# Investigation of characteristics of new customers and devices performance

New customers are mainly from lower median household income neighbourhoods; darker skins report more flags per minute with respect to sleep scores

Report prepared for MINGAR by datanerds

2022-04-11

# Contents

# Executive summary

## Background

Mingar was initially making GPS devices for maritime vehicles and military personnel, then expanded the business into personal GPS devices for runners in the early 2000s. Over the years, they have developed a range of products for outdoor recreation, including high-end fitness tracking wearable devices. The main competitor of Mingar in the area of fitness trackers is Bitfit, which is a company focused on personal fitness tracking devices. The price point of their products is lower than Mingar. As wearable devices are a growing market, to gain more market share in low and medium-end markets, Mingar expanded their products by adding the 'Active' and 'Advance' lines at a more affordable price for the average customers. As the consultant company of Mingar, we will investigate the characteristics of new customers of Mingar and how the new customers who are the buyers of 'Active' and 'Advance' products are different from the traditional customers. Moreover, we would investigate the trend complaints that some devices are not performing well in terms of sleep scores for darker skin users.

## Key findings

- New customers are mainly younger than 30 or older than 60 years old, whereas traditional customers are mainly between 30 and 60 years old.
- Customers who purchased the 'Active' and 'Advanced' lines are more than the previous products.

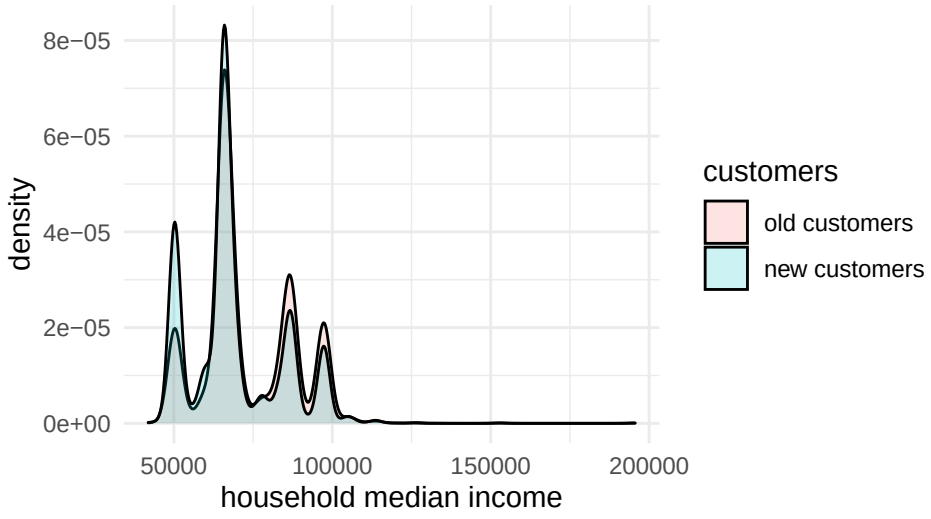


**Figure 1:** Density plot of customers household median income

From figure 1, new customers are mainly from lower median household income neighbourhoods, whereas the traditional customers mainly constitute individuals from wealthy communities.

- 'iDOL' products tend to malfunction with low frequency, while there are more 'Active', 'Advance',

and 'Run' than 'iDOL' among the products that malfunction with higher frequency.

- The average flags per minute of dark skin is around 5 times of default yellow skin; is around 11 times of light skin; is around 3 times of medium skin; is around 1.6 times of medium-dark skin; is around 5 times of medium-light skin.

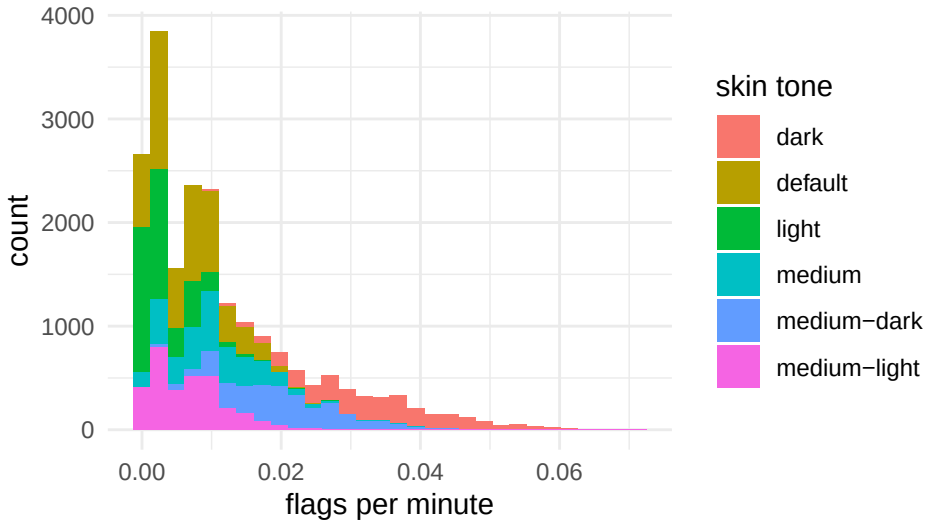


**Figure 2:** Histogram of flags per minute in terms of skin tone

From figure 2, darker skins report more flags per minute with respect to sleep scores. The product design may not take into account the impact of skin color on sensor accuracy.

