Exploratory Data Analysis Module III: Deep Dive

Dr. Mark Williamson

DaCCoTA

University of North Dakota

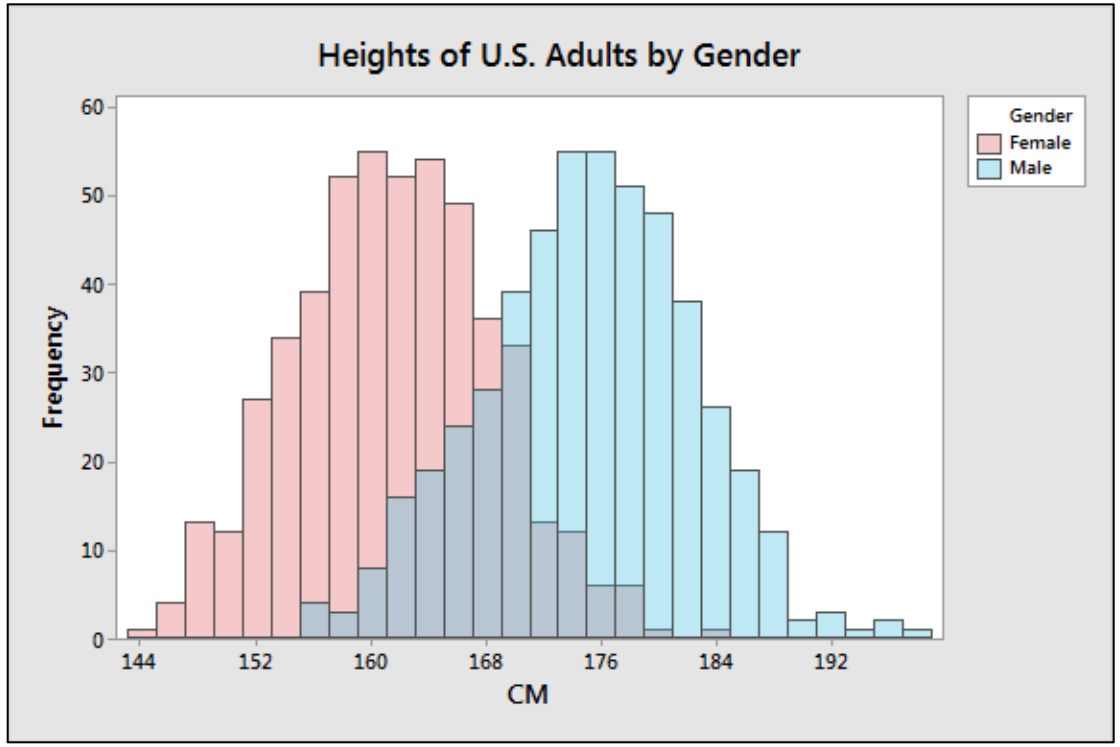
Introduction

Previously:

- Covered a broad overview
- Looked at more detail
- Ran through examples

This time: look at more advanced techniques of exploratory data analysis

- Visualizing more dimensions
- Model selection
- Complex plots



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND UNIVERSITY OF NORTH DAKOTA

Reviewing the Basics

Exploratory Data Analysis

Obs	Species	Weight	Length1	Length2	Length3	Height	Width
1	Bream	242.0	23.2	25.4	30.0	11.5200	4.0200
2	Bream	290.0	24.0	26.3	31.2	12.4800	4.3056
3	Bream	340.0	23.9	26.5	31.1	12.3778	4.6961
4	Bream	363.0	26.3	29.0	33.5	12.7300	4.4555
5	Bream	430.0	26.5	29.0	34.0	12.4440	5.1340
6	Bream	450.0	26.8	29.7	34.7	13.6024	4.9274
7	Bream	500.0	26.8	29.7	34.5	14.1795	5.2785
8	Bream	390.0	27.6	30.0	35.0	12.6700	4.6900
9	Bream	450.0	27.6	30.0	35.1	14.0049	4.8438
10	Bream	500.0	28.5	30.7	36.2	14.2266	4.9594
11	Bream	475.0	28.4	31.0	36.2	14.2628	5.1042
12	Bream	500.0	28.7	31.0	36.2	14.3714	4.8146
13	Bream	500.0	29.1	31.5	36.4	13.7592	4.3680
14	Bream	.	29.5	32.0	37.3	13.9129	5.0728
15	Bream	600.0	29.4	32.0	37.2	14.9544	5.1708
16	Bream	600.0	29.4	32.0	37.2	15.4380	5.5800
17	Bream	700.0	30.4	33.0	38.3	14.8604	5.2854
18	Bream	700.0	30.4	33.0	38.5	14.9380	5.1975
19	Bream	610.0	30.9	33.5	38.6	15.6330	5.1338
20	Bream	650.0	31.0	33.5	38.7	14.4738	5.7276
...							

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	215.9175870	35.9862645	23.47	<.0001
Error	152	233.1080937	1.5336059		
Corrected Total	158	449.0256807			

Levene's Test for Homogeneity of Width Variance
ANOVA of Absolute Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Species	6	38.6585	6.4431	17.04	<.0001
Error	152	57.4674	0.3781		

View data

Summary statistics

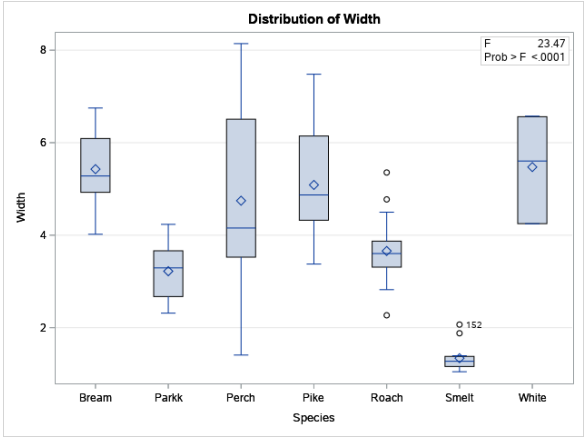
Basic Graphs

Basic Tests

Basic Statistical Measures

Location		Variability	
Mean	4.417486	Std Deviation	1.68580
Median	4.248500	Variance	2.84193
Mode	3.525000	Range	7.09440
		Interquartile Range	2.21340

Species	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Bream	35	22.01	35	22.01
Parkki	11	6.92	46	28.93
Perch	56	35.22	102	64.15
Pike	17	10.69	119	74.84
Roach	20	12.58	139	87.42
Smelt	14	8.81	153	96.23
Whitefish	6	3.77	159	100.00



