The Statistical Software Toolkit

BERDC Special Topics Talk 1

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Dr. Mark Williamson

Biostatistics, Epidemiology, and Research Design Core

Introduction

- There are many statistical software options
- Some are better at some things than others
- At the end of the day, they are all tools to aid you in your research
- It is good to know your way around multiple tools
- Here, we'll be counting down the top 5 tools for statistical software

Inclusion Criteria

- Software is free, has a free version, or has a version that can be accessed through UND for free
- Software I've had at least some experience with

Number 5

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

STATA

UND
UNIVERSITY OF NORTH DAKOTA

Overview:

- general-purpose package created by StataCorp
- 1985

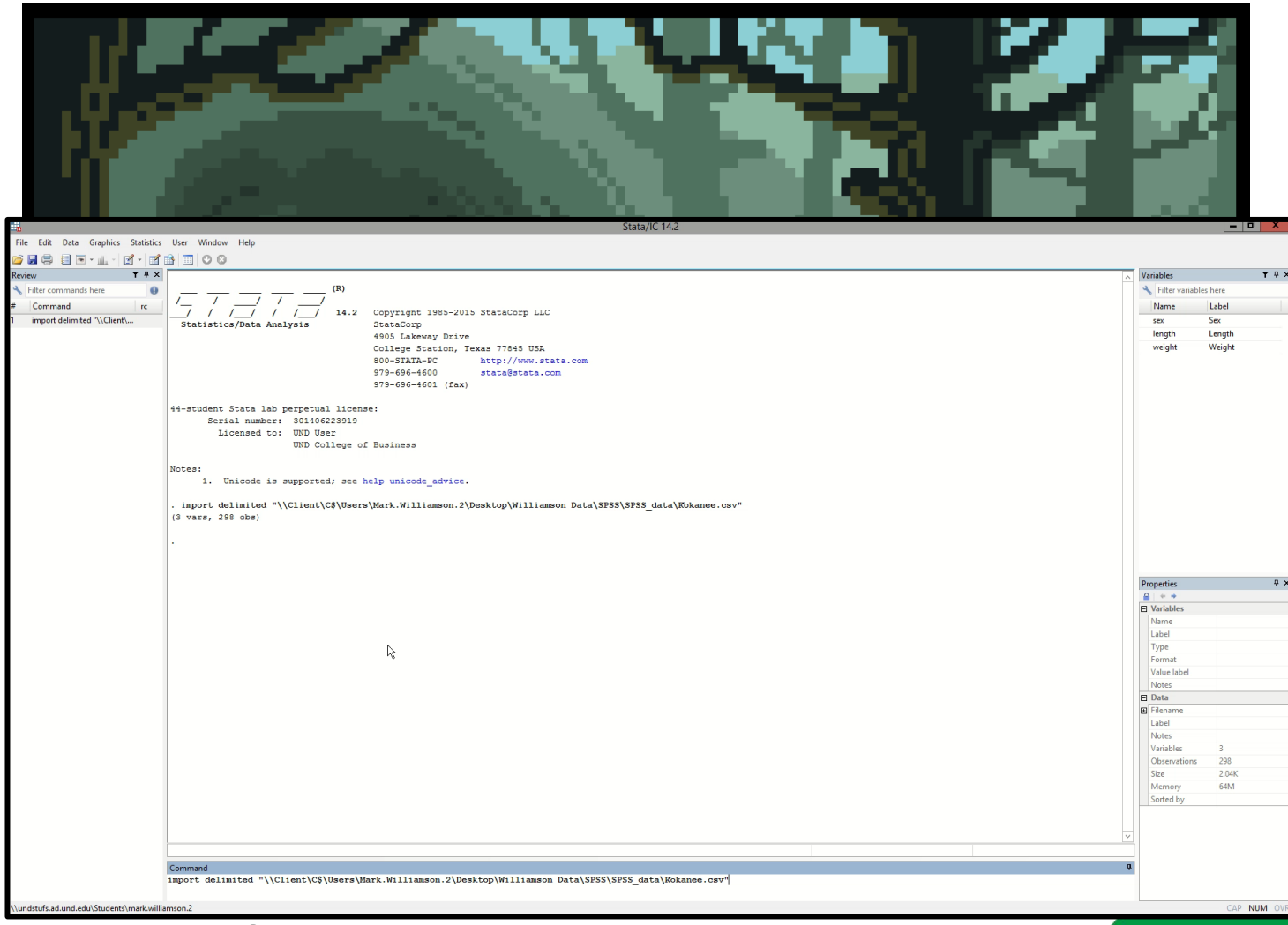
Tool: Rachet wrench

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Stata

UND UNIVERSITY OF NORTH DAKOTA.

- Access
  - Need Citrix Workspace first
    - https://und.teamdynamix.com/TDClient/2048/Portal/KB/ArticleDet?ID=58677
  - Add Stata-64 to the Citrix App
- Features
  - Command
  - Pull down
  - Allows for user-written programs
- Downside
  - Plug-and-chug problem
  - Coding learning curve



Suggested Uses: All purpose data

Number 4

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA

SPSS

Overview:

- Interactive statistical analysis package created by SPSS Inc. then acquired by IBM
- 1968 (SPSS Inc.); 2009 (IBM)

