# Program Summary - Data_prep_numeric.sas

## Execution Environment

| | |
|---|---|
| Author: | u63876948 |
| File: | /home/u63876948/Portfolio/Numerical variable/Data_prep_numeric.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, 7:54:41 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_numeric.sas

```
*1. Examine the list of numerical attributes;

libname mylib '/home/u63876948/Portfolio/Numerical variable';

proc means data=mylib.customer_all;
run;

proc means data=mylib.customer_all nmiss;
run;

*Interpret:
pdays: number of days that passed by after the client was last contacted from a previous campaign (Numeric)
and -1 means client was not previously contacted)
Examine the range of values for day. (last contact day of the month) (numeric)
Examine the column N Miss for the variable "age". 20 missing;


*Examine the variable "age"
use PROC UNIVARIATE to examine the numeric variable "age" by showing tabular and graphical information.;

proc univariate data=mylib.customer_all;
id customer_id;
var age;
histogram / normal;
run;

* Output the customer_id whose age is missing. Use the function missing within if statements;
data mylib.missing_age;
    set mylib.customer_all;
    if missing(age) then output;
run;

proc print data=mylib.missing_age;
    var customer_id;
    title 'Customers with Missing Age';
run;


* Output the customer_id whose age is missing. Use the function missing within if statements;
data mylib.missing_age;
set mylib.customer_all;
if missing(age) then output;
run;

proc print data=mylib.missing_age;
var customer_id age;
title 'Customers with Missing Age';
run;

*Apply imputation to replace missing values for age with the mean age.;
proc stdize data=mylib.customer_all out=mylib.customer_all_Imputed
      oprefix=Orig_         /* prefix for original variables */
      reponly               /* only replace; do not standardize */
      method=MEAN;          /* or MEDIAN, MINIMUM, MIDRANGE, etc. */
   var age;              /* you can list multiple variables to impute */
run;

title 'imputed dataset';
proc print data=mylib.customer_all_Imputed (obs=10);
run;
title;


*Use proc means to check the list of numerical attributes in customer_all_Imputed;
```

```
proc means data= mylib.customer_all_Imputed nmiss;
title 'number of mission for imputed dataset';
run;

*Rename SAS dataset to its original name customer_all.;
data mylib.customer_all;
 set mylib.customer_all_Imputed;
run;

proc datasets library=mylib;
delete customer_all_Imputed;
run;

*To evaluate if age has an influence on balance?;
* use sgplot to draw a scatter plot and regression line ;
title 'Influence of age on balance';
proc sgplot data=mylib.customer_all;
reg x=age y=balance / lineattrs=(color=red thickness=2);
run;

*Binning
discretize the variable age by creating a new cat variable named age_cat:;
data mylib.customer_all;
set mylib.customer_all;
if 18 <=AGE <=35 then age_cat = 'Young_adult';
else if 36 <=AGE <= 55 then age_cat ='Middle_age';
else if AGE >= 56 then age_cat ='Old'; /* if the >56 for old., there will be 178 missing value, so chage data t
run;

*show a simple frequency table for age_cat;
proc freq data=mylib.customer_all;
table age_cat;
title 'frequency of each age';
run;

*Here is the code to create a bar chart of balance by age.;
proc sgplot data=mylib.CUSTOMER_ALL;
    vbar age_cat / response=balance group=balance groupdisplay=cluster stat=mean;
    yaxis grid;
run;
title;



/*Examine the variable campaign
campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact

Use proc univariate on campaign.*/
proc univariate data=mylib.CUSTOMER_ALL nextrobs=10;
   id customer_id;
   var campaign;
   histogram / normal;
run;


* Based on quantiles table, the variable campaign seems more categorical in nature than continuous.
Convert the variable campaign into a categorical variable name "campaign_cat" with ordinal values { 1, 2, 3, >3

proc contents data=mylib.customer_all;
run;

data mylib.customer_all;
set mylib.customer_all;
if campaign =1 then campaign_cat = '1';
else if campaign =2 then campaign_cat = '2';
else if campaign =3 then campaign_cat = '3';
else if campaign >3 then campaign_cat = '>3';
run;

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

/*Examine the variable "balance"
Investigate the distribution of balance. use proc univariate to get the statistics along a histogram for the va
title "Running PROC UNIVARIATE on balance";
proc univariate data=mylib.customer_all noprint;
   id customer_id;
   var balance;
   histogram ;
run;

/*4.2. Have a look at those two graphs. both show the balance by customers who did or did not purchase a CD. Wh
Answer: The first graph(box plot) is good to give overall picture and it is easy to see the different between ba
while the histogarm provide the distribution detail.
Conclusion based on the graph: There is higher chance to by CD in the comtomer with higher balance group. */
```

```
proc sgplot data=mylib.CUSTOMER_ALL;
    vbar y / response=balance group=balance groupdisplay=cluster stat=mean;
    yaxis grid;
run;

title 'distribution of balance by y';
proc univariate data=mylib.customer_all noprint;
class y;
histogram balance;
run;


/*Examine the variable pdays
pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; -1

Use proc univariate on the variable pdays.*/
title 'pday histrogram';
proc univariate data=mylib.customer_all;
var pdays;
histogram /;
run;

/*creating a derived variable
By checking the quantiles table in the proc univariate output,
clearly it is better to create a new categorical variable named "contacted_before"
that takes the value 'yes' if the customer has been contacted before and 'no'
if the customer was not contacted before in a previous campaign (pdays=-1)*/

data mylib.customer_all;
set mylib.customer_all;
if pdays = -1 then contacted_before = 'No';
else contacted_before = 'Yes';
run;

*print the first 5 observations where pdays>0;
title 'first 5 observations where pdays>0';
proc print data=mylib.customer_all (obs =5);
where pdays >0;
run;


*drop the column pdays;
title;
data mylib.customer_all;
    set mylib.customer_all ;
run;

*use proc means and make sure pdays is not there;

proc means data=mylib.customer_all;
run;

*Listing the 10 Highest and Lowest Values of balance;

proc univariate data=mylib.CUSTOMER_ALL nextrobs=10;
   id customer_id;
   var balance;
run;

*Using data cleaning techniques for numeric data.;
proc sort data=mylib.CUSTOMER_ALL
out= top10_high;
by descending balance;
run;

proc sort data=mylib.CUSTOMER_ALL
out= top10_low;
by balance;
run;

title 'top10 high balance';
proc print data= top10_high (obs=10);
var customer_id balance;
run;

title 'top10 low balance';
proc print data= top10_low (obs=10);
var customer_id balance;
run;
```

### Log: Data_prep_numeric.sas

Warnings (4)

