PROJECT – Data Science Project Using SAS

Credit card Details Binary Classification

- **By: Xue Liu(Alexia)**

- **Date: 2024-01-06**

Table of Contents

```
libname project "/home/u63693354/myproj";
*libname project "D:\DataScience\07SAS\SAS_Library";
```

1, Intorduction

Data Source : Data is taken from Kaggle: https://www.kaggle.com/datasets/rohitudageri/credit-card-details/data.
The dataset contains 19 variables and 1,548 observations. There exist missing values, no duplicate data are found.
Describe : Analyzing the customer's information from the dataset, including credit history, income, employment status, and other relevant features, to predict whether an applicant is likely to be a responsible cardholder.
Purpose : To predict the approvement of the customer's credit card application..

2, Objectives of this analysis

/* Business Problem */

```
/* What factors influence credit card approval? */
/* Clients with higher annual incomes are more likely to be approved.
Clients who own property are less likely to be high credit risks.
Education level is positively correlated with creditworthiness.
Gender may influence the choice of housing type and education level.
Clients with a longer employment history are more likely to be approved for credit cards. */
```

/* Import Data */

```
PROC IMPORT OUT= project.CreditCard
        DATAFILE= '/home/u63693354/myproj/Credit_card.csv'
        DBMS=CSV REPLACE;
    GETNAMES=YES;
    GUESSINGROWS=1000;
    *MIXED=NO;
    *SCANTEXT=YES;
    *USEDATE=YES;
    *SCANTIME=YES;
RUN;
*DATAFILE= '/home/u63517891/Library/Credit_card.csv';
Proc Print data=project.CreditCard;
Run;

/* variable identification */
Proc Contents data= project.CreditCard;
Run;
/* ================================================================================ */
```

3, Methodology

/* Missing Values */

```
proc freq data=project.CreditCard;
```

```
tables _CHARACTER_ / missing;
run;

proc means data=project.CreditCard nmiss;
var _numeric_;
run;
*have missing value:
GENDER Type_Occupation
Annual_income Birthday_count;

/* Replace Numeric Missing Values with Median*/
* Calculate Median;
proc summary data=project.CreditCard nway;
   var Annual_income Birthday_count;
   output out=MedianValues median=Annual_income_median Birthday_count_median;
run;

data project.CreditCard2;
   set project.CreditCard;
   if _N_ = 1 then set MedianValues;
   if missing(Annual_income) then Annual_income = Annual_income_median;
   if missing(Birthday_count) then Birthday_count = Birthday_count_median;
run;
proc means data=project.CreditCard2 nmiss;
var _numeric_;
run;

/* Replace Missing Values for Gender with Mode */
proc freq data=project.CreditCard2 noprint;
   table GENDER / out=Mode(drop=percent count) noprint;
   where GENDER is not missing;
run;

data project.CreditCard2;
   set project.CreditCard2;
   if missing(GENDER) then GENDER = 'F';
run;

proc freq data=project.CreditCard2;
tables _CHARACTER_ / missing;
run;

/* Replace Missing Values for Gender with "Unknown" */
data project.CreditCard2;
   set project.CreditCard2;
   if missing(Type_Occupation) then Type_Occupation = "Unknown";
run;
```

```
/* Replace the value for less than 100 in each category with "Others" */
proc sql;
  create table OccupationCounts as
  select Type_Occupation, count(*) as count
  from project.CreditCard2
  group by Type_Occupation;
quit;

data project.CreditCard2;
   if _N_ = 1 then do;
      declare hash h(dataset: "OccupationCounts");
      h.defineKey("Type_Occupation");
      h.defineData("count");
      h.defineDone();
   end;

   set project.CreditCard2;

   rc = h.find();
   if rc = 0 then do;
      if count < 100 then Type_Occupation = "Others";
   end;
run;
/* ========================================================================== */
```

4, Study Variables

Ind_ID: Client ID
Gender: Gender information
Car_owner: Having car or not
Propert_owner: Having property or not
Children: Count of children
Annual_income: Annual income
Type_Income: Income type
Education: Education level
Marital_status: Marital_status
Housing_type: Living style
Birthday_count: Use backward count from current day (0), -1 means yesterday.
Employed_days: Start date of employment. Use backward count from current day (0). Positive value means, individual is currently unemployed.
Mobile_phone: Any mobile phone
Work_phone: Any work phone
Phone: Any phone number
EMAIL_ID: Any email ID
Type_Occupation: Occupation
Family_Members: Family size
Label: 0 is application approved and 1 is application rejected.

/* ========== */

5, Descriptive Analysis

/* Univariate Analysis*/

```
data project.CreditCard3;
set project.CreditCard2;
Age = int(abs(Birthday_count) / 365.25);
if Employed_days = 365243 then do;
  Employed_Day = 0;
  Employed_Year = 0;
end;
else do;
  Employed_Day = abs(Employed_days);
  Employed_Year = round(Employed_Day / 365.25, 0.1);
end;
drop Ind_ID _TYPE_ _FREQ_ Annual_income_median Birthday_count_median count rc Mobile_phone;
run;
Proc Print data=project.CreditCard3;
Run;
```

* Categorical variables;

```
proc freq data=project.CreditCard3;
table GENDER Car_Owner Propert_Owner CHILDREN Type_Income EDUCATION Marital_status
Housing_type Work_Phone Phone EMAIL_ID Type_Occupation Family_Members label/missing;
run;
```

The FREQ Procedure

GENDER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
F	980	63.31	980	63.31
M	568	36.69	1548	100.00

Car_Owner	Frequency	Percent	Cumulative Frequency	Cumulative Percent
N	924	59.69	924	59.69
Y	624	40.31	1548	100.00

Propert_Owner	Frequency	Percent	Cumulative Frequency	Cumulative Percent
N	538	34.75	538	34.75
Y	1010	65.25	1548	100.00

CHILDREN	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1091	70.48	1091	70.48
1	305	19.70	1396	90.18
2	134	8.66	1530	98.84
3	16	1.03	1546	99.87
4	1	0.06	1547	99.94
14	1	0.06	1548	100.00

Type_Income	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Commercial associate	365	23.58	365	23.58
Pensioner	269	17.38	634	40.96
State servant	116	7.49	750	48.45
Working	798	51.55	1548	100.00

EDUCATION	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Academic degree	2	0.13	2	0.13
Higher education	426	27.52	428	27.65
Incomplete higher	68	4.39	496	32.04
Lower secondary	21	1.36	517	33.40
Secondary / secondary special	1031	66.60	1548	100.00

Marital_status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Civil marriage	101	6.52	101	6.52
Married	1049	67.76	1150	74.29
Separated	96	6.20	1246	80.49
Single / not married	227	14.66	1473	95.16
Widow	75	4.84	1548	100.00

Housing_type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Co-op apartment	5	0.32	5	0.32
House / apartment	1380	89.15	1385	89.47
Municipal apartment	53	3.42	1438	92.89
Office apartment	9	0.58	1447	93.48
Rented apartment	21	1.36	1468	94.83
With parents	80	5.17	1548	100.00

Work_Phone	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1226	79.20	1226	79.20
1	322	20.80	1548	100.00

Phone	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1069	69.06	1069	69.06
1	479	30.94	1548	100.00

EMAIL_ID	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1405	90.76	1405	90.76
1	143	9.24	1548	100.00

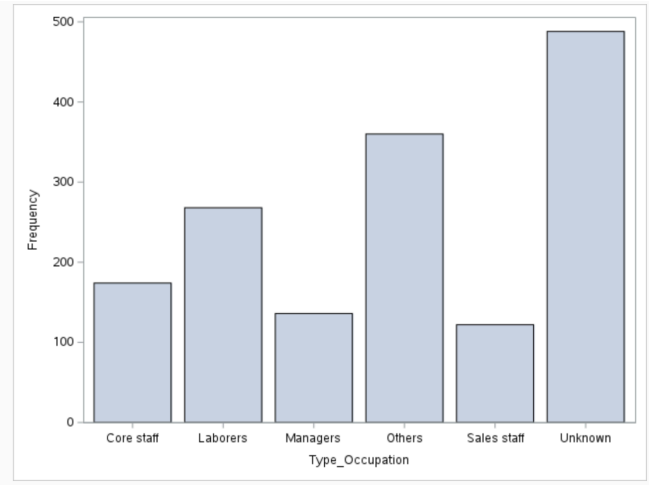
Type_Occupation	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Core staff	174	11.24	174	11.24
Laborers	268	17.31	442	28.55
Managers	136	8.79	578	37.34
Others	360	23.26	938	60.59
Sales staff	122	7.88	1060	68.48
Unknown	488	31.52	1548	100.00

Family_Members	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	334	21.58	334	21.58
2	802	51.81	1136	73.39
3	268	17.31	1404	90.70
4	127	8.20	1531	98.90
5	15	0.97	1546	99.87
6	1	0.06	1547	99.94
15	1	0.06	1548	100.00

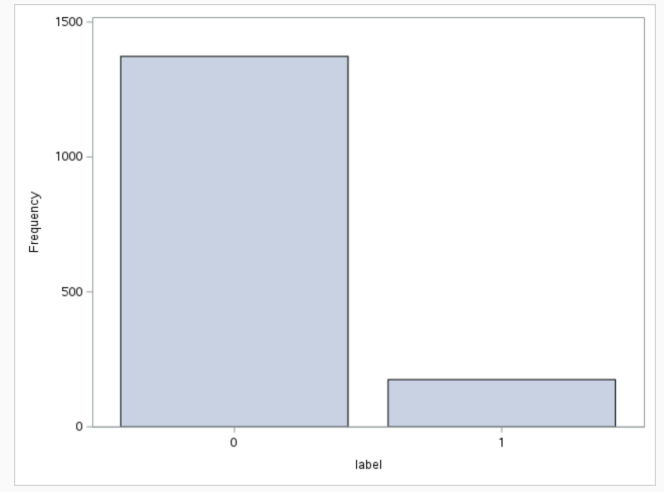
label	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1373	88.70	1373	88.70
1	175	11.30	1548	100.00

```
proc sgplot data=project.CreditCard3;
vbar Car_Owner / missing;
run;
proc sgplot data=project.CreditCard3;
vbar Propert_Owner / missing;
run;
proc sgplot data=project.CreditCard3;
vbar CHILDREN / missing;
run;
proc sgplot data=project.CreditCard3;
vbar Type_Income / missing;
run;
proc sgplot data=project.CreditCard3;
vbar EDUCATION / missing;
run;
proc sgplot data=project.CreditCard3;
vbar Marital_status / missing;
run;
proc sgplot data=project.CreditCard3;
vbar Housing_type / missing;
run;
```

```
proc sgplot data=project.CreditCard3;
vbar Work_Phone / missing;
run;
proc sgplot data=project.CreditCard3;
vbar Phone / missing;
run;
proc sgplot data=project.CreditCard3;
vbar EMAIL_ID / missing;
run;
proc sgplot data=project.CreditCard3;
vbar Type_Occupation / missing;
run;
```



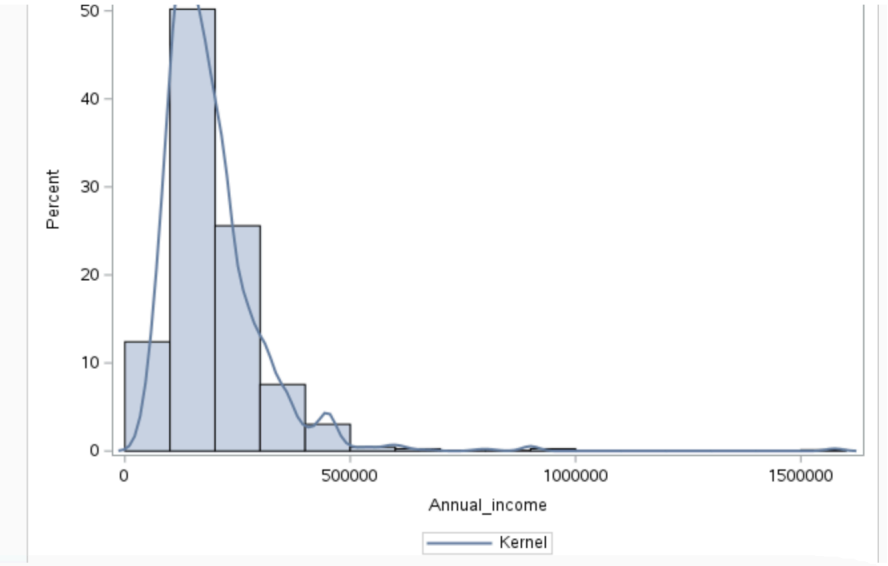

```
proc sgplot data=project.CreditCard3;
vbar Family_Members / missing;
run;
proc sgplot data=project.CreditCard3;;
vbar label / missing;
run;
```



* Numerical variables;

```
title "Annual Income Distribution";
proc sgplot data=project.CreditCard3;
histogram Annual_income;
```

```
density Annual_income /type=kernel;
run;
```




```
title "Birthday_count Distribution";
proc sgplot data=project.CreditCard3;
histogram Birthday_count;
density Birthday_count /type=kernel;
run;

title "Age Distribution";
proc sgplot data=project.CreditCard3;
histogram Age;
density Age /type=kernel;
run;

title "Employed Days Distribution";
proc sgplot data=project.CreditCard3;
histogram Employed_Day;
density Employed_Day /type=kernel;
run;

title "Employed Years Distribution";
proc sgplot data=project.CreditCard3;
histogram Employed_Year;
density Employed_Year /type=kernel;
run;

proc univariate data=project.CreditCard3;
var Annual_income;
histogram Annual_income/ normal (mu=est sigma=est color=black)
kernel (color=blue);
qqplot Annual_income/ normal (mu=est sigma=est color=black);
run;
```

The UNIVARIATE Procedure
Variable: Annual_income

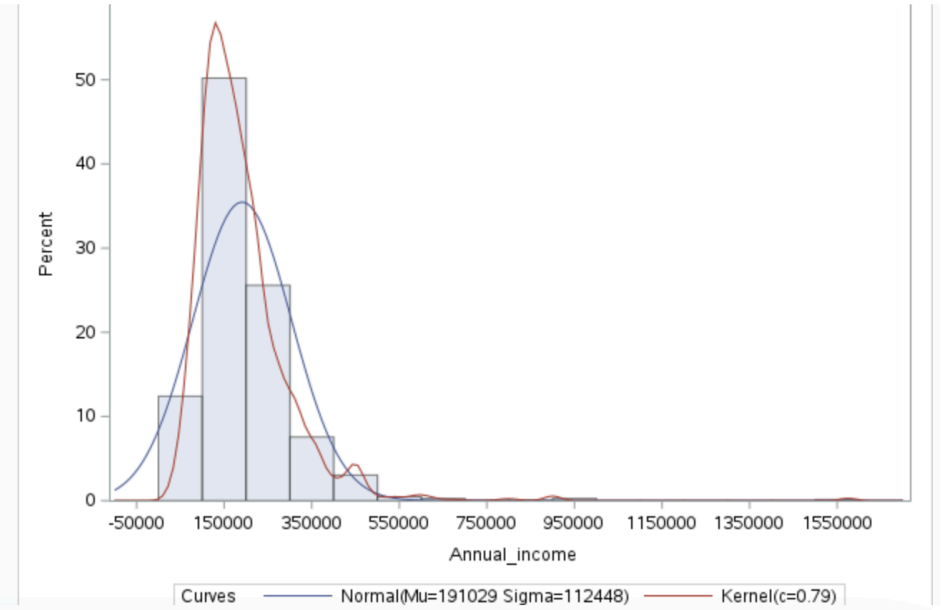
Moments			
N	1548	Sum Weights	1548
Mean	191029.375	Sum Observations	295713473
Std Deviation	112448.335	Variance	1.26446E10
Skewness	3.95945329	Kurtosis	34.4466083
Uncorrected SS	7.60512E13	Corrected SS	1.95612E13
Coeff Variation	58.8644208	Std Error Mean	2858.03506

Basic Statistical Measures			
Location		Variability	
Mean	191029.4	Std Deviation	112448
Median	166500.0	Variance	1.26446E10
Mode	135000.0	Range	1541250
		Interquartile Range	103500

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	66.83941	Pr > |t|	<.0001
Sign	M	774	Pr >= |M|	<.0001
Signed Rank	S	599463	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	1575000
99%	585000
95%	360000
90%	315000
75% Q3	225000
50% Median	166500
25% Q1	121500
10%	90000
5%	76500
1%	54000
0% Min	33750

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
33750	1039	900000	687
36000	395	900000	812
37800	647	900000	1069
40500	1090	1575000	212
40500	856	1575000	234

