Marketing and Product Considerations for MINGAR

MINGAR's new product lines 'Active' and 'Advance' appeal to consumers outside of the traditional age group and higher income base. However, green sensors could result in poorer sleep scores for user's with darker skin tones.

Report prepared for MINGAR by The Outliers

2022-04-07

Contents

Executive summary

The current report provides key insights for marketing and product considerations for MINGAR's new product lines, 'Active' and 'Advance'. In order to compete with Bitfit, MINGAR's new products appeal to consumers with lower incomes outside of their traditional age bracket. This study shows that the new product lines have expanded MINGAR's consumer base to include these *new* consumers. Since these two lines have attracted customers outside of MINGAR's traditionally higher income base, we recommend that continued marketing should be tailored to less wealthy individuals interested in a healthy lifestyle.

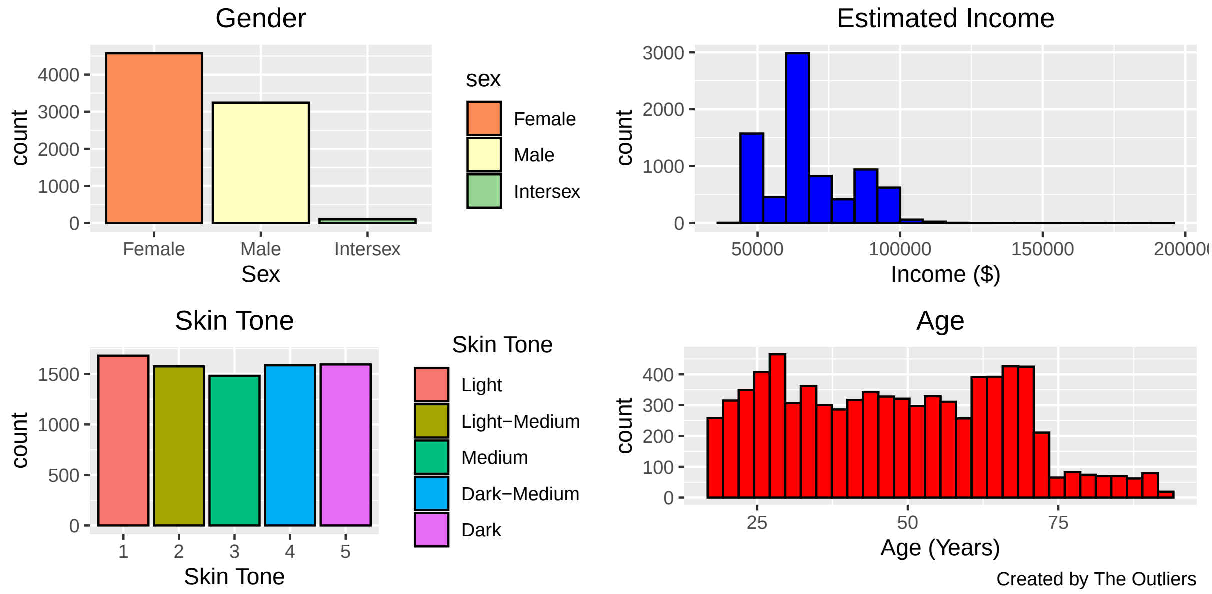
As for product considerations, this report also investigates the claim that MINGAR's products are performing poorly for users with darker skin; particularly, with respect to sleep scores. We estimate that there is evidence to support this claim as individuals with a darker skin tone reported, on average, more complaints than individuals with lighter skin tones. It is suggested to review the green sensors that are currently equipped on the devices, as these have demonstrated poorer performance on darker skin due to melanin. Key results are summarized below.

- The new customers were primarily female; however, the new product lines have more female, male, and intersex customers than MINGAR's traditional products.
- The majority of new consumers have an estimated income between $40,000 and $80,000, with an average of $68,835. This compares to the majority of traditional consumers who have an estimated income between $50,000 and $100,000, with an average of $73,100.
- MINGAR's new product lines have **attracted customers outside of MINGAR's traditionally higher income base.**
- The new products have captured more consumers under the age of 30, as well as consumers over 60. This may be due to the lower price that attracts younger individuals and seniors.
- Newer consumers are more likely to be even in numbers by skin tone category compared to users of MINGAR's traditional product lines, who saw a vast majority of individuals identify their Emojis with light skin tones.
- **MINGAR's new product lines attract, on average, consumers with lower incomes than their traditional consumers. Even though the new lines have increased the number of consumers under the age of 30 and above 60, the average consumer age between new and traditional products remains approximately the same.**
- Consumers with darker skin tones reported, on average, more complaints in their sleep scores than consumers with lighter skin. This is assuming the Emoji modifier skin tone acccurately predicts a user's skin tone.
- Dark skinned consumers reported, on average, 12 flags per sleep session. This compares to light skinned consumers, who only reported on average 1 flag per sleep session.

MINGAR's new product lines have, on average, attracted consumers from lower incomes outside of MINGAR's traditional age bracket. However, there does appear to be a link between darker skin tones and complaints per sleeping session.

