Visualization Group lab1

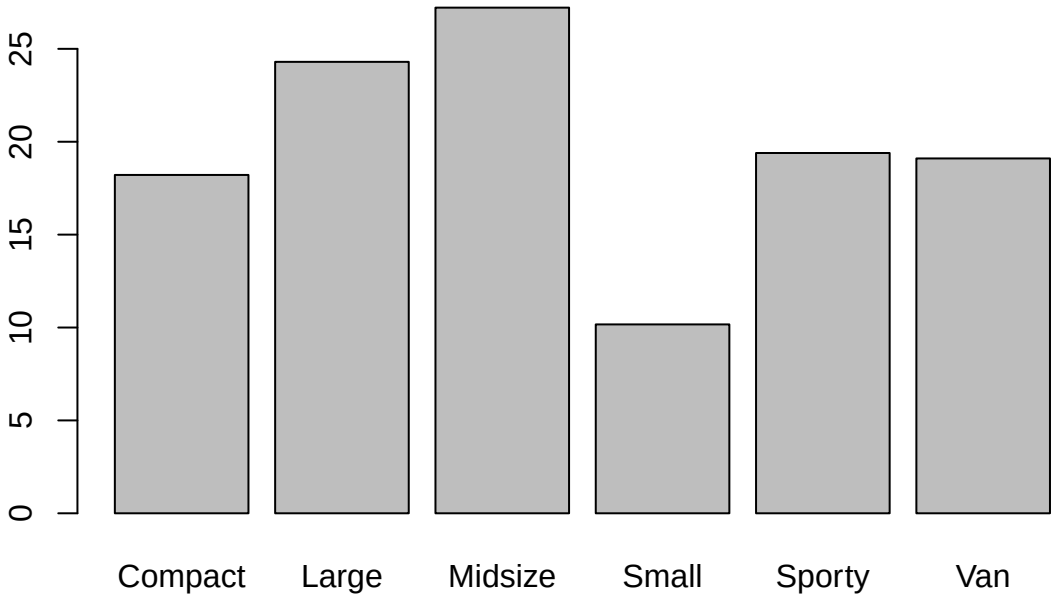
Andrea Bruzzone, Araya Eamrurksiri, Maxime Bonneau, and Thomas Zhang

August 31, 2016

Assignment1

Using the data set `Cars93`, in which cars were selected at random among 1993 passenger car models, a study of the prices of the cars based on their type is conducted.

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



Use `Inkscape` to enhance the plot as follows:

a. Increase fonts
b. Make proper axis labels
c. Rotate Y axis values 90 degrees
d. Add caption
e. Change the color of the bars
f. Add some graphics that would emphasize that mid-sized cars cost much more than small cars

The resulting plot is shown below.

