Satoshi Ido

34788706

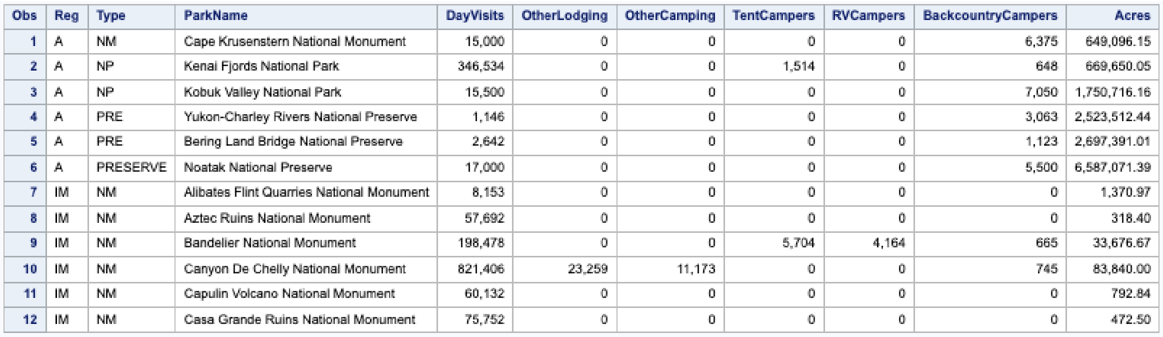
09/15/2023

STAT506

HW2

1. Initial Exploration of the National Parks Summary Data with Various Procedures
    a. Write a PROC PRINT step to display only the first 12 observations in pg1.np_summary. Show your code and the output.

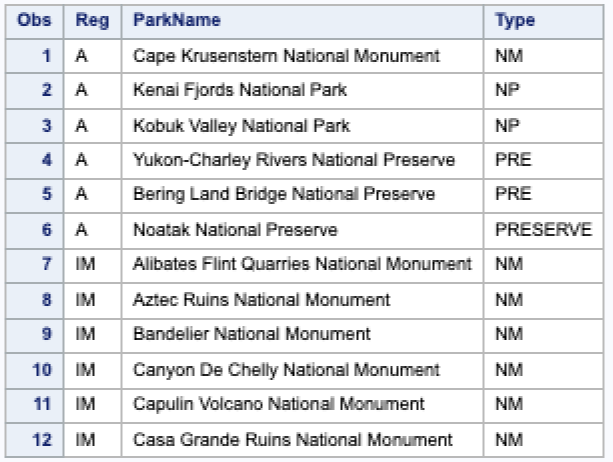
```
proc print data=pg1.np_summary (obs=12);
       run;
```



Obs	Reg	Type	ParkName	DayVisits	OtherLodging	OtherCamping	TentCampers	RVCampers	BackcountryCampers	Acres
1	A	NM	Cape Krusenstern National Monument	15,000	0	0	0	0	6,375	649,096.15
2	A	NP	Kenai Fjords National Park	346,534	0	0	1,514	0	648	669,650.05
3	A	NP	Kobuk Valley National Park	15,500	0	0	0	0	7,050	1,750,716.16
4	A	PRE	Yukon-Charley Rivers National Preserve	1,146	0	0	0	0	3,063	2,523,512.44
5	A	PRE	Bering Land Bridge National Preserve	2,642	0	0	0	0	1,123	2,697,391.01
6	A	PRESERVE	Noatak National Preserve	17,000	0	0	0	0	5,500	6,587,071.39
7	IM	NM	Alibates Flint Quarries National Monument	8,153	0	0	0	0	0	1,370.97
8	IM	NM	Aztec Ruins National Monument	57,692	0	0	0	0	0	318.40
9	IM	NM	Bandelier National Monument	198,478	0	0	5,704	4,164	665	33,676.67
10	IM	NM	Canyon De Chelly National Monument	821,406	23,259	11,173	0	0	745	83,840.00
11	IM	NM	Capulin Volcano National Monument	60,132	0	0	0	0	0	792.84
12	IM	NM	Casa Grande Ruins National Monument	75,752	0	0	0	0	0	472.50

    b. Add a VAR statement to the PROC PRINT step (from part a) to include only the variables Reg, ParkName, and Type (in that order). Notice that Type is based on ParkName. Do you observe any possible inconsistencies in the Type abbreviations used for the different types of parks? Show your code and output, and answer the question.

```
proc print data=pg1.np_summary (obs=12);
       var Reg ParkName Type;
       run;
```

