Exam 1 - Take home

STA 199 - Spring 2024

Edmond Niu

I hereby state that I have not communicated with or gained information in any way from my classmates or any other humans during this exam, and that all work is my own.

Initials: EN

```
library(tidyverse)
library(palmerpenguins)
library(knitr)
```

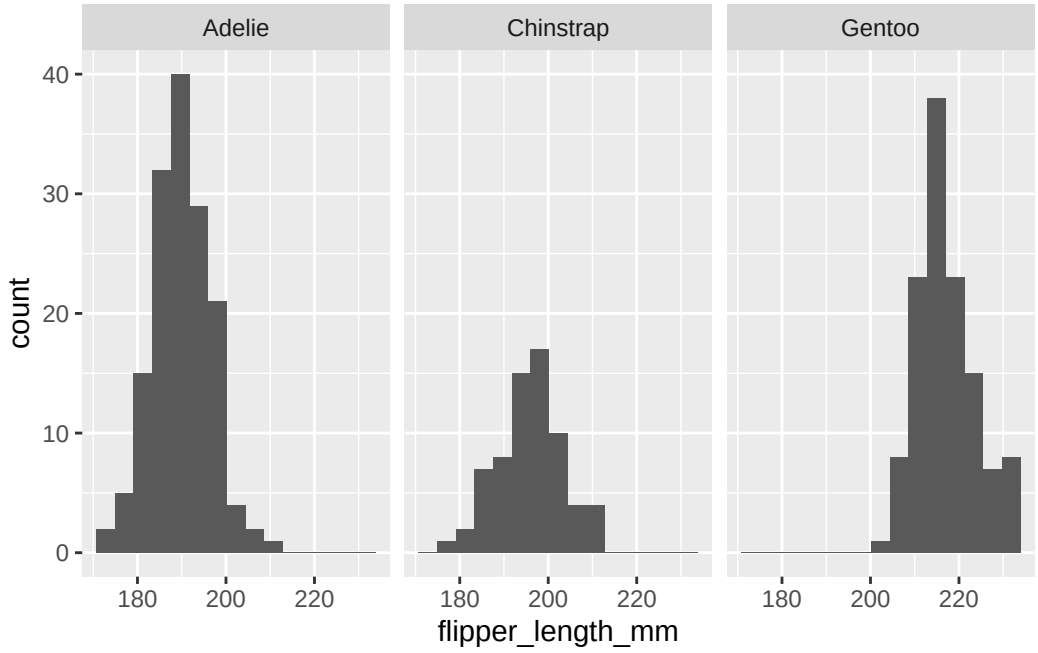
Déjà vu

Question 1

An explanatory factor for this atypical bimodal distribution of flipper lengths can be attributed to species. As can be seen by this visualization, where the original flipper length histogram has been split into individual histograms based on species, it can be shown that each species is contributing to a different portion of the total histogram. The distribution of flipper lengths of Adelie penguins is unimodal and normally distributed, with its mode at around 190 mm, while the distribution for Gentoo penguins' flipper lengths is also a unimodal, normal distribution with its mode at around 210mm. This creates the bimodal structure of the data when looking at these two species combined. The Chinstrap penguin fills in the gap between these two modes (190 and 210mm), with its little mini mode at 200mm and its data normally distributed around it, which ultimately does not change the bimodal structure of the data because there aren't enough data points for this species of penguin. This results in the bimodal distribution when looking at all 3 species combined.

```
ggplot(penguins, aes(x = flipper_length_mm)) +
  geom_histogram(bins = 15) +
  facet_wrap(~species)
```

```
Warning: Removed 2 rows containing non-finite values (`stat_bin()`).
```



