Stat 5810, Section 003
Statistical Visualization I
Fall 2018
Homework 1

ShaunMicheal Bartschi

A01975136

October 17, 2018

Stat 5810 Statistical Visualization I Fall 2018

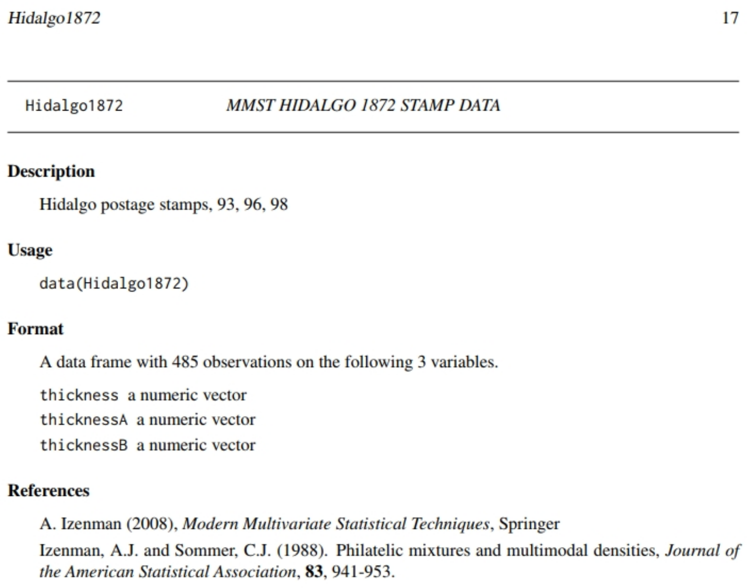
Homework Assignment 1 (10/6/2018)

60 Points — Due Wednesday 10/17/2018 (via Canvas by 11:59pm)

(i) (48 Points) In this question, you have to work with the *Hidalgo1872.rda* data set that is provided in Canvas. You can also directly download the data from `https://github.com/cran/MMST/tree/master/data`. Do not use any other version of this data set you may find on the internet or your results may differ slightly as different versions of the data set seem to exist. This data set was originally a part of the *MMST* R package, but that package was removed from the CRAN repository.

The data set consists of 485 rows and 3 columns. The first column (*thickness*) contains the thickness of 485 Mexican stamps, measured to a precision of a thousandth of a millimeter. It is my understanding that column 2 (*thicknessA*) repeats these measurements if the stamp was pressed in 1872 (and contains NA otherwise). Column 3 (*thicknessB*) repeats these measurements if the stamp was pressed in 1873/74 (and contains NA otherwise). There are a few data entry errors where both columns 2 and 3 contain a measurement.

The full (?!?) description of the data set, as originally provided in the *MMST* R package, is shown in Figure 1.



Hidalgo1872 17

`Hidalgo1872` *MMST HIDALGO 1872 STAMP DATA*

Description

Hidalgo postage stamps, 93, 96, 98

Usage

`data(Hidalgo1872)`

Format

A data frame with 485 observations on the following 3 variables.

`thickness` a numeric vector
`thicknessA` a numeric vector
`thicknessB` a numeric vector

References

A. Izenman (2008), *Modern Multivariate Statistical Techniques*, Springer

Izenman, A.J. and Sommer, C.J. (1988). Philatelic mixtures and multimodal densities, *Journal of the American Statistical Association*, **83**, 941-953.

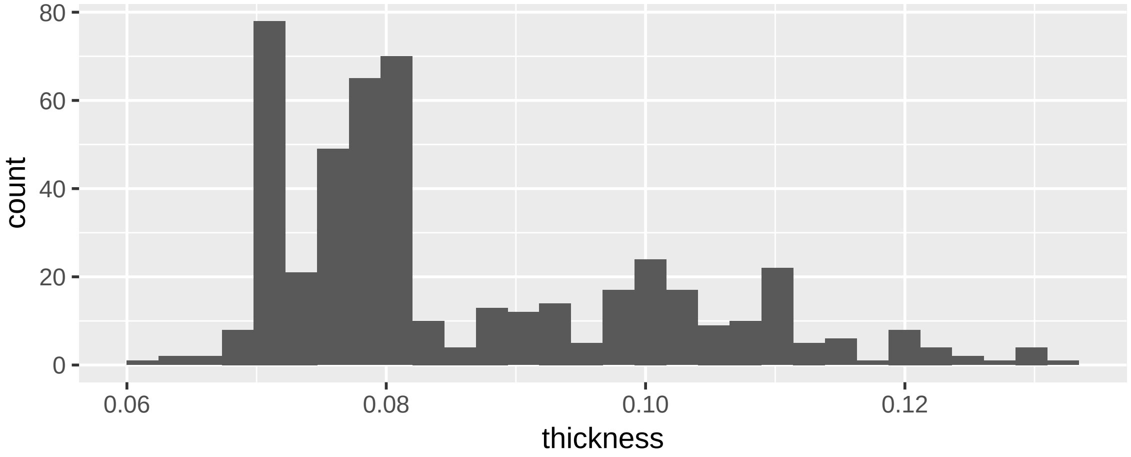
Figure 1: Description of the Hidalgo 1872 data set, obtained from `https://mran.microsoft.com/snapshot/2014-09-30/web/packages/MMST/MMST.pdf`.

(a) (1 Point) Load the Hidalgo1872 data set and all required R packages to answer this question. Show your R code.

```
> setwd("C:/Users/Shaun/Desktop/StatVis/HW1")
> load("Hidalgo1872.rda")
> library(ggplot2)
> library(ggthemes)
> library(grid)
> library(gridExtra)
```

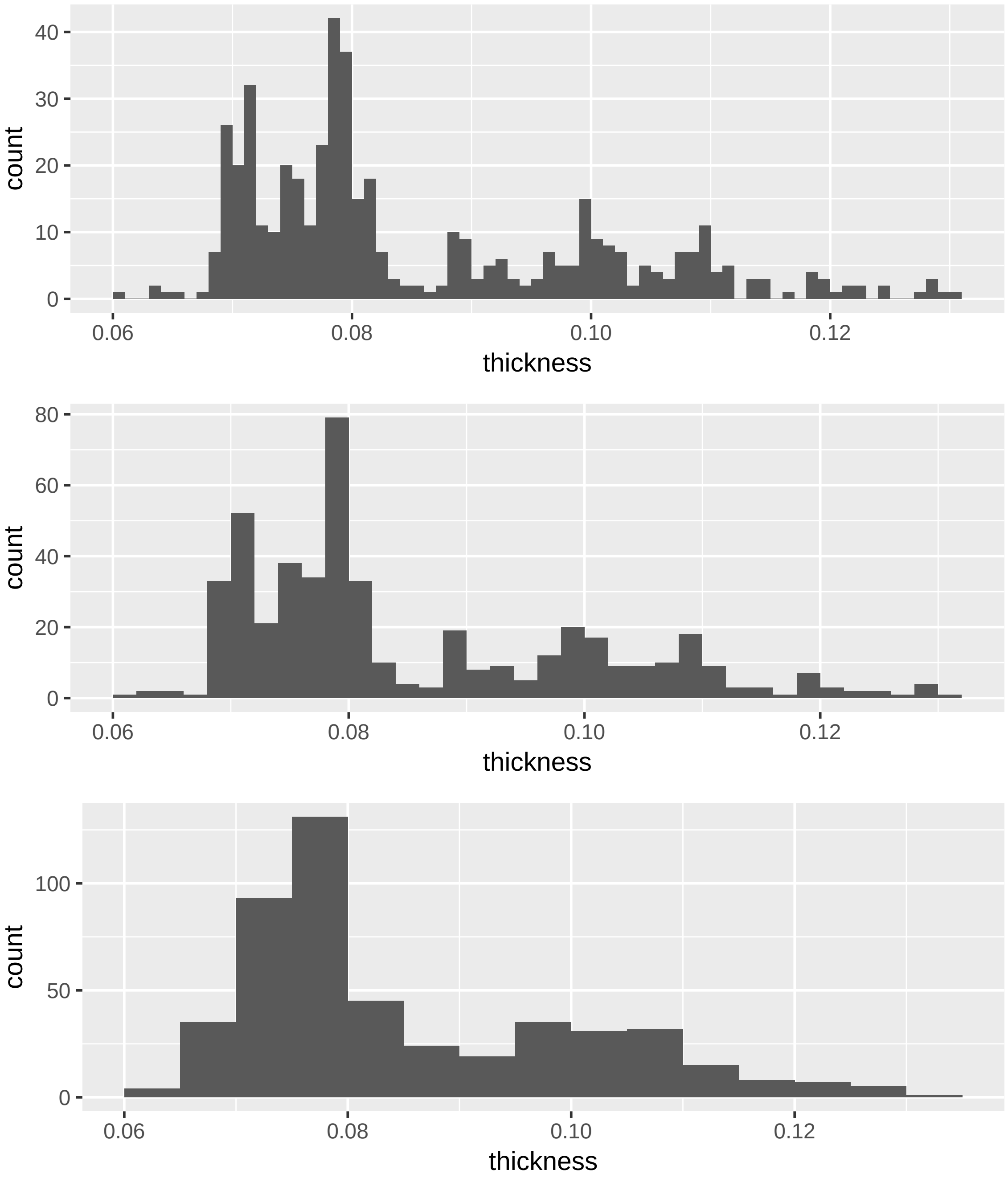
(b) (2 Points) Draw a basic histogram for *thickness* using ggplot2. Include your R code and the resulting graph.

```
> ggplot(Hidalgo1872,aes(x=thickness)) + geom_histogram()
```



