

```
FILENAME project '/home/u63417899/BAN110ZBB/Netflix Userbase 2.xlsx';
```

```
PROC IMPORT DATAFILE= project  
  DBMS=XLSX  
  OUT=WORK.project3;  
  GETNAMES=YES;  
RUN;
```

```
PROC CONTENTS DATA=WORK.project3; RUN;
```

```
proc print data=WORK.project3 (obs=10);  
run;
```

```
proc contents data=WORK.project3;  
run;
```

```
data UserID;  
SET WORK.project3;  
user_ID2 = put (user_ID, 8.);  
run;
```

```
proc format;  
  value $missfmt ' ' = 'Missing' other = 'Not Missing';  
  value missfmt . = 'Missing' other = 'Not Missing';  
run;
```

```
proc freq data=WORK.project3;  
format _CHAR_ $missfmt.;  
tables _CHAR_ / missing missprint nocum nopercnt;  
run;
```

```
proc freq data=WORK.project3;  
tables country device gender plan_duration subscription_type / nocum nopercnt;  
run;
```

```
data _null_;  
set WORK.project3;  
file print;  
if device not in ('Laptop' 'Smart TV' 'Smartphone' 'Tablet' '') then put User_Id= device=;  
if gender not in ('Female' 'Male' '') then put User_Id= Gender=;  
if Plan_Duration not in ('1 Month' '') then put User_Id= Plan_duration=;  
run;
```

```
data WORK.project3;  
set WORK.project3;  
if country in (' ', '.') then country='United States';  
if Plan_Duration in (' ', '.') then plan_duration='1 Month';  
if subscription_type in (' ', '.') then Subscription_type='Basic';  
run;
```

```
proc freq data=WORK.project3;  
tables country plan_duration subscription_type / nocum nopercnt;  
run;
```

```
data WORK.project3;  
set WORK.project3;  
    if cmiss(of _all_) then delete;  
run;  
  
proc freq data=WORK.project3;  
tables country device gender plan_duration subscription_type/ nocum nopercnt;  
run;  
  
data WORK.project3;  
    set WORK.project3;  
    gender = Propcase(gender);  
run;  
  
data WORK.project3;  
set WORK.project3;  
if device in ('35') then Device='Laptop';  
if Plan_duration in ('1','0') then plan_duration='1 Month';  
run;  
  
data WORK.project3;  
set WORK.project3;  
if gender in ('M','Malee','Ma') then gender='Male';  
if gender in ('F','Fem')then gender='Female';  
run;  
  
proc freq data=WORK.project3;  
tables device plan_duration gender / nocum nopercnt;  
run;
```

The CONTENTS Procedure

Data Set Name	WORK.PROJECT3	Observations	2424
Member Type	DATA	Variables	10
Engine	V9	Indexes	0
Created	08/11/2023 22:08:55	Observation Length	88
Last Modified	08/11/2023 22:08:55	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	2
First Data Page	1
Max Obs per Page	1486
Obs in First Data Page	1450
Number of Data Set Repairs	0
Filename	/saswork/SAS_work0B5C00016AA7_odaws01-usw2.oda.sas.com/SAS_workA0ED00016AA7_odaws01-usw2.oda.sas.com/project3.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	536873136
Access Permission	rw-r--r--
Owner Name	u63417899
File Size	384KB
File Size (bytes)	393216

Alphabetic List of Variables and Attributes						
#	Variable	Type	Len	Format	Informat	Label
7	Age	Num	8	BEST.		Age
6	Country	Char	14	\$14.	\$14.	Country
9	Device	Char	10	\$10.	\$10.	Device
8	Gender	Char	6	\$6.	\$6.	Gender
4	Join_Date	Char	8	\$8.	\$8.	Join_Date
5	Last_Payment_Date	Char	8	\$8.	\$8.	Last_Payment_Date
3	Monthly_Revenue	Num	8	BEST.		Monthly_Revenue
10	Plan_Duration	Char	7	\$7.	\$7.	Plan_Duration
2	Subscription_Type	Char	8	\$8.	\$8.	Subscription_Type
1	User_ID	Num	8	BEST.		User_ID

Obs	User_ID	Subscription_Type	Monthly_Revenue	Join_Date	Last_Payment_Date	Country	Age	Gender	Device	Plan_Duration
1	1	Basic	10	15-01-22	45205	United States	28	Male	Smartphone	1 Month
2	2	Premium	15	44325	22-06-23	Canada	35	Female	Tablet	1 Month
3	3	Standard	12	28-02-23	27-06-23	United Kingdom	42	Male	Smart TV	1 Month

Obs	User_ID	Subscription_Type	Monthly_Revenue	Join_Date	Last_Payment_Date	Country	Age	Gender	Device	Plan_Duration
4	6	Premium	15	18-03-22	27-06-23	France	29	Female	Smart TV	1 Month
5	8	Basic	10	44961	24-06-23	Mexico	39	Female	Laptop	1 Month
6	9	Standard	12	20-10-22	23-06-23	Spain	37	Male	Smartphone	1 Month
7	10	Premium	15	45108	22-06-23	Italy	44	Female	Smart TV	1 Month
8	12	Premium	15	23-03-23	28-06-23	Canada	45	Male	Tablet	1 Month
9	13	Standard	12	30-11-21	27-06-23	United Kingdom	48	Female	Laptop	1 Month
10	15	Standard	12	45174	28-06-23	Germany	38	Female	Smart TV	1 Month

The CONTENTS Procedure

Data Set Name	WORK.PROJECT3	Observations	2424
Member Type	DATA	Variables	10
Engine	V9	Indexes	0
Created	08/11/2023 22:08:55	Observation Length	88
Last Modified	08/11/2023 22:08:55	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	2
First Data Page	1
Max Obs per Page	1486
Obs in First Data Page	1450
Number of Data Set Repairs	0
Filename	/saswork/SAS_work0B5C00016AA7_odaws01-usw2.oda.sas.com/SAS_workA0ED00016AA7_odaws01-usw2.oda.sas.com/project3.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	536873136
Access Permission	rw-r--r--
Owner Name	u63417899
File Size	384KB
File Size (bytes)	393216

Alphabetic List of Variables and Attributes						
#	Variable	Type	Len	Format	Informat	Label
7	Age	Num	8	BEST.		Age
6	Country	Char	14	\$14.	\$14.	Country
9	Device	Char	10	\$10.	\$10.	Device
8	Gender	Char	6	\$6.	\$6.	Gender
4	Join_Date	Char	8	\$8.	\$8.	Join_Date
5	Last_Payment_Date	Char	8	\$8.	\$8.	Last_Payment_Date
3	Monthly_Revenue	Num	8	BEST.		Monthly_Revenue
10	Plan_Duration	Char	7	\$7.	\$7.	Plan_Duration

Alphabetic List of Variables and Attributes						
#	Variable	Type	Len	Format	Informat	Label
2	Subscription_Type	Char	8	\$8.	\$8.	Subscription_Type
1	User_ID	Num	8	BEST.		User_ID

---

#### The FREQ Procedure

Subscription_Type	
Subscription_Type	Frequency
Not Missing	2424

Join_Date	
Join_Date	Frequency
Not Missing	2424

Last_Payment_Date	
Last_Payment_Date	Frequency
Not Missing	2424

Country	
Country	Frequency
Not Missing	2424

Gender	
Gender	Frequency
Not Missing	2424

Device	
Device	Frequency
Not Missing	2424

Plan_Duration	
Plan_Duration	Frequency
Not Missing	2424

---

#### The FREQ Procedure

Country	
Country	Frequency
Australia	176
Brazil	175
Canada	311
France	179
Germany	175
Italy	176
Mexico	174

Country	
Country	Frequency
Spain	435
United Kingdom	178
United States	445

Device	
Device	Frequency
Laptop	623
Smart TV	585
Smartphone	598
Tablet	618

Gender	
Gender	Frequency
Female	1220
Male	1204

Plan_Duration	
Plan_Duration	Frequency
1 Month	2424

Subscription_Type	
Subscription_Type	Frequency
Basic	975
Premium	706
Standard	743

---

#### The FREQ Procedure

Country	
Country	Frequency
Australia	176
Brazil	175
Canada	311
France	179
Germany	175
Italy	176
Mexico	174
Spain	435
United Kingdom	178
United States	445

Plan_Duration	
Plan_Duration	Frequency

Plan_Duration	
Plan_Duration	Frequency
1 Month	2424

Subscription_Type	
Subscription_Type	Frequency
Basic	975
Premium	706
Standard	743

---

### The FREQ Procedure

Country	
Country	Frequency
Australia	176
Brazil	175
Canada	311
France	179
Germany	175
Italy	176
Mexico	174
Spain	435
United Kingdom	178
United States	445

Device	
Device	Frequency
Laptop	623
Smart TV	585
Smartphone	598
Tablet	618

Gender	
Gender	Frequency
Female	1220
Male	1204

Plan_Duration	
Plan_Duration	Frequency
1 Month	2424

Subscription_Type	
Subscription_Type	Frequency
Basic	975
Premium	706
Standard	743

**The FREQ Procedure**

Device	
Device	Frequency
Laptop	623
Smart TV	585
Smartphone	598
Tablet	618

Plan_Duration	
Plan_Duration	Frequency
1 Month	2424

Gender	
Gender	Frequency
Female	1220
Male	1204



```

1      OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69      FILENAME project '/home/u63417899/BAN110ZBB/Netflix Userbase 2.xlsx';
70
71      PROC IMPORT DATAFILE= project
72      DBMS=XLSX
73      OUT=WORK.project3;
74      GETNAMES=YES;
75      RUN;

```

NOTE: Import cancelled. Output dataset WORK.PROJECT3 already exists. Specify REPLACE option to overwrite it.

NOTE: The SAS System stopped processing this step because of errors.

NOTE: PROCEDURE IMPORT used (Total process time):

real time	0.00 seconds
user cpu time	0.00 seconds
system cpu time	0.00 seconds
memory	910.65k
OS Memory	27540.00k
Timestamp	08/12/2023 02:10:18 AM
Step Count	673 Switch Count 0
Page Faults	0
Page Reclaims	139
Page Swaps	0
Voluntary Context Switches	0
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	0

76

```

77      PROC CONTENTS DATA=WORK.project3; RUN;

```

NOTE: PROCEDURE CONTENTS used (Total process time):

real time	0.04 seconds
user cpu time	0.04 seconds
system cpu time	0.00 seconds
memory	2114.15k
OS Memory	27564.00k
Timestamp	08/12/2023 02:10:18 AM
Step Count	674 Switch Count 0
Page Faults	0
Page Reclaims	93
Page Swaps	0
Voluntary Context Switches	0
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	16

78

79

```

80      proc print data=WORK.project3 (obs=10);
81      run;

```

NOTE: There were 10 observations read from the data set WORK.PROJECT3.

NOTE: PROCEDURE PRINT used (Total process time):

real time	0.02 seconds
user cpu time	0.03 seconds
system cpu time	0.00 seconds
memory	999.09k
OS Memory	27304.00k
Timestamp	08/12/2023 02:10:18 AM
Step Count	675 Switch Count 0
Page Faults	0
Page Reclaims	63
Page Swaps	0
Voluntary Context Switches	0

Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	24

```

82
83      proc contents data=WORK.project3;
84      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	1182.09k
OS Memory	27564.00k
Timestamp	08/12/2023 02:10:18 AM
Step Count	676 Switch Count 0
Page Faults	0
Page Reclaims	93
Page Swaps	0
Voluntary Context Switches	0
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	24

```

85
86      data UserID;
87      SET WORK.project3;
88      user_ID2 = put (user_ID, 8.);
89      run;

```

NOTE: There were 2424 observations read from the data set WORK.PROJECT3.

NOTE: The data set WORK.USERID has 2424 observations and 11 variables.

NOTE: DATA statement used (Total process time):

real time	0.00 seconds
user cpu time	0.00 seconds
system cpu time	0.00 seconds
memory	1319.84k
OS Memory	27564.00k
Timestamp	08/12/2023 02:10:18 AM
Step Count	677 Switch Count 2
Page Faults	0
Page Reclaims	101
Page Swaps	0
Voluntary Context Switches	11
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	520

```

90
91
92      proc format;
93      value $missfmt ' '= 'Missing' other= 'Not Missing';
NOTE: Format $MISSFMT is already on the library WORK.FORMATS.
NOTE: Format $MISSFMT has been output.
94      value missfmt . = 'Missing' other= 'Not Missing';
NOTE: Format MISSFMT is already on the library WORK.FORMATS.
NOTE: Format MISSFMT has been output.
95      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	354.34k
OS Memory	27044.00k
Timestamp	08/12/2023 02:10:18 AM
Step Count	678 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	32

```

96
97     proc freq data=WORK.project3;
98     format _CHAR_ $missfmt.;
99     tables _CHAR_ / missing missprint nocum nopercent;
100    run;

```

NOTE: There were 2424 observations read from the data set WORK.PROJECT3.

NOTE: PROCEDURE FREQ used (Total process time):

real time	0.04 seconds
user cpu time	0.04 seconds
system cpu time	0.00 seconds
memory	1047.21k
OS Memory	27564.00k
Timestamp	08/12/2023 02:10:18 AM
Step Count	679 Switch Count 2
Page Faults	0
Page Reclaims	141
Page Swaps	0
Voluntary Context Switches	15
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	264

```

101
102     proc freq data=WORK.project3;
103     tables country device gender plan_duration subscription_type/ nocum nopercent;
104    run;

```

NOTE: There were 2424 observations read from the data set WORK.PROJECT3.

NOTE: PROCEDURE FREQ used (Total process time):

real time	0.03 seconds
user cpu time	0.04 seconds
system cpu time	0.00 seconds
memory	1129.84k
OS Memory	27564.00k
Timestamp	08/12/2023 02:10:18 AM
Step Count	680 Switch Count 2
Page Faults	0
Page Reclaims	129
Page Swaps	0
Voluntary Context Switches	12
Involuntary Context Switches	1
Block Input Operations	0
Block Output Operations	280

```

105
106     data _null_;
107     set WORK.project3;
108     file print;
109     if device not in ('Laptop' 'Smart TV' 'Smartphone' 'Tablet' '') then put User_Id= device=;
110     if gender not in ('Female' 'Male' '') then put User_Id= Gender=;
111     if Plan_Duration not in ('1 Month' '') then put User_Id= Plan_duration=;
112    run;

```

NOTE: 0 lines were written to file PRINT.

NOTE: There were 2424 observations read from the data set WORK.PROJECT3.

NOTE: DATA statement used (Total process time):

real time	0.00 seconds
user cpu time	0.00 seconds

```

system cpu time    0.01 seconds
memory            892.68k
OS Memory         27304.00k
Timestamp         08/12/2023 02:10:18 AM
Step Count        681  Switch Count  0
Page Faults       0
Page Reclaims     67
Page Swaps        0
Voluntary Context Switches  1
Involuntary Context Switches 0
Block Input Operations  0
Block Output Operations  0

```

```

113
114     data WORK.project3;
115     set WORK.project3;
116     if country in (' ', '.') then country='United States';
117     if Plan_Duration in (' ', '.') then plan_duration='1 Month';
118     if subscription_type in (' ', '.') then Subscription_type='Basic';
119     run;

```

NOTE: There were 2424 observations read from the data set WORK.PROJECT3.

NOTE: The data set WORK.PROJECT3 has 2424 observations and 10 variables.

NOTE: DATA statement used (Total process time):

```

real time          0.00 seconds
user cpu time      0.00 seconds
system cpu time    0.00 seconds
memory            1321.81k
OS Memory         27564.00k
Timestamp         08/12/2023 02:10:18 AM
Step Count        682  Switch Count  2
Page Faults       0
Page Reclaims     97
Page Swaps        0
Voluntary Context Switches  12
Involuntary Context Switches 0
Block Input Operations  0
Block Output Operations  520

```

```

120
121     proc freq data=WORK.project3;
122     tables country plan_duration subscription_type/ nocum nopercnt;
123     run;

```

NOTE: There were 2424 observations read from the data set WORK.PROJECT3.

NOTE: PROCEDURE FREQ used (Total process time):

```

real time          0.02 seconds
user cpu time      0.03 seconds
system cpu time    0.00 seconds
memory            1159.12k
OS Memory         27564.00k
Timestamp         08/12/2023 02:10:18 AM
Step Count        683  Switch Count  2
Page Faults       0
Page Reclaims     129
Page Swaps        0
Voluntary Context Switches  19
Involuntary Context Switches 0
Block Input Operations  0
Block Output Operations  264

```

```

124
125     data WORK.project3;
126     set WORK.project3;
127     if cmiss(of _all_) then delete;
128     run;

```

NOTE: There were 2424 observations read from the data set WORK.PROJECT3.

NOTE: The data set WORK.PROJECT3 has 2424 observations and 10 variables.

NOTE: DATA statement used (Total process time):

real time	0.00 seconds
user cpu time	0.00 seconds
system cpu time	0.00 seconds
memory	1318.75k
OS Memory	27564.00k
Timestamp	08/12/2023 02:10:18 AM
Step Count	684 Switch Count 2
Page Faults	0
Page Reclaims	98
Page Swaps	0
Voluntary Context Switches	13
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	520

129

```
130      proc freq data=WORK.project3;
131      tables country device gender plan_duration subscription_type/ nocum nopercnt;
132      run;
```

NOTE: There were 2424 observations read from the data set WORK.PROJECT3.

