# Report title

Subtitle that indicates findings

Report prepared for MINGAR by Simplicity

2022-04-11

# Contents



# General comments (you can delete this section)

*Before making any changes, knit this Rmd to PDF and change the name of the PDf to something like 'original-instructions.pdf', or whatever you like (it is just for your reference).. Then you can delete this section and if you want to check what it said, just open the other PDF. You don't HAVE to use this particular template, but you DO need to write you report in RMarkdown and include a cover page.*

*The cover page must be a single stand alone page and have:*

- *A title and subtitle (that indicate your findings)*
- *"Report prepared for MINGAR by" your company name*
- *Date (assessment submission date is fine)*

*You can change the colour of this cover to any colour you would like by replacing 6C3082 in the YAML above (`titlepage-color:`) to another hex code. You could use this tool to help you:* https://htmlcolorcodes.com/color-picker/

*Note: There should NOT be a table of contents on the cover page. It should look like a cover.*

## Executive summary

*Guidelines for the executive summary:*

- *No more than two pages*
- *Language is appropriate for a non-technical audience*
- *Bullet points are used where appropriate*
- *A small number of key visualizations and/or tables are included*
- *All research questions are addressed*

*The module 4 writing prompt provides some tips and information about writing executive summaries.*

# Technical report

*This part of the report is much more comprehensive than the executive summary. The audience is statistics/data-minded people, but you should NOT include code or unformatted R output here.*

## Introduction

In this report, Simplicity will tackle Mingar's most pressing questions to date, i.e. who are Mingar's customers of their newer and more cost-efficient wearable lines? And do the currently trending damaging claims of "racist" Mingar wearables hold any weight? Read on to find out.

Our company has gone through the exhaustive work of selecting the 2 best fitting models of the data that best answer each question by considering any predictors that may influence our variable of interest, within the realm of common sense and reason backed by statistical tests, handy plots and colourful illustrations. Due to racial and ethnic privacy issues that prevent Mingar from collecting such data, we have rolled up our sleeves, wrangled the data and gotten creative with our approach to solving our client's questions. And in the spirit of transparency, we will explain any limitations that we ran into in the process.

### Research questions

Our report aims to address the 2 questions that Mingar has asked of us. More specifically, Question 1 targets primarily who Mingar's new customers of the affordable lines of wearables, "Active" and "Advance" are. We aim to highlight the specific significant factors that do and do not identify such customers.

Question 2 asks to find any device incompetencies found during sleep score flag records due to user skin color likely stemming from sensor quality issue or other data quality issue to determine whether there is racial bias existing within Mingar's wearables.

## Who are the "Active" and "Advance" Line Wearables Customers?

*For each research question, you will want to briefly describe any data manipulation, show some exploratory plots/summary tables, report on any methods you use (i.e. models you fit) and the conclusions you draw from these*

