Great Learning SAS Assessment 2
Ishita Sarkar

1.Import dataset in the SAS environment

```
FILENAME REFFILE '/home/u62305191/Datasets/Life+Insurance+Dataset.csv';

PROC IMPORT DATAFILE=REFFILE
        DBMS=CSV
        OUT=SAS.DataInsurance;
        GETNAMES=YES;
RUN;

PROC CONTENTS DATA=SAS.DataInsurance; RUN;
```

Check top 10 record of import dataset

```
data Top10_Records;
        set SAS.DataInsurance(obs=10);
run;
```

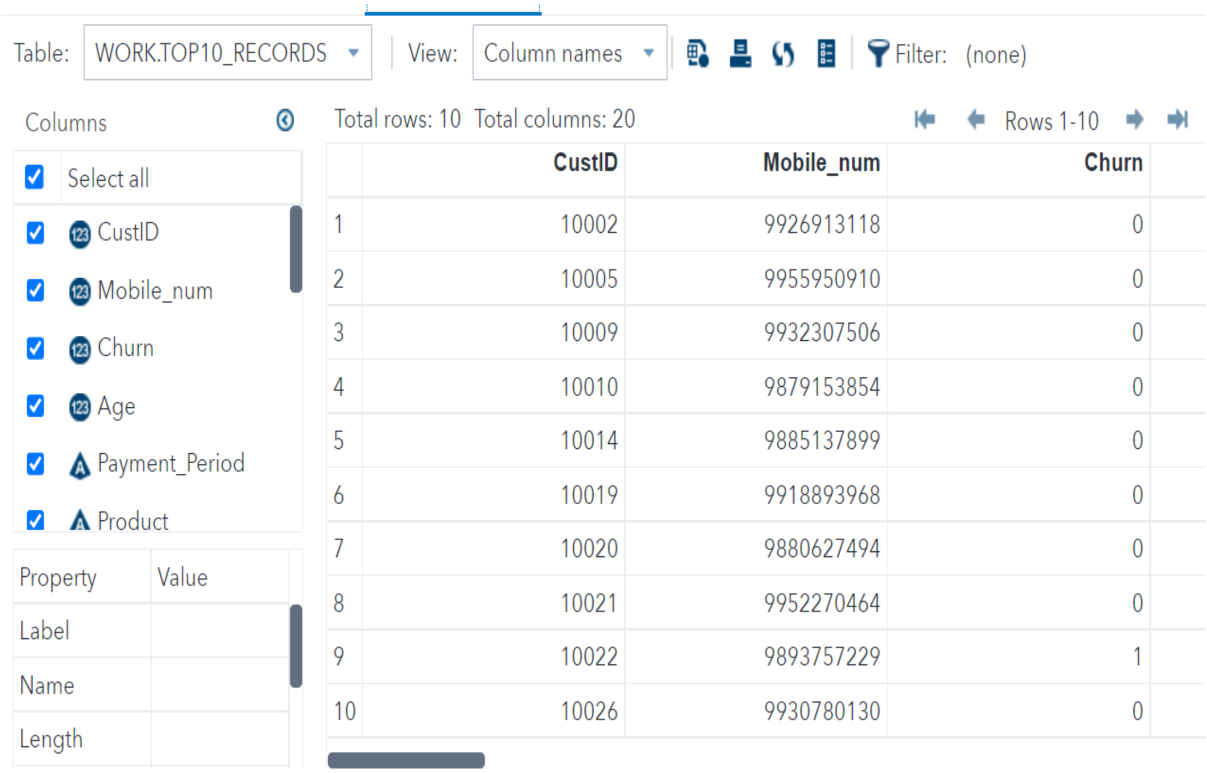


Table: WORK.TOP10_RECORDS   View: Column names   Filter: (none)

Columns

☑ Select all
☑ CustID
☑ Mobile_num
☑ Churn
☑ Age
☑ Payment_Period
☑ Product

Property	Value
Label	
Name	
Length	

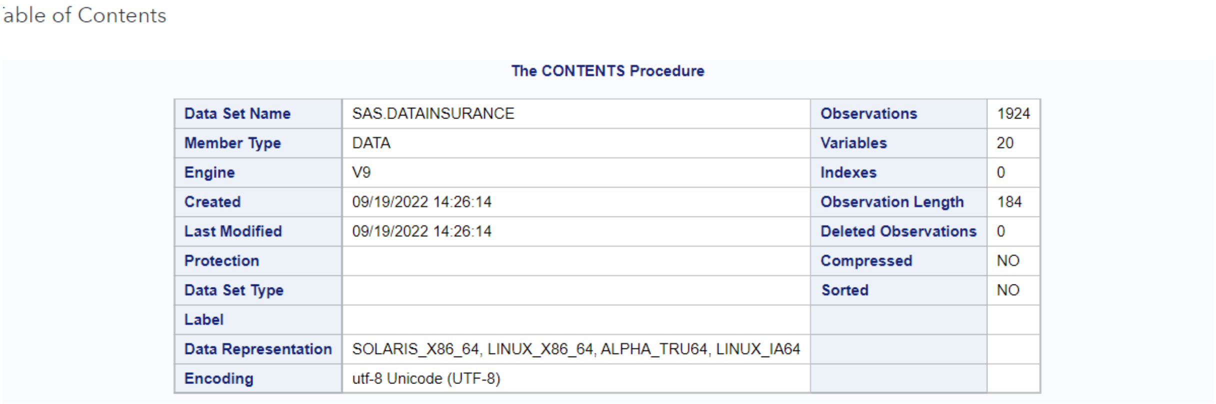
Total rows: 10 Total columns: 20   Rows 1-10

