



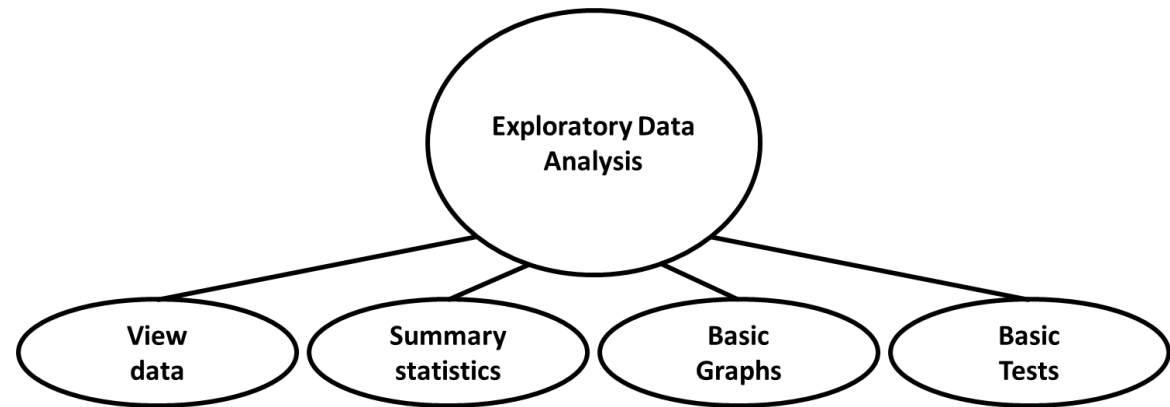
# Exploratory Data Analysis Module II: Leaves and Trees

Dr. Mark Williamson  
DaCCoTA  
University of North Dakota

# Introduction



- Exploration of datasets to summarize main characteristics
- Last time:
  - viewing data
  - summary statistics
  - basic graphs
  - basic tests
- Coming up:
  - Rational and descriptions
  - Step-by-step examples and assessments
  - Caveats and real-world examples



# Rationales

Why should we perform exploratory data analysis?

1. Get to know your data



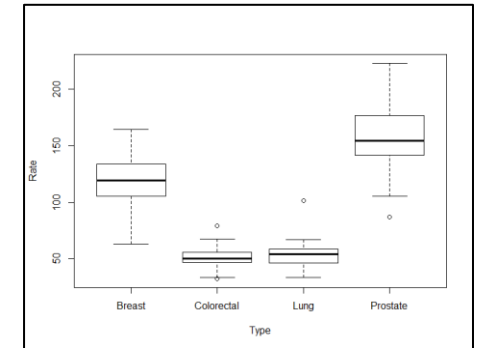
Descriptive Statistics					
Variable	Obs	Mean	Std.Dev.	Min	Max
price	74	6165.257	2949.496	3291	15906
mpg	74	21.297	5.786	12	41
rep78	69	3.406	.99	1	5
headroom	74	2.993	.846	.846	5
trunk	74	13.757	4.277	5	23
weight	74	3019.459	777.194	1760	4840
length	74	187.932	22.266	142	233
turn	74	39.649	4.399	31	51
displacement	74	197.297	91.837	79	425
gear_ratio	74	3.015	.456	2.19	3.89
foreign	74	.297	.46	0	1

2. Save time and effort in the long run

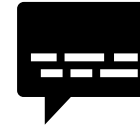


SAS: model Temp= Weight  $\leftrightarrow$  model Weight= Temp  
R: Temp ~ Weight  $\leftrightarrow$  Weight ~ Temp

