Project 2

Cristian Sigala crs4565

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: Today we'll be using the bank_churners dataset in order to explore some data about credit card customers in a bank. We'll be using this data to discover if attrition rate is related to income level of these customers. For this problem, we will be looking at multiple variables such as "Attrition_Flag" which is the status of the customer encoded in Existing Customer/Attrited Customer. The next variable is "Income_Category" which describes the income level of the customer in five groups "Less than 40K, 40-60k, 60-80k, 80-120k, and 120k +."

Approach: Before we can start answering the question, we need to create a summary table. Therefore, the data must first be wrangled to view counts of these different customers in their respective income level. After that we are going to group together these two columns and then count the different statuses of these customers. After that we will then pivot wider in order to get the desired summary table. In order to answer this question, we will need to create a stacked bargraph to compare the proportions of attrition and existing customers based on income level. Stacked bargraphs are the most effective to compare distributions because it allows us to easily compare the proportions between attrited and existing customers on one bar while still separating by income level. This visualization makes it easy to compare overall rates with the different groups.

Analysis:

```
data <- bank_churners %>%
  filter(Income_Category != "Unknown") %>% #Delete Unknowns
  select(Income_Category, Attrition_Flag)%>% #Selected desired variables
  mutate(Income_Category= fct_relevel(
    Income_Category, "Less than $40K", "$40K - $60K", "$60K - $80K",
    "$80K - $120K", "$120K +")) %>% #Change Ordering
  group_by(Income_Category, Attrition_Flag) %>% #Keep order of data
  summarise(n=n()) %>% #adding count variable
  pivot_wider(names_from="Attrition_Flag", values_from="n") #pivoting based on Attrition_Flag
```

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

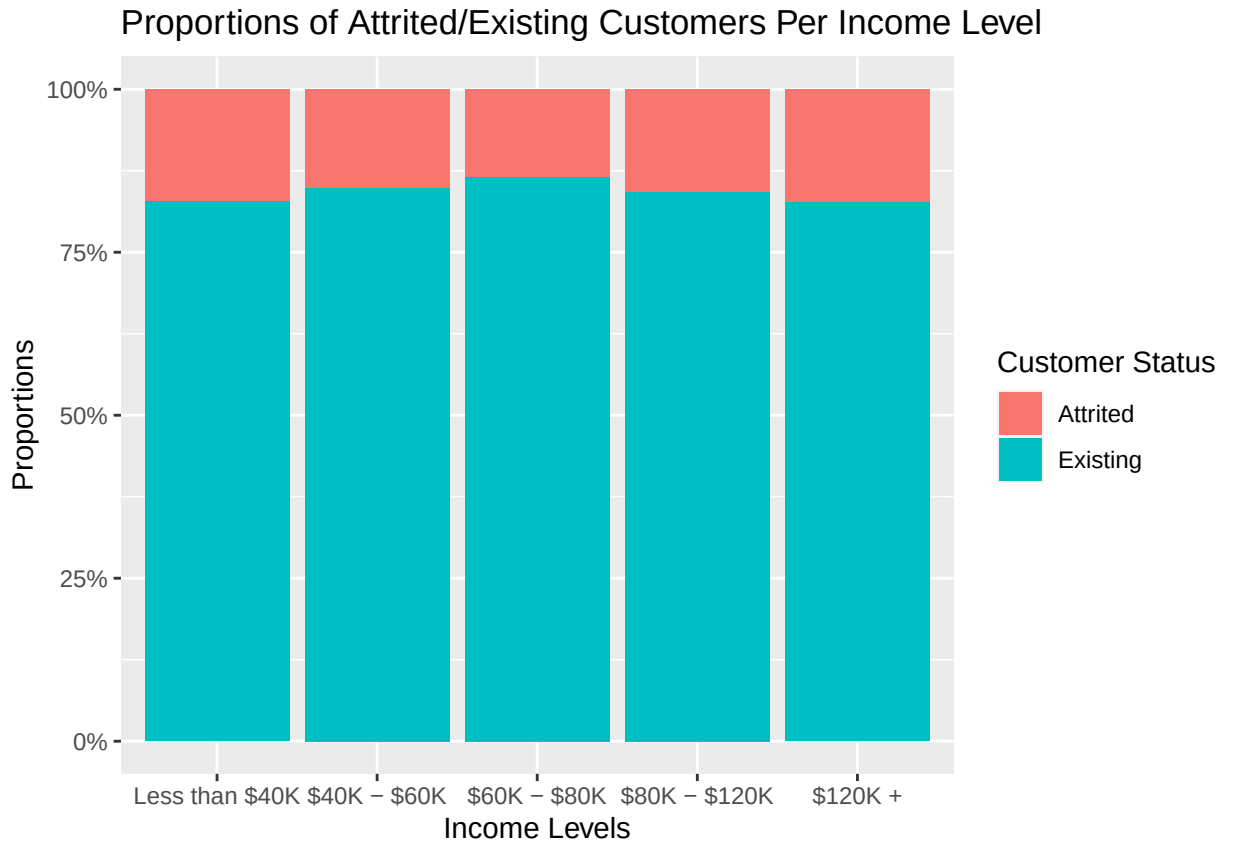
```
head(data)
```

```
## # A tibble: 5 x 3
## # Groups:   Income_Category [5]
##   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
```

```
#Graph Manipulation
data1 <- bank_churners %>%
  filter(Income_Category != "Unknown") %>%
  mutate(Income_Category = fct_relevel(
    Income_Category, "Less than $40K", "$40K - $60K", "$60K - $80K",
    "$80K - $120K", "$120K +" ))


data1 %>% ggplot(aes(x = Income_Category, fill = Attrition_Flag)) + #Graph specifications
  geom_bar(position = "fill") +
  #Make into stacked bargraph and position=fill automatically calculates proportions
```

```
  ylab("Proportions") + #Change y axis
  xlab("Income Levels") + #Change x axis
  ggtitle("Proportions of Attrited/Existing Customers Per Income Level") + #change title
  scale_fill_discrete(name = "Customer Status", labels = c("Attrited", "Existing")) +
  #Change legend title/labels
  scale_y_continuous(labels = percent) #convert y axis ticks to percents
```



