New customers characteristics and issue for particular customers

By observing the data, the consumers who are aged, intersex or male, with less income, from a rural area and are not willing to have sleep tracking serves will buy our new products and the customers with darker skin color are acutally facing the poor ferforming while using our devices.

Report prepared for MINGAR by Uoft superteam

2022-04-11

Contents

Executive summary

Company introduction

Originally, MINGAR is the provider company for marine vehicles and military personnel by using our flagship product, the GPS units. In the early 2000s, our company started to introduce personal GPS units for runners and branched out to sports-related products. After the developments in these years, we promote the products for outdoor recreation and wearable devices that investigate fitness.

Main findings for the characteristics of our new customers

Since the new products "Active" and "Advance" have been added to our offering options, in order to better serve and explore our future potential customers, we are willing to know what factors and customers' characteristics affect the willingness of consuming the "Active" and "Advance" product lines. In this case, we focus on the characteristics that may have relations with "Active" and "Advance", refers to the graph of age for all of the consumers, we can see a clear different distribution for the traditional and new customers, thus, the age may influence their preference of choosing devices. For the purpose of evaluating more influences that may affect our prediction, we build up the most applied models to find our future customers' characteristics. The results show a rough picture of our future new product consumers, older males or intersex people with less income and come from a small place are more willing to buy our new products and sleep tracking service seems not necessary for them.

Whether having issue of poor performing for customers with darker skin color

In order to investigate whether the devices are performing poorly for consumers with darker skin color, we firstly set the graph above to show that customers with dark or medium-dark skin color have more number of flags occurring over the duration of sleep, which means consumers with darker skin color have a higher number of flags occurrence over the duration of sleep and they having more chance to get a quality flag on their advice during the sleep session. In this way, the probability of missing data and unusual errors showing on the devices will increase for them to cause the poor performance. To verify this observation, four models with different fixed and random effects are created to determine the final fitted model. The result of the models all indicates that people with other skin colors are less likely to get the quality flags on their devices and people with darker skin color are facing the issue of poor performance while sleeping.

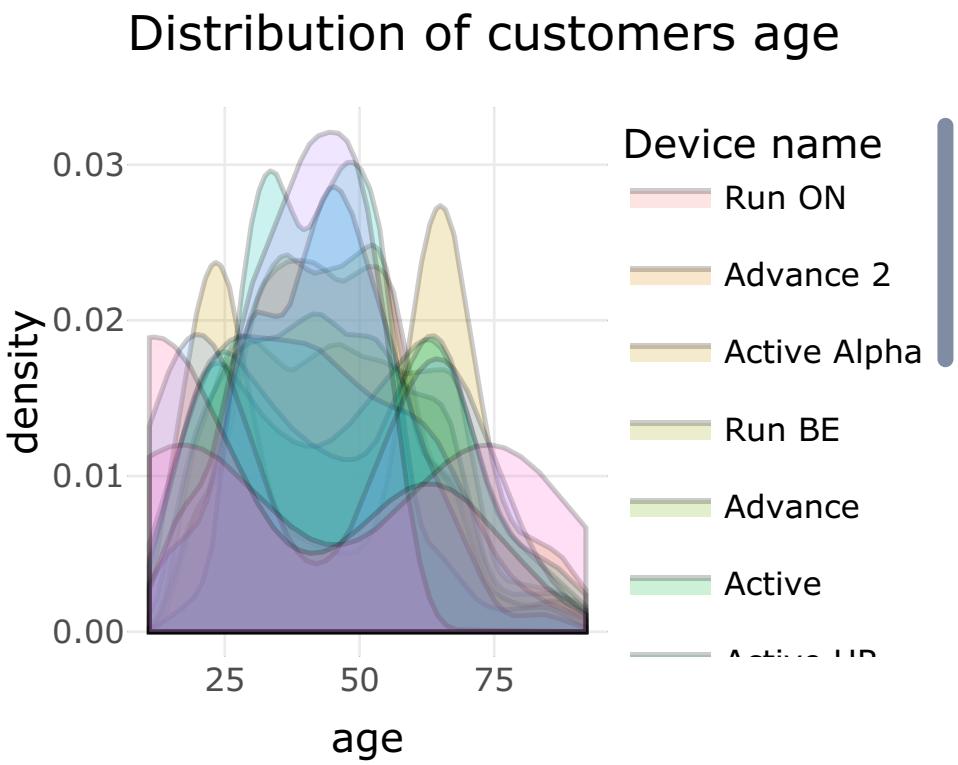

Figure 1: Distribution of customers age

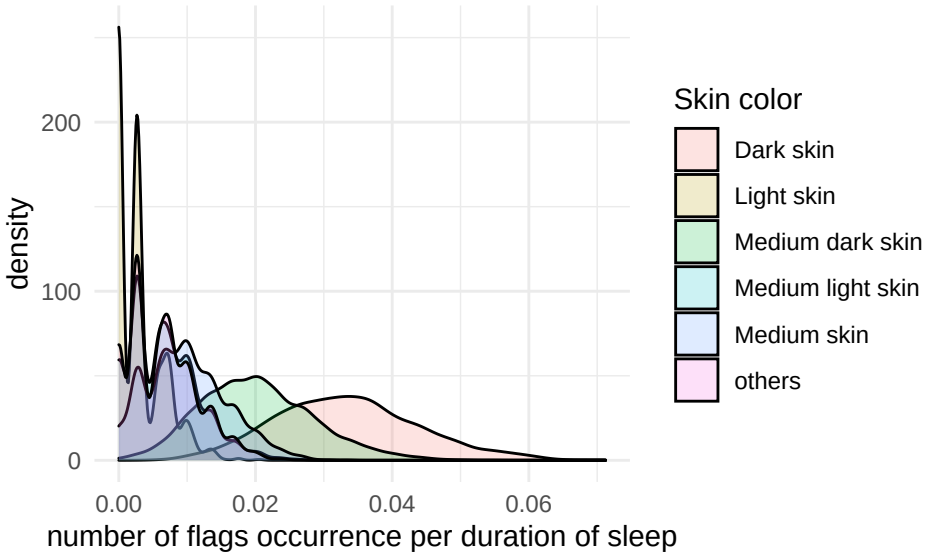

Figure 2: The number of flags occurrence per duration of sleep for differrent skin color types

Technical report

Introduction

The rapid rise in demand for multimedia devices and smartphones, along with an increase in the use of fitness trackers and health-related wearable, is expected to propel the wearable technology industry. Hence, many companies start to produce wearable devices and become our competitors in the fitness tracker space. To compete with other companies, we introduce our new lines "Active" and "Advance" with more approachable prices, therefore, the characteristic of our new consumers and the difference between the new customers and the traditional customers becomes important to explore and analyze. In addition, from the social media team, a complaint that our devices' performance is poor for consumers with darker skin color is promoted. Aiming to investigate whether or not our devices are having this problem becomes necessary for us.