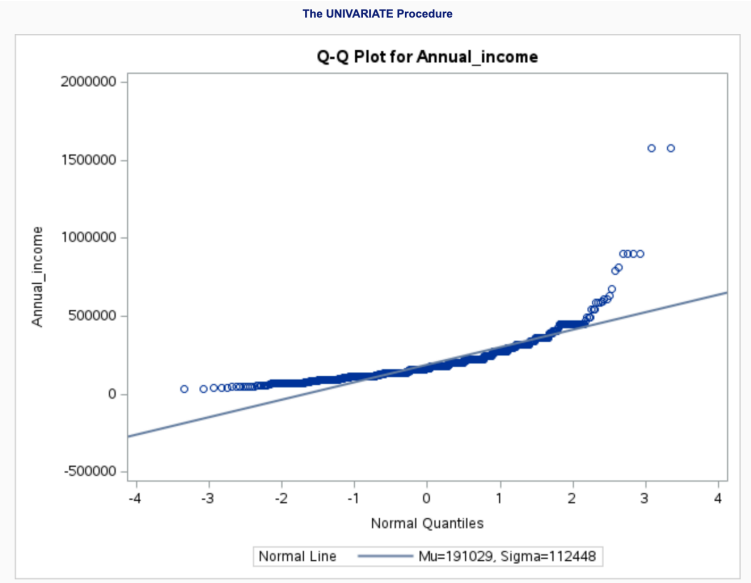


The UNIVARIATE Procedure
Fitted Normal Distribution for Annual_income

Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	191029.4
Std Dev	Sigma	112448.3

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.1610037	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	10.7909356	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	62.9826178	**Pr > A-Sq**	<0.005

Quantiles for Normal Distribution		
	Quantile	
Percent	**Observed**	**Estimated**
1.0	54000.0	-70564.57
5.0	76500.0	6068.32
10.0	90000.0	46921.04
25.0	121500.0	115184.13
50.0	166500.0	191029.38
75.0	225000.0	266874.62
90.0	315000.0	335137.71
95.0	360000.0	375990.43
99.0	585000.0	452623.32

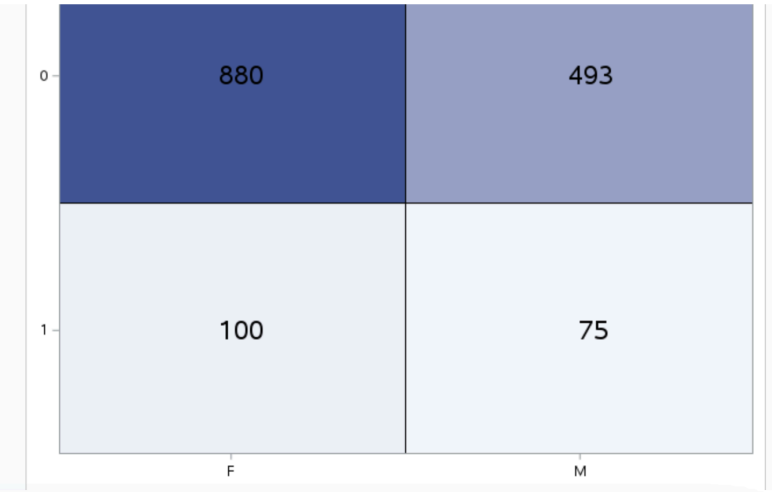



```
proc univariate data=project.CreditCard3;
var Age;
histogram Age/ normal (mu=est sigma=est color=black)
kernel (color=blue);
qqplot Age/ normal (mu=est sigma=est color=black);
run;

proc univariate data=project.CreditCard3;
var Employed_Day;
histogram Employed_Day/ normal (mu=est sigma=est color=black)
kernel (color=blue);
qqplot Employed_Day/ normal (mu=est sigma=est color=black);
run;

proc univariate data=project.CreditCard3;
var Employed_Year;
histogram Employed_Year/ normal (mu=est sigma=est color=black)
kernel (color=blue);
qqplot Employed_Year/ normal (mu=est sigma=est color=black);
run;
/* ============================================================================== */

/* Bivariate Analysis */
* GENDER VS label ;
proc freq data=project.CreditCard3;
 table GENDER * label / norow nocol nopct
   out=project.FreqOut(where=(percent^=.));
run;
```

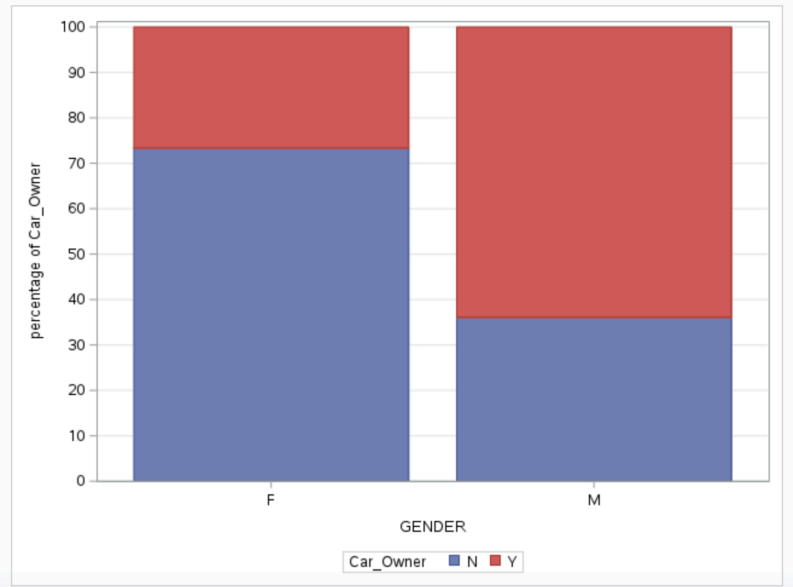
The FREQ Procedure

Frequency	Table of GENDER by label		
	label		
GENDER	0	1	Total
F	880	100	980
M	493	75	568
Total	1373	175	1548

```
* heatmap;
proc sgplot data=project.FreqOut noautolegend;
heatmap x=GENDER y=label / freq=Count
        discretex discretey
        colormodel=TwoColorRamp outline;
text x=GENDER y=label text=Count / textattrs=(size=16pt);
yaxis display=(nolabel) reverse;
xaxis display=(nolabel);
run;
```




```
proc freq data=credit;
tables GENDER*Car_Owner / norow nocol nopercent;
run;
proc freq data=credit;
tables GENDER*Car_Owner / chisq fisher;
run;
proc sort data=credit
          out=credit2;
by GENDER;
proc freq data=credit2 noprint;
by GENDER;
tables Car_Owner / out=freqout;
run;
proc sgplot data=freqout;
vbar GENDER / response=percent
          group=Car_Owner groupdisplay=stack;
xaxis discreteorder=data;
          yaxis grid values=(0 to 100 by 10)
          label="percentage of Car_Owner";
run;
```



Statistics for Table of GENDER by Car_Owner

Statistic	DF	Value	Prob
Chi-Square	1	207.6551	<.0001
Likelihood Ratio Chi-Square	1	208.6365	<.0001
Continuity Adj. Chi-Square	1	206.1088	<.0001
Mantel-Haenszel Chi-Square	1	207.5210	<.0001
Phi Coefficient		0.3663	
Contingency Coefficient		0.3439	
Cramer's V		0.3663	

```
* Marital_status VS label ;
proc freq data=project.CreditCard3;
 table Marital_status * label / norow nocol nopct
   out=project.FreqOut(where=(percent^=.));
run;
* heatmap;
proc sgplot data=project.FreqOut noautolegend;
heatmap x=Marital_status y=label / freq=Count
        discretex discretey
        colormodel=TwoColorRamp outline;
text x=Marital_status y=label text=Count / textattrs=(size=16pt);
yaxis display=(nolabel) reverse;
xaxis display=(nolabel);
run;
```

```
* Housing_type VS label ;
proc freq data=project.CreditCard3;
 table Housing_type * label / norow nocol nopct
   out=project.FreqOut(where=(percent^=.));
```

```
run;
* heatmap;
proc sgplot data=project.FreqOut noautolegend;
heatmap x=Housing_type y=label / freq=Count
        discretex discretey
        colormodel=TwoColorRamp outline;
text x=Housing_type y=label text=Count / textattrs=(size=16pt);
yaxis display=(nolabel) reverse;
xaxis display=(nolabel);
run;
```

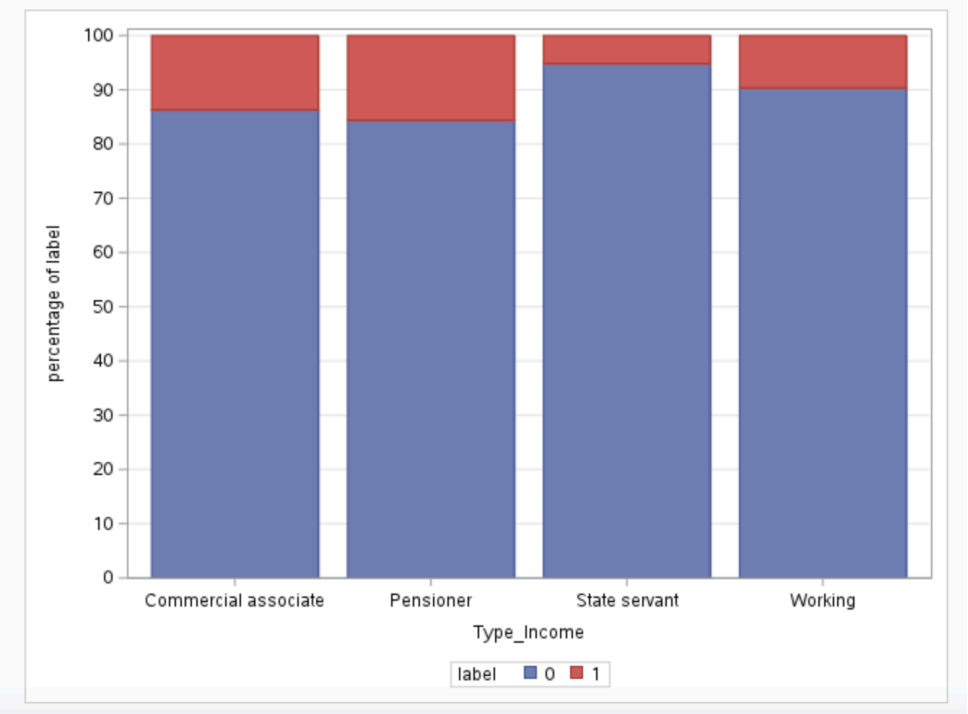
```
* Family_Members VS label ;
proc freq data=project.CreditCard3;
 table Family_Members * label / norow nocol nopct
   out=project.FreqOut(where=(percent^=.));
run;
* heatmap;
proc sgplot data=project.FreqOut noautolegend;
heatmap x=Family_Members y=label / freq=Count
        discretex discretey
        colormodel=TwoColorRamp outline;
text x=Family_Members y=label text=Count / textattrs=(size=16pt);
yaxis display=(nolabel) reverse;
xaxis display=(nolabel);
run;
```

```
* vbar and chisq;
proc freq data=project.CreditCard3;
tables CHILDREN*label / norow nocol nopercent;
run;
proc freq data=project.CreditCard3;
tables CHILDREN*label / chisq fisher;
run;
proc sort data=project.CreditCard3
          out=credit2;
by CHILDREN;
proc freq data=credit2 noprint;
by CHILDREN;
tables label / out=freqout;
run;
proc sgplot data=freqout;
vbar CHILDREN / response=percent
          group=label groupdisplay=stack;
xaxis discreteorder=data;
          yaxis grid values=(0 to 100 by 10)
          label="percentage of label";
run;
```

```
proc freq data=project.CreditCard3;
tables Family_Members*label / norow nocol nopercent;
run;
proc freq data=project.CreditCard3;
tables Family_Members*label / chisq fisher;
run;
proc sort data=project.CreditCard3
        out=credit2;
by Family_Members;
proc freq data=credit2 noprint;
by Family_Members;
tables label / out=freqout;
run;
proc sgplot data=freqout;
vbar Family_Members / response=percent
        group=label groupdisplay=stack;
xaxis discreteorder=data;
        yaxis grid values=(0 to 100 by 10)
        label="percentage of label";
run;

proc freq data=project.CreditCard3;
tables Work_Phone*label / norow nocol nopercent;
run;
proc freq data=project.CreditCard3;
tables Work_Phone*label / chisq fisher;
run;
proc sort data=project.CreditCard3
        out=credit2;
by Work_Phone;
proc freq data=credit2 noprint;
by Work_Phone;
tables label / out=freqout;
run;
proc sgplot data=freqout;
vbar Work_Phone / response=percent
        group=label groupdisplay=stack;
xaxis discreteorder=data;
        yaxis grid values=(0 to 100 by 10)
        label="percentage of label";
run;

proc freq data=project.CreditCard3;
tables EMAIL_ID*label / norow nocol nopercent;
run;
proc freq data=project.CreditCard3;
tables EMAIL_ID*label / chisq fisher;
run;
proc sort data=project.CreditCard3
```

```
        out=credit2;
by EMAIL_ID;
proc freq data=credit2 noprint;
by EMAIL_ID;
tables label / out=freqout;
run;
proc sgplot data=freqout;
vbar EMAIL_ID / response=percent
        group=label groupdisplay=stack;
xaxis discreteorder=data;
        yaxis grid values=(0 to 100 by 10)
        label="percentage of label";
run;

proc freq data=project.CreditCard3;
tables GENDER*label / norow nocol nopercent;
run;
proc freq data=project.CreditCard3;
tables GENDER*label / chisq fisher;
run;
proc sort data=project.CreditCard3
        out=credit2;
by GENDER;
proc freq data=credit2 noprint;
by GENDER;
tables label / out=freqout;
run;
proc sgplot data=freqout;
vbar GENDER / response=percent
        group=label groupdisplay=stack;
xaxis discreteorder=data;
        yaxis grid values=(0 to 100 by 10)
        label="percentage of label";
run;

proc freq data=project.CreditCard3;
tables Car_Owner*label / norow nocol nopercent;
run;
proc freq data=project.CreditCard3;
tables Car_Owner*label / chisq fisher;
run;
proc sort data=project.CreditCard3
        out=credit2;
by Car_Owner;
proc freq data=credit2 noprint;
by Car_Owner;
tables label / out=freqout;
run;
proc sgplot data=freqout;
```

```
vbar Car_Owner / response=percent
        group=label groupdisplay=stack;
xaxis discreteorder=data;
        yaxis grid values=(0 to 100 by 10)
        label="percentage of label";
run;

proc freq data=project.CreditCard3;
tables Propert_Owner*label / norow nocol nopercent;
run;
proc freq data=project.CreditCard3;
tables Propert_Owner*label / chisq fisher;
run;
proc sort data=project.CreditCard3
        out=credit2;
by Propert_Owner;
proc freq data=credit2 noprint;
by Propert_Owner;
tables label / out=freqout;
run;
proc sgplot data=freqout;
vbar Propert_Owner / response=percent
        group=label groupdisplay=stack;
xaxis discreteorder=data;
        yaxis grid values=(0 to 100 by 10)
        label="percentage of label";
run;

proc freq data=project.CreditCard3;
tables Type_Income*label / norow nocol nopercent;
run;
proc freq data=project.CreditCard3;
tables Type_Income*label / chisq fisher;
run;
proc sort data=project.CreditCard3
        out=credit2;
by Type_Income;
proc freq data=credit2 noprint;
by Type_Income;
tables label / out=freqout;
run;
proc sgplot data=freqout;
vbar Type_Income / response=percent
        group=label groupdisplay=stack;
xaxis discreteorder=data;
        yaxis grid values=(0 to 100 by 10)
        label="percentage of label";
run;
```

