Report title

Subtitle that indicates findings

Report prepared for MINGAR by [MANGO]

2022-04-07

Contents

General comments (you can delete this section) **2**

Executive summary **4**

Technical report **5**

Introduction 5
What is the demographic of different lines of product? 5
How does race affect the accuracy of wearable devices? 12
What are the main differences between different lines of the device? 28
Discussion 34

Consultant information **35**

Consultant profiles 35
Code of ethical conduct 35

References **36**

Appendix **37**

Web scraping industry data on fitness tracker devices 37
Accessing Census data on median household income 37
Accessing postcode conversion files 37

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*

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.

```
device_cust <-read.csv("./data/device_cust.csv")
```

```
province <- c(10, 11, 12, 13, 24, 35, 46, 47, 48, 59, 60, 61, 62)
province_name <- c("Newfoundland and Labrador", "Prince Edward Island","Nova Scotia",
↪  "New Brunswick", "Quebec", "Ontario", "Manitoba", "Saskatchewan", "Alberta",
↪  "British Columbia", "Yukon", "Northwest Territories", "Nunavut")
latitude <- c(53.5, 46, 45, 46.498390, 53, 50, 56.415211, 55, 55, 53.726669, 64,
↪  62.135189, 70.453262)
longtitude <- c(-60, -63, -63, -66.159668, -70, -85,-98.739075, -106, -115,
↪  -127.647621, -135, -122.792473, -86.798981)
canada_pro<- data.frame(province = province, province_name = province_name,
↪  latitude=latitude,longtitude= longtitude)

device_cust$province <- substr(device_cust$CSDuid,1,2)
device_cust <- merge(device_cust, canada_pro, by="province")
province_data <- device_cust %>% group_by(province, line) %>% summarise(num = n(),
↪  province_name, latitude, longtitude) %>% unique()
```

```
## `summarise()` has grouped output by 'province', 'line'. You can override using
## the `.groups` argument.
```

```
province_data_active <- province_data[province_data$line == "Active",]

province_data_advance <- province_data[province_data$line == "Advance",]
province_data_run <- province_data[province_data$line == "Run",]
province_data_idol <- province_data[province_data$line == "iDOL",]
```

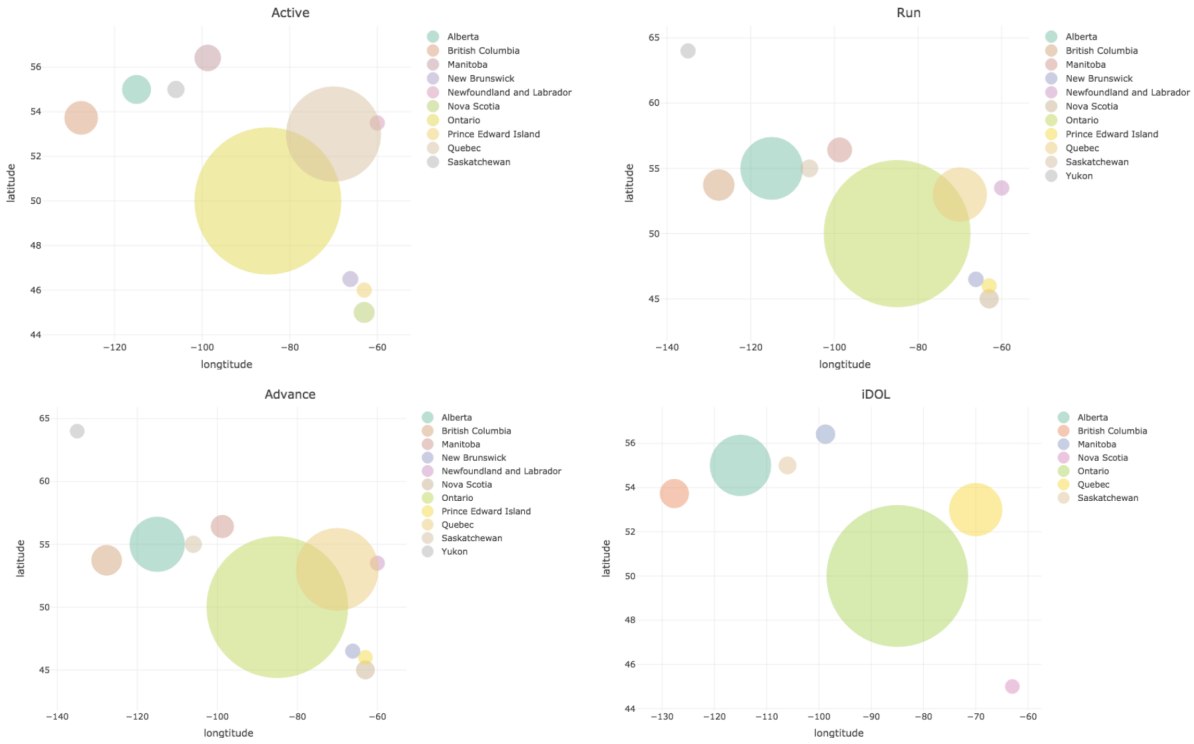


Figure 1: Customer location by device line

Age of customers

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 below, 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.

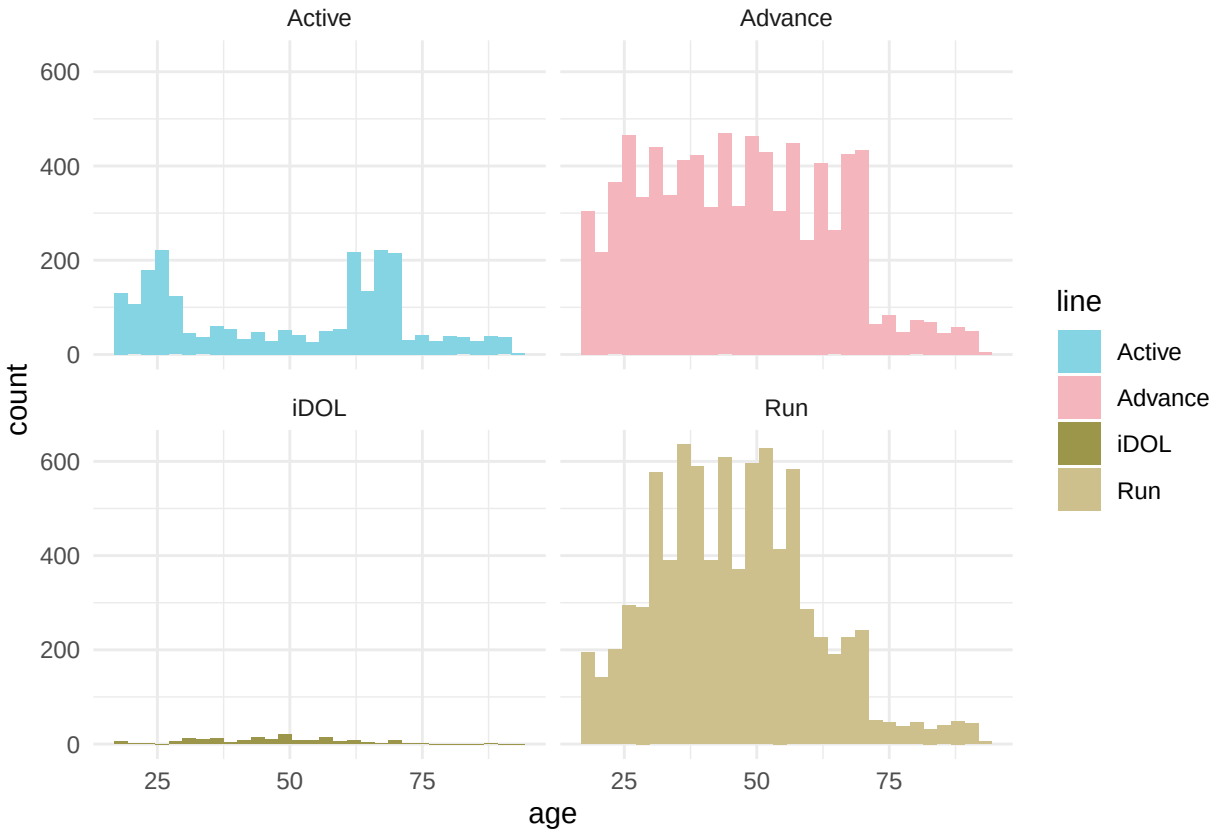


Figure 2: Age of customer by device line

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.

```r
kruskal.test(age ~ line, data = device_cust)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  age by line
## Kruskal-Wallis chi-squared = 60.247, df = 3, p-value = 5.206e-13
```

Gender of customers

From the bar plot below, we can notice that there are more female customers in all four lines. To test whether the customers gender in new and traditional lines are different. We created a new variable `sex_binary` which have a value 1 for female and 0 for other genders and a new variable `line_binary` which have a value `new` for active and advance line, and `old` for other lines.

