Initial Analysis and Model Performance

Sanatan Das

April 16, 2018

Contents

Initial Data Analysis

Load the Data

```
# load the data set from excel file
default_rates <- read_excel("C:/view/opt/apps/git/compscix-415-1-assignments/data/peps3xx.xls")
```

The Variables

```
# take a look at the data
glimpse(default_rates)
```

```
## Observations: 22,965
## Variables: 20
## $ RecordId   <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, ...
## $ OPEID      <chr> "001002", "001002", "001002", "001003", "001003", "...
## $ Name       <chr> "ALABAMA AGRICULTURAL & MECHANICAL UNIVERSITY", "AL...
## $ Address    <chr> "4900 MERIDIAN STREET", "4900 MERIDIAN STREET", "49...
## $ City       <chr> "NORMAL", "NORMAL", "NORMAL", "MONTGOMERY", "MONTGO...
## $ State      <chr> "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL...
## $ StateDesc  <chr> "ALABAMA", "ALABAMA", "ALABAMA", "ALABAMA", "ALABAM...
## $ ZipCode    <chr> "35762", "35762", "35762", "36109", "36109", "36109...
## $ ZipExt     <chr> "1357", "1357", "1357", "3398", "3398", "3398", "60...
## $ ProgLength <chr> "8", "8", "8", "8", "8", "8", "8", "8", "8", "8", "...
## $ SchoolType <chr> "1", "1", "1", "2", "2", "2", "1", "1", "1", "1", "...
## $ Year       <chr> "2014", "2013", "2012", "2014", "2013", "2012", "20...
## $ Num        <chr> "332", "300", "326", "192", "143", "143", "64", "57...
## $ Denom      <chr> "1753", "1812", "1895", "1470", "1491", "1417", "79...
## $ Drate      <chr> "18.9", "16.5", "17.2", "13.0", "9.5", "10.0", "8.0...
## $ Prate      <chr> "A", "A", "A", "A", "A", "A", "A", "A", "A", "P", "...
## $ EthnicCode <chr> "2", "2", "2", "5", "5", "5", "5", "5", "5", "2", "...
## $ CongDis    <chr> "D", "D", "D", "D", "D", "D", "D", "D", "D", "D", "...
## $ Region     <chr> "05", "05", "05", "02", "02", "02", "06", "06", "06...
## $ Avg        <chr> "04", "04", "04", "04", "04", "04", "04", "04", "04...
```

(Data Source : Federal Student Aid)

(Data Definition : Instructions for Using the Data Files)

Problem Category (Regression)

Our target variable is *default rate* or *drate* which is a numerical (double) variable. So our problem is categorically regression problem. We will create different kind of refression models to predict our data.

Predictors Data Types

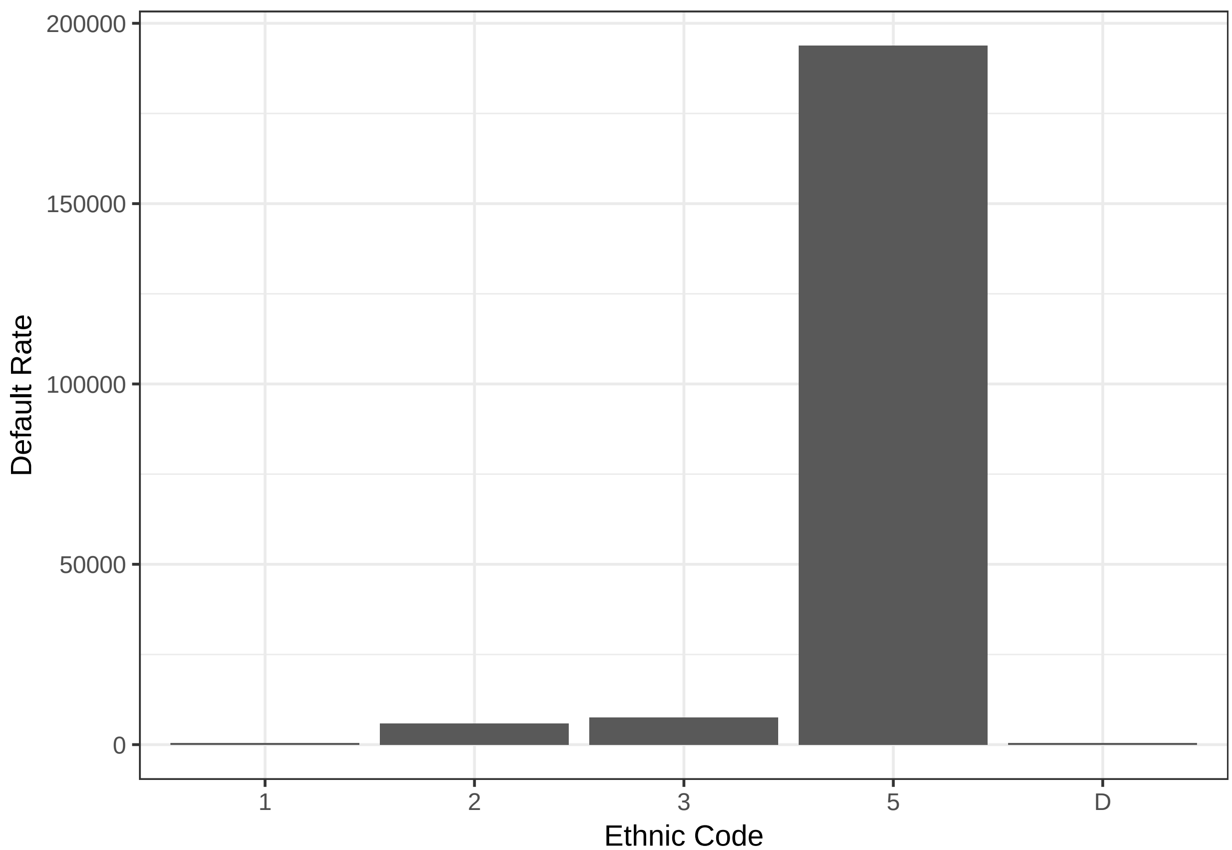
```
# add factor to the 'char' columns
default_rates$Name <- as.factor(default_rates$Name)
default_rates$State <- as.factor(default_rates$State)
default_rates$ZipCode <- as.factor(default_rates$ZipCode)
```

```
default_rates$ProgLength <- as.factor(default_rates$ProgLength)
default_rates$SchoolType <- as.factor(default_rates$SchoolType)
default_rates$EthnicCode <- as.factor(default_rates$EthnicCode)
default_rates$Prate <- as.factor(default_rates$Prate)
default_rates$CongDis <- as.factor(default_rates$CongDis)
# convert the columns to 'double' data type
default_rates$Drate <- as.double(default_rates$Drate)
default_rates$Num <- as.double(default_rates$Num)
default_rates$Denom <- as.double(default_rates$Denom)
```

