Lab report 1

Thomas Zhang

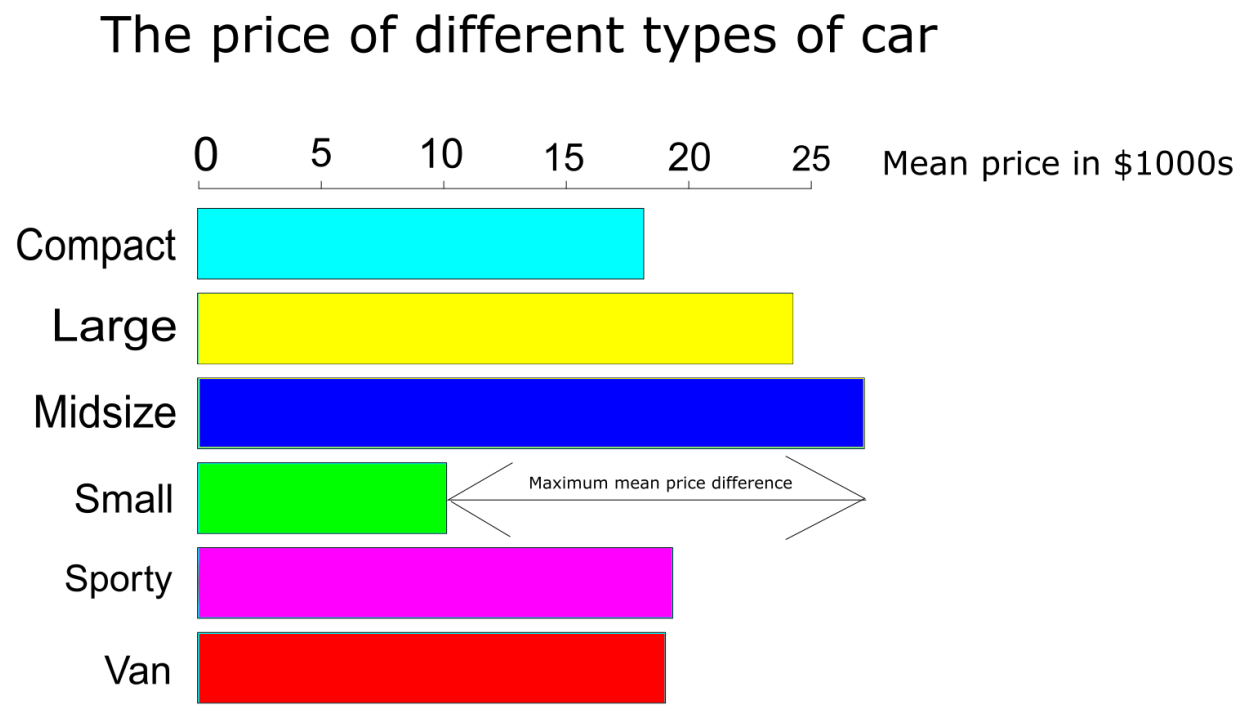
2016 M08 30

Assignment 1

OK, I opened the `cars93` data file in package `MASS`. Then I executed code snippet

```
library(MASS)
df1=aggregate(Price~Type, data=Cars93, FUN=mean)
barplot(df1$Price, names.arg=df1$Type)
```