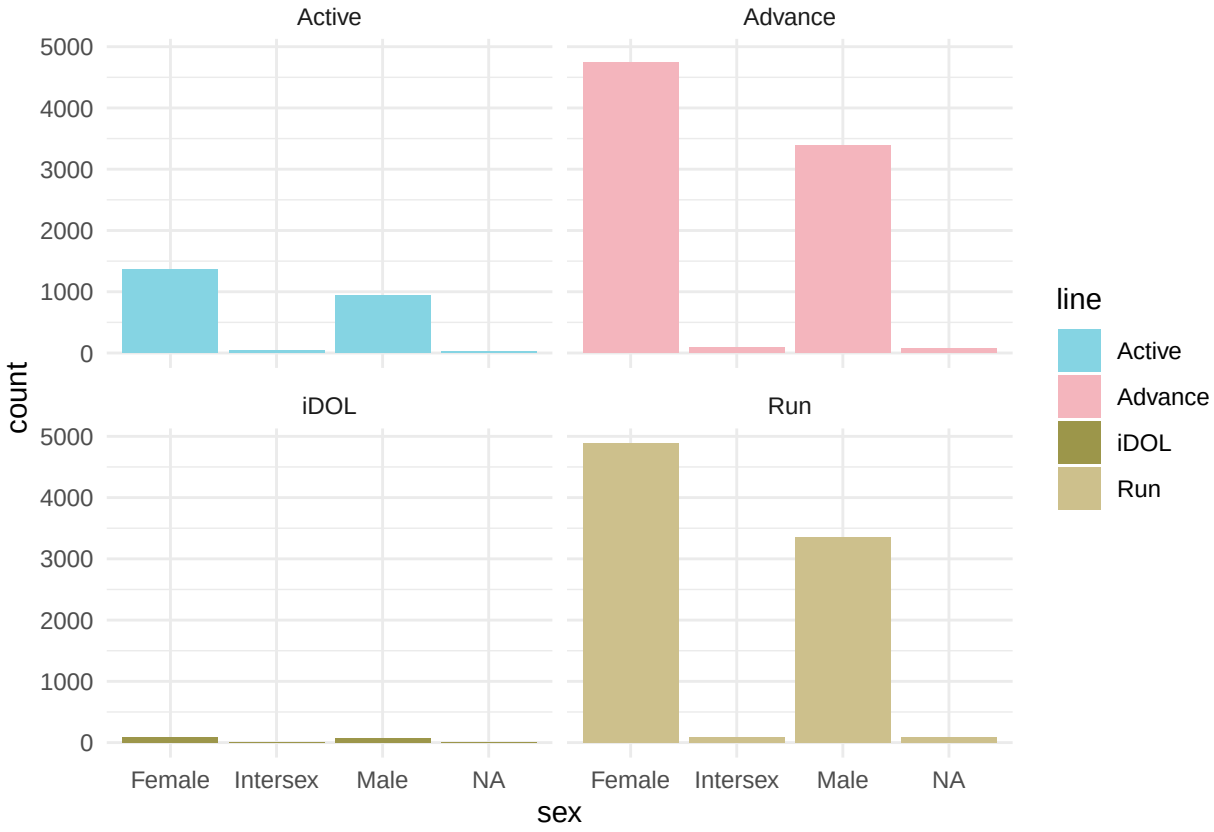


Figure 3: Sex of customer by device line

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.

```
device_cust$sex_binary <-  ifelse(device_cust$sex == "Female", 1, 0)
device_cust$line_binary <- ifelse(device_cust$line %in% c("Advance", "Active"), "new",
↪  "old")
t.test(sex_binary~ line_binary, data = device_cust, var.equal = TRUE)
```

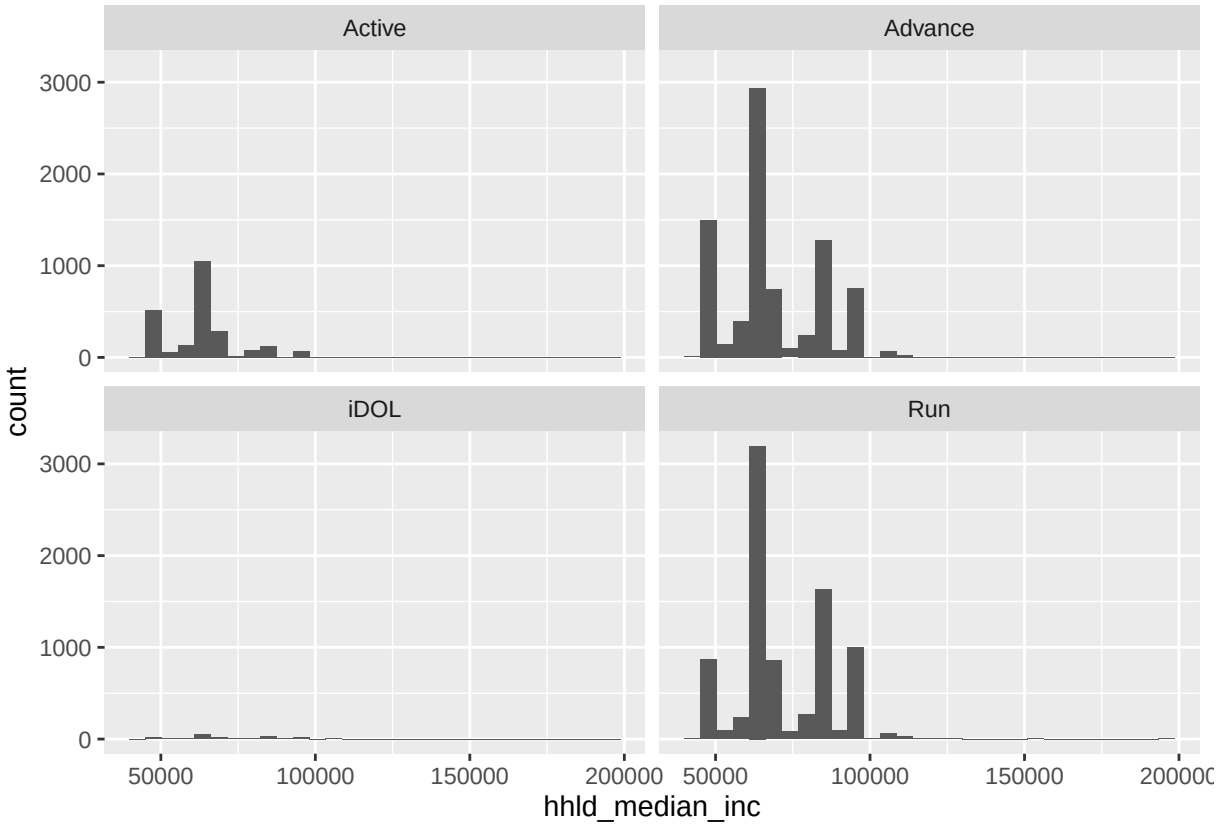
```
##
##  Two Sample t-test
##
## data:  sex_binary by line_binary
## t = -1.127, df = 19055, p-value = 0.2598
## alternative hypothesis: true difference in means between group new and group old is not e
## 95 percent confidence interval:
##  -0.022194593  0.005989754
## sample estimates:
## mean in group new mean in group old
##         0.5783850         0.5864874
```

```
income <-  read.csv("./data/median_income.csv")
group_income <- inner_join(device_cust, income) %>% select(c("line",
↪  "hhld_median_inc"))
```

```
## Joining, by = "CSDuid"
```

```
group_income %>% ggplot(aes(x= hhld_median_inc)) + geom_histogram() +
↪  facet_wrap("line")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
mod <- lm(hhld_median_inc ~ line, data = group_income)
summary(mod)
```

```
##
## Call:
## lm(formula = hhld_median_inc ~ line, data = group_income)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31266  -7317  -4115  12835 125626
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  64845.2      299.2 216.739  < 2e-16 ***
## lineAdvance   5098.5      339.0  15.038  < 2e-16 ***
## lineiDOL      9389.6     1187.8   7.905 2.83e-15 ***
## lineRun       8300.7      338.6  24.517  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14540 on 19249 degrees of freedom
## Multiple R-squared:  0.0328, Adjusted R-squared:  0.03265
## F-statistic: 217.6 on 3 and 19249 DF,  p-value: < 2.2e-16
```

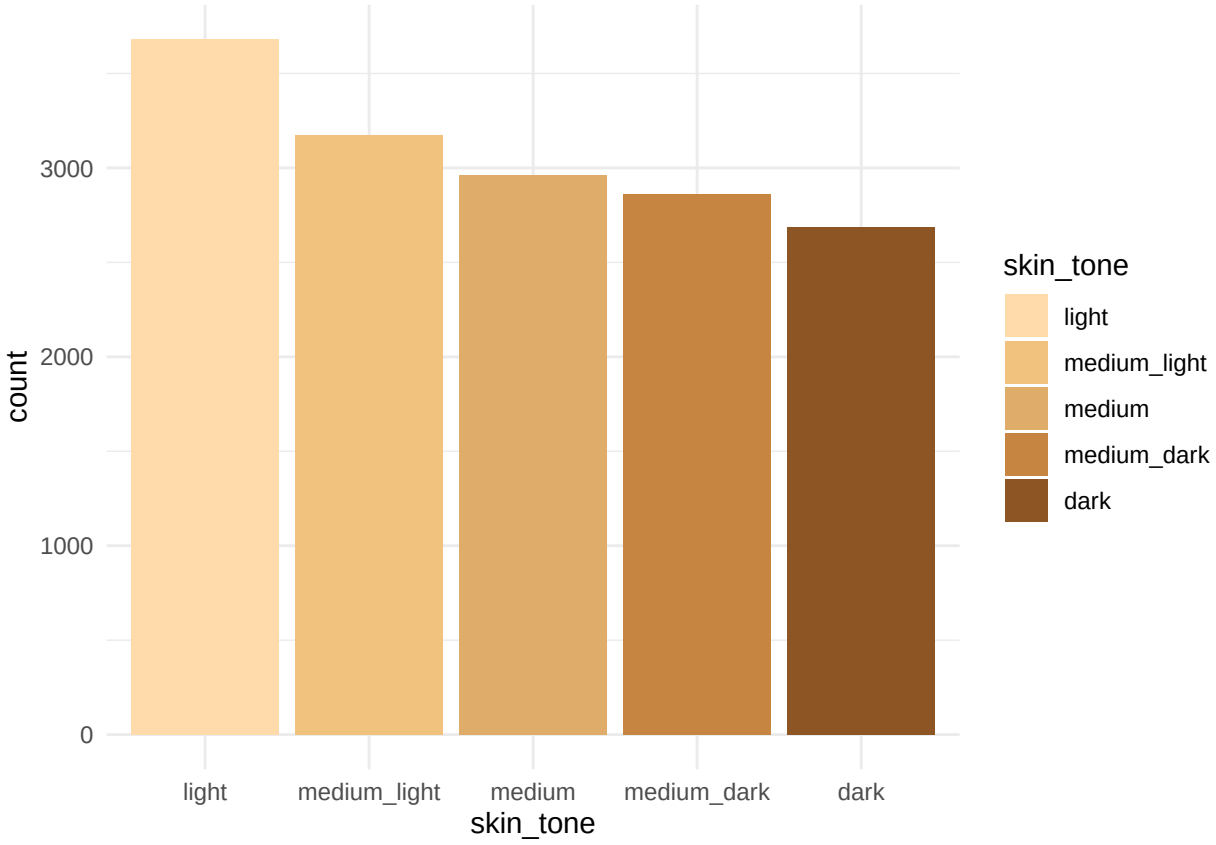
```
summary(aov(hhld_median_inc ~ line, data = group_income))
```

```
##                 Df   Sum Sq   Mean Sq F value Pr(>F)
## line             3 1.38e+11 4.601e+10   217.6 <2e-16 ***
## Residuals    19249 4.07e+12 2.114e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How does race affect the accuracy of wearable devices?