	CustID	Mobile_num	Churn
1	10002	9926913118	0
2	10005	9955950910	0
3	10009	9932307506	0
4	10010	9879153854	0
5	10014	9885137899	0
6	10019	9918893968	0
7	10020	9880627494	0
8	10021	9952270464	0
9	10022	9893757229	1
10	10026	9930780130	0

Great Learning SAS Assessment 2
Ishita Sarkar

2.Check variable type of the import dataset

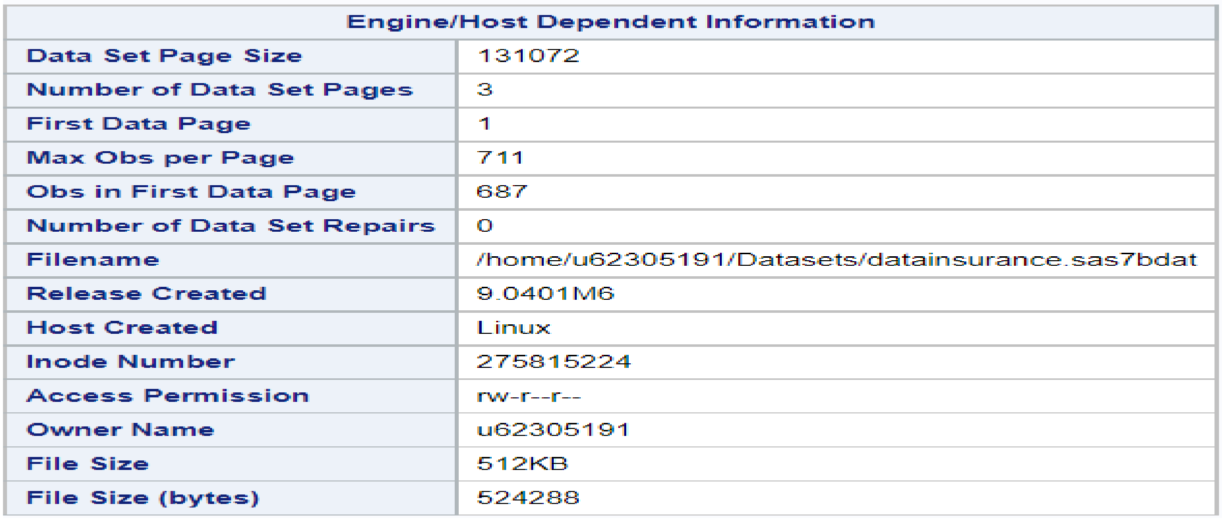
```
proc contents data=SAS.DataInsurance;
run;
```