NOTE: PROCEDURE FREQ used (Total process time):

real time	0.03 seconds
user cpu time	0.04 seconds
system cpu time	0.00 seconds
memory	1129.84k
OS Memory	27564.00k
Timestamp	08/12/2023 02:10:18 AM
Step Count	685 Switch Count 2
Page Faults	0
Page Reclaims	132
Page Swaps	0
Voluntary Context Switches	12
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	280

133

```
134      data WORK.project3;
135      set WORK.project3;
136      gender = Propcase(gender);
137      run;
```

NOTE: There were 2424 observations read from the data set WORK.PROJECT3.

NOTE: The data set WORK.PROJECT3 has 2424 observations and 10 variables.

NOTE: DATA statement used (Total process time):

real time	0.00 seconds
user cpu time	0.01 seconds
system cpu time	0.00 seconds
memory	1206.12k
OS Memory	27564.00k
Timestamp	08/12/2023 02:10:18 AM
Step Count	686 Switch Count 2
Page Faults	0
Page Reclaims	100
Page Swaps	0
Voluntary Context Switches	13
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	520

138

```
139      data WORK.project3;
140      set WORK.project3;
```

```

141      if device in ('35') then Device='Laptop';
142      if Plan_duration in ('1','0') then plan_duration='1 Month';
143      run;

```

NOTE: There were 2424 observations read from the data set WORK.PROJECT3.

NOTE: The data set WORK.PROJECT3 has 2424 observations and 10 variables.

NOTE: DATA statement used (Total process time):

real time	0.00 seconds
user cpu time	0.00 seconds
system cpu time	0.00 seconds
memory	1318.46k
OS Memory	27564.00k
Timestamp	08/12/2023 02:10:18 AM
Step Count	687 Switch Count 2
Page Faults	0
Page Reclaims	100
Page Swaps	0
Voluntary Context Switches	13
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	520

```

144
145      data WORK.project3;
146      set WORK.project3;
147      if gender in ('M','Malee','Ma') then gender='Male';
148      if gender in ('F','Fem')then gender='Female';
149      run;

```

NOTE: There were 2424 observations read from the data set WORK.PROJECT3.

NOTE: The data set WORK.PROJECT3 has 2424 observations and 10 variables.

NOTE: DATA statement used (Total process time):

real time	0.00 seconds
user cpu time	0.00 seconds
system cpu time	0.00 seconds
memory	1206.59k
OS Memory	27564.00k
Timestamp	08/12/2023 02:10:18 AM
Step Count	688 Switch Count 2
Page Faults	0
Page Reclaims	101
Page Swaps	0
Voluntary Context Switches	17
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	520

```

150
151      proc freq data=WORK.project3;
152      tables device plan_duration gender / nocum nopercnt;
153      run;

```

NOTE: There were 2424 observations read from the data set WORK.PROJECT3.

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	1161.50k
OS Memory	27564.00k
Timestamp	08/12/2023 02:10:18 AM
Step Count	689 Switch Count 2
Page Faults	0
Page Reclaims	129
Page Swaps	0
Voluntary Context Switches	12
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	264

```
154  
155     OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;  
165
```

```

FILENAME project '/home/u63417899/BAN110ZBB/Netflix Userbase 1.xlsx';

PROC IMPORT DATAFILE=project
  DBMS=XLSX
  replace
  OUT=WORK.project2;
  GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.project2; RUN;

proc print data=WORK.project2(obs=10);
run;

/* Using Nmiss to ascertain the amount of missing values in each variable */
Title 'Means output before removing missing values';
proc means data=work.project2 N Nmiss Mean Min Max Maxdec=3;
run;

/*Replacing missing variables with its mean*/
proc standard data=work.project2 replace out=work.project2_nomiss;
var Age monthly_revenue;
run;

Title 'Means output after removing missing values';
proc means data=work.project2_nomiss N Nmiss Mean Min Max Maxdec=3;
run;*/

/* OUTLIERS */
/* Outlier detection for some variables using box plots */;
proc sgplot data=work.project2;
vbox monthly_revenue;
run;

/*Histogram of the variable monthly_revenue*/
proc sgplot data=work.project2;
  histogram Monthly_Revenue;
  density Monthly_Revenue;
run;

title "outliers affect";
proc sgplot data=work.project2;
  scatter x=country y=monthly_revenue;
  xaxis grid;
  yaxis grid;
run;

Title 'Q-Q plot';
proc univariate data = work.project2;
ppplot monthly_revenue;
run;

/*Check Distribution and Choose Outlier Detection Method*/;
proc univariate data=work.project2;
  var monthly_revenue;
  histogram monthly_revenue / normal; /* Visualize distribution */
run;

title "Output the outliers for the variable monthly_revenue based on the interquantile range method";
proc means data=work.project2 noprint;
  var monthly_revenue;
  output out=revenue
    Q1=
    Q3=
    QRange= / autoname;
run;

data null;
  file print;
  set work.project2(keep=monthly_revenue);
  if n = 1 then set revenue;
  if monthly_revenue le monthly_revenue_Q1 - 1.5*monthly_revenue_QRange and not missing(monthly_revenue) or
    monthly_revenue ge monthly_revenue_Q3 + 1.5*monthly_revenue_QRange then
    put "Possible Outlier for instant " instant "Value of monthly_revenue is " monthly_revenue;
run;

```



```

title 'Dataset without the outliers';
/* Step 1: Calculate the interquartile range (IQR) for the variable "monthly_revenue" */
proc means data=work.project2 noprint;
  var monthly_revenue;
  output out=revenue
    Q1=monthly_revenue_Q1
    Q3=monthly_revenue_Q3
    QRange=monthly_revenue_QRange / autoname;
run;

/* Step 2: Identify and remove the outliers */
data revenue_OutliersRemoved;
  set work.project2;
  if _n_ = 1 then set revenue; /* Read the IQR values from the temporary dataset */

  /* Check if the value of 'monthly_revenue' is within the acceptable range (Q1 - 1.5*IQR, Q3 + 1.5*IQR) */
  if monthly_revenue >= monthly_revenue_Q1 - 1.5 * monthly_revenue_QRange and monthly_revenue <= monthly_revenue_Q3 + 1.5 *
run;

/* Step 3: Print the dataset without the outliers */
proc print data=revenue_OutliersRemoved (obs=30);
run;

/*Check to ensure outliers were removed using the histogram plot*/
/*Histogram without outliers*/
title "Histogram (with outliers)";
proc sgplot data=work.project2;
  histogram Monthly_Revenue;
  density Monthly_Revenue / type=normal;
run;

/*Histogram without outliers*/
title "Histogram (without outliers)";
proc sgplot data=revenue_OutliersRemoved;
  histogram Monthly_Revenue;
  density Monthly_Revenue / type=normal;
run;

/*Check to ensure outliers were removed using the Q-Q plot*/
title "Box plot from revenue dataset";
proc sgplot data=revenue_OutliersRemoved;
  vbox monthly_revenue;
run;

Title 'Q-Q plot with outliers';
proc univariate data = work.project2;
  ppplot monthly_revenue;
run;

Title 'Q-Q plot without outliers';
proc univariate data = revenue_OutliersRemoved;
  ppplot monthly_revenue;
run;

/*
/*AGE OUTLIERS */
/* Outlier detection for some variables using box plots */;
proc sgplot data=work.project2;
  vbox age;
run;

/*Histogram of the variable age*/
proc sgplot data=work.project2;
  histogram age;
  density age;
run;

Title 'Q-Q plot';
proc univariate data = work.project2;
  ppplot age;
run;

/*Check Distribution and Choose Outlier Detection Method*/;
proc univariate data=work.project2;

```

```

var age;
histogram age / normal; /* Visualize distribution */
run;

title "Output the outliers for the variable monthly_revenue based on the interquantile range method";
proc means data=work.project2 noprint;
var age;
output out=Age_range
      Q1=
      Q3=
      QRange= / autoname;
run;

data null;
file print;
set work.project2(keep=monthly_revenue);
if n = 1 then set age_range;
if age le age_Q1 - 1.5*age_QRange and not missing(age) or
   age ge age_Q3 + 1.5*age_QRange then
  put "Possible Outlier for instant " instant "Value of monthly_revenue is " age;
run;

title 'Dataset without the outliers';
/* Step 1: Calculate the interquartile range (IQR) for the variable age */
proc means data=work.project2 noprint;
var age;
output out=age_range
      Q1=age_Q1
      Q3=age_Q3
      QRange=age_QRange / autoname;
run;

/* Step 2: Identify and remove the outliers */
data age_range_OutliersRemoved;
set work.project2;
if _n_ = 1 then set age_range; /* Read the IQR values from the temporary dataset */

/* Check if the value of 'age' is within the acceptable range (Q1 - 1.5*IQR, Q3 + 1.5*IQR) */
if age >= age_Q1 - 1.5 * age_QRange and age <= age_Q3 + 1.5 * age_QRange and not missing(age);
run;

/* Step 3: Print the dataset without the outliers */
proc print data=age_range_OutliersRemoved (obs=30);
run;

/*Check to ensure outliers were removed using the histogram plot*/
/*Histogram without outliers*/
title "Histogram (with outliers)";
proc sgplot data=work.project2;
  histogram age;
  density age / type=normal;
run;

/*Histogram without outliers*/
title "Histogram (without outliers)";
proc sgplot data=age_range_OutliersRemoved;
  histogram age;
  density age / type=normal;
run;

/*Check to ensure outliers were removed using the Q-Q plot*/
title "Box plot from revenue dataset";
proc sgplot data=age_range_OutliersRemoved;
  vbox age;
run;

Title 'Q-Q plot with outliers';
proc univariate data = work.project2;
ppplot age;
run;

Title 'Q-Q plot without outliers';
proc univariate data = age_range_OutliersRemoved;
ppplot age;
run;

```



The CONTENTS Procedure

Data Set Name	WORK.PROJECT2	Observations	2500
Member Type	DATA	Variables	9
Engine	V9	Indexes	0
Created	08/11/2023 22:16:18	Observation Length	80
Last Modified	08/11/2023 22:16:18	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	2
First Data Page	1
Max Obs per Page	1635
Obs in First Data Page	1596
Number of Data Set Repairs	0
Filename	/saswork/SAS_work0B5C00016AA7_odaws01-usw2.oda.sas.com/SAS_workA0ED00016AA7_odaws01-usw2.oda.sas.com/project2.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	536873122
Access Permission	rw-r--r--
Owner Name	u63417899
File Size	384KB
File Size (bytes)	393216

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
6	Age	Num	8	BEST.		Age
5	Country	Char	14	\$14.	\$14.	Country
8	Device	Char	10	\$10.	\$10.	Device
7	Gender	Char	6	\$6.	\$6.	Gender
3	Join_Date	Char	8	\$8.	\$8.	Join_Date
4	Last_Payment_Date	Char	8	\$8.	\$8.	Last_Payment_Date
2	Monthly_Revenue	Num	8	BEST.		Monthly_Revenue
9	Plan_Duration	Char	7	\$7.	\$7.	Plan_Duration
1	Subscription_Type	Char	8	\$8.	\$8.	Subscription_Type

Obs	Subscription_Type	Monthly_Revenue	Join_Date	Last_Payment_Date	Country	Age	Gender	Device	Plan_Duration
1	Basic	10	15-01-22	45205	United States	28	Male	Smartphone	1 Month
2	Premium	15	44325	22-06-23	Canada	35	Female	Tablet	1 Month
3	Standard	12	28-02-23	27-06-23	United Kingdom	42	Male	Smart TV	1 Month
4	Standard	12	44841	26-06-23	Australia	51	Female	Laptop	1 Month
5	Basic	.	44931	28-06-23	Germany	33	Male	Smartphone	1 Month
6	Premium	15	18-03-22	27-06-23	France	29	Female	Smart TV	1 Month
7	Standard	.	44451	25-06-23	Brazil	46	Male	Tablet	1 Month
8	Basic	10	44961	24-06-23	Mexico	39	Female	Laptop	1 Month
9	Standard	12	20-10-22	23-06-23	Spain	37	Male	Smartphone	1 Month
10	Premium	15	45108	22-06-23	Italy	44	Female	Smart TV	1 Month

Means output before removing missing values

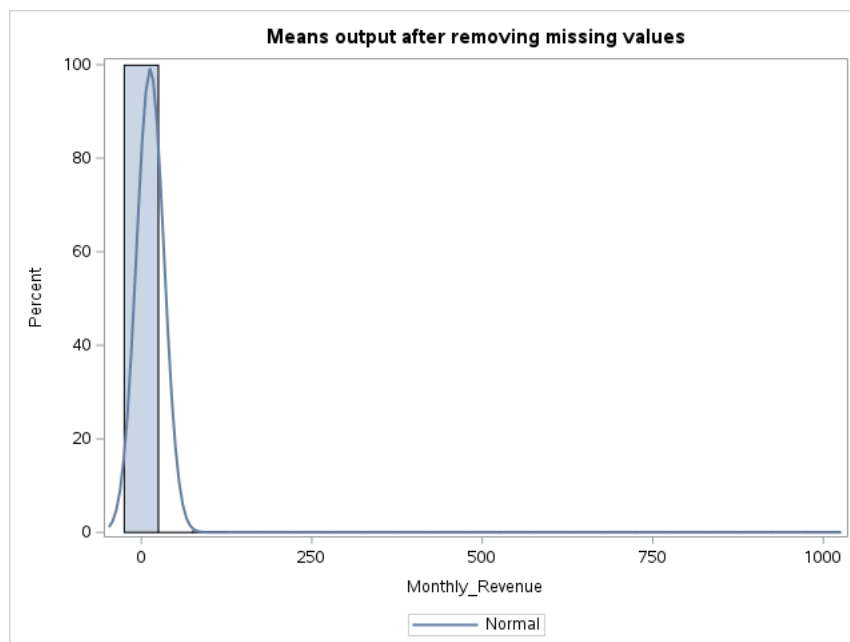
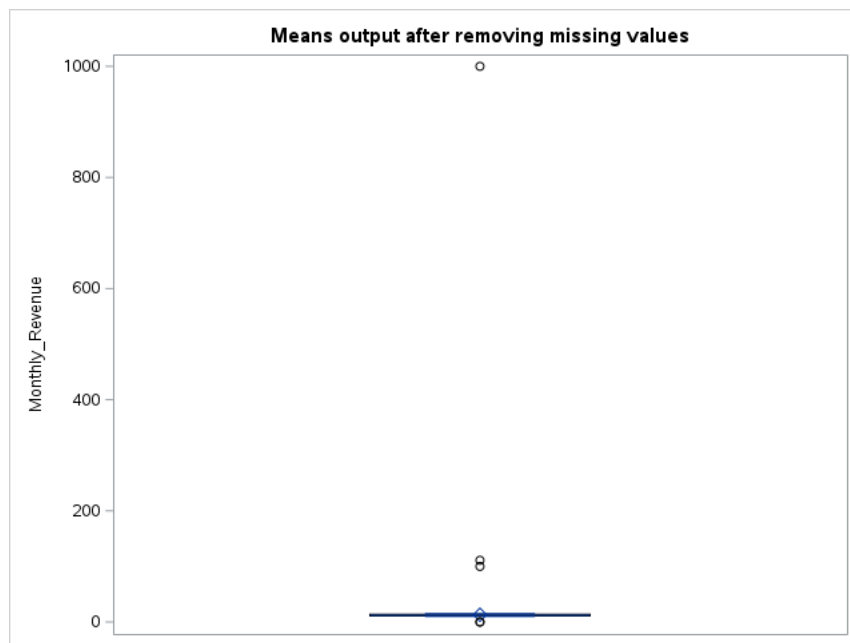
The MEANS Procedure

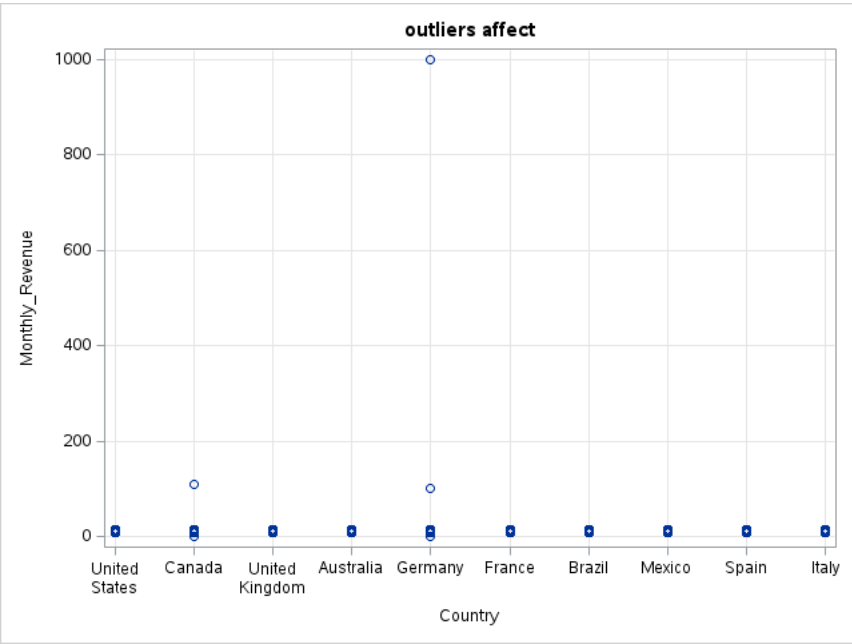
Variable	Label	N	N Miss	Mean	Minimum	Maximum
Monthly_Revenue	Monthly_Revenue	2462	38	12.976	0.000	1000.000
Age	Age	2500	0	38.993	0.000	300.000