```
race_sleep_data <- read.csv("./data/race_sleep_data.csv")
skin_tone_palette <- c("#ffdbac", "#f1c27d", "#e0ac69", "#c68642", "#8d5524")

race_sleep_data <- race_sleep_data %>%
  mutate(skin_tone = fct_relevel(skin_tone, "light", "medium_light", "medium",
↪  "medium_dark", "dark"))
```

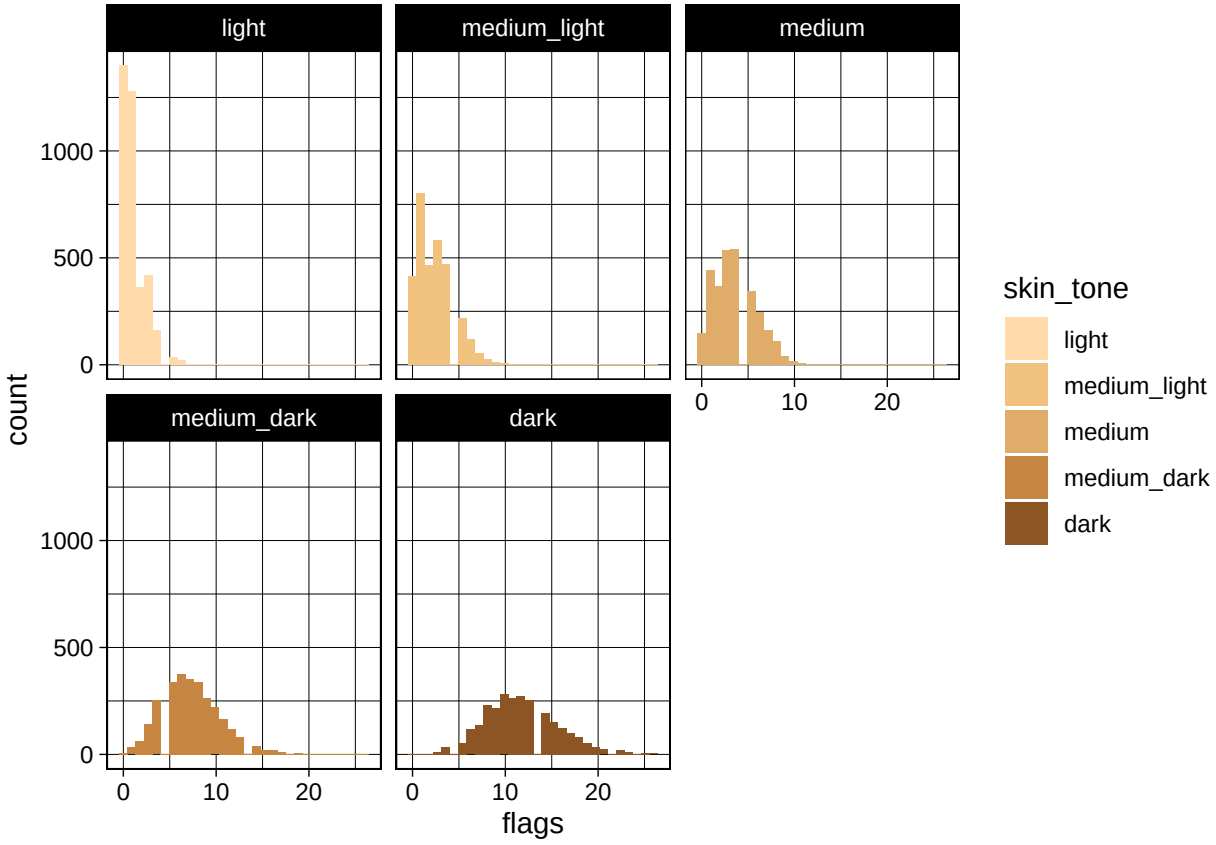
```
# total number of individuals for each race
race_sleep_data %>% ggplot(aes(skin_tone, fill = skin_tone)) + geom_bar() +
↪  scale_fill_manual(values = skin_tone_palette) +theme_minimal()
```




```
race_sleep_data %>% group_by(skin_tone) %>% summarise(count = n())
```

```
## # A tibble: 5 x 2
##   skin_tone    count
##   <fct>        <int>
## 1 light         3680
## 2 medium_light  3173
## 3 medium        2962
## 4 medium_dark   2862
## 5 dark          2687
```

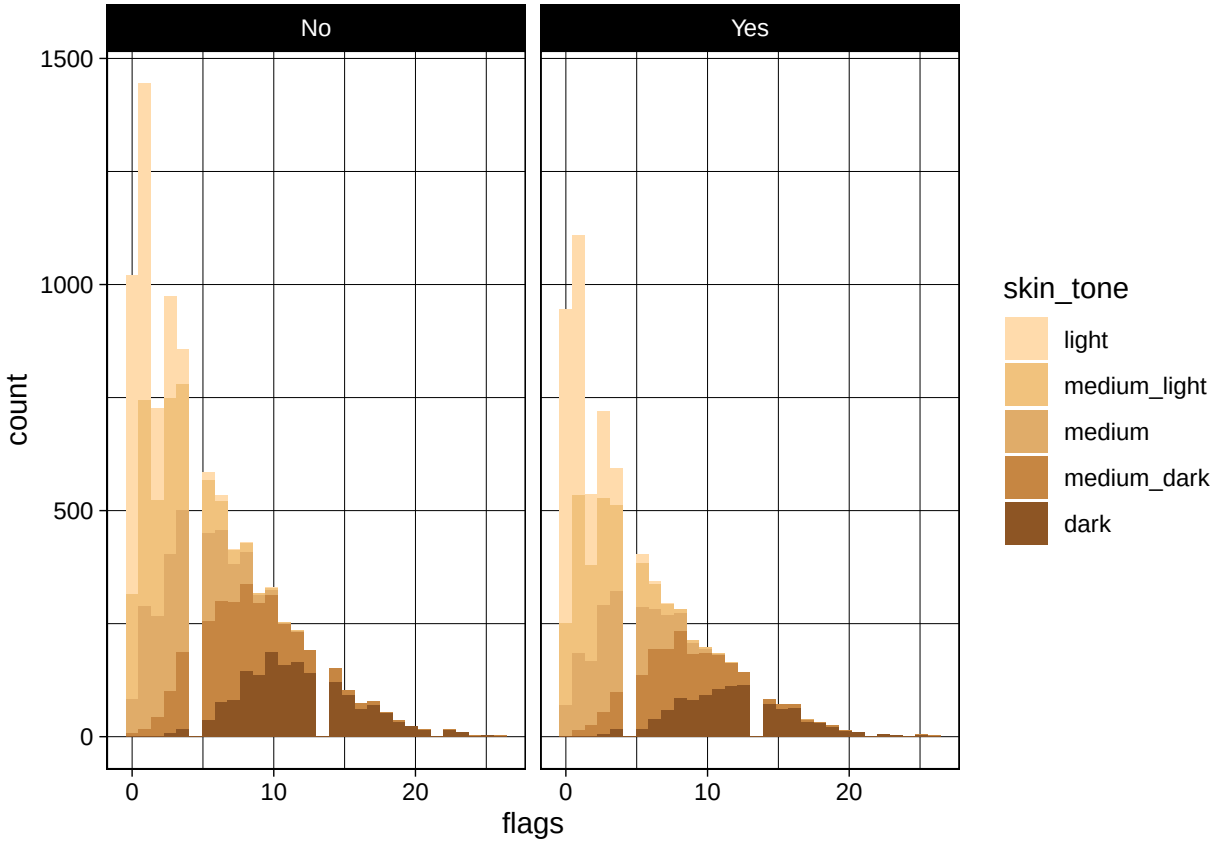
```
# distribution of number of flags for each race
race_sleep_data %>% ggplot(aes(x = flags, fill=skin_tone)) + geom_histogram(bins = 30)
↪  + scale_fill_manual(values = skin_tone_palette) + facet_wrap("skin_tone") +
↪  theme_linedraw()
```



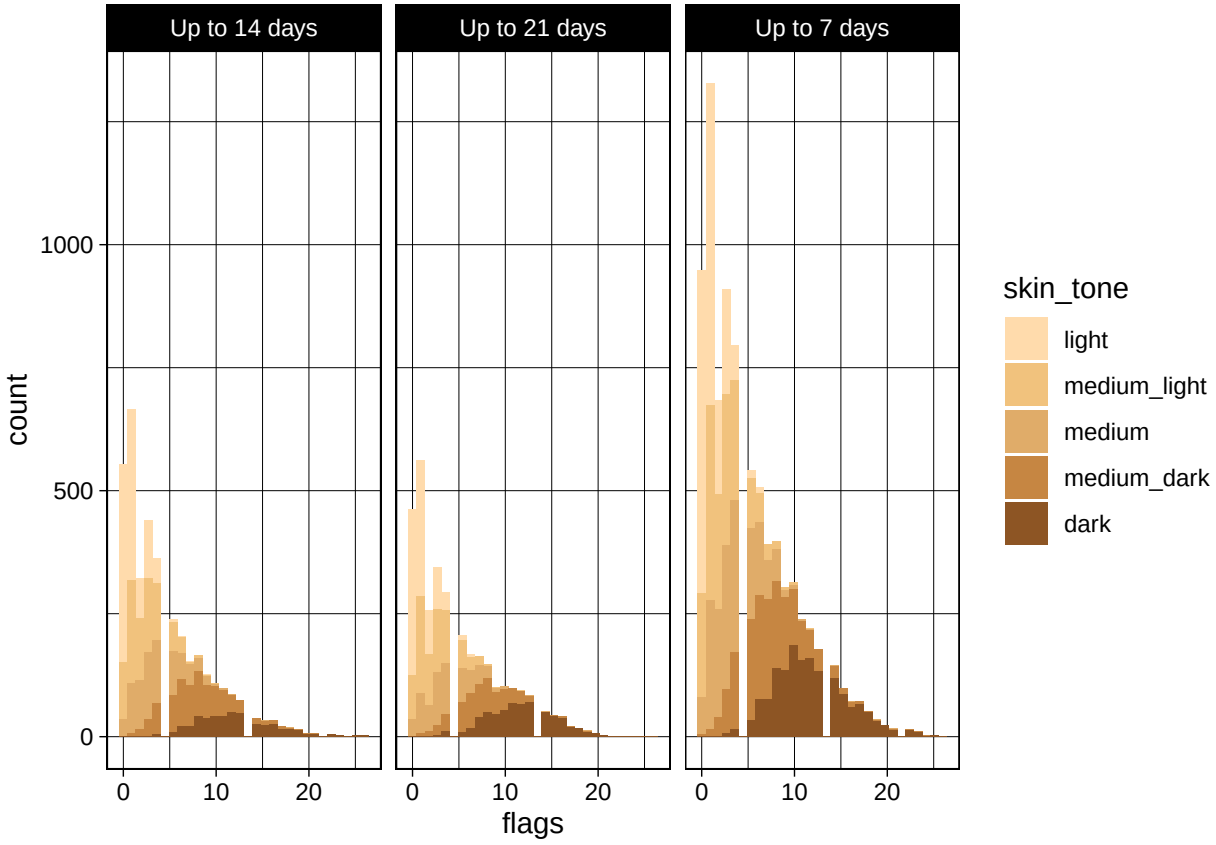

