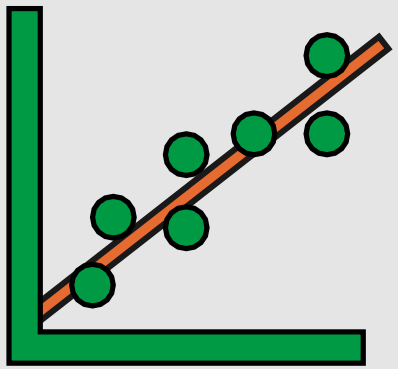
Making Magnificently Good Graphs: R

BERDC Special Topics Talk 3, Part 2



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Dr. Mark Williamson

Biostatistics, Epidemiology, and Research Design Core

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Overview

UNIVERSITY OF NORTH DAKOTA

- We'll cover how to make great looking graphs in R, primarily using the package ggplot2
- We'll start by creating basic graphs, then explore how to upgrade by modifying various elements
- Take the pre-test here https://und.qualtrics.com/jfe/form/SV_56jgWM3TxwYKZRI
- Get the R-code here https://med.und.edu/daccota/_files/docs/berdc_docs/mmgg_r_code.txt
- Get the PDF version here https://med.und.edu/daccota/_files/pdfs/berdc_resource_pdfs/mmgg_r.pdf
- Stay tuned for an extra-special treat at the end

Elements:

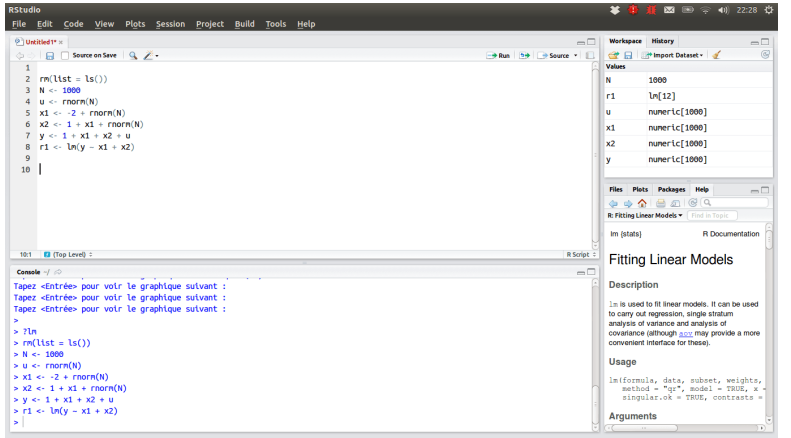
I. **Labels**
II. **Axes**
III. **Colors and Shapes**
IV. **Dots, Lines, and Text**
V. **Other**

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Getting Set Up

UNIVERSITY OF NORTH DAKOTA.

- I always suggest using R-studio or a script in a word editor
- Install package 'ggplot2', 'MASS', 'gmodels', and 'dplyr' if you haven't already
- ggplot2 is a structured way to create great graphs using 'the Grammar of Graphics'
- Here is a great reference sheet:
  - https://rstudio.com/wp-content/uploads/2016/11/ggplot2-cheatsheet-2.1.pdf



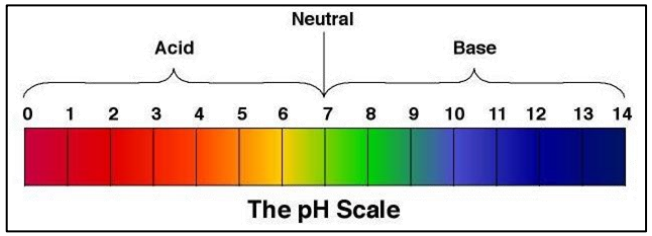
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

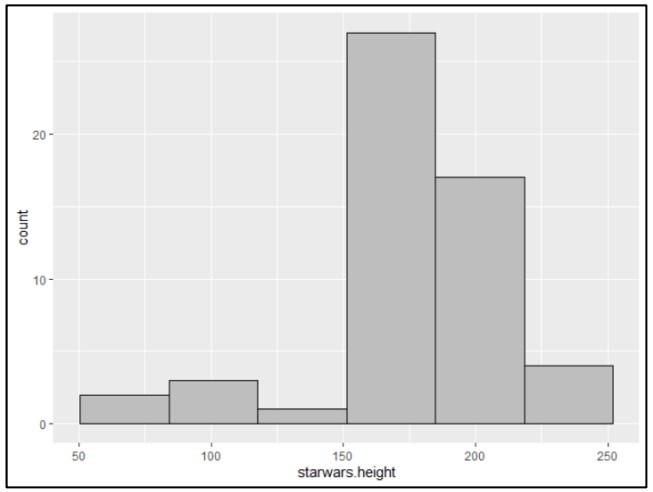
Histograms

UNIVERSITY OF NORTH DAKOTA

I. Basic histogram

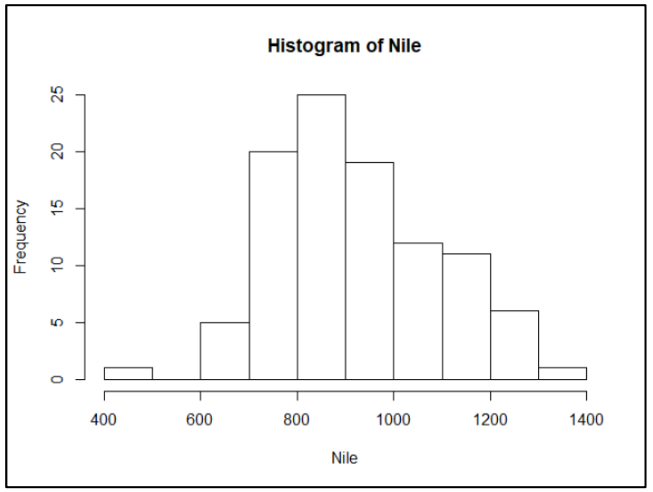
```
h1 <-hist(Nile)
h1
```





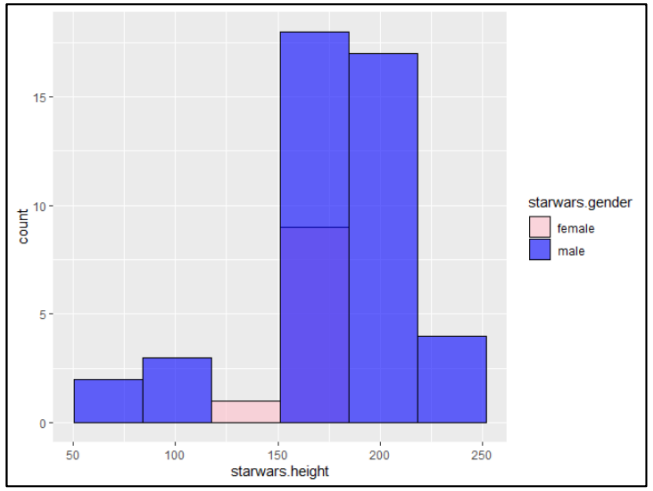
II. Simple histogram in ggplot

```
h2 <- ggplot(data=starwars2, aes(starwars.height))+
  geom_histogram(bins=6, fill="grey", col="black")
h2
```



