Market and Product Investigation of MINGAR Wearables

Exploration of new MINGAR customer characteristics and potential racial bias in sleep tracking performance.

Report prepared for MINGAR by FutureGadget, LLC.

2022-04-07

Contents

Executive summary

Background & Aim

MINGAR, a company that traditionally focused on high-end fitness tracking wearable technologies, has recently introduced more affordable Active and Advanced wearable product lines for average consumers, so there may be differences in the characteristics between the customers who purchased new wearable lines and the customers who purchased more high-end, traditional wearable lines of MINGAR. Simultaneously, there have been concerns on social media platforms regarding the poor sleep tracking performance of MINGAR wearables on people with darker skin tones. Therefore, this study aimed to use the customer-level data and external product and census data to explore the differences in characteristics between new and traditional customers of MINGAR. Another goal of this study was to investigate the racial bias in the sleep tracking function quality of wearables and identify any MINGAR wearable lines with this problem.

Key findings

- Table 1 indicates that new customers with Active or Advanced wearables appear to be older on average and are from neighbourhoods with lower average median incomes when compared to traditional customers of MINGAR.
- It is plausible that customers who preferred she/her or they/them pronouns do not have different odds of being a customer of Active or Advance lines than the customers with he/him pronouns, after adjusting for median income, age and battery life.
- Customers 10 years older have 10.1% higher odds of purchasing wearables from Active or Advanced lines when controlled for median income, age and pronouns. Customers from a neighbourhood with $1,000 higher median income have 3.3% lower odds of purchasing the more affordable wearables, when age, pronouns and battery life are constant.
- Customers purchasing wearables with a battery that can last longer than one week have 86% lower odds of choosing the new, affordable MINGAR wearables when median income, age and pronouns are constant.
- The number of quality issues per 100 minutes of sleep for median-aged user is expected to be 0.31, but for a dark-skinned user of the same age, the number of quality issues per 100 minutes is expected to be 10.91 times higher.
- Among all four product lines, the Line Run has the most severe disparity in performance, as there is a nearly 12-time difference in quality issues between dark color and light users was estimated.
- For the Line Advance and Line Active, the difference in quality issues between dark and light skinned-users is nearly 10 times, given that they are of the same age.

- Line iDOL users with medium-light skin color are expected to experience 43% less quality issues than that of medium-skinned users with the same age.
- Age only impacts the number of quality issues in Line Active, where the number of quality issue per 100 minutes is estimated to decrease by 1% for every 10-year increase in age.
- Figure 1 shows the distribution of the number of quality issues per 100 minutes in five skin color categories, which suggests that individuals with darker skin are likely to experience more quality issues during sleep.

Limitations

- For the same postal code, we took the average median income over all the areas with different CSDuid, which could lead to inaccurate neighbourhood median incomes of customers.
- Categories of the original variable battery life were merged into two levels of either having battery life below one week or more than one week, leading to potentially biased results.
- Observations with missing values for pronouns, sex, sleep-tracking related variables, and those who used default yellow color of emoji were removed, which may introduce bias in the analysis because we dropped customer information that might be relevant.
- The color of the emoji used by the customers might not match their actual skin color because the emoji color is only a proxy of the user's skin color. Thus, our results about the impact of skin color on the number of errors would be more reliable if we had the information on customers' race or ethnicity.

Table 1: Average age and median income for new and traditional customers

Customer	Age	Median income
new	50	64829
traditional	47	71586

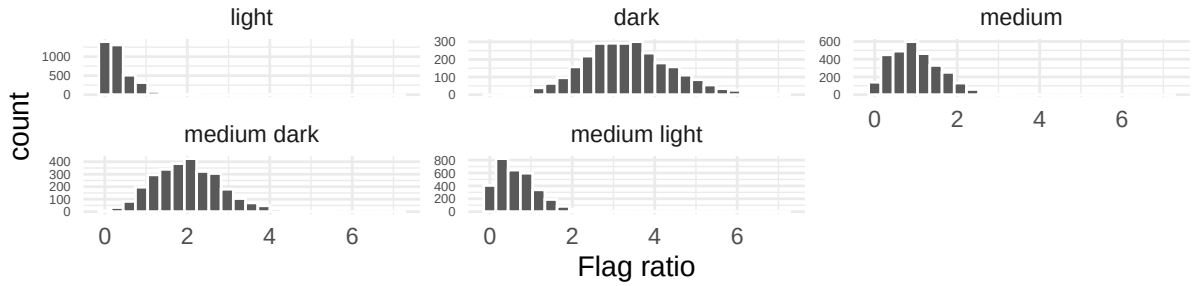

Figure 1: Histograms for flag ratio in different skin color groups.

Technical report

Introduction

As people are paying more attention to personal health and fitness over the past few years, the fitness and activity tracking wearables market have expanded beyond outdoor recreations. Recently, more affordable and compact devices appeared in the market, focusing on providing insights into everyday health for average consumers. MINGAR, a company that primarily focuses on producing high-end fitness tracking wearables, has also introduced Active and Advanced wearable lines with a more approachable price point, to grow and gain a market share for average consumers. As a result, it is essential to investigate the target customers of these two new product lines and elucidate the difference between the traditional customers of high-end MINGAR products and the newer customers that purchased Active or Advanced wearables.

Simultaneously, there have been increasing concerns about wearable devices' poor performance on darker skin tones, especially for sleep scores. Colvonen et al. (2020) reported that wearables might have lower accuracy or do not work on people with darker skin tones due to technological limitations. This difference in wearable performance could make the health constructs inaccessible to selective populations and result in growing health care disparities. Therefore, we aimed to study whether the number of errors or data quality flags during sleep is related to the color of the emoji used by the customer. In this way, we can gain insights into the product considerations regarding racial bias and disparities in personal health assessment.

Therefore, the aim of the present research was to investigate the nuance in the target consumers in the wearable market and the potential racial bias in products for MINGAR. This study is critical because it may help the marketing team to better target their promotions and advertisements to both traditional and newer consumers of MINGAR products. Also, social media team may better address the concerns regarding the performance of sleep tracking wearables and take appropriate actions that are in line with the company values.

Research questions

- The first research question is how the characteristics of traditional and newer customers of MINGAR products differ. Are traditional customers of MINGAR more likely to live in a wealthier neighbourhood with a higher median family income? Are the customers who purchased the new Active or Advanced lines younger than the traditional customers? Are preferred pronouns different for MINGAR's new and traditional customers? Is a battery life of more than one week an essential wearable feature for the new customers compared to the traditional users?

- The second research question is whether users with darker skin colors and greater ages will experience more performance problems regarding sleep scores. And if so, which lines of MINGAR product have such problems?

Data summary

The data used in this analysis were composed of customer-level data, customer-device linkage data, and device data provided by MINGAR and external data relevant to our research, obtained by web scraping, public API, and restricted access via the University of Toronto library. The details of accessing external data and ethical considerations were documented in the Appendix.

Datasets from all sources were merged, and any identifying information of customers in the raw data were removed. Only customer id was kept to represent the different individuals involved in this analysis. Note that during the dataset merging, we found that some postal codes corresponded to more than one CSDuid, and some CSDuid matched two postal codes. CSD, or census subdividsion, represents areas that were treated as equivalent regions for statistical purposes ("Census subdivision (CSD)", 2015). Because we were eventually matching with the postal code in the customer data, we calculated the mean of median household income and summed up the population of regions with different CSDuids for each postal code, and recorded them as the average median household income and total population of the area with this postal code.

Some changes to make the data suitable for analysis were also made. Specifically, each customer's Unicode of emoji characters was converted into one of the levels of the new color variable that represented the color of emoji, including levels light, medium-light, medium, medium-dark, dark and default ("Full emoji modifier sequences, V14.0 - unicode", n.d.). Note that if the user did not intentionally change the color of the emoji, the emoji characters would appear as yellow, which was the default level of the color variable. This variable was thus used as a proxy of the skin color of customers in the analysis. Another variable customer was created to take either level new or traditional, where the level new represented customers who purchased either Active or Advanced lines of wearables and the level traditional represented the customers who purchased traditional MINGAR wearables from all other lines besides Active and Advanced. Then, to remove the sensitive information of the customer birthday, the date of birth was converted into the age by calculating the difference in dates between March 26, 2022 and the date of birth for each customer and dividing the difference by 365. Finally, any sensitive or identifying information were removed from the final dataset used in the analysis, including postal code, emoji Unicode modifier, and date of birth. Other variable that were not relevant to this analysis were also removed from the dataset.

