Statistical Modeling and Analysis Results on Mingar's Fitness Tracking Product

With an Emphasis on Customer Demographics Targeting and Sleep Tracking Function Performance

Report prepared for MINGAR by L.S. Consulting Inc.

2022-04-07

Contents

Executive summary

The following excerpt summarizes the technical analysis performed on Mingar's fitness tracking product data.

Methods

- **Exploratory data analysis (EDA)** performed on Mingar's fitness tracking wearable's customer data, with an emphasis on exploring customer demographic.
- Assessment of the differences between new and old customers purchasing Mingar's product using **Generalized Linear Mixed Models (GLMM)**. Applying both the **Akaike information criterion (AIC)** and the **Bayesian information criterion (BIC)**, a final preferred model was determined.
- Assessment of Mingar products' inadequate sleep tracking performance and potential influencing factors using **Generalized Linear Mixed Models (GLMM)**.

Findings and conclusions

Regarding the new product's customer demographics:

- **Income level:** On average, **the median annual income level for new customers falls between ~60k($) to ~70k.** Most customers have income levels that fall near this interval in a normal-like distribution. Some outliers exist of especially wealthy individuals at >200k.
- **Age: Most new customers are evenly distributed to being within the range of 17 to ~72 years old**, with a drastic fall in interest for customers beyond 72 years old, consisting of only ~1/5 in number at each given age interval.
- **Sex and pronouns: The majority of Mingar's new customers are biological females and males**, where biological females account for ~60% of purchases and males at around ~40%. Intersex customers account for only ~1%. **Most of Mingar's new customers identify as their biological gender**, at ~97% for both male and female biological sexes.

Regarding the new product's customer demographics when compared against traditional product customers:

Interpretation of selected final generalized linear mixed model:

- On average, when median annual income is fixed, an increase in actual age from 17 to 92 increases the odds of being a new customer by 47%. **The customer's interest in new and cheaper products slightly increases with age**.
- On average, when age is fixed, an increase in median income from ~40k(41880) to ~195k(195570) decreases the odds of being a new customer by 95.7%. **Customers with a high income base is very unlikely to become a new customer**.
- Therefore, **the majority of new customers is likely having a low-to- middle income level, which differentiates them from the traditional higher-income customers**.

Regarding the new product's inadequate performance on sleep tracking and its associated potential factors:

Figure 1 below illustrates the distribution of flag (interruption) counts color-coded on skin color:

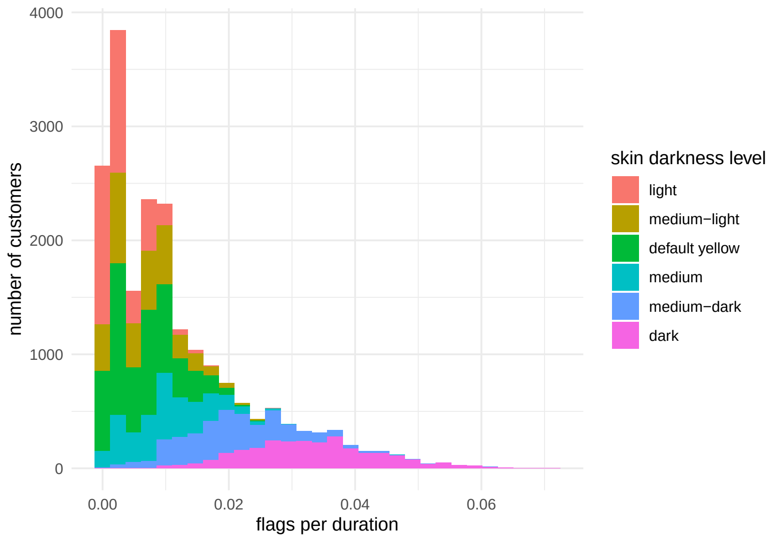

Figure 1: the number of customers with corresponding flags per duration, shown in histogram

Interpretation of selected final generalized linear mixed model:

- The average number of flags recorded for customers with light skin color is 0.003.
- The average number of flags recorded for customers with medium-light skin color is 2.178.
- We obtain the average number of flags recorded for customers with darker skin colors. The results suggest an increasing trend, from 0.003 to 10.940, in the number of flags as the customer's skin gets darker. Thus, **skin darkness is likely a major factor in the product's poor performance on sleep tracking feature**.

