Final Project SAS – CRM

Report

METRO COLLEGE OF TECHNOLOGY - DATA SCIENCE AND APPLICATION

ANA CLARA TUPINAMBÁ FREITAS, mentored by Professor HAMID RAJAEE

Presented at 22nd of July, 2021

Today’s marketplace demands that businesses reduce customer turnover. This project analyses 2 years' worth of customers data of a telecommunications company with the goal of getting insights into customer’s behaviours and identify which features are key to design the best marketing strategy.

The CONTENTS Procedure

Data Set Name	PERM.CRM	Observations	102255
Member Type	DATA	**Variables**	10
Engine	V9	**Indexes**	0
Created	07/22/2021 08:57:51	**Observation Length**	72
Last Modified	07/22/2021 08:57:51	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information

Data Set Page Size	65536
Number of Data Set Pages	113
First Data Page	1
Max Obs per Page	908
Obs in First Data Page	878
Number of Data Set Repairs	0
ExtendObsCounter	YES
Filename	D:\1_Metro College\Courses\Advanced SAS\Project\Data\crm.sas7bdat
Release Created	9.0401M7
Host Created	X64_10PRO
Owner Name	ANACTF-1608\anacl
File Size	7MB
File Size (bytes)	7471104

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Label
1	Acctno	Char	14		A/c Number
2	Actdt	Num	8	MMDDYY10.	A/c Activation Date
8	Age	Num	8		Age
4	DeactReason	Char	4		Deactivation Reason
3	Deactdt	Num	8	MMDDYY10.	A/c Deactivation Date
7	DealerType	Char	2		Dealer Type
5	GoodCredit	Num	8	G_CREDIT_F.	Good Credit?
9	Province	Char	2		Province
6	RatePlan	Num	8		Rate plan
10	Sales	Num	8	DOLLAR8.2	Sales Amount

FIRST VIEW OF DATA SET

FIRST AND LAST OBSERVATIONS

Obs	Acctno	Actdt	Deactdt	DeactReason	GoodCredit	RatePlan	DealerType	Age	Province	Sales
1	1176913194483	06/20/1999	.		N	1	A1	58	BC	$128.00
2	1176914599423	10/04/1999	10/15/1999	NEED	Y	1	A1	45	AB	$72.00
3	1176951913656	07/01/2000	.		N	1	A1	57	BC	$593.00
4	1176954000288	05/30/2000	.		Y	2	A1	47	ON	$83.00
5	1176969186303	12/13/2000	.		Y	1	C1	82	BC	.
102251	2673974127660	12/29/2000	.		Y	1	A2	50		$112.00
102252	2674189951308	01/15/2001	.		Y	2	A1	40	BC	$87.00
102253	2674548796918	01/15/2001	.		Y	1	A1	16	NS	$316.00
102254	2675119766018	01/15/2001	.		Y	2	B1	76	ON	.
102255	2675135410256	01/17/2001	.		Y	1	A1	46	BC	$319.00

DATA AFTER INCLUDING SEGMENTS*

The CONTENTS Procedure

Data Set Name	**WORK.SEG**	**Observations**	**102255**
Member Type	DATA	**Variables**	16
Engine	V9	**Indexes**	0
Created	07/22/2021 08:57:52	**Observation Length**	112
Last Modified	07/22/2021 08:57:52	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information

Data Set Page Size	65536
Number of Data Set Pages	176
First Data Page	1
Max Obs per Page	584
Obs in First Data Page	559
Number of Data Set Repairs	0
ExtendObsCounter	YES
Filename	C:\Users\anacl\AppData\Local\Temp\SAS Temporary Files_TD17336_ANACTF-1608_\seg.sas7bdat
Release Created	9.0401M7
Host Created	X64_10PRO
Owner Name	ANACTF-1608\anacl
File Size	11MB
File Size (bytes)	11599872

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Label
12	ACTIVE	Char	1		
14	AGE_SEG	Num	8	AGE_F.	
1	Acctno	Char	14		A/c Number
2	Actdt	Num	8	MMDDYY10.	A/c Activation Date
9	Age	Num	8		Age
4	DeactReason	Char	4		Deactivation Reason
3	Deactdt	Num	8	MMDDYY10.	A/c Deactivation Date
8	DealerType	Char	2		Dealer Type
6	GoodCredit	Num	8	G_CREDIT_F.	Good Credit?
10	Province	Char	2		Province
7	RatePlan	Num	8		Rate plan
13	SALES_SEG	Num	8	SALES_F.	
11	Sales	Num	8	DOLLAR8.2	Sales Amount
15	TENURE	Num	8		Tenure(Days)
5	TENURE_AUX	Num	8	MMDDYY10.	
16	TENURE_SEG	Num	8	TENURE_F.	

DATA AFTER INCLUDING SEGMENTS*

Obs	Acctno	Actdt	Deactdt	DeactReason	TENURE_AUX	GoodCredit	RatePlan	DealerType	Age	Province	Sales	ACTIVE	SALES_SEG	AGE_SEG	TENURE	TENURE_SEG
1	1176913194483	06/20/1999	.		01/21/2001	N	1	A1	58	BC	$128.00	Y	$100 - $500	41 - 60 years	581	> 1 year
2	1176914599423	10/04/1999	10/15/1999	NEED	10/15/1999	Y	1	A1	45	AB	$72.00	N	< $100	41 - 60 years	11	30 days
3	1176951913656	07/01/2000	.		01/21/2001	N	1	A1	57	BC	$593.00	Y	$500 - $800	41 - 60 years	204	61 - 365 days
4	1176954000288	05/30/2000	.		01/21/2001	Y	2	A1	47	ON	$83.00	Y	< $100	41 - 60 years	236	61 - 365 days
5	1176969186303	12/13/2000	.		01/21/2001	Y	1	C1	82	BC	.	Y	Missing	> 60 years	39	31 - 60 days
102251	2673974127660	12/29/2000	.		01/21/2001	Y	1	A2	50		$112.00	Y	$100 - $500	41 - 60 years	23	30 days
102252	2674189951308	01/15/2001	.		01/21/2001	Y	2	A1	40	BC	$87.00	Y	< $100	21 - 40 years	6	30 days
102253	2674548796918	01/15/2001	.		01/21/2001	Y	1	A1	16	NS	$316.00	Y	$100 - $500	<20 years	6	30 days
102254	2675119766018	01/15/2001	.		01/21/2001	Y	2	B1	76	ON	.	Y	Missing	> 60 years	6	30 days
102255	2675135410256	01/17/2001	.		01/21/2001	Y	1	A1	46	BC	$319.00	Y	$100 - $500	41 - 60 years	4	30 days

*FIRST AND LAST OBSERVATIONS

EXPLORATORY DATA ANALYSIS(EDA)
Account Number
DUPLICATES?

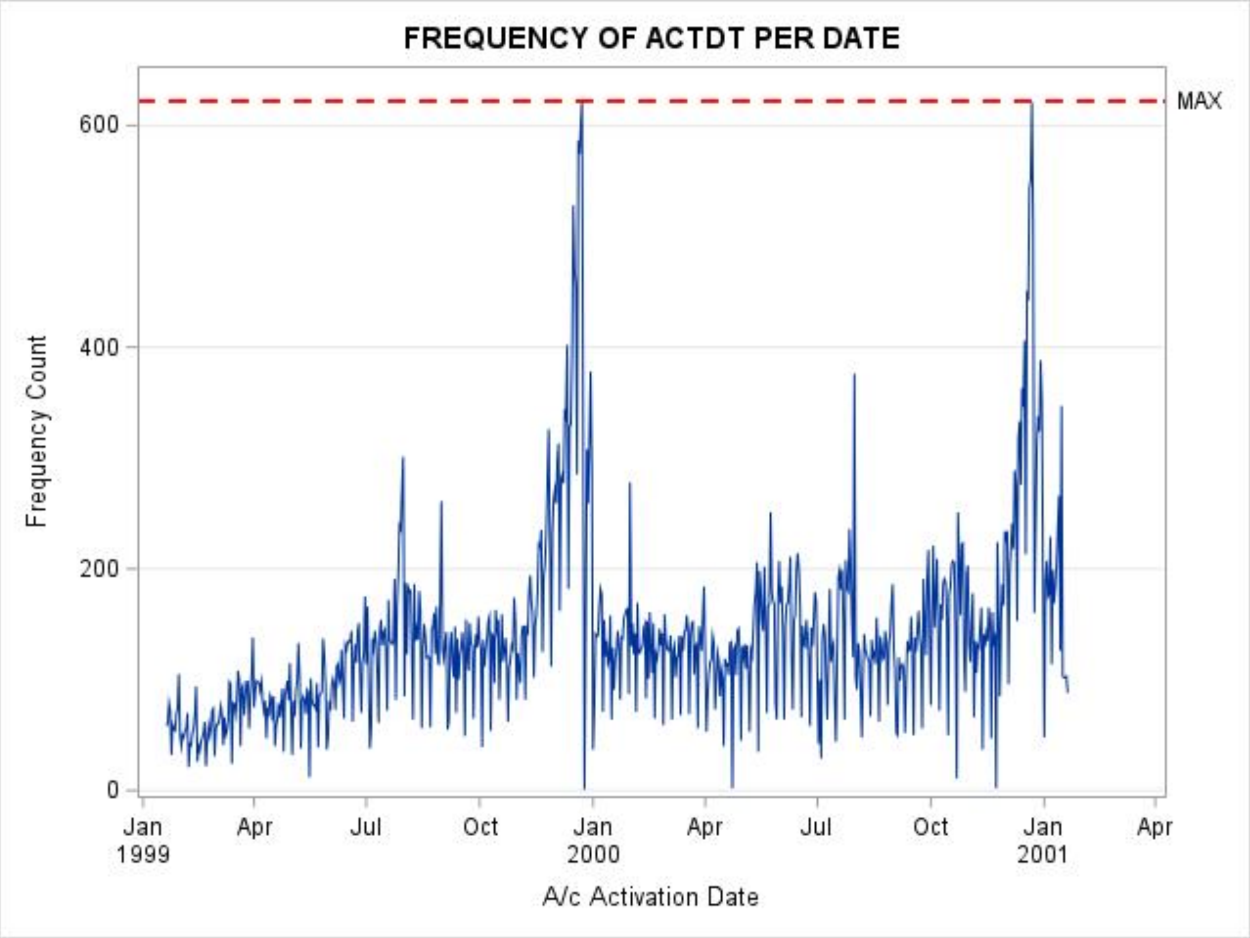
TOTAL # ACCOUNTS	TOTAL # DISTINCT ACCOUNTS	DUPLICATES?
102255	102255	There's no duplicates in data set.

ANALYSIS OF ACTDT

Obs	MIN	MAX
1	01/20/1999	01/20/2001

FREQUENCY OF ACTDT*(FIRST 10 OBSERVATIONS)

Obs	Actdt	COUNT
1	01/20/1999	58
2	01/21/1999	61
3	01/22/1999	79
4	01/23/1999	72
5	01/24/1999	32
6	01/25/1999	59
7	01/26/1999	55
8	01/27/1999	54
9	01/28/1999	69
10	01/29/1999	74



DAY WITH MOST

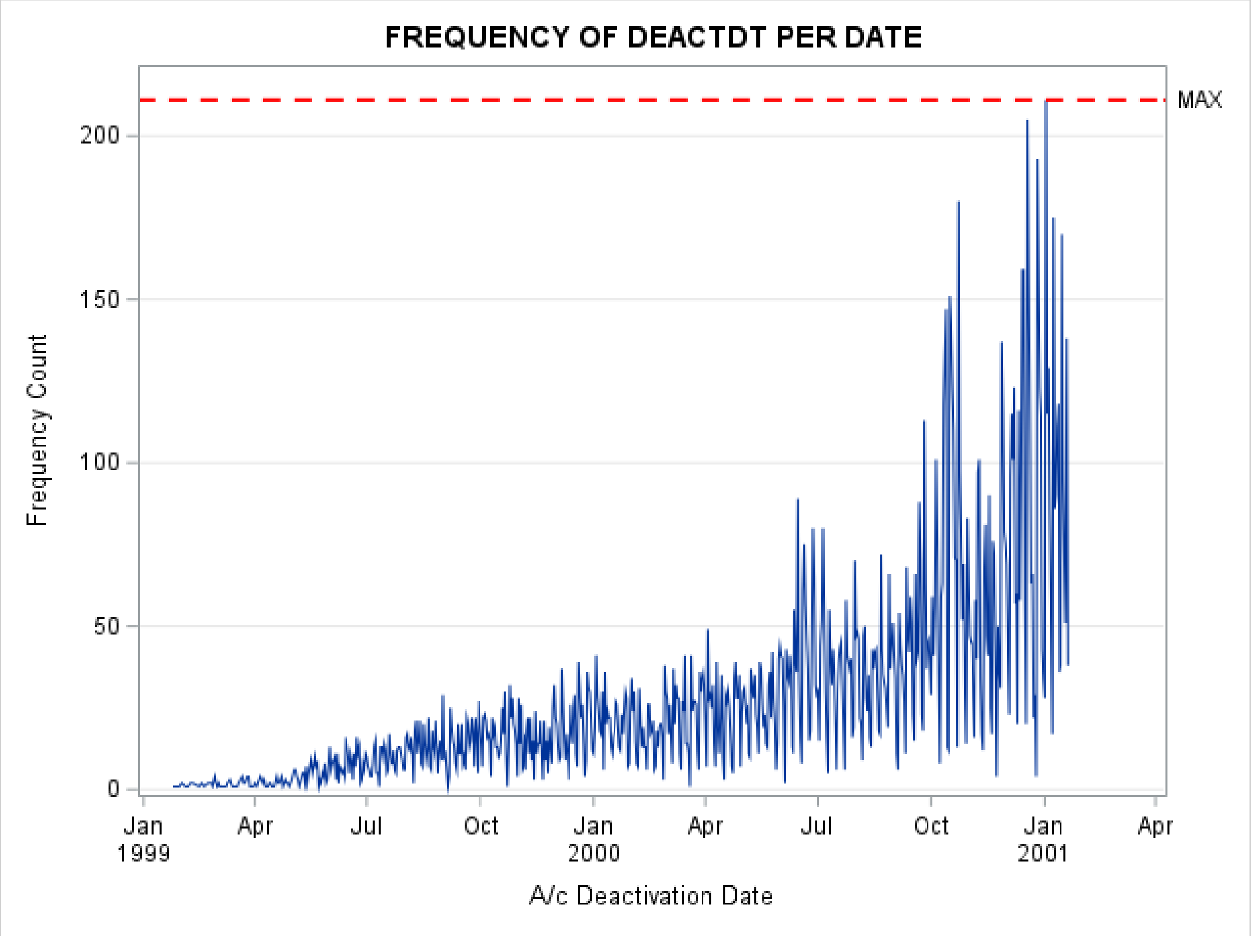
A/c Activation Date	Frequency Count
12/23/1999	622

ANALYSIS OF DEACTDT
MINIMUM AND MAXIMUM DEACTDT DATES

Obs	MIN	MAX
1	01/25/1999	01/20/2001

FREQUENCY OF DEACTDT (FIRST 10 OBSERVATIONS)

Obs	Deactdt	COUNT
1	01/25/1999	1
2	01/30/1999	1
3	02/01/1999	2
4	02/04/1999	1
5	02/06/1999	1
6	02/08/1999	2
7	02/10/1999	2
8	02/15/1999	1
9	02/17/1999	2
10	02/19/1999	1



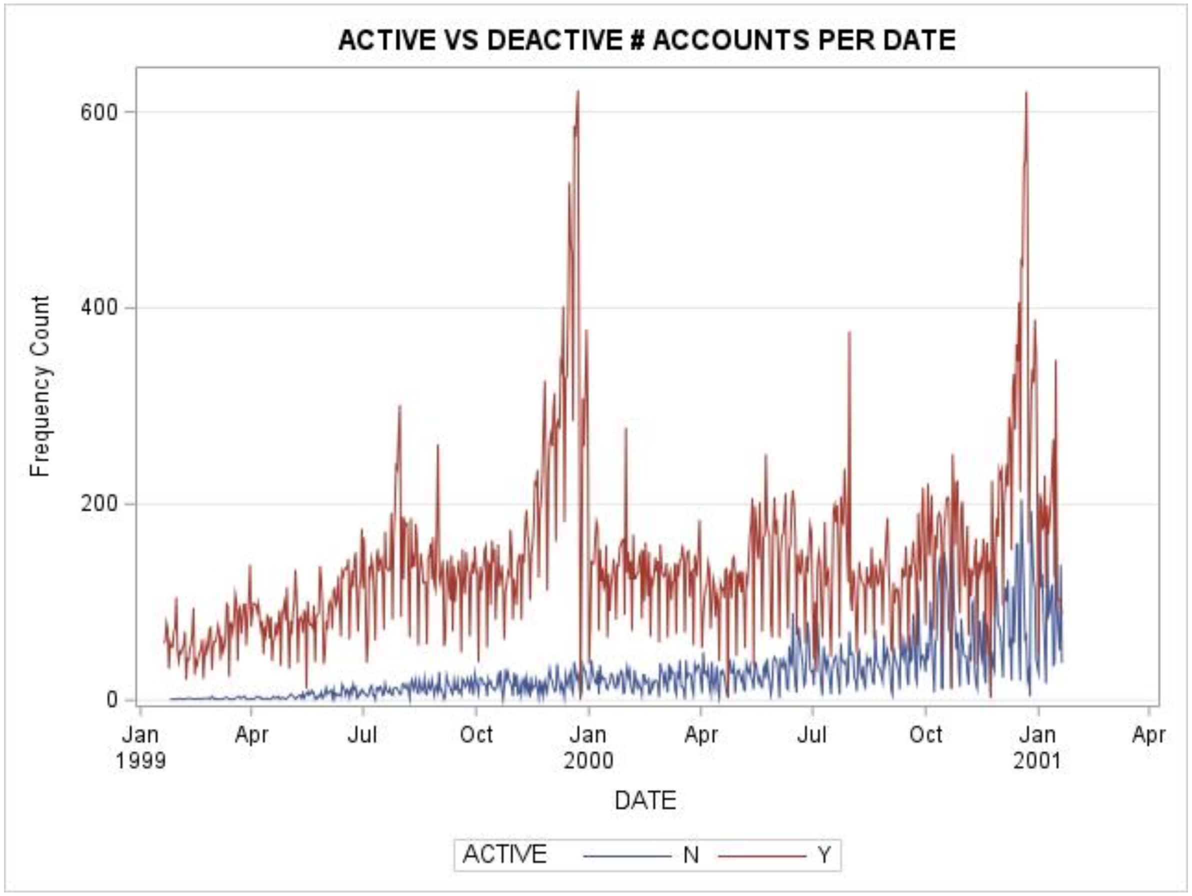
DAY WITH MOST

A/c Deactivation Date	Frequency Count
01/02/2001	211

ACTIVE VS DEACTIVE # ACCOUNTS PER DATE

Obs	DATE	ACTIVE	COUNT
1	01/25/1999	N	1
2	01/30/1999	N	1
3	02/01/1999	N	2
4	02/04/1999	N	1
5	02/06/1999	N	1
6	02/08/1999	N	2
7	02/10/1999	N	2
8	02/15/1999	N	1
9	02/17/1999	N	2
10	02/19/1999	N	1

*10 FIRST OBSERVATIONS



UNIVARIATE ANALYSIS

UNIVARIATE ANALYSIS OF DeactReason FOR SEG

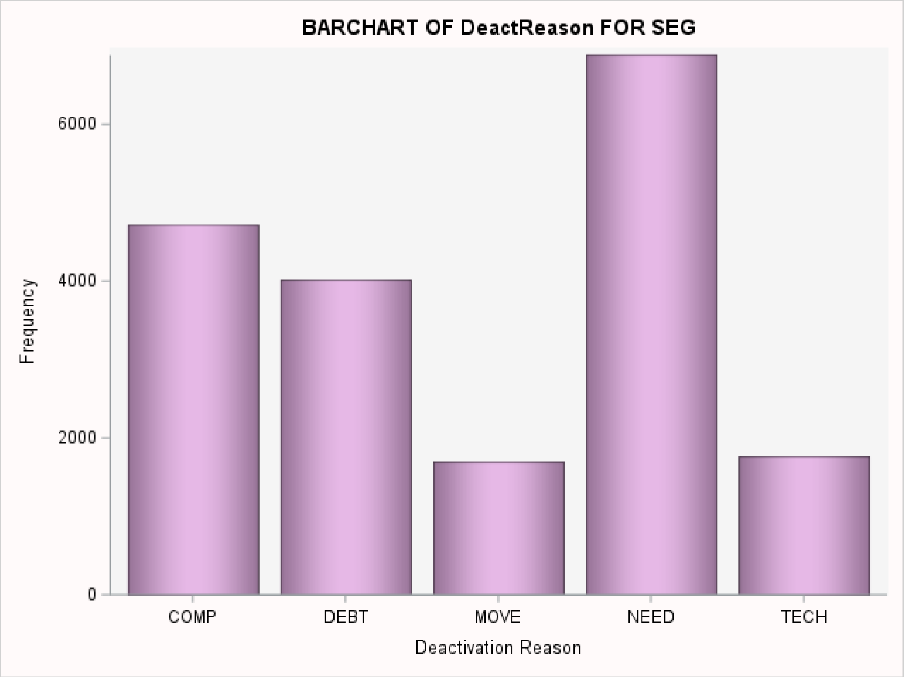
The FREQ Procedure

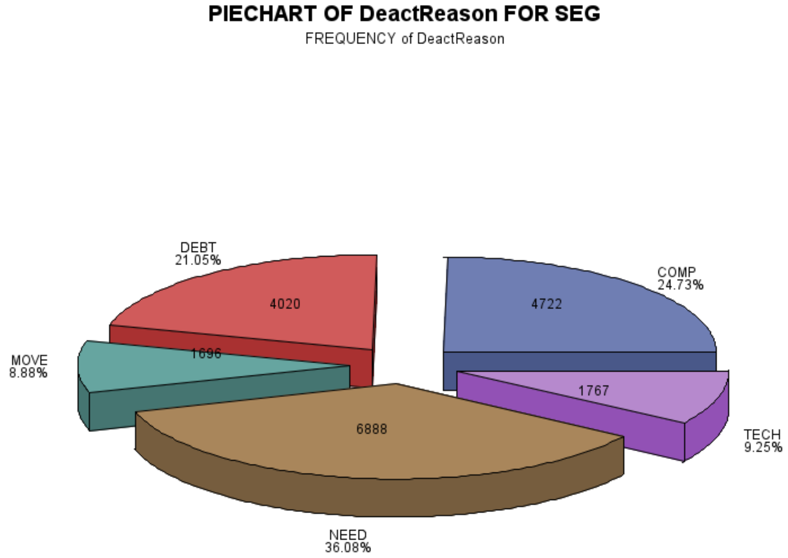
Number of Variable Levels

Variable	Label	Levels	Missing Levels	Nonmissing Levels
DeactReason	Deactivation Reason	6	1	5

Deactivation Reason

DeactReason	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	83162	81.33	83162	81.33
COMP	4722	4.62	87884	85.95
DEBT	4020	3.93	91904	89.88
MOVE	1696	1.66	93600	91.54
NEED	6888	6.74	100488	98.27
TECH	1767	1.73	102255	100.00





UNIVARIATE ANALYSIS OF GoodCredit FOR SEG

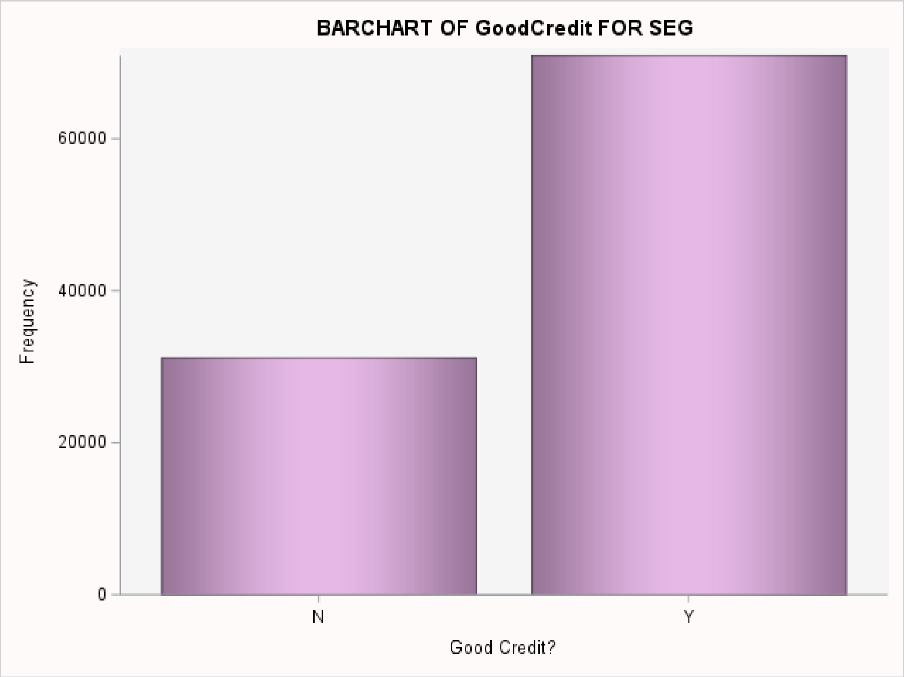
The FREQ Procedure

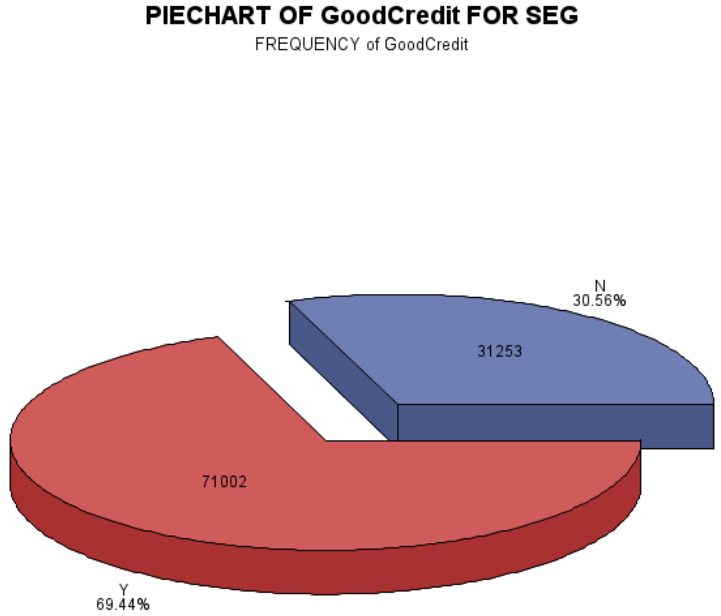
Number of Variable Levels

Variable	Label	Levels
GoodCredit	Good Credit?	2

Good Credit?

