Categorical Variable Markdown

Everett

2024-07-03

R Markdown Categorical Variables

Import dataset and required libraries, prepare variables for analysis

```
setwd("C:/Users/escra/OneDrive/Documents/Job Stuff/DA Project Logistic Regression/Dataset")
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
#Read the dataset and view it
df <- read_csv("dataset.csv")
```

```
## Rows: 9709 Columns: 20
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (5): Income_type, Education_type, Family_status, Housing_type, Occupati...
## dbl (15): ID, Gender, Own_car, Own_property, Work_phone, Phone, Email, Unemp...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(df)
```

```
## # A tibble: 6 x 20
##        ID Gender Own_car Own_property Work_phone Phone Email Unemployed
##     <dbl>  <dbl>   <dbl>        <dbl>      <dbl> <dbl> <dbl>      <dbl>
## 1 5008804      1       1            1          1     0     0          0
## 2 5008806      1       1            1          0     0     0          0
## 3 5008808      0       0            1          0     1     1          0
## 4 5008812      0       0            1          0     0     0          1
```

```
## 5 5008815      1       1            1          1     1     1          0
## 6 5008819      1       1            1          0     0     0          0
## # i 12 more variables: Num_children <dbl>, Num_family <dbl>,
## #   Account_length <dbl>, Total_income <dbl>, Age <dbl>, Years_employed <dbl>,
## #   Income_type <chr>, Education_type <chr>, Family_status <chr>,
## #   Housing_type <chr>, Occupation_type <chr>, Target <dbl>
```

```
#Define the categorical variables within a dataset
df_qual <- df[, c("Gender", "Own_car", "Own_property", "Work_phone", "Phone", "Email", "Unemployed", "In
```

Variable Analysis for Gender Variable

Binary feature that represents the gender of the customer. 0 represents males and 1 represents females. In the US, gender cannot be used when determining whether to extend credit to an applicant (CFPB). This will be an important consideration when creating the machine learning model.

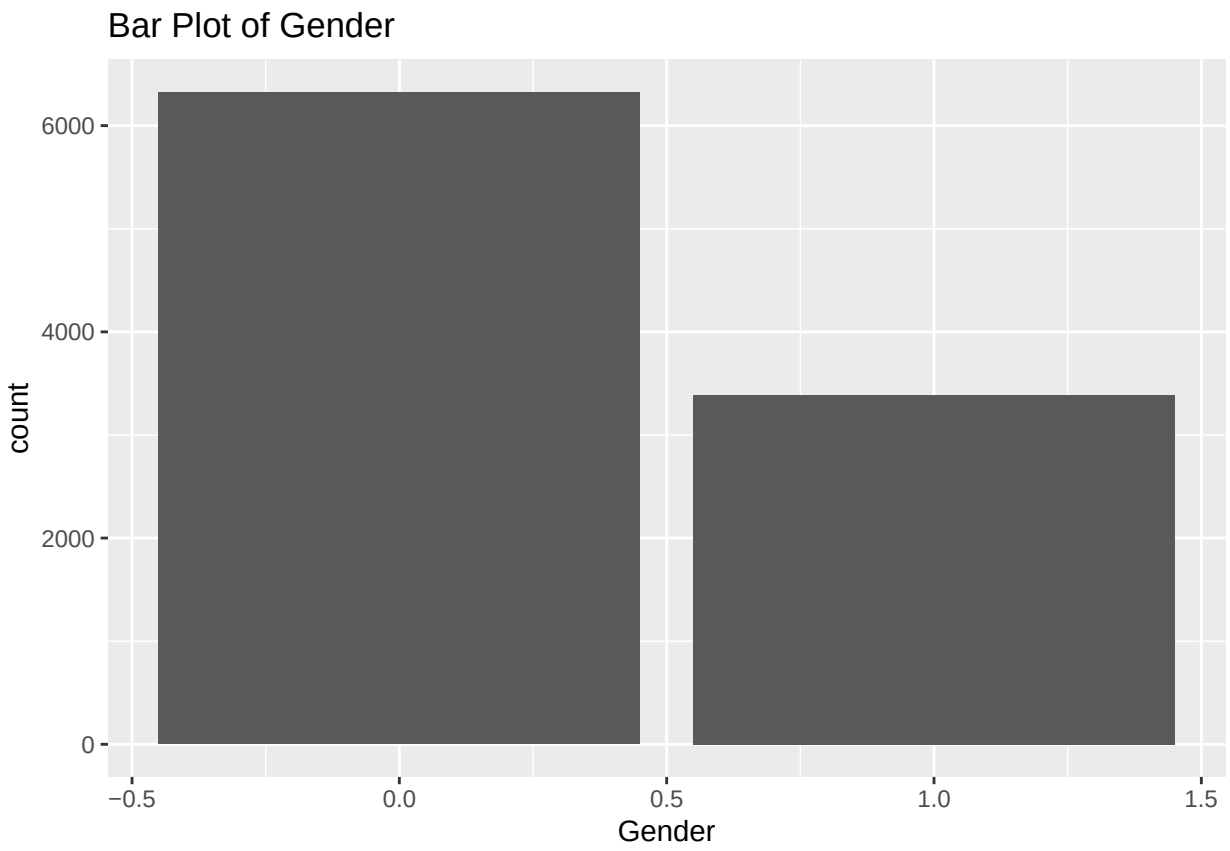
```
table(df_qual$Gender)
```

```
##
##    0    1
## 6323 3386
```

```
prop.table(table(df_qual$Gender))
```

```
##
##         0         1
## 0.6512514 0.3487486
```

```
ggplot(df_qual, aes(x=Gender)) +
  geom_bar() +
  ggtitle("Bar Plot of Gender")
```



Variable Analysis for Own_car Variable

Binary feature that indicates if a customer owns a car. 0 represents no car ownership and 1 represents there is car ownership. In this context, car ownership can show financial stability.