Means output after removing missing values

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Minimum	Maximum
Monthly_Revenue	Monthly_Revenue	2500	0	12.976	0.000	1000.000
Age	Age	2500	0	38.993	0.000	300.000





Q-Q plot

The UNIVARIATE Procedure  
Variable: Monthly\_Revenue (Monthly\_Revenue)

Moments			
N	2462	Sum Weights	2462
Mean	12.9756296	Sum Observations	31946
Std Deviation	20.1501529	Variance	406.028662
Skewness	47.874674	Kurtosis	2342.37983
Uncorrected SS	1413756	Corrected SS	999236.538
Coeff Variation	155.292295	Std Error Mean	0.40610125

Basic Statistical Measures			
Location		Variability	
Mean	12.97563	Std Deviation	20.15015
Median	12.00000	Variance	406.02866
Mode	12.00000	Range	1000
		Interquartile Range	3.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	31.95171	Pr >  t	<.0001
Sign	M	1230	Pr >=  M	<.0001
Signed Rank	S	1513515	Pr >=  S	<.0001

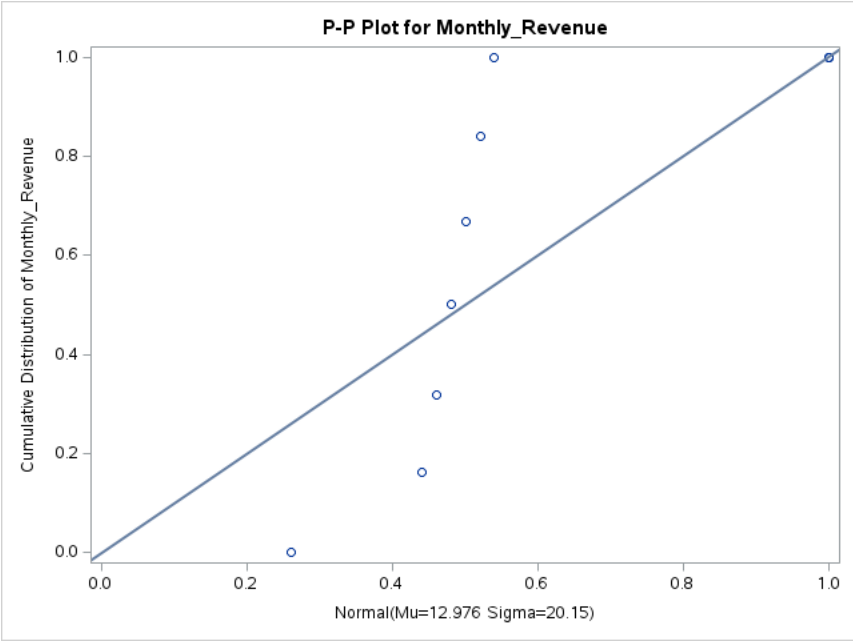
Quantiles (Definition 5)	
Level	Quantile
100% Max	1000
99%	15
95%	15
90%	15
75% Q3	14
50% Median	12
25% Q1	11
10%	10
5%	10
1%	10
0% Min	0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	780	15	2497
0	717	15	2500
10	2490	100	335
10	2486	111	372
10	2481	1000	2310

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	38	1.52	100.00

Q-Q plot

The UNIVARIATE Procedure



Q-Q plot

The UNIVARIATE Procedure  
Variable: Age (Age)

Moments			
N	2500	Sum Weights	2500
Mean	38.9928	Sum Observations	97482
Std Deviation	10.1222592	Variance	102.460132
Skewness	12.2308609	Kurtosis	303.775967
Uncorrected SS	4057144	Corrected SS	256047.87
Coeff Variation	25.9593034	Std Error Mean	0.20244518

Basic Statistical Measures			
Location		Variability	
Mean	38.99280	Std Deviation	10.12226
Median	39.00000	Variance	102.46013
Mode	30.00000	Range	300.00000
		Interquartile Range	13.00000

Note: The mode displayed is the smallest of 2 modes with a count of 116.

Tests for Location: Mu0=0			
Test	Statistic		p Value
Student's t	t	192.6092	Pr >  t  <.0001
Sign	M	1249.5	Pr >=  M  <.0001
Signed Rank	S	1561875	Pr >=  S  <.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	300
99%	51
95%	50
90%	49
75% Q3	45
50% Median	39
25% Q1	32
10%	29
5%	28
1%	27
0% Min	0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	796	51	2426
26	21	51	2449
27	2472	51	2479
27	2447	280	745
27	2412	300	727

Q-Q plot

The UNIVARIATE Procedure  
Variable: Monthly\_Revenue (Monthly\_Revenue)

Moments			
N	2462	Sum Weights	2462
Mean	12.9756296	Sum Observations	31946
Std Deviation	20.1501529	Variance	406.028662
Skewness	47.874674	Kurtosis	2342.37983
Uncorrected SS	1413756	Corrected SS	999236.538
Coeff Variation	155.292295	Std Error Mean	0.40610125

Basic Statistical Measures			
Location		Variability	
Mean	12.97563	Std Deviation	20.15015
Median	12.00000	Variance	406.02866
Mode	12.00000	Range	1000
		Interquartile Range	3.00000

Tests for Location: Mu0=0			
Test	Statistic		p Value
Student's t	t	31.95171	Pr >  t  <.0001
Sign	M	1230	Pr >=  M  <.0001
Signed Rank	S	1513515	Pr >=  S  <.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	1000
99%	15
95%	15
90%	15
75% Q3	14
50% Median	12
25% Q1	11
10%	10
5%	10
1%	10
0% Min	0

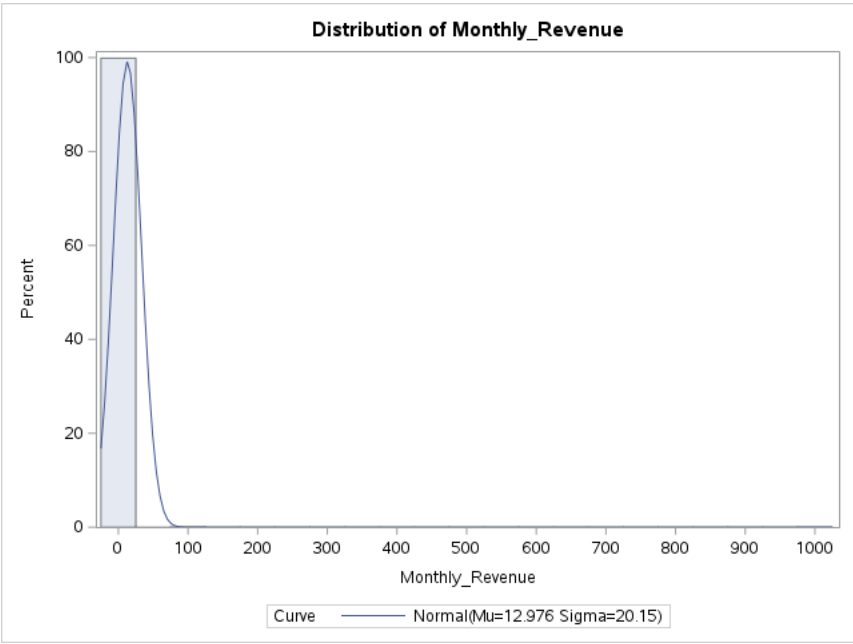
Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	780	15	2497
0	717	15	2500
10	2490	100	335
10	2486	111	372
10	2481	1000	2310

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	38	1.52	100.00

Q-Q plot

The UNIVARIATE Procedure





**Q-Q plot**

The UNIVARIATE Procedure  
Fitted Normal Distribution for Monthly\_Revenue (Monthly\_Revenue)

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	12.97563
Std Dev	Sigma	20.15015

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.458769	Pr > D	<0.010
Cramer-von Mises	W-Sq	160.554749	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	771.367922	Pr > A-Sq	<0.005

Quantiles for Normal Distribution		
Percent	Quantile	
	Observed	Estimated
1.0	10.00000	-33.90064
5.0	10.00000	-20.16842
10.0	10.00000	-12.84783
25.0	11.00000	-0.61544
50.0	12.00000	12.97563
75.0	14.00000	26.56670
90.0	15.00000	38.79909
95.0	15.00000	46.11968
99.0	15.00000	59.85189

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

```
Possible Outlier for instant . Value of monthly_revenue is 10
Possible Outlier for instant . Value of monthly_revenue is 15
Possible Outlier for instant . Value of monthly_revenue is 12
Possible Outlier for instant . Value of monthly_revenue is 12
Possible Outlier for instant . Value of monthly_revenue is .
Possible Outlier for instant . Value of monthly_revenue is 15
Possible Outlier for instant . Value of monthly_revenue is .
Possible Outlier for instant . Value of monthly_revenue is 10
Possible Outlier for instant . Value of monthly_revenue is 12
Possible Outlier for instant . Value of monthly_revenue is 15
Possible Outlier for instant . Value of monthly_revenue is 10
Possible Outlier for instant . Value of monthly_revenue is 15
Possible Outlier for instant . Value of monthly_revenue is 12
Possible Outlier for instant . Value of monthly_revenue is .
Possible Outlier for instant . Value of monthly_revenue is 12
Possible Outlier for instant . Value of monthly_revenue is 15
Possible Outlier for instant . Value of monthly_revenue is 10
Possible Outlier for instant . Value of monthly_revenue is 12
Possible Outlier for instant . Value of monthly_revenue is 15
Possible Outlier for instant . Value of monthly_revenue is 10
Possible Outlier for instant . Value of monthly_revenue is 15
Possible Outlier for instant . Value of monthly_revenue is 12
Possible Outlier for instant . Value of monthly_revenue is .
Possible Outlier for instant . Value of monthly_revenue is 12
Possible Outlier for instant . Value of monthly_revenue is 15
Possible Outlier for instant . Value of monthly_revenue is 10
Possible Outlier for instant . Value of monthly_revenue is 12
Possible Outlier for instant . Value of monthly_revenue is 15
Possible Outlier for instant . Value of monthly_revenue is 10
Possible Outlier for instant . Value of monthly_revenue is 15
Possible Outlier for instant . Value of monthly_revenue is .
Possible Outlier for instant . Value of monthly_revenue is 12
Possible Outlier for instant . Value of monthly_revenue is 12
Possible Outlier for instant . Value of monthly_revenue is 10
Possible Outlier for instant . Value of monthly_revenue is 15
Possible Outlier for instant . Value of monthly_revenue is 10
```

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

7/74

Output the outliers for the variable `monthly_revenue` based on the interquantile range method

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

8/74

Output the outliers for the variable `monthly_revenue` based on the interquartile range method

Output the outliers for the variable monthly\_revenue based on the interquartile range method

9/74

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

10/74

Output the outliers for the variable `monthly_revenue` based on the interquantile range method

[illegible]

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

Output the outliers for the variable `monthly_revenue` based on the interquantile range method

12/74

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

13/74



**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

14/74

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

15/74

Output the outliers for the variable `monthly_revenue` based on the interquartile range method

Output the outliers for the variable monthly\_revenue based on the interquartile range method

16/74

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

Output the outliers for the variable `monthly_revenue` based on the interquantile range method

17/74

Output the outliers for the variable `monthly_revenue` based on the interquartile range method

18/74

[illegible][illegible]

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

20/74

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

21/74



**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

22/74

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

23/74

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

Output the outliers for the variable `monthly_revenue` based on the interquantile range method

24/74

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

25/74

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

26/74

[illegible][illegible]

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

28/74

```
Possible Outlier for instant . Value of monthly_revenue is 14
Possible Outlier for instant . Value of monthly_revenue is 12
Possible Outlier for instant . Value of monthly_revenue is 15
Possible Outlier for instant . Value of monthly_revenue is 15
Possible Outlier for instant . Value of monthly_revenue is 12
Possible Outlier for instant . Value of monthly_revenue is 11
Possible Outlier for instant . Value of monthly_revenue is 10
Possible Outlier for instant . Value of monthly_revenue is 12
Possible Outlier for instant . Value of monthly_revenue is 11
Possible Outlier for instant . Value of monthly_revenue is 10
Possible Outlier for instant . Value of monthly_revenue is 11
Possible Outlier for instant . Value of monthly_revenue is 14
Possible Outlier for instant . Value of monthly_revenue is 14
```

Output the outliers for the variable monthly\_revenue based on the interquantile range method

```
Possible Outlier for instant . Value of monthly_revenue is 11
Possible Outlier for instant . Value of monthly_revenue is 12
Possible Outlier for instant . Value of monthly_revenue is 10
Possible Outlier for instant . Value of monthly_revenue is 14
Possible Outlier for instant . Value of monthly_revenue is 10
Possible Outlier for instant . Value of monthly_revenue is 11
Possible Outlier for instant . Value of monthly_revenue is 13
Possible Outlier for instant . Value of monthly_revenue is 15
Possible Outlier for instant . Value of monthly_revenue is 15
Possible Outlier for instant . Value of monthly_revenue is 14
Possible Outlier for instant . Value of monthly_revenue is 11
Possible Outlier for instant . Value of monthly_revenue is 14
Possible Outlier for instant . Value of monthly_revenue is 13
Possible Outlier for instant . Value of monthly_revenue is 10
Possible Outlier for instant . Value of monthly_revenue is 13
Possible Outlier for instant . Value of monthly_revenue is 15
Possible Outlier for instant . Value of monthly_revenue is 14
Possible Outlier for instant . Value of monthly_revenue is 15
Possible Outlier for instant . Value of monthly_revenue is 10
Possible Outlier for instant . Value of monthly_revenue is 10
Possible Outlier for instant . Value of monthly_revenue is 13
Possible Outlier for instant . Value of monthly_revenue is 11
Possible Outlier for instant . Value of monthly_revenue is 12
Possible Outlier for instant . Value of monthly_revenue is 12
Possible Outlier for instant . Value of monthly_revenue is 15
Possible Outlier for instant . Value of monthly_revenue is 13
Possible Outlier for instant . Value of monthly_revenue is 11
Possible Outlier for instant . Value of monthly_revenue is 10
Possible Outlier for instant . Value of monthly_revenue is 12
Possible Outlier for instant . Value of monthly_revenue is 15
Possible Outlier for instant . Value of monthly_revenue is 12
Possible Outlier for instant . Value of monthly_revenue is 15
Possible Outlier for instant . Value of monthly_revenue is 14
Possible Outlier for instant . Value of monthly_revenue is 12
Possible Outlier for instant . Value of monthly_revenue is 13
Possible Outlier for instant . Value of monthly_revenue is 10
Possible Outlier for instant . Value of monthly_revenue is 14
Possible Outlier for instant . Value of monthly_revenue is .
Possible Outlier for instant . Value of monthly_revenue is 10
Possible Outlier for instant . Value of monthly_revenue is 12
Possible Outlier for instant . Value of monthly_revenue is 13
Possible Outlier for instant . Value of monthly_revenue is 12
Possible Outlier for instant . Value of monthly_revenue is 10
Possible Outlier for instant . Value of monthly_revenue is 11
Possible Outlier for instant . Value of monthly_revenue is 13
Possible Outlier for instant . Value of monthly_revenue is 10
Possible Outlier for instant . Value of monthly_revenue is 13
Possible Outlier for instant . Value of monthly_revenue is 15
Possible Outlier for instant . Value of monthly_revenue is 11
Possible Outlier for instant . Value of monthly_revenue is 14
Possible Outlier for instant . Value of monthly_revenue is 10
Possible Outlier for instant . Value of monthly_revenue is 11
Possible Outlier for instant . Value of monthly_revenue is 11
Possible Outlier for instant . Value of monthly_revenue is 11
Possible Outlier for instant . Value of monthly_revenue is 10
Possible Outlier for instant . Value of monthly_revenue is 13
Possible Outlier for instant . Value of monthly_revenue is 14
Possible Outlier for instant . Value of monthly_revenue is 11
Possible Outlier for instant . Value of monthly_revenue is .
```

Output the outliers for the variable monthly\_revenue based on the interquantile range method

