Statistical Analysis on New Customers' Features and Sleep Monitoring Function of MINGAR's Devices

Young, old people, and people live in low income area are more likely to be new customers and devices perform poorly on dark skin customers

Report prepared for MINGAR by team4.0

2022-04-07

Contents

Executive summary

Background

Mingar is a GPS unit production company and aims to develop fitness tracking wearable-related devices in recent years. In order to gain market share from competitors with cheaper and smaller products, it expands its product lines and wishes to gain some understanding in potential customers based on data analysis. Whether the new product lines would boost their market share with new customers is a key factor they focus on. Also, fitness tracking wearable GPS devices have recently received complaints about performing poorly on people with darker skin. Mingar wants to investigate the relationship between skin colors and sleeping scores in order to show their serious attitude on this.

Aims

This investigation will eliminate the least important features and features containing missing value and focus on:

1. The characteristics of new customers of Mingar's new product lines.
2. The relationship between sleeping quality and skin color with the number of flags being the target variable.

The results of question 1 are summarized below

- Based on figure 1, Customers of "Advanced" and "Active" products have a wider range compared to other products. Sales of these new products of Young teenagers, young adults (18 ~ 28 years old), and old people (60 ~ 92 years old) are significantly higher compared to the market for other products.
- Sales of "Advanced" and "Active" products are significantly higher for people living in low-income area (median income is lower than 50000) compare to the market for other products.

The results of question 2 are summarized below

- The number of flags per minute would decrease by 22% with 1 unit increase in age (starting at 18 years old).
- Based on figure 2, the average number of flags of dark-skin customers is significantly higher than customers with other skin colors.
- The mean number of flags in the dataset across all genders, skin colors, ages, and other variables is 4.32159.

Limitations

- The number of flags cannot be the only determinant to test a person's sleep quality. When there is missing data or the data is unusual, it is probably due to a sensor error or that we collected low quality data. Thus, the data may be inaccurate.
- Data for only male and female are used for this investigation. Removing data of intersex customers(222 customers) would lead to data missingness.
- Due to data collection (survey) methods, the data may not fully satisfy the assumptions.

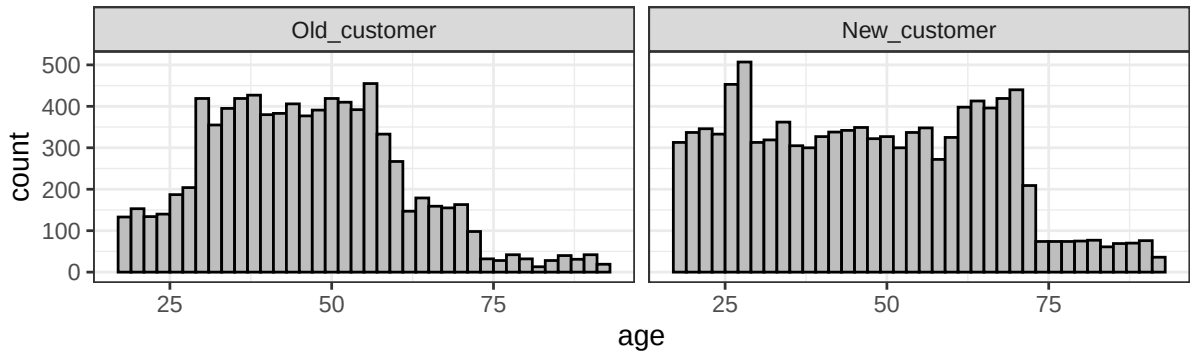

Figure 1: Histograms of Number of Customers on Old Products and New Products with respect to their Age

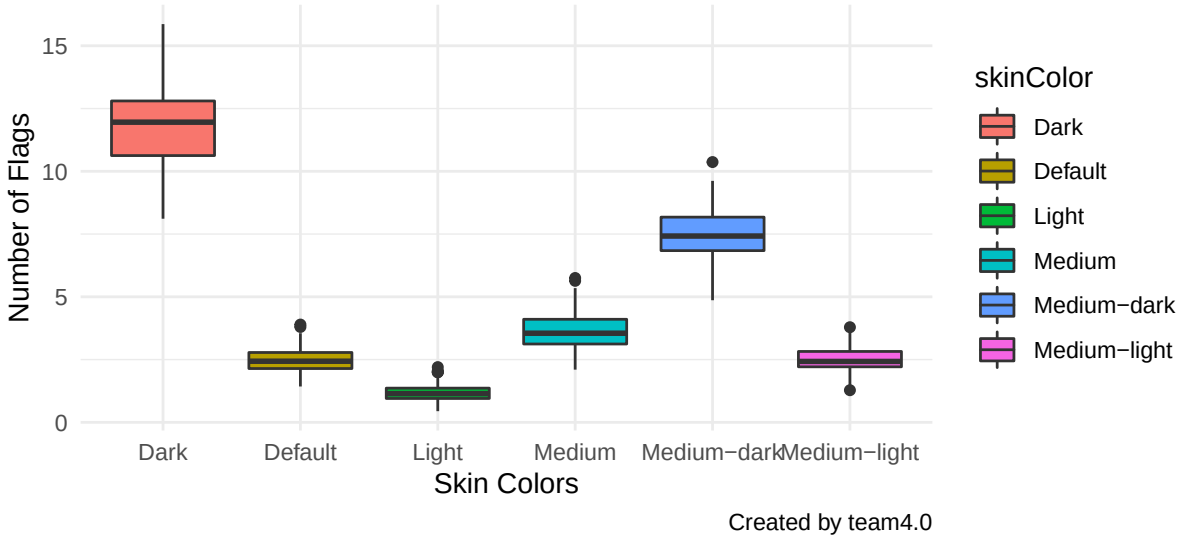

Figure 2: Box Plots of Flags (Took mean for each customer's observations) with Respect to Different Skin Colors in sleep_data

Technical report

Introduction

The GPS units company Mingar has recently entered the market for fitness tracking wearable devices. The competitor's devices are less expensive with smaller sizes. In order to gain market share from the competitor, Mingar decided to add "Active" and "Advance" lines at a more affordable price for new potential customers. For promoting the new products in the best way, this report will suggest the characteristics of Mingar's new target customer. By comparing the old customers and new customers on the more price-competitive lines, this report will analyze whether the new lines would increase Mingar's market share among people with less willingness to spend money on devices. The trend in complaints on social media shows that Mingar's fitness devices tend to perform poorly on customers with darker skin color. Further investigation on the factors that influence the sleeping quality is required for Mingar to prove that its ways of collecting data are not "racist". Therefore, the other purpose of this report is to investigate whether skin colors have an influence on sleep scores. This report will also include the investigation's responsibility to both stakeholders and society and will explain the steps and considerations when web scraping and accessing data from other public resources.

Research questions

- What is difference between new customers with "Active" and "Advance" lines of products and traditional customers?
- What are the effects of various skin colors and other potential factors on sleep score of Mingar's wearable personal GPS devices?

Investigation on the characteristics of new customers of the new products

For this research question, the original data is *customer_data.Rds*. we used web broaches to obtain information about Wearables Fitness Trackers, including customer-level data and Device linkage Data and Device data.
For research convenience and realistic interpretation purpose, we merged and cleaned datasets given and other datasets scraped. The detailed changes are the following:

- Since we need to fit model with features (age). So, we rescaled ages by subtracting minimum age and dividing the difference between maximum and minimum ages.

