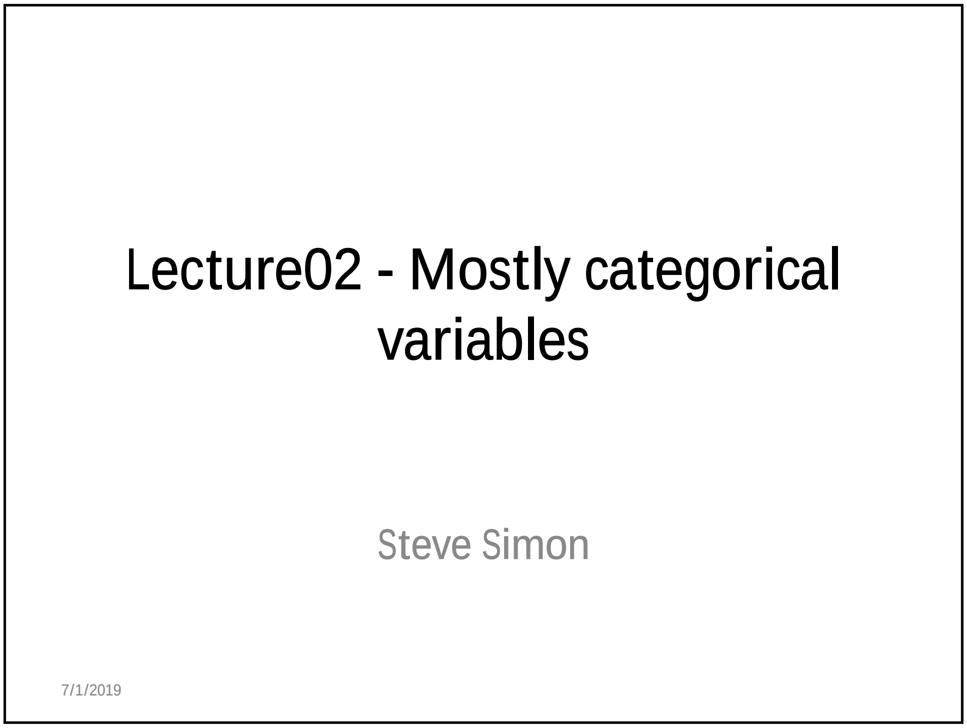
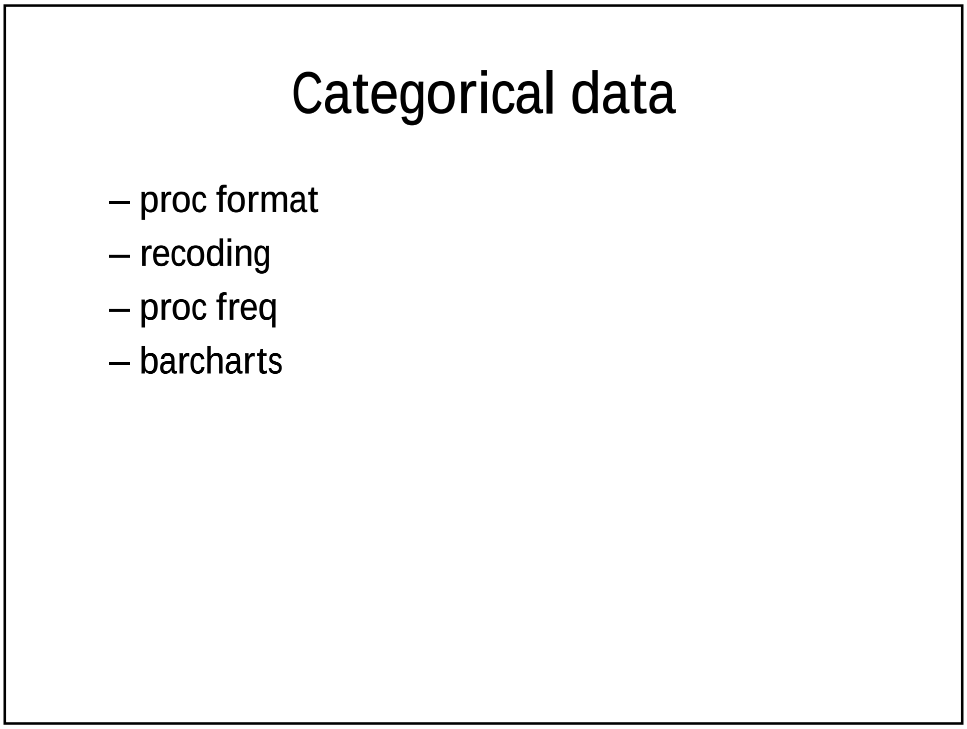