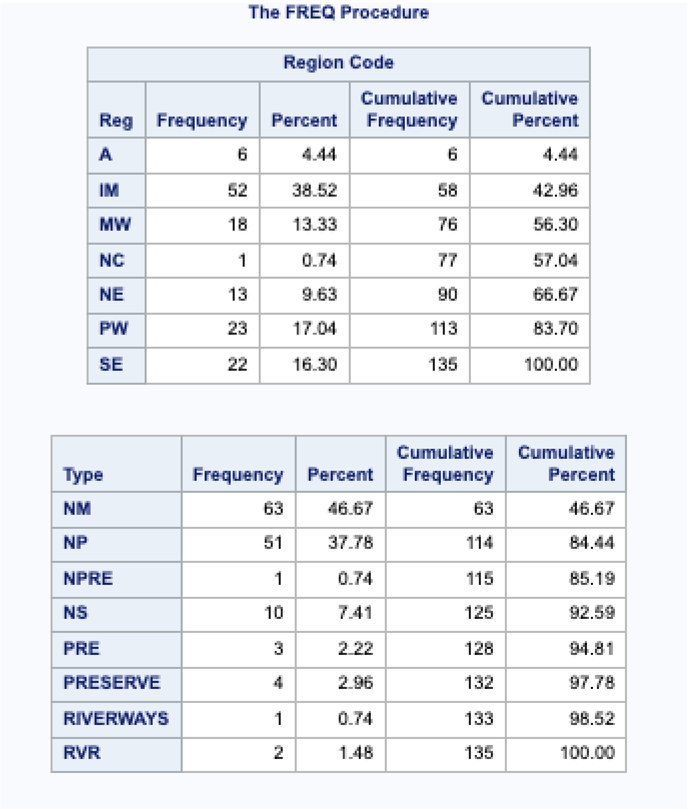


Obs	Reg	ParkName	Type
1	A	Cape Krusenstern National Monument	NM
2	A	Kenai Fjords National Park	NP
3	A	Kobuk Valley National Park	NP
4	A	Yukon-Charley Rivers National Preserve	PRE
5	A	Bering Land Bridge National Preserve	PRE
6	A	Noatak National Preserve	PRESERVE
7	IM	Alibates Flint Quarries National Monument	NM
8	IM	Aztec Ruins National Monument	NM
9	IM	Bandelier National Monument	NM
10	IM	Canyon De Chelly National Monument	NM
11	IM	Capulin Volcano National Monument	NM
12	IM	Casa Grande Ruins National Monument	NM

Yes, there is some inconsistency in the “Type” column. In row 6, Type value should be an abbreviation of ParkName. Therefore, it should be “PRE”, yet it is actually “PRESERVE.”

c. Now using all the observations in pg1.np_summary, write a PROC FREQ step that uses a TABLES statement to produce separate frequency tables for Reg and Type. Which codes/values appear only once each in pg1.np_summary for these variables? Show your code and output, and answer the question.

```
proc freq data=pg1.np_summary;
       table Reg Type;
       run;
```



The FREQ Procedure

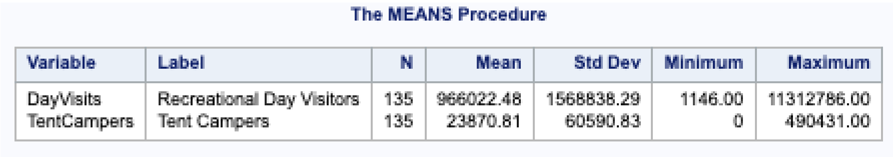
Region Code				
Reg	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	6	4.44	6	4.44
IM	52	38.52	58	42.96
MW	18	13.33	76	56.30
NC	1	0.74	77	57.04
NE	13	9.63	90	66.67
PW	23	17.04	113	83.70
SE	22	16.30	135	100.00

Type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NM	63	46.67	63	46.67
NP	51	37.78	114	84.44
NPRE	1	0.74	115	85.19
NS	10	7.41	125	92.59
PRE	3	2.22	128	94.81
PRESERVE	4	2.96	132	97.78
RIVERWAYS	1	0.74	133	98.52
RVR	2	1.48	135	100.00

From the output, we can see “NC” appeared only once in Region Code, while “NPRE” and “RIVERWAYS” appeared only once in Type respectively.

d. Write a PROC MEANS step for all the observations in pg1.np_summary. Calculate summary statistics for just the DayVisits and TentCampers columns. What are the minimum values for the number of recreational day visitors and for the number of tent campers? Show your code and output, and answer the questions.

```
proc means data=pg1.np_summary;
       var DayVisits TentCampers;
       run;
```



