Canadian Wearables Product and Market Report

Key findings on marketing and product development issues for MINGAR

Report prepared for MINGAR by Beavertail Consulting Group (BCG)

2022-04-07

Contents

Executive summary

As a former military equipment supplier, MINGAR has developed various products aimed at outdoor recreation, including high-end fitness tracking wearable devices. Through research with data analysis and model construction, this study offers key insights for marketing and product considerations for MINGAR. The methods, findings, recommendations, and limitations are provided in this report prepared by Beavertail Consulting Group (BCG).

Marketing suggestions on new customers and traditional customers

To compete with Bitfit, which offers lower price devices, MINGAR has developed two product lines with a more approachable price: "Active" and "Advance." Through data visualization and a generalized active model, we estimate the characteristics of the new customers compared to the traditional customers.

The key findings of the study on customers are summarized below:

- According to the summary of new customers and traditional customers in table 1, new customers have higher average age and lower average median income compared to the traditional customers.
- In new customers, female customers are greater than male customers which are 6967 and 4336 receptively.
- Based on the generalized linear model, we estimate that new female customers are 1.3 times greater than male customers. We also estimate that new customers are more likely to have medium skin color.
- The model estimates that new customers have lower medium income and older than traditional customers.

Therefore, based on our findings, MINGAR's new product lines have attracted customers with lower income. The marketing team may focus on advertising towards older people, female, and diverse customers. The more affordable devices may help MINGAR in the low-end market and gain more market shares.

Product development suggestions to address complaints from social media team

The devices have received complaints about its poorly performance for users with darker skin, especially with respect to sleep scores. We construct a generalized linear model to explore this issue. We estimate that flags are more likely to happen to customers with darker skin.

Table 1: Key Statistics for New and Traditional Customers

Customers	Number	Average Age	Average Median Income
New customer	11496	48.19555	69581.64
Traditional customer	8916	44.96209	71777.37

Table 2: Average Flags per Duration for Different Athnicity

Skin	Average Flags	Average Flags per Duration
Dark	11.792948	0.0333994
Medium-dark	7.435916	0.0202140
Medium	3.653179	0.0099133
Medium-light	2.506692	0.0066393
Default	2.473206	0.0065259
Light	1.153909	0.0030615

The key findings of the study on devices performance are summarized below:

- According to table 1, average flags and average flags per duration for darker skin customers are significantly higher than light skin customers.
- Based on data visualization, the number of quality flags is likely associated with customer's ethnicity.
- The generalized linear model estimates that darker skin customers are more likely to have flags.

Therefore, our findings indicate that MINGAR's devices have poor performance for customers with darker skin. The results prove the complaints from the social media team. MINGAR's future product development should consider addressing the problems regarding skin color.

In conclusion, this report provides insights into MINGAR's product development and marketing strategies through statistical analysis. There are also some limitations regarding our statistical methods, which may potentially affect the p-values of the results. The key findings of the report are summarized in the tables.

Technical report

Introduction

This report summarizes all the statistical modeling and major analysis results associated with new lines of MINGAR's fitness tracking wearable devices and customers. The primary purpose of this report is to investigate the buyer persona of traditional products compared with new line products "Active" and "Advance" and the issue of device performance related to customer's ethnicity. Additionally, this report is designed to analyze the current problem and situation to gain market share and improve competitiveness with Bitfit. This strong competitor only focuses on individual fitness wearable devices. The customer data are provided by MINGAR, with a total of 20412 customers and 13 variables. The source of device data is from Fitness Tracker Info Hub, with 26 devices from MINGAR and Bitfit. Median income data is procured through the Census Mapper API. Postcode data is sourced through U of T Libraries.

Research questions

The primary research questions of this report are organized as follow:

- What are the characteristics of the new customers of MINGAR and how are they more affordable "Active" and "Advance" product different to our traditional customers?
- How are MINGAR's devices perform on customers from different ethnicity with respect to sleep score?

