Analysis of FDA-Regulated Industry Audits

A few days back, I was informed by my boss that I have to make up a team of analytics and then has to bring out insights from data provided by audits. This was a crucial step for FDA because after looking into the analyses they can make reasonable adjustments so that vendors comply with the standards set by them. Also, they can understand as to what results we can get if we combine data analytics with auditing methods.

Our team has completed the analysis in R programming and compiled a report juxtaposing code and their respective output. In the end we have written our best observation, which our boss can forwards to seniors with proper recommendation.

Packages required: “dplyr”, “gridExtra”

Part 1

Good_x_Practices

install.packages("gridExtra")

install.packages("dplyr")

library(gridExtra) library(dplyr)

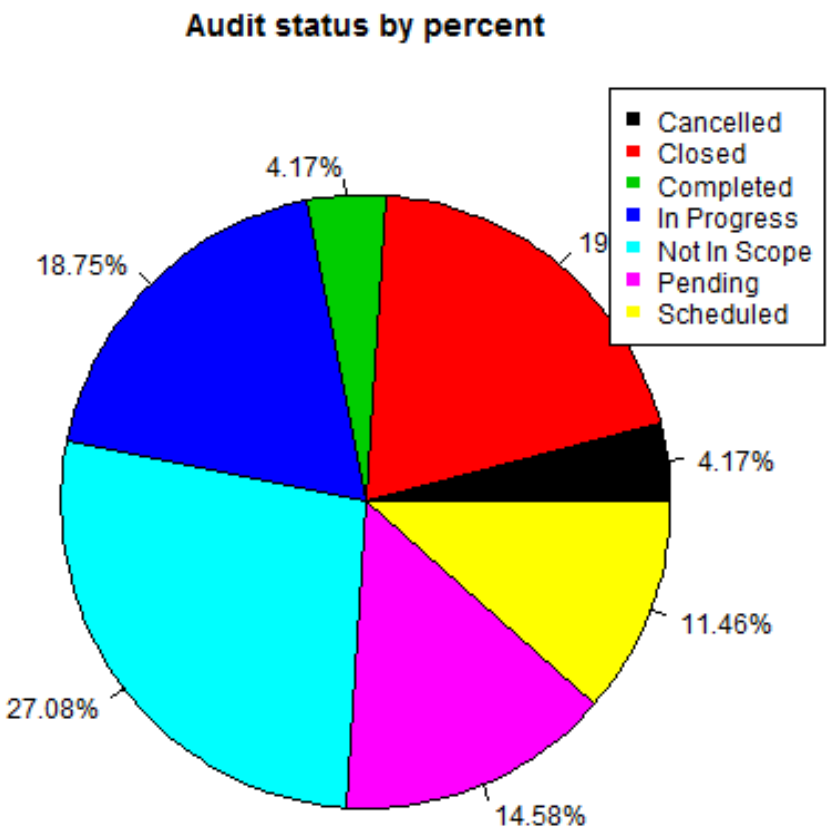
```
getdata <- read.csv("GXP.csv",sep=",",header=TRUE)
df <- getdata[1:96,c(2,3,6,7,8,9)]
a1<- table(as.character(df[,1]))
a2 <- table(as.character(df[,2]))
a3 <- table(as.character(df[,3]))
a4 <- table(as.character(df[,4]))
a5 <- table(as.character(df[,5]))
a6 <- table(as.character(df[,6]))
```

(A)

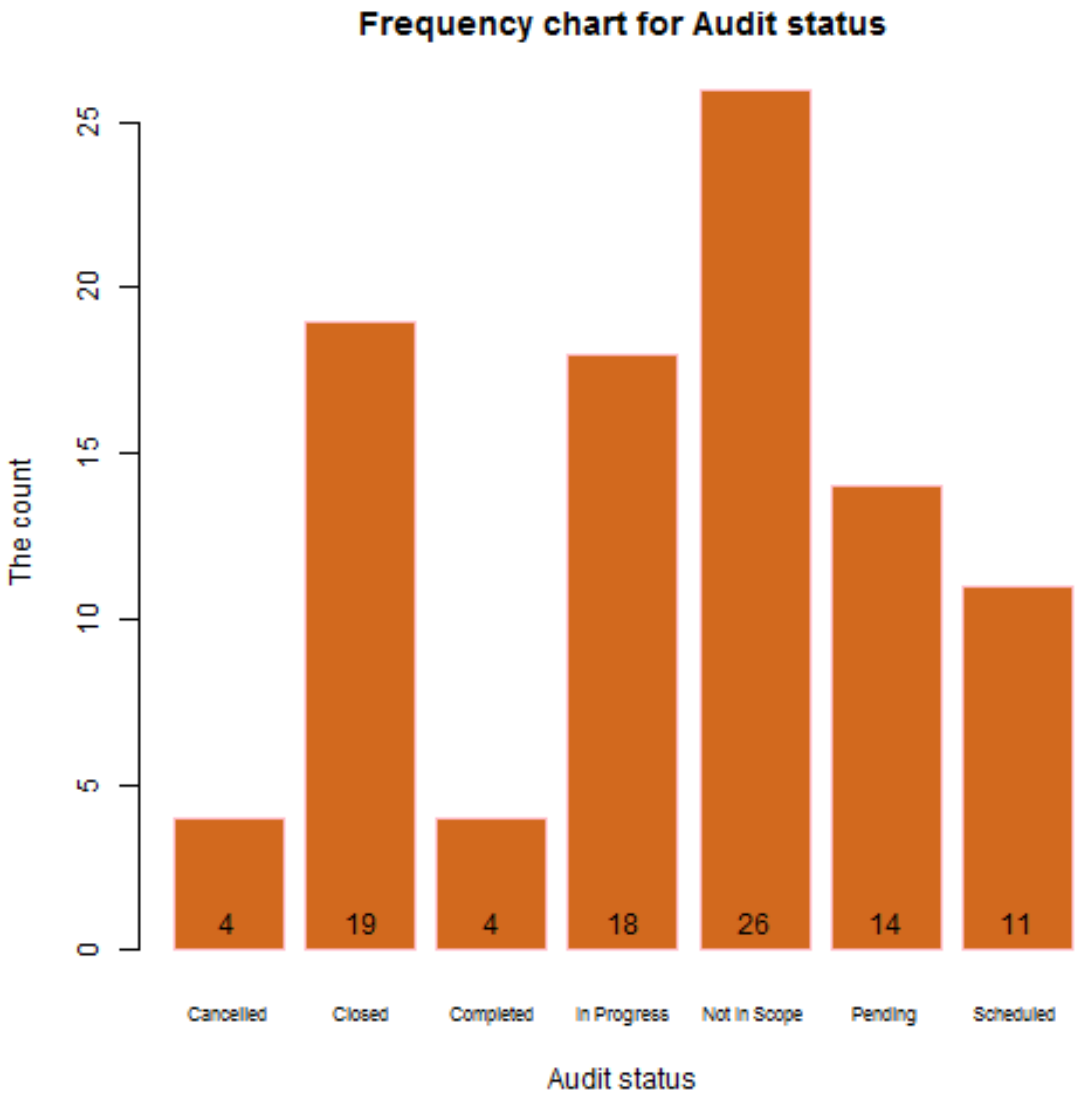
```
library(gridExtra)
## Warning: package 'gridExtra' was built under R version 3.1.3
tb1<- data.frame(a1)
colnames(tb1)<- c("Audit status","Freq")
grid.table(tb1)
```