3. Defendable results

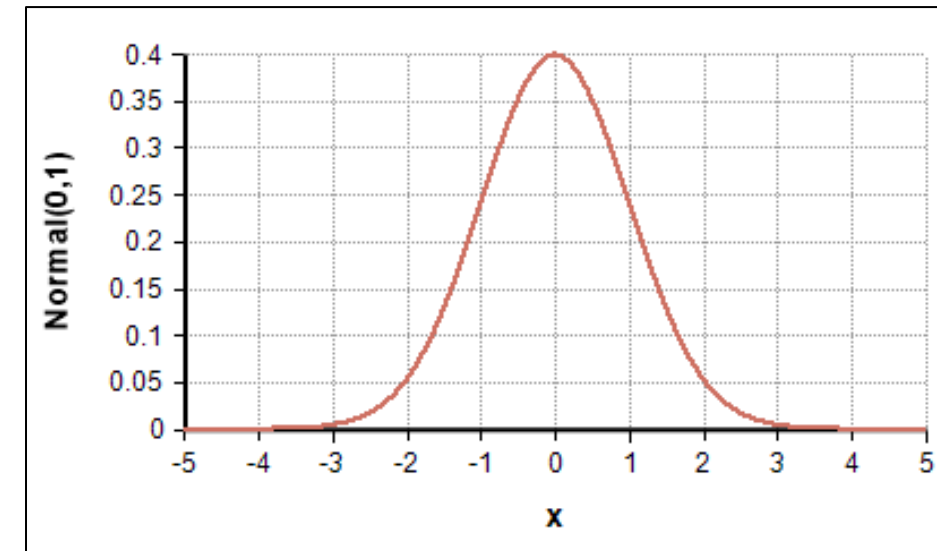


# Descriptions

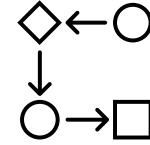


- **Statistical models** – mathematical description of how data conceivably can be produced
- **Parametric data** – fits a normal distribution, assumed for many statistical tests
- **Paired data**-two measurements not independent (ex. before/after)
- **Repeated measures**-two or more measurements not independent (ex. time intervals)
- **Independent variable**-does not depend on another variable; causative, predictor, X
- **Dependent variable**-variable of interest, depends on other variables; response, Y

$$y = \beta_0 + \beta_1 x + e$$

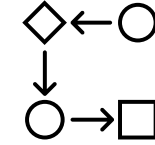


# Step-by-step Example 1



- Software used: R
  - In the datasets package, I'll use data set trees
  - Contains the diameter, height, and volume for Black Cherry Trees
- Research Question:
  - Can we use girth or height to accurately predict volume?
  - Useful because getting volume is difficult--girth and height much easier

# Step-by-step Example 1



## 1) Look at data

> `print(trees)`

3 variables

All numerical

No missing data

## 2) Summary stats

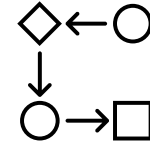
> `summary(trees)`

Girth Height Volume

1	8.3	70	10.3
2	8.6	65	10.3
3	8.8	63	10.2
4	10.5	72	16.4
5	10.7	81	18.8
6	10.8	83	19.7
7	11.0	66	15.6
8	11.0	75	18.2
9	11.1	80	22.6
10	11.2	75	19.9
11	11.3	79	24.2
12	11.4	76	21.0
13	11.4	76	21.4
14	11.7	69	21.3
15	12.0	75	19.1
16	12.9	74	22.2
17	12.9	85	33.8
18	13.3	86	27.4
19	13.7	71	25.7
20	13.8	64	24.9
21	14.0	78	34.5
22	14.2	80	31.7
23	14.5	74	36.3
24	16.0	72	38.3
25	16.3	77	42.6
26	17.3	81	55.4
27	17.5	82	55.7
28	17.9	80	58.3
29	18.0	80	51.5
30	18.0	80	51.0
31	20.6	87	77.0

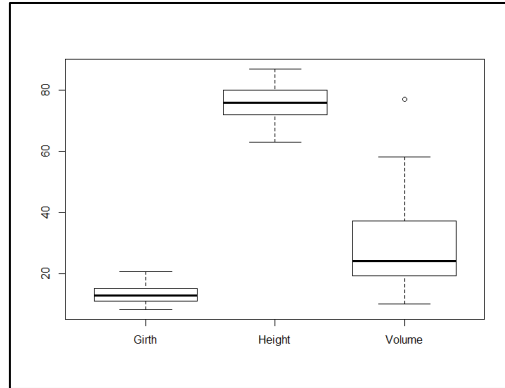
Girth	Height	Volume
Min. :8.30	Min. :63	Min. :10.20
1st Qu.:11.05	1st Qu.:72	1st Qu.:19.40
Median :12.90	Median :76	Median :24.20
Mean :13.25	Mean :76	Mean :30.17
3rd Qu.:15.25	3rd Qu.:80	3rd Qu.:37.30
Max. :20.60	Max. :87	Max. :77.00

