Customer Persona and Factors Affects Sleep Tracking Performance

Investigating The Active/Advance Buyers' Characteristics and Determining Factors Influence Sleep Tracking Function Performance

Report prepared for MINGAR by Mega Consulting Company

2022-04-07

Contents

Executive summary

This report is responsible for analyzing two questions regarding the Mingar wearable fitness tracker. The first question is about the differences between customers who purchase Active or Advance products and customers who purchase traditional products such as devices from the Run product line. The second question is about whether darker skin color of users affects the performance of the sleep score function, or there some additional variables will affect it as well. In terms of the first questions, we have two findings:

- By looking at Table 1 and Figure 1, we can conclude that new customers care mostly about the low price, the latest version product in the market, and whether the device has GPS while traditional customers care about having the best set of functions for outdoor activities and have higher budgets.
- After fitting the statistical model, we conclude that when given an arbitrary neighborhood and a fixed household income level, the odds of a customer being a new customer increased by 45.6% approximately as age increased from 17 years old to 92 years old. Thus, older people are likely to become the new customers under these given conditions. At a given neighborhood and a fixed age level, the odds of a customer being a new customer increased by 7.1% approximately as household median income increased from 41,880 dollars to 195,570 dollars. Hence, people from higher income households are more likely to become new customers under the given conditions.

In terms of the second question, we have three findings:

- Since we are interested in the relationship between users' skin color and the number of quality flags that occurred during the sleep session, Figure 2 showed that darker-skinned users had the largest range of the occurrence of the number of flags per minute with respect to a sleep session.
- From the fitted statistical model, we observed that users' age and skin color affected the number of flags that occurred during the sleep session. Specifically, the average number of quality flags for a certain user whose skin color is dark and who is 17 years old was 0.034. The average number of quality flags that would occur during a sleep session decreases by 0.06% when the users' age increases by one, for users with the same skin color.
- Keeping users' age the same, the average number of quality flags that occurred during a sleep session for users with medium-dark skin, medium skin, medium-light skin and light skin would be 39.36%, 70.24%, 80.07% and 90.83% less than the average number of quality flags for a dark-skinned user, respectively. Thus, darker-skinned users have a higher average number of quality flags. In other words, the darker skin color of users affects the performance of the sleep score function.

Table 1: Industry data for Mingar products

Device name	Line	Price	Battery life	Water resitance	Pulse oximiter	GPS	Sleep tracking	Released	Heart-rate sensor	Notifications	No-contact pay
Run ON	Run	349.99	Up to 21 days	Waterproof, 10 ATM	Y	Y	Y	2021-12-04	Y	Y	Y
Run BE	Run	299.99	Up to 14 days	Waterproof, 10 ATM	Y	Y	Y	2020-11-20	Y	Y	Y
Run 875	Run	350.00	Up to 14 days	Waterproof, 5 ATM	N	Y	Y	2019-09-12	Y	Y	Y
Run 875 X	Run	399.99	Up to 14 days	Waterproof, 5 ATM	Y	Y	Y	2019-09-12	Y	Y	Y
Run 7	Run	399.99	Up to 14 days	Waterproof, 5 ATM	N	Y	Y	2018-03-09	Y	Y	Y
Run 7 Plus	Run	435.00	Up to 14 days	Waterproof, 5 ATM	Y	Y	Y	2018-03-09	Y	Y	Y
Run HYYH	Run	420.00	Up to 7 days	Waterproof, 5 ATM	N	Y	N	2017-02-18	Y	Y	N
Run Leader	Run	479.99	Up to 7 days	Waterproof, 5 ATM	N	Y	N	2016-09-12	Y	Y	N
Run	Run	450.00	Up to 5 days	Waterproof, 5 ATM	N	N	N	2015-08-01	Y	N	N
iDOL	iDOL	199.99	Up to 14 days	Waterproof, 10 ATM	N	Y	Y	2018-08-24	Y	Y	Y
Advance 2	Advance	145.00	Up to 7 days	Resistant	N	Y	Y	2021-07-08	Y	Y	Y
Advance	Advance	120.00	Up to 7 days	Resistant	N	Y	Y	2020-08-20	Y	Y	Y
Active Alpha	Active	99.99	Up to 7 days	Resistant	N	N	Y	2020-12-30	Y	Y	Y
Active	Active	39.99	Up to 14 days	Resistant	N	N	N	2019-10-13	N	N	N
Active HR	Active	79.99	Up to 7 days	Resistant	N	N	N	2019-10-13	Y	N	N

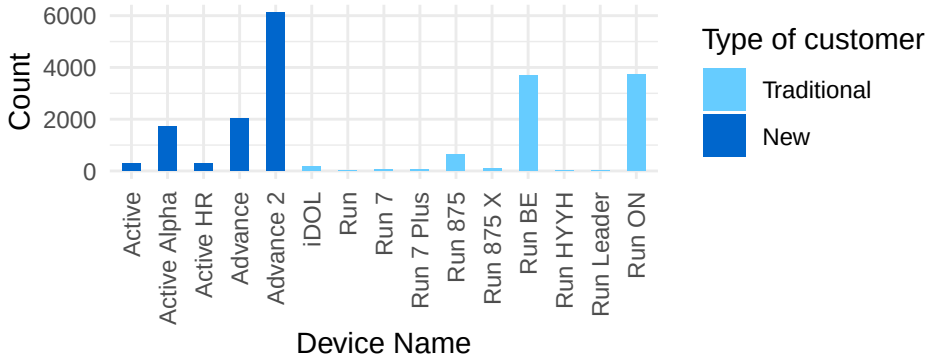

Figure 1: Count number of customers for each type of device

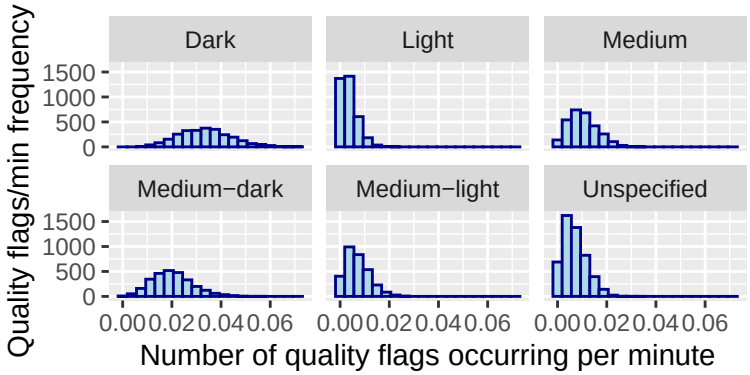

Figure 2: The histogram about the number of quality flags per minute during the sleep session in different skin color customers

Technical report

Introduction

After Mingar released the Active and Advance product line, there are an increasing number of customers who chose to own one of these products. Since these newly released products are extremely popular and attract lots of new customers, it is important to understand the factors or characteristics that differentiate the new customers from the traditional ones.

In this report, we will investigate this question by analyzing the preference with respect to device functions and price of new customers and traditional customers directly from data and by building a statistical model (e.g. Generalized Linear Mixed Model).

After the social media team of Mingar received some complaints about the sleep score functionality of devices, they also would like to find out if the trigger reason for the poor performance of sleep score is the user's skin color, especially for darker skin color, or anything else.

The rest of the report of each question is organized as follows. In **Data** section, we discussed how we collected our data from various sources and the cleaning process. Also, we displayed the results of exploratory data analysis in this section. In **Method** section, we discussed the methods and model assumptions as well as how to assess them to get the best results. Then, in **Result** section, for the first research question, we showed the results of analyzing the differences in customers' characteristics directly from data and models as well as the conclusion we drew from these two methods. And for the second research question, the model parameters were analyzed and discussed according to the research question. Lastly, **Discussion** section summarized our finding to the question, as well as pointing out the strengths and limitations.

