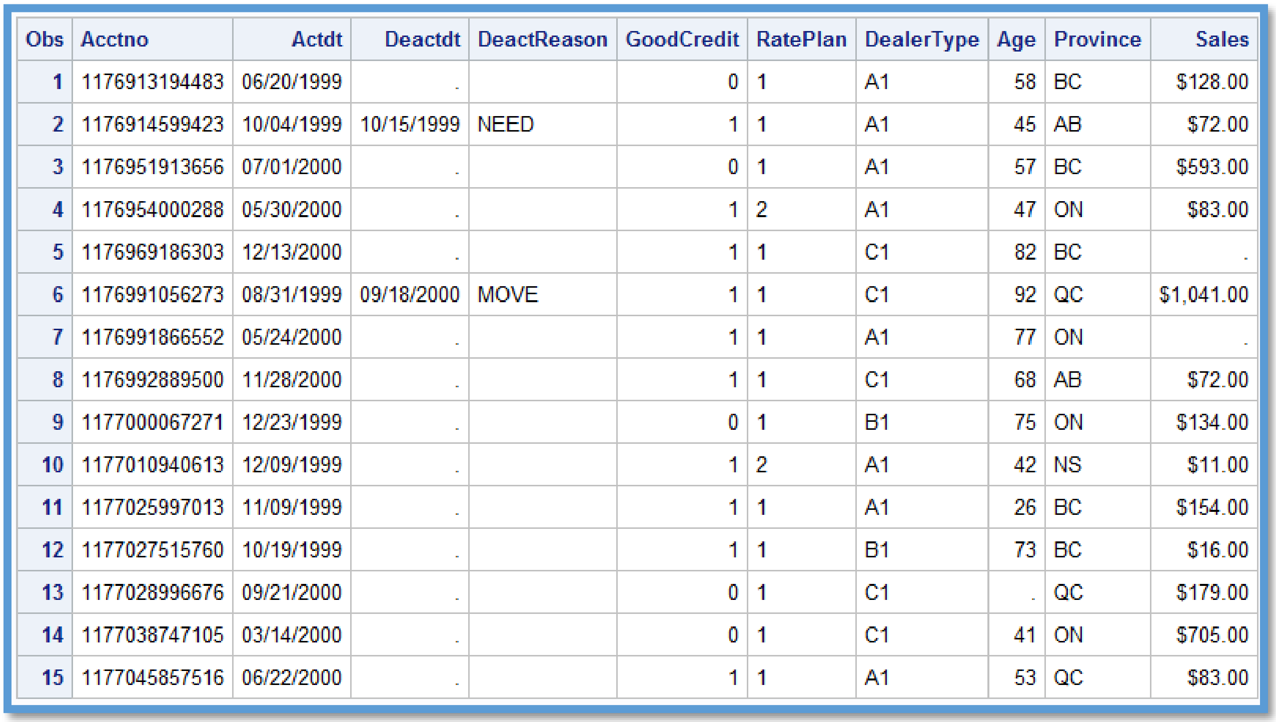
EXPLORATORY DATA ANALYSIS –TELECOM DATASET

Objective-

- Analyzing the CRM data of a wireless company for 2 years to investigate the customer distribution and business behaviors.
- Gain insightful understanding about the customers, and to forecast the deactivation trends for the next 6 months

Dataset Overview-

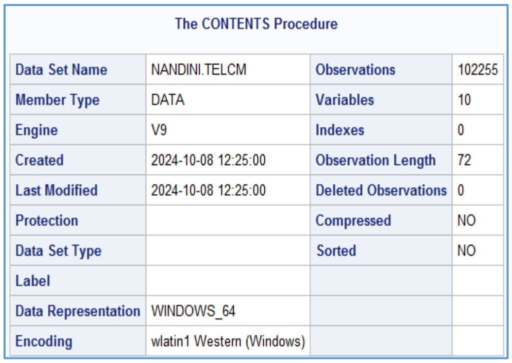


Obs	Acctno	Actdt	Deactdt	DeactReason	GoodCredit	RatePlan	DealerType	Age	Province	Sales
1	1176913194483	06/20/1999	.		0	1	A1	58	BC	$128.00
2	1176914599423	10/04/1999	10/15/1999	NEED	1	1	A1	45	AB	$72.00
3	1176951913656	07/01/2000	.		0	1	A1	57	BC	$593.00
4	1176954000288	05/30/2000	.		1	2	A1	47	ON	$83.00
5	1176969186303	12/13/2000	.		1	1	C1	82	BC	.
6	1176991056273	08/31/1999	09/18/2000	MOVE	1	1	C1	92	QC	$1,041.00
7	1176991866552	05/24/2000	.		1	1	A1	77	ON	.
8	1176992889500	11/28/2000	.		1	1	C1	68	AB	$72.00
9	1177000067271	12/23/1999	.		0	1	B1	75	ON	$134.00
10	1177010940613	12/09/1999	.		1	2	A1	42	NS	$11.00
11	1177025997013	11/09/1999	.		1	1	A1	26	BC	$154.00
12	1177027515760	10/19/1999	.		1	1	B1	73	BC	$16.00
13	1177028996676	09/21/2000	.		0	1	C1	.	QC	$179.00
14	1177038747105	03/14/2000	.		0	1	C1	41	ON	$705.00
15	1177045857516	06/22/2000	.		1	1	A1	53	QC	$83.00

The Telecom Dataset demonstrates the distribution of 102255 observation which are categorized into 10 variables out of which 5 are categorical and 5 are numerical which are as follows

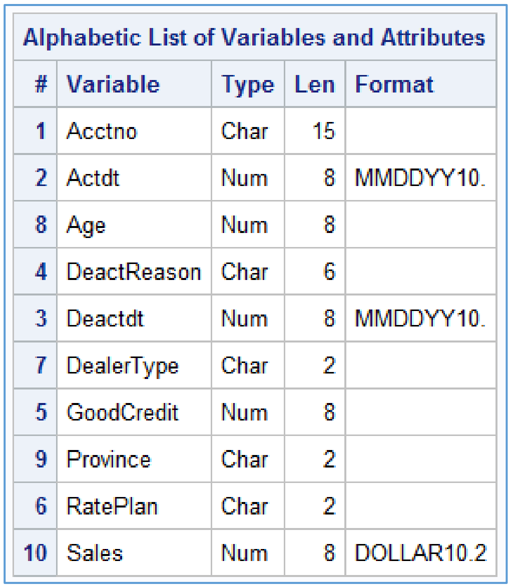
Categorical Variables	Numerical Variables
Acct No	Act dt
Deact Reason	Age
Dealer Type	Deact Dt
Province	Good Credit
Rate Plan	Sales

`



The CONTENTS Procedure

Data Set Name	NANDINI.TELCM	Observations	102255
Member Type	DATA	Variables	10
Engine	V9	Indexes	0
Created	2024-10-08 12:25:00	Observation Length	72
Last Modified	2024-10-08 12:25:00	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		



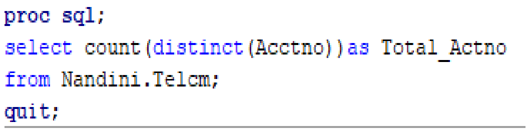
Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format
1	Acctno	Char	15	
2	Actdt	Num	8	MMDDYY10.
8	Age	Num	8	
4	DeactReason	Char	6	
3	Deactdt	Num	8	MMDDYY10.
7	DealerType	Char	2	
5	GoodCredit	Num	8	
9	Province	Char	2	
6	RatePlan	Char	2	
10	Sales	Num	8	DOLLAR10.2

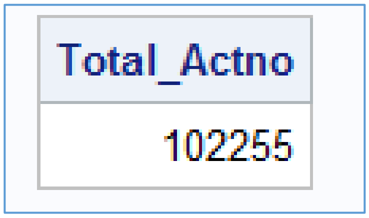
UNIVARIATE ANALYSIS- Categorical Variables

1. Acct No

- It’s the account no of the customer associated with the Telecom Company
- No Missing or duplicate account numbers found
- The Data Reveals the total of 102255 Customer account numbers.



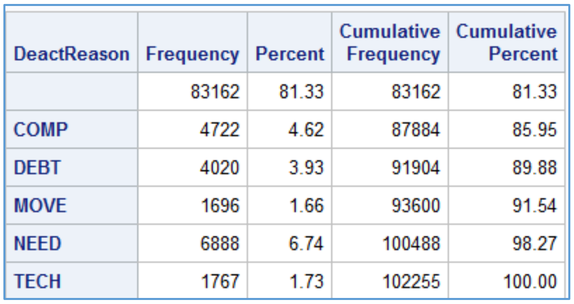
```
proc sql;
select count(distinct(Acctno))as Total_Actno
from Nandini.Telcm;
quit;
```



Total_Actno
102255

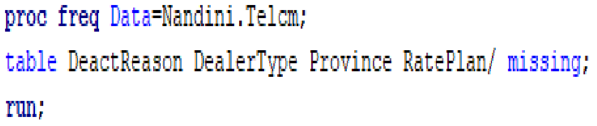
2 . Deact Reason

- The Analysis Focuses on distribution of the variable ‘Deact Reason’ within the dataset that means reason for service deactivation.
- Summarisation- Proc Freq
- Visualisation- Bar Chart
- The data reveals out of total 102255 observations 4722(4.62%) left due ‘COMP’ reason, 4020(3.93%) left due to ‘DEBT’ reason, 1696(1.66%)left due to ‘MOVE’ reason,6888(6.74%)left due to ‘NEED’reason,1767(1.73%) left due to ‘TECH’ reason. 83162(81.33%) reason is missing. This can be either active customer or those who have not mention reason to leave the service.

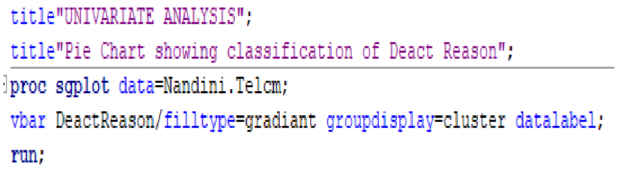


DeactReason	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	83162	81.33	83162	81.33
COMP	4722	4.62	87884	85.95
DEBT	4020	3.93	91904	89.88
MOVE	1696	1.66	93600	91.54
NEED	6888	6.74	100488	98.27
TECH	1767	1.73	102255	100.00

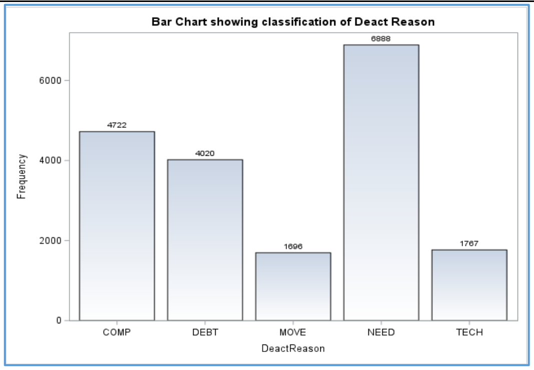
`



```
proc freq Data=Nandini.Telcm;
table DeactReason DealerType Province RatePlan/ missing;
run;
```

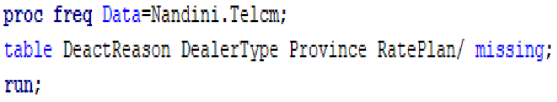


```
title"UNIVARIATE ANALYSIS";
title"Pie Chart showing classification of Deact Reason";
proc sgplot data=Nandini.Telcm;
vbar DeactReason/filltype=gradiant groupdisplay=cluster datalabel;
run;
```

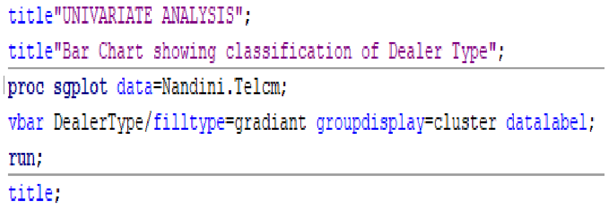


3.Dealer Type-

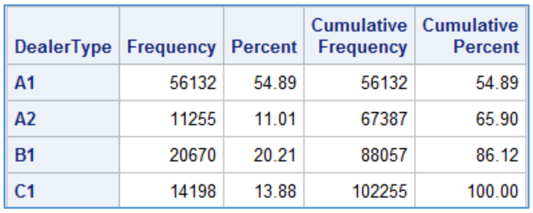
- The Analysis Focuses on distribution of the variable ‘Dealer Type’ within the dataset .
- Summarisation- Proc Freq
- Visualisation- Bar Chart
- The data reveals the distribution of 102255 observations categorized into 4 levels of Dealer type
    - A1- 56132- 54.89%
    - A2-11255-11.01%
    - B1-20670-20.21%
    - C1-14198-13.88%
- It indicates Dealer A1 has maximum customer base



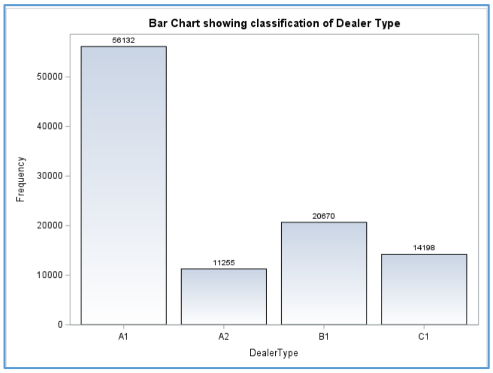
```
proc freq Data=Nandini.Telcm;
table DeactReason DealerType Province RatePlan/ missing;
run;
```



```
title"UNIVARIATE ANALYSIS";
title"Bar Chart showing classification of Dealer Type";
proc sgplot data=Nandini.Telcm;
vbar DealerType/filltype=gradiant groupdisplay=cluster datalabel;
run;
title;
```



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

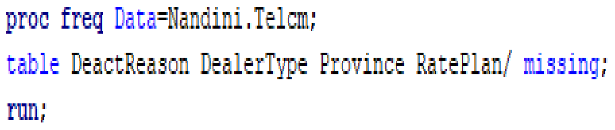


4. Province

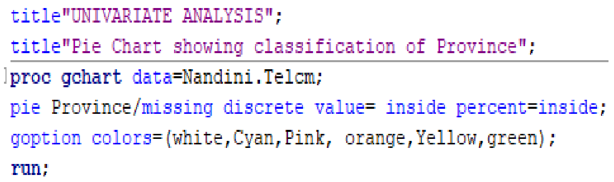
- The Analysis Focuses on distribution of the variable ’Province’ within the dataset
- Summarisation- Proc Freq
- Visualisation- Pie Chart
- The data reveals total 102255 observations which are grouped as follows-

`

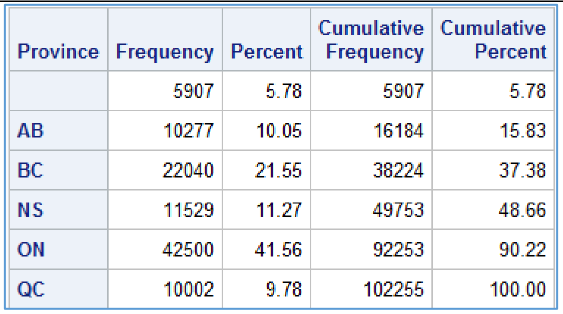
- AB- 10277 (10.05%)
- BC- 22040(21.55%)
- NS- 11529(11.27%)
- ON- 42500(41.56%)
- QC-10002(9.78%)
- No Province Info-5907(5.78%)
- Maximum Customer base is in ON. Minimum Customer Base in QC



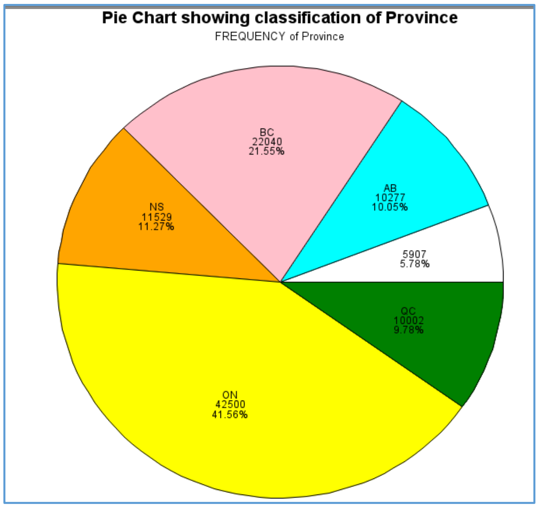
```
proc freq Data=Nandini.Telcm;
table DeactReason DealerType Province RatePlan/ missing;
run;
```



```
title"UNIVARIATE ANALYSIS";
title"Pie Chart showing classification of Province";
proc gchart data=Nandini.Telcm;
pie Province/missing discrete value= inside percent=inside;
goption colors=(white,Cyan,Pink, orange,Yellow,green);
run;
```



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

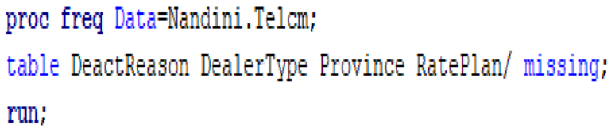


Pie Chart showing classification of Province

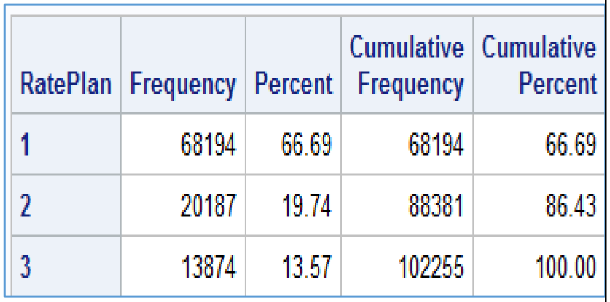
FREQUENCY of Province

5. Rate Plan

- The Analysis Focuses on distribution of the variable 'Rate Plan ' for the customer within the dataset
- Summarisation- Proc Freq
- Visualisation- Pie Chart
- The data reveals the distribution of 102255 observations categorized into 3 Types of Rate Plan
  - 1-68194-(66.69%)
  - 2-20187-(19.74%)
  - 3-13874-(13.57%)
- It indicates majority people -66.69% prefer Rate plan 1

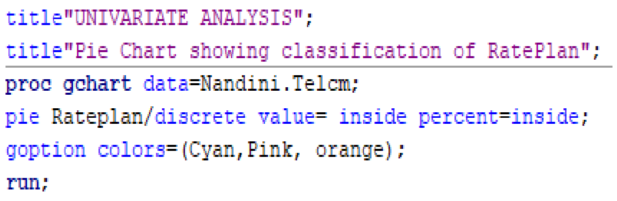


```
proc freq Data=Nandini.Telcm;
table DeactReason DealerType Province RatePlan/ missing;
run;
```

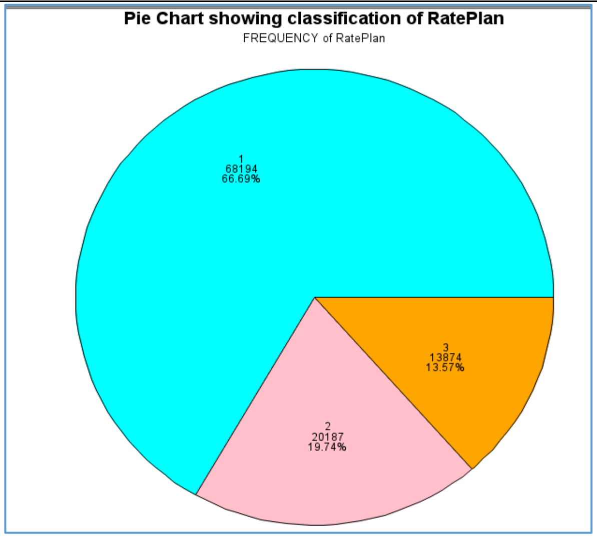


RatePlan	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	68194	66.69	68194	66.69
2	20187	19.74	88381	86.43
3	13874	13.57	102255	100.00

`

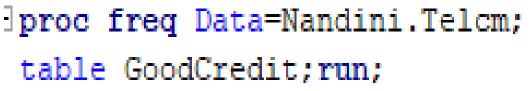


```
title"UNIVARIATE ANALYSIS";
title"Pie Chart showing classification of RatePlan";
proc gchart data=Nandini.Telcm;
pie Rateplan/discrete value= inside percent=inside;
goption colors=(Cyan,Pink, orange);
run;
```

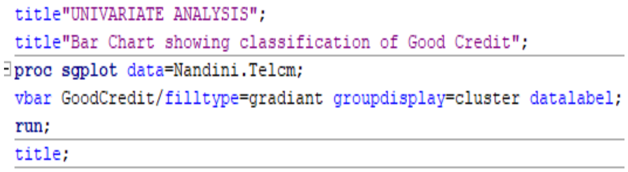


6. Good Credit

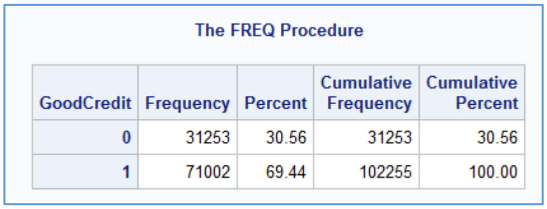
- The Analysis Focuses on distribution of the variable ‘Good Credit’ within the dataset that specifies if customer has Good Credit or not
- It’s a Numerical variable .But being descrete Numerical variable with only 2 levels we can treat this as categorical.
- Summarisation- Proc Freq
- Visualisation- Bar Chart
- The data reveals the distribution of 102255 observations categorized into 2 Levels of Good Credit
  - 0 (No Good Credit)-31253-30.56%
  - 1-(Has Good Credit)-71002-69.44%
- It indicates majority people -69.44% have Good Credit



```
proc freq Data=Nandini.Telcm;
 table GoodCredit;run;
```

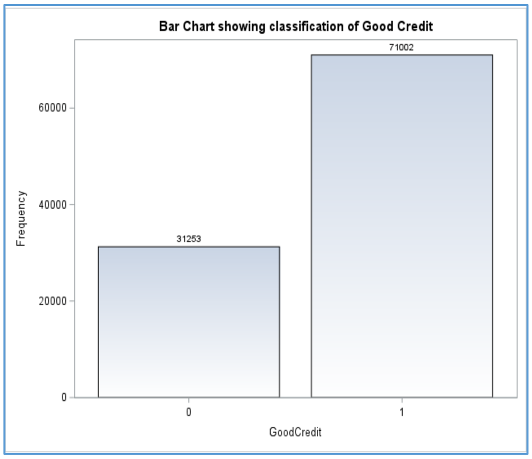


```
title"UNIVARIATE ANALYSIS";
title"Bar Chart showing classification of Good Credit";
proc sgplot data=Nandini.Telcm;
vbar GoodCredit/filltype=gradiant groupdisplay=cluster datalabel;
run;
title;
```



The FREQ Procedure

GoodCredit	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	31253	30.56	31253	30.56
1	71002	69.44	102255	100.00

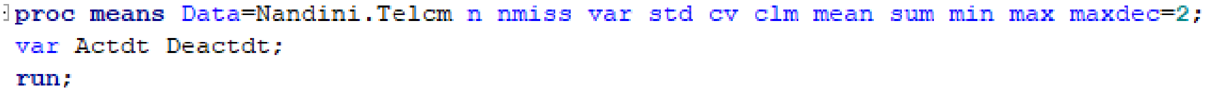


`

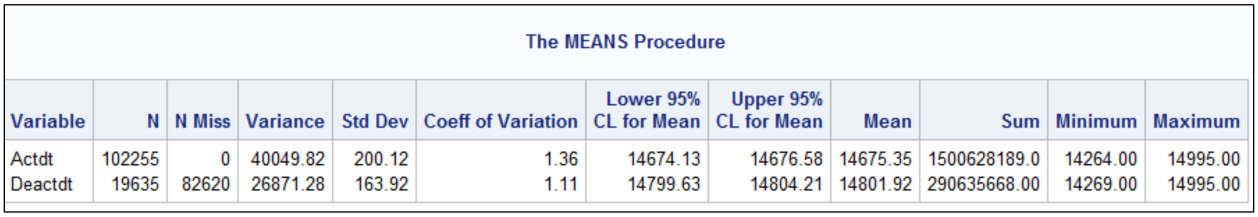
UNIVARIATE ANALYSIS- Numerical Variables

1 .Activation and Deactivation Dates

- The Analysis Focuses on distribution of the continuous variable ‘Actdt;(Account activation Date) and ‘Deactdt’(Account Deactivation Date).
- Summarisation- Proc Means,Proc Univariate
- Visualisation- QQ plot
- Summary estimate is as shown in the image-‘The Means Procedure’
- Latest Activation Date is 20th January 2001 and Latest Deactivation Date is also 20th January 2001.
- Kolmogorov-Smirnov Test of Normality shows that both variables are not normally distributed
- QQ plot displays that distribution is not uniform.

