Project 2

Name: Jeremy Ulfohn // jau392

This is the dataset you will be working with:

```
bank_churners <- readr::read_csv("https://wilkelab.org/SDS375/datasets/bank_churners.csv")
```

More information about the dataset can be found here: https://www.kaggle.com/sakshigoyal7/credit-card-customers

Part 1

Question: Is attrition rate related to income level?

To answer this question, create a summary table and one visualization. The summary table should have three columns, income category, existing customers, and attrited customers, where the last two columns show the number of customers for the respective category.

The visualization should show the relative proportion of existing and attrited customers at each income level.

For both the table and the visualization, make sure that income categories are presented in a meaningful order. For simplicity, you can eliminate the income level "Unknown" from your analysis.

Hints:

1. To make sure that the income levels are in a meaningful order, use `fct_relevel()`. Note that `arrange()` will order based on factor levels if you arrange by a factor.
2. To generate the summary table, you will have to use `pivot_wider()` at the very end of your processing pipeline.

Introduction: *Main Dataset Intro:* The models to follow are based on the bank_churners dataset — a dataset compiled in an effort to predict which credit card holders would attrite in order to go to them before attrition with offers of improved service. To do this, we have tabulated attritions and non-attritions against various factors, namely age, gender, number of dependents, income and education levels, marital status, card tier, and length of customership. With these data, we will find numerous correlations (or lacks thereof) between factors and ultimately build a predictive, empirical model.

Modified Dataset Intro: To see whether attrition rate is related to income level, we will trim the dataset to include only 2 columns for the summary table:

1. Income Category (ordered lowest to highest, from <40K to >120K)
2. Attrition Flag == "Existing Customer" and (3) Attrition Flag == "Attrited Customer", the inverse of column 2

In our analysis, we will find the counts of these 3 discrete variables.

Approach: For this question, we will filter out any unknown-income entries since they are inherently ambiguous and misleading. After ordering the income levels from least to greatest, we will use pivot_wider() to generate a neat, legible summary table for the counts of each attrition flag at each level. This will provide the necessary background information, i.e. counts, for the bar chart.

The summary table shows the count, but fails to show the relative proportions by income level without manual calculation. To visualize this, we use a bar chart with the proportion on the y axis, where the

proportion of attrited and that of existing add to 1 and are distinguished by fill color. The proportions are too similar to each other for a pie chart to have been of use, and a position "stack" bar chart would have been indecipherable without extensive, unsightly labeling of relative proportion at each level (or the same manual calculation we were trying to avoid from the chart.) Therefore, side-by-side bars with position "fill" was the best choice, as it clearly and legibly shows relative proportion better than all the others.

