Statistical Analysis of Mingar Customer Base

Analysing characteristics to better understand customers of new line of products

Report prepared for MINGAR by R3M

2022-04-10

Contents

Executive summary

Background & aim

Since its inception, Mingar has evolved from its initial focus on GPS devices to be a provider of products and experiences that help transform peoples lives so they can achieve their health and fitness goals. Their value proposition is based on exceptional customer experience and high quality of their offerings regardless of age, sex or ethnicity. The recent product line expansion is focused on increasing market share while attracting new clients outside of the traditional Mingar customer affordability levels. We conducted a market analysis and specific points with a focus on devices exhibiting racial bias. As a result of that work, we listed our findings below.

Key findings

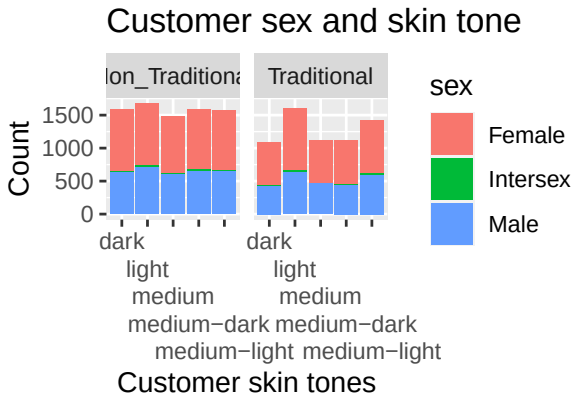
- The majority of non-traditional Mingar customers, the primary consumers of the "Active" and "Advance" lines, are older, live in more populated areas, and have lower median incomes than traditional Mingar customers, which have higher incomes
- Mingar's traditional clientele primarily consists of people with light and medium-light complexions compared to non-traditional consumers
- Assuming clients use non-traditional lines and could be female, clients with medium-dark complexions will report issues with their Mingar product 0.628 times less than Mingar product users with darker complexions
- Mingar customers with light/medium-light/medium complexions are 0.098/0.212/ 0.309 less likely to report issues with their Mingar devices, respectively, which indicates a bias of how Mingar technology performs based on the perception of client skin complexions
- There is an almost 90% discrepancy between the amounts of reported issues with Mingar devices from customers with darker complexions than those with lighter complexions

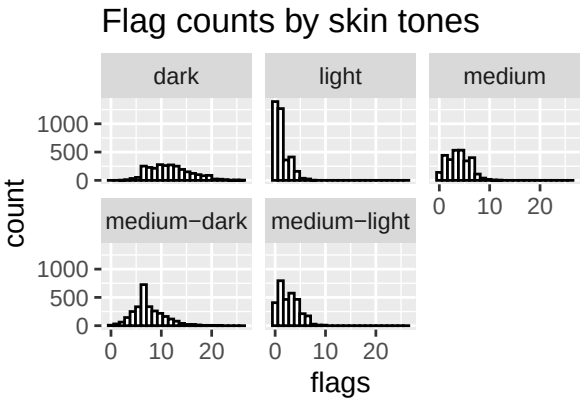
Limitations

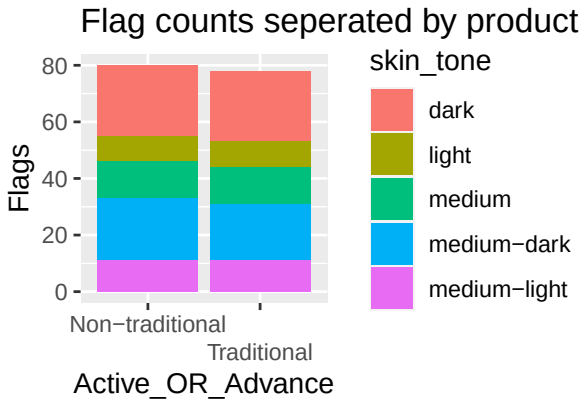
- The datasets provided for analysis had more non-traditional customer info than traditional customers, which would not provide accurate feedback for all Mingar products but only the new "Active" and "Advance" lines
- Most of the customer data provided were for females compared to males and inter-sex customers, which could provide great insight for female clients but fail to address concerns of other clients that are not
- Our team was uncertain if the linearity assumption was satisfied due to the categorical predictor variables, and as result, the model was discontinuous

Data Visualization

Below are key statistical summaries and visualizations of our analysis.







-As for the top left graph we see the different characteristics of customers in the data set. These characteristics include the skin tone and sex and these characteristics are separated based on if they are traditional or non-traditional customers. We see that in both lines women and light skin toned people make up the majority of customer base. It is notable that non-traditional customers have a more even distribution of skin tones.

-The top right graph shows the flag counts separated by skin tones of different customers. We see that lighter skin tones tend to be more dense around lower flag counts on average meaning their sleep scores are less affected, while darker skin tones are shown have a higher spread in flag count with a higher average which affects their sleep scores drastically.

-Lastly the bottom left graph shows the flag count by skin tone of different product lines. In both cases we see that dark skin tone people have the highest count while light skin people have the lowest count.