Question 2

```
tv <- read_csv("data/tv.csv")
```

a.

```
csi <- tv |>
  filter(grepl("CSI", title))

#Documentation for grepl:
#https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/grep

csi |>
  slice(1:10)
```

```
# A tibble: 10 x 9
   season title                year month   day av_rating genre_1 genre_2 genre_3
    <dbl> <chr>               <dbl> <dbl> <dbl>     <dbl> <chr>   <chr>   <chr>
 1      1 CSI: Crime Scene ~  2001     1    20      8.32 Crime   Drama   Mystery
 2      2 CSI: Crime Scene ~  2002     1    10      8.26 Crime   Drama   Mystery
 3      3 CSI: Crime Scene ~  2003     1    15      8.30 Crime   Drama   Mystery
 4      4 CSI: Crime Scene ~  2004     1    18      8.33 Crime   Drama   Mystery
 5      5 CSI: Crime Scene ~  2005     1    24      8.38 Crime   Drama   Mystery
 6      6 CSI: Crime Scene ~  2006     1    16      8.21 Crime   Drama   Mystery
 7      7 CSI: Crime Scene ~  2007     1    14      8.43 Crime   Drama   Mystery
 8      8 CSI: Crime Scene ~  2008     1     7      8.07 Crime   Drama   Mystery
 9      9 CSI: Crime Scene ~  2009     1    27      7.80 Crime   Drama   Mystery
10     10 CSI: Crime Scene ~  2010     1    23      7.69 Crime   Drama   Mystery
```

b. CSI: Crime Scene (15); CSI: Miami (10); CSI: NY (9); CSI: Cyber (2)

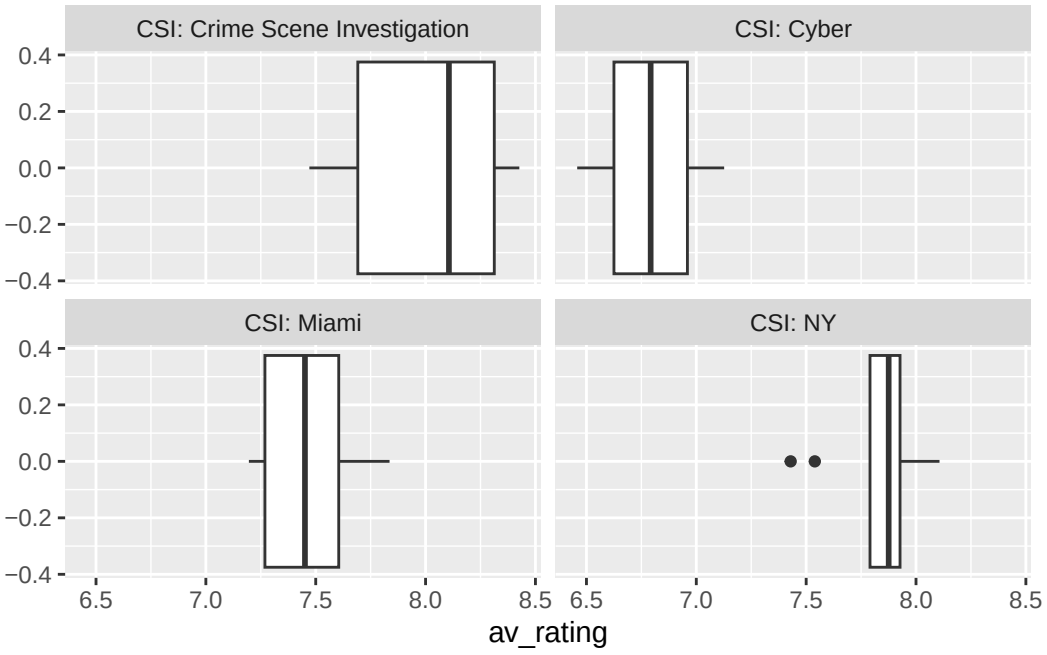
```
csi |>
  count(title) |>
  arrange(desc(n))
```

```
# A tibble: 4 x 2
  title                              n
  <chr>                          <int>
1 CSI: Crime Scene Investigation    15
2 CSI: Miami                        10
```

```
3 CSI: NY                          9
4 CSI: Cyber                       2
```

