Overview Categorical to Categorical Quantitative to Quantitative to Categorical

# Data cleaning and new variable creation, formalized

Shannon Pileggi

**STAT 330** 

STAT 330: Lecture 9 1 / 21

Overview Categorical to Categorical Quantitative to Quantitative to Categorical

#### **OUTLINE**

Overview

Categorical to Categorical

Quantitative to Quantitative

Quantitative to Categorical

STAT 330: Lecture 9 2 / 21

Overview Categorical to Categorical Quantitative to Quantitative Quantitative to Categorical

•000 00000 00000

# Steps to data cleaning/new variable creation

#### Step 1: Get to know your data.

- a. Identify existing values and/or unusual values
- b. Identify if missing values are present
- c. Identify how many observations had the unusual values
- d. Identify which observations had the unusual values

#### Step 2: Create clean new variables with desired result.

Over-writing existing variable values could be problematic down the line

Step 3: Verify that coding was done correctly.

# What's wrong with PROC PRINT for verification?

- ▶ Viewing your data with PROC PRINT, or otherwise like in the data table viewer, is prone to human error. Especially with large data sets, it would be very time consuming to visually inspect *all* the data to verify correctness.
- ► Too much PROC PRINT eats SAS's memory! (Think printing thousands of observations, multiple times...) SAS will dramatically slow down, and maybe even crash on you.
- ► If you get caught where you have used too much PROC PRINT and SAS is slow, try:
  - ▶ the special submit F9
  - ► close SAS and open it again

4 章 ト 章 ト 章 シスペ 3 / 21 STAT 330: Lecture 9 4 / 21

 ✓□ト ✓
 ●ト ✓

Overview

Categorical to Categorical

Quantitative to Quantitative

Quantitative to Categorical

DD00 DD711

# Limiting PROC PRINT

You can use PROC PRINT to get a quick glance at your data, but limit the observations printed.

obs =

specifies the *last* observation that SAS processes in a data set.

PROC PRINT DATA = mydata (obs=10); RUN; prints the first 10 observations

PROC PRINT DATA = mydata (firstobs=5 obs=10); RUN; prints observations 5 through 10

STAT 330: Lecture 9

Overview

Categorical to Categorical

Quantitative to Quantitative

Quantitative to Categorical

5 / 21

Categorical to Categorical

Quantitative to Quantitative

Quantitative to Categorica

Overview

Categorical to Categorical

Quantitative to Quantitative

Quantitative to Categorical

## On your own:

For each of the following questions, identify the scenario as: (1) categorical to categorical, (2) quantitative to quantitative, or (3) quantitative to categorical.

- Lab 4 Q6: Create a new variable called GPA\_clean that is a copy of the GPA variable. Re-code the unusual values missing.
- Lab 4 Q8: Create a new variable called prev\_stats which has a value of yes if students have previous experience with statistics (Q03a = 0) and a value of no if the student does not have previous experience with statistics (Q03a = 1).
- Lab 4 Q11: Create a new variable called class that classifies students as "lower" class (first years and second years) and "upper" class (third years, fourth years, etc.).
- Lab 4 Q13: Use the GPA\_clean variable to create a new variable called honors that classifies students according to their current GPA; students who do not yet achieve honors should be classified as "none".

 4 □ > 4 ∅ > 4 ½ > 4 ½ > ½ 

 6 / 21

STAT 330: Lecture 9

Overview

Quantitative to Quantitative

Quantitative to Categorical

### Step 1: Get to know your data.

Categorical to Categorical

<u>Lab 4 Q11:</u> Create a new variable called class that classifies students as "lower" class (first years and second years) and "upper" class (third years, fourth years, etc.).

- a. Identify existing values and/or unusual values
- b. Identify if missing values are present
- c. Identify how many observations had the unusual values
- d. Identify which observations had the unusual values

PROC FREQ DATA = survey ;
TABLES Q02 ;
RUN ;

\_ SAS Code -

 Q02
 Frequency
 Percent Frequency
 Cumulative Frequency
 Cumulative Frequency

 First year
 9
 25.71
 9
 25.71

 Fourth year
 1
 2.86
 10
 28.57

 Second year
 17
 48.57
 27
 77.14

 Third year
 8
 22.86
 35
 100.00

◆□▶◆□▶◆□▶◆□▶ □ 900

7 / 21

STAT 330: Lecture 9

Overview Categorical to Categorical

Quantitative to Quantitative

Quantitative to Categorical

# Step 2: Create clean new variables with desired result.

<u>Lab 4 Q11:</u> Create a new variable called class that classifies students as "lower" class (first years and second years) and "upper" class (third years, fourth years, etc.).

