Project 1

Group 10

10/24/2021

Title: SBA Loan Analysis: Should the Loan be Approved?

Overview:

The U.S. Small Business Administration (SBA) was founded in 1953. Small businesses have been a primary source of job creation in the US. SBA assists these small businesses through a loan guarantee program which is designed to encourage banks to grant loans to small businesses. SBA reduces the risk for a bank by guaranteeing a portion of the loan. So, if in case a loan goes into default, SBA then covers the amount they guaranteed. There have been many success stories of start-ups receiving SBA loan guarantees and there have also been stories of small businesses and/or start-ups that have defaulted on their SBA-guaranteed loans. The rate of default on these loans has been a source of controversy for decades. Therefore, banks are still faced with difficulty if they should grant such a loan because of the high risk of default involved. One way to inform their decision-making is through analyzing relevant historical data such as the datasets and then classifying the loan into higher risk or lower risks to off the loan. Consequently, through this project, we would like to use the Data Wrangling Process for easy access to analyze and address how the real-world data affects the Loan Approvals/ Disbursements and the Fraud factors related to it. From this project we hope to accomplish how the Data wrangling can be used to analyse the operations of SBA. We'll be addressing a few business questions to anaylse information for Loan Approval in various situations.

Data Acquisition

Libraries

The libraries are attached using the library() function.

```
##
## 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

##
## Attaching package: 'magrittr'
```

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

##
## Attaching package: 'lubridate'

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

##
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:lemon':
##
##     CoordCartesian, element_render

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine

## Linking to GEOS 3.8.1, GDAL 3.2.1, PROJ 7.2.1

##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble 3.1.4     v purrr  0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x purrr::%||%()             masks lemon::%||%()
## x lubridate::as.difftime()  masks base::as.difftime()
## x gridExtra::combine()      masks dplyr::combine()
```

```
## x lubridate::date()          masks base::date()
## x ggplot2::element_render() masks lemon::element_render()
## x magrittr::extract()        masks tidyr::extract()
## x plotly::filter()           masks dplyr::filter(), stats::filter()
## x lubridate::intersect()     masks base::intersect()
## x dplyr::lag()               masks stats::lag()
## x purrr::set_names()         masks magrittr::set_names()
## x lubridate::setdiff()       masks base::setdiff()
## x lubridate::union()         masks base::union()
```

Importing the Dataset

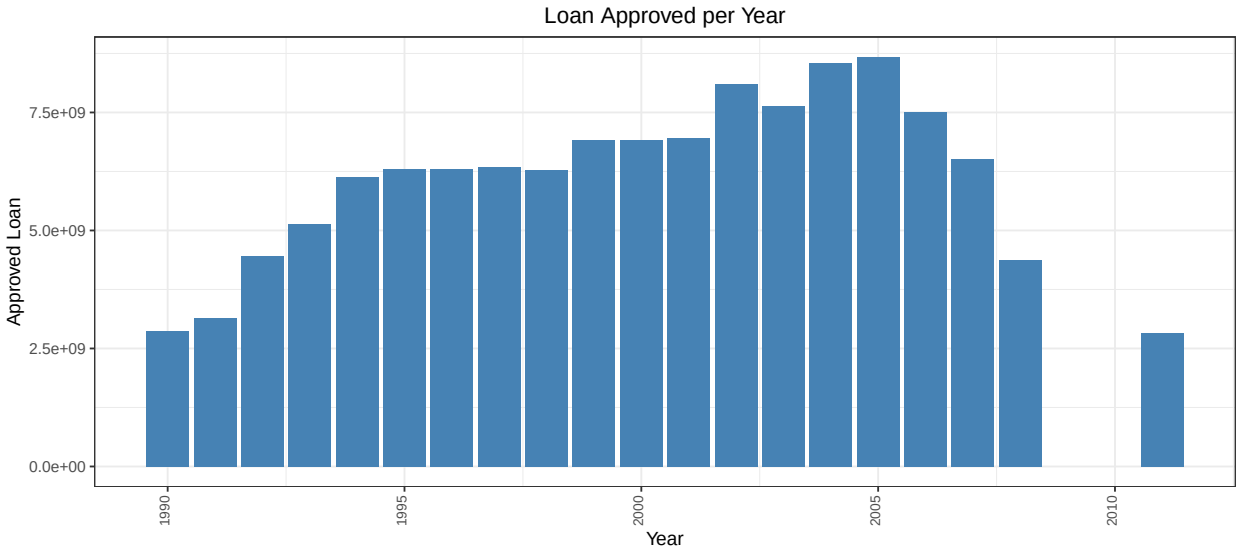
The dataset SBAnational is imported using read_csv().