```
# number of individuals using each line for each race by sex
race_sleep_data %>% ggplot(aes(line, fill = skin_tone)) + facet_wrap("sex") +
↪  geom_bar() + scale_fill_manual(values = skin_tone_palette) + theme_minimal()
```


```
# race_sleep_data %>% ggplot(aes(sex, fill = skin_tone)) + geom_bar() +
↪  scale_fill_manual(values = skin_tone_palette) + theme_minimal()
```

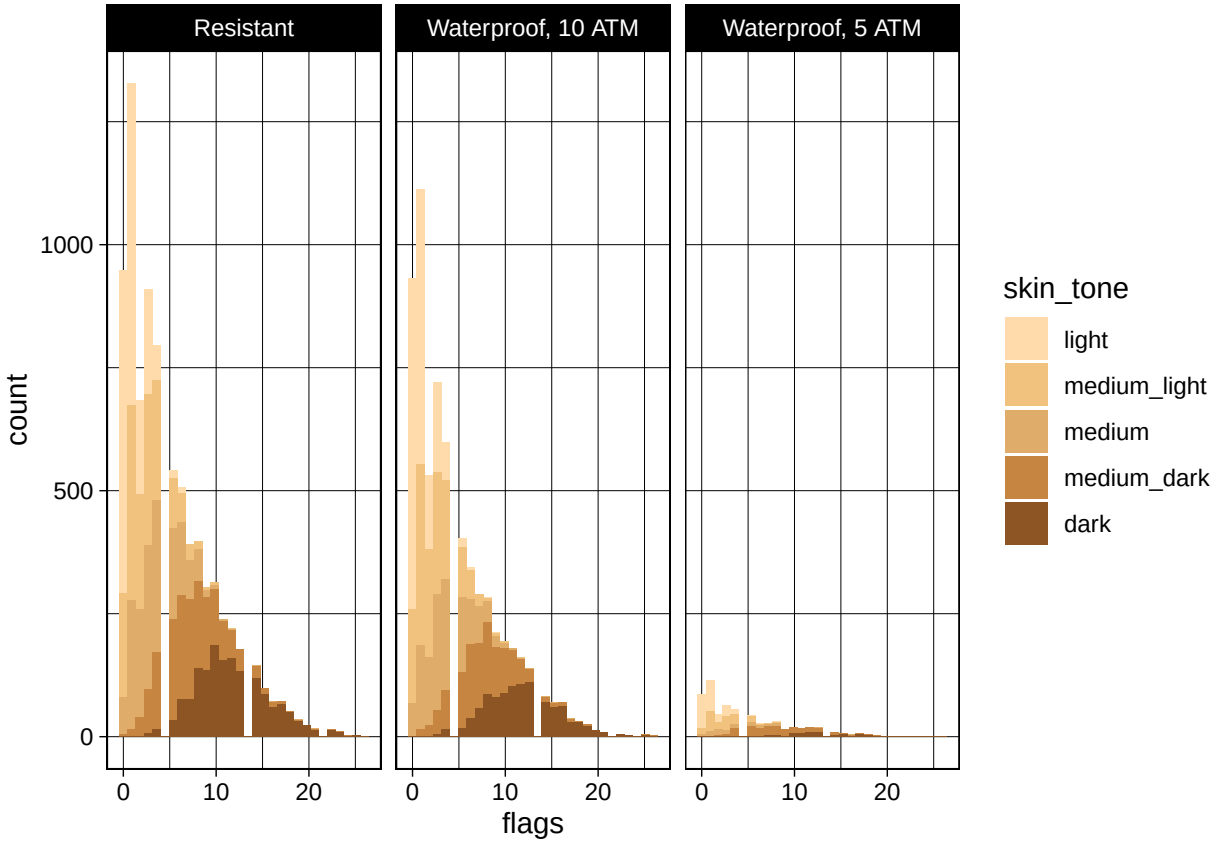
```
# number of flags for whether there is a pulse oximiter, coloured by skin tone
race_sleep_data %>% ggplot(aes(x = flags, fill = skin_tone)) +
↪  scale_fill_manual(values = skin_tone_palette)+ geom_histogram(bins = 30) +
↪  facet_wrap("pulse_oximiter") + theme_linedraw()
```



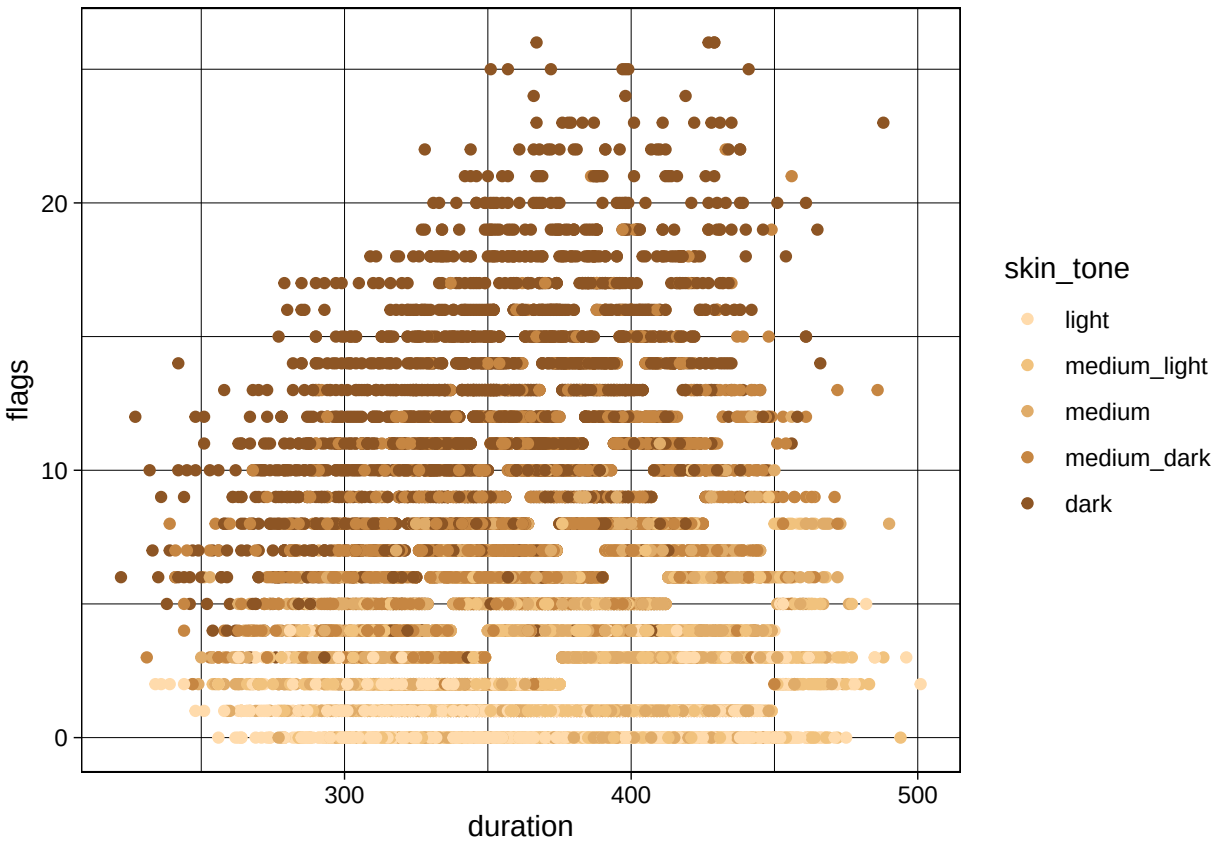

