MINGAR: Products & Customers Analysis

An investigation of new and traditional customers and issues in device performance

Report prepared for MINGAR by Magic Conch

2022-04-07

Contents

Executive Summary

Mingar, a company that develops fitness tracking wearable devices, is interested in the characteristics of their customers, and device performance issues around their new lines of wearables fitness trackers/smartwatches – "Active" and "Advance" lines. More specifically, the marketing team in the company intends to understand new customers who have bought devices from the recently added lines and investigate any existing differences between the new and traditional customers. Additionally, the company aims to investigate any significant relationship between the defective devices' performances, particularly with respect to sleep score, and the skin colours of all of their customers. This study performs analyses to explore the characteristics of the customers from the available data and develops models to investigate factors associated with purchase preferences and factors related to sleep quality flags.

The key results of the study are summarized below.

- The data consists of 8,476 old customers who choose to purchase the "Run" and "iDOL" lines and 10,569 new customers who choose to buy the newly added lines: "Active" and "Advance".
- According to Table1, the new customer market is dominated by female customers, who comprise about 57.8%. Also, there are 16.9% fewer male customers than female customers in the new customer market.
- According to Table 1, there are 2,093 more new customers than traditional ones, and new customers are on average 2 years older than the traditional customers; the average population for the new customers is 41,315 higher than the average population for the traditional customers; in addition, the average median income of new customers is $4,354 lower than the average income of traditional customers.
- There are significant increases in the number of dark, medium-dark, and medium-skinned new customers; and there are more new customers for all three genders: Male, Female, and Intersex.
- Customers with minimum age (17 years old) and maximum age (92 years old) preferred "Active" and "Advanced" products. The elder customers are more likely to become new customers.
- A significant relationship was identified between skin colour and sleep score errors, which users with darker skin are more likely to encounter sleep score errors than users with other skin colours.

The limitations of the analysis are below.

Table 1: Charactersitics of New and Old Customers

Customer Type	Customer Number	Avg Age	Avg Population	Avg Median Income	Female	Male	Line
Old	8476	46	1478529	73168	58.6%	40.3%	Run/iDOL
New	10569	48	1519844	68814	57.8%	40.9%	Active/Advanced

Table 2: The Distribution of Skin Color

Skin Color	Num Clients	Proportion
Dark	2666	0.13
Default	5169	0.25
Medium	2942	0.14
Medium-Light	3138	0.15
Light	3658	0.18
Medium-Dark	2840	0.14
NA's	210	0.01

- Since the skin colour is a sensible data and some people refuse to provide it, around $\frac{1}{4}$ of data is missing skin colour. Therefore, this data may not fully represent the whole population.
- This sample size is reduced after removing all missing values, resulting a reduced data set. This implies that the conclusion we drew might not be applicable to all customers.
- There might exist some confounding variables that are not included in the client information data set, such as the health condition. For example, clients suffering from high anxiety might have worse sleeping quality and irregular sleep duration, therefore would have an impact on the client sleep scores.

In conclusion, purchasing decision (traditional devices/devices from recently added lines) is related to customers' ages and dark skin leads to a high probability of sleep quality flags.

Technical Report

Introduction

Mingar, a company developing fitness tracking wearable devices, currently expanded product offerings by adding two lines – "Active" and "Advance" – in order to be more competitive in the fitness tracker market. The marketing team intended to understand new customers who have bought devices from the newly added lines and investigate the differences between new and traditional customers. Additionally, the market team was interested in whether the expanded product lines have attracted customers outside the traditionally higher-income base. Meanwhile, the social media team noticed defective devices' performance, particularly with respect to sleep score, for users with darker skin.

Research Questions

- Who are our new customers?
  - Identifying who are new customers and their corresponding characteristics for new customers.
- What are the differences between new and traditonal customers?
  - Compare the difference between old and new customers by their characteristics.
- How skin colors affect the sleep quality flags?
  - Identify the proportion skin color.
  - Investigate the dark skin color customers and their sleep quality flag.