GoodCredit	Frequency	Percent	Cumulative Frequency	Cumulative Percent
N	31253	30.56	31253	30.56
Y	71002	69.44	102255	100.00





UNIVARIATE ANALYSIS OF RatePlan FOR SEG

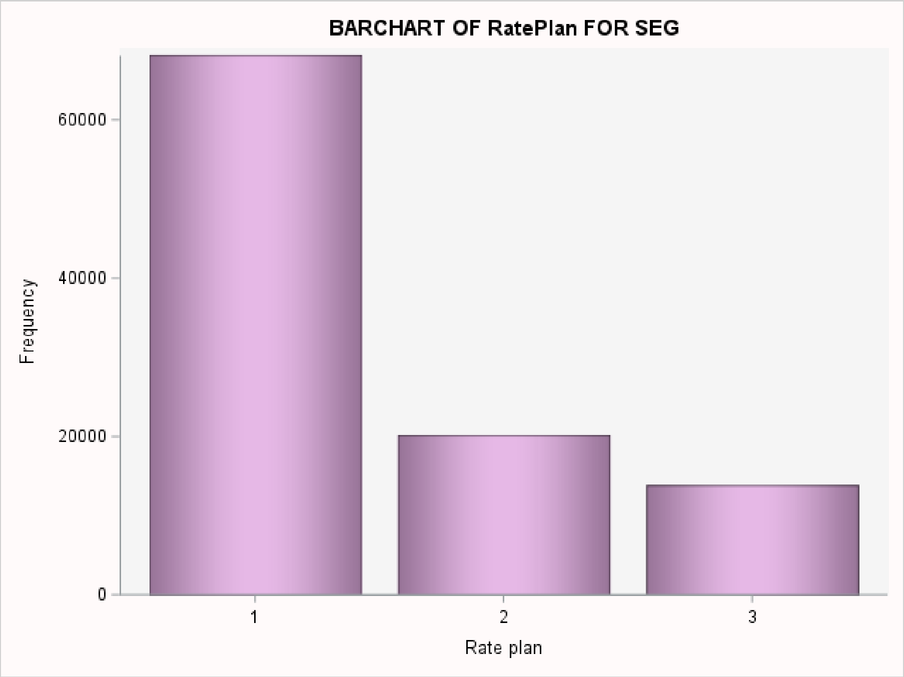
The FREQ Procedure

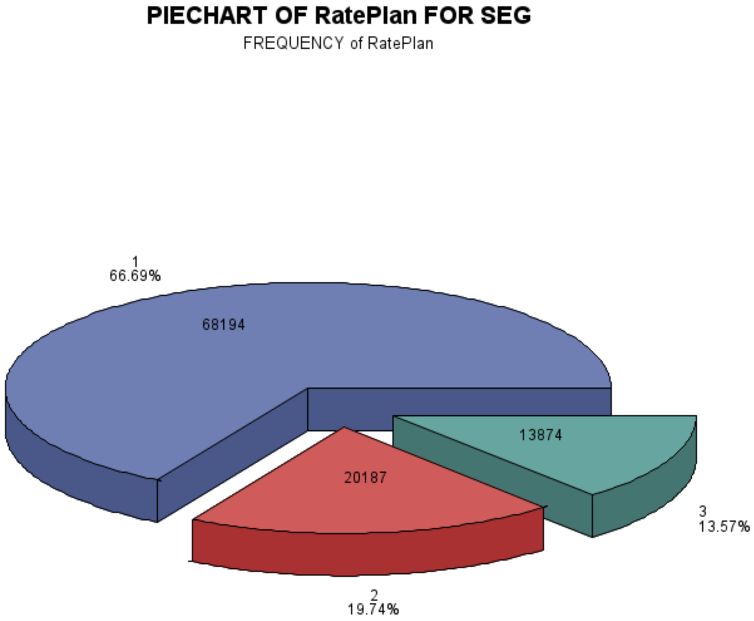
Number of Variable Levels

Variable	Label	Levels
RatePlan	Rate plan	3

Rate plan

RatePlan	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	68194	66.69	68194	66.69
2	20187	19.74	88381	86.43
3	13874	13.57	102255	100.00





UNIVARIATE ANALYSIS OF DealerType FOR SEG

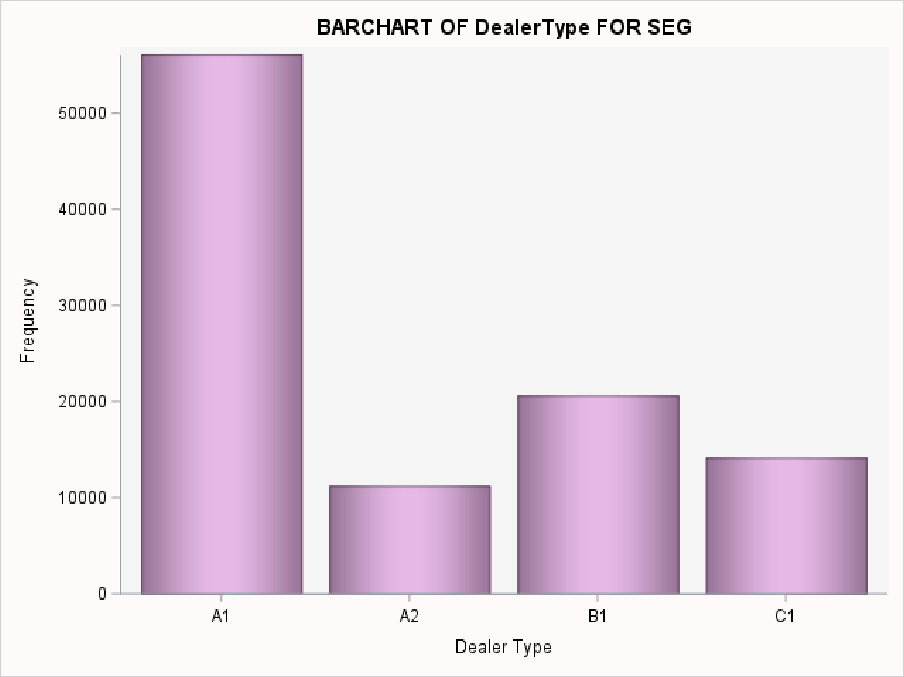
The FREQ Procedure

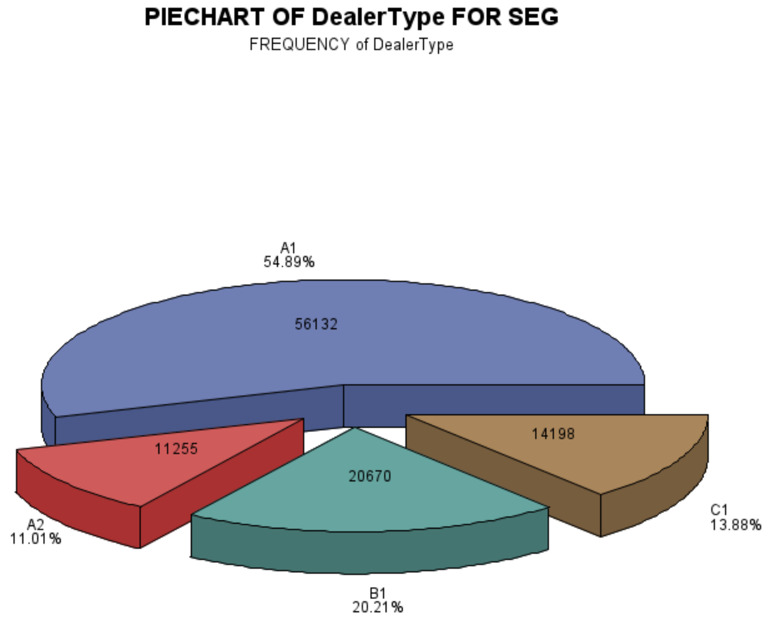
Number of Variable Levels

Variable	Label	Levels
DealerType	Dealer Type	4

Dealer Type

DealerType	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A1	56132	54.89	56132	54.89
A2	11255	11.01	67387	65.90
B1	20670	20.21	88057	86.12
C1	14198	13.88	102255	100.00





UNIVARIATE ANALYSIS OF Province FOR SEG

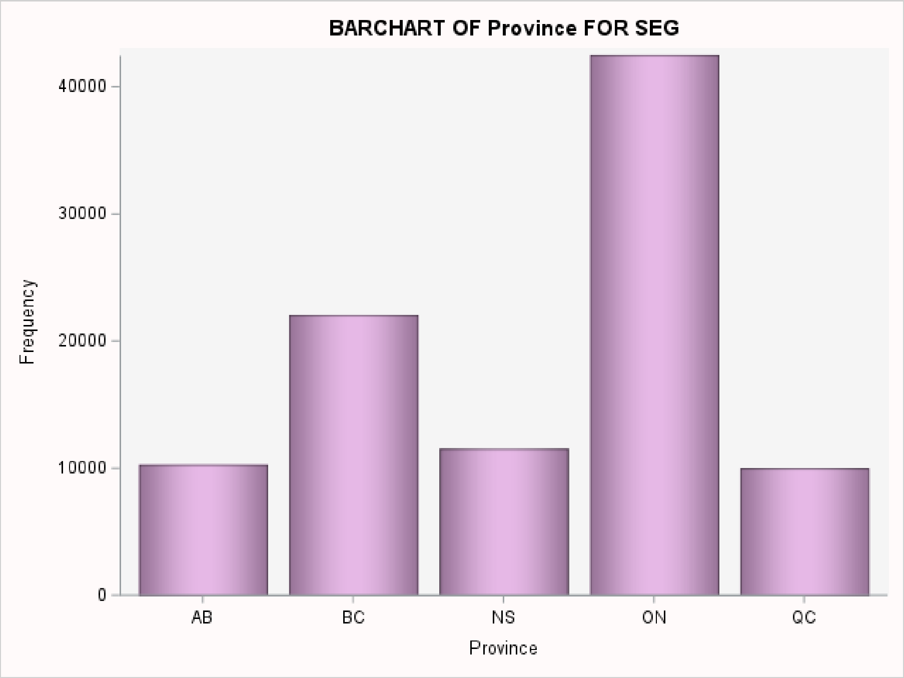
The FREQ Procedure

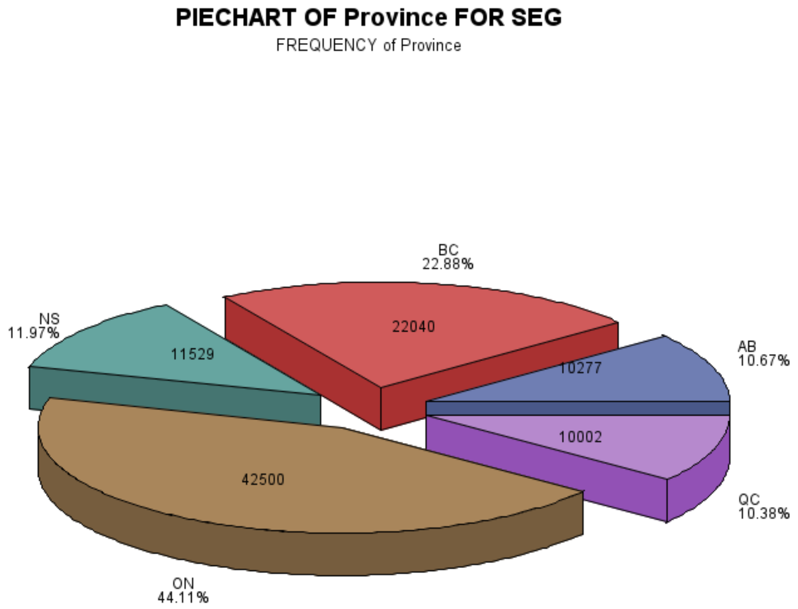
Number of Variable Levels

Variable	Label	Levels	Missing Levels	Nonmissing Levels
Province	Province	6	1	5

Province

Province	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	5907	5.78	5907	5.78
AB	10277	10.05	16184	15.83
BC	22040	21.55	38224	37.38
NS	11529	11.27	49753	48.66
ON	42500	41.56	92253	90.22
QC	10002	9.78	102255	100.00





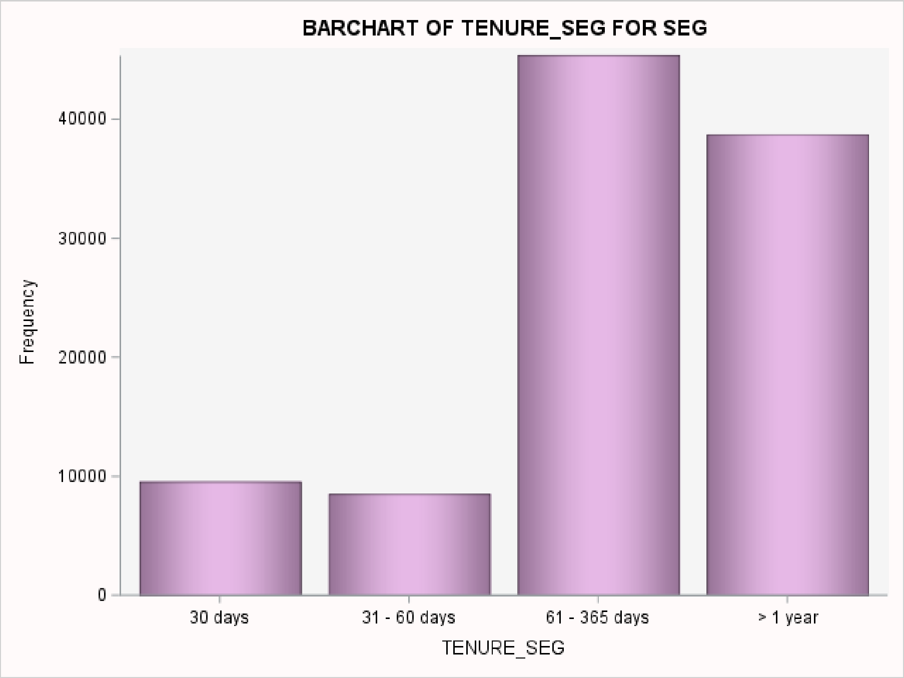
UNIVARIATE ANALYSIS OF TENURE_SEG FOR SEG

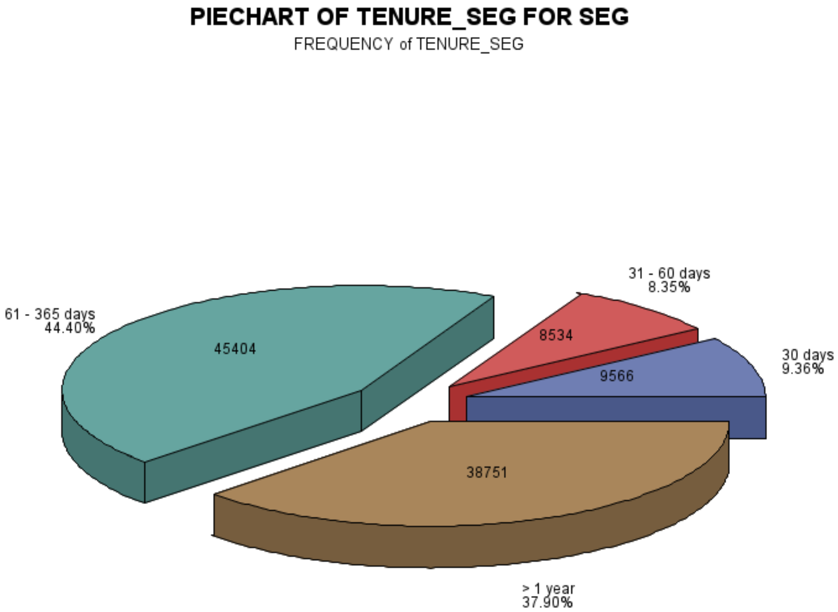
The FREQ Procedure

Number of Variable Levels

Variable	Levels
TENURE_SEG	4

TENURE_SEG	Frequency	Percent	Cumulative Frequency	Cumulative Percent
30 days	9566	9.36	9566	9.36
31 - 60 days	8534	8.35	18100	17.70
61 - 365 days	45404	44.40	63504	62.10
> 1 year	38751	37.90	102255	100.00





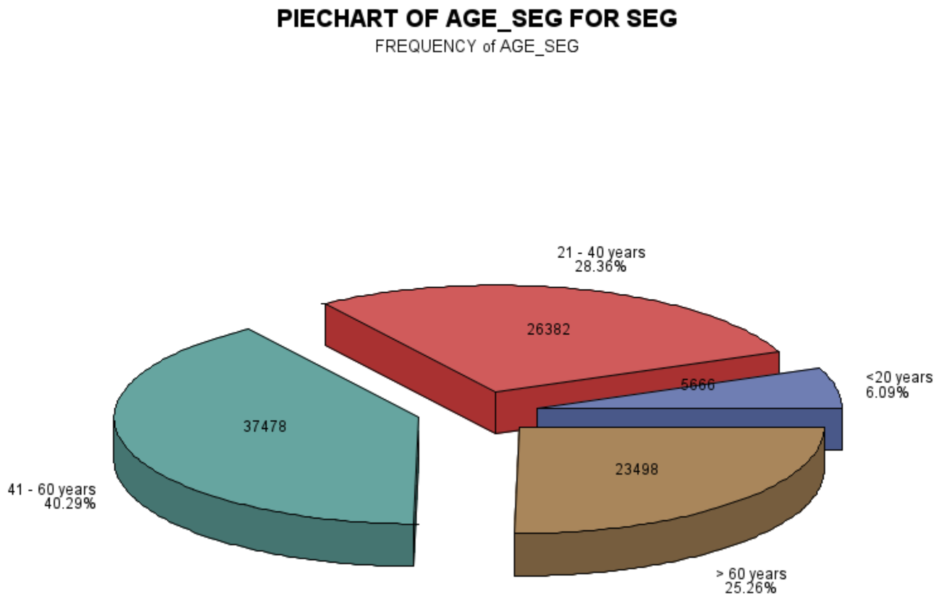
UNIVARIATE ANALYSIS OF AGE_SEG FOR SEG

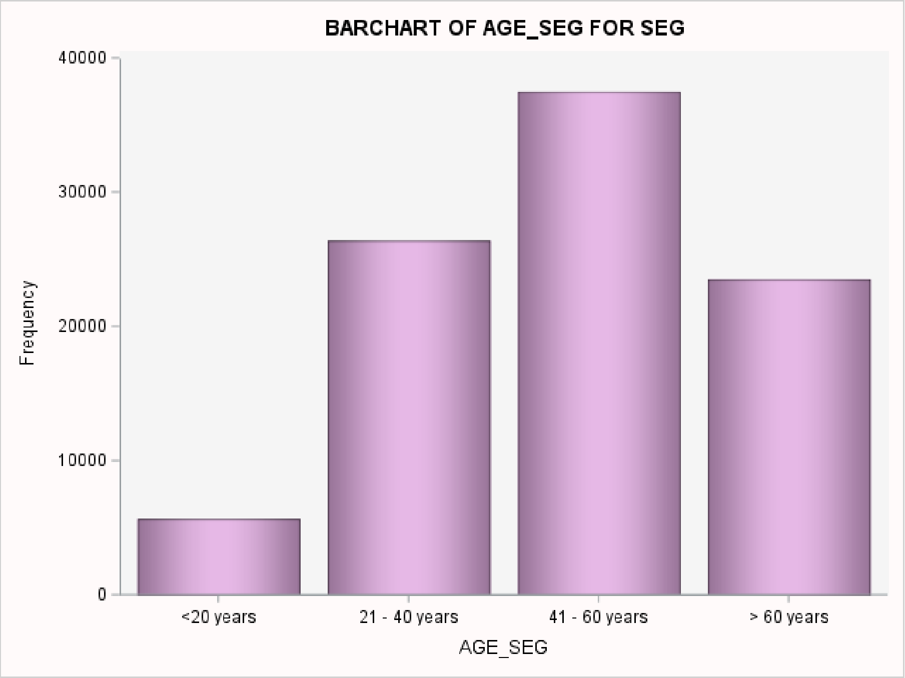
The FREQ Procedure

Number of Variable Levels

Variable	Levels	Missing Levels	Nonmissing Levels
AGE_SEG	5	1	4

AGE_SEG	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing	9231	9.03	9231	9.03
<20 years	5666	5.54	14897	14.57
21 - 40 years	26382	25.80	41279	40.37
41 - 60 years	37478	36.65	78757	77.02
> 60 years	23498	22.98	102255	100.00





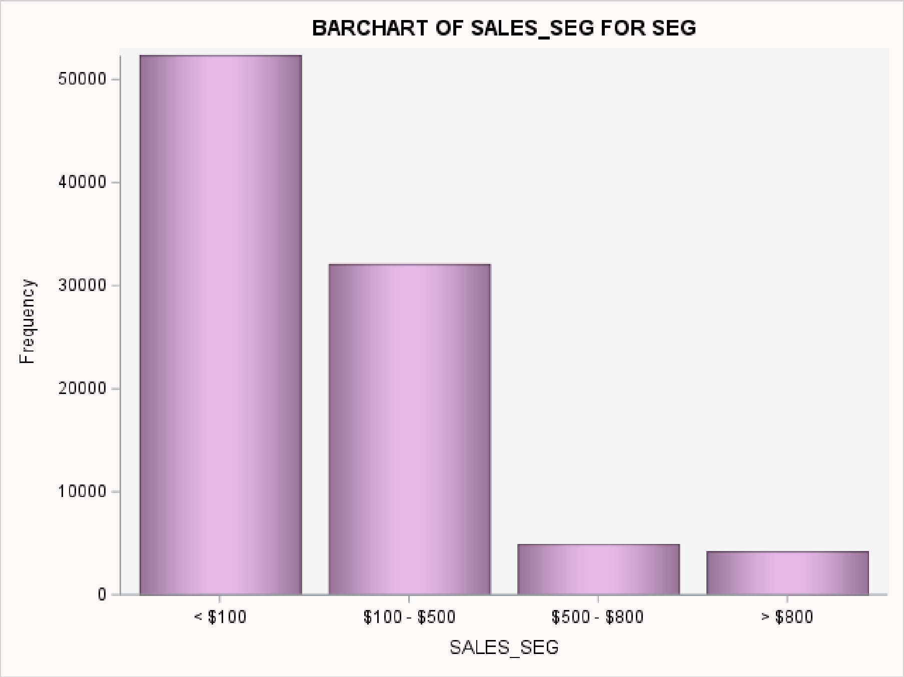
UNIVARIATE ANALYSIS OF SALES_SEG FOR SEG

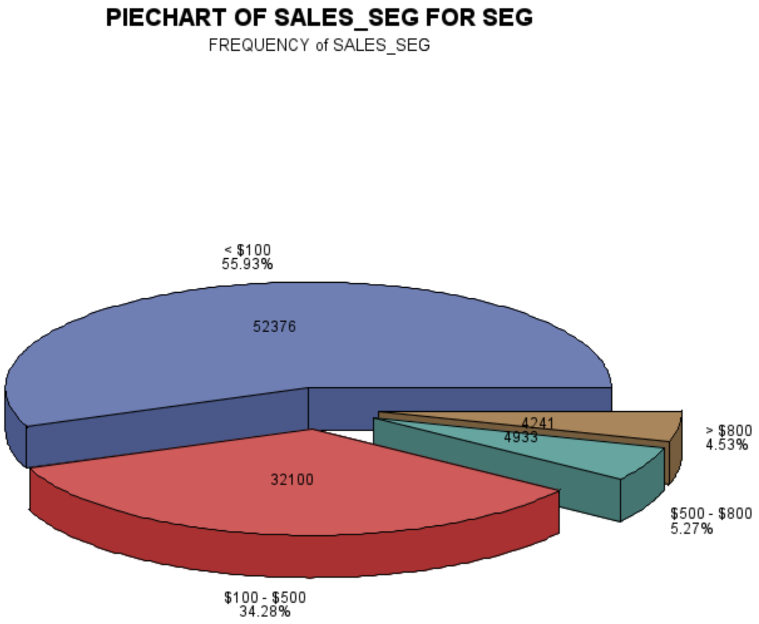
The FREQ Procedure

Number of Variable Levels

Variable	Levels	Missing Levels	Nonmissing Levels
SALES_SEG	5	1	4

SALES_SEG	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing	8605	8.42	8605	8.42
< $100	52376	51.22	60981	59.64
$100 - $500	32100	31.39	93081	91.03
$500 - $800	4933	4.82	98014	95.85
> $800	4241	4.15	102255	100.00



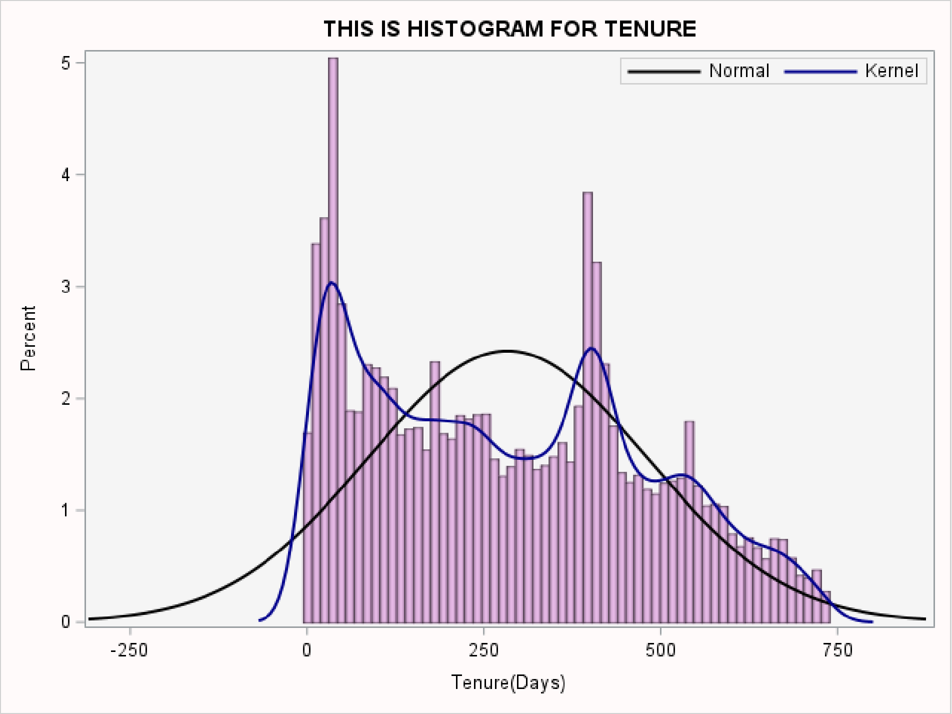


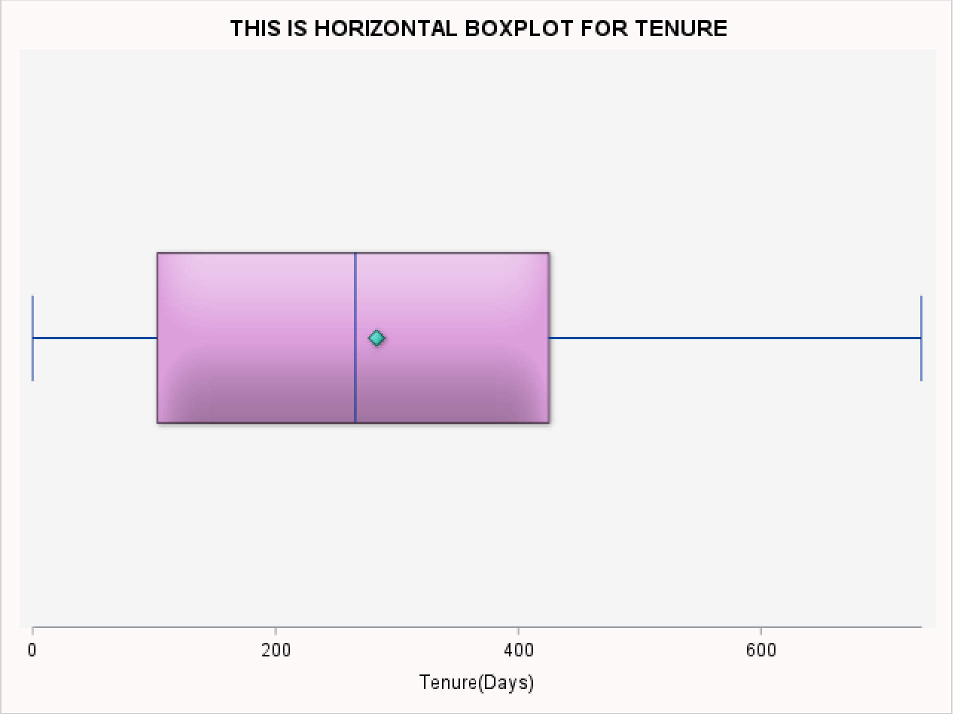
TENURE

The MEANS Procedure

Analysis Variable : TENURE Tenure(Days)

N	N Miss	Mean	Median	Mode	Minimum	Maximum	Std Dev	Variance	Range	Quartile Range
102255	0	283.38	266.00	30.00	0.00	732.00	197.39	38963.98	732.00	324.00