```
# number of flags for each battery_life, coloured by skin tone
race_sleep_data %>% ggplot(aes(x = flags, fill = skin_tone)) +
↪  scale_fill_manual(values = skin_tone_palette)+ geom_histogram(bins = 30) +
↪  facet_wrap("battery_life") + theme_linedraw()
```



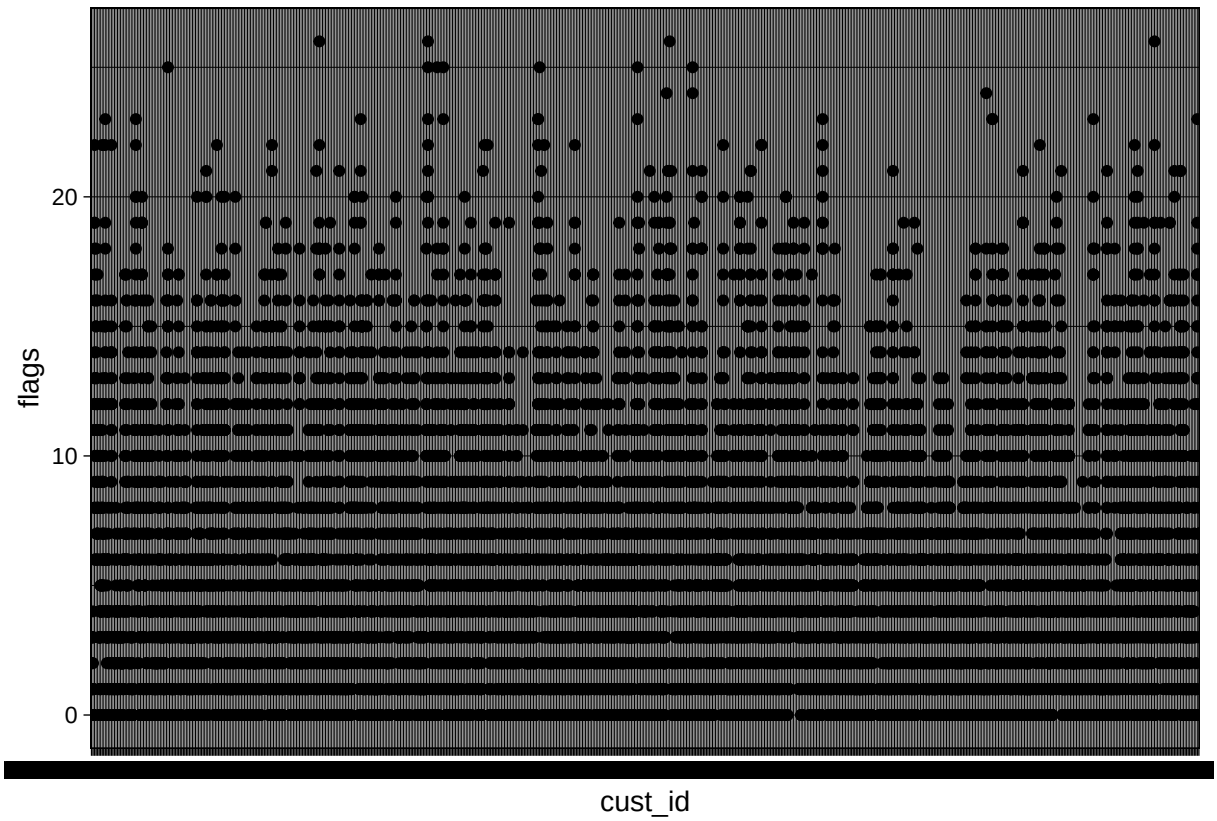

```
# number of flags for each water resistence type, coloured by skin tone
race_sleep_data %>% ggplot(aes(x = flags, fill = skin_tone)) +
↪  scale_fill_manual(values = skin_tone_palette)+ geom_histogram(bins = 30) +
↪  facet_wrap("water_resitance") + theme_linedraw()
```



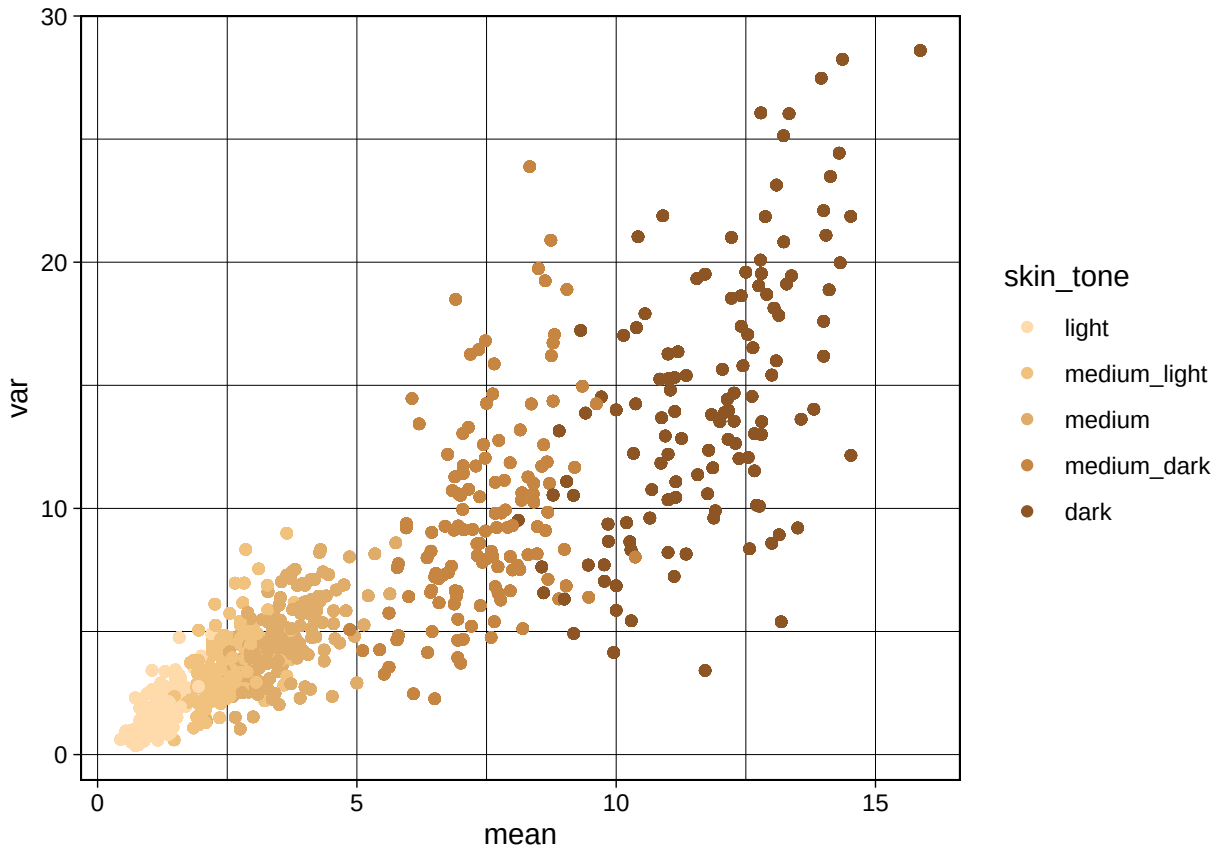

```
# number of flags v.s. duration, coloured by skin tone
race_sleep_data %>% ggplot(aes(x = duration, y = flags, colour = skin_tone)) +
↪  geom_point()+ scale_colour_manual(values = skin_tone_palette) + theme_linedraw()
```




```
race_sleep_data %>% ggplot(aes(x = cust_id, y = flags)) + geom_point() +
↪  theme_linedraw()
```




```
temp_race_sleep_data <- race_sleep_data %>% group_by(cust_id) %>% mutate(mean =
↪  mean(flags)) %>% mutate(var = var(flags))

temp_race_sleep_data %>% ggplot(aes(x = mean, y = var, colour = skin_tone)) +
↪  geom_point()+ scale_colour_manual(values = skin_tone_palette) + theme_linedraw()
```



Observations: - more light skinned individuals than dark skinned individuals

- darker skinned individuals seem to have a larger spread and larger mean of number of flags than compared to lighter skinned individuals
- more females than males wearing devices (might need to do this in a different dataset)
- most individuals wear the Advance or Run line
- battery life, water resistance, and whether or not there is a pulse oximiter seem to not affect the number of flags
- not much relationship between duration of sleep and number of flags

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 come from a normal distribution ????

3. Random Effect errors have constant variance

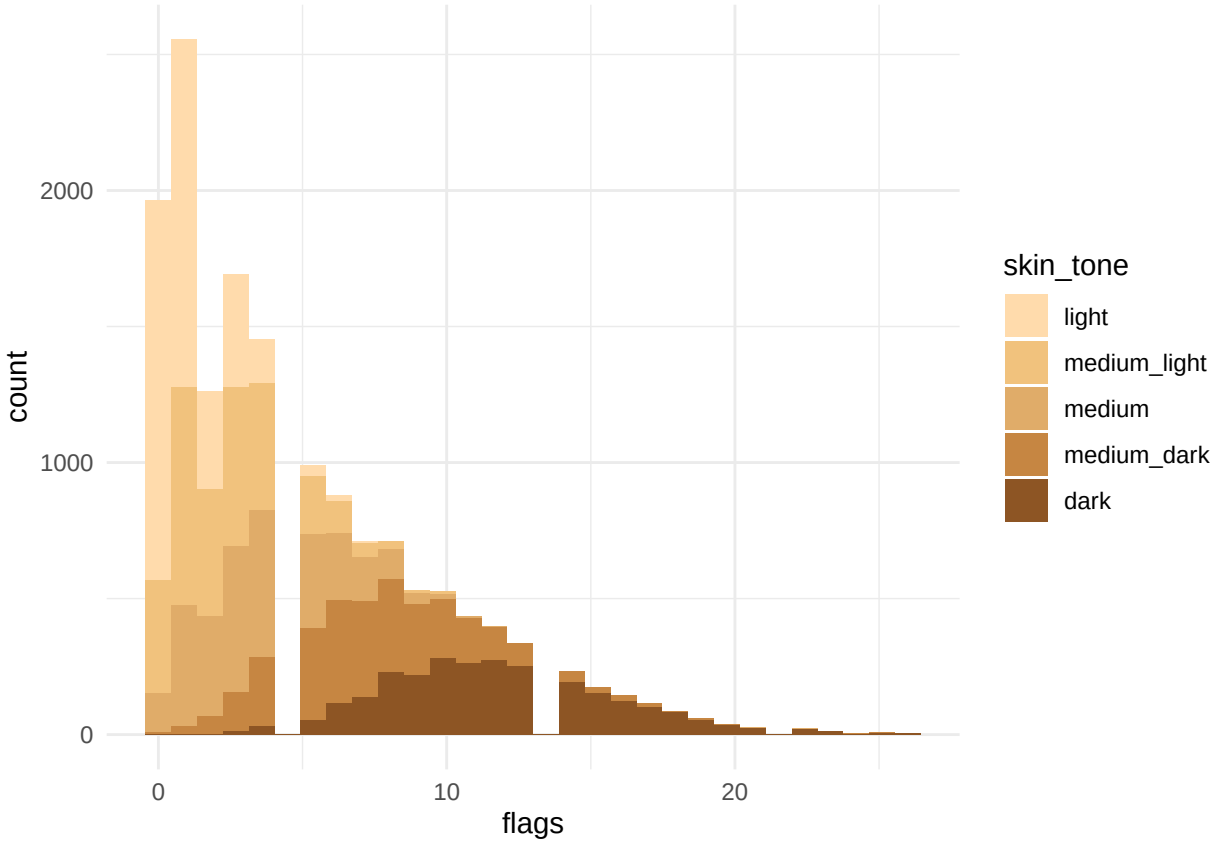
```
temp_race_sleep_data <- race_sleep_data %>%
  filter(!(flags == 0)) %>%
  group_by(skin_tone) %>%
  summarise(var_of_log_flags = var(log(flags)))