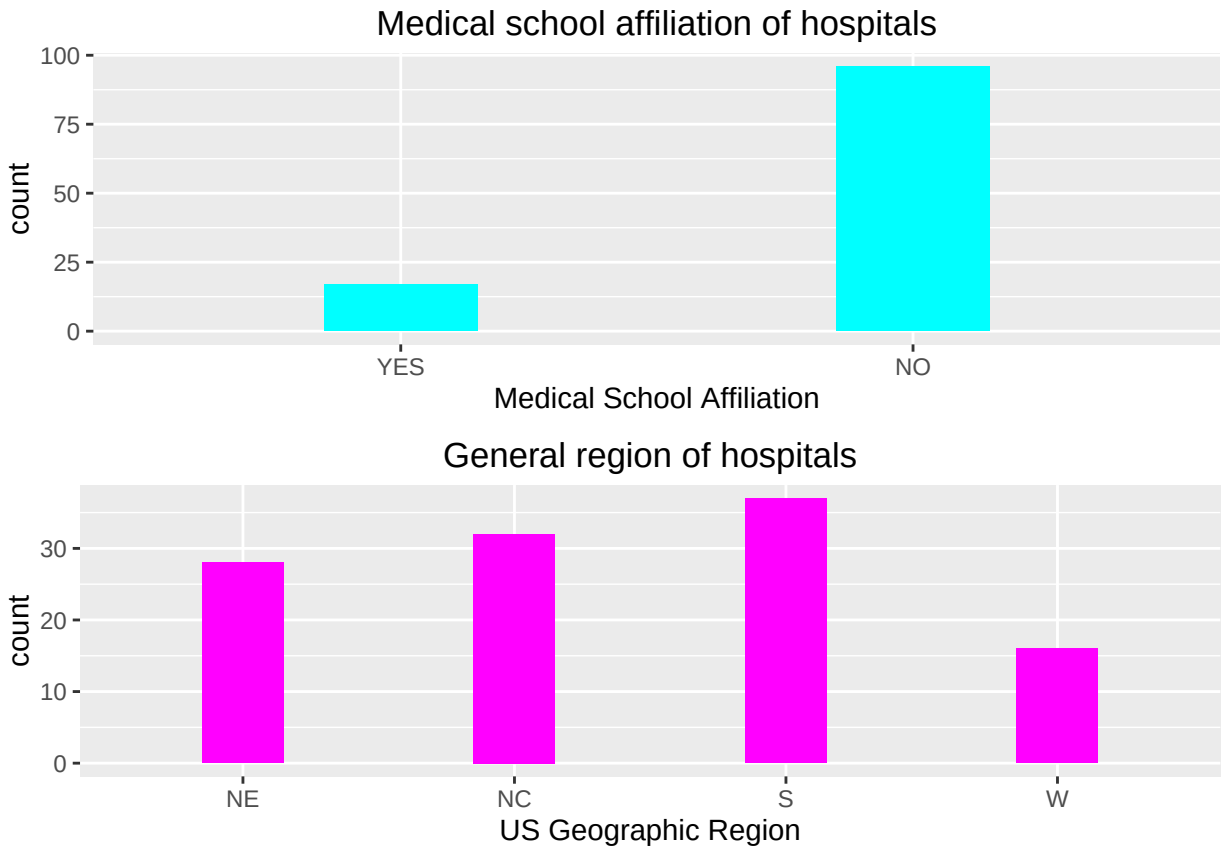
and I enhanced the resulting bar chart using inkscape. The final product looks like this:



