An Investgation into MINGAR's New Customers and Possible Issues with Sensors

An exploration of the customers of the new Advance and Active lines, as well as a study into an issue with MINGAR's sensors and skin color.

Report prepared for MINGAR by Amo Mutant

2022-04-11

Contents

Executive summary

The questions we seek to answer in this paper are in response to the problems posed by MINGAR in regards to their devices. The first probelm was regarding the differences MINGAR can expect between the customers of their new, Active and Advance line, customers and their other customers. The second question answered is in regards to the issue of skin color influencing the sleep sensors of MINGAR's devices, causing some errors and flags to be returned.

Concerning the first question regarding differences between customers, we took the approach of comparing the Advance and Active line customers to each other, then grouping both of these customers and comparing to customers of the older device lines. The following results were found,

- Given the choice between Active and Advance devices and given a customers income is constant, customers born between 1970 and 1990 prefer Advance Devices more than customers born in other time periods (in comparison to Active Devices only).
- Similarly, if a customers age is kept constant, then the more household income the customer has, the more they are probable to buy an Advance Device (in comparison to Active Devices only).
- When grouping the Active and Advance line together and comparing them to older lines, we find that similar patterns emerge, i.e. keeping the customers age constant, the higher the houselhold income of that customer, the more they are expected to prefer the old device lines.
- This is also the case for customer age, as keeping a customers income constant, customers born between 1970 and 1990 are anticipated to buy more devices from the older lines.

In general, income is believed to be the main motivator for the choice of which line a customer will buy a device from. Age also seems to have an effect on this choice, however, a large portion of its effect can still be explained by differences in income. This is because customers born between 1970 and 1990 will tend to have stable jobs and higher income than customers born in another time, who might be young or retired.

For the second question, we ended up confirming MINGAR's concerns that skin color did indeed affect stability of the sleep sensors on their devices. From our studies, we found that skin color and date of birth were the main factors that influenced the amount of flags, with skin color having a more substantial effect. Our findings were as follows,

- As a customers skin tone gets darker, the customers device would produce increasingly more sleep flags per hour, given their age is kept constant. This can be clearly seen in the graphic represented bellow.

- It was also observed, though much less substantial, that if a customers skin tone is kept constant, an increase in a customers age would be expected to cause a decrease in the number of sleep flags per hour.

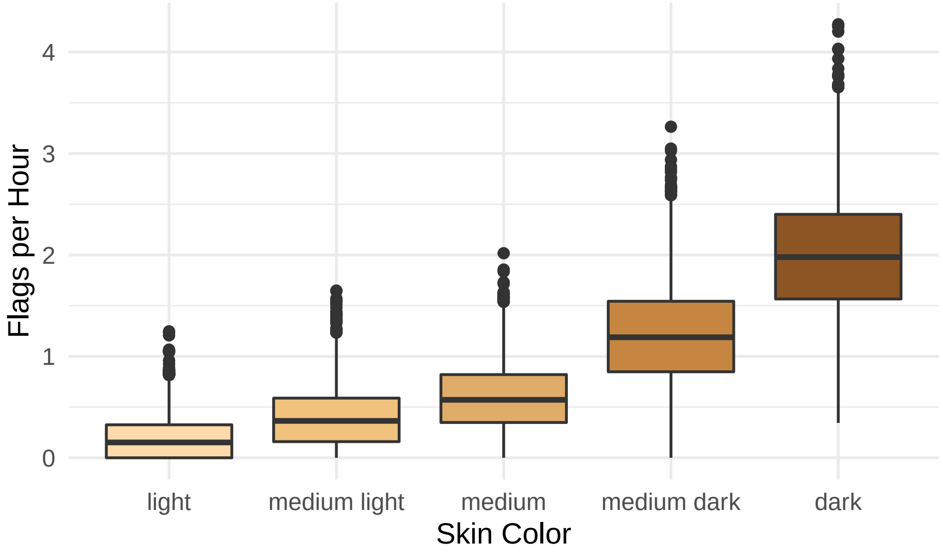

Figure 1: Boxplots for the number of flags per hour for each skin tone.

However, this study is not without its limitations, we were faced with following problems when completing the study,

- Multiple important the actual characteristics of consumers were approximated, which included household income of customers and their skin tones, and the accuracy of the approximations can influence the results.
- There are many confounding variables that could not be accounted for, such as any discount history of the devices or any sleep-related diseases a customer might have.