## Limitations

- The customer's skin tone was inferred by the emoji modifier, which may not accurately describe the actual skin tone of customers as some customers may have different skin tones, but still use the default yellow emoji modifier.
- The household median income is the median household income of the community rather than the customer's household income. Thus, it may lead to inaccurate results based on the household median income.

# Technical report

## Introduction

We are Datanerds, the consultant company of Mingar. Mingar is an outdoor recreation company specializing in GPS production, and Datanerds have taken on the task of doing data analytics for them, specifically for their wearable products and their consumers. Mingar wants insight into the consumers and their new products to strategize for the Canadian market. Therefore, one research question of the task is: "how do the customers of the new products, namely Active and Advance, differ from the customers of the old products?" Since Mingar has recently received some controversies regarding the different product performances on various skin tones, we also need to address another question: "how is the wearer's skin tone associated with performance errors of products?"
To address the two matters, we get the data by web scraping industry data on fitness tracker devices, accessing Census data on median household income and accessing postcode conversion files. We merge the raw dataset into larger datasets for the main report.
Then we will show some exploratory data analysis on the data we collected by doing numerical and graphical summaries through tables, histograms, and density plots. In the following steps, two generalized linear mixed models (GLMM) will be applied; one will be used to solve the first research question regarding differences in customers of old and new products, and the other one will be used to solve concerns on performance errors on various skin tones. Lastly, a conclusion will be reached and reported to Mingar company.

### Research questions

- What is the characteristic of new customers in the Canadian market, and how are new customers different from the traditional customers.
- How do the devices perform in terms of sleep scores on different skin colours.

## Research question 1: What is the characteristic of new customers in the Canadian market and how new customers different from the traditional customers

## Exploratory data analysis

### Numerical summaries

**Table 1:** Count and proportion of customers

| customers | count | proportion |
|---|---|---|
| new customers | 10569 | 0.5549488 |
| old customers | 8476 | 0.4450512 |

As table 1 shows above, new customers count for 10569, and old customers count for 8476. The new customer count is more than 20% of that of the old customer. Table 1 also contains the proportion of each type of customer, new customers count for nearly 55.5% of all customers, and old customers count for 44.5%; it implies that the people who purchased 'Active' and 'Advanced' lines are more than the previous products. Since we want to investigate the characteristics of new customers, having a larger proportion provides us with a larger sample population.

**Table 2:** Numerical summaries of age and household median income

| Variable | Minimum | 1st Quartile | Median | Mean | Standard Deviation | 3rd Quartile | Maximum |
|---|---|---|---|---|---|---|---|
| Age | 18 | 34 | 47 | 47.31 | 16.89 | 60 | 92 |
| Household Median Income | 41880 | 65327 | 65829 | 70752 | 14782.91 | 85981 | 195570 |

Table 2 is the numerical summary of the variables Age and Household Median Income. The minimum age of the customer is 18 years old, and the maximum age of the customer is 92 years old. The median age of customers is 47 years old, and the mean age is 47.31 years old. The mean and median age of customers is very similar. The standard deviation of the ages is 16.89, which means the fluctuation of the age of customers is large. The first and third quartile ages are 34 and 60, respectively. It means that the IQR of the age of customers is 26.
For Household Median Income, it has the minimum of $41880, the wealthiest neighbourhood that customers have a median income of $195570 The mean of this median is $70752 while the standard deviation is 14782.91, which means the fluctuation of household median income among customers is large. The neighbourhood in the first quartile has a median income of $65327, whereas the one in the third quartile earns $85981 at the median. The median of median income is $65829. Since the mean is much greater than the median and the maximum is exceptionally high, the distribution of household median income should be right-skewed.

### Graphical summaries

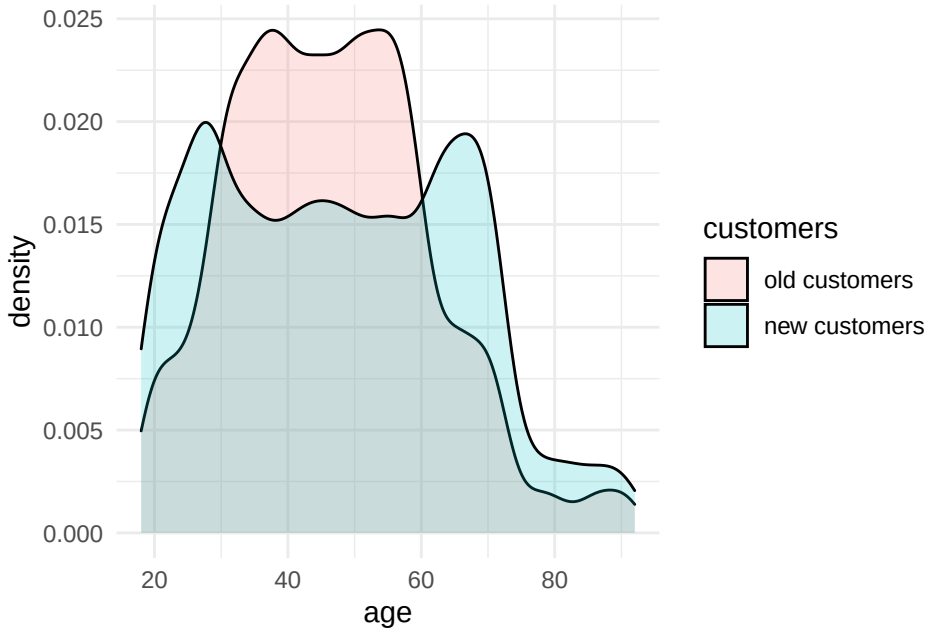


**Figure 3:** Density plot of customers' age

From figure 3, the density plot shows the distribution of age by new and old customers. Both distributions are slightly right-skewed. The distribution of customers who buy old products is unimodal, whereas customers who buy new products are bimodal. For customers younger than 30, there are more 'Active' or 'Advance' products buyers than old products buyers. Old products are more popular for customers between 30 and 60 years old as the density is much higher. For customers older than 60, more of them buy the new products than those who buy the old ones. Besides, the spread of the age of new customers is more extensive than old customers. It implies that our products have become more accessible to a broader range of people of different ages.

