# MINGAR Customer Analysis

The Features of the New Customers and the Reason Behind Flag Raising

Report prepared for MINGAR by Four Point Zero, LLC

2022-04-07

# Contents

## Executive summary

**Background & Aim**

As a company specializing in high-end fitness tracking devices, Mingar wants information on their strategy for the Canadian market so they can better compete with Bitfit and gain more market share. Bitfit also offers fitness tracking devices but their devices are more affordable and their apps offer more insights. To compete with Bitfit, Mingar launched the "Active" and "Advance" lines at a more approachable price. In this study, we will find out how the buyers of these affordable lines differ from the traditional customers to see if these new lines have attracted customers outside the traditionally higher-income base. In response to recent complaints, we will also investigate whether Mingar's devices are performing poorly for darker skin users, particularly about sleep scores.

**Key Findings**

- The age, median income, and population contribute to the main difference between our new customers and traditional customers. Our new customers' groups were elder with lesser median income and population size than the traditional customer (Fig 1.1).
- Mingar should continuously make the products more affordable (with the same quality) to attract more new customers. Mingar may also consider upgrading and installing products to become more elder-friendly. Finally, since new customers are generally living in lower population level subdivisions, the advertisements and marketing are important for Mingar to promote new products lines and increase the market size.
- Racist label is credible based on our statistical analysis. Flags counts are an indicator of quality issues, which reflects the performance of the device. The mean flag count for dark skin emoji users is the highest among all skin-color users, with light skin users the least. The skin color of emoji is a significant predictor of the rate of flag counts. As skin color goes from light to dark, the rate of flag counts increases (Fig 2.1).
- Darker-skin users tend to be young. The user's age is highly correlated to the number of quality flags that occurred during the sleeping session. They possess a negative relationship. This suggests that the behaviors and characteristics of young users may be causing the device to perform poorly.

**Limitations**

- The response variables are selected only based on the explanatory data analysis and a general understanding of the research topic. The model selection process may not be as accurate as of the other statistical methods where all cases are considered in that case.
- Postal code and census subdivisions (CSDuids) do not match in their size. One postal code has several CSDuids. The median income used is the subdivision's median income, instead

of the individual's income. Such demographic information contains inaccurate information about our customers and may result in biased estimates.

- For the sleep model constructed, the existence of random effect is a paradox. Without random effect, the independence assumption of the model is violated. Incorporating the random term doesn't explain the variance in response better.
- We removed missing values, which are default yellow emoji in the data used to fit the model. A user using a specific emoji does not ensure the user herself is that skin color. Such observations along with obscure definition lessen the training sample size and increase the inaccuracy of the model.

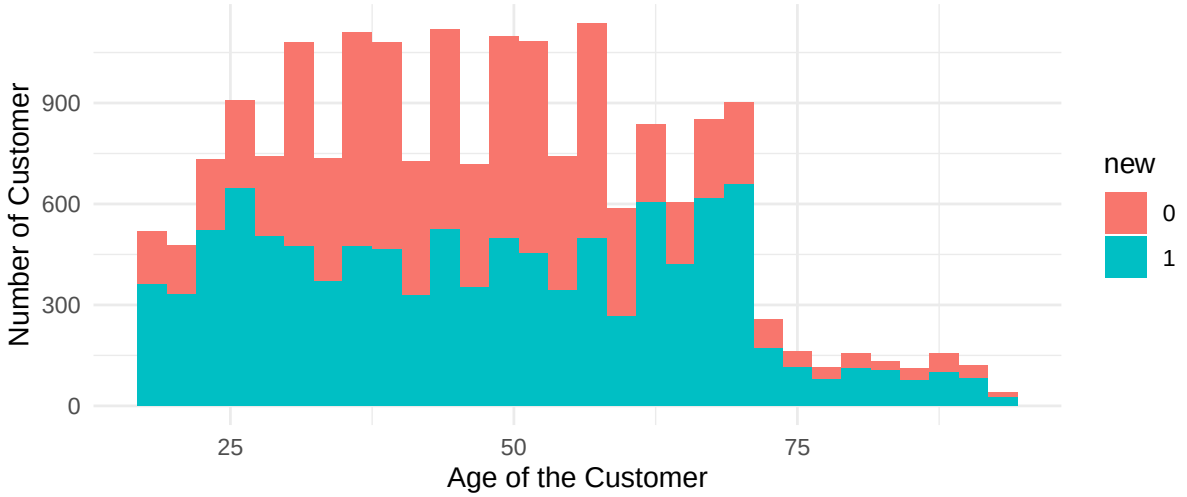


Fig 1.1: Histogram of the distribution of the customer ages between new and old customer