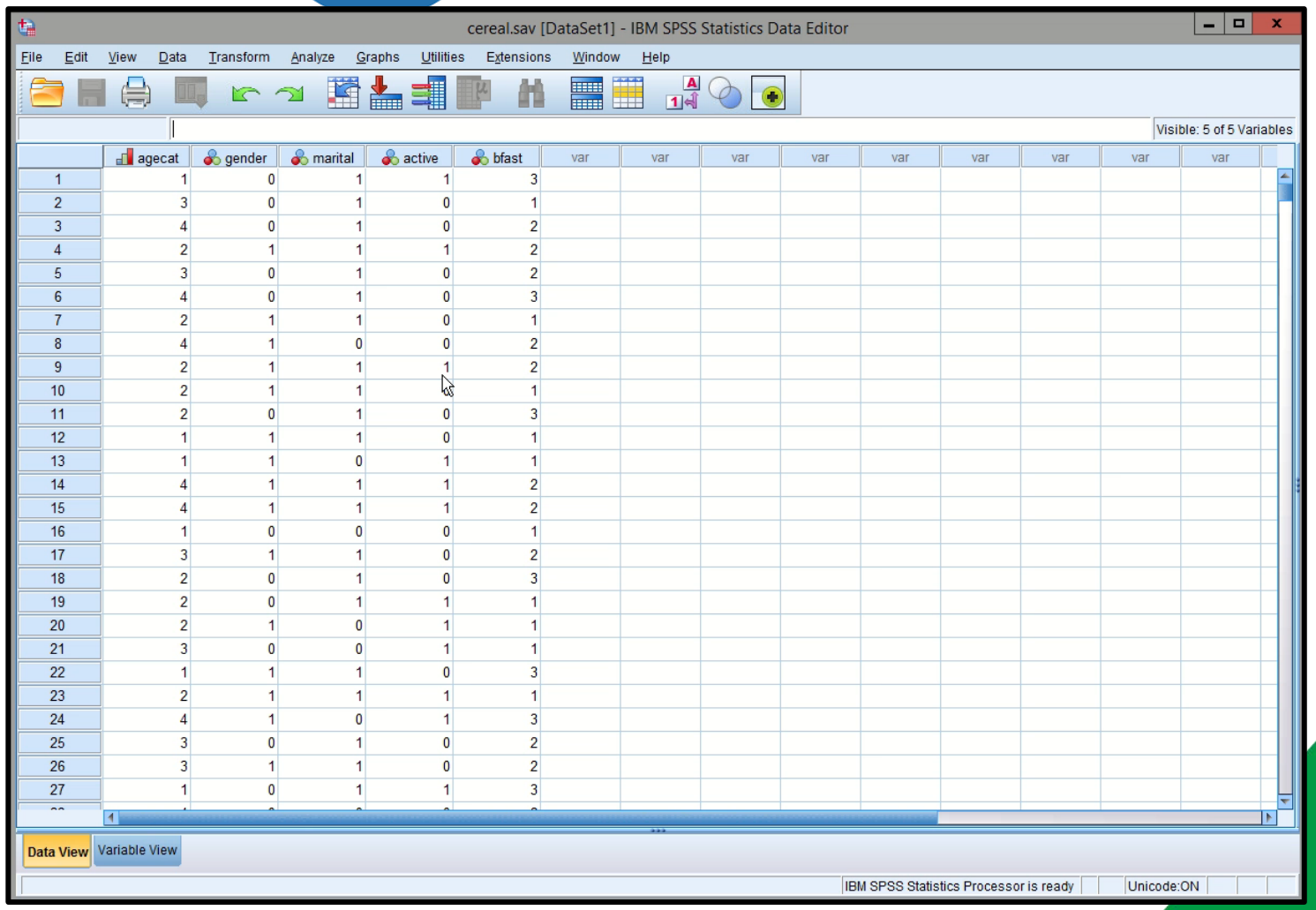
Tool: Power Drill

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND UNIVERSITY OF NORTH DAKOTA.

SPSS®

- Access: Citrix App
- Features
  - All pull-down except a few
  - In-package data working
  - Sample Datasets
- Downsides
  - Pain to get data in
  - Plug-and-chug
  - No coding



	agecat	gender	marital	active	bfast
1	1	0	1	1	3
2	3	0	1	0	1
3	4	0	1	0	2
4	2	1	1	1	2
5	3	0	1	0	2
6	4	0	1	0	3
7	2	1	1	0	1
8	4	1	0	0	2
9	2	1	1	1	2
10	2	1	1	1	1
11	2	0	1	0	3
12	1	1	1	0	1
13	1	1	0	1	1
14	4	1	1	1	2
15	4	1	1	1	2
16	1	0	0	0	1
17	3	1	1	0	2
18	2	0	1	0	3
19	2	0	1	1	1
20	2	1	0	1	1
21	3	0	0	1	1
22	1	1	1	0	3
23	2	1	1	1	1
24	4	1	0	1	3
25	3	0	1	0	2
26	3	1	1	0	2
27	1	0	1	1	3

Suggested Uses: Very Simple or very complex data

Number 3

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Excel

UND
UNIVERSITY OF NORTH DAKOTA

Overview:

- Spreadsheet developed by Microsoft
- 1987

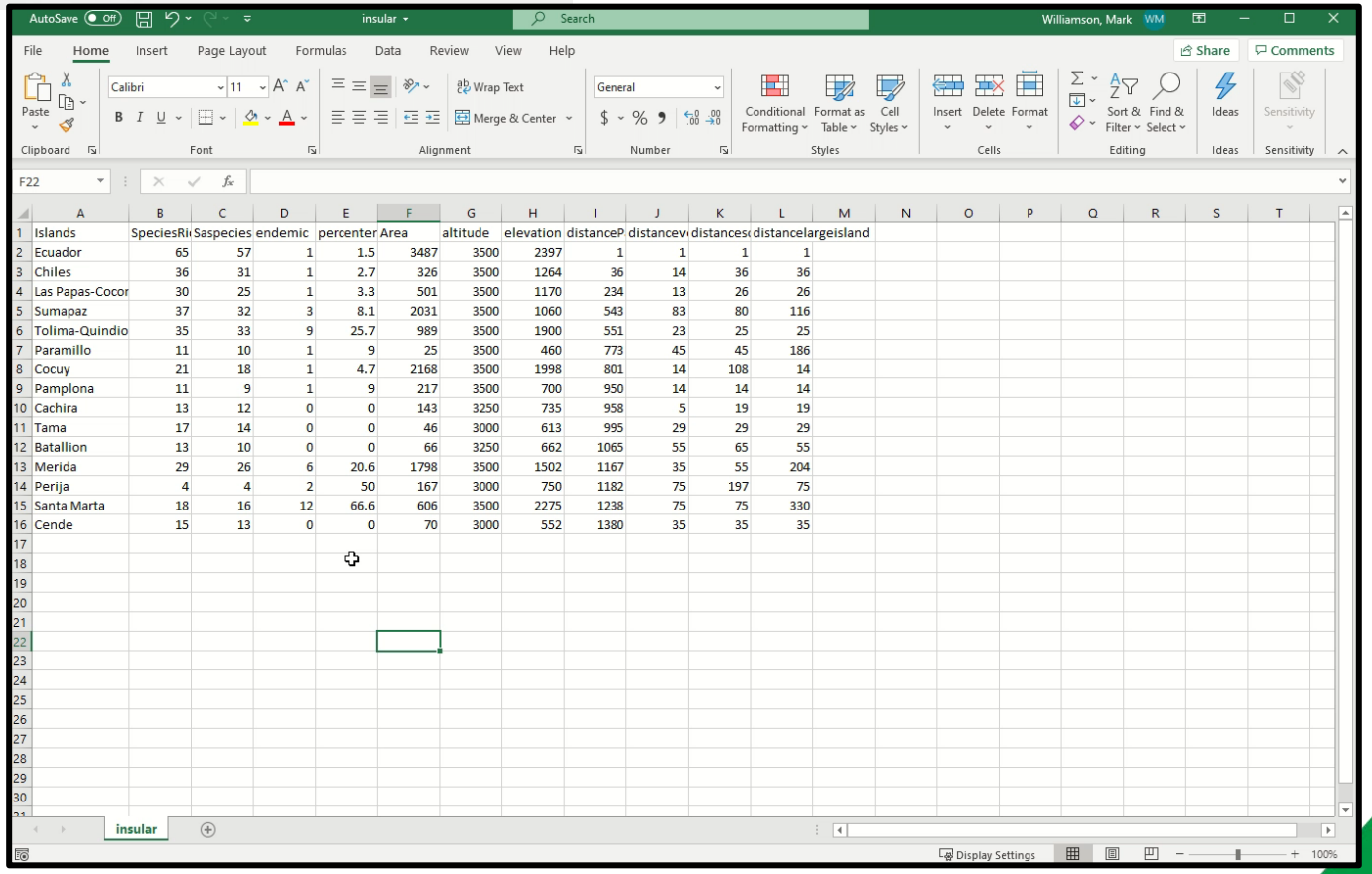
Tool: Hammer

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Excel

UND UNIVERSITY OF NORTH DAKOTA.

- Access: typically built-in
- Features
  - Interactive spreadsheet
  - Functions
- Downsides
  - Not technically statistical software package
  - Basic



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Islands	SpeciesRi	Saspecies	endemic	percenter	Area	altitude	elevation	distanceP	distancev	distances	distancelarge	island
2	Ecuador	65	57	1	1.5	3487	3500	2397	1	1	1	1	
3	Chiles	36	31	1	2.7	326	3500	1264	36	14	36	36	
4	Las Papas-Cocor	30	25	1	3.3	501	3500	1170	234	13	26	26	
5	Sumapaz	37	32	3	8.1	2031	3500	1060	543	83	80	116	
6	Tolima-Quindio	35	33	9	25.7	989	3500	1900	551	23	25	25	
7	Paramillo	11	10	1	9	25	3500	460	773	45	45	186	
8	Cocuy	21	18	1	4.7	2168	3500	1998	801	14	108	14	
9	Pamplona	11	9	1	9	217	3500	700	950	14	14	14	
10	Cachira	13	12	0	0	143	3250	735	958	5	19	19	
11	Tama	17	14	0	0	46	3000	613	995	29	29	29	
12	Batallion	13	10	0	0	66	3250	662	1065	55	65	55	
13	Merida	29	26	6	20.6	1798	3500	1502	1167	35	55	204	
14	Perija	4	4	2	50	167	3000	750	1182	75	197	75	
15	Santa Marta	18	16	12	66.6	606	3500	2275	1238	75	75	330	
16	Cende	15	13	0	0	70	3000	552	1380	35	35	35	

