



## Module 6 Cleaning Data

### **PROC PRINT**

Prints data to the output window in SAS.

```
PROC PRINT DATA= sasdataset;  
RUN;
```

To select variables, add a VAR statement to the PROC PRINT

```
PROC PRINT DATA= sasdataset;  
VAR variablename1, variablename2;  
RUN;
```

### **PROC SORT**

Sorts the data in ascending order.

```
PROC SORT DATA= sasdataset;  
BY variablename(s);  
RUN;
```

More than one variable can be included in the BY statement. Variables will be sorted in the order listed (e.g., variable1 variable2, etc.).

### **OUT Statements:**

Adding an OUT option to the PROC SORT statement will create a new data set.

```
PROC SORT DATA= originalsasdataset OUT=newsasdataset;  
BY variablename(s);  
RUN;
```

### **Missing Data**

Missing SAS character values are represented with blanks.

Missing numeric values are displayed as a dot/period (.).

Use the MISSING() function to find missing data. The function returns two values:

```
1= missing  
0=non-missing values
```

To find missing data:

```
PROC PRINT DATA= sasdataset;  
WHERE MISSING (variablename) = 1;  
RUN;
```

To find non-missing data:

```
PROC PRINT data= sasdataset;  
WHERE MISSING (variablename) =0;  
RUN;
```

### **Duplicate Data**

```
PROC SORT data= sasdataset;  
NODUPKEY OUT=dataset_noduplicates DUPOUT=dataset_duplicaterecords;  
BY variable(s);  
RUN;
```

Review the output and determine if the duplicate data is in error or not. Make changes in accordance to organizations practices.

### **PROF FREQ**

Calculates the frequencies of variables.

```
PROC FREQ DATA= sasdataset;  
TABLES variablename(s);  
RUN;
```

To add an option statement to the tables section of the program.

```
TABLES variablename(2) / OPTION;
```

Options include:

- Missing – to include missing values in the totals
- Nocum- to not include cumulative totals
- Nopercent- to not include percent calculations