Report title

Subtitle that indicates findings

Report prepared for MINGAR by [MANGO]

2022-04-07

Contents

General comments (you can delete this section)

Before making any changes, knit this Rmd to PDF and change the name of the PDf to something like "original-instructions.pdf", or whatever you like (it is just for your reference).. Then you can delete this section and if you want to check what it said, just open the other PDF. You don't HAVE to use this particular template, but you DO need to write you report in RMarkdown and include a cover page.

The cover page must be a single stand alone page and have:

- *A title and subtitle (that indicate your findings)*
- *"Report prepared for MINGAR by" your company name*
- *Date (assessment submission date is fine)*

You can change the colour of this cover to any colour you would like by replacing 6C3082 in

the YAML above (`titlepage-color:`) to another hex code. You could use this tool to help you: https://htmlcolorcodes.com/color-picker/

Note: There should NOT be a table of contents on the cover page. It should look like a cover.

Executive summary

Guidelines for the executive summary:

- *No more than two pages*
- *Language is appropriate for a non-technical audience*
- *Bullet points are used where appropriate*
- *A small number of key visualizations and/or tables are included*
- *All research questions are addressed*

Table 1: Demographics of customers in old and new line

	Old Line	New Line
Location	Ontario, followed by Quebec, Alberta and British Columbia	Ontario, followed by Alberta, Quebec and British Columbia
Gender	58.6% female	57.8% female
Mean Income	73166.18	68815.47

The module 4 writing prompt provides some tips and information about writing executive summaries.

Technical report

This part of the report is much more comprehensive than the executive summary. The audience is statistics/data-minded people, but you should NOT include code or unformatted R output here.

Introduction

Provide a brief introduction to your report and outline what the report will cover. This section is valuable for setting scope and expectations.

Research questions

- How do the demographic for the traditional lines compare with that of the "Active and "Advance" lines?
- Are the devices "racist"?
- What are the main differences between different lines of the device?

What is the demographic of different lines of product?

Location of customers

To create the plot that demonstrate the location of users in different production line, we combined the user dataset with web script Census Canada Postal Code Conversion dataset by postcode. We can therefore generate the province that the user belongs to by extracting the first two digit of CSDuid (Census subdivision unique identifier) from the Census dataset.

In Figure 1, the size of the bubble is proportional to percentage of customers in each province in that line and the position of each bubble represents the latitude and longtitude of that province. From the plot, we can notice that Ontario has the most users in all four lines. In Active line and Advance line, The users in Quebec is slightly less than ones in Ontario, but in Run line, the number of users in Alberta exceeds the number of users in Quebec. From the result, we are able to see where the target customers is located for product in different lines.