Marketing Studies and Suggestions

Data Wrangling

Some process of data manipulation for marketing research has been taken, data read, and cleaning is the first step for manipulation, saved as new_customer database. Using four different factors as our prediction factors to predict how they affect the customer behavior in the product.

Data Visualization

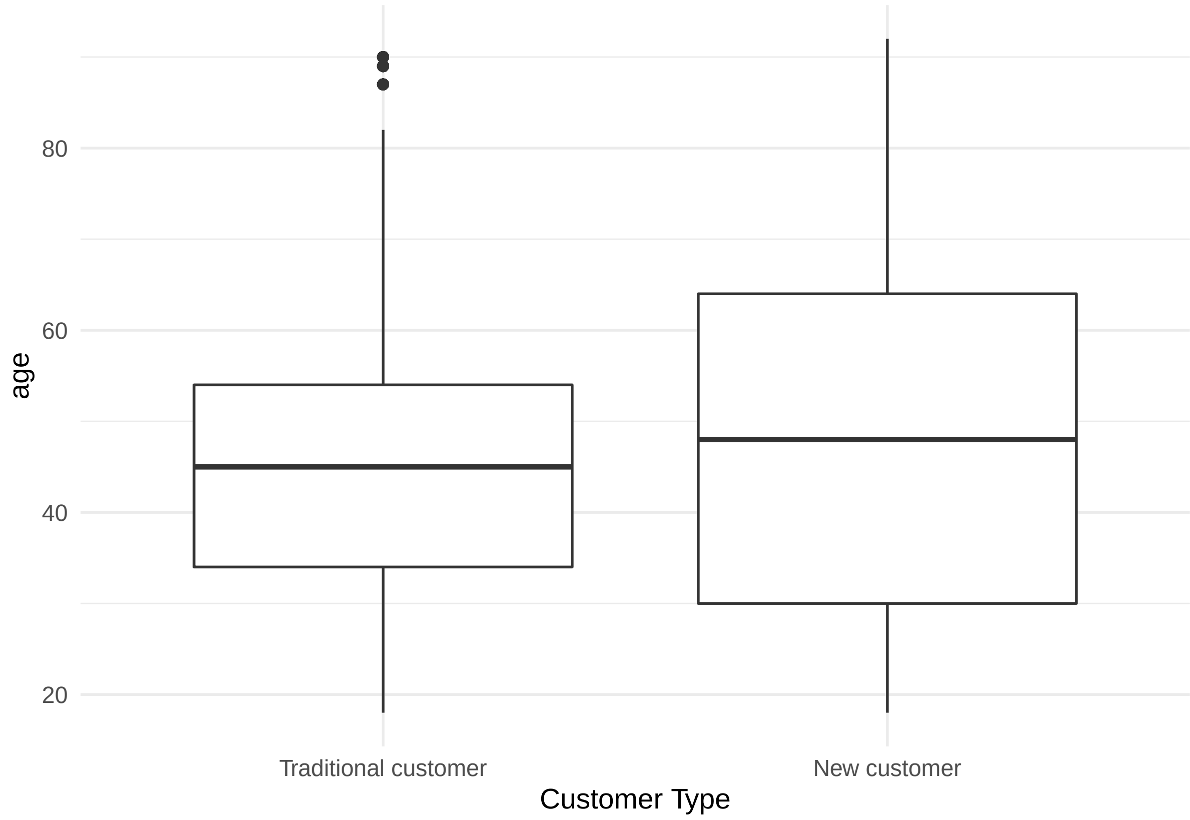

Figure 1: Boxplot for Customer's Age

By looking at the box plot, we find that new customers' age is slightly greater than traditional customers. Age conventional customer range is narrower than new customers. Based on the bar plot, females are more than males for the number of new customers, and the number of new customers is increasing by about 34% for females and 1.7% for male new customers. Also, we analyze the other four factors: age, sex, income, and skin. We find out that age and sex follow the plot above. For income, the new customer tends to spend more money on a product with less income which means the product is more affordable.

