Homework Editing Data

1. The data set shoes\_tracker is in the STA5066 sub directory.
2. Use a PROC PRINT step to list the observations that do not meet both of the following requirements

Product\_Category must not be missing.

Supplier\_Country must have a value of GB or US.

1. Use a PROC FREQ step with a TABLES statement to create frequency tables for Supplier\_Name and Supplier\_ID. Include the NLEVELS option. The data requirements for these variables are:

Supplier\_Name must be 3Top Sports or Greenline Sports Ltd.

Supplier\_ID must be 2963 or 14682.

1. Use a PROC PRINT step with a WHERE statement to display observations for which Product\_Name is not in proper case (use the propcase function)
2. The following table identifies data errors that need to be corrected. The table also notes the correct values . Use a data step to create a file work.shoes\_tracker in which the errors have been corrected.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Obs | Invalid Value | Correct Value | Reference Variable |
| Supplier\_Country | 10 | mixed case | upper case |  |
|  | 5 | UT | US | Supplier\_Country='UT' |
| Product\_Category | 2 |  | Shoes | Product\_Category=' ' |
| Supplier\_ID | 1 | . | 2963 | Supplier\_ID=. |
| Supplier\_Name | 3, 7 | 3op Sports | 3Top Sports | Supplier\_Name = '3op Sports' |
| Product\_ID | 4 | 22020030007 | 220200300079 | \_N\_=4 |
| Product\_ID | 8 | 2202003001290 | 220200300129 |  |
| Product\_Name | 3 | not proper case | proper case |  |
| Supplier\_Name | 9 | 3Top Sports | Greenline Sports Ltd | Supplier\_ID=14682 and Supplier\_Name = '3Top Sports' |

1. The SAS data set qtr2\_2007 in in the STA5066 sub directory.
2. Use a data step to create a new dataset work.qtr2
3. Use a print step to print observations on work.qtr2 that do not meet both of the requirements:

Delivery\_Date values must be equal to or greater than Order\_Date values.

Order\_Date values must be in the range of April 1, 2007 – June 30, 2007.

1. Use a PROC FREQ step with work.qtr2 and a TABLES statement to create frequency tables for Order\_ID and Order\_Type. Include the NLEVELS option.
2. The data on work.qtr2 should meet the following requirements:

Order\_ID must be unique (36 distinct values) and not missing. Use a proc freq step to check this requirement.

Order\_Type must have a value of 1, 2, or 3. Use a PROC PRINT step to list observations for which order\_type does not have a value of 1, 2, or 3 (use the in operator).

1. Correct the invalid data on work.qtr2. Use a data step to create a new version of work.qtr2 by setting work.qtr2. The following table gives the information to make the necessary corrections.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Obs | Invalid Value | Correct Value | Reference Variable |
| Delivery\_Date | 5 | 12APR2007 | 12MAY2007 | Order\_ID=1242012259 |
| Order\_Date | 22 | 26JUL2007 | 26JUN2007 | Order\_ID=1242449327 |
| order\_id | 18 | . | 1241895587 | \_n\_=18 |
| order\_id | 19 | . | 1241895564 | \_n\_=19 |
| order\_type | 2 | 0 | 3 | \_n\_=2 |
| order\_type | 10 | 4 | 3 | \_n\_=10 |
|  |  |  |  |  |

1. Use your previously written programs (above) to verify that the data on work.qtr meet all of the data set requirements.
2. The SAS data set price\_current is in the STA5066 sub directory.
3. Use a PROC CONTENTS step to examine the variables on the price\_current data set, in particular, their type (numeric or character).
4. Use a PROC PRINT step to print observations with missing values for any of the variables Unit\_Cost\_Price, Unit\_Sales\_Price, or Factor.
5. Use a PROC MEANS step to determine whether there are observations with the values of the variables Unit\_Cost\_Price, Unit\_Sales\_Price, and Factor that do not meet the following requirements. Use a VAR statement to restrict the analyses to these variables.

Unit\_Cost\_Price must be in the numeric range of 1-400.

Unit\_Sales\_Price must be in the numeric range of 3-800.

Factor must be in the numeric range of 1-1.05.

1. Use a PROC UNIVARIATE step with a VAR statement to obtain the extreme observations of Unit\_Sales\_Price and Factor. Use an ODS SELECT statement to only display the extreme values.
2. Use a DATA step to create a data set work.price\_current that corrects the errors shown in the following table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Obs | Invalid Value | Correct Value | Reference Variable |
| Unit\_Sales\_Price | 41 | 5730 | 57.30 | Product\_ID=220200200022 |
| Unit\_Sales\_Price | 103 | . | 41.20 | Product\_ID=240200100056 |
| Factor | 14 | 100.0 | 1.0 | \_N\_=14 |
| Factor | 170 | 10.20 | 1.02 | \_N\_=170 |
| Factor | 134 | .01 | 1.0 | \_N\_=134 |

1. The SAS data set labsub1 is in the STA5066 sub directory. This file contains the following information.

|  |  |
| --- | --- |
| Variable Name | Description |
| seqn | sequence number |
| hgp | hemoglobin (g/dl) |
| htp | hematocrit (%) |
| tcp | cholesterol (mg/dl) |
| tgp | triglycerides (mg/dl) |
| lcp | low density lipoprotein (mg/dl) |
| hdp | high density lipoprotein (mg/dl) |
| fbpsi | fibrinogen (mg/dl) |
| crp | C reactive protein (mg/dl) |
| sgp | plasma glucose (mg/dl) |
| urp | urinary creatinine (mg/dl) |

1. Use a PROC CONTENTS step with the position option to find the order of the variables on the data set.
2. Use a PROC MEANS step to find the mean of all the variables except seqn on the data set. In the VAR statement use a varlist to denote all of the variables other than seqn.
3. Examine maximum values for the variables in the results of the PROC MEANS step to identify “fill values.” A fill value will be a whole number consisting entirely of 8’s (e.g. 8, 888, etc.) and are used to designate missing values.
4. Use a DATA step to create a new dataset work.labsub2 that replaces all of the fill values for all of the variables with ., the SAS missing value.
5. Use a PROC MEANS step with work.labsub2 to verify that the fill values have been replaced.