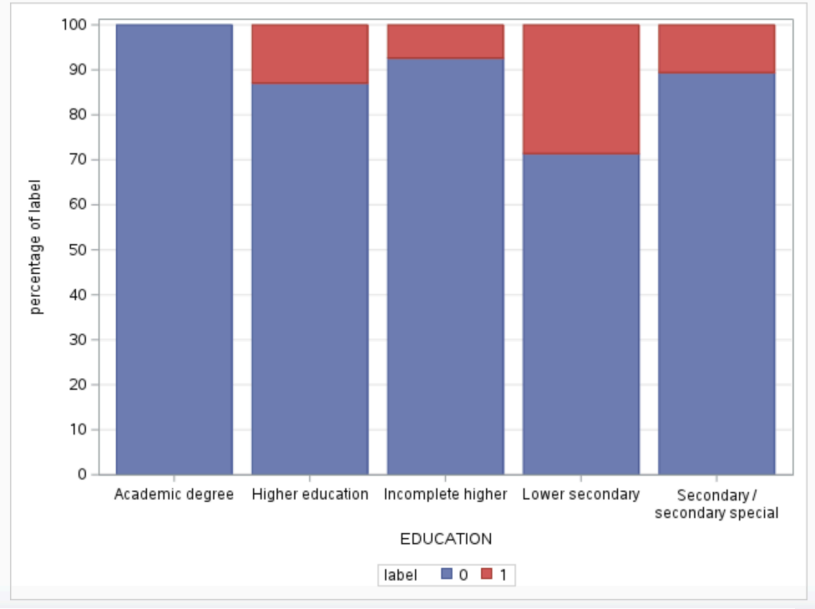


Statistics for Table of Type_Income by label

Statistic	DF	Value	Prob
Chi-Square	3	13.5986	0.0035
Likelihood Ratio Chi-Square	3	14.0933	0.0028
Mantel-Haenszel Chi-Square	1	7.1231	0.0076
Phi Coefficient		0.0937	
Contingency Coefficient		0.0933	
Cramer's V		0.0937	

```
proc freq data=project.CreditCard3;
tables EDUCATION*label / norow nocol nopercent;
run;
proc freq data=project.CreditCard3;
tables EDUCATION*label / chisq fisher;
run;
proc sort data=project.CreditCard3
        out=credit2;
by EDUCATION;
proc freq data=credit2 noprint;
by EDUCATION;
tables label / out=freqout;
run;
proc sgplot data=freqout;
vbar EDUCATION / response=percent
        group=label groupdisplay=stack;
xaxis discreteorder=data;
        yaxis grid values=(0 to 100 by 10)
```

```
        label="percentage of label";
run;
```

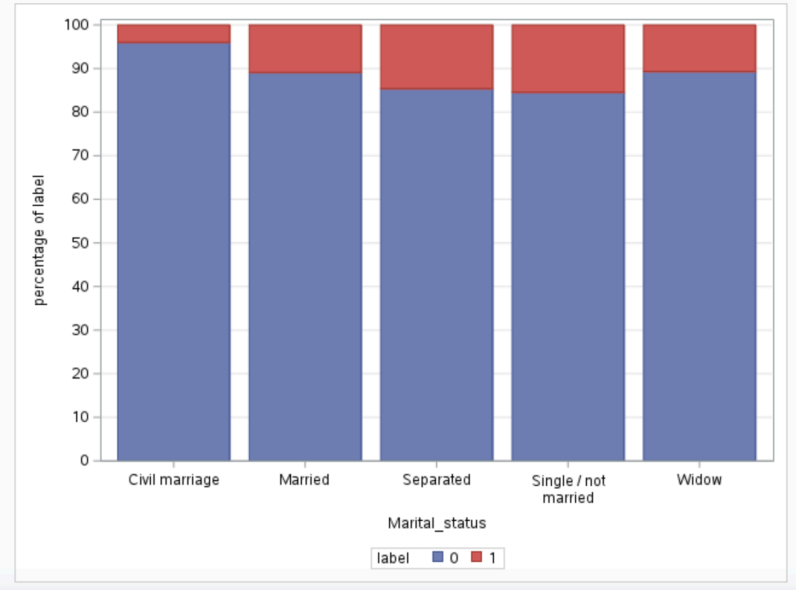


Statistics for Table of EDUCATION by label

Statistic	DF	Value	Prob
Chi-Square	4	9.2057	0.0562
Likelihood Ratio Chi-Square	4	7.9181	0.0946
Mantel-Haenszel Chi-Square	1	1.1311	0.2875
Phi Coefficient		0.0771	
Contingency Coefficient		0.0769	
Cramer's V		0.0771	

```
proc freq data=project.CreditCard3;
tables Marital_status*label / norow nocol nopercent;
run;
proc freq data=project.CreditCard3;
tables Marital_status*label / chisq fisher;
run;
proc sort data=project.CreditCard3
        out=credit2;
by Marital_status;
proc freq data=credit2 noprint;
by Marital_status;
tables label / out=freqout;
run;
proc sgplot data=freqout;
vbar Marital_status / response=percent
        group=label groupdisplay=stack;
xaxis discreteorder=data;
        yaxis grid values=(0 to 100 by 10)
        label="percentage of label";
```

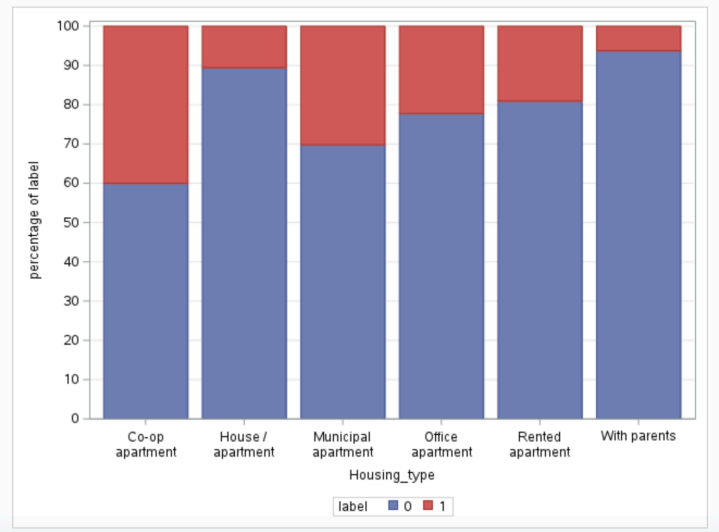
run;



Statistics for Table of Marital_status by label

Statistic	DF	Value	Prob
Chi-Square	4	10.5241	0.0325
Likelihood Ratio Chi-Square	4	11.7168	0.0196
Mantel-Haenszel Chi-Square	1	5.1835	0.0228
Phi Coefficient		0.0825	
Contingency Coefficient		0.0822	
Cramer's V		0.0825	

```
proc freq data=project.CreditCard3;
tables Housing_type*label / norow nocol nopercent;
run;
proc freq data=project.CreditCard3;
tables Housing_type*label / chisq fisher;
run;
proc sort data=project.CreditCard3
        out=credit2;
by Housing_type;
proc freq data=credit2 noprint;
by Housing_type;
tables label / out=freqout;
run;
proc sgplot data=freqout;
vbar Housing_type / response=percent
        group=label groupdisplay=stack;
xaxis discreteorder=data;
        yaxis grid values=(0 to 100 by 10)
        label="percentage of label";
run;
```

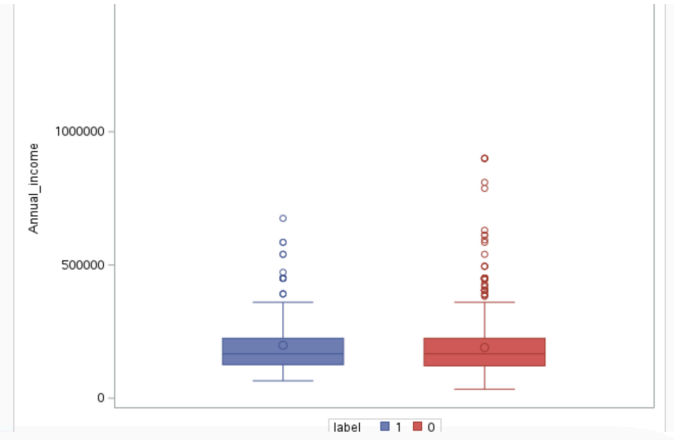


Statistics for Table of Housing_type by label

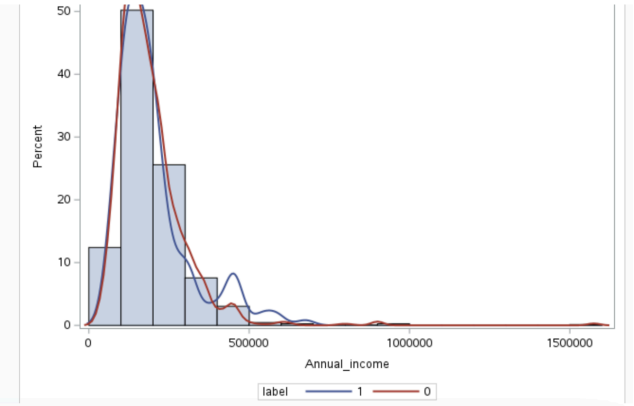
Statistic	DF	Value	Prob
Chi-Square	5	28.0428	<.0001
Likelihood Ratio Chi-Square	5	21.4823	0.0007
Mantel-Haenszel Chi-Square	1	0.0040	0.9495
Phi Coefficient		0.1346	
Contingency Coefficient		0.1334	
Cramer's V		0.1346	

```
proc freq data=project.CreditCard3;
tables Type_Occupation*label / norow nocol nopercent;
run;
proc freq data=project.CreditCard3;
tables Type_Occupation*label / chisq fisher;
run;
proc sort data=project.CreditCard3
        out=credit2;
by Type_Occupation;
proc freq data=credit2 noprint;
by Type_Occupation;
tables label / out=freqout;
run;
proc sgplot data=freqout;
vbar Type_Occupation / response=percent
        group=label groupdisplay=stack;
xaxis discreteorder=data;
        yaxis grid values=(0 to 100 by 10)
        label="percentage of label";
run;
```

```
* mumerical;
* Employed_Day VS label ;
proc freq data=project.CreditCard3;
 table Employed_Day * label / norow nocol nopct
   out=project.FreqOut(where=(percent^=.));
run;
* heatmap;
proc sgplot data=project.FreqOut noautolegend;
heatmap x=Employed_Day y=label / freq=Count
        discretex discretey
        colormodel=TwoColorRamp outline;
text x=Employed_Day y=label text=Count / textattrs=(size=16pt);
yaxis display=(nolabel) reverse;
xaxis display=(nolabel);
run;

*Annual_income VS label;
proc sgplot data=project.CreditCard3;
vbox Annual_income/group=label;
run;
```




```
proc sgplot data=project.CreditCard3;
histogram Annual_income;
density Annual_income/group=label  type=kernel;
run;
```



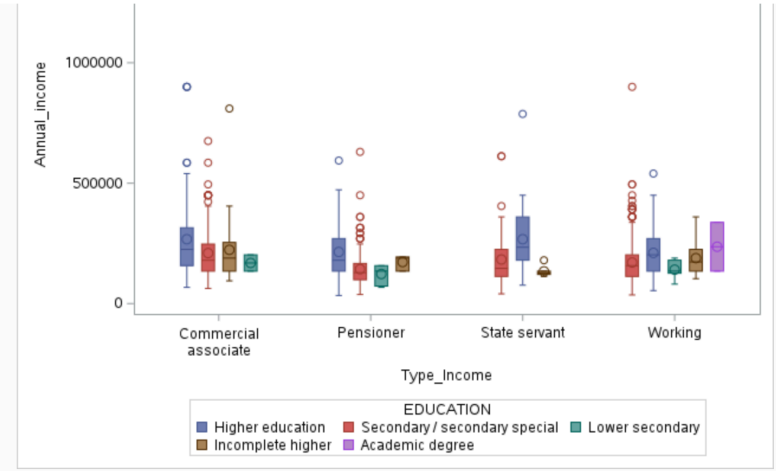

```
proc means data=project.CreditCard3 n nmiss min max mean median p1 p10 q1 median mean q3 p90 p99
maxdec=2;
```

```
class label;
var Annual_income;
run;
```

The MEANS Procedure

Analysis Variable : Annual_income													
label	N Obs	N	N Miss	Minimum	Maximum	Mean	Median	1st Pctl	10th Pctl	Lower Quartile	Upper Quartile	90th Pctl	99th Pctl
0	1373	1373	0	33750.00	1575000.00	190049.14	166500.00	49500.00	90000.00	121500.00	225000.00	315000.00	540000.00
1	175	175	0	65250.00	675000.00	198720.00	166500.00	67500.00	90000.00	126000.00	225000.00	391500.00	585000.00

```
*Annual_income VS Gender;
proc sgplot data=project.CreditCard3;
vbox Annual_income/group=GENDER;
run;
proc sgplot data=project.CreditCard3;
histogram Annual_income;
density Annual_income/group=GENDER  type=kernel;
run;
proc means data=project.CreditCard3 n nmiss min max mean median p1 p10 q1 median mean q3 p90 p99
maxdec=2;
class GENDER;
var Annual_income;
run;

*Annual_income VS CHILDREN;
proc sgplot data=project.CreditCard3;
vbox Annual_income/group=CHILDREN;
run;
proc sgplot data=project.CreditCard3;
histogram Annual_income;
density Annual_income/group=CHILDREN  type=kernel;
run;
proc means data=project.CreditCard3 n nmiss min max mean median p1 p10 q1 median mean q3 p90 p99
maxdec=2;
class CHILDREN;
var Annual_income;
run;

*Annual_income VS Type_Income;
proc sgplot data=project.CreditCard3;
vbox Annual_income/group=Type_Income;
run;
proc sgplot data=project.CreditCard3;
vbox Annual_income/category=Type_Income group=EDUCATION groupdisplay=cluster;
run;
```




```
proc sgplot data=project.CreditCard3;
histogram Annual_income;
density Annual_income/group=Type_Income  type=kernel;
run;
proc means data=project.CreditCard3 n nmiss min max mean median p1 p10 q1 median mean q3 p90 p99
maxdec=2;
class Type_Income;
var Annual_income;
run;
```

*Annual_income VS EDUCATION;

```
proc sgplot data=project.CreditCard3;
vbox Annual_income/group=EDUCATION;
run;
proc sgplot data=project.CreditCard3;
vbox Annual_income/category=EDUCATION group=label groupdisplay=cluster;
run;
proc sgplot data=project.CreditCard3;
histogram Annual_income;
density Annual_income/group=EDUCATION  type=kernel;
run;
proc means data=project.CreditCard3 n nmiss min max mean median p1 p10 q1 median mean q3 p90 p99
maxdec=2;
class EDUCATION;
var Annual_income;
run;
```

*Annual_income VS Marital_status;

```
proc sgplot data=project.CreditCard3;
vbox Annual_income/group=Marital_status;
run;
proc sgplot data=project.CreditCard3;
histogram Annual_income;
density Annual_income/group=Marital_status  type=kernel;
run;
```

```
proc means data=project.CreditCard3 n nmiss min max mean median p1 p10 q1 median mean q3 p90 p99
maxdec=2;
class Marital_status;
var Annual_income;
run;

*Annual_income VS Housing_type;
proc sgplot data=project.CreditCard3;
vbox Annual_income/group=Housing_type;
run;
proc sgplot data=project.CreditCard3;
histogram Annual_income;
density Annual_income/group=Housing_type  type=kernel;
run;
proc means data=project.CreditCard3 n nmiss min max mean median p1 p10 q1 median mean q3 p90 p99
maxdec=2;
class Housing_type;
var Annual_income;
run;

*Annual_income VS Type_Occupation;
proc sgplot data=project.CreditCard3;
vbox Annual_income/group=Type_Occupation;
run;
proc sgplot data=project.CreditCard3;
histogram Annual_income;
density Annual_income/group=Type_Occupation  type=kernel;
run;
proc means data=project.CreditCard3 n nmiss min max mean median p1 p10 q1 median mean q3 p90 p99
maxdec=2;
class Type_Occupation;
var Annual_income;
run;

*Age VS label;
proc sgplot data=project.CreditCard3;
vbox Age/group=label;
run;
proc sgplot data=project.CreditCard3;
histogram Age;
density Age/group=label  type=kernel;
run;
proc means data=project.CreditCard3 n nmiss min max mean median p1 p10 q1 median mean q3 p90 p99
maxdec=2;
class label;
var Age;
run;
```

*Employed_Year VS label;

```
proc sgplot data=project.CreditCard3;
vbox Employed_Year/group=label;
run;
proc sgplot data=project.CreditCard3;
histogram Employed_Year;
density Employed_Year/group=label  type=kernel;
run;
proc means data=project.CreditCard3 n nmiss min max mean median p1 p10 q1 median mean q3 p90 p99
maxdec=2;
class label;
var Employed_Year;
run;
/* ========================================================================== */
```

6, Diagnostics Analysis

*Chi-Square;

```
proc freq data=project.CreditCard3 order=data;
table  (GENDER Car_Owner Propert_Owner CHILDREN Type_Income EDUCATION Marital_status Housing_type
Work_Phone Phone EMAIL_ID Type_Occupation Family_Members) * label / chisq OUT=resptabl;
output all out= outfreq chisq;
run;
```

