Strategy Analysis of MINGAR's New Product Lines and Device Performances

Broader age range and lower income new customers, improved device performance for darker skin users

Report prepared for MINGAR by Live Infinitely

2022-04-07

Contents

Executive Summary

Mingar seeks guidance and data to support the strategy development of two new product lines, "Active" and "Advance," to increase Mingar's market share in the growing wearables market. Besides, Mingar seeks assistance in solving the complaints that the devices perform poorly for users with darker skin. Thus, this report aims to analyze Mingar's marketing, social media, and product considerations of the new "Active" and "Advance" product lines.

We assess that **(1)**customers containing three characteristics, age between 30-60, female and lower-income level, respectively, are key marketing targets for our new product. This report not only gives the key characteristics of the consumers for the new product, but also compares the differences between the company's traditional customers and the new product customers. **(2)**The products of MINGAR company have poor performance on sleep scores for customers with different skin tones. Customers with darker skin color face the problem of more frequent quality marks during a sleep session.

The results of the study are summarized below.

- Based on Table.1, among the new customers who bought "Active" and "Advance" products, female and middle-aged(30-60) customers accounted for a larger proportion.
- The differences in income and age group distribution between old and new customers who bought "Active" and "Advance" products are obvious, while people with different gender and skin color do not differ much in their probability of purchasing the new products.
- The estimated average income(Canadian dollars) of new customers who purchase "Active" and "Advance" is 68787.4, which is 4651.2 lower than traditional customers.
- Based on our prediction model, every time a customer's income level increases by 1000, their odds[1] of purchasing "Active" and "Advance" decrease by 2.15%, holding other conditions constant.
- The estimated *odds*[1] of buying the new "Active" and "Advance" products for non-middle-aged customers is 201% higher than the *odds*[1] for middle-aged customers, assuming other variables keep constant.
- We are 95% confident that the true *odds*[1] of buying the new "Active" and "Advance" products are 183% to 220% higher than the *odds*[1] for middle-aged customers, assuming other variables keep constant.
- Refer to Table.2, customers with darker skin are more likely to experience more device flags, where flags refer to the occurrences of bugs due to missing data or due to data being recorded but sufficiently unusual.
- The expected average number of flags for people with dark skin color is 0.03339.

[1] Odds calculated as the ratio of the number of events that produce that outcome to the number that do not.

- There exists a 91% reduction in the estimated average number of flags for people with light skin color compared to people with dark skin color.
- However, the accuracy of some factors is limited due to privacy and ethnicity. The analysis uses the median income level of the customer's neighborhood as the individual income. We used the color of emoji MINGAR's customers frequently use to represent their skin colors, but the choices of emoji color depend on personal preference instead of their actual skin colors.

Overall, compared with MINGAR's traditional customers, the more affordable "Active" and "Advance" lines appeal to a lower income level group and non-middle-aged customers. Also, MINGAR should seriously deal with the problem of poor device performance for users with darker skin.

Key results of the study are summarized in the following tables.

Summary Table 1. Comparisons among the age, income(in Canadian dollars), and sex distributions of new customers and traditional customers

Customers group	Count	Average age	Middle-aged Proportion	Average Income	Proportion Female
New	8476	46.25307	0.7282916	73168.02	0.5864795
Traditional	10569	47.70461	0.4776232	68813.94	0.5781058

Summary Table 2. The Number of Flags Per Duration of Different Skin Color Comparison

skin color	Mean of Flags Per Duration
Dark	0.0334690
Light	0.0030644
Medium	0.0099200
Medium-Dark	0.0202633
Medium-Light	0.0066273
Not given	0.0065263

Technical Report

Introduction: Research Background

From our client company Mingar's standpoint, its product positioning ranges from a maritime/military GPS device to a fitness tracking wearable that goes to the high-end consumer market. In order to increase the market share of fitness-tracking wearables, the client company Mingar has expanded the price positioning of fitness-tracking wearables, to target the demand of low-income customer groups. On behalf of Live Infinitely, we have compiled a statistical report on the marketing and product considerations of the client company Mingar, using data collected from the client company and the extranet. The subsequent technical report is divided into the following general sections: Method, Result, and Conclusion. The Method section summarizes the data and introduces all the methods used in this report, the Result section describes the modeling results, and the conclusion session provides an overall discussion and identifies some limitations.

Introduction: Research Questions

The report addressed two issues in total:

- First, we aim to provide the client's marketing team with data analysis of the new "Active" and "Advance" product's audience and the difference between the new and old product's customers. To serve the needs of our client companies, we also focused on analyzing the disparity between the income levels of the old and new client groups.
- The other was to provide the social media team with an analysis of the trends in complaints that the team was tracking, focusing on the differences in devices performances by the skin color of the customers, to help improve the users' experiences of Mingar's devices.

Data Wrangling and Data descriptions

Data Collection

We initially obtained the cust_dev, cust_sleep, customer, and device datasets from our client, which provides us with the data of Mingar customers, including their device ID, general profile information, and their sleep data. Through website scraping, we obtained some external website databases from some authoritative websites, including the industry information of the devices and the median income levels of Canadian neighborhoods from the Canadian Census. Since the database is very large, data cleaning and integration is an essential step for our analysis. Detailed information about our data collection process will be shown in the Appendix section.

Data Cleaning

In general, we have organized 3 data sets to facilitate our subsequent data processing from a raw-data folder.

- For the customer dataset, firstly, we web scaping the 2016 census data through the Census Mapper API to collect the average income level and postcode of the customers' neighborhoods. We merged the census data with the customer dataset by the 2016 postcode conversion file to obtain the customer dataset which includes general profiles data of the customers along with their neighborhood's median income. We also paired the specific devices used by the customers with their device ID. Then we processed the date of birth of the customer, into the age of the customer. Since information on the race of the customers is not provided, we determine the skin color of the customer by their choices of the emoji color in their chat setting. In order to facilitate the subsequent data modeling without the problem of convergence, we converted all the non-numerical variables in the dataset into the form of a factor, and divided all income data by 1000 to rescale the income into units of thousands. To compare new customers attracted by the "Active" and "Advance" lines with the traditional customers, we also mutate a new variable to indicate whether the customers bought the new products. The customer group using the new products "Active" and "Advance" will be indicated as 1, and 0 otherwise. Also, based on the age of the customers, we divided them into middle-aged(30-60) and non-middle-aged.
- The device_data_study dataset includes the devices information of both Mingar and Bitfit in the wearables market. This dataset gives us an intuitive analysis of the data purely from a product perspective. We mutate a new column named *New Line*, which is intended to split our product line into Mingar's new products("Active and 'Advance"), Mingar's old products, and our competitor's products(Bitfit). We mutated new variables to convert all product lines that indicate a feature as "Yes" to 1 and "No" to 0, and converted them all to numerical form. We added up the numerical data for each row horizontally, so that we could mutate a new column named "Number of functions" which represents how many kinds of functions each different device will have (such as GPS functions, etc.)
- Finally, for the cust_sleep dataset, we aim to investigate the problem of poor device performance for darker skin customers. Thus, variables related to the number of device failures, the sleeping duration, and the customer's skin color were selected for this dataset. This dataset also helps to facilitate our analysis of the customers' preferences on devices from the perspective of the products they choose, and to facilitate subsequent data summary. We merged this dataset with the original cust_sleep dataset through cust_id for modeling. We also changed all non-numeric variables to factors in order to facilitate subsequent data modeling without converge problems.

  For all three datasets, we've removed the variables that were unnecessary in our data analysis and only kept information that can be disclosed to the public. Additional information on the collections of external datasets will be discussed in the Appendix section.

Variable Description

Table.1 Table of detailed descriptions of variables used in the analysis

Variable	Description	Types
cust_id	Unique ID for each customer	Chr
sex	Biological sex, used for calculations of health metrics, like body-mass index and base metabolic rate.	Factor
income	The median income of the customer's neighborhood in Canadian dollars (unit: thousand)	Factor
line	Line of products this device belongs to	Factor
age	Age of customers	numerical
age_range	Middle-aged(aged between 30-60) customers and non-middle-aged customers	Binary
duration	Duration, in minutes, of sleep session	numerical
flags	Number of times there was a quality flag during the sleep session	numerical
skin_color	Skin color of each customer	Factor
new_target	Customers who buy our "Active" and "Advance" products and customers who buy our old products	Binary

After data cleaning, a total of 3 datasets were obtained. The cust_sleep dataset contains 21233 observations and 10 variables, the customer dataset contains 19045 observations and 10 variables, and the device_data dataset contains 26 observations and 15 variables. As we want to study the audience of the new product and further study the difference between its audience and the customers of the old product. We selected gender, age, age range, income, and whether the customer bought the new product (new target) as the interest variable. Secondly, in order to track the complaint that our device does not perform well for dark-skinned customers, we chose flags, duration, and skin color as the interest variables.

