STAT 385 Homework Assignment 04

Tianli Ding

Grader Comment

Efficiency:

The code and document is written with lots of logical thought about the tasks with few unnecessary or unnecessarily verbose lines and is written in a least brute force way.

Correctness:

The files run with no errors and the coding is written in a most accurate way.

Documentation:

The explanations and comments contain reasonable clarity and exhibit mid-tier (to be expected) fundamentals of usefulness, grammar, smoothness, and structure.

Beauty:

The document is most visually appealing and has the strongest readability. The coding output and visualizations are very attractive.

Please note that the summary above applies to the completed work. If the submission was incomplete, the ultimate mark for the submission will be scaled down proportionally.

Please take care to ensure that one's repo includes the image files that are included in one's .Rmd file. Otherwise, the graded file that is returned to the student may not contain the images.

You may find that your returned work (.Rmd and .pdf) has had sections removed. This is only for accelerating the grading process by removing code chunks that are impeding knitting. If you believe that you were incorrectly penalized for missing work, please do contact the TA and/or professor.

Please see comments below for specifics.

HW 4 Problems

Below you will find problems for you to complete as an individual. It is fine to discuss the homework problems with classmates, but cheating is prohibited and will be harshly penalized if detected.

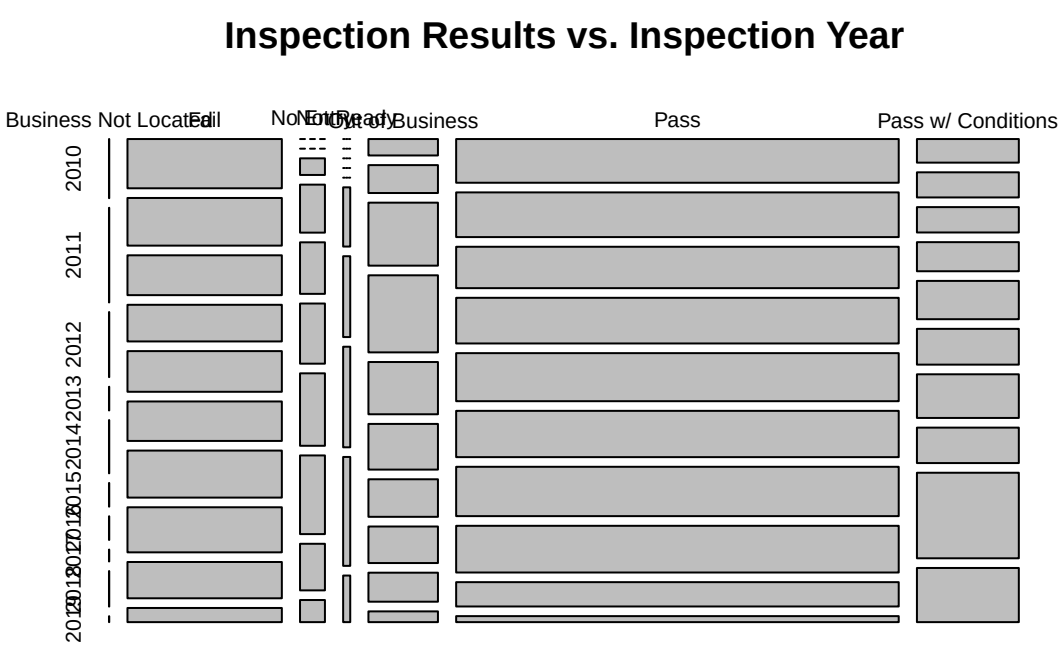
For all problems below, only use base R plotting methods. Do not use tidyverse plotting methods.

1. Read https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5/ for more information about the Chicago Food Inspections Data and do the following:

a. create a visualization (plot) of at least two variables (excluding ID or label variables) using this dataset

```
cfid = read.csv("https://uofi.box.com/shared/static/5637axblfhajotail80yw7j2s4r27hxd.csv")
```

```
new <- cfid
new$Inspection.Date <- as.Date(new$Inspection.Date, "%m/%d/%Y")
new$year <- format(new$Inspection.Date, "%Y")
t1 = table(new$Results, new$year)
plot(t1, main = "Inspection Results vs. Inspection Year", sub("Years 2010-2019"))
```



In this question, I chose two variables, Inspection Date and Inspection Results. I use year and the results generated a plot showing the number of different results of inspected restaurants versus the year.