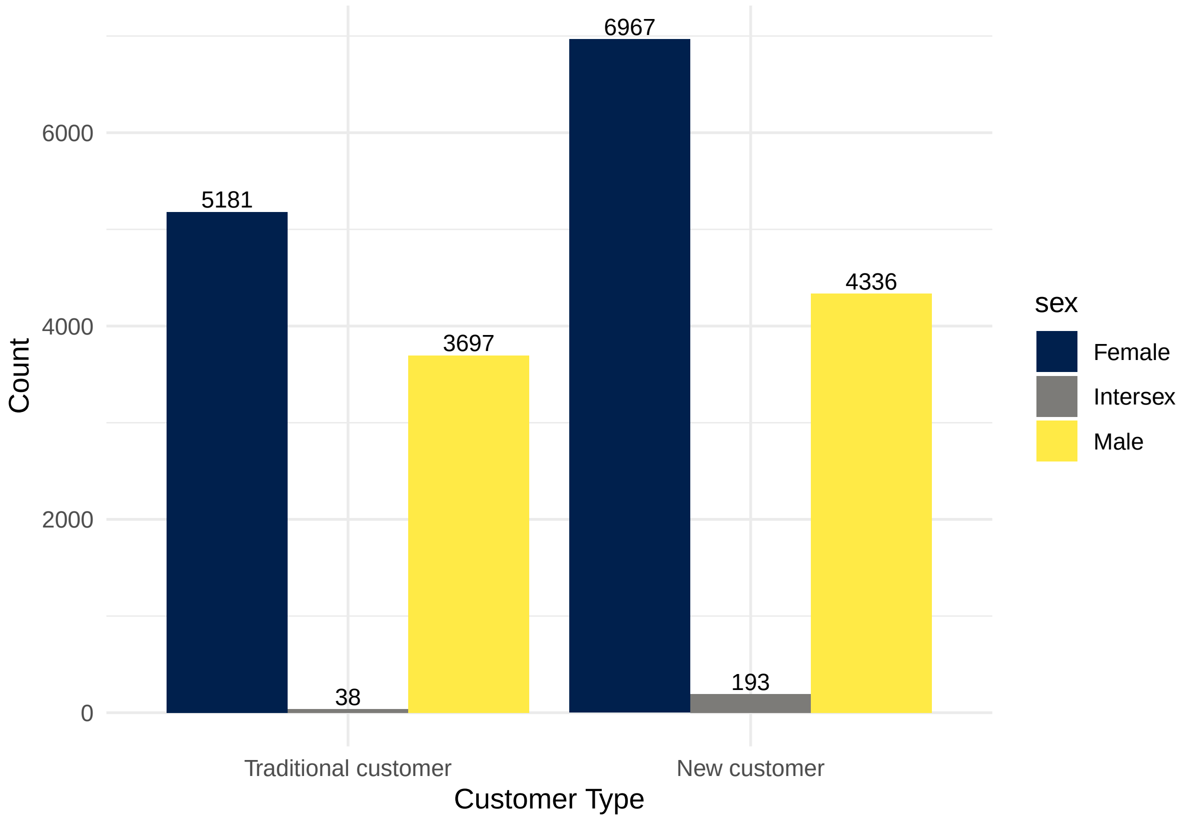

Figure 2: Barplot of Customer's Sex

Methods and Models

We use generalized additive model to predict who are the new customers and using some factors to predict the affection for the new customers behaviors. Key assumptions need to consider for estimating the gam_model are listed below: (1) Assumptions cannot be made on specific link function for error distribution. (2) Non-linearity in partial residual plots may suggest semi-parametric modeling. (3) Prior hypothesis or theory suggest non-linear or skewed relationship among variables. (4) Shape of predictor functions is determined by the data.

Our fitted model is:

$$g(E[y]) = \beta_0 + f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4)$$

where $\beta_0$is the constant, x_1 is the first factor sex, x_2 is the second factor age, x_3 is the third factor skin, x_4 is the last factor income.