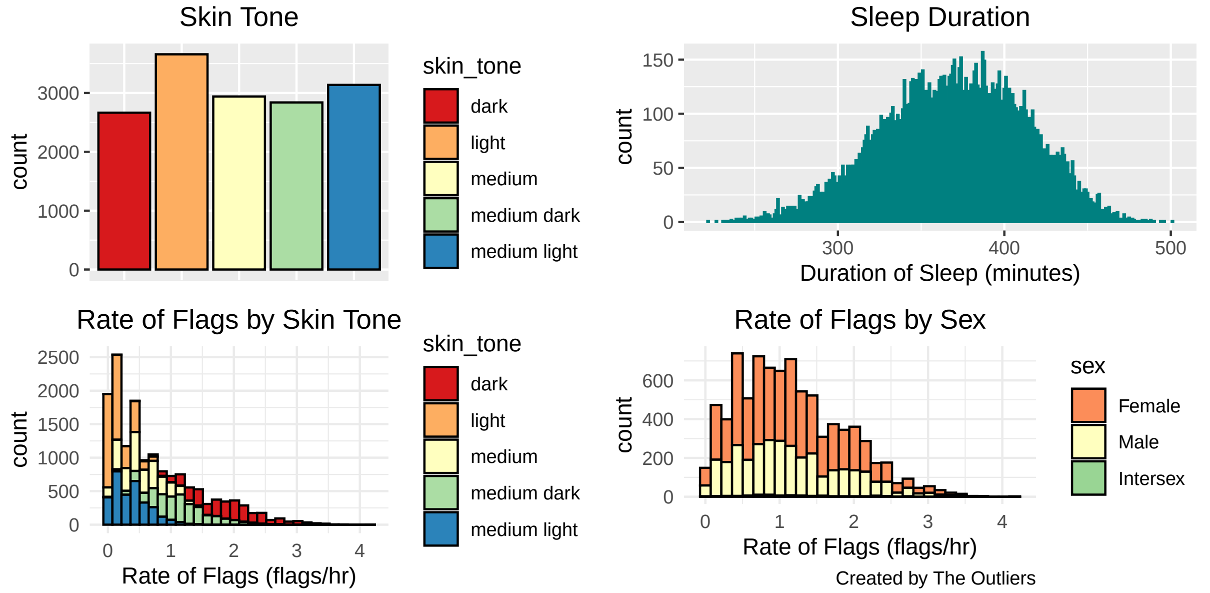
Key results of the study are summarized in the following visualizations.

Figure 1: Summary of MINGAR's New Customers ('Active' and 'Advance' Product Lines)



Created by The Outliers

Figure 2: Exploratory Data of MINGAR's Customer Base (Old and New)



Created by The Outliers

Technical report

Introduction

This report was created to investigate marketing and product considerations for MINGAR's new product lines 'Active' and 'Advance', as well as investigating complaints that their products produce more errors for consumers with darker skin tones. As a whole, the report should provide guidance for MINGAR's business executives on how to capture a larger market share in the growing Wearables market by highlighting MINGAR's new consumer base and product issues.

MINGAR provided The Outliers with customer-level data (age, gender, etc.), customer-device linkage data, and device data, which is an industry updated resource. As a note, the Outliers want to reiterate their and MINGAR's shared commitment to the ethical treatment of customer data. This is taken into consideration throughout the report, and any assumptions or citations are explicitly made clear to the reader. See *Code of ethical conduct* for more information.

This report is primarily focused on investigating two questions relating to MINGAR's consumer base and their products' efficacy. The first section will examine the consumers that compose MINGAR's new consumer base after the launch of their 'Active' and 'Advance' lines. This is intended to provide MINGAR's marketing team with information to inform their strategy in the Canadian markets. Specifically, the considerations of MINGAR's new consumer base also highlights their capture of a larger share of the wearables market; effectively winning over part of Bitfit's less wealthy consumer base. The second section will investigate a trend in complaints that MINGAR's devices are performing poorly for users with darker skin in respect to sleep scores. MINGAR's social media team has been tracking this trend, and this report suggests that business executives should review the green sensors in their devices which has been linked to poorer vital readings for user's with darker skin.

Research questions

This report will primarily answer two research questions.

1. Who are MINGAR's new consumers and how do they differ from the traditional consumer base in regard to age, gender, and income?
2. Is there a statistically significant relationship between the color of a user's skin and the number of complaints/flags they report per sleep session?

The first question will be valuable to MINGAR's marketing team in order to inform their marketing strategy. The second question will be of interest to business executives in planning device development as it will clarify whether there is significant evidence to support the claim that the device performs poorly for users with darker skin.

MINGAR's New Consumer Base and Marketing Considerations

Data Manipulation and Aggregation

In order to first examine MINGAR's new consumer base, we created an aggregated dataset that included our variables of interest. We loaded the basic customer data containing 6 variables (customer ID, date of birth, postcode, sex, pronouns, and Emoji modifiers) and omitted any null observations. Because MINGAR's marketing team is primarily interested in understanding whether their new product lines have attracted customers outside of their traditionally higher income base, we obtained median income data in each neighborhood in Canada. We then compiled the income data together with our customer data on postcode, assuming that each customer's income is the same as their neighborhood's median income. We thereafter calculated the age of the customers from their date of birth, and renamed the Emoji modifiers to represent skin color on a discrete scale: 1 being lightest and 5 being darkest. For easier data handing and manipulation, we created a factor gender variable: 0 being female, 1 being male, and 2 being intersex.

In order to track who MINGAR's *new* consumer base is, we added a variable for device ID which specified whether the consumer was a user of the traditional product lines, or the newer 'Active' and 'Advance' lines. A binary factor variable was then created to determine whether a consumer was *new* or not. Finally, we extracted a subset of our dataset that only included new consumers in order to conduct our exploratory data analysis.

Exploratory Data Analysis

Below is a summary table for the average consumer of MINGAR's new product lines.

