# Program Summary - Data_prep_Character_variables.sas

## Execution Environment

| | |
|---|---|
| Author: | u63876948 |
| File: | /home/u63876948/Portfolio/Character variable/Data_prep_Character_variables.sas |
| SAS Platform: | Linux LIN X64 5.14.0-284.30.1.el9_2.x86_64 |
| SAS Host: | ODAWS02-USW2-2.ODA.SAS.COM |
| SAS Version: | 9.04.01M7P08062020 |
| SAS Locale: | en_US |
| Submission Time: | 11/10/2024, 7:06:43 PM |
| Browser Host: | 135.0.146.25 |
| User Agent: | Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/130.0.0.0 Safari/537.36 |
| Application Server: | ODAMID00-USW2-2.ODA.SAS.COM |

## Code: Data_prep_Character_variables.sas

```
libname mylib '/home/u63876948/Portfolio/Character variable';

/*Examine the target variable y:*/
*Use PROC FREQ to list a simple frequency table for the variable y.;

proc freq data = mylib.customer_all;
table y;
run;

/*Examine the variable "contact" and study its dependency with the target variable y.*/
*Use PROC FREQ to list a simple frequency table for the variable "contact".
Examine the output for invalid values.;
proc freq data = mylib.customer_all ;
table contact;
run;

/*Contiengency table Contact by y and mosaic plot:*/
*create a 2x2 contingency table along with a mosaic plot. Show the statistics for Table of contact by y.;

proc freq data=mylib.customer_all;
table contact*y / out = mylib.frequency norow chisq plots=mosaicplot;
title 'Cross tab contact and Y';
run;

/*Interpret:
Based on the mosaic plot, there is an association between the two variable.
Based on the Contingency coefficient, there is a week to moderate association between the two variable. */


*define a new format, name it education_Check and use it to identify invalid values for the variable education.
Valid values are 'primary', 'secondary', 'tertiary', 'unknown'.;

proc format;
value $education_Check  'primary', 'secondary', 'tertiary', 'unknown' = 'valid'
' ' = 'missing'
other = 'invalid';
run;

Title 'Check invalid value of education';
proc freq data=mylib.customer_all;
table education / nocum;
format education $education_Check.;
run;
title;

/* Use the function lowcase on education column. use the same dataset name for output dataset.*/

data mylib.customer_all;
    set mylib.customer_all;
    education = lowcase(education);
run;

Title 'Check invalid value of updated education';
proc freq data=mylib.customer_all;
table education;
format education $education_Check.;
run;
title;

/* show the simple frequency table after the change. */

Title 'Frequency of each education level';
proc freq data=mylib.customer_all;
table education;
run;
```

```
title;

/*Examine the variable "marital".*/

*Use PROC print with a where statement to check for data errors in the variable marital.
Consider the valid values as "single", "married", "divorced".;

title 'Check invalid value of marital';
proc print data=mylib.customer_all;
where marital not in ('single', 'married', 'divorced');
run;
title;

/*Use the function lowcase on the variable marital.*/
data mylib.customer_all;
    set mylib.customer_all;
    marital = lowcase(marital);
run;

/*show the simple frequency table after the change. */
proc freq data=mylib.customer_all;
table marital;
title 'Check invalid of updated marital';
run;

/* Examine the variable "Job".*/
*Use PROC FREQ to list a simple frequency table.;
proc freq data=mylib.customer_all;
table job;
title 'Check frequecy of each job';
run;

* Write a code to combine the categories "admin." and "ADMINISTRATION" for the job variable as "admin".;
data mylib.customer_all;
    set mylib.customer_all;
    if job in ('admin.', 'ADMINISTRATION') then job = 'admin';
run;

* Show the simple frequency table after the change.;
title 'Check frequecy of each job after grouping admin job';
proc freq data=mylib.customer_all;
table job;
run;
title;

/*Checking missing values*/

title "Checking Missing Character Values";
proc format;
value $Count_Missing ' ' = 'Missing'
        other = 'Nonmissing';
run;

proc freq data=mylib.customer_all;
tables _character_ / nocum missing;
format _character_ $Count_Missing.;
title 'Check missing character variable';
run;

/* Create a new variable named jobMF to indicate the most frequent job category */

*check the most frequent job category based on the output of proc freq.;
proc freq data=mylib.customer_all order=freq;
table job;
title 'Listing of frequency of each job';
run;

*create the new variable jobMF;
data mylib.customer_allMF;
    set mylib.customer_all;
    if job = 'management' then jobMF = 1;
    else jobMF = 0;
run;

*print the first few observations.;
proc print data=mylib.customer_allMF (obs=5);
title 'New dataset with new column_jobMF';
run;

/*Removing units from a value and standarizing*/
*step1 use the approriate function to keep only digits. name the new variable "digits";
*step2 use the function findc on length to search for the character 'm' (stands for meter),
if m is found, keep the value as it is,
*step3 if not, make a foot to meter conversion.;

data assign2.units;
    input Length $ 10. ;
```

```
    Digits = compress(Length,,'kd'); /*step1*/
    if findc(Length,'m','i') then /* Step 2 */
    Length_m = input(Digits,5.);
    else if not missing(Digits) then
    Length_m = input(Digits,5.)/3.281; /* Step 3 */
datalines;
100m.
110 ft.
50M.
70 Ft
180
;
run;

title "Reading Length Values with Unit Conversion";
proc print data=mylib.units;
run;
```

