Key Insights into MINGAR's Newest Devices and Customers

Demographics of New Users and Performance Issues

Report prepared for MINGAR by SAAS

2022-04-11

Contents

Executive summary

Wearables are a growing and competitive market. To gain market share, businesses must continuously adapt to customer demand and improve the products that they offer. This study analyzes the customer data offered by MINGAR to better understand the target market of their newly launched product lines (Active & Advance). At the same time, the study also seeks to identify potential product improvement by investigating the product complaints recently received by the media team.

Given that the newly launched product line is marketed as more affordable, we estimate that customers of the active and advance wearables have lower incomes. We also predict that quality issue arises from a specific line of product or wearables with a specific type of feature.

The key findings of the study are summarized below.

- On average customers who are older by 1 year has a 0.842% higher odd of being a user of the active and/or advance product; older people tend to like the new product lines more.
- Customer with every $1000 more annual income on average has 2% lower odd of being a user of the new and/or advance product; this products attracts people with lower income.
- Intersex customers on average have a 88.3% higher odd of being a consumer of the active and/or advance product in comparison to females; intersex people really like the new product lines over the old ones relative to men and women.
- Males customers have 4.6% lower odds of being a user of the active and/or advance product than females; more men prefers the old product lines in comparison to women
- The difference in demographics of the active and advance product line customers suggests that the new product lines attracts new customer from the market
- Darker skin people have more flag counts in terms of sleep duration, which indicates their sleep score is low
- Customers of age 15 to 40 reported poor sleep performance than customers of ages 40 and above by 0.505%
- By comparing participants of 16814 females and 11530 males, there showed a similar trend in their sleep score, where the mean score of females were 68.4% and the mean score of males were 67.6%
- Based on customer dataset which includes the code for skin modifier, users with darker skin reported twice as many flag counts than users with lighter skin, proving users with darker skin have lower sleep scores

Technical Report

Introduction

The purpose of this analysis is to explore factors related to devices that MINGAR's customers are using. We analyzed the portion of MINGAR's customers that use traditional devices and the portion of them that use the new devices (Advanced, Active). Since the new devices were designed to be more affordable than the traditional devices, the strategy was to lower the retail price of them. To identify whether the new lines of devices attracted more customers who are not in high-income brackets, we explored the data set of income. Along with income, we examined other customer level data such as age and sex to identify relationships between customers' traits and their preferred device types. We also compared the features of new devices and traditional ones to provide feedback to MINGAR regarding what features may be preferred by customers.

In addition, because the social team of MINGAR reported that customers with darker skin color had more complaints about their device's poor performance than customers with lighter skin color, we analyzed customer level data sets to identify whether their sleep score is linked with certain types of devices. To analyze this concern without collecting information on customer race and ethnicity, a customer's skin tone was approximated through the selected emoji modifier in their device. Through our analysis, we noticed people using darker skin emoji modifiers have lower sleep scores, since they have higher numbers of flags over given sleep duration than users of other emoji modifiers.

Research Questions

- Are customers of the new products (Active, Advanced) from a different income group than the customer of traditional products since the new products are cheaper?
- Are the customers of active and advance products new to MINGAR, or are they MINGAR's traditional customers?
- If consumers of active and advance products are new, how do their demographics differ in terms of age and gender?
- Do users of darker skin exhibit lower sleep scores?
- Do users of darker skin share other common traits? If so, do users of other skin tones with lower sleep scores share the same traits?
- Do people with darker skin tend to prefer a specific device with a specific type of feature?
- Are there specific types of devices or certain features that are linked to lower sleep score?

Methods

Data description and wrangling

We first collected data for the device information by web scraping. We collected census data from the Canadian census using postal codes provided by MINGAR customer data. Next, we did some data wrangling. First, we cleaned the column names of the data which we scraped from the web earlier so that the data type of the column indicating the release dates of devices was of type "date". We then did the same thing for MINGAR's provided dataset containing information on all unique devices owned by customers. Next, we joined those two data to create a new data frame so that we have all information pertaining to all devices in the data.

Then, we merged this data frame with the remainder of MINGAR's provided data, creating a new data frame that links customer information, customer sleep data and device data. This is the data frame that will be used for the analysis we are going to do with the goal of answering the question brought up by your social media team.

Aside from the data frame just mentioned, we also created another data frame that combines all the information we have, including customer and device information. This data frame will be used to answer the questions by your marketing team. We dropped data that does not contain a customer ID because it is how we ascertain what device is owned by them, which causes issues when we are interested in who your customers are and what devices they use. We also rearranged the data so that the customer ID comes in the first column of the data frame.

Next, we cleaned the names of the postal code and median income data. Since there are duplicate values in the postcode dataset, we removed the duplicates. We combined those two datasets and added them to both two full data frames we are using for the analysis since we are interested in the income of your customers.

