Mini Project 1

Justine Filion and Sean Casey

January 28 2022

Step 1 : Describe the Dataset

The data used in the following analysis was provided to us by the Department of Information Management from Chung Hua University and the Department of Civil Engineering from Tamkang University which are both located in Taiwan. It was donated as of January 26th 2016 but refers to data collected from April to September 2005. The purpose of this dataset is to provide insight on the relationship between different explanatory variables and the probability of default. It contains 23 explanatory variables, all of which are integers, as well as a response variable which takes a value of 1 if the client defaulted in the following month and 0 if not. This dataset was made available to us via the UCI Machine Learning Repository which can be accessed through this link : link.

Step 2 : Load the Dataset

```
df <- read_excel('defaultofcreditcardclients.xls', skip = 1)
```

Step 3 : Explore the Dataset

```
str(df)
```

```
## tibble [30,000 x 25] (S3: tbl_df/tbl/data.frame)
##  $ ID                        : num [1:30000] 1 2 3 4 5 6 7 8 9 10 ...
##  $ LIMIT_BAL                 : num [1:30000] 20000 120000 90000 50000 50000 50000 500000 100000 1400
##  $ SEX                       : num [1:30000] 2 2 2 2 1 1 1 2 2 1 ...
##  $ EDUCATION                 : num [1:30000] 2 2 2 2 2 1 1 2 3 3 ...
##  $ MARRIAGE                  : num [1:30000] 1 2 2 1 1 2 2 2 1 2 ...
##  $ AGE                       : num [1:30000] 24 26 34 37 57 37 29 23 28 35 ...
##  $ PAY_SEP                   : num [1:30000] 2 -1 0 0 -1 0 0 0 0 -2 ...
##  $ PAY_AUG                   : num [1:30000] 2 2 0 0 0 0 0 -1 0 -2 ...
##  $ PAY_JUL                   : num [1:30000] -1 0 0 0 -1 0 0 -1 2 -2 ...
##  $ PAY_JUN                   : num [1:30000] -1 0 0 0 0 0 0 0 0 -2 ...
##  $ PAY_MAY                   : num [1:30000] -2 0 0 0 0 0 0 0 0 -1 ...
##  $ PAY_APR                   : num [1:30000] -2 2 0 0 0 0 0 -1 0 -1 ...
##  $ BILL_AMT_SEP              : num [1:30000] 3913 2682 29239 46990 8617 ...
##  $ BILL_AMT_AUG              : num [1:30000] 3102 1725 14027 48233 5670 ...
##  $ BILL_AMT_JUL              : num [1:30000] 689 2682 13559 49291 35835 ...
##  $ BILL_AMT_JUN              : num [1:30000] 0 3272 14331 28314 20940 ...
##  $ BILL_AMT_MAY              : num [1:30000] 0 3455 14948 28959 19146 ...
##  $ BILL_AMT_APR              : num [1:30000] 0 3261 15549 29547 19131 ...
##  $ PAY_AMT_SEP               : num [1:30000] 0 0 1518 2000 2000 ...
##  $ PAY_AMT_AUG               : num [1:30000] 689 1000 1500 2019 36681 ...
```

```
##  $ PAY_AMT_JUL               : num [1:30000] 0 1000 1000 1200 10000 657 38000 0 432 0 ...
##  $ PAY_AMT_JUN               : num [1:30000] 0 1000 1000 1100 9000 ...
##  $ PAY_AMT_MAY               : num [1:30000] 0 0 1000 1069 689 ...
##  $ PAY_AMT_APR               : num [1:30000] 0 2000 5000 1000 679 ...
##  $ default payment next month: num [1:30000] 1 1 0 0 0 0 0 0 0 0 ...
```

```
summary(df)
```

```
##        ID           LIMIT_BAL            SEX          EDUCATION
##  Min.   :    1   Min.   :  10000   Min.   :1.000   Min.   :0.000
##  1st Qu.: 7501   1st Qu.:  50000   1st Qu.:1.000   1st Qu.:1.000
##  Median :15000   Median : 140000   Median :2.000   Median :2.000
##  Mean   :15000   Mean   : 167484   Mean   :1.604   Mean   :1.853
##  3rd Qu.:22500   3rd Qu.: 240000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :30000   Max.   :1000000   Max.   :2.000   Max.   :6.000
##     MARRIAGE          AGE            PAY_SEP           PAY_AUG
##  Min.   :0.000   Min.   :21.00   Min.   :-2.0000   Min.   :-2.0000
##  1st Qu.:1.000   1st Qu.:28.00   1st Qu.:-1.0000   1st Qu.:-1.0000
##  Median :2.000   Median :34.00   Median : 0.0000   Median : 0.0000
##  Mean   :1.552   Mean   :35.49   Mean   :-0.0167   Mean   :-0.1338
##  3rd Qu.:2.000   3rd Qu.:41.00   3rd Qu.: 0.0000   3rd Qu.: 0.0000
##  Max.   :3.000   Max.   :79.00   Max.   : 8.0000   Max.   : 8.0000
##     PAY_JUL           PAY_JUN           PAY_MAY           PAY_APR
##  Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000
##  1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000
##  Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000
##  Mean   :-0.1662   Mean   :-0.2207   Mean   :-0.2662   Mean   :-0.2911
##  3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000
##  Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000
##   BILL_AMT_SEP      BILL_AMT_AUG      BILL_AMT_JUL      BILL_AMT_JUN
##  Min.   :-165580   Min.   :-69777   Min.   :-157264   Min.   :-170000
##  1st Qu.:   3559   1st Qu.:  2985   1st Qu.:   2666   1st Qu.:   2327
##  Median :  22382   Median : 21200   Median :  20089   Median :  19052
##  Mean   :  51223   Mean   : 49179   Mean   :  47013   Mean   :  43263
##  3rd Qu.:  67091   3rd Qu.: 64006   3rd Qu.:  60165   3rd Qu.:  54506
##  Max.   : 964511   Max.   :983931   Max.   :1664089   Max.   : 891586
##   BILL_AMT_MAY      BILL_AMT_APR      PAY_AMT_SEP       PAY_AMT_AUG
##  Min.   :-81334   Min.   :-339603   Min.   :     0   Min.   :      0
##  1st Qu.:  1763   1st Qu.:   1256   1st Qu.:  1000   1st Qu.:    833
##  Median : 18105   Median :  17071   Median :  2100   Median :   2009
##  Mean   : 40311   Mean   :  38872   Mean   :  5664   Mean   :   5921
##  3rd Qu.: 50191   3rd Qu.:  49198   3rd Qu.:  5006   3rd Qu.:   5000
##  Max.   :927171   Max.   : 961664   Max.   :873552   Max.   :1684259
##   PAY_AMT_JUL       PAY_AMT_JUN       PAY_AMT_MAY        PAY_AMT_APR
##  Min.   :     0   Min.   :     0   Min.   :     0.0   Min.   :     0.0
##  1st Qu.:   390   1st Qu.:   296   1st Qu.:   252.5   1st Qu.:   117.8
##  Median :  1800   Median :  1500   Median :  1500.0   Median :  1500.0
##  Mean   :  5226   Mean   :  4826   Mean   :  4799.4   Mean   :  5215.5
##  3rd Qu.:  4505   3rd Qu.:  4013   3rd Qu.:  4031.5   3rd Qu.:  4000.0
##  Max.   :896040   Max.   :621000   Max.   :426529.0   Max.   :528666.0
##  default payment next month
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2212
```

```
##  3rd Qu.:0.0000
##  Max.   :1.0000
```

```
sapply(3:5, function(x) table(df[,x])) # Sex, Education, Marriage
```

```
## [[1]]
##
##     1     2
## 11888 18112
##
## [[2]]
##
##     0     1     2     3     4     5     6
##    14 10585 14030  4917   123   280    51
##
## [[3]]
##
##     0     1     2     3
##    54 13659 15964   323
```