	Audit status	Freq
1	Cancelled	4
2	Closed	19
3	Completed	4
4	In Progress	18
5	Not In Scope	26
6	Pending	14
7	Scheduled	11

```
percent <- (100*tb1[,2])/(sum(tb1[,2]))
percent<- round(percent,2)
labels <- paste(percent,"%",sep="")
pie(a1,labels=labels,main="Audit status by percent",col=unique(tb1[,1]))
legend("topright", legend=tb1[,1],col=unique(tb1[,1]),pch=15)
```




```
p1<-
barplot(a1,col="chocolate",border="pink",cex.names=0.7,xpd=FALSE,xlab="Audit
status",ylab="The count",main="Frequency chart for Audit status")
text(p1, 0, round(a1, 1), pos=3,xpd=FALSE)
```

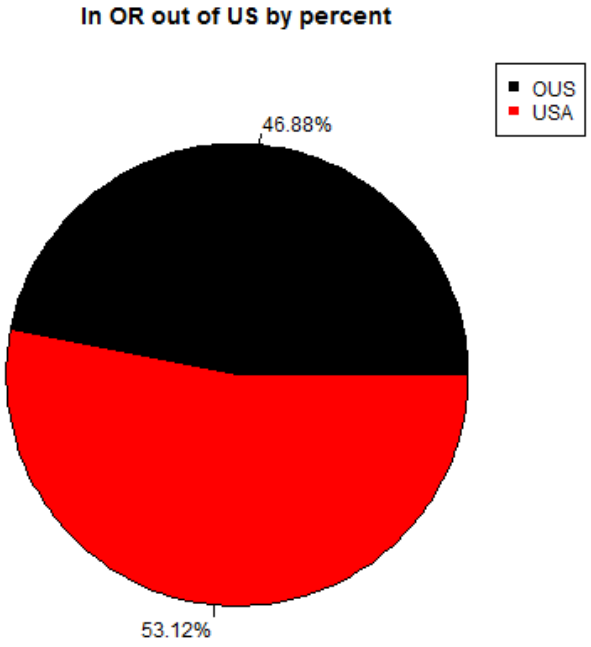


(B)

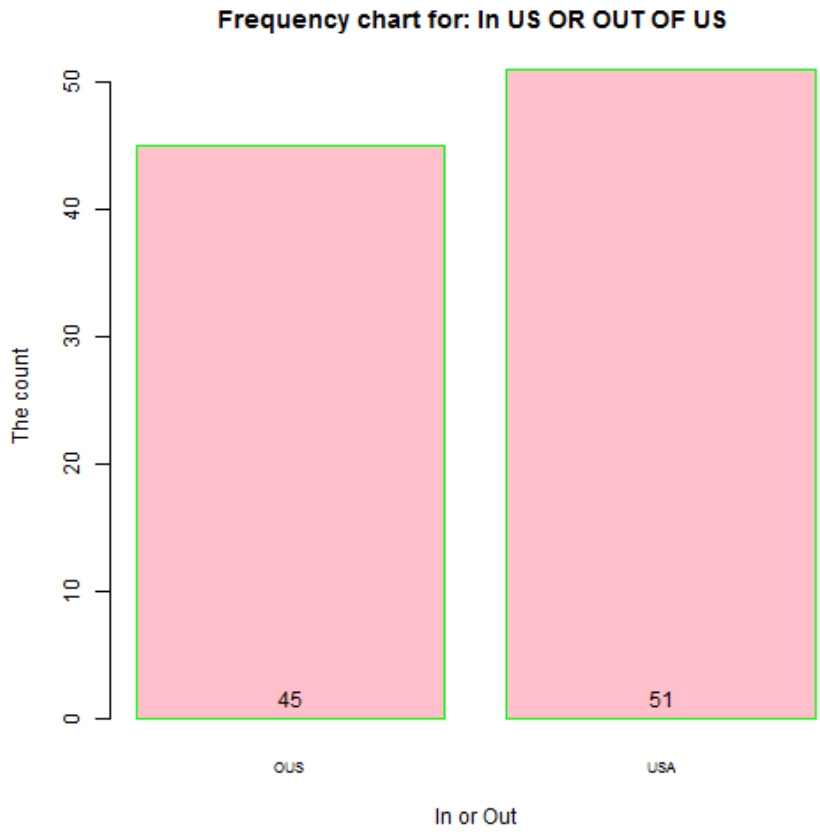
```
tb2<- data.frame(a2)
colnames(tb2)<- c("In or Out of US","Freq")
grid.table(tb2)
```

	In or Out of US	Freq
1	OUS	45
2	USA	51

```
percent <- (100*tb2[,2])/(sum(tb2[,2]))
percent<- round(percent,2)
labels <- paste(percent,"%",sep="")
pie(a2,labels=labels,main="In OR out of US by percent",col=unique(tb2[,1]))
legend("topright", legend=tb2[,1],col=unique(tb2[,1]),pch=15)
```




```
p2<- barplot(a2,col="pink",cex.names=0.7,border="green",xpd=FALSE,xlab="In or
Out",ylab="The count",main="Frequency chart for: In US OR OUT OF US ")
text(p2, 0, round(a2, 1), pos=3,xpd=FALSE)
```

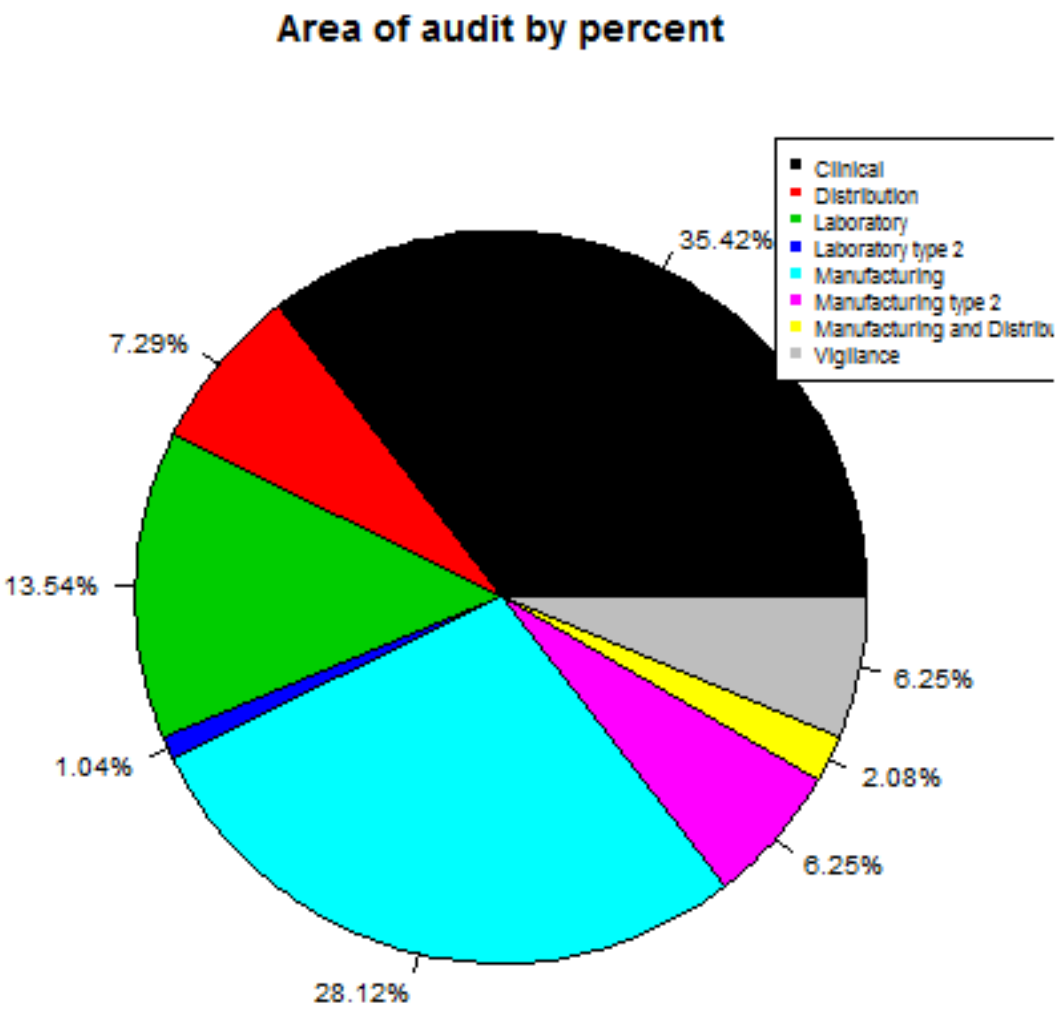


(C)