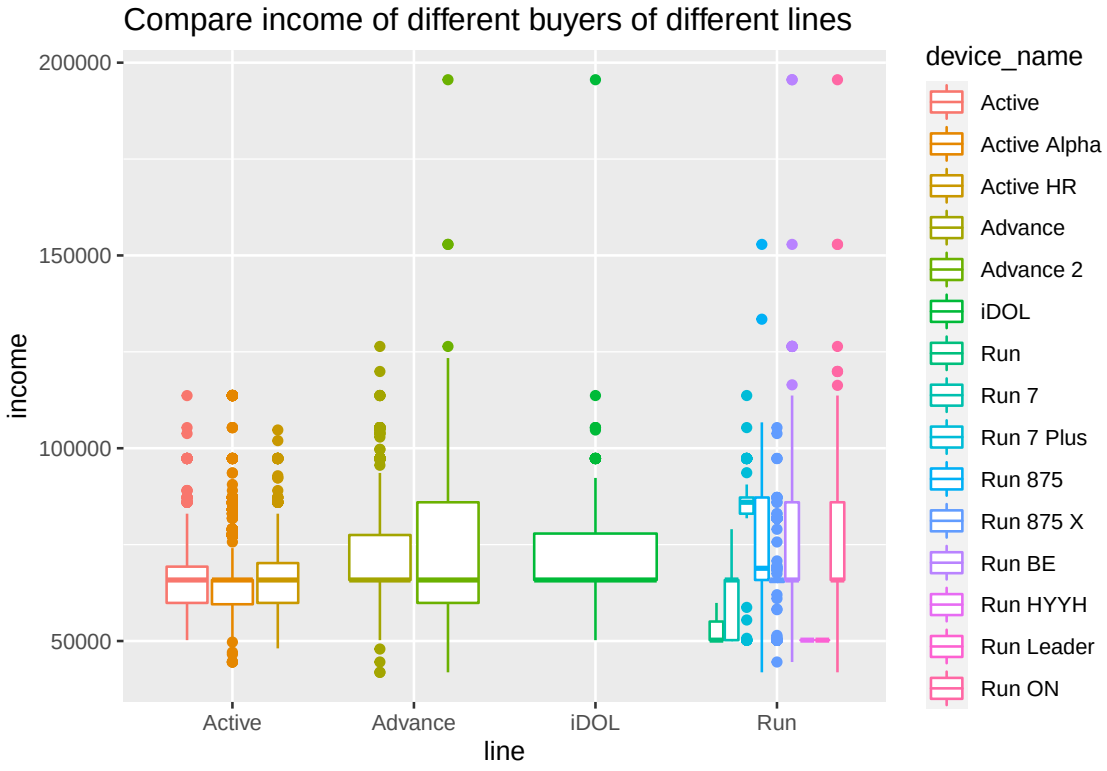
In order to find out how the buyers of the newer products differ from those of the traditional customers, we grouped the customers into two groups. We created a new column in the full data frame we are using in the analysis for your marketing team. This column has a value of 1 if the product is "Active" or "Advance" while the value is 0 if the product is not those two types.

For both final data frames we used, we added a new column that indicates the age of the customers. To figure out why devices are performing poorly for users with darker skin, we were to identify which devices people with darker skin are using and calculate their sleep score. To avoid collecting data on race and ethnicity, a new variable was used for predicting users with darker skin. First, our assumption was based on variables from the customer data such as customer's age, sex, income, and emoji modifier (code for skin color). We assumed poor device performance was linked to low sleep scores, which we can detect using flags. For this assumption, we used a Poisson regression because we considered the average number of flags per 60 minutes for each variables, and determined their significance level.

Analyzing who the Customers are

Building Generalized Linear Model

To analyze the customer profile, we took the 2016 Census data (newest available) to acquire additional information of the customer by matching each customer's postal codes. We used the variable "new" as the binary response and different characteristics of the customer as the predictor. Given that the response variable was binary, we decided to create a link function to the response variable and used a generalized linear model for analysis. In the exploratory analysis of the data, we noticed that the average pricing of the new and active products was lower than older products, which was an indication that the product is targeting customers with lower price points. Therefore, we expected customers who purchased the new and active product to have a lower income.



In an initial exploratory analysis of the data, we notice the mean incomes of Active and Advance customers are relatively similar to the other customers. There are two plausible explanations to this situation: 1. Most customers of the Active and Advance products were customers of the IDOL or Run products 2. Customers of these wearable devices have low price sensitivity in general, which is reasonable as wearable devices are considered luxury products.

To better understand the customers, we analyzed their other characteristics such as age and gender. When fitting the generalized linear model, we believed different genders had different preferences for products. We began by applying a logit link as it allowed us to analyze the % change in probability of being a customer of the new Active and Advance products versus traditional products.

Characteristic	**OR**	**95% CI**	**p-value**
age	1.00843	1.00722, 1.00964	<0.001
income	0.99998	0.99998, 0.99998	<0.001
sex			
Female	—	—	
Intersex	11.4006	3.59088, 37.9642	<0.001
Male	0.66980	0.54575, 0.82212	<0.001
income * sex			
income * Intersex	0.99997	0.99996, 0.99999	0.002
income * Male	1.00000	1.00000, 1.00001	<0.001

After fitting a generalized linear model with all predictors, customers of the active and advanced products exhibited different characteristics in income, gender and age at a 5% significance level. We also explored some possible interaction terms based on intuition.

- More experienced and older people tend to have a higher income
- Income differs between man and woman

However, we could see from the exponentiated results that the coefficients for both interaction terms are extremely close to 1. Although they are both statistically significant, the impact on the odds of being customers of new products is negligible, so we decided not to include the interaction terms.

Discussion

Identify customer demographics through general linear model

The model is recentered based on median age and income for a female customer, so the intercept represents the odds of a 46 years old female customer with $65,829 annual income being a consumer of the active and/or advance product line. Based on this analysis, on average customers who are older by 1 year have 0.842% higher odds of being a user of the Active and/or Advance products. For every $1 increase in annual income, customers have 2% lower odds of being a user of the new Active and/or Advance products. In terms of gender, intersex customers on average have 88.3% higher odds of being a consumer of the active and/or advance product in comparison to females. And male customers have 4.6% lower odds of being a user of the Active and/or Advance products.

Characteristic	**OR**	**95% CI**	**p-value**
age	1.00842	1.00721, 1.00963	<0.001
income	0.99998	0.99998, 0.99999	<0.001
sex			
Female	—	—	
Intersex	1.88279	1.53147, 2.32858	<0.001
Male	0.9540	0.91516, 0.9945	0.026

Looking at the 95% confidence interval, we are confident the coefficient for age is above one, which indicates that the new product line is more frequently purchased than the traditional product lines by older individuals. The coefficient for income being negative on both ends supports our initial prediction in that customers of the new product lines tend to have lower incomes. The coefficient for intersex significantly above 1, showing a strong preference of the new product lines over the traditional product by the intersex relative to females. The coefficient for males being both lower than 1 signifies that more men prefer the active and advanced product over other product line in comparison to females. These differences in demographics suggest that the Active and Advance product lines attract new consumers that were different from their traditional product lines.

Exploratory Analysis of Sleep Scores

Since the social media team of MINGAR had mentioned that they had been tracking a trend in complaints saying that MINGAR's devices were performing poorly for users with darker skin, our SAAS team has investigated customers' general information that can be linked to sleep score. This includes their date of birth, income, sex, and emoji modifiers that they use in their devices.