In general, this report is composed of our research questions, data manipulation, method, graphs of the important variables, assumptions to our research questions, model building and results. There are two main goals, one is looking for the potential future customers in Canada purchasing our new products "Active" and "Advance" and another is to find out the influences that affect the poor performance of the sleep tracking function of our products. Therefore, we collected the data from the websites of the Fitness tracker info hub, the Census Mapper API and U of T libraries to analyze the factors that may be related to our questions.

After plotting the graphs, we found that for finding the potential new customers in Canada for our new products buyers, the gender, income level and age of the customers may have an effect on purchasing our new products "Active" and "Advance", thus, we expect these three characteristics contributed to new products. Moreover, the graphs of sleep tracking and the prediction of the skin color indicate that the darker skin color of the customers may reflect more on the bad performance of the sleep tracking function of our products. Therefore, we want to further explore how skin color is related to this pheromone. In this case, we will use a generalized linear model and a generalized linear mix model to help us solve the problem. During this process, the purchasing of our new products can apply logistic regression and the record of the bad performance on the sleep tracking function can use Poisson distribution.

Research questions

- In order to inform the marketing team's strategy and look for the potential consumers of our new products in Canada, we want to know the characteristics of our new customers and the difference in the personalities between new customers that purchase on 'Active' and 'Advance' products and our traditional customers. And make predictions of our potential

future new product consumers by selected factors.

- Based on the complaints that our devices perform poorly on sleep scores for customers with darker skin color, we want to know if this truly happens and whether other factors affect the result of sleep scores.

Data and important variables information

Our data is made up by using web scraping from other websites. The devices information is provided by web scraped device data from the website of the Fitness tracker info hub; the median income data is procured through the Census Mapper API; the postcode conversion file is provided by U of T Libraries. By using the left join, the data sets from each website will be joined together as our final data set. In order to explore more about the data, the variable of age is created by using the time length function on the original variable date of birth and is rounded to integers. We also created the variable device name to identify the devices name for each customer. As we focus on who is our new customer that uses "Active" and "Advance", the variable of whether or not the line of the product belongs to these two lines is also important for us to consider. For the second part of the report, we want to figure out whether the devices perform poorly for people with darker skin color. However, to avoid the concern about racism, we are not able to collect the skin color directly. Instead, the code for skin tone modifier for emojis has the ability to distinguish the skin colors of each of the customers, so we created another variable skin color by grouping the emoji modifiers.

Variables explanation

Table: Summary Table for all of the variables and their description Variable description:

Variables	Description
Active Advance recode	Whether customers buy Active and Advance products
age	Age in 2022
hhld median inc	Normalized median income
Median Income	Normalized median income
Popu	Normalized population
Sleep tracking	Whether the device contains sleep tracking function
cust id	Unique ID for each customer
date	Time when sleep session started, denoted as year, month, day

dev id	Unique ID for each device
duration	Duration (in minutes) of a sleep session
flags	Number of quality flags during the sleep session
line	Line of products this device belongs to
sex	Biological gender
skin color	An estimate of skin color based on skin tone for emojis used

Determine the characteristics of our new customers and the differences compared to traditional customers

Methods: EDA

To compete with other companies, we should aware of the characteristics of our new customers who consume the products from the "Active" and "Advance" product lines. The purpose of manifesting the graphs and tables for related variables is to visually present the features of our new customers buying "Active" and "Advance" devices. Refers to the graphs of age and income levels for customers consuming our new products, we can see a significantly different distribution for the traditional and new consumers, so we assume age and income level are two elements that we should discuss. Also, with the consideration of the graphs showing how sex affects the sales of products, gender will be an influential element for our model. Nevertheless, there may exist other factors that are related to our new product, so we add sleep tracking function, population distribution and prediction of skin color as more predictors to find out our new products' potential users.

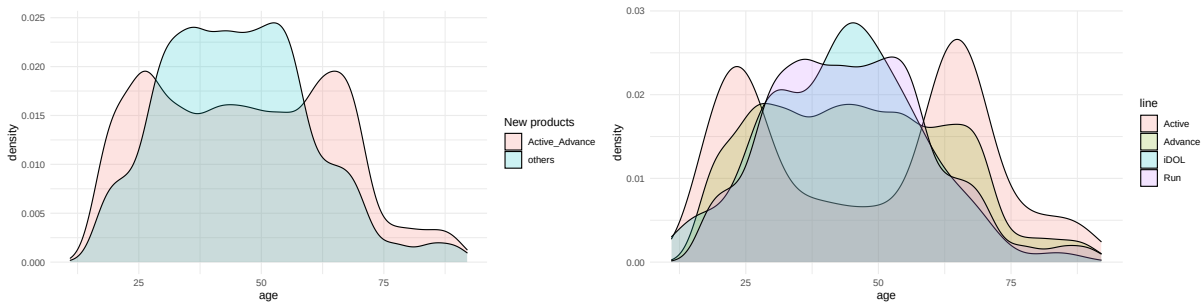

Figure 3: L:Distribution of customers age with new products R:Distribution of customers age with different lines

As observed, the left graph in figure 3 indicates that more affordable buyers who purchase "Active" and "Advance" products are the middle-aged population between 32 and 42 years old.

The bi-modal distribution shows that traditional customers are mainly young adults and aged population over 60 years old. However, the middle-aged customers prefer our new products. When we look into more details, we find out that, for "Advance" products, the density curve can be roughly regarded as uniformly distributed but kids and the elders over 75 years old are not that willing to buy our product. Age has little effect on buying an "Advance" line of products. The bi-modal distribution for products from "Active" line shows that young adults and people around 60 are attracted by devices from the "Active" line.

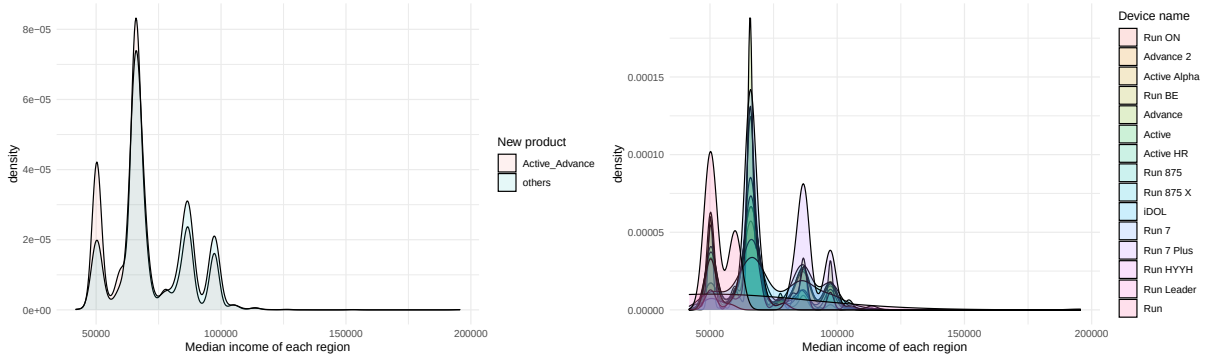

Figure 4: L: Customer's income level with new products, R: Customer's income level with devices