Data Prepration

By merging the customer id with their income by their corresponding customer id, we could analyze the differences between new and traditional customers. Then we re-scaled age, population and income to the range of minimum 0 and maximum 1, and used them in our model to make their values meaningful and to avoid non-convergence.

To match the customer id with their corresponding skin color and analyze the factor affecting sleep quality, we merged the customer information with the sleeping data by the customer id.

As always we took stock of the amount of data, after merging, there were 20622 observations. Large sample sizes were preferred for the type of model we would consider, and n = 20622 was

considered as large. However, to be noted, there were 5169, around $\frac{1}{4}$ of clients who set default as their choice on skin color.

Analysis 1: Who are our new customers?

First of all, the report aimed to discover who are the new customers. The new customers of the MINGAR company are the ones who chose to purchase the recently added lines: "Active" and "Advance", which were denoted as "New" under the newly added column in the data set. To gain better insights into the new customers, we filtered out the data of new customers in the data set and then examined factors including sex, skin color, age, population, median income, and new products chosen by using this data set.

SEX: The sex percentage of three categories: Female, Male, and Intersex was calculated by keeping 3 decimal places. The sex information could be visualized in a pie chart with the percentage format (multiplying 100). From the pie chart, we noticed that female customer accounted for 57.8% of our new customers, whereas intersex accounted for only 1.3%. The remaining 40.9% comprised male customers. Thus, we concluded that the majority of the MINGAR company's new customers were females, which was about 17% larger than the male group.

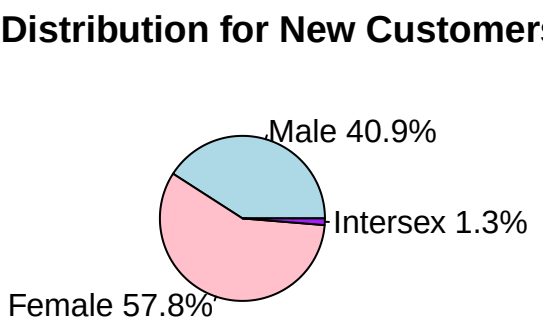

Figure 1: Pie Chart: Sex Percentage For New Customers(n=10569)

SKIN COLOR: In order to investigate the skin color among the new customer, we calculated each skin color's percentage by keeping 3 decimal places and saved skin color's percentage in a table. Each skin color's percentage data in the skin information data set was demonstrated in a pie

Table 3: Average Age, Population, Median Income

avg age	avg population	avg median income
48	1519844	68814

chart with the percentage format (multiplying 100). Then, we discovered an interesting situation from the pie chart, whereby the default option constituted the largest portion (25.1%), meaning that most new customers did not comfortable with sharing their skin color information.

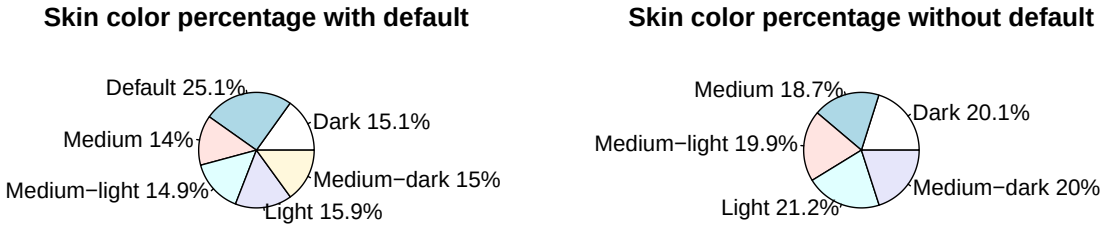

Figure 2: Pie Chart: Skin Color Percentage For New Customers(n=10569) With Default and Without Default

After observing that, we removed the new customers that did not fill out the information about their skin color, and calculated the percentage of the remaining skin colors: Light, Medium-light, Medium, Medium-dark, and Dark by calculating the number of each color type over the sum of five skin color types. Then, these new skin colors' percentages without the default values were shown in the pie chart. Considering that the pie chart was almost evenly divided into Light, Medium-light, Medium, and Medium-dark with approximately 20% each, we concluded that there was little difference in proportion of skin colors among all new customers.