Technical report

Introduction

The technical report will consist of two sections each of which will cover some research questions that will serve to answer our clients needs. The first section will analyze the differences in the type of customers between new affordable device lines and the more traditional device lines. It will also discuss the possibility on whether the newer devices are attracting more customers outside the traditional customer base. The second section will investigate the issue on poor device performance for dark skin toned users and if there exists any potential racial biases that affect their sleep scores. It will also explore if this problem remains the same when we include other factors such as the sex of the user or the device type they use.

In all cases, we will thoroughly present and describe the suitableness of the statistical methods used to tackle each research question. Further, we will present any statistical results produced from our exploratory data analysis and model summaries of coefficient values, and confidence intervals of p-values. At the end we will provide a conclusion to our clients needs while also discussing the limitations and improvements to strengthen future analyses.

Research questions

- What are the differences between traditional and non-traditional customers?
- Do non-traditional customers come outside of traditional customers?
- Does the skin tone of a user affect their sleep score? If so, does it remain the same including other factors?
- Who are our the non-traditional customers?

For the first two questions we wanted to explore the characteristics of customers such as sex, age, income and skin tone to see if there is difference between traditional and non-traditional. We want to know if the new line of products that's meant to be affordable be is actually attracting new customers or different kinds of people. Finally the last question will use available data particularly focusing on sleep data which be used to check the sleep duration and flag count of an individual, and see how it relates to their skin tone.

Analyzing the difference in customers between the new and old lines of fitness tracking wearable devices

Introduction

With the advent of the 2 new active and advance lines, this section will primarily focus on how the customer response has been in relation to the new product lines. We will delve into the multiple factors and characteristics of the customers such as their income, skin tone, geographical location and other important factors to determine how the customers of these new lines differ from the customers of the other lines of products. We will examine the differences and similarities of customers of different lines of product and examine if the new active and advance lines are attracting customers outside of the companies traditional customer base. This will help better understand the value that these more affordable lines have on the customer base and if its worth further pursuing.

Terminology

Before continuing on with the report it is first important to define terminology that will be used further on in the report. As such, it is important to define what is meant by a traditional customer. In essence a traditional customer is a customer who has bought a device that is in the iDOL and Run line of products. As the company has asked us to submit an analyze of the new more affordable lines of products, those who purchase the more expensive lines of products are defined as traditional customers. Conversely, those customers who have bought a device from the new lines of products (Advance/Active) will be defined as non-traditional customer.

Central Research Question

Using the data given to us by the company as well as census and third party data, we will aim to identity in this section to analyze the differences between traditional and non-traditional customers and identify if the new line of devices is attracting new customers that differ outside the traditional customer base. Using these questions we will be able to gain a further insight on who the new customers are and if the new line of products are attracting new customers.

Data Collection Process

Before providing appropriate analysis of with the data it is important to understand where the data used in the technical report was sourced from. The data that will be used in this report was sourced from three different locations. The first location of where the data comes from is the Mingar corporation itself, the company for the purposes of the report has provided us with information on the customers and devices. In addition, to be able to gain other extrapolated data on the customers of different types of devices, information was sourced from the 2016 national Canadian Census. The 2016 census was used as it is the most latest released census and thus we are able to use the most up to date information. We can use this data to understand the

differences between the traditional and non-traditional customers. Finally the last source of data, API web scraping was used from the URL https://fitnesstrackerinfohub.netlify.app/ to get specific data on the various devices in addition to the information soured from the company itself. While collecting data it is important to note that privacy of the company and customers was a top priority and all data was sourced ethically.

Data Cleaning Process

In addition to understand where the data came from it is also important to know how the data was prepared for subsequent analysis. The first step was to collect the data from the aforementioned three sources. From then the data was all combined together primarily based of the customer information (i.e median income) and the device data. Then with one combined data set, new columns were added to make the data easier to work with. Examples of this include a new column of age of customer which was attained by date of birth and skin tone attained by emoji color of the customer. Furthermore, the data was checked for any duplicates and any missing key data points. Duplicate entries were deleted and any observation that contained missing data was removed. Finally, a new variable was added we segregated based of of being a traditional customer or non-traditional customer based on their device line.

Numerical Summaries

Using numerical summaries we can analyze how are buyers of the newer and more affordable "Active" and "Advance" products different to our traditional customers and if the 2 lines are attracting customers outside of the traditional customer base.

Table 1: The trimmed mean is trimmed by 10 percent

custType	Count	Min	Median	Max	Mean	Trimmean	Var	Range
Non_Traditional	7918	41880	65829	195570	68834.94	67554.64	210448849	153690
Traditional	6349	41880	65829	195570	73106.29	72655.26	219898204	153690

From Table 1 we are able to analyze the income of both traditional and non-traditional customers. As we see from the table, the min median and max values of both traditional and non-traditional customers are equal, which means customers from a large income array are customers of both the cheaper line of products and the more expensive line of products. However, both the mean and trimmed mean show that on average those are non-traditional customers have a lower median income compared to the traditional customers. This makes intuitive sense as traditional customers are purchasing the more expensive device and it makes sens that they have a greater median income. Lastly, we see that traditional customers have a greater variance and standard

deviation.