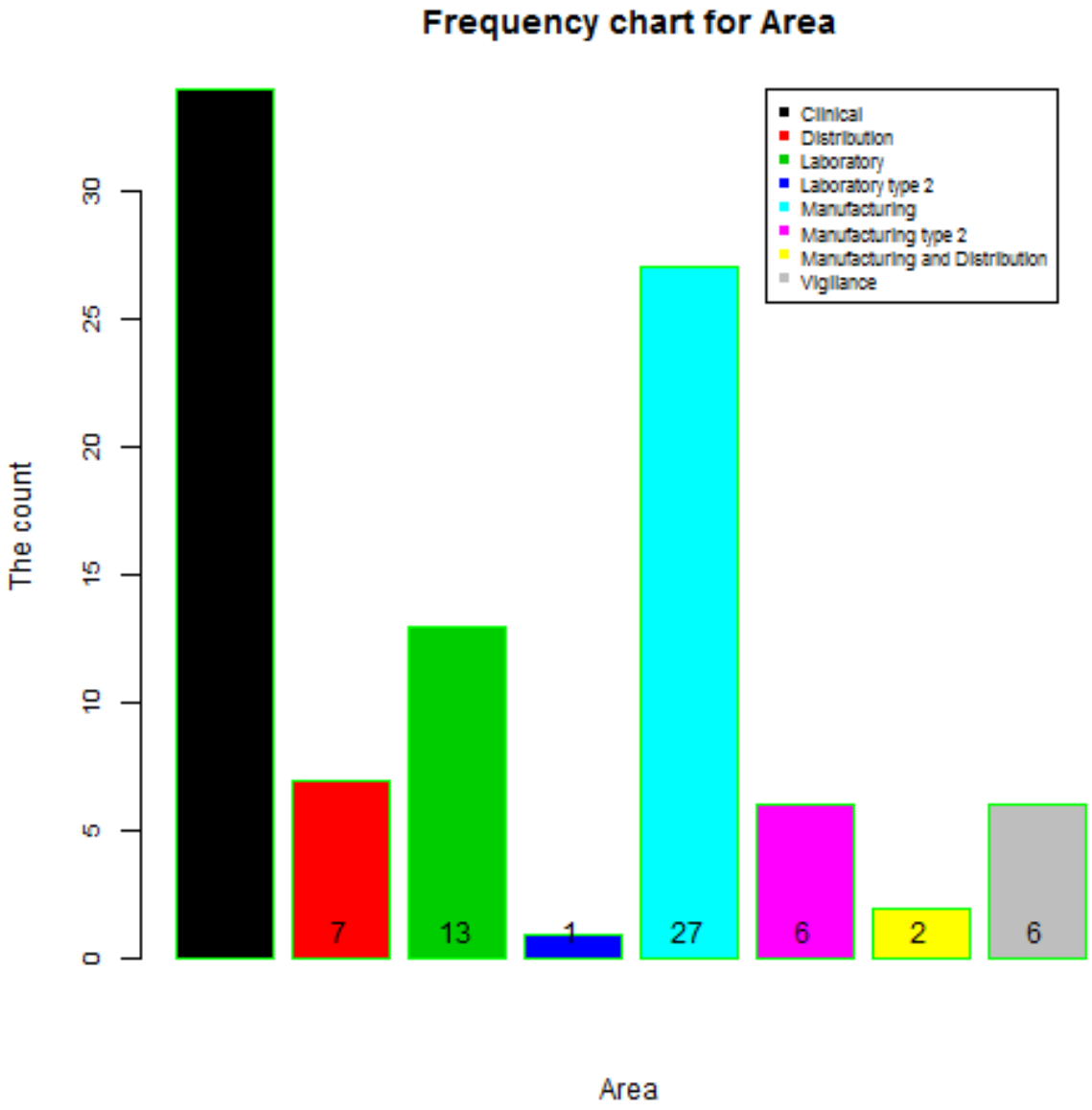
```
names(a3)<- c("Clinical","Distribution","Laboratory","Laboratory type
2","Manufacturing","Manufacturing type 2","Manufacturing and
Distribution","Vigilance")

tb3<- data.frame(a3)

colnames(tb3)<- c("AREA OF AUDIT ","Freq")
grid.table(tb3)
```

	AREA OF AUDIT	Freq
1	Clinical	34
2	Distribution	7
3	Laboratory	13
4	Laboratory type 2	1
5	Manufacturing	27
6	Manufacturing type 2	6
7	Manufacturing and Distribution	2
8	Vigilance	6

```
percent <- (100*tb3[,2])/(sum(tb3[,2]))
percent<- round(percent,2)
labels <- paste(percent,"%",sep="")

pie(a3,labels=labels,main="Area of audit by
percent",col=unique(tb3[,1]),cex=0.8)
legend(.6,1, legend=tb3[,1],col=unique(tb3[,1]),pch=15,cex=.7)
```




```
p3<-
barplot(a3,col=tb3[,1],cex.names=0.7,cex.axis=.8,border="green",xpd=FALSE,xla
b="Area",ylab="The count",main="Frequency chart for Area",names.arg=FALSE)
text(p3, 0, round(a3, 1), pos=3,xpd=FALSE)
legend(6.3,34, legend=tb3[,1],col=unique(tb3[,1]),pch=15,cex=0.7)
```

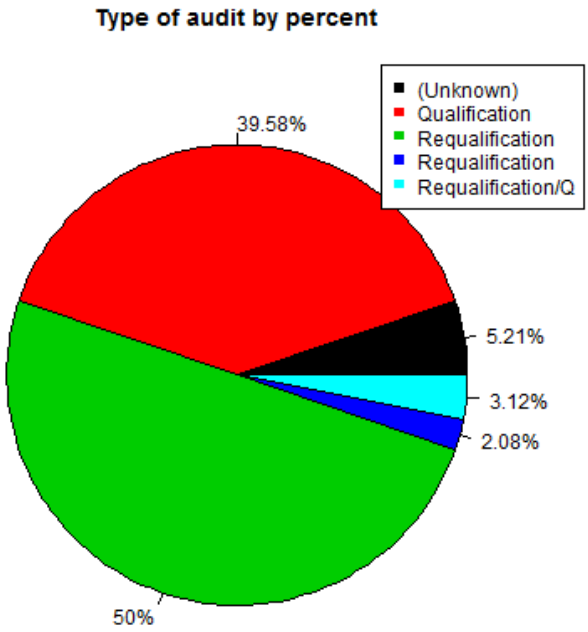

Frequency chart for Area
Clinical
Distribution
Laboratory
Laboratory type 2
Manufacturing
Manufacturing type 2
Manufacturing and Distribution
Vigilance
The count
0
5
10
15
20
25
30
7
13
1
27
6
2
6
Area