Above all, these limitations do not trivialize or nullify the results of this analysis, instead it opens avenues so that problems could be further explored in a more controlled setting. The results provided in this study offer realistic and accurate information on what the expected answer is to these questions.

Technical report

Introduction

This report will be divided into two parts, where each part concentrates on answering one research question.

The first part is related to the first problem discussed in the email. It will try to detail some information about the differences between the new and old customers using the provided data, as well as publicly available data. The study will involve an explanatory analysis into this question as well as some model fitting.

The second part will investigate the issues behind the complaints MINGAR is receiving in relation to errors with their sleep sensors. This section will explore some of the possible factors that could cause these, with the primary factor in question being the users skin color. This section will also have an explanatory analysis and model fitting, that will summarize the possible causes of the sensor issues.

Research questions

- How are the new customers of the Active and Advance line different compared to MINGAR's other customers and each other?
- Does there exist an association between sleep sensors failing on MINGAR devices and the characteristics, mainly skin color, of their users?

Difference between New and Old Customers

In order to begin talking about the differences in customers, we first have to talk about the characteristics of these customers. Provided to us by MINGAR, we have most customers' date of birth, postcode, sex, devices, etc. One major detail we are not given is the income of these customers. In order to approximate the incomes of these customers, we can use publicly available data about the median household income around the customers postcode (household). To do this we have used a dataset containing median income using Census Mapper and we combined it with a 2021 (most recent) postcode dataset from the University of Toronto. In combination with the customer data, the MINGAR device data, and the customer device data, we adequate information to start studying the differences in consumers.

The new dataset will contain all the available details about the customers including the median household income, their devices' names, and the line their devices belong to. Moreover, we remove all the rows with missing values and we rescale age, such that 1 is the oldest and 0 is the youngest, so the data is suitable to analyze.

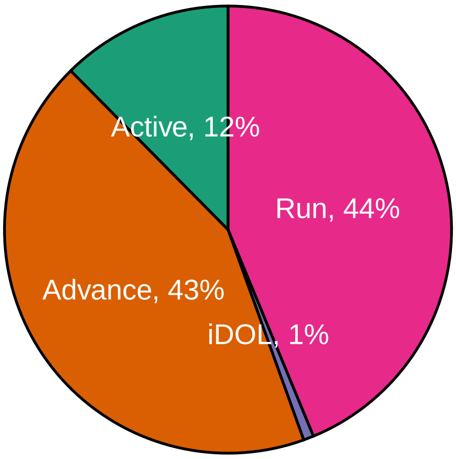

Figure 2: Pie Chart showing the distribution of the devices owned.

First we can perform an explanatory analysis of our data to get familiar with and find any areas of interest. As shown in the figure above most consumers own either a Run model or an Advance model, with a very small amount of people owning an IDOL model. It can also be seen that most people own models from the new Active and Advance line, making up 55% of the consumers being studied.

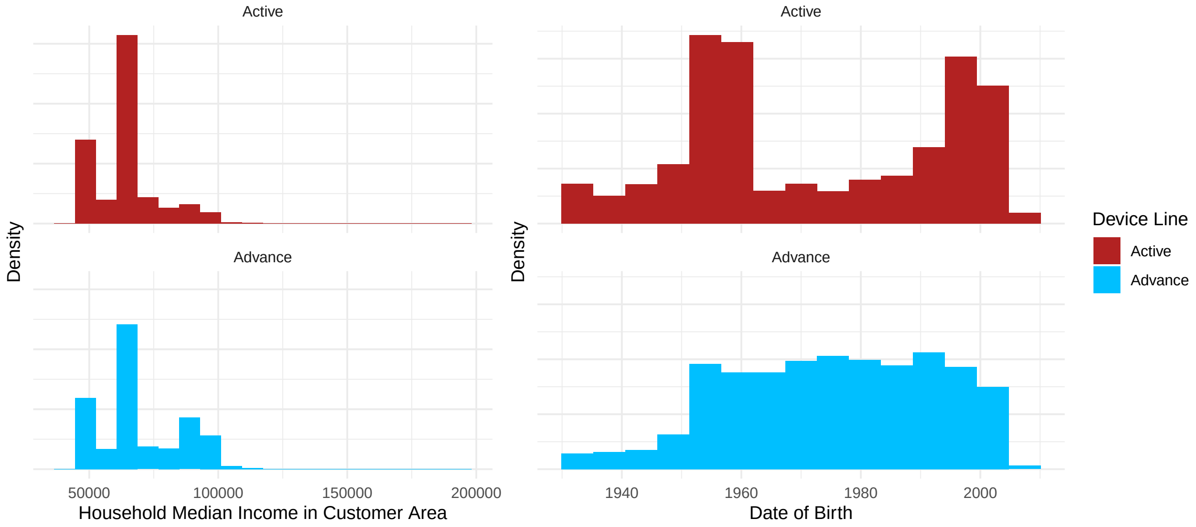

Figure 3: Density Histogram of Household Median Income and Date of Birth, colored in by line of the device bought.

