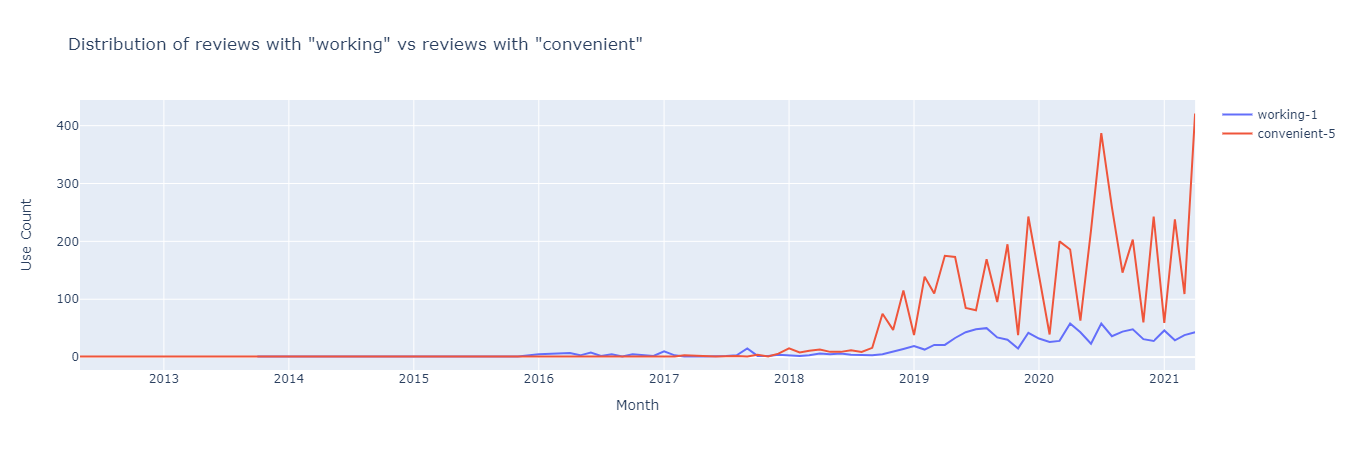




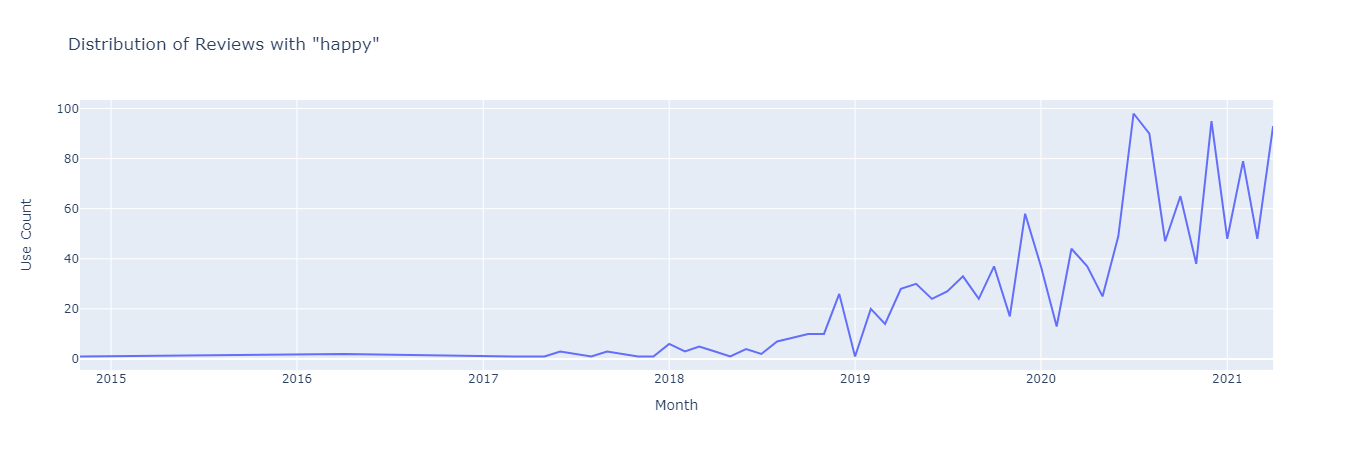
* “**Maintenance**” evidently increased especially during the most recent period. (users might have recently been experiencing troubles in using the app during maintenance periods without prior notice from Gcash) -> will cross check pa in the dashboard if this indeed reflects on rating 1 reviews **yup it reflects on rating 1 reviews! Super similar lang yung graph if rating 1 or all reviews, mas steep/pointed lang yung sa rating 1 -pau**