```
## `summarise()` has grouped output by 'State'. You can override using the `.groups` argument.
```

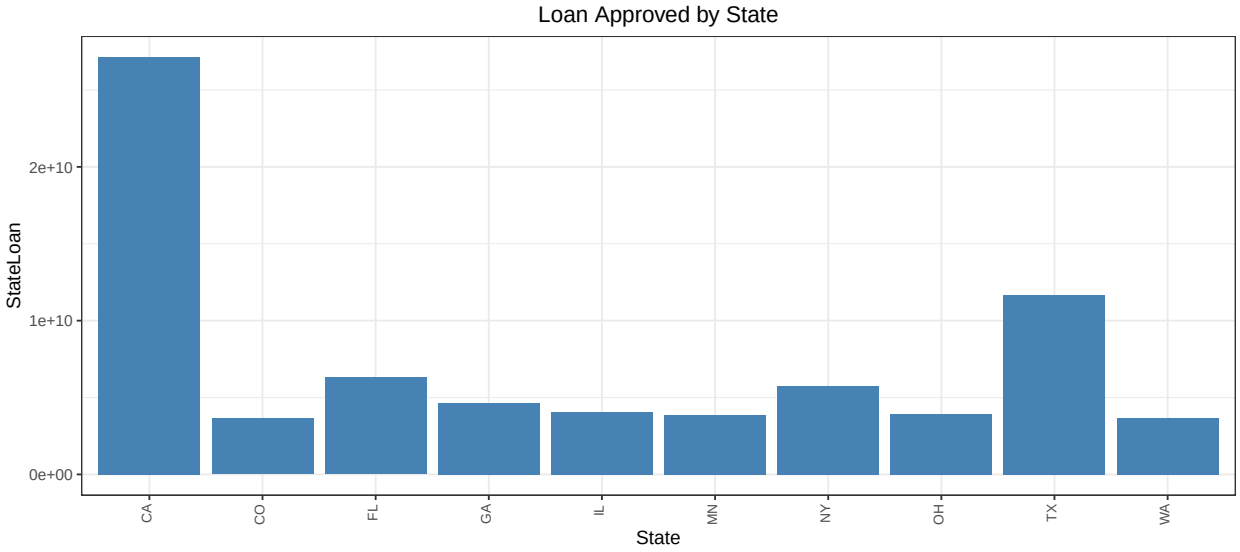
Business Questions

1.1 Classify the Loan Amount on the basis of Years, to identify the number of Loans approved in a Fiscal Year.

```
## Selecting by ApprovedLoan
```

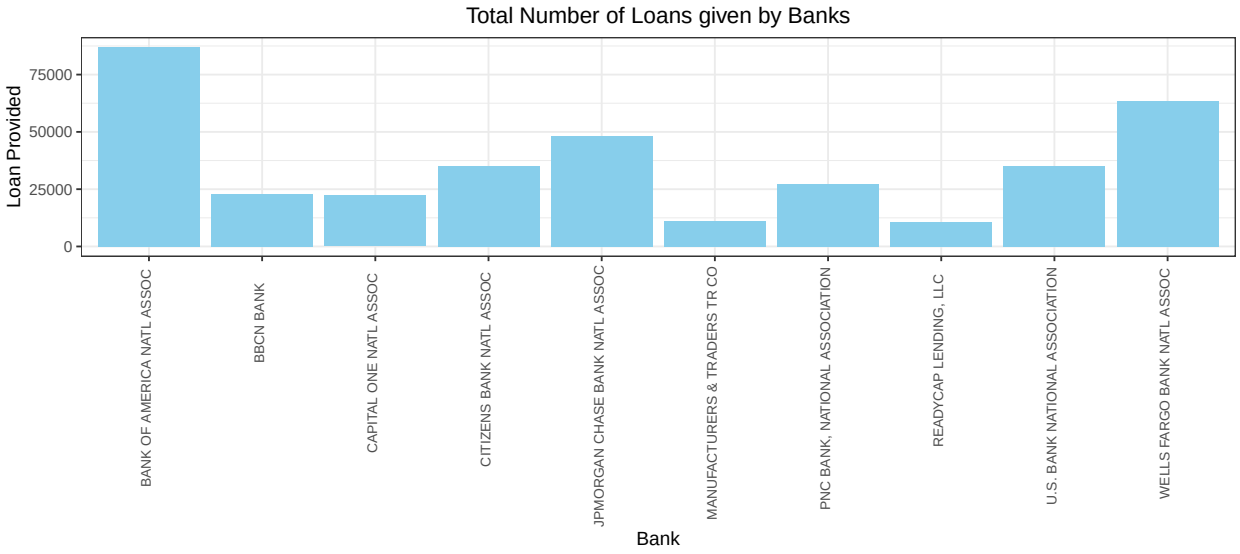


#1.2 Classify the Loan Amount on the basis of states, to identify the Loan amount each state is giving out.



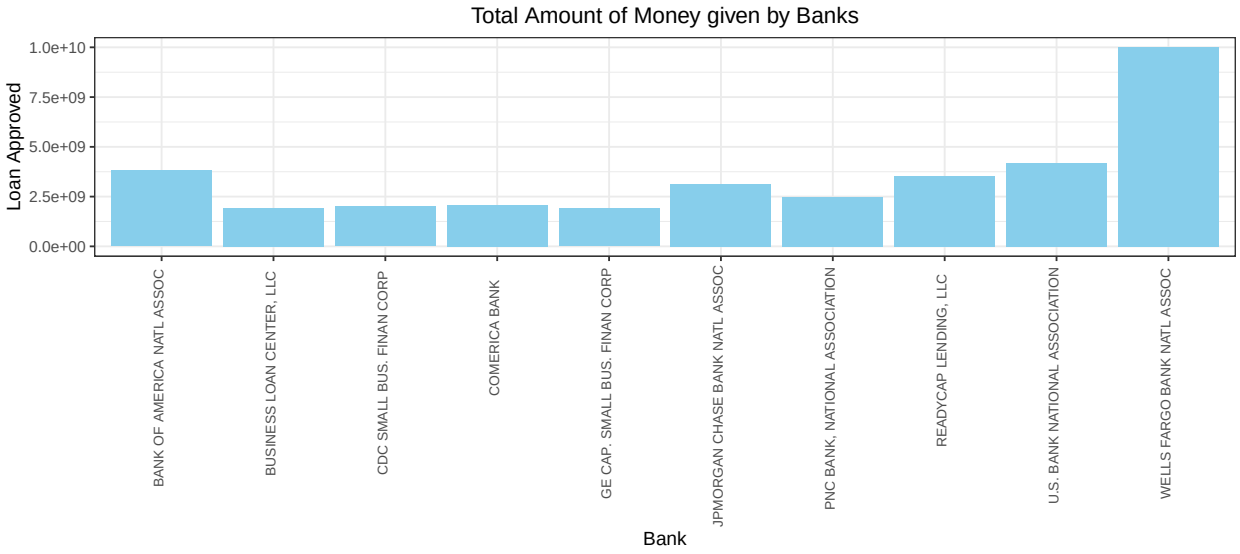
Conclusion The given data provides information about the total amount of loan approved by The U.S. Small Business Administration (SBA) between the years 1962-1984. Its important to analyse reasons for the differences in the low and high risks of default rates. The given data shows the total amount of loan approved by The U.S. Small Business Administration (SBA) on the basis of Top 10 States, to better understand the growth of businesses by states.

#2.1 Amount of Loans given by Banks and the total number of Loans given. #2.2 Banks face with a difficult choice as to whether they should grant a loan because of the high risk of default. Show the total number of records for Banks to analyse the risk factor on the basis of money lended.