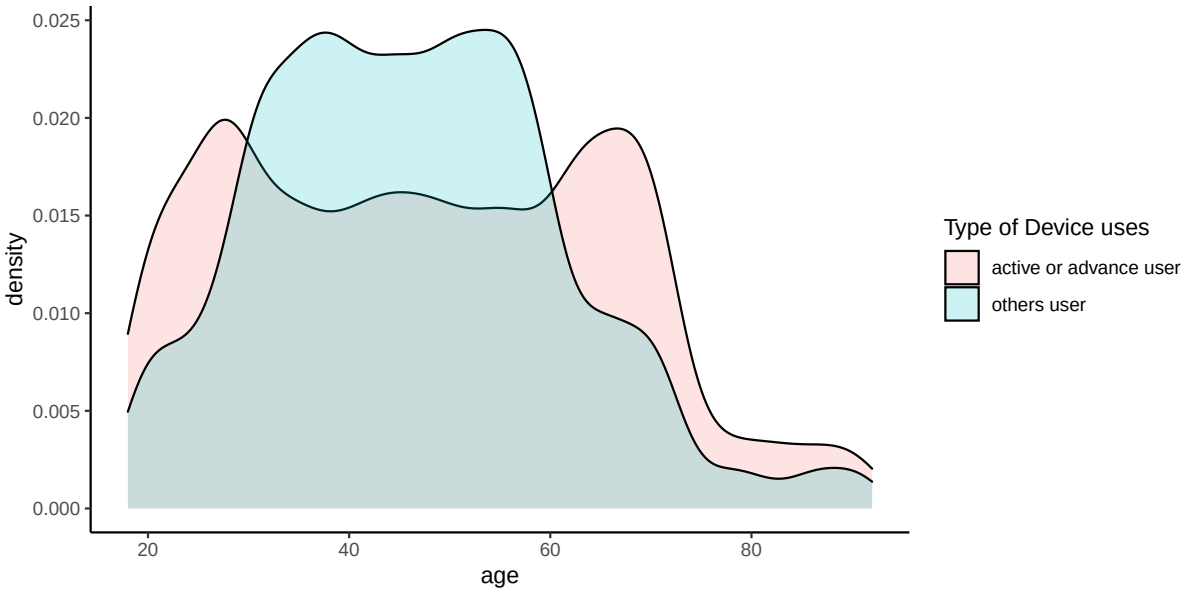



**Figure 1:** Density Distribution of Active or Advance Users vs. Other Device Users by Age

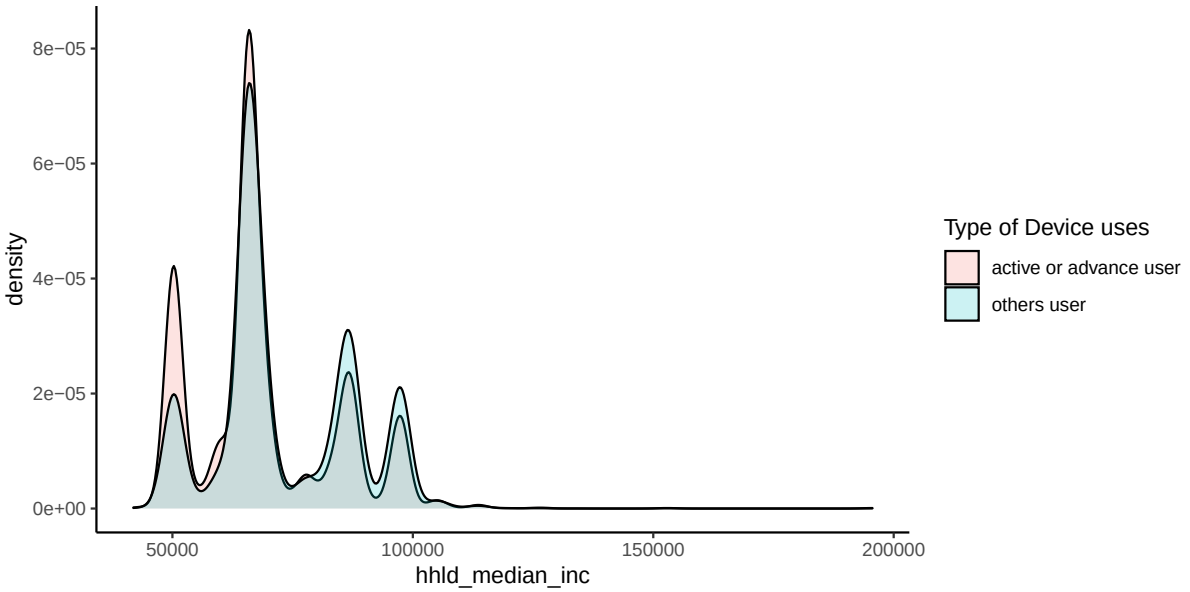



**Figure 2:** Density Distribution of Active or Advance Users vs. Other Device Users by Median Income

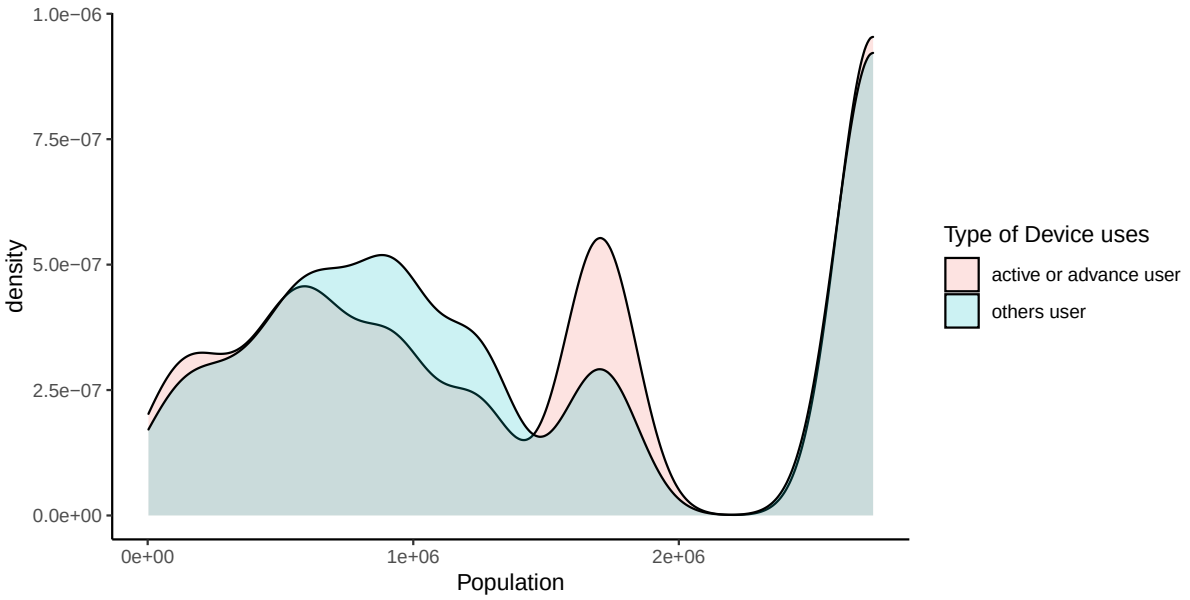



**Figure 3:** Density Distribution of Active or Advance Users vs. Other Device Users by Post Code Population

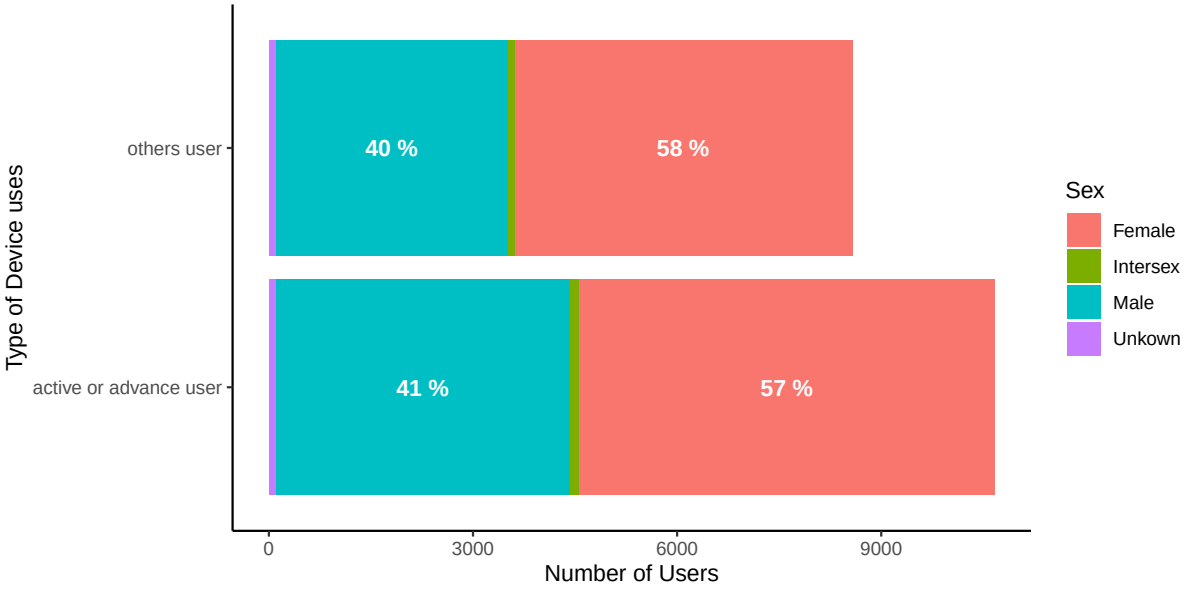



**Figure 4:** Percent Distribution of Active or Advance Users vs. Other Device Users by Sex

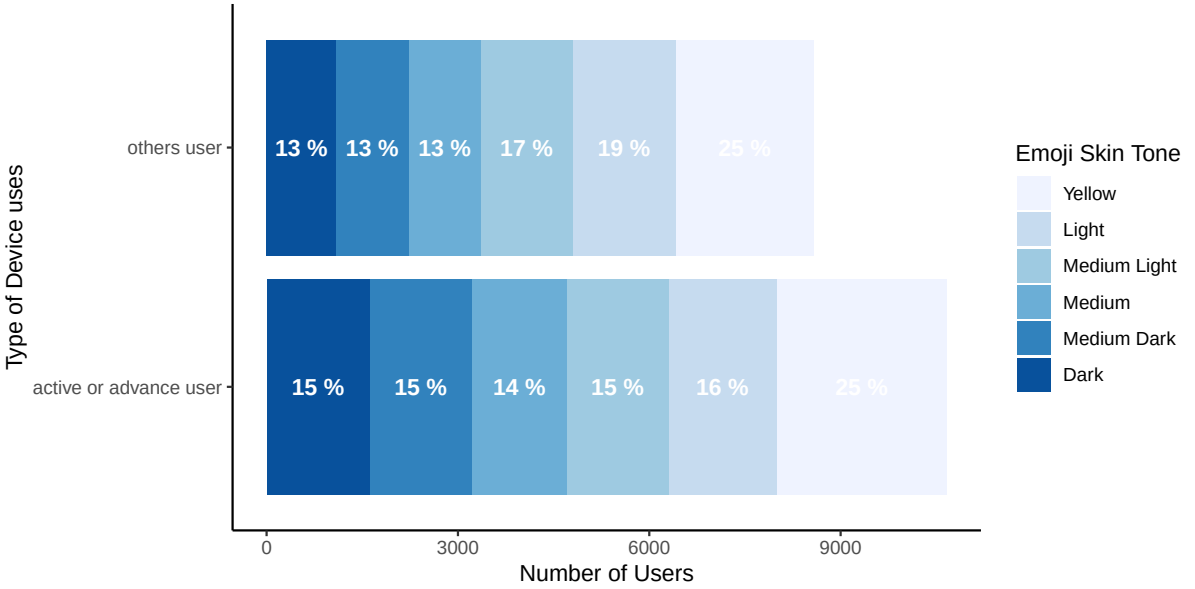



**Figure 5:** Density Distribution of Active or Advance Users vs. Other Device Users by Prefered Emoji Skin Tone

Model1 is given by:

