[redacted]

[redacted]

by

Scott Greenberg

[redacted]

Date: [redacted]

[redacted]

[redacted]

[redacted]

[redacted] [redacted]

[redacted]

[redacted]

[redacted]

[redacted]

In this question, you have to work with the [redacted] data set from the *MASS* package. See the [redacted] help page for further details. In case you are interested in further details of the original study, read the article by Campbell and Mahon (1974), published in the Australian Journal of Zoology (`https://doi.org/10.1071/ZO9740417`). For convenience, the pdf is made available in Canvas.

(a) (3 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 and how many variables are included in this data set overall?

Use something like the following to incorporate results from your R code directly into your LaTeX text: "Apparently, there are 50 observations and 2 variables in the cars data set."

Answer:

```
> library(MASS)
> library(ggplot2)
> library(GGally)
> library(RColorBrewer)
> library(BiocManager)
> library(graph)
> library(PairViz)
> library(gplots)
> library(autoimage)
```

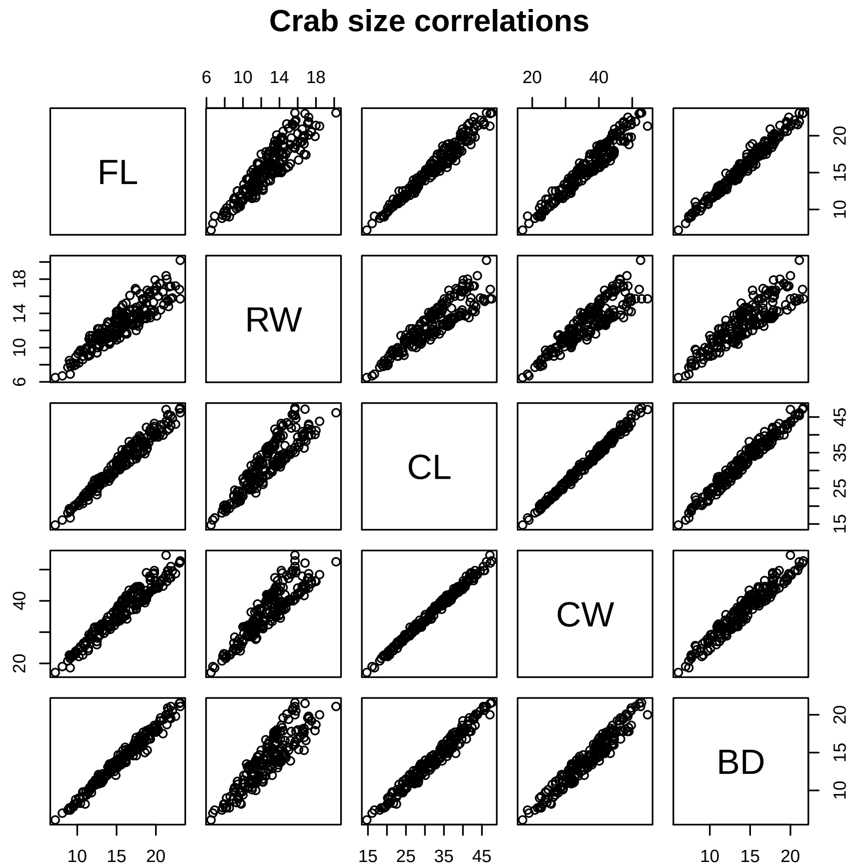
How many observations and how many variables are included in this data set overall?

“Apparently, there are 200 observations and 8 variables in the cars data set.”

(b) (4 Points) Create a default scatterplot matrix of the five quantitative variables (omitting *sp, sex & index*) via *baseR*. Do not optimize this scatterplot matrix and do not use any colors or symbols. Describe this scatterplot matrix. Which questions naturally come to your mind? What would you initially anticipate as the answers to your questions? Do not modify your answer here, even if it turns out later on that your anticipation was not correct! Hint: Think of which variables have been ignored in your scatterplot matrix. Include your figure and your R code.

Answer:

```
> crabs.quant <- crabs[, -1:-3]
> pairs(crabs.quant, main = "Crab size correlations")
```



Comment:

Describe this scatterplot matrix.

We look at a scatterplot for pairs of variables, the plots below the diagonal have an equivalent above the diagonal just with switched axis

Which questions naturally come to your mind?

Why are some of the plots heteroskedastic?

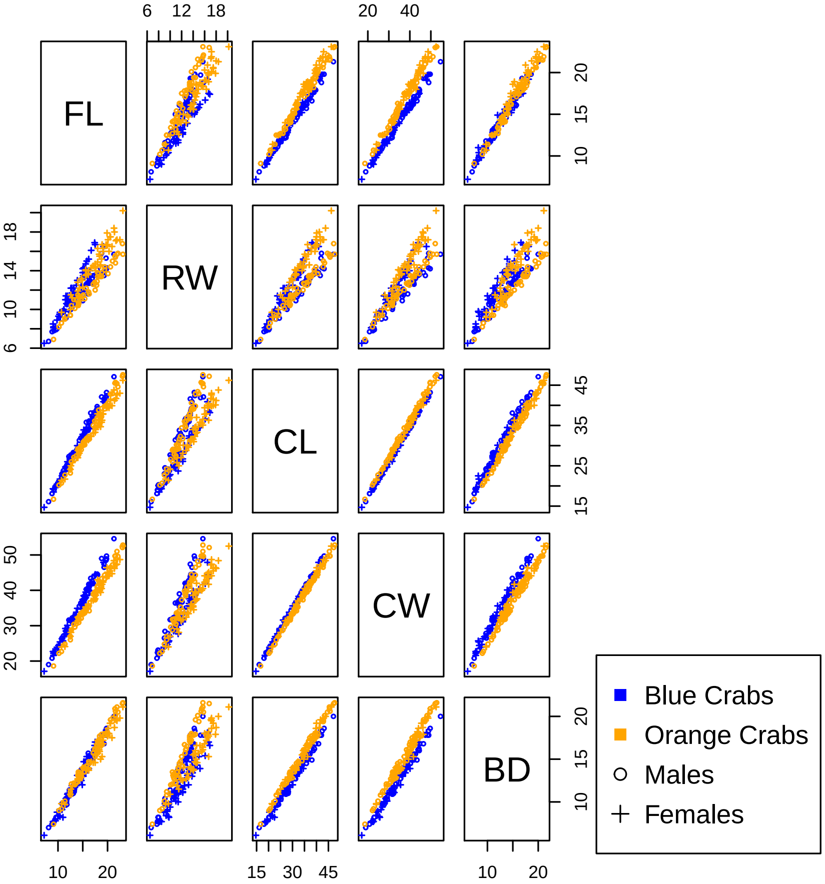
What would you initially anticipate as the answers to your questions?

I would initially anticipate the difference in species or sex is the reason for the heteroskedasticity in some of the scatterplots.

(c) (5 Points) Redo your scatterplot matrix from part (b) of the five quantitative variables (omitting *sp, sex & index*) via *baseR*. Now use different colors and symbols. Recall that "B" represents blue crabs and "O" represents orange crabs, so use these two colors. Use "o" to represent male crabs and "+" to represent female crabs. This is a perfect opportunity to make use of "ifelse" expressions in your R code! Also reduce the symbol size (*cex*) to 0.5. Describe this scatterplot matrix. Does this answer your previous questions from part (b)? Was your anticipation correct? Include your figure and your R code.

