STAT 440: Homework 02 Supplement (HW02S)

Spring 2021, D. Unger

Due: Thursday, March 18 by 11:00 PM CT

Exercise 1

(Return to SBA Business Loans Data)

This Exercise will revisit HW02, Ex2-3 and the SBA Business Loans Data.

- Data link: https://uofi.box.com/shared/static/liz915g0j5zeg67of0ykhoa1shwnr2at
- SBADataKey

The subset here (a file with no extension) contains 29669 observations and 27 columns. This is historical data about actual business loans covered by the Small Business Administration (SBA) from years 1970-2013. The observations are small businesses based in Illinois that seek loans to fund their operations, start-up costs, materials, payroll, rent, etc. The SBA works with banks by guaranteeing a portion of the loan to relieve banks of assuming all financial risk. The original sources are the SBA, Min Li, Amy Mickel, and Stanley Taylor.

Here is the code from the HW02 Solutions to read in the data.

```
library(tidyverse)
```

```
sba_data <- read_delim("https://uofi.box.com/shared/static/liz915g0j5zeg67of0ykhoa1shwnr2at",
            "\t",
            escape_double = FALSE,
            col_types = cols(ApprovalDate = col_date(format = "%d-%b-%y"),
                             DisbursementDate = col_date(format = "%d-%b-%y"),
                             DisbursementGross = col_number(),
                             BalanceGross = col_number(),
                             ChgOffPrinGr = col_number(),
                             GrAppv = col_number(),
                             SBA_Appv = col_number(),
                             NewExist = col_factor()),
            trim_ws = TRUE)
names(sba_data)
```

```
##  [1] "LoanNr_ChkDgt"     "Name"              "City"
##  [4] "State"             "Zip"               "Bank"
##  [7] "BankState"         "NAICS"             "ApprovalDate"
## [10] "ApprovalFY"        "Term"              "NoEmp"
## [13] "NewExist"          "CreateJob"         "RetainedJob"
## [16] "FranchiseCode"     "UrbanRural"        "RevLineCr"
## [19] "LowDoc"            "ChgOffDate"        "DisbursementDate"
## [22] "DisbursementGross" "BalanceGross"      "MIS_Status"
## [25] "ChgOffPrinGr"      "GrAppv"            "SBA_Appv"
```

That code is somewhat incomplete in terms of managing the data because it does not address factors (i.e., categorical variables).

By any means you choose (most likely by updating the above code or by adding coerced assignment statements after the import), create a data set called `sba_data2` so that all categorical variables are of class `factor`.

Further, update those factor variables so that each factor value has a label using the information in the SBADataKey. For example, the UrbanRural variable will have any value of 1 appear as "Urban", 2 appear as "Rural", and 0 as "Undefined". I have not specifically shown how to do this in class, but some very simple searches online will show how. Missing values for any such variable may stay as is. You are not required to perform this task for the NAICS variable.

