630Project.R

Loyd

2020-11-18

```
# Sam Loyd
#
# Final Project
#
# The analysis of this data set will be performed in R.
#
# The data set was obtained as a compressed archive from Kaggle.

# Data Understanding

# LoanNr_ChkDgt Text Identifier – Primary key, I don't want this to confuse the model
# Name Text Borrower name - Remove for privacy concerns.
# City Text Borrower city
# State Text Borrower state
# Zip Text Borrower zip code
# Bank Text Bank name
# BankState Text Bank state
# NAICS Text North American industry classification system code
# ApprovalDate Date/Time Date SBA commitment issued
# ApprovalFY Text Fiscal year of commitment
# Term Number Loan term in months
# NoEmp Number Number of business employees
# NewExist Text 1 D Existing business, 2 D New business
# CreateJob Number Number of jobs created (Target variable)
# RetainedJob Number Number of jobs retained (*)
# FranchiseCode Text Franchise code, (00000 or 00001) D Nofranchise
# UrbanRural Text 1 D Urban, 2 D rural, 0 D undefined
# RevLineCr Text Revolving line of credit: Y D Yes, N D No
# LowDoc Text LowDoc Loan Program: Y D Yes, N D No
# ChgOffDate Date/Time The date when a loan is declared to be in default (*)
# DisbursementDate Date/Time Disbursement date (*)
# DisbursementGross Currency Amount disbursed (*)
# BalanceGross Currency Gross amount outstanding (*)
# MIS_Status Text Loan status charged off D CHGOFF, Paid in full D PIF (Target variable - Convert to logica
l)
# ChgOffPrinGr Currency Charged-off amount (*)
# GrAppv Currency Gross amount of loan approved by bank
# SBA_Appv Currency SBA's guaranteed amount of approved loan
# * warning - future information - most will need to be removed

# Given the nature of this data set it is important that we not let future data
# be used for our model. For example, several fields are based on information after the
# loan was approved. Disbursement is an example.

library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(pastecs)
```

```
##
## Attaching package: 'pastecs'
```

```
## The following objects are masked from 'package:dplyr':
##
##     first, last
```

```
library(psych)
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(purrr)
library(tidyr)
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:pastecs':
##
##     extract
```

```
library(dummies)
```

```
## dummies-1.5.6 provided by Decision Patterns
```

```
library(mltools)
```

```
##
## Attaching package: 'mltools'
```

```
## The following object is masked from 'package:tidyr':
##
##     replace_na
```

```
library(naniar)
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
## The following objects are masked from 'package:pastecs':
##
##     first, last
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
library(binaryLogic)
```

```
##
## Attaching package: 'binaryLogic'
```

```
## The following object is masked from 'package:purrr':
##
##     negate
```

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following objects are masked from 'package:data.table':
##
##     dcast, melt
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(binaryLogic)
library(RColorBrewer)
library(superml)
```

```
## Warning: package 'superml' was built under R version 4.0.3
```

```
## Loading required package: R6
```

```
# Turn off scientinfic notation

options(scipen=999)

sbaloan_data <- read.csv('SBAnational.csv',na.strings=c("NA","NaN", " ","") )

# glimpse(sbaloan_data)
summary(sbaloan_data)
```

```
##  LoanNr_ChkDgt            Name               City              State
##  Min.   :1000014003   Length:899164      Length:899164      Length:899164
##  1st Qu.:2589757508   Class :character   Class :character   Class :character
##  Median :4361439006   Mode  :character   Mode  :character   Mode  :character
##  Mean   :4772612311
##  3rd Qu.:6904626505
##  Max.   :9996003010
##
##       Zip            Bank             BankState             NAICS
##  Min.   :    0   Length:899164      Length:899164      Min.   :     0
##  1st Qu.:27587   Class :character   Class :character   1st Qu.:235210
##  Median :55410   Mode  :character   Mode  :character   Median :445310
##  Mean   :53804                                         Mean   :398661
##  3rd Qu.:83704                                         3rd Qu.:561730
##  Max.   :99999                                         Max.   :928120
##
##  ApprovalDate        ApprovalFY             Term           NoEmp
##  Length:899164      Length:899164      Min.   :  0.0   Min.   :   0.00
##  Class :character   Class :character   1st Qu.: 60.0   1st Qu.:   2.00
##  Mode  :character   Mode  :character   Median : 84.0   Median :   4.00
##                                        Mean   :110.8   Mean   :  11.41
##                                        3rd Qu.:120.0   3rd Qu.:  10.00
##                                        Max.   :569.0   Max.   :9999.00
##
##     NewExist      CreateJob        RetainedJob     FranchiseCode
##  Min.   :0.00   Min.   :   0.00   Min.   :   0.0   Min.   :    0
##  1st Qu.:1.00   1st Qu.:   0.00   1st Qu.:   0.0   1st Qu.:    1
##  Median :1.00   Median :   0.00   Median :   1.0   Median :    1
##  Mean   :1.28   Mean   :   8.43   Mean   :  10.8   Mean   : 2754
##  3rd Qu.:2.00   3rd Qu.:   1.00   3rd Qu.:   4.0   3rd Qu.:    1
##  Max.   :2.00   Max.   :8800.00   Max.   :9500.0   Max.   :99999
##  NA's   :136
##    UrbanRural      RevLineCr            LowDoc           ChgOffDate
##  Min.   :0.0000   Length:899164      Length:899164      Length:899164
##  1st Qu.:0.0000   Class :character   Class :character   Class :character
##  Median :1.0000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :0.7577
##  3rd Qu.:1.0000
##  Max.   :2.0000
##
##  DisbursementDate   DisbursementGross  BalanceGross        MIS_Status
##  Length:899164      Length:899164      Length:899164      Length:899164
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  ChgOffPrinGr          GrAppv            SBA_Appv
##  Length:899164      Length:899164      Length:899164
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

```
# Clean up any non - unique data - None found
sbaloan_data <- sbaloan_data %>% distinct()
summary(sbaloan_data)
```

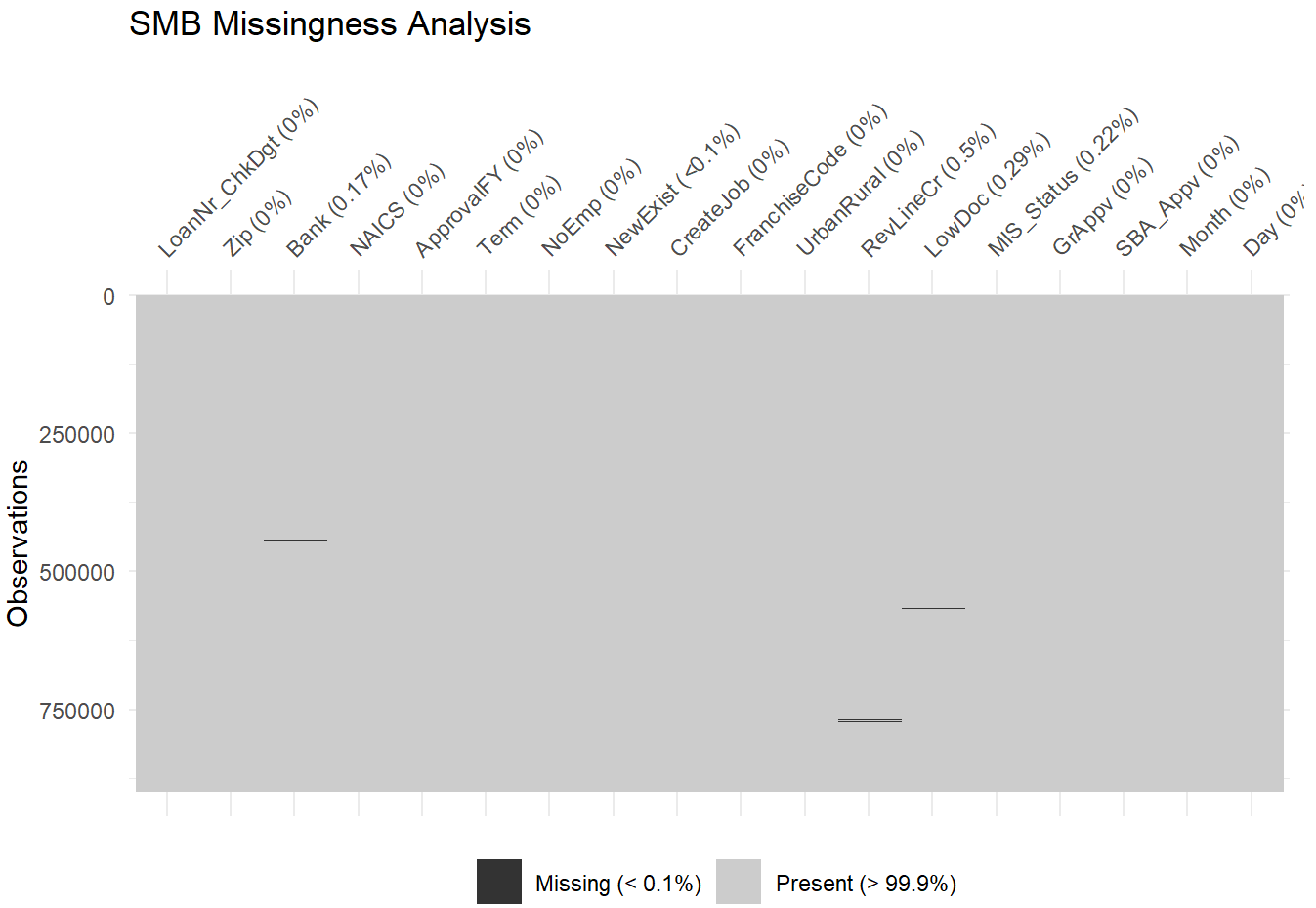
```
##  LoanNr_ChkDgt            Name               City              State
##  Min.   :1000014003   Length:899164      Length:899164      Length:899164
##  1st Qu.:2589757508   Class :character   Class :character   Class :character
##  Median :4361439006   Mode  :character   Mode  :character   Mode  :character
##  Mean   :4772612311
##  3rd Qu.:6904626505
##  Max.   :9996003010
##
##       Zip            Bank             BankState             NAICS
##  Min.   :    0   Length:899164      Length:899164      Min.   :     0
##  1st Qu.:27587   Class :character   Class :character   1st Qu.:235210
##  Median :55410   Mode  :character   Mode  :character   Median :445310
##  Mean   :53804                                         Mean   :398661
##  3rd Qu.:83704                                         3rd Qu.:561730
##  Max.   :99999                                         Max.   :928120
##
##  ApprovalDate        ApprovalFY             Term           NoEmp
##  Length:899164      Length:899164      Min.   :  0.0   Min.   :   0.00
##  Class :character   Class :character   1st Qu.: 60.0   1st Qu.:   2.00
##  Mode  :character   Mode  :character   Median : 84.0   Median :   4.00
##                                        Mean   :110.8   Mean   :  11.41
##                                        3rd Qu.:120.0   3rd Qu.:  10.00
##                                        Max.   :569.0   Max.   :9999.00
##
##     NewExist      CreateJob        RetainedJob     FranchiseCode
##  Min.   :0.00   Min.   :   0.00   Min.   :   0.0   Min.   :    0
##  1st Qu.:1.00   1st Qu.:   0.00   1st Qu.:   0.0   1st Qu.:    1
##  Median :1.00   Median :   0.00   Median :   1.0   Median :    1
##  Mean   :1.28   Mean   :   8.43   Mean   :  10.8   Mean   : 2754
##  3rd Qu.:2.00   3rd Qu.:   1.00   3rd Qu.:   4.0   3rd Qu.:    1
##  Max.   :2.00   Max.   :8800.00   Max.   :9500.0   Max.   :99999
##  NA's   :136
##    UrbanRural      RevLineCr            LowDoc           ChgOffDate
##  Min.   :0.0000   Length:899164      Length:899164      Length:899164
##  1st Qu.:0.0000   Class :character   Class :character   Class :character
##  Median :1.0000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :0.7577
##  3rd Qu.:1.0000
##  Max.   :2.0000
##
##  DisbursementDate   DisbursementGross  BalanceGross        MIS_Status
##  Length:899164      Length:899164      Length:899164      Length:899164
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  ChgOffPrinGr          GrAppv            SBA_Appv
##  Length:899164      Length:899164      Length:899164
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

```r
# Get rid of $ sign and commas
sbaloan_data$GrAppv <- gsub('[^a-zA-Z0-9.]', '', sbaloan_data$GrAppv)
sbaloan_data$ApprovalFY <- gsub('[^a-zA-Z0-9.]', '', sbaloan_data$ApprovalFY)


# Do we have job creation data for each Year.
# unique(sbaloan_data$ApprovalFY)

CheckFY <- filter(sbaloan_data, `CreateJob` > 0)

NoCheckFY <- filter(sbaloan_data, `CreateJob` == 0)