Lecture02 - Mostly categorical variables

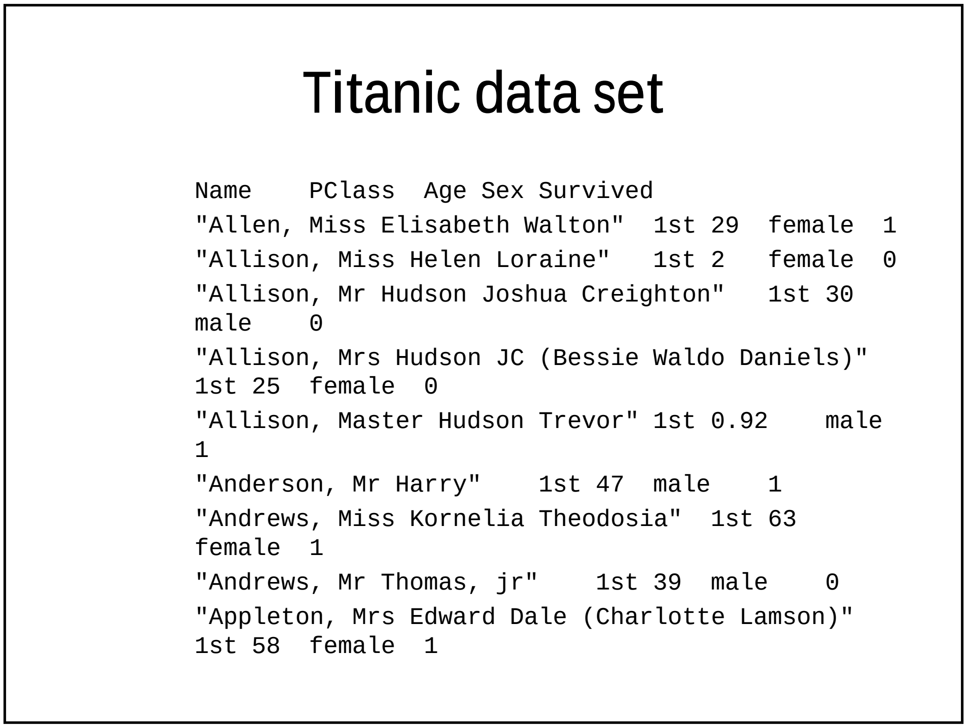
Steve Simon

7/1/2019



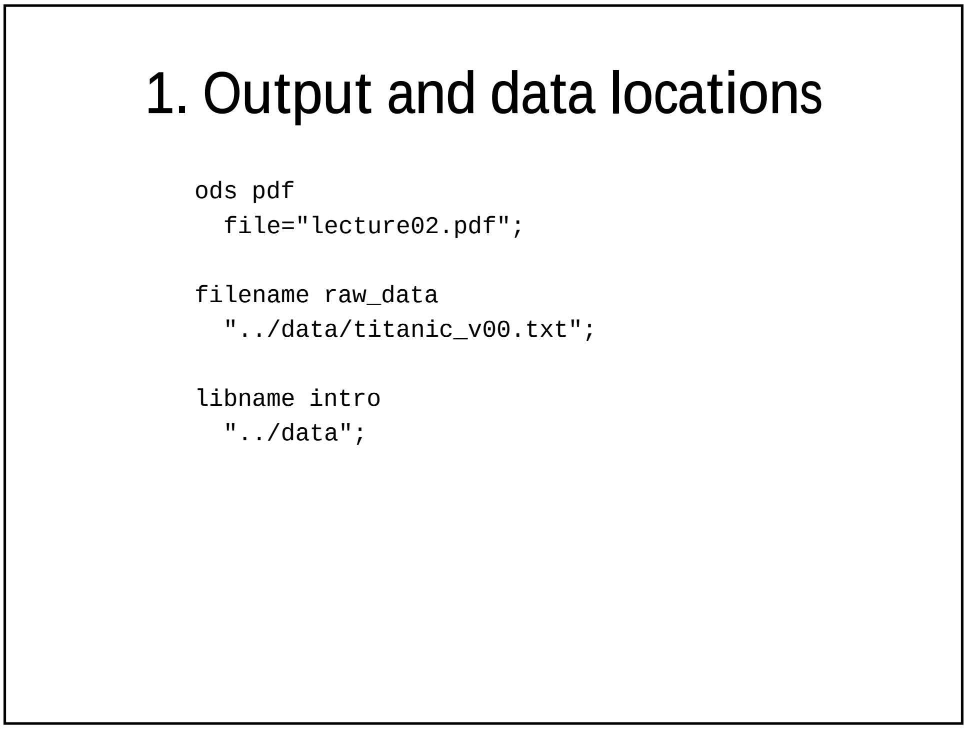
Categorical data

- – proc format
- – recoding
- – proc freq
- – barcharts



Titanic data set

```
Name    PClass  Age Sex Survived
"Allen, Miss Elisabeth Walton"  1st 29  female  1
"Allison, Miss Helen Loraine"   1st 2   female  0
"Allison, Mr Hudson Joshua Creighton"   1st 30
male    0
"Allison, Mrs Hudson JC (Bessie Waldo Daniels)"
1st 25  female  0
"Allison, Master Hudson Trevor" 1st 0.92    male
1
"Anderson, Mr Harry"    1st 47  male    1
"Andrews, Miss Kornelia Theodosia"  1st 63
female  1
"Andrews, Mr Thomas, jr"    1st 39  male    0
"Appleton, Mrs Edward Dale (Charlotte Lamson)"
1st 58  female  1
```