- Since we need to fit model with features (household median incomes). So, we rescaled ages by subtracting minimum age and dividing the difference between maximum and minimum ages.
- For the feature(skinColor), there are 5 classes, dark, medium dark, default, medium light, and light. Since we don't decide to use one-hot to split the column. We will not change this column. The problem if default class will be explained later.
- For the NA data, use mean value to fill it is one of the good ways to solve this problem. However, we remove all the NA data because of sufficient amount of data.
- We mutated a new column called status, because we have no features as our target feature for research. The value of status column is binary, either 0 or 1. Thus, status is a indicator for each customer in the dataset. 0 indicates this customer is "other customer" who did not buy "Advance" and "Active" product, 1 indicates this customer is "new customer" who bought "Advance" and "Active" product.
- Since some of the columns are unnecessary, we decide to remove them, such as **postcode** which could be represented by CSDuid

For our final dataset, we have 12 columns: "cust_id", "sex","device_name" ,"line","CSDuid" ,"hhld_median_inc" "Population" ,"age" ,"age_scaled" ,"income_scaled", "skinColor", "status", here are the table new variables introduced through data cleaning process:

Table 1: Table of New Columns in sleep_data Which Were Not in Given Datasets

Column Name	Description
CSDuid	Census subdivision unique identifier from Canadian census
hhld_median_inc	Household median income from Canadian census
age	Ages of customers, calculated based on birth date with respect to 2022
skinColor	Estimated skin color of the customer based on emoji modifier provided
age_scaled	Ages of customers, rescaled to values in 0 to 1
income_scaled	Household median income, rescaled to values in 0 to 1
status	Indicator for customers. 1 indicates customers bought new products, 0 otherwise

These variables are separated as following:

- CSDuid: This could be a grouping factor for our data since customer ids are unique.
- sex, age, household median incomes, skin colors: Features would be predictors in modeling
- line, device_name: These two variables has already contributed in creating status variable, so they are not considering in future modeling.
- status: Feature would be response variable in modeling.

Next we will conduct preliminary analysis on images of CSDuid, sex, Age (age_scaled), household median incomes (income_scaled) features.