Notes (73)

```
 1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
 68         
 69         
 70         *1. Examine the list of numerical attributes;
 71         
 72         libname mylib '/home/u63876948/Portfolio/Numerical variable';
NOTE: Libref MYLIB was successfully assigned as follows: 
      Engine:        V9 
      Physical Name: /home/u63876948/Portfolio/Numerical variable
 73         
 74         proc means data=mylib.customer_all;
 75         run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.03 seconds
      user cpu time       0.03 seconds
      system cpu time     0.00 seconds
      memory              7680.34k
      OS Memory           33724.00k
      Timestamp           11/11/2024 12:54:39 AM
      Step Count                        303  Switch Count  1
      Page Faults                       0
      Page Reclaims                     1434
      Page Swaps                        0
      Voluntary Context Switches        52
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           8
      

 76         
 77         proc means data=mylib.customer_all nmiss;
 78         run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.02 seconds
      user cpu time       0.01 seconds
      system cpu time     0.01 seconds
      memory              6697.90k
      OS Memory           33724.00k
      Timestamp           11/11/2024 12:54:39 AM
      Step Count                        304  Switch Count  1
      Page Faults                       0
      Page Reclaims                     1451
      Page Swaps                        0
      Voluntary Context Switches        32
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           0
      

 79         
 80         *Interpret:
 81         pdays: number of days that passed by after the client was last contacted from a previous campaign (Numeric)
 82         and -1 means client was not previously contacted)
 83         Examine the range of values for day. (last contact day of the month) (numeric)
 84         Examine the column N Miss for the variable "age". 20 missing;
 85         
 86         
 87         *Examine the variable "age"
 88         use PROC UNIVARIATE to examine the numeric variable "age" by showing tabular and graphical information.;
 89         
 90         proc univariate data=mylib.customer_all;
 91         id customer_id;
 92         var age;
 93         histogram / normal;
 94         run;

NOTE: PROCEDURE UNIVARIATE used (Total process time):
      real time           0.36 seconds
      user cpu time       0.11 seconds
      system cpu time     0.01 seconds
      memory              14255.31k
      OS Memory           40452.00k
      Timestamp           11/11/2024 12:54:39 AM
      Step Count                        305  Switch Count  0
      Page Faults                       0
      Page Reclaims                     3380
      Page Swaps                        0
      Voluntary Context Switches        808
      Involuntary Context Switches      10
      Block Input Operations            0
      Block Output Operations           712


95         
96         * Output the customer_id whose age is missing. Use the function missing within if statements;
97         data mylib.missing_age;
98             set mylib.customer_all;
99             if missing(age) then output;
100        run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.
NOTE: The data set MYLIB.MISSING_AGE has 20 observations and 17 variables.
NOTE: DATA statement used (Total process time):
      real time           0.01 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              1627.12k
      OS Memory           39084.00k
      Timestamp           11/11/2024 12:54:39 AM
      Step Count                        306  Switch Count  1
      Page Faults                       0
      Page Reclaims                     204
      Page Swaps                        0
      Voluntary Context Switches        34
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


101        
102        proc print data=mylib.missing_age;
103            var customer_id;
104            title 'Customers with Missing Age';
105        run;

NOTE: There were 20 observations read from the data set MYLIB.MISSING_AGE.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.01 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              704.50k
      OS Memory           38824.00k
      Timestamp           11/11/2024 12:54:39 AM
      Step Count                        307  Switch Count  1
      Page Faults                       0
      Page Reclaims                     94
      Page Swaps                        0
      Voluntary Context Switches        19
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           8


106        
107        
108        * Output the customer_id whose age is missing. Use the function missing within if statements;
109        data mylib.missing_age;
110        set mylib.customer_all;
111        if missing(age) then output;
112        run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.
NOTE: The data set MYLIB.MISSING_AGE has 20 observations and 17 variables.
NOTE: DATA statement used (Total process time):
      real time           0.01 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              1514.87k
      OS Memory           39340.00k
      Timestamp           11/11/2024 12:54:39 AM
      Step Count                        308  Switch Count  1
      Page Faults                       0
      Page Reclaims                     200
      Page Swaps                        0
      Voluntary Context Switches        40
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


113        
114        proc print data=mylib.missing_age;
115        var customer_id age;
116        title 'Customers with Missing Age';
117        run;

NOTE: There were 20 observations read from the data set MYLIB.MISSING_AGE.
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              607.59k
      OS Memory           38824.00k
      Timestamp           11/11/2024 12:54:39 AM
      Step Count                        309  Switch Count  1
      Page Faults                       0
      Page Reclaims                     63
      Page Swaps                        0
      Voluntary Context Switches        19
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           16


118
119        *Apply imputation to replace missing values for age with the mean age.;
120        proc stdize data=mylib.customer_all out=mylib.customer_all_Imputed
121              oprefix=Orig_         /* prefix for original variables */
122              reponly               /* only replace; do not standardize */
123              method=MEAN;          /* or MEDIAN, MINIMUM, MIDRANGE, etc. */
124           var age;              /* you can list multiple variables to impute */
125        run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.
NOTE: The data set MYLIB.CUSTOMER_ALL_IMPUTED has 10578 observations and 18 variables.
NOTE: PROCEDURE STDIZE used (Total process time):
      real time           0.01 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              2648.81k
      OS Memory           40364.00k
      Timestamp           11/11/2024 12:54:39 AM
      Step Count                        310  Switch Count  1
      Page Faults                       0
      Page Reclaims                     302
      Page Swaps                        0
      Voluntary Context Switches        45
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           2824


126
127        title 'imputed dataset';
128        proc print data=mylib.customer_all_Imputed (obs=10);
129        run;

NOTE: There were 10 observations read from the data set MYLIB.CUSTOMER_ALL_IMPUTED.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.03 seconds
      user cpu time       0.03 seconds
      system cpu time     0.00 seconds
      memory              2092.71k
      OS Memory           39592.00k
      Timestamp           11/11/2024 12:54:39 AM
      Step Count                        311  Switch Count  0
      Page Faults                       0
      Page Reclaims                     265
      Page Swaps                        0
      Voluntary Context Switches        10
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           16


130        title;
131
132
133        *Use proc means to check the list of numerical attributes in customer_all_Imputed;
134        proc means data= mylib.customer_all_Imputed nmiss;
135        title 'number of mission for imputed dataset';
136        run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL_IMPUTED.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.02 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              7682.46k
      OS Memory           44732.00k
      Timestamp           11/11/2024 12:54:39 AM
      Step Count                        312  Switch Count  2
      Page Faults                       0
      Page Reclaims                     1555
      Page Swaps                        0
      Voluntary Context Switches        51
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           0


137
138        *Rename SAS dataset to its original name customer_all.;
139        data mylib.customer_all;
140         set mylib.customer_all_Imputed;
141        run;
```

```
NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL_IMPUTED.
NOTE: The data set MYLIB.CUSTOMER_ALL has 10578 observations and 18 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              3683.15k
      OS Memory           41644.00k
      Timestamp           11/11/2024 12:54:39 AM
      Step Count                        313  Switch Count  1
      Page Faults                       0
      Page Reclaims                     526
      Page Swaps                        0
      Voluntary Context Switches        50
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           2824


142
143        proc datasets library=mylib;
144        delete customer_all_Imputed;
145        run;

NOTE: Deleting MYLIB.CUSTOMER_ALL_IMPUTED (memtype=DATA).
146
147        *To evaluate if age has an influence on balance?;
148        * use sgplot to draw a scatter plot and regression line ;
149        title 'Influence of age on balance';

NOTE: PROCEDURE DATASETS used (Total process time):
      real time           0.02 seconds
      user cpu time       0.02 seconds
      system cpu time     0.01 seconds
      memory              528.56k
      OS Memory           38824.00k
      Timestamp           11/11/2024 12:54:40 AM
      Step Count                        314  Switch Count  2
      Page Faults                       0
      Page Reclaims                     52
      Page Swaps                        0
      Voluntary Context Switches        37
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           8


150        proc sgplot data=mylib.customer_all;
151        reg x=age y=balance / lineattrs=(color=red thickness=2);
152        run;

NOTE: PROCEDURE SGPLOT used (Total process time):
      real time           0.15 seconds
      user cpu time       0.04 seconds
      system cpu time     0.00 seconds
      memory              4115.59k
      OS Memory           40496.00k
      Timestamp           11/11/2024 12:54:40 AM
      Step Count                        315  Switch Count  5
      Page Faults                       0
      Page Reclaims                     728
      Page Swaps                        0
      Voluntary Context Switches        218
      Involuntary Context Switches      3
      Block Input Operations            0
      Block Output Operations           1752

NOTE: Marker and line antialiasing has been disabled for at least one plot because the threshold has been reached. You can set
      ANTIALIASMAX=10600 in the ODS GRAPHICS statement to enable antialiasing for all plots.
NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.

153
154        *Binning
155        discretize the variable age by creating a new cat variable named age_cat:;
156        data mylib.customer_all;
157        set mylib.customer_all;
158        if 18 <=AGE <=35 then age_cat = 'Young_adult';
159        else if 36 <=AGE <= 55 then age_cat ='Middle_age';
160        else if AGE >= 56 then age_cat ='Old'; /* if the >56 for old., there will be 178 missing value, so chage data to >=
161        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 19 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              3573.75k
      OS Memory           41388.00k
      Timestamp           11/11/2024 12:54:40 AM
      Step Count                        316  Switch Count  1
      Page Faults                       0
      Page Reclaims                     496
      Page Swaps                        0
      Voluntary Context Switches        55
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           3080


162
163        *show a simple frequency table for age_cat;
164        proc freq data=mylib.customer_all;
165        table age_cat;
166        title 'frequency of each age';
167        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     0.01 seconds
      memory              2169.71k
      OS Memory           39852.00k
      Timestamp           11/11/2024 12:54:40 AM
      Step Count                        317  Switch Count  3
      Page Faults                       0
      Page Reclaims                     311
      Page Swaps                        0
      Voluntary Context Switches        52
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


168
169        *Here is the code to create a bar chart of balance by age.;
170        proc sgplot data=mylib.CUSTOMER_ALL;
171            vbar age_cat / response=balance group=balance groupdisplay=cluster stat=mean;
172            yaxis grid;
173        run;

NOTE: PROCEDURE SGPLOT used (Total process time):
      real time           0.40 seconds
      user cpu time       0.07 seconds
      system cpu time     0.06 seconds
      memory              4870.40k
      OS Memory           41916.00k
      Timestamp           11/11/2024 12:54:40 AM
      Step Count                        318  Switch Count  2
      Page Faults                       0
      Page Reclaims                     902
      Page Swaps                        0
      Voluntary Context Switches        16396
      Involuntary Context Switches      5
      Block Input Operations            0
      Block Output Operations           664

WARNING: GROUP=BALANCE on the BARCHARTPARM statement is ignored because the GROUPMAX threshold has been reached. You can set
         GROUPMAX=3,800 on the ODS GRAPHICS statement to enable the GROUP variable.
WARNING: The data for a BARCHARTPARM statement are not appropriate. The BARCHARTPARM statement expects summarized data. The ba
         chart might not be drawn correctly.
NOTE: Marker and line antialiasing has been disabled for at least one plot because the threshold has been reached. You can set
      ANTIALIASMAX=5500 in the ODS GRAPHICS statement to enable antialiasing for all plots.
NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.

174        title;
175
176
177
178        /*Examine the variable campaign
179        campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
180
181        Use proc univariate on campaign.*/
