

## QUANTITATIVE ANALYSIS

---

# INITIAL DATA CLEANING

# AGENDA

1. Missing Data
2. Recoding Data
3. Other Data Operations

# 1 MISSING DATA

## 1. MISSING DATA

---

# “PERFECT” DATA

	make	price	mpg
1	AMC Concord	4,099	22
2	AMC Pacer	4,749	17
3	AMC Spirit	3,799	22
4	Buick Century	4,816	20
5	Buick Electra	7,827	15
6	Buick LeSabre	5,788	18
7	Buick Opel	4,453	26
8	Buick Regal	5,189	20
9	Buick Riviera	10,372	16
10	Buick Skylark	4,082	19
11	Cad. Deville	11,385	14
12	Cad. Eldorado	14,500	14
13	Cad. Seville	15,906	21
14	Chev. Chevette	3,299	29
15	Chev. Impala	5,705	16
16	Chev. Malibu	4,504	22

## 1. MISSING DATA

# WHERE HAVE ALL THE DATA GONE?

	make	price	mpg	rep78	headroom
1	AMC Concord	4,099	22	3	2.5
2	AMC Pacer	4,749	17	3	3.0
3	AMC Spirit	3,799	22	.	3.0
4	Buick Century	4,816	20	3	4.5
5	Buick Electra	7,827	15	4	4.0
6	Buick LeSabre	5,788	18	3	4.0
7	Buick Opel	4,453	26	.	3.0
8	Buick Regal	5,189	20	3	2.0
9	Buick Riviera	10,372	16	3	3.5
10	Buick Skylark	4,082	19	3	3.5
11	Cad. Deville	11,385	14	3	4.0
12	Cad. Eldorado	14,500	14	2	3.5
13	Cad. Seville	15,906	21	3	3.0
14	Chev. Chevette	3,299	29	3	2.5
15	Chev. Impala	5,705	16	4	4.0
16	Chev. Malibu	4,504	22	3	3.5

# FUNDAMENTAL QUESTION:

**ARE MISSING VALUES RELATED TO  
THE UNDERLYING VALUES OF THE DATASET?**


# A THREAT TO GENERALIZABILITY

**MISSING COMPLETELY AT RANDOM**  
MISSINGNESS IS NOT A FUNCTION OF  
EITHER OBSERVED OR UNOBSERVED  
FACTORS IN THE DATASET

**MISSING AT RANDOM**

**NOT MISSING AT RANDOM**

# A THREAT TO GENERALIZABILITY



MISSING COMPLETELY AT RANDOM  
“MCAR”

**Example:**  $n$  respondents have their depression symptoms scored, but a random subset of respondents do not have data on their emergency services use recorded.



# A THREAT TO GENERALIZABILITY

**MISSING COMPLETELY AT RANDOM**

**NOT MISSING AT RANDOM**

**MISSING AT RANDOM  
MISSINGNESS DEPENDS ON  
OBSERVED FACTORS IN  
THE DATASET**

# A THREAT TO GENERALIZABILITY

**Example:**  $n$  respondents have their depression symptoms scored and only those with high depression scores have data on their emergency services use recorded.



MISSING AT RANDOM  
“MAR”

# A THREAT TO GENERALIZABILITY

**MISSING COMPLETELY AT RANDOM**

**MISSING AT RANDOM**

**NOT MISSING AT RANDOM  
MISSINGNESS DEPENDS ON  
UNOBSERVED FACTORS IN THE  
DATASET**

# A THREAT TO GENERALIZABILITY

**Example:**  $n$  respondents have their depression symptoms scored and are assessed for suicidality, but only respondents with high suicidality scores have their suicidality data entered in the dataset.



NOT MISSING AT RANDOM

# AN OUNCE OF PREVENTION...

- ▶ When you design your study and conduct it, work to ensure that missing data is minimized to every extent possible.
  - ▶ However, it is ethically important that missingness is allowed and methodologically important that it is anticipated - how will you account for refusals or "I don't know" answers?
- ▶ Can you predict ahead of time who will refuse certain questions?
  - ▶ Be thorough in your design - try to eliminate unobserved sources of missingness, and try to collect data that may help predict missingness.

# ... AND A POUND OF CURE

- ▶ When (and not if!) you have missing data, begin by “declaring” data missing.
- ▶ Then look at the percentage of missing cases in a given variable - a quick rule of thumb is that we would like to have  $< 5\%$  of  $n$  missing values per variable.
- ▶ There are advanced techniques for both assessing MCAR vs MAR and for “recovering” missing data (beyond scope of class).