# Step-by-step Example 1

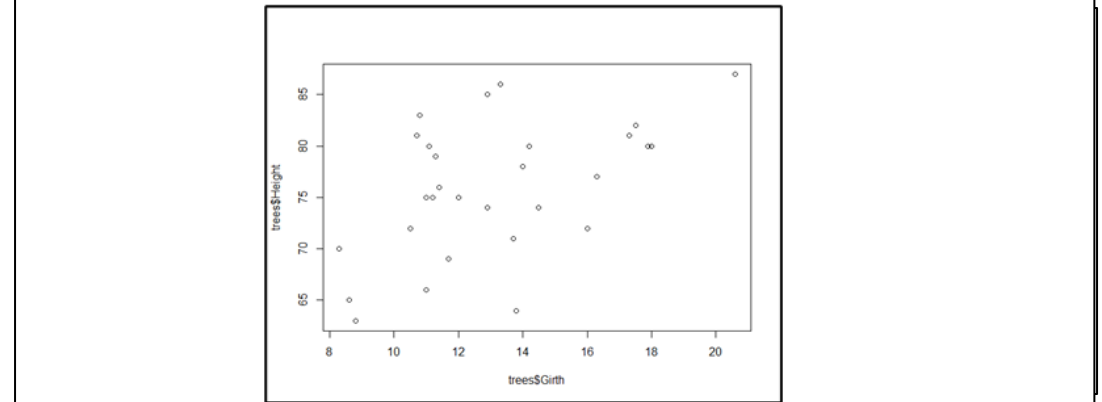
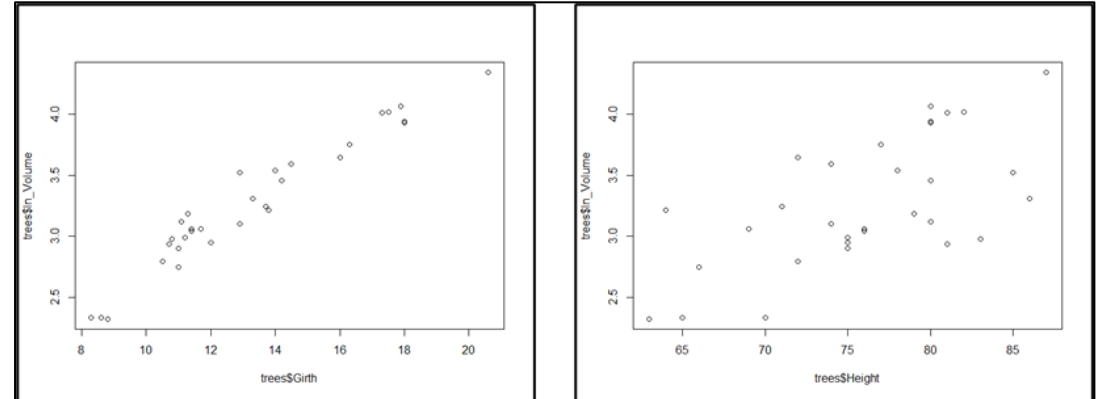


## 3) Graphing

- > `boxplot(trees)`
- > `hist(trees$Girth)`
- > `hist(trees$Height)`
- > `hist(trees$Volume)`

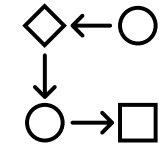


- > `trees$ln_Volume <- log(trees$Volume)`
- > `hist(trees$ln_Volume)`
- > `qqnorm(trees$Volume);qqline(trees$Volume)`
- > `qqnorm(trees$ln_Volume);qqline(trees$ln_Volume)`



- > `plot(trees$ln_Volume~trees$Girth)`
- > `plot(trees$ln_Volume~trees$Height)`
- > `plot(trees$Height~trees$Girth)`

# Step-by-step Example 1



## 4) Simple Tests

```
> cor(trees$ln_Volume, trees$Girth) [1] 0.9693838
> cor(trees$ln_Volume, trees$Height) [1] 0.6482742
> cor(trees$Girth, trees$Height) [1] 0.5192801

> lm1<-lm(trees$ln_Volume~trees$Girth)
> summary(lm1)

> lm2<-lm(trees$ln_Volume~trees$Height)
> summary(lm2)
```

Conclusion: run regression

```
Call:
lm(formula = trees$ln_Volume ~ trees$Height)

Residuals:
    Min     1Q   Median     3Q    Max
-0.66691 -0.26539 -0.06555  0.42608  0.58689

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.79652   0.89053  -0.894   0.378
trees$Height  0.05354   0.01168   4.585 8.03e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4076 on 29 degrees of freedom
Multiple R-squared:  0.4203,    Adjusted R-squared:  0.4003
F-statistic: 21.02 on 1 and 29 DF, p-value: 8.026e-05
```



# Assessment 1



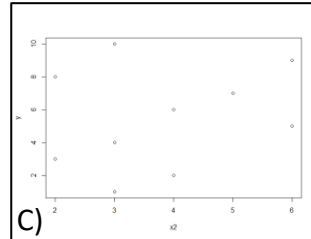
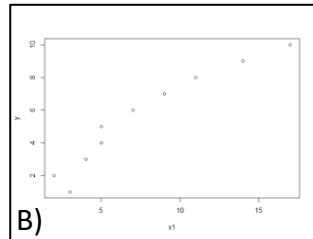
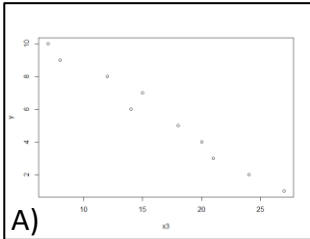
1. What variable is the X variable in the following R equation? What variable is the Y?

```
>scatter(leaf_number ~ branch_number)
```