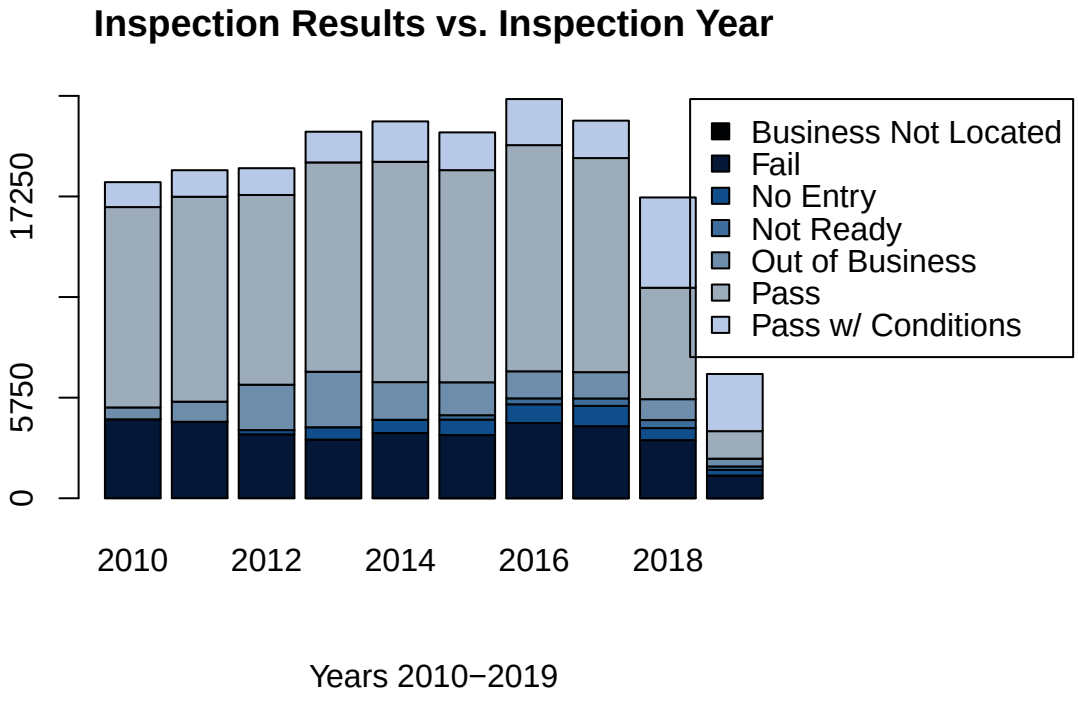
b. explain what is good and what is bad about the visualization

The pros of this visualization is that it can be easily to identify "Pass" takes the largest percentage in the results throughout each year. And we could also tell the difference between each result within each

year. However, there are still some cons. First of all, it is hard to see some part of the text, since some of the section is too small, thus leading to overlaps of the text. Besides, we cannot know the difference of the total number of results. Therefore, more could be done to better virtualize the plot.

c. show a substantially improved visualization

```
levs <- levels(factor(new$Results))
par(mar=c(8, 4, 4, 10), xpd=TRUE)
barplot(t1, axes = FALSE, col = c('#000102','#041737',"#104E8B",'#3F6D9B','#6E8DAB','#9DACBB','#B7C9E6'
title("Inspection Results vs. Inspection Year", "Years 2010-2019")
axis(side = 2, at=c(0, 5750,11500, 17250, 23000), xpd=TRUE)
par(mar=c(6, 4, 4, 4), xpd=TRUE)
legend("topright", inset=c(-0.4,0), legend=levs, fill=c('#000102','#041737',"#104E8B",'#3F6D9B','#6E8DA
```



d. describe the improvement and why the improved plot in **part 1c** helps the reader or viewer more than the original plot in **part 1a**.

In this plot, I use a barchart to better virtualize the data I choose (data and results). In the barplot, it kept the advantages from the plot in 1a, which shows "Pass" has the most frequency in results. Towards the cons from last plot, from the barplot, the total amount of inspected results can be compared. We can see the 2018 has the lowest number of inspected results throughout 2010-2018(not including 2019). The graph is more clear and reasonable by giving legend and color, and reasonable custom axis.