## 1. MISSING DATA

---

# MISSING DATA IN THE WILD

```
. use autoMissing.dta  
  
. tabulate rep78
```

Repair Record 1978	Freq.	Percent	Cum.
1	2	2.90	2.90
2	8	11.59	14.49
3	30	43.48	57.97
4	18	26.09	84.06
5	11	15.94	100.00
Total	69	100.00	

## 1. MISSING DATA

# MISSING DATA IN THE WILD

```
. use autoMissing.dta  
  
. tabulate rep78
```

**WE KNOW  
N = 74**

Repair Record 1978	Freq.	Percent	Cum.
1	2	2.90	2.90
2	8	11.59	14.49
3	30	43.48	57.97
4	18	26.09	84.06
5	11	15.94	100.00
Total	69	100.00	



# MISSING DATA IN THE WILD

```
. tabulate rep78, missing
```

Repair Record 1978	Freq.	Percent	Cum.
1	2	2.70	2.70
2	8	10.81	13.51
3	30	40.54	54.05
4	18	24.32	78.38
5	11	14.86	93.24
.	5	6.76	100.00
Total	74	100.00	

1. MISSING DATA

# MISSING DATA IN THE WILD

```
. tabulate rep78, missing
```

Repair Record 1978	Freq.	Percent	Cum.
1	2	2.70	2.70
2	8	10.81	13.51
3	30	40.54	54.05
4	18	24.32	78.38
5	11	14.86	93.24
.	5	6.76	100.00
Total	74	100.00	

1. MISSING DATA

MISSING DATA IN THE WILD

```
. tabulate mpg, missing
```

Mileage (mpg)	Freq.	Percent	Cum.
12	2	2.70	2.70
14	6	8.11	10.81
15	2	2.70	13.51
17	4	5.41	18.92
18	9	12.16	31.08
19	8	10.81	41.89
20	3	4.05	45.95
21	5	6.76	52.70
22	5	6.76	59.46
23	3	4.05	63.51
24	4	5.41	68.92
25	5	6.76	75.68
26	3	4.05	79.73
28	3	4.05	83.78
29	1	1.35	85.14
31	1	1.35	86.49
34	1	1.35	87.84
35	2	2.70	90.54
41	1	1.35	91.89
unknown	4	5.41	97.30
not measured	2	2.70	100.00
Total	74	100.00	

1. MISSING DATA

MISSING DATA IN THE WILD

```
. tabulate mpg, missing
```

Mileage (mpg)	Freq.	Percent	Cum.
12	2	2.70	2.70
14	6	8.11	10.81
15	2	2.70	13.51
17	4	5.41	18.92
18	9	12.16	31.08
19	8	10.81	41.89
20	3	4.05	45.95
21	5	6.76	52.70
22	5	6.76	59.46
23	3	4.05	63.51
24	4	5.41	68.92
25	5	6.76	75.68
26	3	4.05	79.73
28	3	4.05	83.78
29	1	1.35	85.14
31	1	1.35	86.49
34	1	1.35	87.84
35	2	2.70	90.54
41	1	1.35	91.89
unknown	4	5.41	97.30
not measured	2	2.70	100.00
Total	74	100.00	

1. MISSING DATA

MISSING DATA IN THE WILD

```
. tabulate mpg, missing nolabel
```

Mileage (mpg)	Freq.	Percent	Cum.
12	2	2.70	2.70
14	6	8.11	10.81
15	2	2.70	13.51
17	4	5.41	18.92
18	9	12.16	31.08
19	8	10.81	41.89
20	3	4.05	45.95
21	5	6.76	52.70
22	5	6.76	59.46
23	3	4.05	63.51
24	4	5.41	68.92
25	5	6.76	75.68
26	3	4.05	79.73
28	3	4.05	83.78
29	1	1.35	85.14
31	1	1.35	86.49
34	1	1.35	87.84
35	2	2.70	90.54
41	1	1.35	91.89
.a	4	5.41	97.30
.b	2	2.70	100.00
Total	74	100.00	