```
## # A tibble: 10 x 2
##    Bank                              Loans
##    <chr>                             <dbl>
##  1 WELLS FARGO BANK NATL ASSOC    9999879985
##  2 U.S. BANK NATIONAL ASSOCIATION 4197786866
##  3 BANK OF AMERICA NATL ASSOC     3805406540
##  4 READYCAP LENDING, LLC          3534783474
##  5 JPMORGAN CHASE BANK NATL ASSOC 3120242326
```

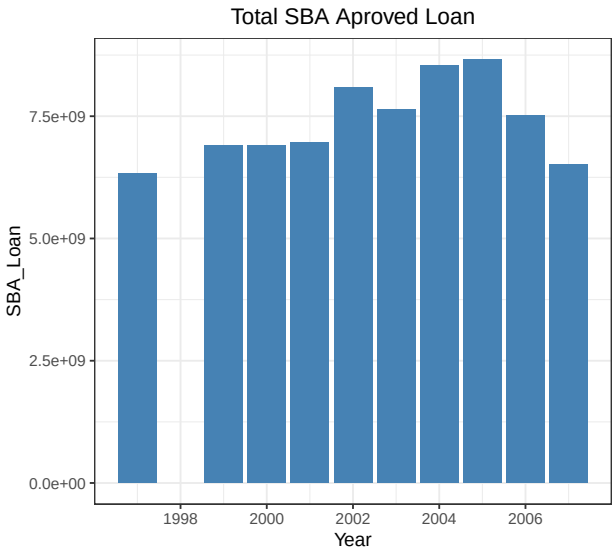
```
##  6 PNC BANK, NATIONAL ASSOCIATION 2490202143
##  7 COMERICA BANK                  2087331694
##  8 CDC SMALL BUS. FINAN CORP      2001560279
##  9 BUSINESS LOAN CENTER, LLC      1915668751
## 10 GE CAP. SMALL BUS. FINAN CORP  1903375222
```

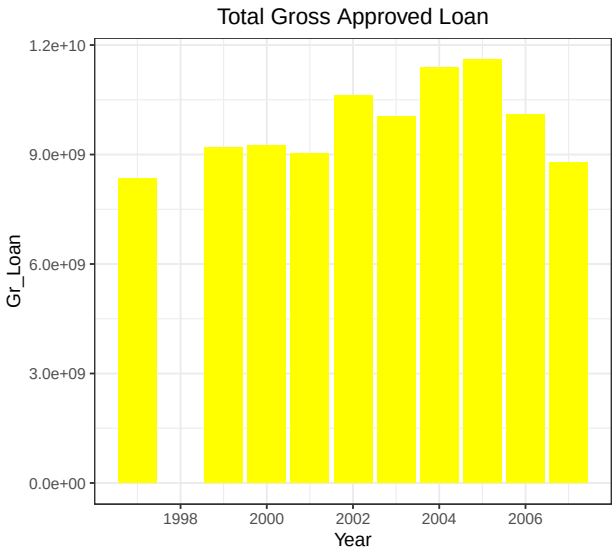


Conclusion

Since SBA only guarantees a portion of the entire loan balance, banks faced with a difficult choice as to whether they should grant such a loan because of the high risk of defaulters. One way to inform their decision making is through analyzing relevant data. The Graph above shows number of Loans granted and the total amount of money provided by various banks

#Q3. What is the Gross amount of Loan Approved by the banamount of the Approved Loan.-by bank,by year



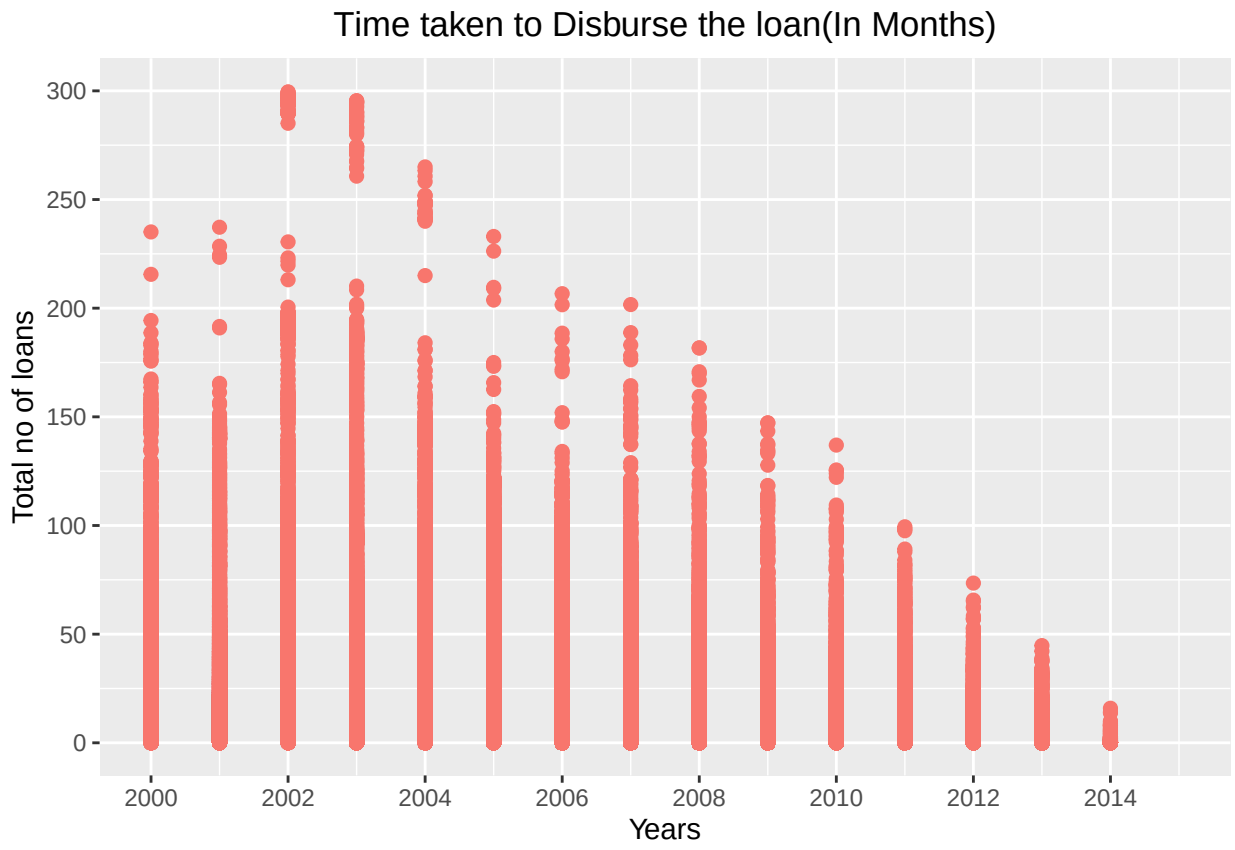


