dsci_project_nirushanbhag

Niru Shanbhag

4/23/2023

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.2     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.2     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflic
ts to become errors
```

```
library(dplyr)
library(ggplot2)
library(lubridate)
```

FINAL PROJECT

dataset = https://www.kaggle.com/datasets/rodsaldanha/arketing-campaign/discussion/182997 (https://www.kaggle.com/datasets/rodsaldanha/arketing-campaign/discussion/182997)

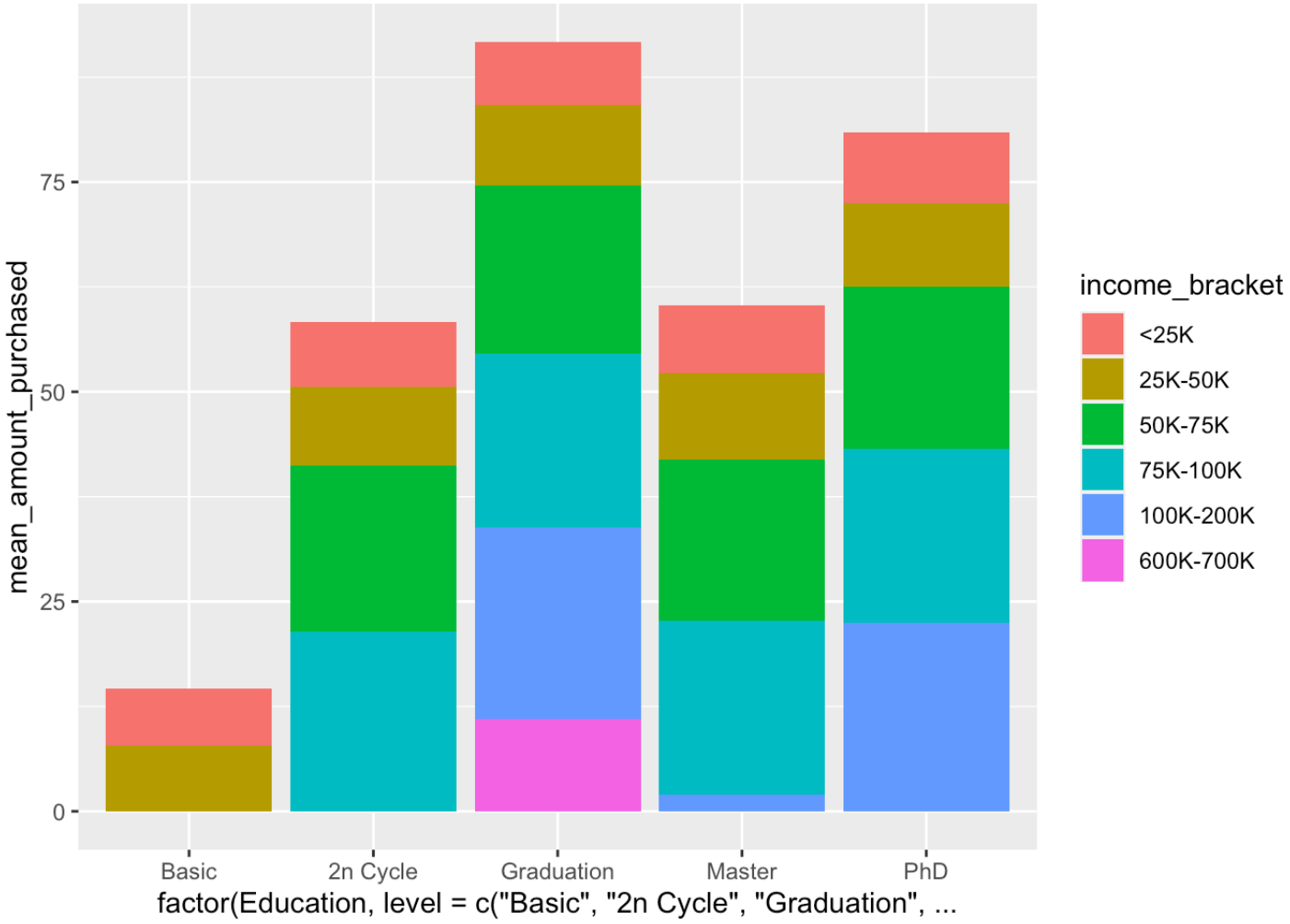
```
#Preparing the data
cust <- read.csv("/Users/nirushanbhag/Downloads/DSCI101/DSCI101/data/marketing_campai
gn.csv", sep = "\t")
cust <- cust %>% filter(!is.na(Income)) #remove all the Incomes that are NA
cust$Dt_Customer = dmy(cust$Dt_Customer)
cust <- cust %>% mutate(income_bracket =
                  cut(Income, breaks = c(-Inf, 25000, 50000, 75000, 100000, 200000, 3
00000, 400000, 500000, 600000,700000),
                  include.lowest = FALSE,
                  labels = c("<25K", "25K-50K", "50K-75K", "75K-100K", "100K-200K", "
200K-300K", "300K-400K", "400K-500K", "500K-600K", "600K-700K")),
                  no_offspring = Teenhome + Kidhome,
                  month_of_enrollment = month(Dt_Customer),
                  year_at_enrollment= year(Dt_Customer),
                  age_at_enrollment = year_at_enrollment - Year_Birth,
                  age_bracket =
                  cut(age_at_enrollment, breaks = c(-Inf, 19, 29, 39, 49, 59, 69, 79,
