

# B3.1 - Exploring Data


<p style="text-align: center;">SAS Programming Process</p> <p style="text-align: center;">3</p>	<p>Now that we know how to access data, we need to explore it a little and see what we have to work with. Exploring data can include learning about the columns and values we have, as well as validating data to look for incorrect or inconsistent values. In this lesson, you learn to use some procedures that give you some of this insight. You also learn to subset the data so you can focus on particular segments, format data so you can easily understand it, sort data, and identify and clean up duplicate values.</p>
<p style="text-align: center;">Exploring Data with Procedures</p> <p style="text-align: center;">4</p>	<p>After you are able to access your data, the next step is to make sure you understand it. Certainly just looking over the table is helpful, and PROC CONTENTS enables you to confirm column attributes, but often the data is too large or complex for a visual review to be sufficient.</p> <p>SAS has several procedures that can be used to quickly and easily explore your data. Later we use some of these same procedures with additional options to analyze and report on data.</p>

## Proc Print

### PRINT Procedure


```
PROC PRINT DATA=input-table (OBS=n);
RUN;
```

use to specify the last observation (row) to read




By default, PROC PRINT lists all columns and rows in the input table.

By default, the PRINT procedure lists all columns and rows in the input table. To limit the number of rows listed, especially if it's a large table, you can use the OBS= data set option in parentheses after the table name.

5
p103d01 

## Setup for the following Question:

- Go to [support.sas.com/documentation](https://support.sas.com/documentation/)  [\(https://support.sas.com/documentation/\)](https://support.sas.com/documentation/). Click **9.4** after **SAS Procedures** by **Name and Product**.
- Click **SAS Procedures by Name** and find **PRINT**. Examine the syntax and the table of procedure tasks and examples.

### PRINT Procedure

Syntax ▾
Overview
Concepts
Using ▾
Examples ▾

**Interaction:** A common practice is to sort a data set using PROC SORT before you use the PROC PRINT BY statement. If you sort a CAS table with VARCHAR variables using PROC SORT, VARCHAR variables are converted to CHAR variables.

**Note:** PROC PRINT supports the VARCHAR data type for CAS tables.

**Tips:** Each password and encryption key option must be coded on a separate line to ensure that they are properly blotted in the log.

Supports the Output Delivery System. For details, see [Output Delivery System: Basic Concepts in SAS Output Delivery System: User's Guide](#).

You can use the ATTRB, FORMAT, LABEL, TITLE, and WHERE statements. See [SAS DATA Step Statements: Reference](#). For more information, see [Statements with the Same Function in Multiple Procedures](#).

Syntax

[Table of Procedure Tasks and Examples](#)

**Syntax**

```
PROC PRINT <option(s)>;
  BY <DESCENDING> variable-1 <<DESCENDING> variable-2 ...> <NOTSORTED>;
  PAGEBY <f>variable;
  SUMVAR <f>variable;
```

*Note: The following activity is for you to answer questions and for you to get your response. It also serves as a checking point for if you are understanding the material. The following activity is not graded.*

## 3.01 Question

saransh707@gmail.com [Switch accounts](#)



Not shared

Which statement in PROC PRINT selects variables that appear in the report and determines their order? 1 point

- ☐ BY
- ☐ ID
- ☐ SUM
- ☐ VAR

Submit

Clear form

Google Forms

This form was created inside Clemson University.



### PRINT Procedure

```
proc print data=sashelp.cars (obs=10);  
  var Make Model Type MSRP;  
run;
```

Obs	Make	Model	Type	MSRP
1	Acura	MDX	SUV	\$36,945
2	Acura	RSX Type S 2dr	Sedan	\$23,820
3	Acura	TSX 4dr	Sedan	\$26,990
4	Acura	TL 4dr	Sedan	\$33,195
5	Acura	3.5 RL 4dr	Sedan	\$43,755
6	Acura	3.5 RL w/Navigation 4dr	Sedan	\$46,100
7	Acura	NSX coupe 2dr manual S	Sports	\$89,765
8	Audi	A4 1.8T 4dr	Sedan	\$25,940
9	Audi	A41.8T convertible 2dr	Sedan	\$35,940
10	Audi	A4 3.0 4dr	Sedan	\$31,840

This PROC PRINT step lists the first 10 rows, or observations, from the **sashelp.cars** table and displays only the **Make**, **Model**, **Type**, and **MSRP** columns.

## Proc Means

## MEANS Procedure

```
PROC MEANS DATA=input-table;
  VAR col-name(s);
RUN;
```

use to specify the  
numeric columns  
to analyze

By default, PROC MEANS  
generates simple  
summary statistics for  
each numeric column  
in the input data.



10

Copyright © SAS Institute Inc. All rights reserved.

p103d01



The MEANS procedure generates simple summary statistics for each numeric column in the input data. You can use the VAR statement to limit the columns, or variables, that SAS analyzes.

