Investigating Mingar Customer Demographic and Device Performance

An analysis of the buyers of the newer products and how skin tone affects device performance

Report prepared for MINGAR by PACK Consulting

2022-04-07

Contents

Executive summary **3**

Technical report **5**

Introduction 5

Analyzing the buyers of the newer and more affordable `Active` and `Advance` products 6

Analyzing the performance for sleep scores of Mingar devices relevant to skin tone . . 12

Discussion 20

Strengths **21**

Limitations **21**

Future Considerations **22**

Consultant information **23**

Consultant profiles 23

Code of ethical conduct 23

References **25**

Appendix **26**

Web scraping industry data on fitness tracker devices 27

Accessing Census data on median household income 27

Executive summary

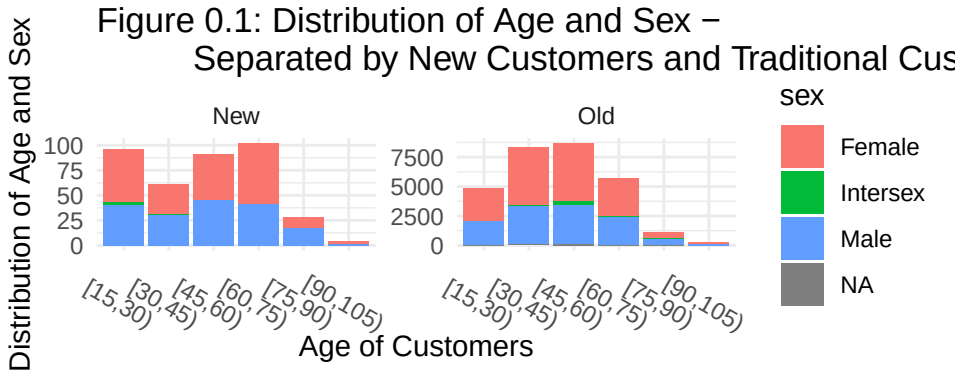
Through statistical support for the technology brand, Mingar, PACK consulting produced a thorough, and detailed analysis of Mingar's customers and technologies to provide key insights for marketing and product considerations. In our analysis, we address two core questions, who are the buyers of Mingar's newer, and more affordable "Advance" and "Active" devices and how do they differ from traditional customers, and whether performance for sleep scores of Mingar devices are related to skin tone. As well, throughout the analysis PACK ensured that the ethical treatment of all data was held to Mingar's high standards.

To address Mingar's first concern, who are the new customers and is there a difference between new and traditional customers, PACK made use of figures and statistical models to ensure a thorough understanding of Mingar's customers. The results of these statistical methods communicated:

- That the customers of Mingar's "Active" and "Advance" devices are primarily women between the ages of 15 to 30 years old, and 60 to 75 years old.
- Newer devices tend to be most popular between customers whose median household income based on their postal regions in the range $40,000 to $80,000 (all income quoted in CAD), followed by those with median household income of $80,000 to $120,000.
- Both household median income and sex were not significant predictors for what type of customer a person is.
- Age is a significant predictor for whether a customer will purchase a newer device, or one of Mingar's other devices.

We can understand this through the help of the visualization with respect to the first concern:

Figure 0.1 below plots the Distribution of Age and Sex, with separation by the type of customer. It can clearly be seen that for both new and traditional customers, it is majority females purchasing Mingar devices. As well, we can see that is mainly 60 to 75 year olds and 15 to 30 year olds purchasing Mingar's newer devices, whereas it is mainly customers between the ages of 30 to 60 years old purchasing Mingar's other devices.



Secondly, to undertake Mingar's second concern, determining if sleep score performance is affected by skin tone, a series of analyses were performed. The results of these analyses revealed that:

- There appears to be a strong correlation between skin tone and sleep scores, with the darker the skin tone experiencing poorer sleep scores.
- The average customer with the darkest skin tone experiences approximately 12 flags per sleep, while the average customer with the lightest skin tone experiences approximately 1 flag per sleep.
- The sex of the customer does not appear to impact the sleep score of the customer.
- There appears to be a strong negative correlation between sleep scores and the age of the customer, meaning that older customers are experiencing poorer sleep scores.

We can understand Mingar's second concern through the help of a visualization:

Figure 0.2 below plots the number of flags for each skin tone. It can clearly be seen that the number of flags increases for darker skin tones. This visualization appears to support the idea that there is a correlation between the number of reported flags and the tone of one's skin.

Technical report

Introduction

The following report will outline useful analysis of Mingar's company data, in order to most accurately answer the questions and concerns provided. The first question will be regarding customers of Mingar's newer, more affordable "Active" and "Advance" devices. Specifically, we will analyze who these customers are and how they differ from Mingar's traditional customers. To perform this analysis, we will first begin by compiling data of Mingar customers and their corresponding devices into a single dataset. We will then, use this data to form figures that will allow us to visually observe the sex, age, and median household income of both newer customers and traditional customers. Lastly, using the data we will implement a generalized linear regression model to determine whether there is a statistical difference between new and traditional customers.

The second question is in regards to the numerous complaints that the devices are performing poorly for users with darker skin, particularly with respect to sleep scores. To analyze this question, we will aggregate the data into a single dataset which contains each unique customer identifier, their personal characteristics and corresponding sleep information. This data will be used to implement a generalized linear regression model to determine whether or not there is statistical reason to believe that darker skin tones face more problems with Mingar devices, holding all else constant. We discuss how this model is chosen, and other potential factors that can influence the outcome of one's sleep score.

In the following sections we will discuss the data used for each question, some important summaries and trends in the data, as well as the methods and model choices implemented with the results of each analysis. The limitations to these findings will also be discussed, as well as how to potentially overcome these drawbacks.

Research questions

- How do the buyers of the newer and more affordable `Active` and `Advance` products differ from the traditional customer?
- Are Mingar devices performing poorly for users with darker skin, particularly with respect to sleep scores?