```
## TableGrob (1 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
```

#Conclusion

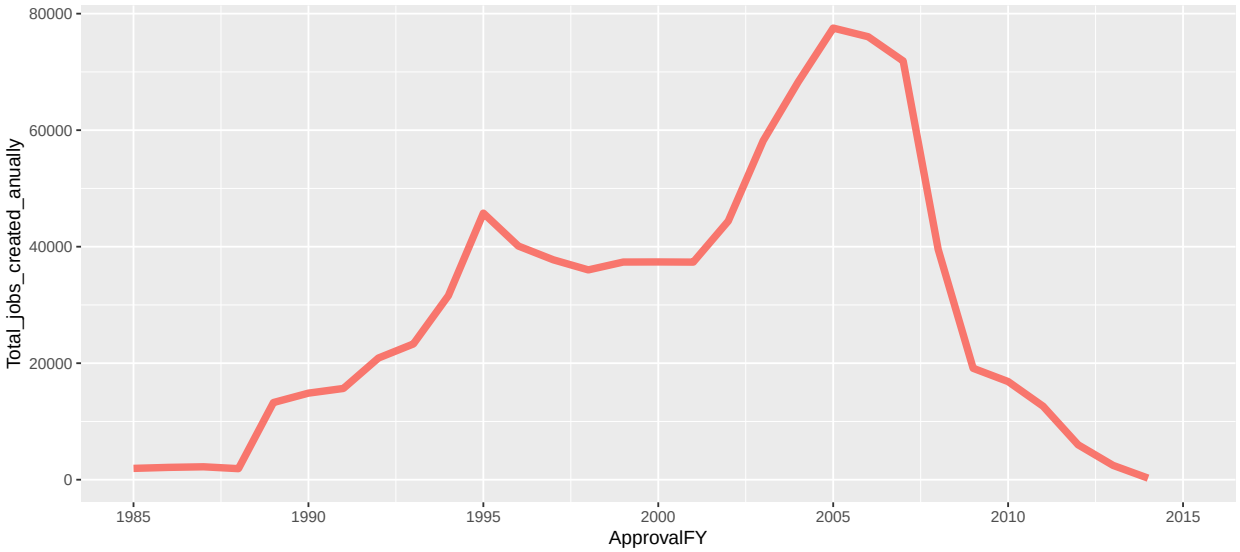
SBA acts much like an insurance provider to reduce the risk for a bank by taking on some of the risk through guaranteeing a portion of the loan. In the case that a loan goes into default, SBA then covers the amount they guaranteed. The graph shows the Gross loan amount approved by SBA Per year and the Gross amount approved by the Bank, since SBA loans only guarantee a portion of the entire loan balance, banks will incur some losses if a small business defaults on its SBA-guaranteed loan.

#5. How long did the bank take to disburse the loans once it was approved?



#Conclusion The duration of days between once loan was approved and bank disbursed the loan was quite high initially and it has improved with years the duration has reduced though the number of applications have increased over the time.

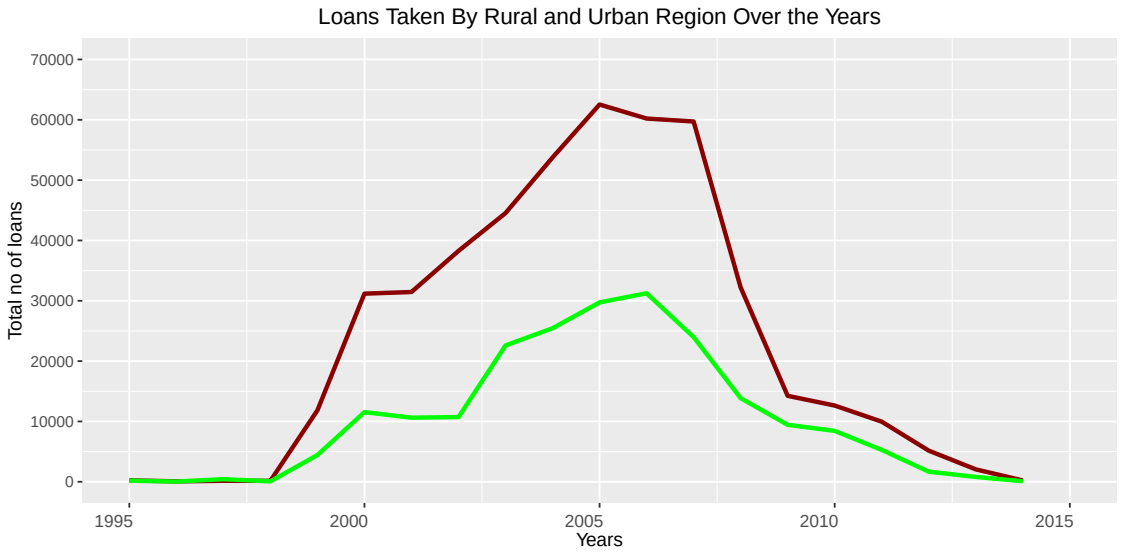
#6. What are the toal number of Jobs created by a Firm annually?



#Conclusion Economists and statisticians use several methods to track economic growth. One of the most significant of these are the Employment growth. The graph above shows the increase in number of employement between the years 2001-2005 and then a dip. This could be because of the Great Recession that took place in 2008 in the United States, leading in increase of unemployment.

#7. Show the number of Loans taken by business in rural and urban regions over the years. Plot a line graph and conclude the analysis