## MEANS Procedure

```
proc means data=sashelp.cars;
  var EngineSize Horsepower MPG_City MPG_Highway;
run;
```

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
EngineSize	Engine Size (L)	428	3.1967290	1.1085947	1.3000000	8.3000000
Horsepower		428	215.8855140	71.8360316	73.0000000	500.0000000
MPG_City	MPG (City)	428	20.0607477	5.2382176	10.0000000	60.0000000
MPG_Highway	MPG (Highway)	428	26.8434579	5.7412007	12.0000000	66.0000000

11

Copyright © SAS Institute Inc. All rights reserved.

p103d01



This PROC MEANS step calculates the default statistics – frequency count (N), mean, standard deviation, minimum, and maximum – for each of the Unithat is listed in the VAR statement.

By examining the PROC MEANS results, you can identify average values, or values that might be outside of an expected range.

# Univariate Procedure

## UNIVARIATE Procedure

```
PROC UNIVARIATE DATA=input-table;
  VAR col-name(s);
RUN;
```

use to specify the  
numeric columns  
to analyze

By default, PROC  
UNIVARIATE generates  
summary statistics for  
each numeric column  
in the input data.



12

Copyright © SAS Institute Inc. All rights reserved.

p103d01



The UNIVARIATE procedure also generates summary statistics, but it includes more detailed statistics related to distribution and extreme values. Notice that it also uses a VAR statement like PROC PRINT and PROC MEANS to select the columns to analyze.

## UNIVARIATE Procedure

```
proc univariate data=sashelp.cars;
  var MPG_Highway;
run;
```

The UNIVARIATE Procedure  
Variable: MPG\_Highway (MPG [highway])

Moments			
N	428	Sum Weights	428
Mean	26.84365	Sum Observations	11480
Std Deviation	5.7412007	Variance	32.9613867
Skewness	1.26239627	Kurtosis	6.04651088
Uncorrected SS	322479	Corrected SS	14074.5117
Coeff Variation	21.3877092	Std Error Mean	0.27751141

Basic Statistical Measures			
Location		Variability	
Mean	26.84365	Std Deviation	5.74120
Median	26.00000	Variance	32.96139
Mode	26.00000	Range	54.00000
	Interquartile Range		5.00000

Tests for Location: Mu=0		
Test	Statistic	p Value
Student's t	1	96.7292 Pr >  t  < .0001
Sign	11	214 Pr >=  S  < .0001
Signed Rank	5	40303 Pr >=  Z  < .0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	66
99%	44
95%	36
90%	34
75% Q3	29
50% Median	26
25% Q1	24
10%	20
5%	18
1%	16
0% Min	12

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
12	167	44	156
13	119	46	405
14	252	51	150
16	217	51	374
16	216	66	151

This PROC UNIVARIATE step analyzes **MPG\_Highway** and provides several summary statistics, including the five lowest and highest extreme values and their observation numbers.

p103d01 SAS

## Freq Procedure

## FREQ Procedure

```
PROC FREQ DATA=input-table;
  TABLES col-name(s);
RUN;
```

use to specify the frequency tables to include in the results

By default, PROC FREQ creates a frequency table for each column in the input table.



The FREQ procedure creates a frequency table for each column in the input table by default, or you can use the TABLES statement to limit the columns that SAS analyzes.

p103d01 SAS

## FREQ Procedure

```
proc freq data=sashelp.cars;
  tables Origin Type DriveTrain;
run;
```

The FREQ Procedure

Origin	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Asia	158	36.92	158	36.92
Europe	123	28.74	281	65.65
USA	147	34.35	428	100.00

Type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Hybrid	3	0.70	3	0.70
SUV	60	14.02	63	14.72
Sedan	262	61.21	325	75.93
Sports	49	11.45	374	87.38
Truck	24	5.61	398	92.99
Wagon	30	7.01	428	100.00

DriveTrain	Frequency	Percent	Cumulative Frequency	Cumulative Percent
All	92	21.50	92	21.50
Front	226	52.80	318	74.30
Rear	110	25.70	428	100.00

This PROC FREQ step creates a separate table for **Origin**, **Type**, and **DriveTrain**. Each table includes a list of the distinct values for the column along with a frequency count, percent, and cumulative frequency and percent. This is a great way to validate the data in your columns. For example, you might notice unexpected values or values that appear in both uppercase and lowercase.

p103d01 SAS

## Demo: Exploring Data with SAS Procedures

File:

**[3\\_1 - Demo - Exploring Data with SAS Procedures](#)** 

<https://clemson.box.com/s/mrubmmu1c45mgsbelozpggf39eee61cw> (Box folder link)

**[3\\_1 - Demo - Exploring Data with SAS Procedures-1.pdf](#)**

<https://clemson.instructure.com/courses/237270/files/23074687?wrap=1> 

[https://clemson.instructure.com/courses/237270/files/23074687/download?download\\_frd=1](https://clemson.instructure.com/courses/237270/files/23074687/download?download_frd=1)   
(Canvas link)