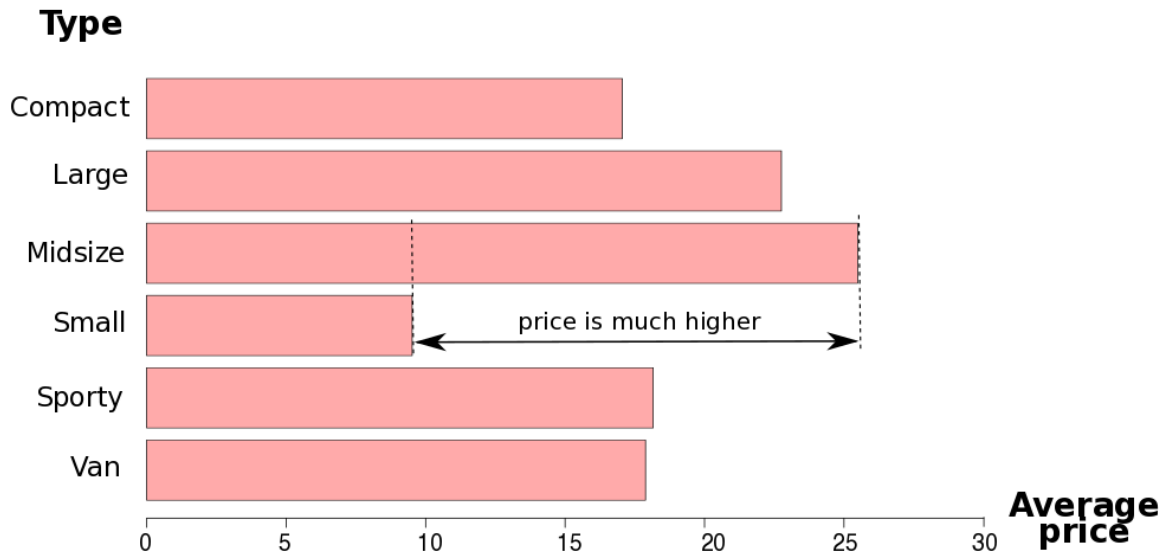
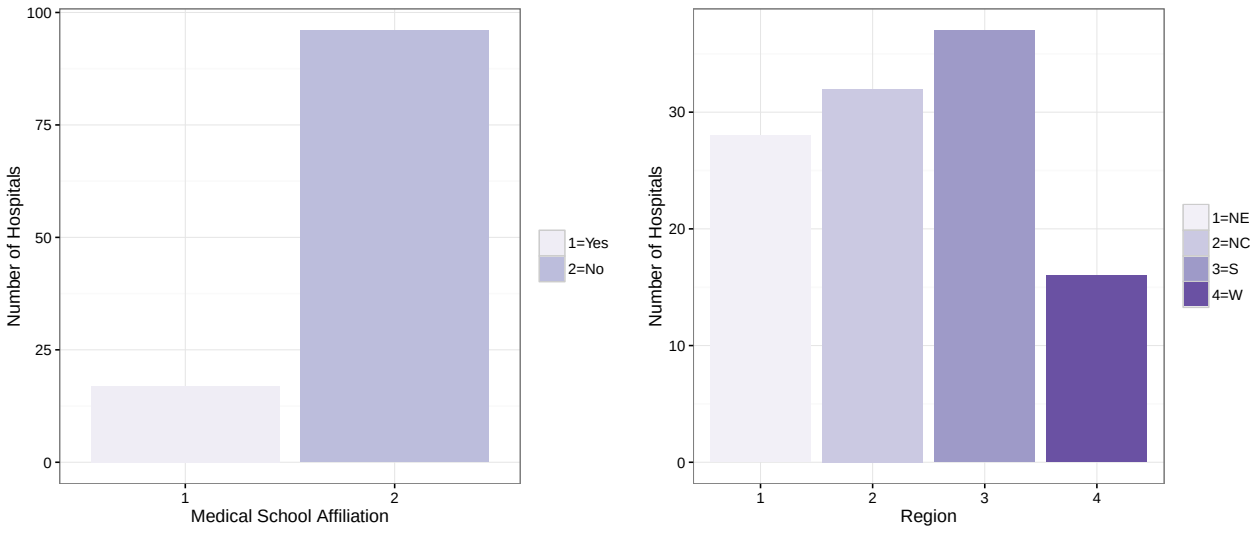


