EXPLORATORY DATA ANALYSIS OF LOANS FOR

PURPOSE OF THIS ANALYSIS

Ascertain what factors determine whether a small business loan will be paid in full or written off

Clarify whether the Small Business Agency (SBA) does indeed help small businesses create/retain jobs

VISION

To uncover what makes a small business succeed or fail to pay off its loan, but using consumer language, for the understanding of the general public. The earlier study on this same dataset was rather technical.

ASK

What are my limitations?

No stats or finance background, rely solely on Google and, of course, my classes in R and EDA

What data do I have on hand? SBA National.csv from https://www.kaggle.com/datasets/mirbektoktogaraev/should-this-loan-be-approved-or-denied

This was an earlier study of the same dataset, though with different parameters and by more experienced data scientists

About the data set

- From U.S. Small Business Administration (SBA)
- Guarantees a percentage of small business loans
- In case a small business can't pay, SBA will pay the bank instead
- Without the guarantee (up to 85% of loan), traditional banks often consider a small biz "too risky"

Why does SBA guarantee loans?

- Small businesses create jobs and lower unemployment; they're the primary job creator in the US
- SBA guaranteed these two start ups once upon a time
- But of course, not every business can make it big
- That's why SBA continues to be there for them

STRATEGY

The plan is to clean, wrangle, visualize and analyze the data on R

Explore the data visually to select variables

EXECUTE

Use R, particularly ggplot2, to aid understanding of the different factors affecting the outcome of a small business loan and create predictive models

EVALUATE

Conduct a train-test set to evaluate the model/s (or attempt to)

DATA CLEANING/WRANGLING IN R

BEFORE

899,164

observations

27

variables

- Previous cleaning in Excel took at least one week – and it wasn't even complete. First cleaning in R took one day and accomplished more, allowing time to review the code again and again and refine EDA, feature selection and model building.
- Imputed NAs, blanks, invalid entries (e.g. a T or R where it should be Y or N)
- Removed unnecessary columns: LoanNr_ChkDgt, Zip, Bank State, ApprovalFY, FranchiseCode, etc.
- Renamed columns so they're easier to understand at a glance, see list of variables for a clearer idea of this
- Transformed chr to int or factors or dates, as needed
- Removed outliers using the IQR method
- Removed commas in values to avoid them being coerced to NA when converted to numbers/integers
- Trimmed industry levels to just first two digits, to limit to just main categories, not hundreds of subcategories
- Removed rows of loans paid in full that had balances and written off amounts, as these should be zero, therefore those with balances and written off amounts were clearly errors - unless they were signs of fraud hidden in plain sight
- As loans paid in full didn't have a written off date, it was not possible to convert the variable written off date from character to date without forcing NAs on the blank entries. To determine the timediff between the approval date of the loan and its written off date, I created a subset of written off loans and then converted the relevant variables to date format and analyze from there.

AFTER

196959

observations

18

variables

BEFORE

27

variables

Before	After
LoanNr_ChkDgt (loan code number)	
Name	name (of small business)
City	
State	state
Zip	
Bank	bank
BankState	
NAICS	
ApprovalDate	industry
ApprovalFY	approval date
Term	loan_term
NoEmp	employees
NewExist	new_or_existing_biz
CreateJob	jobs_created
RetainedJob	jobs_retained
FranchiseCode	
UrbanRural	urban_or_ rural
RevLineCr	revolving_credit _line
LowDoc (low doc loan, faster/easier approval loans under $150K)	
ChgOffDate	written_off_date
ApprovalDate	approval_date
DisbursementDate	
DisbursementGross	amount_disbursed
BalanceGross	
MIS_Status	loan_status: paid in full or written off
ChgOffPrinGr	written_off_amount
GrAppv	bank_approved_loan
SBA_Appv	sba_guaranteed_amount

AFTER

19

variables

though more would be removed during EDA

THE RELEVANT VARIABLES FOR EDA

1. name (of small business)
2. state
3. bank
4. industry
5. approval date
6. loan_term
7. employees
8. new_or_existing_biz
9. jobs_created
10. jobs_retained
11. urban_or_ rural
12. revolving_credit _line
13. written_off_date
14. approval_date
15. amount_disbursed
16. loan_status: paid in full or written off
17. written_off_amount
18. bank_approved_loan
19. sba_guaranteed_amount

THIS DATA SET IS UNUSUAL:
IT HAS TWO POTENTIAL TARGET VARIABLES

Written_off_amount, a continuous variable: a loan paid in full would have a zero written off amount while a loan written off would have a value.

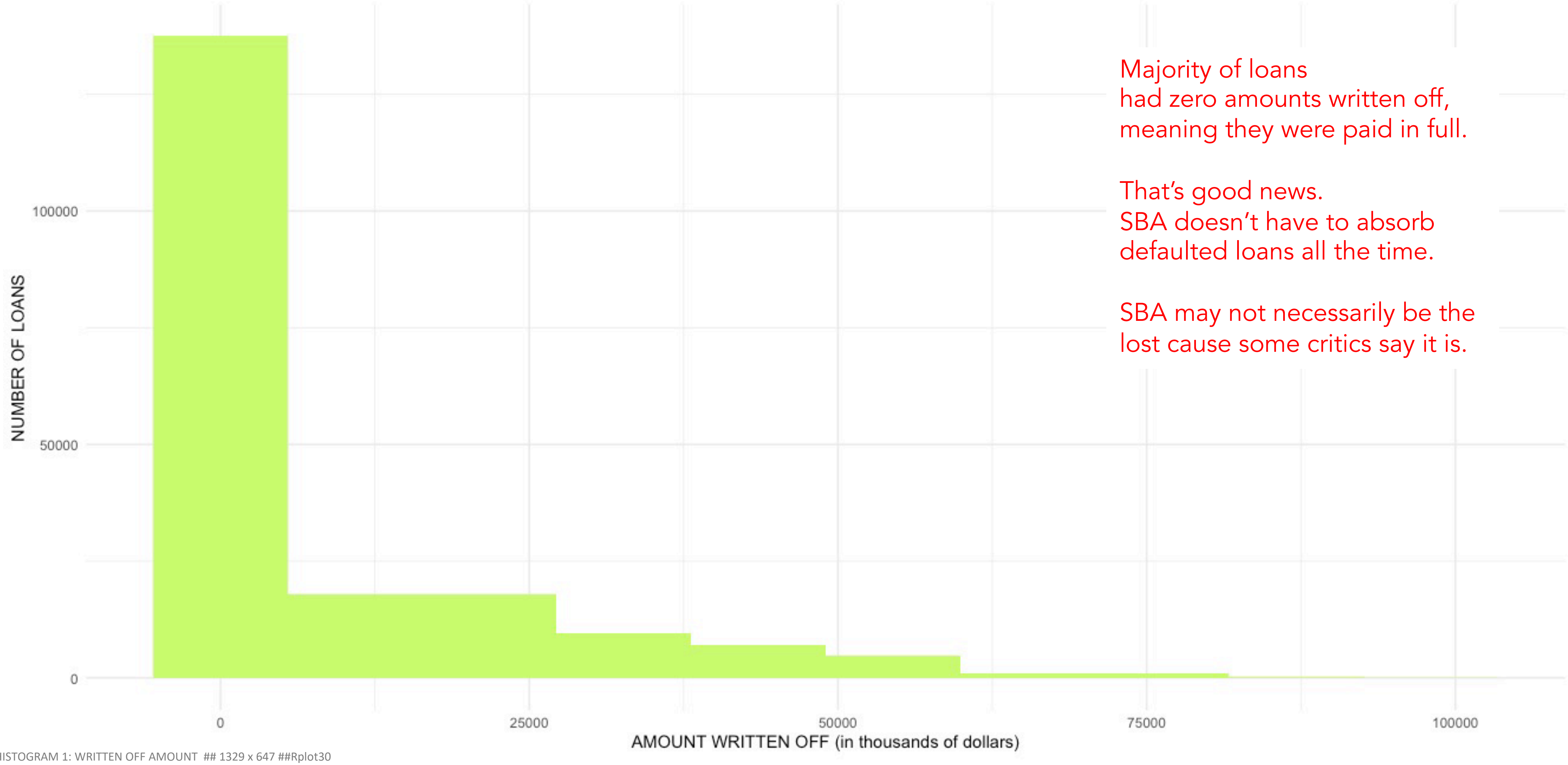
Loan_status, a binomial variable: 1) paid in full or 2) charged off/written off

One or the other can be used as a target variable, though both cannot be used in a model, since they're highly correlated as they stand for the same thing.

With two potential target variables, one continuous and one categorical , I can explore a linear AND a logistic regression.

VISUALIZING DATA FOR INSIGHTS

HISTOGRAM 1 : AMOUNT WRITTEN OFF (TARGET VARIABLE)



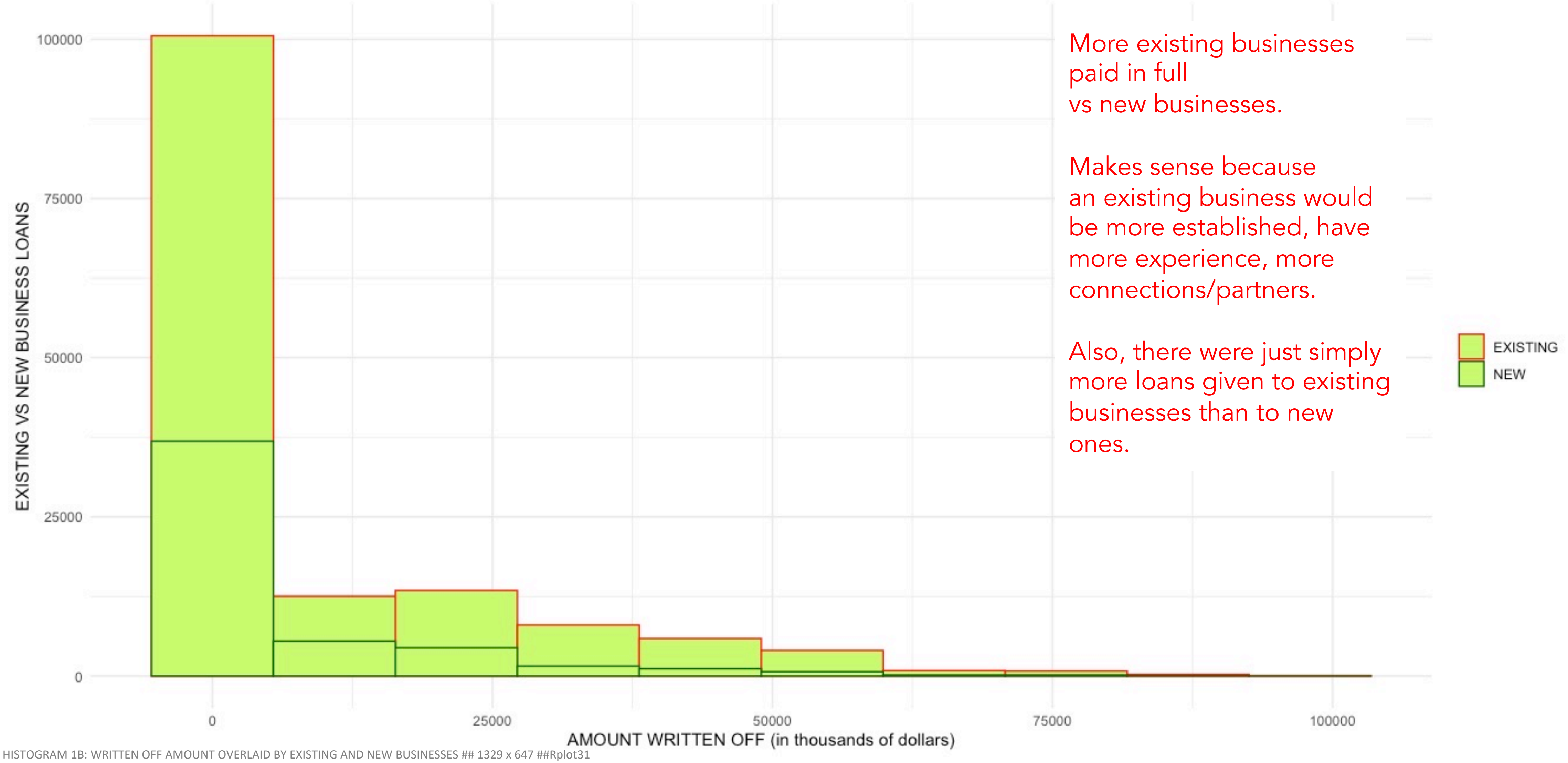
Majority of loans
had zero amounts written off,
meaning they were paid in full.

That's good news.
SBA doesn't have to absorb
defaulted loans all the time.

SBA may not necessarily be the
lost cause some critics say it is.

```
## HISTOGRAM 1: WRITTEN OFF AMOUNT  ## 1329 x 647 ##Rplot30
ggplot(loandata, aes(x=written_off_amount)) +
  geom_histogram(fill="darkolivegreen1", alpha=10, position="identity", bins=10)  +
  theme_minimal() +
  labs(y="NUMBER OF LOANS", x="AMOUNT WRITTEN OFF (in thousands of dollars)")
```

HISTOGRAM 1B : AMOUNT WRITTEN OFF FOR EXISTING VS NEW BUSINESSES



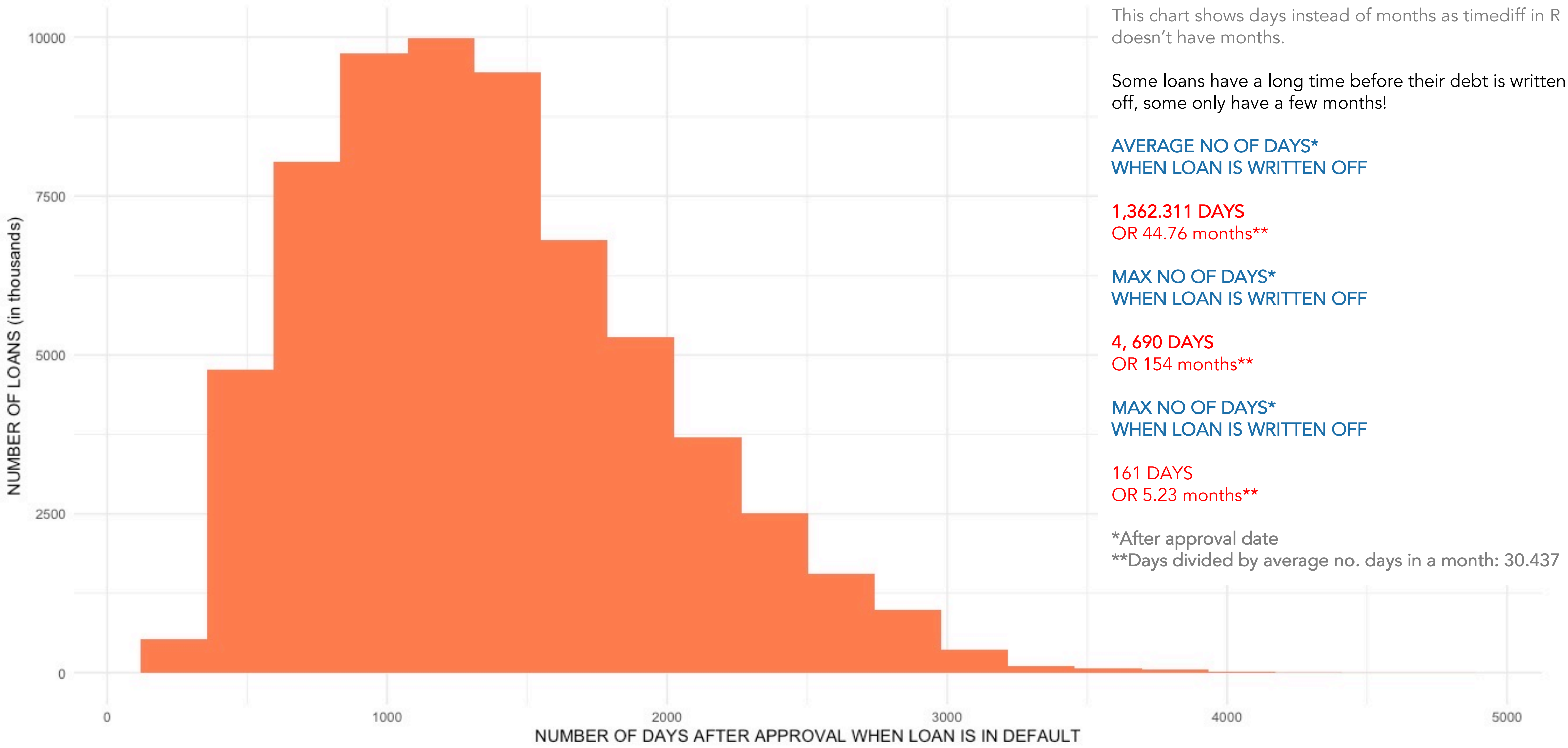
More existing businesses paid in full vs new businesses.

Makes sense because an existing business would be more established, have more experience, more connections/partners.

Also, there were just simply more loans given to existing businesses than to new ones.

```
## HISTOGRAM 1B: WRITTEN OFF AMOUNT OVERLAID BY EXISTING AND NEW BUSINESSES ## 1329 x 647 ##Rplot31
ggplot(loandata, aes(x=written_off_amount, color=new_or_existing_biz)) +
  geom_histogram(fill="darkolivegreen1", alpha=10, position="identity", bins=10) +
  theme_minimal() +
  labs(y="EXISTING VS NEW BUSINESS LOANS", x="AMOUNT WRITTEN OFF (in thousands of dollars)") +
  theme(legend.title=element_blank()) +
  scale_color_manual(labels = c("EXISTING", "NEW"),
              values = c("red", "darkgreen"))
```

HISTOGRAM A : HOW MANY DAYS BEFORE A LOAN IS CONSIDERED IN DEFAULT?



This chart shows days instead of months as timediff in R doesn't have months.

Some loans have a long time before their debt is written off, some only have a few months!

AVERAGE NO OF DAYS*
WHEN LOAN IS WRITTEN OFF

1,362.311 DAYS
OR 44.76 months**

MAX NO OF DAYS*
WHEN LOAN IS WRITTEN OFF

4, 690 DAYS
OR 154 months**

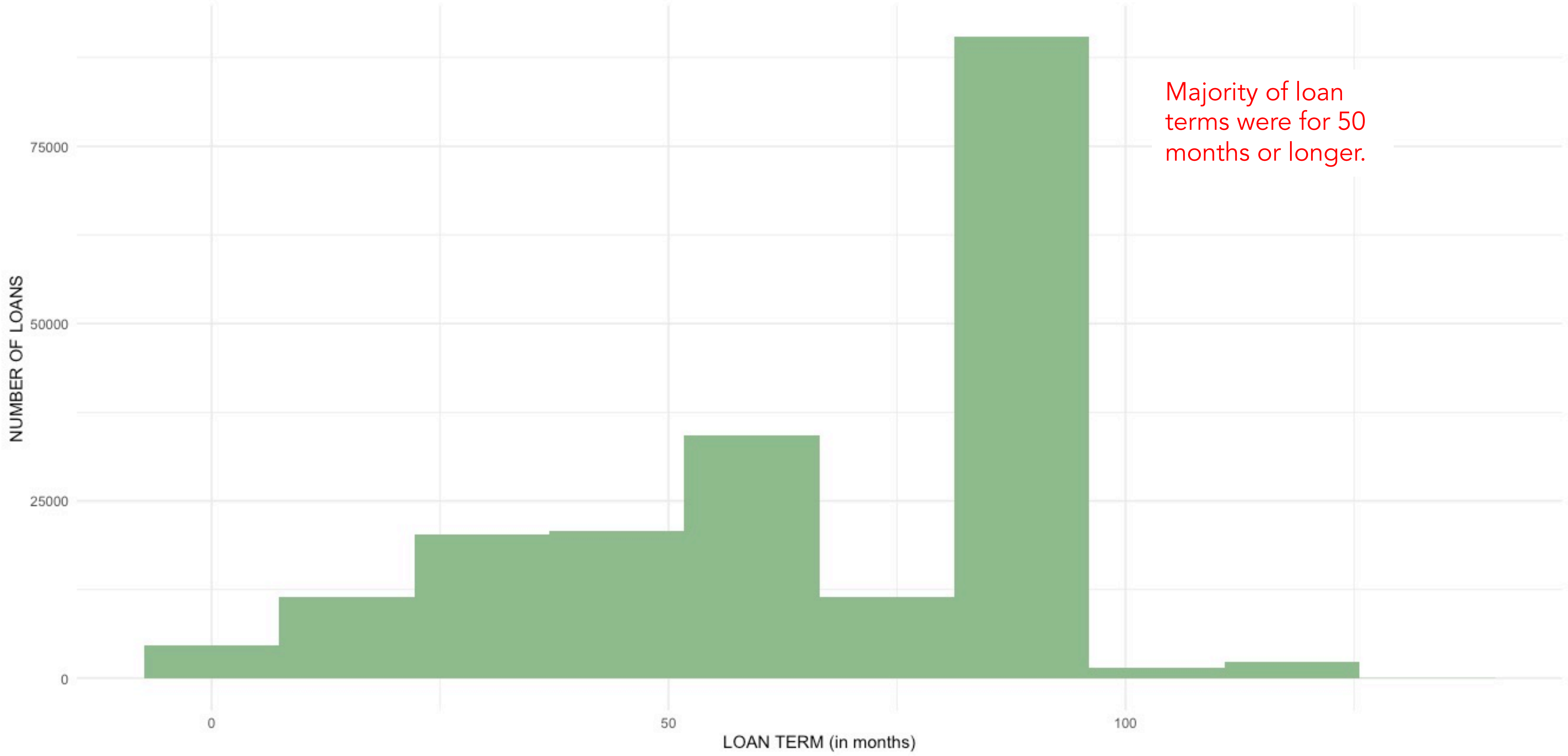
MAX NO OF DAYS*
WHEN LOAN IS WRITTEN OFF