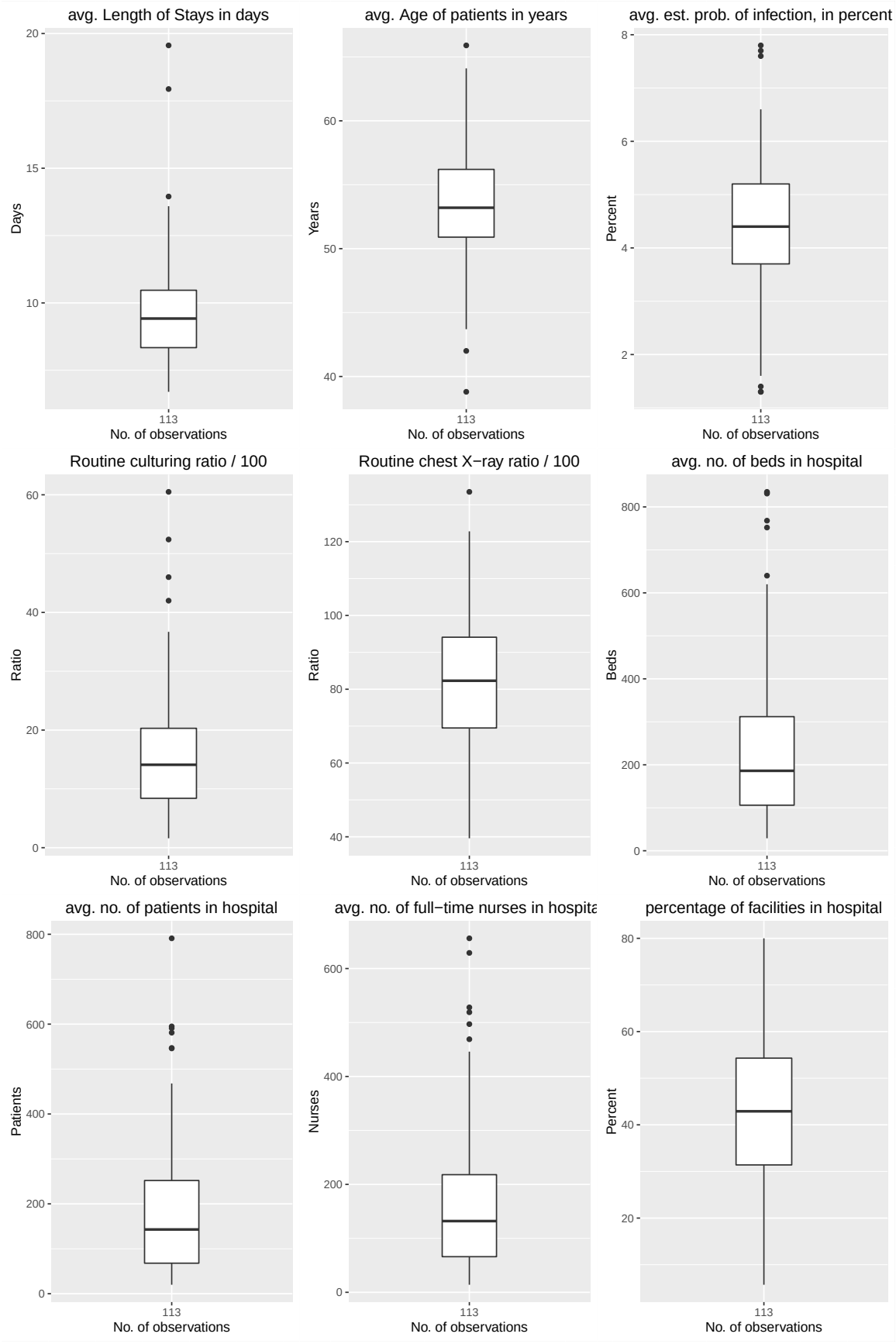
I believe it looks very nice.

Assignment 2

Next, we read U.S. hospital data on `SENIC.csv` and split the variables into qualitative and quantitative variables and produce nice bar charts and boxplots,respectively, using package `ggplot2`.



As we can see, most U.S. hospitals do not have a medical school affiliation.


avg. Length of Stays in days
Days
20
15
10
113
No. of observations
avg. Age of patients in years
Years
60
50
40
113
No. of observations
avg. est. prob. of infection, in percent
Percent
8
6
4
2
113
No. of observations
Routine culturing ratio / 100
Ratio
60
40
20
0
113
No. of observations
Routine chest X−ray ratio / 100
Ratio
120
100
80
60
40
113
No. of observations
avg. no. of beds in hospital
Beds
800
600
400
200
0
113
No. of observations
avg. no. of patients in hospital
Patients
800
600
400
200
0
113
No. of observations
avg. no. of full−time nurses in hospita
Nurses
600
400
200
0
113
No. of observations
percentage of facilities in hospital
Percent
80
60
40
20
113
No. of observations

The interesting question here is whether we have correlations in outliers between the boxplots. A quick check indicates that so may be the case. For instance, Hospital id numbers 47, 53 and 104 all have top ten positions in variables length of stay, age and infeciton risk.

