Introduction to SAS, Working with categorical variables

Steve Simon

Created: 2021-05-30

Overview

- Frequency counts
- Convert string to numeric
- Labels for number codes
- Drawing bar charts
- Converting continuous to categorical
- Modifying categorical variables
- Crosstabulations

SAS code: Documentation header

```
m04-5507-simon-categorical-variables
author: Steve Simon
Date created: 2018-10-22

Purpose: To illustrate how to work with datasets
with mostly continuous variables.

License: public domain;
```

SAS code: Tell SAS where to find and store things

```
options papersize=(6in 4in);
* This needed to have the output fit on
PowerPoint;

%let path=q:/introduction-to-sas;

ods pdf
  file="&path/results/m04-5507-simon-
categorical.pdf";

filename raw_data
  "&path/data/titanic_v00.txt";

libname perm
  "&path/data";
```

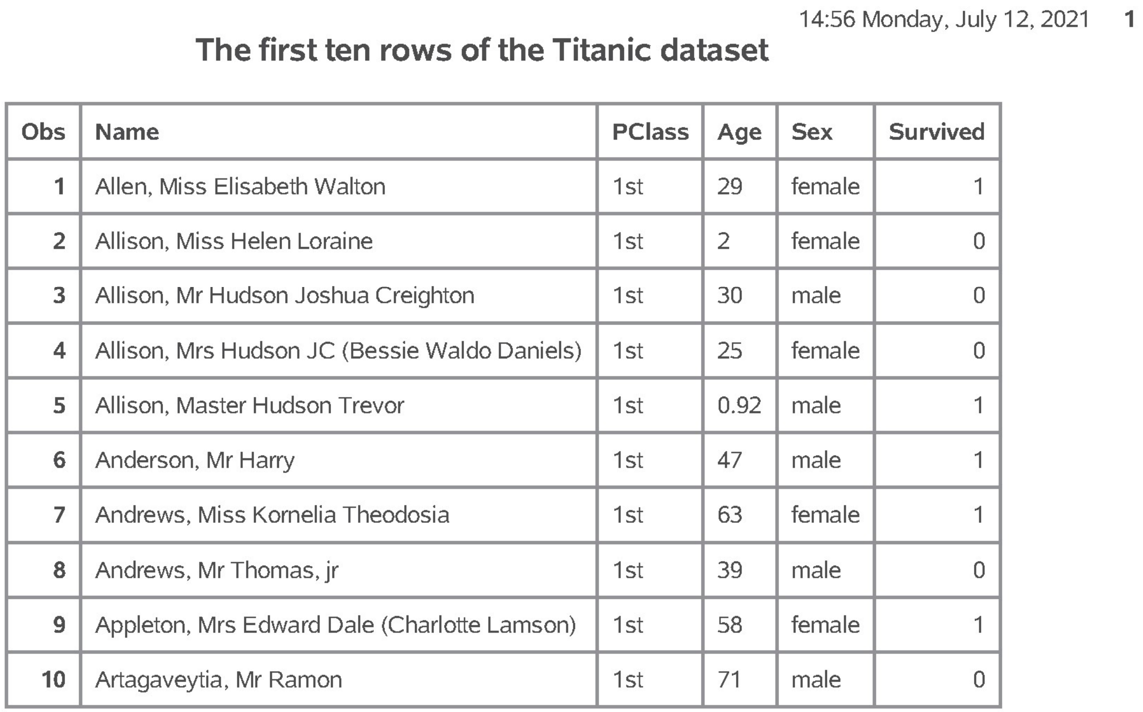
SAS code: Reading using proc import

```
proc import
    datafile=raw_data
    out=perm.titanic
    dbms=dlm
    replace;
  delimiter='09'x;
  getnames=yes;
run;
```

SAS code: Print the first ten lines

```
proc print
    data=perm.titanic(obs=10);
  title1 "The first ten rows of the Titanic
dataset";
run;
```

SAS output: Print the first ten lines



14:56 Monday, July 12, 2021 **1**

The first ten rows of the Titanic dataset

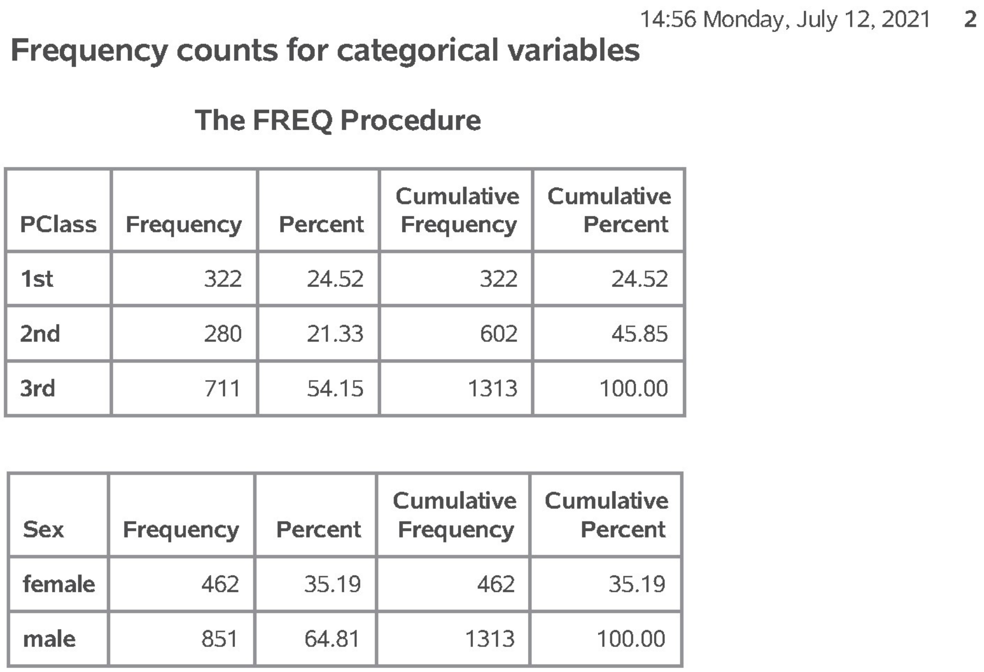
Obs	Name	PClass	Age	Sex	Survived
1	Allen, Miss Elisabeth Walton	1st	29	female	1
2	Allison, Miss Helen Loraine	1st	2	female	0
3	Allison, Mr Hudson Joshua Creighton	1st	30	male	0
4	Allison, Mrs Hudson JC (Bessie Waldo Daniels)	1st	25	female	0
5	Allison, Master Hudson Trevor	1st	0.92	male	1
6	Anderson, Mr Harry	1st	47	male	1
7	Andrews, Miss Kornelia Theodosia	1st	63	female	1
8	Andrews, Mr Thomas, jr	1st	39	male	0
9	Appleton, Mrs Edward Dale (Charlotte Lamson)	1st	58	female	1
10	Artagaveytia, Mr Ramon	1st	71	male	0