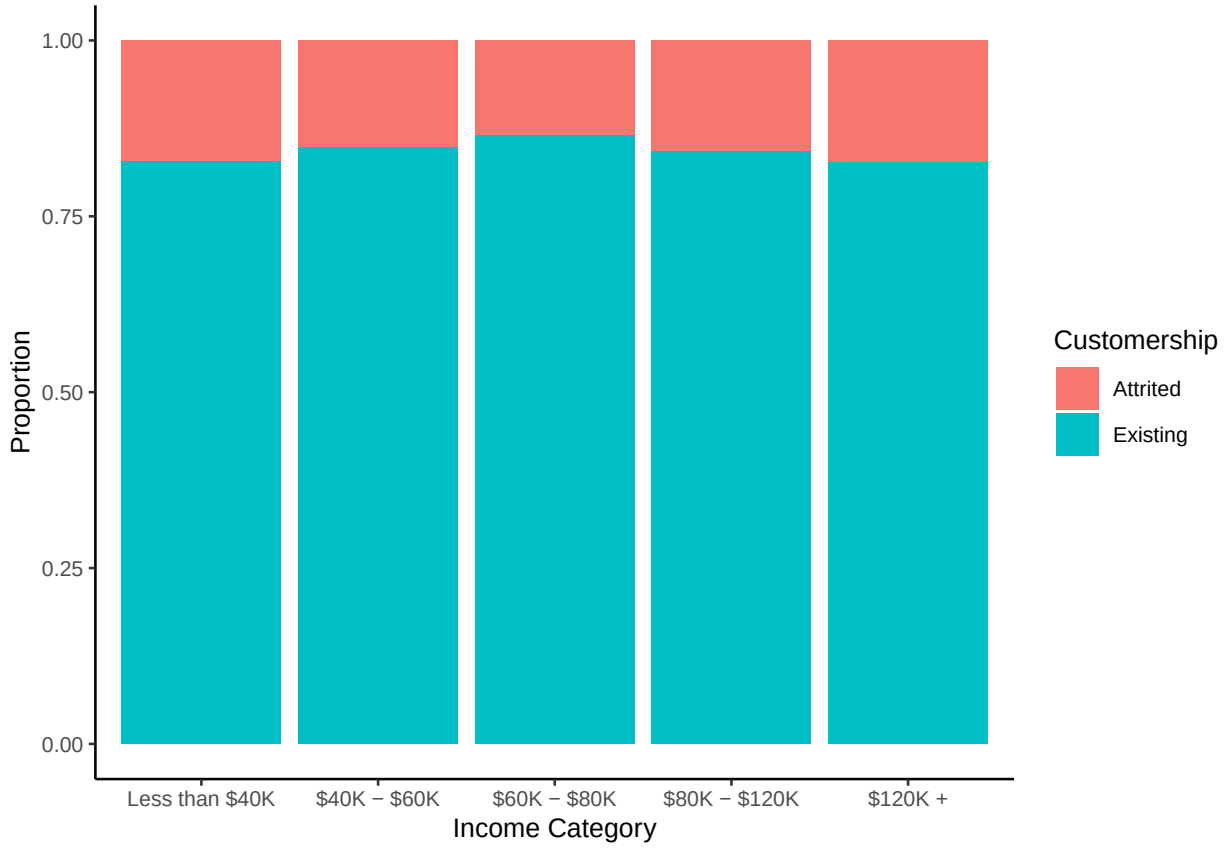
Analysis:

```
summary <- bank_churners %>%
  filter(Income_Category != "Unknown") %>%
  mutate(Income_Level = fct_relevel(Income_Category, "Less than $40K", "$40K - $60K", "$60K - $80K", "$8
  count(Income_Level, Attrition_Flag) %>%
  pivot_wider(names_from = "Attrition_Flag", values_from = "n")
summary
```

```
## # A tibble: 5 x 3
##   Income_Level   `Attrited Customer` `Existing Customer`
##   <fct>                        <int>               <int>
## 1 Less than $40K                 612                2949
## 2 $40K - $60K                    271                1519
## 3 $60K - $80K                    189                1213
## 4 $80K - $120K                   242                1293
## 5 $120K +                        126                 601
```

```
to_plot <- bank_churners %>%
  filter(Income_Category != "Unknown") %>%
  mutate(Existing = Attrition_Flag != "Attrited Customer") # Boolean

# Visualization - stacked bars
library(ggplot2)
# FIXME: Add percentages with label = "___" in aes, or MANUALLY
ggplot(to_plot, aes(
  x = fct_relevel(Income_Category, "Less than $40K", "$40K - $60K", "$60K - $80K", "$80K - $120K", "$12
  scale_x_discrete(name = "Income Category") +
  scale_fill_discrete(name = "Customership", labels = c("Attrited", "Existing")) +
  geom_bar(position = "fill") + # stack
  scale_y_continuous(name = "Proportion", limits = c(0, 1)) +
  theme_classic(10)
```



Discussion: First and foremost, we observe that there is *not very much* difference in relative attrition proportion between the income categories. For what differences there are, the chart is almost completely symmetric, with the lowest attrition rates of ~13.5% in the 60-80K bracket increasing as income decreases to <$40K and as it increases to >120K (each to about 17.2%). We note, however, that the <40K figures are more statistically significant than the >120K ones simply because there are almost 5 times more customers in the low bracket.