In this analysis, we used the median income level(in thousands) of the customer's neighborhood to indicate the income level of that customer. Also, it is important to note that we determined the skin color of the customer by their choices of the emoji color in their chat setting. These are only rough estimations given the data we have, and may introduce some bias into this study. Further discussion about this limitation will be shown in the Conclusion section.

Visualizations of the characterisctics of wearables market and customers group

Age Groups and Sex Groups Distributions of Mingar's New Customers

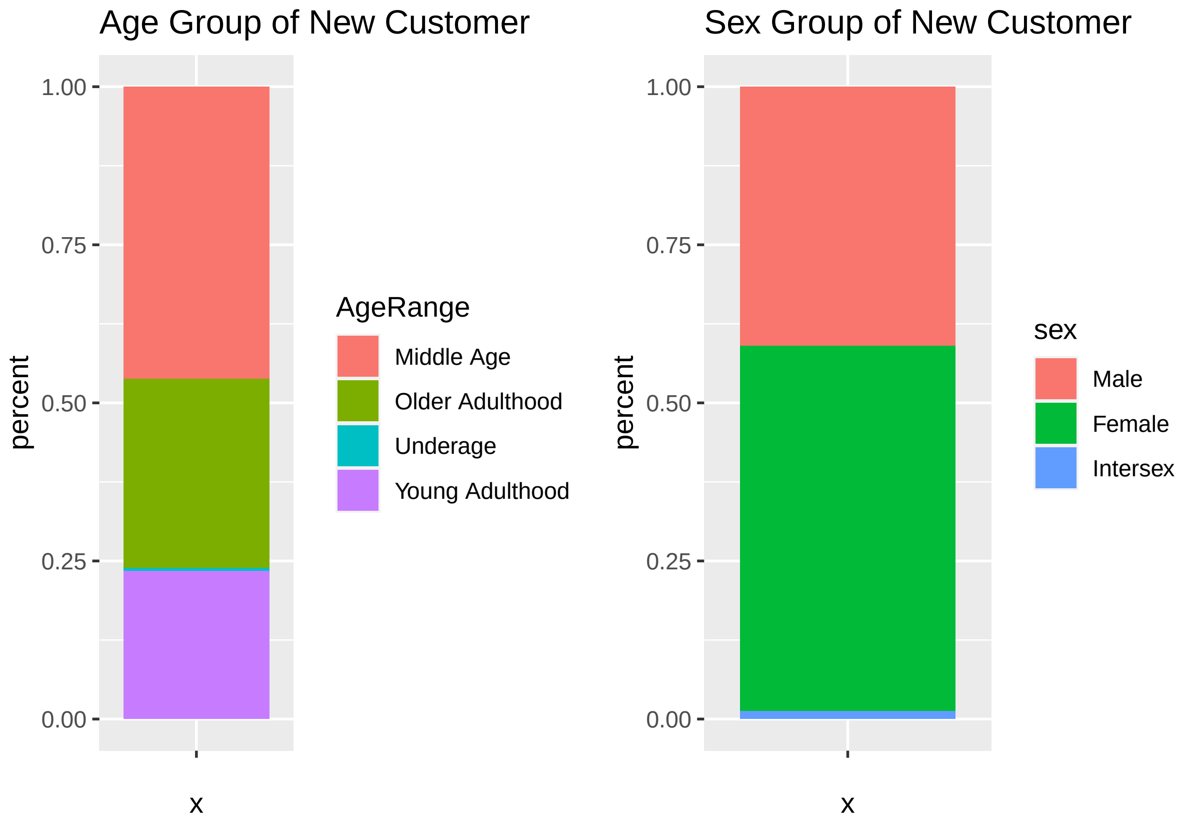

Figure 1: Age and Sex Group Pie Chart

According to the information we have accessed, we define people under 18 years old as underage, people between 18 and 30 years old as young adulthood, people between 30 and 60 years old as middle age, and people over 60 years old as older adulthood[1]. Based on Figure.1, the new product has the most middle-aged audience(45%), followed by young adulthood(24%) and older adulthood(30%). According to the histogram of gender, we can conclude that female customers take up a larger portion(60%) of the new customers than male customers do (39%). Intersex customers account for the smallest proportion(1%), but we cannot ignore the fact that the group size of intersex itself in society is small. In summary, according to the above graph, we can clearly understand that our new product's customer base is largely focused on women and the middle-aged.

Compare the New/Old Products' Customer Groups Regarding Income and Gender

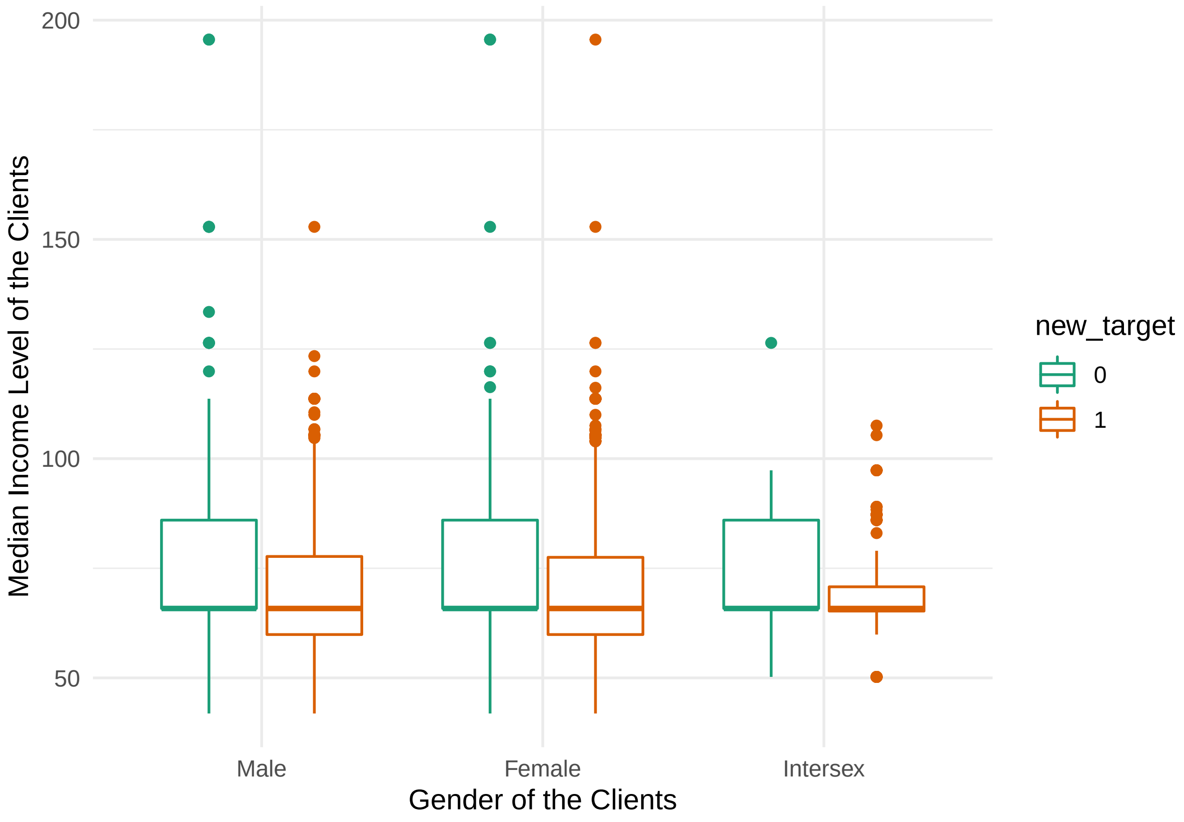

Figure 2: Comparison of New/Old Products' Customer Groups Regarding Income and Gender

Compares the new customers who bought "Active" and "Advance" products with other customers who bought traditional Mingar products regarding their income level and sex. The group of customers who bought the new "Active" and "Advance" products will be indicated as the number "1" in new target, which is shown in orange. On the contrary, the group of customers who buy other products of Mingar is set to the number "0" in new target, which is shown as green. In general, the median income of all customer groups, regardless of gender, is generally right-skewed distributed. Comparing the customer groups who bought the new product with those who bought the old one, we found almost no difference in their median income level. However, the first quartile of the income of the old product group almost overlaps with its median, as can be seen from the boxplot. In contrast, only the first quartile of the median income of customers with gender intersex overlaps with their median for the new product group. We found that although the median income of the customer group did not change much, the overall IQR of the box plot for the new customer group is lower than the IQR of the box plot for the old customer group, indicating that among all sex groups, the income of most of the new customers distributed at a relatively lower level than that of Mingar's old customers.