161 DAYS
OR 5.23 months**

*After approval date
**Days divided by average no. days in a month: 30.437

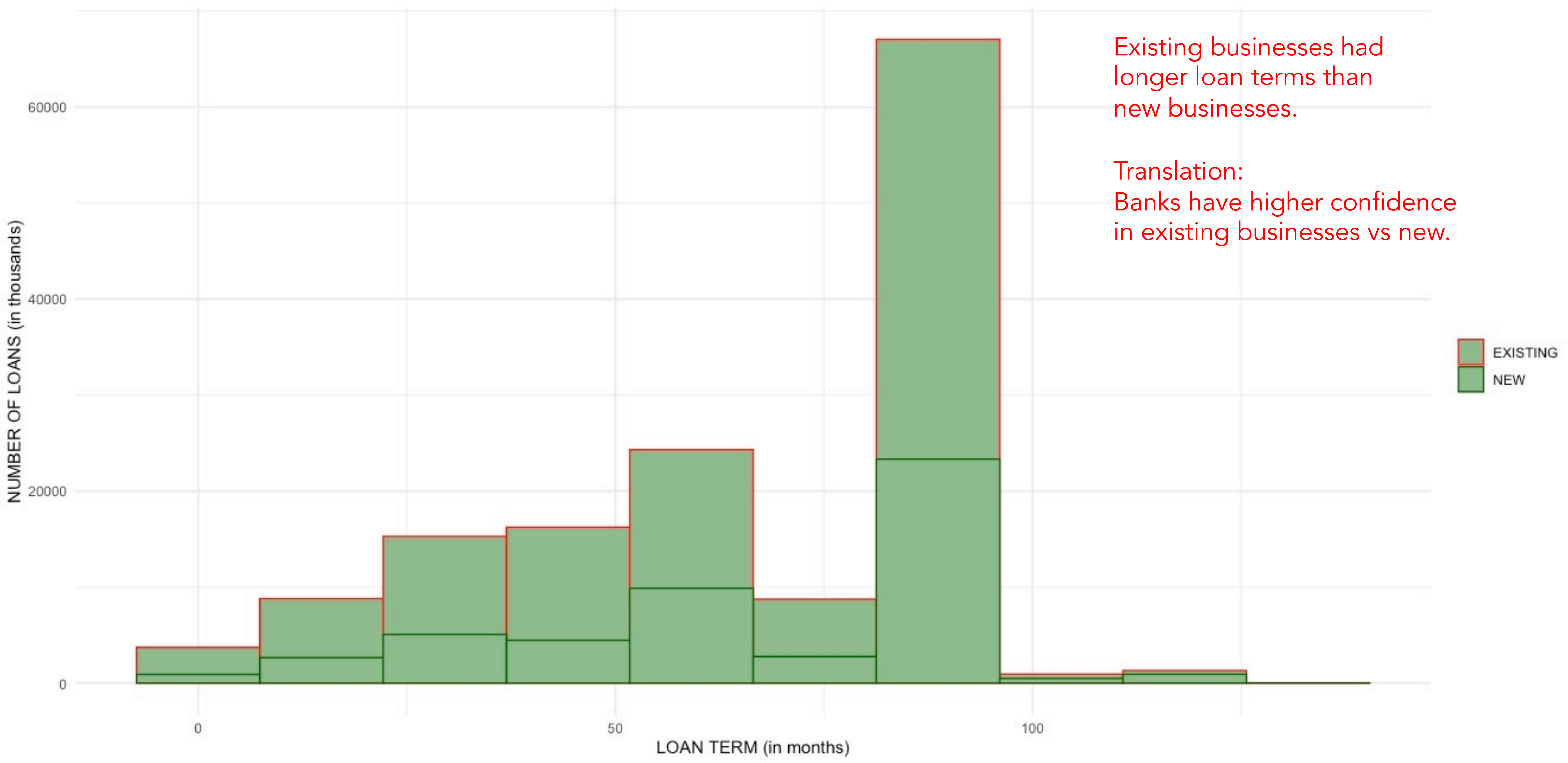
```
## HISTOGRAM A: WRITTEN OFF LOANS DEFAULT DAYS  ## 1329 x 647 ##Rplot71
ggplot(written_off_loans, aes(x=defaultdays)) +
  geom_histogram(fill="coral", alpha=10, position="identity", bins=20)  +
  theme_minimal() +
  labs(x="NUMBER OF DAYS AFTER APPROVAL WHEN LOAN IS IN DEFAULT", y="NUMBER OF LOANS (in thousands)")
```

HISTOGRAM 2 : LOAN TERM




```
## HISTOGRAM 2: LOAN TERM ## 1329 x 647 ##Rplot32
ggplot(loandata, aes(x=loan_term)) +
  geom_histogram(fill="darkseagreen", alpha=10, position="identity", bins=10)  +
  theme_minimal() +
    labs(y="NUMBER OF LOANS", x="LOAN TERM (in months)")
```

HISTOGRAM 2B : LOAN TERM FOR EXISTING VS NEW BUSINESSES

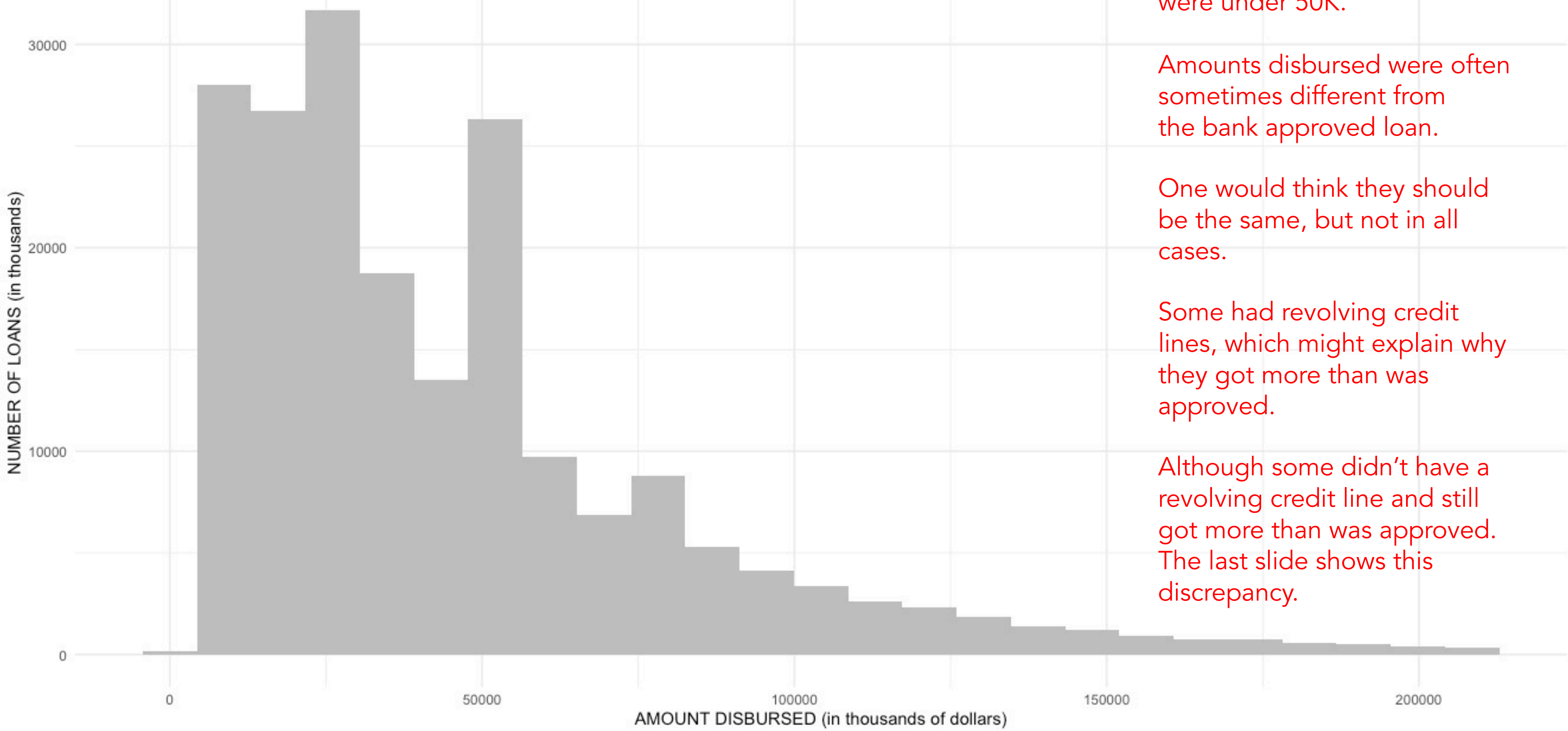


Existing businesses had longer loan terms than new businesses.

Translation:
Banks have higher confidence in existing businesses vs new.

```
## HISTOGRAM 2B: LOAN TERM OVERLAID BY EXISTING AND NEW BUSINESSES ## 1329 x 647 ##Rplot33
ggplot(loandata, aes(x=loan_term, color=new_or_existing_biz)) +
  geom_histogram(fill="darkseagreen", alpha=10, position="identity", bins=10) +
  theme_minimal() + labs(y="NUMBER OF LOANS (in thousands)", x="LOAN TERM (in months)") +
  theme(legend.title=element_blank()) +
  scale_color_manual(labels = c("EXISTING", "NEW"), values = c("red", "darkgreen"))
```

HISTOGRAM 3 : AMOUNT DISBURSED



Majority of amounts disbursed were under 50K.

Amounts disbursed were often sometimes different from the bank approved loan.

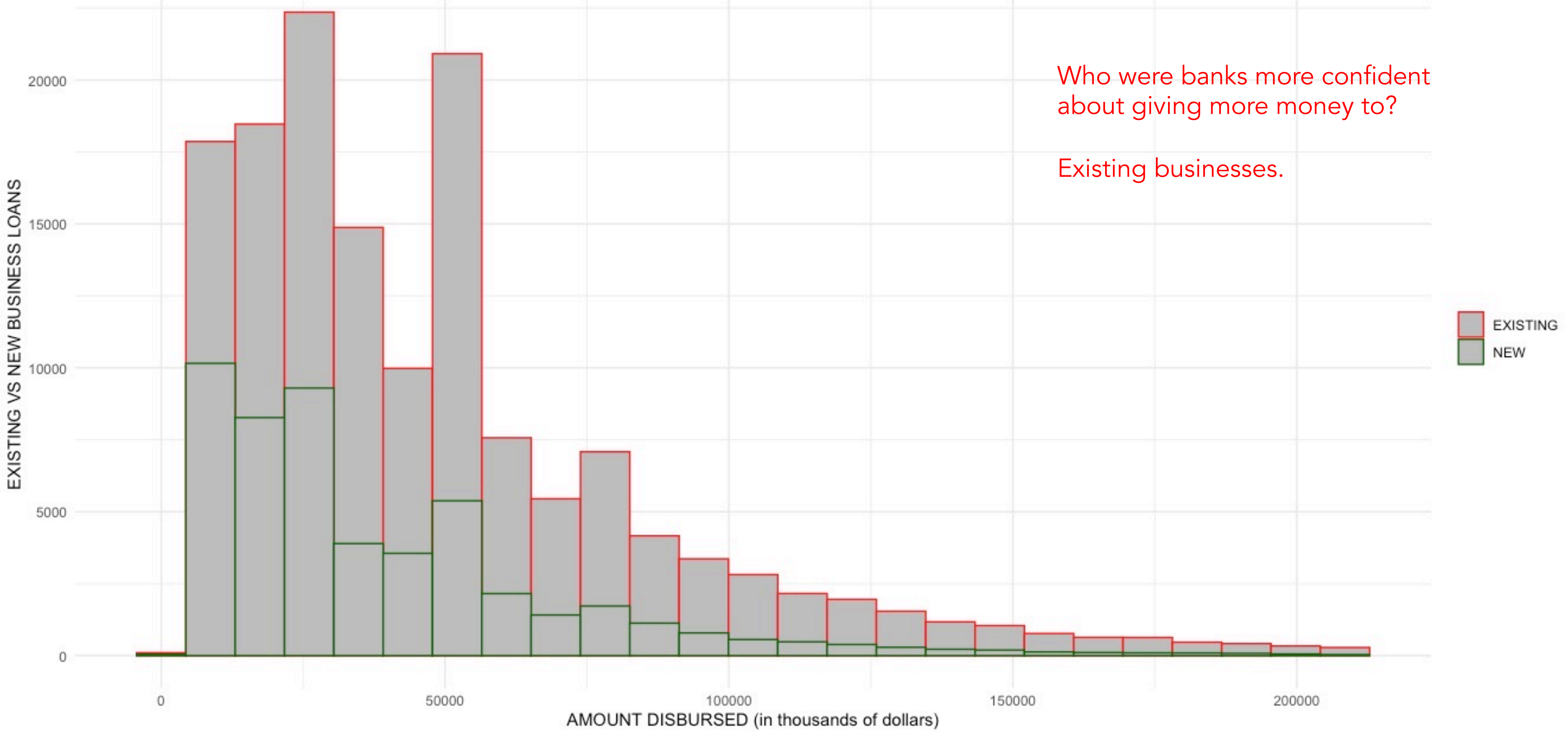
One would think they should be the same, but not in all cases.

Some had revolving credit lines, which might explain why they got more than was approved.

Although some didn't have a revolving credit line and still got more than was approved. The last slide shows this discrepancy.

```
## HISTOGRAM 3: AMOUNT DISBURSED ## 1329 x 647 ##Rplot34
ggplot(loandata, aes(x=loan_term, color=new_or_existing_biz)) +
ggplot(loandata, aes(x=amount_disbursed)) +
  geom_histogram(fill="grey", alpha=10, position="identity", bins=25) +
  theme_minimal() + labs(y="NUMBER OF LOANS (in thousands)", x="AMOUNT DISBURSED (in thousands of dollars)") +
  theme(legend.title=element_blank())
```

HISTOGRAM 3B : AMOUNT DISBURSED TO EXISTING VS NEW BUSINESSES

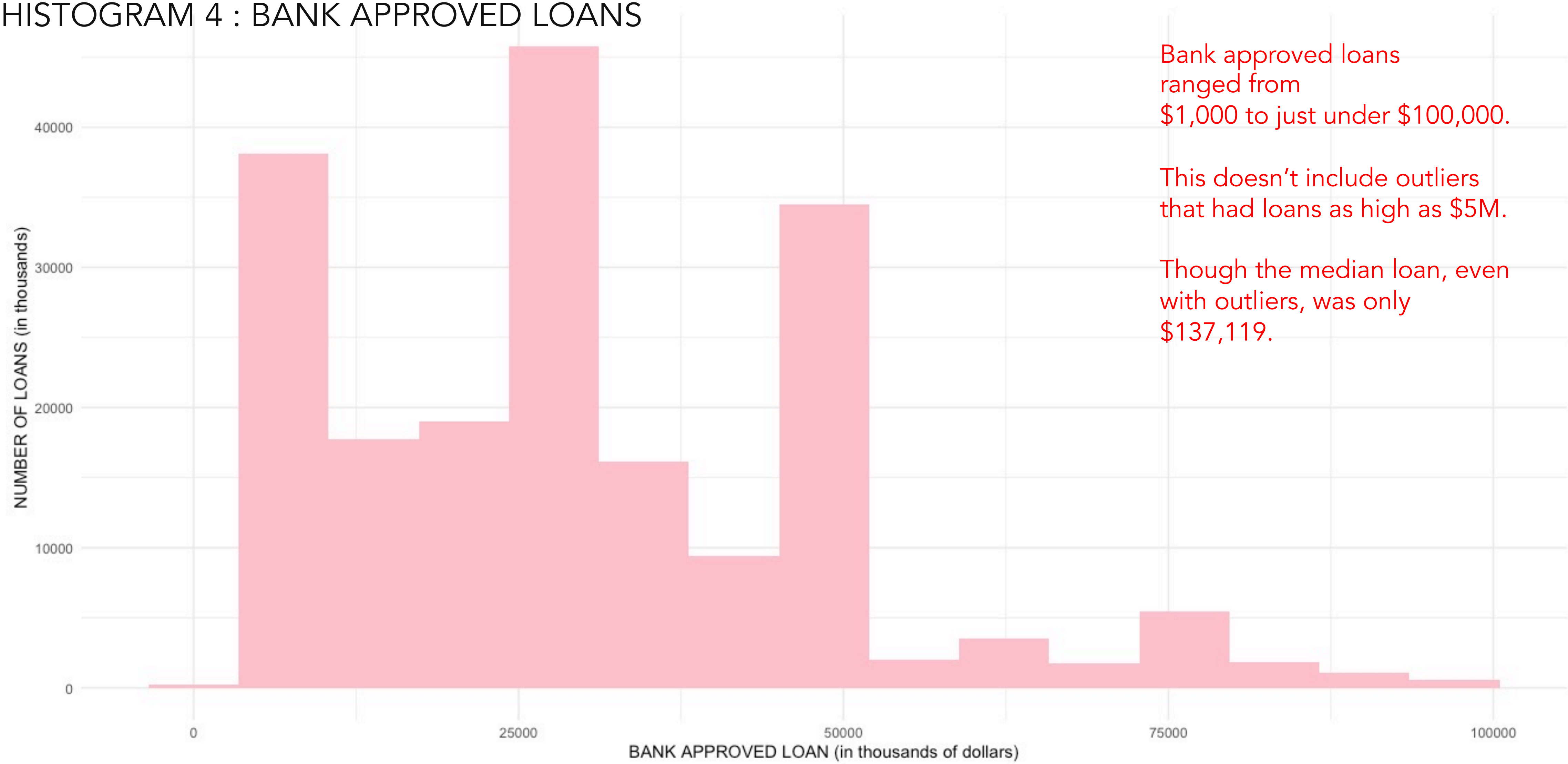


Who were banks more confident about giving more money to?

Existing businesses.

```
## HISTOGRAM 3B: AMOUNT DISBURSED OVERLAID BY EXISTING OR NEW BUSINESS  ## 1329 x 647 ##Rplot35
ggplot(loandata, aes(x=amount_disbursed, color=new_or_existing_biz)) +
  geom_histogram(fill="grey", alpha=10, position="identity", bins=25) +
  theme_minimal() + labs(y="EXISTING VS NEW BUSINESS LOANS", x="AMOUNT DISBURSED (in thousands of dollars)") +
  theme(legend.title=element_blank()) +
  scale_color_manual(labels = c("EXISTING", "NEW"), values = c("red", "darkgreen"))
```

HISTOGRAM 4 : BANK APPROVED LOANS



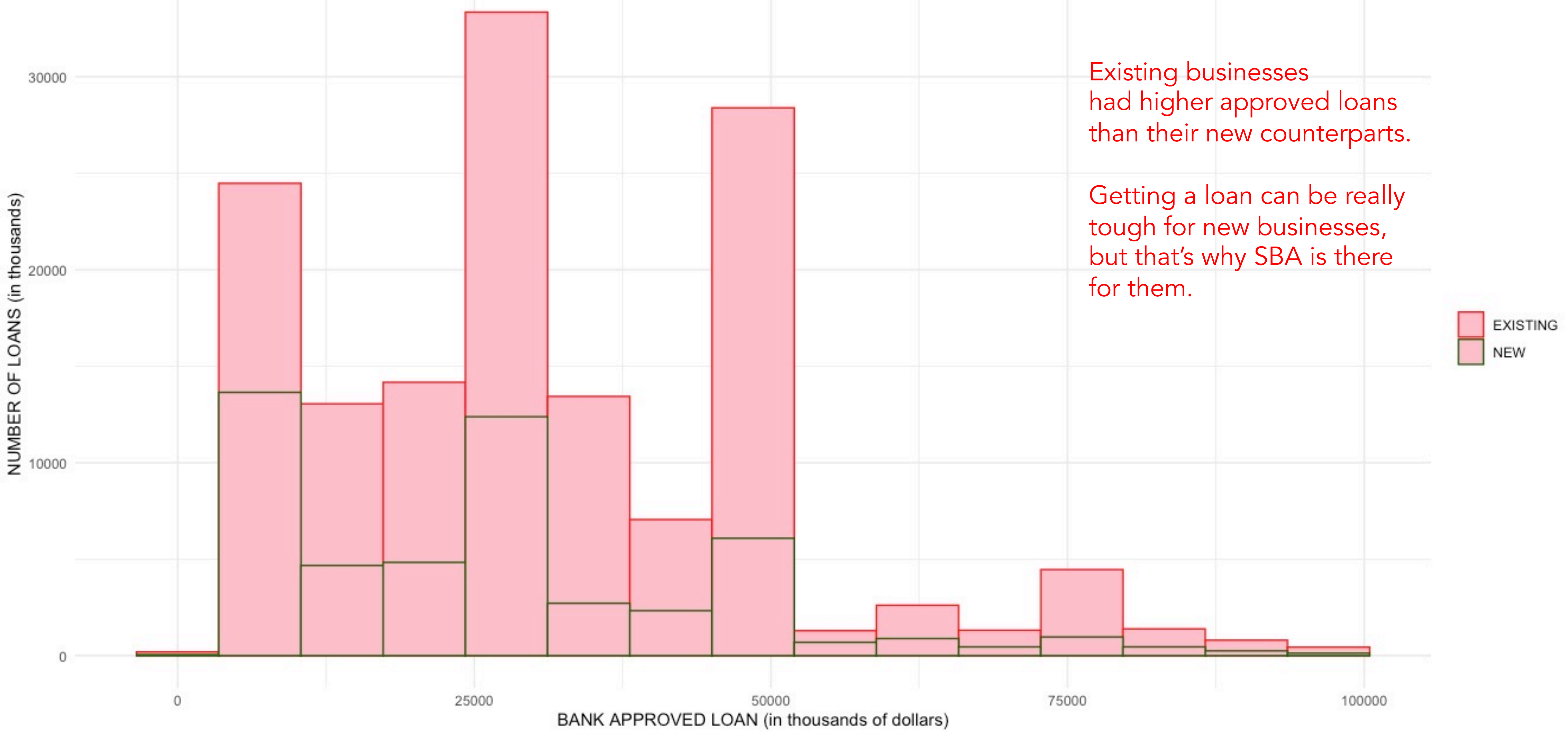
Bank approved loans ranged from $1,000 to just under $100,000.

This doesn't include outliers that had loans as high as $5M.

Though the median loan, even with outliers, was only $137,119.

```
## HISTOGRAM 4: BANK APPROVED LOAN ## 1329 x 647 ##Rplot68
ggplot(loandata, aes(x=bank_approved_loan)) +
  geom_histogram(fill="pink", alpha=10, position="identity", bins=15) +
  theme_minimal() + labs(y="NUMBER OF LOANS (in thousands)", x="BANK APPROVED LOAN (in thousands of dollars)") +
  theme(legend.title=element_blank())
```

HISTOGRAM 4B: BANK APPROVED LOANS TO EXISTING VS NEW BUSINESSES



Existing businesses had higher approved loans than their new counterparts.