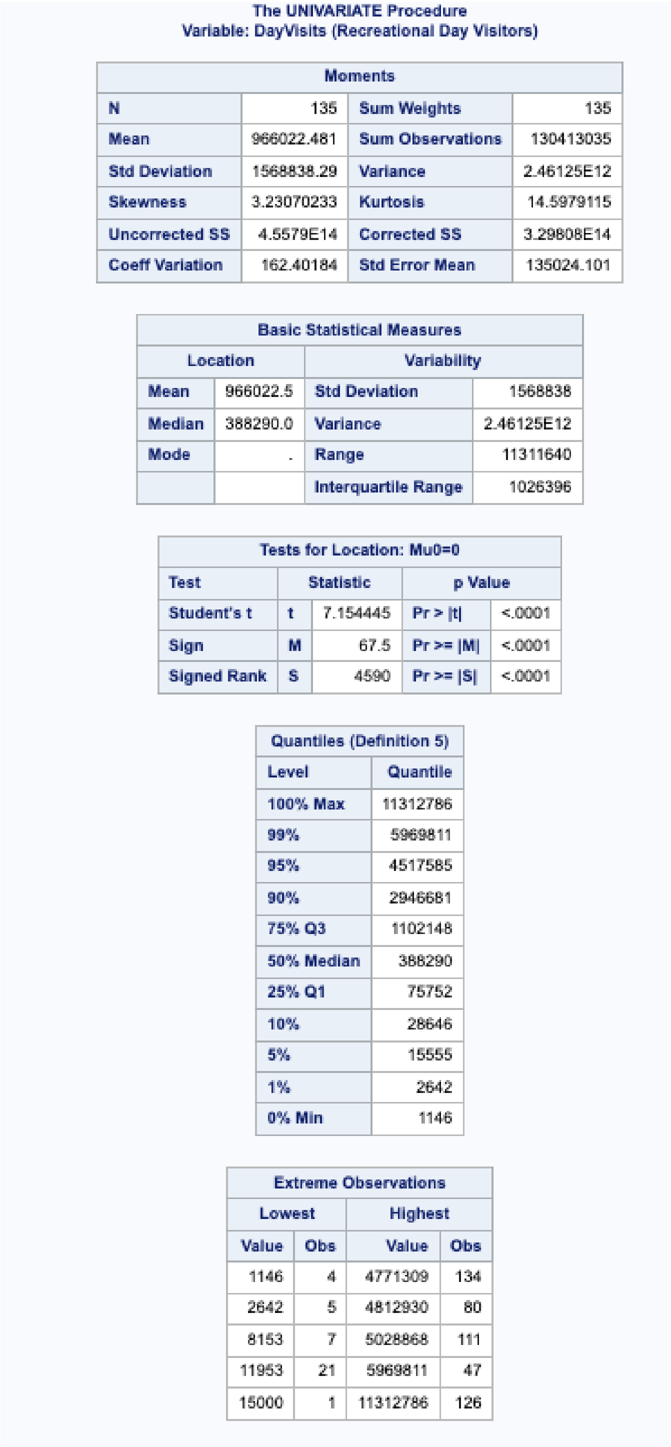
The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
DayVisits TentCampers	Recreational Day Visitors Tent Campers	135 135	966022.48 23870.81	1568838.29 60590.83	1146.00 0	11312786.00 490431.00

The minimum value for the number of recreational day visitors is 1146.00, and the minimum value for the number of tent campers is 0.

e. Write a PROC UNIVARIATE step for all the observations in pg1.np_summary. Calculate summary statistics for just the DayVisits variable. What are the two lowest values and two highest values of DayVisits? Show your code and the relevant part of the output, and answer the questions.

```
proc univariate data=pg1.np_summary;
       var DayVisits;
       run;
```



The UNIVARIATE Procedure
Variable: DayVisits (Recreational Day Visitors)

Moments			
N	135	Sum Weights	135
Mean	966022.481	Sum Observations	130413035
Std Deviation	1568838.29	Variance	2.46125E12
Skewness	3.23070233	Kurtosis	14.5979115
Uncorrected SS	4.5579E14	Corrected SS	3.29808E14
Coeff Variation	162.40184	Std Error Mean	135024.101

Basic Statistical Measures			
Location		Variability	
Mean	966022.5	Std Deviation	1568838
Median	388290.0	Variance	2.46125E12
Mode	.	Range	11311640
		Interquartile Range	1026396

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	7.154445	Pr > |t|	<.0001
Sign	M	67.5	Pr >= |M|	<.0001
Signed Rank	S	4590	Pr >= |S|	<.0001

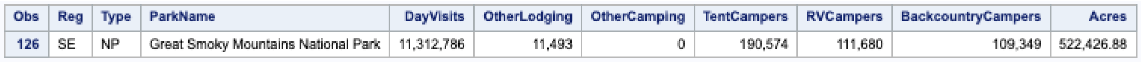
Quantiles (Definition 5)	
Level	Quantile
100% Max	11312786
99%	5969811
95%	4517585
90%	2946681
75% Q3	1102148
50% Median	388290
25% Q1	75752
10%	28646
5%	15555
1%	2642
0% Min	1146

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1146	4	4771309	134
2642	5	4812930	80
8153	7	5028868	111
11953	21	5969811	47
15000	1	11312786	126