Compare the Difference in Skin Color between Old and New Customers

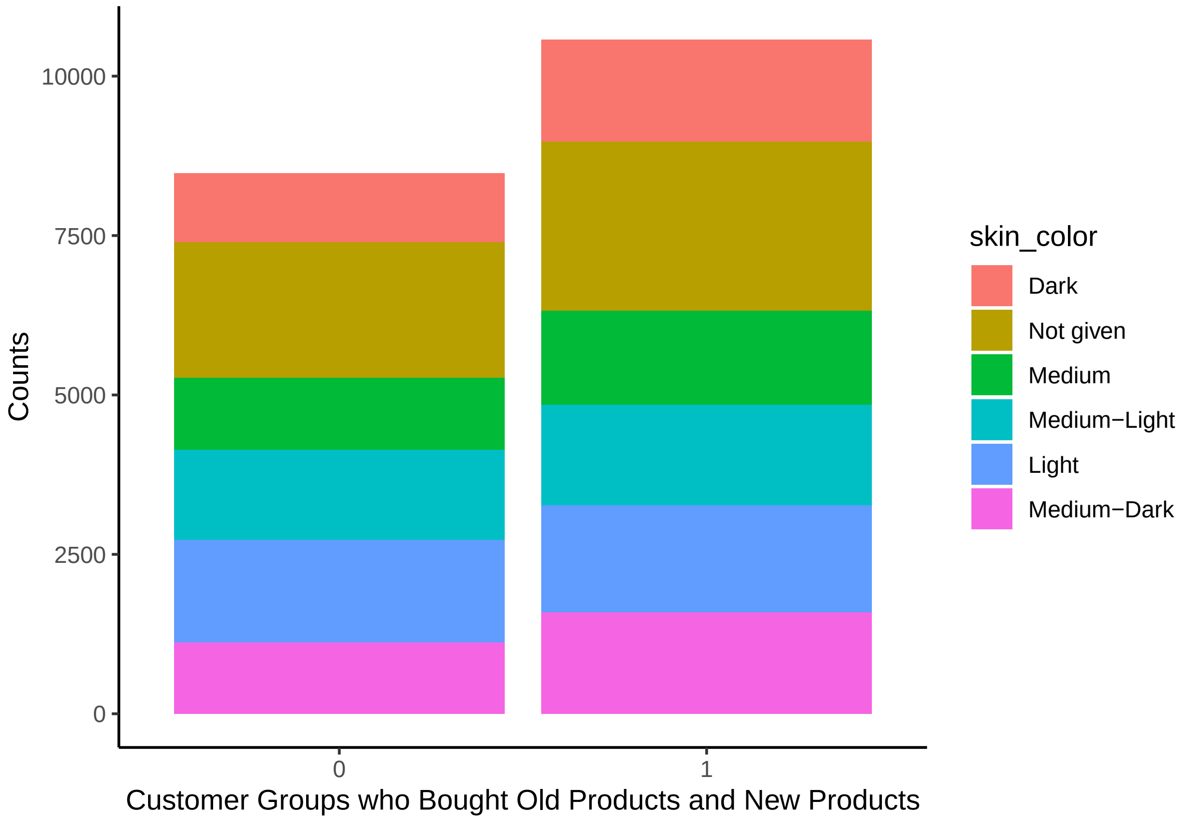

Figure 3: Compare the Difference in Skin Color and the Number of Products between Customers who Bought New Products and Old Products

For Figure.3, the x-axis represents the group of customers who bought the old products and the group of customers who bought the new products ("Active" and "Advance"), which are represented by "0" and "1", respectively. We find that the number of customers who buy new products is about 11,000, which is relatively larger than the number of customers who buy old products, which is about 9200. From the figure, the skin color distribution of both customer groups is fairly even. A closer look at the graph shows that the group of customers with darker skin color may have a slightly higher preference for the new product than for the old one.

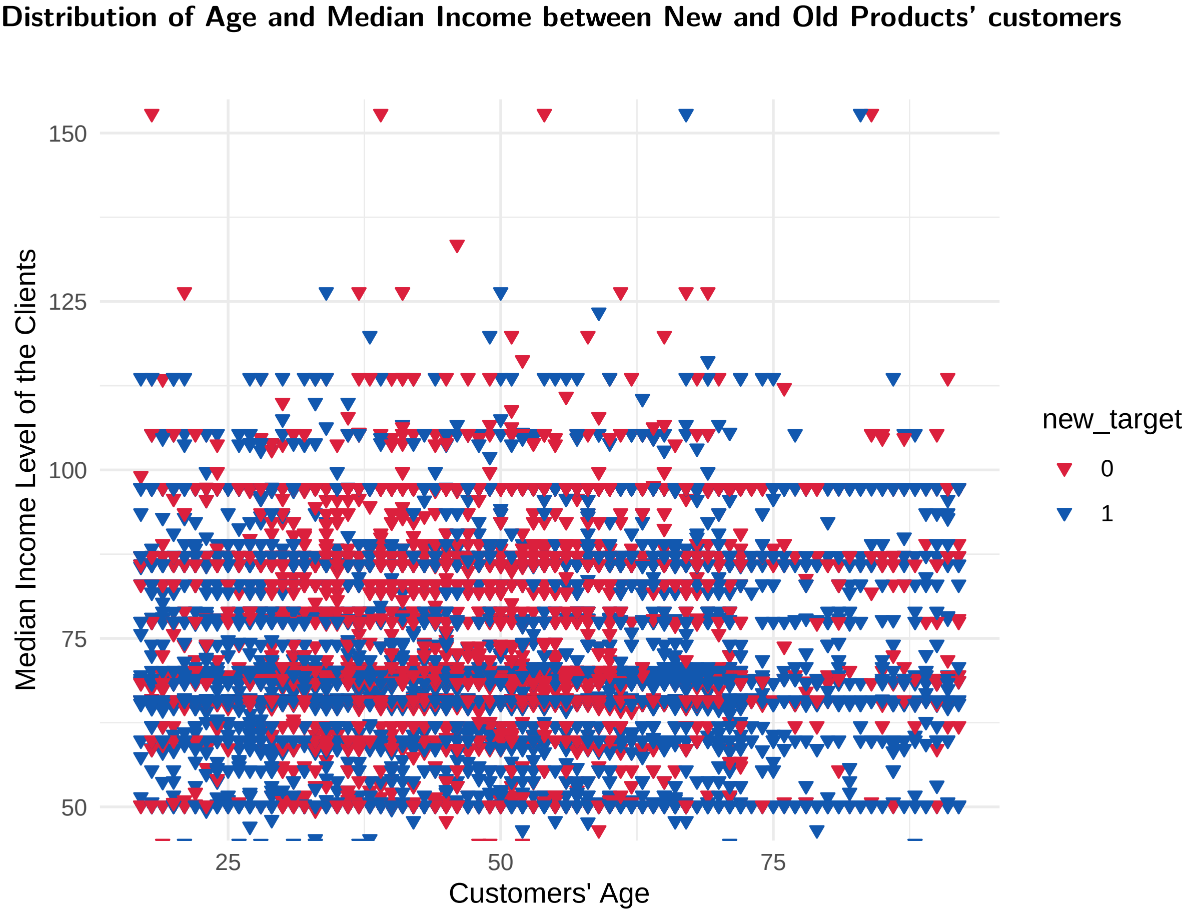

Figure 4: the Distribution of Age and Median Income between New/Old Products' Clients

In Figure.4, customers who buy our new products, "Active" and "Advance", are labeled as "1" in NewTarget, showing the blue inverted triangle. On the contrary, customers who buy other products of our company are labeled as "0" in NewTarget, showing a red inverted triangle. We found that the age distribution of customers who bought the two new products "Active" and "Advance" was significantly wider, covering almost all age groups. Our traditional customers are mainly distributed between 30 and 60 years old, while the new products have attracted some customers from other age groups.