We have defined sleep scores using the sleep data for each customer in terms of duration and flags. Since flags represent the number of times there was a quality issue during a sleep session, not adjusted for errors by the device, we assumed more flags were indicative of a lower quality in sleep, which meant the sleep score would be lower.

We began by defining the response variable as the number of flags recorded in a single sleep session. Our response was a count, so a Poisson regression was considered where the parameter of interest was the average number of flags reported. We primarily explored the relationship between the flag count and the emoji modifier used by a customer, controlling for other covariates such as duration, sex, age, income, and the device used.

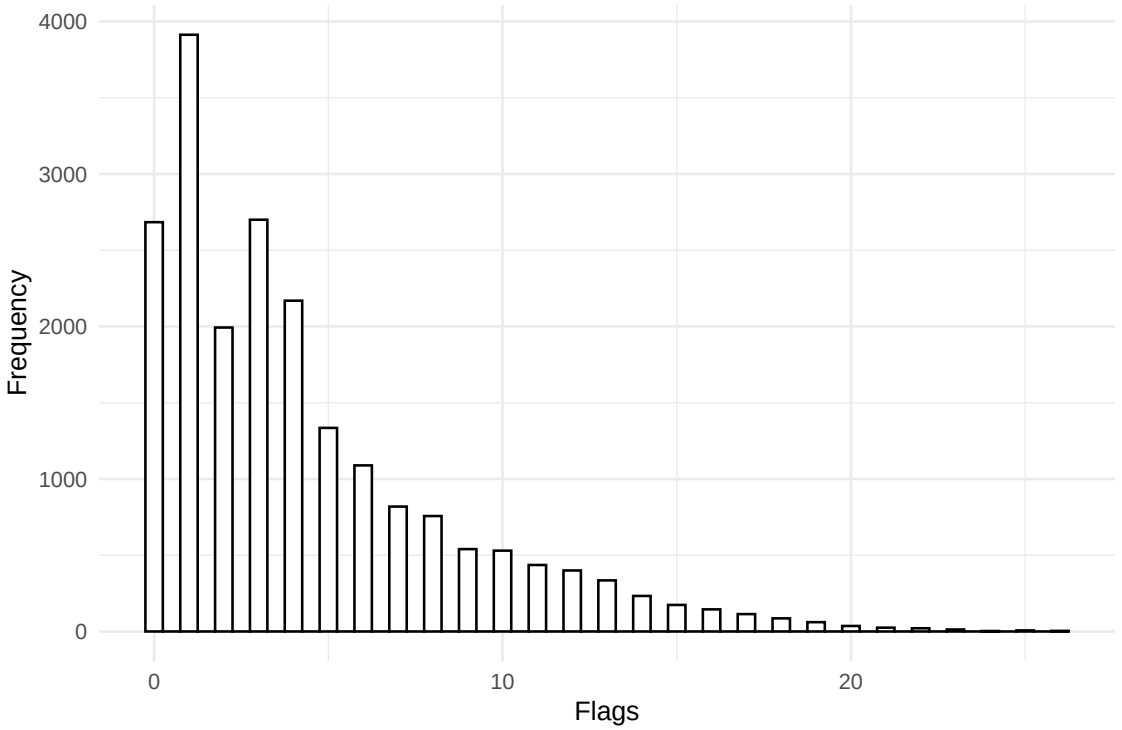


Figure 2.1: Distribution of Flags for all Sleep Sessions (Histogram)

This figure reveals that there is a fair amount of variability present in the number of flags recorded during each sleep session. Responses ranged from 0 to 26 with many reporting between 0 and 4 flags. As in typical Poisson distributions, this graph is visibly right skewed. It clearly suggests that the number of flags is not a normally distributed response, and that a Poisson regression is preferred instead.

However, it is important to note that the flag counts are not directly comparable because they come from different sleep durations. We expect sleep sessions with longer durations to have more flags since a longer time frame allows for more chances for flags to occur. It is possible that there are also relationships between sleep duration and other variables such

as age, so it is worth investigating them in our EDA. For the remainder of this EDA, we can take differences in duration into account by looking at flag counts as a rate of per 60 minutes, where applicable.

Variable Selection

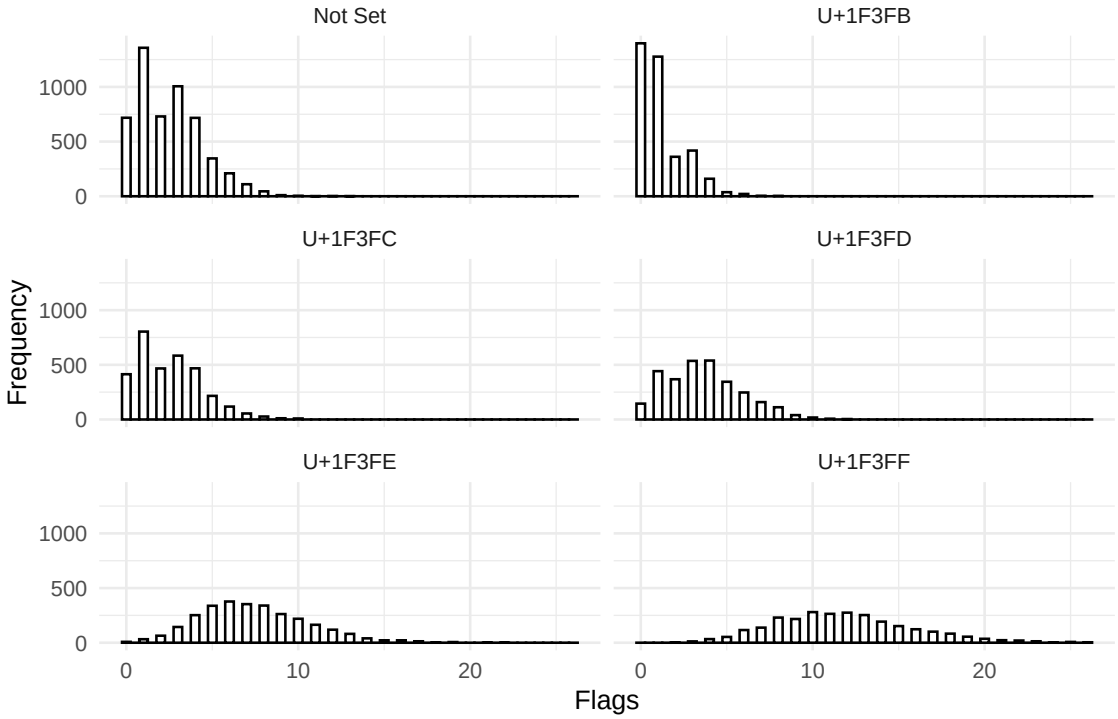


Figure 2.2: Distributions of Flags by Emoji Modifier (Histogram)