SAS Output

SAS code: Counts, proc freq

```
proc freq
    data=perm.titanic;
  table PClass Sex Survived;
  title1 "Frequency counts for categorical
variables";
run;
```

SAS output: Counts, proc freq



14:56 Monday, July 12, 2021 2

Frequency counts for categorical variables

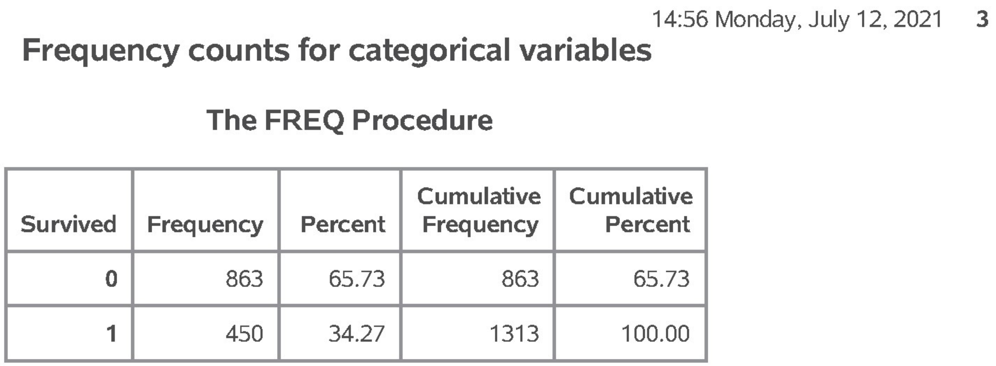
The FREQ Procedure

PClass	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1st	322	24.52	322	24.52
2nd	280	21.33	602	45.85
3rd	711	54.15	1313	100.00

Sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
female	462	35.19	462	35.19
male	851	64.81	1313	100.00

SAS Output

SAS output: Counts, proc freq



14:56 Monday, July 12, 2021 3

Frequency counts for categorical variables

The FREQ Procedure

Survived	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	863	65.73	863	65.73
1	450	34.27	1313	100.00

SAS Output

Break #1

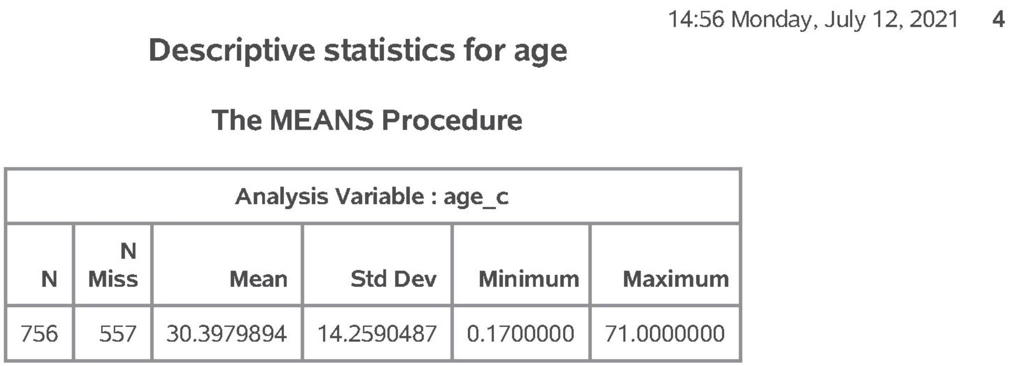
- What have you learned
  - Frequency counts
- What's coming next
  - Convert string to numeric

SAS code: Convert string to numeric, data step

```
data perm.titanic;
  set perm.titanic;
  age_c = input(age, ?? 8.);
run;

proc means
    n nmiss mean std min max
    data=perm.titanic;
  var age_c;
  title1 "Descriptive statistics for age";
run;
```

SAS output: Convert string to numeric, data step



14:56 Monday, July 12, 2021 4

Descriptive statistics for age

The MEANS Procedure

Analysis Variable : age_c					
N	N Miss	Mean	Std Dev	Minimum	Maximum
756	557	30.3979894	14.2590487	0.1700000	71.0000000

SAS Output

Break #2

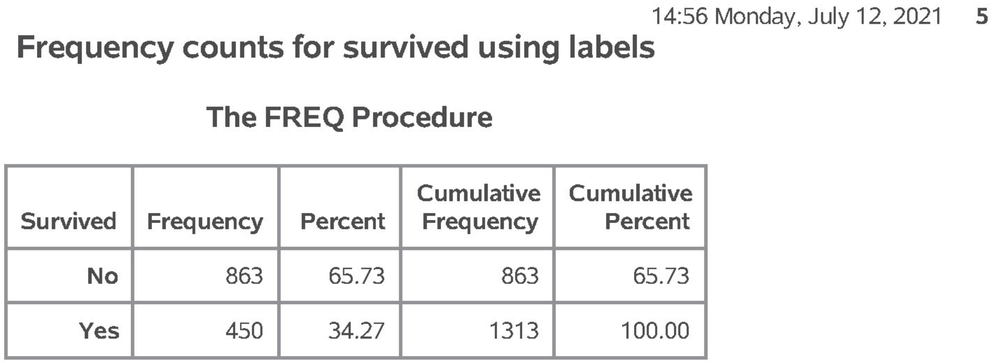
- What you have learned
  - Convert string to numeric
- What's coming next
  - Labels for number codes