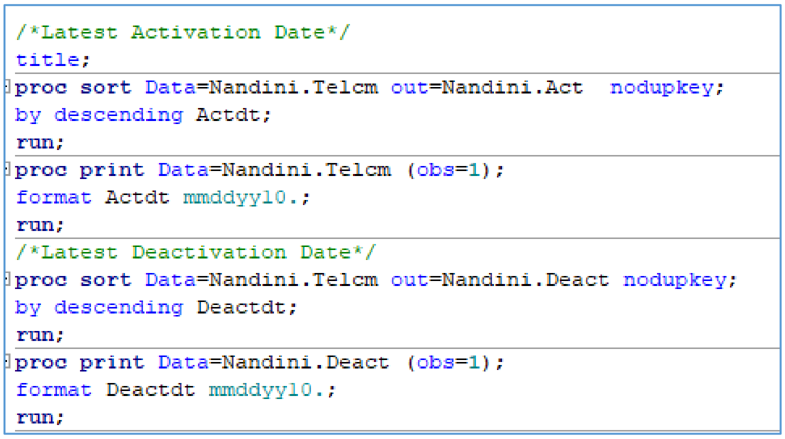


```
proc means Data=Nandini.Telcm n nmiss var std cv clm mean sum min max maxdec=2;
var Actdt Deactdt;
run;
```

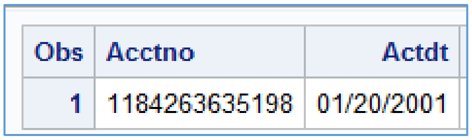


The MEANS Procedure

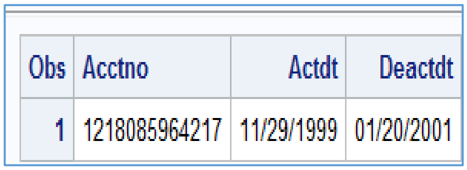
Variable	N	N Miss	Variance	Std Dev	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean	Mean	Sum	Minimum	Maximum
Actdt	102255	0	40049.82	200.12	1.36	14674.13	14676.58	14675.35	1500628189.0	14264.00	14995.00
Deactdt	19635	82620	26871.28	163.92	1.11	14799.63	14804.21	14801.92	290635668.00	14269.00	14995.00



```
/*Latest Activation Date*/
title;
proc sort Data=Nandini.Telcm out=Nandini.Act  nodupkey;
by descending Actdt;
run;
proc print Data=Nandini.Telcm (obs=1);
format Actdt mmddyy10.;
run;
/*Latest Deactivation Date*/
proc sort Data=Nandini.Telcm out=Nandini.Deact nodupkey;
by descending Deactdt;
run;
proc print Data=Nandini.Deact (obs=1);
format Deactdt mmddyy10.;
run;
```

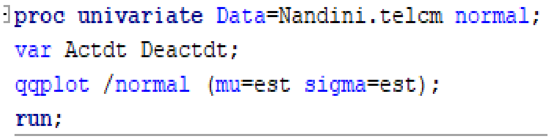


Obs	Acctno	Actdt
1	1184263635198	01/20/2001

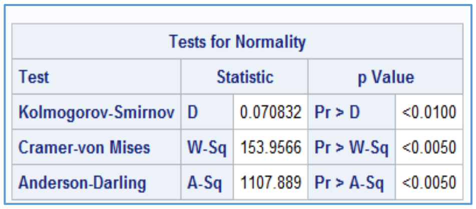


Obs	Acctno	Actdt	Deactdt
1	1218085964217	11/29/1999	01/20/2001

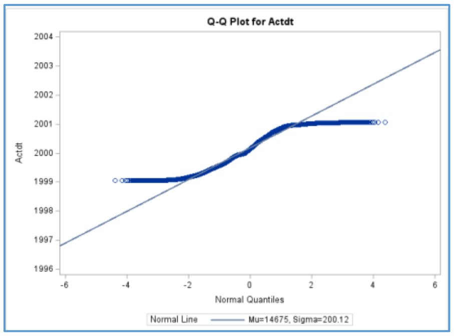
Activation and Deactivation Dates Distribution

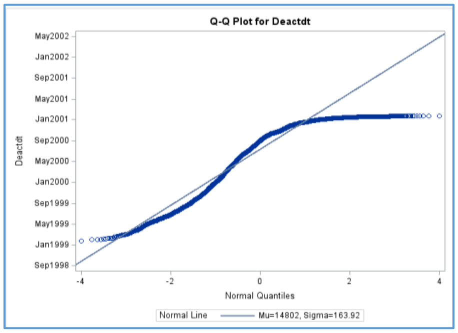


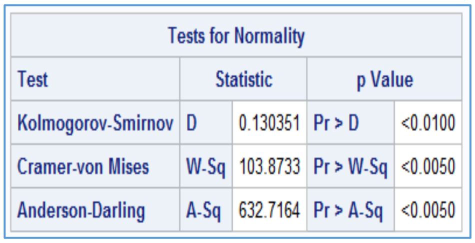
```
proc univariate Data=Nandini.telcm normal;
var Actdt Deactdt;
qqplot /normal (mu=est sigma=est);
run;
```



Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.070832	Pr > D	<0.0100
Cramer-von Mises	W-Sq	153.9566	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1107.889	Pr > A-Sq	<0.0050



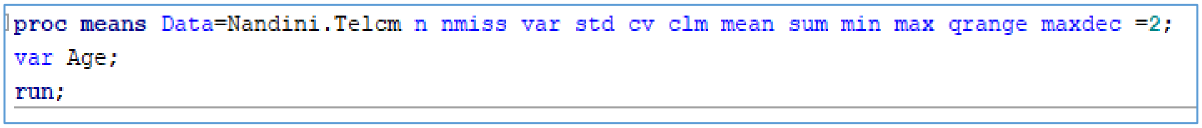




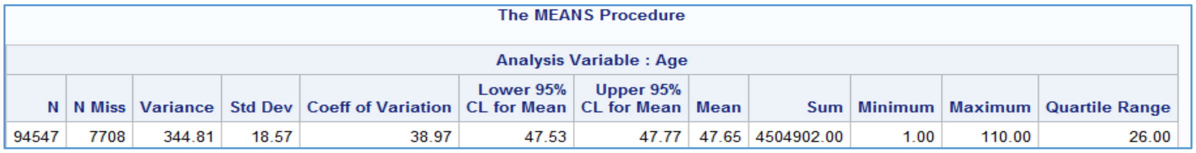
Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.130351	Pr > D	<0.0100
Cramer-von Mises	W-Sq	103.8733	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	632.7164	Pr > A-Sq	<0.0050

2.Age

- The Analysis Focuses on distribution of the continuous variable ‘Age’ within the dataset
- Summarisation- Proc Means,Proc Univariate
- Visualisation- QQplot
- The data reveals the Age of total 94547 out of 102255 people with following summary estimates
  - Mean- 47.65 years
  - Standard Deviation- 18.57
  - Minimum Age-1 Year
  - Maximum Age-110 Years
  - Mode-48 year- This Age appears more often than others
  - Skewness-(0.04) This Means it is Positive or slightly Right skewed
  - Kurtosis-(-0.40) This Means It is Platykurtic .i.e. Has Negative Kurtosis . Peak is Flatter than normal and Tails are longer than normal
  - Presence of outliers. Maximum value is more than Q3+3IQR(upper outer Fence)
  - P value of Kolmogorov Smirnov Test of Normality is less than 0.05. So we reject Null Hypothesis of Normality and conclude Age is not normally distributed .However as per CLT we can assume Age is normally distributed



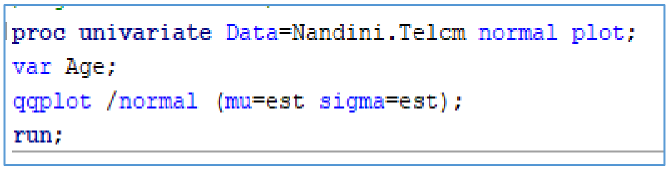
```
proc means Data=Nandini.Telcm n nmiss var std cv clm mean sum min max qrange maxdec =2;
var Age;
run;
```



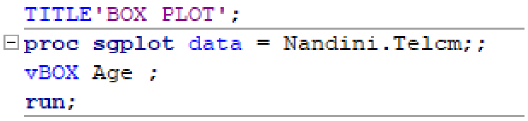
The MEANS Procedure

Analysis Variable : Age

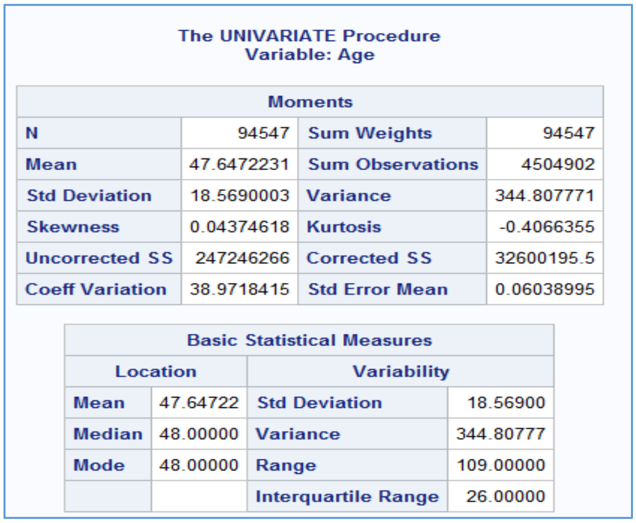
N	N Miss	Variance	Std Dev	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean	Mean	Sum	Minimum	Maximum	Quartile Range
94547	7708	344.81	18.57	38.97	47.53	47.77	47.65	4504902.00	1.00	110.00	26.00



```
proc univariate Data=Nandini.Telcm normal plot;
var Age;
qqplot /normal (mu=est sigma=est);
run;
```



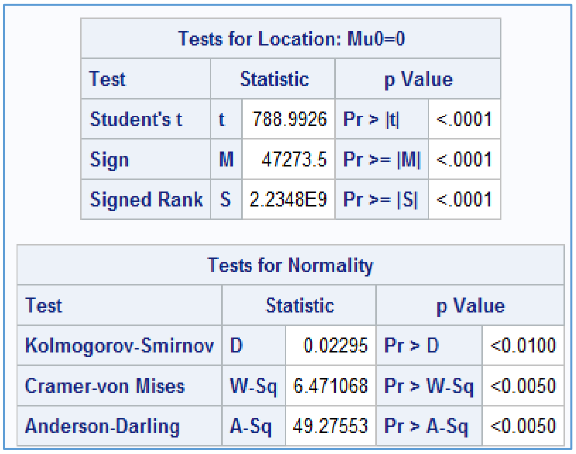
```
TITLE'BOX PLOT';
proc sgplot data = Nandini.Telcm;;
vBOX Age ;
run;
```



The UNIVARIATE Procedure
Variable: Age

Moments			
N	94547	Sum Weights	94547
Mean	47.6472231	Sum Observations	4504902
Std Deviation	18.5690003	Variance	344.807771
Skewness	0.04374618	Kurtosis	-0.4066355
Uncorrected SS	247246266	Corrected SS	32600195.5
Coeff Variation	38.9718415	Std Error Mean	0.06038995

Basic Statistical Measures			
Location		Variability	
Mean	47.64722	Std Deviation	18.56900
Median	48.00000	Variance	344.80777
Mode	48.00000	Range	109.00000
		Interquartile Range	26.00000



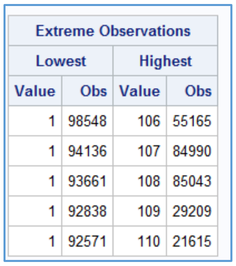
Tests for Location: Mu0=0

Test	Statistic		p Value	
Student's t	t	788.9926	Pr > |t|	<.0001
Sign	M	47273.5	Pr >= |M|	<.0001
Signed Rank	S	2.2348E9	Pr >= |S|	<.0001

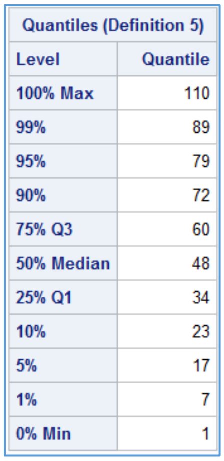
Tests for Normality

Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.02295	Pr > D	<0.0100
Cramer-von Mises	W-Sq	6.471068	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	49.27553	Pr > A-Sq	<0.0050

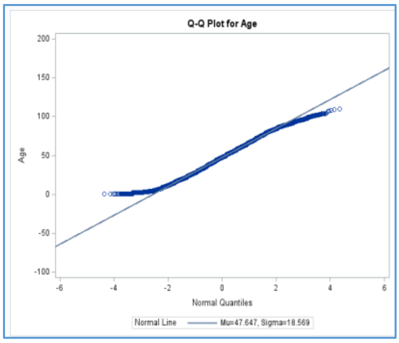
`

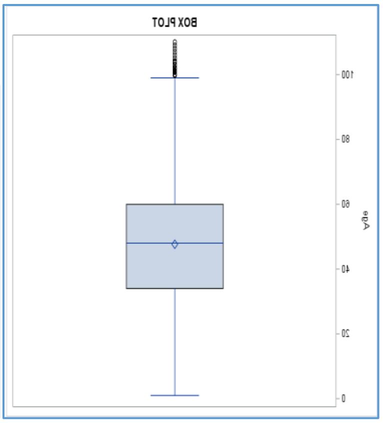


Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1	98548	106	55165
1	94136	107	84990
1	93661	108	85043
1	92838	109	29209
1	92571	110	21615



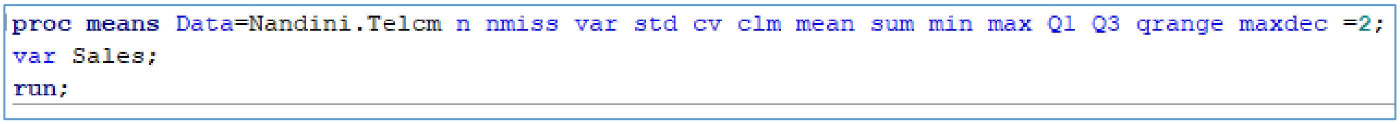
Quantiles (Definition 5)	
Level	Quantile
100% Max	110
99%	89
95%	79
90%	72
75% Q3	60
50% Median	48
25% Q1	34
10%	23
5%	17
1%	7
0% Min	1



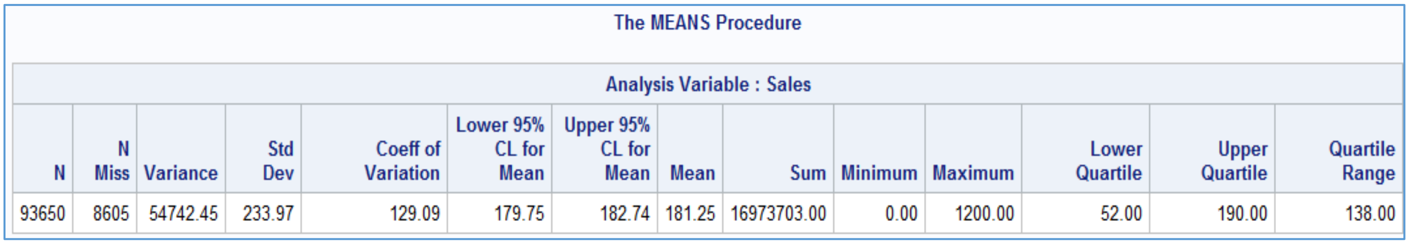


3. Sales

- The Analysis Focuses on distribution of the continuous variable ‘Sales’ within the dataset
- Summarisation- Proc Means,Proc Univariate
- Visualisation- QQplot
- The data reveals the Age of total 93650 out of 102255 people with following summary estimates
  - Mean- $181.25
  - Standard Deviation- 233.97
  - Minimum Sale-0
  - Maximum Sale-$1200
  - Mode-$92 This Amount appears more often than others
  - Skewness-(2.366) This Means it is Positive or Right skewed
  - Kurtosis-(5.28) This Means It is Leptokurtic .i.e. Has Positive Kurtosis . Peak is sharper than normal and Tails are heavier than normal
  - Presence of outliers. Maximum value is more than Q3+3IQR(upper outer Fence)
  - P value of Kolmogorov Smirnov Test of Normality is less than 0.05. So we reject Null Hypothesis of Normality and conclude Sales is not normally distributed .However as per CLT we can assume Age is normally distributed



```
proc means Data=Nandini.Telcm n nmiss var std cv clm mean sum min max Q1 Q3 qrange maxdec =2;
var Sales;
run;
```

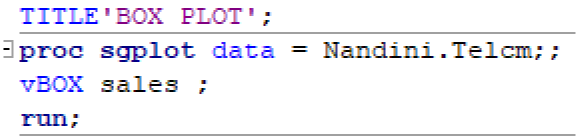


The MEANS Procedure

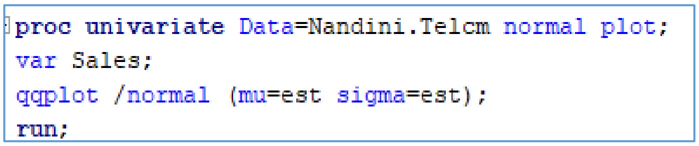
Analysis Variable : Sales

N	N Miss	Variance	Std Dev	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean	Mean	Sum	Minimum	Maximum	Lower Quartile	Upper Quartile	Quartile Range
93650	8605	54742.45	233.97	129.09	179.75	182.74	181.25	16973703.00	0.00	1200.00	52.00	190.00	138.00

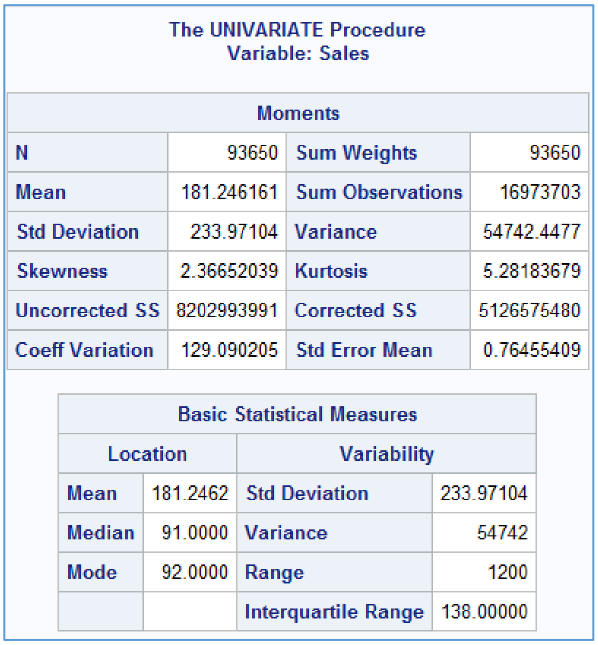
`



```
TITLE'BOX PLOT';
proc sgplot data = Nandini.Telcm;;
vBOX sales ;
run;
```



```
proc univariate Data=Nandini.Telcm normal plot;
var Sales;
qqplot /normal (mu=est sigma=est);
run;
```

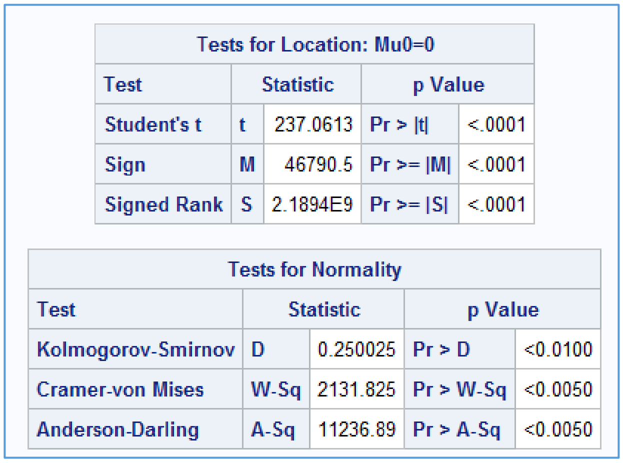


The UNIVARIATE Procedure
Variable: Sales

Moments			
N	93650	Sum Weights	93650
Mean	181.246161	Sum Observations	16973703
Std Deviation	233.97104	Variance	54742.4477
Skewness	2.36652039	Kurtosis	5.28183679
Uncorrected SS	8202993991	Corrected SS	5126575480
Coeff Variation	129.090205	Std Error Mean	0.76455409

Basic Statistical Measures

Location		Variability	
Mean	181.2462	Std Deviation	233.97104
Median	91.0000	Variance	54742
Mode	92.0000	Range	1200
		Interquartile Range	138.00000

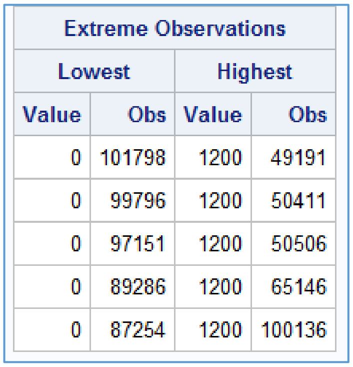


Tests for Location: Mu0=0

Test	Statistic		p Value	
Student's t	t	237.0613	Pr > |t|	<.0001
Sign	M	46790.5	Pr >= |M|	<.0001
Signed Rank	S	2.1894E9	Pr >= |S|	<.0001

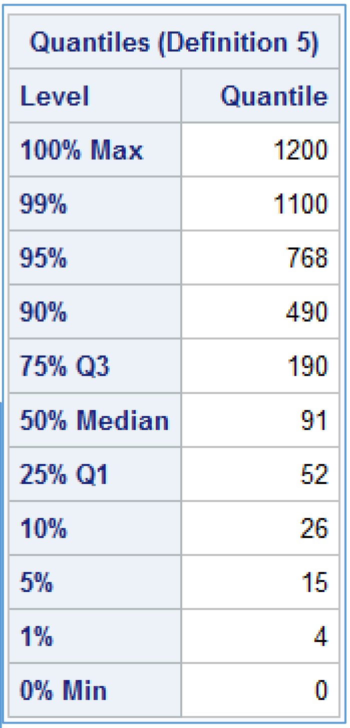
Tests for Normality

Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.250025	Pr > D	<0.0100
Cramer-von Mises	W-Sq	2131.825	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	11236.89	Pr > A-Sq	<0.0050



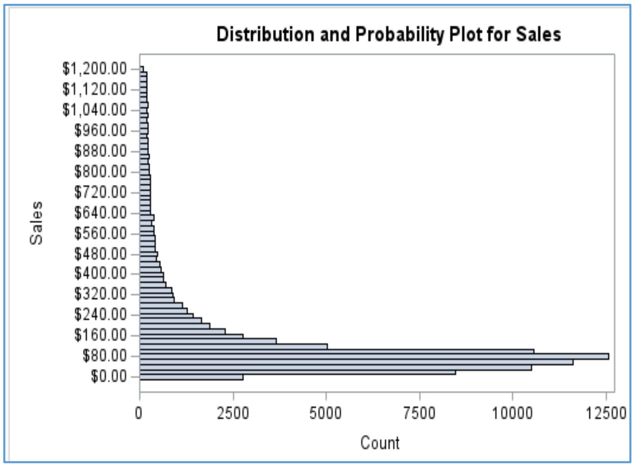
Extreme Observations

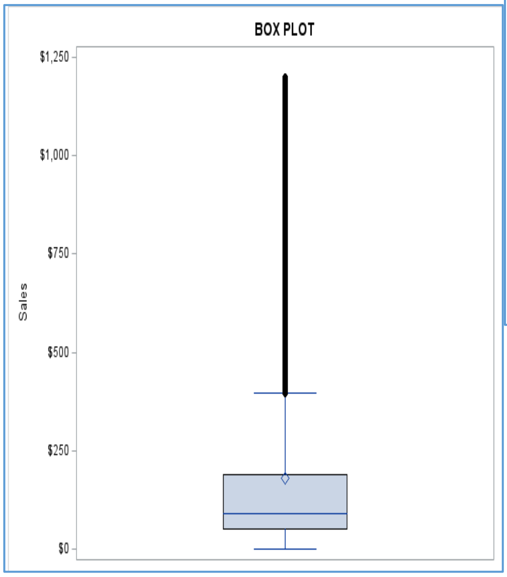
Lowest		Highest	
Value	Obs	Value	Obs
0	101798	1200	49191
0	99796	1200	50411
0	97151	1200	50506
0	89286	1200	65146
0	87254	1200	100136



Quantiles (Definition 5)

Level	Quantile
100% Max	1200
99%	1100
95%	768
90%	490
75% Q3	190
50% Median	91
25% Q1	52
10%	26
5%	15
1%	4
0% Min	0





`

1.2 What are the age and province distributions of active and deactivated customers? Cont vs categorical.

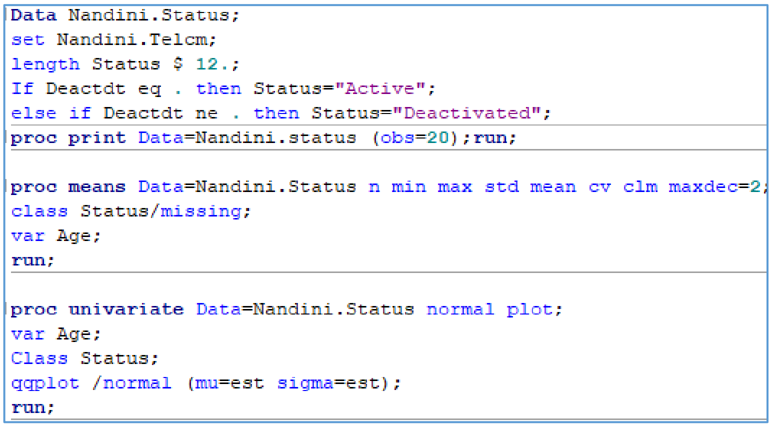
A) Age Distribution of Active and Deactivated Customers

H0- Means of Age is equal in both groups i.e Age equally distributed in both Active and Deactivated customers group.

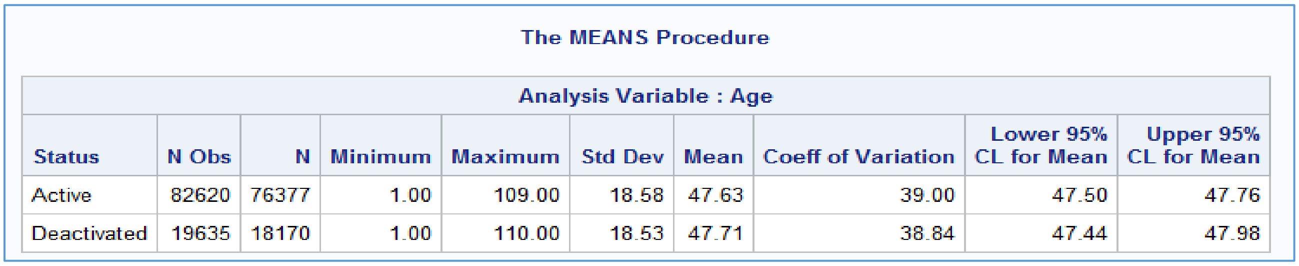
H1- Means of Age is not equal in both groups i.e Age not equally distributed in both Active and Deactivated customers group.

Approach- since Age is Numerical variable and Status -Active/Deactivated is a categorical Variable we will use:-

- For Summarisation- Proc Means
- For Normality-Proc Univariate
- For Visualisation- Grouped Box Plot
- For Independency- Proc T test



```
Data Nandini.Status;
set Nandini.Telcm;
length Status $ 12.;
If Deactdt eq . then Status="Active";
else if Deactdt ne . then Status="Deactivated";
proc print Data=Nandini.status (obs=20);run;

proc means Data=Nandini.Status n min max std mean cv clm maxdec=2;
class Status/missing;
var Age;
run;

proc univariate Data=Nandini.Status normal plot;
var Age;
Class Status;
qqplot /normal (mu=est sigma=est);
run;
```



The MEANS Procedure

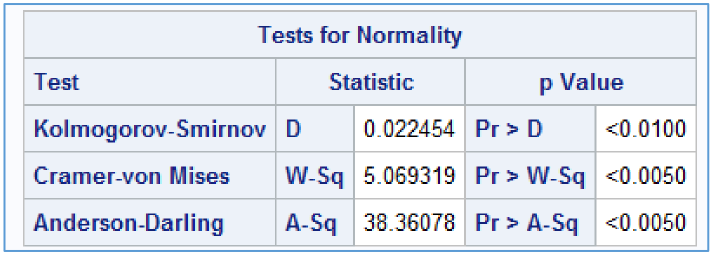
Analysis Variable : Age

Status	N Obs	N	Minimum	Maximum	Std Dev	Mean	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean
Active	82620	76377	1.00	109.00	18.58	47.63	39.00	47.50	47.76
Deactivated	19635	18170	1.00	110.00	18.53	47.71	38.84	47.44	47.98

- From the results of Means Procedure, we see that Mean and standard deviation of Active and Deactivated customers almost same or with minimum difference.
- To statistically prove this we will use normality test using Proc Univariate, Homoscedasticity test using Proc GLM and finally Proc Ttest to prove means are equal in both groups

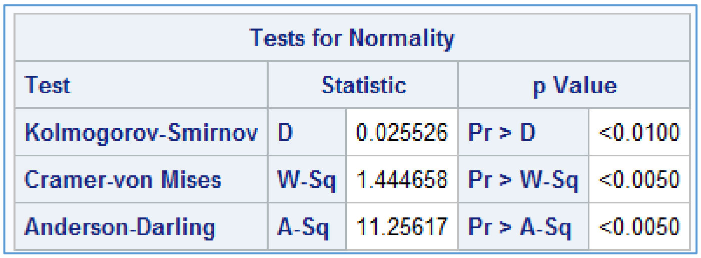
- Test of Normality-

H0- Age is normally distributed
H1- Age is not normally distributed



Tests for Normality

Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.022454	Pr > D	<0.0100
Cramer-von Mises	W-Sq	5.069319	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	38.36078	Pr > A-Sq	<0.0050



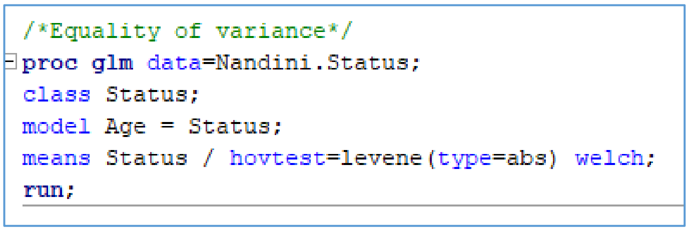
Tests for Normality

Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.025526	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.444658	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	11.25617	Pr > A-Sq	<0.0050

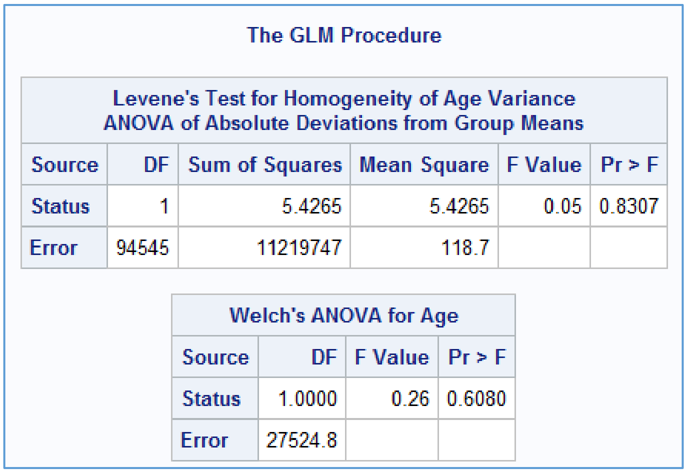
P value in Kolmogorov-Smirnov of Normality is less than 0.05 significance level.So we reject Null Hypothesis of normality and conclude that Age is not normally distributed ,However as per CLT, since sample size is more than>30, we can assume Age is normally distributed

- Test of Homoscedasticity for equality of variance-

H0- Variance of Age is equal in both Groups
H1- Variance of Age is not equal in both Groups



```
/*Equality of variance*/
proc glm data=Nandini.Status;
class Status;
model Age = Status;
means Status / hovtest=levene(type=abs) welch;
run;
```



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
Status	1	5.4265	5.4265	0.05	0.8307
Error	94545	11219747	118.7		

Welch's ANOVA for Age

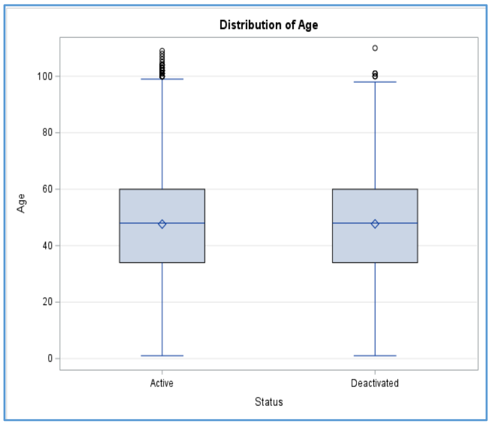
Source	DF	F Value	Pr > F
Status	1.0000	0.26	0.6080
Error	27524.8		

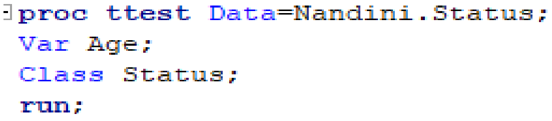
P value in Levene's Test for Equality of Variance 0.8307 is greater than 0.05 significance level.