2. Read https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5/ for more information about the Chicago Food Inspections Data and do the following:

a. create a table of 3 or more descriptive statistics of your choice

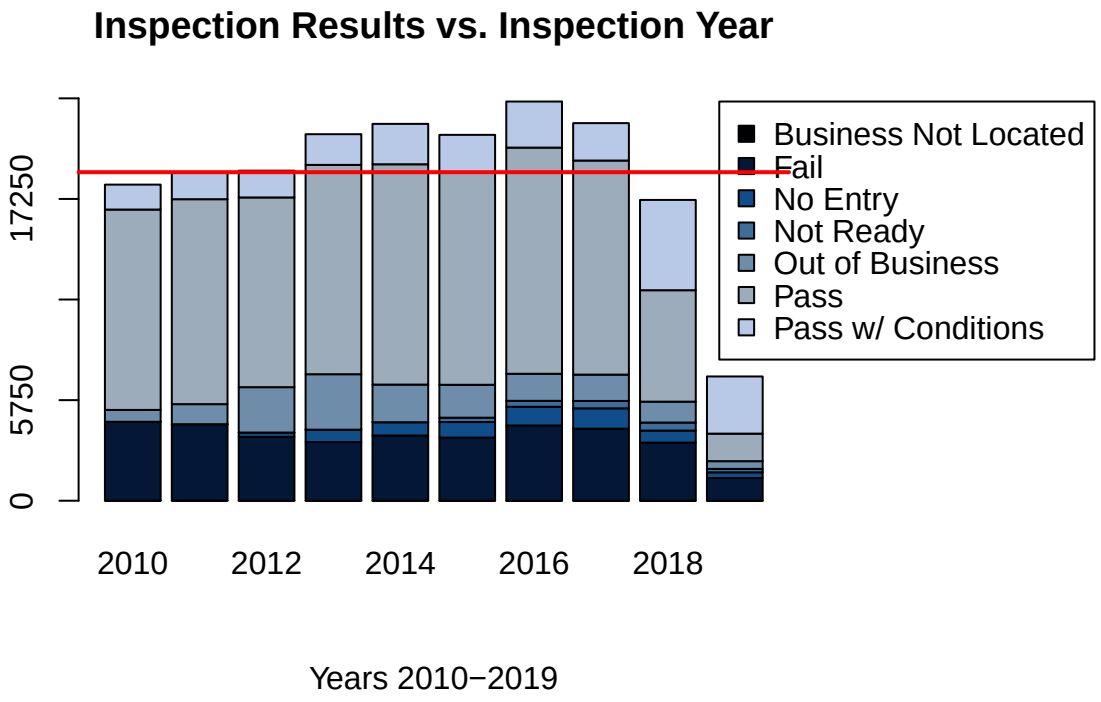
```
averageTotal = mean(colSums(t1))
averagePass = mean(t1["Pass",])
averageFail = mean(t1["Fail",])
averagePC = mean(t1["Pass w/ Conditions",])
data.frame(averageTotal, averagePass, averageFail,averagePC, row.names = "through year 2010-2019")
```

```
##                         averageTotal averagePass averageFail averagePC
## through year 2010-2019      18778.7     10380.4      3620.9    2389.6
```

This table summarizes the average number of Results of inspection in past 10 years from 2010 to 2019, as well as average number of Pass, average number of fail, and average number of Pass with conditions.

b. redo the plot in **part 1c** but add one of the descriptive statistics from **part 2a** to the plot

```
par(mar=c(8, 4, 4, 10), xpd=TRUE)
barplot(t1, axes = FALSE, col = c('#000102','#041737',"#104E8B",'#3F6D9B','#6E8DAB','#9DACBB','#B7C9E6'
title("Inspection Results vs. Inspection Year", "Years 2010-2019")
axis(side = 2, at=c(0, 5750,11500, 17250, 23000), xpd=TRUE)
par(mar=c(6, 3, 4, 2), xpd=TRUE)
legend("topright", inset=c(-0.43,0), legend=levs, fill=c('#000102','#041737',"#104E8B",'#3F6D9B','#6E8D
average = mean(colSums(t1))
par(xpd=FALSE)
abline(h=average, col=2, lwd=2)
```



In this question, I added a line to indicate the mean of the number of results throughout the past 10

years from 2010 to 2019.

c. write a brief explanatory narrative of the visualization in **part 2b**. In your explanation, be convincing and persuasive about your visualization. Attempt to highlight why this visualization is crucial to your imaginary supervisor.