The two lowest values are `1146` and `2642.` The two highest values are `11312786` and `5969811.`

f. Write a PROC PRINT step and use a WHERE statement to display only the row/observation that had the maximum number of DayVisits. (It's OK to just hardcode in a value here.) Show your code and output.

```
proc print data=pg1.np_summary;
	where DayVisits = 11312786;
	run;
```

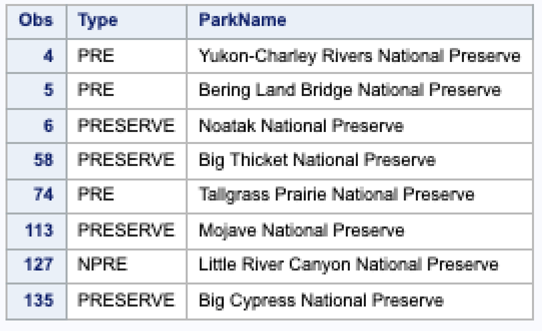


Obs	Reg	Type	ParkName	DayVisits	OtherLodging	OtherCamping	TentCampers	RVCampers	BackcountryCampers	Acres
126	SE	NP	Great Smoky Mountains National Park	11,312,786	11,493	0	190,574	111,680	109,349	522,426.88

2. Further exploring the National Parks Summary Data

a. Open the program p103p04.sas (from the "practices" folder). Add a WHERE statement to print only the rows where ParkName includes the word "Preserve" anywhere in the name of the park using wildcards. What codes (in Type) are currently being used to denote Preserves? Show your code and output, and answer the question.

```
proc print data=pg1.np_summary;
	var Type ParkName;
	*Add a WHERE statement;
	where ParkName like "%Preserve%";
run;
```



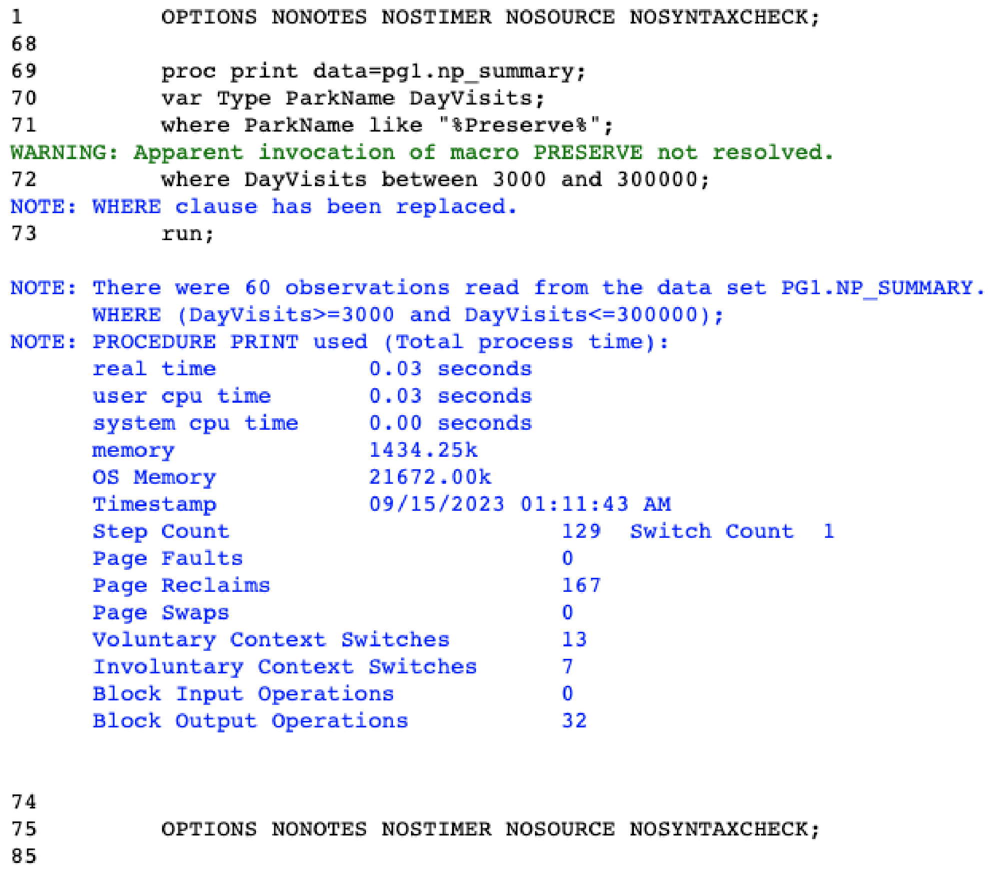
Obs	Type	ParkName
4	PRE	Yukon-Charley Rivers National Preserve
5	PRE	Bering Land Bridge National Preserve
6	PRESERVE	Noatak National Preserve
58	PRESERVE	Big Thicket National Preserve
74	PRE	Tallgrass Prairie National Preserve
113	PRESERVE	Mojave National Preserve
127	NPRE	Little River Canyon National Preserve
135	PRESERVE	Big Cypress National Preserve

