



Exploratory Data Analysis Module I: A Bird's Eye View

Dr. Mark Williamson

DaCCoTA

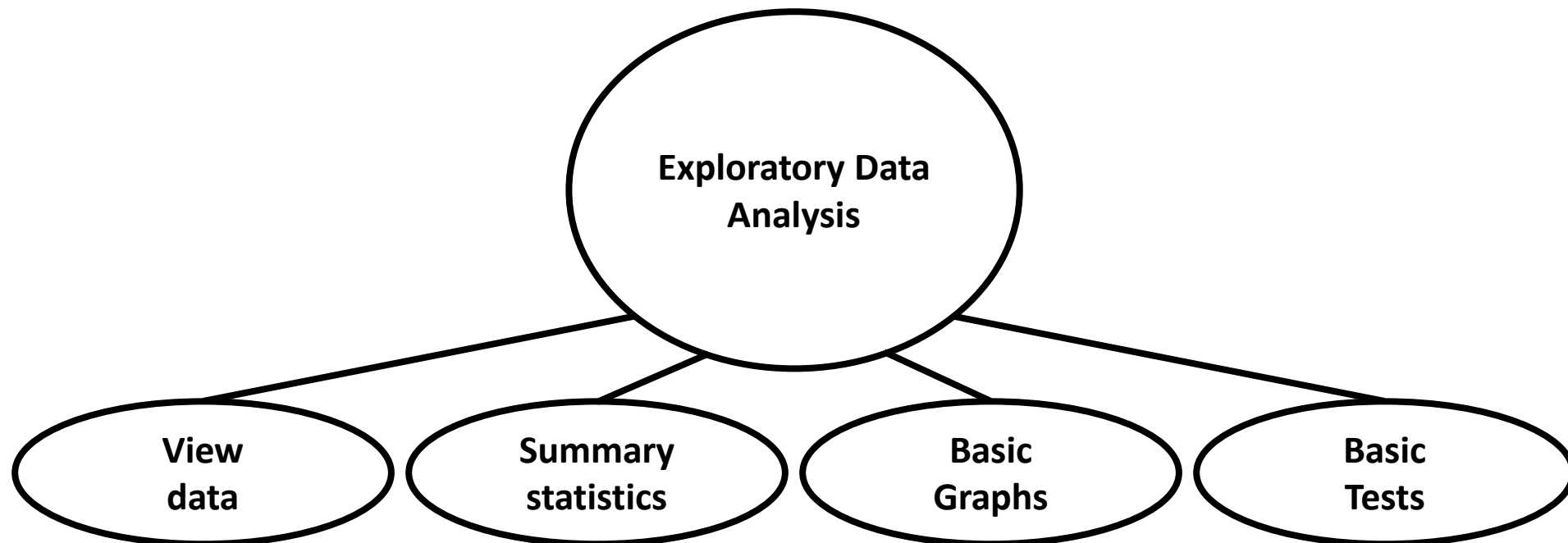
University of North Dakota

Introduction

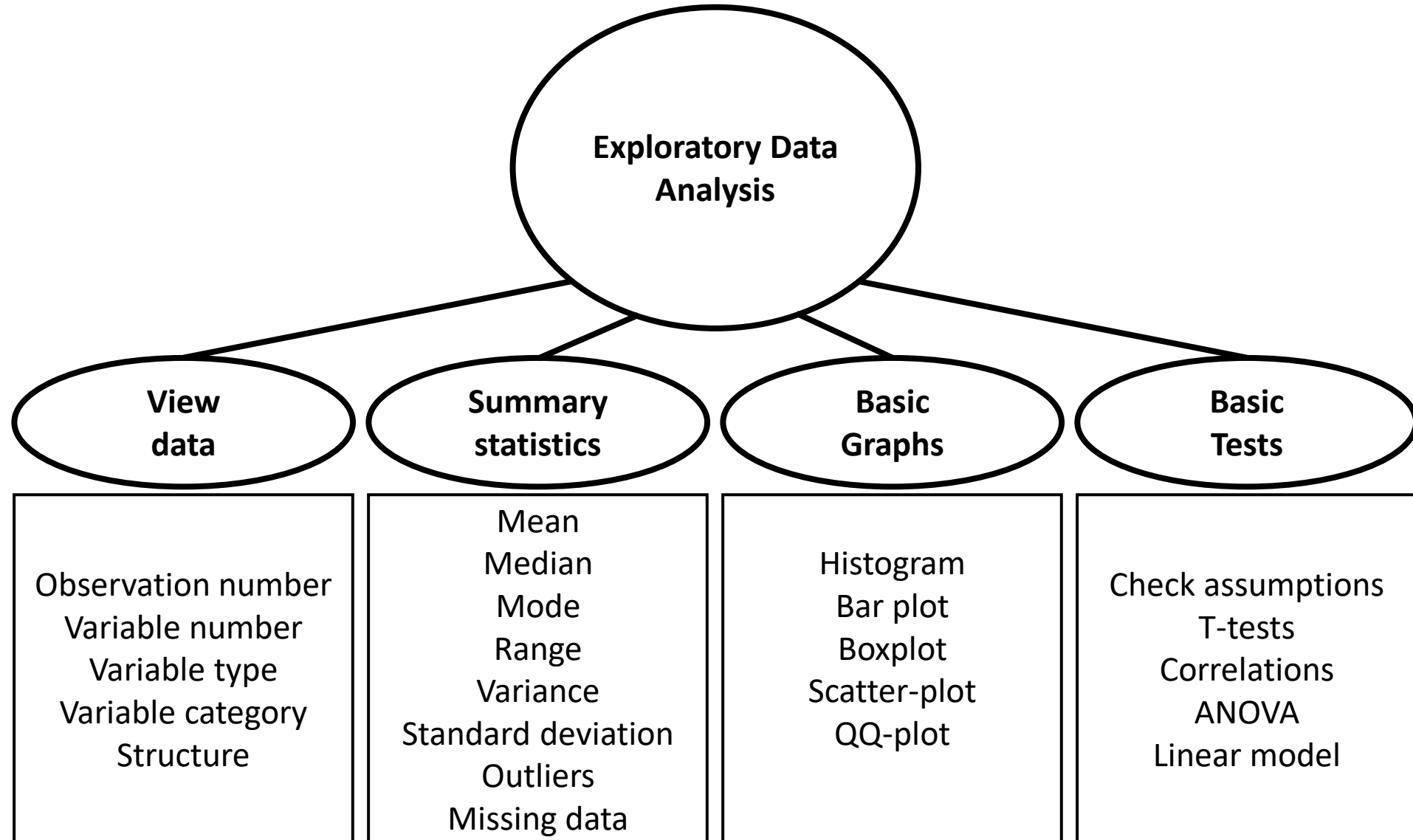
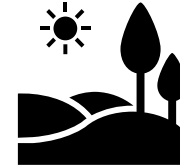


Exploratory data analysis:

- The approach that explores datasets to summarize their main characteristics using summary statistics and graphical methods
- First step for successful statistical analyses



Landscape



Structures and Uses

View data

Observation number
 Variable number
 Variable type
 Variable category
 Structure

Variable category:
 Nuisance/bookkeeping
 Dependent variable
 Independent variable

Variable number:
 Number of columns

Variable type:
 Numerical discrete
 Numerical continuous
 Categorical nominal
 Categorical ordinal

Structure:
 Long or wide
 Paired data
 Repeated measures

Observation number:
 number of rows
 number of samples

Number	Body Temp.	Age	Weight	Major	Grade
1	98.3	19	200.1	Math	Junior
2	99.1	19	245.3	Chemistry	Sophomore
3	98.6	23	197.5	Biology	Senior
4	98.0	20	156.0	Math	Junior
5	95.4	23	220.5	Physics	Junior
6	98.2	18	185.4	Physics	Sophomore
7	98.4	21	172.0	Biology	Senior
8	98.6	18	280.3	Chemistry	Freshman
9	99.0	24	129.0	Biology	Senior
10	98.8	17	210.1	Biology	Freshman



Structures and Uses

Summary statistics

- Mean
- Median
- Mode
- Range
- Variance
- Standard deviation
- Outliers
- Missing data

Mean	Also know as the average; the sum of all values divided by number of values
Median	Also known as the middle value; less influenced by outliers than mean
Mode	The most common value among observations
Range	The distance between the highest and lowest value among observations
Variance	How much the values are spread out; average all the numbers differ from the mean
Standard deviation	Another measures of how much the values are spread out; square root of variance
Outliers	Very high or very low values; may be true or some sort of measurement/writing error
Missing data	Data that should be there but is not; may be error or simple not measured