1. Output and data locations

```
ods pdf
  file="lecture02.pdf";

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

libname intro
  "../data";
```

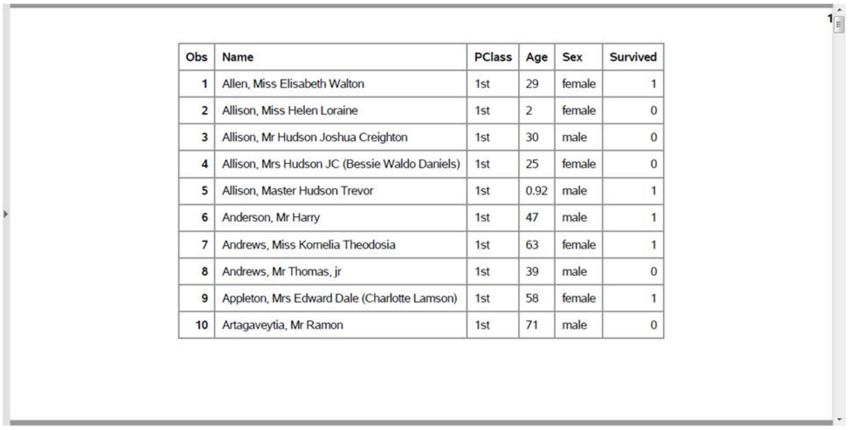
2. Reading, proc import

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

3. First ten lines, proc print

```
proc print
    data=intro.titanic(obs=10);
  title1 " ";
run;
```

First ten rows of the Titanic data set



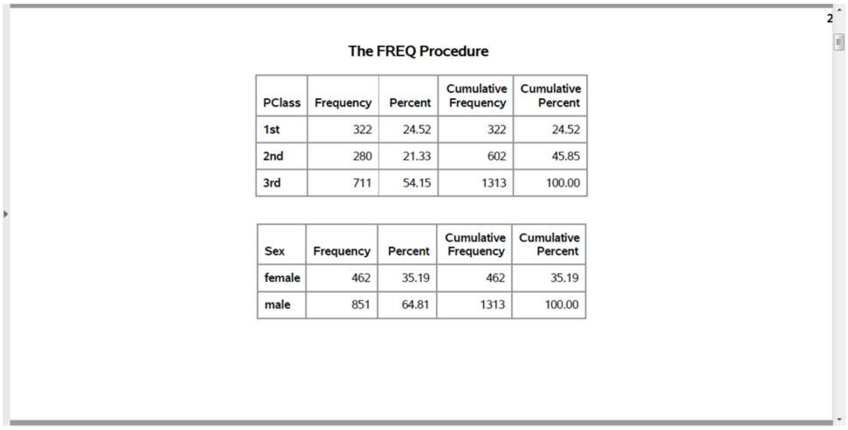
Obs	Name	PClass	Age	Sex	Survived
1	Allen, Miss Elisabeth Walton	1st	29	female	1
2	Allison, Miss Helen Loraine	1st	2	female	0
3	Allison, Mr Hudson Joshua Creighton	1st	30	male	0
4	Allison, Mrs Hudson JC (Bessie Waldo Daniels)	1st	25	female	0
5	Allison, Master Hudson Trevor	1st	0.92	male	1
6	Anderson, 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

Output, proc print

4. Counts, proc freq

```
proc freq
    data=intro.titanic;
  tables PClass Sex Survived;
run;
```

Counts for categorical data (1/2)



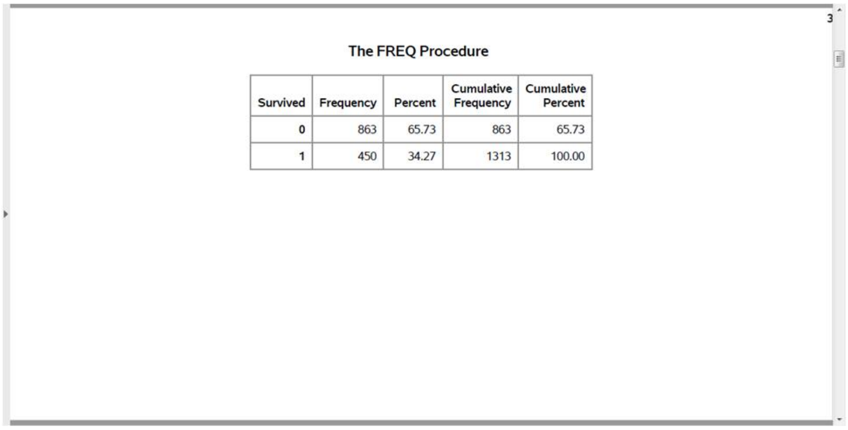
The FREQ Procedure

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

Output, proc freq

Counts for categorical data (2/2)