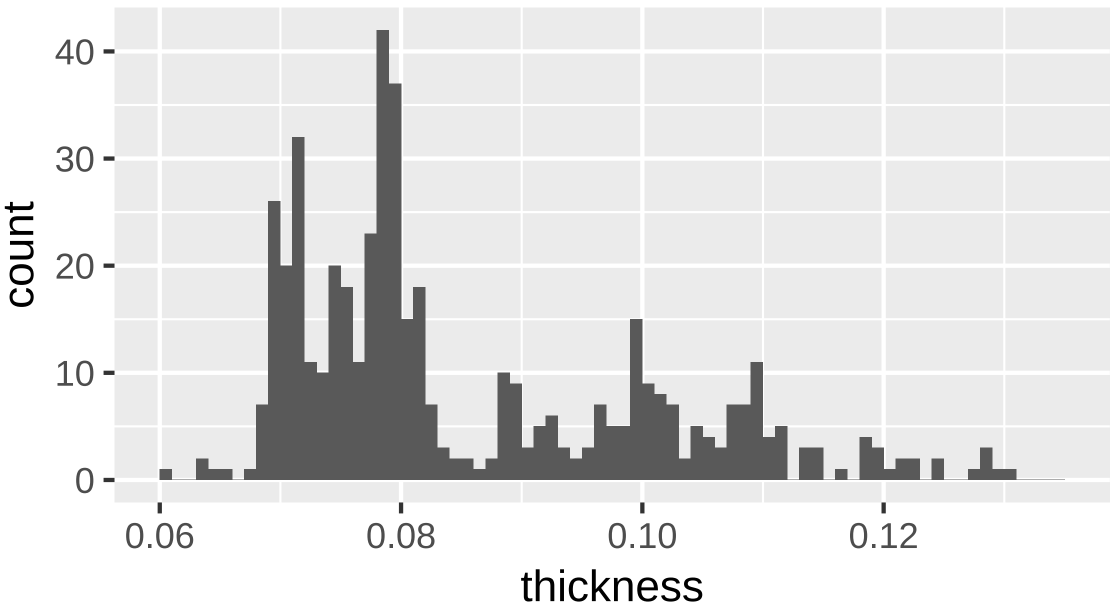
(c) (6 Points) Further improve your histogram from (b). Try three different binwidths: 0.001, 0.002, and 0.005 and use 0.0005, 0.001, and 0.0025 as the center, respectively. You should also adjust the range of the horizontal and vertical axes, labels, title, etc. **Clearly indicate which changes you made and why you made these changes.** As in class, make these changes step by step and continue with further adjustments until your graph is ready for publication. Include your three final graphs for these three binwidths and the R code for these finals graph. No need to include any intermediate graphs and the R code for those. Hint: When you get warnings from ggplot2, check carefully. It is easy to cut off parts of the histogram on the horizontal or vertical axis, in particular when you copy and paste your R code and forget to adjust some of the arguments.

```
> #Binwidth = 0.001
> p1 <- ggplot(Hidalgo1872,aes(x=thickness)) +
+   geom_histogram(binwidth=0.001, center = 0.0005)
> #Binwidth = 0.002
> p2 <- ggplot(Hidalgo1872,aes(x=thickness)) +
+   geom_histogram(binwidth=0.002, center = 0.001)
> #Binwidth = 0.005
> p3 <- ggplot(Hidalgo1872,aes(x=thickness)) +
+   geom_histogram(binwidth=0.005, center = 0.0025)
> #plot
> grid.arrange(p1, p2, p3, nrow = 3)
```

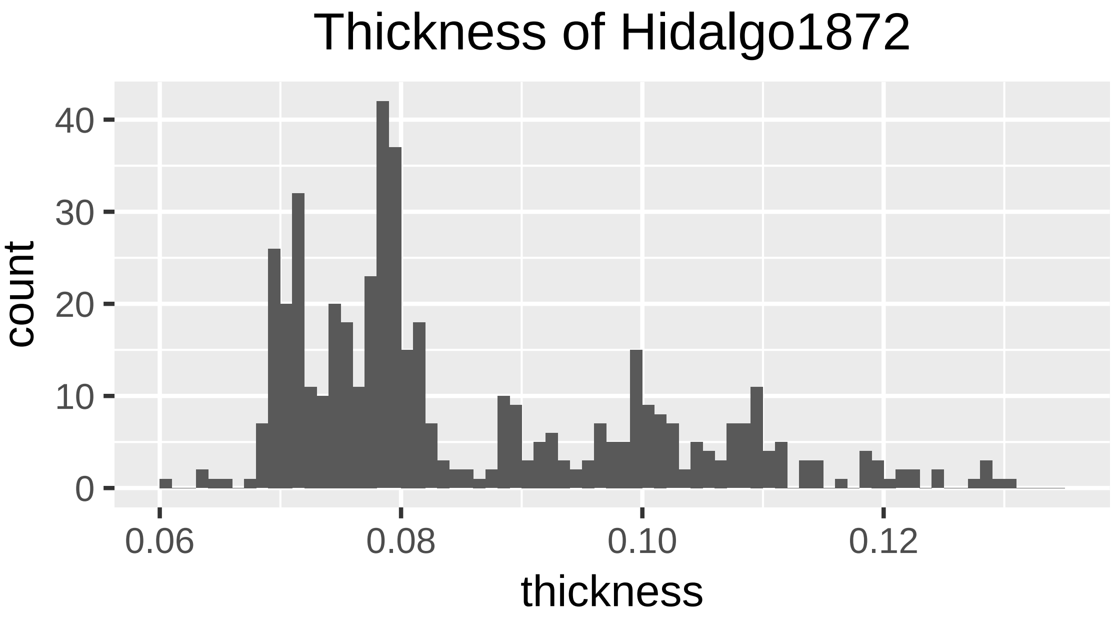


Now to improve, start with improving a single graph:

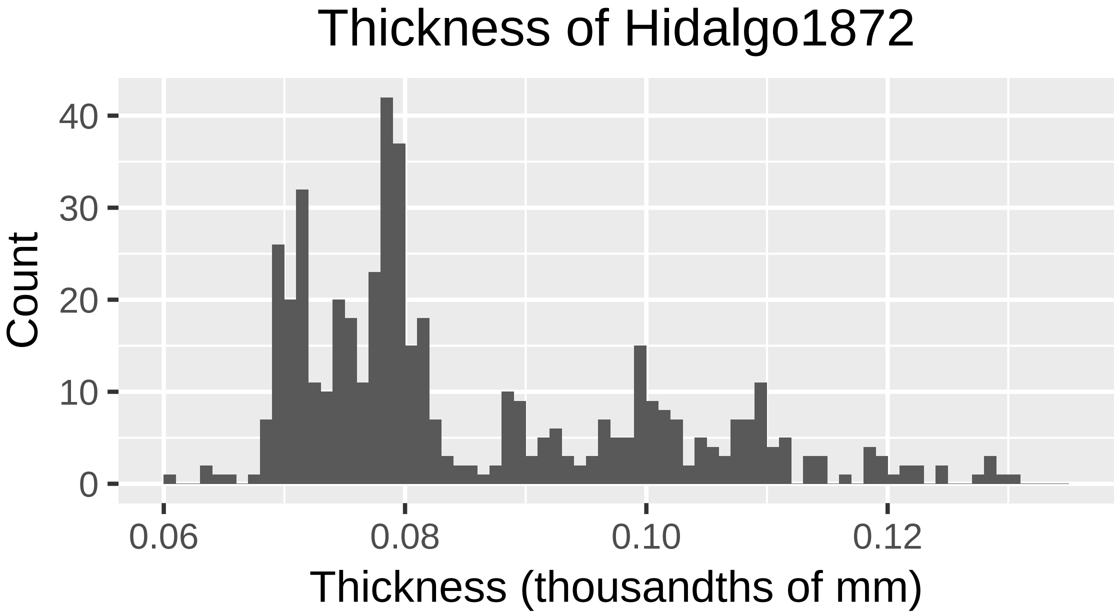
```
> #Binwidth = 0.001
> p1 <- ggplot(Hidalgo1872,aes(x=thickness)) +
+   geom_histogram(binwidth=0.001, center = 0.0005) +
+   xlim(0.06,0.135)
> p1
```



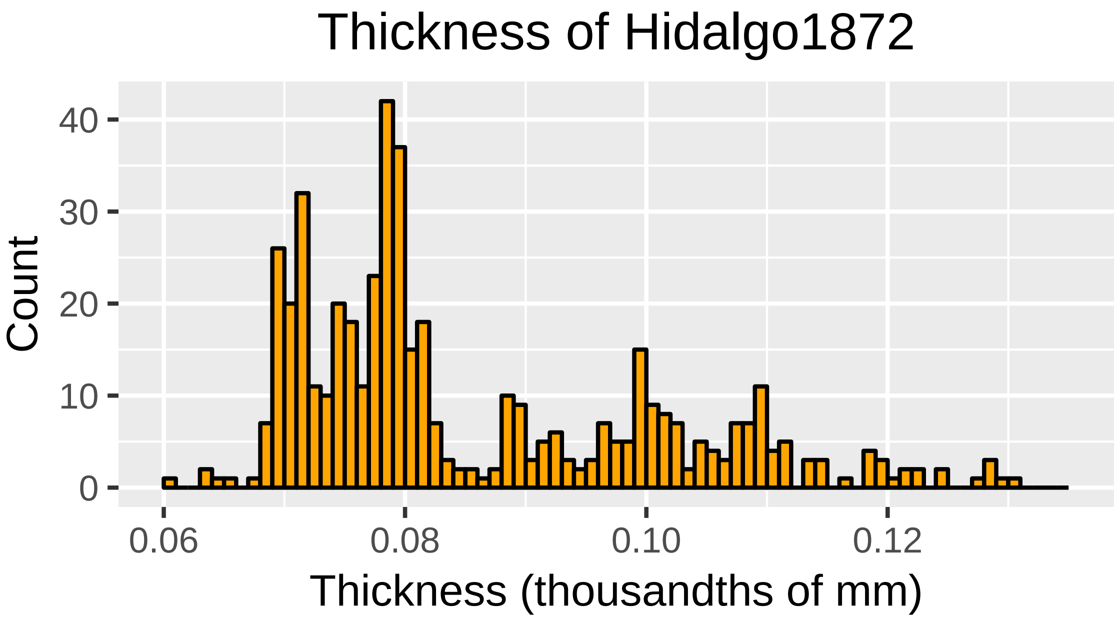
```
> #Add title
> ggplot(Hidalgo1872,aes(x=thickness)) +
+   geom_histogram(binwidth=0.001, center = 0.0005) +
+   xlim(0.06,0.135) +
+   ggtitle("Thickness of Hidalgo1872") +
+   theme(plot.title=element_text(hjust=0.5))
```



```
> #Add better axis titles
> ggplot(Hidalgo1872,aes(x=thickness)) +
+   geom_histogram(binwidth=0.001, center = 0.0005) +
+   xlim(0.06,0.135) +
+   ggtitle("Thickness of Hidalgo1872") +
+   theme(plot.title=element_text(hjust=0.5)) +
+   xlab("Thickness (thousandths of mm)") +
+   ylab("Count")
```