SAS code: Using proc format to code categorical data

```
proc format;
  value f_survived
    0 = "No"
    1 = "Yes";
run;

proc freq
    data=perm.titanic;
  tables Survived;
  format Survived f_survived.;
  title1 "Frequency counts for survived using
labels";
run;
```

SAS output: Using proc format to code categorical data



14:56 Monday, July 12, 2021 5

Frequency counts for survived using labels

The FREQ Procedure

Survived	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	863	65.73	863	65.73
Yes	450	34.27	1313	100.00

SAS Output

Break #3

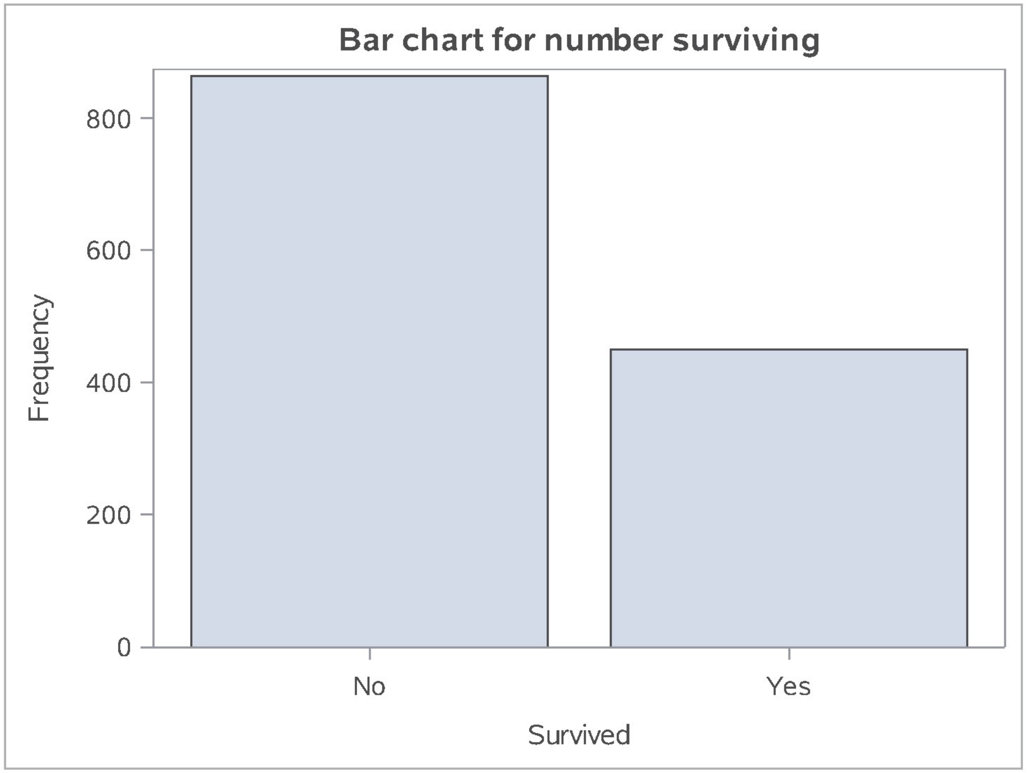
- – What you have learned
  - Labels for number codes
- – What's coming next
  - Drawing bar charts

SAS code: Bar charts, proc sgplot (1/3)

```
proc sgplot
    data=perm.titanic;
  vbar Survived;
  format Survived f_survived.;
  title1 "Bar chart for number surviving";
run;
```

SAS output: Bar charts, proc sgplot (1/3)

14:56 Monday, July 12, 2021 **6**



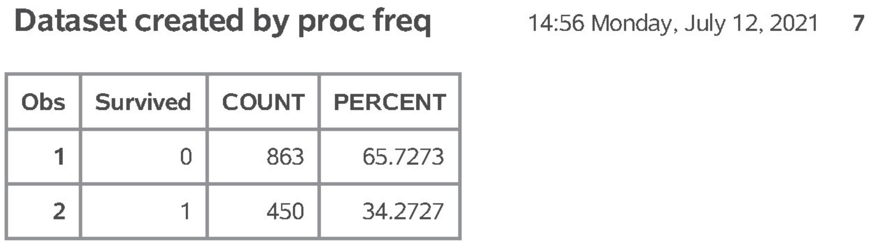
SAS Output

SAS code: Bar charts, proc sgplot (2/3)

```
proc freq
    noprint
    data=perm.titanic;
  tables Survived / out=pct_survived;
run;

proc print
    data=pct_survived;
  title1 "Dataset created by proc freq";
run;
```

SAS output: Bar charts, proc sgplot (2/3)