Figure 1: Data from 93 Cars on Sale in the USA in 1993

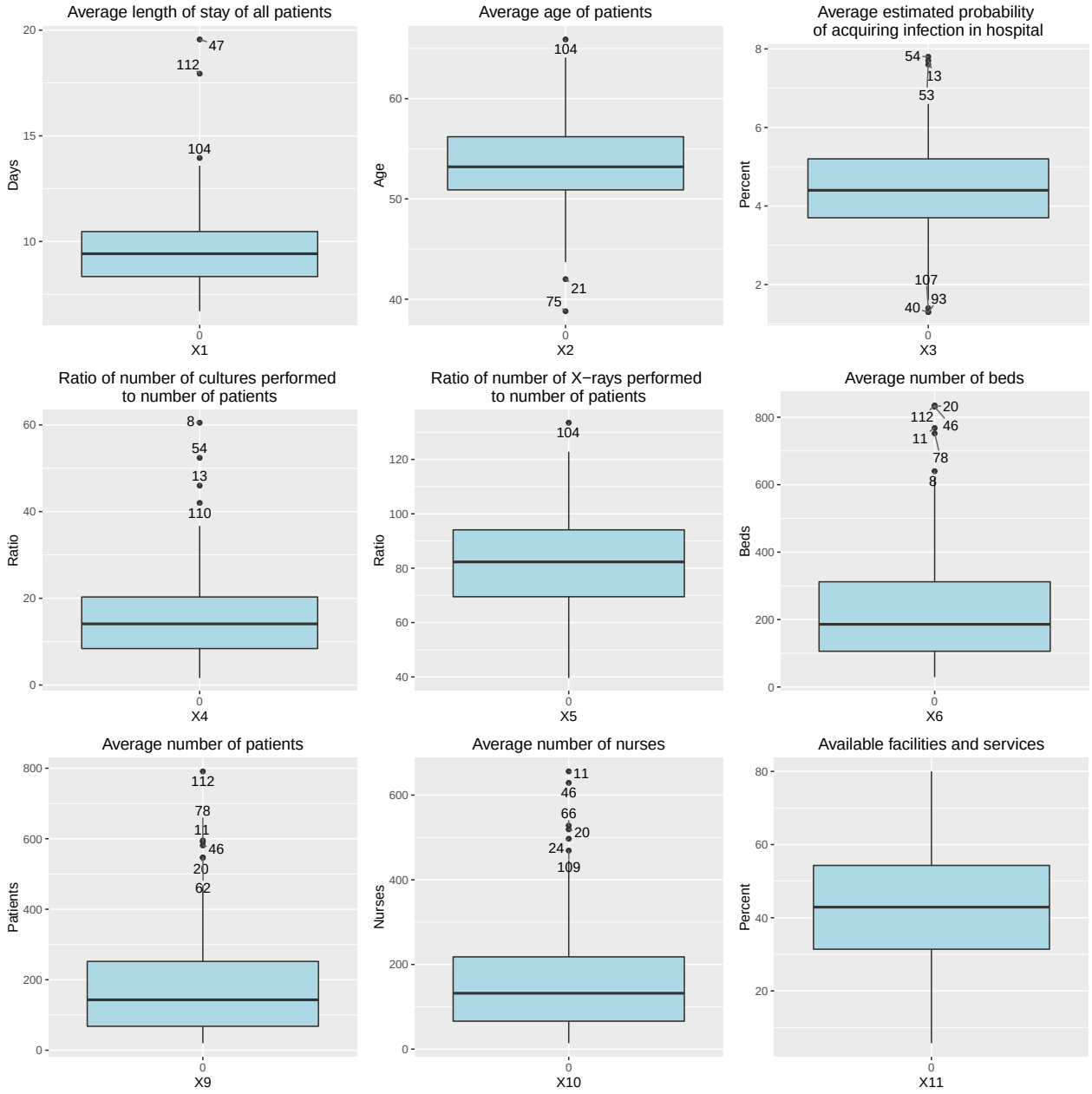
Assignment2

Data set `SENIC.csv` describes the results of measurements taken at different US hospitals. It contains a random sample of $n = 113$, the ID number and 11 other variables.

Then, split the variables into qualitative and quantitative variables and produce nice bar charts and boxplots, respectively, using package `ggplot2`.



It can be seen that most hospitals do not have a medical school affiliation. Less than 25 percent of hospitals have operated a medical school affiliation. From the second bar chart, it is possible to see the distribution of the hospitals selected in this study, most of them are from the South, less than 20 instead for the hospitals from the West US.



It can be seen that an outlier seems to be present for a lot of variables. Among all the variables in this data set, $X11$, or available facilities and services, is the only variable with no outlier.