$ActiveAdvanceUsers \sim Binomial(N, p)$

$log(\frac{p}{1-p}) = \beta_0 + \beta_1 MedianIncome + \beta_2 Age + \beta_3 Population$

Where $N$ represents the number of Device Users, and $p$ represents the probability of a device user uses Active or Advance given the information on median income and population.

**Table 1:** Summary Table of The Generalized Linear Model

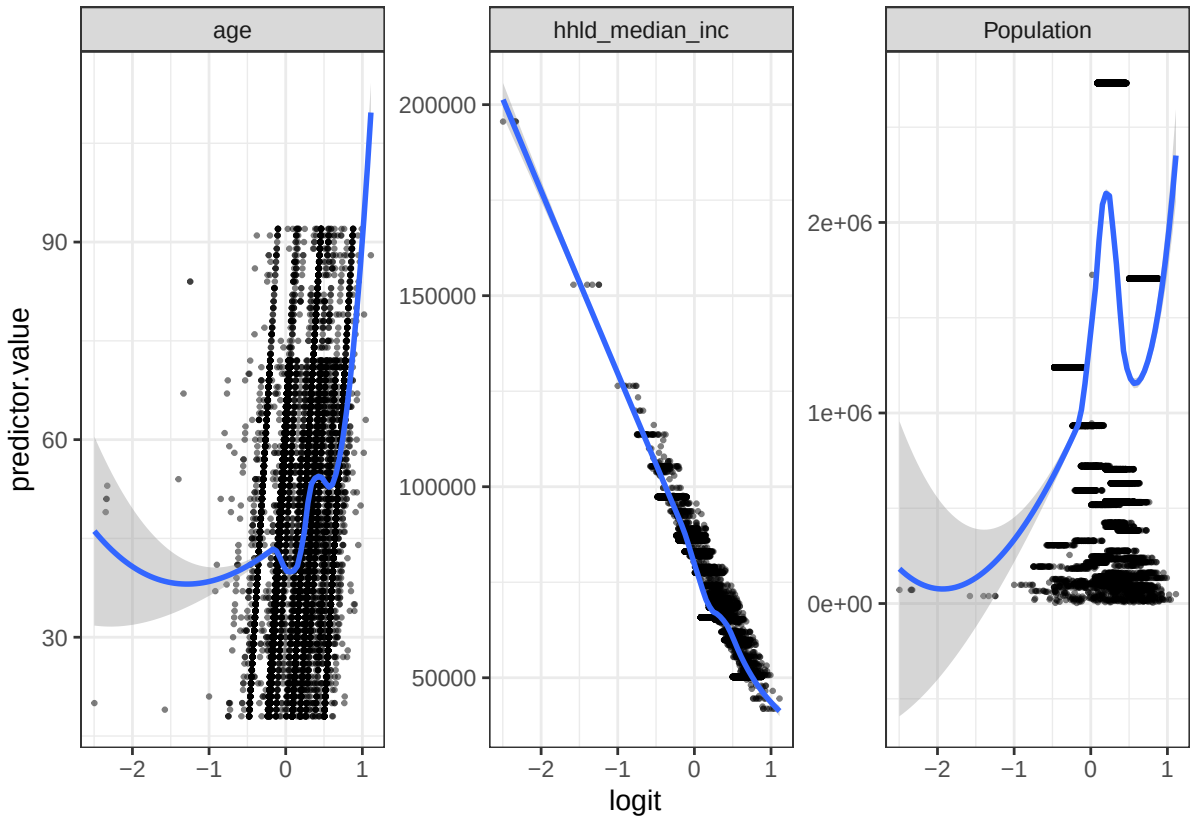
| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 0.817 | 0.063 | 12.934 | 0 |
| scales::rescale(hhld_median_inc) | -3.318 | 0.192 | -17.261 | 0 |
| scales::rescale(age) | 0.377 | 0.075 | 5.041 | 0 |
| Population | 0.000 | 0.000 | -4.185 | 0 |