Therefore, we fail to reject Null Hypothesis and Conclude the variance of Age is equal in both active and deactivated status

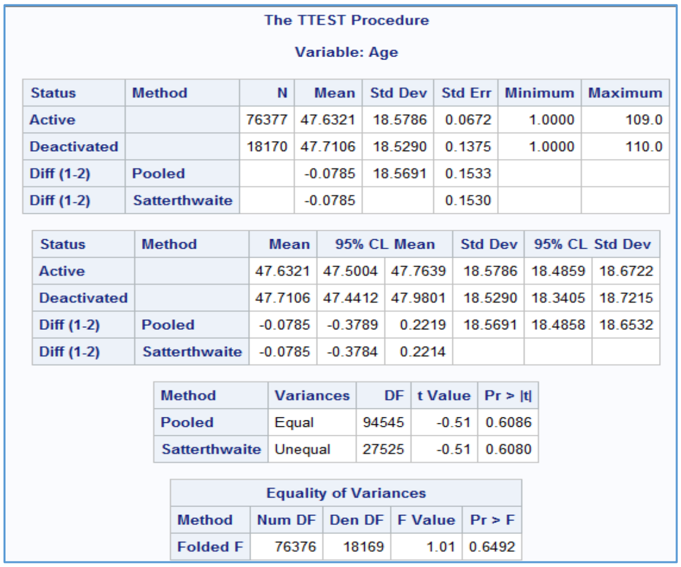
- Test of difference-

H0- Mean of Age is Equal in both Active and Deactivated groups
H1- Mean of Age is not Equal in both Active and Deactivated groups





```
proc ttest Data=Nandini.Status;
Var Age;
Class Status;
run;
```



The TTEST Procedure

Variable: Age

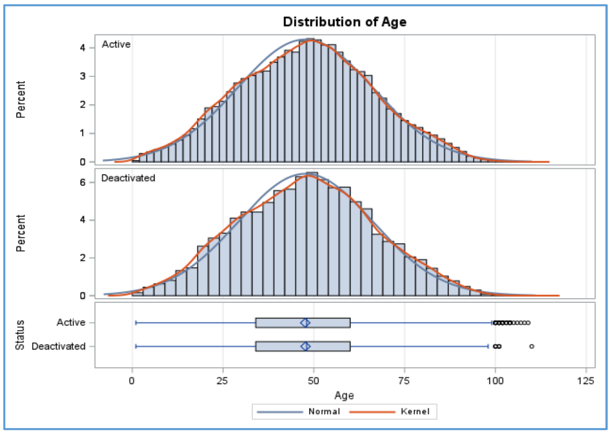
Status	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Active		76377	47.6321	18.5786	0.0672	1.0000	109.0
Deactivated		18170	47.7106	18.5290	0.1375	1.0000	110.0
Diff (1-2)	Pooled		-0.0785	18.5691	0.1533		
Diff (1-2)	Satterthwaite		-0.0785		0.1530		

Status	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Active		47.6321	47.5004	47.7639	18.5786	18.4859	18.6722
Deactivated		47.7106	47.4412	47.9801	18.5290	18.3405	18.7215
Diff (1-2)	Pooled	-0.0785	-0.3789	0.2219	18.5691	18.4858	18.6532
Diff (1-2)	Satterthwaite	-0.0785	-0.3784	0.2214			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	94545	-0.51	0.6086
Satterthwaite	Unequal	27525	-0.51	0.6080

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	76376	18169	1.01	0.6492



`

- Diagram –Grouped Box Plot- The Total Length of Boxplot or Interquartile range (Distance between Q1 and Q3) is very similar for Active and Deactivated groups. This is what we had expected. The groups have similar variance. We can see they line up and Diamond that represents the mean is aligned equally as well. There are more outliers In active group than deactivated group . This are extreme outliers (greater than Q3+3IQR- upper outer fence)
- Diagram Histogram and Density plot -The plots compare the distribution of Age with 2 categories of Active and Deactivated customers. From Histogram, we can see the distribution of Age for both categories of status very symmetric or bell shaped, so we can say the distribution is normal.

  .
- T- Test-

  H0- Means of Age is equal in both groups i.e Age equally distributed in both Active and Deactivated customers group.

  H1- Means of Age is not equal in both groups i.e Age not equally distributed in both Active and Deactivated customers group.

- The First table contain the valid sample size, mean, standard deviation, min and max.

  In this case the mean of Age- has very minimum difference between two groups. We need to check this difference is statistically or happens by chance. For this we need to check second table that has Two other Test Pooled and Satterwaite .
- Pooled Test assumes that both groups have the same variance in Age whereas Satterwaite test does not make this assumption.
- It can be done by checking the last table that it is folded f test:
  - Folded f test hypothesis:H0 is : Variance are equal
  - H1 is- Variance are not equal
- p value for f test it is 0.6492 > 0.05 so we fail to reject null hypothesis and we will say the variance are equal. Therefore, we will the see results from Pooled Test.

Conclusion-

Since P values in Pooled Test 0.6492 >0.05 ,

- we fail to reject the Null Hypothesis at 5% Significance Level
- Average Age of Active customers is equal to average age of Deactivated customers

B) Province Distribution of Active and Deactivated Customers

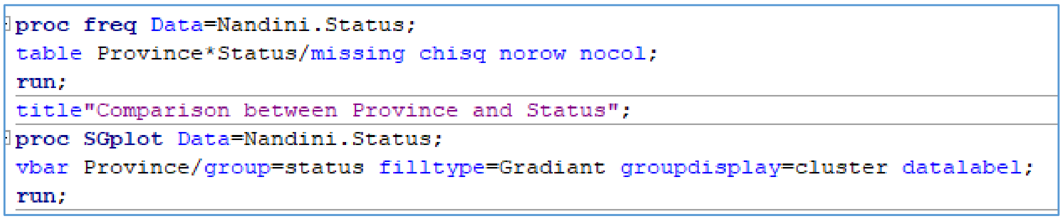
Ho- Null Hypothesis-There is no association between Province and Status(Active /Deactivated)

H1- Alternate Hypothesis-There is association between Province and Status(Active /Deactivated)

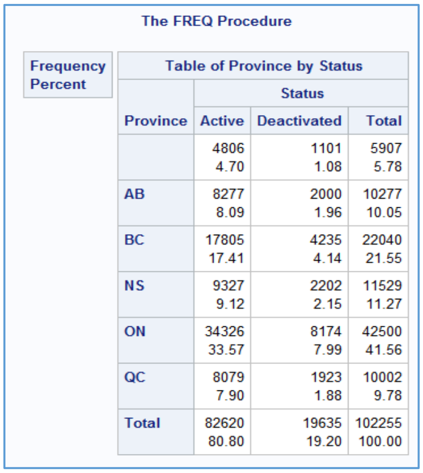
Approach- since both are categorical Variables we will use:-

- For Summarisation- Frequency Table
- For Visualisation- Grouped Bar Chart
- For Independency- Chi sq test

`



```
proc freq Data=Nandini.Status;
table Province*Status/missing chisq norow nocol;
run;
title"Comparison between Province and Status";
proc SGplot Data=Nandini.Status;
vbar Province/group=status filltype=Gradiant groupdisplay=cluster datalabel;
run;
```

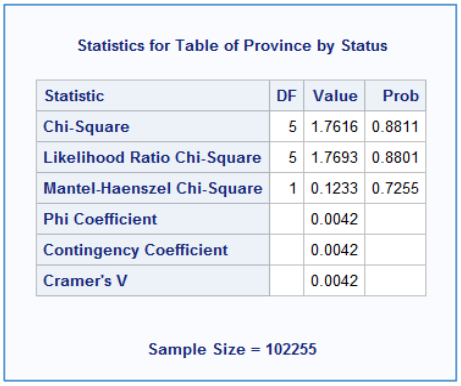


The FREQ Procedure

Frequency
Percent

Table of Province by Status

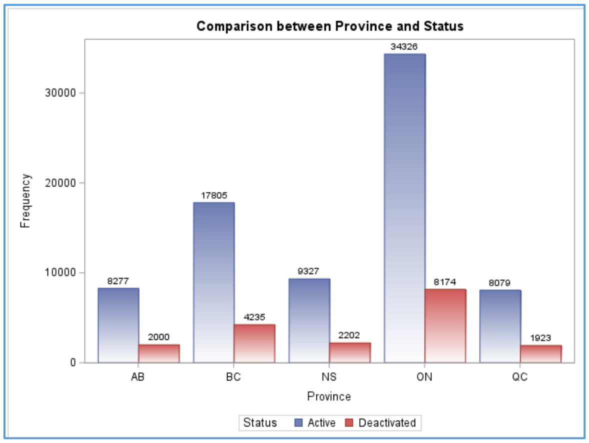
Province	Status Active	 Deactivated	 Total
	4806 4.70	1101 1.08	5907 5.78
AB	8277 8.09	2000 1.96	10277 10.05
BC	17805 17.41	4235 4.14	22040 21.55
NS	9327 9.12	2202 2.15	11529 11.27
ON	34326 33.57	8174 7.99	42500 41.56
QC	8079 7.90	1923 1.88	10002 9.78
Total	82620 80.80	19635 19.20	102255 100.00



Statistics for Table of Province by Status

Statistic	DF	Value	Prob
Chi-Square	5	1.7616	0.8811
Likelihood Ratio Chi-Square	5	1.7693	0.8801
Mantel-Haenszel Chi-Square	1	0.1233	0.7255
Phi Coefficient		0.0042	
Contingency Coefficient		0.0042	
Cramer's V		0.0042	

Sample Size = 102255



- Grouped Bar Chart shows the distribution of Province with Active and Deactivated Customers.
- From first table, we can say out of 102255
  AB has 8.09% of Active Customer and 1.96% of Deactivated Customers
  BC has 17.41% of Active Customer and 4.14% of Deactivated Customers
  NS has 9.12% of Active Customer and 2.15% of Deactivated Customers
  ON has 33.57% of Active Customer and 7.99% of Deactivated Customers
  QC has 7.9% of Active Customer and 1.88% of Deactivated Customers
  No Province info- 4.7% of Active Customer and 1.08% of Deactivated Customers
- Ontario has maximum number of Active and Deactivated customers. Whereas Quebec has Minimum Active and Deactivated Customers
- Second table shows chi-square P value is0.8811 >0.05 and Cramer's V is 0.0042, which means there is no association between Province and Status

Conclusion-

- ➢ Since P values >0.05 we do not have enough evidence l to reject the Null Hypothesis at 5% Significance Level
- ➢ We accept There is no statistically significant association between Province and Active or Deactivated Customers
- ➢ In other words, Province and Status are Independent of each other.

`

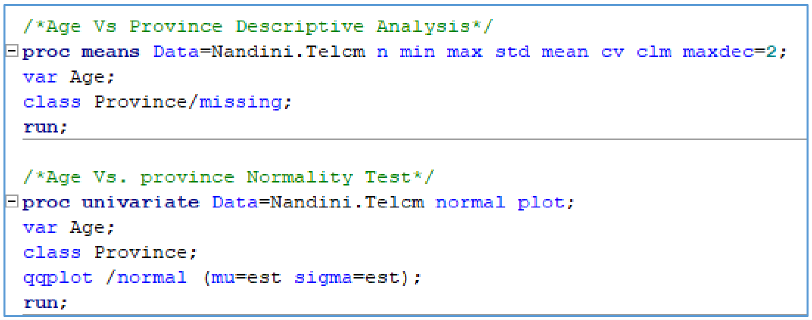
C) Age Distribution Across Provinces

H0- Means of Age is equal in all Provinces i.e Age equally distributed in all Provinces

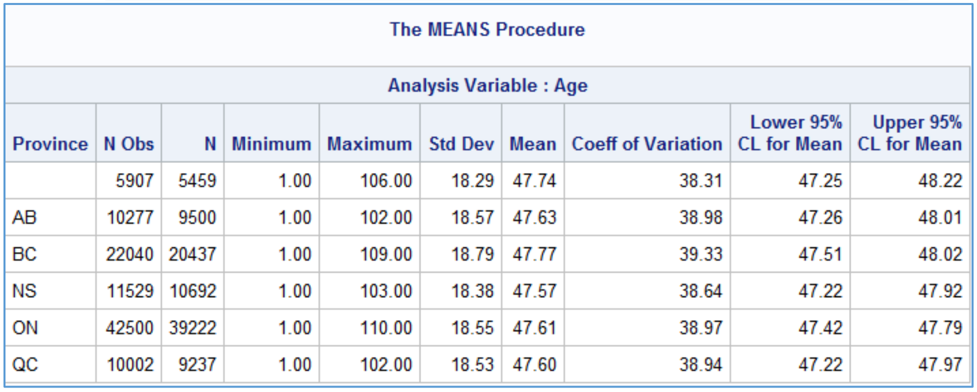
H1- Means of Age is not equal in all Provinces i.e Age is not equally distributed in all Provinces.

Approach- since Age is Numerical variable and Province is a categorical Variable with more than 2 levels we will use:-

D) For Summarisation- Proc Means
E) For Normality-Proc Univariate
F) For Visualisation- Grouped Box Plot
G) For Independency- Proc Anova



```
/*Age Vs Province Descriptive Analysis*/
proc means Data=Nandini.Telcm n min max std mean cv clm maxdec=2;
var Age;
class Province/missing;
run;

/*Age Vs. province Normality Test*/
proc univariate Data=Nandini.Telcm normal plot;
var Age;
class Province;
qqplot /normal (mu=est sigma=est);
run;
```



The MEANS Procedure

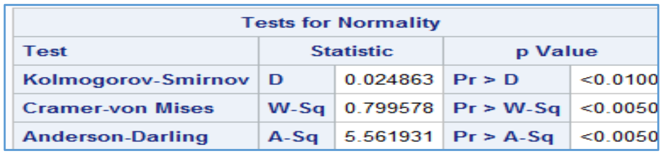
Analysis Variable : Age

Province	N Obs	N	Minimum	Maximum	Std Dev	Mean	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean
	5907	5459	1.00	106.00	18.29	47.74	38.31	47.25	48.22
AB	10277	9500	1.00	102.00	18.57	47.63	38.98	47.26	48.01
BC	22040	20437	1.00	109.00	18.79	47.77	39.33	47.51	48.02
NS	11529	10692	1.00	103.00	18.38	47.57	38.64	47.22	47.92
ON	42500	39222	1.00	110.00	18.55	47.61	38.97	47.42	47.79
QC	10002	9237	1.00	102.00	18.53	47.60	38.94	47.22	47.97

- From the results of Means Procedure, we see that Mean and standard deviation of Age is all provinces almost same or with minimum difference.
- To statistically prove this we will use normality test using Proc Univariate, Homoscedasticity test using Proc GLM and finally Proc Anova to prove means are equal in all groups

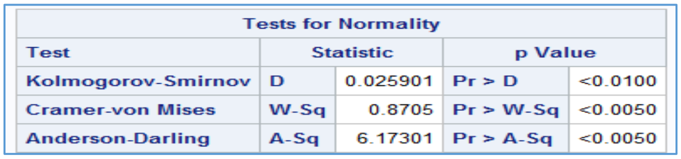
- Test of Normality-

H0- Age is normally distributed
H1- Age is not normally distributed



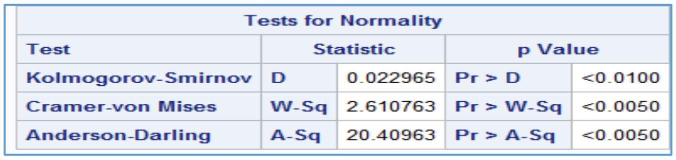
Tests for Normality

Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.024863	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.799578	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	5.561931	Pr > A-Sq	<0.0050



Tests for Normality

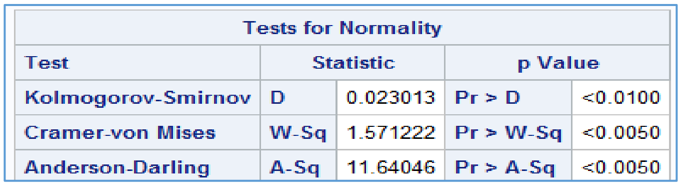
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.025901	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.8705	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	6.17301	Pr > A-Sq	<0.0050



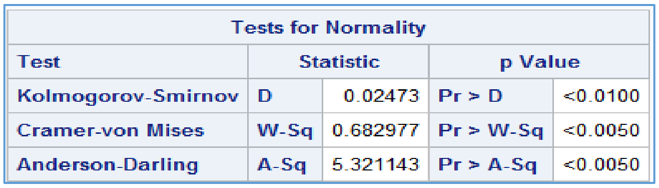
Tests for Normality

Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.022965	Pr > D	<0.0100
Cramer-von Mises	W-Sq	2.610763	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	20.40963	Pr > A-Sq	<0.0050

`



Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.023013	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.571222	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	11.64046	Pr > A-Sq	<0.0050

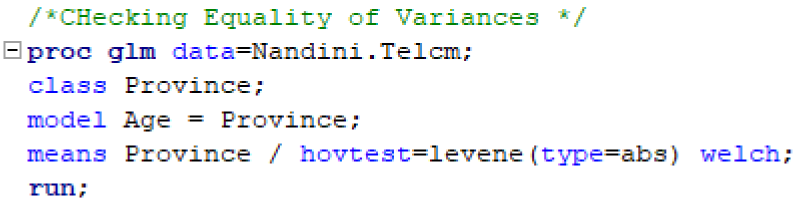


Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.02473	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.682977	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	5.321143	Pr > A-Sq	<0.0050

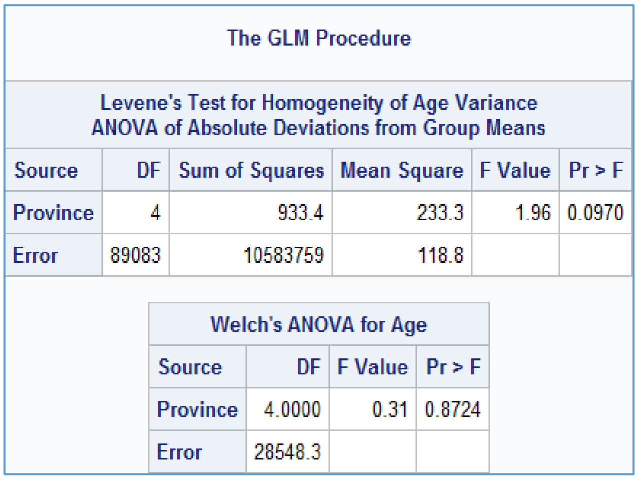
P value in Kolmogorov-Smirnov of Normality is less than 0.05 significance level for all provinces. So we reject Null Hypothesis of normality and conclude that Age is not normally distributed , However as per CLT, since sample size is more than>30, we can assume Age is normally distributed

- Test of Homoscedasticity for equality of variance-

H0- Variance of Age is equal in all Groups
H1- Variance of Age is equal in a;; Groups



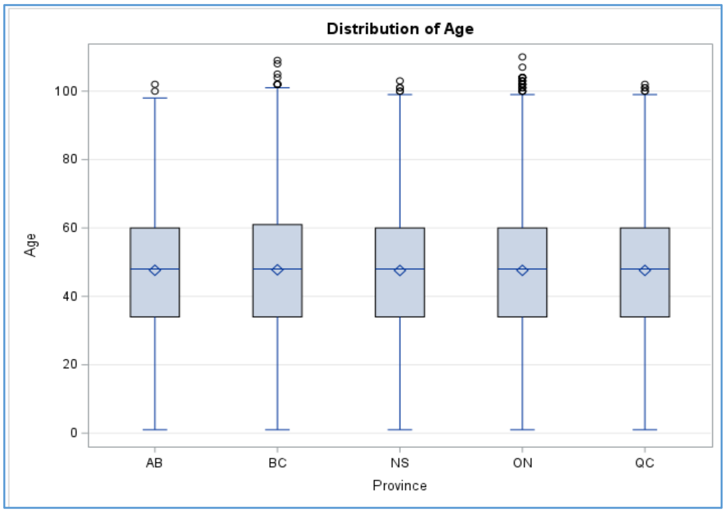
```
/*CHecking Equality of Variances */
proc glm data=Nandini.Telcm;
class Province;
model Age = Province;
means Province / hovtest=levene(type=abs) welch;
run;
```



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
Province	4	933.4	233.3	1.96	0.0970
Error	89083	10583759	118.8		

Welch's ANOVA for Age			
Source	DF	F Value	Pr > F
Province	4.0000	0.31	0.8724
Error	28548.3		

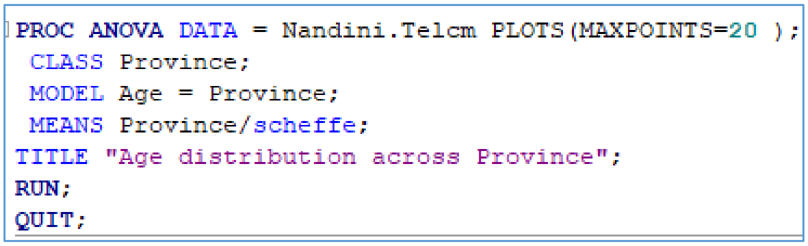


P value in Levene's Test for Equality of Variance 0.0970 is greater than 0.05 significance level.

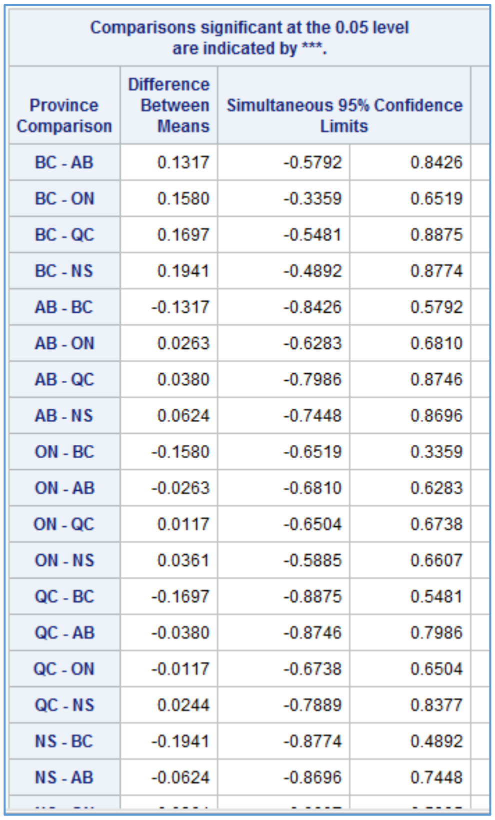
Therefore, we fail to reject Null Hypothesis and Conclude the variance of Age is equal in all Provinces

- Test of difference-
  H0- Mean of Age is Equal in all Provinces
  H1- Mean of Age is not Equal in all Provinces

`

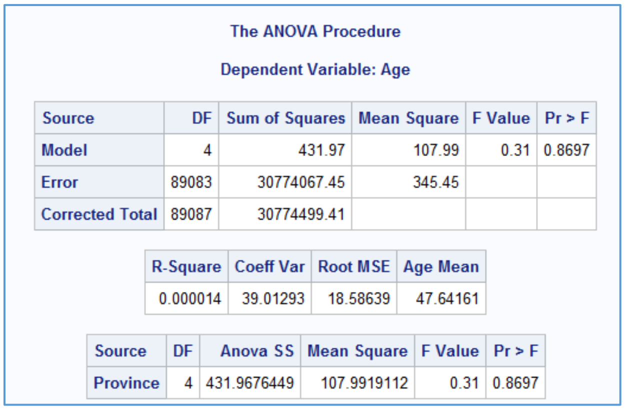


```
PROC ANOVA DATA = Nandini.Telcm PLOTS(MAXPOINTS=20 );
 CLASS Province;
 MODEL Age = Province;
 MEANS Province/scheffe;
TITLE "Age distribution across Province";
RUN;
QUIT;
```



Comparisons significant at the 0.05 level are indicated by ***.

Province Comparison	Difference Between Means	Simultaneous 95% Confidence Limits	
BC - AB	0.1317	-0.5792	0.8426
BC - ON	0.1580	-0.3359	0.6519
BC - QC	0.1697	-0.5481	0.8875
BC - NS	0.1941	-0.4892	0.8774
AB - BC	-0.1317	-0.8426	0.5792
AB - ON	0.0263	-0.6283	0.6810
AB - QC	0.0380	-0.7986	0.8746
AB - NS	0.0624	-0.7448	0.8696
ON - BC	-0.1580	-0.6519	0.3359
ON - AB	-0.0263	-0.6810	0.6283
ON - QC	0.0117	-0.6504	0.6738
ON - NS	0.0361	-0.5885	0.6607
QC - BC	-0.1697	-0.8875	0.5481
QC - AB	-0.0380	-0.8746	0.7986
QC - ON	-0.0117	-0.6738	0.6504
QC - NS	0.0244	-0.7889	0.8377
NS - BC	-0.1941	-0.8774	0.4892
NS - AB	-0.0624	-0.8696	0.7448



The ANOVA Procedure

Dependent Variable: Age

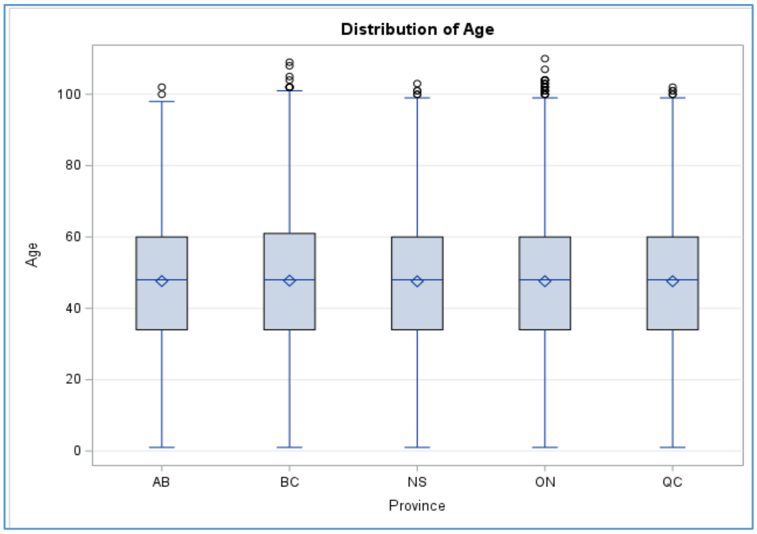
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	431.97	107.99	0.31	0.8697
Error	89083	30774067.45	345.45		
Corrected Total	89087	30774499.41			

R-Square	Coeff Var	Root MSE	Age Mean
0.000014	39.01293	18.58639	47.64161

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Province	4	431.9676449	107.9919112	0.31	0.8697

- F value: the overall f statistic is calculated by using mean square model/mean square error,
  107.99/345.45 =0.31
  F value: It is the ratio of mean square model/mean square error, it is used to determine the variance explained by model is significantly greater than the unexplained variance
- The p value: (>0.05)-this means we fail to reject null hypothesis and the mean is statistically equal between the groups.
- R squared: It is the proportion of variance in Age explained by the model. 0.000014 percent shows that variability of Age can not be explained by Province, it is a low, so our model is not sufficient to explain the variability of Age using Province.