Table 2: The trimmed mean is trimmed by 10 percent

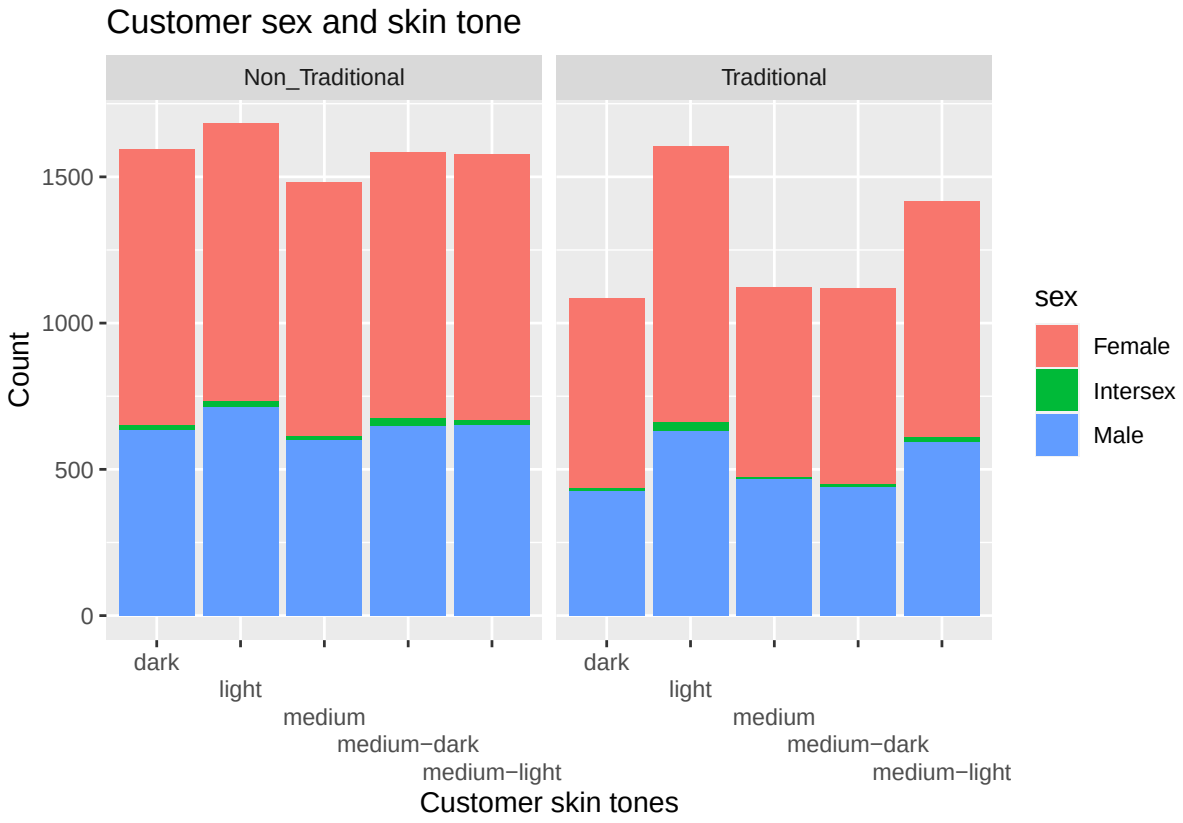
custType	Count	Min	Median	Max	IQR	Mean	Trimmean	Var	SD	Range
Non_Traditional	7918	18	47	92	32	47.96	47.46	341	18.46	74
Traditional	6349	18	46	92	21	46.50	45.89	218	14.76	74

From Table 2 we are able to analyze the age of both traditional and non-traditional customers. For both traditional and non-traditional customers the ages range from 18 to 92, but we see that the median, mean and trimmed mean for traditional customers is slightly lower. Lastly, we can see that variance and standard deviation is much higher for non-traditional customers. From this we can surmise that, traditional customers are slightly younger on average compared to non-traditional customers and have a smaller spread among their ages. This makes intuitive sense as those who are younger are more likely to use the more expensive features leading the age average of traditional customers to be higher.

Graphical Summaries

In addition to numerical summaries we are able to use graphical summaries to better understand both traditional customers and non-traditional customers.

Figure 1:



From Figure 1 we are able to see the makeup of customer's sex and skin tone separated by if they are traditional or non-traditional customers. From the table we can see that the customer base for both traditional and non-traditional customers are both dominated by females compared to men and inter-sex individuals. Furthermore, as for skin tone of the customer base its apparent that all the skin tones for non-traditional customers are more evenly situated compared to traditional customers where there is a noticeably higher amount of customers who are light and medium light skin toned compared to the other skin tones. As for non-traditional customers there is a noticeably higher count for all skin tones aside from light skin tone individual.