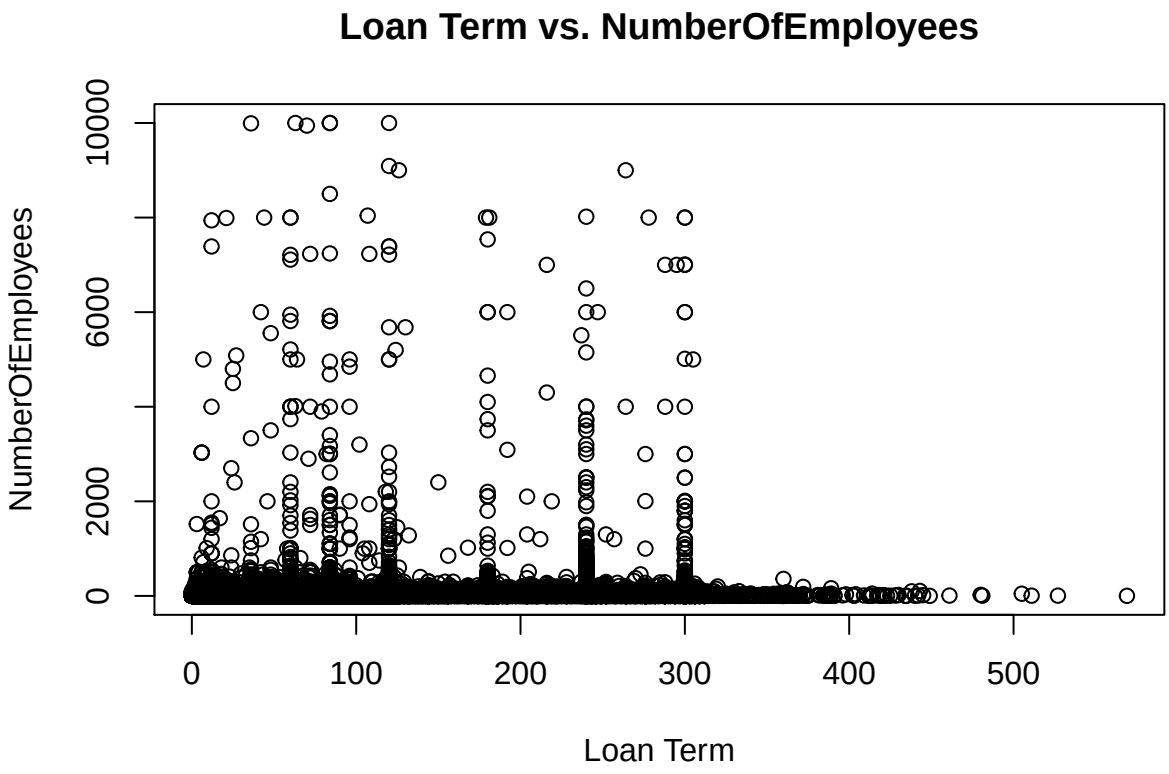
From this barplot, my supervisor could see and compare the total inspection results throughout year 2010 to 2019. He could also evaluate the the number of results every year by comparing with the average number of results of the past 10 years. The color each indicates a different inspection result, which can be easily told the difference of each results within a year.

3. Read https://www.tandfonline.com/doi/full/10.1080/10691898.2018.1434342 for information about the SBA Business Loans Data and do the following:

a. create a visualization (plot) of at least two variables (excluding ID or label variables) using this dataset

```
bld = read.csv("https://uofi.box.com/shared/static/vi37omgitiaa2yyplrom779qvwk1g14x.csv")
```

```
plot(bld$Term, bld$NoEmp, xlab = "Loan Term", ylab = "NumberOfEmployees")
title("Loan Term vs. NumberOfEmployees")
```



In this plot, I chose Loan Term(Term) and number of business employees(NoEmp) as my variables.