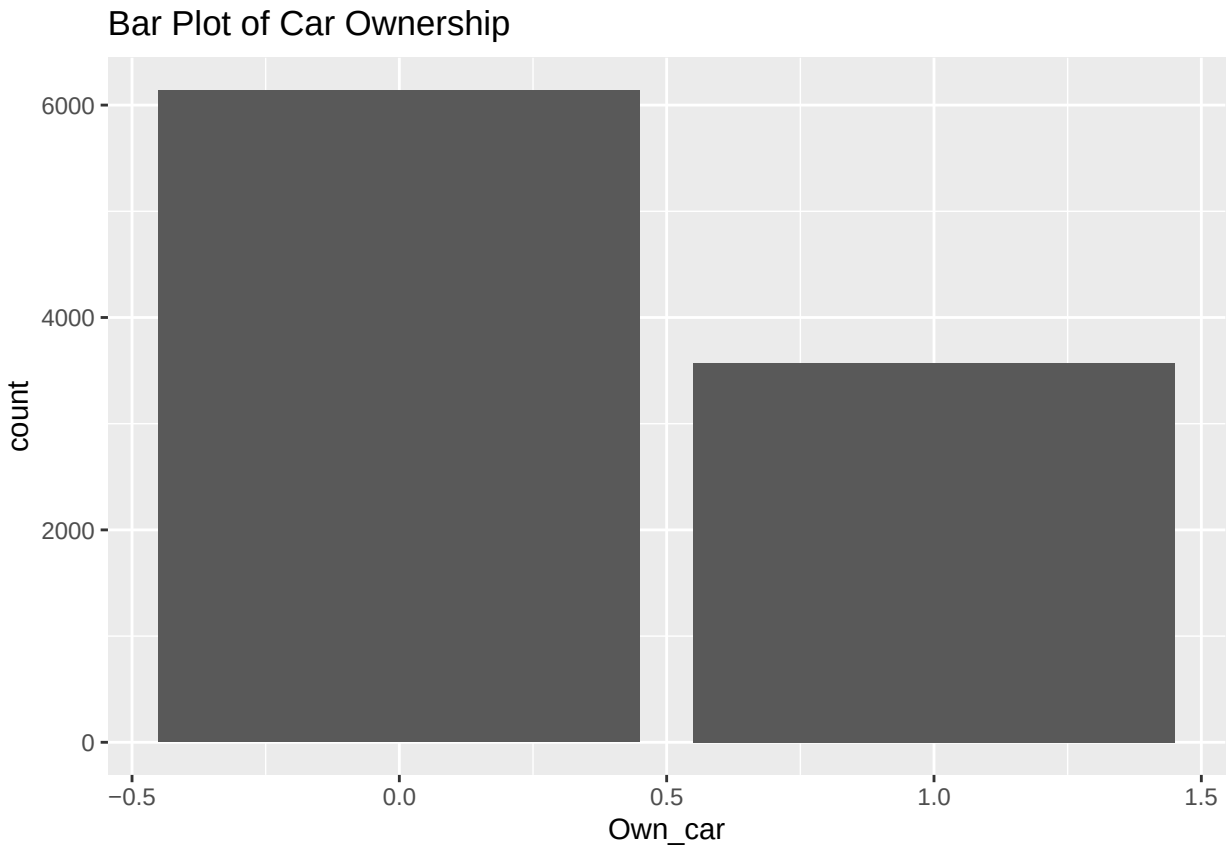
```
table(df_qual$Own_car)
```

```
##
##    0    1
## 6139 3570
```

```
prop.table(table(df_qual$Own_car))
```

```
##
##         0         1
## 0.6322999 0.3677001
```

```
ggplot(df_qual, aes(x=Own_car)) +
  geom_bar() +
  ggtitle("Bar Plot of Car Ownership")
```



Variable Analysis for the Own_property Variable

Binary feature that indicates if the customer owns property. 0 represents no property ownership and 1 shows the customer owns property. Property ownership shows financial stability and a solid credit score.

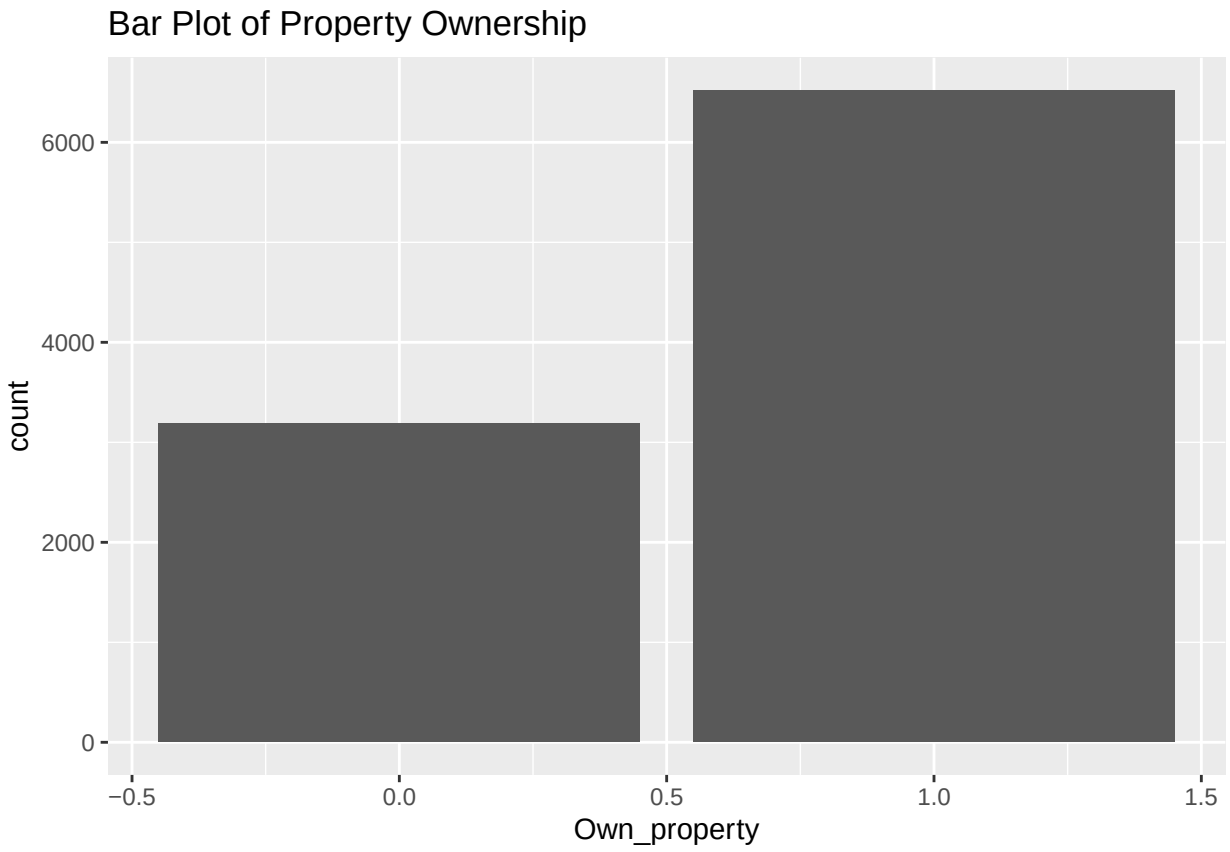
```
table(df_qual$Own_property)
```

```
##
##    0    1
## 3189 6520
```

```
prop.table(table(df_qual$Own_property))
```

```
##
##         0         1
## 0.3284581 0.6715419
```

```
ggplot(df_qual, aes(x=Own_property)) +
  geom_bar() +
  ggtitle("Bar Plot of Property Ownership")
```



Variable Analysis for the Work_phone Variable

Binary feature that indicates if a customer owns a work phone. 0 indicates no and 1 indicates yes. Owning a work phone can be a sign of employment stability.

