2022 Customer and Product Report

New Line Customer Profile and Wearbale Skin Colour Complaint

Report prepared for MINGAR by Data Analytics Inc.

2022-04-07

Executive Summary

Results for the differences between new customers and old customers show that age, income and population have effects on the chances of attracting new customers, as well as indicate that the new Active line of products are more attractive to this category, as seen in Figure 1.

Current scientific literature suggests that wearable devices and their various tracking methods (heart rate, sleep-tracking, etc.) may not be as accurate for those with darker skin tones due to physiological factors and perhaps negligence of inclusive design. In addition, wearable devices thus far have been typically targeted towards, and consumed by a wealthier customer base who can afford these expensive products. The current study aims to update these socioeconomic and racial findings specific to Mingar's products. It will consider the effects of customer skin tone (using the customers preferred emoji colour) and how error-prone their wearable is. The analysis also intends to explore characteristic trends (e.g. age, household income, etc.) among Mingar's new customer base who purchase from newer, more affordable product lines and its traditional customer base.

```
## Warning: Removed 10 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 4 rows containing missing values (geom_bar).
```

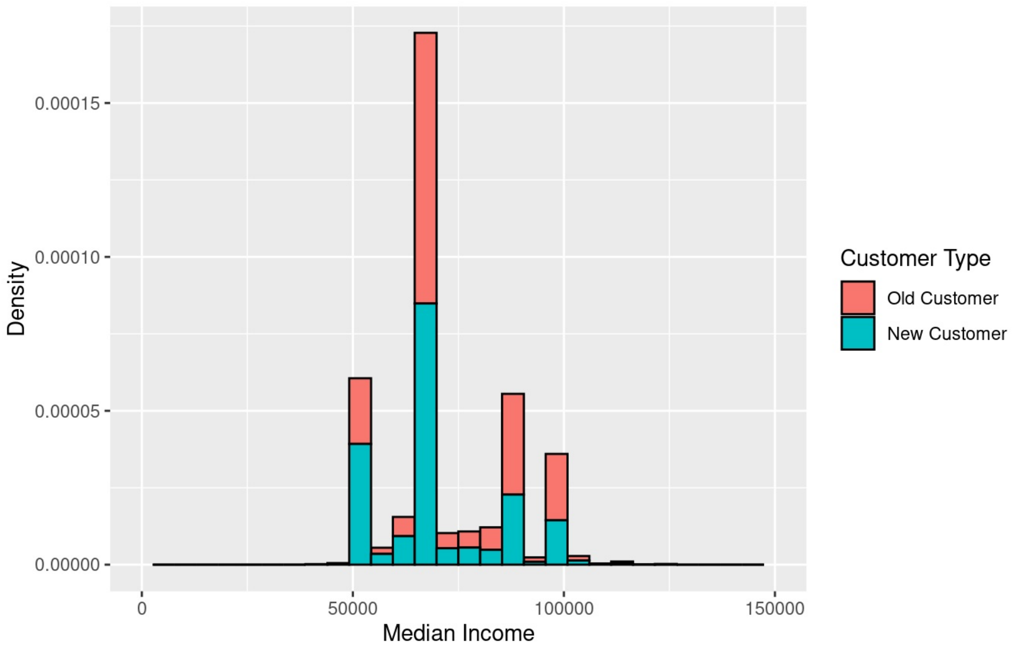


Fig 1:Distribution of Median Income based on customer types

Key findings

- New customers tend to be older with a 0.44% times increase in purchasing odds for every increase in age yearly
- New customers tend to be located in less densely populated areas with a 0.000006% times reduction in odds per population unit increase
- New customers tend to be lower in income with a 0.02% reduction in purchasing odds per unit of income increase
- 24% of the study sample population uses the lightest emoji and 18% uses the darkest emoji

Limitations

- Emoji colour may not be indicative of actual customer race, and neglects customers who choose to use default colored emojis

Technical report

Introduction

This report detailed by Data Analytics Inc aims to provide thorough statistical analyses consumable by both a more technical statistician audience and a professional business audience for Mingar and their wearable devices. The report will discuss in detail and attempt to answer two main overarching questions. One addressing a potential socio-racial issue with Mingar's devices. The other question explores differences in characteristics between customer bases who purchase from two newer and more affordable different product lines sold by Mingar.

Research questions (bullet points): How do the new customers differ from the traditional customers? With a focus on: whether the customers using the "Active" and "Advanced" line come from a different income base compared to the traditional customers?