Analysis of Features by plot and Visualization

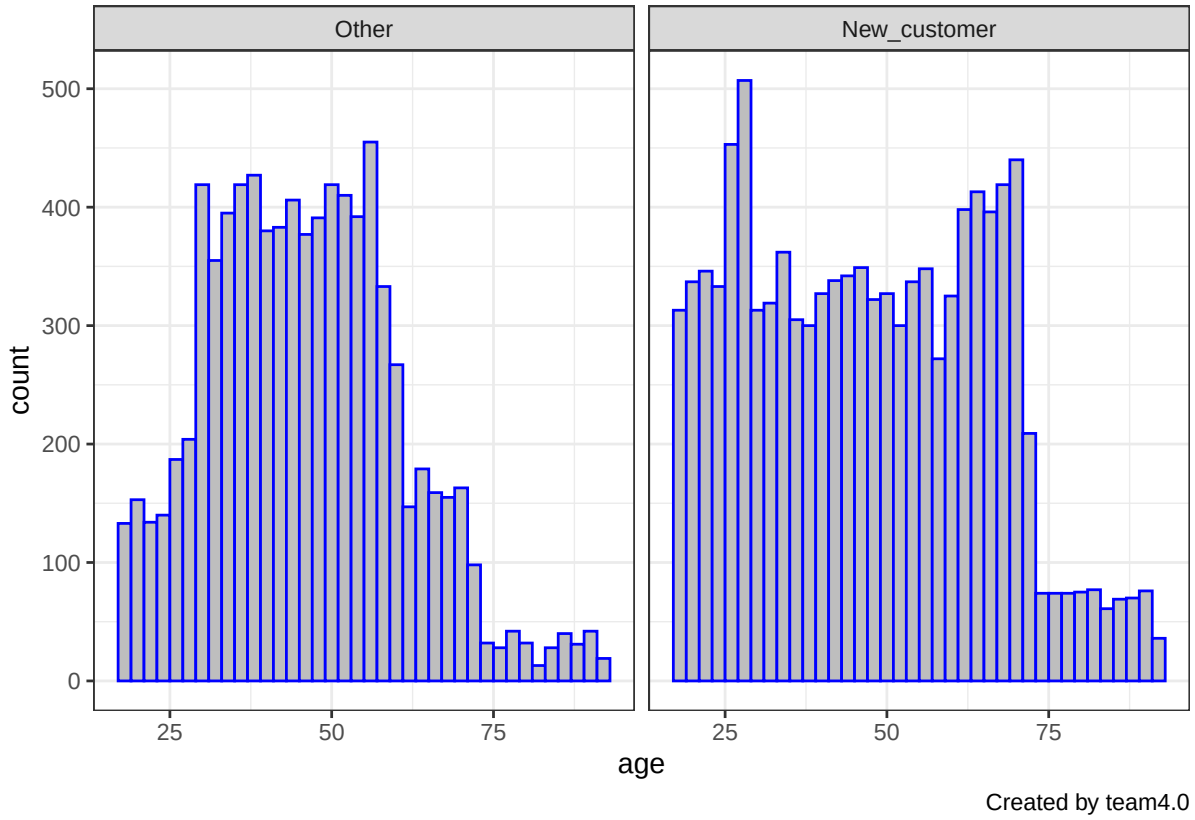

Figure 3: Histogram of Age by product

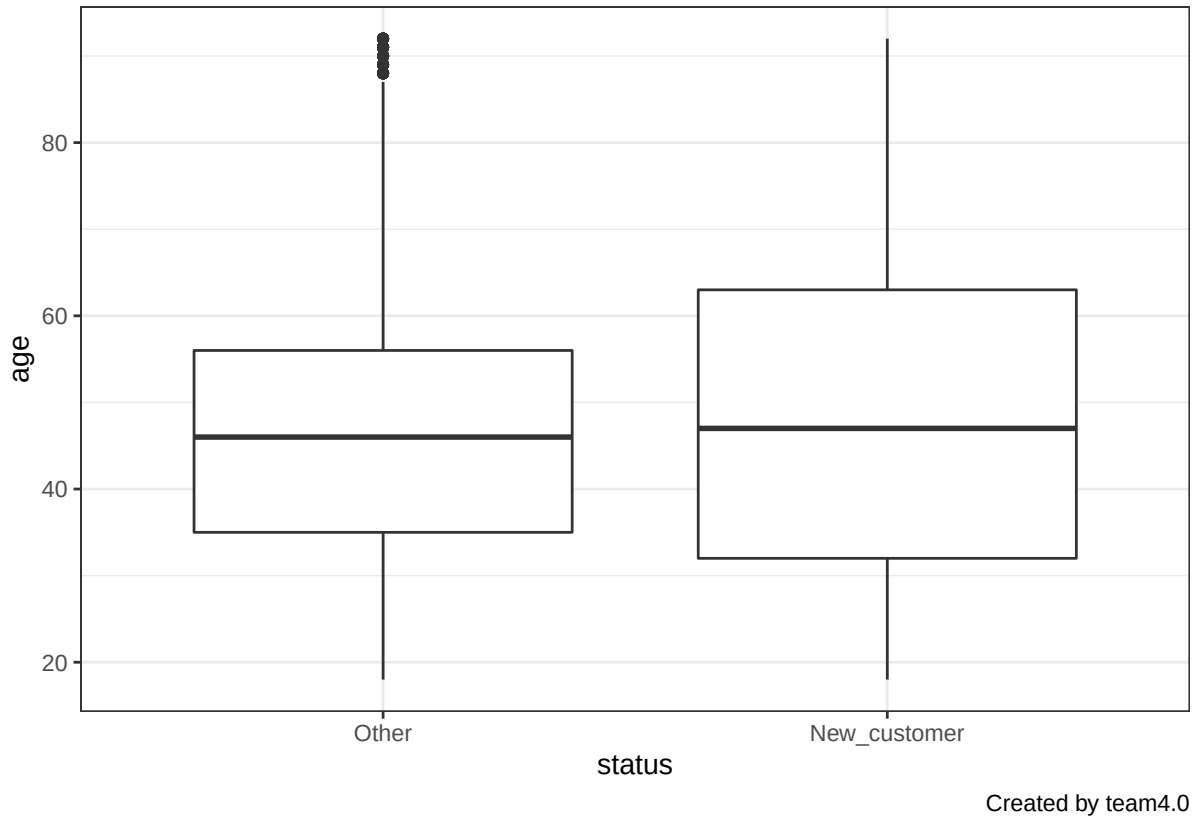

Figure 4: Boxplot of Age by product

Firstly, we used histogram to analyze the situation of age in the two categories. As can be seen from the figure 3, the age distribution of other products is slightly skew normal distributed. The difference between the two is that the new product has more customers in the area with lower age. It is worth mentioning that in other age groups, the population distribution of the two products is similar, indicating that the new product has not completely lost its appeal to other age groups. For the older age groups, the number of customers are also increase.We also used Boxplot to look at the data more visually and found that the age distribution of customers for the new product is not that concentrate compared to other products. Their median is relatively close. However, we still cannot confirm how the age changes from the image, so we need to continue the test in the following step.