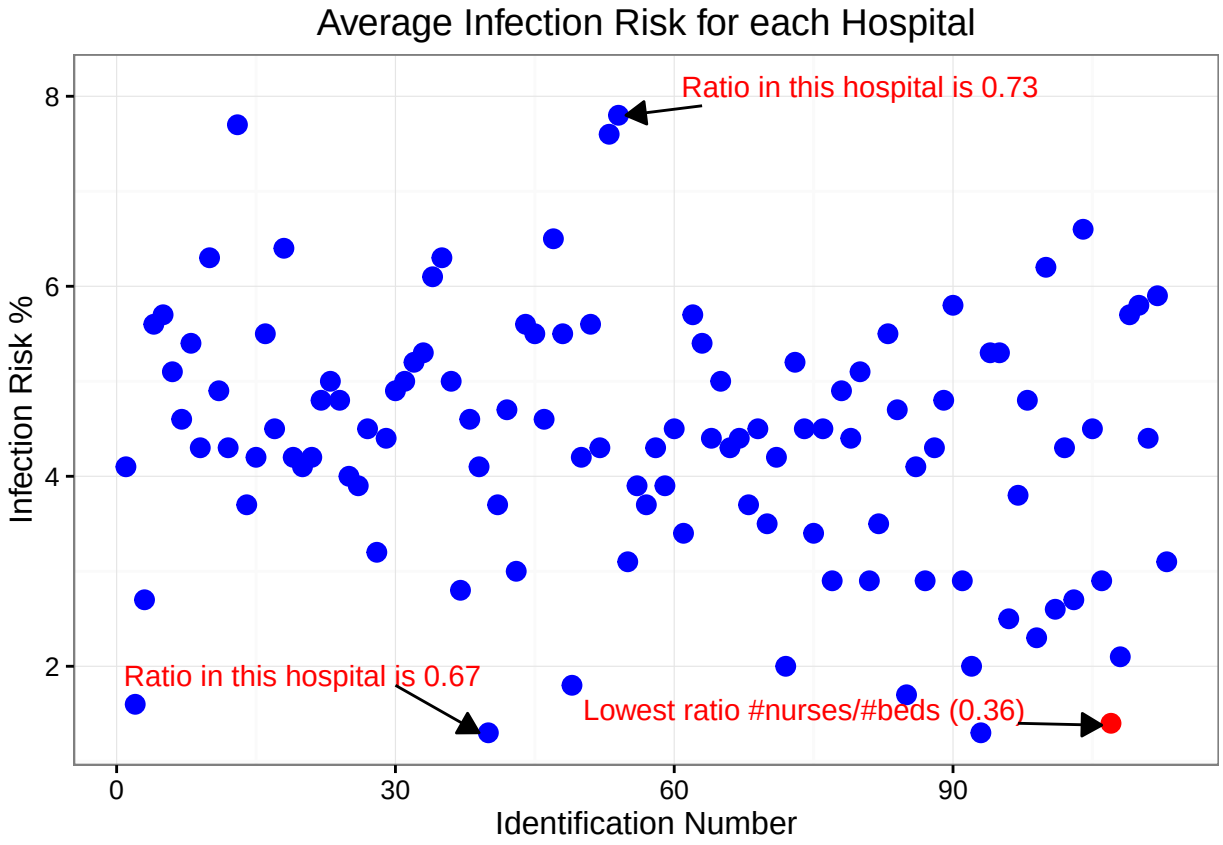
The interesting question here is whether we have correlations in outliers between the boxplots. A quick check indicates that so may be the case. For instance, Hospital id numbers 47, 53 and 104 all have top ten positions in variables length of stay, age and infeciton risk.

Furthermore, we can see that patients who get a treatment from hospital id numbers 112 stay in the hospital almost three weeks. This hospital seems to be a district hospital as it has a large number of beds and patients for intensive and long-term care.