Technical report

Introduction

The primary objective of this report is to analyze customer and device data on Mingar company's newly developed "Advance" or "Active" fitness tracking wearable product, with a focus on the company's request to investigate the new product's customer targeting and inadequate performance in sleep tracking technology. The report consists of 3 main parts: the first two parts contain EDA, model construction and model interpretation for customer targeting analysis, whereas the third part contains model construction and model interpretation for the potential factors associated with the product's disturbed sleep tracking function. A conclusive discussion, along with any strength and limitation of the models presented are listed at the end of the report.

Research questions

- regarding the new product's customer targeting (Q1):
  - (Q1A) What are some characteristics, such as income level, age, sex and pronouns, of the new product's customers?
  - (Q1B) How are the new product's customers different from the company's traditional customers in terms of characteristics, especially income level?
- regarding the product's sleep score performance:
  - (Q2) Does skin darkness, along with other potential factors (sex, device type, date), have a significant association with the number of flags, which indicates the number of interruptions to sleep tracking?

Data introduction

The majority of data used is provided by the company, which includes:

- data file [customer]: customer information, such as sex, pronouns, date of birth, emoji modifier and postal code
- data file [cust_dev]: device id and corresponding customer id
- data file [cust_sleep]: records of sleep tracking related data, such as number of flags (interruptions) recorded, date recorded and duration of record
- data file [device]: basic device information, such as device id, device name,.etc

In addition, there is also external data from different sources used, which includes:

- data file [web scraping industry data]: contains device information, such as device name, product line, released date recommended retail price and whether the device has certain functions
- data file [Census API data]: contains median annual income level, CSDuid and population
- data file [postcode conversion data]: contains postal code and CSDuid

Data wrangling and manipulation

We first modified a few data file:

- modify [customer]: create variable age from date of birth, rescale age to be in range of 0-1 for model building purposes, drop date of birth to eliminate repetitive information
- modify [postcode conversion data]: clean repetitive observations from postcode data. We find that one postal code corresponds to multiple CSDuid and decided to keep the first CSDuid appeared to be the CSDuid associated with every unique postal code

In consideration of customer's privacy and industry regulations, 2 datasets are generated based on the data file listed above. Note that the purpose of merging data files is to combine information needed for data analysis unless mentioned specifically.

- dataset for Q1:
  - contains all necessary information for new customer targeting analysis.
  - data wrangling process
    - merge [customer] and [cust_dev] by the same customer id
      - [cust_dev] only serves a connector purpose and does not provide information for analysis
    - merge resulted dataset in last step and [device] by the same device id
    - merge resulted dataset in last step and [postcode conversion data] by the same postal code
    - merge resulted dataset in last step and [Census API data] by the same CSDuid
    - filter out and remove the observations containing no information on sex, pronouns, age and median annual income
    - create binary variable "new customer" by checking if their purchased product line is "Active" or "Advance"

        - drop irrelevant variables and variables not allowed to be published, such as customer id and CSDuid

- data set for Q2:
    - contains all necessary information for sleep tracking disturbance analysis.
    - data wrangling process
        - merge [customer] and [cust_dev] by the same customer id
            - [cust_dev] only serves a connector purpose and does not provide information for analysis
        - merge resulted dataset in last step and [device] by the same device id
        - merge resulted dataset in last step and [cust_sleep] by the same customer id
        - filter out and remove the observations containing no information on sex, pronouns, age, duration of records, date recorded and number of flags recorded
        - create variable skin darkness level: a skin color indicator that can be obtained by checking the Fitzpatrick scale with the corresponding unicode in emoji modifier
        - make skin darkness level a factor and change the order of its level to be from the lightest skin color to the darkest
        - change variable date from specific date records to be a binary indicator of weekdays or weekends
        - drop irrelevant variables, repetitive variables and variables not allowed to be published, such as emoji_modifier, customer id and CSDuid

Both datasets are exported as RDS file for data analysis.

Analysis of customer features of Mingar's new wearable product through EDA

In response to the question, we made a subset of the data with only new customers (customer who purchased the "Active" or "Advanced" product) and performed exploratory data analysis on new customers' sex (biological and self-identified), age, median annual income level and device information.