Step 4 : Initial Thoughts

- Why are there 0, 5 and 6s for Education if supposed to be from 1 to 4 –> put them in 4 = others ?
- Why is there 0 for Marital status if supposed to be from 1 to 3 –> put it in 3 = others ?
- Make column to count the amount of times the client defaulted (evaluate how good/bad of a credit the client has)

Step 5 : Wrangling

For the `Education` variable, the data description does not mention what 0, 5 and 6 represent. For our analysis, we chose to classify these values as others and so attributed them the value 4. We also thought it would be interesting to add a column that represented the number of months out of the 6 included in the data where the client defaulted. To do this, we calculated the number of months out of the 6 where a payment was delayed (values from 1 to 9 in columns PAY_SEP through PAY_APR). Another column called `REVOLVING_CREDIT` was added. It counts the number of months where a client was offered revolving credit. Finally, we decided to discretize the age variable. We created the following age intervals : [20, 30), [30, 40), [40, 50), [50, 60), [60, 70), [70, 80) and [80, 90).

```
# Move everything not in HighSchool, Undergrad, or GradSchool to Other category
df[df$EDUCATION %in% c(0, 5, 6), "EDUCATION"] <- 4

# Count number of times defaulted on monthly paymnt, and number of times offered
# revolving credit within the 6-month observation period
for (i in 1:nrow(df)){
  default_count <- 0
  revolving_credit_count <- 0
  for (j in 7:12){
    if (df[i, j] > 0) {
      default_count <- default_count + 1
    } else if (df[i, j] == 0) {
      revolving_credit_count <- revolving_credit_count + 1
    }
  }
  df[i, 'QTY_DEFAULT'] <- default_count
```

```
  df[i, 'REVOLVING_CREDIT'] <- revolving_credit_count
}

# Discretize age by making bins (20s, 30s, etc.)
df$AGE_BINNED <- cut(df$AGE, breaks = c(20, 30, 40, 50, 60, 70, 80, 90), include.lowest = TRUE)
```

Step 6 : Research Questions

1- Do people with a higher education (graduate school, university), on average, default less frequently than people with a lower level of education (high school)?