Getting a loan can be really tough for new businesses, but that's why SBA is there for them.

```
## HISTOGRAM 4B: BANK APPROVED LOAN OVERLAID BY EXISTING OR NEW BUSINESS  ## 1329 x 647 ##Rplot69
ggplot(loandata, aes(x=bank_approved_loan, color=new_or_existing_biz)) +
  geom_histogram(fill="pink", alpha=10, position="identity", bins=15) +
  theme_minimal() + labs(y="NUMBER OF LOANS (in thousands)", x="BANK APPROVED LOAN (in thousands of dollars)") +
  theme(legend.title=element_blank()) +
  scale_color_manual(labels = c("EXISTING", "NEW"), values = c("red", "darkgreen"))
```

HISTOGRAM 5 : AMOUNT GUARANTEED BY SBA

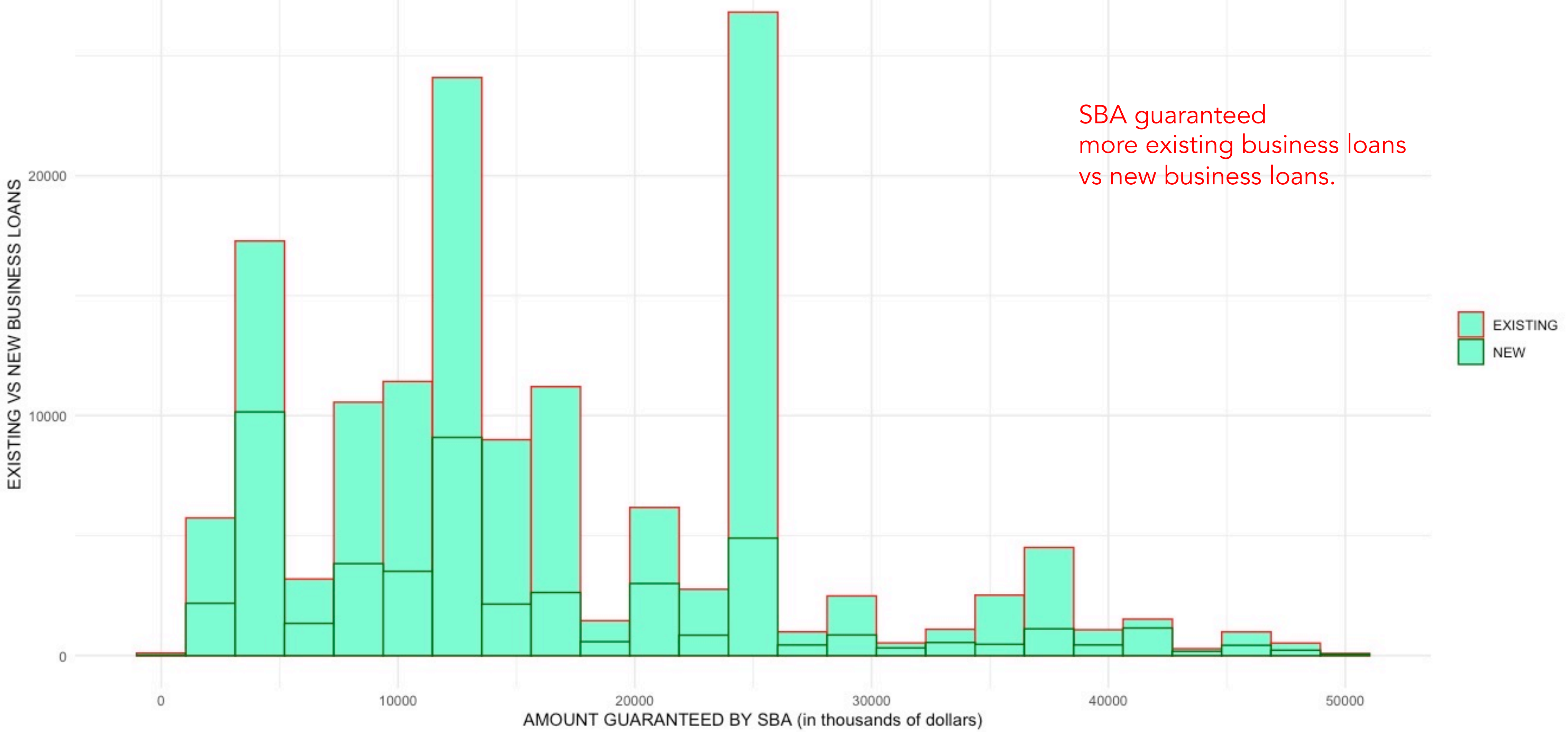
SBA guaranteed amounts from as little as $1,000 to $50,000

One of the categorical variables, low_doc_loan, indicates whether a small business can get a loan with just a one-page application.
As the low doc loan only offered $150,000, not many businesses may have availed of it (based on the low number of YESes under the variable).

Therefore, I didn't consider low_doc_loan a predictor variable.

```
## HISTOGRAM 5: AMOUNT GUARANTEED BY SBA ## 1329 x 647 ##Rplot38
ggplot(loandata, aes(x=sba_guaranteed_amount)) +
  geom_histogram(fill="aquamarine", alpha=10, position="identity", bins=25) +
  theme_minimal() + labs(y="NUMBER OF LOANS (in thousands)", x="AMOUNT GUARANTEED BY SBA (in thousands of dollars)") +
  theme(legend.title=element_blank())
```

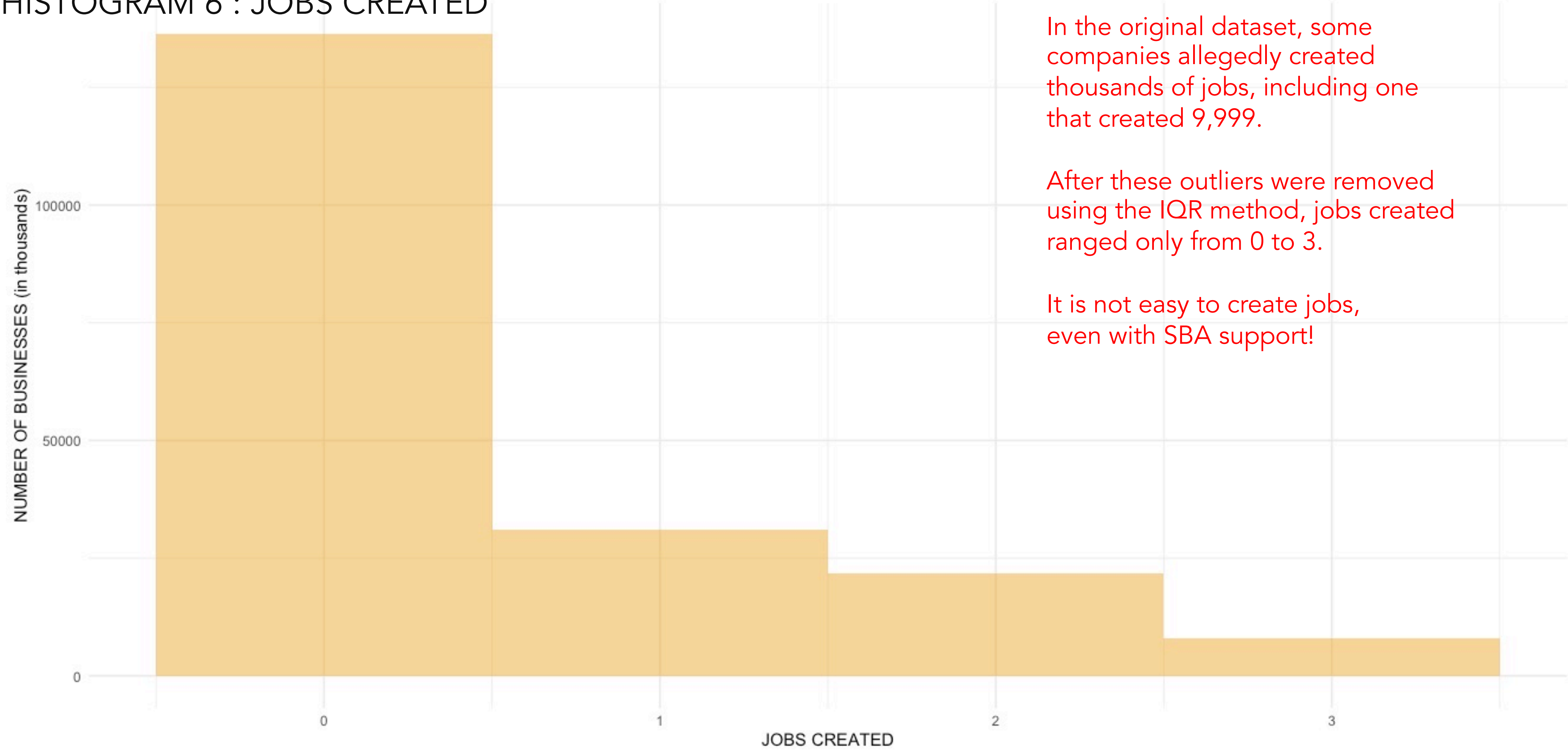
HISTOGRAM 5B: AMOUNT GUARANTEED BY SBA FOR EXISTING VS NEW BUSINESSES



SBA guaranteed
more existing business loans
vs new business loans.

```
## HISTOGRAM 5B: AMOUNT GUARANTEED BY SBA FOR EXISTING VS NEW BUSINESSES ## 1329 x 647 ##Rplot39
ggplot(loandata, aes(x=sba_guaranteed_amount, color=new_or_existing_biz)) +
  geom_histogram(fill="aquamarine", alpha=10, position="identity", bins=25) +
  theme_minimal() + labs(y="EXISTING VS NEW BUSINESS LOANS", x="AMOUNT GUARANTEED BY SBA (in thousands of dollars)") +
  theme(legend.title=element_blank()) +
  scale_color_manual(labels = c("EXISTING", "NEW"), values = c("red", "darkgreen"))
```

HISTOGRAM 6 : JOBS CREATED



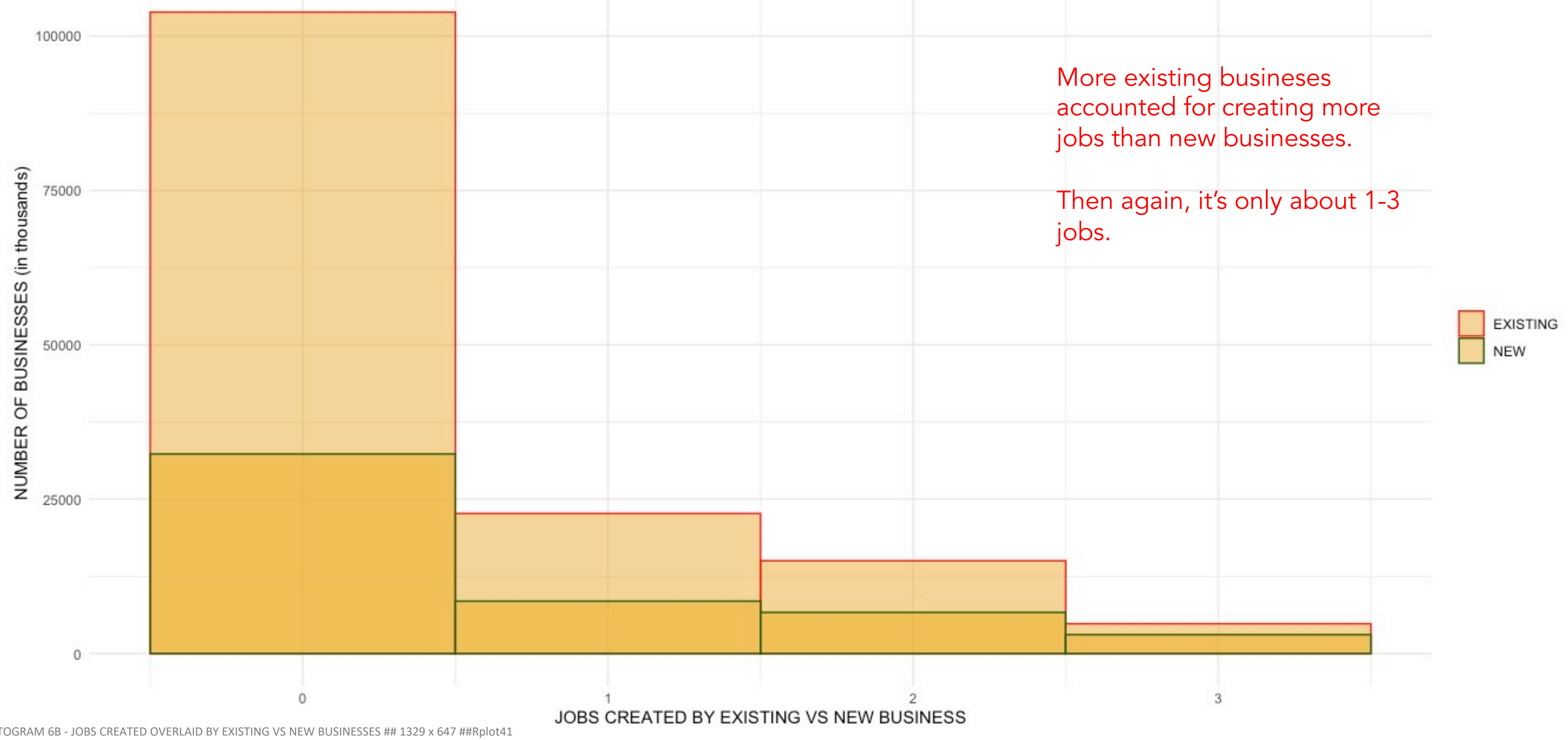
In the original dataset, some companies allegedly created thousands of jobs, including one that created 9,999.

After these outliers were removed using the IQR method, jobs created ranged only from 0 to 3.

It is not easy to create jobs, even with SBA support!

```
## HISTOGRAM 6 - JOBS CREATED ## 1329 x 647 ##Rplot40
ggplot(loandata, aes(x=jobs_created)) +
  geom_histogram(fill="darkgoldenrod2", alpha=0.5, position="identity", bins=4) +
  theme_minimal() +
  labs(x="JOBS CREATED", y="NUMBER OF BUSINESSES (in thousands)") +
  theme(legend.title=element_blank())
```

HISTOGRAM 6B : JOBS CREATED BY EXISTING VS NEW BUSINESSES

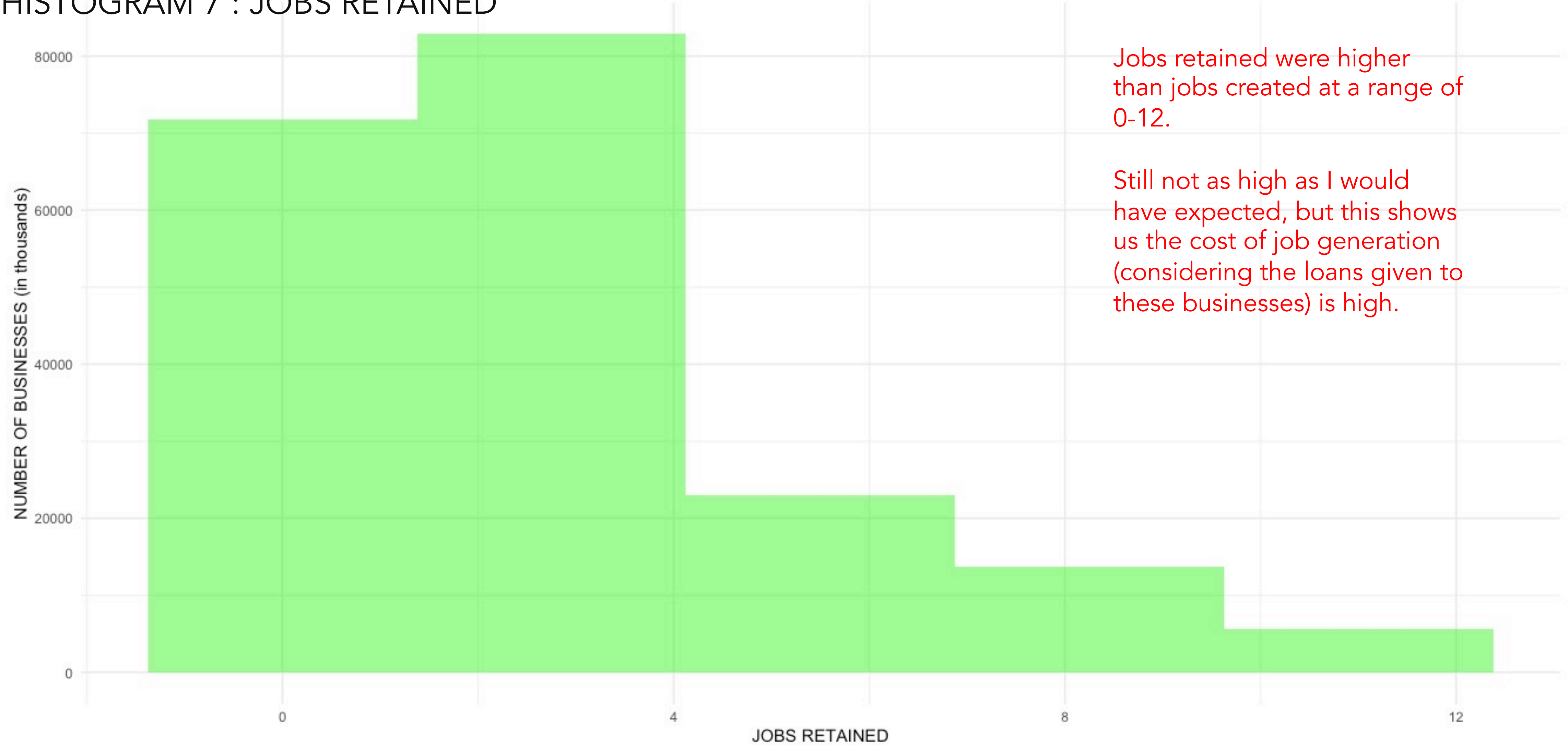


More existing busineses accounted for creating more jobs than new businesses.

Then again, it's only about 1-3 jobs.

```
## HISTOGRAM 6B - JOBS CREATED OVERLAID BY EXISTING VS NEW BUSINESSES ## 1329 x 647 ##Rplot41
ggplot(loandata, aes(x=jobs_created, color=new_or_existing_biz)) +
  geom_histogram(fill="darkgoldenrod2", alpha=0.5, position="identity", bins=4) +
  theme_minimal() +
  labs(x="JOBS CREATED BY EXISTING VS NEW BUSINESS", y="NUMBER OF BUSINESSES (in thousands)") +
  theme(legend.title=element_blank()) +
  scale_color_manual(labels = c("EXISTING", "NEW"), values = c("red", "darkgreen"))
```

HISTOGRAM 7 : JOBS RETAINED

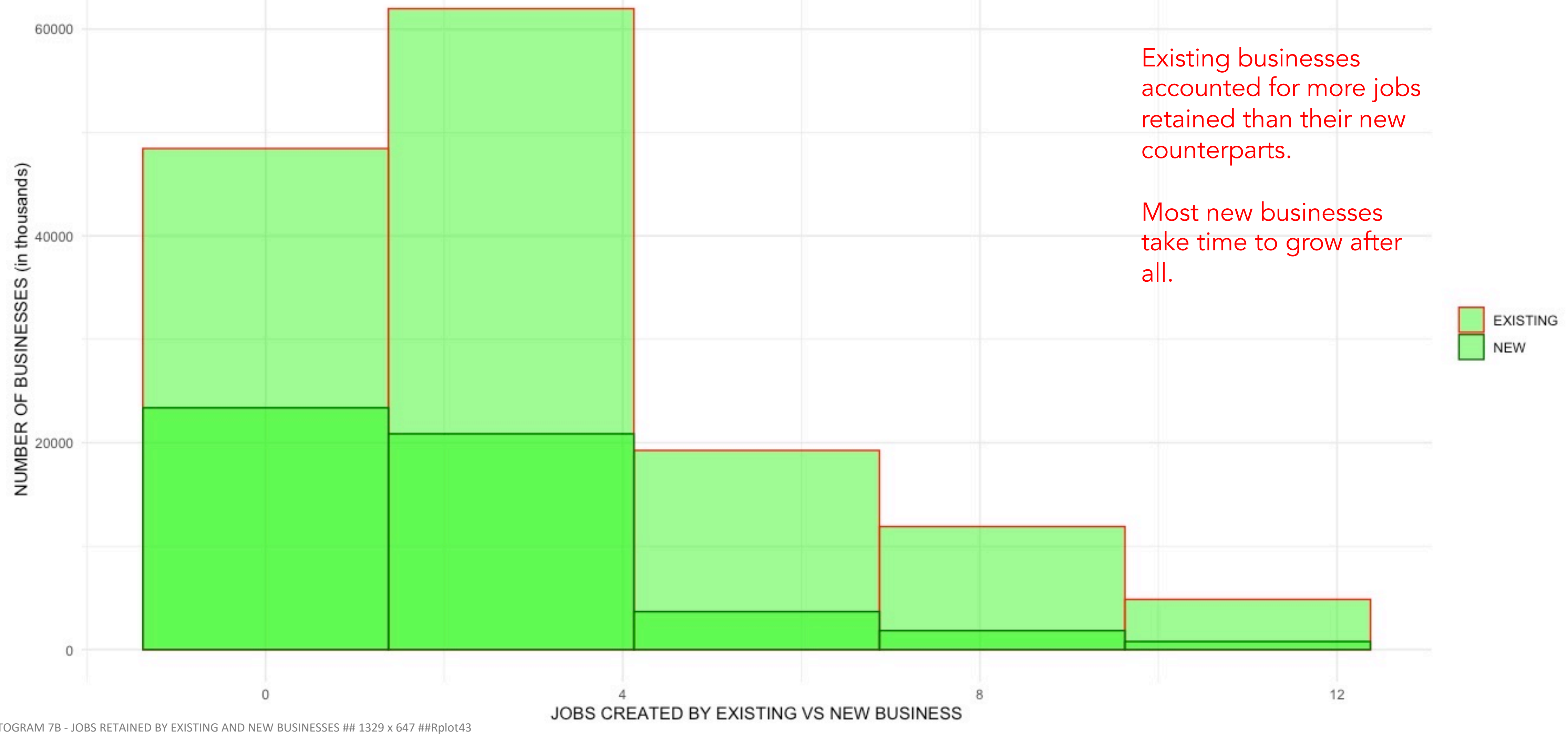


Jobs retained were higher than jobs created at a range of 0-12.