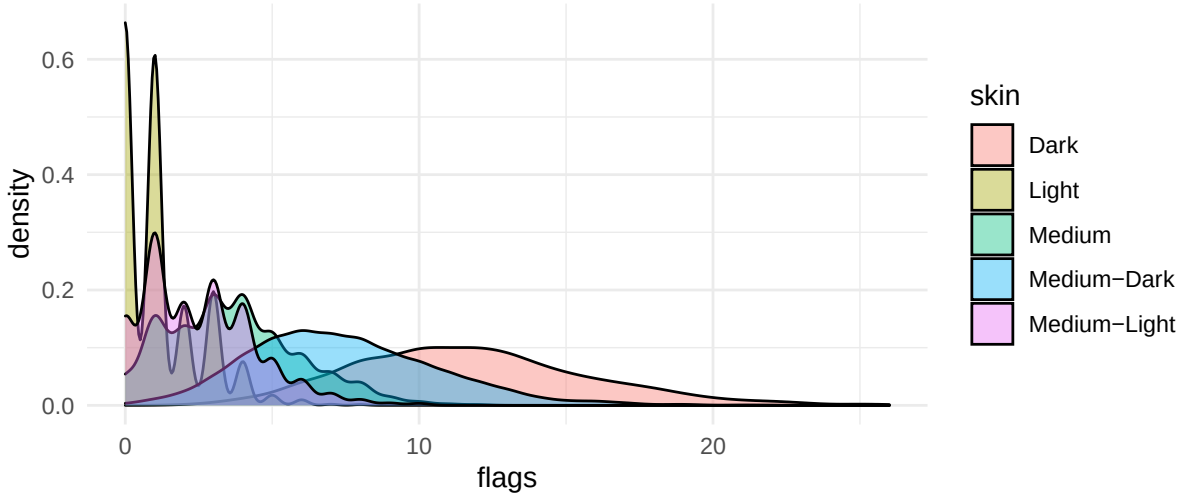


Figure 2.1: The distribution of number of quality flags

# Technical report

## Introduction

Our team, Four Point Zero, LLC, conducted a statistical analysis to investigate two claims from our client MINGAR, LLC (see below in section Research questions). Based on two claims, the goal of our team is first to conduct a statistical analysis to provide information on their Canadian marketing strategy of Advance and Active line products; and second to investigate the reason behind the racist charge on their products. In this report, we provide a detailed explanation of how we collected our data for statistical usage, including data accessing, and wrangling. There is a complete exploratory data analysis. Statistical methods employed in answering two questions along with their results are presented. The statistical interpretation and realistic application of the model results are emphasized. Lastly, we summarize the limitation that existed in the two models and draw a conclusion on the objectives. After statistical analysis, we expect the client's claims can be addressed.

### Research questions

- Mingar recently published two new lines of products, Advance and Active. The request from the marketing team is to examine the purchasing group behind new lines of products and provide information for Mingar to construct its strategy in the Canadian market. Especially, what kind of customers should be their target? What are the features of customers who purchase new products compared to those who buy traditional products?
- The products of Mingar are censured with the racist label as the devices are poorly performed for darker skin users. The request from the social media team is to explore the reason behind such criticism.

## Data Manipulation

Based on the dataset given, the customer-level data was left joined with the postcode conversion file to obtain their CSDuid. The first pair of CSDuid and postcode were kept to make sure one CSDuid corresponds to the postcode. Next, the data was further left-joined with the Census Mapper API to obtain customers' regional median income and population. Then, we merged the dataset with sleep data for each customer. We further left-joined the dataset with customer-device linkage information data to match each customer to their current device. Lastly, the dataset was left joined by device data to obtain device information. To prevent re-identification risk and the disclosure of private information, the postcode of customers was eliminated from the dataset to prevent disclosure of private information.

We filtered the customer data by removing all the missing values from the sex. We then calculated the age of each customer by using today's date minus the date of birth. The age is rescaled to a minimum of 0 and a maximum of 1. 0 corresponds to the minimum age which is 13.94, while 1 corresponds to the maximum age, which is 90.87. We defined the customers as new if the customer purchases products from the Active or Advance line.

In the dataset, there is no direct variable that tells the race of the users. Rather, we used "code for skin tone modifier for emojis the user used" as a proxy variable to race. We create a variable "skin" to match the color code to its actual skin color based on the Fitzpatrick scale. Missing values of the skin color indicate the user uses the default yellow emoji. However, the default emojis do not provide an explicit clue about the users' race. Thus, they are removed from the dataset. Lastly, we selected the variables that are useful in further analysis. After data wrangling, the final dataset we used to fit models contains 15243 observations with 13 variables.

In the entire analytical process, variables contain private information about our customers. It is possible to identify the customer even if their names are removed. To prevent re-identification risk and the disclosure of private information, we shall not release our original data to prevent disclosure of private information.

**Table 1:** Description of the important variables

| Variable | Description |
|---|---|
| sex | The sex of the customer |
| CSDuid | Census Division Unique Identifier of the |
| hhld_median_inc | Regional household median income of the customer |
| Population | Population number corresponding to the CSDUid living by the customer |
| age | The age of the customer |
| new_customer | The binary indicator that indicates whether the customer is new or not |
| cust_id | the id of the customer |
| flags | number of quality flags during sleeping session |
| duration | the duration of the sleeping session |
| skin | skin color of the user mutated based on emoji the user used |
| line | the line of device that the user used |

## Features that Differentiate the New and Traditional Customers of Mingar

### Exploratory Data Analysis

**Table 2:** Numerical Summaries of the data

| | Minimum | | Maximum | | Median | | Mean | |
|---|---|---|---|---|---|---|---|---|
| Variables | New | Old | New | Old | New | Old | New | Old |
| Income | 41880 | 41880 | 195570 | 195570 | 65829 | 65829 | 68814 | 73168 |
| Population | 3914 | 2813 | 2731571 | 2731571 | 1704694 | 1239220 | 1519844 | 1478529 |
| Age | 17 | 17 | 92 | 92 | 47 | 47 | 47 | 46 |

From the numerical summary, we observed that the minimum and maximum income, population, and age are almost the same. This suggests that there is little or no difference in the range of new and old customers. The mean and the median income for the old customers are higher than the new. It matches the new product offerings that the new product is more affordable for the lower median income group. The mean and median values of the age of the new customer are higher than those of the old customer. Thus, the new product is more attractive to the elder customers.