Statistics for Table of Type_Income by label

Statistic	DF	Value	Prob
Chi-Square	3	13.5986	0.0035
Likelihood Ratio Chi-Square	3	14.0933	0.0028
Mantel-Haenszel Chi-Square	1	13.1240	0.0003
Phi Coefficient		0.0937	
Contingency Coefficient		0.0933	
Cramer's V		0.0937	

Statistics for Table of Marital_status by label

Statistic	DF	Value	Prob
Chi-Square	4	10.5241	0.0325
Likelihood Ratio Chi-Square	4	11.7168	0.0196
Mantel-Haenszel Chi-Square	1	0.0202	0.8869
Phi Coefficient		0.0825	
Contingency Coefficient		0.0822	
Cramer's V		0.0825	

Statistics for Table of Housing_type by label

Statistic	DF	Value	Prob
Chi-Square	5	28.0428	<.0001
Likelihood Ratio Chi-Square	5	21.4823	0.0007
Mantel-Haenszel Chi-Square	1	18.2524	<.0001
Phi Coefficient		0.1346	
Contingency Coefficient		0.1334	
Cramer's V		0.1346	
WARNING: 33% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Statistics for Table of EDUCATION by label

Statistic	DF	Value	Prob
Chi-Square	4	9.2057	0.0562
Likelihood Ratio Chi-Square	4	7.9181	0.0946
Mantel-Haenszel Chi-Square	1	1.2366	0.2661
Phi Coefficient		0.0771	
Contingency Coefficient		0.0769	
Cramer's V		0.0771	
WARNING: 30% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Statistics for Table of Type_Occupation by label

Statistic	DF	Value	Prob
Chi-Square	5	2.3612	0.7972
Likelihood Ratio Chi-Square	5	2.4475	0.7844
Mantel-Haenszel Chi-Square	1	1.0793	0.2989
Phi Coefficient		0.0391	
Contingency Coefficient		0.0390	
Cramer's V		0.0391	

*Correlation;

```
proc corr data=project.CreditCard3;
var Annual_income Age  Employed_Year;
run;
```

The CORR Procedure

3 Variables:	Annual_income Age Employed_Year

Simple Statistics

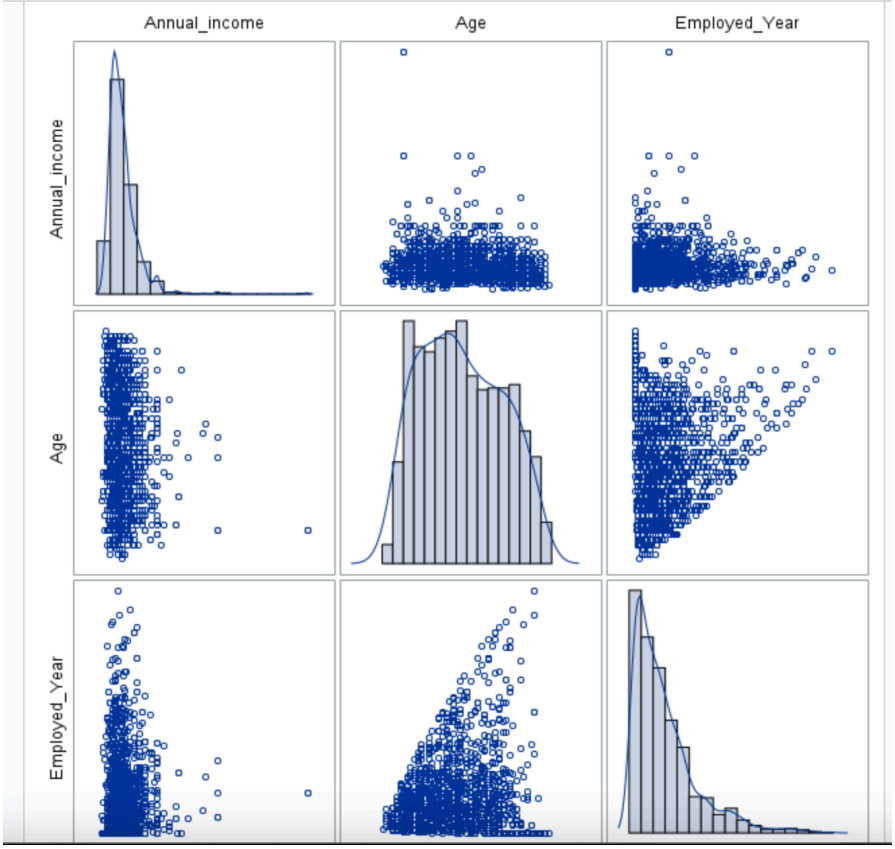
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Annual_income	1548	191029	112448	295713473	33750	1575000
Age	1548	43.39922	11.50136	67182	21.00000	68.00000
Employed_Year	1548	6.06951	6.57997	9396	0	40.80000

Pearson Correlation Coefficients, N = 1548
Prob > |r| under H0: Rho=0

	Annual_income	Age	Employed_Year
Annual_income	1.00000	-0.10975 <.0001	0.05125 0.0438
Age	-0.10975 <.0001	1.00000	-0.02552 0.3156
Employed_Year	0.05125 0.0438	-0.02552 0.3156	1.00000

```
proc sgscatter data=project.CreditCard3;
```

```
matrix Annual_income Age  Employed_Year / diagonal=(histogram kernel);
run;
```




```
/*================================================================================== */
```

/* Outliers */

* PCA;

```
proc princomp data=project.CreditCard3 out=project.pcaout;
var Annual_income Age  Employed_Year;
run;
```

Correlation Matrix

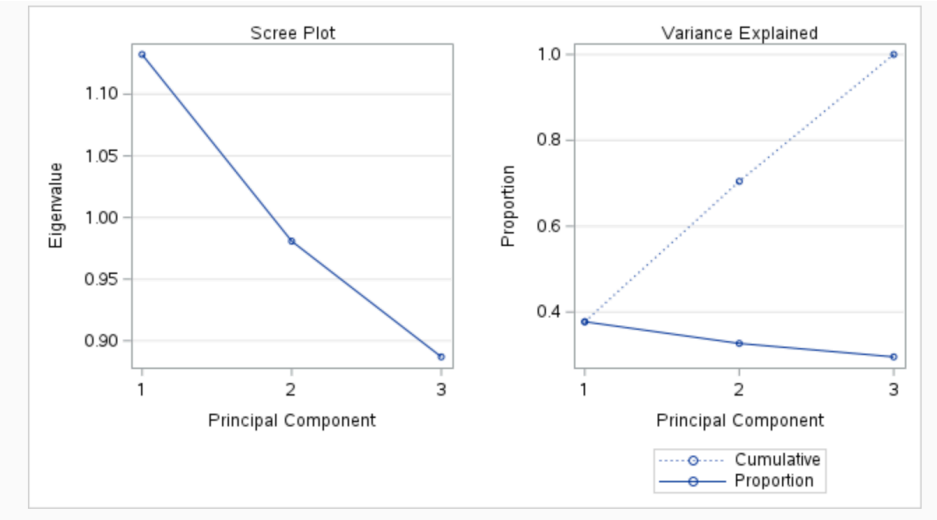
	Annual_income	Age	Employed_Year
Annual_income	1.0000	-.1097	0.0513
Age	-.1097	1.0000	-.0255
Employed_Year	0.0513	-.0255	1.0000

Eigenvalues of the Correlation Matrix

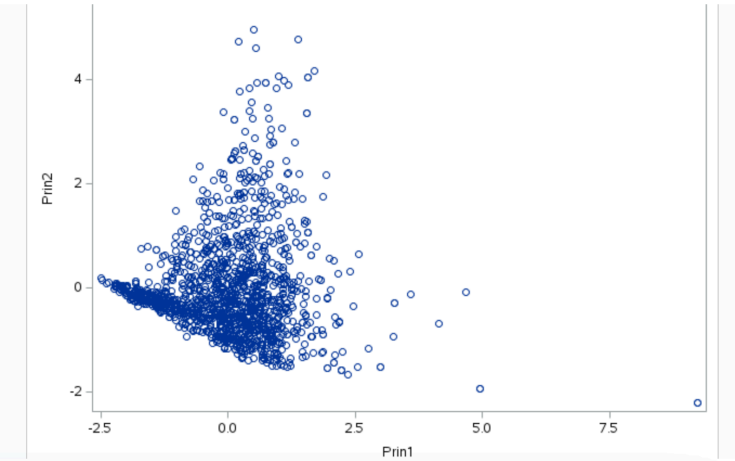
	Eigenvalue	Difference	Proportion	Cumulative
1	1.13226221	0.15146174	0.3774	0.3774
2	0.98080047	0.09386315	0.3269	0.7044
3	0.88693732		0.2956	1.0000

Eigenvectors

	Prin1	Prin2	Prin3
Annual_income	0.673246	-.141024	0.725846
Age	-.632532	0.398546	0.664127
Employed_Year	0.382941	0.906241	-.179117




```
* Visualize PCA scores;
proc  sgscatter data=project.pcaout;
 plot prin2 * prin1;
run;
```




```
data outliers1 outliers2;
set project.pcaout;
obsnumber=_n_;
if prin1>5 then output outliers1;
if prin2>4 then output outliers2;
run;
proc print data=outliers1;run;
proc print data=outliers2;run;
```

Table of Contents

t	Employed_days	Work_Phone	Phone	EMAIL_ID	Type_Occupation	Family_Members	label	Age	Employed_Day	Employed_Year	Prin1	Prin2	Prin3	obsnumber
2	-2479	0	0	0	Managers	2	0	27	2479	6.8	9.23046	-2.20333	7.96660	212
2	-2479	0	0	0	Managers	2	0	27	2479	6.8	9.23046	-2.20333	7.96660	234

t	Employed_days	Work_Phone	Phone	EMAIL_ID	Type_Occupation	Family_Members	label	Age	Employed_Day	Employed_Year	Prin1	Prin2	Prin3	obsnumber
3	-12332	0	1	1	Others	1	0	53	12332	33.8	1.55866	4.05289	0.30926	276
9	-13382	0	0	0	Unknown	2	0	64	13382	36.6	0.49699	4.94950	0.20013	312
4	-14887	0	0	0	Laborers	2	0	64	14887	40.8	0.68754	5.53924	0.02771	348
3	-12621	0	1	0	Core staff	1	0	54	12621	34.6	1.68494	4.16951	0.49046	400

```
proc corr data=project.pcaout;
var Annual_income Age  Employed_Year;
with prin1 prin2 prin3;
run;
```

The CORR Procedure

3 With Variables:	Prin1 Prin2 Prin3
3 Variables:	Annual_income Age Employed_Year

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Prin1	1548	0	1.06408	0	-2.49966	9.23046
Prin2	1548	0	0.99035	0	-2.20333	5.53924
Prin3	1548	0	0.94177	0	-1.86938	7.96660
Annual_income	1548	191029	112448	295713473	33750	1575000
Age	1548	43.39922	11.50136	67182	21.00000	68.00000
Employed_Year	1548	6.06951	6.57997	9396	0	40.80000

Pearson Correlation Coefficients, N = 1548
Prob > |r| under H0: Rho=0

	Annual_income	Age	Employed_Year
Prin1	0.71639 <.0001	-0.67306 <.0001	0.40748 <.0001
Prin2	-0.13966 <.0001	0.39470 <.0001	0.89750 <.0001
Prin3	0.68358 <.0001	0.62546 <.0001	-0.16869 <.0001

```
/* ========================================================================== */
```

/* Hypothesis Testing */

* Centering & Standardizing Values;

```
proc standard data=project.CreditCard3 out=project.CreditCard3stand mean=0 std=1;
var Annual_income Age Employed_Year;
run;
proc print data=project.CreditCard3stand;run;
```

* Mean and 95% Confidence Interva;

```
proc means data=project.CreditCard3 lclm mean uclm maxdec=2 alpha=0.05;
var Annual_income Age Employed_Year;
```

run;

The MEANS Procedure

Variable	Lower 95% CL for Mean	Mean	Upper 95% CL for Mean
Annual_income	185423.34	191029.38	196635.41
Age	42.83	43.40	43.97
Employed_Year	5.74	6.07	6.40

```
proc ttest data=project.CreditCard3 alpha=0.05 H0=0;
 var Annual_income Age Employed_Year;
run;
```

The null hypothesis for each variable's mean to be tested. In this case, the null hypothesis is that the mean is equal to 0. A small p-value (typically less than 0.05) is often interpreted as evidence against the null hypothesis.

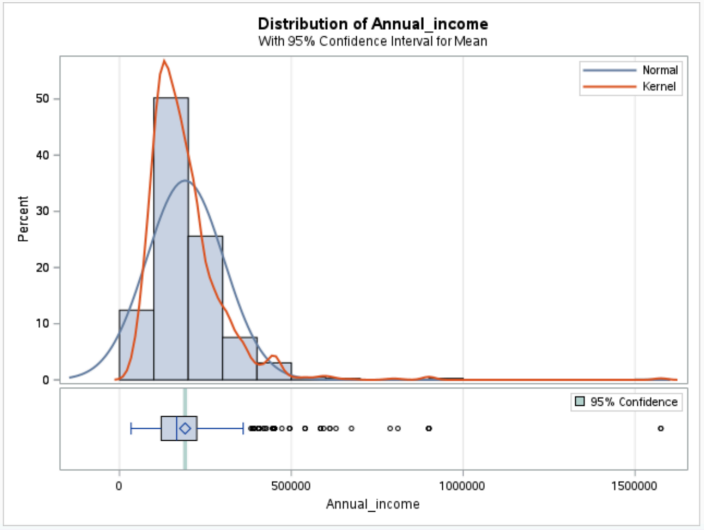
The TTEST Procedure

Variable: Annual_income

N	Mean	Std Dev	Std Err	Minimum	Maximum
1548	191029	112448	2858.0	33750.0	1575000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
191029	185423	196635	112448	108622	116556

DF	t Value	Pr > |t|
1547	66.84	<.0001

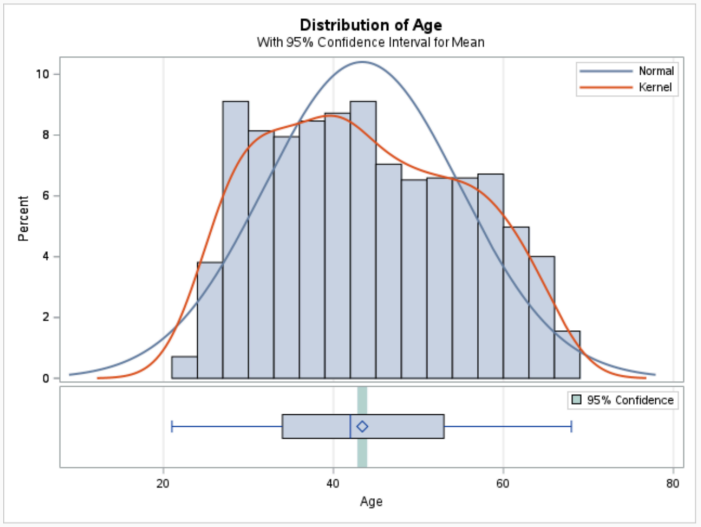


Variable: Age

N	Mean	Std Dev	Std Err	Minimum	Maximum
1548	43.3992	11.5014	0.2923	21.0000	68.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
43.3992	42.8258	43.9726	11.5014	11.1100	11.9215

DF	t Value	Pr > |t|
1547	148.46	<.0001

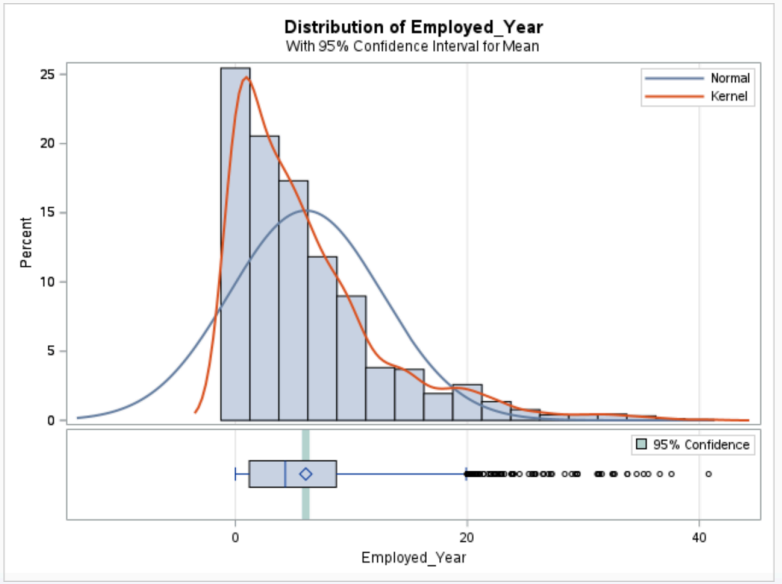


Variable: Employed_Year

N	Mean	Std Dev	Std Err	Minimum	Maximum
1548	6.0695	6.5800	0.1672	0	40.8000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
6.0695	5.7415	6.3975	6.5800	6.3561	6.8203