2. Which variable (Fig, Chestnut, and Oak) has the strongest relationship to Apple?

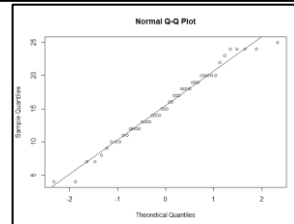
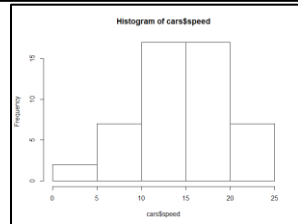
```
>Cor(Apple, Fig) -----> 0.56
>Cor(Apple, Chestnut) ----> 0.24
>Cor(Apple, Oak) -----> -0.82
```

3. Is there a relationship between the two variables in the graphs below? If so, what kind?



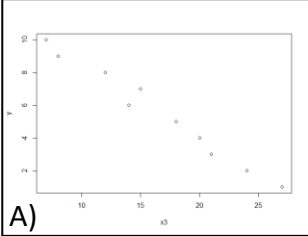
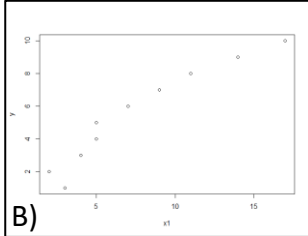
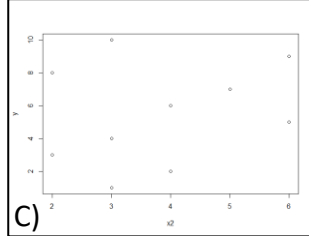
4. What are two graphs you can use to visualize if data is normally distributed?

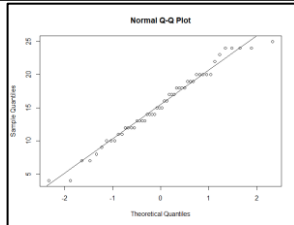
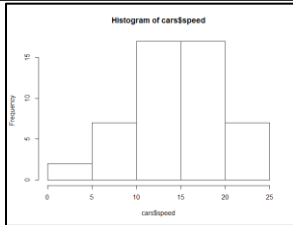
5. Is this data normally distributed?



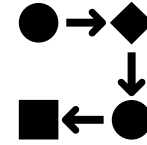
# Assessment 1



<p>1. What variable is the X variable in the following R equation? What variable is the Y?  <code>&gt;scatter(leaf_number ~ branch_number)</code></p>	<p>leaf_number is Y (dependent) variable          branch_number is X (independent) variable</p>
<p>2. Which variable (Fig, Chestnut, and Oak) has the strongest relationship to Apple?  <code>&gt;Cor(Apple, Fig) -----&gt; 0.56</code>  <code>&gt;Cor(Apple, Chestnut) ----&gt; 0.24</code>  <code>&gt;Cor(Apple, Oak) -----&gt; -0.82</code></p>	<p>Oak has the strongest relationship to Apple</p>
<p>3. Is there a relationship between the two variables in the graphs below? If so, what kind?</p> <div style="display: flex; justify-content: space-around;"> <div data-bbox="264 796 570 1029">  <p>A)</p> </div> <div data-bbox="596 796 901 1029">  <p>B)</p> </div> <div data-bbox="927 796 1233 1029">  <p>C)</p> </div> </div>	<p>A) Yes, negative relationship          B) Yes, positive relationship          C) No</p>
<p>4. What are two graphs you can use to visualize if data is normally distributed?</p>	<p>Histogram, qq-plot</p>
<p>5. Is this data normally distributed?</p>	<p>Yes, looks to be so</p>

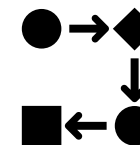


# Step-by-step Example 2



- Software used: SAS
  - In the sashelp library, I'll use data set fish
  - Contains the Weight, Length (3 measurements), Height, and Width of 7 species of fish caught in Finland
- Research Question:
  - Is there a width difference between the species of fish?

# Step-by-step Example 2



## 1) Look at data

**PROC PRINT** data=fish;

7 variables Species is categorial nominal, rest are numerical continuous

Weight has a missing value (observation 14)

We'll only use Species and Width (ignore the rest)

## 2) Summary stats

**PROC UNIVARIATE** data=fish;

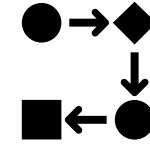
var Width;

**PROC FREQ** data=fish;

tables Species;

