PROC TABULATE

Shannon Pileggi

STAT 330

OUTLINE

Overview

Structure

Statistics

Display

Overview

PROC	Detail	Summary	Control	N	sum	mean	std	%
PRINT	✓	✗	✓	✓	✓	✗	✗	✗
MEANS	✗	✓	✗	✓	✓	✓	✓	✗
FREQ	✗	✓	✗	✓	✗	✗	✗	✓
REPORT	✓	✓	✓	✓	✓	✓	✓	✓
TABULATE	✗	✓	✓	✓	✓	✓	✓	✓
SQL	✓	✓	✗	✓	✓	✓	✓	✓

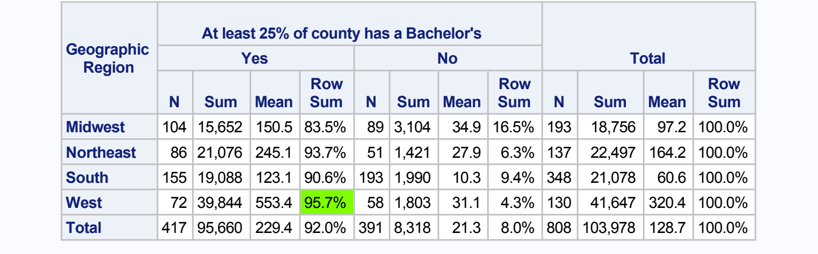
- Detail: display a row for each observation
- Summary: display a row for a group of observations
- Control: many layout/format/display options in output
- SQL: can additionally combine and sort data

Patents data

- number of utility patent ("patents for inventions") grants from 2011, by county
- demographic variables from the American Community Survey
  - some variables may be missing for smaller counties
- San Jose, CA (Santa Clara County)
  - 3^{rd} largest city in CA, 10^{th} largest city in US
  - leads all US cities in generating patents

On your own: Explore the `patents` data in SAS.

Goal



Geographic Region	At least 25% of county has a Bachelor's								Total			
	Yes				No							
	N	Sum	Mean	Row Sum	N	Sum	Mean	Row Sum	N	Sum	Mean	Row Sum
Midwest	104	15,652	150.5	83.5%	89	3,104	34.9	16.5%	193	18,756	97.2	100.0%
Northeast	86	21,076	245.1	93.7%	51	1,421	27.9	6.3%	137	22,497	164.2	100.0%
South	155	19,088	123.1	90.6%	193	1,990	10.3	9.4%	348	21,078	60.6	100.0%
West	72	39,844	553.4	95.7%	58	1,803	31.1	4.3%	130	41,647	320.4	100.0%
Total	417	95,660	229.4	92.0%	391	8,318	21.3	8.0%	808	103,978	128.7	100.0%

- Region along rows, education status along columns
- Row and column totals
- Various statistics reported, formatted values in cells
- Highlighted cell: in the west region, 95.7% of all patents come from counties with higher education levels
- Style modified and exported to a pdf

Overview

Structure

Statistics

Display

Syntax

SAS Code

```
PROC TABULATE DATA = dataset ;
   CLASS catvar1 catvar2...  ;
   VAR   quantvar1 quantvar2... ;
   TABLE page-var, row-var, col-var;
RUN;
```

SAS Code

Each variable listed in TABLE statement **must** also be listed in either CLASS or VAR.

`TABLE` *var1*; one-dimensional table with *var1* on columns

`TABLE` *var2*, *var1*; two-dimensional table with *var2* on rows, *var1* on columns

`TABLE` *var3*, *var2*, *var1*; three-dimensional table with page by *var3*, *var2* on rows, *var1* on columns

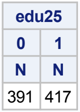
One-dimensional table

- columns = edu25

SAS Code

```
PROC TABULATE DATA = patents ;
   CLASS edu25 ;
   TABLE edu25 ;
RUN ;
```

SAS Code



edu25	
0	1
N	N
391	417

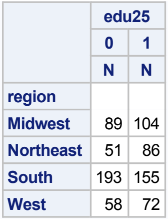
Two-dimensional table

- columns = edu25
- rows = region

SAS Code

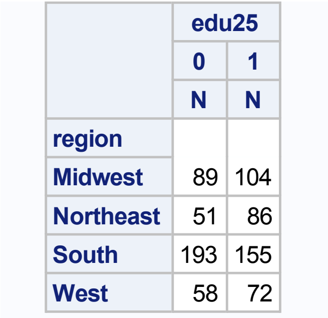
```
PROC TABULATE DATA = patents ;
   CLASS edu25 region ;
   TABLE region, edu25 ;
RUN ;
```