DF	t Value	Pr > |t|
1547	36.29	<.0001



* Values by Label_Means;

```
proc means data=project.CreditCard3 lclm mean uclm maxdec=2 alpha=0.05;
class label;
var Annual_income Age Employed_Year;
run;
```

The MEANS Procedure

label	N Obs	Variable	Lower 95% CL for Mean	Mean	Upper 95% CL for Mean
0	1373	Annual_income	184116.09	190049.14	195982.20
		Age	42.61	43.22	43.82
		Employed_Year	5.94	6.30	6.65
1	175	Annual_income	181498.72	198720.00	215941.28
		Age	43.11	44.84	46.57
		Employed_Year	3.51	4.29	5.07

* Values by Label_T-Test;

```
proc ttest data=project.CreditCard3 alpha=0.05;
 class label;
 var Annual_income Age Employed_Year;
run;
```

The TTEST Procedure

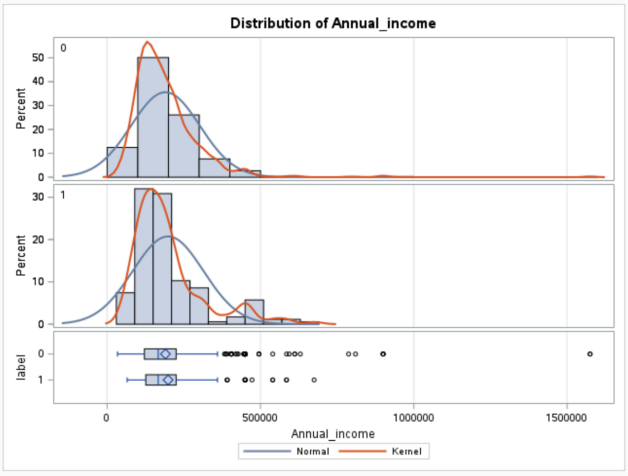
Variable: Annual_income

label	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		1373	190049	112068	3024.5	33750.0	1575000
1		175	198720	115426	8725.4	65250.0	675000
Diff (1-2)	Pooled		-8670.9	112451	9026.0		
Diff (1-2)	Satterthwaite		-8670.9		9234.7		

label	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		190049	184116	195982	112068	108028	116425
1		198720	181499	215941	115426	104468	128973
Diff (1-2)	Pooled	-8670.9	-26375.3	9033.6	112451	108624	116560
Diff (1-2)	Satterthwaite	-8670.9	-26871.7	9530.0			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	1546	-0.96	0.3369
Satterthwaite	Unequal	217.92	-0.94	0.3488

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	174	1372	1.06	0.5810



Variable: Age

label	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		1373	43.2156	11.4815	0.3099	21.0000	68.0000
1		175	44.8400	11.5881	0.8760	24.0000	67.0000
Diff (1-2)	Pooled		-1.6244	11.4936	0.9225		
Diff (1-2)	Satterthwaite		-1.6244		0.9292		

label	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		43.2156	42.6077	43.8234	11.4815	11.0676	11.9279
1		44.8400	43.1111	46.5689	11.5881	10.4880	12.9482
Diff (1-2)	Pooled	-1.6244	-3.4340	0.1852	11.4936	11.1024	11.9135
Diff (1-2)	Satterthwaite	-1.6244	-3.4556	0.2068			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	1546	-1.76	0.0785
Satterthwaite	Unequal	219.83	-1.75	0.0818

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	174	1372	1.02	0.8476

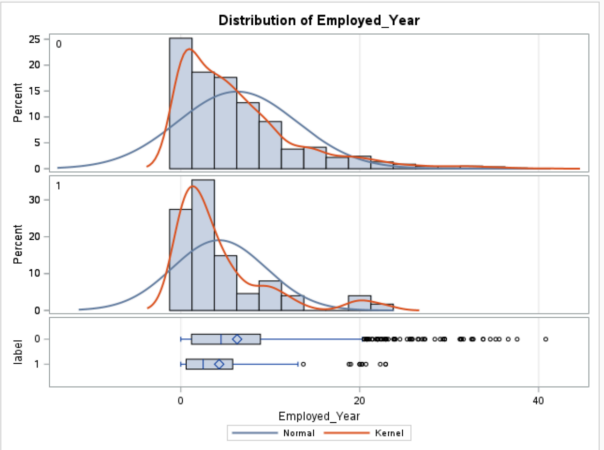
Variable: Employed_Year

label	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		1373	6.2966	6.6999	0.1808	0	40.8000
1		175	4.2880	5.2337	0.3956	0	22.9000
Diff (1-2)	Pooled		2.0086	6.5513	0.5258		
Diff (1-2)	Satterthwaite		2.0086		0.4350		

label	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		6.2966	5.9419	6.6513	6.6999	6.4583	6.9603
1		4.2880	3.5071	5.0689	5.2337	4.7369	5.8480
Diff (1-2)	Pooled	2.0086	0.9771	3.0400	6.5513	6.3283	6.7906
Diff (1-2)	Satterthwaite	2.0086	1.1519	2.8652			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	1546	3.82	0.0001
Satterthwaite	Unequal	252.88	4.62	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	1372	174	1.64	<.0001



* Mean comparison for more than 2 groups: Proc Anova: ;

```
proc anova data=project.CreditCard3 ;
 class label;
 model Annual_income=label;
 means label;
run;
```

The ANOVA Procedure

Dependent Variable: Annual_income

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	11669755005	11669755005	0.92	0.3369
Error	1546	1.954957E13	12645258636		
Corrected Total	1547	1.956124E13			

R-Square	Coeff Var	Root MSE	Annual_income Mean
0.000597	58.86589	112451.1	191029.4

Source	DF	Anova SS	Mean Square	F Value	Pr > F
label	1	11669755005	11669755005	0.92	0.3369

```
proc anova data=project.CreditCard3 ;
 class Type_Occupation;
 model Annual_income=Type_Occupation;
 means Type_Occupation;
run;
```

Class Level Information		
Class	Levels	Values
Type_Occupation	6	Core staff Laborers Managers Others Sales staff Unknown

Number of Observations Read	1548
Number of Observations Used	1548

The ANOVA Procedure

Dependent Variable: Annual_income

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1.5740422E12	314808445587	26.99	<.0001
Error	1542	1.7987197E13	11664849142		
Corrected Total	1547	1.956124E13			

R-Square	Coeff Var	Root MSE	Annual_income Mean
0.080467	56.53787	108003.9	191029.4

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Type_Occupation	5	1.5740422E12	314808445587	26.99	<.0001

```
* proc glm_Same as above;
proc glm data=project.CreditCard3;
 class label;
 model Annual_income=label;
 means label;
run;

* proc mixed;
proc mixed data=project.CreditCard3;
 class label;
 model Annual_income=label;
 lsmeans label;
run;

* Test for Equal Variance & Normality;
* Test for Normality;
ods select testsfornormality;
proc univariate data=project.CreditCard3 normal;
class Label;
var Annual_income;
run;
ods select off;
```

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	62.8375	Pr > |t|	<.0001
Sign	M	686.5	Pr >= |M|	<.0001
Signed Rank	S	471625.5	Pr >= |S|	<.0001

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.731937	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.156156	Pr > D	<0.0100
Cramer-von Mises	W-Sq	9.227633	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	54.54285	Pr > A-Sq	<0.0050

```
* Test for equal variance for Annual_income by Label;
proc glm data=project.CreditCard3;
 class Label;
 model Annual_income=Label;
 means Label / hovtest=levene;
run;
```

The GLM Procedure

Dependent Variable: Annual_income

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	11669755005	11669755005	0.92	0.3369
Error	1546	1.954957E13	12645258636		
Corrected Total	1547	1.956124E13			

R-Square	Coeff Var	Root MSE	Annual_income Mean
0.000597	58.86589	112451.1	191029.4

Source	DF	Type I SS	Mean Square	F Value	Pr > F
label	1	11669755005	11669755005	0.92	0.3369

Source	DF	Type III SS	Mean Square	F Value	Pr > F
label	1	11669755005	11669755005	0.92	0.3369

* Nonparametric Test;

```
proc npar1way data=project.CreditCard3 wilcoxon ;
 class Label;
 var age;
run;
```

Wilcoxon Two-Sample Test					
				t Approximation	
Statistic	Z	Pr > Z	Pr > |Z|	Pr > Z	Pr > |Z|
145324.0	1.7578	0.0394	0.0788	0.0395	0.0790
Z includes a continuity correction of 0.5.					

Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
3.0901	1	0.0788

* Proportion with Confidence Interval;

```
proc freq data=project.CreditCard3;
 table Label/ binomial;
run;
```

The binomial test is typically used to assess whether the observed distribution of a categorical variable deviates significantly from an expected distribution.

Binomial Proportion	
label = 0	
Proportion	0.8870
ASE	0.0080
95% Lower Conf Limit	0.8712
95% Upper Conf Limit	0.9027
Exact Conf Limits	
95% Lower Conf Limit	0.8701
95% Upper Conf Limit	0.9023

Test of H0: Proportion = 0.5	
ASE under H0	0.0127
Z	30.4489
One-sided Pr > Z	<.0001
Two-sided Pr > |Z|	<.0001

Sample Size = 1548

* Bivariate analysis for categorical (with proportion) with confidence interval;

```
proc freq data=project.CreditCard3 order=data;
 table Label* GENDER/ chisq relrisk ;
run;
```

Statistics for Table of label by GENDER

Statistic	DF	Value	Prob
Chi-Square	1	3.2279	0.0724
Likelihood Ratio Chi-Square	1	3.1711	0.0750
Continuity Adj. Chi-Square	1	2.9356	0.0866
Mantel-Haenszel Chi-Square	1	3.2258	0.0725
Phi Coefficient		0.0457	
Contingency Coefficient		0.0456	
Cramer's V		0.0457	

```
proc logistic data=project.CreditCard3;
class GENDER;
model Label(event="1") = GENDER;
ods output PredictedProbabilities=pred;
run;
```

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	3.1711	1	0.0750
Score	3.2279	1	0.0724
Wald	3.2129	1	0.0731

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
GENDER	1	3.2129	0.0731

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.0289	0.0814	621.3817	<.0001
GENDER	F	1	-0.1459	0.0814	3.2129	0.0731

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
GENDER F vs M	0.747	0.543	1.028

Association of Predicted Probabilities and Observed Responses

Percent Concordant	27.5	Somers' D	0.070
Percent Discordant	20.5	Gamma	0.145
Percent Tied	52.0	Tau-a	0.014
Pairs	240275	c	0.535

```
proc logistic data=project.CreditCard3;
class GENDER;
model Label(event="1") = GENDER;
estimate 'Gender Effect' GENDER 1 / ilink;
run;
```

```
proc logistic data=project.CreditCard3;
class Type_Income (ref="State servant");
model Label(event="1") = Type_Income;
run;
```

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Type_Income Commercial associate vs State servant	2.909	1.214	6.973
Type_Income Pensioner vs State servant	3.391	1.399	8.218
Type_Income Working vs State servant	1.957	0.833	4.601

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	40.8	Somers' D	0.156
Percent Discordant	25.2	Gamma	0.236
Percent Tied	33.9	Tau-a	0.031
Pairs	240275	c	0.578

```
proc logistic data=project.CreditCard3;
class Housing_type (ref="With parents");
model Label(event="1") = Housing_type;
run;
```

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Housing_type Co-op apartment vs With parents	10.000	1.346	74.274
Housing_type House / apartment vs With parents	1.775	0.706	4.459
Housing_type Municipal apartment vs With parents	6.486	2.206	19.074
Housing_type Office apartment vs With parents	4.286	0.699	26.281
Housing_type Rented apartment vs With parents	3.529	0.856	14.547

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	17.9	Somers' D	0.111
Percent Discordant	6.7	Gamma	0.453
Percent Tied	75.4	Tau-a	0.022
Pairs	240275	c	0.556

```
proc logistic data=project.CreditCard3;
class Marital_status (ref=" Civil marriage ");
model Label(event="1") = Marital_status;
```

run;

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Marital_status Married vs Civil marriage	2.957	1.068	8.189
Marital_status Separated vs Civil marriage	4.140	1.312	13.068
Marital_status Single / not married vs Civil marriage	4.421	1.527	12.796
Marital_status Widow vs Civil marriage	2.896	0.838	10.005

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	31.7	Somers' D	0.114
Percent Discordant	20.3	Gamma	0.220
Percent Tied	48.0	Tau-a	0.023
Pairs	240275	c	0.557

/* ========== */

7, Predictive Analysis

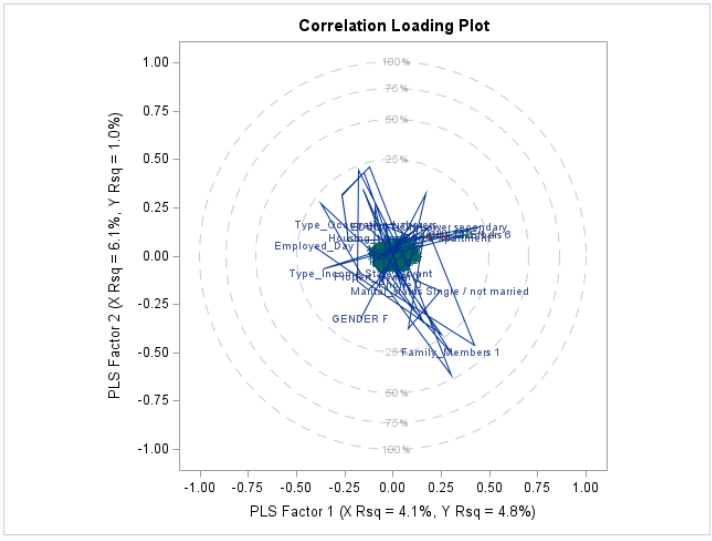
/* Model Building */

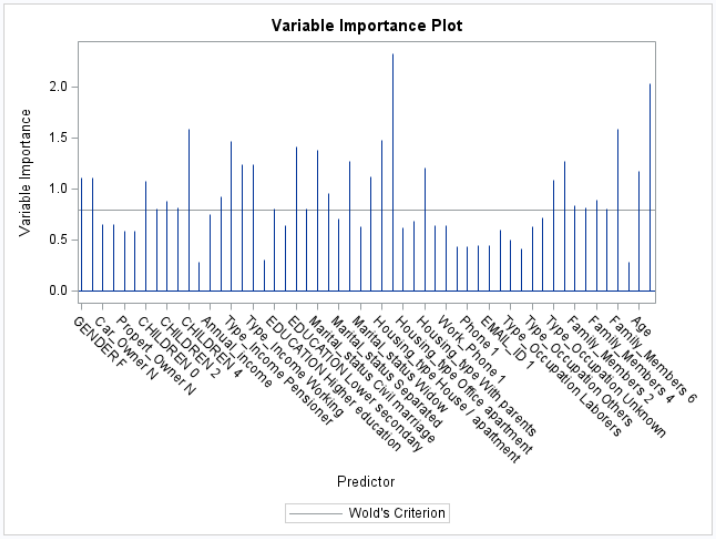
```
proc pls data=project.CreditCard3 plots=all;
class  GENDER Car_Owner Propert_Owner CHILDREN Type_Income EDUCATION Marital_status
Housing_type Work_Phone Phone EMAIL_ID Type_Occupation Family_Members;
model label= GENDER Car_Owner Propert_Owner CHILDREN Annual_income Type_Income EDUCATION
Marital_status
Housing_type Work_Phone Phone EMAIL_ID Type_Occupation Family_Members Age Employed_Day/
solution  ;
run;
```

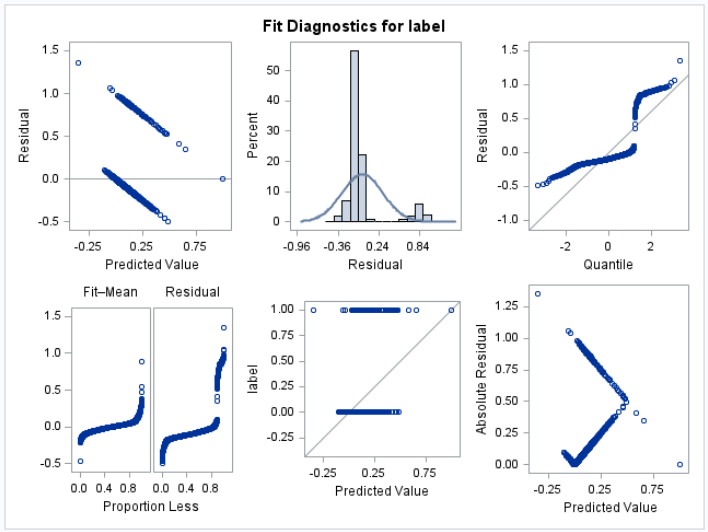
Use the Proc PLS procedure to build a multivariate statistical model and perform Partial Least Squares (PLS) analysis.