Do the devices' sleep tracking feature perform worse for users with darker skin?

Research questions

- Do the customers using the "Active" and "Advanced" line come from a different income base compared to the traditional customers
- Do the devices' sleep tracking feature perform worse for users with darker skin?

Data Description and Preparation

There are three files sourced externally. There are four other files provided by Mingar. They include the customer information, customer ID matched to their device ID, device information associated with each device ID, and customer sleep tracking history. All the files, excluding sleep tracking history, are merged together into one dataset for ease of analysis. Specifically, the customer postal codes are matched with their associated census subdivisions through the Postal Code Conversion File, then matched to the median household total income at the specific census subdivisions. The other internal data files are joined together either by the customer ID or device ID.

Some changes are made to the variables. We create three additional variables: age, skin tone, and a boolean variable identifying whether a customer is new versus traditional. The age is calculated from customers' date of birth, referencing against April.7, 2022, and the skin tone is identified based on the emoji type the customers use. We categorise five types of skin tones: light, medium light, medium, medium dark, and dark. The customers who don't have emojis are removed from the dataset used in answering the second research question, because we are not able to identify their race based on the emojis for the analysis. The customers using devices from the Advanced or Active line are the new customers; the ones using devices from the Run line are the traditional customers. The new customers are given a value of one, while the traditional customers are assigned a value of zero, for the identifier variable.