SAS Code



	edu25	
	0	1
	N	N
region		
Midwest	89	104
Northeast	51	86
South	193	155
West	58	72

Discussion



	edu25	
	0	1
	N	N
region		
Midwest	89	104
Northeast	51	86
South	193	155
West	58	72

Which of the following is a correct interpretation?

1. 72 people in the Western region with higher education levels received patents
2. 72 counties in the Western region with higher education levels received patents
3. 72 patents come from the Western region with higher education levels
4. 72% of patents come the Western region with higher education levels

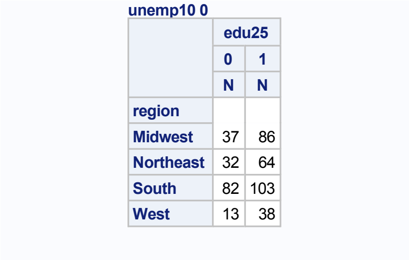
Three-dimensional table

SAS Code

```
PROC TABULATE DATA = patents ;
   CLASS unemp10 edu25 region ;
   TABLE unemp10, region, edu25 ;
RUN ;
```

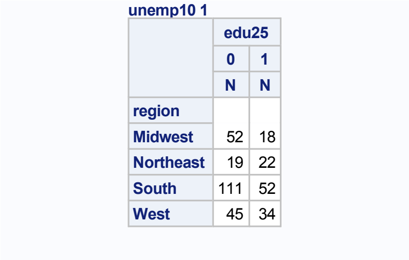
SAS Code

- columns = edu25
- rows = region
- page = unemp10



unemp10 0

	edu25	
	0	1
	N	N
region		
Midwest	37	86
Northeast	32	64
South	82	103
West	13	38



unemp10 1

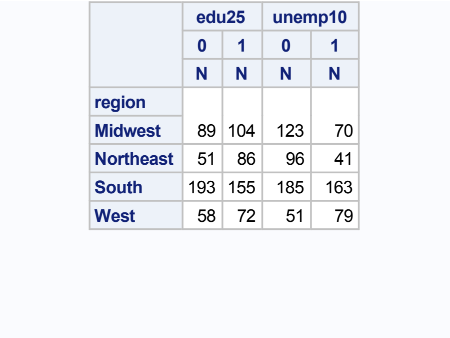
	edu25	
	0	1
	N	N
region		
Midwest	52	18
Northeast	19	22
South	111	52
West	45	34

Concatenate

SAS Code

```
PROC TABULATE DATA = patents ;
   CLASS edu25 region unemp10 ;
   TABLE region, edu25 unemp10 ;
RUN ;
```

SAS Code



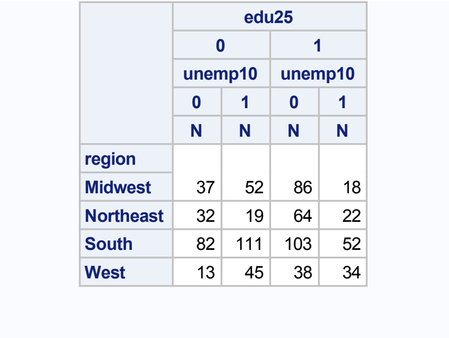
	edu25		unemp10	
	0	1	0	1
	N	N	N	N
region				
Midwest	89	104	123	70
Northeast	51	86	96	41
South	193	155	185	163
West	58	72	51	79

Cross

SAS Code

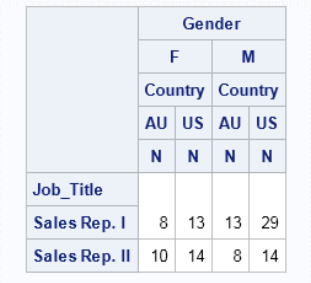
```
PROC TABULATE DATA = patents ;
   CLASS edu25 region unemp10 ;
   TABLE region, edu25*unemp10 ;
RUN ;
```

SAS Code



	edu25			
	0		1	
	unemp10		unemp10	
	0	1	0	1
	N	N	N	N
region				
Midwest	37	52	86	18
Northeast	32	19	64	22
South	82	111	103	52
West	13	45	38	34

Discussion



	Gender			
	F		M	
	Country		Country	
	AU	US	AU	US
	N	N	N	N
Job_Title				
Sales Rep. I	8	13	13	29
Sales Rep. II	10	14	8	14

On your own: This is a (one/two/three) dimensional table where the variables gender and country are (crossed/concatenated).