Answer:

```
> pairs(crabs.quant, oma = c(3, 3, 3, 17),
+       col = ifelse(crabs[, 1] == "B", "Blue", "Orange"),
+       pch = ifelse(crabs[, 2] == "M", 1, 3), cex = 0.5,
+       main = "Crab size correlations by Sex and Species")
> par(xpd = TRUE)
> legend("bottomright",
+        legend = c("Blue Crabs", "Orange Crabs", "Males", "Females"),
+        pch = c(15, 15, 1, 3), col = c("Blue", "Orange", "Black",
+                                       "Black"))
> reset.par()
```



Comment:

Describe this scatterplot matrix.

The same as the previous scatterplot matrix except that we have colors and symbols to represent species and sex.

Does this answer your previous questions from part (b)?

Yes because I can see that if I subset by sex or species, the heteroskedastic plots become homoskedastic.

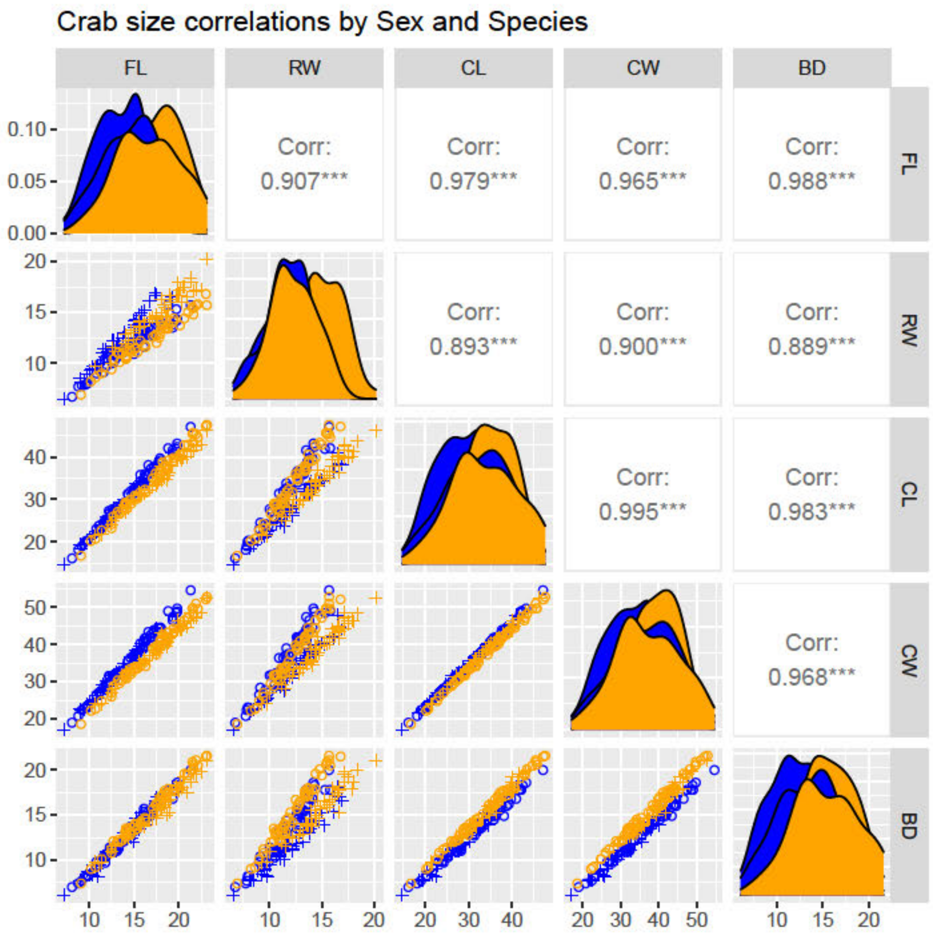
Was your anticipation correct?

Yes.

(d) (4 Points) Redo your scatterplot matrix from part (c) of the five quantitative variables (omitting *sp, sex & index*) via *ggpairs*. Use the same colors and symbols as in part (c). For everything else, use the default settings and show the four densities on the diagonal and the correlations in the upper triangular matrix. Include your figure and your R code.

Answer:

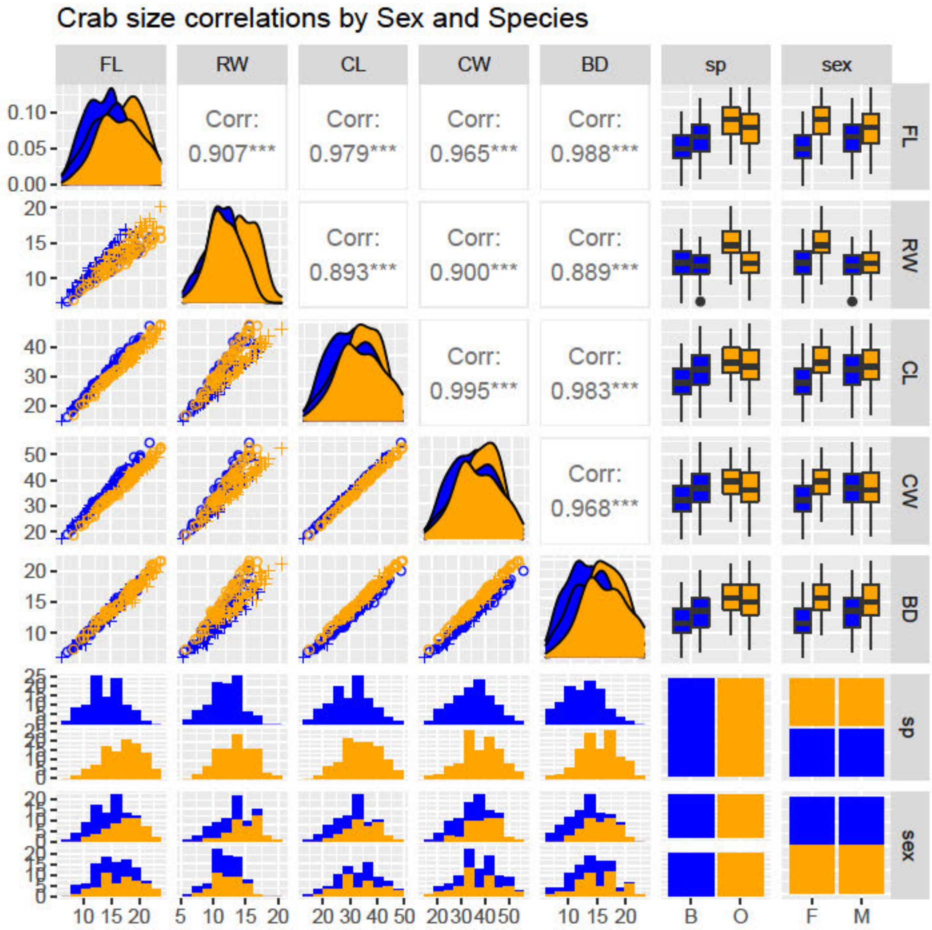
```
> l.w.col <- list(continuous = wrap("points",
+                                   color = ifelse(crabs[, 1] == "B",
+                                                 "Blue", "Orange")))
> ggpairs(crabs, columns = 4:8,
+         lower = l.w.col,
+         aes(shape = crabs[, 2], fill = crabs[, 1])) +
+ scale_fill_manual(values = c("B" = "Blue", "O" = "Orange")) +
+ scale_shape_manual(values = c("M" = 1, "F" = 3)) +
+ ggtitle("Crab size correlations by Sex and Species")
>
```



(e) (5 Points) Redo your scatterplot matrix from part (d) of the five quantitative variables including *sp & sex* (but still omitting *index*) via *ggpairs*. Place *sp & sex* after the five quantitative variables so that the histograms and box plots appear on the bottom and on the right. Use the same colors and symbols as in part (d). Choose 10 bins for all histograms. For everything else, use the default settings and show the four densities on the diagonal and the correlations in the upper triangular matrix. Include your figure and your R code.