---

**Log: Data_prep_Character_variables.sas**

Notes (51)

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         libname mylib '/home/u63876948/Portfolio/Character variable';
NOTE: Libref MYLIB was successfully assigned as follows:
      Engine:        V9
      Physical Name: /home/u63876948/Portfolio/Character variable
70
71         /*Examine the target variable y:*/
72         *Use PROC FREQ to list a simple frequency table for the variable y.;
73
74         proc freq data = mylib.customer_all;
75         table y;
76         run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.03 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              2779.46k
      OS Memory           28332.00k
      Timestamp           11/11/2024 12:06:42 AM
      Step Count                        125  Switch Count  2
      Page Faults                       0
      Page Reclaims                     318
      Page Swaps                        0
      Voluntary Context Switches        28
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           272


77
78         /*Examine the variable "contact" and study its dependency with the target variable y.*/
79         *Use PROC FREQ to list a simple frequency table for the variable "contact".
80         Examine the output for invalid values.;
81         proc freq data = mylib.customer_all ;
82         table contact;
83         run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.01 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              2059.37k
      OS Memory           28332.00k
      Timestamp           11/11/2024 12:06:42 AM
      Step Count                        126  Switch Count  2
      Page Faults                       0
      Page Reclaims                     318
      Page Swaps                        0
      Voluntary Context Switches        19
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           264


84
85         /*Contiengency table Contact by y and mosaic plot:*/
```

```
86         *create a 2x2 contingency table along with a mosaic plot. Show the statistics for Table of contact by y.;
87
88         proc freq data=mylib.customer_all;
89         table contact*y / out = mylib.frequency norow chisq plots=mosaicplot;
90         title 'Cross tab contact and Y';
91         run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.
NOTE: The data set MYLIB.FREQUENCY has 6 observations and 4 variables.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.17 seconds
      user cpu time       0.06 seconds
      system cpu time     0.01 seconds
      memory              10720.93k
      OS Memory           36156.00k
      Timestamp           11/11/2024 12:06:42 AM
      Step Count                        127  Switch Count  6
      Page Faults                       0
      Page Reclaims                     2346
      Page Swaps                        0
      Voluntary Context Switches        260
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           1320


92
93         /*Interpret:
94         Based on the mosaic plot, there is an association between the two variable.
95         Based on the Contingency coefficient, there is a week to moderate association between the two variable. */
96
97
98         *define a new format, name it education_Check and use it to identify invalid values for the variable education.
99         Valid values are 'primary', 'secondary', 'tertiary', 'unknown'.;
100
101        proc format;
102        value $education_Check  'primary', 'secondary', 'tertiary', 'unknown' = 'valid'
103        ' ' = 'missing'
104        other = 'invalid';
NOTE: Format $EDUCATION_CHECK is already on the library WORK.FORMATS.
NOTE: Format $EDUCATION_CHECK has been output.
105        run;

