Statistics in SAS

In this article, we will show you 7 different ways to analyze your data using the FREQ procedure.

You will learn how to see frequencies of different variables, find the most/least commonly occurring values in your data, check for missing values,...

Let's get started!

Software

Data Sets

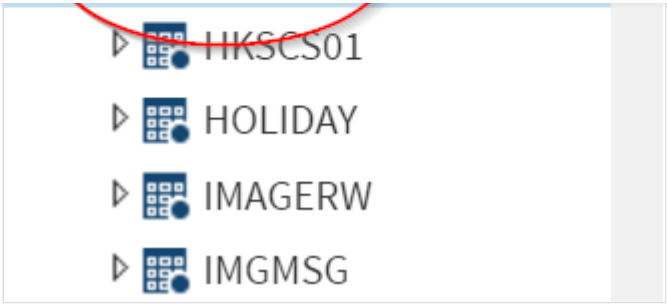
The examples used in this article are based on the CLASS and CLASSFIT data sets from the SASHelp library.

13	Audi	A6 3.0 4dr	Sedan	Europe	Front	$36,640	$33,129	3	6	220	20	27	3561	109	192
14	Audi	A6 3.0 Qu	Sedan	Europe	All	$39,640	$35,992	3	6	220	18	25	3880	109	192
15	Audi	A4 3.0 cor	Sedan	Europe	Front	$42,490	$38,325	3	6	220	20	27	3814	105	180
16	Audi	A4 3.0 Qu	Sedan	Europe	All	$44,240	$40,075	3	6	220	18	25	4013	105	180
17	Audi	A6 2.7 Tur	Sedan	Europe	All	$42,840	$38,840	2.7	6	250	18	25	3836	109	192
18	Audi	A6 4.2 Qu	Sedan	Europe	All	$49,690	$44,936	4.2	8	300	17	24	4024	109	193

Total rows: 5209 Total columns: 17 Data Set: SASHelp.Heart Rows 1-100

	Status	Death...	AgeCH...	Sex	AgeAtS...	Height	Weight	Diastolic	Systolic	MRW	Smoking	AgeAt...	Choles...	Chol_S
1	Dead	Other	.	Female	29	62.5	140	78	124	121	0	55	.	
2	Dead	Cancer	.	Female	41	59.75	194	92	144	183	0	57	181	Desi
3	Alive		.	Female	57	62.25	132	90	170	114	10	.	250	High
4	Alive		.	Female	39	65.75	158	80	128	123	0	.	242	High
5	Alive		.	Male	42	66	156	76	110	116	20	.	281	High
6	Alive		.	Female	58	61.75	131	92	176	117	0	.	196	Desi
7	Alive		.	Female	36	64.75	136	80	112	110	15	.	196	Desi
8	Dead	Other	.	Male	53	65.5	130	80	114	99	0	77	276	High
9	Alive		.	Male	35	71	194	68	132	124	0	.	211	Bord
10	Dead	Cerebral V	.	Male	52	62.5	129	78	124	106	5	82	284	High
11	Alive		.	Male	39	66.25	179	76	128	133	30	.	225	Bord
12	Alive		57	Male	33	64.25	151	68	108	118	0	.	221	Bord
13	Alive		55	Male	33	70	174	90	142	114	0	.	188	Desi
14	Alive		79	Male	57	67.25	165	76	128	118	15	.	.	
15	Alive		66	Male	44	69	155	90	130	105	30	.	292	High
16	Alive		.	Female	37	64.5	134	76	120	108	10	.	196	Desi
17	Alive		.	Male	40	66.25	151	72	132	112	30	.	192	Desi

You can find the CARS and HEART data sets from the sashelp library:



[Don't have the software yet? Download **SAS Studio** here for free.]

1. Basic Usage

The most basic usage of Proc Freq is to determine the frequency (number of occurrences) for all values found within each variable of your dataset.

```
Run;
```

The code above creates a frequency table for each of the variable in the data set.

For example, below is a frequency table for the variable MAKE.