able of Contents

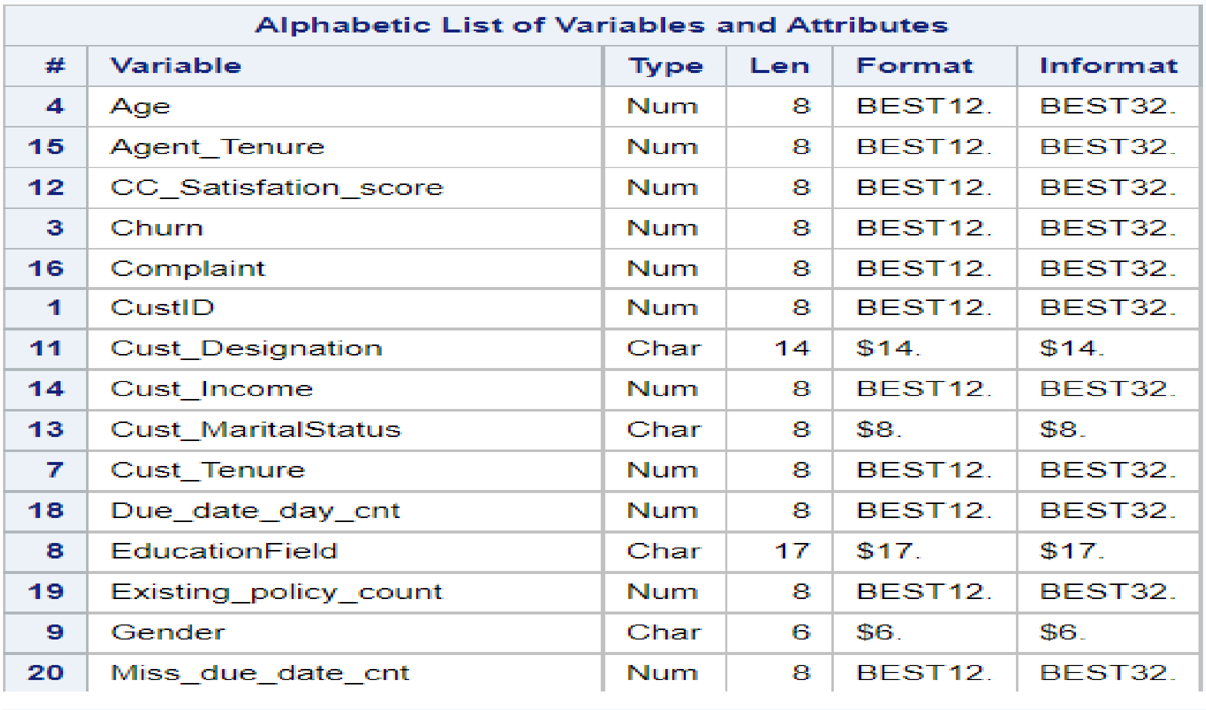
The CONTENTS Procedure

Data Set Name	SAS.DATAINSURANCE	Observations	1924
Member Type	DATA	Variables	20
Engine	V9	Indexes	0
Created	09/19/2022 14:26:14	Observation Length	184
Last Modified	09/19/2022 14:26:14	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	3
First Data Page	1
Max Obs per Page	711
Obs in First Data Page	687
Number of Data Set Repairs	0
Filename	/home/u62305191/Datasets/datainsurance.sas7bdat
Release Created	9.0401M6
Host Created	Linux
Inode Number	275815224
Access Permission	rw-r--r--
Owner Name	u62305191
File Size	512KB
File Size (bytes)	524288

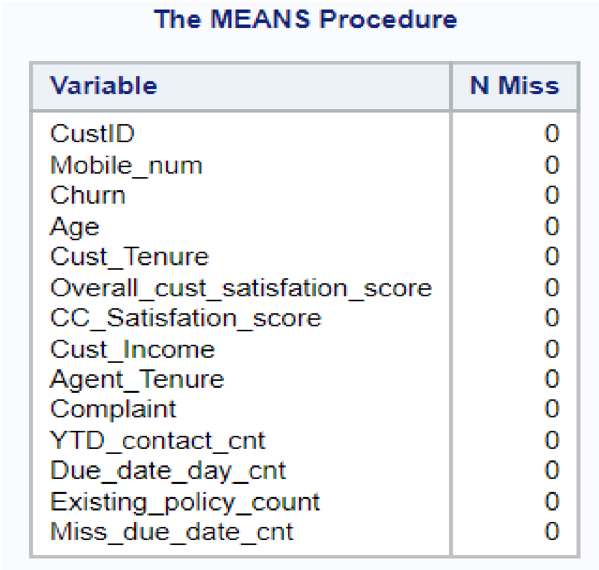
Great Learning SAS Assessment 2
Ishita Sarkar



Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
4	Age	Num	8	BEST12.	BEST32.
15	Agent_Tenure	Num	8	BEST12.	BEST32.
12	CC_Satisfation_score	Num	8	BEST12.	BEST32.
3	Churn	Num	8	BEST12.	BEST32.
16	Complaint	Num	8	BEST12.	BEST32.
1	CustID	Num	8	BEST12.	BEST32.
11	Cust_Designation	Char	14	$14.	$14.
14	Cust_Income	Num	8	BEST12.	BEST32.
13	Cust_MaritalStatus	Char	8	$8.	$8.
7	Cust_Tenure	Num	8	BEST12.	BEST32.
18	Due_date_day_cnt	Num	8	BEST12.	BEST32.
8	EducationField	Char	17	$17.	$17.
19	Existing_policy_count	Num	8	BEST12.	BEST32.
9	Gender	Char	6	$6.	$6.
20	Miss_due_date_cnt	Num	8	BEST12.	BEST32.

3.Checks if any variables have missing values, if yes then do treatment?

```
proc means data=SAS.DataInsurance nmiss;
run;
```



The MEANS Procedure

Variable	N Miss
CustID	0
Mobile_num	0
Churn	0
Age	0
Cust_Tenure	0
Overall_cust_satisfation_score	0
CC_Satisfation_score	0
Cust_Income	0
Agent_Tenure	0
Complaint	0
YTD_contact_cnt	0
Due_date_day_cnt	0
Existing_policy_count	0
Miss_due_date_cnt	0

4.Check summary and percentile distribution of all numerical variables for churners and non-churners?

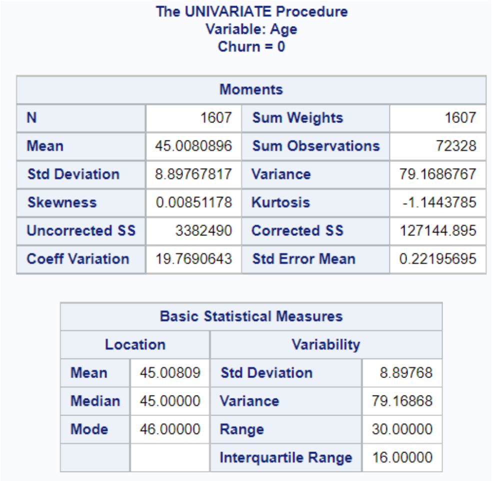
```
proc univariate data= SAS.DataInsurance;

var Age Cust_Tenure Overall_cust_satisfation_score CC_Satisfation_score Cust_Income Agent_Tenure
```

Great Learning SAS Assessment 2
Ishita Sarkar

YTD_contact_cnt Due_date_day_cnt Existing_policy_count Miss_due_date_cnt;

class churn;

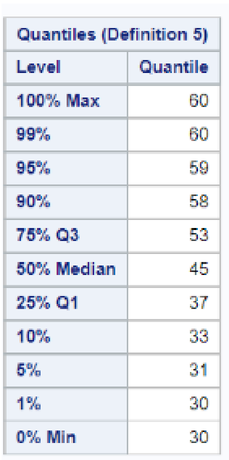
run;