182        proc univariate data=mylib.CUSTOMER_ALL nextrobs=10;
183           id customer_id;
184           var campaign;
185           histogram / normal;
186        run;

NOTE: PROCEDURE UNIVARIATE used (Total process time):
      real time           0.21 seconds
      user cpu time       0.12 seconds
      system cpu time     0.01 seconds
      memory              8780.93k
      OS Memory           42500.00k
      Timestamp           11/11/2024 12:54:40 AM
      Step Count                        319  Switch Count  0
      Page Faults                       0
      Page Reclaims                     996
      Page Swaps                        0
      Voluntary Context Switches        802
      Involuntary Context Switches      7
      Block Input Operations            0
      Block Output Operations           384


187
188
189        * Based on quantiles table, the variable campaign seems more categorical in nature than continuous.
190        Convert the variable campaign into a categorical variable name "campaign_cat" with ordinal values { 1, 2, 3, >3};
191
192        proc contents data=mylib.customer_all;
193        run;
```

```
NOTE: PROCEDURE CONTENTS used (Total process time):
      real time           0.03 seconds
      user cpu time       0.04 seconds
      system cpu time     0.00 seconds
      memory              2277.25k
      OS Memory           40364.00k
      Timestamp           11/11/2024 12:54:40 AM
      Step Count                        320  Switch Count  0
      Page Faults                       0
      Page Reclaims                     286
      Page Swaps                        0
      Voluntary Context Switches        7
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           24


194
195        data mylib.customer_all;
196        set mylib.customer_all;
197        if campaign =1 then campaign_cat = '1';
198        else if campaign =2 then campaign_cat = '2';
199        else if campaign =3 then campaign_cat = '3';
200        else if campaign >3 then campaign_cat = '>3';
201        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 20 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              3688.93k
      OS Memory           41644.00k
      Timestamp           11/11/2024 12:54:40 AM
      Step Count                        321  Switch Count  1
      Page Faults                       0
      Page Reclaims                     491
      Page Swaps                        0
      Voluntary Context Switches        53
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           3080


202
203        proc freq data =mylib.customer_all;
204        table campaign_cat;
205        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              2168.84k
      OS Memory           40108.00k
      Timestamp           11/11/2024 12:54:40 AM
      Step Count                        322  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


206
207        /*Examine the variable "balance"
208        Investigate the distribution of balance. use proc univariate to get the statistics along a histogram for the variab
208      ! balance.*/
209        title "Running PROC UNIVARIATE on balance";
210        proc univariate data=mylib.customer_all noprint;
211           id customer_id;
212           var balance;
213           histogram ;
214        run;

NOTE: PROCEDURE UNIVARIATE used (Total process time):
      real time           0.09 seconds
      user cpu time       0.04 seconds
      system cpu time     0.00 seconds
      memory              8386.09k
      OS Memory           41988.00k
      Timestamp           11/11/2024 12:54:41 AM
      Step Count                        323  Switch Count  0
      Page Faults                       0
      Page Reclaims                     989
      Page Swaps                        0
      Voluntary Context Switches        135
      Involuntary Context Switches      6
      Block Input Operations            0
      Block Output Operations           360


215
216        /*4.2. Have a look at those two graphs. both show the balance by customers who did or did not purchase a CD. Which
```

```
 216      ! is more informative? What conclusion can you formulate based on the graph?
 217        Answer: The first graph(box plot) is good to give overall picture and it is easy to see the different between balan
 217      !  buy or not by CD,
 218        while the histogarm provide the distribution detail.
 219        Conclusion based on the graph: There is higher chance to by CD in the comtomer with higher balance group. */
 220
 221        proc sgplot data=mylib.CUSTOMER_ALL;
 222            vbar y / response=balance group=balance groupdisplay=cluster stat=mean;
 223            yaxis grid;
 224        run;

 NOTE: PROCEDURE SGPLOT used (Total process time):
       real time           0.32 seconds
       user cpu time       0.06 seconds
       system cpu time     0.06 seconds
       memory              4694.50k
       OS Memory           41916.00k
       Timestamp           11/11/2024 12:54:41 AM
       Step Count                        324  Switch Count  2
       Page Faults                       0
       Page Reclaims                     821
       Page Swaps                        0
       Voluntary Context Switches        15059
       Involuntary Context Switches      6
       Block Input Operations            0
       Block Output Operations           640

 WARNING: GROUP=BALANCE on the BARCHARTPARM statement is ignored because the GROUPMAX threshold has been reached. You can set
          GROUPMAX=3,800 on the ODS GRAPHICS statement to enable the GROUP variable.
 WARNING: The data for a BARCHARTPARM statement are not appropriate. The BARCHARTPARM statement expects summarized data. The ba
          chart might not be drawn correctly.
 NOTE: Marker and line antialiasing has been disabled for at least one plot because the threshold has been reached. You can set
       ANTIALIASMAX=5000 in the ODS GRAPHICS statement to enable antialiasing for all plots.
 NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.

 225
 226        title 'distribution of balance by y';
 227        proc univariate data=mylib.customer_all noprint;
 228        class y;
 229        histogram balance;
 230        run;

 NOTE: PROCEDURE UNIVARIATE used (Total process time):
       real time           0.21 seconds
       user cpu time       0.04 seconds
       system cpu time     0.00 seconds
       memory              4693.96k
       OS Memory           40752.00k
       Timestamp           11/11/2024 12:54:41 AM
       Step Count                        325  Switch Count  0
       Page Faults                       0
       Page Reclaims                     500
       Page Swaps                        0
       Voluntary Context Switches        405
       Involuntary Context Switches      1
       Block Input Operations            0
       Block Output Operations           376


 231
 232
 233        /*Examine the variable pdays
 234        pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; -1 mean
 234      ! client was not previously contacted)
 235
 236        Use proc univariate on the variable pdays.*/
 237        title 'pday histrogram';
 238        proc univariate data=mylib.customer_all;
 239        var pdays;
 240        histogram /;
 241        run;

 NOTE: PROCEDURE UNIVARIATE used (Total process time):
       real time           0.11 seconds
       user cpu time       0.07 seconds
       system cpu time     0.01 seconds
       memory              8466.43k
       OS Memory           42244.00k
       Timestamp           11/11/2024 12:54:41 AM
       Step Count                        326  Switch Count  0
       Page Faults                       0
       Page Reclaims                     1066
       Page Swaps                        0
       Voluntary Context Switches        135
       Involuntary Context Switches      3
       Block Input Operations            0
       Block Output Operations           360


 242
 243        /*creating a derived variable
 244        By checking the quantiles table in the proc univariate output,
 245        clearly it is better to create a new categorical variable named "contacted_before"
 246        that takes the value 'yes' if the customer has been contacted before and 'no'
 247        if the customer was not contacted before in a previous campaign (pdays=-1)*/
 248
 249        data mylib.customer_all;
 250        set mylib.customer_all;
```

```
           if pdays = -1 then contacted_before = 'No';
           else contacted_before = 'Yes';
           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 21 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              3572.84k
      OS Memory           41900.00k
      Timestamp           11/11/2024 12:54:41 AM
      Step Count                        327  Switch Count  1
      Page Faults                       0
      Page Reclaims                     526
      Page Swaps                        0
      Voluntary Context Switches        53
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           3080



           *print the first 5 observations where pdays>0;
           title 'first 5 observations where pdays>0';
           proc print data=mylib.customer_all (obs =5);
           where pdays >0;
           run;

NOTE: There were 5 observations read from the data set MYLIB.CUSTOMER_ALL.
      WHERE pdays>0;
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.02 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              2445.12k
      OS Memory           40108.00k
      Timestamp           11/11/2024 12:54:41 AM
      Step Count                        328  Switch Count  0
      Page Faults                       0
      Page Reclaims                     293
      Page Swaps                        0
      Voluntary Context Switches        15
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           8




           *drop the column pdays;
           title;
           data mylib.customer_all;
               set mylib.customer_all ;
           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 21 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              3571.31k
      OS Memory           41644.00k
      Timestamp           11/11/2024 12:54:41 AM
      Step Count                        329  Switch Count  1
      Page Faults                       0
      Page Reclaims                     491
      Page Swaps                        0
      Voluntary Context Switches        63
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           3080



           *use proc means and make sure pdays is not there;

           proc means data=mylib.customer_all;
           run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.03 seconds
      user cpu time       0.03 seconds
      system cpu time     0.00 seconds
      memory              7566.12k
      OS Memory           44988.00k
      Timestamp           11/11/2024 12:54:41 AM
      Step Count                        330  Switch Count  1
      Page Faults                       0
      Page Reclaims                     1538
      Page Swaps                        0
      Voluntary Context Switches        52
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           0
```

```
272
273        *Listing the 10 Highest and Lowest Values of balance;
274
275        proc univariate data=mylib.CUSTOMER_ALL nextrobs=10;
276           id customer_id;
277           var balance;
278        run;

NOTE: PROCEDURE UNIVARIATE used (Total process time):
      real time           0.04 seconds
      user cpu time       0.04 seconds
      system cpu time     0.00 seconds
      memory              2333.31k
      OS Memory           39848.00k
      Timestamp           11/11/2024 12:54:41 AM
      Step Count                        331  Switch Count  0
      Page Faults                       0
      Page Reclaims                     245
      Page Swaps                        0
      Voluntary Context Switches        7
      Involuntary Context Switches      3
      Block Input Operations            0
      Block Output Operations           32


279
280        *Using data cleaning techniques for numeric data.;
281        proc sort data=mylib.CUSTOMER_ALL
282        out= top10_high;
283        by descending balance;
284        run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.
NOTE: The data set WORK.TOP10_HIGH has 10578 observations and 21 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              4852.53k
      OS Memory           42432.00k
      Timestamp           11/11/2024 12:54:41 AM
      Step Count                        332  Switch Count  2
      Page Faults                       0
      Page Reclaims                     766
      Page Swaps                        0
      Voluntary Context Switches        19
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           3096


285
286        proc sort data=mylib.CUSTOMER_ALL
287        out= top10_low;
288        by balance;
289        run;

NOTE: There were 10578 observations read from the data set MYLIB.CUSTOMER_ALL.
NOTE: The data set WORK.TOP10_LOW has 10578 observations and 21 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time           0.00 seconds
      user cpu time       0.01 seconds
      system cpu time     0.01 seconds
      memory              4852.15k
      OS Memory           42432.00k
      Timestamp           11/11/2024 12:54:41 AM
      Step Count                        333  Switch Count  2
      Page Faults                       0
      Page Reclaims                     766
      Page Swaps                        0
      Voluntary Context Switches        20
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           3088


290
291        title 'top10 high balance';
292        proc print data= top10_high (obs=10);
293        var customer_id balance;
294        run;

NOTE: There were 10 observations read from the data set WORK.TOP10_HIGH.
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              2017.00k
      OS Memory           39848.00k
      Timestamp           11/11/2024 12:54:41 AM
      Step Count                        334  Switch Count  0
      Page Faults                       0
      Page Reclaims                     254
      Page Swaps                        0
      Voluntary Context Switches        0
      Involuntary Context Switches      0
      Block Input Operations           0
      Block Output Operations          0


295
296        title 'top10 low balance';
297        proc print data= top10_low (obs=10);
298        var customer_id balance;
299        run;

NOTE: There were 10 observations read from the data set WORK.TOP10_LOW.
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              2129.75k
      OS Memory           39848.00k
      Timestamp           11/11/2024 12:54:41 AM
      Step Count                        335  Switch Count  0
      Page Faults                       0
      Page Reclaims                     254
      Page Swaps                        0
      Voluntary Context Switches        0
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           0


300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316        OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
326
```