Answer:

```
> l.w.c.bin <- list(continuous = wrap("points",
+                                      color = ifelse(crabs[, 1] == "B",
+                                                     "Blue", "Orange")),
+                       combo = wrap("facethist", bins = 10))
> ggpairs(crabs, column = c(4:8, 1:2),
+         lower = l.w.c.bin,
+         aes(shape = crabs[, 2], fill = crabs[, 1])) +
+ scale_fill_manual(values = c("B" = "Blue", "O" = "Orange")) +
+ scale_shape_manual(values = c("M" = 1, "F" = 3)) +
+ ggtitle("Crab size correlations by Sex and Species")
>
>
```


Crab size correlations by Sex and Species
FL
RW
CL
CW
BD
sp
sex
Corr: 0.907***
Corr: 0.979***
Corr: 0.965***
Corr: 0.988***
Corr: 0.893***
Corr: 0.900***
Corr: 0.889***
Corr: 0.995***
Corr: 0.983***
Corr: 0.968***
0.00
0.05
0.10
10
15
20
20
30
40
50
B
O
F
M
5
10 15 20
20 30 40 50
20304050

(f) (6 Points) Frankly, I have a problem with the interpretation of the (extended) scatterplot matrix from part (e). Can you easily identify which boxplot and histogram belongs to which species or which sex? Therefore, do the following: Add an "interaction variable" of *sp.sex* to the crabs data set. Recall how we did this for "SexClassSurvived" for the Titanic data set. **Reorder your factor levels as B.M, B.F, O.M, and O.F. Do not leave the factor levels in the order that was produced by R.** We have done such reorderings of factor levels in *Statistical Visualization I.*

Redo your scatterplot matrix from part (e), but now with *sp.sex* as the only categorical variable. Use a diverging 4–class *RdYlBu* color scheme from the *RColorBrewer* R package where "blue" represents blue crabs and "red" represents orange crabs. Use the two darker (outside) colors for the male crabs and the two fainter (inside) colors for the female crabs. If you haven't used the *RColorBrewer* R package so far, glance forward in the lecture notes how this R package can be used. Or, extract the RGB color values from `https://colorbrewer2.org/#type=diverging&scheme=RdYlBu&n=4` and use them as input for the *rgb()* function from baseR.

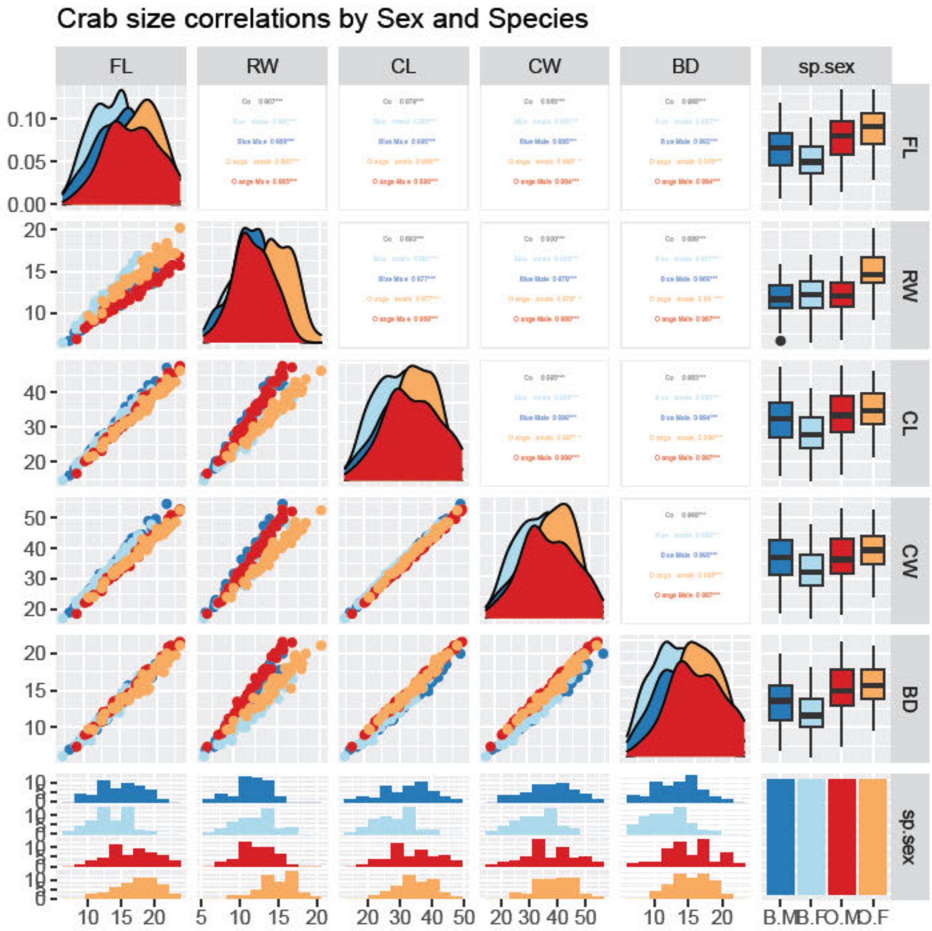
Warning: Make sure that you correctly map these colors to the species and sex. If you are not certain, compare with your scatterplot matrix from *baseR*.

As before, choose 10 bins for all histograms. For everything else, use the default settings and show the four densities on the diagonal and the correlations in the upper triangular matrix. Reduce alpha to 0.5 to address the overplotting problem. Likely, your font size will be too large for the values of the correlations. Also check and adjust the font sizes on the axes if necessary. Recall that this has to look good in your final pdf version and not in the Plots window in RStudio. If necessary, google for a solution how to adjust the font sizes in *ggpairs*. Include your figure and your R code.

Answer:

```
> crabs.gg <- crabs
> crabs.gg$sp.sex <- factor(with(crabs,
+                                interaction(crabs[, 1], crabs[, 2])),
+                           levels = c("B.M", "B.F", "O.M", "O.F"))
> sp.sex.colors <- brewer.pal(4, "RdYlBu")
> ggplot <- function(...) ggplot2::ggplot(...) +
+ scale_color_manual(values = c(sp.sex.colors[3], sp.sex.colors[4],
+                               sp.sex.colors[2], sp.sex.colors[1])) +
+ scale_fill_manual(values = c(sp.sex.colors[3], sp.sex.colors[4],
+                               sp.sex.colors[2], sp.sex.colors[1]))
```