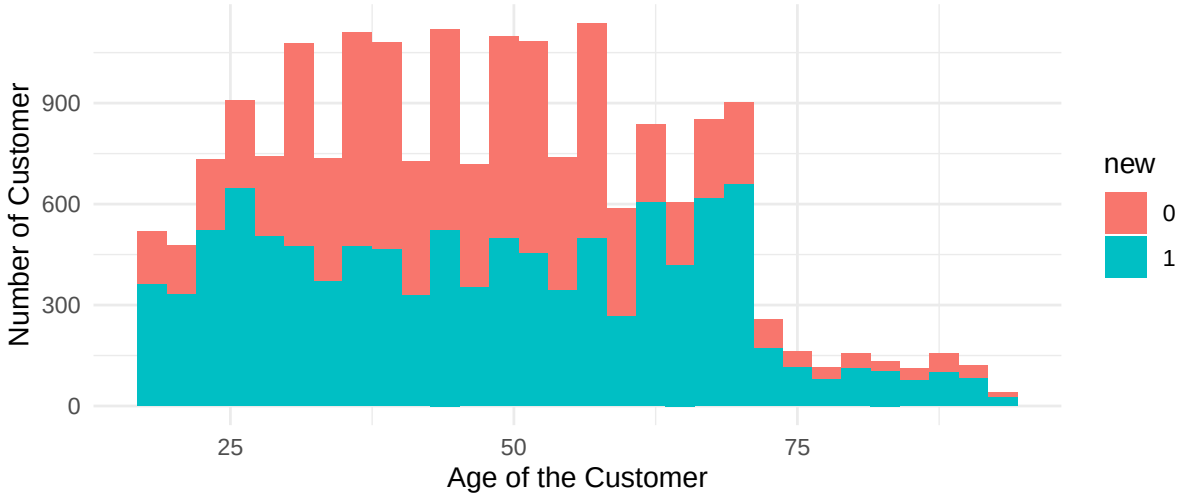


Fig 1.1: Histogram of the distribution of the customer ages between new and old customer

Fig 1.1 is the histogram of the customer's age compared between new and old customers. From the graph, we could observe that for both new and old customers, the graph is slightly right-skewed. This shows that the main customer is still young adults. However, for the elderly customer, the proportion of new customers is significantly higher than the old customer. This means that our

new products attracted more elder customers compared to the young.

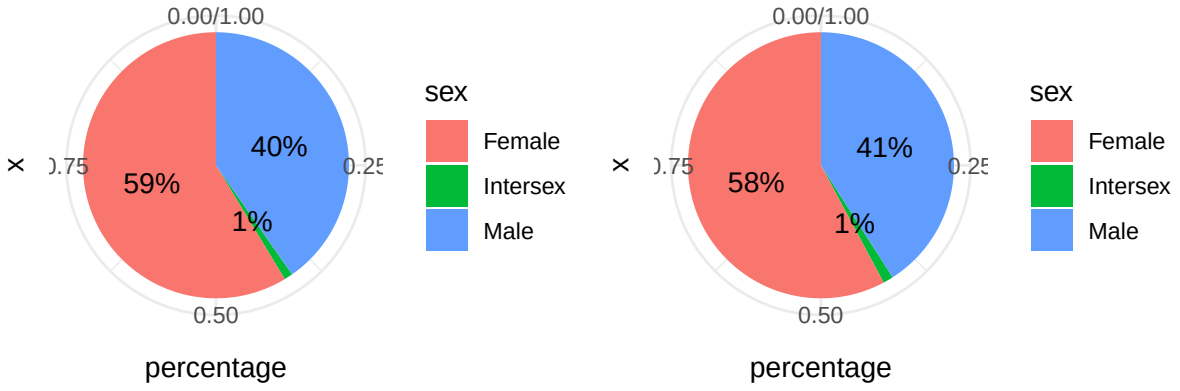


Figure 1.2: Pie charts of the gender percentage between new customer and old customer

Fig 1.2 is the pie chart of the proportion of customers from different gender compared between new customers and old customers. Compared to the new customer with the old, we do not observe any significant change in the proportion of the gender. There is a slight increase in the proportion of the male customer which demonstrated that our new product may be slightly more attractive to the male customer.

### Statistical Methods

We relied on the exploratory data analysis, real-life rationals, and the client's claim to construct a statistical model. Based on data visualization, we observe that sex, age, median income, and the population of subdivisions seem to influence whether a customer buys new line products. They were chosen as potential fixed effects in our model. Other original variables might be invalid for us to approach the research aim, such as device name and the released date. Besides, since the number of census subdivision unique identifiers is 256, extremely smaller than the total sample observations of 19045, it was selected as the random effect in the initial model. The response, new customers, is binary; thus, the initial model is the generalized linear mixed model. To avoid inaccurate analysis, a log transformation was taken on population and median income.