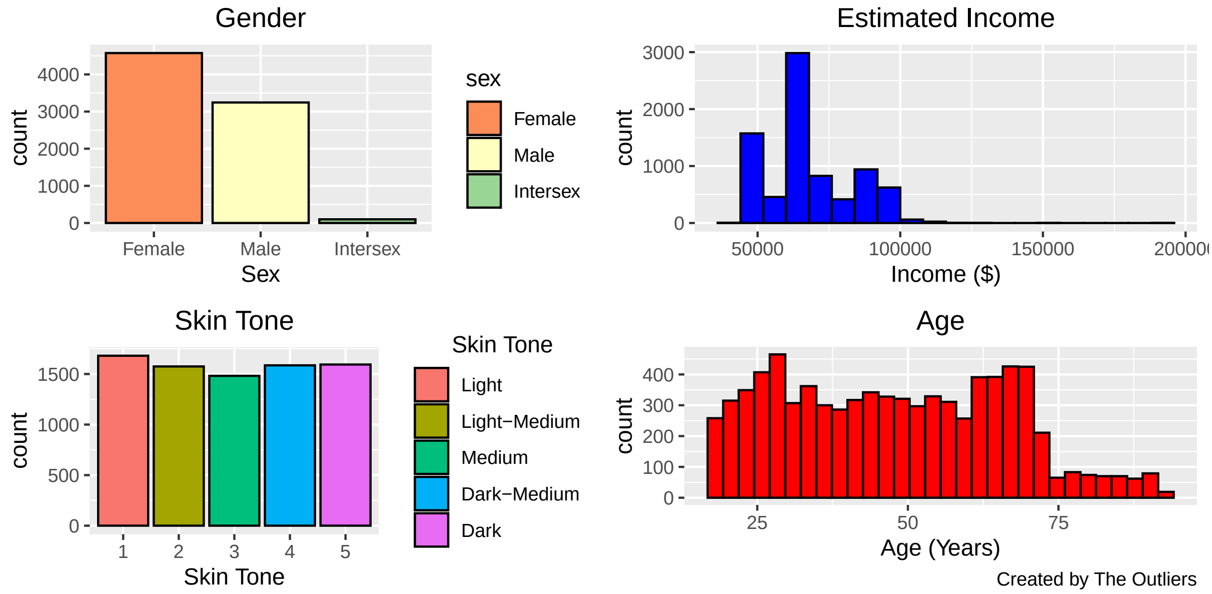
Table 1: Average New User

Variable	Average Statistic
Sex	Female
Age	48
Income	68835
Device	Advance 2

Variable	Average Statistic
Skin Tone	Light (1)

The above information is useful in determining who the average new MINGAR user is. The data shows that the average new user is a white, 48 year-old female with an average income of approximately $69,000. It is also notable that the Advance 2 device is the most popular device. Despite this information's usefulness in determining the average user, MINGAR's marketing team should be more interested in viewing the distribution of these variables in order to inform their marketing strategy in Canada. Therefore, the below visualizations represent the distribution of gender, income, age, and skin tones in MINGAR's new consumer base.

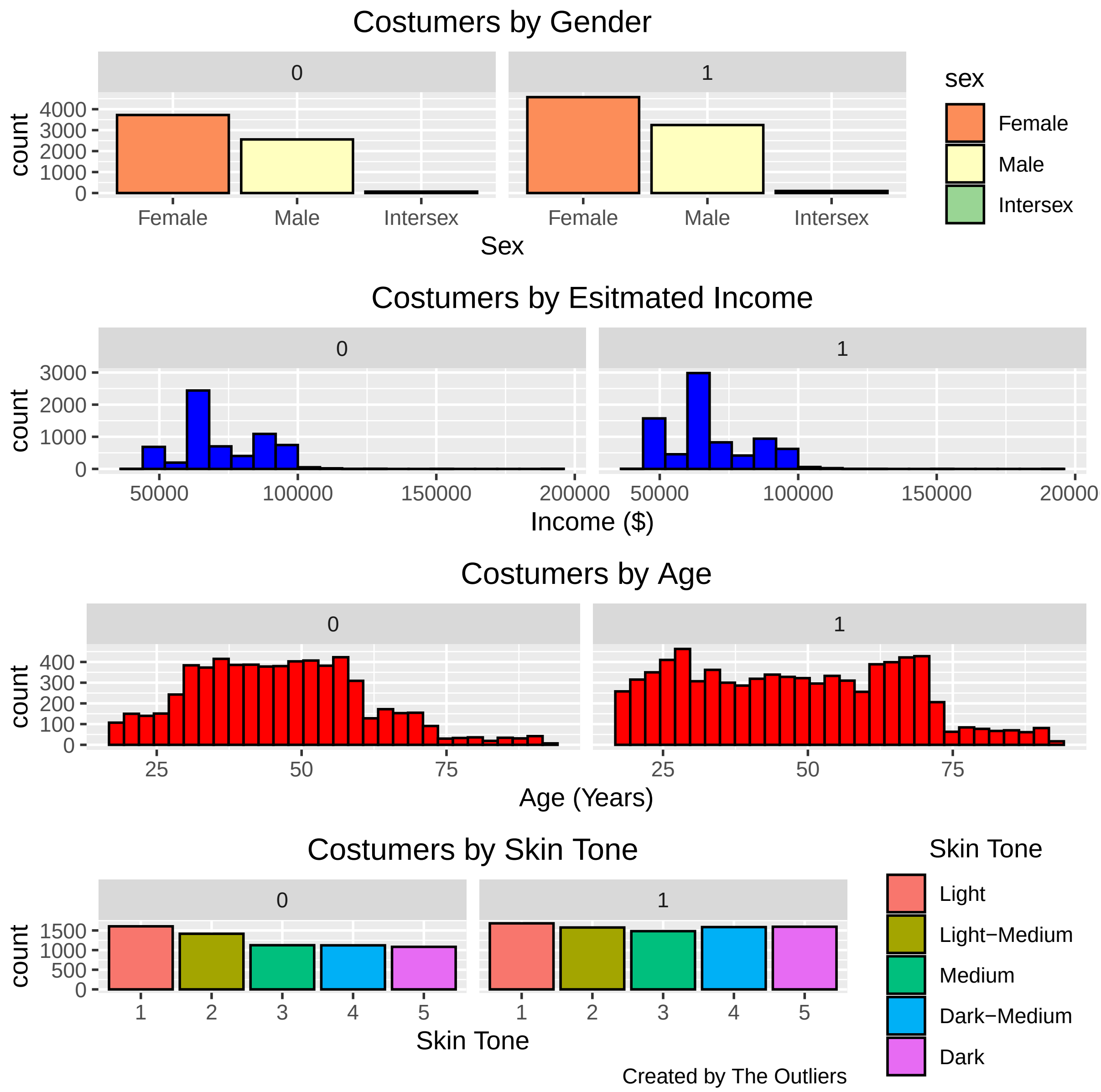
Figure 3: Summary Visualizations of New Customer Base



As is evident from the visualizations above, in the consumer base for newer 'Advance' and 'Active' device users, there are more female compare to male customers. The vast majority of costumers also appear to have an estimated income ranging from $40,000 to $80,000 per year. It is noticeable that the mean income across all new consumers is $68,835, whereas the median income across all new consumers is $65,829. This highlights that the majority of new users belong to a lower income bracket, emphasizing that the more affordable product lines have managed to capture less wealthy individuals - as is also shown by the distribution in the histogram. As for age, most newer customers belong in the 18-75 year age group, with a very small proportion of customers being older than 75. It is interesting to note the cluster in age distribution for ages below 30 years and above 60. This highlights how MINGAR's new product lines have attracted both younger and older users outside of their traditional age bracket. Finally, MINGAR's new consumer base is uniformly distributed on skin tone. We used the user's Emoji modifiers to indicate their skin tone. This assumption was necessary because MINGAR did not track information on user's skin color.