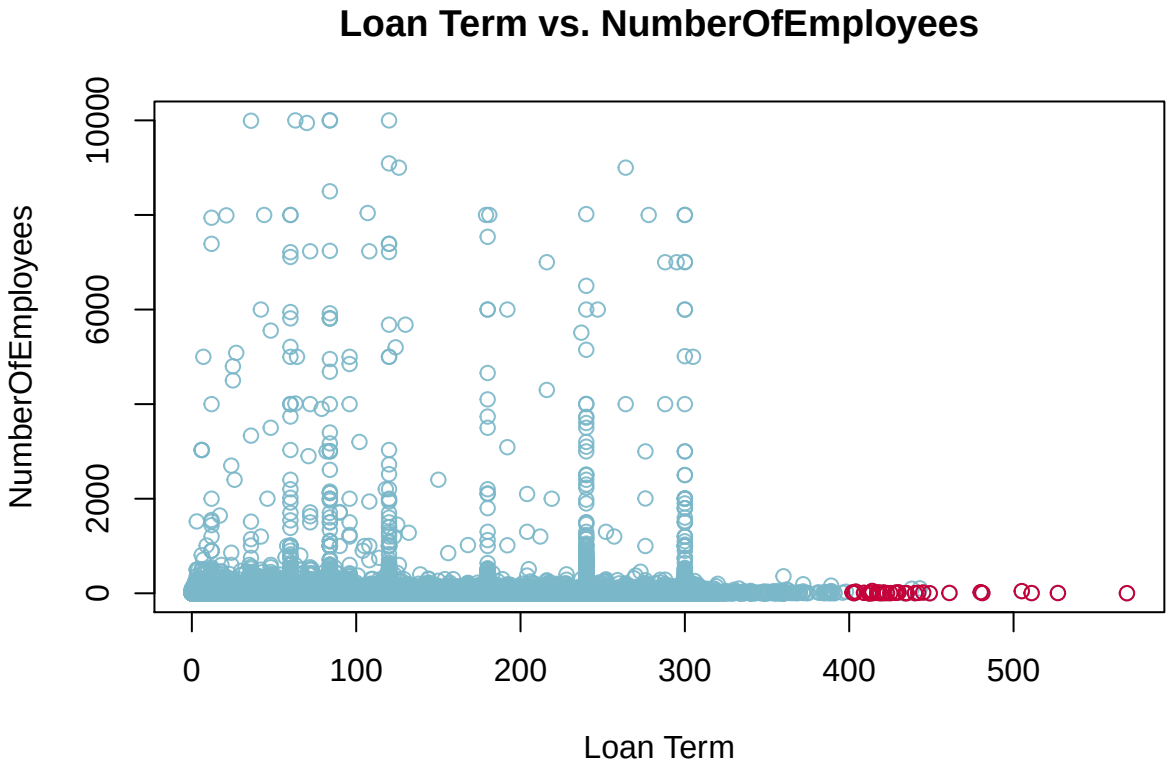
b. explain what is good and what is bad about the visualization

I plotted Loan Term versus NumberOfEmployees to see the connection between the loan term and number of employees. From the plot, it is clear to see the relationship. And we can also tell that there are a large amount of

loan ranging from 0 to 300 are from companies with number of employees less than 2000. But there are some significant points should be highlighted as outliers, which will be improved in the next plot.

c. show a substantially improved visualization

```
plot(bld$Term, bld$NoEmp, xlab = "Loan Term", ylab = "NumberOfEmployees", col='#7AB7C9E6')
title("Loan Term vs. NumberOfEmployees")
points(bld$Term[bld$Term > 400 & bld$NoEmp < 50], bld$NoEmp[bld$Term > 400 & bld$NoEmp < 50], col="#C70
```



d. describe the improvement and why the improved plot in **part 3c** helps the reader or viewer more than the original plot in **part 3a**.