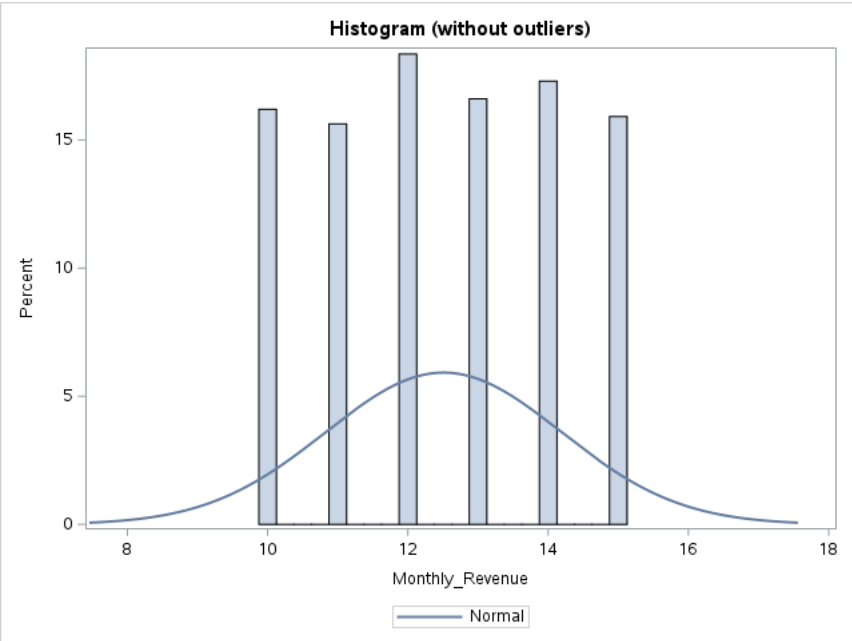
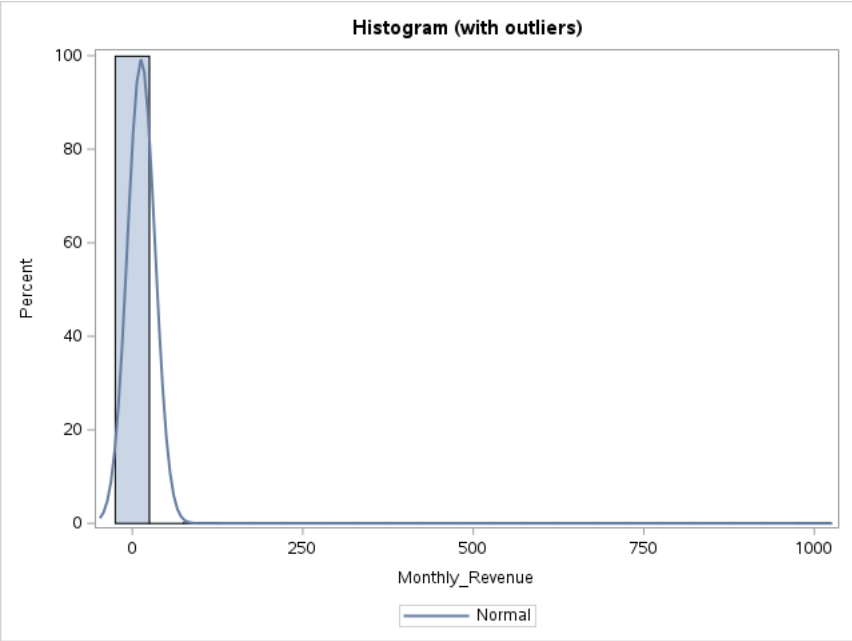
```
Possible Outlier for instant . Value of monthly_revenue is 15
Possible Outlier for instant . Value of monthly_revenue is 14
Possible Outlier for instant . Value of monthly_revenue is 15
Possible Outlier for instant . Value of monthly_revenue is 12
Possible Outlier for instant . Value of monthly_revenue is 13
Possible Outlier for instant . Value of monthly_revenue is 15
```

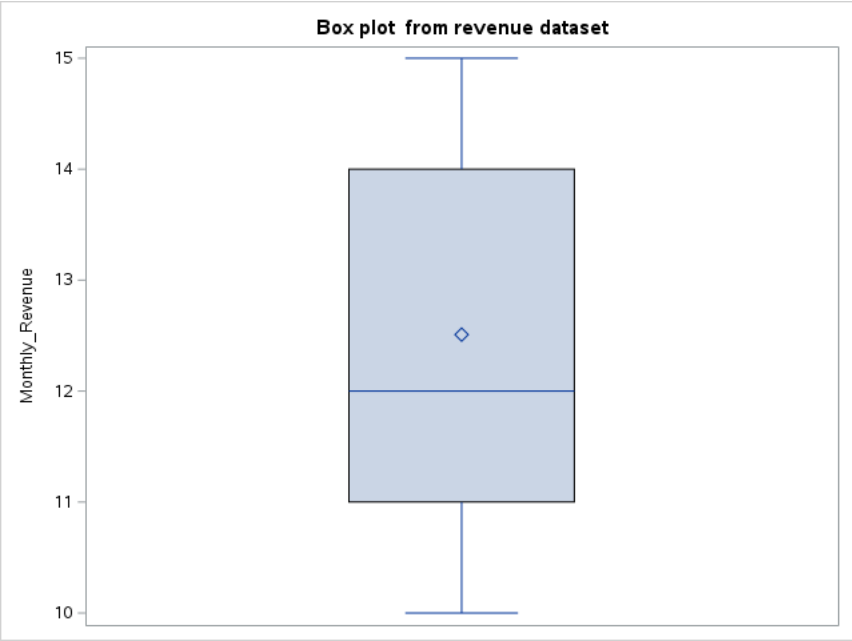
Dataset without the outliers

Obs	Subscription_Type	Monthly_Revenue	Join_Date	Last_Payment_Date	Country	Age	Gender	Device	Plan_Duration	_TYPE_	_FREQ_	monthly_revenue_Q1	monthly_revenue_Q3	monthly_reven
1	Basic	10	15-01-22	45205	United States	28	Male	Smartphone	1 Month	0	2500	11	14	
2	Premium	15	44325	22-06-23	Canada	35	Female	Tablet	1 Month	0	2500	11	14	
3	Standard	12	28-02-23	27-06-23	United Kingdom	42	Male	Smart TV	1 Month	0	2500	11	14	
4	Standard	12	44841	26-06-23	Australia	51	Female	Laptop	1 Month	0	2500	11	14	
5	Premium	15	18-03-22	27-06-23	France	29	Female	Smart TV	1 Month	0	2500	11	14	
6	Basic	10	44961	24-06-23	Mexico	39	Female	Laptop	1 Month	0	2500	11	14	
7	Standard	12	20-10-22	23-06-23	Spain	37	Male	Smartphone	1 Month	0	2500	11	14	
8	Premium	15	45108	22-06-23	Italy	44	Female	Smart TV	1 Month	0	2500	11	14	
9	Basic	10	16-05-22	22-06-23	United States	31	Female	Smartphone	1 Month	0	2500	11	14	
10	Premium	15	23-03-23	28-06-23	Canada	45	Male	Tablet	1 Month	0	2500	11	14	
11	Standard	12	30-11-21	27-06-23	United Kingdom	48	Female	Laptop	1 Month	0	2500	11	14	
12	Standard	12	45174	28-06-23	Germany	38	Female	Smart TV	1 Month	0	2500	11	14	



Obs	Subscription_Type	Monthly_Revenue	Join_Date	Last_Payment_Date	Country	Age	Gender	Device	Plan_Duration	_TYPE_	_FREQ_	monthly_revenue_Q1	monthly_revenue_Q3	monthly_reven
13	Premium	15	44746	27-06-23	France	36	Male	Tablet	1 Month	0	2500	11	14	
14	Basic	10	24-01-22	25-06-23	Brazil	30	Female	Laptop	1 Month	0	2500	11	14	
15	Standard	12	18-10-21	24-06-23	Mexico	43	Male	Smartphone	1 Month	0	2500	11	14	
16	Premium	15	15-02-23	23-06-23	Spain	32	Female	Smart TV	1 Month	0	2500	11	14	
17	Basic	10	27-05-23	22-06-23	Italy	41	Male	Tablet	1 Month	0	2500	11	14	
18	Premium	15	45205	22-06-23	United States	26	Female	Laptop	1 Month	0	2500	11	14	
19	Standard	12	44328	27-06-23	United Kingdom	49	Female	Smart TV	1 Month	0	2500	11	14	
20	Standard	12	44624	26-06-23	Australia	31	Male	Tablet	1 Month	0	2500	11	14	
21	Basic	10	14-03-23	28-06-23	Germany	40	Female	Laptop	1 Month	0	2500	11	14	
22	Premium	15	44896	27-06-23	France	29	Male	Smartphone	1 Month	0	2500	11	14	
23	Basic	10	29-08-22	25-06-23	Brazil	47	Female	Smart TV	1 Month	0	2500	11	14	
24	Standard	12	27-09-21	24-06-23	Mexico	33	Male	Tablet	1 Month	0	2500	11	14	
25	Premium	15	19-12-22	23-06-23	Spain	36	Female	Laptop	1 Month	0	2500	11	14	
26	Basic	10	17-05-23	22-06-23	Italy	42	Male	Smartphone	1 Month	0	2500	11	14	
27	Basic	10	44749	28-06-23	Canada	37	Male	Tablet	1 Month	0	2500	11	14	
28	Standard	12	21-01-22	27-06-23	United Kingdom	41	Female	Laptop	1 Month	0	2500	11	14	
29	Standard	12	25-05-22	26-06-23	Australia	29	Male	Smartphone	1 Month	0	2500	11	14	
30	Basic	10	45234	28-06-23	Germany	39	Female	Smart TV	1 Month	0	2500	11	14	





Q-Q plot with outliers

The UNIVARIATE Procedure  
Variable: Monthly\_Revenue (Monthly\_Revenue)

Moments			
N	2462	Sum Weights	2462
Mean	12.9756296	Sum Observations	31946
Std Deviation	20.1501529	Variance	406.028662
Skewness	47.874674	Kurtosis	2342.37983
Uncorrected SS	1413756	Corrected SS	999236.538
Coeff Variation	155.292295	Std Error Mean	0.40610125

Basic Statistical Measures			
Location		Variability	
Mean	12.97563	Std Deviation	20.15015
Median	12.00000	Variance	406.02866
Mode	12.00000	Range	1000
		Interquartile Range	3.00000

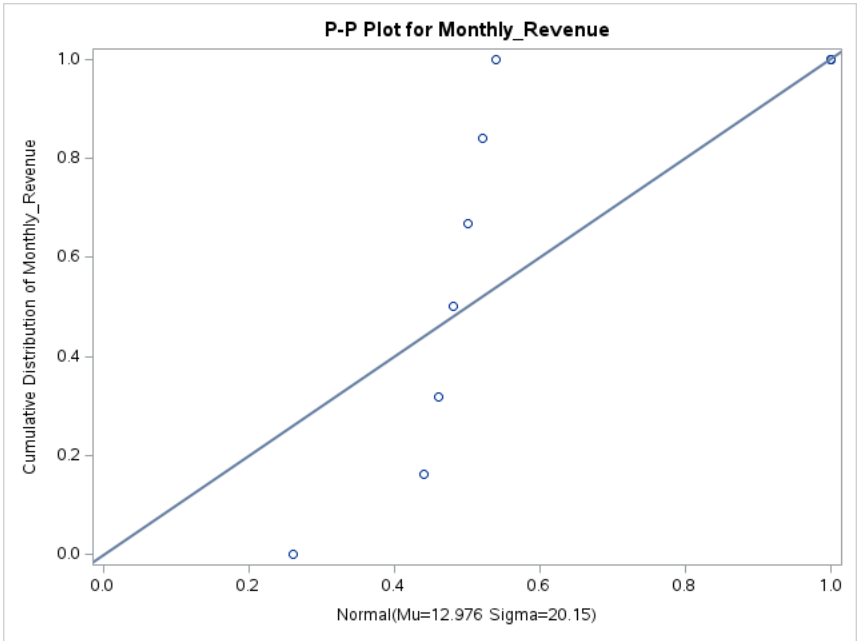
Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	31.95171	Pr >  t	<.0001
Sign	M	1230	Pr >=  M	<.0001
Signed Rank	S	1513515	Pr >=  S	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	1000
99%	15
95%	15
90%	15
75% Q3	14
50% Median	12
25% Q1	11
10%	10
5%	10
1%	10
0% Min	0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	780	15	2497
0	717	15	2500
10	2490	100	335
10	2486	111	372
10	2481	1000	2310

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	38	1.52	100.00

Q-Q plot with outliers  
The UNIVARIATE Procedure



Q-Q plot with outliers  
The UNIVARIATE Procedure  
Variable: Age (Age)

Moments			
N	2500	Sum Weights	2500
Mean	38.9928	Sum Observations	97482
Std Deviation	10.1222592	Variance	102.460132
Skewness	12.2308609	Kurtosis	303.775967
Uncorrected SS	4057144	Corrected SS	256047.87
Coeff Variation	25.9593034	Std Error Mean	0.20244518

Basic Statistical Measures			
Location		Variability	
Mean	38.99280	Std Deviation	10.12226
Median	39.00000	Variance	102.46013
Mode	30.00000	Range	300.00000
		Interquartile Range	13.00000

Note: The mode displayed is the smallest of 2 modes with a count of 116.

Tests for Location: Mu0=0			
Test	Statistic	p Value	
Student's t	t 192.6092	Pr >  t	<.0001
Sign	M 1249.5	Pr >=  M	<.0001
Signed Rank	S 1561875	Pr >=  S	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	300
99%	51
95%	50
90%	49
75% Q3	45
50% Median	39
25% Q1	32
10%	29
5%	28
1%	27
0% Min	0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	796	51	2426
26	21	51	2449
27	2472	51	2479
27	2447	280	745
27	2412	300	727

Q-Q plot without outliers

The UNIVARIATE Procedure  
Variable: Monthly\_Revenue (Monthly\_Revenue)

Moments			
N	2457	Sum Weights	2457
Mean	12.5091575	Sum Observations	30735
Std Deviation	1.68414284	Variance	2.83633712
Skewness	-0.0143105	Kurtosis	-1.2293155
Uncorrected SS	391435	Corrected SS	6966.04396
Coeff Variation	13.4632795	Std Error Mean	0.03397632

Basic Statistical Measures			
Location		Variability	
Mean	12.50916	Std Deviation	1.68414
Median	12.00000	Variance	2.83634
Mode	12.00000	Range	5.00000
		Interquartile Range	3.00000

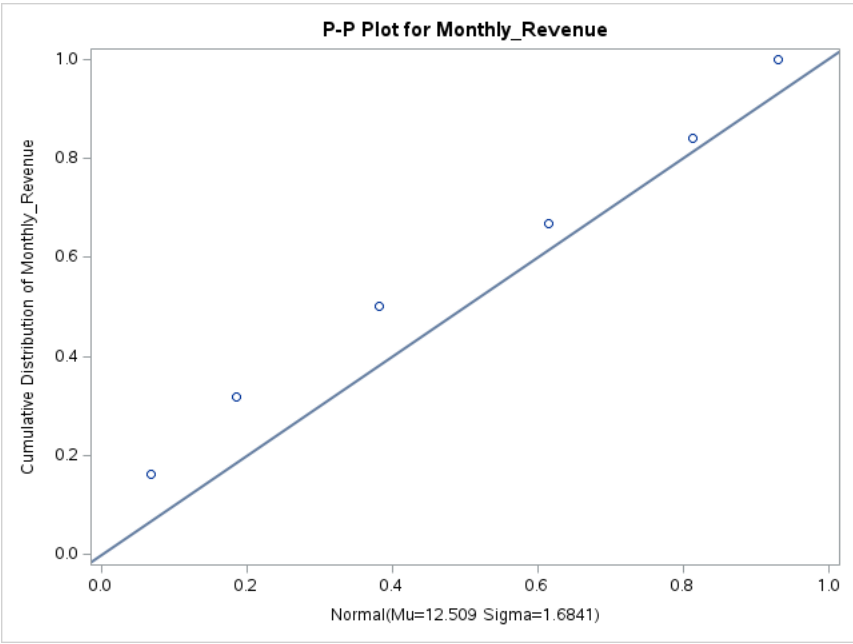
Tests for Location: Mu0=0			
Test	Statistic		p Value
Student's t	t	368.1728	Pr >  t  <.0001
Sign	M	1228.5	Pr >=  M  <.0001
Signed Rank	S	1509827	Pr >=  S  <.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	15
99%	15
95%	15
90%	15
75% Q3	14
50% Median	12
25% Q1	11
10%	10
5%	10
1%	10
0% Min	10

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
10	2448	15	2426
10	2444	15	2441
10	2439	15	2452
10	2436	15	2454
10	2432	15	2457

Q-Q plot without outliers

The UNIVARIATE Procedure



Q-Q plot without outliers

The UNIVARIATE Procedure  
Variable: Age (Age)

Moments			
N	2457	Sum Weights	2457
Mean	39.020757	Sum Observations	95874
Std Deviation	10.1723331	Variance	103.476361
Skewness	12.2570034	Kurtosis	302.916749
Uncorrected SS	3995214	Corrected SS	254137.941
Coeff Variation	26.0690306	Std Error Mean	0.2052192

Basic Statistical Measures			
Location		Variability	
Mean	39.02076	Std Deviation	10.17233
Median	39.00000	Variance	103.47636
Mode	30.00000	Range	300.00000
		Interquartile Range	13.00000

Tests for Location: Mu0=0			
Test		Statistic	p Value
Student's t	t	190.1418	Pr >  t  <.0001
Sign	M	1228	Pr >=  M  <.0001
Signed Rank	S	1508598	Pr >=  S  <.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	300
99%	51
95%	50
90%	49
75% Q3	45
50% Median	39
25% Q1	32
10%	29
5%	28
1%	27
0% Min	0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	777	51	2385
26	18	51	2408
27	2431	51	2437
27	2406	280	728
27	2372	300	710

Q-Q plot without outliers

The UNIVARIATE Procedure  
Variable: \_TYPE\_

Moments			
N	2457	Sum Weights	2457
Mean	0	Sum Observations	0
Std Deviation	0	Variance	0
Skewness	.	Kurtosis	.
Uncorrected SS	0	Corrected SS	0
Coeff Variation	.	Std Error Mean	0

Basic Statistical Measures			
Location		Variability	
Mean	0	Std Deviation	0
Median	0	Variance	0
Mode	0	Range	0
		Interquartile Range	0

Tests for Location: Mu0=0			
Test	Statistic	p Value	
Student's t	t	Pr >  t	.
Sign	M	Pr >=  M	.
Signed Rank	S	Pr >=  S	.