Dataset created by proc freq 14:56 Monday, July 12, 2021 **7**

Obs	Survived	COUNT	PERCENT
1	0	863	65.7273
2	1	450	34.2727

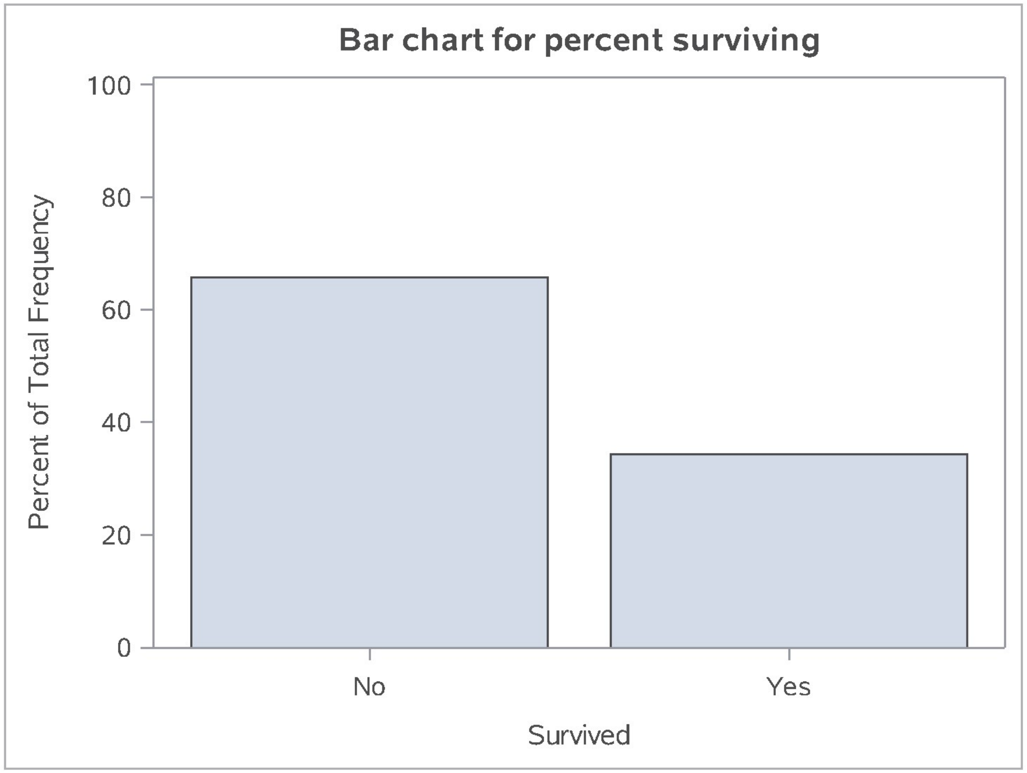
SAS Output

SAS code: Bar charts, proc sgplot (3/3)

```
proc sgplot
    data=pct_survived;
  vbar Survived / response=Percent;
  yaxis max=100;
  format Survived f_survived.;
  title1 "Bar chart for percent surviving";
run;
```

SAS output: Bar charts, proc sgplot (3/3)

14:56 Monday, July 12, 2021 8



SAS Output

Break #4

– What you have learned
  - Drawing bar charts
– What’s coming next
  - Converting continuous to categorical

SAS code: Converting continuous to categorical (1/5)

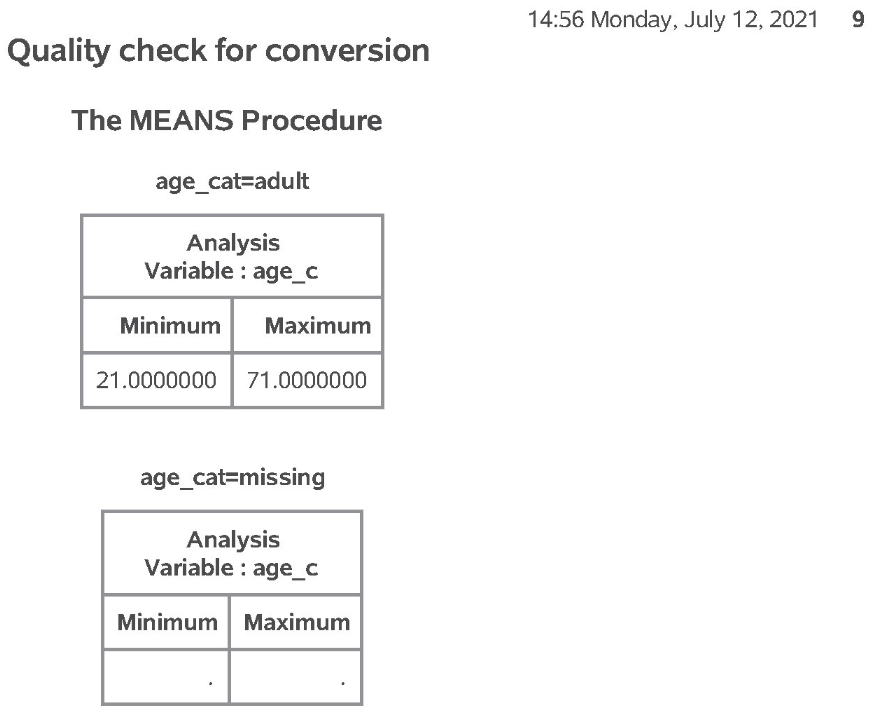
```
data age_categories;
  set perm.titanic;
  if age_c = .
    then age_cat = "missing ";
  else if age_c < 6
    then age_cat = "toddler ";
  else if age_c < 13
    then age_cat = "pre-teen";
  else if age_c < 21
    then age_cat = "teenager";
  else age_cat   = "adult   ";
run;
```

SAS code: Converting continuous to categorical (2/5)

```
proc sort
    data=age_categories;
  by age_cat;
run;

proc means
    min max
    data=age_categories;
  by age_cat;
  var age_c;
  title1 "Quality check for conversion";
run;
```

SAS output: Converting continuous to categorical (2/5)