```
## `summarise()` has grouped output by 'UrbanRural'. You can override using the `.groups` argument.
```

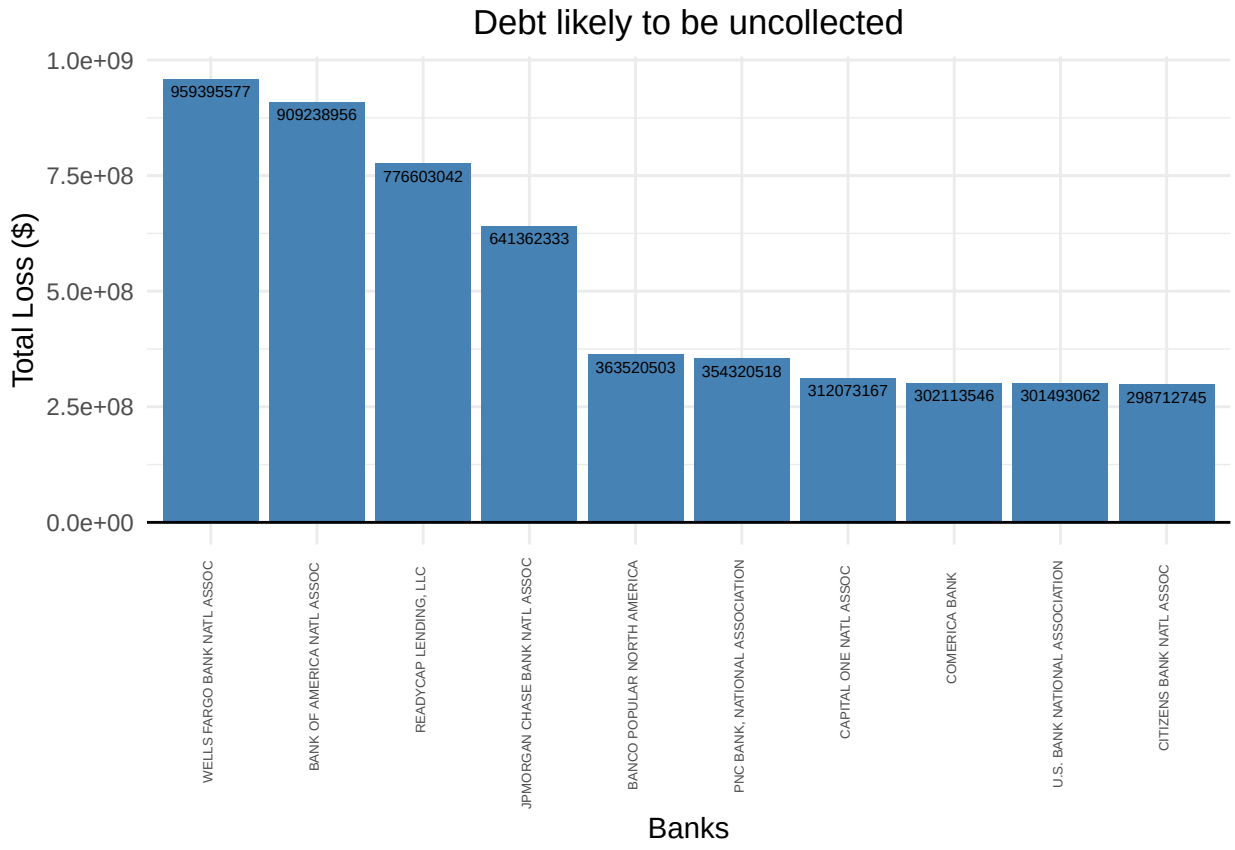


#Plot to support statistics

Conclusion

In the above graph, we can see the Total number of loans taken by businesses in rural and urban areas over the period of years. We can anyse the Employment growth of economy o=looking at the employement generation in the urban and Rural aread.

8. What is the loss incurred by banks based on the bad loans?



Conclusion

A chargeoff is a declaration by a creditor that an amount of debt is unlikely to be collected. This occurs when a consumer becomes severely delinquent on a debt. Traditionally, creditors make this declaration at the point of six months without payment. The banks in this scrutinize the defaulters into high-risk defaulter list, if not met with the deadline.

9. What are High and Low Doc Programs? Was it useful to the customers?