c. Across the various CSI titles, this is the median average rating across all seasons from highest to lowest: Crime Scene Investigation, NY, Miami, and Cyber. The IQR across various CSI titles from largest to smallest: Crime Scene Investigation, Cyber/Miami (too similar to tell), and NY. NY also is the only CSI title to have two outliers in its average rating across all seasons. Overall (excluding outliers), this showcases that Crime Scene Investigation had the highest ratings but had the greatest variability, while Cyber had the lowest ratings, and NY had the lowest variability. All CSI titles show minimal skewed-ness.

```
ggplot(csi, aes(x = av_rating)) +
  geom_boxplot() +
  facet_wrap(~title)
```



Credit cards

```
credit <- read_csv("data/credit.csv")
```

Question 3

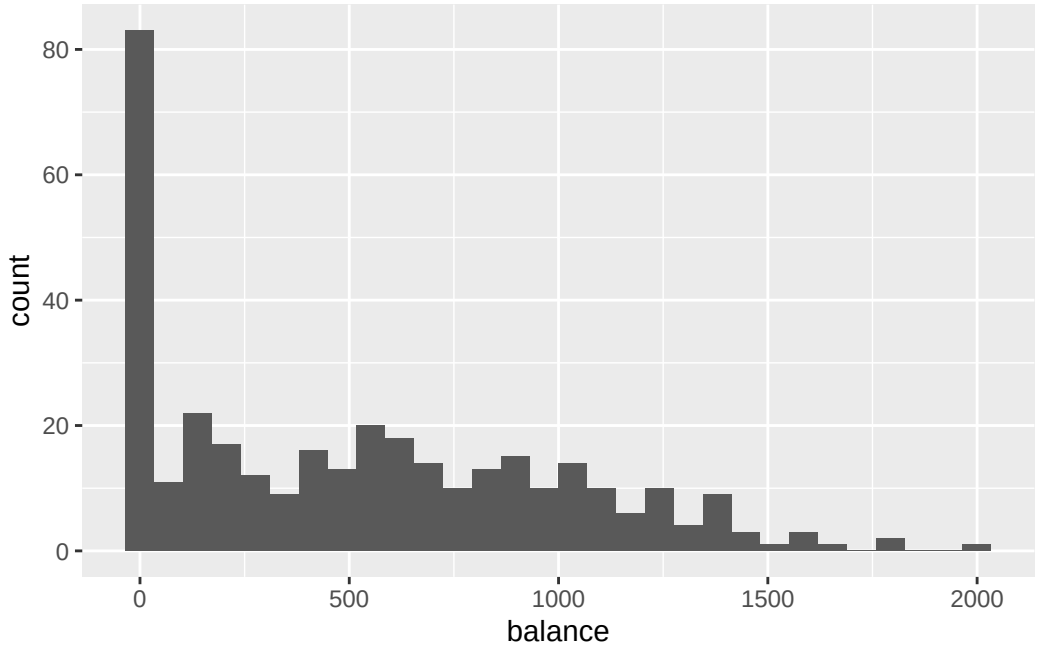
The `credit` dataset has 347 rows and 5 columns. Each row represents a credit card customer. The primary variable of interest in the dataset is `balance`, credit card balances in US Dollars. The shape of the distribution of credit card balances is unimodal and right-skewed. The typical credit card balance is 467. 50% of the customers in the dataset have credit card balances between 68.5 and 865.5.

```
glimpse(credit)
```

```
Rows: 347
Columns: 5
$ balance         <dbl> 890, 194, 298, 47, 0, 912, 0, 155, 1587, 637, 1176, 73~
$ income          <dbl> 49.927, 24.314, 21.153, 34.537, 25.974, 23.283, 16.529~
$ student_status  <chr> "Not student", "Not student", "Not student", "Not stud~
$ marriage_status <chr> "Married", "Married", "Not married", "Married", "Not m~
$ limit           <dbl> 6396, 3409, 3736, 3271, 2308, 5443, 1357, 3388, 7582, ~
```

```
ggplot(credit, aes(x = balance)) +
  geom_histogram()
```

