Project 2

Wenxuan Zhou wz4388

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: We are working with the `bank_churners` dataset, which contains information about 10127 customers of a bank. Each row of the dataset represents information about a single customer. The dataset contains 21 columns that provide information like gender, education level, income, marital status, etc.

To determine whether attrition rate is related to income level or not, we will be working with the following columns:

1. `Income_Category`: Annual Income Category of the account holder
2. `Attrition_Flag`: whether the customer is an attrited customer or existing customer

Approach: Our approach is to first create a summary table containing the number of attrited and existing customers at each income level. Next, we will visualize the relative proportion of attrited and existing customers at each income category using a bar chart scaled to 100%. With the scaling, we can see the relative proportion based on the number of attrited and existing customers. The alternative ways to show the relative proportion is pie chart. However, the attrition rate at each income level is very similar. Pie chart is not the most appropriate visualization to show the subtle difference.

To create the summary table containing the number of attrited and existing customers at each income level, these functions will be applied:

1. `filter()` to eliminate the unknown income level
2. `group_by()` and `summarize()` to count the number of customers based on `Income_Category` and `Attrition_Flag`
3. `pivot_wider()` to extract the the number of customers based on `Attrition_Flag` and put them in new columns
4. `arrange()` to order `Income_Category` by value

To plot the relative proportion of attrited and existing customers at each income level, we use the following functions: 1. `filter()` to eliminate the income level "Unknown"

2. `fct_relevel()` to relevel `Income_Category` by value
3. `geom_bar()` to create a bar plot of the relative proportion of attrited and existing customers at each income level

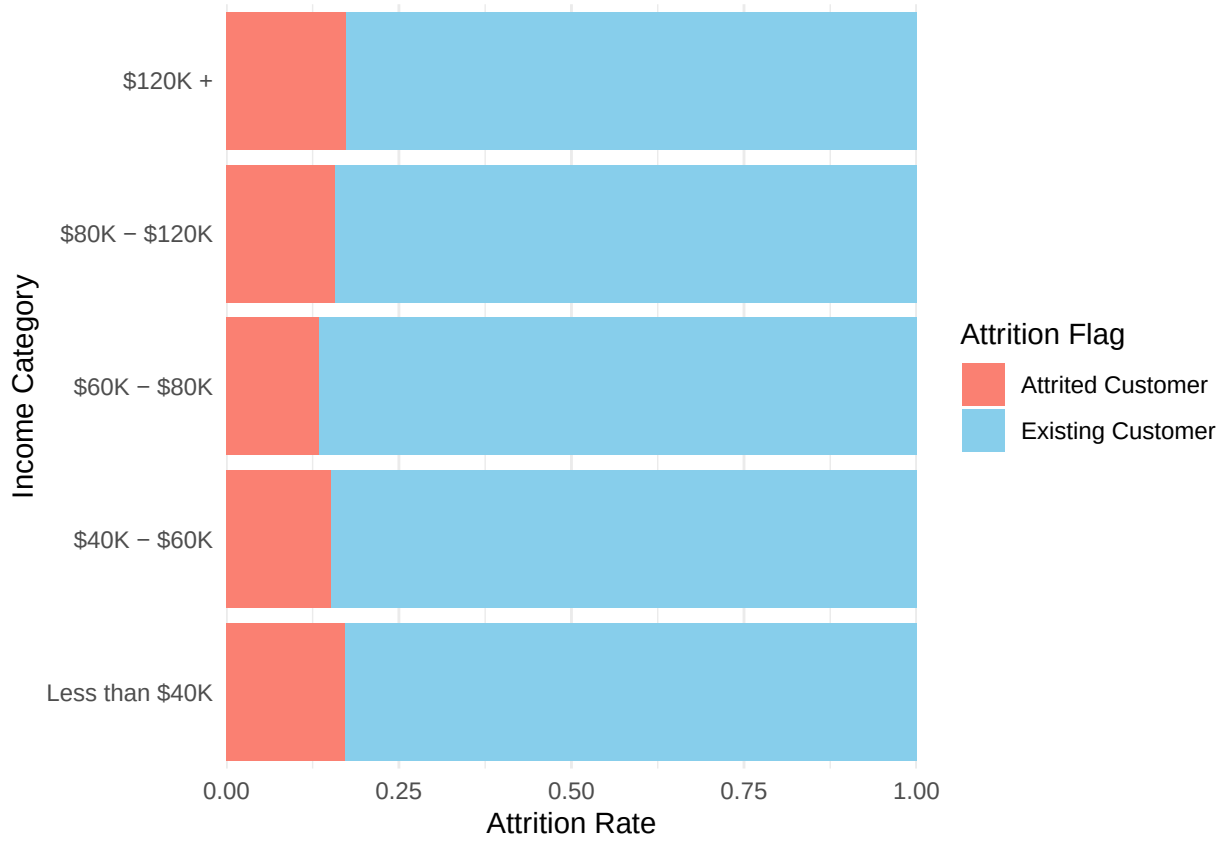
Analysis:

```
bank_churners %>% filter(Income_Category != 'Unknown') %>%
  group_by(Income_Category, Attrition_Flag) %>%
  summarize(n = n()) %>%
  pivot_wider(names_from = "Attrition_Flag", values_from = "n") %>%
  arrange(match(Income_Category, c('$120K +', '$80K - $120K', '$60K - $80K',
                                   '$40K - $60K', 'Less than $40K')))
```

```
## `summarise()` regrouping output by 'Income_Category' (override with `.groups` argument)
```