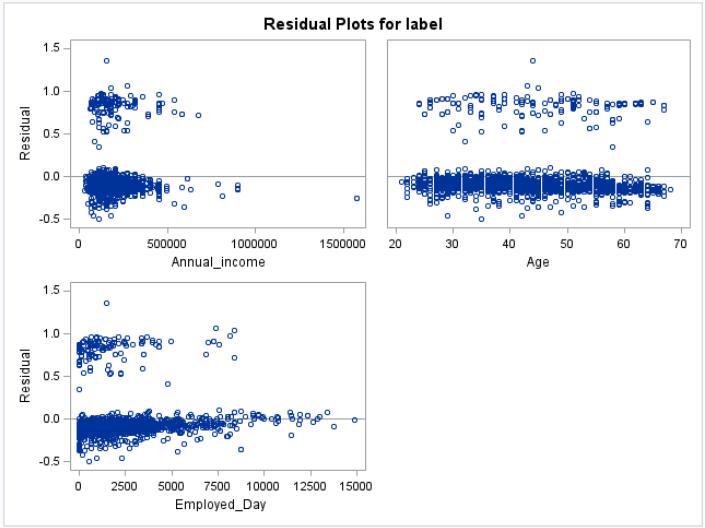
According to the Variable Importance Plot, the importance of the variables "Car_Owner" , "Propert_Owner", “Work_Phone”, “Phone” and “EMAIL_ID” to the target variable is lower than Wold's Criterion.

At the same time, we also know the importance of each category in each variable to the target variable.










```
proc logistic data=project.CreditCard3 plots(only)=(effect oddsratio);
class GENDER (ref="M") CHILDREN (ref="0") Type_Income (ref="Pensioner") EDUCATION (ref="Lower
secondary")
        Marital_status (ref="Single / not married") Housing_type (ref="Municipal apartment")
        Type_Occupation (ref="Core staff") Family_Members (ref="1") ;
model  label (event="1")= GENDER CHILDREN Type_Income EDUCATION Marital_status Housing_type
Family_Members;
run;
```

Use the Proc Logistic procedure to perform logistic regression analysis to process data with binary classification results. Based on the importance of each classification, specify a reference group for each classification variable to compare to the baseline category.

The P-value of CHILDREN is 0.4961 and the P-value of EDUCATION is 0.1323, indicating that these two variables are not significant in the model.

The p-values for Likelihood Ratio, Score and Wald tests are all less than 0.0001, strongly rejecting the null hypothesis that all coefficients are zero simultaneously. This means that at least some of the independent variables have a significant impact on the dependent variable.

The c value of 0.696 indicates that the overall predictive power of the model is moderate

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
GENDER	1	6.0773	0.0137
CHILDREN	5	4.3803	0.4961
Type_Income	3	14.6652	0.0021
EDUCATION	4	7.0681	0.1323
Marital_status	4	20.7377	0.0004
Housing_type	5	24.7811	0.0002
Family_Members	4	18.3007	0.0011

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	86.4250	26	<.0001
Score	98.7284	26	<.0001
Wald	71.6032	26	<.0001

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	68.6	Somers' D	0.393
Percent Discordant	29.3	Gamma	0.401
Percent Tied	2.0	Tau-a	0.079
Pairs	240275	c	0.696

```
proc logistic data=project.CreditCard3 plots(only)=(effect oddsratio);
class GENDER (ref="M") Type_Income (ref="Pensioner") Marital_status (ref="Single / not married")
        Housing_type (ref="Municipal apartment") Type_Occupation (ref="Core staff") Family_Members
(ref="1") ;
model  label (event="1")= GENDER Type_Income Marital_status Housing_type Family_Members;
run;
```

After removing CHILDREN and EDUCATION, two variables that were not significant in the previous model, the remaining variables in the model all showed statistical significance. However, the overall prediction accuracy of the model (as represented by the c-value) decreased slightly, from 0.696 previously to 0.679.

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
GENDER	1	6.3001	0.0121
Type_Income	3	16.3219	0.0010
Marital_status	4	25.0537	<.0001
Housing_type	5	24.1545	0.0002
Family_Members	6	17.0445	0.0091

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	66.1	Somers' D	0.358
Percent Discordant	30.3	Gamma	0.371
Percent Tied	3.6	Tau-a	0.072
Pairs	240275	c	0.679

* Proc GLM;

```
proc glm data=project.CreditCard3;
class  GENDER Type_Income Marital_status Housing_type Family_Members;
model label= GENDER Annual_income Type_Income Marital_status Housing_type Family_Members Age
Employed_Day/ solution  ss3 clparm;
lsmeans GENDER Type_Income Marital_status Housing_type Family_Members/ pdiff stderr cl;
output out = outstat1
 p = Predicted
 r = Residual
 stdr = se_resid
 student = RStudent
 h = Leverage
 cookd = CooksD
```

```
 lcl=lcl_label;
 ods output ParameterEstimates=ParamEst;
run;
quit;
```

Use the GLM model to analyze the impact of factors such as gender, income type, marital status, housing type, and number of family members on the target variables.

Almost all variables (gender, income type, marital status, housing type, and number of family members) are statistically significant, except for the category with 15 family members. This indicates that these variables have a significant impact on the model dependent variable.

GENDER	label LSMEAN	Standard Error	H0:LSMEAN=0 Pr > |t|	H0:LSMean1=LSMean2 Pr > |t|
F	0.34359589	0.07353119	<.0001	0.0341
M	0.38042909	0.07394279	<.0001	

Type_Income	label LSMEAN	Standard Error	Pr > |t|	LSMEAN Number
Commercial associate	0.39289296	0.07430782	<.0001	1
Pensioner	0.37327195	0.07772663	<.0001	2
State servant	0.32464123	0.07831168	<.0001	3
Working	0.35724382	0.07335083	<.0001	4

Marital_status	label LSMEAN	Standard Error	Pr > |t|	LSMEAN Number
Civil marriage	0.20574172	0.07874885	0.0091	1
Married	0.27193162	0.07292299	0.0002	2
Separated	0.44893441	0.08013762	<.0001	3
Single / not married	0.46619410	0.07953881	<.0001	4
Widow	0.41726060	0.08554310	<.0001	5

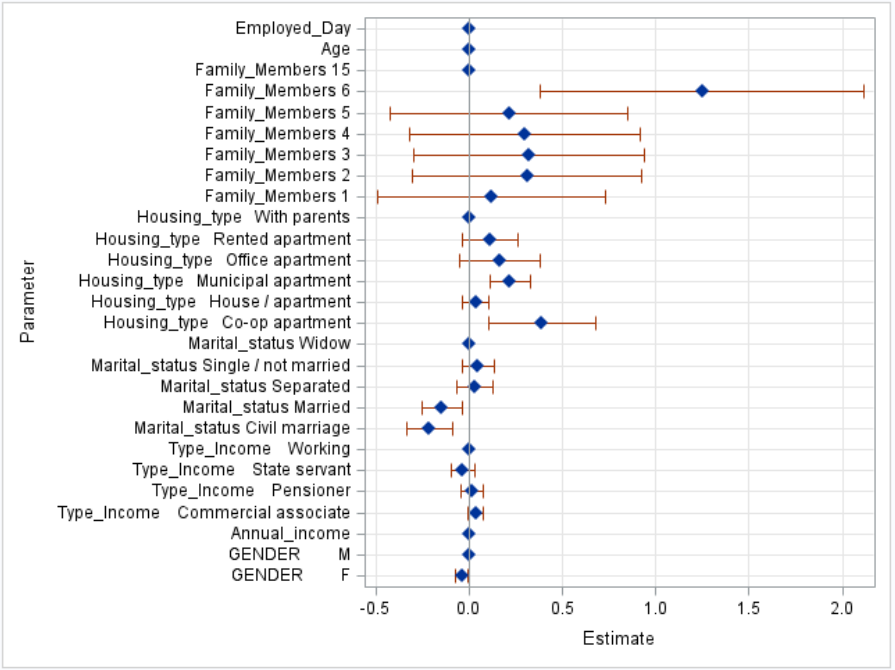
Housing_type	label LSMEAN	Standard Error	Pr > |t|	LSMEAN Number
Co-op apartment	0.59885557	0.15465264	0.0001	1
House / apartment	0.24366564	0.06610578	0.0002	2
Municipal apartment	0.42745947	0.07824310	<.0001	3
Office apartment	0.37295032	0.12290163	0.0024	4
Rented apartment	0.32154219	0.09468551	0.0007	5
With parents	0.20760176	0.07485778	0.0056	6

Family_Members	label LSMEAN	Standard Error	Pr > |t|	LSMEAN Number
1	0.12228862	0.03880571	0.0017	1
2	0.31420651	0.04121613	<.0001	2
3	0.32357524	0.04457289	<.0001	3
4	0.30132270	0.05014830	<.0001	4
5	0.21728330	0.09092050	0.0170	5
6	1.25185026	0.31322878	<.0001	6
15	0.00356080	0.31311316	0.9909	7

* visualization of coeffcient;

```
title "Parameter Estimates with 95% Confidence Limits";
proc sgplot data=ParamEst;
where Parameter ne "Intercept";
 scatter y=Parameter x=Estimate / xerrorlower=LowerCl xerrorupper=UpperCl
markerattrs=(symbol=diamondfilled) ;
 refline 0 / axis=x;
 xaxis grid;
 yaxis grid;
run;
title;
```

Use the Proc Sgplot procedure to create a scatter plot with error bars that displays the estimates of model parameters and their confidence intervals.




```
proc corr data=project.CreditCard3;
var Annual_income Age Employed_Day;
run;
```

<u>Annual_income and Age:</u>

Correlation coefficient: -0.10975, p-value: <.0001

This indicates a slight negative correlation between annual income and age, and this result is statistically significant.

<u>Annual_income and Employed_Day:</u>

Correlation coefficient: 0.05111, p-value: 0.0444

Showing a slight positive correlation between annual earnings and days of employment, this result is statistically significant, but the strength of the correlation is weak.

<u>Age and Employed_Day:</u>

Correlation coefficient: -0.02551, p-value: 0.3159

It is shown that the correlation between age and days of employment is very weak and this result is not statistically significant.

<u>Summarize</u>

A slight negative correlation between annual income and age may mean that annual income decreases slightly as age increases. However, the strength of this relationship is weak.

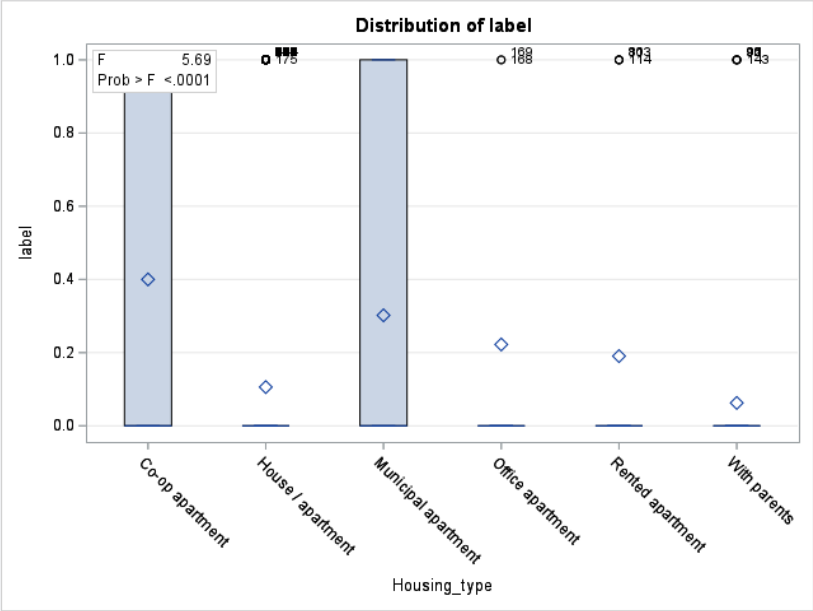
The slight positive correlation between annual earnings and days of employment may indicate that annual earnings slightly increase with longer working hours. Again, the strength of this relationship is weak.

There is little correlation between age and days employed and it is not statistically significant.

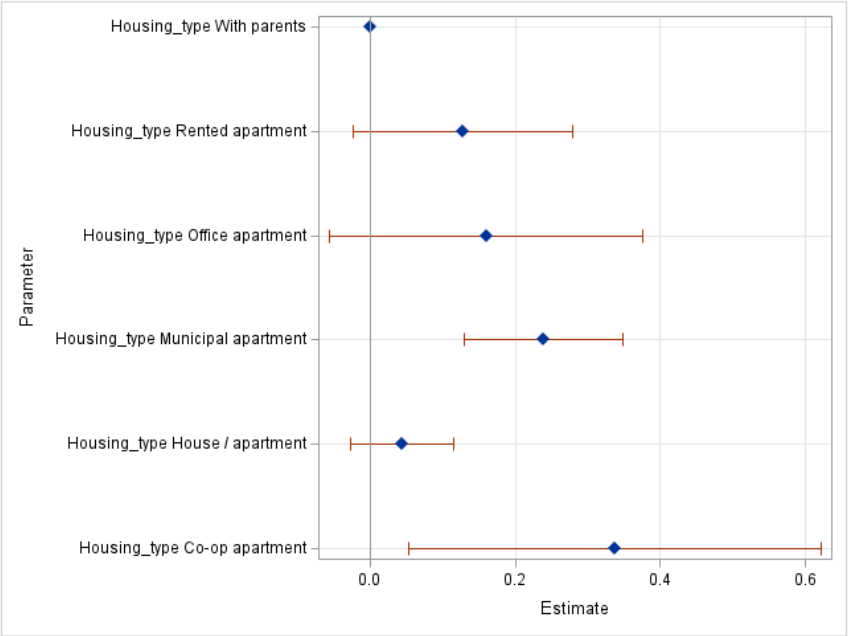
Pearson Correlation Coefficients, N = 1548 Prob > |r| under H0: Rho=0			
	Annual_income	Age	Employed_Day
Annual_income	1.00000	-0.10975 <.0001	0.05111 0.0444
Age	-0.10975 <.0001	1.00000	-0.02551 0.3159
Employed_Day	0.05111 0.0444	-0.02551 0.3159	1.00000

```
* relationship label vs Housing_type for bivariate interpretation purpose;
proc glm data=project.CreditCard3;
class Housing_type;
model label=Housing_type/ solution  clparm ;
lsmeans Housing_type;
ods output ParameterEstimates=ParamEstType;
run;
```

Dependent Variable: label					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2.8118230	0.5623646	5.69	<.0001
Error	1542	152.4045853	0.0988357		
Corrected Total	1547	155.2164083			




```
proc sgplot data=ParamEstType;
where Parameter ne "Intercept";
 scatter y=Parameter x=Estimate / xerrorlower=LowerCl xerrorupper=UpperCl
markerattrs=(symbol=diamondfilled) ;
 refline 0 / axis=x;
 xaxis grid;
 yaxis grid;
run;
```




```
/* Split Data_Train Data and Test Data*/
*  use of surveyselect for sampling;
proc surveyselect data=project.CreditCard3 rate=0.70 outall out=result seed=1234;
run;
data traindata testdata;
set result;
if selected=1 then output traindata;
else output testdata;
run;
```

Use the Proc Surveyselect procedure to randomly select samples and split the selected samples into training and test sets.

Selection Method	Simple Random Sampling

Input Data Set	CREDITCARD3
Random Number Seed	1234
Sampling Rate	0.7
Sample Size	1084
Selection Probability	0.700258
Sampling Weight	0
Output Data Set	RESULT

```
/* Data Selection */
* use of glmselect ;
proc glmselect data=traindata testdata=testdata plots(stepaxis=number)=ASEPlot;
class GENDER Type_Income Marital_status Housing_type Family_Members;
model label= GENDER Annual_income Type_Income Marital_status Housing_type Family_Members Age
Employed_Day;
score data=testdata out=testpred;
output out=outputedata p=prob_predicted r=residual;
```

run;

Use the Proc Glmselect procedure for variable selection and modeling, and the score statement for score prediction on test data.