Quantiles (Definition 5)	
Level	Quantile
100% Max	0
99%	0
95%	0
90%	0
75% Q3	0
50% Median	0
25% Q1	0
10%	0
5%	0
1%	0
0% Min	0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	2457	0	2453
0	2456	0	2454
0	2455	0	2455
0	2454	0	2456
0	2453	0	2457

Q-Q plot without outliers

The UNIVARIATE Procedure  
Variable: \_FREQ\_

Moments			
N	2457	Sum Weights	2457
Mean	2500	Sum Observations	6142500
Std Deviation	0	Variance	0
Skewness	.	Kurtosis	.
Uncorrected SS	1.53563E10	Corrected SS	0
Coeff Variation	0	Std Error Mean	0

Basic Statistical Measures			
Location		Variability	
Mean	2500.000	Std Deviation	0
Median	2500.000	Variance	0
Mode	2500.000	Range	0
		Interquartile Range	0

Tests for Location: Mu0=0			
Test	Statistic	p Value	
Student's t	t	Pr >  t	.
Sign	M 1228.5	Pr >=  M	<.0001
Signed Rank	S 1509827	Pr >=  S	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	2500

Quantiles (Definition 5)	
Level	Quantile
100% Max	2500
99%	2500
95%	2500
90%	2500
75% Q3	2500
50% Median	2500
25% Q1	2500
10%	2500
5%	2500
1%	2500
0% Min	2500

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
2500	2457	2500	2453
2500	2456	2500	2454
2500	2455	2500	2455
2500	2454	2500	2456
2500	2453	2500	2457

Q-Q plot without outliers

The UNIVARIATE Procedure  
Variable: monthly\_revenue\_Q1 (Monthly\_Revenue)

Moments			
N	2457	Sum Weights	2457
Mean	11	Sum Observations	27027
Std Deviation	0	Variance	0
Skewness	.	Kurtosis	.
Uncorrected SS	297297	Corrected SS	0
Coeff Variation	0	Std Error Mean	0

Basic Statistical Measures			
Location		Variability	
Mean	11.00000	Std Deviation	0
Median	11.00000	Variance	0
Mode	11.00000	Range	0
		Interquartile Range	0

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	.	Pr >  t	.
Sign	M	1228.5	Pr >=  M	<.0001
Signed Rank	S	1509827	Pr >=  S	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	11
99%	11
95%	11
90%	11
75% Q3	11
50% Median	11
25% Q1	11
10%	11
5%	11
1%	11
0% Min	11

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
11	2457	11	2453
11	2456	11	2454
11	2455	11	2455
11	2454	11	2456
11	2453	11	2457

Q-Q plot without outliers

The UNIVARIATE Procedure  
Variable: monthly\_revenue\_Q3 (Monthly\_Revenue)

Moments			
N	2457	Sum Weights	2457

Moments			
Mean	14	Sum Observations	34398
Std Deviation	0	Variance	0
Skewness	.	Kurtosis	.
Uncorrected SS	481572	Corrected SS	0
Coeff Variation	0	Std Error Mean	0

Basic Statistical Measures			
Location		Variability	
Mean	14.00000	Std Deviation	0
Median	14.00000	Variance	0
Mode	14.00000	Range	0
		Interquartile Range	0

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	.	Pr >  t	.
Sign	M	1228.5	Pr >=  M	<.0001
Signed Rank	S	1509827	Pr >=  S	<.0001

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

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
14	2457	14	2453
14	2456	14	2454
14	2455	14	2455
14	2454	14	2456
14	2453	14	2457

Q-Q plot without outliers

The UNIVARIATE Procedure  
Variable: monthly\_revenue\_QRange (Monthly\_Revenue)

Moments			
N	2457	Sum Weights	2457
Mean	3	Sum Observations	7371
Std Deviation	0	Variance	0
Skewness	.	Kurtosis	.
Uncorrected SS	22113	Corrected SS	0
Coeff Variation	0	Std Error Mean	0

Basic Statistical Measures			
Location		Variability	
Mean	3.000000	Std Deviation	0
Median	3.000000	Variance	0
Mode	3.000000	Range	0
		Interquartile Range	0

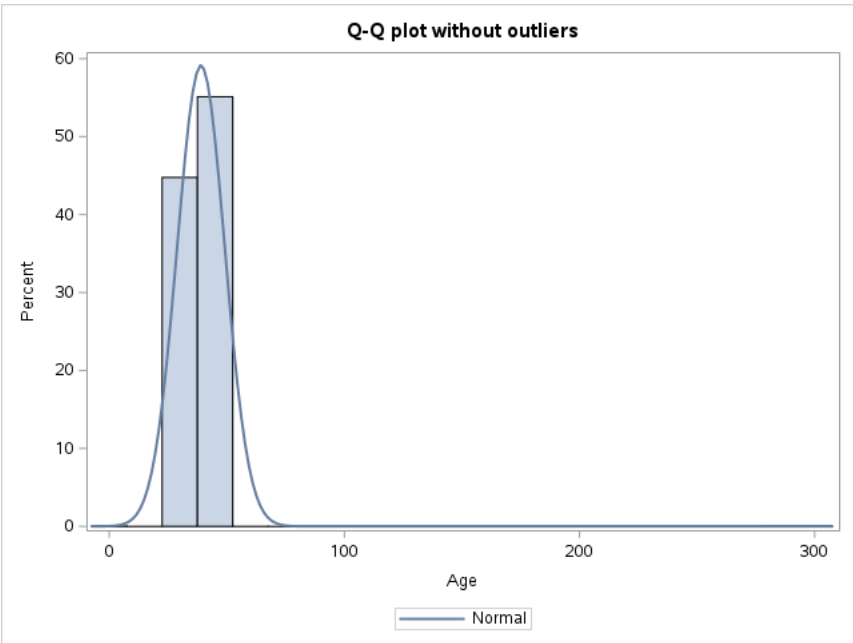
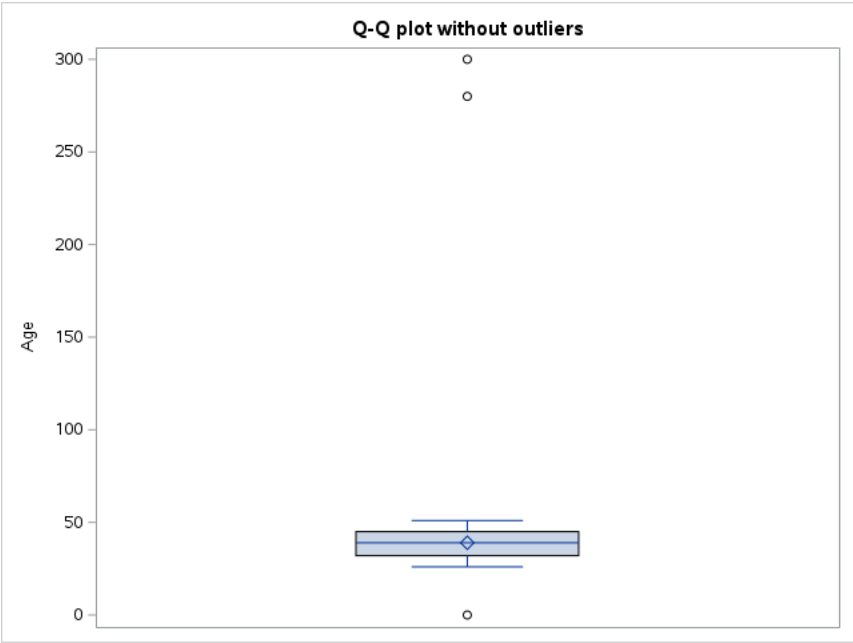
Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	.	Pr >  t	.
Sign	M	1228.5	Pr >=  M	<.0001
Signed Rank	S	1509827	Pr >=  S	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	3
99%	3
95%	3
90%	3
75% Q3	3



Quantiles (Definition 5)	
Level	Quantile
50% Median	3
25% Q1	3
10%	3
5%	3
1%	3
0% Min	3

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
3	2457	3	2453
3	2456	3	2454
3	2455	3	2455
3	2454	3	2456
3	2453	3	2457



Q-Q plot

The UNIVARIATE Procedure

Variable: Monthly\_Revenue (Monthly\_Revenue)

Moments			
N	2462	Sum Weights	2462
Mean	12.9756296	Sum Observations	31946
Std Deviation	20.1501529	Variance	406.028662
Skewness	47.874674	Kurtosis	2342.37983
Uncorrected SS	1413756	Corrected SS	999236.538
Coeff Variation	155.292295	Std Error Mean	0.40610125

Basic Statistical Measures			
Location		Variability	
Mean	12.97563	Std Deviation	20.15015
Median	12.00000	Variance	406.02866
Mode	12.00000	Range	1000
		Interquartile Range	3.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	31.95171	Pr >  t	<.0001
Sign	M	1230	Pr >=  M	<.0001
Signed Rank	S	1513515	Pr >=  S	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	1000
99%	15
95%	15
90%	15
75% Q3	14
50% Median	12
25% Q1	11
10%	10
5%	10
1%	10
0% Min	0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	780	15	2497
0	717	15	2500
10	2490	100	335
10	2486	111	372
10	2481	1000	2310

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	38	1.52	100.00

Q-Q plot

The UNIVARIATE Procedure  
Variable: Age (Age)

Moments			
N	2500	Sum Weights	2500
Mean	38.9928	Sum Observations	97482
Std Deviation	10.1222592	Variance	102.460132
Skewness	12.2308609	Kurtosis	303.775967
Uncorrected SS	4057144	Corrected SS	256047.87
Coeff Variation	25.9593034	Std Error Mean	0.20244518

Basic Statistical Measures			
Location		Variability	
Mean	38.99280	Std Deviation	10.12226
Median	39.00000	Variance	102.46013
Mode	30.00000	Range	300.00000
		Interquartile Range	13.00000

Note: The mode displayed is the smallest of 2 modes with a count of 116.

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	192.6092	Pr >  t	<.0001
Sign	M	1249.5	Pr >=  M	<.0001

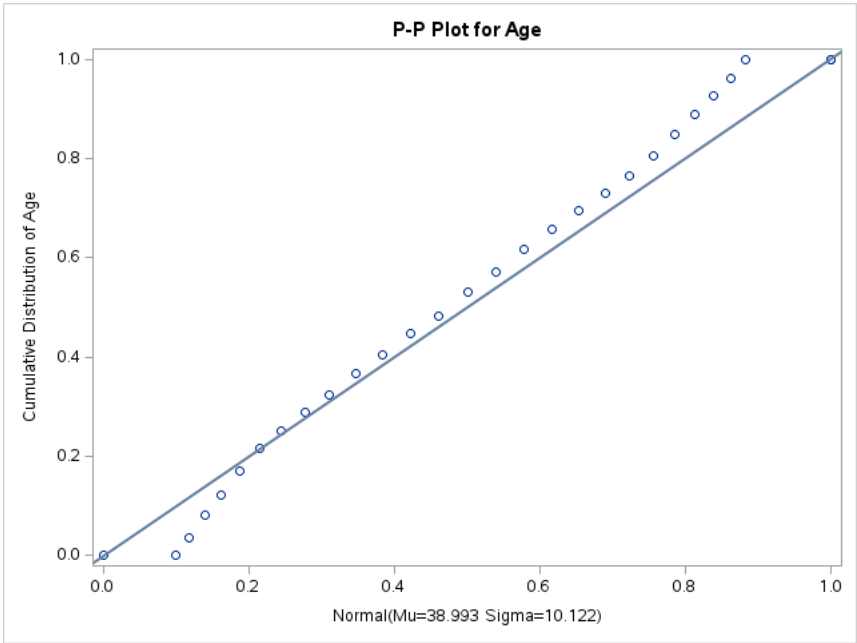
Tests for Location: Mu0=0			
Test	Statistic		p Value
Signed Rank	S	1561875	Pr >=  S  <.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	300
99%	51
95%	50
90%	49
75% Q3	45
50% Median	39
25% Q1	32
10%	29
5%	28
1%	27
0% Min	0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	796	51	2426
26	21	51	2449
27	2472	51	2479
27	2447	280	745
27	2412	300	727

Q-Q plot

The UNIVARIATE Procedure



Q-Q plot

The UNIVARIATE Procedure  
Variable: Age (Age)

Moments			
N	2500	Sum Weights	2500
Mean	38.9928	Sum Observations	97482
Std Deviation	10.1222592	Variance	102.460132
Skewness	12.2308609	Kurtosis	303.775967
Uncorrected SS	4057144	Corrected SS	256047.87
Coeff Variation	25.9593034	Std Error Mean	0.20244518

Basic Statistical Measures			
Location		Variability	
Mean	38.99280	Std Deviation	10.12226
Median	39.00000	Variance	102.46013
Mode	30.00000	Range	300.00000
		Interquartile Range	13.00000

Note: The mode displayed is the smallest of 2 modes with a count of 116.

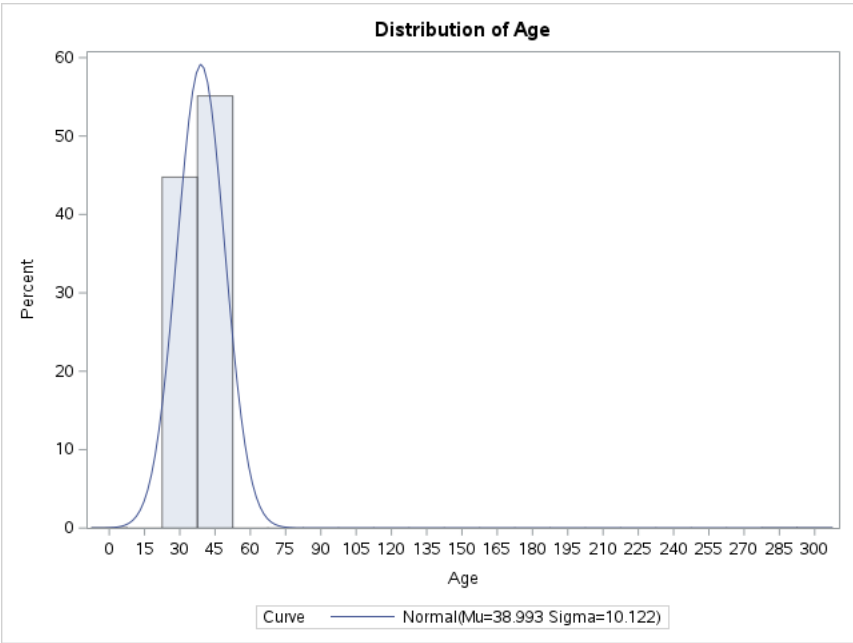
Tests for Location: Mu0=0				
Test		Statistic	p Value	
Student's t	t	192.6092	Pr >  t	<.0001
Sign	M	1249.5	Pr >=  M	<.0001
Signed Rank	S	1561875	Pr >=  S	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	300
99%	51
95%	50
90%	49
75% Q3	45
50% Median	39
25% Q1	32
10%	29
5%	28
1%	27
0% Min	0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	796	51	2426
26	21	51	2449
27	2472	51	2479
27	2447	280	745
27	2412	300	727

Q-Q plot

The UNIVARIATE Procedure



Q-Q plot

The UNIVARIATE Procedure  
Fitted Normal Distribution for Age (Age)

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	38.9928
Std Dev	Sigma	10.12226

Goodness-of-Fit Tests for Normal Distribution			
Test		Statistic	p Value
Kolmogorov-Smirnov	D	0.1172493	Pr > D <0.010
Cramer-von Mises	W-Sq	5.4118936	Pr > W-Sq <0.005
Anderson-Darling	A-Sq	57.2377082	Pr > A-Sq <0.005

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	27.0000	15.4449
5.0	28.0000	22.3432
10.0	29.0000	26.0206
25.0	32.0000	32.1654
50.0	39.0000	38.9928
75.0	45.0000	45.8202
90.0	49.0000	51.9650
95.0	50.0000	55.6424
99.0	51.0000	62.5407

[illegible][illegible]

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

43/74

Output the outliers for the variable `monthly_revenue` based on the interquantile range method

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

44/74

Output the outliers for the variable `monthly_revenue` based on the interquartile range method

Output the outliers for the variable monthly\_revenue based on the interquartile range method

45/74



Output the outliers for the variable `monthly_revenue` based on the interquartile range method

46/74

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

[illegible]

**Output the outliers for the variable monthly\_revenue based on the interquartile range method**

**Output the outliers for the variable monthly\_revenue based on the interquartile range method**

48/74

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

49/74

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

50/74

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

Output the outliers for the variable `monthly_revenue` based on the interquartile range method

Output the outliers for the variable `monthly_revenue` based on the interquartile range method

52/74

Output the outliers for the variable `monthly_revenue` based on the interquantile range method