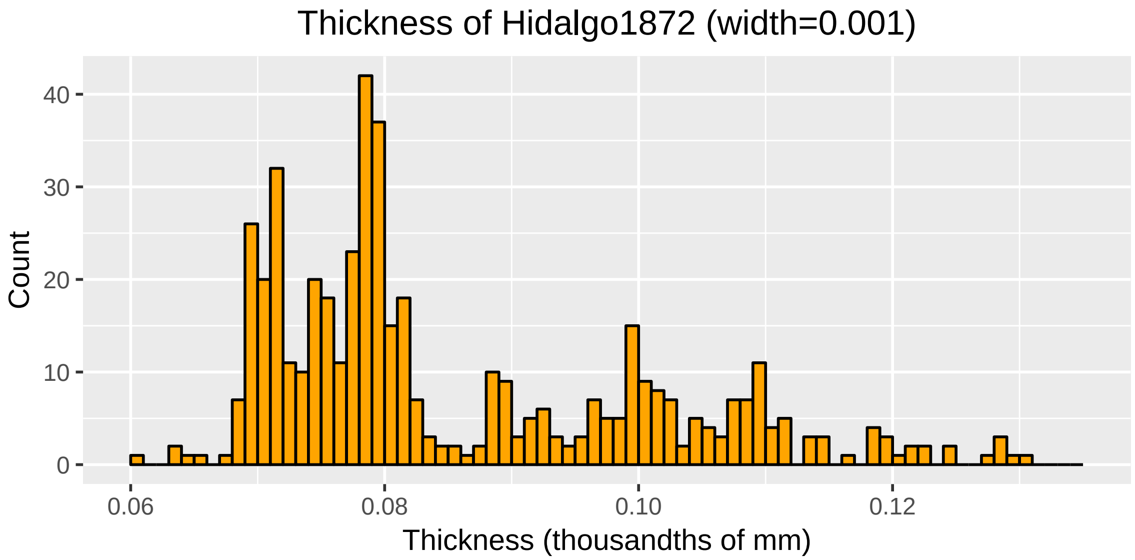


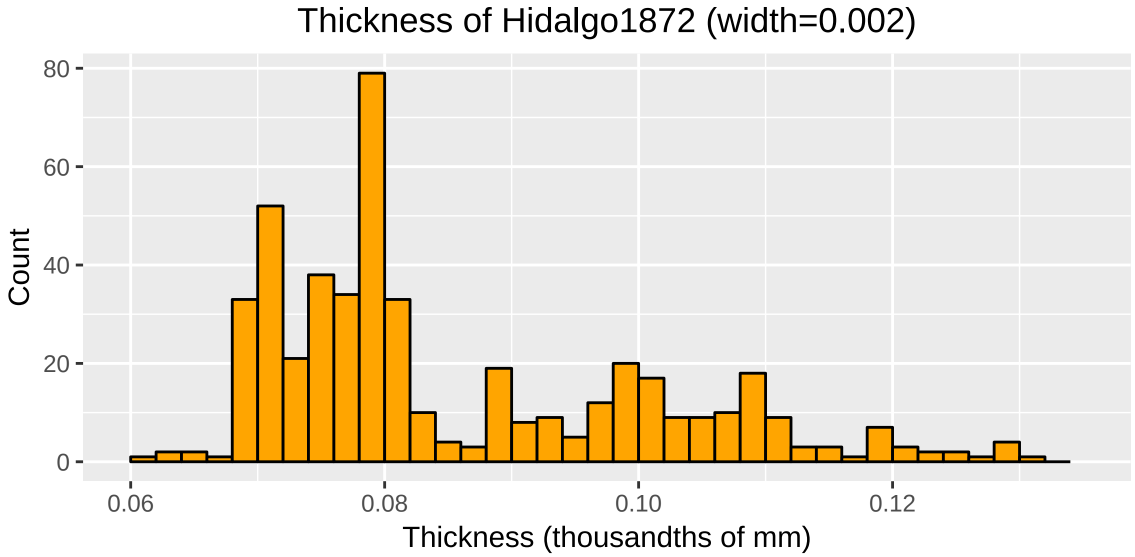
```
> #Add now color the plot
> ggplot(Hidalgo1872,aes(x=thickness)) +
+   geom_histogram(binwidth=0.001, center = 0.0005, color='black',
+                  fill="orange") + xlim(0.06,0.135) +
+   ggtitle("Thickness of Hidalgo1872") +
+   theme(plot.title=element_text(hjust=0.5)) +
+   xlab("Thickness (thousandths of mm)") +
+   ylab("Count")
```

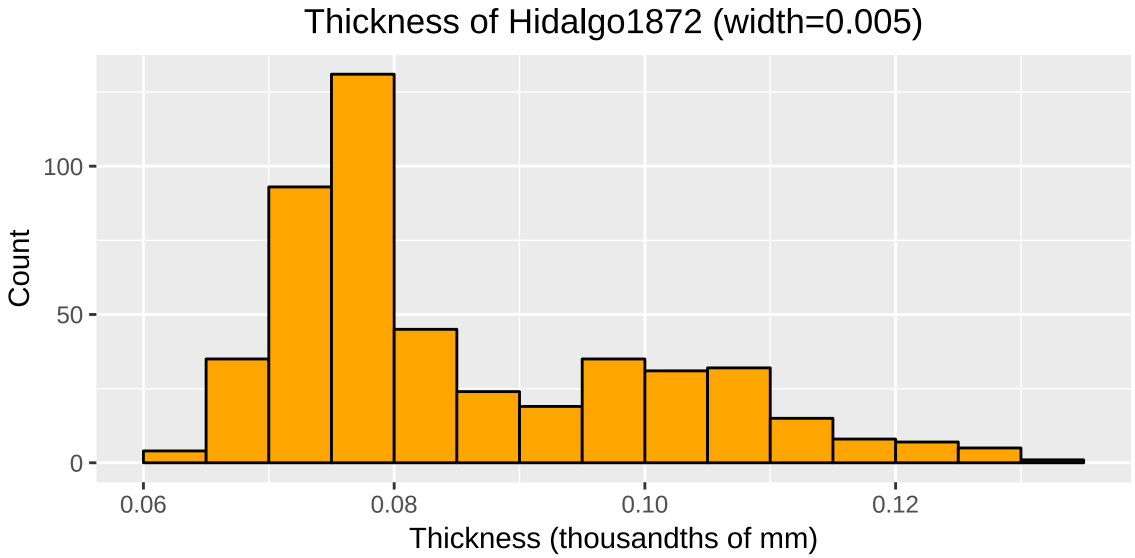


Now, applying this to our final ouputs for all three graphs, we get:

```
> #Binwidth = 0.001
> p1 <- ggplot(Hidalgo1872,aes(x=thickness)) +
+   geom_histogram(binwidth=0.001, center = 0.0005, color='black',
+                  fill="orange") + xlim(0.06,0.135) +
+   ggtitle("Thickness of Hidalgo1872 (width=0.001)") +
+   theme(plot.title=element_text(hjust=0.5)) +
+   xlab("Thickness (thousandths of mm)") + ylab("Count")
> #Binwidth = 0.002
> p2 <- ggplot(Hidalgo1872,aes(x=thickness)) +
+   geom_histogram(binwidth=0.002, center = 0.001, color='black',
+                  fill="orange") + xlim(0.06,0.135) +
+   ggtitle("Thickness of Hidalgo1872 (width=0.002)") +
+   theme(plot.title=element_text(hjust=0.5)) +
+   xlab("Thickness (thousandths of mm)") + ylab("Count")
> #Binwidth = 0.005
> p3 <- ggplot(Hidalgo1872,aes(x=thickness)) +
+   geom_histogram(binwidth=0.005, center = 0.0025, color='black',
+                  fill="orange") + xlim(0.06,0.135) +
+   ggtitle("Thickness of Hidalgo1872 (width=0.005)") +
+   theme(plot.title=element_text(hjust=0.5)) +
+   xlab("Thickness (thousandths of mm)") + ylab("Count")
> #plot
> grid.arrange(p1, p2, p3, nrow = 3)
```

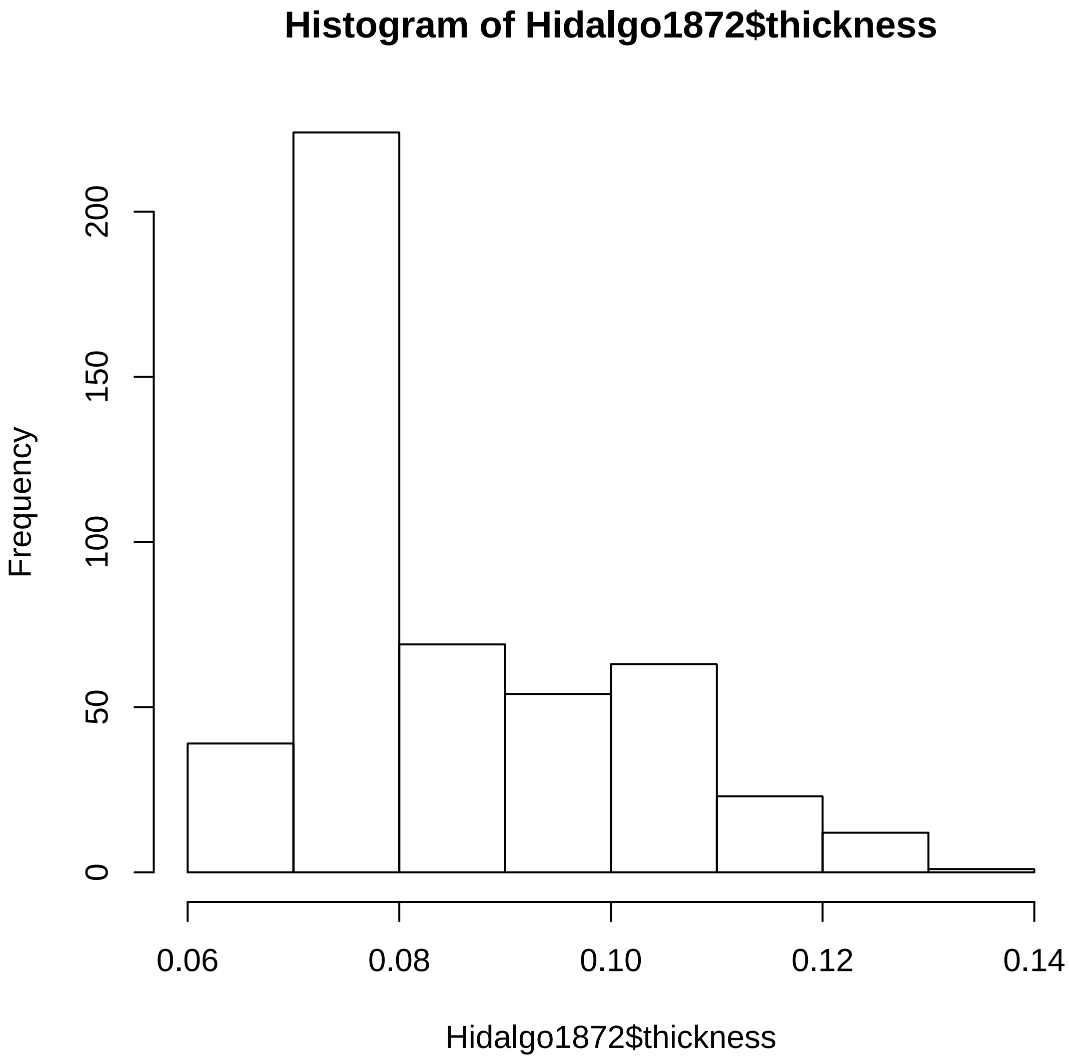

Thickness of Hidalgo1872 (width=0.001)
Count
40
30
20
10
0
0.06
0.08
0.10
0.12
Thickness (thousandths of mm)


Thickness of Hidalgo1872 (width=0.002)
Count
80
60
40
20
0
0.06
0.08
0.10
0.12
Thickness (thousandths of mm)


Thickness of Hidalgo1872 (width=0.005)
Count
100
50
0
0.06
0.08
0.10
0.12
Thickness (thousandths of mm)

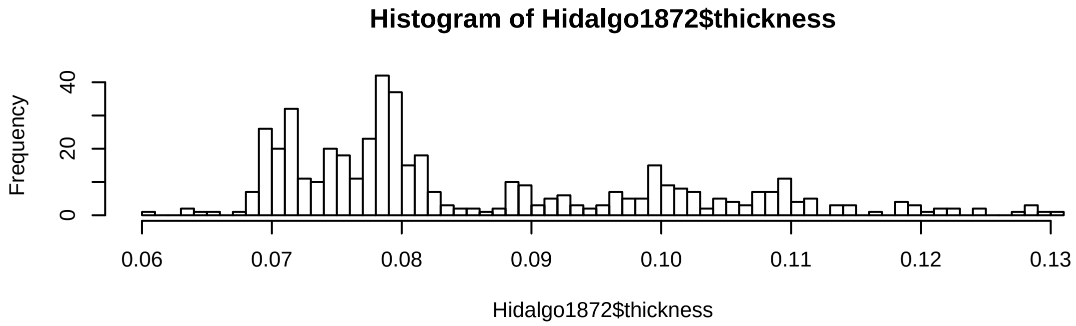
(d) (1 Point) Repeat (b) from above, now using the *hist* function from baseR.