III. Two-sample histogram in ggplot

```
h3 <- ggplot(data=starwars2, aes(x=starwars.height, fill=starwars.gender))+
  geom_histogram(bins=6, col="black", alpha=0.6, position='identity')+
  scale_fill_manual(values=c("pink", "blue"))
h3
```



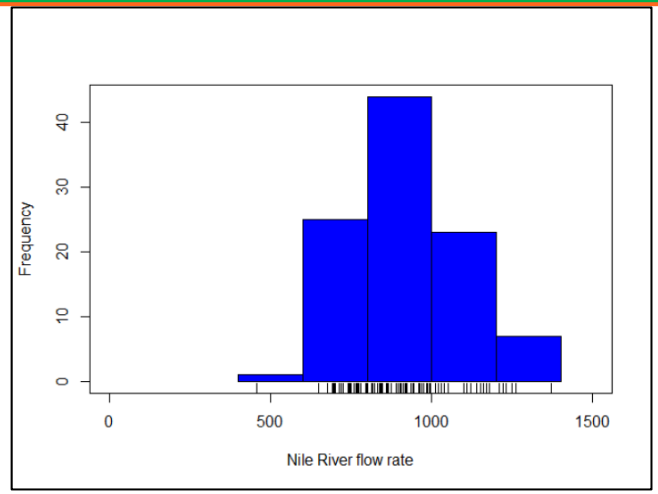
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Histograms cont.

UNIVERSITY OF NORTH DAKOTA

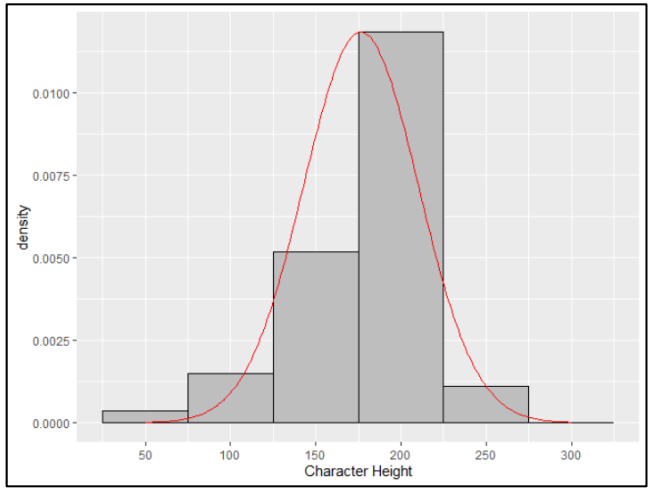
I. Basic histogram upgrade

```
h4 <- hist(Nile, col="blue", border="black", xlab="Nile River flow rate",
    xlim=c(0,1500), breaks=6, main='')
h4
rug(Nile)
box()
```



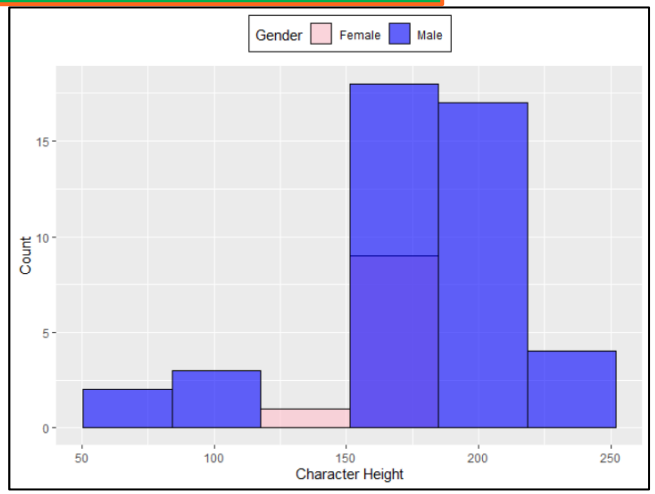
II. Simple histogram in ggplot upgrade

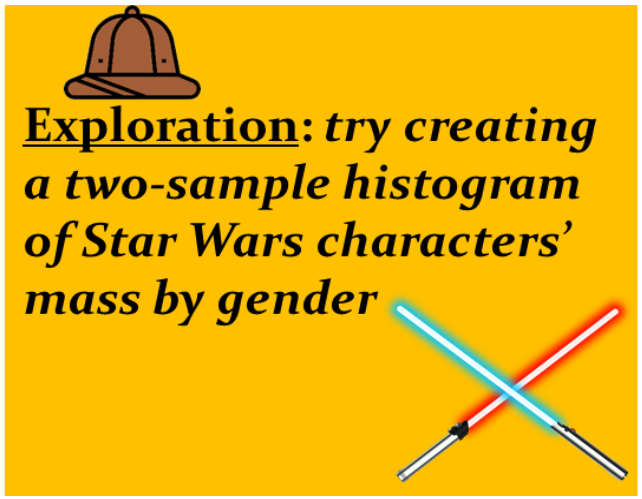
```
x <- seq(50, 300, length.out=250)
df <- with(starwars2, data.frame(x = x, y = dnorm(x, mean(starwars.height),
                    sd(starwars.height))))
h5 <- ggplot(data=starwars2, aes(x=starwars.height))+
 geom_histogram(aes(y=stat(density)), bins=6, fill="grey", col="black") +
 labs(x="Character Height") +
 scale_x_continuous(breaks=c(50, 100, 150, 200, 250, 300)) +
 geom_line(data=df, aes(x=x, y=y), color="red")
h5
```



III. Two-sample histogram in ggplot upgrade

```
h6 <- ggplot(data=starwars2, aes(x=starwars.height, fill=starwars.gender))+
  geom_histogram(bins=6, col="black", alpha=0.6, position='identity') +
  scale_fill_manual(values=c("pink", "blue"), name="Gender",
                    labels=c("Female","Male")) +
  labs(x="Character Height", y="Count") +
  scale_x_continuous(breaks=c(50, 100, 150, 200, 250, 300)) +
  theme(legend.background=element_rect(colour="black"), legend.position ="top")
h6
```





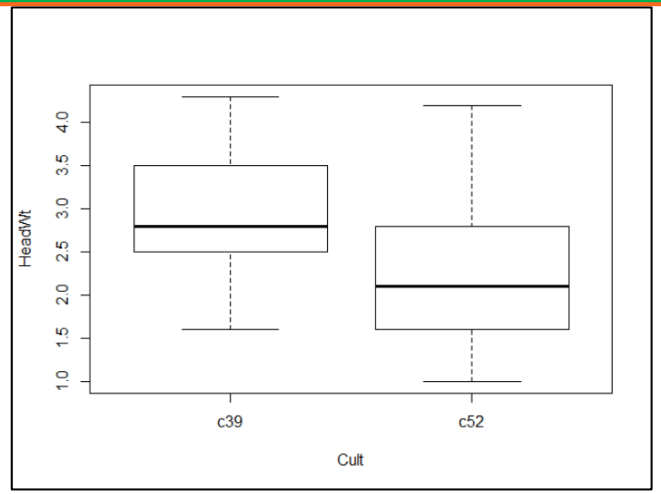
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA.