# Commented out for length
# CheckFY %>% count(ApprovalFY)
# NoCheckFY %>% count(ApprovalFY)

# How many unique banks
#unique(sbaloan_data$Bank)


# Convert GrAppv to numeric
sbaloan_data$GrAppv <- as.numeric(as.character(sbaloan_data$GrAppv))
sbaloan_data$SBA_Appv <- gsub('[^a-zA-Z0-9.]', '', sbaloan_data$SBA_Appv)
sbaloan_data$RevLineCr <- gsub('[^a-zA-Z0-9]', 'UNKNOWN', sbaloan_data$RevLineCr)

# sbaloan_data$RevLineCr
# unique(sbaloan_data$RevLineCr)

# Convert GrAppv to numeric
sbaloan_data$SBA_Appv <- as.numeric(as.character(sbaloan_data$SBA_Appv))

# Domain Knowledge
# Retrieved from https://www.investopedia.com/terms/c/chargeoff.asp
# Quote:
# A charge-off refers to debt that a company believes it will no longer
# collect as the borrower has become delinquent on payments.
# This would be future knowledge as I am interested in loans that
# Retrieved from https://www.investopedia.com/terms/d/disbursement.asp
# Quote:
# A student loan disbursement is the paying out of loan proceeds to a borrower
# Rebrieved from https://www.census.gov/eos/www/naics/faqs/faqs.html#q1
# Quote:
# The North American Industry Classification System (NAICS,
# pronounced Nakes) was developed under the direction
# and guidance of the Office of Management and Budget (OMB)
# as the standard for use by Federal statistical agencies
# in classifying business establishments for the collection,
# tabulation, presentation, and analysis of statistical
# data describing the U.S. economy. Use of the standard provides
# uniformity and comparability in the presentation of
# these statistical data. NAICS is based on a production-oriented
# concept, meaning that it groups establishments
# into industries according to similarity in the processes
# used to produce goods or services.
# NAICS replaced the Standard Industrial
# Classification (SIC) system in 1997.
```

```
# head(sbaloan_data)
summary(sbaloan_data)
```

```
##  LoanNr_ChkDgt            Name               City              State
##  Min.   :1000014003   Length:899164      Length:899164      Length:899164
##  1st Qu.:2589757508   Class :character   Class :character   Class :character
##  Median :4361439006   Mode  :character   Mode  :character   Mode  :character
##  Mean   :4772612311
##  3rd Qu.:6904626505
##  Max.   :9996003010
##
##       Zip            Bank            BankState             NAICS
##  Min.   :    0   Length:899164      Length:899164      Min.   :     0
##  1st Qu.:27587   Class :character   Class :character   1st Qu.:235210
##  Median :55410   Mode  :character   Mode  :character   Median :445310
##  Mean   :53804                                         Mean   :398661
##  3rd Qu.:83704                                         3rd Qu.:561730
##  Max.   :99999                                         Max.   :928120
##
##  ApprovalDate        ApprovalFY             Term           NoEmp
##  Length:899164      Length:899164      Min.   :  0.0   Min.   :   0.00
##  Class :character   Class :character   1st Qu.: 60.0   1st Qu.:   2.00
##  Mode  :character   Mode  :character   Median : 84.0   Median :   4.00
##                                        Mean   :110.8   Mean   :  11.41
##                                        3rd Qu.:120.0   3rd Qu.:  10.00
##                                        Max.   :569.0   Max.   :9999.00
##
##     NewExist      CreateJob        RetainedJob     FranchiseCode
##  Min.   :0.00   Min.   :   0.00   Min.   :   0.0   Min.   :    0
##  1st Qu.:1.00   1st Qu.:   0.00   1st Qu.:   0.0   1st Qu.:    1
##  Median :1.00   Median :   0.00   Median :   1.0   Median :    1
##  Mean   :1.28   Mean   :   8.43   Mean   :  10.8   Mean   : 2754
##  3rd Qu.:2.00   3rd Qu.:   1.00   3rd Qu.:   4.0   3rd Qu.:    1
##  Max.   :2.00   Max.   :8800.00   Max.   :9500.0   Max.   :99999
##  NA's   :136
##    UrbanRural      RevLineCr            LowDoc           ChgOffDate
##  Min.   :0.0000   Length:899164      Length:899164      Length:899164
##  1st Qu.:0.0000   Class :character   Class :character   Class :character
##  Median :1.0000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :0.7577
##  3rd Qu.:1.0000
##  Max.   :2.0000
##
##  DisbursementDate   DisbursementGross  BalanceGross        MIS_Status
##  Length:899164      Length:899164      Length:899164      Length:899164
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  ChgOffPrinGr           GrAppv           SBA_Appv
##  Length:899164      Min.   :    200   Min.   :    100
##  Class :character   1st Qu.:  35000   1st Qu.:  21250
##  Mode  :character   Median :  90000   Median :  61250
##                     Mean   : 192687   Mean   : 149489
##                     3rd Qu.: 225000   3rd Qu.: 175000
##                     Max.   :5472000   Max.   :5472000
##
```

```
sbaloan_data$Month = substr(sbaloan_data$ApprovalDate,3,6)
sbaloan_data$Month <- gsub('[^a-zA-Z]', '', sbaloan_data$Month)
# sbaloan_data$Month
sbaloan_data$Day = substr(sbaloan_data$ApprovalDate,1,2)
sbaloan_data$Day <- gsub('[^[:alnum:]]', '', sbaloan_data$Day)
# sbaloan_data$Day

# Remove future data by dropping columns.
# Future data: BalanceGross, DisbursmentGross, DisbursementDate
#              ChgOffDate, ChgOffPrinGr, RetainedJob, and BalanceGross
#              ApprovalDate, ApprovalFY, DisbursmentGross, Disbursement Date
# Name - removed for privacy and security
# Approval Fiscal Date and Year removed as future data
# and to avoid problems with predictions at the beginning of a new year.
# City and state - removed as redundant with zip
# Domain Knowledge
# It is also illegal to use local zip in determining loans -
# can lead to racial bias
# Bank geographical information as well .... page 81 (Seigel, 2016)

# Column Removal
sbaloan_data <- subset(sbaloan_data, select = -c(`BankState`,`City`,`State`,`ApprovalDate`,`BalanceGross`,`
RetainedJob`,`Name`,`DisbursementGross`,`DisbursementDate`,`ChgOffDate`,`ChgOffPrinGr`))


vis_miss(sbaloan_data,warn_large_data=FALSE)  + ggtitle("SMB Missingness Analysis")
```



```
# Data preparation

# I am going to keep most of these as categories
# Converting here as its a pain after they are factors
sbaloan_data[is.na(sbaloan_data)]<-'UNKNOWN'


# sbaloan_data$Bank[is.na(sbaloan_data$Bank)] <- " "

# Convert variable as appropriate for analysis
sbaloan_data$Bank = as.factor(sbaloan_data$Bank)
sbaloan_data$NAICS = as.factor(sbaloan_data$NAICS)
sbaloan_data$FranchiseCode = as.factor(sbaloan_data$FranchiseCode)
sbaloan_data$UrbanRural = as.factor(sbaloan_data$UrbanRural)
sbaloan_data$RevLineCr = as.factor(sbaloan_data$RevLineCr)
sbaloan_data$ApprovalFY = as.factor(sbaloan_data$ApprovalFY)
sbaloan_data$Day = as.factor(sbaloan_data$Day)
sbaloan_data$RevLineCr = as.factor(sbaloan_data$RevLineCr)
sbaloan_data$LowDoc = as.factor(sbaloan_data$LowDoc)
sbaloan_data$MIS_Status = as.factor(sbaloan_data$MIS_Status)
sbaloan_data$NewExist = as.factor(sbaloan_data$NewExist)
sbaloan_data$ApprovalFY = as.integer(sbaloan_data$ApprovalFY)
sbaloan_data$Day = as.integer(sbaloan_data$Day)

# Remove loans with unknown status as this is a target variable
sbaloan_data <- filter(sbaloan_data, `MIS_Status` == "CHGOFF" | `MIS_Status` =="P I F")


# df1$IS_PASS = as.factor(df1$IS_PASS)
#glimpse(sbaloan_data)
summary(sbaloan_data)
```

```
##  LoanNr_ChkDgt            Zip                                   Bank
##  Min.   :1000014003   Min.   :    0   BANK OF AMERICA NATL ASSOC    : 86773
##  1st Qu.:2593070004   1st Qu.:27612   WELLS FARGO BANK NATL ASSOC   : 63461
##  Median :4363894001   Median :55416   JPMORGAN CHASE BANK NATL ASSOC: 48131
##  Mean   :4774981605   Mean   :53857   U.S. BANK NATIONAL ASSOCIATION: 35112
##  3rd Qu.:6908644007   3rd Qu.:83706   CITIZENS BANK NATL ASSOC      : 33770
##  Max.   :9996003010   Max.   :99999   PNC BANK, NATIONAL ASSOCIATION: 27340
##                                       (Other)                       :602580
##      NAICS          ApprovalFY         Term           NoEmp
##  0      :201667   Min.   : 3.00   Min.   :  0.0   Min.   :   0.00
##  722110 : 27941   1st Qu.:35.00   1st Qu.: 60.0   1st Qu.:   2.00
##  722211 : 19435   Median :40.00   Median : 84.0   Median :   4.00
##  811111 : 14539   Mean   :39.14   Mean   :110.8   Mean   :  11.41
##  621210 : 14034   3rd Qu.:44.00   3rd Qu.:120.0   3rd Qu.:  10.00
##  624410 : 10092   Max.   :52.00   Max.   :569.0   Max.   :9999.00
##  (Other):609459
##     NewExist        CreateJob        FranchiseCode    UrbanRural
##  0      :  1028   Min.   :   0.000   1      :637395   0:322826
##  1      :643446   1st Qu.:   0.000   0      :208040   1:469281
##  2      :252559   Median :   0.000   78760  :  3373   2:105060
##  UNKNOWN:   134   Mean   :   8.444   68020  :  1921
##                   3rd Qu.:   1.000   50564  :  1034
##                   Max.   :8800.000   21780  :  1001
##                                      (Other): 44403
##    RevLineCr          LowDoc         MIS_Status         GrAppv
##  N      :419252   N      :780997   CHGOFF :157558   Min.   :   1000
##  0      :257431   Y      :110171   P I F  :739609   1st Qu.:  35000
##  Y      :200660   UNKNOWN:  2578   UNKNOWN:     0   Median :  90000
##  T      : 15239   0      :  1490                    Mean   : 193060
##  UNKNOWN:  4534   C      :   758                    3rd Qu.: 225000
##  1      :    23   S      :   603                    Max.   :5472000
##  (Other):    28   (Other):   570
##     SBA_Appv          Month                Day
##  Min.   :    500   Length:897167      Min.   : 1.00
##  1st Qu.:  21250   Class :character   1st Qu.: 8.00
##  Median :  62050   Mode  :character   Median :16.00
##  Mean   : 149781                      Mean   :16.01
##  3rd Qu.: 175000                      3rd Qu.:23.00
##  Max.   :5472000                      Max.   :31.00
##
```

```
# Domain Knowledge
# from https://www.sba.gov/sites/default/files/SDOLoanFactSheet_Oct_2011.pdf
# $5 million
# Quote:
# The exact percentage of the guaranty depends on a variety of factors such as size of
# loan and which SBA program is to be used. This will be worked out between the

# SBA and your bank. Amounts - The maximum loan amount is $5 million. The total
# SBA guarantee for any one borrower may not exceed $3,750,000.
# The data shows outliers. Given the information above anything with over 3,750,000 and be imputed
# as likely a data error.
# Correcting for this.
sum(sbaloan_data$SBA_Appv > 3750000)
```

```
## [1] 57
```

```
# Given the large data set removing 57 rows should be acceptable

sbaloan_data <- filter(sbaloan_data, SBA_Appv < 3750001)