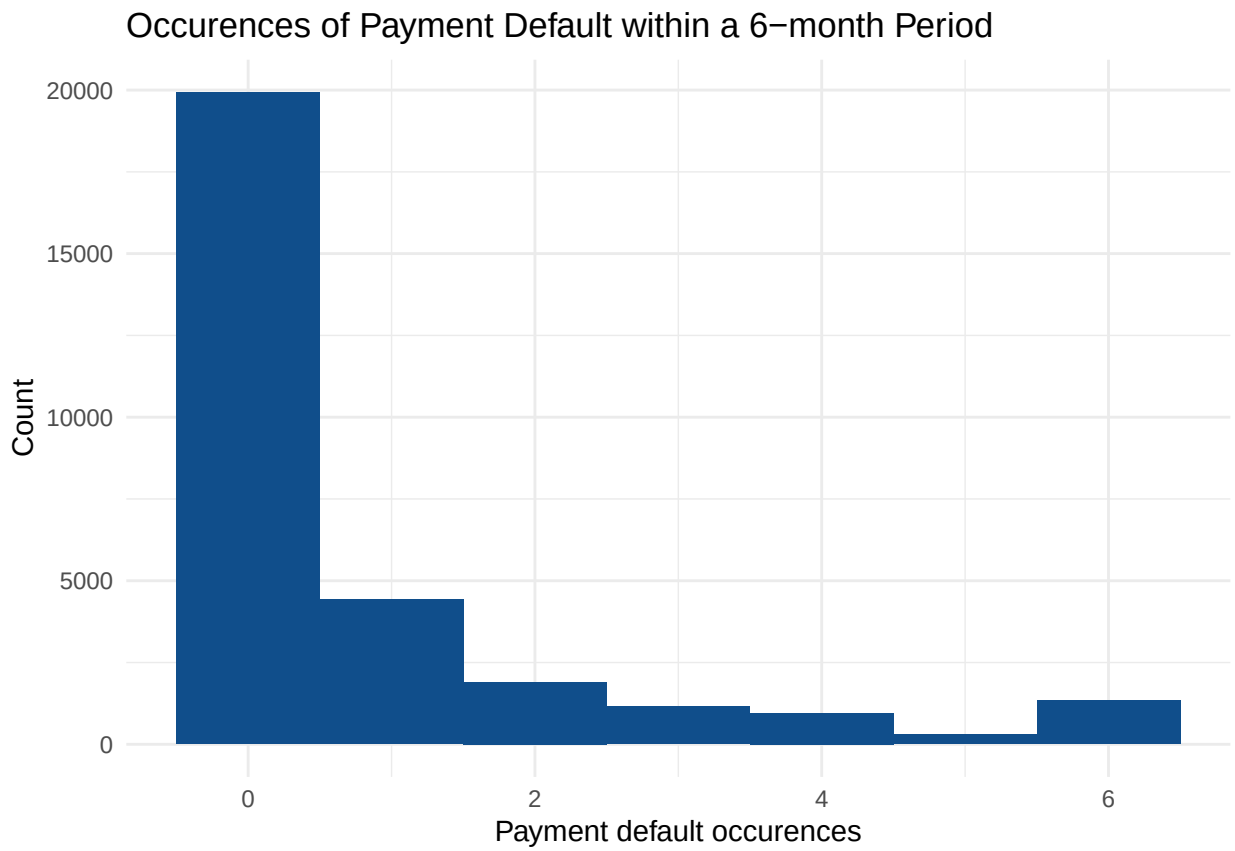
2- Are the people who are given higher amounts of credit, also people who, on average, default less frequently?

3- What is the frequency of default payments in a 6 month period for the data collected?

4- Are the people who have defaulted more frequently in the past 6 months more likely to default in October (the following month: `default payment next month`)?