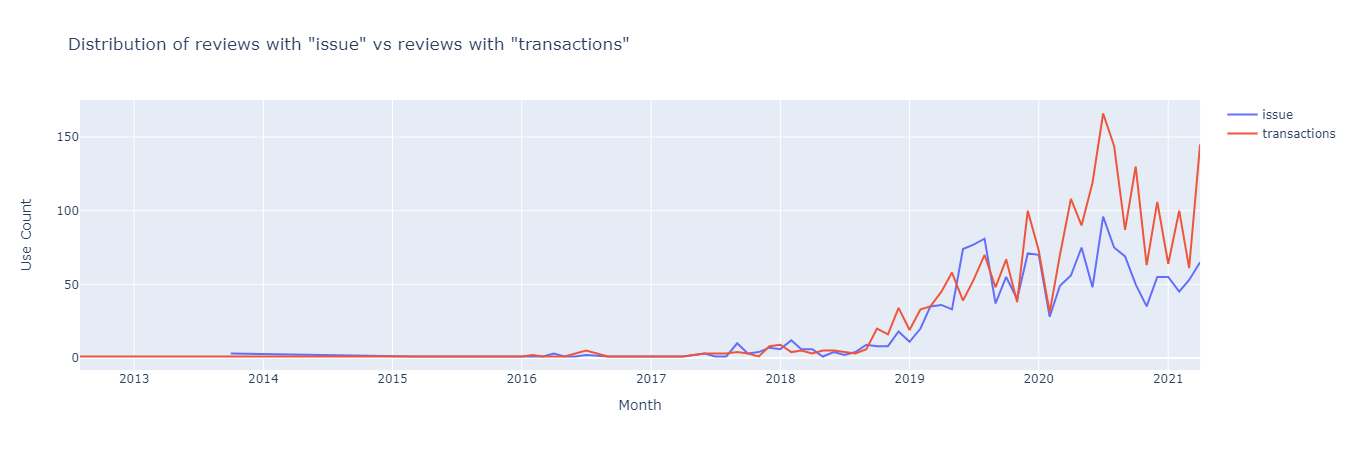
* Times that **“working”** and **“convenient”** were mentioned in rating 1 and rating 5 reviews respectively were increasingly rising since late 2018 up to now. This could mean that although some customers were finding the app to be convenient in their daily, weekly, or monthly transactions, others were still facing some problems when it comes to system errors.
  + Some users mentioned that when they needed to use the app, the system was not working or they were encountering bugs.



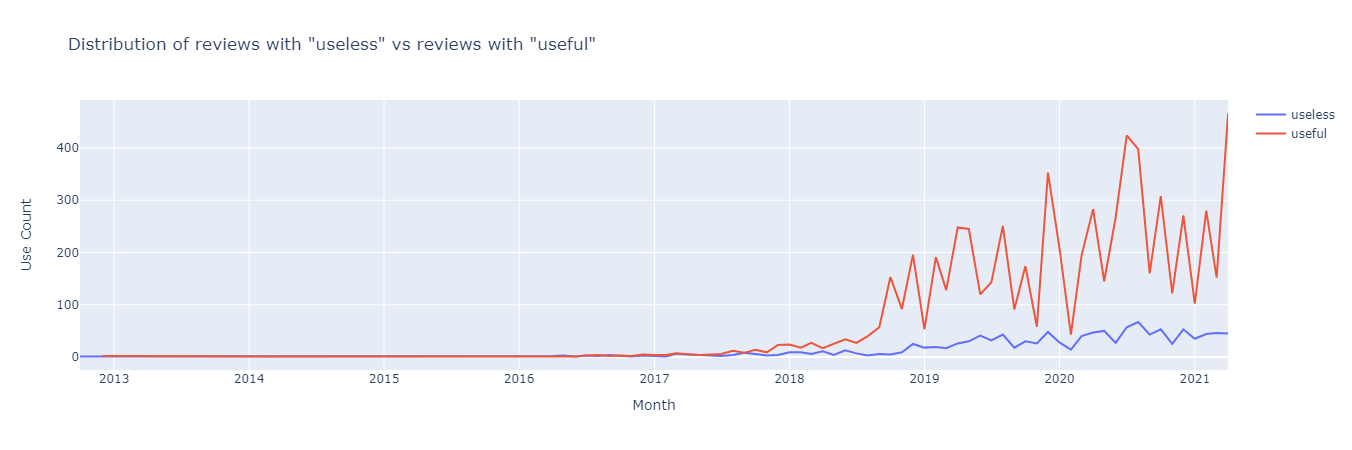
* There is an increasing trend in terms of how **“happy”** the customers are with the app. As seen in the graph, there was a sudden spike in the year 2020, probably because people were in quarantine at the time and they were using the app more in transacting.