In order to provide MINGAR's marketing team with as much useful information as possible, we now compare their new consumer base with their traditional one across the same variables. This provides interesting results which indicate how the 'Active' and 'Advance' lines have managed to attract both younger and older users more interested in affordability. This also gives MINGAR an insight into how they have managed to compete with Bitfit, which boasts devices at a lower price point. As is evident from the visualizations below, MINGAR's new product lines have expanded their consumer base to include both younger and older, less wealthy individuals compared to their traditional consumer base which has increased MINGAR's competitiveness.

Figure 4: Comparison Visualizations of Old (0) and New (1) Consumer Bases



Costumers by Gender

Costumers by Esitmated Income

Costumers by Age

Costumers by Skin Tone

Created by The Outliers

The above visualization provides useful information for the comparison between MINGAR's old and new consumer base. This is particuarly important in order for the marketing team to determine which individuals their 'Active' and 'Advance' product lines are most appealing to. It appears that the majority of consumers are female in both cases (old consumer base and new consumer base). However, it is visible that the new product lines have in total more users which could be because of its affordability. Further, though the income range is fairly similar across both groups, it is notable that there are more users with an income below $50,000 in the new group compared to MINGAR's traditional consumer base. This is interesting as it directly relates to our research question. The age distribution in each group also shows that the new product lines have attracted younger users, particularly under the age of 30. However, there is also a significant increase in the number of users above 60. A reason for this could be because of the product's increased affordability and simplicity compared to MINGAR's more expensive and complex products. Finally, whereas the traditional MINGAR products were predominantly owned by users with lighter skin colors, the new group seems to exhibit a uniform distribution in skin tones.

Modeling the relationship between user characterstics and MINGAR's new consumer base

To truly understand how the customers of the new 'Active' and 'Advance' devices differ from the traditional users, we propose a question: Given a person is a customer, what are the odds that he/she is using a device from the newer line? Also, how does the odds change if he/she have a specific set of traits, such as a higher income or darker skin tone?

To answer these questions and provide the marketing team with the best results, we construct a model with the generation of device they are using as a response. For the response, 1 means the costumer uses either an 'Active' or 'Advance' device. 0 means he/she is using an older generation, such as the Run or iDOL device.

Each observation should be independent in theory. There might be some special cases where someone recommends a device to another customer, but in general we assume that the observations are independent. Further, the response follows a binomial distribution, as explained above.

Table 2: Modeling New Consumers on User Base Characteristics

Predictors	Coefficients	Std. Error	P-value	Odds Ratio	95% C.I. Odds Ratio
(Intercept)	1.286	1.163e-01	< 2e-16 ***	3.618	2.881 - 4.546
sexMale	4.147e-02	3.475e-02	0.234	1.042	0.974 - 1.116
sexIntersex	1.704e-01	1.607e-01	0.289	1.186	0.867 - 1.631
income	-1.928e-05	1.285e-06	< 2e-16 ***	0.999	0.999 - 1.000
age	5.082e-03	1.010e-03	4.91e-07 ***	1.005	1.003 - 1.007
light-medium	4.794e-02	5.113e-02	0.348	1.049	0.949 - 1.160

Predictors	Coefficients	Std. Error	P-value	Odds Ratio	95% C.I. Odds Ratio
medium	6.652e-02	5.435e-02	0.221	1.069	0.961 - 1.189
dark-medium	3.794e-02	5.557e-2	0.495	1.039	0.931 - 1.158
dark	8.092e-02	5.579e-02	0.147	1.084	0.972 - 1.210

The response of the model exponentiated is the log odds of a person owning a newer device (‘Active’ and ‘Advance’) compared to an older one (Run and iDOL), given that he or she is a customer of the company. From the summary of the model, we can see that we have very strong evidences against the null hypothesis that a change in a person’s age or income does not affect the odds of he/she owning a new device, assuming all other factors remain constant. However, we have no evidence against the null hypothesis for gender or skin tones. Therefore, there is no evidence that gender or skin tone affect the odds of a user owning a new MINGAR device, assuming all other variables are constant.

In order to interpret the model and coefficients as log odds, since it is a logistic regression, we first need to exponentiate our coefficient estimates. Using income as an example, suppose the gender, age and the skin tone of a person remains constant, the model suggest a one dollar increase in yearly income will, on average, decrease the odds of he/she owning a new MINGAR device by a factor of 0.9999807. Because log odds are inherently relative, this also means that an increase in income by $10,000 dollar will decrease the odds by a factor of $0.9999807^{10000} = 0.8244804$. The base gender is female, and the base skin tone is light (whitest).

Looking at the confidence intervals, using income as an example again, we observe that we are 95% confident that a $1 increase in a person’s yearly estimated income will cause the odds of he/she owning a new device (‘Active’ and ‘Advance’) compared to an old device decrease by a factor between 0.9999782 and 0.9999832.

In short, a customer with a lower income is more likely to buy a device from the ‘Active’ or ‘Advance’ product lines than to buy a Run or iDOL one. Further, an older customer is more likely to buy a device from the ‘Active’ or ‘Advance’ product lines, compared to Run or iDOL.

