Data Analysis for Mingar

The new customers have lower median income and people with darker emoji have more flags during sleep sessions

Report prepared for MINGAR by CJYB

2022-04-07

Contents

Executive summary

We are a consulting firm that specialize in statistical consulting. Mingar has given us two problems to solve. The first problem is to find out who are the new customers that use the "Active" and "Advanced" product and compare them to the traditional users. With emphasis on whether the two new lines have attracted customers outside of the traditional line or not. The second problem is that the social media team of Mingar found a trend that suggests that customers with darker skins result in poorer performance on the devices. We were asked to investigate that.

Problem 1:

- Median income is the only variable that really impacts the outcome of whether a customer is using "Active and Advanced" or traditional products.
- The average median income for customers who bought the "Active and Advanced" line is $67,487.65CAD, the average median income for the traditional customer is $73,550.55CAD. The average median income for traditional is 8.98% higher than the median income for "Active and Advanced".
- The average median income for traditional is 8.98% higher than the median income for "Active" and "Advanced".

Problem 2:

- We have observed that users who have darker skin have a higher number of flags in contrast to those who have lighter skin.
- Young users who have dark skin have the highest number of flags.
- We would suggest the development team of Mingar to take a look at the censor of each device as that may be the cause of why users with darker skins have more flags.

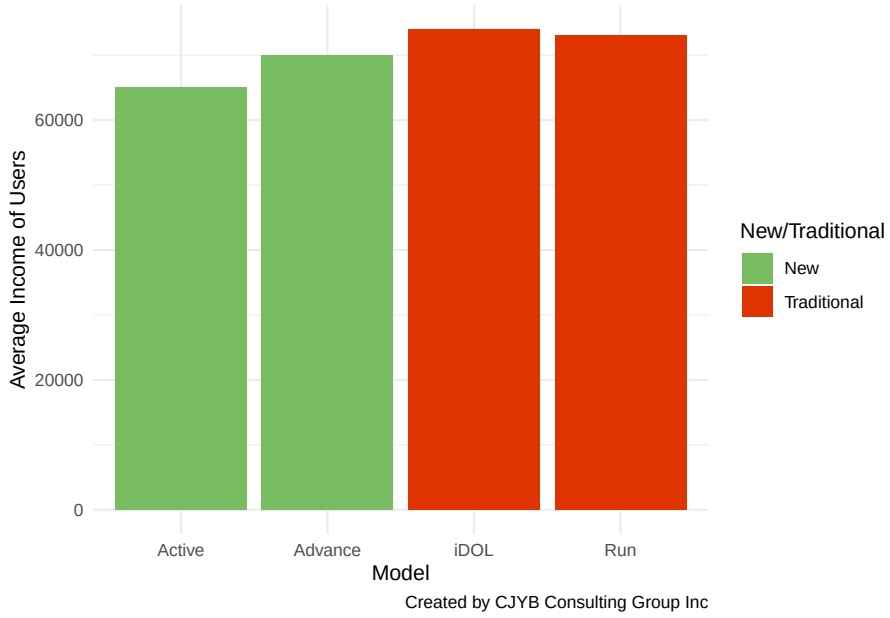


Figure 1: Comparison of Average Income of Different Device Users

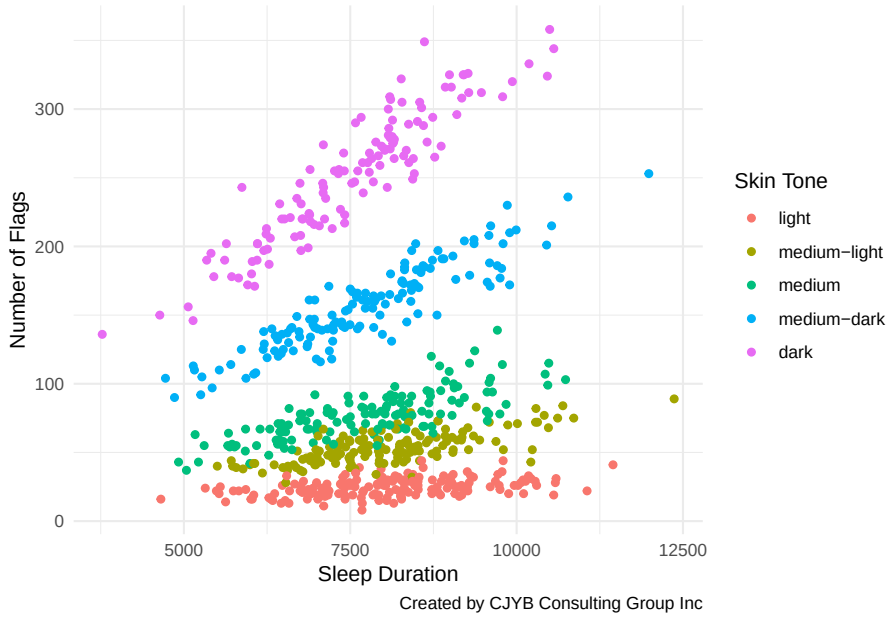


Figure 2: Plot the relationship between flags and duration grouped by skin tone pf users

Technical report

Introduction

Famous Statistician Karl Pearson once said "statistics is the grammar of science". Even though statistics can be misleading sometimes, it is still one of the best ways at supporting the reasoning behind a cause when you can't find the true reason behind said cause. CJYB Consulting Group Inc. is a consulting firm that specializes in statistical consulting. Mingar is a company that produces products aimed for outdoor recreation, such as high-end fitness tracking devices. They recently added two new product lines, "Active" and "Advanced", at a more approachable price point for the average consumers in contrast to their traditional customers.