Focusing on just the people who Active and Advance lines, we can see that people with low to mid-range incomes tend to prefer the active lines, while those with a bit higher incomes prefer Advance lines. As for date of birth, it seems that youger and older people prefer the active line, while the Advance line is equally preferred by most ages. One reason, for these observation is that the active line is usually cheaper than the advance line. This means the people who earn less money (students and retirees) prefer to spend less and buy device from the Active line. Other variables such as sex and gender were considered, but they didn't seem to have much impact on whether someone owned an Active or Advance device.

This is further supported after building a generalized mixed model that takes a binomial response that is 1 (true) if a customer owns a device active line and 0 if they own a device from the advance line. This model is,

$$log(\frac{p}{1-p}) = 1.36 - 0.507 \cdot d(1950-70) - 1.73 \cdot d(1970-90) - 0.367 \cdot d(1990+) - 2.81 \cdot 10^{-5} \cdot i$$

The variables mean,

- p: probability a customer owns the active model
- d(1950-70): a value that is 1 is someone is born in between 1950 to 1970, 0 otherwise.
- d(1970-90): a value that is 1 is someone is born in between 1970 to 1990, 0 otherwise.
- d(1990+): a value that is 1 is someone is born after 1990, 0 otherwise.

- i: median household income in the customers area

As can be seen from the model, generally age seemed to have the biggest impact on the odds of someone owning an Active device. Holding median income constant, the log odds of someone owning an active device for people born between 1950-1970 is 0.507 less than this born before 1950. Similarly, there are 0.367 less log odds for people born after 1990 compared to customers born before 1950. Meanwhile, people born between 1970 and 1990 have 1.73 less log odds of owning an active device. This supports what we observed in the histograms as people born 1970-1990 are a lot less likely to own an active device. Median household income also plays a role in deciding which device someone buys, which states that as income increases by one unit the log odds of owning an active device decrease by $2.81 \cdot 10^{-5}$ (all else held constant), which means they are less likely to own an active model as there income increases.

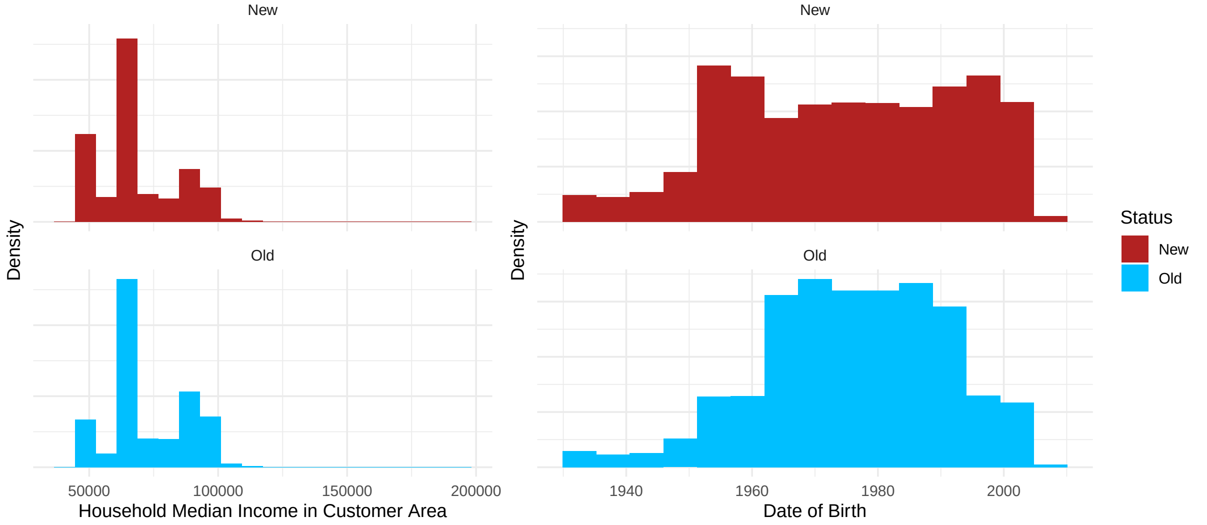

Figure 4: Density Histogram of Household Median Income and Date of Birth, colored in by whether the line of the device owned is new or old.