1. MISSING DATA

MISSING DATA IN THE WILD

```
. tabulate mpg, missing nolabel
```

Mileage (mpg)	Freq.	Percent	Cum.
12	2	2.70	2.70
14	6	8.11	10.81
15	2	2.70	13.51
17	4	5.41	18.92
18	9	12.16	31.08
19	8	10.81	41.89
20	3	4.05	45.95
21	5	6.76	52.70
22	5	6.76	59.46
23	3	4.05	63.51
24	4	5.41	68.92
25	5	6.76	75.68
26	3	4.05	79.73
28	3	4.05	83.78
29	1	1.35	85.14
31	1	1.35	86.49
34	1	1.35	87.84
35	2	2.70	90.54
41	1	1.35	91.89
.a	4	5.41	97.30
.b	2	2.70	100.00
Total	74	100.00	

1. MISSING DATA

MISSING DATA IN THE WILD

```
. tabulate trunk
```

Trunk space (cu. ft.)	Freq.	Percent	Cum.
-1	4	5.41	5.41
5	1	1.35	6.76
6	1	1.35	8.11
7	3	4.05	12.16
8	5	6.76	18.92
10	5	6.76	25.68
11	8	10.81	36.49
12	3	4.05	40.54
13	4	5.41	45.95
14	4	5.41	51.35
15	5	6.76	58.11
16	12	16.22	74.32
17	8	10.81	85.14
18	1	1.35	86.49
20	6	8.11	94.59
21	2	2.70	97.30
22	1	1.35	98.65
23	1	1.35	100.00
Total	74	100.00	

1. MISSING DATA

MISSING DATA IN THE WILD

```
. tabulate trunk
```

Trunk space (cu. ft.)	Freq.	Percent	Cum.
-1	4	5.41	5.41
5	1	1.35	6.76
6	1	1.35	8.11
7	3	4.05	12.16
8	5	6.76	18.92
10	5	6.76	25.68
11	8	10.81	36.49
12	3	4.05	40.54
13	4	5.41	45.95
14	4	5.41	51.35
15	5	6.76	58.11
16	12	16.22	74.32
17	8	10.81	85.14
18	1	1.35	86.49
20	6	8.11	94.59
21	2	2.70	97.30
22	1	1.35	98.65
23	1	1.35	100.00
Total	74	100.00	



1. MISSING DATA

MISSING DATA IN THE WILD

. tabulate turn

Turn Circle (ft.)	Freq.	Percent	Cum.
31	1	1.35	1.35
32	1	1.35	2.70
33	2	2.70	5.41
34	6	8.11	13.51
35	6	8.11	21.62
36	9	12.16	33.78
37	4	5.41	39.19
39	1	1.35	40.54
40	6	8.11	48.65
41	4	5.41	54.05
42	7	9.46	63.51
43	12	16.22	79.73
44	3	4.05	83.78
45	3	4.05	87.84
46	3	4.05	91.89
48	2	2.70	94.59
51	1	1.35	95.95
999	3	4.05	100.00
Total	74	100.00	

1. MISSING DATA

MISSING DATA IN THE WILD

```
. tabulate turn
```

Turn Circle (ft.)	Freq.	Percent	Cum.
31	1	1.35	1.35
32	1	1.35	2.70
33	2	2.70	5.41
34	6	8.11	13.51
35	6	8.11	21.62
36	9	12.16	33.78
37	4	5.41	39.19
39	1	1.35	40.54
40	6	8.11	48.65
41	4	5.41	54.05
42	7	9.46	63.51
43	12	16.22	79.73
44	3	4.05	83.78
45	3	4.05	87.84
46	3	4.05	91.89
48	2	2.70	94.59
51	1	1.35	95.95
999	3	4.05	100.00
Total	74	100.00	

# MISSING DATA IN THE WILD

HETENURE

HETENURE

Location:

29-30 (width: 2; decimal: 0)

Variable Type:

numeric

Question:

ARE YOUR LIVING QUARTERS...

<i>Value</i>	<i>Label</i>	<i>Unweighted Frequency</i>	<i>%</i>
-1	BLANK	14852	9.8 %
1	OWNED OR BEING BOUGHT BY A HH MEMBER	94178	62.2 %
2	RENTED FOR CASH	40568	26.8 %
3	OCCUPIED WITHOUT PAYMENT OF CASH RENT	1710	1.1 %

Based upon 151308 valid cases out of 151308 total cases.