Average Age, Population, Median Income: we achieved the original values of average age, average population, and average median income, which were 48 years old, 1519844 people, and 68814 dollars respectively.

Analysis 2: What are the differences between new and traditional customers?

Compared to the traditional devices, the new devices from the expanded lines, "Active" and "Advance", generally had more functionalities with better prices.

Table 4: New Products

Line	Recommended retail price	Battery life	Water resitance	Released	Brand	Num of function
Advance	145.00	Up to 7 days	Resistant	2021-07-08	Mingar	5
Active	99.99	Up to 7 days	Resistant	2020-12-30	Mingar	4
Advance	120.00	Up to 7 days	Resistant	2020-08-20	Mingar	5
Active	39.99	Up to 14 days	Resistant	2019-10-13	Mingar	0
Active	79.99	Up to 7 days	Resistant	2019-10-13	Mingar	1

Table 5: Old Products

Line	Recommended retail price	Battery life	Water resitance	Released	Brand	Num of function
Advance	145.00	Up to 7 days	Resistant	2021-07-08	Mingar	5
Active	99.99	Up to 7 days	Resistant	2020-12-30	Mingar	4
Advance	120.00	Up to 7 days	Resistant	2020-08-20	Mingar	5
Active	39.99	Up to 14 days	Resistant	2019-10-13	Mingar	0
Active	79.99	Up to 7 days	Resistant	2019-10-13	Mingar	1

GENDER: The bar graph indicated that there were more new customers for all three genders: Male, Female, and Intersex. Moreover, the majority of customers are females for both new and old customer populations.

SKIN COLOR: The bar graph compared 6 skin types for both old and new customers: Dark, Default, Medium, Medium-light, Light, and Medium-dark. We observed from the graph that overall there were more new customers among all skin colours. Other than the default, for both old and new customers, the majority of customers were light skins, and the second-largest population was medium-light. Compared to the old customers, there were significant increases in counts for dark, medium-dark, and medium skin colours in new customers. “Active” and “Advance”, generally had more functionalities with better prices.

Table 6: The number, average age, and median income of new and old customers

Customer Type	num of customers	avg age	avg population	avg median income
Old	8476	46	1478529	73168
New	10569	48	1519844	68814

Table 7: Number of Old and New Customers

Customer Type	Counts
Old	8476
New	10569

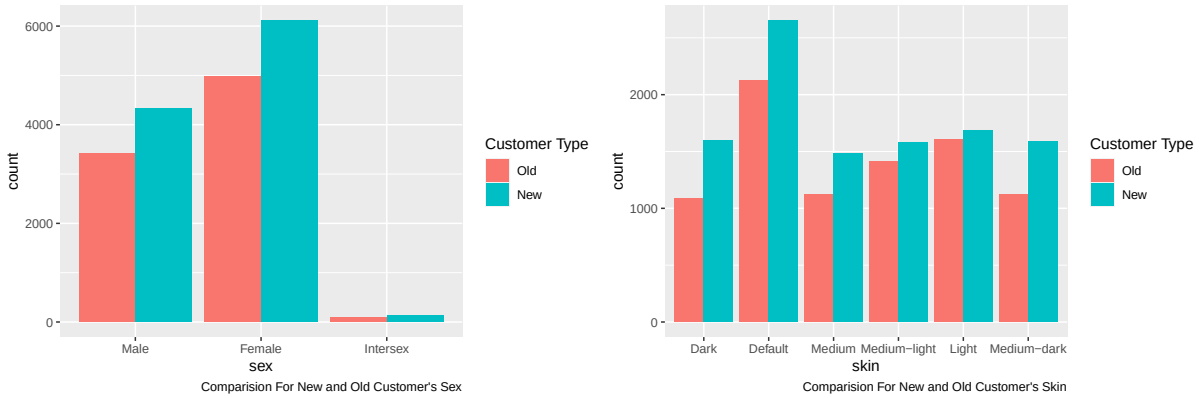

Figure 3: Comparision For New and Old Customer's Sex and Skin

Noticed that we re-scaled the age and population to the range of minimum 0 and maximum 1, and used them in our model to make their value become meaningful and avoid non-convergent.