Data Visualization

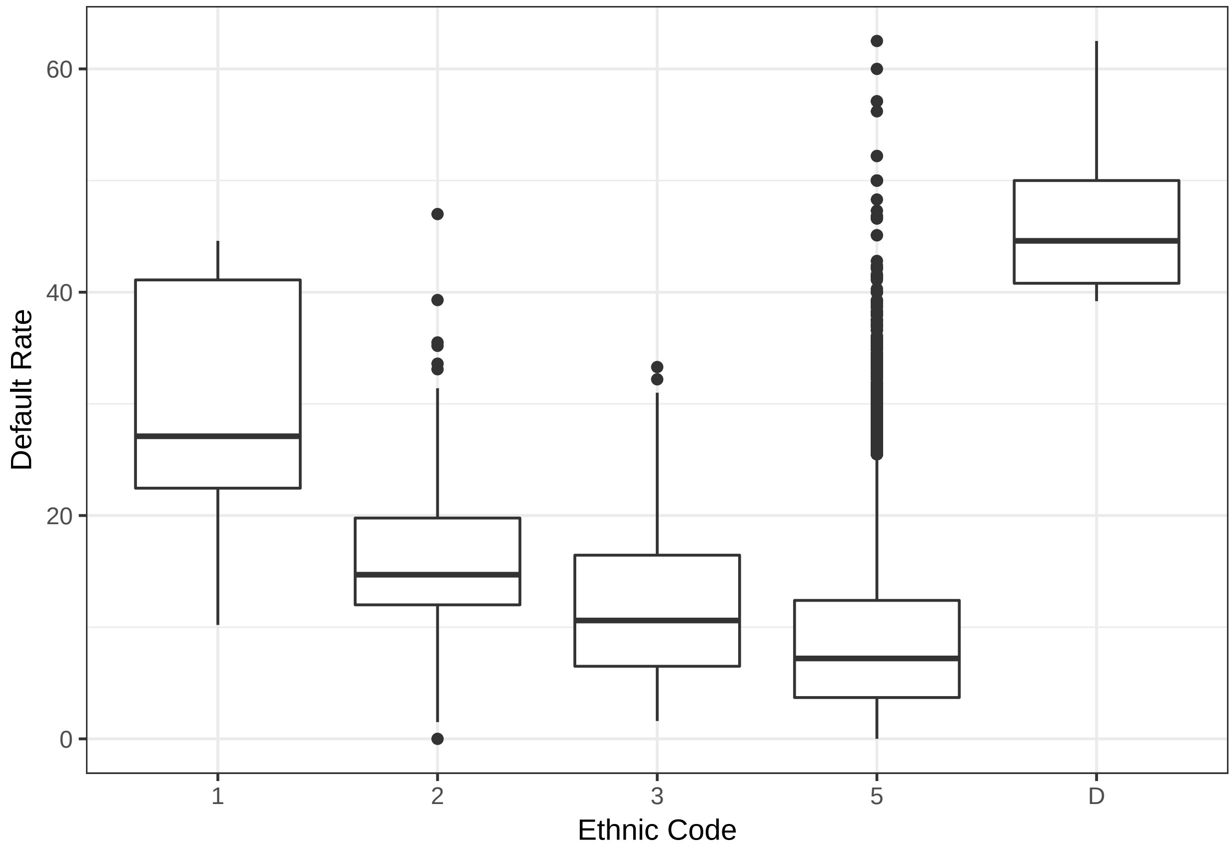
Relationship between Ethnic Code and the Default Rate

```
# plot the relationship between Ethnic Code and the Default Rate (Bar Graph)
default_rates %>%
ggplot() +
  geom_bar(aes(x = EthnicCode, y = Drate), stat = 'identity') +
  labs(x="Ethnic Code", y="Default Rate") +
  theme_bw()
```