Analyzing the buyers of the newer and more affordable `Active` and `Advance` products

Data

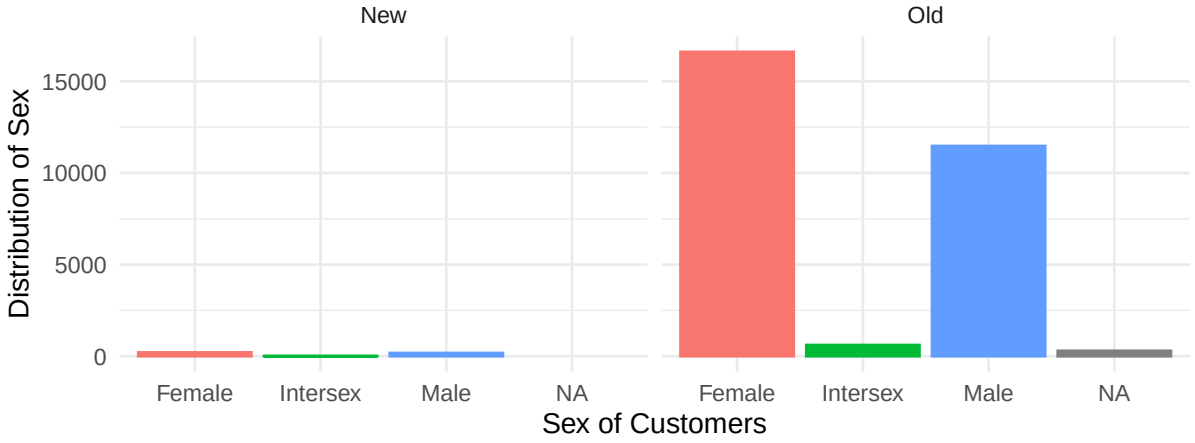
To observe whether there is a difference between customers of the newer, more affordable "Active" and "Advance" devices and traditional customers, we first began by collecting the necessary data through the merging of datasets. We merged the Customer, Customer Device, Device, Postcode, and Median Income datasets. The merging of these datasets is imperative so that we may get information on the Customers and the Mingar devices that they are purchasing.

Once we had our new dataset with all the required data, we began cleaning and mutating variables so that we may have more meaningful data. We first began by making a new variable that categorizes the new customers and the traditional customers based on which device they purchased. Then, we created a variable that cut the customer's ages into 15-year ranges, and we created a variable that cut the customer's household median income into $40,000 ranges. While cleaning our data, we choose to keep NA values as they provide insight into the Mingar customers. Keeping the NA values was imperative to the analysis as the NA values were present only in the variable for customer sex, and an NA in this variable may represent a customer who does not identify their sex as being Male, Female, or Intersex. Thus, to ensure no customers were misrepresented, or not represented at all, we have chosen to keep the NA values in our dataset.

After we completed the cleaning and mutation process, we could begin moving on to the last step in our dataset generation process which was extracting the variables of interest. The variables that provided the most insight into our analysis were the variables containing the information for customer ids, whether a customer is new or traditional, customers age, customers age range, customer's household median income, customers household median income range, customers sex, and the population of the region in which the customers lives. Thus, we were left with eight variables in our new dataset. Once we generated our data, we began creating plots so that we could begin visually observing who Mingar's new customers are, and how they differ from the more traditional customers.

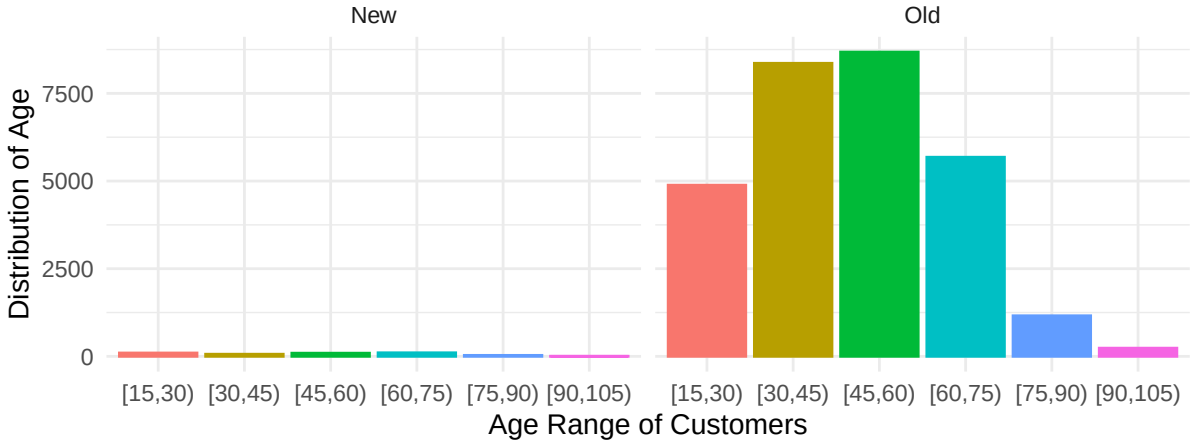
First, we began by creating figure 1.1 which displays the distribution of sex for new customers and traditional customers. We created this figure by plotting the sex of customers on the x-axis, and the distribution of sex on the y-axis. We then separated this plot by new and traditional customers so that we may visually compare how the sex of new customers compares to the sex of traditional customers.

Figure 1.1: Distribution of Sex − Separated by New Customers and Traditional Customers