Moving our focus away from just the new devices, we can now compare how customers of the older an newer models differ. Similar to the previous histogram, we can see that age here has a some impact in deciding whether someone will buy a new or old device. It follows the same pattern as the previous histogram, where customers who are younger or older prefer the cheaper new devices. Moreover in general, it seems that customers with lower incomes are a lot more likely to buy the newer more affordable models in comparison to the people with higher incomes, who prefer some of the old models more.

Building a model similar to our previous analysis, but with a response that tells us if someone owns a new model, we find,

$$log(\frac{q}{1-q}) = 2.25 - 0.419 \cdot d(1950-70) - d(1970-90) - 0.203 \cdot d(1990+) - 2.04 \cdot 10^{-5} \cdot i$$

The variables mean,

- q: probability a customer owns a device from the new lines (active or advance)
- d(1950-70): a value that is 1 is someone is born in between 1950 to 1970, 0 otherwise.
- d(1970-90): a value that is 1 is someone is born in between 1970 to 1990, 0 otherwise.
- d(1990+): a value that is 1 is someone is born after 1990, 0 otherwise.
- i: median household income in the customers area

It can be seen that there the logs odds of someone owning a new model decrease by $1.98 \cdot 10^{-5}$ as income increases by one unit. This supports our finding that people with higher incomes tend to prefer the older device lines more. Similarly to our previous model, we can also see that people born between 1970-1990 have much lower log odds, and consequently much lower probability, of owning a new Active or Advance device keeping all other values constant. This could be a result of the factor previously discussed, where younger and older people, who earn less money would prefer the more affordable Active and Advance devices in comparison to other customers.

Issues with the Devices' Sleep Sensors

To investigate the answer to this question we will be using the data provided to us by MINGAR that tells us how many flags a users device returned during their different times they were asleep. Moreover, it is possible that a devices' release date or price might affect the qualities of the sensors. In order to acquire this data we can scrape the information from the Fitness tracker info hub, which contains all the specifications of these devices. We can combine this with the customer data, the customer device data, and the sleep data, to get all the information needed to start analysing the issues.

To make this dataset suitable for analysis, we must remove any missing values. Next, we need a way to guess the skin color of a customer based on the information we have. To do this we can reference their emoji identifier and map it to a suitable skin tone, which we could understand. We can also rescale the date of birth of the customer and release date of their device so that they are in the range of 1 to 0 (1 is the oldest). Finally, we can add a new column that contains the flags per hour a user's device woudl return as it will be useful to our analysis.

```
## `geom_smooth()` using formula 'y ~ x'
```