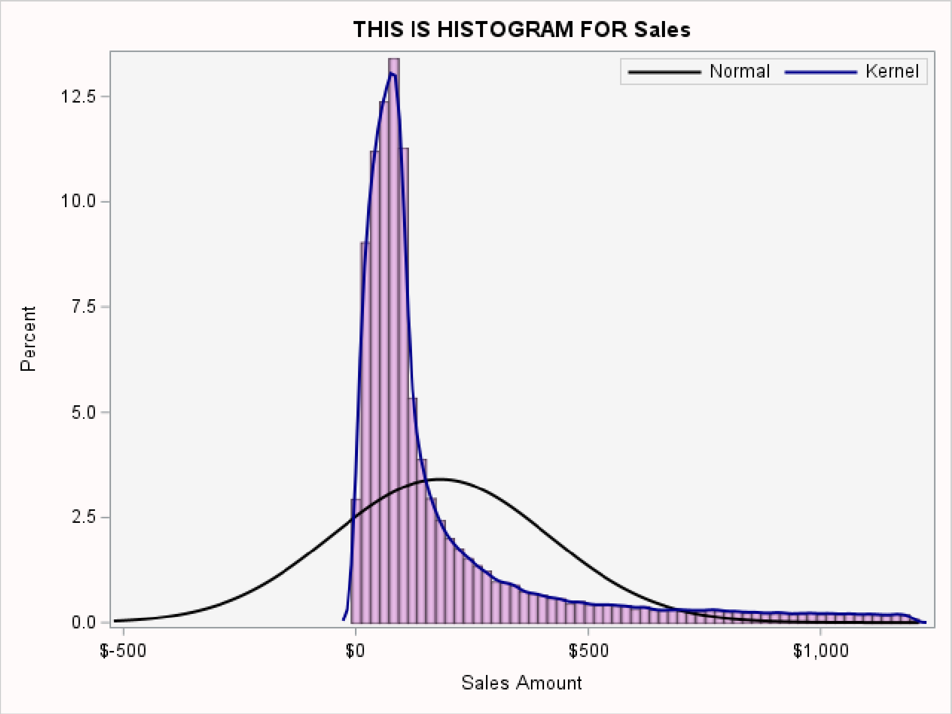
*The base date for calculation of tenure is: 21Jan2001

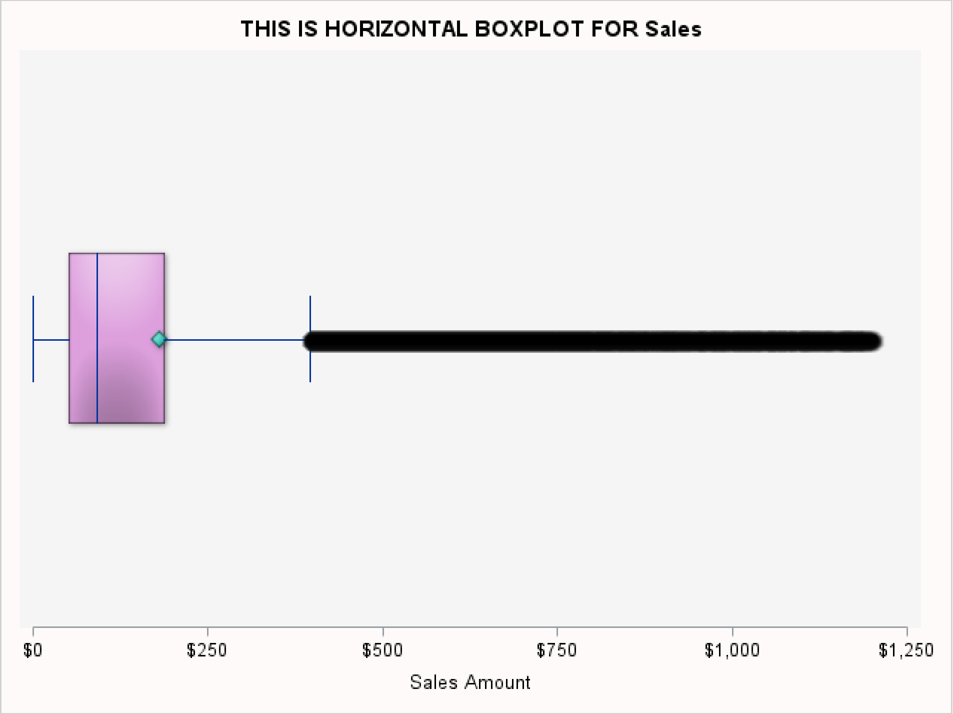
Sales

The MEANS Procedure

Analysis Variable : Sales Sales Amount

N	N Miss	Mean	Median	Mode	Minimum	Maximum	Std Dev	Variance	Range	Quartile Range
93650	8605	181.25	91.00	92.00	0.00	1200.00	233.97	54742.45	1200.00	138.00



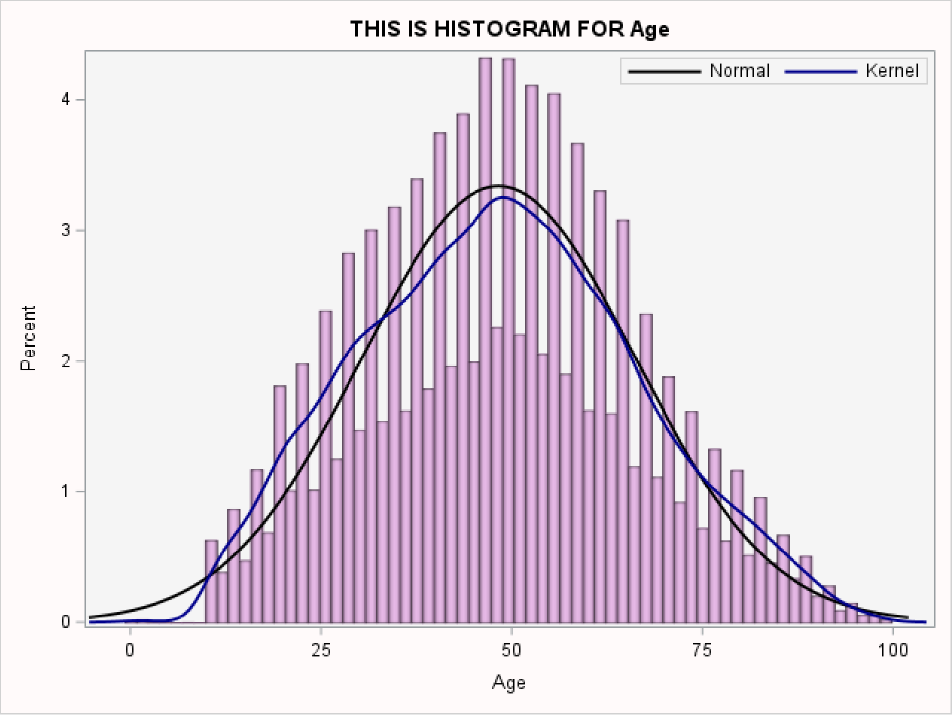


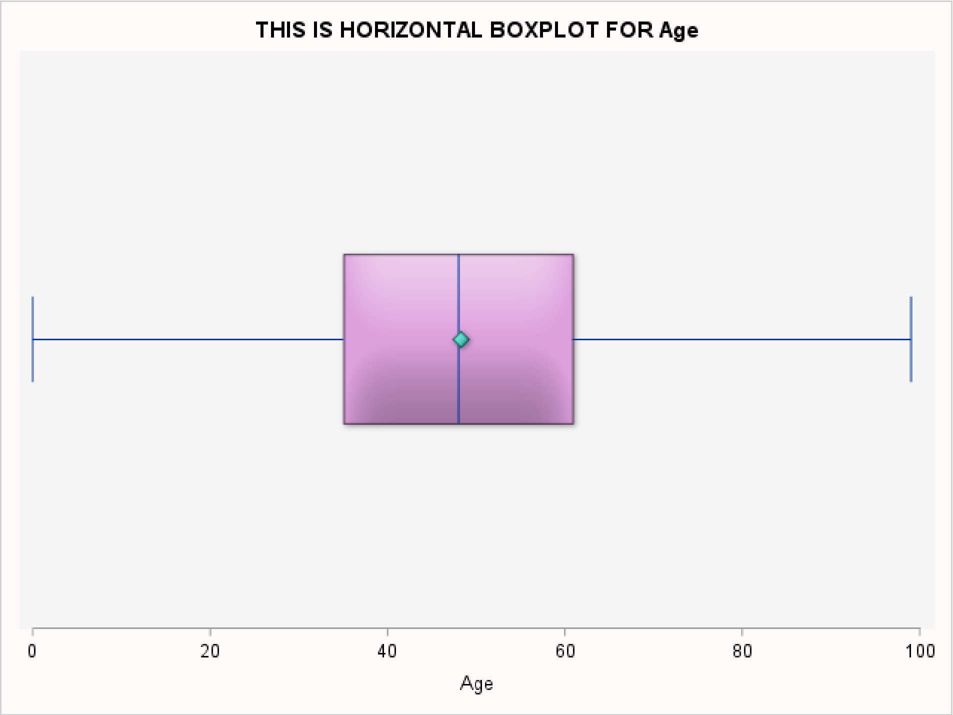
THIS IS HISTOGRAM FOR Age

The MEANS Procedure

Analysis Variable : Age Age

N	N Miss	Mean	Median	Mode	Minimum	Maximum	Std Dev	Variance	Range	Quartile Range
93024	9231	48.28	48.00	48.00	0.00	99.00	17.91	320.86	99.00	26.00





DROPPING OBSERVATIONS

We can see age as low as 0, since the goal of this analysis is to investigate customers' distribution and behaviours,I'll drop any observations with the age of 18, the usually legal age.It was decided that since this is a behavioural study to trimm observations below the legal age and missing provinces and missing sales.Keeping even then 75% of absevations.

PERCENTAGE OF PRESERVED DATA

Obs	id	Total_obs_BEFORE_dropping	Total_obs_AFTER_dropping	PERC
1	1	102255	76877	75%

UNIVARIATE ANALYSIS - After Trimming observations

UNIVARIATE ANALYSIS OF DeactReason FOR SEG

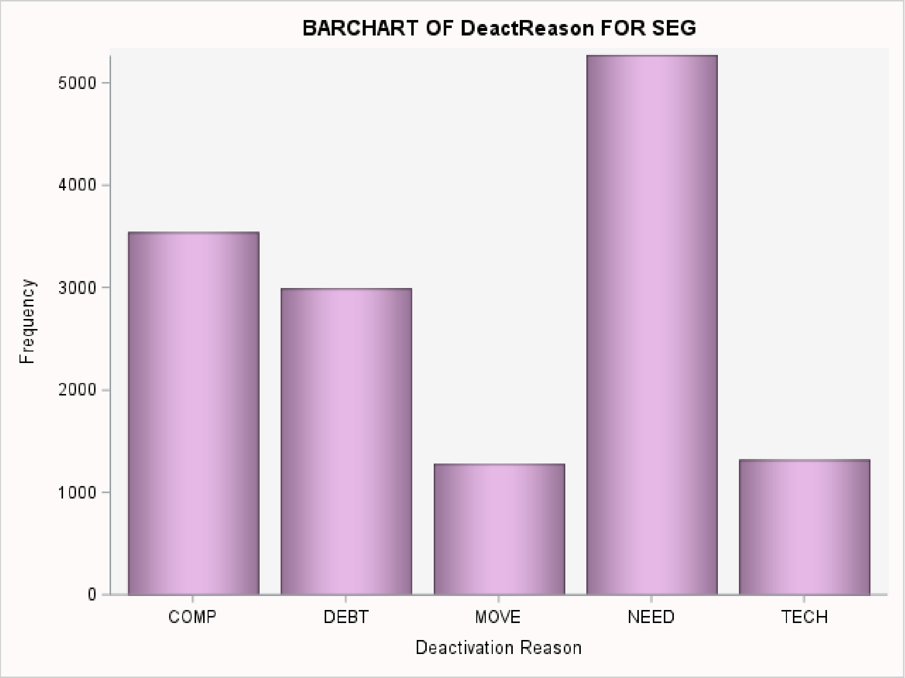
The FREQ Procedure

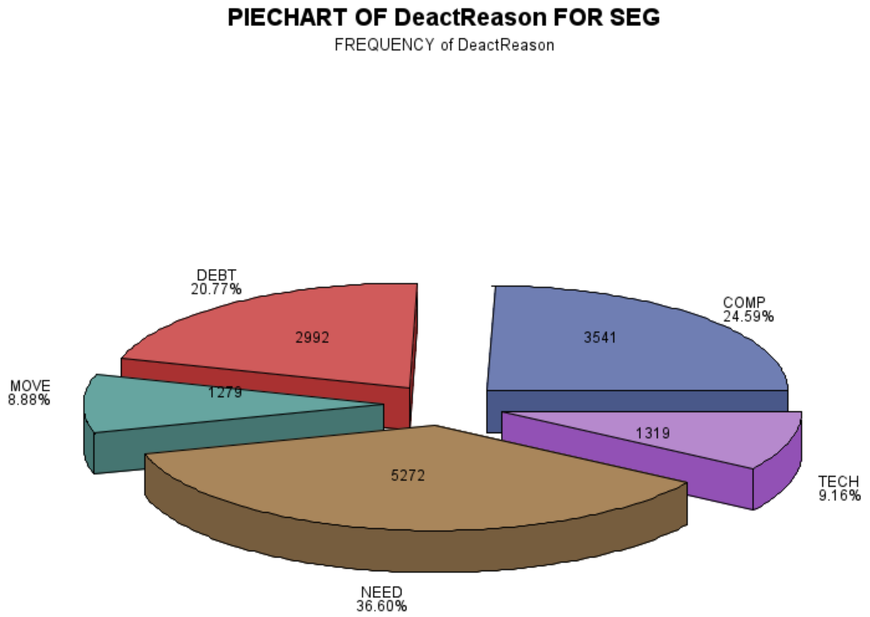
Number of Variable Levels

Variable	Label	Levels	Missing Levels	Nonmissing Levels
DeactReason	Deactivation Reason	6	1	5

Deactivation Reason

DeactReason	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	62474	81.26	62474	81.26
COMP	3541	4.61	66015	85.87
DEBT	2992	3.89	69007	89.76
MOVE	1279	1.66	70286	91.43
NEED	5272	6.86	75558	98.28
TECH	1319	1.72	76877	100.00





UNIVARIATE ANALYSIS OF GoodCredit FOR SEG

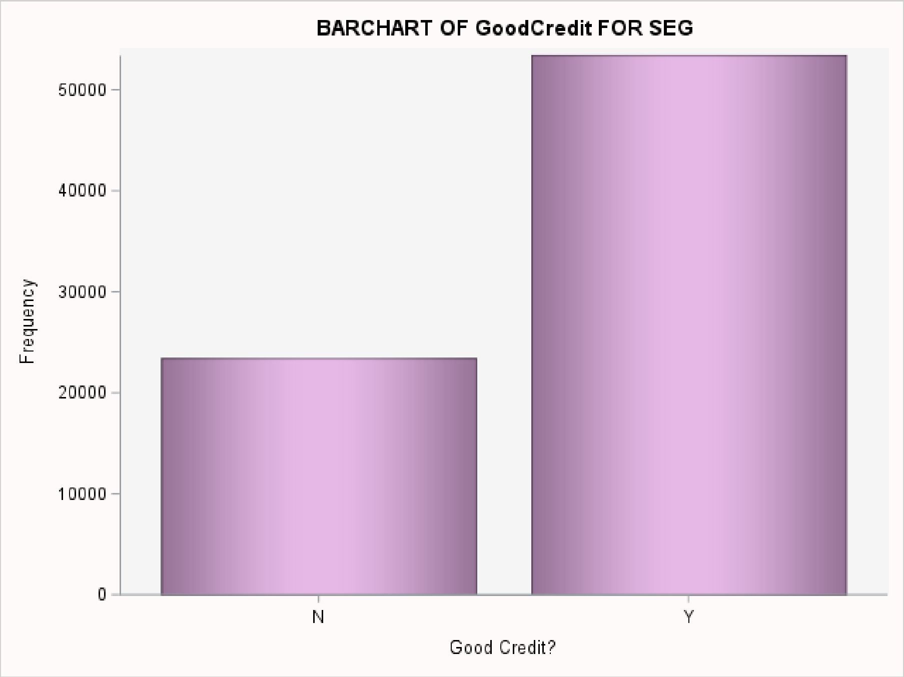
The FREQ Procedure

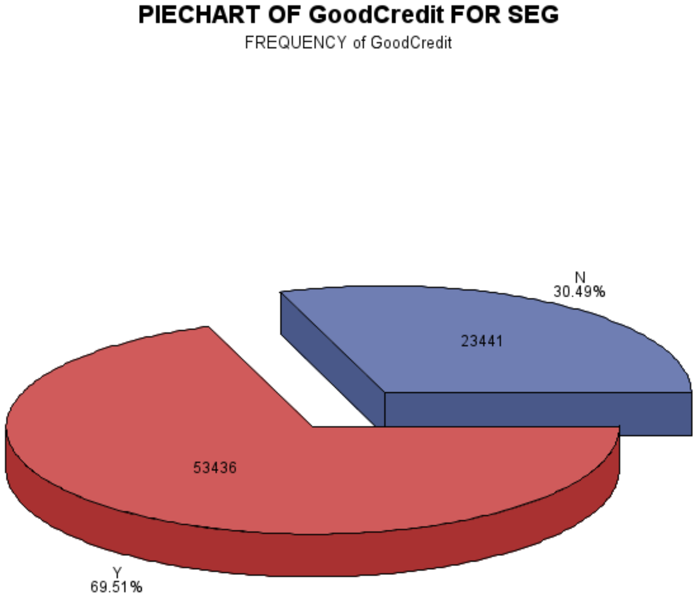
Number of Variable Levels

Variable	Label	Levels
GoodCredit	Good Credit?	2

Good Credit?

GoodCredit	Frequency	Percent	Cumulative Frequency	Cumulative Percent
N	23441	30.49	23441	30.49
Y	53436	69.51	76877	100.00





UNIVARIATE ANALYSIS OF RatePlan FOR SEG

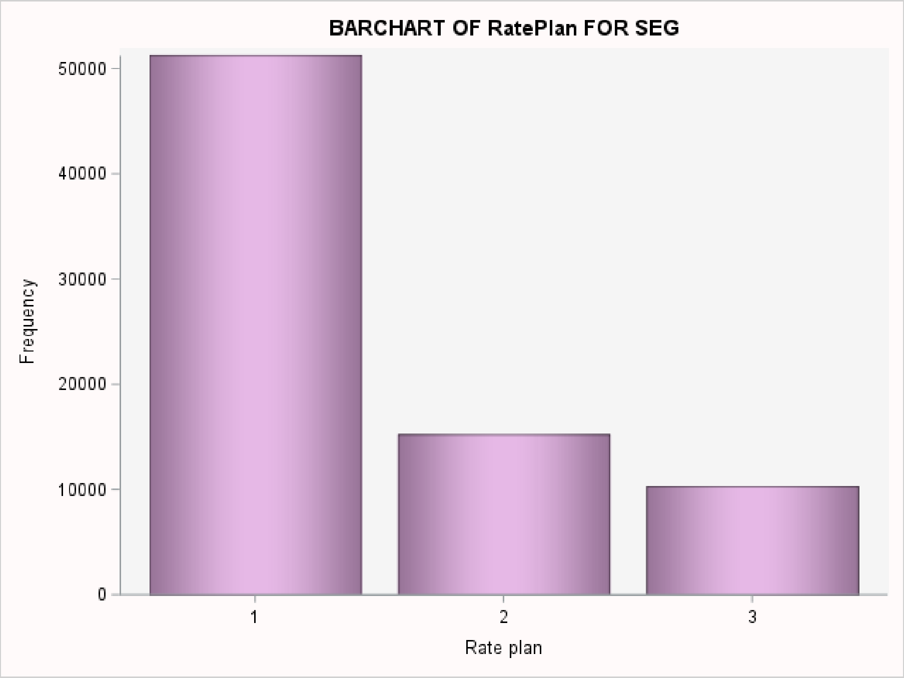
The FREQ Procedure

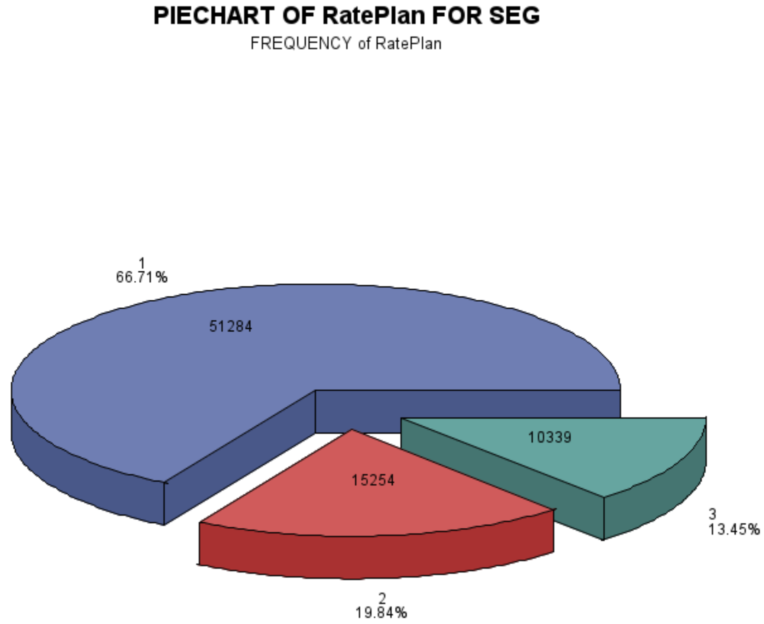
Number of Variable Levels

Variable	Label	Levels
RatePlan	Rate plan	3

Rate plan

RatePlan	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	51284	66.71	51284	66.71
2	15254	19.84	66538	86.55
3	10339	13.45	76877	100.00





UNIVARIATE ANALYSIS OF DealerType FOR SEG

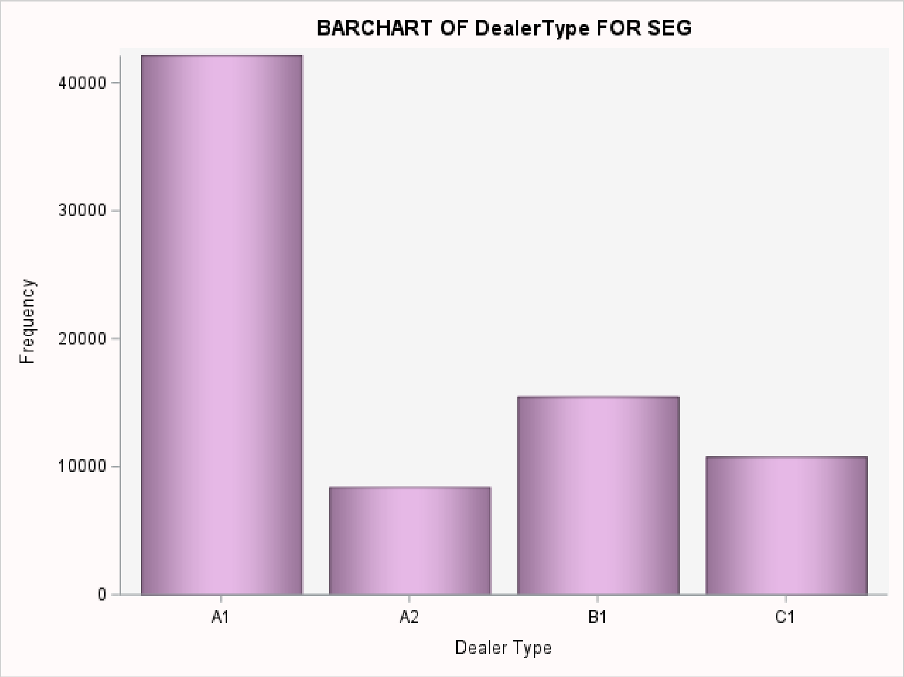
The FREQ Procedure

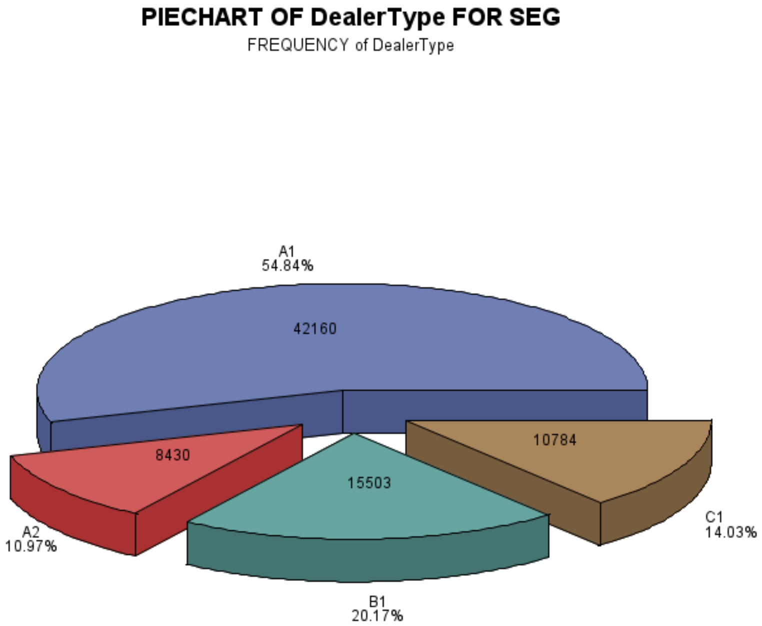
Number of Variable Levels

Variable	Label	Levels
DealerType	Dealer Type	4

Dealer Type

DealerType	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A1	42160	54.84	42160	54.84
A2	8430	10.97	50590	65.81
B1	15503	20.17	66093	85.97
C1	10784	14.03	76877	100.00





UNIVARIATE ANALYSIS OF Province FOR SEG

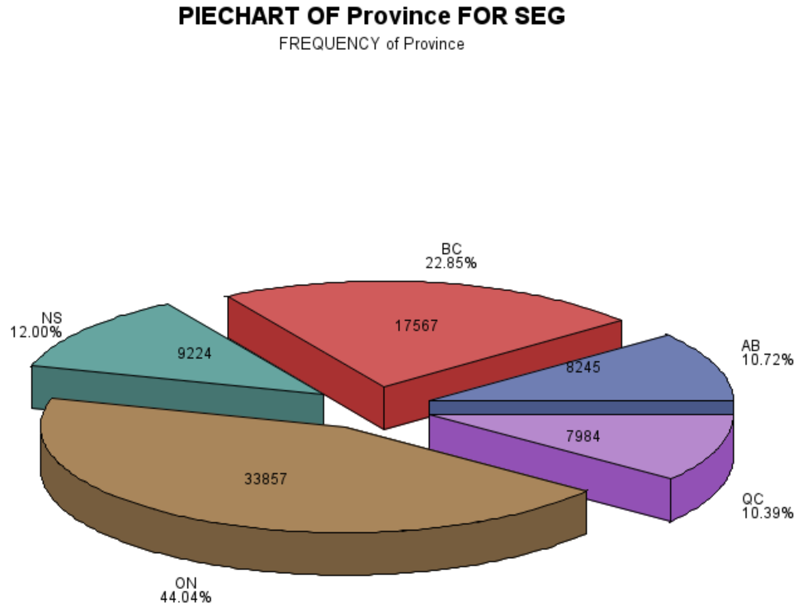
The FREQ Procedure

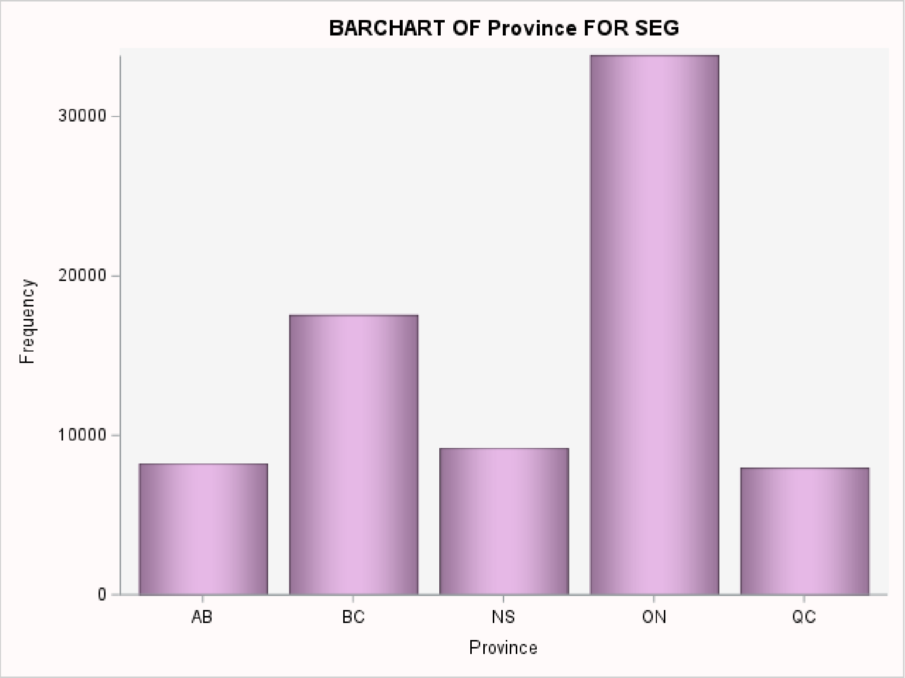
Number of Variable Levels

Variable	Label	Levels
Province	Province	5

Province

Province	Frequency	Percent	Cumulative Frequency	Cumulative Percent
AB	8245	10.72	8245	10.72
BC	17567	22.85	25812	33.58
NS	9224	12.00	35036	45.57
ON	33857	44.04	68893	89.61
QC	7984	10.39	76877	100.00



