

Statistics 135 – Lab Project

April 27, 2015

1 Background

```
library("ggplot2")
library("grid")
library("gridExtra")

data.df <- na.omit(read.csv("diabetes.csv"))
```

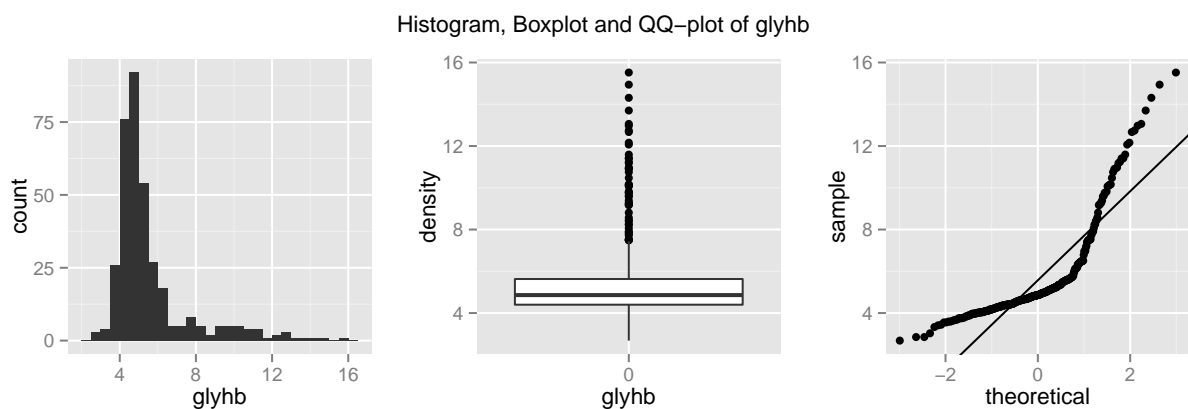
2 Accessing Data, Visualization and Summarization

```
1. glyhb.histogram <- ggplot(data.df) +
  geom_histogram(aes(x=glyhb), binwidth=0.5)

glyhb.boxplot <- ggplot(data.df) +
  geom_boxplot(aes(x=factor(0), y=glyhb)) +
  labs(x="glyhb", y="density")

glyhb.qqplot <- ggplot(data.df) +
  stat_qq(aes(sample=glyhb)) +
  geom_abline(aes(intercept=mean(glyhb), slope=sd(glyhb)))

grid.arrange(glyhb.histogram, glyhb.boxplot, glyhb.qqplot, ncol=3,
  main="Histogram, Boxplot and QQ-plot of glyhb")
```



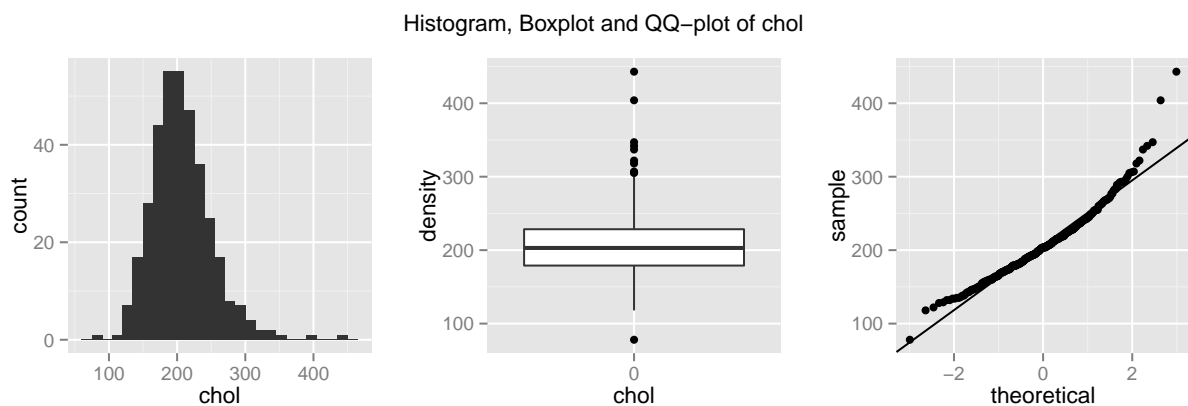
```
2. chol.histogram <- ggplot(data.df) +
  geom_histogram(aes(x=chol), binwidth=15)

chol.boxplot <- ggplot(data.df) +
```

```
geom_boxplot(aes(x=factor(0), y=chol)) +
labs(x="chol", y="density")

chol.qqplot <- ggplot(data.df) +
  stat_qq(aes(sample=chol)) +
  geom_abline(aes(intercept=mean(chol), slope=sd(chol)))

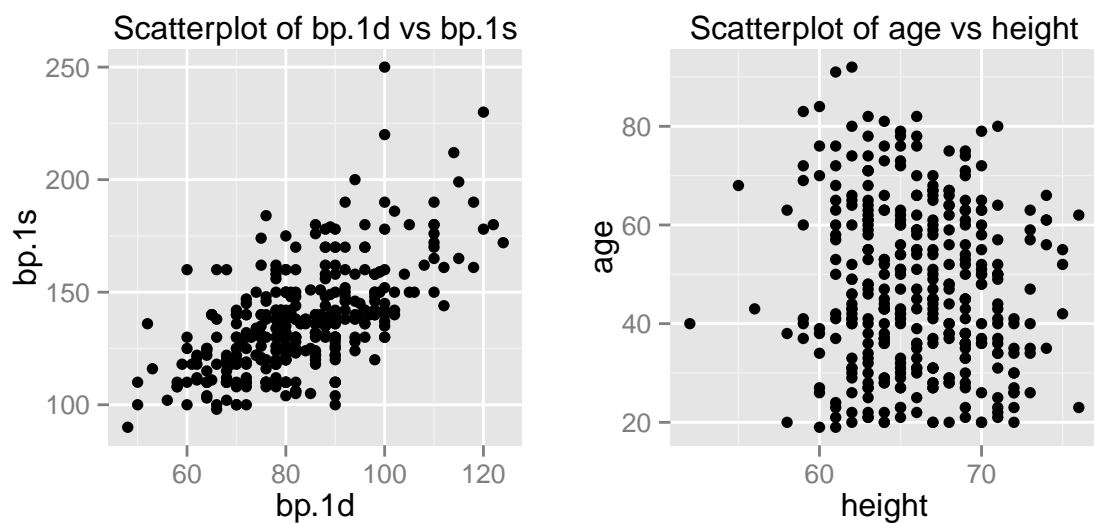
grid.arrange(chol.histogram, chol.boxplot, chol.qqplot, ncol=3,
  main="Histogram, Boxplot and QQ-plot of chol")
```



```
3. bp.1d.bp.1s.scatterplot <- ggplot(data.df) +
  geom_point(aes(x=bp.1d, y=bp.1s)) +
  labs(title="Scatterplot of bp.1d vs bp.1s") +
  theme(plot.title=element_text(size=rel(1)))

height.age.scatterplot <- ggplot(data.df) +
  geom_point(aes(x=height, y=age)) +
  labs(title="Scatterplot of age vs height") +
  theme(plot.title=element_text(size=rel(1)))

grid.arrange(bp.1d.bp.1s.scatterplot, height.age.scatterplot, ncol=2)
```