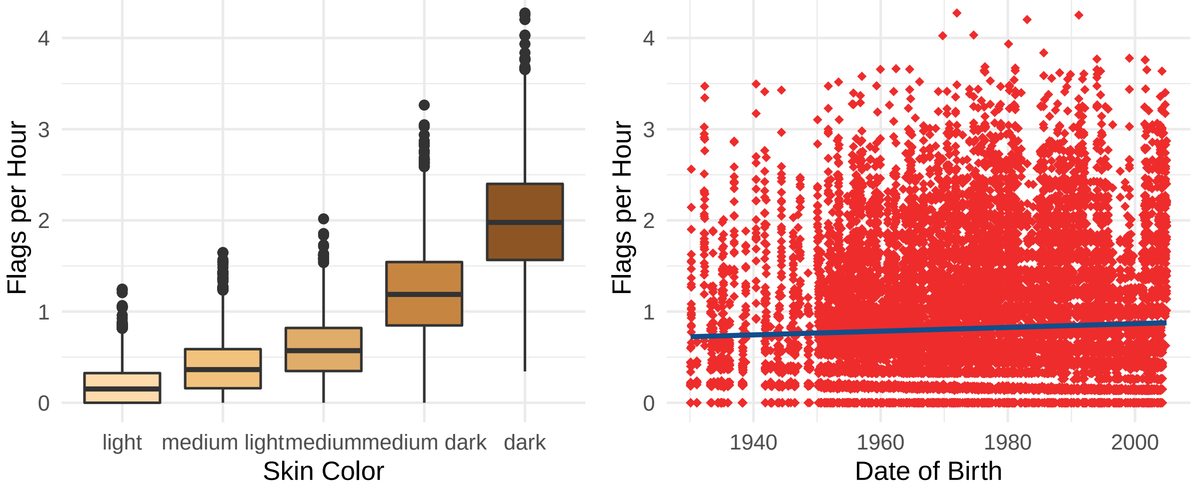

Figure 5: Boxplots of the number flags per hour for each persons predicted skin color and a scatterplot showing flags' association to date of birth.

A quick look at the flags per hour a device has against each definition of skin color we have, we find that the flags per hour consistently rises as the skin tones get darker. This hints toward the sleep sensors working worse with darker skin tones, but more analysis is needed before a conclusion. We can also that age might play some role in influencing the number of flags per hours, where older customers might be experiencing less flags.

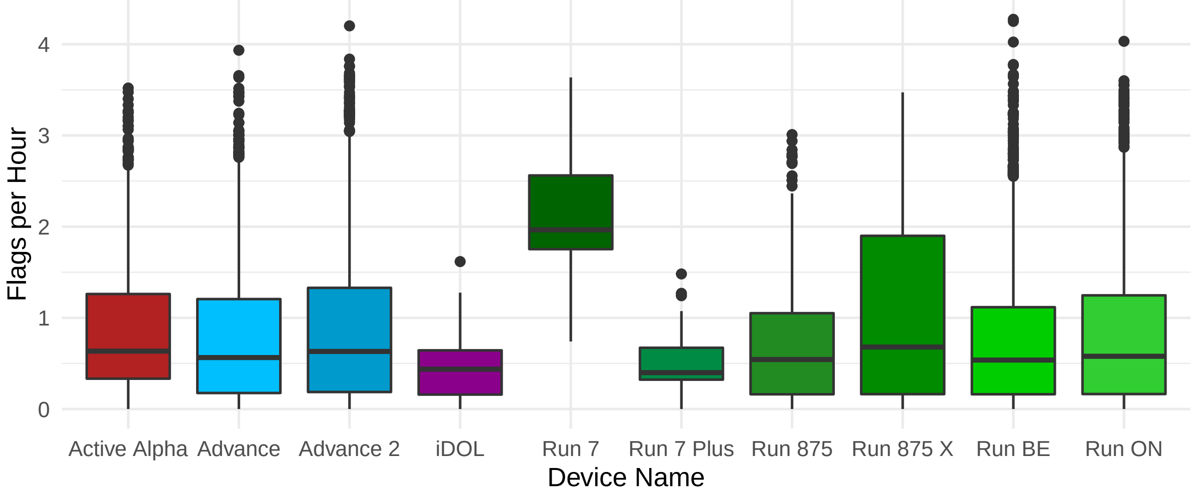

Figure 6: Boxplots of the number flags per hour for each device.

In comparison, looking at the devices owned we can see that the devices generally don't have much of an impact on the number of flags per hour. One outlier from this graph is the Run 7 device, which seems to have more flags per hour. However, further research into this shows that there is only one such device owned by customer, who turned out to have a darker skin tone. This means that the cause of this increase is most likely not the device, but the customers skin tone. As such, the devices type, release date, and price will all be ignored here on out as the type of device doesn't seem to play much of a role in the number of flags per hour.