# Further validation
sum(sbaloan_data$GrAppv > 5000000)
```

```
## [1] 0
```

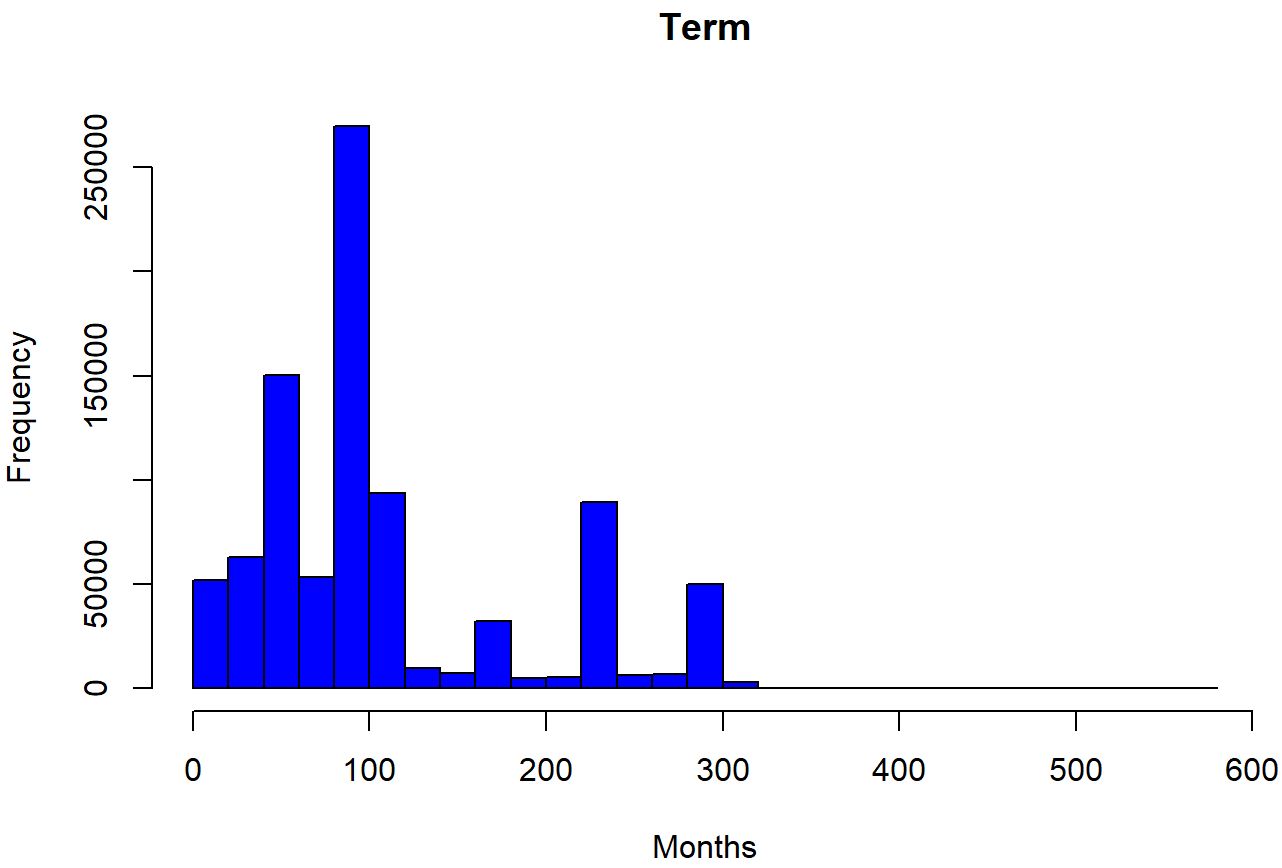
```
# Returned 0 so the above truncate took care of these as well


# Numeric only for statistics
num_sbaloan_data <- sbaloan_data[,c("Term", "NoEmp", "CreateJob", "SBA_Appv", "GrAppv")]

format(round(stat.desc(num_sbaloan_data,basic = TRUE, norm = FALSE), digits = 3), scientific=FALSE)
```

```
##                      Term        NoEmp   CreateJob         SBA_Appv
## nbr.val        897110.000   897110.000  897110.000       897110.000
## nbr.null          806.000     6617.000  627597.000            0.000
## nbr.na              0.000        0.000       0.000            0.000
## min                 0.000        0.000       0.000          500.000
## max               569.000     9999.000    8800.000      3750000.000
## range             569.000     9999.000    8800.000      3749500.000
## sum          99440200.000 10235833.000 7574913.000 134134716152.000
## median             84.000        4.000       0.000        62050.000
## mean              110.845       11.410       8.444       149518.695
## SE.mean             0.083        0.078       0.250          238.793
## CI.mean.0.95        0.163        0.153       0.490          468.027
## var              6224.705     5445.367   56148.933  51155277756.997
## std.dev            78.897       73.793     236.958       226175.325
## coef.var            0.712        6.467      28.063            1.513
##                       GrAppv
## nbr.val           897110.000
## nbr.null               0.000
## nbr.na                 0.000
## min                 1000.000
## max              5000000.000
## range            4999000.000
## sum          172944830593.000
## median            90000.000
## mean             192779.961
## SE.mean             296.931
## CI.mean.0.95        581.975
## var           79096532668.695
## std.dev          281241.058
## coef.var              1.459
```

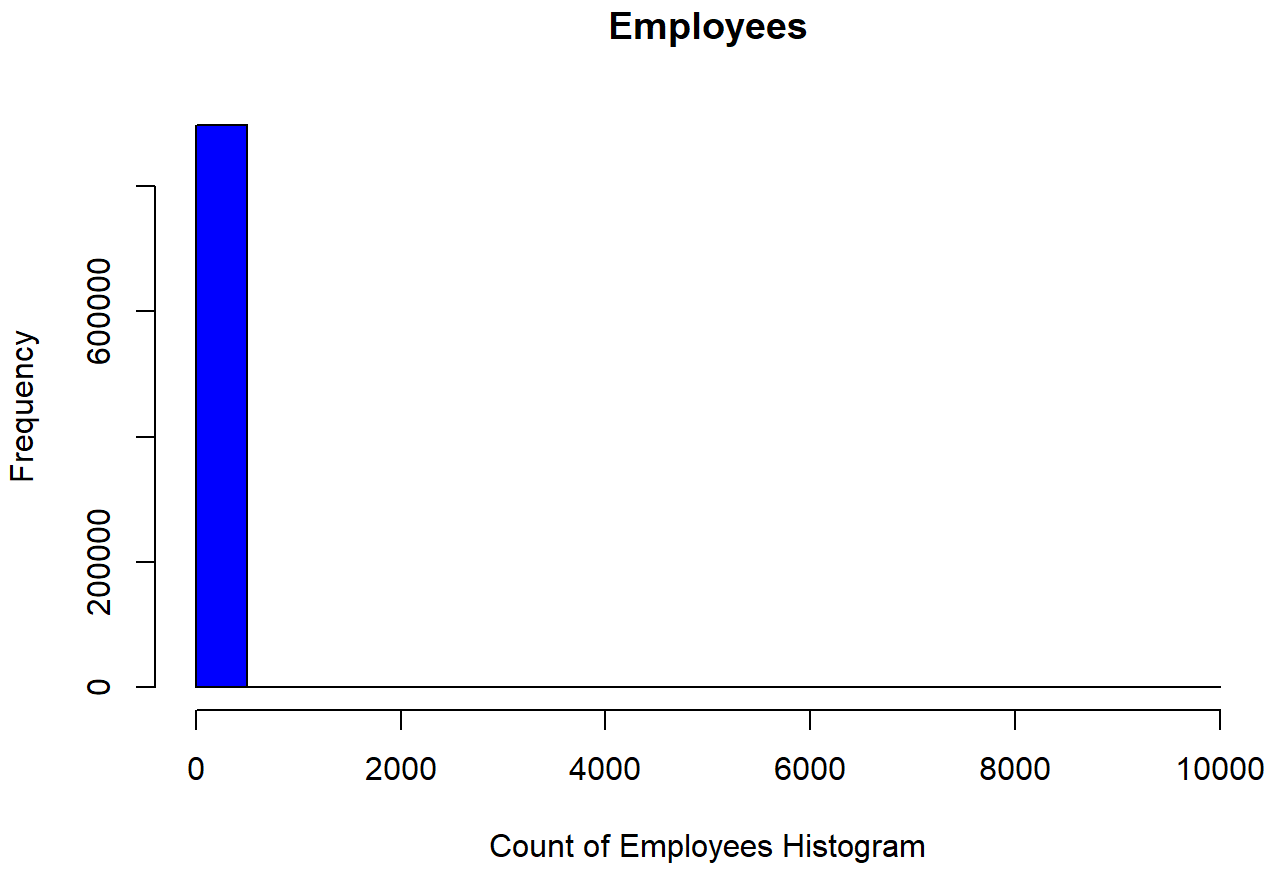
```
sample_num_sbaloan_data <- num_sbaloan_data[sample(1:nrow(num_sbaloan_data), 5000,
                                        replace=FALSE),]

format(round(stat.desc(sample_num_sbaloan_data,basic = FALSE, norm = TRUE), digits = 3), scientific=FALSE)
```

```
##                 Term     NoEmp CreateJob        SBA_Appv         GrAppv
## median       84.000     4.000     0.000       60000.000       91450.000
## mean        110.689    14.537     5.549      150184.248      193511.687
## SE.mean       1.117     2.114     2.495        3172.212        3923.145
## CI.mean.0.95  2.190     4.144     4.891        6218.927        7691.085
## var        6239.676 22346.193 31124.306 50314637378.361 76955332679.311
## std.dev      78.992   149.486   176.421      224309.245      277408.242
## coef.var      0.714    10.283    31.791           1.494           1.434
## skewness      1.130    43.905    49.568           3.239           3.091
## skew.2SE     16.318   633.911   715.666          46.760          44.623
## kurtosis      0.217  2053.892  2467.074          16.232          13.273
## kurt.2SE      1.568 14830.098 17813.472         117.201          95.840
## normtest.W    0.841     0.033     0.009           0.643           0.648
## normtest.p    0.000     0.000     0.000           0.000           0.000
```

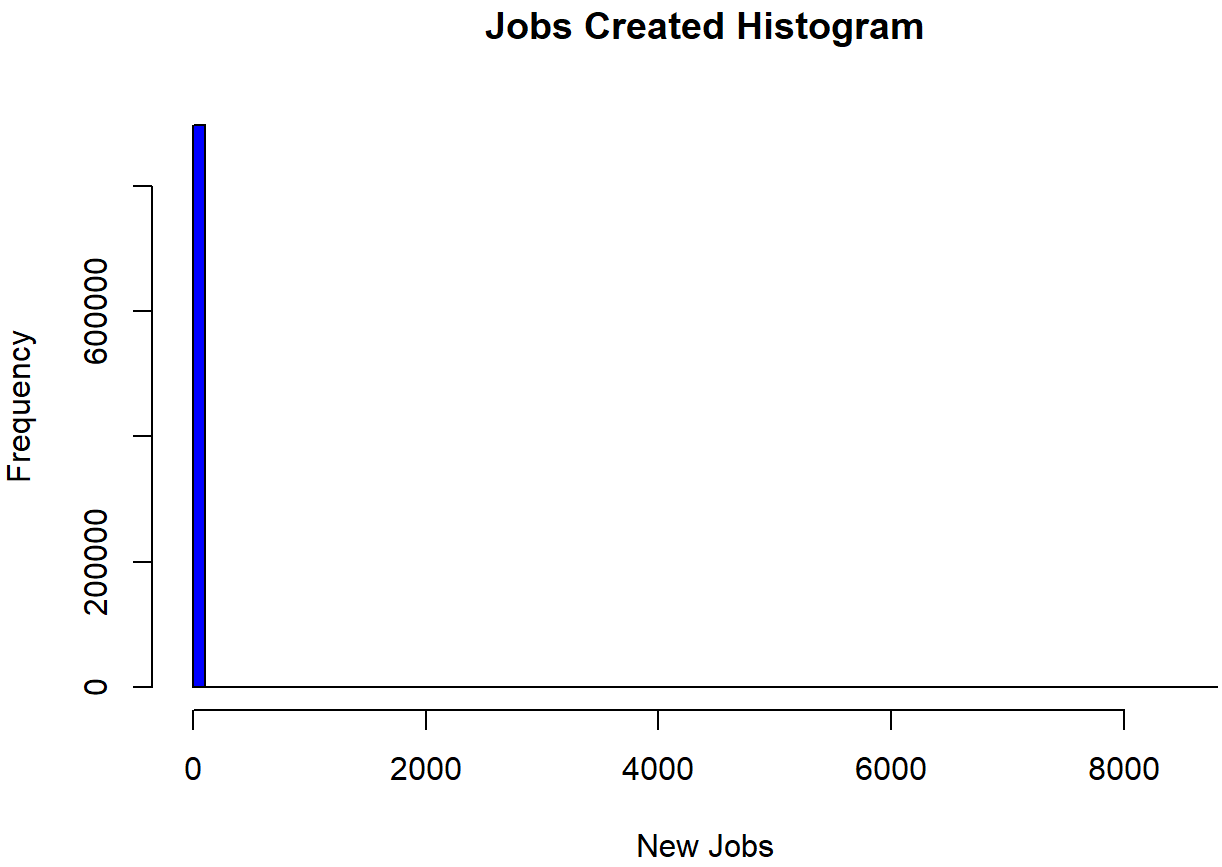
```
# Significant skew and kurtosis. Do not use models that assume a normal distribution.