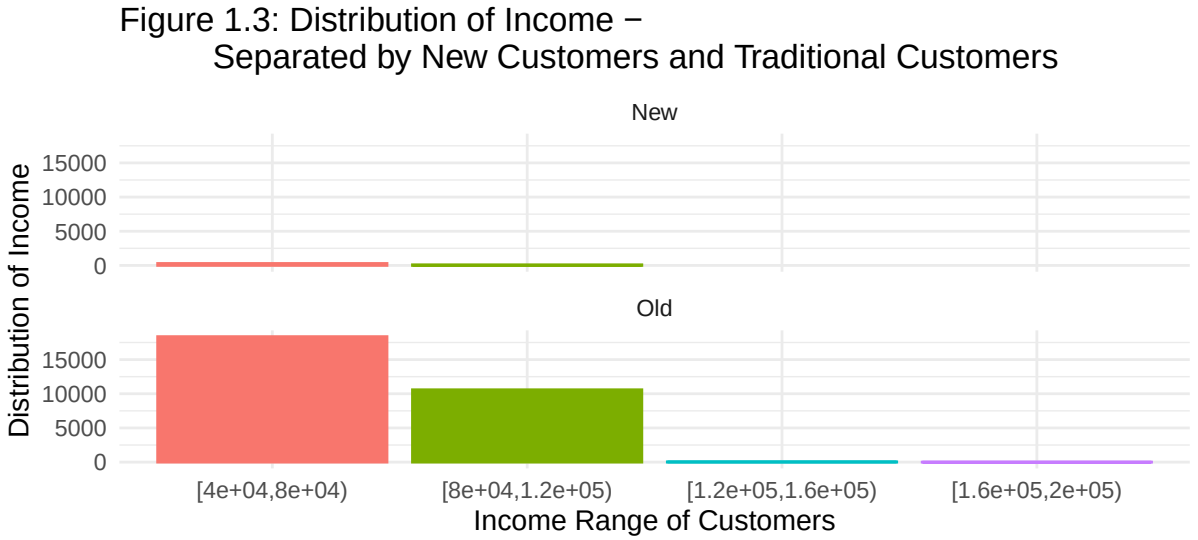


Secondly, we repeated this process in figure 1.2 shown below, which displays the distribution of age for new customers and traditional customers. Similarly to figure 1.1, we created figure 1.2 by making a histogram that plots the age range of the customers on the x-axis, and the distribution of age on the y-axis.

Figure 1.2: Distribution of Age − Separated by New Customers and Traditional Customers



Lastly, we repeated this process once more in figure 1.3 (shown below), which displays the distribution of household median income for new customers and traditional customers. Similar to figures 1.1 and 1.2, we created figure 1.3 by making a histogram that plots the household median income range of the customers on the x-axis, and the distribution of the household median income on the y-axis.



Methods

Next, to confirm whether there exists a difference between newer customers and traditional customers, we create a generalized linear mixed model. Since the use of this model is to determine whether there is a difference between customers, we will use the variable containing the information on whether a customer is new or traditional as our response. We choose a generalized linear model for our analysis as we have a binary response variable, as a customer can be classified into one of two categories; a new customer or a traditional customer. As well, we choose a generalized linear mixed model specifically as we need to account for random effects. Our main variables of interest are age and sex of the Mingar customers, and thus we will use them as predictors in our model. We use these two variables as we want to explore if age or sex has an effect on what devices Mingar customers are purchasing. From figure 1.1, we see that the distribution of sex for newer customers does follow a relatively similar distribution for traditional customers, and thus including sex in the model allows us to further explore this relationship. As well, from figure 1.2 we can also observe that the distribution of age range for newer customers does not follow the same distribution for traditional customers, but again, applying this model will help us further understand. As well, we wondered if the population of the region a customer resides in could possibly have an effect on what device a customer purchased and thus we formed our model according. The generalized linear model we have chosen to represent whether there is a difference between new and traditional customers in terms of age and income is:

Model is: $Y_i \sim Binomial(N_i, p_i)$

$$log(\frac{p_i}{1-p_i}) = X_i\beta + Ui$$

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

where,

- Y_i is the type of the i^{th} Mingar customer (New or Traditional)
- X_i includes the indicator variables for the variables age and sex
- β represents the coefficient matrix for fixed effects above
- p_i is the probability that the type of customers differs dependent on X_i
- N_i is the number of Mingar customers
- $U_{Population}$ represents the individual-level random effect that the population of the postal region a customer resides in has on the model.

We can perform the generalized linear model under the following assumptions. The first assumption is that the type of each customer (new or traditional) is independent of one another. The second assumption is that the random effect population comes from a normal distribution. The third assumption is that the random effect errors and residual errors have constant variance. Lastly, it is assumed that the log link function is appropriate for this model.

Results

As per the goal of this research question, we aim to discover how the users of newer more affordable devices differ from those who use older or more traditional devices. The variables of interest are the sex of the user, their age, median household income, and the population of the region which they habit.

As displayed in figure 1.1, which shows the distribution of sex separated by new customers and traditional customers, we can see that the distribution appears to be follow a similar pattern, regardless of if the customer is new to Mingar products, or a returning customer. That is, both newer and traditional devices were most popular with women, followed by men, intersex, and then NA. Next, figure 1.2 displays the distribution of age separated by new customers and traditional customers. Here we see that the newer models tend to be more popular between age groups of 15-30 and 60-75 years. Whereas, the more traditional models are most popular between age groups of 30-45 and 45-60. Finally, as seen in figure 1.3, the newer and traditional models tend to be most popular between those people who identify their median household income as in the range [40000, 80000), followed by those who identify their median household income as in the range [80000, 120000).

Given the relationship between newer and traditional customers identified above, we look to further analyze each by creating a generalized linear model, as discussed in the methods section. We hypothesized that:

H0: There is no difference between customers (newer) of the "Active" and "Advance" devices and customers (older) of the traditional devices

H1: There is a difference between customers (newer) of the "Active" and "Advance" devices and customers (older) of the traditional devices

Since our response variable follows the binomial distribution, we chose to use a generalized linear mixed model. After forming the necessary models we chose to perform a Likelihood Ratio Test to compare the models in order to decide the best predictors for our model. One model contained the Population variable (Let this be called Model 2) as a random effect and the other did not (Let this be called Model 1). To choose one of the model, we ran a likelihood ratio test, and below are the results.

