Project 2

Sookja Kang, sk26949

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:

`bank_churners`dataset is used to answer part 1 question. This dataset contains 10,127 bank consumers with 21 categories of their personal data. In this dataset, each row represents an individual bank consumer and each column represents a different type of bank related personal data (a total of 21: client number, attrition flag, age, gender, etc.).
To understand the relationship between attrition rate and income level, I am going to work with the two columns of Income_Category and Attrition_Flag.

1. Income_Category: There are 5 different income levels (Less than $40K, $40K - $60K, $60K - $80K, $80K - $120K, $120K +) as well as unknown data.
2. Attrition_Flag: Two types of customers (Existing and Attrited Customer) were reported.

Approach:

My approach is to understand whether the attrition rate is related to the income levels. First, I am going to present a table that shows the numbers of attrited customers and existing customers for the 5 different income levels. Next, I am going to visualize comparing the proportion of attrited customers and existing customers for each income level using horizontal stacked bars. These stacked bars allow to compare overall portion changes of existing and attrited costumers. However, it is difficult to make a proportion comparison in attrited customers for $40K - $60K, $60K - $80K, and $80K - $120K since the changes are similar and have different starting locations.

To present a table for the numbers of attrited customers and existing customers across the 5 different income levels

1. filter(): extract only 5 income level without "Unknown" from Income_Category
2. count(): count numbers for each subcategory of Income_Category and Attrition_Flag
3. pivot_wider(): making a wider table (Income_Category as far left column and Attrition_Flag as top row in the table)

To make horizontal stacked bars, the following functions will be used:

1. filter(): extract only 5 income level without "Unknown" from Income_Category
2. mutate(): rewrite the Income_Category column in a new order
3. fct_relevel(): reorder the Income_Category column by hand: "Less than $40K","$40K - $60K", "$60K - $80K","$80K - $120K","$120K +")
4. geom_bar(): creating stacked bars after counting each cell for Income_Category by Attrition_Flag

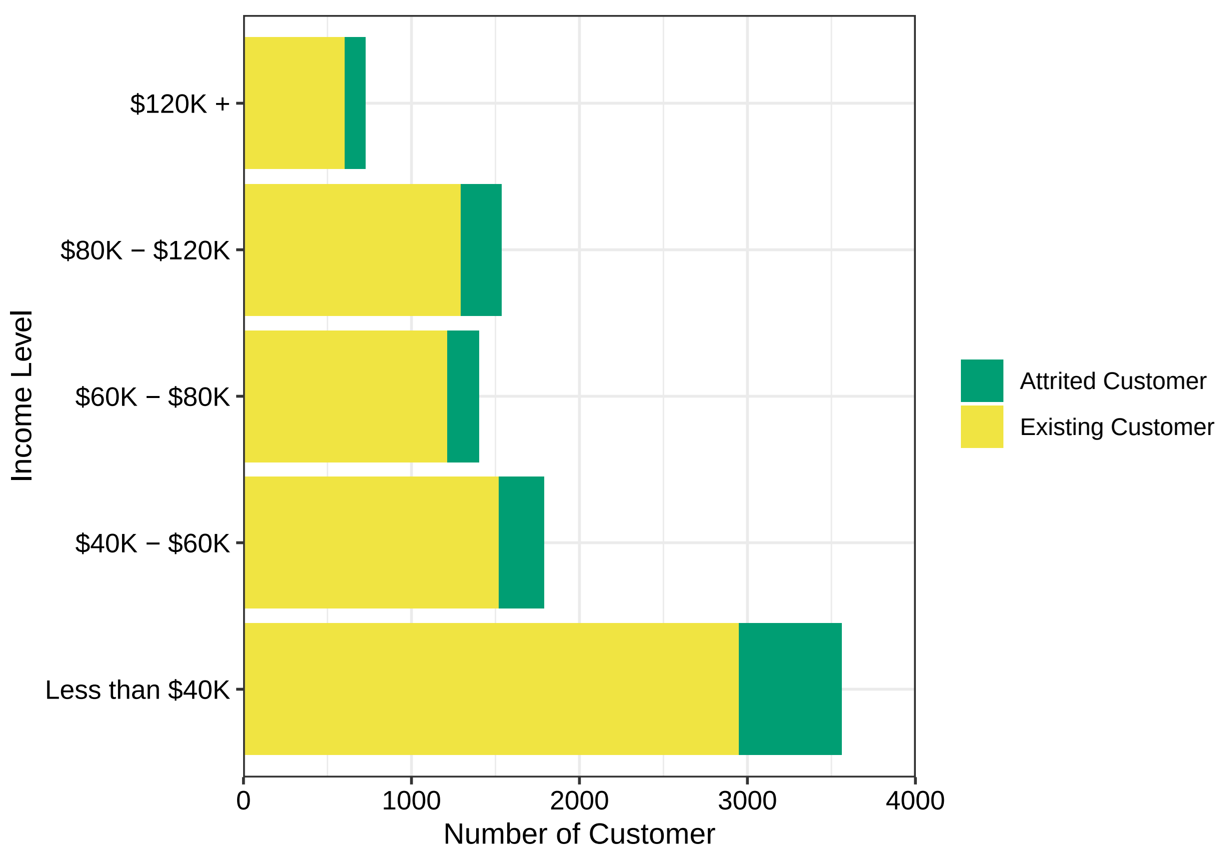
Analysis:

```
# To present a table for the numbers of attrited customers and existing customers across the 5 differen
Bank_SummaryTable1 <- bank_churners %>%
  filter(Income_Category != "Unknown") %>%
  count(Income_Category, Attrition_Flag) %>%
  pivot_wider(names_from = "Attrition_Flag", values_from = "n")


Bank_SummaryTable1
```