89, 99, Inf), labels = c("Teens", "20s", "30s", "40s", "50s", "60s", "70s", "80s", "9
0s", "100s")),
                  total_amount_purchased = NumWebPurchases+NumCatalogPurchases+NumSto
rePurchases+NumDealsPurchases
)
```

1a. Visualize the education level and mean amount purchased in a stacked bar chart, and color code by marital status

From this you can see that graduates make the most purchases, and they also have the most diverse income bracket.

```
cust %>% group_by(Education, income_bracket) %>% summarize(mean_amount_purchased = me
an(total_amount_purchased)) %>% ggplot(aes(x = factor(Education, level = c("Basic", "
2n Cycle", "Graduation", "Master", "PhD")), y = mean_amount_purchased, fill = income_
bracket)) + geom_bar(stat = "identity")
```

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



1b.Find the group within education, marital_status, and income that have the highest average amount purchased.
People in the bracket who have a PhD, are married, and are in the income bracket of 100-200K who have 1 kid have the highest average amount purchased. This makes sense because having a high degree and being married usually means you tend to have a stable source of income, since at least one partner is contributing to the income. Having a child at home also means you probably make more purchases

```
cust %>% group_by(Education, Marital_Status, income_bracket, no_offspring) %>% summar
ize(mean_amount_purchased = mean(total_amount_purchased)) %>% arrange(-mean_amount_pu
rchased)
```