Table 1: Likelihood ratio test for model 1 and model 2

Model	Df	LogLik	Chisq	Pr(>Chisq)
1: customer type ~ sex + age + household median income	5	-2020	-	-
2: customer type ~ sex + age + household median income + (1|population)	6	-2008	24.237	<0.001

Since $p < 0.001$, we have very strong evidence against the null hypothesis that the simpler model (Model 1) explains the data just as well as the more complicated model (Model 2). Thereby, suggestive of using Model 2 (the full/complex model) as our final model for analysis. After deciding on using the model with Population as a random effect, we come to the conclusion that household median should be dropped from the model since it was not significant (Let this be called Model 3). So, we used the likelihood ratio test to reach this decision. The results of this likelihood ratio test can be seen below:

Table 2: Likelihood ratio test for model 2 and model 3

Model	Df	LogLik	Chisq	Pr(>Chisq)
2: customer type ~ sex + age + household median income + (1|population)	6	-2008	-	-
3:customer type ~ sex + age + (1|population)	5	-2010	3.4378	0.0637

Upon closer inspection of the results, we can see that the p-value is 0.06 approx, which is between

0.05 and 0.1. This tells us that we have weak evidence against the null hypothesis that the simpler model (Model 3) explains the data just as good as the more complicated model (Model 2). This suggests we should hint towards using Model 3 which is the simpler model.Therefore, we will use Population as a random effect in our model and we will not use household income in our model.

So, our final model becomes: $Y_i \sim Binomial(N_i, \mu_i)$

$$log(\frac{\mu_i}{1-\mu_i}) = X_i\beta$$

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

In the above model:

- N_i is the total number Mingar customers
- Y_i is the type of customer, new or old
- β represents the coefficient matrix for the fixed effects
- X_i represents the covariates such as age
- $\mu_i$probability of a ring becoming damaged given X_i
- $U_{population}$ represents the the random effect that the population

Next, we want to look at the chosen model, whose summary can be seen below:

Table 3: Generalized linear model output

Variable	Estimate	Standard Error	P-value	Exponentiated Estimate	Exponentiated Confidence Interval
Intercept	5.0073	0.2396	<0.001	0.006	(93.663, 245.634)
age	-0.5845	0.2276	<0.05	0.557	(0.352, 0.859)
Intersex	0.3129	0.5115	>0.5	1.36738	(0.5017, 3.7263)
Malesex	-0.1986	0.1045	>0.05	0.819877	(0.6680, 1.0063)

Upon inspecting the summary of our chosen model, we can see a significant negative relation of age with p-value < 0.01 of the estimated odds ratio of -0.5845 (e^{-0.5845} = 0.55738). This can be interpreted as for each increase in age, the difference between old and new customers declines by 44.4%.The other variables do not have a significant relation with the number of old and new customers. Hence, we can conclude that age serves as a strong predictor for the odds number of new and old device customers.

We can also say that there is a statistically significant association between the age and the odds of them being new and old device customers since 1 is not included in e^{β_1}. The inclusion of 1 in the confidence interval for age would mean that odds of them being new and old device customers would not be affected by their age. We can be 95% confident that a 10% increase in the proportion of ages is related to a 0% to 22% decline in the difference between new and old device customers for the `Active` and `Advance` products.

In conclusion, the results of our analysis revealed that Mingar's new customers are primarily women between ages 15 to 30 years old and 60 to 75 years old, with a household median income of $40,000 to $80,000. As well, our analysis communicated that a customer's age is a strong predictor of whether they purchased the "Active" and "Advance devices or one of Mingar's other devices, specifically Mingar's younger customers are more likely than other customers to purchase the Mingar's new products.

Analyzing the performance for sleep scores of Mingar devices relevant to skin tone

Data

To analyze whether or not there is a correlation between the sleep score performance and skin tone, the appropriate data must be extracted and cleaned. To accomplish this, the customer device data is initially merged with the device description data. This updated dataset provides

which product each customer purchased and the description of that product. Next, this new dataset is merged with the customer's sleep data in order to discover the duration of sleep and number of reported faulty incidents with each product. Finally, this dataset is then merged with the customer information dataset, in order to add the characteristics of each individual.

Once the data has properly been merged into a single dataset, we could clean the data. The main factor we had to consider was how to handle the NA values. Given the significant number of observations, we chose to remove the observations with NA values entirely. We have no reason to believe that this decision forms any bias, and therefore removing these values should have no impact on the result of this study. The last alteration performed was to create an age variable to display the age of each customer. To do so, we subtracted the date of each observation by each customer's date of birth. This new variable displays the age of the at the time of the reported observation.

The last step in the data cleaning process is to extract only the variables of interest. Upon further examination of the research topic, the variables indicating the customer's identification, age, sex, skin tone identifier, along with their sleep duration and number of reported flags appear to be the only variables of immediate use. Therefore, only these six variables are kept for further analysis.

Figure 2.1 below plots the number of flags for each skin tone, with separation by the sex of the individual. It can clearly be seen that the number of flags increases for darker skin tones. This visualization appears to support the idea that there is a correlation between the number of reported flags and the tone of one's skin.