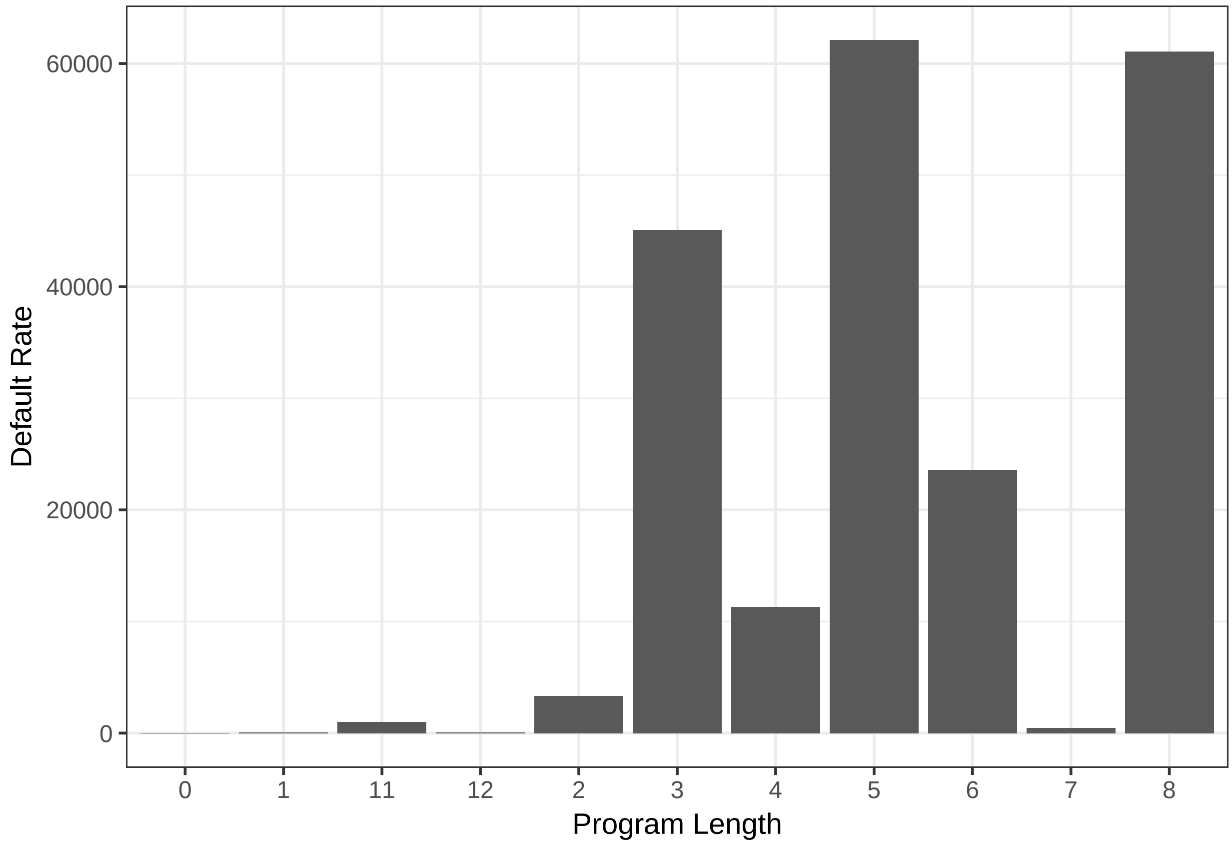

```
# plot the relationship between Ethnic Code and the Default Rate (Box Plot)
default_rates %>%
```

```r
ggplot(aes(x = EthnicCode, y = Drate), group=EthnicCode) +
  geom_boxplot() +
  labs(x="Ethnic Code", y="Default Rate") +
  theme_bw()
```