**Table 2:** Confidence Interval Table of The Generalized Linear Model

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | 0.693 | 0.941 |
| scales::rescale(hhld_median_inc) | -3.695 | -2.942 |
| scales::rescale(age) | 0.230 | 0.523 |
| Population | 0.000 | 0.000 |

For a generalized linear model to be valid, following assumptions needs to hold.

1. The model's response, the probability of Active Advance Users among the users should be independently distributed, and the errors are also assumed to be independent
2. The model assumes a linear relationship between the transformed response with the predictor variables. Which in this case is the $log(\frac{p}{1-p})$ and $age, median\ income$, and $population$. Where $p$ represents the probability of a device user uses Active or Advance given the information on median income and population.



As seen from the figure, both age and median income have clear linear relationship with the transformed response, while population fails to reach the assumptions.

Although Population is significant predictor to the model, due to its low treatment effect on the probability of active or advance users, and its failure to reach the model assumptions. The predictor is removed from the model.

The new model is given by:

$ActiveAdvanceUsers \sim Binomial(N, p)$

$log(\frac{p}{1-p}) = \beta_0 + \beta_1 MedianIncome + \beta_2 Age$

Where $N$ represents the number of Device Users, and $p$ represents the probability of a device user uses Active or Advance given the information on median income and population.

**Table 3:** Summary Table of The Generalized Linear Model

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 0.647 | 0.048 | 13.450 | 0 |
| scales::rescale(hhld_median_inc) | -3.043 | 0.180 | -16.907 | 0 |
| scales::rescale(age) | 0.376 | 0.075 | 5.042 | 0 |

**Table 4:** Confidence Interval Table of The Generalized Linear Model

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | 0.553 | 0.741 |
| scales::rescale(hhld_median_inc) | -3.396 | -2.691 |
| scales::rescale(age) | 0.230 | 0.523 |

The model summary and confidence interval are almost identical to the previously proposed model, just without population as one of the model's predictor.