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

# Step-by-step Example 2



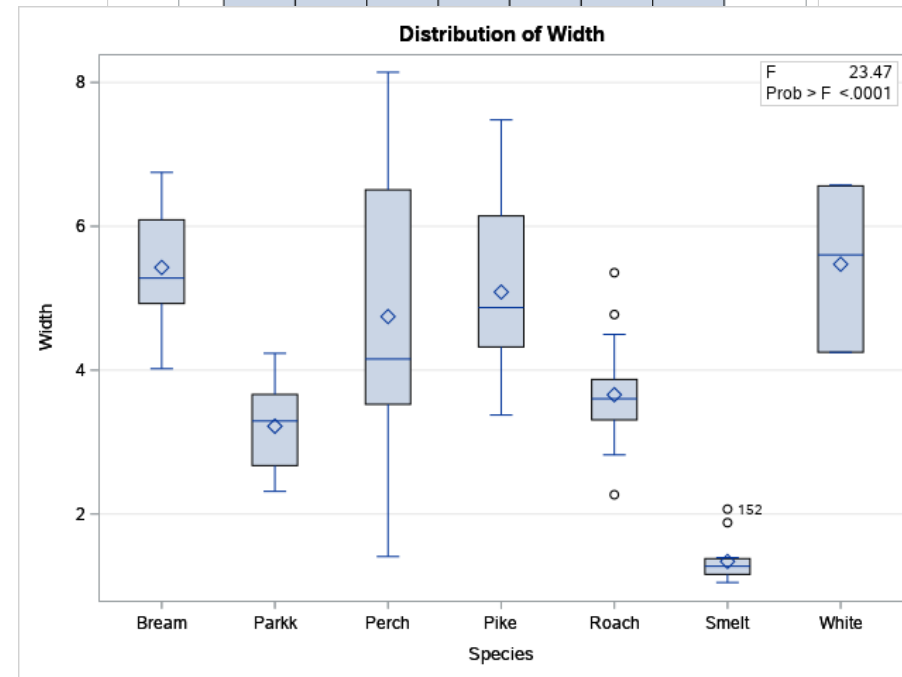
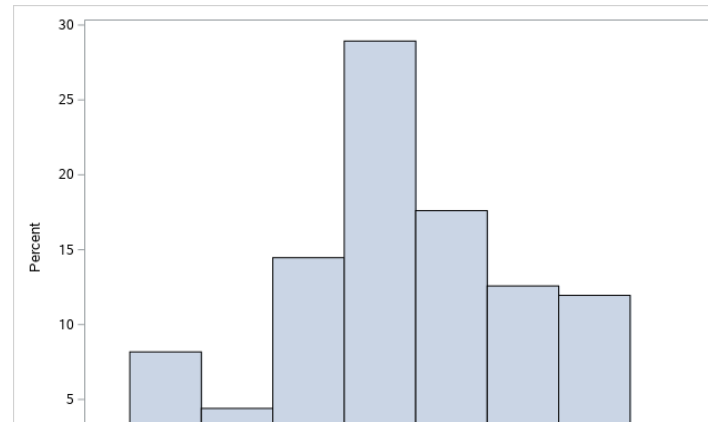
## 3) Graphing

**PROC SGPLOT data=fish;**

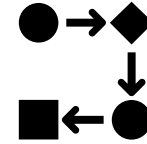
**histogram Width;**

**PROC SGPLOT data=fish;**

**vbox Width / category=Species;**



# Step-by-step Example 2



## 4) Simple Tests

```
PROC SORT data=fish;
```

```
  by Species;
```

```
PROC UNIVARIATE data=fish normal;
```

```
  by Species;
```

```
  var Width;
```

```
  qqplot /normal (mu=est sigma=est);
```

```
  histogram / normal;
```

```
PROC GLM data=fish;
```

```
  class Species;
```

```
  model Width=Species;
```

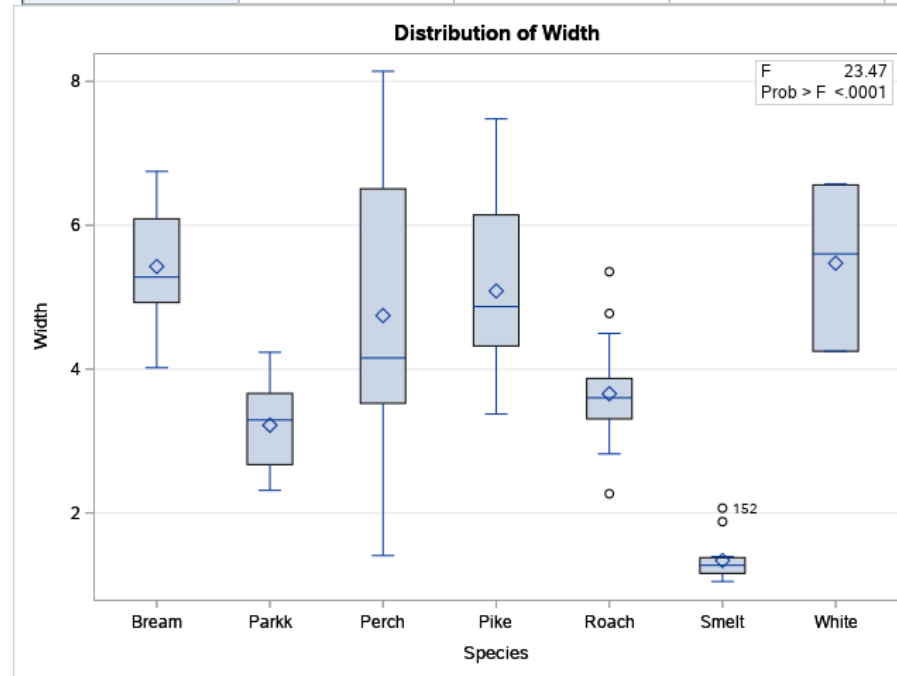
```
  means Species / hovtest=levене(type=abs);
```