# Histograms of numeric features
hist(sbaloan_data$Term,
                    main="Term",
                    xlab="Months",
                    col="blue")
```



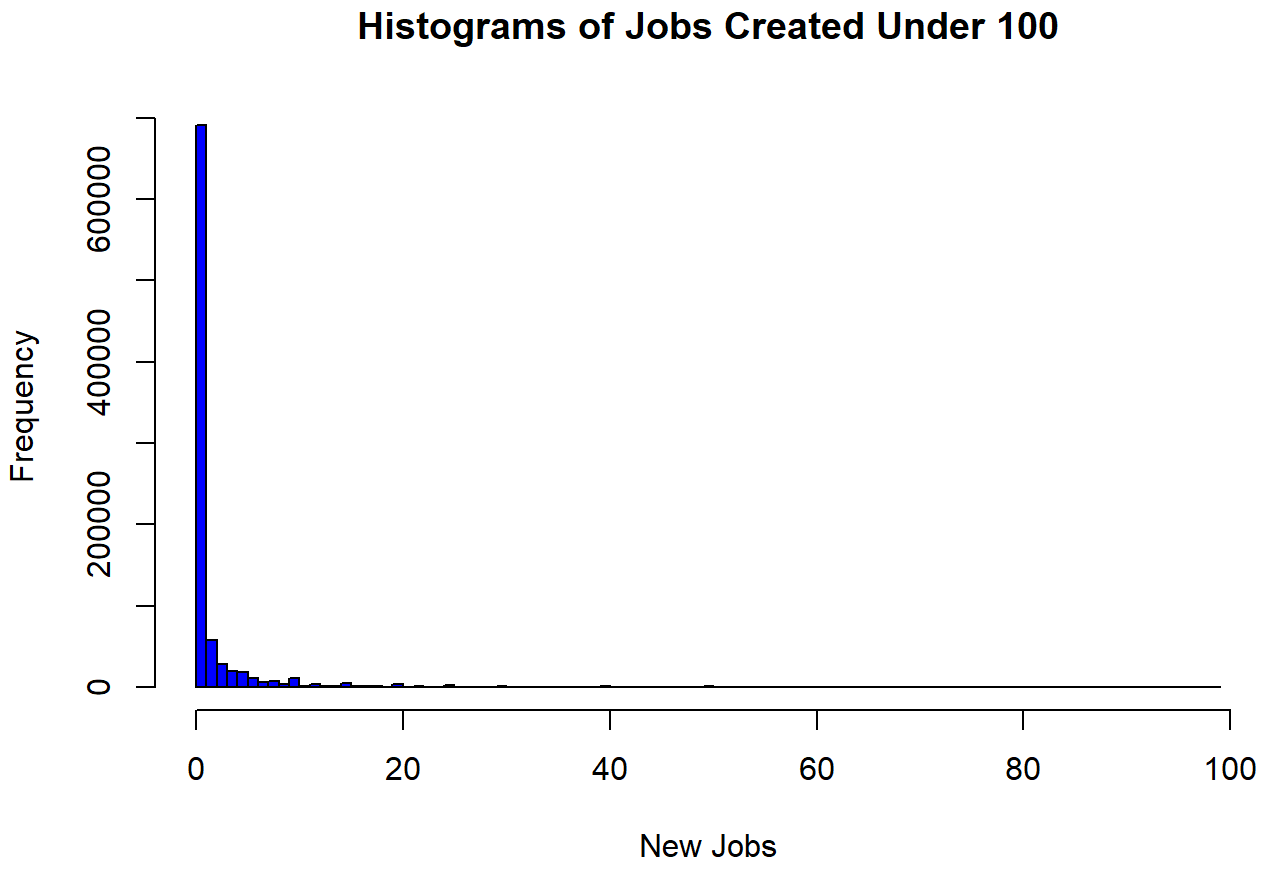

```
hist(sbaloan_data$NoEmp,
                   main="Employees",
                   xlab="Count of Employees Histogram",
                   col="blue")
```



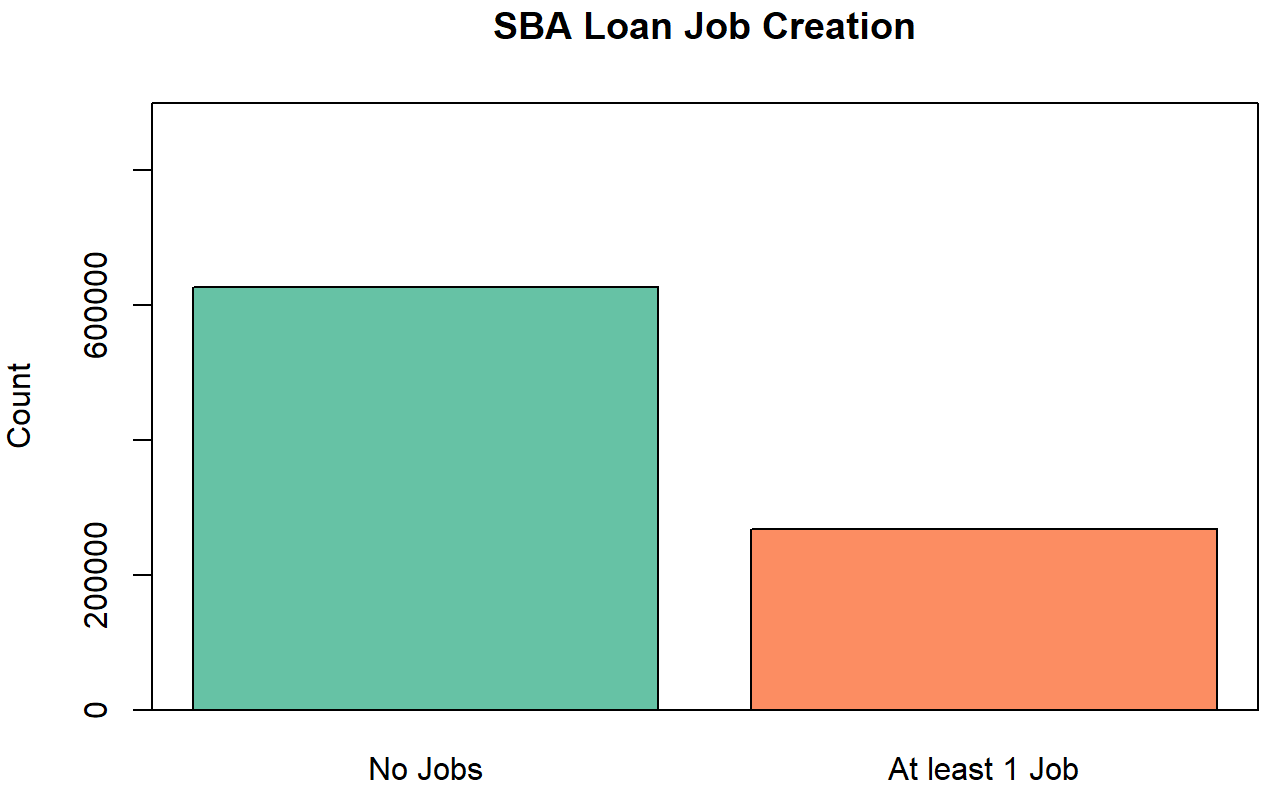

```
hist(sbaloan_data$CreateJob,
                   main="Jobs Created Histogram",
                   xlab="New Jobs",
                   breaks = 100,
                   col="blue")
```



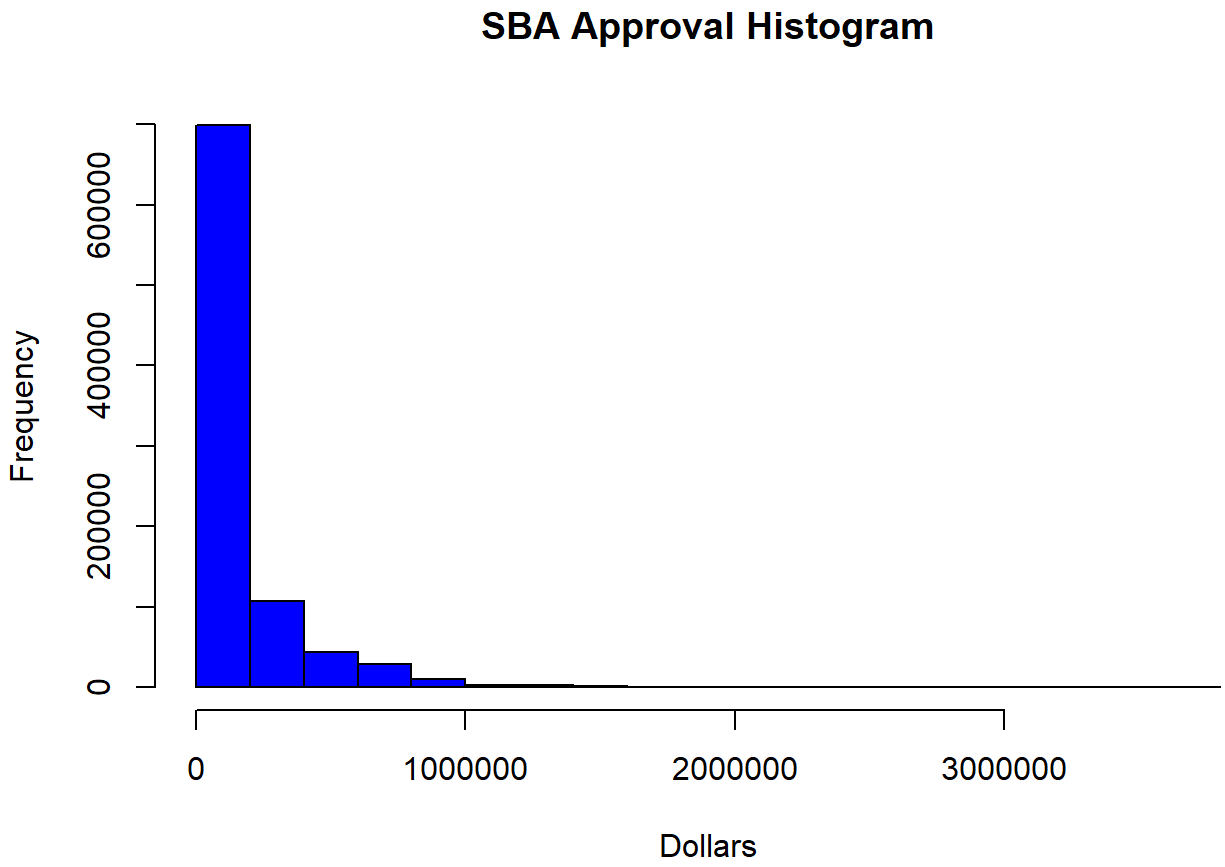

```
# Lets filter out the extremes and look
CreateJobAnalysis <- filter(sbaloan_data, `CreateJob` < 100)

hist(CreateJobAnalysis$CreateJob,
     main="Histograms of Jobs Created Under 100",
     xlab="New Jobs",
     breaks = 100,
     col="blue")
```



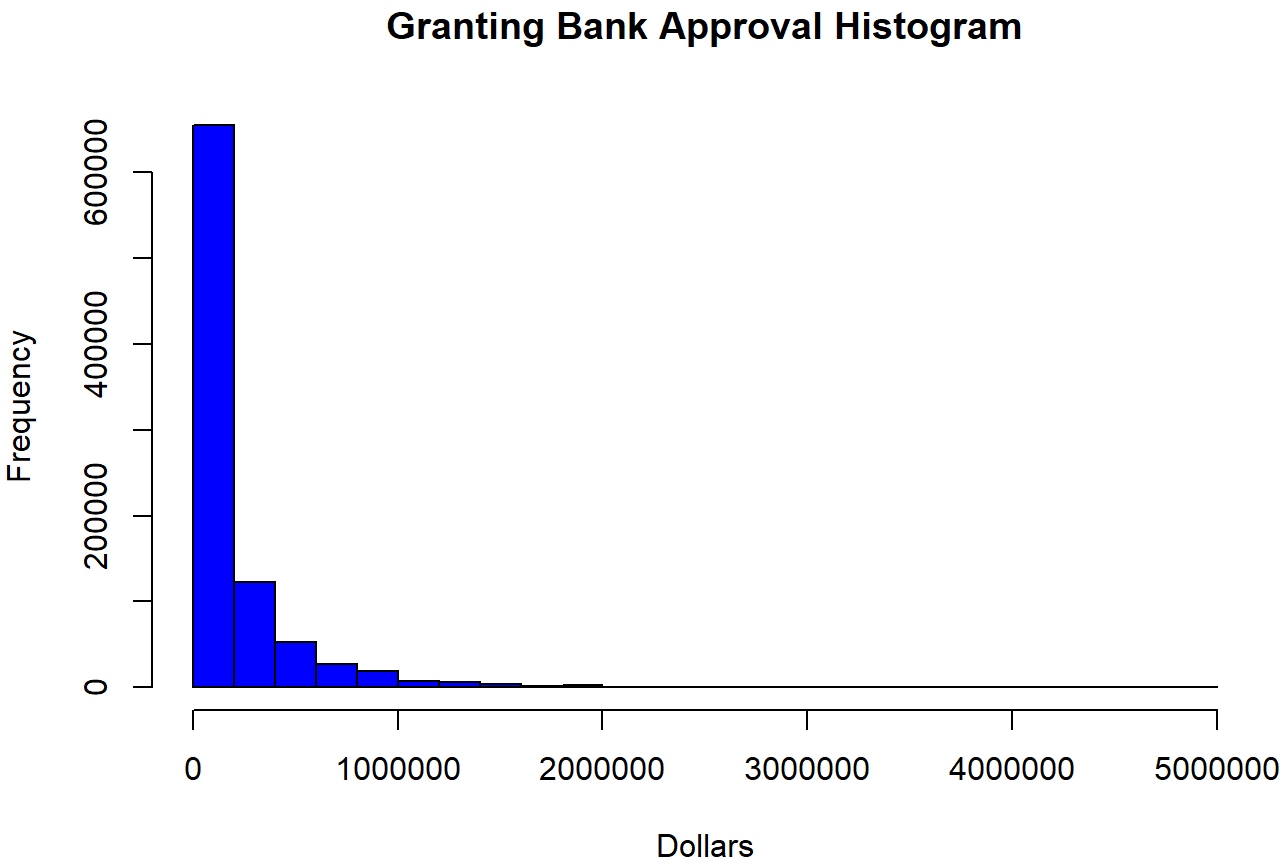