Then, we compared the new customer with the old customer. There were 19036 participants in our study, 8476 of whom were old customers and 10569 of whom were new customers.

Then, we conducted a comparison of the difference between the old customer and the old customer. There were 19036 participants in our study, of which 8476 of participants were old customers and the rest 10569 participants were new customers. Based on the data set, there were five factors for us to consider: skin, sex, age (re-scaled), population (re-scaled), and median income (re-scaled). First of all, we tried to build a generalized linear mixed model (GLMM) to explore the difference between the new and old customers (response variable) based on all of these factors, since the mixed effects existed among these factors and the response variable was a binomial distribution. However, it was not hard to observe that there was a high correlation between the population (re-scaled) and median income (re-scaled) in the graph, so they could not both exist in the same model to prevent the model from becoming non-convergent.

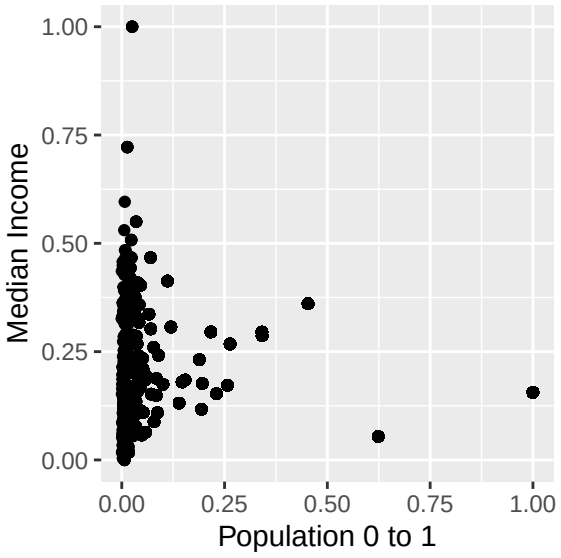

Figure 4: The graph shows the highly correlation between populationa and median income

Then all of the factors skin, sex, age (re-scaled), and median income (re-scaled) were included as predictors, and new or old customers were included as responses. For fixed effects, the mixed effects could be categorized as skin color, sex, and age (re-scaled), and median income for random effects. Specifically, the model could be presented as follows:

Model 1

$$Y_{ijkl} \sim Bernoulli(\rho_{ijkl})$$

$$logit(\rho_{ijkl}) = \mu + X_{ijkl}\beta + U_i$$

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

, where Y_{ijkl}: Being new or old customers had skin color k with median income i at age j and sex l. X_{ijkl}: indicator variables for age, skin color and sex. U_i: each median income random effect.

After observing this model, we found that only the result for the age(re-scaled) was significant since its p-value was $8.08e-09 \leq 0.05$. Then, we tried to build a simpler model with only age(re-scaled) and population as predictors and new/old customers as responses. To be more specific, the model could be represented as the following:

Model 2

$$Y_{ij} \sim Bernoulli(\rho_{ij})$$

Table 8: Comparison between two models

#Df	LogLik	Df	Chisq	Pr(>Chisq)
10	-12847.95	NA	NA	NA
3	-12850.47	-7	5.049033	0.6539795

Table 9: Model 2 Summary

	Estimate	Std. Error	z value	Pr(>|z|)
Baseline(log)	0.1390077	0.0555527	2.502267	0.0123401
Log Odd Ratio	0.3807597	0.0658897	5.778740	0.0000000

$$logit(\rho_{ij}) = \mu + X_{ij}\beta + U_i$$

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

, where Y_{ij}: Being new or old customers with median income i at age j. X_{ij}: indicator variables for age. U_i: each median income random effect.