From the output, we can see that codes PRE, NPRE and PRESERVE are used as Type to denote Preserves.

b. Edit the VAR statement to additionally include the DayVisits variable. Add a second WHERE statement (below the previous one) to include only observations that had between 3,000 and 300,000 (inclusive) Recreational Day Visitors. Run the code to see if you get the expected results. Show your code and the corresponding Log notes.

```
proc print data=pg1.np_summary;
	var Type ParkName DayVisits;
	where ParkName like "%Preserve%";
```

```
      where DayVisits between 3000 and 300000;
      run;
```



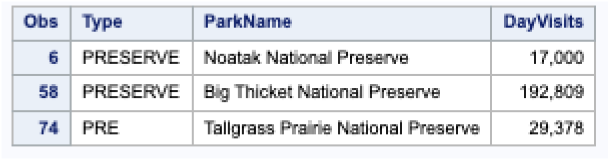
```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         proc print data=pg1.np_summary;
70         var Type ParkName DayVisits;
71         where ParkName like "%Preserve%";
WARNING: Apparent invocation of macro PRESERVE not resolved.
72         where DayVisits between 3000 and 300000;
NOTE: WHERE clause has been replaced.
73         run;

NOTE: There were 60 observations read from the data set PG1.NP_SUMMARY.
      WHERE (DayVisits>=3000 and DayVisits<=300000);
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.03 seconds
      user cpu time       0.03 seconds
      system cpu time     0.00 seconds
      memory              1434.25k
      OS Memory           21672.00k
      Timestamp           09/15/2023 01:11:43 AM
      Step Count                        129  Switch Count  1
      Page Faults                       0
      Page Reclaims                     167
      Page Swaps                        0
      Voluntary Context Switches        13
      Involuntary Context Switches      7
      Block Input Operations            0
      Block Output Operations           32


74
75         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
85
```

c. Combine the two previous WHERE statements into one WHERE statement that uses both conditions (Preserves with between 3,000 and 300,000 visitors) for subsetting. Show your code and output.

```
proc print data=pg1.np_summary;
      var Type ParkName DayVisits;
      where ParkName like "%Preserve%" and DayVisits between 3000 and
300000;
      run;
```



Obs	Type	ParkName	DayVisits
6	PRESERVE	Noatak National Preserve	17,000
58	PRESERVE	Big Thicket National Preserve	192,809
74	PRE	Tallgrass Prairie National Preserve	29,378

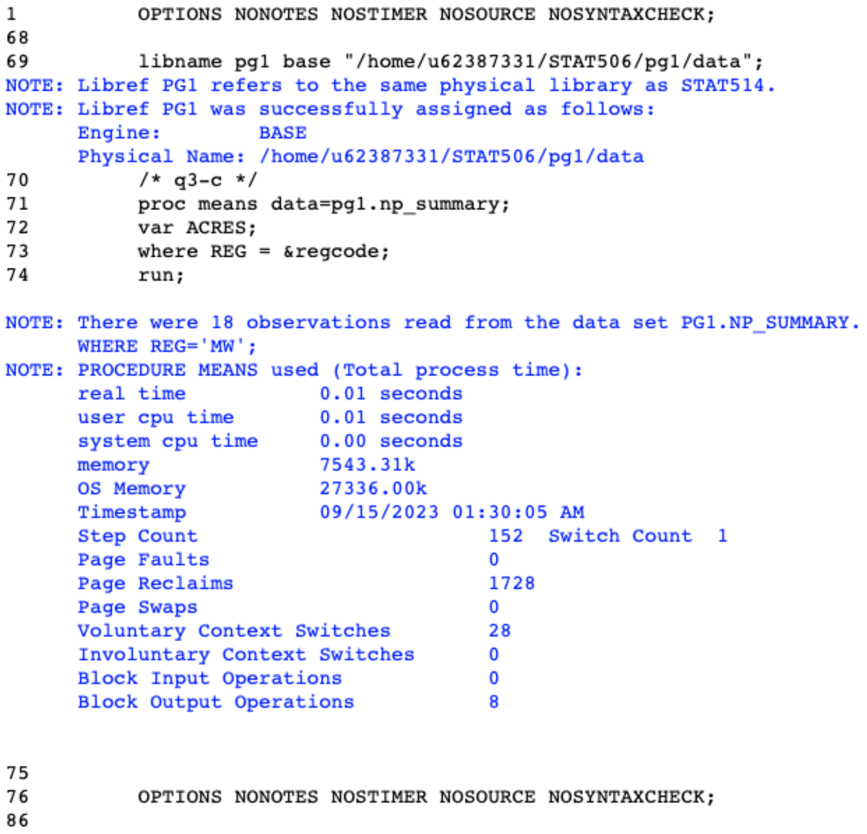
3. Using a Macro Variable

   a. Create a macro variable named regcode and use it to store the text “MW”. Show your code.

```
%let regcode="MW";
```

   b. Write a PROC MEANS step to calculate summary statistics for the variable ACRES in pg1.np_summary. Use a WHERE statement to only include observations with the variable REG equal to your macro variable regcode. If done properly, this should be 18 observations. Show your code, corresponding log notes, and output.