**Are Mingar Wearables Racist?**

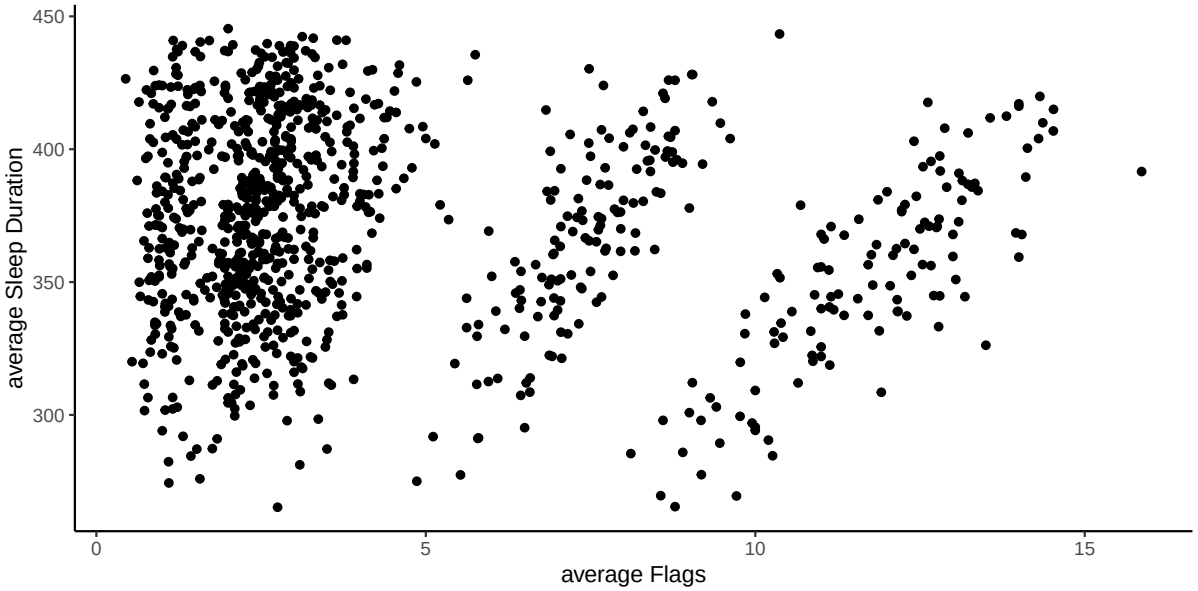



**Figure 6:** Relationship of average Sleep Duration and average Flags per User's Sleep

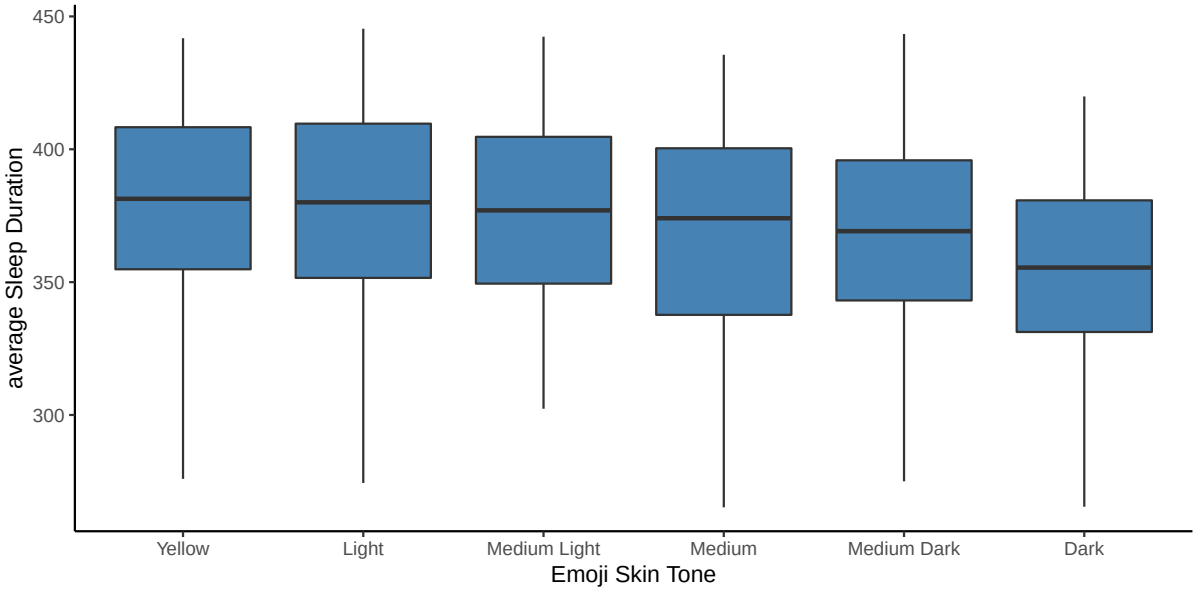



**Figure 7:** Distribution of average Sleep Duration per User by preferred Emoji Skin Tone

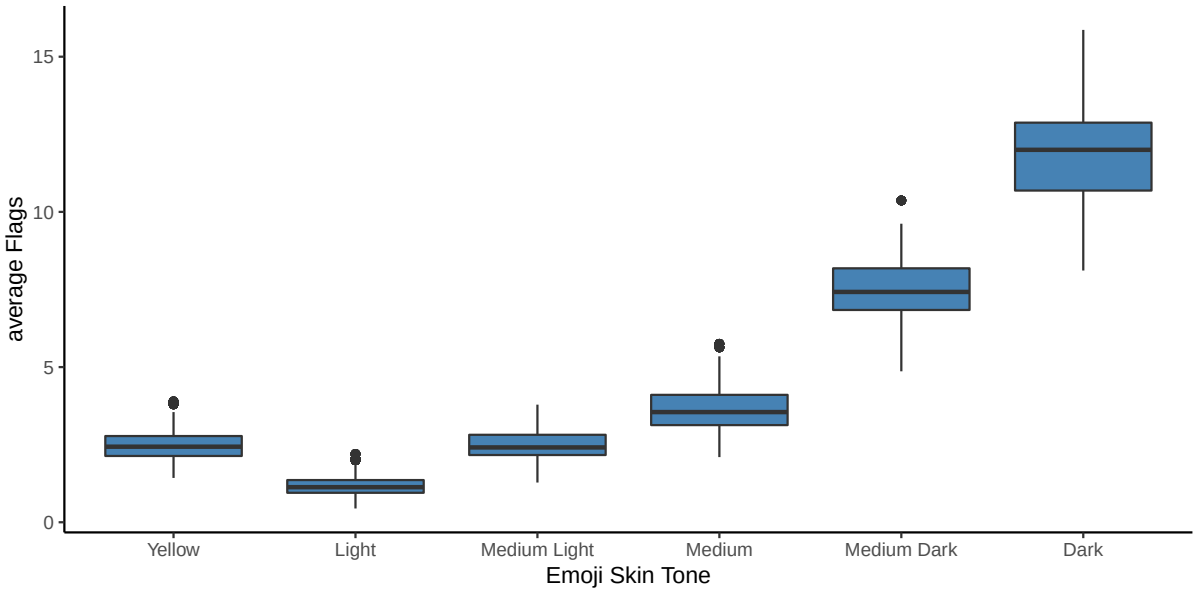



**Figure 8:** Distribution of average Flags per User's sleep by preferred Emoji Skin Tone

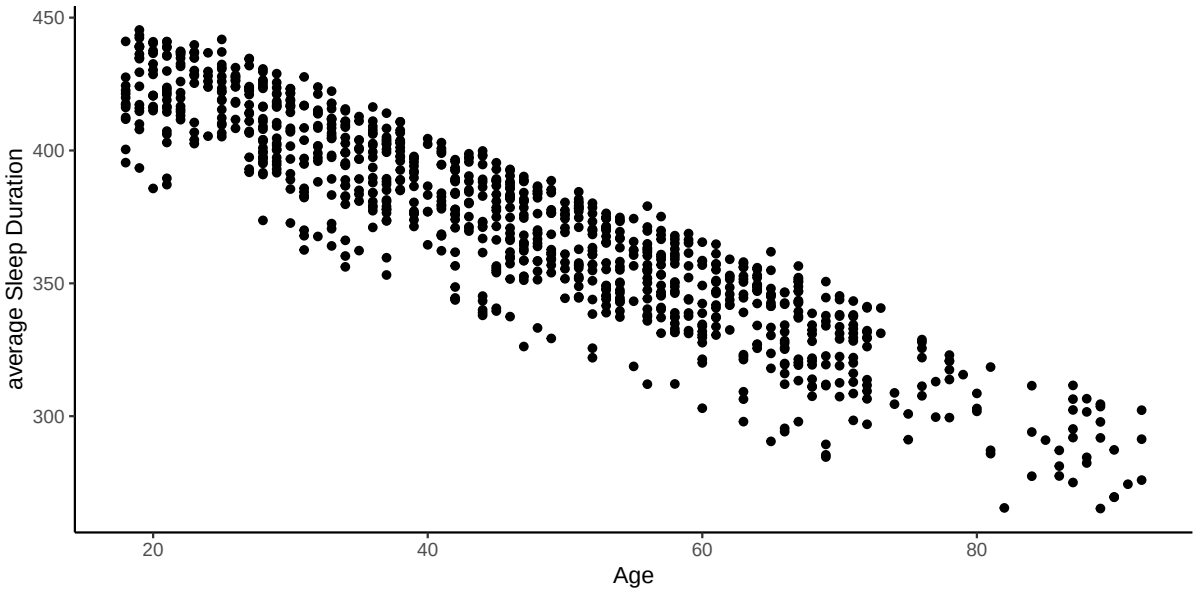



**Figure 9:** Relationship of average Sleep Duration and Age per User

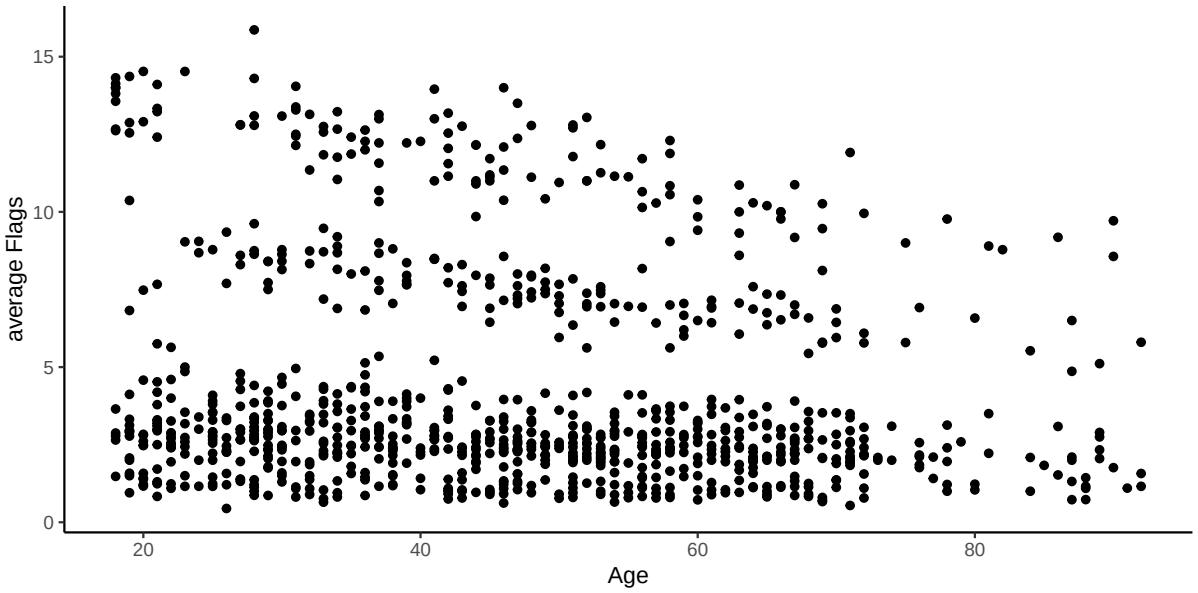



**Figure 10:** Relationship of average Flags and Age per User's sleep

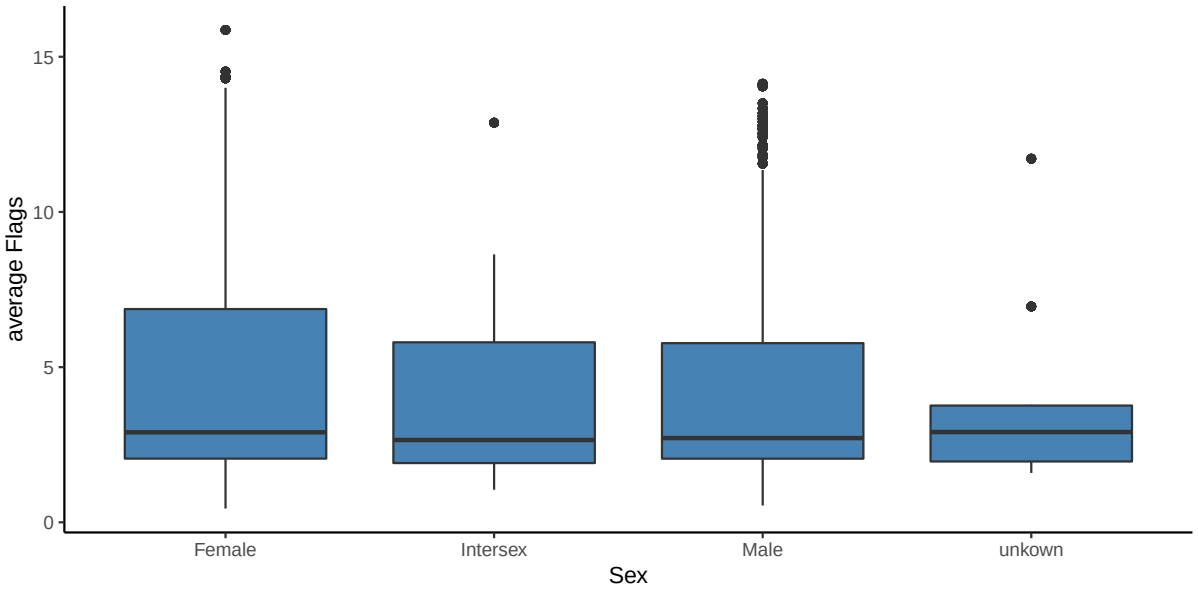



**Figure 11:** Distribution of average flags per User's Sleep by Sex

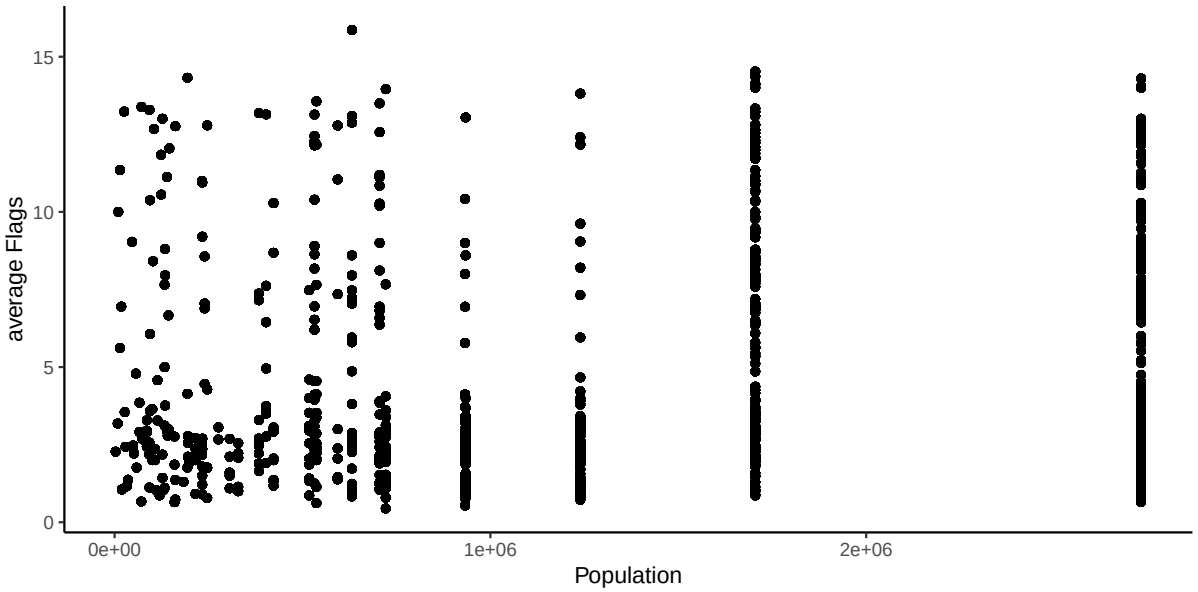



**Figure 12:** Relationship of average flags per User's Sleep and User's Postal Code Population

Model2 is given by:

$log(flags) = log(duration) + \beta_1 emoji\_colorYellow + \beta_2 emoji\_colorLight + \beta_3 emoji\_colorMediumLight + \beta_4 emoji\_colorMedium + \beta_5 emoji\_colorMediumDark + \beta_6 emoji\_colorDark + U_i$

Where $U_i$ represents random effect for $Customer_i$. Because only dummy variables were used as predictors, the report decides not to use intercept for this model, and there will be no reference level.

**Table 5:** Summary Table of The Poisson Regression Model

| effect | term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|---|
| fixed | emoji_colorYellow | -5.034 | 0.009 | -536.011 | 0 |
| fixed | emoji_colorLight | -5.790 | 0.016 | -365.287 | 0 |
| fixed | emoji_colorMedium Light | -5.017 | 0.012 | -416.901 | 0 |
| fixed | emoji_colorMedium | -4.614 | 0.011 | -435.055 | 0 |
| fixed | emoji_colorMedium Dark | -3.903 | 0.008 | -474.205 | 0 |