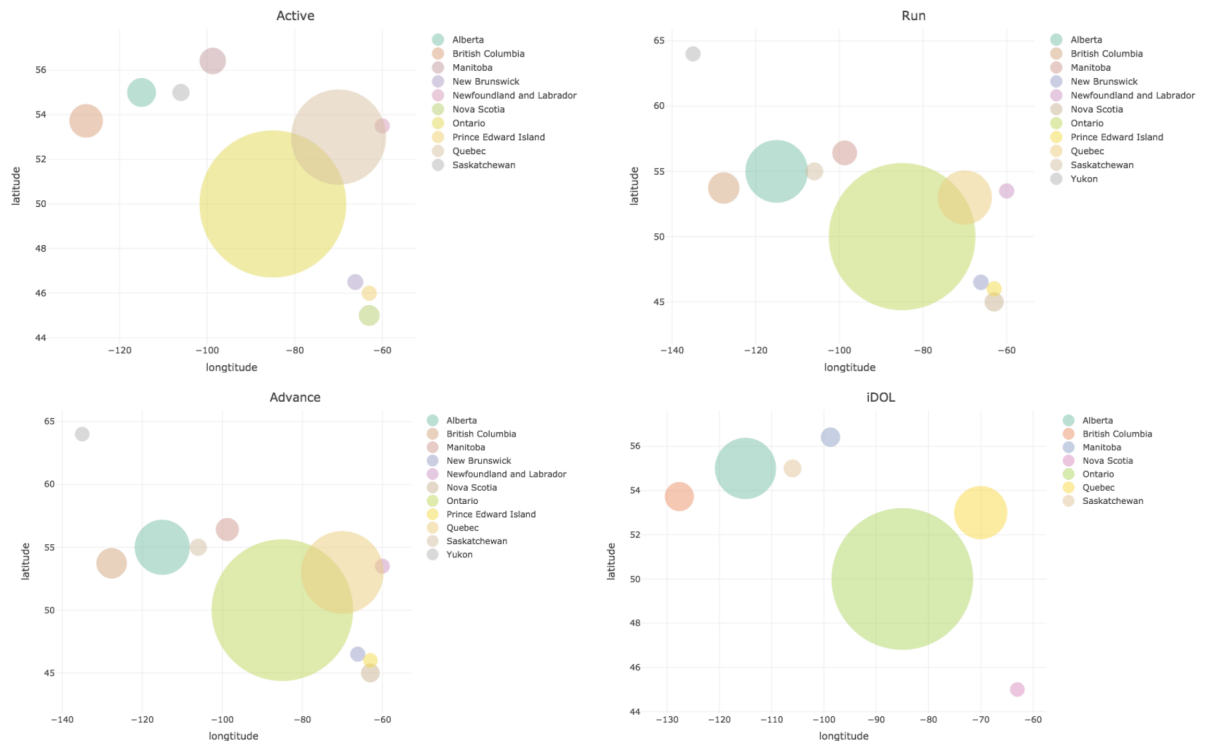


Figure 1: Customer location by device line

Age of customers

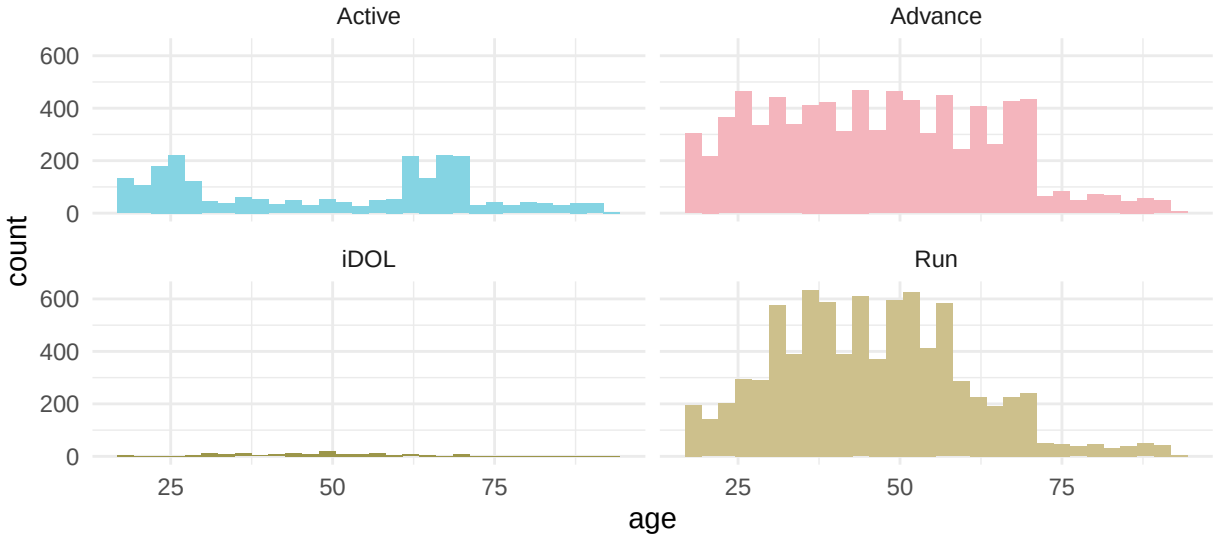


Figure 2: Age of customer by device line

To get a sense of how the age of customers is distributed in different lines, we created a histogram plot of the Age of customer by device line. In the plot above, we can see that Advance, Run and iDOL line have a similar distribution that is slightly right-skewed and there is a bimodal distribution in Active line. We deside to use Kruskal Wallis test since the response do not have a normal distribution and there are more than two categories for the independent variable, `line`.