****

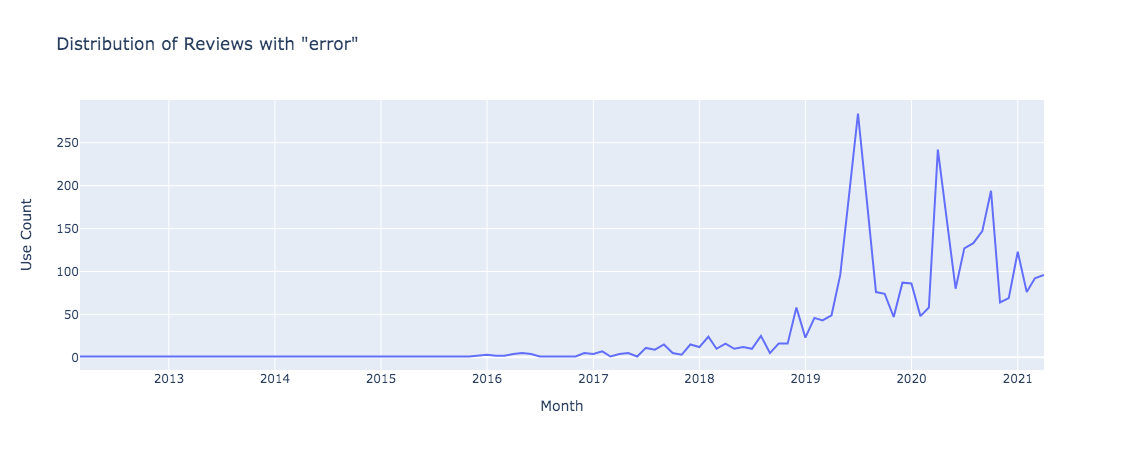
* Looking at the graph, we can see that both graphs for **“issue”** and **“transactions”** have similar behaviors especially in the period between 2020 and 2021. We can assume that some customers experience problems when it comes to transactions (sending/receiving money or paying bills).

****

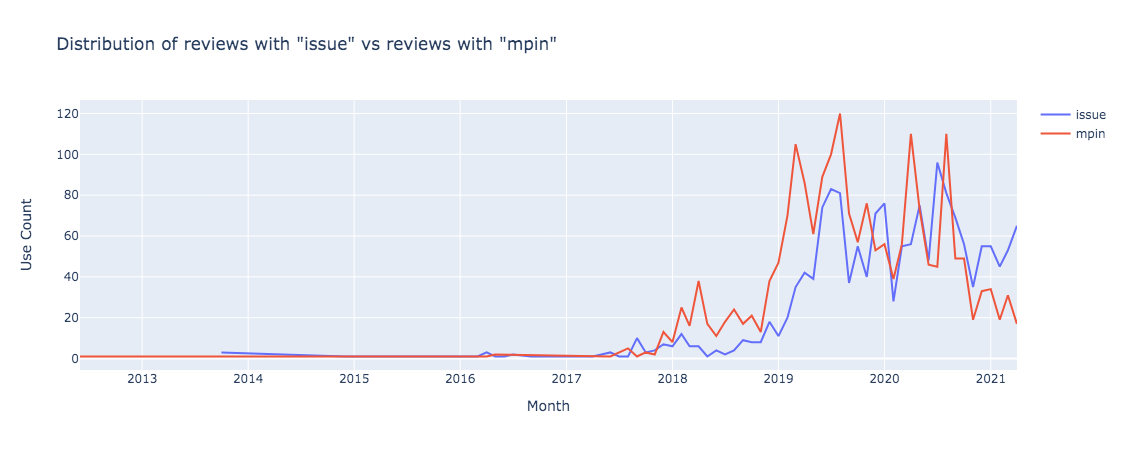
* **“Useless”** vs **“useful”** ratings → show that a lot more users find the app useful which can be backed up by the data on having more high-star ratings (3 to 5) compared to low-star ratings (1 and 2).



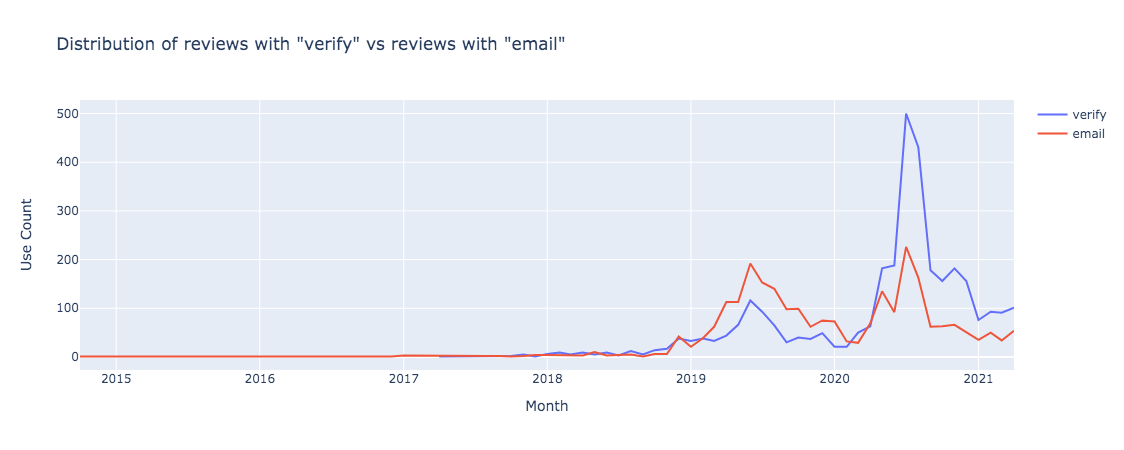
* **“Error”** ratings -> when it comes to error it can be observed that Gcash responds quickly. This can be seen through the huge drops.