Suggested Uses: Data formatting and exploration

Number 2

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

sas®

UND
UNIVERSITY OF NORTH DAKOTA.

Overview:

- Multi-purpose package developed by SAS Institute
- 1976

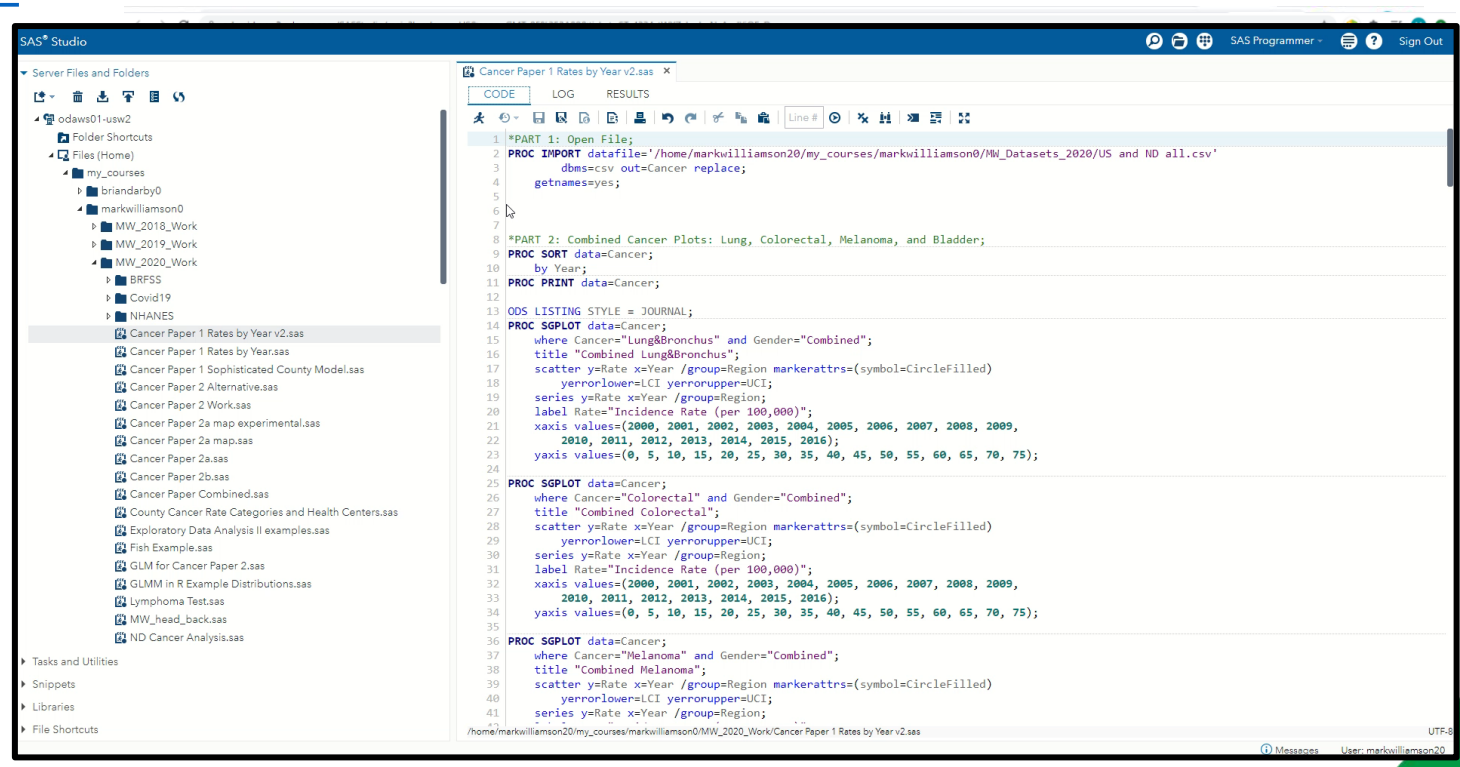
Tool: Table Saw

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

sas

UND
UNIVERSITY OF NORTH DAKOTA.

- Access: SAS Studio
  - https://www.sas.com/en_us/software/on-demand-for-academics.htmlFeatures
- Features
  - Consistent structure
  - Directories