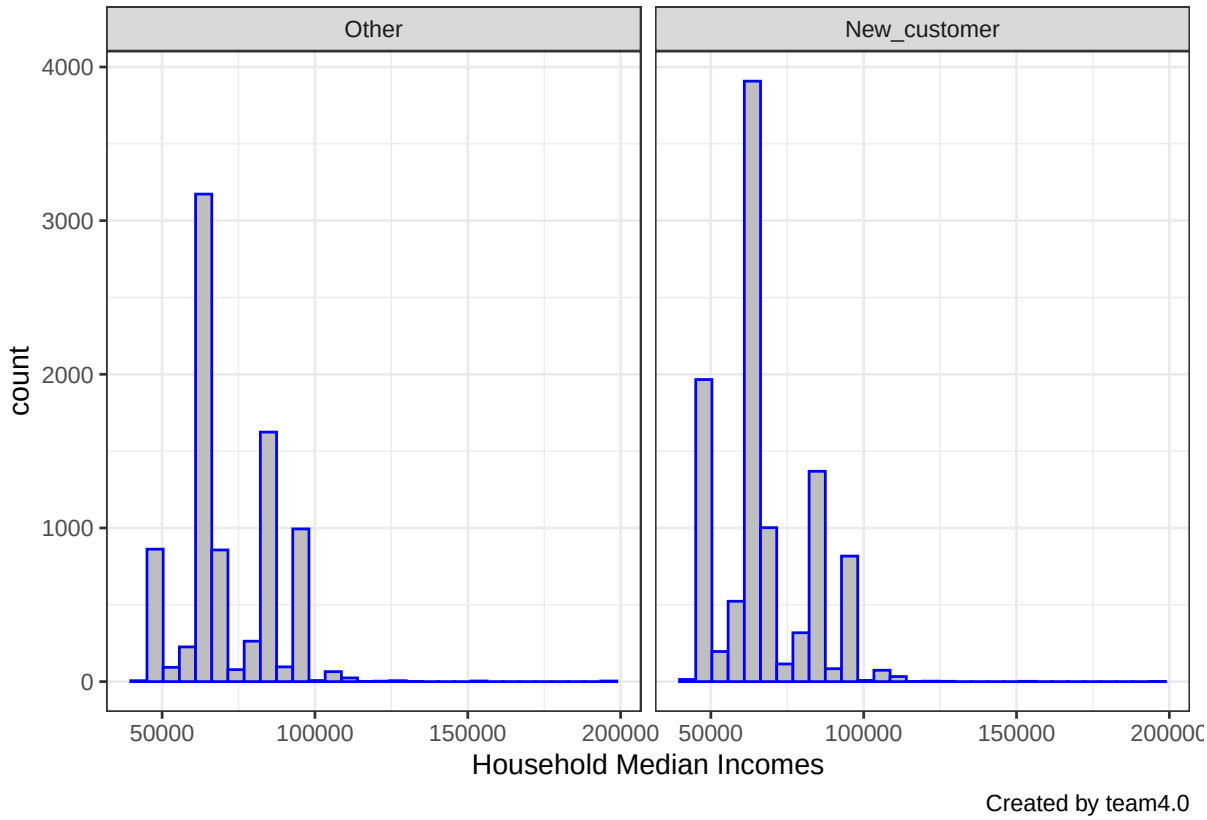

Figure 5: Histogram of Income by product

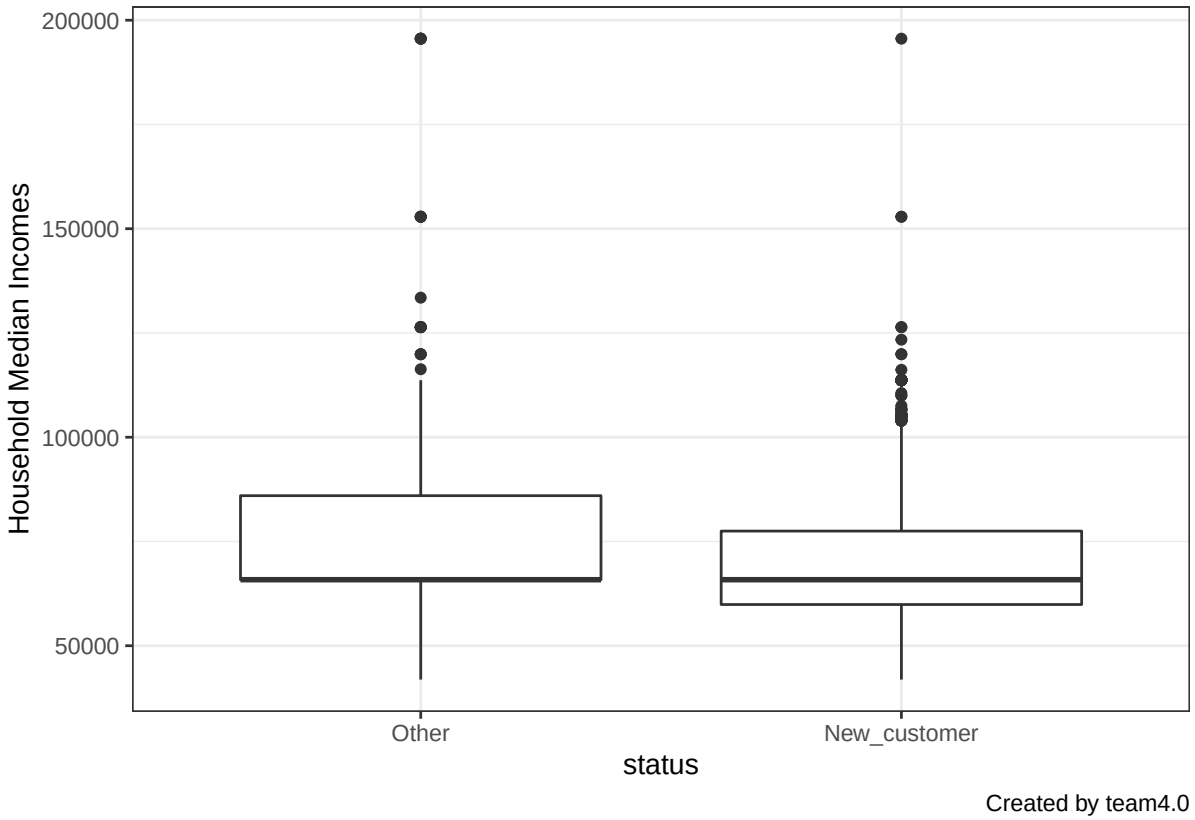

Figure 6: Boxplot of Income by product

Secondly, we continued to use histogram to analyze income in the two categories. The results given by figure 5 show that the clients whose income is between 50,000-100000 still occupy the largest proportion. In the data of customers with income below 50000, we find a huge increase in the number of customers, which indicates that low-income customers have a strong intention to buy new products. As can be seen from figure 6, when median is almost the same, data distribution moves to salary, which is also consistent with the conclusion drawn from Histogram.

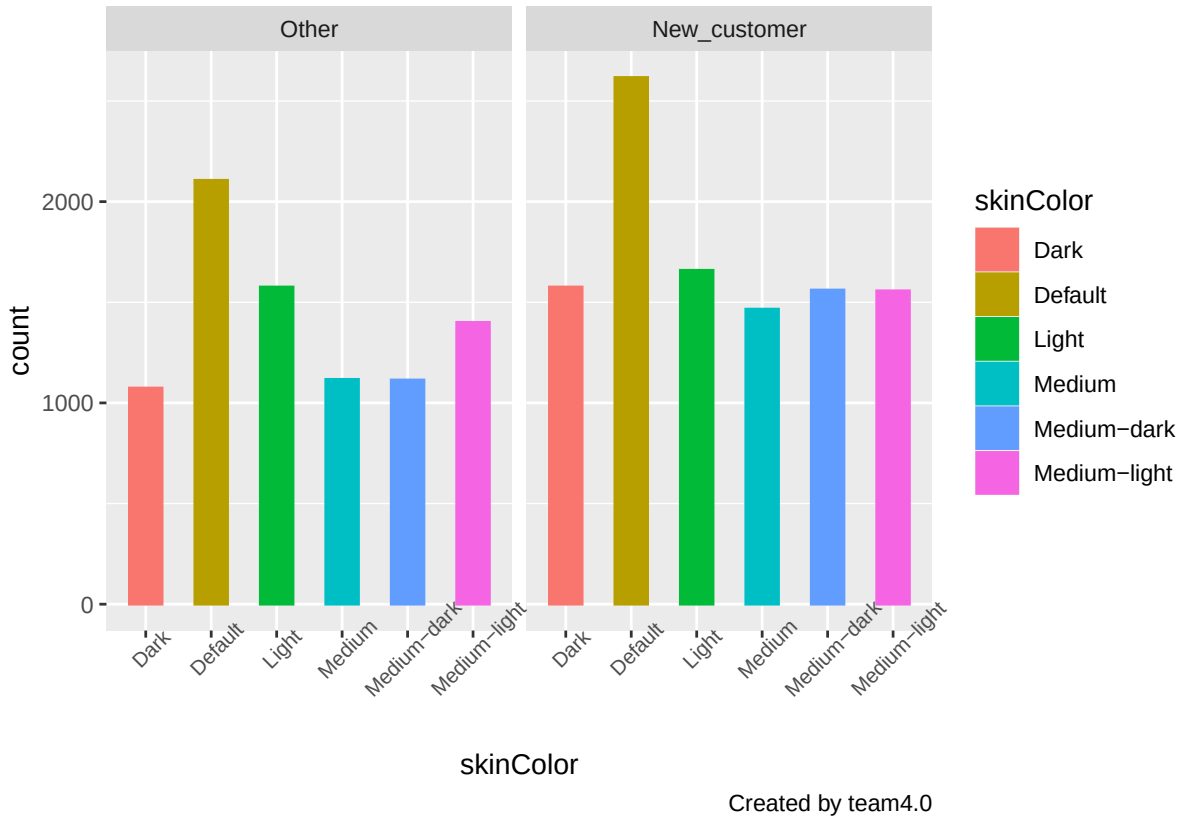

Figure 7: Histogram of Skin color by product

Later, we also analyzed customers with different skin color. No matter from the perspective of histogram or box plot, we all believed that Skin color was not a factor of purchasing power. As showed in figure 7, if the number of people in default group is the largest, we suspect that this is because customers are less willing to answer this question when collecting data, so the number of people in Default is particularly large. We will not use boxplot to visualize this feature because of the modeling part.