```
# Look at loans that created at least 1 job.
CreateJobAnalysis <- CreateJobAnalysis %>%
        mutate(CreateJob = ifelse( CreateJob == 0,0,1))

coul <- brewer.pal(5, "Set2")
barplot(table(CreateJobAnalysis$CreateJob),names.arg=c("No Jobs", "At least 1 Job"), col=coul, main="SBA Lo
an Job Creation", ylab="Count",ylim=c(0,900000))
box()
```



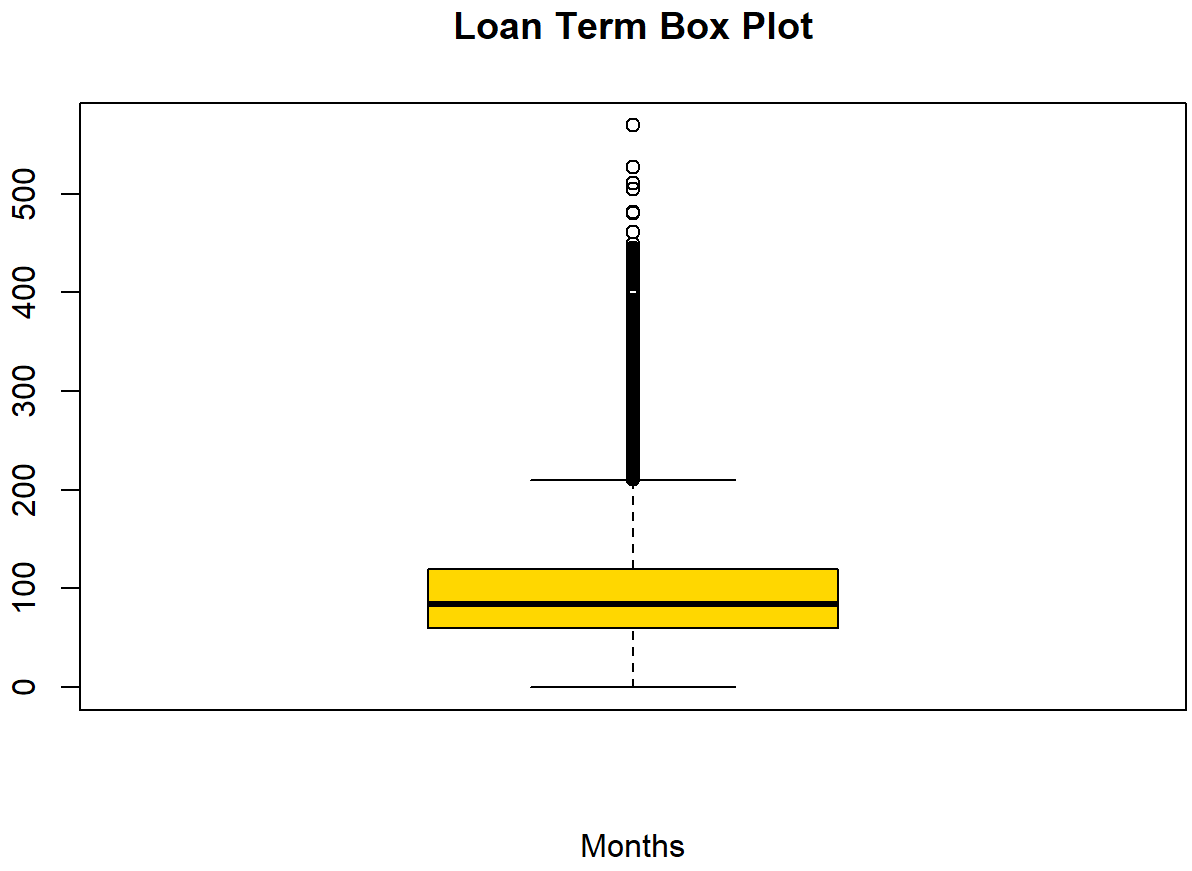

```
hist(sbaloan_data$SBA_Appv,
                   main="SBA Approval Histogram",
                   xlab="Dollars",
                   col="blue")
```



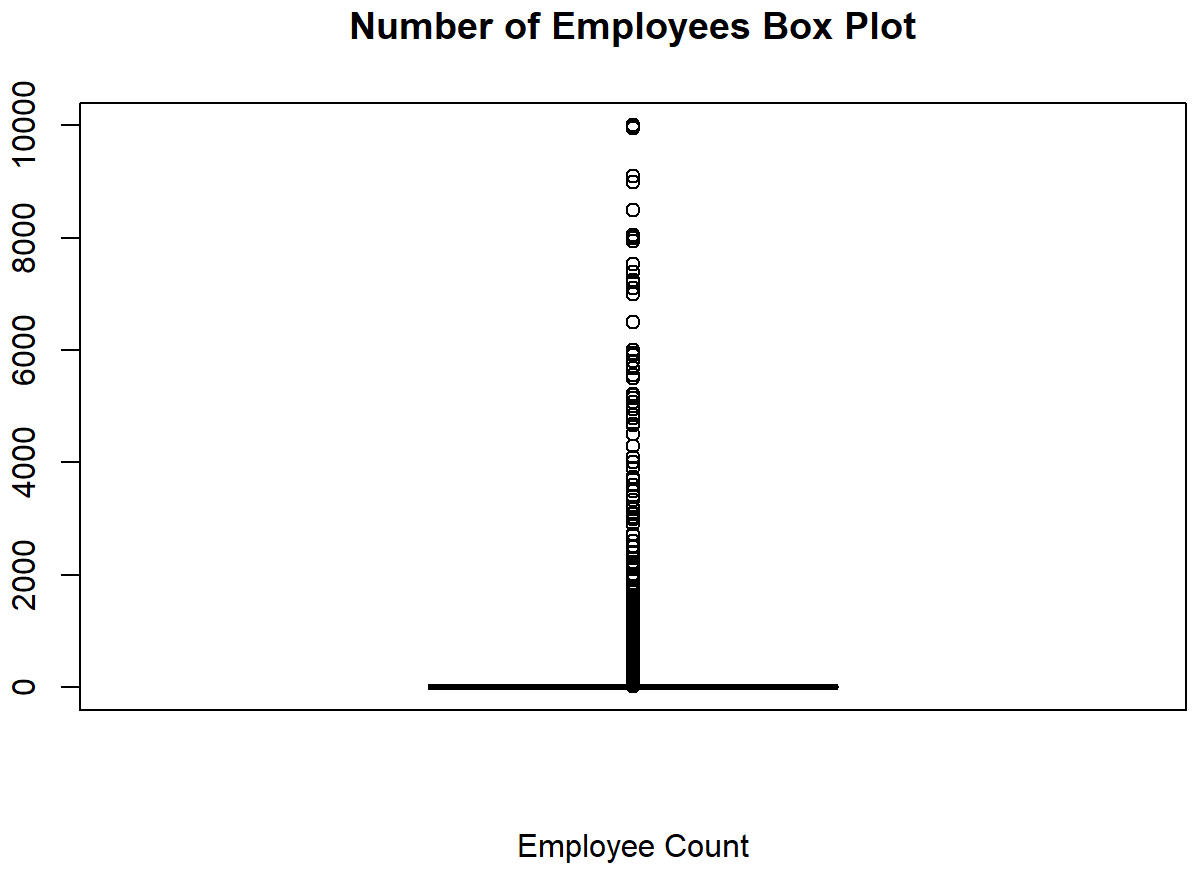

```
hist(sbaloan_data$GrAppv,
                   main="Granting Bank Approval Histogram",
                   xlab="Dollars",
                   col="blue")
```



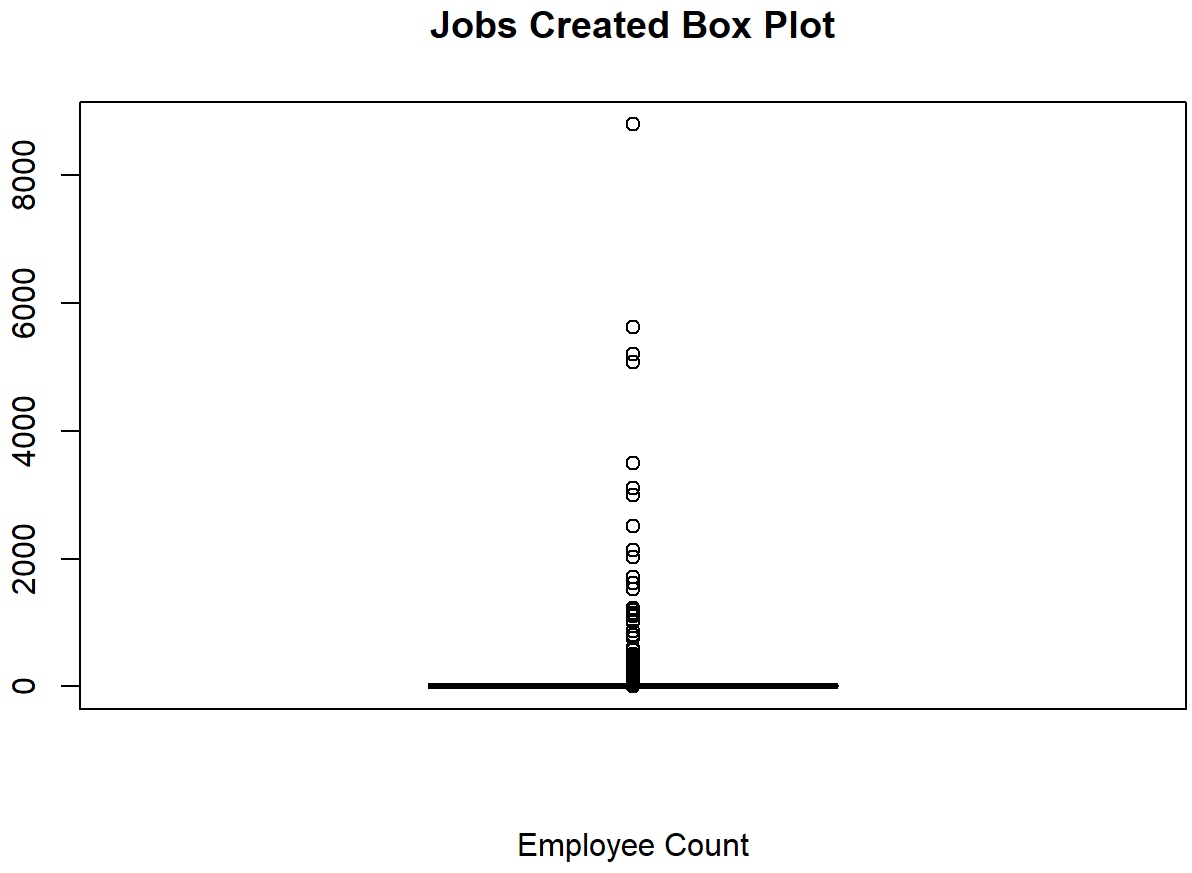

```
# Skew and kertosis confirmed by visuals



# Box plots of one variable
boxplot(sbaloan_data$Term,
        col=(c("gold","darkgreen")),
        main="Loan Term Box Plot", xlab="Months")
```



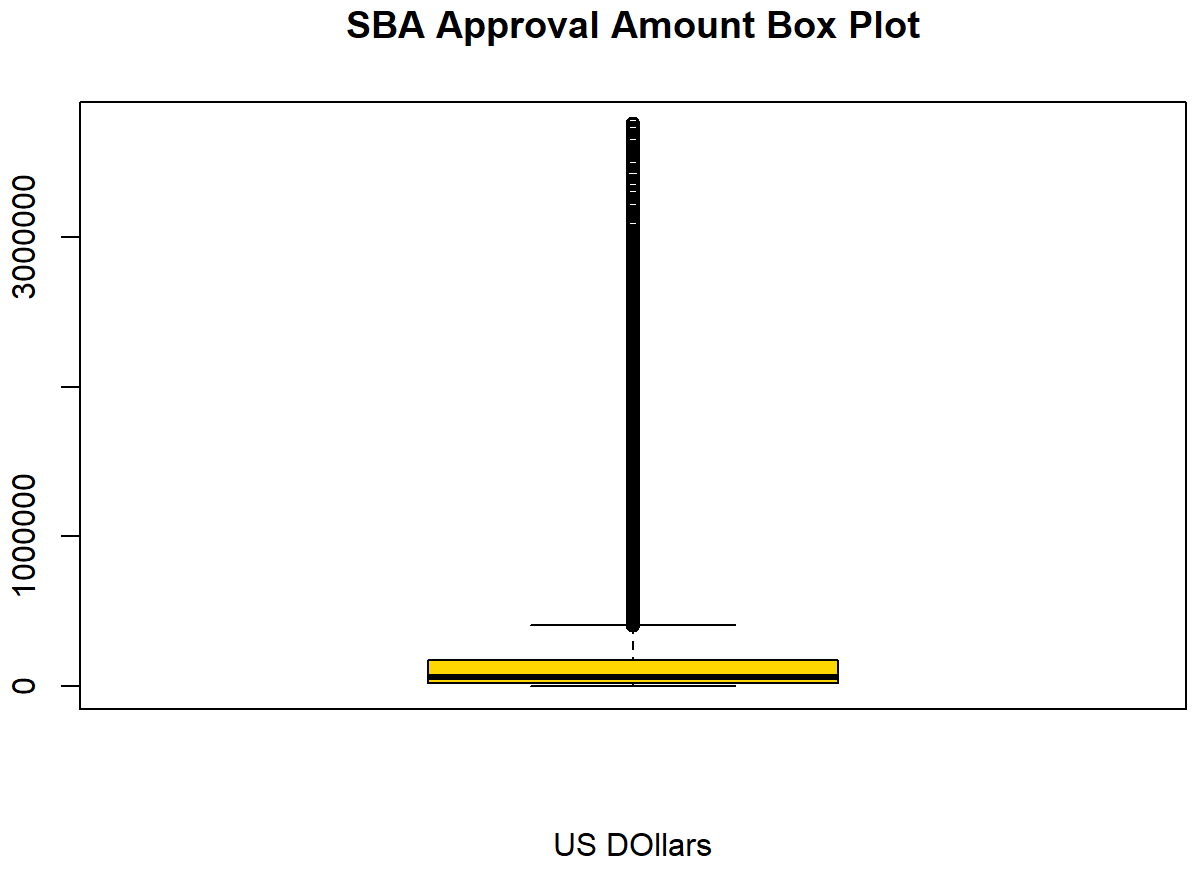