Boxplots

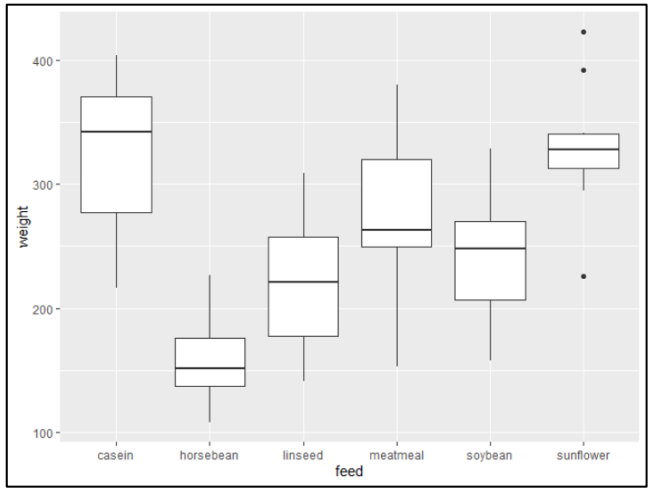
I. Basic boxplot

```
bx1 <-boxplot(data=cabbages, HeadWt~Cult)
bx1
```



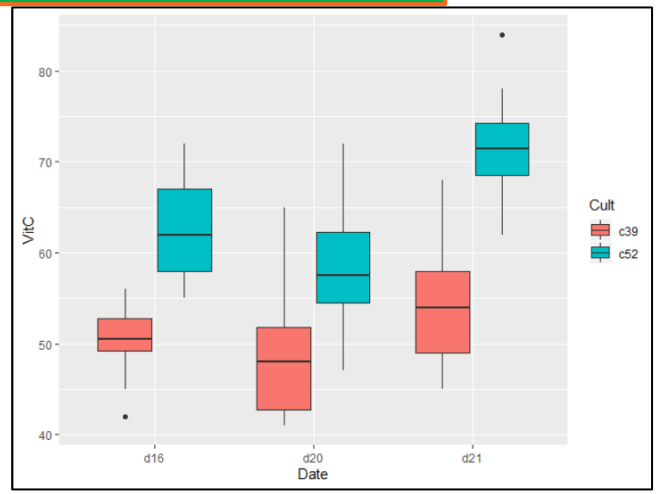
II. Simple boxplot in ggplot

```
bx2 <-ggplot(data=chickwts, aes(x=feed, y=weight)) +
  geom_boxplot()
bx2
```



III. Two-way boxplot in ggplot

```
bx3 <-ggplot(data=cabbages, aes(y=VitC, x=Date, fill=Cult)) +
  geom_boxplot()
bx3
```



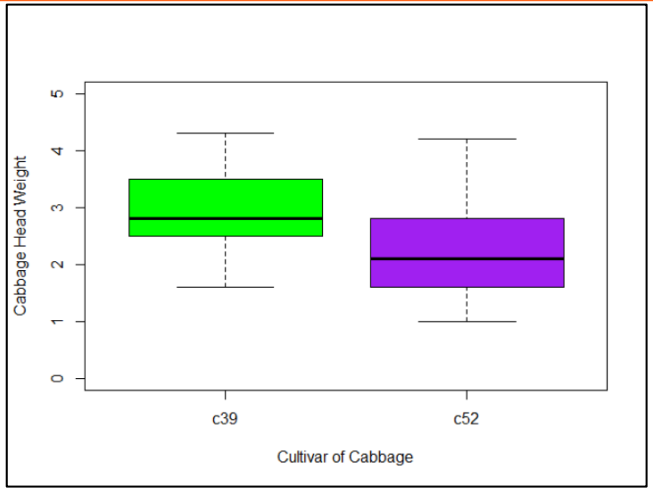
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Boxplots cont.

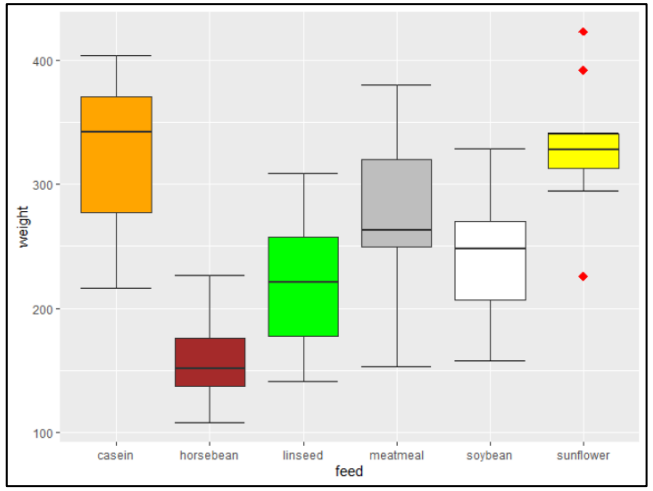
I. Basic boxplot upgrade

```
bx4 <-boxplot(data=cabbages, HeadWt~Cult, xlab="Cultivar of Cabbage",
              col=c("green","purple"),  ylab="Cabbage Head Weight",
              ylim=c(0,5))
bx4
```



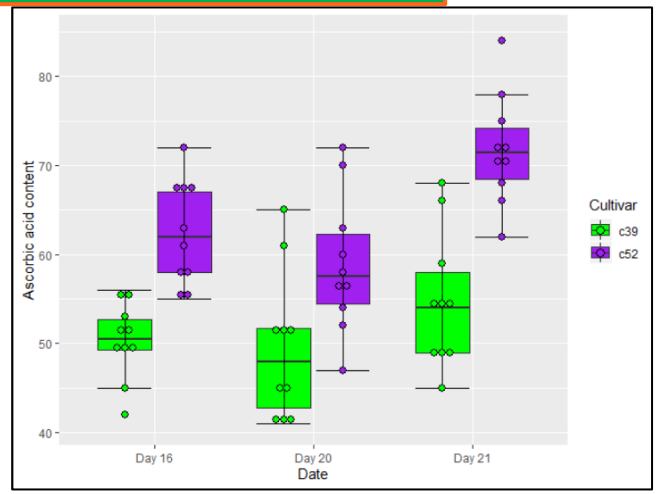
II. Simple boxplot in ggplot upgrade

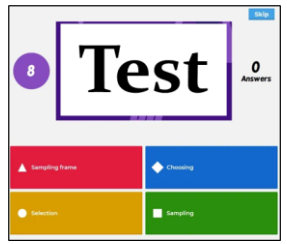
```
bx5 <-ggplot(data=chickwts, aes(x=feed, y=weight, fill=feed)) +
  stat_boxplot(geom ='errorbar') +
  geom_boxplot(outlier.colour="red", outlier.shape=18, outlier.size=3) +
  scale_fill_manual(values=c("orange","brown","green","grey","white","yellow")) +
  theme(legend.position="none")
bx5
```