The FREQ Procedure

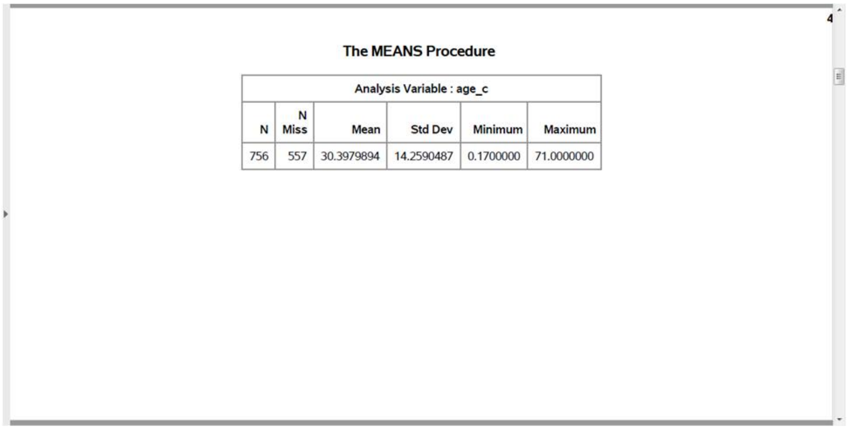
Survived	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	863	65.73	863	65.73
1	450	34.27	1313	100.00

Output, proc freq

5. Convert string to numeric, data step

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

proc means
    n nmiss mean std min max
    data=intro.titanic;
  var age_c;
run;
```

Means and standard deviations for age



The MEANS Procedure

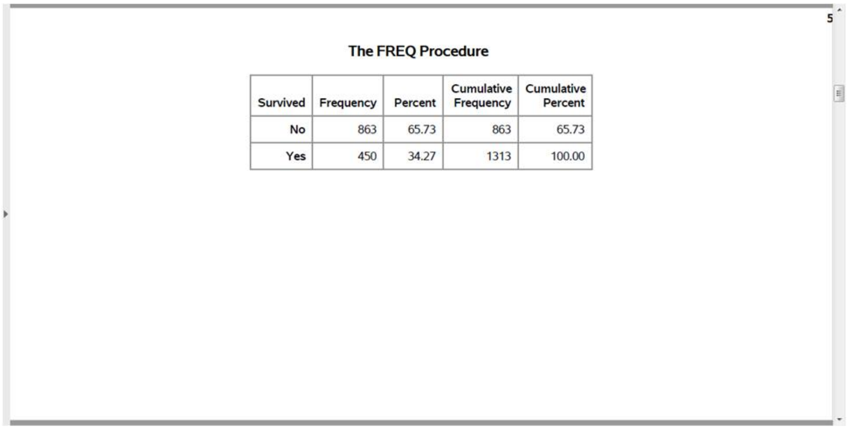
Analysis Variable : age_c					
N	N Miss	Mean	Std Dev	Minimum	Maximum
756	557	30.3979894	14.2590487	0.1700000	71.0000000

Output, proc freq

6. Using proc format to code categorical data

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

proc freq
    data=intro.titanic;
  tables Survived;
  format Survived f_survived.;
run;
```

Nicely formatted counts for survival



The FREQ Procedure

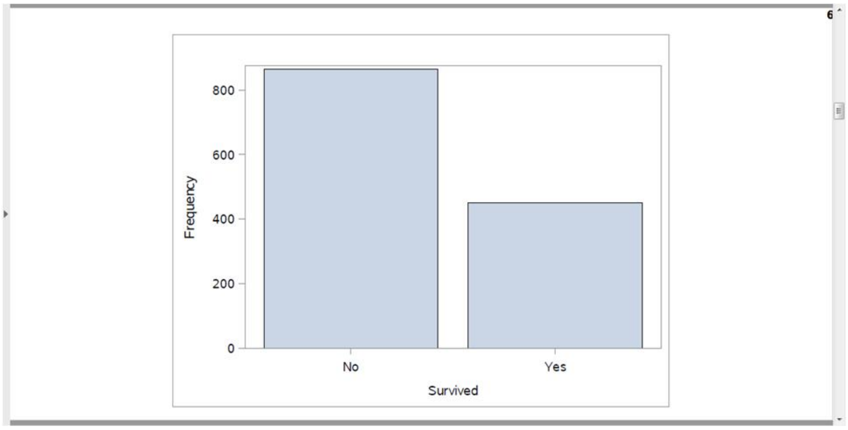
Survived	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	863	65.73	863	65.73
Yes	450	34.27	1313	100.00

Output, proc freq

7. Bar charts, proc sgplot

```
proc sgplot
    data=intro.titanic;
  vbar Survived;
  format Survived f_survived.;
run;
```