From this information, we glean that there might be a tendency for low and high earners to be more capricious with their choice of bank, probably for reasons to do with better interest rates elsewhere. This could tie in to a larger trend of the "extreme" ends of earners being more focused on their finances than the average person and therefore more likely to seek change. Again, this is more conjecture than anything since the differences are relatively small. Combining this data with those of other banks and extending the income ranges down to <10K and >150K would allow us further insight into the ostensible trend.

Part 2

Question: How does the relationship (if any) between customer age and credit limit change by marital status?

Introduction: For this question, we will utilize the following 3 columns of the same bank_churners dataset described above:

1. Marital Status
2. Customer Age
3. Credit Limit

Marital Status is a categorical variable of either "Single," "Married", or "Divorced", and the latter two are

continuous (numerical) variables.

Approach: Some modifications to bank_churners are due. First, we remove "Unknown" marital status as we did with income level in Part 1. We then order the dataset manually by marital status and summarize, populating the summary table with the average customer age and average credit limit for each of the 3 marital statuses. Finally, we widen the summary table for readability.

Since we have two quantitative, continuous variables, we will perform a smoothed least sum of squares regression with the independent variable of 'age' on the x axis and 'credit limit' on the y axis. Since Marital Status is the only discrete categorical variable, it is the perfect opportunity to use the "color" feature of geom_point() to distinguish between the three models. Because the credit limit can vary very widely and might make the trend line unreadable, we switch the y axis scale to log_10(Credit Limit).

Geom_point is ideal since we have the strictly-increasing quantity of age on the x-axis and the highly-variable quantity of credit limit on the y. Similarly, using a smoothed line of interconnected points will be able to accurately and visibly represent whatever trends, if any, exist between age and credit limit.

Analysis:

```
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
# Create summary table
summ2 <- bank_churners %>%
  filter(Marital_Status != "Unknown") %>%
  mutate(Marriage_Status = fct_relevel(Marital_Status, "Single", "Married", "Divorced")) %>%
  group_by(Marriage_Status) %>%
  summarize(
    n = n(),
    Avg_Cust_Age = mean(Customer_Age),
    Avg_Cred_Limit = mean(Credit_Limit) ## Check text
  ) %>%
  pivot_wider()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
summ2
```

```
## # A tibble: 3 x 4
##   Marriage_Status     n Avg_Cust_Age Avg_Cred_Limit
##   <fct>           <int>        <dbl>          <dbl>
## 1 Single           3943         46.2          9000.
## 2 Married          4687         46.7          8077.
## 3 Divorced          748         45.1          9359.
```

```
second_plot <- bank_churners %>%
  filter(Marital_Status != "Unknown")


ggplot(second_plot, aes(y = Credit_Limit, x = Customer_Age, color = Marital_Status)) +
```