III. Two-way boxplot in ggplot upgrade

```
bx6 <-ggplot(data=cabbages, aes(y=VitC, x=Date, fill=Cult)) +
  geom_boxplot() +  stat_boxplot(geom ='errorbar') +
scale_fill_manual(values=c("green", "purple"), name="Cultivar",
                  labels=c("c39","c52"))+
  labs(x="Date", y="Ascorbic acid content") +
  scale_x_discrete(labels=c("Day 16", "Day 20", "Day 21")) +
  geom_dotplot(binaxis="y", stackdir="center", dotsize=0.60,position=position_dodge(0.75))
bx6
```





https://create.kahoot.it/share/mmgg-in-r-quick-test-1/7fd94595-028c-45d1-b35e-19a4e44b5059

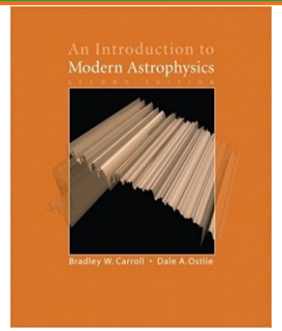
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

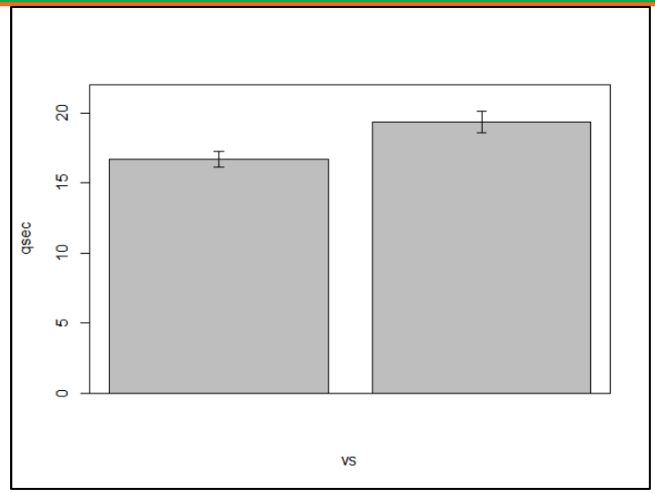
Bar plots

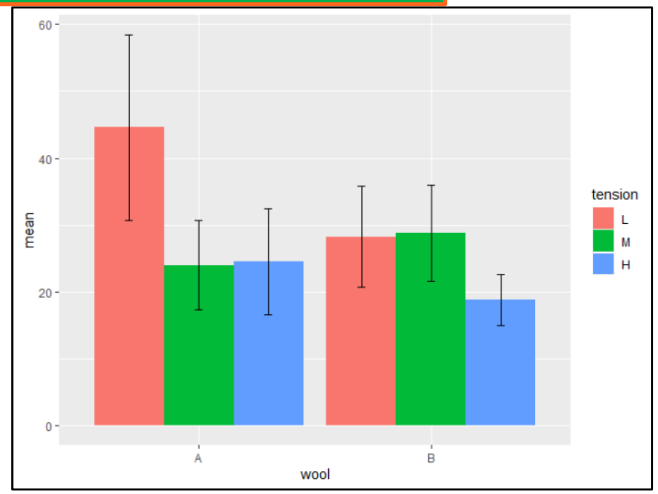
Pitfall Alerts

I. Basic bar plot




```
V <-mtcars[mtcars$vs==0,]; S <-mtcars[mtcars$vs==1,]
V.mean <-mean(V$qsec); S.mean <-mean(S$qsec)
ll<-c(ci(V$qsec)[2], ci(S$qsec)[2])
ul<-c(ci(V$qsec)[3], ci(S$qsec)[3])
x<-c(1,2); y<-c(V.mean,S.mean); x<-barplot(y)
br1<-barplot(y, xlab='vs', ylab='qsec', ylim=c(0,22))
br1 + arrows(x,ul,x,ll, length=0.05, angle=90, code=3) +box()
```

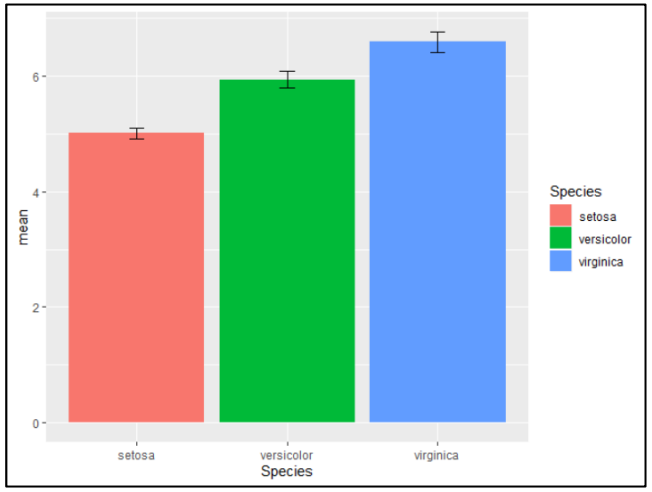




Stat Hacks

II. Simple barplot in ggplot

```
iris_sum<-iris %>% group_by(Species) %>%
  summarise(mean=mean(Sepal.Length),
            sd = sd(Sepal.Length), error = qt(0.975,df=n()-1)*sd/sqrt(n()),
            ul = mean + error, ll = mean - error)
br2 <-ggplot(data=iris_sum, aes(x=Species, y=mean, fill=Species)) +
geom_bar(stat="identity") + geom_errorbar(aes(ymin=ll, ymax=ul), width=0.1)
br2
```



III. Two-way bar plot in ggplot

```
warpbreaks_sum<-warpbreaks %>% group_by(wool,tension) %>%
  summarise(mean=mean(breaks),
            sd = sd(breaks), error = qt(0.975,df=n()-1)*sd/sqrt(n()),
            ul = mean + error, ll = mean - error)
br3 <-ggplot(data=warpbreaks_sum, aes(x=wool, y=mean)) +
  geom_bar(aes(fill=tension), stat="identity", position=position_dodge()) +
  geom_errorbar(aes(ymin=ll, ymax=ul, group=tension), width=0.1,
position=position_dodge(0.9));
br3
```

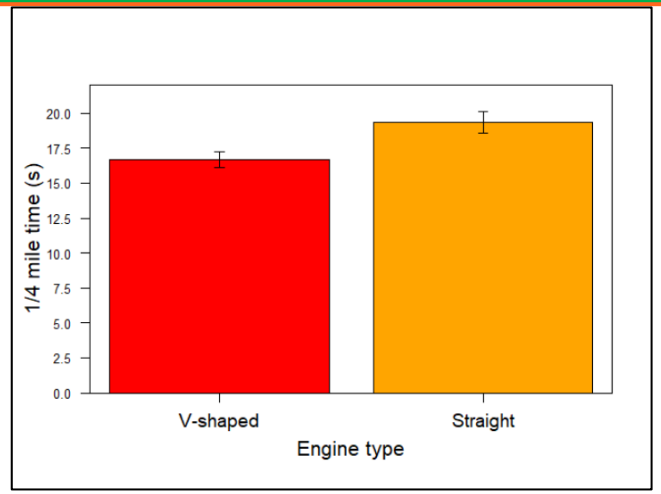
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Bar plots cont.