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

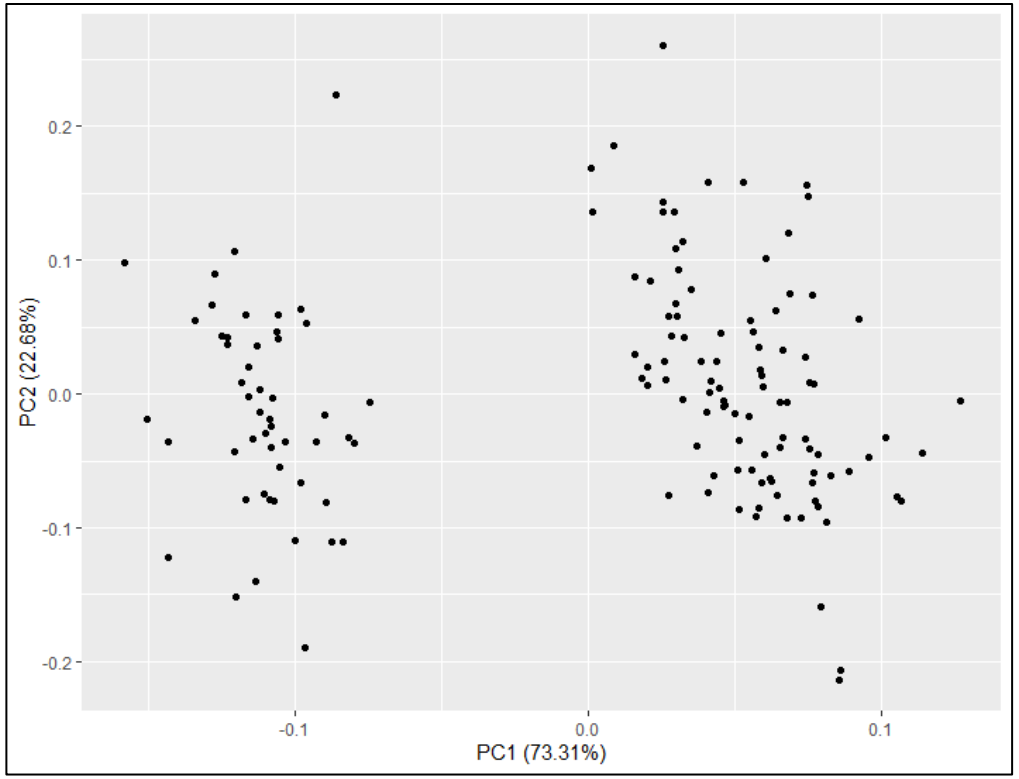
UND
UNIVERSITY OF NORTH DAKOTA

Topics Covered

- Visualizing more dimensions
  - Bubble graph
  - 3D scatterplot
  - Principle Components Analysis (PCA)
- Variable selection
  - Model selection
- Complex plots
  - Scatterplot Matrix
  - Multiple bar plots
  - Scatterplot with factors

Visualizing More Dimensions

- Three numerical variables on one graph
  - Bubble graph
    - Typical X-Y scatterplot of first two variables
    - Third variables is scaled by size of the point (bubble)
  - 3D scatterplot
    - Scatter plot runs in 3 dimensions
    - X, Y, and Z
- Many numerical variables on one graph
  - Principle Components Analysis (PCA)



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND UNIVERSITY OF NORTH DAKOTA

Visualizing More Dimensions

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Exploratory Data Analysis III examples.R

```r
library(rgl)
##########################################################################################


#Bubble Plots

#iris
head(iris)

ggplot(data=iris, aes(x=Sepal.Length, y=Sepal.Width, size=Petal.Length))+
  geom_point(alpha=0.7)
ggplot(data=iris, aes(x=Sepal.Length, y=Sepal.Width, size=Petal.Length, color=Species))+
  geom_point(alpha=0.7)

#insular
insular<-read.csv("insular.csv")
attach(insular)
print(insular)

ggplot(data=insular, aes(x=Area ,y=elevation, size=SpeciesRichness)) +
  geom_point(alpha=0.7)
##########################################################################################


#3D Scatterplots

#iris
par(mar=c(0,0,0,0))
plot3d(
  x=iris$Sepal.Length, y=iris$Sepal.Width, z=iris$Petal.Length)
```

```
R version 3.6.2 (2019-12-12) -- "Dark and Stormy Night"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[workspace loaded from ~/.RData]

>
```

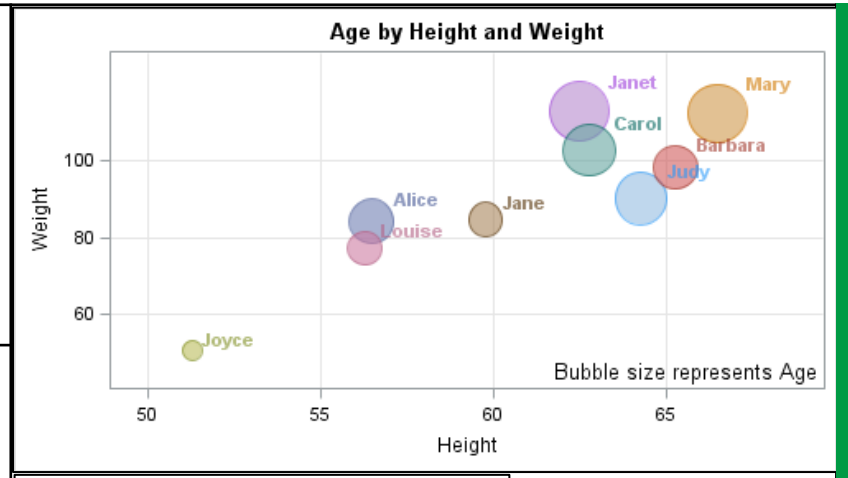
Environment is empty

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

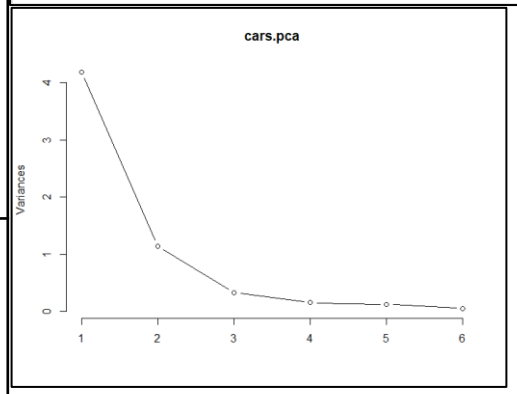
UND UNIVERSITY OF NORTH DAKOTA

Assessment 1

1) *What does this picture tell us about the relationship between height, weight and age?*