NOTE: PROCEDURE FORMAT used (Total process time):
      real time           0.00 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              246.15k
      OS Memory           33444.00k
      Timestamp           11/11/2024 12:06:42 AM
      Step Count                        128  Switch Count  0
      Page Faults                       0
      Page Reclaims                     15
      Page Swaps                        0
      Voluntary Context Switches        0
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           32


106
107        Title 'Check invalid value of education';
108        proc freq data=mylib.customer_all;
109        table education / nocum;
110        format education $education_Check.;
111        run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.01 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              2028.03k
      OS Memory           34732.00k
      Timestamp           11/11/2024 12:06:42 AM
      Step Count                        129  Switch Count  2
      Page Faults                       0
      Page Reclaims                     325
      Page Swaps                        0
      Voluntary Context Switches        23
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


112        title;
113
114        /* Use the function lowcase on education column. use the same dataset name for output dataset.*/
115
116        data mylib.customer_all;
117            set mylib.customer_all;
118            education = lowcase(education);
119        run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.
NOTE: The data set MYLIB.CUSTOMER_ALL has 10578 observations and 17 variables.
NOTE: DATA statement used (Total process time):
      real time           0.02 seconds
```

```
      user cpu time       0.00 seconds
      system cpu time     0.01 seconds
      memory              3425.37k
      OS Memory           36268.00k
      Timestamp           11/11/2024 12:06:42 AM
      Step Count                        130  Switch Count  1
      Page Faults                       0
      Page Reclaims                     503
      Page Swaps                        0
      Voluntary Context Switches        46
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           2568


120
121       Title 'Check invalid value of updated education';
122       proc freq data=mylib.customer_all;
123       table education;
124       format education $education_Check.;
125       run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.02 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              2194.93k
      OS Memory           34732.00k
      Timestamp           11/11/2024 12:06:42 AM
      Step Count                        131  Switch Count  2
      Page Faults                       0
      Page Reclaims                     311
      Page Swaps                        0
      Voluntary Context Switches        40
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           264


126       title;
127
128       /* show the simple frequency table after the change. */
129
130       Title 'Frequency of each education level';
131       proc freq data=mylib.customer_all;
132       table education;
133       run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.01 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              2024.06k
      OS Memory           34732.00k
      Timestamp           11/11/2024 12:06:42 AM
      Step Count                        132  Switch Count  2
      Page Faults                       0
      Page Reclaims                     311
      Page Swaps                        0
      Voluntary Context Switches        25
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           264


134       title;
135
136       /*Examine the variable "marital".*/
137
138       *Use PROC print with a where statement to check for data errors in the variable marital.
139       Consider the valid values as "single", "married", "divorced".;
140
141       title 'Check invalid value of marital';
142       proc print data=mylib.customer_all;
143       where marital not in ('single', 'married', 'divorced');
144       run;

NOTE: No observations were selected from data set MYLIB.CUSTOMER_ALL.
NOTE: There were 0 observations read from the data set MYLIB.CUSTOMER_ALL.
      WHERE marital not in ('divorced', 'married', 'single');
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.00 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              2139.31k
      OS Memory           34732.00k
      Timestamp           11/11/2024 12:06:42 AM
      Step Count                        133  Switch Count  0
      Page Faults                       0
      Page Reclaims                     276
      Page Swaps                        0
      Voluntary Context Switches        6
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           16
```

```
145        title;
146
147        /*Use the function lowcase on the variable marital.*/
148        data mylib.customer_all;
149            set mylib.customer_all;
150            marital = lowcase(marital);
151        run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.
NOTE: The data set MYLIB.CUSTOMER_ALL has 10578 observations and 17 variables.
NOTE: DATA statement used (Total process time):
      real time           0.02 seconds
      user cpu time       0.00 seconds
      system cpu time     0.01 seconds
      memory              3453.34k
      OS Memory           36268.00k
      Timestamp           11/11/2024 12:06:42 AM
      Step Count                        134  Switch Count  1
      Page Faults                       0
      Page Reclaims                     499
      Page Swaps                        0
      Voluntary Context Switches        34
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           2568


152
153        /*show the simple frequency table after the change. */
154        proc freq data=mylib.customer_all;
155        table marital;
156        title 'Check invalid of updated marital';
157        run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.01 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              2024.06k
      OS Memory           34732.00k
      Timestamp           11/11/2024 12:06:42 AM
      Step Count                        135  Switch Count  3
      Page Faults                       0
      Page Reclaims                     311
      Page Swaps                        0
      Voluntary Context Switches        44
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           264