The statement that generated this table is:

1. `TABLE country*gender, job_title ;`
2. `TABLE job_title, gender*country ;`
3. `TABLE gender, country, job_title ;`
4. `TABLE job_title, country gender ;`
5. `TABLE country gender, job_title ;`

Creating totals

The keyword ALL can be used to create *overall* summarizations.

- ALL can be included in any table dimension

```
TABLE region ALL, edu25 ALL;
```

- ALL can be included with concatenated variables

```
TABLE region, edu25 ALL unemp10 ALL;
```

- ALL can be included with crossed variables

```
TABLE region, edu25*unemp10 ALL;
```

- use parentheses to summarize within group(s)

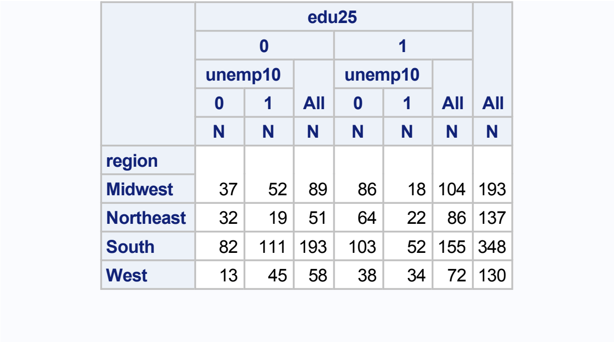
```
TABLE region, edu25*(unemp10 ALL) ALL;
```

Example with ALL

SAS Code

```
PROC TABULATE DATA = patents;
   CLASS edu25 region unemp10;
   TABLE region, edu25*(unemp10 ALL) ALL ;
RUN;
```

SAS Code



	edu25						
	0			1			
	unemp10			unemp10			
	0	1	All	0	1	All	All
	N	N	N	N	N	N	N
region							
Midwest	37	52	89	86	18	104	193
Northeast	32	19	51	64	22	86	137
South	82	111	193	103	52	155	348
West	13	45	58	38	34	72	130

Overview

Structure

Statistics

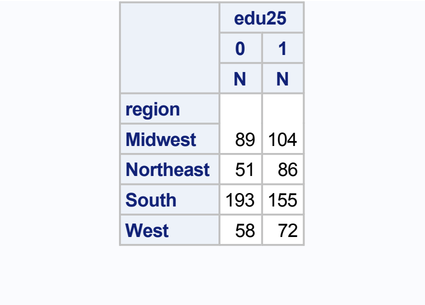
Display

Categorical variables - default statistics

SAS Code

```
PROC TABULATE DATA = patents ;
   CLASS edu25 region ;
   TABLE region, edu25 ;
RUN ;
```

SAS Code



	edu25	
	0	1
	N	N
region		
Midwest	89	104
Northeast	51	86
South	193	155
West	58	72

SAS Code

```
PROC TABULATE DATA = patents ;
   CLASS edu25 region ;
   TABLE region, edu25*N ;
RUN ;
```

SAS Code

- categorical variables go in CLASS
- default statistic is N
- N can be explicitly specified with $*$

Quantitative variables - default statistics

SAS Code

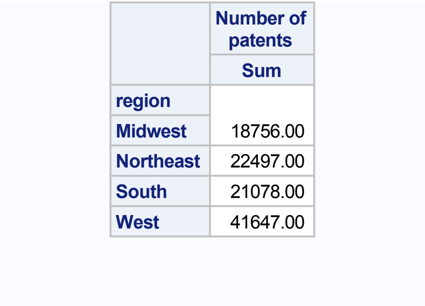
```
PROC TABULATE DATA = patents ;
   CLASS region ;
   VAR patents ;
   TABLE region, patents ;
RUN;
```