2) *Based on the PCA summary of 6 car variables and the plot, how many components capture the majority of variance?*



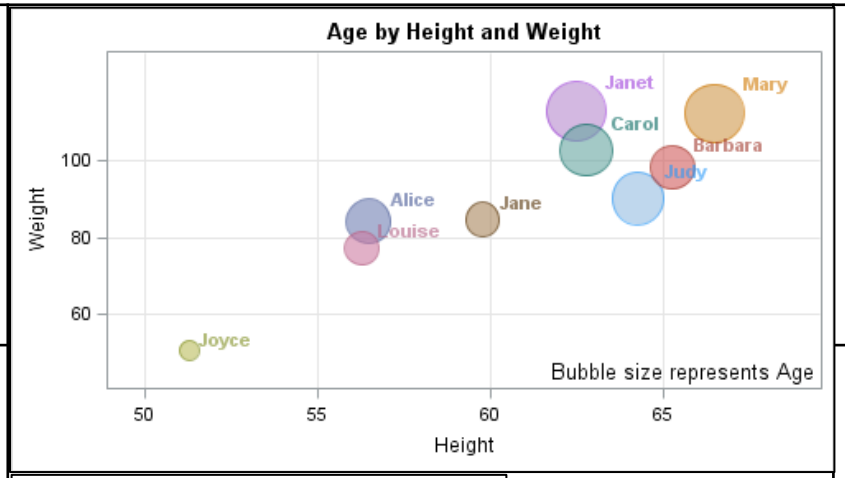
Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	2.0463	1.0715	0.57737	0.39289	0.3533	0.22799
Proportion of Variance	0.6979	0.1913	0.05556	0.02573	0.0208	0.00866
Cumulative Proportion	0.6979	0.8892	0.94481	0.97054	0.9913	1.00000

3) *Which of the following variables could be colored in a 3D plot: height, weight, age, college major*

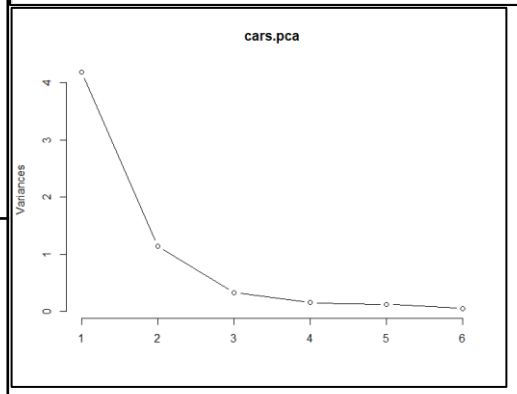
Assessment 1

1) *What does this picture tell us about the relationship between height, weight and age?*



As one variable increases (such as age), the other two tend to do so as well.

2) *Based on the PCA summary of 6 car variables and the plot, how many components capture the majority of variance?*



Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	2.0463	1.0715	0.57737	0.39289	0.3533	0.22799
Proportion of Variance	0.6979	0.1913	0.05556	0.02573	0.0208	0.00866
Cumulative Proportion	0.6979	0.8892	0.94481	0.97054	0.9913	1.00000

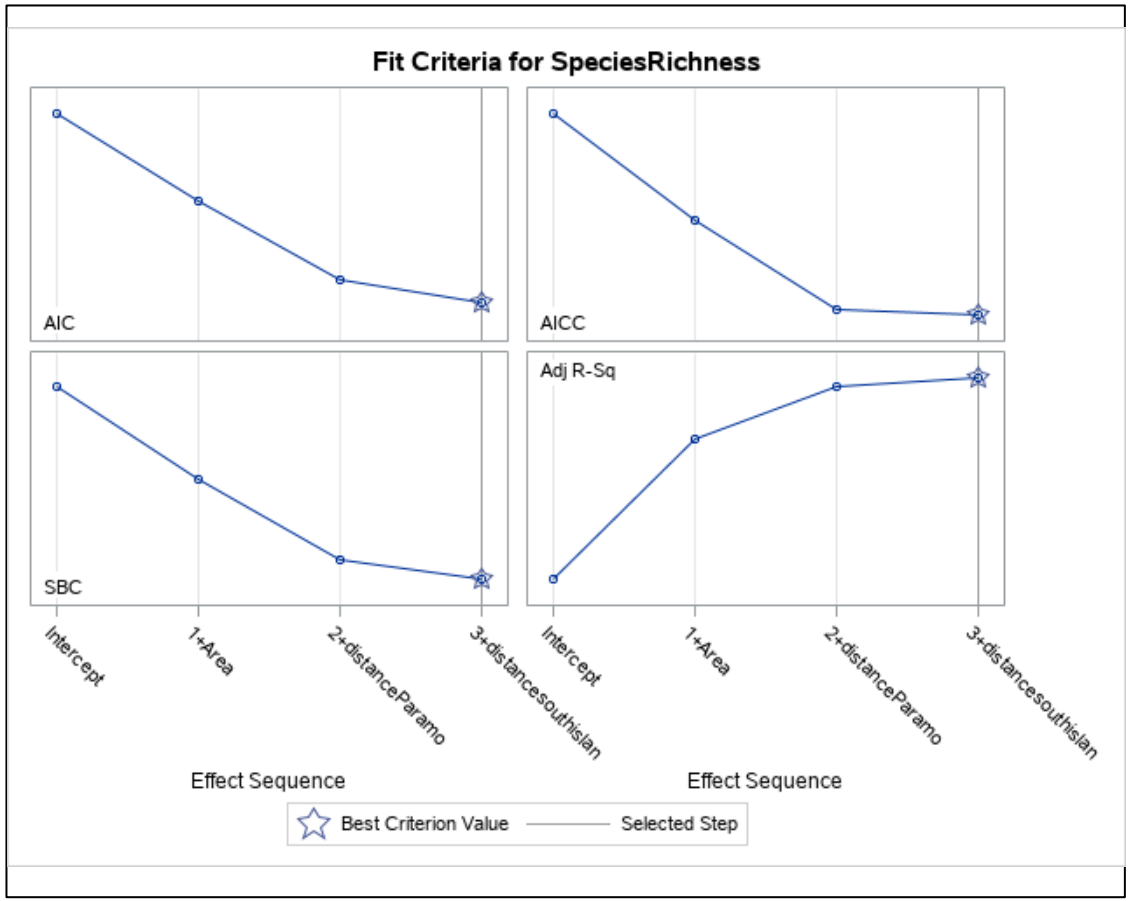
2 (together, they cover almost 90%)