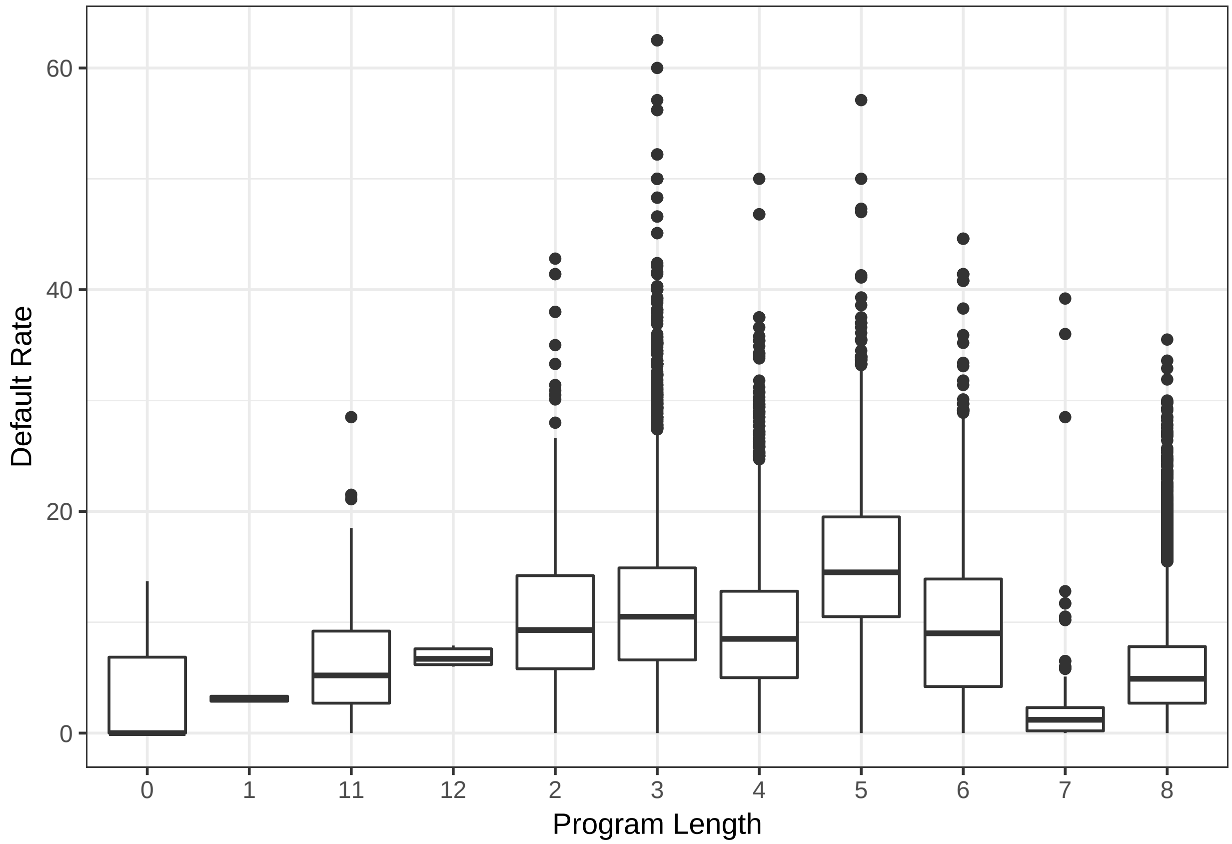


Relationship between Program Length and the Default Rate

```r
# plot the relationship between Program Length and the Default Rate (Bar Graph)
default_rates %>%
ggplot() +
  geom_bar(aes(x = ProgLength, y = Drate), stat = 'identity') +
  labs(x="Program Length", y="Default Rate") +
  theme_bw()
```

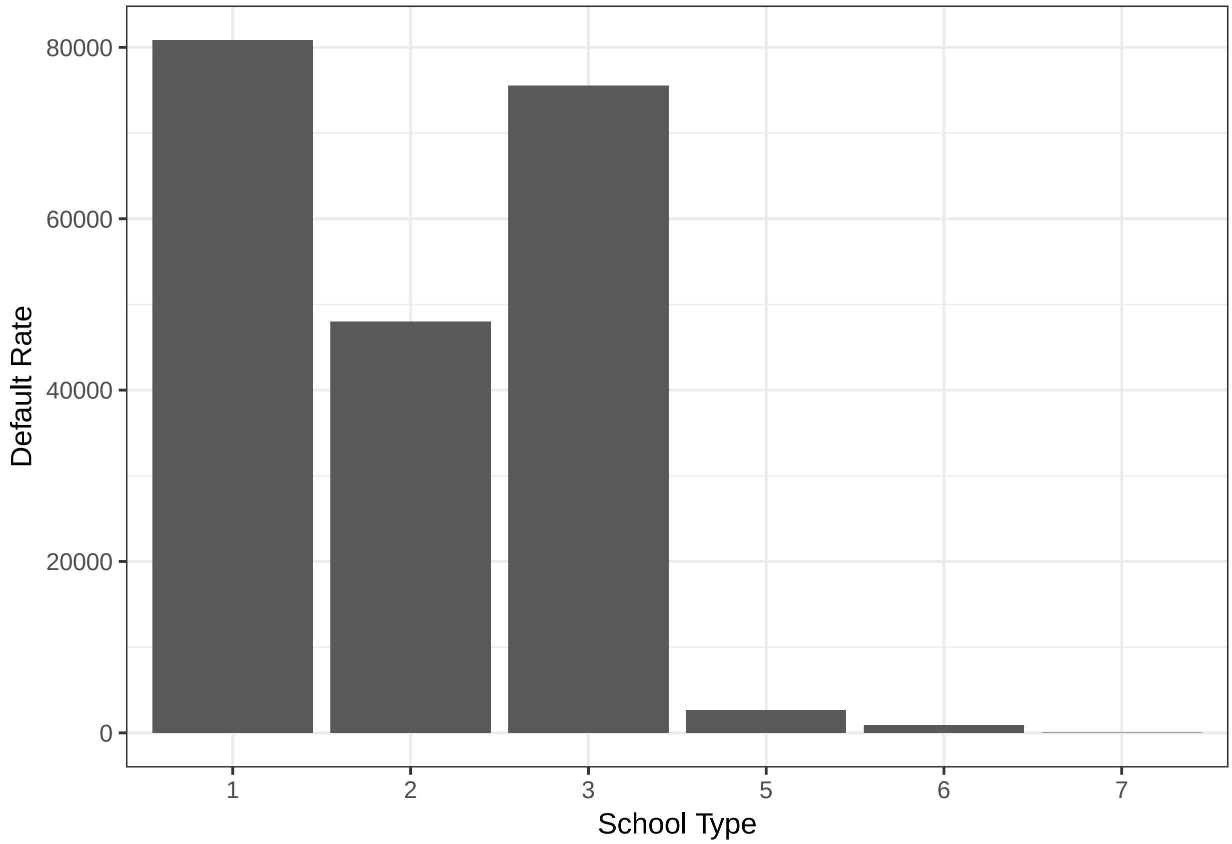



```r
# plot the relationship between Prog Length and the Default Rate (Box Plot)
default_rates %>%
ggplot(aes(x = ProgLength, y = Drate), group=ProgLength) +
  geom_boxplot() +
  labs(x="Program Length", y="Default Rate") +
  theme_bw()
```