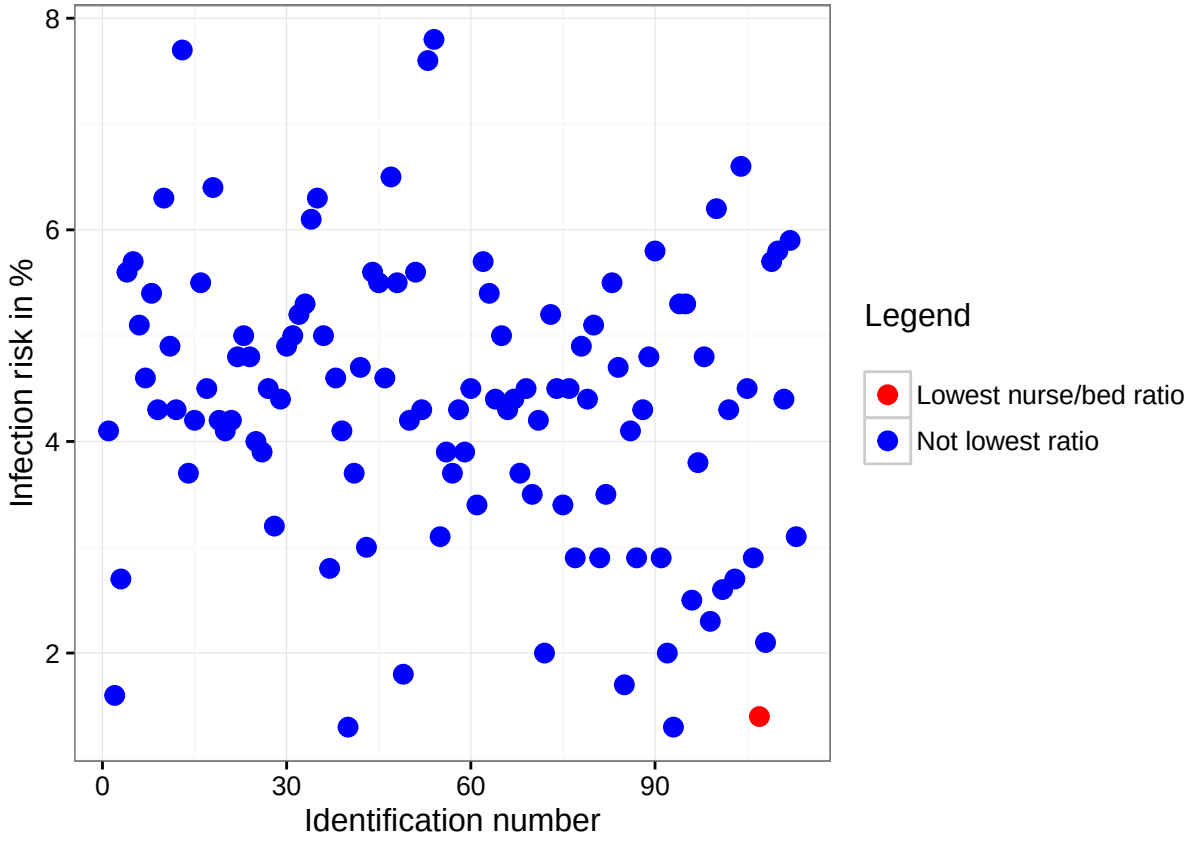
Assigment 3

Keeping with data set `SENIC.csv`, we write a code that finds out in which hospital the ratio "Number of nurses/Number of beds" is the lowest. This ratio tells us how well staffed the nurses at the hospital are relative to number of hospital beds. The code looks like this:

```
senicdata <- cbind(senicdata, senicdata$X10 / senicdata$X6)
sort(senicdata[,13],index.return = TRUE)$ix
```