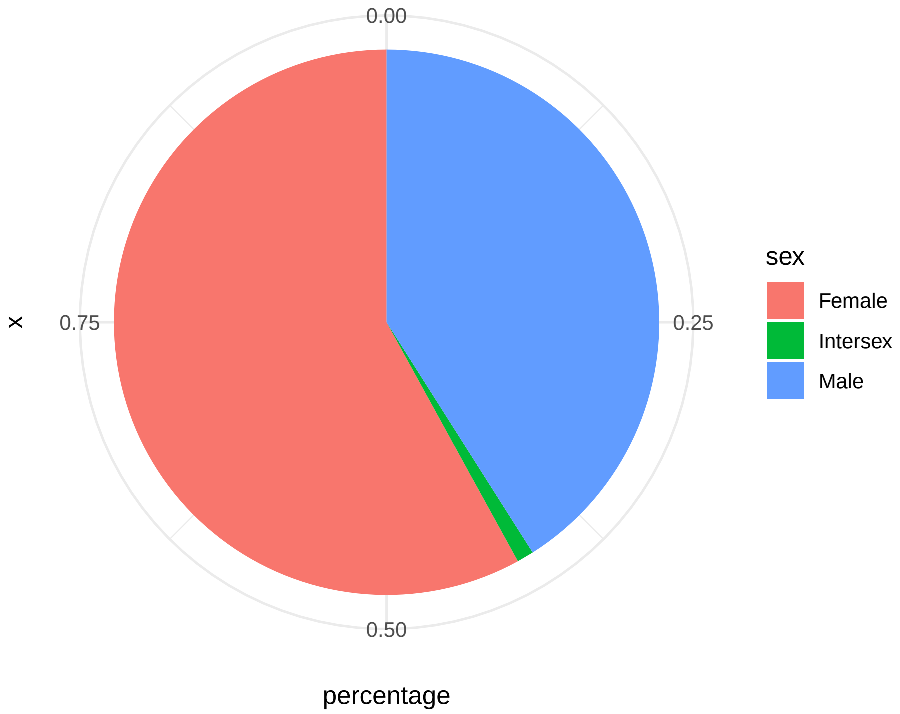

Figure 2: new customers' biological sex distribution, shown in pie chart

Table 1: biological sex ratio of the new customers

sex	number of new customers	percentage among all new customers
Female	6110	0.58
Intersex	133	0.01
Male	4326	0.41

Both **Figure 2** and **Table 1** suggest that the majority of new customers are biological females and males, where females account for ~60% of the purchases and male accounts for ~40%. Intersex customers accounts for only ~1%.

Table 2: preferred pronouns ratio of the new customers who are biologically female

pronouns	number of new customers (bio-female)	percentage among all new customers (bio-female)
he/him	58	0.01
she/her	5924	0.97
they/them	128	0.02

Table 3: preferred pronouns ratio of the new customers who are biologically male

pronouns	number of new customers (bio-male)	percentage among all new customers (bio-male)
he/him	4194	0.97
she/her	41	0.01
they/them	91	0.02

Both **Table 2** and **Table 3** suggest that 97% of the new customers identify themselves as their biological gender. From this respect, considering both biological sex and pronouns (self-identified sex) is, in some cases, unnecessary as they contain a substantial amount of overlapped information.