3) *Which of the following variables could be colored in a 3D plot: height, weight, age, college major*

College major

Variable selection

- Model Selection
  - Lots of predictor variables
  - Need to trim down to only ones that are important
  - Can’t try every combination if there are lots of variables
  - Forwards, backwards, and stepwise selection



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND UNIVERSITY OF NORTH DAKOTA

Variable selection

SAS Studio

odamid-usw2.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT-05%253A00&ticket=ST-144801-oPFowoJ4YpQ4TU4ZNOxU-cas

SAS® Studio | SAS Programmer | Sign Out

Server Files and Folders

- odaws03-usw2
  - Folder Shortcuts
  - Files (Home)
    - my_courses
      - briandarby0
        - biometrydatasets
          - 2019exampractice.csv
          - amphipods.csv
          - anhydrobiosis.csv
          - antnests.csv
          - Ashton.csv
          - barley142.csv
          - beeduration.csv
          - bots1.csv
          - cad.csv
          - calibration.csv
          - cobralily.csv
          - corticosterone.csv
          - crickets.csv
          - eggpore.csv
          - exam1.csv
          - exam2.csv
          - exam3.csv
          - exam4.csv
          - FDA.csv
          - fish.csv
          - galapagos.csv
          - helmet.csv
          - honeydew.csv
          - insular.csv
- Tasks and Utilities
- Snippets
- Libraries
- File Shortcuts

Exploratory Data Analysis Module 3 examples.sas | exam1.csv | exam2.csv

CODE | LOG | RESULTS

```
*Model Selection;

*Iris Data;
PROC PRINT data=sashelp.iris;

PROC GLM data=sashelp.iris;
    model PetalWidth=SepalLength;
PROC GLM data=sashelp.iris;
    model PetalWidth=SepalWidth;
PROC GLM data=sashelp.iris;
    model PetalWidth=SepalLength*SepalWidth;
PROC GLM data=sashelp.iris;
    model PetalWidth=SepalLength SepalWidth;
PROC GLM data=sashelp.iris;
    model PetalWidth=SepalLength SepalLength*SepalWidth;
PROC GLM data=sashelp.iris;
    model PetalWidth=SepalWidth SepalLength*SepalWidth;
PROC GLM data=sashelp.iris;
    model PetalWidth=SepalLength SepalWidth SepalLength*SepalWidth;

PROC GLMSELECT data=sashelp.iris plots=(Criteria Candidates);
    model PetalWidth=SepalLength|SepalWidth/
        selection=stepwise(select=AICC);

*Insular Data;
PROC IMPORT datafile='/home/markwilliamson20/my_courses/markwilliamson0/MW_Datasets_2020/insular.csv'
    dbms=csv out=insular replace; getnames=yes;
PROC PRINT data=insular;

PROC GLMSELECT data=insular plots=(Criteria Candidates);
    model SpeciesRichness=endemic percentendemic Area   altitude elevation
        distanceParamo distancevegisland distancesouthisland distancelargeisland/
        selection=forward(select=AICC);

PROC GLMSELECT data=insular plots=(Criteria Candidates);
    model SpeciesRichness=endemic percentendemic Area   altitude elevation
        distanceParamo distancevegisland distancesouthisland distancelargeisland/
        selection=backward(select=AICC);

PROC GLMSELECT data=insular plots=(Criteria Candidates);
    model SpeciesRichness=endemic percentendemic Area   altitude elevation
```

/home/markwilliamson20/my_courses/markwilliamson0/MW_2020_Work/Module Examples/Exploratory Data Analysis Module 3 examples.sas | Line 110, Column 1 | UTF-8

Messages: 1 | User: markwilliamson20

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND UNIVERSITY OF NORTH DAKOTA