Research questions

- As the new products Active and Advance are targeting customers that are different from customers that purchase traditional products, how are the characteristics of customers purchasing Active and Advance different from the ones of customers buying traditional products?
- Many reports from users claimed that the sleep score feature performs poorly for users with darker skin. Does darker skin color really affect the performance of sleep score function?

Research Question 1

Data

Data Wrangling The data we used for analysis comes from various sources. Firstly, the data about customers personal information, customers devices, and device information were provided by the client. These data sets were merged together. Secondly, using private access to the census postal code conversion files, we chose to download the one for 2016 since it is the latest version. By inspecting this data, we found that one postal code could be corresponding to multiple Census subdivision unique identifiers. Thus, to remove the duplication, we chose to keep the first record of each postal code. Thirdly, through Cancensus API, we extracted the Canadian Census data for 2016 which was the latest version. Then, we only kept columns that were used for the analysis (Census subdivision unique identifier, household median income, and the population of the area for the postal code). Fourthly, in order to get the data for all the wearable fitness tracker in the Canadian market, we used web-scraping ethically to get the industry data for all brands of wearable fitness trackers available in Canada.

After collecting data, we merged the customer-device data, census postal code conversion data, and household median income data together to form a larger data set. Lastly, we dropped the variables that were not important for the analysis (e.g. postal code, pronouns, product release date, and device id).

Next, in the final data set, we cleaned the final data that observations that contain missing sex value or contain intersex. Also, we used date of birth to calculate the age of the customer and remove the date of birth variable. Next, we imputed the color of customer's skin by mapping the value in `emoji modifier` to the skin tone of the emoji and remove the `emoji modifier` variable. Also, we removed the Population variable since we kept the household median income variable which reflects the same information about the neighborhood as Population does.

Finally, we recoded some categorical variables in the following way:

- `Male` was coded as 1 and `Female` was coded as 0 in `sex` variable
- Created a new variable `new customer`. In the `line` variable, if customer uses Active or Advance products then the value in `new customer` is 1 indicating new customers. If customer uses Run or iDOL products then the value in `new customer` is coded 0 indicating traditional customers.

Data Summary From Table 2, we can conclude that there are 18,823 observations in the data. Approximately 41% of the observations are male customers. The average age of the 18,823 customers is around 47 years old. The age of the youngest customer is 17 and the oldest customer

Table 2: Summary statistics for numerical variables in the data

	Number of Observation	Mean	Standard Deviation	Min	Max
sex	18823	0.41	0.49	0	1
age	18823	47.12	16.89	17	92
dark skin color	18823	0.14	0.35	0	1
medium-dark skin color	18823	0.14	0.35	0	1
medium skin color	18823	0.14	0.34	0	1
medium-light skin color	18823	0.16	0.36	0	1
light skin color	18823	0.17	0.38	0	1
new customer	18823	0.55	0.50	0	1
household median income	18823	70762.41	14790.24	41880	195570

is 92 years old. Moreover, there are 55% of customers purchased Active or Advance line product, thus 55% of our customers are new. The average household median income for household under the same postal code is approximately 70,762.41 in dollars. The lowest median income for the household under one postal code is 41,880 dollars while the highest is 195,570 dollars. customers with dark skin color, medium-dark skin color, and medium skin color are all 14% respectively among all customers. Customers with medium-light skin color and light skin color are 16% and 17% respectively among all customers.

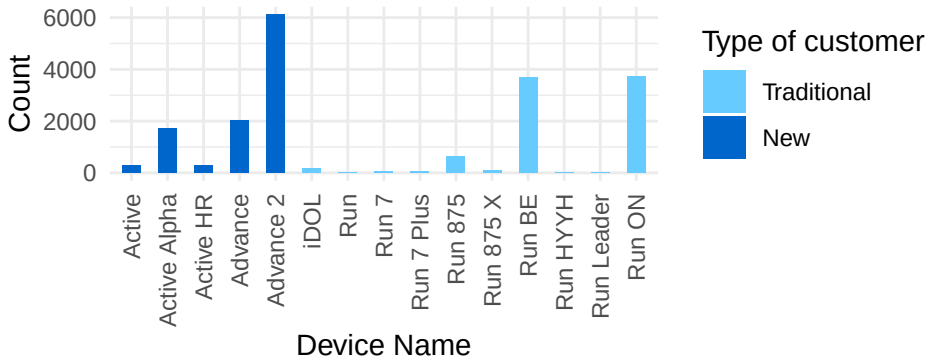

Figure 3: Count number of customers for each type of device

Figure 3 demonstrates the number of customers that own each type of device. It is obvious that around 1/3 customers own Advance 2. For traditional customers, most of them either own the Run ON or Run BE and the both type of devices have similar shares among all traditional customers, while only a small portion of customers own iDOL or other devices in the Run product line. Among new customers, besides the huge share of Advance 2, Advance has the second largest customers and the Active Alpha is the third. Compared to the Advance 2, Advance, and Active Alpha, the other two new products (e.g. Active HR and Active) have trivial shares.

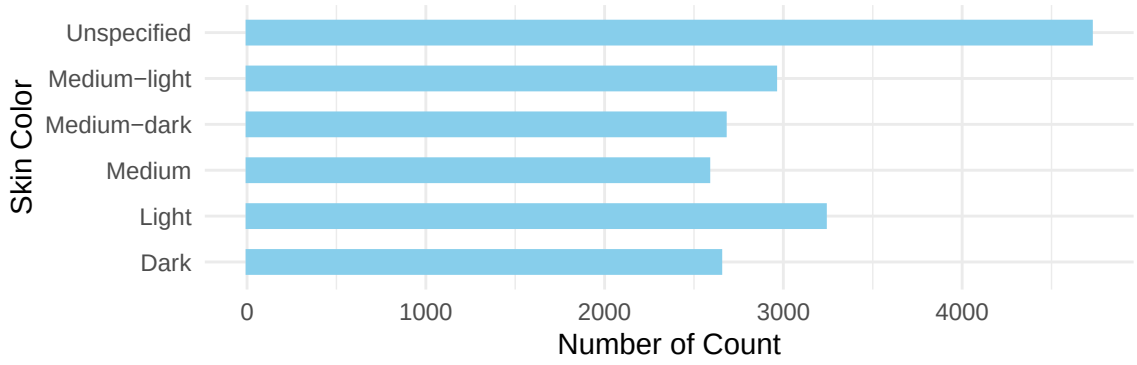

Figure 4: Count number of each type of skin color

Figure 4 shows the the number of customers for different imputed skin colors. Most customers do not specify the skin tone of the emoji and use the default color. Other than these customers, most of other customers are imputed as light-skin customer. Medium-dark and Dark are nearly the same whereas Medium has the lowest customer share.

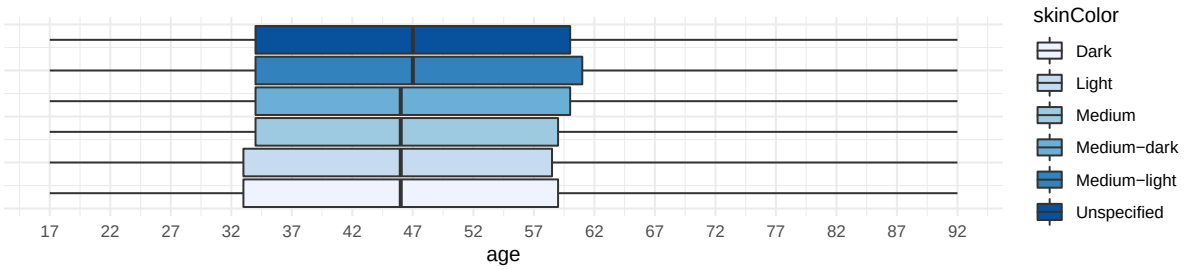

Figure 5: The boxplot for age for each type of skincolor

Figure 5 illustrates the distribution of age for each skin color group. The mean age for each skin color group are all around 47 years old.

