Demo: Exploring Data with Procedures

Scenario

Use the PRINT, MEANS, UNIVARIATE, and FREQ procedures to explore and validate data.

Files
- **p103d01.sas**
- **storm_summary** – a SAS table that contains one row per storm for the 1980 through 2016 storm seasons

Syntax

```
PROC PRINT DATA=input-table(OBS=n);
     VAR col-name(s);
RUN;

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

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

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

Notes
- PROC PRINT lists all columns and rows in the input table by default. The OBS= data set option limits the number of rows read from the input data. The VAR statement limits and orders the columns that are listed.
- PROC MEANS generates simple summary statistics for each numeric column in the input data by default. The VAR statement limits the columns to analyze.
- PROC UNIVARIATE also generates summary statistics for each numeric column in the data by default, but it includes more detailed statistics related to distribution and extreme values. The VAR statement limits the columns to analyze.
- PROC FREQ creates a frequency table for each column in the input table by default. You can limit the columns that are analyzed by using the TABLES statement.

Demo

1. Open **p103d01.sas** from the **demos** folder and find the **Demo** section of the program. Complete the PROC PRINT statement to list the data in **pg1.storm_summary**. Print the first 10 observations. Highlight the step and run the selected code.

```
proc print data=pg1.storm_summary (obs=10);
run;
```

Obs	Season	Name	Basin	Type	MaxWindMPH	MinPressure	StartDate	EndDate	Hem_NS	Hem_EW	Lat	Lon
1	1980		na	TS	35	.	17JUL1980	18NOV1980	N	W	25.7	-91.2
2	1980		SP	NR	.	998	27MAR1980	30MAR1980	S	E	19.1	137.0
3	1980	AGATHA	EP	TS	115	.	09JUN1980	15JUN1980	N	W	12.8	-118.7
4	1980	ALBINE	SI	ET	.	.	27NOV1979	06DEC1979	S	E	19.1	137.0
5	1980	ALEX	WP	TS	40	998	09OCT1980	14OCT1980	N	E	27.2	140.5
6	1980	ALLEN	NA	TS	190	899	31JUL1980	11AUG1980	N	W	21.8	-86.4
7	1980	AMY	SI	NR	132	915	04JAN1980	12JAN1980	S	E	-19.4	119.6
8	1980	BERENICE	SI	TS	.	.	15DEC1979	21DEC1979	S	E	-19.4	119.6
9	1980	BETTY	WP	ET	115	925	28OCT1980	08NOV1980	N	E	14.3	127.5
10	1980	BLAS	EP	TS	58	.	16JUN1980	19JUN1980	N	W	12.3	-110.5

2. Add a VAR statement to include only the following columns: **Season**, **Name**, **Basin**, **MaxWindMPH**, **MinPressure**, **StartDate**, and **EndDate**. Add **list first 10 rows** as a comment before the PROC PRINT statement. Highlight the step and run the selected code.

   SAS Studio Note: To easily add column names, place your cursor after the keyword VAR. Use the Library section of the navigation pane to find the **pg1** library. Expand the **storm_summary** table to see a list of column names. Hold down the Ctrl key and select the columns in the order in which you want them to appear in the statement. Drag the selected columns to the VAR statement.

```
/*list first 10 rows*/
proc print data=pg1.storm_summary(obs=10);
     var Season Name Basin MaxWindMPH MinPressure StartDate
         EndDate;
run;
```

Obs	Season	Name	Basin	MaxWindMPH	MinPressure	StartDate	EndDate
1	1980		na	35	.	17JUL1980	18NOV1980
2	1980		SP	.	998	27MAR1980	30MAR1980
3	1980	AGATHA	EP	115	.	09JUN1980	15JUN1980
4	1980	ALBINE	SI	.	.	27NOV1979	06DEC1979
5	1980	ALEX	WP	40	998	09OCT1980	14OCT1980
6	1980	ALLEN	NA	190	899	31JUL1980	11AUG1980
7	1980	AMY	SI	132	915	04JAN1980	12JAN1980
8	1980	BERENICE	SI	.	.	15DEC1979	21DEC1979
9	1980	BETTY	WP	115	925	28OCT1980	08NOV1980
10	1980	BLAS	EP	58	.	16JUN1980	19JUN1980

3. Copy the PROC PRINT step and paste it at the end of the program. Change **PRINT** to **MEANS**. Remove the OBS= data set option to analyze all observations. Modify the VAR statement to calculate summary statistics for **MaxWindMPH** and **MinPressure**. Add **calculate summary statistics** as a comment before the PROC MEANS statement. Highlight the step and run the selected code.

```
/*calculate summary statistics*/
proc means data=pg1.storm_summary;
     var MaxWindMPH MinPressure;
run;
```

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
MaxWindMPH	3095	79.3179321	31.6853937	6.0000000	213.0000000
MinPressure	2922	961.8545517	288.6582966	-9999.00	1012.00

4. Copy the PROC MEANS step and paste it at the end of the program. Change **MEANS** to **UNIVARIATE**. Add **examine extreme values** as a comment before the PROC UNIVARIATE statement. Highlight the step and run the selected code.

```
/*examine extreme values*/
proc univariate data=pg1.storm_summary;
     var MaxWindMPH MinPressure;
run;
```

The UNIVARIATE Procedure Variable: MaxWindMPH			

Moments

N	3095	Sum Weights	3095
Mean	79.3179321	Sum Observations	245489
Std Deviation	31.6853937	Variance	1003.96417
Skewness	0.5963944	Kurtosis	-0.3710172
Uncorrected SS	22577945	Corrected SS	3106265.15
Coeff Variation	39.947327	Std Error Mean	0.56954597

Extreme Observations

Lowest		Highest	
Value	Obs	Value	Obs
6	2659	184	702
17	1960	184	1477
23	2757	184	2164
23	1366	190	6
23	1103	213	3017

Missing Values

Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	23	0.74	100.00

5. Copy the PROC UNIVARIATE step and paste it at the end of the program. Change **UNIVARIATE** to **FREQ**. Change the VAR statement to a TABLES statement to produce frequency tables for **Basin**, **Type**, and **Season**. Add **list unique values and frequencies** as a comment before the PROC FREQ statement. Highlight the step and run the selected code.

```
/*list unique values and frequencies*/
proc freq data=pg1.storm_summary;
     tables Basin Type Season;
run;
```

The FREQ Procedure

Basin	Frequency	Percent	Cumulative Frequency	Cumulative Percent
EP	671	21.52	671	21.52
NA	472	15.14	1143	36.66
NI	84	2.69	1227	39.35
SI	588	18.86	1815	58.21
SP	359	11.51	2174	69.72
WP	928	29.76	3102	99.49
na	16	0.51	3118	100.00

Type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
DS	293	9.40	293	9.40
ET	761	24.41	1054	33.80

End of Demonstration