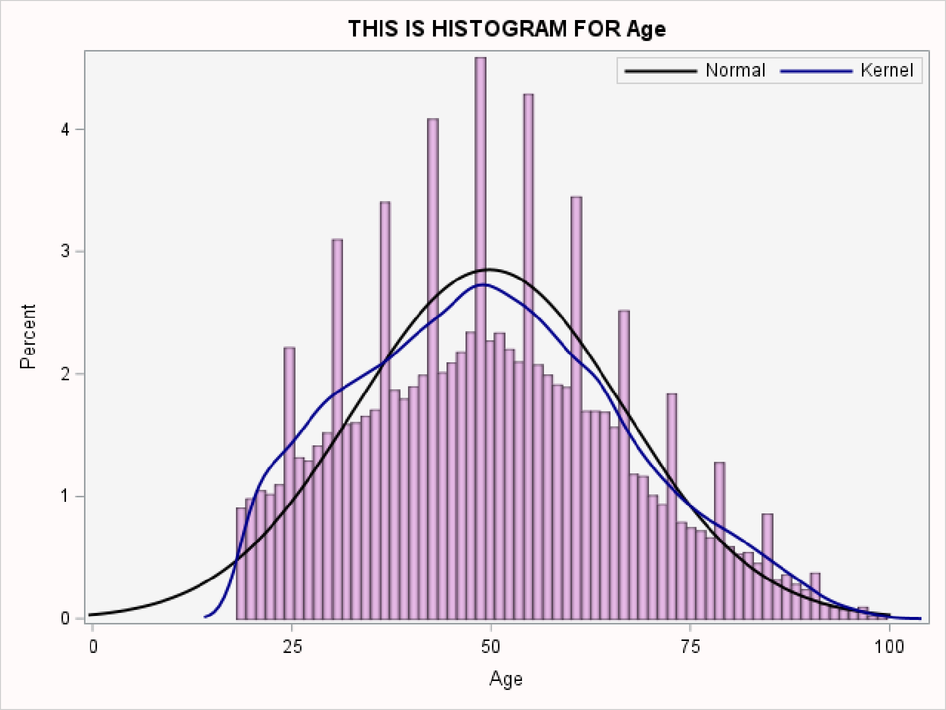


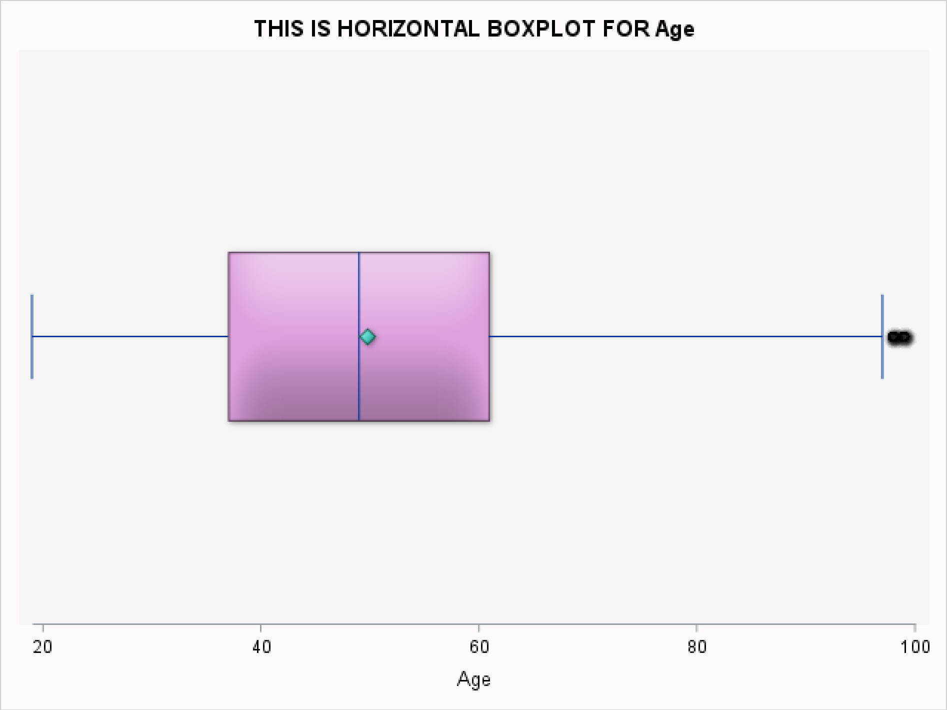
THIS IS HISTOGRAM FOR Age

The MEANS Procedure

Analysis Variable : Age Age

N	N Miss	Mean	Median	Mode	Minimum	Maximum	Std Dev	Variance	Range	Quartile Range
76877	0	49.78	49.00	48.00	19.00	99.00	16.78	281.73	80.00	24.00



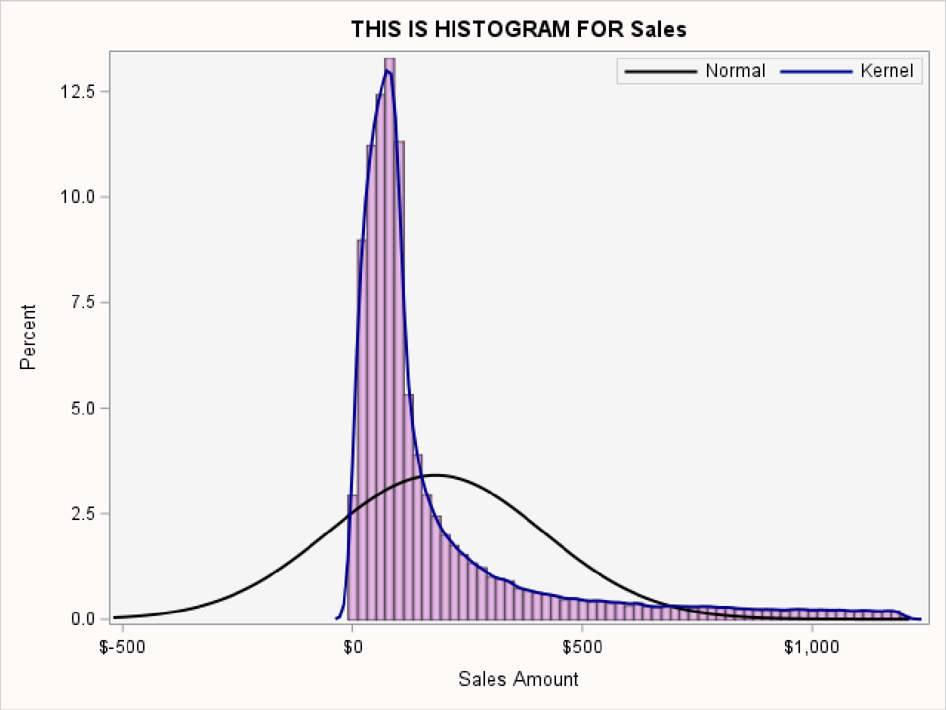


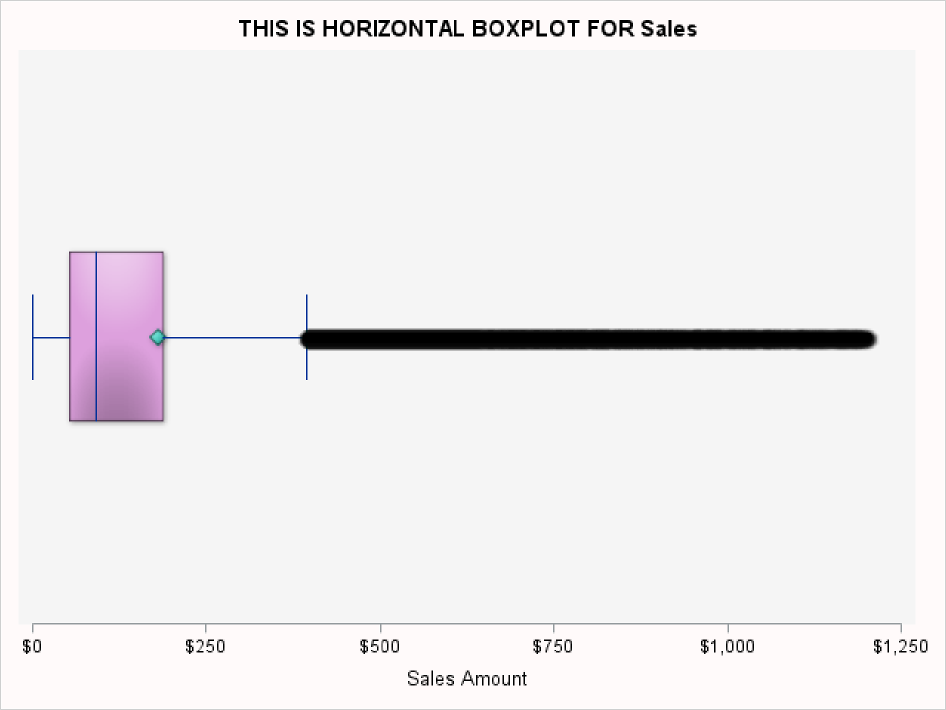
THIS IS HISTOGRAM FOR Sales

The MEANS Procedure

Analysis Variable : Sales Sales Amount

N	N Miss	Mean	Median	Mode	Minimum	Maximum	Std Dev	Variance	Range	Quartile Range
76877	0	181.32	91.00	94.00	0.00	1200.00	233.87	54693.34	1200.00	137.00



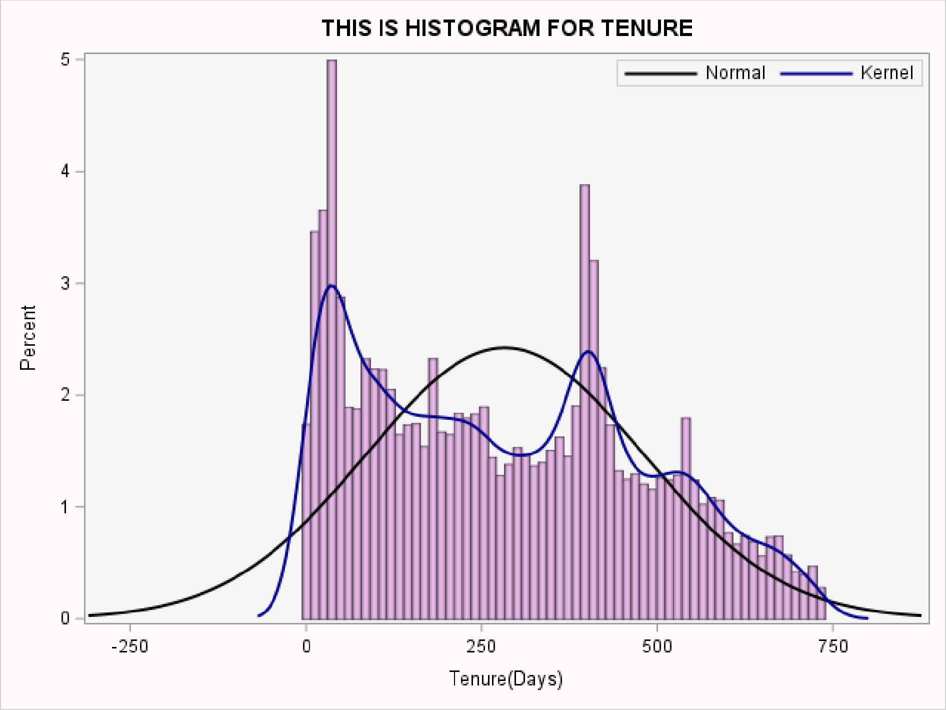


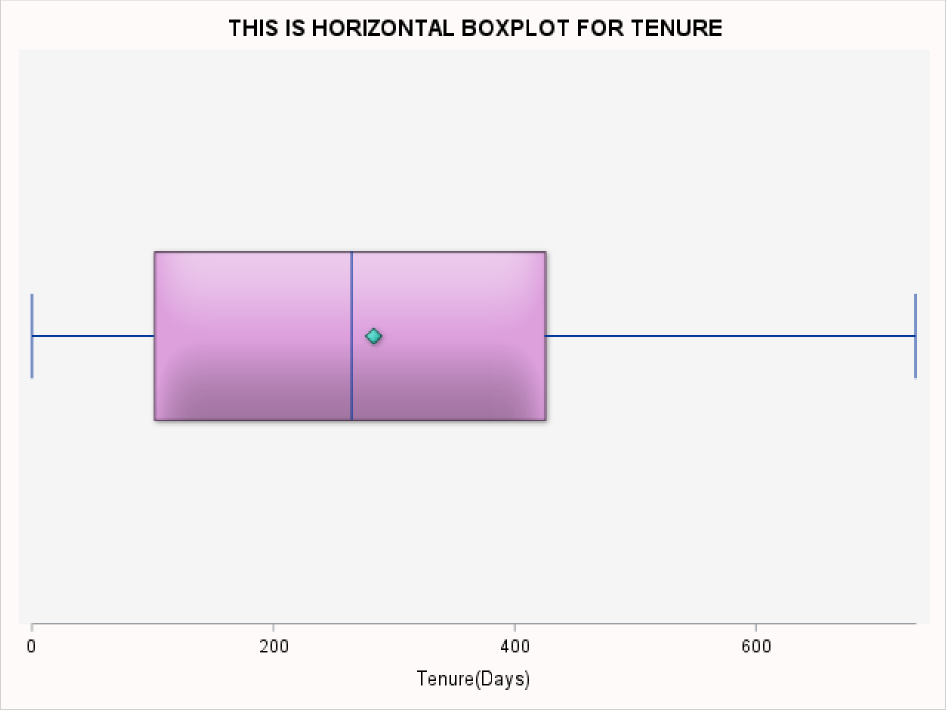
THIS IS HISTOGRAM FOR TENURE

The MEANS Procedure

Analysis Variable : TENURE Tenure(Days)

N	N Miss	Mean	Median	Mode	Minimum	Maximum	Std Dev	Variance	Range	Quartile Range
76877	0	282.98	265.00	30.00	0.00	732.00	197.64	39060.11	732.00	325.00





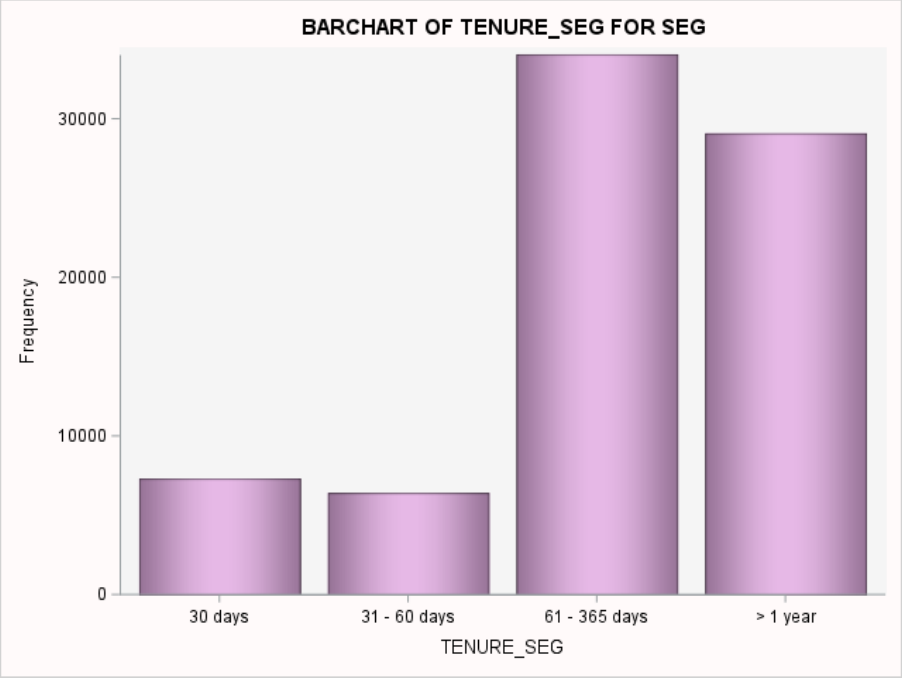
UNIVARIATE ANALYSIS OF TENURE_SEG FOR SEG

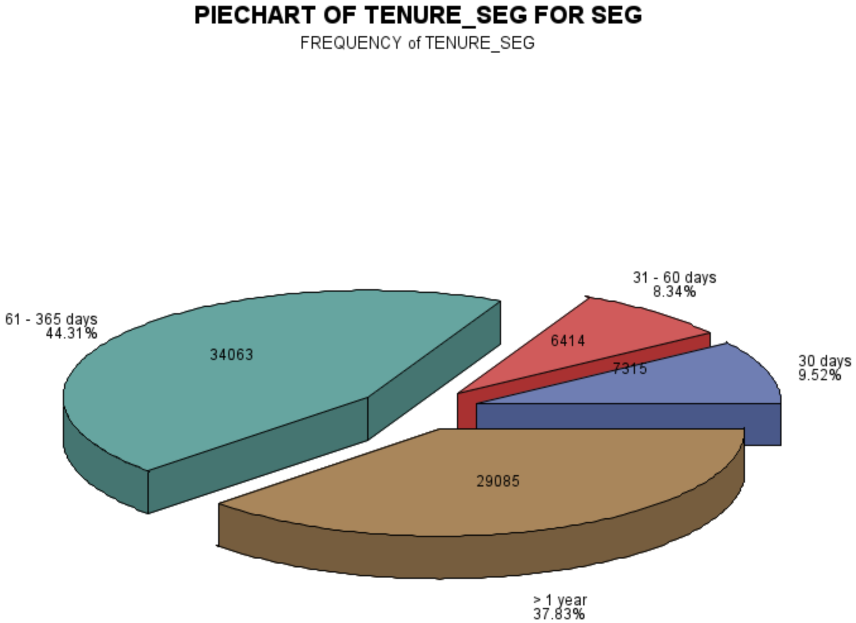
The FREQ Procedure

Number of Variable Levels

Variable	Levels
TENURE_SEG	4

TENURE_SEG	Frequency	Percent	Cumulative Frequency	Cumulative Percent
30 days	7315	9.52	7315	9.52
31 - 60 days	6414	8.34	13729	17.86
61 - 365 days	34063	44.31	47792	62.17
> 1 year	29085	37.83	76877	100.00





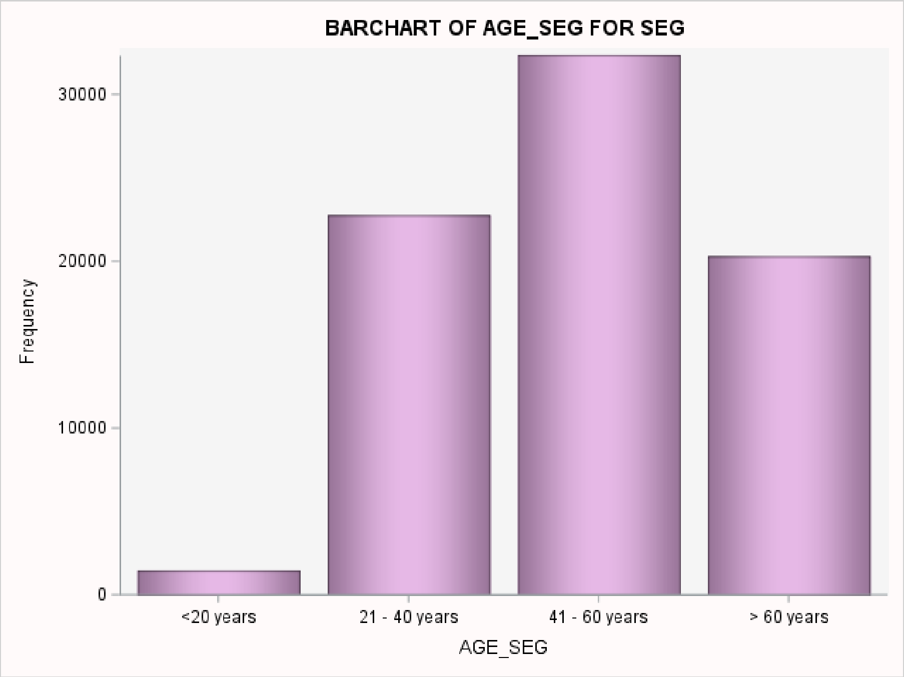
UNIVARIATE ANALYSIS OF AGE_SEG FOR SEG

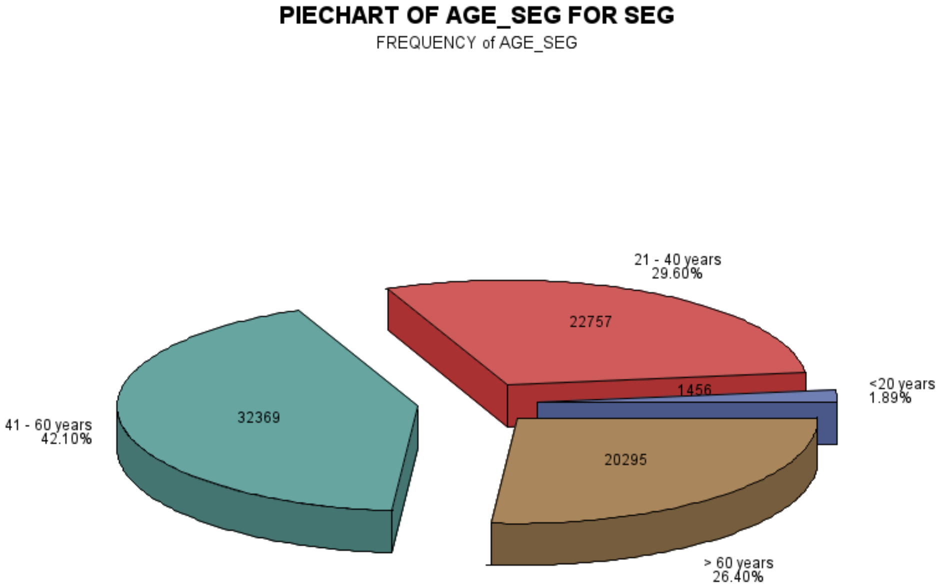
The FREQ Procedure

Number of Variable Levels

Variable	Levels
AGE_SEG	4

AGE_SEG	Frequency	Percent	Cumulative Frequency	Cumulative Percent
<20 years	1456	1.89	1456	1.89
21 - 40 years	22757	29.60	24213	31.50
41 - 60 years	32369	42.10	56582	73.60
> 60 years	20295	26.40	76877	100.00





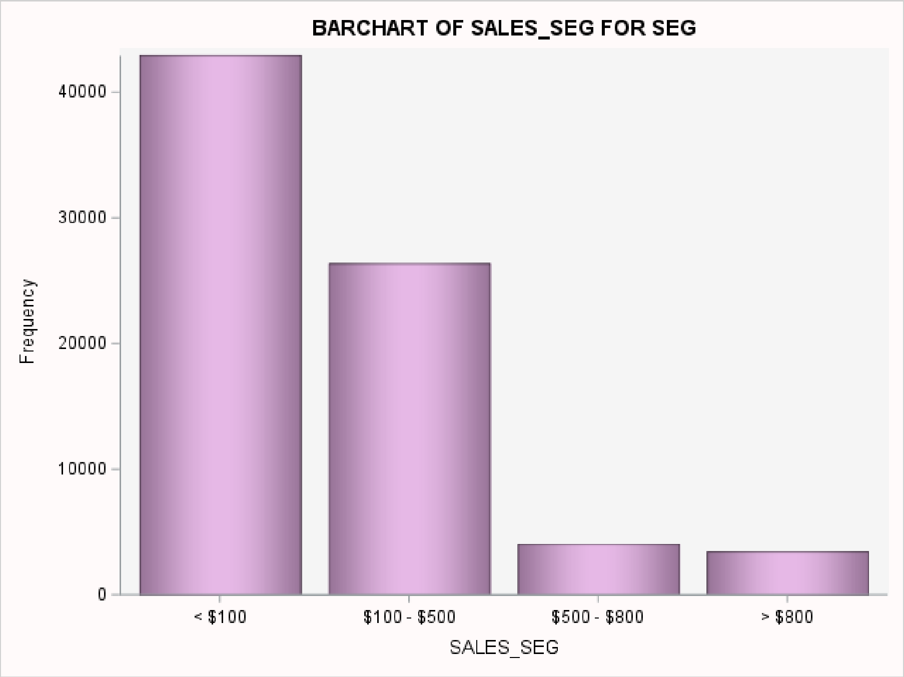
UNIVARIATE ANALYSIS OF SALES_SEG FOR SEG

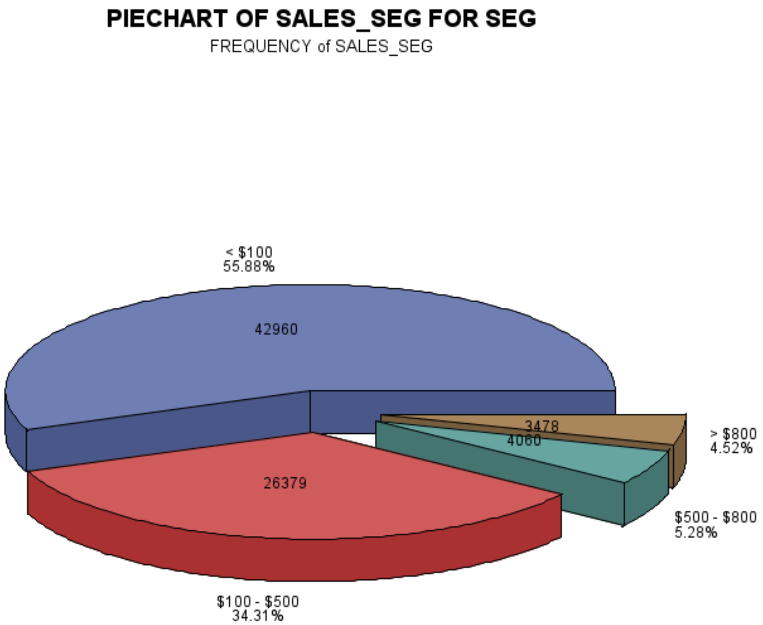
The FREQ Procedure

Number of Variable Levels

Variable	Levels
SALES_SEG	4

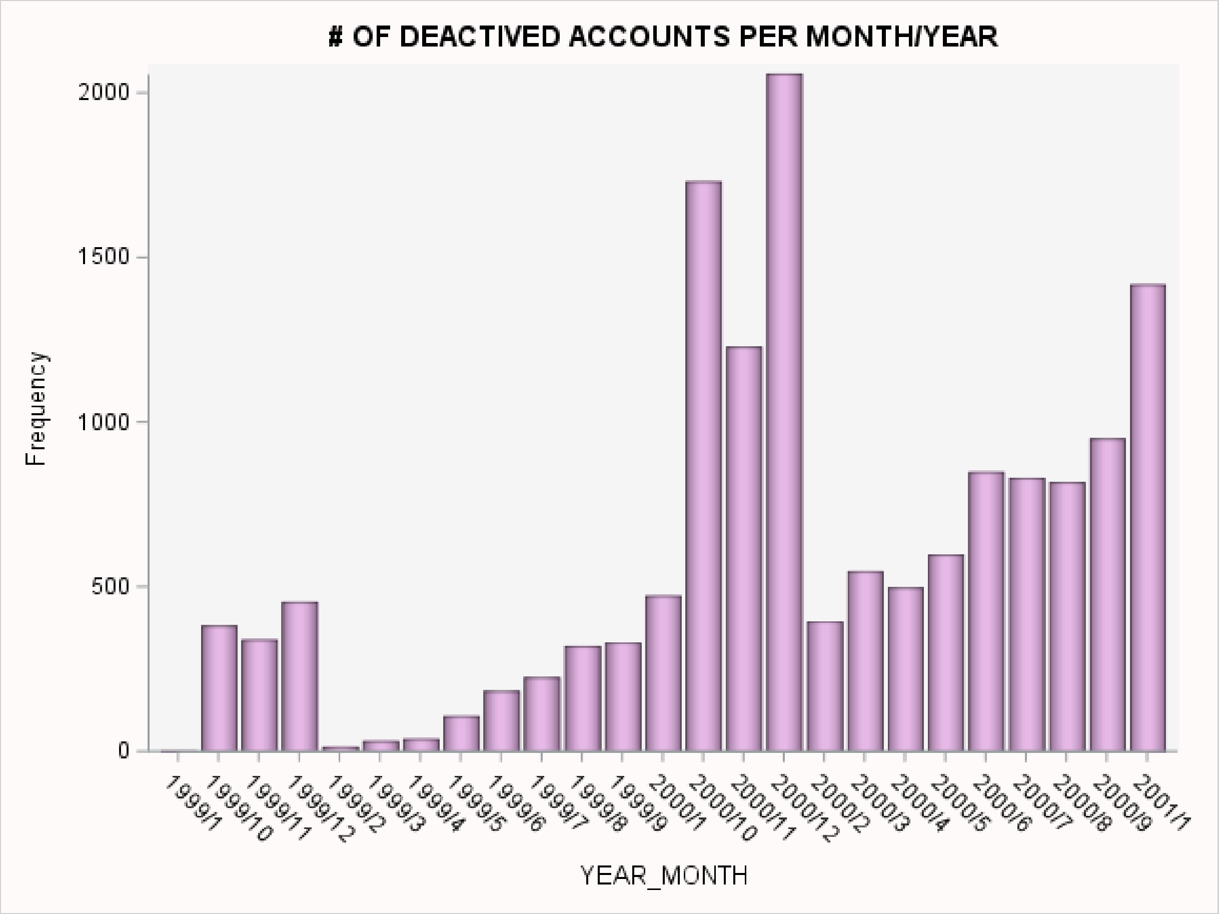
SALES_SEG	Frequency	Percent	Cumulative Frequency	Cumulative Percent
< $100	42960	55.88	42960	55.88
$100 - $500	26379	34.31	69339	90.19
$500 - $800	4060	5.28	73399	95.48
> $800	3478	4.52	76877	100.00





# OF DEACTIVED ACCOUNTS PER MONTH/YEAR

Obs	YEAR_MONTH	N_ACCOUNTS
1	1999/1	2
2	1999/10	382
3	1999/11	340
4	1999/12	454
5	1999/2	14
6	1999/3	32
7	1999/4	38
8	1999/5	107
9	1999/6	184
10	1999/7	225
11	1999/8	320
12	1999/9	331
13	2000/1	473
14	2000/10	1731
15	2000/11	1229
16	2000/12	2059
17	2000/2	395
18	2000/3	547
19	2000/4	498
20	2000/5	598
21	2000/6	849
22	2000/7	830
23	2000/8	818
24	2000/9	950
25	2001/1	1418



BIVARIATE ANALYSIS
CATEGORICAL VS CATEGORICAL

Null hypothesis:

1. N, the total frequency, should be reasonably large (greater than 50)

2. The sample observations should be independent. No individual item should be included twice or more in the sample"

3. No expected frequencies should be small. Preferably each expected frequency should be larger than 10 but in any case not less than 5.