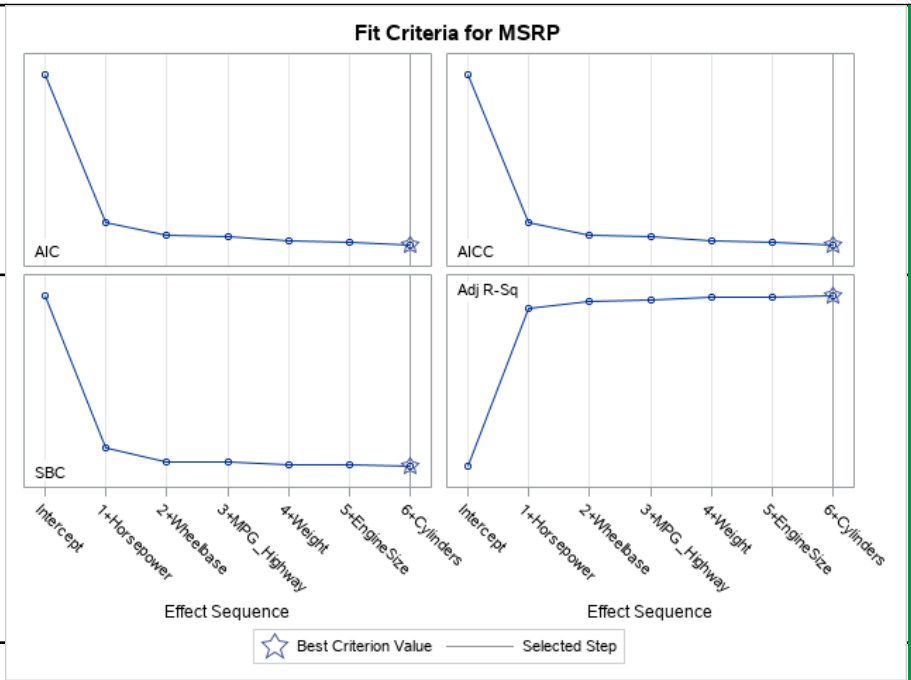
Assessment 2

1) *Based on the graph to the right, which model has the best fit for MSRP?*

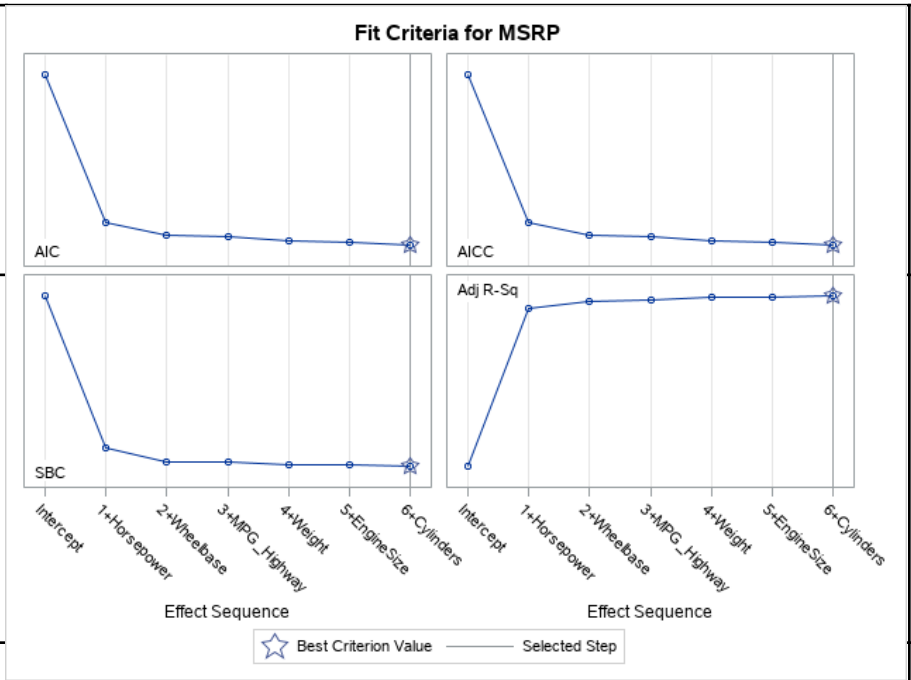
2) *How will the <u>stepwise</u> selection method select a model?*

a) Start with the null model and add until best fit
b) Start with the full model and subtract until best fit
c) Start with the null model and add or subtract until best fit
d) Start with the null model and subtract or add until best fit

3) *Should you use the R-squared to compared models?*

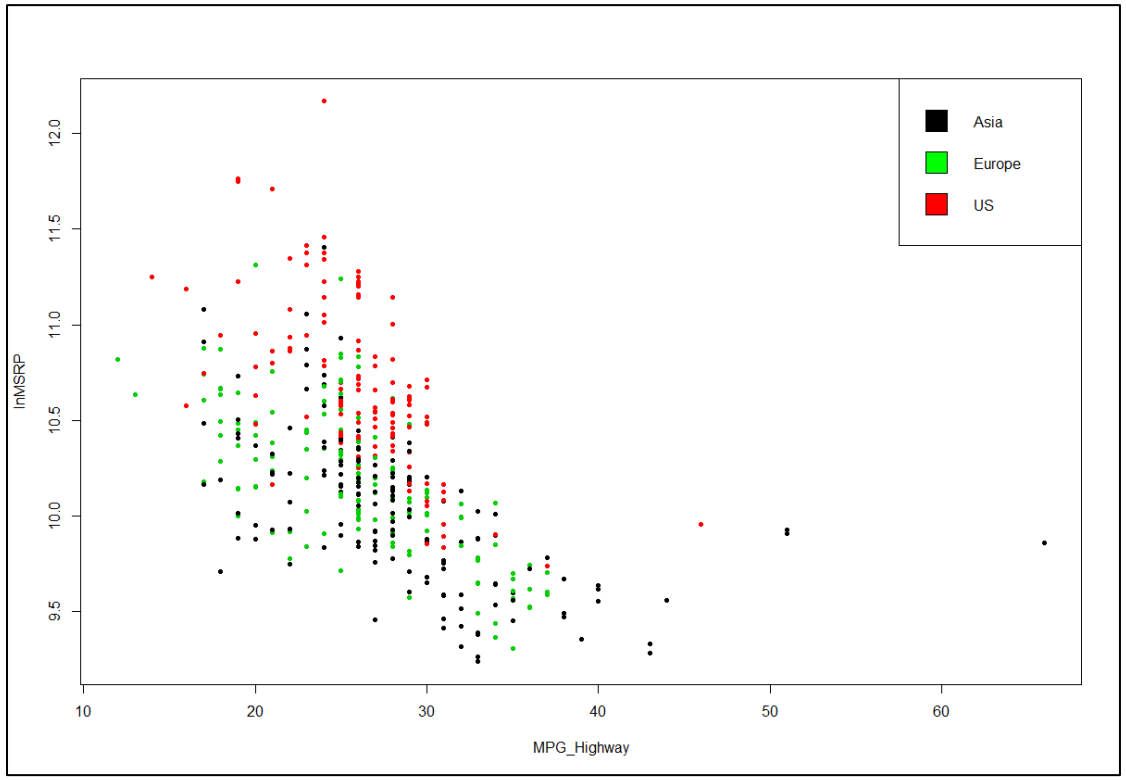


Assessment 2

1) *Based on the graph to the right, which model has the best fit for MSRP?*		Model with 6 variables (up to Cylinders)
2) *How will the <u>stepwise</u> selection method select a model?* *a) Start with the null model and add until best fit* *b) Start with the full model and subtract until best fit* *c) Start with the null model and add or subtract until best fit* *d) Start with the null model and subtract or add until best fit*		C) Start with null and add/subtract
3) *Should you use the R-squared to compared models?*		Only for single variable models, otherwise it is biased by adding more variables

Complex Plots

- Scatterplot Matrix
  - Compare many numerical variables at once
- Multiple bar plots
  - Compare numerical variable across multiple categorical variables
- Scatterplot with factors
  - Compare two numerical variables across categories



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND UNIVERSITY OF NORTH DAKOTA.