(D)

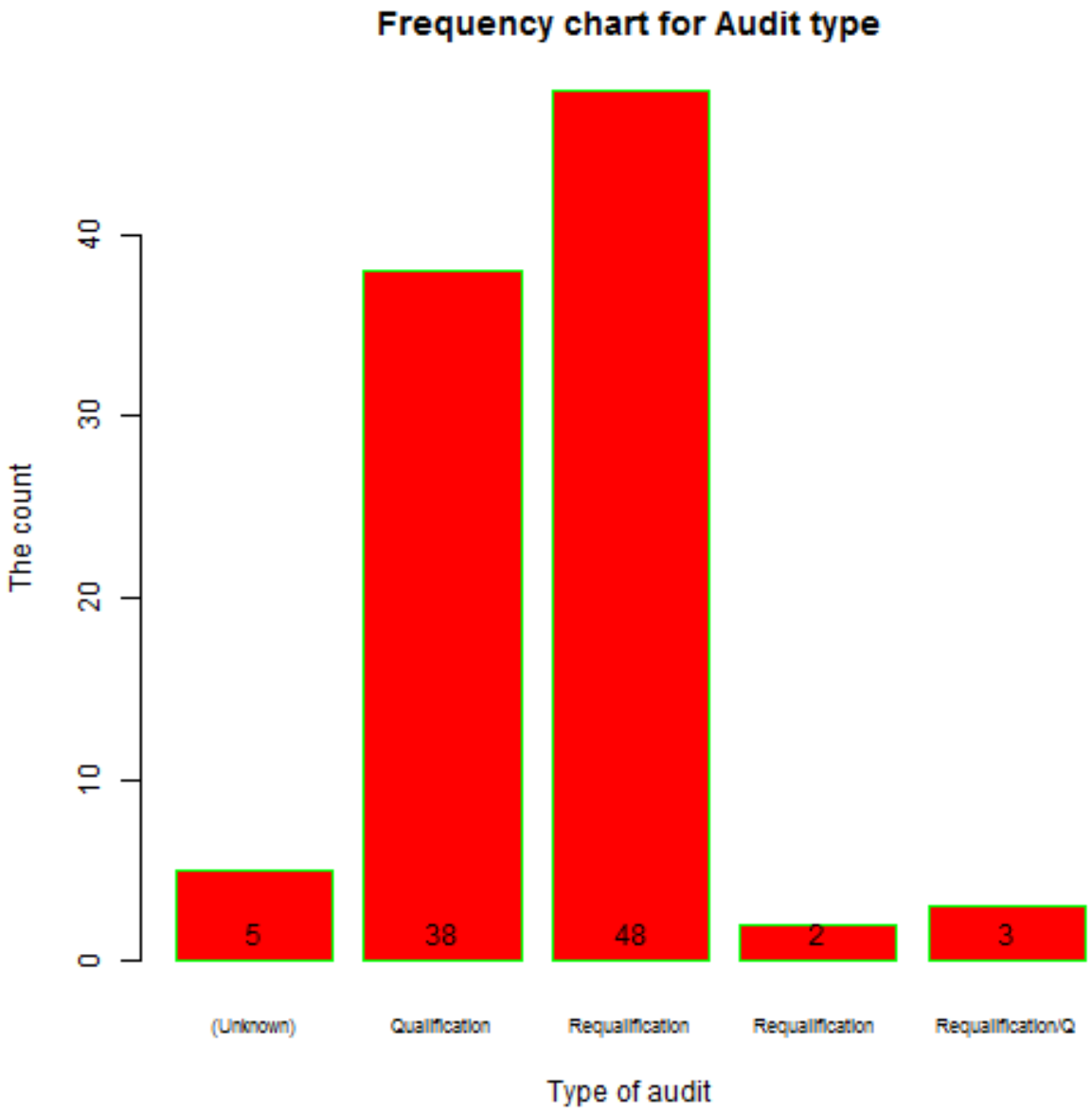
```
tb4<- data.frame(a4)
colnames(tb4)<- c("TYPE OF AUDIT","Freq")
grid.table(tb4)
```

	TYPE OF AUDIT	Freq
1	(Unknown)	5
2	Qualification	38
3	Requalification	48
4	Requalification	2
5	Requalification/Q	3

```
percent <- (100*tb4[,2])/(sum(tb4[,2]))
percent<- round(percent,2)
labels <- paste(percent,"%",sep="")
pie(a4,labels=labels,main="Type of audit by percent",col=unique(tb4[,1]))
legend("topright", legend=tb4[,1],col=unique(tb4[,1]),pch=15)
```




```
p4<- barplot(a4,col="red",cex.names=0.7,border="green",xpd=FALSE,xlab="Type
of audit",ylab="The count",main="Frequency chart for Audit type")
text(p4, 0, round(a4, 1), pos=3,xpd=FALSE)
```

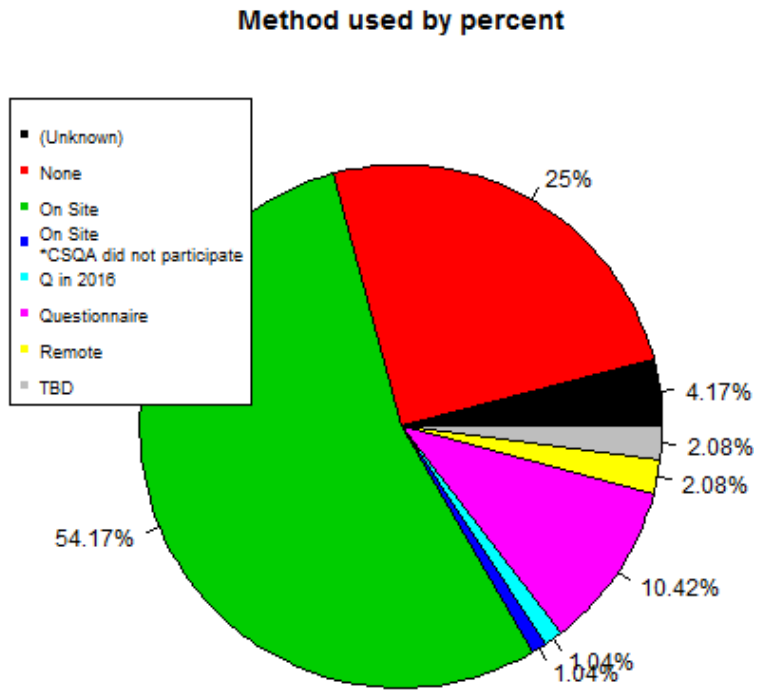

Frequency chart for Audit type
The count
0
10
20
30
40
5
38
48
2
3
(Unknown)
Qualification
Requalification
Requalification
Requalification/Q
Type of audit

(E)

