Suyash Mhetre A02317398

Homework Assignment 2 (11/21/2020)

Stat 5550 Statistical Visualization I Fall 2020

Homework Assignment 2 (10/03/2020)

65 Points — Due Friday 11/13/2020 (via Canvas by 11:59pm)

(i) (39 Points) **Olive Oils from Italy:** In this question, you have to work with the *oliveoil* data set from the *pdfCluster* R package. We are only interested in the variables *palmitic* and *macro.area* in this question. Ignore all other variables. See the *oliveoil* help page for further details.

(a) (2 Points) Load all required R packages to answer this question. Show your R code. Do not just blindly trust the information on the help page! How many observations are included in this data set overall? And how many are there in each of the three macro areas?

Use something like the following to incorporate results from your R code directly into your LaTeXtext: "Apparently, there are 50 observations in the cars data set."

Answer:

```
# Place your answer here
library(pdfCluster)

## pdfCluster 1.0-3

data("oliveoil")
str(oliveoil)

## 'data.frame': 572 obs. of  10 variables:
##  $ macro.area : Factor w/ 3 levels "South","Sardinia",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ region     : Factor w/ 9 levels "Apulia.north",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ palmitic   : int  1075 1088 911 966 1051 911 922 1100 1082 1037 ...
##  $ palmitoleic: int  75 73 54 57 67 49 66 61 60 55 ...
##  $ stearic    : int  226 224 246 240 259 268 264 235 239 213 ...
##  $ oleic      : int  7823 7709 8113 7952 7771 7924 7990 7728 7745 7944 ...
##  $ linoleic   : int  672 781 549 619 672 678 618 734 709 633 ...
##  $ linolenic  : int  36 31 31 50 50 51 49 39 46 26 ...
##  $ arachidic  : int  60 61 63 78 80 70 56 64 83 52 ...
##  $ eicosenoic : int  29 29 29 35 46 44 29 35 33 30 ...
```

```
names(oliveoil)
```

```
##  [1] "macro.area"  "region"      "palmitic"    "palmitoleic" "stearic"
##  [6] "oleic"       "linoleic"    "linolenic"   "arachidic"   "eicosenoic"
```

```
dim(oliveoil)
```

```
## [1] 572  10
```

```
nrow(oliveoil)
```

```
## [1] 572
```

```
ncol(oliveoil)
```

```
## [1] 10
```

Comments:

Place your answer here There are 572 observations and 10 columns/variables. There are three levels of macr area.

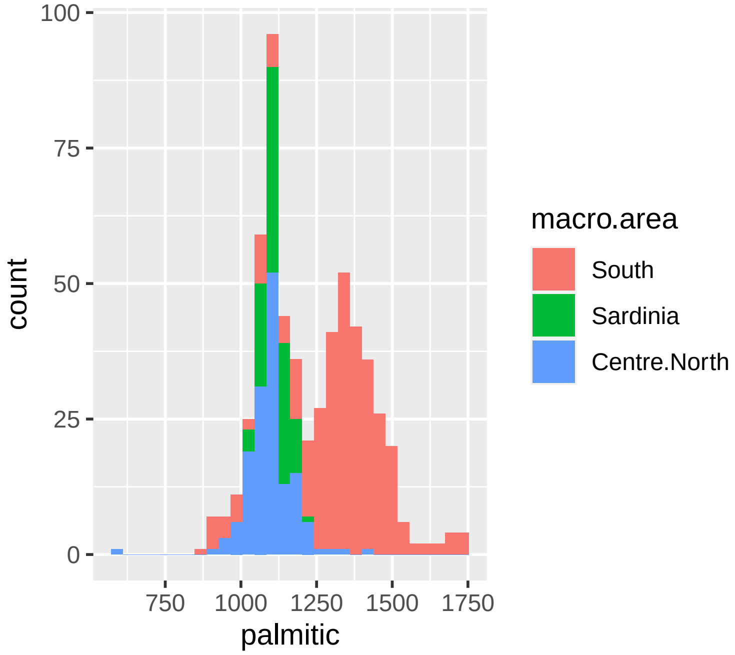
Reference :

- Help page of pdf cluster

(b) (3 Points) Create a default histogram of *palmitic* for all three macro areas combined via *ggplot2*. Do not optimize this histogram. Describe this histogram. Include your figure and your R code.

Answer:

```
# Place your answer here
library(ggplot2)
oliveoildf <- as.data.frame(oliveoil)

ggplot(oliveoildf, aes(x = palmitic, fill = macro.area)) +
        geom_histogram()

## ‘stat_bin()‘ using ‘bins = 30‘.  Pick better value with ‘binwidth‘.
```



Comments:

Place your answer here If we look at the palmitic and macro areas we can say that maximum number of macro areas lie in the 1000-1250 region of palmitic.Also, South macro area has more palmitic spead compared to the other two.

Reference :

- See http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visualization

(c) (4 Points) Create four histogram of *palmitic* for all three macro areas combined via *ggplot2*, using Sturges, sqrt(n), Scott, and FD breaks. How many intervals are there in each of the four histograms? Consider them as small multiples. So enforce the same scale for the horizontal and vertical axes. Describe these histograms. What is similar, what is different?

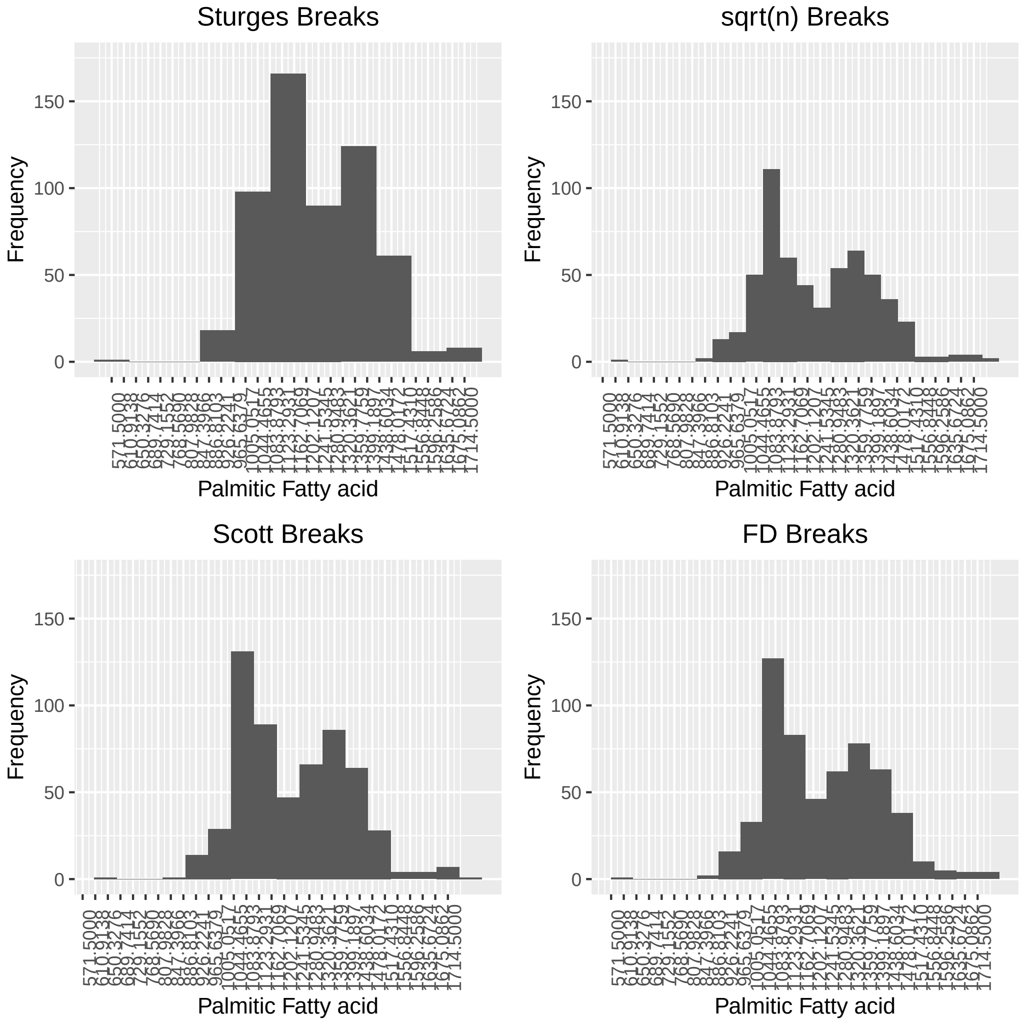
Note: We would like to see how poor the original *ggplot2* defaults are for these four histograms. Therefore, first create a dummy ggplot object for each of the four break types, but do not plot any of these histograms. Then extract the actual intervals used by *ggplot2* from these four dummy objects using the `ggplot_build` function. Finally, use these extracted intervals and create four new *ggplot2* histograms and plot them. In each of your four plots, you have to use the extracted intervals twice: once for the actual breaks and once for the scales. Using `scale_x_continuous` and `scale_y_continuous` may be helpful to enforce he same scale for the horizontal and vertical axes for all four of these histograms.