| fixed | emoji_colorDark | -3.400 | 0.007 | -463.509 | 0 |

```
##                                2.5 %    97.5 %
```

```
## emoji_colorYellow       -5.052770 -5.015953
## emoji_colorLight        -5.820857 -5.758727
## emoji_colorMedium Light -5.040940 -4.993764
## emoji_colorMedium       -4.634678 -4.593106
## emoji_colorMedium Dark  -3.918840 -3.886579
## emoji_colorDark         -3.414575 -3.385819
```

## Discussion

After selecting the most appropriate model for each question, we can now answer the 2 questions at hand.

Who Are the "Active" and "Advance" Line Wearables Customers? First, we have to answer what the distinct characteristics of the "Active" and "Advance" customers are. Through the final selected generalized linear model, we see the most significant character traits to be income, age and number of residents in the same neighbourhood. Furthermore, we see that there is a negative relationship between the active and advance customers binary response variable and income, which reasons that wealthier individuals buy less of these 2 lines, meanwhile lower income individuals buy more of them. In one of our plots above, we see that customers less than 20 and over 60 buy more of the active and advance lines than the others. This suggests that once again adolescents or young adults and seniors are more the target audience for these lines. This is congruent to our prior conclusions regarding income, i.e. individuals in those age groups are generally regarded as having lower income looking for more affordable products. As for the last characteristic, population within a neighbourhood, conclusions are hard to draw as the relationship is unclear, though notably it is negative.