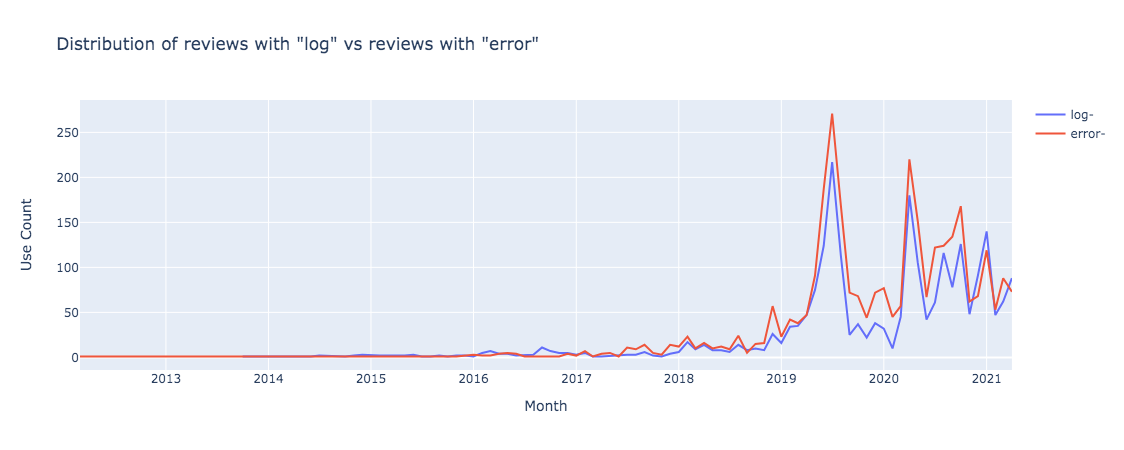
* **“Issue” vs “mpin” -** issues may be related to the mpin ?