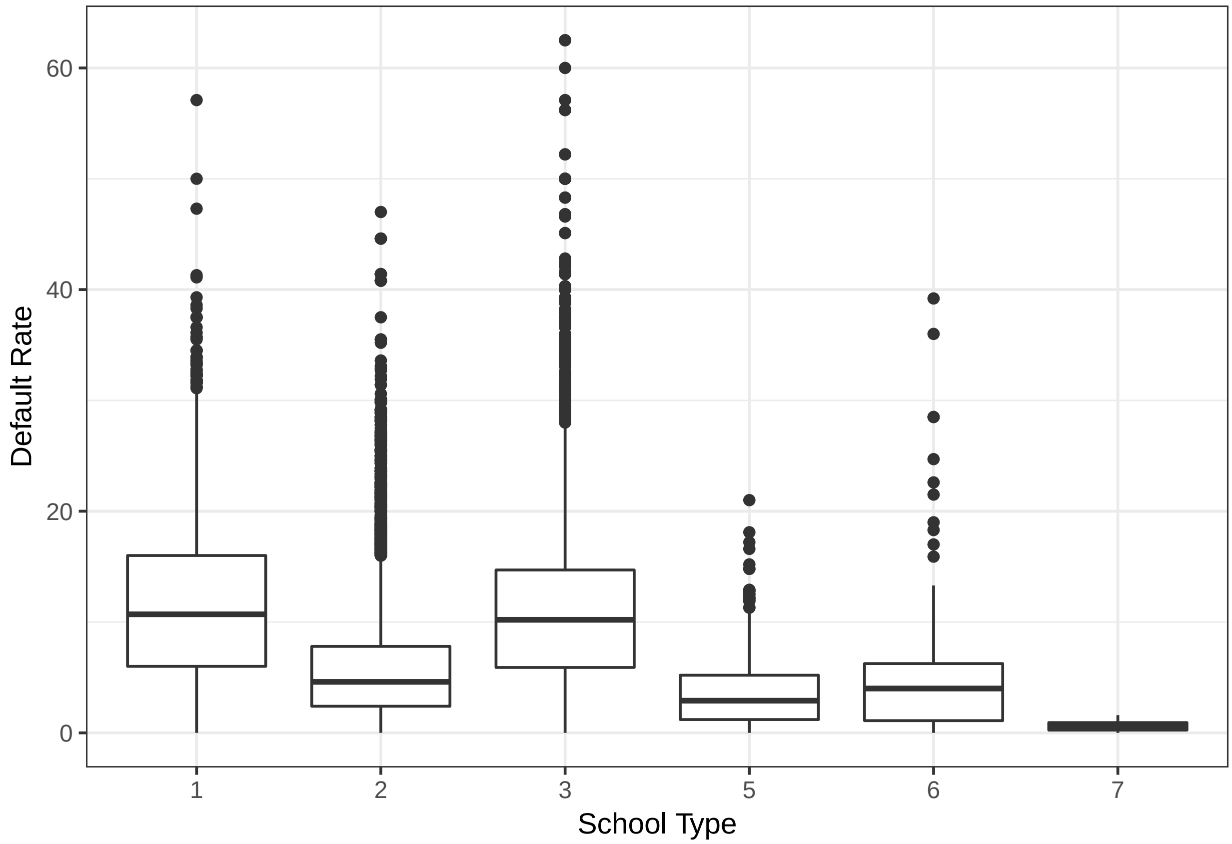


Relationship between School Type and the Default Rate

```
# plot the relationship between School Type and the Default Rate (Bar Graph)
default_rates %>%
ggplot() +
  geom_bar(aes(x = SchoolType, y = Drate), stat = 'identity') +
  labs(x="School Type", y="Default Rate") +
  theme_bw()
```