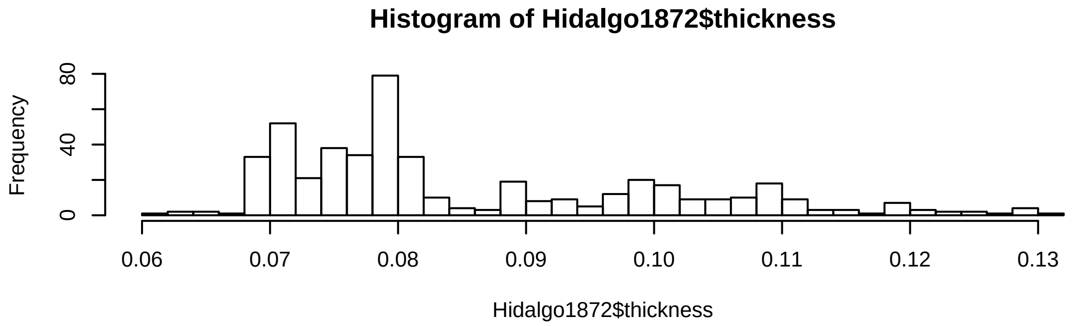
```
> hist(Hidalgo1872$thickness)
```

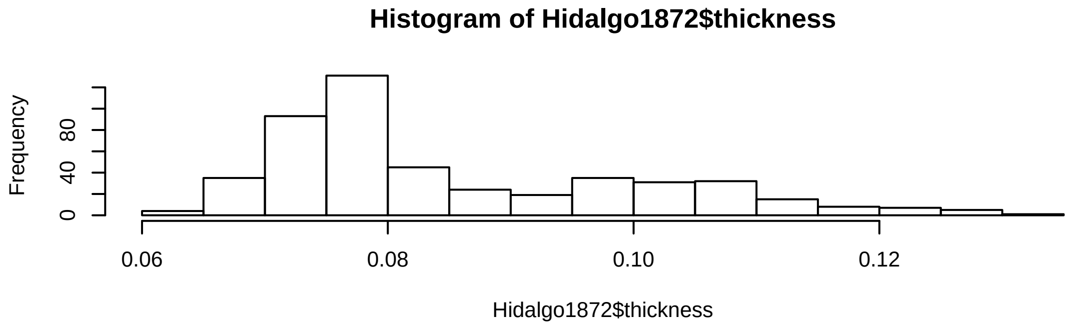


(e) (6 Points) Repeat (c) from above, now using the *hist* function from baseR.

```
> # help from statmethods.net
> # https://www.statmethods.net/advgraphs/layout.html
> attach(mtcars)
> par(mfrow=c(3,1))
> hist(Hidalgo1872$thickness, breaks = (0.14-0.06)/0.001)
> hist(Hidalgo1872$thickness, breaks = (0.14-0.06)/0.002)
> hist(Hidalgo1872$thickness, breaks = (0.14-0.06)/0.005)
```

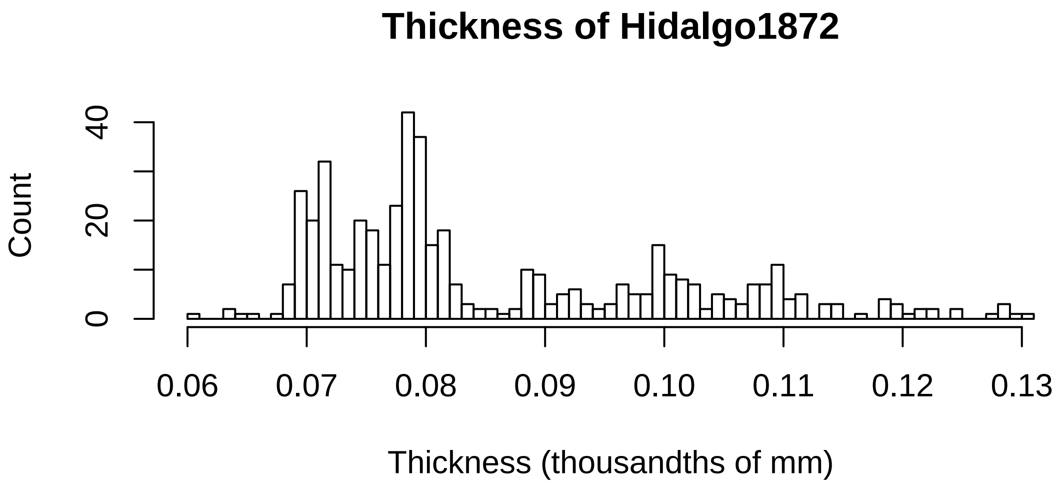


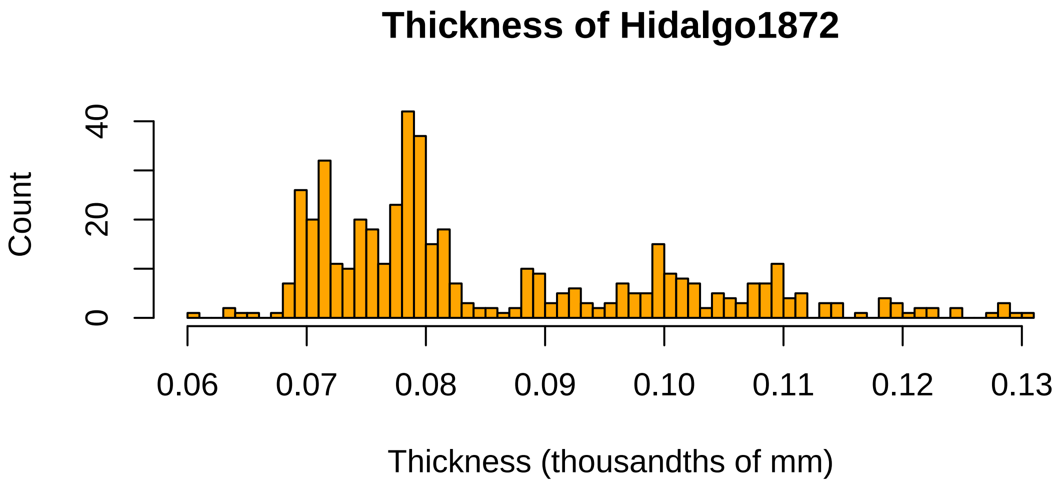




Now to begin improving these plots, lets start with renaming titles, then adding color.

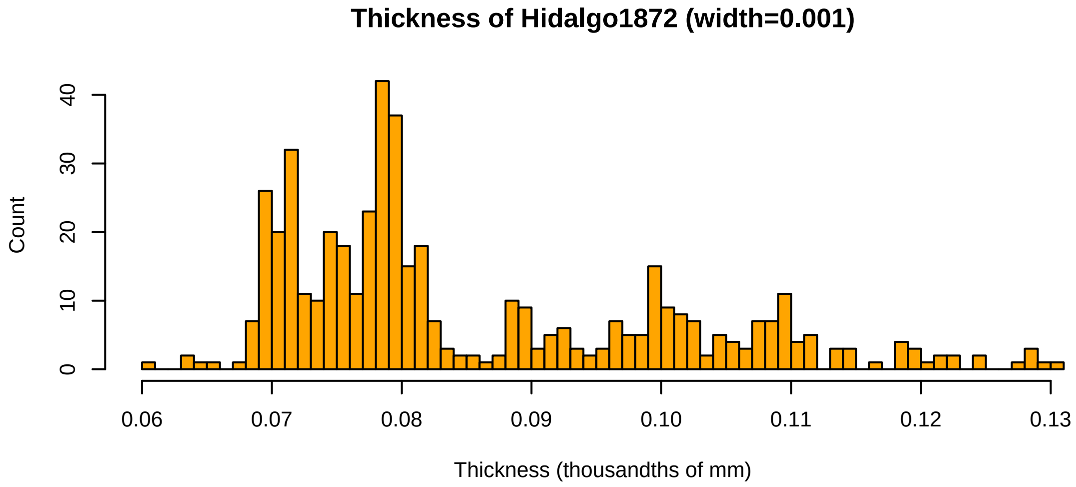
```
> attach(mtcars)
> par(mfrow=c(2,1))
> hist(Hidalgo1872$thickness, breaks = (0.14-0.06)/0.001,
+      main = "Thickness of Hidalgo1872", xlab =
+        "Thickness (thousandths of mm)", ylab ="Count")
> hist(Hidalgo1872$thickness, breaks = (0.14-0.06)/0.001,
+      main = "Thickness of Hidalgo1872", xlab =
+        "Thickness (thousandths of mm)", ylab ="Count",col="orange")
```