```
boxplot(sbaloan_data$NoEmp,
        col=(c("gold","darkgreen")),
        main="Number of Employees Box Plot", xlab="Employee Count"
        )
```



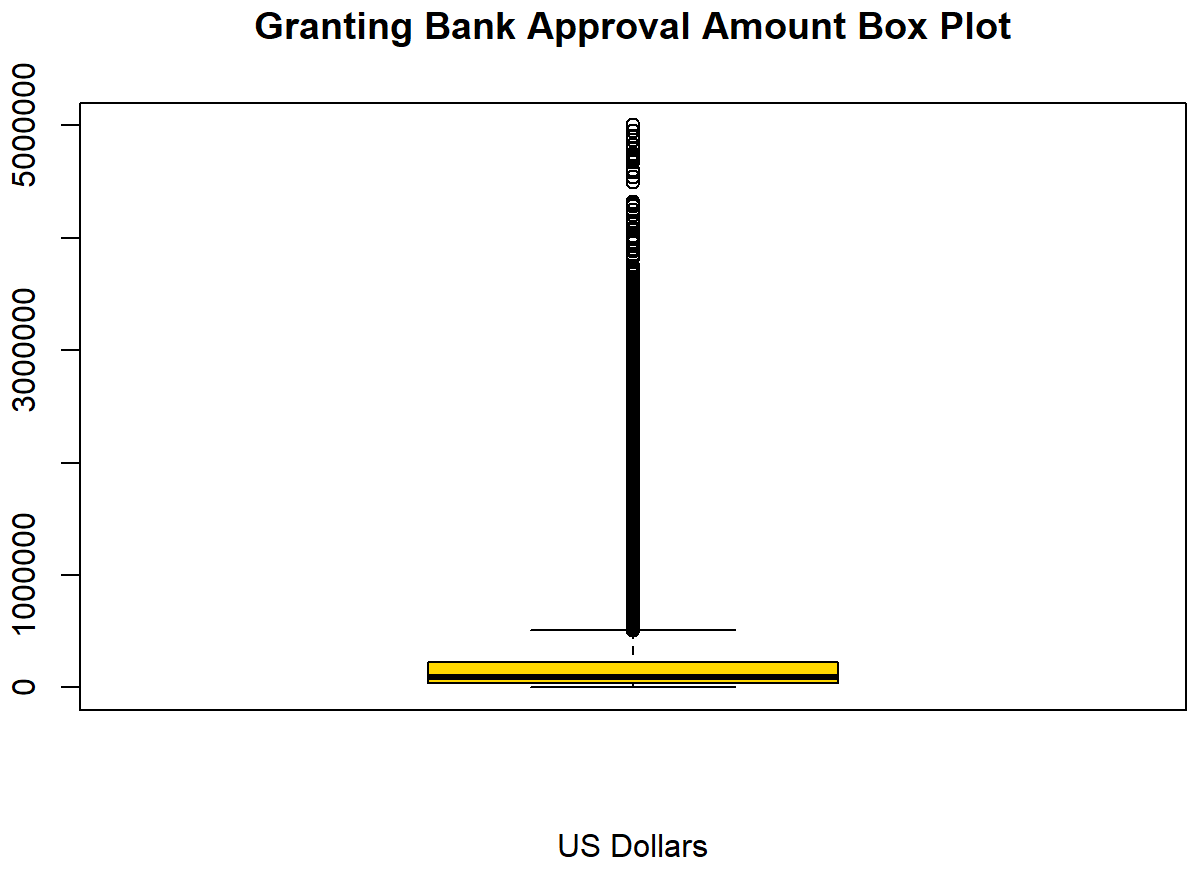

```
boxplot(sbaloan_data$CreateJob,
        col=(c("gold","darkgreen")),
        main="Jobs Created Box Plot", xlab="Employee Count")
```



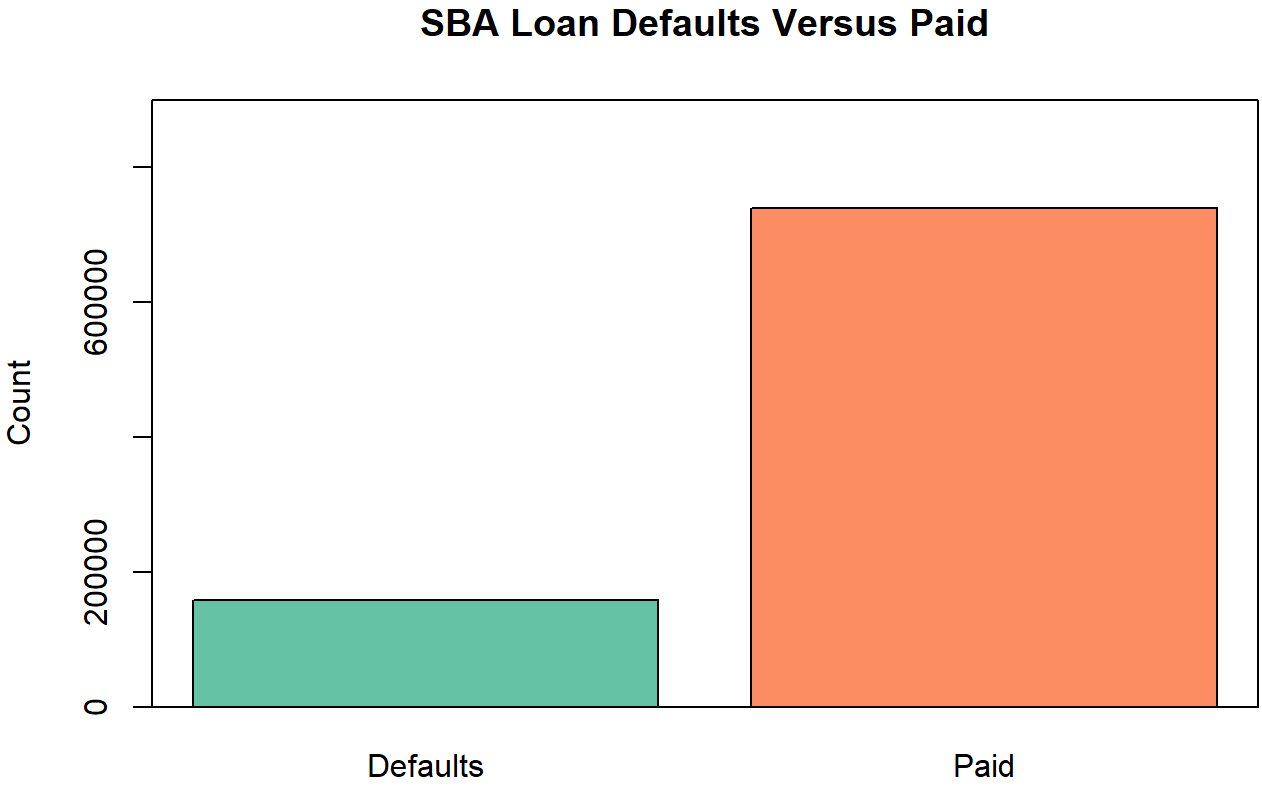
```
boxplot(sbaloan_data$SBA_Appv,
        col=(c("gold","darkgreen")),
        main="SBA Approval Amount Box Plot", xlab="US DOllars")
```




```
boxplot(sbaloan_data$GrAppv,
        col=(c("gold","darkgreen")),
        main="Granting Bank Approval Amount Box Plot", xlab="US Dollars")
```


Granting Bank Approval Amount Box Plot
0
1000000
3000000
5000000
US Dollars

```
# Data shows significant outliers but domain knowledge does not indicate justification for removal.
# We would not want to not account for some of our largest loans in the model.

# 1 Hot
encoded_sbaloan_data <- sbaloan_data %>% mutate(value = 1)  %>% spread(NewExist, value,  fill = 0 )
names(encoded_sbaloan_data)[names(encoded_sbaloan_data) == '0'] <- 'NewExist_0'
names(encoded_sbaloan_data)[names(encoded_sbaloan_data) == '1'] <- 'NewExist_1'
names(encoded_sbaloan_data)[names(encoded_sbaloan_data) == '2'] <- 'NewExist_2'
names(encoded_sbaloan_data)[names(encoded_sbaloan_data) == 'UNKNOWN'] <- 'NewExist_U'

encoded_sbaloan_data <- encoded_sbaloan_data %>% mutate(value = 1)  %>% spread(UrbanRural, value,  fill = 0
)
names(encoded_sbaloan_data)[names(encoded_sbaloan_data) == '0'] <- 'UrbanRural_0'
names(encoded_sbaloan_data)[names(encoded_sbaloan_data) == '1'] <- 'UrbanRural_1'
names(encoded_sbaloan_data)[names(encoded_sbaloan_data) == '2'] <- 'UrbanRural_2'

encoded_sbaloan_data <- encoded_sbaloan_data %>% mutate(value = 1)  %>% spread(Month, value,  fill = 0 )
names(encoded_sbaloan_data)[names(encoded_sbaloan_data) == 'Jan'] <- 'Month_Jan'
names(encoded_sbaloan_data)[names(encoded_sbaloan_data) == 'Feb'] <- 'Month_Feb'
names(encoded_sbaloan_data)[names(encoded_sbaloan_data) == 'Mar'] <- 'Month_Mar'
names(encoded_sbaloan_data)[names(encoded_sbaloan_data) == 'Apr'] <- 'Month_Apr'
names(encoded_sbaloan_data)[names(encoded_sbaloan_data) == 'May'] <- 'Month_May'
names(encoded_sbaloan_data)[names(encoded_sbaloan_data) == 'Jun'] <- 'Month_Jun'
names(encoded_sbaloan_data)[names(encoded_sbaloan_data) == 'Jul'] <- 'Month_Jul'
names(encoded_sbaloan_data)[names(encoded_sbaloan_data) == 'Aug'] <- 'Month_Aug'
names(encoded_sbaloan_data)[names(encoded_sbaloan_data) == 'Sep'] <- 'Month_Sep'
names(encoded_sbaloan_data)[names(encoded_sbaloan_data) == 'Oct'] <- 'Month_Oct'
names(encoded_sbaloan_data)[names(encoded_sbaloan_data) == 'Nov'] <- 'Month_Nov'
names(encoded_sbaloan_data)[names(encoded_sbaloan_data) == 'Dec'] <- 'Month_Dec'


# Create a new field that changes these to binary
encoded_sbaloan_data <- encoded_sbaloan_data %>%
        mutate(MIS_logical = ifelse( MIS_Status == "P I F",1,0))

# Bar Graph of Defaults to Paid

coul <- brewer.pal(5, "Set2")
barplot(table(encoded_sbaloan_data$MIS_logical),names.arg=c("Defaults", "Paid"), col=coul, main="SBA Loan D
efaults Versus Paid", ylab="Count",ylim=c(0,900000))
box()
```