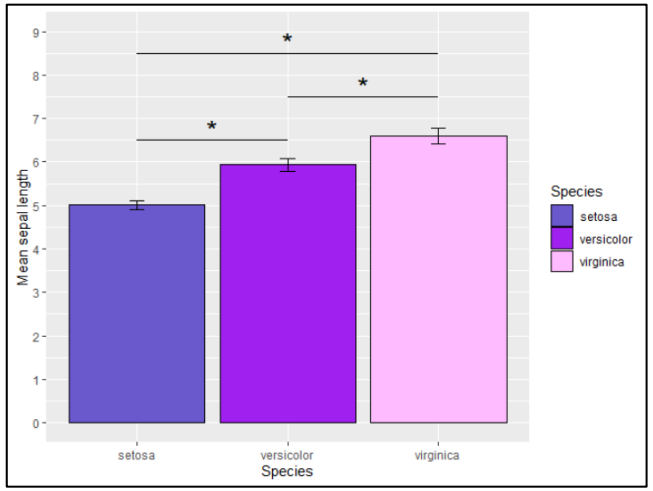
I. Basic bar plot upgrade

```
br4<-barplot(y, col=c("red","orange"), ylim=c(0,22), axes=F)
axis(side=1, at=x, labels=c("V-shaped","Straight"), cex.lab=1.5, cex.axis=1.2)
axis(2, at=c(0,2.5,5,7.5,10,12.5,15,17.5,20,22.5), las=2, cex.lab=1.5, cex.axis=0.80)
mtext(side=1, line=2.5, "Engine type", cex=1.3)
mtext(side=2, line=2.5, "1/4 mile time (s)", cex=1.3)
br4 + arrows(x,ul,x,ll, length=0.05, angle=90, code=3) + box()
```



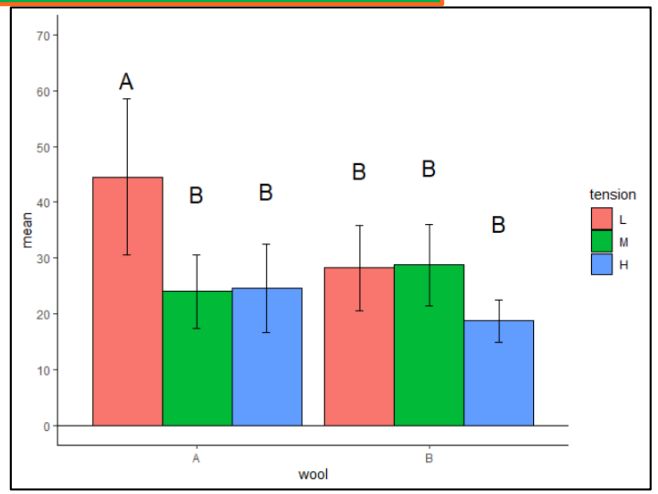
II. Simple bar plot in ggplot upgrade

```
br5 <-ggplot(data=iris_sum, aes(x=Species, y=mean, fill=Species)) +
  geom_bar(stat="identity", color="black") +
  geom_errorbar(aes(ymin=ll, ymax=ul), width=0.1) +
  labs(y="Mean sepal length", x="Species") +
  scale_y_continuous(limits=c(0,9),breaks=c(0,1,2,3,4,5,6,7,8,9)) +
  scale_fill_manual(values=c("slateblue3", "purple", "plum1")) +
  geom_segment(aes(x=1,xend=2,y=6.5,yend=6.5)) +
  geom_segment(aes(x=2,xend=3,y=7.5,yend=7.5)) +
  geom_segment(aes(x=1,xend=3,y=8.5,yend=8.5)) +
  geom_text(aes(label="*"), x=c(1.5,2,2.5), y=c(6.75,8.75,7.75), size=7)
br5
```



III. Two-way bar plot in ggplot upgrade

```
br6 <-ggplot(data=warpbreaks_sum, aes(x=wool, y=mean)) +
 geom_bar(aes(fill=tension), stat="identity", color="black", position=position_dodge()) +
 geom_errorbar(aes(ymin=ll, ymax=ul, group=tension), width=0.1,
                position=position_dodge(0.9)) +
 scale_y_continuous(limits=c(0,70),breaks=c(0,10,20,30,40,50,60,70)) +
 geom_text(aes(label=c("A","B","B","B","B","B"), group=tension), vjust=-5, size=6,
                position=position_dodge(0.9))+
 geom_hline(aes(yintercept=0))  + theme_classic()
br6
```



Animaker Summary

https://app.animaker.com/anim0/4fgb5LNoGWAsnkyy/

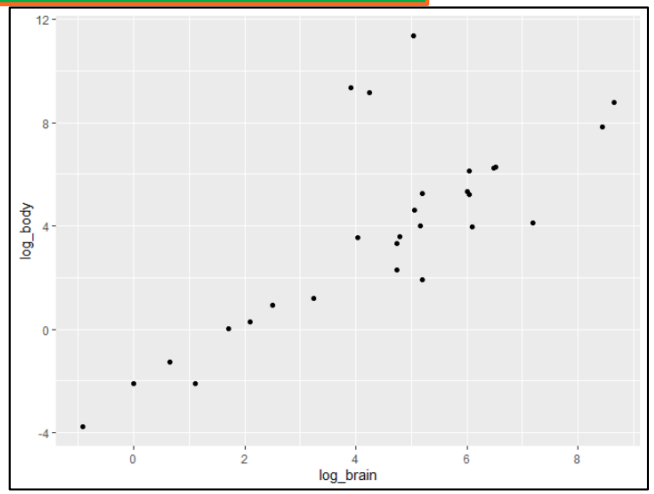
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Scatter plots