When finished, output the descriptor portion and the first 5 observations of only the following variables: Name, City, State, UrbanRural, NewExist, and LowDoc.

```
sba_data2 <- read_delim("https://uofi.box.com/shared/static/liz915g0j5zeg67of0ykhoa1shwnr2at",
            "\t",
            escape_double = FALSE,
            col_types = cols(ApprovalDate = col_date(format = "%d-%b-%y"),
                             DisbursementDate = col_date(format = "%d-%b-%y"),
                             DisbursementGross = col_number(),
                             BalanceGross = col_number(),
                             ChgOffPrinGr = col_number(),
                             GrAppv = col_number(),
                             SBA_Appv = col_number(),
                             NewExist = col_factor(),
                             UrbanRural = col_factor(),
                             RevLineCr = col_factor(),
                             LowDoc = col_factor(),
                             MIS_Status = col_factor()),
            trim_ws = TRUE)
```

```
sba_data2$NewExist <- factor(sba_data2$NewExist, levels = c(1,2), labels = c("Existing business","New bu
sba_data2$UrbanRural <- factor(sba_data2$UrbanRural, levels = c(1,2,0), labels = c("Urban","Rural","und
sba_data2$RevLineCr <- factor(sba_data2$RevLineCr, levels = c('Y','N',0), labels = c("Yes","No",0))
sba_data2$LowDoc <- factor(sba_data2$LowDoc, levels = c('Y','N'), labels = c("Yes","No"))
sba_data2$MIS_Status <- factor(sba_data2$MIS_Status, levels = c('P I F','CHGOFF'), labels = c("Paid in
```

```
sba_data2.0 = sba_data2%>%
  select(Name, City, State, UrbanRural, NewExist,LowDoc)
str(sba_data2.0)
```

```
## tibble [29,669 x 6] (S3: tbl_df/tbl/data.frame)
##  $ Name      : chr [1:29669] "PROFESSIONAL ELEVATOR SERVICES" "M.A.S. TRUCKING, INC." "RAYMIES GROCE
##  $ City      : chr [1:29669] "CHICAGO" "SPRINGFIELD" "Chicago" "SCHAUMBURG" ...
##  $ State     : chr [1:29669] "IL" "IL" "IL" "IL" ...
##  $ UrbanRural: Factor w/ 3 levels "Urban","Rural",..: 3 3 3 3 2 3 3 1 1 3 ...
##  $ NewExist  : Factor w/ 2 levels "Existing business",..: 1 1 1 1 1 1 2 1 2 1 ...
##  $ LowDoc    : Factor w/ 2 levels "Yes","No": 2 2 2 2 2 2 2 2 2 2 ...
##  - attr(*, "problems")= tibble [1 x 5] (S3: tbl_df/tbl/data.frame)
##   ..$ row     : int 24663
##   ..$ col     : chr "ApprovalFY"
##   ..$ expected: chr "no trailing characters"
##   ..$ actual  : chr "1976A"
##   ..$ file    : chr "'https://uofi.box.com/shared/static/liz915g0j5zeg67of0ykhoa1shwnr2at'"
```

```
head(sba_data2.0, n=5)
```

```
## # A tibble: 5 x 6
##   Name                          City       State UrbanRural NewExist        LowDoc
##   <chr>                         <chr>      <chr> <fct>      <fct>           <fct>
## 1 PROFESSIONAL ELEVATOR SERV~ CHICAGO    IL    undefined  Existing busin~ No
## 2 M.A.S. TRUCKING, INC.       SPRINGFIE~ IL    undefined  Existing busin~ No
## 3 RAYMIES GROCERIES INC.      Chicago    IL    undefined  Existing busin~ No
## 4 RICCARDO'S PIZZERIA & COCK~ SCHAUMBURG IL    undefined  Existing busin~ No
## 5 HASTY EQUIPMENT INC         EDWARDS    IL    Rural      Existing busin~ No
```

Exercise 2

Using the new and improved `sba_data2` from **Exercise 1**, create a table that compares a few basic statistics for GrAppv between the Rural and Urban business. The statistics appearing in the table should be the minimum, first quartile, median, mean, third quartile, maximum, and standard deviation. Comment on any interesting similarities or disparities.

```
Rural = c(summary(sba_data2$GrAppv[sba_data2$UrbanRural=='Rural']),sd(sba_data2$GrAppv[sba_data2$UrbanRu
names(Rural) = c(names(summary(sba_data2$GrAppv[sba_data2$UrbanRural=='Rural'])),"Std Dev")
Urban = c(summary(sba_data2$GrAppv[sba_data2$UrbanRural=='Urban']),sd(sba_data2$GrAppv[sba_data2$UrbanRu
names(Urban) = c(names(summary(sba_data2$GrAppv[sba_data2$UrbanRural=='Urban'])),"Std Dev")
ans = rbind(Rural,Urban)
ans
```

```
##       Min. 1st Qu. Median     Mean 3rd Qu.    Max.  Std Dev
## Rural 1000   30000  50000 132545.4  130000 3017000 235729.6
## Urban  800   25000  50000 163289.1  150000 5000000 294627.3
```

While majority of the values differ the median of both Rural and Urban is exactly the same at 50000. Additionally we notice while Rural has a higher 1st Quartile it has lower values in comparison to Urban in Mean, 3rd Quatile, max and std Dev

Exercise 3

Using the new and improved `sba_data2` from **Exercise 1** and the information in the SBADataKey, create a subset of the data named "illinois" such that all of the following are addressed:

a. the bank issuing the small business loan is in Illinois

b. the approval fiscal year is between 2000 and 2010

c. the NAICS corresponding only to Utilities, Construction, or Manufacturing

d. the SBA guaranteed approved amount is greater than 50% of the gross amount approved by the bank

e. create a new variable called "SBApercent" which equals the SBA guaranteed approved amount as a proportion of the gross amount approved by the bank

f. keep the following variables: Name, City, State, Bank State, ApprovalFY, and SBApercent.