$EmojiModifier$: To extrapolate the skin colors of customers, the emoji modifier selected in a customer's device was assumed to represent their skin tone. Codes U+1F3FB to U+1F3FF represent the lightest to darkest skin tones. If an emoji modifier was not set, it is defaulted to the yellow skin tone.

Looking at figure 2.2, it can be seen that distributions of flag counts for darker skinned customers are noticeably different, with dark tones correlating to higher flag counts. Looking at data summaries across different groups further elaborates on this. It can be seen that those who use modifier U+1F3FF have a median flag rate that is over 13 times that of those who use modifier U+1F3FB, with modifiers in between following a similar trend to lesser extremes. These results suggest that users with darker skin have considerably lower sleep quality and that their devices perform poorly on them.

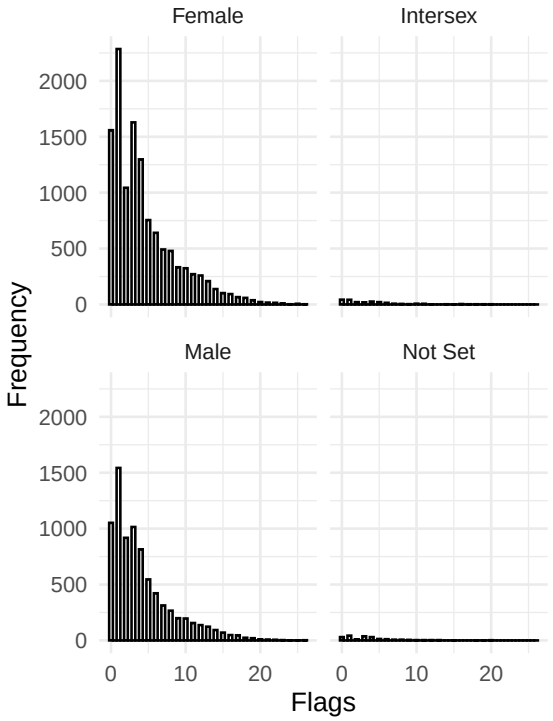


Figure 2.3: Distributions of Flags by Sex (Histogram)

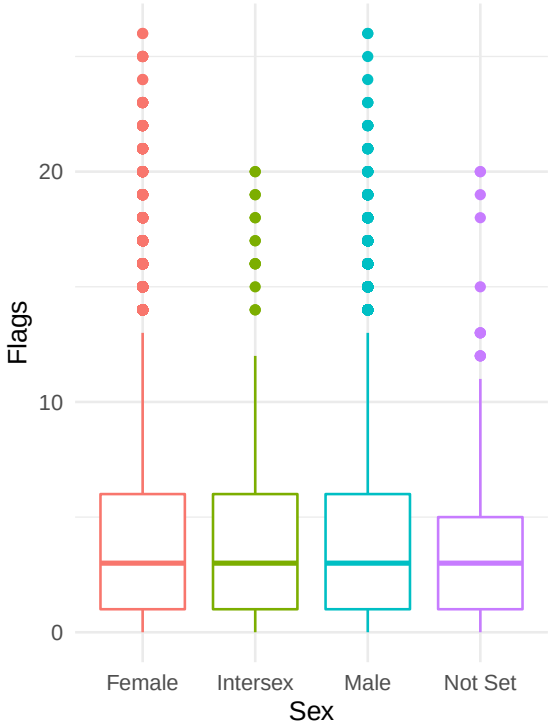


Figure 2.4: Distributions of Flags by Sex (Boxplot)

Sex: Figures 2.3 and 2.4 compare the four groups of sex (female, intersex, male, not set) in terms of flags. There appears to be no significant differences in sleep score between all groups and their distribution patterns are consistent with each other, which indicates that sex is likely not a differentiating factor for sleep scores. Any discrepancies with intersex customers, as well as those who have not set a gender, can be explained by their relatively small sample size.