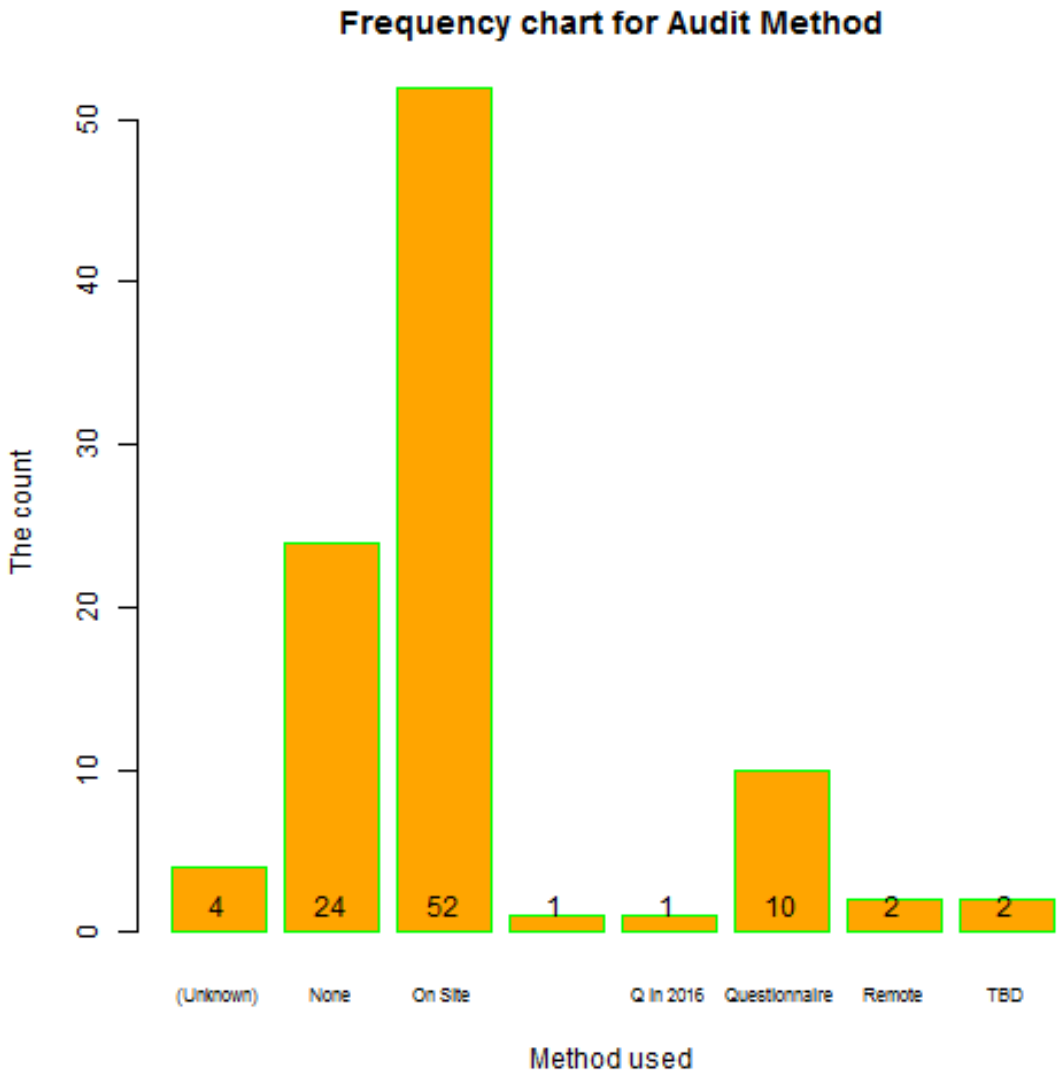
```
tb5<- data.frame(a5)
colnames(tb5)<- c("METHOD USED","Freq")
grid.table(tb5)
```

	METHOD USED	Freq
1	(Unknown)	4
2	None	24
3	On Site	52
4	On Site *CSQA did not participate	1
5	Q in 2016	1
6	Questionnaire	10
7	Remote	2
8	TBD	2

```
percent <- (100*tb5[,2])/(sum(tb5[,2]))
percent<- round(percent,2)
labels <- paste(percent,"%",sep="")
pie(a5,labels=labels,main="Quarter by percent",col=unique(tb5[,1]))
legend(-1.2,1, legend=tb5[,1],col=unique(tb5[,1]),pch=15,cex=.75)
```




```
p5<-
barplot(a5,col="orange",border="green",cex.names=0.7,xpd=FALSE,xlab="Method
used",ylab="The count",main="Frequency chart for Audit Method")
text(p5, 0, round(a5, 1), pos=3,xpd=FALSE)
```

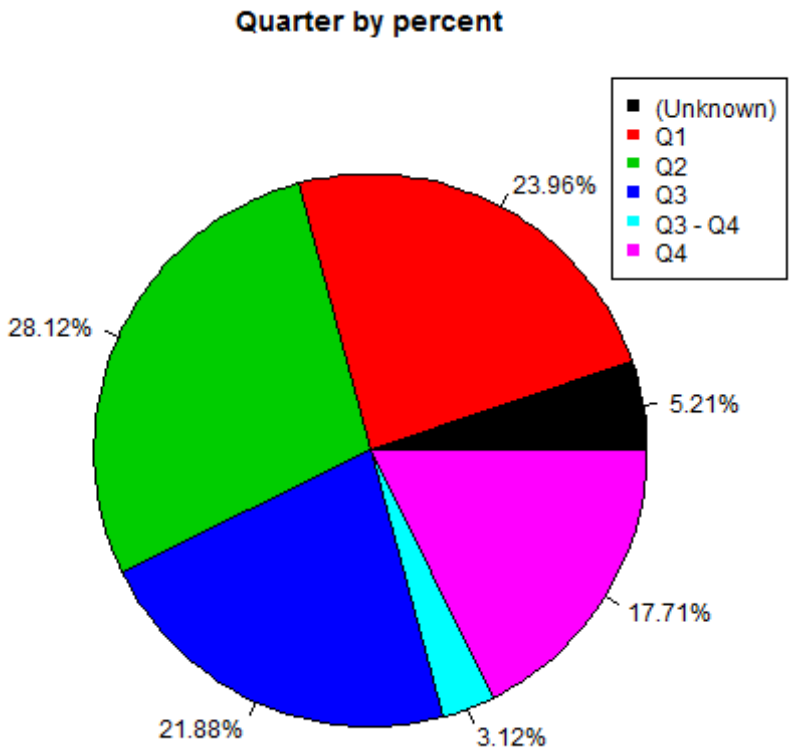

Frequency chart for Audit Method
The count
0
10
20
30
40
50
4
24
52
1
1
10
2
2
(Unknown)
None
On Site
Q In 2016
Questionnaire
Remote
TBD
Method used

(F)

```
tb6<- data.frame(a6)
colnames(tb6)<- c("PROPOSED QUARTER","Freq")
grid.table(tb6)
```