Assignment3

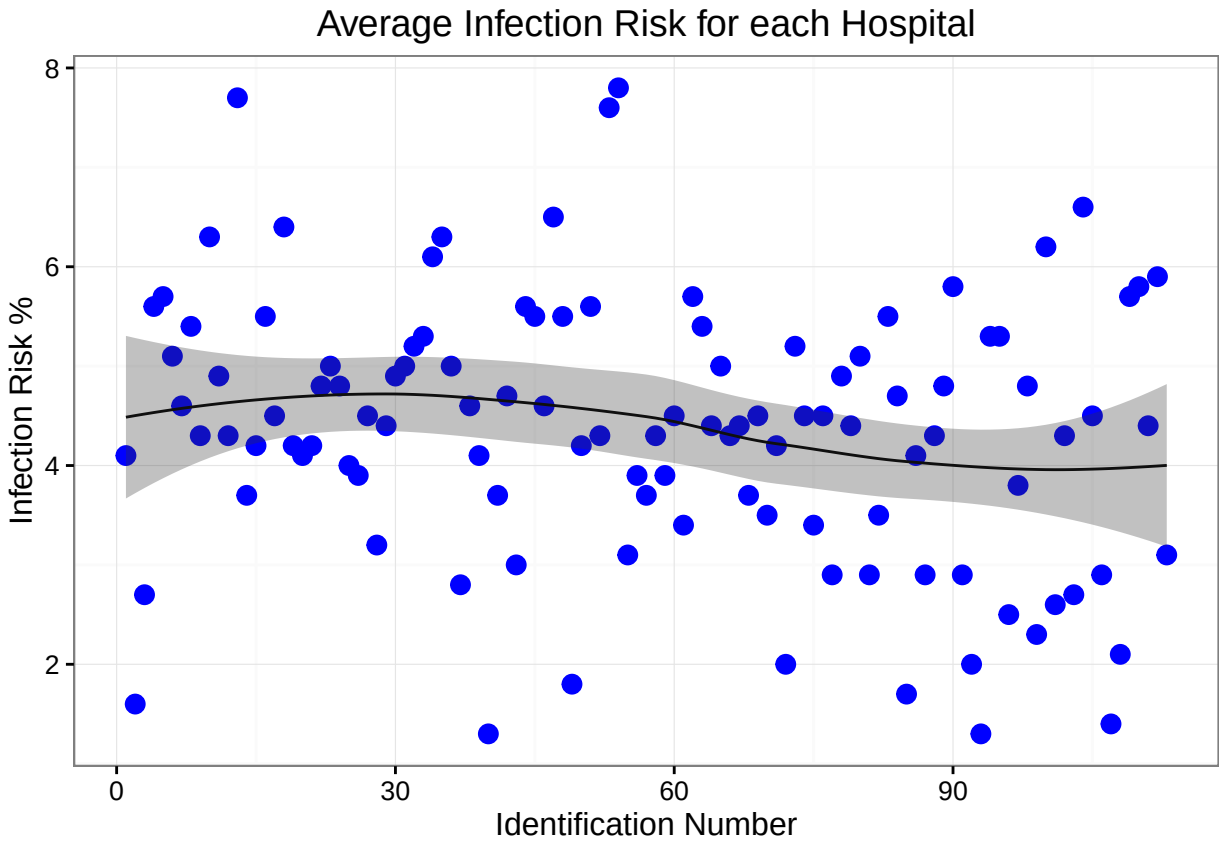
We write a code that finds out in which hospital the ratio "Number of nurses/Number of beds" is the lowest. This ratio tells us how well staffed the nurses at the hospital are relative to number of hospital beds. It also shows for each bed how many nurses there are in the hospital, that is the same as saying how many nurses for each patient

Hospital ID number 107 has got the lowest ratio, which is 0.1652174.



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

Contribution

....

R Code

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

## ----echo=FALSE, fig.height=5, fig.width=12-----------------------------
library(png)
library(grid)
img <- readPNG("drawing.png")
grid.raster(img)
```

```
## ---- echo=FALSE, fig.height=5, fig.width=12-----------------------------
#1
library(ggplot2)
library(gridExtra)
senic <- read.csv2("/Users/lynn/Documents/LiU/732A98 Visualization/Lab1/Senic.csv")

#2
#qualitative variable
g1 <- ggplot(data=senic) +
    geom_bar(aes(as.factor(X7), fill=as.factor(X7))) +
    xlab("Medical School Affiliation") + ylab("Number of Hospitals") +
    scale_fill_brewer(name = "", labels = c("1=Yes", "2=No"), palette="Purples") +
    theme_bw()

g2 <- ggplot(data=senic) +
    geom_bar(aes(as.factor(X8), fill=as.factor(X8))) +
    xlab("Region") + ylab("Number of Hospitals") +
    scale_fill_brewer(name = "", labels = c("1=NE", "2=NC", "3=S", "4=W"), palette="Purples") +
    theme_bw()

grid.arrange(g1, g2, ncol=2)

## ---- echo=FALSE, fig.height=12, fig.width=12----------------------------
#3
#quantitative variable
library(ggrepel)

is_outlier <- function(x) {
    return(x < quantile(x, 0.25) - 1.5 * IQR(x) | x > quantile(x, 0.75) + 1.5 * IQR(x))
}

pl <- list()
num <- c(1:6,9:11)
var <- paste0("X",num)
y <- c("Days","Age","Percent","Ratio","Ratio","Beds","Patients","Nurses","Percent")
title <- c("Average length of stay of all patients","Average age of patients",
           "Average estimated probability \nof acquiring infection in hospital",
           "Ratio of number of cultures performed \nto number of patients",
           "Ratio of number of X-rays performed \nto number of patients","Average number of beds",
           "Average number of patients","Average number of nurses",
           "Available facilities and services")