- Mean: 1.09
- Median: 1.00
- Mode: 1.00
- Minimum: -1.00
- Maximum: 3.00
- Standard Deviation: 0.83

# MISSING VALUES IN STATA

```
misstable summarize [varlist]
```

```
. misstable summarize
```

				Obs<.		
Variable	Obs=.	Obs>.	Obs<.	Unique values	Min	Max
mpg		6	68	19	12	41
rep78	5		69	5	1	5

# MISSING VALUES IN STATA

```
misstable tree [varlist], [frequency]
```

```
. misstable tree, frequency
```

MAY BE  
"MCAR"

Nested pattern of missing values			
mpg	rep78	trunk	turn
-----			
6	0	0	0
			0
			0
			0
	6	0	0
			0
			0
			0
			0
			0
68	5	0	6
			0
			0
			0
	63	5	2
			3
			0
			4
			1
			58
-----			
number missing listed first)			

## 1. MISSING DATA

---

# MISSING VALUES IN STATA

- `use` <http://www.stata-press.com/data/r13/studentsurvey>, `clear`
- `misstable tree`, `frequency`

Nested pattern of missing values

dept	age	female
------	-----	--------

9	3	3
		0
	6	0
		6
116	0	0
		0
	116	0
		116

(number missing listed first)

MAY BE  
"MAR"

# FUNDAMENTAL QUESTION:

**AT THIS STAGE IN THE RESEARCH  
PROCESS, DO YOU NEED TO KNOW  
WHY DATA ARE MISSING?**

# MISSING VALUES IN STATA

1. The period ( `.` ) used to represent missing data; can be combined with letters a through z
2. (all positive values)  $< . < \infty$
3. (all positive values)  $< . < .a < .b < \dots < .z < \infty$



# RECODING MISSING VALUES

```
recode varname oldVal = newVal [oldVal = newVal]
```

- `use autoMissing.dta`
- `recode turn (999 = .)`  
(turn: 3 changes made)
- `recode trunk (-1 = .a)`  
(trunk: 4 changes made)
- `recode mpg (.a/.b = .)`  
(mpg: 6 changes made)

# RECODING MISSING VALUES

```
recode varname oldVal = newVal [oldVal = newVal]
```

- `use autoMissing.dta`
- `recode turn (999 = .)`  
(turn: 3 changes made)
- `recode trunk (-1 = .a)`  
(trunk: 4 changes made)
- `recode mpg (.a/.b = .)`  
(mpg: 6 changes made)

## Tips:

- ▶ Apply value labels if using .a, .b, etc.
- ▶ Forward slash allows for a range of values to be recoded - i.e. all values between .a and .b