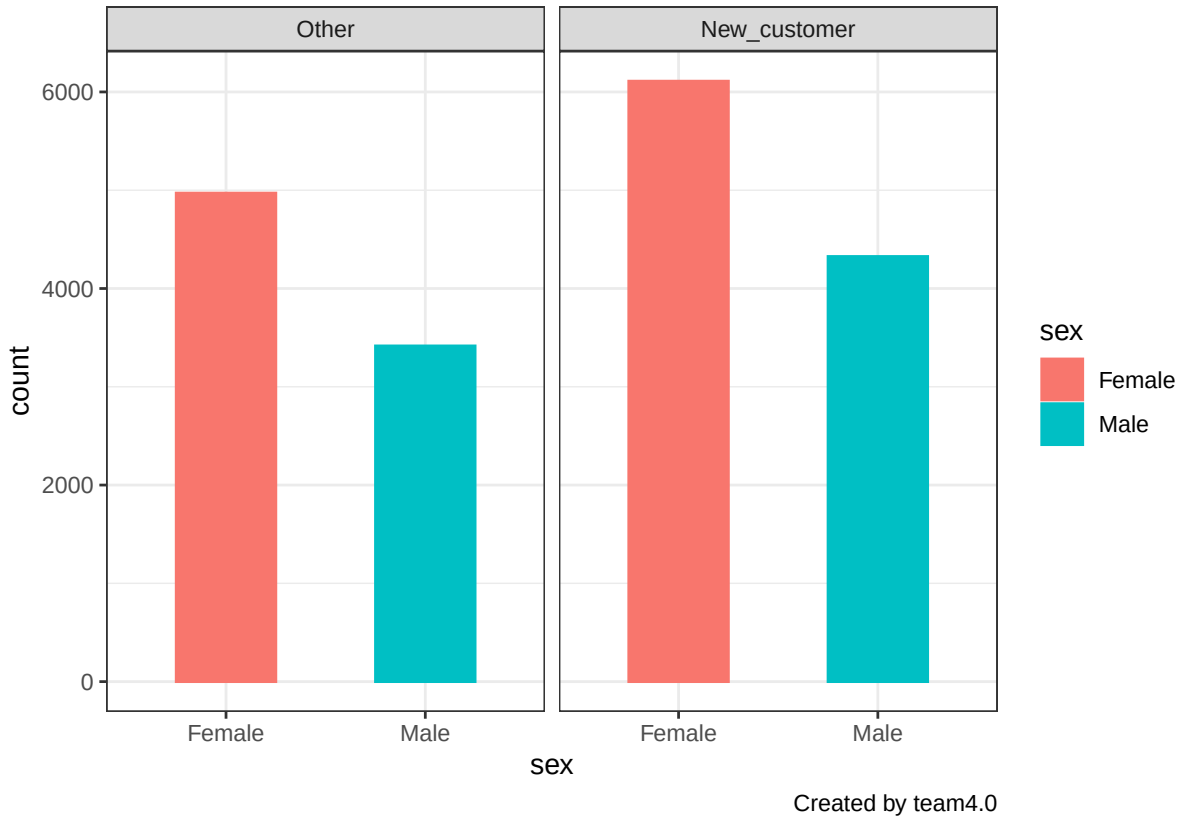

Figure 8: Histogram of sex by product

In the case of sex, there is no way to accurately see the effect from figure 8, so we will continue to study it later in the modeling process. We will not use boxplot to visualize this feature because of the modeling part.

Until now, we found from the graphs that People with lower income prefer new products. For new products, the age range of target population may change. Skin Color and sex do not contribute on choice of products. For Skin color, default does not mean all are yellow, someone may be too lazy to set up emoji color.

Model Selection

For this part, we would use the modeling method to analyze. We will refer to the information for each model in the following steps. After all the operation, we will make a summary.

For the Model, we chose to use generalized linear mixed model. There few reasons for it to be considered here:

- Since we could group customers by CSDuid, so a random effect should be considered.
- We would use status as our response variable and status is binary, which is discrete, so a linear mixed model can not be applied here.

A general form of logistic regression with generalized linear mixed model is (Without subscript):

$$Y \sim Binomial(N, \mu)$$

$$\log(\frac{\mu}{1-\mu}) = X\beta + U$$

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

where:

- Y is the binomial response variable
- X is the matrix of predictors
- β coefficient matrix of predictors
- U is the random effect part follow normal distribution

We chose to try logistic regression and test two different sets of predictors.

The first model we fit is: **mod_log**

- status is the response variable
- age (rescaled), skin color, sex, and household median incomes (rescaled) are the predictors
- A random intercept of CSDuid is introduced

After fitting the model: **mod_log**, we obtain following results of p-values:

Table 2: Table of Variables in Full Model and Corresponding P-values

Variables	P-Values
Intercept	1.42e-11 (< 0.001)
age(rescaled)	1.66e-08 (< 0.001)
Default tone skin color	0.853
Light tone skin color	0.654
Medium tone skin color	0.831
Medium dark tone skin color	0.395
Medium light tone skin color	0.797
Sex Male	0.216
Household median incomes(rescaled)	1.70e-12 (< 0.001)

Based on the p-values, and 5% significant level standard:

- We have strong evidence that household median incomes (rescaled) is significant to our model.
- We have strong evidence that age (rescaled) is significant to our model.
- We have no evidence that sex and skin-color are significant to our model.

Thus, sex and skin color, regardless of the class, neither feature is significant at significant level, which means that the two features will not be affected by product changes.

The second model we fit is: **mod_log_1**, it is the reduced version of **mod_log**, we only saved significant variables: age (rescaled) and household median incomes (rescaled) as our fixed effect.

Since these two models are nested on fixed effects, so we decided to use likelihood ratio test for comparison.

The p-value of the test is 0.7553, we have no evidence that the more complicated model: **mod_log** should be used. Therefore, the simpler model: **mod_log_1** should be used.

Final Model

Finally, we have use **mod_log_1** as the final model

$$\log(\frac{\mu_i}{1-\mu_i}) = \hat{\beta}_0 + \hat{\beta}_1 X_{Age(scaled)i} + \hat{\beta}_2 X_{Median\ Income(scaled)i} + U_i$$

$$U_i \sim N(0, 0.2211^2)$$

where:

- μ_i is the probability of a customer is defined as new customer
- $X_{Age(scaled)i}$ is the fixed effect which is the age (rescaled) of customer i
- $X_{Median\ Income(scaled)i}$ is the fixed effect which is the household median incomes (rescaled) of customer i
- U_i is the random intercept of CSDuid which follows a normal distribution

Model Diagonstics

Generalized Linear Mixed Model has four assumptions:

- Our units are independent, but the observations within each unit are taken not to be.
- Random effects come from a normal distribution.
- The random effects errors and within-unit errors have constant variances.
- The link function is correctly chose.

Refer to our final model, here are the results of checking four assumptions:

- We separated groups based on CSDuid. Therefore, observations between each census division are independent, while the observations in the same census division would be dependent.
- With current knowledge of generalized linear mixed model, it is hard to assess the second and third assumption, so we assumed they are satisfied here.
- The link function we chose it the log function which is commonly used in the logistic regression with generalized linear mixed model. Thus, it is appropriate to be used.

Final Model Interpretation

Table 3: Table of Values, Exponentialed Value, and 95% Confidence Interval of Coefficients of Variables in Final Model

Coefficient	Corresponding Variable	Value	Exponentialed	95% CI
$\hat{\beta}_0$ (Intercept)	Fix effect	0.671	1.956	(1.635, 2.360)
$\hat{\beta}_1$	Age(recale)	0.370	1.448	(1.274, 1.647)
$\hat{\beta}_2$	Income(recale)	-2.652	0.071	(0.033, 0.148)

From the table of coefficients above, we have that:

- $\hat{\beta}_0$ (Intercept): 1.9561 is the odd of a 18 years old customer being our new customer.
- $\hat{\beta}_1$: While keeping household median incomes fixed, the odds of being a new customer would increase by 44.828%, when a customer's age increases from 18 to 92,
- $\hat{\beta}_2$: While keeping age fixed, the odds of being a new customer decreased by 92.95%, when a customer's income increases from 41880 to 195570.

Check difference between old and new customer and disscusion

To find more accurate differences between new customer and old customer, we decided to fit a simple linear regression between fixed effects and status. A typical simple linear regression model with only one dummy variable is as the following:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Where:

- Y is the response variable
- β_0 is Intercept
- β_1 is coefficient of X, greater than 0 implies positive effect
- X is indicator variable (dummy variable), either be 0 or 1
- ϵ is the residual term

Here is the summary table of coefficients we obtained from two models:

Table 4: Table of Coefficients and Estimated Values of Linear Models of Age vs Status and Household Median Incomes vs Status

Linear Model	Intercept	$\hat{\beta}_1$
age ~ status	46.516	1.4356
hhld_median_inc ~ status	73182.000	-4364.1000

As showed in table 4, we could conclude the following:

- The average age of customers who bought the new product is 1.4356 unit higher than that of customers who bought other products.
- The average median income of customers who bought the new product is 4364.1 unit lower than that of customers who bought other products

Counclusion

In conclusion, according to the previous results, we can know that the new product launched is for the new customer composed of less income and older people. In this data set, low income refers to low-income areas, so the result shows low income area prefer to it. This does not conflict with our conclusion that the purchasing power of low-income areas increases due to the purchase of low-income people. The skin color and sex are not part of the influence of the new product, that is to say, they have nothing to do with the new product. Compare with others, new customer with younger age also prefer the product. Therefore, Mingar's new products should focus on people with less income and older age, but we can't ignore the young customers.