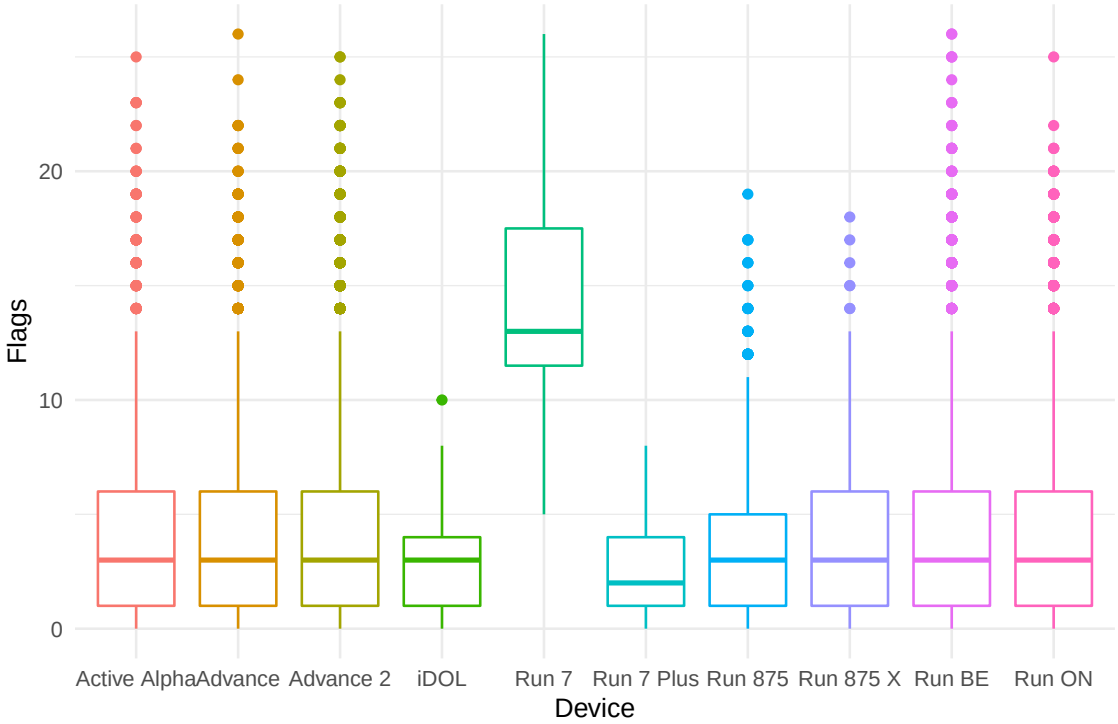


Figure 2.5: Distributions of Flags by Device (Boxplot)

DeviceName: There are ten different types of MINGAR's devices, where three out of

ten are from newer product lines, and the rest being from traditional ones. Figure 2.5 illustrates how different devices compare in the number of flags raised. Similarly to sex, there are no significant differences between the products, as "Run 7" and "Run 875 X" devices have relatively small sample sizes, explaining their more unusual boxplots. However, it is worth noting that devices from "Active" and "Advance" lines have somewhat higher mean flag rates, so this will be further investigated when modeling.

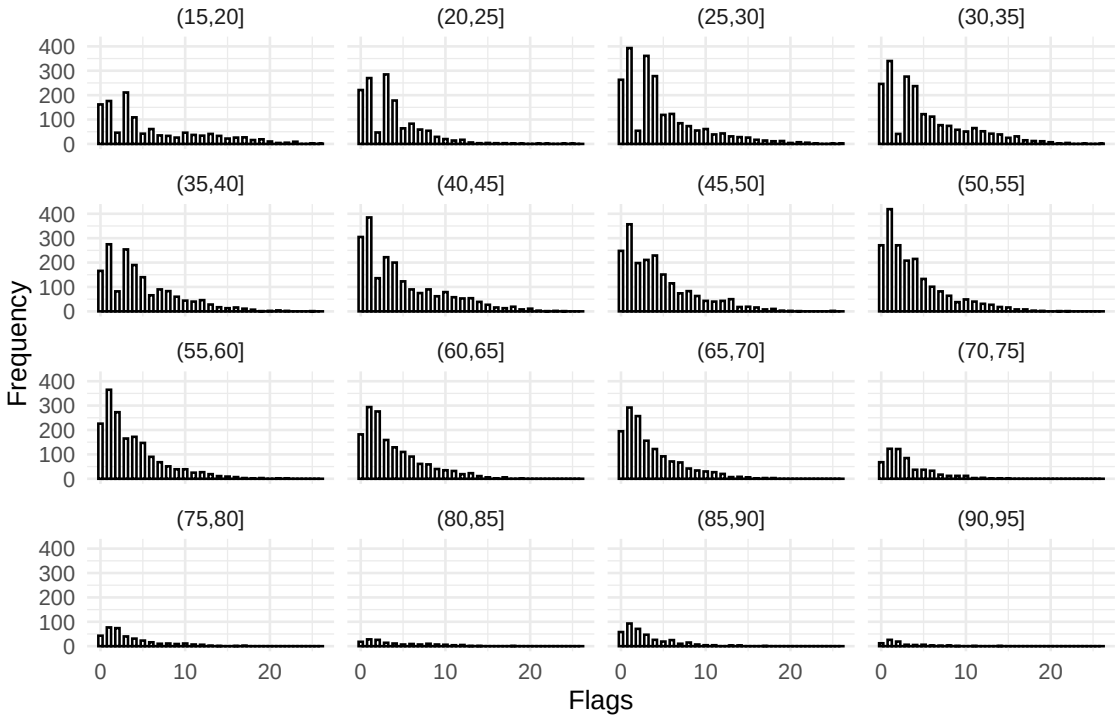


Figure 2.6: Distributions of Flags By Age Bracket (Histogram)