SAS Code

SAS Code

```
PROC TABULATE DATA = patents ;
   CLASS region ;
   VAR patents ;
   TABLE region, patents*SUM ;
RUN ;
```

SAS Code



	Number of patents
	Sum
region	
Midwest	18756.00
Northeast	22497.00
South	21078.00
West	41647.00

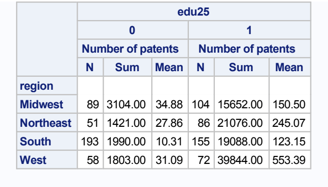
- quantitative variables go in VAR
- default statistic is *SUM*
- *SUM* can be explicitly specified with *

Specifying Statistics

SAS Code

```
PROC TABULATE data = patents ;
   CLASS edu25 region ;
   VAR patents ;
   TABLE region,
         edu25*patents*(N SUM MEAN) ;
RUN ;
```

SAS Code



	edu25					
	0			1		
	Number of patents			Number of patents		
	N	Sum	Mean	N	Sum	Mean
region						
Midwest	89	3104.00	34.88	104	15652.00	150.50
Northeast	51	1421.00	27.86	86	21076.00	245.07
South	193	1990.00	10.31	155	19088.00	123.15
West	58	1803.00	31.09	72	39844.00	553.39

- A statistic is specified in TABLE dimension with *

```
TABLE quantvar*statistic;
```

- Nest statistic within *catvar*

```
TABLE catvar*quantvar*statistic;
```

- Multiple statistics can be specified with parentheses

```
TABLE region, edu25*patents*(N SUM MEAN);
```

TABLE statistics

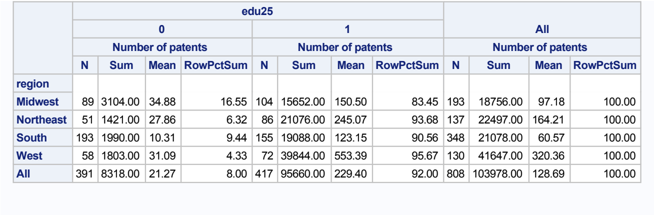
CSS	CV	KURTOSIS	LCLM	MAX
MEAN	MIN	MODE	N	NMISS
RANGE	SKEWNESS	STDEV	STDERR	SUM
SUMWGT	UCLM	USS	VAR	
PCTN	PCTSUM	REPPCTN	REPPCTSUM	PAGEPCTN
PAGEPCTSUM	ROWPCTN	ROWPCTSUM	COLPCTN	COLPCTSUM
MEDIAN	P1	P5	P10	P25
P75	P90	P95	P99	QRANGE

Statistics with ALL

SAS Code

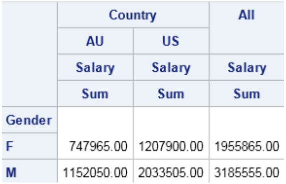
```
PROC TABULATE DATA = patents ;
  CLASS edu25 region ;
  VAR patents ;
  TABLE region ALL,
        edu25*patents*(N SUM MEAN ROWPCTSUM)
        ALL*patents*(N SUM MEAN ROWPCTSUM) ;
RUN ;
```

SAS Code



	edu25								All			
	0				1							
	Number of patents				Number of patents				Number of patents			
	N	Sum	Mean	RowPctSum	N	Sum	Mean	RowPctSum	N	Sum	Mean	RowPctSum
region												
Midwest	89	3104.00	34.88	16.55	104	15652.00	150.50	83.45	193	18756.00	97.18	100.00
Northeast	51	1421.00	27.86	6.32	86	21076.00	245.07	93.68	137	22497.00	164.21	100.00
South	193	1990.00	10.31	9.44	155	19088.00	123.15	90.56	348	21078.00	60.57	100.00
West	58	1803.00	31.09	4.33	72	39844.00	553.39	95.67	130	41647.00	320.36	100.00
All	391	8318.00	21.27	8.00	417	95660.00	229.40	92.00	808	103978.00	128.69	100.00

Discussion



	Country		All
	AU	US	
	Salary	Salary	Salary
	Sum	Sum	Sum
Gender			
F	747965.00	1207900.00	1955865.00
M	1152050.00	2033505.00	3185555.00