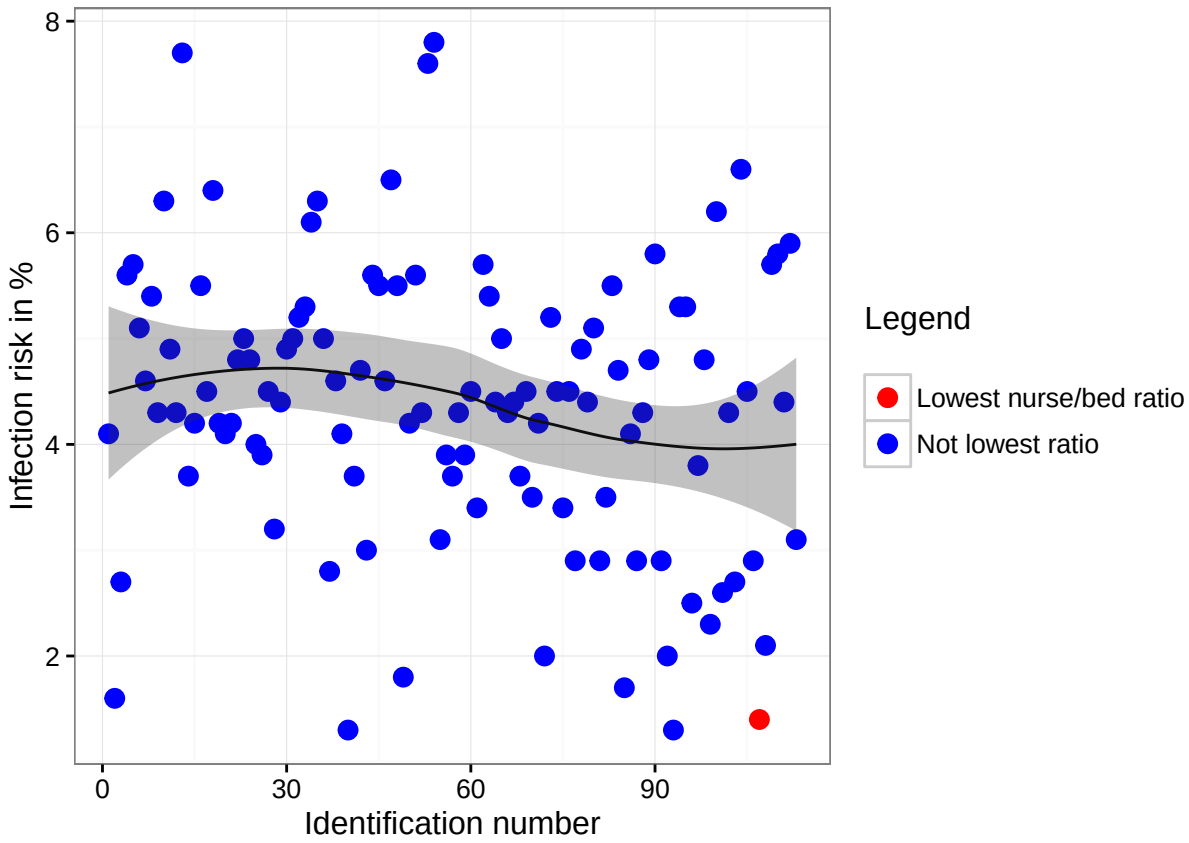
The hospital id 107 is the least well staffed hospital of the lot.

Now we make a scatterplot over infection risk against hospital identification number. We also mark the least well staffed hospital in the scatterplot.



It seems as if the least well staffed hospital has one of the lowest infection risks, according to the data. It could be the case that nurses are good infectious disease vectors.

Finally, we plot the predicted values of a smoother for the scatterplot and its 95% pointwise confidence band.



It would appear as if a horizontal line could fit inside the band. It could mean that the Infection risk is not correlated with hospital identification number.