Include your final graphs (arranged in a single figure) and your R code.

Answer:

```
# Place your answer here


library(gridExtra)

palmitic_length <- oliveoil[, 3]
h <- ggplot(oliveoil, aes(x = palmitic_length)) +
        geom_histogram()
pg <- ggplot_build(h)

## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.

n <- length(palmitic_length)


# Sturges breaks
h1 <- ggplot(oliveoil, aes(x = palmitic_length)) +
        geom_histogram(bins = nclass.Sturges(palmitic_length)) +
        ggtitle("Sturges Breaks") +
        scale_x_continuous(breaks = pg$data[[1]]$xmin) +
        scale_y_continuous(limits = c(0, 175)) +
        theme(plot.title = element_text(hjust = 0.5),
              axis.text.x = element_text(angle = 90)) +
        xlab("Palmitic Fatty acid") +
        ylab("Frequency")

# Place your answer here
```

```r
# sqrt(n) breaks
h2 <- ggplot(oliveoil, aes(x = palmitic_length)) +
        geom_histogram(bins = as.integer(sqrt(n))) +
        ggtitle("sqrt(n) Breaks") +
        theme(plot.title = element_text(hjust = 0.5),
              axis.text.x = element_text(angle = 90)) +
        scale_x_continuous(breaks = pg$data[[1]]$xmin) +
        scale_y_continuous(limits = c(0, 175)) +
        xlab("Palmitic Fatty acid") +
        ylab("Frequency")
# Place your answer here

# Scott breaks
h3 <- ggplot(oliveoil, aes(x = palmitic_length)) +
        geom_histogram(bins = nclass.scott(palmitic_length)) +
        ggtitle("Scott Breaks") +
        theme(plot.title = element_text(hjust = 0.5),
              axis.text.x = element_text(angle = 90)) +
        scale_x_continuous(breaks = pg$data[[1]]$xmin) +
        scale_y_continuous(limits = c(0, 175)) +
        xlab("Palmitic Fatty acid") +
        ylab("Frequency")
# Place your answer here

# FD breaks
h4 <- ggplot(oliveoil, aes(x = palmitic_length)) +
        geom_histogram(bins = nclass.FD(palmitic_length)) +
        ggtitle("FD Breaks") +
        theme(plot.title = element_text(hjust = 0.5),
              axis.text.x = element_text(angle = 90)) +
        scale_x_continuous(breaks = pg$data[[1]]$xmin) +
        scale_y_continuous(limits = c(0, 175)) +
        xlab("Palmitic Fatty acid") +
        ylab("Frequency")

grid.arrange(h1, h2, h3, h4, nrow = 2)
```



<u>Comments:</u>

Place your answer here If we look at the sturges graph it is a smoothened graph compared to the other 3 but at the same time it also misses some data. It also has more binwidth compared to other 3 graphs.The scale on y axis that we have extracted does not make much sense and not useful for user.All the 4 graphs are similar in terms of detecting the outliers at the 571-600 range.

<u>Reference :</u>

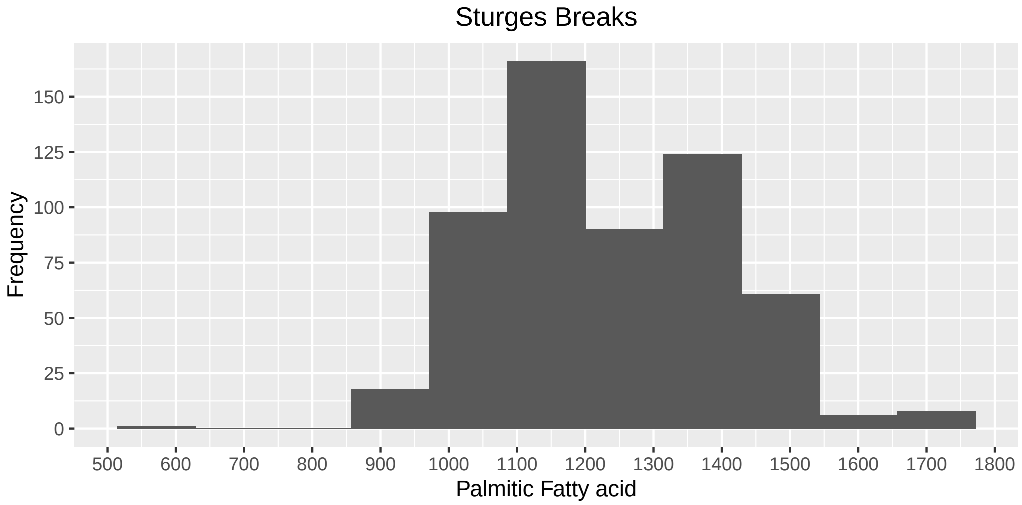
- See `http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visualization`
- Referred the lecture notes as well

(d) (3 Points) Choose one of your four histograms and optimize it for a human reader via *ggplot2*. Make sure to select meaningful starting values for the intervals and meaningful interval widths. 610 (as starting value) and 33.5 (as interval width) are not meaningful for a human reader. You may end up with a slightly different number of intervals than what you started with. Indicate which numbers you selected for the starting value, the number of intervals, and the interval widths. Further optimize this histogram. Include your final graph and your R code.