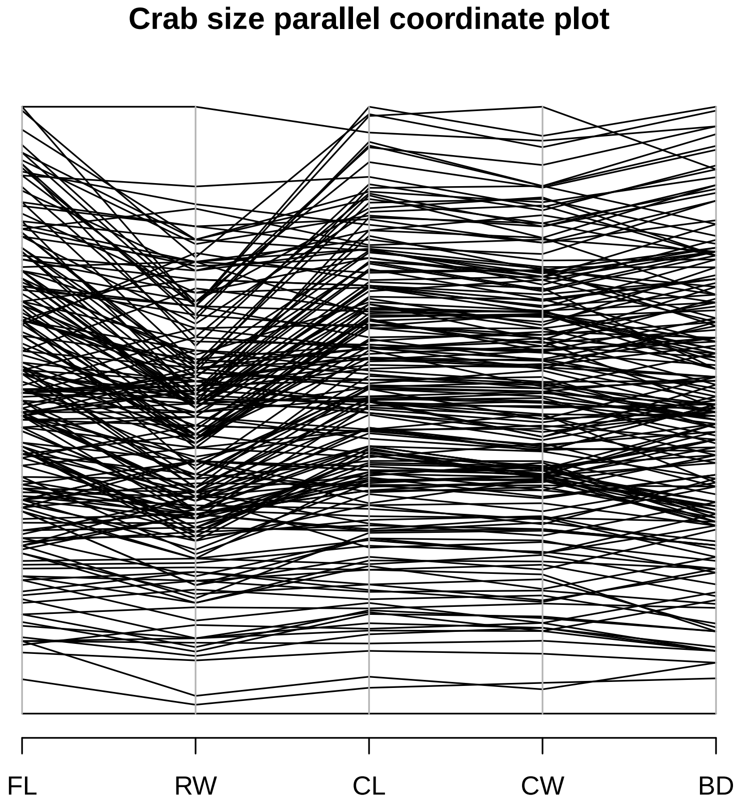
```
> unlockBinding("ggplot", parent.env(asNamespace("GGally")))
> assign("ggplot", ggplot, parent.env(asNamespace("GGally")))
> ggpairs(crabs.gg, columns = 4:9,
+         lower = list(combo = wrap("facethist", bins = 10)),
+         upper = list(continuous = wrap("cor",
+                                        alpha = 0.5, size = 1)),
+         aes(color = gsub("B.M", "Blue Male",
+                          gsub("B.F", "Blue Female",
+                               gsub("O.M", "Orange Male",
+                                    gsub("O.F", "Orange Female",
+                                         crabs.gg[, 9])))))) +
+ ggtitle("Crab size correlations by Sex and Species")
> ggplot <- function(...) ggplot2::ggplot(...)
> unlockBinding("ggplot", parent.env(asNamespace("GGally")))
> assign("ggplot", ggplot, parent.env(asNamespace("GGally")))
>
>
```



(g) (4 Points) Create a default parallel coordinates plot (PCP) of the five quantitative variables (omitting *sp, sex & index*) via *baseR*. Do not optimize this PCP and do not use any colors. Describe this PCP. Are there any variables that allow to separate one of the species or sexes? Include your figure and your R code.

Answer:

```
> parcoord(crabs[, 4:8], main = "Crab size parallel coordinate plot")
```



Comment:

Describe this PCP.

We have lines that have different standardized values across the variables (not relatively level lines) and some lines that have similar standardized values across the variables (relatively level lines). Most of the values are in the middle.

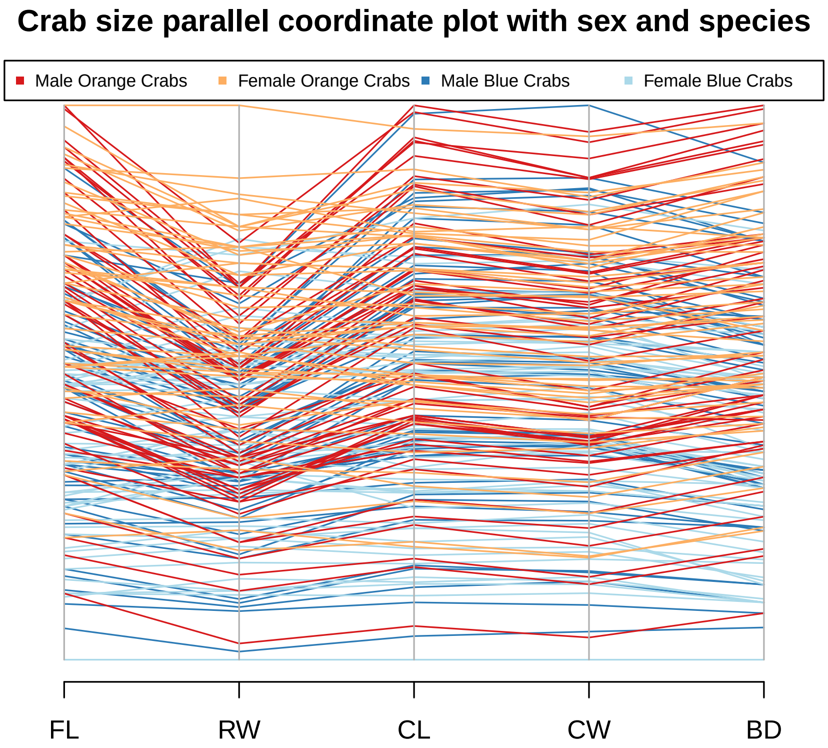
Are there any variables that allow to separate one of the species or sexes?

Blue Crabs tend to have lower values of BD [and Orange Crabs tend to have higher values]. Male crabs tend to have lower values of RW [and female crabs tend to have higher values].

(h) (5 Points) Redo your PCP from part (g) of the five quantitative variables (omitting *sp, sex & index*) via *baseR*. Now use the same colors as in part (f). Are there any variables that allow to partially separate one of the species or sexes? Include your figure and your R code.

Answer:

```
> par(oma = c(1, 0, 1, 0), mar = c(5, 4, 4, 2))
> col.title <- "Crab size parallel coordinate plot with sex and species"
> parcoord(crabs.gg[, 4:8],
+          col = gsub("B.M", sp.sex.colors[4],
+                     gsub("B.F", sp.sex.colors[3],
+                          gsub("O.M", sp.sex.colors[1],
+                               gsub("O.F", sp.sex.colors[2],
+                                    crabs.gg[, 9])))),
+          main = col.title)
> par(oma = c(0, 0, 0, 0), xpd = TRUE)
> legend("top", ncol = 4, pch = c(15, 15, 15, 15), cex = .65,
+        legend = c("Male Orange Crabs", "Female Orange Crabs",
+                   "Male Blue Crabs", "Female Blue Crabs"),
+        col = c(sp.sex.colors[1], sp.sex.colors[2],
+                sp.sex.colors[4], sp.sex.colors[3]))
> reset.par()
```



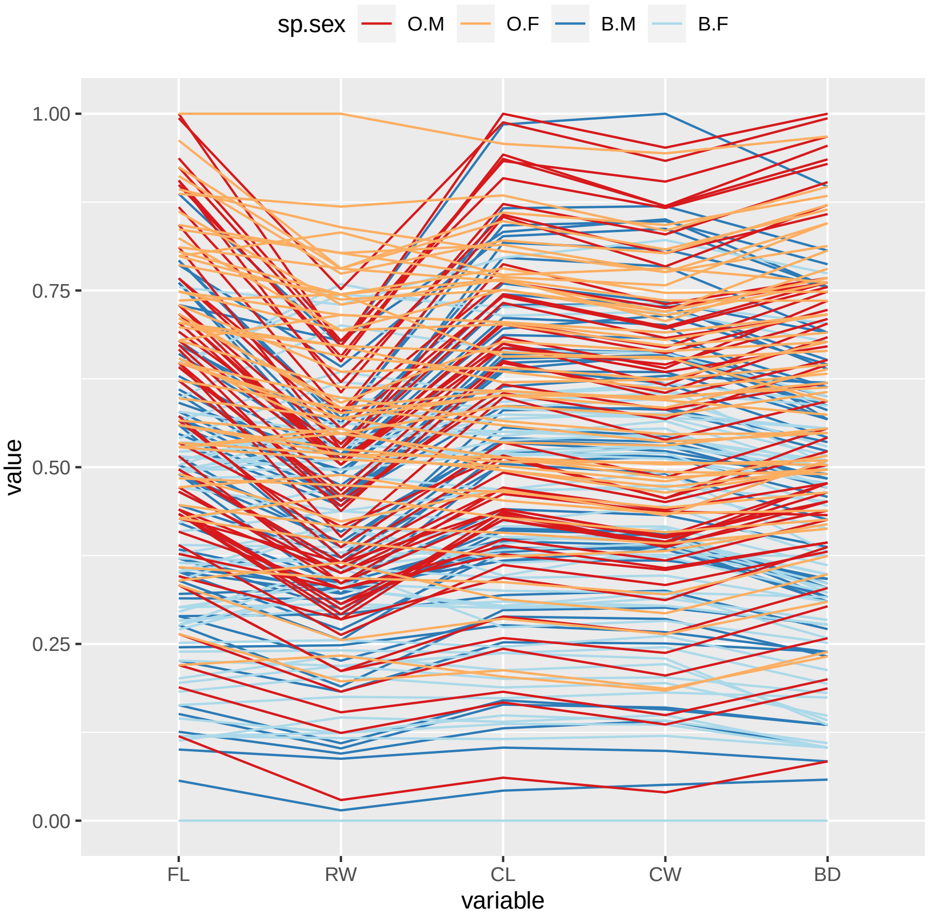
Comment:

Are there any variables that allow to partially separate one of the species or sexes?

