Project 2

Xiangrui Pan

xp742

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.

Introduction: We are working with `bank_churchers` dataset. This dataset contains 10127 customers' information such as their age, salary, marital_status, credit card limit, credit card category, etc. Those are stored in about 18 columns (features) and each row represents a single customer.

To answer questions asked in part 1, we need to create a summary table and a corresponding visualization graph. The summary table needs to have three columns that include income category, existing customers, and attrited customers. From the instruction above, we will only need two columns from the original dataset: 1) `Income_Category`: `Income_Category` shows the range of each customer's yearly salary and there are five types of salary range: less than $40K, $40K-$60K, $60K-$80K, $80K-$120K, and above $120K+. 2) `Attrition_Flag`: `Attrition_Flag` denotes two types of customers: existing customers and attrited customers. Our visualization will show the relative proportion of existing and attrited customers at each income level.

Approach: Our approach to this question is to first create a summary table that contains income category, existing customer, and attrited customer. Next, we will visualize the relative proportion of existing customer and attrited customer based on their salaries so that we can see how the attrited customers get distributed based on their income level. Notice that there are some customers' salaries are categorized as Unknown. For the purpose of simplicity, we will remove those customer from our analysis. In my analysis, I will separate the data wrangling part and visualization part, so we can run them independently.

To look at the summary table (data wrangling part), I will use the following functions.

1) `filter()`: to remove the customers who have "unknown" salaries from our dataset.
2) `select()`: to pick the columns that we will be working with
3) `mutate()`: to rewrite `Income_Category` column in a new order
4) `fct_relevel()`: to manually set income level from the lowest to the highest
5) `group_by()`: to take `Income_Categoty` for the further operation
6) `summarize()`: receive the column from `group_by` and add two more columns into the dataset that includes `Income_Category`

To make the graph of relative proportion of existing and attrited customers at each income level, besides all the functions I used earlier for the summary table, I will introduce the new functions here.

1) `geom_col(position = "fill")`: I use this function to draw the stacked bar plot for attrited customer and existing customer for each salary category. The argument `position = "fill"` is for setting up the relative proportion. Without `position = "fill"`, `geom_col()` will only show stacked bar for counts.
2) `xlab()`: rename the x axis
3) `ylab()`: rename the y axis
4) `lab()`: rename the legend
5) `scale_fill_manual()`: manually select color for the plot
6) `theme_minimal()`: make the plot with no background annotations
7) `theme()`: allows us to manually adjust the features of axis text.

Analysis:

```
#Data wrangling to show the summary table
summary_table = bank_churners%>%
filter(Income_Category != "Unknown")%>%
select(Attrition_Flag, Income_Category)%>%
mutate(
        Income_Category = fct_relevel(Income_Category,
        "Less than $40K","$40K - $60K",
        "$60K - $80K","$80K - $120K","$120K +"),
       )%>%

group_by(Income_Category)%>%
```

```
  summarize(
   # n = n(),
    Attrited_Customer = sum(Attrition_Flag == "Attrited Customer"),
    Existing_Customer = sum(Attrition_Flag == "Existing Customer"),
    #can check the rate by uncommenting the line of code below
    #Attrited_Rate = Attrited_Customer/(Attrited_Customer+Existing_Customer)
  )
```