Complex Plots

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Exploratory Data Analysis III examples.R

```
#Scatterplot matrices:

# Basic Scatterplot Matrix
pairs(~Sepal.Length+Sepal.Width+Petal.Length+Petal.Width, data=iris)
pairs(~SpeciesRichness+endemic+Area+altitude+elevation+distanceParamo, data=insular)

#Multiplots:
Scars <-read.csv("SAS cars.csv")
head(Scars)
attach(Scars)

boxplot(MSRP)
boxplot(MSRP~DriveTrain)
boxplot(MSRP~DriveTrain*Origin,
  col=c('blue','red','green'))
boxplot(MSRP~DriveTrain*Origin*Type,
  col=c('blue','red','green'), cex.axis=0.5)

par(mfrow=c(3,2))
boxplot(MSRP[Type=='Hybrid']~DriveTrain[Type=='Hybrid']*Origin[Type=='Hybrid'], main='Hybrid',cex.axis=0.75)
boxplot(MSRP[Type=='SUV']~DriveTrain[Type=='SUV']*Origin[Type=='SUV'], main='SUV',cex.axis=0.75)
boxplot(MSRP[Type=='Sedan']~DriveTrain[Type=='Sedan']*Origin[Type=='Sedan'], main='Sedan',cex.axis=0.75)
boxplot(MSRP[Type=='Sports']~DriveTrain[Type=='Sports']*Origin[Type=='Sports'], main='Sports',cex.axis=0.75)
boxplot(MSRP[Type=='Truck']~DriveTrain[Type=='Truck']*Origin[Type=='Truck'], main='Truck',cex.axis=0.75)
boxplot(MSRP[Type=='Wagon']~DriveTrain[Type=='Wagon']*Origin[Type=='Wagon'], main='Wagon',cex.axis=0.75)
par(mfrow=c(1,1))

#Monstrosity
boxplot(MSRP~DriveTrain*Origin*Type*as.factor(Cylinders))

#Scatter Plot Colored by categories
lnMSRP <-log(MSRP)
plot(MPG_Highway, lnMSRP, pch=16, cex=0.70)
plot(MPG_Highway, lnMSRP, pch=16, cex=0.70, col=Origin)
  legend("topright", legend=c("Asia","Europe","US"), fill=c('black','green','red'))
```

Console Terminal Jobs

C:/Users/Mark.Williamson.2/Desktop/Willliamson Presentations/Statistical Modules/

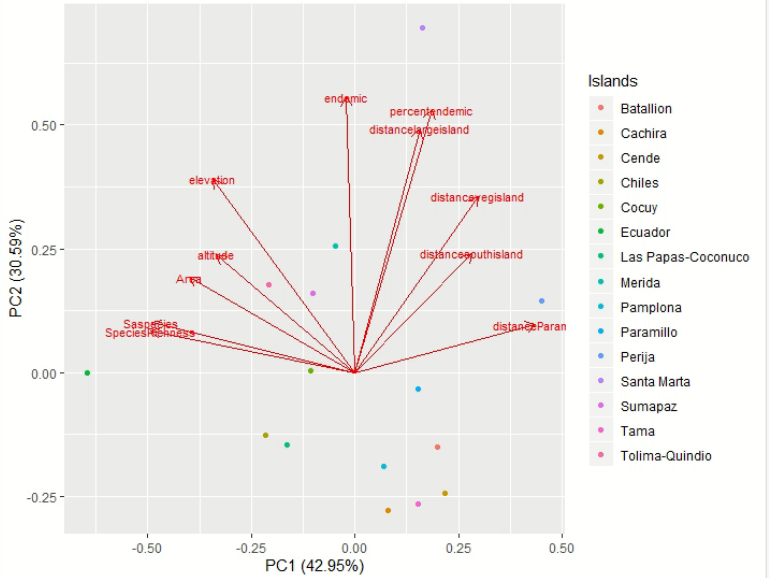
```
> boxplot(MSRP[Type=='Sports']~DriveTrain[Type=='Sports']*Origin[Type=='Sports'], main='Sports',cex.axis=0.75)
> boxplot(MSRP[Type=='Truck']~DriveTrain[Type=='Truck']*Origin[Type=='Truck'], main='Truck',cex.axis=0.75)
> boxplot(MSRP[Type=='Wagon']~DriveTrain[Type=='Wagon']*Origin[Type=='Wagon'], main='Wagon',cex.axis=0.75)
> par(mfrow=c(1,1))
> #Monstrosity
> boxplot(MSRP~DriveTrain*Origin*Type*as.factor(Cylinders))
> #Scatter Plot Colored by categories
> lnMSRP <-log(MSRP)
> plot(MPG_Highway, lnMSRP, pch=16, cex=0.70)
> plot(MPG_Highway, lnMSRP, pch=16, cex=0.70, col=Origin)
>   legend("topright", legend=c("Asia","Europe","US"), fill=c('black','green','red')))
Error: unexpected ')' in "  legend("topright", legend=c("Asia","Europe","US"), fill=c('black','green','red')))"
> autoplot(in.pca, data=insular, colour='Islands',
+          loadings=TRUE, loadings.label=TRUE, loadings.label.size=3)
> pairs(~Sepal.Length+Sepal.Width+Petal.Length+Petal.Width, data=iris)
> pairs(~SpeciesRichness+endemic+Area+altitude+elevation+distanceParamo, data=insular)
> autoplot(in.pca, data=insular, colour='Islands',
+          loadings=TRUE, loadings.label=TRUE, loadings.label.size=3)
>
```

Environment History Connections

Global Environment

Data	
in.pca	List of 5
insular	15 obs. of 12 variables
ir.pca	List of 5
iris	150 obs. of 6 variables
log.ir	150 obs. of 4 variables
Scars	428 obs. of 14 variables
Values	
in.islands	Factor w/ 15 levels "Batallion","Cachira",..: 6 4 7 13 15 10…
ir.species	Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1…
lnMSRP	num [1:428] 10.5 10.1 10.2 10.4 10.7 ...
mycolors	chr [1:3] "royalblue1" "darkcyan" "oldlace"