From the two figures for the median income level of regions in Canada, the median income of buyers of "Advance" and "Active" is generally lower than that for traditional customers. People with lower median income are more attracted by our "Active" or "Advance" products. The four significant modals in our graph represent the different income levels for people in different regions of Canada. For the first two modals with lower income levels, we can easily see that our new products sell better than other devices. Conversely, for the left two modals with higher income levels, our products sell worse than other products. Consumers with about 70000 income

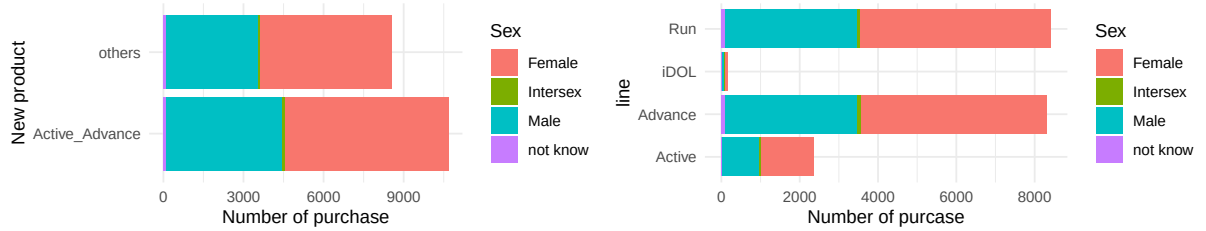

Figure 5: L: Customer sex ratio of Advance and Active product and other products, R: Customer sex ratio of different lines of products

As observed, in the first bar plot, the number of buyers for our new products is over 1000 more

than the number of consumers for our traditional devices. Also, we can easily find that around 40% of our new consumers are males and 50% of them are females which illustrates that the number of female buyers is greater than any other gender group. In this situation, our company could consider more services for females to develop their experience of using our devices. The other figure also indicates females are more likely to buy our products not only for the line of "Active" and "Advance" but also for iDOL and Run. Significantly, the number of purchases for the product line "Run" is the highest one and "Advance" is following with just a little bit difference. However, compared with the sales of them, the product lines "iDOL" and "Active" are facing a problem of minimal purchasing, so we could explore more services that are similar to "Run" and "Advance" products on the other two lines to solve improve our general sales.

Methods: Applying the models to our data

From the EDA part, by plotting the factors that may affect the predictions of our potential customers, we selected the following variables, income level, age, gender, population, skin color and if our products have the function for sleep tracking. The purpose of applying the models is to examine and evaluate the actual effects for each variables, we choose to make a generalized linear model to explore the characteristic of our new consumers. Moreover, our responding variable is whether the consumers buy our new products which follows Logistic regression. To satisfy the assumption for the generalized linear model, our data need to be independently distributed and we assume a linear relation between the transformed expected response in terms of the link function and the explanatory variables.

Firstly, we build up the generalized linear model between the probability of buying our new products and the customers' income level, age, gender, population for the region and skin color as fixed effects. The following is the corresponding formula:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 Median\ \ Income + \beta_2 Rescaled\ \ age + \beta_3 Sex_{Intersex} + \beta_4 Sex_{Male} + \beta_5 population + \beta_6 Skin\ color_{Light\ skin} + \beta_7 Skin\ color_{Medium\ dark\ skin} + \beta_8 Skin\ color_{Medium\ light\ skin} + \beta_9 Skin\ color_{Medium\ skin} + \beta_{10} Skin\ color_{Others}$$

Secondly, we build up the generalized linear model between the probability of buying our new products and the customers' income level, age, gender and population for the region as fixed effects. The following is the corresponding formula:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 Median\ \ Income + \beta_2 Rescaled\ \ age + \beta_3 Sex_{Intersex} + \beta_4 Sex_{Male} + \beta_5 population$$

Thirdly, we build up the generalized linear model between the probability of buying our new products and the customers' income level, the ages, the gender, besides as fixed effects. The following is the corresponding formula:

$log(\frac{p}{1-p}) = \beta_0 + \beta_1 Median\ Income + \beta_2 Rescaled\ age + \beta_3 Sex_{Intersex} + \beta_4 Sex_{Male}$

In the end, we build up the generalized linear model between the probability of buying our new products and the customers' income level, the ages, the gender, the distribution of population and if our products have the function for sleep tracking as the fixed effects. The following is the corresponding formula:

$log(\frac{p}{1-p}) = \beta_0 + \beta_1 Median\ Income + \beta_2 Rescaled\ age + \beta_3 Sex_{Intersex} + \beta_4 Sex_{Male} + \beta_5 Population + \beta_6 Sleep\ tracking_{Yes}$

By applying the likelihood ratio test between each two of them we could finally conclude our final model to figure out the characteristics of our new customers.

Result

From the first model, model 1, the p-values for customers' median income level, age and population are 0, which indicates that we have very strong evidence to say the probability of buying our new products is related to these elements, however, the p-values for sex and skin color are greater than 0.05, which means sex and skin color are not significant for our model. For the next model, we will consider to reduce one of sex and skin color to fit a better model. Therefore, after adding the coefficients for each predictor, the formula for model 1 will be the following:

$log(\frac{\hat{p}}{1-\hat{p}}) = 0.0613 - 0.3183 Median\ Income + 0.4069 Rescaled\ age + 0.1758 Sex_{Intersex} + 0.0364 Sex_{Male} - 0.0653 Population - 0.0691 Skin\ color_{Light\ skin} - 0.0443 Skin\ color_{Medium\ dark\ skin} - 0.017 Skin\ color_{Medium\ lightskin} - 0.0068 Skin\ color_{Medium\ skin} - 0.0268 Skin\ color_{Others}$

By reducing the fixed effect of skin color, we get our second model with income level, age, gender and population for the region. For model 2, the p-values for customers' median income level, age and population are 0, which means we have very strong evidence to show the probability of buying our new products is related to these characteristics. Therefore, after adding the coefficients of each predictor, model 2 can be the following:

$log(\frac{\hat{p}}{1-\hat{p}}) = 0.0326 - 0.322 Median\ Income + 0.4074 Rescaled\ age + 0.1715 Sex_{Intersex} + 0.0363 Sex_{Male} - 0.0654 Population$

In order to determine the better fitted model for our data, we do the likelihood ratio test to solve it and the corresponding result table is following:

From the table of testing the likelihood ratio of model1 and model2, we can find the p_value is 0.801111 larger than 0.05, which means reducing the fixed effect of skin color does explain the data better and we have no evidence against the hypothesis that the simpler model explains the data. Thus, we have to choose the reduced model and model 2 is better.