`



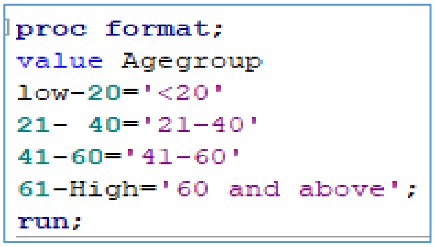
- Diagram – Grouped Box Plot- The Total Length of Boxplot or Interquartile range (Distance between Q1 and Q3) is similar for all 5 Provinces. We can see they line up and Diamond that represents the mean shows that the mean of all Provinces is almost at similar level.
- There are extreme outliers in all Provinces

Conclusion-

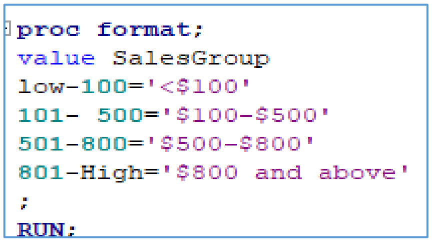
- ➢ Since P values 0.8697>0.05 ,we fail to reject the Null Hypothesis at 5% Significance Level
- ➢ Mean of age is significantly similar in all Provinces.
- ➢ Province explains 0% of variability of Age. Therefore, we can say this is not a good model.

1.3 Segment the customers based on age, province, and sales amount:

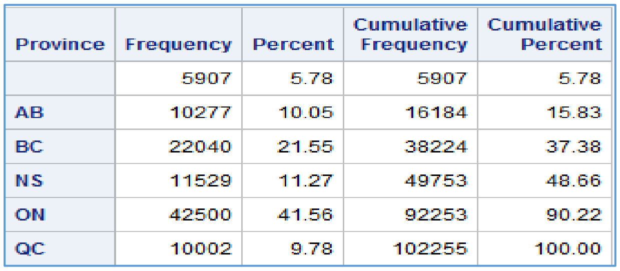
Sales segment: < $100, $100-$500, $500-$800, $800 and above.
Age segments: < 20, 21-40, 41-60, 60 and above.



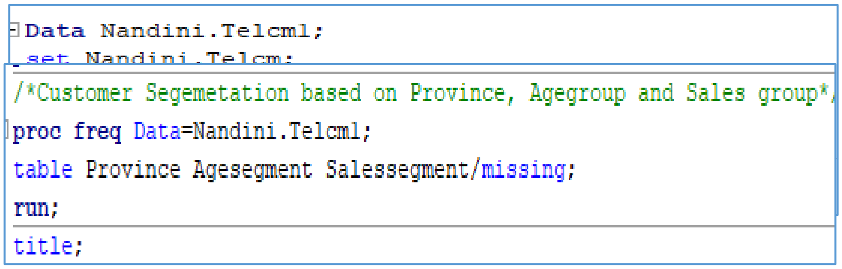
```
proc format;
value Agegroup
low-20='<20'
21- 40='21-40'
41-60='41-60'
61-High='60 and above';
run;
```



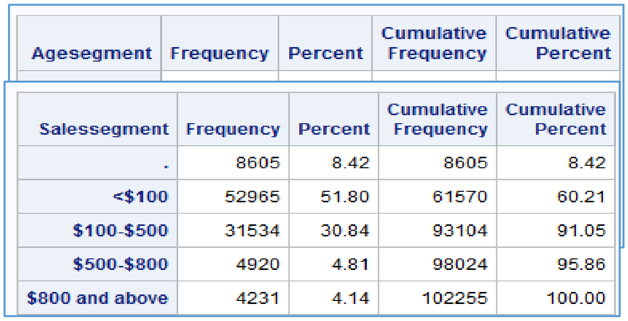
```
proc format;
value SalesGroup
low-100='<$100'
101- 500='$100-$500'
501-800='$500-$800'
801-High='$800 and above'
;
RUN;
```



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



```
Data Nandini.Telcml;
set Nandini.Telcm;
/*Customer Segemetation based on Province, Agegroup and Sales group*/
proc freq Data=Nandini.Telcml;
table Province Agesegment Salessegment/missing;
run;
title;
```



Agesegment	Frequency	Percent	Cumulative Frequency	Cumulative Percent

Salessegment	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	8605	8.42	8605	8.42
<$100	52965	51.80	61570	60.21
$100-$500	31534	30.84	93104	91.05
$500-$800	4920	4.81	98024	95.86
$800 and above	4231	4.14	102255	100.00

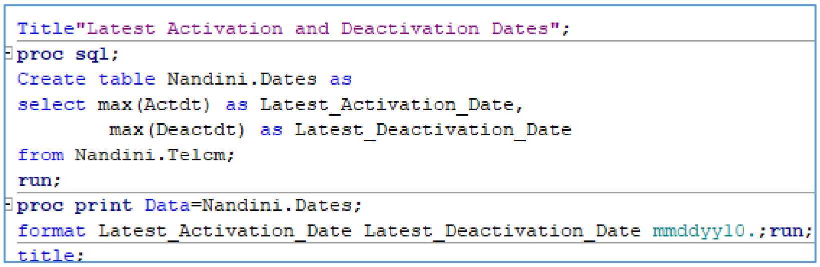
`

Observations from the Customer Segmentation

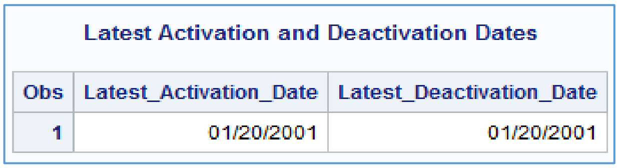
- ON Has Maximum Customer Base whereas QC has Minimum Customer Base
- Maximum Customers are from Age Group 41-60 whereas age group <20 has minimum customer base. This also indicates Young customers do not stick to the particular telecom service for longer.
- Maximum customers has taken the service from this telecom company for amount less that $100. There is a possibility that maximum customers are not happy with service.

1.4. Statistical Analysis:

1) Calculate the tenure in days for each account and give its simple statistics.

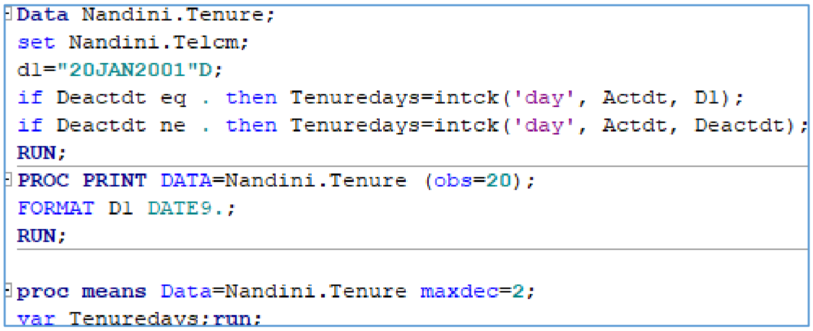


```
Title"Latest Activation and Deactivation Dates";
proc sql;
Create table Nandini.Dates as
select max(Actdt) as Latest_Activation_Date,
       max(Deactdt) as Latest_Deactivation_Date
from Nandini.Telcm;
run;
proc print Data=Nandini.Dates;
format Latest_Activation_Date Latest_Deactivation_Date mmddyy10.;run;
title;
```

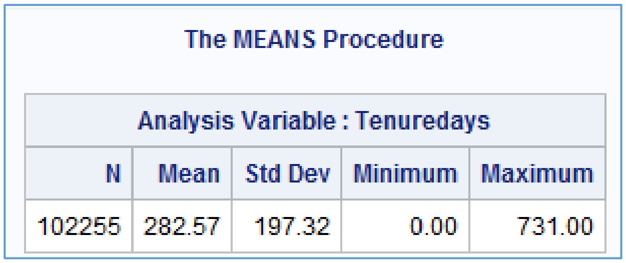


Latest Activation and Deactivation Dates

Obs	Latest_Activation_Date	Latest_Deactivation_Date
1	01/20/2001	01/20/2001



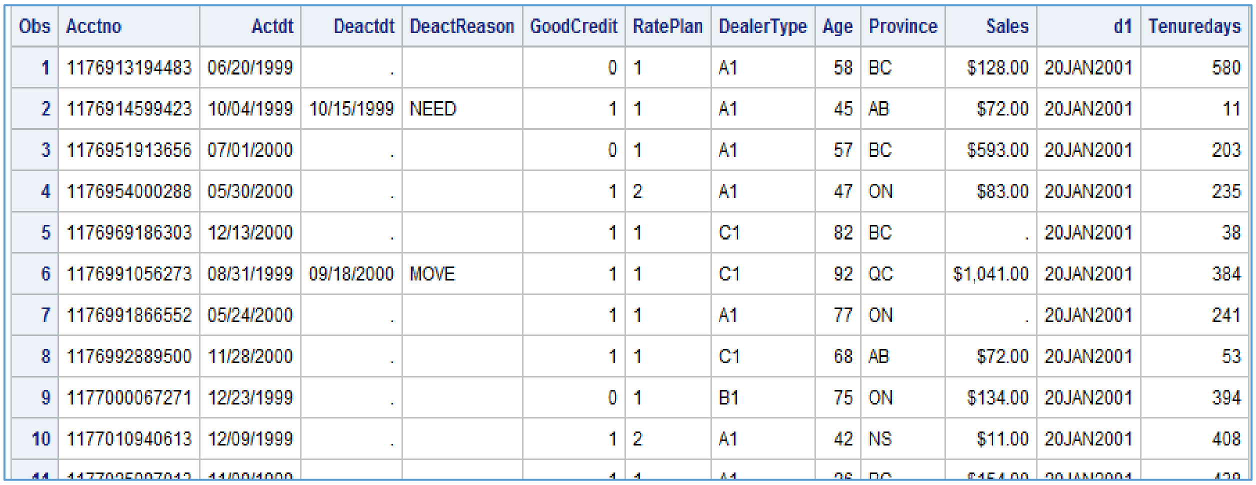
```
Data Nandini.Tenure;
set Nandini.Telcm;
d1="20JAN2001"D;
if Deactdt eq . then Tenuredays=intck('day', Actdt, D1);
if Deactdt ne . then Tenuredays=intck('day', Actdt, Deactdt);
RUN;
PROC PRINT DATA=Nandini.Tenure (obs=20);
FORMAT D1 DATE9.;
RUN;

proc means Data=Nandini.Tenure maxdec=2;
var Tenuredays;run;
```



The MEANS Procedure

Analysis Variable : Tenuredays

N	Mean	Std Dev	Minimum	Maximum
102255	282.57	197.32	0.00	731.00



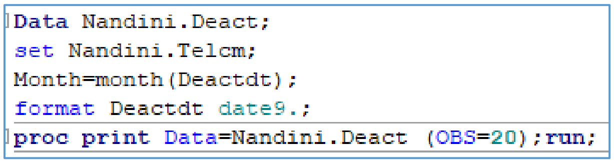
Obs	Acctno	Actdt	Deactdt	DeactReason	GoodCredit	RatePlan	DealerType	Age	Province	Sales	d1	Tenuredays
1	1176913194483	06/20/1999	.		0	1	A1	58	BC	$128.00	20JAN2001	580
2	1176914599423	10/04/1999	10/15/1999	NEED	1	1	A1	45	AB	$72.00	20JAN2001	11
3	1176951913656	07/01/2000	.		0	1	A1	57	BC	$593.00	20JAN2001	203
4	1176954000288	05/30/2000	.		1	2	A1	47	ON	$83.00	20JAN2001	235
5	1176969186303	12/13/2000	.		1	1	C1	82	BC	.	20JAN2001	38
6	1176991056273	08/31/1999	09/18/2000	MOVE	1	1	C1	92	QC	$1,041.00	20JAN2001	384
7	1176991866552	05/24/2000	.		1	1	A1	77	ON	.	20JAN2001	241
8	1176992889500	11/28/2000	.		1	1	C1	68	AB	$72.00	20JAN2001	53
9	1177000067271	12/23/1999	.		0	1	B1	75	ON	$134.00	20JAN2001	394
10	1177010940613	12/09/1999	.		1	2	A1	42	NS	$11.00	20JAN2001	408

`

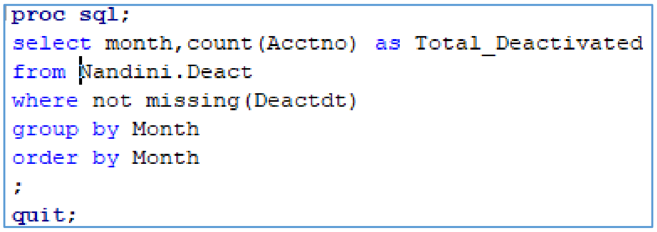
Observations:

- For tenure of customer who are still active , we have considered end date as the latest actvation date . For the tenure of customers who have deactivated service, end date is same as deactivation date.
- For active Customer- Tenure= from Activation date to Latest Activation date
- For Deactivated Customers-Tenure- From Activation date till Deactivation Date
- Latest activation as well as Deactivation Date is 20th Jan 2001
- Mean of Tenure is 282.57. Maximum Tenure is 731 days.
- Minimum tenure is 0 days which indicates there are customers who has deactivated the service on the same day

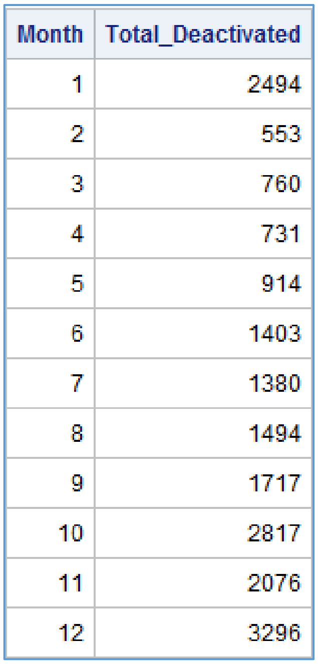
2) Calculate the number of accounts deactivated for each month.



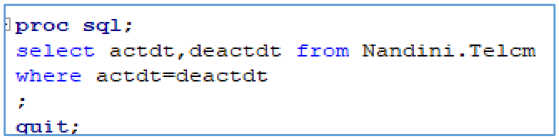
```
Data Nandini.Deact;
set Nandini.Telcm;
Month=month(Deactdt);
format Deactdt date9.;
proc print Data=Nandini.Deact (OBS=20);run;
```



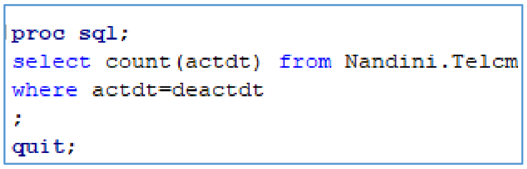
```
proc sql;
select month,count(Acctno) as Total_Deactivated
from Nandini.Deact
where not missing(Deactdt)
group by Month
order by Month
;
quit;
```



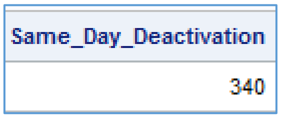
Month	Total_Deactivated
1	2494
2	553
3	760
4	731
5	914
6	1403
7	1380
8	1494
9	1717
10	2817
11	2076
12	3296



```
proc sql;
select actdt,deactdt from Nandini.Telcm
where actdt=deactdt
;
quit;
```



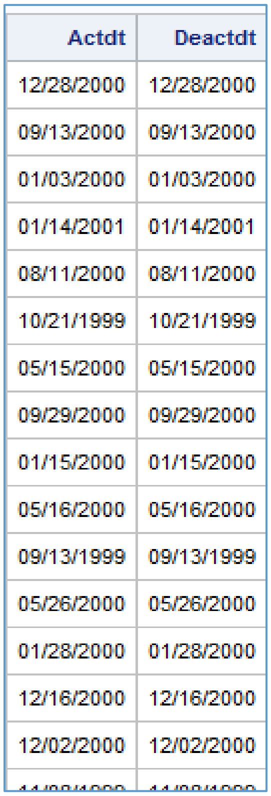
```
proc sql;
select count(actdt) from Nandini.Telcm
where actdt=deactdt
;
quit;
```



Same_Day_Deactivation
340

Observations:

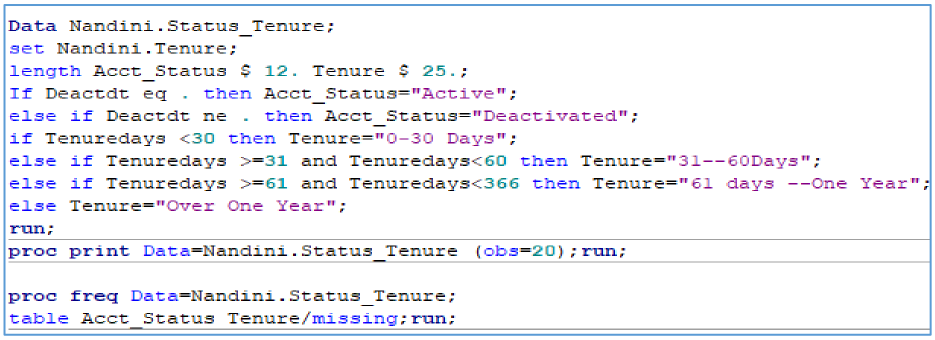
- Maximum Deactivation Occurred in Winter- From October to January
- 340 customers deactivated service on the same day



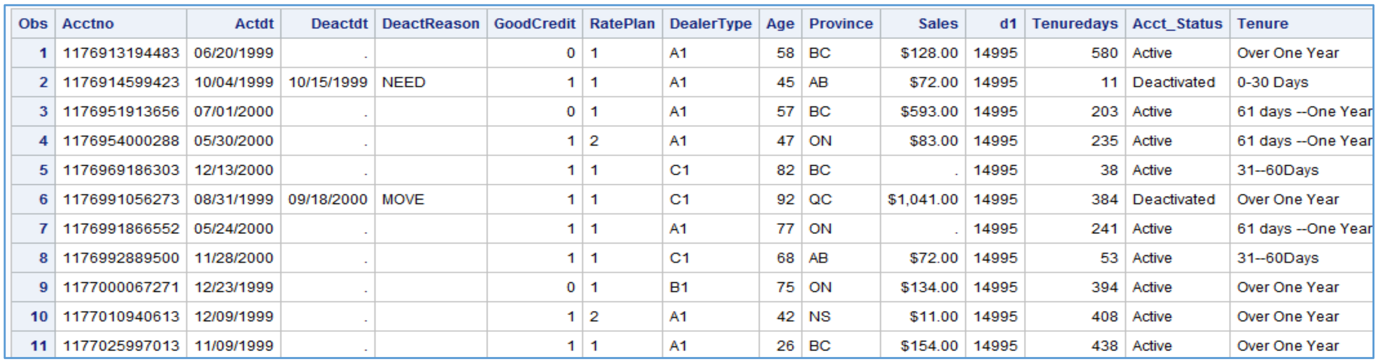
Actdt	Deactdt
12/28/2000	12/28/2000
09/13/2000	09/13/2000
01/03/2000	01/03/2000
01/14/2001	01/14/2001
08/11/2000	08/11/2000
10/21/1999	10/21/1999
05/15/2000	05/15/2000
09/29/2000	09/29/2000
01/15/2000	01/15/2000
05/16/2000	05/16/2000
09/13/1999	09/13/1999
05/26/2000	05/26/2000
01/28/2000	01/28/2000
12/16/2000	12/16/2000
12/02/2000	12/02/2000

3) Segment the account, first by account status “Active” and “Deactivated”, then byTenure: < 30 days, 31---60 days, 61 days--- one year, over one year. Report the number of accounts of percent of all for each segment.

`



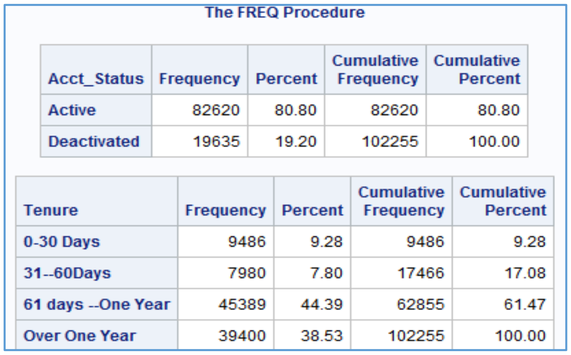
```
Data Nandini.Status_Tenure;
set Nandini.Tenure;
length Acct_Status $ 12. Tenure $ 25.;
If Deactdt eq . then Acct_Status="Active";
else if Deactdt ne . then Acct_Status="Deactivated";
if Tenuredays <30 then Tenure="0-30 Days";
else if Tenuredays >=31 and Tenuredays<60 then Tenure="31--60Days";
else if Tenuredays >=61 and Tenuredays<366 then Tenure="61 days --One Year";
else Tenure="Over One Year";
run;
proc print Data=Nandini.Status_Tenure (obs=20);run;

proc freq Data=Nandini.Status_Tenure;
table Acct_Status Tenure/missing;run;
```



Obs	Acctno	Actdt	Deactdt	DeactReason	GoodCredit	RatePlan	DealerType	Age	Province	Sales	d1	Tenuredays	Acct_Status	Tenure
1	1176913194483	06/20/1999	.		0	1	A1	58	BC	$128.00	14995	580	Active	Over One Year
2	1176914599423	10/04/1999	10/15/1999	NEED	1	1	A1	45	AB	$72.00	14995	11	Deactivated	0-30 Days
3	1176951913656	07/01/2000	.		0	1	A1	57	BC	$593.00	14995	203	Active	61 days --One Year
4	1176954000288	05/30/2000	.		1	2	A1	47	ON	$83.00	14995	235	Active	61 days --One Year
5	1176969186303	12/13/2000	.		1	1	C1	82	BC	.	14995	38	Active	31--60Days
6	1176991056273	08/31/1999	09/18/2000	MOVE	1	1	C1	92	QC	$1,041.00	14995	384	Deactivated	Over One Year
7	1176991866552	05/24/2000	.		1	1	A1	77	ON	.	14995	241	Active	61 days --One Year
8	1176992889500	11/28/2000	.		1	1	C1	68	AB	$72.00	14995	53	Active	31--60Days
9	1177000067271	12/23/1999	.		0	1	B1	75	ON	$134.00	14995	394	Active	Over One Year
10	1177010940613	12/09/1999	.		1	2	A1	42	NS	$11.00	14995	408	Active	Over One Year
11	1177025997013	11/09/1999	.		1	1	A1	26	BC	$154.00	14995	438	Active	Over One Year

Observation-

- Tenure of Maximum customers is between 61 days- One year
- Minimum customers have tenure between 31 days- 60 days
- Active customers are 4 times greater than Deactivated customers , which is a good thing.



The FREQ Procedure

Acct_Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Active	82620	80.80	82620	80.80
Deactivated	19635	19.20	102255	100.00

Tenure	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0-30 Days	9486	9.28	9486	9.28
31--60Days	7980	7.80	17466	17.08
61 days --One Year	45389	44.39	62855	61.47
Over One Year	39400	38.53	102255	100.00

4) Test the general association between the tenure segments and “Good Credit” “RatePlan ” and “DealerType.”

A) Association Between Tenure segments and Good Credit

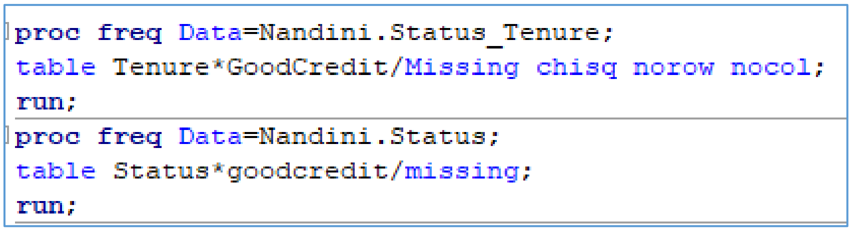
Ho- Null Hypothesis-There is no association between Tenure and Good Credit
H1- Alternate Hypothesis-There is association between Tenure and Good Credit

Approach- since both are categorical Variables we will use:-

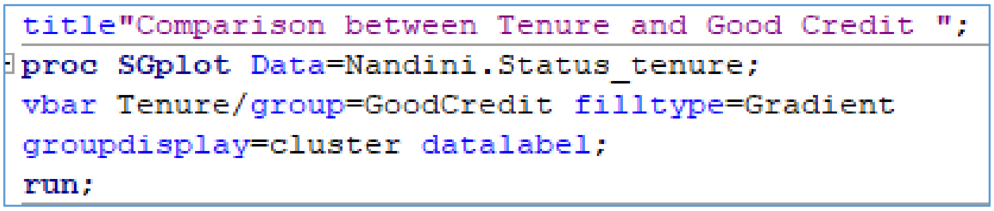
- For Summarisation- Frequency Table

`

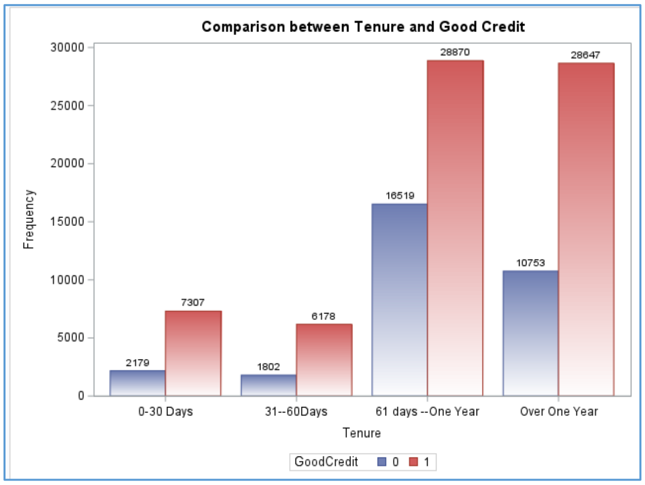
- For Visualisation- Grouped Bar Chart
- For Independency- Chi sq test

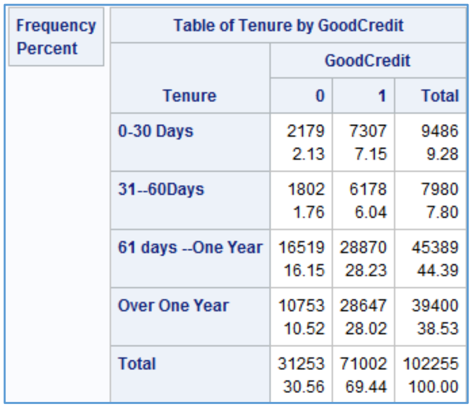


```
proc freq Data=Nandini.Status_Tenure;
table Tenure*GoodCredit/Missing chisq norow nocol;
run;
proc freq Data=Nandini.Status;
table Status*goodcredit/missing;
run;
```

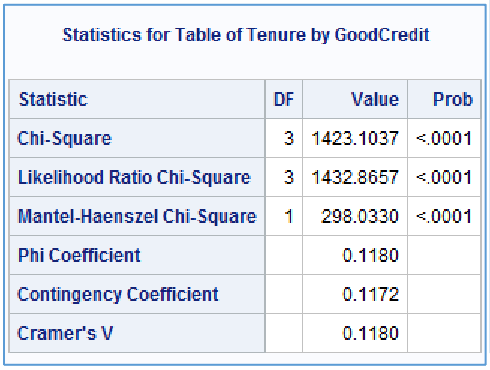


```
title"Comparison between Tenure and Good Credit ";
proc SGplot Data=Nandini.Status_tenure;
vbar Tenure/group=GoodCredit filltype=Gradient
groupdisplay=cluster datalabel;
run;
```