Comparison of Different Product Line Regarding Flags Per Duration and Skin Color

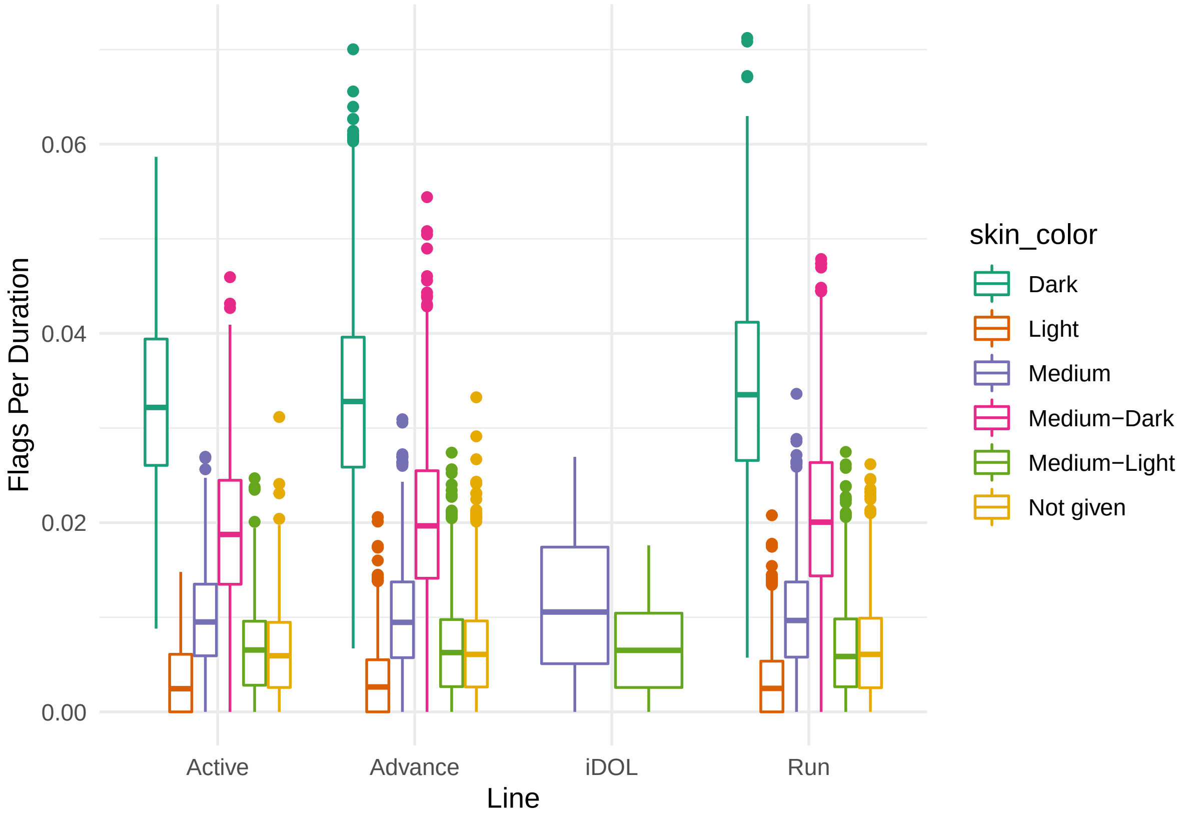

Figure 5: Comparison of Different Product Line Regarding Flags Per Duration and Skin Color

From our previous analysis, we concluded that our products do perform significantly differently in terms of flags per duration between skin tones. We wanted to further investigate the reasons for this, so we categorized the product line by boxplot as well. We grouped our products into four product lines: "Active", "Advance", "iDOL", and "Run". From the graph, it is clear that in the three product lines "Active", "Advance" and "Run", the flag per duration of customers with black skin is significantly higher than that of customers with other skin tones. In particular, "Advance" and "Run" product lines have more outliers. The "iDOL" product line has no purchases from black-skinned customers.

Research Methodology

Our clients consult the detailed description of the personal fitness tracking devices market in Canada and the difference between traditional customers and new customers who buy the more affordable "Active" and "Advance" products. They also want to know the reason for the device's poor performance problem for users with darker skin. Based on the client's request, we explored deep into the data and tried to find the answers, applying methods of hypothesis tests, linear models, linear mixed models, generalized linear models, and model selection.

Comparisons between characteristics of new and traditional customers groups

To provide useful insights for the Marketing team, we built several models to compare the characteristics of new customers attracted by the "Active" and "Advance" products with Mingar's traditional customers. Besides, we will focus on investigating the difference in the average income level between the new and traditional customers. Before any model construction and analysis, we hypothesize that the distribution of age groups will be different between our new and traditional customers, more specifically, **we predicted that compared to the middle-aged concentrated traditional customer base, non-middle-aged customers will be more attracted to the new products. We also hypothesize that the income level of our new customers will be lower than traditional customers, given the lower prices of "Active" and "Advance" products.**

Hypothesis Test

To see if the results informed by the models are significant, we set up hypothesis testing for each model. We will conduct hypothesis tests separately for each of the model parameters, to check if the corresponding predictor is significantly related to the response variable. Thus, the null hypothesis for a single parameter would be $\beta_i = 0, i = 0, 1, 2...$ The alternative hypothesis is $\beta_i \neq 0, i = 0, 1....$ For each hypothesis test, we calculate the t-test statistics by comparing the estimated model parameters with 0. Subsequently, we generate a conclusion about the significance of a parameter according to its corresponding p-values. For the first request, p-values help compare whether there is any significant difference between new and traditional customers. For the second request, p-values indicate whether the selected independent variable significantly influences the device's performance on sleep score. We will estimate the confidence interval for parameters in different models to provide information about the real situation based on the models' results.

Linear Regression Models and Linear Mixed Model

The crucial purpose of our study is to trace the characteristics of Mingar's new customers and compare those with traditional customers. After exploring the data given by our client, **we are most interested in two factors: age and income**. Since age and income are both continuous response variables, we will construct linear regression models for further analysis. The assumptions of the models are satisfied as customers in the sample data are independent. Therefore, we first build a simple linear regression model to set age as the response and analyze the age difference between new and traditional customers.

The linear regression model for age is

$$y_{age} = \beta_0 + \beta_1 x_{new\_target} + \epsilon$$

Age is the continuous response variable. It is the age of each customer. X_{new_target} represents whether a customer will or will not buy the new "Active" or "Advance" products. It is a binary categorical variable. If the customer buys the new products, then $X_{new\_target} = 1$. Otherwise, $X_{new\_target} = 0$. β_0 represents the intercept of the model. It means the estimated average age of a customer if the customer does not buy the new "Active" or "Advance" products. β_1 represents the slope of the model. It means the difference in average age level between the customers who buy the new products and customers who do not buy the new products. ϵ represents the residuals (error terms) of the model.

From our point of view, the customer's income is not the only factor that may influence whether people buy the new products or not. Based on Figure.4 in the Data section, **we hypothesize that the income may also have a potential relationship with their age range**. We observe that the age of traditional customers is roughly concentrated between 30 and 60 years old, but the new products have attracted additional customers from younger and older age groups. To test our estimation, we have created two candidate linear models.The first candidate model is $y_{income} = \beta_0 + \beta_1 X_{new\_target} + \epsilon$. The second candidate model is $y_{income} = \beta_0 + \beta_1 X_{new\_target} + \beta_2 x_{age\_range} + \epsilon$

We apply ANOVA to check if customers' age range will correlate with their income level. In the ANOVA test, the null hypothesis is that the model with a new customer alone fits the data as well as the model with both new customer and age range. Thus, we will take customers' age range into account only if the p-value of the test is greater than 0.1. If the ANOVA test suggests we consider the possible effect from customers' age range, we will only consider the customers' age range to be a random effect since the main purpose of this income model is to study the difference in the income level between new and traditional customers. Hence, the linear mixed model for income with the random effect of age range can be expressed as:

$$y_{income} = \beta_0 + \beta_1 X_{new\_target} + b_{age\_range} + \epsilon$$

$Income$ is the continuous response variable. It is the income of each customer. X_{new_target} represents whether a customer will or will not buy the new "Active" or "Advance" products. It is a binary categorical variable. If the customer buy the new products, then $X_{new\_target} = 1$. Otherwise, $X_{new\_target} = 0$. As for the coefficients, β_0 represents the estimated average income level of customers who do not buy the new "Active" or "Advance" products. β_1 represents the average difference in income between the customers who buy the new products and customers who do not buy the new products. b_{age_range} is the random effect for age range. It represents the difference in income between middle-aged and non-middle-aged customers. ϵ represents the random noise (and potentially other unmeasured confounders). Both b_{age_range} and ϵ are random variables which follows a Normal distribution. $b_{age\_range} \sim N\left(0, \sigma_b^2\right)$ and $\epsilon \sim N\left(0, \sigma^2\right)$.

Logistic Regression Model

To inform about the customer's behavior based on their characteristics, we build logistic regression models to predict the odds of a customer buying the "Active" and "Advance" products of Mingar. We choose to build logistic regression models because the response of whether customers buy our new products or not is a categorical variable with a binary response. Besides, we did not consider generalized linear mixed models as no random effects were found to provide additional information in our case.

The possible related factors we selected are sex, age range, skin color, and income level. From the Data section, we use the color of emoji they use to represent their race or ethnicity in this analysis. This replacement may cause inaccuracy and bias because people may use the default emoji color and the emoji colors they use may only reflect their emoji preference instead of their true skin color. For this reason, we create two logistic regression models, one with skin color and one without skin color to project the customers' behaviors. We will use BIC in model selection and choose the model with a lower BIC value to be our final model. The final model can be expressed as:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{sexIntersex} + \beta_2 x_{sexMale} + \beta_3 x_{age\_rangeNotMiddleAge} + \beta_4 x_{median\_income\_level}$$

In this model, p generally represents the probability of a customer purchasing the Mingar product in lines of "Active" or "Advance". Here are the specific interpretations of our independent variables. $x_{sexIntersex}$ and $x_{sexMale}$ are two different levels of the sex variable. The sex variable has three levels: female, intersex, and male. These two variables will equal 1 if the customer belongs to the indicated sex group, and 0 otherwise. $x_sexFemale$ is the base level that does not appear in the model. $x_{age\_rangeNotMiddleAge}$ is a binary variable. $x_{age\_rangeMiddleAge}$ is the base level that does not appear in the model. $x_{median\_income\_level}$ is a numerical variable indicating the income level of the customer.