  - Sample data
  - Colors
  - Anticipatory
  - Log
  - Transforming data
  - Great support
- Downsides
  - Not all available in Studio
  - Upload limits



Suggested Uses: All purpose data

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA

Honorable Mentions

Minitab®

Wolfram Mathematica

MathWorks®

python™

Number 1

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA

Overview:

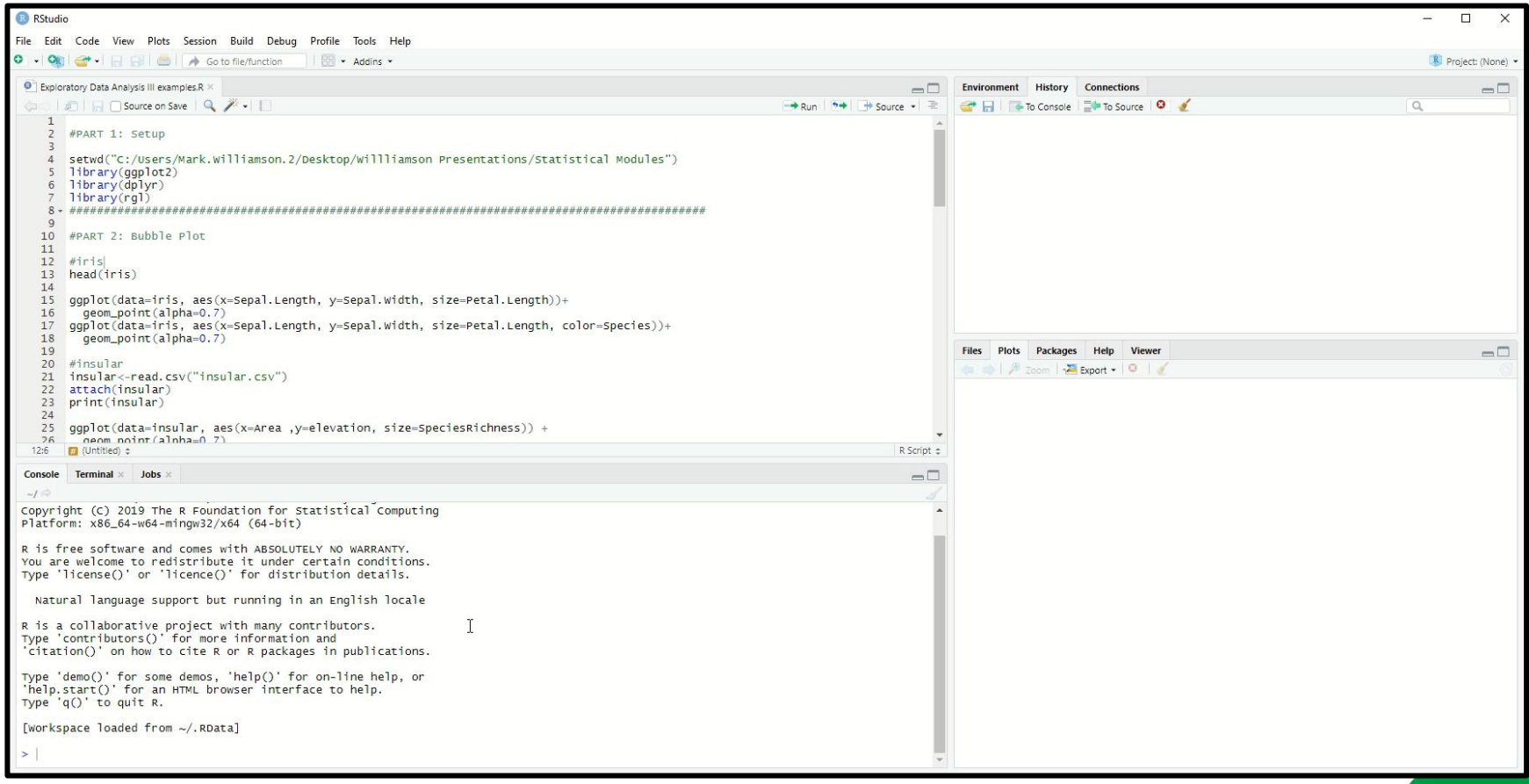
- Programming language and software environment developed by the R Core Team
- 1993