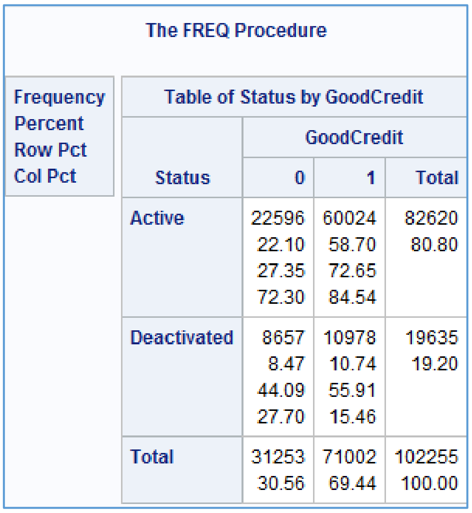


Frequency Percent	Table of Tenure by GoodCredit			
		GoodCredit		
	Tenure	0	1	Total
	0-30 Days	2179 2.13	7307 7.15	9486 9.28
	31--60Days	1802 1.76	6178 6.04	7980 7.80
	61 days --One Year	16519 16.15	28870 28.23	45389 44.39
	Over One Year	10753 10.52	28647 28.02	39400 38.53
	Total	31253 30.56	71002 69.44	102255 100.00



Statistics for Table of Tenure by GoodCredit

Statistic	DF	Value	Prob
Chi-Square	3	1423.1037	<.0001
Likelihood Ratio Chi-Square	3	1432.8657	<.0001
Mantel-Haenszel Chi-Square	1	298.0330	<.0001
Phi Coefficient		0.1180	
Contingency Coefficient		0.1172	
Cramer's V		0.1180	



The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of Status by GoodCredit			
		GoodCredit		
	Status	0	1	Total
	Active	22596 22.10 27.35 72.30	60024 58.70 72.65 84.54	82620 80.80
	Deactivated	8657 8.47 44.09 27.70	10978 10.74 55.91 15.46	19635 19.20
	Total	31253 30.56	71002 69.44	102255 100.00

- Grouped Bar Chart shows the distribution of Tenure segments with 2 levels of Good Credit
- From first table, we can say out of 102255

  0 to 30 days- has 7.15% of Good Credit customers and 2.13% of Customers do not have good credit

`

- 31 days to 60 days- has 6.04% of Good Credit customers and 1.76% of Customers do not have good credit
- 61 days to One year - has 28.23% of Good Credit customers and 16.15% of Customers do not have good credit. This group has maximum customers with Good Credit .
- over One year - has 28.02% of Good Credit customers and 10.52% of Customers do not have good credit
- Simultaneously Third table indicates that Maximum active customers have Good Credit
- Second table shows chi-square P value is <0.05 and Cramer's V is 0.1180, which means there is a statistically significant association between Tenure and Good Credit

Conclusion-

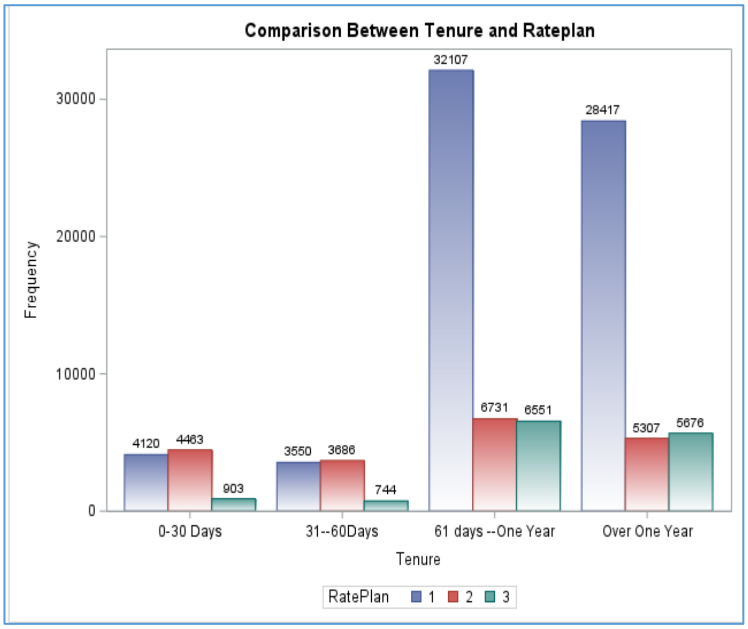
- Since P values >0.05 reject the Null Hypothesis at 5% Significance Level
- We accept there is statistically significant association between Tenure and Good Credit Customers
- In other words we can say Customer with more tenure has Good Credit.

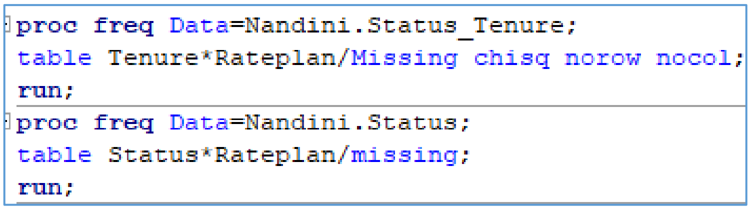
B) Association Between Tenure segments and RatePlan

Ho- Null Hypothesis-There is no association between Tenure and Rate Plan
H1- Alternate Hypothesis-There is association between Tenure and Rate Plan

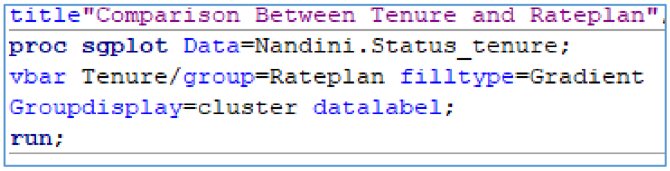
Approach- since both are categorical Variables we will use:-

- For Summarisation- Frequency Table
- For Visualisation- Grouped Bar Chart
- For Independency- Chi sq test

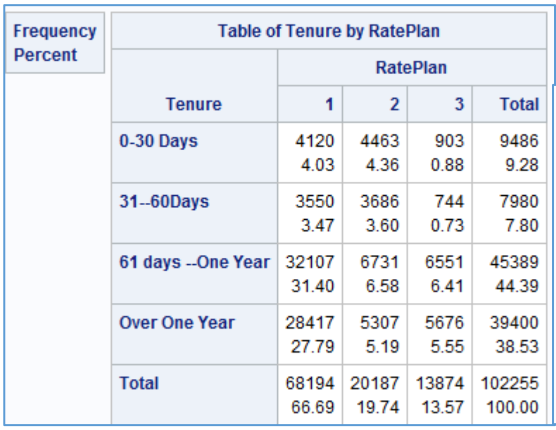




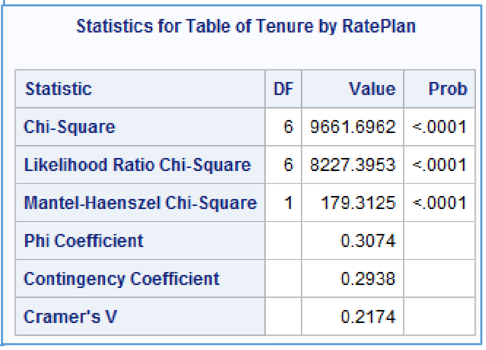
```
proc freq Data=Nandini.Status_Tenure;
table Tenure*Rateplan/Missing chisq norow nocol;
run;
proc freq Data=Nandini.Status;
table Status*Rateplan/missing;
run;
```



```
title"Comparison Between Tenure and Rateplan";
proc sgplot Data=Nandini.Status_tenure;
vbar Tenure/group=Rateplan filltype=Gradient
Groupdisplay=cluster datalabel;
run;
```

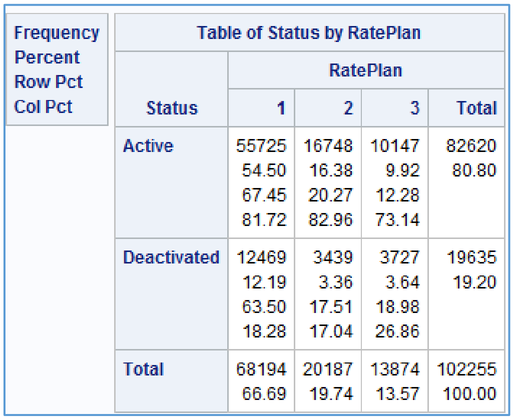


Frequency Percent	Table of Tenure by RatePlan			
	RatePlan			
Tenure	1	2	3	Total
0-30 Days	4120 4.03	4463 4.36	903 0.88	9486 9.28
31--60Days	3550 3.47	3686 3.60	744 0.73	7980 7.80
61 days --One Year	32107 31.40	6731 6.58	6551 6.41	45389 44.39
Over One Year	28417 27.79	5307 5.19	5676 5.55	39400 38.53
Total	68194 66.69	20187 19.74	13874 13.57	102255 100.00



Statistics for Table of Tenure by RatePlan

Statistic	DF	Value	Prob
Chi-Square	6	9661.6962	<.0001
Likelihood Ratio Chi-Square	6	8227.3953	<.0001
Mantel-Haenszel Chi-Square	1	179.3125	<.0001
Phi Coefficient		0.3074	
Contingency Coefficient		0.2938	
Cramer's V		0.2174	



Frequency Percent Row Pct Col Pct	Table of Status by RatePlan			
	RatePlan			
Status	1	2	3	Total
Active	55725 54.50 67.45 81.72	16748 16.38 20.27 82.96	10147 9.92 12.28 73.14	82620 80.80
Deactivated	12469 12.19 63.50 18.28	3439 3.36 17.51 17.04	3727 3.64 18.98 26.86	19635 19.20
Total	68194 66.69	20187 19.74	13874 13.57	102255 100.00

- Grouped Bar Chart shows the distribution of Tenure segments with 3 levels of Rate Plan
- From first table, we can say out of 102255 customers
  For tenure -0 to 30 days- 4.03% Customer have Rateplan 1,4.36 % has Rateplan 2,0.88% customers have Rateplan 3
- For Tenure 31 days to 60 days- - 3.47% Customer have Rateplan 1,3.6% has Rateplan 2,0.73% customers have Rateplan 3
- For Tenure 61 days to One year – 31.40% Customer have Rateplan 1, 6.58 % has Rateplan 2, 6.41% customers have Rateplan 3.
- For tenure over One year – 27.79% Customer have Rateplan 1,5.19% has Rateplan 2,5.55% customers have Rateplan 3.
- Simultaneously Third table indicates that Maximum active customers 54.50% prefer Rate plan 1.
- Second table shows chi-square P value is <0.05 and Cramer's V is 0.2174, which means there is a statistically significant association between Tenure and Rate Plan

Conclusion-

- Since P values >0.05 reject the Null Hypothesis at 5% Significance Level
- We accept there is statistically significant association between Tenure and Rate Plan
- Customers with Rate plan 1 have tenure more than 60 days.
- Rate Plan 2 has minimum churn followed by rate plan
- Rate Plan 3 has Maximum Churn.
- Over all rate plan 1 is better than other 2
- Majority Active customers are likely to have Rate plan 1

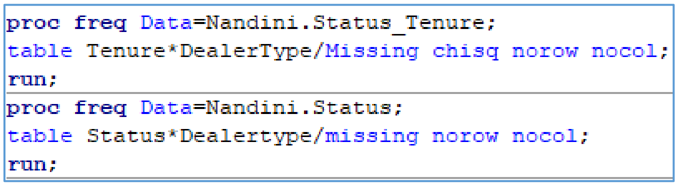
C) <u>Association Between Tenure segments and Dealer Type</u>

Ho- Null Hypothesis-There is no association between Tenure and Dealer Type
H1- Alternate Hypothesis-There is association between Tenure and Dealer Type

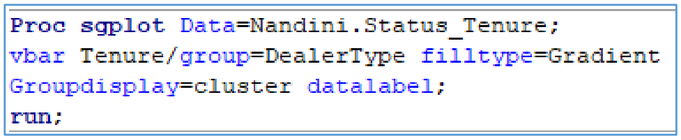
`

Approach- since both are categorical Variables we will use:-

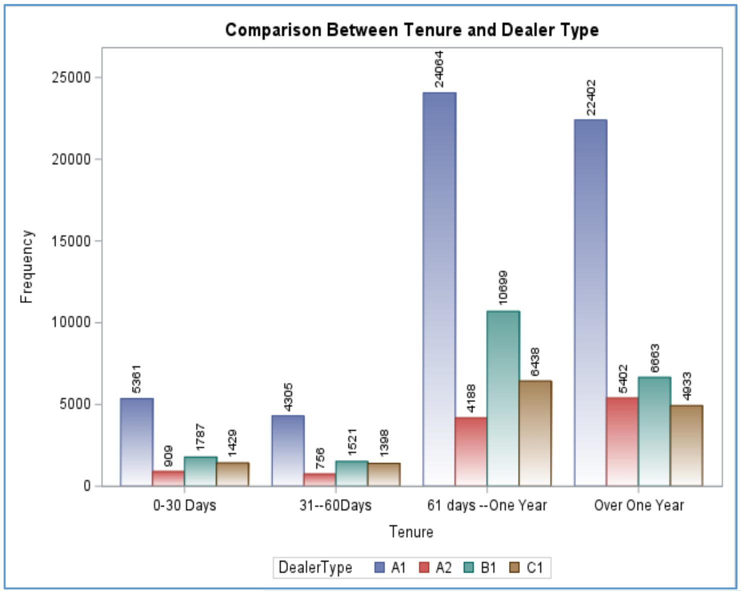
- For Summarization- Frequency Table
- For Visualisation- Grouped Bar Chart
- For Independency- Chi sq test

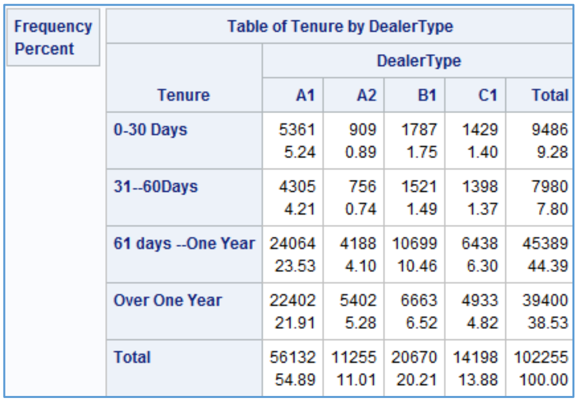


```
proc freq Data=Nandini.Status_Tenure;
table Tenure*DealerType/Missing chisq norow nocol;
run;
proc freq Data=Nandini.Status;
table Status*Dealertype/missing norow nocol;
run;
```

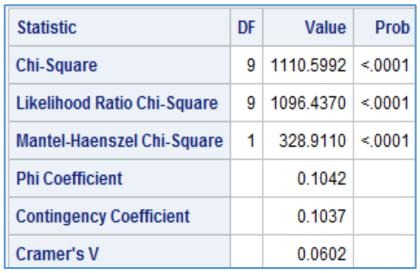


```
Proc sgplot Data=Nandini.Status_Tenure;
vbar Tenure/group=DealerType filltype=Gradient
Groupdisplay=cluster datalabel;
run;
```

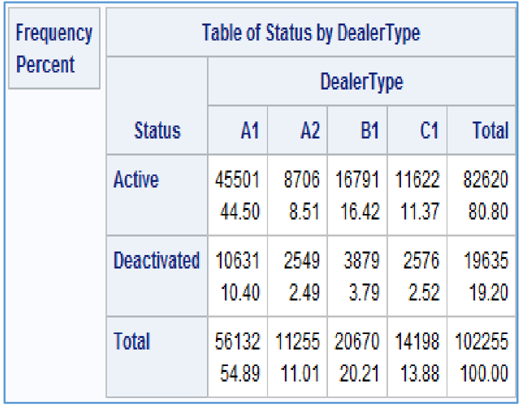




Frequency Percent	Table of Tenure by DealerType					
		DealerType				
	Tenure	A1	A2	B1	C1	Total
	0-30 Days	5361 5.24	909 0.89	1787 1.75	1429 1.40	9486 9.28
	31--60Days	4305 4.21	756 0.74	1521 1.49	1398 1.37	7980 7.80
	61 days --One Year	24064 23.53	4188 4.10	10699 10.46	6438 6.30	45389 44.39
	Over One Year	22402 21.91	5402 5.28	6663 6.52	4933 4.82	39400 38.53
	Total	56132 54.89	11255 11.01	20670 20.21	14198 13.88	102255 100.00



Statistic	DF	Value	Prob
Chi-Square	9	1110.5992	<.0001
Likelihood Ratio Chi-Square	9	1096.4370	<.0001
Mantel-Haenszel Chi-Square	1	328.9110	<.0001
Phi Coefficient		0.1042	
Contingency Coefficient		0.1037	
Cramer's V		0.0602	



Frequency Percent	Table of Status by DealerType					
		DealerType				
	Status	A1	A2	B1	C1	Total
	Active	45501 44.50	8706 8.51	16791 16.42	11622 11.37	82620 80.80
	Deactivated	10631 10.40	2549 2.49	3879 3.79	2576 2.52	19635 19.20
	Total	56132 54.89	11255 11.01	20670 20.21	14198 13.88	102255 100.00

- Grouped Bar Chart shows the distribution of Tenure segments with 4 Dealer Types
- From first table, we can say out of 102255 customers
  For tenure -0 to 30 days- 5.24% Customer Deal with A1,0.89% Customer Deal with A2,1.75% Customer Deal with B1 and 1.4% Customer deal with C1
- For Tenure 31 days to 60 days- 4.21% Customer Deal with A1,0.74% Customer Deal with A2,1.49% Customer Deal with B1 and 1.37% Customer deal with C1
- For Tenure 61 days to One year – 23.53% Customer Deal with A1,4.1% Customer Deal with A2,10.46% Customer Deal with B1 and 6.3% Customer deal with C1
- For tenure over One year –21.91% Customer Deal with A1,5.28 % Customer Deal with A2,6.52% Customer Deal with B1 and 4.82% Customer deal with C1
- Simultaneously Third table indicates that Maximum active customers 44.50% prefer to deal with Dealer type A1
- Second table shows chi-square P value is <0.05 and Cramer's V is 0.0602, which means there is a statistically significant association between Tenure and Dealer Type

Conclusion-

➢ Since P values >0.05 reject the Null Hypothesis at 5% Significance Level

`

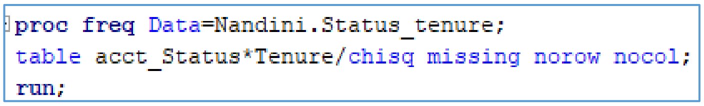
- We accept there is statistically significant association between Tenure and Dealer Type
- Customers Who deal with A1 1 have tenure more than 60 days.
- Majority Active customers are likely to Deal with A1.

5) Is there any association between the account status and the tenure segments?

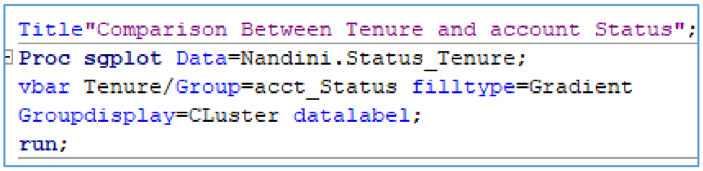
Ho- Null Hypothesis-There is no association between Tenure and Dealer Type
H1- Alternate Hypothesis-There is association between Tenure and Dealer Type

Approach- since both are categorical Variables we will use:-

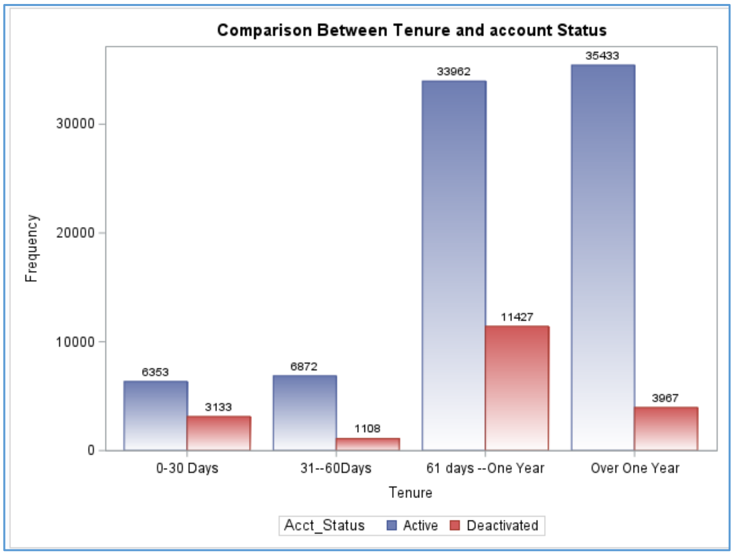
- For Summarization- Frequency Table
- For Visualisation- Grouped Bar Chart
- For Independency- Chi sq test
-

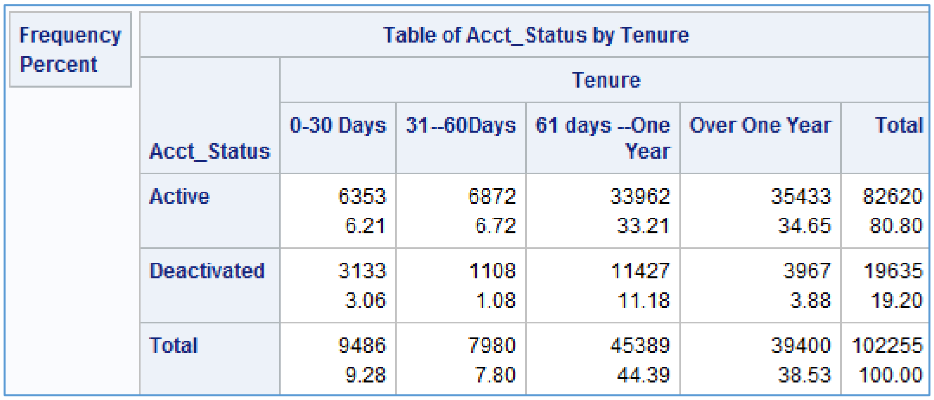


```
proc freq Data=Nandini.Status_tenure;
table acct_Status*Tenure/chisq missing norow nocol;
run;
```

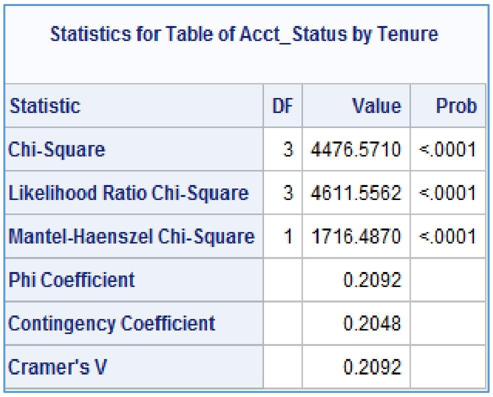


```
Title"Comparison Between Tenure and account Status";
Proc sgplot Data=Nandini.Status_Tenure;
vbar Tenure/Group=acct_Status filltype=Gradient
Groupdisplay=CLuster datalabel;
run;
```





Frequency Percent	Table of Acct_Status by Tenure				
	Tenure				
Acct_Status	0-30 Days	31--60Days	61 days --One Year	Over One Year	Total
Active	6353 6.21	6872 6.72	33962 33.21	35433 34.65	82620 80.80
Deactivated	3133 3.06	1108 1.08	11427 11.18	3967 3.88	19635 19.20
Total	9486 9.28	7980 7.80	45389 44.39	39400 38.53	102255 100.00



Statistics for Table of Acct_Status by Tenure

Statistic	DF	Value	Prob
Chi-Square	3	4476.5710	<.0001
Likelihood Ratio Chi-Square	3	4611.5562	<.0001
Mantel-Haenszel Chi-Square	1	1716.4870	<.0001
Phi Coefficient		0.2092	
Contingency Coefficient		0.2048	
Cramer's V		0.2092	

- Grouped Bar Chart shows the distribution of Tenure segments with Account Status Active and Deactivated
- From first table, we can say out of 102255 customers

`

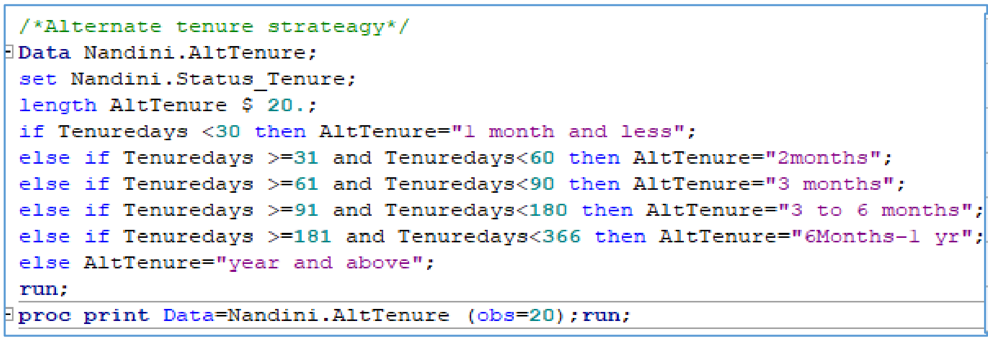
For tenure -0 to 30 days- 6.21% customer are active and 3.06% are deactivated

- For Tenure 31 days to 60 days- 6.72% customer are active and 1.08% are deactivated
- For Tenure 61 days to One year –33.21% customer are active and 11.18% are deactivated
- For tenure over One year –34.65% customer are active and 3.88% are deactivated
- Simultaneously Third table indicates that Maximum active customers 34.65% have Tenure More than one year
- Second table shows chi-square P value is <0.05 and Cramer’s V is 0.2092, which means there is a statistically significant association between Tenure and Account Status

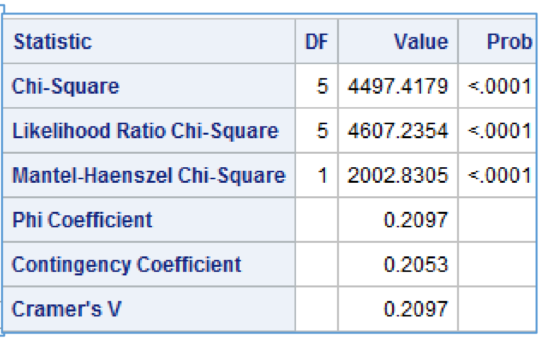
Conclusion-

- ➢ Since P values >0.05 reject the Null Hypothesis at 5% Significance Level
- ➢ We accept there is statistically significant association between Tenure and Account Status
- ➢ Maximum Active Customers have tenure more than 60 days.
- ➢ Customer churn is seen more with tenure <30 days.(33%)(3133 customers out of 9486 for Tenure less than 30 days)

<u>Could you find a better tenure segmentation strategy that is more associated with the account status?</u>



```
/*Alternate tenure strateagy*/
Data Nandini.AltTenure;
set Nandini.Status_Tenure;
length AltTenure $ 20.;
if Tenuredays <30 then AltTenure="1 month and less";
else if Tenuredays >=31 and Tenuredays<60 then AltTenure="2months";
else if Tenuredays >=61 and Tenuredays<90 then AltTenure="3 months";
else if Tenuredays >=91 and Tenuredays<180 then AltTenure="3 to 6 months";
else if Tenuredays >=181 and Tenuredays<366 then AltTenure="6Months-1 yr";
else AltTenure="year and above";
run;
proc print Data=Nandini.AltTenure (obs=20);run;
```



Statistic	DF	Value	Prob
Chi-Square	5	4497.4179	<.0001
Likelihood Ratio Chi-Square	5	4607.2354	<.0001
Mantel-Haenszel Chi-Square	1	2002.8305	<.0001
Phi Coefficient		0.2097	
Contingency Coefficient		0.2053	
Cramer's V		0.2097	