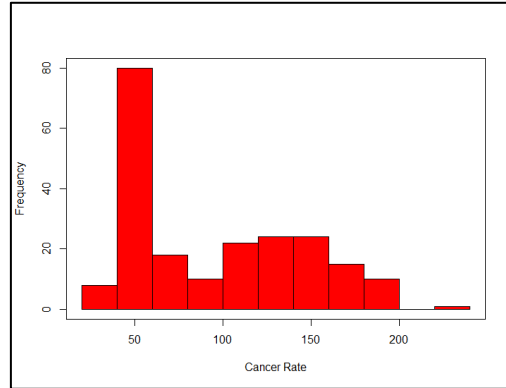
Structures and Uses

Basic Graphs

Histogram
Bar-plot
Boxplot
Scatter-plot
QQ-plot

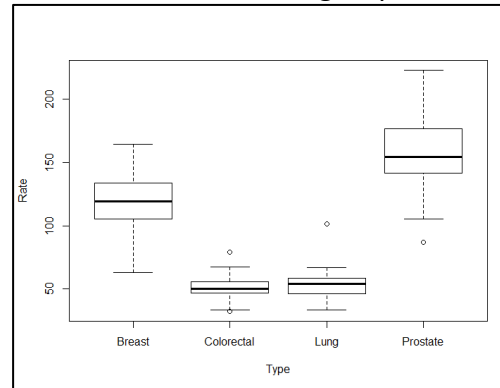
Histogram:

Distribution of values



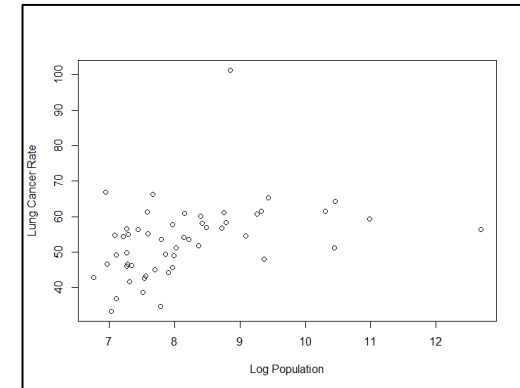
Boxplot:

Median, quartiles,
outliers across groups



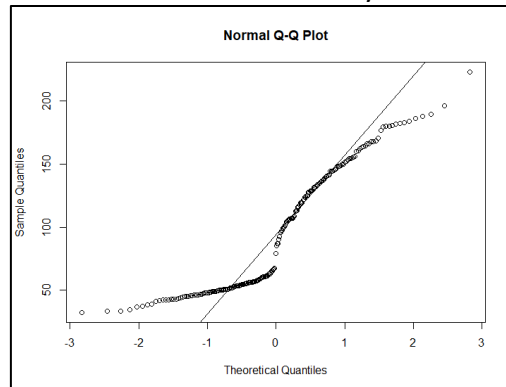
Scatter-plot

Relationship between to
numerical variables



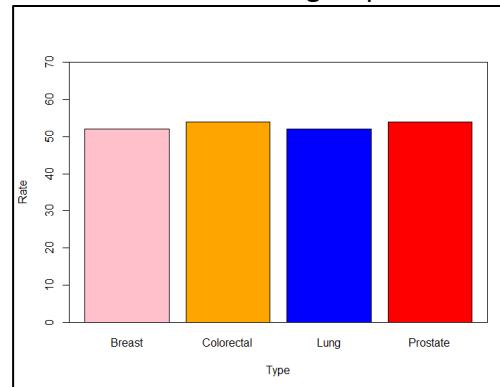
QQ-plot:

Test for normality



Bar-plot:

Mean across groups



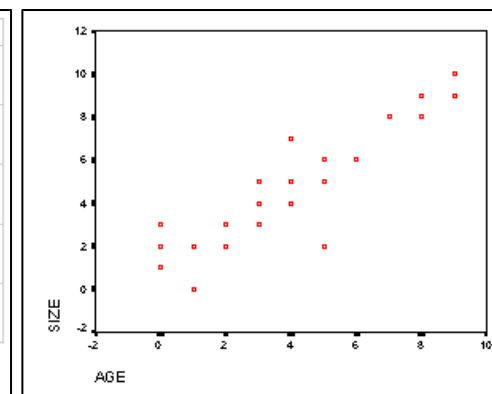
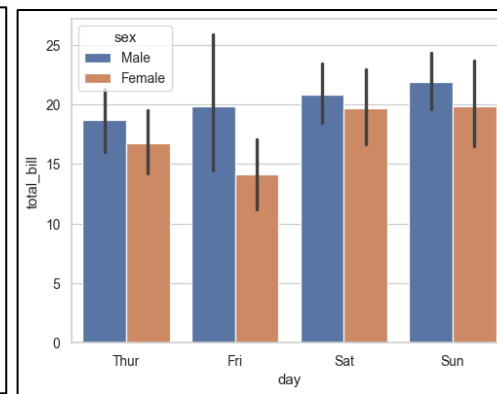
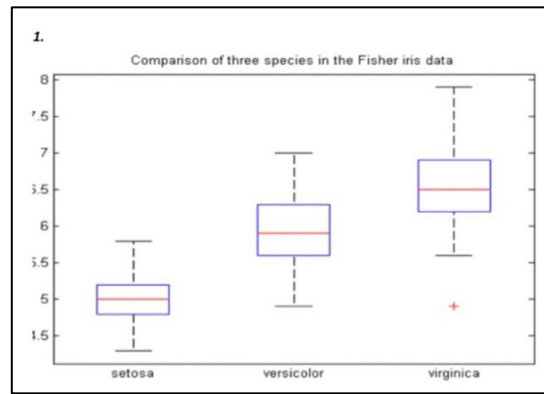
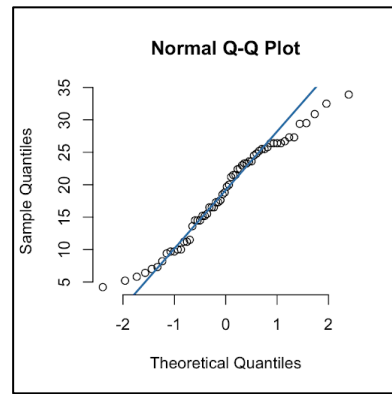
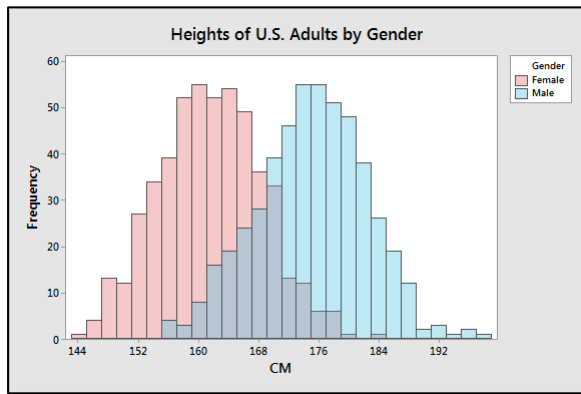
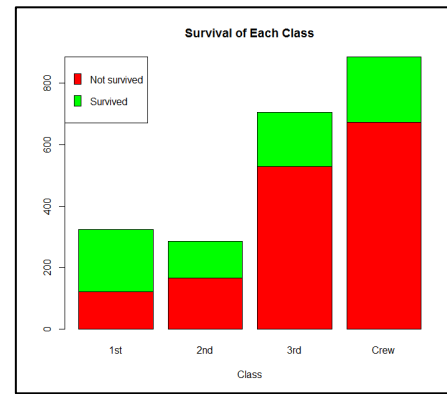
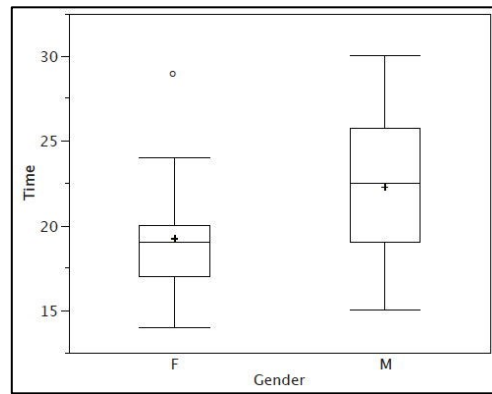
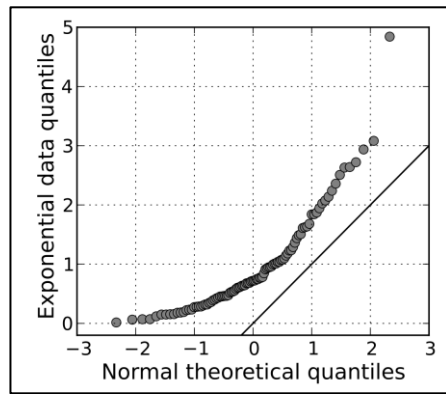
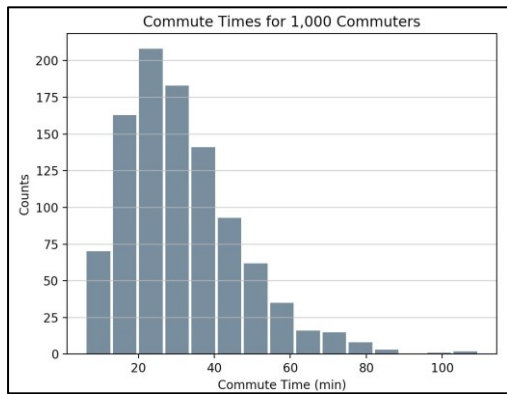
Structures and Uses