```
summary(encoded_sbaloan_data$MIS_logical)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  1.0000  1.0000  0.8244  1.0000  1.0000
```

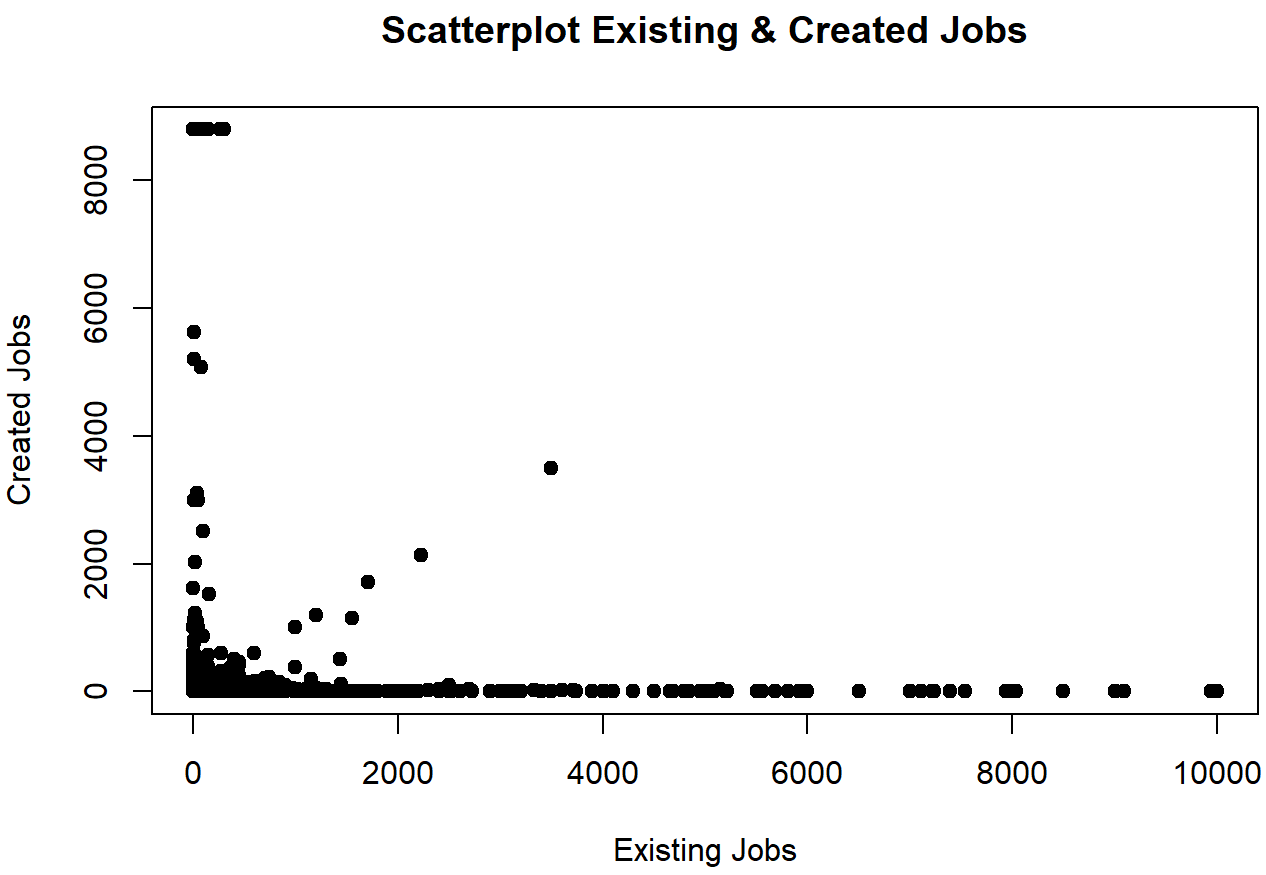
```
# Label Encoding
sbaloan_data$Bank <- as.integer(sbaloan_data$Bank)
# sbaloan_data$Bank


# Binary Encoding when too many values
# This code segment was taken from https://www.r-bloggers.com/2020/02/a-guide-to-encoding-categorical-features-using-r/
encode_binary <- function(x, order = unique(x), name = "v_") {
        x <- as.numeric(factor(x, levels = order, exclude = NULL))
        x2 <- as.binary(x)
        maxlen <- max(sapply(x2, length))
        x2 <- lapply(x2, function(y) {
                l <- length(y)
                if (l < maxlen) {
                        y <- c(rep(0, (maxlen - l)), y)
                }
                y
        })
        d <- as.data.frame(t(as.data.frame(x2)))
        rownames(d) <- NULL
        colnames(d) <- paste0(name, 1:maxlen)
        d
}

# Binary Encode calling function above
encoded_sbaloan_data <- cbind(encoded_sbaloan_data, encode_binary(encoded_sbaloan_data[["RevLineCr"]], name = "RevLineCr_"))
encoded_sbaloan_data <- cbind(encoded_sbaloan_data, encode_binary(encoded_sbaloan_data[["FranchiseCode"]], name = "FranchiseCode_"))
encoded_sbaloan_data <- cbind(encoded_sbaloan_data, encode_binary(encoded_sbaloan_data[["LowDoc"]], name = "LowDoc_"))
encoded_sbaloan_data <- cbind(encoded_sbaloan_data, encode_binary(encoded_sbaloan_data[["NAICS"]], name = "NAICS_"))
encoded_sbaloan_data <- cbind(encoded_sbaloan_data, encode_binary(encoded_sbaloan_data[["Bank"]], name = "Bank_"))

# Weak correlation and likely illegal for loan consideration so I am removing
# encoded_sbaloan_data <- cbind(encoded_sbaloan_data, encode_binary(encoded_sbaloan_data[["BankState"]], name = "BankState_"))
# encoded_sbaloan_data <- cbind(encoded_sbaloan_data, encode_binary(encoded_sbaloan_data[["Zip"]], name = "Zip_"))

# Remove columns
encoded_sbaloan_data <- subset(encoded_sbaloan_data, select = -c(`LoanNr_ChkDgt`,`Zip`,`NAICS`,`FranchiseCode`,`RevLineCr`,`LowDoc`))


# Remove origial column
encoded_sbaloan_data <- subset(encoded_sbaloan_data, select = -c(`MIS_Status`))

# Remove Bank since it was enocded and Day as it was just a test and didn't help in later processing
encoded_sbaloan_data <- subset(encoded_sbaloan_data, select = -c(`Bank`,`Day`))
# encoded_sbaloan_data <- subset(encoded_sbaloan_data, select = -c(`Month`,`NewExist`))
# glimpse(encoded_sbaloan_data)
summary(encoded_sbaloan_data)
```

```
##    ApprovalFY         Term           NoEmp           CreateJob
##  Min.   : 3.00   Min.   :  0.0   Min.   :   0.00   Min.   :   0.000
##  1st Qu.:35.00   1st Qu.: 60.0   1st Qu.:   2.00   1st Qu.:   0.000
##  Median :40.00   Median : 84.0   Median :   4.00   Median :   0.000
##  Mean   :39.14   Mean   :110.8   Mean   :  11.41   Mean   :   8.444
##  3rd Qu.:44.00   3rd Qu.:120.0   3rd Qu.:  10.00   3rd Qu.:   1.000
##  Max.   :52.00   Max.   :569.0   Max.   :9999.00   Max.   :8800.000
##      GrAppv           SBA_Appv         NewExist_0         NewExist_1
##  Min.   :   1000   Min.   :    500   Min.   :0.000000   Min.   :0.0000
##  1st Qu.:  35000   1st Qu.:  21250   1st Qu.:0.000000   1st Qu.:0.0000
##  Median :  90000   Median :  62050   Median :0.000000   Median :1.0000
##  Mean   : 192780   Mean   : 149519   Mean   :0.001146   Mean   :0.7172
##  3rd Qu.: 225000   3rd Qu.: 175000   3rd Qu.:0.000000   3rd Qu.:1.0000
##  Max.   :5000000   Max.   :3750000   Max.   :1.000000   Max.   :1.0000
##    NewExist_2       NewExist_U         UrbanRural_0     UrbanRural_1
##  Min.   :0.0000   Min.   :0.0000000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.0000000   1st Qu.:0.0000   1st Qu.:0.000
##  Median :0.0000   Median :0.0000000   Median :0.0000   Median :1.000
##  Mean   :0.2815   Mean   :0.0001471   Mean   :0.3599   Mean   :0.523
##  3rd Qu.:1.0000   3rd Qu.:0.0000000   3rd Qu.:1.0000   3rd Qu.:1.000
##  Max.   :1.0000   Max.   :1.0000000   Max.   :1.0000   Max.   :1.000
##   UrbanRural_2      Month_Apr        Month_Aug         Month_Dec
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.0000   Median :0.0000   Median :0.00000   Median :0.00000
##  Mean   :0.1171   Mean   :0.0892   Mean   :0.08762   Mean   :0.07777
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##    Month_Feb         Month_Jan         Month_Jul         Month_Jun
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.07376   Mean   :0.07459   Mean   :0.08509   Mean   :0.08711
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##    Month_Mar         Month_May         Month_Nov         Month_Oct
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.0000
##  Mean   :0.09299   Mean   :0.08589   Mean   :0.07608   Mean   :0.0776
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##    Month_Sep        MIS_logical      RevLineCr_1        RevLineCr_2
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.000000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:0.000000   1st Qu.:0.00000
##  Median :0.00000   Median :1.0000   Median :0.000000   Median :0.00000
##  Mean   :0.09231   Mean   :0.8244   Mean   :0.000029   Mean   :0.02207
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.000000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.000000   Max.   :1.00000
##   RevLineCr_3      RevLineCr_4    FranchiseCode_1    FranchiseCode_2
##  Min.   :0.0000   Min.   :0.000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :1.0000   Median :1.000   Median :0.000000   Median :0.000000
##  Mean   :0.5107   Mean   :0.696   Mean   :0.001716   Mean   :0.005145
##  3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.0000   Max.   :1.000   Max.   :1.000000   Max.   :1.000000
##  FranchiseCode_3   FranchiseCode_4   FranchiseCode_5   FranchiseCode_6
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
```

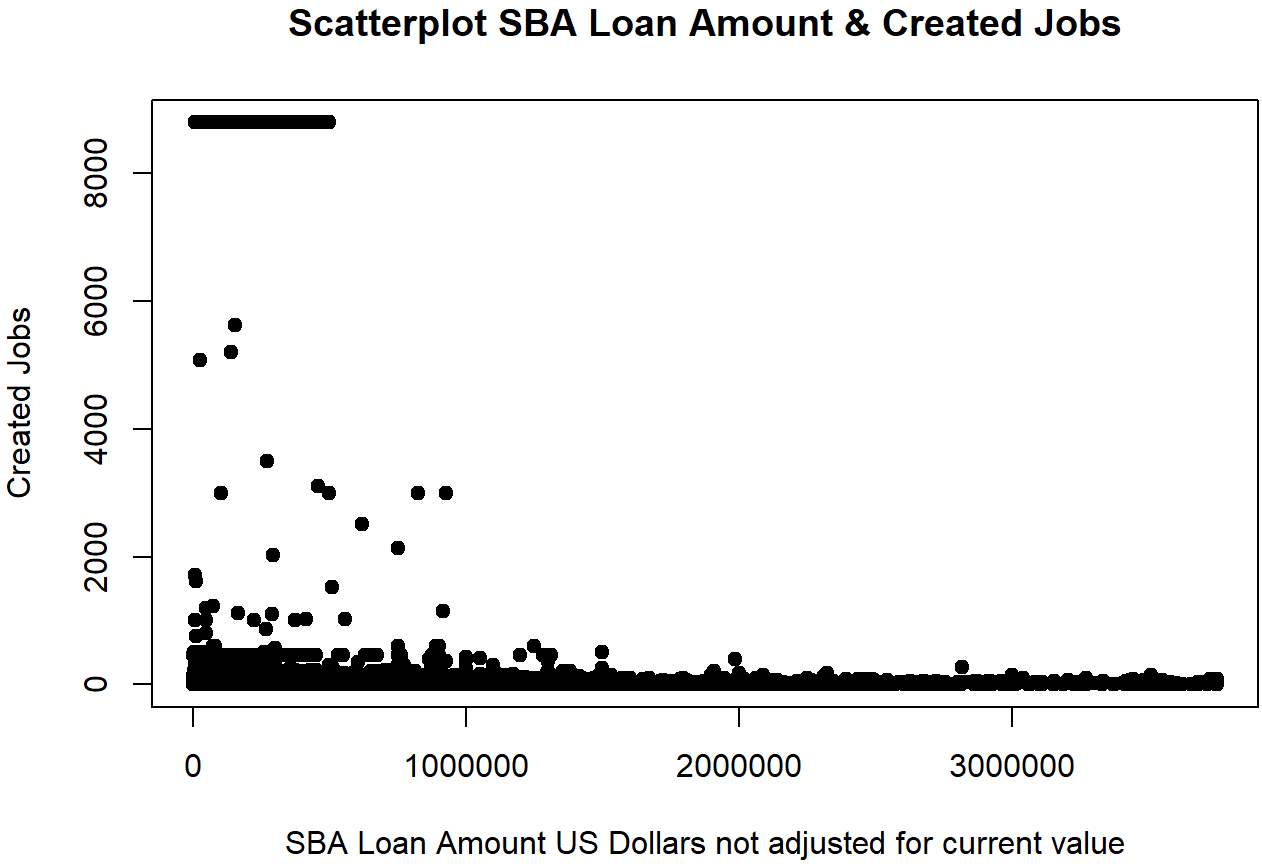
```
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.01001   Mean   :0.01568   Mean   :0.01845   Mean   :0.02627
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##  FranchiseCode_7   FranchiseCode_8   FranchiseCode_9   FranchiseCode_10
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.02538   Mean   :0.03061   Mean   :0.02547   Mean   :0.02789
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##  FranchiseCode_11 FranchiseCode_12    LowDoc_1          LowDoc_2
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.0000   Median :1.0000   Median :0.000000   Median :0.000000
##  Mean   :0.2591   Mean   :0.7405   Mean   :0.002213   Mean   :0.003629
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.000000   Max.   :1.000000
##     LowDoc_3         LowDoc_4         NAICS_1           NAICS_2
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000000   Min.   :0.00000
##  1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:0.00000
##  Median :1.0000   Median :0.0000   Median :0.000000   Median :0.00000
##  Mean   :0.8721   Mean   :0.1283   Mean   :0.009581   Mean   :0.07659
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.000000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.000000   Max.   :1.00000
##     NAICS_3          NAICS_4          NAICS_5          NAICS_6
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.1853   Mean   :0.2745   Mean   :0.3007   Mean   :0.3595
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##     NAICS_7          NAICS_8          NAICS_9          NAICS_10
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :1.0000   Median :0.0000
##  Mean   :0.3927   Mean   :0.3908   Mean   :0.6294   Mean   :0.3923
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##     NAICS_11         Bank_1            Bank_2            Bank_3
##  Min.   :0.000   Min.   :0.000000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.000   1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.000   Median :0.000000   Median :0.00000   Median :0.00000
##  Mean   :0.341   Mean   :0.007908   Mean   :0.04277   Mean   :0.07864
##  3rd Qu.:1.000   3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.000   Max.   :1.000000   Max.   :1.00000   Max.   :1.00000
##      Bank_4           Bank_5           Bank_6           Bank_7
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.1407   Mean   :0.1674   Mean   :0.2415   Mean   :0.2897
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      Bank_8           Bank_9           Bank_10          Bank_11
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :1.0000
```