Still not as high as I would have expected, but this shows us the cost of job generation (considering the loans given to these businesses) is high.

```
## HISTOGRAM 7 - JOBS RETAINED ## 1329 x 647 ##Rplot42
ggplot(loandata, aes(x=jobs_retained)) +
  geom_histogram(fill="green", alpha=0.5, position="identity", bins=5) +
  theme_minimal() +
  labs(x="JOBS RETAINED", y="NUMBER OF BUSINESSES (in thousands)") +
  theme(legend.title=element_blank())
```

HISTOGRAM 7B : JOBS RETAINED BY EXISTING VS NEW BUSINESSES

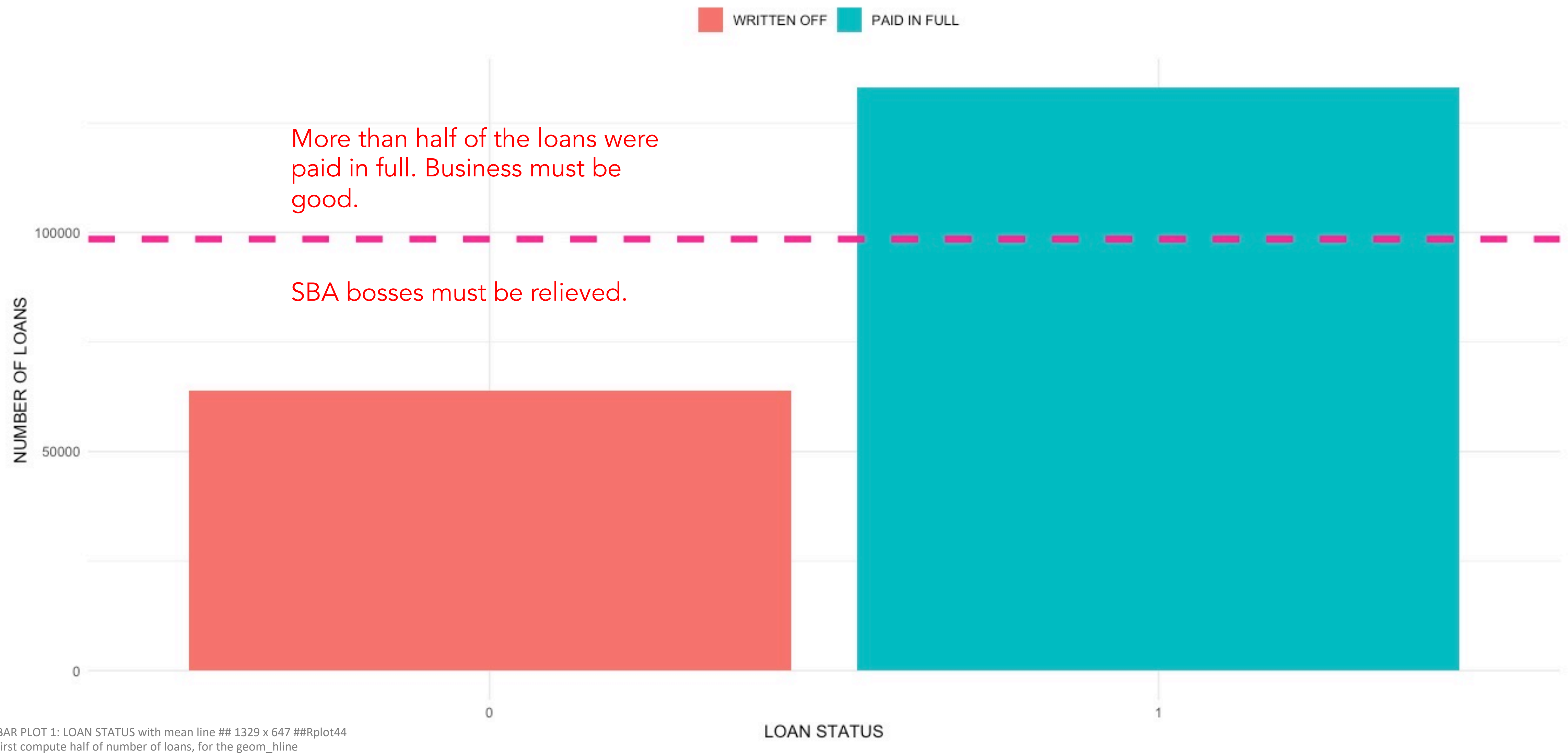


Existing businesses accounted for more jobs retained than their new counterparts.

Most new businesses take time to grow after all.

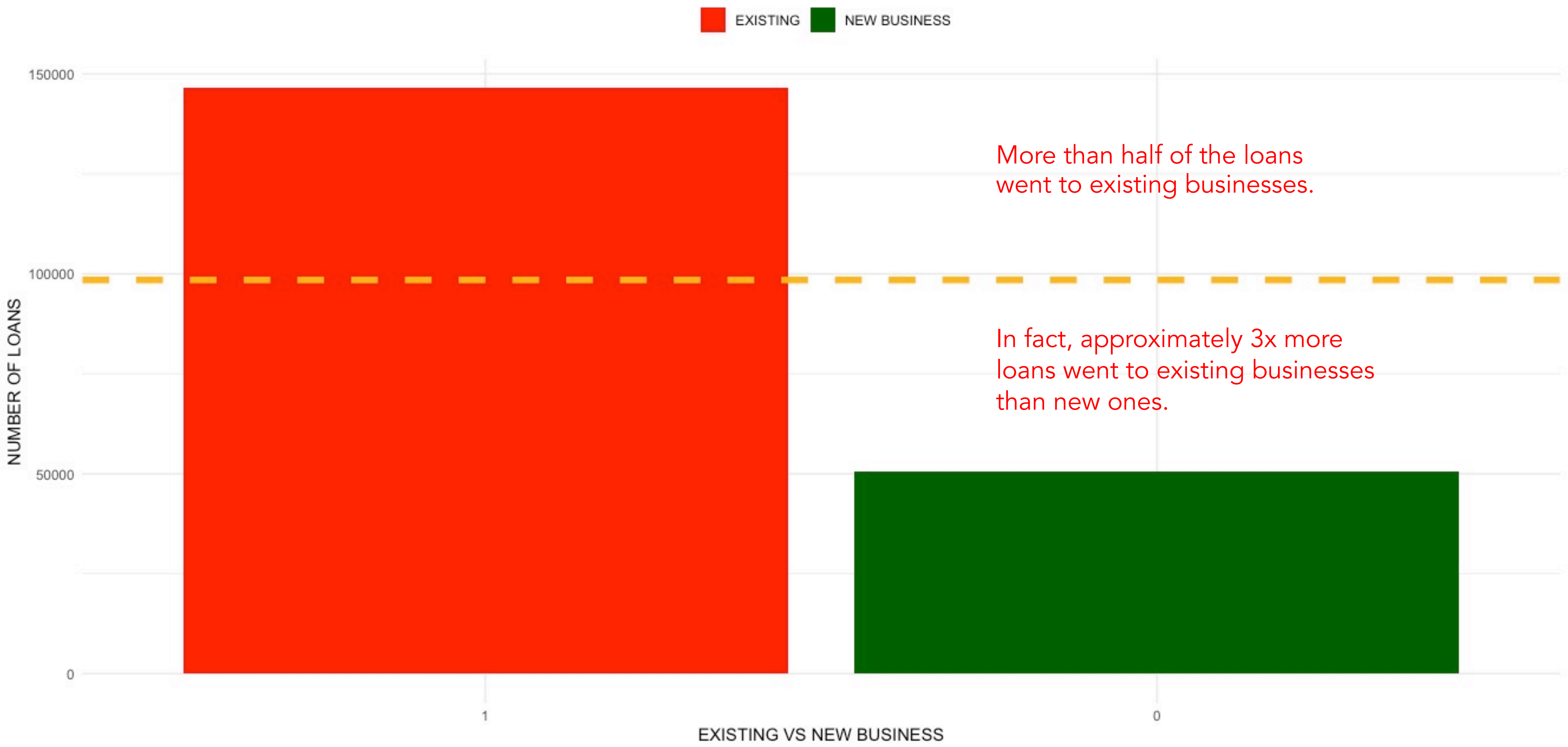
```
## HISTOGRAM 7B - JOBS RETAINED BY EXISTING AND NEW BUSINESSES ## 1329 x 647 ##Rplot43
ggplot(loandata, aes(x=jobs_retained, color=new_or_existing_biz)) +
  geom_histogram(fill="green", alpha=0.5, position="identity", bins=5) +
  theme_minimal() +
  labs(x="JOBS CREATED BY EXISTING VS NEW BUSINESS", y="NUMBER OF BUSINESSES (in thousands)") +
  theme(legend.title=element_blank()) +
  scale_color_manual(labels = c("EXISTING", "NEW"), values = c("red", "darkgreen"))
```

BAR PLOT: LOANS X LOAN STATUS (THE OTHER TARGET VARIABLE CANDIDATE)



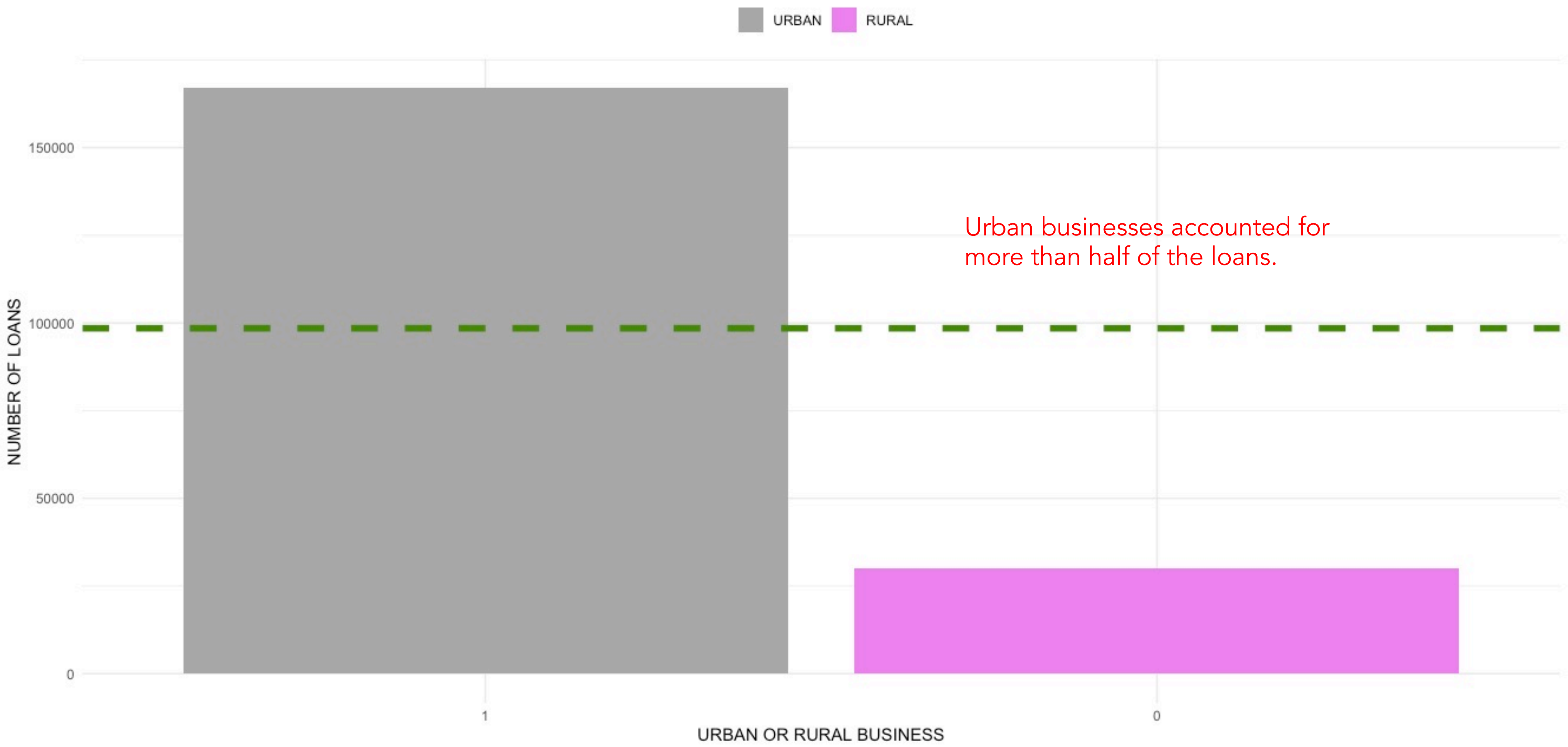

```
## BAR PLOT 1: LOAN STATUS with mean line ## 1329 x 647 ##Rplot44
## first compute half of number of loans, for the geom_hline
z<-196959/2 ## value for half of loans
## create barplot
ggplot(loandata, aes(x = loan_status, fill = loan_status)) +
  geom_bar(stat = "count") + theme_minimal() +
  labs(x= "LOAN STATUS", y="NUMBER OF LOANS") + theme(legend.position = "top")+
  geom_hline(yintercept = mean(z, na.rm=TRUE),color="deeppink", linetype="dashed", linewidth=2)+
  scale_fill_discrete(labels=c('WRITTEN OFF', 'PAID IN FULL'), name = "")
```

BAR PLOT: LOANS TO EXISTING VS NEW BUSINESSES




```
##BAR PLOT 2: LOANS FOR EXISTING AND NEW BUSINESSES with mean line  ## 1329 x 647 ##Rplot45
ggplot(loandata, aes(x = new_or_existing_biz, fill = new_or_existing_biz)) +
  geom_bar(stat = "count") + theme_minimal() +
  labs(x= "EXISTING VS NEW BUSINESS", y="NUMBER OF LOANS") + theme(legend.position = "top")+
  geom_hline(yintercept = mean(z, na.rm=TRUE),color="darkgoldenrod1", linetype="dashed", linewidth=2)+
  scale_fill_discrete(labels=c('EXISTING', 'NEW BUSINESS'), name = "") +
  scale_fill_manual(values = c("red", "darkgreen"), name = "",
              labels=c('EXISTING', 'NEW BUSINESS'))
```

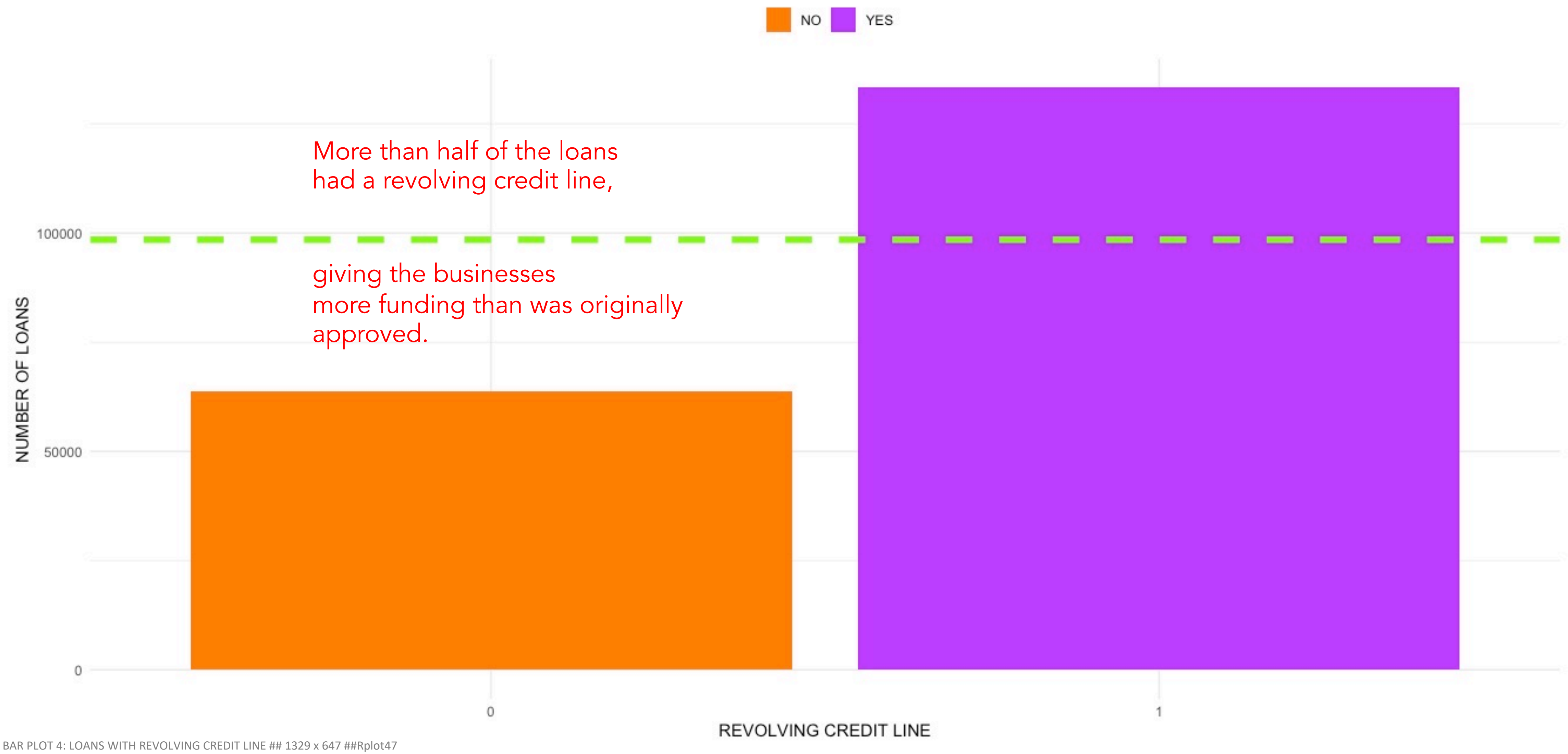
BAR PLOT: LOANS TO URBAN VS RURAL BUSINESSES



Urban businesses accounted for more than half of the loans.

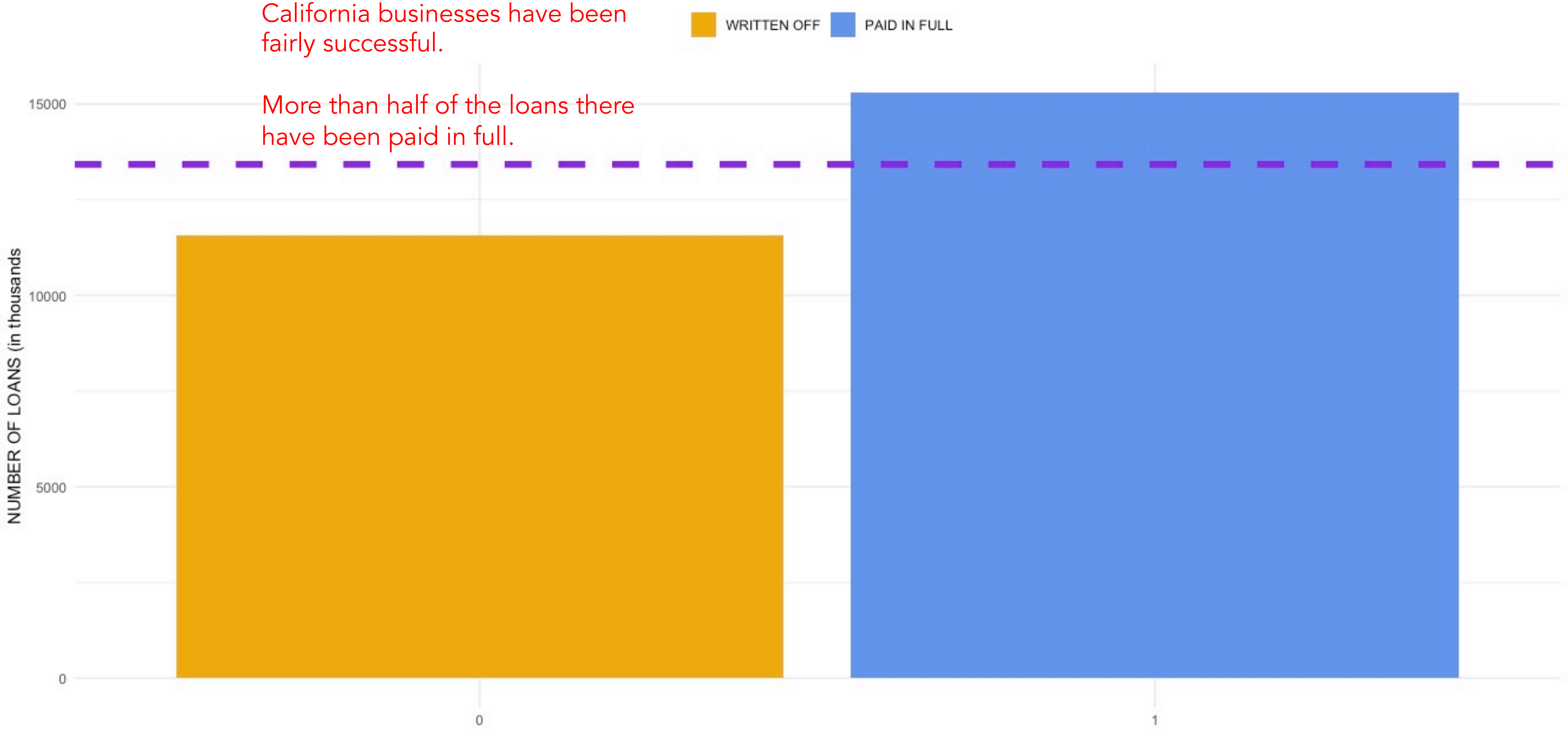
```
## BAR PLOT 3: LOANS FOR URBAN OR RURAL BUSINESSES ## 1329 x 647 ##Rplot46
ggplot(loandata, aes(x= urban_or_rural, fill = urban_or_rural)) +
  geom_bar(stat = "count") + theme_minimal() +
  labs(x= "URBAN OR RURAL BUSINESS", y="NUMBER OF LOANS") + theme(legend.position = "top")+
  geom_hline(yintercept = mean(z, na.rm=TRUE),color="chartreuse4", linetype="dashed", linewidth=2)+
  scale_fill_manual(values = c("darkgrey", "violet"), name = "",
                labels=c('URBAN', 'RURAL'))
```

BAR PLOT: LOANS WITH REVOLVING CREDIT LINE




```
## BAR PLOT 4: LOANS WITH REVOLVING CREDIT LINE ## 1329 x 647 ##Rplot47
ggplot(loandata, aes(x= revolving_credit_line, fill = revolving_credit_line)) +
  geom_bar(stat = "count") + theme_minimal() +
  labs(x= "REVOLVING CREDIT LINE", y="NUMBER OF LOANS") + theme(legend.position = "top")+
  geom_hline(yintercept = mean(z, na.rm=TRUE),color="chartreuse", linetype="dashed", linewidth=2)+
  scale_fill_manual(values = c("darkorange1", "darkorchid1"), name = "",
            labels=c('NO', 'YES'))
```

