AIM: SAS for EDA Analysis

CODE:

```
%web_drop_table(WORK.IMPORT);
FILENAME REFFILE '/home/u62333421/sasuser.v94/credit_train.csv';
PROC IMPORT DATAFILE=REFFILE
      DBMS=CSV
      OUT=WORK.IMPORT;
      GETNAMES=YES;
RUN;
PROC CONTENTS DATA=WORK.IMPORT; RUN;
%web_open_table(WORK.IMPORT);
```

The CONTENTS Procedure

Data Set Name	WORK.IMPORT	Observations	100514
Member Type	DATA	Variables	19
Engine	V9	Indexes	0
Created	11/24/2022 10:40:59	Observation Length	224
Last Modified	11/24/2022 10:40:59	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	173
First Data Page	1
Max Obs per Page	584
Obs in First Data Page	564
Number of Data Set Repairs	0
Filename	/saswork/SAS_workCBFC0001C37C_odaws01-apse1.oda.sas.com/SAS_work46220001C37C_odaws01-apse1.oda.sas.com/import.sas7bdat
Release Created	9.0401M6
Host Created	Linux
Inode Number	536883539
Access Permission	rw-r--r--
Owner Name	u62333421
File Size	22MB
File Size (bytes)	22806528

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
7	Annual Income	Num	8	BEST12.	BEST32.
18	Bankruptcies	Num	8	BEST12.	BEST32.
6	Credit Score	Num	8	BEST12.	BEST32.
16	Current Credit Balance	Num	8	BEST12.	BEST32.
4	Current Loan Amount	Num	8	BEST12.	BEST32.
2	Customer ID	Char	36	$36.	$36.
9	Home Ownership	Char	13	$13.	$13.
1	Loan ID	Char	36	$36.	$36.
3	Loan Status	Char	11	$11.	$11.
17	Maximum Open Credit	Num	8	BEST12.	BEST32.
11	Monthly Debt	Num	8	BEST12.	BEST32.
13	Months since last delinquent	Char	2	$2.	$2.
15	Number of Credit Problems	Num	8	BEST12.	BEST32.
14	Number of Open Accounts	Num	8	BEST12.	BEST32.
10	Purpose	Char	18	$18.	$18.
19	Tax Liens	Num	8	BEST12.	BEST32.
5	Term	Char	10	$10.	$10.
8	Years in current job	Char	9	$9.	$9.
12	Years of Credit History	Num	8	BEST12.	BEST32.

```
PROC means DATA=WORK.IMPORT mean median mode std var min max;
run;
```

The MEANS Procedure

Variable	Mean	Median	Mode	Std Dev	Variance	Minimum	Maximum
Current Loan Amount	11760447.39	312246.00	99999999.00	31783942.55	1.010219E15	10802.00	99999999.00
Credit Score	1076.46	724.0000000	747.0000000	1475.40	2176816.35	585.0000000	7510.00
Annual Income	1378276.56	1174162.00	1162572.00	1081360.20	1.1693399E12	76627.00	165557393
Monthly Debt	18472.41	16220.30	0	12174.99	148230445	0	435843.28
Years of Credit History	18.1991410	16.9000000	16.0000000	7.0153236	49.2147659	3.6000000	70.5000000
Number of Open Accounts	11.1285300	10.0000000	9.0000000	5.0098704	25.0988010	0	76.0000000
Number of Credit Problems	0.1683100	0	0	0.4827050	0.2330041	0	15.0000000
Current Credit Balance	294637.38	209817.00	0	376170.93	141504572088	0	32878968.00
Maximum Open Credit	760798.38	467874.00	0	8384503.47	7.0299898E13	0	1539737892
Bankruptcies	0.1177402	0	0	0.3514238	0.1234987	0	7.0000000
Tax Liens	0.0293129	0	0	0.2581824	0.0666582	0	15.0000000

```
/* Missing values in our dataset */
PROC means DATA=WORK.IMPORT nmiss;
run;
```