SAS Code

IF Q02 IN ("First year", "Second year") THEN class = "lower";

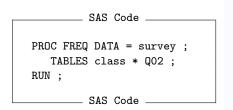
ELSE class = "upper";

SAS Code

 Overview Categorical to Categorical Quantitative to Quantitative to Categorical

# Step 3: Verify that coding was done correctly.

<u>Lab 4 Q11:</u> Create a new variable called class that classifies students as "lower" class (first years and second years) and "upper" class (third years, fourth years, etc.).



Percent Row Pct Col Pct	Table of class by Q02								
			Q02						
	class	First year	Fourth year	Second year	Third year	Total			
	lower	9 25.71 34.62 100.00	0.00 0.00 0.00	17 48.57 65.38 100.00	0.00 0.00 0.00	26 74.29			
	upper	0.00 0.00 0.00	2.86 11.11 100.00	0.00 0.00 0.00	22.86 88.89 100.00	9 25.71			
	Total	9 25.71	1 2.86	17 48.57	8 22.86	35 100.00			

10 / 21

STAT 330: Lecture 9

Overview Categorical to Categorical Quantitative to Quantitative Quantitative to Categorical

# Step 3: Verify that coding was done correctly, better.

<u>Lab 4 Q11:</u> Create a new variable called class that classifies students as "lower" class (first years and second years) and "upper" class (third years, fourth years, etc.).

#### Step 3: Verify that coding was done correctly.

PROC FREQ DATA = survey;
TABLES class \* QO2
/ LIST MISSING;
RUN;

\_\_\_ SAS Code \_

	The FREQ Procedure						
class	Q02	Frequency	Percent	Cumulative Frequency	Cumulative Percent		
lower	First year	9	25.71	9	25.71		
lower	Second year	17	48.57	26	74.29		
upper	Fourth year	1	2.86	27	77.14		
upper	Third year	8	22.86	35	100.00		

Overview Categorical to Categorical Quantitative to Quantitative to Categorical

OOOO Quantitative to Quantitative to Categorical

Ovorviou

Categorical to Categorical

Quantitative to Quantitative

Quantitative to Categorical

 ★□▶★□▶★□▶★□▶★□
 ●
 \*○○

 STAT 330: Lecture 9
 12 / 21

 Overview

Categorical to Categorical

Quantitative to Quantitative

Quantitative to Categorical

## Step 1: Get to know your data.

Lab 4 Q6: Create a new variable called GPA\_clean that is a copy of the GPA variable. Re-code the unusual values that you identified in the previous question to missing.

- a. Identify existing values and/or unusual values
- b. Identify if missing values are present
- c. Identify how many observations had the unusual values
- d. Identify which observations had the unusual values

\_\_\_\_\_ SAS Code \_\_\_\_\_ PROC MEANS DATA = work.survey2 VAR Q04; RUN; PROC UNIVARIATE DATA = work.survey2; VAR Q04; RUN; PROC FREQ DATA = work.survey2; TABLES Q04; RUN; PROC PRINT DATA = work.survey2; WHERE Q04 > 4; RUN; SAS Code STAT 330: Lecture 9 13 / 21

Categorical to Categorical Quantitative to Categorical Overview Quantitative to Quantitative

# Step 2: Create clean new variables with desired result.

Lab 4 Q6: Create a new variable called GPA\_clean that is a copy of the GPA variable. Re-code the unusual values that you identified in the previous question to missing.

```
_____ SAS Code _____
GPA_clean = Q04 ;
IF GPA_clean > 4 THEN GPA_clean = . ;
          _____ SAS Code ____
```

◆□ → ◆□ → ◆ き → ◆ き → り へ ○ STAT 330: Lecture 9 14 / 21

Quantitative to Quantitative

Categorical to Categorical Quantitative to Quantitative Quantitative to Categorical

# Step 3: Verify that coding was done correctly.

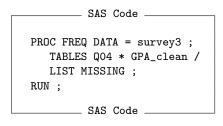
Lab 4 Q6: Create a new variable called GPA\_clean that is a copy of the GPA variable. Re-code the unusual values that you identified in the previous question to missing.

\_\_\_ SAS Code \_\_\_ PROC MEANS DATA = survey3 ; VAR QO4 GPA\_clean ; RUN ; \_\_ SAS Code \_\_\_\_

Variable N Mean Std Dev Minimum Maximum 35 4.8446571 3.1000485 2.0000000 10.0000000 GPA clean 26 3.0624231 0.4726480 2.0000000 3.9000000

# Step 3: Verify that coding was done correctly, better.

Lab 4 Q6: Create a new variable called GPA\_clean that is a copy of the GPA variable. Re-code the unusual values that you identified in the previous question to missing.