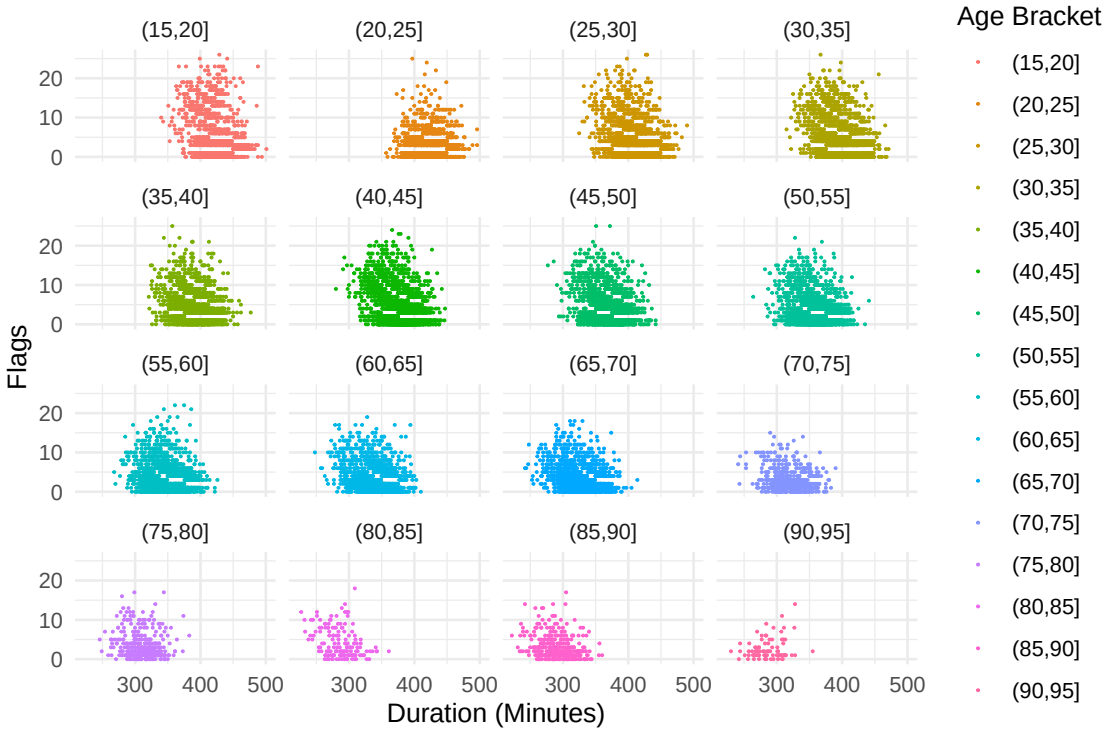


Figure 2.7: Flags vs Duration, by Age Bracket

Age: Age was considered as a covariate in the model, as people of different ages may experience different quality of sleep. Looking at figure 2.6, when comparing flags with

respect to age, younger customers appear to have more flags on average than those who are older. As age gradually increases, the expected number of flags decreases, suggesting that relatively younger customers, particularly from ages 15 to 45, have a worse sleep score.

However, figure 2.7 offers crucial insight as to why this is the case. Figure 2.7 shows that as age increases, sleep duration decreases, which explains the previously established relationship between age and flag counts. This gives us strong incentive to compensate for different sleep durations in our model.

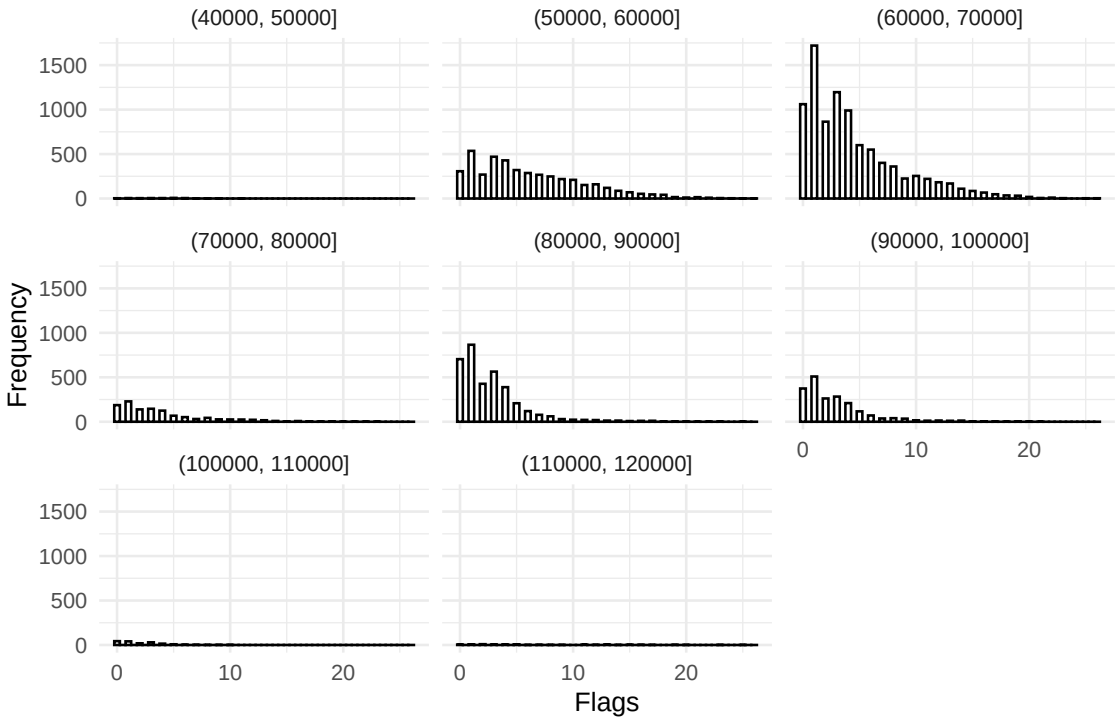


Figure 2.8: Distribution of Flags By Income Bracket (Histogram)

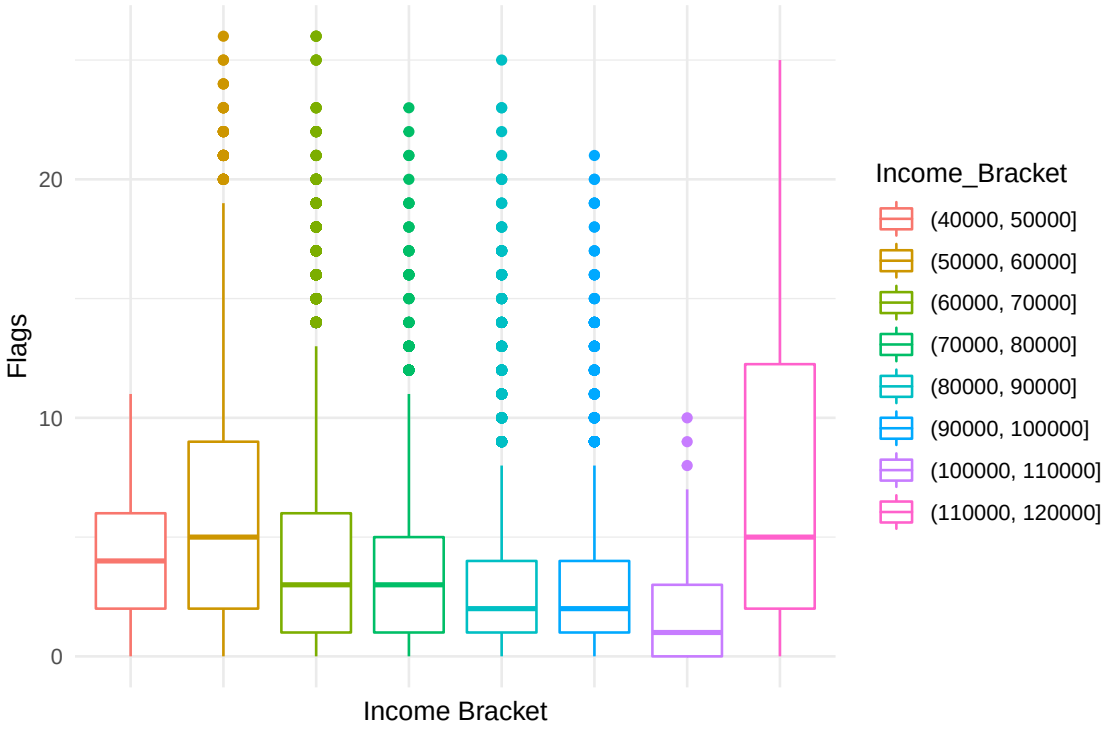


Figure 2.9: Distribution of Flags By Income Bracket (Boxplot)