Methods

With our data cleaned and analyzed through numerical and graphical summaries we can move onto using the data in models to analyze the underlying question of this report on how are buyers of the newer and more affordable "Active" and "Advance" products different to our traditional

customers and if the 2 lines are attracting customers outside of the traditional customer base?

In order to accomplish this we will be using a logistical generalized linear model. The model is the following:

$$is\_tradional \sim Bernoulli(\mu_i)$$

$$log(\frac{\mu_i}{1-\mu_i}) = \beta_1 sex_i + \beta_2 age_i + \beta_3 income_i + \beta_1 skintone_i$$

where:

- μ_i : is the probability of being a traditional customer
- log[μ_i / (1 - μ_i)] : is a log odds
- μ_i / (1 - μ_i): is an odds
- if $\mu_i \approx 0$ then $\mu_i \approx \mu_i/(1-\mu_i)$

The reason logistic generalized linear models were used has to primarily do with the fact that the response variable is if the customer is a traditional or non-traditional customer (a binary value). In addition, it is assumed that the predictor variables of sex, age, income and skin tone are all independent. Using this logistic generalized linear model we will be able to analyze the differences between the traditional and non-traditional customers.

Also, before moving onto results it is important to center our quantitative parameters of age and median income to make interpreting the coefficients more meaningful. Currently our intercept is log odds when age and median income is 0. To alleviate this, both parameter values were all subtracted from their median quantile values (age: 47 and hhld_median_inc = 65829). It is important to note that it does not change the relationship between the values.

Lastly, for the purposes of logistic regression, if the customer is a non-traditional customer then the customer has been assigned the value of 1 and if the customer is a non-traditional customer then the customer has been assigned the value of 0. This is die to the fact that we are able to interpret the results as an “even occurring” if the customer is a non-traditional customer.

Results

Table 3: Results from the generalized linear model (converted from log odds)

	estimates	p.values	lower.bound	upper.bound
(Intercept)	3.9265944	0.0000000	3.2051604	4.8137909
age	1.0050834	0.0000005	1.0030985	1.0070752
hhld_median_inc	0.9999807	0.0000000	0.9999782	0.9999832
as.factor(skin_tone)light	0.9225714	0.1485979	0.8269843	1.0291587
as.factor(skin_tone)medium	0.9860574	0.8042622	0.8824193	1.1018656
as.factor(skin_tone)medium-dark	0.9581546	0.4441931	0.8587602	1.0690245
as.factor(skin_tone)medium-light	0.9679646	0.5661454	0.8660569	1.0818437
as.factor(sex)Intersex	1.1848450	0.2913314	0.8666265	1.6292419
as.factor(sex)Male	1.0421574	0.2360471	0.9733708	1.1158514

From Table 3 we are able to compare the attributes between traditional customers and non-traditional customers. As a reminder, when interpreting the results non-traditional customers were assigned the value of 1 as to be given the "occurring event" interpretation in the model. As we can see from the summary table, the skin tone of the customers and sex of the customers had p values of greater than 0.05 and so we can not definitively conclude if these factors are statistically significant between traditional and non-traditional customers. In addition, we found that household income and age were statistically significant with p values lower than 0.05. When we examine the coefficient estimate of age we see that it is 1.0050834 meaning that compared to traditional customers that non-traditional customers seem to be generally older in age. In addition as for income the estimated coefficient was 0.9999807 meaning that those who are non-traditional customers generally have lower median income. As for the age, with the dummy variable being female that inter-sex and males have a higher odds of being non-traditional customers and as for skin tone we see that all other skin tones have a smaller odds of being a non-traditional customer compared to people with dark skin. These results follow the data analysis in the data section.

Conclusion

Using what we have learned from this analysis we are able to analyze how are buyers of the newer and more affordable "Active" and "Advance" products different to our traditional customers and if the 2 lines are attracting customers outside of the traditional customer base.