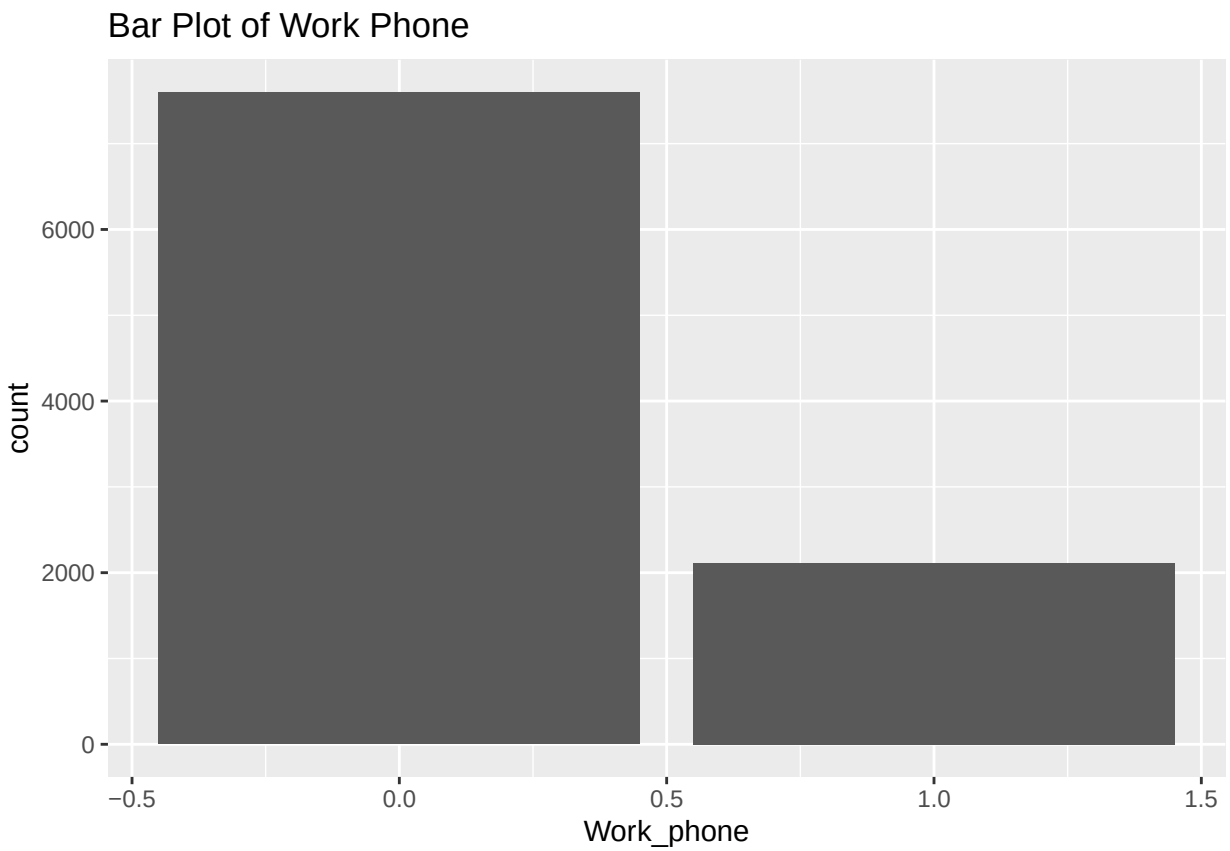
```
table(df_qual$Work_phone)
```

```
##
##    0    1
## 7598 2111
```

```
prop.table(table(df_qual$Work_phone))
```

```
##
##         0         1
## 0.7825729 0.2174271
```

```
ggplot(df_qual, aes(x=Work_phone)) +
  geom_bar() +
  ggtitle("Bar Plot of Work Phone")
```



Variable Analysis for the Phone Variable

Binary feature that indicates if a customer has a phone. 0 indicates they do not have a phone, while 1 indicates they own a phone. Owning a phone is another indicator of financial stability.

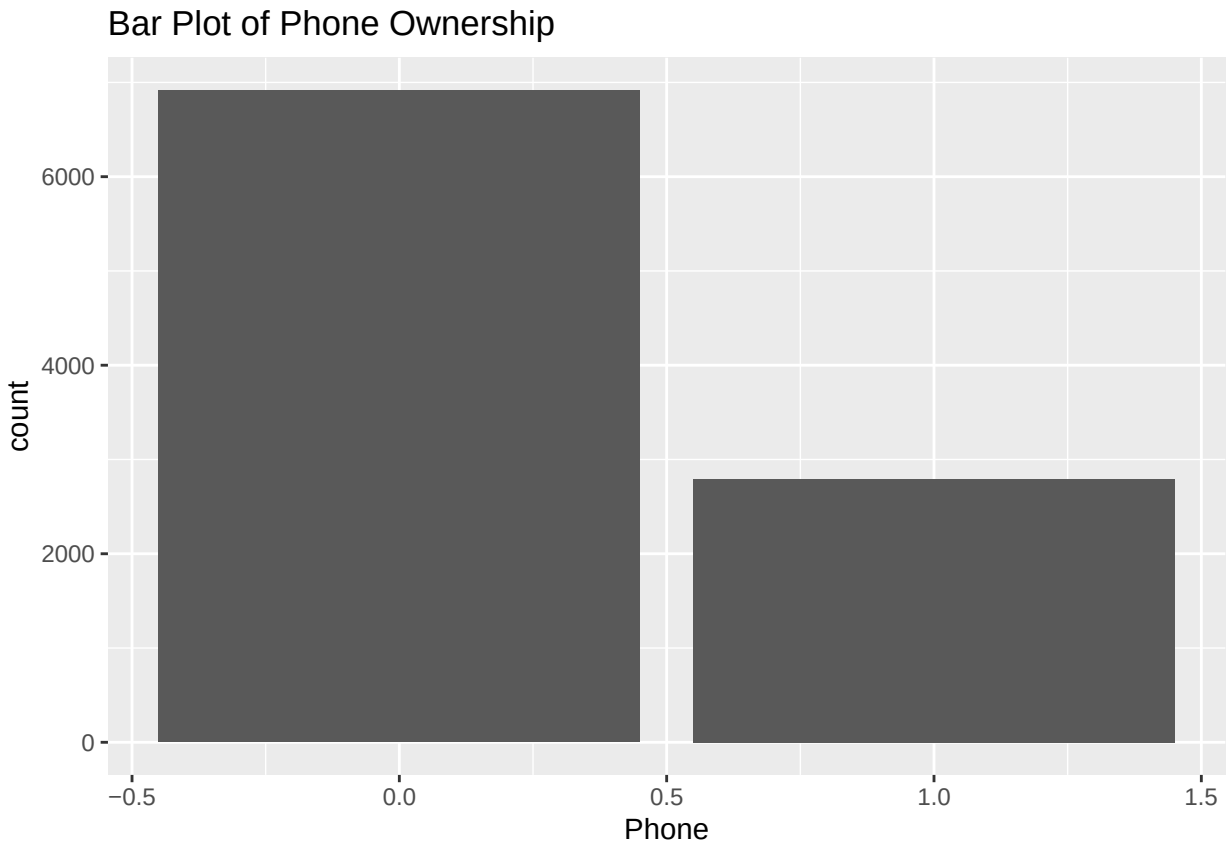
```
table(df_qual$Phone)
```

```
##
##    0    1
## 6916 2793
```

```
prop.table(table(df_qual$Phone))
```

```
##
##         0         1
## 0.7123288 0.2876712
```

```
ggplot(df_qual, aes(x=Phone)) +
  geom_bar() +
  ggtitle("Bar Plot of Phone Ownership")
```



Variable Analysis for the Email Variable

Binary feature that indicates whether an applicant provided an email on their application. 0 indicates no email while 1 indicates a provided email.