Files Plots Packages Help Viewer

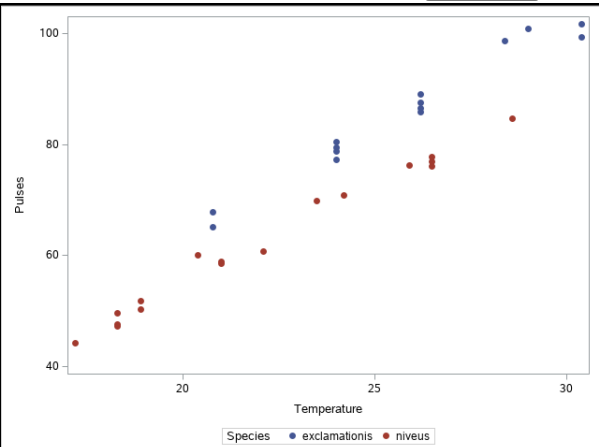


DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

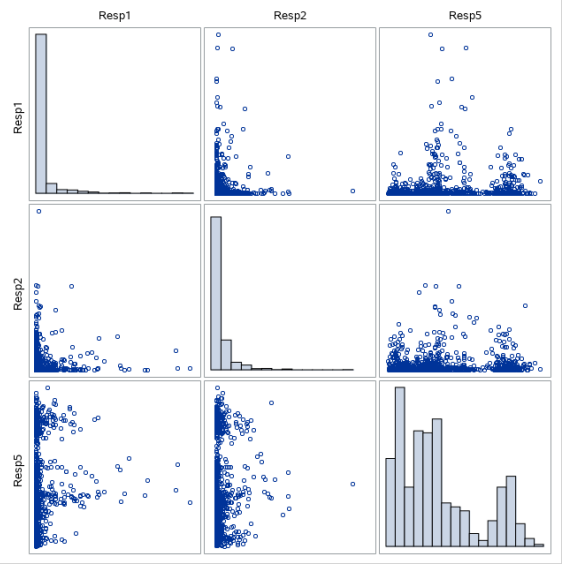
UND UNIVERSITY OF NORTH DAKOTA

Assessment 3

1) *The scatterplot to the right displays the number of pulses per hour against the outside temperature for crickets. What does the graph tell out about the two species?*



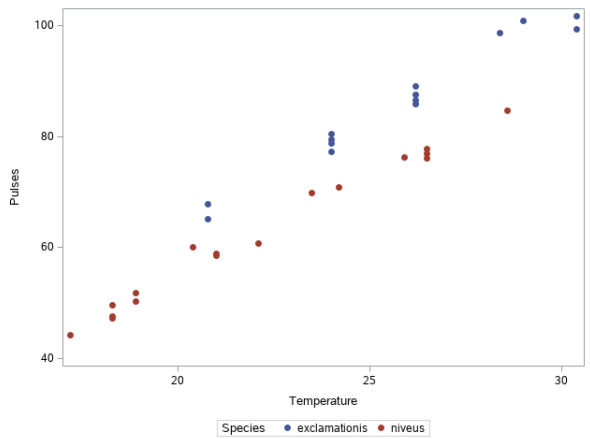
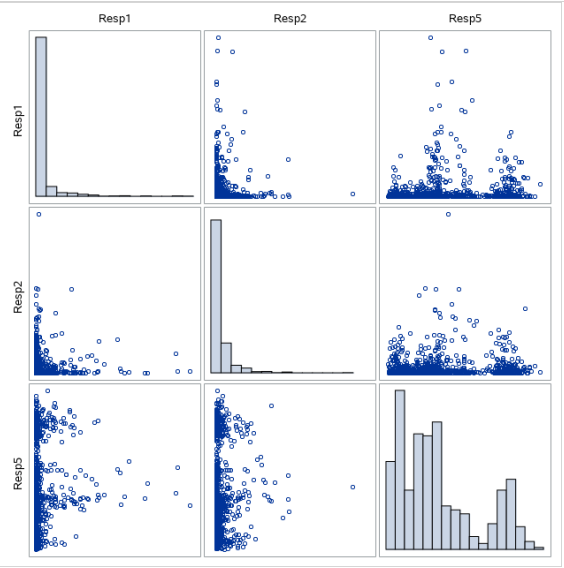
2) *To the right is a three-way scatterplot matrix of three response variables. Does there appear to be a relationship/correlation between any of the three? If so, why?*



3) *Suppose you have a dataset with a response variable (Weight) and three categorical variables (Gender, Ethnicity, Occupation) and want to use a graph to visualized possible differences in Weight across those variables. What R-code would work best?*

a) pairs(~Weight + Gender + Ethnicity + Occupation)
*b) boxplot(Weight~Gender*Ethnicity*Occupation)*
*c) plot(Weight, Gender, col=Ethnicity*Occupation)*

Assessment 3

1) *The scatterplot to the right displays the number of pulses per hour against the outside temperature for crickets. What does the graph tell out about the two species?*		It appears that exclamationis crickets tend to haver more pulses per hour as the temperature increases
2) *To the right is a three-way scatterplot matrix of three response variables. Does there appear to be a relationship/correlation between any of the three? If so, why?*		Not really. There may be some sort of negative relationship between Resp1 and Resp2, but it is hard to tell.
3) *Suppose you have a dataset with a response variable (Weight) and three categorical variables (Gender, Ethnicity, Occupation) and want to use a graph to visualized possible differences in Weight across those variables. What R-code would work best?* *a) pairs(~Weight + Gender + Ethnicity + Occupation)* *b) boxplot(Weight~Gender*Ethnicity*Occupation)* *c) plot(Weight, Gender, col=Ethnicity*Occupation)*		b) boxplot

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA

Caveats and Concerns

- Variables should be relevant to research questions
  - If you look at enough variables, you're bound to find correlations by chance (mining for significance)
  - Scatterplot matrices can help identify correlated covariates
- Limitations to visualizing complex data
  - Tables are appropriate alternatives
- Exploratory data visualization is not analysis
  - Need to follow up visualization with appropriate statistical analyses

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND UNIVERSITY OF NORTH DAKOTA.

Real World Examples

Siegel, R. L., Miller, K.D., Jemal, A. (2020). "Cancer statistics, 2020." CA Cancer Journal for Clinicians **70**(1).