Cadillac	8	1.67	63	14.72
Chevrolet	27	6.31	90	21.03
Chrysler	15	3.50	105	24.53
Dodge	13	3.04	118	27.57
Ford	23	5.37	141	32.94
GMC	8	1.87	149	34.81
Honda	17	3.97	166	38.79
Hummer	1	0.23	167	39.02
Hyundai	12	2.80	179	41.82
Infiniti	8	1.87	187	43.69
Isuzu	2	0.47	189	44.16
Jaguar	12	2.80	201	46.96
Jeep	3	0.70	204	47.66

If you scroll down, you will also see the frequency tables for the variable ORIGIN and DRIVETRAIN:

Cadillac	8	1.67	63	14.72
Chevrolet	27	6.31	90	21.03
Chrysler	15	3.50	105	24.53
Dodge	13	3.04	118	27.57
Ford	23	5.37	141	32.94
GMC	8	1.87	149	34.81
Honda	17	3.97	166	38.79
Hummer	1	0.23	167	39.02
Hyundai	12	2.80	179	41.82
Infiniti	8	1.87	187	43.69
Isuzu	2	0.47	189	44.16
Jaguar	12	2.80	201	46.96
Jeep	3	0.70	204	47.66

By default, the TABLES statement used with Proc Freq will output a table which lists the values found within the variable(s) specified, the frequency of each value, the percentage of that value relative to all other value as well as the cumulative frequencies and cumulative percentages.

Audi	19	4.44	26	6.07
BMW	20	4.67	46	10.75
Buick	9	2.10	55	12.85
Cadillac	8	1.87	63	14.72
Chevrolet	27	6.31	90	21.03
Chrysler	15	3.50	105	24.53
Dodge	13	3.04	118	27.57
Ford	23	5.37	141	32.94
GMC	8	1.87	149	34.81
Honda	17	3.97	166	38.79
Hummer	1	0.23	167	39.02
Hyundai	12	2.80	179	41.82
Infiniti	8	1.87	187	43.69
Isuzu	2	0.47	189	44.16
Jaguar	12	2.80	201	46.96
Jeep	3	0.70	204	47.66

However, using Proc Freq in this manner without any options is usually

Model	Frequency	Percent	Cumulative Frequency	Cumulative Percent
3.5 RL 4dr	1	0.23	1	0.23
3.5 RL w/Navigation 4dr	1	0.23	2	0.47
300M 4dr	1	0.23	3	0.70
300M Special Edition 4dr	1	0.23	4	0.93
325Ci 2dr	1	0.23	5	1.17
325Ci convertible 2dr	1	0.23	6	1.40
325i 4dr	1	0.23	7	1.64
325xi 4dr	1	0.23	8	1.87
325xi Sport	1	0.23	9	2.10
330Ci 2dr	1	0.23	10	2.34
330Ci convertible 2dr	1	0.23	11	2.57
330i 4dr	1	0.23	12	2.80
330xi 4dr	1	0.23	13	3.04
350Z Enthusiast convertible 2dr	1	0.23	14	3.27
350Z coupe 2dr	1	0.23	15	3.50
4Runner SR5 V6	1	0.23	16	3.74
525i 4dr	1	0.23	17	3.97
530i 4dr	1	0.23	18	4.21

A more efficient and effective use of Proc Freq is to use the TABLES statement to limit the variables that are reported on.

```
Proc freq data = sashelp.cars;
 Tables Origin;
Run;
```

The resulting table from this code is shown here:

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

Using the order option, you can easily see the most or least commonly occurring values of both Type and Origin variables:

```
Proc freq data = sashelp.cars order=freq;
 Tables type origin;
Run;
```

The resulting tables shows the frequency of each variable sorted with the most common variable on top and the least common on the bottom:

Origin	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Asia	158	36.92	158	36.92
USA	147	34.35	305	71.26
Europe	123	28.74	428	100.00

3. Check for Missing Values

Proc freq is an excellent tool to check for missing values in your dataset.