Table 1: Comparison of model 1 and model 2 using the likelihood ratio test for question 1

#Df	LogLik	Df	Chisq	Pr(>Chisq)
11	-12860.36	NA	NA	NA
6	-12861.53	-5	2.335013	0.801111

Table 2: Comparison of model 2 and model 3 using the likelihood ratio test for question 1

#Df	LogLik	Df	Chisq	Pr(>Chisq)
6	-12861.53	NA	NA	NA
5	-12870.27	-1	17.49717	2.88e-05

Then we introduce model 3 by considering median income level, age and gender. For model 3, the p-value for customers' median income level and age are 0, which we have very strong evidence to say the probability of buying our new products is related to these characteristics. After adding the coefficients of each predictor, the model3 can be the following:

$log(\frac{\hat{p}}{1-\hat{p}}) = 0.0321 - 0.2992 Median\ \ Income + 0.4075 Rescaled\ \ age + 0.1728 Sex_{Intersex} + 0.0363 Sex_{Male}$

In order to determine the final fitted model, we need to do the likelihood ratio test for model 2 and model 3 and the result table is the following:

Given the table of testing the likelihood ratio of model2 and model3, we can find the p_value is so small around 0, which is less than 0.05. Including the term of population for the scenario does explain the data better and we have very strong evidence against the hypothesis that the simpler model fits the data. Thus, model 2 is better.

Finally, we introduce the model 4 with fixed effects: median income level, age, gender, population of location and whether or not having sleep tracking function of the products.

Table 3: A summary of model 4 for research question 1

term	estimate	std.error	statistic	p.value
(Intercept)	4.3515	0.4107	10.5954	0.0000
Median_Income	-0.3166	0.0162	-19.5590	0.0000
scales::rescale(age)	0.4058	0.0722	5.6215	0.0000
sexIntersex	0.1532	0.1412	1.0856	0.2777

term	estimate	std.error	statistic	p.value
sexMale	0.0318	0.0305	1.0420	0.2974
Popu	-0.0666	0.0158	-4.2062	0.0000
Sleep_trackingYes	-4.3708	0.4096	-10.6698	0.0000

From the table of the model4, the p-value for customers' median income level, age, population and our products of function sleep tracking are all 0, which we have very strong evidence to say the probability of buying our new products is related to these characteristics. Even though the p-value for sex is not that significant, we can still keep this variable in our model since we discovered the graphs in EDA sections the probability of consumers for each gender is totally different. Therefore, after adding the coefficients of each predictor, the model4 can be the following.

$log(\frac{\hat{p}}{1-\hat{p}}) = 4.3515 - 0.3166 Median\ \ Income + 0.4058 Rescaled\ \ age + 0.1532 Sex_{Intersex} + 0.0318 Sex_{Male} - 0.0666 Population - 4.3708 Sleep\ tracking_{Yes}$

After that we need to do the likelihood ratio test for model 2 and model 4 to test the more proper model as our final model.

The p-value for the likelihood ratio test of model2 and model4 is 0, which is less than 0.05. Including the fixed effect, whether or not having sleep tracking in our devices, does explain the data better and we have very strong evidence against the hypothesis that the simpler model fits the data. So by comparing these two models, we get a more applying model as following and model 4 is the final model.

$log(\frac{\hat{p}}{1-\hat{p}}) = 4.3515 - 0.3166 Median\ \ Income + 0.4058 Rescaled\ \ age + 0.1532 Sex_{Intersex} + 0.0318 Sex_{Male} - 0.0666 Population - 4.3708 Sleep\ tracking_{Yes}$

In specific, when we keep rescaled age, population of location, whether or not having sleep tracking function on the devices and gender fixed, 0.7286 is when median income level increases 1 unit, the odds of probability of buying new products will decrease 0.2714 unit. It means that if we look at the median income level, the probability is negatively related to consuming the new

Table 4: Comparison of model 2 and model 4 using the likelihood ratio test for question 1

#Df	LogLik	Df	Chisq	Pr(>Chisq)
6	-12861.53	NA	NA	NA
7	-12556.35	1	610.3479	0

products. Then, we keep median income level, gender, population of location, whether or not having sleep tracking function on the devices fixed, 1.5005 is when rescaled age increases 1 unit, the odds of probability of purchasing new products will increase 0.5005 unit, which means that if we focus on the age, the probability is positively related to consuming the new products. If we keep the median income level, rescaled age, gender and whether or not having sleep tracking function on the devices fixed, 0.9356 is when population of the location increases 1 unit, the odds of probability of buying new products will decrease 0.0644 unit, which indicates that if we focus on the population, the probability is negatively related to buying the new products. Also, when we keep rescaled age, median income level, gender and population of location fixed, when we compare with the products without the sleep tracking function, the devices with sleep tracking function will lower the odds of probability 0.9874 multiples. Therefore, the devices with sleep tracking function is less popular than without this function.

When we discuss how gender effects the odds of probability of purchasing the new products. When we keep rescaled age, median income level, population of location and whether with or without sleep tracking function fixed, when we compare with the customers who are female, the customers who are intersex will larger the odds of probability 0.1656 multiples. Therefore, customers who are intersex will better accept the new products. Moreover, when we compare with the customers who are female, the male customers will larger the odds of probability 1.0323 multiples. In this case, male customers will favor our new products more than female customers.

From the final model, we can find our potential new customers purchasing new products are related to their income level, ages, gender, the population of their location and the production function with sleep tracking during the night. Thus, we can make predictions about the characteristics of consumers for our new products: The consumers who are aged, intersex or male with less income, comes from a rural area and are not willing to have sleep tracking services will buy our new products.

Determine whether the devices perform poorly for consumers with darker skin color

Methods: EDA

In order to solve if customers with dark skin color are facing the problem of poor-performing on sleep scores during the time of using the device, we concentrate on the number of times there was a quality flag during the sleep session and duration, in minutes, of sleep session. In particular, a quality flag may occur due to missing data, or due to data being recorded but sufficiently unusual to suggest it may be a sensor error or other data quality issue, therefore, more flags mean the poorer performance on sleep scores. Moreover, by common sense, we notice

that as the duration of sleep session increases, the possibility of showing the accidental error or missing data on the devices will also increases which indicate that the number of quality flags will increase as well. By showing the graph of the relation between the number of flags and the duration of sleep sessions with the consideration of position, we could verify if there is a positive relationship between them. Thus, we could show the relationship between the mean value of the duration and the mean value of the number of times there was a quality flag among all of the positions in Canada by using postcodes.