temp_race_sleep_data
```

```
## # A tibble: 5 x 2
##   skin_tone    var_of_log_flags
##   <fct>                   <dbl>
## 1 light                   0.307
## 2 medium_light            0.412
## 3 medium                  0.408
## 4 medium_dark             0.233
## 5 dark                    0.133
```

We see that the variance of the number of flags transformed by the link function (log) is homogeneous across skin tones. This suggests that the homoscedasiticy of random effect errors assumption is satisfied

4. Poisson link function is appropriate:

```
# check distribution of response variable
race_sleep_data %>% ggplot(aes(flags, fill = skin_tone)) + geom_histogram(bins = 30) +
↪  scale_fill_manual(values = skin_tone_palette) +theme_minimal()
```



From the plot, we see that the number of flags (response variable) is ranges from 0 to over 20, is right skewed, and can be modeled by Poisson distribution.

```
# check mean and variance of number of flags for each skin tone (main effect)
race_sleep_data %>% group_by(skin_tone) %>% dplyr::summarize(mean = mean(flags), var =
↪  var(flags), n= n())
```

```
## # A tibble: 5 x 4
##   skin_tone     mean   var     n
##   <fct>        <dbl> <dbl> <int>
## 1 light         1.16  1.68  3680
## 2 medium_light  2.50  3.60  3173
## 3 medium        3.65  4.98  2962
## 4 medium_dark   7.43 10.2   2862
## 5 dark         11.8  16.1   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.

Check that log(response) is a linear function of the predictor???

GLMM Model Fitting

```
race_sleep_data <- race_sleep_data %>% filter(!is.na(sex))

race1<- lme4::glmer(flags ~ skin_tone + (1 | cust_id),
        family='poisson', data=race_sleep_data)
summary(race1)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: flags ~ skin_tone + (1 | cust_id)
##    Data: race_sleep_data
##
##      AIC      BIC   logLik deviance df.resid
##  65651.1  65696.9 -32819.5  65639.1    15237
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.8465 -0.9410 -0.1404  0.7205  5.9705
##
## Random effects:
##  Groups  Name        Variance Std.Dev.
##  cust_id (Intercept) 0.01629  0.1276
## Number of obs: 15243, groups:  cust_id, 719
##
## Fixed effects:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            0.13640    0.01825   7.473 7.81e-14 ***
## skin_tonemedium_light  0.77566    0.02391  32.435  < 2e-16 ***
## skin_tonemedium        1.14922    0.02331  49.296  < 2e-16 ***
## skin_tonemedium_dark   1.86223    0.02240  83.130  < 2e-16 ***
## skin_tonedark          2.32202    0.02226 104.293  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Correlation of Fixed Effects:
##             (Intr) skn_tnmdm_l skn_tnm skn_tnmdm_d
## skn_tnmdm_l -0.762
## skin_tonmdm -0.782  0.597
## skn_tnmdm_d -0.814  0.621       0.637
## skin_tondrk -0.819  0.625       0.641    0.667
```

```
race2<- lme4::glmer(flags ~ skin_tone + sex + (1 | cust_id),
       family='poisson', data=race_sleep_data)
summary(race2)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: flags ~ skin_tone + sex + (1 | cust_id)
##    Data: race_sleep_data
##
##      AIC      BIC   logLik deviance df.resid
##  65626.3  65687.3 -32805.1  65610.3    15235
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.8542 -0.9445 -0.1466  0.7185  6.1136
##
## Random effects:
##  Groups  Name        Variance Std.Dev.
##  cust_id (Intercept) 0.01507  0.1228
## Number of obs: 15243, groups:  cust_id, 719
##
## Fixed effects:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            0.16504    0.01877   8.793  < 2e-16 ***
## skin_tonemedium_light  0.77520    0.02359  32.861  < 2e-16 ***
## skin_tonemedium        1.14868    0.02297  50.001  < 2e-16 ***
## skin_tonemedium_dark   1.86061    0.02204  84.411  < 2e-16 ***
## skin_tonedark          2.32240    0.02188 106.124  < 2e-16 ***
## sexIntersex           -0.09662    0.06563  -1.472    0.141
```

```
## sexMale               -0.06918    0.01300  -5.323 1.02e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) skn_tnmdm_l skn_tnm skn_tnmdm_d skn_tnd sxIntr
## skn_tnmdm_l -0.736
## skin_tonmdm -0.756  0.601
## skn_tnmdm_d -0.790  0.626       0.643
## skin_tondrk -0.791  0.631       0.648   0.675
## sexIntersex -0.055  0.010       0.011  -0.013       0.004
## sexMale     -0.271  0.000       0.001   0.012      -0.009   0.077
```

```
# race3<- glm(flags ~ skin_tone + sex + line,
#        family='poisson', data=race_sleep_data)
# summary(race3)

lmtest::lrtest(race1, race2)
```

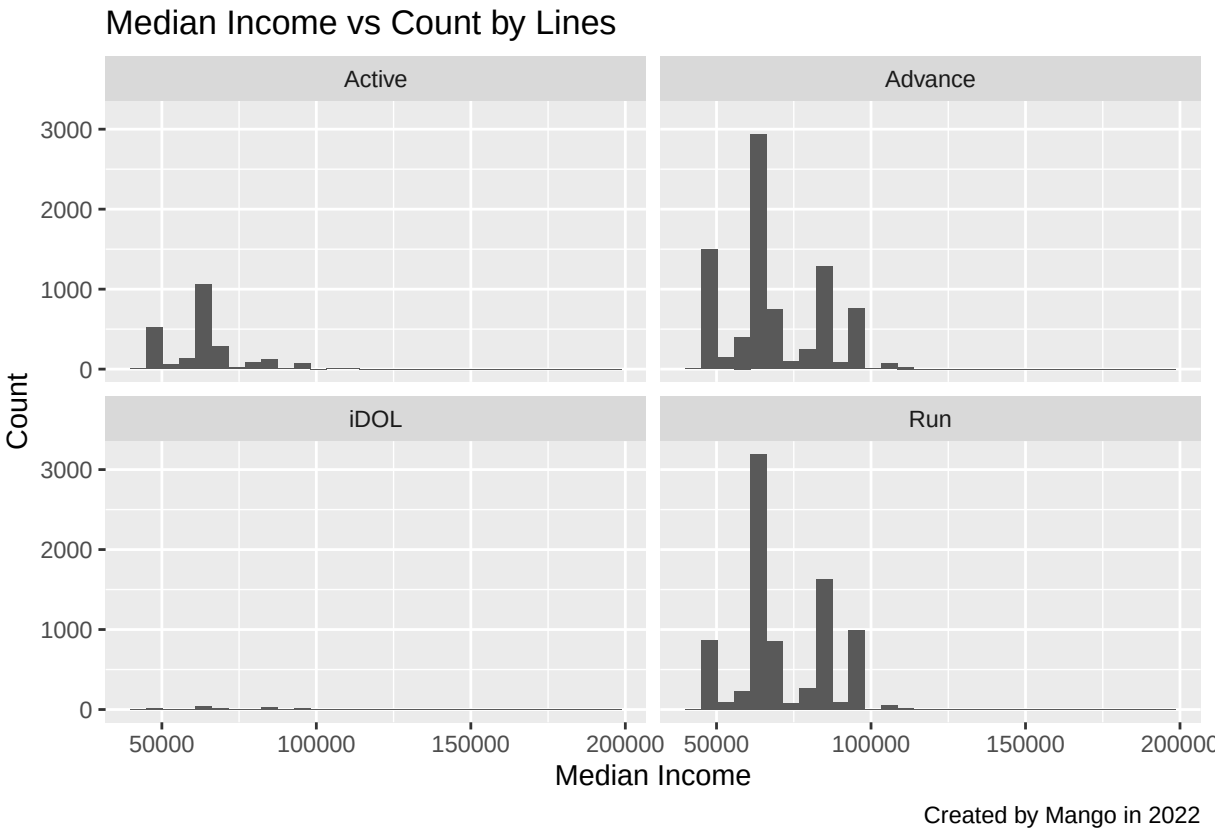
```
## Likelihood ratio test
##
## Model 1: flags ~ skin_tone + (1 | cust_id)
## Model 2: flags ~ skin_tone + sex + (1 | cust_id)
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   6 -32820
## 2   8 -32805  2 28.831  5.488e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Likelihood Ratio Test shows that we have significant evidence (p value = 5.49e-07) against the fact that the simpler model explains the data just as well as the complex model. Hence, we prefer race2 over race1. From race2, INTERPRET THE MODEL HEREEE

```
income <-  read.csv("./data/median_income.csv")
group_income <- inner_join(device_cust, income) %>% select(c("line",
↪  "hhld_median_inc"))
```

```
## Joining, by = "CSDuid"
```

```
group_income %>%
  ggplot(aes(x= hhld_median_inc)) +
  geom_histogram(bins = 30) +
  facet_wrap("line")+
  labs(title = "Median Income vs Count by Lines",
       x = "Median Income",
       y = "Count",
       caption = "Created by Mango in 2022")
```