14:56 Monday, July 12, 2021 **9**

Quality check for conversion

The MEANS Procedure

age_cat=adult

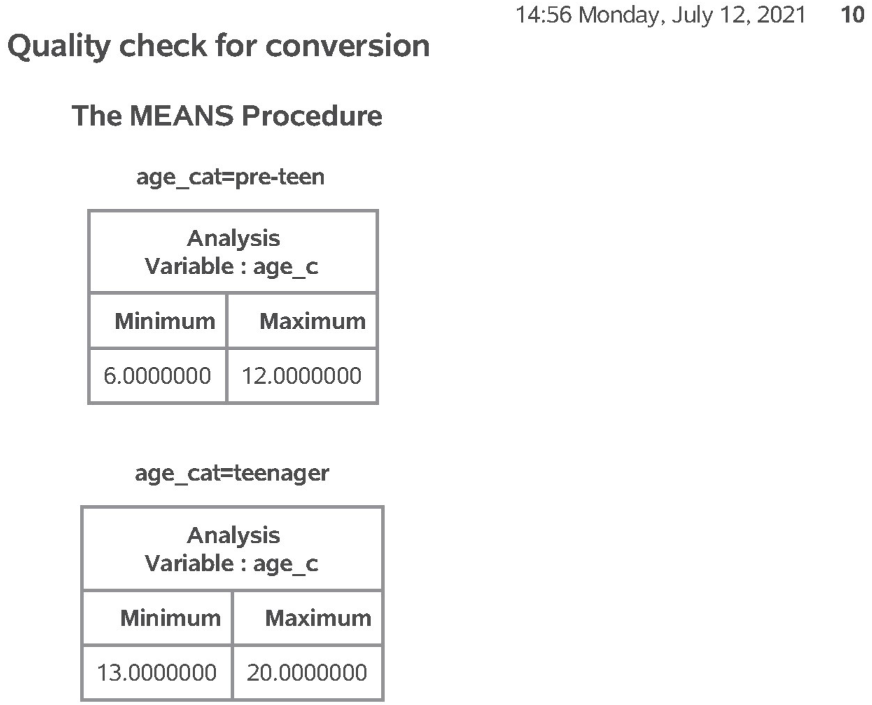
Analysis Variable : age_c	
Minimum	**Maximum**
21.0000000	71.0000000

age_cat=missing

Analysis Variable : age_c	
Minimum	**Maximum**
.	.

SAS Output

SAS output: Converting continuous to categorical (2/5)



14:56 Monday, July 12, 2021 **10**

Quality check for conversion

The MEANS Procedure

age_cat=pre-teen

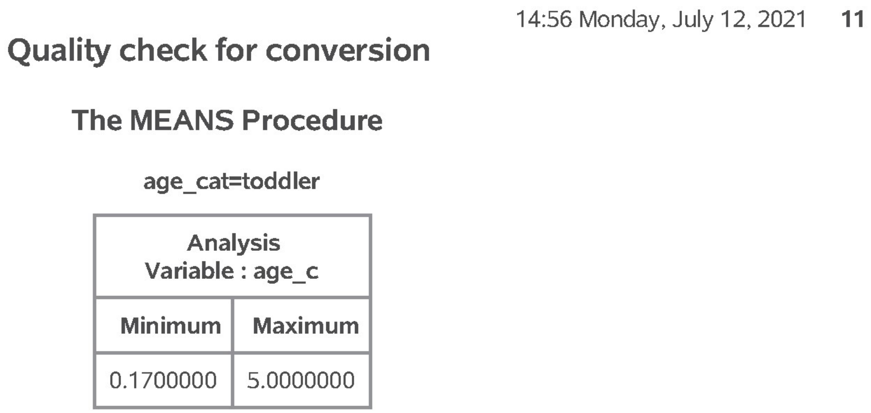
Analysis Variable : age_c	
Minimum	**Maximum**
6.0000000	12.0000000

age_cat=teenager

Analysis Variable : age_c	
Minimum	**Maximum**
13.0000000	20.0000000

SAS Output

SAS output: Converting continuous to categorical (2/5)



14:56 Monday, July 12, 2021 **11**

Quality check for conversion

The MEANS Procedure

age_cat=toddler

Analysis Variable : age_c	
Minimum	**Maximum**
0.1700000	5.0000000

SAS Output

SAS code: Converting continuous to categorical (3/5)

```
data age_codes;
  set perm.titanic;
  if age_c = .
    then age_cat = 9;
  else if age_c < 6
    then age_cat = 1;
  else if age_c < 13
    then age_cat = 2;
  else if age_c < 21
    then age_cat = 3;
  else age_cat = 4;
run;
```

SAS code: Converting continuous to categorical (4/5)

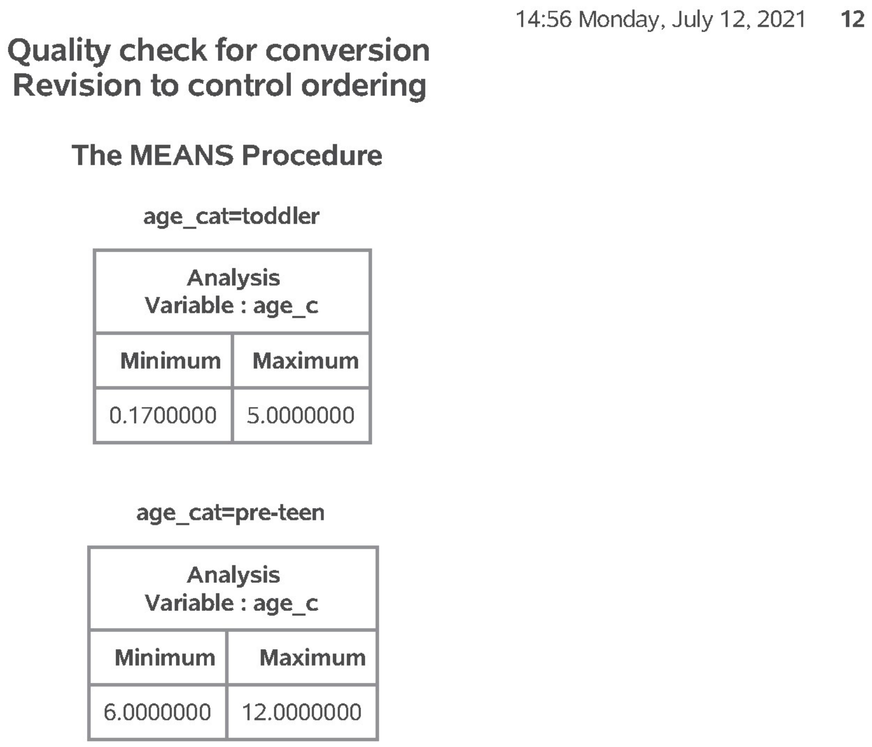
```
proc format;
  value f_age
    1 = "toddler"
    2 = "pre-teen"
    3 = "teenager"
    4 = "adult"
    9 = "unknown";
run;
```