To select the statistically significant fixed predictors, we relied on several likelihood ratio tests and p-values of fixed effects. Since the models being compared in each test differ only in fixed effects, the likelihood ratio test is feasible to compare two nested models' goodness of fit. We concluded the better model in each comparison based on the test's p-value. A p-value less than the significance level of 0.05 indicates that adding the additional fixed effect is meaningless. After four pairs of likelihood ratio tests, we could obtain a preferred model with the most ideal fixed

effects. We could also observe the individual p-value from the t-test of each fixed effect in the final model.

After finalizing the fixed effects of our model, we checked the validity of random effects. In the model summary, we could observe the proportion of variance of the response explained by the random effects. If the variance explained by the random term occupies a relatively high proportion among all variance, then that random effect shall have remained. Otherwise, we would remove the random term and the model shall be converted to a generalized linear model to reduce the complexity of the model. If the random term is removed, the resulting final model is the generalized linear model (GLM). Finally, we checked the assumptions of GLM.

## Results

### Model Comparison

All models have CSDuid as a random effect. Model 1 is intercept-only. Model 2 has sex as a fixed effect. Model 3 is with fixed factor age. model 4 has age and median income as fixed effects. Model 5 has age, median income, and population as fixed factors.

**Table 3:** A summary of p-values from likelihood ratio tests for research question 1

| Comparison | P-value | Preferred Model |
|---|---|---|
| Model 1 & 2 | 0.2197 | Model 1 |
| Model 1 & 3 | $7.187 * 10^{-9}$ | Model 3 |
| Model 3 & 4 | $3.829 * 10^{-12}$ | Model 4 |
| Model 4 & 5 | 0 | Model 5 |

The table conducted above reveals the results of 4 likelihood ratio tests. After rolling comparison, model 5 is the best fit model. The independent t-test corroborates the significance of fixed effects. Therefore, model 5 is the preferred model based on the results of the likelihood ratio test and p-values. However, the variance of the random effect in model 5 is 0.0277, which is quite small compared to the whole variance in the response; therefore, the random term was removed from the model, which alters the final model into a generalized linear model.

**Final model equation:**

$$log(\frac{\hat{p}}{1-\hat{p}}) = 18.678 + 0.005X_{age} - 1.597 * log(X_{income}) - 0.064 * logX_{population}$$

where p is the probability of customer who buy new products, 1-p is the probability of the traditional customer who does not buy new product;
$X_{age}$: represents the difference between maximum age and minimum age of customers (Note: Rescale age to (age-min(age))/max(age)-min(age));
$X_{income}$: represents the median income of subdivision of customers;
$X_{population}$: is the population size of the area that the customer lives in.

**Model Interpretation:**
$\hat{\beta}_0 = 18.678$: while all other predictors equals to zero, the intercept is the odds of customer buying new products, which is $1.293458 \times 10^8$ ($e^{18.678}$). However, interpretation of $\hat{\beta}_0$ is meaningless here, because population never takes a value of 0.

$\hat{\beta}_1 = 0.005$: with other predictors constant, for every one unit increase in age of the user, we expect the odd of customers who would buy new line products will be 1.005 times greater than the original odd ($e^{0.005} = 1.005$). Elder people are more likely to become our new customers, as the price of two lines products are more affordable for them.

$\hat{\beta}_2 = -1.597$: holding everything else fixed, for every one unit increase in the log of the median income of subdivision, we expect the odd of customers who would buy new line products will be 0.203 times greater than the original odd ($e^{-1.597} = 0.203$). Due to the more approachable price of products from "Active" and "Advanced" lines, customers with a less median income in the subdivisions are our new customers.

$\hat{\beta}_3 = -0.064$: for every one unit increase in the log of the population, with other predictors constant, we expect the odd of customers who would buy new line products will be 0.938 times greater than the original odd ($e^{-0.064} = 0.938$). New customers are more likely from the subdivisions with less population.

**Table 4:** A summary of statistics of fixed effects of the final model

| | Estimate | P-value | 95% CI | VIF |
|---|---|---|---|---|
| Intercept | 18.678 | $2 * 10^{-16}$ | (16.890, 20.473) | |
| Age | 0.005 | $6.22 * 10^{-9}$ | (0.003, 0.007) | 1.0001 |
| log(median income) | -1.597 | $2 * 10^{-16}$ | (-1.745, -1.449) | 1.0023 |
| log(population) | -0.064 | $2.09 * 10^{-5}$ | (-0.094, -0.035) | 1.0024 |

The table above summarizes the statistics for our final model. The p-values for three predictors are all less than 0.05, indicating the significance of age, median income, and population on

whether the customer buys the new products. The 95% CI of all significant terms does not contain 0, confirming their significance.

**Assumption checking**

First, the data shall be independently distributed by observing VIF values. From Table 4, we see that all predictors have a VIF well below 5. Thus, the independence assumption is satisfied. Second, our response to the new customer takes the value of 1, and old consumers are 0. The assumption of the binomial distribution is satisfied. Thirdly, our final model is a linear combination of predictors to a link function of the response; thus, the linearity assumption holds.