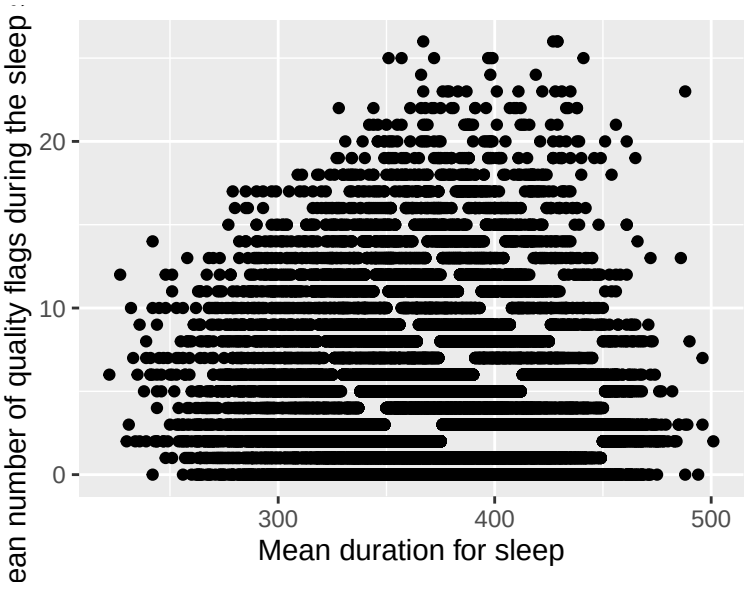

Figure 6: Realtionship between mean duration and mean flags

From the figure above, we can clearly see a pattern between the number of occurrences of quality flags and the duration of sleep sessions and the relation is approximately positive. As the duration increases, the number of flags showing up will also increases. Even though the points on the graph is not concentrate on a line, we can still see a weak positive patter generally.

To clarify the duration is not related we should also analyze whether the duration of sleep sessions is longer for people with darker skin color since the poor performance of the devices for people with darker skin occurring may be related to their longer sleeping duration for them.

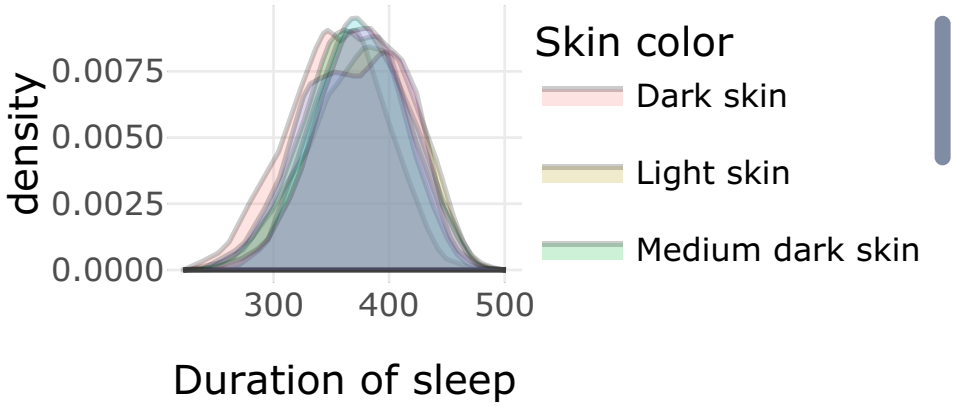

Figure 7: The duration of sleep session for people with all different skin colors

As observed, the figure of the duration of sleep for different skin color types shows that the duration is roughly normally distributed for customers with all types of skin colors, which means the duration of sleep sessions is pretty similar among all people. We can also see on the graph that the duration of sleep sessions for people with darker skin color is actually a little bit less than the duration of sleep sessions for other people which indicates that the possibility of having the flags during sleep sessions should be less for customers with darker skin. In this case, we can conclude that the statement, the poor performance of the devices for people with darker skin occurring is caused by longer duration of sleep sessions for them, is not reasonable, but there exists a positive relationship between the number of times there was a quality flag and duration.

To clearly verify that consumers with darker skin color actually have more time of facing the issue of missing data and technical errors when using the devices, we can also have a look at the number of flags per minute of duration. The reason for considering the mean value of the number of flags' occurrence over duration is to reduce the effect of a positive relationship between the number of times there was a quality flag and the duration of the sleep session. Therefore we will focus on the value of the number of flags' occurrence over the duration as our responding variable.

From the table above, we can clearly see that customers with darker skin color actually having more times there was a quality flag showing up per minute during the sleep session. The mean value of times there was a quality flag showing up per minute for people with dark skin is 0.0333971 and 0.0201984 for consumers with medium dark skin, which are the top two high mean value in the table above. All the mean value of number of flags showing up over duration for people with other skin color are under 0.01. Specially, for light skin consumer, they have the

Table 5: Table for the mean value of number of times there was a quality flag over duration of sleep session for all types of skin colors

skin_color	mean(flags/duration)
Dark skin	0.0333994
Light skin	0.0030615
Medium dark skin	0.0202140
Medium light skin	0.0066393
Medium skin	0.0099133
others	0.0065259

lowest mean value of number of times there was a quality flag per minute.

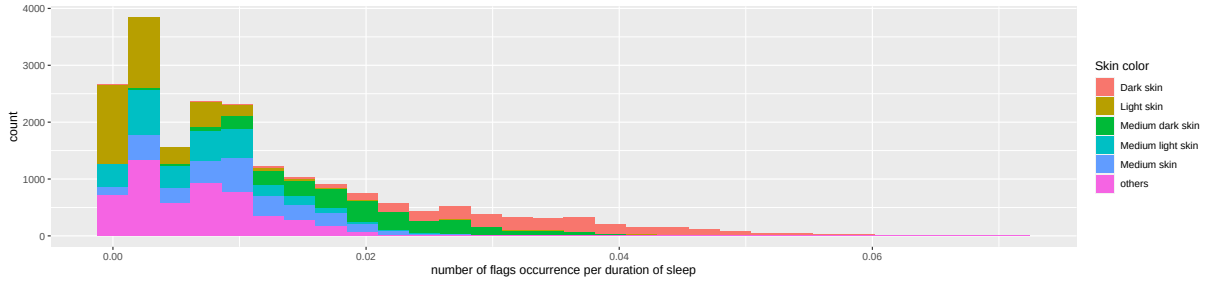

Figure 8: The histogram of number of times there was a quality flag over duration of sleep session for people with all types of different skin colors

Consequently, from the histogram of the number of times, there was a quality flag over the duration of the sleep session, clearly shows that for a higher value of the number of times there was a quality flag over the duration, almost all of the observations are with dark skin or medium-dark skin on the right tail of the graph. The overall histogram seems to be right-skewed with a tail on the right side which means there are fewer observations with a high value of the number of times there was a quality flag over the duration. In our cases, the highest value of the number of times there was a quality flag over duration is approximately 0.07 and belongs to the group of people with dark skin color. The lowest value of the number of times there was a quality flag over duration is 0 and belongs to groups of people with light or medium-light or medium or other skin colors. Generally, the trend of the graph indicates that people with darker skin colors are actually having the problem of more times there being a quality flag over the duration. Thus, we can conclude that the devices are performing poorly for users with darker skin.

Then we consider whether the types of devices affect the result of the value of the number of flags over the duration. If all of the devices have a similar problem of having quality flags during

the sleeping session, then we should also consider the effect of the type of devices on the result of the number of flags.