158
159        /* Examine the variable "Job".*/
160        *Use PROC FREQ to list a simple frequency table.;
161        proc freq data=mylib.customer_all;
162        table job;
163        title 'Check frequecy of each job';
164        run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.02 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              2024.78k
      OS Memory           34732.00k
      Timestamp           11/11/2024 12:06:42 AM
      Step Count                        136  Switch Count  3
      Page Faults                       0
      Page Reclaims                     311
      Page Swaps                        0
      Voluntary Context Switches        28
      Involuntary Context Switches      3
      Block Input Operations            0
      Block Output Operations           280


165
166        * Write a code to combine the categories "admin." and "ADMINISTRATION" for the job variable as "admin".;
167        data mylib.customer_all;
168            set mylib.customer_all;
169            if job in ('admin.', 'ADMINISTRATION') then job = 'admin';
170        run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.
NOTE: The data set MYLIB.CUSTOMER_ALL has 10578 observations and 17 variables.
NOTE: DATA statement used (Total process time):
      real time           0.02 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              3453.71k
      OS Memory           36268.00k
      Timestamp           11/11/2024 12:06:42 AM
      Step Count                        137  Switch Count  1
      Page Faults                       0
      Page Reclaims                     499
```

```
      Page Swaps                        0
      Voluntary Context Switches        34
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           2576


171
172       * Show the simple frequency table after the change.;
173       title 'Check frequecy of each job after grouping admin job';
174       proc freq data=mylib.customer_all;
175       table job;
176       run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.03 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              2026.65k
      OS Memory           34732.00k
      Timestamp           11/11/2024 12:06:42 AM
      Step Count                        138  Switch Count  2
      Page Faults                       0
      Page Reclaims                     311
      Page Swaps                        0
      Voluntary Context Switches        36
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           264


177       title;
178
179       /*Checking missing values*/
180
181       title "Checking Missing Character Values";
182       proc format;
183       value $Count_Missing ' ' = 'Missing'
184       other = 'Nonmissing';
NOTE: Format $COUNT_MISSING is already on the library WORK.FORMATS.
NOTE: Format $COUNT_MISSING has been output.
185       run;

NOTE: PROCEDURE FORMAT used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              246.18k
      OS Memory           33444.00k
      Timestamp           11/11/2024 12:06:42 AM
      Step Count                        139  Switch Count  0
      Page Faults                       0
      Page Reclaims                     14
      Page Swaps                        0
      Voluntary Context Switches        0
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           0


186
187       proc freq data=mylib.customer_all;
188       tables _character_ / nocum missing;
189       format _character_ $Count_Missing.;
190       title 'Check missing character variable';
191       run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.04 seconds
      user cpu time       0.04 seconds
      system cpu time     0.01 seconds
      memory              2338.21k
      OS Memory           34732.00k
      Timestamp           11/11/2024 12:06:42 AM
      Step Count                        140  Switch Count  3
      Page Faults                       0
      Page Reclaims                     317
      Page Swaps                        0
      Voluntary Context Switches        29
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           280


192
193       /* Create a new variable named jobMF to indicate the most frequent job category */
194
195       *check the most frequent job category based on the output of proc freq.;
196       proc freq data=mylib.customer_all order=freq;
197       table job;
198       title 'Listing of frequency of each job';
199       run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.01 seconds
```

```
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              2052.90k
      OS Memory           34732.00k
      Timestamp           11/11/2024 12:06:42 AM
      Step Count                        141  Switch Count  3
      Page Faults                       0
      Page Reclaims                     311
      Page Swaps                        0
      Voluntary Context Switches        31
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           280


200
201        *create the new variable jobMF;
202        data mylib.customer_allMF;
203            set mylib.customer_all;
204            if job = 'management' then jobMF = 1;
205            else jobMF = 0;
206        run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.
NOTE: The data set MYLIB.CUSTOMER_ALLMF has 10578 observations and 18 variables.
NOTE: DATA statement used (Total process time):
      real time           0.02 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              3453.28k
      OS Memory           36268.00k
      Timestamp           11/11/2024 12:06:43 AM
      Step Count                        142  Switch Count  1
      Page Faults                       0
      Page Reclaims                     491
      Page Swaps                        0
      Voluntary Context Switches        47
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           2824