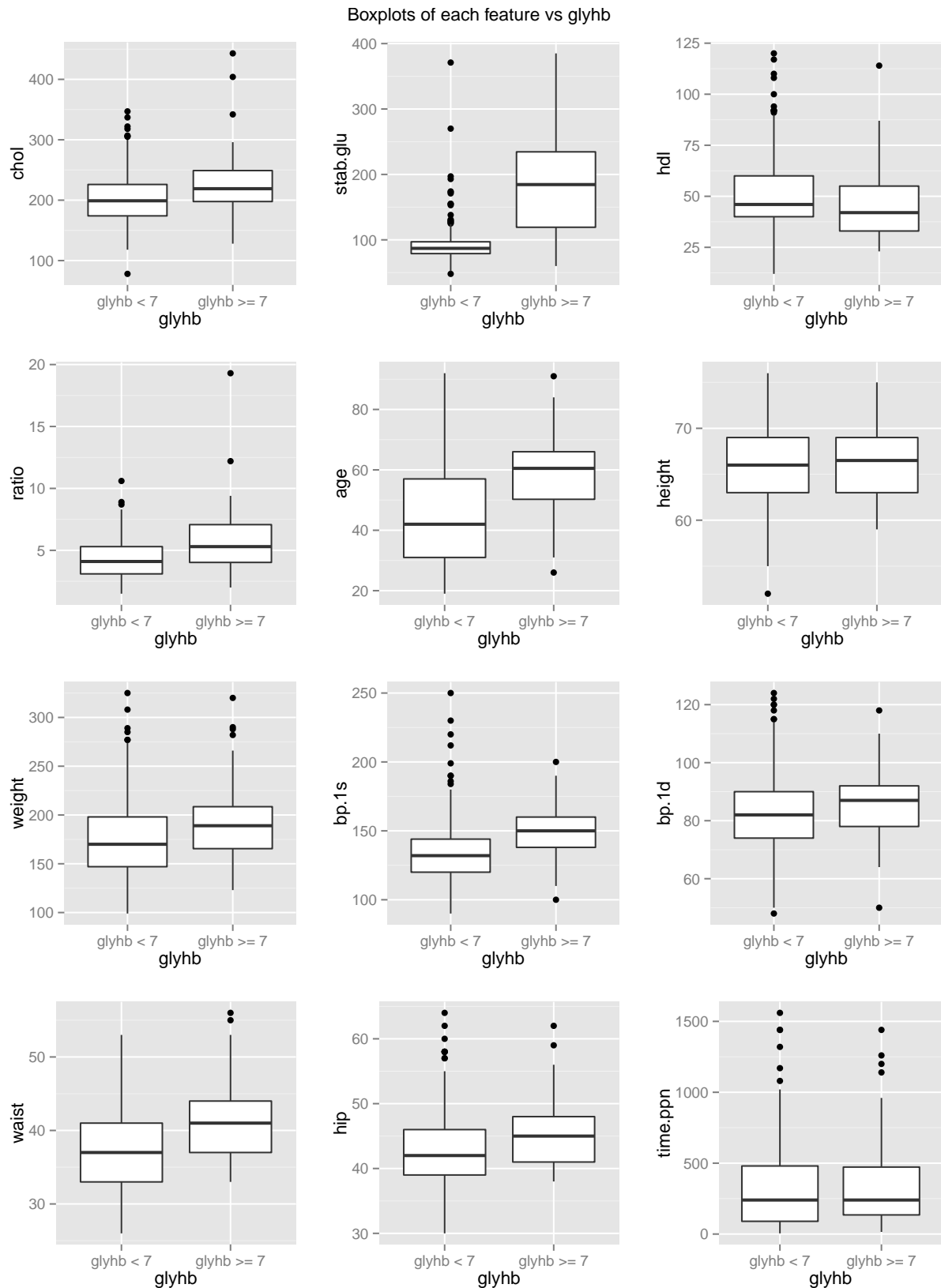
```

4. data.df$glyhb.cond <- NA
data.df[data.df$glyhb>=7, ]$glyhb.cond <- "glyhb >= 7"
data.df[data.df$glyhb<7, ]$glyhb.cond <- "glyhb < 7"

features <- c("chol", "stab.glu", "hdl", "ratio", "age", "height", "weight",
              "bp.1s", "bp.1d", "waist", "hip", "time.ppn")
feature.boxplots = list()
for (i in 1:length(features)) {
  feature <- features[i]
  feature.boxplot <- ggplot(data.frame(glyhb.cond=data.df$glyhb.cond,
                                       feature=data.df[, feature])) +
    geom_boxplot(aes(x=glyhb.cond, y=feature)) +
    labs(x="glyhb", y=feature)
  feature.boxplots[[i]] <- feature.boxplot
}

do.call(grid.arrange, c(feature.boxplots, ncol=3,
                        main="Boxplots of each feature vs glyhb"))