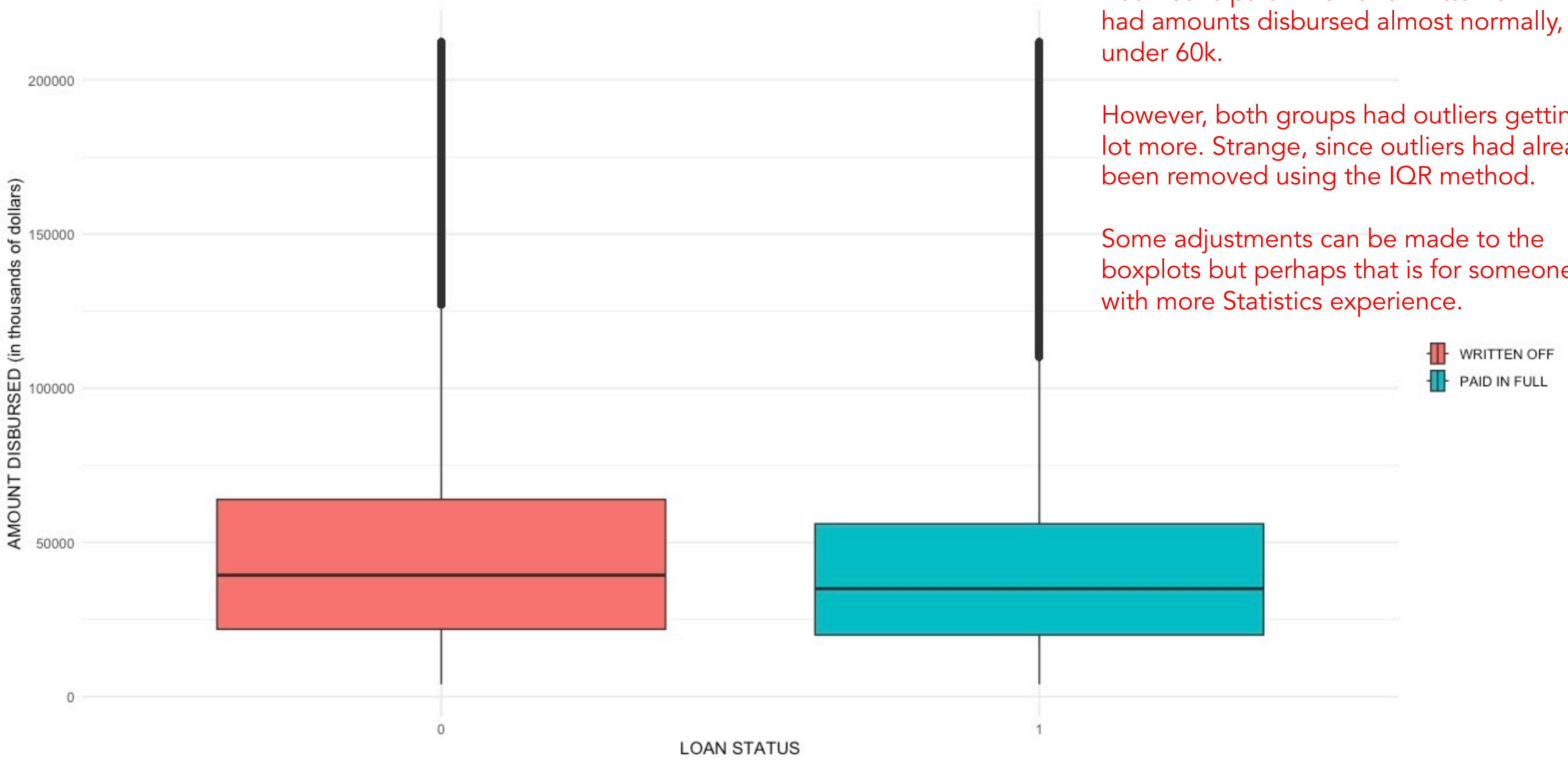
BAR PLOT: STATUS OF SMALL BUSINESS LOANS IN CALIFORNIA




```
## BAR PLOT 5: STATUS OF LOANS IN CALIFORNIA ## 1329 x 647 ##Rplot48
## first, create subset of CALIFORNIA loans
CAloans <- subset(loandata, (loandata$state == "CA"))
## 26843 observations, 18 variables
## remove state in CAloans
CAloans = subset(CAloans, select = -c(state))
## 26843 observations, 17 variables
```

```
## get mean count of CAloans for hline
c<-26843/2
## get barplot for CAloans with hline
ggplot(CAloans, aes(x = loan_status, fill = loan_status))+ geom_bar(stat = "count") + theme_minimal() +
  labs(x="STATUS OF LOANS IN CALIFORNIA", y="NUMBER OF LOANS (in thousands")+theme(legend.position = "top") +
  geom_hline(yintercept = mean(c, na.rm=TRUE),color="blueviolet", linetype="dashed", linewidth=2)+
  scale_fill_manual(values = c("darkgoldenrod2", "cornflowerblue"), name = "", labels=c('WRITTEN OFF', 'PAID IN FULL'))
```

BOX PLOT: AMOUNT DISBURSED X LOAN STATUS



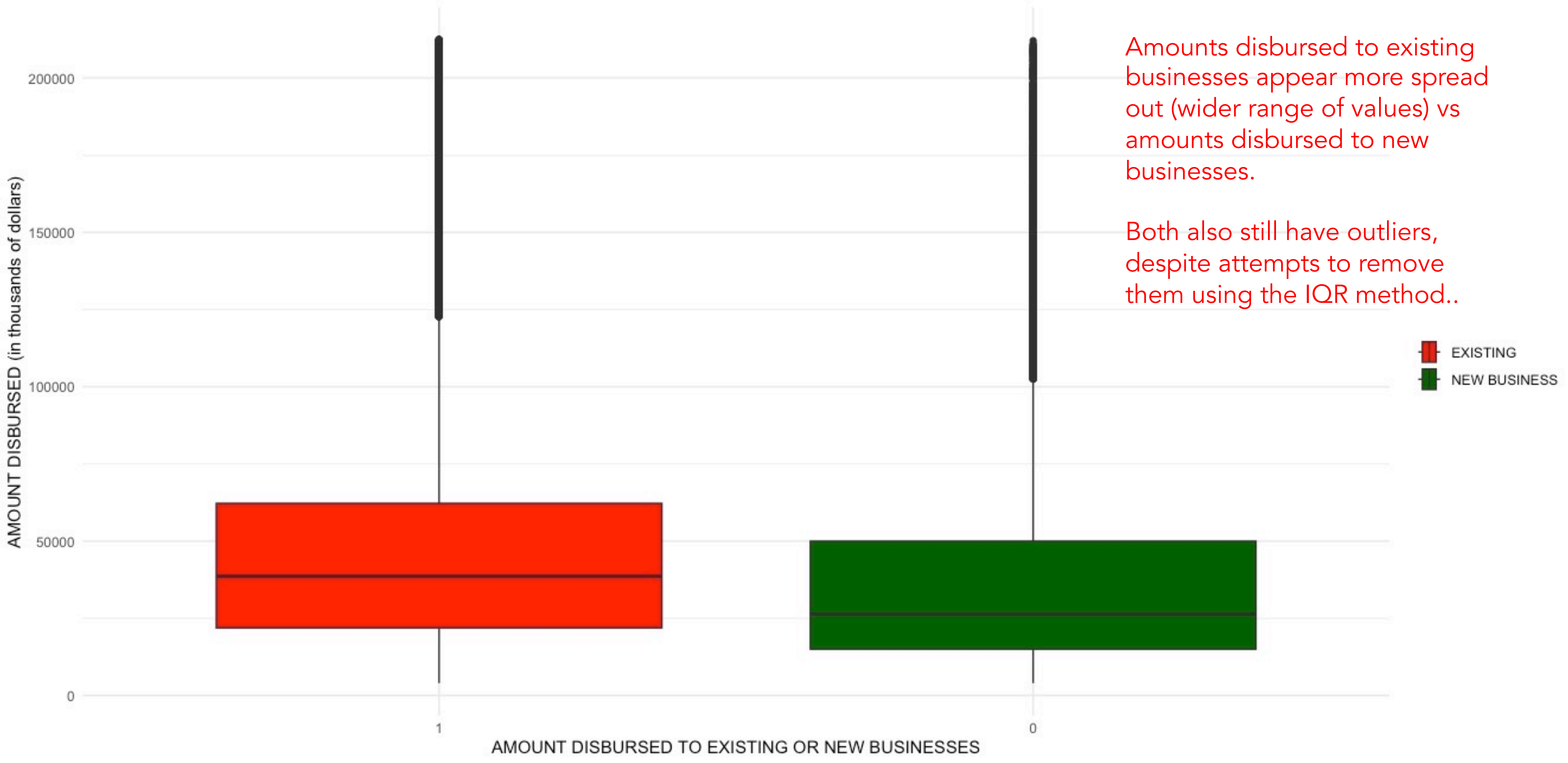
Both loans paid in full and written off had amounts disbursed almost normally, under 60k.

However, both groups had outliers getting a lot more. Strange, since outliers had already been removed using the IQR method.

Some adjustments can be made to the boxplots but perhaps that is for someone with more Statistics experience.

```
##BOXPLOT 1: AMOUNT DISBURSED BY LOAN STATUS ## 1329 x 647 ##Rplot49
ggplot(loandata, aes(x=amount_disbursed, y=loan_status,fill = loan_status))+
  geom_boxplot() + theme_minimal() + coord_flip() +
  labs(y="LOAN STATUS", x="AMOUNT DISBURSED (in thousands of dollars)") +
  theme(legend.title=element_blank()) +
  scale_fill_discrete(labels=c('WRITTEN OFF', 'PAID IN FULL'), name = "")
```

BOX PLOT: AMOUNT DISBURSED X EXISTING VS NEW BUSINESSES

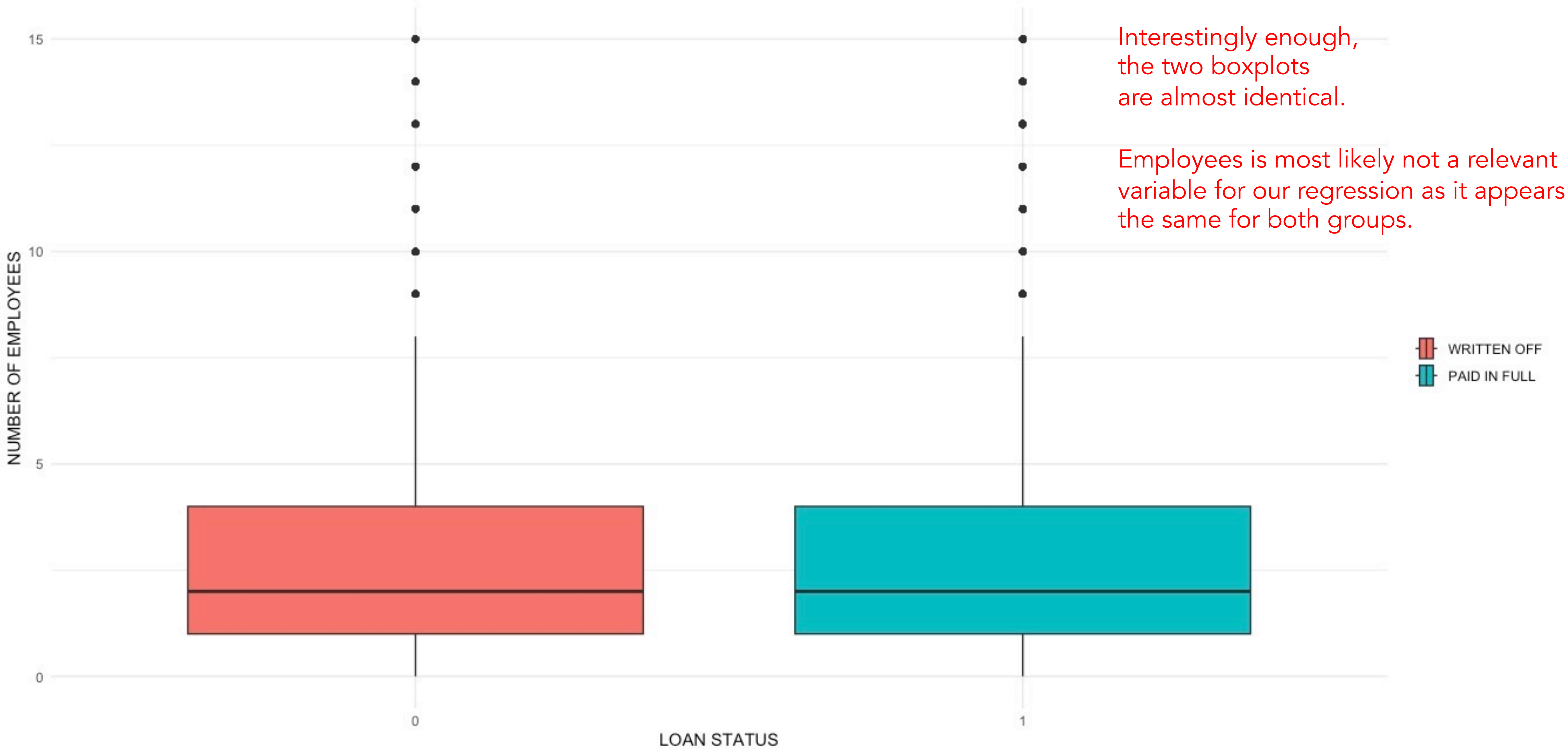


Amounts disbursed to existing businesses appear more spread out (wider range of values) vs amounts disbursed to new businesses.

Both also still have outliers, despite attempts to remove them using the IQR method..

```
##BOXPLOT 2: AMOUNT DISBURSED TO EXISTING AND NEW BUSINESSES ## 1329 x 647 ##Rplot50
ggplot(loandata, aes(x=amount_disbursed, y=new_or_existing_biz,fill = new_or_existing_biz))+
  geom_boxplot() + theme_minimal() + coord_flip() +
  labs(y="AMOUNT DISBURSED TO EXISTING OR NEW BUSINESSES", x="AMOUNT DISBURSED (in thousands of dollars)") +
  theme(legend.title=element_blank()) + scale_fill_manual(values = c("red", "darkgreen"), name = "",
              labels=c('EXISTING', 'NEW BUSINESS'))
```

BOX PLOT: EMPLOYEES X LOAN STATUS

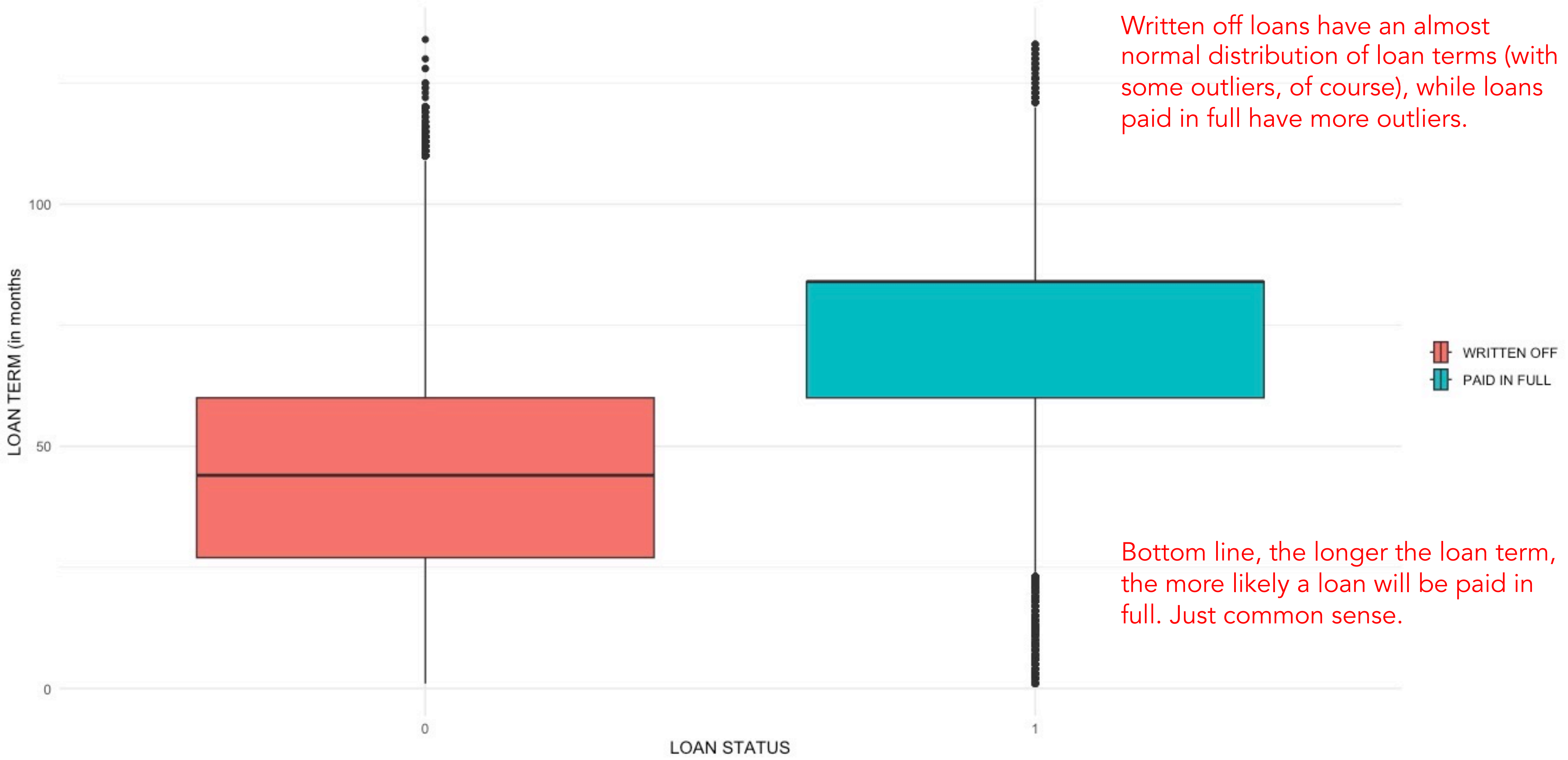


Interestingly enough,
the two boxplots
are almost identical.

Employees is most likely not a relevant variable for our regression as it appears the same for both groups.

```
##BOXPLOT 3: EMPLOYEES BY LOAN STATUS ## 1329 x 647 ##Rplot51
ggplot(loandata, aes(x=employees, y=loan_status,fill = loan_status))+
  geom_boxplot() + theme_minimal() + coord_flip() + labs(y= "LOAN STATUS", x="NUMBER OF EMPLOYEES") +
  theme(legend.title=element_blank()) +
  scale_fill_discrete(labels=c('WRITTEN OFF', 'PAID IN FULL'), name = "")
## PRACTICALLY IDENTICAL, EMPLOYEES IS LIKELY NOT A PREDICTOR VARIABLE
```

BOX PLOT: LOAN TERMS X LOAN STATUS

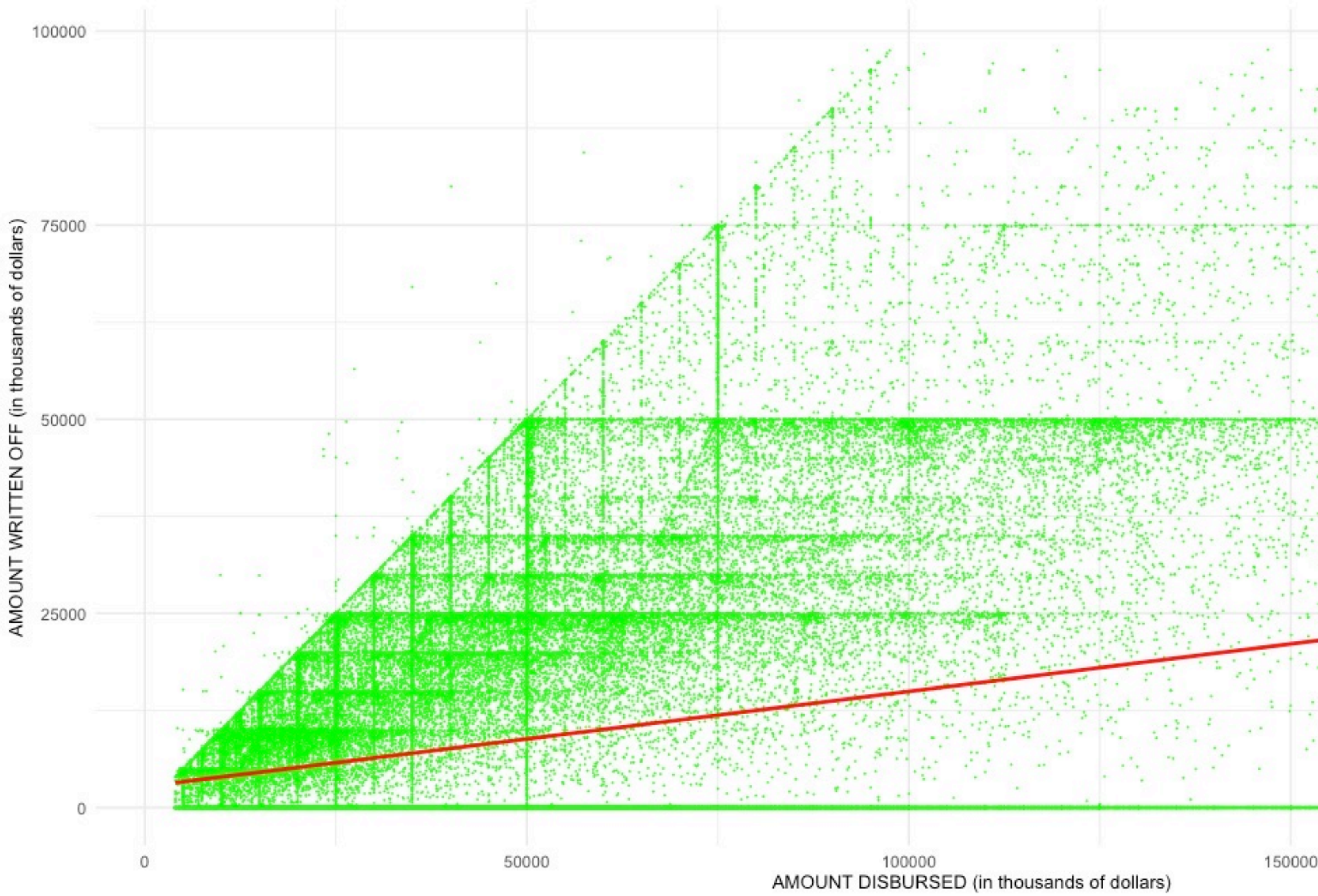


Written off loans have an almost normal distribution of loan terms (with some outliers, of course), while loans paid in full have more outliers.