## Results: Data_prep_numeric.sas

**The MEANS Procedure**

| Variable | Label | N | Mean | Std Dev | Minimum | Maximum |
|---|---|---|---|---|---|---|
| customer_id | | 10578 | 127278.17 | 13660.22 | 100103.00 | 145309.00 |
| day | day | 10578 | 15.4758934 | 8.4137946 | 1.0000000 | 31.0000000 |
| campaign | campaign | 10578 | 2.4747589 | 2.6151781 | 1.0000000 | 50.0000000 |
| pdays | pdays | 10578 | 51.9548119 | 109.3471124 | -1.0000000 | 854.0000000 |
| previous | previous | 10578 | 0.8525241 | 3.4721156 | 0 | 275.0000000 |
| balance | | 10578 | 1548.53 | 3130.57 | -3058.00 | 81204.00 |
| AGE | AGE | 10558 | 41.2641599 | 12.1483452 | 18.0000000 | 146.0000000 |

**The MEANS Procedure**

| Variable | Label | N Miss |
|---|---|---|
| customer_id | | 0 |
| day | day | 0 |
| campaign | campaign | 0 |
| pdays | pdays | 0 |
| previous | previous | 0 |
| balance | | 0 |
| AGE | AGE | 20 |

**The UNIVARIATE Procedure**
**Variable: AGE (AGE)**

| Moments | | | |
|---|---|---|---|
| N | 10558 | Sum Weights | 10558 |
| Mean | 41.2641599 | Sum Observations | 435667 |
| Std Deviation | 12.1483452 | Variance | 147.582292 |
| Skewness | 1.00818411 | Kurtosis | 2.05204285 |
| Uncorrected SS | 19535459 | Corrected SS | 1558026.26 |
| Coeff Variation | 29.4404279 | Std Error Mean | 0.11822962 |

| Basic Statistical Measures | | | |
|---|---|---|---|
| Location | | Variability | |
| Mean | 41.26416 | Std Deviation | 12.14835 |
| Median | 39.00000 | Variance | 147.58229 |
| Mode | 31.00000 | Range | 128.00000 |
| | | Interquartile Range | 17.00000 |

| Tests for Location: Mu0=0 | | | | |
|---|---|---|---|---|
| Test | Statistic | | p Value | |
| Student's t | t | 349.0171 | Pr > \|t\| | <.0001 |
| Sign | M | 5279 | Pr >= \|M\| | <.0001 |

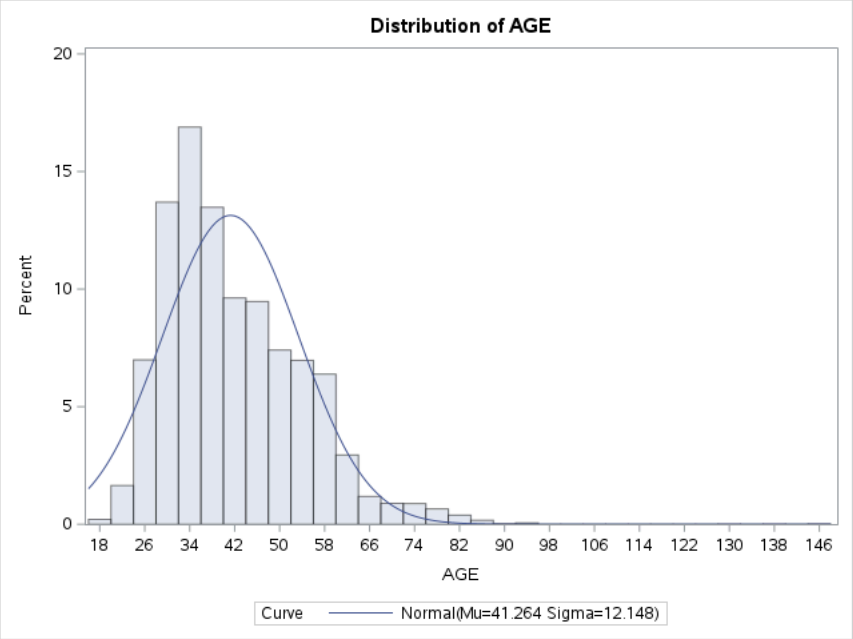
| Tests for Location: Mu0=0 | | | | |
|---|---|---|---|---|
| Test | Statistic | | p Value | |
| Signed Rank | S | 27870481 | Pr >= \|S\| | <.0001 |

| Quantiles (Definition 5) | |
|---|---|
| Level | Quantile |
| 100% Max | 146 |
| 99% | 77 |
| 95% | 62 |
| 90% | 58 |
| 75% Q3 | 49 |
| 50% Median | 39 |
| 25% Q1 | 32 |
| 10% | 28 |
| 5% | 26 |
| 1% | 22 |
| 0% Min | 18 |

| Extreme Observations | | | | | |
|---|---|---|---|---|---|
| Lowest | | | Highest | | |
| Value | customer_id | Obs | Value | customer_id | Obs |
| 18 | 144745 | 10273 | 95 | 141764 | 8550 |
| 18 | 143738 | 9679 | 130 | 120217 | 3490 |
| 18 | 143055 | 9240 | 139 | 107284 | 1152 |
| 18 | 142375 | 8880 | 144 | 109385 | 1531 |
| 18 | 141588 | 8455 | 146 | 102598 | 402 |

| Missing Values | | | |
|---|---|---|---|
| | | Percent Of | |
| Missing Value | Count | All Obs | Missing Obs |
| . | 20 | 0.19 | 100.00 |