```
## `summarise()` has grouped output by 'State'. You can override using the `.groups` argument.

## Selecting by HighDocLoan

## Selecting by LowDocLoan
```

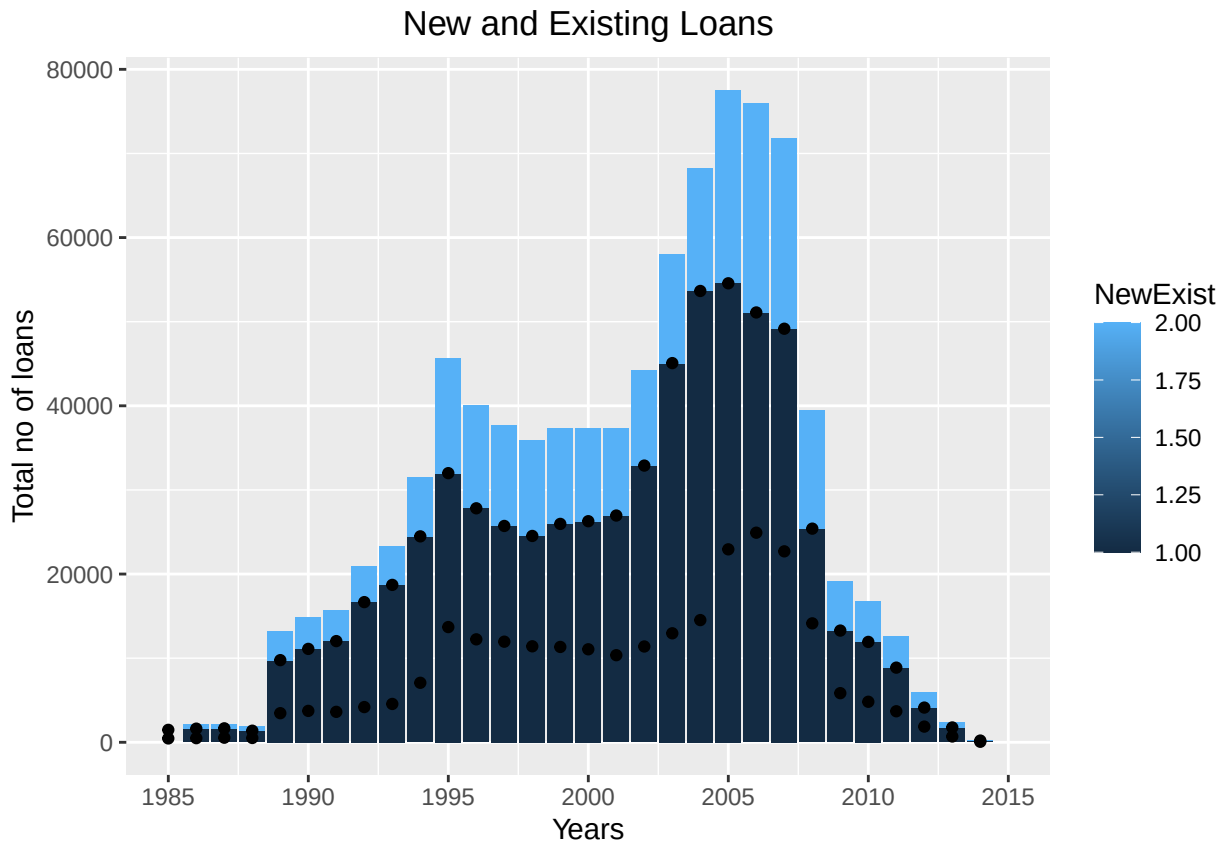

```
## TableGrob (2 x 1) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (2-2,1-1) arrange gtable[layout]
```

Conclusion

LowDoc (Y = Yes, N = No): In order to process more loans efficiently, a "LowDoc Loan" program was implemented where loans under $150,000 can be processed. It's a government supported scheme to startup small busniness and uplift the society, it's smooth process hence can be taken up easily. "Yes" indicates loans with a one-page application, and "No" indicates loans with more information attached to the application. From the graph, we can conclude that the number of states that has opted for more Low Doc Loans and High Doc Loans. The highest is California for High Doc Loans and Texas for Low Doc Loans. Over the years, number of Low Doc Loans have been increased and it was a smooth process and which in turn brought more business to the banks.

10. How many loans where given on the basis of New and Existing Business over the years?

```
## `summarise()` has grouped output by 'ApprovalFY'. You can override using the `.groups` argument.
```



#Conclusion The New and Existing Loans graph shows the Number of loans taken for Existing Business(1) and New Business(2) and the loans given in both the types of Businesses have increased over the years and there was a spike from 2004 to 2008 after which Bank started scrutinizing the documents due to deregulation in the financial industry.

Summary

```
df_SBA<-df_SBA %>%
mutate(
DisbursementDate=as.Date(DisbursementDate, format="%m/%d/%Y"),                #Mutate DisbursementDate, A
ApprovalDate=as.Date(ApprovalDate, format="%m/%d/%Y"),
ChgOffDate=as.Date(ChgOffDate, format="%m/%d/%Y"),
) %>%
mutate(
Disburse_month=month(DisbursementDate, label = TRUE),
Disburse_year=year(DisbursementDate),
ApprovalDate_month=month(ApprovalDate, label = TRUE),
ApprovalDate_year=year(ApprovalDate),
ChgOffDate_month=month(ChgOffDate, label = TRUE),
ChgOffDate_year=year(ChgOffDate)
)