The statement that generated this table is:

1. `TABLE gender, country, ALL ;`
2. `TABLE gender, country, ALL*salary ;`
3. `TABLE gender, country*salary ALL ;`
4. `TABLE gender, country*salary ALL*salary ;`
5. `TABLE gender, country*SUM, ALL*SUM ;`

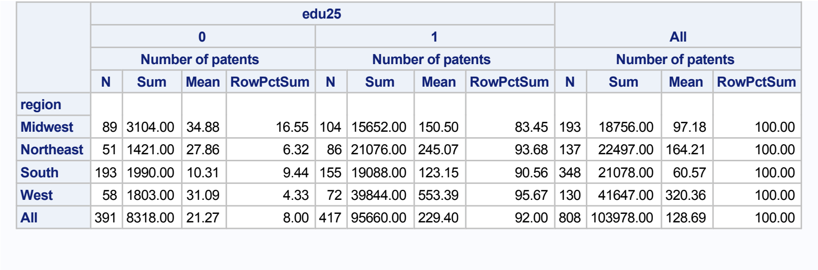
Overview

Structure

Statistics

Display

Discussion



	edu25											
	0				1				All			
	Number of patents				Number of patents				Number of patents			
	N	Sum	Mean	RowPctSum	N	Sum	Mean	RowPctSum	N	Sum	Mean	RowPctSum
region												
Midwest	89	3104.00	34.88	16.55	104	15652.00	150.50	83.45	193	18756.00	97.18	100.00
Northeast	51	1421.00	27.86	6.32	86	21076.00	245.07	93.68	137	22497.00	164.21	100.00
South	193	1990.00	10.31	9.44	155	19088.00	123.15	90.56	348	21078.00	60.57	100.00
West	58	1803.00	31.09	4.33	72	39844.00	553.39	95.67	130	41647.00	320.36	100.00
All	391	8318.00	21.27	8.00	417	95660.00	229.40	92.00	808	103978.00	128.69	100.00

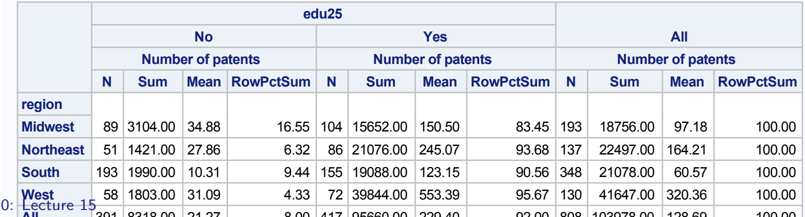
On your own: What are some things you would like to change about this table?

Apply formats to variable values

SAS Code

```
PROC FORMAT ;  VALUE yn 1 = "Yes" 0 = "No" ; RUN ;
PROC TABULATE DATA = patents ;
   CLASS edu25 region;
   VAR patents;
   TABLE region ALL,
         edu25*patents*(N SUM MEAN ROWPCTSUM)
         ALL*patents*(N SUM MEAN ROWPCTSUM);
    FORMAT edu25 yn. ;
RUN;
```

SAS Code



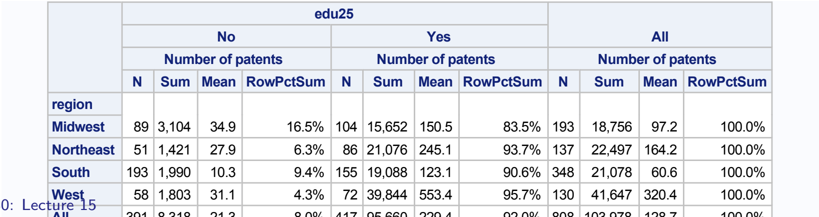
	edu25								All			
	No				Yes							
	Number of patents				Number of patents				Number of patents			
	N	Sum	Mean	RowPctSum	N	Sum	Mean	RowPctSum	N	Sum	Mean	RowPctSum
region												
Midwest	89	3104.00	34.88	16.55	104	15652.00	150.50	83.45	193	18756.00	97.18	100.00
Northeast	51	1421.00	27.86	6.32	86	21076.00	245.07	93.68	137	22497.00	164.21	100.00
South	193	1990.00	10.31	9.44	155	19088.00	123.15	90.56	348	21078.00	60.57	100.00
West	58	1803.00	31.09	4.33	72	39844.00	553.39	95.67	130	41647.00	320.36	100.00
All	391	8318.00	21.27	8.00	417	95660.00	229.40	92.00	808	103978.00	128.69	100.00