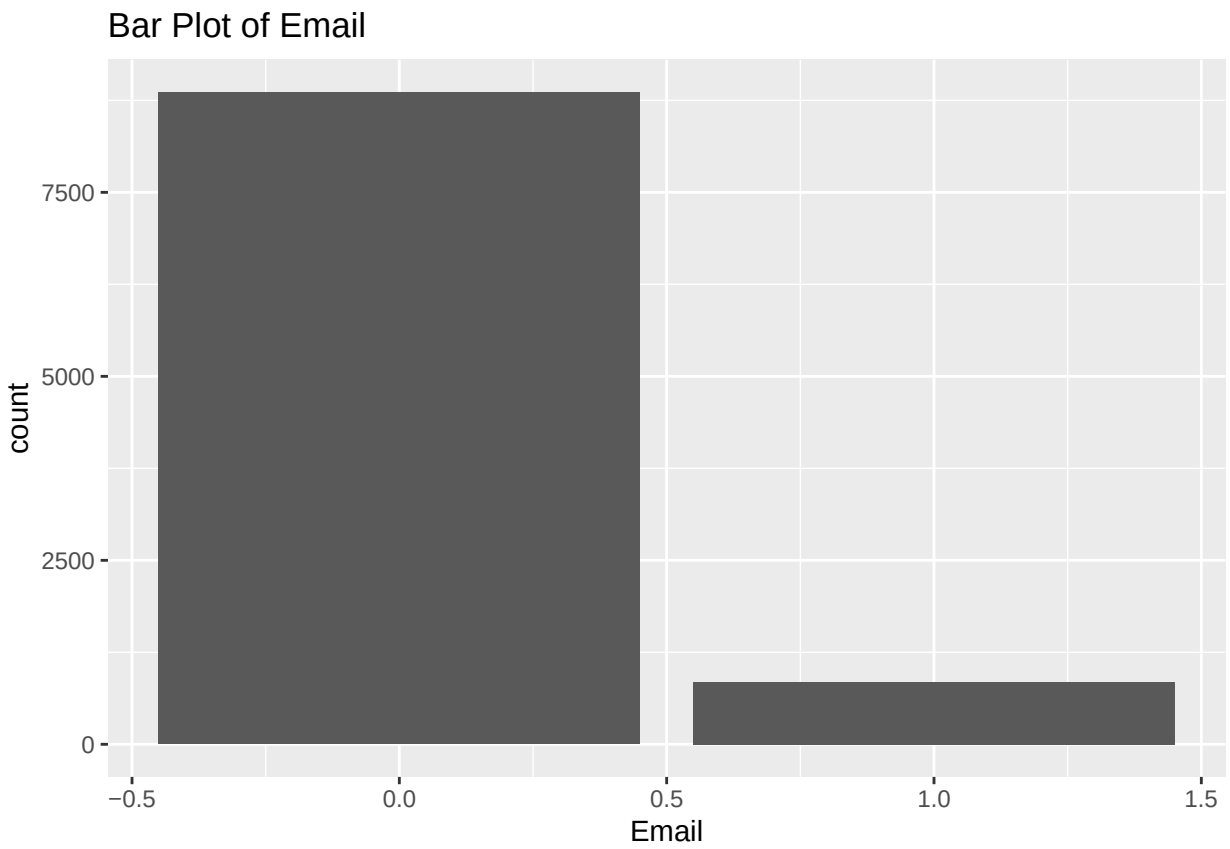
```
table(df_qual$Email)
```

```
##
##    0    1
## 8859  850
```

```
prop.table(table(df_qual$Email))
```

```
##
##          0          1
## 0.91245236 0.08754764
```

```
ggplot(df_qual, aes(x=Email)) +
  geom_bar() +
  ggtitle("Bar Plot of Email")
```



Variable Analysis for the Unemployed Variable

A binary feature that indicates if a customer is unemployed. 0 means the customer is employed and 1 indicates a person is unemployed. Employment is important to prove financial stability and a constant income.

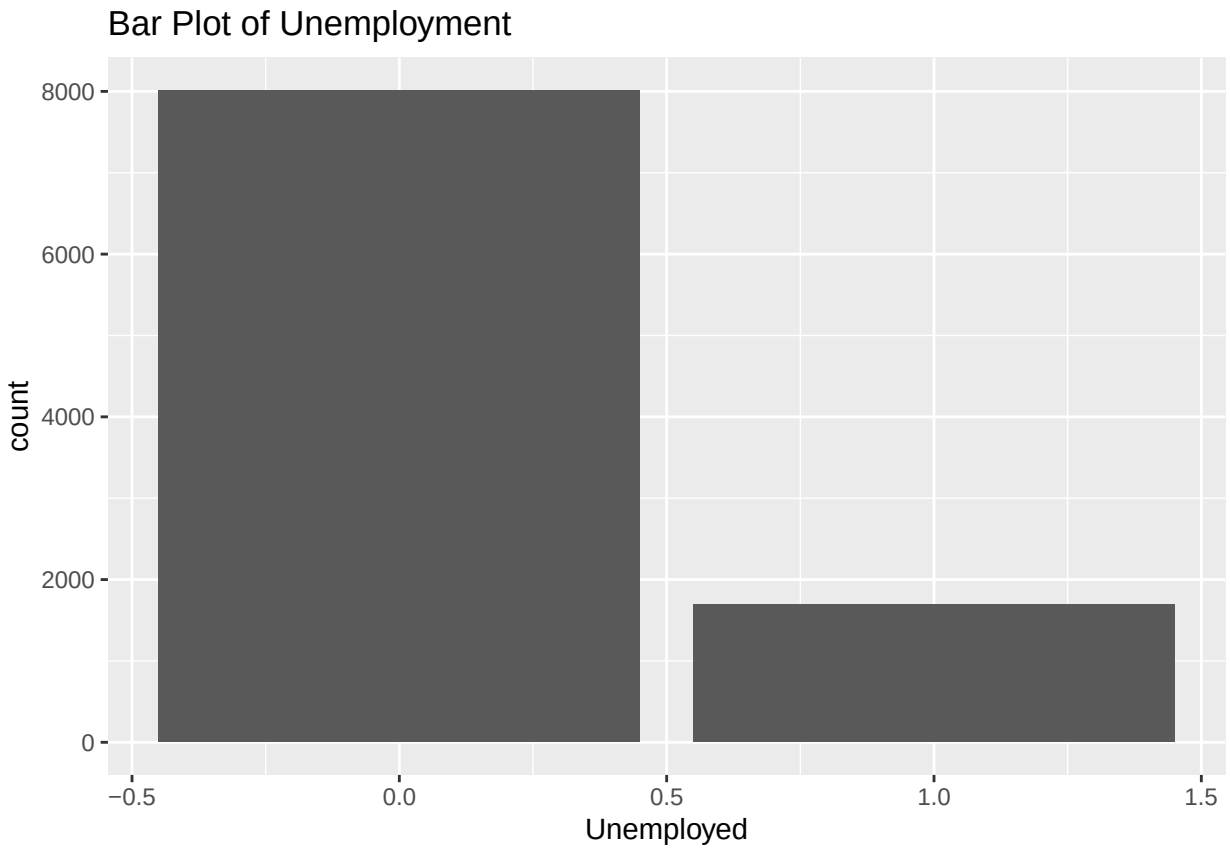
```
table(df_qual$Unemployed)
```

```
##
##    0    1
## 8013 1696
```

```
prop.table(table(df_qual$Unemployed))
```

```
##
##         0         1
## 0.8253167 0.1746833
```

```
ggplot(df_qual, aes(x=Unemployed)) +
  geom_bar() +
  ggtitle("Bar Plot of Unemployment")
```