Method First of all, we inspect the difference between preferences of new customers and traditional customers when choosing their device to buy. Then, using the industry data for wearable fitness tracker devices, we conclude what features of the device each type of customers care about the most.

Secondly, since the question we are investigating is the difference between traditional customers and new customers who buy Active and Advance products and we have a binary variable to indicate the type of customer, it is reasonable to use logistic regression. Also, we consider that there exists a random effect that observations with same Census Subdivision Unique Identifier (`CSDuid`) are people who live under the same area have correlations with each other. Thus, in this

situation, the Generalized Linear Mixed Model (GLMM) is strongly preferred with the `CSDuid` as the random intercept.

The starting model is displayed below.

$$(\text{Being a new customer})_{ij} \sim Binomial(N_i, \rho_{ij})$$

$$\begin{aligned} log(\frac{\rho_{ij}}{1-\rho_{ij}}) = \beta_0 &+ \beta_1 * \text{sex}_{ij} + \beta_2 * \text{age}_{ij} + \beta_3 * \text{skin color dark}_{ij} + \beta_4 * \text{skin color medium-dark}_{ij} \\ &+ \beta_5 * \text{skin color medium}_{ij} + \beta_6 * \text{skin color medium-light}_{ij} + \beta_7 * \text{skin color light}_{ij} \\ &+ \beta_8 * \text{household median income}_{ij} + U_j \end{aligned}$$

- N_i represents the number of observations
- ρ_{ij} represents the probability of i-th customer is new customer who owns a Active or Advance device under the j-th Census SubDivision Unique Identifier
- sex_{ij} represents the sex of i-th customer under the j-th Census SubDivision Unique Identifier
- age_{ij} represents the scaled age of i-th customer under the j-th Census SubDivision Unique Identifier. The age is scaled by the formula $\frac{\text{age}-\text{minimum age}}{\text{maximum age}-\text{minimum age}}$ thus the scaled age is ranged from 0 to 1. When age equals 0, it means the customer is at the minimum age among all customers, in our case, 17 years old. When age equals 1, it means the customer is at the maximum age which is 92 years old. The reasons why we did this scaling are that firstly age = 0 would have appropriate meaning when interpreting the estimated coefficients and secondly it can reduce the chance that the statistical software fails to compute the estimation of the coefficients of the model.

Since there are different type of customer's skin color, we construct 5 dummy variables with the base value as "Unspecified".

- skin color dark_{ij}, a dummy variable represents the skin tone of i-th customer is dark under the j-th Census SubDivision Unique Identifier
- Other dummy variables are skin color medium-dark_{ij}, skin color medium_{ij}, skin color medium-light_{ij}, and skin color light_{ij}. They all represents the same thing as skin color dark_{ij} does but for their own corresponding skin colors
- household median income_{ij} represents the scaled household median income for the i-th customer under the j-th Census SubDivision Unique Identifier. The reason of scaling this variable is similar to the one of scaling age.
- U_j represents the random effect caused by the j-th Census SubDivision Unique Identifier.
- β_0 represents the log odd of being a new customer given that this female customer is at the minimum age (18 years old), has minimum household median income, and does not specify her skin tone under the same Census SubDivision Unique Identifier

- β_1 represents, under the same Census SubDivision Unique Identifier, the difference between the log odd of a male customer becomes a new customer, when he is at the minimum age (18 years old), has minimum household median income also does not specify his skin tone, and the log odd of a female customer becomes a new customer when she is at the minimum age (18 years old), has minimum household median income, and does not specify her skin tone
- β_2 represents, under the same Census SubDivision Unique Identifier, the difference between the log odd of a 92 year-old female customer being a new customer, given that her household median income is at the minimum as well as does not specify her skin color, and the log odd of a 17 year-old female customer being a new customer, given that her household median income is at the minimum and does not specify her skin color.
- β_3 represents, under the same Census SubDivision Unique Identifier, the difference between the log odd of a dark-skin female customer who is 17 years old and has minimum household median income and the log off of a unspecified-skin-color female customer who is 17 years old and has minimum household median income
- β_4, β_5, β_6, and β_7 represents similar meanings as β_3 but for their own corresponding skin color.
- β_8 represents, under the same Census SubDivision Unique Identifier, the difference between the log odd of a 17 year-old, unspecified skin tone female customer that has maximum household median income and the log odd of a 17 year-old, unspecified skin tone female customer that has minimum household median income

Before implementing the model, there are 4 assumptions that need to be checked.

- Grouping units are independent, even though observations in each group are taken not to be
- Random Effects come from a normal distribution
- The random effects errors and within-unit residual errors have constant variance
- The chosen link function is appropriate

For assumption 1, since the grouping unit is household median income which is independent for one area to another, it is satisfied. For second assumption, it is fairly hard to verify, given the tools and knowledge we have currently. To verify the third assumption, we generate residual plots to see if there are constant variance. The last assumption is satisfied since we have a binary outcome variable and choose the logit to be the link function can be easier to interpret the results.

After checking the assumptions, we start with fitting the starting model and check the estimated coefficients and their p-values. Then, we drop statistically insignificant variables and use Likelihood Ratio Test to check if it is appropriate to drop that variable. Eventually, We keep the

best model and build interpretation on it.

Result From Figure 3, we already discussed that, among new customers, the most popular device is Advance 2. And for traditional customers, the most popular devices are Run ON and Run BE. Then, we need to compare which features make these devices stand out among their customers.

From Table 11 in the Table Appendix, we can see that traditional customers care more about whether a device has a full set of functions and these functions need to be the best among all devices. For example, the most popular devices among traditional customers are RUN BE and RUN ON. Both of them have the ability of waterproof under maximum 10 atmosphere pressure, the longest or second longest battery life, and with all available functions for a fitness tracker. These customers are relatively not sensitive to price but more sensitive to whether the device is newer in the market.

On the other hand, for new customers, the most popular device is Advance 2. Hence, they are relatively sensitive to price comparing to traditional customers and do not need the best version in terms of functions. For instance, these new customers do not care too much about battery life, waterproof, or pulse oximiter. However, they care about GPS and whether the device is new to the market. For instance, Advance and Advance 2 are the most popular two and are totally similar in terms of functions. Because Advance 2 is more recent product, it has the highest sales.

Next, we fit a Generalized Linear Mixed Model to further analyze the characteristics of new customers. First of all, we need to check the second assumption of GLMM that random effects come from a normal distribution. By plotting a histogram of all the estimates of the random effect, we observed that the distribution of the estimated random effects follows a Normal distribution. Thus, we can alternatively say that the assumption that Random effects come from Normal distribution is satisfied.

Then, we plot the residual vs. fitted plot and the pattern we observed is normal for a Binomial Distribution. Hence, we can continue fitting the model as all assumptions are satisfied.