Research questions

Mingar has given two questions to CJYB Consulting Group Inc. to resolve:

- The first question is from their marketing team, asking who the new customers are. How are the buyers for the newer "Active" and "Advanced" different from the traditional customers with focus on whether the two new lines had attracted new customers outside of the traditionally higher income base.
- The second question is from their social media team. They have picked up a trend in complaints regarding customers with darker skins having poor performance from their devices, especially with the sleep score. We were asked to investigate this.

Methods

Data description and wrangling

For our statistical analysis, we will prepare two data sets, one to support our analysis on research question 1 and another to support our analysis on research question 2. For the first question we are interested in analyzing the median income difference between new customers and old customers, so we would be using the customer's data, median income to predict whether the customer would buy the new line of device or the old line of device. For the second question, we are interested in analyzing the number of flags, which indicates quality issues in data by the customer's data, especially the skin tone of the customer which we would use the emoji modifier to approximate [5].

For both data sets, we will use the customer ID, date of birth, biological sex, postal code, device ID, and the emoji modifier from the customer list dataset and combine it with the device data to get the corresponding device name and line of device of each device ID so we can obtain each customer's specific device and line of their device [6]. We will also generate a new variable called release, which is a binary variable indicating whether the line of device is one of the newer lines represented by 1, or the traditional line represented by 0 in order to differentiate the new customers from the traditional customers [6].

Note that the date of birth was converted to a discrete variable current age which is rounded down to the closest year. Also note that the emoji modifier was converted into a discrete variable named skin tone which used conversion of the unicode to a skin tone color from the emoji chart from unicode.org [6].

For the first dataset we will use the median income and CSduid data from web scraping CensusMapper [8]. The CSduid is a location code used in the census, we will use the conversion file provided by Statistics Canada to convert each CSduid to the first corresponding postal code, for postal codes with two CSduid, we will take the mean of the median income of the same postal code so there is one median household income corresponding to each postal code. Then we will combine the customer data with the median income and postal code data by matching the postal code to get a predicted household median income for each customer.

For the second data set, we will take the customer ID, sleep duration, and the number of flags which indicate quality issue of the data or missing data from the sleep log data, and sum the sleep duration and flags of every individual customer based on each sleep log's customer ID in order get a sleep log dataset which each observation is independent to each other. Then this new summed sleep log dataset would be combined with the customer data set so each customer would have their own summed sleep log if their sleep data was logged [6].

Note that we omitted any observations that is missing a data point for both datasets, so specifically, we omitted any customer that had NA for sex but we included intersex, and we ommited customers that had NA for their emoji modifier, which means they had their skin tone set as the default yellow color which is not a real skin tone. After the omissions we have 14267 customer's data for the first dataset and 719 customer's data for the second data set.