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

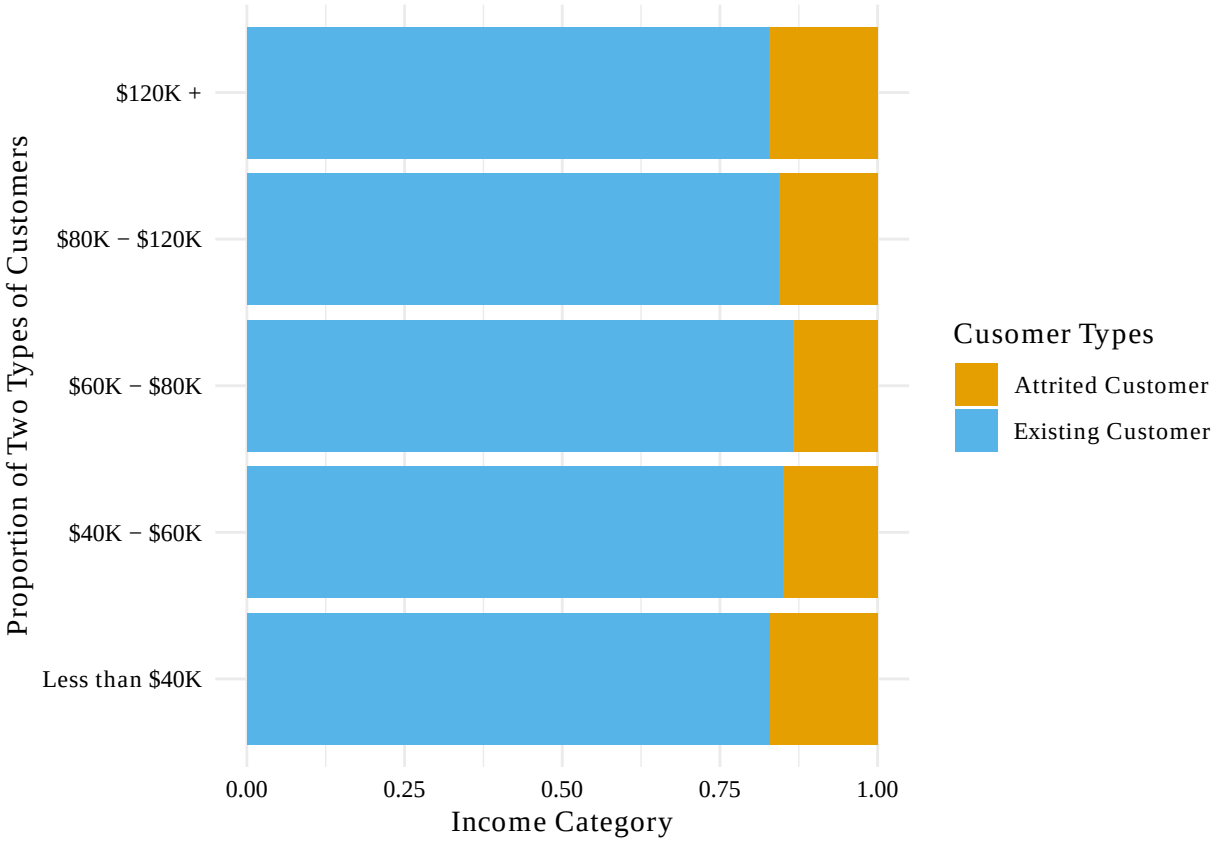
```
  #call the summary table that contains 3 columns
  summary_table
```

```
## # 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
```

```
  #Visualization
  bank_churners%>%
  filter(Income_Category != "Unknown")%>%
  select(Attrition_Flag, Income_Category)%>%

  mutate(
         Income_Category = fct_relevel(Income_Category,
          "Less than $40K","$40K - $60K",
          "$60K - $80K","$80K - $120K","$120K +")
         )%>%
  #need to group by both column, otherwise cannot call later
  group_by(Attrition_Flag, Income_Category)%>%
  summarize(
    n=n(),
    Attrited_Customer = sum(Attrition_Flag == "Attrited Customer"),
    )%>%
  ggplot(aes(y=Income_Category, x=n, fill=factor(Attrition_Flag)))+
    geom_col(position = "fill")+
    xlab("Income Category")+
    ylab("Proportion of Two Types of Customers")+
    labs(fill = "Cusomer Types")+
    scale_fill_manual(values = c( "#E69F00", "#56B4E9"))+
    theme_minimal()+
    theme(
        axis.text = element_text(
                    size = 8.5,
                    color = "black"
                    ),
        text = element_text(
                    family = "Palatino")
        )
```

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



Discussion: From the summary table, we can easily see that the total number of customers who's salary is less than $40k is the greatest, which is 3561 and takes approximately 39.5% of all customers. We can also observe the decreasing trend of the number of customers as income category go from "less than $40k" to "$120k", so we can conclude the distribution of this summary table is skewed to the right. However, since we did not include proportion in our summary table, it is a bit hard to answer our question.

In the visualization, we can see that relative proportion of existing and attrited customers at each income level. Clearly, existing customers at each level of income category is more than 75% and there is only sightly difference in each relative proportion. This indicates that there is a very similar proportion of customers leaving their credit card service in every level. In fact, we can see the relative proportion for attrited customer by uncommenting line 106 in the code. With all the information we have so far, we conclude that income level is not an issue that affects the attrited rate.

Part 2

Question: Is there a relationship between marital status and being a customer with this bank based on three most-commonly-seen degrees?

Introduction In part 2, we are interested in seeing the relationship between marital status and being a customer with this bank according to 3 different degree levels. To answer the question in part 2, we will create a summary table and a corresponding visualization graph. This summary table will have the customers' marital status and the corresponding degree levels. To get started, we need pick several columns from the "'bank_churners"".