Answer:

Place your answer here

```r
# Place your answer here
ggplot(oliveoil, aes(x = palmitic_length)) +
        geom_histogram(bins = nclass.Sturges(palmitic_length)) +
        ggtitle("Sturges Breaks") +
        theme(plot.title = element_text(hjust = 0.5)) +
        scale_x_continuous(breaks = seq(500, 2000, 100)) +
        scale_y_continuous(breaks = seq(0, 200, 25)) +
        xlab("Palmitic Fatty acid") +
        ylab("Frequency")
```



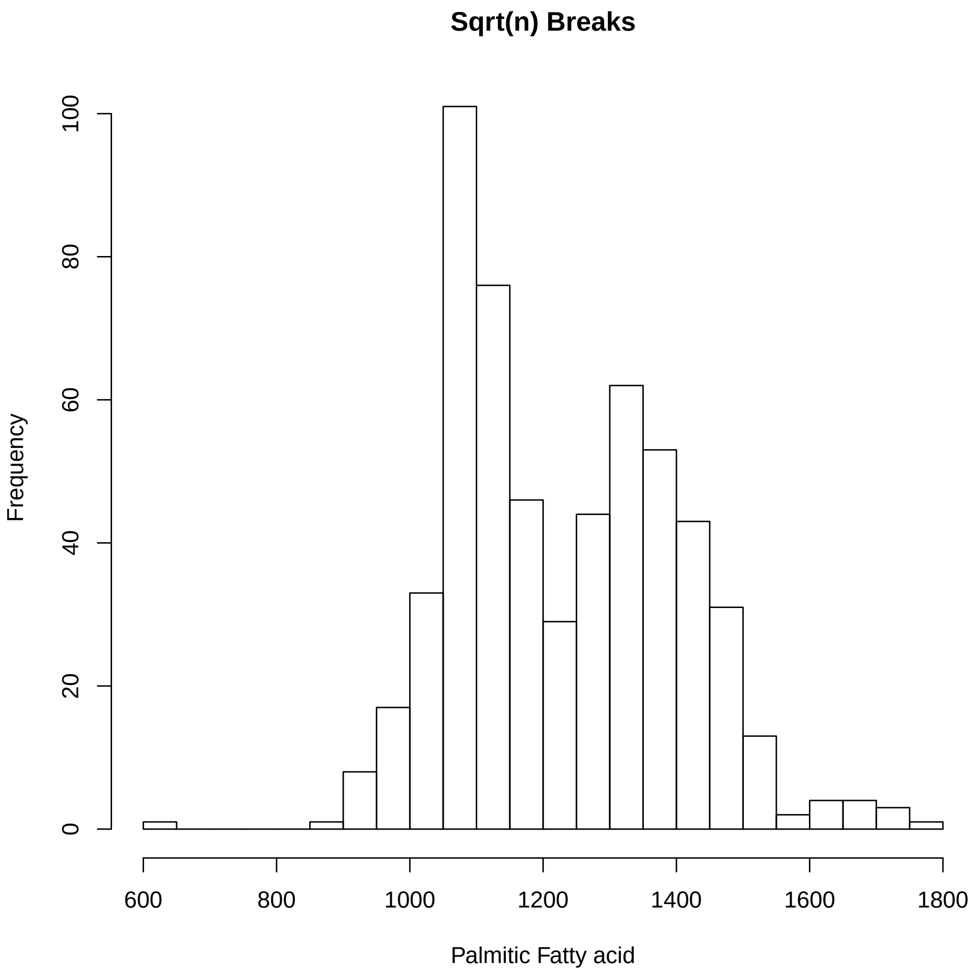
Reference :

- See http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visualization

(e) (3 Points) What does baseR do? Work with the same method you used for your manually optimized histogram in *ggplot2*, i.e., Sturges, sqrt(n), Scott, or FD breaks. How does the resulting baseR histogram of *palmitic* compare to your optimized one from *ggplot2* with respect to starting value, the number of intervals, and interval widths? Include your final figure and your R code.

Answer:

```
# Place your answer here
hist(
        oliveoil$palmitic,
        xlab = "Palmitic Fatty acid",
        main = "Sqrt(n) Breaks",
        breaks = as.integer(sqrt(n))
)
```



Comments:

Place your answer here BaseR removes the scale breaks that are not necessary and is more readable even when we draw the default graph. Also, it does not miss the bins and data on the edges.

Reference :

- See `https://statisticsglobe.com/histogram-in-base-r-hist-function`

(f) (6 Points) Continue with your manually optimized histogram in *ggplot2*. Make sure to switch your histogram to a density scale. Then create six different plots that overlay six different density curves, using the default, nrd, ucv, bcv, SJ-ste, and SJ-dpi bandwidths. Do not further modify the multiplicative bandwidth adjustment (just keep the default for this). Add a jittered rug plot underneath. Jitter once and then use the same jittering for all other rug plots as well. Indicate the amount you use for jittering and why you choose that amount.

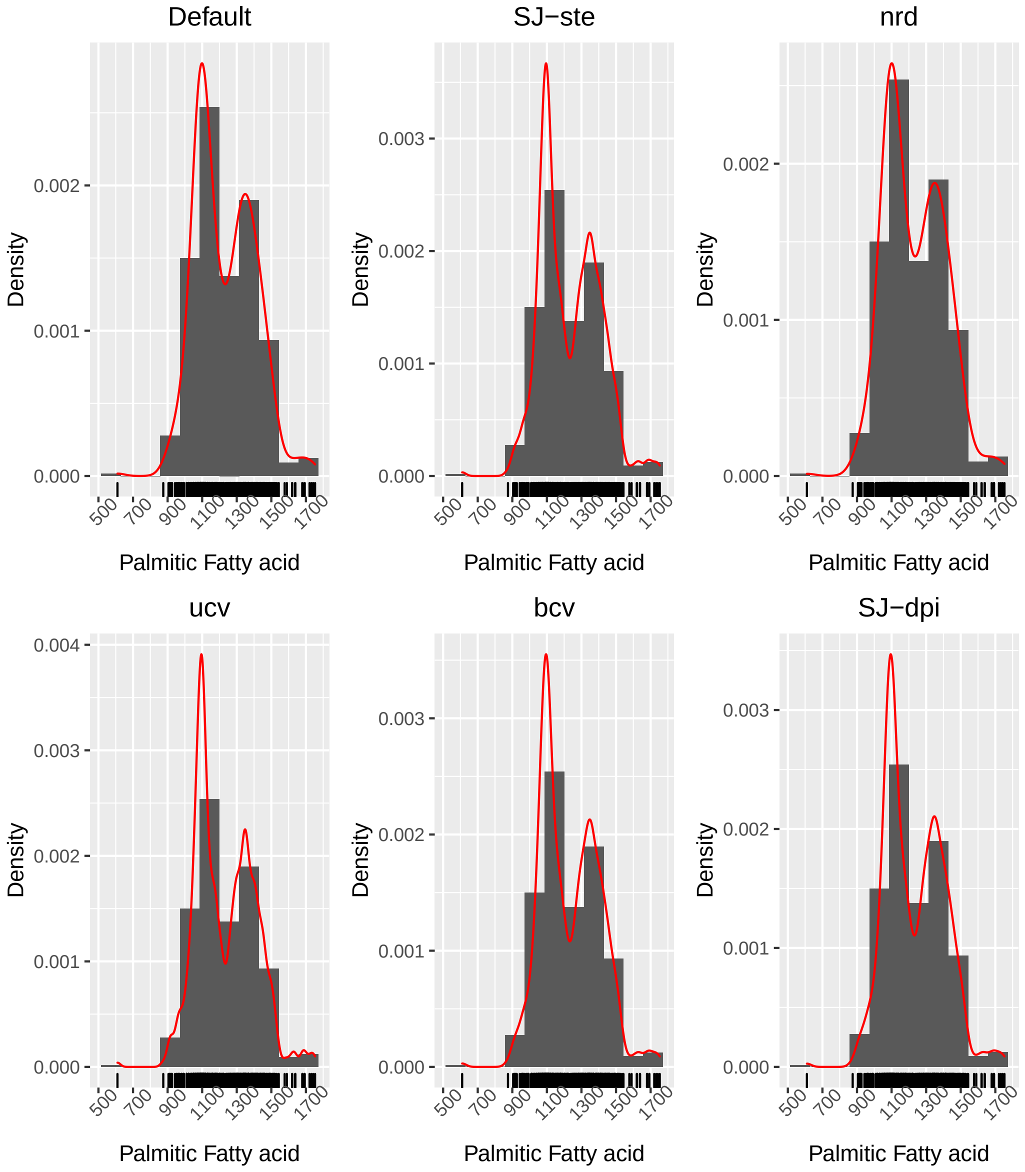
Be careful: Your histograms and density curves must make use of the original data and not of the jittered data.