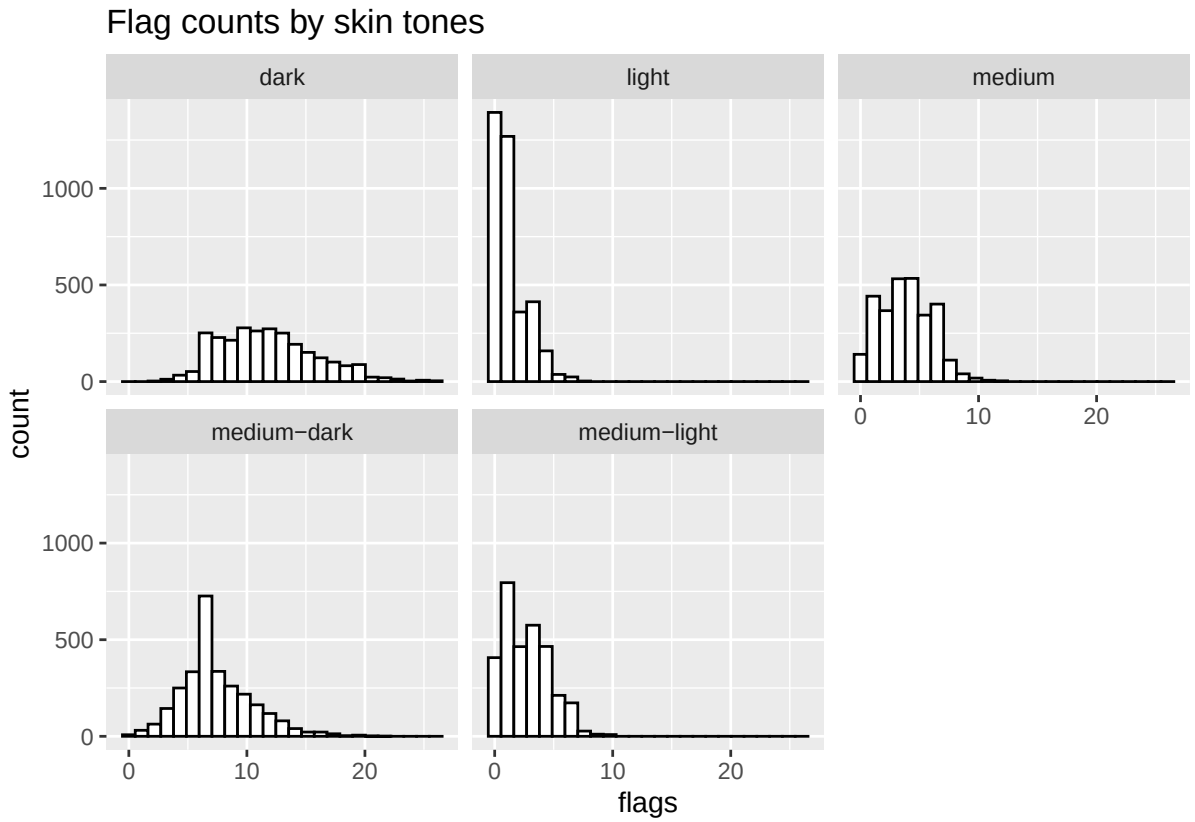
Starting off with the differences between traditional customers and non-traditional customers from the data analysis we see that starting off the data set used had a greater number of non-traditional customers compared to traditional customers. Further, more we see that for both traditional customers and non-traditional customers there were more females compared to males and inter-sex people. In addition we see that the traditional customers consists of more light and medium light skin people compared to non-traditional customers. Also, we see from the numerical summaries we see that non-traditional customers are generally older, live in more populated areas and have lower median incomes than traditional customers. Lastly, from our modeling we see that the statistically significant differences between their ages and incomes. This helps us understand if the non-traditional customers are from outside the traditional customer base. As we see from the report findings, non-traditional customers are generally outside of the regular customer base in terms of age and income levels. This makes sense as the new line was designed as a cheaper alternatives so it follows those we are in the non-traditional customer base would be older and have less median income. To conclude, in terms of who are these new customers we know its statistically significantly those who have lower median income and are generally older.

Investigation: Worse sleep scores for darker skin tones

For this analysis we were not given a clear variable on sleep score and so we calculated sleep score based on the flag count of an individual. Furthermore, since Mingar isn't authorized to extract information about their users' individual ethnicities, we instead received the variable called the "emoji modifier", which represents the skin tone of the emojis they use or react feature of the user. We will use this to conduct some of our exploratory data analysis and statistical modeling for this section to investigate if there exists any potential racial biases.

Methods

Figure 2:



In the histogram shown in Figure 2 above we can see that there exists quite a bit of evidence that for lighter skin tones having less flag counts since their graphs tend to be more dense around 0 and is right skewed. Darker skin tones on the other hand seem to have a greater overall mean and with greater variance in their densities, which seem to indicate that these groups have significantly more flags on average in general compared to the lighter skin toned groups.

The research question for section will be to conduct an analysis on whether or not the skin tone of a user does in fact have an affect on their respective sleep scores. We will explore some

potential variables that may have a have some relationship with the amount of flags the device gives, and then we will produce a model that will hopefully best explain this issue.

Out of the data gathered, there was a total of 20622 device users with their own unique ID. (We chose to exclude the emojis that identified yellow to focus more and make a better comparison within each skin tone group). Mingar has kindly provided us with their customer data which kept a record of each of the 20622 users including the date, sleep duration in minutes, and the amount of flags during their sleep session. Since Mingar wants us to discuss how sleep score is affected, we will instead use the flag variable as flags give us information about the device's behaviour. The more flags occurring on a sleep session will indicate a worse performance of the device, which would in turn directly imply a worse sleep score reliability.