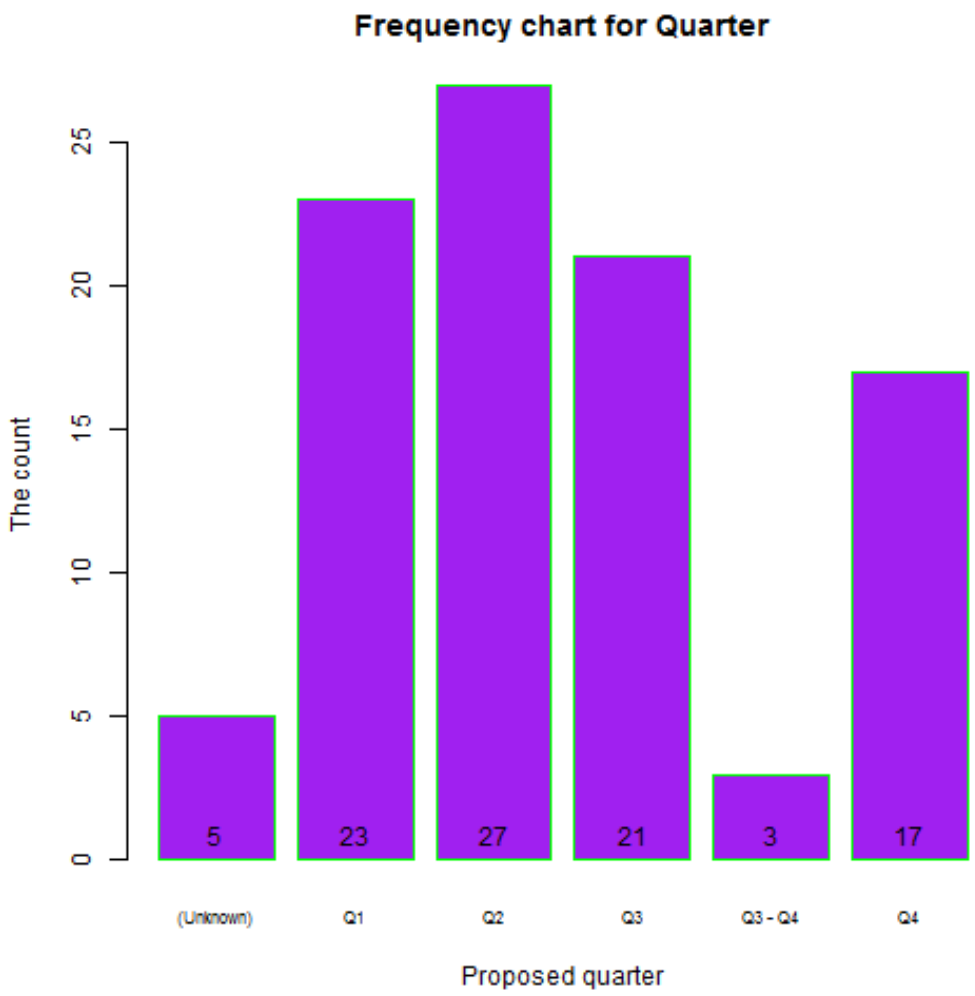
	PROPOSED QUARTER	Freq
1	(Unknown)	5
2	Q1	23
3	Q2	27
4	Q3	21
5	Q3 - Q4	3
6	Q4	17

```
percent <- (100*tb6[,2])/(sum(tb6[,2]))
percent<- round(percent,2)
labels <- paste(percent,"%",sep="")
pie(a6,labels=labels,main="Quarter by percent",col=unique(tb6[,1]))
legend("topright", legend=tb6[,1],col=unique(tb6[,1]),pch=15)
```




```
p6<-
barplot(a6,col="purple",border="green",xpd=FALSE,cex.names=0.7,xlab="Proposed
quarter",ylab="The count",main="Frequency chart for Quarter")
```

```
text(p6, 0, round(a6, 1), pos=3,xpd=FALSE)
```



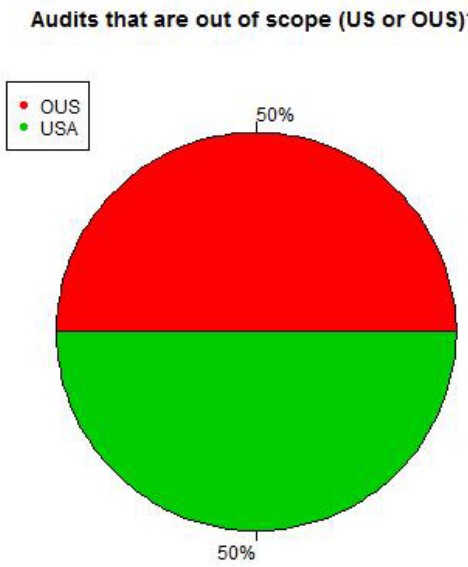
Observations

- More than 50% vendors are out of US.
- Onsite auditing method are generally practiced because about 50% of audits are shown to be onsite.
- 50% of our vendors has already been audited before, that is why requalification is 50% green. This implies that rest are operating for the first time, that is, half of the vendors are new.
- Anomaly in this analysis is that two vendors that had requalified and qualifies respectively are “Q in 2016” and “CSQA did not participate” and they have their auditing status out of scope. I am now able to assume that there are some secondary authority that might be handling some of these vendors.

More on GxP

(A)

```
scope <- df %>% filter(Audit.Status=="Not In Scope") %>%
group_by(In.USA.or.OUS) %>% summarise(count=n())

scope<- data.frame(scope)

pie(scope[,2],col=unique(scope[,1]),labels=c("50%","50%"),main="Audits that
are out of scope (US or OUS)?")

legend(-1,1,legend=scope[,1],col=unique(scope[,1]),pch=19)
```

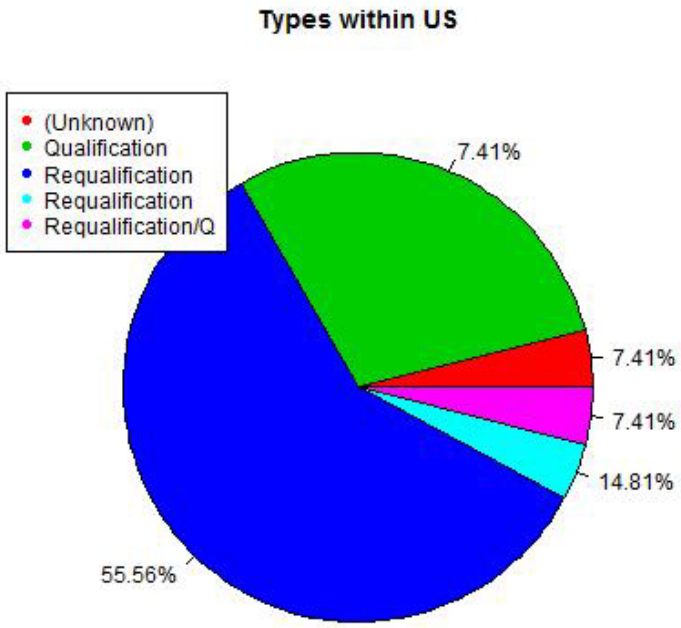


Observation: Being out of scope does not necessarily means that they are out of US, there are vendors who are in scope while they are out of US. Which makes it reasonable to assume that the auditing authority have people working from other countries, OR, this could also mean that vendors from US has established their business outside of US but they must comply with US regulatory rules if both of them have some affiliation.

(B)

```
getdata <- read.csv("GXP.csv",sep=",",header=TRUE)
df <- getdata[1:96,c(2,3,6,7,8,9)]
install.packages("dplyr")
```

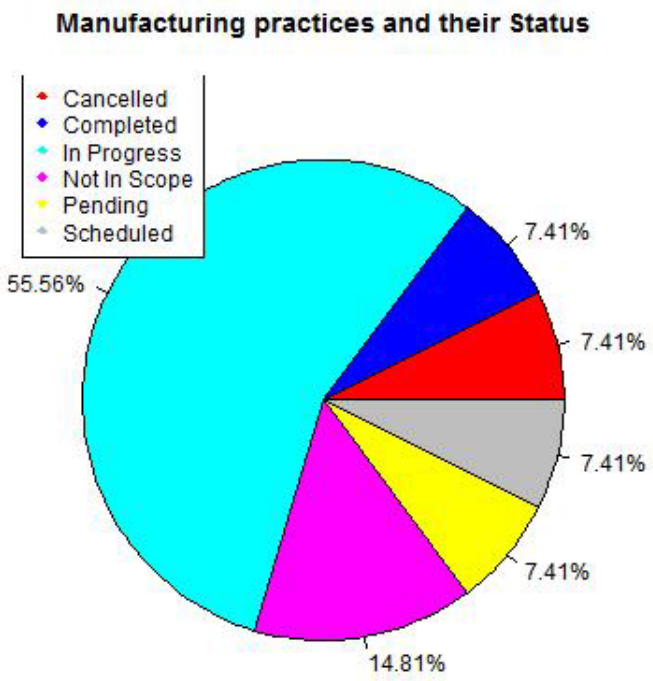
```
us <- df %>% filter(In.USA.or.OUS=="USA") %>% group_by(Audit.Type) %>%
summarise(count=n())

us <- data.frame(us)
percent <- round((100*us[,2])/(sum(us[,2])),2)
labels <- paste(percent,"%",sep="")
pie(us[,2],labels=labels,main="Types within US",col=unique(us[,1]))
legend(-1.2,1,legend=us[,1],col=unique(us[,1]),pch=19)
```



Observation: We see that within US, about 60% are requalified, which means that vendors that has established their business within US has maintained the quality of product or service they are providing. Vendors that are outside US can be studied and determined as to conclude what are the factors that makes it possible to perform well.