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




```
credit |>
  summarize(
    median = median(balance),
    q1 = quantile(balance, 0.25),
    q3 = quantile(balance, 0.75)
  )
```

```
# A tibble: 1 x 3
  median    q1    q3
   <dbl> <dbl> <dbl>
1    467  68.5  866.
```

```
#ggplot(credit, aes(x = balance)) +
#  geom_boxplot()
```

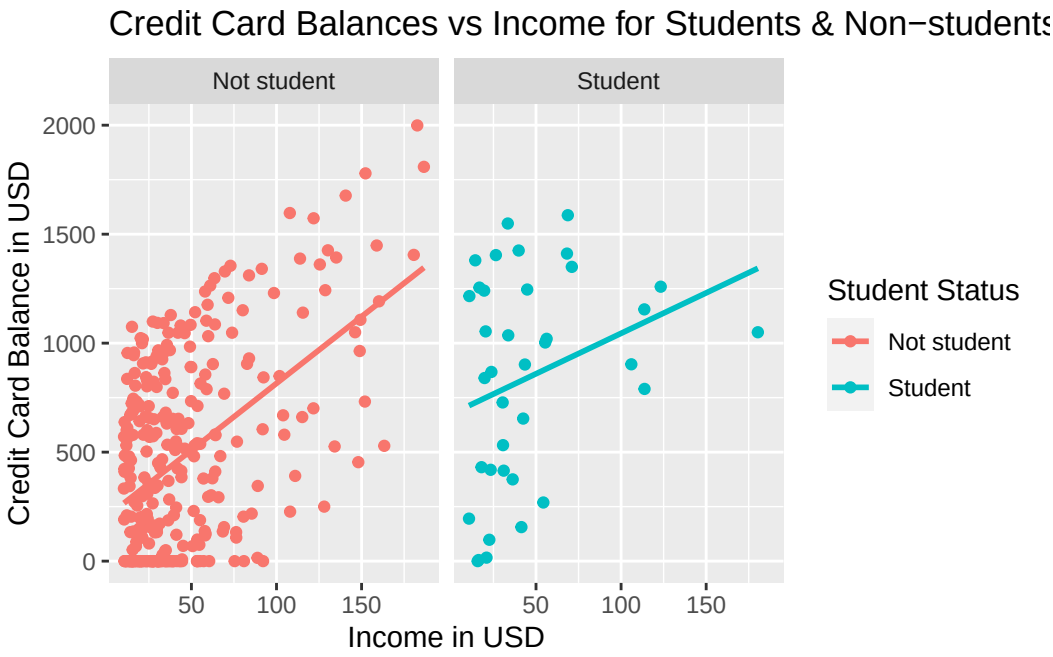
Question 4

The relationship between credit card balance and income is weak, linear, and positive. As income increases, the credit card balance tends to increase as well. However, the points are all over the place and the data do not fit the line of best fits nicely (hence, weak).

When it comes to differences by student status, the relationship for balance vs income for students is less positive (flatter line) than for non-students. In other words, the balance of students' credit cards do not increase as much as the balance of non-students' credit cards do, per dollar increase in income. That is to say, for every dollar that a non-student earns, they tend to have a higher credit card balance than a student would if a student earned that same dollar.

```
ggplot(credit, aes(x = income, y = balance, color = student_status)) +
  geom_point() +
  facet_wrap(~student_status) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Credit Card Balances vs Income for Students & Non-students",
    x = "Income in USD",
    y = "Credit Card Balance in USD",
    color = "Student Status"
  )
```

```
`geom_smooth()` using formula = 'y ~ x'
```


Credit Card Balances vs Income for Students & Non−students
Not student
Student
Credit Card Balance in USD
2000
1500
1000
500
0
50
100
150
Income in USD
Student Status
Not student
Student