Bar chart

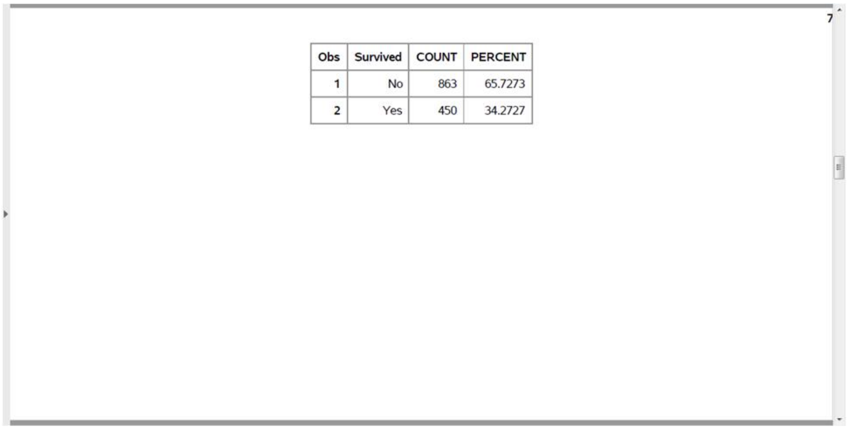


Output, proc sgplot

8. Percentages for bar chart

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

proc print
    data=pct_survived;
  format Survived f_survived.;
run;

proc sgplot
    data=pct_survived;
  vbar Survived / response=Percent;
  yaxis max=100;
  format Survived f_survived.;
run;
```

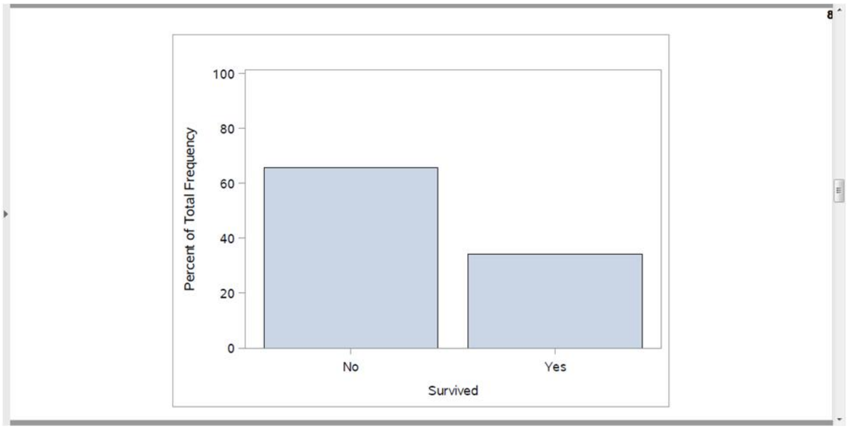
Percentages, proc freq



Obs	Survived	COUNT	PERCENT
1	No	863	65.7273
2	Yes	450	34.2727

Output, proc freq

Percentages in a bar chart

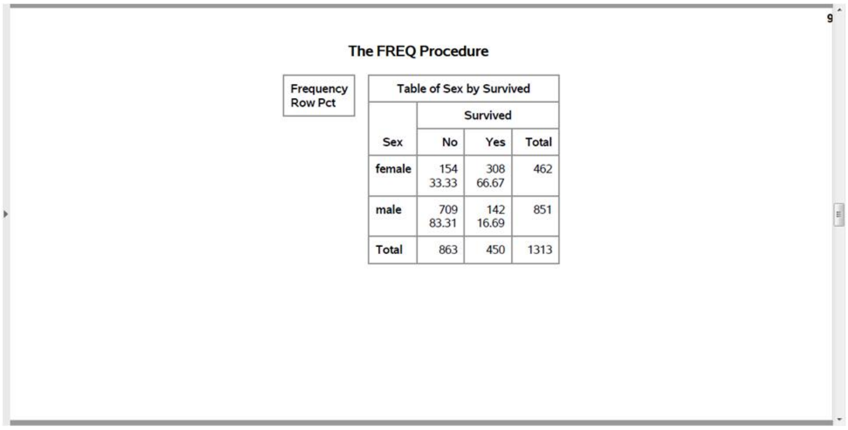


Output, proc sgplot

9. Crosstabulation

