Potential Consumer Analysis of MINGAR's New Products

MINGAR's New Products Active and Advance

Report prepared for MINGAR by [TianyeWang]

2022-04-07

Contents

Executive summary

The company MINGAR is mainly engaged in the production of GPS devices for runners. Recently, in order to compete with competitor Bitfit company, new products of "Active" and "Advance" have been launched at lower prices. The purpose of this report is to help Company MINGAR analyze the personal characteristics of new customers and how they differ from traditional customers in order to develop potential new customers in the future. The second task is to help Company MINGAR address whether sleep quality is related to skin color.

Finding 1

Table 1: The difference between new customers and old customers

Varibles	Old customers	New customers
Age	46.29	47.72
Income	73182	68818
Population	1478851	1519423

- Customers using new products are mainly concentrated in densely populated areas.
- Customers using new products are 1.43 older than old customers.
- The income of new customers are lower than old customers.

Finding 2

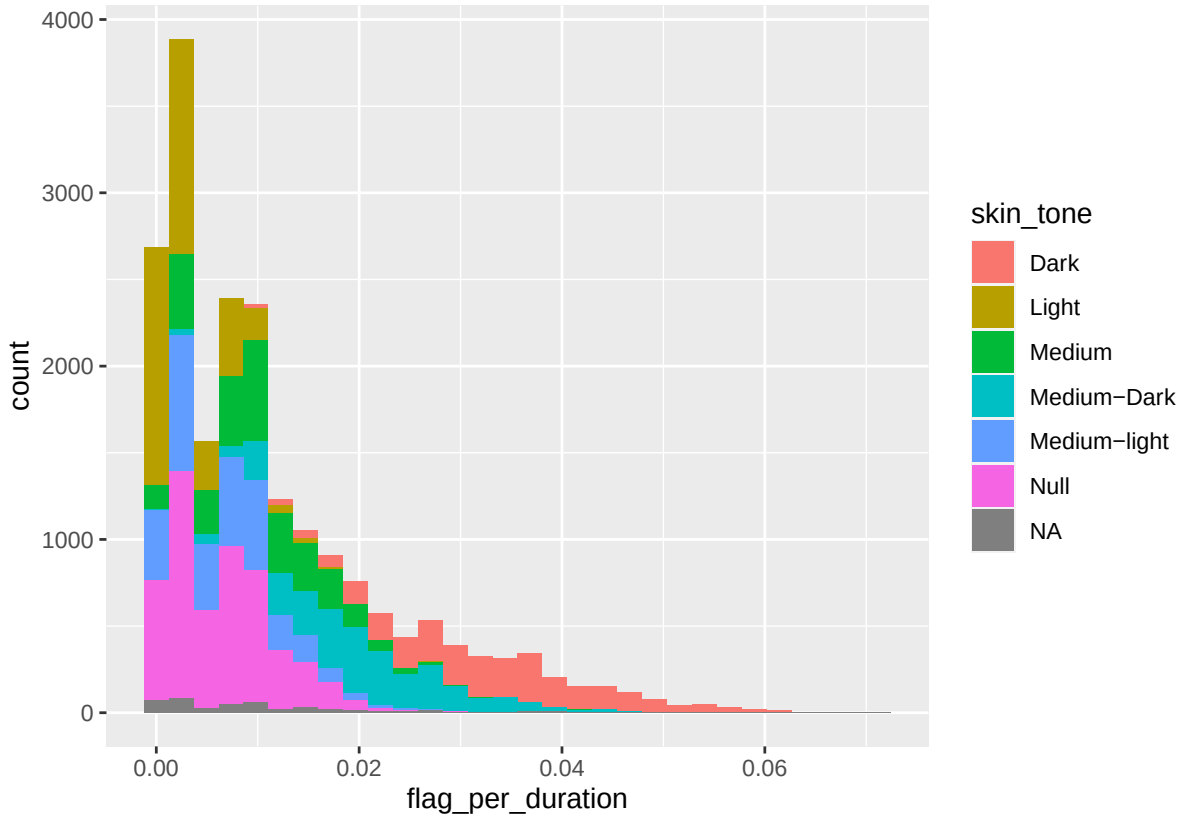


Figure 1: The Boxplot of Regional population between new customers and old customers

- The darker skin tone customers has lower sleep scores than others. There is a trend in complaints that devices are performing poorly for users with darker skin, particularly with respect to sleep scores.

Suggestion

We recommend that new products can be promoted in areas with relatively concentrated populations. Purchasing multiple products or buying a family product package will be more favorable.And a separate new product is set up for dark skin to ensure that people with dark skin can also improve the quality of sleep after using the product.