Figure 2.1: Number Of Flags For Each Skin Tone − Separated by Sex

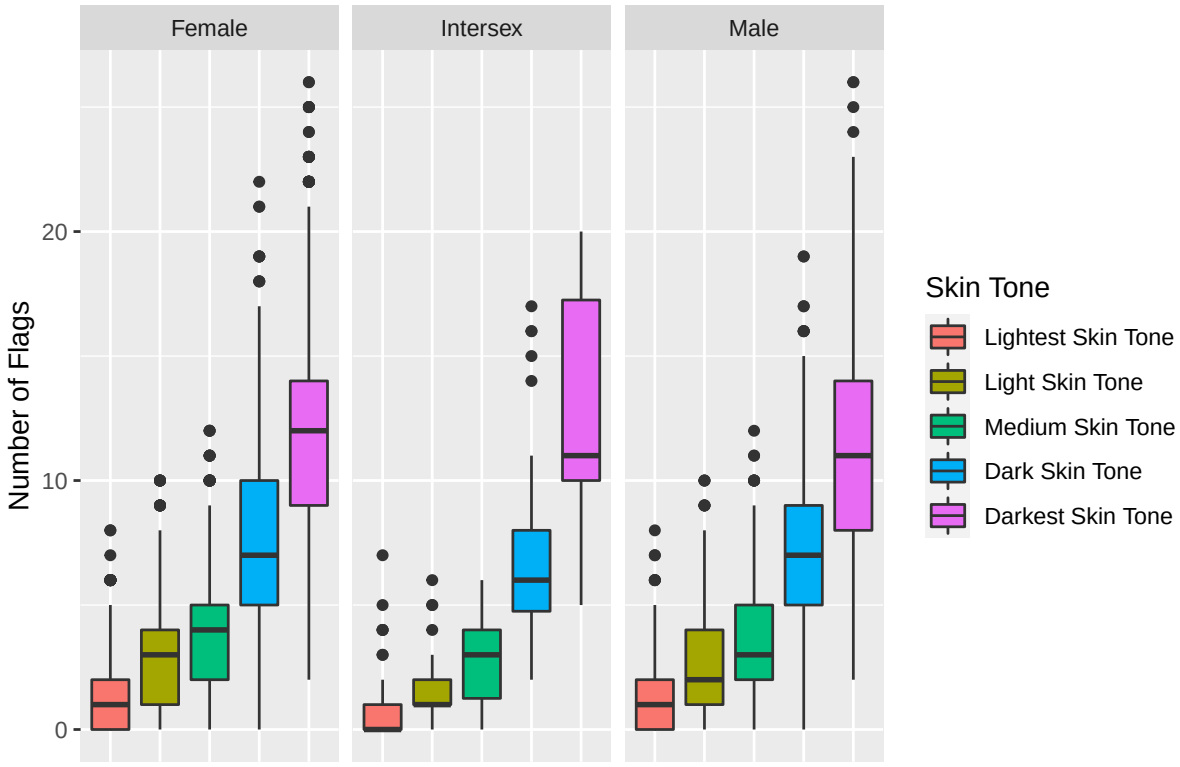
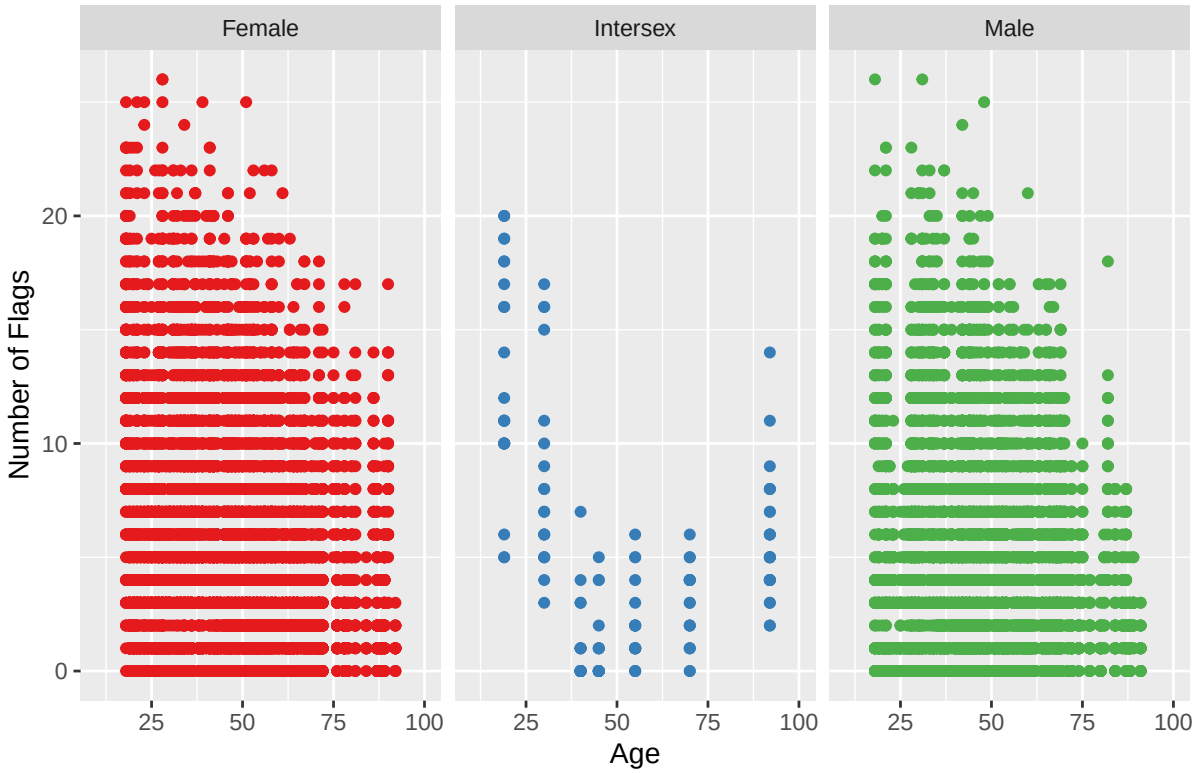


Figure 2.2 below plots the number of flags compared to the age of the age of the customer, with separation by sex. There appears to be a negative correlation between the two variables. This implies that there tends to be fewer reported flags for older customers.

Figure 2.2: Number Of Flags For Each Age Group − Separated by Sex



Methods

To further analyze the relationship between each customer's sleep score and their respective skin tone, a generalized linear model is performed. As the goal of this topic is to determine what influences a customer's sleep score, the variable representing the number of flags will be the response variable. A generalized linear model is chosen because the number of flags is a count variable, and therefore Poisson regression is needed for the analysis. This model is also appropriate because we need to account for random effects in the estimation.

The variable indicating the skin tone of the customer is the main variable of interest, so this will be included as a predictor in the model. Furthermore, 3 provides graphical rationale for including the age of the individual as a predictor for the number of flags. It is also reasonable to wonder whether the number of flags has anything to do with the sex of the individual.

Therefore, the generalized linear model to estimate the mean number of flags at time t will have the following form:

Our model: $Y_{ij} \sim Poisson(\lambda_{ij})$

$$log(\lambda_{ij}) = \mu + X_{ij}\beta + Ui + log(d)$$

We have: $Ui \sim N(0, \sigma^2)$

In the above model:

- X_{ij} includes the indicator variables for skin tone and sex, along with the numeric variable age.
- β represents the coefficient matrix for fixed effects above Y_{ij} is the number of flags for individual i for sleep scores observation j per unit of time (in minutes, since duration is our offset).
- U_i represents the individual level random effects for each customer
- log(d) is the offset by sleep duration.