207
208        *print the first few observations.;
209        proc print data=mylib.customer_allMF (obs=5);
210        title 'New dataset with new column_jobMF';
211        run;

NOTE: There were 5 observations read from the data set MYLIB.CUSTOMER_ALLMF.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.02 seconds
      user cpu time       0.02 seconds
      system cpu time     0.01 seconds
      memory              2100.00k
      OS Memory           34472.00k
      Timestamp           11/11/2024 12:06:43 AM
      Step Count                        143  Switch Count  1
      Page Faults                       0
      Page Reclaims                     255
      Page Swaps                        0
      Voluntary Context Switches        20
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           0


212
213        /*Removing units from a value and standarizing*/
214        *step1 use the approriate function to keep only digits. name the new variable "digits";
215        *step2 use the function findc on length to search for the character 'm' (stands for meter),
216        if m is found, keep the value as it is,
217        *step3 if not, make a foot to meter conversion.;
218
219        data assign2.units;
220            input Length $ 10. ;
221            Digits = compress(Length,,'kd'); /*step1*/
222           if findc(Length,'m','i') then /* Step 2 */
223           Length_m = input(Digits,5.);
224         else if not missing(Digits) then
225           Length_m = input(Digits,5.)/3.281; /* Step 3 */
226        datalines;

NOTE: The data set ASSIGN2.UNITS has 5 observations and 3 variables.
NOTE: DATA statement used (Total process time):
      real time           0.02 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              686.78k
      OS Memory           33704.00k
      Timestamp           11/11/2024 12:06:43 AM
      Step Count                        144  Switch Count  2
      Page Faults                       0
      Page Reclaims                     86
      Page Swaps                        0
      Voluntary Context Switches        51
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           272
```

```
232        ;
233        run;
234
235        title "Reading Length Values with Unit Conversion";
236        proc print data=mylib.units;
237        run;

NOTE: There were 5 observations read from the data set MYLIB.UNITS.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.01 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              609.93k
      OS Memory           33704.00k
      Timestamp           11/11/2024 12:06:43 AM
      Step Count                        145  Switch Count  0
      Page Faults                       0
      Page Reclaims                     62
      Page Swaps                        0
      Voluntary Context Switches        10
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           0