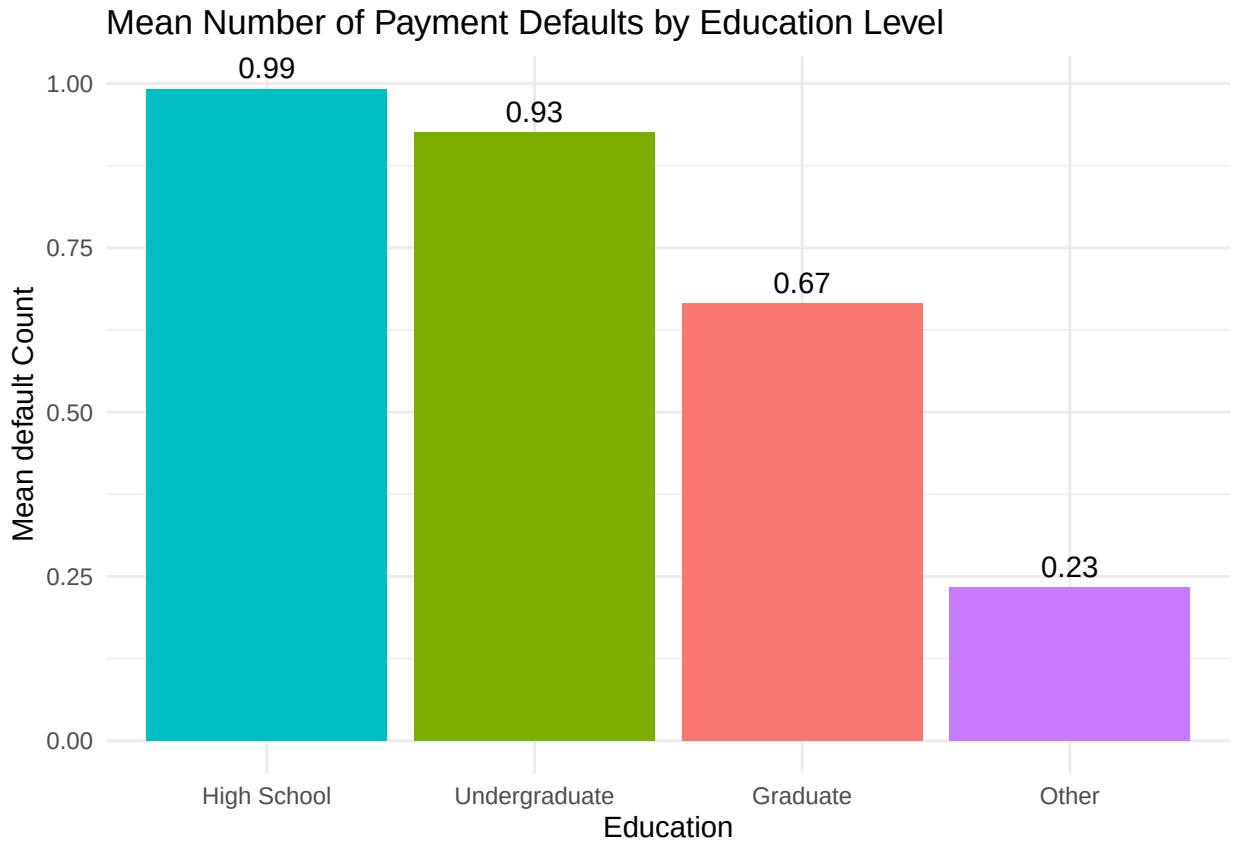
Step 7 : Data Analysis & Visualizations

```
ggplot(df) +
  aes(x = `QTY_DEFAULT`) +
  geom_histogram(binwidth = 1, fill = "dodgerblue4") +
  ggtitle("Occurences of Payment Default within a 6-month Period") +
  labs(x = "Payment default occurences", y = "Count") +
  theme_minimal()
```