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

```r
bank_churners %>% filter(Income_Category != 'Unknown') %>%
  ggplot(aes(y = fct_relevel(Income_Category, 'Less than $40K', '$40K - $60K',
                             '$60K - $80K', '$80K - $120K', '$120K +'),
             fill = Attrition_Flag)) +
  geom_bar(position = position_fill(reverse = TRUE)) +
  scale_x_continuous(name = 'Attrition Rate', expand = c(0, 0)) +
  scale_y_discrete(name = 'Income Category', expand = c(0, 0.5)) +
  scale_fill_manual(values = c('Existing Customer' = "#87CEEB",
                               'Attrited Customer' = "#FA8072"),
                    name = 'Attrition Flag') +
  theme_minimal()
```



Discussion: Looking at the bar plot, we can see that the attrition rates of each income level are similar. Customers in the income level of $60K - $80K have the lowest attrition rate, and it increases as the income increases or decreases from that category. The attrition rate is below 0.2 for every income level. For conclusion, I think attrition rate is not related to income level.

Part 2

Question: Does a customer's education level affect the card type he/she owns?

Introduction: To determine whether education level is related to card category or not, we will be working with the following columns:

1. `Education_Level`: Educational Qualification of the account holder
2. `Card_Category`: Type of card (Blue, Silver, Gold, Platinum)

Approach: Our approach is to first create a summary table containing the number of customers with different education levels for each card category. Next, we will visualize the relative proportion of customers with different education levels at each card category by using faceted pie charts. There are not many card types and education levels, so pie charts are appropriate for comparing the proportions. The alternative visualization is stacked bar plot. However, pie chart is more visually appealing for small datasets, so faceted pie charts are more appropriate.

To create the summary table containing the number of customers with different education levels for each card category, these functions will be applied:

1. `filter()` to eliminate unknown education level
2. `group_by()` and `summarize()` to count the number of customers based on `Card_Category` and `Education_Level`
3. `pivot_wider()` to extract the the number of customers based on `Education_Level` and put them in new columns
4. `replace_na()` replace the NAs in the dataset with 0
5. `arrange()` to order `Card_Category` by card type
6. `select()` to order the columns based on education level

To plot the relative proportion of customers with different education levels at each card category, we use the following functions:

1. `filter()` to eliminate unknown education level
2. `group_by()` and `summarize()` to count the number of customers based on `Card_Category` and `Education_Level`
3. `pivot_wider()` to extract the the number of customers based on `Education_Level` and put them in new columns
4. `replace_na()` replace the NAs in the dataset with 0
5. `arrange()` and `desc()`: to sort the table by descending count
6. `transform()` to reorder the pie charts
7. `geom_arc_bar()` to create pie charts of the relative proportions
8. `face_wrap()` to create pie charts facets for each card category
9. `coord_fixed()` to fix the coordinate system

Analysis:

```
bank_churners %>% filter(Education_Level != 'Unknown') %>%
  group_by(Education_Level, Card_Category) %>%
  summarize(n = n()) %>%
  pivot_wider(names_from = "Education_Level", values_from = "n") %>%
  replace_na(list('College' = 0)) %>%
  arrange(match(Card_Category, c('Blue', 'Silver', 'Gold', 'Platinum'))) %>%
  select(Card_Category, Uneducated, 'High School', College, Graduate,
         'Post-Graduate', Doctorate)
```

```
## `summarise()` regrouping output by 'Education_Level' (override with `.groups` argument)
```