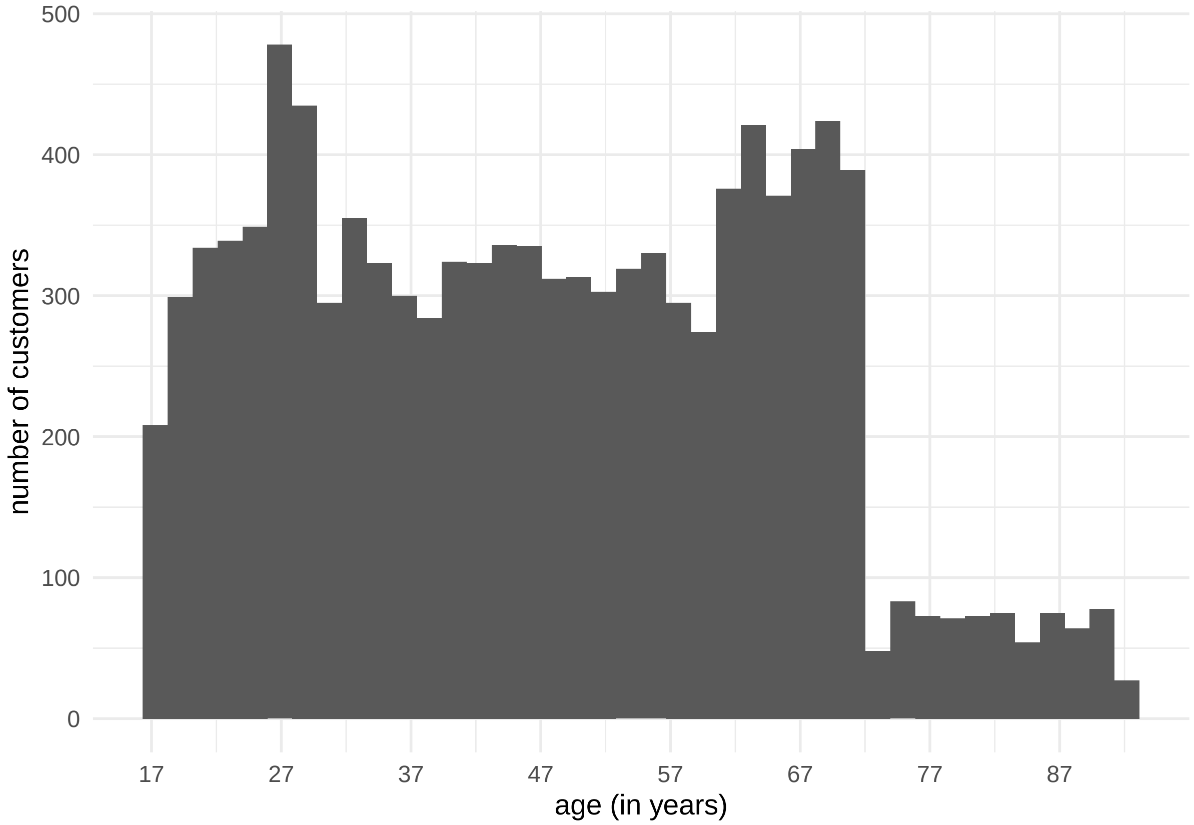

Figure 3: new customers' age distribution, shown in histogram

Figure 3 indicates that the age of new customers are evenly distributed from 17 to ~72 years old, after which the number of customers decreases to only ~1/5 at each age.

Table 4: common statistics measure of the new customers' median annual income level

	mean	median	standard deviation
median annual income ($)	0.1752485	0.1558267	0.094051

Table 4 indicates that, on average, new customer's median annual income level is ~60k to ~70k. Note that the standard deviation is largely affected by a few outliers (~190k), the majority of data falls around the mean in a Normal-like distribution.

Table 5: device popularity among the new customers

device name	number sold	sales in percentage	released date
Advance 2	6185	0.59	2021-07-08
Advance	2043	0.19	2020-08-20
Active Alpha	1763	0.17	2020-12-30
Active	279	0.03	2019-10-13
Active HR	299	0.03	2019-10-13

Table 6: newly developed device information

Device name	Recommended retail price	Battery life	Heart rate sensor	GPS
Advance 2	145.00	Up to 7 days	Yes	Yes
Active Alpha	99.99	Up to 7 days	Yes	No
Advance	120.00	Up to 7 days	Yes	Yes
Active	39.99	Up to 14 days	No	No
Active HR	79.99	Up to 7 days	Yes	No

Table 7: newly developed device information (continued)

Device name	Sleep tracking	Smart notifications	Contactless payments
Advance 2	Yes	Yes	Yes
Active Alpha	Yes	Yes	Yes
Advance	Yes	Yes	Yes
Active	No	No	No
Active HR	No	No	No

Both **Table 6** and **Table 7** suggest similar features of two products named "Advance 2" and "Advance", which account for ~80% sales to new customer in **Table 5**. The newest-released and most expensive "Advance 2" accounts for ~60% of all sales to new customers, which is ~3 time of the sales from the second popular product "Advance".