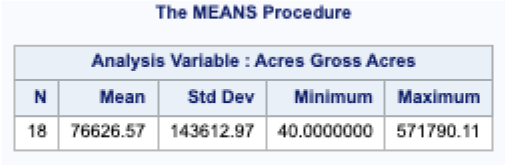
```
proc means data=pg1.np_summary;
        var ACRES;
        where REG = &regcode;
        run;
```



```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         libname pg1 base "/home/u62387331/STAT506/pg1/data";
NOTE: Libref PG1 refers to the same physical library as STAT514.
NOTE: Libref PG1 was successfully assigned as follows:
      Engine:        BASE
      Physical Name: /home/u62387331/STAT506/pg1/data
70         /* q3-c */
71         proc means data=pg1.np_summary;
72         var ACRES;
73         where REG = &regcode;
74         run;

NOTE: There were 18 observations read from the data set PG1.NP_SUMMARY.
      WHERE REG='MW';
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.01 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              7543.31k
      OS Memory           27336.00k
      Timestamp           09/15/2023 01:30:05 AM
      Step Count                        152  Switch Count  1
      Page Faults                       0
      Page Reclaims                     1728
      Page Swaps                        0
      Voluntary Context Switches        28
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           8


75
76         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
86
```



The MEANS Procedure

Analysis Variable : Acres Gross Acres				
N	Mean	Std Dev	Minimum	Maximum
18	76626.57	143612.97	40.0000000	571790.11

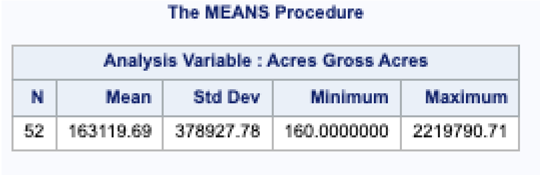
c. Change the value stored in the regcode macro variable to “IM”. Rerun that statement and rerun the same PROC MEANS step as before. This time, there should be 52 observations included. Show your code, corresponding log notes, and output.

```
%let regcode="IM";
proc means data=pg1.np_summary;
       var ACRES;
       where REG = &regcode;
       run;
```

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         %let regcode="IM";
70         proc means data=pg1.np_summary;
71         var ACRES;
72         where REG = &regcode;
73         run;

NOTE: There were 52 observations read from the data set PG1.NP_SUMMARY.
      WHERE REG='IM';
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.01 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              7538.96k
      OS Memory           27336.00k
      Timestamp           09/15/2023 01:32:32 AM
      Step Count                        158  Switch Count  1
      Page Faults                       0
      Page Reclaims                     1704
      Page Swaps                        0
      Voluntary Context Switches        26
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           8


74
75         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
85
```

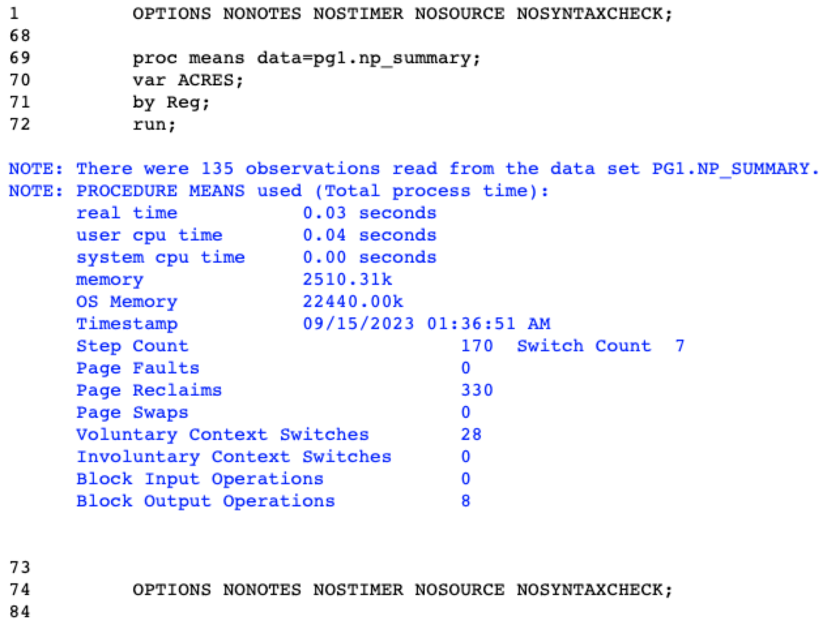


The MEANS Procedure

Analysis Variable : Acres Gross Acres				
N	Mean	Std Dev	Minimum	Maximum
52	163119.69	378927.78	160.0000000	2219790.71

d. Remove the WHERE statement from the PROC MEANS step and replace it with the statement: BY reg; Run the edited step and observe the output. Show just your code and corresponding log notes.