Then, a full dataset was created and observations with missing values were not dropped. Note

that for the two research questions, two different datasets were created from the full dataset and discussed separately in the following sections.

The influence of Median income, age, pronouns and battery life on the choice of wearables

Methods

In the first research question, the sleep tracking information from the full dataset was not of concern, so irrelevant variables were removed from the dataset. Because there were customers with sleep tracking information for multiple days while all other information was the same, duplicated observations with the same customer ids were also removed. Also, all observations with at least one missing value were moved, then a complete-case dataset was created for the first research question.

Because we were interested in the difference between new and traditional customers of MINGAR, the model chosen was a generalized linear mixed model (GLMM), where the response was the binary variable customer with two levels, new and traditional, and the fixed effects were age, median income, pronouns, and battery life. We also recognized that there would be random individual differences in their preferences of purchasing either the newer or the traditional lines of wearables, so we treated individual customers as intercept-only random effects and expected different customers to have different baseline odds of purchasing the new wearables. Interestingly, upon inspection of the data, we found that the property of having a pulse oximeter perfectly predicted the response, where all wearables with pulse oximeters were from the traditional lines of MINGAR wearables. As a result, pulse oximeter would not be appropriate to be included in the model as a potential confounding variable.

We defined a new response variable with two levels, 0 and 1, where level 1 represented the customer purchased either Active or Advanced wearables. To accommodate the binary response variable and better fit the model, two new variables, age centered and median income centered, were created by subtracting the median age and median income and dividing by 10 and 1,000, respectively. This rescaling was also for a more meaningful interpretation of the model intercept. The potential confounding variable population was also centered by subtracting the median population and dividing by 1,000. In the original full dataset, the variable batter life had four levels corresponding to battery life up to 5, 7, 14, and 21 days. However, the differences in the battery life of those four levels were not consistent, so a new binary battery life variable was created to label each wearable as either having a battery that could last less than 1 week or more than 1 week. The variable pronouns had three levels of “she/her”, “he/him” and “they/them”, and it was not transformed.

Table 2: Summary statistics for age and median income, grouped by customer status and pronouns

		Mean (standard deviation)			
Customer	Pronouns	Age	Median income	Count	Proportion
new	he/him	50.37 (21.67)	64868.57 (11512.88)	935	0.40
new	she/her	49.88 (21.67)	64581.36 (11276.86)	1357	0.58
new	they/them	49.16 (23.17)	70955.1 (12574.07)	49	0.02
traditional	he/him	46.76 (16.24)	71746.31 (15067.89)	6787	0.41
traditional	she/her	46.58 (15.98)	71443.67 (14972.56)	9576	0.57
traditional	they/them	46.89 (15.75)	72395.29 (14794.05)	341	0.02

After data manipulation, we performed exploratory data analysis. As seen in Table 2, the new customers who bought either Active or Advanced lines of wearables had a lower average median income than the traditional customers, which suggested that median income might be a significant predictor in the model for customer status. Regardless of pronouns, it also appears that the new customers had greater average ages than the traditional customers, suggesting that age might be significant in the model as well. However, across new and traditional customer status, the proportion of individuals with each pronoun appears to be very close.

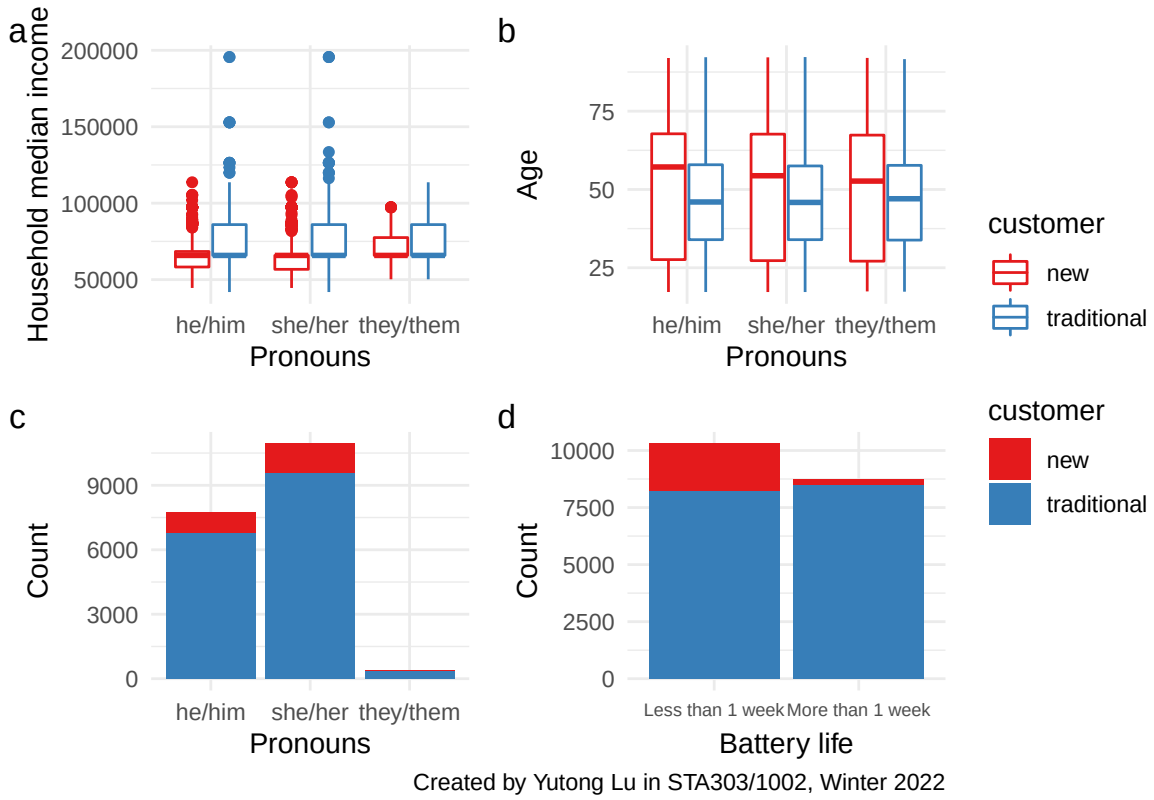

Figure 2: Boxplots plots and barplots for the variables of interest and their interaction with new and traditional customers. The distribution of median household income, age, pronouns and battery life are shown in the plots below, where each of the plot is color coded with the binary response variable, customer.

As indicated in Figure 2, for each category of pronouns, the range of household median income for new and traditional customers appears to be different (Fig 2a) and median age also appears different for different customer statuses (Fig 2b). The categorical bar plots indicated that the proportion of new and traditional customers across pronouns were similar (Fig 2c), yet the customer status proportion appears to be more different across wearables with either less or more than one week of battery life (Fig 2d). Figure 2 suggests that pronouns might not be significant, and the battery life might be significant in determining the customer status.