Discussion: After constructing our stacked bar graph, attrition rates between all these groups seem very similar. Therefore, there is not a significant difference of attrition rate between the different income levels. It is important to note that counts are not displayed so we don't know how different the counts are between the groups.

Part 2

Question: Is there a difference in credit line distribution between different education levels and gender?

Introduction: While still using the bank_churners dataset, we will now use "Credit_Limit", "Education_Level", and "Gender" to see if there are any differences of credit limit per education level and gender. "Credit_Limit" is a numerical variable that tells us the credit limit of the individual. "Education_Level" which describes the education level of the customer in six groups "Uneducated, High School, College, Graduate, Post-Graduate, and Doctorate." Lastly, "Gender" gives us the gender of the customer with either "Male" or "Female."

Approach: To answer our question, I wanted to create boxplot series based on education level. After that I want to facet it over gender to get a clear visualization of credit line distributions. Boxplots are the best graph to use because it gives us the ability to see the summary statistics of credit line per education level all

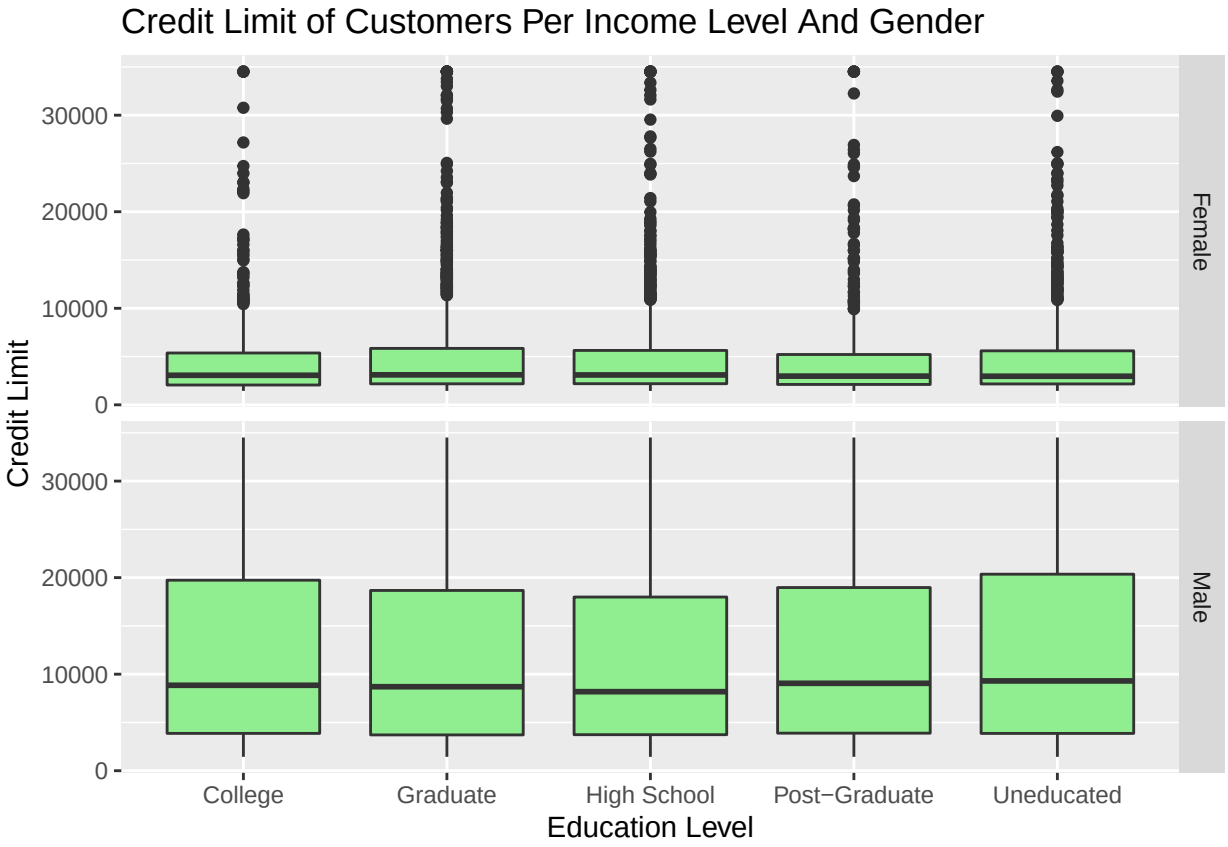
in one graph. After faceting over gender, we should be able to see if there are any differences of credit line between education level and gender.

Analysis:

```
data3 <- bank_churners %>%
  filter(Marital_Status != "Unknown") %>% #Delete Unknowns
  filter(Education_Level != "Unknown") %>% #Delete Unknowns
  select(Gender, Education_Level, Credit_Limit) # Deleting other variables


data3$Education_Level[data3$Education_Level=="Doctorate"] <- "Post-Graduate" #Fix data flaw
data3$Gender[data3$Gender=="M"] <- "Male" #Data rename
data3$Gender[data3$Gender=="F"] <- "Female" #Data rename

data3 %>% ggplot(aes(x = Education_Level, y = Credit_Limit)) + #Graph specifications
  geom_boxplot(fill = "Light Green") + #Creating boxplots
  facet_grid(vars(Gender)) + #Faceting over Gender
  ylab("Credit Limit") + #Change y axis
  xlab("Education Level") + #Change x axis
  ggtitle("Credit Limit of Customers Per Income Level And Gender") #Give title
```



Discussion: To answer our question, there is not a difference in credit line distributions between education level but there are differences based on gender. When only looking at education level, credit line distributions are the same for both men and women. However, when factoring gender, there is a clear difference in credit line distributions between men and women. Consistently when controlling education level, men have a higher average and range of credit limit compared to woman. From this data, we can conclude there is a difference

in credit line distribution based on gender but not education level.