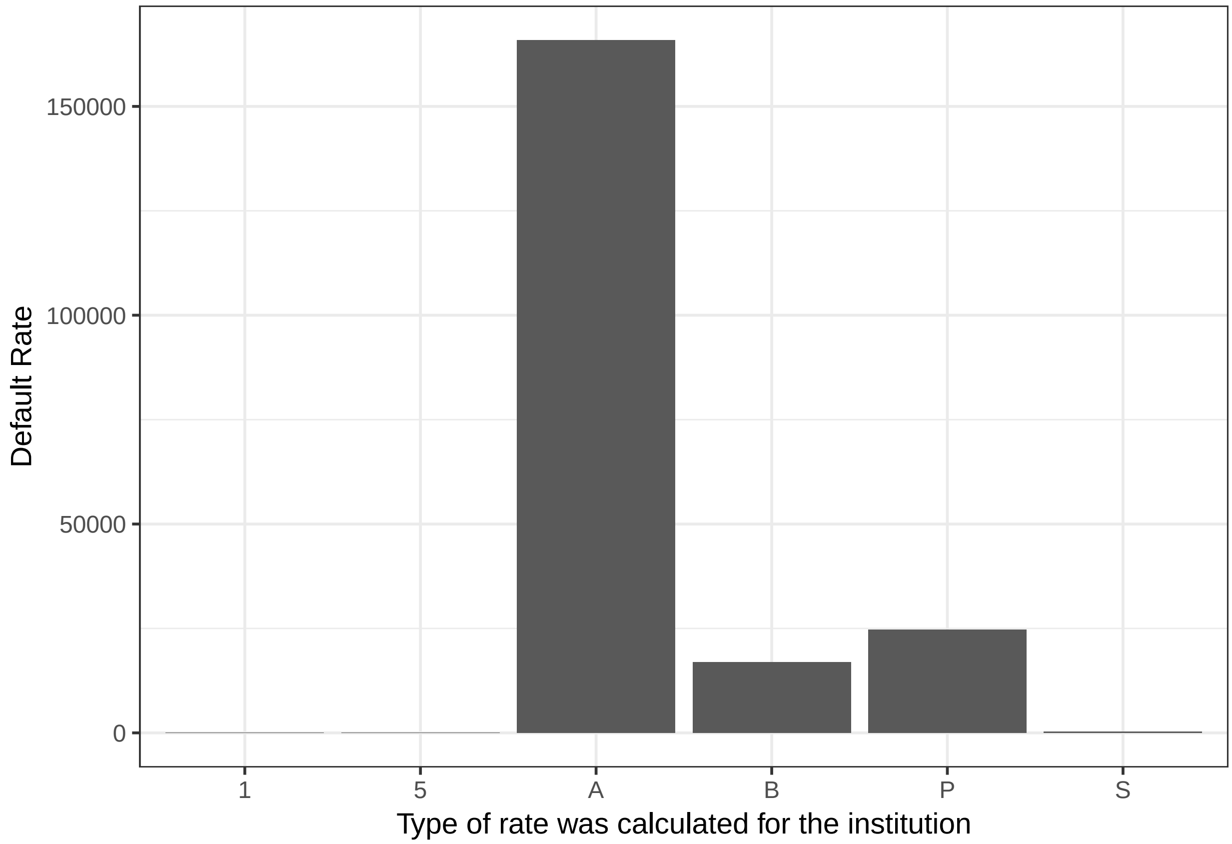



```r
# plot the relationship between School Type and the Default Rate (Box Plot)
default_rates %>%
ggplot(aes(x = SchoolType, y = Drate), group=SchoolType) +
  geom_boxplot() +
  labs(x="School Type", y="Default Rate") +
  theme_bw()
```

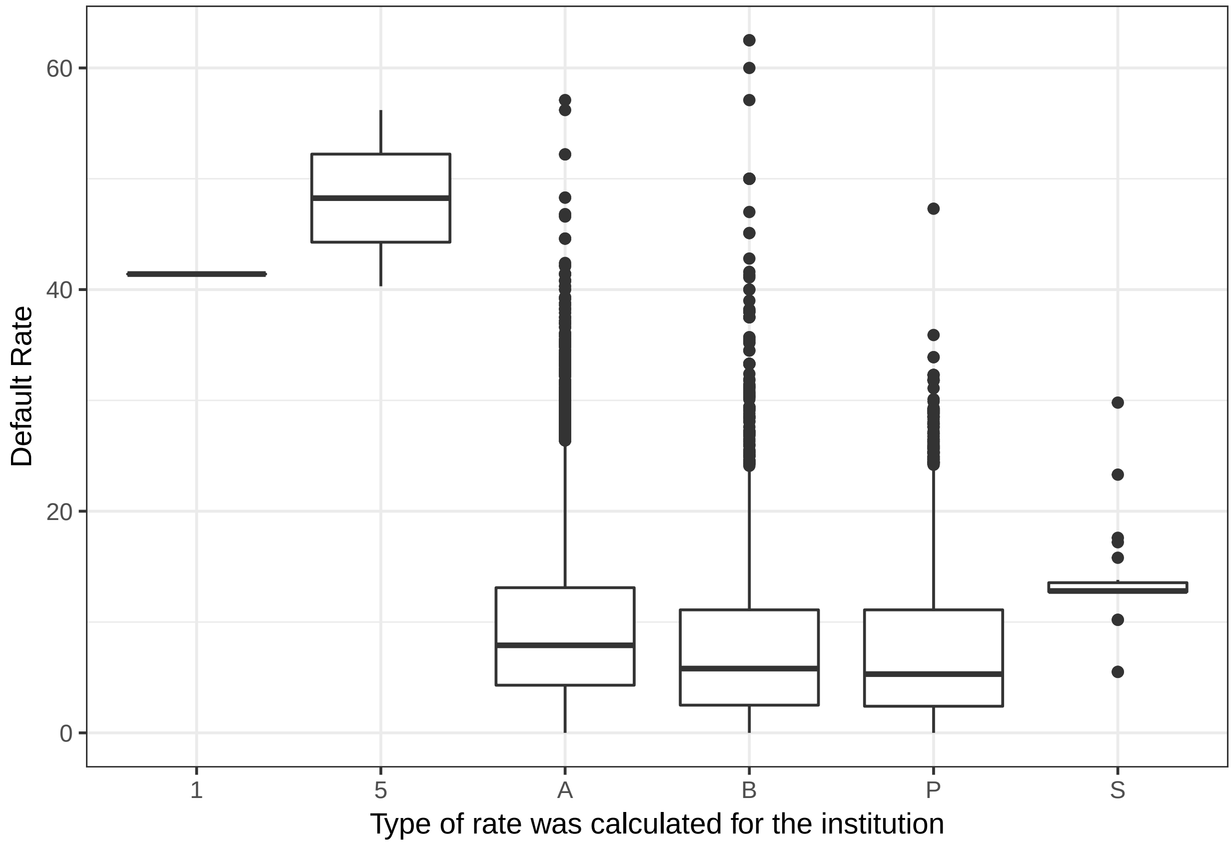


Relationship between Prate (Type of rate calculated for the institution.) and the Default Rate

```
# plot the relationship between Prate (Type of rate was calculated for the institution)
# and the Default Rate (Bar Graph)
default_rates %>%
ggplot() +
  geom_bar(aes(x = Prate, y = Drate), stat = 'identity') +
  labs(x="Type of rate was calculated for the institution", y="Default Rate") +
  theme_bw()
```