SAS code: Converting continuous to categorical (5/5)

```
proc sort
    data=age_codes;
  by age_cat;
run;

proc means
    min max
    data=age_codes;
  by age_cat;
  var age_c;
  format age_cat f_age.;
  title1 "Quality check for conversion";
  title2 "Revision to control ordering";
run;
```

SAS output: Converting continuous to categorical (5/5)



14:56 Monday, July 12, 2021 **12**

Quality check for conversion
Revision to control ordering

The MEANS Procedure

age_cat=toddler

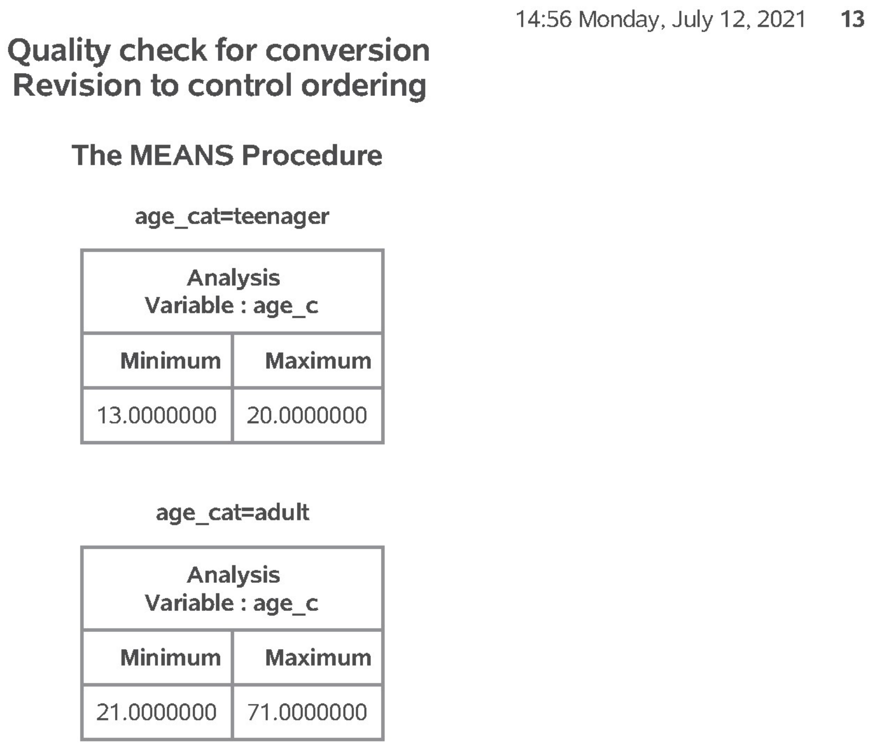
Analysis Variable : age_c	
Minimum	**Maximum**
0.1700000	5.0000000

age_cat=pre-teen

Analysis Variable : age_c	
Minimum	**Maximum**
6.0000000	12.0000000

SAS Output

SAS output: Converting continuous to categorical (5/5)



14:56 Monday, July 12, 2021 **13**

Quality check for conversion
Revision to control ordering

The MEANS Procedure

age_cat=teenager

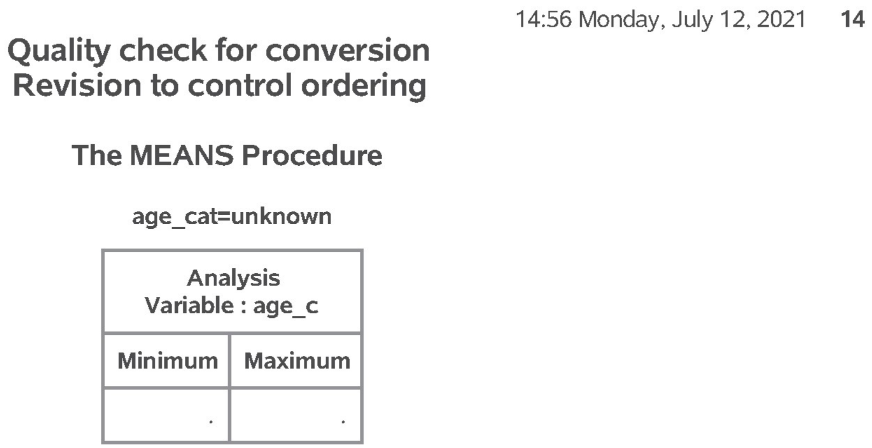
Analysis Variable : age_c	
Minimum	**Maximum**
13.0000000	20.0000000

age_cat=adult

Analysis Variable : age_c	
Minimum	**Maximum**
21.0000000	71.0000000

SAS Output

SAS output: Converting continuous to categorical (5/5)



14:56 Monday, July 12, 2021 **14**

Quality check for conversion
Revision to control ordering

The MEANS Procedure

age_cat=unknown

Analysis Variable : age_c	
Minimum	**Maximum**
.	.