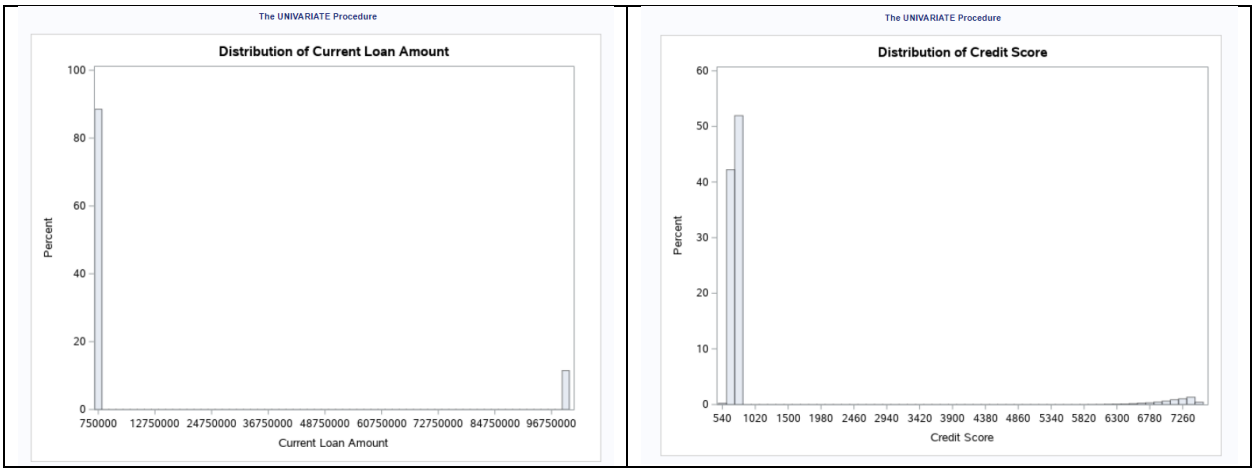
The MEANS Procedure

Variable	N Miss
Current Loan Amount	514
Credit Score	19668
Annual Income	19668
Monthly Debt	514
Years of Credit History	514
Number of Open Accounts	514
Number of Credit Problems	514
Current Credit Balance	514
Maximum Open Credit	516
Bankruptcies	718
Tax Liens	524

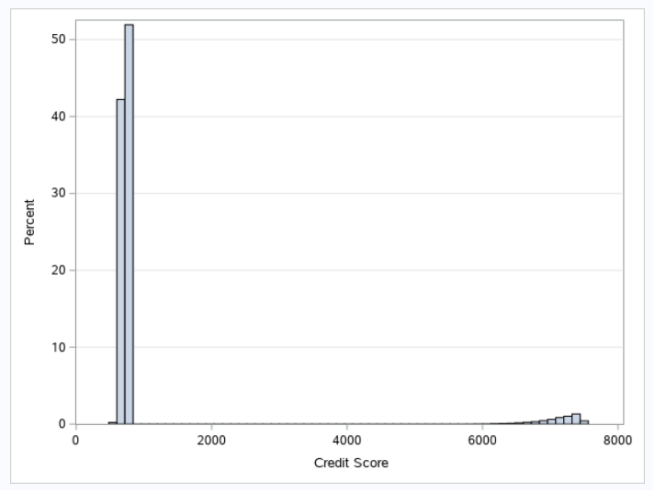
```
PROC SQL;
select count(distinct 'Loan Status'n) as 'Loan Status'n,
       count(distinct Bankruptcies) as Bankruptcies,
       count(distinct Term) as Term,
       count(distinct 'Credit Score'n) as 'Credit Score'n,
       count(distinct 'Monthly Debt'n) as 'Monthly Debt'n
  from WORK.IMPORT;
QUIT;
```

Loan Status	Bankruptcies	Term	Credit Score	Monthly Debt
2	8	2	324	65765

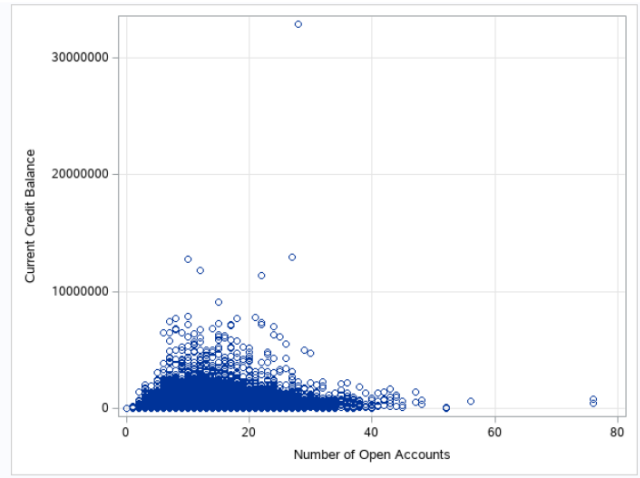
```
/* You can give multiple variables in this procedure to create
histograms */
PROC univariate DATA=WORK.IMPORT  novarcontents;
histogram 'Current Loan Amount'n 'Credit Score'n / ;
RUN;
```



```
/* Creating histogram with only one variable (i.e Credit Score) */
ods graphics / reset width=6.4in height=4.8in imagemap;
proc sgplot DATA=WORK.IMPORT;
      histogram 'Credit Score'n /;
      yaxis grid;
RUN;
```




```
/* Checking Relationship between two variables by using scatter plot
*/
ods graphics / reset width=6.4in height=4.8in imagemap;
PROC sgplot DATA=WORK.IMPORT;
       scatter x='Number of Open Accounts'n y='Current Credit
Balance'n /;
       xaxis grid;
       yaxis grid;
RUN;
ods graphics / reset;
```



```
/* Correaltion among numeric variables */
ods noproctitle;
ods graphics / imagemap=on;
PROC corr DATA=WORK.IMPORT pearson nosimple noprob plots=none;
       var 'Current Loan Amount'n 'Credit Score'n 'Annual Income'n
'Monthly Debt'n
             'Years of Credit History'n 'Number of Open Accounts'n
             'Number of Credit Problems'n 'Current Credit Balance'n
'Maximum Open Credit'n
             Bankruptcies 'Tax Liens'n;
RUN;
```