Research Question 1: marketing analysis on new and traditional customer

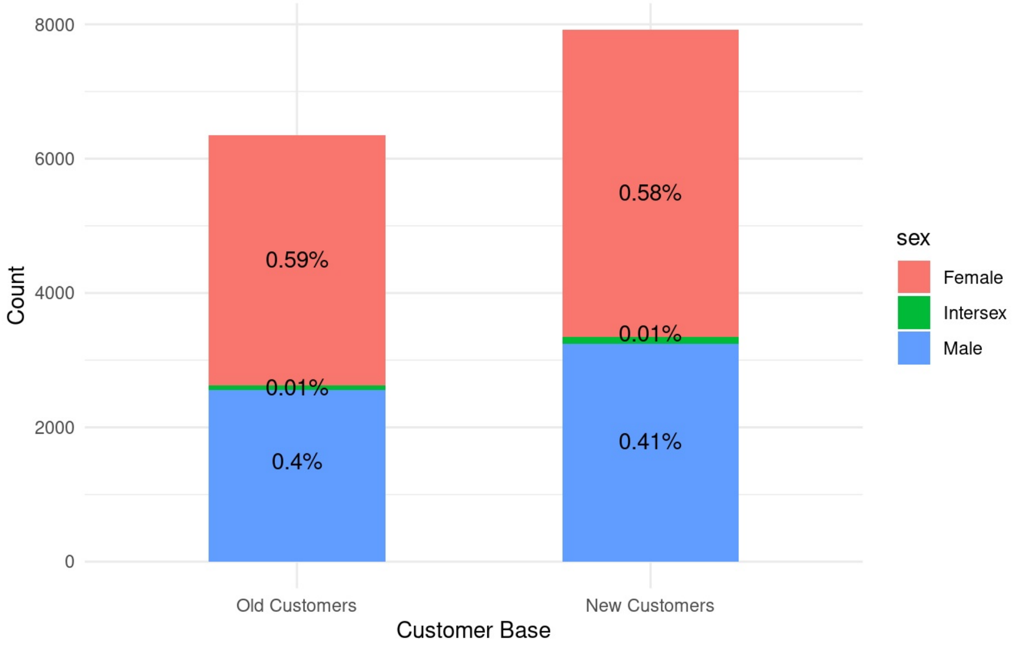


Fig 2: Count Distribution based on customer base and sex