## To Investigate Why Mingar Devices Perform Poorly for Users with Darker Skin

### Exploratory Data Analysis

**Table 5:** Numerical summary of flag numbers under users with different skin color

| skin | Number of User | Mean | Variance | Flag/Duration |
|---|---|---|---|---|
| Dark | 2666 | 11.79 | 16.07 | 0.033 |
| Light | 3658 | 1.15 | 1.68 | 0.003 |
| Medium | 2941 | 3.65 | 4.98 | 0.010 |
| Medium-Dark | 2840 | 7.44 | 10.17 | 0.020 |
| Medium-Light | 3138 | 2.51 | 3.61 | 0.007 |

Table 5 corroborates the information indicated in Fig 2.1 below. The mean flag is 11.79 for dark-skin users, and 1.68 for light-skin users. Users with darker skin have more quality flags on their devices. In addition, since the mean and variance are roughly equal in each group, the flags follow a Poisson distribution. The column "Flag/Duration" is the mean of the ratio between flag numbers and duration for each skin color group. Such division is helpful in the later part.

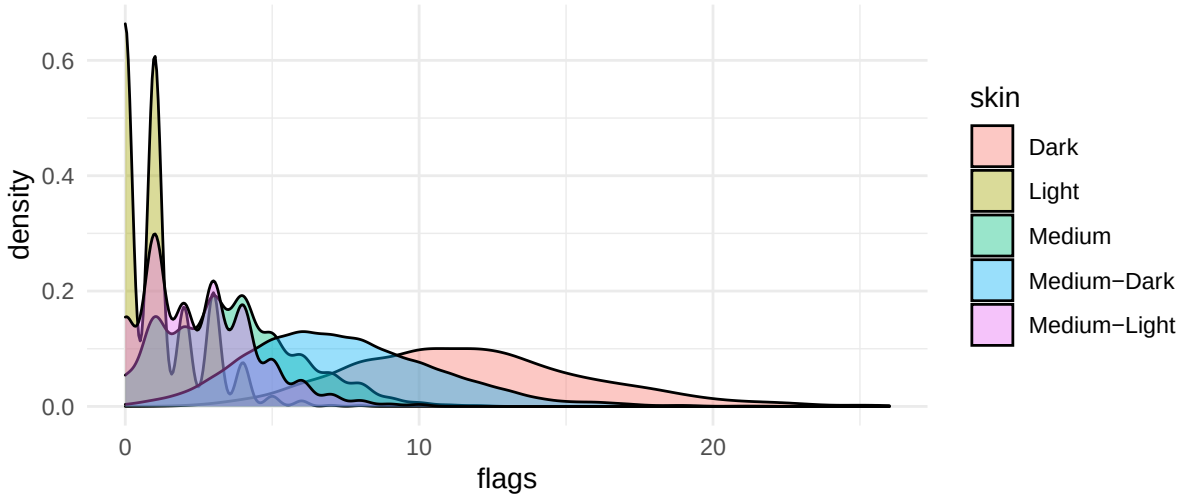


Figure 2.1: The distribution of number of quality flags

Figure 2.1 shows the distribution of flag numbers under users with different skin-colors. As displayed, dark skin and medium-dark skin users indeed have a greater number of flags than users with lighter skin. The charge that Mingar company facing is credible.

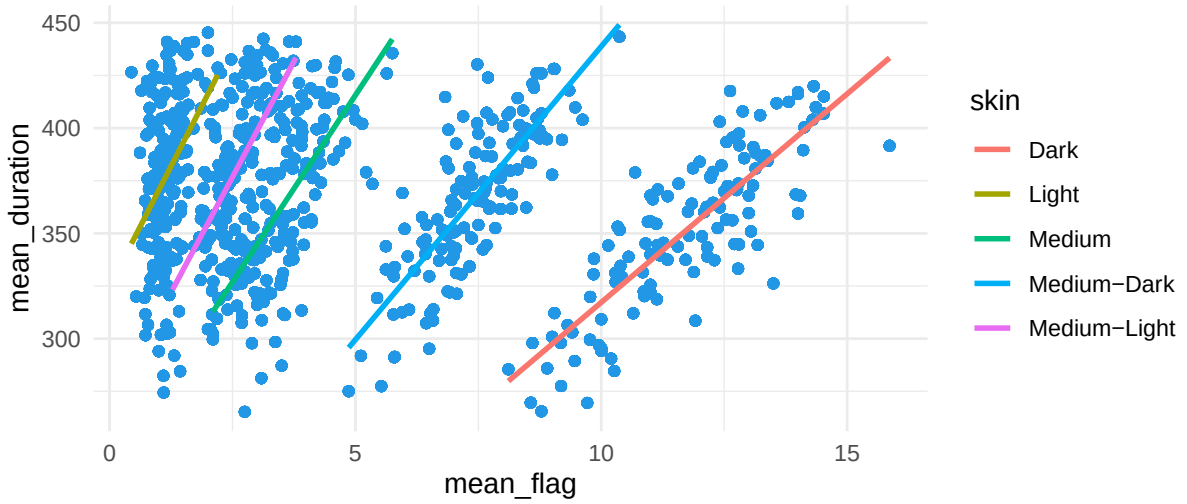


Figure 2.2: Mean flags and mean duration under different postal code

Figure 2.2 is a scatterplot between mean flag numbers and mean duration under each postal code, grouped by skin color. By the graph, quality flags and duration seem to have an obvious positive relationship in every skin color group. Additionally, Since each postcode has a different mean flag and duration, the feature of the postcode is potentially related to flag numbers.