```
proc freq
    data=intro.titanic;
  tables Sex*Survived / nocol nopercent;
  format Survived f_survived.;
run;
```

Percentages, proc freq



The FREQ Procedure

Frequency
Row Pct

Table of Sex by Survived

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

Output, proc freq

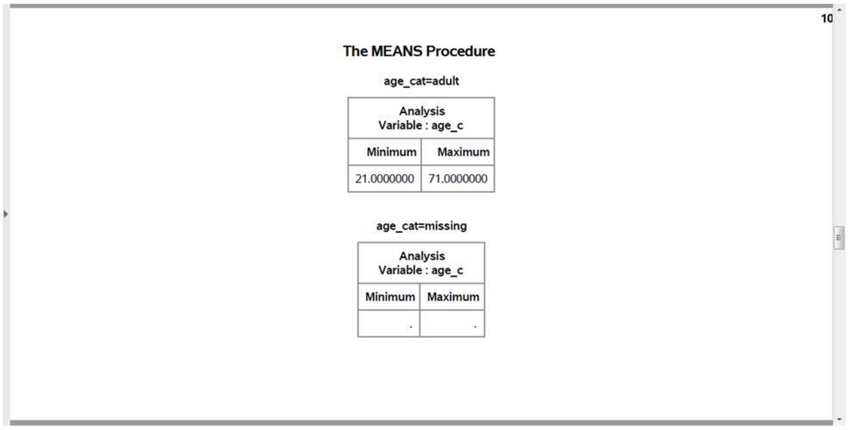
10. Converting a continuous variable to categorical

```
data age_categories;
  set intro.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;
```

11. Quality check

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

proc means
    min max
    data=age_categories;
  by age_cat;
  var age_c;
run;
```

Recoding age (1 / 3)



10

The MEANS Procedure

age_cat=adult

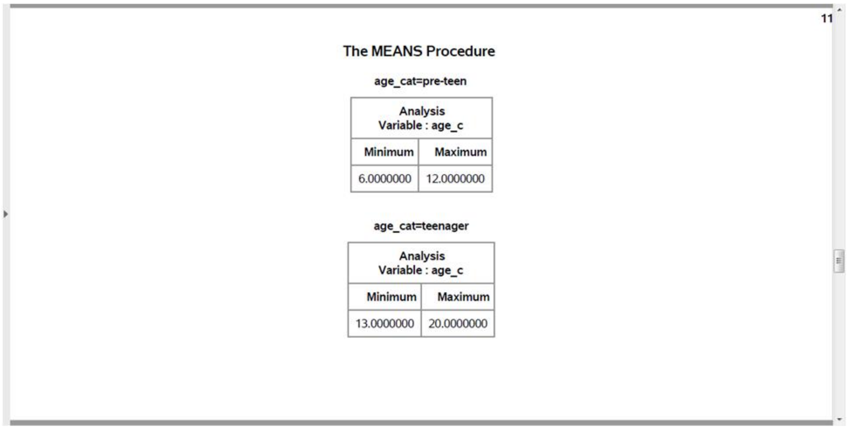
Analysis Variable : age_c	
Minimum	Maximum
21.0000000	71.0000000

age_cat=missing

Analysis Variable : age_c	
Minimum	Maximum
.	.

Output, proc means

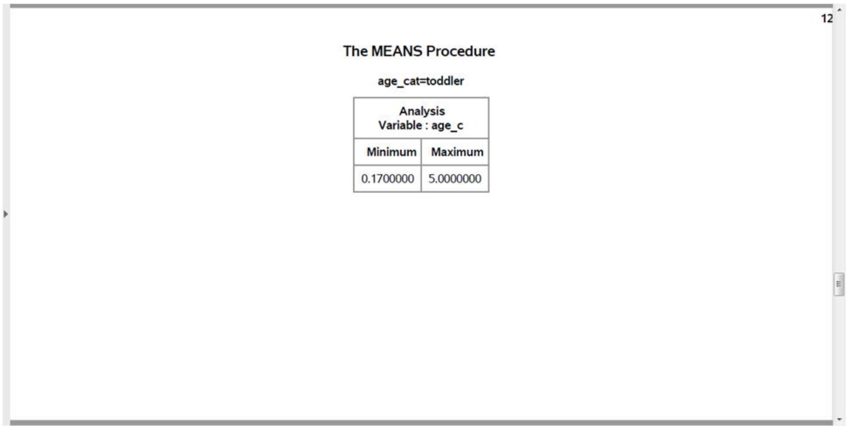
Recoding age (2 / 3)



```
                                                        11
              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
```