Our generalized additive model shows that the interception is 1.303, which means new females

Table 3: Coefficient of estimated GAM model

	Estimate	Std. Error	z value	Pr(>|z|)
intercept	0.0941095	0.0618399	1.522	<0.005
sexMale	-0.1407536	0.0293073	-4.803	<0.005
age	0.0115961	0.0008394	13.814	<0.005
skinDefault	0.1669971	0.0497202	3.359	<0.005
skinLight	-0.3407874	0.0548365	-6.381	<0.005
skinMedium	-0.1172738	0.0548365	-2.139	<0.005
skinMedium_dark	-0.1987899	0.0551054	-3.607	<0.005
skinMedium_light	-0.2989961	0.0545989	-5.467	<0.005

are 1.3 times greater than male customers. We can see that intercept of the age factor is 0.01, which means the age of the new customer is slightly more significant than the old customer. Looking at skin factor, we find out people with skin medium is has more amount with our new customers. Then, by looking at the income factor, new customers have less than traditional customers.

To find out the changing of behavior of new customer on income, we using linear mixed effects model to predict the behavior. key assumptions for linear mixed effects model are: (1) Homogeneity of variance. (2) Normality of error term. (3) Normality of random effect. The expression of our fitted model is:

$$y = X\beta + Zb + \epsilon$$

$$b \sim N(0, \psi_\theta), \epsilon \sim N(0, \Lambda_\theta)$$

where y is β, an unknown vector of fixed effects.b is an unknown vector of random effects, with mean $E(b) = 0$ and variance-covariance matrix $var(b) =$ G.ϵ is an unknown vector of random errors. X and Y are known design matrices relating the observations.

Results

We use the generalized additive model to predict who are the new customers and use some factors to indicate the affection for the new customers' behaviors. By looking at the box plot, we find that new customers' age is slightly greater than traditional customers. The age range of traditional customers is narrower than new customers. Based on the bar plot, females are more than males for the number of new customers, and the number of new customers is increasing by about 34% for females and 1.7% for male new customers.

Looking at the gam model, we can see the intercept is 0.0941095, which means a positive correlation. The p-value is 2eˆ-16 which is significant, and the test hypothesis is false and should be rejected. R-Squared is 0.024, which indicates how much the independent variable explains the variation of a dependent variable in a regression model. The intercept shows that age is the most significant factor out of four with 0.0115961. For income, we use the LMER model to predict the affection. The interception is -1.857e-09, which means income is negatively corrected with new customers. We conclude that new customers are willing to buy products with less income, products are more affordable to them.

Product Development Studies and Suggestions

Data Wrangling

We conducted a study to investigate the performing of our fitness tracking wearable devices in respect to the ethnicity of users. Data manipulations are done by removing all NA values in flags variable, and creating a new variable flags_duration which offsets the influences by duration of sleep session.