Technical report

Introduction

Overview

This business data analysis report shows the purchasing trend of consumers when it comes to Active and Advance projects, and the difference between old customers and new customers. We also analyzed whether the two devices had differences in performance in terms of different skin colors, particularly with respect to sleep scores.

Scope

This report focuses on the differences between new and existing users buying on new products. And pay attention to the various characteristics of potential users. For example, age income, skin color, region, and so on.

Methodology

The Company collects personal information from MINGAR's customers and information about the products used by customers. We also crawled the web with specific information about MINGAR's various products.

Research questions

- Who are our new customers?
- How are buyers of the newer and more affordable "Active" and "Advance" products different to our traditional customers?
- Do the company's products perform differently between people of different skin tones,particularly with respect to sleep scores?

Describe Data

Table 2: Describe Customer Data

Varibles	Meaning
cust_id	Unique ID for each customer.
sex	Biological sex, used for calculations of health metrics, like body-mass index and base metabolic rate.

Varibles	Meaning
dev_id	Unique ID for each device registered with our app.
device_name	Name of device type.
line	Line of products this device belongs to.
released	Release date for this particular device type. Year, month, day format.
CSDuid	hhld_median_inc
Population	The number of people in the customer's area
age	Age calculated based on the customer's date of birth
skin_tone	The client's skin tone
new_customer	If it is new customer,set 1.

Table 3: Describe Customer Sleep Data

Varibles	Meaning
cust_id	Unique ID for each customer.
date	For sleep data, date sleep session started. Year, month, day format.
flag	Number of times there was a quality flag during the sleep session. Flags may occur due to missing data, or due to data being recorded but sufficiently unusual to suggest it may be a sensor error or other data quality issue.
duration	Duration, in minutes, of sleep session.

Variable special instruction: We subtract the age variable by 17 years (the youngest of the clients).This is used to facilitate the establishment of models to illustrate meaning.We convert gender and product categories into factors. We set the gender to 1 for male and 0 for female.

We set customers who use Advance and Enhance devices as new customers, and set this variable to 1. Other customers set to 0.

Characteristics of new customers

In this section We provide some visualizations and accurate results of variables using relevant data analysis models to illustrate the difference between new and old consumers.The main visualization tools used are box plots and histograms. The models we use in the field of data analysis are mainly mixed linear logistic regression models.

Visualizations

We first look at the number of new and old customers in the dataset.

Table 4: The number of new customer and old customer

customer type	number
0	8387
1	10436

We can see from table 1 that the number of old customers is 8387 and the number of new customers is 10436. The number of new customers is 2049 more than the number of old customers.Let's take a look at the differences between the different personal characteristics of new and old customers.