Data Visualizations and Summary

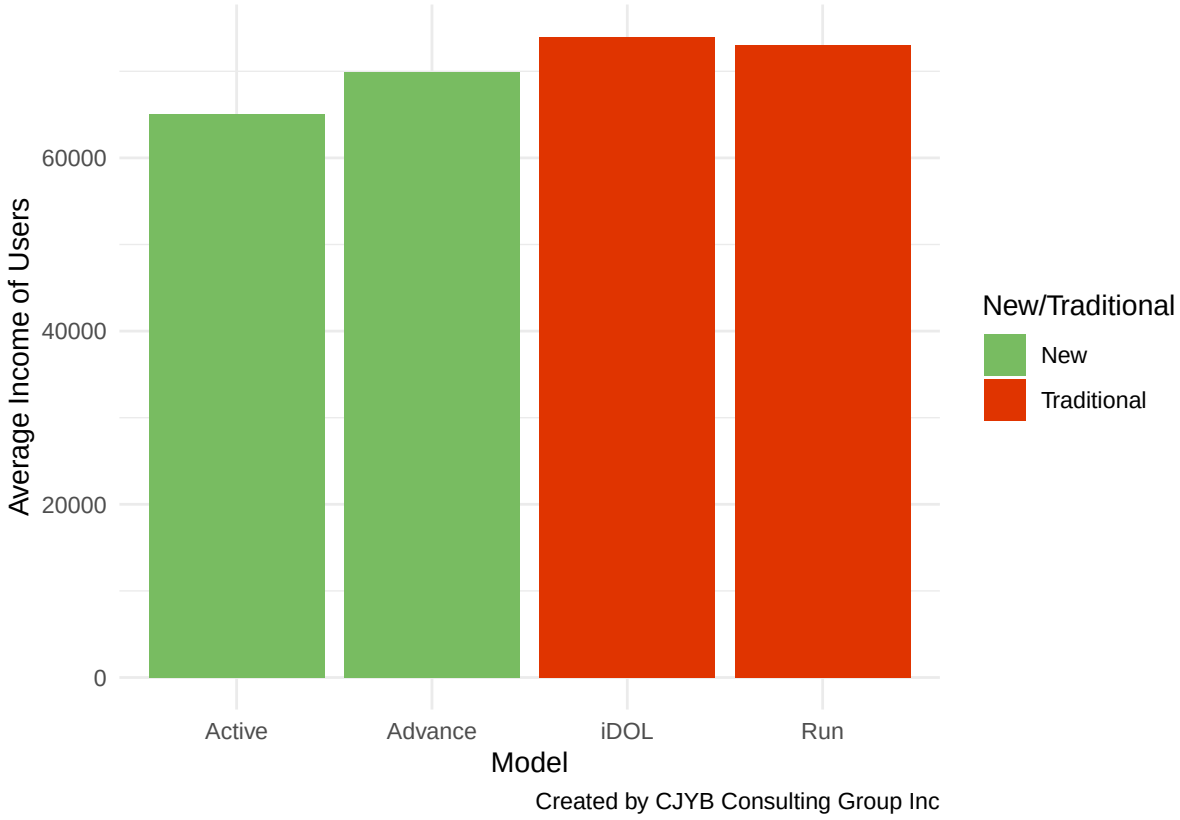


Figure 3: Comparison of Average Income of Different Device Users

Figure 3 is a bar graph showing the average income of the users for each line of device, and using color to indicate the traditional lines and the new lines. We can observe that the two traditional lines have higher average income than the two new lines.

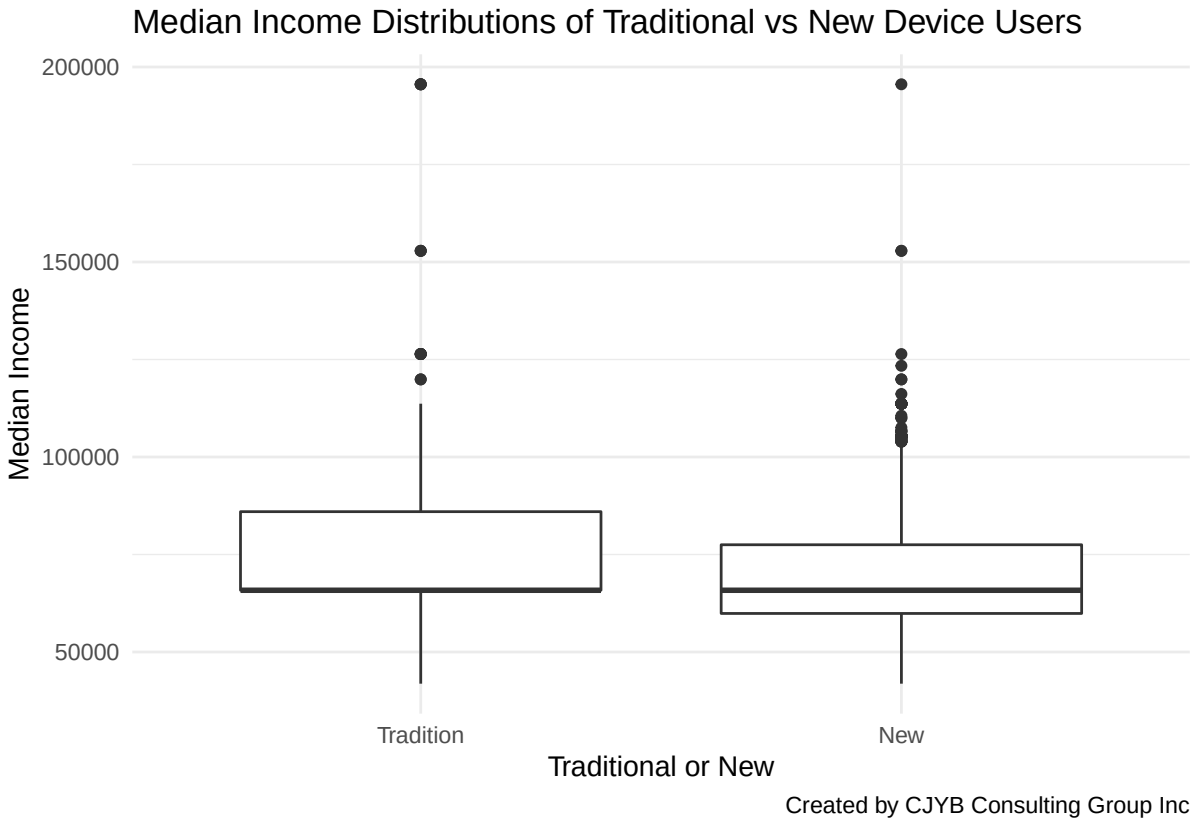


Figure 4: Median Income Distributions of Different Device Users

Figure 4 is a box plot depicting the income distribution of the old lines of devices and the new line of devices. We can observe that the box representing the income distribution of the new line of devices is lower than the box representing the income distribution of the old lines of devices.