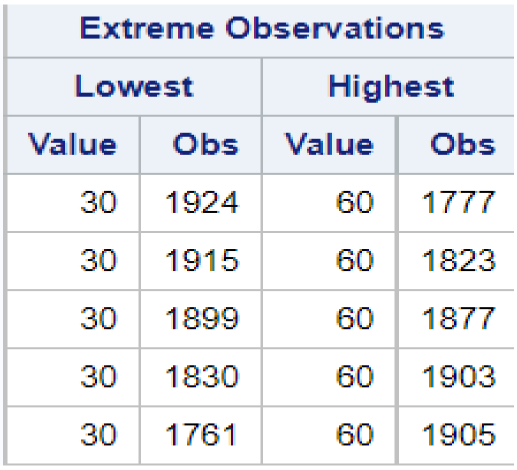
The UNIVARIATE Procedure
Variable: Age
Churn = 0

Moments			
N	1607	Sum Weights	1607
Mean	45.0080896	Sum Observations	72328
Std Deviation	8.89767817	Variance	79.1686767
Skewness	0.00851178	Kurtosis	-1.1443785
Uncorrected SS	3382490	Corrected SS	127144.895
Coeff Variation	19.7690643	Std Error Mean	0.22195695

Basic Statistical Measures			
Location		Variability	
Mean	45.00809	Std Deviation	8.89768
Median	45.00000	Variance	79.16868
Mode	46.00000	Range	30.00000
		Interquartile Range	16.00000



Quantiles (Definition 5)	
Level	Quantile
100% Max	60
99%	60
95%	59
90%	58
75% Q3	53
50% Median	45
25% Q1	37
10%	33
5%	31
1%	30
0% Min	30



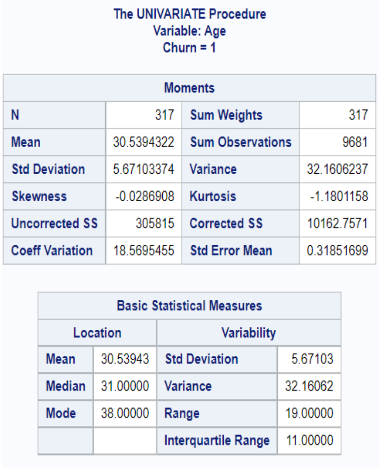
Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
30	1924	60	1777
30	1915	60	1823
30	1899	60	1877
30	1830	60	1903
30	1761	60	1905

proc means data= SAS.DataInsurance n nmiss min p1 p5 p10 p25 p50 p75 p90 p95 p99 max;

var Age Cust_Tenure Overall_cust_satisfation_score CC_Satisfation_score Cust_Income Agent_Tenure

YTD_contact_cnt Due_date_day_cnt Existing_policy_count Miss_due_date_cnt;

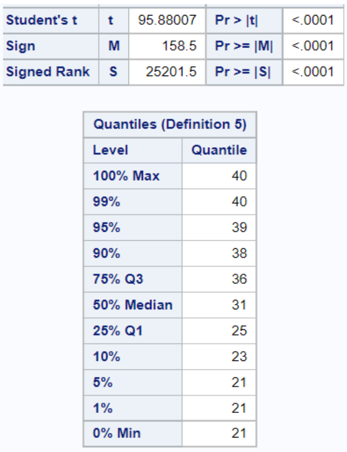
run;



The UNIVARIATE Procedure
Variable: Age
Churn = 1

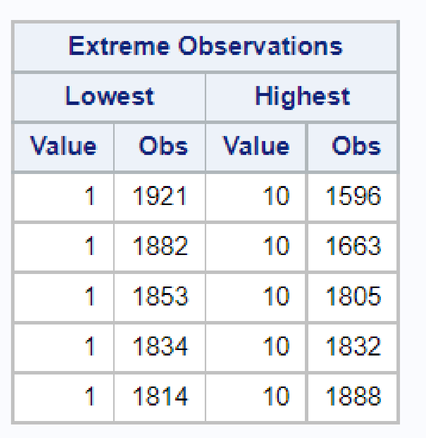
Moments			
N	317	Sum Weights	317
Mean	30.5394322	Sum Observations	9681
Std Deviation	5.67103374	Variance	32.1606237
Skewness	-0.0286908	Kurtosis	-1.1801158
Uncorrected SS	305815	Corrected SS	10162.7571
Coeff Variation	18.5695455	Std Error Mean	0.31851699

Basic Statistical Measures			
Location		Variability	
Mean	30.53943	Std Deviation	5.67103
Median	31.00000	Variance	32.16062
Mode	38.00000	Range	19.00000
		Interquartile Range	11.00000



Student's t	t	95.88007	Pr > |t|	<.0001
Sign	M	158.5	Pr >= |M|	<.0001
Signed Rank	S	25201.5	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	40
99%	40
95%	39
90%	38
75% Q3	36
50% Median	31
25% Q1	25
10%	23
5%	21
1%	21
0% Min	21



Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1	1921	10	1596
1	1882	10	1663
1	1853	10	1805
1	1834	10	1832
1	1814	10	1888

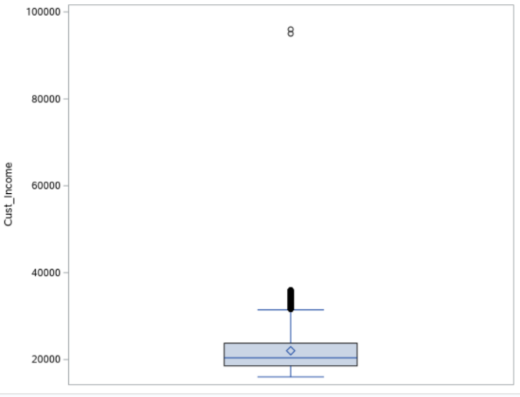
5.Check for outlier, if yes then do treatment?