If condition of chi-square are satisfied and p-value is less than significant level (5%), reject null hypothesis:

- There is a relationship between them at 5% significant level.

BIVARIATE ANALYSIS OF ACTIVE AND PROVINCE FOR SEG

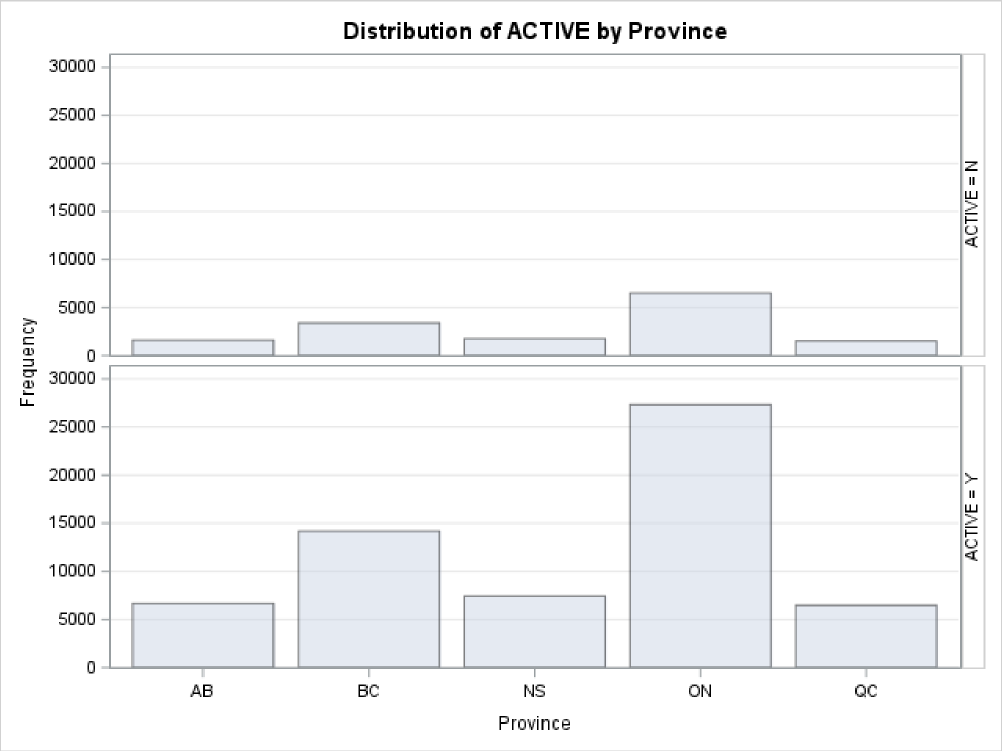
Null hypothesis: ACTIVE is independent of the PROVINCE

The FREQ Procedure

Table of ACTIVE by Province

Frequency
Percent
Row Pct
Col Pct

ACTIVE	Province(Province)					
	AB	BC	NS	ON	QC	Total
N	1611	3400	1781	6514	1518	14824
	2.10	4.42	2.32	8.47	1.97	19.28
	10.87	22.94	12.01	43.94	10.24	
	19.54	19.35	19.31	19.24	19.01	
Y	6634	14167	7443	27343	6466	62053
	8.63	18.43	9.68	35.57	8.41	80.72
	10.69	22.83	11.99	44.06	10.42	
	80.46	80.65	80.69	80.76	80.99	
Total	8245	17567	9224	33857	7984	76877
	10.72	22.85	12.00	44.04	10.39	100.00



Statistics for Table of ACTIVE by Province

Statistic	DF	Value	Prob
Chi-Square	4	0.8235	0.9353
Likelihood Ratio Chi-Square	4	0.8236	0.9353
Mantel-Haenszel Chi-Square	1	0.7190	0.3965
Phi Coefficient		0.0033	
Contingency Coefficient		0.0033	
Cramer's V		0.0033	

Sample Size = 76877

If condition of chi-square are satisfied and p-value is less than significant level (5%),reject null hypothesis:There is a relationship between them at 5% significant level.

We can see that the assumptions for chi-square test are met, with p-value of 0.9353, we fail to reject the null hypothesis and can't say that there's a relationship between the features.

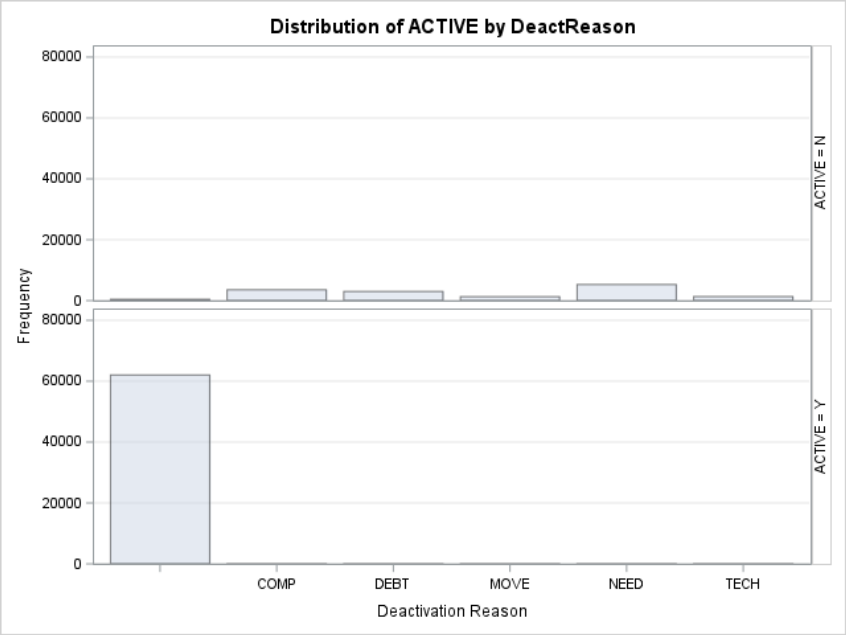
BIVARIATE ANALYSIS OF ACTIVE AND DEACTREASON FOR SEG1
Null hypothesis: ACTIVE is independent of the DEACTREASON

The FREQ Procedure

Table of ACTIVE by DeactReason

Frequency
Percent
Row Pct
Col Pct

ACTIVE	DeactReason(Deactivation Reason)						
		COMP	DEBT	MOVE	NEED	TECH	Total
N	421 0.55 2.84 0.67	3541 4.61 23.89 100.00	2992 3.89 20.18 100.00	1279 1.66 8.63 100.00	5272 6.86 35.56 100.00	1319 1.72 8.90 100.00	14824 19.28
Y	62053 80.72 100.00 99.33	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	62053 80.72
Total	62474 81.26	3541 4.61	2992 3.89	1279 1.66	5272 6.86	1319 1.72	76877 100.00



Statistics for Table of ACTIVE by DeactReason

Statistic	DF	Value	Prob
Chi-Square	5	74190.3553	<.0001
Likelihood Ratio Chi-Square	5	70336.0750	<.0001
Mantel-Haenszel Chi-Square	1	57598.9003	<.0001
Phi Coefficient		0.9824	
Contingency Coefficient		0.7008	
Cramer's V		0.9824	

Sample Size = 76877

If condition of chi-square are satisfied and p-value is less than significant level (5%),reject null hypothesis:There is a relationship between them at 5% significant level.

We can see that the assumptions for chi-square test are met, with p-value of <0.0001, we fail to reject the null hypothesis and can't say that there's a relationship between the features.

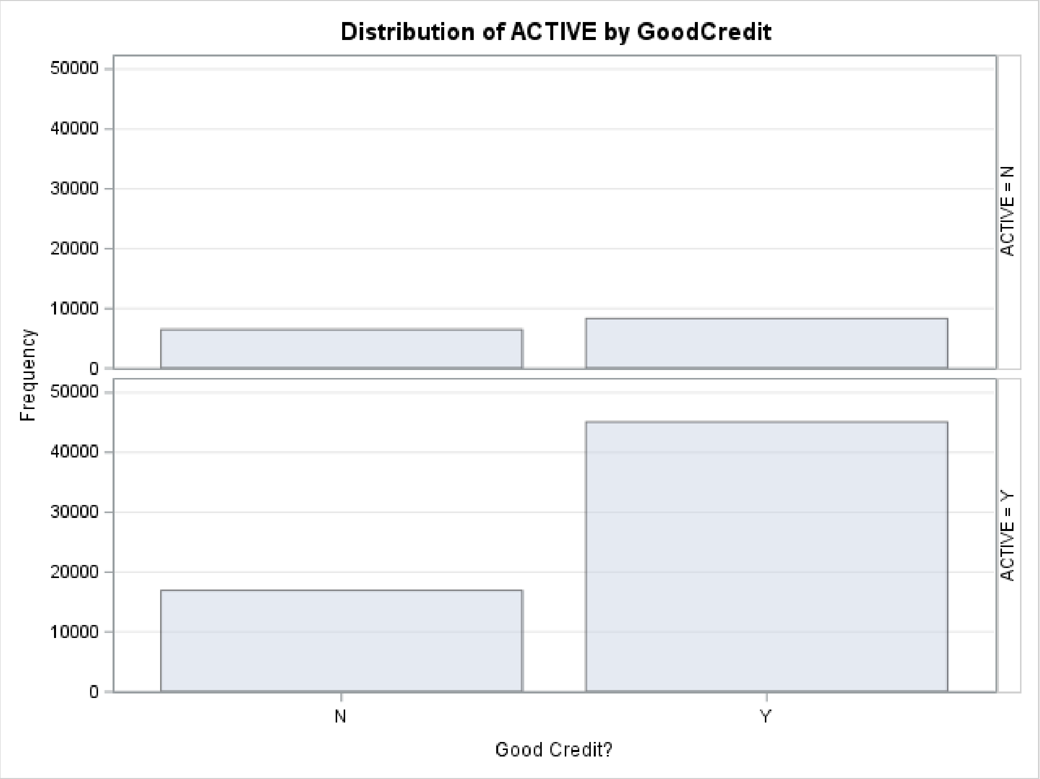
BIVARIATE ANALYSIS OF ACTIVE AND GOODCREDIT FOR SEG
Null hypothesis: ACTIVE is independent of the GOODCREDIT

The FREQ Procedure

Table of ACTIVE by GoodCredit

Frequency
Percent
Row Pct
Col Pct

ACTIVE	GoodCredit(Good Credit?)		
	N	Y	Total
N	6488 8.44 43.77 27.68	8336 10.84 56.23 15.60	14824 19.28
Y	16953 22.05 27.32 72.32	45100 58.67 72.68 84.40	62053 80.72
Total	23441 30.49	53436 69.51	76877 100.00



Statistics for Table of ACTIVE by GoodCredit

Statistic	DF	Value	Prob
Chi-Square	1	1527.1107	<.0001
Likelihood Ratio Chi-Square	1	1457.0037	<.0001
Continuity Adj. Chi-Square	1	1526.3348	<.0001
Mantel-Haenszel Chi-Square	1	1527.0909	<.0001
Phi Coefficient		0.1409	
Contingency Coefficient		0.1396	
Cramer's V		0.1409	

Fisher's Exact Test

Cell (1,1) Frequency (F)	6488
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	<.0001
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Sample Size = 76877

If condition of chi-square are satisfied and p-value is less than significant level (5%),reject null hypothesis:There is a relationship between them at 5% significant level.

We can see that the assumptions for chi-square test are met, with p-value of <0.0001, we can reject the null hypothesis and say that there's a relationship between the features.

BIVARIATE ANALYSIS OF ACTIVE AND RATEPLAN FOR SEG

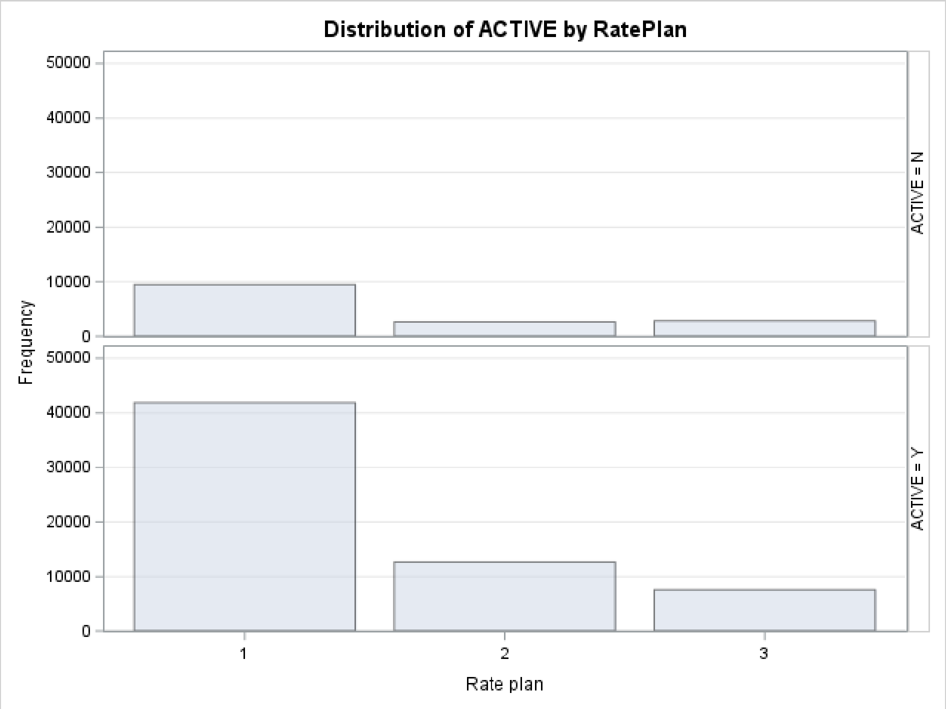
Null hypothesis: ACTIVE is independent of the RATEPLAN

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of ACTIVE by RatePlan

ACTIVE	RatePlan(Rate plan)			
	1	2	3	Total
N	9437	2597	2790	14824
	12.28	3.38	3.63	19.28
	63.66	17.52	18.82	
	18.40	17.03	26.99	
Y	41847	12657	7549	62053
	54.43	16.46	9.82	80.72
	67.44	20.40	12.17	
	81.60	82.97	73.01	
Total	51284	15254	10339	76877
	66.71	19.84	13.45	100.00



Statistics for Table of ACTIVE by RatePlan

Statistic	DF	Value	Prob
Chi-Square	2	469.6421	<.0001
Likelihood Ratio Chi-Square	2	438.5681	<.0001
Mantel-Haenszel Chi-Square	1	251.4548	<.0001
Phi Coefficient		0.0782	
Contingency Coefficient		0.0779	
Cramer's V		0.0782	

Sample Size = 76877

If condition of chi-square are satisfied and p-value is less than significant level (5%),reject null hypothesis:There is a relationship between them at 5% significant level.

We can see that the assumptions for chi-square test are met, with p-value of <0.0001, we can reject the null hypothesis and say that there's a relationship between the features.

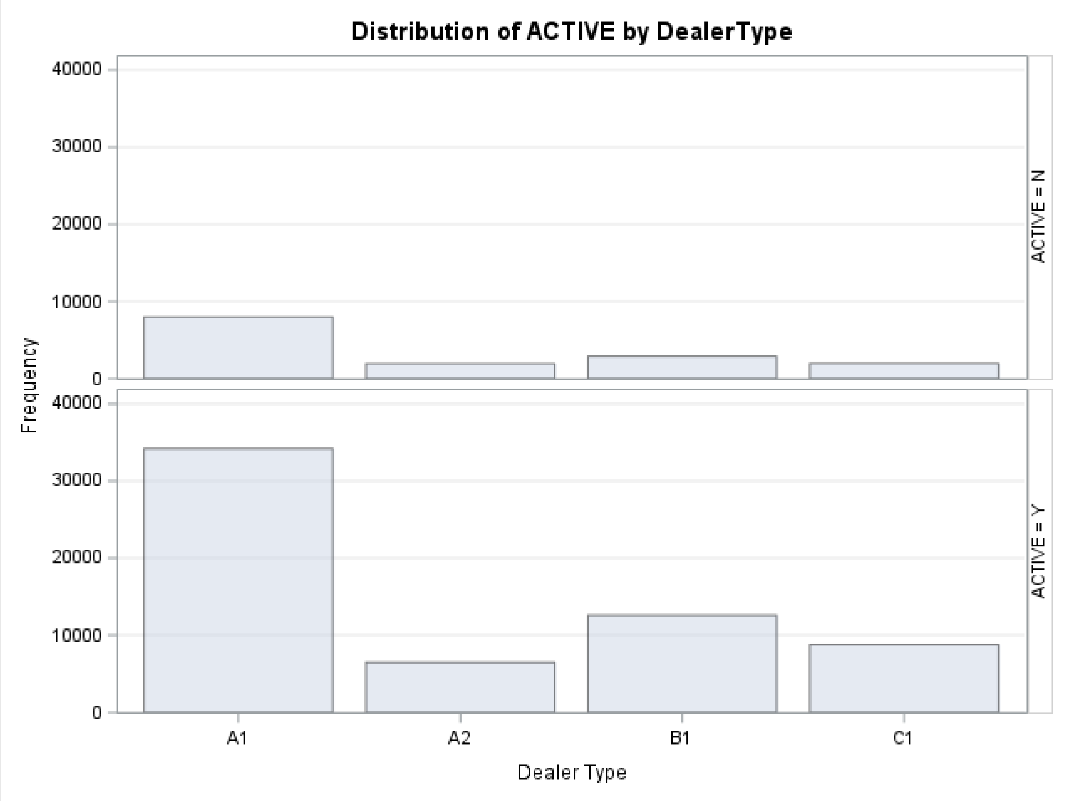
BIVARIATE ANALYSIS OF ACTIVE AND DEALERTYPE FOR SEG
Null hypothesis: ACTIVE is independent of the DEALERTYPE

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of ACTIVE by DealerType

ACTIVE	DealerType(Dealer Type)				
	A1	A2	B1	C1	Total
N	7991	1946	2909	1978	14824
	10.39	2.53	3.78	2.57	19.28
	53.91	13.13	19.62	13.34	
	18.95	23.08	18.76	18.34	
Y	34169	6484	12594	8806	62053
	44.45	8.43	16.38	11.45	80.72
	55.06	10.45	20.30	14.19	
	81.05	76.92	81.24	81.66	
Total	42160	8430	15503	10784	76877
	54.84	10.97	20.17	14.03	100.00



Statistics for Table of ACTIVE by DealerType

Statistic	DF	Value	Prob
Chi-Square	3	90.0092	<.0001
Likelihood Ratio Chi-Square	3	86.5759	<.0001
Mantel-Haenszel Chi-Square	1	1.3391	0.2472
Phi Coefficient		0.0342	
Contingency Coefficient		0.0342	
Cramer's V		0.0342	

Sample Size = 76877

If condition of chi-square are satisfied and p-value is less than significant level (5%),reject null hypothesis:There is a relationship between them at 5% significant level.

We can see that the assumptions for chi-square test are met, with p-value of <0.0001, we can reject the null hypothesis and say that there's a relationship between the features.

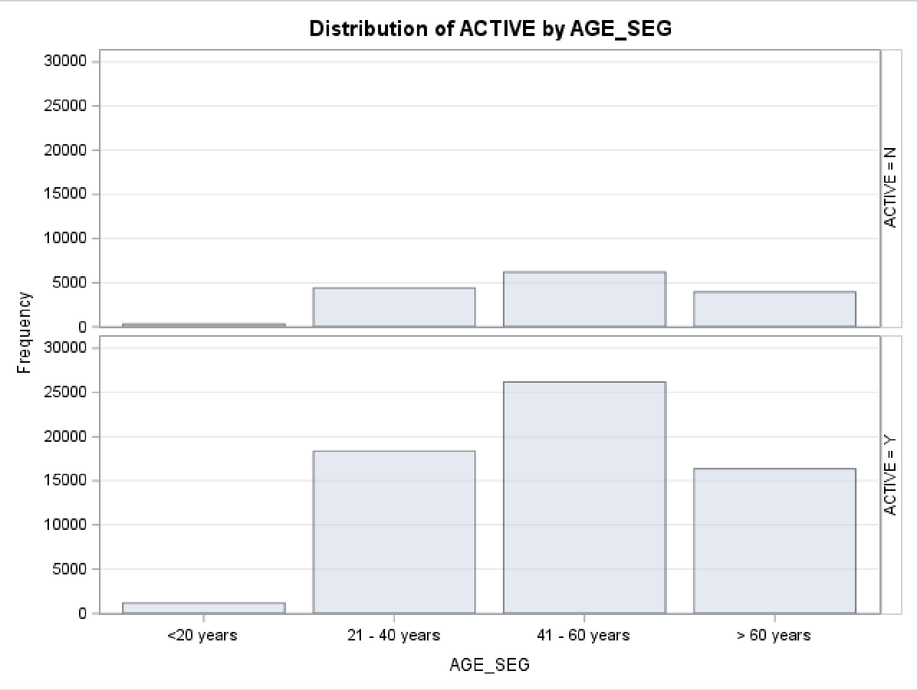
BIVARIATE ANALYSIS OF ACTIVE AND AGE_SEG FOR SEG
Null hypothesis: ACTIVE is independent of the AGE_SEG

The FREQ Procedure

Table of ACTIVE by AGE_SEG

Frequency
Percent
Row Pct
Col Pct

ACTIVE	AGE_SEG				
	<20 years	21 - 40 years	41 - 60 years	> 60 years	Total
N	297	4392	6198	3937	14824
	0.39	5.71	8.06	5.12	19.28
	2.00	29.63	41.81	26.56	
	20.40	19.30	19.15	19.40	
Y	1159	18365	26171	16358	62053
	1.51	23.89	34.04	21.28	80.72
	1.87	29.60	42.18	26.36	
	79.60	80.70	80.85	80.60	
Total	1456	22757	32369	20295	76877
	1.89	29.60	42.10	26.40	100.00



Statistics for Table of ACTIVE by AGE_SEG

Statistic	DF	Value	Prob
Chi-Square	3	1.7221	0.6320
Likelihood Ratio Chi-Square	3	1.7059	0.6356
Mantel-Haenszel Chi-Square	1	0.0005	0.9821
Phi Coefficient		0.0047	
Contingency Coefficient		0.0047	
Cramer's V		0.0047	

Sample Size = 76877

If condition of chi-square are satisfied and p-value is less than significant level (5%),reject null hypothesis:There is a relationship between them at 5% significant level.

We can see that the assumptions for chi-square test are met, with p-value of 0.6320, we fail to reject the null hypothesis and can't say that there's a relationship between the features.

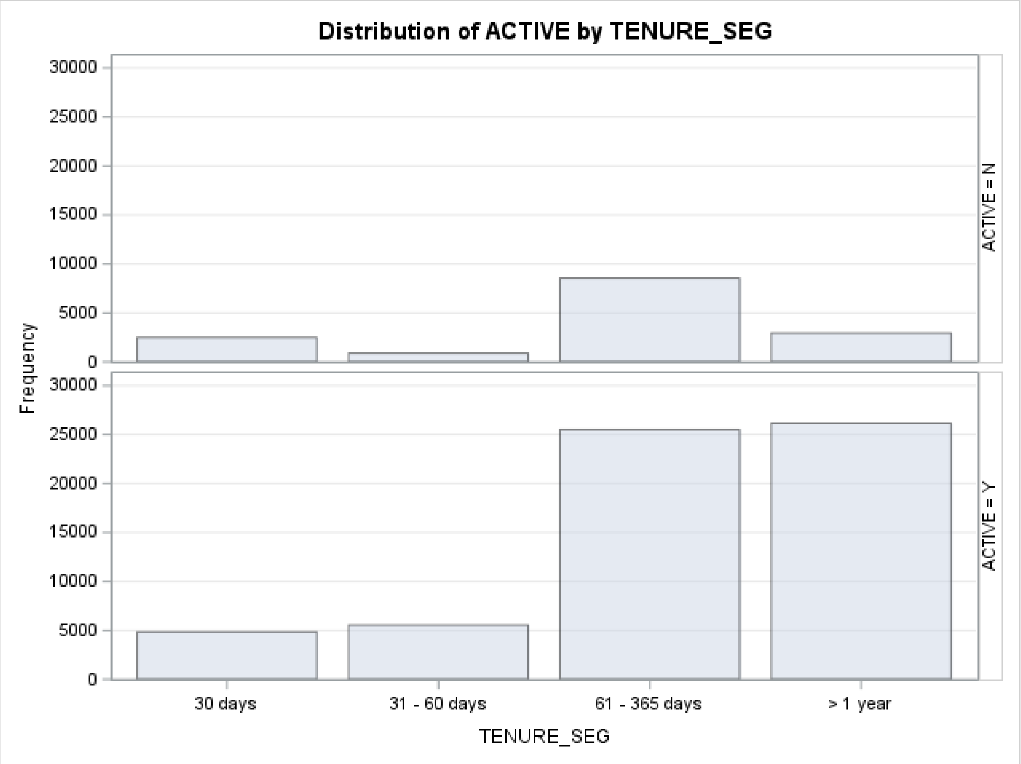
BIVARIATE ANALYSIS OF ACTIVE AND TENURE_SEG FOR SEG
Null hypothesis: ACTIVE is independent of the TENURE_SEG

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of ACTIVE by TENURE_SEG

ACTIVE	TENURE_SEG				
	30 days	31 - 60 days	61 - 365 days	> 1 year	Total
N	2476	878	8546	2924	14824
	3.22	1.14	11.12	3.80	19.28
	16.70	5.92	57.65	19.72	
	33.85	13.69	25.09	10.05	
Y	4839	5536	25517	26161	62053
	6.29	7.20	33.19	34.03	80.72
	7.80	8.92	41.12	42.16	
	66.15	86.31	74.91	89.95	
Total	7315	6414	34063	29085	76877
	9.52	8.34	44.31	37.83	100.00



Statistics for Table of ACTIVE by TENURE_SEG

Statistic	DF	Value	Prob
Chi-Square	3	3455.5756	<.0001
Likelihood Ratio Chi-Square	3	3545.6239	<.0001
Mantel-Haenszel Chi-Square	1	2636.4859	<.0001
Phi Coefficient		0.2120	
Contingency Coefficient		0.2074	
Cramer's V		0.2120	

Sample Size = 76877

If condition of chi-square are satisfied and p-value is less than significant level (5%),reject null hypothesis:There is a relationship between them at 5% significant level.