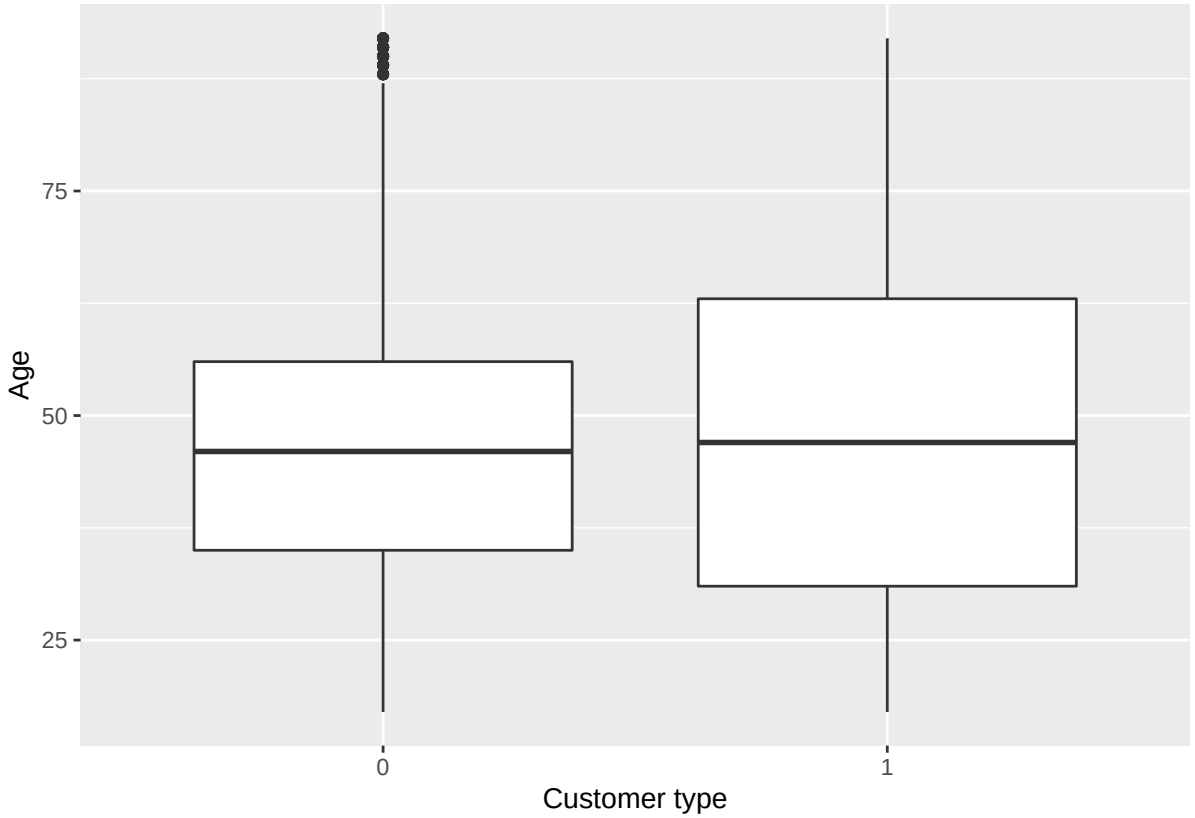
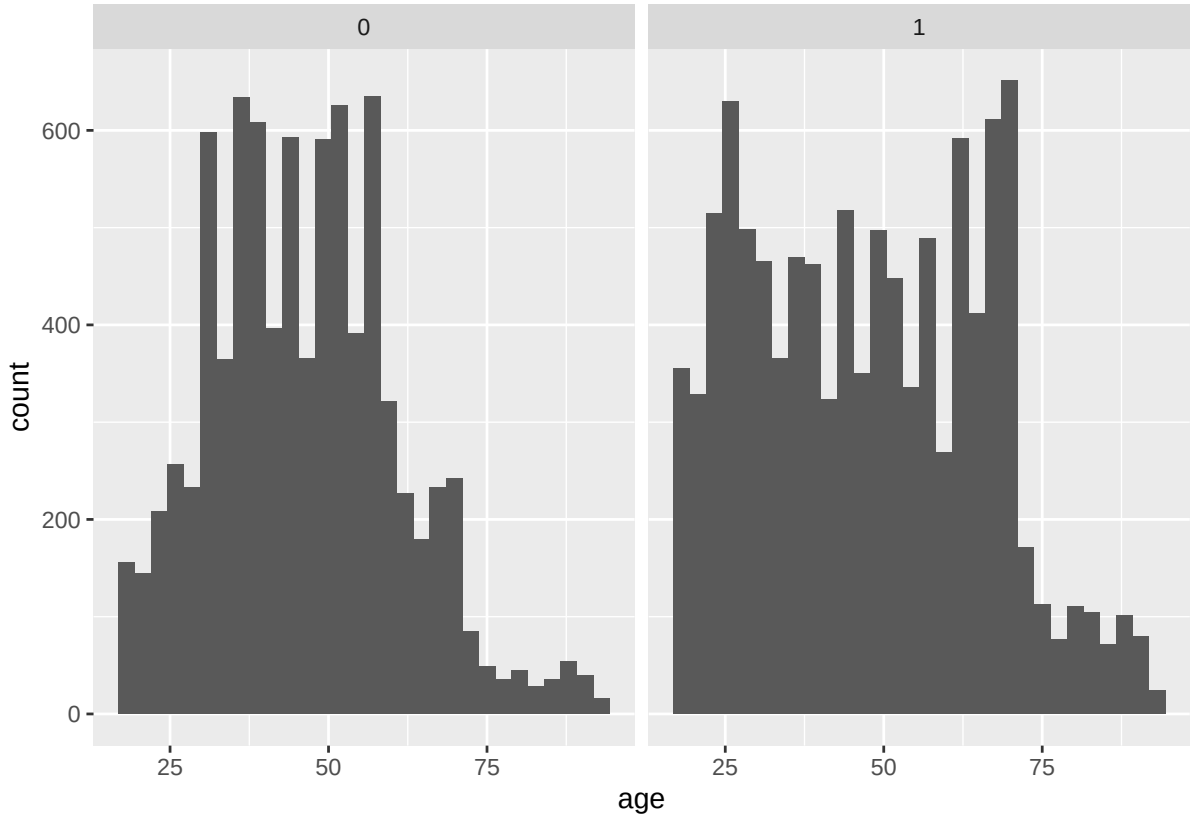