Assumption Checking

- It satisfies the assumptions that there are two or more levels in the independent variable `line`, the dependent variable,`age`, is on a ratio scale and the observations are independent.
- The fourth assumption is that all groups should have same shape distributions. In our case, three of the groups have same shape distribution, while Active line have a bimodal distribution, which may be caused by the lack of data. It is a limitation in the model assumption checking process.

Kruskal Wallis Test After running the Kruskal Wallis Test, we get a p-value of 5.206^{-13}. It is smaller than the significant level of 0.05, thus we can conclude that the customers in the four lines do not have the same median age.

Gender of customers

To test whether the customers gender in new and traditional lines are different. We created a new variable `line_binary` which have a value `new` for active and advance line, and `old` for other lines.From the bar plot below, we can notice that there are more female customers in both new line and old lines. Since there are more females in either case, we created a new variable `sex_binary` which have a value 1 for female and 0 for other genders, in order to do the two sample t-test.

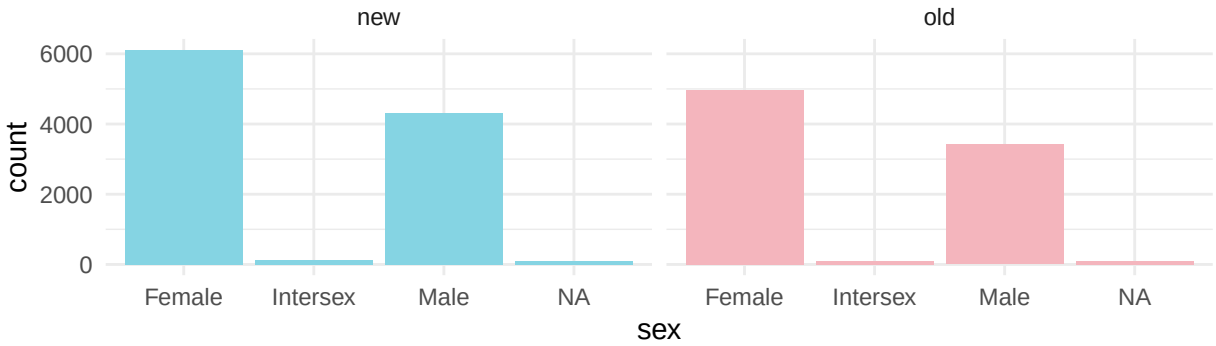


Figure 3: Sex of customer by device line

Two Sample t-test We conducted a two sample t-test on it. The mean in group `new` and group `old` is 0.578 and 0.586 respectively. Which indicates that there are approximatly 58% customers that are female in all lines. Since the p-value is 0.2598, we do not reject the hypothesis that the difference in mean gender equal to 0 between two groups of lines. The result indicates that the gender of customers do not very for old and new lines.