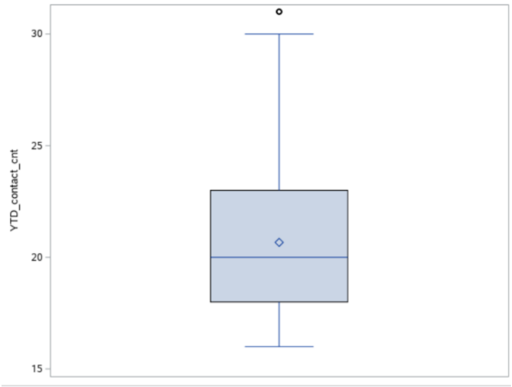
Great Learning SAS Assessment 2
Ishita Sarkar

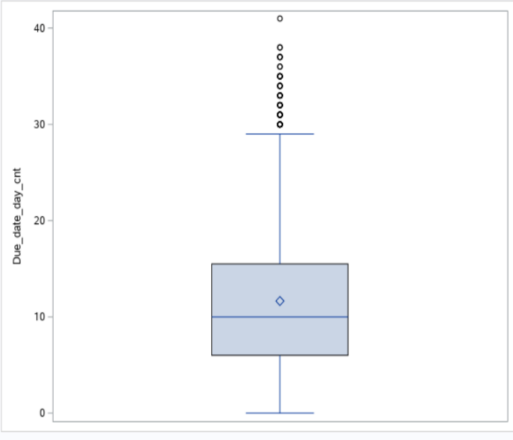
```
proc sgplot data= SAS.DataInsurance;
vbox Age;
run;

proc sgplot data= SAS.DataInsurance;
vbox Cust_Tenure;
run;

proc sgplot data = SAS.DataInsurance;
vbox Overall_cust_satisfation_score;
run;

proc sgplot data = SAS.DataInsurance;
vbox CC_Satisfation_score;
run;

proc sgplot data = SAS.DataInsurance;
vbox Cust_Income;
run;

proc sgplot data = SAS.DataInsurance;
vbox Agent_Tenure;
run;

proc sgplot data = SAS.DataInsurance;
vbox YTD_contact_cnt;
run;

proc sgplot data = SAS.DataInsurance;
vbox Due_date_day_cnt;
run;

proc sgplot data = SAS.DataInsurance;
vbox Existing_policy_count;
run;

proc sgplot data = SAS.DataInsurance;
vbox Miss_due_date_cnt;
run;

proc univariate data= SAS.DataInsurance;
```

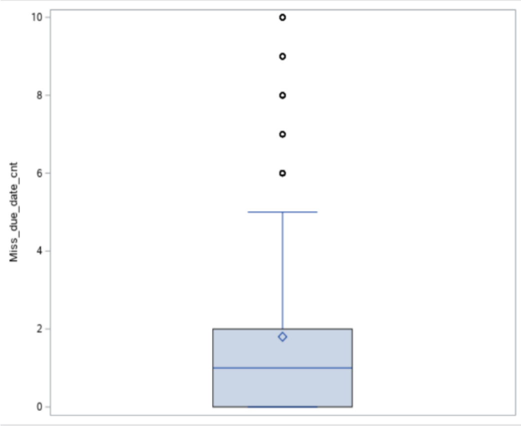
Great Learning SAS Assessment 2
Ishita Sarkar

```
var Age Cust_Tenure Overall_cust_satisfation_score CC_Satisfation_score Cust_Income
Agent_Tenure

 YTD_contact_cnt Due_date_day_cnt Existing_policy_count Miss_due_date_cnt;
run;

/*there are outliers in
1).Miss_due_date_cnt,
2).Due_date_day_cnt,
3).YTD_contact_cnt,
4).Cust_Income;
thus we will be using flooring and capping techniques for these variables*/

data insure;
   set SAS.DataInsurance;
   if Cust_Income >  31585.5 then Cust_Income = 31585.5;
   if YTD_contact_cnt > 30.5 then YTD_contact_cnt = 30.5;
   if Due_date_day_cnt > 29.75 then Due_date_day_cnt = 29.75 ;
   if Miss_due_date_cnt > 5 then Miss_due_date_cnt = 5;
run;
```