A stepwise regression method was used, and the selection criterion was the Schwarz Bayesian criterion (SBC).

The first step introduces Employed_Day, resulting in a decrease in SBC (indicating model improvement).

An attempt was made to introduce Age, but SBC did not improve and was therefore not included in the model.

Attempted to remove Employed_Day but caused SBC to increase, so it remains in the model.

The F value is 8.85, indicating that the model is statistically significant.

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	Number Parms In	SBC	ASE	Test ASE
0	Intercept		1	1	-2546.5141	0.0948	0.1119
1	Employed_Day		2	2	-2548.3576*	0.0941	0.1103

* Optimal Value of Criterion

Stop Details

Candidate For	Effect	Candidate SBC		Compare SBC
Entry	Age	-2544.2436	>	-2548.3576
Removal	Employed_Day	-2546.5141	>	-2548.3576

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	1	0.83416	0.83416	8.85
Error	1082	101.96566	0.09424	
Corrected Total	1083	102.79982		

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	0.131350	0.012611	10.42
Employed_Day	1	-0.000011478	0.000003858	-2.98

```
* use of glmselect with Backward Selection;
proc glmselect data=traindata testdata=testdata plots(stepaxis=number)=ASEPlot;
class GENDER Type_Income Marital_status Housing_type Family_Members;
model label= GENDER Annual_income Type_Income Marital_status Housing_type Family_Members Age
Employed_Day /
        selection=backward(select=sl slstay=0.01);
score data=testdata out=testpred;
output out=outputedata p=prob_predicted r=residual;
run;
quit;
```

Using the backward selection method, non-significant variables are gradually removed until all remaining variables have a significance level below 0.01.

Removed Annual_income, GENDER, Age, Family_Members, Marital_status, and Employed_Day in order.

When trying to remove Type_Income, its significance is 0.0020, which is below the stopping criterion of 0.01, so this variable remains in the model.
The final model contains 8 independent variables. The F value is 4.64, indicating that the model is statistically significant.

Backward Selection Summary

Step	Effect Removed	Number Effects In	Number Parms In	ASE	Test ASE	F Value	Pr > F
0		9	21	0.0900	0.1062		
1	Annual_income	8	20	0.0900	0.1062	0.00	0.9957
2	GENDER	7	19	0.0902	0.1067	2.36	0.1244
3	Age	6	18	0.0903	0.1068	2.11	0.1467
4	Family_Members	5	14	0.0910	0.1092	2.02	0.0891
5	Marital_status	4	10	0.0914	0.1099	1.12	0.3478
6	Employed_Day	3	9	0.0917	0.1109	3.20	0.0738

Stop Details

Candidate For	Effect	Candidate Significance		Compare Significance	
Removal	Type_Income	0.0020	<	0.0100	(SLS)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	8	3.43140	0.42892	4.64
Error	1075	99.36842	0.09244	
Corrected Total	1083	102.79982		

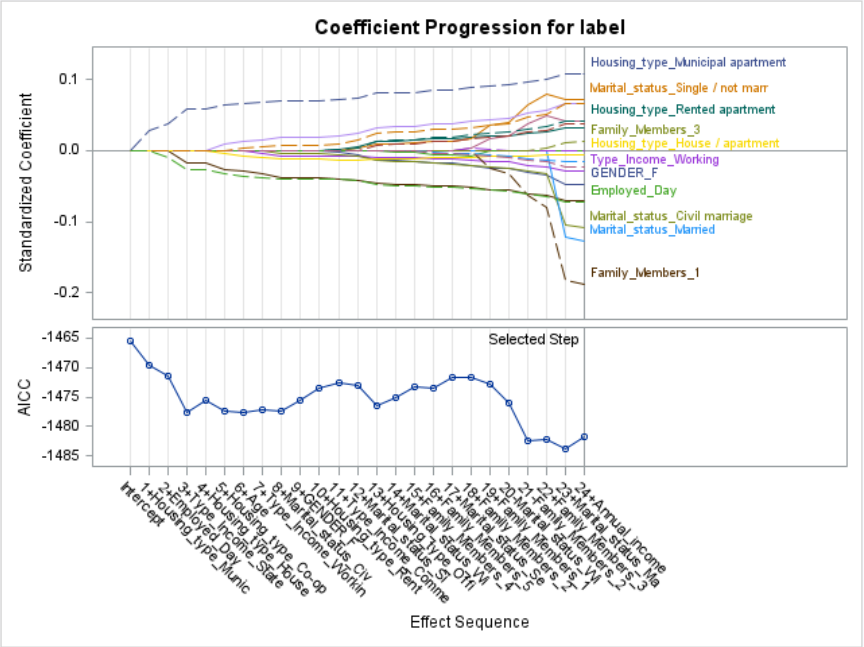
```
* use of glmselect with LASSO selectiont;
proc glmselect data=traindata testdata=testdata plots=all;
class GENDER Type_Income Marital_status Housing_type Family_Members;
model label= GENDER Annual_income Type_Income Marital_status Housing_type Family_Members Age
Employed_Day /
        selection=lasso( stop=none);
score data=testdata out=testpred;
output out=outputedata p=prob_predicted r=residual ;
run;
```

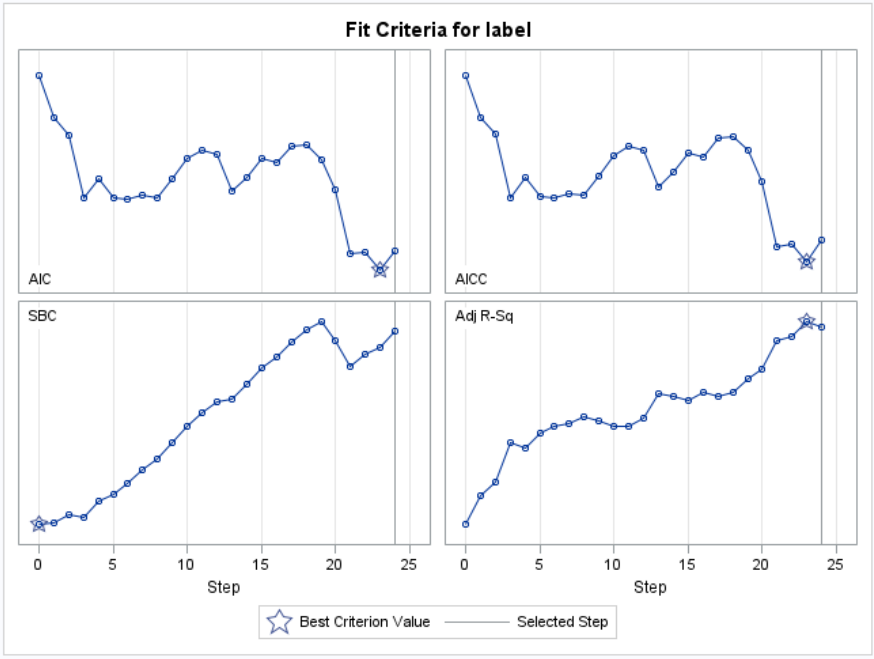
Using LASSO regression analysis, the model was fitted while retaining all variables. The model contains 20 independent variables. The F value is 2.88, indicating that the model is statistically significant and explains a certain proportion of the variability of the dependent variable.

<u>Model comparison</u>

Explanatory power: Judging from the R-squared value, the third model (LASSO regression) has the highest explanatory power.
Model complexity: The first model is the simplest and includes only one variable. The third model is the most complex and includes all variables. The second model is somewhere in between.
Statistical significance: F-values for all models indicate that the model is statistically significant, but this does not mean that the model has strong predictive power.
So choose the feature selected by the third model

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	20	5.29170	0.26458	2.88
Error	1063	97.50812	0.09173	
Corrected Total	1083	102.79982		





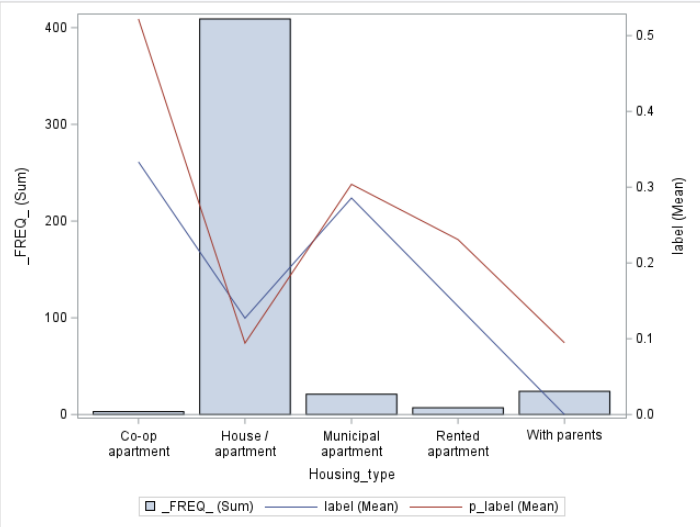
/* Categorical variable prediction */

* visualize prediction by categorical var;

```
proc summary data=testpred;
  class Housing_type;
  var label p_label;
output out= preddata  mean=;
*output out= preddata (drop= _freq_ _type_ _type_) sum=;
```

```
quit;
proc sgplot data=preddata ;
 vbar Housing_type / response=_freq_;
 vline Housing_type  / response=label  y2axis stat=mean;
 vline Housing_type /  response=p_label y2axis stat=mean;
run;
```

The mean actual label values and predicted p_label values differ for some housing types, indicating that there may be differences in the model's predictive accuracy on these categories.
There seems to be a large difference between the actual and predicted values for the housing types Co-op apartment and With parents.



* comparison predictions and output by variables;

```
proc means data=testpred;
class Housing_type;
var label p_label;
run;
```

Housing_type	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
Co-op apartment	3	label	3	0.3333333	0.5773503	0	1.0000000
		p_label	3	0.5219979	0.0529971	0.4631629	0.5659937
House / apartment	409	label	409	0.1271394	0.3335370	0	1.0000000
		p_label	407	0.0942036	0.0548182	-0.0590684	0.2564444
Municipal apartment	21	label	21	0.2857143	0.4629100	0	1.0000000
		p_label	21	0.3038317	0.0588212	0.1479016	0.4133109
Rented apartment	7	label	7	0.1428571	0.3779645	0	1.0000000
		p_label	7	0.2308804	0.0440757	0.1704352	0.2953666
With parents	24	label	24	0	0	0	0
		p_label	24	0.0946220	0.0518564	-0.0026875	0.2506401

* Data Partitioning with Lasso Regression;

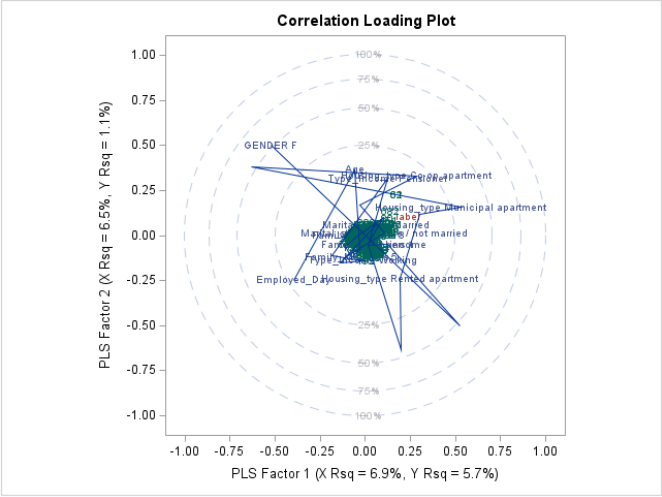
```
proc glmselect data=project.CreditCard3 plots=all;
 partition fraction (test=0.25 validate=0.25);
 class Housing_type;
model label= Annual_income Age Employed_Day Housing_type / selection=lasso(choose=cv stop=none)
cvmethod=random(3);
output out=outDataForward;
run;
```

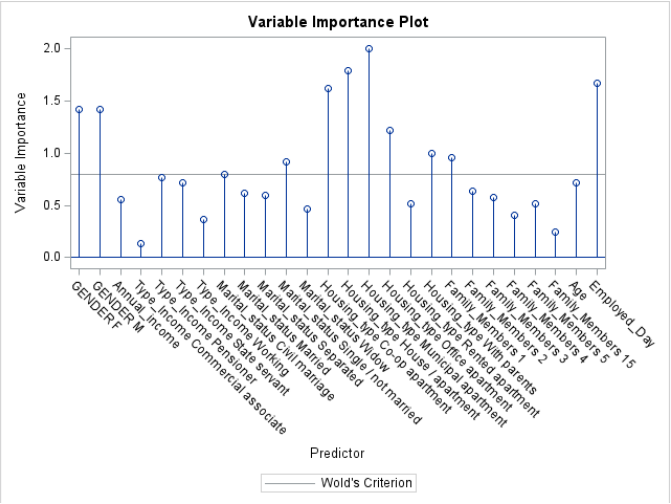
Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	2	0.82279	0.41140	3.87
Error	810	86.12186	0.10632	
Corrected Total	812	86.94465		

ASE (Train)	0.10593
ASE (Validate)	0.08237
ASE (Test)	0.10243
CV PRESS	85.18090

```
/* Split Data_Train Data, Test Data and Valid Data*/
data traindata2 testdata2 validate2;
set outDataForward;
if _ROLE_="VALIDATE" then output validate2;
else if _ROLE_="TEST" then output testdata2;
else output traindata2;
run;
```

```
/* Partial Least Squares Regression */
proc pls data=traindata2 plots=all;
class  GENDER Type_Income Marital_status Housing_type Family_Members;
model label= GENDER Annual_income Type_Income Marital_status Housing_type Family_Members Age
Employed_Day/ solution  ;
run;
quit;
```



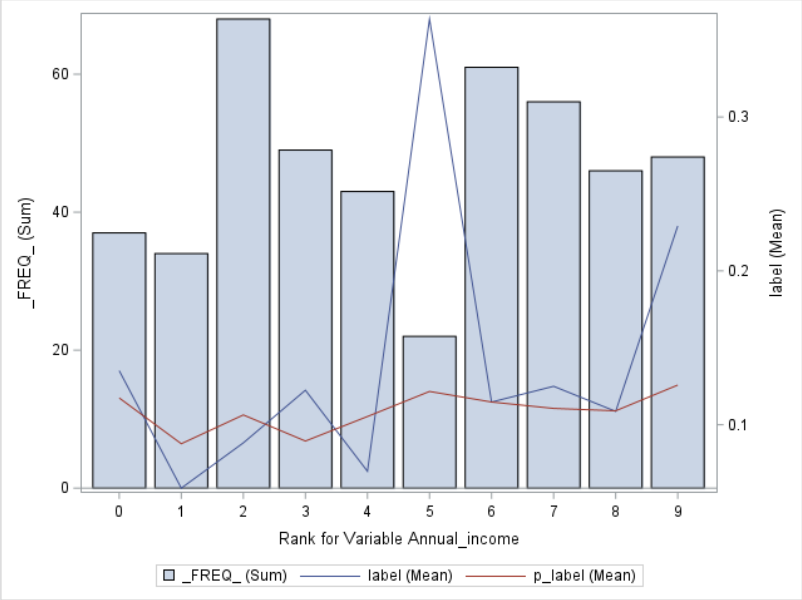



```
* use of continous variables;
* compute percentiles and compare predictions;
proc means data=testpred n min p10 p20 p30 p40 p50 p60 p70 p80 p90 max maxdec=2;
var Annual_income; run;
proc rank data=testpred out=testpred_percent groups=10;
 var Annual_income;
 ranks rank; run;
```

Analysis Variable : Annual_income											
N	Minimum	10th Pctl	20th Pctl	30th Pctl	40th Pctl	50th Pctl	60th Pctl	70th Pctl	80th Pctl	90th Pctl	Maximum
464	36000.00	90000.00	112500.00	135000.00	135000.00	162000.00	180000.00	202500.00	247500.00	315000.00	1575000.00

```
proc summary data=testpred_percent;
  class rank;
  var label p_label;
output out= preddata  mean=;
quit;

proc sgplot data=preddata ;
 vbar rank / response=_freq_;
 vline rank  / response=label  y2axis stat=mean;
 vline rank /  response=p_label y2axis stat=mean;
run;
```

The bar chart shows the frequency for each Annual_income quantile group. These groupings are created based on the results of proc rank, with each group containing a similar number of observations.
The graph shows how the average of actual label values and predicted p_label values changes across different groups of income quantiles. In some groups, the mean actual and predicted values are very close, while in other groups such as rank_5 and rank_9 there is a significant difference.
This chart evaluates the model's performance at different income levels, with the goal of understanding whether the model has the same accuracy for all income levels.




```
* comparison predictions and output by variables;
proc means data=preddata n  mean ;
class rank;
var label p_label;
run;
```

These averages differ for different income quantile groups, and this information can be used to evaluate the model's performance at different income levels.