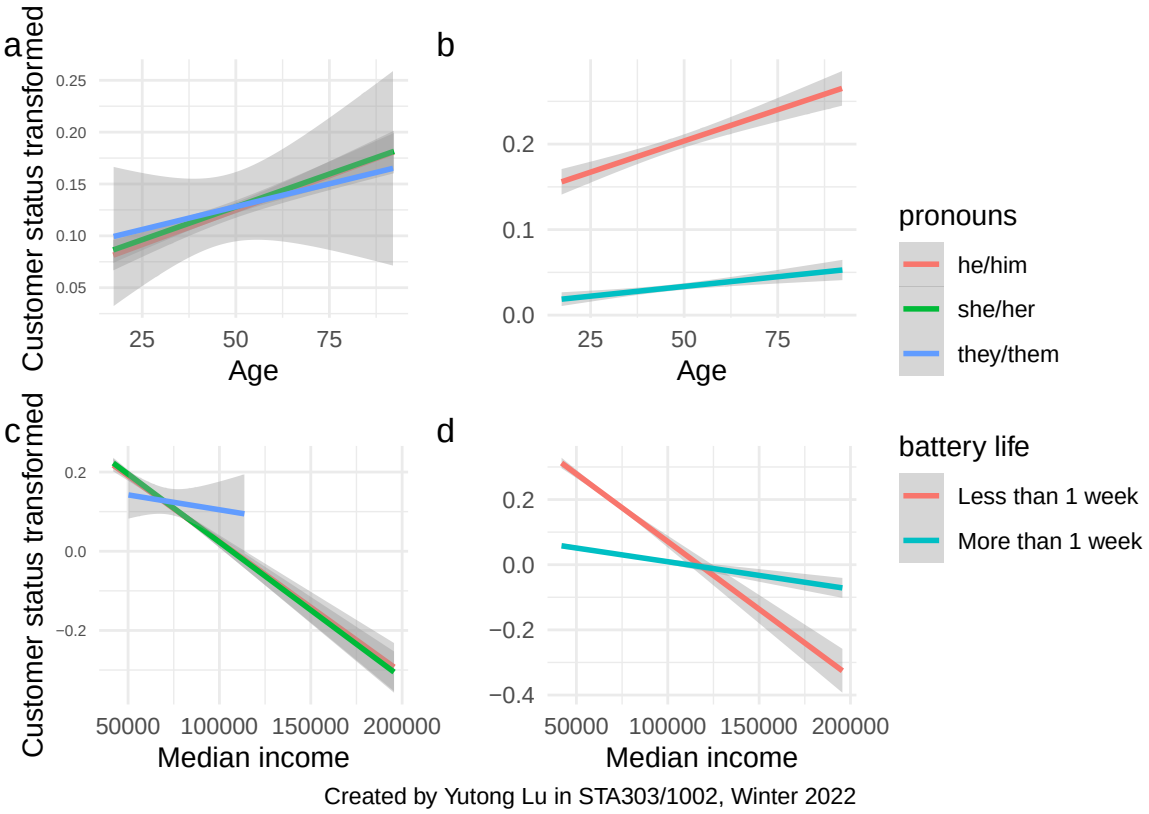

Figure 3: Examination of the interactions between the categorical variables and continuous variables in the model. Note that generalized linear model was used as a method to fit the smooth lines. The figure suggests that there might be significant interactions bewteen the median income and pronouns, and between median income and battery life.

We then examined the interactions between variables in the model, as seen in Figure 3. There appears to be no interaction between the age and the pronouns of the customers (Fig 3a), nor between the age and the battery life of the wearables purchased by the customers (Fig 3b). However, there might be interactions between the median income and the pronouns (Fig 3c), as well as between the median income and the battery life (Fig 3d). As a result, Figure 3 suggests that we might need to consider the interaction between the median income and the two categorical variables in the model.

After examining the data, we fitted an initial generalized linear mixed model (GLMM) with binary response customer status, numerical predictors median income centered and age centered, and categorical predictor pronouns and battery life. The response variable followed a Bernoulli distribution with two possible outcomes, which were being a new customer or being a traditional customer. Median income centered, age centered, pronouns and battery life were fixed effects

since we might be interested in predicting an individual's odds of being a new customer using these aspects. Because we wanted to extend our results beyond the existing customers but we still wanted to adjust for the individual differences in the tendencies of purchasing different wearables, we included an intercept-only random effect of customer id, which we assumed to be normally distributed with mean 0. This model can be specified as follows:

$$Y_i \sim Bernoulli(p_i)$$

$$log(\frac{p_i}{1-p_i}) = \mu + X_i\beta + U_i$$

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

- Y_i is the customer status of individual i, or alternatively, the choice of customer i purchasing a wearable from the new Active or Advanced lines of MINGAR. It comes from a Bernoulli distribution with the probability of choosing the new lines of p_i.
- $log(\frac{p_i}{1-p_i})$ is a log-odds for individual i to be a customer of the new Active or Advanced lines.
- X_i has variables including median income centered, age centered, pronouns and battery life.
- β has the coefficients for each variable (and intercept).
- μ is the grand mean of the log-odds.
- U_i is an individual-level random effect that follows a Normal distribution with mean 0 and variance σ^2.

Then, we would perform ANOVA tests and likelihood ratio tests for model selection. For all the following tests, we would access the strength of evidence against the null hypothesis, where we would have very strong evidence against the null hypothesis for the tests if the p-value was smaller than 0.001 and no evidence against the null hypothesis if the p-value was 1. In order to adjust for potential confounding variables, population and color, we fitted another two models with one extra predictor, population centered and color, respectively. Based on our exploratory data analysis, we also fitted two more models to examine the interaction effects, one with median income centered and battery life interaction, and the other with median income centered and pronouns interaction. Then, we used ANOVA tests to compare four more complex models with the model specified initially, where the null hypothesis was that the smaller model is as good as the more complex model in explaining the data. Note that all five models were nested in fixed effects and had the same random effect of customer id.

We also wanted to include the appropriate random effects and adjust for them in our model. As a result, if we failed to add the fixed effects of population centered and color into the model, we

would use the likelihood ratio tests to compare the model selected in the previous ANOVA tests with another two models with the same fixed effects but one additional intercept-only random effects of color and population centered, respectively. We would also use likelihood ratio tests to compare the model selected from the previous step with other models with the same fixed effect, but different intercept and slope random effects with respect to the slope of each fixed effect. The null hypothesis of the likelihood ratio tests was that the simpler model explains the data as well as the complex model. All models in this section were fitted using `lme4` R package (Bates et al.,2015).

Then, we would check model diagnostics to identify any violations in the model assumptions and address any potential problems in the limitation section. Specifically, we would check the distribution of random effect variables if possible. Also, we would check the variance of model data after being transformed by the link function across categories by separating the dataset into three using pronouns, predicting the log odds using the fitted model and evaluating the variance of the predicted values. If the errors of random effects and the within-unit residual were constant, we would expect the variance of predicted values to be similar.

Results

After fitting all the fixed effects and random effects in the first research question in a GLMM, we compared this initial model with another two GLMMs with additional population centered and color variables, respectively. Based on the results of the ANOVA test, we had no evidence against the null hypothesis that the smaller model can explain the data as well as the more complex model ($p = 1$). On the other hand, we also had no evidence against the null hypothesis that the initial model with no interaction terms is as good as the other two GLMMs fitted with interaction terms ($p = 1$). As a result, we opted for the initial model with fixed effects of median income centered, age centered, pronouns, and battery life.

Then, we fitted two GLMMs with the same fixed effects as the initial model, each with a different random effect. One of the GLMMs had one extra intercept-only random effect of population centered, and the other had the intercept-only random effect of color centered. We performed likelihood ratio tests on our initial model and each of the two GLMMs. The results suggested that we had no evidence against the null hypothesis that the initial model with only random customer id effect is good as the more complex model with both intercept-only random effects of customer id and color ($p = 1$). However, we had very strong evidence against the null hypothesis that the model with only random effect of customer id is as good as the model with intercept-only random effects of both customer id and population centered ($p < .001$). We failed to fit the other nested model with the same fixed effects but different intercept and slope random effects because it was suggested that the random effects are probably unidentifiable, and we did not

choose these models. As a result, we selected the model with both intercept-only random effects of customer id and population centered to adjust for the individual and population differences in the preference of newer or traditional lines of wearables. The likelihood ratio tests were performed using `lmtest` R package (Zeileis & Hothorn, 2002).