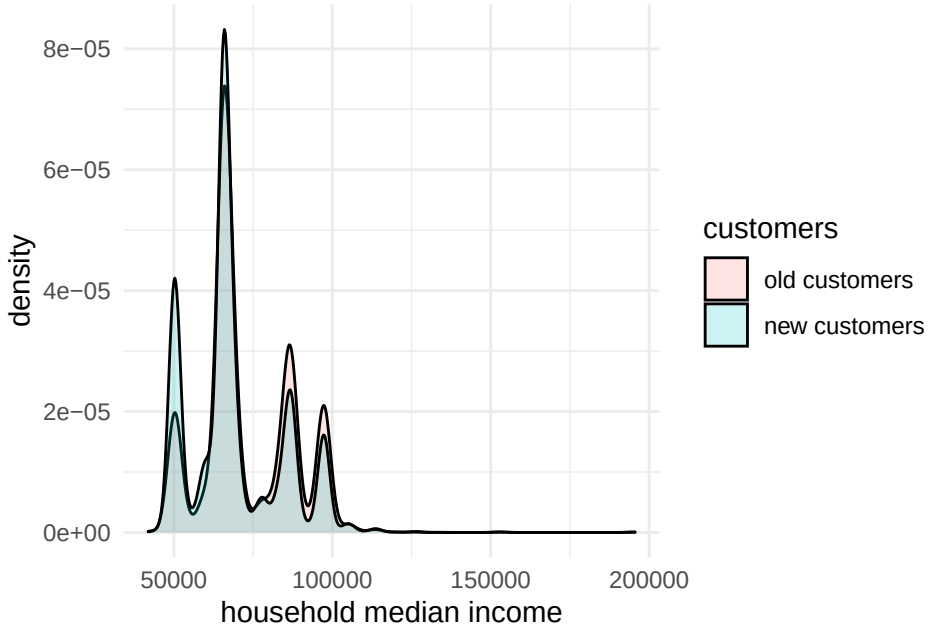


**Figure 4:** Density plot of customers' household median income

From figure 4, the density plot above shows the distributions of the household median income of the neighbourhood that the customers are from by the type of products they buy. The right tail is very long, so some customers are richer than the rest. The two distributions are multimodal; there are four peaks each. The first peak is around $50000, and more customers buy the new products than those who buy the old products. The second peak is around $70000; there are also slightly more new customers than the old customers. However, as household median income increases, the third peak, around $85000 shows more customers buy the old products than those who buy the new Active or Advance products. At the last peak, which is a little less than $100000, there are also more old customers than new customers. Therefore, buyers from poorer neighbourhoods tend to buy new products: 'Active' or 'Advance'.

## Method

**Data description and wrangling:**

**Table 3:** Variable description for research question 1

| Variable | Description |
|---|---|
| CSDuid | Census subdivision unique identifier |
| age | Customers' age until 2022 |
| household median income | Household median income of the region by postcode area |

| Variable | Description |
|---|---|
| new customer | the buyers of 'Active' and 'Advance' products |

Table 3 shows the variables and their description used for the Generalized Linear Mixed Model to investigate the factors that may affect the odds of being a new customer. In addition, to make the age equal to 0 is interpretable, we rescale the age by subtracting all ages by the minimum age in the dataset(18 years old). Also, to make the household median income equal to 0 is interpretable, we resale the household median income by subtracting all household median income by the minimum household median income in the dataset($41880). The new customers are defined as the buyers of 'Active' and 'Advance' products. Thus, we chose the customer information from the 'Active' and 'Advance' lines and defined them as 1 (new customer). The other lines are defined as 0 (old customer). The missing values were removed from the dataset.
The total sample size of the customer dataset used in this model has 19045 observations with 11 variables. The customer dataset is a merged dataset from web scrapping and API. The response variable is *new customer*, and the predictor variables are *age*, and *household median income*.

**The Generalized Linear Mixed Model:**

$$Y_{ij}|U \sim Bernoulli(\pi_i)$$

$$log(\frac{\pi_i}{1-\pi_i}) = X\beta + U_i$$

$$U_i \sim N(0, \sigma^2)$$

Where:

- $Y_{ij}$ represents new or old customer of 'Active' and 'Advance' line in postcode area i at measurement j.
- $\pi_i$ represents the probability of the new customer.
- $X$ represents the fixed effect of the model.
- $U_i$ is the individual-level random effect.
- $U_i > 0$ if i is more likely than the average to have a new customer.

The purpose of the Generalized Linear Mixed Model is to investigate the factors that may affect the odds of being a new customer.