The UNIVARIATE Procedure



The UNIVARIATE Procedure
Fitted Normal Distribution for AGE (AGE)

| Parameters for Normal Distribution | | |
|---|---|---|
| Parameter | Symbol | Estimate |
| Mean | Mu | 41.26416 |
| Std Dev | Sigma | 12.14835 |

| Goodness-of-Fit Tests for Normal Distribution | | | | |
|---|---|---|---|---|
| Test | Statistic | | p Value | |
| Kolmogorov-Smirnov | D | 0.106237 | Pr > D | <0.010 |
| Cramer-von Mises | W-Sq | 25.931403 | Pr > W-Sq | <0.005 |
| Anderson-Darling | A-Sq | 150.997380 | Pr > A-Sq | <0.005 |

| Quantiles for Normal Distribution | | |
|---|---|---|
| | Quantile | |
| Percent | Observed | Estimated |
| 1.0 | 22.0000 | 13.0029 |
| 5.0 | 26.0000 | 21.2819 |
| 10.0 | 28.0000 | 25.6954 |
| 25.0 | 32.0000 | 33.0702 |
| 50.0 | 39.0000 | 41.2642 |

**Quantiles for Normal Distribution**

| | Quantile | |
|---|---|---|
| Percent | Observed | Estimated |
| 75.0 | 49.0000 | 49.4581 |
| 90.0 | 58.0000 | 56.8329 |
| 95.0 | 62.0000 | 61.2464 |
| 99.0 | 77.0000 | 69.5254 |

## Customers with Missing Age

| Obs | customer_id |
|---|---|
| 1 | 100898 |
| 2 | 103782 |
| 3 | 104872 |
| 4 | 108581 |
| 5 | 112972 |
| 6 | 113317 |
| 7 | 114933 |
| 8 | 115167 |
| 9 | 117338 |
| 10 | 122821 |
| 11 | 127452 |
| 12 | 128062 |
| 13 | 128123 |
| 14 | 131745 |
| 15 | 134418 |
| 16 | 134663 |
| 17 | 135384 |
| 18 | 135695 |
| 19 | 143464 |
| 20 | 143512 |

## Customers with Missing Age

| Obs | customer_id | AGE |
|---|---|---|
| 1 | 100898 | . |
| 2 | 103782 | . |
| 3 | 104872 | . |
| 4 | 108581 | . |
| 5 | 112972 | . |
| 6 | 113317 | . |
| 7 | 114933 | . |
| 8 | 115167 | . |
| 9 | 117338 | . |
| 10 | 122821 | . |
| 11 | 127452 | . |
| 12 | 128062 | . |
| 13 | 128123 | . |
| 14 | 131745 | . |
| 15 | 134418 | . |
| 16 | 134663 | . |
| 17 | 135384 | . |
| 18 | 135695 | . |
| 19 | 143464 | . |
| 20 | 143512 | . |

## imputed dataset

| Obs | customer_id | contact | day | month | campaign | pdays | previous | poutcome | y | default | balance | housing | loan | Education | Orig_AGE | marital | JOB | AGE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 100103 | unknown | 5 | may | 1 | -1 | 0 | unknown | no | no | 2 | yes | yes | secondary | 33 | married | entrepreneur | 33 |
| 2 | 100106 | unknown | 5 | may | 1 | -1 | 0 | unknown | no | no | 231 | yes | no | tertiary | 35 | married | management | 35 |
| 3 | 100118 | unknown | 5 | may | 1 | -1 | 0 | unknown | no | no | 52 | yes | no | primary | 57 | married | blue-collar | 57 |
| 4 | 100119 | unknown | 5 | may | 1 | -1 | 0 | unknown | no | no | 60 | yes | no | primary | 60 | married | retired | 60 |
| 5 | 100121 | unknown | 5 | may | 1 | -1 | 0 | unknown | no | no | 723 | yes | yes | secondary | 28 | married | blue-collar | 28 |
| 6 | 100126 | unknown | 5 | may | 1 | -1 | 0 | unknown | no | no | -372 | yes | no | secondary | 44 | married | admin. | 44 |
| 7 | 100130 | unknown | 5 | may | 1 | -1 | 0 | unknown | no | no | 265 | yes | yes | secondary | 36 | single | technician | 36 |
| 8 | 100141 | unknown | 5 | may | 1 | -1 | 0 | unknown | no | no | 2586 | yes | no | secondary | 44 | divorced | services | 44 |
| 9 | 100161 | unknown | 5 | may | 1 | -1 | 0 | unknown | no | no | 0 | yes | no | tertiary | 32 | married | admin. | 32 |
| 10 | 100168 | unknown | 5 | may | 1 | -1 | 0 | unknown | no | no | 59 | yes | no | tertiary | 59 | divorced | management | 59 |

## number of mission for imputed dataset

**The MEANS Procedure**

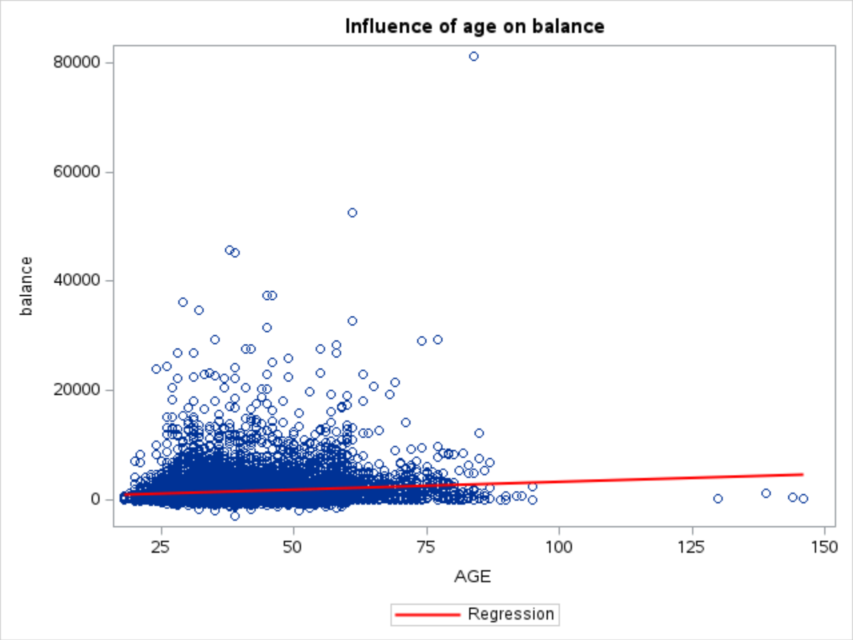
| Variable | Label | N Miss |
|---|---|---|

| Variable | Label | N Miss |
|---|---|---|
| customer_id | | 0 |
| day | day | 0 |
| campaign | campaign | 0 |
| pdays | pdays | 0 |
| previous | previous | 0 |
| balance | | 0 |
| Orig_AGE | AGE | 20 |
| AGE | AGE | 0 |

## number of mission for imputed dataset

| Directory | |
|---|---|
| Libref | MYLIB |
| Engine | V9 |
| Physical Name | /home/u63876948/Portfolio/Numerical variable |
| Filename | /home/u63876948/Portfolio/Numerical variable |
| Inode Number | 14163067942 |
| Access Permission | rwxr-xr-x |
| Owner Name | u63876948 |
| File Size | 0KB |
| File Size (bytes) | 149 |