For this example, the SASHELP.HEART dataset is used. The SASHELP.HEART dataset can be accessed in the same way as the CARS dataset described above.

```
Run;
```

Here, you can see the missing values highlighted at the bottom of the table:

The FREQ Procedure

Cause of Death

DeathCause	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Cancer	539	27.07	539	27.07
Cerebral Vascular Disease	378	18.99	917	46.06
Coronary Heart Disease	605	30.39	1522	76.44
Other	357	17.93	1879	94.37
Unknown	[illegible]	[illegible]	1991	100.00

Frequency Missing = 3218

```
Proc freq data=sashelp.heart;
 Tables deathcause /missing;
Run;
```

With the MISSING option, you can see the frequencies and the percentage of missing values within the table:

The FREQ Procedure

Cause of Death				
DeathCause	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	3218	61.78	3218	61.78
Cancer	539	10.35	3757	72.13
Cerebral Vascular Disease	378	7.26	4135	79.38
Coronary Heart Disease	605	11.61	4740	91.00
Other	357	6.85	5097	97.85
Unknown	112	2.15	5209	100.00

```
Proc freq data=sashelp.heart;
 Tables deathcause /missprint;
Run;
```

Notice that in this table, the percentages of each value are lower than the percentages in the previous table, as the missing values are not factored into this calculation.

Using the Unknown value as an example, the percentage of records that have an Unknown value for Cause of Death is 5.63% with MISSPRINT, compared to only 2.15% in the previous table with the MISSING option:

Are you totally new to SAS?

Take our **Practical SAS Training Course for Beginners** and learn how to code your first SAS program!

Frequencies and percentages calculated using Proc Freq can also be saved to an output dataset using the OUT option combined with the TABLES statement.

The OUTCUM option can also be added to include the cumulative frequencies in the output dataset if desired:

```
Proc freq data = sashelp.cars order=freq;
 Tables type /out=cars_freq outcum;
Run;
```

5. Use the FORMAT statement to categorize and analyze data

When combined with Proc Format and a FORMAT statement, Proc Freq also becomes a powerful tool to categorize and subsequently analyze continuous variables (or variables with a large number of unique values).

Using the MSRP (Manufacturer's Suggested Retail Price) variable in the Cars dataset as an example, you can see that the standard Proc Freq output shown below does not produce very useful information for a variable such as MSRP:

MSRP	Frequency	Percent	Frequency	Percent
$10,280	1	0.23	1	0.23
$10,539	1	0.23	2	0.47
$10,760	1	0.23	3	0.70
$10,995	1	0.23	4	0.93
$11,155	1	0.23	5	1.17
$11,290	1	0.23	6	1.40
$11,560	1	0.23	7	1.64

However, buy using Proc Format you can create categories (or groups) of MSRPs to see, for example, how many cars fall within a particular price range.

```
Proc format;
 Value msrp_groups
  10000-19999 = '10,000-19,999'
  20000-29999 = '20,000-29,999'
  30000-39999 = '30,000-39,999'
  40000-high = '40,000+'
```

distribution of MSRPs across the different groups:

```
Proc freq data = sashelp.cars;
 Tables msrp;
 Format msrp msrp_groups.;
Run;
```

As you can see, this produces a much more useful and informative table:

Proc freq can also be used to produce 2x2 or higher nxn multi-way tables to determine the distribution (or frequency) of records that fall into 2 or more combinations of categories.

For example, if you would like to compare the different car DriveTrain types by the continent of Origin from the Cars dataset, you could use the following code:

```
Proc freq data=sashelp.cars;
 Tables origin*drivetrain;
Run;
```

In this example, both Origin and DriveTrain each have 3 possible values. As a result, the cross-tabulation produces a 3x3 table which includes a

	7.94 21.52 36.96	23.13 62.66 43.81	5.84 15.82 22.73	36.92
Europe	36 8.41 29.27 39.13	37 8.64 30.08 16.37	50 11.68 40.65 45.45	123 28.74
USA	22 5.14 14.97 23.91	90 21.03 61.22 39.82	35 8.18 23.81 31.82	147 34.35
Total	92 21.50	226 52.80	110 25.70	428 100.00

While this table may seem overwhelming at first, let's walk through it step-by-step to understand what each component refers to.

As shown in the legend, the first row corresponds to the frequencies. For example, the 34 in the top left box indicates that there are 34 cars from Asia that have an "All" for DriveTrain.

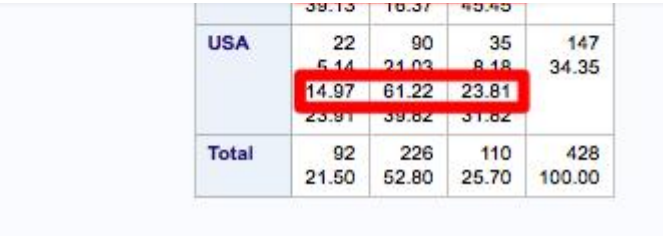
Moving from left to right, the 99 in the top middle box indicates that there are 99 cars from Asia that have a "Front" drivetrain, and so on.