**The assumptions for Generalized Linear Mixed Model:**

- Random effects come from a normal distribution.
- The random effects errors and within-unit residual errors have constant variance.

## Result

**Final model:**
*Model 1:*

$$log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 \text{ age} + \beta_2 \text{ household median income} + U_i$$

Where:

- $U_i$ is the individual-level random effect, CSDuid is the random effect in these models as CSDuid is a grouping unit.

Since we wonder whether age or household median income would affect the odds of being a new customer, thus, age and household median income are the fixed effect.

**Model selection:**
Assume the significant level is 0.05. Since the p-value of likelihood ratio test is 0.6700434, which is larger than 0.05. Thus, we prefer the simpler model 1 presented above as the final model for research question 1.

**Table 4:** Estimate and Confidence interval of exponential coefficients of model 1

| | Estimate | 95% CI |
|---|---|---|
| intercept | 0.6598200 | (0.5278393, 0.7918008) |
| age | 0.0050605 | (0.0033380, 0.0067831) |
| household median income | -0.0000170 | (-0.0000208, -0.0000132) |

From table 4, we can get our **Estimated Generalized Linear Mixed Model 1:**

$$log(\frac{\hat{\pi_i}}{1-\hat{\pi_i}}) = 0.65982 + 0.005061X_{\text{age}} - 0.000017X_{\text{household median income}}$$

Where:

- $\hat{\pi_i}$ represents the estimated probability the new customer.
- For $\hat{\beta_0} = 0.65982$, that is the intercept of the model. It means when age, and household median income equal 0, the expected log odds of the new customer is 0.65982. In this case, when age is 0 represents the age of 18, since we resale the age by subtracting all age by the minimum age in the dataset(18 years old). When household median income is 0, represents the household median income is 41880, since we resale the household median income by subtracting all household median income by the minimum household median income in the dataset($41880). Also, we are 95% confident that when age, and household median income equal 0, the population mean log odds of the new customer is between 0.5278393 and 0.7918008 from table 4.
- For $\hat{\beta_1} = 0.005061$, for every one-unit increase in age, the average log odds of the new customer will increase by 0.005061 when holding other predictor constant. Also, from table 4, we are 95% confident that for every one-unit increase in age, the population mean log odds of the new customer will increase between 0.0033380 and 0.0067831, when holding other predictor constant.

- For $\hat{\beta}_2 = -0.000017$, for every one-unit increase in household median income, the average log odds of the new customer will decrease by 0.000017, when holding other predictor constant. Also, from table 4, we are 95% confident that for every one-unit increase in household median income, the population mean log odds of the new customer will decrease between 0.0000208 and 0.0000132, when holding other predictor constant.

**Model assumption check:**

- The residuals versus fitted values plot of the model 1 shows two levels of data, indicated by the two lines of dots. Both levels have a downward trend, which suggests that the random effects and residual errors do not have constant variance.
- The dots of the normal Q-Q plot of model 1 are not spread along the diagonal line, so the normality assumption is violated.

#### Conclusion for Research question 1

The results suggest that the new customers are mainly younger than 30 or older than 60. Also, the newer and more affordable 'Active' and 'Advance' products attract more buyers with lower median household incomes. More specifically, for every one-unit increase in age, the average log odds of the new customer will increase by 0.005061; for every one-unit increase in household median income, the average log odds of the new customer will decrease by 0.000017. New customers have a wider range of age than the old customers. New customers are mainly younger than 30 or older than 60 years old, whereas old customers are mainly between 30 and 60 years old. The new customers are mainly from lower median household income neighbourhoods, whereas the old customers are primarily from wealthier communities. We notice that there are more new customers than old customers, our new products are more popular with females no matter in old or new customers.
In summary, younger or older customers are the main source of new customers because these groups are more price-sensitive. Lower prices are more likely to attract these customers. Therefore, applying the price strategy, it is easy to attract young people under the age of 30 and the elderly over the age of 60. In this way, more new customers and markets can be obtained.

## Research question 2: How the devices perform in terms of sleep scores on different color of skin

## Exploratory data analysis

#### Numerical summaries

**Table 5:** Numerical summaries of flags and flags per minute

| Variables | Mean | Median | Min | Max | Variance |
|---|---|---|---|---|---|
| Flags | 4.32 | 3.00 | 0.00 | 26.00 | 17.87258 |

| Variables | Mean | Median | Min | Max | Variance |
|---|---|---|---|---|---|
| Flags per minute | 0.011825 | 0.00787 | 0.00 | 0.07123 | 0.00014 |

Table 5 is the numerical summaries of flags and flags per minute. The sample mean of flags is 4.32, and the sample median is 3.00, lower than its sample mean. The sample mean of flags per minute is 0.011825, and the sample median is 0.00787, lower than its sample mean. The range of flags is from 0 to 26, and the range of flags per minute is from 0 to 0.07123, which is smaller than the range of flags. The sample variance of flags is 17.87258. The sample variance of flags per minute is at 0.00014.
The sample variance of flags is much larger than the sample mean of flags, which implies that there may exist overdispersion.