Tool: Swiss-army knife made of Swiss-army knives

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

- Access: Citrix App or direct download (https://cran.r-project.org/)
- Features
  - Command line
  - R commander
  - R studio
  - Packages and support
- Downsides
  - Learning curve
  - No standardized notation



Suggested Uses: All purpose data and weird stuff

Example

Kokanee Fish

- 298 observations
- Length and weight of salmon
- Male and female (will not use)

Summary Handout

- Available online at https://med.und.edu/daccota/berdc-resources.html
  - Under Special Topics Talks -> The Statistical Software Toolkit

General Info	STATA	SPSS	Excel	SAS	R
Access	Citrix Workspace -> Stata-64 app	Citrix Workspace -> SPSS Statistics 26 app	(installed on most computers)	SAS website -> On demand for academics -> SAS Studio	https://cran.r-project.org https://rstudio.com/
Features	Command line; drop-down menus; user-written programs	Drop-down menus; within-software data spreadsheet	Interactive spreadsheet; functions; add-on functionality	Within-software code sheet; consistent structure; high support	Command line; R commander; R Studio; additional packages
Resources	**STATA** • https://www.stata.com/links/resources-for-learning-stata/ • https://data.princeton.edu/stata/ • https://dsc.gmu.edu/files/Advanced_Stata_Skills.pdf • https://www.stata.com/bookstore/statacheatsheets.pdf **SPSS** • https://www.spss-tutorials.com/basics/ • statisticshowto.com/probability-and-statistics/spss-tutorial-beginners/ • https://students.shu.ac.uk/lits/it/documents/pdf/analysing_data_using_spss.pdf • https://rslblissett.com/resources/spss_cheat_sheet/ **XCEL** • https://www.excel-easy.com/data-analysis/analysis-toolpak.html • https://www.excel-easy.com/data-analysis/charts.html • https://edu.gcfglobal.org/en/excel2016/functions/1/ • https://www.customguide.com/cheat-sheet/excel-cheat-sheet.pdf **SAS** • https://documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.3&docsetId=pgmsashome&docsetTarget=home.htm&locale=en • https://www.tutorialspoint.com/sas/index.htm • listendata.com/p/sas-tutorials.html • https://sites.ualberta.ca/~ahamann/teaching/renr480/SAS-Cheat.pdf **R** • statmethods.net/r-tutorial/index.html • https://www.tutorialspoint.com/r/index.htm • https://daviddalpiaz.github.io/appliedstats/applied_statistics.pdf • https://rstudio.com/wp-content/uploads/2016/10/r-cheat-sheet-3.pdf				