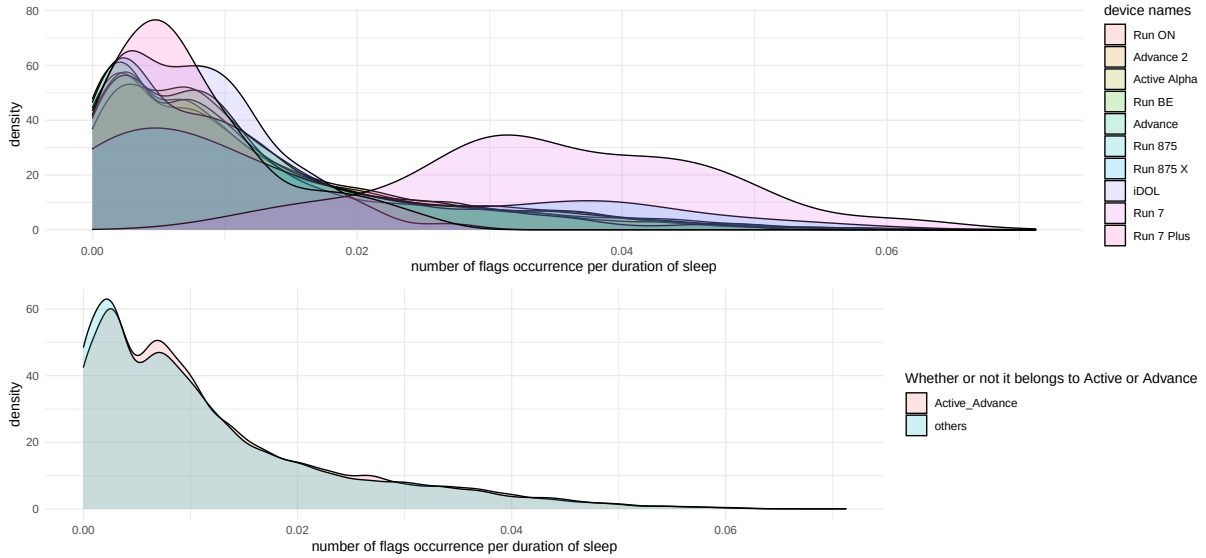

Figure 9: Above: Number of flags occurrence per duration of sleep for different devices, Below: Number of flags occurrence per duration of sleep for our devices

In spite of the above graph in figure 7, the only device "Run 7" have a higher number of quality flags over duration during the sleep session and the result of the number of flags over duration for all the other devices are pretty similar compared to each other with a right-skewed distribution. For the device "Run 7", the graph is almost normal distributed with two modals at about 0.03 and 0.045. Thus, the type of devices does not affect the number of appearances of error for the devices and we can ignore the effect of the types of devices.

In addition, the following graph also represents that the appearance of the number of quality flags over duration for the new devices 'Active' and 'Advance' has an almost same right-skewed distribution as the number of quality flags over duration for other devices, which also demonstrate that the number of quality flags over the duration of sleep session is not associated with the type of devices. As observed, there are two significant modals on the right side of the graph. When the number of quality flags over the duration of a sleep session is less than about 0.0025, the number of quality flags over the duration of a sleep session for other devices is greater than our new devices 'Active' and 'Advance'. Regardless, when the number of quality flags over the duration of a sleep session is greater than 0.05, the number of quality flags over the duration of sleep session for our new devices 'Active' and 'Advance' is almost all greater than other devices. Therefore, as the value of the number of quality flags over the duration of a sleep session becomes greater, our devices are more likely to face the missing value or unusual errors while using.

Methods: Applying the models to our data

Even though we have a lot of evidence above to show the poor performance of the devices for people with darker skin color, we should make some models to make our proper final conclusion. As our responding variable is the number of times over the duration of the sleep session there was a quality flag, the response type for our scenario is Poisson distributed, so we need to use the generalized linear model or the generalized linear mixed model to solve the problem. In this case, we assume the variance of the model is equal to the mean of the model.

Firstly, we set up the generalized linear model between skin color and the number of times there was a quality flag with age, sex and the line of products this device belongs to as confounding variables. $log(\frac{E(flags)}{duration}) = \beta_0 + \beta_1 skincolor_{light} + \beta_2 skincolor_{Mediumdark} + \beta_3 skincolor_{Mediumlight} + \beta_4 skincolor_{Medium} + \beta_5 skincolor_{others} + \beta_6 RescaledAge + \beta_7 Sex_{Intersex} + \beta_8 Sex_{Male} + \beta_9 line_{Advance} + \beta_{10} line_{iDOL} + \beta_{11} line_{Run}$ Secondly, we set up the generalized linear model between skin color and the number of times there was a quality flag with age and sex as confounding variables.

$log(\frac{E(flags)}{duration}) = \beta_0 + \beta_1 skincolor_{light} + \beta_2 skincolor_{Mediumdark} + \beta_3 skincolor_{Mediumlight} + \beta_4 skincolor_{Medium} + \beta_5 skincolor_{others} + \beta_6 RescaledAge + \beta_7 Sex_{Intersex} + \beta_8 Sex_{Male}$

Thirdly, we set up the generalized linear model with the fixed effect skin color, age and sex and random effect customer id.

$log(\frac{E(flags)_i}{duration}) = \beta_{0i} + \beta_{1i} skincolor_{light} + \beta_{2i} skincolor_{Mediumdark} + \beta_{3i} skincolor_{Mediumlight} + \beta_{4i} skincolor_{Medium} + \beta_{5i} skincolor_{others} + \beta_{6i} Rescaledage + \beta_{7i} sex_{Intersex} + \beta_{8i} sex_{Male} + U_{CustomerIDi}$

Fourthly, we set up the generalized linear model with the fixed effect skin color, age and sex and random effect customer ID and device ID. $log(\frac{E(flags)_i}{duration}) = \beta_{0i} + \beta_{1i} skincolor_{light} + \beta_{2i} skincolor_{Mediumdark} + \beta_{3i} skincolor_{Mediumlight} + \beta_{4i} skincolor_{Medium} + \beta_{5i} skincolor_{others} + \beta_{6i} Rescaledage + \beta_{7i} sex_{Intersex} + \beta_{8i} sex_{Male} + U_{CustomerIDi} + U_{DeviceIDi}$

By applying the likelihood ratio test between each two of them we could finally conclude our final model to answer the question of whether or not the customers with dark skin color are facing the problem of poor-performing on sleep scores during the time of using the device.