Table 1: Comparing different device users based on median income

	Traditional	New
[40000,79999]	4195	6068
(80000,119999]	2141	1845
(120000,159999]	9	4
(160000, 200000]	4	1

Table 2: Comparing proportions of median income of users

	Traditional	New
[40000, 79999]	0.66	0.77
[80000, 119999]	0.34	0.23
[120000, 159999]	0	0
[160000, 200000]	0	0

Table 1 and 2 shows the frequency and proportion respectively of users between the traditional and new line of 4 different income levels. We can observe that there are more new users in the lowest level of income, specifically 6068 compared to 4194, and from table 2 we can observe that 66% of the traditional users are in the lowest income level while 77% of the new users are in this income level . We can also observe that there are more traditional users in the second level of income which is the higher level of income than the new users, specifically 2141 to 1845, and from table 2 we can observe that 34% of the traditional users are in this higher level of income while 23% of the new users are in this higher level of income. The third and fourth level of income has too few users that the data is not relevant as we can see they are both 0% rounded to 2 decimal places in table 2.

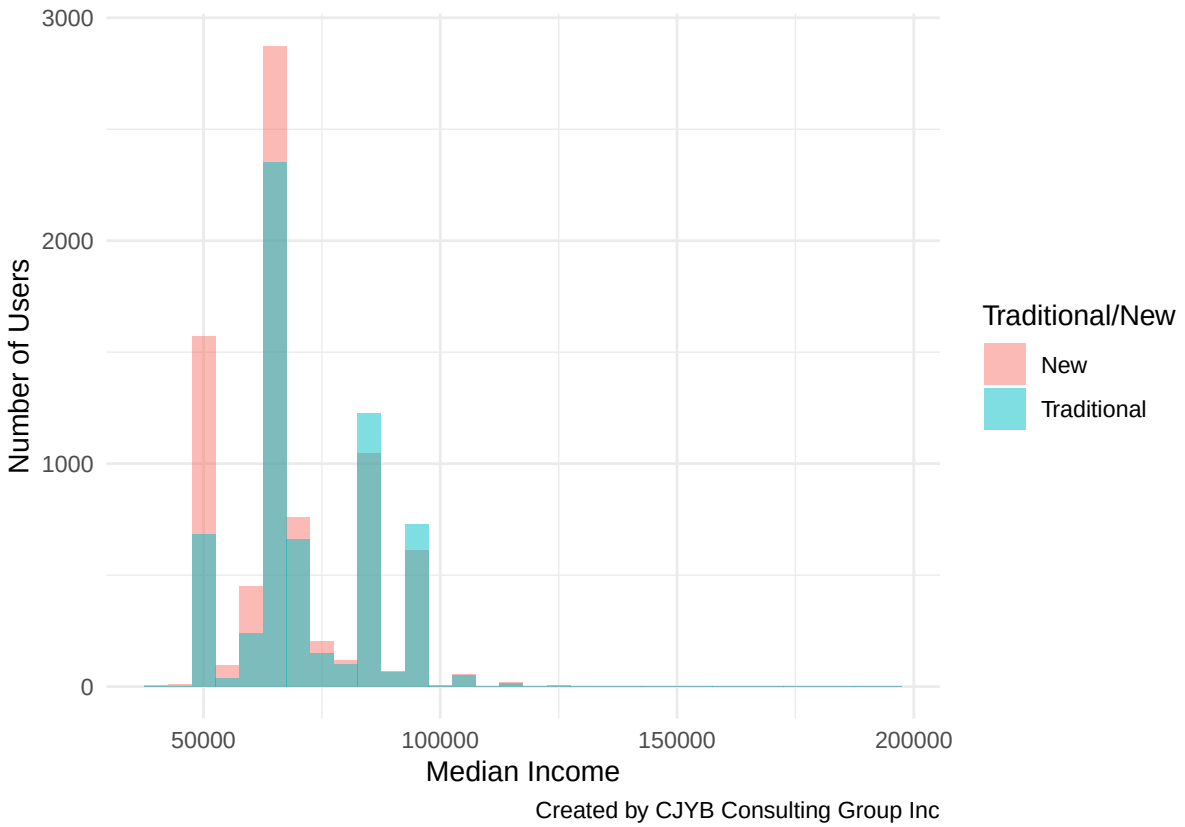


Figure 5: Comparison of Median Income of Different Device Users

Figure 5 is a histogram showing the distribution of median income, differentiating the distribution from the traditional line of users from the new line of users. We can observe that all the bins of income have both new users and old users, meaning that there are new and old customers in all income ranges. But we can see that there are more new users in the left bins meaning there are more new users that have lower median incomes, while there are more traditional users in the right bins meaning that there are more traditional users that have higher income.