```
proc means data=pg1.np_summary;
       var ACRES;
       by Reg;
       run;
```



```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         proc means data=pg1.np_summary;
70         var ACRES;
71         by Reg;
72         run;

NOTE: There were 135 observations read from the data set PG1.NP_SUMMARY.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.03 seconds
      user cpu time       0.04 seconds
      system cpu time     0.00 seconds
      memory              2510.31k
      OS Memory           22440.00k
      Timestamp           09/15/2023 01:36:51 AM
      Step Count                        170  Switch Count  7
      Page Faults                       0
      Page Reclaims                     330
      Page Swaps                        0
      Voluntary Context Switches        28
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           8


73
74         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
84
```

4. Using Formats

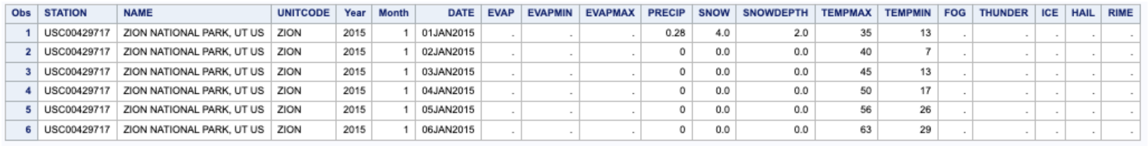
    a. Write a step to examine the descriptor portion of the pg1.np_westweather table. Which format is currently being used to display the DATE variable? Show your code and answer the question.

```
proc contents data=pg1.np_westweather;
       run;
```

They use “YYMMDD10” format to display the DATE variable.

    b. Write a PROC PRINT step to display the first 6 observations of pg1.np_westweather. Use the DATE9. format to display DATE, and use the 4.1 format to display both SNOW and SNOWDEPTH. Show your code and output.

```
proc print data=pg1.np_westweather(obs=6);
       format DATE DATE9. SNOW SNOWDEPTH 4.1;
       run;
```



Obs	STATION	NAME	UNITCODE	Year	Month	DATE	EVAP	EVAPMIN	EVAPMAX	PRECIP	SNOW	SNOWDEPTH	TEMPMAX	TEMPMIN	FOG	THUNDER	ICE	HAIL	RIME
1	USC00429717	ZION NATIONAL PARK, UT US	ZION	2015	1	01JAN2015	.	.	.	0.28	4.0	2.0	35	13	.	.	.	.	.
2	USC00429717	ZION NATIONAL PARK, UT US	ZION	2015	1	02JAN2015	.	.	.	0	0.0	0.0	40	7	.	.	.	.	.
3	USC00429717	ZION NATIONAL PARK, UT US	ZION	2015	1	03JAN2015	.	.	.	0	0.0	0.0	45	13	.	.	.	.	.
4	USC00429717	ZION NATIONAL PARK, UT US	ZION	2015	1	04JAN2015	.	.	.	0	0.0	0.0	50	17	.	.	.	.	.
5	USC00429717	ZION NATIONAL PARK, UT US	ZION	2015	1	05JAN2015	.	.	.	0	0.0	0.0	56	26	.	.	.	.	.
6	USC00429717	ZION NATIONAL PARK, UT US	ZION	2015	1	06JAN2015	.	.	.	0	0.0	0.0	63	29	.	.	.	.	.

5. Sorting the National Parks Summary Data

a. Write a PROC SORT step to read pg1.np_summary and create a temporary sorted table named np_sorted. Include a BY statement to order the data by first by Reg and then by descending DayVisits. Add a WHERE statement to select Type equal to either "NP" or "NS". Show your code and the corresponding log notes.