Data description & wrangling

Here are the following changes we made to prepare for this research question:

- Used the customer device data to create a new column variable called skin tone, which will be the response variable in our modeling and it includes 6 skin tone types ranging from light to dark
- Wanted to know whether the device line was "Active" or "Advance" and created a new column variable called "Active_Advance", which indicates if the device line is Active/Advance or other types
- Joined the customer device data with the customer sleep data with respect to individual customer ID.
- Removed customers with NA values in the sex column variable
- Removed the skin tone type "yellow" in skin tone variable to give a better comparison strictly between actual skin tone types

Purpose/Statistical methods

We decided to use a Poisson regression generalized linear model to determine our response variable to compare the differences between the flag count and skin tone. To use this model we would need to check if our data satisfies the assumptions for a Poisson model. First we checked if the response follows a Poisson response which clearly does since flags is a discrete value that take whole values starting from zero. We can also assume independence since none of the observations have any direct influence over each other and so we can assume that flag counts are independent to one another. Another important assumption we needed to check is if the mean is equal to the variance. We see that in the table _, the observed variance for flag is larger than the observed mean for each skin tone type. Therefore we have some over-dispersion and instead of using a traditional Poisson model, we will instead be using a negative binomial regression model to loosen this assumption. Finally for the linearity assumption, with the fact

that the predictors in this analysis are not continuous variables, we can say that the assumption of linearity in log(lambda) is not something worth considering, but could still be a potential problem when drawing conclusions from the model.

Table 4: mean and variance for skin tones

skin tone	Mean	Variance
light	1.15	1.68
medium	3.65	4.98
medium light	2.51	3.61
medium dark	7.44	10.17
dark	11.80	16.07

As we can see from Table 4, both the mean and variance increase as the shade of skin tone gets darker.

This negative binomial model can also be considered as a Poisson model that follows a Gamma distribution, where λ is random. It has the following mathematical expression:

$$log(\lambda/sleepduration) = \beta_0 + \beta_1 * (skintone) + \beta_2 * (ActiveAdvance) + \beta_3 * (sex)$$

we can equivalently write this as,

$$log(\lambda) = \beta_0 + \beta_1 * (skintone) + \beta_2 * (ActiveAdvance) + \beta_3 * (sex) + log(sleepduration)$$

where λ is the mean count of flags during a sleep session and sleep duration is the total amount in minutes of a sleep session. Active Advance represents if it is a Active/Advance or other, and finally sex represents if the person is Male, Female, or Inter-sex. The offset term in this model is the log of sleep duration and we chose this to be the offset because every customer has a different amount of flags for different amounts of sleep duration. For example, we can have a customer who sleeps 10 hours and gets 4 flags, while another customer could sleep for 5 hours and get no flags in their session.

For our model selection process, we started the negative binomial model off with only one predictor variable skin tone to see if it produced anything of statistical significance. We found that for every skin tone it produced a significant p-value of less than 0.05. Next we wanted to see

if non-traditional device lines have any effect and so we added the "Active or Advance" variable in the model to see if there is a difference in flag counts for these traditional and non-traditional lines. We then used a likelihood ratio test to compare the two models and the model with the two predictors produced a higher p-value, meaning that we should add this extra term. We also wanted to see if the sex of a customer has an effect as perhaps there could exist biases in sexes, so we performed another likelihood test with another additional sex term added. The likelihood ratio test showed that we should accept this model with the additional term and so we chose to go with a model with three predictor variables. These variables are skin tone, active or advance, and sex.

Results

Table 5: Results for model

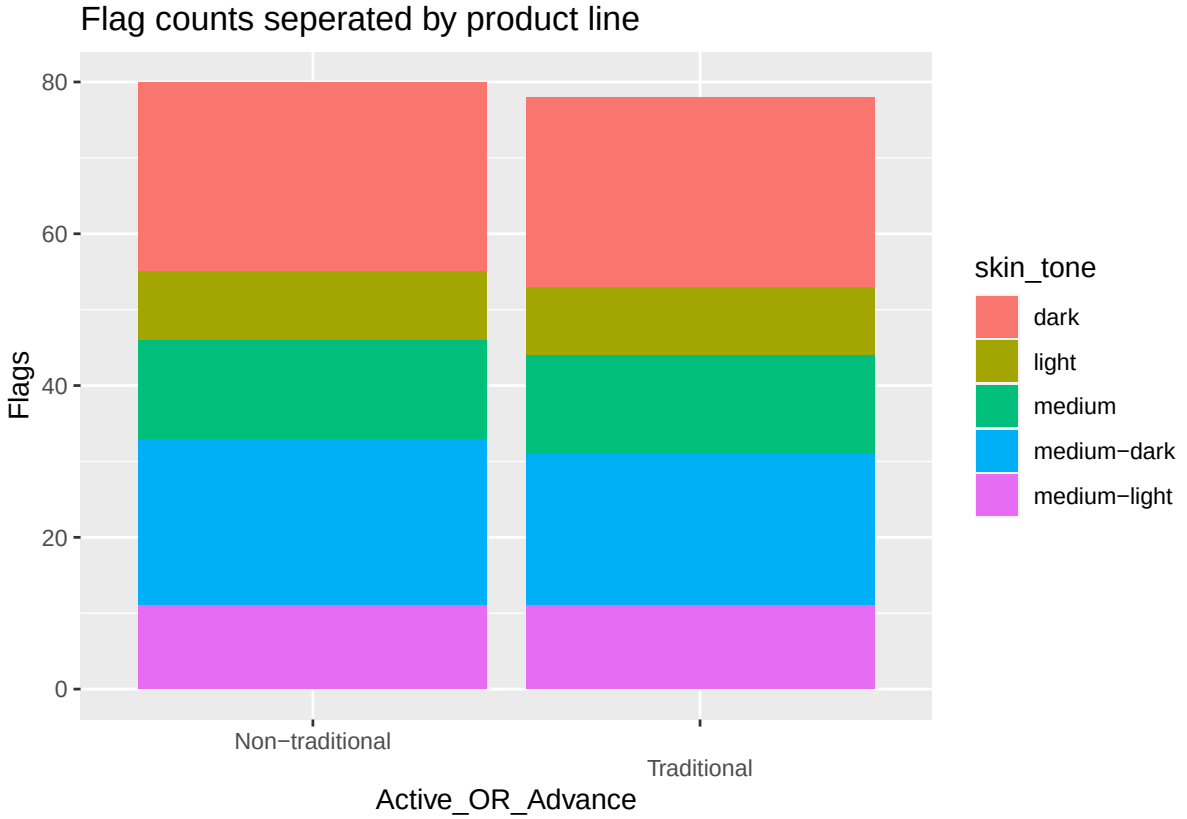
skin tone	Estimate	p-value	Upper Bound	Lower Bound
intercept	12.06	1.3210^{-9}	12.146	11.974
skin_tonelight	0.098	3.9710^{-6}	0.099	0.096
skin_tonemedium	0.309	2.3310^{-5}	0.313	0.306
skin_tonemedium-light	0.212	1.2110^{-4}	0.214	0.210
skin_tonemedium-dark	0.628	6.7210^{-7}	0.634	0.623
Active_Advanceothers	1.020	3.2010^{-8}	1.027	1.013
sexintersex	0.939	4.6710^{-4}	0.972	0.906
sexMale	0.933	1.8710^{-7}	0.940	0.926

