Project 2

Derian Lee (dtl634)

This is the dataset you will be working with:

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

bank_churners
```

```
## # A tibble: 10,127 x 21
##    CLIENTNUM Attrition_Flag Customer_Age Gender Dependent_count Education_Level
##        <dbl> <chr>                 <dbl> <chr>            <dbl> <chr>
##  1 768805383 Existing Cust~           45 M                    3 High School
##  2 818770008 Existing Cust~           49 F                    5 Graduate
##  3 713982108 Existing Cust~           51 M                    3 Graduate
##  4 769911858 Existing Cust~           40 F                    4 High School
##  5 709106358 Existing Cust~           40 M                    3 Uneducated
##  6 713061558 Existing Cust~           44 M                    2 Graduate
##  7 810347208 Existing Cust~           51 M                    4 Unknown
##  8 818906208 Existing Cust~           32 M                    0 High School
##  9 710930508 Existing Cust~           37 M                    3 Uneducated
## 10 719661558 Existing Cust~           48 M                    2 Graduate
## # ... with 10,117 more rows, and 15 more variables: Marital_Status <chr>,
## #   Income_Category <chr>, Card_Category <chr>, Months_on_book <dbl>,
## #   Total_Relationship_Count <dbl>, Months_Inactive_12_mon <dbl>,
## #   Contacts_Count_12_mon <dbl>, Credit_Limit <dbl>, Total_Revolving_Bal <dbl>,
## #   Avg_Open_To_Buy <dbl>, Total_Amt_Chng_Q4_Q1 <dbl>, Total_Trans_Amt <dbl>,
## #   Total_Trans_Ct <dbl>, Total_Ct_Chng_Q4_Q1 <dbl>,
## #   Avg_Utilization_Ratio <dbl>
```

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:

The *bank_churners* dataset includes various information about customers at a bank, such as age, gender, income, and more. The question to answer is whether or not attrition rate is related to the income level of a customer. To answer this question, several variables from the dataset will be needed. The following variables will be taken from the dataset: *Income_Category* and *Attrition_Flag.*

Note: Because *bank_churners* has a value of "Unknown" for some income categories, a new dataset, *bank_churners1,* will be used to remove these values. Additional transformations to *bank_churners1* are described below.

Approach:

DATA WRANGLING / SUMMARY TABLE: One limitation with the *bank_churners* dataset is that there are unknown values for income level. Since this is deadweight to our analysis, rows possessing a value "Unknown" will be removed through the *filter()* function. Next, we will subset the data by using the *select()* function, which allows us to focus solely on our variables of interest. Lastly, the *fct_relevel()* argument is used within the *mutate()* function to manually reorder income categories from least to greatest, allowing for better interpretability when reading the summary table and visualization. The transformed data should now use *count()* and *pivot_wider()* to return attrition count per income category.

PLOT: Although both a faceted pie chart and a stacked bar chart was considered for this analysis, (these are visually appealing for a small number of subsets, in this case, *Attrited* and *Existing*) the latter was thought to be more appropriate here. A stacked bar chart was chosen because the probabilities for attrition are very similar throughout all income categories, so having these probabilities in close proximity (stacked on top of each other) allows for an easy visual distinction between these minor differences. Likewise, I would argue that the distance between each pie in a faceted pie chart would make these minor distinctions more difficult to recognize.

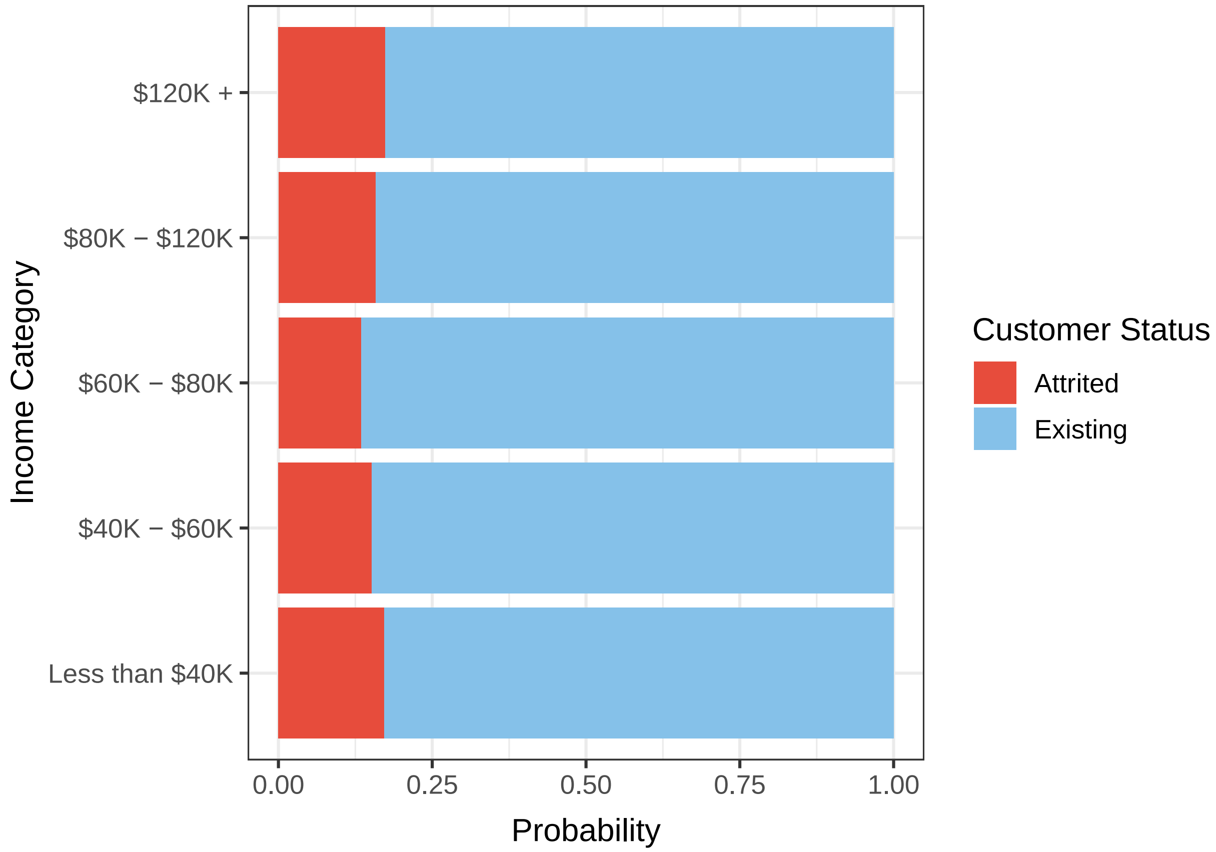
Analysis:

```r
#Question 1: Data Wrangling
bank_churners1 <- bank_churners %>%
  filter(Income_Category != "Unknown") %>% #remove unknown values
  select(Income_Category, Attrition_Flag) %>% #focus on 2 variables
  mutate(Income_Category = fct_relevel(Income_Category, #meaningful reordering
  "Less than $40K","$40K - $60K","$60K - $80K","$80K - $120K","$120K +"))

#Question 1: Summary Table
bank_churners1 %>% #return attrition count per income category
  count(Income_Category, Attrition_Flag) %>%
  pivot_wider(names_from = "Attrition_Flag", values_from = "n")
```

```
## # A tibble: 5 x 3
##   Income_Category `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
```

```r
#Question 1: Visualization
bank_churners1 %>%
  ggplot(aes(y = Income_Category, fill = Attrition_Flag)) +
  geom_bar(position = position_fill(rev = TRUE)) + #reverse stacking order
  scale_x_continuous( #renaming x axis
    name = "Probability"
  ) +
  scale_y_discrete( #renaming y axis
    name = "Income Category"
  ) +
  scale_fill_manual( #renaming legend
    name = "Customer Status",
    labels = c('Attrited Customer' = "Attrited", 'Existing Customer' = "Existing"),
    values = c('Attrited Customer' = "#E74C3C", 'Existing Customer' = "#85C1E9")
  ) +
  theme_bw(12) +
  theme( #resizing, shifting axis titles
    axis.title.x = element_text(size = 11.5, vjust = -0.5),
    axis.title.y = element_text(size = 11.5, vjust = 1)
  )
```



Discussion:

According to the visualization, it appears that attrition rate is highest among the two opposite ends of income category, and this rate lowers as income approaches the middle category ($60k - $80k) from both ends of the spectrum. Thus, a manager should proactively focus on customers who fall into the ends of this income spectrum if they'd like to reduce attrition rate. One could argue change in income plays a factor in credit card attrition. Perhaps at lower and higher income categories, fluctuations in income are more likely than middle income categories. For example, you might be more likely to get a raise at lower income levels. And higher income levels have a higher ability to invest in the market and thus, have fluctuating income from stock returns. Since card category (blue, silver, gold, platinum) broadly represent economic status, changes in income can serve as incentives to upgrade or downgrade cards.

Part 2

Question: Out of those who are married, how does the proportion of card categories differ across sex, and does sex and card type have an effect on transaction count?

Introduction:

The *bank_churners* dataset is referenced again for Question 2; information about the dataset is described above. The question to answer is, out of those who are married, (1) how does the proportion of card categories differ across sex, and (2) does sex and card category have an effect on total transaction count. To answer this question, several variables from the dataset will be needed.The following variables will be taken from the dataset: *Marital_Status*, *Card_Category*, *Gender*, and *Total_Trans_Ct*.

Note: Because this question focuses on married customers, *bank_churners2* will be used to filter marital status. Additional transformations to *bank_churners2* are described below.

Approach:

DATA WRANGLING / SUMMARY TABLE: Since this question focuses on married customers, this demographic will be subsetted through the *filter()* function. For our eventual summary table and visualization, we want a logical order for card category, so the *fct_relevel()* argument will be used within *mutate()* to manually change this into an ascending order. We can then use *count()* followed by *pivot_wider()* to return card category count per sex. Next, we'll use the *mutate()* function along with mathematical operators to add our columns of interest, specifically the proportion of each card category among the sexes. Lastly, we'll use the *select()* function to return both the proportions and counts of each card category. Though the question is interested in proportions, the count is returned as well, since it is necessary to answer the visualization part of the question.

Additional note on summary table: To answer the proportion part of the question, a table was created rather than a visualization because the blue category is overwhelmingly popular for both sexes and thus, a visualization was not deemed as helpful to answer this question (some categories may not be visible as percentages).

PLOT: For the second part of the question, a faceted boxplot will be used. This was chosen because faceted boxplots allow us to visualize the spread of total transactions against card category, then facet across the two sexes. It's also a solid choice because boxplots provide a quick way to communicate visual information, and can thus be used over many categories without being difficult to interpret. For example: One can simply look at how the median transaction count changes across cards within a sex, or how median transaction count changes across sex within a card; either approach is an easy visual task as our visual acuity is good at measuring horizontal displacement, which can ultimately be used to compare medians in this analysis.

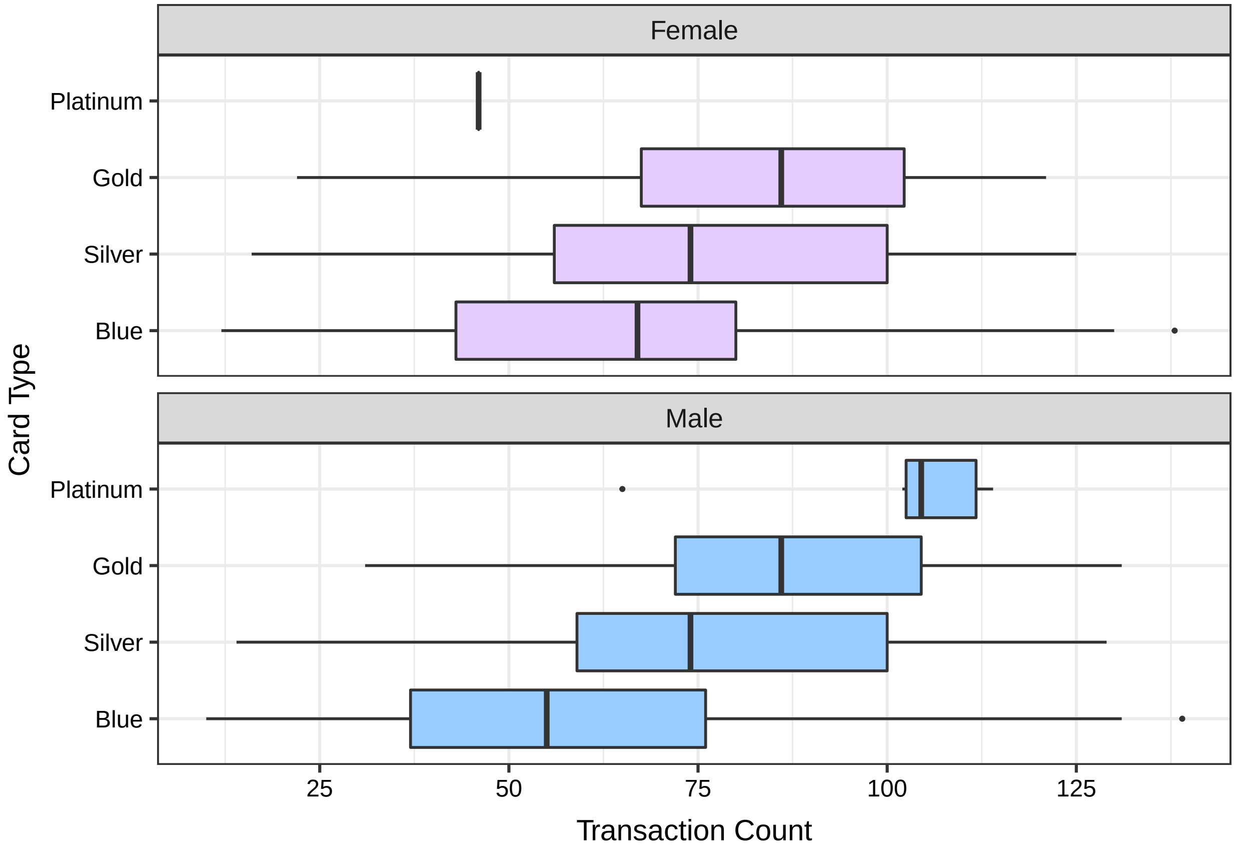
Analysis:

```r
#Question 2: Data Wrangling
bank_churners2 <- bank_churners %>%
  filter(Marital_Status == "Married") %>% #focus only on married customers
  mutate(Card_Category = fct_relevel(Card_Category,
  "Blue","Silver","Gold","Platinum")) #meaningful reordering

#Question 2: Data Wrangling + Summary Table
bank_churners2 %>%
  count(Card_Category, Gender) %>%
  pivot_wider(names_from = "Card_Category", values_from = "n") %>% # return card count per gender
  mutate( #proportion calculations: adding new columns
    Blue_Prop. = (Blue/(Blue+Silver+Gold+Platinum)),
    Silver_Prop. = (Silver/(Blue+Silver+Gold+Platinum)),
    Gold_Prop. = (Gold/(Blue+Silver+Gold+Platinum)),
    Platinum_Prop. = (Platinum/(Blue+Silver+Gold+Platinum))
    ) %>%
  select( #return these specific columns only
    Gender, Blue, Silver, Gold, Platinum, Blue_Prop., Silver_Prop.,
    Gold_Prop., Platinum_Prop.
    )
```

```
## # A tibble: 2 x 9
##   Gender  Blue Silver  Gold Platinum Blue_Prop. Silver_Prop. Gold_Prop.
##   <chr>  <int>  <int> <int>    <int>      <dbl>        <dbl>      <dbl>
## 1 F       2363     77    10        1      0.964       0.0314    0.00408
## 2 M       2070    129    31        6      0.926       0.0577    0.0139
## # ... with 1 more variable: Platinum_Prop. <dbl>
```

```
#Question 2: Visualization
bank_churners2 %>%
  ggplot(aes(x = Total_Trans_Ct, y = Card_Category, fill = Gender)) + #setting axes
  geom_boxplot(outlier.size = 0.4) +
  labs(x = "Transaction Count", y ="Card Type") + #relabeling axes titles
  facet_wrap(vars(Gender), #separate by gender
  ncol =1, labeller = as_labeller
  (c(`F` = "Female", `M` = "Male"))
  ) +
  scale_x_continuous( #setting numerical scale
    breaks = c(25, 50, 75, 100, 125)
  ) +
  scale_fill_manual(values=c("#E5CCFF", "#99CCFF")) + #color by gender
  theme_bw(12) +
  theme(
    axis.text = element_text(size = 9, color = "black"), #resizing and recoloring axes labels
    axis.title.x = element_text(size = 11, vjust = -0.5), #resizing, shifting axis title
    axis.title.y = element_text(size = 11, vjust = 2) #resizing, shifting axis title
  ) +
  theme(legend.position="none") #removing legend for 'Gender'
```



Discussion:

The first part of the question (summary table) looks at the proportion of card categories across sex. The proportion table reveals that both sexes overwhelmingly prefer the blue card against all other cards, at 93% and 96% blue card usage for males and females, respectively. For both sexes, silver is second most popular, followed by gold, then platinum. The proportions of card categories across sex differ only slightly: the greatest deviation is seen in proportions of blue, which is only a 3% difference. One explanation for this deviation could perhaps be explained by the societal wealth gap. As colors have different operational costs, perhaps men are more readily able to afford non-blue cards than women. That is, within this dataset, men own non-blue cards at higher proportions at each consecutive level when compared to women.

The second part of the question (visualization) looks at the effects of sex and card category on transaction count. Though it appears transaction count increases with card level, based on our summary table, only blue and silver card categories will be studied in this analysis due to sample size (n > 50). On average, male and female silver card holders have similar transaction counts. For blue card holders, on average, females have a higher transaction count than males. And lastly, on average, silver card holders have a higher transaction count than blue card holders for both sexes. One explanation for these trends is that, since card level might be associated with economic class, users with a higher card level have more transactions because they have more buying power / ability to spend. Secondly, out of those who own blue cards, females are more likely than males to use their card more. A reason for this may be, if blue represents the common class, perhaps females have higher transaction counts due to the fashion industry / are more likely to buy cosmetic products (ie. shop more than men) to model wealthier women. This explanation would simultaneously explain why the average transaction count has a sexual discrepancy only at the blue card level and not other card levels.