(i) (5 Points) Redo your PCP from part (h) of the five quantitative variables (omitting *sp, sex & index*) via the *ggparcoord* function from the *GGally* R package. Use the same colors as in part (f). Use a 0–1–scale for all parallel axes. Hint: The legend of this graph should reveal whether you correctly use the colors as specified in part (f). If something goes wrong here, likely you use the colors incorrectly in some of your previous parts as well. Include your figure and your R code.

Answer:

```
> ggplot <- function(...) ggplot2::ggplot(...) +
+ scale_color_manual(values = c(sp.sex.colors[1], sp.sex.colors[2],
+                               sp.sex.colors[4], sp.sex.colors[3]))
> unlockBinding("ggplot", parent.env(asNamespace("GGally")))
> assign("ggplot", ggplot, parent.env(asNamespace("GGally")))
> crabs.gg.p <- crabs
> crabs.gg.p$sp.sex <- factor(with(crabs,
+                                  interaction(crabs[, 1],
+                                              crabs[, 2])),
+                           levels = c("O.M", "O.F", "B.M", "B.F"))
> ggparcoord(crabs.gg.p, columns = 4:8, groupColumn = 9,
+            scale = "uniminmax") +
+
+ theme(legend.position="top")
>
```


sp.sex
O.M
O.F
B.M
B.F
1.00
0.75
0.50
0.25
0.00
value
FL
RW
CL
CW
BD
variable

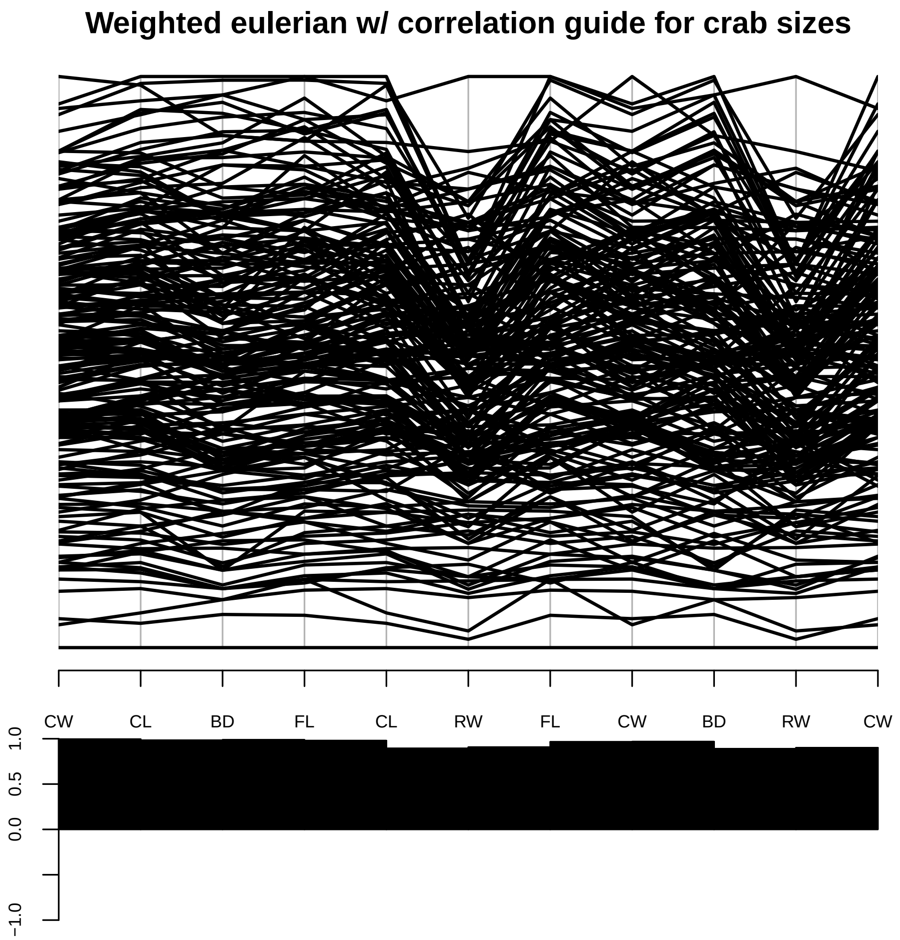
(j) (6 Points) Let's try something: First install the *BiocManager* R package from CRAN. Then install the *graph* R package from Bioconductor as follows:

```
BiocManager::install("graph", version = "3.12")
```

Finally install the *PairViz* R package from CRAN. Read the PCP vignette at `https://cran.r-project.org/web/packages/PairViz/vignettes/pcp.html`. Adapt the example code to create a "Weighted Eulerian with Correlation Guide," placing the bars for the correlations below the PCP. Do not color–code the lines in the PCP. Describe and interpret this version of the PCP (recall the made–up examples from class). It will also be helpful to revisit the correlations in the scatterplot matrix in part (f). Include your figure and your R code.

Answer:

```
> data <- crabs.gg[, 4:8]
> corw <- as.dist(cor(data))
> o <- eulerian(-corw)
> corw <- dist2edge(corw)
> edgew <- cbind(corw*(corw>0), corw*(corw<0))
> par(cex.axis=.7)
> guided_pcp(crabs.gg[, 4:9], edgew, path=o, lwd=2,
+          main="Weighted eulerian w/ correlation guide for crab sizes",
+          bar.ylim=c(-1,1), bar.axes=TRUE)
>
```



Comment:

Describe and interpret this version of the PCP (recall the made–up examples from class).

Correlations between RW and any other variable are relatively low.