Include your final graphs (arranged in a single figure) and your R code. Describe which of the six bandwiths seems to be the best option to create a density curve for this variable (use the histogram and rug plot for comparison).

Answer:

```
# Place your answer here
library(ggplot2)
library(gridExtra)

h1 <- ggplot(oliveoil, aes(x = palmitic_length)) +
        geom_histogram(aes(y = ..density..), bins = nclass.Sturges(palmitic_length)) +
        ggtitle("Default") +
        theme(plot.title = element_text(hjust = 0.5),
              axis.text.x = element_text(angle = 45)) +
        scale_x_continuous(breaks = seq(500, 2000, 200)) +
        xlab("Palmitic Fatty acid") +
        ylab("Density") +
        geom_density(col = "red") +
        geom_rug()

h2 <- ggplot(oliveoil, aes(x = palmitic_length)) +
        geom_histogram(aes(y = ..density..), bins = nclass.Sturges(palmitic_length)) +
        ggtitle("SJ-ste") +
        theme(plot.title = element_text(hjust = 0.5),
              axis.text.x = element_text(angle = 45)) +
        scale_x_continuous(breaks = seq(500, 2000, 200)) +
        xlab("Palmitic Fatty acid") +
        ylab("Density") +
        geom_density(bw = "SJ-ste", col = "red") +
        geom_rug()

h3 <- ggplot(oliveoil, aes(x = palmitic_length)) +
        geom_histogram(aes(y = ..density..), bins = nclass.Sturges(palmitic_length)) +
        ggtitle("nrd") +
        theme(plot.title = element_text(hjust = 0.5),
```

```r
              axis.text.x = element_text(angle = 45)) +
        scale_x_continuous(breaks = seq(500, 2000, 200)) +
        xlab("Palmitic Fatty acid") +
        ylab("Density") +
        geom_density(bw = "nrd", col = "red") +
        geom_rug()

h4 <- ggplot(oliveoil, aes(x = palmitic_length)) +
        geom_histogram(aes(y = ..density..), bins = nclass.Sturges(palmitic_length)) +
        ggtitle("ucv") +
        theme(plot.title = element_text(hjust = 0.5),
              axis.text.x = element_text(angle = 45)) +
        scale_x_continuous(breaks = seq(500, 2000, 200)) +
        xlab("Palmitic Fatty acid") +
        ylab("Density") +
        geom_density(bw = "ucv", col = "red") +
        geom_rug()

h5 <- ggplot(oliveoil, aes(x = palmitic_length)) +
        geom_histogram(aes(y = ..density..), bins = nclass.Sturges(palmitic_length)) +
        ggtitle("bcv") +
        theme(plot.title = element_text(hjust = 0.5),
              axis.text.x = element_text(angle = 45)) +
        scale_x_continuous(breaks = seq(500, 2000, 200)) +
        xlab("Palmitic Fatty acid") +
        ylab("Density") +
        geom_density(bw = "bcv", col = "red") +
        geom_rug()

h6 <- ggplot(oliveoil, aes(x = palmitic_length)) +
        geom_histogram(aes(y = ..density..), bins = nclass.Sturges(palmitic_length)) +
        ggtitle("SJ-dpi") +
        theme(plot.title = element_text(hjust = 0.5),
              axis.text.x = element_text(angle = 45)) +
        scale_x_continuous(breaks = seq(500, 2000, 200)) +
        xlab("Palmitic Fatty acid") +
        ylab("Density") +
        geom_density(bw = "SJ-dpi", col = "red") +
        geom_rug()

grid.arrange(h1, h2, h3, h4, h5, h6, ncol = 3)
```



Comments:

Place your answer here According to my obsesrvation the default histogram also looks good to be used for density curve. but the best option would be to go with the nrd as the curve is really soomth compared to others where the density curve gets distorted. Also, I have used the default rug and jitters as it shows the noramlly distributed bins perfectly.

Reference :

- See `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/bandwidth.html`

- Also referred lecture notes.

(g) (6 Points) Now focus on the three different macro areas: Create three graphs side-by-side that show (a) boxplots for the three macro areas; (b) violin plots for the three macro areas; and (c) letter-valued boxplots for the three macro areas.

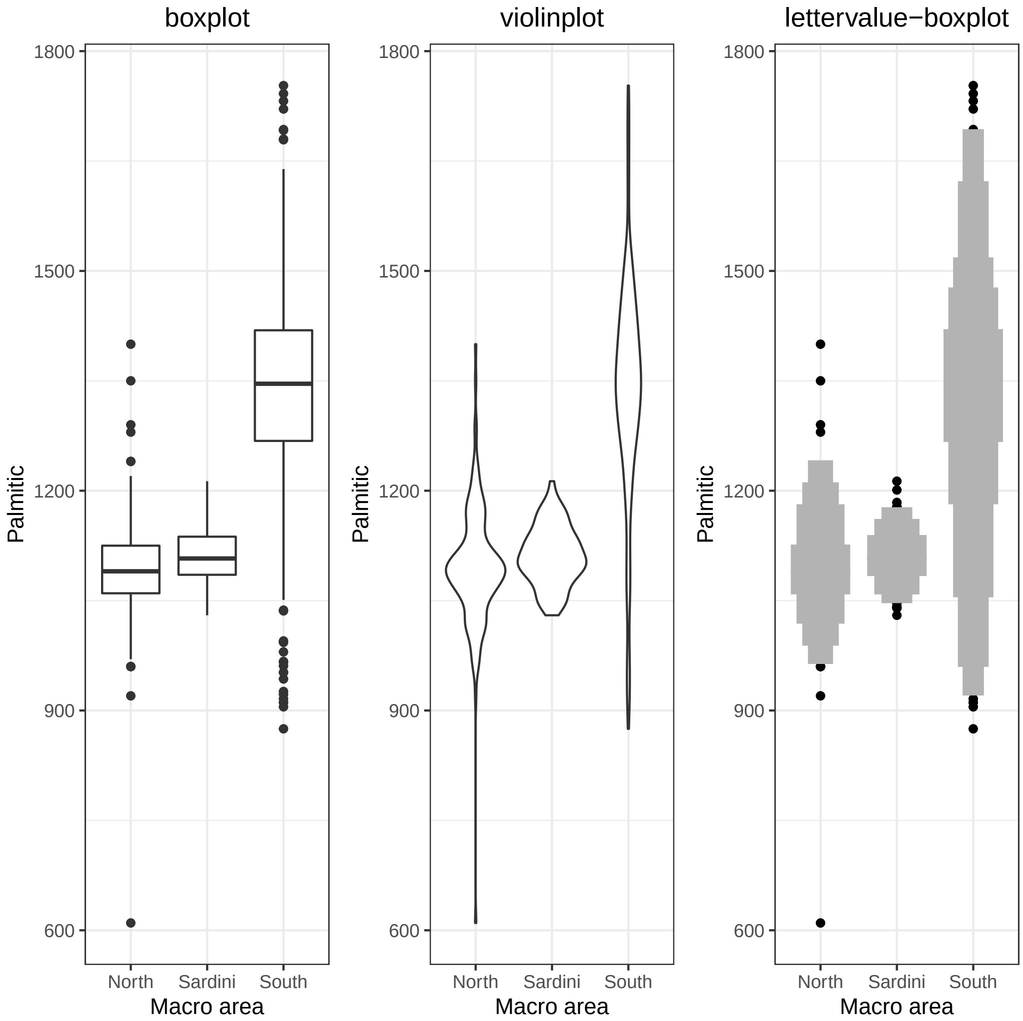
Make sure that the macro areas are ordered as Centre.North (left), Sardinia (center), and South (right) in your three graphs. You may have to resort the factor levels to obtain this ordering when you work with *ggplot2*. Moreover, simplify "Centre.North" to "North" in all of your graphs and legends from now on and also refer to this macro area just as "North" in your text answers.