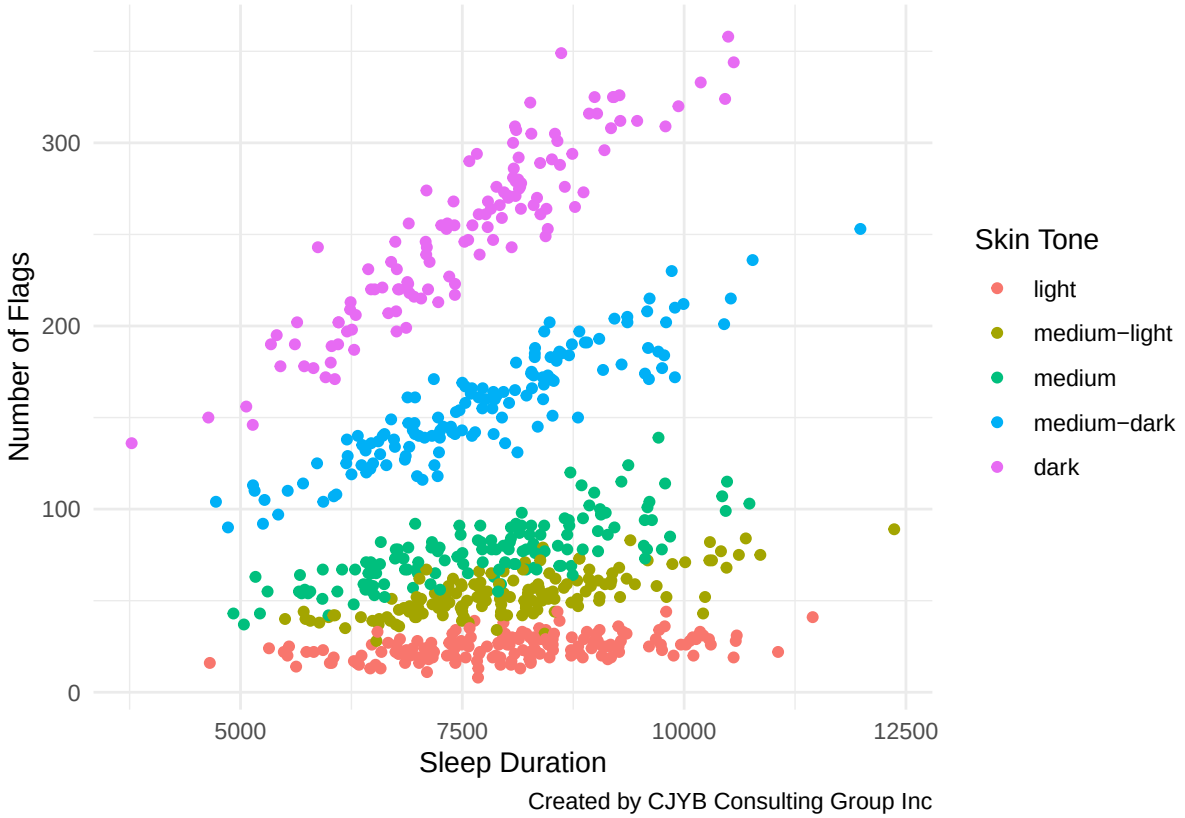


Figure 6: Plot the relationship between flags and duration grouped by skin tone pf users

Figure 6 shows a scatter plot graphing the duration versus the number of flags with different skin tones indicating different colors. From this scatter plot, we can observe that there are 5 very obvious separation of points which are separated by the colors of skin tone, where the darker the skin tone the higher the number of flags are. We can also observe a slight positive slope between the duration and the flags, but it is not obvious in the lighter skin tones.