Result

Firstly, we discuss the generalized linear model between skin color and the number of times there was a quality flag with age, sex and the line of products this device belongs to as confounding variables. From the graph showing the relation between the duration of sleep session and the number of flags' occurrences, we already notice a positive relationship between them, therefore,

Table 6: The result for the likelihood ratio test for model 1 and model 2

#Df	LogLik	Df	Chisq	Pr(>Chisq)
12	-42778.43	NA	NA	NA
9	-42781.69	-3	6.534254	0.0883214

we should set the offset of the model to avoid the effect on the result. After getting the estimate for this model we have the following formula for this model:

$log(\frac{E(\hat{flags})}{duration}) = -3.4036 - 2.3919 skincolor_{light} - 0.4992 skincolor_{Mediumdark} - 1.6160 skincolor_{Mediumlight} - 1.2116 skincolor_{Medium} - 1.6308 skincolor_{others} - 0.0479 RescaledAge - 0.0367 Sex_{Intersex} + 0.0049 Sex_{Male} + 0.0185 line_{Advance} + 0.0630 line_{iDOL} + 0.0284 line_{Run}$

The p-value for all types of skin colors and rescaled age are 0 which indicates very strong evidence to show that the number of times there was a quality flag has a relation with the skin color and age. However, for sex and line, the p-values are almost all above 0.05, which means the effect for sex and line of products is not that significant to consider, so we could ignore one of them to set a new model.

By ignoring the line of products this device belongs to, we set up model 2 which is the generalized linear model between skin color and the number of times there was a quality flag with age and sex as confounding variables. The formula for model 2 can be written as:

$log(\frac{E(\hat{flags})}{duration}) = -3.383 - 2.390 skincolor_{light} - 0.499 skincolor_{Mediumdark} - 1.614 skincolor_{Mediumlight} - 1.212 skincolor_{Medium} - 1.632 skincolor_{others} - 0.050 RescaledAge - 0.052 Sex_{Intersex} + 0.006 Sex_{Male}$

In order to figure out the fitted model, the likelihood ratio test is made and the result table is the following:

Given the table of likelihood ratio test for model 1 and model 2, the p-value is 0.0883214 which is greater than 0.05 which means reducing the fixed effect the line of products does explain the data better and we have no evidence against the hypothesis that the simpler model explains the data just as well. Thus, model 2 is better for our scenario.

Our next step is adding the random intercept into the model to see how fit is our new model with the variable customer ID. The following table indicates the estimate and p-value for each variable. The corresponding formula is following:

$log(\frac{E(\hat{flags})_i}{duration}) = -3.3838 - 2.3899 skincolor_{light} - 0.4988 skincolor_{Mediumdark} - 1.6138 skincolor_{Mediumlight} - 1.2121 skincolor_{Medium} - 1.6310 skincolor_{others} - 0.0501 Rescaledage - 0.0535 sex_{Intersex} + 0.0066 sex_{Male} + U_{CustomerIDi}$

Table 7: The result for the likelihood ratio test for model 2 and model 3

#Df	LogLik	Df	Chisq	Pr(>Chisq)
9	-42781.69	NA	NA	NA
10	-42769.14	1	25.10175	5e-07

Table 8: The result for the likelihood ratio test for model 3 and model 4

#Df	LogLik	Df	Chisq	Pr(>Chisq)
10	-42769.14	NA	NA	NA
11	-42768.94	1	0.3989498	0.5276321

For model 3, the p-values for all of the skin color types are all still nearly 0, which illustrates the relation between flags' occurrences and skin colors is significant to consider. In order to determine which model is better, we do the likelihood ratio test and the result is the following.

The previous table for the result of the likelihood ratio test of model 2 and model 3 shows that the p-value is 5e-07 which is smaller than 0.05, so we believe that model 3 (full model) better explains our data. Including a random effect for the scenario does explain the data better and we have very strong evidence against the hypothesis that the simpler model fits the data.

Finally, according to the information we get in the data, we can also add another random effect, the device ID, to see if there is a better model. The following table would be the summary table for model 4. The formula in this case can be written as:

$log(\frac{E(\hat{flags})_i}{duration}) = -3.3834 - 2.3901 skincolor_{light} - 0.4983 skincolor_{Mediumdark} - 1.6145 skincolor_{Mediumlight} - 1.2118 skincolor_{Medium} - 1.6306 skincolor_{others} - 0.0481 Rescaledage - 0.0514 sex_{Intersex} + 0.0069 sex_{Male} + U_{CustomerIDi} + U_{DeviceIDi}$

Compared model 3 and model 4, the p-value is 0.5276321 which is greater than 0.05, which means the random effect is an unnecessary complication to our model and we have no evidence against the null hypothesis that the simpler model explains the data. As described, we can finally get our final model is the third model we set, and the mathematical formula is the following:

$log(\frac{E(\hat{flags})_i}{duration}) = -3.3838 - 2.3899 skincolor_{light} - 0.4988 skincolor_{Mediumdark} - 1.6138 skincolor_{Mediumlight} - 1.2121 skincolor_{Medium} - 1.6310 skincolor_{others} - 0.0501 Rescaledage - 0.0535 sex_{Intersex} + 0.0066 sex_{Male} + U_{CustomerIDi}$

Table: Summary table for the fixed effects of model 3

term	estimate	p-values
(Intercept)	-3.3838	<2e-16
skin color Light skin	-2.3899	<2e-16
skin color Medium dark skin	-0.4988	<2e-16
skin color Medium light skin	-1.6138	<2e-16
skin color Medium skin	-1.2121	<2e-16
skin color others	-1.6310	<2e-16
rescaled age	-0.0501	0.0051
sexIntersex	-0.0535	0.1592
sexMale	0.0066	0.4116

Be specific, keep sex, the mean number of quality flags divide by duration for people with light skin color is 0.0916 times lower than the reference group (people with dark skin); the mean number of quality flags over duration for people with medium dark skin color is 0.6073 lower than the reference group (people with dark skin); the the number of quality flags over duration for people with medium light skin color is 0.1982 lower than the reference group (people with dark skin); the mean number of quality flags over duration for people with medium skin color is 0.8089 lower than the reference group (people with dark skin); the mean number of quality flags over duration for people with other skin color is 0.1957 lower than the reference group(people with dark skin).

Moreover, keep skin color, sex and Line of products this device belongs to fixed, when rescaled age increases 1 percent range of age, the mean value of number of flags over duration of sleep session will change to 0.9511 times the original the mean value of number of flags over duration. Keep skin color, rescaled age and line of products this device belongs to fixed, the mean number of quality flags over duration for people with other skin color is 0.9479 lower than the reference group (females); the the number of quality flags over duration for people with other skin color is 1.0066 lower than the reference group (females).

From all of the summary tables for these four models above, it is easy to observe a similarity which is that for the variable skin color, the estimate is always negative and the reference group for all of them is people with dark skin color. In this case, we could conclude that the number of times there was a quality flag during the sleep session is less for people with lighter skin colors and people with darker skin colors are actually facing the issue of poorly performing on their devices. By observing the EDA, we also notice that this is a problem for all of the devices

not just our new devices. Therefore, if we could develop our devices to solve this issue, our consumers may more likely to purchase our devices in the future and this action may also help to enhance the reputation of our company on the market.