All individual graphs should extend in vertical (top-bottom) direction. Thus, use the same scale for the vertical axis. Ensure that the graphs follow the small multiples principle. So, when you use a specific color for a macro area, you have to use the same color in all other graphs for that macro area. Include your final graphs (arranged side-by-side) and your R code. What can we learn about similarities and differences of the distributions in the three different macro areas from these graphs?

Answer:

```r
# Place your answer here
library(ggplot2)
library(gridExtra)
library(lvplot)

h1 <-
        ggplot(oliveoil, aes(x = reorder(macro.area, palmitic), y = palmitic)) +
        geom_boxplot() +
        scale_x_discrete(labels = c("North", "Sardini", "South")) +
        theme_bw() +
        ggtitle("boxplot") +
        theme(plot.title = element_text(hjust = 0.5)) +
        xlab("Macro area") +
        ylab("Palmitic")

h2 <-
        ggplot(oliveoil, aes(x = reorder(macro.area, palmitic), y = palmitic)) +
        geom_violin() +
        scale_x_discrete(labels = c("North", "Sardini", "South")) +
        theme_bw() +
        ggtitle("violinplot") +
        theme(plot.title = element_text(hjust = 0.5)) +
        xlab("Macro area") +
        ylab("Palmitic")
```

```
h3 <-
        ggplot(oliveoil, aes(x = reorder(macro.area, palmitic), y = palmitic)) +
        geom_lv() +
        scale_x_discrete(labels = c("North", "Sardini", "South")) +
        theme_bw() +
        ggtitle("lettervalue-boxplot") +
        theme(plot.title = element_text(hjust = 0.5)) +
        xlab("Macro area") +
        ylab("Palmitic")

grid.arrange(h1, h2, h3, ncol = 3)
```



Comments:

Place your answer here To get better understanding of above three graphs we should look at the very first histogram, where we can see 3 marco areas overlapped over each other. Above 3 graphs explain the histogram really well. We can see that the South macro area has the median between 1300-

1500 of palmitic fat. Aslo, if we look at the violin graph North and Sardini maro areas shows higher concentration at the mdeian.

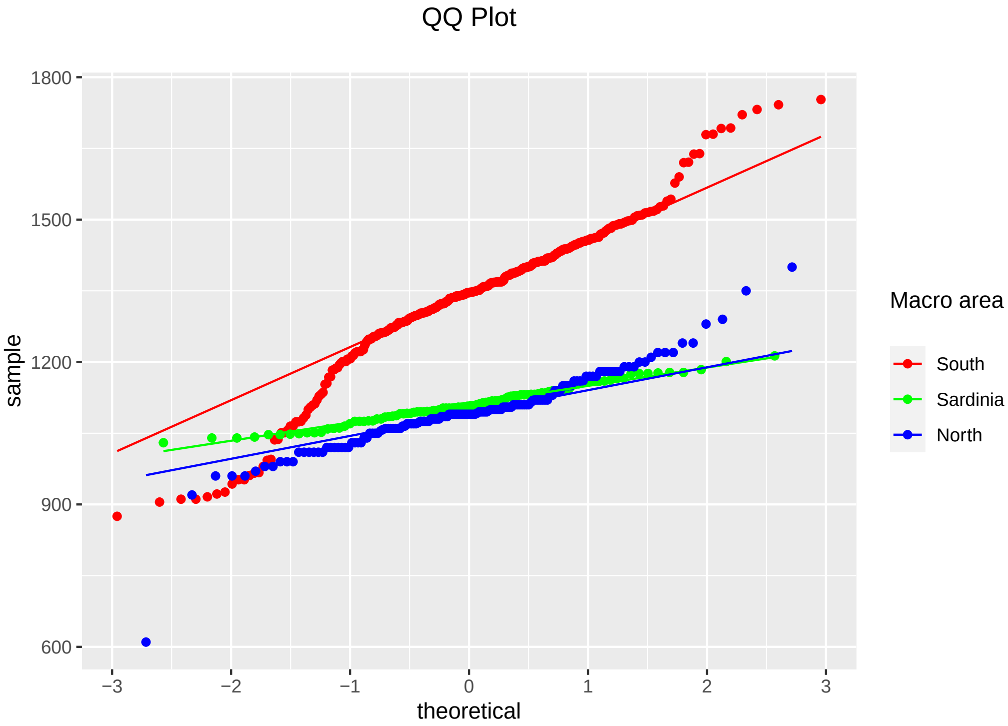
Reference :

- See http://www.sthda.com/english/wiki/ggplot2-box-plot-quick-start-guide-r-software-and-data-visualization
- See http://www.sthda.com/english/wiki/ggplot2-violin-plot-quick-start-guide-r-software-and-data-visualization
- See https://www.rdocumentation.org/packages/lvplot/versions/0.2.0/topics/geom_lv

(h) (3 Points) Are the distributions in the three different macro areas approximately Normal? Construct Q-Q plots and answer this question. Include your final graphs (arranged in a meaningful way) and your R code.

Answer:

```
# Place your answer here
ggplot(oliveoil, aes(sample = palmitic, colour = factor(macro.area))) +
        geom_qq() +
        geom_qq_line() +
        labs(title = "QQ Plot\n", color = "Macro area\n") +
        scale_color_manual(
                labels = c("South", "Sardinia", "North"),
                values = c("red", "green", "blue")
        ) +
        theme(plot.title = element_text(hjust = 0.5))
```



Comments:

Place your answer here If we have to order according to the noramlity then Sardini is most nomral macro area , then second would be the South which can be said as approximately normal. For north macro area we might need a litlle standardization.