The following are introductions of our coefficients.β_0 represents the estimated log odds of purchasing "Active" and "Advance" products when the individual is female in her middle age with zero income. β_1 represents the average difference in the log odds of purchasing "Active" and "Advance" products between Intersex and Females when holding other independent variables constant. β_2 represents the average difference in the logarithmic odds of purchasing "Active" and "Advance" products between Male and Female when holding other independent variables constant. β_3 represents the average difference in the logarithmic odds of purchasing "Active" and "Advance" products between middle-aged and non-middle-aged customers when holding other independent variables constant. β_4 indicated that for every one-unit increase in income(in thousands), we expect an β_4 increase in the log odds of buying "Active" and "Advance" products, holding other variables at fixed values.

There are several model assumptions. First, the binary logistic regression requires the dependent variable to be binary. In our case, it is whether or not customers buy "Active" or "Advance" products. Second, logistic regression presupposes that independent variables are linear and that log odds are positive. Although the dependent and independent variables do not have to be linearly related in this analysis, the independent variables must be linearly related to the log chances. Third, the observations are independent of each other. Every customer makes the choice independently. Fourth, the independent variables are not highly correlated with each other. Our predictors sex, age range, and income have no direct relations. Fifth, it assumes that the variance of the residuals is equal across individuals. Last, logistic regression necessitates a high number of samples. The customer dataset contains 19045 observations, which are big enough to set the logistics regression model.

Analysis of the poor performance of sleep scores for users with darker skin

User experiences considerably affect the reputation of a product. The poor device performance of sleep scores for users with darker skin is our key concern. Our team applies the methods of hypothesis tests and generalized linear mixed models to investigate this issue and will give modification suggestions based on the model results. Based on our summarization in the Data section(Figure.5), we hypothesize that skin color and lines of products will correlate with the performance of devices reporting a quality flag per sleep session. Also, **we hypothesize that users with darker skin will experience more device flags per sleep session**.

Hypothesis Test

Similar to what we did in the first research question, we set up hypothesis tests for each parameter in different models. We use the p-values to confirm whether the parameter significantly influences the response. The null hypothesis for a single parameter would be $\beta_i = 0, i = 0, 1, 2....$ The alternative hypothesis is $\beta_i \neq 0, i = 0, 1....$

Generalized Linear Model (GLM)

Flags is a count variable recording the number of times there was a quality flag during the sleep session for each customer. The number of times recorded flags is how we evaluate the device performance for a customer. Except for the true sleeping quality records of customers, flags may occur due to various reasons, such as missing data, a sensor error, or other data quality issues. **We predict that other than skin colors, there may be additional confounders that correlate with the number of device flags.** Exploring the given data, we consider different lines of products may influence the number of flags. Refer Figure.5 in the Data section, we observe that the flag per duration of customers with black skin is much higher than that of customers with other skin tones in the three product lines “Active”, “Advance” and “Run”.

To find whether the correlation between the number of flags and product lines is significant, a generalized linear model is built. We used Poisson regression to model the number of times there was a quality flag during the sleep session using product lines as a predictor.

Here is the equation of the generalized linear model:

$$Y \sim \text{Poi}(\mu)$$

$$\log(\mu) = \beta_0 + \beta_1 x_{LineAdvance} + \beta_2 x_{\text{LineiDOL}} + \beta_3 x_{\text{LineRun}}$$

Y is the total number of flags for one customer. μ is the expected flag rates for a given line category per duration of sleep for one customer. $x_{LineAdvance}$, $x_{LineiDOL}$, and $x_{LineRun}$ are three different levels of the categorical variable Line. The variable line has four categories in total: "Active", "Advance", "iDOL", and "Run". $x_{LineActive}$ is considered as the base level for comparison, so it does not appear in the written model. Notice every device can only be produced in one line. For example, if the device is in the "Advance" line, then $x_{LineiDOL}$ and $x_{LineRun}$ will be indicated as 0 and our estimation in flags will then depends on β_0 and β_1. β_0 is the intercept of the model. It means the log odds of the expected number of flags when the customer uses the device produced in the line of "Active". β_1 represents the line of "Advance" has β_1 times more flags compared to the reference line of "Active" in the log scale. β_2 represents the line of "iDOL" has β_2 times more flags compared to the reference line of "Active" in the log scale. β_3 represents the line of Run has β_3 times more flags compared to the line of "Active" in the log scale.

Generalized Linear Mixed Models (GLMM)

Our main purpose is to investigate the association between flags per duration and customers' skin tones. Thus, our outcome of interest is the number of times that a quality flag is reported by the device during the sleep session. Our predictors are the customer's skin color and the lines of their products.

The method we applied is generalized linear mixed models (GLMM) in Poisson regression. First, the response variable is the flag counts per sleep session for each customer. It shows the counts when a quality flag happened, so it is numeric but not a continuous variable nor a categorical variable. Thus, we would build a Poisson regression instead of a linear regression model. Second, the observations are not completely independent given the nature of our sample data. In our data, some customers may several measurements on a different date. Thus, there may be correlations of flags counts data for the same customers, and we need to control its influence by setting it as a random effect. Other variables, like skin color and the product lines, are independent for different customers, and will be considered as a fixed effect in our model.

The generalized linear mixed model can be expressed as:

$$Y_{ij} \sim \text{Poi}\left(\mu_{ij}\right)$$
$$\log\left(\mu_{ij}\right) = X_{ij}\beta + U_i$$

Y_{ij} is the total number of flags for customer i in measurement j. μ_{ij} is the expected number of flags recorded per duration for customer i in measurement j. X_{ij} has indicator variables for i^{th} customer and j^{th} measurement. They are skin colors with six levels and product lines with four levels. x_{skin_color} has six categories: dark, light, medium, medium-dark, medium-light, and Not given. According to alphabetic order, $x_{skin\_colorDark}$ is our base level for comparison which does not appear in the model. Every category of x_{skin_color} has two levels: 0 and 1. Since every customer only has one skin color, for example, if the customer has light skin color ($x_{skin\_colorLight} = 1$), then all other categories will be 0. x_{Line} has four categories: "Active", "Advance", "iDOL", and "Run". $x_{LineActive}$ is our base level. Same as x_{skin_color}, every category of x_{Line} has two levels: 0 and 1. U_i is an individual-level random effect. It is the effect brought by the various measurements of the same customers towards the estimation of flags.

There will be a parameter(β) for every level of independent variables. For this model, there are nine in total from β_0 to β_8. β_0 represents the intercept of the model. This represents the estimated log odds of the mean number of flags recorded per duration for customers who have dark skin color and use the "Active" product. $\beta_m form = 1, 2 \ldots 5$ is the coefficient of the corresponding level of x_{skin_color} from medium-dark to light. For example, β_1 is the coefficient of $x_{skin\_colorMedium-Dark}$. It represents we expect a β_1 increase in log odds of the average flags counts recorded for customers with medium-dark skin color than customers with dark skin color when they both use products from the same line. Every $\beta_m form = 1, 2 \ldots 5$ has the same pattern of interpretation as β_1 . $\beta_n forn = 6, 7, 8$ is the coefficient of the corresponding level of x_{Line} of Advance, iDOL, and Run. For example, β_6 is the coefficient of $x_{LineAdvance}$. It represents the average difference in the log odds of flags recorded per duration between "Advance" and "Active" lines for customers with certain skin color. Every $\beta_n forn = 6, 7, 8$ has the same pattern of interpretation as β_6 .

Model Assumptions for GLM and GLMM

For both GLM and GLMM, our grouping units are independent, even though observations within each group are taken not to be. In our case, the chosen link function is appropriate and the model is correctly specified. Thus, we have satisfied these assumptions in both models.

GLMM has some extra assumptions about the random effects. It is assumed that random effects should come from a normal distribution, which means $Random\ effect \sim N(0, \theta^2)$. Besides, the random effects errors and within-unit residual errors should have constant variance. There are also some assumptions for the Poisson regression model. First, the response variable is a count per unit of time, described by a Poisson distribution. This is satisfied as the response variable we chose is the flag count recorded per sleep session. Second, the observations must be independent of one another. This assumption of independence is also met as every customer in our dataset is independent, and we've controlled the dependence within the measurements of a customer by setting a random effect. Third, by definition, the mean of a Poisson random variable must be equal to its variance, and the log of the mean rate $log(\mu)$ must be a linear function of x. However, some of the model assumptions were hard to test, for example, linearity with respect to $log(\mu)$ is difficult to discern without continuous predictors. This may introduce some bias to our analysis and further discussions about these limitations will be shown in the Conclusion section.