Table 3: Summary table for starting model

	Estimate	Std..Error	z.value	p_value
(Intercept)	0.6556	0.0979	6.6958	0.0000
sex	0.0376	0.0303	1.2407	0.2147
age	0.3754	0.0663	5.6627	0.0000
skin color dark	0.0094	0.0506	0.1863	0.8522
skin color medium-dark	-0.0387	0.0503	-0.7686	0.4421
skin color medium	0.0216	0.0502	0.4308	0.6666
skin color medium-light	0.0240	0.0482	0.4977	0.6187
skin color light	-0.0156	0.0470	-0.3328	0.7393
household median income	-2.6711	0.3795	-7.0383	0.0000

After fitting the starting model, we can observed that, from Table 3, the coefficient of the `age` is statistically significant since its p-value is smaller than 0.001 which means we have very strong evidence against the null hypothesis that the coefficient of `age` is 0. Also, the coefficient of the `household median income` is statistically significant since its p-value is smaller than 0.001 which means we have very strong evidence against the null hypothesis that the coefficient of `household median income` is 0. On the other hand, `sex` and all of the dummy variables indicating skin colors are not statistically significant, thus we would firstly remove the `sex` from the starting model and use Likelihood Ratio Test (LRT) to assess if `sex` should be dropped.

After running the LRT, the p-value of this test is around 0.2147 which suggests that there is no evidence against the null hypothesis that the reduced model (e.g. the model without `sex`) is as good as the more complex one (e.g. the starting model). Hence, we accept the following model as the best model so far.

$$\begin{aligned} log(\frac{\rho_{ij}}{1-\rho_{ij}}) = \beta_0 &+ \beta_1 * \text{age}_{ij} + \beta_2 * \text{skin color dark}_{ij} + \beta_3 * \text{skin color medium-dark}_{ij} \\ &+ \beta_4 * \text{skin color medium}_{ij} + \beta_5 * \text{skin color medium-light}_{ij} + \beta_6 * \text{skin color light}_{ij} \\ &+ \beta_7 * \text{household median income}_{ij} + U_j \end{aligned}$$

Then, we inspect the summary table after fitting this new model, we still observe that `age` and `household median income` is statistically significant while all dummy variables that indicate the type of skin color of a customer are not. Hence, we produce a reduced model by removing all the dummy variables from current best model. The LRT again shows that the p-value for this test is 0.866, indicating the reduced model is as good as the more complex one. Therefore, we have the final best model that is displayed here:

$$log(\frac{\rho_{ij}}{1-\rho_{ij}}) = \beta_0 + \beta_1 * \text{age}_{ij} + \beta_2 * \text{household median income}_{ij} + U_j$$

Thus, plug in estimated coefficient from Table 4 to the final model, we have

$$log(\frac{\rho_{ij}}{1-\rho_{ij}}) = 0.6662 + 0.3765 * \text{age}_{ij} - 2.6514 * \text{household median income}_{ij} + U_j$$

Table 4: Summary table for model without sex and skin color

	Estimate	Std..Error	z.value	p_value
(Intercept)	0.6663	0.0928	7.1836	0
age	0.3762	0.0663	5.6766	0
household median income	-2.6513	0.3746	-7.0783	0

Table 5: Exponential of the estimate of final model coefficient

	Estimate	95% CI: 2.5%	95% CI: 97.5%
(Intercept)	1.9470200	1.6233528	2.3351355
age	1.4567385	1.2793231	1.6588689
household median income	0.0705594	0.0338617	0.1470199

For easier interpretation, we need to consider the exponential of estimated coefficients and the results are displayed in Table 5. The intercept is around 1.95 which represents when holding the random effect fixed the odd of being a new customer for a customer who is 17 years old and has the minimum household median income. The coefficient for age is 1.46 approximately which represents that, holding random effect constant and fixing at the minimum level of household median income, the odd of being a new customer for a 92 year-old customer is 1.45 times larger than the odd of being a new customer for a 17 year-old customer. The coefficient for household median income is around 0.071 which means that, holding random effect constant and fixing at the minimum age 17, the odd of being a new customer for a customer whose household income is at maximum level is 7.1% larger than the odd of being a new customer for a customer whose household income is at the minimum level. The 95% confidence interval for each estimate does not include 1, then it again proves that the estimations are statistically significant. The 95% confidence interval for the estimate of age shows that we are 95% confident that the true magnitude of the effect of age on the odd of being a new customer is between 1.28 and 1.66. The 95% confidence interval for the estimate of household median income shows that we are 95% confident that the true magnitude of the effect of household median income on the odd of being

a new customer is between 0.034 and 0.147.

Conclusion In conclusion, by combining the analysis done with the industry data and the result of the GLMM model, we can see that the new customers are the people who has little need of a powerful and expensive outdoor fitness tracker unlike traditional customers but care about having the GPS function, relatively low price, and whether it is the newest in the market. Moreover, from the result of the GLMM, it is more likely for older people to become new customers. It is reasonable because as people ages it is hard for them to participate in aggressive outdoor activities in general. Thus, they are unlikely in need of a fitness tracker that can be fully functioning under water with 10 ATM pressure. Hence, Advance and Active devices are more attractive for them than the traditional products. Another conclusion we can draw from the GLMM model is that people with higher household median income may be more likely to become a new customer. This proves the claim that new customers are sensitive to price because if one device is from the Run product line while another device is from Advance or Active product line and both device overall performance are similar people with higher income household still purchase the device with lower price. Nevertheless, there is a few caveats about the result of the model which are discussed in the Limitation section.

Research Question 2

Data

Data Wrangling Since the subject of our study is the device's sleep function aspect, we would focus on the provided dataset containing customers' sleep data. Both the datasets of customers' personal information and devices purchased by customers were merged with this dataset by customer ID and device ID. All combined data were stored in a new dataset *merged customer sleep* and unrelated variables postal code, customer's pronouns for social profile, device ID, and release date of the particular device were removed.

In addition, removed missing values in the user's sex based on the merged dataset and only included males and females since the proportion of intersex was relatively small overall. Then a new variable representing the user's age was created based on their date of birth. Created a new variable for the user's skin color to represent the corresponding skin tone for their emoji modifier when using the chat and react features of the social component of the company's app [17]. The missing value of the user's emoji modifier represented that they used the default yellow color, which was classified as "Unspecified" skin color. Moreover, rescaled users' age into the range from 0 to 1 by subtracting the minimum age level in the dataset (17 years old) and then dividing by the difference between the maximum age level (92 years old) and the minimum age level

in the dataset. This would allow fitting any model more efficiently and conveniently later on. Besides, created two variables: the number of quality flags per unit(minute) during sleep session; the mean of the number of quality flags per minute for each age level. These two variables would be used in data exploration to construct figures or tables.

In the end, the cleaned-up *merged customer sleep* dataset was output and read in as the *sleep* dataset we would explore and analyze later on.

Variables of Interest

- Skin color: the skin color of a user, transferred from the code for skin tone modifier for emojis when using the chat and react features of the social component of the company's app.
- Flags: the number of times there was a quality flag during a user's sleep session for their devices. This might due to device's missing data, sensor error or other data quality issues. We would use it to assess the sleep score function of a device.
- Duration: the length of sleep session in minutes.
- Age: the age of a user.
- Customer ID: unique ID for each user.
- Device name: Name of device type.

Data Exploration Firstly, since we are mainly interested in the sleep score function for users with different skin colors, then we would explore it through graphical and numerical summaries.