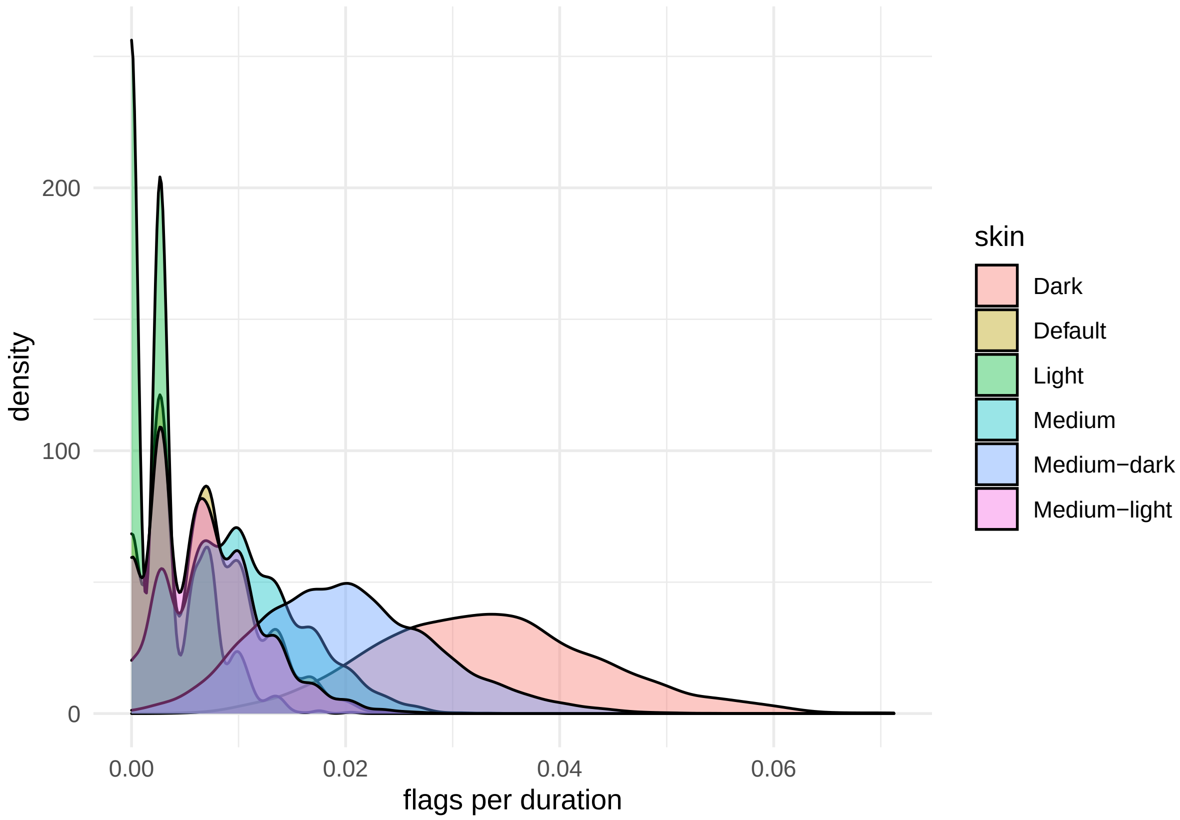

Figure 3: Density Plot of Flags per Duration Across Different Ethnicity

Table 4: Coefficient of estimated GLM

	Estimate	Std. Error	z value	Pr(>|z|)
intercept	-3.398	0.006	-602.6	<0.001
skinDefault	-1.634	0.010	-155.7	<0.001
skinLight	-2.392	0.016	-145.9	<0.001
skinMedium	-1.214	0.011	-108.6	<0.001
skinMedium_dark	-0.502	0.009	-56.4	<0.001
skinMedium_light	-1.615	0.013	-128.1	<0.001

Data Visualization

In the above kernel density estimate plot, there are significant right-skewed distributions for skin tone with "Default", "Light", "Medium", and "Medium-light", peaking at 0~0.01 flags per duration. Whereas the distribution for "Medium-dark" and "Dark" skin tone are bell-shaped, which peak in the middle of the distribution at 0.02 and 0.03 flags per duration respectively. This result illustrates that the number of quality flags might associated with customer's ethnicity.

Methods and Models

To compare longitudinal change for quality flags between the group of customers with different races, we fitted a generalized linear model in R version 4.0.5 (R Core Team, 2021) using the lme4 package (Bates, Maechker, Bolker, & Walker, 2015). There are several key assumptions required for estimating a Poisson GLM: (1) The data $Y_1, Y_2, ..., Y_n$ are independently distributed. (2) Y_i follows a Poisson distribution. (3) $log(\mu)$ must be a linear function of the explanatory variables. (4) The mean of the distribution is equal to the variance. Note that the homogeneity of variance and normality of residuals does not need to be satisfied in GLM.

The model for Y_i (number of quality flags detected for i^{th} customer) is as follows:

$$Y_i \sim Poisson(\mu_i)$$

$$log(\mu_i) = log(duration) + \beta_0 + \beta_1 skin_i$$

where $log(duration)$ is the offset term, μ_i represents the average number of quality flags for ethnicity i, $skin_i$ denotes the i^{th} customer's ethnicity for $i = 1, ..., 6$, β_0, β_i are coefficients of their corresponding independent variable. Note that the Dark skin tone is the reference level in our model.