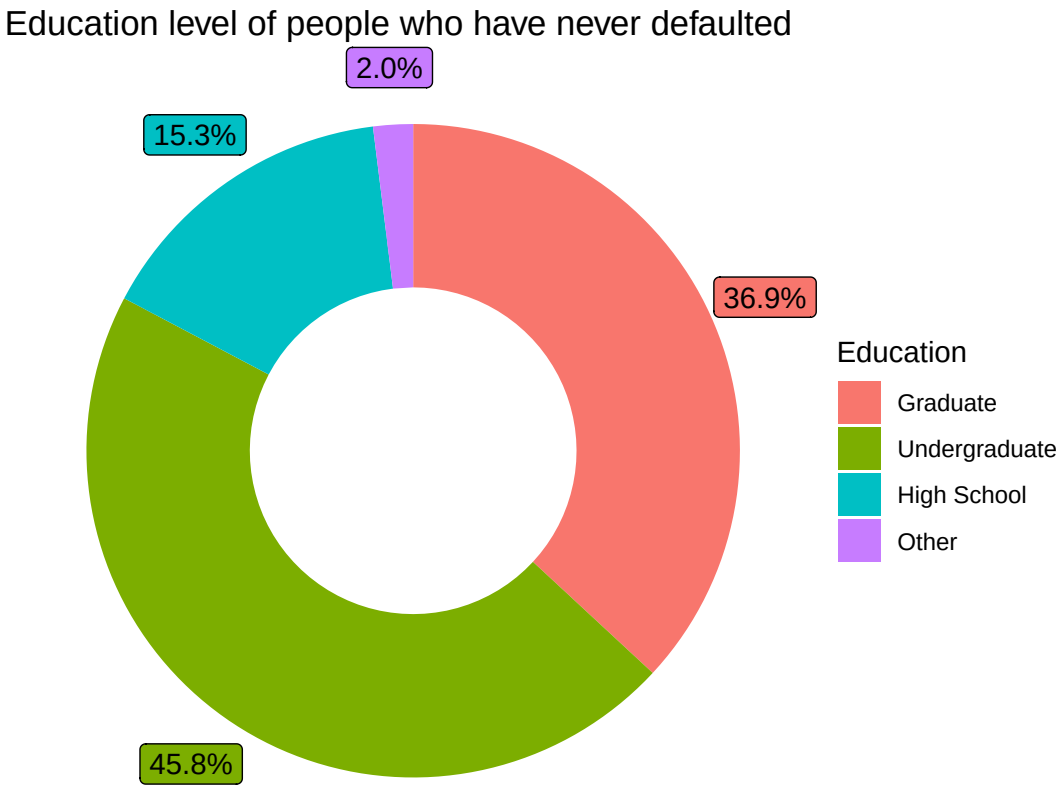
For this visualization, we opted for a simple histogram because all we wanted to accomplish here was to show the distribution of the number of missed payments within the 6-month observation period. As we can see, the majority of people in our dataset had no payments in default within this period.

```r
default_by_education  <- df %>% group_by(`EDUCATION`) %>% summarise(mean_default = mean(`QTY_DEFAULT`))
default_by_education$EDUCATION <- factor(
  default_by_education$EDUCATION,
  levels = c(1, 2, 3, 4),
  labels = c("Graduate", "Undergraduate", "High School", "Other")
)

ggplot(default_by_education) +
  aes(x = reorder(`EDUCATION`, -`mean_default`), y = `mean_default`, fill = `EDUCATION`) +
  geom_bar(show.legend = FALSE, stat="identity", position="dodge") +
  geom_text(aes(label=sprintf("%.2f", `mean_default`)), position=position_dodge(width = 0.2), vjust=-0.5
  ggtitle("Mean Number of Payment Defaults by Education Level") +
  labs(x = "Education", y = "Mean default Count") +
  theme_minimal()
```



For this visualization, we chose to use a bar chart so we could easily see which `Education` group had the highest average number of defaults. It also makes it easy to compare the different groups to each other. As we can see, on average, people with a lower level of education tend to have the highest number of payment defaults.