	39.13	16.37	45.45	
USA	22 14.97 23.91	90 61.22 39.82	35 23.81 31.82	147 34.35
Total	92 21.50	226 52.80	110 25.70	428 100.00

The second row contains the percentages relative to the other 8 combinations. Using the top left box again as an example, the 7.94% indicates that out of the 9 possible combinations of Origin and DriveTrain, 7.94% of records have Origin=Asia *and* DriveTrain=All.

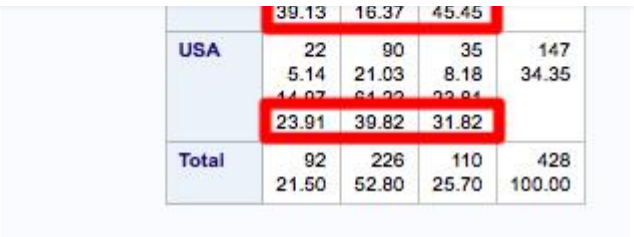
	39.13	16.37	45.45	
USA	22 5.14	90 21.03	35 8.18	147 34.35
	23.91	39.82	31.82	
Total	92 21.50	226 52.80	110 25.70	428 100.00

The third row contains what is known as the row percentages. Starting with the top left box as an example, the 21.52 indicates that of **those records with Origin=Asia,** 21.52% have a DriveTrain=All. Moving across the row from left to right, you can see that for Origin=Asian cars, 62.66% have DriveTrain=Front, and 15.82% have a DriveTrain=Rear. Notice that these 3 percentages total 100% when summed (added together) across the row.



USA	22 5.14	90 21.03	35 8.18	147 34.35
	14.97	61.22	23.81	
Total	92 21.50	226 52.80	110 25.70	428 100.00

The fourth row contains what is known as the column percentages. Starting with the top left box as an example, the 36.96 indicates that of **those records with DriveTrain=All**, 36.96% have Origin=Asia. Moving down the column from left to right, you can see that for DriveTrain=All cars, 39.13% have Origin=Europe and 23.91% have Origin=USA. Notice that these 3 percentages total 100% when summed (added together) down the column.



	39.13	16.37	45.45	
USA	22 5.14	90 21.03	35 8.18	147 34.35
	23.91	39.82	31.82	
Total	92 21.50	226 52.80	110 25.70	428 100.00

Depending on the desired results, you can choose to suppress some of these numbers from the output. The NOCOL, NOROW, NOFREQ and NOPERCENT options can be used to suppress the column percentages, row percentages, frequencies and overall percentages from your output. These options can be used independently or in different combinations together.

For example, if you wanted to suppress the row and column percentages, but keep the frequencies and overall percentages, you would use the following code:

overall percentages:

The FREQ Procedure

Frequency Percent	Table of Origin by DriveTrain			
	DriveTrain			
Origin	All	Front	Rear	Total
Asia	34 7.94	99 23.13	25 5.84	158 36.92
Europe	36 8.41	37 8.64	50 11.68	123 28.74
USA	22 5.14	90 21.03	35 8.18	147 34.35
Total	92 21.50	226 52.80	110 25.70	428 100.00

Two-way or multi-way tables can also be displayed in more of a list format for improved readability. This is especially useful when there are many possible combinations between the two variables. To display a cross tabulation in the long form “list” format, you can simply use the LIST

The results are identical to those produced without the LIST option, the only change is in how the information is displayed:

The FREQ Procedure

Origin	DriveTrain	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Asia	All	34	7.94	34	7.94
Asia	Front	99	23.13	133	31.07
Asia	Rear	25	5.84	158	36.92
Europe	All	36	8.41	194	45.33
Europe	Front	37	8.64	231	53.97
Europe	Rear	50	11.68	281	65.65
USA	All	22	5.14	303	70.79
USA	Front	90	21.03	393	91.82
USA	Rear	35	8.18	428	100.00