Results

From our model results shown in table 3, there are 6 different ethnicity, we represented the effects of different ethnicity through 5 indicator variables. For example, $\beta_1 = -1.634$ indicates that the mean flags per duration in the default group is $0.195(e^{-1.634})$ times than of the dark group. $\beta_2 = -2.392$ means that the mean flags per duration in the light group is $0.091(e^{-2.392})$ times than of the dark group. The intercept term $\beta_0 = -3.398$ indicates that the predicted mean number of flags per duration for dark skin tone is $0.033(e^{-3.398})$. Since all p-values are smaller than 0.05, we conclude that our result are statistically significant.

Based on the results of generalized linear model, we are confident to conclude that the customer's ethnicity is associated with the number of flags in respect to sleep scores. The evidence from confidence interval shows rate of flags are reduced for all other skin tone compare to dark skin. Thus, the quality flags during sleep session tend to be grater in customer with darker skin tone.

Discussion

Base on our estimated model and result from research question two, we conclude that the customer's ethnicity does influence the mean numbers of quality flags during sleep session. The device perform weaker for users with darker skin particularly in sleep scores.

Strengths and limitations

For the second research question, the use of a generalized linear model is one of our study's strengths. It is a flexible model that allows a broader range of distribution, including any exponential distribution type. In addition, it can efficiently deal with categorical variables and is less susceptible to overfitting.

One of the limitations is that the residual deviance is slightly greater than residual degrees of freedom, indicating the model is slightly over-dispersed. In other words, standard errors might be falsely small, and our p-values might be significantly small. This can be adjusted by performing a quasi-likelihood, negative binomial regression, or a zero-inflated Poisson.

Conclusion and Recommendations

Through analyzing customers data, we have summarized the characteristics of MINGAR's new customers compared to traditional customers. Therefore, we suggest that the marketing team focus on advertising towards females, older people, and diverse customers. Also, the advertisement should focus on the affordable feature of the devices to compete with Bitfit in the low-end market. This strategy would potentially help MINGAR gain more market shares in the wearable market.

Moreover, this report shows that the customer's ethnicity is associated with the number of flags regarding sleep scores. It is similar to the complaints from the social media team. Customers with darker skins are more likely to have higher flags numbers. We believe it is a critical defect of MINGAR's product. MINGAR should address this issue in the future product development process.

Consultant Information

Consultant Profiles

Jixuan Huang. Jixuan Huang is a junior consultant with Beavertail Consulting Group (BCG). She is a highly analytical and process-oriented consultant with a solid understanding of data analytics. Jixuan has two years of experience analyzing data to drive successful business solutions and engaging with various senior stakeholders in the business. She specializes in data mining, manipulation, visualization, and modeling to provide insights and implement action-oriented solutions to improve business performance. Jixuan earned her Bachelor of Arts and Science, majoring in Statistics and Economics, from the University of Toronto in 2023.

Ziyuan Xu. Ziyuan Xu is a senior consultant with Beavertail Consulting Group (BCG). Ziyuan completed a bachelor's degree in Finance & Economics and Statistics at the University of Toronto. He also has a master's degree in Financial Engineering from Harbin Business School (HBS). Ziyuan had project experiences with many technology companies such as Apple, Microsoft, etc. He is familiar with business analytics and data analysis skills with R programming. He is proven to be a fast and attentive learner through project experiences in three different countries.

Haocheng Xu. Haocheng Xu is one of the consultants on the team, providing professional consulting reports to clients. Haocheng completed a bachelor's degree and a master's degree in Applied Statistics at the University of Toronto. He is good at positioning the target audience for the company's products and providing development advice to the company. In 2020, Haocheng completed a user report on MALO CLINIC, identified future potential clients, and analyzed the possibility of its IPO listing.

Maoyuan Gao. Maoyuan Gao is a consultant with Eminence Analytics. He had three years of analytic training in school, including solid data cleaning, data manipulation, and reporting skills. Besides, he has solid verbal skills, is detail-oriented, and multitasks ability. Besides that, Maoyuan also has fields experience that gained more valuable work experience such as business insight, team collaboration, and analytical thinking. Maoyuan gets his honor bachelor's degree in Arts and Science, majoring in Statistics and Mathematics, from the University of Toronto in 2022.