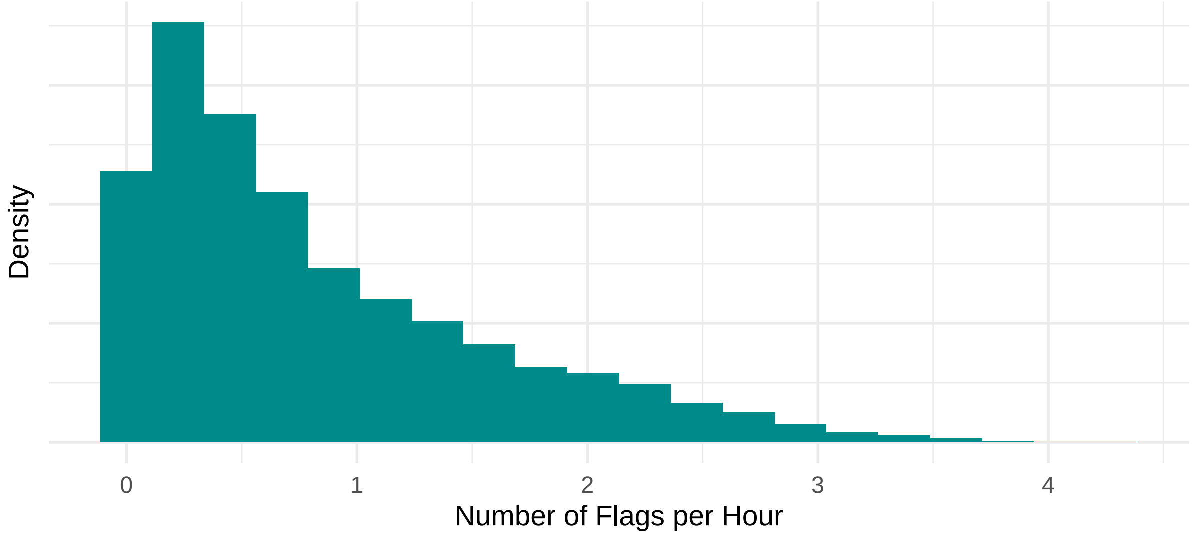

Figure 7: Density Histogram for the number of flags per hour.

This leaves us with fitting a model. A poisson model with flags as a response and duration as an offset seems to be an adequate model in our case. But first, we need to make that the flags per hour follow a poisson model. Looking at our density histogram we can see that the flags per hour does indeed follow a poisson distribution meaning that a poisson model is within reason.

Table 1: Table comparing the Mean Number of Flags per Hour to its Variance for each skin color

Skin Color	Mean Flags per Hour	Variance of Flags per Hour
light	0.1836871	0.0408311
medium light	0.3983598	0.0884292
medium	0.5947989	0.1241681
medium dark	1.2128428	0.2452702

Skin Color	Mean Flags per Hour	Variance of Flags per Hour
dark	2.0039631	0.3950364

However, looking at the mean and variance of the flags per hour under each category of skin color, we see that they are not equal. This deviates from the behavior we would expect from a poisson, so a negative binomial distribution will be fit instead in order to offset this effect. The resulting model will take the number of flags as a response and duration as an offset. Skin color and date of birth will be included as predictors, since skin color has been shown to have an effect on the result, and someones age would could cause changes in their sleeping habits. Moreover, a random effect for each customer will be included in order to account for their different sleeping habits. To do this each customer is assigned a numeric ID based on their customer ID to allow for model fitting.

The resulting model is,

$$\begin{aligned} log(flags) = &-5.77 - 0.0499 \cdot date + 0.777 \cdot (mediumlight) + 1.18 \cdot (medium) \\ &+ 1.89 \cdot (mediumdark) + 2.39 \cdot (dark) + ID + log(duration) \end{aligned}$$

The variables mean,

- flags: The number of flags
- duration: duration of the sleep
- date: the date of birth of the customer (rescaled from 1 to 0, with 1 being oldest)
- (medium light): a value that is 1 if someone has a medium light skin tone, 0 otherwise.
- (medium): a value that is 1 if someone has a medium skin tone, 0 otherwise.
- (medium dark): a value that is 1 if someone has a medium dark skin tone, 0 otherwise.
- (dark): a value that is 1 if someone has a dark skin tone, 0 otherwise.
- ID: It is a random effect with ID ~ N(0, 0.0003393). It separates each customer based on their numerical ID.

From our model, we can conclude that keeping all other factors the same, as someone ages they tend to get less flags per hour. However, mainly we can see, that keeping all else the same, as someones skin tone gets darker they tend to experience more flags per hour. In fact, people with dark skin tones experience 10.9 times more flags per hour than people with light skin tones. We are also 95% confident that, keeping all the other factors equal, people with darker skin tones experience 10.55 to 11.29 more flags per hour.