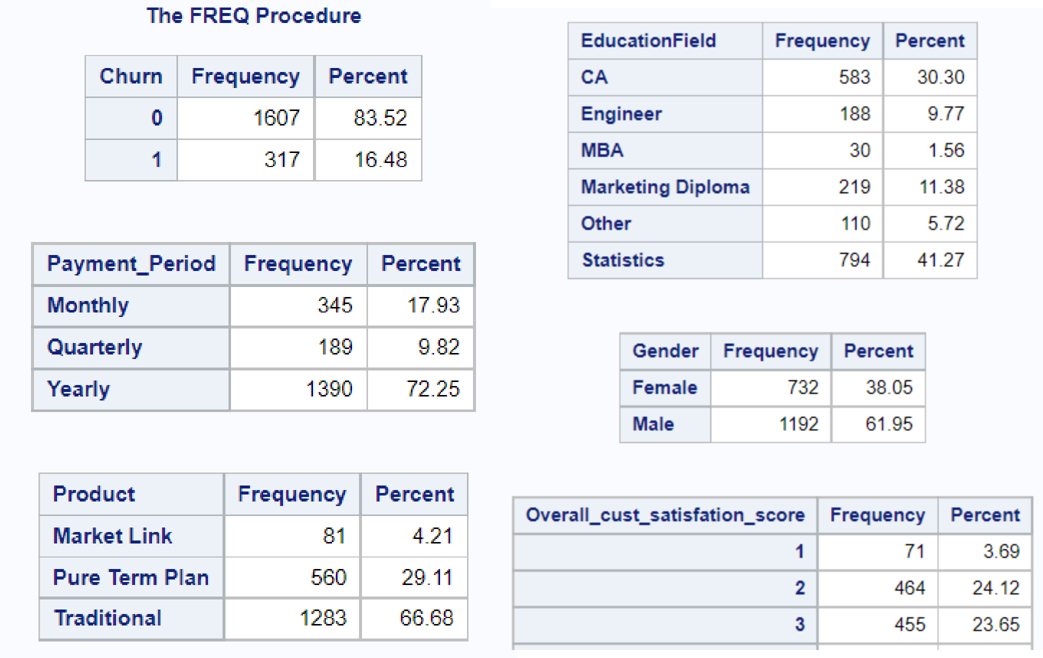




Great Learning SAS Assessment 2
Ishita Sarkar

6.Check the proportion of all categorical variables and extract percentage contribution of each class in respective variables?

```
proc freq data = insure;
table
churn
Payment_period
Product
EducationField
Gender
Overall_cust_satisfation_score
Cust_Designation
CC_Satisfation_score
Cust_MaritalStatus
Complaint
/ nocum;
run;
```



The FREQ Procedure

Churn	Frequency	Percent
0	1607	83.52
1	317	16.48

Payment_Period	Frequency	Percent
Monthly	345	17.93
Quarterly	189	9.82
Yearly	1390	72.25

Product	Frequency	Percent
Market Link	81	4.21
Pure Term Plan	560	29.11
Traditional	1283	66.68

EducationField	Frequency	Percent
CA	583	30.30
Engineer	188	9.77
MBA	30	1.56
Marketing Diploma	219	11.38
Other	110	5.72
Statistics	794	41.27

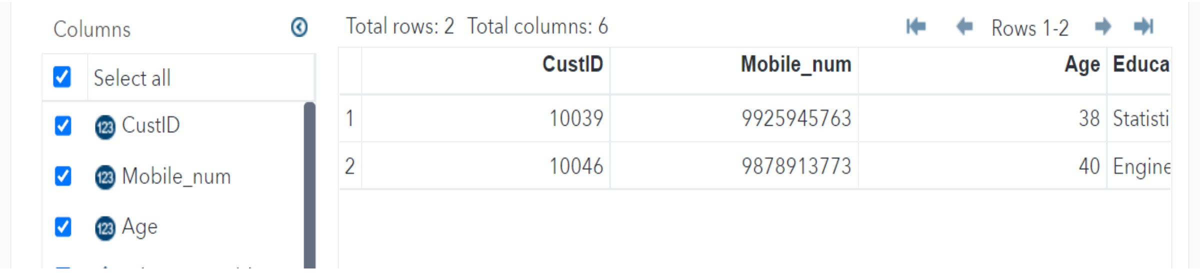
Gender	Frequency	Percent
Female	732	38.05
Male	1192	61.95

Overall_cust_satisfation_score	Frequency	Percent
1	71	3.69
2	464	24.12
3	455	23.65

7.Customer service management want you to create a macro where they will just put mobile number and they will get all the important information like Age, Education, Gender, Income and CustID.

Great Learning SAS Assessment 2
Ishita Sarkar

```
%MACRO customer_information();
DATA macro_insurance (keep = Mobile_num CustID Age EducationField Gender
Cust_Income);

SET SAS.DataInsurance;
where Mobile_num in (&Mobile_num.);
RUN;
proc print data=output;
run;
%MEND;
/* input mobile number */
%let Mobile_num = 9878913773,9925945763;
/* run macro for output */
%customer_information;
```