A<-df_SBA %>%
```

```
select(ApprovalDate_month)%>%
group_by(ApprovalDate_month)%>%
filter(is.na(ApprovalDate_month)==FALSE)%>%
summarise(Total_no_of_loans_approved=n())

D <-df_SBA %>%
select(Disburse_month)%>%
group_by(Disburse_month)%>%
filter(is.na(Disburse_month)==FALSE)%>%
summarise(Total_no_of_loans_disbursed=n())

C <-df_SBA %>%
select(ChgOffDate_month)%>%
group_by(ChgOffDate_month)%>%
filter(is.na(ChgOffDate_month)==FALSE)%>%
summarise(Total_no_of_loans_ChgOff=n())
df_month <- data.frame (col1 = A, col2 = D, col3 = C)
colnames(df_month)[1] <- "Months"
colnames(df_month)[2] <- "Loans_approved"
colnames(df_month)[4] <- "Loans_Disbursed"
colnames(df_month)[6] <- "Loans_ChgOff"
kable
```

```
## function (x, format, digits = getOption("digits"), row.names = NA,
##     col.names = NA, align, caption = NULL, label = NULL, format.args = list(),
##     escape = TRUE, ...)
## {
##     format = kable_format(format)
##     if (!missing(align) && length(align) == 1L && !grepl("[^lcr]",
##         align))
##         align = strsplit(align, "")[[1]]
##     if (inherits(x, "list")) {
##         format = kable_format_latex(format)
##         res = lapply(x, kable, format = format, digits = digits,
##             row.names = row.names, col.names = col.names, align = align,
##             caption = NA, format.args = format.args, escape = escape,
##             ...)
##         return(kables(res, format, caption, label))
##     }
##     caption = kable_caption(label, caption, format)
##     if (!is.matrix(x))
##         x = as.data.frame(x)
##     if (identical(col.names, NA))
##         col.names = colnames(x)
##     m = ncol(x)
##     isn = if (is.matrix(x))
##         rep(is.numeric(x), m)
##     else sapply(x, is.numeric)
##     if (missing(align) || (format == "latex" && is.null(align)))
##         align = ifelse(isn, "r", "l")
##     digits = rep(digits, length.out = m)
##     for (j in seq_len(m)) {
##         if (is_numeric(x[, j]))
```

```
##             x[, j] = round(x[, j], digits[j])
##     }
##     if (any(isn)) {
##         if (is.matrix(x)) {
##             if (is.table(x) && length(dim(x)) == 2)
##                 class(x) = "matrix"
##             x = format_matrix(x, format.args)
##         }
##         else x[, isn] = format_args(x[, isn], format.args)
##     }
##     if (is.na(row.names))
##         row.names = has_rownames(x)
##     if (!is.null(align))
##         align = rep(align, length.out = m)
##     if (row.names) {
##         x = cbind(‘ ‘ = rownames(x), x)
##         if (!is.null(col.names))
##             col.names = c(" ", col.names)
##         if (!is.null(align))
##             align = c("l", align)
##     }
##     n = nrow(x)
##     x = replace_na(to_character(x), is.na(x))
##     if (!is.matrix(x))
##         x = matrix(x, nrow = n)
##     x = trimws(x)
##     colnames(x) = col.names
##     if (format != "latex" && length(align) && !all(align %in%
##         c("l", "r", "c")))
##         stop("’align’ must be a character vector of possible values ’l’, ’r’, and ’c’")
##     attr(x, "align") = align
##     if (format == "simple" && nrow(x) == 0)
##         format = "pipe"
##     res = do.call(paste("kable", format, sep = "_"), list(x = x,
##         caption = caption, escape = escape, ...))
##     structure(res, format = format, class = "knitr_kable")
## }
## <bytecode: 0x7fbfad91dcb8>
## <environment: namespace:knitr>
```

```
(df_month %>%
select(Months,Loans_approved,Loans_Disbursed,Loans_ChgOff))
```

```
##    Months Loans_approved Loans_Disbursed Loans_ChgOff
## 1     Jan          67084           95717        12146
## 2     Feb          66342           57054        12352
## 3     Mar          83628           64165        15424
## 4     Apr          80207          106006        13442
## 5     May          77194           62981        15627
## 6     Jun          78290           62414        18491
## 7     Jul          76487           98464        14285
## 8     Aug          78776           60692        16046
## 9     Sep          83068           64805        14956
## 10    Oct          69757          101803        10682
```

```
## 11     Nov           68400           59317          8699
## 12     Dec           69931           63378         10549
```

Depicted below is the month on month summary of the loans applied, sanctioned, disbursed and charged off by the banks in USA for the period 1960 to 2015 in each months Conclusion of loans approved The number of loans approved is over 60K for each month and less than 85K while major loans approved before the summer and Fall and the banks have disbursed the loans during the peak season summer and after fall during Oct . The frequency of charge-off loans over the months has been pretty much the same except the spike in May,June and August