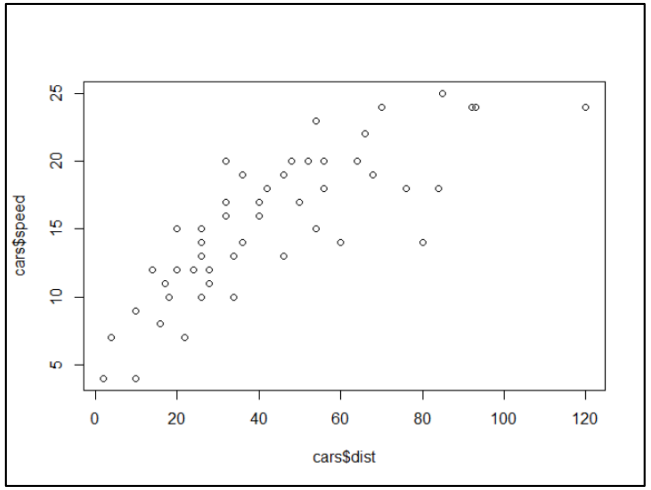
I. Basic scatter plot

```
s1 <-plot(cars$speed~cars$dist)
s1
```



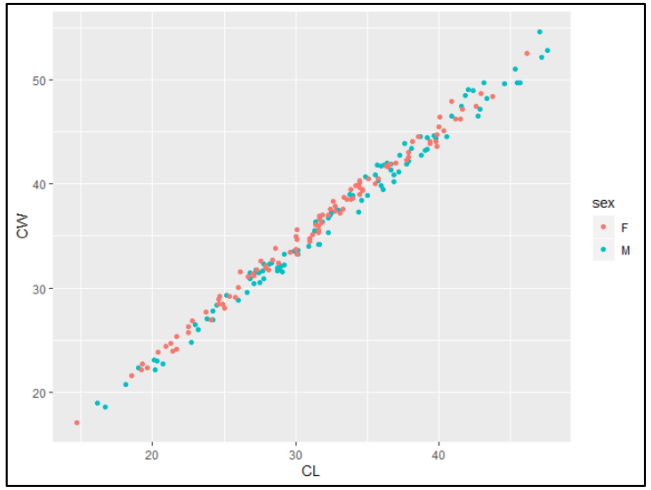
II. Simple scatter plot in ggplot

```
Animals$log_brain <-log(Animals$brain)
Animals$log_body <-log(Animals$body)
s2 <-ggplot(data=Animals, aes(x=log_brain, y=log_body)) +
  geom_point()
s2
```



III. Two-way scatter plot in ggplot

```
s3 <-ggplot(data=crabs, aes(x=CL, y=CW, fill=sex, color=sex)) +
  geom_point()
s3
```



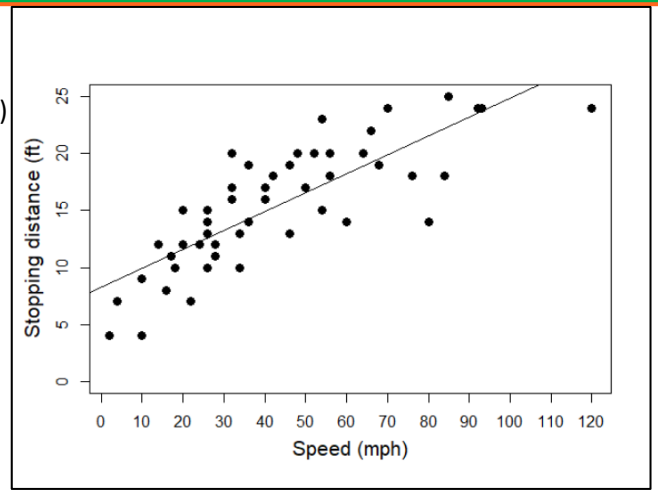
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Scatter plots cont.

UNIVERSITY OF NORTH DAKOTA

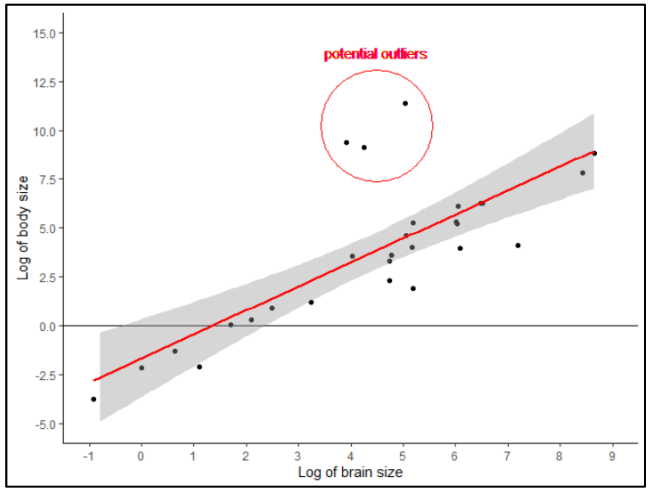
I. Basic scatter plot upgrade

```
s4 <-plot(cars$speed~cars$dist,ylim=c(0,25),axes=F, xlab='',ylab='',pch=19,col="black",cex=1.2)
axis(side=1, at=c(0,10,20,30,40,50,60,70,80,90,100,110,120))
axis(side=2, at=c(0,5,10,15,20,25))
mtext(side=1, line=2.5, "Speed (mph)", cex=1.3)
mtext(side=2, line=2.5, "Stopping distance (ft)", cex=1.3)
s4; box(); abline(lm(cars$speed~cars$dist))
```



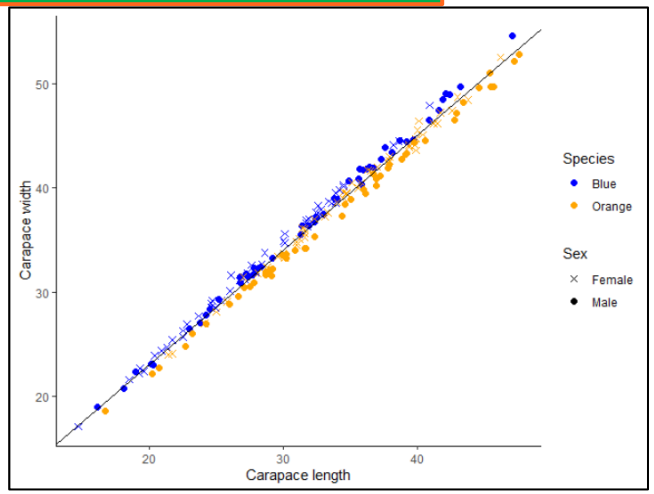
II. Simple scatter plot in ggplot upgrade