Income of customers

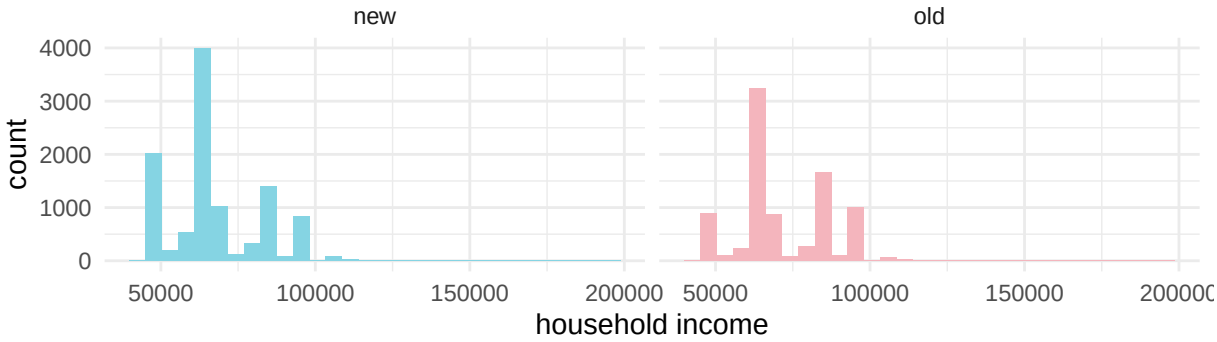


Figure 4: Household income of customer by device line

Two Sample t-test From the two sample t-test, the mean income in group `new` and group `old` is 68815.47 and 73166.18 respectively with a p-value of 2.2^{-16}. Thus, we reject the hypothesis that the two groups have the same mean in income and we can see that the customers who buy products from the new line have around 4000 less mean income than those who buy from the old line.

How does skin color affect the accuracy of wearable devices?

Data Manipulation

The dataset "race_sleep_data" is obtained through combining the customer profiles, the sleep tracking, and the device data. The names are cleaned and only variables relating to quality and customer attributes are selected. The newly added variable "skin_tone" divides the users to different skin tone categories depending on their selected emoji modifier.

Exploratory Plots

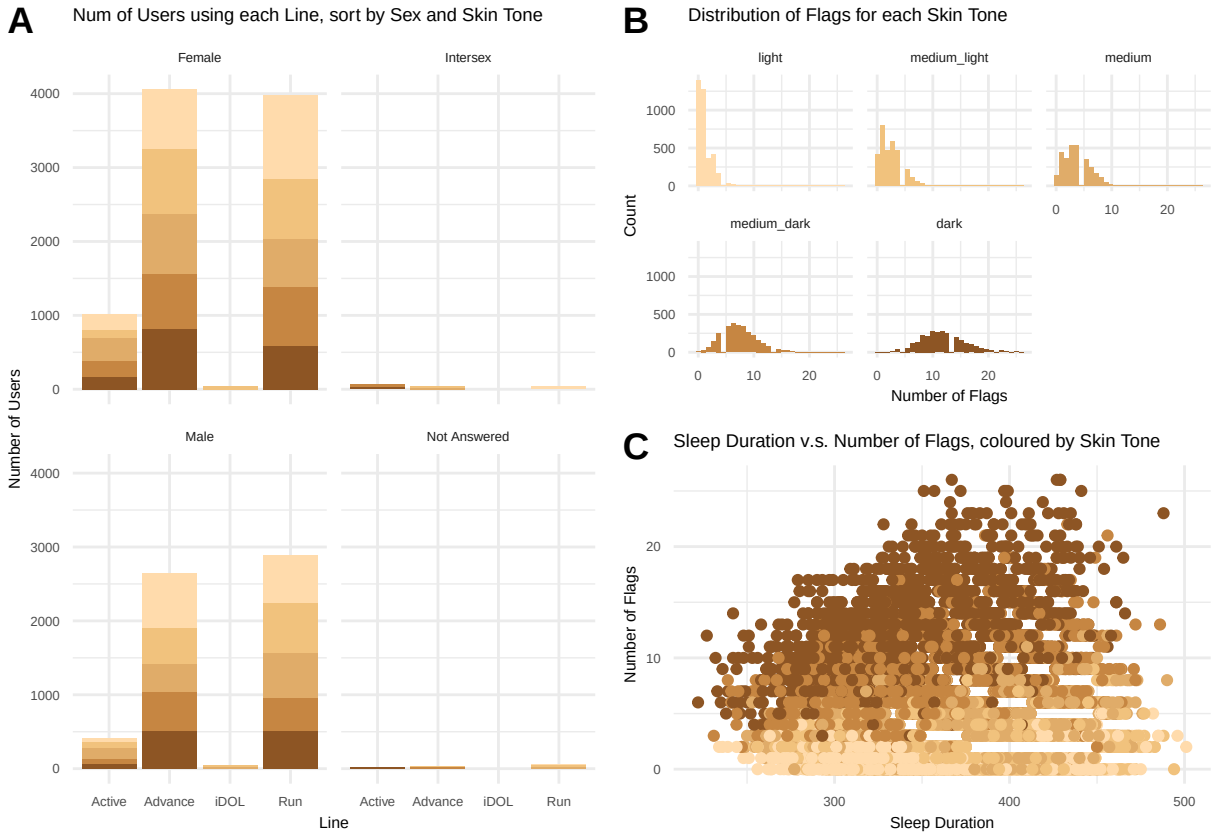


Figure A shows that there are more light skinned users than dark skinned users and there are more females than males and intersex users. Most users use the Advance or Run lines.