```
kruskal.test(hhld_median_inc ~ line, data = group_income)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  hhld_median_inc by line
## Kruskal-Wallis chi-squared = 610.04, df = 3, p-value < 2.2e-16
```

```r
mod <- lm(hhld_median_inc ~ line, data = group_income)
summary(mod)
```

```
##
## Call:
## lm(formula = hhld_median_inc ~ line, data = group_income)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31266  -7317  -4115  12835 125626
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  64845.2      299.2 216.739  < 2e-16 ***
## lineAdvance   5098.5      339.0  15.038  < 2e-16 ***
## lineiDOL      9389.6     1187.8   7.905 2.83e-15 ***
## lineRun       8300.7      338.6  24.517  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14540 on 19249 degrees of freedom
## Multiple R-squared:  0.0328, Adjusted R-squared:  0.03265
## F-statistic: 217.6 on 3 and 19249 DF,  p-value: < 2.2e-16
```

```r
#
# summary(aov(hhld_median_inc ~ line, data = group_income))
```

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

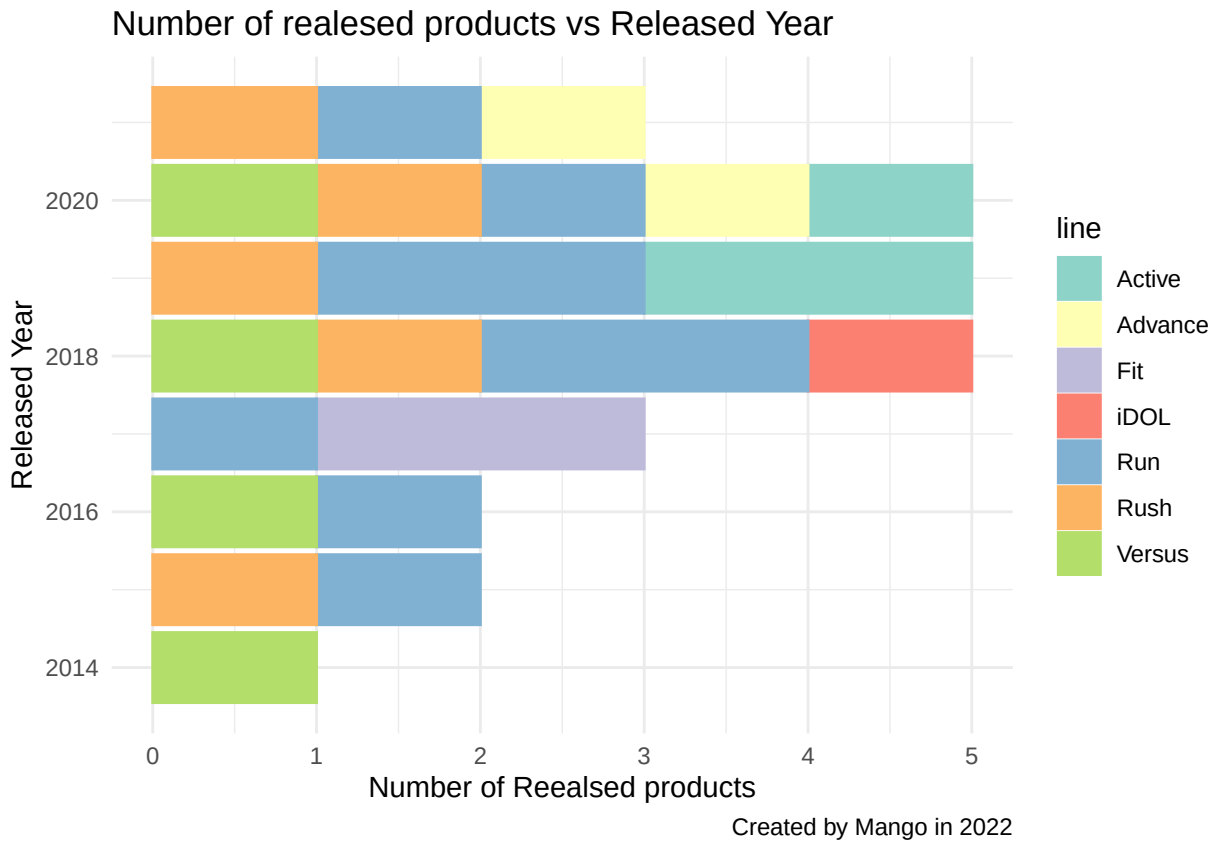
- Use wearable device data
- performance/price difference among the different lines of services, combined with the different demographic of the users between lines, we can give a report summary/advice for what has changed and how the change affects the target audience

```r
device_data <- read.csv("./data/device_data.csv")
device_cust <- read.csv("./data/device_cust.csv")
device_performance <- read.csv("./data/device_performance.csv")
```

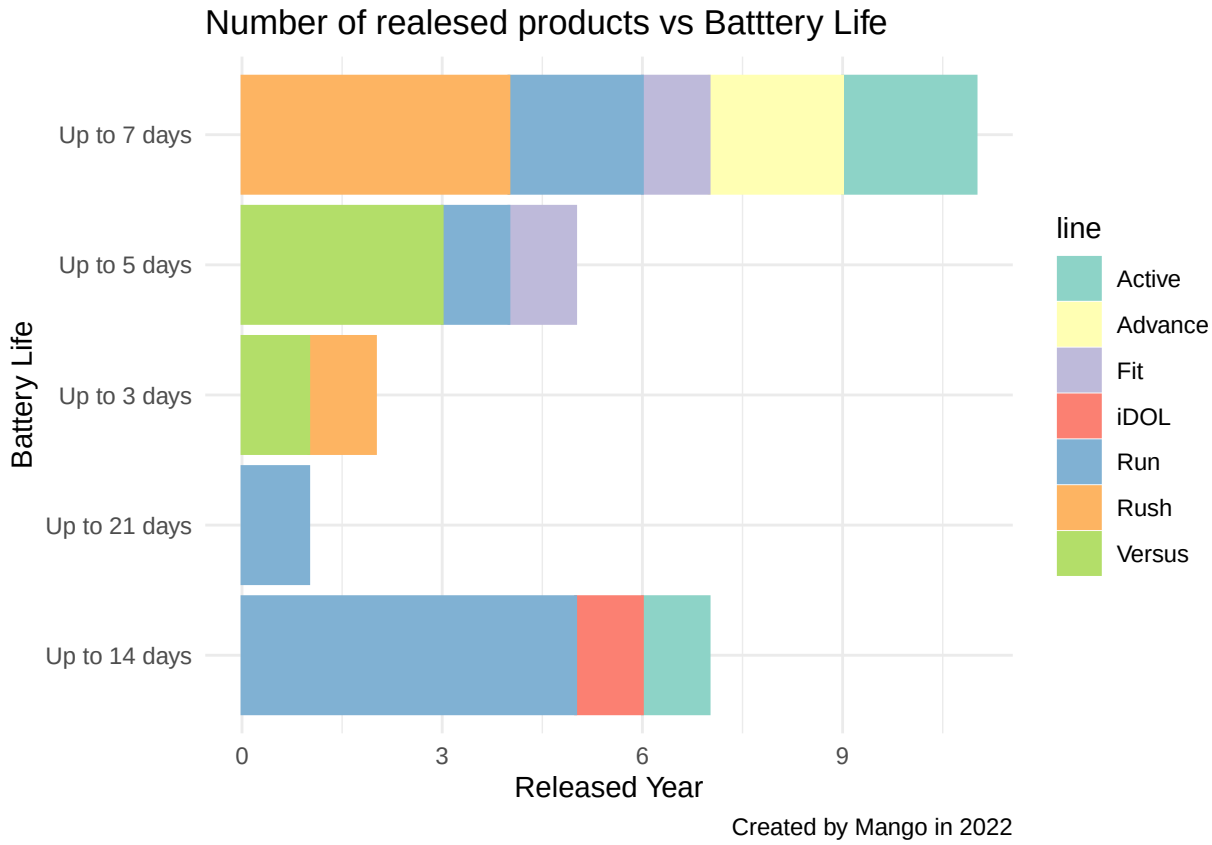
```
device_data %>%
  ggplot(aes(x = released_yr, y = recommended_retail_price, color = line, shape =
↪  brand)) +
  geom_point(size = 4) +
  scale_color_brewer(palette="Set3") +
  labs(title = "Recommeded Retail price vs Released Date",
       x = "Released Date",
       y = "Recommended Retail Price",
       caption = "Created by Mango in 2022")
```