(C)

```
manf <- df %>% filter(GxP.Area=="GMP") %>% group_by(Audit.Status) %>%
summarise(count=n())
manf <- data.frame(manf)
percent= round((100*manf[,2])/(sum(manf[,2])),2)
labels <- paste(percent,"%",sep="")
pie(manf[,2],col=unique(manf[,1]),labels=labels,main="Manufacturing practices
and their Status")
legend(-1,1, leged=manf[,s1],col=manf[,1],pch=18)
```



Observation: We see that most of the manufacturing audits are in progress and very less has been completed. This insight could be used to study as to why this is the case. Do manufacturing vendors need more time to get audited than others, or somethings there that is not allowing the process to flow smoothly.

CSQA analysis

In this analysis we have less number of observation in comparison to the GxP analysis, for that same reason I have avoided the use of any graph or chart because the same message can be conveyed by simple tabular representation of information.

```
getdata2 <- read.csv("CSQA.csv",sep=",",header=TRUE)
df2 <- getdata2[,c(3,4,7,8,9,10)]
a1<- table(as.character(df2[,1]))
a2 <- table(as.character(df2[,2]))
a3 <- table(as.character(df2[,3]))
a4 <- table(as.character(df2[,4]))
a5 <- table(as.character(df2[,5]))
a6 <- table(as.character(df2[,6]))
library(gridExtra)
td1<- data.frame(a1)
colnames(td1)<- c("Audit status","Freq")
grid.table(td1)
```

```
td2<- data.frame(a2)
colnames(td2)<- c("In or Out of US","Freq")
grid.table(td2)
```

```
td3<- data.frame(a3)

colnames(td3)<- c("AREA OF AUDIT ","Freq")
grid.table(td3)
```

```
tb4<- data.frame(a4)
colnames(tb4)<- c("TYPE OF AUDIT","Freq")
grid.table(tb4)
```

```
tb5<- data.frame(a5)
colnames(tb5)<- c("METHOD USED","Freq")
grid.table(tb5)
```

```
tb6<- data.frame(a6)
colnames(tb6)<- c("PROPOSED QUARTER","Freq")
grid.table(tb6)
```

	Audit status	Freq
1	Closed	3
2	Completed	9

	In or Out of US	Freq
1	OUS	3
2	USA	9

	AREA OF AUDIT	Freq
1	GCP	1
2	GIS	10
3	GMP	1

	TYPE OF AUDIT	Freq
1	On Site	10
2	Questionnaire	2

	METHOD USED	Freq
1	Internal	1
2	Qualification	2
3	Requalification	9

	PROPOSED QUARTER	Freq
1	Q1	4
2	Q2	1
3	Q3	7

From the analysis of CSQA I have concluded that all the vendors in CSQA are performing first class in terms of qualifying and meeting the requirement criteria. My claim is backed by the fact that audits of all of them has been completed, which signifies that there were no discrepancies from the side of vendors that held up the process of auditing.

Moreover, in almost all cases, the assessment was done on site which tells us that US has better output when auditing with onsite implementation. In addition to that, most of them are in requalified category, which means they all have been performing very well in past — that again proofs the same point— that these vendors are doing a great job of meeting the guidelines set by US authorities.

Part 2

GxP Analysis

1

```
date_1<- getdata[,c(14,15)]
intake<- as.character(date_1[,1])
sent <- as.character(date_1[,2])
intake <- as.Date(intake,format = "%m/%d/%Y")
sent <- as.Date(sent,format = "%m/%d/%Y")
Days_Intake_QSent <- sent[1:100] - intake[1:100]
x<- data.frame(Days_Intake_QSent)
mean_1 <- mean(Days_Intake_QSent,na.rm=TRUE)
median_1<- median(Days_Intake_QSent,na.rm=TRUE)
print(paste("Mean is:",round(mean_1,2), "and Median is:",median_1))

## [1] "Mean is: 17.36 and Median is: 2"
```

Result1:

2

```
date_2 <- getdata[,c(15,16)]
sent <- as.character(date_2[,1])
received <- as.character(date_2[,2])
sent <- as.Date(sent,format = "%m/%d/%Y")
received <- as.Date(received,format = "%m/%d/%Y")
Days_QSent_QReceived <- received[1:100] - sent[1:100]
x<-data.frame(Days_QSent_QReceived)
mean_2 <- mean(Days_QSent_QReceived,na.rm=TRUE)
median_2<- median(Days_QSent_QReceived,na.rm=TRUE)
print(paste("Mean is:",round(mean_2,2), "and Median is:",median_2))

## [1] "Mean is: 29.2 and Median is: 33"
```

Result 2: It takes around a month for audit report to reach the office for analysis.

3

```
date_3 <- getdata[,c(17,18)]
scheduled <- as.character(date_3[,1])
actual_start <- as.character(date_3[,2])
scheduled <- as.Date(scheduled,format = "%m/%d/%Y")
actual_start <- as.Date(actual_start,format = "%m/%d/%Y")
Days_AUDIT_scheduled_vs_AUDIT_actual <- actual_start[1:100] -
scheduled[1:100]
x<-data.frame(Days_AUDIT_scheduled_vs_AUDIT_actual)
mean_3<- mean(Days_AUDIT_scheduled_vs_AUDIT_actual,na.rm=TRUE)
median_3<- median(Days_AUDIT_scheduled_vs_AUDIT_actual,na.rm=TRUE)
```

```
print(paste("Mean is:",round(mean_3,2), "and Median is:",median_3))

## [1] "Mean is: 54.49 and Median is: 47"
```

Result 3: On average, the auditing starts very late in comparison to its schedule.

4

```
date_4 <- getdata[,c(18,19)]
start <- as.character(date_4[,1])
end <- as.character(date_4[,2])
start <- as.Date(start,format = "%m/%d/%Y")
end <- as.Date(end,format = "%m/%d/%Y")
Days_StartDate_EndDate <- end[1:100] - start[1:100]
x<-data.frame(Days_StartDate_EndDate)
mean_4<- mean(Days_StartDate_EndDate,na.rm=TRUE)
median_4<- median(Days_StartDate_EndDate,na.rm=TRUE)
print(paste("Mean is:",round(mean_4,2), "and Median is:",median_4))

## [1] "Mean is: 1.06 and Median is: 1"
```

Result 4: These values indicated that it takes 1 day on average to complete auditing of a vendor.