Categorical to Categorical

Q04	GPA_clean	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2	2	2	5.71	2	5.71
2.3	2.3	1	2.86	3	8.57
2.589	2.589	1	2.86	4	11.43
2.83	2.83	1	2.86	5	14.29
2.84	2.84	1	2.86	6	17.14
3	3	8	22.86	14	40.00
3.1	3.1	1	2.86	15	42.86
3.167	3.167	1	2.86	16	45.71
3.2	3.2	1	2.86	17	48.57
3.204	3.204	1	2.86	18	51.43
3.233	3.233	1	2.86	19	54.29
3.3	3.3	2	5.71	21	60.00
3.5	3.5	1	2.86	22	62.86
3.69	3.69	1	2.86	23	65.71
3.7	3.7	1	2.86	24	68.57
3.77	3.77	1	2.86	25	71.43
3.9	3.9	1	2.86	26	74.29
9.99		6	17.14	32	91.43
10		3	8.57	35	100.00

The EDEO Broodure

Quantitative to Categorical

Overview

 Overview Categorical to Categorical Quantitative to Quantitative to Categorical

### Step 1: Get to know your data.

<u>Lab 4 Q13</u>: Use the GPA\_clean variable to create a new variable called honors that classifies students according to their current GPA; students who do not yet achieve honors should be classified as "none".

- a. Identify existing values and/or unusual values
- b. Identify if missing values are present
- c. Identify how many observations had the unusual values
- d. Identify which observations had the unusual values

erview Categorical to Categorical Quantitative to Quantitative **Quantitative Quantitative <b>Quantitative Quantitative Quantitative Quantitative <b>Quantitative Quantitative Quantitative Quantitative <b>Quantitative Quantitative Quantitative Quantitative <b>Quanti** 

# Step 2: Create clean new variables with desired result, method 1.

<u>Lab 4 Q13:</u> Use the GPA\_clean variable to create a new variable called honors that classifies students according to their current GPA; students who do not yet achieve honors should be classified as "none".

```
LENGTH honors $ 20 ;

IF GPA_clean = . THEN honors = "" ;

ELSE IF GPA_clean >= 3.85 THEN honors = "Summa cum laude" ;

ELSE IF 3.70 <= GPA_clean < 3.85 THEN honors = "Magna cum laude" ;

ELSE IF 3.50 <= GPA_clean < 3.70 THEN honors = "Cum laude" ;

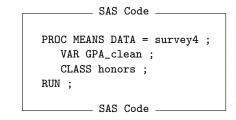
ELSE honors = "none" ;
```

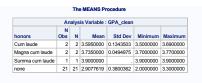
 Overview
 Categorical to Categorical
 Quantitative to Quantitative
 Quantitative to Categorical

 0000
 00000
 0000€0

# Step 3: Verify that coding was done correctly, method 1.

Lab 4 Q13: Use the GPA\_clean variable to create a new variable called honors that classifies students according to their current GPA; students who do not yet achieve honors should be classified as "none".





◆ロト ◆御 ト ◆ 恵 ト ◆ 恵 ・ 夕 Q ○

4 D > 4 B > 4 E > 4 E > 900

STAT 330: Lecture 9

Overview Categorical to Categorical Quantitative to Quantitative to Categorical

OOO OOO OOOO

# Step 3: Verify that coding was done correctly, method 2.

Lab 4 Q13: Use the GPA\_clean variable to create a new variable called honors that classifies students according to their current GPA; students who do not yet achieve honors should be classified as "none".

\_ SAS Code \_\_\_

PROC FREQ DATA = survey4 ;
 TABLES honors \* GPA\_clean /
 LIST MISSING;
RUN ;

SAS Code \_

honors	GPA_clean	Frequency	Percent	Cumulative Frequency	Cumulative Percent
		9	25.71	9	25.71
Cum laude	3.5	- 1	2.86	10	28.57
Cum laude	3.69	- 1	2.86	- 11	31.43
Magna cum laude	3.7	1	2.86	12	34.29
Magna cum laude	3.77	1	2.86	13	37.14
Summa cum laude	3.9	1	2.86	14	40.00
none	2	2	5.71	16	45.71
none	2.3	1	2.86	17	48.57
none	2.589	1	2.86	18	51.43
none	2.83	1	2.86	19	54.29
none	2.84	1	2.86	20	57.14
none	3	8	22.86	28	80.00
none	3.1	1	2.86	29	82.86
none	3.167	1	2.86	30	85.71
none	3.2	1	2.86	31	88.57
none	3.204	1	2.86	32	91.43
none	3.233	1	2.86	33	94.29
none	3.3	2	5.71	35	100.00



STAT 330: Lecture 9

21 / 21