Bottom line, the longer the loan term, the more likely a loan will be paid in full. Just common sense.

```
##BOXPLOT 3: LOAN TERM BY LOAN STATUS ## 1329 x 647 ##Rplot52
ggplot(loandata, aes(x=loan_term, y=loan_status,fill = loan_status))+
  geom_boxplot() + theme_minimal() + coord_flip() + labs(y= "LOAN STATUS", x="LOAN TERM (in months") +
  theme(legend.title=element_blank()) +
  scale_fill_discrete(labels=c('WRITTEN OFF', 'PAID IN FULL'), name = "")
```

SCATTERPLOT: AMOUNT WRITTEN OFF X AMOUNT DISBURSED



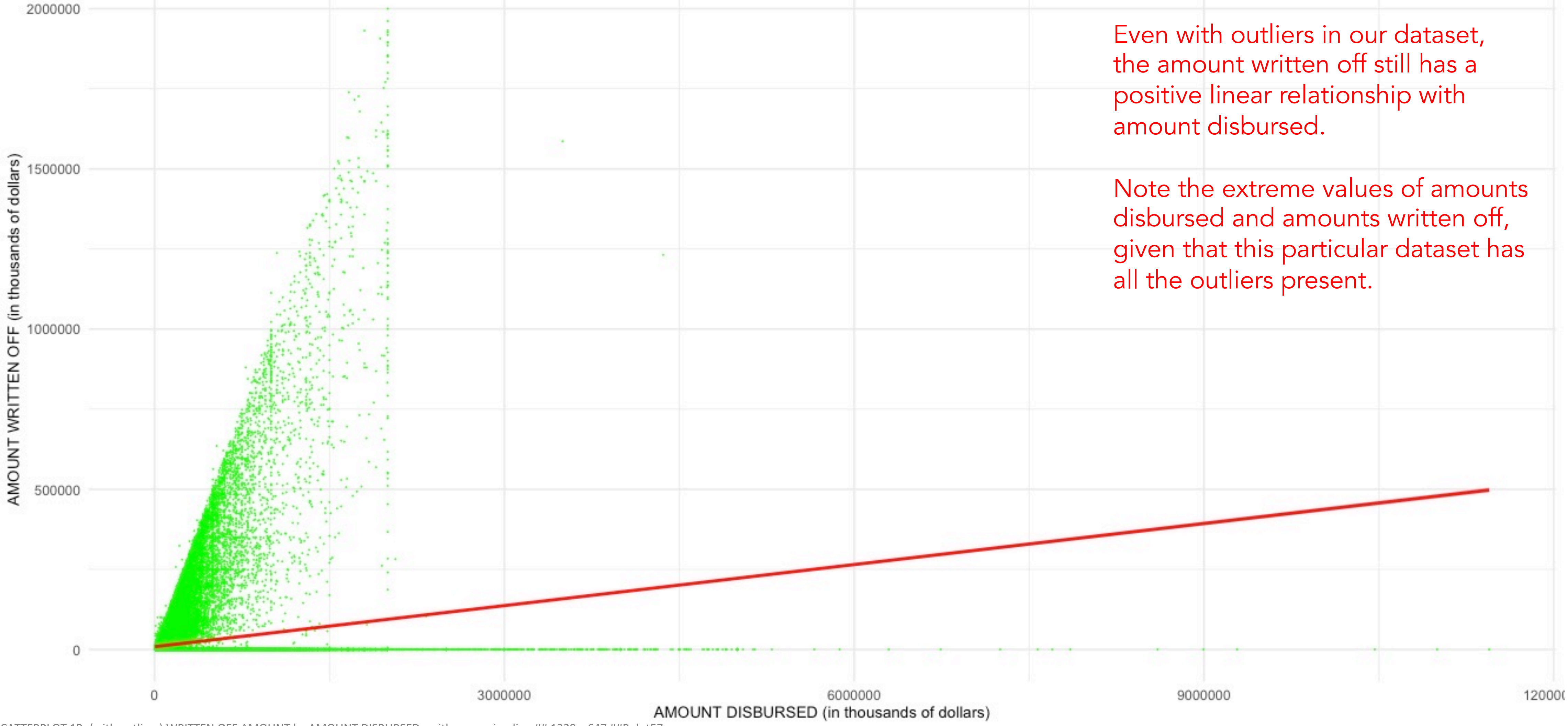
Amount written off, one potential target variable, has a positivelinear relationship with amount disbursed.

That's unfortunate. The bigger the amount disbursed, the more likely the business can't pay it off.

You'd think more cash would help a business thrive and eventually pay off its loan.

```
## SCATTERPLOT 1: WRITTEN OFF AMOUNT by AMOUNT DISBURSED,  with regression line  ## 1329 x 647 ##Rplot53
ggplot(data = loandata, aes(x = amount_disbursed, y = written_off_amount))+
  geom_point(size=.5, shape=18, color="green") + geom_smooth(method="lm",color="red") +
  theme_minimal() +
  labs(x="AMOUNT DISBURSED (in thousands of dollars)",
     y="AMOUNT WRITTEN OFF (in thousands of dollars)") +
  theme(axis.title = element_text(size = 10))
```

SCATTERPLOT: AMOUNT WRITTEN OFF X AMOUNT DISBURSED (WITH OUTLIERS)r

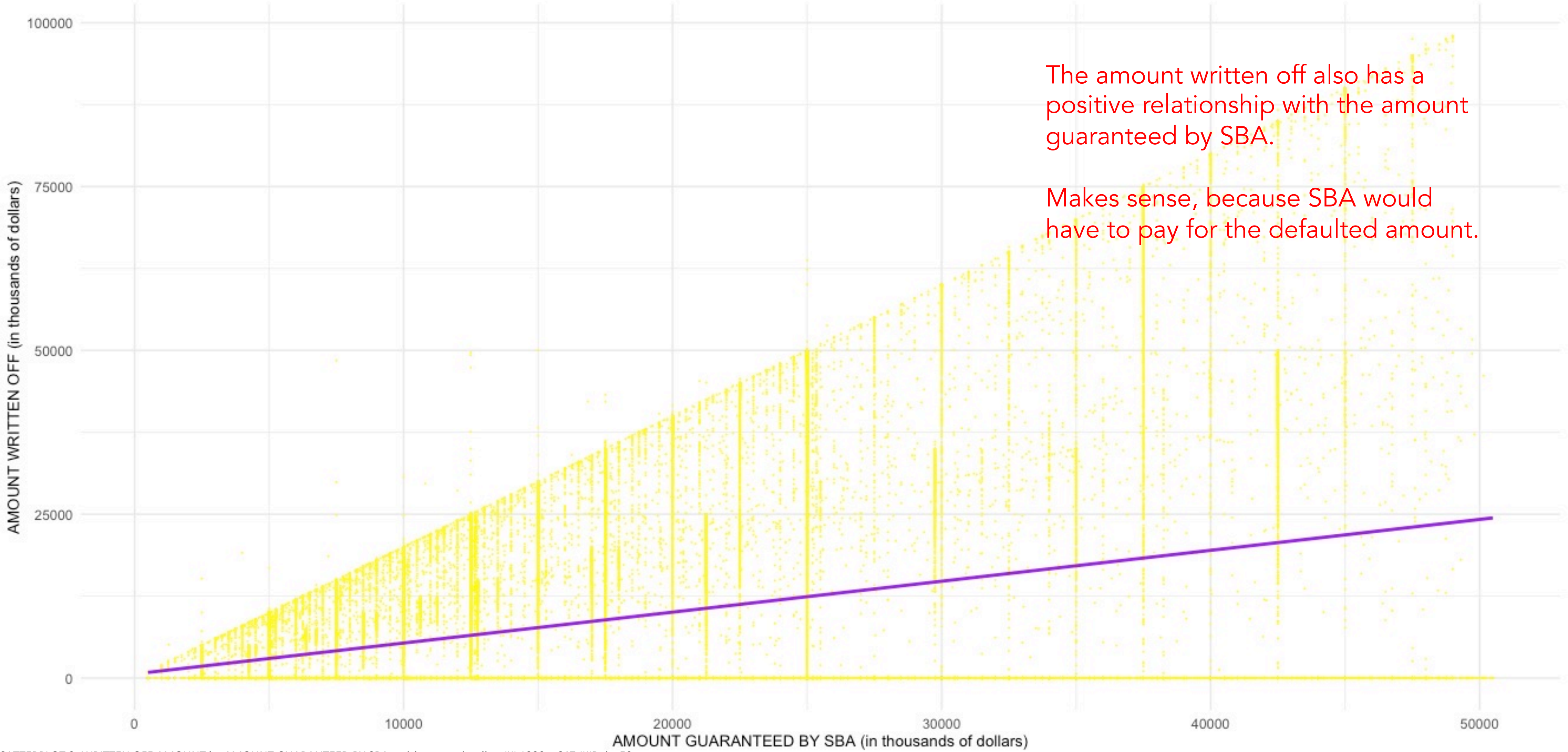


Even with outliers in our dataset, the amount written off still has a positive linear relationship with amount disbursed.

Note the extreme values of amounts disbursed and amounts written off, given that this particular dataset has all the outliers present.

```
## SCATTERPLOT 1B: (with outliers) WRITTEN OFF AMOUNT by AMOUNT DISBURSED,  with regression line ## 1329 x 647 ##Rplot57
ggplot(data = loandataWITHoutliers, aes(x = amount_disbursed, y = written_off_amount))+
  geom_point(size=.5, shape=18, color="green") + geom_smooth(method="lm",color="red") +
  theme_minimal() +
  labs(x="AMOUNT DISBURSED (in thousands of dollars)",
      y="AMOUNT WRITTEN OFF (in thousands of dollars)") +
  theme(axis.title = element_text(size = 10))
```

SCATTERPLOT: AMOUNT WRITTEN OFF X AMOUNT GUARANTEED BY SBA

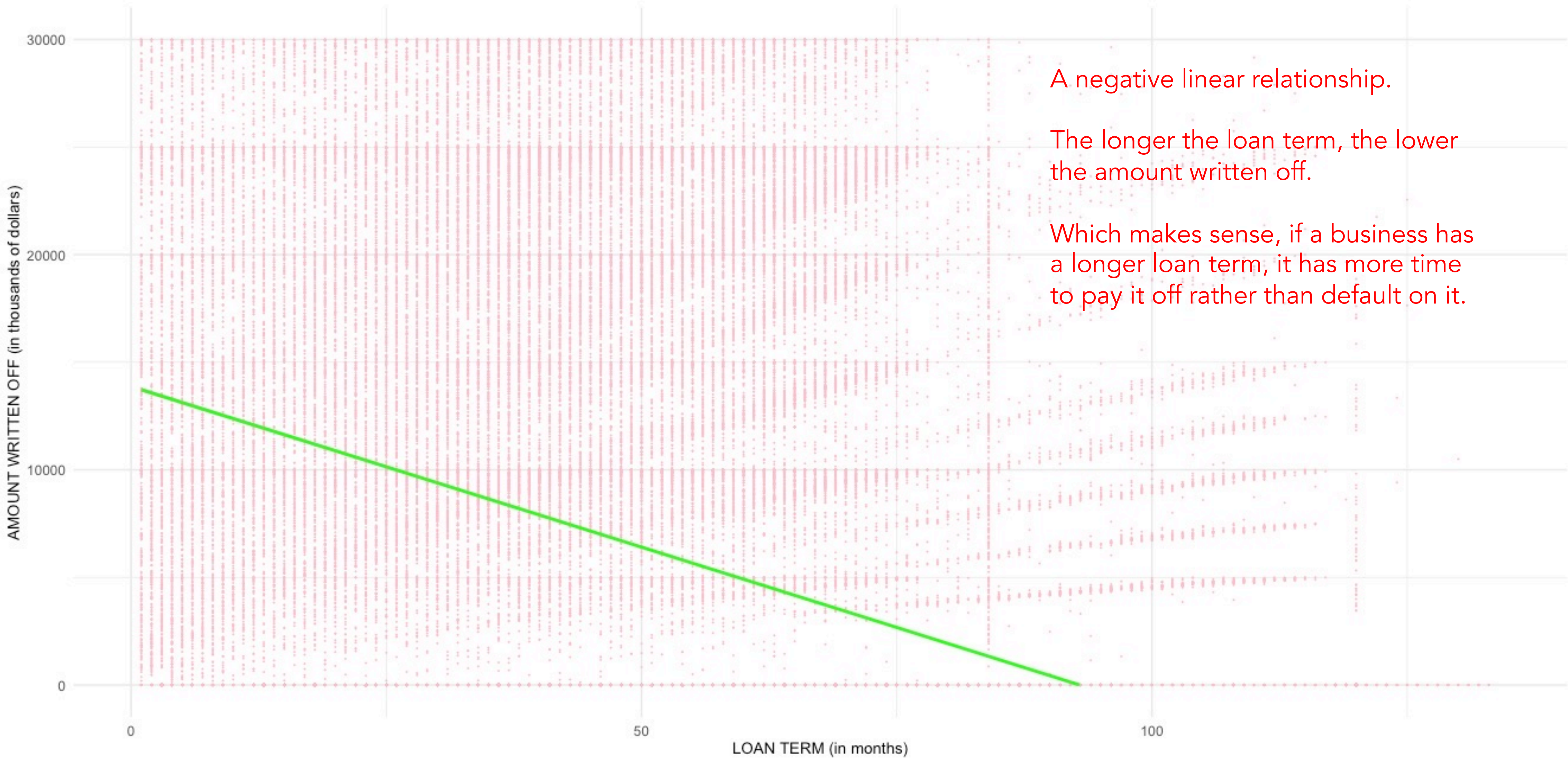


The amount written off also has a positive relationship with the amount guaranteed by SBA.

Makes sense, because SBA would have to pay for the defaulted amount.

```
## SCATTERPLOT 2: WRITTEN OFF AMOUNT by AMOUNT GUARANTEED BY SBA,  with regression line ## 1329 x 647 ##Rplot59
ggplot(data = loandata, aes(x = sba_guaranteed_amount, y = written_off_amount)) +
  geom_point(size=.5, shape=18, color="yellow") +
  geom_smooth(method="lm",color="purple") +
  theme_minimal() + labs(x="AMOUNT GUARANTEED BY SBA (in thousands of dollars)",
                       y="AMOUNT WRITTEN OFF (in thousands of dollars)") +
  theme(axis.title = element_text(size = 10))
```

SCATTERPLOT: AMOUNT WRITTEN OFF X LOAN TERM



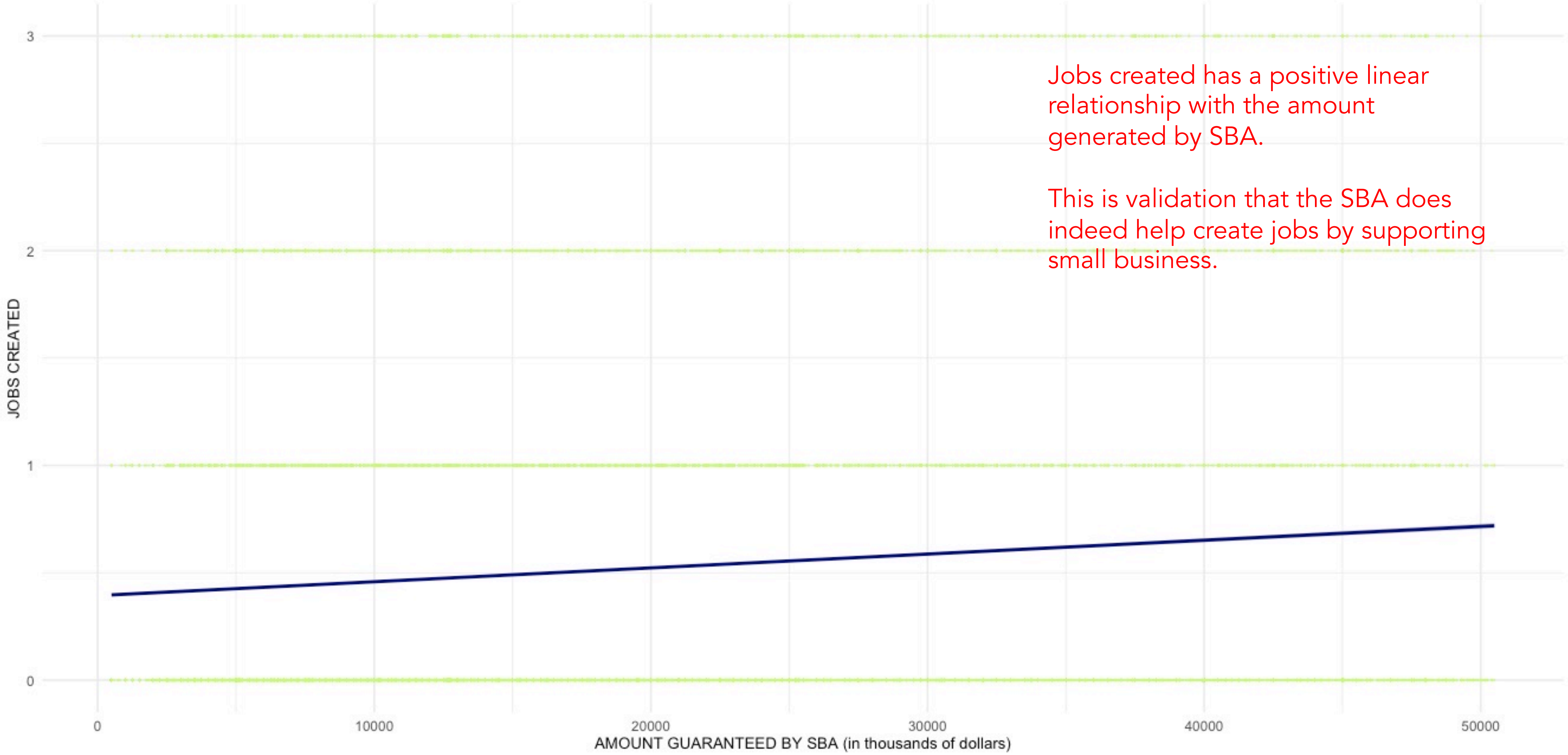
A negative linear relationship.

The longer the loan term, the lower the amount written off.

Which makes sense, if a business has a longer loan term, it has more time to pay it off rather than default on it.

```
## SCATTERPLOT 3: WRITTEN OFF AMOUNT AND LOAN TERM,  with regression line ## 1329 x 647 ##Rplot60
ggplot(data = loandata, aes(x = loan_term, y = written_off_amount)) +
  ylim(0,30000) + geom_point(size=.5, shape=18, color="pink") +
  geom_smooth(method="lm",color="green") + theme_minimal() +
  theme_minimal() + labs(x="LOAN TERM (in months)",
                    y="AMOUNT WRITTEN OFF (in thousands of dollars)") +
  theme(axis.title = element_text(size = 10))
```

SCATTERPLOT: JOBS CREATED X AMOUNT GUARANTEED BY SBA

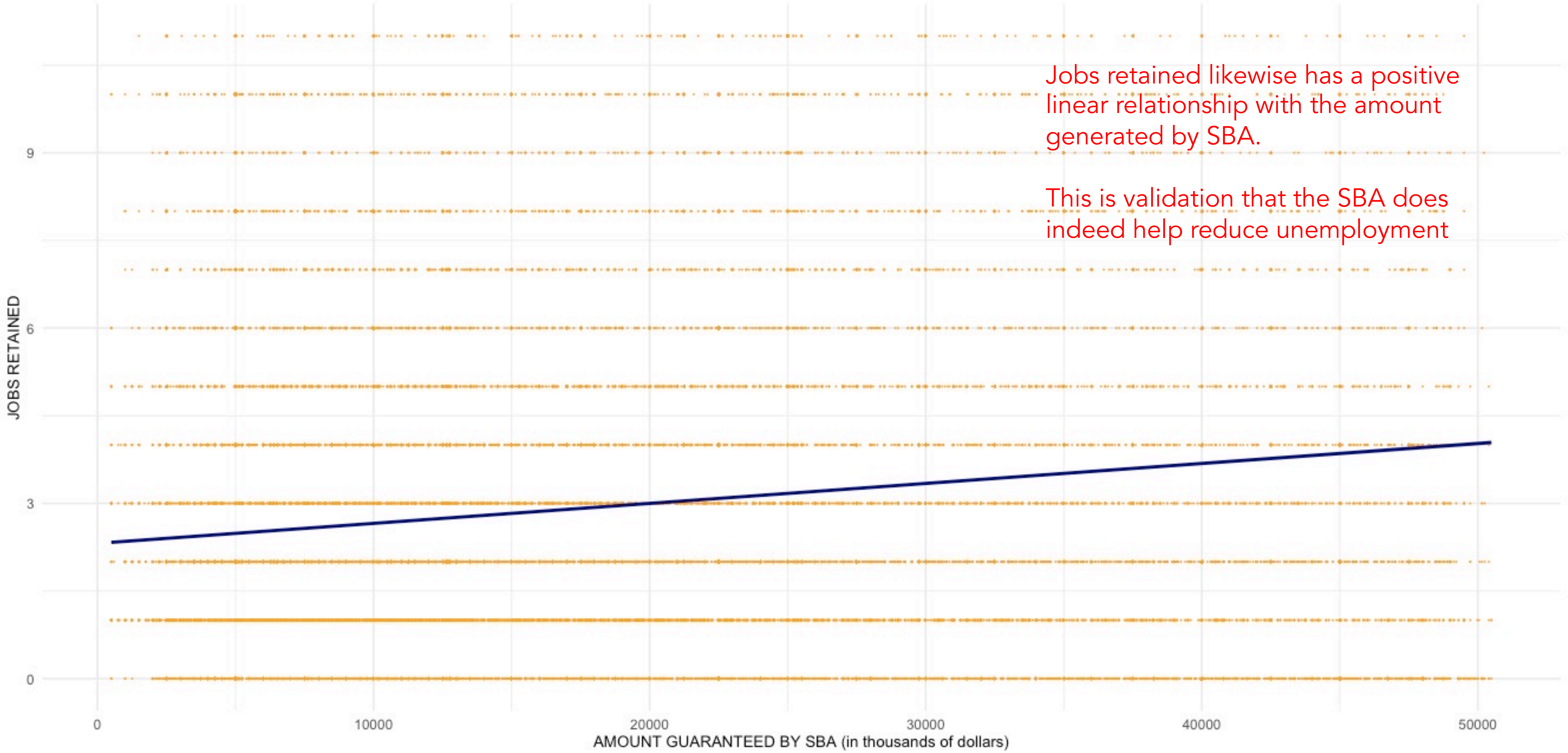


Jobs created has a positive linear relationship with the amount generated by SBA.