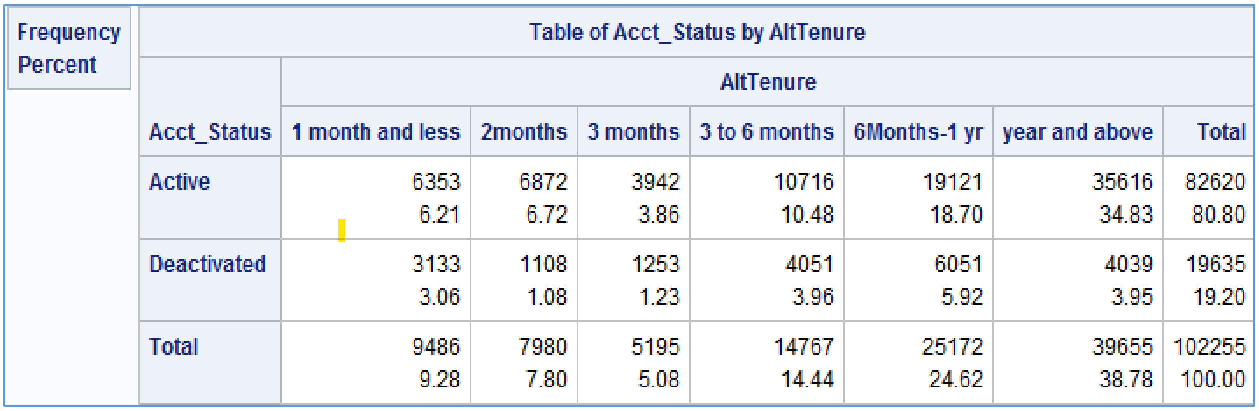
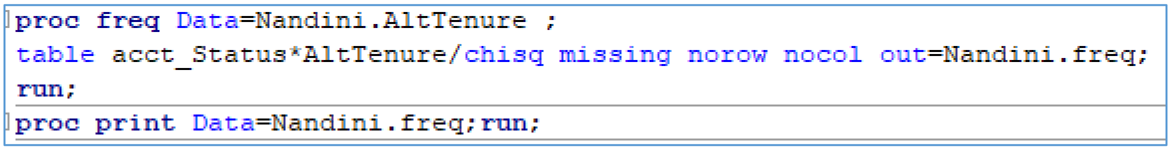


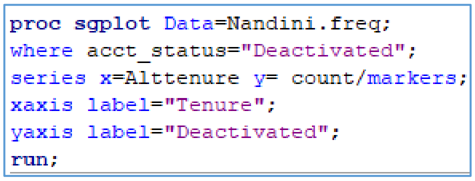
Table of Acct_Status by AltTenure

Frequency Percent	AltTenure						
Acct_Status	1 month and less	2months	3 months	3 to 6 months	6Months-1 yr	year and above	Total
Active	6353 6.21	6872 6.72	3942 3.86	10716 10.48	19121 18.70	35616 34.83	82620 80.80
Deactivated	3133 3.06	1108 1.08	1253 1.23	4051 3.96	6051 5.92	4039 3.95	19635 19.20
Total	9486 9.28	7980 7.80	5195 5.08	14767 14.44	25172 24.62	39655 38.78	102255 100.00



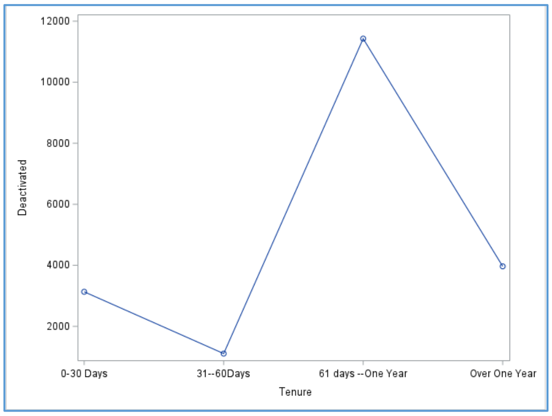
```
proc freq Data=Nandini.AltTenure ;
table acct_Status*AltTenure/chisq missing norow nocol out=Nandini.freq;
run;
proc print Data=Nandini.freq;run;
```

`

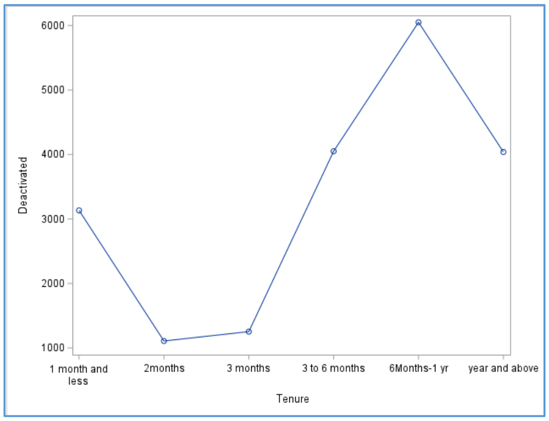


```
proc sgplot Data=Nandini.freq;
where acct_status="Deactivated";
series x=Alttenure y= count/markers;
xaxis label="Tenure";
yaxis label="Deactivated";
run;
```

Original Tenure



Alternate Tenure Segmentation



- Frequency Table Shows that Maximum Active customers have tenure more than 1 year and above
- Diagram-Alternate Tenure segmentation indicates the Trend of Customer churn with each alternative Tenure segment. Customer churn was more in first month which reduced in next 2 months. It again increased after month 3 till year end. It finally decreased after one year.
- Diagram –Original Tenure does not explain the customer churn well.

6) Does the Sales amount differ among different account statuses, GoodCredit, and customer age segments?

A) Sales Distribution across Age Segments

H0- Means of Sales is equal All Age groups i.e Sale equally distributed in all age groups

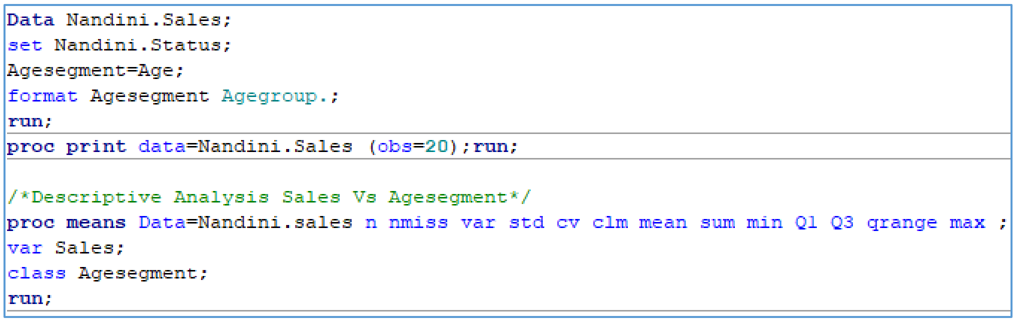
H1 Means of Sales is not equal All Age groups i.e Sale Is not equally distributed in all age groups

.

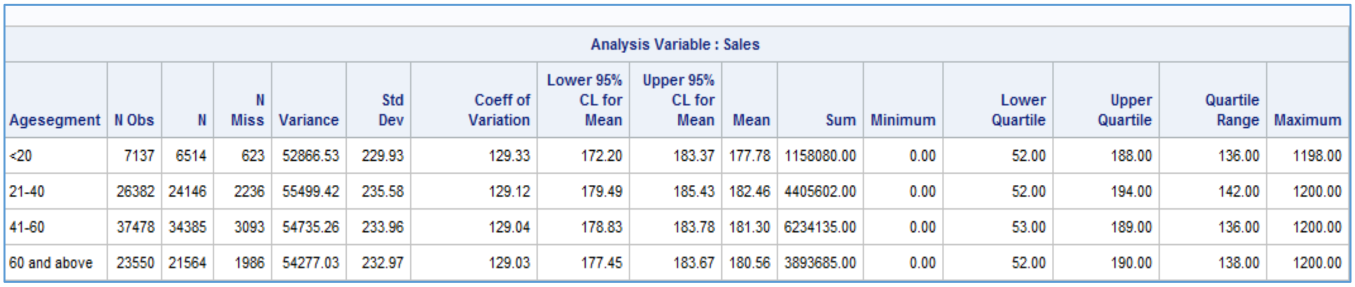
Approach- since Sale is Numerical variable and AgeSegment is a categorical Variable we will use:-

- For Summarisation- Proc Means
- For Normality-Proc Univariate
- For Visualisation- Grouped Box Plot
- For Independency- Proc Anova

`



```
Data Nandini.Sales;
set Nandini.Status;
Agesegment=Age;
format Agesegment Agegroup.;
run;
proc print data=Nandini.Sales (obs=20);run;

/*Descriptive Analysis Sales Vs Agesegment*/
proc means Data=Nandini.sales n nmiss var std cv clm mean sum min Q1 Q3 qrange max ;
var Sales;
class Agesegment;
run;
```



Analysis Variable : Sales

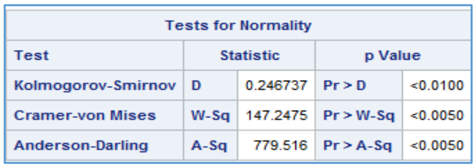
Agesegment	N Obs	N	N Miss	Variance	Std Dev	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean	Mean	Sum	Minimum	Lower Quartile	Upper Quartile	Quartile Range	Maximum
<20	7137	6514	623	52866.53	229.93	129.33	172.20	183.37	177.78	1158080.00	0.00	52.00	188.00	136.00	1198.00
21-40	26382	24146	2236	55499.42	235.58	129.12	179.49	185.43	182.46	4405602.00	0.00	52.00	194.00	142.00	1200.00
41-60	37478	34385	3093	54735.26	233.96	129.04	178.83	183.78	181.30	6234135.00	0.00	53.00	189.00	136.00	1200.00
60 and above	23550	21564	1986	54277.03	232.97	129.03	177.45	183.67	180.56	3893685.00	0.00	52.00	190.00	138.00	1200.00

- From the results of Means Procedure, we see that Mean and standard deviation of all age segments are very much closer to each other
- To statistically prove this we will use normality test using Proc Univariate, Homoscedasticity test using Proc GLM and finally Proc Anova to prove means are equal in both groups

- Test of Normality-

H0- Sales is normally distributed
H1- Sales is not normally distributed

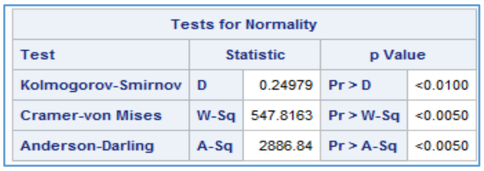
Age <20-P value <0.05



Tests for Normality

Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.246737	Pr > D	<0.0100
Cramer-von Mises	W-Sq	147.2475	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	779.516	Pr > A-Sq	<0.0050

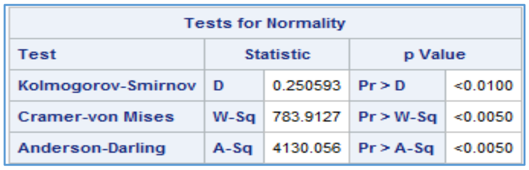
Age 21-40-P value <0.05



Tests for Normality

Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.24979	Pr > D	<0.0100
Cramer-von Mises	W-Sq	547.8163	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2886.84	Pr > A-Sq	<0.0050

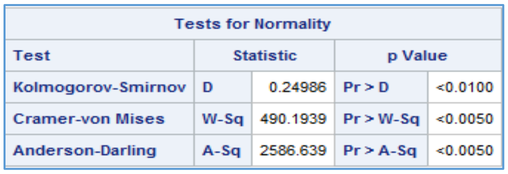
Age 41-60-P value <0.05



Tests for Normality

Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.250593	Pr > D	<0.0100
Cramer-von Mises	W-Sq	783.9127	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	4130.056	Pr > A-Sq	<0.0050

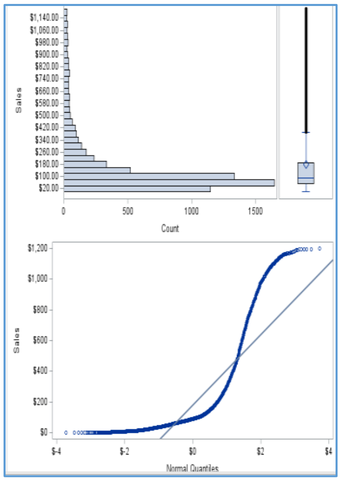
Age 60&above-P value <0.05

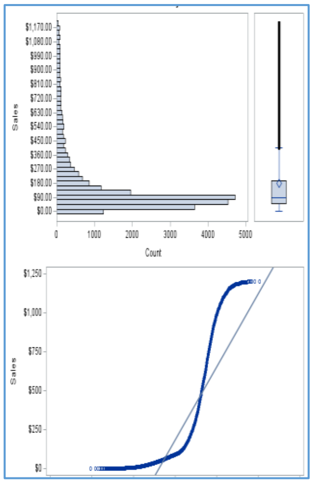


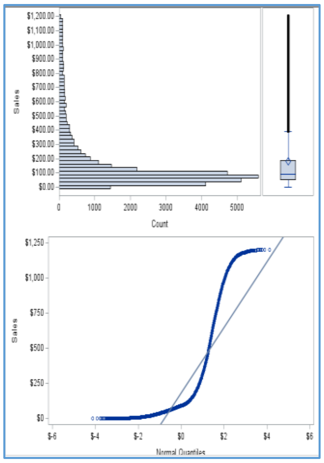
Tests for Normality

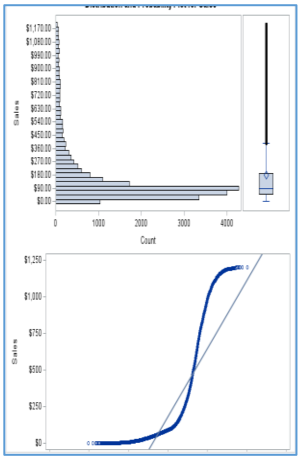
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.24986	Pr > D	<0.0100
Cramer-von Mises	W-Sq	490.1939	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2586.639	Pr > A-Sq	<0.0050

`





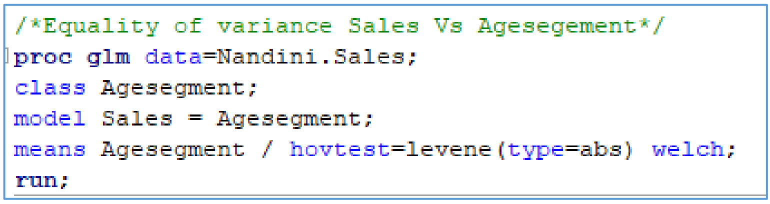




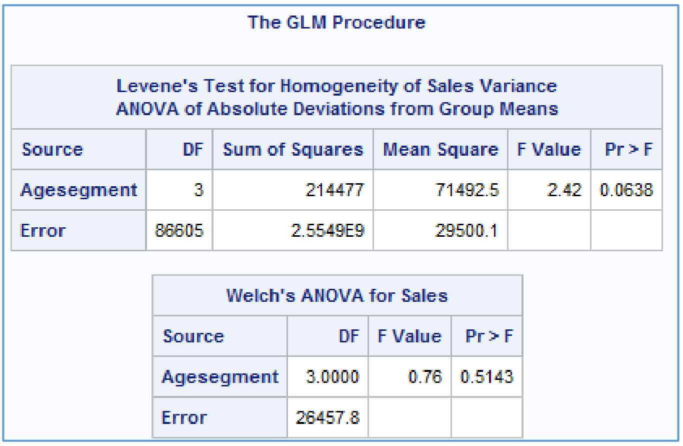
- P value in Kolmogorov-Smirnov of Normality is less than 0.05 significance level for all age segments.
- So we reject Null Hypothesis of normality and conclude that Sales is not normally distributed,
- However as per CLT, since sample size is more than>30, we can assume Sales is normally distributed
- No bell shape visible -Data is highly skewed-Sales not normally distributed in any age group

- Test of Homoscedasticity for equality of variance-

  H0- Variance of Sales is equal in all Age Groups
  H1- Variance of Sales is not equal in all Age Groups



```
/*Equality of variance Sales Vs Agesegement*/
proc glm data=Nandini.Sales;
class Agesegment;
model Sales = Agesegment;
means Agesegment / hovtest=levene(type=abs) welch;
run;
```



The GLM Procedure

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

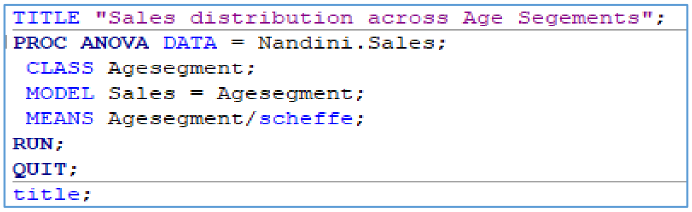
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Agesegment	3	214477	71492.5	2.42	0.0638
Error	86605	2.5549E9	29500.1		

Welch's ANOVA for Sales

Source	DF	F Value	Pr > F
Agesegment	3.0000	0.76	0.5143
Error	26457.8		

P value in Levene's Test for Equality of Variance 0.0638 is greater than 0.05 significance level.

Therefore, we fail to reject Null Hypothesis and Conclude the variance of Sale is equal All Age segments.



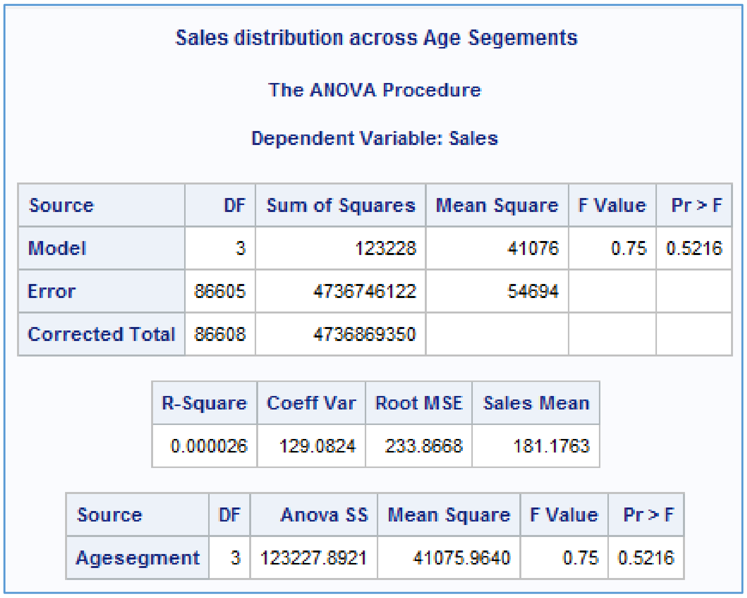
```
TITLE "Sales distribution across Age Segements";
PROC ANOVA DATA = Nandini.Sales;
 CLASS Agesegment;
 MODEL Sales = Agesegment;
 MEANS Agesegment/scheffe;
RUN;
QUIT;
title;
```

- Test of difference-

  H0- Mean of Sale is Equal in all Age groups

`

H1- Mean of Sale is not Equal in all Age groups



Sales distribution across Age Segements

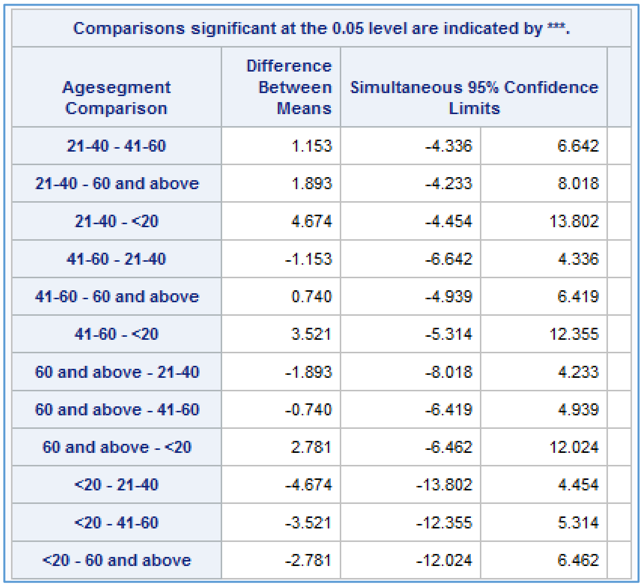
The ANOVA Procedure

Dependent Variable: Sales

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	123228	41076	0.75	0.5216
Error	86605	4736746122	54694		
Corrected Total	86608	4736869350			

R-Square	Coeff Var	Root MSE	Sales Mean
0.000026	129.0824	233.8668	181.1763

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Agesegment	3	123227.8921	41075.9640	0.75	0.5216

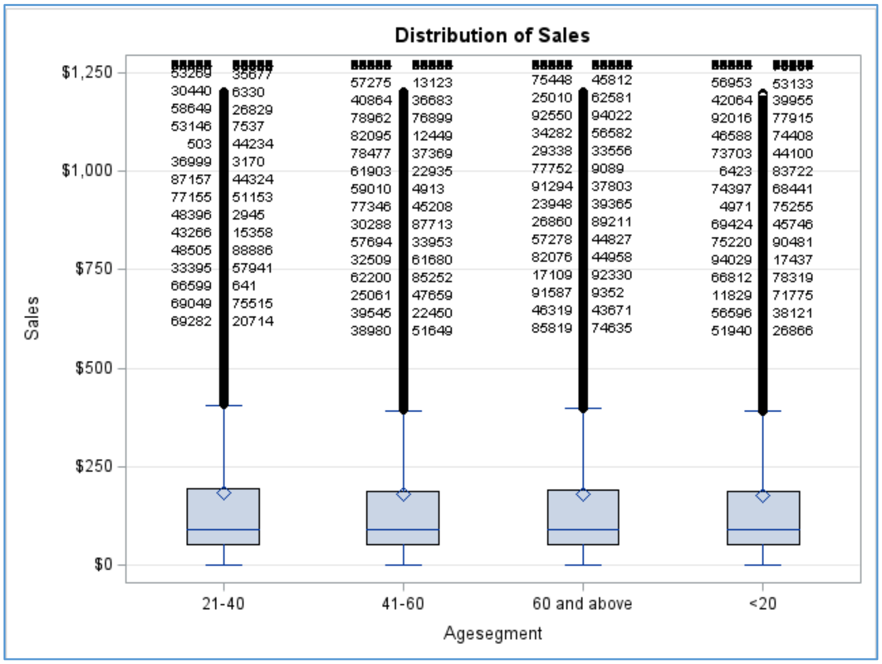


Comparisons significant at the 0.05 level are indicated by ***.

Agesegment Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
21-40 - 41-60	1.153	-4.336	6.642	
21-40 - 60 and above	1.893	-4.233	8.018	
21-40 - <20	4.674	-4.454	13.802	
41-60 - 21-40	-1.153	-6.642	4.336	
41-60 - 60 and above	0.740	-4.939	6.419	
41-60 - <20	3.521	-5.314	12.355	
60 and above - 21-40	-1.893	-8.018	4.233	
60 and above - 41-60	-0.740	-6.419	4.939	
60 and above - <20	2.781	-6.462	12.024	
<20 - 21-40	-4.674	-13.802	4.454	
<20 - 41-60	-3.521	-12.355	5.314	
<20 - 60 and above	-2.781	-12.024	6.462	

- F value: the overall f statistic is calculated by using mean square model/mean square error, 41078/56494 =0.75
  F value: It is the ratio of mean square model/mean square error, it is used to determine the variance explained by model is significantly greater than the unexplained variance
- The p value: (>0.05)-this means we fail to reject null hypothesis and the mean is statistically equal between the groups.
- Sceffe's results indicate very small difference between each age segments
- R squared: It is the proportion of variance in Age explained by the model. 0.000026 percent shows that variability of Sales can not be explained by Age segments, it is a low, so our model is not sufficient to explain the variability of Sales using Age segments

`



- Diagram – Grouped Box Plot- The Total Length of Boxplot or Interquartile range (Distance between Q1 and Q3) is similar for all 4 age segments. We can see they line up and Diamond that represents the mean shows that the mean of all age segments is almost at similar level.
- There are extreme outliers in all age segments . Outliers are greater than upper outer fence (Q3+3IQR)

Conclusion-

➢ Since P values 0.5216>0.05 ,we fail to reject the Null Hypothesis at 5% Significance Level

➢ Mean of Sales is significantly similar in all age segments

➢ Age segment explains 0% of variability of Sales. Therefore, we can say this is not a good model.

B) <u>Sales Distribution between active and deactivated customers</u>

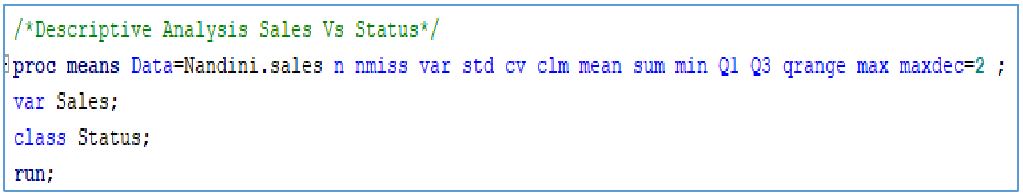
H0- Means of Sales is equal in both groups i.e Sales equally distributed in both Active and Deactivated customers group.

H1- Means of Sales is not equal in both groups i.e Sales not equally distributed in both Active and Deactivated customers group.