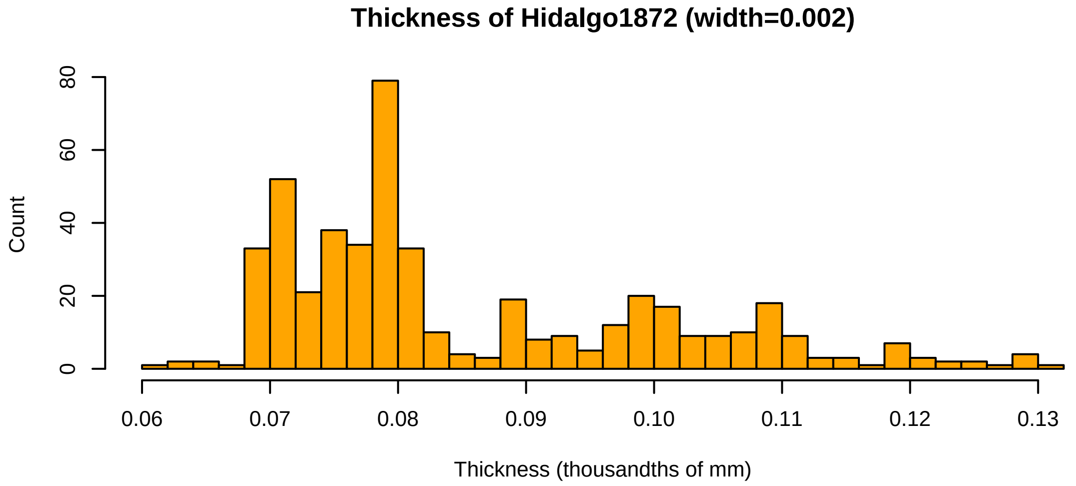


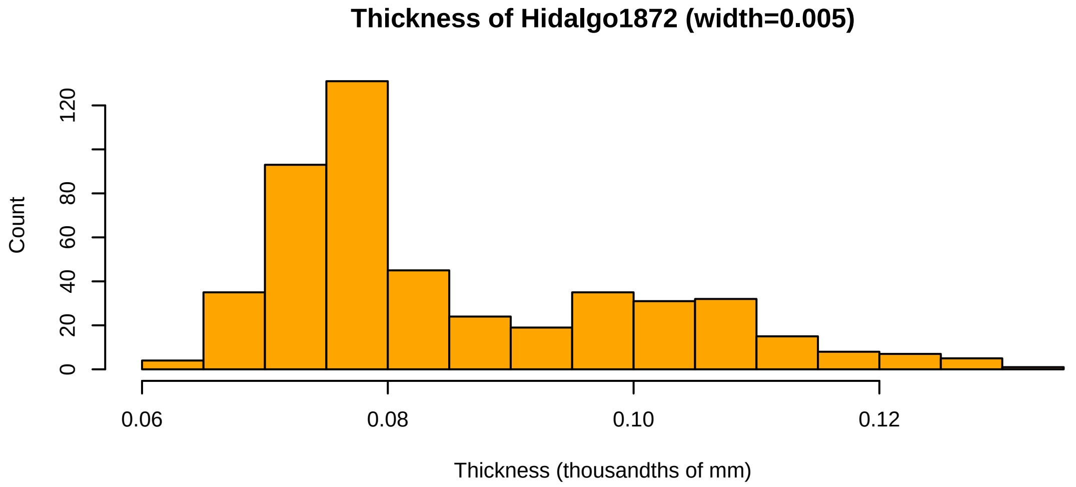


Now for the to apply these to our three graphs:

```
> # help from statmethods.net
> # https://www.statmethods.net/advgraphs/layout.html
> attach(mtcars)
> par(mfrow=c(3,1))
> hist(Hidalgo1872$thickness, breaks = (0.14-0.06)/0.001,
+      main = "Thickness of Hidalgo1872 (width=0.001)", xlab =
+        "Thickness (thousandths of mm)", ylab ="Count",col="orange")
> hist(Hidalgo1872$thickness, breaks = (0.14-0.06)/0.002,
+      main = "Thickness of Hidalgo1872 (width=0.002)", xlab =
+        "Thickness (thousandths of mm)", ylab ="Count",col="orange")
> hist(Hidalgo1872$thickness, breaks = (0.14-0.06)/0.005,
+      main = "Thickness of Hidalgo1872 (width=0.005)", xlab =
+          "Thickness (thousandths of mm)", ylab ="Count",col="orange")
```







(f) (4 Points) Based on your final histograms in (c) and (e), how many modes does the Hidalgo

1872 data set seem to have? Answer this question separately for your three different binwidths. What is your overall conclusion regarding the number of modes?

Answer:

I would say that for bin width of 0.001, there are 6 modes, for 0.002, there are 5, and for 0.005, there are two. Ultimately, I would be inclined to state that there are between three and five modes.

(g) (4 Points) Recall that the stamps originate from the years 1872 and 1873/74. The help page shown in Figure 1 is not very helpful, so we have to make our own assumptions: If the third column (*thicknessB*) contains a value, then this is a measurement from 1873/74. Otherwise, it is a measurement from 1872. Using any approach in R you are familiar with, add a column called *Year* to the Hidalgo1872 data frame. You cannot modify the data outside of R, e.g., via Excel. Show your R code, the first 6 lines of your modified data frame, and a table that summarizes the new *Year* column. Hint: There should be 289 measurements for 1872 and 196 for 1873/74.

```
> Hidalgo1872$Year <- ifelse(!is.na(Hidalgo1872$thicknessB), "1873/74", "1872")
> First6 <- head(Hidalgo1872,6)
> First6
```

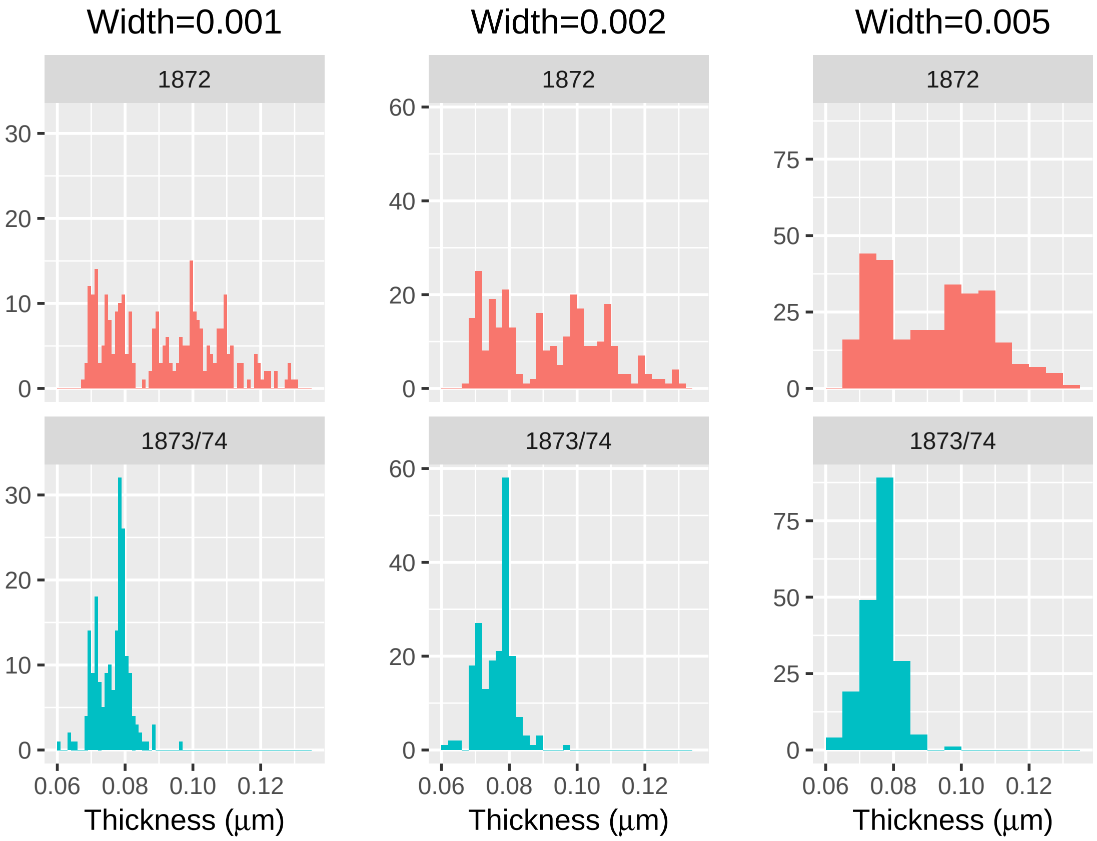
```
  thickness thicknessA thicknessB Year
1     0.068      0.068         NA 1872
2     0.069      0.069         NA 1872
3     0.069      0.069         NA 1872
4     0.069      0.069         NA 1872
5     0.070      0.070         NA 1872
6     0.070      0.070         NA 1872
```

```
> table(Hidalgo1872$Year)
```

```
   1872 1873/74
    289     196
```

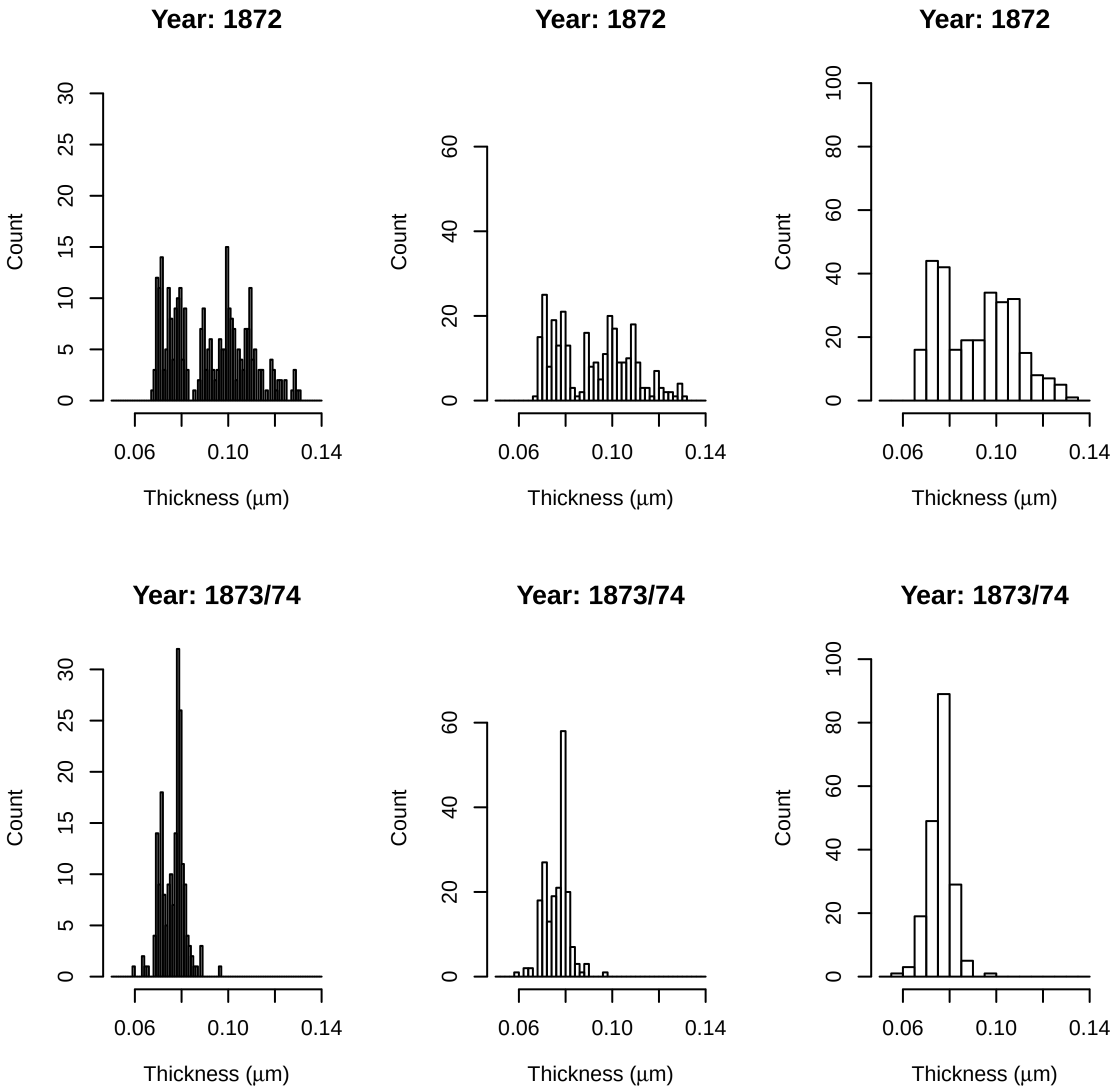
(h) (6 Points) Start with a basic histogram for *thickness*, conditioned on the two options for *Year*, using ggplot2. Optimize this graph in multiple steps. Use a layout that shows the two resulting histograms above each other for better comparison. As before, try three different binwidths: 0.001, 0.002, and 0.005 and use 0.0005, 0.001, and 0.0025 as the center, respectively. Include your R code and the resulting final figure that consists of six histograms overall: 1872 on top and 1873/74 at the bottom and binwidths 0.001, 0.002, and 0.005 from left to right. The two histograms above each other should follow the small multiple principle, i.e., have the same ranges for the horizontal and vertical axes. The histograms besides each other do not have to follow this principle

```
> #Binwidth = 0.001
> p1 <- ggplot(Hidalgo1872,aes(x=thickness, fill=Year)) +
+   geom_histogram(binwidth=0.001, center = 0.0005) + xlim(0.06,0.135) +
+   ggtitle("Width=0.001") +
+   theme(plot.title=element_text(hjust=0.5)) +
+   xlab(expression(paste("Thickness (", mu,"m)"))) + ylab("") + facet_wrap(~Year,
> #Binwidth = 0.002
> p2 <- ggplot(Hidalgo1872,aes(x=thickness, fill=Year)) +
+   geom_histogram(binwidth=0.002, center = 0.001) + xlim(0.06,0.135) +
+   ggtitle("Width=0.002") +
+   theme(plot.title=element_text(hjust=0.5)) +
+   xlab(expression(paste("Thickness (", mu,"m)"))) + ylab("") + facet_wrap(~Year,
> #Binwidth = 0.005
> p3 <- ggplot(Hidalgo1872,aes(x=thickness, fill=Year)) +
+   geom_histogram(binwidth=0.005, center = 0.0025) + xlim(0.06,0.135) +
+   ggtitle("Width=0.005") +
+   theme(plot.title=element_text(hjust=0.5)) +
+   xlab(expression(paste("Thickness (", mu,"m)"))) + ylab("") + facet_wrap(~Year,
> #plot
> grid.arrange(p1, p2, p3, nrow = 1)
```