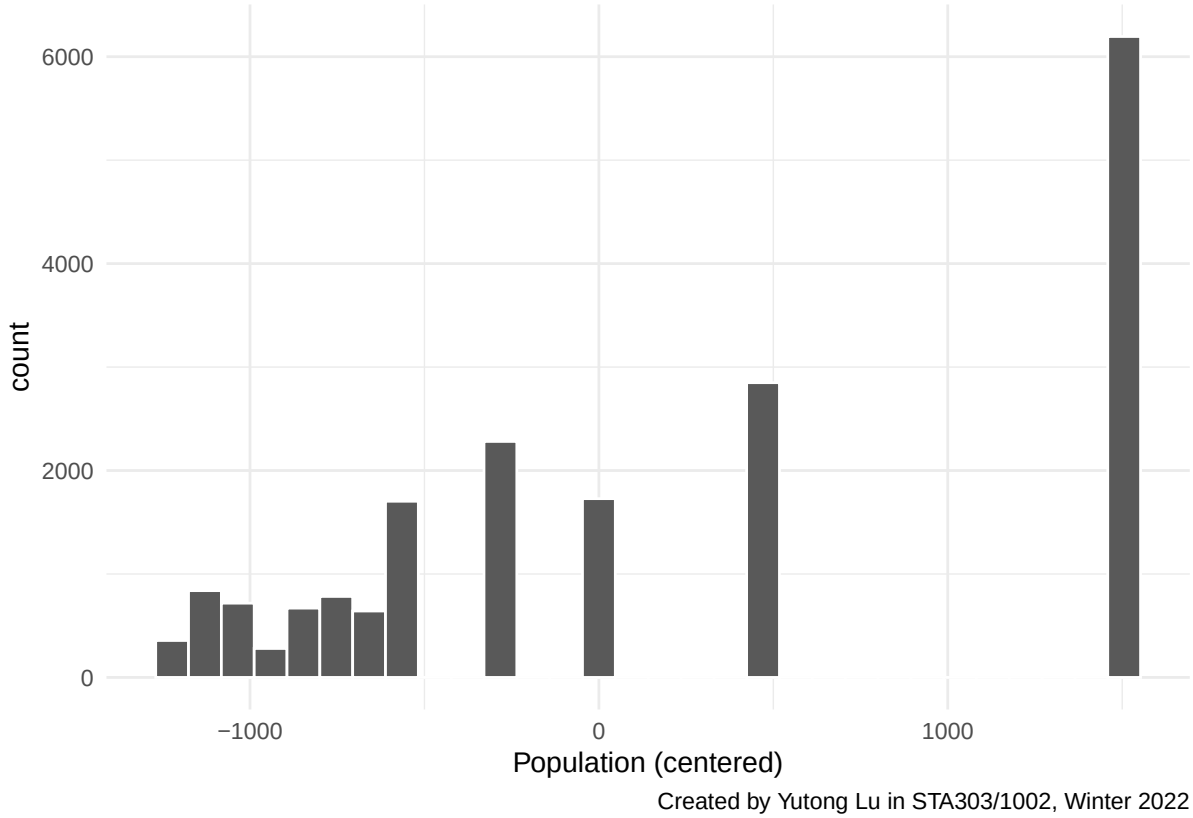

Figure 4: Histogram for the variable population centered. It indicates that the distribution of this random effect did not come from a Normal distribution.

Table 3: Comparing the variance of predicted values for each pronun

Sub-categories of pronouns	Variance of response after transformed by the link function
he/him	1.444565
she/her	1.443990
they/them	1.430629

For model diagnostics, we first assumed that the individual observations were independent, con-

sidering that there should be no influence on the purchase decisions between different customers. Also, we had already removed all the duplicated individual information used for the second research question. Although we could not check whether the random effect of customer id came from a Normal distribution, we checked the distribution of the random effect population centered. As seen in Figure 4, the distribution of population centered does not appear to be Normal, suggesting a potential violation of the assumption for the normality of population centered random effect. The third assumption for the constant variance of random effect errors was checked by comparing the variance of predicted response among different categories of pronouns, which appears to be very close to constant as seen in Table 3. Finally, we defined the response as a binary variable indicating the customer status, so the chosen logit link function was assumed to be appropriate for our model.

Table 4: Model Summary for generalized linear mixed model (exponentiated).

		95% Confidence interval		
	Odds Ratio estimates	Lower bound	Upper bound	p-Value
Intercept	0.2550	0.2219	0.2930	0.0000
Median income (centered)	0.9676	0.9594	0.9759	0.0000
Age (centered)	1.1014	1.0739	1.1296	0.0000
Pronouns (she/her)	1.0310	0.9389	1.1320	0.5227
Pronouns (they/them)	1.0557	0.7644	1.4579	0.7421
Battery life (more than 1 week)	0.1430	0.1256	0.1629	0.0000

Table 4 summarises our chosen model's estimates, 95% confidence intervals, and p-values. At the 95% level, there is no evidence that customers who preferred she/her and they/them pronouns have higher odds of being a customer of the new, affordable lines of wearables after controlling for median income, age, and battery life. On the other hand, a customer who has the median age, prefers he/him pronouns, lives in a neighbourhood with the median of median income, and has purchased a wearable with less than one week of battery life, has a baseline odds of being a customer who purchased the new lines from MINGAR of 0.255. Customers who are ten years older have 10.1% higher odds of purchasing wearables from Active or Advanced lines when adjusted for median income, pronouns, and battery life. After controlling for age, pronouns and battery life, customers from a neighbourhood with $1,000 higher median income have 3.3% lower odds of purchasing new wearables with more affordable price points. Interestingly, customers who purchased wearables with more than one week of battery life have 86% lower odds of being a

customer of the new lines of wearables after controlling for median income, age, and pronouns.

Therefore, these findings may help us to understand the difference between the new and traditional customers of MINGAR wearables. The results of this study indicate that it is plausible that customers who purchased the affordable Active or Advanced wearables are older, from neighbourhoods with higher median income, and do not value long battery life as an essential feature of wearables. Also, the preferred pronouns of customers do not seem to significantly influence the choice of purchasing either new or traditional wearables.

Research into the relation between skin colors and sleep tracking issues

Methods

For the second research question, we would only look at the following variables in the dataset: customer id, date, age, line of product, skin color, flag count, duration of sleeping in minutes, median incomes of the customer's community, and sex. Flags represented the count of the quality issues, which could indicate the frequency of performance problems in sleeping scores. All other variables in the original full dataset were not related to the research question. After removing all observations with missing values in variables sex and sleep duration, we made sure that only the customers with valid sleep tracking records were kept in the dataset. We also removed all observations with their emoji color being default since the default yellow did not necessarily imply the actual skin color. Thus in this analysis, we would use the emoji color as a proxy of skin color and would refer to color as in skin tones or skin colors in the following section. Then we re-scaled and re-centered the age variable so that it was centered in median age and in the unit of every 10 years. By doing this, we can interpret the model's estimate of intercept in a more meaningful way. The duration was also scaled to the unit of every 100 minutes and denoted as duration100. And we used duration100 as an offset later so that the flag counts were adjusted to be comparable across users of different sleep duration. During the exploratory data analysis, we calculated the flag ratio as flags divided by duration100 and explored this ratio's relations with other variables instead. The exploratory data analysis included the data summary of median incomes, ages, duration100, flag ratio, line, skin color, and sex. We also provided histograms and box plots for flag ratios in different skin color groups. Moreover, we provided interaction plots which help to explore potential random effects needed in the model.

Then we fitted different general linear mixed models (GLMM) on the full data, with flag being the response, skin color and age being the fixed effect, and adjusted for different sets of random effects. We chose GLMM because our response variable did not come from Normal distribution, and we would like to take other necessary variables into account as well, in the form of random effects. As mentioned before, we introduced an offset being duration100 on the model.

Meanwhile, the response should be in the family of Poisson, since flags indicated the number of quality issues. The model can be specified as:

$$Y_i \sim Poisson(\lambda_i)$$

$$log(\frac{\lambda_i}{duration100}) = \mu + X_i\beta + U_i$$

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

- Y_i is the number of quality flags raised during the sleep of user i. It comes from a Poisson distribution with the mean number of quality flags per sleep λ_i.
- $duration100$ is an offset indicating how long the sleep is in units of 100 minutes and $\frac{\lambda_i}{duration100}$ represent the mean number of flags per 100 minutes, so that the count is adjusted to be comparable across sleeps of different durations.
- X_i has variables including skin color and age centered.
- β has the coefficients for each variable (and intercept).
- μ is the grand mean of the log mean count.
- U_i is an individual-level random effect that follows a Normal distribution with mean 0 and variance σ^2.

We then used likelihood test to compare the pair of nested GLMMs, where sets of random effects being tested are {customer id, date, median income, sex}, {customer id, date, median income}, {customer id, date, sex}, {customer id, date}, {customer id}, and {date}. By comparing these models, we could find the optimal set of random effects among them, then we used anova test to see if an interaction term between color and age is necessary. We would only keep the more complicated model if the p-value provided by the test was significant with a significance level of 0.05 ($\alpha = 0.05$), and therefore kept the simplest model without losing any necessary random effects in the end.