```
##  Mean   :0.4432   Mean   :0.3277   Mean   :0.4854   Mean   :0.5921
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##     Bank_12          Bank_13
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :0.0000
##  Mean   :0.5339   Mean   :0.4879
##  3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000
```

```
# Plot did not exhibit strong linear correlation
plot(encoded_sbaloan_data$NoEmp, encoded_sbaloan_data$CreateJob, main="Scatterplot Existing & Created Jobs"
,
     xlab="Existing Jobs", ylab="Created Jobs", pch=19)
```



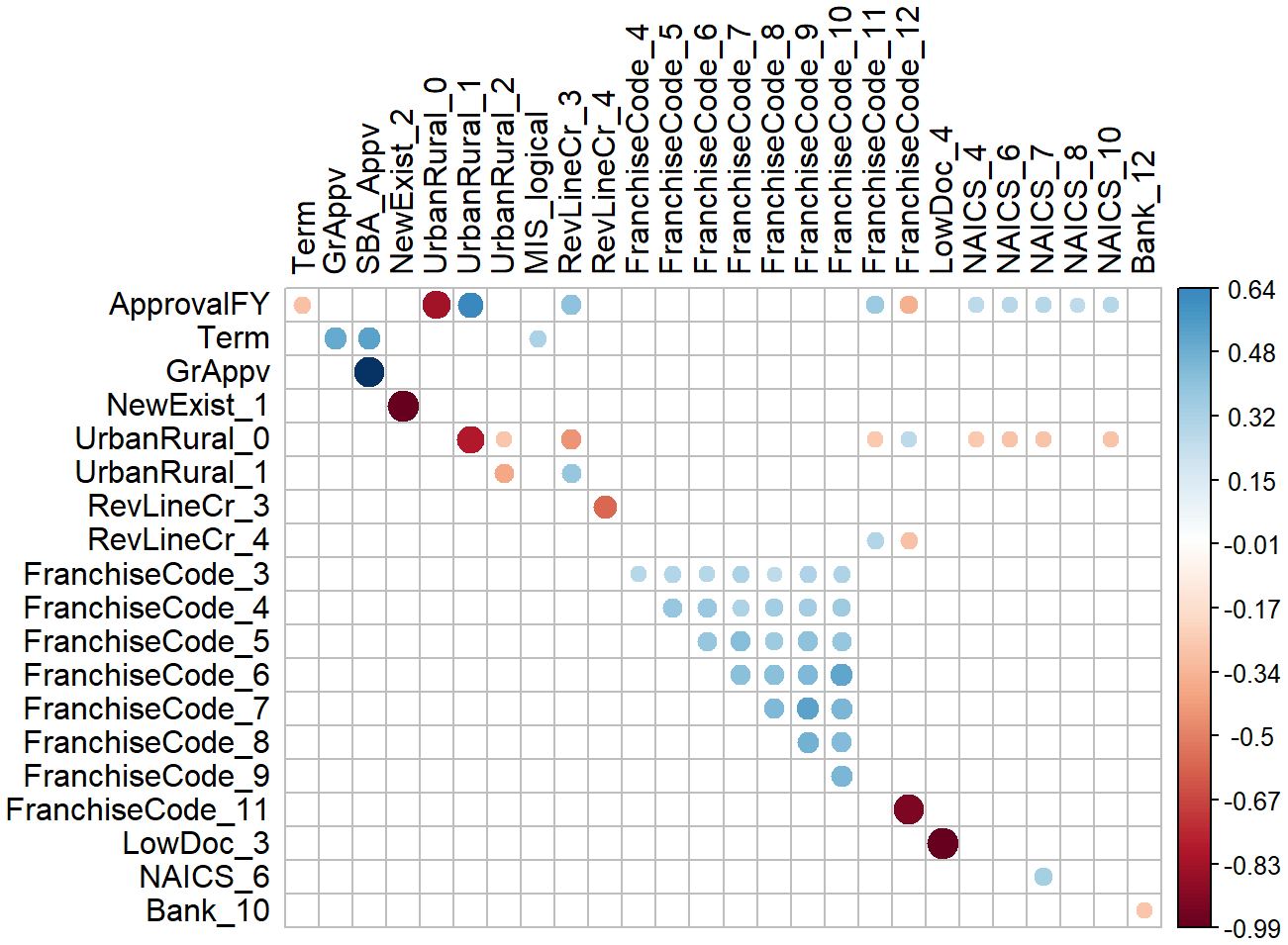

```
# Plot did not exhibit strong linear correlation
plot(encoded_sbaloan_data$SBA_Appv, encoded_sbaloan_data$CreateJob, main="Scatterplot SBA Loan Amount & Cre
ated Jobs",
     xlab="SBA Loan Amount US Dollars not adjusted for current value", ylab="Created Jobs", pch=19)
```


Scatterplot SBA Loan Amount & Created Jobs
Created Jobs
0
2000
4000
6000
8000
0
1000000
2000000
3000000
SBA Loan Amount US Dollars not adjusted for current value

```r
# Spearman was selected for correlation analysis due to non-parametric data.
cor_matrix <- cor(encoded_sbaloan_data,method="spearman")

# Commented out for length will summarize below
# cor_matrix
# eROUT_CIR <- corrplot(cor_matrix, method = "circle", title = "Circle Correlation Matrix Heat Map", mar=c
(0,0,1,0), tl.cex=.75)

# Too much to look at, I want a quick scan of relevant values
# The following code does that

# Code modified slightly from
# https://towardsdatascience.com/how-to-create-a-correlation-matrix-with-too-many-variables-309cc0c0a57
# Look at highest values
corr_simple <- function(data=encoded_sbaloan_data,sig=0.25){
        #convert data to numeric in order to run correlations
        #convert to factor first to keep the integrity of the data - each value will
        #become a number rather than turn into NA
        df_cor <- data %>% mutate_if(is.character, as.factor)
        df_cor <- df_cor %>% mutate_if(is.factor, as.numeric)
        #run a correlation and drop the insignificant ones
        corr <- cor(df_cor)
        #prepare to drop duplicates and correlations of 1
        corr[lower.tri(corr,diag=TRUE)] <- NA
        #drop perfect correlations
        corr[corr == 1] <- NA
        #turn into a 3-column table
        corr <- as.data.frame(as.table(corr))
        #remove the NA values from above
        corr <- na.omit(corr)
        #select significant values
        corr <- subset(corr, abs(Freq) > sig)
        #sort by highest correlation
        corr <- corr[order(-abs(corr$Freq)),]
        #print table
        print(corr)
        #turn corr back into matrix in order to plot with corrplot
        mtx_corr <- reshape2::acast(corr, Var1~Var2, value.var="Freq")

        #plot correlations visually
        corrplot(mtx_corr, is.corr=FALSE, tl.col="black", na.label=" ")
}
corr_simple()
```

```
##                 Var1            Var2       Freq
## 568        NewExist_1       NewExist_2 -0.9968132
## 3195         LowDoc_3         LowDoc_4 -0.9933737
## 355            GrAppv         SBA_Appv  0.9740971
## 2911 FranchiseCode_11 FranchiseCode_12 -0.9240874
## 701         ApprovalFY     UrbanRural_0 -0.8321875
## 781       UrbanRural_0     UrbanRural_1 -0.7851538
## 771         ApprovalFY     UrbanRural_1  0.6438458
## 2059      RevLineCr_3      RevLineCr_4 -0.5730451
## 352              Term         SBA_Appv  0.5295493
## 2697  FranchiseCode_7  FranchiseCode_9  0.5281540
## 2766  FranchiseCode_6 FranchiseCode_10  0.5176347
## 282              Term           GrAppv  0.5057383
## 2698  FranchiseCode_8  FranchiseCode_9  0.4680634
## 2767  FranchiseCode_7 FranchiseCode_10  0.4514295
## 2769  FranchiseCode_9 FranchiseCode_10  0.4500151
## 1971     UrbanRural_0      RevLineCr_3 -0.4430693
## 2627  FranchiseCode_7  FranchiseCode_8  0.4409770
## 2696  FranchiseCode_6  FranchiseCode_9  0.4406931
## 2768  FranchiseCode_8 FranchiseCode_10  0.4277323
## 2555  FranchiseCode_5  FranchiseCode_7  0.4238976
## 2556  FranchiseCode_6  FranchiseCode_7  0.4139644
## 2626  FranchiseCode_6  FranchiseCode_8  0.4132253
## 2695  FranchiseCode_5  FranchiseCode_9  0.4061890
## 1961       ApprovalFY      RevLineCr_3  0.4041849
## 2414  FranchiseCode_4  FranchiseCode_5  0.3860626
## 2765  FranchiseCode_5 FranchiseCode_10  0.3846901
## 852      UrbanRural_1     UrbanRural_2 -0.3813796
## 2485  FranchiseCode_5  FranchiseCode_6  0.3794746
## 1972     UrbanRural_1      RevLineCr_3  0.3777353
## 2484  FranchiseCode_4  FranchiseCode_6  0.3687704
## 2801       ApprovalFY FranchiseCode_11  0.3598823
## 2625  FranchiseCode_5  FranchiseCode_8  0.3564138
## 2871       ApprovalFY FranchiseCode_12 -0.3560775
## 2764  FranchiseCode_4 FranchiseCode_10  0.3542669
## 2694  FranchiseCode_4  FranchiseCode_9  0.3458109
## 2624  FranchiseCode_4  FranchiseCode_8  0.3435690
## 3692          NAICS_6          NAICS_7  0.3352409
## 2553  FranchiseCode_3  FranchiseCode_7  0.3166439
## 1752             Term      MIS_logical  0.3141285
## 2554  FranchiseCode_4  FranchiseCode_7  0.3022303
## 2693  FranchiseCode_3  FranchiseCode_9  0.2993248
## 2763  FranchiseCode_3 FranchiseCode_10  0.2983977
## 2413  FranchiseCode_3  FranchiseCode_5  0.2978764
## 2900      RevLineCr_4 FranchiseCode_12 -0.2942971
## 71          ApprovalFY             Term -0.2927300
## 2830      RevLineCr_4 FranchiseCode_11  0.2915160
## 3651     UrbanRural_0          NAICS_7 -0.2871106
## 2343  FranchiseCode_3  FranchiseCode_4  0.2868328
## 3641       ApprovalFY          NAICS_7  0.2860179
## 2483  FranchiseCode_3  FranchiseCode_6  0.2839049
## 3851       ApprovalFY         NAICS_10  0.2833578
## 3861     UrbanRural_0         NAICS_10 -0.2798738
## 3581     UrbanRural_0          NAICS_6 -0.2787678
## 4827          Bank_10          Bank_12 -0.2780266
## 3571       ApprovalFY          NAICS_6  0.2775135
## 851       UrbanRural_0     UrbanRural_2 -0.2730515
## 2811     UrbanRural_0 FranchiseCode_11 -0.2649130
```

```
## 3441     UrbanRural_0          NAICS_4 -0.2612105
## 2881     UrbanRural_0 FranchiseCode_12  0.2603173
## 3431       ApprovalFY          NAICS_4  0.2593809
## 2623  FranchiseCode_3  FranchiseCode_8  0.2515352
## 3711       ApprovalFY          NAICS_8  0.2510739
```




```
# Term had the strongest correlation to loan repayment
# None were noticeable for Jobs Created
# Nothing significant for either



# Write out for python
write.csv(encoded_sbaloan_data,"sbapython.csv", row.names = FALSE)

summary(sbaloan_data$CreateJob)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.000    0.000    0.000    8.444    1.000 8800.000
```