Accessment of Mingar's customer differences between its traditional and new product through GLMM

Model Building and Model selection

In consideration of the model's validity and significance of the potential factors causing the difference, we constructed 5 statistical models:

- model 1:
    - response variable: new customer indicator (yes or no)
    - fixed effect: customer's age, sex, pronouns and median annual income level
    - random intercept: interaction term of customer's age and median annual income level
        * based on common knowledge: income usually peaks at middle age and is much lower for young and old people.
    - model type: GLMM with Binomial Regression (based on binary response variable and random effect)
- model 2: model 1 with no fixed effect on sex
- model 3: model 1 with no fixed effect on pronouns
- model 4: model 1 with no fixed effect on sex, pronouns
- model 5 (GLM): model 1 with no fixed effect on sex, pronouns and no random effect on age-income interaction

Model Selection Process: We applied AIC and BIC on the models indicated above, the result indicates that model 4 has the smallest AIC, BIC and the largest log-likelihood. Therefore, we prefer model 4 for further analysis.

We arrived at out final model:

- model 4 (final model–GLMM with Binomial Regression):
    - response variable: new customer indicator(yes or no)
    - fixed effect: customer's age and median annual income level
    - random intercept: interaction term of customer's age and median annual income level

Model Estimates and Confidence Intervals

Table 8: MLEs for odds and odds ratios of being a new customer, with 95 percent confidence intervals

	estimates	2.5%	97.5%
new customer indicator(yes or no)	1.945	1.788	2.117
age (in years)	1.470	1.290	1.675
median annual income level	0.043	0.031	0.059

Table Interpretation:

- the estimated intercept (row 1):
    - for 17 years old customer with median annual income level ~$40k($41880), the probability implied by the odds of being a new customer (purchasing an "Active" or "Advance" product) is 1.945/(1+1.945) = 0.66.
        - ∗ not very informative as very few people earns ~$40k annually at age 17.
    - CI: we are 95% confident that, for 17 years old customer with median annual income level ~$40k, the probability of being a new customer is between 1.788/(1+1.788) = 0.64 and 2.117/(1+2.117) = 0.68.
- the estimated adjusted age (row 2):
    - on average, when median annual income is fixed, an increase in actual age from 17 to 92 increases the odds of being a new customer by 1.470 - 1 = 47%.
        - ∗ suggesting that customer's interest in new and cheaper products slightly increases with age.
    - CI: when median annual income is fixed, we are 95% confident that, on average, an increase in age from 17 to 92 increases the odds of being a new customer between 29% and 67.5%.
- the estimated adjusted median annual income level (row 3):
    - on average, when age is fixed, an increase in median income from ~40k(41880) to ~195k(195570) decreases the odds of being a new customer by 1 - 0.043 = 95.7%.
        - ∗ suggesting that customers with a high income base is very unlikely to become a new customer. Therefore, the majority of new customers is likely having a low-to-middle income level, which differentiates them from the traditional higher-income customers.

  - CI: when age is fixed, we are 95% confident that, on average, an an increase in median income from ~40k(41880) to ~195k(195570) decreases the odds of being a new customer between 94.1% and 96.9%.

Analysis of the new product's inadequate performance on sleep tracking and its associated potential factors through GLMM

Model Building

In consideration of the model's validity and significance of the potential factors causing the flags, we constructed 4 statistical models:

- model 1 (final model–GLMM with Poisson Regression):
  - response variable: number of flags recorded (times of sleep tracking interruption)
  - fixed effect: customer's skin darkness level (indicated by emoji modifiers), sex, device name
  - random intercept: date indicator(weekday or weekend)
    - difference in individual's behavioral characteristics in weekdays and weekends may have an effect on the device's sleep tracking performance.
  - offset term: duration
    - longer sleeping duration is likely associated with more occurrence of flags and vice versa.
  - model type: GLMM with Poisson Regression (based on discrete response variable and random effect)
- model 2: model 4 with no fixed effect on sex
- model 3: model 4 with no fixed effect on sex, device name
- model 4 (GLM): model 4 with no fixed effect on sex, device name and no random effect on date indicator

Model Selection Process: We applied AIC and BIC on the models indicated above, the result indicates that, although model 4 has smaller AIC and BIC, the difference in AIC, BIC and log-likelihood between model 3 & 4 is not substantial. Therefore, considering the importance of the random intercept, we prefer model 3, the complexier model, for further analysis.