# MISSING DATA WORKFLOW

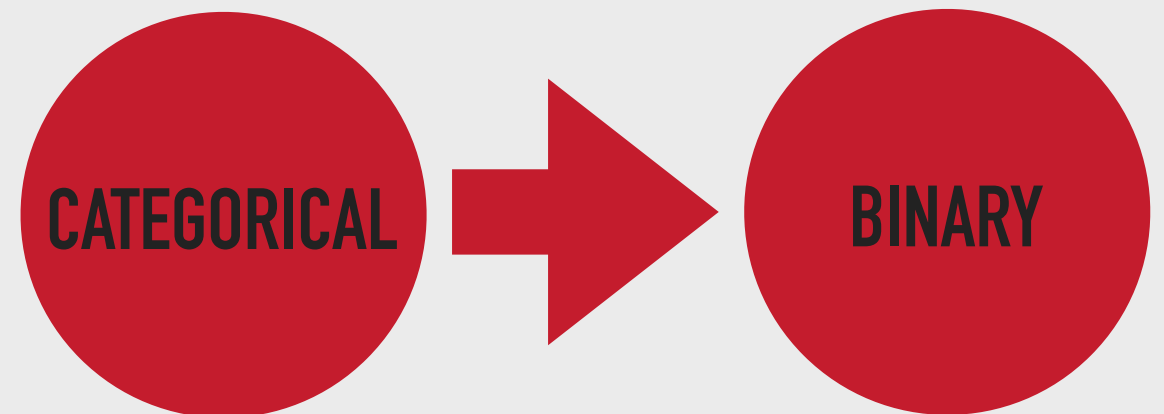
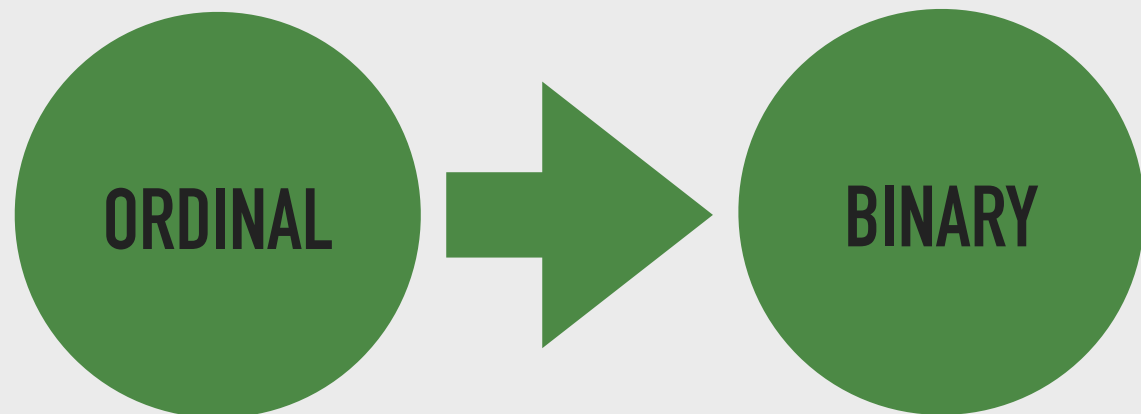
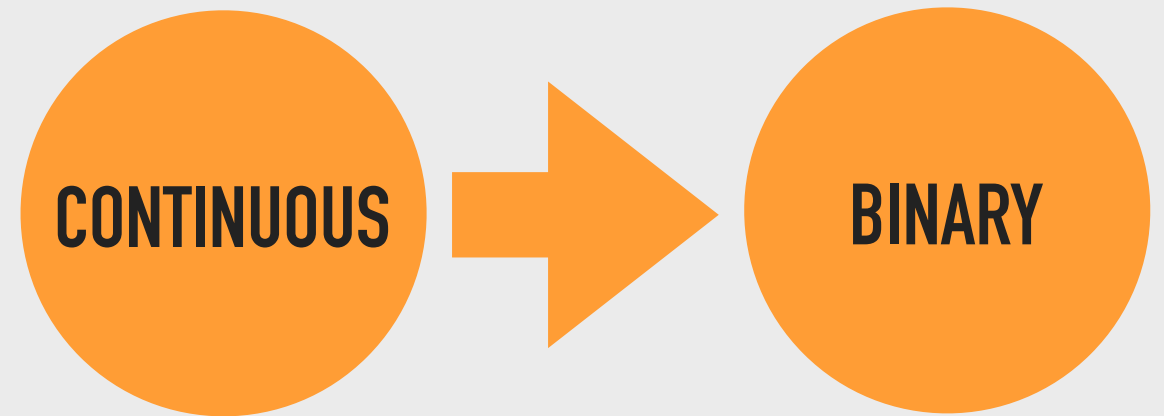
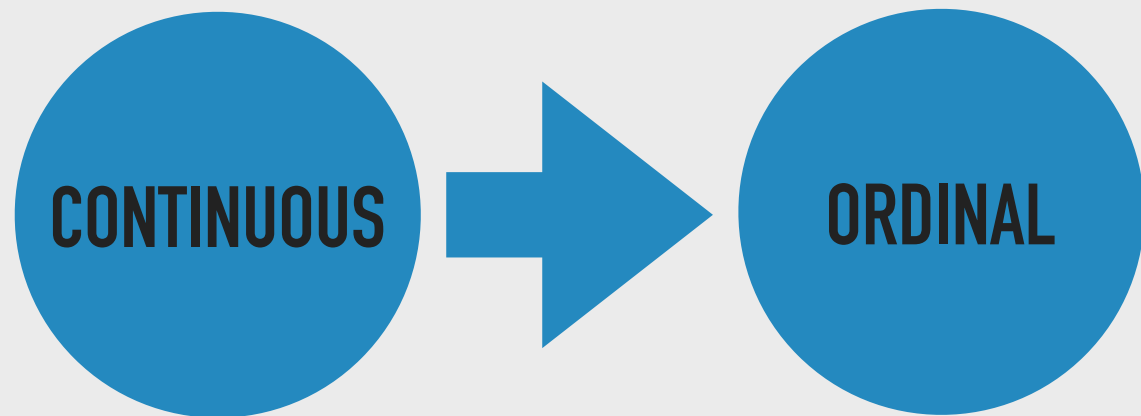
1. Determine if missing data need to be recoded within variables.
2. Determine if you want to preserve different types of missing data, or you only want to identify missing values without differentiating between them.
3. Recode values accordingly using the `recode` command.
4. Use `misstable` commands to look for patterns in missingness.