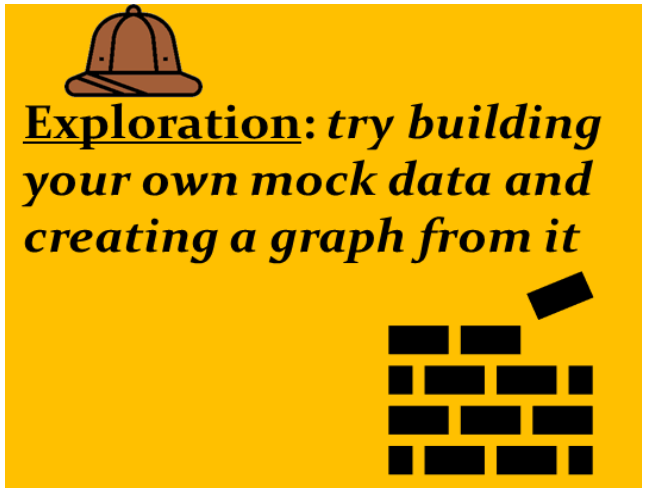
```
circ = data.frame(x=4.5, y=10.25)
s5 <-ggplot(data=Animals, aes(x=log_brain, y=log_body)) +
  geom_point() + geom_hline(aes(yintercept=0)) +
  geom_smooth(method=lm, color="red") +
  theme_classic() + labs(y="Log of body size", x="Log of brain size") +
  scale_y_continuous(limits=c(-5,15),breaks=c(-5,-2.5,0,2.5,5,7.5,10,12.5,15)) +
  scale_x_continuous(limits=c(-1,9),breaks=c(-1,0,1,2,3,4,5,6,7,8,9)) +
  geom_point(aes(x=x, y=y), data=circ, size=40, shape=1, color="red") +
  geom_text(aes(label="potential outliers"), x=c(4.5), y=c(14), size=4, color="red")
s5
```



III. Two-way scatter plot in ggplot upgrade

```
summary(lm(CW~CL, data=crabs))
s6 <-ggplot(data=crabs, aes(x=CL, y=CW, color=sp, shape=sex)) +
  geom_point(size=2) +
  labs(x="Carapace length", y="Carapace width", color="Species",  shape="Sex") +
  scale_color_manual(labels=c("Blue","Orange"), values=c("blue","orange")) +
  scale_shape_manual(labels=c("Female","Male"), values=c(4,16)) +
  geom_abline(aes(intercept=1.09, slope=1.1)) +
  theme_classic() + guides(fill="none")
s6
```





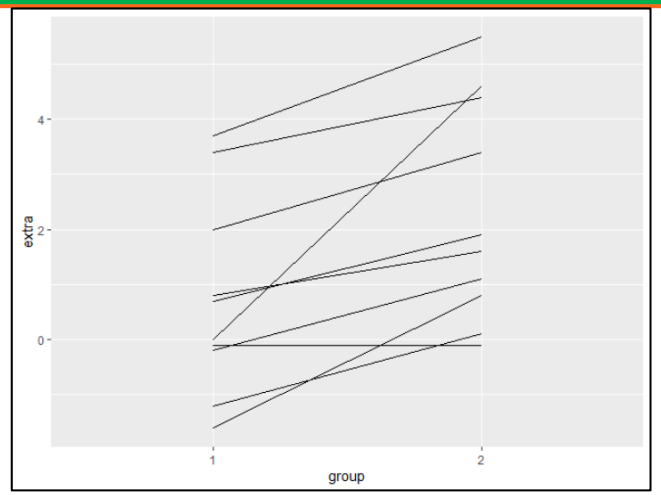
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Other plots

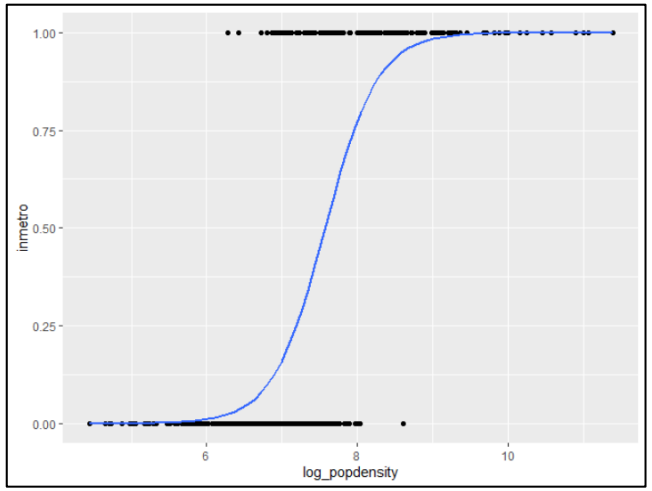
I. Spaghetti plot in ggplot

```
o1 <-ggplot(data=sleep, aes(x=group, y=extra, group=ID)) +
  geom_line()
o1
```



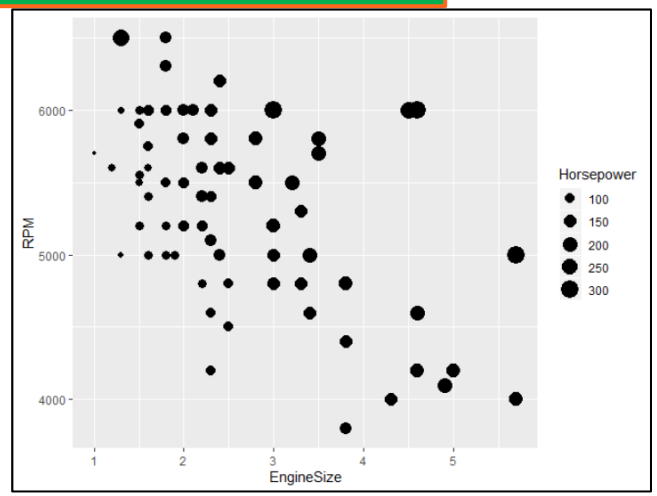
II. Logistic regression plot in ggplot

```
midwest$log_popdensity <-log(midwest$popdensity)
o2 <-ggplot(data=midwest, aes(x=log_popdensity, y=inmetro)) +
  geom_point() +
  stat_smooth(method="glm", se=FALSE, method.args=list(family=binomial))
o2
```



III. Bubble plot in ggplot

```
o3 <-ggplot(data=Cars93, aes(x=EngineSize, y=RPM, size=Horsepower))+
  geom_point()
o3
```



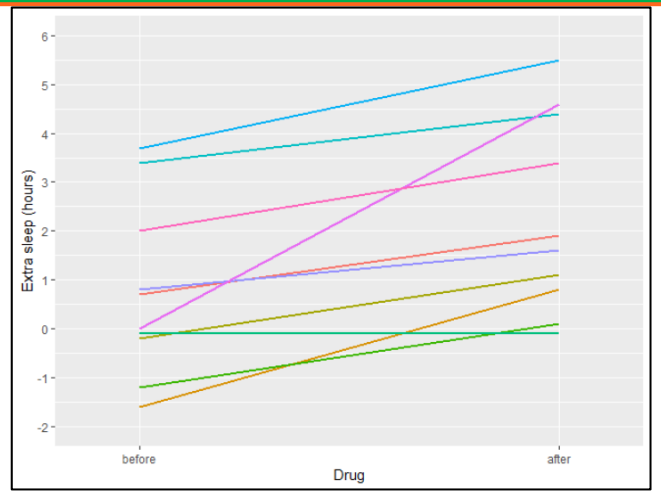
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Other plots cont.