From this model, we saw that the result for the age (re-scaled) was statistically significant as well since its p-value was $7.44e-09 \leq 0.05$. The ML test was used to compare the previous two models in order to find a better model. The null hypothesis was that a simpler model (Model 2) explained the data just as well as the more complicated one (Model 1). We chose the simpler model (Model 2) since the p-value was > 0.05, therefore there was no evidence against the null hypothesis.

This Model 2 investigated the association between the age (Re-scale: 0-1) and the new or old customers. The baseline odds: 1.148 in this situation represented the odds for the customer who had the minimum age (17 years old), which was achieved by applying the exponential function on the log odds. Based on the probability formula $\frac{odds}{1+odds}$, we estimated that people with a minimum age of 17 years old have a 53.5% chance to be new customers, while people with an age of 17 years old would be 7% less likely to be traditional customers. Then, the odds ratio between maximum age and the minimum age was shown as 1.463, which could be gained by applying the exponential function on log odds of 0.381. Thus, based on the odds ratio formula

Table 10: Estimates(odds) and confidence interval

	Estimate	95% CI
Baseline(MIN) Odds	1.149	(1.03,1.28)
Odds Ratio(MAX/MIN)	1.463	(1.29,1.67)

Table 11: Odds of Min and Max Age

	Odds
Min Age: 17	1.149000
Max Age: 92	1.680987

Table 12: Probability of Being New and Traditional Customers

	Probability of new customer	Probability of old customer
Min Age: 17	0.535	0.465
Max Age: 92	0.627	0.373

$\frac{odds\ \ for\ \ maximum\ \ age}{odds\ \ for\ \ minimum\ \ age} = 1.463$, the odds for new customers who have the maximum age (92 years old) was about 1.681. Then, according to the formula $\frac{odds}{1+odds}$, we discovered the probability that people with a maximum age of 92 years old would be new customers at 62.7%, which was about 25.4% more than the probability of having traditional customers with maximum age 92 years old. Therefore, we could conclude that both the younger customers with minimum age and the elder customers with maximum age preferred to become new customers who were more affordable "Active" and "Advanced" products. Besides, since the probability of being a new customer at 17 years old was smaller than that of being a new customer at 92 years old, then we concluded that older people were more likely to be the new customers in the Mingar company.

Analysis 3: How skin colors affect the sleep quality flags?

We fitted variables in glmer model since a person could have multiple sleeping records. In conclusion, poisson regression was a good first choice because the response variables were counted per year.

Customer id had no impact on the color of the skin. Since people had multiple records affecting the initial value of the flag, Customer ID was treated as a random intercept in the model. We

Table 13: The distribution of skin color

	Clients Counts	Proportion
Dark	2666	0.1292794
Default	5169	0.2506546
Medium	2941	0.1426147
Medium-light	3138	0.1521676
Light	3658	0.1773834
Medium-dark	2840	0.1377170
NA's	210	0.0101833

Table 14: Comparison between model 1 and model 2

#Df	LogLik	Df	Chisq	Pr(>Chisq)
7	-42774.37	NA	NA	NA
9	-42773.04	2	2.663176	0.2640576

added an offset term since a person's duration of sleep session denoted the exposure period in our model.

In order to analyze the potential factors that might affect the number of sleeping quality flags. We generated two models: the first model only considered skin color as the fixed effect, and in the second model, we took sex as an additional factor to examine the potential confounding variables.

To be more specific, the two glmer models' formula could be written as:

Model 1

$$Y_{ij} \sim \text{Possion}(\mu_{ij})$$

$$\log(\mu_{ij}) = \log(T_{ij}) + \text{skin}_{ij} + U_i$$

.

Model 2

$$Y_{ij} \sim \text{Possion}(\mu_{ij})$$

$$\log(\mu_{ij}) = \log(T_{ij}) + \text{skin}_{ij} + \text{sex}_{ij} + U_i$$

.

, where U_i represented i^{th} customer ID as a random effect. skin_{ij} was the skin color of i^{th} customer in j^{th} measurement. $\log(T_{ij})$ was the offset term which presented the duration for i^{th} customer in j^{th} measurement. sex_{ij} was the sex of i^{th} customer in j^{th} measurement.