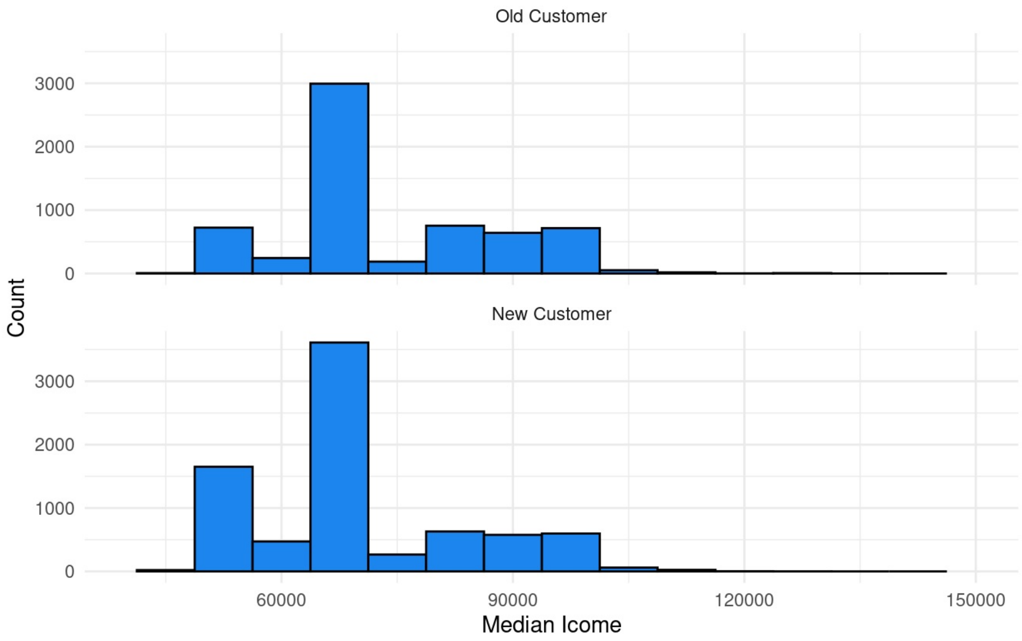


Fig 3:Distribution of Median Income based on customer types

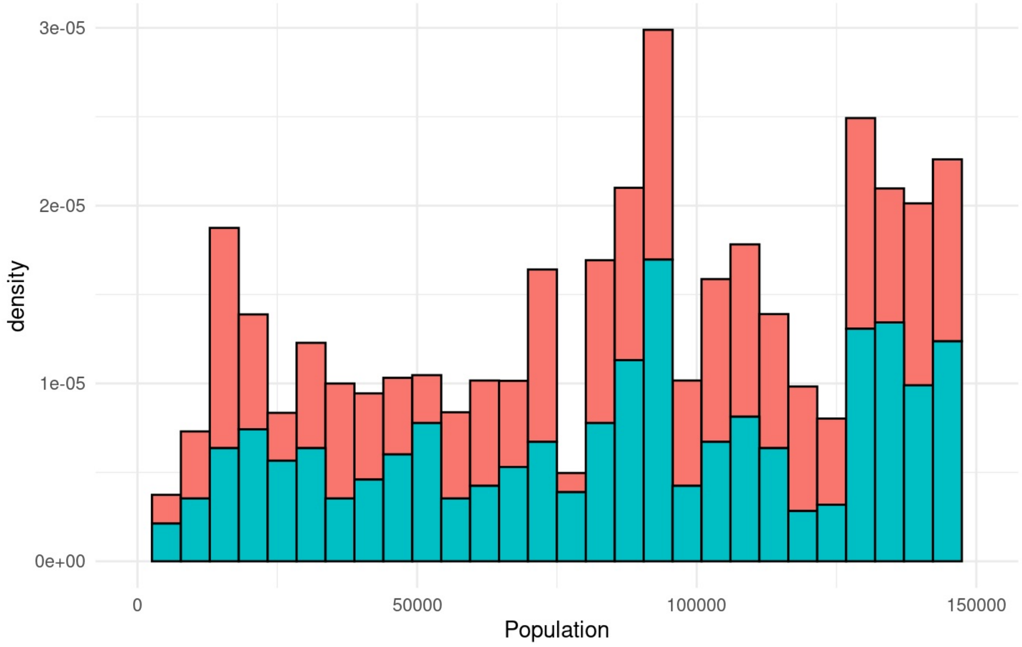
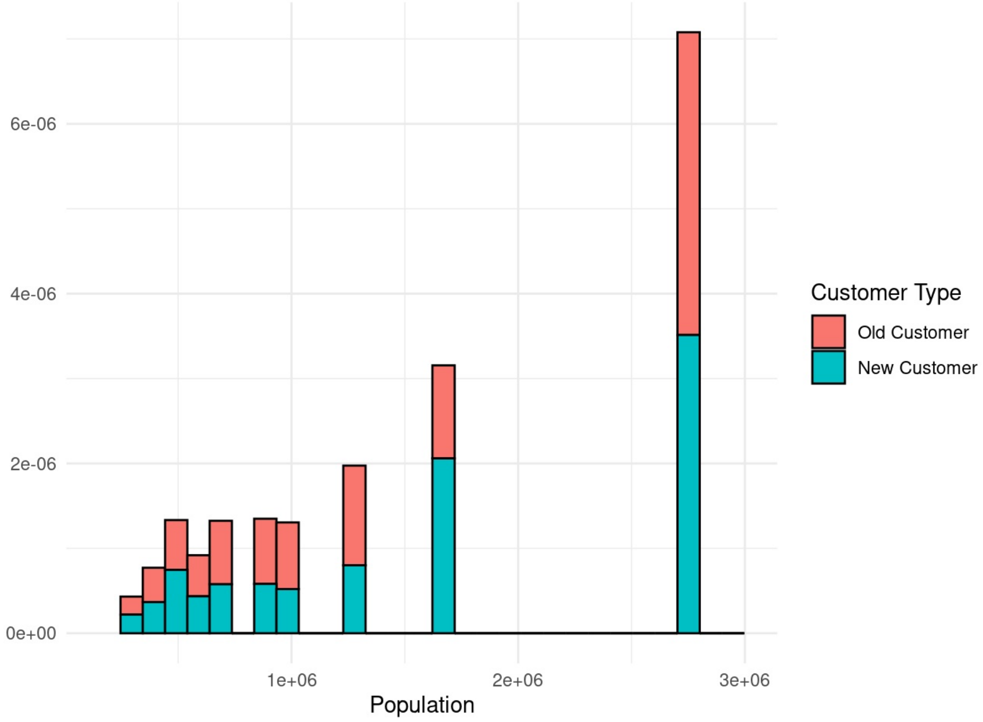