Investigating Poor App Performance for Users with Darker Skin Tone

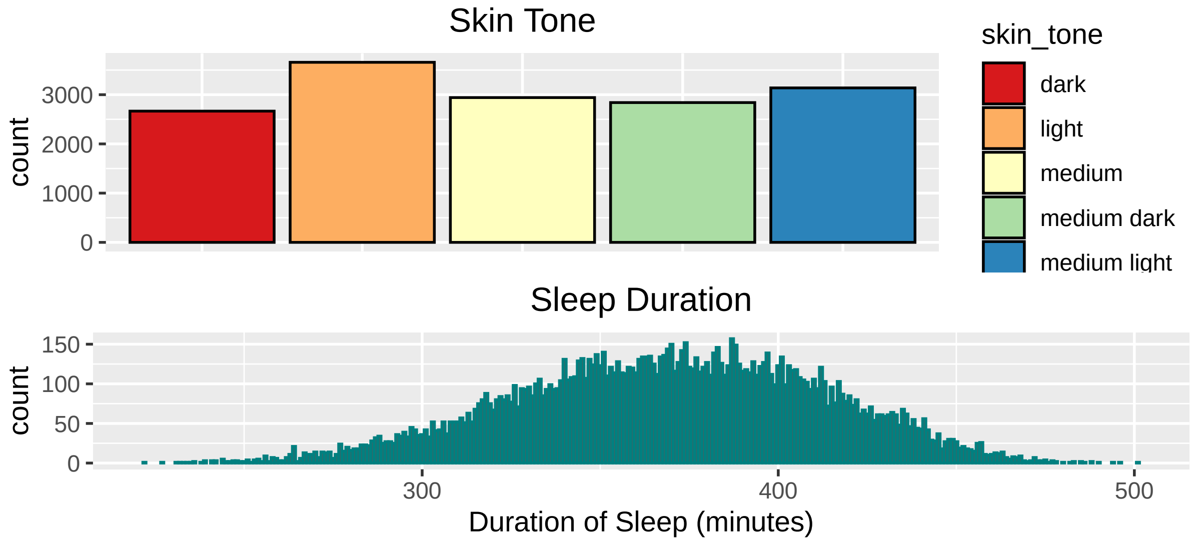
Data Manipulation and Aggregation

To investigate the claim that MINGAR’s devices are performing poorly for users with darker skin in respect to sleep scores, we merged customer data with sleep data to model app performance. The sleep and device data was taken from the Fitness Tracker Info Hub ("Data tracking MINGAR and Bitfit wearable fitness trackers", 2022). We removed the observations with null values for Emoji modifiers & gender to focus on these two variables. We added a variable flag_rate which represents the amounts of flags/complaints per hour during a user's sleep session. We, again, included a factor variable for gender, calculated age from date of birth, and created a skin tone variable that generalizes the emoji modifiers into 5 different categories: light, light-medium, medium, dark-medium, and dark skin tone. This information on emoji modifiers was taken from the Unicode CLDR Project ("Full Emoji Modifier Sequences, v14.0", 2022). Age was then factorized into three different categories, youth, adulthood, and seniors, in accordance with the Government of Canada's life cycle groupings ("Age Categories, Life Cycle Groupings", 2017).

This study is based on a few assumptions. Firstly, to investigate whether the app performs poorly for darker skinned users, we assume that 'darker skin' includes dark, dark-medium, and medium skin tones. Secondly, one key assumption is that the user's emoji modifier is an accurate representation of their skin color. Finally, we assume that sleep scores are determined by the number of flags, where more flags represent worse scores, with respect to the duration of sleep.

Exploratory Data Analysis

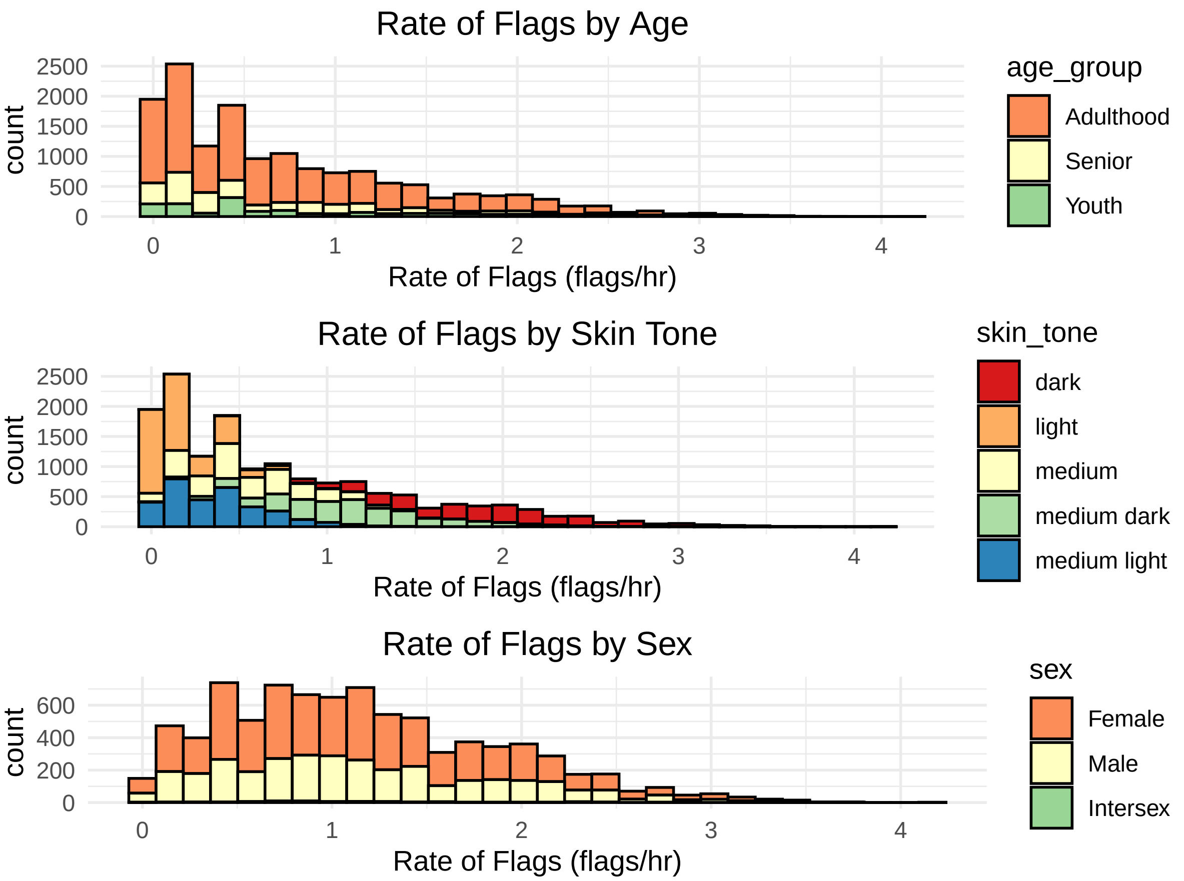
Figure 6: Distribution of Skin Tone and Sleep Duration



The above visualisation presents some summary distributions for various variables. Particularly, note the distribution of MINGAR consumers by their skin-tone. We can observe that the distribution of skin tones is approximately uniform, with a few more consumers falling under the light category. We did this to check if there was an uneven distribution of customers by skin tones which could skew our model results. This does not appear to be the case. The distribution of sleep duration appears to follow the normal with an average sleep duration of 369.3 minutes (6.2 hours).