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

```
## 5 Less than $40K                    612                 2949
```

```r
# To make horizontal stacked bars
Bank_stacked_bars <- bank_churners %>%
  filter(Income_Category != "Unknown") %>%
  mutate(Income_Category = fct_relevel(Income_Category, "Less than $40K","$40K - $60K",
          "$60K - $80K","$80K - $120K","$120K +")) %>%
  ggplot(aes(y = Income_Category, fill = Attrition_Flag)) +
  geom_bar() +
  theme_bw() +
  scale_x_continuous("Number of Customer",
                     limits = c(0, 4000),
                     breaks = c(0, 1000, 2000, 3000, 4000),
                     expand = expansion(mult = c(0, 0))
                     ) +
  scale_y_discrete("Income Level") +
  scale_fill_manual(name = NULL,
                    values = c('Attrited Customer' = "#009E73", 'Existing Customer' = "#F0E442")) +
  theme(axis.text = element_text(color = "black", size = 10))

Bank_stacked_bars
```



Discussion:

These stacked bars show total numbers of customers across the 5 income levels. The numbers of customers tend to decrease as the income level increases. In these stack bars, the raw number of attrited customers in lower income levels are higher compared to ones in the higher income levels, which could show a relationship between income levels and the attrition rate. However, unless I compare the actual attrition rate per each *income level, I cannot make a definite conclusion of the relationship between income levels and the attrition rate from this analysis.*

Part 2

Question:

How do the numbers of different card types change across the income levels?

Introduction:

I am going to use the same dataset of `bank_churners` (including 10,127 bank consumers with 21 variables) that I used for the part 1 question. To answer this question, I am going to use the following two variables: 1. Card_Category: four different card types (Platinum, Gold, Silver, and Blue) were reported 2. Income_Category: reported using 5 different income levels (Less than \$40K, \$40K - \$60K, \$60K - \$80K, \$80K - \$120K, and \$120K +).

Approach:

My approach is to understand the proportion changes of different card types by the income levels. First, I will show each card type's number across the income levels in a wide table. Next, I will make a pie chart to visually compare different card types' proportion changes across the income levels. This pie chart allows showing each slice's proportion (representing card types) in a whole circle across the different income levels. However, it is difficult to present a very small portion of slices in a pie chart.