This is validation that the SBA does indeed help create jobs by supporting small business.

```
## SCATTERPLOT 6 JOBS CREATED AND AMOUNT GUARANTEED BY SBA,  with regression line ##1329 x 647 ## Rplot61
ggplot(data = loandata, aes(x = sba_guaranteed_amount, y = jobs_created)) +
  geom_point(size=.5, shape=18, color="darkolivegreen1") +geom_smooth(method="lm",color="navy") +
  theme_minimal() + labs(x="AMOUNT GUARANTEED BY SBA (in thousands of dollars)",
                   y="JOBS CREATED") +
  theme(axis.title = element_text(size = 10))
```

SCATTERPLOT: JOBS RETAINED X AMOUNT GUARANTEED BY SBA

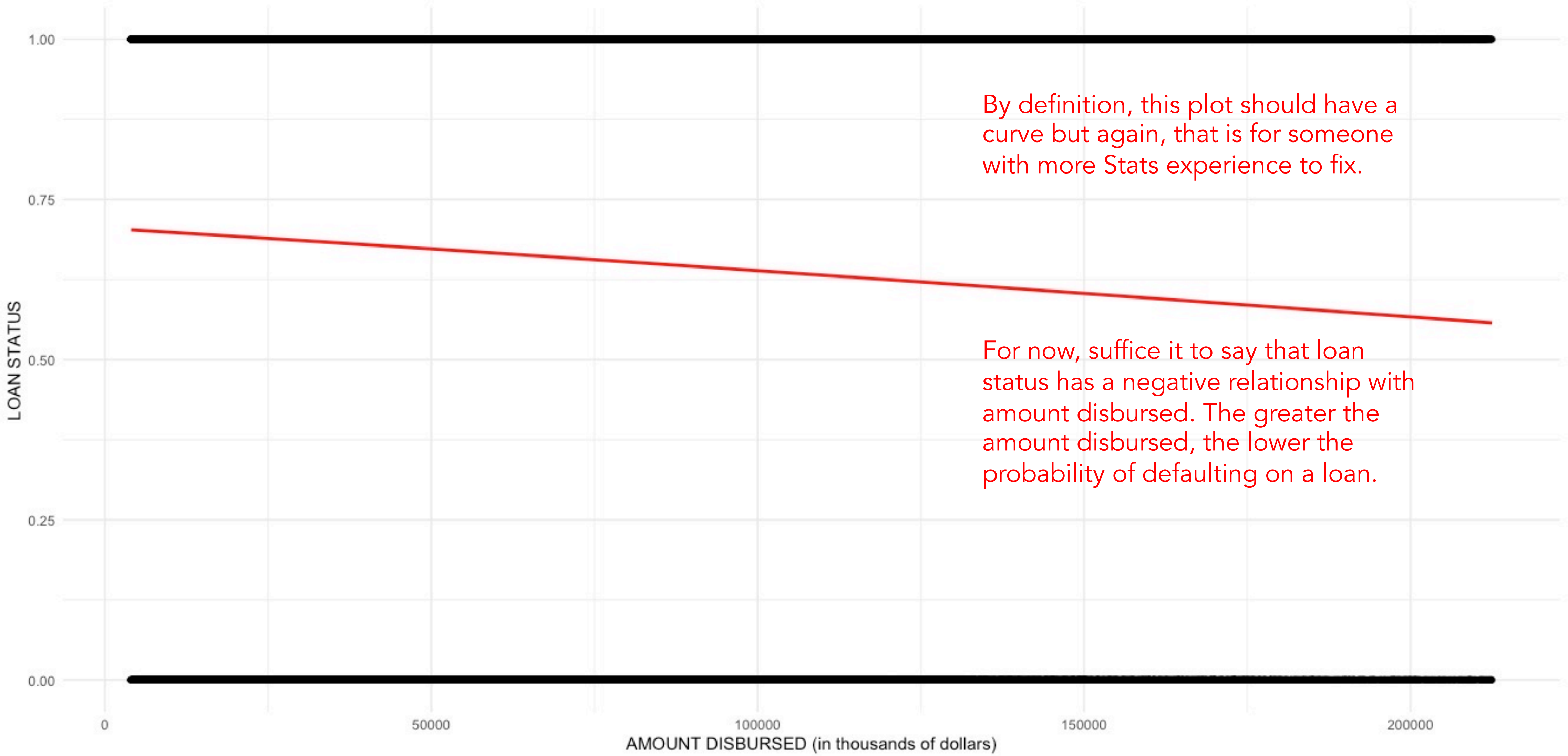


Jobs retained likewise has a positive linear relationship with the amount generated by SBA.

This is validation that the SBA does indeed help reduce unemployment

```
## SCATTERPLOT 7 JOBS RETAINED AND AMOUNT GUARANTEED BY SBA,  with regression line ##1329 x 647 ## Rplot62
ggplot(data = loandata, aes(x = sba_guaranteed_amount, y = jobs_retained)) +
  geom_point(size=.5, shape=18, color="orange") +geom_smooth(method="lm",color="navy") +
  theme_minimal() + labs(x="AMOUNT GUARANTEED BY SBA (in thousands of dollars)",
                     y="JOBS RETAINED") +
  theme(axis.title = element_text(size = 10))
```

LOGISTIC REGRESSION PLOT: LOAN STATUS BY AMOUNT DISBURSED




```
## LOGISTIC REGRESSION PLOT: LOAN STATUS by AMOUNT DISBURSED,  with regression line ## 1329 x 647 ##Rplot58
## first, turn loan_status from factor to number again for plotting
loandata$loan_status <- as.integer(loandata$loan_status)
## check levels table(loandata$loan_status)
## change levels 1/2 to 0/1 loandata$loan_status<-ifelse(loandata$loan_status==2,1,0)
## create plot with curve line, except it doesn't come out as a curve ##Rplot56 ggplot(data = loandata, aes(x = amount_disbursed, y = loan_status))+geom_point(alpha=.5) +
  theme_minimal() + stat_smooth(method="glm", se=FALSE, method.args = list(family=binomial),
                    col="red") +  labs(x="AMOUNT DISBURSED (in thousands of dollars)",    y="LOAN STATUS")
```

CREATE TRAIN AND TEST SETS

```
#R code to make this example reproducible
set.seed(1)

#Use 70% of dataset as training set and remaining 30% as testing set
sample <- sample(c(TRUE, FALSE), nrow(loandata), replace=TRUE, prob=c(0.7,0.3))
train <- loandata[sample, ]
test <- loandata[!sample, ]
```

FEATURE SELECTION

WITH CORRELATION MATRIX
FOR LINEAR REGRESSION
USING WRITTEN OFF AMOUNT AS THE TARGET VARIABLE

CORRELATION PLOT 1 WITH WRITTEN OFF AMOUNT AS TARGET VARIABLE

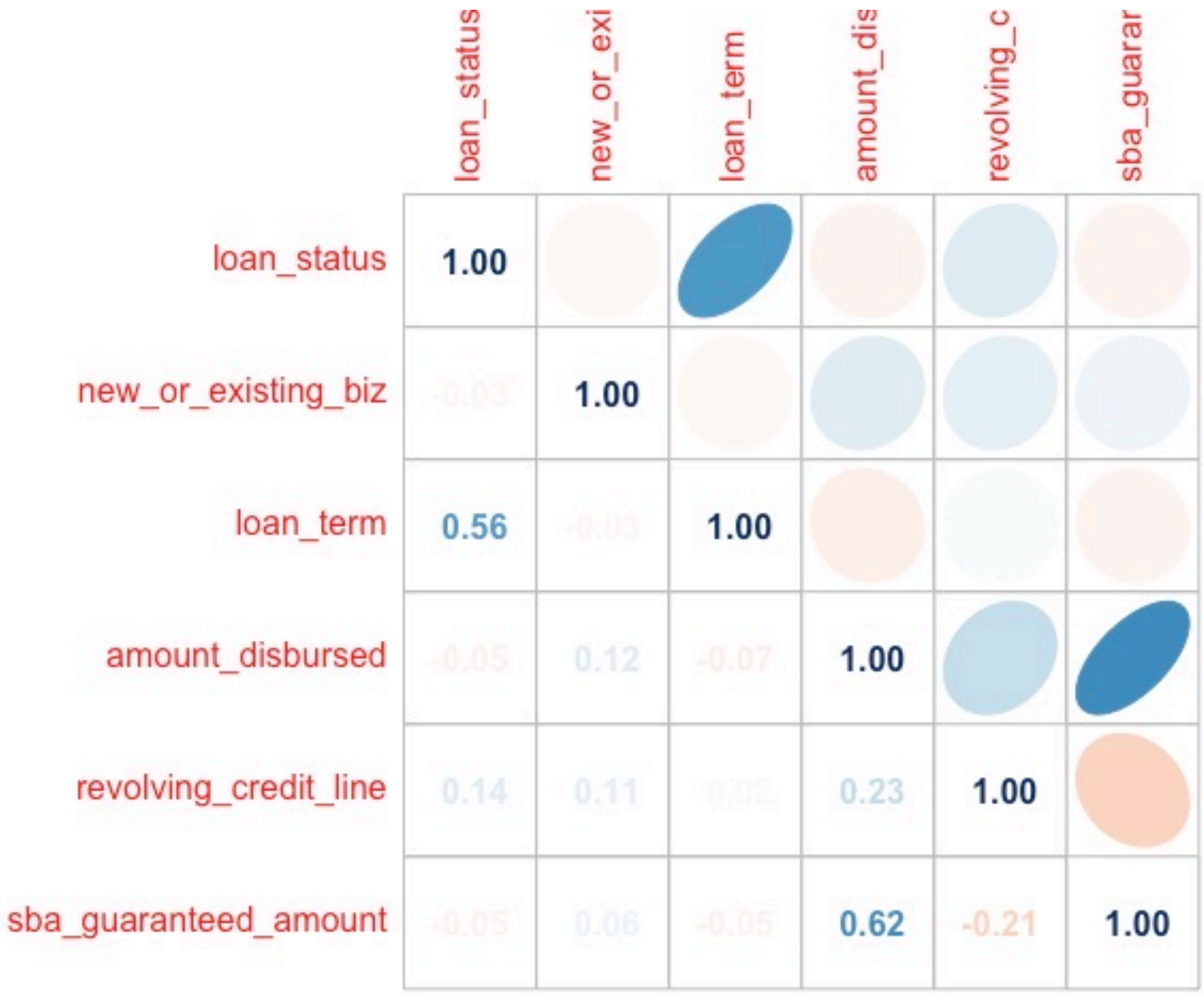
- Narrow and elongated: strong correlations
- Rounder, less elongated: weaker associations
- Darker colors: stronger correlations vs lighter colors
- Positive correlation: slope from lower left to upper right
- Negative if it slopes from the upper left to the lower right.

Negative correlation but not too high as it's rounded: written off amount and loan term

Written off amount doesn't have a strong nor weak correlation with the rest

Strong correlation because it's narrow: amount_disbursed and sba_guaranteed_amount but this is expected, the latter amount is directly related to the former

CORRELATION PLOT 2 WITH LOAN STATUS (A DUMMY VARIABLE) AS TARGET VARIABLE



- Narrow and elongated: strong correlations
- Rounder, less elongated: weaker associations
- Darker colors: stronger correlations vs lighter colors
- Positive correlation: slope from lower left to upper right
- Negative if it slopes from the upper left to the lower right.

Positive and high correlation: loan status and loan term, the opposite of the relationship between written off amount and loan term.

Loan status doesn't have a strong nor weak correlation with the rest

Strong correlation again between: amount_disbursed and sba_guaranteed_amount but this is expected, the latter amount is directly related to the former

Negative and somewhat high correlation: sba guaranteed amount and revolving credit line. Perhaps the higher the guaranteed amount, the greater the chance of getting a revolving credit line.

FITTING THE LINEAR REGRESSION MODEL

```
## R code for logistic regression model
model <- lm(formula = written_off_amount ~ amount_disbursed +  loan_term +
            sba_guaranteed_amount + new_or_existing_biz + revolving_credit_line, data=loandataM2)

## to get coefficients and other values
summary(model)
```

OUTPUT

```
Call:
lm(formula = written_off_amount ~ amount_disbursed + loan_term +
sba_guaranteed_amount + new_or_existing_biz + revolving_credit_line,
   data = loandataM2)

Residuals:     Min      1Q        Median      3Q        Max
               -39116   -7766     -1695       3667      79222

Coefficients:
                        Estimate       Std. Error    t value   Pr(>ltl)
(Intercept)             21787.012482   185.595424    117.39    <0.0000000000000002 ***
amount_disbursed        0.067712       0.001173      57.71     <0.0000000000000002 ***
loan_term               -247.453369    1.175291      -210.55   <0.0000000000000002 ***
sba_guaranteed_amount   0.269296       0.004109      65.54     <0.0000000000000002 ***
new_or_existing_biz     -1283.744956   67.813718     -18.93    <0.0000000000000002 ***
revolving_credit_line   -2069.170211   73.124312     -2830     <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13000 on 196953 degrees of freedom
Multiple R-squared:  0.2817,      Adjusted R-squared:  0.2817
F-statistic: 1.545e+04 on 5 and 196953 DF,  p-value: < 0.00000000000000022
```

P VALUES ALL BELOW 0.05, THE VARIABLES ARE NOT INDEPENDENT OF EACH OTHER

BUT RSquared and adjusted Rsquared are still closer to 0 than to 1 so this may not be a very accurate model

After exploring linear regression, I checked if logistic regression might work better

FEATURE SELECTION

WITH CHISQUARED TEST

FOR LOGISTIC REGRESSION

USING LOAN STATUS AS THE TARGET VARIABLE

IN LIEU OF CORRPLOT, A CHISQUARED TEST
TO CHECK DEPENDENCE OF CATEGORICAL VARIABLES

Z <- chisq.test(loandata$new_or_existing_biz, loandata$loan_status)

Z has a p.value less than 0.05, the two variables are not independent of each other

P <- chisq.test(loandata$loan_term, loandata$loan_status)

P has a p.value less than 0.05, (though R warned it may be incorrect) the two variables are not independent of each other

K <- chisq.test(loandata$revolving_credit_line, loandata$loan_status)

K has a p.value less than 0.05, the two variables are not independent of each other

S <- chisq.test(loandata$amount_disbursed, loandata$loan_status)

S has a p.value less than 0.05, (though R warned it may be incorrect) the two variables are not independent of each other

T <- chisq.test(loandata$sba_guaranteed_amount, loandata$loan_status)

T has a p.value less than 0.05, (though R warned it may be incorrect)the two variables are not independent of each other

FITTING THE LOGISTIC REGRESSION MODEL

```
## R code for logistic regression model
logmodel1 <- glm(loan_status~new_or_existing_biz+loan_term+amount_disbursed+revolving_credit_line+
                 sba_guaranteed_amount, family="binomial", data=train)
```

OUTPUT of summary(logmodel1)

```
Call:
glm(formula = loan_status ~ new_or_existing_biz + loan_term +
    amount_disbursed + revolving_credit_line + sba_guaranteed_amount,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1194  -0.6908   0.4079   0.6159   2.5798

Coefficients:
                          Estimate       Std. Error   z value   Pr(>|z|)
(Intercept)           -4.5507946136   0.0462008143   -98.50   <0.0000000000000002 ***
new_or_existing_biz    0.1810120743   0.0165669465    10.93   <0.0000000000000002 ***
loan_term              0.0568797886   0.0003209593   177.22   <0.0000000000000002 ***
amount_disbursed      -0.0000084539   0.0000002831   -29.86   <0.0000000000000002 ***
revolving_credit_line  1.0100790919   0.0174158369    58.00   <0.0000000000000002 ***
sba_guaranteed_amount  0.0000215387   0.0000009914    21.73   <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 173871  on 137890  degrees of freedom
Residual deviance: 124822  on 137885  degrees of freedom
AIC: 124834

Number of Fisher Scoring iterations: 4
```

P VALUES ALL BELOW 0.05, THE VARIABLES ARE NOT INDEPENDENT OF EACH OTHER BUT WE WILL CHECK FOR MULTICOLLINEARITY AS WELL

ASSESSING MODEL FIT with McFadden's R-squared

```
## R code to calculate McFadden's R-squared for logmodel1
pR2(logmodel1)["McFadden"]
```

OUTPUT

```
fitting null model for pseudo-r2
McFadden
0.2821005
```

```
## another way to get McFadden's Rsquared
with(summary(logmodel1), 1 - deviance/null.deviance)
```

OUTPUT

```
[1] 0.2821005
```

Values close to 0 : model has no predictive power
Values over 0.40 : model fits the data very well.

THEREFORE: logmodel1 fits the data, though perhaps not very well

The analysis continues…

COMPUTE VARIABLE IMPORTANCE and CHECK FOR MULTICOLLINEARITY

```
## R code to calculate variable importance
varImp(logmodel1)
```

OUTPUT

Overall	
new_or_existing_biz	11.79186
loan_term	177.25188
amount_disbursed	29.73239
revolving_credit_line	58.41831
sba_guaranteed_amount	22.03936

Higher values indicate more importance.

CHECK FOR MULTICOLLINEARITY

```
## R code to calculate VIF values of each variable to see if multicollinearity is a problem
vif(logmodel1)
```

OUTPUT

new_or_existing_biz	loan_term	amount_disbursed	revolving_credit_line
1.024287	1.014594	2.109762	1.436702

sba_guaranteed_amount
2.009353

VIF values above 5 indicate severe multicollinearity; not an issue in our model: logmodel1

USING THE MODEL TO MAKE PREDICTIONS

```
# R code to predict with "new data"
## "new data" is actually from an actual written off loan for CHICAGO BRICK UNLIMITED INC from the dataframe

probability <- predict(logmodel1,
      newdata = data.frame(amount_disbursed=51440,
                  loan_term =84,
                  sba_guaranteed_amount= 17500,
                  new_or_existing_biz= 1,
                  revolving_credit_line= 1), type = "response")
```

OUTPUT

```
1
0.9144996  ## high probability it is going to be written off, which it was
```

```
# Predict again with "new data"
## "new data" is actually from a paid in full loan  for LILY DAY GARDENS from the dataframe
probability <- predict(logmodel1,
            newdata = data.frame(amount_disbursed=60859,
                        loan_term =26,
                        sba_guaranteed_amount= 10000,
                        new_or_existing_biz= 0,
                        revolving_credit_line= 1), type = "response")
```

OUTPUT

```
1
0.2722522   ##low probability it is going to be written off and it was actually paid in full
```

CONCLUSION: Logistic regression is a more accurate model for this dataset. And based on the above, it does seem fairly accurate.
I tested the model with the test set in R.

CONCLUSION:

Logistic regression is a more accurate model for this dataset.
Moreover, based on the previous slide, it does seem fairly accurate.
I also tried the model against the test set (last code in R).

LEARNINGS:

These factors do have a bearing on whether a small business defaults on its loan or pays it off:

- A loan term
- having a revolving credit line
- being an existing or new business
- the amount disbursed for a loan
- the amount guaranteed by SBA

- Moreover, these loans supported by SBA do create and retain jobs. Not enough to make headlines however.
- But it just shows how much of an investment it takes to generate employment.