(i) (6 Points) Repeat (h) from above, now using the *hist* function from baseR. In particular, use the same layout as described above. Hint: Now you have to be really careful to use the proper ranges for the horizontal and vertical axes.

```
> #Arrange histograms into a Grid
> par(mfrow = c(2, 3))
> hist(Hidalgo1872$thickness[Hidalgo1872$Year == "1872"],
+   breaks = seq(.05, .14, by = .001),
+   ylim = c(0, 31),
+   xlab = expression(paste("Thickness (", mu,"m)")),
+   ylab = "Count",
+   main = "Year: 1872")
> hist(Hidalgo1872$thickness[Hidalgo1872$Year == "1872"],
+   breaks = seq(.05, .14, by = .002),
+   ylim = c(0, 75),
+   xlab = expression(paste("Thickness (", mu,"m)")),
+   ylab = "Count",
+   main = "Year: 1872")
> hist(Hidalgo1872$thickness[Hidalgo1872$Year == "1872"],
+   breaks = seq(.05, .14, by = .005),
+   ylim = c(0, 100),
+   xlab = expression(paste("Thickness (", mu,"m)")),
+   ylab = "Count",
+   main = "Year: 1872")
> hist(Hidalgo1872$thickness[Hidalgo1872$Year == "1873/74"],
+   breaks = seq(.05, .14, by = .001),
+   ylim = c(0, 31),
+   xlab = expression(paste("Thickness (", mu,"m)")),
+   ylab = "Count",
+   main = "Year: 1873/74")
> hist(Hidalgo1872$thickness[Hidalgo1872$Year == "1873/74"],
+   breaks = seq(.05, .14, by = .002),
+   ylim = c(0, 75),
+   xlab = expression(paste("Thickness (", mu,"m)")),
+   ylab = "Count",
+   main = "Year: 1873/74")
> hist(Hidalgo1872$thickness[Hidalgo1872$Year == "1873/74"],
+   breaks = seq(.05, .14, by = .005),
+   ylim = c(0, 100),
+   xlab = expression(paste("Thickness (", mu,"m)")),
+   ylab = "Count",
+   main = "Year: 1873/74")
```



(j) (4 Points) Based on your final histograms in (h) and (i), how many modes does the Hidalgo 1872 data set seem to have for each year? Answer this question separately for your two different years (1872 and 1873/74) and your three different binwidths. What is your overall conclusion regarding the number of modes?

Answer:

I would say the following:

For 1872:

At Binwidth = 0.001 –> there are 5 modes

At Binwidth = 0.002 –> there are 4 modes

At Binwidth = 0.005 –> there are 2 modes

For 1873/74:

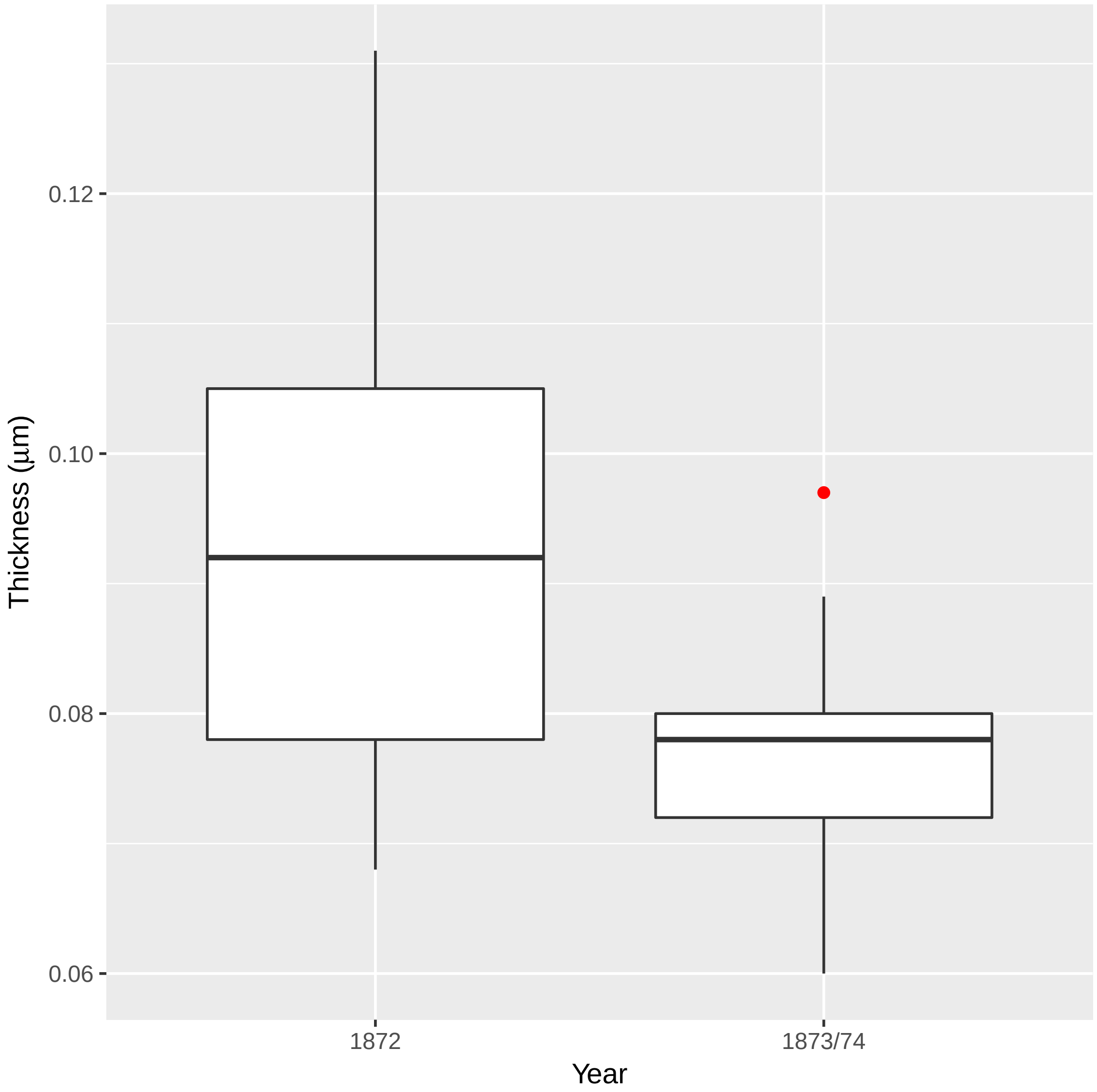
At Binwidth = 0.001 –> there are 2 modes
At Binwidth = 0.002 –> there are 2 modes
At Binwidth = 0.005 –> there is 1 mode
Thus, for my final conclusion, there is a significant visual difference between each plot, thus I will give difference conclusions for each.
For 1872, I would say that it is safe to assume that there are around 3 or 4 modes.
For 1873/74, I would say that it is reasonable to assume that we have 2 modes.

(k) (4 Points) Would boxplots be good replacements for the two histograms for the two years? Compare carefully what can be seen in the box plots and what cannot be seen. First create two basic boxplots with a package of your choice. Then refine them. Add labels as needed and make sure your boxplots follow the small multiple principle. As always, include your R code and the final resulting graphs. Then answer **yes** (they are good replacements) or **no** (they are not good replacements). Justify your answer!

Answer

To find out, lets run a simple box plot first.

```
> ggplot(Hidalgo1872, aes(x=Year, y=thickness)) +
+   geom_boxplot(outlier.color = "red", outlier.shape=16,
+   outlier.size=2, notch=FALSE) +
+   xlab("Year") +
+   ylab(expression(paste("Thickness (", mu,"m)")))
```