Figure B shows that darker skinned users seem to have a larger spread and larger mean of number of flags than compared to lighter skinned users, suggesting some evidence of a performance bias.

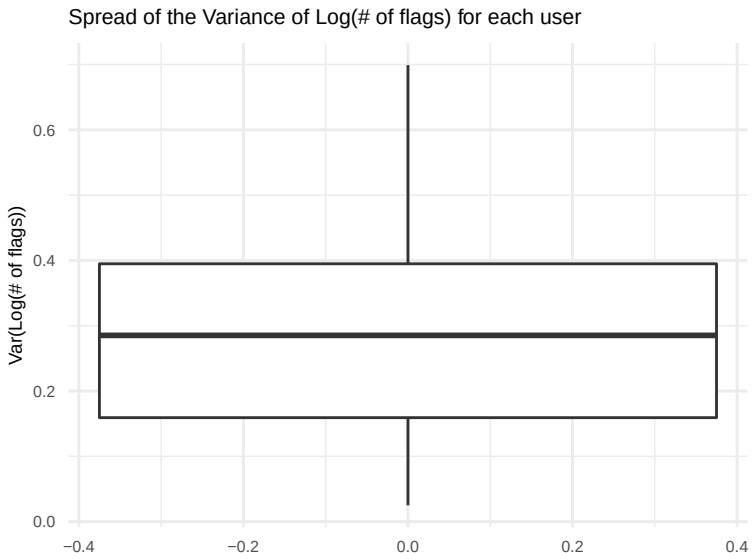
Figure C shows that there is not much relationship between duration of sleep and number of flags, indicating that duration may not be an important predictor for the number of flags.

Fitting a GLMM with Poisson Link Function

Model Assumption Checks

1. Independence of Subjects: the independence of subjects assumption is satisfied as we assume that every customer is independent from each other and are randomly selected from the population
2. Random Effects (each individual) come from a normal distribution

3. Random Effect errors have constant variance:



We see that the variance of the number of flags transformed by the link function (log) is mostly the same with little spread across individuals. This suggests that the homoscedasiticy of random effect errors assumption is satisfied.

4. Poisson link function is appropriate:

From the Figure B above, we see that the number of flags (response variable) ranges from 0 to over 20, is right skewed, and can be modeled by Poisson distribution. We also check if the mean and variance of number of flags is equal for each skin tone:

Skin Tone	Mean(# of flags)	Var(# of flags)	Number of Observations
light	1.156522	1.682480	3680
medium_light	2.502364	3.596858	3173
medium	3.653950	4.982877	2962
medium_dark	7.432215	10.163701	2862
dark	11.792333	16.089398	2687

We do observe some evidence of a violation of the mean=variance assumption; however, any violations are modest. Hence, the Poisson model is appropriate in this case.

Model Fitting

$$Y_{irs} \sim Poisson(\lambda_{irs})$$
$$\log(\lambda_{irs}) = \beta_0 + X_{irs}\beta + U_i$$
$$U_i \sim N(0, \sigma^2)$$

- Y_{irs} is the number of flags for individual i with skin tone r and sex s
- X_{irs} has indicator variables for skin tone and sex
- U_i is the individual level random effect

	2.5%	97.5%
(Intercept)	1.1397195	1.2263361
skin_tonemedium_light	2.0651302	2.2639955
skin_tonemedium	3.0099785	3.2920486
skin_tonemedium_dark	6.1408829	6.6919149
skin_tonedark	9.7524545	10.6208188
sexIntersex	0.7986575	1.0322661
sexMale	0.9097382	0.9571648
sexNot Answered	0.8353090	1.1004299

From the model, we see that the general trend is the darker one is, the higher the log of the mean number of flags their devices report. For example, compared to the base line of the light skin tone, we are 95% confident that, compared to light skinned users, the mean number of flags for a user with the dark skin tone changes by a factor of between (9.77, 10.65), which is about ten times more. The p-values for each level of skin tone predictors are extremely small, suggesting strong evidence against the fact that the mean number of flags are equal for different types of skin tones. We also observe that there is a (1 - 0.91) x 100% = 9% decrease in the mean number of flags for males compared to females; however, this result may just be due to the relatively fewer data for male users than for female users.