Figure 2: The Boxplot of age between new customers and old customers

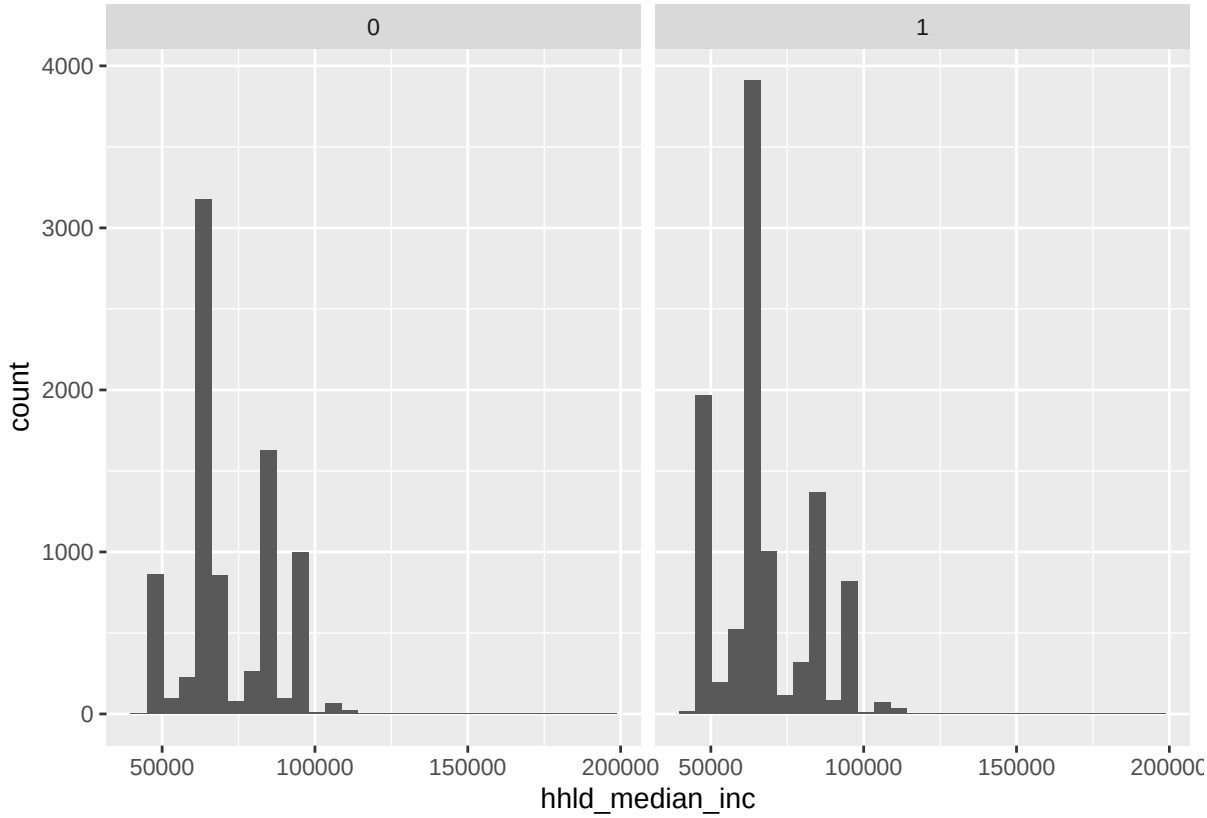


We can see from the box chart and histogram that the average age of new customers is slightly higher than the average age of old customers. But the age range for new customers is larger than for older customers.

Table 5: The average age of new customer and old customer

customer type	Average age	Variance of age
0	46.28508	218.5959
1	47.72346	337.5329

Customer type 0 refers to old customers and customer type 1 refers to new customers. We can see from Table 1 that the average age of new customers is older than the average age of old customers. But the age distribution of new customers fluctuates greatly.



We can see from the box chart and histogram that the average income of new customers is slightly lower than the average income of old customers. And the income range for new customers is smaller than for older customers.The income distribution of new customers is more concentrated in low-income groups.

Table 6: The average income of new customer and old customer

customer type	Average income	Variance of income
0	73181.99	220106858
1	68817.89	209195907

We can see from Table 2 that the average income of new customers is about 4364.1 lower than the average income of old customers. And the age distribution of new customers fluctuates greatly.We can also see from the variance that the income range of the new customer group is more concentrated.