Conclusion: run modified ANOVA

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		



# Assessment 2

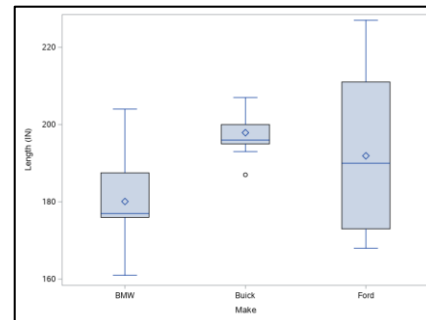


1. What variable is the X variable in the following SAS equation? What variable is the Y?  
`model Length = Species`

2. Based on the SAS output, is there equal variance?

Levene's Test for Homogeneity of Length Variance ANOVA of Absolute Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Make	2	63.5302	31.7651	0.61	0.5479
Error	42	2185.7	52.0408		

3. Based on the boxplot, would you expect the categories of cars to have equal variance? Why or why not?



4. How can the assumption of independent sampling be tested?

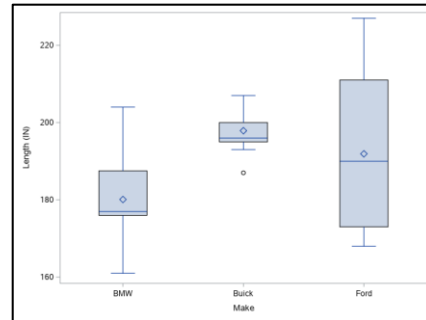
5. Suppose your data consists of fuel efficiency (miles per gallon) across four different car makes (Ford, Honda, Nissan, and Dodge). How should you test for normality to run an ANOVA (aka, is there a difference in fuel efficiency across makes)?

**a)** check normality over all makes,    **b)** check normality for each make individually

# Assessment 2



<p>1. What variable is the X variable in the following SAS equation? What variable is the Y?  <code>model Length = Species</code></p>	<p>Length is Y (dependent) variable                  Species is X (independent) variable</p>																								
<p>2. Based on the SAS output, is there equal variance?</p> <table border="1" data-bbox="206 522 1352 732"> <thead> <tr> <th colspan="6">Levene's Test for Homogeneity of Length Variance ANOVA of Absolute Deviations from Group Means</th> </tr> <tr> <th>Source</th> <th>DF</th> <th>Sum of Squares</th> <th>Mean Square</th> <th>F Value</th> <th>Pr &gt; F</th> </tr> </thead> <tbody> <tr> <td>Make</td> <td>2</td> <td>63.5302</td> <td>31.7651</td> <td>0.61</td> <td>0.5479</td> </tr> <tr> <td>Error</td> <td>42</td> <td>2185.7</td> <td>52.0408</td> <td></td> <td></td> </tr> </tbody> </table>	Levene's Test for Homogeneity of Length Variance ANOVA of Absolute Deviations from Group Means						Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	Make	2	63.5302	31.7651	0.61	0.5479	Error	42	2185.7	52.0408			<p>Yes, because the p-value (0.5479) is greater than 0.05, so we fail to reject the hypothesis that the variances are equal</p>
Levene's Test for Homogeneity of Length Variance ANOVA of Absolute Deviations from Group Means																									
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F																				
Make	2	63.5302	31.7651	0.61	0.5479																				
Error	42	2185.7	52.0408																						
<p>3. Based on the boxplot, would you expect the categories of cars to have equal variance? Why or why not?</p>	<p>No, the quartile lengths are very different.</p>																								
<p>4. How can the assumption of independent sampling be tested?</p>	<p>It can't. Good sampling design ensures the assumption is met.</p>																								
<p>5. Suppose your data consists of fuel efficiency (miles per gallon) across four different car makes (Ford, Honda, Nissan, and Dodge). How should you test for normality to run an ANOVA (aka, is there a difference in fuel efficiency across makes)?  <b>a)</b> check normality over all makes,      <b>b)</b> check normality for each make individually</p>	<p><b>b)</b></p>																								