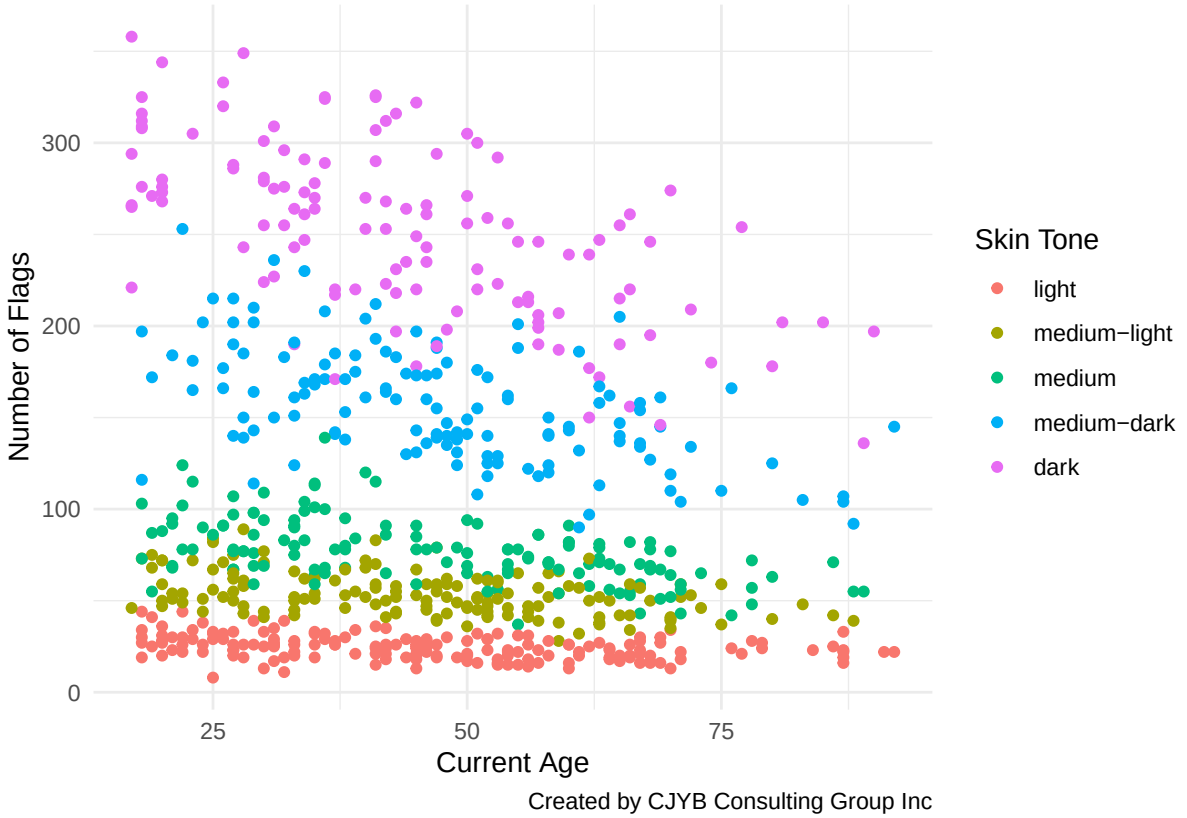


Figure 7: Plot the relationship between flags and age grouped by skin tones of users

Figure 7 shows a scatter plot graphing the current age versus the number of flags with different skin tones indicating different colors. From this scatter plot, we can observe again the 5 obvious separation of points which are separated by the colors of the skin tone, we can also observe a slight decreasing slope between the current age and the number of flags.

Method Description for Question 1

For the nature of the question we have decided to use linear regression as a main method to approach the data analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable. These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.

For question 1, we are trying to analyze how the buyers for the newer “Active” and “Advanced”

are different from the traditional customers. Therefore, our response variable is Release, which is the indication of whether the device is the traditional one or the new one (Active and Advance) indicated by 0 and 1 respectively. The most appropriate model to use was the general linear model, specifically logistic regression model since our response is a binary variable [4]. We have many predictor variables and to see which ones are significant with the response we run several models with different predictors. This process is called forward selection. Forward selection is a type of stepwise regression which begins with an empty model and adds in variables one by one. We have used the AIC value and the likelihood ratio test to determine the best model. For models with the same number of variables, we used AIC values to compare. After we choose the model with higher AIC, we used likelihood ratio tests to compare models with different numbers of variables. Likelihood ratio tests can help us determine whether or not adding complexity to our model makes our model significantly more accurate. The result indicated that only median income was significant with the response.

$$\log(\frac{\hat{\mathrm{P}}_{release}}{1-\hat{\mathrm{P}}_{release}}) = \beta_0 + \beta_1 x_{\text{median\_income}}$$

Where $\hat{\mathrm{P}}_{release}$ represents the proportion of product being "Active" or "Advanced", β_0 represents the intercept, and β_1 represents the slope of median income.

Method Description for Question 2

For question 2, we are trying to analyze what is affecting the poor performance of the device, especially sleep scores. Therefore, the response variable for this question is flag, which indicates the number of times there was a quality flag during the sleep session. For this question, we have decided to use a generalized linear mixed model because our response is count and we used poisson regression [4]. In order to determine which predictors are significant we modeled the response against the predictors and the result can tell us which predictors are fixed effects. This is because statistical significance indicates that changes in the independent variables correlate with shifts in the dependent variable. Therefore, our fixed effects are skin tone and current age. Given this information we use the generalized linear mixed model to assess which random effects are significant. The method to pick the best model follows the same procedure as question 1 where we add the random effects one by one. For models with the same number of variables, we used AIC values to compare. After we choose the model with higher AIC, we used likelihood ratio tests to compare models with different numbers of variables. Likelihood ratio tests can help us determine whether or not adding complexity to our model makes our model significantly more

accurate [1]. The result indicates that the model with skin tone and current age as fixed effect and gender as random effect resulted in the best model.

Model for 2:

$$
\begin{aligned}
Y_{\text{flags}\,i} &\sim \text{Poisson}(\lambda_i) \\
\log(\lambda_i) &= \alpha_{ji} + \beta_1 x_{\text{skin\_tone}_{\text{light}}} + \beta_2 x_{\text{skin\_tone}_{\text{medium}}} + \\
&\qquad \beta_3 x_{\text{skin\_tone}_{\text{medium-dark}}} + \beta_4 x_{\text{skin\_tone}_{\text{medium-light}}} + \beta_5 x_{Current\_age} \\
\alpha_j &\sim N\left(\mu_{\alpha_j}, \sigma^2_{\alpha_j}\right), \text{ for sex j} = 1, \ldots, \text{J}
\end{aligned}
$$