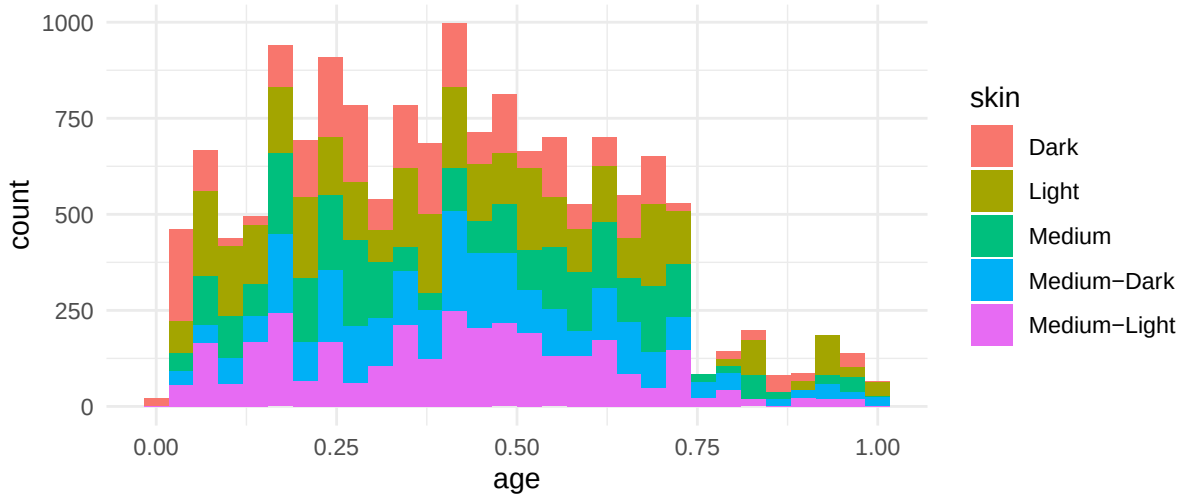


Figure 2.3: The distribution of flags stacked by age

Age is another potential factor that influences the flag numbers. Younger people might be less careful than elder individuals. In Fig 2.3, for ages ranging below 25, dark skin users occupy the largest numbers among all users, indicated by the relative height of the red bar. For the ages ranging above 60, lighter skin users occupy the biggest amount, shown by the olive green bar. Therefore, the plot raises an expectation that age might contribute more to the poorly performed sleeping quality of devices than skin color as criticized.

### Statistical Methods

We selected the number of quality flags of the sleeping session as our response. More flag numbers indicate worse device behavior and worse sleep score reliability. The response follows a Poisson distribution, so a generalized linear model or generalized linear mixed model is appropriate.
Potential predictors are selected based on results of exploratory data analysis and general sense. As displayed in Fig 2.3, age is a potential fixed effect. Duration could be an offset in our model. An offset accounts for different denominators in rates while allowing for counts to still be the response. Commonsensically, gender might affect the flag numbers. Men tend to have a better understanding of handling devices in the social norm. Their flag numbers might be less than women. The feature of the device that the user purchased might also influence the performance of the device. For example, the Advance products are all water-resistant, containing heart rate sensors but not pulse oximeters. Their functionality might influence the quality flags. The regional median income might also affect the flag numbers. Consequently, age, gender, line of device, and skin color are all potential fixed effects.

We started with the simplest model with no fixed effect and added in a potential predictor as a fixed effect one at a time. All models have flag counts as the response, duration as the offset, and

customer id as a random intercept. Since the models created are nested, we used the likelihood ratio test here. We assessed the best model by looking at the test's p-value in each comparison. At obtaining the final model, we checked the assumptions of the Poisson model.

### Result

**Model Comparison:**
All models have a random effects of customer id. Model 1 is an intercept-only model. Model 2 has a fixed effect of skin. Model 3 is with skin and age fixed factors. Model 4 has skin, age, and line of devices as fixed effects. Model 5 has skin, age, and sex as fixed predictors.

**Table 6:** A summary of p-values from likelihood ratio tests for research question 2

| Comparison | P-value | Preferred Model |
|---|---|---|
| model 1 & 2 | $2.2 * 10^{-16}$ | model 2 |
| model 2 & 3 | 0.007 | model 3 |
| model 3 & 4 | 0.327 | model 3 |
| model 3 & 5 | 0.374 | model 3 |

Table 6 summarizes the results of 4 likelihood ratio tests. A p-value less than 0.05 indicates that the full model is better. As shown, model 3 is the best after rolling comparisons with the best model after each comparison. Only fixed effects on skin color and age of users shall be chosen in the final model.

**Final model equation:**

$$log(\hat{\lambda}_x) = -3.38 - 2.39X_L - 1.21X_M - 0.50X_{MD} - 1.61X_{ML} - 0.05X_A + log(duration)$$

where $\lambda_x$ is the expectation of count of flags;
$X_L$ is 1 if the skin color of the user is light, 0 if not;
$X_M$ is 1 if the skin color of the user is medium, 0 if not;
$X_{MD}$ is 1 if the skin color of the user is medium-dark, 0 if not;
$X_{ML}$ is 1 if the skin color of the user is medium-light, 0 if not;
$X_A$ is the age of the user.

**Model Interpretation:**
$\hat{\beta}_0 = -3.38$: the average number of flags for customers with a minimum age 13.94 and dark skin

is $e^{-3.38} = 0.034$.