Apply formats to statistics

SAS Code

```
PROC FORMAT ;  PICTURE pct(ROUND) low-high = ‘009.9%’;  RUN;
PROC TABULATE DATA = patents ;
   CLASS edu25 region;
   VAR patents;
   TABLE region ALL,
   edu25*patents*(N SUM*F=COMMA7. MEAN*F=COMMA5.1 ROWPCTSUM*F=PCT.)
   ALL*patents*(N SUM*F=COMMA7. MEAN*F=COMMA5.1 ROWPCTSUM*F=PCT.);
   FORMAT edu25 yn. ;
RUN;
```

SAS Code



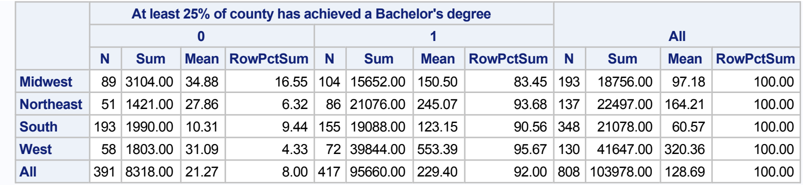
	edu25								All			
	No				Yes							
	Number of patents				Number of patents				Number of patents			
	N	Sum	Mean	RowPctSum	N	Sum	Mean	RowPctSum	N	Sum	Mean	RowPctSum
region												
Midwest	89	3,104	34.9	16.5%	104	15,652	150.5	83.5%	193	18,756	97.2	100.0%
Northeast	51	1,421	27.9	6.3%	86	21,076	245.1	93.7%	137	22,497	164.2	100.0%
South	193	1,990	10.3	9.4%	155	19,088	123.1	90.6%	348	21,078	60.6	100.0%
West	58	1,803	31.1	4.3%	72	39,844	553.4	95.7%	130	41,647	320.4	100.0%
All	391	8,318	21.3	8.0%	417	95,660	229.4	92.0%	808	103,978	128.7	100.0%

Basic Labels

SAS Code

```
PROC TABULATE DATA = patents;
   CLASS edu25 region ;
   VAR patents;
   TABLE region=" "  ALL,
     edu25*patents=" "*(N SUM MEAN ROWPCTSUM)
     ALL*patents=" "*(N SUM MEAN ROWPCTSUM) ;
   LABEL edu25="At least 25% of county has a Bachelor’s";
RUN;
```

SAS Code



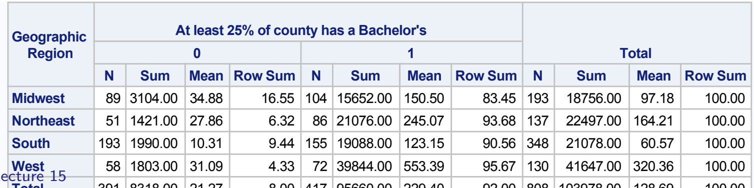
	At least 25% of county has achieved a Bachelor's degree								All			
	0				1							
	N	Sum	Mean	RowPctSum	N	Sum	Mean	RowPctSum	N	Sum	Mean	RowPctSum
Midwest	89	3104.00	34.88	16.55	104	15652.00	150.50	83.45	193	18756.00	97.18	100.00
Northeast	51	1421.00	27.86	6.32	86	21076.00	245.07	93.68	137	22497.00	164.21	100.00
South	193	1990.00	10.31	9.44	155	19088.00	123.15	90.56	348	21078.00	60.57	100.00
West	58	1803.00	31.09	4.33	72	39844.00	553.39	95.67	130	41647.00	320.36	100.00
All	391	8318.00	21.27	8.00	417	95660.00	229.40	92.00	808	103978.00	128.69	100.00

KeyLabel and Box

SAS Code

```
PROC TABULATE DATA = patents;
   CLASS edu25 region ;
   VAR patents;
   TABLE region=" "  ALL,
         edu25*patents=" "*(N SUM MEAN ROWPCTSUM)
         ALL*patents=" "*(N SUM MEAN ROWPCTSUM) /
         BOX = "Geographic Region";
   LABEL edu25="At least 25% of county has a Bachelor's";
   KEYLABEL ALL="Total" ROWPCTSUM="Row Sum" ;
RUN;
```