Given that the number of flags recorded do not occur over equal sleep duration periods, the duration variable is chosen as an offset to account for this. Furthermore, given that there are several reported incidents for each customer in the dataset, the customer identification variable is chosen as a random effect. This is because each observation by the same customer is likely dependent of one another, and therefore treating this variable as a random effect will remove this bias from the mode

To determine whether or not the sex variable should be included in the model, a likelihood ratio test is performed on two models; one with the sex variable as a predictor and one without it. A likelihood ratio test assesses the goodness of fit of two models, where the predictors of one model are a subset of the other, and therefore the results of this test will aid in deciding which model is better suited for the given data.

We can perform the generalized linear model under the following assumptions. The first assumption is that each customer is independent of one another. The second assumption is that the random effect for customer identification comes from a normal distribution. The third assumption is that the random effect errors and residual errors have constant variance. Lastly, it is assumed that the log link function is appropriate for this model.

Results

Before delving into the results section, let us state our null hypothesis for this research question:

H0: There is no difference between skin tones with respect to Mingar device performance.

H1: There is a difference in skin tones with respect to Mingar device performance.

As aforementioned, the goal of this research question is to discover whether darker skin tones affect Mingar devices. Our statistical consulting team has also decided to include other variables that could be of interest, such as, age and sex. As discussed in the methods section, to determine which set of predictors will be used for a model,we performed a Likelihood Ratio Test. This test compares two models, one with the sex variable included as a predictor and one without.

To mention about our first plot,boxplot from figure 2.1 shown in the above section: We see the median number of flags is the lowest for the lightest skin tone and the median number of flags is the highest for darkest skin tone (gradually increases amongst darker skin tones). Therefore, we can deduce from this plot that more flags for darker skin tones indicates worse performance of the device which means worse sleep score reliability. At the same time, we can see some minimal relation between sex and skin tone for Mingar device performance. The boxplot indicates some evidence of a higher median of flags for darker skin tones with respect to females as opposed to males or intersex. This means that there may be worse performance of Mingar devices for females of darker skin tones compared to males or intersex. However, through model selection in the above section, `sex` variable is not a significant predictor for number of flags per unit of time (duration in minutes).

Furthermore, through figure 2.2, we again see that for females the number of flags for the age group between 25 to 100 is more compared to males or intersex. As a general trend, we can say that the number of flags for the younger age group which is from 25 to 40 has more flags, that is, worse sleep scores compared to the older age group of say 65+ which shows lesser flags, that is, better sleep score performance by Mingar devices.

Although these plots give us some early indications for variables of interest. We need to perform in-depth analysis to find out variables of significance in prediction. As mentioned in the methods section, we performed an LRT test for which the table is below:

Table 4: Likelihood ratio test for model 1 and model 2

Model	Df	LogLik	Chisq	Pr(>Chisq)
1: flags ~ skin Tone + sex + age + (1 | customer ID)	9	-32841	-	-
2: flags ~ skin Tone + age + (1 | customer ID)	7	-32482	1.961	0.375

Table (4) shows LRT being used to compare two models, one with the sex variable included as a predictor and one without. The p-value of approximately, 0.375 (not significant) indicates no evidence against the null that the simpler model (Model 1) explains the data just as well as the more complicated model (Model 2). More specifically, we have reason to believe that the model

without the sex variable is best suited, and therefore, our Model 2 (without the `sex` variable) which is a simpler model is the best one to proceed with.

After performing LRT, and deciding our final model. Here is our final model formula to estimate the mean number of flags at time t:

Our model: $Y_{ij} \sim Poisson(\lambda_{ij})$

$$log(\lambda_{ij}) = \mu + X_{ij}\beta + Ui + log(d)$$

We have: $Ui \sim N(0, \sigma^2)$

In the above model:

- X_{ij} includes the indicator variables for skin tone, along with the numeric variable age.
- β represents the coefficient matrix for fixed effects above
- Y_{ij} is the number of flags for individual i for sleep scores observation j per unit of time (in minutes, since duration is our offset).
- U_i represents the individual level random effects for each customer
- log(d) is the offset by sleep duration.

Now that an appropriate set of predictors is chosen, the final model is conducted. We used the `glmer()` function to determine the log-odds estimate and the confidence interval in response to the number of flags per unit of time (in minutes). The summary of the model is included below in Table (2b):

Table 5: Generalized linear model output

Variable	Estimate	Standard Error	P-value	Exponentiated Estimate	Exponentiated 95% Confidence Interval
Intercept	-5.773	0.017	<0.001	0.003	(0.0030, 0.0032)
Light Skin Tone	0.777	0.012	<0.001	2.174	(2.0912, 2.2599)
Medium Skin Tone	1.178	0.019	<0.001	3.248	(3.1301, 3.3713)
Dark Skin Tone	1.891	0.018	<0.001	6.624	(6.3988, 6.8588)
Darkest Skin Tone	2.390	0.017	<0.001	10.914	(10.5909, 11.2913)
Age	0.050	0.019	0.007	0.952	(0.9174, 0.9868)

This table (5) displays the results of the generalized linear model, which estimates the odds ratio for the number of flags at time t (in minutes). It can quickly be noted that each factor level for skin tone has a significant p-value (p-value<.0001) and therefore is a reliable predictor in the model. More specifically, the darker skin tones are positively associated with odds of Mingar devices performing poorly, with respect to sleep scores per minute. This finding follows the trend identified in figure 2.1. It can also be seen that the age variable has a significant p-value (p-value < 0.001), and therefore we have strong reason to believe that the age of the customer is a reliable predictor for the odds of Mingar devices performing poorly, with respect to sleep scores per unit of time (in minutes). Again, this finding directly follows the trend identified in figure 2.2.

Given the confidence intervals displayed in table (4), the interval for the estimates of skin tones, and age, respectively, e^{β_1} to e^{β_4} and e^{β_5} does not include 1, and therefore the model with age and skin tone is preferred to a model without these predictors. That is, age and skin tone is significantly associated with sleep scores (flags) per minute (duration) for mingar device performance.

From the table, we should note that the reference level is the lightest skin tone, so, with 95% confidence, we can claim that the odds ratio of the number of flags (sleep scores) per minute (duration) from the darkest skin tone is about 10 to 11 times higher than that of the lightest skin tone. As a note, by exponentiation of confidence intervals we obtain the multiplicative factor by which the mean count changes. We also exponentiated each coefficient above for ease of interpretation. The similar approach applies for the rest of the skin tone intensities as well.