SAS Output

Break #5

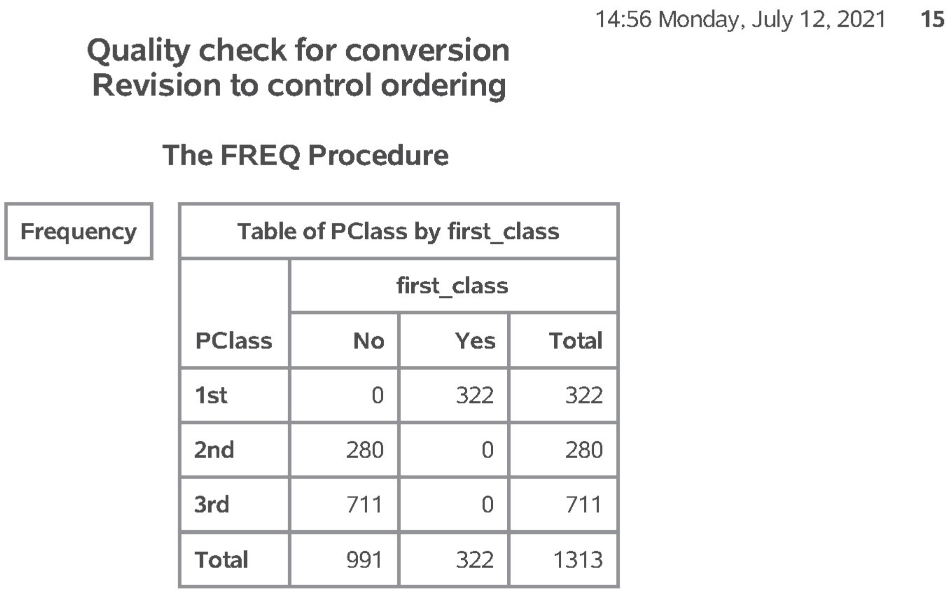
- What you have learned
  - Converting continuous to categorical
- What's coming next
  - Modifying categorical variables

SAS code: Modifying a categorical variable

```
data first_class;
  set perm.titanic;
  if PClass = "1st"
    then first_class = "Yes";
    else first_class = "No";
run;

proc freq
    data=first_class;
  table PClass*first_class /
    norow nocol nopercent;
run;
```

SAS output: Modifying a categorical variable



14:56 Monday, July 12, 2021 15

Quality check for conversion
Revision to control ordering

The FREQ Procedure

Frequency

Table of PClass by first_class

	first_class		
PClass	No	Yes	Total
1st	0	322	322
2nd	280	0	280
3rd	711	0	711
Total	991	322	1313

SAS Output

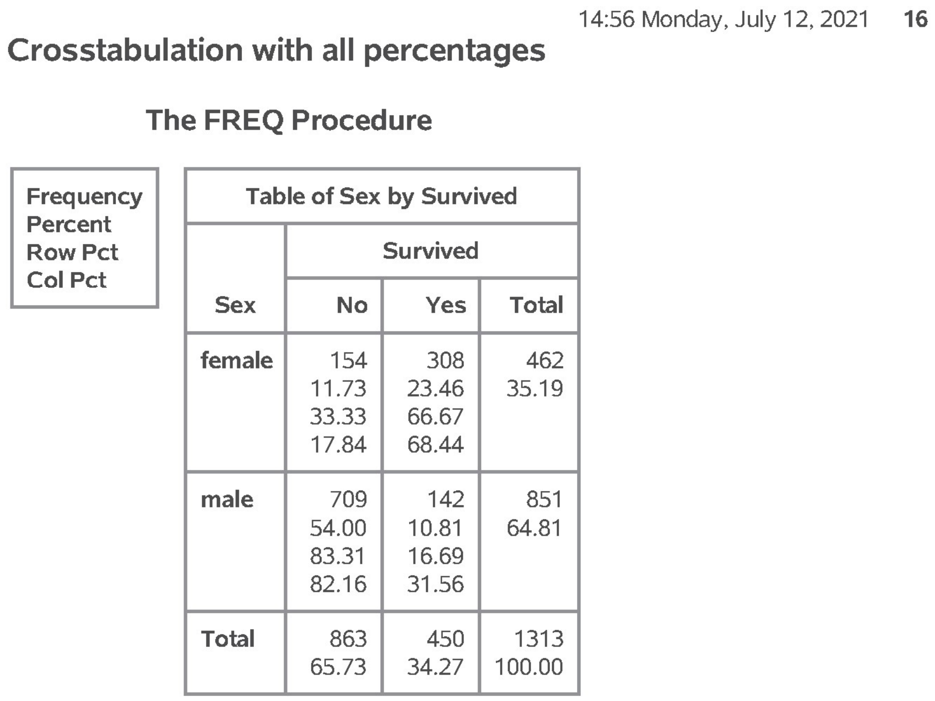
Break #6

- What you have learned
  - Modifying categorical variables
- What's coming next
  - Crosstabulation

SAS code: Crosstabulation (1/4)

```
proc freq
    data=perm.titanic;
  tables Sex*Survived;
  format Survived f_survived.;
  title1 "Crosstabulation with all percentages";
run;
```

SAS output: Crosstabulation (1/4)



14:56 Monday, July 12, 2021 16

Crosstabulation with all percentages

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

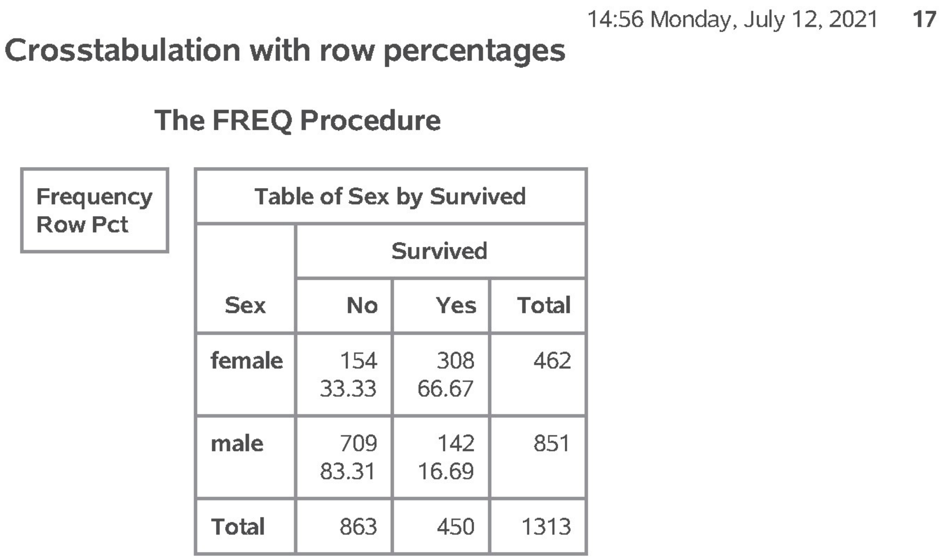
Table of Sex by Survived			
	Survived		
Sex	No	Yes	Total
female	154 11.73 33.33 17.84	308 23.46 66.67 68.44	462 35.19
male	709 54.00 83.31 82.16	142 10.81 16.69 31.56	851 64.81
Total	863 65.73	450 34.27	1313 100.00