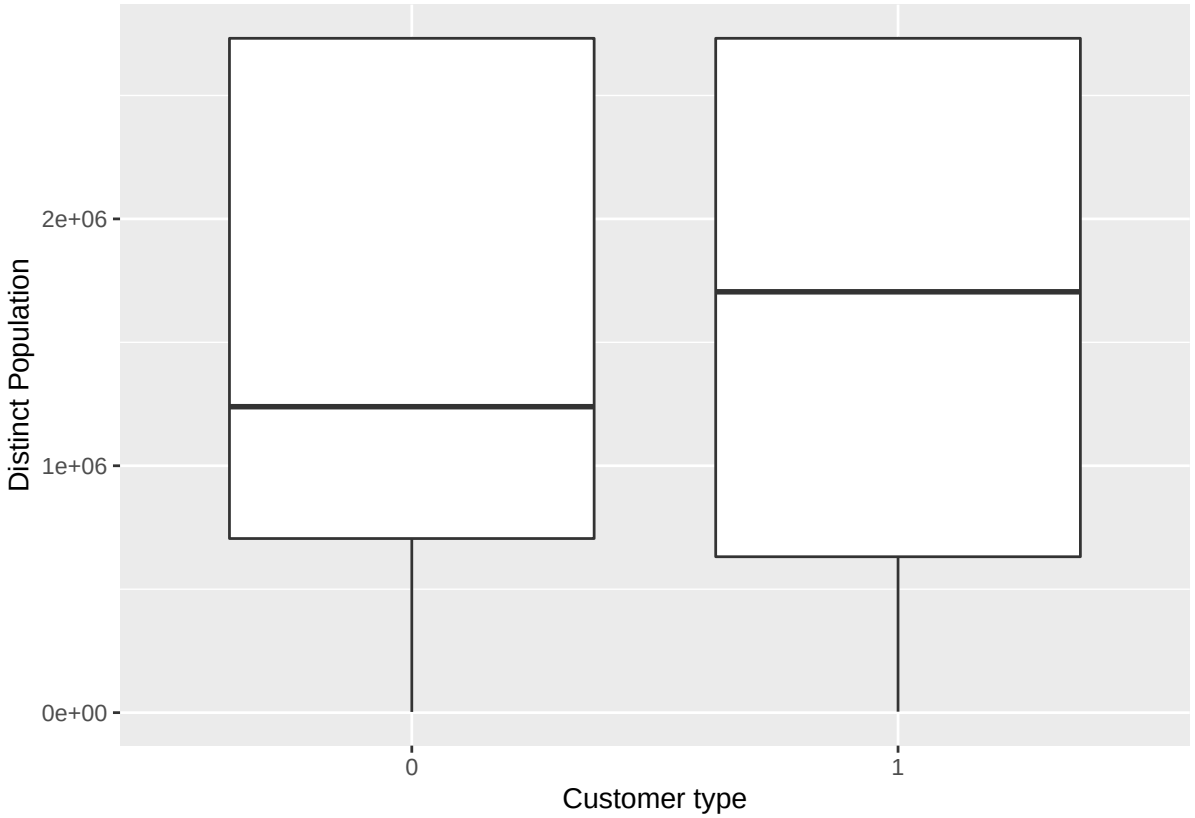
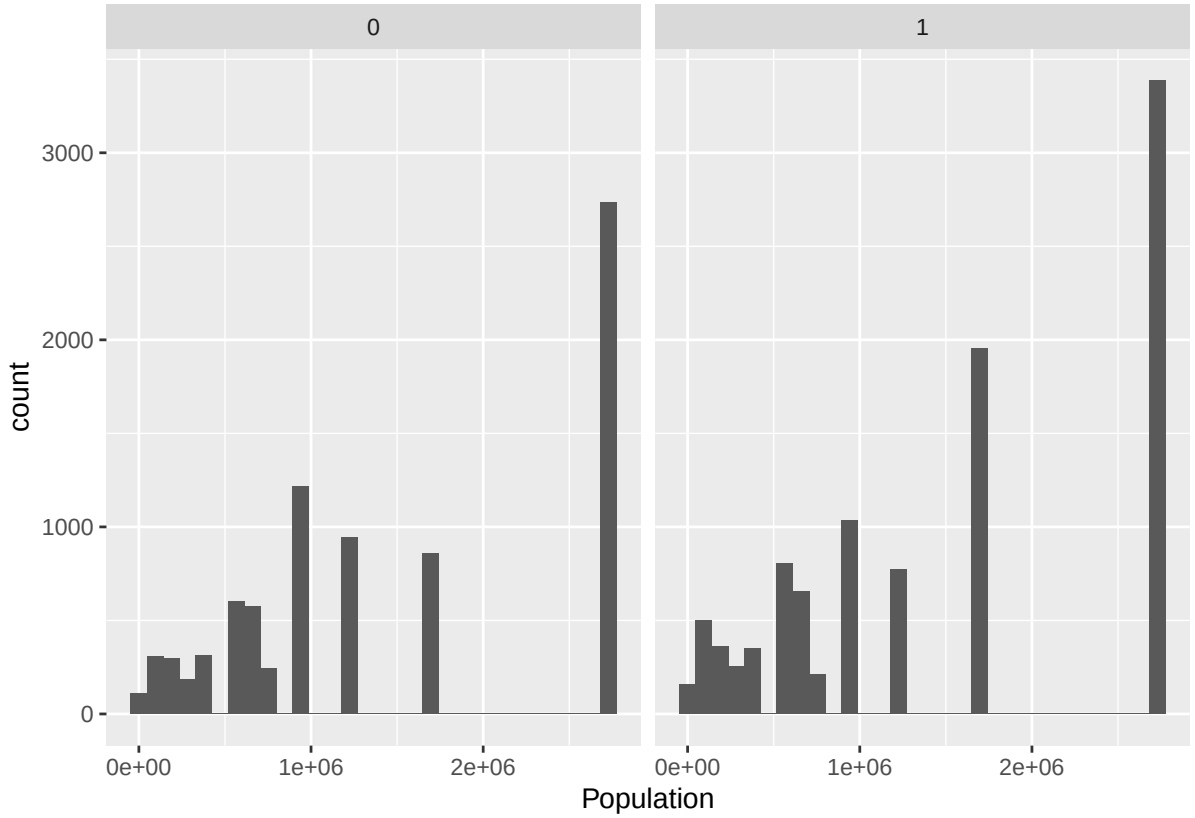