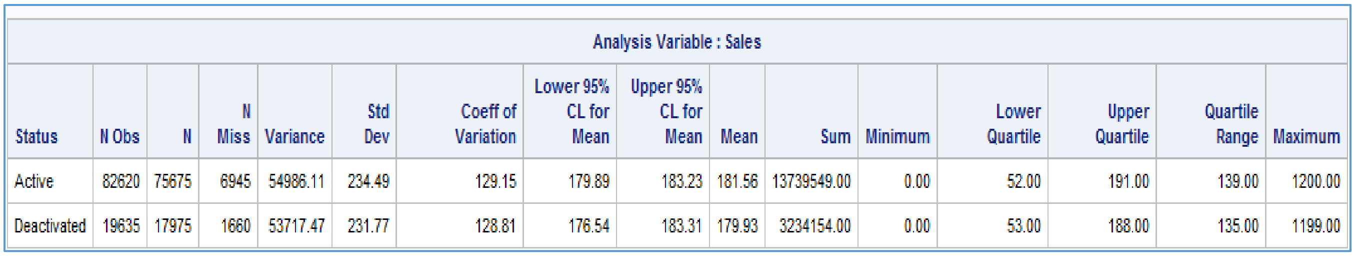
Approach- since Sales is Numerical variable and Status -Active/Deactivated is a categorical Variable we will use:-

- For Summarisation- Proc Means
- For Normality-Proc Univariate
- For Visualisation- Grouped Box Plot
- For Independency- Proc T test

`



```
/*Descriptive Analysis Sales Vs Status*/
proc means Data=Nandini.sales n nmiss var std cv clm mean sum min Q1 Q3 qrange max maxdec=2 ;
var Sales;
class Status;
run;
```

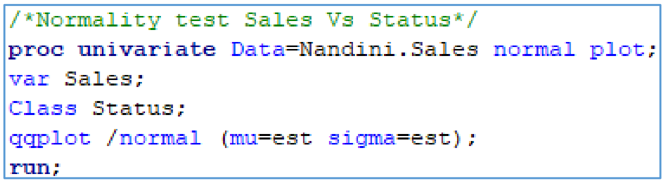


Analysis Variable : Sales

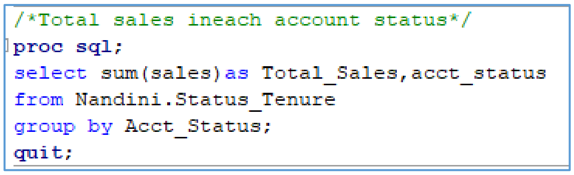
Status	N Obs	N	N Miss	Variance	Std Dev	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean	Mean	Sum	Minimum	Lower Quartile	Upper Quartile	Quartile Range	Maximum
Active	82620	75675	6945	54986.11	234.49	129.15	179.89	183.23	181.56	13739549.00	0.00	52.00	191.00	139.00	1200.00
Deactivated	19635	17975	1660	53717.47	231.77	128.81	176.54	183.31	179.93	3234154.00	0.00	53.00	188.00	135.00	1199.00

- From the results of Means Procedure, we see that Mean and standard deviation of Active and Deactivated customers are almost with minimum difference.
- To statistically prove this we will use normality test using Proc Univariate, Homoscedasticity test using Proc GLM and finally Proc Ttest to prove means are equal in both groups

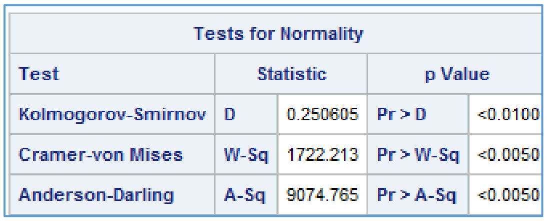
  - Test of Normality-
    H0- Sales is normally distributed
    H1- Sales is not normally distributed



```
/*Normality test Sales Vs Status*/
proc univariate Data=Nandini.Sales normal plot;
var Sales;
Class Status;
qqplot /normal (mu=est sigma=est);
run;
```



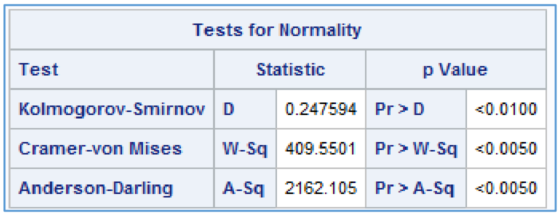
```
/*Total sales ineach account status*/
proc sql;
select sum(sales)as Total_Sales,acct_status
from Nandini.Status_Tenure
group by Acct_Status;
quit;
```



Tests for Normality

Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.250605	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1722.213	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	9074.765	Pr > A-Sq	<0.0050

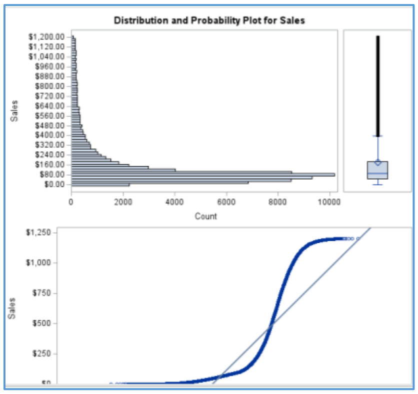
Status-Active

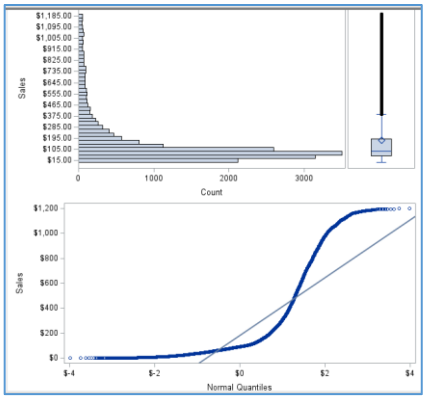


Tests for Normality

Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.247594	Pr > D	<0.0100
Cramer-von Mises	W-Sq	409.5501	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2162.105	Pr > A-Sq	<0.0050

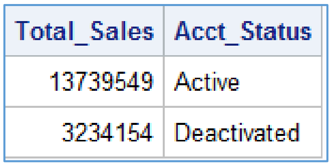
Status- Deactivated





`

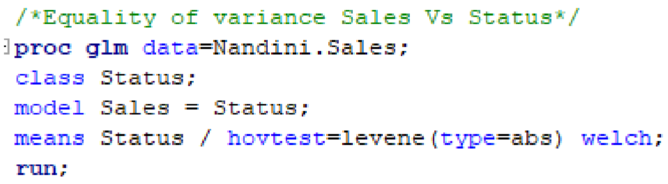
- P value in Kolmogorov-Smirnov of Normality is less than 0.05 significance level.
- So we reject Null Hypothesis of normality and conclude that Sales is not normally distributed
- ,However as per CLT, since sample size is more than>30, we can assume Sales is normally distributed
- Total Sales for Active customers is $13,739,549 while for Deactivated customers is $32,34,154



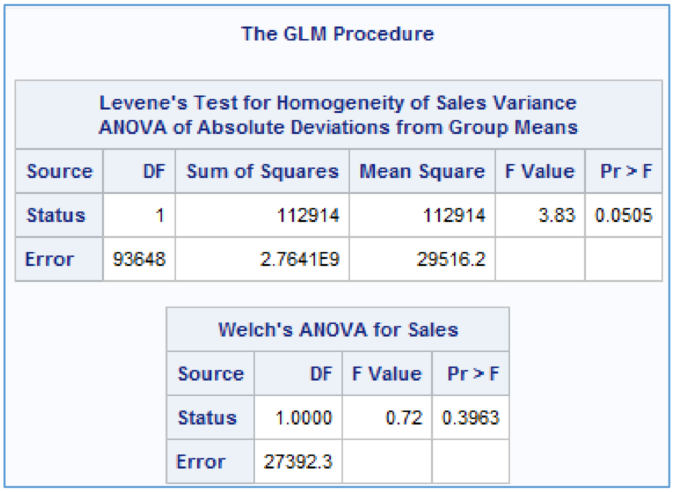
Total_Sales	Acct_Status
13739549	Active
3234154	Deactivated

- Test of Homoscedasticity for equality of variance-

  H0- Variance of Sales is equal in both Groups
  H1- Variance of Sales is not equal in both Groups



```
/*Equality of variance Sales Vs Status*/
proc glm data=Nandini.Sales;
class Status;
model Sales = Status;
means Status / hovtest=levene(type=abs) welch;
run;
```



The GLM Procedure

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

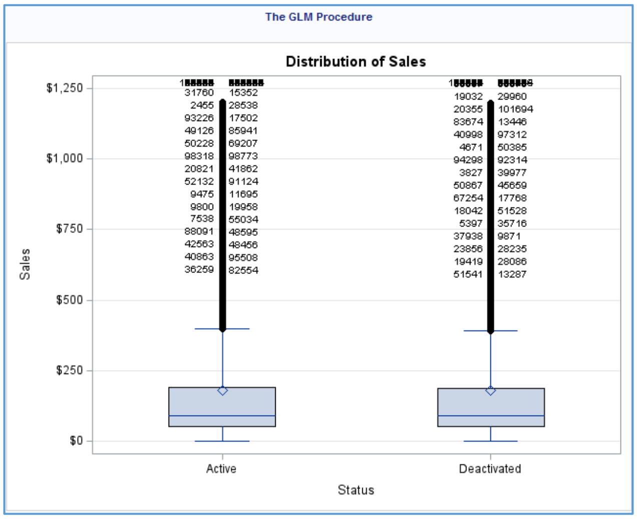
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Status	1	112914	112914	3.83	0.0505
Error	93648	2.7641E9	29516.2		

Welch's ANOVA for Sales

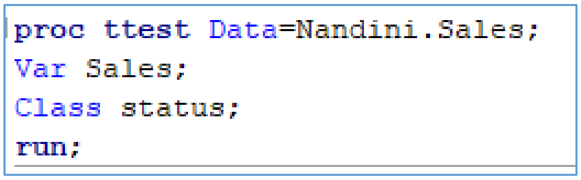
Source	DF	F Value	Pr > F
Status	1.0000	0.72	0.3963
Error	27392.3		

P value in Levene's Test for Equality of Variance 0.0505 is greater than 0.05 significance level.

Therefore, we fail to reject Null Hypothesis and Conclude the variance of Sales is equal in both active and deactivated status

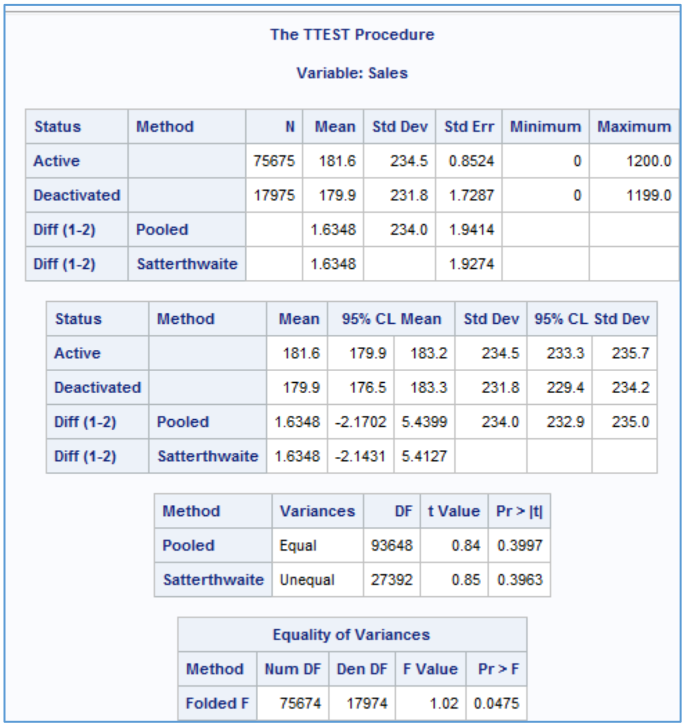


- Test of difference-
  H0- Mean of Sales is Equal in both Active and Deactivated groups
  H1- Mean of Sales is not Equal in both Active and Deactivated groups



```
proc ttest Data=Nandini.Sales;
Var Sales;
Class status;
run;
```

`



The TTEST Procedure

Variable: Sales

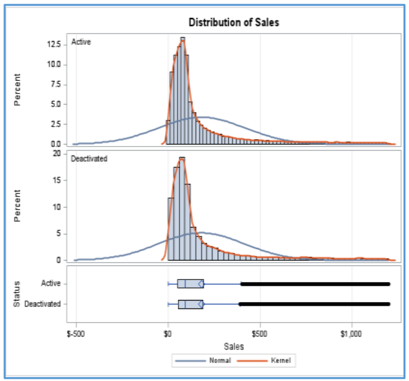
Status	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Active		75675	181.6	234.5	0.8524	0	1200.0
Deactivated		17975	179.9	231.8	1.7287	0	1199.0
Diff (1-2)	Pooled		1.6348	234.0	1.9414		
Diff (1-2)	Satterthwaite		1.6348		1.9274		

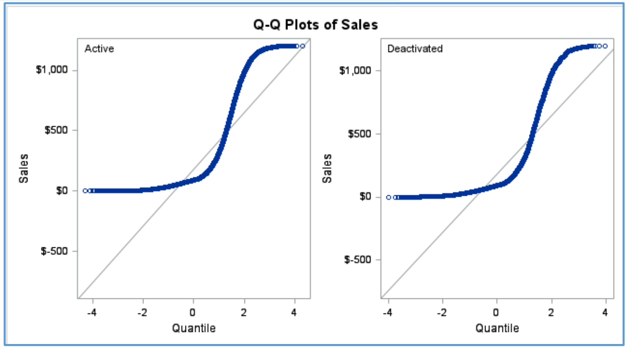
Status	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Active		181.6	179.9	183.2	234.5	233.3	235.7
Deactivated		179.9	176.5	183.3	231.8	229.4	234.2
Diff (1-2)	Pooled	1.6348	-2.1702	5.4399	234.0	232.9	235.0
Diff (1-2)	Satterthwaite	1.6348	-2.1431	5.4127			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	93648	0.84	0.3997
Satterthwaite	Unequal	27392	0.85	0.3963

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	75674	17974	1.02	0.0475





- Diagram –Grouped Box Plot-
  The Total Length Boxplot or Interquartile range (Distance between Q1 and Q3) is very similar for Active and Deactivated groups. This is what we had expected. The groups have similar variance. We can see they line up and Diamond that represents the mean is aligned equally as well. There are more outliers both group . This are extreme outliers (greater than Q3+3IQR- upper outer fence)
- Diagram Histogram and Density plot -The plots compare the distribution of Sales with 2 categories of Active and Deactivated customers. From Histogram, we can see the distribution of Sales for both categories of status is not symmetric or bell shaped, it is right skewed.so we can say the distribution is not normal.

  .
- T- Test-
  H0- Means of Sales is equal in both groups i.e Sales equally distributed in both Active and Deactivated customers group.
  H1- Means of Sales is not equal in both groups i.e Sales not equally distributed in both Active and Deactivated customers group.
- The First table in ttest contain the valid sample size, mean, standard deviation, min and max. In this case the mean of Sales- has very minimum difference between two groups. We need to check this difference is statistically or happens by chance. For this we need to check second table that has Two other Test Pooled and Satterwaite .
- Pooled Test assumes that both groups have the same variance in Age whereas Satterwaite test does not make this assumption.
- It can be done by checking the last table that it is folded f test:
  - Folded f test hypothesis:H0 is : Variance are equal
  - H1 is- Variance are not equal

`

- p value for f test it is 0.0475 < 0.05 so we reject null hypothesis and we will say the variance are not equal. Therefore, we will the see results from Satterwaite Test.

Conclusion-

Since P values in Satterwaite Test 0.3963 >0.05 ,

- we fail to reject the Null Hypothesis at 5% Significance Level
- Average Sales ge of Active customers is equal to average Sales of Deactivated customers
- Mean of Sales is equal in both Active Deactivated customer groups

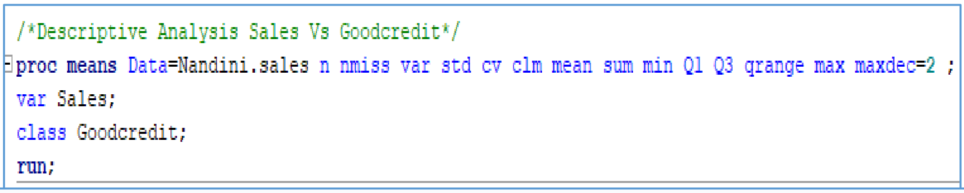
C) <u>Sales Distribution between Good Credit Category</u>

H0- Means of Sales is equal in both groups i.e Sales equally distributed in both categories of GoodCredit

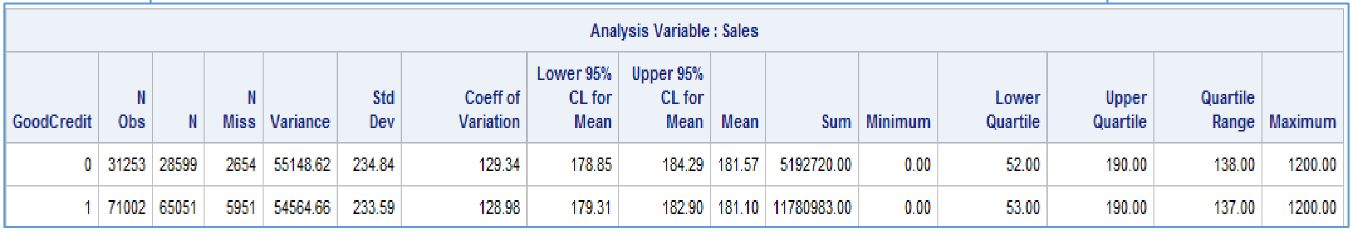
H1- Means of Sales is not equal in both groups i.e Sales not equally distributed in both categories of good credit

Approach- since Sales is Numerical variable and Good Credit Yes/No -is a categorical Variable we will use:-

- For Summarisation- Proc Means
- For Normality-Proc Univariate
- For Visualisation- Grouped Box Plot
- For Independency- Proc T test



```
/*Descriptive Analysis Sales Vs Goodcredit*/
proc means Data=Nandini.sales n nmiss var std cv clm mean sum min Q1 Q3 qrange max maxdec=2 ;
var Sales;
class Goodcredit;
run;
```



Analysis Variable : Sales

GoodCredit	N Obs	N	N Miss	Variance	Std Dev	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean	Mean	Sum	Minimum	Lower Quartile	Upper Quartile	Quartile Range	Maximum
0	31253	28599	2654	55148.62	234.84	129.34	178.85	184.29	181.57	5192720.00	0.00	52.00	190.00	138.00	1200.00
1	71002	65051	5951	54564.66	233.59	128.98	179.31	182.90	181.10	11780983.00	0.00	53.00	190.00	137.00	1200.00

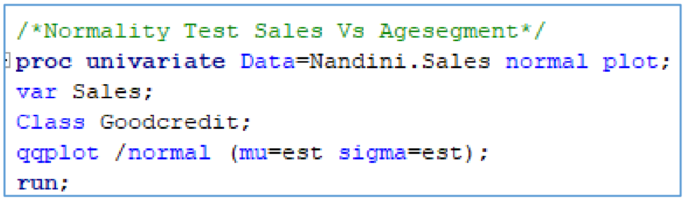
- From the results of Means Procedure, we see that Mean and standard deviation of Sales in Good Credit Yes and Good Credit No category are almost with minimum difference.
- To statistically prove this we will use normality test using Proc Univariate, Homoscedasticity test using Proc GLM and finally Proc Ttest to prove means are equal in both groups

- Test of Normality-

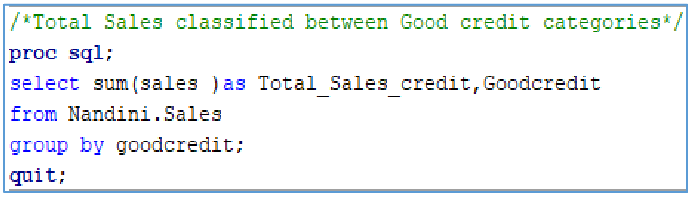
H0- Sales is normally distributed

H1- Sales is not normally distributed

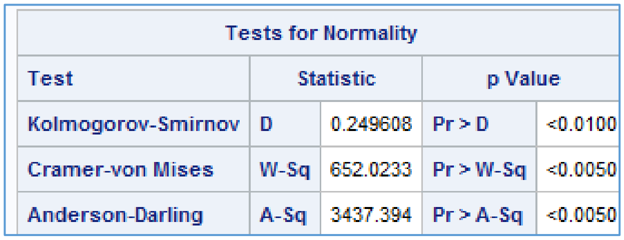
`



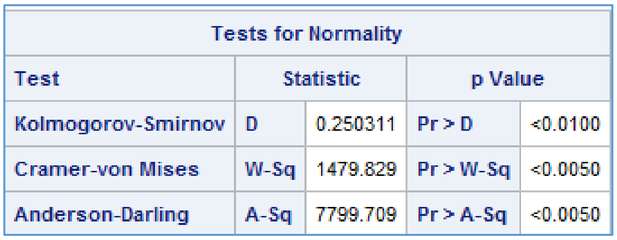
```
/*Normality Test Sales Vs Agesegment*/
proc univariate Data=Nandini.Sales normal plot;
var Sales;
Class Goodcredit;
qqplot /normal (mu=est sigma=est);
run;
```



```
/*Total Sales classified between Good credit categories*/
proc sql;
select sum(sales )as Total_Sales_credit,Goodcredit
from Nandini.Sales
group by goodcredit;
quit;
```



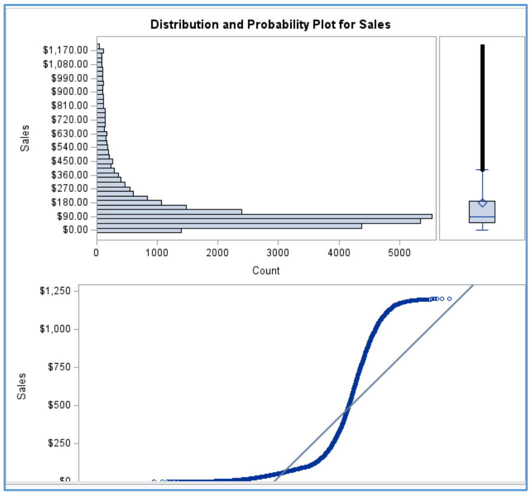
Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.249608	Pr > D	<0.0100
Cramer-von Mises	W-Sq	652.0233	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	3437.394	Pr > A-Sq	<0.0050

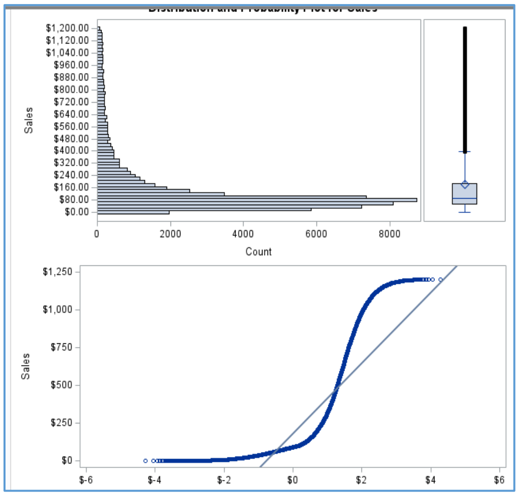


Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.250311	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1479.829	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	7799.709	Pr > A-Sq	<0.0050

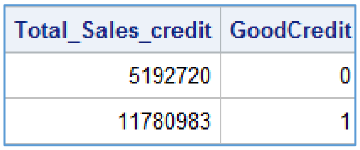
No Good Credit

Has Good Credit





- P value in Kolmogorov-Smirnov of Normality is less than 0.05 significance level.
- So we reject Null Hypothesis of normality and conclude that Sales is not normally distributed
- No Bell Shape distribution observed .Data is highly right skewed
- However as per CLT, since sample size is more than>30, we can assume Sales is normally distributed
- Total Sales for Customers with Good Credit is $1,17,80,983 while Customers without Good credit is $51,92,72

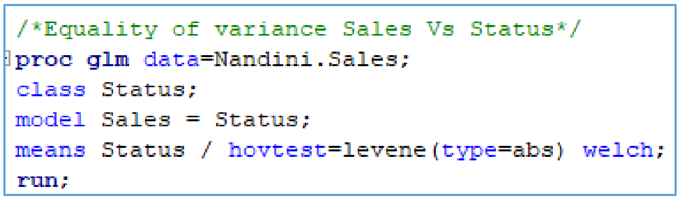


Total_Sales_credit	GoodCredit
5192720	0
11780983	1

- Test of Homoscedasticity for equality of variance-

H0- Variance of Sales is equal in both Groups of customers with Good Credit and Without Good Credit
H1- Variance of Sales is not equal in both Groups of customers with Good Credit and Without Good Credit

`

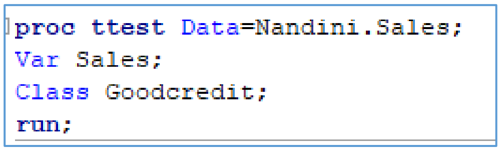


```
/*Equality of variance Sales Vs Status*/
proc glm data=Nandini.Sales;
class Status;
model Sales = Status;
means Status / hovtest=levene(type=abs) welch;
run;
```

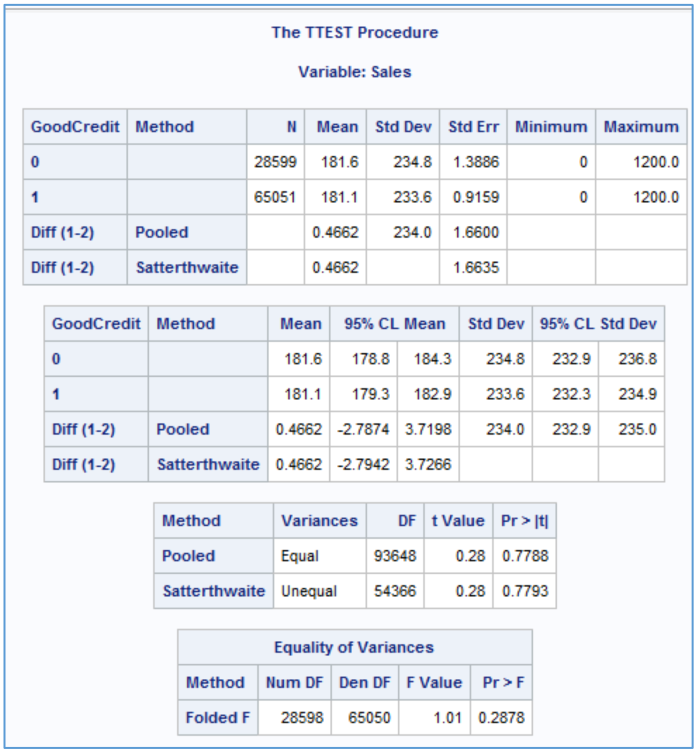
P value in Levene's Test for Equality of Variance 0.6795 is greater than 0.05 significance level.

Therefore, we fail to reject Null Hypothesis and Conclude the variance of Sales is equal in customers with Good Credit and without Good Credit Categories.

- Test of difference-
  H0- Mean of Sales is Equal in both Customer with Good Credit and Without Good Credit groups
  H1- Mean of Sales is not Equal in both Customer with Good Credit and Without Good Credit groups



```
proc ttest Data=Nandini.Sales;
Var Sales;
Class Goodcredit;
run;
```



The TTEST Procedure

Variable: Sales

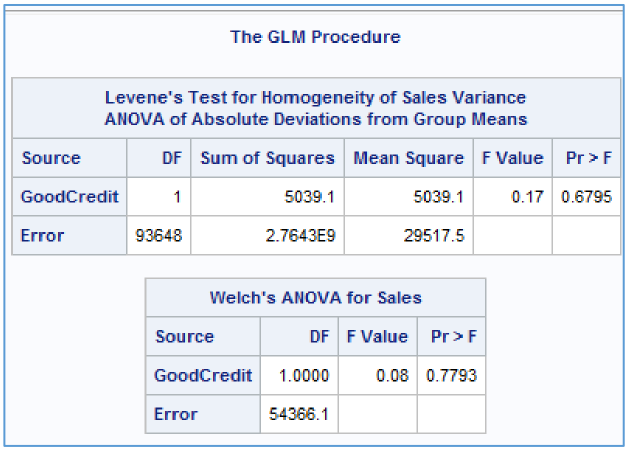
GoodCredit	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		28599	181.6	234.8	1.3886	0	1200.0
1		65051	181.1	233.6	0.9159	0	1200.0
Diff (1-2)	Pooled		0.4662	234.0	1.6600		
Diff (1-2)	Satterthwaite		0.4662		1.6635		

GoodCredit	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		181.6	178.8	184.3	234.8	232.9	236.8
1		181.1	179.3	182.9	233.6	232.3	234.9
Diff (1-2)	Pooled	0.4662	-2.7874	3.7198	234.0	232.9	235.0
Diff (1-2)	Satterthwaite	0.4662	-2.7942	3.7266			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	93648	0.28	0.7788
Satterthwaite	Unequal	54366	0.28	0.7793

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	28598	65050	1.01	0.2878



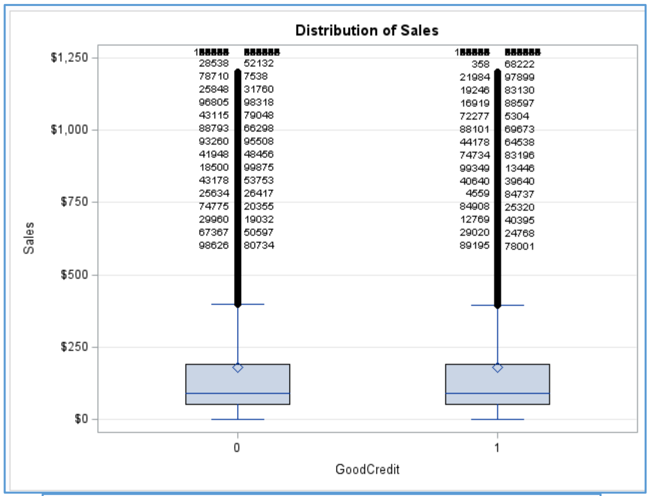
The GLM Procedure

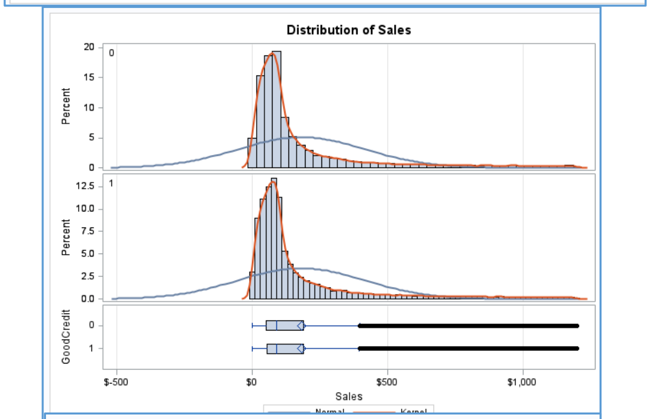
Levene's Test for Homogeneity of Sales Variance
ANOVA of Absolute Deviations from Group Means

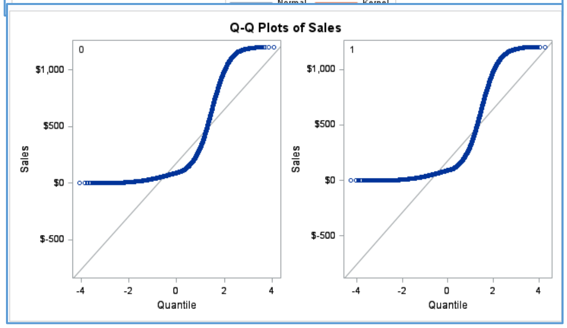
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
GoodCredit	1	5039.1	5039.1	0.17	0.6795
Error	93648	2.7643E9	29517.5		

Welch's ANOVA for Sales

Source	DF	F Value	Pr > F
GoodCredit	1.0000	0.08	0.7793
Error	54366.1		







`

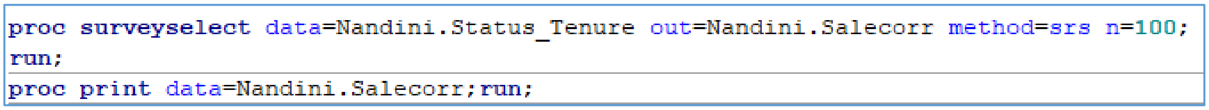
- Diagram –Grouped Box Plot-
  The Total Length Boxplot or Interquartile range (Distance between Q1 and Q3) is very similar for With Good Credit and Without Good Credit groups. This is what we had expected. The groups have similar variance. We can see they line up and Diamond that represents the mean is aligned equally as well. There are more outliers both group . This are extreme outliers (greater than Q3+3IQR- upper outer fence)
- Diagram Histogram and Density plot -The plots compare the distribution of Sales with 2 categories of Customers with and without Good Credit. From Histogram, we can see the distribution of Sales for both categories of Good Credit is not symmetric or bell shaped, it is right skewed.so we can say the distribution is not normal.
  .
- T- Test-
  H0- Means of Sales is equal in both groups i.e Sales equally distributed in both With Good Credit and Without Good Credit Categories.
  H1- Means of Sales is not equal in both groups i.e Sales not equally distributed in both With Good Credit and Without Good Credit Categories.