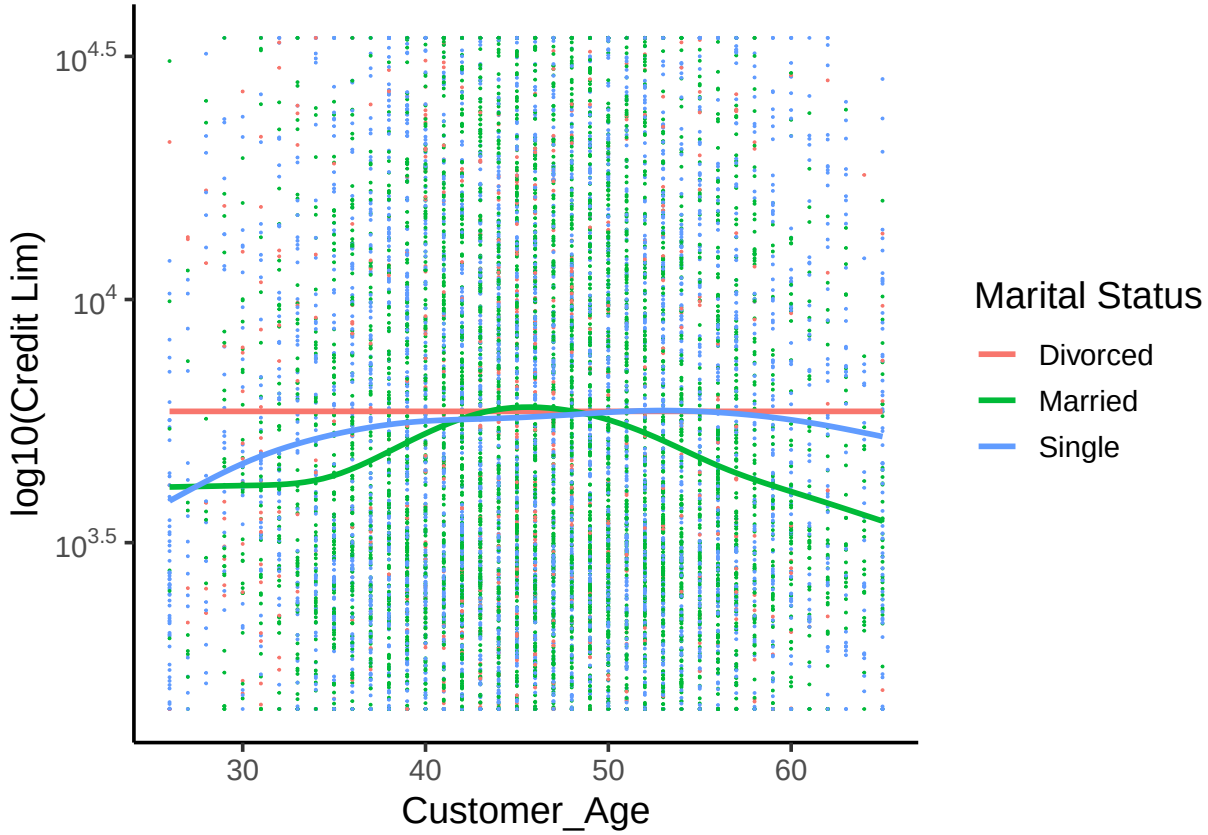
```
  geom_point(size = 0.01) +
  theme_classic(14) +
  scale_color_discrete(name = "Marital Status") +
  # Change y axis to reflect 10^y = Credit Limit (log10 transform)
  scale_y_continuous(name = "log10(Credit Lim)", trans='log10', breaks=trans_breaks("log10", function(x)
  scale_x_continuous(name = "Customer Age") +
  xlim(26, 65) + # Change x to stop at 65, since all unmarried data ends
  geom_smooth(se = FALSE, alpha = 1/100) # Adjust alpha for size of data
```

```
## Scale for 'x' is already present. Adding another scale for 'x', which will
## replace the existing scale.
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

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

```
## Warning: Removed 10 rows containing missing values (geom_point).
```



Discussion: We see that by and large, divorced status has the highest credit limits, especially at ages younger than 40 or older than 50. It is also by far the least variant trend line of the three, despite interestingly having the far-lowest count of only 748. Single status sees the largest average hike in credit limit of the 3, briefly cresting slightly above divorced's level in the early fifties. Conversely, married status has the largest average fall by age in credit limit of the 3, reaching **its** apex above both other levels in the mid forties.

We observe that the density of bank customers is much more sparce at the extreme young and old ends of the spectrum, so we can speak with the most confidence in saying that there's not much of a difference in credit limits by marital status between 40 and 50. The marital status differences we observe at early ages might be because the the majority of those 26 to 30 are single, but the fact that married credit limits actually **fall**

before they rise casts doubt on this conclusion so we cannot say for sure.