### Graphical summaries

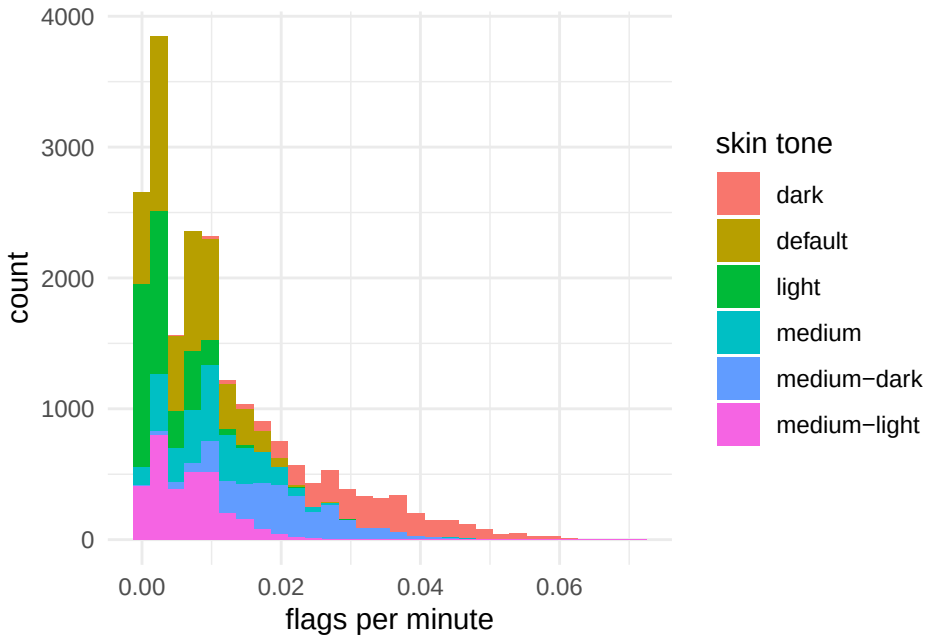


**Figure 5:** Histogram of flags per minute in terms of skin tone

Figure 5 shows the distributions of the number of flags per minute of the devices on different skin tones. All distributions are highly right-skewed. While some people use the default, others still use specific kinds of skin tones. As shown in green, most light-skin-toned users report flags less than 0.02 per minute. Medium-light skin tone (pink) customers report even fewer cases of flags per minute than light-skinned customers. However, as the skin tone gets darker, the distribution's right tail gets longer as well. For medium skin-toned users (light blue), many report flags per minute between 0 and 0.02, while others report flags per minute higher than 0.02. Moving onto medium-dark skin-toned users (dark blue), they report fewer cases between 0 and 0.02 flags per minute than the medium skin-toned ones, but many of

them report more than 0.02 flags per minute. Some say they have more than 0.04 flags per minute. When the skin tone is dark (red), which is the darkest of all, the tail of the distribution is the longest of all. Among those who experience 0.02 to 0.04 flags per minute, there are darker skin-toned than light, medium-light, medium, or medium-dark skin-toned. Most of those who experience more than 0.04 flags per minute have dark skin tone. There are even some extreme cases beyond 0.06, which are also dark-skin-toned. Generally speaking, darker skins seem to report more flags per minute.

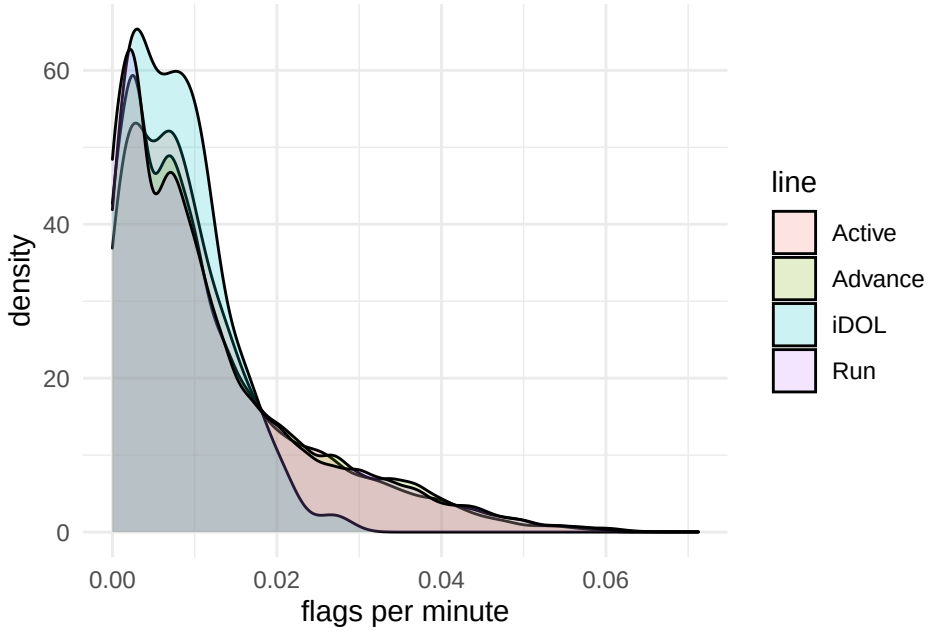


**Figure 6:** Density plot of flags per duration in terms of product line

Figure 6 the distributions of flags per minute of four products lines. All four distributions have long right tails, so they are right-skewed. 'Active' and 'Advance' are new products, whereas 'iDOL' and 'Run' are older ones. Among the products that do not cause malfunction (0 flag per minute), there are more 'Run' products than 'Active', 'Advance' or 'iDOL.' Between 0 and 0.017 flags per minute, more errors are coming from the 'iDOL' line than the rest. However, after 0.017 flags per minute, there are not as many malfunctioning 'iDOL' lines. There are barely any 'iDOL' products with flags more than 0.03 per minute. The other three lines, though, still send out more than 0.03 per minute errors. Some even average more than 0.07 flags per minute. Generally speaking, 'iDOL' products tend to malfunction with low frequency, while among the products that malfunction with higher frequency, there are more 'Active', 'Advance', and 'Run' than 'iDOL'.