Income: Defined as the median household income for a customer. This variable was selected with the intention of predicting the relationship between income and flags as well as investigating the impact income has on customer's device types, if any. To elaborate, since the new devices are more affordable than traditional devices, our aim was to examine whether people with lower income are purchasing the new devices at a higher volume than people with higher income, which is important if device is a significant predictor.

Figures 2.8 and 2.9 reveal a clear trend showing that those in higher income brackets record a lower number of flags on average. Unlike with age, adjusting for duration is not helpful here. Thus, further investigation is required.

More on Model Assumptions

A Poisson regression model assumes that the response variable can be described by a Poisson distribution. As we have already seen, there is sufficient reasoning to believe this is the case.

It is also assumed that the mean is equal to the variance at all groups of X. Comparing the means and variances across each age group and income bracket shows some evidence of assumption violations, although modest.

Linear mixed models assume that our customers are independent, even through observations from the same customer are taken not to to be. Other assumptions state that random effects come from a normal distribution, and random effects errors and within-unit residual errors have constant variance. A Normal Q-Q plot and a plot of residuals vs fitted values allow us to see that is indeed the case, with small violations.

The Poisson regression model also assumes that the log mean number of flags is a linear function of covariates such as age and household median income. To check this assumption, we would plot the log mean number of flags by age, as well as by income. However, the true rate parameter is unknown as we do not actually have all the observations available to us, so the best we can do is to use the observed means, take their logs, and plot them.

Plotting these curves suggests weakly linear relationships between the log of the mean number of flags and age and income, if any relationship exists at all. This behaviour is consistent across all emoji modifiers and suggests that it may not be necessary to control for either these covariates in our model.

Model Diagnostic

GLMM: $log(Flags) = -0.939 - 0.764 \cdot \text{U+1F3FB} + 0.00817 \cdot \text{U+1F3FC} + 0.421 \cdot \text{U+1F3FD} + 1.128 \cdot \text{U+1F3FE} + 1.632 \cdot \text{U+1F3FF} + log(Flags * 60/Duration)$

In our final model, we have chosen the variable flags as a response (dependent) variable, with emoji skin color code (emoji modifier) as the predictor (independent) variable. We used a generalized linear mixed model (glmm) to fit our model because emoji skin color code was treated as a fixed effect, and unique customers were treated as random effects. From the EDA, it became apparent that responses were not normally distributed, so we selected a Poisson distribution for our analysis.

Since we assumed all customers to be independent, but observations from the same customer to not be independent, the customer id was used to set a random effect in the glmm model. Then, the emoji modifier was set as a fixed effect as the goal is to investigate how flag rates vary across each of them. In addition, a term was included on the right side as an offset, representing the number of flags per 60 minutes. This term was included to account for the fact that longer sleep durations would result in higher flag counts. Duration itself is not a significant predictor variable, so we have no problems doing this.

Furthermore, the p-values for emoji modifier codes are <2e-16, making our results statistically significant. We concluded that income, age, sex, and device type were not significant predictors in estimating sleep score, and that there is no significant relationship between the four variables and sleep score i.e. the rate of flags.

Table 3: Table 2.1: Coefficients and Confidence Intervals Final Model

	Estimate	95% CI
Baseline rate	0.39	(0.38, 0.40)
U+1F3FB	0.47	(0.45, 0.49)
U+1F3FC	1.02	(0.99, 1.05)
U+1F3FD	1.52	(1.48, 1.57)
U+1F3FE	3.10	(3.03, 3.18)
U+1F3FF	5.13	(5.01, 5.25)

We can interpret the coefficients in our model in the following manner: The estimated coefficient value of -0.764 indicates that the mean flags per 60 minutes for those who use the emoji modifier U+1F3FB is about 0.466 ($e^{-0.764}$) times that of those who do not have an emoji modifier set. On the hand, those using U+1F3FE can report 5.11 ($e^{1.632}$) times as many flags as those who do not have an emoji modifier set.

Limitations

Analyzing the active and advance product line together as one may ignore some differences between them. The active wearables tend to be more sustainable while the advance product line focuses more on the advance features. As a result, the two product line may attracts different type of customers. It is recommended to further analyze them separately to provides the most accurate customer profile for each line.

While the general linear model enables us to compare the customers of the new product lines with the traditional product lines. It is measured in the relative basis. It is possible that the majority of customers of the new product lines are males despite a smaller portion in comparison to the old product line. In this case, the decrease in odds of using new product for male customers can be misleading when identifying the target customer of the new product line.

Emoji modifier is an indirect representation of populations of different skin tones, and may not be an accurate representation of a customer's skin tone. Furthermore, customers may not bother changing their emoji modifier for various reasons, such as not feeling represented by any of the given skin tones. As such, it is difficult to interpret this group. However, this group is a significant portion of users, so they should not be disregarded. It is suggested that more data should be gathered relating to ethnicity in order to perform a more accurate analysis, if possible.

Due to computational resources restrictions, more complex models could not be evaluated. Possibly significant predictors such as age and household median income were difficult to formally evaluate. This may violate the assumption that the model is correctly specified, with covariates not being controlled for when estimating coefficients for emoji modifier.

Consultant profiles

Sallie (Kyong Ei) Kim. Sallie is a senior consultant with SAAS. She specializes in cyber security analytics and cyber risk consulting. She earned her Bachelor of Science, Major in Statistics and Mathematics, from the University of Toronto in 2022.