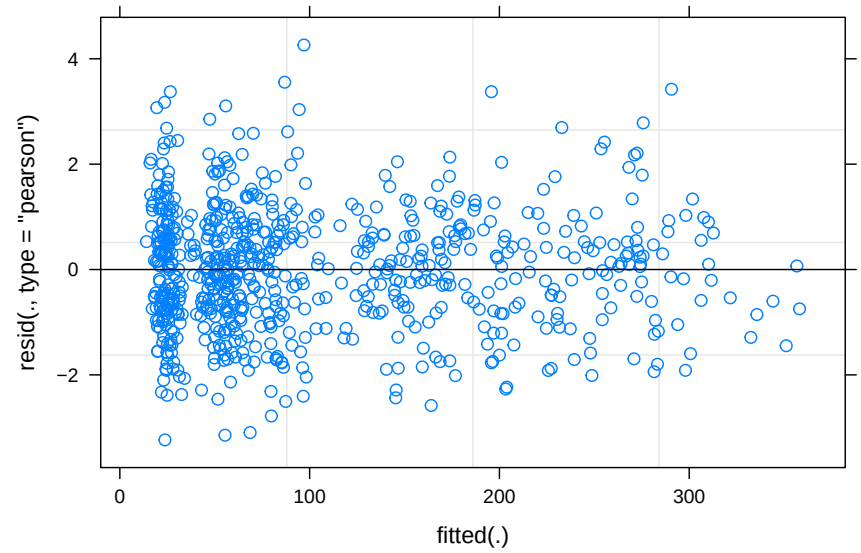
Where $Y_{\text{flags}\,i}$ represents the number of flags for individual i, α_{ji} represents the intercept of customer i at jth sex, $\beta_1 \ldots \beta_4$ represents the slope for skin tones of customer, and β_5 represents the slope for the current age of customer.

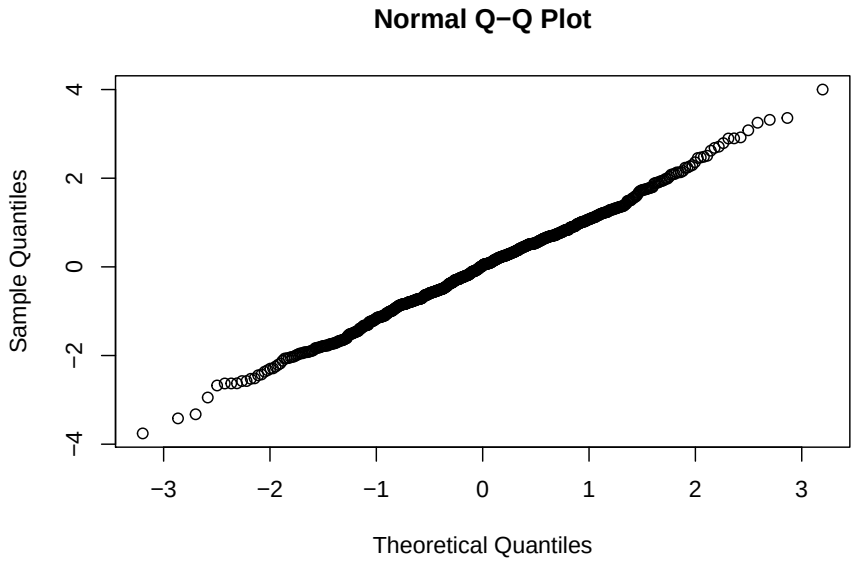
	(1)
(Intercept)	1.626 ***
	(0.085)
median_income	-0.000 ***
	(0.000)
N	14267
logLik	-9654.981
AIC	19313.961

*** p < 0.001; ** p < 0.01; * p < 0.05.

	(1)
(Intercept)	-5.772 ***
	(0.017)
skin_tonemedium-light	0.777 ***
	(0.019)
skin_tonemedium	1.179 ***
	(0.018)
skin_tonemedium-dark	1.891 ***
	(0.017)
skin_tonedark	2.390 ***
	(0.016)
scales::rescale(Current_age)	-0.049 **
	(0.016)
sd___(Intercept)	0.000
	(NA)
N	719
logLik	-2691.752
AIC	5397.503

*** p < 0.001; ** p < 0.01; * p < 0.05.





We checked the assumption for our model. In the residual versus fitted value plot, the residuals don't seem to have any pattern. The qqplot shows that the data seems to be normally distributed. Therefore, our assumptions are satisfied.

Result

Question 1

From figure 3, figure 4 and figure 5, we can see that people who have the new model have lower median income whereas people with the traditional model tend to have higher median income. The average median income of people who have the "Active and Advance" is 67,487.65 CAD whereas the average median income of people who have the traditional device is 73,550.55 CAD. Our final model for question 1 has a p-value of 2eˆ-16 for median income which suggests it has a significant impact. This clearly shows that the new customers have lower median income. The confidence interval of the median income at a 5% significance level is (-2.211634e-05, -1.752901e-05). This suggests that the true parameter value is inside this interval.