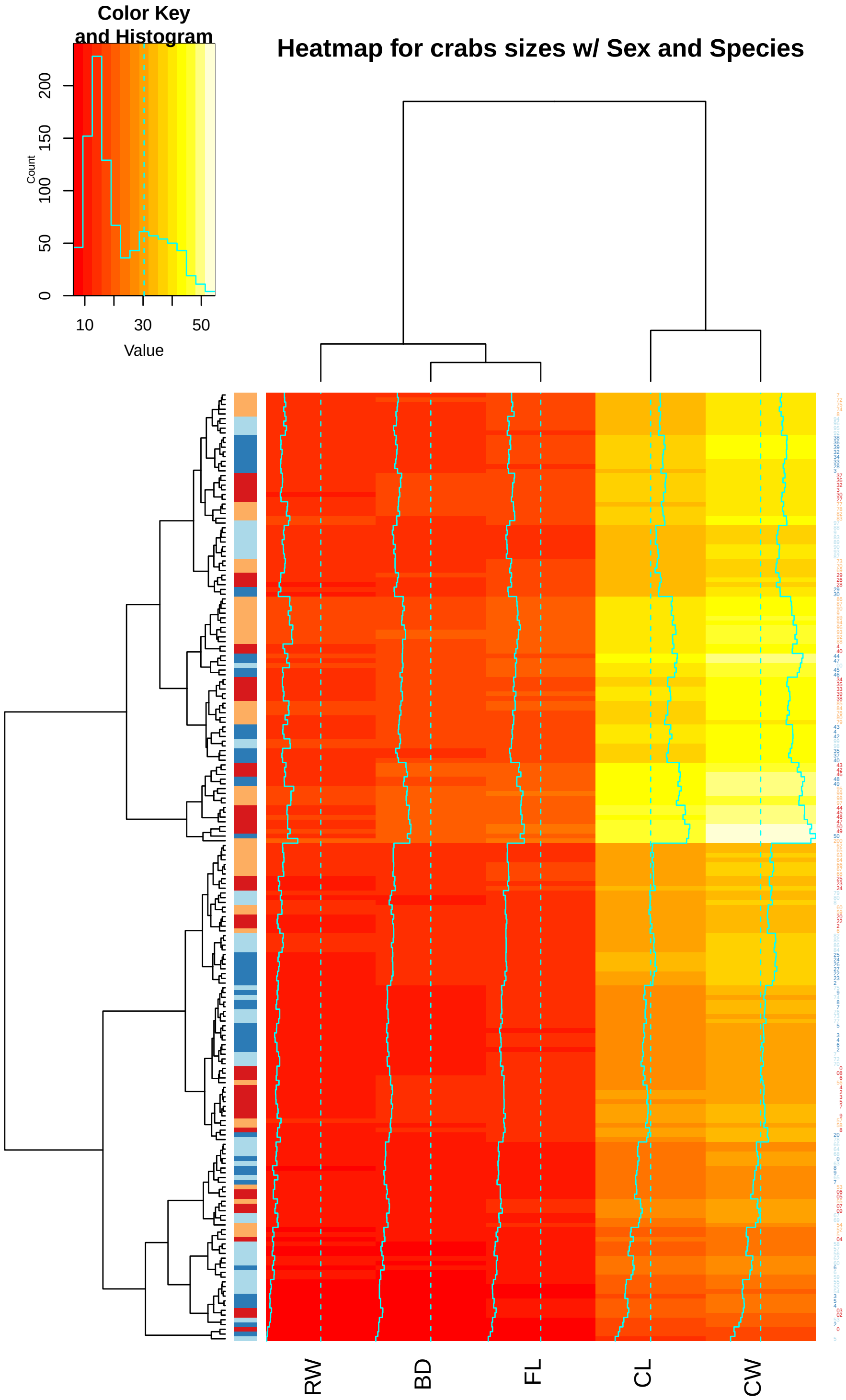
(k) (5 Points) Create a heatmap of the five quantitative variables (omitting *sp, sex & index*) via the *heatmap.2()* function from the *gplots* R package. Make sure to scale your variables!

Show the index variable on the right side of the heatmap and color it, using the same colors as in the previous question parts. Also display the colors for the four groups via the `RowSideColors` argument on the left side of the heatmap. Make sure that your colors on the left (for the boxes) and on the right (for the fonts) are matching.

Also make sure that your heatmap extends (in height) over a full output page or otherwise, it will be difficult to distinguish results and labels. Look up help pages if necessary to obtain all of these features. What can we learn from this heatmap (if anything at all)? Include your figure and your R code.

Answer:

```
> heatmap.2(as.matrix(crabs[, 4:8]),
+           RowSideColors = gsub("B.M", sp.sex.colors[4],
+                                gsub("B.F", sp.sex.colors[3],
+                                     gsub("O.M", sp.sex.colors[1],
+                                          gsub("O.F",
+                                               sp.sex.colors[2],
+                                               crabs.gg[, 9])))),
+           colRow = gsub("B.M", sp.sex.colors[4],
+                         gsub("B.F", sp.sex.colors[3],
+                              gsub("O.M", sp.sex.colors[1],
+                                   gsub("O.F", sp.sex.colors[2],
+                                        crabs.gg[, 9])))),
+           cexRow = 0.4, labRow = rownames(crabs),
+           main = "Heatmap for crabs sizes w/ Sex and Species")
```


Color Key
and Histogram
Count
0
50
100
150
200
10
30
50
Value
Heatmap for crabs sizes w/ Sex and Species
RW
BD
FL
CL
CW

Comment:

What can we learn from this heatmap (if anything at all)?

From the column clustering we can see that CL and CW are very much corolated with each other.

(l) (8 Points) Start with a basic description of the data set and its variables. Then provide a detailed interpretation of your graphical results. Point out differences and similarities for the two species and the two sexes. Also comment on similarities of the variables. Do this for the univariate graphical summaries of the five quantitative variables, the pairwise scatterplots of these variables, the PCPs, and the heatmap. Compare and combine results obtained from the various graphs. Ideally, different graphs should result in similar interpretations, but, if necessary, point out when different graphs result in contradictory interpretations. Be specific and refer back to specific question parts and do not just say "the PCP or the scatterplot matrix shows …". This summary should be at least 1 page in length.

Answer:

From the scatterplot matricies, we can see that Orange Crabs typically have larger sizes than the Blue Crabs. Interestingly, the Female Orange Crab is typically larger than the Male Orange Crab; but, the Male Blue Crab is typically larger than the Female Blue Crab in all dimensions except RW. Continuing with this, we see that RW is a significantly smaller size for all the male crabs from our Parallel Coordinate plot. From the Weighted Eulerian with Correlation guide of the crabs, we see that RW has the smallest correlation between all the other variables, and a significant amount have a dip in RW between each of their other dimensions. Finally with the heatmap, we see that dispite the differences between sex and species of the crab, all types of crabs appear in all types of sizes.

(ii) [redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

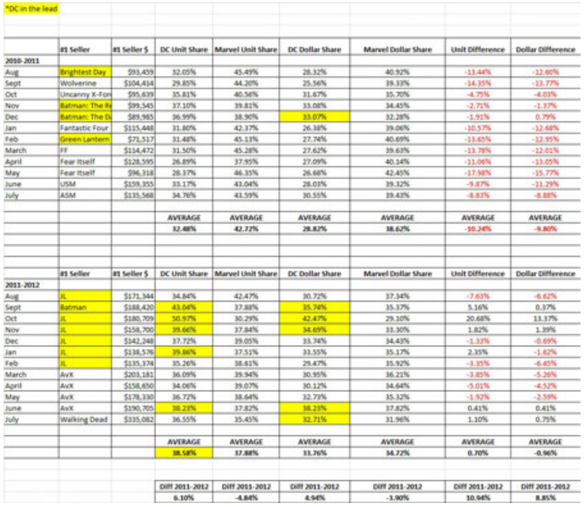
[redacted]

[redacted]

[redacted]

[redacted]