**Basic
 Tests**

Check assumptions
 T-tests
 Correlations
 ANOVA
 Linear model

Check assumptions	Is the data normally distributed? (histogram, qq-plot) How large is the variance? (summary statistics) Are there equal observations across groups? (summary statistics)
T-tests	Is there is a difference between one group and a set value or between two groups? (boxplot, bar-plot)
Correlations	Is there a relationship between two numerical variables? (scatter plot)
ANOVA	Is there a difference between three or more groups? (boxplot, bar-plot)
Linear model	Can the dependent variable be predicted from one or more independent variables?

Examples



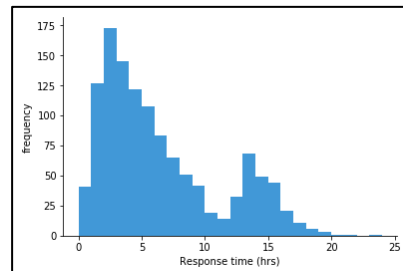
	Mean	SD	Median	Minimum	Maximum	Total
Reading	60.80	17.63	61	27	92	25
Writing	60.00	19.02	60	25	99	25
Verbal	60.00	14.31	62	40	91	25
History	55.20	19.56	51	26	95	25
Math	54.40	19.29	50	23	92	25
Science	50.80	17.96	52	20	82	25

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

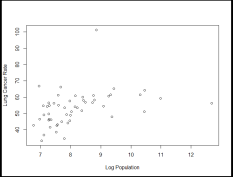
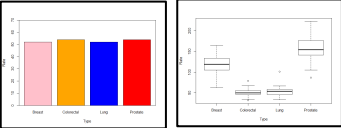
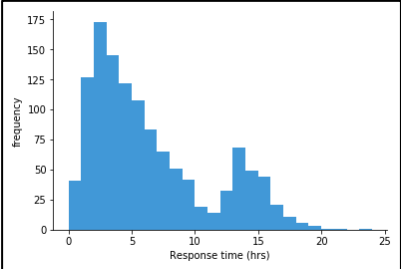
Quick Assessment

1. What are the four basic parts of exploratory data analysis?
2. What graph would you use to see if there might be a relationship between two numerical variables?
3. What exploration part helps determine the variable number, type, and category?
4. What summary statistic is also called the average?
5. If you ran an ANOVA, what sort of graph(s) would you use to display the results?

6. What type of graph is this?



Quick Assessment

<p>1. What are the four basic parts of exploratory data analysis?</p>	<p>View Data, Summary Statistics, Simple Graphs, Simple Tests</p>
<p>2. What graph would you use to see if there might be a relationship between two numerical variables?</p>	<p>Scatter Plot </p>
<p>3. What exploration part helps determine the variable number, type, and category?</p>	<p>Viewing the data</p>
<p>4. What summary statistic is also called the average?</p>	<p>Mean</p>
<p>5. If you ran an ANOVA, what sort of graph(s) would you use to display the results?</p>	<p>Bar-plot and/or boxplot </p>
<p>6. What type of graph is this? </p>	<p>Histogram</p>

Summary and Conclusion



- Exploratory data analysis is the first step in analyzing data
 - Viewing data helps determine the type and structure of the data
 - Summary statistics helps summarize the data numerically
 - Simple graphs helps visualize structure and relationships of the data
 - Simple tests provide a guide for further data analysis
-
- Tune in next time for a stroll through the core components of Exploratory Data Analysis in Module II: Leaves and Trees