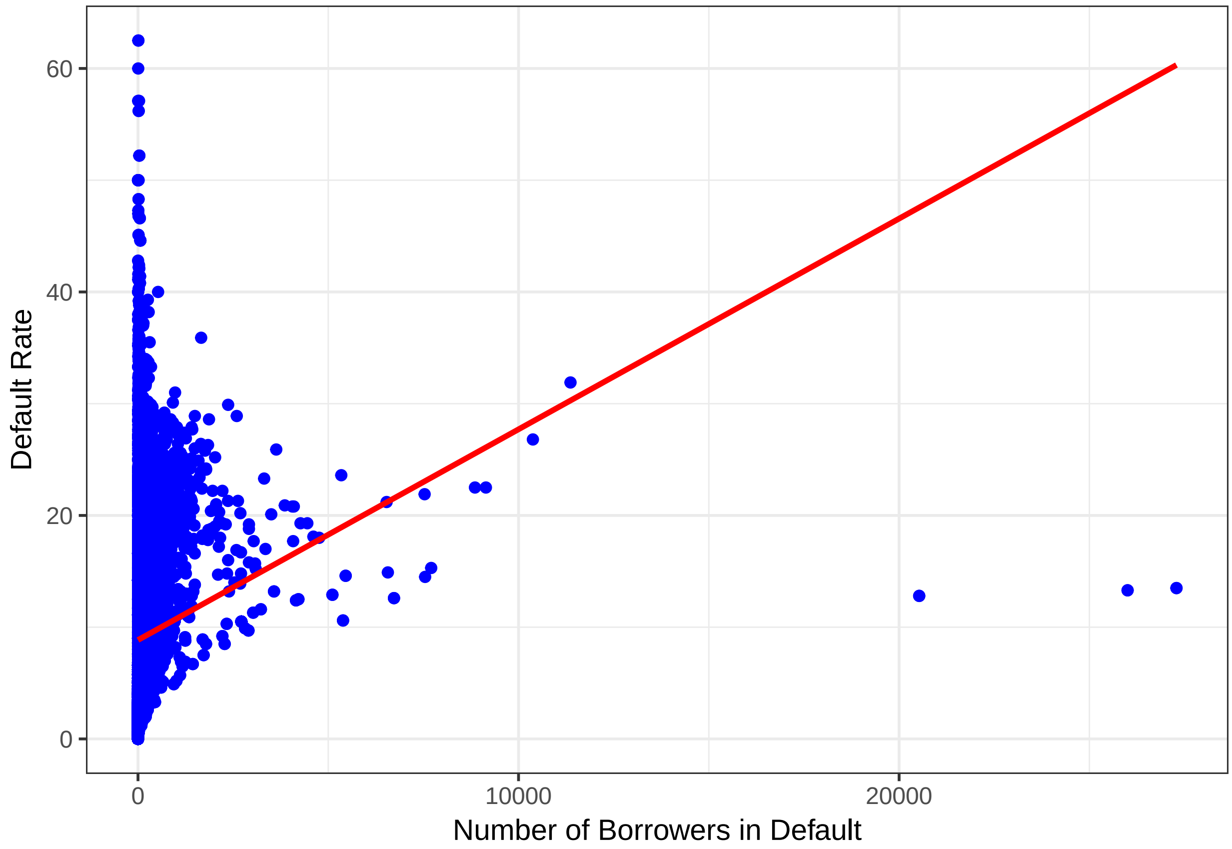



```
# plot the relationship between Prate (Type of rate was calculated for the institution)
# and the Default Rate (Box Plot)
default_rates %>%
ggplot(aes(x = Prate, y = Drate), group=Prate) +
  geom_boxplot() +
  labs(x="Type of rate was calculated for the institution", y="Default Rate") +
  theme_bw()
```