We performed a likelihood ratio test to compare model 1 and model 2. The null hypothesis for this test assume simpler model (Model 1) explained the data just as well as the more complicated (Model 2). As the result of this test, p value > 0.05, therefore, there was no evidence against the null hypothesis, we should use the simpler model.

Table 15: Estimates of the different skin color and their confidence interval

	Estimate	95% CI
Dark	0.033	(0.03,0.03)
Default	0.195	(0.19,0.20)
Medium	0.297	(0.29,0.30)
Medium Light	0.199	(0.19,0.20)
Light	0.092	(0.09,0.09)
Medium Dark	0.605	(0.59,0.62)

Therefore, we concluded that Model 1 perform better on fitting, we took model 1 as our final model.

From the table, we concluded that customers with dark skin were more likely to experience sleeping problems.

Based on the estimate for each category of customers' skin colors, we concluded that dark skin customers would have 0.033 flags every duration. The number of flags in every duration of customers with default skin was 0.092 times that of customers with dark skin, which meant that it was less likely to observe flags in default skin customer's sleep duration than in dark skin customer's duration. The number of flags in every duration of customers with default skin was 0.195 times that of customers with dark skin, which meant that it was less likely to observe flags in default skin customer's sleep duration than in dark skin customer's duration. Similarly, we found that it was less likely to observe flags in medium, medium-light, and medium-dark customers' duration than dark skin people's duration. Therefore, we concluded that dark skin could lead to a higher probability of observing sleeping flags.

We created a new column storing flags appearing during every sleep duration.

As you can see in the figure, the sleeping flag appeared more frequently in the sleep duration of customers with dark skin. Thus, we concluded that customers with dark skin were more likely to have sleeping flags.