In the graph, I made a colored plot to have more beatiful view. Besides, I highlighted the spots which have loan term longer than 400 months with the business employees less than 50. It should be significant to the readers since apparently they can see that most of the loan that larger than 400 are from companies that has number of employees less than 50. These significant points will show useful information for readers

4. Read https://www.tandfonline.com/doi/full/10.1080/10691898.2018.1434342 for information about the SBA Business Loans Data and do the following:

a. create a table of descriptive statistics of your choice

```
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
```

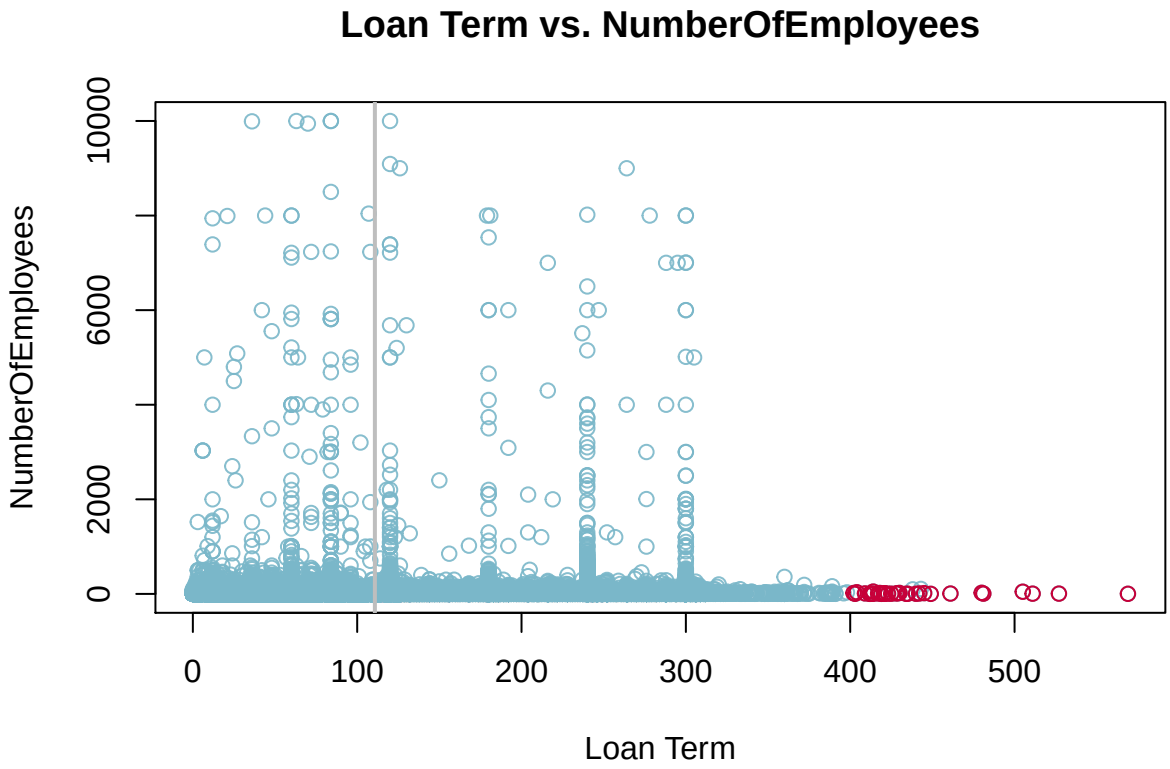
```
termMean = mean(bld$Term)
termMedian= median(bld$Term)
termMode = getmode(bld$Term)

empMean = mean(bld$NoEmp)
empMedian = median(bld$NoEmp)
empMode = getmode(bld$NoEmp)
data.frame(Term = c(termMean, termMedian, termMode), NoEmp = c(empMean, empMedian, empMode), row.names =
```

```
##             Term    NoEmp
## MEAN   110.7731 11.41135
## MEDIAN  84.0000  4.00000
## MODE    84.0000  1.00000
```

b. redo the plot in **part 3c** but add one of the descriptive statistics from **part 4a** to the plot

```
plot(bld$Term, bld$NoEmp, xlab = "Loan Term", ylab = "NumberOfEmployees", col='#7AB7C9E6')
title("Loan Term vs. NumberOfEmployees")
points(bld$Term[bld$Term > 400 & bld$NoEmp < 50], bld$NoEmp[bld$Term > 400 & bld$NoEmp < 50], col="#C70
abline(v=termMean, lwd=2, col = "grey")
```



c. write a brief explanatory narrative of the visualization in **part 4b**. In your explanation, be convincing and persuasive about your visualization. Attempt to highlight why this visualization is crucial to your imaginary supervisor.

The plot highlights the points which have loan term that longer than 400 months and companies with business employees less than 50 people. The plot also gives a grey line that indicates the mean of the loan term, which may give the supervisor useful information.