Variable Analysis for the Income_type Variable

A categorical variable that indicates the type of income for a customer (e.g., Working, Commercial Associate, Pensioner, etc.). Different income types indicate different amounts of financial stability.

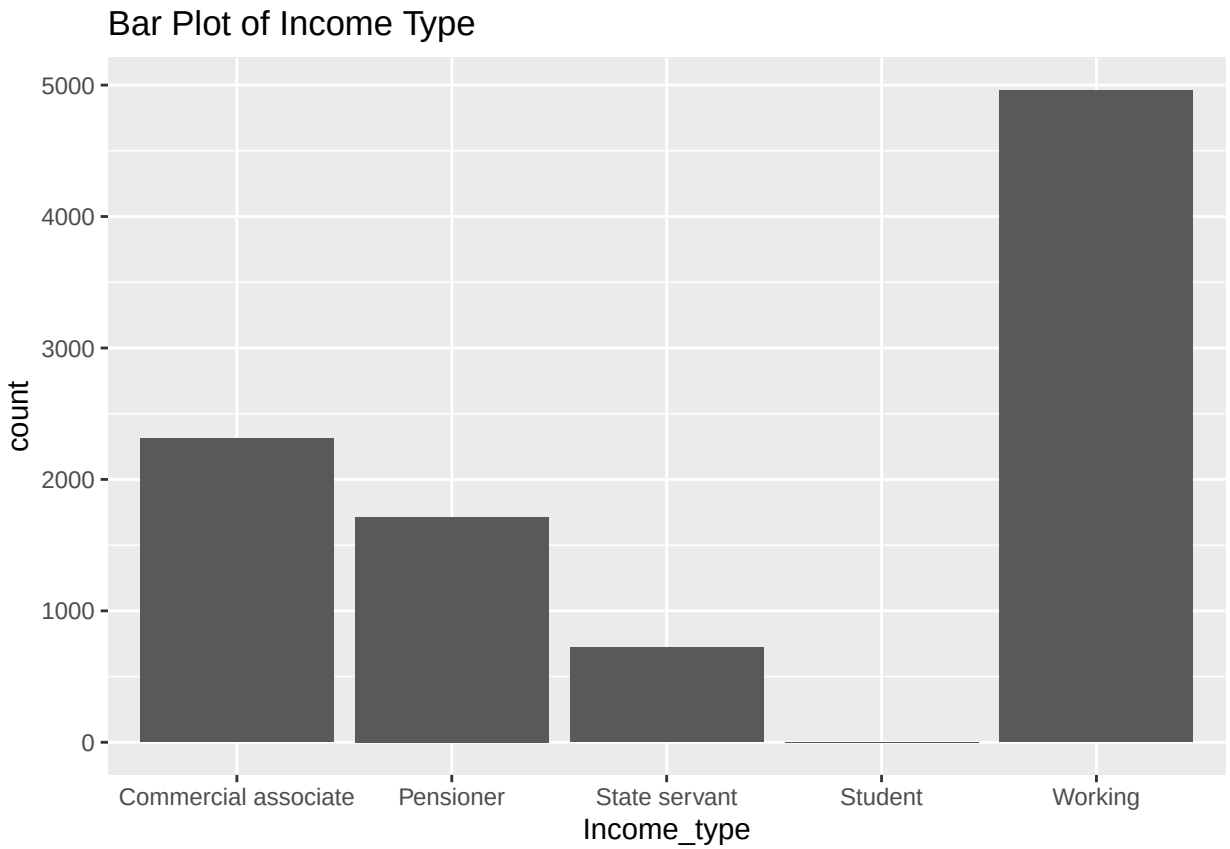
```
table(df_qual$Income_type)
```

```
##
## Commercial associate           Pensioner       State servant
##                 2312                1712                 722
##              Student             Working
##                    3                4960
```

```
prop.table(table(df_qual$Income_type))
```

```
##
## Commercial associate           Pensioner       State servant
##         0.2381295705        0.1763312391        0.0743639922
##              Student             Working
##         0.0003089917        0.5108662066
```

```
ggplot(df_qual, aes(x=Income_type)) +
  geom_bar() +
  ggtitle("Bar Plot of Income Type")
```



Variable Analysis for the Education_type Variable

A categorical variable that shows the education level of a customer (e.g., Secondary/secondary special, Higher education, etc.). Higher levels of education can demonstrate greater reliability or potentially student loan debt.