238
239
240
241
242
243
244        OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
254
```

---

**Results: Data_prep_Character_variables.sas**

The FREQ Procedure

| y | | | | |
|---|---|---|---|---|
| y | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| no | 5289 | 50.00 | 5289 | 50.00 |
| yes | 5289 | 50.00 | 10578 | 100.00 |

---

The FREQ Procedure

| contact | | | | |
|---|---|---|---|---|
| contact | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| cellular | 7682 | 72.62 | 7682 | 72.62 |
| telephone | 712 | 6.73 | 8394 | 79.35 |
| unknown | 2184 | 20.65 | 10578 | 100.00 |

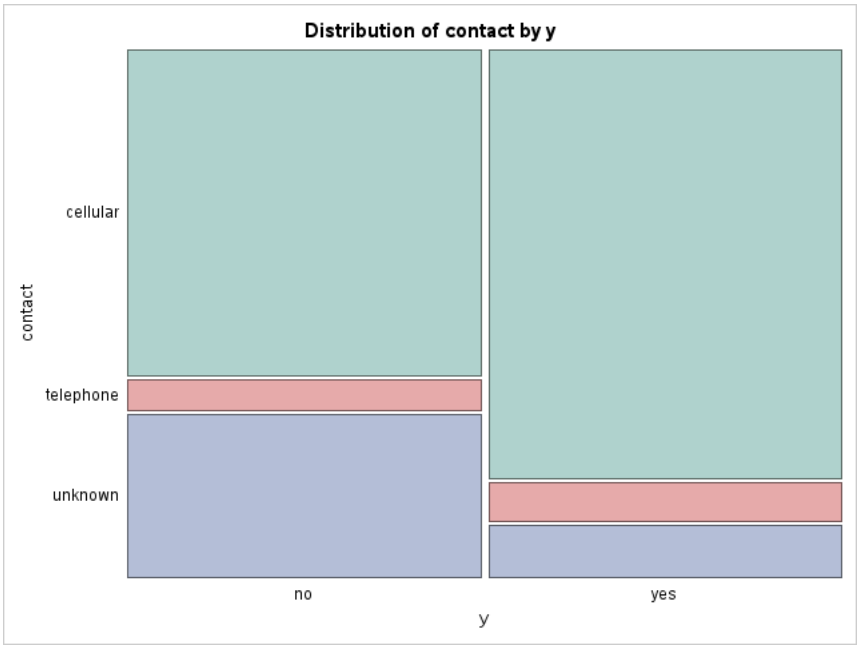
---

**Cross tab contact and Y**

The FREQ Procedure

| Frequency<br>Percent<br>Col Pct |
|---|

| Table of contact by y | | | |
|---|---|---|---|
| | y(y) | | |
| contact(contact) | no | yes | Total |
| cellular | 3313<br>31.32<br>62.64 | 4369<br>41.30<br>82.61 | 7682<br>72.62 |
| telephone | 322<br>3.04<br>6.09 | 390<br>3.69<br>7.37 | 712<br>6.73 |
| unknown | 1654<br>15.64<br>31.27 | 530<br>5.01<br>10.02 | 2184<br>20.65 |
| Total | 5289<br>50.00 | 5289<br>50.00 | 10578<br>100.00 |




**Statistics for Table of contact by y**

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 2 | 730.1254 | <.0001 |
| Likelihood Ratio Chi-Square | 2 | 759.2990 | <.0001 |
| Mantel-Haenszel Chi-Square | 1 | 678.0393 | <.0001 |
| Phi Coefficient | | 0.2627 | |
| Contingency Coefficient | | 0.2541 | |
| Cramer's V | | 0.2627 | |

**Sample Size = 10578**

## Check invalid value of education

**The FREQ Procedure**

| Education | | |
|---|---|---|
| Education | Frequency | Percent |
| valid | 10578 | 100.00 |

## Check invalid value of updated education

**The FREQ Procedure**

| Education | | | | |
|---|---|---|---|---|
| Education | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| valid | 10578 | 100.00 | 10578 | 100.00 |

## Frequency of each education level

**The FREQ Procedure**

| Education | | | | |
|---|---|---|---|---|
| Education | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| primary | 1440 | 13.61 | 1440 | 13.61 |
| secondary | 5204 | 49.20 | 6644 | 62.81 |
| tertiary | 3470 | 32.80 | 10114 | 95.61 |
| unknown | 464 | 4.39 | 10578 | 100.00 |

## Check invalid of updated marital

**The FREQ Procedure**

| marital | | | | |
|---|---|---|---|---|
| marital | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| divorced | 1243 | 11.75 | 1243 | 11.75 |
| married | 5942 | 56.17 | 7185 | 67.92 |
| single | 3393 | 32.08 | 10578 | 100.00 |

## Check frequecy of each job

**The FREQ Procedure**

| JOB | | | | |
|---|---|---|---|---|
| JOB | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| admin | 1185 | 11.20 | 1185 | 11.20 |
| blue-collar | 1914 | 18.09 | 3099 | 29.30 |
| entrepreneur | 291 | 2.75 | 3390 | 32.05 |
| housemaid | 262 | 2.48 | 3652 | 34.52 |
| management | 2391 | 22.60 | 6043 | 57.13 |
| retired | 757 | 7.16 | 6800 | 64.28 |
| self-employed | 367 | 3.47 | 7167 | 67.75 |
| services | 850 | 8.04 | 8017 | 75.79 |
| student | 375 | 3.55 | 8392 | 79.33 |
| technician | 1768 | 16.71 | 10160 | 96.05 |
| unemployed | 353 | 3.34 | 10513 | 99.39 |
| unknown | 65 | 0.61 | 10578 | 100.00 |

## Check frequecy of each job after grouping admin job

The FREQ Procedure

| JOB | | | | |
|---|---|---|---|---|