To present a wide table for each card type's number across the income levels:

1. filter(): extract only 5 income levels of Income_Category without "Unknown"
2. mutate(): rewrite the Income_Category column in a new order
3. fct_relevel(): reorder the Income_Category column by hand: "Less than \$40K","\$40K - \$60K", "\$60K - \$80K","\$80K - \$120K","\$120K +")
4. count(): count numbers for each subcategory of Card_Category and Income_Category
5. pivot_wider(): making a wide table (Card_Category as far left column and Income_Category as top row in the table)

To make pie charts, the following functions will be used:

1. filter(): extract only 5 income levels of Income_Category without "Unknown"
2. mutate(): rewrite the Income_Category column in a new order
3. fct_relevel(): reorder the Income_Category column by hand: "Less than \$40K","\$40K - \$60K", "\$60K - \$80K","\$80K - \$120K","\$120K +")
4. count(): count numbers for each subcategory of Card_Category and Income_Category
5. mutate(): make a new column (total_number) using the `n` column that created from count()
6. arrange() and -desc(): to sort the total_number by ascending count
7. fct_reorder(): to reorder the Card_Category column by the total_number
8. group_by(): to group by the Income_Category
9. mutate(): make new columns
    - the end_angle, start_angle, mid_angle for each pie slice
    - horizontal and vertical justifications for outer labels
10. ggplot(): to plot the pie_data
11. geom_arc_bar() to specify the exact location of the pie center in the x-y plane
12. coord_fixed(): to ensure that the pie is round
13. facet_wrap(): to create pie chart facets for each income level
14. theme_void(): to remove the x-y plane

Analysis:

```
# To present a wide table for each card type's number across the income levels:
Bank_Summary <- bank_churners %>%
```

```r
  filter(Income_Category != "Unknown") %>%
  mutate(Card_Category = fct_relevel(Card_Category,
                                    "Blue", "Silver", "Gold", "Platinum")) %>%
  mutate(Income_Category = fct_relevel(Income_Category, "Less than $40K","$40K - $60K",
        "$60K - $80K","$80K - $120K","$120K +")) %>%
  count(Card_Category, Income_Category)

Bank_Summary
```