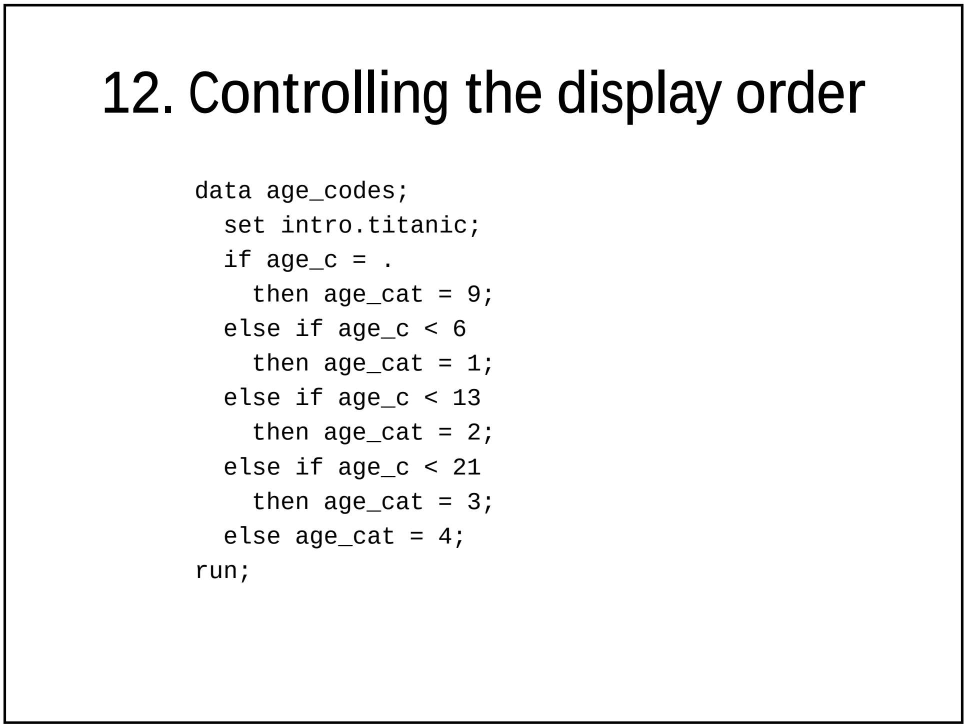
Output, proc means

Recoding age (3 / 3)



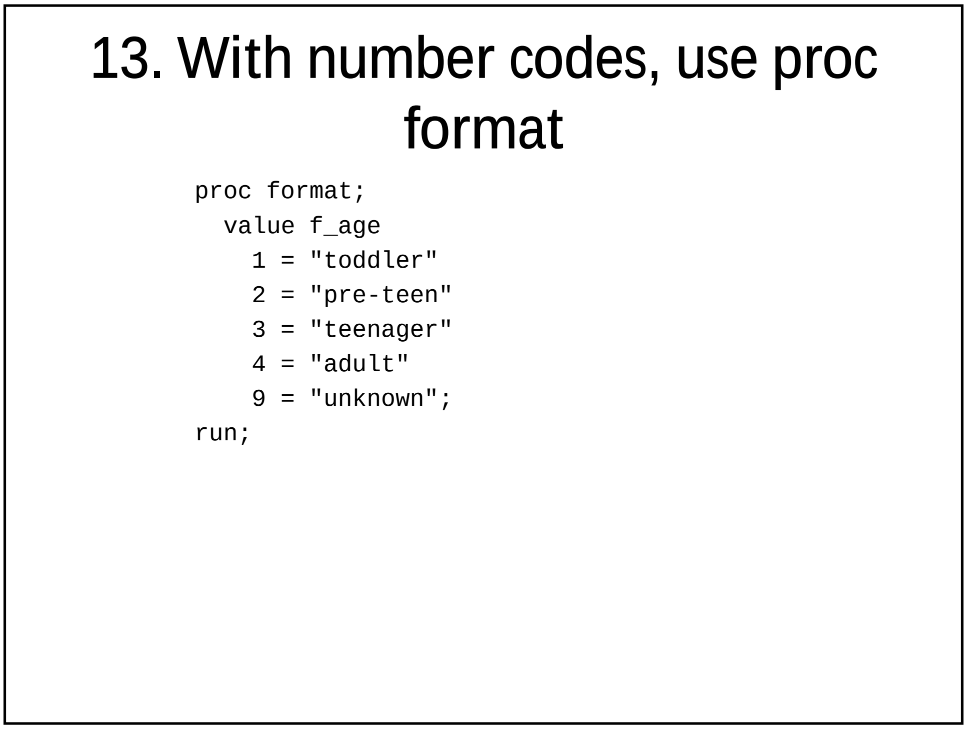
```
                                                        12
              The MEANS Procedure

               age_cat=toddler

                  Analysis
              Variable : age_c
          Minimum       Maximum
        0.1700000     5.0000000
```

Output, proc means



12. Controlling the display order

```
data age_codes;
  set intro.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;
```



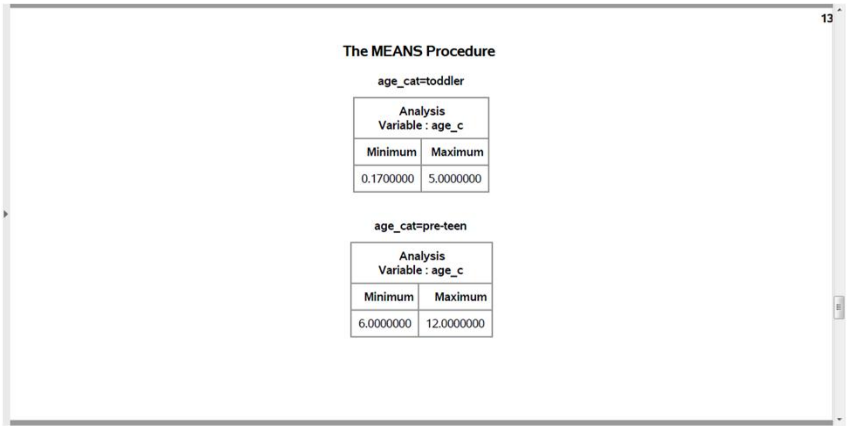
13. With number codes, use proc format

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

14. Quality check

```
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.;
run;
```

Better age recode (1 /3)