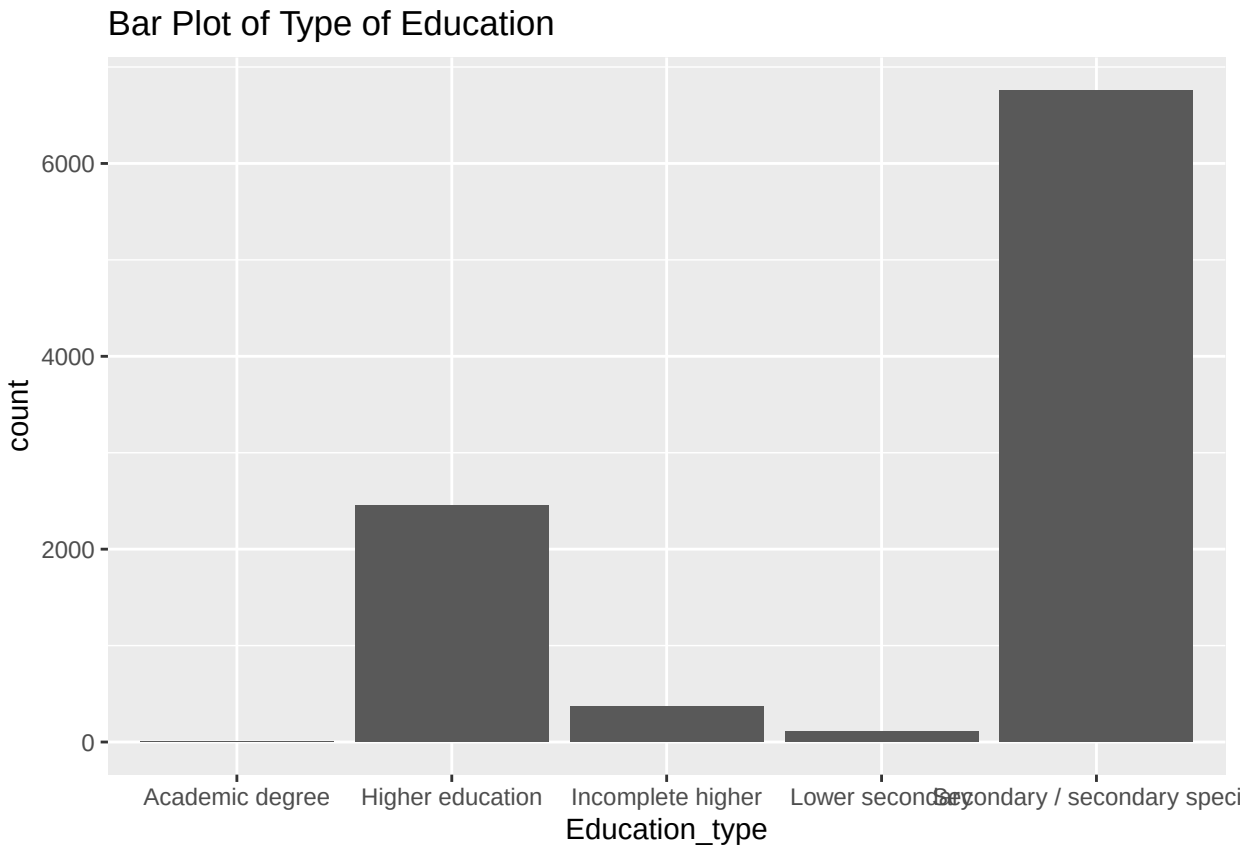
```
table(df_qual$Education_type)
```

```
##
##               Academic degree             Higher education
##                             6                         2457
##             Incomplete higher              Lower secondary
##                           371                          114
## Secondary / secondary special
##                          6761
```

```
prop.table(table(df_qual$Education_type))
```

```
##
##               Academic degree             Higher education
##                  0.0006179833                 0.2530641673
##             Incomplete higher              Lower secondary
##                  0.0382119683                 0.0117416830
## Secondary / secondary special
##                  0.6963641982
```

```
ggplot(df_qual, aes(x=Education_type)) +
  geom_bar() +
  ggtitle("Bar Plot of Type of Education")
```



Variable Analysis for the Family_status Variable

A categorical variable that shows the family status of an individual. In the US, it is important to understand that marital status cannot be used to determine whether to extend credit. Marital status can be considered in certain cases such as relying on a spouse for income (CFPB). These are important considerations to avoid discrimination.

```
table(df_qual$Family_status)
```

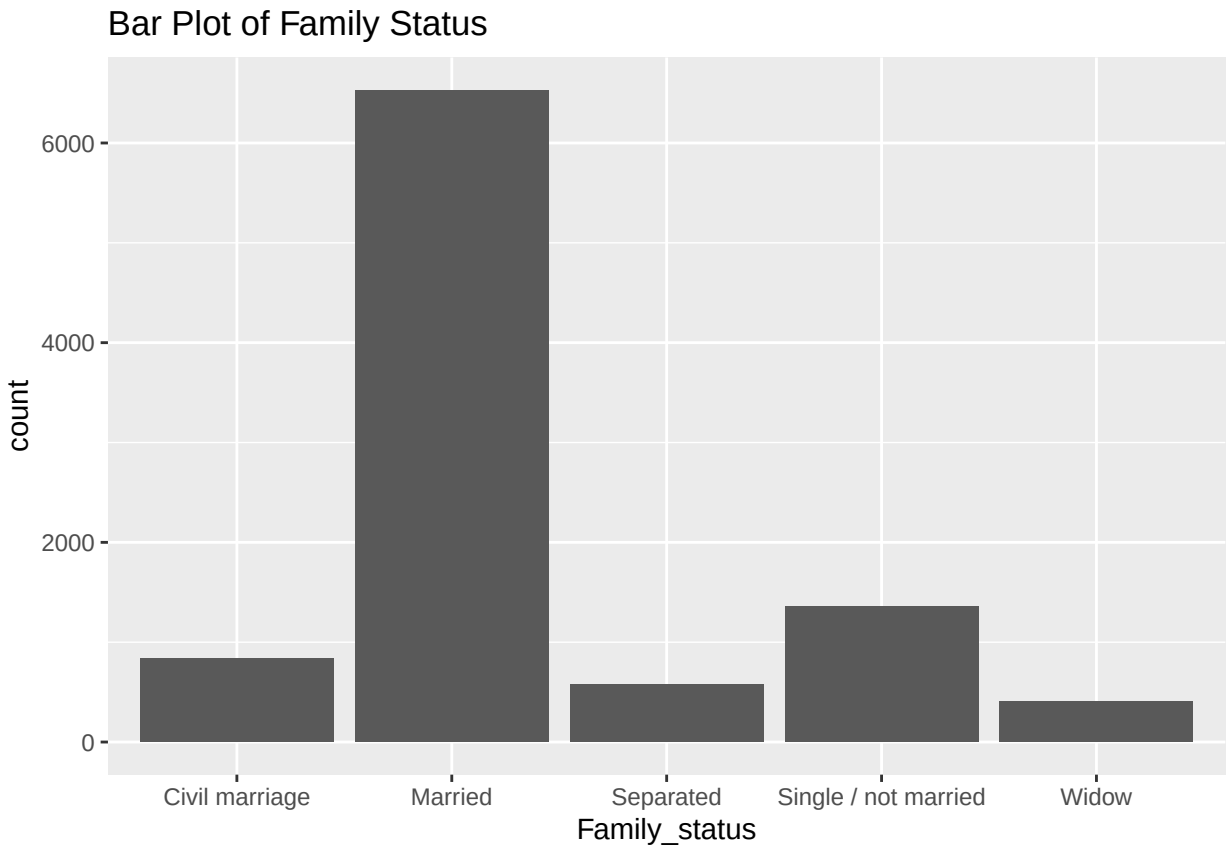
```
##
##       Civil marriage              Married            Separated
##                  836                 6530                  574
## Single / not married               Widow
##                 1359                  410
```

```
prop.table(table(df_qual$Family_status))
```

```
##
##       Civil marriage              Married            Separated
```

```
##           0.08610568           0.67257184           0.05912040
## Single / not married                Widow
##           0.13997322           0.04222886
```

```
ggplot(df_qual, aes(x=Family_status)) +
  geom_bar() +
  ggtitle("Bar Plot of Family Status")
```



Variable Analysis for Housing_type

A categorical variable that indicates the type of housing that a customer lives in (e.g., House/apartment, With parents, etc.).