Analysis Results: Comparisons between characteristics of new and traditional customers groups

Difference in average age level of new and traditional customers groups

Table.2 Table of linear regression model summary

Variables	Parameter	Estimate value	p_value	Upper CI bound	Lower CI bound
intercept	β_0	46.2531	<2e-16	45.8937	46.6125
new target	β_1	1.4515	3.76e-09	0.9691	1.9340

- Table.2 summarizes the values of the estimated parameters $\hat{\beta_0}$ and $\hat{\beta_1}$ and their corresponding p-value in the linear regression model. Based on the results, the p-value of $\hat{\beta_1}$ is smaller than 0.01, indicating that there is strong evidence against the null hypothesis that there is no difference in the true average age between buyers of the newer and more affordable "Active" and "Advance" products and traditional customers. According to Table.2, **the estimated average age of people who buy the newer and more affordable "Active" and "Advance" products is 47.7046, which is greater than that of the traditional customers of Mingar(46.2531)**. Based on the confidence intervals, the true average age of traditional customers is approximately around 46-47, while the true average age of new customers who buy "Active" and "Advance" products is 1 to 2 ages higher.
  However, based on Figure.4, while the age of traditional customers is roughly concentrated in the middle age(30-60), the new "Active" and "Advance" lines attract some additional customers from both younger and older age groups. Therefore, although the average age difference between the new and old customer groups is not large, we will build a logistic model to compare the age group distribution of customers who buy "Active" and "Advance" products with that of traditional customers.

Lower income level of new customers compare to the tradional customers groups

Besides age, we also want to compare the average income level(in thousands) of traditional customers and customers who buy new the "Active" and "Advance" products. Recall that we hypothesized there will be correlations of income level within an age group, and we built two candidate models with income level as the response variable to investigate whether we should also consider the possible relationships between the customers' age group and their income level. Based on the ANOVA test results, the p-value is $3.664 * 10^{-8}$, meaning that we can't ignore the possible relationship between the age range and the average income level. Since we aim to determine only the difference in average income level between traditional customers and customers who buy new lines of products, **we will consider the age range of customers as a random effect for our analysis of the income difference**.

Table.3 Table of linear regression model summary

Variable	Parameter	Estimated value
intercept	β_0	1.317891
new_target	β_1	0.088789

- Note that the unit of the response variable in the model is in thousands. From the model, the estimated average income level for people who buy the newer and more affordable "Active" and "Advance" products is 68787.4, which is 4651.2 less than that of the traditional customers of Mingar(73438.6). **This is in line with our hypothesis that the average income level of customers attracted by the new "Active" and "Advance" lines is lower than that of traditional higher-income base customers.** The results are reasonable as the prices of the new "Active" and "Advance" products are generally lower than the retailed price level of the previous high-end products, thus new consumers with lower income may be attracted by the more affordable "Active" and "Advance" products of Mingar. However, the estimated income level indicated by this model may simplify the real income distribution of customers and this limitation will be addressed in detail in the following Conclusion section.

Predicting customers' tendency on buying new products given their age range and income level

Recall from the Method section, that our main goal is to compare the characteristics between traditional customers and new customers attracted by the "Active" and "Advance" lines of Mingar. Besides the age and income difference that we've investigated, we are also interested in the difference in age group, skin color, and sex distribution between traditional and new customers. Our initial models include *age group, sex, skin color, and income level(in thousands)* as predictors. However, based on the summarized results of the model coefficients, the p-values of all coefficients for predictors *sex* and *skin color* are significantly greater than 0.1, meaning that both *sex* and *skin color* are not significantly related to the log odds of buying the "Active" and "Advance" products. Therefore, we next performed model selection to determine whether we should keep these two predictors in our final model.

Table.4 Table comparing the BIC value of candidate models

Model predictors	BIC
sex, age group, skin color, income level	24579.1
sex, age group, income level	24532.3

- According to Table.4, the value of BIC of the model without including the predictor *skin color* (24532.3) is lower than the value of BIC of the full model(24579.1). Therefore, we removed the predictor *skin color* from our model. Although we did not find a significant difference in the sex groups distribution between traditional and new customers, we will not remove *sex* from the model as it will over-simplified the model and reduce the reliability of our analysis of the customers' characteristics. Also, knowing what sex groups a customer belongs to can still provide some information for the model to predict the probability that the customer buys the new "Active" and "Advance" products.

Table.5 Table of final logistic regression model summary

Variable	Parameter	Estimated value	Exponential of Estimated value
intercept	β_0	1.317891	3.7355357
sexIntersex	β_1	0.088789	1.0928501

Variable	Parameter	Estimated value	Exponential of Estimated value
sexMale	β_2	0.028107	1.0285060
age_rangeNotMiddleAge	β_3	1.102045	3.0103146
median_income_level	β_4	-0.021743	0.9784919

- From Table.5, the p-values of the coefficient of both age group and income level(in thousands) are significantly smaller than 0.001, so we have strong evidence against the null hypothesis that there is no significant difference in the log odds of buying the new "Active" and "Advance" products between middle-aged customers and customers not belongs to the middle-aged group. Also, the results show that there are significant correlations between the income level and the log odds of buying the new "Active" and "Advance" products.
- Based on the model coefficients, the estimated odds of buying the new "Active" and "Advance" products of a middle-aged female customer who earns a 0 income will be 3.7355357. The estimated odds of buying the new Active and Advance products for an Intersex customer is 9.285% higher than that of female customers, keeping other variables constant. The estimated odds of buying the new "Active" and "Advance" products for male customers is 2.851% higher than that of female customers, holding other variables at a fixed value. Also, the estimated odds of buying the new "Active" and "Advance" products for non-middle-aged customers is 201% higher than the odds for middle-aged customers, assuming other variables keep constant. Furthermore, for each one-unit increase in the income level(in thousands), that is, when the customer's income level increases by 1000, we will see a 2.15% decrease in the estimated odds of buying the new products, holding other variables constant.

Table.6 Table summarizes the exponential 95

Variable	Parameter	Exponential upper CI bound
intercept	β_0	3.2127471
age_rangeNotMiddleAge	β_3	2.8292947
median_income_level	β_4	0.9764734

- As per Table.6, we do not have evidence that the odds of buying the new "Active" and "Advance" products among different sex groups are different. Therefore, we only summarized the exponential 95% confidence intervals of the other model parameters. According to Table.6, we are 95% confident that the true odds of a middle-aged female customer who earns a 0 buying the new A'Active' and "Advance" products will be between 3.2127471 and 4.3451855. For non-middle-aged customers, we are 95% confident that the true odds of buying the new "Active" and "Advance" products are 183% to 220% higher than the odds for middle-aged customers, assuming other variables keep constant. **This supports our hypotheses from the Data section that compared to the traditional high-end products, the new "Active" and "Advance" lines may be more attractive to non-middle-aged customers.** Besides, we are 95% confident that, when the customer's income level increases by 1000, the odds of buying the new products will decrease by 1.95% to 2.35%, holding other variables constant. **Therefore, customers outside of Mingar's traditional higher-income base may be more interested in the more affordable "Active" and "Advance" lines.**

- These results seem reasonable because the new "Active" and "Advance" products are characterized by low prices, so younger and older customers with lower income levels may be inclined to buy the more affordable new products, while some middle-aged customers may be more demanding in terms of product features and have more stable incomes, and thus may prefer to buy traditional high-end products.

Analysis results: the poor performance of sleep scores for users with darker skin

Table 7. Values of Estimated parameters of Line and their Corresponding p-values

Variables	coefficients	Exponential of estimated values	p-values
Intercept	$\hat{\beta}_0$	0.01145452	< 2e-16
Line Advance	$\hat{\beta}_1$	1.03903462	0.000868
Line iDOL	$\hat{\beta}_2$	0.67333871	3.7e-10
Line Run	$\hat{\beta}_3$	0.99459344	0.637063

- Table.7 illustrates the values of estimated parameters ($\hat{\beta}$) of predictor *Line* and their corresponding p-values. Based on the table, the p-value of β_1 is smaller than 0.001, which means we have strong evidence against the null hypothesis that compared to products in the "Active" line, the expected number of flags of products in the "Advance" line is different. Thus, we concluded that the expected flag rates of "Advance" products are 3.903% higher than that of the "Active" products. Similarly, the p-value of β_2 is smaller than 0.001, indicating that there is no evidence against the null hypothesis that there is no difference between the number of flags for products in "iDOL" and "Active". According to the estimated value in the table, the expected flag rates of "iDOL" products are 32.666% lower than that of the "Active" products. Although the p-value of Line Run (0.66558) is greater than 0.1, we still take the variable Line into consideration when investigating the factors that can affect flags rate since the "Run" line only stands for a small portion of all product lines of Mingar. **Therefore, we would consider the variable Line as a possible confounder when we investigate how the difference in skin color correlates with the number of flags of the devices**.