# Caveats and Concerns

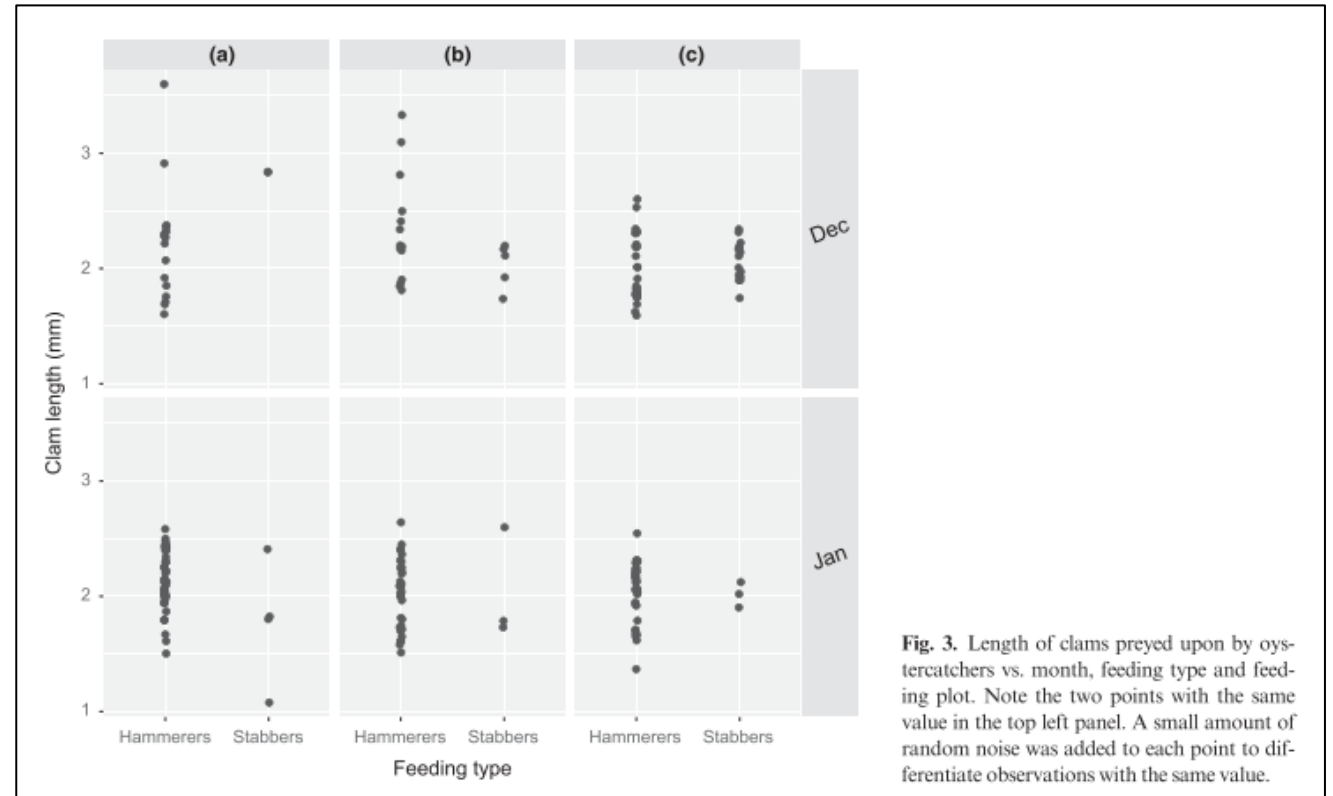
- Normality tests are an art
  - Suggest using histograms and qq-plots over tests for normality
- There is more than one way of doing things
- Code output can be confusing
- Data can be problematic by nature and design
  - Uneven samples sizes
  - Unequal variances

# Real World Examples

Zuur, A. F., et al. (2016). "A protocol for conducting and presenting results of regression-type analyses." Methods in Ecology and Evolution **7**(6): 636-645.

## Protocol for conducting and presenting results of regression-type analyses

1. State appropriate questions
2. Visualize the experimental design
3. Conduct data exploration
4. Identify the dependency structure in the data
5. Present the statistical model
6. Fit the model
7. Validate the model
8. Interpret and present the numerical output of the model
9. Create a visual representation of the model
10. Simulate from the model



**Fig. 3.** Length of clams preyed upon by oystercatchers vs. month, feeding type and feeding plot. Note the two points with the same value in the top left panel. A small amount of random noise was added to each point to differentiate observations with the same value.