Are Mingar Wearables "Racist"? In short, yes. Let's talk about it... Through our generalized linear mixed model, it was evident that the number of flags, which are quality issues that occur during sleep, related to sensor error or other data quality issues, increased for darker skin tones of overall the emoji modifiers; in exactly the order of Light, Medium Light, Medium, Medium Dark to Dark (least flags to most flags respectively). Moreover, out of the six models we considered, the most suitable model found the emoji modifier predictor to be the only significant one for the flag response variable, leaving us with the belief that the only consequential factor of these flags is the individual's skin tone. Our recommendation is that further tests on the hardware of the wearables should be conducted to determine the exact cause behind the skin tone sensitivity and lack of quality as a result. ### Strengths and limitations

Simplicity believes in transparency and as such, we find it important to disclose the limitations of our research and findings in order for our clients to feel confident with what they are receiving. Since we do not have information on individual income, we had to improvise. We cleverly pulled

income information based on postal code from the postal code conversion file (2016 census geography, August 2021 postal codes) provided to us as University of Toronto students. Thus, we only have a general sense of the wealth of a customer based on the neighbourhood they live in. It may not be exactly accurate per consumer considering that high income individuals may reside in low income neighbourhoods and the opposite is true too; however we believe that our analysis provides insights as to the average buying behaviours of your wearables customers, important when answering you first set of questions.