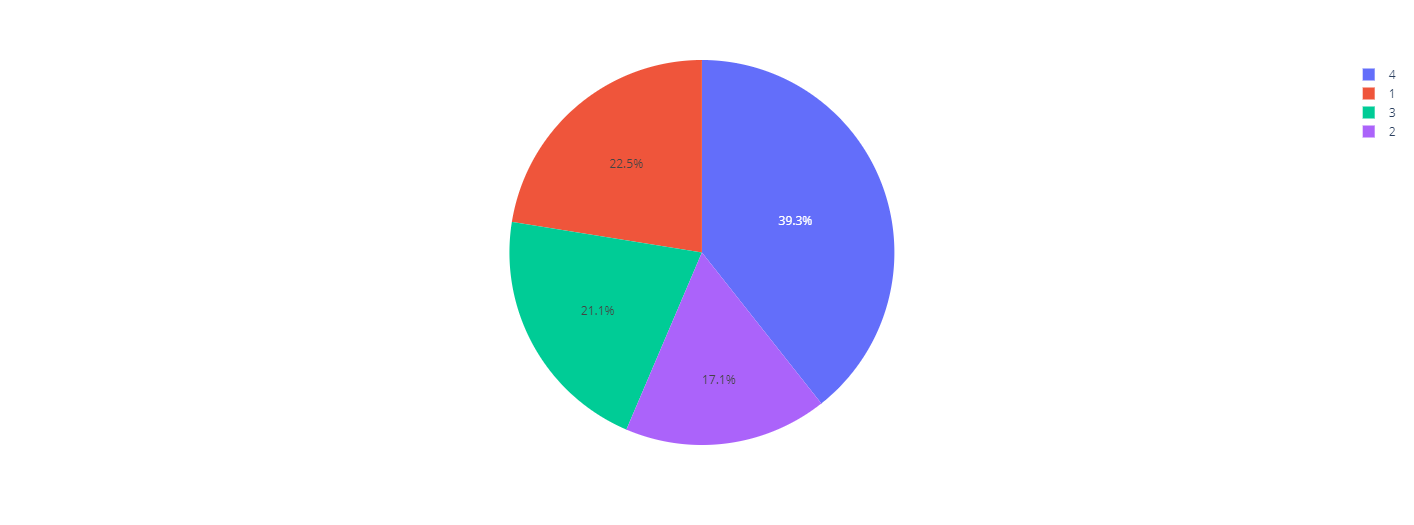
****

* **“Verify” and “email” -**

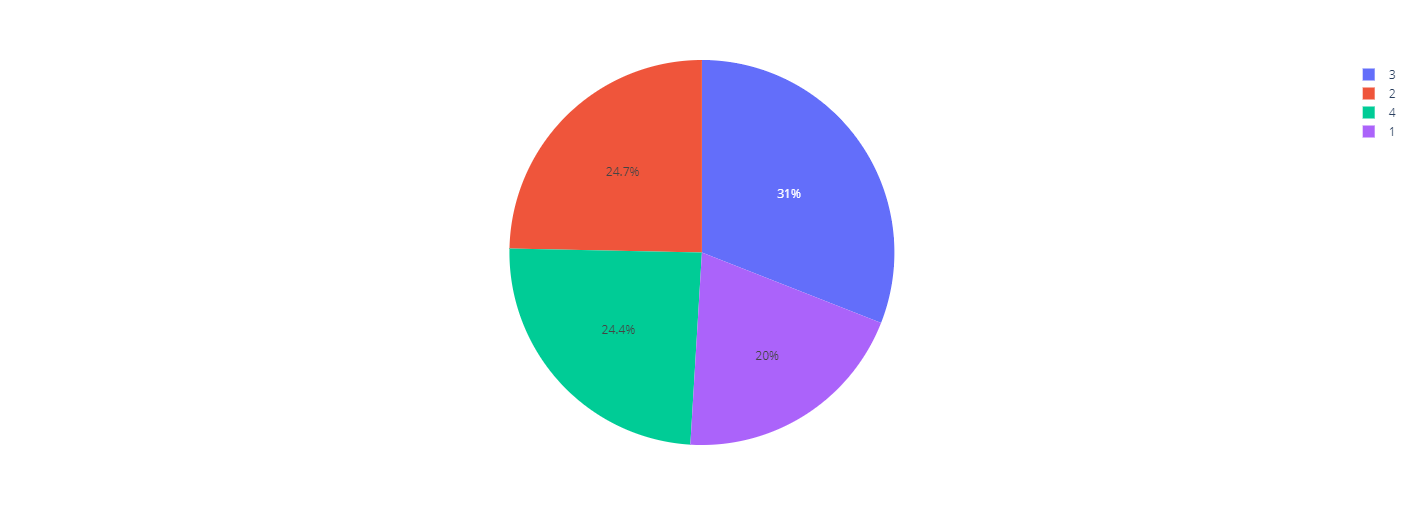
****

* **“Log” and “error” -** most of the errors may occur from log-in errors.



**Rating 1 Overall Topic Probability Proportions**

**Rating 5 Overall Topic Probability Proportions**

****

* 1. **Decisions Improved**

From currently only being able to generate qualitative insights from user reviews and report changes in the app’s ratings, the structured data can now guide the marketing team’s decisions, specifically regarding what features they can promote for their next marketing campaign based on quantitative evidence.

The company can also gauge the app’s performance better and compare it across time periods. In particular, the product development team can have a better understanding of when they need to start developing the next app update and what issues need to be addressed based on classified topics of customer insights.

* 1. **What does this optimize? (optional)**

1. **SUSTAINABILITY**
   1. **Application to Long-term Business Processes**

In the long run, they could keep using this solution by updating the dataset and adding additional data (aka reviews). They could simply edit and run the code to generate information from the dashboard.