Figure 3: The Boxplot of Regional population between new customers and old customers



We can see from the box plot and histogram that most of the new customer groups live in more densely populated areas. Old customers, on the other hand, live mainly in places with low population density. The difference is very pronounced.

We can see from Table 4 that the proportion of women in new clients is much greater than that of men. And the proportion of women in new customers is higher than the proportion of women in old customers.

Data analysis

We used new_customer as the dependent variables and sex,age,income as the independent variable and CSDuid as control varible to construct the mixed linear logistic model.Generalized linear mixed models (or GLMMs) are an extension of linear mixed models to allow response variables from different distributions, such as binary responses. Alternatively, you could think of GLMMs as an extension of generalized linear models (e.g., logistic regression) to include both fixed and random effects (hence mixed models).

Table 7: Mixed Effects Logistic Model

Varibles	Estimate	Std.Error	Z-value	Pr(>z)
(intercept)	1.365e+00	6.830e-02	19.988	< 2e-16
sex	3.763e-02	3.028e-02	1.243	0.214
age	4.999e-03	8.843e-04	5.653	1.58e-08
hhld_median_inc	-1.715e-05	9.022e-07	-19.006	< 2e-16

From the results of the output above, age has a positive and significant prediction of the probability of becoming a new customer. Revenue, on the other hand, has a significant negative projection. Specifically, older and lower-income people are more likely to become new customers. Below we explain the different variables one by one:

We used age as the independent variables and new_customer as the dependent variable to construct the regression model.A regression model provides a function that describes the relationship between one or more independent variables and a response, dependent, or target variable.

Table 8: Study age vs new_customer

Varibles	Estimate	Std.Error	Z-value	Pr(>z)
(intercept)	29.2788	0.1842	158.957	< 2e-16
new_customer1	1.4392	0.2474	5.818	6.05e-09

We can see from table 6 that the age of new customers is 1.4392 years more than the age of old customers.

We used income as the dependent variables and new_customer as the independent variable to construct the linear mixed-effects models model.Linear mixed models are an extension of simple linear models to allow both fixed and random effects, and are particularly used when there is non independence in the data, such as arises from a hierarchical structure.

Table 9: Study income vs new_customer

Varibles	Estimate	Std.Error	t-value
(intercept)	7.413e+04	1.458e+02	508.6
new_customer1	-2.137e-09	1.471e-04	0.0

We can see from table t that the income of new customers is less than the income of old customers.

We used Population as the dependent variables and new_customer as the independent variable to construct the linear mixed-effects models model.Linear mixed models are an extension of simple linear models to allow both fixed and random effects, and are particularly used when there is non independence in the data, such as arises from a hierarchical structure.

Table 10: Study Population vs new_customer

Varibles	Estimate	Std.Error	t-value
(intercept)	9.872e+04	1.782e+03	55.41
new_customer1	1.220e-07	4.120e-03	0.0

We can see from Table 8 that there is a high probability of new customers in densely populated areas.