Further, examining the distribution of flag rates by skin tone and gender can provide important information to MINGAR's executive team in order to determine whether their app is performing poorly for some users with specific characteristics.

Figure 7: Distribution of Flag Rates by Skin Tone, Gender, and Age Group



Created by The Outliers

From rate of flags by age, we note that the adulthood level has the most individuals, understandably as it has the most entries among the three groups. Based on this histogram, it does appear that age has an impact on the rate of flags as this graph's shape is identical to rate of flags by skin tone. Since we hypothesize that skin tone has a relationship with flag rate, and the two graphs observe the same distribution, we also hypothesize that the relationship of flag rates and age is statistically significant.

From the rate of flags by skin tone, we can observe that there does seem to be a correlation between skin tone and rate of flags. From this, it appears that most dark skinned people are getting at least 1 flag / hr when they sleep. This is concerning because lighter skin people don't have this issue. This could hint at a potential product issue, a trend which MINGAR's social media team have been tracking. From the graph by gender, we note that the rate of flags is more prevalent in females than males. This is expected because there are more females than males in MINGAR's consumer base, however, the effect size seems large. More analysis in the subsequent subsection should clarify this relationship.

Modeling Sleep Score by Skin Tone, Gender, and Age Group

We will use a generalized linear mixed model (GLMM) in order to estimate the relationship between flags and skin tone, gender, and age group. The assumptions for this model as satisfied. We assume that cases are independently distributed. This is reasonable as the sleep score of one individual is unlikely to affect that of another, assuming those individuals do not sleep in the same location. Also, from the visualizations above, it is clear that the dependent variable follows a distribution from an exponential family.

We conduct the GLMM to see the impact of skin tone on the frequency of flags with respect to customer ID (the individual user). There is a non-independence issue with the observations because there are multiple observations from the same user. However, adding customer ID as a random intercept solves the non-independence issue in the data that comes from having multiple responses by the same subject. Using the offset function on duration accounts for the fact that duration influences the frequency of flags. The base gender is female and base skin tone is dark.

Table 3a: Modeling Flag Rate on Skin Tone and Gender, controlling for individual users.

Predictors	Coefficients	Std. Error	P-value	Odds Ratio	95% C.I. Odds Ratio
(Intercept)	-3.399828	0.007106	< 2e-16 ***	0.033	0.0329 - 0.0338
light	-2.391344	0.017345	< 2e-16 ***	0.092	0.0884 - 0.0947
light-medium	-1.614883	0.013904	< 2e-16 ***	0.120	0.194 - 0.204
medium	-1.213693	0.012654	< 2e-16 ***	0.297	0.290 - 0.304

Predictors	Coefficients	Std. Error	P-value	Odds Ratio	95% C.I. Odds Ratio
dark-medium	-0.501792	0.010723	< 2e-16 ***	0.605	0.593 - 0.618
sexMale	-0.001762	0.008772	0.841	0.998	0.981 - 1.016
sexIntersex	-0.060557	0.043509	0.164	0.941	0.863 - 1.024

Table 3b: Modeling Flag Rate on Skin Tone and Age, controlling for individual users.

Predictors	Coefficients	Std. Error	P-value	Odds Ratio	95% C.I. Odds Ratio
(Intercept)	-3.382801	0.009449	< 2e-16 ***	0.0340	0.0333 - 0.0346
light	-2.390008	0.017296	< 2e-16 ***	0.0916	0.0886 - 0.0948
light-medium	-1.613542	0.013838	< 2e-16 ***	0.199	0.194 - 0.205
medium	-1.211918	0.012587	< 2e-16 ***	0.298	0.290 - 0.305
dark-medium	-0.499291	0.010663	< 2e-16 ***	0.607	0.594 - 0.620
age	-0.049703	0.018573	0.00745 **	0.952	0.917 - 0.987

We fit two regression models in order to determine the relationship between flag rate and skin tone, gender, and age. Model A, represented in **Table 3a**, is a GLMM with flag rate as the response and skin tone and gender as the model parameters. Model B, shown in **Table 3b**, is a GLMM with flag rate as the dependent variables and skin tone and age group as our variables of interest. Model C is a GLMM with flag rate as the response and only skin tone as our paremeter. We added customer ID as a random intercept in all models in order to control for non-independence issues.

From both the models A and B, we note that the level 'dark' is used as the base factor for the other levels of the fixed effect skin tone. We can observe that, based on the estimates, a change in skin tone from 'dark' to a lighter skin tone leads to a lower frequency of flags per sleep session. This effect was found to be statistically significant in both models.

Further, for both models, the low p-values indicate that there is strong evidence against the null hypothesis that changes in user skin tone will not result in any change in the frequency of flags, keeping all other factors constant. However, in Model A, we observe that there is no evidence against the null hypothesis that a change in sex leads to no change in the frequency of flags, while skin tone remains constant. Based on this, we can omit the variables sex as they don't seem to influence the rate of flags during sleep sessions. Model B shows that all variables, including age, are statistically significant. Because the AIC and BIC values for models B and C were sufficiently similar to one another, this justifies the inclusion of the age factor: see **Table 4**. Therefore, we are left with our final model B that only includes fixed effects for varying levels of skin tone, age, and the random intercept for customer ID.

Table 4: Comparison of AIC and BIC in Models A, B, and C.