After obtaining the optimal GLMM, we first checked the mean equal to variance assumption for Poisson regression by looking at mean and variance in each skin color group. We also checked error constant variance assumption for GLMMs by looking at the residual plot. We would expect no fanning patterns in the residual plot. The full dataset was then split into four subsets according to which line of product the observation was from. Then we would fit the GLMM on each sub dataset and see if different skin colors and age would have significantly different means of flags in each dataset. Lastly, we presented the summary and confidence interval of coefficients for each line of product to see individuals with which skin colors experienced the most number of quality issues in this line, if skin colors were indeed a significant predictor for flags.

Results

As part of the exploratory data analysis, summary statistics for continuous variables are given below in the Table 5. Also, the majority of observations were in either Run or Advance line and observations from Active line were only about one fifth of the Run line or Advance line. Interestingly, observations were distributed more evenly when categorized into skin colors. In terms of sex, there were 9095 female, 6002 male, and 146 intersex individuals in the cleaned dataset.

Table 5: Summary statistics in sleep score dataset

Variable	Mean (s.d.)	Median
Age	46.597 (17.45)	45.924
Flags	4.946 (4.6)	4
Median income	7.0135746×10^4 (1.4582968×10^4)	6.5829×10^4
Duration of sleep per 100 min	3.693 (0.428)	3.7

Figure 5 shows the distribution of flag ratio in the five skin color categories. We can see that the mode of the distribution moves further to the right as skin color goes from light to dark. This may suggest that number of quality flags is actually related to skin color such that a darker skin is more likely to experience more sleep tracking issues.

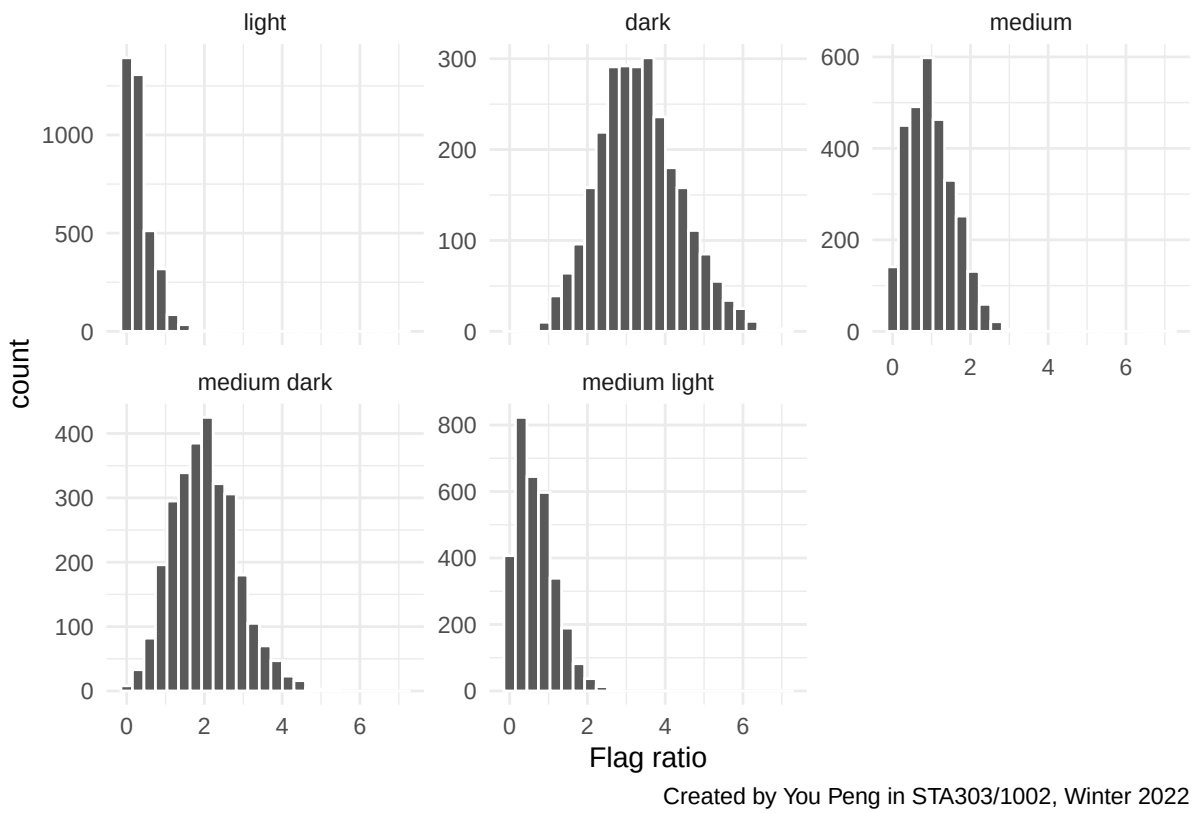

Figure 5: Histograms for flag ratio in different skin color groups.

We can get similar conclusion from the box plots in Figure 6. Observations with dark and medium dark skin color had the highest two medians of flag ratios. This pattern is universal in all three lines of product except iDOL since there were only two different skin tones for customers who used iDOL line. However, for the two categories in iDOL, we can still observe a higher median of flag ratio for the darker skinned observations, which is the medium skin tone in this case.

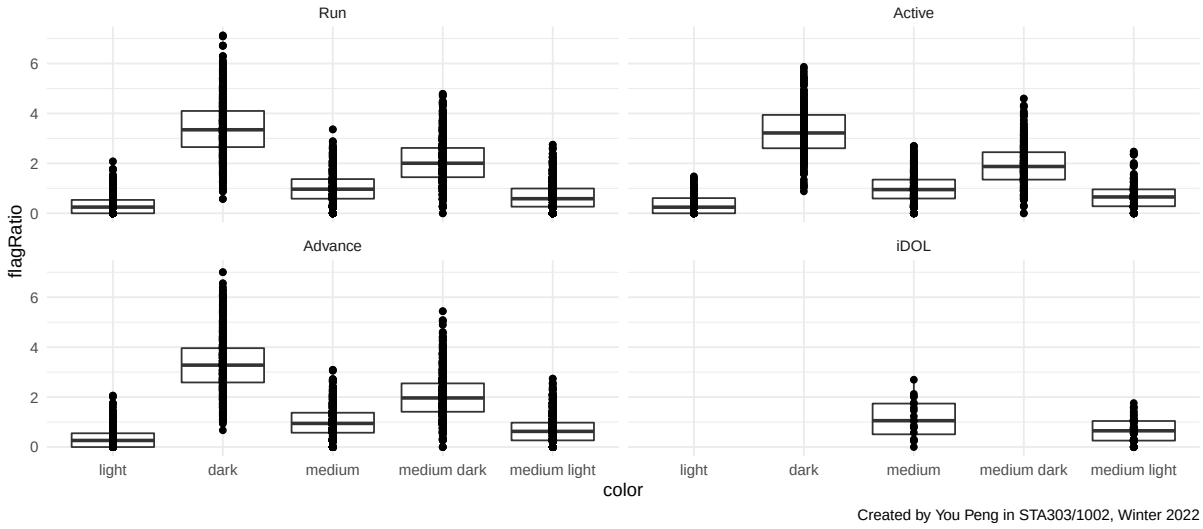

Figure 6: Box plot showing Flag ratio by emoji color as a proxy of skin color

From the following three interaction plots (Figure 6, 7, 8), we can see that there are no obvious difference in intercepts and slopes between flag ratio and skin color. This indicates that sex and median income may not be significant random effects that we need to take into account later. Moreover, the effect of skin color on flags may be similar across four lines of product since there are no significant difference between four lines in the plot.