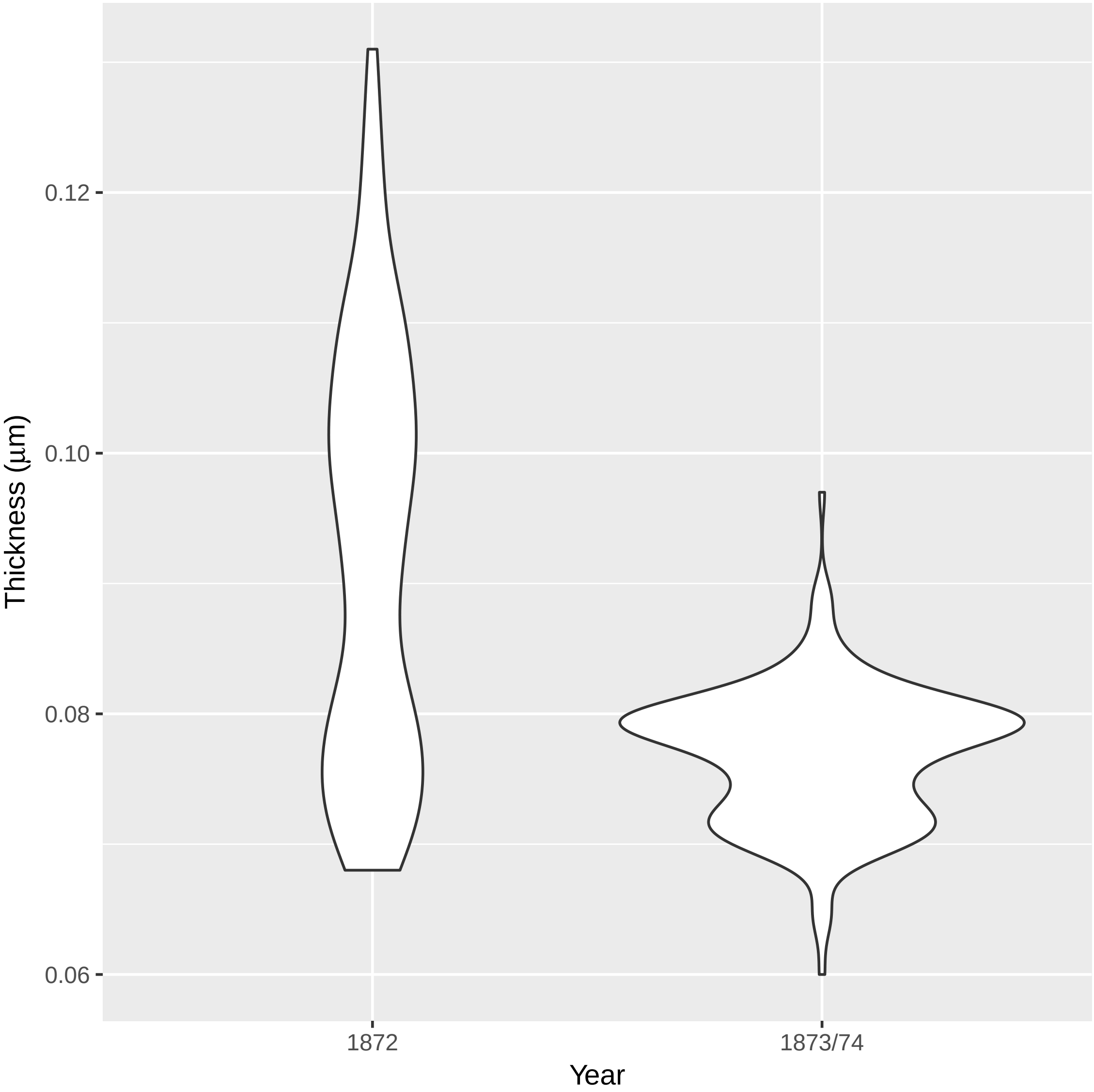
Given how much information we lose in the box plot compaired to the histogram (specifically, we lose most of the shape, multimodality, overall data trend, and have no way to compair the number of observations with a basic boxplot), I think that it is safe to conclude that **no**, *they are not good replacements, although they may be a useful suplement to the histograms.*

(l) (4 Points) Have you ever heard of violin plots? If not, google them! Find a suitable R package that creates violin plots or see how they can be created in ggplot2. Would violin plots be good replacements for the two histograms? First create two basic violin plots with a package of your choice. As always, include your R code and the final resulting graphs. Then answer **yes** (they are good replacements) or **no** (they are not good replacements). Justify your answer!

Answer

To find out, lets run the violin plots first.

```
> ggplot(Hidalgo1872, aes(x=Year, y=thickness)) +
+   geom_violin(mapping = NULL, data = NULL, stat = "ydensity",
+   position = "dodge", draw_quantiles = NULL, trim = TRUE,
+   scale = "area", na.rm = FALSE, show.legend = NA,
+   inherit.aes = TRUE) +
+   xlab("Year") +
+   ylab(expression(paste("Thickness (", mu,"m)")))
```

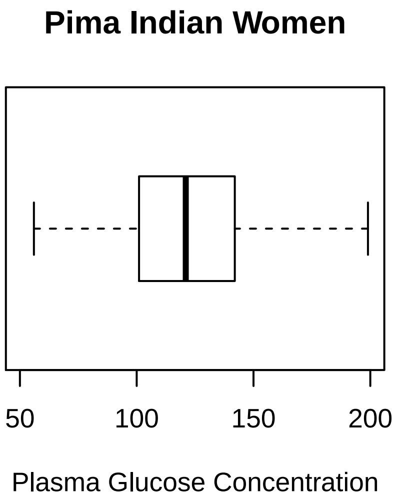


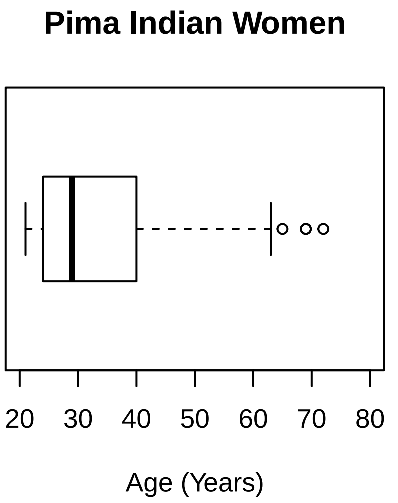
Given that the violin plot shows the shape, spread, multimodality, and the variability, I would say that **yes***, a violin plot would make a good substitute. However, it doesn't seem to note the discrepancy between the number of observations, so that should be kept in mind.*

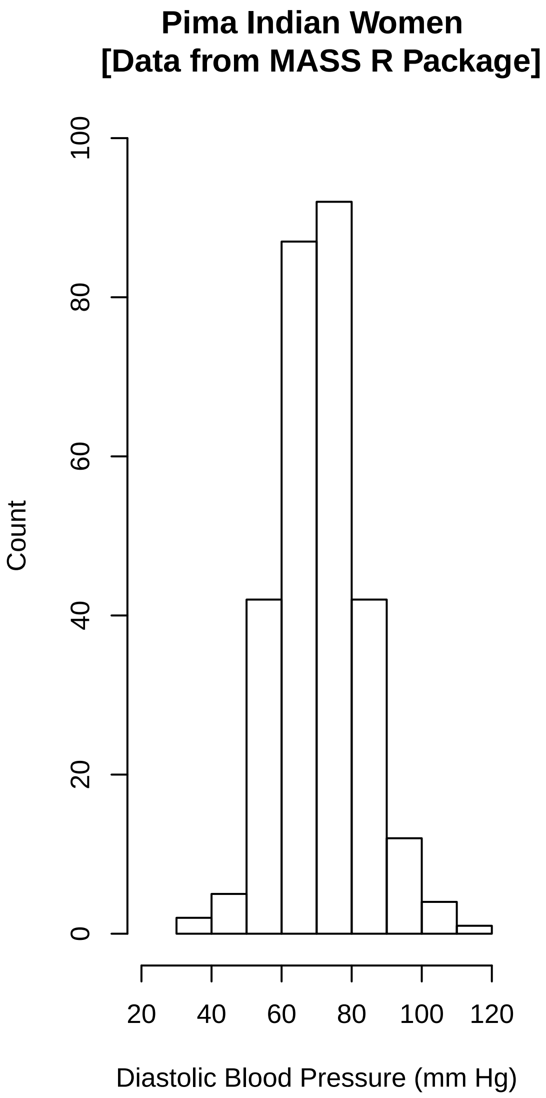
(ii) (12 Points) This question makes use of the *Pima.tr2* data set from the *MASS* R package again. These graphs may not be perfect and may need some further adjustments, but those are not required to get full points in this question.

(a) (6 Points) Recreate the graphs (and layout) shown in Figure 2 using baseR. Include your R code and the resulting graphs. Hint: You can create a new line via `\n` without any extra spaces before/after `\n`.

```
> #load the appropriate data
> data(Pima.tr2, package="MASS")
> layout(matrix(c(1,3,2,3), 2, 2, byrow = TRUE), widths=c(1,1), heights=c(1,1))
> # Boxplot of Glucose
> boxplot(Pima.tr2$glu, ylim=c(50, 200),
+   main="Pima Indian Women",
+   xlab="Plasma Glucose Concentration",
+   ylab=" ", horizontal=TRUE)
> # Boxplot of Age
> boxplot(Pima.tr2$age, ylim=c(20, 80),
+   main="Pima Indian Women",
+   xlab="Age (Years)",
+   ylab=" ",
+   horizontal=TRUE)
> # Histogram of Blood Pressure
> hist(Pima.tr2$bp,
+   main="Pima Indian Women \n [Data from MASS R Package]",
+   xlab="Diastolic Blood Pressure (mm Hg)",
+   ylab="Count",
+   xlim=c(20, 120),
+   ylim=c(0, 100))
```


Pima Indian Women
50
100
150
200
Plasma Glucose Concentration


Pima Indian Women
20
30
40
50
60
70
80
Age (Years)


Pima Indian Women
[Data from MASS R Package]
Count
0
20
40
60
80
100
20
40
60
80
100
120
Diastolic Blood Pressure (mm Hg)