We can see that the assumptions for chi-square test are met, with p-value of <0.0001, we can reject the null hypothesis and say that there's a relationship between the features.

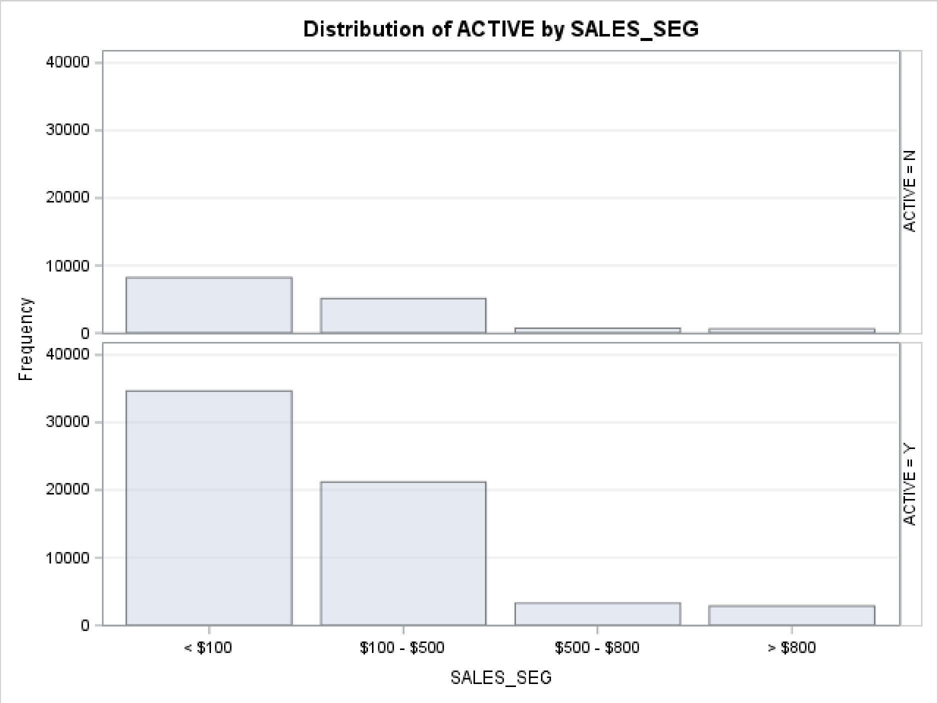
BIVARIATE ANALYSIS OF ACTIVE AND SALES_SEG FOR SEG
Null hypothesis: ACTIVE is independent of the SALES_SEG

The FREQ Procedure

Table of ACTIVE by SALES_SEG

Frequency
Percent
Row Pct
Col Pct

ACTIVE	SALES_SEG				
	< $100	$100 - $500	$500 - $800	> $800	Total
N	8249	5146	775	654	14824
	10.73	6.69	1.01	0.85	19.28
	55.65	34.71	5.23	4.41	
	19.20	19.51	19.09	18.80	
Y	34711	21233	3285	2824	62053
	45.15	27.62	4.27	3.67	80.72
	55.94	34.22	5.29	4.55	
	80.80	80.49	80.91	81.20	
Total	42960	26379	4060	3478	76877
	55.88	34.31	5.28	4.52	100.00



Statistics for Table of ACTIVE by SALES_SEG

Statistic	DF	Value	Prob
Chi-Square	3	1.6519	0.6477
Likelihood Ratio Chi-Square	3	1.6531	0.6474
Mantel-Haenszel Chi-Square	1	0.2960	0.5864
Phi Coefficient		0.0046	
Contingency Coefficient		0.0046	
Cramer's V		0.0046	

Sample Size = 76877

If condition of chi-square are satisfied and p-value is less than significant level (5%),reject null hypothesis:There is a relationship between them at 5% significant level.

We can see that the assumptions for chi-square test are met, with p-value of <0.6477, we cannot reject the null hypothesis . Thus, we say that there's not a relationship between the features.

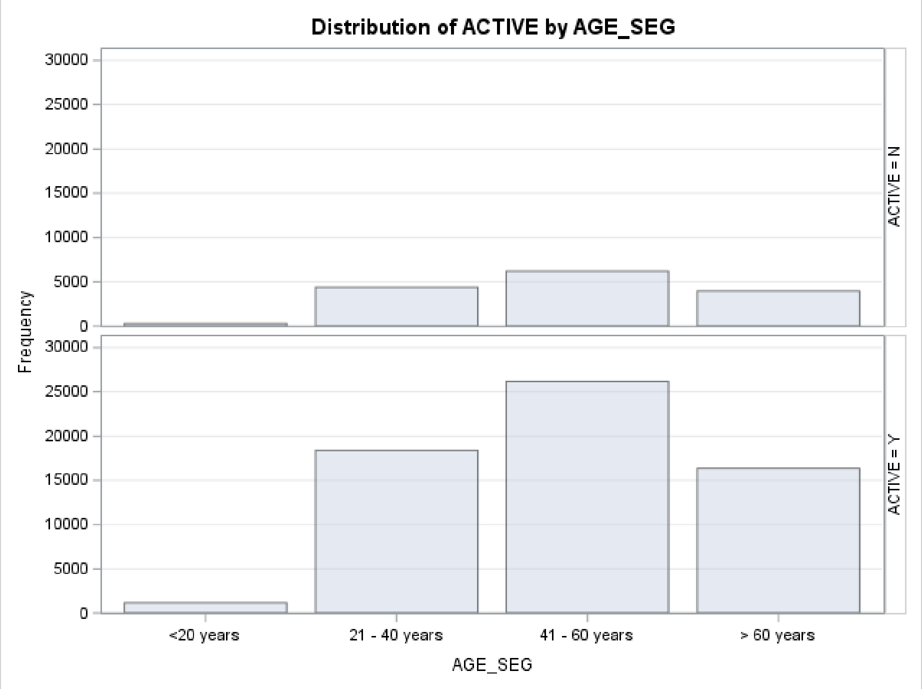
BIVARIATE ANALYSIS OF ACTIVE AND AGE_SEG FOR SEG
Null hypothesis: ACTIVE is independent of the AGE_SEG

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of ACTIVE by AGE_SEG

ACTIVE	AGE_SEG				
	<20 years	21 - 40 years	41 - 60 years	> 60 years	Total
N	297	4392	6198	3937	14824
	0.39	5.71	8.06	5.12	19.28
	2.00	29.63	41.81	26.56	
	20.40	19.30	19.15	19.40	
Y	1159	18365	26171	16358	62053
	1.51	23.89	34.04	21.28	80.72
	1.87	29.60	42.18	26.36	
	79.60	80.70	80.85	80.60	
Total	1456	22757	32369	20295	76877
	1.89	29.60	42.10	26.40	100.00



Statistics for Table of ACTIVE by AGE_SEG

Statistic	DF	Value	Prob
Chi-Square	3	1.7221	0.6320
Likelihood Ratio Chi-Square	3	1.7059	0.6356
Mantel-Haenszel Chi-Square	1	0.0005	0.9821
Phi Coefficient		0.0047	
Contingency Coefficient		0.0047	
Cramer's V		0.0047	

Sample Size = 76877

If condition of chi-square are satisfied and p-value is less than significant level (5%),reject null hypothesis:There is a relationship between them at 5% significant level.

We can see that the assumptions for chi-square test are met, with p-value of 0.6320, we fail to reject the null hypothesis and can't say that there's a relationship between the features.

CATEGORICAL VS CONTINUOUS

Null hypothesis: There’s no difference in means

1.Sample distribution must be normal:

CLT :

If looks normal each group must have more than 30 observations – no need for Shapiro’s test

If moderately skewed, each group must have more than 100 observations – no need for Shapiro’s test

2.Groups are independent of one another.

3.There are no major outliers.

4.A check for unequal variances will help determine which version of an independent samples t-test is most appropriate:

(Levene’s test, null hypothesis: equal variance)

a.If variances are equal, then a pooled t-test is appropriate

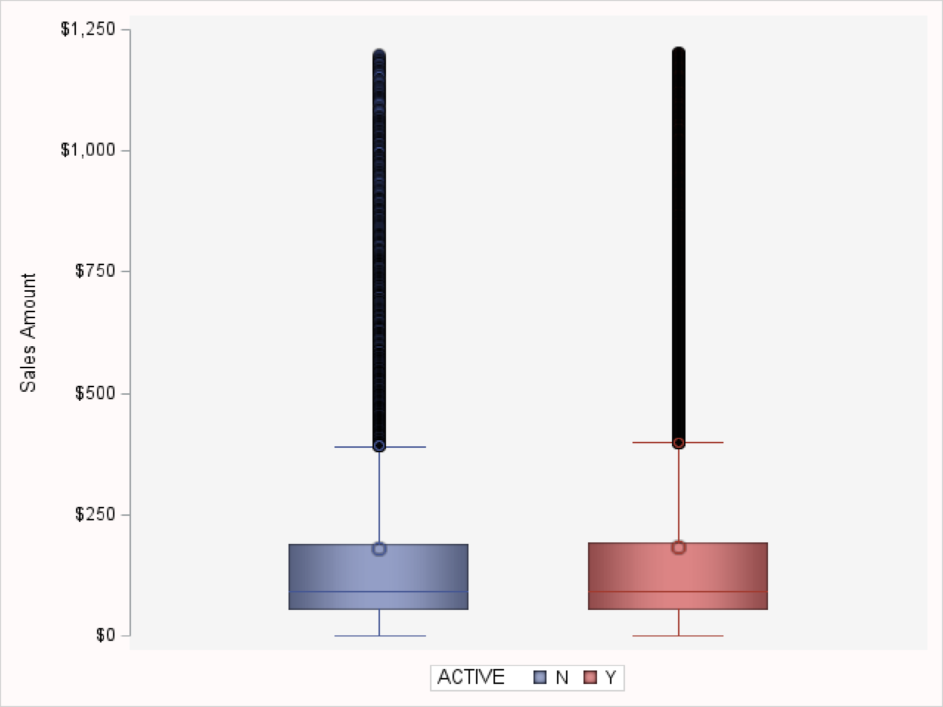
b.If variances are unequal, then a Satterthwaite (also known as Welch’s) t-test is appropriate

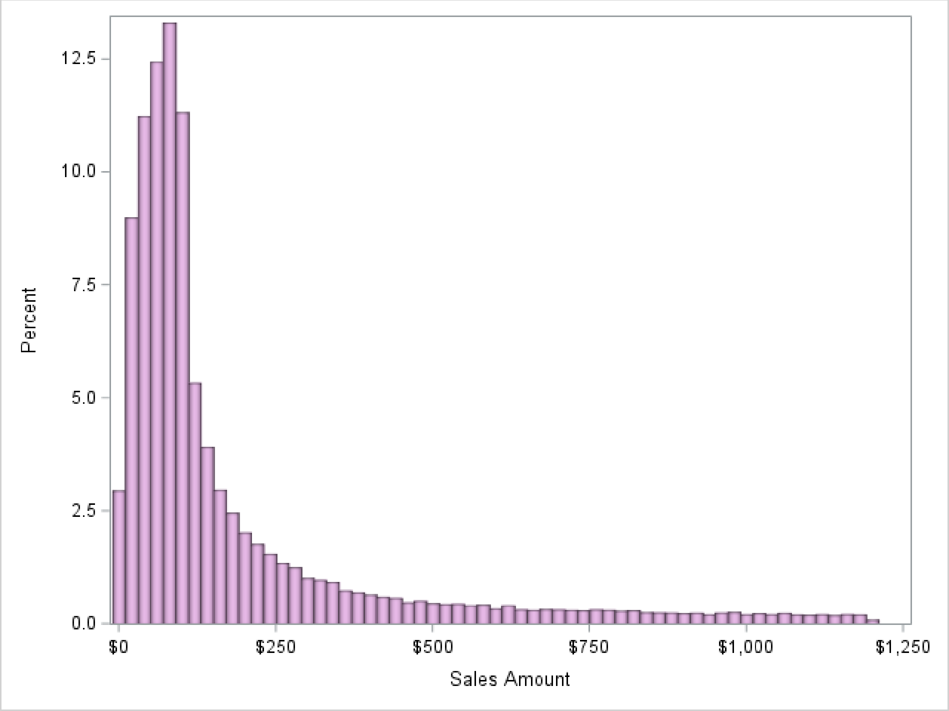
RELATION BETWEEN SALES AND ACTIVE

The MEANS Procedure

Analysis Variable : Sales Sales Amount

ACTIVE	N Obs	N	N Miss	Minimum	Lower Quartile	Median	Mean	Upper Quartile	Maximum	Quartile Range	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean
N	**14824**	14824	0	0.00	53.00	92.00	180.02	188.00	1197.00	135.00	128.65	176.30	183.75
Y	**62053**	62053	0	0.00	52.00	91.00	181.63	191.00	1200.00	139.00	129.06	179.78	183.47





We can see a great number of major outliers, so as the data is, it's not possible to use t-test for sales and active features

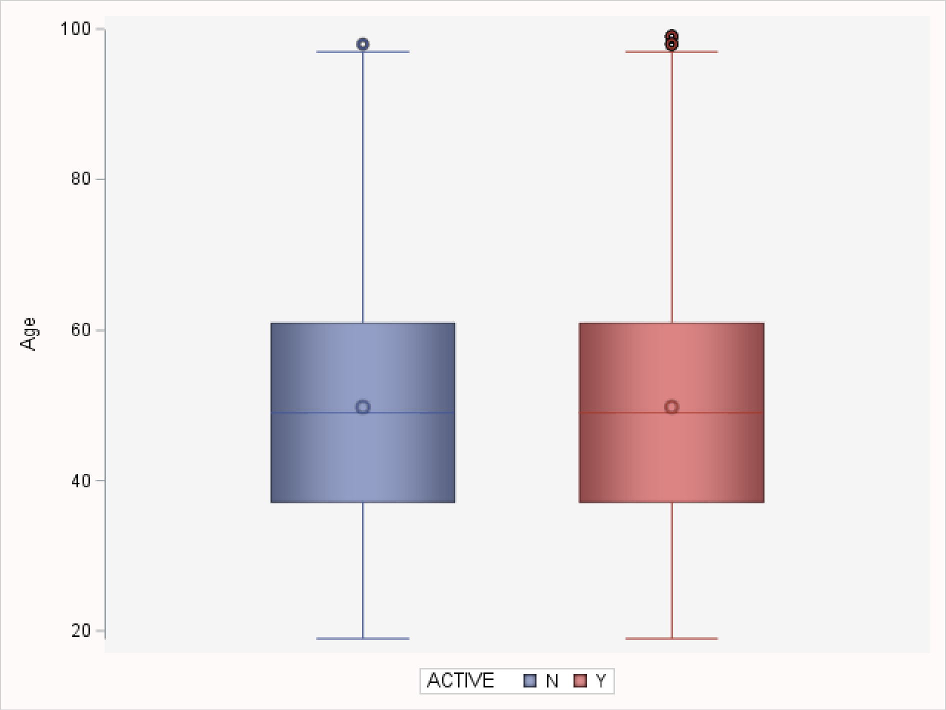
This test will be performed in due time, after trimming major outliers.

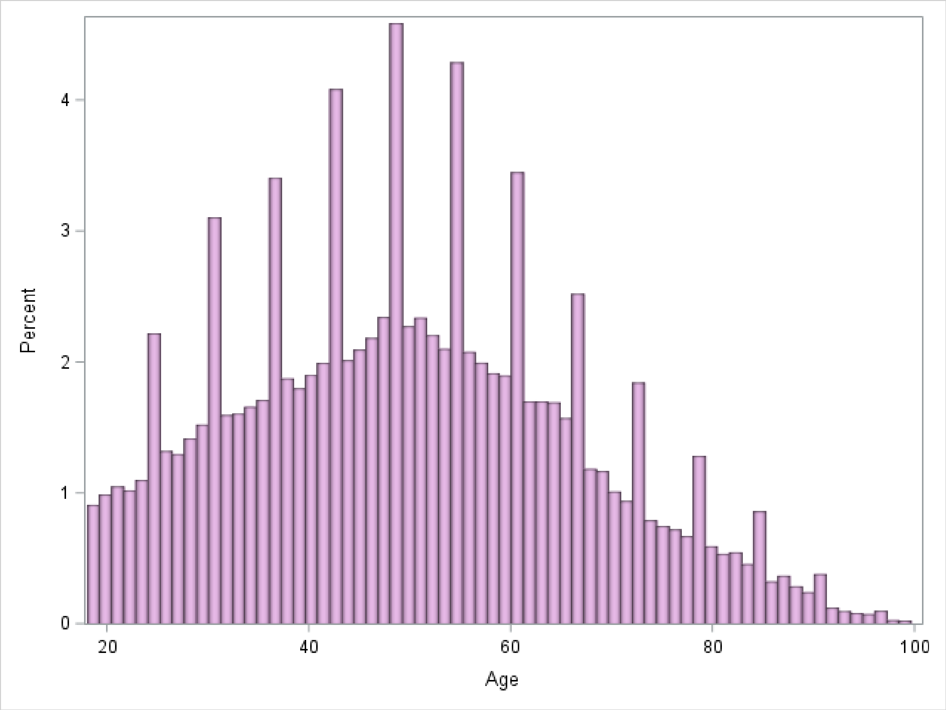
RELATION BETWEEN AGE AND ACTIVE

The MEANS Procedure

Analysis Variable : Age Age

ACTIVE	N Obs	N	N Miss	Minimum	Lower Quartile	Median	Mean	Upper Quartile	Maximum	Quartile Range	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean
N	**14824**	14824	0	19.00	37.00	49.00	49.77	61.00	98.00	24.00	33.92	49.50	50.04
Y	**62053**	62053	0	19.00	37.00	49.00	49.78	61.00	99.00	24.00	33.67	49.65	49.91





We can see that in each group between active and age features we have more than 100 observations, so there is no need to test for normal distribution. Let us test for homogeneity of variances:

The GLM Procedure

Class Level Information

Class	Levels	Values
ACTIVE	2	N Y

Number of Observations Read	**76877**
Number of Observations Used	76877

The GLM Procedure

Dependent Variable: Age Age

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.02	1.02	0.00	0.9520
Error	76875	21657940.05	281.73		
Corrected Total	76876	21657941.07			

R-Square	Coeff Var	Root MSE	Age Mean
0.000000	33.71869	16.78479	49.77891

Source	DF	Type I SS	Mean Square	F Value	Pr > F
ACTIVE	1	1.02065420	1.02065420	0.00	0.9520

Source	DF	Type III SS	Mean Square	F Value	Pr > F
ACTIVE	1	1.02065420	1.02065420	0.00	0.9520

The GLM Procedure

Levene's Test for Homogeneity of Age Variance
ANOVA of Absolute Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
ACTIVE	1	148.5	148.5	1.59	0.2072
Error	76875	7174743	93.3300		

Welch's ANOVA for Age

Source	DF	F Value	Pr > F
ACTIVE	1.0000	0.00	0.9522
Error	22327.9		

The GLM Procedure

Level of ACTIVE	N	Age	
		Mean	Std Dev
N	**14824**	49.7714517	16.8843539
Y	**62053**	49.7806875	16.7609233

We can see that Levene's test points to equal variances (pvalue of 0.2072), since we fail to reject null hypothesis at 5% significance

The TTEST Procedure

Variable: Age (Age)

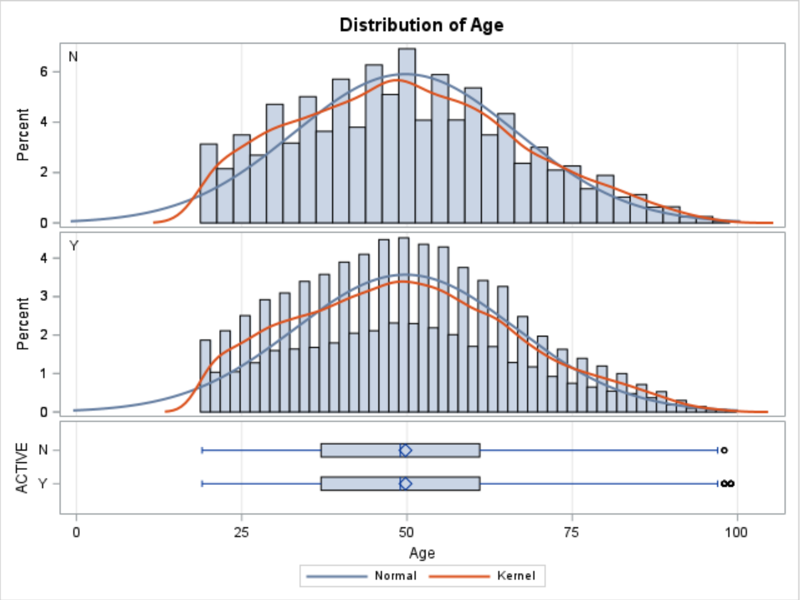
ACTIVE	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
N		14824	49.7715	16.8844	0.1387	19.0000	98.0000
Y		62053	49.7807	16.7609	0.0673	19.0000	99.0000
Diff (1-2)	**Pooled**		-0.00924	16.7848	0.1534		
Diff (1-2)	**Satterthwaite**		-0.00924		0.1541		

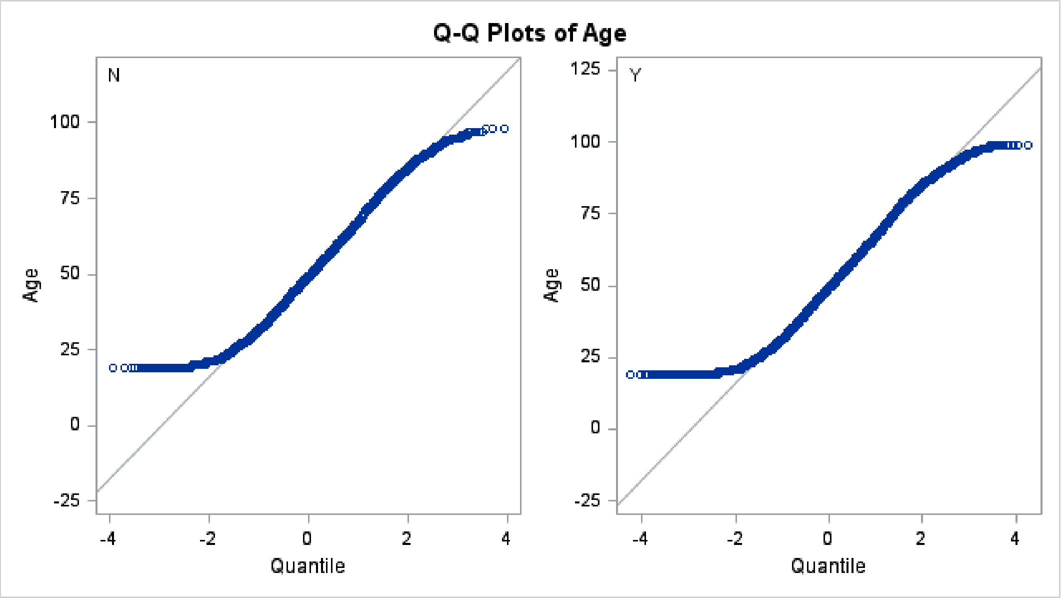
ACTIVE	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
N		49.7715	49.4996	50.0433	16.8844	16.6943	17.0788
Y		49.7807	49.6488	49.9126	16.7609	16.6682	16.8547
Diff (1-2)	**Pooled**	-0.00924	-0.3100	0.2915	16.7848	16.7013	16.8691
Diff (1-2)	**Satterthwaite**	-0.00924	-0.3114	0.2929			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	76875	-0.06	0.9520
Satterthwaite	Unequal	22328	-0.06	0.9522

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	14823	62052	1.01	0.2546





We can see that with a pvalue of 0.9520, we failed to reject null hypothesis. Thus, the two groups are equal.

Testing association - TENURE SEGMENT AND GOODCREDIT, RATEPLAN, AND DEALERTYPE

Assumptions:

2. The sample observations should be independent. No individual item should be included twice or more in the sample"

3. No expected frequencies should be small. Preferably each expected frequency should be larger than 10 but in any case not less than 5.

If condition of chi-square are satisfied and p-value is less than significant level (5%), reject null hypothesis:

- There is a relationship between them at 5% significant level.

BIVARIATE ANALYSIS OF TENURE_SEG AND GOODCREDIT FOR SEG

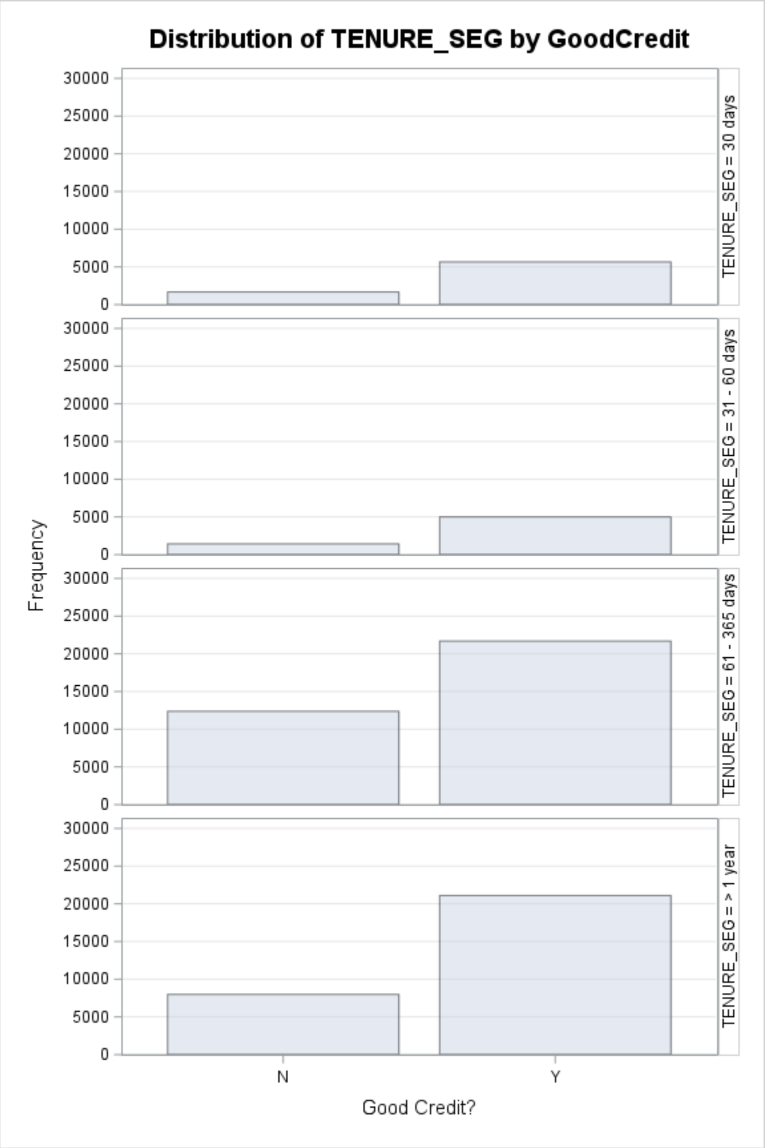
Null hypothesis: TENURE_SEG is independent of the GOODCREDIT

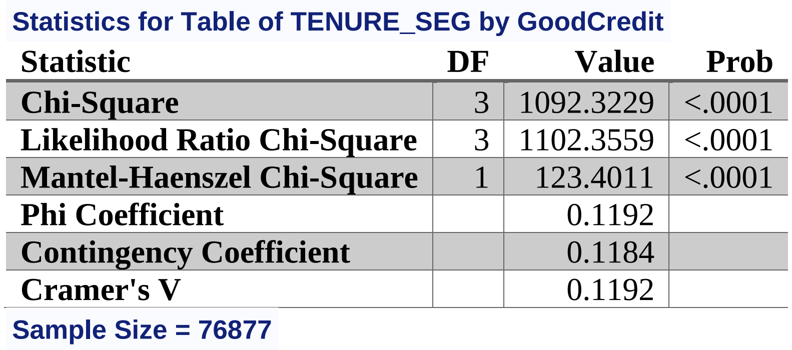
The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of TENURE_SEG by GoodCredit

TENURE_SEG	GoodCredit(Good Credit?)		
	N	Y	Total
30 days	1663	5652	7315
	2.16	7.35	9.52
	22.73	77.27	
	7.09	10.58	
31 - 60 days	1423	4991	6414
	1.85	6.49	8.34
	22.19	77.81	
	6.07	9.34	
61 - 365 days	12375	21688	34063
	16.10	28.21	44.31
	36.33	63.67	
	52.79	40.59	
> 1 year	7980	21105	29085
	10.38	27.45	37.83
	27.44	72.56	
	34.04	39.50	
Total	23441	53436	76877
	30.49	69.51	100.00





Statistics for Table of TENURE_SEG by GoodCredit

Statistic	DF	Value	Prob
Chi-Square	3	1092.3229	<.0001
Likelihood Ratio Chi-Square	3	1102.3559	<.0001
Mantel-Haenszel Chi-Square	1	123.4011	<.0001
Phi Coefficient		0.1192	
Contingency Coefficient		0.1184	
Cramer's V		0.1192	

Sample Size = 76877

If condition of chi-square are satisfied and p-value is less than significant level (5%),reject null hypothesis:There is a relationship between them at 5% significant level.