- Recall from the Data section(Figure.5), we summarized that among all product lines of Mingar, the average number of flags per duration for dark skin customers is noticeably larger than that of customers with lighter skin color. We surprisingly find that the average number of flags per duration for dark skin customers who purchased "Advance" products is a bit higher than that of dark skin customers who purchased "Active" products. Also, the "iDOL" line has no purchases from black skin customers, which may explain why the expected flag rates of "iDOL" products are 32.666% lower than that of the "Active" products. Therefore, we hypothesize that the pattern shown in Table.7 could be driven by the correlation between skin color and the device's flag rates. Next, we will construct a generalized linear mixed model to investigate how skin color and line correlate with the number of flags under the control of the correlation of flags among different measurements of the same customer. **We consider the customer's ID as a random effect since one customer may have various measurements in the given data set**.

Table 8. Values of Estimated Parameters of Skin Color and Line, and their corresponding p-values

Variables	Estimated values of the corresponding	p-values
Intercept	-3.42295	<2e-16
skin_color Medium-Dark	-0.50160	<2e-16
skin_color Medium	-1.21359	<2e-16
skin_color Medium-Light	-1.61923	<2e-16
skin_color Light	-2.39240	<2e-16
skin_color Not given	-1.63277	<2e-16
Line Advance	0.02224	0.1019

Variables	Estimated values of the corresponding	p-values
Line iDOL	0.06871	0.3237
Line Run	0.03057	0.0252

- The table above illustrates the estimated parameters of skin color and Line and their corresponding p-values after controlling the random effects for customers. Notably, the p-values of variable Line Advance and Line iDOL are greater than 0.1, indicating that compared to the flag rates for "Active" products, there is no significant difference in the flag rates for both "Advance" and "iDOL" products. This supports our hypothesis from Table.7 that the pattern shown in Table.8 could be driven by the correlation between skin color and the device's flag rates. **Therefore, we would exclude the predictor Line from our generalized linear mixed model and refit the model**.

Table 9. Values of Estimated Parameters ($\hat{\beta}$) of Skin Color and their Corresponding p-values

Variables	Coefficients	Estimated Values	Rate Ratios e^{β}	P-values
(Intercept)	$\hat{\beta}_0$	-3.399485	0.033390	<2e-16
skin_color Medium-Dark	$\hat{\beta}_1$	-0.501458	0.605647	<2e-16
skin_color Medium	$\hat{\beta}_2$	-1.214176	0.296955	<2e-16
skin_color Medium-Light	$\hat{\beta}_3$	-1.616923	0.198509	<2e-16
skin_color Light	$\hat{\beta}_4$	-2.390913	0.091546	<2e-16
skin_color Not given	$\hat{\beta}_5$	-1.633564	0.1952325	<2e-16

- From the table above, the intercept $\hat{\beta}_0$ (-3.399485) shows that the expected number of flags for people has dark skin color is 0.03339. Moreover, $\hat{\beta}_1$ (-0.501458) demonstrates that the estimated average number of flags for people with medium-dark skin color reduces by 39% compared to dark skin customers. Similarly, the estimated average number of flags for people with medium skin color reduces by 70% compared to people with dark skin color. Furthermore, $\hat{\beta}_3$ (-1.616923) demonstrates that the estimated average number of flags drops by 80% compared to that for people with dark skin color. Lastly, $\hat{\beta}_4$ (-2.390913) demonstrates that there is a 91% reduction in the estimated average number of flags for people with light skin color compared to people with dark skin color. Since all of the p-values shown in the table above are less than 0.001, there is very strong evidence against the null hypothesis that there is no difference between the flag rates for people with corresponding skin color and those for people with dark skin color. As skin color changes from medium-dark to light stepwise, the differences in estimated flags rates between customers with the corresponding skin color and dark-skin customers become larger. **Therefore, we can conclude that the number of device flags is correlated with the customer's skin color, and customers with darker skin are expected to experience higher device flag rates, meaning that there are problems of poor device performances for users with darker skin.**

Conclusions

To serve the needs of our client company, we mainly focus on summarizing the characteristics of new customers who buy the "Active" and "Advance" products, and compare these customers with Mingar's traditional customer base. In this analysis, we built a linear regression model to compare the difference in age and income level between new customers and traditional customers. A logistic model is also built for predicting the odds of customers with different characteristics buying the new products. Besides, we also construct a generalized linear mixed model to investigate the problem of poor performance of research devices with dark-skinned customers.

Based on our analysis, **among the new customers who bought "Active" and "Advance" products, female and middle-aged (30-60) customers accounted for a larger proportion**. By building the regression model, we found that the difference in income and age group distribution between old and new customers is obvious, while people with different gender and skin color do not differ much in their odds of purchasing the new products. Further, based on our prediction model, every time a customer's income level increases by 1000, their odds of purchasing "Active" and "Advance" decrease by 2.15%, holding other variables constant. In addition to the gap in income level between the old and new customer groups, the estimated odds of buying the new "Active" and "Advance" products for non-middle-aged customers is 201% higher than the odds for middle-aged customers, assuming other variables keep constant. For non-middle-aged customers, we are 95% confident that the true odds of buying the new Active and Advance products are 183% to 220% higher than the odds for middle-aged customers, assuming other variables keep constant.

Regarding the first research question, **we conclude that unlike customers in Mingar's traditional high-income base, the more affordable "Active" and "Advance" lines appeal to a lower income level group.** In addition, **non-middle-aged customers are**

more likely to be attracted to the new "Active" and "Advance" lines than to Mingar's traditional high-end products. These results are reasonable because the new "Active" and "Advance" products are characterized by low prices, so younger and older customers with lower income levels may be inclined to buy the more affordable new products, while some middle-aged customers may be more demanding in terms of product features and have more stable incomes, and thus may prefer to buy traditional high-end products.

The second purpose of this report is to investigate the differences in the devices' performances driven by customers with different skin colors. Our hypothesis is that skin color and lines of products are both associated with the performance of devices. However, after building and interpreting the generalized linear mixed model, we found that there is no relationship between lines of products and the number of flags. **Our key conclusion is that the number of device flags is correlated with the customer's skin color.** More specifically, **for customers with darker skin, our devices are more likely to have more flags during a sleep session**. For instance, compared to people with dark skin color, there is a 91% reduction in the estimated average number of flags for people with light skin color. This key result demonstrates the unsolved problems of poor device performance for users with darker skin. Therefore, to optimize our product, we strongly recommend the department of technology investigate the reasons behind these defects and optimize the users' experiences with respect to the sleep tracking function.

Strengths and limitations

- The strengths of our company can be shown by the outstanding ability of our employees to construct and interpret the appropriate model, produce reproducible outputs, and precisely communicate the central ideas of the analysis to different groups of people. We are always dedicated to providing high-quality data analysis to our clients. Furthermore, we have accomplished substantial research reports for our clients in the past years, contributing to an incredible increase in their sales.
- Though we have tried our best to use various models and reduce personal bias for analysis, there are three inevitable limitations. The first one is that the income of every customer is private. Mingar is very concerned about protecting the privacy of its customers. We use the median income level of the customer's neighborhood as the individual income of the customer. However, the income levels may be very different from person to person even though they live in the same neighborhood. This raises the inaccuracy and bias when we describe our new customers who buy "Active" or "Advance" products. Second, we use the color of emoji our customers frequently use to represent their skin colors. The skin colors of customers are crucial information for our analysis of the poorly performing devices

for darker skin users. Mingar opposes any racial discrimination and extremely not want our devices labeled as "racist". The emoji colors are the most related data we can get to estimate their skin colors, but it is not accurate. Many people use the default emoji color, and the emoji color they use may depend on their personal preference with no direct relationship with their actual skin colors. In addition, there are a few model assumptions that are hard to check. For example, when fitting the generalized linear mixed model in Poisson regression, it is hard to check the normality assumption of random effect and to check the linearity of $log(\mu)$ without continuous predictors. Therefore, the results of the model may not accurately reflect the information in the data when the model assumptions are not perfectly met.

Consultant information

Consultant profiles

- Yifan Cheng is a senior consultant with Live Infinitely. She specializes in data visualization, and Yifan Cheng received a double degree in statistical visualization studies and financial modeling from the University of Toronto and Harvard Research School in 2023. She is an expert in quantitative data analysis and visualization research.
- Yiwen Li is a senior data analyst with Live Infinitely. She specializes in computer modeling and predictive analytics, including designing and providing data solutions and tools to enable product-related inquiries for companies. Yiwen earned her Bachelor of Science, Specialist in Statistics Methods and Practice, from the University of Toronto in 2023.
- Jiayi Jin is a senior consultant with Live Infinitely. She specializes in data visualization, statistical communication, and reproducible analysis. Jiayi Jin received her double major Bachelor's degree in Statistics and Economics from the University of Toronto in 2023.
- Yuxin Du is a senior consultant with Live Infinitely. She has expert skills in analyzing quantitative data and scenario analyses in relation to budgets and forecasts. She has firm leadership to present appropriate solutions and provide project marketing advice. Yuxin earned her Bachelor of Science, major in Statistics, Economics, and Mathematics from the University of Toronto in 2023.