1) `Months_on_books`: This columns shows the period of relationship that a customer with the bank, unit in month.

2) `Education_Level`: the customers' degree level

3) `Marital_Status`: the customers' marital status

Approach: Our approach in part 2 is same to part 1. We will first have a summary table that contain the information about the customers' marital status, degree level, and the median value (in month) of being in relationship with the bank. Since there are unknown variables in both `Marital_Status` and `Education_Level`, we will remove all the unknowns for the sake of simplicity from both columns. Moreover, there are too many types of degree levels, so we only want three of them that are most representative and most seen in everyday life: "High School","Graduate",and "Doctorate". Just like part 1, I will separate the table and the graph.

To look at the summary table (data wrangling part), I will use the following functions.

1) `filter()`: only pick customers who have High School degree, Bachelor degree (Graduate), and PhD degree (Doctorate).

2) `select()`: select the three columns from `bank_churner` dataset

3) `mutate()`: arrange the degree levels in ascending order

4) `group_by()`: to take `Education_Level` for the further operation

5) `summarize()`: put new columns with education level. `Average_Booking` is the median period that the customers has the relationship with the bank at each degree level. The rest of columns are marital statuses: `Married`,`Single`, and `Divorced`

In the visualization part, I used some new functions in the presence of previous functions. I will introduce these new functions:

1) `geom_boxplot()`: this functions give us box plot of martial status and period of relationship

2) `xlab()`: allows us to change name in x axis

3) `ylab()`: allows us to change name in y axis

4) `facet_wrap()`: to create box plot facets for each degree level

5) `theme_bw()`: to change the background of the box plots

6) `theme()`: allows us to manually adjust the features of axis text.

7) `scale_fill_manual()`: to manually select color for the box plots

Analysis:

```
#Summary Table
bank_churners%>%
  #we only need these three degree level
  filter(
         Education_Level == c("High School","Graduate","Doctorate")
         )%>%
  select(
    Months_on_book,
    Education_Level,
    Marital_Status
    )%>%
  mutate(
    Education_Level =
      fct_relevel(Education_Level,
                  "High School",
                  "Graduate","Doctorate")
    )%>%
```

```
  group_by(Education_Level)%>%
  summarize(
    Average_Booking = median(Months_on_book),
    Married = sum(Marital_Status == "Married"),
    Single = sum(Marital_Status == "Single"),
    Divorced = sum(Marital_Status == "Divorced")
  )
```

```
## Warning in Education_Level == c("High School", "Graduate", "Doctorate"): longer
## object length is not a multiple of shorter object length
```