```

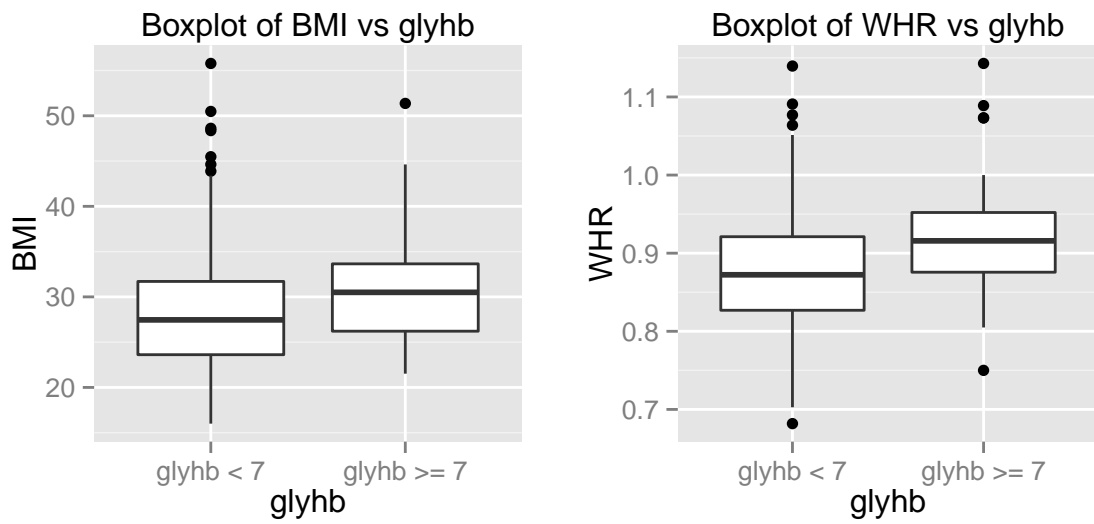


```
5. data.df$BMI <- 703*data.df$weight/data.df$height^2
   data.df$WHR <- data.df$waist/data.df$hip
```

```
features <- c("BMI", "WHR")
feature.boxplots = list()
for (i in 1:length(features)) {
  feature <- features[i]
```

```
feature.boxplot <- ggplot(data.frame(glyhb.cond=data.df$glyhb.cond,
                                     feature=data.df[, feature])) +
  geom_boxplot(aes(x=glyhb.cond, y=feature)) +
  labs(title=paste(c("Boxplot of", feature, "vs glyhb"), collapse=" "),
       x="glyhb", y=feature) +
  theme(plot.title=element_text(size=rel(1)))
feature.boxplots[[i]] <- feature.boxplot
}
```

```
do.call(grid.arrange, c(feature.boxplots, ncol=2))
```



6.

3 Parametric Inference

1.

2.

3.

4 Testing

1.

2.

3.

4.

5.

6.

5 Regression

- 1.
- 2.
- 3.
- 4.
- 5.
- 6.
- 7.