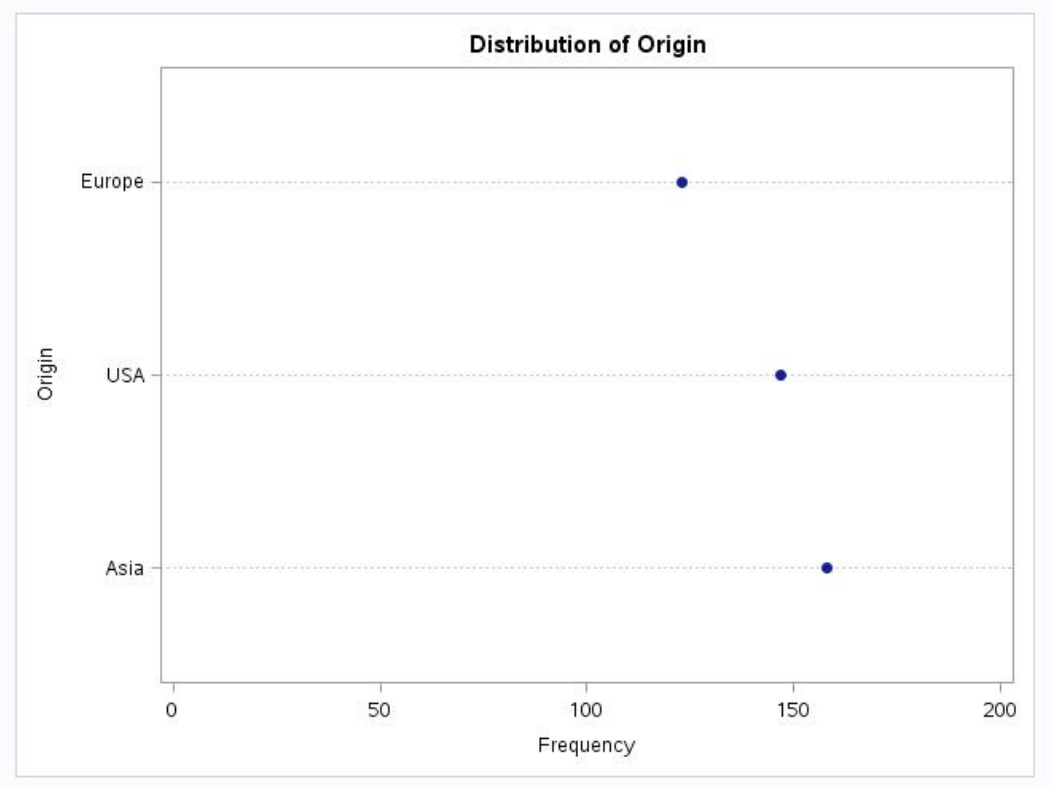
Within Proc Freq, you have the ability to create either dot or bar plots, which can be created based on either the frequencies or the overall percentages.