In Table 5 shown above, we can see that all estimates exhibit a p-value far less than 0.05 which proves that all of them are indeed statistically significant. We interpret these estimates in relation to dark skin tone females who use non-traditional device lines (Active or Advanced). For example, the skin tone medium-dark has an intercept of 0.628 and we can interpret this by saying we expect medium-dark skinned customers to have less flags than dark skin tones by about 0.628 times, assuming they are also female and use a non-traditional device line. We can say the same interpretation for the rest of the skin tones, but notice how the value decreases as the skin tone gets lighter. We see that medium, medium-light, and light skins tone customers are 0.309, 0.212, 0.098 times of dark skin tones respectively, which clearly shows there exists bias here. When we strictly compare light to dark then dark skinned customers experience a 90% difference in flag counts more than light skinned customers.

As for the other estimates, there isn't much of a difference as their values are approximately or

at least close to 1. We also see in figure _, that the proportion of flag counts does not defer too much from each other between traditional and non-traditional device lines. Therefore for other cases where we perhaps have a Male customer with a traditional device line, there won't be much of a difference from our results explained earlier and the skin tone ratio remains approximately the same.

Figure 3:



In Figure 3 we see that for both Active and Advance and others, that those with dark skin tone have the highest number of flags and those with light skin tones have the lowest amount.

Conclusion

For the skin tone investigation, we quickly explored through graphs and summary statistics that darker skin tones tend have a higher flag count on average which led us to believe they have overall worse sleep scores. Consequently, the results from our model show that skin tone does have a great affect on flag count and also provides us with each skin tone estimate having considerable differences when compare light to dark as darker customers are a lot more likely to accumulate flags throughout their sleep sessions which in turn affect their sleep scores. We also found that these racial biases stay the same when we consider the person's sex and their device type.

Discussion

Over the course of this report we were able to go into a through investigation of traditional and non-traditional customers. From our analysis we found that non-traditional customers and traditional statistically contrast in their ages and median income. We found that non-traditional customers generally have lower median income levels and are generally older than traditional customers. We also found that sex and skin tone were not statistically signifant difference between traditional and non-traditional customers. From our analysis of skin tones we found that those have dark skin tones tend to have higher flag count which lead us to believe they have lower sleep scores compared to other people with other skin tones.

Strengths and limitations

Limitations

A limitation we had for the skin tone investigation was that we were not completely sure if the linearity assumption was satisfied. All the predictor variables in our model were categorical and thus it is not continuous. We would advise to use this model with caution as we cannot make any definite conclusions, but it may still be used to provide evidence.

For future analysis we would consider adding other predictor variables to the model to explore how racial biases might have changed in other scenarios with other factors involved, but we wanted to focus solely on studying on how skin tone effects on sleep score and so we made the model as simple as possible.

Another limitation we found was that there was an uneven number of traditional and non-traditional customers leading to an unbalanced experimental design. This imbalance leads to believe there is potential for bias in our analysis. In addition, it was assumed that the predictors for the model were independent. Lastly, we assumed the skin tone based of the customers emoji preference which led to us deleting observation of those who used yellow deafult emoji as we were not able to ascertain their skin tone.