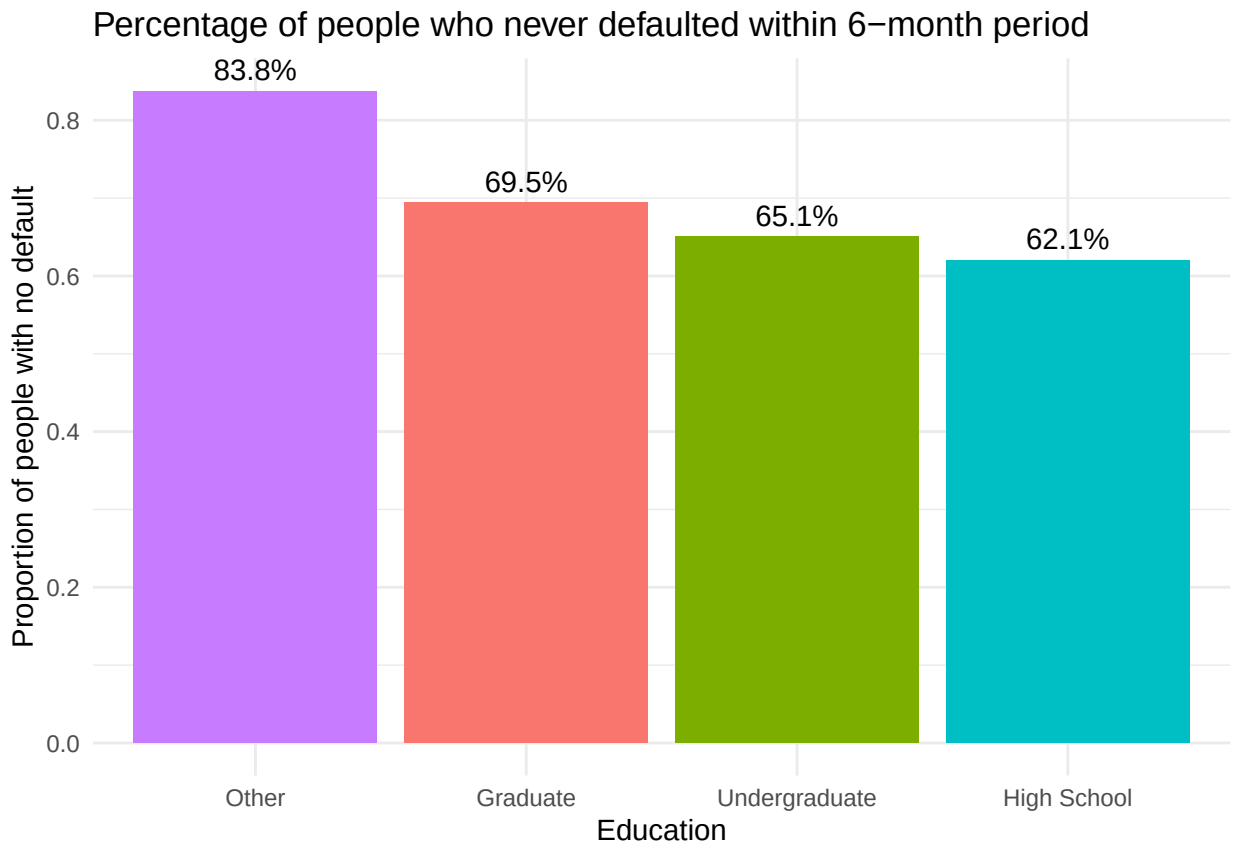
```
zero_count <- count(df, QTY_DEFAULT)[[1,2]]
zero_default_education <- df %>% filter(QTY_DEFAULT == 0) %>% count(EDUCATION)
zero_default_education$EDUCATION <- factor(
  zero_default_education$EDUCATION,
  levels = c(1, 2, 3, 4),
  labels = c("Graduate", "Undergraduate", "High School", "Other")
)
zero_default_education$fraction <- zero_default_education$n / zero_count
zero_default_education$ymax <- cumsum(zero_default_education$fraction)
zero_default_education$ymin <- c(0, head(zero_default_education$ymax, n=-1))
zero_default_education$labelPosition <- (zero_default_education$ymax + zero_default_education$ymin) / 2

ggplot(zero_default_education) +
  aes(ymax=ymax, ymin=ymin, xmax=5, xmin=4, fill=EDUCATION) +
  geom_rect() +
  geom_label(
    x = 5.35,
    aes(y=labelPosition, label=sprintf("%.1f%%", fraction * 100)),
    show.legend = FALSE
  ) +
  coord_polar(theta = "y") +
  xlim(c(3,5)) +
  ggtitle("Education level of people who have never defaulted") +
  labs(fill = "Education") +
  theme_void()
```



```
  # theme(legend.position = "none")
```

For this visualization, we opted for a donut chart because we wanted to visualize the proportional breakdown by education of people who've never defaulted, and a donut chart is considered a good choice for displaying proportions. This graph illustrates that out of all the clients in our data who never defaulted, people with an undergraduate degree form the biggest category followed by individuals with a graduate and high school education.