## Method

**Data description and wrangling:**

**Table 6:** Variable description for research question 2

| Variable | Description |
|---|---|
| cust id | Unique ID for each customer |
| duration | Duration, in minutes, of sleep session. |
| skin tone | Infer the skin tone of customers by emoji modifier |
| flags | Number of times there was a quality flag during the sleep session |

Table 6 shows the variables and their description used for the Generalized Linear Mixed Model to investigate the factors that may affect the number of times there was a quality flag during the sleep session. In addition, we created the skin tone in the data-prep file by using the emoji modifier. The emoji modifier was interpreted by the standard Unicode modifiers based on the Fitzpatrick scale. Since the flags are the number of times, there was a quality flag during the sleep session. Flags may occur due to missing data or data being recorded but sufficiently unusual to suggest it may be a sensor error or other data quality issue. We also have the duration variable, the sleep session in minutes. For example, to mitigate the effect of duration, a longer duration may have more flags than shorter duration. Thus, we create a new variable called adjusted flags by rescaling the flags by dividing by the duration. The missing values were removed from the dataset.
The total sample size of the customer sleep dataset used in this model has 20412 observations with 43 variables. The customer dataset is a merged dataset from web scrapping and API. The response variable is *flags*, and the predictor variable is *skin tone*.

**The Generalized Linear Mixed Model:**

$$Y_{ij}|U \sim Poisson(\mu_i)$$

$$log(\mu) = log(duration) + X\beta + U_i$$

$$U_i \sim N(0, \sigma^2)$$

Where:

- $Y_{ij}$ represents number of times there was a quality flag during the sleep session for customer i at sleep session j.
- $\mu$ represents the flags per duration unit.
- $X$ represents the fixed effect.
- $log(duration)$: the offset term. The coefficient estimate for an offset term is fixed to be 1.
- $U_i$ is the individual-level random effect.

The Generalized Linear Mixed Model aims to investigate the factors that may affect the number of times there was a quality flag during the sleep session. Moreover, the duration of the sleep session is also relative to the flags. For example, a longer duration may have more flags than shorter duration. Thus duration is the offest term for this model.

**The assumptions for Poisson Generalized Linear Mixed Model:**

- Random effects come from a normal distribution.
- The random effects errors and within-unit residual errors have constant variance.

- The mean and variance of the response are equal.

## Result

**Final model:**
*Model 2*:
$log(\hat{\mu}) = log(duration) + \beta_0$
$+ \beta_1 * I(skin\,tone = default)$
$+ \beta_2 * I(skin\,tone = light)$
$+ \beta_3 * I(skin\,tone = medium)$
$+ \beta_4 * I(skin\,tone = medium\,dark)$
$+ \beta_5 * I(skin\,tone = medium\,light)$
$+ U_i$

Where:

- $U_i$ is the individual-level random effect, cust id is the random effect in these models as cust id is a grouping unit.

Since we wonder how skin tone would affect the flags per duration unit, thus, skin tone are the fixed effect.

**Model selection:**

Assume the significant level is 0.05. The p-value of likelihood ratio test is 0.2640583, which is larger than 0.05. Thus, we prefer the simpler model 2 as the final model for research question 2.

**Table 7:** Estimate of coefficients of model 2

| | Estimate |
|---|---|
| Intercept | -3.4000276 |
| default skin tone | -1.6327651 |
| light skin tone | -2.3912784 |
| medium skin tone | -1.2137395 |
| medium dark skin tone | -0.5018046 |
| medium light skin tone | -1.6148868 |

**Estimated Generalized Linear Mixed Model 2:**
From table 7, we found the estimate coefficients for the final model:

$$log(\hat{\mu}) = -3.4000276 - 1.6327651 * I(skin\,tone = default)$$
$$- 2.3912784 * I(skin\,tone = light)$$
$$- 1.2137395 * I(skin\,tone = medium)$$
$$- 0.5018046 * I(skin\,tone = medium\,dark)$$
$$- 1.6148868 * I(skin\,tone = medium\,light)$$

**Table 8:** Estimate and Confidence interval of exponential coefficients of model 2

| | Estimate | 95% CI |
|---|---|---|
| Intercept | 0.0333723 | (0.0328969, 0.0338547) |
| default skin tone | 0.1953886 | (0.1908650, 0.2000193) |
| light skin tone | 0.0915126 | (0.0884270, 0.0947059) |
| medium skin tone | 0.2970843 | (0.2896749, 0.3046831) |
| medium dark skin tone | 0.6054371 | (0.5925232, 0.6186325) |
| medium light skin tone | 0.1989132 | (0.1934875, 0.2044910) |

To better interpret the model, we have the estimated exponential coefficients from table 8.
$$\hat{\mu} = 0.0333723 + 0.1953886 * I(skin\,tone = default)$$
$$+ 0.0915126 * I(skin\,tone = light)$$
$$+ 0.2970843 * I(skin\,tone = medium)$$
$$+ 0.6054371 * I(skin\,tone = medium\,dark)$$
$$+ 0.1989132 * I(skin\,tone = medium\,light)$$