We arrived at out final model:

- model 3 (final model–GLMM with Poisson Regression):
  - response variable: number of flags recorded (times of sleep tracking interruption)
  - fixed effect: customer's skin darkness level (indicated by emoji modifiers)
  - random intercept: date indicator(weekday or weekend)

    - difference in individual's behavioral characteristics in weekdays and weekends may have an effect on the device's sleep tracking performance.
  - offset term: duration
    - longer sleeping duration is likely associated with more occurrence of flags and vice versa.

Model Estimates and Confidence Intervals

Table 9: MLEs for odds and odds ratios of sleep tracking interruptions at different skin darkness level (lightest to darkest), with 95 percent confidence intervals.

	estimates	2.5%	97.5%
light	0.003	0.003	0.003
medium-light	2.175	2.095	2.258
default yellow	2.135	2.062	2.211
medium	3.249	3.136	3.367
medium-dark	6.621	6.407	6.845
dark	10.937	10.592	11.295

Table Interpretation:

- the estimated intercept (row 1):
  - the average number of flags recorded for customers with light skin color is 0.003.
  - CI: both the upper CI and lower CI are rounded to 0.003, indicating that we are 95% confident the average number of flags recorded for customers with light skin color falls in a range, within which any value ≈ 0.003.
- the estimated flags of customers with medium-light skin (row 2):
  - the average number of flags recorded for customers with medium-light skin color is 0.003 + 2.175 = 2.178.
  - CI: we are 95% confident that the average number of flags recorded for customers with medium-light skin color falls between 0.003 + 2.095 = 2.098 and 0.003 + 2.258 = 2.261.

Following a similar approach, we can obtain the average number of flags recorded for customers with darker skin colors (row 3-6). These numbers suggest an increasing trend, from 0.003 to 10.940, in the number of flags as the customer's skin gets darker. Thus, skin darkness is likely a major factor in the product's poor performance on sleep tracking feature.

Discussion

In response to Mingar's request for their new wearable product's customer targeting, our EDA and model suggest an even distribution of biological sex among its new customers, with female accounting for ~10% more than male's share in product sales. 97% of the new customers' self-identified sex aligns with their biological sex. The age of new customer also distributes evenly. Comparing to traditional customers, new customer is likely to be older as odds of becoming a new customer increases with aging. On average, new customers' median annual income is $60k to $70k. This is inferably lower than that of the traditional customers due to the inverse relationship between becoming a new customer and individual's median annual income. Around 60% of new customers purchased the new product named "Advance 2", without further information, the potential factor associated to its popularity is unclear.

As for the new product's sleep tracking problem, our model implies that, in consideration of the behavioral difference in sleeping habit between weekdays and weekends, more sleep tracking interruptions are associated with higher level of skin darkness. It is reasonable to infer that skin darkness is a major cause of this problem.

Strengths and limitations

Strengths:

- Generalized Linear Mixed Models (GLMMs), the main tool used in this analysis to construct models for client use, is a powerful class of models that combines the characteristics of both GLMs and LMMs. As demonstrated in this analysis, the model can be used across multiple common distributions, furthering its versatility.
- Using visual representations on sleep quality analysis with an emphasis on skin color clearly illustrates and highlights problematic properties of the new product as per client instruction.

Limitations:

- use of median annual income level as indicator of customer's personal income
  - this is applied in both the EDA and model construction. The median annual income level is postal code-based, thus may not reflect the customer's actual income. However, with a large sample size of ~20k, it is reasonable to infer that the median annual income level is representative of the average of customer's personal income.
- use of emoji identifier as indicator of customer's skin color

  - this is applied in analysis of the sleep tracking problem. Emoji identifier is entirely dependent on the customer's preference, thus may not reflect the customer's real skin color. Although customers are likely to select an emoji identifier that reflects their skin color, it is advised to make use of the result of this report with caution and collect information on a wider range of potential causes to the problem.

- assumption of non-specified emoji identifier to be "default yellow", a skin color with darkness level between medium-light and medium
  - this is applied in analysis of the sleep tracking problem. Customer's who did not specify an emoji identifier may not have a "default yellow" skin color. However, the model suggests an estimate for "default yellow" close to the estimate of medium skin color, thus reduces its negative effect on the model's validity.