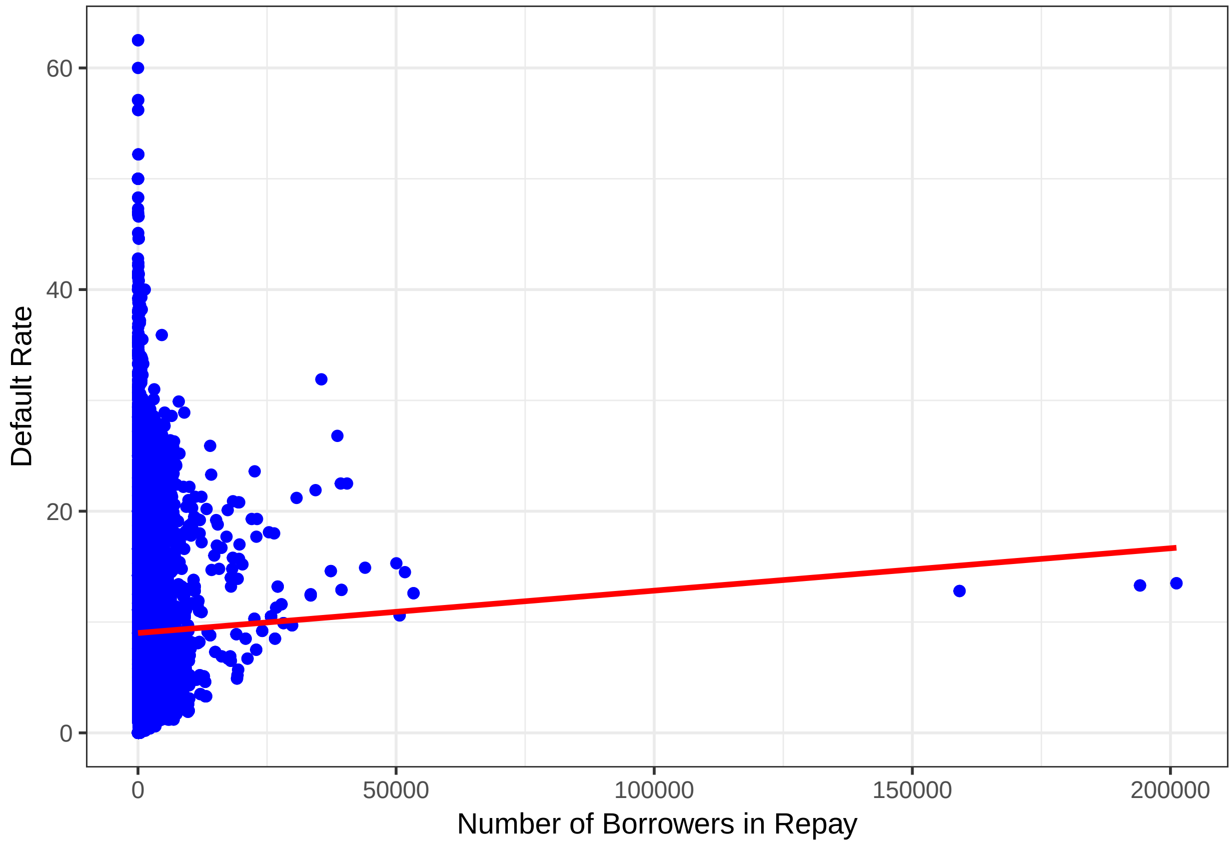
Relationship between Num (Number of Borrowers in Default) and the Default Rate

```
# plot the relationship between Num (Number of Borrowers in Default)
# and the Default Rate
ggplot(default_rates, aes(x = Num, y = Drate)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(x="Number of Borrowers in Default", y="Default Rate") +
  theme_bw()
```



Relationship between Denom (Number of Borrowers in Repay) and the Default Rate

```r
# plot the relationship between Denom (Number of Borrowers in Repay)
# and the Default Rate
ggplot(default_rates, aes(x = Denom, y = Drate)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(x="Number of Borrowers in Repay", y="Default Rate") +
  theme_bw()
```



Default Rate
60
40
20
0
0
50000
100000
150000
200000
Number of Borrowers in Repay

Model Performance (initial naive models)

Model evaluation on 'ProgLength' feature

```r
# Linear model on 'ProgLength' feature
lm_4 <- lm(Drate ~ ProgLength, data = default_rates)
glance(lm_4)
```

```
##   r.squared adj.r.squared    sigma statistic p.value df    logLik      AIC
## 1 0.2786509     0.2783366 5.964083  886.6931       0 11 -73590.29 147204.6
##        BIC deviance df.residual
## 1 147301.1 816480.4       22954
```

Model evaluation on 'SchoolType' feature

```r
# Linear model on 'SchoolType' feature
lm_5 <- lm(Drate ~ SchoolType, data = default_rates)
glance(lm_5)
```

```
##   r.squared adj.r.squared   sigma statistic p.value df    logLik      AIC
## 1 0.1726394     0.1724592 6.38662  958.1381       0  6 -75164.74 150343.5
##        BIC deviance df.residual
## 1 150399.8 936472.6       22959
```

Model evaluation on 'Num' feature

```r
# Linear model on 'Num' feature
lm_6 <- lm(Drate ~ Num, data = default_rates)
tidy(lm_6)
```

```
##          term    estimate    std.error statistic      p.value
## 1 (Intercept) 8.851168401 4.689508e-02  188.7441 0.000000e+00
## 2         Num 0.001885717 9.048456e-05   20.8402 1.434798e-95
```

Model evaluation on 'Denom' feature

```r
# Linear model on 'Denom' feature
lm_7 <- lm(Drate ~ Denom, data = default_rates)
tidy(lm_7)
```

```
##          term     estimate    std.error  statistic     p.value
## 1 (Intercept) 9.009393e+00 4.823076e-02 186.797640 0.000000000
## 2       Denom 3.823208e-05 1.214074e-05   3.149074 0.001639986
```

Model evaluation on 'EthnicCode' feature

```r
# Linear model on 'EthnicCode' feature
lm_9 <- lm(Drate ~ EthnicCode, data = default_rates)
glance(lm_9)
```

```
##    r.squared adj.r.squared    sigma statistic       p.value df    logLik
## 1 0.03850229    0.03833478 6.884752   229.853 7.804687e-194  5 -76890.01
##      AIC      BIC deviance df.residual
## 1 153792 153840.3  1088300       22960
```

Multiple Linear Regression Model

```r
# Final model using all features
lm_10 <- lm(Drate ~ ProgLength + SchoolType + Num + Denom + Prate + EthnicCode, data = default_rates)
glance(lm_10)
```

```
##   r.squared adj.r.squared    sigma statistic p.value df    logLik    AIC
## 1  0.405219     0.4045448 5.417525  601.0559       0 27 -71374.98 142806
##        BIC deviance df.residual
## 1 143031.1 673220.5       22938
```