| JOB | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| admin | 1185 | 11.20 | 1185 | 11.20 |
| blue-collar | 1914 | 18.09 | 3099 | 29.30 |
| entrepreneur | 291 | 2.75 | 3390 | 32.05 |
| housemaid | 262 | 2.48 | 3652 | 34.52 |
| management | 2391 | 22.60 | 6043 | 57.13 |
| retired | 757 | 7.16 | 6800 | 64.28 |
| self-employed | 367 | 3.47 | 7167 | 67.75 |
| services | 850 | 8.04 | 8017 | 75.79 |
| student | 375 | 3.55 | 8392 | 79.33 |
| technician | 1768 | 16.71 | 10160 | 96.05 |
| unemployed | 353 | 3.34 | 10513 | 99.39 |
| unknown | 65 | 0.61 | 10578 | 100.00 |

## Check missing character variable

The FREQ Procedure

| contact | | |
|---|---|---|
| contact | Frequency | Percent |
| Nonmissing | 10578 | 100.00 |

| month | | |
|---|---|---|
| month | Frequency | Percent |
| Nonmissing | 10578 | 100.00 |

| poutcome | | |
|---|---|---|
| poutcome | Frequency | Percent |
| Nonmissing | 10578 | 100.00 |

| y | | |
|---|---|---|
| y | Frequency | Percent |
| Nonmissing | 10578 | 100.00 |

| default | Frequency | Percent |
|---|---|---|
| Nonmissing | 10578 | 100.00 |

| housing | Frequency | Percent |
|---|---|---|
| Nonmissing | 10578 | 100.00 |

| loan | Frequency | Percent |
|---|---|---|
| Nonmissing | 10578 | 100.00 |

| Education | | |
|---|---|---|
| Education | Frequency | Percent |
| Nonmissing | 10578 | 100.00 |

| marital | | |
|---|---|---|
| marital | Frequency | Percent |
| Nonmissing | 10578 | 100.00 |

| JOB | | |
|---|---|---|
| JOB | Frequency | Percent |

| JOB | | |
|---|---|---|
| JOB | Frequency | Percent |
| Nonmissing | 10578 | 100.00 |

## Listing of frequency of each job

The FREQ Procedure

| JOB | | | | |
|---|---|---|---|---|
| JOB | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| management | 2391 | 22.60 | 2391 | 22.60 |
| blue-collar | 1914 | 18.09 | 4305 | 40.70 |
| technician | 1768 | 16.71 | 6073 | 57.41 |
| admin | 1185 | 11.20 | 7258 | 68.61 |
| services | 850 | 8.04 | 8108 | 76.65 |
| retired | 757 | 7.16 | 8865 | 83.81 |
| student | 375 | 3.55 | 9240 | 87.35 |
| self-employed | 367 | 3.47 | 9607 | 90.82 |
| unemployed | 353 | 3.34 | 9960 | 94.16 |
| entrepreneur | 291 | 2.75 | 10251 | 96.91 |
| housemaid | 262 | 2.48 | 10513 | 99.39 |
| unknown | 65 | 0.61 | 10578 | 100.00 |

## New dataset with new column_jobMF

| Obs | customer_id | contact | day | month | campaign | pdays | previous | poutcome | y | default | balance | housing | loan | Education | AGE | marital | JOB | jobMF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 100103 | unknown | 5 | may | 1 | -1 | 0 | unknown | no | no | 2 | yes | yes | secondary | 33 | married | entrepreneur | 0 |
| 2 | 100106 | unknown | 5 | may | 1 | -1 | 0 | unknown | no | no | 231 | yes | no | tertiary | 35 | married | management | 1 |
| 3 | 100118 | unknown | 5 | may | 1 | -1 | 0 | unknown | no | no | 52 | yes | no | primary | 57 | married | blue-collar | 0 |
| 4 | 100119 | unknown | 5 | may | 1 | -1 | 0 | unknown | no | no | 60 | yes | no | primary | 60 | married | retired | 0 |
| 5 | 100121 | unknown | 5 | may | 1 | -1 | 0 | unknown | no | no | 723 | yes | yes | secondary | 28 | married | blue-collar | 0 |

## Reading Length Values with Unit Conversion

| Obs | Length | Digits | Length_m |
|---|---|---|---|
| 1 | 100m. | 100 | 100.000 |
| 2 | 110 ft. | 110 | 33.526 |
| 3 | 50M. | 50 | 50.000 |
| 4 | 70 Ft | 70 | 21.335 |
| 5 | 180 | 180 | 54.861 |