```
## `summarise()` has grouped output by 'Education', 'Marital_Status',
## 'income_bracket'. You can override using the `.groups` argument.
```

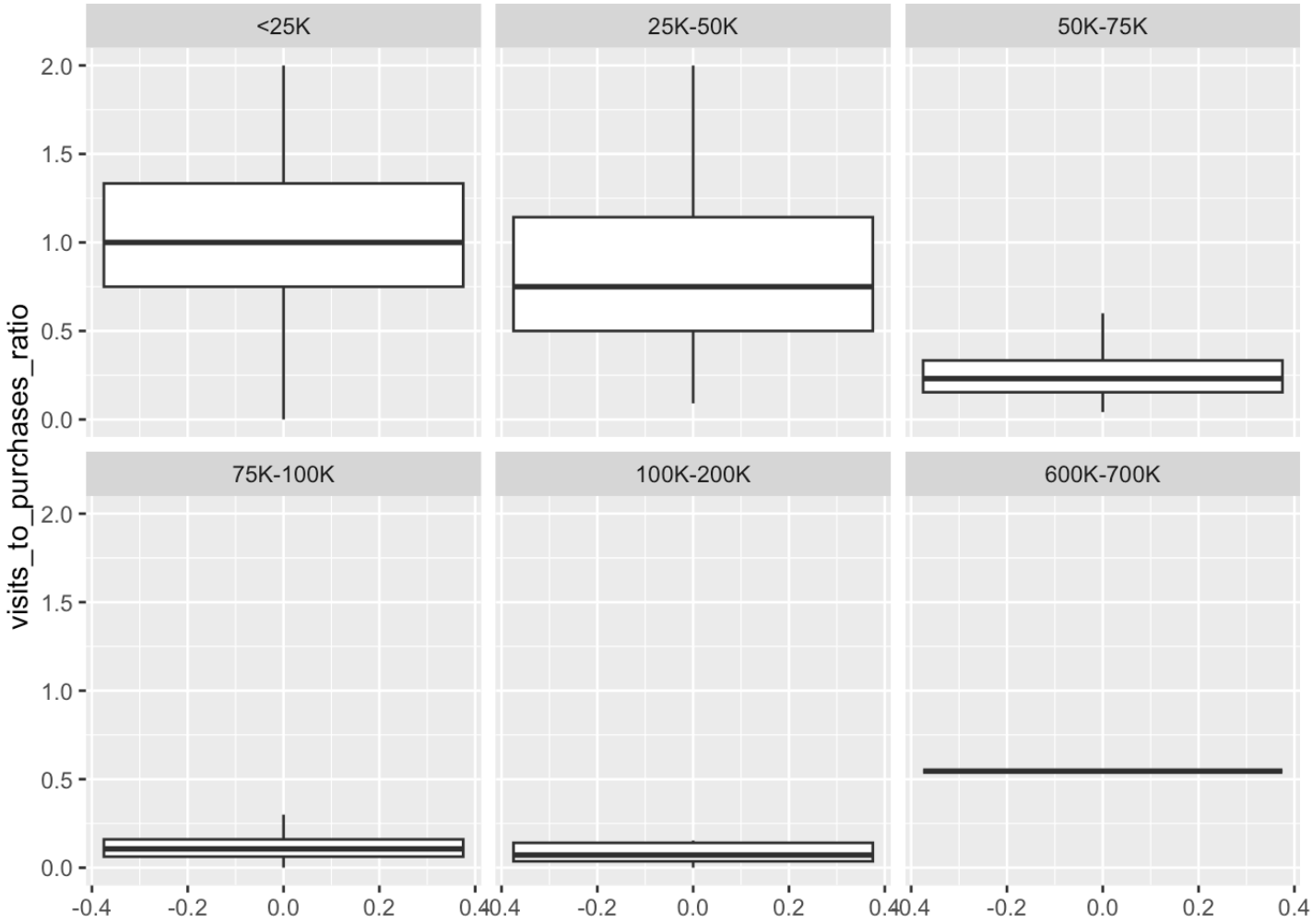
```
## # A tibble: 247 × 5
## # Groups:   Education, Marital_Status, income_bracket [97]
##    Education  Marital_Status income_bracket no_offspring mean_amount_purchased
##    <chr>      <chr>          <fct>                 <int>                  <dbl>
##  1 PhD        Married        100K-200K                 1                     37
##  2 Graduation Married        75K-100K                  2                     29
##  3 Graduation Single         75K-100K                  3                     29
##  4 Graduation Together       100K-200K                 0                     29
##  5 2n Cycle   Divorced       25K-50K                   0                     28
##  6 2n Cycle   Widow          50K-75K                   1                     27
##  7 Graduation Single         100K-200K                 0                     27
##  8 PhD        Single         100K-200K                 0                     27
##  9 2n Cycle   Divorced       50K-75K                   0                     26
## 10 PhD        Together       100K-200K                 0                     26
## # ℹ 237 more rows
```

2. Visualize the visits to purchase ratio based on the income bracket

It looks like for the most part, the median of visits to purchase ratio drastically decreases (meaning that they make more purchases than visits) as we go up the income bracket.