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

53/74



**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

54/74

[illegible][illegible]

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

56/74

[illegible]

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

58/74

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

59/74

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

Output the outliers for the variable `monthly_revenue` based on the interquantile range method

60/74

Output the outliers for the variable `monthly_revenue` based on the interquartile range method

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

61/74



**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

62/74

[illegible][illegible]

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

**Output the outliers for the variable monthly\_revenue based on the interquantile range method**

64/74

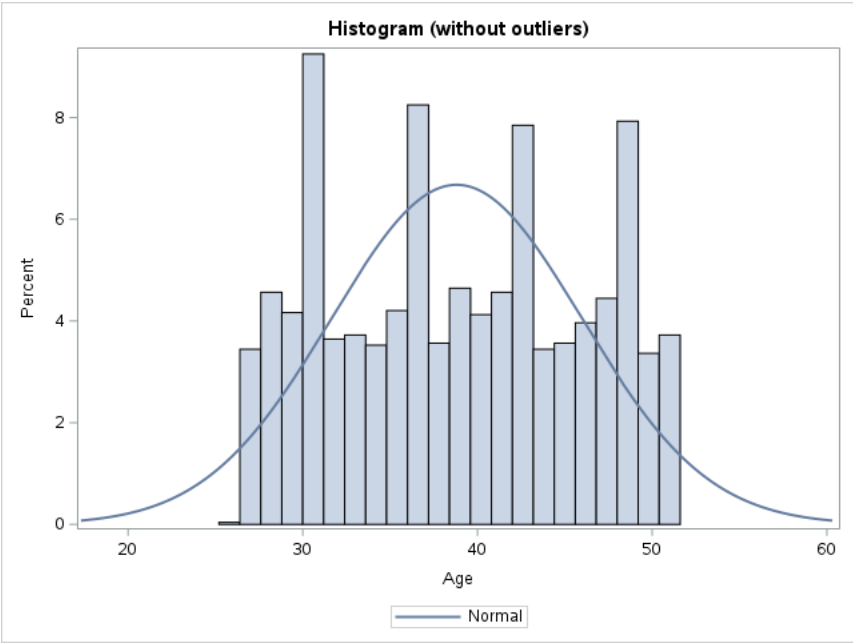
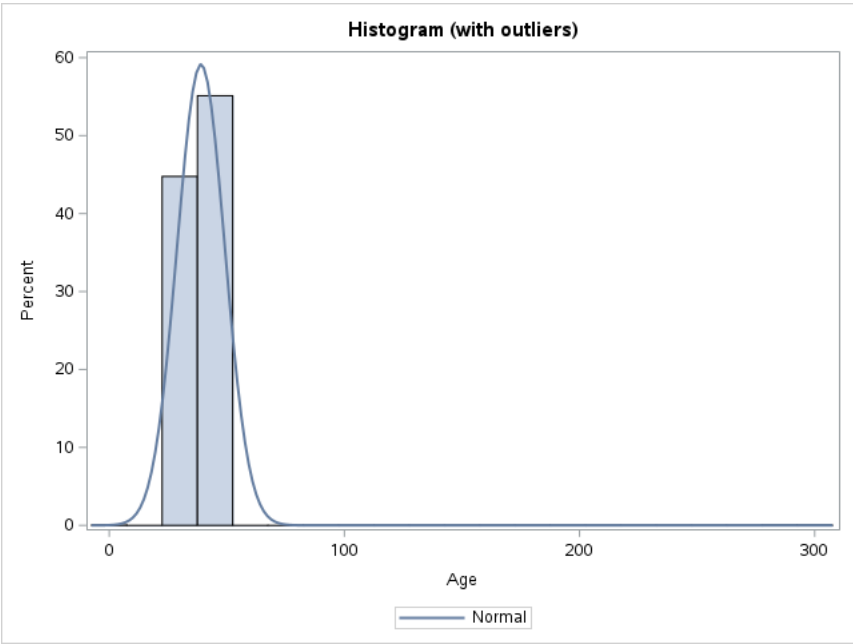
Possible Outlier for instant . Value of monthly\_revenue is .  
Possible Outlier for instant . Value of monthly\_revenue is .  
Possible Outlier for instant . Value of monthly\_revenue is .  
Possible Outlier for instant . Value of monthly\_revenue is .  
Possible Outlier for instant . Value of monthly\_revenue is .  
Possible Outlier for instant . Value of monthly\_revenue is .  
Possible Outlier for instant . Value of monthly\_revenue is .  
Possible Outlier for instant . Value of monthly\_revenue is .  
Possible Outlier for instant . Value of monthly\_revenue is .  
Possible Outlier for instant . Value of monthly\_revenue is .  
Possible Outlier for instant . Value of monthly\_revenue is .  
Possible Outlier for instant . Value of monthly\_revenue is .  
Possible Outlier for instant . Value of monthly\_revenue is .  
Possible Outlier for instant . Value of monthly\_revenue is .  
Possible Outlier for instant . Value of monthly\_revenue is .

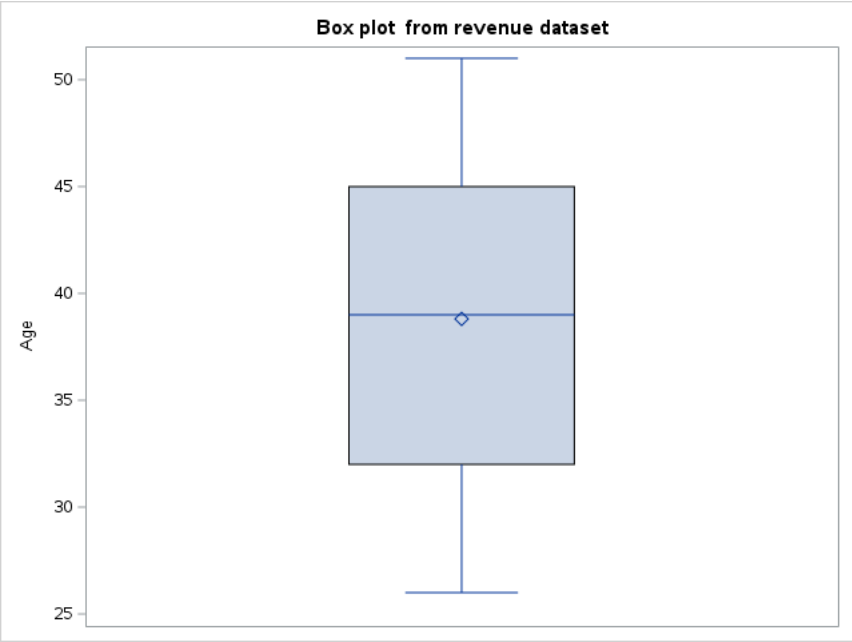
Output the outliers for the variable monthly\_revenue based on the interquantile range method

Possible Outlier for instant . Value of monthly\_revenue is .  
Possible Outlier for instant . Value of monthly\_revenue is .  
Possible Outlier for instant . Value of monthly\_revenue is .  
Possible Outlier for instant . Value of monthly\_revenue is .  
Possible Outlier for instant . Value of monthly\_revenue is .  
Possible Outlier for instant . Value of monthly\_revenue is .

Dataset without the outliers

Obs	Subscription_Type	Monthly_Revenue	Join_Date	Last_Payment_Date	Country	Age	Gender	Device	Plan_Duration	_TYPE_	_FREQ_	age_Q1	age_Q3	age_QRange
1	Basic	10	15-01-22	45205	United States	28	Male	Smartphone	1 Month	0	2500	32	45	13
2	Premium	15	44325	22-06-23	Canada	35	Female	Tablet	1 Month	0	2500	32	45	13
3	Standard	12	28-02-23	27-06-23	United Kingdom	42	Male	Smart TV	1 Month	0	2500	32	45	13
4	Standard	12	44841	26-06-23	Australia	51	Female	Laptop	1 Month	0	2500	32	45	13
5	Basic	.	44931	28-06-23	Germany	33	Male	Smartphone	1 Month	0	2500	32	45	13
6	Premium	15	18-03-22	27-06-23	France	29	Female	Smart TV	1 Month	0	2500	32	45	13
7	Standard	.	44451	25-06-23	Brazil	46	Male	Tablet	1 Month	0	2500	32	45	13
8	Basic	10	44961	24-06-23	Mexico	39	Female	Laptop	1 Month	0	2500	32	45	13
9	Standard	12	20-10-22	23-06-23	Spain	37	Male	Smartphone	1 Month	0	2500	32	45	13
10	Premium	15	45108	22-06-23	Italy	44	Female	Smart TV	1 Month	0	2500	32	45	13
11	Basic	10	16-05-22	22-06-23	United States	31	Female	Smartphone	1 Month	0	2500	32	45	13
12	Premium	15	23-03-23	28-06-23	Canada	45	Male	Tablet	1 Month	0	2500	32	45	13
13	Standard	12	30-11-21	27-06-23	United Kingdom	48	Female	Laptop	1 Month	0	2500	32	45	13
14	Basic	.	44569	26-06-23	Australia	27	Male	Smartphone	1 Month	0	2500	32	45	13
15	Standard	12	45174	28-06-23	Germany	38	Female	Smart TV	1 Month	0	2500	32	45	13
16	Premium	15	44746	27-06-23	France	36	Male	Tablet	1 Month	0	2500	32	45	13
17	Basic	10	24-01-22	25-06-23	Brazil	30	Female	Laptop	1 Month	0	2500	32	45	13
18	Standard	12	18-10-21	24-06-23	Mexico	43	Male	Smartphone	1 Month	0	2500	32	45	13
19	Premium	15	15-02-23	23-06-23	Spain	32	Female	Smart TV	1 Month	0	2500	32	45	13
20	Basic	10	27-05-23	22-06-23	Italy	41	Male	Tablet	1 Month	0	2500	32	45	13
21	Premium	15	45205	22-06-23	United States	26	Female	Laptop	1 Month	0	2500	32	45	13
22	Basic	.	22-07-22	28-06-23	Canada	34	Male	Smartphone	1 Month	0	2500	32	45	13
23	Standard	12	44328	27-06-23	United Kingdom	49	Female	Smart TV	1 Month	0	2500	32	45	13
24	Standard	12	44624	26-06-23	Australia	31	Male	Tablet	1 Month	0	2500	32	45	13
25	Basic	10	14-03-23	28-06-23	Germany	40	Female	Laptop	1 Month	0	2500	32	45	13
26	Premium	15	44896	27-06-23	France	29	Male	Smartphone	1 Month	0	2500	32	45	13
27	Basic	10	29-08-22	25-06-23	Brazil	47	Female	Smart TV	1 Month	0	2500	32	45	13
28	Standard	12	27-09-21	24-06-23	Mexico	33	Male	Tablet	1 Month	0	2500	32	45	13
29	Premium	15	19-12-22	23-06-23	Spain	36	Female	Laptop	1 Month	0	2500	32	45	13
30	Basic	10	17-05-23	22-06-23	Italy	42	Male	Smartphone	1 Month	0	2500	32	45	13





**Q-Q plot with outliers**

The UNIVARIATE Procedure  
Variable: Monthly\_Revenue (Monthly\_Revenue)

Moments			
N	2462	Sum Weights	2462
Mean	12.9756296	Sum Observations	31946
Std Deviation	20.1501529	Variance	406.028662
Skewness	47.874674	Kurtosis	2342.37983
Uncorrected SS	1413756	Corrected SS	999236.538
Coeff Variation	155.292295	Std Error Mean	0.40610125

Basic Statistical Measures			
Location		Variability	
Mean	12.97563	Std Deviation	20.15015
Median	12.00000	Variance	406.02866
Mode	12.00000	Range	1000
		Interquartile Range	3.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	31.95171	Pr >  t	<.0001
Sign	M	1230	Pr >=  M	<.0001
Signed Rank	S	1513515	Pr >=  S	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	1000
99%	15
95%	15
90%	15
75% Q3	14
50% Median	12
25% Q1	11
10%	10
5%	10
1%	10
0% Min	0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	780	15	2497
0	717	15	2500
10	2490	100	335
10	2486	111	372
10	2481	1000	2310

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	38	1.52	100.00

Q-Q plot with outliers

The UNIVARIATE Procedure  
Variable: Age (Age)

Moments			
N	2500	Sum Weights	2500
Mean	38.9928	Sum Observations	97482
Std Deviation	10.1222592	Variance	102.460132
Skewness	12.2308609	Kurtosis	303.775967
Uncorrected SS	4057144	Corrected SS	256047.87
Coeff Variation	25.9593034	Std Error Mean	0.20244518

Basic Statistical Measures			
Location		Variability	
Mean	38.99280	Std Deviation	10.12226
Median	39.00000	Variance	102.46013
Mode	30.00000	Range	300.00000
		Interquartile Range	13.00000

Note: The mode displayed is the smallest of 2 modes with a count of 116.

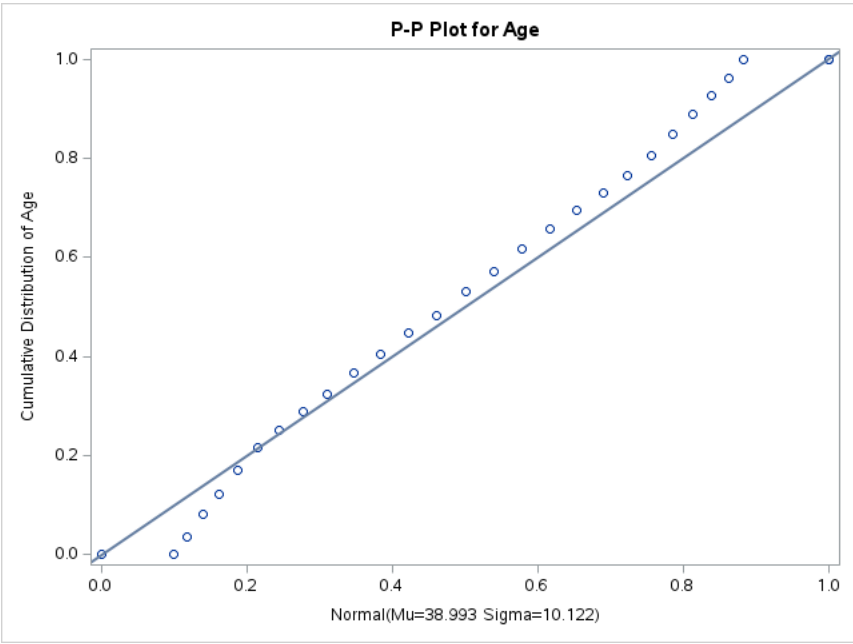
Tests for Location: Mu0=0				
Test		Statistic	p Value	
Student's t	t	192.6092	Pr >  t	<.0001
Sign	M	1249.5	Pr >=  M	<.0001
Signed Rank	S	1561875	Pr >=  S	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	300
99%	51
95%	50
90%	49
75% Q3	45
50% Median	39
25% Q1	32
10%	29
5%	28
1%	27
0% Min	0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	796	51	2426
26	21	51	2449
27	2472	51	2479
27	2447	280	745
27	2412	300	727

Q-Q plot with outliers

The UNIVARIATE Procedure



Q-Q plot without outliers

The UNIVARIATE Procedure  
Variable: Monthly\_Revenue (Monthly\_Revenue)

Moments			
N	2459	Sum Weights	2459
Mean	12.9776332	Sum Observations	31912
Std Deviation	20.1623574	Variance	406.520655
Skewness	47.8458647	Kurtosis	2339.54979
Uncorrected SS	1413370	Corrected SS	999227.77
Coeff Variation	155.362361	Std Error Mean	0.40659501

Basic Statistical Measures			
Location		Variability	
Mean	12.97763	Std Deviation	20.16236
Median	12.00000	Variance	406.52065
Mode	12.00000	Range	1000
		Interquartile Range	3.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	31.91784	Pr >  t	<.0001
Sign	M	1228.5	Pr >=  M	<.0001
Signed Rank	S	1509827	Pr >=  S	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	1000
99%	15
95%	15
90%	15
75% Q3	14
50% Median	12
25% Q1	11
10%	10
5%	10
1%	10
0% Min	0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	778	15	2494
0	717	15	2497
10	2487	100	335
10	2483	111	372
10	2478	1000	2307



Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	38	1.52	100.00

Q-Q plot without outliers

The UNIVARIATE Procedure  
Variable: Age (Age)

Moments			
N	2497	Sum Weights	2497
Mean	38.8073688	Sum Observations	96902
Std Deviation	7.16764522	Variance	51.375138
Skewness	0.02742358	Kurtosis	-1.1988272
Uncorrected SS	3888744	Corrected SS	128232.344
Coeff Variation	18.4698047	Std Error Mean	0.14343899

Basic Statistical Measures			
Location		Variability	
Mean	38.80737	Std Deviation	7.16765
Median	39.00000	Variance	51.37514
Mode	30.00000	Range	25.00000
		Interquartile Range	13.00000

Note: The mode displayed is the smallest of 2 modes with a count of 116.