Appendix

R code

```
## library(MASS)
## df1=aggregate(Price~Type, data=Cars93, FUN=mean)
## barplot(df1$Price, names.arg=df1$Type)
library(ggplot2)
library(fANCOVA)
library(gridExtra)
senicdata <- read.csv2("SENIC.csv")
#head(senicdata)
#X7 and X8 are the qualitative variables
qualplots1 <- ggplot(senicdata, aes(x = factor(X7))) + geom_bar(fill = "cyan", width = 0.3) +
  xlab("Medical School Affiliation") +
ggtitle("Medical school affiliation of hospitals") + scale_x_discrete(labels = c("YES","NO"))
qualplots2 <- ggplot(senicdata, aes(x = factor(X8))) + geom_bar(fill = "magenta", width = 0.3) +
  xlab("US Geographic Region") +
  ggtitle("General region of hospitals") + scale_x_discrete(labels = c("NE","NC","S","W"))
listplots <- list(qualplots1,qualplots2)
grobplot <- arrangeGrob(grobs = listplots, nrow = 2)
```

```r
plot(grobplot)

senicquants <- senicdata[,-(8:9)]
senicquants$Obs <- rep(113,113)
senicquants$Obs <- as.factor(senicquants$Obs)


quantplots1 <- ggplot(senicquants) + geom_boxplot(aes(x = Obs, y = X1),width = 0.3) +
  ggtitle("avg. Length of Stays in days") +
ylab("Days") + xlab("No. of observations")
quantplots2 <- ggplot(senicquants) + geom_boxplot(aes(x = Obs, y = X2),width = 0.3) +
  ggtitle("avg. Age of patients in years") +
  ylab("Years") + xlab("No. of observations")
quantplots3 <- ggplot(senicquants) + geom_boxplot(aes(x = Obs, y = X3),width = 0.3) +
  ggtitle("avg. est. prob. of infection, in percent") +
  ylab("Percent") + xlab("No. of observations")
quantplots4 <- ggplot(senicquants) + geom_boxplot(aes(x = Obs, y = X4),width = 0.3) +
  ggtitle("Routine culturing ratio / 100") +
  ylab("Ratio") + xlab("No. of observations")
quantplots5 <- ggplot(senicquants) + geom_boxplot(aes(x = Obs, y = X5),width = 0.3) +
  ggtitle("Routine chest X-ray ratio / 100") +
  ylab("Ratio") + xlab("No. of observations")
quantplots6 <- ggplot(senicquants) + geom_boxplot(aes(x = Obs, y = X6),width = 0.3) +
  ggtitle("avg. no. of beds in hospital") +
  ylab("Beds") + xlab("No. of observations")
quantplots9 <- ggplot(senicquants) + geom_boxplot(aes(x = Obs, y = X9),width = 0.3) +
  ggtitle("avg. no. of patients in hospital") +
  ylab("Patients") + xlab("No. of observations")
quantplots10 <- ggplot(senicquants) + geom_boxplot(aes(x = Obs, y = X10),width = 0.3) +
  ggtitle("avg. no. of full-time nurses in hospital") +
  ylab("Nurses") + xlab("No. of observations")
quantplots11 <- ggplot(senicquants) + geom_boxplot(aes(x = Obs, y = X11),width = 0.3) +
  ggtitle("percentage of facilities in hospital") +
  ylab("Percent") + xlab("No. of observations")
listquantplots <- list(quantplots1,quantplots2,quantplots3,quantplots4,quantplots5,quantplots6,
                       quantplots9,quantplots10,quantplots11)
grobquantplot <- arrangeGrob(grobs = listquantplots, nrow = 3)

plot(grobquantplot)
## senicdata <- cbind(senicdata, senicdata$X10 / senicdata$X6)
## sort(senicdata[,13],index.return = TRUE)$ix
senicdata <- cbind(senicdata, senicdata$X10 / senicdata$X6)
#sort(senicdata[,13],index.return = TRUE)$ix
#this shows how well staffed the nurses are relative to no of beds
#hospital 107 is the least well staffed
ident <- rep("red",113)
ident[107] <- "blue"
scatterpl <- ggplot(data = senicdata, aes(Obs,X3) ) + geom_point(aes(colour = ident),size = 3) +
  labs(y = "Infection risk in %", x ="Identification number",color="Legend\n") +
  scale_color_manual(labels = c("Lowest nurse/bed ratio", "Not lowest ratio"), values = c("red", "blue")
  theme_bw()
scatterpl
mod=loess.as(senicdata$Obs, senicdata$X3, criterion="gcv", degree=2,
```

```
             plot=FALSE)
result=predict(mod, se=TRUE)
#head(result)
newframe <-cbind(senicdata, data.frame(Obs = senicdata$Obs, fitted = result$fit,
                         upper = result$fit + 1.96*result$se.fit,
                         lower = result$fit - 1.96*result$se.fit))

scatterpl <- scatterpl + geom_line(data = newframe, mapping = aes(Obs,fitted)) +
  geom_ribbon(data = newframe,mapping = aes(ymin = lower,ymax = upper),alpha=0.3)
scatterpl
##
```