Where:

- $\hat{\mu}$ represents the estimated average flags per duration unit.
- $e^{\hat{\beta}_0} = 0.0333723$ is the intercept of the model. It means for dark skin, the expected flags per duration unit is 0.0333723. Also, we are 95% confident that the population mean flags per duration unit for dark skin is between 0.0328969 and 0.0338547 from table 8.
- $e^{\hat{\beta}_1} = 0.1953886$ represents the relative rate between default and dark skin is 0.1953886, that is the expected flags per duration unit of dark skin is around 5 times of default skin. Thus, the expected flags per duration unit for default skin is around 0.006. Also, we are 95% confident that the population mean relative rate between default and dark skin is is between 0.1908650 and 0.2000193 from table 8.
- $e^{\hat{\beta}_2} = 0.0915126$ represents the relative rate between light and dark skin is 0.0915126, that is the expected flags per duration unit of dark skin is around 11 times of light skin. Thus, the expected flags per duration unit for light skin is around 0.0027. Also, we are 95% confident that the population mean relative rate between light and dark skin is is between 0.0884270 and 0.0947059 from table 8.
- $e^{\hat{\beta}_3} = 0.2970843$ represents the relative rate between medium and dark skin is 0.2970843, that is the expected flags per duration unit of dark skin is around 3 times of medium skin. Thus, the

expected flags per duration unit for medium skin is around 0.009. Also, we are 95% confident that the population mean relative rate between medium and dark skin is is between 0.2896749 and 0.3046831 from table 8.

- $e^{\hat{\beta}_4} = 0.6054371$ represents the relative rate between medium-dark and dark skin is 0.6054371 , that is the expected flags per duration unit of dark skin is around 1.6 times of medium-dark skin. Thus, the expected flags per duration unit for medium-dark skin is around 0.0183. Also, we are 95% confident that the population mean relative rate between medium-dark and dark skin is is between 0.5925232 and 0.6186325 from table 8.
- $e^{\hat{\beta}_5} = 0.1989132$ represents the relative rate between medium-light and dark skin is 0.1989132, that is the expected flags per duration unit of dark skin is around 5 times of medium-light skin. Thus, the expected flags per duration unit for medium-light skin is around 0.006. Also, we are 95% confident that the population mean relative rate between medium-light and dark skin is is between 0.1934875 and 0.2044910 from table 8.

**Model assumption check:**

- The residual versus fitted plot of the model 2 have three clusters on the plot, so the observations are not independent. However, since independency is not an assumption for GLMM, nothing has been violated yet. Because the dots do not have discernible pattern, and the distance from the top to the bottom of the trend is almost constant throughout the observations, the homoscedasticity assumption is reasonably satisfied.
- The dots of the normal Q-Q plot are almost spread along the diagonal line, the normality assumption of random effects of model 2 is reasonably satisfied.
- From the numerical summaries of the response variable *flags*, the sample variance is greater than the sample mean; thus, we have overdispersion assumption is violated. The standard errors might be falsely small, meaning we'll probably have falsely small p-values.

### Conclusion for Research question 2

The results demonstrate that darker skins report more flags per minute with respect to sleep scores. In addition, 'iDOL' products tend to malfunction with low frequency, while among the products that malfunction with higher frequency, there are more 'Active', 'Advance', and 'Run' than 'iDOL'. Moreover, we notice that the expected flags per minute is 0.03 for dark skin. The expected flags per minute of dark skin is around 5 times default skin; around 11 times light skin; is around 3 times medium skin; is around 1.6 times medium-dark skin, and is around 5 times medium-light skin. Thus, darker skins reported more times there was a quality flag during the sleep session.
In summary, the product design may not take into account the impact of skin color on sensor accuracy. Mingar could improve the device for different skin colors.

## Discussion

### Strengths and limitations

Overall, we addressed the client's questions by providing the results base on the exploratory data analysis and generalized linear mixed model. **We found that**:

- New customers are mainly younger than 30 or older than 60 years old, whereas traditional customers are mainly between 30 and 60 years old.
- For every one-unit increase in age, the average log odds of the new customer will increase by 0.005061.
- New customers are mainly from lower median household income neighbourhoods, whereas the old customers are primarily from wealthier communities.

- For every one-unit increase in household median income, the average log odds of the new customer will decrease by 0.000017.
- Customers who purchased the 'Active" and "Advanced' lines are more than the previous products.
- 'iDOL' products tend to malfunction with low frequency, while there are more 'Active', 'Advance', and 'Run' than 'iDOL' among the products that malfunction with higher frequency.
- The average flags per minute of dark skin is around 5 times of default yellow skin; is around 11 times of light skin; is around 3 times of medium skin; is around 1.6 times of medium-dark skin; is around 5 times of medium-light skin.

**However, there are some limitations in the analysis:**

- The customer's skin tone was inferred by the emoji modifier, which may not accurately describe the actual skin tone of customers as some customers may have different skin tones, but still use the default yellow emoji modifier.
- The household median income is the region's median household income rather than the customer's. Thus, it may lead to inaccurate results based on the household median income.
- Postcode and CSDuid were not one-to-one. Sometimes the same postcode has different CSDuids, and we only keep one of them in this case. Thus, the results base on CSDuid may not be that representative.
- The assumptions of constant variance and normality were violated in model 1 of research question 1. Thus the p-value may not be that reliable.
- Overdispersion was presented in model 2 of research question 2, as the sample variance of flags is much larger than the sample mean of flags (response variable); thus, the standard errors might be falsely small, meaning we'll probably have falsely small p-values.
- We used the Wald confidence interval instead of the profile likelihood-based confidence interval. It may lead to the results of the confidence interval are not that accurate.