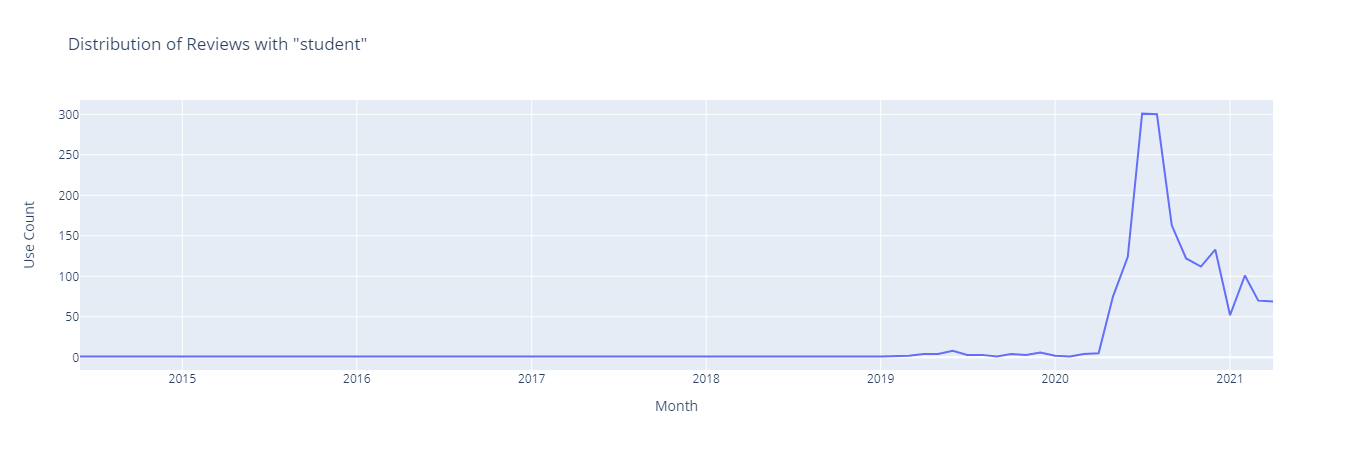
The company can continuously monitor their performance through the dashboard the group has created and provided. They could also look at how the consumers respond to the app’s features and updates and garner feedback and meaningful insights from there. By looking at certain time periods, they can also determine if certain business decisions such as marketing campaigns or feature updates had certain effects, may it be positive or negative.

NOTES

* **Last app update: (based on play store)**
  + April 25 2021 version 5.40.0
  + May 24 2021 version 5.41.0 -- navigation bar, merchant categories, more billers
  + Users are required to update the app to its latest version
  + haven't found info on how often Gcash updates their app
* **mpin** = Mobile Personal Identification Number (a 4-digit passcode that you can use to authenticate all your GCash transactions)
* **ticket** = Gcash app users send a ticket regarding their concern then the Gcash support team will respond to them
  + “If you have already submitted a ticket using your GCash-registered email address, you should be able to see and track the status of your ticket inside your GCash app. There is no need to create a new ticket to follow up on your concern.”

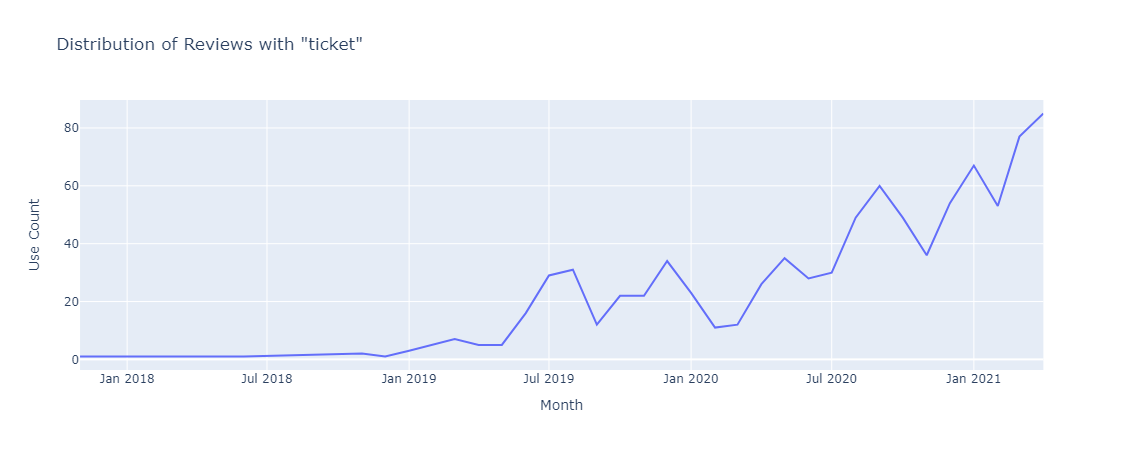
OLD INSIGHTS

* Times that **“student”** was mentioned in rating 1 reviews suddenly rose from 5 to 301 (April vs. July 2020 → **onset of pandemic**) however it can be implied that Gcash was able to resolve it because as of April 2021 it dropped to only 65 mentions among the rating 1 reviews
  + “Not student-friendly”; Hard for students to get verified because they don’t have many valid IDs; doesn’t accept school IDs; school IDs might be most helpful for students during the pandemic because it’s hard to go out to get valid IDs