Question 5

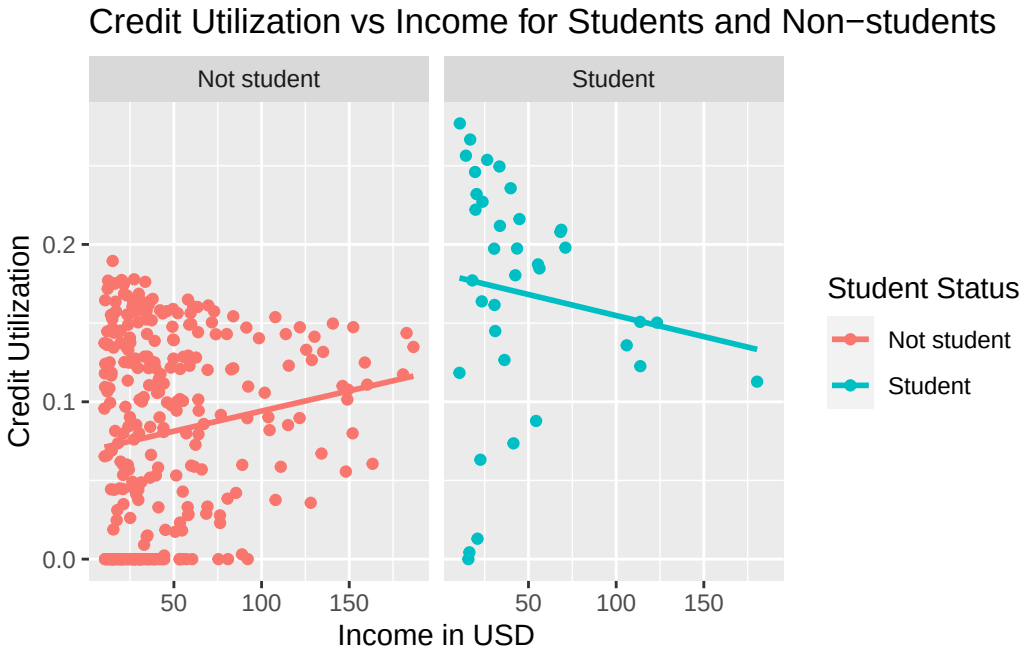
The relationship between income and credit utilization is different from the relationship between income and credit card balance for students because the relationship is now weak, linear, and negative (instead of positive).

On the other hand, the relationship for non-students is still weak, linear, and positive, but if anything, the correlation is even weaker (the points seem even more random and less linear) and even less positive (line is closer to horizontal).

```
credit_plus <- credit |>
  mutate(credit_util = balance / limit)

ggplot(credit_plus, aes(x = income, y = credit_util, color = student_status)) +
  geom_point() +
  facet_wrap(~student_status) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Credit Utilization vs Income for Students and Non-students",
    x = "Income in USD",
    y = "Credit Utilization",
    color = "Student Status"
  )
```

```
`geom_smooth()` using formula = 'y ~ x'
```


Credit Utilization vs Income for Students and Non−students
Not student
Student
Credit Utilization
0.2
0.1
0.0
50
100
150
Income in USD
Student Status
Not student
Student

Question 6

a. The group that had the highest credit card balance were students that were not married.

```
credit_summary_unclean <- credit |>
  group_by(student_status, marriage_status) |>
  summarise(
    mean_credit_card_balance = mean(balance)
  )
```

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

```
credit_summary_unclean
```

```
# A tibble: 4 x 3
# Groups:   student_status [2]
  student_status marriage_status mean_credit_card_balance
  <chr>          <chr>                             <dbl>
1 Not student    Married                            498.
2 Not student    Not married                        453.
3 Student        Married                            792.
4 Student        Not married                        899.
```

b.

```
credit_summary <- credit_summary_unclean |>
  pivot_wider(names_from = marriage_status, values_from = mean_credit_card_balance)

kable(credit_summary, digits = 0, col.names = c("", "Married", "Not married"))
```

	Married	Not married
Not student	498	453
Student	792	899

```
#pivot wider documentation: https://tidyr.tidyverse.org/reference/pivot_wider.html
```