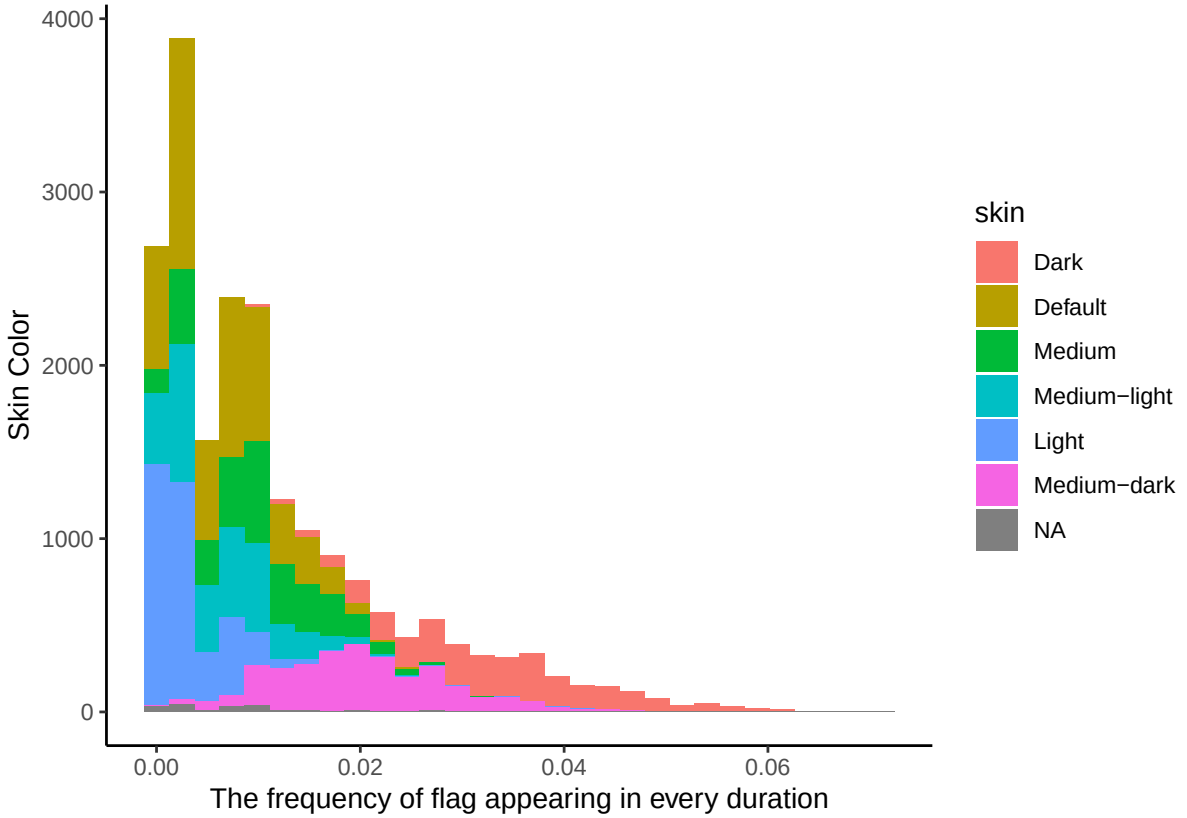

Figure 5: The frequency of sleeping flag appearing in diffent skin color people's duration

Discussion

- As a result, our new customers were those who bought the products in the "Active" and "Advance" line. Based on the above, we observed that the average age of our customers was 48 years old, and females took up a large proportion of being our new clients.
- According to the above, we found that the population size of new customers was larger than the traditional customers. We observed from the above analysis that the elder customers were more likely to purchase the new product lines "Active" and "Advanced" after finding the best fit model.
- According to our analysis, 25% of the customers had default skins, which meant that they were uncomfortable sharing information about their skin. According to the report, the skin color proportions of new and traditional customers (including those with dark skin) were also provided, with 13% having dark skin, 14% having medium-dark skin, 15% having medium-light skin, 18% having light skin, and 14% having medium-dark skin. One percent of the customers were unknown. Skin color was related to the high probability of sleep

quality flags, especially for the dark skin which led to a high probability of sleep quality flags.

Strengths and limitations

Strength:

- Since customer id was a grouping unit and was not meaningful in our experiment, we made customer id a random effect in the regression. Otherwise, we were violating the independence assumption.
- To avoid the non-convergence issue during model fitting, we re-scaled the age and population to the range of minimum 0 and maximum 1 to make their value become meaningful.
- We handled the duration as an offset term, then the coefficient estimate for an offset term was fixed to be one. The obvious relationship between duration and sleep flag led us to set duration as an offset term so we could compare the relationships between skin colour and sleep flags in different durations.

Limitations:

- A limitation of this analysis was that the sample size was reduced after removing all missing values and having a rather smaller data set. This implied that the conclusion we drew might not be fully applicable to all customers.
- Given that a model that contained many variables could not execute successfully due to the limited computational efficiency, only skin colour and sex were used to fit the models, and we found that sex had no significant demonstration of the response variable. There were other factors such as age not included in the model and these variables might affect sleeping as well.
- In addition to these concerns, some other limitations needed to be noted. We must be concerned about whether this data represented a random sample of the whole population. Since the skin colour was sensible data and some people refused to provide it, around a quarter of the data is without skin colour. Therefore, this data might not reflect the whole population precisely.
- At times, we also needed to consider some confounding variables that were included in the client information data set, such as the health condition. For example, clients suffering from high anxiety might have worse sleeping quality and irregular sleep duration.

Consultant information

Consultant profiles

Yinuo Chen. Yinuo is a junior consultant with Magic Conch. He specializes in data visualization and statistical modeling. Yinuo earned his Bachelor of Science majoring in Statistics and Economics from the University of Toronto in 2023.

Nanyi Wang. Nanyi is a senior consultant with Magic Conch. She specializes in sales forecasting and reproducing analysis. Nanyi earned her Bachelor of Science majoring in Statistics and Computational Cognitive Science and minoring in Mathematics from the University of Toronto in 2023.

Meiyi Wu. Meiyi is a senior consultant with Magic Conch. She specializes in Data Cleaning and Statistical Programming. Meiyi earned her Bachelor of Science majoring in Statistics and Mathematics and minoring in Computer Science from the University of Toronto in 2023.