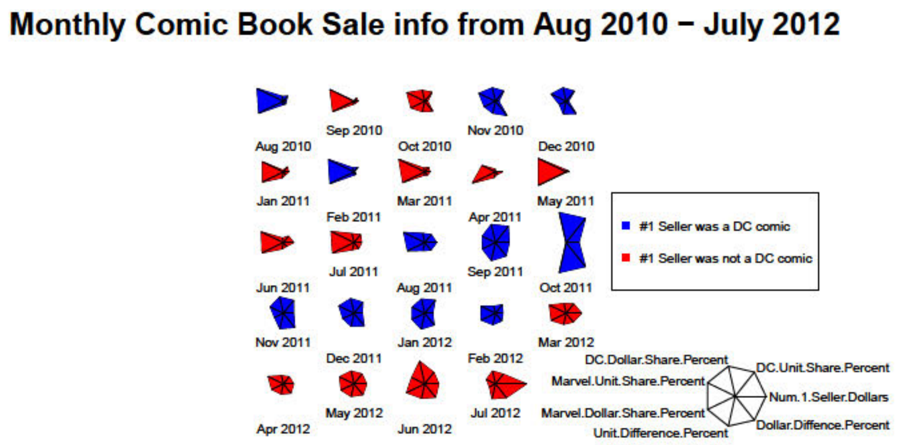
```
> par(mfrow = c(2, 1))
> dc.v.o.sales <- data.frame("Month" = c("Aug", "Sep", "Oct", "Nov",
+                                         "Dec", "Jan", "Feb", "Mar",
+                                         "Apr", "May", "Jun", "Jul",
+                                         "Aug", "Sep", "Oct", "Nov",
+                                         "Dec", "Jan", "Feb", "Mar",
+                                         "Apr", "May", "Jun", "Jul"),
+                            "Num 1 Seller" = c("Brightest Day",
+                                               "Wolverine",
+                                               "Uncanny X-Force",
+                                               "Batman: The Return",
+                                               paste("Batman: The",
+                                                     "Dark Knight"),
```

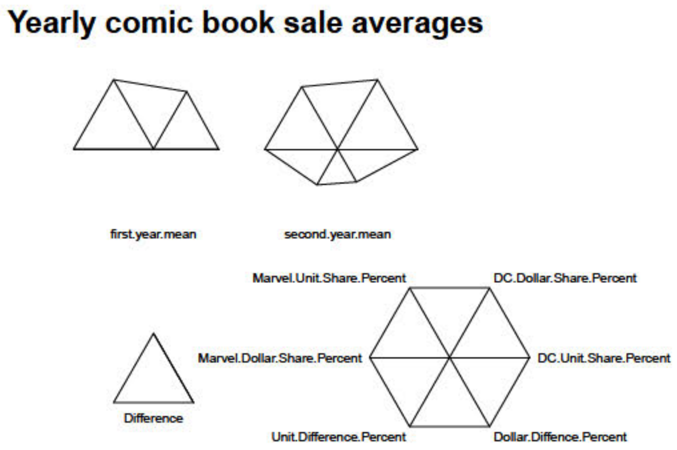


*DC in the lead								
	#1 Seller	#1 Seller $	DC Unit Share	Marvel Unit Share	DC Dollar Share	Marvel Dollar Share	Unit Difference	Dollar Difference
2010-2011								
Aug	Brightest Day	$93,459	32.05%	45.49%	28.32%	40.92%	-13.44%	-12.60%
Sept	Wolverine	$104,414	29.85%	44.20%	25.56%	39.33%	-14.35%	-13.77%
Oct	Uncanny X-For	$95,639	35.81%	40.56%	31.67%	35.70%	-4.75%	-4.03%
Nov	Batman: The Re	$99,545	37.10%	39.81%	33.08%	34.45%	-2.71%	-1.37%
Dec	Batman: The Da	$89,985	36.99%	38.90%	33.07%	32.28%	-1.91%	0.79%
Jan	Fantastic Four	$115,448	31.80%	42.37%	26.38%	39.06%	-10.57%	-12.68%
Feb	Green Lantern	$71,517	31.48%	45.13%	27.74%	40.69%	-13.65%	-12.95%
March	FF	$114,472	31.50%	45.28%	27.62%	39.63%	-13.78%	-12.01%
April	Fear Itself	$128,595	26.89%	37.95%	27.09%	40.14%	-11.06%	-13.05%
May	Fear Itself	$96,318	28.37%	46.35%	26.68%	42.45%	-17.98%	-15.77%
June	USM	$159,355	33.17%	43.04%	28.03%	39.32%	-9.87%	-11.29%
July	ASM	$135,568	34.76%	43.59%	30.55%	39.43%	-8.83%	-8.88%
			AVERAGE	AVERAGE	AVERAGE	AVERAGE	AVERAGE	AVERAGE
			32.48%	42.72%	28.82%	38.62%	-10.24%	-9.80%
	#1 Seller	#1 Seller $	DC Unit Share	Marvel Unit Share	DC Dollar Share	Marvel Dollar Share	Unit Difference	Dollar Difference
2011-2012								
Aug	JL	$171,344	34.84%	42.47%	30.72%	37.34%	-7.63%	-6.62%
Sept	Batman	$188,420	43.04%	37.88%	35.74%	35.37%	5.16%	0.37%
Oct	JL	$180,709	50.97%	30.29%	42.47%	29.10%	20.68%	13.37%
Nov	JL	$158,700	39.66%	37.84%	34.69%	33.30%	1.82%	1.39%
Dec	JL	$142,248	37.72%	39.05%	33.74%	34.43%	-1.33%	-0.69%
Jan	JL	$138,576	39.86%	37.51%	33.55%	35.17%	2.35%	-1.62%
Feb	JL	$135,374	35.26%	38.61%	29.47%	35.92%	-3.35%	-6.45%
March	AvX	$203,181	36.09%	39.94%	30.95%	36.21%	-3.85%	-5.26%
April	AvX	$158,650	34.06%	39.07%	30.12%	34.64%	-5.01%	-4.52%
May	AvX	$178,330	36.72%	38.64%	32.73%	35.32%	-1.92%	-2.59%
June	AvX	$190,705	38.23%	37.82%	38.23%	37.82%	0.41%	0.41%
July	Walking Dead	$335,082	36.55%	35.45%	32.71%	31.96%	1.10%	0.75%
			AVERAGE	AVERAGE	AVERAGE	AVERAGE	AVERAGE	AVERAGE
			38.58%	37.88%	33.76%	34.72%	0.70%	-0.96%
			Diff 2011-2012	Diff 2011-2012	Diff 2011-2012	Diff 2011-2012	Diff 2011-2012	Diff 2011-2012
			6.10%	-4.84%	4.94%	-3.90%	10.94%	8.85%