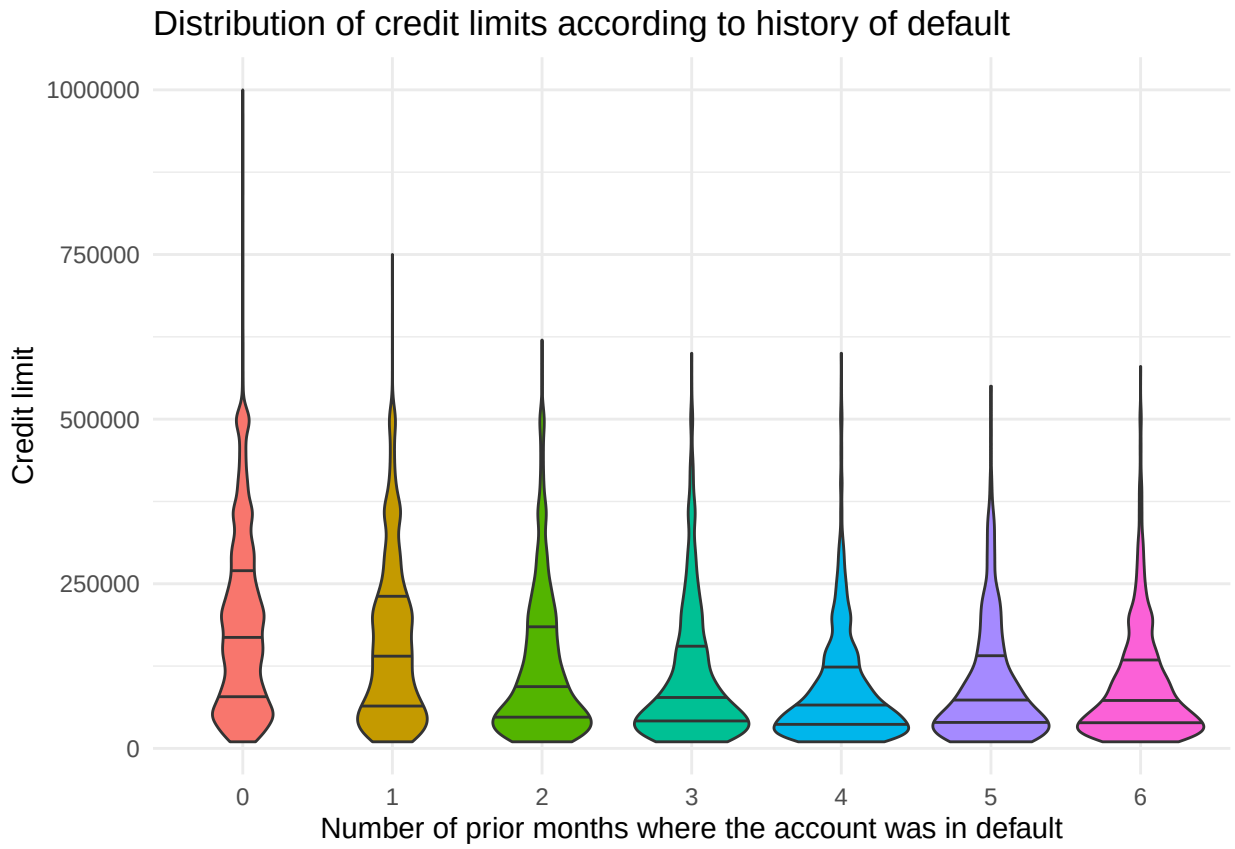
```r
no_default  <- df[df$QTY_DEFAULT == 0,]

no_default_education <- no_default %>% group_by(`EDUCATION`) %>% summarise(no_default = NROW(`EDUCATION`

no_default_education <- cbind(no_default_education, table(df$EDUCATION))[-3]

no_default_education['mean'] <- no_default_education$no_default/no_default_education$Freq

no_default_education$EDUCATION <- factor(
  no_default_education$EDUCATION,
  levels = c(1, 2, 3, 4),
  labels = c("Graduate", "Undergraduate", "High School", "Other")
)

ggplot(no_default_education) +
  aes(x = reorder(`EDUCATION`, -`mean`), y = `mean`, fill = `EDUCATION`) +
  geom_bar(show.legend = FALSE, stat="identity", position="dodge") +
  geom_text(aes(label=sprintf("%.1f%%", `mean`*100)), position=position_dodge(width = 0.2), vjust=-0.5)
  ggtitle("Percentage of people who never defaulted within 6-month period") +
  labs(x = "Education", y = "Proportion of people with no default") +
  theme_minimal()
```



This bar chart illustrates the percentage of people, by education level, that never defaulted within the 6-month period. In the pie chart, we had a higher percentage of people with undergraduate degrees that never defaulted compared to people with graduate degrees.This could potentially be explained by a higher number of people with graduate degrees in our data. In the above bar chart, the size of the groups (graduate,