This link helped with getting rid of outliers and setting a limit on the y axis:
https://stackoverflow.com/questions/5677885/ignore-outliers-in-ggplot2-boxplot
(https://stackoverflow.com/questions/5677885/ignore-outliers-in-ggplot2-boxplot)

```
#any number greater than 1 means that they visited more than they purchased
cust %>% mutate(visits_to_purchases_ratio = (NumWebVisitsMonth/(NumWebPurchases+NumCa
talogPurchases+NumStorePurchases+NumDealsPurchases))) %>% filter(visits_to_purchases_
ratio != "Inf") %>% ggplot(aes(y = visits_to_purchases_ratio)) + geom_boxplot(outlier
.shape = NA) + scale_y_continuous(limits = c(0, 2)) + facet_wrap(~income_bracket)
```

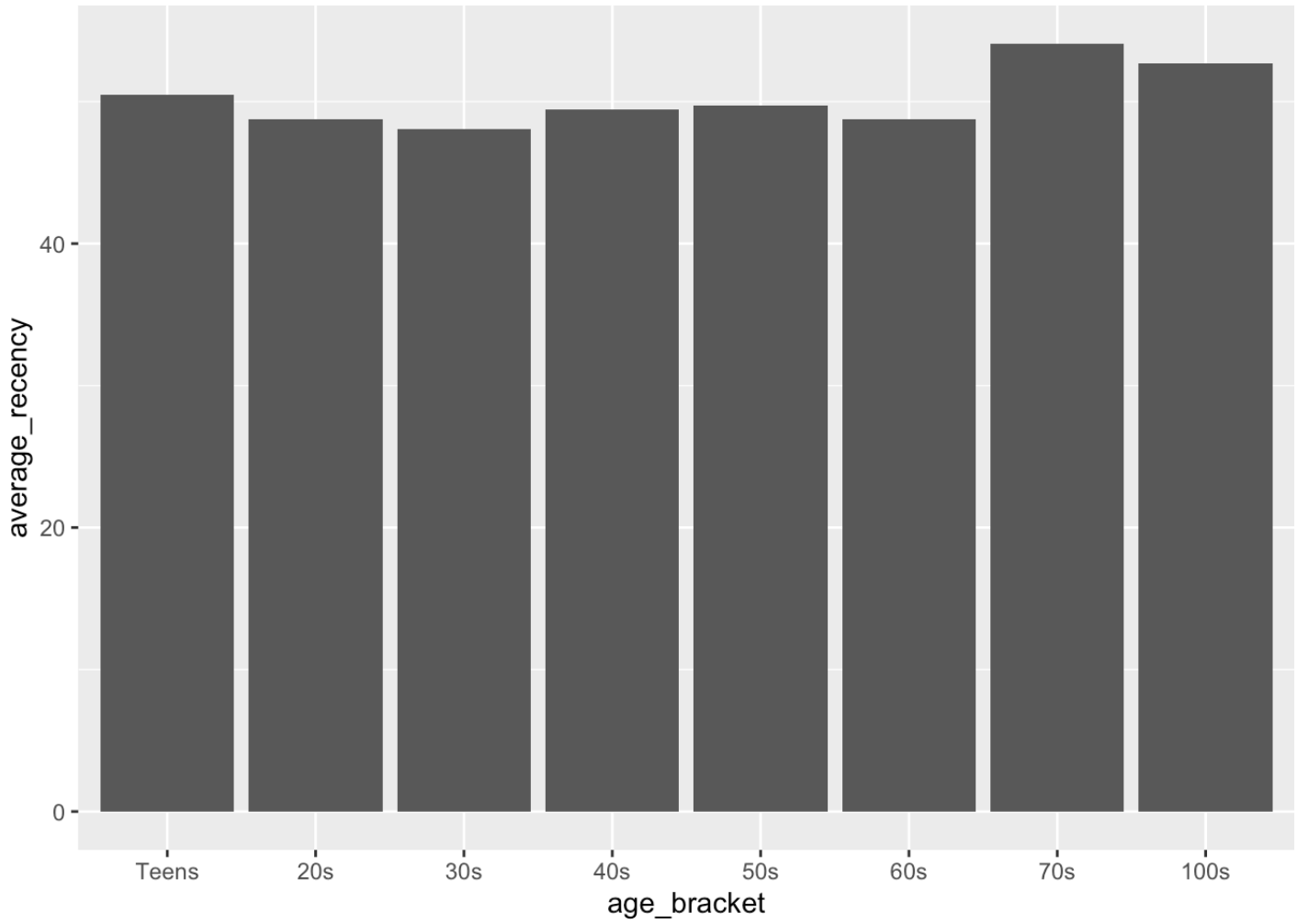
```
## Warning: Removed 10 rows containing non-finite values (`stat_boxplot()`).
```



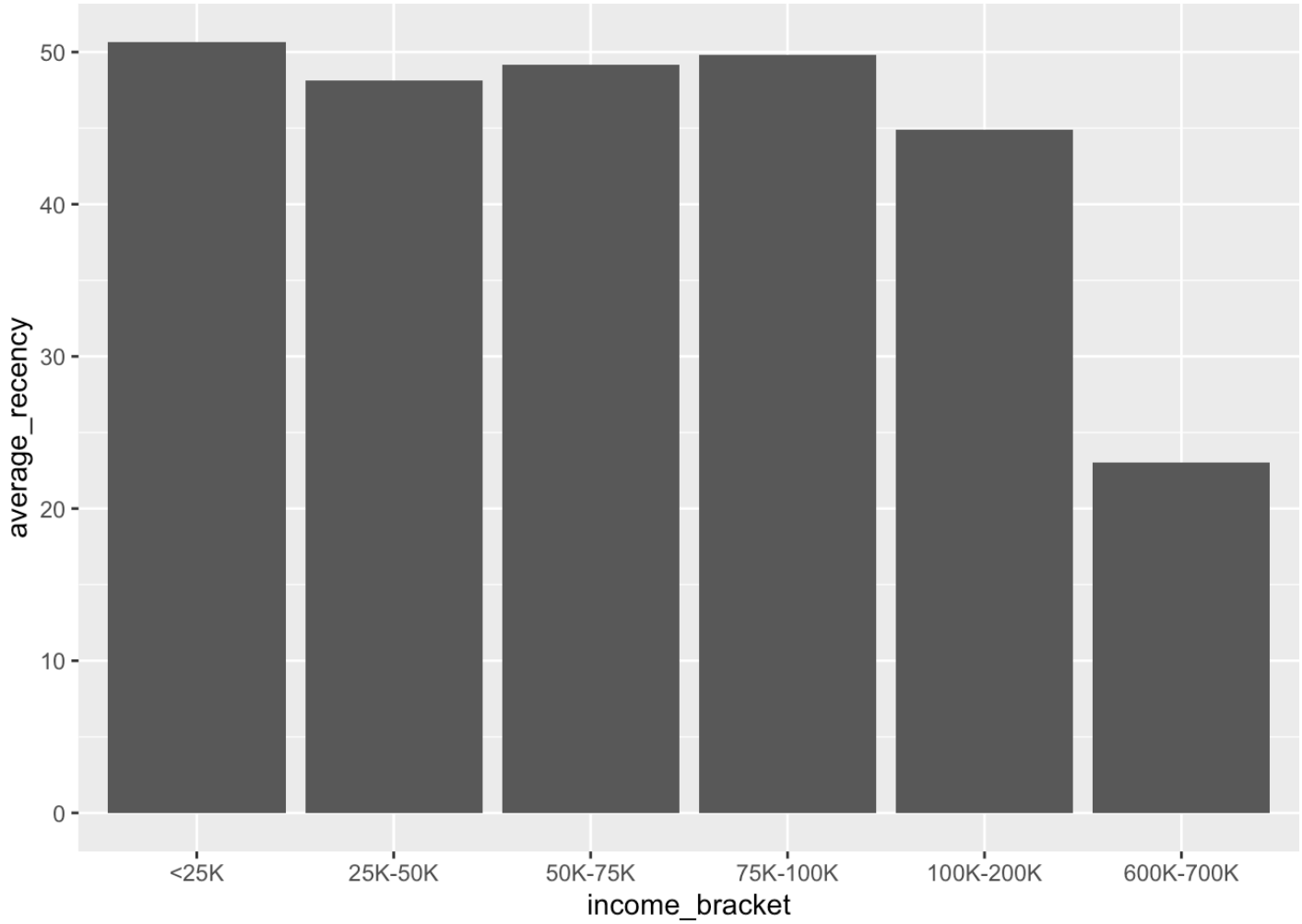
3. Visualize the age bracket and average number of recency (how many days pass between purchases). Do this again for income bracket.