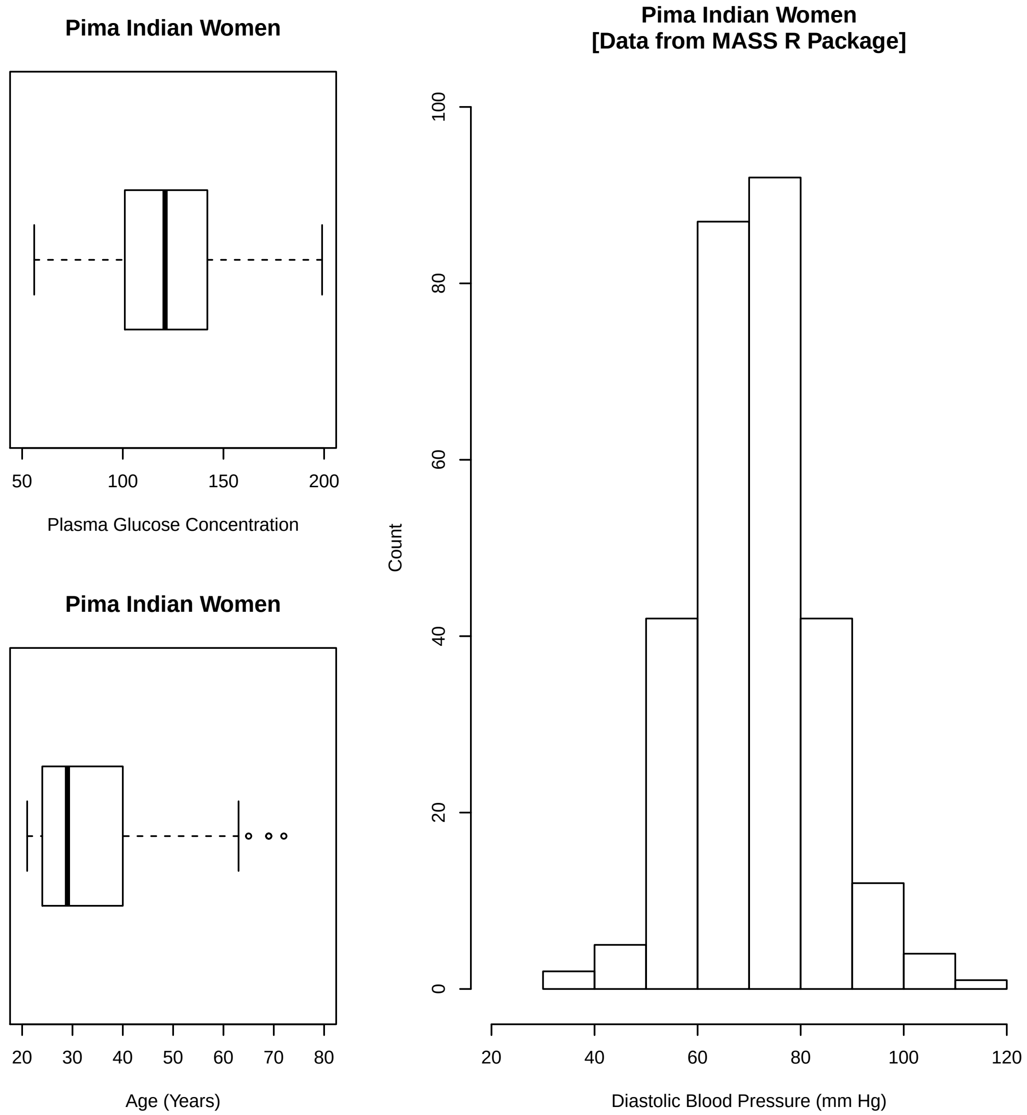
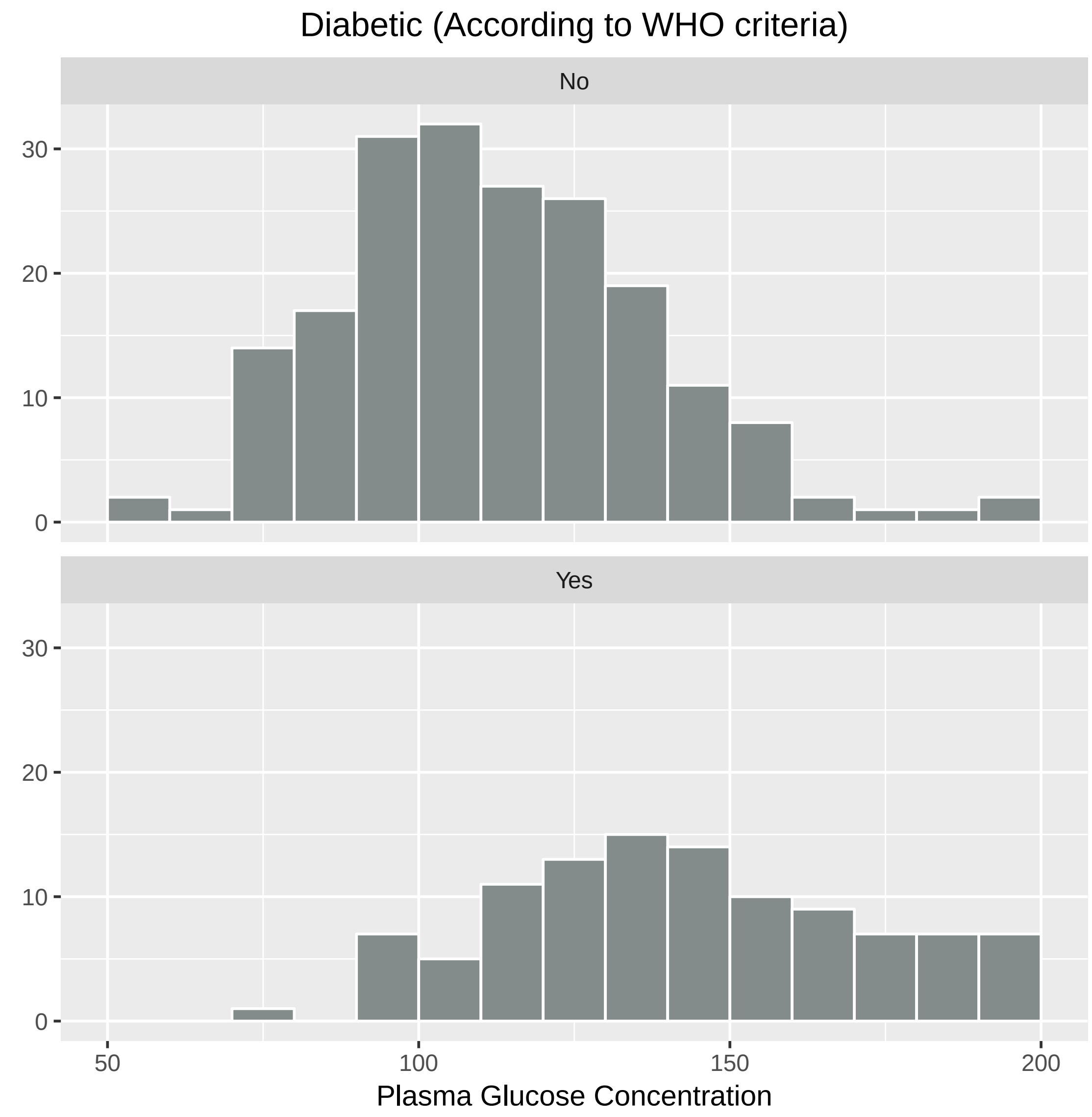


Figure 2: Graphs created with baseR.

(b) (6 Points) Recreate the graph shown in Figure 3 using ggplot2. Include your R code and the resulting graph. Note: You have to change some of the labels and adjust the grid lines so they do not run through the middle of some of your bars. Find suitable help pages or information on stackoverflow.

```
> ggplot(Pima.tr2, aes(x=glu, fill=type)) +
+   xlim(50, 200) +
+   geom_histogram(breaks = seq(50, 200, by=10), color="white", fill="azure4") +
+   xlab("Plasma Glucose Concentration") + ylab(" ") +
+   facet_wrap(˜type, ncol=1) +
+   theme(legend.position="none") +
+   ggtitle("Diabetic (According to WHO criteria)") +
+   theme(plot.title = element_text(hjust = 0.5))
```



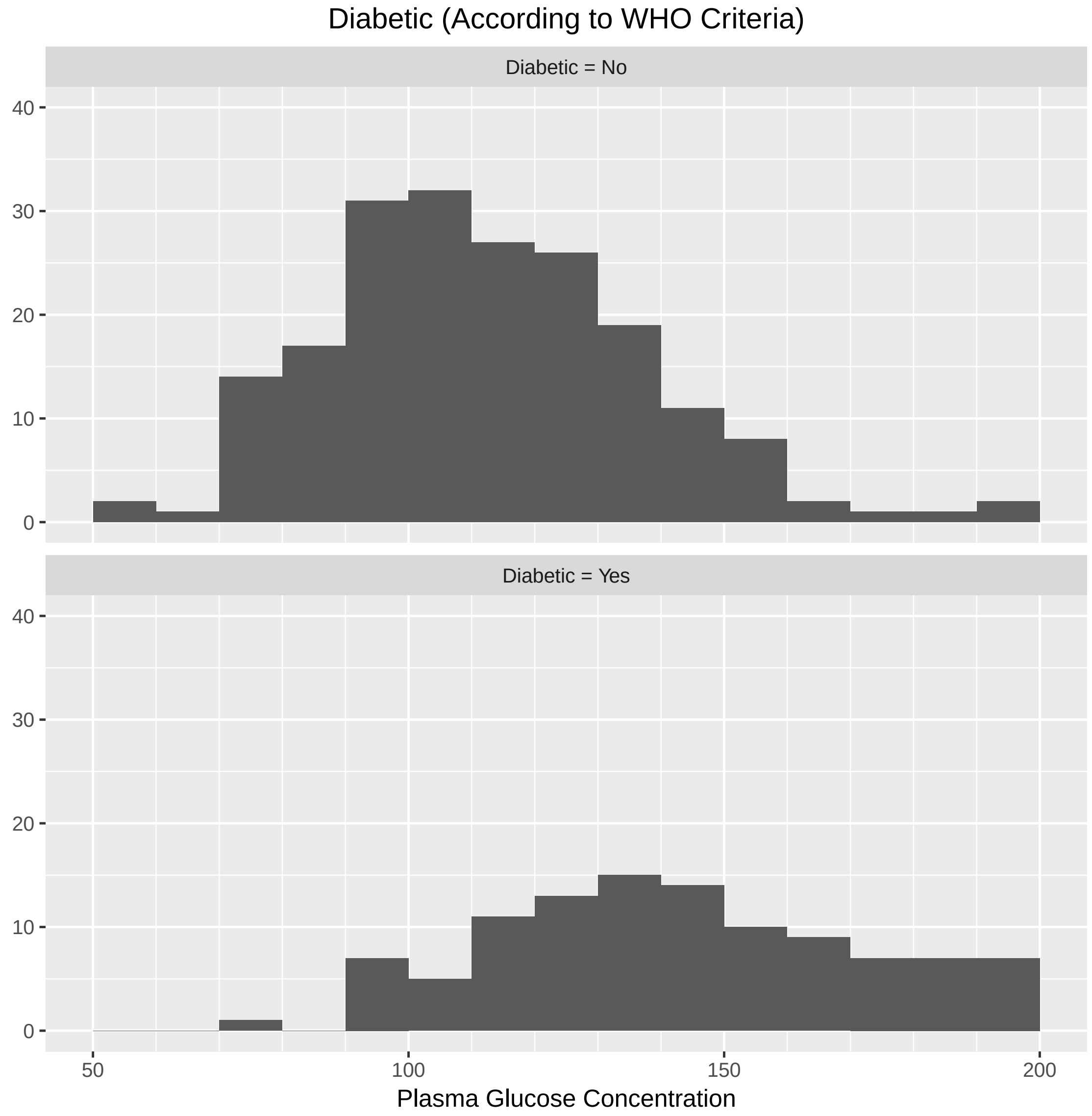


Figure 3: Graph created with *ggplot2*.

General Instructions

(i) Create a single html or pdf document, using R Markdown, Sweave, or knitr. You only have to submit this one document.

(ii) Include a title page that contains your name, your A-number, the number of the assignment, the submission date, and any other relevant information.

(iii) Start your answers to each main question on a new page (continuing with the next part of a question on the same page is fine). Clearly label each question and question part.

(iv) Before you submit your homework, check that you follow all recommendations from Google's R Style Guide (see `https://google.github.io/styleguide/Rguide.xml`). Moreover, make sure that your R code is consistent, i.e., that you use the same type of assignments and the same type of quotes throughout your entire homework.

(v) Give credit to external sources, such as stackoverflow or help pages. Be specific and include the full URL where you found the help (or from which help page you got the information). Consider R code from such sources as "legacy code or third-party code" that does not have to be adjusted to Google's R Style (even though it would be nice, in particular if you only used a brief code segment).

(vi) **Not following the general instructions outlined above will result in point deductions!**

(vii) For general questions related to this homework, please use the corresponding discussion board in Canvas! I will try to reply as quickly as possible. Moreover, if one of you knows an answer, please post it. It is fine to refer to web pages and R commands, but do not provide the exact R command with all required arguments or which of the suggestions from a stackoverflow web page eventually worked for you! This will be the task for each individual student!

(viii) Submit your single html or pdf file via Canvas by the submission deadline. Late submissions will result in point deductions as outlined on the syllabus.