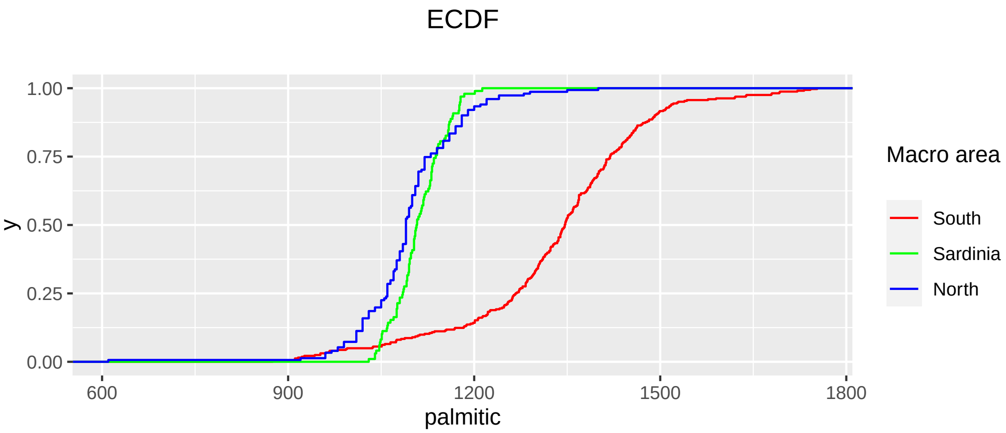
Reference :

- See `https://ggplot2.tidyverse.org/reference/geom_qq.html`

(i) (3 Points) Construct three empirical cdfs (ecdfs) for the three different macro areas and overlay them in the same graph. If you used colors before, use the same colors again for the three macro areas. In any case, include a meaningful legend. How similar (or different) are those ecdfs? Include your final graph and your R code.

Answer:

```
# Place your answer here
ggplot(oliveoil, aes(palmitic, colour = macro.area)) +
        stat_ecdf() +
        labs(title = "ECDF\n", color = "Macro area\n") +
        scale_color_manual(
                labels = c("South", "Sardinia", "North"),
                values = c("red", "green", "blue")
        ) +
        theme(plot.title = element_text(hjust = 0.5))
```



Comments:

Place your answer here North and South macro areas seem to be perfectly similar with respect to the palmitic fat than the South macro area.

Reference :

- See `https://ggplot2.tidyverse.org/reference/stat_ecdf.html`

(j) (6 Points) Summarize your results. What did we learn about the distribution of the palmitic fat overall? And what can we say about the macro areas? Which are similar, which are different (if any). If necessary, repeat some of your previous results and observations here. Your summary should be about 1/2 to 3/4 of a page long.

Comments:

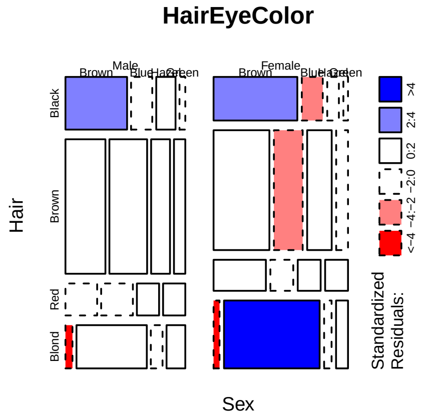
Starting with the first graph, if we look at the overlapping of three macro areas. We can clearly see that most of the oliveoil comes from South followed by North and Sardinia. Also, if we talk about the highest pamiltic count then it ranges from 1000-1250 and the number of macro area for this range of palmitic acid is 50 to 90.50 being the North followed by Sardinia and South on the highest end. Secondly,to prove the above statement we can have a look at the box plots and violin plots. For the South the median is approzimately 1350 and the spread of violin plot concludes that the distribution of the palmitic and South is near this value. From the density curves we can say that the palmitic fatty acid is not normalized as we cannot see perfect bell shaped curve. Hence,w we need some soomethening to do.

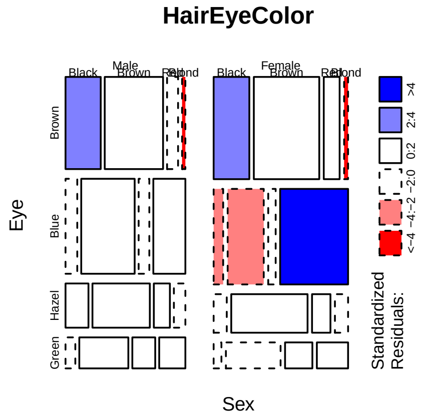
(ii) (16 Points) **Hair and Eye Colors:** In this question, you have to work with the *HairEyeColor* data set (from baseR). It shows the distribution of hair color, eye color, and sex in 592 statistics students. See the *HairEyeColor* help page and any of the cited references for further details.

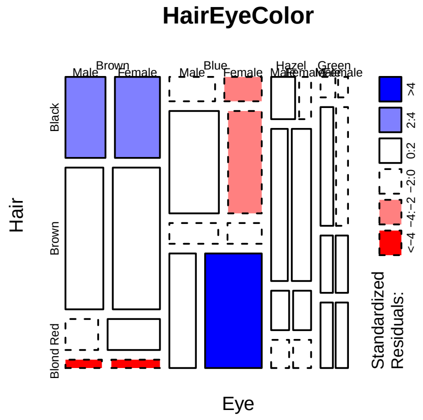
(a) (6 Point) Create six different mosaic plots that show all possible layouts for the three variables, using baseR. Also show the standardized residuals, based on the assumption that all three variables are independent. Include your figures and your R code.

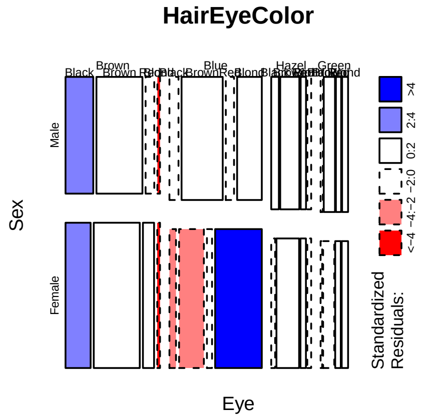
Answer:

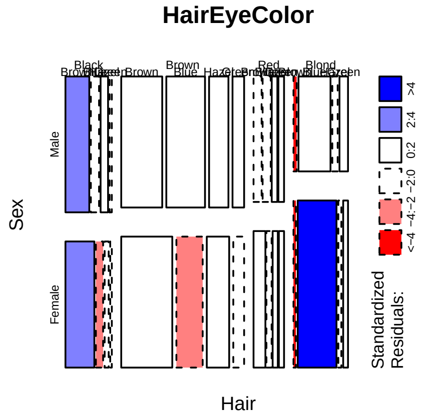
```
# Place your answer here
data("HairEyeColor")
library(vcd)

## Loading required package:  grid

par(mfrow = c(3, 2))
mosaicplot( ~ Sex + Hair + Eye, data = HairEyeColor, shade = TRUE)
mosaicplot( ~ Sex + Eye + Hair, data = HairEyeColor, shade = TRUE)
mosaicplot( ~ Eye + Hair + Sex, data = HairEyeColor, shade = TRUE)
mosaicplot( ~ Eye + Sex + Hair, data = HairEyeColor, shade = TRUE)
mosaicplot( ~ Hair + Sex + Eye, data = HairEyeColor, shade = TRUE)
mosaicplot( ~ Hair + Eye + Sex, data = HairEyeColor, shade = TRUE)
```

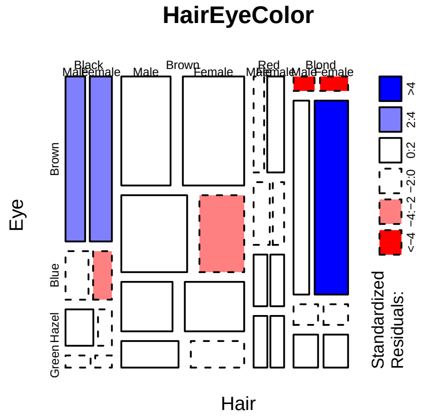












Reference :