The assumptions are met. And, as pvalue is <.0001, we can reject the null hypothesis at 5% significance level and say that there's an association between the features.

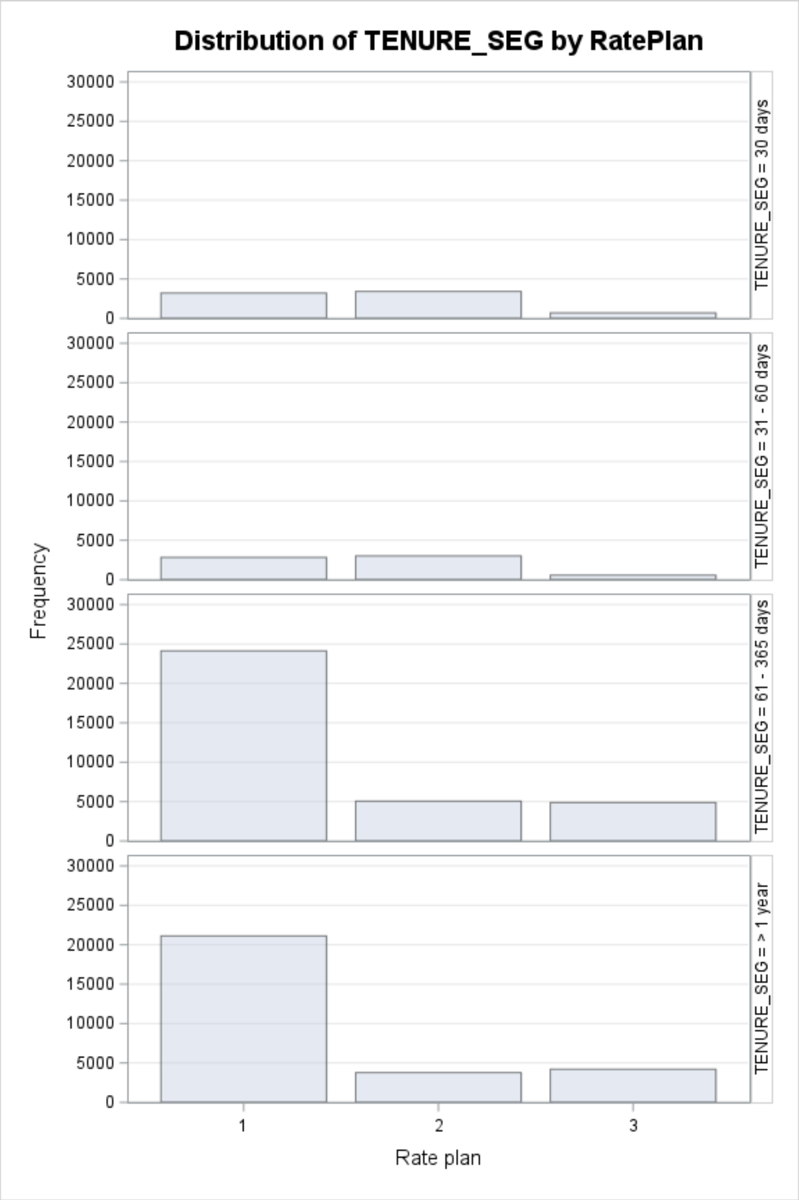
BIVARIATE ANALYSIS OF TENURE_SEG AND RATEPLAN FOR SEG
Null hypothesis: TENURE_SEG is independent of the RATEPLAN

The FREQ Procedure

Table of TENURE_SEG by RatePlan

Frequency
Percent
Row Pct
Col Pct

TENURE_SEG	RatePlan(Rate plan)			
	1	2	3	Total
30 days	3209 4.17 43.87 6.26	3411 4.44 46.63 22.36	695 0.90 9.50 6.72	7315 9.52
31 - 60 days	2829 3.68 44.11 5.52	3016 3.92 47.02 19.77	569 0.74 8.87 5.50	6414 8.34
61 - 365 days	24128 31.39 70.83 47.05	5058 6.58 14.85 33.16	4877 6.34 14.32 47.17	34063 44.31
> 1 year	21118 27.47 72.61 41.18	3769 4.90 12.96 24.71	4198 5.46 14.43 40.60	29085 37.83
Total	51284 66.71	15254 19.84	10339 13.45	76877 100.00



Statistics for Table of TENURE_SEG by RatePlan

Statistic	DF	Value	Prob
Chi-Square	6	7682.7184	<.0001
Likelihood Ratio Chi-Square	6	6579.6396	<.0001
Mantel-Haenszel Chi-Square	1	320.1327	<.0001
Phi Coefficient		0.3161	
Contingency Coefficient		0.3014	
Cramer's V		0.2235	

Sample Size = 76877

If condition of chi-square are satisfied and p-value is less than significant level (5%),reject null hypothesis:There is a relationship between them at 5% significant level.

The assumptions are met. And, as pvalue is <.0001, we can reject the null hypothesis at 5% significance level and say that there's an association between the features

BIVARIATE ANALYSIS OF TENURE_SEG AND DEALERTYPE FOR SEG

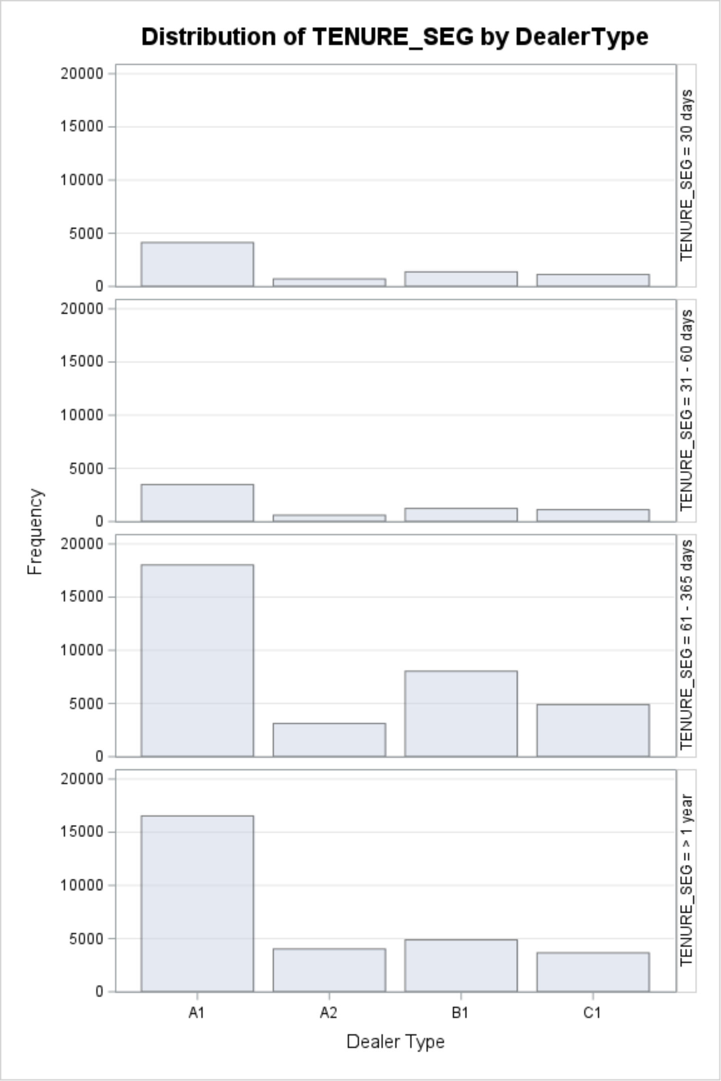
Null hypothesis: TENURE_SEG is independent of the DEALERTYPE

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of TENURE_SEG by DealerType

TENURE_SEG	DealerType(Dealer Type)				
	A1	A2	B1	C1	Total
30 days	4125	701	1367	1122	7315
	5.37	0.91	1.78	1.46	9.52
	56.39	9.58	18.69	15.34	
	9.78	8.32	8.82	10.40	
31 - 60 days	3477	594	1229	1114	6414
	4.52	0.77	1.60	1.45	8.34
	54.21	9.26	19.16	17.37	
	8.25	7.05	7.93	10.33	
61 - 365 days	18024	3114	8031	4894	34063
	23.45	4.05	10.45	6.37	44.31
	52.91	9.14	23.58	14.37	
	42.75	36.94	51.80	45.38	
> 1 year	16534	4021	4876	3654	29085
	21.51	5.23	6.34	4.75	37.83
	56.85	13.82	16.76	12.56	
	39.22	47.70	31.45	33.88	
Total	42160	8430	15503	10784	76877
	54.84	10.97	20.17	14.03	100.00



Statistics for Table of TENURE_SEG by DealerType

Statistic	DF	Value	Prob
Chi-Square	9	879.7876	<.0001
Likelihood Ratio Chi-Square	9	868.8502	<.0001
Mantel-Haenszel Chi-Square	1	231.6874	<.0001
Phi Coefficient		0.1070	
Contingency Coefficient		0.1064	
Cramer's V		0.0618	

Sample Size = 76877

If condition of chi-square are satisfied and p-value is less than significant level (5%),reject null hypothesis: There is a relationship between them at 5% significant level.

The assumptions are met. And, as pvalue is <.0001, we can reject the null hypothesis at 5% significance level and say that there is an association between the features.

CATEGORICAL VS CONTINUOUS
Testing association - SALES AMOUNT AND ACCOUNT STATUS, GOOD CREDIT AND AGE SEGMENTS

Null hypothesis: There’s no difference in means

1.Sample distribution must be normal:

CLT :

If looks normal each group must have more than 30 observations – no need for Shapiro’s test

If moderately skewed, each group must have more than 100 observations – no need for Shapiro’s test

2.Groups are independent of one another.

3.There are no major outliers.

4.A check for unequal variances will help determine which version of an independent samples t-test is most appropriate:

(Levene’s test, null hypothesis: equal variance)

a.If variances are equal, then a pooled t-test is appropriate

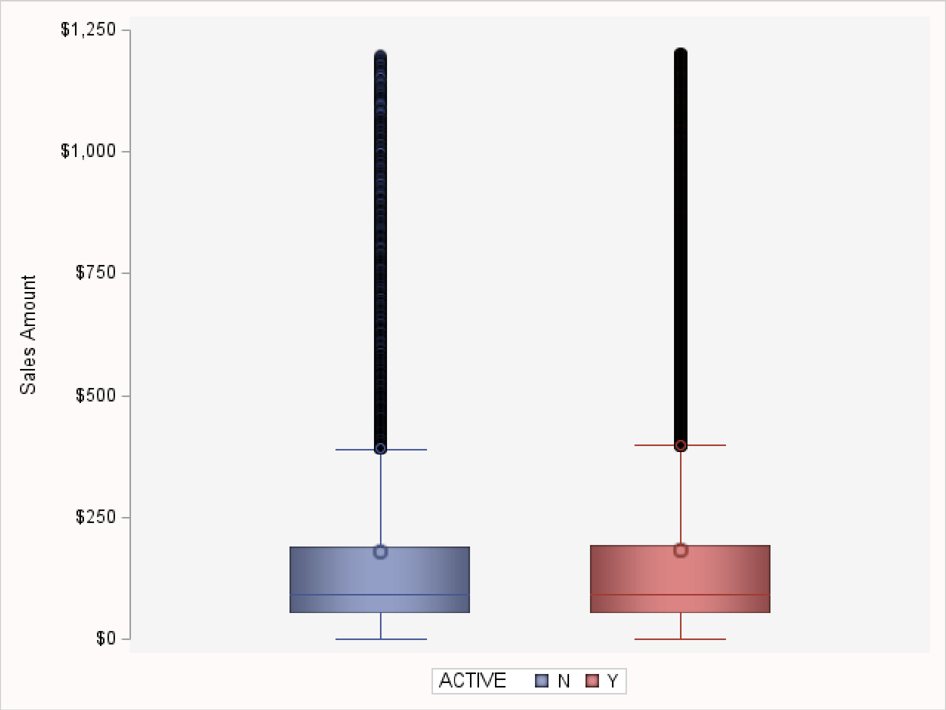
b.If variances are unequal, then a Satterthwaite (also known as Welch’s) t-test is appropriate

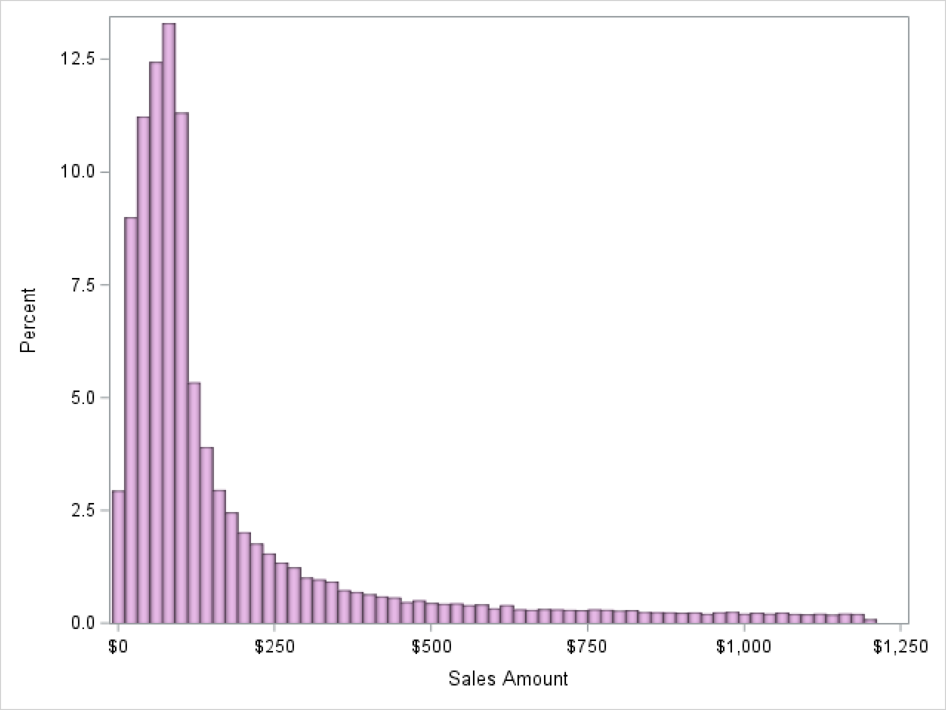
RELATION BETWEEN SALES AND ACTIVE

The MEANS Procedure

Analysis Variable : Sales Sales Amount

ACTIVE	N Obs	N	N Miss	Minimum	Lower Quartile	Median	Mean	Upper Quartile	Maximum	Quartile Range	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean
N	**14824**	14824	0	0.00	53.00	92.00	180.02	188.00	1197.00	135.00	128.65	176.30	183.75
Y	**62053**	62053	0	0.00	52.00	91.00	181.63	191.00	1200.00	139.00	129.06	179.78	183.47





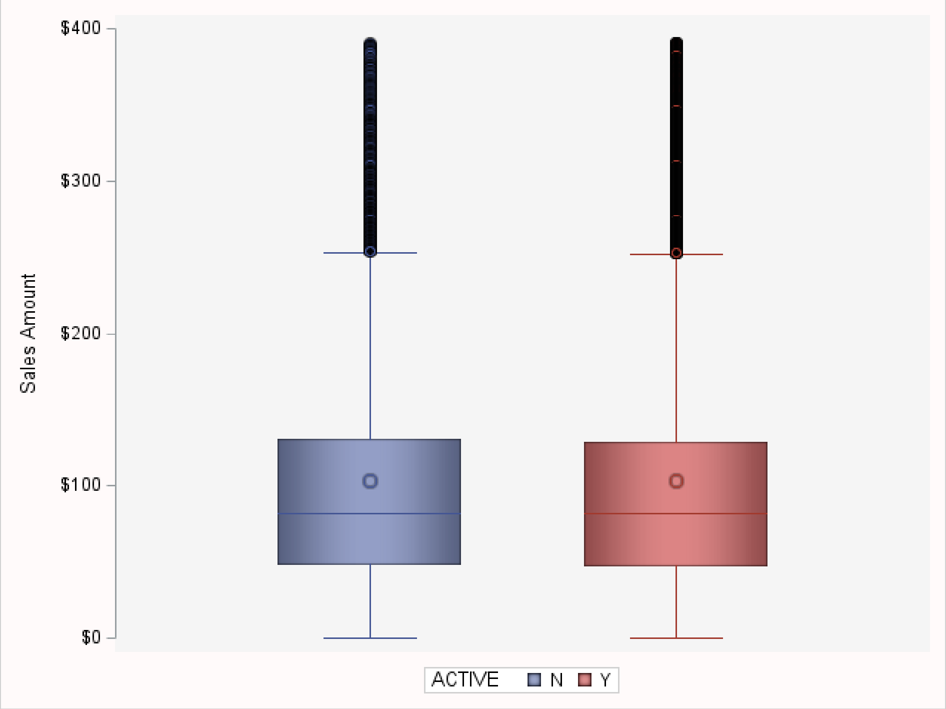
Sales greater than $390 (~Q3+3*IQR) will be dropped to perform the test.

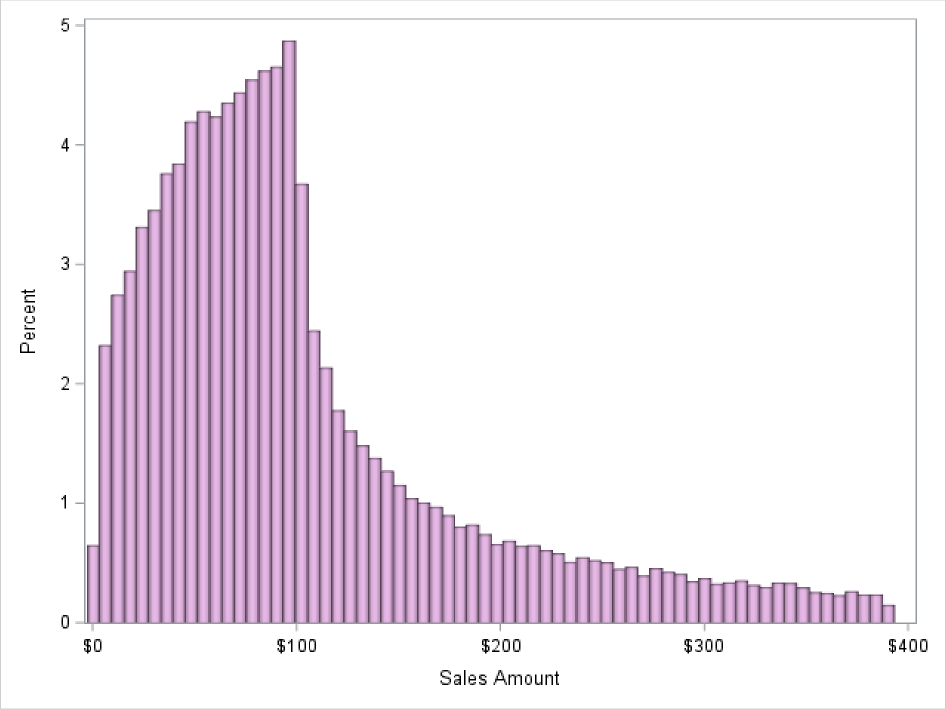
RELATION BETWEEN SALES AND ACTIVE

The MEANS Procedure

Analysis Variable : Sales Sales Amount

ACTIVE	N Obs	N	N Miss	Minimum	Lower Quartile	Median	Mean	Upper Quartile	Maximum	Quartile Range	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean
N	**12973**	12973	0	0.00	48.00	82.00	103.33	130.00	390.00	82.00	78.79	101.93	104.74
Y	**54056**	54056	0	0.00	47.00	82.00	102.75	129.00	390.00	82.00	79.59	102.07	103.44





We can see that in each group between active and age features we have more than 100 observations, so there's no need to test for normality.

Let's test for homogeneity of variances:

The GLM Procedure

Class Level Information

Class	Levels	Values
ACTIVE	2	N Y

Number of Observations Read	67029
Number of Observations Used	67029

The GLM Procedure

Dependent Variable: Sales Sales Amount

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3519.3	3519.3	0.53	0.4678
Error	67027	447480229.2	6676.1		
Corrected Total	67028	447483748.5			

R-Square	Coeff Var	Root MSE	Sales Mean
0.000008	79.43046	81.70752	102.8667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
ACTIVE	1	3519.315269	3519.315269	0.53	0.4678

Source	DF	Type III SS	Mean Square	F Value	Pr > F
ACTIVE	1	3519.315269	3519.315269	0.53	0.4678

The GLM Procedure

Levene's Test for Homogeneity of Sales Variance
ANOVA of Absolute Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
ACTIVE	1	117.5	117.5	0.04	0.8427
Error	67027	1.9985E8	2981.6		

Welch's ANOVA for Sales

Source	DF	F Value	Pr > F
ACTIVE	1.0000	0.53	0.4666
Error	19736.7		

The GLM Procedure

Level of ACTIVE	N	Sales	
		Mean	Std Dev
N	12973	103.334464	81.4151231
Y	54056	102.754477	81.7775346

We can see that Levene's test points to equal variances (pvalue of 0.8427), since we fail to reject null hypothesis at 5% significance level.

The TTEST Procedure

Variable: Sales (Sales Amount)

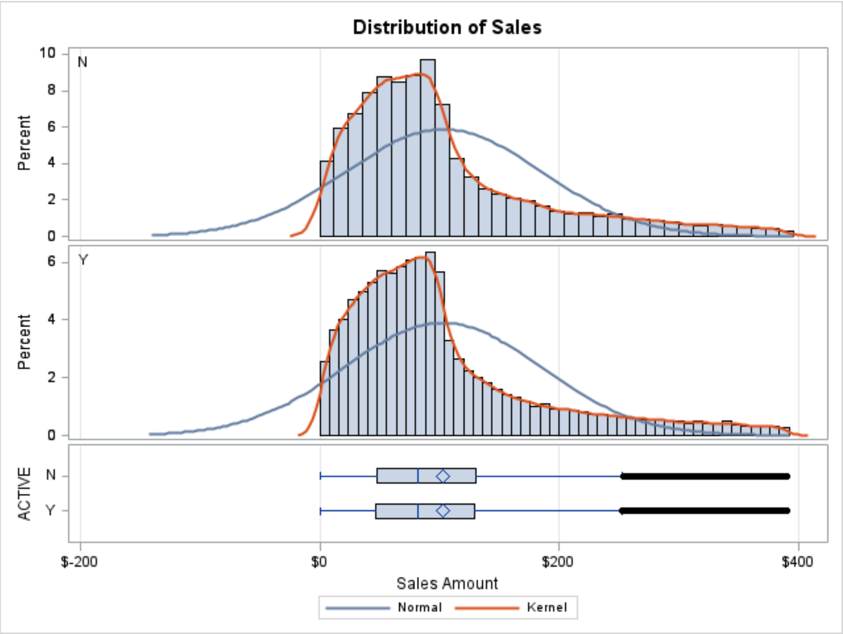
ACTIVE	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
N		12973	103.3	81.4151	0.7148	0	390.0
Y		54056	102.8	81.7775	0.3517	0	390.0
Diff (1-2)	Pooled		0.5800	81.7075	0.7988		
Diff (1-2)	Satterthwaite		0.5800		0.7967		

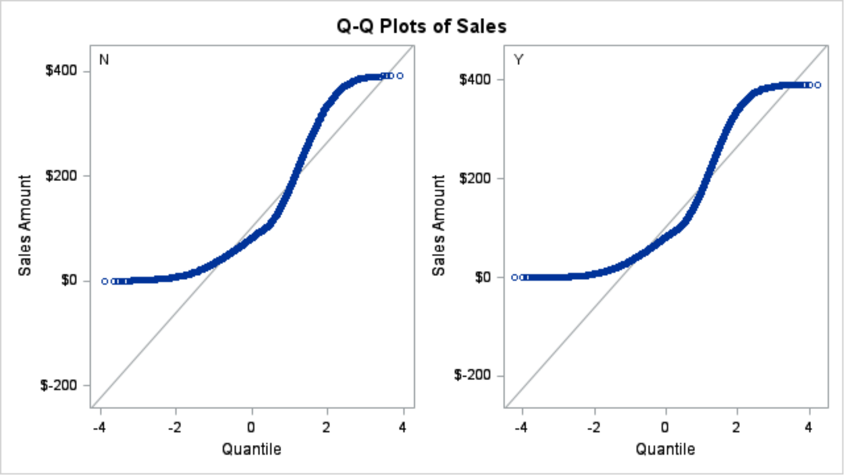
ACTIVE	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
N		103.3	101.9	104.7	81.4151	80.4364	82.4181
Y		102.8	102.1	103.4	81.7775	81.2930	82.2679
Diff (1-2)	Pooled	0.5800	-0.9857	2.1457	81.7075	81.2725	82.1473
Diff (1-2)	Satterthwaite	0.5800	-0.9815	2.1415			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	67027	0.73	0.4678
Satterthwaite	Unequal	19737	0.73	0.4666

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	54055	12972	1.01	0.5228





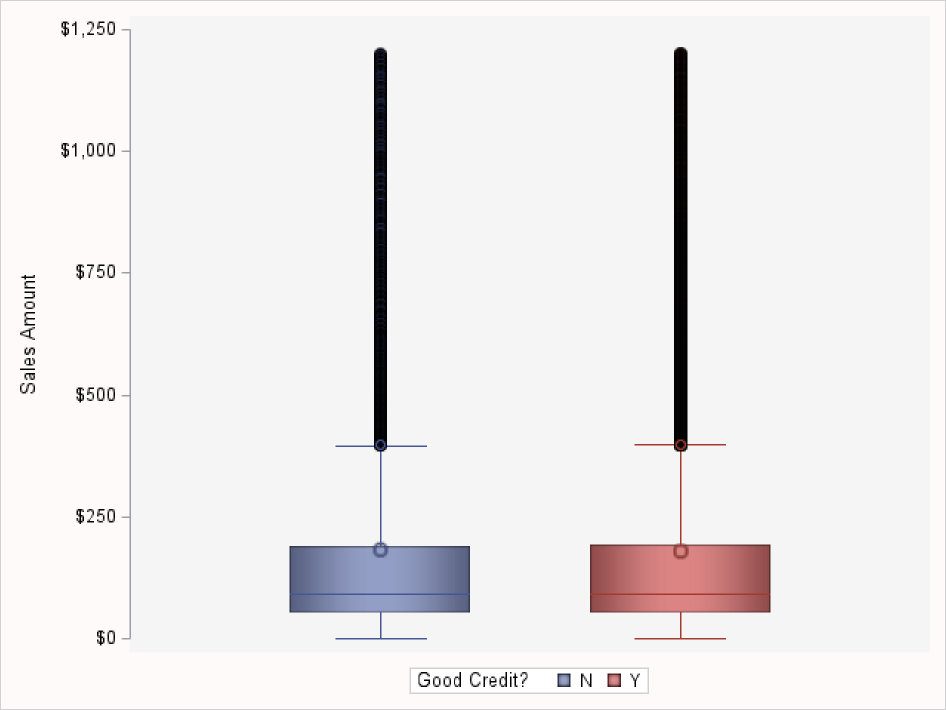
We can see that with a pvalue of 0.4678, we failed to reject null hypothesis. Thus, the two groups are equal

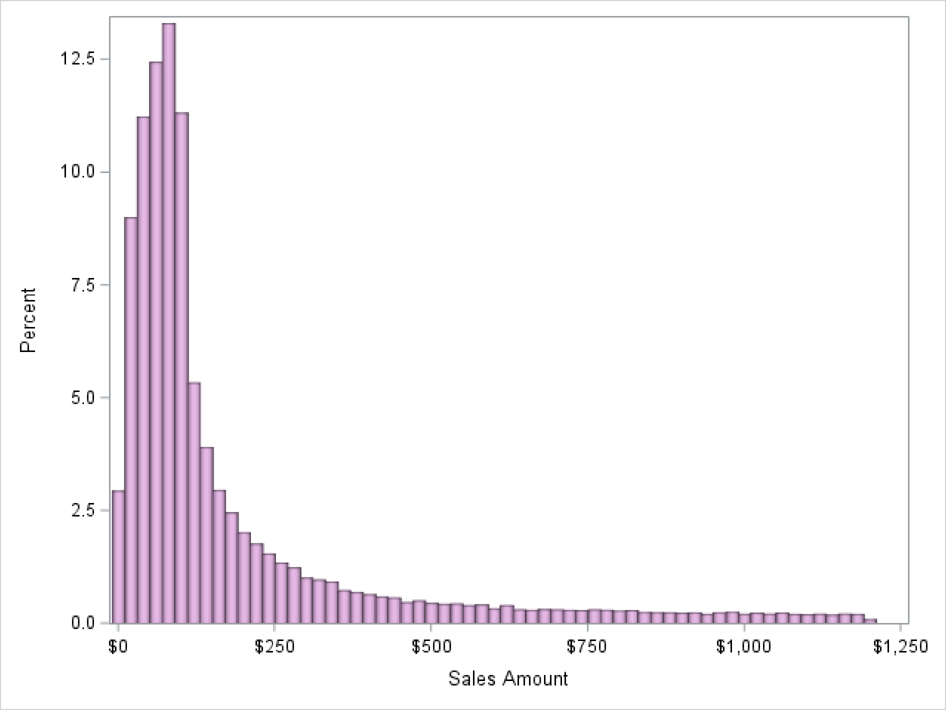
RELATION BETWEEN SALES AND GOODCREDIT

The MEANS Procedure

Analysis Variable : Sales Sales Amount

Good Credit?	N Obs	N	N Miss	Minimum	Lower Quartile	Median	Mean	Upper Quartile	Maximum	Quartile Range	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean
N	**23441**	23441	0	0.00	52.00	91.00	181.71	190.00	1200.00	138.00	129.41	178.70	184.72
Y	**53436**	53436	0	0.00	53.00	91.00	181.15	191.00	1200.00	138.00	128.79	179.17	183.13





We can see a great number of major outliers, so as the data is, it's not possible to use t-test for sales and active features.