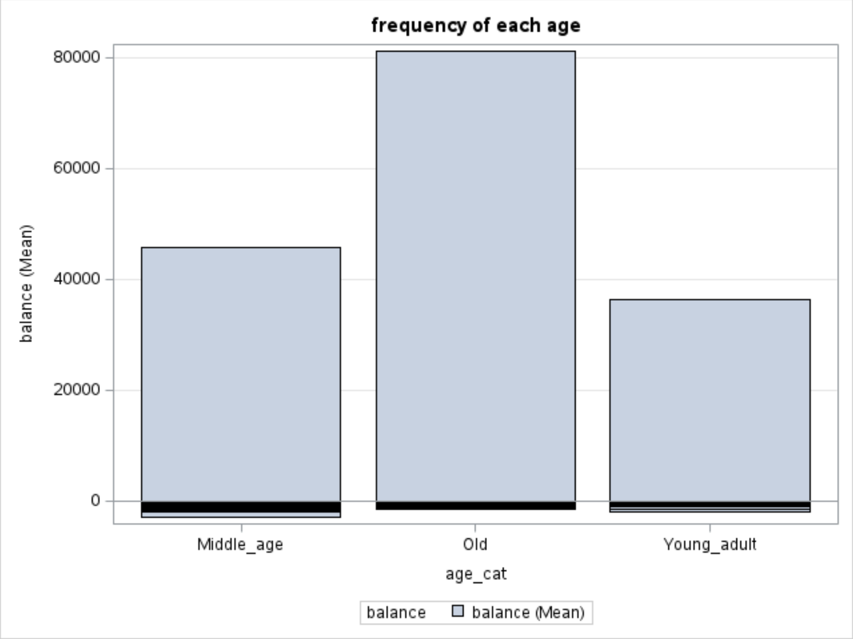
| # | Name | Member Type | File Size | Last Modified |
|---|---|---|---|---|
| 1 | CUSTOMER_ALL | DATA | 2MB | 11/11/2024 00:54:39 |
| 2 | CUSTOMER_ALL_IMPUTED | DATA | 2MB | 11/11/2024 00:54:39 |
| 3 | MISSING_AGE | DATA | 256KB | 11/11/2024 00:54:39 |



## frequency of each age

### The FREQ Procedure

| age_cat | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| Middle_age | 4977 | 47.05 | 4977 | 47.05 |
| Old | 1438 | 13.59 | 6415 | 60.64 |
| Young_adult | 4163 | 39.36 | 10578 | 100.00 |



**The UNIVARIATE Procedure**
**Variable: campaign (campaign)**

| Moments | | | |
|---|---|---|---|
| N | 10578 | Sum Weights | 10578 |
| Mean | 2.47475893 | Sum Observations | 26178 |
| Std Deviation | 2.61517814 | Variance | 6.83915672 |
| Skewness | 5.0976061 | Kurtosis | 44.6295296 |
| Uncorrected SS | 137122 | Corrected SS | 72337.7606 |
| Coeff Variation | 105.674056 | Std Error Mean | 0.02542726 |

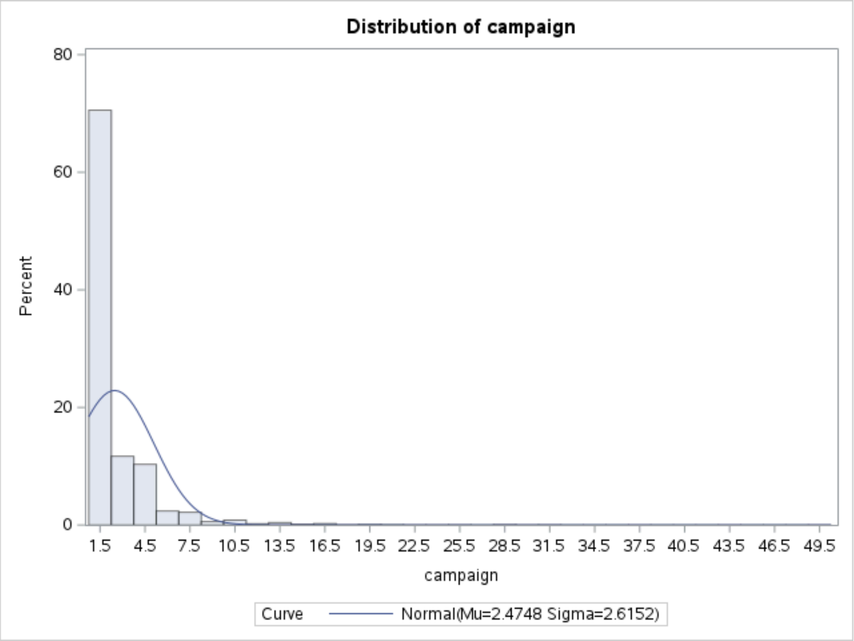
| Basic Statistical Measures | | | |
|---|---|---|---|
| Location | | Variability | |
| Mean | 2.474759 | Std Deviation | 2.61518 |
| Median | 2.000000 | Variance | 6.83916 |
| Mode | 1.000000 | Range | 49.00000 |
| | | Interquartile Range | 2.00000 |

| Tests for Location: Mu0=0 | | | | |
|---|---|---|---|---|
| Test | Statistic | | p Value | |
| Student's t | t | 97.32702 | Pr > \|t\| | <.0001 |
| Sign | M | 5289 | Pr >= \|M\| | <.0001 |
| Signed Rank | S | 27976166 | Pr >= \|S\| | <.0001 |

| Quantiles (Definition 5) | |
|---|---|
| Level | Quantile |
| 100% Max | 50 |
| 99% | 14 |
| 95% | 7 |
| 90% | 5 |
| 75% Q3 | 3 |
| 50% Median | 2 |
| 25% Q1 | 1 |
| 10% | 1 |
| 5% | 1 |
| 1% | 1 |
| 0% Min | 1 |

| Extreme Observations | | | | | |
|---|---|---|---|---|---|
| Lowest | | | Highest | | |
| Value | customer_id | Obs | Value | customer_id | Obs |
| 1 | 145305 | 10574 | 29 | 118799 | 3210 |
| 1 | 145304 | 10573 | 30 | 113035 | 2137 |
| 1 | 145303 | 10572 | 30 | 117173 | 2945 |
| 1 | 145302 | 10571 | 31 | 111495 | 1888 |
| 1 | 145298 | 10569 | 31 | 115970 | 2726 |
| 1 | 145297 | 10568 | 31 | 118202 | 3114 |
| 1 | 145296 | 10567 | 32 | 103432 | 543 |
| 1 | 145293 | 10565 | 37 | 110065 | 1643 |
| 1 | 145292 | 10564 | 43 | 113776 | 2276 |
| 1 | 145291 | 10563 | 50 | 118814 | 3214 |

**The UNIVARIATE Procedure**



**The UNIVARIATE Procedure**
**Fitted Normal Distribution for campaign (campaign)**

| Parameters for Normal Distribution | | |
|---|---|---|
| **Parameter** | **Symbol** | **Estimate** |
| **Mean** | Mu | 2.474759 |
| **Std Dev** | Sigma | 2.615178 |

| Goodness-of-Fit Tests for Normal Distribution | | | | |
|---|---|---|---|---|
| **Test** | **Statistic** | | **p Value** | |
| **Kolmogorov-Smirnov** | **D** | 0.28640 | **Pr > D** | <0.010 |
| **Cramer-von Mises** | **W-Sq** | 228.76726 | **Pr > W-Sq** | <0.005 |
| **Anderson-Darling** | **A-Sq** | 1214.01226 | **Pr > A-Sq** | <0.005 |

| Quantiles for Normal Distribution | | |
|---|---|---|
| | **Quantile** | |
| **Percent** | **Observed** | **Estimated** |
| **1.0** | 1.00000 | -3.60906 |
| **5.0** | 1.00000 | -1.82683 |
| **10.0** | 1.00000 | -0.87673 |
| **25.0** | 1.00000 | 0.71085 |
| **50.0** | 2.00000 | 2.47476 |
| **75.0** | 3.00000 | 4.23867 |
| **90.0** | 5.00000 | 5.82624 |
| **95.0** | 7.00000 | 6.77634 |
| **99.0** | 14.00000 | 8.55857 |

**The CONTENTS Procedure**

| | | | |
|---|---|---|---|
| **Data Set Name** | MYLIB.CUSTOMER_ALL | **Observations** | 10578 |
| **Member Type** | DATA | **Variables** | 19 |
| **Engine** | V9 | **Indexes** | 0 |
| **Created** | 11/10/2024 19:54:40 | **Observation Length** | 144 |
| **Last Modified** | 11/10/2024 19:54:40 | **Deleted Observations** | 0 |
| **Protection** | | **Compressed** | NO |
| **Data Set Type** | | **Sorted** | NO |
| **Label** | | | |
| **Data Representation** | SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64 | | |
| **Encoding** | utf-8 Unicode (UTF-8) | | |

| Engine/Host Dependent Information | |
|---|---|
| **Data Set Page Size** | 131072 |
| **Number of Data Set Pages** | 12 |
| **First Data Page** | 1 |
| **Max Obs per Page** | 909 |
| **Obs in First Data Page** | 879 |
| **Number of Data Set Repairs** | 0 |
| **Filename** | /home/u63876948/Portfolio/Numerical variable/customer_all.sas7bdat |
| **Release Created** | 9.0401M7 |
| **Host Created** | Linux |
| **Inode Number** | 14160407297 |
| **Access Permission** | rw-r--r-- |
| **Owner Name** | u63876948 |
| **File Size** | 2MB |

| Engine/Host Dependent Information | |
|---|---|
| File Size (bytes) | 1703936 |