Now, print the first 5 observations and the last 5 observations of the "illinois" data set. *All columns should be visible in the print out.*

```
names(sba_data2)
```

```
##  [1] "LoanNr_ChkDgt"     "Name"              "City"
##  [4] "State"             "Zip"               "Bank"
##  [7] "BankState"         "NAICS"             "ApprovalDate"
## [10] "ApprovalFY"        "Term"              "NoEmp"
## [13] "NewExist"          "CreateJob"         "RetainedJob"
## [16] "FranchiseCode"     "UrbanRural"        "RevLineCr"
## [19] "LowDoc"            "ChgOffDate"        "DisbursementDate"
## [22] "DisbursementGross" "BalanceGross"      "MIS_Status"
## [25] "ChgOffPrinGr"      "GrAppv"            "SBA_Appv"
```

```
sba_data_3.0 = sba_data2%>%
  filter(BankState == 'IL')%>%
  filter(ApprovalFY>=2000 & ApprovalFY<=2010)%>%
  filter(SBA_Appv > (0.5*GrAppv))%>%
  filter(NAICS>=220000|340000>=NAICS)%>%
  mutate(SBApercent = SBA_Appv/GrAppv)%>%
  select(Name, City, BankState, ApprovalFY, SBApercent)
head(sba_data_3.0,n=5)
```

```
## # A tibble: 5 x 5
##   Name                          City         BankState ApprovalFY SBApercent
##   <chr>                         <chr>        <chr>          <dbl>      <dbl>
## 1 HASTY EQUIPMENT INC           EDWARDS      IL              2006      1
## 2 MARION COUNTY AMBULANCE, INC. ODIN         IL              2006      0.85
## 3 BRAZ TRANS INC                CHICAGO      IL              2006      1
## 4 CONTRACT PACKAGING PLUS INC   BENSENVILLE  IL              2006      0.75
## 5 BULLY'S SMOKEHOUSE            EDWARDSVILLE IL              2006      0.750
```

```
tail(sba_data_3.0,n=5)
```

```
## # A tibble: 5 x 5
##   Name                           City       BankState ApprovalFY SBApercent
##   <chr>                          <chr>      <chr>          <dbl>      <dbl>
## 1 MIDWEST WEALTH MANAGMENT LLC   ROSELLE    IL              2006       0.85
## 2 REVERMANN CHIROPRACTIC, P.C.   BREESE     IL              2006       0.75
## 3 TAYLOR BUSINESS INSTITUTE      CHICAGO    IL              2006       1
## 4 Siteline Interior Carpentry, I LEMONT     IL              2006       0.75
## 5 SHADHESHWAIR MATAJI, INC       BRIDGEVIEW IL              2006       1
```

Exercise 4

Using the new and improved `sba_data2` from **Exercise 1**, determine the number of small businesses that have one of the following abbreviations in their Name.

- Inc
- LLC
- Ltd

Keep in mind that the case of the words should not matter. Also, some businesses may represent these abbreviations with periods. For example, "inc." or "L.L.C." or "ltd." or other derivations.

In a table, report the frequency count of each of the three abbreviations as well as the proportion as compared to the total number of businesses in the `sba_data2` data set.

```
#install.packages("sjmisc")
library(sjmisc)
#?str_contains
x="Inc"
y="LLC"
z="Ltd"
a=0
b=0
c=0
m=length(sba_data2$Name)
sba_data2$Name=str_replace(sba_data2$Name, '.','')
for (i in 1:m){
  if(str_contains(sba_data2$Name[i], x,, ignore.case = TRUE)){
      a=a+1
  }
  if(str_contains(sba_data2$Name[i], y, ignore.case = TRUE)){
      b=b+1
  }
  if(str_contains(sba_data2$Name[i], z, ignore.case = TRUE)){
    c=c+1
  }
}
freq=c(a,b,c)
names(freq) = c("Inc","LLC","Ltd")
freq
```

```
##   Inc   LLC   Ltd
## 11492  1579   722
```