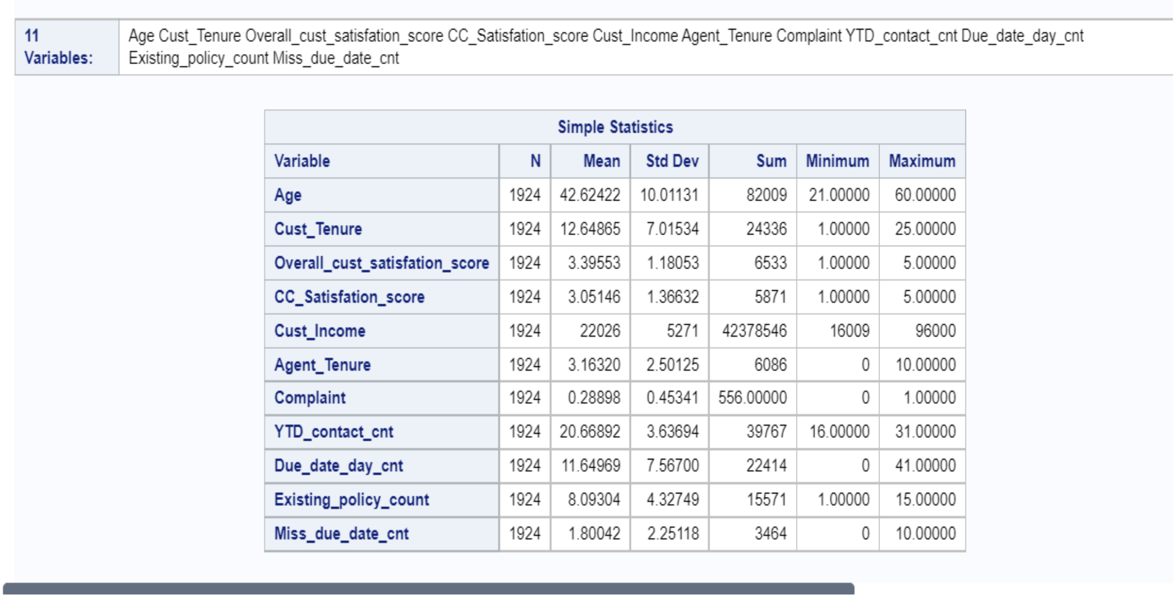


Total rows: 2 Total columns: 6

	CustID	Mobile_num	Age	Educa
1	10039	9925945763	38	Statisti
2	10046	9878913773	40	Engine

8.Check correlation of all numerical variables before building model, because we cannot add correlated variables in model?

```
proc corr data= SAS.DataInsurance noprob;
var Age Cust_Tenure Overall_cust_satisfation_score CC_Satisfation_score Cust_Income
Agent_Tenure Complaint YTD_contact_cnt Due_date_day_cnt Existing_policy_count
Miss_due_date_cnt;
run;
```



11 Variables:	Age Cust_Tenure Overall_cust_satisfation_score CC_Satisfation_score Cust_Income Agent_Tenure Complaint YTD_contact_cnt Due_date_day_cnt Existing_policy_count Miss_due_date_cnt

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Age	1924	42.62422	10.01131	82009	21.00000	60.00000
Cust_Tenure	1924	12.64865	7.01534	24336	1.00000	25.00000
Overall_cust_satisfation_score	1924	3.39553	1.18053	6533	1.00000	5.00000
CC_Satisfation_score	1924	3.05146	1.36632	5871	1.00000	5.00000
Cust_Income	1924	22026	5271	42378546	16009	96000
Agent_Tenure	1924	3.16320	2.50125	6086	0	10.00000
Complaint	1924	0.28898	0.45341	556.00000	0	1.00000
YTD_contact_cnt	1924	20.66892	3.63694	39767	16.00000	31.00000
Due_date_day_cnt	1924	11.64969	7.56700	22414	0	41.00000
Existing_policy_count	1924	8.09304	4.32749	15571	1.00000	15.00000
Miss_due_date_cnt	1924	1.80042	2.25118	3464	0	10.00000

9.Train and test (70:30) dataset from the existing data set. Put seed 1234?

```
proc surveyselect data=SAS.DataInsurance method=srs reps=1 sampsize=500 seed=1234
out=test;
run;

proc contents data=test varnum;  /* data=test */
run;

proc freq data=test;
table Churn /nocum;
run;

proc sql;
create table train as select tes.* from insurance as tes
where CustID not in (select CustID from test);
quit;

proc freq data=train;
table Churn /nocum;
run;
```

10.Develop linear regression model first on the target variable to extract VIF information to check multicollinearity?

Great Learning SAS Assessment 2
Ishita Sarkar

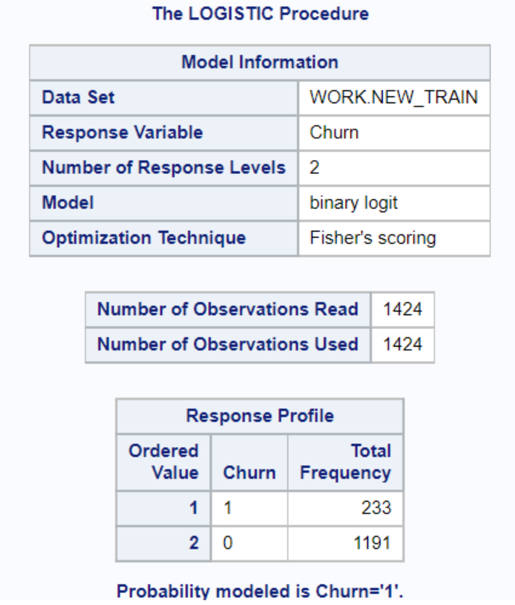
```
proc contents data= train;
run;

proc freq data=train;
tables Churn * Overall_cust_satisfation_score;
run;

proc freq data=train;
tables (Age Cust_Tenure Overall_cust_satisfation_score CC_Satisfation_score
Cust_Income
Agent_Tenure Complaint YTD_contact_cnt Due_date_day_cnt Existing_policy_count
Miss_due_date_cnt)
* Churn / chisq;
run;

data new_train (keep = CustID Churn Age Cust_Tenure Overall_cust_satisfation_score
CC_Satisfation_score Cust_Income Agent_Tenure Complaint YTD_contact_cnt
Due_date_day_cnt

Existing_policy_count Miss_due_date_cnt);
set train;
run;

proc freq data=new_train;
tables ( Age Cust_Tenure Overall_cust_satisfation_score CC_Satisfation_score
Cust_Income
Agent_Tenure Complaint YTD_contact_cnt Due_date_day_cnt Existing_policy_count
Miss_due_date_cnt)
 * Churn / chisq;
run;

proc logistic data = new_train;
class Churn Overall_cust_satisfation_score / param=ref;
model Churn = Overall_cust_satisfation_score / lackfit rsq;
title 'Churn vs Overall_cust_satisfation_score';
run;

proc logistic data = new_train;
class Churn Age Cust_Tenure Overall_cust_satisfation_score CC_Satisfation_score
Cust_Income Agent_Tenure Complaint YTD_contact_cnt Due_date_day_cnt
Existing_policy_count Miss_due_date_cnt / param=ref;
model Churn = Age Cust_Tenure Overall_cust_satisfation_score CC_Satisfation_score
Cust_Income
```

```
Agent_Tenure Complaint YTD_contact_cnt Due_date_day_cnt Existing_policy_count
Miss_due_date_cnt
/ lackfit rsq;
title 'Churn vs Overall_cust_satisfation_score - Multivariable Logistic Regression';
run;

proc corr data=new_train;
var Churn Age Cust_Tenure Overall_cust_satisfation_score
CC_Satisfation_score Cust_Income Agent_Tenure Complaint YTD_contact_cnt
Due_date_day_cnt
Existing_policy_count Miss_due_date_cnt;
run;
```