#Boxplot for each variable
for(i in 1:9){
    pl[[i]] <- ggplot(data=senic, aes_string(x=factor(0),y=var[i])) +
        geom_boxplot(fill="lightblue") + xlab(var[i]) + ylab(y[i]) +
        ggtitle(title[i])
}

#Indicate outlier in each variable
outlier <- matrix(NA,9,113)
for(k in 1:9){
    outlier[k,] <- ifelse( is_outlier(senic[[var[k]]]), senic$Obs, as.numeric(NA) )
```

```
}

p1 <- pl[[1]] + geom_text_repel(aes(label=outlier[1,]), na.rm=TRUE)
p2 <- pl[[2]] + geom_text_repel(aes(label=outlier[2,]), na.rm=TRUE)
p3 <- pl[[3]] + geom_text_repel(aes(label=outlier[3,]), na.rm=TRUE)
p4 <- pl[[4]] + geom_text_repel(aes(label=outlier[4,]), na.rm=TRUE)
p5 <- pl[[5]] + geom_text_repel(aes(label=outlier[5,]), na.rm=TRUE)
p6 <- pl[[6]] + geom_text_repel(aes(label=outlier[6,]), na.rm=TRUE)
p7 <- pl[[7]] + geom_text_repel(aes(label=outlier[7,]), na.rm=TRUE)
p8 <- pl[[8]] + geom_text_repel(aes(label=outlier[8,]), na.rm=TRUE)
p9 <- pl[[9]] + geom_text_repel(aes(label=outlier[9,]), na.rm=TRUE)

grid.arrange(p1, p2, p3, p4, p5, p6, p7, p8, p9, ncol=3)


## ---- echo=FALSE-------------------------------------------------------------
#1
dat <- senic$X10 / senic$X6

## ---- echo=FALSE-------------------------------------------------------------
attach(senic)
#X12[which.max(X3)] #ratio of the hospital with highest Infection Risk (a little bit more than 1 bed pe
#X12[which.min(X3)] #ratio of the hospital with minimum Infection Risk

color <- rep("blue", dim(senic)[1])
color[107] <- "red"

ggplot(senic, aes(x = Obs, y = X3)) + geom_point(size = 3, col = color) +
  ylab("Infection Risk %") + xlab("Identification Number") +
  ggtitle("Average Infection Risk for each Hospital") + theme_bw() +
  annotate("text", x = 74, y = 1.55, size = 4,
           label = "Lowest ratio #nurses/#beds (0.36)", color="red") +
  annotate("segment", x = 97, xend = 106, y = 1.4, yend = 1.38 , arrow = arrow(length = unit(.3, "cm"),
  annotate("text", x = 80, y = X3[which.max(X3)] + 0.3, size = 4,
           label = "Ratio in this hospital is 0.73", color="red") +
  annotate("segment", x = 63, xend = 54.8, y = 7.9, yend = X3[which.max(X3)], arrow = arrow(length = uni
  annotate("text", x = 20, y = X3[which.min(X3)] + 0.6, size = 4,
           label = "Ratio in this hospital is 0.67", color="red") +
  annotate("segment", x = 30, xend = 39, y = 1.8, yend = X3[which.min(X3)], arrow = arrow(length = unit


## ---- echo=FALSE, warning=FALSE, message=FALSE-------------------------------
library(fANCOVA)
mod <- loess.as(Obs, X3, criterion="gcv", degree=2)
result <- predict(mod, se=TRUE)

predframe <- data.frame(Obs, Pred. = result$fit, lwr = result$fit - 1.96*result$se.fit,
                        upr = result$fit + 1.96*result$se.fit)


ggplot(senic, aes(x = Obs, y = X3)) +
  geom_point(size = 3, color = "blue") +
  geom_line(data = predframe, aes(x = Obs, y = predframe$Pred.)) +
```

```
  ylab("Infection Risk %") + xlab("Identification Number") +
  ggtitle("Average Infection Risk for each Hospital") + theme_bw() +
  geom_ribbon(data = predframe, aes(ymin = lwr,ymax = upr), alpha = 0.3)


## ----code=readLines(knitr::purl('group5_lab1.Rmd', documentation = 1)), eval = FALSE----
## NA
```