- See `https://www.rdocumentation.org/packages/graphics/versions/`

`3.6.2/topics/mosaicplot`

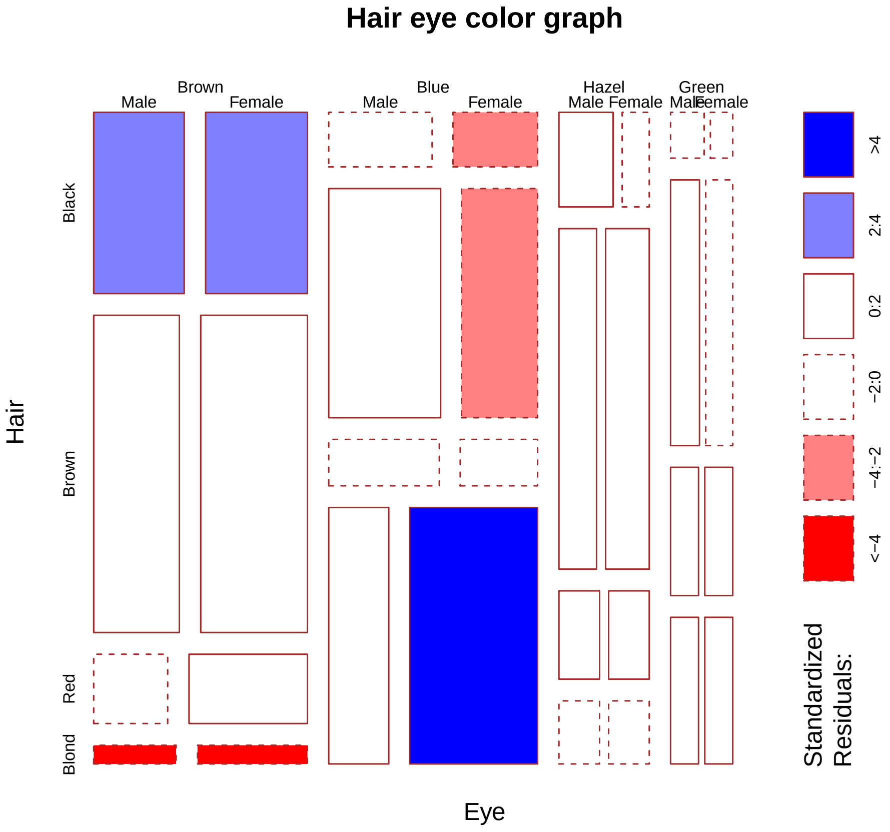
- Also,referred lecture notes

(b) (5 Point) Overall, each of your mosaic plots above should show seven colored areas. These may relate to hair color, eye color, or sex. Six of the seven shaded areas come in pairs (i.e., three pairs of two related areas each) and one is a unique combination of the three variables. Optimize (i.e., add labels, etc.) the mosaic plot that best displays the pairs and the unique combination. Pairs should be located next to each other and not in different regions of the plot. There are two (of the six) mosaic plots that meet this condition and could be optimized. You only have to optimize one. Include your resulting figure and your R code.

Answer:

Place your answer here

```
# Place your answer here
mosaicplot(
        ~ Eye + Hair + Sex,
        data = HairEyeColor,
        main = "Hair eye color graph",
        shade = TRUE,
        xlab = "Eye",
        ylab = "Hair",
        cex.axis = 0.66,
        border = "brown"
)
```



<u>Reference :</u>

- See `https://www.rdocumentation.org/packages/graphics/versions/3.6.2/topics/mosaicplot`
- Also,referred lecture notes

(c) (5 Point) Describe and explain your mosaic plot from (b) above. What can be seen? Assume that a reader is not familiar with mosaic plots, so you have to start with the basic layout you used. How can we best interpret the three pairs and the unique combination? Isn't there an important lurking variable that is missing from this data set, but that would help to even better explain the observed pattern? Which variable is this — and how could it be used to explain the pattern?

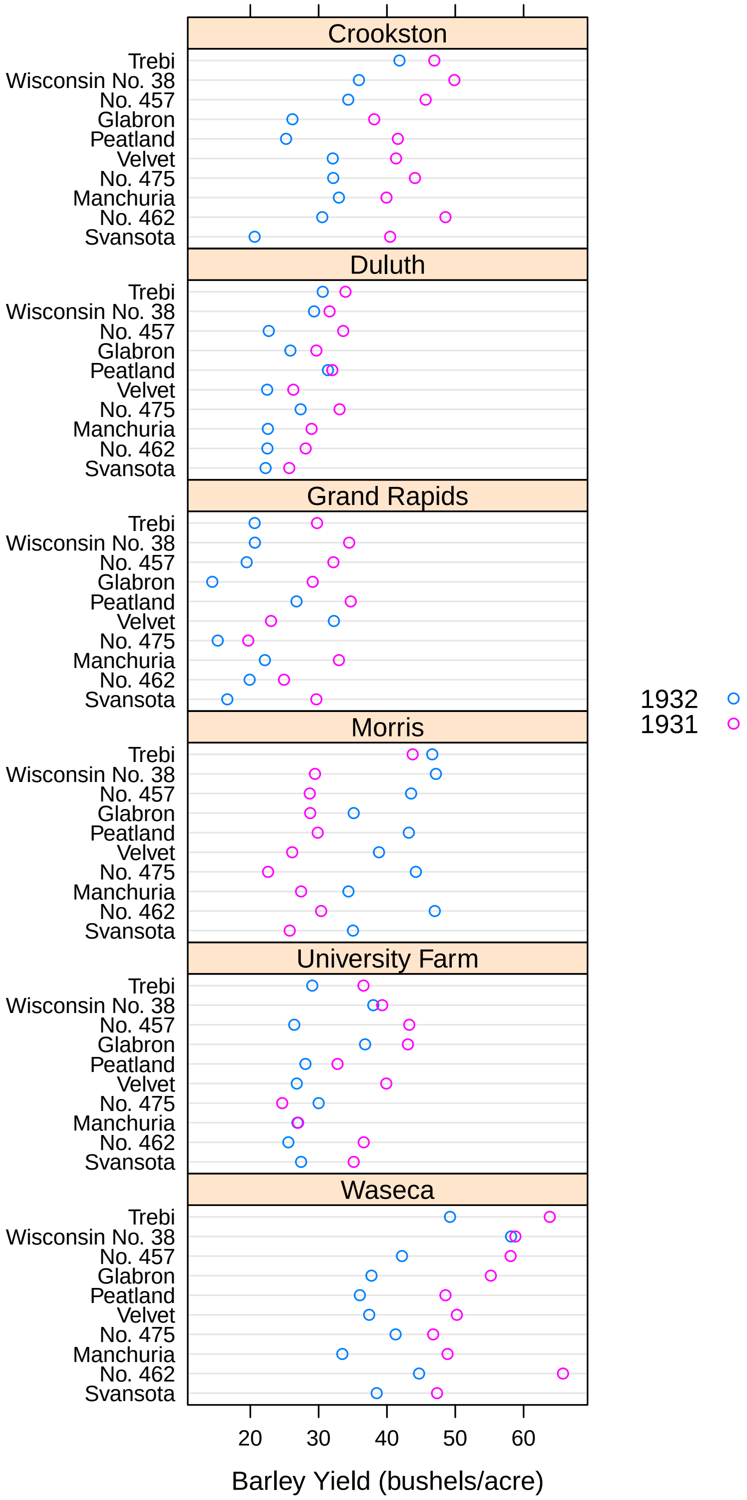
Answer:

A mosaic plot is a graphical display of the cell frequencies of a contingency table in which the area of boxes of the plot are proportional to the cell frequencies of the contingency table. In the above mosaic plot width of the bar indicates the color of eye in male and female while the height of bar indicates the color of hair in male and female. If we look at the brow color eye then we can say that there are slightly more females with brown color eye than males from the width of the bar. While the there are no male with blue color eye but we do have female with blue color eye.The hair color for blonde and black for male and female is the same. The missing variable is blue color eye for male.