$\hat{\beta}_1 = -2.39$: holding everything else constant, the average number of flags for customers using light skin emoji is about $e^{-2.39} = 0.092$ times that for customer using dark skin emoji.

$\hat{\beta}_2 = -1.21$: holding everything else constant, the average number of flags for customers using medium skin emoji is about $e^{-1.21} = 0.298$ times that for customer using dark skin emoji.

$\hat{\beta}_3 = -0.50$: holding everything else constant, the average number of flags for customers using medium-dark skin emoji is about $e^{-0.5} = 0.607$ times that for customer using dark skin emoji.

$\hat{\beta}_4 = -1.61$: holding everything else constant, the average number of flags for customers using medium-light skin emoji is about $e^{-1.61} = 0.2$) times that for customer using dark skin emoji.

$\hat{\beta}_5 = -0.05$: the exponentiated coefficient $e^{-0.05}$ =0.95 represents that the number of flags changes by a factor of 0.95 between the minimum and maximum age. The estimate of average change with each additional year is $\frac{1-0.95}{90.87-13.94}$, 0.06%. We predict a 0.06% decrease in the average number of flags for each additional year older the customer is.

We recognize that the probability of obtaining a flag rises as the skin color of the user gets darker and darker. Younger people tend to have more quality flags. The majority of Dark skin users have a younger age than other skin-color users (shown in Fig 2.3).

The table below computes the estimate for each coefficient after exponentiating them. 95% confidence interval for each estimates is also exponentiated $\hat{\beta} \pm Z * SE(\hat{\beta})$. All of the confidence intervals above do not include 1, so we know that emoji skin color and age are significant predictors that should be included in the model.

**Table 7:** Results of final model of flag counts

| Factor | Estimate | 95% CI |
|---|---|---|
| Baseline | 0.034 | (0.033,0.035) |
| Light | 0.092 | (0.089,0.095) |
| Medium | 0.298 | (0.290,0.305) |
| Medium-Dark | 0.607 | (0.594,0.620) |
| Medium-Light | 0.199 | (0.194,0.205) |
| Age | 0.949 | (0.914,0.986) |

**Assumption Checking**

To make interpretation meaningful from the model, the assumptions of a Poisson model shall

be satisfied. First, our response is flag counts, which are Poisson-distributed. Secondly, the observations must be independent of one another. Our dataset doesn't satisfy independent observations, because the total observations in the dataset are 15243 while there are only 719 customer ids and 10 device ids. Adding random effects (customer id or device id) to the model addressed the issue. Thirdly, the mean of response must be equal to its variance. As shown in Table 5, this assumption holds for each level of skin color group. Finally, we ensured the linearity assumptions hold by making $log\lambda$ a linear function of fixed effects and random effects.

## Discussion

**Conclusion:**

To analyze the main features of Mingar's new customers who buy the "Active" and "Advance" lines' new products, we selected potential predictors based on the data visualization results and real-life rationales. Age, median income level, and the population of subdivisions of customers play significant roles in predicting new customer groups. Mingar's new products attract the elder age group who focus on healthier lifestyles more than young customers, thus Mingar could consider elderly-focus products to further aggrandize this advantage. To target elderly customers, Mingar could add a bigger font size and a safety feature, and make their product easy to keep. Moreover, the group with less income is more likely to purchase the two new lines. To capture more market share, Mingar should consider continuously making the products more affordable to attract more new customers and better match the price offered by Bitfit. Lastly, the new customers tend to come from small population subdivisions; hence, various advertisements and community marketing strategies could be considered.

Using summary statistics and distribution graphs, we found that the average flag count rate for dark-skin users is significantly higher than for light-skin users. We also found that flag count rates tend to decrease as customer age increases, especially for dark-skin users by plotting the two variables. Using a likelihood ratio test, we have fitted the most appropriate model to model the flag count rate. Our results indicate that skin color is a significant predictor of flag count rate. The flag count rate increase as skin color goes from light to dark. Thus, the remark that Mingar device is performing poorly for dark skin users is valid. Consistent with our explanatory data analysis, we found that age is also a significant predictor of flag count rate and they are negatively associated. Among young users, dark skin users occupy the largest proportion. Perhaps, due to some characteristics of the device, the product is more attractive to younger black users and young users are using the device in a way that causes quality issues. Therefore, to address the racist criticism, Mingar company shall detailedly inform their young customers on how to handle the device and examine the reason behind each flag (missing data, quality issues etc).

**Limitations:**

One postal code may have multiple CSDuids. Furthermore, the median income used is the subdivision's median income, instead of the individual's income. Therefore, our demographic information may contain inaccurate information about our customers and result in biased estimates. To better understand customers, Mingar can consider collecting demographic information on each customer. However, this action should be carefully taken as it may raise privacy concerns.

Missing values in the emoji are the default yellow emoji and they are removed. However, users who use default emoji must possess a specific skin color. Moreover, even if the user used a specific color emoji, this does not imply the user has that skin color. Thus, the variable itself also decreases the size of the effective sample size. We could prevent such a problem if the client provides straightforward variables about ethnicity and race or any other proxy variables along with the emoji modifier.

For the second model, there is a trade-off of adding a random intercept effect. Initially, the second independence assumption was not satisfied. After adding a random intercept effect with customer id, the model fulfills the independence request. However, the random effect only explains the 0.0021 of variation in response, which indicates the random effect shall not be added. In future analysis, any client shall consider providing a bigger sample size to prevent such at full steam.