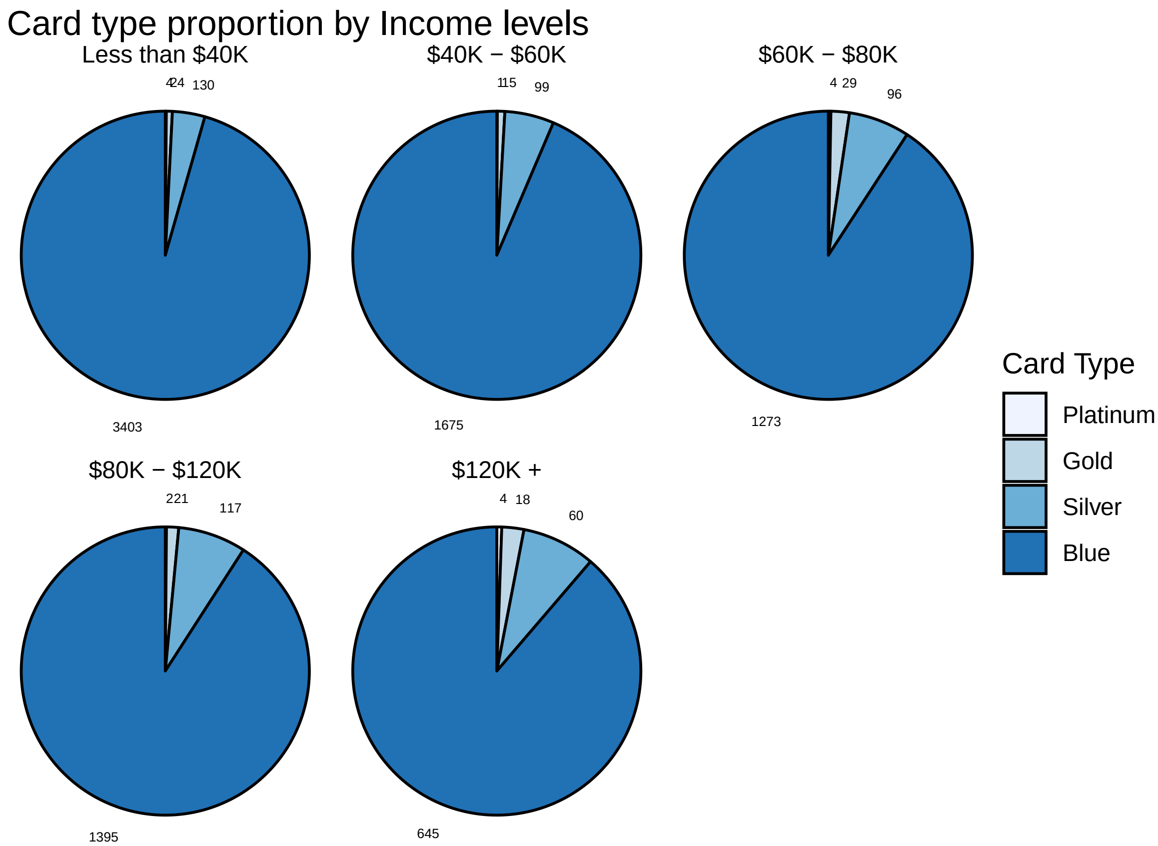
```
## # A tibble: 20 x 3
##    Card_Category Income_Category     n
##    <fct>         <fct>           <int>
##  1 Blue          Less than $40K   3403
##  2 Blue          $40K - $60K      1675
##  3 Blue          $60K - $80K      1273
##  4 Blue          $80K - $120K     1395
##  5 Blue          $120K +           645
##  6 Silver        Less than $40K    130
##  7 Silver        $40K - $60K        99
##  8 Silver        $60K - $80K        96
##  9 Silver        $80K - $120K      117
## 10 Silver        $120K +            60
## 11 Gold          Less than $40K     24
## 12 Gold          $40K - $60K        15
## 13 Gold          $60K - $80K        29
## 14 Gold          $80K - $120K       21
## 15 Gold          $120K +            18
## 16 Platinum      Less than $40K      4
## 17 Platinum      $40K - $60K         1
## 18 Platinum      $60K - $80K         4
## 19 Platinum      $80K - $120K        2
## 20 Platinum      $120K +             4
```

```r
Bank_SummaryTable2 <- Bank_Summary %>%
  pivot_wider(names_from = "Income_Category", values_from = "n")

Bank_SummaryTable2
```

```
## # A tibble: 4 x 6
##   Card_Category `Less than $40K` `$40K - $60K` `$60K - $80K` `$80K - $120K`
##   <fct>                    <int>         <int>         <int>          <int>
## 1 Blue                      3403          1675          1273           1395
## 2 Silver                     130            99            96            117
## 3 Gold                        24            15            29             21
## 4 Platinum                     4             1             4              2
## # ... with 1 more variable: `$120K +` <int>
```

```r
# To make pie charts:
Bank_data <- bank_churners %>%
  filter(Income_Category != "Unknown") %>%
  mutate(Income_Category = fct_relevel(Income_Category,
                                      "Less than $40K","$40K - $60K", "$60K - $80K","$80K - $120K","$1
  count(Card_Category, Income_Category) %>%
  mutate(total_number = n ) %>%
  arrange(-desc(total_number)) %>%
```

```r
  mutate(Card_Category = fct_reorder(Card_Category, total_number))



pie_data<- Bank_data %>%
  group_by(Income_Category) %>%
  mutate(end_angle = 2*pi*cumsum(n)/sum(n),
         start_angle = lag(end_angle, default = 0),
         mid_angle = 0.5*(start_angle + end_angle),
         hjust = ifelse(mid_angle > pi, 1, 0),
         vjust = ifelse(mid_angle < pi/2 | mid_angle > 3*pi/2, 0, 1))


ggplot(pie_data, aes(x0 = 0, y0 = 0, r0 = 0, r = 1,
                     start = start_angle, end = end_angle,
                     fill = Card_Category))+
  geom_arc_bar() +
  geom_text(size = 1.8,
            aes(x = 1.15 * sin(mid_angle),
                y = 1.2 * cos(mid_angle),
                label = total_number,
                hjust = hjust)) +
  coord_fixed() +
  facet_wrap(~Income_Category, ncol=3) +
  theme_void() +
  scale_fill_brewer(name = "Card Type") +
  ggtitle("Card type proportion by Income levels")
```



Discussion:

These pie charts show the proportion changes of different card types across the 5 income levels. As you see, higher income levels show higher proportions of Platinum, Gold, and Silver cards compared to the lowest income level (less than 40K). Thus, the income levels are related to the card types. However, it is challenging to precisely compare the proportions when the sizes of pie slices are similar or too small. As a result, it is difficult to compare proportions of card types for $60K - $80K and $80K - $120K.