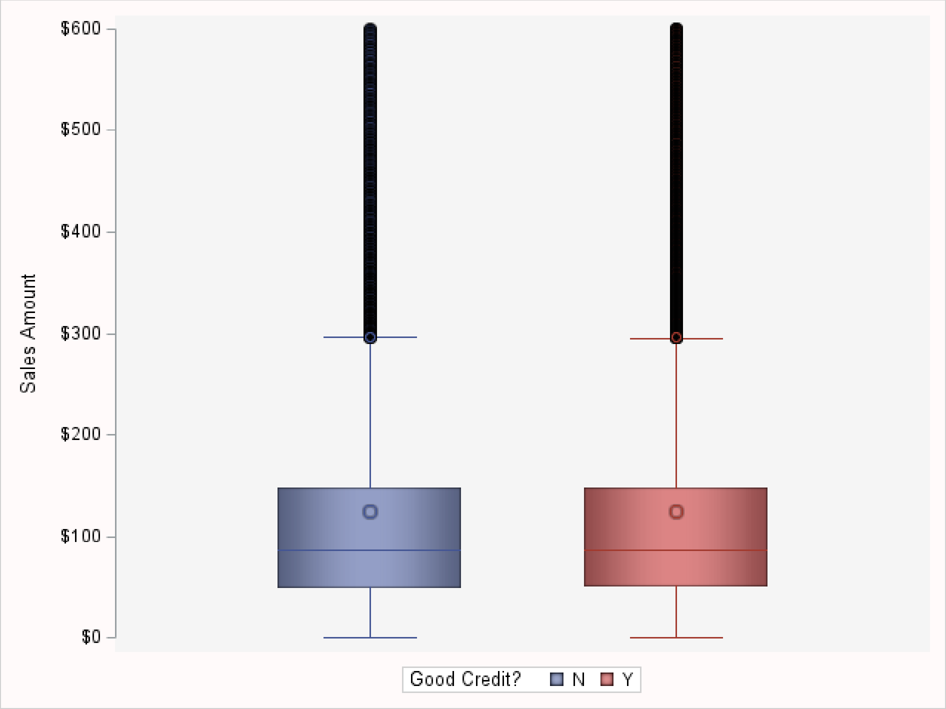
Sales greater than $600 (~Q3+3*IQR) will be dropped to perform the test.

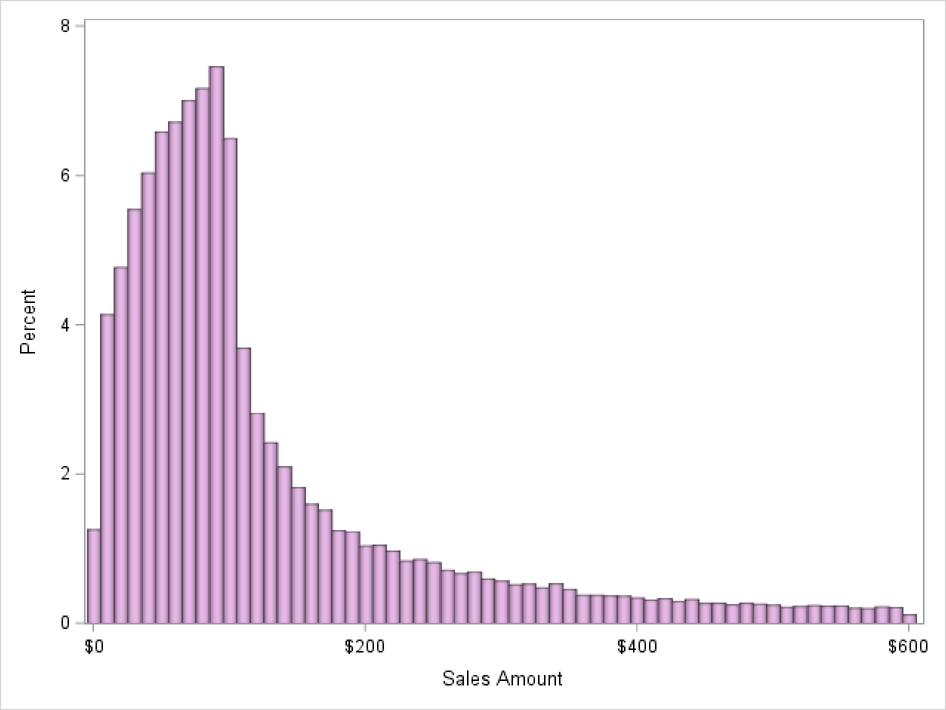
RELATION BETWEEN SALES AND GOODCREDIT

The MEANS Procedure

Analysis Variable : Sales Sales Amount

Good Credit?	N Obs	N	N Miss	Minimum	Lower Quartile	Median	Mean	Upper Quartile	Maximum	Quartile Range	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean
N	**21625**	21625	0	0.00	49.00	86.00	123.82	148.00	600.00	99.00	95.86	122.24	125.40
Y	**49347**	49347	0	0.00	50.00	86.00	124.30	148.00	600.00	98.00	96.10	123.25	125.35





.

We can see that in each group between active and age features we have more than 100 observations, so there's no need to test for normality

Let's test for homogeneity of variances:

The GLM Procedure

Class Level Information

Class	Levels	Values
GoodCredit	2	N Y

Number of Observations Read	70972
Number of Observations Used	70972

The GLM Procedure

Dependent Variable: Sales Sales Amount

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3471	3471	0.24	0.6212
Error	70970	1008759435	14214		
Corrected Total	70971	1008762905			

R-Square	Coeff Var	Root MSE	Sales Mean
0.000003	96.02857	119.2220	124.1526

Source	DF	Type I SS	Mean Square	F Value	Pr > F
GoodCredit	1	3470.566132	3470.566132	0.24	0.6212

Source	DF	Type III SS	Mean Square	F Value	Pr > F
GoodCredit	1	3470.566132	3470.566132	0.24	0.6212

The GLM Procedure

Levene's Test for Homogeneity of Sales Variance
ANOVA of Absolute Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
GoodCredit	1	2651.5	2651.5	0.39	0.5345
Error	70970	4.8763E8	6871.0		

Welch's ANOVA for Sales

Source	DF	F Value	Pr > F
GoodCredit	1.0000	0.25	0.6203
Error	41500.1		

The GLM Procedure

Level of GoodCredit	N	Sales	
		Mean	Std Dev
N	21625	123.818590	118.686942
Y	49347	124.299025	119.455713

We can see that Levene's test points to equal variances (pvalue of 0.5345), since we fail to reject null hypothesis at 5% significance level.

The TTEST Procedure

Variable: Sales (Sales Amount)

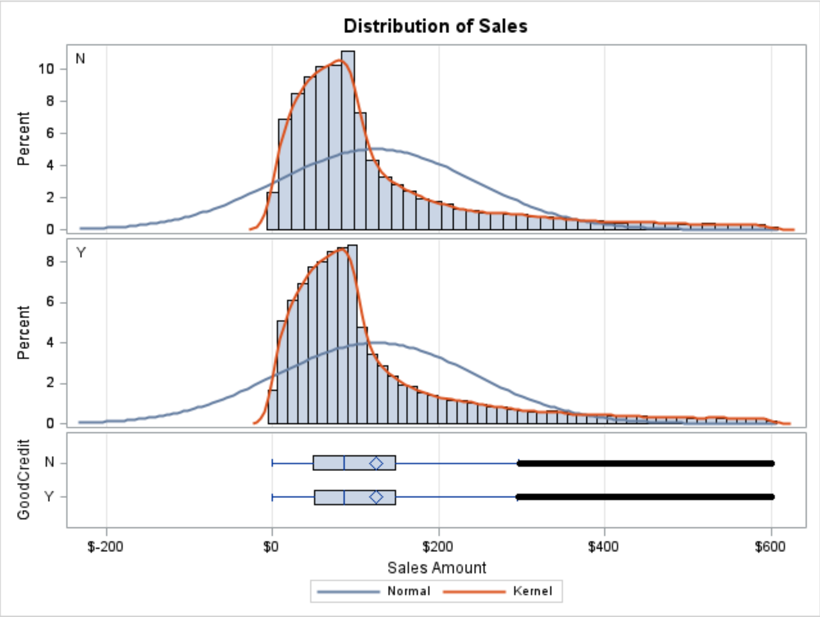
GoodCredit	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
N		21625	123.8	118.7	0.8071	0	600.0
Y		49347	124.3	119.5	0.5377	0	600.0
Diff (1-2)	**Pooled**		-0.4804	119.2	0.9723		
Diff (1-2)	**Satterthwaite**		-0.4804		0.9698		

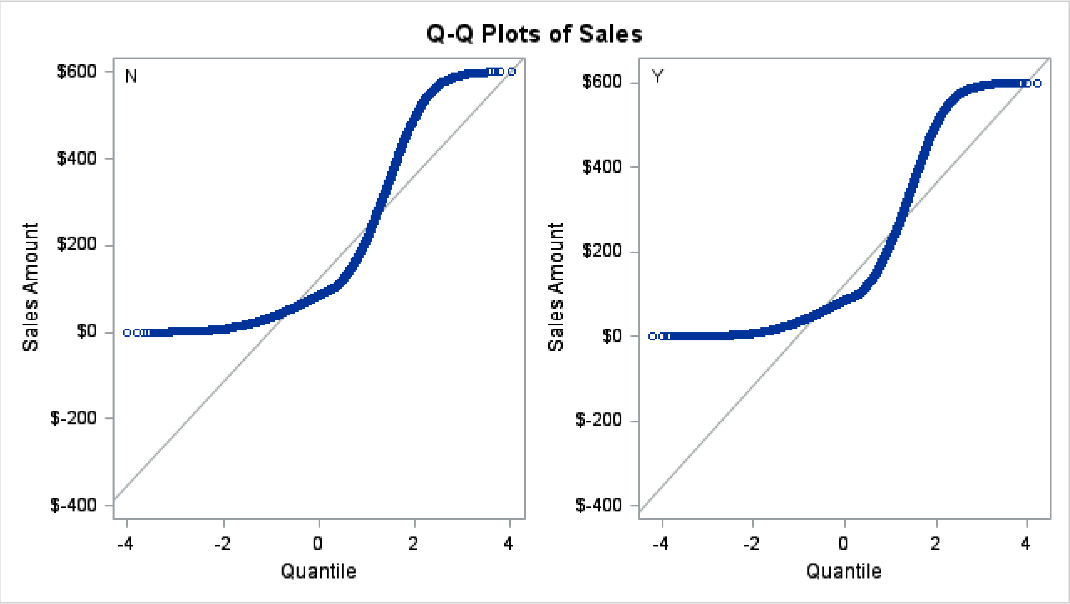
GoodCredit	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
N		123.8	122.2	125.4	118.7	117.6	119.8
Y		124.3	123.2	125.4	119.5	118.7	120.2
Diff (1-2)	**Pooled**	-0.4804	-2.3861	1.4252	119.2	118.6	119.8
Diff (1-2)	**Satterthwaite**	-0.4804	-2.3813	1.4205			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	70970	-0.49	0.6212
Satterthwaite	Unequal	41500	-0.50	0.6203

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	49346	21624	1.01	0.2639





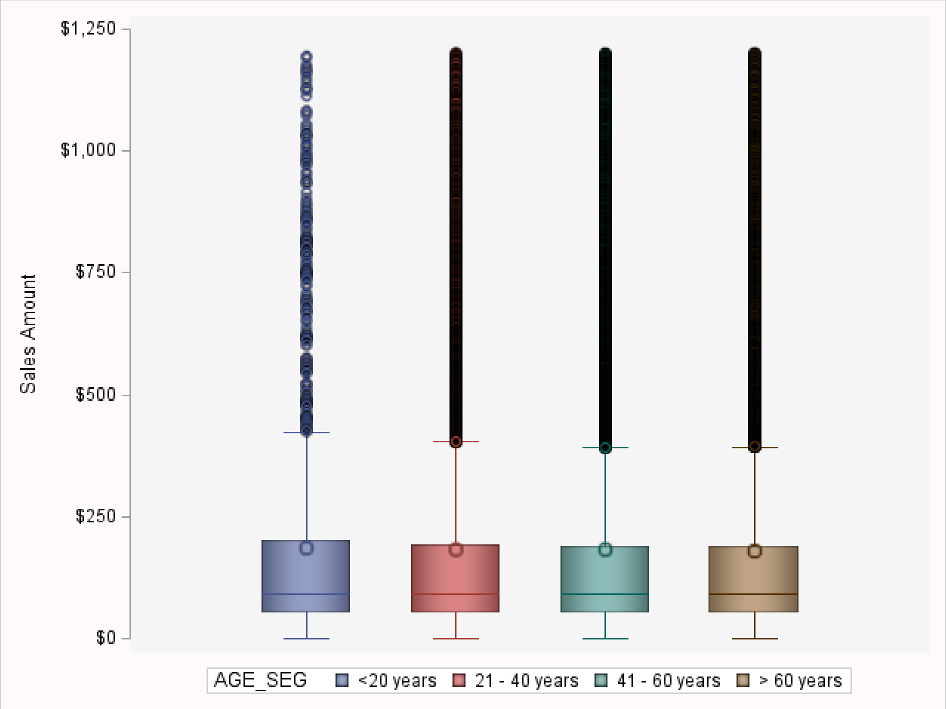
We can see that with a pvalue of 0.6212, we failed to reject null hypothesis. Thus, the two groups are equal.

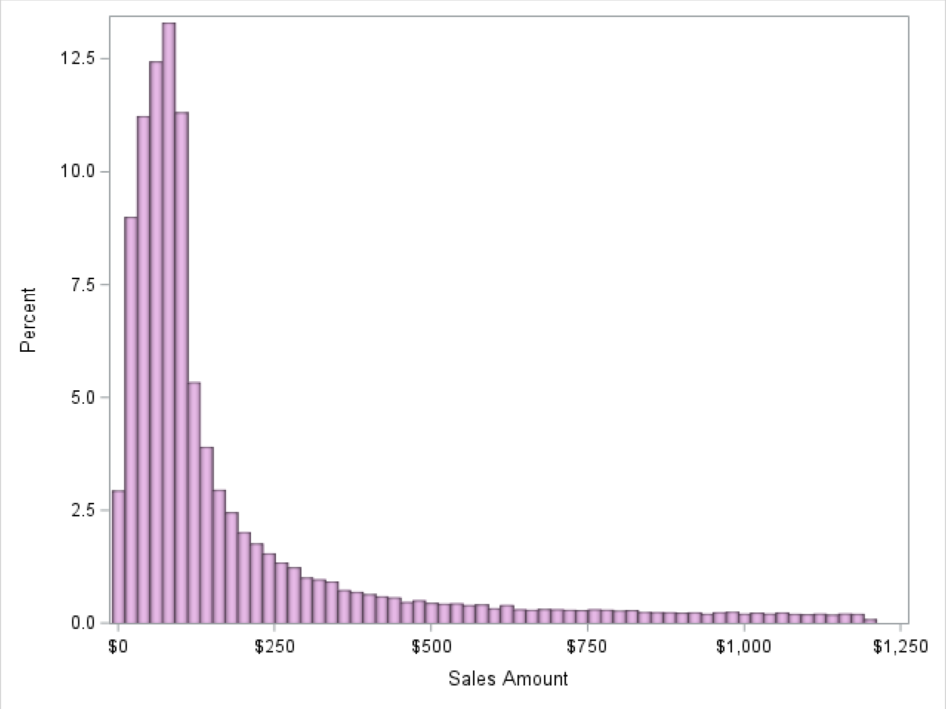
RELATION BETWEEN SALES AND AGE_SEG

The MEANS Procedure

Analysis Variable : Sales Sales Amount

AGE_SEG	N Obs	N	N Miss	Minimum	Lower Quartile	Median	Mean	Upper Quartile	Maximum	Quartile Range	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean
<20 years	**1456**	1456	0	0.00	53.00	90.00	184.53	201.00	1195.00	148.00	128.59	172.33	196.72
21 - 40 years	**22757**	22757	0	0.00	52.00	91.00	182.09	193.00	1200.00	141.00	129.14	179.04	185.15
41 - 60 years	**32369**	32369	0	0.00	53.00	92.00	181.31	189.00	1200.00	136.00	128.93	178.77	183.86
> 60 years	**20295**	20295	0	0.00	52.00	91.00	180.23	189.00	1200.00	137.00	128.91	177.04	183.43





We can see a great number of major outliers, so as the data is, it's not possible to use t-test for sales and active features.

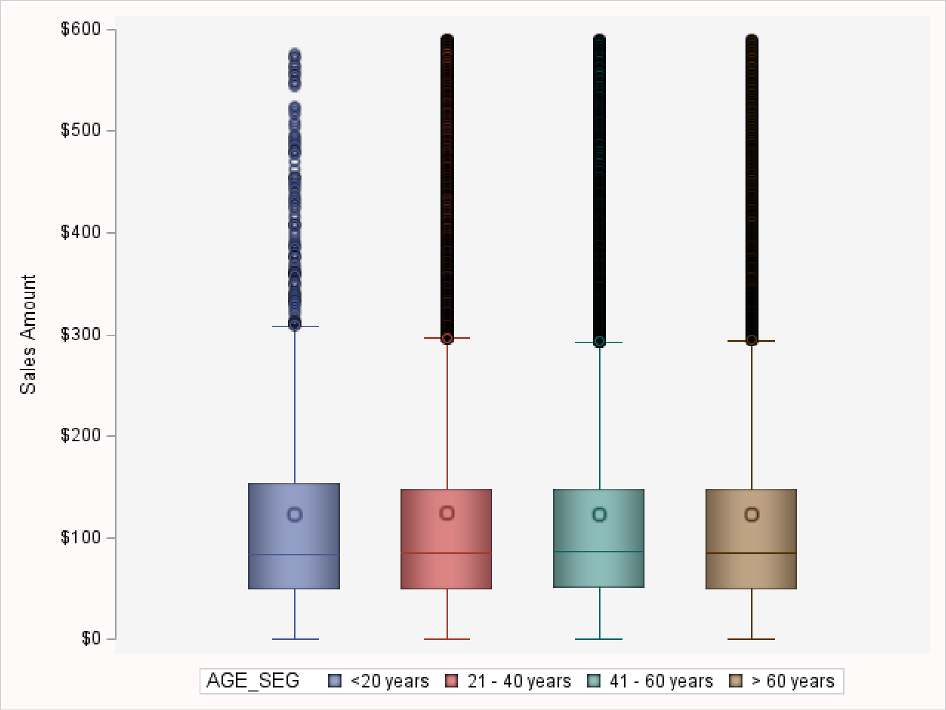
Sales greater than $590 (~Q3+3*IQR) will be dropped to perform the test.

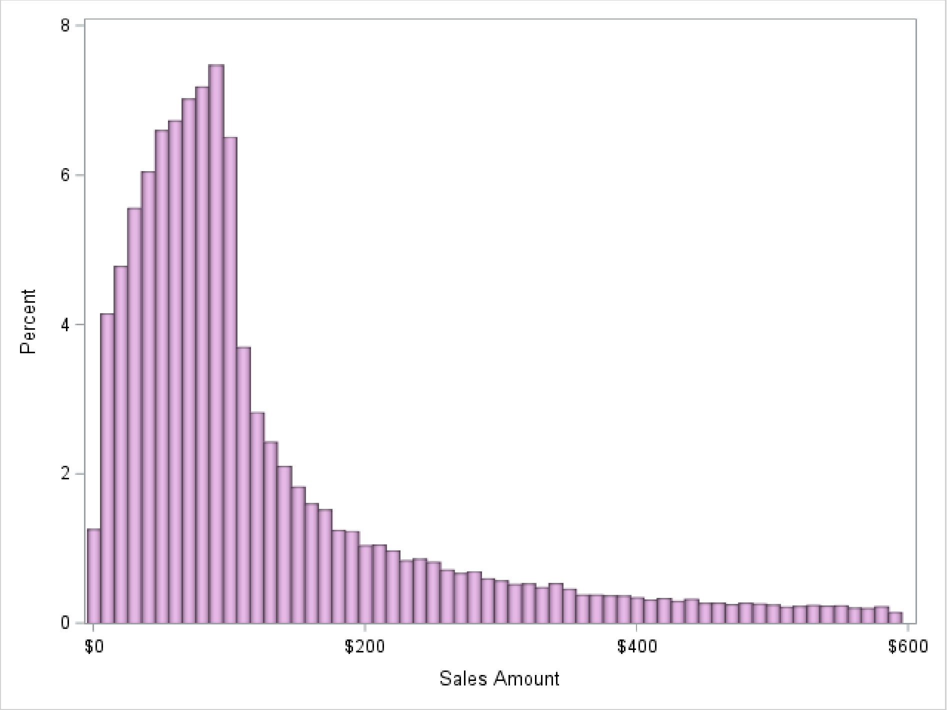
RELATION BETWEEN SALES AND AGE_SEG

The MEANS Procedure

Analysis Variable : Sales Sales Amount

AGE_SEG	N Obs	N	N Miss	Minimum	Lower Quartile	Median	Mean	Upper Quartile	Maximum	Quartile Range	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean
<20 years	**1332**	1332	0	0.00	49.00	83.00	122.24	153.00	576.00	104.00	94.99	116.00	128.48
21 - 40 years	**20969**	20969	0	0.00	49.00	85.00	123.87	148.00	590.00	99.00	96.26	122.26	125.49
41 - 60 years	**29824**	29824	0	0.00	50.00	86.00	123.18	147.00	590.00	97.00	94.94	121.85	124.51
> 60 years	**18712**	18712	0	0.00	49.00	85.00	122.75	147.00	590.00	98.00	95.05	121.07	124.42





We can see that in each group between active and age features we have more than 100 observations, so there's no need to test for normality. Let's test for homogeneity of variances:

The GLM Procedure

Class Level Information

Class	Levels	Values
AGE_SEG	4	<20 years 21 - 40 years 41 - 60 years > 60 years

Number of Observations Read	**70837**
Number of Observations Used	70837

The GLM Procedure

Dependent Variable: Sales Sales Amount

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	14420.5	4806.8	0.35	0.7907
Error	70833	978671478.7	13816.6		
Corrected Total	70836	978685899.2			

R-Square	Coeff Var	Root MSE	Sales Mean
0.000015	95.36736	117.5440	123.2540

Source	DF	Type I SS	Mean Square	F Value	Pr > F
AGE_SEG	3	14420.49490	4806.83163	0.35	0.7907

Source	DF	Type III SS	Mean Square	F Value	Pr > F
AGE_SEG	3	14420.49490	4806.83163	0.35	0.7907

The GLM Procedure

Levene's Test for Homogeneity of Sales Variance
ANOVA of Absolute Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
AGE_SEG	3	45955.9	15318.6	2.30	0.0751
Error	70833	4.7163E8	6658.3		

Welch's ANOVA for Sales

Source	DF	F Value	Pr > F
AGE_SEG	3.0000	0.34	0.7929
Error	6356.7		

The GLM Procedure

Level of	N	Sales	
AGE_SEG		Mean	Std Dev
<20 years	**1332**	122.242492	116.121697
21 - 40 years	**20969**	123.874481	119.244211
41 - 60 years	**29824**	123.181532	116.950778
> 60 years	**18712**	122.746045	116.665526

We can see that Levene's test points to equal variances (pvalue of 0.0751), since we fail to reject null hypothesis at 5% significance level.

The GLM Procedure

Class Level Information

Class	Levels	Values
AGE_SEG	4	<20 years 21 - 40 years 41 - 60 years > 60 years

Number of Observations Read	**70837**
Number of Observations Used	70837

The GLM Procedure

Dependent Variable: Sales Sales Amount

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	14420.5	4806.8	0.35	0.7907
Error	70833	978671478.7	13816.6		
Corrected Total	70836	978685899.2			

R-Square	Coeff Var	Root MSE	Sales Mean
0.000015	95.36736	117.5440	123.2540

Source	DF	Type I SS	Mean Square	F Value	Pr > F
AGE_SEG	3	14420.49490	4806.83163	0.35	0.7907

Source	DF	Type III SS	Mean Square	F Value	Pr > F
AGE_SEG	3	14420.49490	4806.83163	0.35	0.7907

We can see that with a pvalue of 0.7907, we failed to reject null hypothesis. Thus, the two groups are equal.

CONCLUSIONS

We can see that from the gathered data, there may be a trend of increase in activations in the beginning of year and in the middle.
It seems to be a threshold of 60 days in tenure that either makes customer's leave or stay in a long term relationship.

The NEED reason is the one being most used.

Account Status it's impacted by:
Good credit
Rate Plan
Dealer Type
Tenure(Segmented)

Segmented Tenure it's impacted by:
Good credit
Rate Plan
Dealer Type

Sales amount it's not impacted by Account Status, Good credit, or even Age.

RECOMMENDATIONS

Observe the increasing of deactivations in the last 6 months of 2000 and beginning of 2001.

Investigate further the type NEED of deactivations reasons to look for a direct marketing strategy.

Investigate further to see the threshold between a finer adjustment the credit score checking would bring benefits.

Investigate further Rate Plan 1 for its success with the customers to replicate its features into the other plans.

Investigate further Dealer Type A1 for its success with the customers to replicate its features with other dealers.

Investigate further the Tenure Segments.

Next steps: do multivariate analysis(when possible) with the features mentioned above to find other associations.