Lastly, flag counts are the number of quality flags during sleep. 1949 observations have 0 flag counts. Some customers do not wear the device when they sleep, and the flag counts of these customers are 0. We are not able to tell if a 0 flag count indicates no quality issue or that the customer is not wearing the device, which potentially underestimates the true quality issue. To address the problem, Mingar can find ways to distinguish the customers who wear the device during sleeping and only study these customers.

# Consultant information

## Consultant profiles

**Lucia Tan**. Lucia is a senior consultant with 4.0. She specializes in statistical communication and she is proficient in R, SQL and SAS. Lucia earned her Bachelor of Science, Specialist in Statistics Methods and Practice, from the University of Toronto in 2025.

**Jinping Liang**. Jinping is a junior consultant with Four Point Zero, LLC Analytics. She specialize in reproducible analysis and statistical communication. Jinping earned her Bachelor of Science, Specialist in Statistical Science (Methods and Practice) and Major in Economics from the University of Toronto in 2024.

**Tong Su**. Tong is a junior data analytics with FOUR POINT ZERO. She specializes in machine learning and data visualization. Tong earned her Bachelor of Science, Specialist in Computer Science and Majoring in Statistics from the University of Toronto in 2024.

**Ruojin Lin**. Ruojin is a junior data analytics with FOUR POINT ZERO. She specializes in statistical models and time series analysis. Ruojin earned her Bachelor of Science, major in Statistics & Mathematics & Economics, from the University of Toronto in 2023.

## Code of ethical conduct

Responsibility to Employers and Clients

- We have carried out the work according to the specification of clients, document every step of statistical analysis, and presents the steps honestly to clients.
- We ensured that the result of the analysis will not be misinterpreted by clients by clearly stating any limitations of assumptions for the result.
- Every member has signed the statements from the Non-Disclosure and Confidentially Agreement to protect confidential information.
- We promise we will not sell or share the client data with any third parties.

Professionalism

- We promise we only take work that falls within our limit of statistical competencies.
- We actively seek opportunities to upgrade professional knowledge and skills.

Responsibility to Other Statistical Practitioners - We fully acknowledge the source of other statisticians. - We provide support to fellow statisticians by encouraging new entrants to the field, giving constructive advice, and directing criticism to procedures rather than persons.

## References


- Achim Zeileis, Torsten Hothorn (2002). Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. URL https://CRAN.R-project.org/doc/Rnews/
- Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra
- Claus O. Wilke (2020). cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2". R package version 1.1.1. https://CRAN.R-project.org/package=cowplot
- Dmytro Perepolkin (2019). polite: Be Nice on the Web. R package version 0.1.1. https://CRAN.R-project.org/package=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 (2021). rvest: Easily Harvest (Scrape) Web Pages. R package version 1.0.2. https://CRAN.R-project.org/package=rvest
- Hadley Wickham and Dana Seidel (2020). scales: Scale Functions for Visualization. R package version 1.1.1. https://CRAN.R-project.org/package=scales
- Hao Zhu (2021). kableExtra: Construct Complex Table with "kable" and Pipe Syntax. R package version 1.3.4. https://CRAN.R-project.org/package=kableExtra
- John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/
- Postal code conversion file. University of Toronto Libraries. https://mdl.library.utoronto.ca/collections/numeric-data/census-canada/postal-code-conversion-file
- R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Statistics Canada, postcode, 3/31/22. Reproduced and distributed on an education basis with the permission of Statistics Canada.
- Unicode. Full Emoji Modifier Sequences, v14.0. https://unicode.org/emoji/charts/full-emoji-modifiers.html
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

# Appendix

### Web scraping industry data on fitness tracker devices

To access industry data on fitness tracker devices, we web scraped a table using the rvest and polite package. Before web scraping, we checked that the website does not have an API. We also checked that no user guideline specifies any restrictions for web scraping. Next, we pass in the URL we want to scrape and introduce ourselves to the host by updating the user-agent string with one of our consultant's email addresses (tong.su@mail.utoronto.ca) to provide our contact information. This is so that the host can contact us in case there are any questions or concerns. After we pass in the information, we see that the crawl delay is 12 seconds so we will adhere to this limit. We also see that there are 2 rules defined for 2 bots and the path is indeed scrapable for us. Then We scraped only the data we need and used it to create new data.

### Accessing Census data on median household income

Median household income data were obtained from the Canadian Census API. We have obtained the API by registering the account with one of our consultant's email addresses (tong.su@mail.utoronto.ca) to provide our contact information in case there are any questions or concerns. We used the cancensus package and the API key we get from the census mapper. We only want to take the data we need, so we only requested data from the 2016 census and selected the three variables needed, which are postal code, median income, and population. The data is governed by the Statistics Canada Open Data Licence and we made sure to adhere to the license grant. It has gone through some processing and may differ from the original data.

### Accessing postcode conversion files

We sourced through the UofT library by logging in with one of our consultant's email addresses (tong.su@mail.utoronto.ca), in which we have access to a Census Canada Postal Code Conversion Files. After accepting a license agreement, we chose the most recent census dataset from August 2016. The data downloaded are in ".sav" format. We uploaded the compressed form of the downloaded dataset into R and read it. When downloading, we only select two variables we needed: postcodes (PC) and CSDuid.