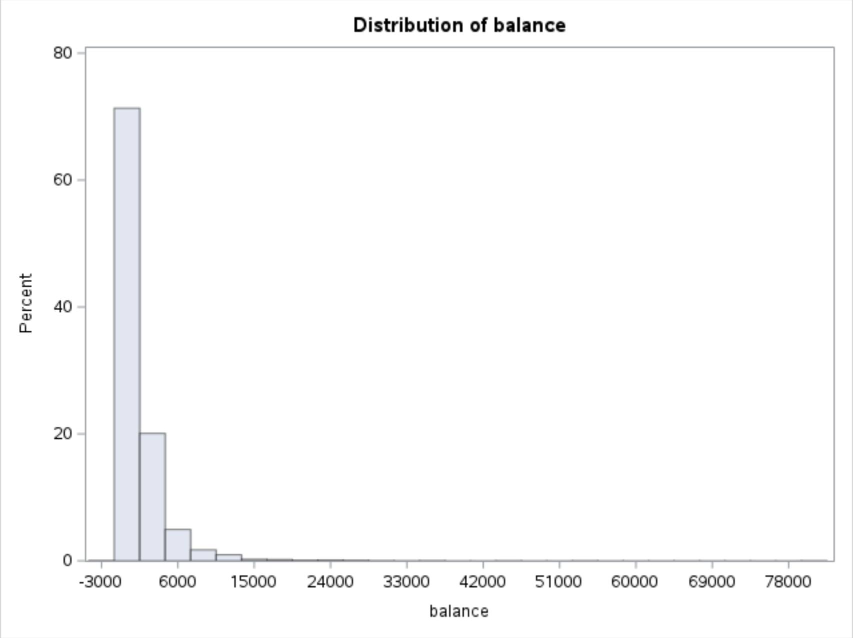
| Alphabetic List of Variables and Attributes | | | | | | |
|---|---|---|---|---|---|---|
| # | Variable | Type | Len | Format | Informat | Label |
| 18 | AGE | Num | 8 | F4. | | AGE |
| 14 | Education | Char | 9 | $CHAR9. | | Education |
| 17 | JOB | Char | 14 | $CHAR14. | | JOB |
| 15 | Orig_AGE | Num | 8 | F4. | | AGE |
| 19 | age_cat | Char | 11 | | | |
| 11 | balance | Num | 8 | BEST12. | BEST32. | |
| 5 | campaign | Num | 8 | BEST. | | campaign |
| 2 | contact | Char | 9 | $9. | $9. | contact |
| 1 | customer_id | Num | 8 | BEST12. | BEST32. | |
| 3 | day | Num | 8 | BEST. | | day |
| 10 | default | Char | 3 | $3. | $3. | |
| 12 | housing | Char | 3 | $3. | $3. | |
| 13 | loan | Char | 3 | $3. | $3. | |
| 16 | marital | Char | 8 | $CHAR8. | | marital |
| 4 | month | Char | 3 | $3. | $3. | month |
| 6 | pdays | Num | 8 | BEST. | | pdays |
| 8 | poutcome | Char | 7 | $7. | $7. | poutcome |
| 7 | previous | Num | 8 | BEST. | | previous |
| 9 | y | Char | 3 | $3. | $3. | y |

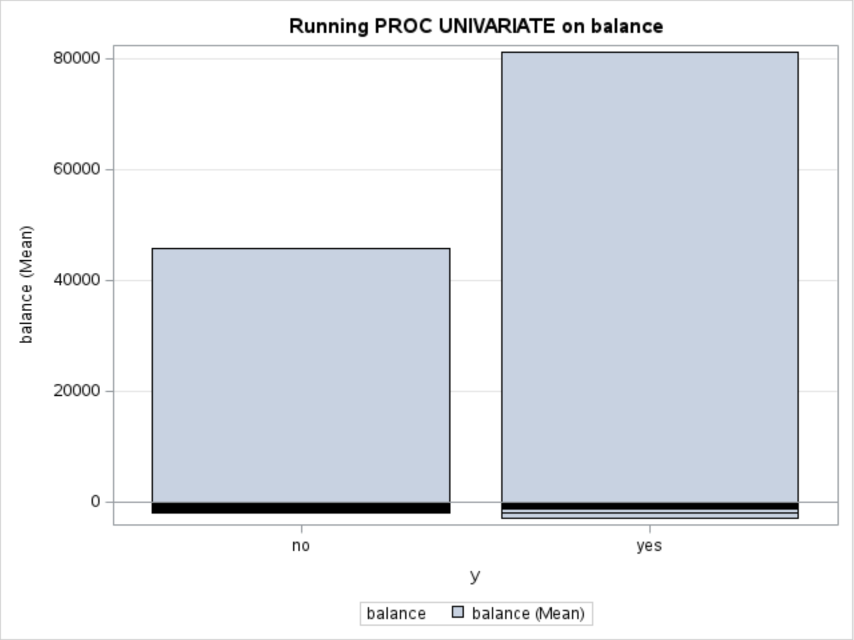
### The FREQ Procedure

| campaign_cat | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 1 | 4556 | 43.07 | 4556 | 43.07 |
| 2 | 2907 | 27.48 | 7463 | 70.55 |
| 3 | 1237 | 11.69 | 8700 | 82.25 |
| > | 1878 | 17.75 | 10578 | 100.00 |

## Running PROC UNIVARIATE on balance

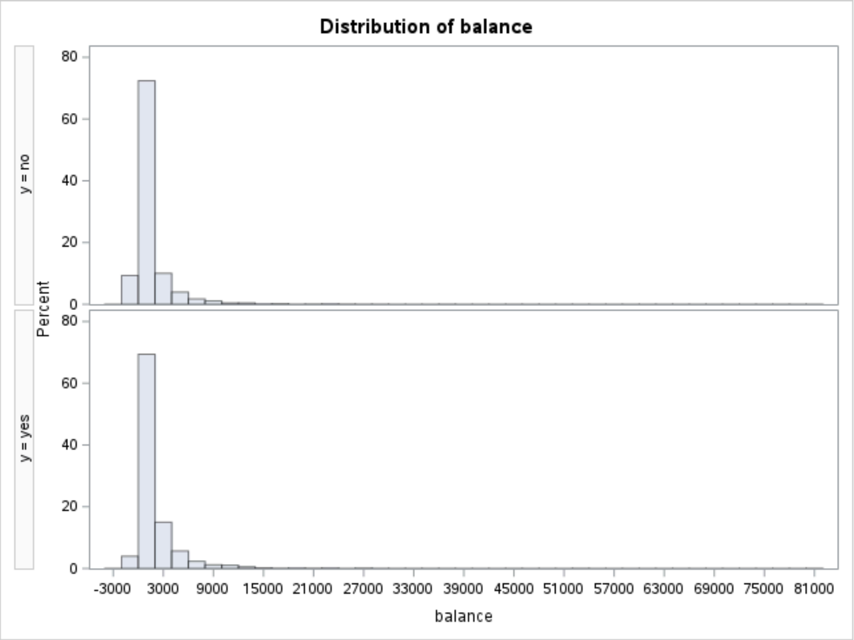
### The UNIVARIATE Procedure





## distribution of balance by y

### The UNIVARIATE Procedure



## pday histrogram

### The UNIVARIATE Procedure
### Variable: pdays (pdays)

| Moments | | | |
|---|---|---|---|
| N | 10578 | Sum Weights | 10578 |
| Mean | 51.9548119 | Sum Observations | 549578 |
| Std Deviation | 109.347112 | Variance | 11956.791 |
| Skewness | 2.41099367 | Kurtosis | 6.46379989 |
| Uncorrected SS | 155020200 | Corrected SS | 126466978 |
| Coeff Variation | 210.465804 | Std Error Mean | 1.06317691 |

| Basic Statistical Measures | | | |
|---|---|---|---|
| Location | | Variability | |
| Mean | 51.95481 | Std Deviation | 109.34711 |
| Median | -1.00000 | Variance | 11957 |
| Mode | -1.00000 | Range | 855.00000 |
| | | Interquartile Range | 50.00000 |