UNIVERSITY OF NORTH DAKOTA

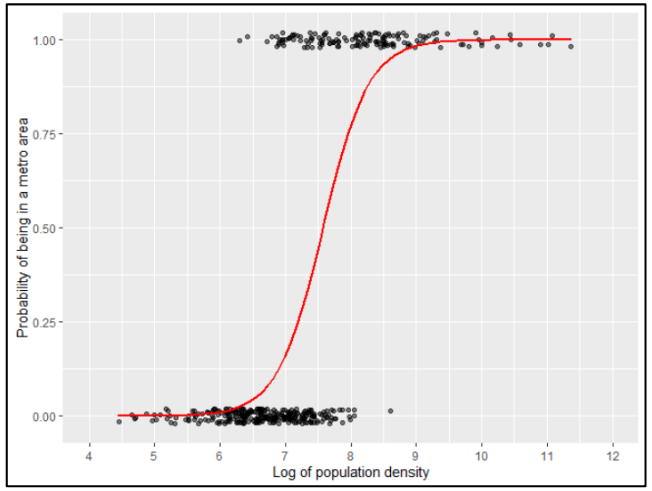
I. Spaghetti plot in ggplot upgrade

```
o4 <-ggplot(data=sleep, aes(x=group, y=extra, group=ID, colour=ID)) +
  geom_line(size=1) + theme(legend.position="none") +
  labs(y="Extra sleep (hours)", x="Drug") +
  scale_y_continuous(limits=c(-2,6),breaks=c(-2,-1,0,1,2,3,4,5,6)) +
  scale_x_discrete(limits=c(1,2), labels=c("before","after"), expand=c(0.1,0.1))
o4
```



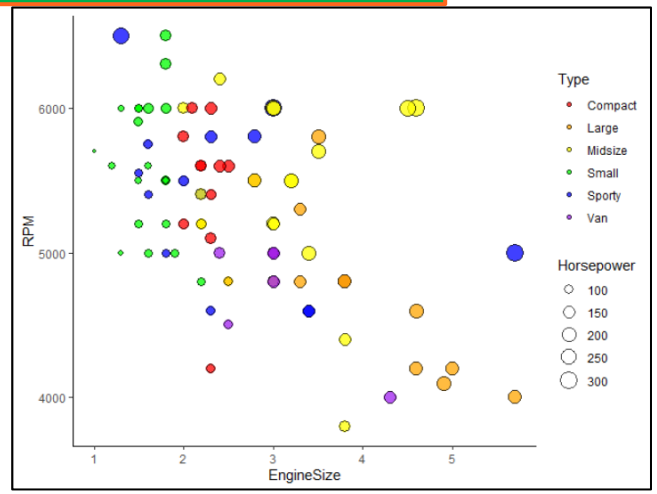
II. Logistic regression plot in ggplot upgrade

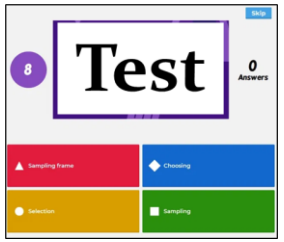
```
o5 <-ggplot(data=midwest, aes(x=log_popdensity, y=inmetro)) +
  geom_point(alpha=0.5, position=position_jitter(width=.02,height=.02)) +
  stat_smooth(method="glm", se=FALSE, method.args=list(family=binomial),
              color="red", size=1)+
  labs(x="Log of population density", y="Probability of being in a metro area")+
  scale_x_continuous(limits=c(4,12), breaks=c(4,5,6,7,8,9,10,11,12))
o5
```



III. Bubble plot in ggplot upgrade

```
o6 <-ggplot(data=Cars93, aes(x=EngineSize, y=RPM, size=Horsepower, fill=Type))+
  geom_point(alpha=0.75, shape=21) +
  scale_fill_manual(values=c("red","orange","yellow","green","blue","purple"))+
  theme_classic()
o6
```





https://create.kahoot.it/share/mmgg-in-r-quick-test-2/0c94e59e-5d0e-4232-98fd-be42f3edbca7

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

Closing

- Please try out the post-test and survey

https://und.qualtrics.com/jfe/form/SV_9AELZY4MWSJksrs

https://und.qualtrics.com/jfe/form/SV_6SfLoElRrjKzUHk

- **Special Treat**: Example R-code contains functions that allow you to prebuild graphs!
  - barchart.2sample -> bar chart for t-test
  - barchart.anova -> bar chart for 1-way ANOVA
  - barchart.2anova -> bar chart for 2-way ANOVA
- You can find the function code and examples at the bottom of the R-code

Animaker Summary

https://app.animaker.com/animo/UnkaurjzkZTVXQuj/

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

References

UNIVERSITY OF NORTH DAKOTA

- ✓ https://www.datacamp.com/community/tutorials/make-histogram-ggplot2
- ✓ https://www.r-graph-gallery.com/histogram_several_group.html
- ✓ https://homepage.divms.uiowa.edu/~luke/classes/STAT4580/histdens.html
- ✓ http://www.sthda.com/english/wiki/ggplot2-box-plot-quick-start-guide-r-software-and-data-visualization
- ✓ http://www.sthda.com/english/wiki/ggplot2-error-bars-quick-start-guide-r-software-and-data-visualization
- ✓ http://sape.inf.usi.ch/quick-reference/ggplot2/shape
- ✓ http://www.sthda.com/english/wiki/ggplot2-dot-plot-quick-start-guide-r-software-and-data-visualization
- ✓ https://www.r-graph-gallery.com/4-barplot-with-error-bar.html
- ✓ https://www.statmethods.net/advgraphs/axes.html
- ✓ http://www.sthda.com/english/wiki/ggplot2-scatter-plots-quick-start-guide-r-software-and-data-visualization
- ✓ http://www.sthda.com/english/wiki/r-plot-pch-symbols-the-different-point-shapes-available-in-r
- ✓ https://stackoverflow.com/questions/6862742/draw-a-circle-with-ggplot2
- ✓ https://stats.idre.ucla.edu/r/faq/how-can-i-visualize-longitudinal-data-in-ggplot2/
- ✓ https://mgimond.github.io/Stats-in-R/Logistic.html
- ✓ http://www.cookbook-r.com/Statistical_analysis/Logistic_regression/

DaCCoTA
DAKOTA CANCER COLLABORATIVE
ON TRANSLATIONAL ACTIVITY

Acknowledgements

- The DaCCoTA is supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729.
- For the labs that use the Biostatistics, Epidemiology, and Research Design Core in any way, including this Module, please acknowledge us for publications. ***"Research reported in this publication was supported by DaCCoTA (the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729)".***

DaCCoTA
DAKOTA CANCER COLLABORATIVE
ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.