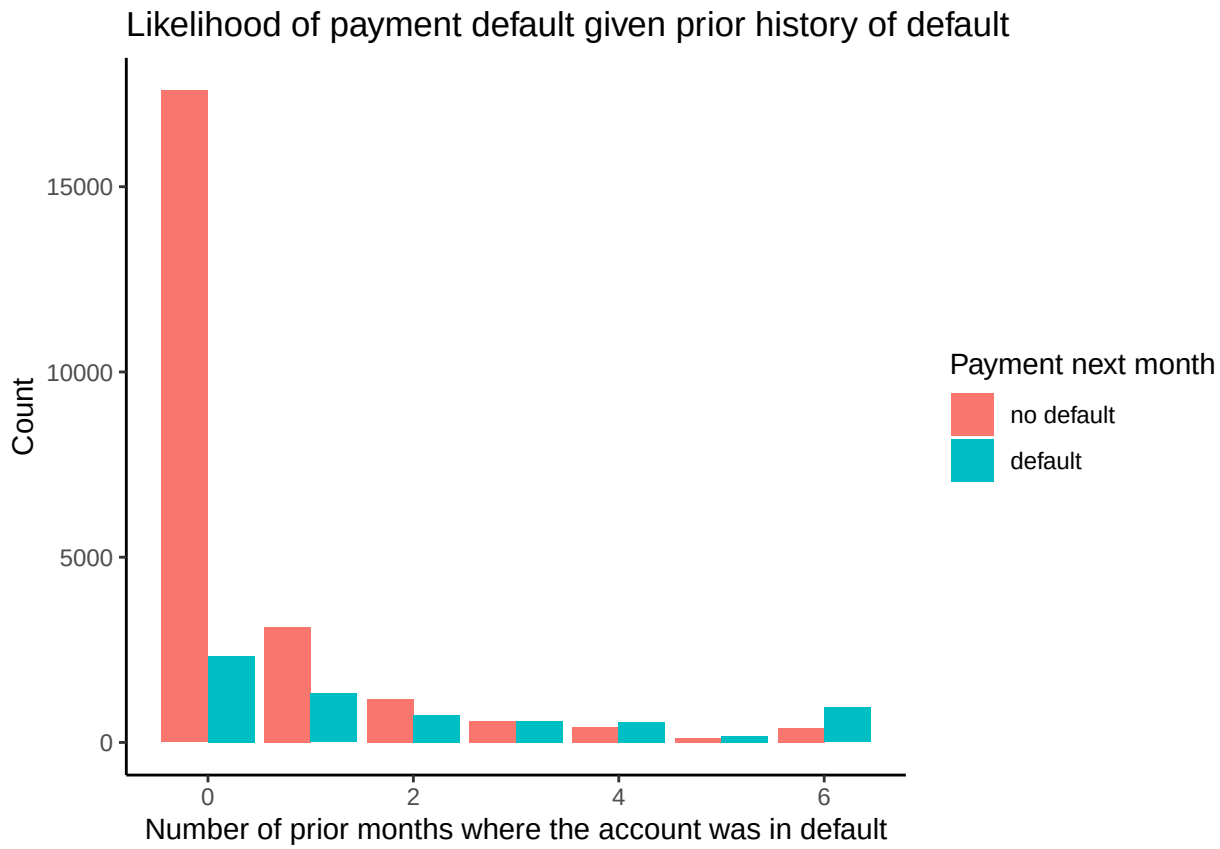
undergraduate etc) are taken into consideration when calculating the percentages. Thus, this graph illustrates that individuals with a graduate degree tend to have better credit habits than those with a high school education. In fact 69.5% of people who have a graduate degree never defaulted within the 6-month period versus 62.1% for people with a high school education.

```
ggplot(df, aes(x = as.factor(QTY_DEFAULT), y = LIMIT_BAL, fill = as.factor(QTY_DEFAULT))) + geom_violin
```



Violinplots were chosen in order to illustrate the relationship between the number of months in default and the amount of credit granted to an individual. Our hypothesis was that higher limits of credit were given to those with better credit habits, thus those with a lower number of months in default. Violinplots were favored to boxplots as their area vary according to the number of observations. Looking at the 3rd quartile of each violin, we notice that as the number of months where the account was in default increase from 0 to 4, the amount of credit that was given decreases. This validates our hypothesis.

```r
df$default <- factor(ifelse(df$`default payment next month` == 0, "no default", "default"))
df$default <- relevel(df$default, "no default")

ggplot(df, aes(x = QTY_DEFAULT, fill = default)) + geom_bar(position = 'dodge') + theme_classic() + labs
```



This bar plot was made in order to explore the following research question: are people who have defaulted more frequently in the past 6 months more likely to default in the following month? We notice that people that never defaulted are a lot more likely to continue paying on time. Furthermore, as the number of months where the account was in default increases, the difference in count between the 2 groups (default vs no default next month) decreases. In fact, for people who defaulted 3 months out of 6, we notice that the probability that they default during the 7-month is approximately 50%.

Research Questions

1. Is there a relationship between the income of an individual, and the number of times within the 6-month observation period that they default? (Answering this question would require a dataset that includes all the columns of the current dataset, plus columns for net monthly income)
2. Is there a relationship between the income of an individual, and the credit limit? (Answering this question would require a dataset that includes all the columns of the current dataset, plus columns for net monthly income)

Step 8 : Summary and Conclusions

We've found that a sizable majority of the people surveyed in this dataset didn't default even a single time within the 6-month observation period in which the data was collected. We also noticed that people with higher levels of education were, on average, more likely to never default within this 6-month period. Similarly, we calculated that people with higher levels of education defaulted, on average, a smaller number of times over the course of the 6-month period. Also, we were able to conclude that higher limits of credit were granted to those who generally defaulted less frequently.

Finally, we looked at the relationship between the number of months where the account was in default for a 6-month period and the probability of default for the subsequent month (month 7). It was shown that people who defaulted less than 3 times within the 6-month period were more likely to pay on time for the 7th month. On the other hand, people who defaulted more than 2 times had a greater chance of defaulting the following month.