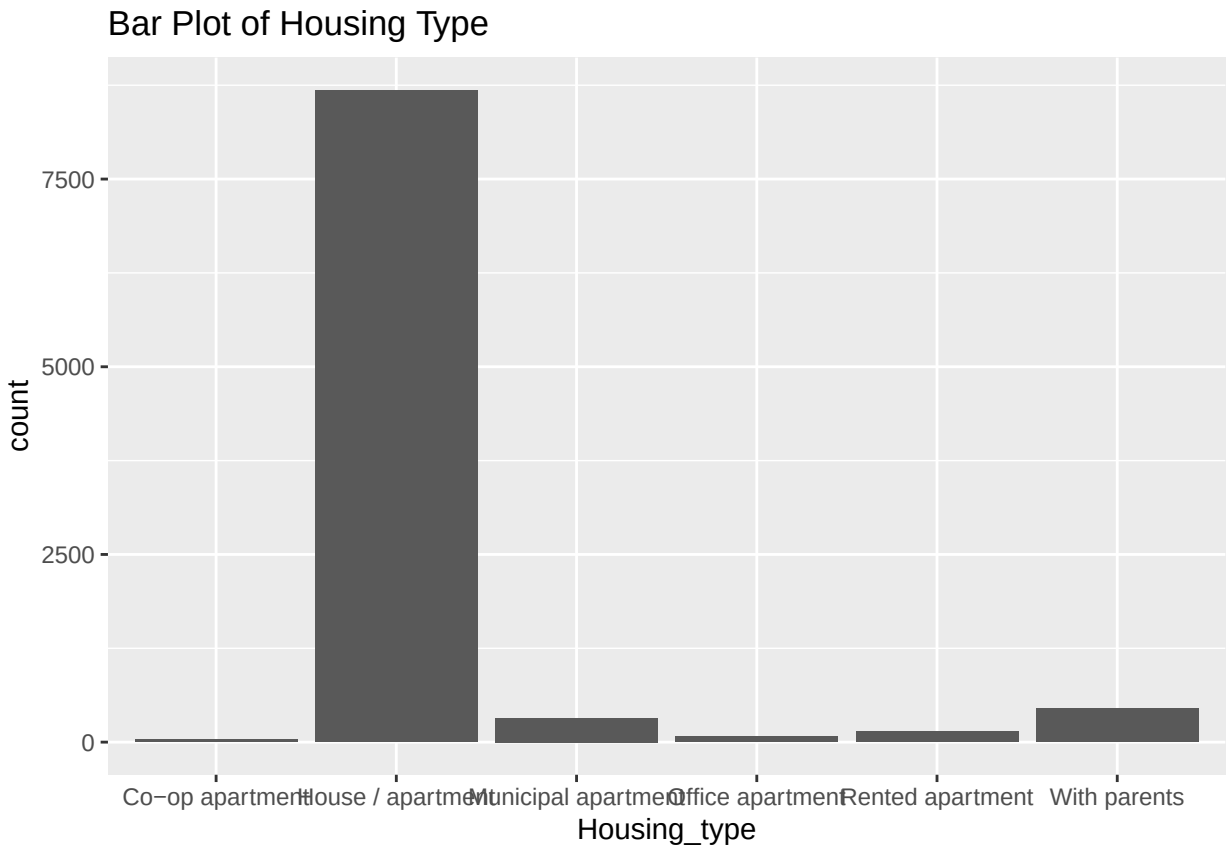
```
table(df_qual$Housing_type)
```

```
##
##     Co-op apartment   House / apartment Municipal apartment    Office apartment
##                  34                8684                 323                  76
##    Rented apartment        With parents
##                 144                 448
```

```
prop.table(table(df_qual$Housing_type))
```

```
##
##     Co-op apartment   House / apartment Municipal apartment    Office apartment
##         0.003501905         0.894427850         0.033268102         0.007827789
##    Rented apartment        With parents
##         0.014831600         0.046142754
```

```
ggplot(df_qual, aes(x=Housing_type)) +
  geom_bar() +
  ggtitle("Bar Plot of Housing Type")
```



Variable Analysis for Occupation_type Variable

A categorical variable that indicates the type of occupation that an individual is engaged in (e.g., Laborers, Sales Staff, Accountants, etc.). Different occupations indicate different levels of financial security.

```
table(df_qual$Occupation_type)
```