# 2 RECODING DATA

## 2. RECODING DATA

---

# WHY RECODE?



# STEP 1: GENERATE A NEW VARIABLE

`generate newVar = oldVar`

- `generate mpg0rd = mpg`

# STEP 1: GENERATE A NEW VARIABLE

`generate newVar = oldVar`

- `generate mpg0rd = mpg`

### Tips:

- ▶ Use short, intuitive names (ex - mpg0rd for the ordinal version of mpg).
- ▶ Keep names short.

## 2. RECODING DATA

---

# STEP 2: RECODE VALUES

```
recode varname oldVal = newVal [oldVal = newVal]
```

```
▪ recode mpg0rd (12/19 = 1) (20/29 = 2) (30/35 = 3) (41 = 4)
```



# STEP 2: RECODE VALUES

```
recode varname oldVal = newVal [oldVal = newVal]
```

```
▪ recode mpg0rd (12/19 = 1) (20/29 = 2) (30/35 = 3) (41 = 4)
```

## Tips:

- ▶ Wrap values in () for clarity and readability
- ▶ Forward slash allows for a range of values to be recoded - i.e. all values between 12 and 19 or all values between 20 and 29

# STEP 2: RECODE VALUES

- `generate mpgHigh = mpg`
- `recode mpgHigh (12/29 = 0) (30/41 = 1)`
  
- `generate mpgBin = mpgOrd`
- `recode mpgBin (1/3 = 0) (4 = 1)`

## 2. RECODING DATA

---

# STEP 3: DEFINE VALUE LABELS

```
label define lblName val [""]lbl[""] [["]lbl[""]]
```

```
. label define mpg0rdVals 1 "< 20" 2 "20 to 29" 3 "30 to 39" 4 ">= 40"
```

# STEP 3: DEFINE VALUE LABELS

```
label define lblName val [""]lbl[""] [[""]lbl[""]]
```

```
. label define mpg0rdVals 1 "< 20" 2 "20 to 29" 3 "30 to 39" 4 ">= 40"
```

### Tips:

- ▶ Keep label name short but descriptive
- ▶ Keep values labels short but descriptive - what does this value represent or measure?

## 2. RECODING DATA

---

# STEP 4: LABEL VALUES

`label values varname lblName`

- `label values mpg0rd mpg0rdVals`

## 2. RECODING DATA

---

# STEP 5: LABEL VARIABLE

`label variable varname "labelText"`

- `label variable mpg0rd "ordinal version of mpg"`

# STEP 5: LABEL VARIABLE

`label variable varname "labelText"`

- `label variable mpg0rd "ordinal version of mpg"`

### Tips:

- ▶ Keep variable labels short but descriptive - what does this variable represent or measure?

# 3 OTHER DATA OPERATIONS



# DROPPING VARIABLES

`drop varlist`

- `drop mpg trunk`

# DROPPING VARIABLES

`drop varlist`

- `drop mpg trunk`

## Tips:

- ▶ Use `drop` for removing *small* numbers of variables from datasets or for removing variables from *smaller* datasets.

# KEEPING VARIABLES

`keep` *varlist*

- `keep` mpg trunk

# KEEPING VARIABLES

`keep varlist`

- `keep mpg trunk`

## Tips:

- ▶ Use `keep` for removing *large* numbers of variables from datasets or for removing variables from *larger* datasets.

# REORDERING VARIABLES

`order varlist [, first last before(varname) after(varname)]`

- `order mpg trunk, first`
- `order mpg trunk, last`
- `order mpg trunk, before(weight)`
- `order mpg trunk, after(weight)`

# SAVING DATA

```
save ["filename.dta"], replace
```

```
. save "mpgAug29.dta", replace
```

# SAVING DATA

```
save ["filename.dta"], replace
```

```
. save "mpgAug29.dta", replace
```

## Tips:

- ▶ Use short filenames
- ▶ Avoid spaces and special characters in filenames
- ▶ Use snake\_case or camelCase to include multiple words
- ▶ Imply a logical order (use dates, numbers, etc.)
- ▶ Never call things final!

# DOCUMENT DETAILS

Document produced by [Christopher Prener, Ph.D](#) for the Saint Louis University course SOC 5050: QUANTITATIVE ANALYSIS - APPLIED INFERENTIAL STATISTICS. See the [course wiki](#) and the repository [README.md](#) file for additional details.



This work is licensed under a [Creative Commons Attribution 4.0 International License](#).