What are the main differences between different lines of the device?

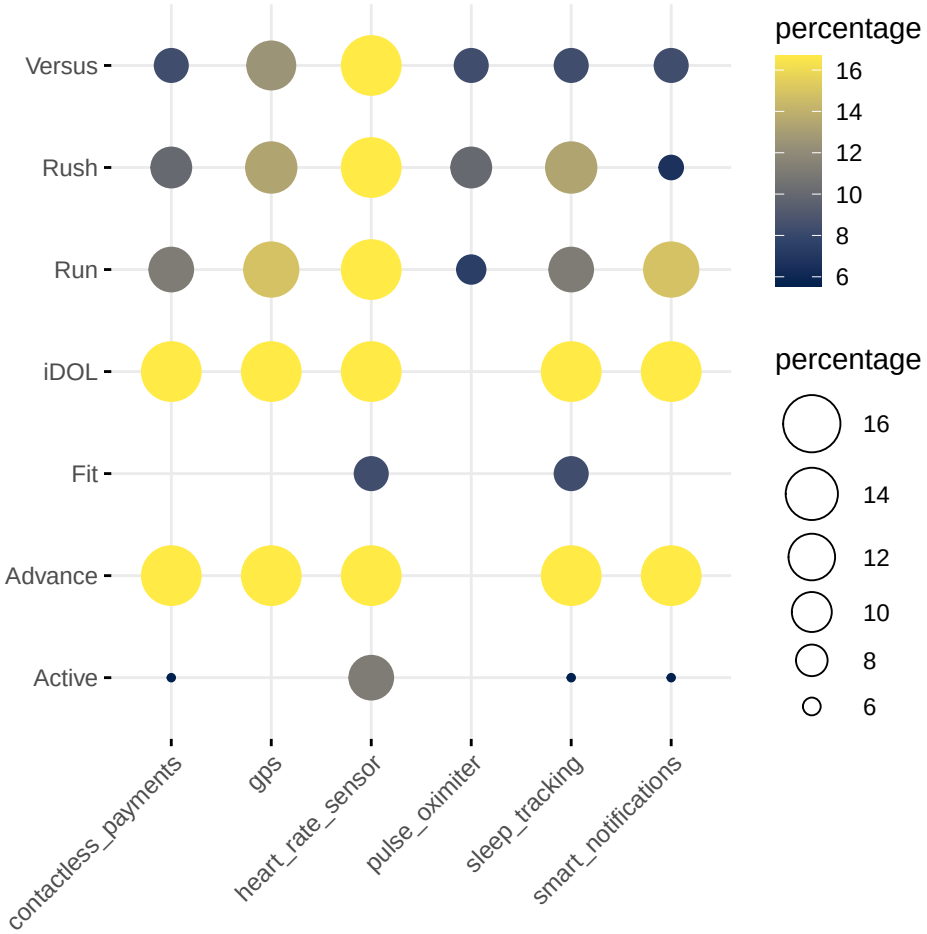
In this research question, we are interested in the differences of the **performance/price** among the different lines of services, combined with the different demographic of the users between

lines, we can give a report summary/advice for what has changed and how the change affects the target audience.