Conclusion

We can see from the visual graphs and the output data of the model that the new customers are older and have lower incomes. Customers using new products are mainly concentrated in densely populated areas. And women are the main customers who use new products.So women who are older, have lower incomes, and live in densely populated areas are more likely to become potential new consumers.

Relationship between skin color and sleep score

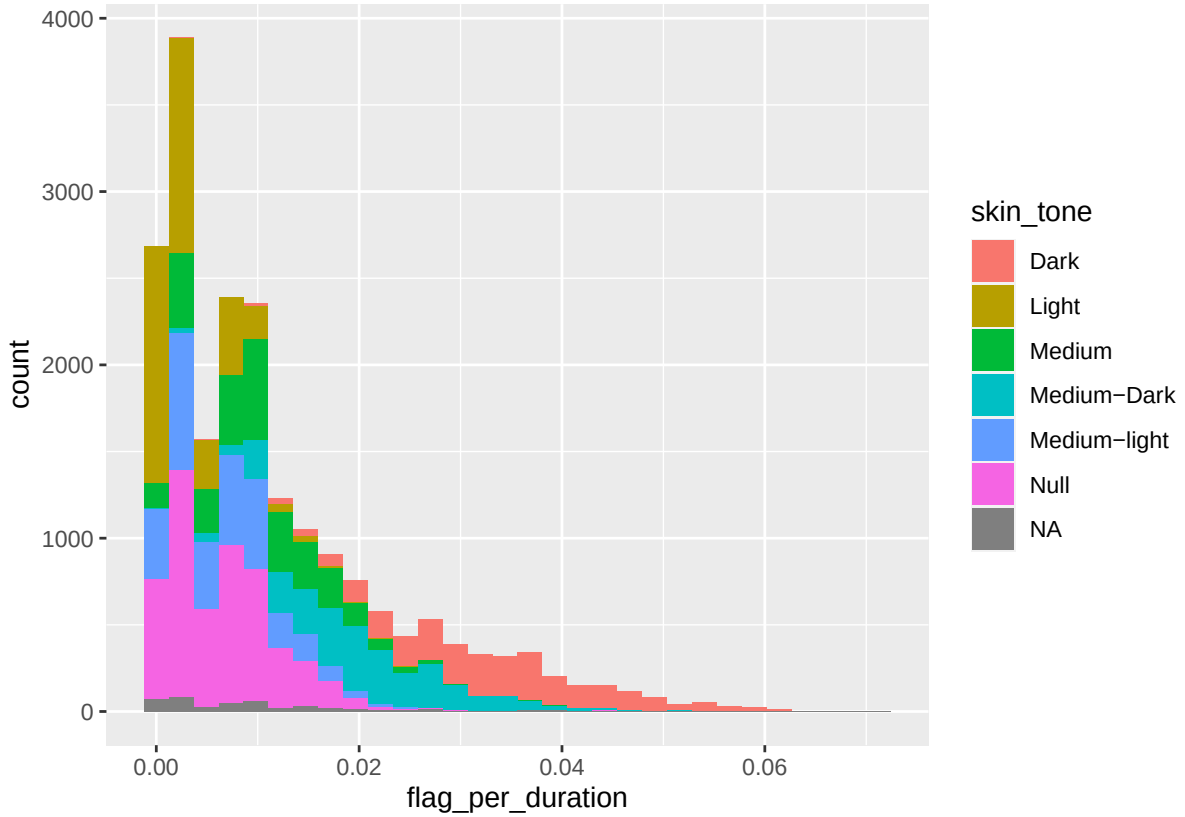
In this section,we mainly focus on the relationship between skin tone and sleep quality. We used visualization images and mixed linear regression models to illustrate the relationship between

the two, respectively.

Visualizations

Since customers use the device for different lengths of time, we create a new variable that divides the tag by the duration to control the device usage time.Average flags per duration reflects the number of times sleep interrupted during a certain period of time.

This table shows that people with dark skin tone have highest average flags per duration. This reflects that sleep scores are lowest.We can see from the table that people with dark skin have lower sleep scores. People with light skin tones have higher sleep scores.



We can see from the pictures that dark-skinned people have more sleep quality flags in a certain period of time.This means that people with darker skin tones have lower sleep scores than people of other skin tones.

Data analysis

We used average flags per duration as the dependent variables and skin_tone as the independent variable and CSDuid as control varible to construct the mixed linear logistic model.Generalized

linear mixed models (or GLMMs) are an extension of linear mixed models to allow response variables from different distributions, such as binary responses. Alternatively, you could think of GLMMs as an extension of generalized linear models (e.g., logistic regression) to include both fixed and random effects (hence mixed models).