Fig 4:Distribution of Population of Customer based on customer types



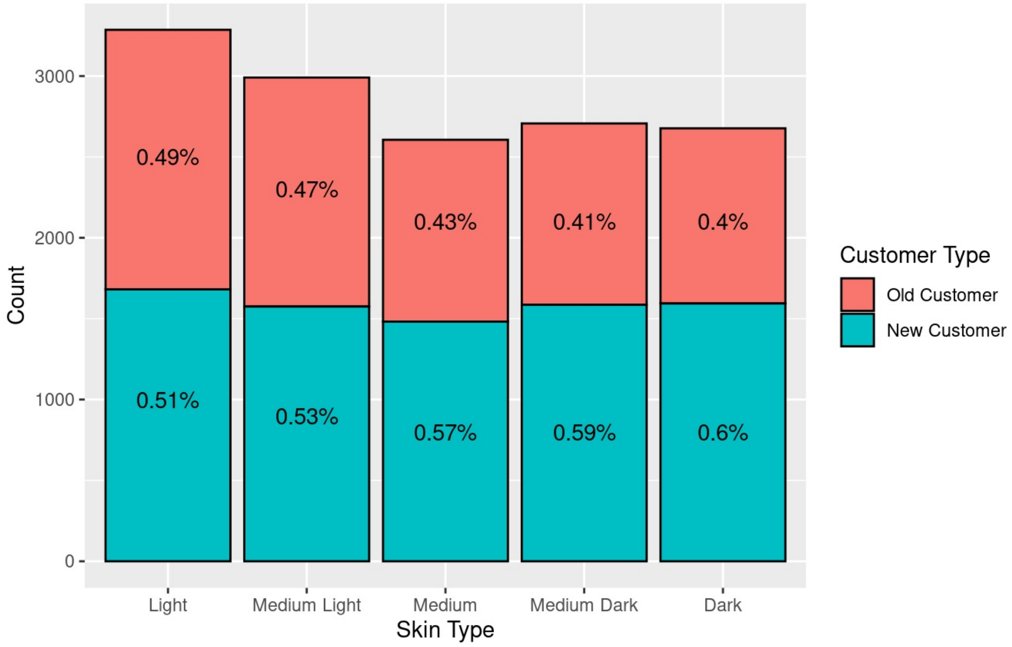


Fig 5: Proportions of customer types per skin type

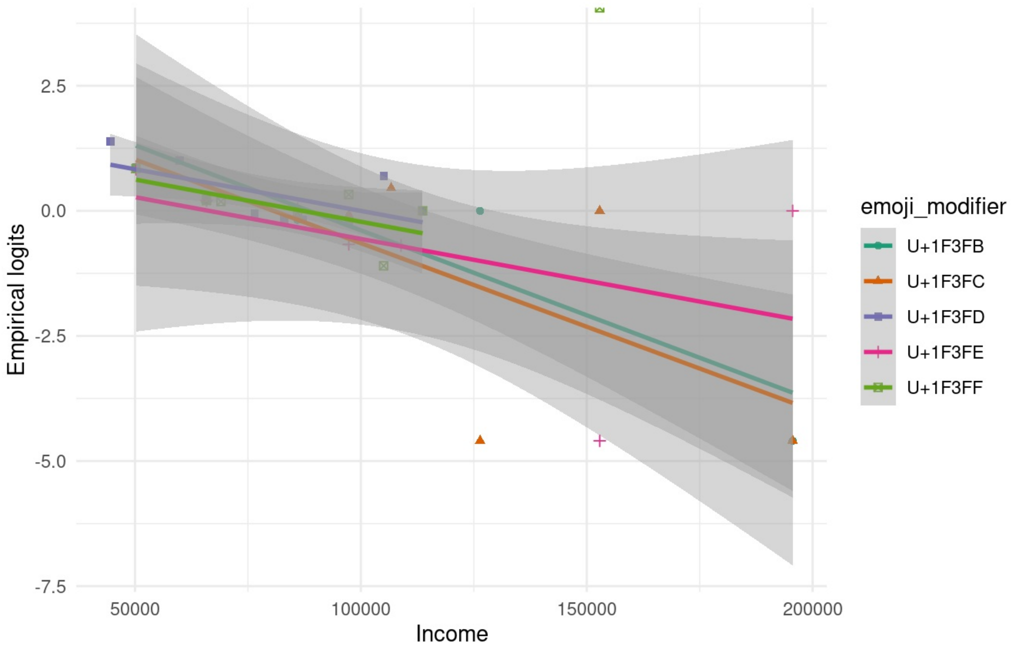


Fig 6:Emprirical Logit plot of Income seperated by Emoji_modifier

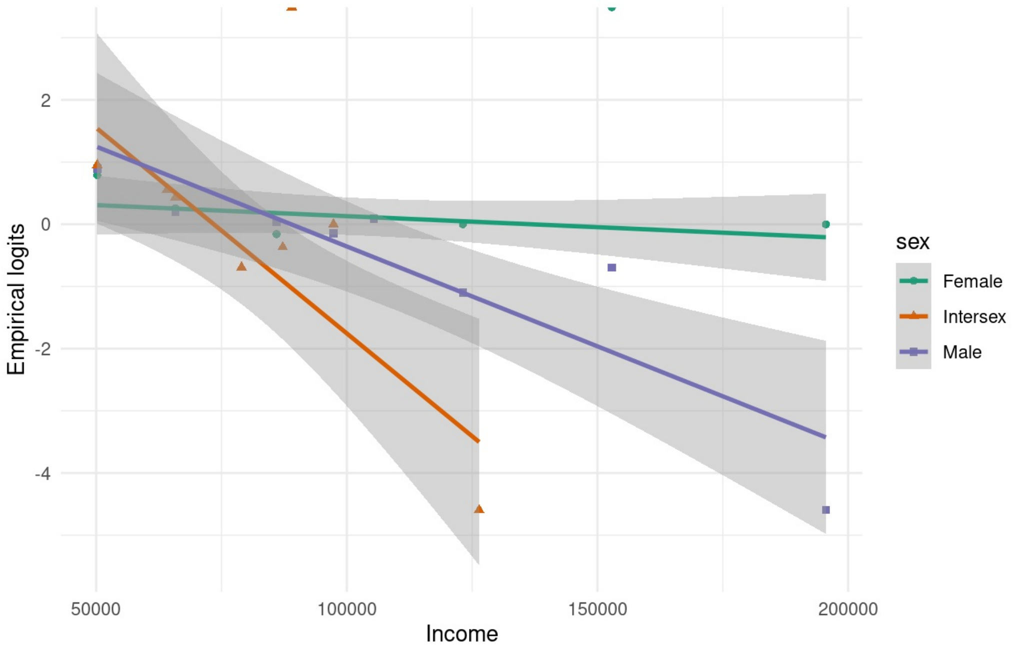


Fig 7: Emprirical Logit plot of Income seperated by sex

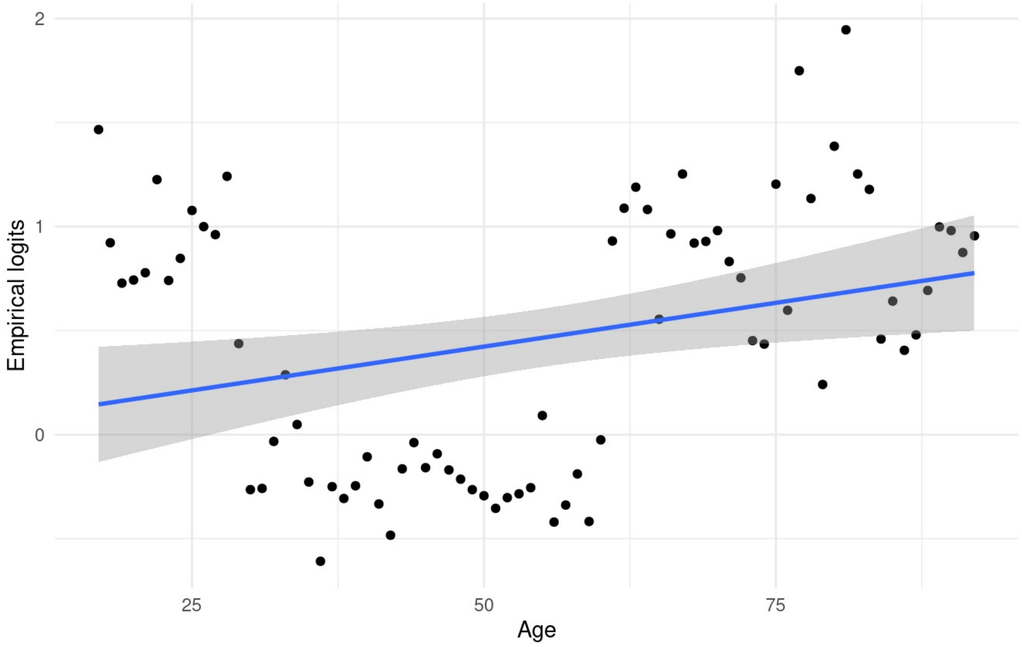


Fig 8: Emprirical Logit plot of Age

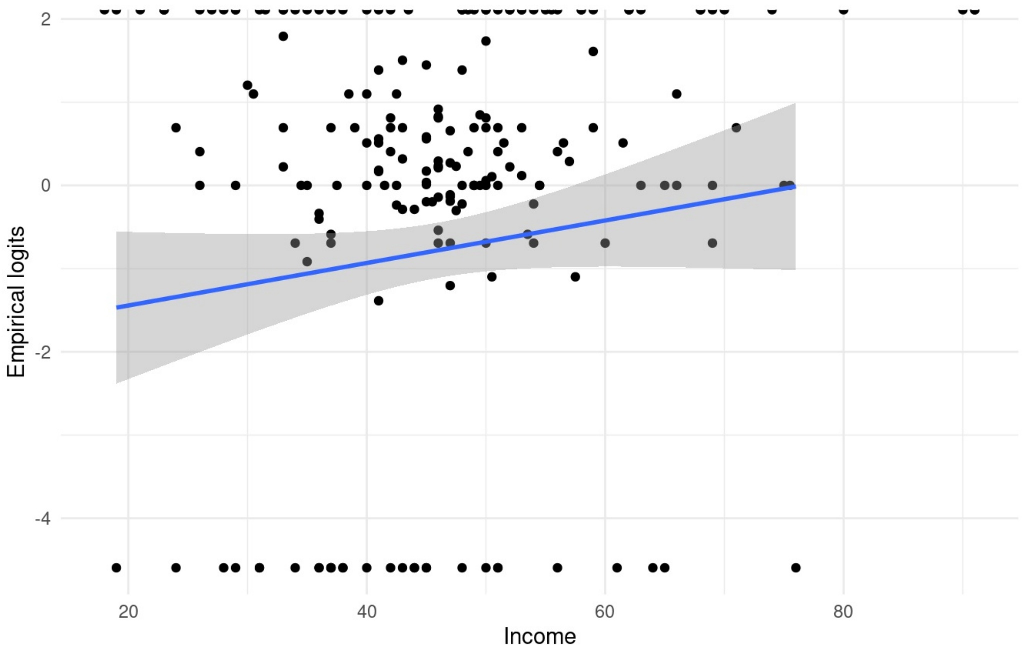


Fig 9: Emprirical Logit plot of Income

EDA suggests that income, population and age should be relevant factors for our model. Furthermore, based on the nature of the data, we opted to use a generalised linear model under the binomial family. Our EDA seemed to be accurate, as initial models suggested that all three of the aforementioned factors play a significant role within the model. Sex was also correctly assumed to be insignificant, although we may note that there exists a slight relationship when looking at the combined effects of income and sex, but this was only relavent for one of the three sexes in our data. Furthermore, we discovered that the emoji used by customers (an indication of skin colour of our customers) has a significant role in regards to the model we have built so far. The interaction between income and skin colour was the most relevant, as we spotted a p-value supported trend of certain emojis (in conjunction with income) having a direct effect on the model. Tests against the emoji model versus non-emoji models indicated that the non-emoji model had a slightly smaller AIC score. Studies have shown that people of certain skin colours are known to have differences in income despite having no other discernible differences in attributes (Devaraj et. al, 2018), meaning that we may gain insight into customers that fall under this category with our current model. However, we chose to reject the emoji model in favour of our current model as the AIC indicated a better model, and p-values were significant for all variables (as opposed to not being the case for the emoji model). All other factors considered did not provide better explainability to our model, or increased accuracy. We also did not find a random effect to be needed in our model, since our sample size is large and we can be confident that we have captured every effect level needed.