SAS Output

SAS code: Crosstabulation (2/4)

```
proc freq
    data=perm.titanic;
  tables Sex*Survived / nocol nopercent;
  format Survived f_survived.;
  title1 "Crosstabulation with row percentages";
run;
```

SAS output: Crosstabulation (2/4)



14:56 Monday, July 12, 2021 17

Crosstabulation with row percentages

The FREQ Procedure

Frequency Row Pct

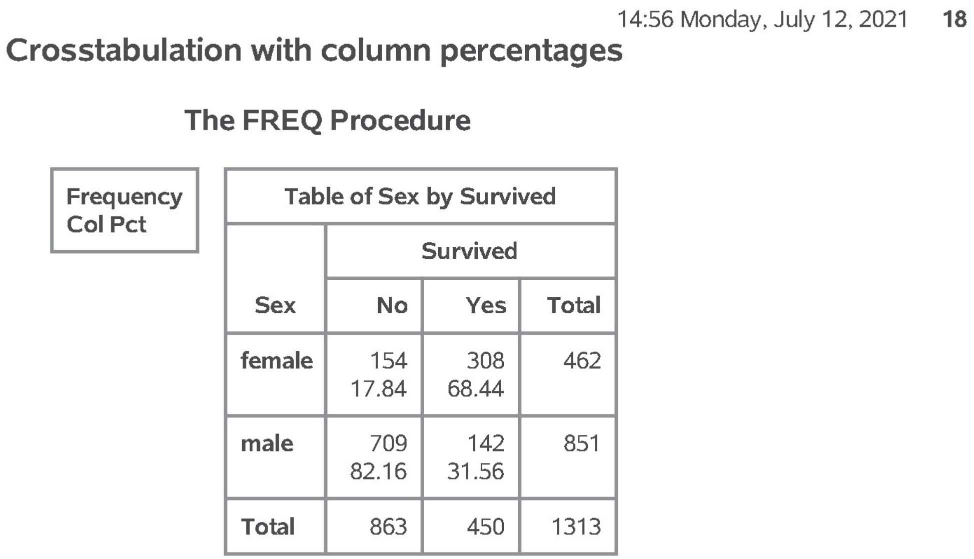
Table of Sex by Survived			
	Survived		
Sex	No	Yes	Total
female	154 33.33	308 66.67	462
male	709 83.31	142 16.69	851
Total	863	450	1313

SAS Output

SAS code: Crosstabulation (3/4)

```
proc freq
    data=perm.titanic;
  tables Sex*Survived / norow nopercent;
  format Survived f_survived.;
  title1 "Crosstabulation with column
percentages";
run;
```

SAS output: Crosstabulation (3/4)



14:56 Monday, July 12, 2021 **18**

Crosstabulation with column percentages

The FREQ Procedure

Frequency
Col Pct

Table of Sex by Survived

	Survived		
Sex	**No**	**Yes**	**Total**
female	154 17.84	308 68.44	462
male	709 82.16	142 31.56	851
Total	863	450	1313

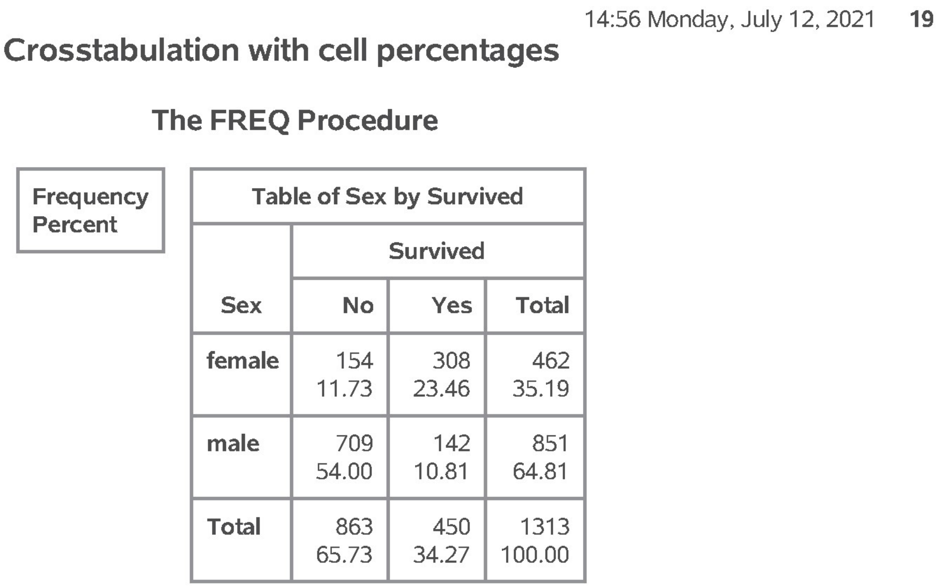
SAS Output

SAS code: Crosstabulation (4/4)

```
proc freq
    data=perm.titanic;
  tables Sex*Survived / norow nocol;
  format Survived f_survived.;
  title1 "Crosstabulation with cell percentages";
run;

ods pdf close;
```

SAS output: Crosstabulation (4/4)



14:56 Monday, July 12, 2021 **19**

Crosstabulation with cell percentages

The FREQ Procedure

Frequency
Percent

Table of Sex by Survived

	Survived		
Sex	**No**	**Yes**	**Total**
female	154 11.73	308 23.46	462 35.19
male	709 54.00	142 10.81	851 64.81
Total	863 65.73	450 34.27	1313 100.00

SAS Output

Review

- – Frequency counts
- – Convert string to numeric
- – Labels for number codes
- – Drawing bar charts
- – Converting continuous to categorical
- – Modifying categorical variables
- – Crosstabulations