Table 11: Generalized linear mixed model Output

Varibles	Estimate	Std.Error	z-value	Pr(z>z)
Intercept	-3.398196	0.006466	-525.54	<2e-16
skin_toneLight	-2.391386	0.016517	-144.79	<2e-16
skin_toneMedium	-1.213827	0.011267	-107.73	<2e-16
skin_toneMedium-Dark	-0.502596	0.009004	-55.82	<2e-16
skin_toneMedium-light	-1.614746	0.012781	-126.34	<2e-16
skin_toneNull	-1.633642	0.012781	0.012781	<2e-16

We can see from the output data that dark skin tones as the basis, people who change to other skin tones have a reduced number of markings. People with dark skin tones who prove to have the worst sleep quality are the worst.

Table 12: Confidence level 95%

Varibles	2.5%	97.5%
Dark skin tone	0.00	0.022
Intercept	-3.41	-3.39
skin_toneLight	-2.42	-2.36
skin_toneMedium	-1.24	-1.19
skin_toneMedium-Dark	-0.52	-0.48
skin_toneMedium-Light	-1.64	-1.59
skin_toneNull	-1.65	-1.61

Conclusion

We can see from the visual graphs and the output data of the model that the darker skin tone customers has lower sleep scores than others. There is a trend in complaints that devices are performing poorly for users with darker skin, particularly with respect to sleep scores.

Discussion

We will answer the three survey questions we mentioned earlier.

- For research question 1 (Who are our new customers?),women who are older, have lower incomes, and live in densely populated areas are more likely to become potential new consumers.
- For research question 2 (How are buyers of the newer and more affordable "Active" and "Advance" products different to our traditional customers?),new customers are older and have lower incomes than old customers and are mainly concentrated in densely populated areas.
- For research question 3 (Do the company's products perform differently between people of different skin tones,particularly with respect to sleep scores?),there is a trend in complaints that devices are performing poorly for users with darker skin, particularly with respect to sleep scores.

Suggestion

We recommend that new products can be promoted in areas with relatively concentrated populations. Purchasing multiple products or buying a family product package will be more favorable.And a separate new product is set up for dark skin to ensure that people with dark skin can also improve the quality of sleep after using the product.

Strengths and limitations

Strengths

The present report analyses the problem from two perspectives. On the one hand, the customer's problem is analyzed from a graphical point of view, so that the customer can intuitively get the desired information visually. On the one hand, the problem is analyzed from the professional perspective of data analysis, and the data is analyzed through statistical methods. It is convenient for non-professionals to understand and for professionals to conduct research and discussion.

Limitations

The limitation of this report is that, due to the limited computing power of computers, the models used are too simple to explore the relationship between several variables in a single model. Models have certain limitations.

Consultant information

Consultant profiles

Experienced and dedicated Data Analyst with several years of experience identifying efficiencies and problem areas within data streams while communicating needs for projects. Adept at receiving and monitoring data from multiple data streams, including Access, SQL, and Excel data sources. Ability to synthesize quantitative information and interact effectively with colleagues and clients. Proven track record of generating summary documents for senior management for monthly and quarterly audit and compliance reporting.

TianyeWang.Tianye Wang is a senior consultant with Eminence Analytics. She specializes in data visualization,and modeling. TianyeWang earned her Master of Data Science, Specialist in Statistics Methods and Practice, from the Harbin Institute of Technology in 2023.

Code of ethical conduct

- Respect the Company's private information and the customer's personal information. Absolutely no information provided by any company will be disclosed.
- Everything in the data analysis report is objective, impartial, and there is no subjective bias.
- Accept responsibility for work and give objective and reliable information on procedures in any professional review or assessment.

References

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

[2] 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.

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

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

[5] Achim Zeileis, Torsten Hothorn (2002). Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. URL https://CRAN.R-project.org/doc/Rnews/ [6] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

[7] 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.

[8] 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.

[9] 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/.

[10] 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

Web scraping industry data on fitness tracker devices

https://fitnesstrackerinfohub.netlify.app

- We grabbed information about the company's equipment from the web. https://unicode.org/emoji/charts/full-emoji-modifiers.html
- We grabbed the emoji information used by customers from the Internet to analyze the customer's skin color.

Accessing Census data on median household income

We select median income CSDuid and Population from Census Data. The census api key is CensusMapper_d6eb158793ff5789d40d5eb55f72b896.

Accessing postcode conversion files

We get this file "data-raw/pccfNat_fccpNat_082021sav.sav" in 2016 from https://mdl.library.utoronto.ca/collections/numeric-data/census-canada/postal-code-conversion-file. And we got PC, CSDuid variables.

Course statistic project content URL

https://sta303-bolton.github.io/sta303-w22-final-project/