# Real World Examples

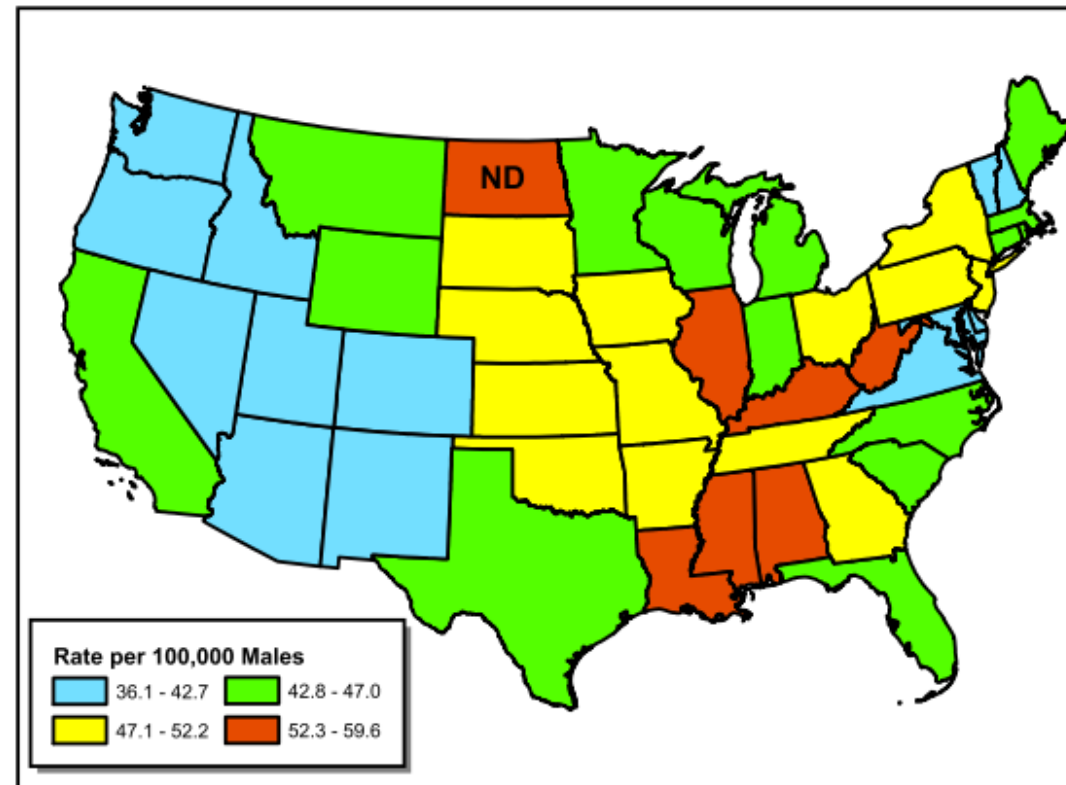
Ahmed, R., et al. (2020). "United States County-level COVID-19 Death Rates and Case Fatality Rates Vary by Region and Urban Status." Healthcare (Basel) **8**(3).

**Table 1.** Factors from the American Community Survey and Definitive Healthcare Hospital Beds data.

Variable	Total US		Midwest		Northeast		Southeast		Southwest		West	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
<b>Demographic</b>												
% White	83.1	89.7	91.4	94.7	85.3	91.2	74.8	78.8	81.3	83.5	83.9	89.8
% Black	9.1	2.3	2.6	0.9	7.1	3.3	20.3	14.2	5.2	2.9	1.3	0.7
% Hispanic	9.3	4.1	4.6	2.8	6.8	3.8	5.0	3.2	30.6	23.5	14.4	9.4
Sex ratio	100.9	98.5	100.9	99.7	99.1	97.1	98.8	96.2	103.8	99.4	104.8	101.6
Age	41.3	41.2	42.0	41.9	42.5	42.5	41.1	41.0	39.4	39.0	41.1	40.2
<b>Socioeconomic</b>												
Family size	3.1	3.0	3.0	2.9	3.0	3.0	3.1	3.1	3.3	3.2	3.2	3.1
Income	51,557	49,886	53,508	52,558	63,285	59,114	45,620	42,621	49,727	48,331	56,827	53,311
% insured	89.9	90.8	92.2	93.3	94.0	94.5	89.1	89.2	83.9	84.3	89.3	90.5
% in poverty	11.2	10.3	8.8	8.2	8.2	8.0	14.5	13.7	12.5	11.8	9.5	9.0
<b>Hospital</b>												
Bed number	305.2	42.0	205.7	25.0	813.4	237.0	251.9	55.0	314.9	25.0	389.5	30.0
Bed utilization	0.3	0.3	0.3	0.3	0.5	0.5	0.3	0.3	0.2	0.2	0.3	0.3
Ventilator number	2.0	2.0	1.5	1.0	3.7	3.0	2.3	2.0	1.4	1.0	1.7	1.5

# Real World Examples

Schwartz, G. G., et al. (2019). "An exploration of colorectal cancer incidence rates in North Dakota, USA, via structural equation modeling." International Journal of Colorectal Disease **34**(9): 1571-1576.



# Summary and Conclusion

- Exploratory Data Analysis is a necessary first step in understanding your data and determining how to analyze it
- Helps to:
  - Get to know your data
  - Save time and effort in the long run
  - End with defensible results
- Many ways to get it done (R, SAS, SPSS, Excel, etc.)
- Tune in next time for a plunge into advanced topics of Exploratory Data Analysis in Module III: Deep Dive