Jiahui Yao. Jiahui is a junior consultant with Magic Conch specializing in statistical communication. Jiahui earned his Bachelor of Science majoring in Statistics and Mathematics from the University of Toronto in 2023.

Code of Ethical Conduct

1. Protects people's privacy as well as the confidentiality of their personal information. Knows and adheres to applicable rules, consents, and guidelines to protect private information.
2. Acknowledges any statistical and substantive assumptions made during the execution and interpretation of any analysis . Recognizes data editing techniques, including any imputation and missing data mechanisms, when reporting on the validity of data utilized.
3. Avoid disclosing or authorizing the revelation of confidential information collected in the course of professional activity for personal advantage or benefit to a third party without the previous written approval of the employer or client, or as required by a court of law.
4. Only take on work and perform services that are within the scope of professional competence; do not claim to have any expertise that do not have. In any professional evaluation or assessment, accept responsibility for the work and provide impartial and truthful information about procedures.
5. Recognize that include statisticians as authors or acknowledging their contributions to projects or publications requires their explicit agreement because it indicates approval of the work.

References

[1] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

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

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

[4] Hadley Wickham (2021). rvest: Easily Harvest (Scrape) Web Pages. 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] Yihui Xie (2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.37.

[8] Fitness tracker info hub. (n.d.). STA303/1002 Winter 2022 Final Project. Retrieved April 9, 2022, from https://fitnesstrackerinfohub.netlify.app/

[9] Postal code conversion file | Map and Data Library. (n.d.). Postal Code Conversion File. Retrieved April 9, 2022, from https://mdl.library.utoronto.ca/collections/numeric-data/census-canada/postal-code-conversion-file

[10] Population Density. (n.d.). CensusMapper. Retrieved April 9, 2022, from https://censusmapper.ca/

[11] Full Emoji Modifier Sequences, v14.0. (n.d.). Unicode. Retrieved April 9, 2022, from https://unicode.org/emoji/charts/full-emoji-modifiers.html

[12] Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3

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

Appendix

Web scraping industry data on fitness tracker devices

Steps:

1. Setting the libraries polite and rvest.
2. Getting the website url for the fitness tracker devices.
3. Using bow function to get any details provided in the robots text on crawl delays and which agents are allowed to scrape.
4. Using scrape function to scrape the content of the authorized page.
5. Getting the fitness tracker devices data set from the scraped website.

Considerations:

Following the web scrape principles:

- Avoid scaping everything all together from a public API, only use the one(s) needed instead.
- Always provide a User Agent string with clear intentions and contact method for any further questions or concerns.
- Request data at a reasonable rate. Strive to never be confused by a DDoS attack.
- Only save the data needed from a page.
- Respect all kept content. Never pass it off without confirmation of permission.
- Look for ways to return value by driving some (real) traffic to the site or credit it in an article or post.
- Respond in a timely fashion to the content owner's outreach and work with him/her towards a resolution.
- Scrape for the purpose of creating new value from the data, not to duplicate it.

Accessing Census data on median household income

Steps:

1. Installing the "cancensus" package.
2. Setting the library cancensus.
3. Signing up for the cancensus API through https://censusmapper.ca/

4. Getting API key and replacing it in the provided code
5. Getting all regions at the 2016 Census. (2020 Census not updated)
6. Filtering out the regions with CSD(Census Subdivision) level.
7. Getting the median income table with each median income in each CSD.
8. Re-scaling the median income to minimum 0 and maximum 1.

Considerations:

When the median income is zero, the median income is not meaningful. Then, we do the re-scale to make the minimum median income 0 and the maximum median income 1. Thus, the median income is meaningful at 0.

Accessing postcode conversion files

Steps:

1. Reading in the postcode conversion data set and saving it as postcode.

Considerations:

About ethical:

- Requiring to accept a license agreement to get access to this data.
- Disallowing to showing on GitHub directly. Need to pay more attention to how you will process this data, saving only the information you need for any data that is part of your final submission.