5

```
date_5 <- getdata[,c(19,20)]
end<- as.character(date_5[,1])
audit_due <- as.character(date_5[,2])
end <- as.Date(end,format = "%m/%d/%Y")
audit_due <- as.Date(audit_due,format = "%m/%d/%Y")
Days_AuditEnd_FinalReportDue <- audit_due[1:100] - end[1:100]
x<-data.frame(Days_AuditEnd_FinalReportDue)
mean_5<- mean(Days_AuditEnd_FinalReportDue,na.rm=TRUE)
median_5<- median(Days_AuditEnd_FinalReportDue,na.rm=TRUE)
print(paste("Mean is:",round(mean_5,2), "and Median is:",median_5))

## [1] "Mean is: 29.94 and Median is: 30"
```

Result 5: All audits completed well before final report needs to be submitted. This way, enough time is given for documentation.

6

```
date_6 <- getdata[,c(20,21)]
audit_due <- as.character(date_6[,1])
audit_completed <- as.character(date_6[,2])
audit_due <- as.Date(audit_due,format = "%m/%d/%Y")
audit_completed <- as.Date(audit_completed,format = "%m/%d/%Y")
Days_FinalReportDue_CompletionDate <- audit_due[1:100] -
audit_completed[1:100]
x<-data.frame(Days_FinalReportDue_CompletionDate)
```

```
mean_6 <- mean(Days_FinalReportDue_CompletionDate,na.rm=TRUE)
median_6<- median(Days_FinalReportDue_CompletionDate,na.rm=TRUE)
print(paste("Mean is:",round(mean_6,2), "and Median is:",median_6))

## [1] "Mean is: 0.16 and Median is: 0"
```

Result 6: Generally speaking, from the average we can say that most of the audits report were completed right on scheduled time.

CSQA Analysis

```
getdata2 <- read.csv("CSQA.csv",sep=",",header=TRUE)
```

1

```
date_1<- getdata2[,c(14,15)]
intake<- as.character(date_1[,1])
sent <- as.character(date_1[,2])
intake <- as.Date(intake,format = "%m/%d/%Y")
sent <- as.Date(sent,format = "%m/%d/%Y")
Days_Intake_QSent <- sent[1:12] - intake[1:12]
x<- data.frame(Days_Intake_QSent)
mean_1 <- mean(Days_Intake_QSent,na.rm=TRUE)
median_1<- median(Days_Intake_QSent,na.rm=TRUE)
print(paste("Mean is:",round(mean_1,2), "and Median is:",median_1))

## [1] "Mean is: 50.5 and Median is: 50.5"
```

2

```
date_2 <- getdata2[,c(15,16)]
sent <- as.character(date_2[,1])
received <- as.character(date_2[,2])
sent <- as.Date(sent,format = "%m/%d/%Y")
received <- as.Date(received,format = "%m/%d/%Y")
Days_QSent_QReceived <- received[1:12] - sent[1:12]
x<-data.frame(Days_QSent_QReceived)
mean_2 <- mean(Days_QSent_QReceived,na.rm=TRUE)
median_2<- median(Days_QSent_QReceived,na.rm=TRUE)
print(paste("Mean is:",round(mean_2,2), "and Median is:",median_2))

## [1] "Mean is: 18 and Median is: 18"
```

Result 2: It takes less time for CSQA audits to reach to reach final office for analysis in comparison with GxP.

3

```
date_3 <- getdata2[,c(17,18)]
scheduled <- as.character(date_3[,1])
actual_start <- as.character(date_3[,2])
scheduled <- as.Date(scheduled,format = "%m/%d/%Y")
actual_start <- as.Date(actual_start,format = "%m/%d/%Y")
Days_AUDIT_scheduled_vs_AUDIT_actual <- actual_start[1:12] - scheduled[1:12]
x<-data.frame(Days_AUDIT_scheduled_vs_AUDIT_actual)
mean_3<- mean(Days_AUDIT_scheduled_vs_AUDIT_actual,na.rm=TRUE)
median_3<- median(Days_AUDIT_scheduled_vs_AUDIT_actual,na.rm=TRUE)
print(paste("Mean is:",round(mean_3,2), "and Median is:",median_3))

## [1] "Mean is: 45.22 and Median is: 48"
```

Result 3: Though the auditing starts late than its scheduled date, but it's still better than GxP dataset.

4

```
date_4 <- getdata2[,c(18,19)]
start <- as.character(date_4[,1])
end <- as.character(date_4[,2])
start <- as.Date(start,format = "%m/%d/%Y")
end <- as.Date(end,format = "%m/%d/%Y")
Days_StartDate_EndDate <- end[1:12] - start[1:12]
x<-data.frame(Days_StartDate_EndDate)
mean_4<- mean(Days_StartDate_EndDate,na.rm=TRUE)
median_4<- median(Days_StartDate_EndDate,na.rm=TRUE)
print(paste("Mean is:",round(mean_4,2), "and Median is:",median_4))

## [1] "Mean is: 0.78 and Median is: 1"
```

Result 4: In some cases it takes even less than a day to get the audit completed.

5

```
date_5 <- getdata2[,c(19,20)]
end<- as.character(date_5[,1])
audit_due <- as.character(date_5[,2])
end <- as.Date(end,format = "%m/%d/%Y")
audit_due <- as.Date(audit_due,format = "%m/%d/%Y")
Days_AuditEnd_FinalReportDue <- audit_due[1:12] - end[1:12]
x<-data.frame(Days_AuditEnd_FinalReportDue)
mean_5<- mean(Days_AuditEnd_FinalReportDue,na.rm=TRUE)
median_5<- median(Days_AuditEnd_FinalReportDue,na.rm=TRUE)
print(paste("Mean is:",round(mean_5,2), "and Median is:",median_5))

## [1] "Mean is: 30 and Median is: 30"
```

Result 5: On average, much more time is given to reports of CSQA for documentation than GxP.

```
date_6 <- getdata2[,c(22,23)]
audit_debrief <- as.character(date_6[,1])
audit_completed <- as.character(date_6[,2])
audit_debrief <- as.Date(audit_debrief,format = "%m/%d/%Y")
audit_completed <- as.Date(audit_completed,format = "%m/%d/%Y")
Days_FinalReportDebrief_CompletionDate <- audit_completed[1:12] -
audit_debrief[1:12]
x<-data.frame(Days_FinalReportDebrief_CompletionDate)
mean_6 <- mean(Days_FinalReportDebrief_CompletionDate,na.rm=TRUE)
median_6<- median(Days_FinalReportDebrief_CompletionDate,na.rm=TRUE)
print(paste("Mean is:",round(mean_6,2), "and Median is:",median_6))

## [1] "Mean is: 29.67 and Median is: 27"
```

Result 6: Because we do not have final report due date in CSQA dataset, I have calculated the difference between the audit completion date and the debrief date. Debrief date is the date when a report is concluded and finalized by everyone in committee.

Part 3

I would **not** recommend merging “2017 GxP Audits” and “2017 CSQA”. There is a straight reasoning behind it— vendors auditing in CSQA outperforms vendor auditing of GxP. From my analysis I have concluded that vendors in CSQA worksheet have better quality in their functionality.

On the other hand, the time analysis of CSQA worksheet signifies that auditing process went smoother than GxP vendors, which again proofs that vendors in CSQA report are the ones that stands differently from vendors in GxP.

Merging both of these sheets would skew the result of whole analysis because the values in CSQA sheet are less in number (only 12) in comparison to GxP (around 100), but at the same time the values CSQA contains have far more weightage than values GxP contains, let alone so many unknown values in GxP.

My recommendation is to set a benchmark based on CSQA reports and when analysis for next year would be executed, the FDA can compare newly calculated values with CSQA values.