```
proc sort data=pg1.np_summary out=pg1.np_sorted;
       by Reg descending DayVisits;
       where Type = "NP" or Type = "NS";
       run;
```

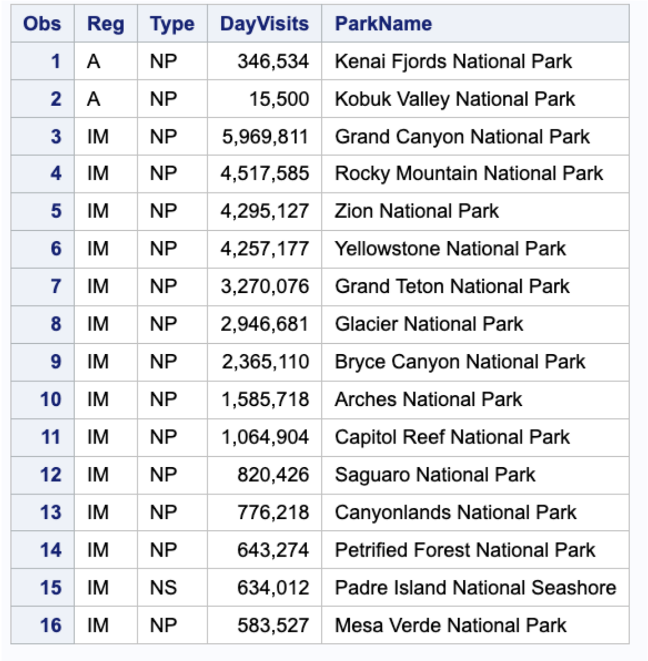
```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         proc sort data=pg1.np_summary out=pg1.np_sorted;
70         by Reg descending DayVisits;
71         where Type = "NP" or Type = "NS";
72         run;

NOTE: There were 61 observations read from the data set PG1.NP_SUMMARY.
      WHERE Type in ('NP', 'NS');
NOTE: The data set PG1.NP_SORTED has 61 observations and 10 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time           0.01 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              1205.46k
      OS Memory           20396.00k
      Timestamp           09/15/2023 07:53:01 PM
      Step Count                        41  Switch Count  2
      Page Faults                       0
      Page Reclaims                     196
      Page Swaps                        0
      Voluntary Context Switches        49
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           272


73
74         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
84
```

b. Write a PROC PRINT step to display only the first 16 observations from np_sorted and the only the variables Reg, Type, DayVisits, and ParkName (in that order). Show your code and output.

```
proc print data=pg1.np_sorted(obs=16);
       var Reg Type DayVisits ParkName;
       run;
```



Obs	Reg	Type	DayVisits	ParkName
1	A	NP	346,534	Kenai Fjords National Park
2	A	NP	15,500	Kobuk Valley National Park
3	IM	NP	5,969,811	Grand Canyon National Park
4	IM	NP	4,517,585	Rocky Mountain National Park
5	IM	NP	4,295,127	Zion National Park
6	IM	NP	4,257,177	Yellowstone National Park
7	IM	NP	3,270,076	Grand Teton National Park
8	IM	NP	2,946,681	Glacier National Park
9	IM	NP	2,365,110	Bryce Canyon National Park
10	IM	NP	1,585,718	Arches National Park
11	IM	NP	1,064,904	Capitol Reef National Park
12	IM	NP	820,426	Saguaro National Park
13	IM	NP	776,218	Canyonlands National Park
14	IM	NP	643,274	Petrified Forest National Park
15	IM	NS	634,012	Padre Island National Seashore
16	IM	NP	583,527	Mesa Verde National Park

6. Using PROC SORT to Subset a Table
    a. Write a PROC SORT step which will split the pg1.np_westweather table into two new temporary tables named newyearsdays and others. The table newyearsdays should include just the first recorded observation for each unique occurrence of the variables NAME and YEAR. For example, the first observation in newyearsdays should be for Death Valley on Jan. 1, 2015. The second observation should be for Death Valley on Jan. 1, 2016, etc. The table others should include all the other observations from the original table.
    [Note that pg1.np_westweather is helpfully already sorted by date. Use the nodupkey option and other corresponding syntax in your PROC SORT. The table newyearsdays should contain 12 observations.]
    Show your code and corresponding log output.

```
proc sort data=pg1.np_westweather out=pg1.newyearsdays
      nodupkey dupout=pg1.others;
      by NAME YEAR;
      run;
```

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         proc sort data=pg1.np_westweather out=pg1.newyearsdays
70         nodupkey dupout=pg1.others;
71         by NAME YEAR;
72         run;

NOTE: There were 4355 observations read from the data set PG1.NP_WESTWEATHER.
NOTE: 4343 observations with duplicate key values were deleted.
NOTE: The data set PG1.NEWYEARSDAYS has 12 observations and 19 variables.
NOTE: The data set PG1.OTHERS has 4343 observations and 19 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time           0.03 seconds
      user cpu time       0.00 seconds
      system cpu time     0.01 seconds
      memory              4582.93k
      OS Memory           25272.00k
      Timestamp           09/15/2023 08:31:44 PM
      Step Count                        83  Switch Count  4
      Page Faults                       0
      Page Reclaims                     1015
      Page Swaps                        0
      Voluntary Context Switches        106
      Involuntary Context Switches      0
      Block Input Operations            1792
      Block Output Operations           2328


73
74         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
84
```