Consultant information

Consultant profiles

Yiyu (Star) Li. Yiyu is a senior consultant with L.S. Consulting Inc.. As a economics analysis specialist, she specializes in working with real world data and adapts her methods based on client needs. She maintains an organized work-flow and prioritizes valid and prudent decision-making with appropriate methodology. Yiyu earned her Bachelor of Science, Major in Statistics and Economics, from the University of Toronto in 2023.

Jiawei (Jae) Shi. Jiawei is a senior consultant with L.S. Consulting Inc.. Jiawei specializes in data visualization, she seeks to understand and mitigate known or suspected limitations of her reports, and focuses on representing data in an intuitive way for general audiences. Jiawei earned her Bachelor of Science, Majoring in Economics and Minoring in Statistics and Mathematics from the University of Toronto in 2023.

Code of ethical conduct

Ethical Conduct Statement:

As ethical statistics practitioners, *L.S. Consulting Inc.*:
1. Is strictly dedicated to procedural integrity and objectivity, use appropriate, valid and relevant methods to produce reproducible and interpretable results.

2. Acknowledges and respects intellectual properties of other individuals and groups, recognizes and strives for the dignity and inclusion of all people.

3. Follows and upholds all applicable guidelines, policies and laws relating to statistical consulting, and takes appropriate actions against deviation from these guidelines without ethical justification.

4. Prioritizes client privacy and assures transparency regarding all client information and data usage. Exercises extreme caution working with proprietary and/or confidential data.

References

> Anon. (2022). Fitness tracker info hub. Fitness Tracker Info Hub. https://fitnesstrackerinfohub.netlify.app/

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

> Census Canada. (2016). Population Density. Censusmapper.ca; Census Canada. https://censusmapper.ca/

> 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.

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

> the Committee on Professional Ethics of the American Statistical Association. (2022, January). Ethical Guidelines for Statistical Practice. American Statistical Association. https://www.amstat.org/your-career/ethical-guidelines-for-statistical-practice#:~:text=The%20ethical%20statistical%20practitioner%20seeks

> Hao Zhu (2021). kableExtra: Construct Complex Table with "kable" and Pipe Syntax. http://haozhu233.github.io/kableExtra/, https://github.com/haozhu233/kableExtra.

> 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/.

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

> Jared E. Knowles (2020). eeptools: Convenience Functions for Education Data. R package version 1.2.4. https://github.com/jknowles/eeptools

> Unicode Inc. (2022). Full Emoji Modifier Sequences, v14.0. Unicode.org. https://unicode.org/emoji/charts/full-emoji-modifiers.html

> UofT Map and Data Library. (2016). Postal code conversion file | Map and Data Library. Mdl.library.utoronto.ca.
> https://mdl.library.utoronto.ca/collections/numeric-data/census-canada/postal-code-conversion-file

> 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.

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

Appendix

Web scraping industry data on fitness tracker devices

We first read the Terms and Conditions of the industry data website to look for an API and make sure this website is scrapable.

Since the website does not provide an API, we web scraped this data by following the common web scraping principles. Note that the R package "polite" and "rvest" are used in the process.

Data accessing process:

- Introduce our company to the host of the industry data website url using functions in the polite package
  - a User Agent string is provided to clarify our intentions and provides contact information of our company to the host
- check the returned web scraping allowance information from the host
  - make sure the path is scrapable and request data at a reasonable rate
- download data using functions in the rvest package

Accessing Census data on median household income

We first read the Terms and Conditions of the Canadian census website to look for an API and make sure this website is scrapable.

Since the Canadian census website does provide an API, we proceed to obtain data using the API. Note that the R package "cancensus" are used in the process.

Data accessing process:

- sign up at the Canadian census website to obtain a personal API key
- using R function from the cancensus package to set a folder to cache the data
  - cache avoid frequent data request to the website and reduce the chance of web crush
- using R function from the cancensus package to get all regions as at the 2016 Census (most up-to-date Census)
- select appropriate region code (CSD) to aquire data
- using R function from the cancensus package to download data

Accessing postcode conversion files

The access of this file is limited to private users at the UofT Map and Data Library website. As legal users of the website, we downloaded the data directly after agreeing to the liscense agreement of data usage on the website.

To match the proper Geographic code (CSDuid) in the Canadian census data, we selected the postcode conversion data in year 2016 to download.

However, loading of the obtained data file is causing crushes in R. This is why we choose to use the provided emergency file instead of the data file obtained from the website.