Andy (Wei Qi) Lin. Andy is a junior consultant with SAAS. He specializes in data analytics and finance. He earned his Bachelor of Science, Major in Economics and Statistics, from the University of Toronto in 2021, and earned his MBA from the University of Toronto in 2023.

Alexander Tran. Alex is a junior data engineer with SAAS. He specializes in data mining and data visualization. He earned his Bachelor of Science, Specialist in Data Science, from the University of Toronto in 2021.

Sharon (Fang-Chu) Lin. Sharon is a senior data scientist with SAAS. She specializes in natural language processing (NLP) machine learning. She earned her Bachelors of Science, Specialist in Computer Science, from the University of Toronto in 2020 and is currently in her last year of master's degree in machine learning, at the University of Toronto in 2022.

Code of ethical conduct

1. Inclusion and respect: Responsible treatment of individuals

- Promote adherence to the Code by other subscribing consultants.
- Consultants must abide by applicable employment standards, labor, non-discrimination, and human rights legislation
- Consultants must maintain workplaces characterized by professionalism and respect for the dignity of every individual with whom their employees interact. Consultants must not tolerate harassment discrimination, violence, retaliation, and other disrespectful and inappropriate behavior.

2. Confidentiality: Data (Consensus API, postcode) are confidential (web scraping)

- Respect the intellectual rights of professional peers and not use their copyrighted information, materials, or methodology without permission.
- Refrain from misrepresenting or denigrating professional peers

3. Objectivity: Our statistical analysis is based solely on facts (e.g. data collected from MINGAR's customers, 2016 Consensus)

- Serve their clients with integrity, honesty, competence, independence, objectivity, and professionalism, helping them to build their philanthropic expertise and capacity.
- Only accept assignments that they, their staff and their colleagues have the necessary experience and competence to perform effectively.

References

Census subdivision types by province and territory, 2016 Census from the Dictionary, Census of Population, 2016. Retrieved from www12.statcan.gc.ca/census-recensement/2016/ref/dict/tab/t1_5-eng.cfm on 27 March, 2022

Dmytro Perepolkin (2019). polite: Be Nice on the Web. R package version 0.1.1. https://github.com/dmi3kno/polite

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Hadley Wickham and Evan Miller (2021). haven: Import and Export "SPSS", "Stata" and "SAS" Files. https://haven.tidyverse.org, https://github.com/tidyverse/haven, https://github.com/WizardMac/ReadStat.

Hadley Wickham (2021). rvest: Easily Harvest (Scrape) Web Pages. https://rvest.tidyverse.org/, https://github.com/tidyverse/rvest.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Jared E. Knowles (2020). eeptools: Convenience Functions for Education Data. R package version 1.2.4. https://github.com/jknowles/eeptools

Lüdecke D (2021). *sjPlot: Data Visualization for Statistics in Social Science*. R package version 2.8.10, <URL: https://CRAN.R-project.org/package=sjPlot>.

Sam Firke (2021). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.1.0. https://github.com/sfirke/janitor

Sjoberg DD, Whiting K, Curry M, Lavery JA, Larmarange J. Reproducible summary tables with the gtsummary package. The R Journal 2021;13:570–80. https://doi.org/10.32614/RJ-2021-053.

von Bergmann, J., Dmitry Shkolnik, and Aaron Jacobs (2021). cancensus: R package to access, retrieve, and work with Canadian Census data and geography. V0.4.2. Retrieved from https://censusmapper.ca/ on 27 March,2022

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Yihui Xie (2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.37.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

Code Appendix

Web scraping industry data on fitness tracker devices

```
url <- "https://fitnesstrackerinfohub.netlify.app/"
# Introduce self to owner, and check if we have permission to scrape data, and if
↪  yes, under what conditions
target <- bow(url,
              user_agent = "fangchu.lin@mail.utoronto.ca for STA303/1002
              ↪  project",
              force = TRUE)
target

html <- scrape(target)

device_data <- html %>%
  html_elements("table") %>%
  html_table() %>%
  pluck(1)
```

To do web scraping, we use the bow function to introduce ourselves to the owner of the url, and also check if the website is scrapable. From the result of the bow function, we notice that the craw delay is 12 seconds, which means we should limit our rate to 12 seconds while scraping. The web scraping on fitness tracker devices is for getting information about the devices. We are doing web scraping on a public API and we provided a User Agent string to be clear on why we need this data and also provided the contact information so that the owner could contact us if there are any concerns.

Accessing Census data on median household income

```
options(cancensus.api_key = "CensusMapper_a65403b98524cd275dcf3ad2daef70c7",
        cancensus.cache_path = "cache") # this sets a folder for your cache


# get all regions as at the 2016 Census (2020 not up yet)
regions <- list_census_regions(dataset = "CA16")
```

```
## Reading regions list from local cache.
```

```
regions_filtered <-  regions %>%
  filter(level == "CSD") %>% # Figure out what CSD means in Census data
  as_census_region_list()
```

```
# This can take a while
# We want to get household median income
census_data_csd <- get_census(dataset='CA16', regions = regions_filtered,
                          vectors=c("v_CA16_2397"),
                          level='CSD', geo_format = "sf")
```

```
## Reading vectors data from local cache.
```

```
## Reading geo data from local cache.
```

```
#Simplify to only needed variables

median_income <- census_data_csd %>%
  as_tibble() %>%
  select(CSDuid = GeoUID, contains("median"), Population) %>%
  mutate(CSDuid = parse_number(CSDuid)) %>%
  rename(hhld_median_inc = 2)
```

Since we are interested in the income data, we accessed the Canada CensusMapper and got an API key after signing up. We use our unique API key to get the 2016 census dataset and filter to only get the regions where the level is CSD, which means census division in the dataset. We then use the 2016 census dataset and the regions we are interested in to collect the household median income, simplified to only keep the id, the median income, and the population, which we may want to use in the project.

Accessing postcode conversion files

```
postcode <- read_rds("data-raw/break_glass_in_case_of_emergency.Rds")
```

We obtained the postal code conversion file by accessing the Census Canada Postal Code Conversion Files from the University of Toronto Libraries website. Before downloading the data, we are asked to accept the license agreement. To accept this agreement, we agreed to only use data for our learning purpose and will not make the data public.