Investigation on the relationship between skin colors and sleeping scores

Dataset Description and Glimpse

For this research question, the primary dataset we are using is: **customer_sleep_data**. The **customer_sleep_data** is constructed by joining other datasets given or scraped to **cust_sleep** dataset. It is a dataset that contains 20181 observations with 953 customers.

For research convenience and realistic interpretation purposes, we made the following changes:

- We calculated actual ages based on given birth dates for each customer. Then, we rescaled ages by subtracting the minimum age and dividing the difference between the maximum

and minimum ages. Therefore, the ages are in a range between 0 and 1. 0 indicates the minimum age, which is 18 years old, and 1 indicates the maximum age, which is 92 years old.

- We rescaled household median incomes data by subtracting the minimum and dividing the difference between maximum and minimum. Therefore, the household median incomes are in the range between 0 and 1. 0 indicates the minimum, and 1 indicates the maximum among all customers.
- We estimated customers' skin colors based on the emoji modifier provided, stored the estimations in the column: **skinColor**, and the levels are kept the same as in the emoji modifier: dark, medium-dark, default, medium-light, and light. The default skin color is the most special one because there are chances that some customers didn't set up skin tones for their emoji modifiers. Therefore, some default skin customers may belong to different skin colors, rather than the default skin color.
- We removed some columns that are either unnecessary or include duplicate information, such as **Pronouns** (to **sex**, it is duplicated information).
- Any recorded observation that involves **NA** is removed.
- Since the proportion of intersex customers is relatively small: 1% (222 observations) out of 19241 distinct customers, customers who are identified as intersex gender are removed,

Finally, there are ten columns in total: cust_id, duration, flags, sex, device_name, line, CSDuid, hhld_median_inc, age, skinColor

CSDuid, **hhld_median_inc**, **age**, **skinColor** are four columns were not in given datasets, the descriptions of them are as following (in the order of appearance):

Table 5: Table of New Columns in sleep_data Which Were Not in Given Datasets

Column Name	Description
CSDuid	Census subdivision unique identifier from Canadian census
hhld_median_inc	Household median income from Canadian census
age	Ages of customers, calculated based on birth date with respect to 2022
skinColor	Estimated skin color of the customer based on emoji modifier provided
age_scaled	Ages of customers, rescaled to values in 0 to 1
income_scaled	Household median income, rescaled to values in 0 to 1

Since we are interested in the sleep scores of the customers, and they are not directly provided, we chose flags as our main research target variable. Flags are the number of times there was a quality flag during the sleep session. Therefore, a higher number of flags during one sleep session would refer to a lower sleep score. We looked into some basic statistics of the **sleep_data** dataset.

Table 6: Table of Mean, Maximum, and Minimum of Number of Flags

Mean	Max	Min
4.32159	26	0

Table 7: Table of Mean of Flags with Respect to Different Skin Colors

Skin Color	Mean of Flags
Dark	11.783119
Default	2.473643
Light	1.154972
Medium	3.658228
Medium-dark	7.442418
Medium-light	2.510911

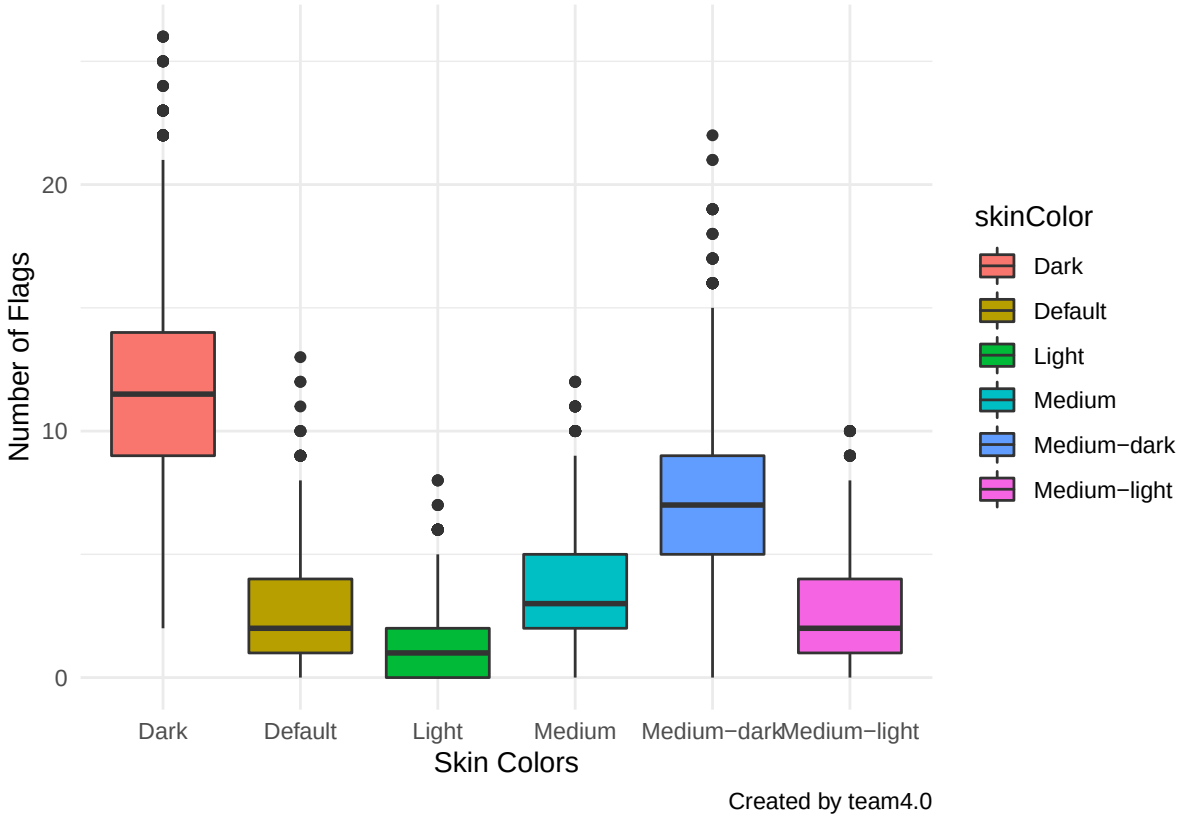

Figure 9: Box Plots of Total Number of Flags with Respect to Different Skin Colors in sleep_data

According to table 6, the mean flags of both dark-skin customer and medium-dark skin customers is higher than that of other customers with lighter skin colors.

According to figure 9, the maximum number of flags is reached by dark skin customers, while the minimum is reached by light skin customers. Also, the median and mean flags of dark skin customers are taking the lead. However, there are multiple observations for one customer, so this box plot is not enough to draw an accurate conclusion. Indeed, the box plot gives us the most basic idea of the result. Thus, we decided to set up models to better investigate the relationship between flags, skin colors, and other factors.

Model Setup and Selection

As mentioned in the dataset introduction, we chose flags as our response variable since it is closely related to sleep scores. More flags would indicate a worse sleeping quality as well as lower sleep scores. For predictors, we chose skin color, age (rescaled), product of line of the current devices, household median incomes of current customer's living area (rescaled), and sex. It would be biased if we include skin color as the only predictor, so we would also want to investigate other factors, which might also affect flags/sleep scores.