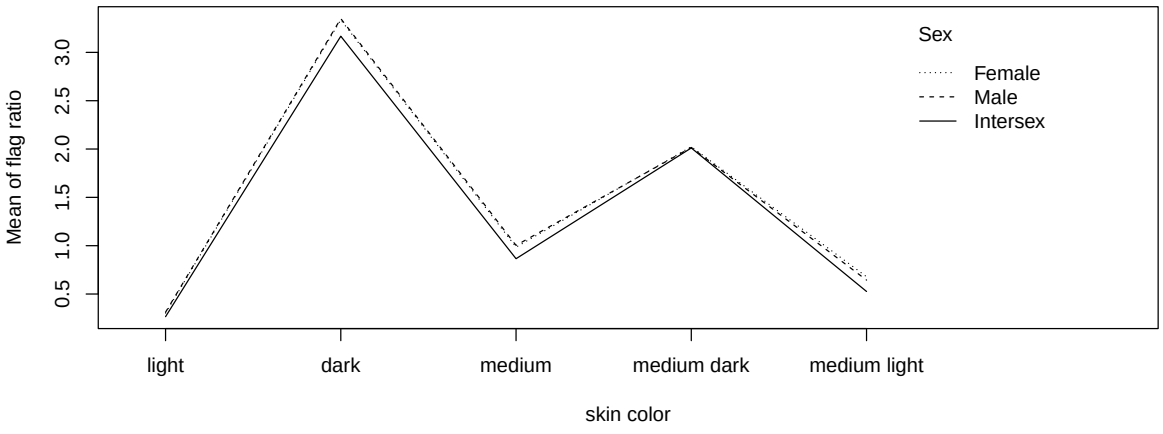

Figure 7: Interaction plot for flag, skin color, and sex

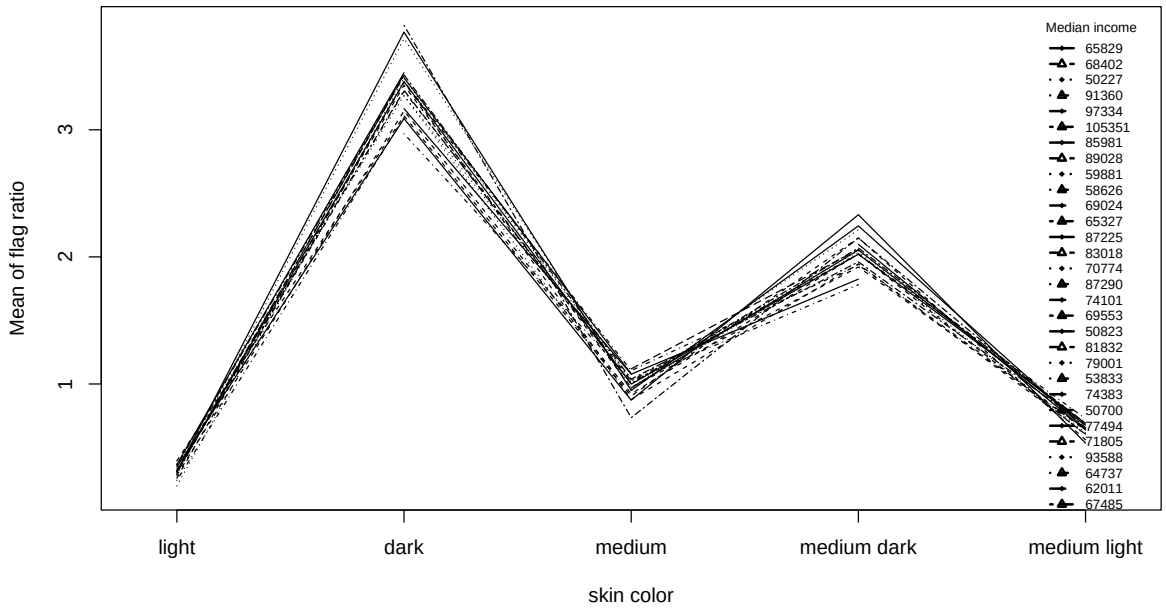

Figure 8: Interaction plot for flag, skin color, and median income

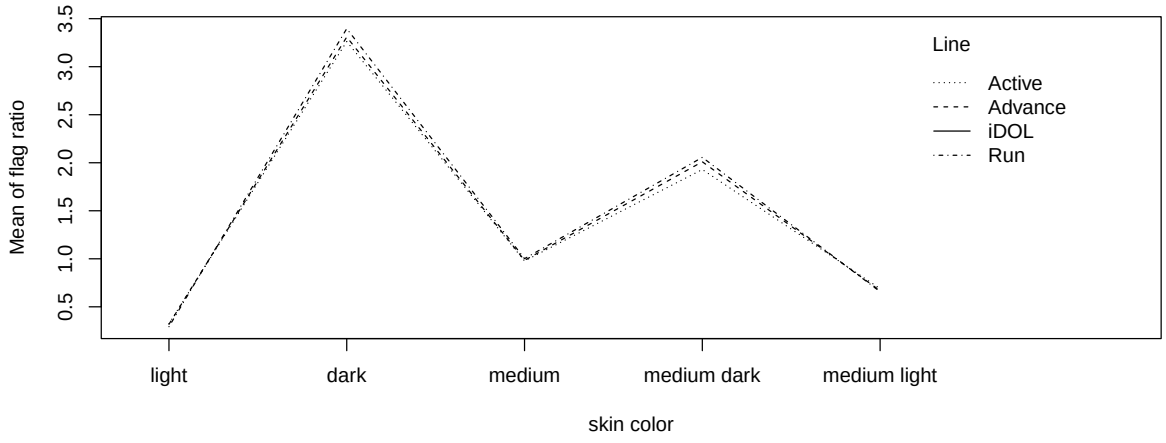

Figure 9: Interaction plot for flag, skin color, and Line

After fitting each model and comparison with the likelihood ratio test, we obtained the final model. The response was the count of flags, skin color and scaled age were fixed effect, and

customer id was the only random effect. Date, sex, and median income were proved to be not necessary according to the test. In addition, the likelihood ratio tests also indicated that random slope for customer id is not necessary. Lastly, the ANOVA tests indicated that an interaction between color and age is not necessary. Table 6 is the summary table for the finalized model. From the summary, it appears that all fixed effects are significant with p-values smaller than a significance level of 0.05. All models in this section were fitted using `lme4` R package (Bates et al.,2015). The likelihood ratio tests were performed using `lmtest` R package (Zeileis & Hothorn, 2002).

Table 6: Model Summary for generalized linear mixed model.

	Flags rate Estimates	p-Value
Intercept	0.3053	0.0000
Dark skin color	10.9135	0.0000
Medium skin color	3.2482	0.0000
Medium dark skin color	6.6242	0.0000
Medium light skin color	2.1738	0.0000
Age (centered)	0.9933	0.0076

For model diagnostics, we first assumed that individual subjects were independent, considering that there should be no correlation on the usage of sleeping tracking function between different customers, although observations within each subject were not considered as independent. Then, we could not check whether the random effect of the discrete variable, customer id, came from a Normal distribution, but we checked the the constant variance of random effect errors by the residual plot in Figure 10. There appeared to be no serious fanning pattern so that we could infer the constant variance assumption was satisfied. We also provided Table 7 to present the variances of data (transformed by the link function) across categories. Thus, we concluded that variances were homogeneous across categories since those variances were very close to each other.

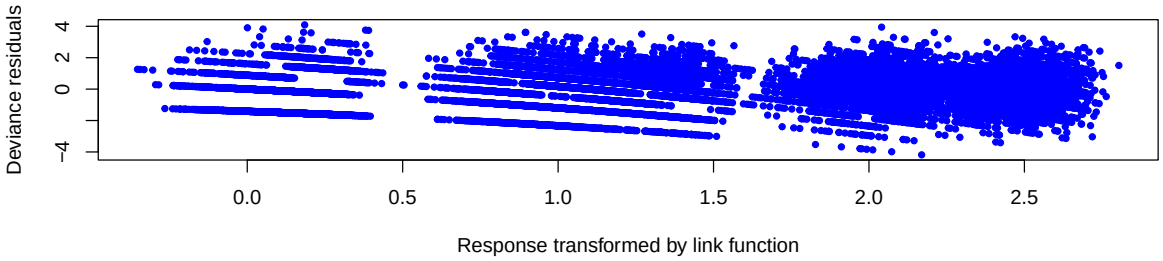

Figure 10: Residual plot on link scale

Table 7: Variance of data across categories

Category	Variance of flags after link transformation
Light	0
Dark	0.001
Medium dark	0.001
Medium light	0
Medium	0.001

Table 8 below shows that means and variances in different skin color groups are not equal. This means that we have some evidence on the violation of mean equal to variance property of Poisson, which was assumed to be the distribution of the number of flags. Finally, for the last assumption of GLMM, we defined the response as a poisson variable indicating the counts of flags during sleep, so the chosen log link function was assumed to be appropriate for our model.