Before I end

I did find something interesting while data cleaning

Name	City	State	Bank	Industry	Urban/Rural	New/Existing	Employees	JobsCreated	JobsRetainer	LoanTerm	ApprovalDat	DisbursementDate	AmtDisbursed	BankApprov	[illegible]	[illegible]	[illegible]	BalanceGros	MIS_Status	CHGOFFamt	LowDoc
BITS & CHIPS	MIDDLETOW	OH	SPRING VALL	333911	1	1	12	0	0	237	19-Oct-04	30-Sep-05	1,070,000	[illegible]	[illegible]	[illegible]	[illegible]	996,262	P I F	0	N
GULF COAST	SCOTT	LA	HOME BANK	213112	2	2	3	0	0	129	1-Mar-06	30-Jun-06	1,080,000	1,080,000	810,000	N		827,875	P I F	0	N
Bella's Resale	Chicago	IL	CENTIER BAN	453310	1	2	8	8	0	60	16-Nov-09	1-Jan-10	90,000	120,000	60,000	N		84,617	P I F	0	N
Planet Beach	Blue bell	PA	PENN LIBERT	812199	1	1	8	0	8	12	9-Oct-09	1-May-10	75,000	75,000	37,500	[illegible]		43,127	P I F	0	N
Local Ocean	Newport	OR	WEST COAST	722110	2	1	30	5	30	60	7-Apr-09	30-Apr-09	166,826	100,000	[illegible]	Y		37,100	P I F	0	N
NAR Enterpri	LAFAYETTE	CO	JPMORGAN	722211	1	2	10	3	13	84	30-Oct-07	30-Nov-07	87,491	25,000	[illegible]	Y		25,000	P I F	0	N
Dinh V. Luor	Branson	MO	LIBERTY BAN	721110	2	1	2	0	2	287	22-Jul-09	31-Aug-09	999,950	999,950	[illegible]	N	29-Jun-11	0	P I F	634,181	N
KING OCEAN	PORT ARTHU	TX	BANK LEUMI	114112	1	1	1	0	0	154	23-Jul-99	31-Aug-99	535,000	535,000	[illegible]	N	19-Jul-02	0	P I F	522,586	N
SUN DEVIL P	CHANDLER	AZ	WELLS FARG	238220	1	1	15	1	15	1	8-Apr-05	30-Apr-05	1,169,267	350,000	[illegible]	Y	12-Jun-10	0	P I F	350,000	N
SUPERIOR SU	MIAMI	FL	OCEAN BANI	541360	1	1	24	0	[illegible]	[illegible]	[illegible]	[illegible]	500,000	500,000	[illegible]	N	2-Oct-09	0	P I F	247,715	N
PSB Smoothi	SAN JOSE	CA	U.S. BANK N.	722213	1	2	2	10	[illegible]	[illegible]	[illegible]	[illegible]	247,000	247,000	[illegible]	N	9-Dec-10	0	P I F	220,606	N
BATH JUNKIE	SUGAR LAND	TX	AMEGY BK N	327111	1	2	10	2	[illegible]	[illegible]	[illegible]	[illegible]	243,000	243,000	[illegible]	N	18-Mar-11	0	P I F	201,616	N
GAUTOLAN I	HACKETTSTO	NJ	PNC BANK, N	332721	1	1	8	0	[illegible]	[illegible]	[illegible]	[illegible]	350,000	200,000	[illegible]	Y	29-Oct-08	0	P I F	200,000	N
RESORT REST	SLATYFORK	WV	CITIZENS BK	722310	2	2	60	0	[illegible]	[illegible]	[illegible]	[illegible]	606,000	606,000	[illegible]	N	31-May-11	0	P I F	194,176	N
PERFORMAN	PASADENA	TX	BANK LEUMI	811111	1	2	5	0	[illegible]	[illegible]	[illegible]	[illegible]	190,000	190,000	[illegible]	N	14-Jun-02	0	P I F	188,415	N
B & H DESIG	LAKEWOOD	NJ	TD BANK, NA	321999	1	1	35	5	[illegible]	[illegible]	[illegible]	[illegible]	1,275,079	250,000	[illegible]	Y	14-Mar-11	0	P I F	167,186	N
RECWIC, LLC	LIMA	OH	JPMORGAN	722211	1	2	0	9	[illegible]	[illegible]	[illegible]	[illegible]	204,900	204,900	[illegible]	N	25-Jun-12	0	P I F	151,139	N
CE FIORE / TE	MONTEREY P	CA	INNOVATIVE	445299	1	2	6	0	[illegible]	[illegible]	[illegible]	[illegible]	150,000	150,000	[illegible]	N	25-Feb-13	0	P I F	150,000	N
PAIRINGS RE	NIXA	MO	LIBERTY BAN	722110	2	2	17	0	[illegible]	[illegible]	[illegible]	[illegible]	150,000	150,000	112,500	N	31-May-11	0	P I F	149,832	N
TWO TALENT	OAK RIDGE	TN	O R N L FCU	454110	1	2	3	0	0	21	30-Oct-03	31-Oct-04	300,000	300,000	225,000	[illegible]	13-Aug-08	0	P I F	147,867	N
COTTMAN TI	NEW HOPE	PA	NCB, FSB	811113	1	2	1	0	0	101	7-Jun-02	31-Oct-02	220,000	220,000	165,000	N	30-Aug-06	0	P I F	134,426	N
ACCURATE P	LOS ANGELE	CA	WELLS FARG	332813	1	1	27	1	1	21	9-Nov-07	30-Nov-07	438,099	250,000	125,000	Y	23-Jun-11	0	P I F	127,911	N
W. Ventures,	Phoenix	AZ	BANCO POPI	722211	1	2	1	7	1	110	31-Oct-08	31-Jan-09	174,000	234,000	175,500	N	13-Apr-11	0	P I F	127,387	N
Temecula HB	Temecula	CA	COMERICA B	722211	1	2	6	8	6	60	15-Feb-06	31-Mar-06	203,000	203,000	152,250	N	20-Jun-11	0	P I F	121,669	N
WASATCH IC	SALT LAKE CI	UT	FIRST UTAH	445299	2	2	20	8	[illegible]	[illegible]	[illegible]	[illegible]	340,182	250,000	125,000	Y	[illegible]	0	P I F	108,709	N
IL JIN TRADIN	NEW YORK	NY	CITIBANK, N.	339920	1	1	5	0	[illegible]	[illegible]	[illegible]	[illegible]	104,000	100,000	50,000	Y	[illegible]	0	P I F	100,000	N
CONCORD IN	PARAMUS	NJ	PNC BANK, N	336413	1	2	2	2	[illegible]	[illegible]	[illegible]	[illegible]	100,000	100,000	50,000	Y	[illegible]	0	P I F	100,000	N
CAPT'N PAUI	OAKWOOD	CA	BANK OF AM	722110	1	2	38	0	[illegible]	[illegible]	[illegible]	[illegible]	100,000	100,000	50,000	Y	24-Aug-04	0	P I F	[illegible]	[illegible]
DUPAN BAKI	FORT LEE	NJ	CITIBANK, N.	311811	1	1	5	0	[illegible]	[illegible]	[illegible]	[illegible]	260,410	100,000	50,000	Y	3-Mar-10	[illegible]	[illegible]	[illegible]	N
MGM WINE	PLYMOUTH	MN	WELLS FARG	445310	1	1	10	0	[illegible]	[illegible]	[illegible]	[illegible]	316,364	100,000	50,000	Y	17-Aug-12	0	P I F	99,587	N
NICE LOOK, I	SAN DIEGO	CA	BANK OF THI	339911	1	1	12	0	[illegible]	[illegible]	[illegible]	[illegible]	106,000	106,000	79,500	N	5-May-03	0	P I F	99,379	N
Dr. Derek K.	Huntington	CA	SOUTH CNTY	624190	1	2	1	2	[illegible]	[illegible]	[illegible]	[illegible]	198,000	100,000	50,000	Y	10-Dec-10	0	P I F	98,967	N
P4 TECHNOL	WALDORF	MD	PNC BANK, N	541519	1	1	30	0	[illegible]	[illegible]	[illegible]	[illegible]	184,750	100,000	50,000	Y	15-Sep-08	0	P I F	98,564	N
Weld Rite Ste	Norco	CA	SOUTH CNTY	811310	2	1	5	2	[illegible]	[illegible]	[illegible]	[illegible]	130,000	130,000	97,500	N	20-Jan-10	0	P I F	98,491	N
R. Gems Inc.	NEW YORK	NY	JPMORGAN	423940	1	1	8	1	[illegible]	[illegible]	[illegible]	[illegible]	356,235	100,000	50,000	Y	16-Nov-10	0	P I F	97,743	N
TRICAS MAN	DALLAS	TX	WELLS FARG	621511	1	1	7	2	[illegible]	[illegible]	[illegible]	[illegible]	165,553	100,000	50,000	Y	2-Feb-09	0	P I F	97,486	N
ST MARTIN B	NEW ORLEAN	LA	JPMORGAN	541310	1	1	5	3	[illegible]	34	29-Jan-08	29-Feb-08	255,865	100,000	50,000	Y	30-Oct-13	0	P I F	96,467	N
DUBRAY FAN	PHOENIX	AZ	NATIONAL B.	621111	1	1	3	3	0	10	23-Jul-01	31-Aug-01	100,000	150,000	75,000	Y	1-Jun-04	0	P I F	94,024	N
VAC L.L.C.	Lee's Summit	MO	COMMERCE	722110	1	2	10	0	10	90	19-Jul-07	31-Jul-07	132,600	132,600	66,300	N	23-Mar-10	0	P I F	92,611	N
KURETI ENTE	SOUTH AMB	NJ	PNC BANK, N	445291	1	1	28	1	29	3	17-Apr-07	31-May-07	192,343	100,000	50,000	Y	17-Mar-09	0	P I F	92,343	N
Modern Elec	NEW MILFOR	CT	JPMORGAN	238210	2	2	28	0	28	69	1-Dec-05	31-Dec-05	106,131	100,000	50,000	Y	29-May-08	0	P I F	90,286	N
FREEWAY TEC	WALNUT	CA	BANK OF AM	334111	1	1	6	5	6	35	10-Apr-02	30-Apr-02	90,000	100,000	50,000	Y	17-Jul-06	0	P I F	90,000	N
MARCHESSA	COLCHESTER	VT	CITIZENS BAN	442210	1	1	4	0	4	39	23-May-07	30-Jun-07	214,765	90,000	45,000	Y	27-Apr-11	0	P I F	89,460	N
LIBARDI'S AU	ROCKLEDGE	FL	CAPITAL ONE	811121	1	1	5	2	5	64	24-Oct-06	30-Nov-06	100,000	100,000	50,000	N	5-Oct-10	0	P I F	89,187	N
ANCORA DEV	MIAMI LAKES	FL	CITIBANK, N.	423390	1	1	4	0	4	62	8-Jun-06	31-Jul-06	100,000	100,000	50,000	Y	12-Aug-08	0	P I F	89,030	N
GEORGIO DI	LOS ANGELE	CA	AMERICANW	339911	1	1	6	0	0	63	3-May-01	30-Jun-01	100,000	100,000	85,000	N	10-Nov-03	0	P I F	87,531	N
Everything Sp	NEW YORK	NY	JPMORGAN	812112	1	2	2	2	2	27	16-Nov-07	30-Nov-07	101,300	101,300	50,650	N	14-Feb-13	0	P I F	86,280	N
Bamboo Kin	HOUSTON	TX	BANK OF AM	442110	1	1	5	2	5	34	18-Sep-07	30-Sep-07	100,000	100,000	50,000	N	15-Jan-10	0	P I F	85,881	N
C.O.W. USA T	NEW YORK	NY	CITIBANK, N.	315991	1	1	8	0	8	69	22-Feb-07	31-Mar-07	136,873	100,000	50,000	Y	18-Aug-08	0	P I F	83,978	N
D E A L S ETC	WATERTOWN	MA	BANK OF AM	561621	1	1	3	2	5	4	31-Aug-01	30-Nov-01	95,671	90,300	45,150	Y	17-Aug-05	0	P I F	83,872	N
SUBURBAN (	HINGHAM	MA	CITIZENS BAN	442210	1	1	3	0	3	29	10-Feb-04	31-May-04	146,100	85,000	42,500	Y	13-Apr-09	0	P I F	83,265	N

Paid in full but have a balance?

AmtDisbursed bigger than BankApproved Amount

Okay, they have RevolvingCredit

Paid in full but have a charge off amount?

83	Neurological	RUTHERFOR	NC	BANK OF AM	541690	1	1	7	0	7	38	20-Dec-05	31-Dec-05	107,400	60,000	30,000	Y	16-Nov-09	0	PIF	59,850	N
84	All Points Ins	AUSTIN	TX	BANK OF AM	541350	1	1	18	0	18	29	18-Aug-05	31-Jan-06	128,694	69,000	34,500	N	21-Sep-10	0	PIF	59,466	N
85	COTTMAN TI	SAN ANTONI	TX	BROADWAY	811113	1	2	1	0	0	24	18-Jul-01	31-Aug-01	152,431	152,431	114,323	N	5-Sep-06	0	PIF	59,462	N
86	VINCI PACIFI	DEL MAR	CA	CALIFORNIA	238910	2	1	80	10	80	12	13-Nov-03	30-Nov-03	369,544	150,000	75,000	Y	2-Jun-10	0	PIF	59,089	N
87	ERNEST JOHI	KULA	HI	CAPITAL ONE	447190	2	1	4	3	4	62	11-Jul-06	30-Sep-06	100,000	100,000	50,000	N	15-Jun-10	0	PIF	58,623	N
88	ALICE LOUISI	Dallas	TX	COMPASS B/	424310	1	2	1	1	1	54	8-Dec-05	31-Dec-05	150,000	150,000	127,500	N	12-Feb-09	0	PIF	58,306	N
89	SAM CRUM '	HEMET	CA	CAPITAL ONE	237110	1	1	3	10	3	32	21-Mar-06	30-Apr-06	100,000	100,000	50,000	N	27-Sep-10	0	PIF	57,894	N
90	BASCOM OP	SAN JOSE	CA	CITIBANK, N.	621320	1	1	1	0	1	47	30-Nov-07	31-Jan-08	242,945	100,000	50,000	Y	6-Apr-11	0	PIF	57,163	N
91	BOB'S AUTO	TRAVERSE CI	MI	IRWIN UNIOI	336322	2	1	3	0	3	19	11-Mar-03	31-Mar-03	86,250	57,500	28,750	Y	5-May-05	0	PIF	56,322	N
92	CENTRAL SQ	CENTRAL SQ	NY	CAPITAL ONE	811411	1	1	3	2	3	48	29-Jan-07	28-Feb-07	75,000	75,000	37,500	N	28-Mar-10	0	PIF	55,991	N
93	A D I DISTRIB	QUEENS VILL	NY	BANK OF AM	421690	1	1	3	0	3	16	16-May-03	30-Jun-03	66,386	60,000	30,000	Y	27-Mar-06	0	PIF	55,664	N
94	Gateway Me	CLARKSVILLE	TN	BANK OF AM	621111	1	1	19	2	21	13	4-May-05	31-May-05	171,627	100,000	50,000	Y	10-May-11	0	PIF	55,370	N
95	Hometex Ent	AUSTIN	TX	BANK OF AM	531311	1	1	9	0	9	25	26-Oct-06	30-Nov-06	138,809	100,000	50,000	Y	8-Nov-11	0	PIF	55,066	N
96	PURE PRESSI	SPRING GRO	IL	CITIZENS BAI	561790	1	1	4	0	4	33	6-Dec-05	31-Dec-05	75,000	75,000	[illegible]		5-Jun-10	0	PIF	54,513	N
97	PATOU	PHILADELPH	PA	SANTANDER	722110	1	2	25	0	0	40	7-Apr-04	30-Sep-04	75,000	55,0[illegible]	27,500	Y	[illegible]13	0	PIF	54,000	N
98	Building Mai	CHARLOTTE	NC	BANK OF AM	561720	1	1	1	0	1	28	1-Mar-05	31-May-05	155,415	[illegible],000	50,000	N	3-Mar-1[illegible]	0	PIF	53,000	N
99	LINCOLN LTI	MALVERN	PA	CITIZENS BAI	541612	1	1	4	0	4	21	13-Sep-06	31-Oct-06	53,800	56,000	28,000	N	8-Feb-12	0	PIF	52,382	N
00	ABSOLUTE FI	MAPLE GRO'	MN	HIGHLAND E	337212	1	1	55	0	0	44	20-Feb-03	30-Apr-03	100,000	100,000	85,000	N	8-Sep-06	[illegible]	PIF	51,016	N
01	MANDAREE	BISMARCK	ND	WELLS FARG	422210	1	1	8	0	0	35	28-Jun-01	31-Aug-01	300,[illegible]	300,000	225,000	N	31-May-07	0	[illegible]F	50,742	N
02	Redford Urg	Canton	MI	PNC BANK, N	621111	1	1	7	1	28	11-Jan-07	30-Apr-07	1[illegible],923	55,000	27,500	Y	2-Feb-10	0	P[illegible]	50,264	N
03	Credit Innov	FORT LAUDE	FL	BANK OF AM	561499	1	1	3	0	0	47	15-Dec-04	30-Apr-05	[illegible]49,469	50,000	25,000	N	15-May-08	0	PIF	50,000	N
04	Thomas M. S	Cutchogue	NY	TD BANK, NA	238120	1	1	3	3	3	56	5-Jul-05	31-Aug-05	50,000	50,000	25,000	Y	14-Feb-08	0	PIF	50,000	N
05	Siel Construc	BAYVILLE	NJ	BANK OF AM	238350	1	1	3	2	5	40	4-Feb-06	28-Feb-06	85,000	50,000	25,000	Y	8-Dec-09	0	PIF	50,000	N
06	PREFERRED I	LAKEWOOD	CA	CAPITAL ONE	811121	1	1	1	1	1	15	17-Jan-07	31-May-07	140,740	50,000	25,000	Y	17-Apr-13	0	PIF	50,000	N
07	EQUITY TRAD	HEMPSTEAD	NY	CAPITAL ONE	236115	1	1	15	0	15	73	15-Jan-08	29-Feb-08	51,417	50,000	25,000	Y	10-Apr-09	0	PIF	50,000	N
08	SPACE ENVIR	PACIFIC PALI	CA	CAPITAL ONE	541990	1	1	18	1	[illegible]	[illegible]	[illegible]	30-Apr-08	126,400	50,000	25,000	Y	13-Mar-10	0	PIF	50,000	N
09	CONVENTIOI	Greenwood '	CO	WELLS FARG	561920	1	1	5	[illegible]	[illegible]	[illegible]	[illegible]	1-Feb-1[illegible]	50,000	50,000	25,000	Y	15-Nov-11	0	PIF	50,000	N
10	MCM MECH	SANTA ROSA	CA	CITIBANK, N.	235110	1	1	5	[illegible]	[illegible]	[illegible]	[illegible]	30-Apr-[illegible]	50,000	50,000	25,000	Y	20-Sep-06	0	PIF	50,000	N
11	SUPER SLEEP	BRADENTON	FL	BANK OF AM	442110	1	1	20	[illegible]	[illegible]	[illegible]	[illegible]	31-Jan-[illegible]4	50,000	50,000	25,000	Y	26-Apr-07	0	PIF	50,000	N
12	TUMBLETOW	READING	PA	PNC BANK, N	713940	1	2	7	[illegible]	[illegible]	[illegible]	[illegible]	31-Jan[illegible]5	50,000	50,000	25,000	Y	28-Mar-06	0	PIF	50,0[illegible]	
13	J&J CONCRE	EL PASO	TX	JPMORGAN	235710	1	1	20	1[illegible]	[illegible]	[illegible]	[illegible]	31-Ju[illegible]02	91,894	90,000	45,000	Y	30-Aug-05	0	PIF	[illegible]	
14	GEOSOURCE	PROVO	UT	ZIONS FIRST	541618	1	1	3	[illegible]	[illegible]	[illegible]	[illegible]	30-Ju[illegible]07	78,381	50,000	25,000	Y	13-Mar-10	0	PIF	[illegible]	
15	Super Surgic	MELVILLE	NY	JPMORGAN	339112	1	1	12	[illegible]	[illegible]	[illegible]	[illegible]	30-Ap[illegible]08	99,254	50,000	25,000	Y	6-Mar-10	0	PIF	[illegible]	
16	J-Land, Inc.	OPA LOCKA	FL	BANK OF AM	452990	2	1	3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]0-No[illegible]04	50,000	50,000	25,000	Y	3-May-07	0	PIF	[illegible]	
17	CLM Associa	HIGHLANDS	CO	WELLS FARG	447110	1	1	5	[illegible]	[illegible]	[illegible]	[illegible]	31-Ju[illegible]06	84,717	50,000	25,000	Y	26-Dec-12	0	PIF	[illegible]	
18	Chapman De	SOUTH GATE	CA	BANK OF AM	423990	1	1	25	[illegible]	[illegible]	[illegible]	[illegible]	30-Jun[illegible]5	145,000	100,000	50,000	N	24-Jan-12	0	PIF	[illegible]	
19	One Love Pe	SALISBURY	NC	BANK OF AM	623220	1	1	98	4[illegible]	[illegible]	[illegible]	[illegible]	30-Sep[illegible]8	84,700	50,000	25,000	Y	16-Mar-11	0	PIF	[illegible]	
20	LUCKY PORK	SAN FRANCIS	CA	BANK OF AM	445210	1	1	8	2	8	39	5-Apr-01	30-Apr-[illegible]	75,250	50,000	25,000	Y	4-Apr-06	0	PIF	[illegible]	
21	Father & Son	Westminster	CO	CALIFORNIA	561730	1	1	5	0	0	24	10-Nov-04	30-Nov-0[illegible]	98,949	74,000	37,000	Y	6-Mar-10	0	PIF	[illegible]	
22	IRISH NATUR	BOSTON	MA	CITIZENS BAI	811111	1	1	2	0	[illegible]	[illegible]	[illegible]	[illegible]1-May-0[illegible]	49,522	50,000	25,000	N	21-Nov-11	0	PIF	[illegible]	
23	DIVENTUS H	HONOLULU	HI	AMERICAN S	812910	1	1	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]1-Jan-09	94,549	50,000	25,000	Y	25-Apr-11	0	PIF	[illegible]	
24	BURKE WOO	:NORTH HOL	CA	WELLS FARG	236118	1	1	8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]0-Nov-04	125,314	50,000	25,000	Y	7-Mar-12	0	PIF	[illegible]	
25	Rami Holdin	Grand Prairie	TX	WELLS FARG	611699	1	2	0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]1-Aug-06	51,000	51,000	25,500	N	29-Apr-09	0	PIF	[illegible]	
26	GRS CONSUI	LOUISVILLE	KY	CITIBANK, N.	541618	1	1	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]0-Apr-07	104,120	50,000	25,000	Y	22-Jan-10	0	PIF	[illegible]	
27	A B B SANIT	SAN DIMAS	CA	BANK OF AM	561990	1	1	11	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]1-Dec-03	[illegible]50,000	100,000	50,000	Y	19-May-08	0	PIF	49,14[illegible]	[illegible]
28	J & M Del Es	BRONX	NY	BANK OF AM	488510	1	2	3	[illegible]	[illegible]	[illegible]	[illegible]	28-Feb-06	[illegible],000	50,000	25,000	N	14-May-07	0	P[illegible]	49,100	N
29	SUZANNE LA	PITTSBURGH	PA	CITIZENS BAI	621610	1	1	5	[illegible]	[illegible]	[illegible]	[illegible]	28-Feb-06	100,[illegible]	100,000	50,000	N	10-Aug-12	0	[illegible]F	49,092	N
30	GREG LEE FO	YANCEYVILLE	NC	CAPITAL ONE	441110	2	1	7	[illegible]	[illegible]	[illegible]	[illegible]	31-Jul-05	108,84[illegible]	50,000	25,000	Y	6-Jan-09	[illegible]	PIF	48,990	N
31	Perfect Pear	Chicago	IL	PNC BANK, N	453998	1	2	1	[illegible]	[illegible]	[illegible]	[illegible]	31-Jan-06	108,517	50,000	25,000	Y	28-May-08	0	PIF	48,959	N
32	EDUARDO RI	VINELAND	NJ	PNC BANK, N	442110	2	2	2	[illegible]	[illegible]	[illegible]	[illegible]	31-Jan-07	75,000	[illegible]0,000	25,000	N	16-Feb-10	0	PIF	48,886	N
33	TECH TECH II	KEY WEST	FL	BANK OF AM	423430	2	1	1	0	[illegible]	[illegible]	[illegible]	30-Nov-03	50,000	50,[illegible]	25,000	Y	12-A[illegible]	0	PIF	48,685	N
34	Phuket Thai	SEATTLE	WA	BANK OF AM	722110	1	1	4	4	4	55	14-Jan-08	31-Jan-08	94,333	50,000	[illegible]	Y	[illegible]-Aug-10	0	PIF	48,666	N

But wait, some have NO Revolving Credit yet still got more

What bank gives more?
We want to know!!!

INNOCENT MISTAKES OR FRAUD HIDDEN IN PLAIN SIGHT?

THANK YOU!