Another limitation we encountered was discovered when tackling your second set of questions regarding the racial complaints against Mingar's devices. Since, as mentioned, Mingar does not collect any data on the racial ethnicity of its users, we had to get creative with identifying darker complexion customers from lighter ones. We utilized the emoji modifier field populated by the customer. We must warn that some users have no data within the emoji_modifier field (labeled as NA) as they have presumably not selected one, so we had to work within a subset of the customer data. Moreover, it is not exactly accurate to suggest that all customers use the same skin tone emoji modifier as their own, because people have complex skin tones and may find it difficult to select the right shade. Or they may simply slect the emoji that they wish to express themselves with regardless of their own shades. However, we believe, once again, that the average consumer will selected a fairly similar emoji modifier to their own skin colour; and that was how we were able to proceed with our analysis and provide some meaningful results.

Not to toot our own horn, but Simplicity prides itself on being expert model selectors, whereby we consider numerous models, six, in fact, per question. Through an open and diverse lens, we consider various predictors when answering both questions of focus to ensure we do not exclude any important factors. Our strong foundational understanding of distributions assisted us in deciding what the underlying distributions of the datasets that we cleverly were able to wrangle, join together and explore for the most meaningful models. Most importantly, our company's strong core beliefs of accountability and integrity allows you to be assured that the final model selection was based on results that were authentic and reproducible with various illustrations, plots and tests to showcase the data and results.

# Consultant information

## Consultant profiles

*Complete this section with a brief bio for each member of your group. If you are completing the project individually, you only need to complete one for yourself. In that case, change the title of this section to 'Consultant profile' instead. Examples below. This section is only marked for completeness, clarity and professionalism, not 'truth' so you can write it as if we're a few years in the future. Put your current degree in as completed and/or add your first choice grad school program, whatever you like. What skills related skills would you most like to highlight? What job title do you want?*

**Aaisha Eid**. Aaisha is a senior consultant at Google. She specializes in reproducible analysis and statistical communication. Aaisha earned her Bachelor of Science, Specialist in Statistics Machine Learning and Data Mining, from the University of Toronto in 2022.

**Charles Lu**. Charles is a junior consultant with Eminence Analytics. He specializes in data visualization. Charles earned their Bachelor of Science, Majoring in Computer Science and Statistics from the University of Toronto in 2024.

## Code of ethical conduct

Simplicity ensures that any information shared between itself and its clients remain confidential and protected against sale for personal gain or benefit. Conflicts of interest will surely be communicated to clients transparently and accordingly. Personal information provided to by clients will be treated anonymously and respectfully. Finally, Simplicity is operated by proud statisticians that uphold the professional statistical standards of procedure and analysis.

## References

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/.

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.

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

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

Hao Zhu (2021). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.3.4. https://CRAN.R-project.org/package=kableExtra

David Robinson, Alex Hayes and Simon Couch (2021). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.6. https://CRAN.R-project.org/package=broom

Ben Bolker and David Robinson (2022). broom.mixed: Tidying Methods for Mixed Models. R package version 0.2.9.3. https://CRAN.R-project.org/package=broom.mixed

Chung-hong Chan, Geoffrey CH Chan, Thomas J. Leeper, and Jason Becker (2021). rio: A Swiss-army knife for data file I/O. R package version 0.5.26.

Hadley Wickham and Evan Miller (2020). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version 2.3.1. https://CRAN.R-project.org/package=haven

Hadley Wickham and Jim Hester (2020). readr: Read Rectangular Text Data. R package version 1.4.0. https://CRAN.R-project.org/package=readr

Hadley Wickham (2021). rvest: Easily Harvest (Scrape) Web Pages. R package version 1.0.2. https://CRAN.R-project.org/package=rvest

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.

Unicode, Inc. (2022). Unicode character table. Retrieved April 11, 2022, from https://unicode-table.com/en/

# Appendix

### Web scraping industry data on fitness tracker devices

The industry data on fitness tracker devices was web scrapped in R from https://fitnesstrackerinfohub.netlify.app/. the whole web scraping process utilizes the R packages "polite" and "rvest". informative user agents details including purpose of web scraping and contact information was provided to website before scraping. It is important to note that to respect the web scrapped contents, follow the terms and services and Robots.txt of the website and only take what is needed. Avoid web scraping at all if the target website provide public API to access the data. Respect the crawl limit suggested by the website. The web scraping industry data on fitness tracker devices are private data saved under the folder raw-data, and will never be shared and published on Github or any websites.

### Accessing Census data on median household income

The Canada Census data on median household income was accessed from https://censusmapper.ca/ through the APIs of R package "cancencus" and the private API key kindly provided by CensusMapper. The data is governed by the Statistic Canada Open Data license and Statistic Canada, and acknowledgment of them is required use the data. The Census data are licensed data saved under the folder raw-data, and will never be shared and published on Github or any websites. The API key provided by CensusMapper should also not be shared anywhere as it is a personal private key.

### Accessing postcode conversion files

The postal code conversion files was downloaded from https://mdl.library.utoronto.ca/collections/numeric-data/census-canada/postal-code-conversion-file through using University of Toronto student credentials. The file is governed by Statistic Canada and stored in University of Toronto online map and data library. Same as the Census data, the file is classified as licensed data saved under the folder raw-data, and will never be shared and published on Github or any websites.