Code of ethical conduct

This study uses relevant and appropriate methods and data without bias or prejudice and gives valid, interpretable, and reproducible results. We do not accept the ethically unqualified work of others. We are truthful about any professional limitations and advise readers to consult other statisticians when necessary.

Our team protects and respects our research subjects' rights and interests at all stages of our participation in the project. This includes data collected in censuses or surveys, such as electronic files, personal information, and physical and mental health testing data.

We make disclosures about known or suspected limitations, deficiencies, or biases in the data that may affect the integrity or reliability of the statistical analysis. This report provides an objective and valid interpretation of the results and requires identification and confirmation of the reliability and integrity of the data.

Statistical practice requires considering various possible interpretations of the observed phenomena, and different observers using their own experiences, may obtain different interpretations and judgments. We, therefore, respect the research results of other statisticians and encourage debate and discussion towards procedures rather than persons.

References

[1] Canada Census. (2021). Population density. Census Mapper. Retrieved April 11, 2022, from https://censusmapper.ca/#12/49.2430/-123.1248

[2]Libraries, U. T. (2021, August). Postal code conversion file: 2016 census geography. Postal code conversion file: 2016 census geography | Map and Data Library. Retrieved April 11, 2022, from https://mdl.library.utoronto.ca/collections/numeric-data/census-canada/postal-code-conversion-file/2016

[3]MINGAR and Bitfit. (n.d.). Fitness tracker info hub. Retrieved April 11, 2022, from https://fitnesstrackerinfohub.netlify.app/

[4] 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/.

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

[6] 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.

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

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

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

[9] Simon Garnier, Noam Ross, Robert Rudis, Antônio P. Camargo, Marco Sciaini, and Cédric Scherer (2021). Rvision - Colorblind-Friendly Color Maps for R. R package version 0.6.2.

[10] Bob Rudis (2020). hrbrthemes: Additional Themes, Theme Components and Utilities for “ggplot2”. R package version 0.8.0. http://github.com/hrbrmstr/hrbrthemes

[11] 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

[12] Erik Clarke and Scott Sherrill-Mix (2017). ggbeeswarm: Categorical Scatter (Violin Point) Plots. R package version 0.6.0. https://github.com/eclarke/ggbeeswarm

[13] Hao Zhu (2021). kableExtra: Construct Complex Table with “kable” and Pipe Syntax. <http://haozhu233.github.io/kableExtra/, https://github.com/haozhu233/kableExtra>.

Appendix

Web scraping industry data on fitness tracker devices

We used rvest package and polite package from R to web scraping industry data on fitness tracker devices. Web scraping is method of getting data from a webpage. Notice that web scraping does not include downloading a pre-made.csv file from a website. Informative user agent details are provided to make our intentions clear, and also leave a contact if they have any issues.

Accessing Census data on median household income

We used a package called cancensus to get the income data from the Canadian census. First we accessed our API(application programming interface) key through https://censusmapper.ca/. Next, we replaced the API key and sets a folder for our cache under function options(). The data of all region in Canada at the 2016 census are accessed, and filter by census subdivision. Then, we accessed the household median income and manipulated the dataset to only needed variables. It is important to note that the data must be request at a reasonable rate when web scraping, otherwise it might be consider as DDoS attack.

Accessing postcode conversion files

We accessed the postcode conversion files as a university of Toronto student. We accepted a license agreement to access this data and downloaded the .sav file of 2016 census, since the income data from Canadian census is also from 2016. We only saved the proper information we need in this data.

95% CI of Mean Flags per Duration

```
##         intercept skinDefault skinLight skinMedium skinMedium_dark
## 2.5%    0.0330628   0.1912536 0.0885318  0.2906290       0.5949593
## 97.5 %  0.0338018   0.1992816 0.0944076  0.3036432       0.6160755
##         skinMedium_light
## 2.5%           0.1939828
## 97.5 %         0.2038101
```