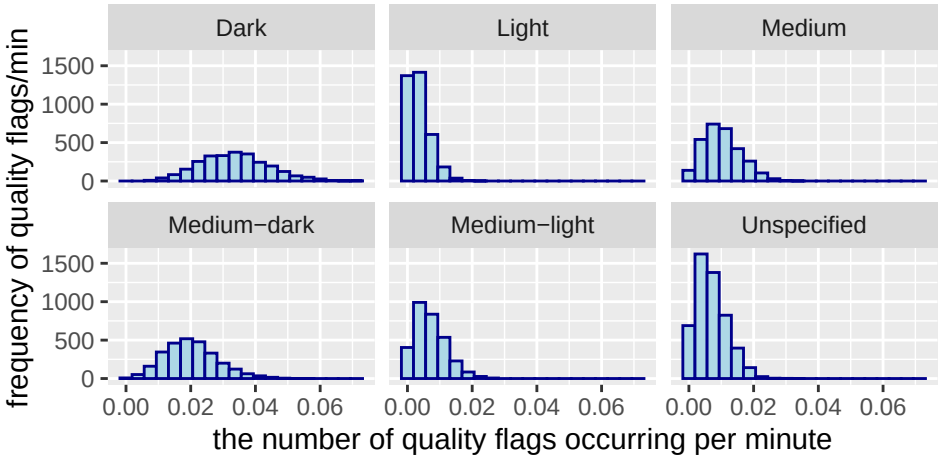

Figure 6: The histogram about the number of quality flags per minute during the sleep session in different skin color users

From Figure 6, each plot has only one mode and almost normally distributed. However, the range of the number of flags per minute on dark-skinned users were the largest, from about 0.007 to 0.063. The range for other skin color users were around 0 to 0.028, except for medium-dark users(approximately 0 to 0.05) and light-skinned users(approximately 0 to 0.015).

Table 6: The average number of flags per minute of devices for users with different skin colors

users' skin color	average flags per minute
Dark	0.0334151
Medium-dark	0.0202156
Medium	0.0099211
Medium-light	0.0066490
Unspecified	0.0065309
Light	0.0030657

Table 6 summarizes the average probability of flags per minute for the devices used by different skin color users. It shows that dark-skinned users have the highest number of flags per minute on average(0.0334) which means their devices have a higher probability of occurring flags during sleep sessions. Meanwhile, the average number of flags per minute occurring on light-skinned users is the smallest(around 0.0031) among six types. Therefore, according to Figure 6 and Table 6, we propose that the number of quality flags that occurred for a device during the sleep session would vary by user's skin color.

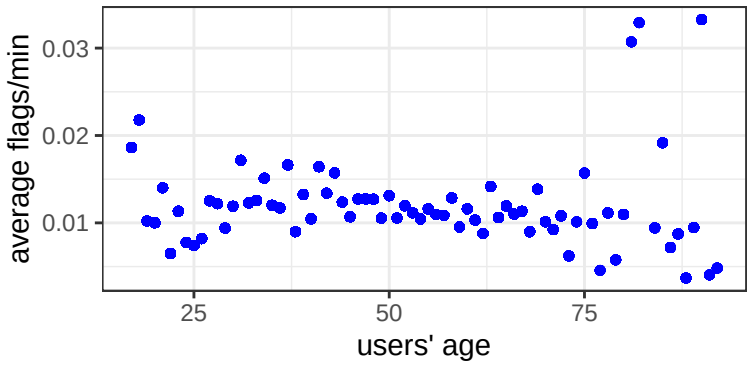

Figure 7: The scatterplot about the average number of quality flags per minute during the sleep session for users at different age levels

Figure 7 shows the distribution of user age and the average number of flags per unit of sleep session(minute). The average number of flags per minute for users aged less than 25 years fluctuated between 0.005 to 0.02. However, users over 75 years had the highest fluctuations in

their average number of flags per minute, reaching a maximum of about 0.037 and a minimum of less than 0.003. Overall, the average number of flags occurred per minute detected by the device decrease slightly when users' age increased. Thus, users' age would be a factor that leads to different numbers of quality flags during a sleep session.

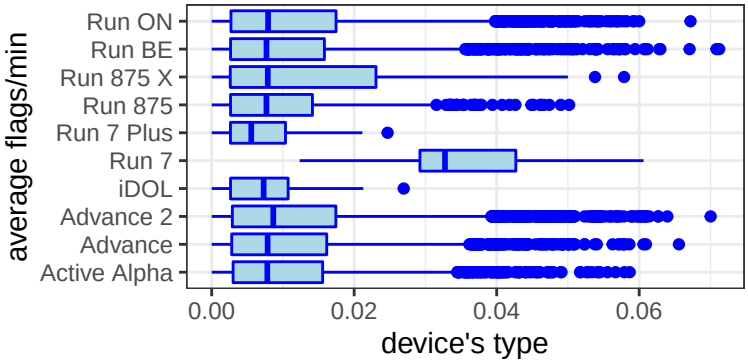

Figure 8: Boxplot about the number of flags per minute during the sleep session for different device types

Figure 8 presents the number of flags per minute during the sleep session for different devices. All devices have the minimum number of flags per minute is 0 and a similar median around 0.006, except for device Run 7. Patterns of device Active Alpha, Advance, and Run BE are similar, while Run BE has the largest outliers (approximately 0.072) among them. Thus, it is likely that different types of devices would have different numbers of quality flags occurring during sleep sessions.

Based on all of the figures and tables discussed above, the user's skin color, age, and the type of device they used would all have an impact on the number of quality flags during the sleep session. These three variables could be used as predictors when fitting the model later.

Method

For this research question, we would like to find out whether there is a relationship between the device's performance of the sleep score function and users' skin color. According to the interesting variables discussed in the Data section, the number of times the quality flag occurred during the sleep session for a user would be a good choice to assess the sleep score function for the device.

To discover the association between the sleep score performance of the device and the user's skin color, we would fit a model with the number of quality flags as the response variable and the user's skin color, age, and the device type as predictors. Meanwhile, each user has recorded the

number of quality flags in multiple sleep sessions, which implies that observations are dependent on each user. In this case, we chose to use a generalized linear mixed model (GLMM) with the customer ID representing a specific user as a random effect.

Since the number of quality flags is a countable data, then the response distribution would be the Poisson distribution, which is the family of distribution we used when fitting the model. The sleep session duration should be an offset since their values are quite different from each other when accounting for the number of quality flags for each observation. Also, we use users' scaled age instead of actual age so that it would be more efficient and computationally convenient when fitting the model.

Model Formula The mathematical formula for the generalized linear mixed model is

$$log(\mu_{ij}) = log(duration) + \beta_0 + \beta_1 skinColor_{ij} + \beta_2 ageScale_{ij} + \beta_3 device\_name_{ij} + U_j$$

and $flags_{ij} \sim Poisson(\mu_{ij})$

- $flags_{ij}$ is the number of times there was a quality flag during ith sleep session for jth user, which is the response variable and follows a Poisson distribution with $\lambda = \mu_{ij}$.
- $log(\mu_{ij})$ is the log count of the number of quality flags that occurred during ith sleep session for jth user. It uses the log link function for transformation. μ_{ij} is the number of quality flags that occurred during ith sleep session for jth user.
- U_j is the **random effect** for jth user and $U_j \sim N(0, \sigma^2)$
- $log(duration)$ is the offset term, which has a fixed coefficient estimate of 1.
- $skinColor_{ij}$ is the skin color of jth user for ith sleep session.
- $ageScale_{ij}$ is the scaled age of jth user for ith sleep session.
- $device\_name$ is the device type used by jth user for ith sleep session.

We would interpret the exponential of each β parameter as it is more meaningful.

- β_0 is the intercept term. The exponential of β_0 (e^{β_0}) represents the count of the number of occurrence of quality flags for a certain user whose skin color is at the reference level and who is 17 years old.
- the exponential of β_1 (e^{β_1}) represents the relative count for the group of users with skin color at the other level (not at the reference level) and the group of users with the skin color at the reference level, while keeping everything else fixed.
- the exponential of β_2 (e^{β_2}) represents the relative count when a user's scaled age increases by 1, which means the scaled age increases from 17 to 92, keeping everything else fixed.
- the exponential of β_3 (e^{β_3}) represents the relative count between the group of users' device type at the other level (not at the reference level) and the group of users' device type at

the reference level while keeping everything else fixed.