Table 8: Comparing the mean and variance of response variable for each skin color

	mean	variance	count
Light	0.3061452	0.1134198	3658
Dark	3.3399384	1.0973234	2666
Medium	0.9913315	0.3449114	2941
Medium dark	2.0214047	0.6813062	2840
Medium light	0.6639329	0.2456368	3138

After checking all the assumption, the same model was fit on four datasets for each line. From the model summary for Line Run in Table 9, we can see that the flags rate per 100 minutes in the light skin users with a median age is nearly 0.29. And the flags rate per 100 minutes in the dark skin users is nearly 11.57 times (95% confidence interval from 11 to 12 times higher) that of the light skin users controlling for the age of user. Flags rate per 100 minutes in the medium dark skin users is nearly 6.99 times higher than that of the light skin users controlling for the age of user. Flags rate per 100 minutes in the medium skin and medium light skin users are nearly 3.41 and 2.24 times higher than that of the light skin users, respectively, after controlling for the age of user. As a result, the flag ratio per 100 minutes is indeed higher for darker skin colors. There appears to be no significant change in flags rate per 100 minutes as age varies.

Table 9: Model Summary for GLMM on Run Line.

	Estimate	95% Confidence Interval	p-Value
Baseline rate	0.29	(0.28, 0.31)	0
Dark skin	11.57	(11.01, 12.16)	0
Medium skin	3.41	(3.23, 3.60)	0
Medium dark skin	6.99	(6.65, 7.36)	0
Medium Light skin	2.24	(2.12, 2.37)	0
Scaled age	1.00	(0.99, 1.00)	0.33

From the model summary for Line Advance in Table 10, we can see that the flags rate per 100 minutes in the light skin users with a median age is nearly 0.32. And the flags rate per 100 minutes in the dark skin users is nearly 10.18 times (95% confidence interval from 9.12 to 11.41

times higher) that of the light skin users controlling for the age of user. Flags rate per 100 minutes in the medium dark skin users is nearly 5.95 times higher than that of the light skin users controlling for the age of user. Flags rate per 100 minutes in the medium skin and medium light users are nearly 3.03 and 2.15 times higher than that of the light skin users, respectively, after controlling for the age of user. As a result, the flag ratio per 100 minutes is indeed higher as for darker skin colors. There appears to be no significant change in flags rate per 100 minutes as age varies.

Table 10: Model Summary for GLMM on Advance Line.

	Estimate	95% Confidence Interval	p-Value
Baseline rate	0.32	(0.29, 0.36)	0
Dark skin	10.18	(9.12, 11.41)	0
Medium skin	3.03	(2.70, 3.41)	0
Medium dark skin	5.95	(5.32, 6.69)	0
Medium Light skin	2.15	(1.88, 2.47)	0
Scaled age	1.00	(0.99, 1.01)	0.55

From the model summary for Line Active in Table 11, we can see that the flags rate per 100 minutes in the light skin users with a median age is nearly 0.32. And the flags rate per 100 minutes in the dark skin users is nearly 10.34 times (95% confidence interval from 9.83 to 10.88 times higher) that of the light skin users controlling for the age of user. Flags rate per 100 minutes in the medium dark skin users is nearly 6.35 times higher than that of the light skin users controlling for the age of user. Flags rate per 100 minutes in the medium skin and medium light users are nearly 3.10 and 2.10 times higher than that of the light skin users, respectively, after controlling for the age of user. As a result, the flag ratio per 100 minutes is indeed higher for darker skin colors. Unlike other lines of product, there is a significant 1% decrease in flags rate per 100 minutes as age increase from median age by 10 years ($p < .001$).

Table 11: Model Summary for GLMM on Active Line.

	Estimate	95% Confidence Interval	p-Value
Baseline rate	0.32	(0.30, 0.33)	0
Dark skin	10.34	(9.83, 10.88)	0
Medium skin	3.10	(2.93, 3.28)	0
Medium dark skin	6.35	(6.03, 6.69)	0
Medium light skin	2.10	(1.98, 2.22)	0
Scaled age	0.99	(0.98, 1.00)	0

From the model summary for Line iDOL in Table 12, there are only two skin color categories. We can see that the flags rate per 100 minutes in the medium skin users with a median age is nearly 1.15. And the flags rate per 100 minutes in the medium light skin users is nearly 43% lower (95% confidence interval from 56% to 25% lower) than that of the medium skin users, after controlling for user age. As a result, the flag ratio per 100 minutes is indeed higher for darker skin colors. There appears to be no significant change in flag rate per 100 minutes as age varies.

Lastly, among all four lines of product, the Line Run products have the most serious disparities in the effects of skin colors on flag rate per 100 minutes. The Line Advance and Line Active have similar level of effects, and Line iDOL only shows the difference between medium and medium light skin colors which are relatively less severe but still significant. In conclusion, a significant correlation between user's skin color and the rate of experiencing sleep racking issues is observed in all four lines of product.

Table 12: Model Summary for GLMM on iDOL Line.

	Estimate	95% Confidence interval	p-Value
Baseline rate	1.15	(0.92, 1.43)	0.2
Medium light skin	0.57	(0.44, 0.75)	0
Scaled age	1.01	(0.86, 1.19)	0.88

The tables displayed above were created using the `R` package `kableExtra` (Zhu, 2021). The figures were created using `R` packages `tidyverse` and `patchwork` (Wickham et al., 2019; see also

Pedersen, 2020).

All analysis for this report was programmed using `R version 4.1.3` (R Core Team, 2022).

Discussion

The introduction of new, affordable Active and Advanced wearable lines has attracted new average customers to MINGAR products, which were traditionally high-end fitness tracking technologies. Simultaneously, it has been reported on social media platforms that the sleep tracking function of MINGAR wearables performed poorly on individuals with darker skin tones. As a result, the present research aimed to examine the differences in characteristics between the traditional customers and the new customers who purchased the affordable Active or Advanced lines of wearables and investigate the potential racial bias in the sleep tracking function of MINGAR wearables to identify any malfunctioning product lines.

The results for the first research question suggest that it is plausible that the preferred pronouns of customers are not significant for the choice of purchasing either traditional or new MINGAR product lines after adjusting for median income, age and battery life. After controlling for age, pronouns, and battery life, for every $1,000-increase in the household median income, the customers have 3.3% lower odds of purchasing products from the affordable new wearable lines. Conversely, for every 10-year older, customers have 10.1% higher odds of purchasing Active or Advanced products after controlling for median income, pronouns and battery life. Interestingly, customers purchasing wearables with a battery that can last longer than one week have 86% lower odds of being a customer of the new, affordable MINGAR wearables when median income, age and pronouns are constant.

For the second research question, after controlling for user age, we found that the skin color of users has a significant effect on the number of quality issues they encountered, such that users with darker skin colors will experience more quality issues during their sleep. This effect is found in all four lines of product, where the Line Run products have the most severe effect that a nearly 12 times difference in quality issues between dark color and light users were estimated. The Line Advance and Line Active have similar degrees of effect, where the flags rate per 100 minutes in the dark skin users is nearly 10 times that of the light skin users controlling for the age of user. Moreover, for the only two types of skin color in Line iDOL, users with medium-light skin color are expected to experience 43% less quality issues than users with medium skin tone, controlling for the user's age. Lastly, the confounding effect caused by the age of users is only observed in Line Active, where a significant 1% decrease in flags rate per 100 minutes as age increases from median age by ten years is estimated. For the rest three lines of product, the effect of skin colors on quality issues is not confounded by age.

Strengths and limitations

This study has several strengths that make the findings valuable for gaining insights into wearable markets and MINGAR products. By applying ANOVA and likelihood ratio tests to compare different pairs of models, we achieved a balance between being too specific on the customer characteristics and targeting to the proper customers for the new Active or Advanced lines and the traditional wearable lines of MINGAR. For the research into the relationship between skin colors and sleep tracking issues, we took a potential confounding factor age into account to provide a more accurate estimation of skin color's effect on quality issues. Moreover, the analysis was performed for each line of products, and the results were summarized separately to identify the lines of product with most severe problem of racial bias.