Final Model Coefficients

Coefficients	Estimate	P values	Significance Level
Intercept	1.589	1.05*10^(-31)	***
hhld_median_inc	-0.000021	3.35*10^(-45)	***
Population	-0.0000000648	3.95*10^(-3)	**

age	0.00473	8.43*10^(-5)	***

Research question 2: sleep tracking quality based on skin colors

The response being modeled is the number of flags. Because it is counts per session, Poisson distribution is the primary option. The number of flags depend on the duration of the sleep session – a longer session will result in more flags than a shorter session. Therefore, logarithm of the duration is applied as an offset in the model. The observations are grouped based on the customer ID, so even though the customers are independent, the observations are not independent from each other. Therefore, a generalized linear mixed model, with the customer ID as a random effect, and Poisson distribution for the response, is appropriate for this research question. After removing the observations with no emojis and thus no skin tone information, the remaining dataset is plotted. Figure 1 shows that the customers' racial backgrounds are evenly distributed, for customers who have data on their emoji choices. Therefore, any bias caused by customers unwilling to represent their actual skin tones when choosing emojis is minimal. This shows that the dataset is not affected by this potential bias.

There are four assumptions for Poisson distribution. 1.Poisson Response The response variable is a count per unit of time or space, described by a Poisson distribution. 2.Independence The observations must be independent of one another. 3.Mean=Variance By definition, the mean of a Poisson random variable must be equal to its variance. 