11 Variables: Current Loan Amount Credit Score Annual Income Monthly Debt Years of Credit History Number of Open Accounts Number of Credit Problems Current Credit Balance Maximum Open Credit Bankruptcies Tax Liens

Pearson Correlation Coefficients
Number of Observations

	Current Loan Amount	Credit Score	Annual Income	Monthly Debt	Years of Credit History	Number of Open Accounts	Number of Credit Problems	Current Credit Balance	Maximum Open Credit	Bankruptcies	Tax Liens
Current Loan Amount	1.00000 100000	-0.09665 80846	0.01311 80846	-0.00664 100000	0.01928 100000	0.00148 100000	-0.00279 100000	0.00388 100000	-0.00127 99998	-0.00061 99796	-0.00205 99990
Credit Score	-0.09665 80846	1.00000 80846	-0.01708 80846	-0.00167 80846	-0.00972 80846	0.00644 80846	-0.00302 80846	-0.00010 80846	-0.00283 80845	-0.00693 80684	0.00515 80840
Annual Income	0.01311 80846	-0.01708 80846	1.00000 80846	0.48523 80846	0.16167 80846	0.14617 80846	-0.01701 80846	0.31234 80846	0.05306 80845	-0.04767 80684	0.04017 80840
Monthly Debt	-0.00664 100000	-0.00167 80846	0.48523 80846	1.00000 100000	0.19929 100000	0.41135 100000	-0.05538 100000	0.48135 100000	0.03927 99998	-0.07898 99796	0.02012 99990
Years of Credit History	0.01928 100000	-0.00972 80846	0.16167 80846	0.19929 100000	1.00000 100000	0.13235 100000	0.06159 100000	0.20847 100000	0.03112 99998	0.06625 99796	0.01724 99990
Number of Open Accounts	0.00148 100000	0.00644 80846	0.14617 80846	0.41135 100000	0.13235 100000	1.00000 100000	-0.01399 100000	0.22814 100000	0.03134 99998	-0.02458 99796	0.00654 99990
Number of Credit Problems	-0.00279 100000	-0.00302 80846	-0.01701 80846	-0.05538 100000	0.06159 100000	-0.01399 100000	1.00000 100000	-0.11252 100000	-0.01207 99998	0.75294 99796	0.58129 99990
Current Credit Balance	0.00388 100000	-0.00010 80846	0.31234 80846	0.48135 100000	0.20847 100000	0.22814 100000	-0.11252 100000	1.00000 100000	0.13920 99998	-0.12260 99796	-0.01565 99990
Maximum Open Credit	-0.00127 99998	-0.00283 80845	0.05306 80845	0.03927 99998	0.03112 99998	0.03134 99998	-0.01207 99998	0.13920 99998	1.00000 99998	-0.01457 99794	-0.00103 99988
Bankruptcies	-0.00061 99796	-0.00693 80684	-0.04767 80684	-0.07898 99796	0.06625 99796	-0.02458 99796	0.75294 99796	-0.12260 99796	-0.01457 99794	1.00000 99796	0.04611 99796
Tax Liens	-0.00205 99990	0.00515 80840	0.04017 80840	0.02012 99990	0.01724 99990	0.00654 99990	0.58129 99990	-0.01565 99990	-0.00103 99988	0.04611 99796	1.00000 99990

```
/* Box plot for checking outliers in the data */
ods graphics / reset width=6.4in height=4.8in imagemap;
proc sgplot  DATA=WORK.IMPORT;
      vbox 'Credit Score'n  / category='Loan Status'n;
      yaxis grid;
run;
ods graphics / reset;
```