- The First table in ttest contain the valid sample size, mean, standard deviation, min and max.
- In this case the mean of Sales- has very minimum difference between two groups. We need to check this difference is statistically or happens by chance. For this we need to check second table that has Two other Test Pooled and Satterwaite.
- Pooled Test assumes that both groups have the same variance in Age whereas Satterwaite test does not make this assumption.
- It can be done by checking the last table that it is folded f test:
  - Folded f test hypothesis:H0 is : Variance are equal
  - H1 is- Variance are not equal
- p value for f test it is 0.2878 > 0.05 so we fail to reject null hypothesis and we will say the variance are equal. Therefore, we will the see results from Pooled Test.

Conclusion-
Since P values in Pooled Test 0.7788 >0.05 ,

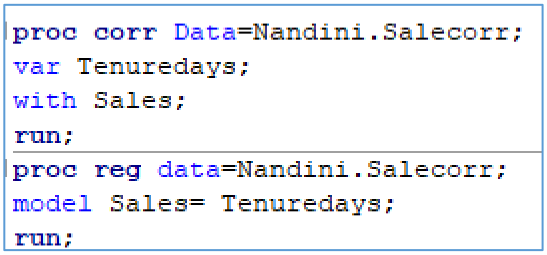
- we fail to reject the Null Hypothesis at 5% Significance Level
- Average Sales of Customers with Good Credit is equal to average Sales of Customers without Good Credit
- Mean of Sales is equal in both With Good Credit and Without Good Credit Categories.

Determine if Tenuredays and Sales are correlated with each Other

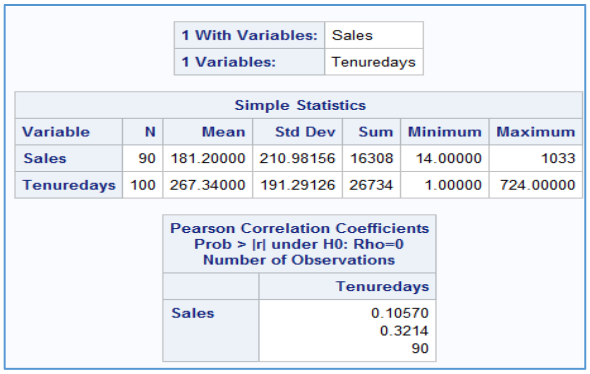


```
proc surveyselect data=Nandini.Status_Tenure out=Nandini.Salecorr method=srs n=100;
run;
proc print data=Nandini.Salecorr;run;
```

`



```
proc corr Data=Nandini.Salecorr;
var Tenuredays;
with Sales;
run;
proc reg data=Nandini.Salecorr;
model Sales= Tenuredays;
run;
```



1 With Variables:	Sales
1 Variables:	Tenuredays

Simple Statistics

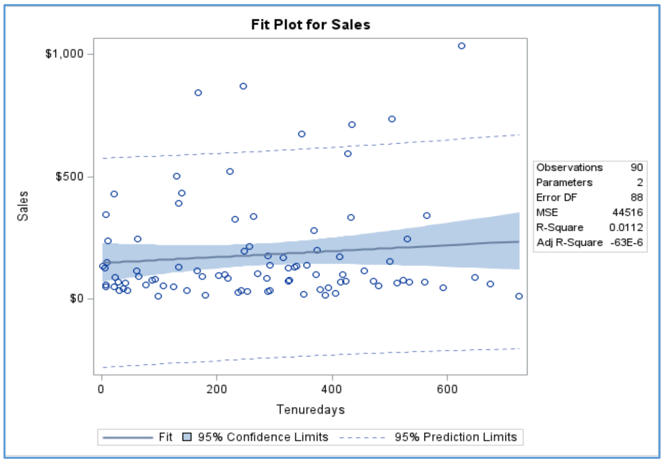
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Sales	90	181.20000	210.98156	16308	14.00000	1033
Tenuredays	100	267.34000	191.29126	26734	1.00000	724.00000

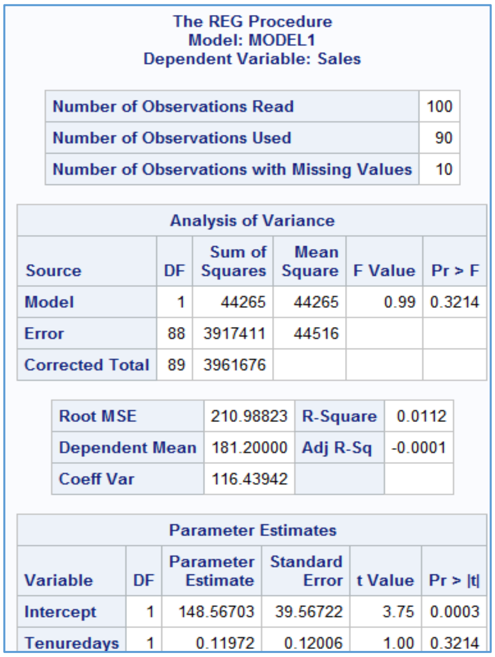
Pearson Correlation Coefficients
Prob > \|r\| under H0: Rho=0
Number of Observations

	Tenuredays
Sales	0.10570 0.3214 90

H0- There is no correlation Between Sales and Tenuredays

H1- There is significant Correlation between Sales and Tenuredays





The REG Procedure
Model: MODEL1
Dependent Variable: Sales

Number of Observations Read	100
Number of Observations Used	90
Number of Observations with Missing Values	10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	44265	44265	0.99	0.3214
Error	88	3917411	44516		
Corrected Total	89	3961676			

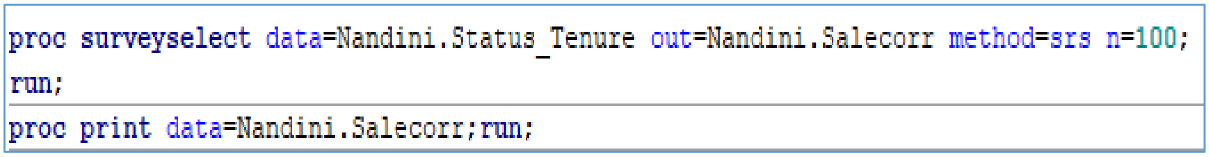
Root MSE	210.98823	R-Square	0.0112
Dependent Mean	181.20000	Adj R-Sq	-0.0001
Coeff Var	116.43942		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	148.56703	39.56722	3.75	0.0003
Tenuredays	1	0.11972	0.12006	1.00	0.3214

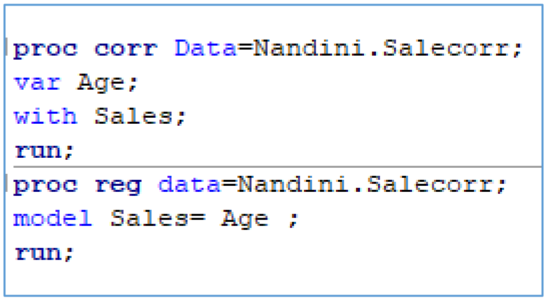
- P value of Tenuredays is 0.3214 >0.05. Therefore, we fail to reject Null Hypothesis and conclude there is no correlation between Sales and Tenuredays.
- R square is very low 0.0112and RMSE is high 210.988
- R Square- 0.0112 indicates only 1% of variation in sales can be explained by Tenuredays.
- Model is not a best fit for the analysis between Sales and Tenuredays

Determine if Age and Sales are correlated with each Other

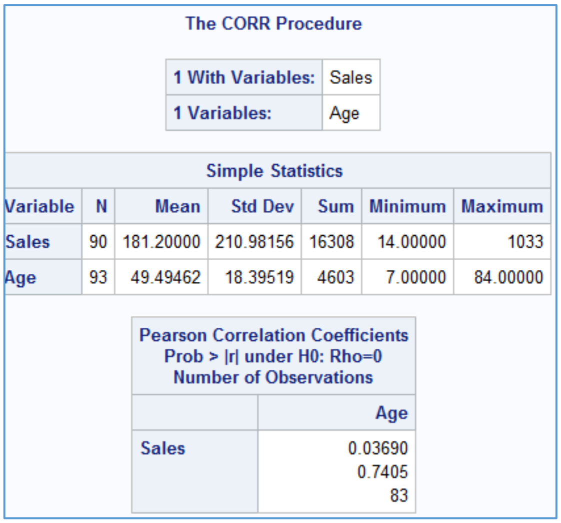


```
proc surveyselect data=Nandini.Status_Tenure out=Nandini.Salecorr method=srs n=100;
run;
proc print data=Nandini.Salecorr;run;
```

`



```
proc corr Data=Nandini.Salecorr;
var Age;
with Sales;
run;
proc reg data=Nandini.Salecorr;
model Sales= Age ;
run;
```



The CORR Procedure

1 With Variables:	Sales
1 Variables:	Age

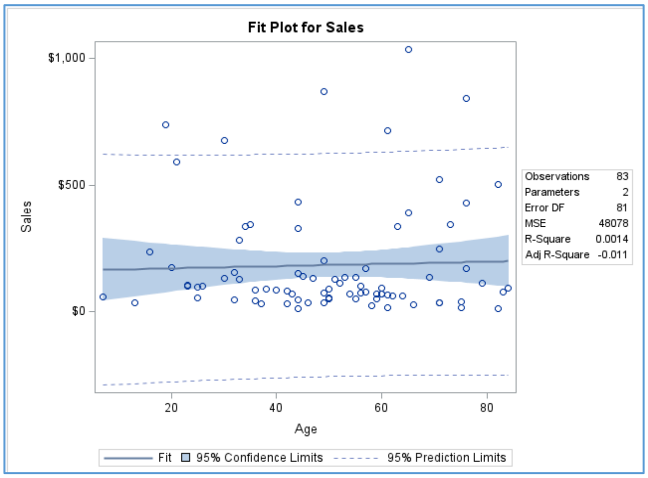
Simple Statistics

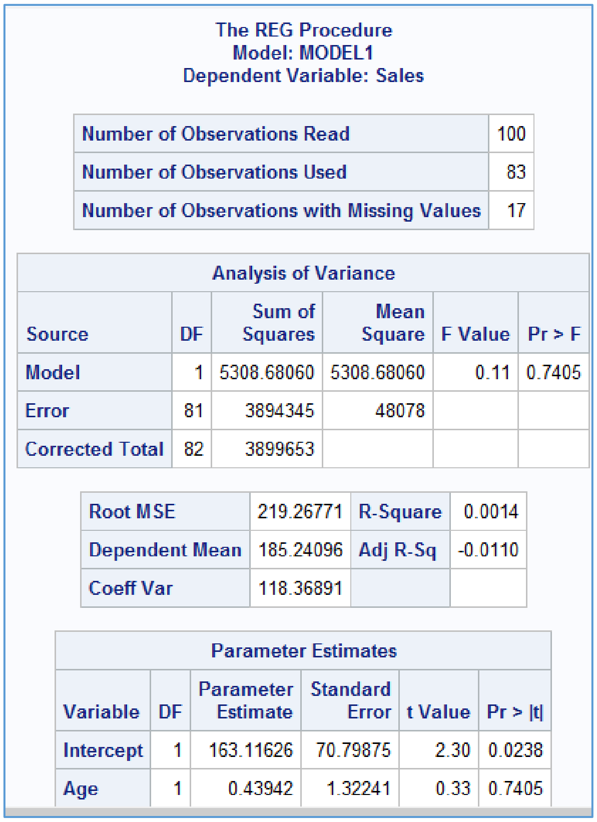
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Sales	90	181.20000	210.98156	16308	14.00000	1033
Age	93	49.49462	18.39519	4603	7.00000	84.00000

Pearson Correlation Coefficients
Prob > \|r\| under H0: Rho=0
Number of Observations

	Age
Sales	0.03690 0.7405 83

H0- There is no correlation Between Sales and Age
H1- There is significant Correlation between Sales and Age





The REG Procedure
Model: MODEL1
Dependent Variable: Sales

Number of Observations Read	100
Number of Observations Used	83
Number of Observations with Missing Values	17

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5308.68060	5308.68060	0.11	0.7405
Error	81	3894345	48078		
Corrected Total	82	3899653			

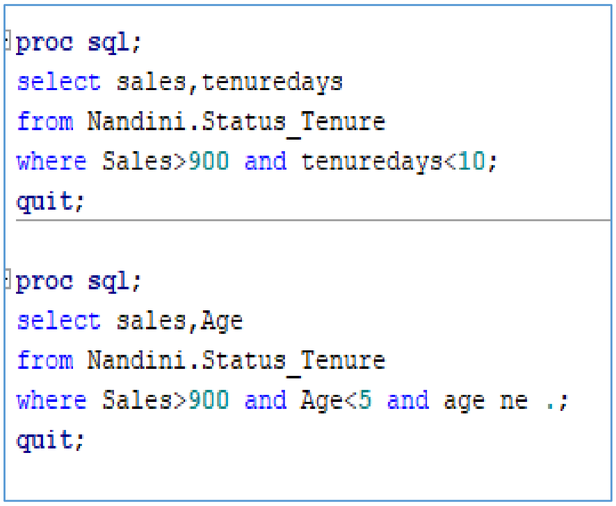
Root MSE	219.26771	R-Square	0.0014
Dependent Mean	185.24096	Adj R-Sq	-0.0110
Coeff Var	118.36891		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	163.11626	70.79875	2.30	0.0238
Age	1	0.43942	1.32241	0.33	0.7405

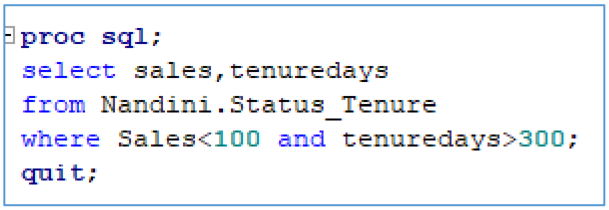
- P value of Tenuredays is 0.7405 >0.05. Therefore, we fail to reject Null Hypothesis and conclude there is no correlation between Sales and Age.
- R square is very low 0.0014and RMSE is high 219.27
- R Square- 0.0014 indicates almost 0% of variation in sales can be explained by Age
- Therefore Model is not a best fit for the analysis between Sales and Age

`

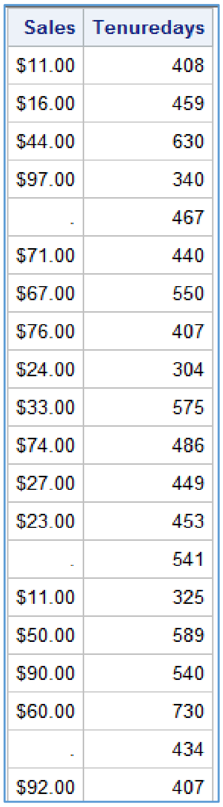


```
proc sql;
select sales,tenuredays
from Nandini.Status_Tenure
where Sales>900 and tenuredays<10;
quit;

proc sql;
select sales,Age
from Nandini.Status_Tenure
where Sales>900 and Age<5 and age ne .;
quit;
```

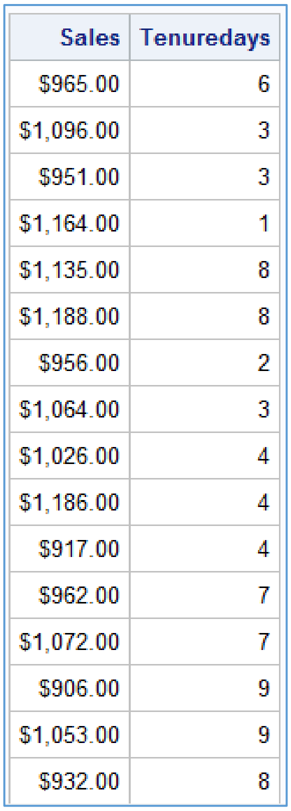
- Sale is missing or very low for tenure more than 1 year
- Sale is very high for even 1 days
- Sale is very high for age of 1 year
- Most places data looks unrealistic More accurate information needed to make this model perfect



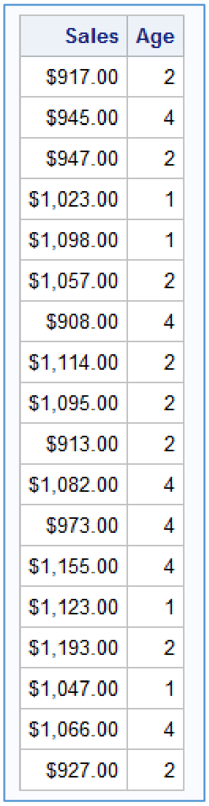
```
proc sql;
select sales,tenuredays
from Nandini.Status_Tenure
where Sales<100 and tenuredays>300;
quit;
```



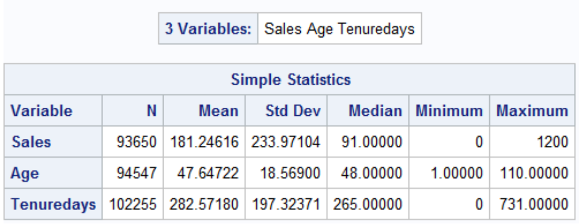
Sales	Tenuredays
$11.00	408
$16.00	459
$44.00	630
$97.00	340
.	467
$71.00	440
$67.00	550
$76.00	407
$24.00	304
$33.00	575
$74.00	486
$27.00	449
$23.00	453
.	541
$11.00	325
$50.00	589
$90.00	540
$60.00	730
.	434
$92.00	407



Sales	Tenuredays
$965.00	6
$1,096.00	3
$951.00	3
$1,164.00	1
$1,135.00	8
$1,188.00	8
$956.00	2
$1,064.00	3
$1,026.00	4
$1,186.00	4
$917.00	4
$962.00	7
$1,072.00	7
$906.00	9
$1,053.00	9
$932.00	8



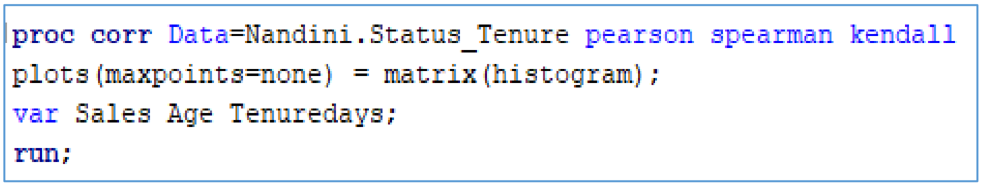
Sales	Age
$917.00	2
$945.00	4
$947.00	2
$1,023.00	1
$1,098.00	1
$1,057.00	2
$908.00	4
$1,114.00	2
$1,095.00	2
$913.00	2
$1,082.00	4
$973.00	4
$1,155.00	4
$1,123.00	1
$1,193.00	2
$1,047.00	1
$1,066.00	4
$927.00	2



3 Variables: Sales Age Tenuredays

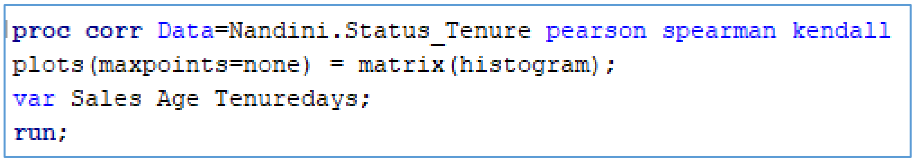
Simple Statistics

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
Sales	93650	181.24616	233.97104	91.00000	0	1200
Age	94547	47.64722	18.56900	48.00000	1.00000	110.00000
Tenuredays	102255	282.57180	197.32371	265.00000	0	731.00000

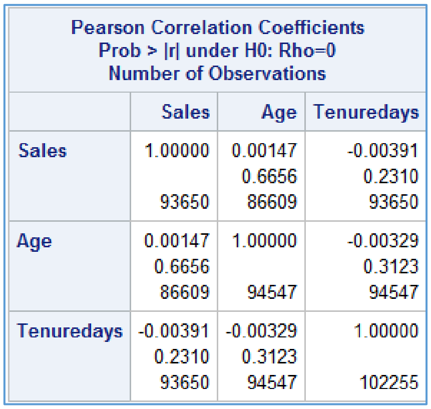


```
proc corr Data=Nandini.Status_Tenure pearson spearman kendall
plots(maxpoints=none) = matrix(histogram);
var Sales Age Tenuredays;
run;
```

`

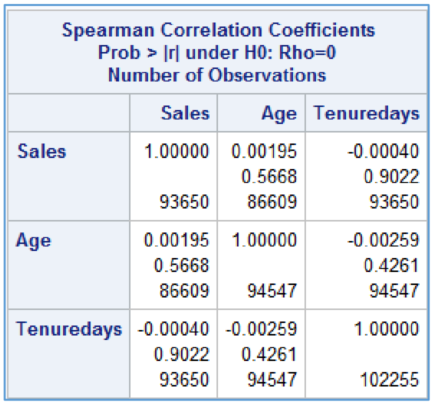


```
proc corr Data=Nandini.Status_Tenure pearson spearman kendall
plots(maxpoints=none) = matrix(histogram);
var Sales Age Tenuredays;
run;
```



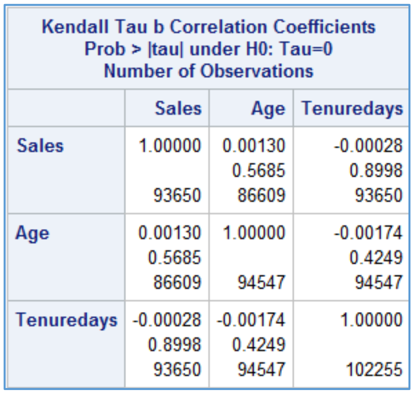
Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Sales	Age	Tenuredays
Sales	1.00000 93650	0.00147 0.6656 86609	-0.00391 0.2310 93650
Age	0.00147 0.6656 86609	1.00000 94547	-0.00329 0.3123 94547
Tenuredays	-0.00391 0.2310 93650	-0.00329 0.3123 94547	1.00000 102255



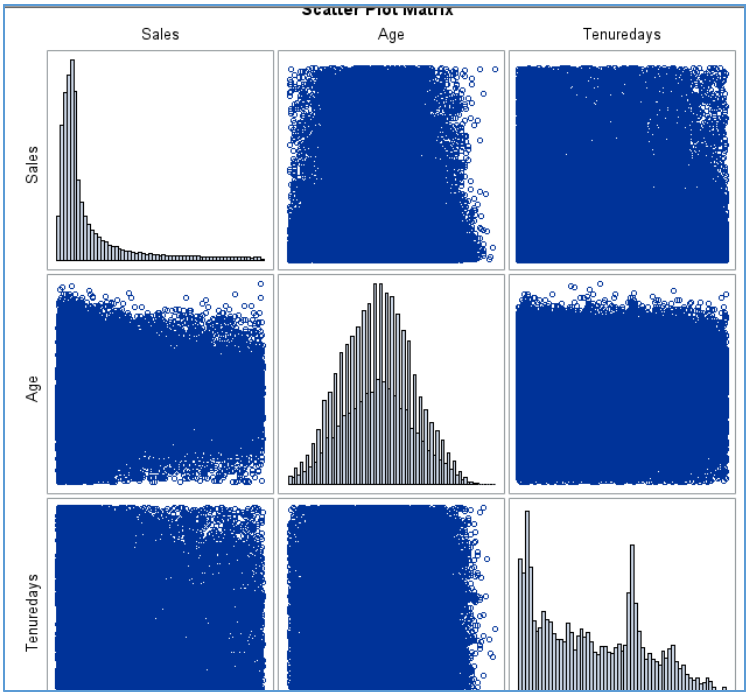
Spearman Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Sales	Age	Tenuredays
Sales	1.00000 93650	0.00195 0.5668 86609	-0.00040 0.9022 93650
Age	0.00195 0.5668 86609	1.00000 94547	-0.00259 0.4261 94547
Tenuredays	-0.00040 0.9022 93650	-0.00259 0.4261 94547	1.00000 102255



Kendall Tau b Correlation Coefficients
Prob > |tau| under H0: Tau=0
Number of Observations

	Sales	Age	Tenuredays
Sales	1.00000 93650	0.00130 0.5685 86609	-0.00028 0.8998 93650
Age	0.00130 0.5685 86609	1.00000 94547	-0.00174 0.4249 94547
Tenuredays	-0.00028 0.8998 93650	-0.00174 0.4249 94547	1.00000 102255



- No Pattern or linear correlation found between any of 3 independent variables.

`

FINDINGS

- Age and Sales has Majority of Outliers, Missing Data, Unrealistic figures
- ON has maximum Customer Base where as QC has Minimum
- Count of Age<20 (Young Customers) is very low
- 52% customers has taken service of less than $100.Only 4 % customers have taken service of more than $800.
- Maximum Deactivation Occurred in Winters. Reason can be interrupted service, outage due to harsh weather
- 340 customers have deactivated service on the same day
- Customers with greater Tenure have good credit
- Majority active customers prefer Rate plan 1 .Rate plan 2 has minimum customer churn. Rate Plan 3 has maximum customer churn
- Dealer type A1 has maximum customer base.
- Customer churn is seen more with tenure <30 days.(33%)
- Customer churn went increasing after March till year end .
- Sales is Maximum in age group 41-60 and minimum in <20.It indicates young people tend to prefer better deals by competitors
- Sales is maximum for active customers than Deactivated Customers.
- Maximum Sales is from the customers with Good Credit
- Model is not best fit to explain sales with the help of Tenure and Age data.

RECCOMENDATIONS

- Need More accurate information for effective analysis.
- More promotions needed in QC
- Good Promotions, offers, deals need to be arranged to attract Young crowd
- Need to reach out to customer for their feedback and understand their need. Offer better solutions. Initiate Rewards for loyal customers.
- Need to find way outs to avoid interruption in winters.
- Deactivation on same day is a major point of concern.Need investigation
- Arrange Loyalty rewards for Good Credit Customers
- Rate Plan 2 can be improved to match rate plan 1.Rate plan 3 needs attention.
- Other Dealers need to match the service of Dealer A1
- Provide best service to avoid losing customers in first month.
- Competitor analysis required .
- Age and Sales figures need to be rechecked .
- Overall the available data is not sufficient to make strong conclusions regarding sale and customer churn. Details like customer income, customer feedback may add value to the analysis.

`