Models	AIC	BIC	logLik	deviance
Model A (skin tone and gender)	64,978.6	65,032.1	-32,482.3	64,964.6
Model B (skin tone and age)	64,985.8	65,046.8	-32,484.9	64,969.8
Model C (skin tone only)	64,983.8	65,029.5	-32,485.9	64,971.8

The ideal model would be Model B as we have evidence against the null for all the effects. This is intuitive, as we have observed in Figure 7 (Rate of Flags by Sex) that the distribution of flags between males and females is roughly the same, as stated prior. Further, both the AIC and BIC are sufficiently low in all models that justify choosing the most statistically significant model. Therefore, this supports the choice of Model B, which includes the effects of skin tone, age, and a random intercept for customer ID.

For our chosen Model B, we can see from the confidence intervals that the log odds ratio is negative for the 'darker' skin tones. The intervals are less negative relative to the reference level compared to the light skin tones which further indicates that as the skin tone gets lighter, the interval that surrounds the true value of each level will be more negative relative to the reference level. In short, this suggests that an individual who has a darker skin tone will be more likely to report a flag compared to a person with lighter skin.

Discussion

This report has shown that MINGAR's new product lines, 'Active' and 'Advance, have attracted customers outside of MINGAR's traditional age bracket and higher income base. There was statistically significant evidence against the null hypothesis that there exists no relationship between owning a new MINGAR device and income and age, holding all other factors constant. However, there was no evidence to suggest that gender or skin tone was related to the odds of owning a new MINGAR device.

This report has also found that a statistical relationship exists between darker skin tones and flag rates per sleeping session. Our results indicate that a user who uses a darker skin tone Emoji modifier will be more likely to report a flag in their sleep session compared to a person using a lighter skin. Possible reasons for this includes the new product watches being equipped with green lights, which are cheaper. According to Stat news, green light sensors don't work well with darker skin due to the high amount of melanin in the skin which blocks the green light, resulting in a poor reading of vitals (Hailu, 2019). It is therefore suggested in this report that MINGAR should review the sensors that are currently equipped on the watches in order to protect against 'racist' claims. The model also showed that a statistically significant relationship exists between the rate of flags and age groups. As the age group gets older, we observe, on average, a lower amount of flags per sleeping session.

Strengths and limitations

This report has multiple strengths and limitations to consider. Firstly, the data visualizations and logistic regression models are constructed using a dataset of approximately 17,500 observations (range is between 14,267 and 20,622). The plethora of observations means that that model was able to construct a more accurate relationship of our variables of interest and response. Further, data cleaning and manipulation has made our report easily digestible and reproducible for any interested readers. Each model was properly calibrated with appropriate assumptions, and the distribution of our response variable and model specifications were relevant. Finally, our strict adherence to ethical statistical principles and data handling attests to this reports appropriateness.

There does, however, exist multiple limitations. In cleaning our data, we decided omitted all rows with null values for our variables of interst. We noticed that there are about 5,379 rows with null valies, about 1/4 of the observations. Since most of these are in the emoji_modifier (skin tone) columns, we could have kept these observations for first model, but to stay consistent removed them for all the analysis. Also, The income of each customer is estimated using the median income of their neighbourhood (on postal codes) Due to this, there is a possibility that the estimations are inaccurate. However, we tried to remediate this by obtaining the best possible information we could gather. Further, each individual observation (customer) may not be independent from other observations. For example, it is possible that a person recommended his friend to get a certain device. Therefore, for simplicity, we assumed every observation is independent. Another limitation stemms from a mismatch between sex and pronouns. Finally, since we are unaware of the accuracy of Emoji modifiers to reflect true skin tone, there may be response bias present in the data.

Consultant information

Consultant profiles

David Lundgren. David is a senior business analyst at The Outliers, working specifically with data analysis and visualizations. He has worked in many industries, however, specializes in the technology and wearables industry. David earned his Honors Bachelor of Arts and Sciences in Economics and Statistics with a Focus in Data Analysis from the University of Toronto in 2023. He went on to complete a Masters in Law and Finance at the University of Oxford in 2025.

Yugong Zhang. Yugong is a mathematic and statistic consultant at The Outliers. He likes to work with complex problems and are profitent in statistical analysis and programming. He earned his Bachelor of Science, Specialist in Mathematics and Statistics from the University of Toronto in 2023.

Muhammad Abdul Mannan. Abdul is a junior data science consultant at The Outliers. He specializes in data manipulation and analysis. He has earned his bachelors of Science, majoring in Computer Science and Statistics from the University of Toronto in 2022. He is also proficient with software & web development. Currently, Abdul is actively looking for a software developer position in Toronto.

Muhammet Rodin Karadeniz. Rodin is currently a lead AI researcher in the Outliers. He specializes in Machine Learning algorithms. He earned his Bachelor of Science, majoring in Computer Science and Statistics from the University of Toronto in 2016.

Code of ethical conduct

Ethical statistical consulting is a crucial aspect of our vision and the following are key guidelines that we strive to follow:

1. Conduct analysis and use methodology that does not result in prejudice or discrimination against anyone. Respect all individuals and condone any acts of harassment or abuse of any power.
2. Formulate a reasonable timeline and deadlines that are also reviewed and agreed upon by the client. Inform the client if there may be any issues present in the dataset that violate ethical boundaries of data collection.
3. All work done will be credited accordingly and any external work will be clearly cited and referenced in order to respect the intellectual property of other individuals.
4. Be open to critique and healthy feedback and all statistical practices and any work done should be transparent and open to being reviewed by other academics. Provide guidelines that promote reproducibility.

5. Making efforts to prevent the manipulation of data in order to influence the results of the analysis. Reject any external influence that desires a specific outcome from the study.
6. Be respectful and discrete when handling sensitive information regarding clients and data and protect this data per privacy laws.
7. Refrain from collecting data unnecessarily and only collect sensitive information with the consent of individuals.
8. Avoid using terminology or phrasing that may result in misinterpretation of results or which would result in controversy.
9. Inform the client regarding any legal limitations that may have been overlooked by client and potential impacts of the study or analysis on the population of interest.

References

Age Categories, Life Cycle Groupings. (2017). Retrieved 11 April 2022, from https://www.statcan.gc.ca/en/concepts/definitions/age2

Augile, Baptiste. (2017). “gridExtra: Miscellaneous Functions for”Grid” Graphics”. R package version 2.3.