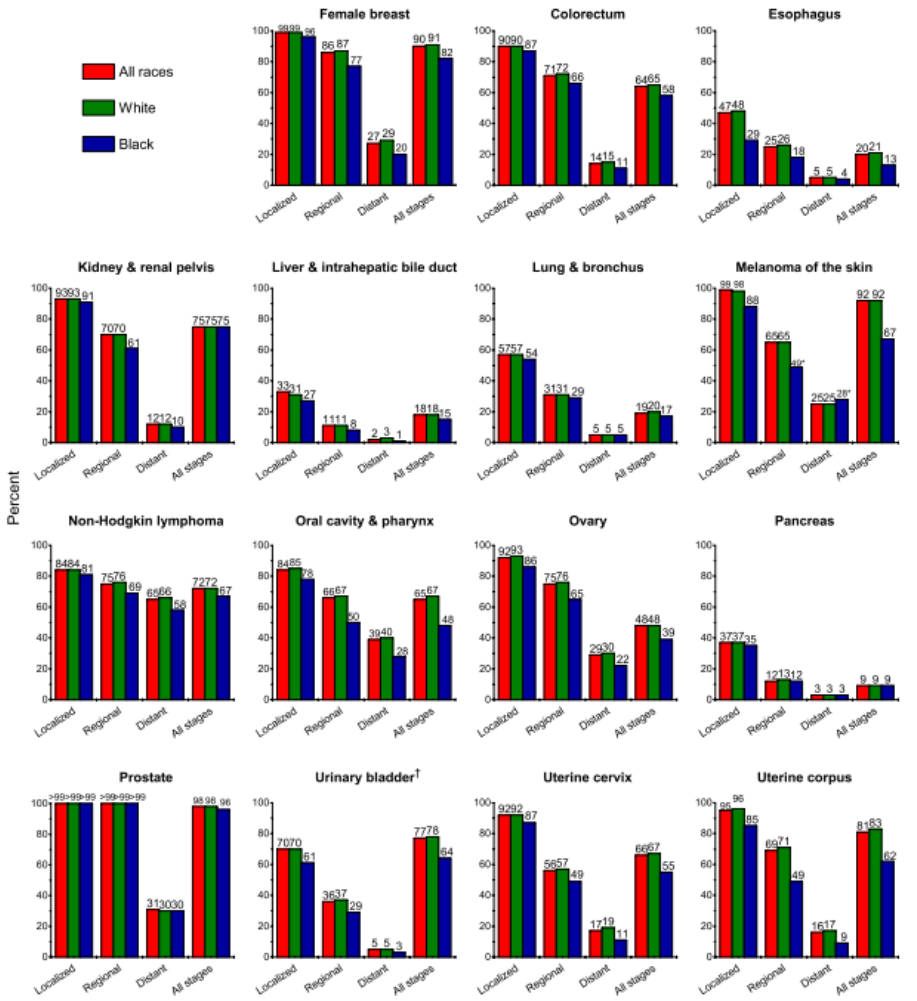


FIGURE 4. Five-Year Relative Survival Rates for Selected Cancers by Race and Stage at Diagnosis, United States, 2009 to 2015. *The standard error of the survival rate is between 5 and 10 percentage points. †The survival rate for carcinoma in situ of the urinary bladder is 95% in all races, 95% in whites, and 91% in blacks.

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND UNIVERSITY OF NORTH DAKOTA.

Real World Examples

(2020). "The GTEx Consortium atlas of genetic regulatory effects across human tissues." Science **369**(6509): 1318.

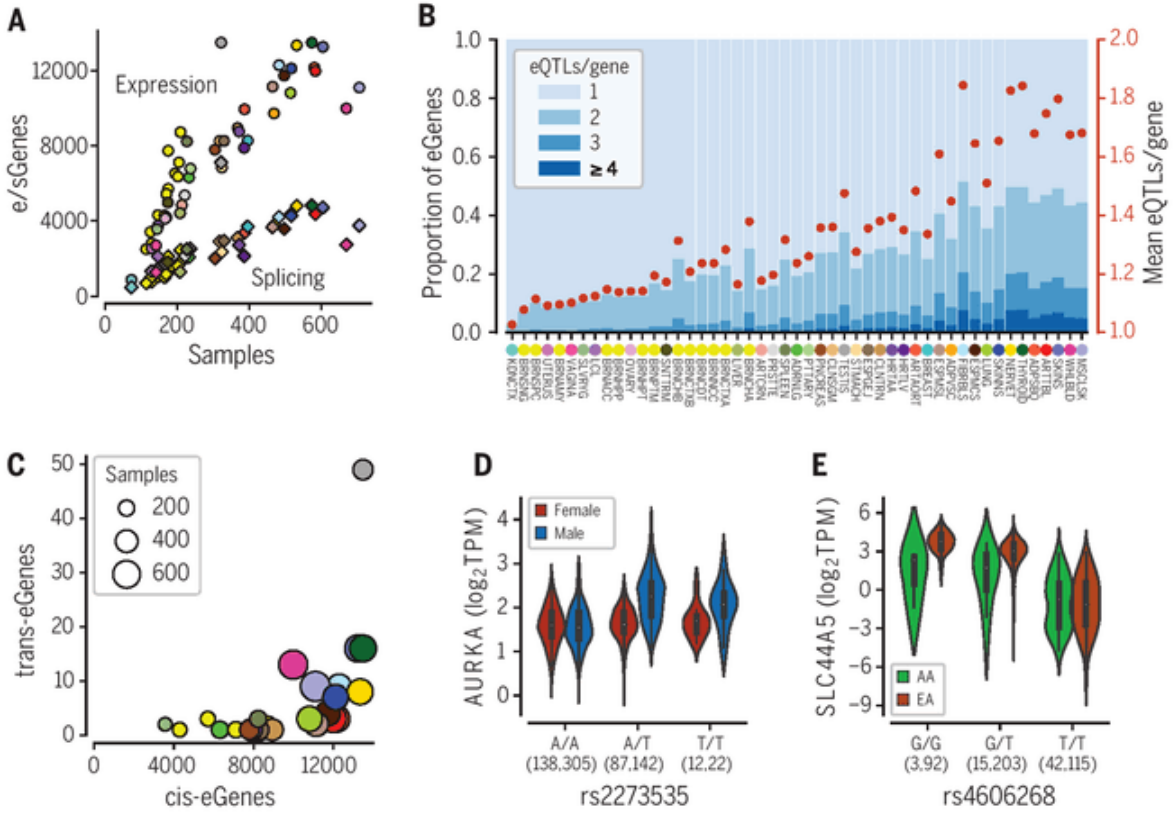


Fig. 2. QTL discovery. (**A**) The number of genes with a cis-eQTL (eGenes) or cis-sQTL (sGenes) per tissue, as a function of sample size. See Fig. 1A for the legend of tissue colors. (**B**) Allelic heterogeneity of cis-eQTLs depicted as proportion of eGenes with one or more independent cis-eQTLs (blue stacked bars; left *y* axis) and as a mean number of cis-eQTLs per gene (red dots; right *y* axis). The tissues are ordered by sample size. (**C**) The number of genes with a trans-eQTL as a function of the number of cis-eGenes. (**D**) Sex-biased cis-eQTL for *AURKA* in skeletal muscle, where rs2273535-T is associated with increased *AURKA* expression in males ($P = 9.02 \times 10^{-27}$) but not in females ($P = 0.75$). (**E**) Population-biased cis-eQTL for *SLC44A5* in esophagus mucosa [aFC = −2.85 and −4.82 and in African Americans (AA) and European Americans (EA), respectively; permutation *P* value = 1.2×10^{-3}]. TPM, transcripts per million.

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA

Summary and Conclusion

- Lots of methods for more advanced data exploration and visualization
- Helps to understand data more
- Increasingly useful in the era of large dataset and complex analyses