It's analogous.

Since flags (our response) tell us about device behavior and as an overall conclusion we can say that more flags indicate worse performance of Mingar devices which means worse sleep scores reliability for darker skin tones compared to lighter ones.

Discussion

To summarize our findings for the first research question, we aimed to find out who Mingar's new customers are and how these new customers differ from traditional customers. To answer this question, first we merged the important datasets and wrangled the data to produce a final dataset containing all necessary information. Afterwards, we performed exploratory data analysis to explore the distributions of our parameters of interest, which were age, sex, and median household income of the customers. We then observed how the distributions for each parameter changed for newer customers versus traditional customers. Since our response variable, type of customer, is a binary variable with new customers being represented by a 0 and traditional customers with a 1, we use generalized linear models to represent our data. When checking the assumptions of our model, we also noted that the binary nature of our response variable had an interesting effects on the residual plot. The residual plot was separated into two clusters with all the new customers being at one section of the plot, while the traditional customers were represented in another section. After running summaries and likelihood tests, using different predictors and adding random effects we came to the conclusion that the model with two fixed effects: age and sex, and one random effect: population, best fit for our data. Since our model makes use of a random effect, we change it from a generalized linear model to a generalized linear mixed model. As well, we felt the need to add random effect, population to account for the variability in the data.

To summarize our findings for the second research question, we started with the merging of necessary datasets to put together a final dataset that incorporates the important variables of interest. In the next step, we built a model formula based on a generalized linear mixed model since our response variable is flags (which tells us about Mingar's device behavior), which is a count variable. To incorporate this type of response, we need a Poisson regression, and at the same time, we also need to incorporate a varying factor to model a real world scenario, that is, customer ID. The reason this is a random effect is because the number of flags (the sleep scores per unit of time (in minutes) differs for each person, and we need to account for that variability. Then, we performed some plots such as boxplot and scatterplot. This was to convey information about the number of flags for each skin tone, with separation by the sex of the individual, and the number of flags compared to the age of the age of the customer, with separation by sex, respectively. We then investigated our full model with all our variables of

interest, namely, age, sex, random effect of customer ID, log(duration) as an offset, and skin tone. However, we used a Likelihood ratio test to reduce the model into just, age, random effect of customer ID, log(duration) as an offset, and skin tone (Sex wasn't a significant predictor to keep). Based on the null hypothesis mentioned in the results section, our conclusion for the results is that, we have strong evidence against the null hypothesis that there is no difference between skin tones with respect to Mingar device performance. That is, more flags indicate worse performance of Mingar devices which means worse sleep scores reliability for darker skin tones compared to lighter skin tones.

Strengths and limitations

Strengths

Throughout the process of completing this report, we have found many strengths in our methods. A strength of our analysis was our use of an application programming interface (API). Through the use of an API we were able to easily make our data request and have it fulfilled by the computers without us having to know how all the background processes work. Another strength in the report is the use of the generalized linear mixed model, which allows us to account for a response variable which follows a Poisson distribution and handles random effects in the model. We performed a proper likelihood ratio test to compare models, in a significant attempt to choose the most reliable predictors.

Limitations

A key limitation in discovering if the devices perform poorer for those with darker skin tones is the method used to extract the color of skin. The skin tone identifier is based on the customer's created emoji and therefore might not accurately reflect the true skin tone of the customer.

Another limitation in the generalized linear models is potentially due to the set of predictors chosen. With numerous potential predictors to choose from given the Mingar data, a subset was initially chosen based on intuitive inference and therefore may not represent the most accurate set available. If further analysis was completed, a more rigorous variable selection method could potentially improve the model. Another limitation is in regards to the generalized linear model, as the residual plot does not appear to be entirely random. It is recommended that the relationship in the plot is further analyzed, to see if a key assumption for performing the model is violated.

As well, a limitation faced when discovering who Mingar's new customers are is, using a binary response variable prevents assumptions from being proved in a clear manner. The model chosen

to analyze the difference in the type of Mingar customer, was a binary variable, thus meaning it could only take on a value of 0 or 1. Because of this, the residual plots were hard to interpret and could not directly communicate whether the model fulfilled the assumptions of a generalized linear mixed model.

Future Considerations

For future work, it is recommended that the NA's (missing values) are better handled in the data cleaning process. For simplicity, we chose to remove all observations with an NA, however, this may have formed some bias in the process. Another potential approach could be to fill the NA value with the estimate or average of that variable, instead of removing the observation altogether. Furthermore, one could look deeper into the manner and discover why the NA values are occurring even occurring in the first place. For example, is a certain demographic of customer not providing an emoji (relevant to skin tones)? This would mean their skin tone cannot be determined, and therefore cannot be used for the purposes of the analysis. This could create a potential confounding variable which can alter the outcome of the analysis.

Furthermore, it is also recommended that one should look more critically into the predictors chosen in each of the generalized linear models. One could look to create a larger set of potential predictors from the given data, and perform a more in-depth model selection process (AIC, BIC, etc) to analyze whether there is a more appropriate subset of predictors for each of the generalized linear models. Another potential limitation is with the model predicting the odds number of flags as there is evidence of overdispersion. For the response variable, flags, the variance was significantly greater than the mean and therefore the assumption that this variable follows a poisson process may not be accurate. For future purposes, one could look to employ a negative binomial generalized linear model which can account for overdispersion.

Hence, for future considerations, more care must be taken into the type of variable the response variable is for ensuring that model assumptions can be clearly checked, and interpreted. This can be remedied by mutating the variable chosen for the response so that it is of a different type that will produce easier interpreted information. This consideration also applies to linear model of other types, i.e. Linear Mixed Models, General Linear Mixed Models.

Consultant information

Consultant profiles

Purumidha Sharma. Purumidha Sharma is a junior data analyst consultant with PACK Consulting. She specializes in reproducible analysis, data visualization, and provides actionable insights for the business industries. Purumidha earned her Bachelor of Science, Majoring in Statistics, minoring in computer science, and minoring in Mathematics from the University of Toronto in 2022.