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

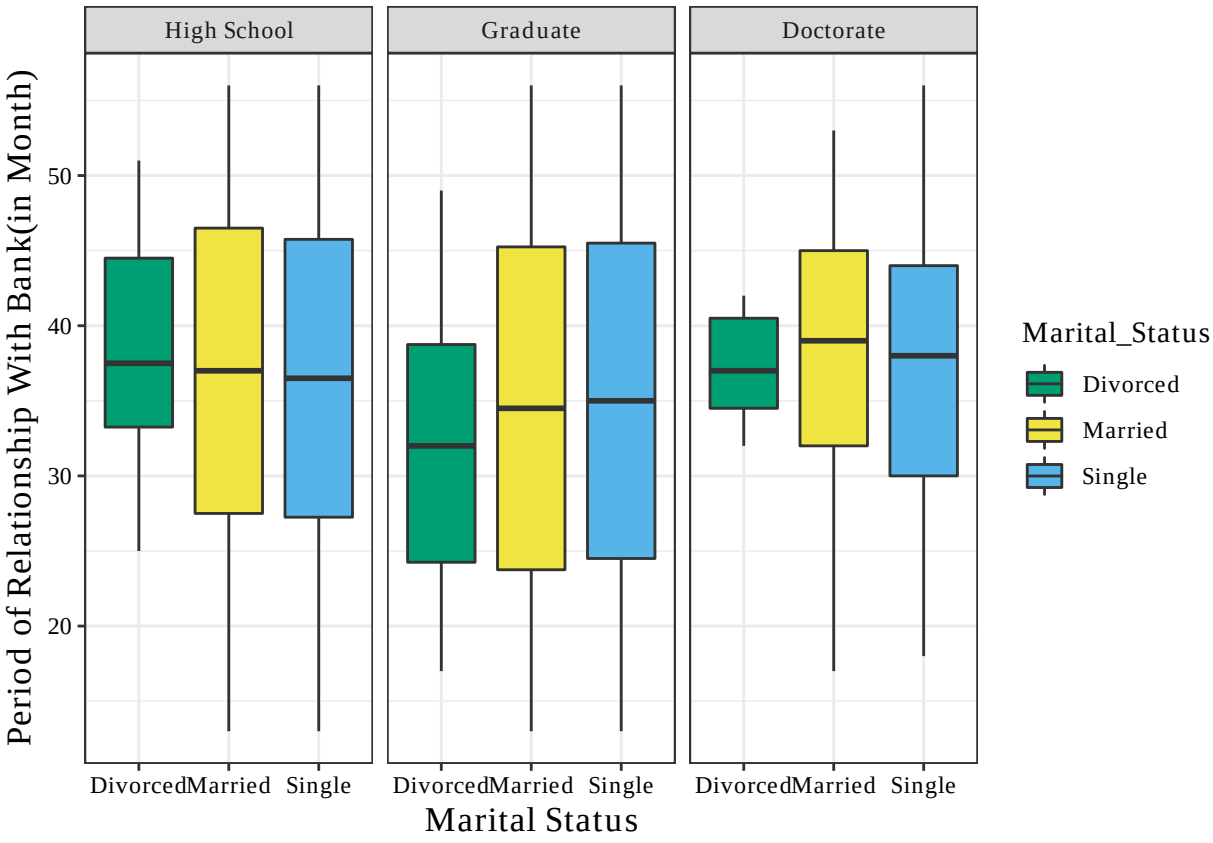
```
## # A tibble: 3 x 5
##   Education_Level Average_Booking Married Single Divorced
##   <fct>                     <dbl>   <int>  <int>    <int>
## 1 High School                  36     319    239       45
## 2 Graduate                     36     504    394       72
## 3 Doctorate                    36      66     57       10
```

```
#Visualization
bank_churners%>%
  filter(Marital_Status != "Unknown",
         Education_Level == c("High School","Graduate","Doctorate")
         )%>%
  select(
    Months_on_book,
    Education_Level,
    Marital_Status
    )%>%
  mutate(
    Education_Level =
      fct_relevel(Education_Level,
                  "High School",
                  "Graduate","Doctorate")
    )%>%
  group_by(Months_on_book, Education_Level,Marital_Status)%>%
  summarize(
    Married = sum(Marital_Status == "Married"),
    Single = sum(Marital_Status == "Single"),
    Divorced = sum(Marital_Status == "Divorced")
  )%>%
  ggplot(aes(y=Months_on_book,x=Marital_Status, fill=Marital_Status))+
  #using boxplot to show the difference
  geom_boxplot()+
  xlab("Marital Status")+
  ylab("Period of Relationship With Bank(in Month)")+
  #using facet function to separete each degree level
  facet_wrap(vars(Education_Level))+
  scale_fill_manual(
    values = c("#009E73", "#F0E442","#56B4E9")
    )+
  theme_bw()+
  theme(
        axis.text = element_text(
```

```
                size = 8.5,
                color = "black"
                ),
        axis.title = element_text(
                size = 13
        ),
        text = element_text(
                family = "Palatino")
        )
```

```
## Warning in Education_Level == c("High School", "Graduate", "Doctorate"): longer
## object length is not a multiple of shorter object length
```

```
## `summarise()` regrouping output by 'Months_on_book', 'Education_Level' (override with `.groups` argu
```



‘‘‘

Discussion: From the summary table, we can see the median period of being in relationship with the bank is 36 months for these 3 degree levels. However, if we look closely at each degree level in the box plot graphs, the median value is differed. In other words, if we observe the marital status at each degree, then we can see the medians are not exact to 36 months. Regardless of the degree level, we are able to see the period of relationship between customers and bank and it shows the box plot of customers who are categorized as “Divorced” is less spread out than the other two. In the language of statistics, we would say the range of period for “Divorced” customers is shorter. In fact, this is true for all three degrees. With those information, we are yet to fully answer the question.

From the first graph(high school), it is not hard to see that median is almost 36 months in all three martial statuses. The customers who are denoted as “Divorced” have the lowest median value in the second

graph(Graduate) and the other two statuses seem to have the same median value. Lastly, all the median values for the three statuses are different for the customers who have PhD degree. For the customers with high school degree and college degree, “Married” and “Single” have almost the same box plots, but that is not true for customers who possess PhD degree. Hence, we can conclude that there is relationship between marital status and period of being a customer with this bank in different degree level, and if the customers have college degree, then he or she will likely to have the lowest median month for being the customer with this bank.