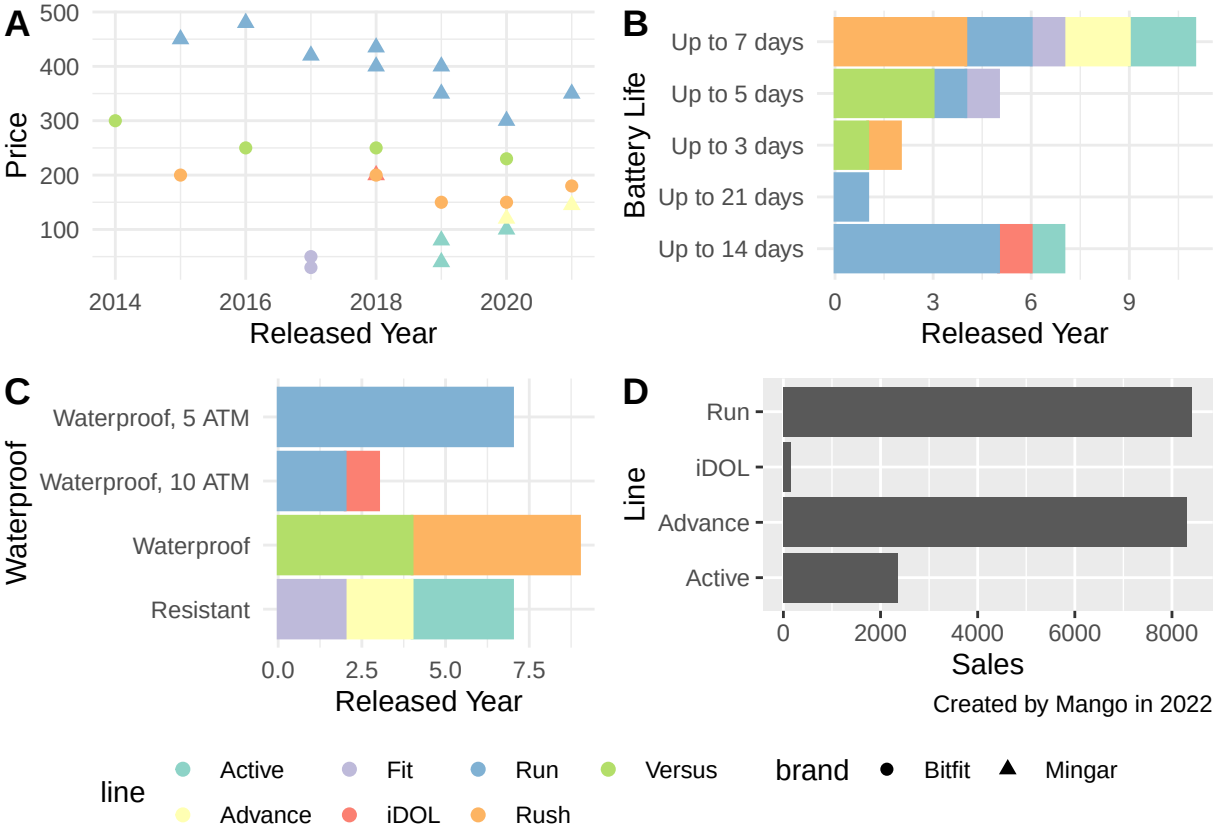
Data Manipulation

We used three database in this part. The first one is ***device_data***, which is retrieved by web scraping from the fitness tracker info hub. This database includes all the information on the devices produced by the company *Mingar* and *Bitfit.* The second database is ***device_cust*** including the data for individuals who bought the devices from *Mingar.* This database is obtained by merging the database *cust_dev* and *customer.* The third database used in this question is ***data_performance,*** which includes the percentage of the devices having some functionality in each line**.** This dataset is built by group the *device_data* by line and functionality. More details about the code of data waggling is in the *data_prep.Rmd* file.

Exploratory plots



The above plot shows the performance of different lines. The size of the circle is the percentage of devices in the line on x-axis that has the functionality on y-axis. Based on the plot, the line *Run* has all of the functionality, while line *Active* and *Advanced* has all except the pulse oximiter. By the size of each circle, the line *Run* has a more diverse functionality than the other two lines. Each device in line *Advanced* almost has the same functionality.