Question 2

From figure 6, we can observe that the darker the skin tone is the higher the number of flags there are. The final model for question 2 suggests that the p-values of the skin tones and age is smaller than 0.05 which means they are significant. From figure 7, we can see that younger people with darker skin tone tend to have a higher number of flags during sleep sessions. Also, as age increases for the medium-dark and dark people, the number of flags decreases. The true parameter values for skin tone and age are within the confidence interval at 5% significance level.

Conclusion

Our objective was to investigate the two requests from Mingar. For the first question, the results of the graphs and models suggest that the new customers buying the "Active and Advance" products have lower median income in contrast to the traditional customers. As for the second question, our model shows that devices of customers with darker skin have more flags during their sleep session. We would suggest the product development team of Mingar to check the censor of their devices as it may be the cause as to why devices of customers with darker skins have poor sleep quality.

Limitations

When working on these questions, we have encountered a few limitations. One of them being when comparing income for each line of products, we used the median income of each customer's area code. Since we don't have the income information of every customer, this is the best way for us to get an idea of how income level would represent each product line. This however may not showcase the correct income level of the customers. For example, it is possible that a group of users live in an area code that contains a small group of extremely wealthy people who use the traditional products, while the rest are around the city median income. This would cause the median income of that area code to simply be around the city median, which would not correctly represent the group of customers.

One other limitation we faced was we don't have a more complete data on the customers. Which caused us to not be able to narrow the customers of each type of product line any further. If we were given a more complete data on the users such as BMI, we could describe a more specific image of customers using each type of line.

Another concern we faced was that we don't have data regarding each customer's race. When brainstorming for ideas on how to approach question 2, we decided the colour of emoji used by the customer would represent their race as this is the best way to relate the data to race when we're not given data regarding each customer's racial identity. The problem with this method is that this may not correctly represent the customer's actual race as there may be some customer who doesn't change their emoji colour at all and uses the default colour. There may also be some customers who would deliberately change to another colour that doesn't reflect their race.

Consultant information

Consultant profiles

Chun Ki Yip. I am one of the co-founders of CJYB Consulting Group Inc. I have more than 20 years of industry experience and I spend my spare time reading about statistics to improve my craft.

James Bai. I am one of the co-founders of CJYB Consulting Group Inc. I have more than 20 years of industry experience. I like to play basketball and badminton in my free time.

Tsogjavkhlant Chuluunzorigt . I am one of the co-founders of CJYB Consulting Group Inc. I have more than 20 years of industry experience. I have worked on many projects.

Yunqi Jia. I am one of the co-founders of CJYB Consulting Group Inc. I have more than 20 years of industry experience. I have worked on many projects.

Code of ethical conduct

Responsibility to Client

- We will not disclose any confidential data provided by the client.
- We will not mislead the client with skewed analysis and data graphics.
- We will declare any conflict of interest to my clients.

Responsibility as a Statistician

- We will abide by the rules of ethical scrapping.
- We will give credit where credit is due.
- We will use the most appropriate statistical method to the best of our knowledge.

Professionalism

- We take responsibility for our work and encourage review and critics to assess our work.
- We will only provide services in areas which we are knowledgeable at.
- We will not misrepresent our work for any interest.

References

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. David Hugh-Jones (2021). *huxtable: Easily Create and Style Tables for LaTeX, HTML and Other Formats.* R package version 5.4.0. https://hughjonesd.github.io/huxtable/
3. Dmytro Perepolkin (2019). *polite: Be Nice on the Web.* R package version 0.1.1. https://github.com/dmi3kno/polite
4. 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.
5. *Full emoji modifier sequences,* V14.0 - unicode. (n.d.)., from https://www.unicode.org/emoji/charts/full-emoji-modifiers.html
6. Hadley Wickham (2021). *rvest: Easily Harvest (Scrape) Web Pages.* https://rvest.tidyverse.org/,https://github.com/tidyverse/rvest.
7. *Postal code conversion file.* Postal code conversion file | Map and Data Library. (n.d.)., from https://mdl.library.utoronto.ca/collections/numeric-data/census-canada/postal-code-conversion-file
8. von Bergmann J, Shkolnik D, Jacobs A (2021).* cancensus: R package to access, retrieve, and work with Canadian Census data and geography.* R package version 0.4.2, https://mountainmath.github.io/cancensus/.

Appendix

The web scraping statement was checked for CensusMapper [8], the statement indicates that web scraping is allowed as long as delay of web scraping is set to 5 seconds which was followed carefully [3]. In order to web scrape on CensusMapper, an account was made to obtain an API key and the API cancensus is used. The population, median household income and CSduid were scraped from the website to create a table. The postal code conversion file was accessed in the University of Toronto database. This conversion file is created based on data licensed by Canada Post Corporation and provided by Statistics Canada, the uses of this conversion data file is limited to teaching and learning purposes, research purposes, and planning purposes of University of Toronto faculty and students [7]. This was considered but not followed because our consulting work is ultimately a private commercial project, but we would release a version of our consulting report with simulated customer data which doesn't break our NDA with our client's data for teaching purposes.

Code appendix

```
## <polite session> https://censusmapper.ca/
##     User-agent: polite R package - https://github.com/dmi3kno/polite
##     robots.txt: 4 rules are defined for 1 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```