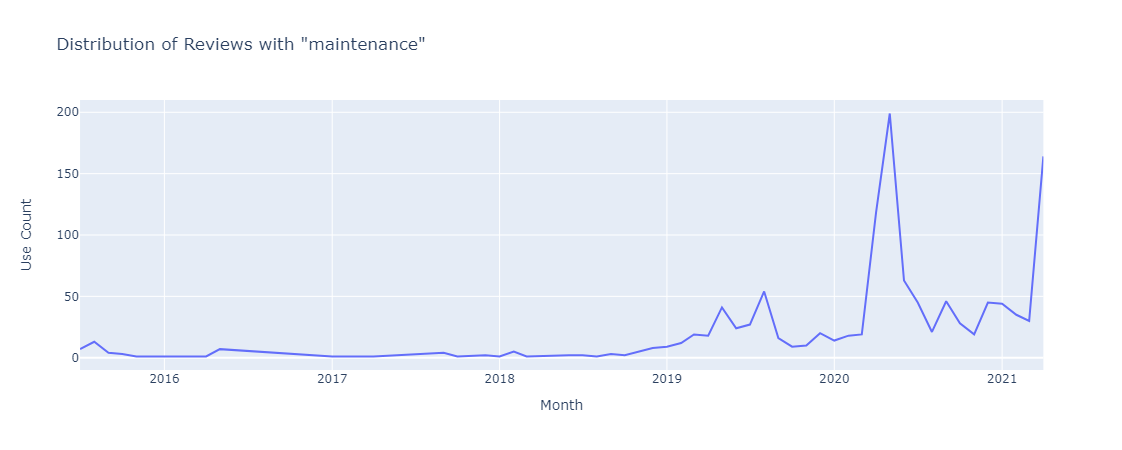


* In the graph below, a comparison in the usage trends of the words “**customer”** and “**service”** in the one-star reviews reveal that the two share similar behaviors which may refer to the overall customer support provided by Gcash for its users. Both started to fluctuate around 2019 and continued thereafter. This could generally mean that the Gcash Customer Support team has not been that consistent and effective in handling and addressing user concerns. An underlying issue might be revealed by how the usage trend of the word “**ticket**” also fluctuates over the periods and started around the same time as well. Gcash users submit a ticket in the app to report their concerns which are handled by the Gcash support team.

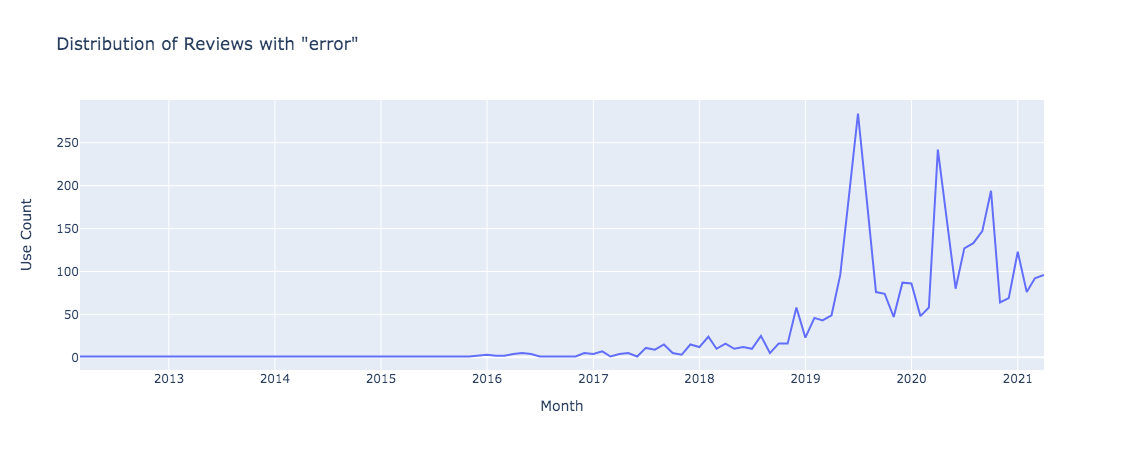




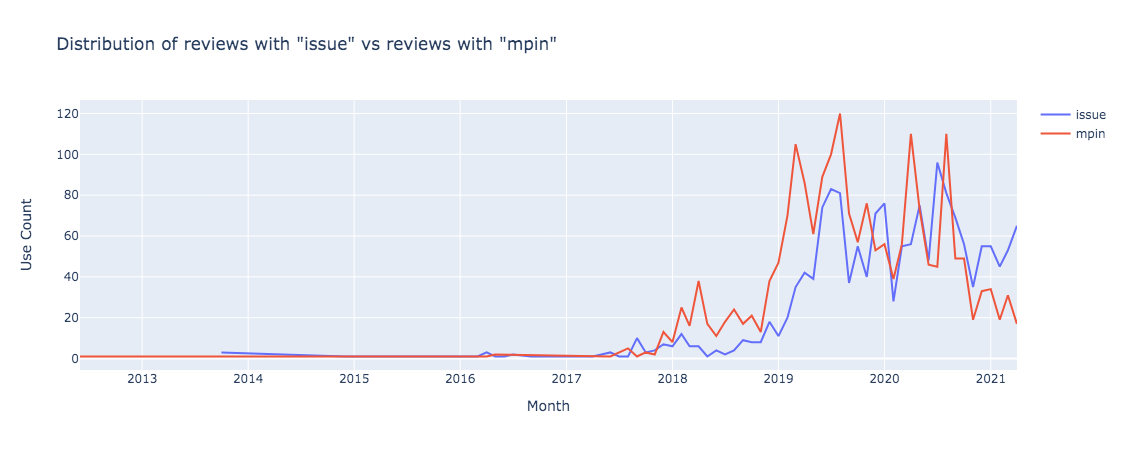
* “**Maintenance**” evidently increased especially during the most recent period. (users might have recently been experiencing troubles in using the app during maintenance periods without prior notice from Gcash)