Plot A above is a scatter plot showing the recommended retail price, released year and the lines. According to the plot, the line *Run* has the highest price in the range of $300 - $500, while line *Advanced* and *Active* has a much lower price from $50 to $150. The average price of all lines in *Mingar* is $284.66. And the average price of all lines in *Bitfit* is $180.90. By the plot A, the released year for the line *Run* is spread through the x-axis, while the line *Advanced* and *Active* only came out in the most recent 3 years.

Plot B is a bar chart showing the battery life for each line. Based on the plot, we noticed that line *Run* usually has the longest battery life, since most of the devices of *Run* has a battery life up to 14 days and more. However, the battery life of line *Advanced* and *Active* is shorter, which is about 5-14 days.

Plot C is a also a bar chart showing the performance of waterproof. Line *Run* has a better waterproof performance, all devices from *Run* has a waterproof of 5ATM-10ATM, while devices

from line *Advanced* and *Active* are only water resistant, which is the lowest level of waterproof.

Plot D shows the popularity of each line in *Mingar*. From the bar chart, we can conclude that the line *Run* and *Advanced* are the two most popular line in *Mingar*. And the sales of *Run* is slightly greater than *Advanced.*

Conclusions

Above all, we noticed that the line *Run* and *Advanced* are the most popular lines at *Mingar*. Line *Run* has an overall better performance and a higher price among all other lines. Run is one of the earliest lines the company started producing, thus, line *Run* has 9 devices, which is the line having most products. Line *Active* and *Advanced* are two new lines that came out in the most recent 3 years. There are only 5 devices for the line *Advanced* and *Active.* However, the total sales of line *Advanced* and *Active* is more than that of line *Run.* The result implies that when the devices with sufficient functionality, price is the major factor for number of sales.

Discussion

In this section you will summarize your findings across all the research questions and discuss the strengths and limitations of your work. It doesn't have to be long, but keep in mind that often people will just skim the intro and the discussion of a document like this, so make sure it is useful as a semi-standalone section (doesn't have to be completely standalone like the executive summary).

Strengths and limitations

Consultant information

Consultant profiles

Complete this section with a brief bio for each member of your group. If you are completing the project individually, you only need to complete one for yourself. In that case, change the title of this section to "Consultant profile" instead. Examples below. This section is only marked for completeness, clarity and professionalism, not "truth" so you can write it as if we're a few years in the future. Put your current degree in as completed and/or add your first choice grad school program, whatever you like. What skills related skills would you most like to highlight? What job title do you want?

Statsy McStatsstats. Statsy is a senior consultant with Eminence Analytics. She specializes in data visualization. Statsy earned her Bachelor of Science, Specialist in Statistics Methods and Practice, from the University of Toronto in 2023.

Datana Scatterplot. Datana is a junior consultant with Eminence Analytics. They specialize in reproducible analysis and statistical communication. Datana earned their Bachelor of Science, Majoring in Computer Science and Statistics from the University of Toronto in 2024.

Code of ethical conduct

This section should be fairly short, no more than half a page. Assume a general audience, much like your executive summary.

- *Make at least three relevant statements about your company's approach to ethical statistical consulting. These should be appropriately in line with professional conduct advice like the (Statistical Society of Canada Code of Conduct)[https://ssc.ca/sites/default/files/data/Members/public/Accreditation/ethics_e.pdf] or the (Ethical Guidelines for Statistical Practice from the American Statistical Society)[https://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx]. For example, "the customer is always right" ISN'T the type of thing an ethical statistical consultant would include.*
- *Be very careful not to just copy and paste from these other documents! Put things in your own words.*

References

You don't need to cite course materials, but consider all the the places you got data from, as well as the packages used and R itself. These are all things you should consider citing. Likewise, you might use some external resources on the emoji skin tones/Fitzpatrick scale, etc.

Appendix

These appendices should outline in more detail the steps taken to access the following datasets. They should NOT include code, but should briefly describe the steps and important considerations. I.e., show that you understand what needs to be considered when web scraping, protecting licensed data, etc.

Web scraping industry data on fitness tracker devices

Accessing Census data on median household income

Accessing postcode conversion files

Final advice: KNIT EARLY AND OFTEN!