Discussion

For the first research question, we find that there are various important factors to contribute to the willingness of buying the product "Active" or "Advance". Referring to the figures for the number of purchases on our new products among customers' age, gender and median income level distributions, we can find a difference between the preference of customers. From the final model of the probability of buying devices from the lines "Active" or "Advance", the income level, the age, the gender, the population of the location of our customers and whether having sleep tracking on the products shows the significant influence on the purchasing. Thus, we can make a prediction about the characteristic of consumers for our new products. The older the customers are, the more likely they will buy new products; customers who are intersex and males will prefer to buy our products; customers who have lower income will favor our new products; the customers who come from a small city will buy our products and our products will be more popular if take out the sleep tracking function. In conclusion, our marketing policies should target the group of people who are male or intersex and come from rural places with lower income levels. Moreover, to enhance the sales volume of new products, it is better to take away our sleep tracking functions.

For the second research question, we find that the there is actually an issue for people with darker skin color while using the devices during sleep. It may more likely cause missing data, unusual data recorded or other data quality issue when customers with darker skin color access our devices. From the figure of number of flags occurrence over duration of sleep, we can easily observe that people with darker skin color have higher number of flags occurrence over duration of sleep which indicates that they having more chance to get a quality flag on their advice during the sleep session. Besides, the final fitted model between the number of quality flags and skin color is also representative for our above conclusion since the meaning for estimate of each variables shows that customers with lighter skin colors have less probability of getting missing data or other quality issue. Therefore, in fact, the devices perform poorly for users with darker skin. If our team could solve this problem and develop the performance for users with darker skin, more customers will join and enjoy our devices and also may also help us enhance the reputation in the market.

Strengths and limitations

In this whole report, we strictly follow the statistical rules and apply statistical methods. To find out the potential new product users, we used logistic regression and a generalized linear model. Moreover, in order to look for the characteristics of groups of consumers that have the worst performance in monitoring the health data during sleep, we used Poisson distribution and a generalized linear mixed model to solve the problem.

For the limitation, through our process of left join or cleaning up, some of the observations data may not be covered in our analysis. As the variable skin color is created by using the emoji modifier which may not accurate for us to identify the personality of each customer. Moreover, the income in our data is not the individual income, instead, it is the average income for the area, so this may cause our result to be not accurate. Finally, the relation between the postcode and the ID of subdivisions in Canada is not one to one corresponded. Finally, when we are crawling data, we are using data from 2016 that is not the most recent data, which also gives us some limitations on the results.

Consultant information

Consultant profiles

Zhuoxuan Li. Zhuoxuan is a senior technical consultant at our company who is responsible for gathering all data sets from the web and helping the company get the final data for each question and providing code for charts and models while adhering to strict ethical boundaries. He will graduate from the University of Toronto, Canada in 2023.

Zihan Zhang. Zihan is a junior consultant in Uoft superteam. For this project, her job is to create tables and graphs in this project. Zihan earned her Bachelor of Science from University of Toronto in 2023.

Ziqi Wang. Ziqi is a junior consultant in Uoft superteam. Her job is to analyze the output of the graphs and tables with statistical thoughts and link it to the reality. Ziqi earned her Bachelor of Science from University of Toronto in 2023.

Ziqing Gao. Ziqing is another junior consultant in Uoft superteam. Her job is to write reports for the output of the graphs and tables with statistical thoughts and applying to the realistic questions. Ziqing earned her Bachelor of Science from University of Toronto in 2023.

Code of ethical conduct

- To inform public opinion and policy, it is essential to maintain objectivity and avoid procedural or personal bias.
- Execute and document work with care and diligence according to the employer's or client's requirements.
- You should not disclose or give authorization to disclose confidential data acquired during professional practice to a third party for personal gain or benefit without the prior written permission of an employer or client, or as directed by a court.

Reference

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

[2] Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

[3] C. Sievert. Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC Florida, 2020.

[4] Dmytro Perepolkin (2019). polite: Be Nice on the Web. R package version 0.1.1. https://CRAN.R-project.org/package=polite

[5] 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.

[6] End-use licence agreement for Postal Codeom Conversion files. my.access - University of Toronto Libraries Portal. (n.d.). Retrieved April 10, 2022, from https://access.library.utoronto.ca/data/access_postalcode.php?url=https%3A%2F%2Fmdl.library.utoronto.ca%2Fsites%2Fdefault%2Fpublic%2Fmdldata%2Frestricted%2Fcanada%2Fnational%2Fstatcan%2Fpostalcodes%2Fpccf%2F2016%2F2021aug%2FpccfNat_fccpNat_082021sav.zip

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

[8] Hadley Wickham and Jim Hester (2021). readr: Read Rectangular Text Data. R package version 2.0.2. https://CRAN.R-project.org/package=readr

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

[10] Kamil Slowikowski (2021). ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. R package version 0.9.1. https://github.com/slowkow/ggrepel

[11] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

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

[13] Kamil Slowikowski (2021). ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. R package version 0.9.1. https://github.com/slowkow/ggrepel

[14] Population density. Census Mapper. (n.d.). Retrieved April 10, 2022, from https://censusmapper.ca/

Appendix

When we try to make research for these two questions, we need many data to help us to continue to do the project. So, we start to collect external data online.

First, we scrap the data of the information of different devices from a website and the URL is "https://fitnesstrackerinfohub.netlify.app/" with the user agent "zhuoxuan.li@mail.utoronto.ca for STA303/1002 project". Then, we can use the code "target <- bow(url, user_agent ="liza. bolton@utoronto.ca for STA303/1002 project", force = TRUE)" to see the web scraping information. We discover that the path is scrapable for the user agent with a crawl delay of 12 sec. It means we are available to scrap information from this website. Moreover, we use this code "html <- scrape(target)" and "device_data <- html %>% html_elements("table") %>% html_table() %>% pluck(1)" let the information of different devices become a simple table that's easier to see and research. Then, we will use this table be our data for the next steps.

Second, we will use the API method to collect the data on people's median income from the website "https://censusmapper.ca/". On this website, we can see the income of people in different areas. Then, we find and download the information on incomes by using the API key from the user's profile with the code support from our professor. In here, we should not scrape data because we must use API if the website has a public API.

Then, we use the code "median_income <- census_data_csd %>% as_tibble() %>% select(CSDuid = GeoUID, contains("median"), Population) %>% mutate(CSDuid = parse_number(CSDuid)) %>% rename(hhld_median_inc = 2)" to get the only information of median income of people in different areas. Until now, we already get the information of income thats' useful in our research questions.

Last, We collect data from a file called "Census Canada Postal Code Conversion Files". We could use this file with a license agreement because we are the University of Toronto students. But we could only use this kind of file by ourselves and not give it to others. In this files, we collect a .sav file called "data-raw/pccfNat_fccpNat_082021sav.sav" by using the code "dataset= read_sav()" and the data from 2016. Then, we use this code "rename(c("postcode"="PC")) %>% group_by(postcode, CSDuid)" to change the name of "PC" in the data and group by this kind of data for good looking. Then, we will use the Collated data be the data we will use in the next steps.

Until now, we use the "Web scraping" and API methods to collect the data we need. But whether the data is very small or not important, we still need to really care to read the term and conditions on a website and make sure these data are public for us to use.