Cameron Dietzel. Cameron Dietzel is a junior consultant with PACK Consulting. He specializes in reproducible analysis and statistical communication. Cameron earned his Bachelor of Science, Specializing in Mathematics while minoring in Statistics and Economics from the University of Toronto in 2022.

Kyra Chow. Kyra Chow is a junior consultant with PACK Consulting. She specializes in reproducible analysis and statistical communication. Kyra earned her Bachelor of Science, Majoring in Cognitive Science and Statistics and minoring in Computer Science, from the University of Toronto in 2022.

Aliza Aziz Lakho. Aliza Lakho is a junior data analyst consultant with PACK Consulting. She specializes in reproducible analysis and visualization to provide real world business solutions. Aliza earned her Bachelor of Science, Majoring in Statistics and Mathematics, and a minor in Computer Science from the University of Toronto in 2022.

Code of ethical conduct

By establishing the code of conduct for PACK Consulting, all our members agree on making this an integral part of this organization and we acknowledge your important role in helping us protect our most valuable assets. The code of conduct below ties with abiding to the ethical workings of this statistical consulting company:

1. The team at PACK Consulting believes in honesty and transparency of data. Any suspected limitation, biases in the data or the methods must be communicated to our stakeholders and clients on the potential impacts in interpretations, recommendations and conclusions.
2. The team at PACK Consulting takes integrity of data and integrity of professionals very seriously. Data analyzed should not be manipulated to get statistically significant results, and professionals will face severe consequences if this occurs.
3. The team at PACK Consulting protects confidentiality of data and respects any additional confidentiality regulations between PACK Consulting and the client. We will not use the

data provided by the client for any other purposes even when our contract ends. We focus on relationship building and trust with our clients.

References

1. Bergmann, J. von, & Cervantes, A. (2021). "Population density". Census Mapper. Retrieved from https://censusmapper.ca/
2. *Postal code conversion file*. Postal code conversion file | Map and Data Library. (n.d.). Retrieved from https://mdl.library.utoronto.ca/collections/numeric-data/census-canada/postal-code-conversion-file
3. Fitness tracker info hub. (2022). Retrieved from https://fitnesstrackerinfohub.netlify.app/
4. 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.
5. 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.
6. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
7. 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/.
8. 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.
9. Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society (B) 73(1):3-36
10. Hadley Wickham (2021). rvest: Easily Harvest (Scrape) Web Pages. https://rvest.tidyverse.org/, https://github.com/tidyverse/rvest.
11. Dmytro Perepolkin (2019). polite: Be Nice on the Web. R package version 0.1.1. https://github.com/dmi3kno/polite
12. 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

As PACK consultants, we are here to provide consultative guidance on client's research focus. We gather input and data from the client because gathering information is needed in order to provide recommendations or develop new strategies. We sometimes need to make use of market data, but make sure it's properly credited and it's allowed to be used. Our entire team works together to then pursue and refine the analysis, to then produce a professional report to explain our recommendations with supporting analysis and visualizations. We then present our recommendations and findings to the client and/or stakeholders.

PACK Consulting takes seriousness in protecting licensed data from the clients/stakeholders. This means an agreement between our company and the client/stakeholder for protecting our confidential data as well as protecting their confidential data. We will prohibit usage of the data apart from its necessary use. Our team focuses on crediting/acknowledging the rights of the data from the licensee and having a document that approves the permitted use of the data.

As part of ethical considerations for web scraping, we have only collected information from the data source, and nothing else. As an important note, if a public API is present, we use that instead of web scraping. This also depends on T & C (Terms and Conditions) for a website we want to scrape. Reading through it can help us understand what's allowed or disallowed for web scraping. To give context, our company web-scraped after thoroughly checking to see if scraping is prohibited or not. We gathered that web-scraping for fitness tracker website is allowed, and we made sure we put a user agent string by one of our PACK team members. We also restricted the "crawl limit" to 5 seconds since nothing was specified on the website. This is a polite default. Furthermore, we also needed access to information about Canada's census geographies, and we were able to get a public API through which we gathered our data. Thereby no need of scraping the website for Census geographies website. We then prepared both the datasets in a useful format to start focusing on the research questions as highlighted by the Mingar company client.

As part of PACK Consulting, we believe in a few ethical scraping rules as follows:

1. We always provide a User Agent string that makes our company's intention clear and provides a means for contacting us with any questions or concerns.
2. We respond promptly to the outreach and work with you towards a solution.
3. We only save the data we absolutely need from the page, and we do not breach privacy or manipulate information known to us in any way.

Web scraping industry data on fitness tracker devices

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

# this code is updated appropriately to provide informative user_agent details
target <- bow(url,
              user_agent = "PACK Consulting purumidha.sharma@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
```

```
## <polite session> https://fitnesstrackerinfohub.netlify.app/
##     User-agent: PACK Consulting purumidha.sharma@mail.utoronto.ca for STA303/1002 project
##     robots.txt: 2 rules are defined for 2 bots
##    Crawl delay: 12 sec
##   The path is scrapable for this user-agent
```

```
html <- scrape(target)

device_data <- html %>%
  html_elements("table") %>%
  html_table() %>%
  pluck(1) # added, in case we get a list format
```

Accessing Census data on median household income

```
# install.packages("cancensus")


options(cancensus.api_key = "CensusMapper_a7378c10c7d7ca6abebf450419b73937",
        cancensus.cache_path = "cache") # this sets a folder for your cache


# get all regions as at the 2016 Census (2020 not up yet)
regions <- list_census_regions(dataset = "CA16")
```

```
## Querying CensusMapper API for regions data...
```

```
regions_filtered <-  regions %>%
  filter(level == "CSD") %>% # Figure out what CSD means in Census data
  as_census_region_list()

# This can take a while
# We want to get household median income
census_data_csd <- get_census(dataset='CA16', regions = regions_filtered,
                          vectors=c("v_CA16_2397"),
                          level='CSD', geo_format = "sf")
```

```
## Reading vectors data from local cache.
```

```
## Reading geo data from local cache.
```

```
# 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)
```