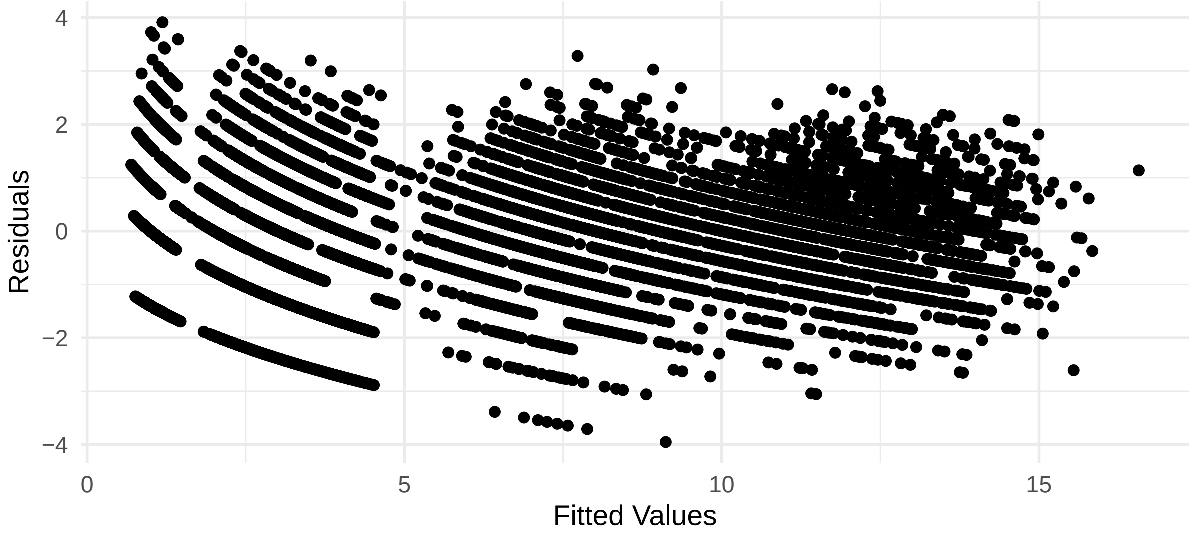

Figure 8: Residual plot of model fitted of the number of flags per hour.

We can also assess some of the faults in our model by studying the residual plot. It seems that a pattern is visible for lower fitted values, however, the pattern subsides as the fitted value increases. In general, this can cause some problems, however, it doesn't trivialize the result we found.

Discussion

In regard to our first research question about the difference in customers, we found that customers with higher incomes and who were born between 1970 and 1990 generally preferred the more expensive devices. This means that given a choice between an Active an Advance device, these customers would most likely get an Advance device as it is more expensive and would probably have more features. Similarly, if the choice was instead between the new affordable device lines and the older expensive lines, they would prefer the older device line for some of the same reasons.

For the second research question regarding issue with sleep sensors, results showed that indeed skin color does play a major role in causing sleep flags in MINGAR devices, with darker skin tones causing more flags. Moreover, we also found that age also play some role in role in the number of flags per hour, where older people tend to get less flags than younger people.

Strengths and limitations

Strengths

- Even without the model, the explanatory analysis and graphs showed very visible patterns that provided with a lot of answers in relation to our research questions.
- The explanatory analysis of the data and the models both agree on the factors affecting the data, which add more credibility to both.
- All the coefficients of the models built were significant at 0.05 significance level, which doesn't mean much in isolation, but it helps show that the models are on the right track.

Limitations

- A lot of the actual characteristics of consumers were approximated. We guessed the household income of customers and their skin tones. The accuracy of analysis is very dependent on whether are guesses were correct.
- There are a lot of confounding variables that could affect the data, especially with sleep scores. For example, some people might suffer from sleep walking, insomnia, or other sleep-related problems.
- Another such confounding variable is the possibility of discounts on devices when analyzing the differences between customers. If the the devices in any specific line were discounted previously, then it is possible that the customers only bought the device due to a discount on it.
- The residual plot shows a pattern for lower fitted values for the sleep data. This means that the model does have flaws and even though its general result can be trusted, its exact values and numbers shouldn't be taken at face value.

Consultant information

Consultant profile

Adam Musa. Adam is a senior data analyst with Amo Mutant. He specializes in Machine Learning and analyzing and documenting patterns in data. Adam earned her Bachelor of Science, Specialist in Computer Science with a Statitics Minor, from the University of Toronto in 2023.

Code of ethical conduct

- We seek to protect our clients and volunteers privacy, so no private information pertaining to these people will be disclosed or publicized without explicit approval.
- We will not be influenced by any monetary gain or public pressure in order to sway the results of our studies and we will state any interest that could possibly affect our judgment.
- We will be transparent in our methods and reasoning, so that our studies can be replicated and critiqued. We will take full responsibility for our studies and carefully consider any such criticism.