When we build the generalized linear mixed model, we need to check these four model assumptions, which are the same as the model assumptions discussed in the **Method** section in Research Question 1.

We would apply the following two model selection techniques (i.e., Hypothesis Testing, Likelihood-Ratio Test) that could be helpful for choosing the final model.

- Hypothesis Testing could statistically check whether our hypothesis meets the condition we state. It needs null hypothesis H_0 and alternative hypothesis H_a. Then calculate the test statistic from sample data and p-value under the assumption that H_0 is true, then using the common statistical threshold (0.05) as an indicator. Therefore, for each p-value, if it is below the threshold (< 0.05), then we would have strong evidence to reject the null hypothesis. Otherwise, we would accept the null hypothesis.
- The Likelihood-Ratio test is a hypothesis test, which is a statistical test of the fit between two models[10]. Specifically, a relatively complex model is compared with a simpler model to see which model fits a particular data set significantly better. We would use Hypothesis Testing: the null hypothesis(H_0) is that the simpler model fits the data as well as the complicated model; the alternative hypothesis(H_a) is that the simpler model does not fit the data as well as the complicated model. When the p-value is below the threshold, we reject the H_0 and consider using the complicated model. Otherwise, we prefer the simpler model.

Result

According to the GLMM we proposed in the **Method** section, we have the fitted model with fixed effects of users' skin color, age, and their device's type, a random intercept of a certain user, and an offset term of the duration.

Table 7: Key Numerical Outputs of the Fixed Effects of the Fitted Model

	Estimate	P-value
(Intercept)	-3.4083	0.0000
skin color light	-2.3902	0.0000
skin color medium	-1.2084	0.0000
skin color medium-dark	-0.4986	0.0000
skin color medium-light	-1.6153	0.0000
skin color unspecified	-1.6288	0.0000
scaled age	-0.0419	0.0187
device Advance	0.0444	0.0097
device Advance 2	0.0142	0.3121
device iDOL	0.0656	0.3389
device Run 7	0.0640	0.3953
device Run 7 Plus	-0.1113	0.2609
device Run 875	0.0431	0.1005
device Run 875 X	0.1344	0.0092
device Run BE	0.0231	0.1326
device Run ON	0.0288	0.0569

From Table 7, the p-values for the all skin colors are ($< 2e^{-16}$), close to 0, which are below the threshold, so we reject the null hypothesis that the corresponding coefficient parameter is 0, respectively. Thus, users' skin color is related to the number of flags that occurred during a sleep session. Similarly, for users' scaled age, the p-value (0.0187169) is below the threshold, and we concluded that it would have an impact on the number of quality flags. The p-values for some device types are above the threshold, so it may not be quite related to the number of flags during a sleep session for the user's device. We may consider not including it in the model. Therefore, we fitted a reduced model, which is exactly the same as the full model, except for it does not have the predictor, the user's device type.

Table 8: Numerical Outputs of Likelihood-ratio Test

Degrees of Freedom	Log Lik	Df	Chi Square	P-value
8	-42261.64			
17	-42253.20	9	16.8799	0.0506

Now, we can apply Likelihood-ratio test for these two nested models. From Table 8, the p-value we got, under the assumption that the null hypothesis is true, is 0.0506305. Thus, we fail to

reject the null hypothesis and conclude that the reduced model explains the data as well as the more complicated model. We preferred the reduced model and used it as our final model.

Table 9: Key Numerical Outputs of the Fixed Effects of the Reduced Model

	Estimate	P-value
(Intercept)	-3.3826	0.0000
skin color light	-2.3892	0.0000
skin color medium	-1.2121	0.0000
skin color medium-dark	-0.5003	0.0000
skin color medium-light	-1.6130	0.0000
skin color unspecified	-1.6314	0.0000
scaled age	-0.0479	0.0071

Based on the values in Table 9, the final fitted generalized linear mixed model is

$log(\hat{\mu}_{ij}) = log(duration) + \hat{\beta}_0 + \hat{\beta}_1 skinColorLight_{ij} + \hat{\beta}_2 skinColorMedium_{ij} + \hat{\beta}_3 skinColorMedium\text{-}dark_{ij} + \hat{\beta}_4 skinColorMedium\text{-}light_{ij} + \hat{\beta}_5 skinColorUnspecified_{ij} + \hat{\beta}_6 ageScale_{ij} + U_j$

and $\hat{flags}_{ij} \sim Poisson(\hat{\mu}_{ij})$.

- $\hat{flags}_{ij}$ is the estimated number of times there was a quality flag during ith sleep session for jth user, which is the response and follows a Poisson distribution with $\lambda = \hat{\mu}_{ij}$.
- $log(\hat{\mu}_{ij})$ is the expected log count of the number of quality flags that occurred during ith sleep session for jth user. It uses the log link function for transformation. $\hat{\mu}_{ij}$ is the expected number of quality flags that occurred during ith sleep session for jth user.
- U_j is the **random effect** for jth user and $U_j \sim N(0, \sigma^2)$
- $log(duration)$ is the offset term, which has a fixed coefficient estimate of 1.
- $skinColor_{ij}$: The model uses dummy variables with dark skin color as the reference level and the other 5 levels as five indicator variables whose value would be 1 if jth user has that skin color for ith sleep session. Otherwise, the value would be 0.
- $ageScale_{ij}$ is the scaled age of jth user for ith sleep session.
- $\hat{\beta}_0$: -3.3826472 is the intercept term. The exponential of $\hat{\beta}_0$ ($e^{\hat{\beta}_0}$) is 0.0339574, which represents the value of the expected count of the number of flags for a certain user whose skin color is dark and who is 17 years old (minimum customer's age level).
- the exponential of $\hat{\beta}_1$ ($e^{\hat{\beta}_1}$): 0.0917021 represents the expected relative count for the group of users with light skin color and the group of users with dark skin color. The expected number of quality flags for a user with light skin would be 9.1702076% times as many as dark skin user while keeping everything else fixed.

- the exponential of $\hat{\beta}_2$ (0.2975803; medium skin color), $\hat{\beta}_3$ (0.6063743; medium-dark skin color), $\hat{\beta}_4$ (0.1992876; medium-light skin color), $\hat{\beta}_5$ (0.1956492; unspecified skin color) would be similar as the above interpretation of the exponential of $\hat{\beta}_1$ except for different skin color levels and values.
- the exponential of $\hat{\beta}_6$ ($e^{\hat{\beta}_6}$): 0.9532721 represents the expected relative count when a user's scaled age when it increases by 1. When the user's scaled age increases from 17 (minimum age level) to 92 (maximum age level), the expected count of the number of quality flags decreases by 4.6727903%, while keeping everything else fixed. In other words, when the user's age increases by 1, the expected count of the number of quality flags would occur during a sleep session decreases by 0.0637862%, while keeping everything else fixed.

Then, we would assess the four model assumptions for the final model:

- The grouping unit is the customer/user, which is independent. This has been discussed in the **Method** section.
- The second assumption, the random effect(customer ID) comes from a Normal distribution, would be hard to assess, which will be discussed in the **Strength and Limitation** section later.
- We checked the residual plot with respect to the fitted values and find no obvious pattern, so the errors from random effect and within-unit residual errors have constant variance.
- Since the number of quality flags for a sleep session is a count data, the response distribution we have chosen is the Poisson distribution. The log function is the link function, which would be appropriate for it. The fourth assumption is satisfied.

Now, we could proceed to discuss more about the estimated coefficients from the final model. From Table 9, since these p-values are all below the threshold, then we choose to keep all predictors in the model as they are all related to the response.