Bates D, Maechler M, Bolker B, and Walker, S. (2015). “Fitting Linear Mixed-Effects Models Using lme4”. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Cooley, David. (2022). “geojsonsf: GeoJSON to Simple Feature Converter”. R package version 2.0.2. Retrieved 11 April 2022, from https://github.com/SymbolixAU/geojsonsf.

Data tracking MINGAR and Bitfit wearable fitness trackers. (2022). Retrieved 11 April 2022, from https://fitnesstrackerinfohub.netlify.app

Firke, Sam. (2021). “janitor: Simple Tools for Examining and Cleaning Dirty Data”. R package version 2.1.0. Retrieved 11 April 2022, from https://github.com/sfirke/janitor.

Full Emoji Modifier Sequences, v14.0. (2022). Retrieved 11 April 2022, from https://unicode.org/emoji/charts/full-emoji-modifiers.html

Grolemund, Garrett, and Wickham, Hadley. (2011). “Dates and Times Made Easy with lubridate”. Journal of Statistical Software. 40(3), pp. 1-25. Retrieved 11 April 2022, from https://www.jstatsoft.org/v40/i03/.

Hailu, R. (2019). Fitbits and other wearables may not accurately track heart rates in people of color. Retrieved 11 April 2022, from https://www.statnews.com/2019/07/24/fitbit-accuracy-dark-skin/

Pebesma, Edzer. (2018). “Simple Features for R: Standardized Support for Spatial Vector Data”. The R Journal. 10(1), pp. 439-446. doi:10.32614/RJ-2018-009. Retrieved April 11, 2022, from https://doi.org/10.32614/RJ-2018-009.

Perepolkin, Dmytro. (2019). “polite: Be Nice on the Web”. R package version 0.1.1. Retrieved 11 April 2022, from https://github.com/dmi3kno/polite.

von Bergmann J, Shkolnik D, and Jacobs A. (2021). “cancensus: R package to access, retrieve, and work with Canadian Census data and geography”. R package version 0.4.2. Retrieved 11 April 2022, from https://mountainmath.github.io/cancensus/.

Wickham, Hadley. (2016). “ggplot2: Elegant Graphics for Data Analysis”. Springer-Verlag New

York. ISBN: 978-3-319-24277-4. Retrieved 11 April 2022, from https://ggplot2.tidyverse.org.

Wickham, Hadley. (2021a). “rvest: Easily Harvest (Scrape) Web Pages”. Retrieved 11 April 2022, from https://rvest.tidyverse.org/, https://github.com/tidyverse/rvest.

Wickham H, Girlich M, Ruiz E. (2021). “dbplyr: A ‘dplyr’ Back End for Databases”. Retrieved from https://dbplyr.tidyverse.org/, https://github.com/tidyverse/dbplyr.

Wickham, Hadley, and Miller, Evan. (2021b). “haven: Import and Export ‘SPSS’, ‘Stata’ and ‘SAS’ files”. Retrieved 11 April 2022, from https://haven.tidyverse.org, https://github.com/tidyverse/haven, https://github.com/WizardMac/ReadStat.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). “Welcome to the tidyverse.” Journal of Open Source Software, 4(43), 1686. doi: 10.21105/joss.01686.

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

Appendix

This appendix includes the methods followed for data preparation to process, clean, and reshape raw data into the data to be used for statistical analysis. Note that the usages of these functions and code for reshaping the data are provided in data-prep.rmd file. This appendix summarizes the procedure.

Setup

Install the necessary packages and open them using the library() function. The libraries used are the following: tidyverse, janitor, polite, rvest, haven, cancensus, sf, geojsonsf, lubridate, gridExtra, dbplyr, lme4, and lmtest.

Client Data

Load the raw data and clean it by first reading in the data with read_rds() function (note that alternative functions also exist in the library for different file types, such as read_csv for .csv files). Then, clean the names using the janitor package. Mutate NA values in numeric columns to 0 with mutate_if() function, and filter rows with NA values if still present with filter() and is.na().

Web scraping industry data on fitness tracker devices

The industry data on devices is scraped and cleaned in the following steps. Use the bow() function to wrap the scraping information – URL is https://fitnesstrackerinfohub.netlify.app/ and set the user_agent to an informative string. Use scrape() function with the return on bow() and save the HTML output to a variable. Receive the table with using the html_elements("table") and html_table() functions respectively. Save retrieved data to the data_raw folder. Clean the table data the same way client data has been cleaned.

Accessing Census data on median household income

After signing up to https://censusmapper.ca/ and receiving the API key, the following steps get the census data and clean it. Set the API key and cache path for the data with options() function. Get the region data and filter it with using list_census_regions() with dataset "CA16", filter() with level "CSD", and as_census_region_list(). With filtered regions, get the median income data using census_data_csd(). Simplify the data to only needed variables using select(), mutate() and rename() functions. Save retrieved data to the data_raw folder. Clean the table data the same way client data has been cleaned.

Accessing postcode conversion files

If you have access to Census Canada Postal Code Conversion Files and accept the license agreement to receive the data, you can create the postcode conversion files. Download the .sav version of the file. Since the 2020 Census has not been released yet, we used the 2016 Census geography. Read the file into R using read_sav() function. Save retrieved data to the data_raw folder. Clean the table data the same way client data has been cleaned.

Additional Mutations for Question 1

Create a table customer_income, assigning customers with the median incomes in their regions. Process customer_income to customer_info, calculate the age for each customer using their date of birth and add to that table. Introduce skin tone for each customer and add to the dataset. Factor the gender variable into “female”, “male”, and “intersex”. Introduce the device each customer uses to the dataset. Introduce a categorical variable on whether the customer is “new” or “old”. Save new customers into dataset new_cust.

Additional Mutations for Question 2

Merge customer and sleep data. Remove any newly created NA values. Add rate of flags to data. Add a variable that generalizes the emoji skin according to skin tone. Calculate the age for customers using their date of birth, and add the age data to the table. According to Statistics Canada, categorize ages into groups “youth”, “adulthood”, “senior”, and “empty”. Factor the gender variable into “female”, “male”, and “intersex”.

Writing Data

Write all the cleaned and mutated data into the data folder using write_rds() function. Note that just like reading the data, there exist multiple methods for multiple file types.