References

Software and Packages Used

- 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/.
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- 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.
- Hadley Wickham (2021). rvest: Easily Harvest (Scrape) Web Pages. https://rvest.tidyverse.org/, https://github.com/tidyverse/rvest.
- Dmytro Perepolkin (2019). polite: Be Nice on the Web. R package version 0.1.1. https://github.com/dmi3kno/polite
- Achim Zeileis, Torsten Hothorn (2002). Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. URL https://CRAN.R-project.org/doc/Rnews/
- Wood, S.N. (2017) Generalized Additive Models: An Introduction with R (2nd edition). Chapman and Hall/CRC.
- 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.
- 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.

Sources of Information

- Fitness Tracker Info Hub. (n.d.). Retrieved April 7, 2022, from https://fitnesstrackerinfohub.netlify.app/
- University of Toronto. (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
- von Bergmann, J., & Cervantes, A. (n.d.). Census Mapper. Retrieved April 9, 2022, from https://censusmapper.ca/

- Unicode. (n.d.). Emoji Charts. Retrieved April 9, 2022, from https://unicode.org/emoji/charts/full-emoji-modifiers.html
- SAPE, Software and Programmer Efficiency Research Group. (n.d.). ggplot2 Quick Reference: colour (and fill). Retrieved April 9, 2022, from http://sape.inf.usi.ch/quick-reference/ggplot2/colour

Appendix

Web scraping industry data on fitness tracker devices

To start scraping the website, I first searched the website itself to check if scraping was allowed or an api was available. There was no such information available. So to confirm I have permission to scrape, I established a connection to the website (https://fitnesstrackerinfohub.netlify.app/) and checked the robots.txt, which stated I have permission to scrape under some conditions.

```r
url <- "https://fitnesstrackerinfohub.netlify.app/"

# Check if we are allowed to scrape data
target <- bow(url,
              user_agent = "adamm.musa@mail.utoronto.ca for STA303/1002 project",
              force = TRUE)

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

Using the rvest and polite packages, I scraped all the information on the main page of the website. I then filtered this information out to include the table in the homepage, which is the only information I needed. I then converted this data into a format I can use for the rest of the study.

```r
# Get website data
html <- scrape(target)

# Filter data to only the needed table
device_data <- html %>%
  html_elements("table") %>%
  html_table() %>%
  pluck(1) %>%
  janitor::clean_names() # clean column names
```

Accessing Census data on median household income

I first visted the website (https://mdl.library.utoronto.ca/collections/numeric-data/census-canada/postal-code-conversion-file) and read through all the information about the data. I then selected the year I wanted the data from. I was prompted to login with my credentials and agree to therms and conditions. This allowed to choose the exact time period I needed for the data and download and Rds file, which I could import into my project. The Rds file was imported with tidyverse and the information saved to be used in the rest of the study.

```
# Load the data
dataset = read_sav("data-raw/pccfNat_fccpNat_082021sav.sav")

# Keep only needed information
postcode <- dataset %>%
  select(PC, CSDuid)
```

Accessing postcode conversion files

Looking at the website it was apparent that an api could be used to obtain the data we needed. After creating an account and agreeing to the terms and conditions, I began using their system in order to choose what data I needed. After selecting the regions and details I needed, I was given an api key and code that gives me the details I needed.

```
# Set up api data
options(cancensus.api_key = "CensusMapper_4085db21aef33e8ae9c3ce0426e411e1",
        cancensus.cache_path = "cache")
```

I constrained the data to the region level of Census Subdivision (level that was needed) and using the cancensus package, I retrieved all the data I needed from the 2016 map.

```
# get all regions as at the 2016 Census
regions <- list_census_regions(dataset = "CA16")

# Select necessary regions
regions_filtered <-  regions %>%
  filter(level == "CSD") %>% # Census Subdivision
  as_census_region_list()

# Get household median income
census_data_csd <- get_census(dataset='CA16', regions = regions_filtered,
                          vectors=c("v_CA16_2397"),
                          level='CSD', geo_format = "sf")
```

I then filtered it down to only information I needed (Household Income, Population, and CSDuid)

```r
# Simplify to only needed variables
median_income <- census_data_csd %>%
  as_tibble() %>%
  select(CSDuid = GeoUID, contains("median"), Population) %>%
  mutate(CSDuid = parse_number(CSDuid)) %>%
  rename(hhld_median_inc = 2)
```