Table 10: 95 percent Exponential CI in Final Model

	Lower Bound 2.5%	Higher Bound 97.5%
(Intercept)	0.0333233	0.0346036
skin color light	0.0885992	0.0949136
skin color medium	0.2901507	0.3052002
skin color medium-dark	0.5933828	0.6196502
skin color medium-light	0.1938453	0.2048826
skin color unspecified	0.1910984	0.2003084
scaled age	0.9206373	0.9870638

Table 10 summarizes the exponential of the 95%-confidence interval for each β parameter.For

exponential of β_1, the exponential 95%-confidence interval is (0.0885992, 0.0949136), then we are 95% certain that the number of quality flags for a light-skinned user is 8.8599182% to 9.4913639% times as many as dark-skinned user while keeping everything else constant. The 95%-confidence interval for β_6 is (0.9206373, 0.9870638), then we are 95% certain that when the user's age increases by 1, the count of the number of quality flags decreases by 0.1101915% to 0.0173593% times while keeping everything else fixed. Basically, these 95%-confidence intervals are all small, so the estimated coefficients we fitted for the final model have high precision and the final model would be appropriate.

Conclusion In conclusion, the average number of quality flags that occurred during the sleep session would be the highest for the dark-skinned users. For users of the same age, the average number of quality flags for users with medium-dark, medium, medium-light, or light skin color would be 39.3625688%, 70.2419677%, 80.0712442%, or 90.8297924% less than the average number of quality flags for a dark-skinned user respectively. For each of the five levels of user's skin color, since they are all compared with the same base level (darker-skinned users), then we conclude that the quality flag would occur more frequently for dark-skinned users. For the group of users whose skin color is unspecified, it is not meaningful to draw any conclusion by its estimated coefficient as we do not have enough information. Also, the average number of quality flags would decrease by 0.0637862% for users with the same skin color if age increases by one. Thus, the quality flag during the sleep session would occur more frequently for younger users.

Therefore, the occurrence of the quality flag would be higher for younger users whose skin color is darker. It is reasonable to conclude that the user's skin color would have an effect on the performance of the functionality of the device's sleep score, and for a younger and darker-skinned user, his/her device would perform poorly for sleep score function.

Discussion

In order to get more insights to answer the question about the difference between the customers who purchase the Active and Advance products and the traditional customers who buy traditional products, we firstly compared the difference in feature preferences and price sensitivities between Active or Advance buyer and traditional buyers and found out that Active/Advance buyers are the people who cares mostly about the low price, the latest version product in the market, and whether the device has GPS. However, traditional product buyers are not price-sensitive but care about whether the device has the best and most powerful functions for outdoor activities. Secondly, we fitted a Generalized Linear Mixed Model and concluded that at a given neighborhood and a given level of household income, the older people have more odds to buy the Active/Advance product than younger people. Also, at a given age and a given neighborhood, people who come

from a higher income household are more likely to buy the Active/Advance product.

For the second research question, we are interested in whether darker skin users affects the performance of the device's sleep score function, or there are some additional effects that will influence the sleep score. We firstly constructed several graphical and numerical summaries to demonstrate the relationship between the number of quality flags occurred during the sleep session and the customer's skin color and other related variables. Then, we built a Generalized Linear Mixed Model and concluded that darker-skinned users are more likely to occur quality flags during sleep sessions. Simultaneously, users' age would have some effect on the sleep score function, but it is not significant. Research states that it is more difficult to obtain accurate data since melanin-rich skin blocks green light which some device companies will use in their sensor[12]. Therefore, the conclusions we use the model to analyze are reasonable and it is consistent with those in the research.

Strengths and limitations

In this section, we would like to point out a few pitfalls and advantages in our models for the first research question.

When fitting the generalized linear mixed model to find the difference between the Active/Advance buyers and traditional buyers, the model suffers from two limitations.

- Given the data at hand, the model suffers from omitted variable bias. For example, we concluded that, at a given age and neighborhood, people with higher income are more likely to purchase Active/Advance product. However, it could potentially be the case that people who exercise frequently are the ones who purchase this product and in the meantime these people are from a relatively rich household. Thus, if possible, further analysis is required and more detailed data should be collected.
- Skin color may have a statistically significant effect on being a new customer. In the provided data, most customers do not use customized emoji color, thus, we cannot impute most customers' skin color and this may potentially cause bias to the model.

In terms of advantages, first of all, our model carefully considers the effect of neighborhood on the probability of being a new customer. We include the Census SubDivision Unique Identifier to account for this random effect in our model. Secondly, we assessed the difference between two groups of customers from two different dimensions through different methods. We first concluded the preference with respect to device features and price of two groups of customers and then utilized an advanced statistical model (Generalized Linear Mixed Model) to model the characteristics of new customers. The results from two methods are aligned with each other. This increases the credibility of the analysis.

Then, we would like to summarize some strengths and limitations of the second research question.

For the strengths, firstly, by using various types of figures and summarized tables, we explored the users' skin color and other variables of interest from the sleep dataset to discover variables that would be correlated with the occurrence of quality flags during the sleep session. Secondly, in the Method and Result sections, the choice of a GLMM with Poisson distribution for the response variable is appropriate. Thirdly, according to Table 10, the p-values of the estimated coefficients in the final model are relatively small, which reveals that it is appropriate and reasonable to include these predictors in the final model from a statistical perspective.

There are three limitations shown as follows.

- Since one group of skin color classified as "Unspecified" because the user did not specifically set an emoji modifier and used the default state directly, some users' information were not successfully obtained, which may affect the model selection and conclusions. In addition, users may have a preference for the emoji modifier, so there may be some bias in inferring the user's skin color from the emoji modifier.
- Although the final model included users' scaled age as one of the predictors, its influence on the average number of quality flags occurred during sleep session was extremely small even if its p-value is smaller than the threshold. Further analysis on whether to include the predictor of the user's scaled age for the model or not may be conducted in future.
- When checking the second assumption of the final model in the Result section, due to the limitation of the scope, it is hard for us to assess it in the current stage. For the third assumption, there is actually some pattern for small fitted values. It may be necessary to further investigate it in future. Besides, since the model we chose is a GLMM with Poisson distribution, then it would be better to check overdispersion. This may be another limitation and we may use some other external packages to check overdispersion to decide whether we need to make some changes to the model in future.

Consultant information

Consultant profiles

Weilin Alex Yin. Alex is a senior Data Scientist in Mega Consulting Company. He specializes in data analysis, statistical modelling, data visualization (e.g. Seaborn, Matplotlib, ggplot, Tableau), and programming (e.g. Python, SAS, R, SQL, STATA). Alex earned her Bachelor of Science, majoring in Statistics and Economics, from the University of Toronto in 2023.

Kexin Selina Sha. Selina is a senior Data Scientist in Mega Consulting Company. She specializes in statistical communication, statistical data analysis, database design, algorithms, machine learning, and programming (Java, Python, JavaScript, SQL, R, C/C++). She earned their Bachelor of Science, Majoring in Computer Science and Statistics from the University of Toronto in 2024.

Qing Lyu. Qing is a junior Data Scientist. She specializes in machine learning, statistical analysis, data wrangling and visualization. Qing earned her Bachelor of Science, Majoring in Computer Science and Statistics from the University of Toronto in 2023.

Xuening Bai. Xuening is a junior consultant with Data Analytics. She specializes in data visualization and analysis. Xuening earned her Bachelor of Science, Majoring in Statistics and Economics from the University of Toronto in 2023.

Code of ethical conduct

The process of data collection and storage and the analytic results will be protected by relevant privacy laws. The private account information (i.e., API key) for accessing the census data will always keep secret.