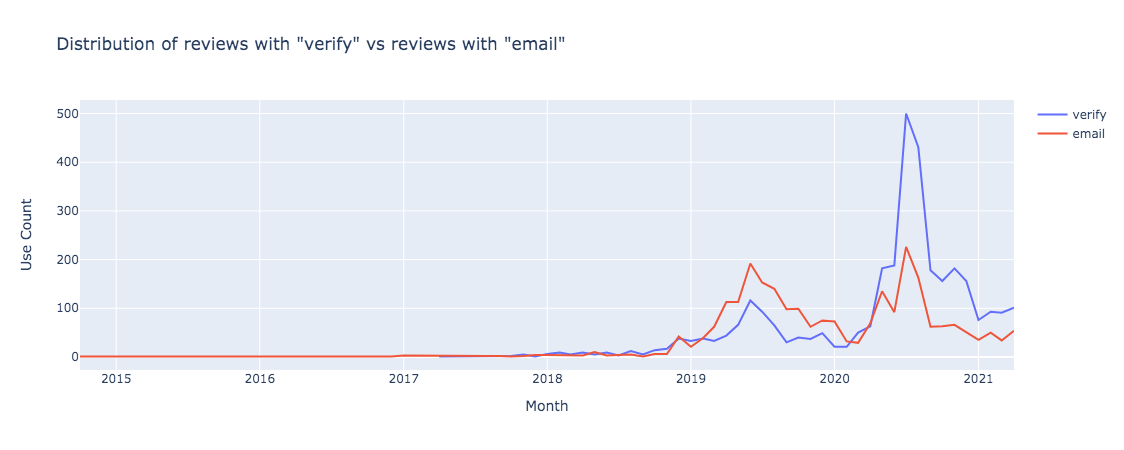
* **“Error”** ratings -> when it comes to error it can be observed that Gcash responds quickly. This can be seen through the huge drops.



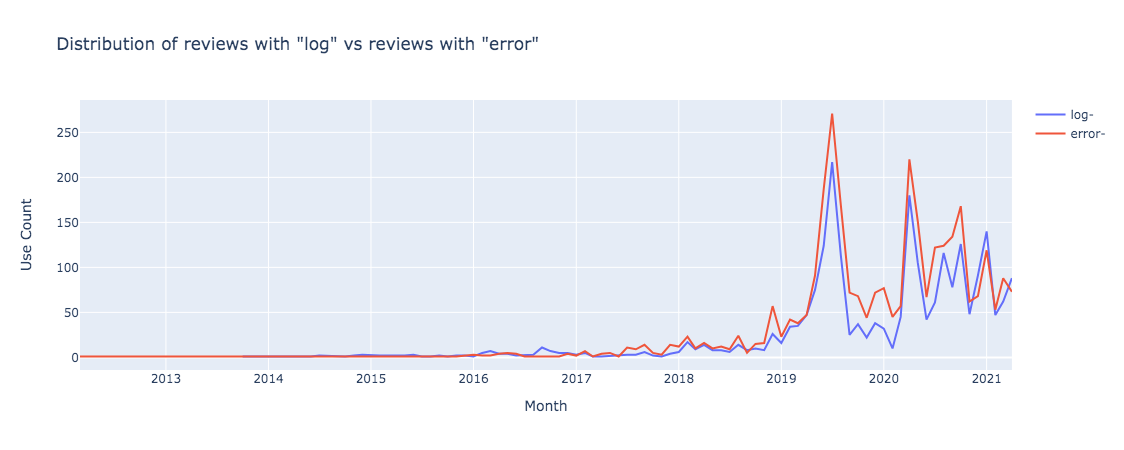
* **“Issue” vs “mpin” -** issues may be related to the mpin ?

****

* **“Verify” and “email” -**

****

* **“Log” and “error” -** most of the errors may occur from log-in errors.



**Rating 1 Overall Topic Probability Proportions**

**Rating 5 Overall Topic Probability Proportions**