```
device_data %>%
  ggplot(aes(y = released_yr, color = line, fill = line)) +
  geom_bar(stat="count") +
  scale_color_brewer(palette="Set3") +
  scale_fill_brewer(palette="Set3") +
  labs(title = "Number of realesed products vs Released Year",
       y = "Released Year",
       x = "Number of Reealsed products",
       caption = "Created by Mango in 2022") +
```

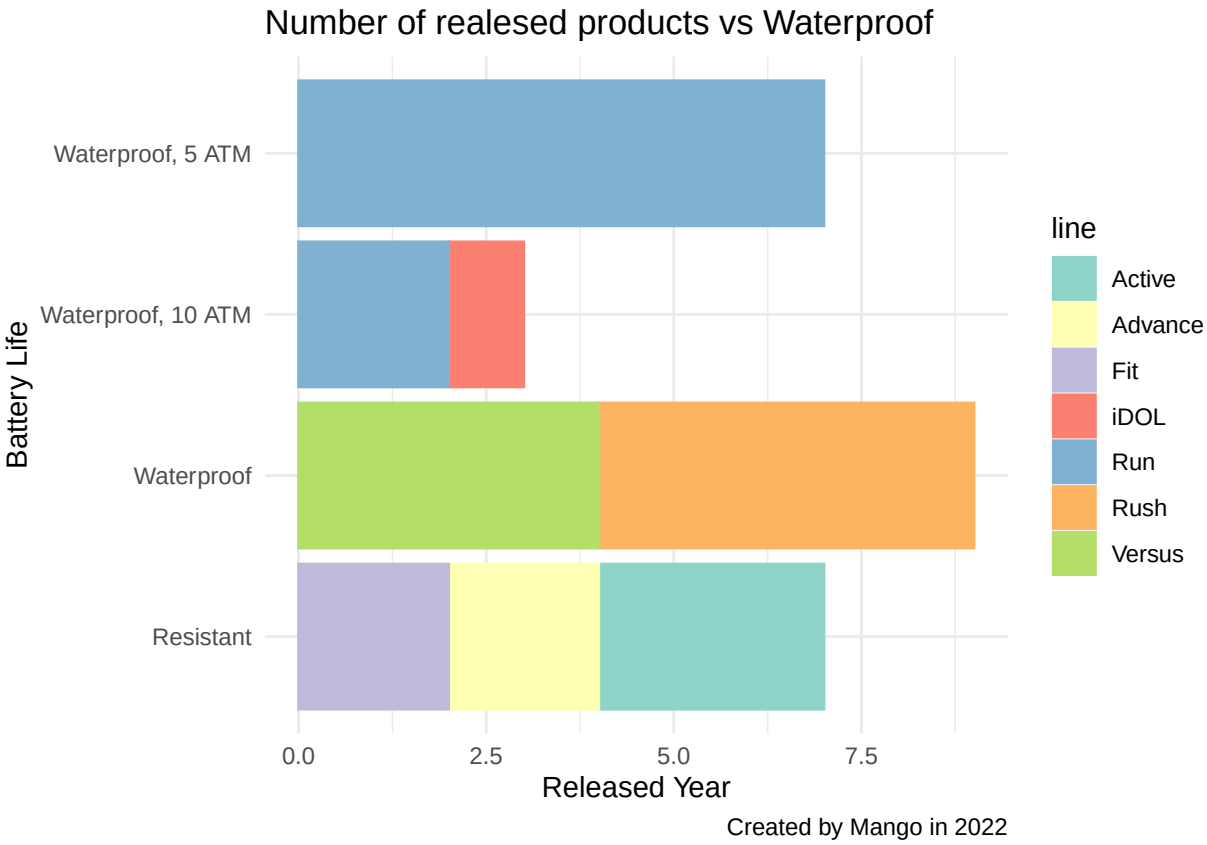
```
  theme_minimal()
```



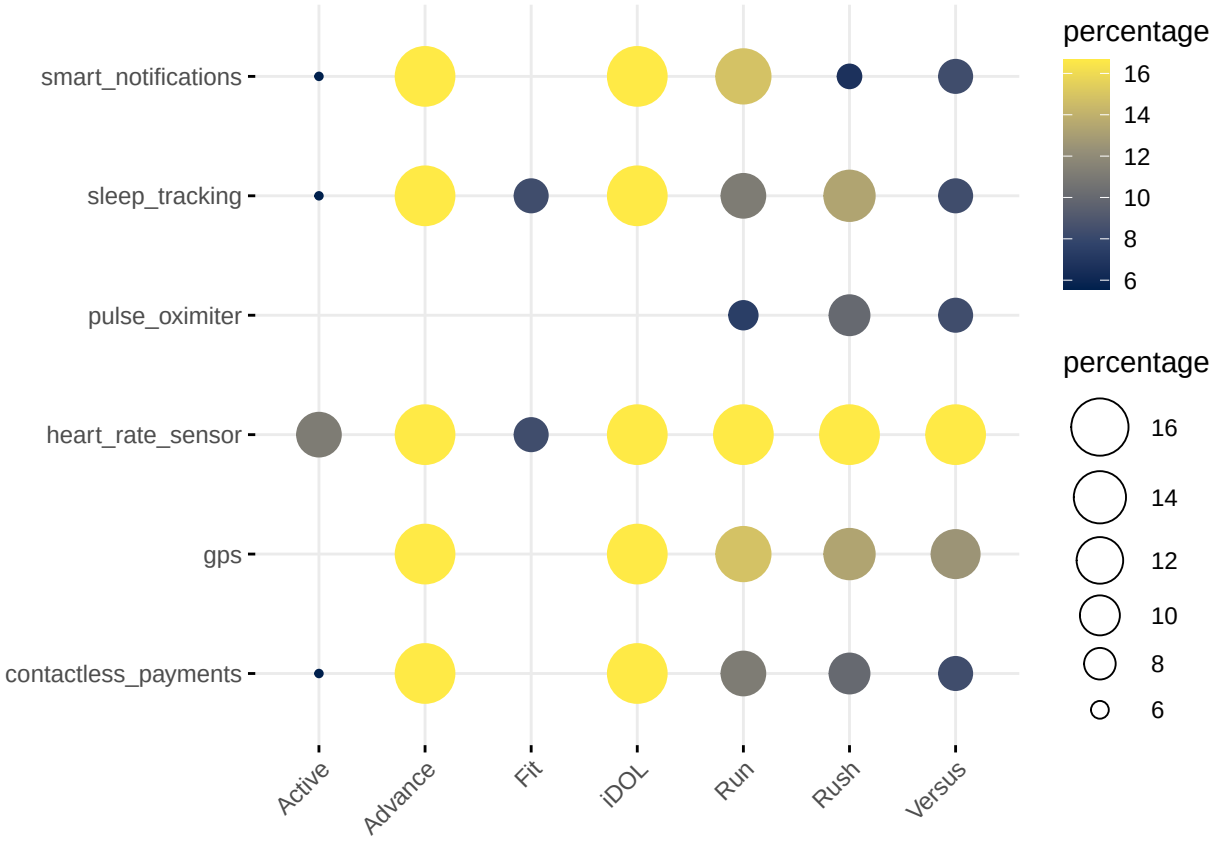

```
# Batterty Lives
device_data %>%
  ggplot(aes(y = battery_life, color = line, fill = line)) +
  geom_bar(stat="count") +
  scale_color_brewer(palette="Set3") +
  scale_fill_brewer(palette="Set3") +
  labs(title = "Number of realesed products vs Batttery Life",
       x = "Released Year",
       y = "Battery Life",
       caption = "Created by Mango in 2022") +
  theme_minimal()
```



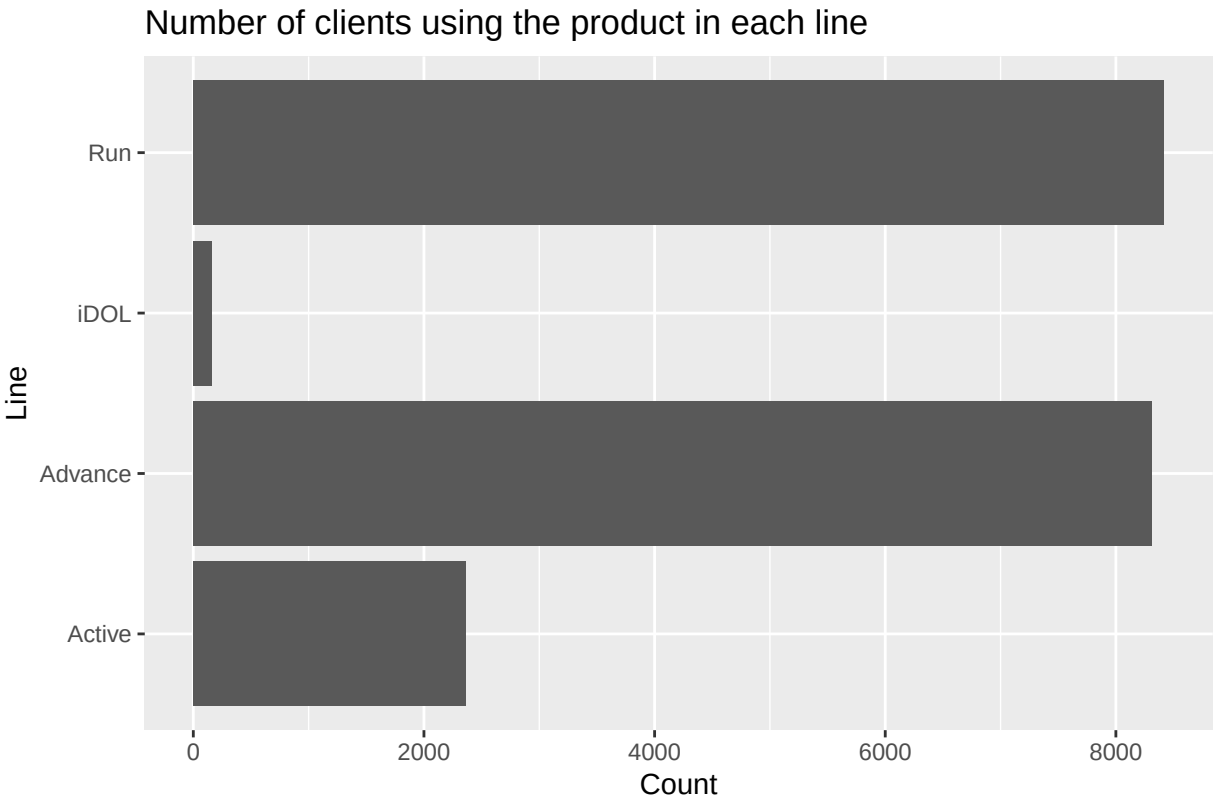

```
# Waterproof
device_data %>%
  ggplot(aes(y = water_resitance, color = line, fill = line)) +
  geom_bar(stat="count") +
  scale_color_brewer(palette="Set3") +
  scale_fill_brewer(palette="Set3") +
  labs(title = "Number of realesed products vs Waterproof",
       x = "Released Year",
       y = "Battery Life",
       caption = "Created by Mango in 2022") +
  theme_minimal()
```




```
# Other performance
library(ggpubr)
ggballoonplot(device_performance,
              x="line",
              y="name",
              size = "percentage",
              fill = "percentage",
              color = "percentage") +
  scale_fill_viridis_c(option = "E") +
  scale_color_viridis_c(option = "E")
```




```
device_cust %>%
  ggplot(aes(y=line))+
  geom_bar() +
  labs(title = "Number of clients using the product in each line",
       x = "Count",
       y = "Line",
       caption = "Created by Mango in 2022")
```



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!