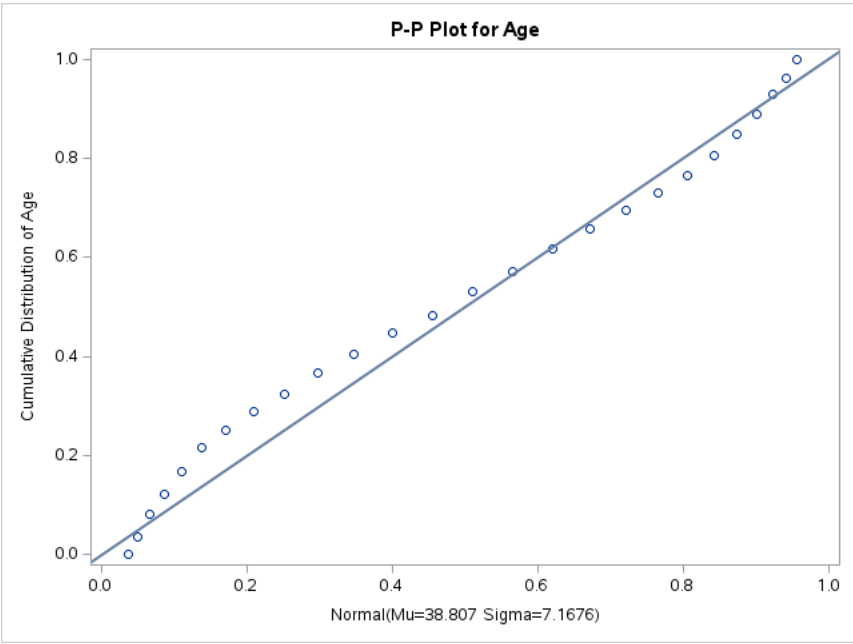
Tests for Location: Mu0=0				
Test		Statistic	p Value	
Student's t	t	270.5496	Pr >  t	<.0001
Sign	M	1248.5	Pr >=  M	<.0001
Signed Rank	S	1559377	Pr >=  S	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	51
99%	51
95%	50
90%	49
75% Q3	45
50% Median	39
25% Q1	32
10%	29
5%	28
1%	27
0% Min	26

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
26	21	51	2375
27	2469	51	2390
27	2444	51	2423
27	2409	51	2446
27	2394	51	2476

Q-Q plot without outliers

The UNIVARIATE Procedure



Q-Q plot without outliers

The UNIVARIATE Procedure  
Variable: \_TYPE\_

Moments			
N	2497	Sum Weights	2497
Mean	0	Sum Observations	0
Std Deviation	0	Variance	0
Skewness	.	Kurtosis	.
Uncorrected SS	0	Corrected SS	0
Coeff Variation	.	Std Error Mean	0

Basic Statistical Measures			
Location		Variability	
Mean	0	Std Deviation	0
Median	0	Variance	0
Mode	0	Range	0
		Interquartile Range	0

Tests for Location: Mu0=0			
Test	Statistic	p Value	
Student's t	t	Pr >  t	.
Sign	M	Pr >=  M	.
Signed Rank	S	Pr >=  S	.

Quantiles (Definition 5)	
Level	Quantile
100% Max	0
99%	0
95%	0
90%	0
75% Q3	0
50% Median	0
25% Q1	0
10%	0
5%	0
1%	0
0% Min	0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	2497	0	2493
0	2496	0	2494
0	2495	0	2495
0	2494	0	2496
0	2493	0	2497

Q-Q plot without outliers

The UNIVARIATE Procedure  
Variable: \_FREQ\_

Moments			
N	2497	Sum Weights	2497
Mean	2500	Sum Observations	6242500
Std Deviation	0	Variance	0
Skewness	.	Kurtosis	.
Uncorrected SS	1.56063E10	Corrected SS	0
Coeff Variation	0	Std Error Mean	0

Basic Statistical Measures			
Location		Variability	
Mean	2500.000	Std Deviation	0
Median	2500.000	Variance	0
Mode	2500.000	Range	0
		Interquartile Range	0

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	.	Pr >  t	.
Sign	M	1248.5	Pr >=  M	<.0001
Signed Rank	S	1559377	Pr >=  S	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	2500
99%	2500
95%	2500
90%	2500
75% Q3	2500
50% Median	2500
25% Q1	2500
10%	2500
5%	2500
1%	2500
0% Min	2500

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
2500	2497	2500	2493
2500	2496	2500	2494
2500	2495	2500	2495
2500	2494	2500	2496
2500	2493	2500	2497

Q-Q plot without outliers

The UNIVARIATE Procedure  
Variable: age\_Q1 (Age)

Moments			
N	2497	Sum Weights	2497
Mean	32	Sum Observations	79904
Std Deviation	0	Variance	0
Skewness	.	Kurtosis	.
Uncorrected SS	2556928	Corrected SS	0
Coeff Variation	0	Std Error Mean	0

Basic Statistical Measures			
Location		Variability	
Mean	32.00000	Std Deviation	0
Median	32.00000	Variance	0
Mode	32.00000	Range	0
		Interquartile Range	0

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	.	Pr >  t	.
Sign	M	1248.5	Pr >=  M	<.0001
Signed Rank	S	1559377	Pr >=  S	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	32

Quantiles (Definition 5)	
Level	Quantile
99%	32
95%	32
90%	32
75% Q3	32
50% Median	32
25% Q1	32
10%	32
5%	32
1%	32
0% Min	32

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
32	2497	32	2493
32	2496	32	2494
32	2495	32	2495
32	2494	32	2496
32	2493	32	2497

Q-Q plot without outliers

The UNIVARIATE Procedure

Variable: age\_Q3 (Age)

Moments			
N	2497	Sum Weights	2497
Mean	45	Sum Observations	112365
Std Deviation	0	Variance	0
Skewness	.	Kurtosis	.
Uncorrected SS	5056425	Corrected SS	0
Coeff Variation	0	Std Error Mean	0

Basic Statistical Measures			
Location		Variability	
Mean	45.00000	Std Deviation	0
Median	45.00000	Variance	0
Mode	45.00000	Range	0
		Interquartile Range	0

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	.	Pr >  t	.
Sign	M	1248.5	Pr >=  M	<.0001
Signed Rank	S	1559377	Pr >=  S	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	45
99%	45
95%	45
90%	45
75% Q3	45
50% Median	45
25% Q1	45
10%	45
5%	45
1%	45
0% Min	45

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
45	2497	45	2493
45	2496	45	2494
45	2495	45	2495
45	2494	45	2496
45	2493	45	2497

Q-Q plot without outliers

The UNIVARIATE Procedure

Variable: age\_QRange (Age)

Moments			
N	2497	Sum Weights	2497

Moments			
Mean	13	Sum Observations	32461
Std Deviation	0	Variance	0
Skewness	.	Kurtosis	.
Uncorrected SS	421993	Corrected SS	0
Coeff Variation	0	Std Error Mean	0

Basic Statistical Measures			
Location		Variability	
Mean	13.00000	Std Deviation	0
Median	13.00000	Variance	0
Mode	13.00000	Range	0
		Interquartile Range	0

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	.	Pr >  t	.
Sign	M	1248.5	Pr >=  M	<.0001
Signed Rank	S	1559377	Pr >=  S	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	13
99%	13
95%	13
90%	13
75% Q3	13
50% Median	13
25% Q1	13
10%	13
5%	13
1%	13
0% Min	13

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
13	2497	13	2493
13	2496	13	2494
13	2495	13	2495
13	2494	13	2496
13	2493	13	2497

```

1      OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69      FILENAME project '/home/u63417899/BAN110ZBB/Netflix Userbase 1.xlsx';
70
71      PROC IMPORT DATAFILE=project
72      DBMS=XLSX
73      replace
74      OUT=WORK.project2;
75      GETNAMES=YES;
76      RUN;

```

NOTE: One or more variables were converted because the data type is not supported by the V9 engine. For more details, run with options MSGLEVEL=I.

NOTE: The import data set has 2500 observations and 9 variables.

NOTE: WORK.PROJECT2 data set was successfully created.

NOTE: PROCEDURE IMPORT used (Total process time):

real time	0.31 seconds
user cpu time	0.32 seconds
system cpu time	0.00 seconds
memory	3527.09k
OS Memory	31740.00k
Timestamp	08/12/2023 02:16:18 AM
Step Count	805 Switch Count 4
Page Faults	0
Page Reclaims	687
Page Swaps	0
Voluntary Context Switches	24
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	760

```

77
78      PROC CONTENTS DATA=WORK.project2; RUN;

```

NOTE: PROCEDURE CONTENTS used (Total process time):

real time	0.04 seconds
user cpu time	0.04 seconds
system cpu time	0.00 seconds
memory	2000.87k
OS Memory	29612.00k
Timestamp	08/12/2023 02:16:18 AM
Step Count	806 Switch Count 0
Page Faults	0
Page Reclaims	96
Page Swaps	0
Voluntary Context Switches	0
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	16

```

79
80      proc print data=WORK.project2(obs=10);
81      run;

```

NOTE: There were 10 observations read from the data set WORK.PROJECT2.

NOTE: PROCEDURE PRINT used (Total process time):

real time	0.02 seconds
user cpu time	0.03 seconds
system cpu time	0.00 seconds
memory	870.50k
OS Memory	29352.00k
Timestamp	08/12/2023 02:16:18 AM
Step Count	807 Switch Count 0
Page Faults	0
Page Reclaims	63
Page Swaps	0
Voluntary Context Switches	0
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	16

```

82
83      /* Using Nmiss to ascertain the amount of missing values in each variable */
84      Title 'Means output before removing missing values';
85      proc means data=work.project2 N Nmiss Mean Min Max Maxdec=3;
86      run;

```

NOTE: There were 2500 observations read from the data set WORK.PROJECT2.

NOTE: PROCEDURE MEANS used (Total process time):

real time	0.02 seconds
user cpu time	0.03 seconds
system cpu time	0.01 seconds
memory	6915.06k
OS Memory	34492.00k
Timestamp	08/12/2023 02:16:18 AM
Step Count	808 Switch Count 1
Page Faults	0
Page Reclaims	1382
Page Swaps	0
Voluntary Context Switches	19
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	0

```

87
88      /*Replacing missing variables with its mean*/
89      proc standard data=work.project2 replace out=work.project2_nomiss;
90      var Age monthly_revenue;
91      run;

```

NOTE: The data set WORK.PROJECT2\_NOMISS has 2500 observations and 9 variables.

NOTE: PROCEDURE STANDARD used (Total process time):

real time	0.00 seconds
user cpu time	0.00 seconds
system cpu time	0.00 seconds
memory	1083.12k
OS Memory	29612.00k
Timestamp	08/12/2023 02:16:18 AM
Step Count	809 Switch Count 2
Page Faults	0
Page Reclaims	85
Page Swaps	0
Voluntary Context Switches	15
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	520

```

92
93      Title 'Means output after removing missing values';
94      proc means data=work.project2_nomiss N Nmiss Mean Min Max Maxdec=3;
95      run;

```

NOTE: There were 2500 observations read from the data set WORK.PROJECT2\_NOMISS.

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	6744.65k
OS Memory	34492.00k
Timestamp	08/12/2023 02:16:18 AM
Step Count	810 Switch Count 1
Page Faults	0
Page Reclaims	1355
Page Swaps	0
Voluntary Context Switches	19
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	0

```

95      !      */
96
97
98      /* OUTLIERS */
99      /* Outlier detection for some variables using box plots */;
100     proc sgplot data=work.project2;
101     vbox monthly_revenue;
102     run;

```

NOTE: PROCEDURE SGPLOT used (Total process time):

real time	0.10 seconds
user cpu time	0.05 seconds
system cpu time	0.02 seconds
memory	8304.09k
OS Memory	35120.00k
Timestamp	08/12/2023 02:16:18 AM
Step Count	811 Switch Count 1

Page Faults	0
Page Reclaims	1636
Page Swaps	0
Voluntary Context Switches	186
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	720

NOTE: There were 2500 observations read from the data set WORK.PROJECT2.

```

103
104      /*Histogram of the variable monthly_revenue*/
105      proc sgplot data=work.project2;
106          histogram Monthly_Revenue;
107          density Monthly_Revenue;
108      run;

```

NOTE: PROCEDURE SGPLOT used (Total process time):

real time	0.08 seconds
user cpu time	0.04 seconds
system cpu time	0.00 seconds
memory	2487.68k
OS Memory	35376.00k
Timestamp	08/12/2023 02:16:18 AM
Step Count	812
Switch Count	1
Page Faults	0
Page Reclaims	399
Page Swaps	0
Voluntary Context Switches	166
Involuntary Context Switches	1
Block Input Operations	0
Block Output Operations	440

NOTE: There were 2500 observations read from the data set WORK.PROJECT2.

```

109
110      title "outliers affect";
111      proc sgplot data=work.project2;
112          scatter x=country y=monthly_revenue;
113          xaxis grid;
114          yaxis grid;
115      run;

```

NOTE: PROCEDURE SGPLOT used (Total process time):

real time	0.38 seconds
user cpu time	0.08 seconds
system cpu time	0.01 seconds
memory	2080.71k
OS Memory	35376.00k
Timestamp	08/12/2023 02:16:18 AM
Step Count	813
Switch Count	1
Page Faults	0
Page Reclaims	293
Page Swaps	0
Voluntary Context Switches	7622
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	384

NOTE: There were 2500 observations read from the data set WORK.PROJECT2.

```

116
117      Title 'Q-Q plot';
118      proc univariate data = work.project2;
119          ppplot monthly_revenue;
120      run;

```

NOTE: PROCEDURE UNIVARIATE used (Total process time):

real time	0.21 seconds
user cpu time	0.15 seconds
system cpu time	0.01 seconds
memory	9128.25k
OS Memory	41728.00k
Timestamp	08/12/2023 02:16:18 AM
Step Count	814
Switch Count	0
Page Faults	0
Page Reclaims	2051
Page Swaps	0
Voluntary Context Switches	130
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	416



```

121
122      /*Check Distribution and Choose Outlier Detection Method*/;
123      proc univariate data=work.project2;
124          var monthly_revenue;
125          histogram monthly_revenue / normal; /* Visualize distribution */
126      run;

```

NOTE: PROCEDURE UNIVARIATE used (Total process time):

```

real time      0.22 seconds
user cpu time   0.15 seconds
system cpu time 0.01 seconds
memory         7436.81k
OS Memory      40708.00k
Timestamp      08/12/2023 02:16:19 AM
Step Count     815  Switch Count  0
Page Faults    0
Page Reclaims 1548
Page Swaps     0
Voluntary Context Switches 787
Involuntary Context Switches 0
Block Input Operations 0
Block Output Operations 416

```

```

127
128      title "Output the outliers for the variable monthly_revenue based on the interquantile range method";
129      proc means data=work.project2 noprint;
130          var monthly_revenue;
131          output out=revenue
132              Q1=
133              Q3=
134              QRange= / autoname;
135      run;

```

NOTE: There were 2500 observations read from the data set WORK.PROJECT2.

NOTE: The data set WORK.REVENUE has 1 observations and 5 variables.

NOTE: PROCEDURE MEANS used (Total process time):

```

real time      0.00 seconds
user cpu time   0.00 seconds
system cpu time 0.00 seconds
memory         7422.93k
OS Memory      41680.00k
Timestamp      08/12/2023 02:16:19 AM
Step Count     816  Switch Count  3
Page Faults    0
Page Reclaims 1672
Page Swaps     0
Voluntary Context Switches 36
Involuntary Context Switches 0
Block Input Operations 0
Block Output Operations 264

```

```

136
137      data null;
138          file print;
139          set work.project2(keep=monthly_revenue);
140          if n = 1 then set revenue;
141          if monthly_revenue le monthly_revenue_Q1 - 1.5*monthly_revenue_QRange and not missing(monthly_revenue) or
142             monthly_revenue ge monthly_revenue_Q3 + 1.5*monthly_revenue_QRange then
143              put "Possible Outlier for instant " instant "Value of monthly_revenue is " monthly_revenue;
144      run;

```

NOTE: Variable n is uninitialized.

NOTE: Variable instant is uninitialized.

NOTE: 2500 lines were written to file PRINT.

NOTE: Missing values were generated as a result of performing an operation on missing values.

Each place is given by: (Number of times) at (Line):(Column).  
 2500 at 141:45 2500 at 141:50 2500 at 142:45 2500 at 142:50

NOTE: There were 2500 observations read from the data set WORK.PROJECT2.

NOTE: The data set WORK.NULL has 2500 observations and 8 variables.

NOTE: DATA statement used (Total process time):

```

real time      0.26 seconds
user cpu time   0.26 seconds
system cpu time 0.00 seconds
memory         1755.40k
OS Memory      35760.00k
Timestamp      08/12/2023 02:16:19 AM
Step Count     817  Switch Count  2
Page Faults    0

```

Page Reclaims	144
Page Swaps	0
Voluntary Context Switches	15
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	1208

```

145
146     title 'Dataset without the outliers';
147     /* Step 1: Calculate the interquartile range (IQR) for the variable "monthly_revenue" */
148     proc means data=work.project2 noprint;
149         var monthly_revenue;
150         output out=revenue
151             Q1=monthly_revenue_Q1
152             Q3=monthly_revenue_Q3
153             QRange=monthly_revenue_QRange / autoname;
154     run;

```

NOTE: There were 2500 observations read from the data set WORK.PROJECT2.

NOTE: The data set WORK.REVENUE has 1 observations and 5 variables.