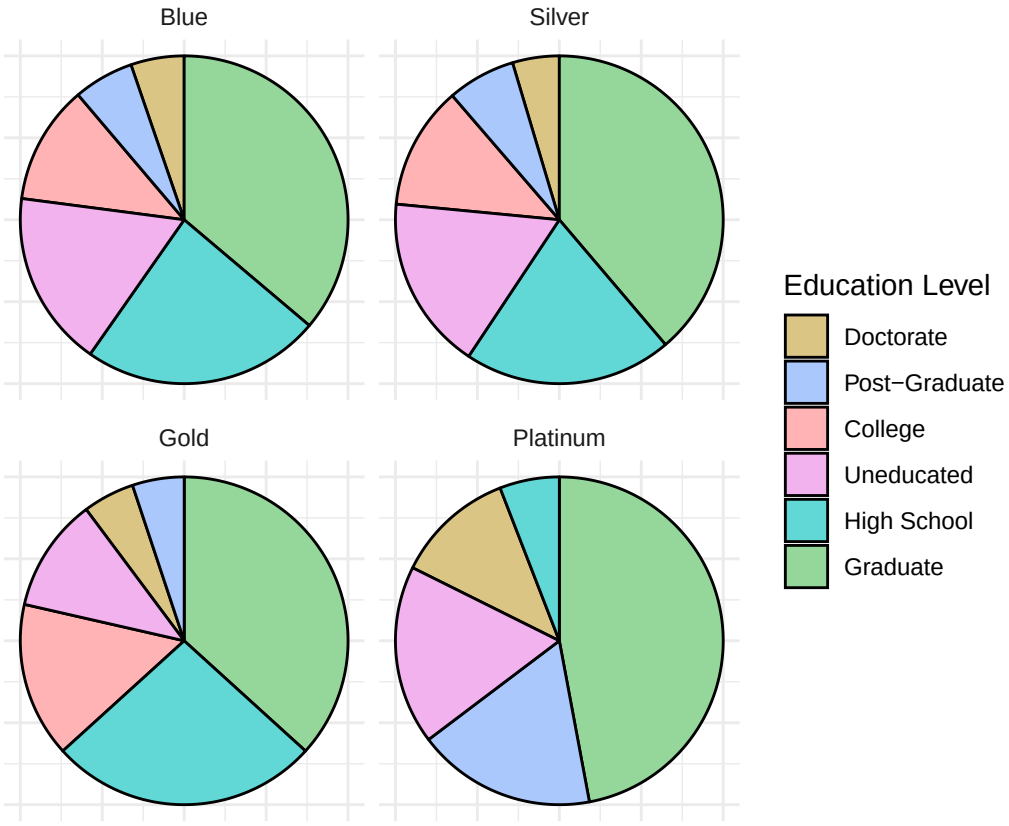
```
## # A tibble: 4 x 7
##   Card_Category Uneducated `High School` College Graduate `Post-Graduate`
##   <chr>              <int>         <int>   <dbl>    <int>           <int>
## 1 Blue                1391          1888     940     2899             476
```

```
## 2 Silver                82            98       58       185              32
## 3 Gold                  11            26       15        36               5
## 4 Platinum               3             1        0         8               3
## # ... with 1 more variable: Doctorate <int>
```

```
bank_churners_card <- bank_churners %>% filter(Education_Level != 'Unknown') %>%
  group_by(Education_Level, Card_Category) %>%
  summarize(n = n()) %>%
  replace_na(list('College' = 0)) %>%
  arrange(desc(n))
```

```
## `summarise()` regrouping output by 'Education_Level' (override with `.groups` argument)
```

```
ggplot(transform(bank_churners_card,
      Card_Category = factor(Card_Category, levels=c('Blue', 'Silver',
                                                     'Gold', 'Platinum')))) +
  aes(
    x0 = 0, y0 = 0, # position of pie center
    r0 = 0, r = 1,  # inner and outer radius
    amount = n, # size of pie slices
    fill = Education_Level
  ) +
  geom_arc_bar(stat = "pie") +
  facet_wrap(~Card_Category) +
  coord_fixed() +
  labs(x = NULL, y = NULL) +
  scale_fill_discrete_qualitative(palette = "Set 3", name = 'Education Level',
                                  breaks = bank_churners_card$Education_Level) +
  guides(fill = guide_legend(reverse = TRUE)) +
  theme_minimal() +
  theme(axis.text.x=element_blank(), axis.text.y=element_blank())
```



Discussion: Looking at the faceted pie charts, we can see that the relative proportions of customers with different education levels are very similar for customers with Blue and Silver cards. For Gold card holders, there are more customers with College and High School educational background. For Platinum card holders, the proportions of customers with Doctorate, Post-Graduate, and Graduate educational background increase compared to other card holders, and no customers with College educational background owns the Platinum card. In conclusion, different card categories have different relative proportions of customers with different educational background, and a customer's education level can affect the card type he/she owns.