The model we decided to use is the Poisson regression with generalized linear regression model. There are two main reasons why we chose it:

- Since there are multiple observations for one customer, so one person could be treated as one group. CSDuid is another potential grouping factor, since customer id can be enough to cover all the groups, the feature CSDuid is unnecessary. The observations between customers are completely independent by default. Meantime, the observations of one customer are dependent. Therefore, a random effect of customer id should be considered, and this leads to the choice of generalized linear mixed model.
- The response variable, number of flags, is the number of times there was a quality flag during the sleep session, which is a non-binary discrete variable. Therefore, logistic regression is not a proper choice for this model. Data of flags follows a Poisson distribution by the definition of count data, therefore, it determines the type of the regression of our model: Poisson.

There are two assumptions of generalized linear mixed model that are already satisfied. However, an offset term should be added to our model since the duration of sleep varies from observation to observation and customer to customer.

A typical Poisson regression (including offset term and link function log) with generalized linear mixed model is in the following format:

$$Y \sim Poi(\mu)$$

$$\log(\mu) = \log(T) + X\beta + U$$

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

where:

- Y is the response variable
- $log(T)$ is the offset term
- X is the matrix of indicator variables of fixed effect, which we are interested in
- β is the matrix of coefficients
- U is the random effect

The initial model we came up with is the full model: **mod_poi_1** which includes all the features we are interested in as predictors. They are: skin color, age, product line, household median income, sex, a random intercept of customer id, and an offset term which is the log of duration of sleep.

The summary of the variables and p-values of the full model is shown below:

Table 8: Table of Variables in Full Model and Corresponding P-values

Variables	P-Values
Intercept	<2e-16 (<0.001)
Default tone skin color	<2e-16 (<0.001)
Light tone skin color	<2e-16 (<0.001)
Medium tone skin color	<2e-16 (<0.001)
Medium dark tone skin color	<2e-16 (<0.001)
Medium light tone skin color	<2e-16 (<0.001)
age(rescaled)	0.00988

Variables	P-Values
Line: Advance	0.12548
Line: iDOL	0.33913
Line: Run	0.04929
Household median incomes(rescaled)	0.77596
Sex Male	0.56197

Based on the P-values, and the 5% significant level standard:

- We have very strong evidence that skin color is significant to our model.
- We have strong evidence that age (rescaled) is significant to our model.
- We have no evidence that household median incomes and sex are significant to our model.
- Since Run product line has a P-value of 0.04929, we decided to keep it. Then we designed other two models based on the above interpretation on P-value:

1. **mod_poi_2**: Skin color, age (rescaled), line, a random intercept of customer id as predictors, and an offset term which is the log of duration of sleep.
2. **mod_poi_3**: Skin color, age (rescaled), a random intercept of customer id as predictors, and an offset term which is the log of duration of sleep.

Finally, since three models are only varied and nested on fixed effects, we used the likelihood ratio test for comparison between the three models.

Table 9: Table of P-values of Likelihood Ratio Tests Between Three Models

Test	P-value
mod_poi_1 vs mod_poi_2	0.7843
mod_poi_2 vs mod_poi_3	0.2157

According to the likelihood ratio test, we have very strong evidence that **mod_poi_3**: skin color + age + random intercept of customer id with an offset term which is the log of duration of sleep is better than the other two models.

Final Model

Based on various test results, the final model is: **mod_poi_3**

$$\log(\mu_{ij}) = \log(Dij) + \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{\beta}_3 X_{i3} + \hat{\beta}_4 X_{i4} + \hat{\beta}_5 X_{i5} + \hat{\beta}_6 Age_i + U_i$$

$$Y_{ij} \sim Poi(\mu_{ij})$$

$$U_i \sim N(0, 0.05094^2)$$

where:

- μ_{ij} is the average number of flags per minute for person i in observation j.
- D_{ij} is the offset term: log duration of sleep of person i in observation j.
- X_{ij1} is a indicator variable that equals to 1 when person i has default skin color (estimated based on emoji modifier), 0 otherwise in observation j.
- X_{ij2} is a indicator variable that equals to 1 when person i has light skin color (estimated based on emoji modifier), 0 otherwise in observation j.
- X_{ij3} is a indicator variable that equals to 1 when person i has medium skin color (estimated based on emoji modifier), 0 otherwise in observation j.
- X_{ij4} is a indicator variable that equals to 1 when person i has medium dark skin color (estimated based on emoji modifier), 0 otherwise in observation j.
- X_{ij5} is a indicator variable that equals to 1 when person i has medium light skin color (estimated based on emoji modifier), 0 otherwise in observation j.
- Y_{ij} is the number of flags of person i in observation j.
- U_i is the random effect of person i.
- $\hat{\beta}_i$ is the coefficient of variables.

Model Diagnostics

Generalized Linear Mixed Model has four assumptions:

- Our units are independent, but the observations within each unit are dependent.
- Random effects come from a normal distribution.

- The random effects errors and within-unit errors have constant variances.
- The link function is correctly chosen.

Refer to our final model, here are the results of checking four assumptions for the Generalized Linear Mixed Model:

- As mentioned in the previous section, we separated groups based on customer id. Therefore, it is natural that the observations between each group are independent, while the observations in the same group would be dependent. Assumption 1 is satisflied.
- With current knowledge of generalized linear mixed model, it is hard to assess the second and third assumptions, so we assumed they are satisfied here.
- The link function we chose is the log function which is commonly used in the Poisson regression with generalized linear mixed model. Thus, it is appropriate to be used.

Final Model Interpretation

Table 10: Table of Values, Exponentialed Value, and 95% Confidence Interval of Coefficients of Variables in Final Model

Coefficient	Corresponding Variable	Value	Exponentialed	95% CI
$\hat{\beta}_0$ (Intercept)	Dark Skin Color	-3.383112	0.033	(0.033, 0.035)
$\hat{\beta}_1$	Default Skin Color	-1.631439	0.196	(0.191, 0.200)
$\hat{\beta}_2$	Light Skin Color	-2.389198	0.092	(0.089, 0.095)
$\hat{\beta}_3$	Medium Skin Color	-1.212070	0.298	(0.290, 0.305)
$\hat{\beta}_4$	Medium Dark Skin Color	-0.500260	0.606	(0.593, 0.620)
$\hat{\beta}_5$	Medium light Skin Color	-1.612974	0.199	(0.194, 0.205)
$\hat{\beta}_6$	Age (rescaled)	-0.047375	0.954	(0.921, 0.987)

Here, we would use exponentialed values of betas to better interpret results:

- $\hat{\beta}_0$ (Intercept): 0.3394166 is the expected number of flags per minute of 18 years old dark skin customers.
- $\hat{\beta}_1$: While keeping age fixed, the number of flags per minute of default skin customers would be 0.19565 times that of dark skin customers.

- $\hat{\beta}_2$: While keeping age fixed, the number of flags per minute of light customers would be 0.09170 times that of dark skin customers.
- $\hat{\beta}_3$: While keeping age fixed, the number of flags per minute of medium skin customers would be 0.29758 times that of dark skin customers.
- $\hat{\beta}_4$: While keeping age fixed, the number of flags per minute of medium dark skin customers would be 0.60637 times that of dark skin customers.
- $\hat{\beta}_5$: While keeping age fixed, the number of flags per minute of medium light skin customers would be 0.19929 times that of dark skin customers.
- $\hat{\beta}_6$: The ratio of average number of flags per minute of 18 years old customer (minimum age) and 92 years old customer (maximum age) is 0.954.