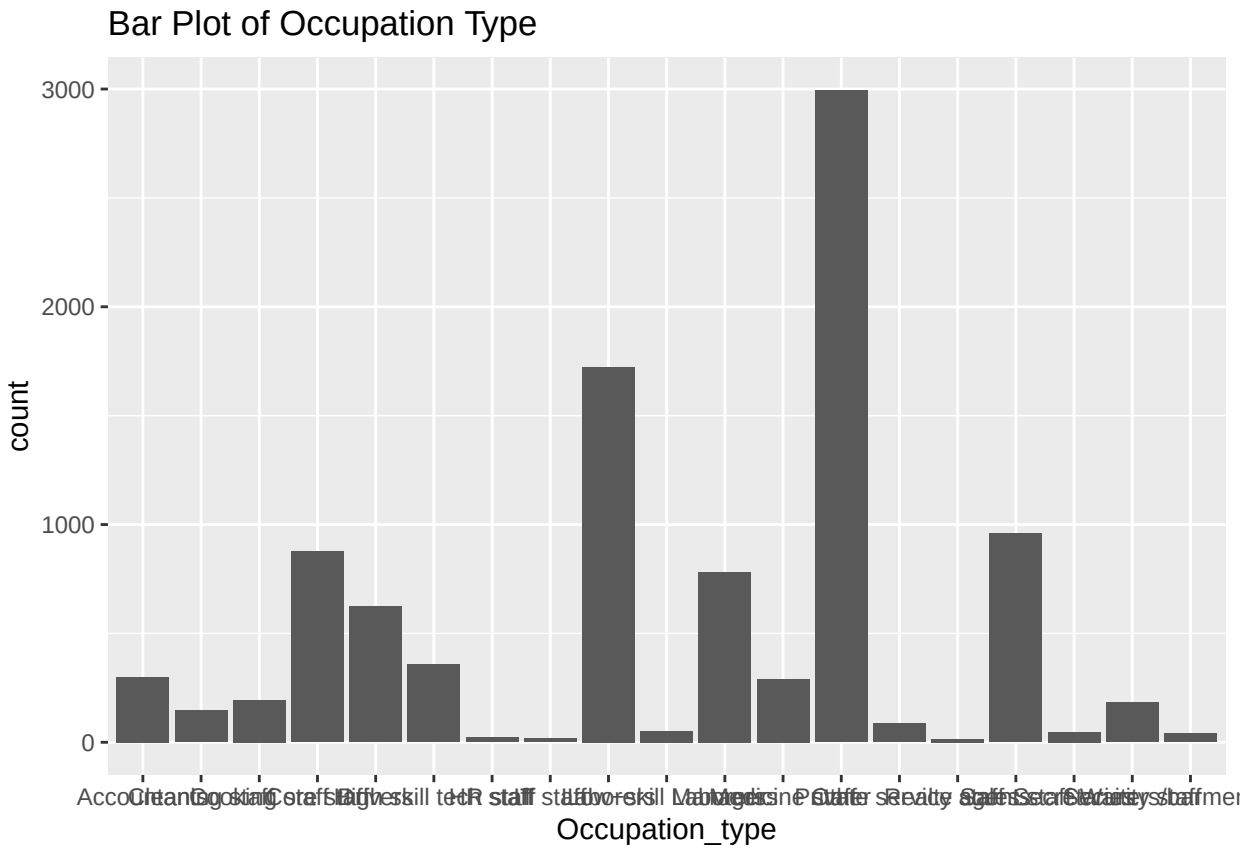
```
##
##           Accountants        Cleaning staff         Cooking staff
##                   300                   146                   193
##            Core staff               Drivers High skill tech staff
##                   877                   623                   357
##              HR staff            IT staff              Laborers
##                    22                    18                  1724
```

```
##    Low-skill Laborers               Managers        Medicine staff
##                    53                    782                   291
##                 Other Private service staff         Realty agents
##                  2994                     86                    16
##           Sales staff            Secretaries        Security staff
##                   959                     46                   182
##  Waiters/barmen staff
##                    40
```

```
prop.table(table(df_qual$Occupation_type))
```

```
##
##           Accountants        Cleaning staff         Cooking staff
##           0.030899166           0.015037594           0.019878463
##            Core staff               Drivers High skill tech staff
##           0.090328561           0.064167267           0.036770007
##               HR staff               IT staff               Laborers
##           0.002265939           0.001853950           0.177567206
##    Low-skill Laborers               Managers        Medicine staff
##           0.005458853           0.080543825           0.029972191
##                 Other Private service staff         Realty agents
##           0.308373674           0.008857761           0.001647956
##           Sales staff            Secretaries        Security staff
##           0.098774333           0.004737872           0.018745494
##  Waiters/barmen staff
##           0.004119889
```

```
ggplot(df_qual, aes(x=Occupation_type)) +
  geom_bar() +
  ggtitle("Bar Plot of Occupation Type")
```



Variable Analysis for Target Variable

A binary target variable that indicates if a person is eligible for a credit card. 0 indicates the customer is not eligible, while 1 indicates the person is eligible.

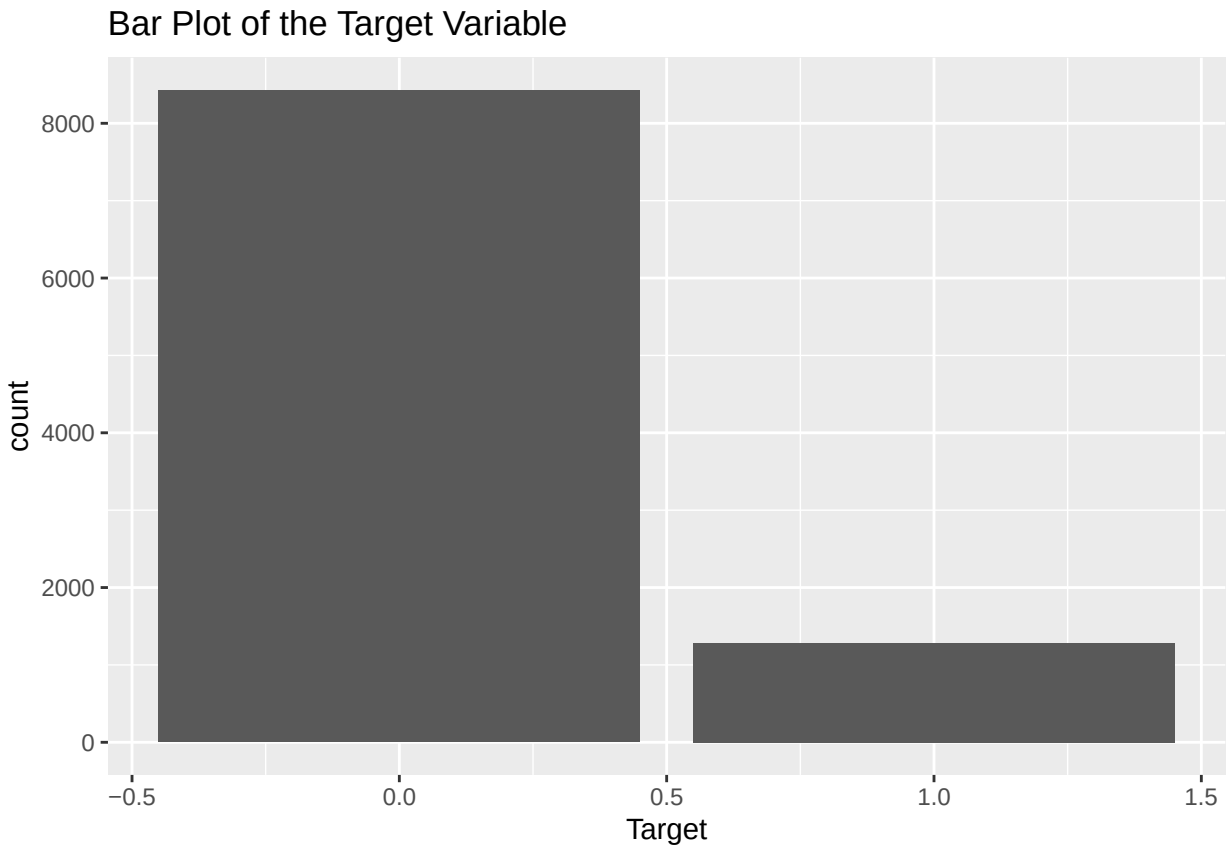
```
table(df_qual$Target)
```

```
##
##    0    1
## 8426 1283
```

```
prop.table(table(df_qual$Target))
```

```
##
##         0         1
## 0.8678546 0.1321454
```

```
ggplot(df_qual, aes(x=Target)) +
  geom_bar() +
  ggtitle("Bar Plot of the Target Variable")
```



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.