Examples	STATA	SPSS	Excel	SAS	R
Summary statistics	Data->Describe Data-> Summary Statistics **OR** summarize ***num_var***	Analyze -> Descriptive Statistics -> Descriptives	=AVERAGE(***num_var***) =MEDIAN(***num_var***) =STDEV.S(***num_var***) ...	PROC UNIVARIATE; var ***num_var;***	summary(***num_var***)
Histogram	Graphics-> Histogram **OR** histogram ***num_var***	Graphs -> Chart Builder -> Histogram	Insert (Charts)-> Histogram	PROC SGPLOT; histogram ***num_var;***	hist(***num_var***)
Boxplot	Graphics-> Box plot **OR** graph box ***num_var***, over(***cat_var***)	Graphs -> Chart Builder -> Boxplot	Insert (Charts)-> Box and Whisker	PROC SGPLOT; vbox ***num_var***/ group=***cat_var;***	plot(***num_var~cat_var***)
Bar plot	Graphics-> Bar Chart **OR** graph bar (mean) ***num_var***, over(***cat_var***)	Graphs -> Chart Builder -> Bar	Insert (Charts)-> Column	PROC SGPLOT; vbarparm category=***cat_var*** treatment=***num_mean;***	***means*** <- c(***mean_cat1, mean_cat2***) barplot(***means***)
Scatterplot	Graphics -> Twoway graph **OR** twoway (scatter ***num_var1 num_var2***)	Graphs -> Chart Builder -> Scatter/Dot	Insert (Charts)-> Scatter	PROC SGPLOT; Scatter y=***num_var1*** x=***num_var2;***	plot(***num_var1, num_var2***)
T-test	Statistics -> Summaries, tables, and tests -> Classical tests of hypotheses -> t tests **OR** ttest ***num_var***, by(***cat_var***)	Analyze -> Compare means-> Independent-Samples T Test	=TTEST(***num_var1, num_var2, tails, type***)	PROC TTEST; var ***num_var;*** class ***cat_var;***	t.test(***num_var~cat_var***)
ANOVA	Statistics-> Linear models and related -> ANOVA/MANOVA -> One-way ANOVA **OR** oneway ***num_var cat_var***	Analyze -> Compare means-> One-Way ANOVA	Data Analysis (add-on) -> Anova: Single Factor	PROC ANOVA; class ***cat_var;*** model ***num_var=cat_var;***	aov(***num_var~cat_var***)
Normal linear regression model	Statistics-> Linear models and related -> Linear regression **OR** regress ***num_var1 num_var2***	Analyze -> Regression-> Linear	Data Analysis (add-on) -> Regression	PROC REG; model ***num_var1= num_var2;***	lm(***num_var1 ~num_var2***)
Logistic regression model	Statistics-> Binary outcomes-> Logistic regression **OR** logit ***binary_var num_var***	Analyze -> Regression-> Binary Logistic	N/A	PROC LOGISTIC; model ***event/trial= num_var2;***	glm(***binary_var ~ num_var***, family=biniomial)
Poisson regression model	Statistics -> Count outcomes-> Poisson regression **OR** Poisson ***count_var num_var***	Analyze -> Regression-> Generalized Linear Models	N/A	PROC GLIMMIX; model ***count_var= num_var*** /dist=Poisson;	glm(***count_var ~ num_var***, family=Poisson)
Generalized linear mixed model	Statistics -> Multilevel mixed-effects models -> Generalized linear model **OR** meglm ***var1 var2*** || ***rand_var_eqn***, family(***distribution***) link(***link_function***)	Analyze-> Mixed Models-> Generalized Linear	N/A	PROC GLIMMIX; class ***cat_var;*** model ***num_var1= num_var2 cat_var rand_var;*** random ***rand_var;***	Package lme4 Lmer(***num_var1~ num_var2 + cat_var*** + (1|***rand_var***)