**Future consideration:**

We may change the models particularly to address some of the limitations shown above for future consideration. For instance, the zero-inflated Poisson model may address the overdispersion. Also, if we can get more detailed information about the customers, we can include more or different predictors in the model to gain other insights.

# Consultant information

## Consultant profiles

**Qianlin Gao**: Qianlin Gao is a junior consultant with Datanerds. She has already worked as an intern in Datanerds in her sophomore year. She specializes in collecting and analyzing data. Qianlin Gao earned her Bachelor of Science, Majoring in Statistics and Economics, from the University of Toronto in 2023.

**Ruizhe Huang**: Ruizhe Huang is a senior consultant with Datanerds. He specializes in data analysis and visualization. Ruizhe Huang earned his Bachelor of Science, majoring in Statistics with an Economics and Mathematics minor, from the University of Toronto in 2023.

**Hang Ren**: Hang Ren is a senior data analyst with Datanerds. He specializes in the reproducible analysis and statistical communication. Hang Ren earned his Bachelor of Science, majoring in Statistics and Economics with a focus in Data Analytics, from the University of Toronto in 2023.

**Xiaoke Zeng**: Xiaoke Zeng is a junior data analyst in the data analysis team at Datanerds. She is an undergraduate student that will graduate in 2024, and this is already her second year of internship with Datanerds. She intended to be a data analyst. Xiaoke is obtaining her Bachelor of Science degree, majoring in statistics and economics, from the University of Toronto in 2024.

## Code of ethical conduct

- The dataset of Datanerd published was organized. The data that were kept in the published version are the data that were permitted to be public. Any of the data that involved customer privacy was unpublicized, which perfectly protected customers' privacy.
- The use of any misleading data or manually modified data was prevented to avoid P-hacking and any other action that will negatively affect the standing of statistics in a good way. For instance, all the values and data in this analysis were actual values directly obtained from the raw data, and no changes were made to the data in order to obtain the expected value.
- Objectivity was maintained throughout this analysis, no judgmental comment was made based on the observed data, personal bias was strictly avoided, public and minority feelings were considered.

## References

[1] Achim Zeileis, Torsten Hothorn (2002). Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. URL https://CRAN.R-project.org/doc/Rnews/

[2] Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

[3] 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.

[4] Dmytro Perepolkin (2019). polite: Be Nice on the Web. R package version 0.1.1. https://CRAN.R-project.org/package=polite

[5] Fitness tracker info hub. (n.d.). Retrieved April 7, 2022, from https://fitnesstrackerinfohub.netlify.app/

[6] Full emoji list, V14.0 - unicode. (n.d.). Retrieved April 7, 2022, from https://unicode.org/emoji/charts/full-emoji-list.html

[7] Hadley Wickham (2021). rvest: Easily Harvest (Scrape) Web Pages. R package version 1.0.2. https://CRAN.R-project.org/package=rvest

[8] Kamil Slowikowski (2021). ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. R package version 0.9.1. https://CRAN.R-project.org/package=ggrepel

[9] Population density. Census Mapper. (n.d.). Retrieved April 7, 2022, from https://censusmapper.ca/

[10] 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/

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

[12] Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society (B) 73(1):3-36

# Appendix

## Web scraping industry data on fitness tracker devices

```
# These are the libraries I find useful for webscraping
library(tidyverse)
library(polite)
library(rvest)

url <- "https://fitnesstrackerinfohub.netlify.app/"

# Make sure this code is updated appropriately to provide
# informative user_agent details
target <- bow(url,
              user_agent = "ruizhe.huang@mail.utoronto.ca for STA303/1002 project",
              force = TRUE)

# Any details provided in the robots text on crawl delays and
# which agents are allowed to scrape
target

html <- scrape(target)

device_data <- html %>%
  html_elements("table") %>%
  html_table() %>%
  pluck(1) # added, in case you're getting a list format
```

The code chunk above shows how we scrapped the industry data on the fitness tracker devices. The data is scrapped from path https://fitnesstrackerinfohub.netlify.app/ with a crawl delay of 12 seconds. The robot text indicates that this path is scrapable by providing informative user agent details, including the email address and the intention of scrapping.

## Accessing Census data on median household income

```
library(cancensus)
options(cancensus.api_key = "API key here",
        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")
```

```
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")

# 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)
```

The code chunk above shows how we access the census data on median household income by API. We sign up for the cancensus API through https://censusmapper.ca/ to get the median household income data for the 2016 Census, as 2020 Census is not up yet. CenusMapper contains a growing set of census-related datasets. To investigate the research questions, we only select CSDuid (Census subdivision unique identifier), median household income, and population in our dataset.

## Accessing postcode conversion files

As the University of Toronto students, we have access to Census Canada Postal Code Conversion Files. The postal code conversion file is a file that allows for the matching of six-digit postal codes to standard census geographies. We download the August 2021 postcode conversion file for 2016 Census data in the .sav file version. We accepted the license agreement to get access to this data. We only save the variables "PC" and "CSDuid" to match the postal codes to standard census geographies.