SAS Code



Geographic Region	At least 25% of county has a Bachelor's								Total			
	0				1							
	N	Sum	Mean	Row Sum	N	Sum	Mean	Row Sum	N	Sum	Mean	Row Sum
Midwest	89	3104.00	34.88	16.55	104	15652.00	150.50	83.45	193	18756.00	97.18	100.00
Northeast	51	1421.00	27.86	6.32	86	21076.00	245.07	93.68	137	22497.00	164.21	100.00
South	193	1990.00	10.31	9.44	155	19088.00	123.15	90.56	348	21078.00	60.57	100.00
West	58	1803.00	31.09	4.33	72	39844.00	553.39	95.67	130	41647.00	320.36	100.00
Total	391	8318.00	21.27	8.00	417	95660.00	229.40	92.00	808	103978.00	128.69	100.00

Cell colors

- To apply a background color to all cells, use the following in a TABLE statement:

  ```
  variable*{STYLE={BACKGROUND=mycolor}}
  ```

- To highlight individual cells based on their values (trafficlighting)
  1. Create a format that specifies color based on values

     ```
     PROC FORMAT; VALUE myhl 95-high="mycolor"; RUN;
     ```

  2. Apply the format to the background style in the TABLE statement

     ```
     statistic*{STYLE={BACKGROUND=myhl.}}
     ```

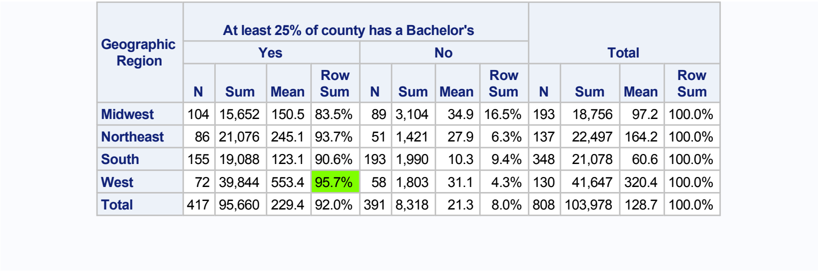
- Predefined SAS colors: http://support.sas.com/documentation/cdl/en/graphref/67881/HTML/default/viewer.htm#n161ukdyz9wpfsn1nh8sihforvyq.htm

Highlight cells

SAS Code

```
PROC FORMAT ;  VALUE hlpct 95-high="Chartreuse" ; RUN ;
PROC TABULATE DATA=patents;
CLASS region;
CLASS edu25 / DESCENDING;
VAR patents;
TABLE region=" "  ALL,
edu25*patents=" "*
  (N SUM*F=COMMA7.
   MEAN*F=COMMA5.1
   ROWPCTSUM*F=PCT.*{STYLE={BACKGROUND=HLPCT.}})
ALL*patents=" "*
  (N SUM*F=COMMA7. MEAN*F=COMMA5.1 ROWPCTSUM*F=PCT.) /
BOX="Geographic Region";
LABEL edu25="At least 25% of county has a Bachelor's";
KEYLABEL ALL="Total" ROWPCTSUM="Row Sum" ;
FORMAT edu25 yn.;
RUN;
```

Final table



Geographic Region	At least 25% of county has a Bachelor's								Total			
	Yes				No							
	N	Sum	Mean	Row Sum	N	Sum	Mean	Row Sum	N	Sum	Mean	Row Sum
Midwest	104	15,652	150.5	83.5%	89	3,104	34.9	16.5%	193	18,756	97.2	100.0%
Northeast	86	21,076	245.1	93.7%	51	1,421	27.9	6.3%	137	22,497	164.2	100.0%
South	155	19,088	123.1	90.6%	193	1,990	10.3	9.4%	348	21,078	60.6	100.0%
West	72	39,844	553.4	95.7%	58	1,803	31.1	4.3%	130	41,647	320.4	100.0%
Total	417	95,660	229.4	92.0%	391	8,318	21.3	8.0%	808	103,978	128.7	100.0%