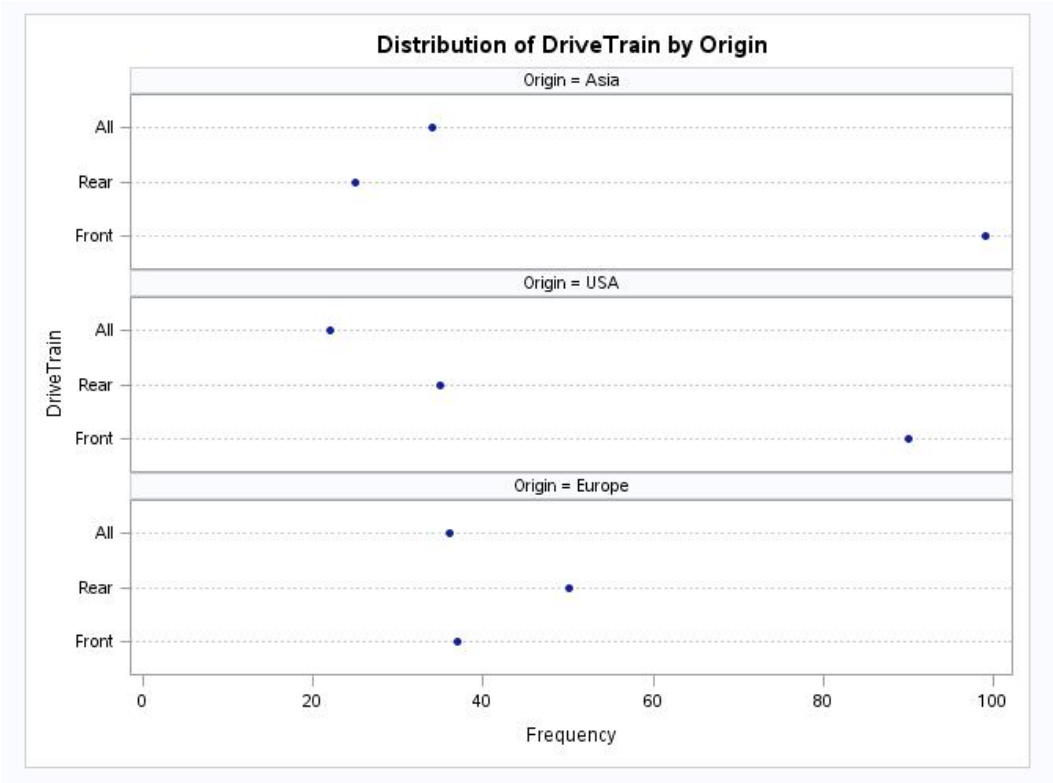
In the following example, the TABLES statement is used to create both a 1-way frequency table for the Origin variable, and a 3x3 frequency table for the DriveTrain variable crossed with Origin.

To produce a dot plot for these variables, the plots=freqplot (type=dot) option is added. In order to produce these graphs, ODS graphics must also be turned ON (and subsequently turned OFF) as shown below:

```
Ods graphics on;
Proc freq data=sashelp.cars order=freq;
 Tables origin drivetrain*origin / plots=freqplot(type=dot);
```

continent of Origin in the Cars dataset:

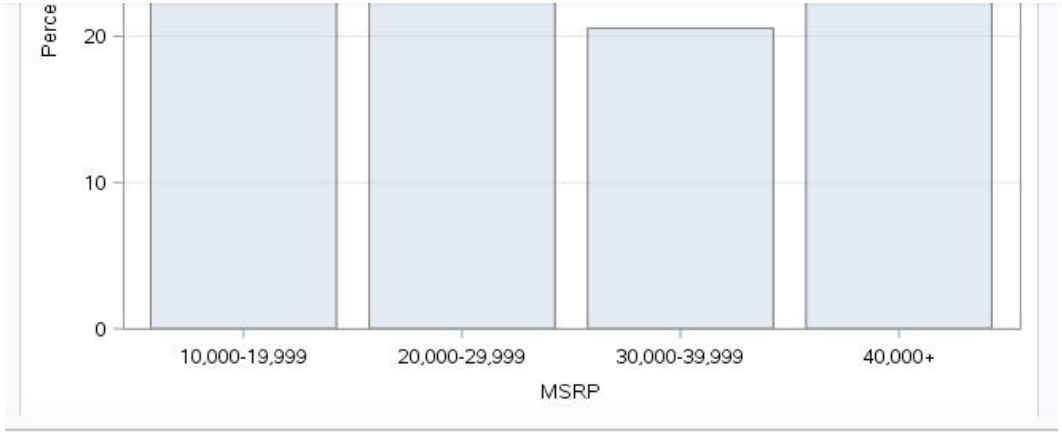




Alternatively, similar code can be used to produce bar plots based on the percentages instead of the frequencies. Of course, you can also mix and match combinations to produce a dot plot of percentages or a bar plot of

```
Proc format;
 Value msrp_groups
  10000-19999 = '10,000-19,999'
  20000-29999 = '20,000-29,999'
  30000-39999 = '30,000-39,999'
  40000-high = '40,000+'
  ;
Run;

Ods graphics on;
Proc freq data=sashelp.cars order=freq;
 Tables msrp / plots=freqplot (type=bar scale=percent);
 Format msrp msrp_groups.;
Run;
Ods graphics off;
```


Perce
20
10
0
10,000-19,999
20,000-29,999
30,000-39,999
40,000+
MSRP

Name

Be the first to comment.

Subscribe Add Disqus to your siteAdd DisqusAdd

Do Not Sell My Data