Conclusion

Through our data modeling and analysis, while age is fixed, customers with darker skin would tend to have lower sleep scores caused by a higher number of flags compared to customers who have lighter skin. However, age would be another influential factor for sleep score. From the analysis of our model, young customers tend to have lower sleep scores.

Discussion

In question 1's data preparation, we removed missing and unnecessary data from the POSTcode dataset and the Customer Sleep dataset. We then rescaled the age and income columns for a more convenient analysis. Based on the clean datasets, we created histograms and boxplots to show the age distribution of both old and new products. It's shown that new products tend to capture more people within the age range of under 25 years old and above 60 years old. Also, the age distribution for the new product is not as concentrated as the old product, which further shows that the new product has a wider market with respect to customers' age. From the histogram of the household median income distribution of both old and new products, we observe that customers living in areas with income below 50000 canadian dollars have increased, which means that the market share of lower-income people has increased. Based on the interpretation of the p-value, we have chosen the household median income and age as the significant variables, and we chose the logistic regression model to fit the data. At last, we used a linear regression model to check the influence of age and income on the purchase count. From our analysis, the new product with cheaper price attracts people of younger age and older age and people with less purchasing ability.
In question 2, from the customer_sleep_data dataset, we rescaled and removed some columns

that are least important or missing, and kept ten columns for further investigation. Based on the column of flags, we proceeded with some statistical analyses. We calculated the maximum, minimum, and mean of the number of flags, and made a boxplot on the number of flags for customers with different types of skin colors. From the boxplot, it's clearly shown that people with darker skin color tend to have more flags during sleeping.
Since the number of flags is a number count in a sleep session and it's not binary, and observations between customers are independent, we decided to use Poisson regression with generalized linear regression model. From comparing the P-values of the variables in the full model, we concluded that skin color, age, and run product line are significant to our model. Through comparing models by likelihood ratio test and checking the assumptions, our final best model is Skin color, age, and a random intercept of customer id as predictors. In conclusion, with fixed age, the number of flags per minute of darker skin and younger customers would be more than lighter skin and older customers respectively, therefore, it leads to a poor sleeping score.

Strengths and limitations

Strength:

- Estimated result with box plot.
- Other than skin color, we found another potential factor that would affect sleeping scores: age.
- Model that includes random effect is more robust to the correlation problem

Limitation:

For Question 1:

- For the assumption of our generalized linear mixed model for question 1, we can't make sure that all the assumptions are satisfied. Not meeting the assumptions might affect the accuracy of our model.
- If we use the mean value to fill the NA values, the result might be more reasonable and capture better information.
- The result obtained from modeling may be different from our graph interpretation.

For Question 2:

- We ignored intersex customers, therefore leading to a loss of data.

- The data does not perfectly satisfies the assumptions, we assumed the second and the third of the generalized linear mixed model are satisfied without testing them, so our model might not be suitable for our data.
- Number of Flags is probably not a determining variable for sleeping quality, they might be raised due to the device's functionality issue.
- There is an overdispersion in flags data, since the variance is 17.84583, and is greater than the mean: 4.32159. A complicated model might be chosen at the end due to large variance.

Consultant information

Consultant profiles

Yizhou Peng. Yizhou is a senior data scientist with team4.0. She specializes in data visualization. Yizhou earned her Master of Mathematical&Industrial engineering, Majoring in data analytics from the University of Toronto in 2023.

Zihao Zeng. Zihao is a senior consultant with team4.0. He specializes in artificial intelligence. Zihao earned his Master of Mathematical&Industrial engineering, Majoring in data analytics from the University of Toronto in 2023.

Jiahao Bai. Jiahao is a senior consultant with team4.0. He specializes in Machine Learning and Cognitive algorithm development. Jiahao earned his Bachelor of Science, Majoring in Statistics from the University of Toronto in 2023.

Code of ethical conduct

Responsibility to stakeholders, including clients and employers:

- Protects customers' personal physical condition data or other personal information, customers' personal information data will not be used for personal commercial gain or benefit from a third party. Any use of clients' information will go through the formal permission of the client.
- Explains the use and potential limitations of Mingar's devices in different contexts to stakeholders, and provide guidance and alternatives from the perspective of cost and scope to stakeholders.

Responsibility to society, including competitors :

- Collects data from all groups of nations and all gender types; ensures that human rights are protected and takes related concerns seriously.
- Recognizes and respects that other fitness tracker companies may have similar target groups of clients. Encourage new entrants to the fitness device space to bring better products to the space, and avoids behaviors that would harm society.

References

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

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

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Fitness tracker info hub. (2022). Fitness tracker info hub. Retrieved April 6, 2022, from https://fitnesstrackerinfohub.netlify.app/

Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL https://www.jstatsoft.org/v40/i03/.

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

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.

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/.

Unicode, Inc. (1991). Full emoji modifier sequences, V14.0 - unicode. Full Emoji Modifier Sequences, v14.0. Retrieved April 6, 2022, from https://www.unicode.org/emoji/charts/full-emoji-modifiers.html

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

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.

Appendix

Web scraping industry data on fitness tracker devices

The sleeping data of fitness tracker devices is scraped from the website: https://fitnesstrackerinfohub.netlify.app/, and the data is stored in a database and is converted into a data frame. We have provided a User-Agent string that makes our intentions clear and provided our email address for possible concerns and questions. Any details provided in the robots.txt on crawl delays and we are allowed to scrap on this website. Some considerations of web scraping include ensuring we make new value out of the data instead of just duplicating it and citing the R packages we used in this part in the reference section.

```
# Scraping the industry data
url <- "https://fitnesstrackerinfohub.netlify.app/"

# Get scrap target, provide relevant contact information
target <- bow(url,
              user_agent = "jiahao.bai@mail.utoronto.ca",
              force = TRUE)

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

# Obtain data, and convert to data frame
html <- scrape(target)
device_data <- html %>%
  html_elements("table") %>%
  html_table() %>%
  pluck(1)
```

Accessing Census data on median household income

The data on median household income is scraped from the 2016 census (2020 not up yet), and we have converted it into a data frame. Then, for the data preparation step, we simplified the data into fewer variables, CSDuid, median income, and population, so we are able to better analyze the problem with neat data. We have cited the R packages we used in this part in the reference section. Some considerations include that since we are provided a public API, we can use it, so we won't scrap together. Through scraping, we want to make sure that the web scrapping is allowed and we assess data at a reasonable rate.

```
# Scraping census data
options(cancensus.api_key = "CensusMapper_b041be566f3982329277d23be6e52e8d",
        cancensus.cache_path = "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") %>%
  as_census_region_list()

# Get the 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)
```

Accessing postcode conversion files

The postal code data has been provided and is loaded into a dataset. We will keep the data for our own investigation purpose and will not let the data be publicly accessed. We will use the data to create connections between postal code and census and develop our investigation based on the data analysis, and we will credit you in our report or article. Our team will respond in a timely manner if there is any concern. As mentioned above, we have cited R packages in the reference.

```
# Load postal code data
dataset = read_rds("data-raw/break_glass_in_case_of_emergency.Rds")

postcode <- dataset %>%
  select(PC, CSDuid)
```