(iii) (10 Points) **Barley Data:** Reconstruct and optimize the final version of the *barley* data dot plot from Section 5.7 (Dot Charts for Univariate Data) in our lecture notes, using *ggplot2*. For convenience, this original graph, based on the *lattice* R package, is shown below.

Make sure that you use the same sorting (of the varieties and of the sites) and colors (for the years) as in our version of this plot that was created via the *lattice* dotplot function. Include your final figure and your R code.

```r
library(lattice)

data(barley)

# alphabetical sorting of sites (top to bottom)
dotplot(
        variety ~ yield | site,
        data = barley,
        groups = year,
        key = simpleKey(levels(barley$year), space = "right"),
        xlab = "Barley Yield (bushels/acre)",
        aspect = 0.5,
        layout = c(1, 6),
        ylab = NULL,
        index.cond = list(c(6, 3, 4, 1, 2, 5))
)
```


Crookston
Duluth
Grand Rapids
Morris
University Farm
Waseca
Trebi
Wisconsin No. 38
No. 457
Glabron
Peatland
Velvet
No. 475
Manchuria
No. 462
Svansota
20
30
40
50
60
Barley Yield (bushels/acre)
1932 ○
1931 ○

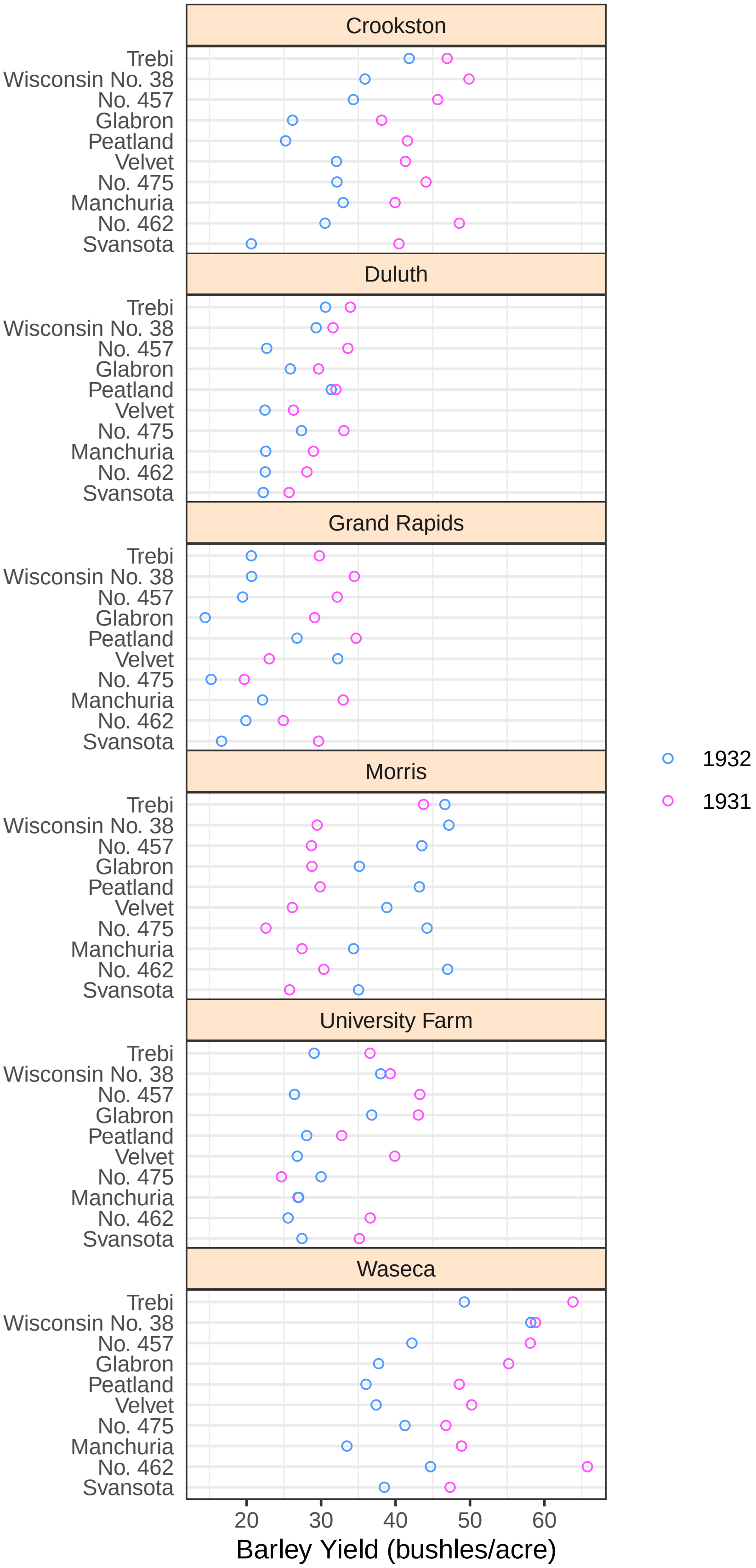
Reference :

- Referred lecture notes

Answer:

Place your answer here

```
# Place your answer here
library(plyr)
library(lattice)
library(ggplot2)
data(barley)

neworder <-
        c("Crookston",
          "Duluth",
          "Grand Rapids",
          "Morris",
          "University Farm",
          "Waseca")
barley <- arrange(transform(barley,
                            site = factor(site, levels = neworder)), site)

ggplot(barley) +
        geom_point(aes(x = yield, y = variety, color = year), shape = 1) +
        scale_color_manual(values = c("1932" = "#4d9cff", "1931" = "#ff53ff")) +
        facet_wrap( ~ site, dir = "v", ncol = 1) +
        theme_bw() +
        xlab("Barley Yield (bushles/acre)") +
        ylab("") +
        theme(
                panel.grid.major.x = element_blank(),
                strip.background = element_rect(fill = "#ffe5cc"),
                panel.margin.y = unit(-0.1, "lines"),
                axis.ticks.y = element_blank(),
                legend.title = element_blank(),
                aspect.ratio = 0.5
        )
```


Crookston
Duluth
Grand Rapids
Morris
University Farm
Waseca
Trebi
Wisconsin No. 38
No. 457
Glabron
Peatland
Velvet
No. 475
Manchuria
No. 462
Svansota
20
30
40
50
60
Barley Yield (bushles/acre)
○ 1932
○ 1931

Reference :

- See `https://ggplot2.tidyverse.org/reference/facet_wrap.html`
- Also,referred lecture notes

General Instructions

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

(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. Your answer to question (i) should start on page 2!

(iv) Show your R code and resulting graph(s) for each question part!

(v) Before you submit your homework, check that you follow all recommendations from Google's R Style Guide (see `http://web.stanford.edu/class/cs109l/unrestricted/resources/google-style.html`). 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.

(vi) 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).

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

(viii) 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!

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