However, these results must be interpreted with caution due to the several limitations of this study. First of all, when we matched postal code and CSDuid, we found that the same postal code can lead to different CSDuid and different median incomes in that area. We took the average median income over these conflicted CSDuid areas. This may lead to an inaccurate median income for users with this postal code. Then, the decision we made on the variable "battery life" also caused the prediction to be less specific since we changed the original multi-class variables into two categories, either having a battery life below one week or more than one week. Moreover, we dropped observations with missing values in the variables pronouns. This exclusion of specific observations might induce more bias because there might be some reasons for not specifying pronouns that might relate to the choice of purchasing wearables.

The findings of the second research question may also be somewhat limited. Firstly, we dropped all observations that used the default yellow emoji and those with missing values in sex and variables related to sleep tracking. However, some customers may have skin colors that fall into one of the categories in the variable color, so we did not take their data into account. Also, we did not know if the missing values of sleep tracking related variables were due to quality issues or the fact that they did not use this function, which may lead to potential bias in the results because we might be losing information on wearable malfunctioning. In addition to this, we could not guarantee the color of the emoji used by the customers exactly matched their actual skin color because the emoji color was only a proxy of the user's skin color. For example, even if a customer used a dark-skinned emoji, this individual may or may not be a dark-skinned person. Thus, our results about the impact of skin color on the number of errors would be more reliable if we had information on customers' race or ethnicity. Another limitation is that we could not check the normality of random effect customer id since this is not a continuous variable. Also, the assumed distribution of the response was Poisson, but our data did not satisfy the property of Poisson distribution where the mean equals to the variance, so the model might not be correctly specified. Lastly, some of the models we fitted in both analyses reported

singularity since random effects' variances were estimated too close to 0. However, we considered those random effects, such as customer id, necessary to be included in the model to adjust for individual differences.

In conclusion, this study set out to explore the customer differences between the traditional MINGAR customers and the customers of the newly-introduced affordable wearable lines and investigate the impact of skin tones on the performance of MINGAR wearables' sleep tracking function while using emoji color as a proxy. Notwithstanding the limitations in data and model diagnostics, the results of this study suggest that older customers who live in neighbourhoods with lower median income and do not necessarily value long battery life as an essential feature of wearables tend to purchase the new Active or Advanced lines of wearables rather than the traditional high-end lines of MINGAR. The results also imply that users with darker skin tones tend to experience more sleep tracking performance issues with MINGAR wearables. Eventually, the results of this analysis may provide some insights into the target customers of the new MINGAR wearable lines and the product considerations regarding the disparities in sleep tracking performances based on the user's skin tone.

Consultant information

Consultant profiles

Yutong Lu. Yutong Lu is a senior consultant with FutureGadget, LLC. She specializes in data visualization, casual inference and Bayesian inference. Yutong earned her Bachelor of Science, Specialist in Statistics Methods and Practice, from the University of Toronto in 2023.

You Peng. You Peng is a senior data scientist with FutureGadget, LLC. He specializes in reproducible data analysis and modeling. You Peng earned his Bachelor of Science, Specialist in Data Science and Statistics from the University of Toronto in 2024.

Code of ethical conduct

The following ethical practices are adapted from code of ethical statistical practice (Statistical Society of Canada, n.d.). As part of the ethical statistical consulting, we took responsibility to society by ensuring that the publication of data and results protects users and the company's privacy. For example, after accessing the census data, we only kept the household median income and population so that the raw data were not made available to the public. Moreover, the postal code conversion file was also protected since we only used this file to match CSDuid with postal codes and dropped CSDuid and postal code in the published dataset to protect licensed information.

We also took our responsibility to our clients by ensuring that all assumptions and limitations of our analysis were fully disclosed to our client. Such information about limitations was provided in the discussion section of our technical report as well as the end of executive summary. Therefore, as a consultant, we answered all the questions given by our client MINGAR with our professional knowledge and within our capacity as statisticians.

Lastly, we abided by professional integrity and accountability. We took a serious and responsible attitude to our work and used appropriate methods during the analysis. We were also honest that we did not tamper with the data nor perform p-hacking to obtain significant results artificially.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01.

Census subdivision (CSD). (2015, November 27). Statistics Canada. Retrieved March 31, 2022, from https://www12.statcan.gc.ca/census-recensement/2011/ref/dict/geo012-eng.cfm

Colvonen, P. J., DeYoung, P. N., Bosompra, N. A., & Owens, R. L. (2020). Limiting racial disparities and bias for wearable devices in health science research. *Sleep, 43*(10), 1–3. https://doi.org/10.1093/sleep/zsaa159

Cooley, D. (2022). geojsonsf: GeoJSON to Simple Feature Converter. R package version 2.0.2. https://CRAN.R-project.org/package=geojsonsf

Full emoji modifier sequences, V14.0 - unicode. (n.d.). Retrieved April 1, 2022, from https://www.unicode.org/emoji/charts/full-emoji-modifiers.html

Ooms, J. (2014). The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. arXiv:1403.2805 [stat.CO] https://arxiv.org/abs/1403.2805.

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal, 10*(1), 439-446, https://doi.org/10.32614/RJ-2018-009

Pedersen, T. L. (2020). patchwork: The Composer of Plots. R package version 1.1.1. https://CRAN.R-project.org/package=patchwork

Perepolkin, D. (2019). polite: Be Nice on the Web. R package version 0.1.1. https://CRAN.R-project.org/package=polite

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Statistical Society of Canada. (n.d.). *Code of ethical statistical practice.* Retrieved April 4, 2022, from https://ssc.ca/sites/default/files/data/Members/public/Accreditation/ethics_e.pdf

von Bergmann, J., Shkolnik, D., & Jacobs, A. (2021). cancensus: R package to access, retrieve, and work with Canadian Census data and geography. v0.4.2.

Wickham, H. (2021). rvest: Easily Harvest (Scrape) Web Pages. R package version 1.0.2. https://CRAN.R-project.org/package=rvest

Wickham, H. et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43),

1686, https://doi.org/10.21105/joss.01686

Wickham, H. & Miller, E. (2021). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version 2.4.3. https://CRAN.R-project.org/package=haven

Zeileis, A. & Hothorn, T. (2002). Diagnostic Checking in Regression Relationships. *R News 2*(3), 7-10. URL https://CRAN.R-project.org/doc/Rnews/

Zhu, H (2021). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.3.4. https://CRAN.R-project.org/package=kableExtra

Appendix

Web scraping industry data on fitness tracker devices

The industry data on fitness tracker devices were web scraped using `polite` and `rvest` R packages (Perepolkin, 2019; see also Wickham, 2021). Since no public API was available to provide the information needed, we included a User Agent string to clarify our intentions of web scraping and provide our contacts for any questions and concerns. Also, we strictly followed the minimum crawl delay of 12 seconds per page, as specified by the website. All data scraped from the website were essential for this study, and no extra information was obtained from the website.

Accessing Census data on median household income

The Census data on median household income was obtained using cancensus API via `cancensus`, `sf`, `jsonlite` and `geojsonsf` R packages (von Bergmann, Shkolnik & Jacobs, 2021; see also Pebesma, 2018; Ooms, 2014; Cooley, 2022). Using a personalized, non-public API key, we could get the information of all the regions recorded in the 2016 census. Then the regions with CSD were filtered from the raw region data, which stands for census subdivisions as explained in the data summary. By specifying the regions with CSD and the 2016 version of the census, we obtained the census data with regions labelled with CSD. We only kept the information needed for the study, and thus only the household median income and population data were selected from the 2016 census data. Note that the raw data were not made available to the public, and only the essential information was retrieved from the raw census data.

Accessing postcode conversion files

The postal code conversion file (PCCF) was accessed via the University of Toronto Map and Data Library and imported into R using `haven` package (Wickham & Miller, 2021). PCCF had restricted access and was only available for download for users who agreed to a license agreement by signing in using the institution account. The version of August 2021 postal codes with 2016 census geography was obtained because the census data we obtained previously was from 2016, and the latest version of the postal code conversion file was from August 2021 by the time of this study. Matching the most up-to-date postal codes to CSDuid would provide us with the most relevant median household income and population data. Since this postal code conversion file was not public data, we only used the data in this file to convert the CSDuid in the census data to postal code. As a result, no information in this conversion file was kept in the final dataset used in the analysis.