4.Linearity The log of the mean rate, log(λ), must be a linear function of x If any of the assumptions are not satisfied, it will interfere with getting accurate analytic result from the model and the best fit. The histogram of the flag counts show reducing bin heights, indicating that the response is indeed described by a Poisson distribution. Therefore the first assumption is satisfied. The observations have been checked to remove any duplicates. Also, there are no two customers living at the same address, so the sleep tracking quality of individual customer is not highly correlated because of physical proximity. We can assume that the second observation is satisfied. The mean and variance of the number of flags for different groups of skin tone are not equal. The difference between the two increases for the groups with darker skin tones. For example, for the group with dark skin tone, the variance is 16.09, and the mean is 11.79. The mean and variance of the frequency for different groups of skin tone are not equal either. The variances are always smaller than the mean, and the difference increases for groups with dark skin tone. For the group with dark skin tone, the variance of the frequency is 0.4, while the mean is 2. To handle dispersion problem, we can either use quasipossion distribution or negative binomial distribution instead of poisson distribution. Figure 4 shows that the log of the mean rate is a linear function of x.

We fit a negative binomial model, with flags as the response, and skin tone, age, sex, and model released date as the predictors, as discussed, and include logarithm of the duration as an offset. The individual significance test indicates that skin tone, sex, and model released date are significant. Next, different nested model combinations are tested by pairs via the Chi-square test. Our final model includes skin tone, sex, and model released date as fixed term predictors. No interaction term is included in the model. The AIC and BIC values for this model is 2098.1 and 2136.1, which are the lowest among the model options. Unfortunately, the value for the Goodness-of-Fit test is 0. Outliers in the model can be a significant cause, and they do exist as seen in Figure 1. With the sleep tracking function, it is likely that once the feature malfunctions for a device, it would produce a high number of flags, as well as if it doesn't track sleep as well for people with darker skin tone compared to people with light skin tone.