NOTE: PROCEDURE MEANS used (Total process time):

real time	0.00 seconds
user cpu time	0.00 seconds
system cpu time	0.00 seconds
memory	7310.43k
OS Memory	41936.00k
Timestamp	08/12/2023 02:16:19 AM
Step Count	818
Page Faults	0
Page Reclaims	1693
Page Swaps	0
Voluntary Context Switches	34
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	264

```

155
156     /* Step 2: Identify and remove the outliers */
157     data revenue_OutliersRemoved;
158         set work.project2;
159         if _n_ = 1 then set revenue; /* Read the IQR values from the temporary dataset */
160
161         /* Check if the value of 'monthly_revenue' is within the acceptable range (Q1 - 1.5*IQR, Q3 + 1.5*IQR) */
162         if monthly_revenue >= monthly_revenue_Q1 - 1.5 * monthly_revenue_QRange and monthly_revenue <= monthly_revenue_Q3 +
162         ! 1.5 * monthly_revenue_QRange and not missing(monthly_revenue);
163     run;

```

NOTE: There were 2500 observations read from the data set WORK.PROJECT2.

NOTE: There were 1 observations read from the data set WORK.REVENUE.

NOTE: The data set WORK.REVENUE\_OUTLIERSREMOVED has 2457 observations and 14 variables.

NOTE: DATA statement used (Total process time):

real time	0.00 seconds
user cpu time	0.00 seconds
system cpu time	0.01 seconds
memory	1810.93k
OS Memory	36016.00k
Timestamp	08/12/2023 02:16:19 AM
Step Count	819
Page Faults	0
Page Reclaims	134
Page Swaps	0
Voluntary Context Switches	11
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	784

```

164
165     /* Step 3: Print the dataset without the outliers */
166     proc print data=revenue_OutliersRemoved (obs=30);
167     run;

```

NOTE: There were 30 observations read from the data set WORK.REVENUE\_OUTLIERSREMOVED.

NOTE: PROCEDURE PRINT used (Total process time):

real time	0.08 seconds
user cpu time	0.08 seconds
system cpu time	0.00 seconds
memory	988.46k
OS Memory	35624.00k
Timestamp	08/12/2023 02:16:19 AM

Step Count	820	Switch Count	0
Page Faults	0		
Page Reclaims	97		
Page Swaps	0		
Voluntary Context Switches	1		
Involuntary Context Switches	1		
Block Input Operations	0		
Block Output Operations	40		

```

168
169
170      /*Check to ensure outliers were removed using the histogram plot*/
171      /*Histogram without outliers*/
172      title "Histogram (with outliers)";
173      proc sgplot data=work.project2;
174          histogram Monthly_Revenue;
175          density Monthly_Revenue / type=normal;
176      run;

```

NOTE: PROCEDURE SGPLOT used (Total process time):

real time	0.09 seconds
user cpu time	0.05 seconds
system cpu time	0.00 seconds
memory	2298.21k
OS Memory	36400.00k
Timestamp	08/12/2023 02:16:19 AM
Step Count	821
Page Faults	0
Page Reclaims	325
Page Swaps	0
Voluntary Context Switches	166
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	440

NOTE: There were 2500 observations read from the data set WORK.PROJECT2.

```

177
178      /*Histogram without outliers*/
179      title "Histogram (without outliers)";
180      proc sgplot data=revenue_OutliersRemoved;
181          histogram Monthly_Revenue;
182          density Monthly_Revenue / type=normal;
183      run;

```

NOTE: PROCEDURE SGPLOT used (Total process time):

real time	0.08 seconds
user cpu time	0.05 seconds
system cpu time	0.00 seconds
memory	2623.34k
OS Memory	36528.00k
Timestamp	08/12/2023 02:16:19 AM
Step Count	822
Page Faults	0
Page Reclaims	340
Page Swaps	0
Voluntary Context Switches	167
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	456

NOTE: There were 2457 observations read from the data set WORK.REVENUE\_OUTLIERSREMOVED.

```

184
185      /*Check to ensure outliers were removed using the Q-Q plot*/
186      title "Box plot from revenue dataset";
187      proc sgplot data=revenue_OutliersRemoved;
188          vbox monthly_revenue;
189      run;

```

NOTE: PROCEDURE SGPLOT used (Total process time):

real time	0.09 seconds
user cpu time	0.04 seconds
system cpu time	0.00 seconds
memory	2598.46k
OS Memory	36528.00k
Timestamp	08/12/2023 02:16:19 AM
Step Count	823
Page Faults	0
Page Reclaims	327
Page Swaps	0

```

Voluntary Context Switches      164
Involuntary Context Switches    0
Block Input Operations           0
Block Output Operations          400

```

NOTE: There were 2457 observations read from the data set WORK.REVENUE\_OUTLIERSREMOVED.

```

190
191      Title 'Q-Q plot with outliers';
192      proc univariate data = work.project2;
193      ppplot monthly_revenue;
194      run;

```

NOTE: PROCEDURE UNIVARIATE used (Total process time):

```

real time      0.20 seconds
user cpu time   0.15 seconds
system cpu time 0.00 seconds
memory         8676.68k
OS Memory      42240.00k
Timestamp      08/12/2023 02:16:20 AM
Step Count     824  Switch Count  0
Page Faults    0
Page Reclaims 1931
Page Swaps     0
Voluntary Context Switches      130
Involuntary Context Switches    0
Block Input Operations           0
Block Output Operations          424

```

```

195
196      Title 'Q-Q plot without outliers';
197      proc univariate data = revenue_OutliersRemoved;
198      ppplot monthly_revenue;
199      run;

```

NOTE: PROCEDURE UNIVARIATE used (Total process time):

```

real time      0.41 seconds
user cpu time   0.35 seconds
system cpu time 0.00 seconds
memory         8732.59k
OS Memory      42368.00k
Timestamp      08/12/2023 02:16:20 AM
Step Count     825  Switch Count  0
Page Faults    0
Page Reclaims 1965
Page Swaps     0
Voluntary Context Switches      130
Involuntary Context Switches    1
Block Input Operations           0
Block Output Operations          512

```

```

200
201      /*
202
203      /*AGE OUTLIERS */
204      /* Outlier detection for some variables using box plots */;
205      proc sgplot data=work.project2;
206      vbox age;
207      run;

```

NOTE: PROCEDURE SGPLOT used (Total process time):

```

real time      0.09 seconds
user cpu time   0.04 seconds
system cpu time 0.01 seconds
memory         2296.62k
OS Memory      36912.00k
Timestamp      08/12/2023 02:16:20 AM
Step Count     826  Switch Count  1
Page Faults    0
Page Reclaims 317
Page Swaps     0
Voluntary Context Switches      171
Involuntary Context Switches    0
Block Input Operations           0
Block Output Operations          424

```

NOTE: There were 2500 observations read from the data set WORK.PROJECT2.

```

208
209      /*Histogram of the variable age*/

```

```

210      proc sgplot data=work.project2;
211          histogram age;
212          density age;
213      run;

```

NOTE: PROCEDURE SGPLOT used (Total process time):

```

real time      0.10 seconds
user cpu time   0.04 seconds
system cpu time 0.00 seconds
memory         2479.59k
OS Memory      36912.00k
Timestamp      08/12/2023 02:16:20 AM
Step Count     827  Switch Count  1
Page Faults    0
Page Reclaims  307
Page Swaps     0
Voluntary Context Switches 167
Involuntary Context Switches 0
Block Input Operations 0
Block Output Operations 440

```

NOTE: There were 2500 observations read from the data set WORK.PROJECT2.

```

214
215      Title 'Q-Q plot';
216      proc univariate data = work.project2;
217          ppplot age;
218      run;

```

NOTE: PROCEDURE UNIVARIATE used (Total process time):

```

real time      0.20 seconds
user cpu time   0.15 seconds
system cpu time 0.00 seconds
memory         8670.21k
OS Memory      42496.00k
Timestamp      08/12/2023 02:16:20 AM
Step Count     828  Switch Count  0
Page Faults    0
Page Reclaims  1886
Page Swaps     0
Voluntary Context Switches 132
Involuntary Context Switches 0
Block Input Operations 0
Block Output Operations 432

```

```

219
220      /*Check Distribution and Choose Outlier Detection Method*/;
221      proc univariate data=work.project2;
222          var age;
223          histogram age / normal; /* Visualize distribution */
224      run;

```

NOTE: PROCEDURE UNIVARIATE used (Total process time):

```

real time      0.21 seconds
user cpu time   0.14 seconds
system cpu time 0.01 seconds
memory         7487.12k
OS Memory      41220.00k
Timestamp      08/12/2023 02:16:21 AM
Step Count     829  Switch Count  0
Page Faults    0
Page Reclaims  1462
Page Swaps     0
Voluntary Context Switches 787
Involuntary Context Switches 0
Block Input Operations 0
Block Output Operations 400

```

```

225
226      title "Output the outliers for the variable monthly_revenue based on the interquantile range method";
227      proc means data=work.project2 noprint;
228          var age;
229          output out=Age_range
230              Q1=
231              Q3=
232              QRange= / autoname;
233      run;

```

NOTE: There were 2500 observations read from the data set WORK.PROJECT2.

NOTE: The data set WORK.AGE\_RANGE has 1 observations and 5 variables.

NOTE: PROCEDURE MEANS used (Total process time):

real time	0.00 seconds
user cpu time	0.01 seconds
system cpu time	0.01 seconds
memory	7422.12k
OS Memory	42704.00k
Timestamp	08/12/2023 02:16:21 AM
Step Count	830 Switch Count 3
Page Faults	0
Page Reclaims	1673
Page Swaps	0
Voluntary Context Switches	36
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	264

```

234
235     data null;
236         file print;
237         set work.project2(keep=monthly_revenue);
238         if n = 1 then set age_range;
239         if age le age_Q1 - 1.5*age_QRange and not missing(age) or
240             age ge age_Q3 + 1.5*age_QRange then
241             put "Possible Outlier for instant " instant "Value of monthly_revenue is " age;
242     run;

```

NOTE: Variable n is uninitialized.

NOTE: Variable age is uninitialized.

NOTE: Variable ageQ3 is uninitialized.

NOTE: Variable instant is uninitialized.

NOTE: 2500 lines were written to file PRINT.

NOTE: Missing values were generated as a result of performing an operation on missing values.

Each place is given by: (Number of times) at (Line):(Column).  
 2500 at 239:21 2500 at 239:26 2500 at 240:20 2500 at 240:25

NOTE: There were 2500 observations read from the data set WORK.PROJECT2.

NOTE: The data set WORK.NULL has 2500 observations and 10 variables.

NOTE: DATA statement used (Total process time):

real time	0.35 seconds
user cpu time	0.36 seconds
system cpu time	0.00 seconds
memory	1746.59k
OS Memory	36784.00k
Timestamp	08/12/2023 02:16:21 AM
Step Count	831 Switch Count 2
Page Faults	0
Page Reclaims	137
Page Swaps	0
Voluntary Context Switches	13
Involuntary Context Switches	1
Block Input Operations	0
Block Output Operations	1200

```

243
244     title 'Dataset without the outliers';
245     /* Step 1: Calculate the interquartile range (IQR) for the variable age */
246     proc means data=work.project2 noprint;
247         var age;
248         output out=age_range
249             Q1=age_Q1
250             Q3=age_Q3
251             QRange=age_QRange / autoname;
252     run;

```

NOTE: There were 2500 observations read from the data set WORK.PROJECT2.

NOTE: The data set WORK.AGE\_RANGE has 1 observations and 5 variables.

NOTE: PROCEDURE MEANS used (Total process time):

real time	0.00 seconds
user cpu time	0.01 seconds
system cpu time	0.00 seconds
memory	7310.00k
OS Memory	42704.00k
Timestamp	08/12/2023 02:16:21 AM
Step Count	832 Switch Count 3
Page Faults	0
Page Reclaims	1675
Page Swaps	0
Voluntary Context Switches	32
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	264

```

253
254      /* Step 2: Identify and remove the outliers */
255      data age_range_OutliersRemoved;
256          set work.project2;
257          if _n_ = 1 then set age_range; /* Read the IQR values from the temporary dataset */
258
259          /* Check if the value of 'age' is within the acceptable range (Q1 - 1.5*IQR, Q3 + 1.5*IQR) */
260          if age >= age_Q1 - 1.5 * age_QRange and age <= age_Q3 + 1.5 * age_QRange and not missing(age);
261      run;

```

NOTE: There were 2500 observations read from the data set WORK.PROJECT2.

NOTE: There were 1 observations read from the data set WORK.AGE\_RANGE.

NOTE: The data set WORK.AGE\_RANGE\_OUTLIERSREMOVED has 2497 observations and 14 variables.

NOTE: DATA statement used (Total process time):

real time	0.00 seconds
user cpu time	0.00 seconds
system cpu time	0.00 seconds
memory	1698.90k
OS Memory	36784.00k
Timestamp	08/12/2023 02:16:21 AM
Step Count	833 Switch Count 2
Page Faults	0
Page Reclaims	131
Page Swaps	0
Voluntary Context Switches	14
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	776

```

262
263      /* Step 3: Print the dataset without the outliers */
264      proc print data=age_range_OutliersRemoved (obs=30);
265      run;

```

NOTE: There were 30 observations read from the data set WORK.AGE\_RANGE\_OUTLIERSREMOVED.

NOTE: PROCEDURE PRINT used (Total process time):

real time	0.08 seconds
user cpu time	0.09 seconds
system cpu time	0.00 seconds
memory	874.59k
OS Memory	36392.00k
Timestamp	08/12/2023 02:16:21 AM
Step Count	834 Switch Count 0
Page Faults	0
Page Reclaims	100
Page Swaps	0
Voluntary Context Switches	0
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	32

```

266
267
268      /*Check to ensure outliers were removed using the histogram plot*/
269      /*Histogram without outliers*/
270      title "Histogram (with outliers)";
271      proc sgplot data=work.project2;
272          histogram age;
273          density age / type=normal;
274      run;

```

NOTE: PROCEDURE SGPLOT used (Total process time):

real time	0.09 seconds
user cpu time	0.04 seconds
system cpu time	0.01 seconds
memory	2309.03k
OS Memory	37168.00k
Timestamp	08/12/2023 02:16:21 AM
Step Count	835 Switch Count 1
Page Faults	0
Page Reclaims	306
Page Swaps	0
Voluntary Context Switches	166
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	456

NOTE: There were 2500 observations read from the data set WORK.PROJECT2.

```

275
276      /*Histogram without outliers*/
277      title "Histogram (without outliers)";
278      proc sgplot data=age_range_OutliersRemoved;
279          histogram age;
280          density age / type=normal;
281      run;

```

NOTE: PROCEDURE SGPLOT used (Total process time):

```

real time          0.11 seconds
user cpu time      0.05 seconds
system cpu time    0.00 seconds
memory            2605.46k
OS Memory          37296.00k
Timestamp          08/12/2023 02:16:21 AM
Step Count         836  Switch Count  1
Page Faults        0
Page Reclaims      339
Page Swaps         0
Voluntary Context Switches 166
Involuntary Context Switches 0
Block Input Operations 0
Block Output Operations 472

```

NOTE: There were 2497 observations read from the data set WORK.AGE\_RANGE\_OUTLIERSREMOVED.

```

282
283      /*Check to ensure outliers were removed using the Q-Q plot*/
284      title "Box plot from revenue dataset";
285      proc sgplot data=age_range_OutliersRemoved;
286          vbox age;
287      run;

```

NOTE: PROCEDURE SGPLOT used (Total process time):

```

real time          0.08 seconds
user cpu time      0.04 seconds
system cpu time    0.01 seconds
memory            2600.03k
OS Memory          37296.00k
Timestamp          08/12/2023 02:16:21 AM
Step Count         837  Switch Count  1
Page Faults        0
Page Reclaims      326
Page Swaps         0
Voluntary Context Switches 164
Involuntary Context Switches 0
Block Input Operations 0
Block Output Operations 408

```

NOTE: There were 2497 observations read from the data set WORK.AGE\_RANGE\_OUTLIERSREMOVED.

```

288
289      Title 'Q-Q plot with outliers';
290      proc univariate data = work.project2;
291          ppplot age;
292      run;

```

NOTE: PROCEDURE UNIVARIATE used (Total process time):

```

real time          0.21 seconds
user cpu time      0.17 seconds
system cpu time    0.01 seconds
memory            8621.00k
OS Memory          42496.00k
Timestamp          08/12/2023 02:16:22 AM
Step Count         838  Switch Count  0
Page Faults        0
Page Reclaims      1806
Page Swaps         0
Voluntary Context Switches 129
Involuntary Context Switches 0
Block Input Operations 0
Block Output Operations 432

```

```

293
294      Title 'Q-Q plot without outliers';
295      proc univariate data = age_range_OutliersRemoved;
296          ppplot age;
297      run;

```

NOTE: PROCEDURE UNIVARIATE used (Total process time):



```
real time          0.51 seconds
user cpu time      0.47 seconds
system cpu time    0.00 seconds
memory             8611.31k
OS Memory          42624.00k
Timestamp          08/12/2023 02:16:22 AM
Step Count         839  Switch Count  0
Page Faults        0
Page Reclaims      1894
Page Swaps         0
Voluntary Context Switches 128
Involuntary Context Switches 2
Block Input Operations 0
Block Output Operations 496
```

```
298
299
300
301
302
303      OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
313
```