Rank for Variable Annual_income	N Obs	Variable	N	Mean
0	1	label p_label	1 1	0.1351351 0.1174185
1	1	label p_label	1 1	0.0588235 0.0876674
2	1	label p_label	1 1	0.0882353 0.1063872
3	1	label p_label	1 1	0.1224490 0.0894264
4	1	label p_label	1 1	0.0697674 0.1053648
5	1	label p_label	1 1	0.3636364 0.1216464
6	1	label p_label	1 1	0.1147541 0.1147737
7	1	label p_label	1 1	0.1250000 0.1106199
8	1	label p_label	1 1	0.1086957 0.1090694
9	1	label p_label	1 1	0.2291667 0.1258001

/* logistic regression */

```
proc logistic data=project.CreditCard3 plots(only)=(effect oddsratio);
class   GENDER (ref="M") Type_Income (ref="Pensioner") Marital_status (ref="Single / not married")
        Housing_type (ref="Municipal apartment") Type_Occupation (ref="Core staff") Family_Members
(ref="1") /param=ref ;
model label(event="1")= Housing_type GENDER Type_Income Marital_status Family_Members
        Annual_income Age Employed_Day / details lackfit;
output out=pred p=phat lower=lcl upper=ucl predprob=(individual crossvalidate);
ods output Association=Association;
run;
quit;
```

The result of Testing Global Null Hypothesis: BETA=0 shows that at least one coefficient in the model is significantly non-zero, which is determined by three different tests (likelihood ratio, score and Wald).
Type 3 Analysis of Effects table shows that Housing_type, GENDER, Marital_status, Family_Members, and Employed_Day are significant in the model, while Type_Income, Annual_income, and Age are not.
The Odds Ratio Estimates table provides odds ratio estimates and 95% confidence intervals for different levels of categorical variables relative to the reference group.
The Association of Predicted Probabilities and Observed Responses table shows the model's prediction accuracy and related statistics. The c value is 0.692
The Partition for the Hosmer and Lemeshow Test table shows the grouped Hosmer and Lemeshow tests used to evaluate the goodness of fit of the model.
The test statistic in the Hosmer and Lemeshow Goodness-of-Fit Test table is 4.6694, the degrees of freedom are 8, and the p-value is 0.7923, which means that the model fit is good and there is no statistical poor fit.
In summary, this logistic regression model is statistically significant and shows that some variables have a significant impact on the target event label="1". The overall prediction accuracy of the model also looks good, and according to the Hosmer and Lemeshow tests, the model fits the data well.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	86.8209	22	<.0001
Score	92.5960	22	<.0001
Wald	71.2188	22	<.0001

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	69.2	Somers' D	0.384
Percent Discordant	30.8	Gamma	0.384
Percent Tied	0.0	Tau-a	0.077
Pairs	240275	c	0.692

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
4.6694	8	0.7923

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Housing_type	5	24.1874	0.0002
GENDER	1	5.2552	0.0219
Type_Income	3	5.9443	0.1144
Marital_status	4	25.7203	<.0001
Family_Members	6	18.1704	0.0058
Annual_income	1	0.0314	0.8593
Age	1	2.4921	0.1144
Employed_Day	1	10.0695	0.0015

Partition for the Hosmer and Lemeshow Test					
		label = 1		label = 0	
Group	Total	Observed	Expected	Observed	Expected
1	155	5	3.36	150	151.64
2	155	6	7.34	149	147.66
3	155	11	9.72	144	145.28
4	156	11	12.14	145	143.86
5	155	10	14.28	145	140.72
6	155	20	16.25	135	138.75
7	155	22	19.06	133	135.94
8	155	23	21.93	132	133.07
9	155	25	25.75	130	129.25
10	152	42	45.18	110	106.82

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Housing_type Co-op apartment vs Municipal apartment	3.727	0.483	28.733
Housing_type House / apartment vs Municipal apartment	0.290	0.152	0.555
Housing_type Office apartment vs Municipal apartment	0.925	0.160	5.362
Housing_type Rented apartment vs Municipal apartment	0.574	0.159	2.068
Housing_type With parents vs Municipal apartment	0.175	0.057	0.536
GENDER F vs M	0.663	0.466	0.942
Type_Income Commercial associate vs Pensioner	1.353	0.715	2.562
Type_Income State servant vs Pensioner	0.532	0.185	1.532
Type_Income Working vs Pensioner	0.915	0.499	1.681
Marital_status Civil marriage vs Single / not married	0.076	0.023	0.253
Marital_status Married vs Single / not married	0.229	0.115	0.453
Marital_status Separated vs Single / not married	0.889	0.436	1.813
Marital_status Widow vs Single / not married	0.616	0.248	1.532
Family_Members 2 vs 1	4.406	2.182	8.896
Family_Members 3 vs 1	4.950	2.194	11.167
Family_Members 4 vs 1	3.757	1.382	10.209
Family_Members 5 vs 1	<0.001	<0.001	>999.999
Family_Members 6 vs 1	>999.999	<0.001	>999.999
Family_Members 15 vs 1	<0.001	<0.001	>999.999
Annual_income	1.000	1.000	1.000
Age	1.016	0.996	1.037
Employed_Day	1.000	1.000	1.000

* ROC Curve and Sensitivity Analysis;

```
proc logistic data=traindata plots=ROC;
class Housing_type(ref="Municipal apartment") GENDER Type_Income Marital_status Family_Members
/param=ref ;
model label(event="1")= Housing_type GENDER Type_Income Marital_status Family_Members
        Annual_income Age Employed_Day / details lackfit outroc=troc;
score data=testdata out=testpred outroc=vroc;
roc; roccontrast;
output out=outputedata p=prob_predicted xbeta=linpred;
run;
quit;
```

Type 3 Analysis of Effects: Shows the Wald chi-square statistic and its corresponding p-value for each covariate in the model. These values indicate that Housing_type, Marital_status, and Employed_Day significantly affect the dependent variable label in the model.
Hosmer and Lemeshow Test: used to evaluate the goodness of fit of the model. The p value is 0.3783, which indicates that the model fits well and there is no sign of insufficient fit.
Model Convergence Status: Shows that the model has successfully converged.
ROC Association Statistics: Provides the area of the ROC curve and related statistics, such as Mann-Whitney statistics, Somers' D, etc.
ROC Contrast Test Results: Comparative test results show the comparison of the model ROC curve with random guessing (AUC=0.5). The performance of the model here is significantly better than random guessing.
The ROC curve AUC shown in the first chart is 0.6752, indicating that the model has certain discrimination ability.
The second chart shows a ROC curve with an AUC of 0.7010, which generally indicates better model performance.
The third chart compares two models, one with an ROC curve AUC of 0.6752 and the other with an AUC of 0.5 (random guessing level with no discriminating power).

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Housing_type	5	19.1111	0.0018
GENDER	1	2.8032	0.0941
Type_Income	3	7.9959	0.0461
Marital_status	4	12.2780	0.0154
Family_Members	4	8.0855	0.0885
Annual_income	1	0.0111	0.9160
Age	1	2.1492	0.1426
Employed_Day	1	4.5785	0.0324

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
8.5877	8	0.3783

Model Convergence Status

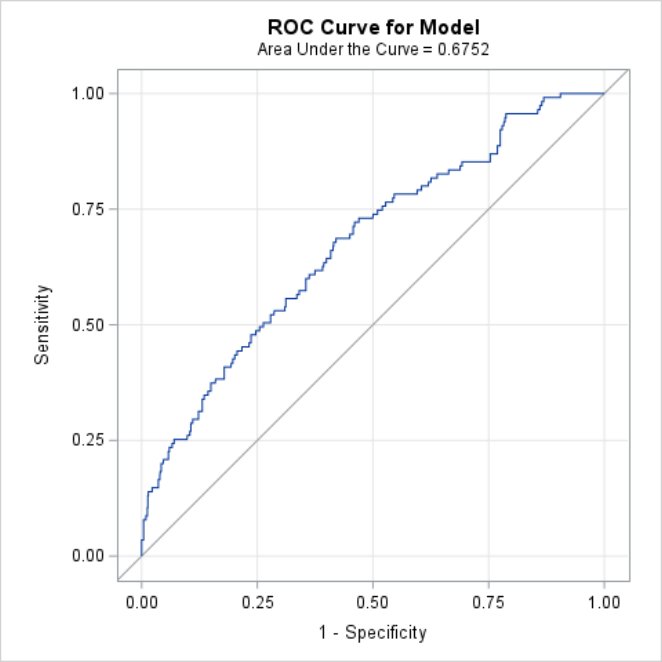
Convergence criterion (GCONV=1E-8) satisfied.

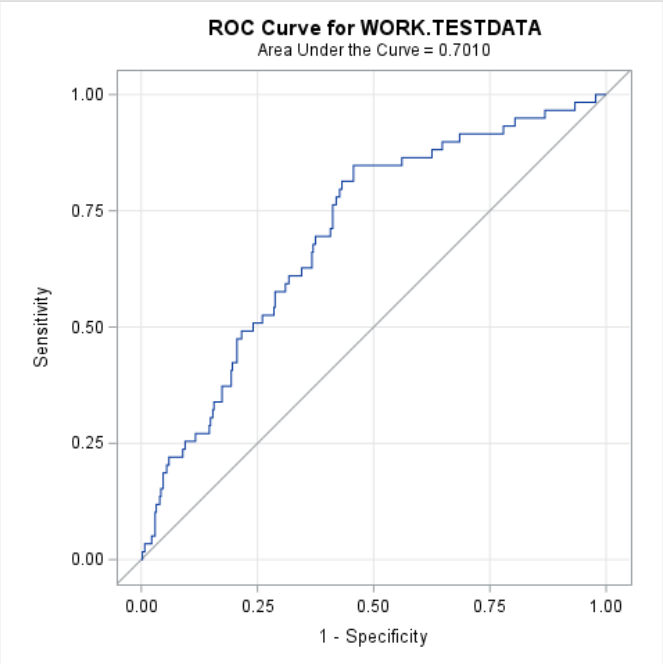
ROC Association Statistics

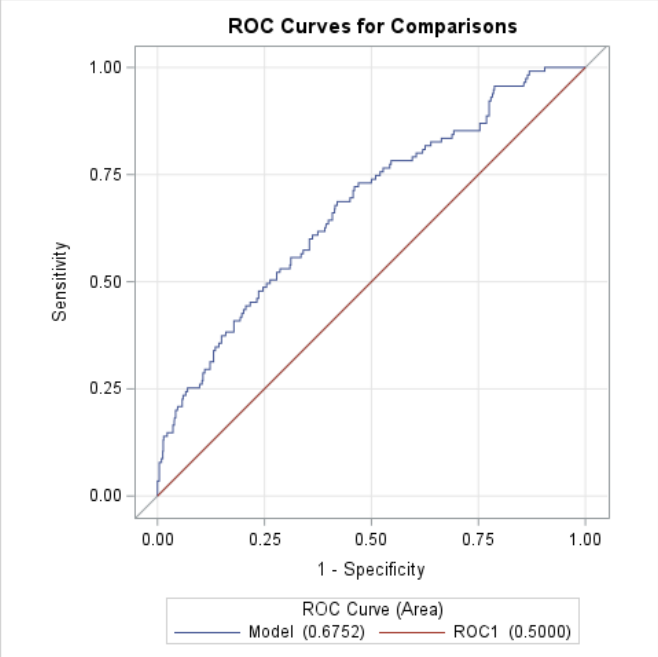
ROC Model	Mann-Whitney				Somers' D	Gamma	Tau-a
	Area	Standard Error	95% Wald Confidence Limits				
Model	0.6752	0.0268	0.6227	0.7276	0.3503	0.3504	0.0665
ROC1	0.5000	0	0.5000	0.5000	0	.	0

ROC Contrast Test Results

Contrast	DF	Chi-Square	Pr > ChiSq
Reference = Model	1	42.7791	<.0001







* Confusion matrix;

```
proc sort data=testpred;
by descending F_label descending I_label;
run;

proc freq data=testpred order=data;
tables  F_label*I_label / senspec out=CellCounts;
run;
```

Evaluate the performance of classification models. The sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of the model are calculated by comparing the predicted label (F_label(From: label)) and the actual label (I_label(Into: label))

Frequency Percent Row Pct Col Pct	Table of F_label by I_label		
	I_label(Into: label)		
F_label(From: label)	1	0	Total
1	1 0.22 1.69 33.33	58 12.55 98.31 12.64	59 12.77
0	2 0.43 0.50 66.67	401 86.80 99.50 87.36	403 87.23
Total	3 0.65	459 99.35	462 100.00

Sensitivity and Specificity				
Statistic	Estimate	Standard Error	95% Confidence Limits	
Sensitivity	0.3333	0.2722	0.0000	0.8668
Specificity	0.8736	0.0155	0.8432	0.9040
Positive Predictive Value	0.0169	0.0168	0.0000	0.0499
Negative Predictive Value	0.9950	0.0035	0.9882	1.0000

```
data CellCounts;
set CellCounts;
Match=0;
if F_label=I_label  then Match=1;
run;
proc means data=CellCounts mean;
freq count;
var Match;
run;
quit;
```

Create a new CellCounts dataset, add a new variable Match to indicate whether the prediction is correct (when F_label equals I_label), and calculate the average of the Match variable, which is the overall accuracy of the model.

The proc means output shows that the average overall accuracy (Match) is approximately 86.64%, which means that the model correctly predicted approximately 86.64% of the samples.

Analysis Variable : Match
Mean
0.8663793

8, Prescriptive Analysis

Prescriptive Solutions:

Enhance Targeted Marketing: Utilize model findings to focus on customer segments more likely to engage, such as specific Housing_type or income groups.

Improve Credit Decision-Making: Adjust credit policies based on key predictors like Employed_Day and Age to reduce default risks.

Customize Customer Services: Offer personalized services to cater to the needs of different customer segments identified by the models.

Impact on Outcome:
Increased Efficiency: Targeted marketing can lead to higher engagement and better use of resources.
Risk Reduction: Improved credit decision-making helps in lowering default rates.
Enhanced Customer Satisfaction: Personalized services lead to better customer experiences.

Deployment in Business:
Integrate Insights into Business Strategy: Apply model insights in marketing and credit risk management.
Train Teams: Educate relevant departments on leveraging these insights.
Regular Updates and Monitoring: Continuously refine strategies based on new data and feedback for ongoing improvement.

9, High Level Findings

Housing Type Significantly Correlates with Credit Card Application Outcomes:
Applicants living in co-op apartments have a higher likelihood of rejection compared to those in municipal apartments, while those living with parents have a lower likelihood. This suggests housing environment may reflect an applicant's financial stability or credit risk.

Marital Status is a Significant Risk Factor:
Single applicants have a lower risk of rejection compared to those who are married or in a civil union. This could relate to different financial circumstances associated with married individuals or those in civil unions.

Family Member Count Significantly Affects Application Outcomes:
Applicants with a higher number of family members are more likely to be rejected compared to single applicants, potentially pointing to the impact of family obligations and responsibilities on credit card approval.

Specific Personal Attributes Correlate with Application Outcomes:
Gender and days employed also show significant effects on the outcome of credit card applications, with females having a lower likelihood of rejection compared to males, and applicants with fewer days of employment having a higher likelihood of rejection.

10, Recommendations

Enhance Risk Assessment Models:
Incorporate housing type as a predictor in credit risk assessment models to better gauge financial stability. Co-op apartment residents may require additional scrutiny, whereas living with parents could be seen as a positive signal.

Marital Status as a Risk Modifier:
Adjust credit scoring algorithms to reflect the marital status of applicants. Given that single applicants tend to have lower rejection rates, marital status should be factored in to assess financial obligations that might affect creditworthiness.

Family Size Consideration:

Develop strategies for applicants with larger family sizes since they show a higher likelihood of rejection. This could involve offering financial planning resources or products tailored to applicants with dependents.

Employment History Verification:

Implement stricter verification processes for employment history as a lower number of employed days is associated with higher rejection rates. Longer employment history could be incentivized or considered a stabilizing factor in the application process.

11, Appendix

1) https://www.kaggle.com/datasets/rohitudageri/credit-card-details/data
2) Credit_card.csv
3) Credit Project_Code.sas