Discussion

Based on the model, age seems to be the biggest indicator of new customers. Similarly, we see that population plays an effect, with people in less population dense representing a part of the new customer base. Similarly, older individuals also seem to encompass the newer customer base. Income also provides great insight, as we see that there is a negative trend between income and being a new customer. This leads to the conclusion that products on the cheaper end of the the spectrum, marketed towards older individuals in rural areas, are the most likely target marker for expanding the customer base. This also implies that new cheaper products from mingar are more likely to attract new customers, as well as customers outside of their usual high income customer base. In terms of specific products, the cheaper Active line of products seems to be the most conducive for attracting new customers. ### Strengths and limitations Model 1: The interpretability of the model is a key strength that we decided to focus on: the limited number of variables and strong p-values give us numerical results that are easy to understand as well as visualise. In terms of difficulties with our model, customers that may be more privacy inclined may not be represented within the model, since we opted to focus our model on users that used specific emojis (thus indicating their skin colour). Customers that opted out of this feature for privacy reasons may not be accurately accounted for in the model. Customers could potentially use emojis that are not actually indicative of their own skin tone (for a variety of reasons).

Consultant information

Consultant profiles

Sherry Xiaoman Lu. Sherry is a Data Scientist at Data Analytics. She specializes in statistical modeling, project management, and business leadership. Sherry earned her Bachelor of Science, Specialist in Mathematics and its applications, with a focus in Probability and Statistics from the University of Toronto in 2023.

Rumteen Taheri Dolatabadi. Rumteen is a Data Scientist in the technology sector with experience as an Actuary. He specializes in theoretical and mathematical probability. Rumteen earned his Bachelor of Science with a specializiation in Applied Mathematics and a concentration in Statistics and Probability from the University of Toronto in 2022.

**Antony Dudnikov. Antony is a Senior Data Analyst at Data Analytics. His specializations include statistical modeling, data wrangling and data visualizations. Antony received his Bachelor of Science, majoring in Statistics and Economics with a focus in Data Analytics, from the University of Toronto St.George in 2023.

**Opal is a Data Scientist at Data Analytics. He specializes in mathematical probability and public policy. Opal earned his Bachelor of Science in Mathematics and Applications (Probability Theory) with a minor in Philosophy from the University of Toronto in 2023.

Code of ethical conduct

We promote three fundamental tenets within our company: 1. Impartial and unbiased practice of statistical methods. All the work we do is free from personal influence and is ensured to have minimal conflict of interest. 2. Impartial and unbiased presentation of research. All the work we do is represented in a manner that reflects a neutral tone, while minimising the possibility for misinformation. 3. Lawful practice of statistical methods. All the work we do complies with federal and local laws, including matters regarding privacy and handling of sensitive data.

References

Devaraj, S., Quigley N, Patel P. The effects of skin tone, height, and gender on earnings. 2018. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0190640 (https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0190640)

https://academic.oup.com/sleep/article/43/10/zsaa159/5902283 (https://academic.oup.com/sleep/article/43/10/zsaa159/5902283)

Hailu R. Fitbits and other wearables may not accurately track heart rates in people of color. 2019. https://www.statnews.com/2019/07/24/fitbit-accuracy-dark-skin/ (https://www.statnews.com/2019/07/24/fitbit-accuracy-dark-skin/). Accessed Apr 18, 2022.

Bent B, et al. Investigating sources of inaccuracy in wearable optical heart rate sensors. NPJ Digit Med. 2022;03:18.

Appendix

The data on the wearable models released by Mingar, specifically with the line name paired with each model to indicate if it belongs to one of the new lines or the traditional line, is on the website https://fitnesstrackerinfohub.netlify.app/ (https://fitnesstrackerinfohub.netlify.app/). It is scrapable with a 12-second crawl delay. This information is useful to answer the first research question, when classifying the devices.

The 2016 census data is the most current census data available. It is sourced through an API provided on https://censusmapper.ca/ (https://censusmapper.ca/) and by using the R package Cancensus. The focus is the 58 census subdivisions, and their associated median household total income.

The Postal Code Conversion File with the reference date August 2021 is the latest conversion file available for the 2016 census. It matches postal codes with the geographical census subdivisions. We use the newest file because it is the most comprehensive with over 142 thousands more postal codes added, compared to the previous version. Some rows are repeating and some postal codes are associated with more than one census subdivision. This problem is solved by removing the duplicates. If the duplicates are present, it would affect the mean and variance values for different groups of customers and violate the independent observation assumptions for some models.