The age bracket that has the highest recency, or number of days that pass between purchases, are people in their 70s and the bracket with the lowest recency are people in their 30s. But all brackets pretty much have around the same average recency. What about income bracket?
The income bracket that has the lowest recency, or number of days that pass between purchases, are people making 600K-700K, and highest recency are people making <25K.

```
cust  %>% group_by(age_bracket) %>% summarize(average_recency = mean(Recency)) %>% ar
range(-average_recency) %>% ggplot(aes(x = age_bracket, y = average_recency)) + geom_
bar(stat = "identity")
```




```
cust  %>% group_by(income_bracket) %>% summarize(average_recency = mean(Recency)) %>%
arrange(-average_recency) %>% ggplot(aes(x = income_bracket, y = average_recency)) +
geom_bar(stat = "identity")
```



4. Which education, marital_status, and income groups engage the most with marketing campaigns and which do not engage with marketing campaigns?

I measured this by summing the columns AcceptedCmp_[1,2,3,4,5] for each individual. The 1,2,3,4,5 represent which campaign they accepted (1st,2nd,3rd,4th, or 5th campaign). The value of the column is binary; 1 means they accepted and 0 means it was not accepted. If the individual had 0s across all of these columns, that means they did not engage in any market campaign at any point in time. If the total sum across the column was 1, that means they engaged in 1 market campaign. For each row, I added the values in across the AcceptedCmp_[1,2,3,4,5]columns, and this represents how many times they engaged in all the marketing campaigns.

The income bracket 75-100K had the largest marketing campaign engagement “score”, and the 600-700K had the lowest.

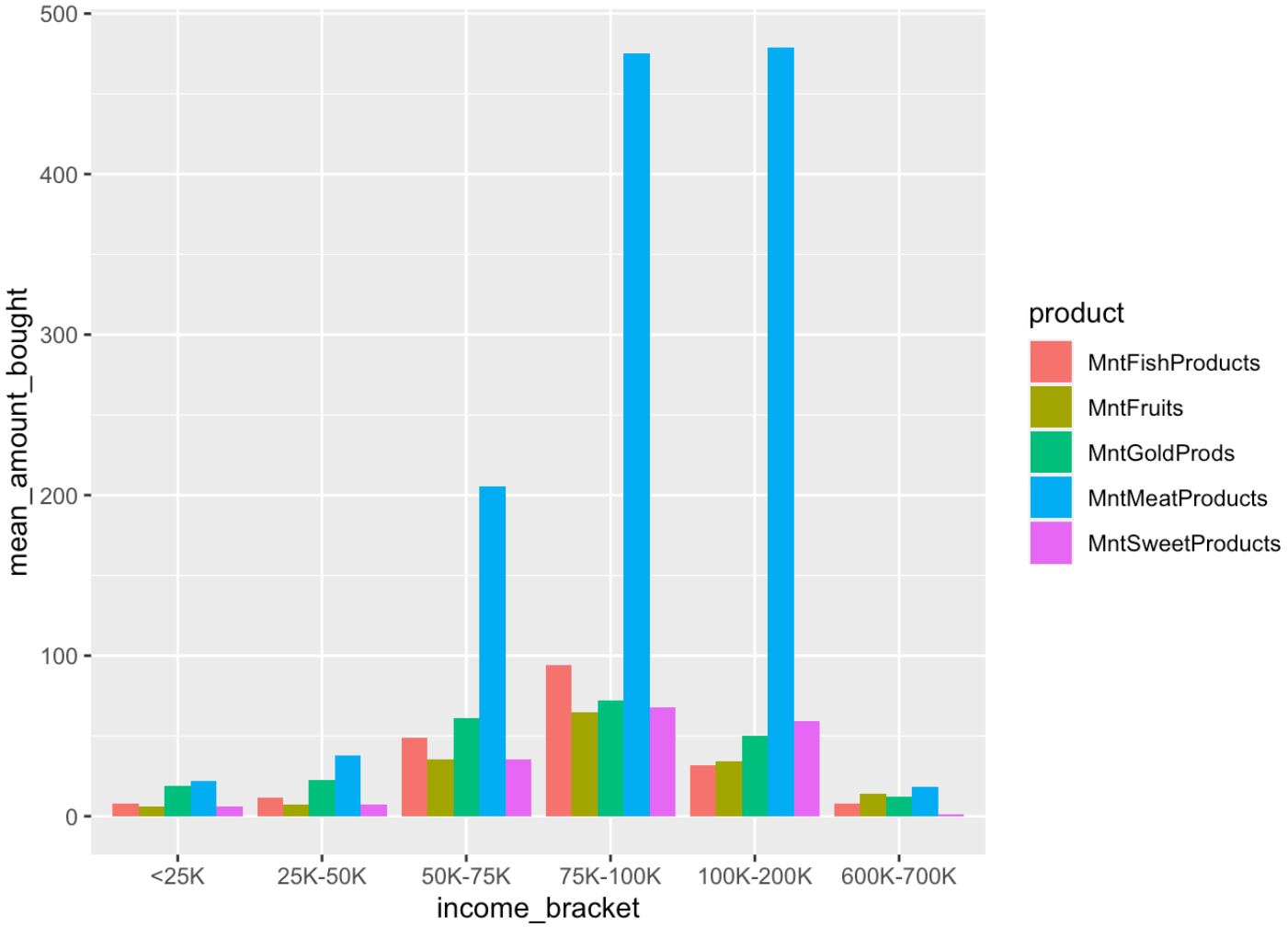
```
cust %>% mutate(engagement_score = rowMeans(across(c(AcceptedCmp1, AcceptedCmp2, Acce
ptedCmp3, AcceptedCmp4, AcceptedCmp5))))  %>% group_by(income_bracket) %>% summarize(
grouped_mean_engagement_sum = sum(engagement_score)) %>% arrange(-grouped_mean_engage
ment_sum)
```

```
## # A tibble: 6 × 2
##   income_bracket grouped_mean_engagement_sum
##   <fct>                                <dbl>
## 1 75K-100K                              65.4
## 2 50K-75K                               41.8
## 3 25K-50K                               18.6
## 4 <25K                                   4
## 5 100K-200K                              2.4
## 6 600K-700K                              0
```

5. Visualize the average number of purchases made for Fruit, Wine, Meat, Fish, Sweets, Gold for each income bracket
   Meat is the product that is bought the most across all income brackets, and fruit and sweets is less popular

```