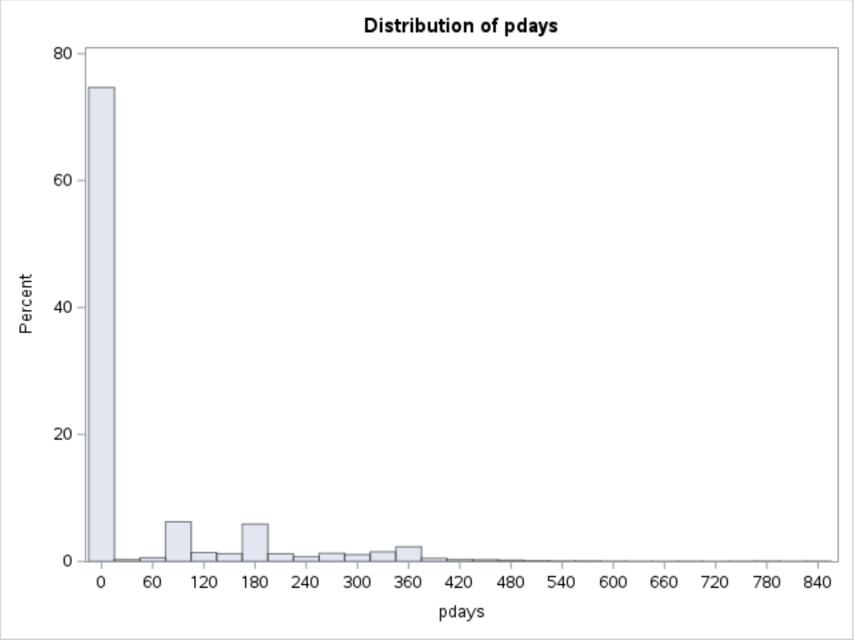
| Tests for Location: Mu0=0 | | | | |
|---|---|---|---|---|
| Test | Statistic | | p Value | |
| Student's t | t | 48.86751 | Pr > \|t\| | <.0001 |
| Sign | M | -2577 | Pr >= \|M\| | <.0001 |
| Signed Rank | S | -2988344 | Pr >= \|S\| | <.0001 |

| Quantiles (Definition 5) | |
|---|---|
| **Level** | **Quantile** |
| **100% Max** | 854 |
| **99%** | 430 |
| **95%** | 329 |
| **90%** | 192 |
| **75% Q3** | 49 |
| **50% Median** | -1 |
| **25% Q1** | -1 |
| **10%** | -1 |
| **5%** | -1 |
| **1%** | -1 |
| **0% Min** | -1 |

| Extreme Observations | | | |
|---|---|---|---|
| **Lowest** | | **Highest** | |
| **Value** | **Obs** | **Value** | **Obs** |
| -1 | 10577 | 805 | 10526 |
| -1 | 10576 | 828 | 10217 |
| -1 | 10575 | 831 | 10299 |
| -1 | 10573 | 842 | 10396 |
| -1 | 10572 | 854 | 10380 |

## pday histrogram

**The UNIVARIATE Procedure**



## first 5 observations where pdays>0

| Obs | customer_id | contact | day | month | campaign | pdays | previous | poutcome | y | default | balance | housing | loan | Education | Orig_AGE | marital | JOB | AGE | age_cat | campaign |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **4210** | 124163 | telephone | 21 | oct | 1 | 166 | 1 | other | yes | no | -247 | yes | yes | secondary | 42 | single | admin. | 42 | Middle_age | 1 |
| **4211** | 124165 | telephone | 21 | oct | 1 | 91 | 4 | failure | yes | no | 3444 | yes | no | secondary | 33 | married | services | 33 | Young_adult | 1 |
| **4218** | 124178 | telephone | 23 | oct | 1 | 143 | 3 | failure | yes | no | 0 | yes | no | tertiary | 36 | married | management | 36 | Middle_age | 1 |
| **4221** | 124181 | unknown | 23 | oct | 1 | 147 | 2 | success | yes | no | 589 | yes | no | secondary | 56 | married | technician | 56 | Old | 1 |
| **4262** | 124237 | unknown | 6 | nov | 1 | 101 | 11 | other | no | no | 1770 | yes | no | tertiary | 34 | married | management | 34 | Young_adult | 1 |

**The MEANS Procedure**

| Variable | Label | N | Mean | Std Dev | Minimum | Maximum |
|---|---|---|---|---|---|---|
| customer_id | | 10578 | 127278.17 | 13660.22 | 100103.00 | 145309.00 |
| day | day | 10578 | 15.4758934 | 8.4137946 | 1.0000000 | 31.0000000 |
| campaign | campaign | 10578 | 2.4747589 | 2.6151781 | 1.0000000 | 50.0000000 |
| pdays | pdays | 10578 | 51.9548119 | 109.3471124 | -1.0000000 | 854.0000000 |
| previous | previous | 10578 | 0.8525241 | 3.4721156 | 0 | 275.0000000 |
| balance | | 10578 | 1548.53 | 3130.57 | -3058.00 | 81204.00 |
| Orig_AGE | AGE | 10558 | 41.2641599 | 12.1483452 | 18.0000000 | 146.0000000 |
| AGE | AGE | 10578 | 41.2641599 | 12.1368542 | 18.0000000 | 146.0000000 |

**The UNIVARIATE Procedure**
**Variable: balance**

| Moments | | | |
|---|---|---|---|
| **N** | 10578 | **Sum Weights** | 10578 |
| **Mean** | 1548.52978 | **Sum Observations** | 16380348 |
| **Std Deviation** | 3130.5653 | **Variance** | 9800439.07 |
| **Skewness** | 7.71681305 | **Kurtosis** | 119.649924 |
| **Uncorrected SS** | 1.29025E11 | **Corrected SS** | 1.03659E11 |

| Moments | | | |
|---|---|---|---|
| Coeff Variation | 202.163713 | Std Error Mean | 30.4383415 |

| Basic Statistical Measures | | | |
|---|---|---|---|
| Location | | Variability | |
| Mean | 1548.530 | Std Deviation | 3131 |
| Median | 566.000 | Variance | 9800439 |
| Mode | 0.000 | Range | 84262 |
| | | Interquartile Range | 1640 |

| Tests for Location: Mu0=0 | | | | |
|---|---|---|---|---|
| Test | Statistic | | p Value | |
| Student's t | t | 50.87432 | Pr > \|t\| | <.0001 |
| Sign | M | 4221.5 | Pr >= \|M\| | <.0001 |
| Signed Rank | S | 22496590 | Pr >= \|S\| | <.0001 |

| Quantiles (Definition 5) | |
|---|---|
| Level | Quantile |
| 100% Max | 81204 |
| 99% | 13118 |
| 95% | 6158 |
| 90% | 3994 |
| 75% Q3 | 1765 |
| 50% Median | 566 |
| 25% Q1 | 125 |
| 10% | 0 |
| 5% | -76 |
| 1% | -542 |
| 0% Min | -3058 |

| Extreme Observations | | | | | |
|---|---|---|---|---|---|
| Lowest | | | Highest | | |
| Value | customer_id | Obs | Value | customer_id | Obs |
| -3058 | 132814 | 6058 | 34646 | 120838 | 3616 |
| -1980 | 120418 | 3530 | 36252 | 134271 | 6501 |
| -1944 | 135493 | 6753 | 37378 | 102879 | 452 |
| -1781 | 119684 | 3396 | 37378 | 141898 | 8605 |
| -1668 | 115067 | 2532 | 45248 | 100547 | 71 |
| -1598 | 118222 | 3117 | 45789 | 115970 | 2726 |
| -1455 | 105635 | 888 | 52587 | 140864 | 8152 |
| -1379 | 106177 | 971 | 52587 | 143154 | 9321 |
| -1350 | 107068 | 1116 | 81204 | 142659 | 9037 |
| -1336 | 112813 | 2098 | 81204 | 143494 | 9526 |

## top10 high balance

| Obs | customer_id | balance |
|---|---|---|
| 1 | 142659 | 81204 |
| 2 | 143494 | 81204 |
| 3 | 140864 | 52587 |
| 4 | 143154 | 52587 |
| 5 | 115970 | 45789 |
| 6 | 100547 | 45248 |
| 7 | 102879 | 37378 |
| 8 | 141898 | 37378 |
| 9 | 134271 | 36252 |
| 10 | 120838 | 34646 |

## top10 low balance

| Obs | customer_id | balance |
|---|---|---|
| 1 | 132814 | -3058 |
| 2 | 120418 | -1980 |
| 3 | 135493 | -1944 |
| 4 | 119684 | -1781 |
| 5 | 115067 | -1668 |
| 6 | 118222 | -1598 |
| 7 | 105635 | -1455 |
| 8 | 106177 | -1379 |
| 9 | 107068 | -1350 |
| 10 | 112813 | -1336 |