Additionally, all the processes and analyses are objective in order to eliminate racial problems and bias. The information or client data about the employer and client should not to be distributed without their consent. Moreover, assumptions about data collection and analysis and limitations of the results need to be adequately presented. Consultants need to fully demonstrate statistical models and methods to analyze results within their capabilities[21].

Besides, we guarantee the consultants strictly follow the terms in the Non-Disclosure Agreement and use project code instead of the actual name of the client to avoid potential information leaking[22].

References

[1] 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/.

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

[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] Hadley Wickham (2021). rvest: Easily Harvest (Scrape) Web Pages. R package version 1.0.1. https://CRAN.R-project.org/package=rvest

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

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

[7] Hao Zhu (2021). kableExtra: Construct Complex Table with "kable" and Pipe Syntax. R package version 1.3.4. https://CRAN.R-project.org/package=kableExtra

[8] John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

[9] Revelle, W. (2022) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 2.2.3,.

[10] Glen, S. (2021, August 1). Likelihood-Ratio Tests (Probability and Mathematical Statistics). Statistics How To. Retrieved April 7, 2022, from https://www.statisticshowto.com/likelihood-ratio-tests/

[11] Glen, S. (2021, June 3). P-Value in Statistical Hypothesis Tests: What is it? Statistics How To. Retrieved April 7, 2022, from https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/p-value/

[12] Hailu, R. (2019, July 24). Fitbits and other wearables may not accurately track heart rates in people of color. STAT. Retrieved April 7, 2022, from https://www.statnews.com/2019/07/24/fitbit-accuracy-dark-skin/

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

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

[15] von Bergmann, J., Dmitry Shkolnik, and Aaron Jacobs (2021). cancensus: R package to access, retrieve, and work with Canadian Census data and geography. v0.4.2.

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

[17] Full Emoji Modifier Sequences, v14.0. (n.d.). UNICODE. Retrieved April 7, 2022, from https://unicode.org/emoji/charts/full-emoji-modifiers.html

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

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

[20] Postal code conversion file: 2016 census geography. Postal code conversion file: 2016 census geography | Map and Data Library. (n.d.). Retrieved April 7, 2022, from https://mdl.library.utoronto.ca/collections/numeric-data/census-canada/postal-code-conversion-file/2016

[21] Code of ethical statistical practice. Statistical Society of Canada. (n.d.). Retrieved April 7, 2022, from https://ssc.ca/sites/default/files/data/Members/public/Accreditation/ethics_e.pdf

[22] Sauder, L. (n.d.). Protecting Client Confidentiality | Consulting And Professional Services Radio. Consulting & Professional Service Ratio. Retrieved April 7, 2022, from https://cpsradio.com/protecting-client-confidentiality/

Appendix

Web scraping industry data on fitness tracker devices

Firstly, since this website does not have a public API that provides data, then we could not use it. Secondly, since we could not find the Terms and Conditions for this site, then we looked up the robots.txt, which did not disallow the page we visited to access the data. Then, when web scraping the industry data, we provided a Use Agent string containing our email address to be contacted and our intention clearly. Simultaneously, we requested data at a reasonable rate with a crawl limit of 12 seconds, which was suggested by robots.txt. Eventually, we scraped the industry data on fitness tracker devices using R.

```
# loading required libraries
library(tidyverse)
library(polite)
library(rvest)

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

# provide informative user_agent details to introduce myself to the host
target <- bow(url,
              user_agent = "alex.yin@mail.utoronto.ca for STA303/1002 Final project",
              force = TRUE)

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

```
## <polite session> https://fitnesstrackerinfohub.netlify.app/
##     User-agent: alex.yin@mail.utoronto.ca for STA303/1002 Final project
##     robots.txt: 2 rules are defined for 2 bots
##    Crawl delay: 12 sec
##   The path is scrapable for this user-agent
```

```
# scrape the content of the authorized page
html <- scrape(target)

# output the scraping content and store as a device_data dataset
device_data <- html %>%
  html_elements("table") %>%
```

```
  html_table() %>%
  pluck(1) # added, in case getting a list format
```

Accessing Census data on median household income

The Canadian census website does not state the Term and Conditions on what we can and cannot do. However, the site contains a file called robots.txt, which is about what "robots" are allowed or not allowed to access this site. By checking the conditions, it allows us to access the data through the website. In this way, we use the public API provided to access the data. 5 seconds is used as the default state since there is no rate limit request for our access. In addition, when accessing the page we should register a unique API key.

```
# loading required library
library(cancensus)

# sets the cancensus.api_key using the registered API key to use public API to
# access data
options(cancensus.api_key = "< The API Key goes here>",
        cancensus.cache_path = "cache") # sets a folder for local 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()

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

Accessing postcode conversion files

There is a license agreement including Terms of Use for the postcode conversion files for use by the University of Toronto Faculty, Students, and Staff. Also, the robots.txt for this site did not disallow the page we visited. Then, we accepted the license agreement to uphold the Terms of Use and logged in using our student account to get the data as a University of Toronto student. One thing to notice is that we chose the postcode conversion files of the 2016 census data. The reason is that we used the 2016 census data for median household income data, so we would like to have these two datasets matching for the same year.

```
# loading required library
library(haven)
library(tidyverse)

# read the downloaded sav file from data-raw folder
dataset = read_rds("data-raw/break_glass_in_case_of_emergency.Rds")

# selected only variables: PC, CSDuid
postcode <- dataset %>%
  select(PC, CSDuid)
```

Table Appendix

Table 11: Industry data for Mingar products

Device name	Line	Price	Battery life	Water resitance	Pulse oximiter	GPS	Sleep tracking	Released	Heart-rate sensor	Notifications	No-contact pay
Run ON	Run	349.99	Up to 21 days	Waterproof, 10 ATM	Y	Y	Y	2021-12-04	Y	Y	Y
Run BE	Run	299.99	Up to 14 days	Waterproof, 10 ATM	Y	Y	Y	2020-11-20	Y	Y	Y
Run 875	Run	350.00	Up to 14 days	Waterproof, 5 ATM	N	Y	Y	2019-09-12	Y	Y	Y
Run 875 X	Run	399.99	Up to 14 days	Waterproof, 5 ATM	Y	Y	Y	2019-09-12	Y	Y	Y
Run 7	Run	399.99	Up to 14 days	Waterproof, 5 ATM	N	Y	Y	2018-03-09	Y	Y	Y
Run 7 Plus	Run	435.00	Up to 14 days	Waterproof, 5 ATM	Y	Y	Y	2018-03-09	Y	Y	Y
Run HYYH	Run	420.00	Up to 7 days	Waterproof, 5 ATM	N	Y	N	2017-02-18	Y	Y	N
Run Leader	Run	479.99	Up to 7 days	Waterproof, 5 ATM	N	Y	N	2016-09-12	Y	Y	N
Run	Run	450.00	Up to 5 days	Waterproof, 5 ATM	N	N	N	2015-08-01	Y	N	N
iDOL	iDOL	199.99	Up to 14 days	Waterproof, 10 ATM	N	Y	Y	2018-08-24	Y	Y	Y
Advance 2	Advance	145.00	Up to 7 days	Resistant	N	Y	Y	2021-07-08	Y	Y	Y
Advance	Advance	120.00	Up to 7 days	Resistant	N	Y	Y	2020-08-20	Y	Y	Y
Active Alpha	Active	99.99	Up to 7 days	Resistant	N	N	Y	2020-12-30	Y	Y	Y
Active	Active	39.99	Up to 14 days	Resistant	N	N	N	2019-10-13	N	N	N
Active HR	Active	79.99	Up to 7 days	Resistant	N	N	N	2019-10-13	Y	N	N