Strength

One strength we had was that we were able to get a diverse amount of data which lead to create a diverse and expansive numerical summaries, graphical summaries and model. This lead our data to be more representative of the real world data. In addition, another one of the strengths is that data data we had was well organized and made modeling and summaries detailed. This allowed are models to very precisely target the research questions.

Consultant information

Consultant profiles

Eric Liu. Eric joined the firms retail sales division in the beginning of 2022 and is now a senior consultant of R3M. He has been effectively leading both the comprehensive data extraction and statistical analysis duties for R3M client needs. Eric graduated in 2023 from the University of Toronto with a Bachelor of Science degree, Majoring in Mathematics and Statistics.

Dhanraj Patel. Dhanraj has been with R3M since early 2022 starting out in retail sales and is currently a senior associate of the firm. Since joining, Dhanraj has successfully supported the implementation of comprehensive statistical analysis of R3M client needs. He studied at the University of Toronto, where he specialized in Statistics, and earned his Bachelor of Science degree in 2023.

Marko Sarenac. Marko has been with R3M as a senior consultant in the retail sales division since 2022. With his expertise he has successfully lead the statistical communications of the division and provided invaluable data visualization insight to R3M and its clients. Marko earned his Bachelor of Science, Majoring in Mathematics and Statistics from the University of Toronto in 2022.

Yanqing Weng. Yanqing has been apart of the R3M retail sales division since 2022 and is currently a senior consultant. He has successfully supported both the data extraction/visualization responsibilities and statistical communications of the retail sales division within R3M. Yanqing has a Bachelor of Science degree where he Majored in Mathematics and Statistics while studying at the University of Toronto, which he graduated from in 2023.

Code of ethical conduct

At R3M, our approach to statistical consulting is centered around consistently demonstrating that we operate at the highest level of ethical standards on a daily basis. All R3M employees are expected to abide by the following principles which are inline with proper business practices.

Honesty

R3M consultants operate with the goal to provide assurance to clients that regardless of what the circumstances may be, they are in fact conducting them selves in a manner that follows the agreed upon goals set, promises made and stated objectives.

Privacy

R3M consultants are required to maintain the privacy of our clients as top priority by offering measures that provide assurance which include but are not limited to, keeping sensitive info pertaining to clients and/or projects private, unless requested by the client or the law.

Equity

Consultants at R3M will treat clients and their interests fairly and not place them above the company nor that of the community. Furthermore, R3M employees will keep their own biases at bay while being objective and not going beyond their own expertise.

References

Data Sources

Fitness Tracker Info Hub. Retrieved from https://fitnesstrackerinfohub.netlify.app/ on 09/04/2022.

Census Canada Postal Code Conversion File: 2016 Census Geography, August 2021 Postal Codes. Retrieved from https://mdl.library.utoronto.ca/collections/numeric-data/census-canada/postal-code-conversion-file/2016 on 09/04/2022.

Canadian Income Census Data. Retrieved from https://censusmapper.ca/ on 09/04/2022.

Packages/Libraries

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

rvest package: Easily Harvest (Scrape) Web Pages (2021). Retrieved from https://rvest.tidyverse.org/, https://github.com/tidyverse/rvest

polite package: Be Nice on the Web (2019). Retrieved from https://github.com/dmi3kno/polite

lubridate package: Grolemund G, Wickham H (2011). "Dates and Times Made Easy with lubridate." Journal of Statistical Software, 40(3), 1–25. https://www.jstatsoft.org/v40/i03/.

lmtest package: Zeileis A, Hothorn T (2002). "Diagnostic Checking in Regression Relationships." R News, 2(3), 7–10. https://cran.r-project.org/doc/Rnews/.

MASS package: Venables WN, Ripley BD (2002). Modern Applied Statistics with S, Fourth edition. Springer, New York. ISBN 0-387-95457-0, https://www.stats.ox.ac.uk/pub/MASS4/.

Resources

Full Emoji Modifier Sequences. Retrieved from https://unicode.org/emoji/charts/full-emoji-modifiers.html on 09/04/2022.

Appendix

Web scraping industry data on fitness tracker devices

In the step of web scraping data, the terms and conditions from the source website must be followed. Robots documents should be consulted to follow instructions in terms of which robots are allowed or not allowed to be used to visit the website. In case case, there is a description of web scraping with 12 seconds crawl delay. Other ethical considerations to be considered are being responsible to the the data from web scraping, making a request in web scraping in an appropriate rate, using the data from web scraping to create our own work.

Accessing Census data on median household income

After census data in 2016 is accessed, the data of median household income is summarized in by selecting relevant data in terms of median household income from the original data and summarizing them into a new data tibble. Only this new data tibble is used towards producing later work in this project. The terms and conditions must be strictly followed at the time to accessing the data from license agreement and only information relevant to the final report shall be saved.

Accessing postcode conversion files

The postal code conversion file is accessed from the census website. Then, the postal code data is summarized by selecting only relevant information including postal code and making it into a new dataset in renaming the postal code variable name to an appropriate one. With the idea of following terms and conditions from the census website in terms of the license agreement, this new dataset with relevant information only will be used to produce further project work.