cust %>% group_by(income_bracket) %>% summarise(across(c(MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds), mean, na.rm = TRUE))  %>% pivot_longer(cols = c(MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds), names_to = 'product', values_to = "mean_amount_bought")  %>%  ggplot(aes(x = income_bracket, y = mean_amount_bought, fill = product)) + geom_bar(stat = "identity", position = position_dodge())
```

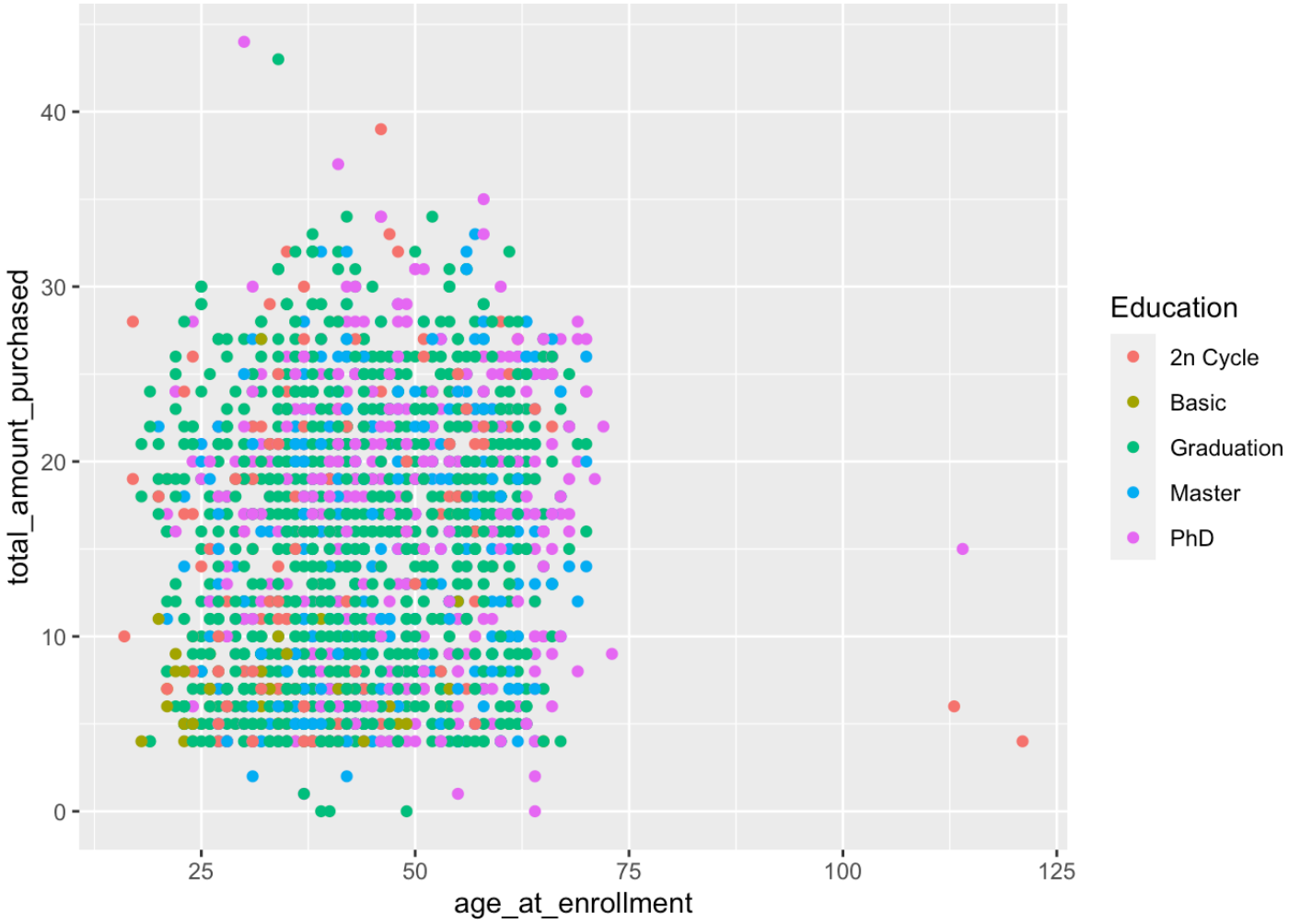
```
## Warning: There was 1 warning in `summarise()`.
## ℹ In argument: `across(...)`.
## ℹ In group 1: `income_bracket = <25K`.
## Caused by warning:
## ! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.
## Supply arguments directly to `.fns` through an anonymous function instead.
##
##   # Previously
##   across(a:b, mean, na.rm = TRUE)
##
##   # Now
##   across(a:b, \(x) mean(x, na.rm = TRUE))
```



6. Visualize the age at enrollment and the total amount of goods purchased, and color by marital status

You can't really tell if there's a trend, but the peak of highest total amount purchased comes from middle aged people and most of the colors at the peak are 2n cycle, graduation, master, and PhD.

```
cust %>% ggplot(aes(x = age_at_enrollment, y = total_amount_purchased, color = Educat
ion)) + geom_point()
```


total_amount_purchased
40
30
20
10
0
25
50
75
100
125
age_at_enrollment
Education
2n Cycle
Basic
Graduation
Master
PhD