Code of ethical conduct

- Live Infinitely promise to avoid selective reporting of data with the intent to mislead or deceive when conducting any analysis work for our clients. Our experts strive to be objective when engaging in statistical practice, avoid personal bias, and will carefully examine relevant assumptions before using any statistical tools. In addition, our company promises to provide valid and reliable information to our clients and the public by honestly informing them of the possible limitations of our analyses.

- Live Infinitely is committed to respecting and protecting human dignity and rights. Our company will refrain from conducting any analysis that involves discrimination. In serving our clients, we will strictly adhere to privacy laws and the privacy guidelines established by the SSC. We guarantee that any non-public data used in the analysis will not be disclosed. We will refrain from disclosing confidential information collected for and derived from the analysis without the prior written permission of the employer or client.
- Live Infinitely is dedicated to providing professional, respectful service to our clients. All results of this study are presented in a way that allows analysis and review, the data cleaning process is clearly discussed in a way that ensures the study's reproducibility. All external public information used in our analysis will be appropriately cited and are listed in the Bibliography section of our analysis report.

References

> [1]Petry, N. M. (2002, February 1). Comparison of young, middle-aged, and older adult treatment-seeking pathological gamblers. OUP Academic. Retrieved April 7, 2022, https://academic.oup.com/gerontologist/article/42/1/92/641498

> [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. https://rvest.tidyverse.org/, https://github.com/tidyverse/rvest.

> [5]Dmytro Perepolkin (2019). polite: Be Nice on the Web. R package version 0.1.1. https://github.com/dmi3kno/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]H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

> [8]Hao Zhu (2021). kableExtra: Construct Complex Table with "kable" and Pipe Syntax. http://haozhu233.github.io/kableExtra/, https://github.com/haozhu233/kableExtra.

> [8]Industry data retreived from https://fitnesstrackerinfohub.netlify.app/

> [9]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

> [10]Census data scraping from Census Mapper API. Census mapper. (n.d.). Retrieved April 7, 2022, <from https://censusmapper.ca/api>

> [11]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/./

Appendix

Appendix.1 Figure comparing the customer purchasing preference

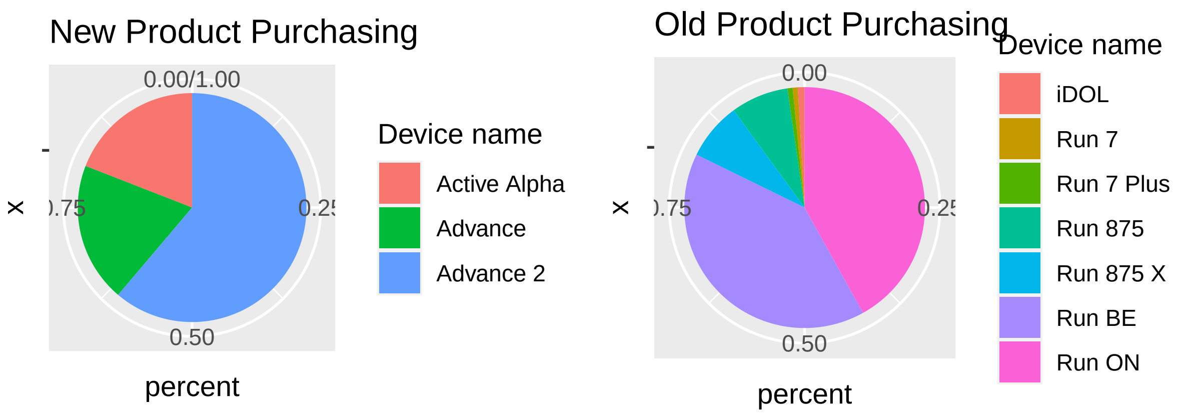

Figure 6: Comparison of Customer Purchasing Preference

Understanding the shopping tendency of our customers is also very helpful for our product promotion department. Compare the table of the percentage of our customers who buy new products and old products. We found that more than half of our customers bought new products, which means that the introduction of new products has greatly increased our customers. Further, we found that "Advance2" and "Advance" are the two most popular products, accounting for almost 3/4 of the new product market. The percentage of customers who bought the old product between the different devices shows that the "Run ON" and "Run BE" devices take the majority of the market.

Device name	Line	Retail price	Battery life	Released	Number of functions
Advance 2	Advance	145	Up to 7 days	2021-07-08	5
Active Alpha	Active	99.99	Up to 7 days	2020-12-30	4
Advance	Advance	120	Up to 7 days	2020-08-20	5
Active	Active	39.99	Up to 14 days	2019-10-13	0
Active HR	Active	79.99	Up to 7 days	2019-10-13	1

Device name	Line	Retail price	Battery life	Released	Number of functions
Rush 5	Rush	179.99	Up to 7 days	2021-03-17	6
Versus 4	Versus	229.99	Up to 5 days	2020-06-08	6
Rush 4	Rush	149.99	Up to 7 days	2020-03-04	6
Rush 3	Rush	149.99	Up to 7 days	2019-03-10	5

Appendix.2 Figure comparing the products features between MINGAR and BITFIT

We analyzed this in conjunction with a chart of customer preferences for our new products. "Advance 2" and "Advance" are the two most popular fitness tracking devices in our new product line (Appendix.1). The "Advance 2" is the fitness tracking device in our latest year lineup, and as we can see from the first table, it has a total of 5 features and a retail price of 145. Compared to Bitfit's "Rush 3", which also has 5 features, the "Rush 3" retails for 149.99. So this product has an advantage in terms of price and trends. It is worth noting that the "Advance" in last year's lineup retails for only 120 and has the same number of features as the "Rush 3". It is inferred that these two products in the new product lineup have a high price advantage in their class in the market. According to the second table, several of Bitfit's products have six features, albeit at a high price point. But in comparison, several of Mingar's new products do not yet have all 6 functions available. This is a breakthrough for Mingar. Developing products with more comprehensive features while keeping the product price at a lower level could be a viable strategy to attract more customers.

Web scraping industry data on fitness tracker devices

```r
# These are the libraries for webscraping
library(tidyverse)
library(polite)
library(rvest)

# the url of the website
url <- "https://fitnesstrackerinfohub.netlify.app/"

# informative user_agent details
target <- bow(url,
              user_agent = "xxx@xxx.com for xxxxx use",
              force = TRUE)

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

# scrape the data
html <- scrape(target)

# save data into device dataset
device_data <- html %>%
  html_elements("table") %>%
  html_table() %>%
  pluck(1)
```

Before we scraped the data from the website of Fitness Tracker Info Hub, we installed the libraries for help and loaded the url of the website. We followed the instructions of the robots text on crawl delays and which agents are allowed to scrape. We provided a User Agent string that makes our intentions clear and provides a way to contact us with questions or concerns. Specifically, we wrote the User Agent string with the purpose of academic project and the email for communication. Then we scraped the data and saved it as a dataset.

Accessing Census data on median household income

```r
# Set up any libraries
# install.packages('cancensus')
library(cancensus)
# provide the private API
options(cancensus.api_key = "Our API here",
        cancensus.cache_path = "cache") # this sets a folder for your cache

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

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

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

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

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

As an ethical scraper, we strictly follow the site's terms and conditions. Before web scraping, we read all the Terms and Conditions and the Robots text. When we scraped the census data, we used the API and we did not scrape all together. Besides, we only save the data that we absolutely need for our report. We filtered the regions with the level of CSD, used CSD to match the median income, and selected the data of CSDuid, median income and population for later use.

Accessing postcode conversion files

```r
# set up the libraries
# install.packages("haven")
library(haven)
library(tidyverse)
# read the sav file
dataset = read_sav("data-raw/pccfNat_fccpNat_082021sav.sav")
# choose only PC and CSDuid as we need
postcode <- dataset %>%
  select(PC, CSDuid)
```

To access Census Canada Postal Code Conversion Files, we accepted a license agreement. To download PCCF data, we selected the 2016 census year. We were interested in data of 2016 because we wanted to merge the census data and customer portfolio data by postcode. Since the census data was collected in 2016, the postcode file should also be 2016 for information of consistency . We downloaded the sav file version and read it into R. We chose PC and CSDuid data that absolutely need for our report, and saved into the postcode dataset.

For any content we scraped from the websites, we used it to create new value from the data, not to duplicate it or to pass it off as our own. We properly cited the sources in the reference section in our report.