Figure 1: Original jpg of Excel sheet

```
+                                               "Fantastic Four",
+                                               "Green Latern", "FF",
+                                               "Fear Itself",
+                                               "Fear Itself", "USM",
+                                               "ASM", "JL", "Batman",
+                                               "JL", "JL", "JL", "JL",
+                                               "JL", "AvX", "AvX",
+                                               "AvX", "AvX",
+                                               "Walking Dead"),
+                            "Num 1 Seller Dollars" = c(93499, 104414,
+                                                       95639, 99545,
+                                                       89985, 115448,
+                                                       71517, 114472,
+                                                       128595, 96318,
+                                                       159355, 135568,
+                                                       171344, 188420,
+                                                       180709, 158700,
+                                                       142248, 138576,
+                                                       135374, 203181,
+                                                       158650, 178330,
+                                                       190705, 335082),
+                            "Num 1 Seller DC" = c(TRUE, FALSE, FALSE,
+                                                  TRUE, TRUE, FALSE,
+                                                  TRUE, FALSE, FALSE,
+                                                  FALSE, FALSE, FALSE,
+                                                  TRUE, TRUE, TRUE,
+                                                  TRUE, TRUE, TRUE,
+                                                  TRUE, FALSE, FALSE,
+                                                  FALSE, FALSE, FALSE),
+                            "DC Unit Share Percent" = c(32.05, 29.85,
+                                                       35.81, 37.1,
+                                                       36.99, 31.8,
+                                                       31.48, 31.5,
+                                                       26.89, 28.37,
+                                                       33.17, 34.76,
+                                                       34.84, 43.04,
+                                                       50.97, 39.66,
+                                                       37.72, 39.86,
+                                                       35.26, 36.09,
+                                                       34.06, 36.72,
+                                                       38.23, 36.55),
+                            "Marvel Unit Share Percent" = c(45.49,
+                                                            44.2,
+                                                            40.56,
+                                                            39.81,
+                                                            38.9,
+                                                            42.37,
+                                                            45.13,
+                                                            45.28,
+                                                            37.95,
+                                                            46.35,
+                                                            43.04,
+                                                            43.59,
+                                                            42.47,
+                                                            37.88,
+                                                            30.29,
+                                                            37.84,
+                                                            39.05,
+                                                             37.51,
+                                                             38.61,
+                                                             39.94,
+                                                             39.07,
+                                                             38.64,
+                                                             37.82,
+                                                             35.45),
+                             "DC Dollar Share Percent" = c(28.32, 25.56,
+                                                           31.67, 33.08,
+                                                           33.07, 26.38,
+                                                           27.74, 27.62,
+                                                           27.09, 26.68,
+                                                           28.03, 30.55,
+                                                           30.72, 35.74,
+                                                           42.47, 34.69,
+                                                           33.74, 33.55,
+                                                           29.47, 30.95,
+                                                           30.12, 32.73,
+                                                           38.23,
+                                                           32.71),
+                             "Marvel Dollar Share Percent" = c(40.92,
+                                                               39.33,
+                                                               35.7,
+                                                               34.45,
+                                                               32.28,
+                                                               39.06,
+                                                               40.69,
+                                                               39.63,
+                                                               40.14,
+                                                               42.45,
+                                                               39.32,
+                                                               39.43,
+                                                               37.34,
+                                                               35.37,
+                                                               29.1,
+                                                               33.3,
+                                                               34.43,
+                                                               35.17,
+                                                               35.92,
+                                                               36.21,
+                                                               34.64,
+                                                               35.32,
+                                                               37.82,
+                                                               31.96),
+                             "Unit Difference Percent" = c(-13.44,
+                                                           -14.35,
+                                                           -4.75, -2.71,
+                                                           -1.91,
+                                                           -10.57,
+                                                           -13.65,
+                                                           -13.78,
+                                                           -11.06,
+                                                           -17.98,
+                                                           -9.87, -8.83,
+                                                           -7.63, 5.16,
+                                                           20.68, 1.82,
+                                                           -1.33, 2.35,
+                                                           -3.35, -3.85,
+                                                           -5.01, -1.92,
+                                                           0.41, 1.1),
+                             "Dollar Diffence Percent" = c(-12.6,
+                                                           -13.77,
+                                                           -4.03, 1.37,
+                                                           -0.79,
+                                                           -12.68,
+                                                           -12.95,
+                                                           -12.01,
+                                                           -13.05,
+                                                           -15.77,
+                                                           -11.29,
+                                                           -8.88, -6.62,
+                                                           0.37, 13.37,
+                                                           1.39, -0.69,
+                                                           -1.62, -6.45,
+                                                           -5.26, -4.52,
+                                                           -2.59, 0.41,
+                                                           0.75),
+                            "Year" = c(2010, 2010, 2010, 2010, 2010,
+                                       2011, 2011, 2011, 2011, 2011,
+                                       2011, 2011, 2011, 2011, 2011,
+                                       2011, 2011, 2012, 2012, 2012,
+                                       2012, 2012, 2012, 2012))
> first.year.mean <- apply(dc.v.o.sales[, c(5, 7, 6, 8:10)][1:12, ],
+                           2, mean)
> second.year.mean <- apply(dc.v.o.sales[, c(5, 7, 6, 8:10)][13:24, ],
+                            2, mean)
> stars(dc.v.o.sales[1:24, c(3, 5, 7, 6, 8:10)], key.loc = c(17, 1.9),
+       xpd = TRUE, labels = paste(dc.v.o.sales[1:24, 1],
+                                  dc.v.o.sales[1:24, 11]), cex = .5,
+       col.stars = ifelse(dc.v.o.sales[, 4] == TRUE, "Blue", "Red"),
+       main = "Monthly Comic Book Sale info from Aug 2010 - July 2012")
> legend(x = 13, y = 8.5, legend = c("#1 Seller was a DC comic",
+                             "#1 Seller was not a DC comic"),
+        col = c("Blue", "Red"), pch = c(15, 15), cex = .5)
> stars(rbind(first.year.mean, second.year.mean,
+             "Difference" = second.year.mean - first.year.mean),
+       key.loc = c(6, 2), cex = .5,
+       main = "Yearly comic book sale averages")
> reset.par()
```





(c) (8 Points) In addition to translating the table into a graph, you have to explain and motivate your resulting graph! Why did you choose that particular graph design? What sorting did you use for the rows and columns of the original table? How (and why) did you combine data from the table, e.g., means and standard deviations from the table into means and confidence intervals in your graph? How (and why) did you choose specific colors and glyphs in your graph, etc.?

Also emphasize what are the most important features that can be seen in your graph. Keep in mind that the general audience may have never seen such a graph type before. So, you may have to explain some of the basic design features of your graph.

So from my original table I changed some things up when I copied it to R, where I replace the highlighting aspect with a column of logical values of whether or not the Number 1 comic selling comic for the month was by DC comics. I also condensed it into just one data.frame for both years and adding a column which tells the year. I also eliminated the averages as I instead calculate them in R from the given table.

I decided to use star plots for each of months and for the yearly averages because the table has a decent amount of variables that seem to be great for a star plot. In addition, the star plot works great for negative values making it a good possible choice. I believe that because of the decent amounts of variables, a star plot works better than a line plot for showing each of the variables as a time series (and the averages part of the table seems to focus more on the differences in the years rather than trends and seasonality we'd expect to identify with time series).

From the Monthly plots I rearranged the variables to see several things : DC is the 2 highest star plot verticies, while Marvel is the leftmost, Differences are the two bottommost and the rightmost is the Dollar amount of the Number 1 seller for that month, represented by DC by a blue star plot, or not DC by a red star plot (not necessarily a Marvel comic though). In addition for each grouping, the clockwisemost value is in number of units, while the other is in terms of dollars.Interestingly, the sale info for the 15 month showed DC comics vastly outperformed Marvel comics the most since the points for the Marvel comics are pretty much missing, despite the number 1 seller being low in dollars. Also, we see that the dollar amount from the number 1 seller in the 24th month vastly outperformed the sales of all the number 1 other comics, which makes sense since the it was "The Walking Dead" which had a very popular TV show still going on.

From the graphs of the yearly averages, they are arranged in a similar manner to the monthly sale info star plots. The right and top right are DC, left and top left are Marvel, and Bottom are the differences; in addition, clockwisemost value is in number of units, while the other is in terms of dollars for each group. With these star plots we see that the DC comics only sold on average a larger amount of comics than Marvel in the 2011-2012 year.

General Instructions

(i) Create a single pdf document, using R Markdown, Sweave, or knitr. When you take this course at the 6000–level, you have to use $\LaTeX$ in combination with 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) [if any] 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.

1 ”References”

- https://stackoverflow.com/a/22201568
- https://stackoverflow.com/a/29966785
- https://stackoverflow.com/a/51685306
- https://www.blopig.com/blog/2019/06/a-brief-introduction-to-ggpairs/
- https://ggobi.github.io/ggally/reference/wrap.html
- https://stackoverflow.com/a/2954610
- http://www.cookbook-r.com/Graphs/Legends_(ggplot2)/
- https://aosmith.rbind.io/2020/07/09/ggplot2-override-aes/
- https://www.biostars.org/p/261685/
- https://stackoverflow.com/a/18420757
- https://stackoverflow.com/a/50340979
- https://stackoverflow.com/a/18420757
- https://r.789695.n4.nabble.com/change-default-output-size-when-using-Sweave-td885499.html