13

The MEANS Procedure

age_cat=toddler

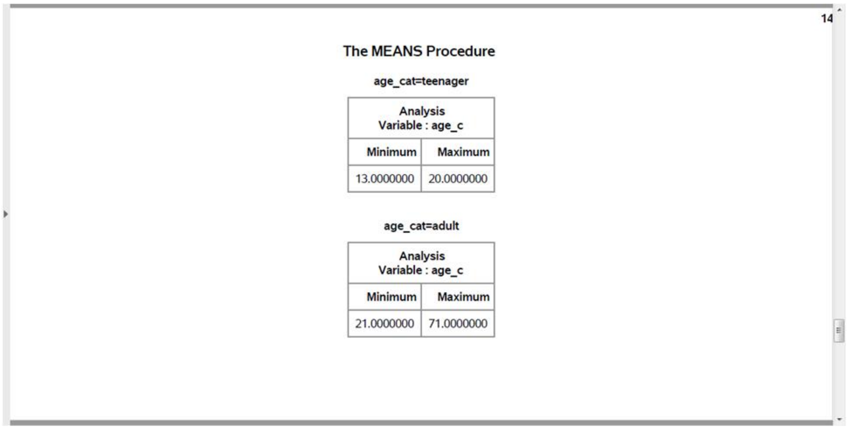
Analysis Variable : age_c	
Minimum	Maximum
0.1700000	5.0000000

age_cat=pre-teen

Analysis Variable : age_c	
Minimum	Maximum
6.0000000	12.0000000

Output, proc means

Better age recode (2 /3)



14

The MEANS Procedure

age_cat=teenager

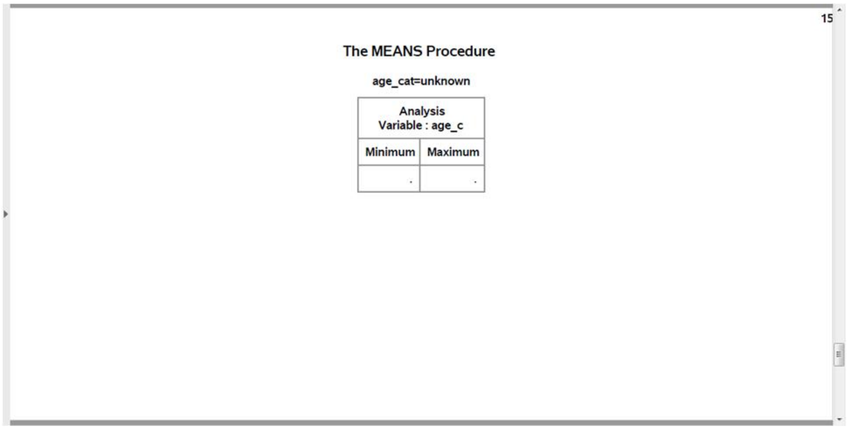
Analysis Variable : age_c	
Minimum	Maximum
13.0000000	20.0000000

age_cat=adult

Analysis Variable : age_c	
Minimum	Maximum
21.0000000	71.0000000

Output, proc means

Better age recode (3 /3)



15

The MEANS Procedure

age_cat=unknown

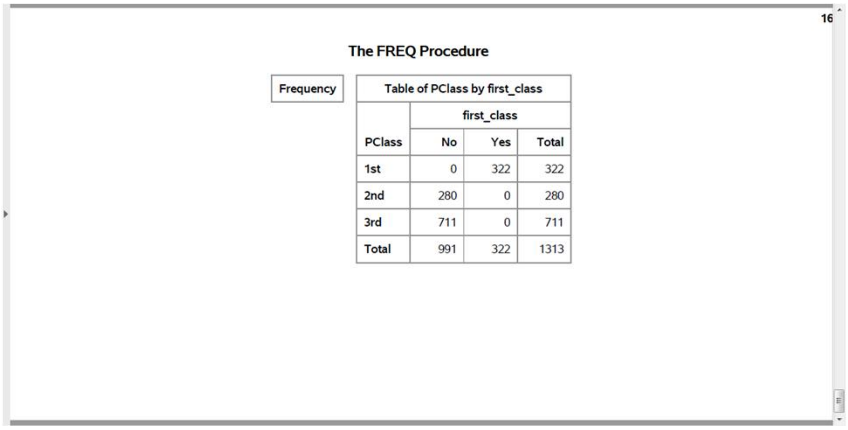
Analysis Variable : age_c	
Minimum	Maximum
.	.

Output, proc means

15. Modifying a categorical variable

```
data first_class;
  set intro.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;
```

Quality check



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

Output, proc means