11.Create clean logistic model on the target variables?

```
%let var = Age Cust_Tenure Overall_cust_satisfation_score CC_Satisfation_score
Cust_Income

Agent_Tenure Complaint YTD_contact_cnt Due_date_day_cnt Existing_policy_count
Miss_due_date_cnt;


proc logistic data=new_train descending outmodel=model;
model Churn = &var / lackfit;
output out = train_output xbeta = coeff stdxbeta = stdcoeff predicted = prob;
run;
```

Great Learning SAS Assessment 2
Ishita Sarkar



The LOGISTIC Procedure

Model Information	
Data Set	WORK.NEW_TRAIN
Response Variable	Churn
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	1424
Number of Observations Used	1424

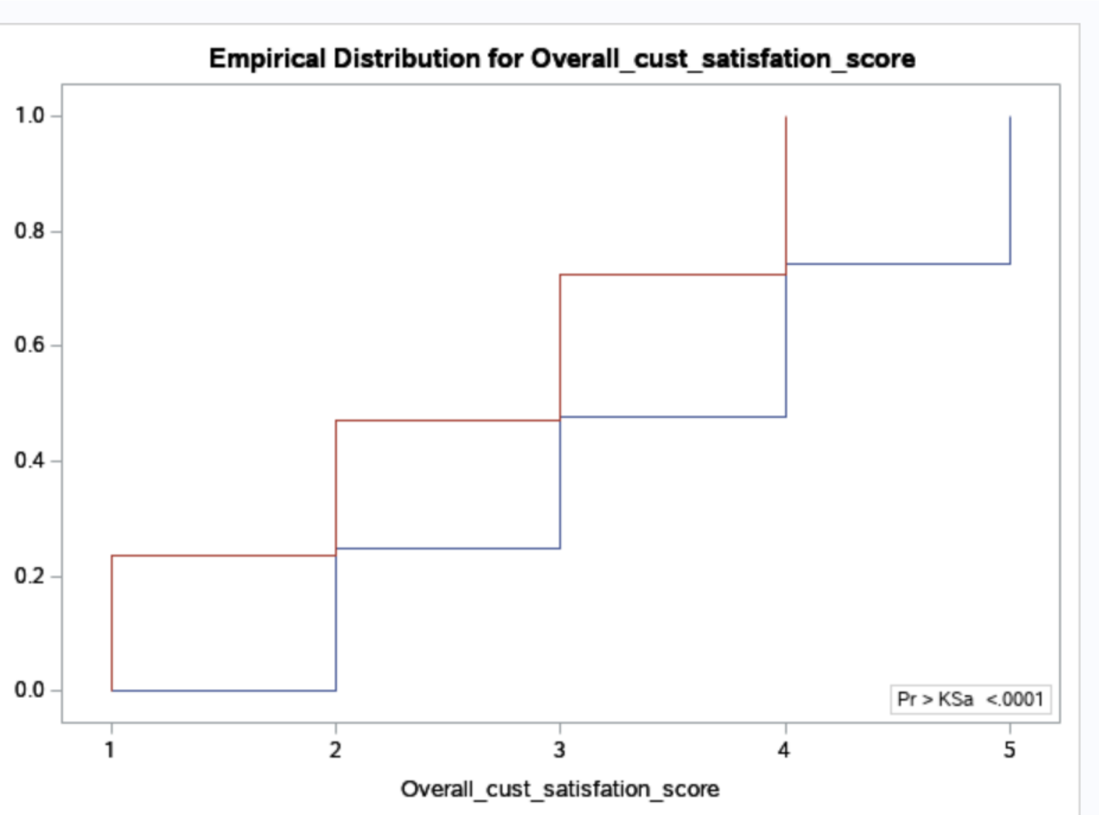
Response Profile

Ordered Value	Churn	Total Frequency
1	1	233
2	0	1191

Probability modeled is Churn='1'.

12.Create a macro and take a KS approach to take a cut off on the calculated scores?

```
ods graphics on;
proc npar1way edf plots= edfplot data= new_train;
class Churn;
var Overall_cust_satisfation_score;
exact ks;
run;
ods graphics off;
```



13.Predict test dataset using created model?

```
data test;
set test;
prob = -0.0226-0.0398*Age+0.4174*Overall_Satisfaction_Score
-
0.00009*Premium+0.0930*Network_hospital_nearby+0.0289*not_passed_percent_claim;
score = exp(problem)/(1+exp(problem));
run;
```