**A close-up of a sign

Description automatically generated**

**INTRODUCTION**

Now a days telecommunications industries has become very popular with large number of operators around the world. This has led to an increased level of competition between the providers. It is extremely hard for companies to survive in this competitive market. They need to implement multiple marketing strategies depending on the customer behavioral pattern. Some of the approaches these companies implement to generate more revenue are to getting more new customers, holding on to existing customers with promotions and other benefits to make sure they don’t leave the company searching for another provider. Customer churn is one of the most important problem every telecom company faces nowadays. Prediction of potential customer churn is very important because most of the time the costs related with acquiring a new customer is much greater when compared to retaining an existing customer. Therefore, by predicting potential customers who are more likely to churn from the company on an early stage will prove to be more profitable and there by increasing the overall revenue of the company. Good Customer Retention Strategy is needed by the companies to minimize customer attrition. This can be achieved by investigating and analyzing the customer distribution and behavioral pattern to get a proper understanding of the customer and the factors that influence customer churn.

**UNDERSTANDING BUSINESS**

The data used in this project contains a CRM data of wireless company for 2 years starting from January 1999 to January 2001. Each row represents a customer with a ClientID, each column contains the customer’s attributes such as the date when the customer activated his account with the company, If the customer is still active with the company or has the customer deactivated his account and the reason for de activation Customer credit status , age, demographic information and sales customer has had during his active period with the company In this project, we will analyze customer level data of a telecom company, investigate their behavioral pattern to identify customers at high risk of churn and recognize the main indicators or reason for their churn. This behavioral analysis will help the company to categorize customers and design strategies that will in turn help them to minimize customer attrition , retain their most valuable customers and help to forecast the deactivation trends for the next 6 months.

**BUSINESS QUESTIONS**

* What is the number of accounts activated and deactivated? When is the earliest and
* latest activation/deactivation dates available?
* What is the age and province distributions of active and deactivated customers?
* What is the tenure in days for each account ?
* How many accounts are deactivated each month in total? Compare it with the accounts getting deactivate by month each year. Is there any increase in customer attrition over the years ?
* Is there any association between tenure and credit status , Rate plan and Dealer Type?
* Is there any association between the account status and the tenure segments?
* Does Sales amount differ among different account status, GoodCredit, and customer age segments?

**ATTRIBUTES**

Acctno: Account number.

Actdt: Account activation date

Deactdt: Account deactivation date

DeactReason: Reason for deactivation.

GoodCredit: Customer’s credit (Good or Bad)

RatePlan: Rate plan for the customer.

DealerType: Dealer type.

Age: Customer age.

Province: Province.

Sales: The amount of sales to a customer

**IMPORTING AND GETTING FAMILIAR WITH DATA**

**A screenshot of a data sheet

Description automatically generated**

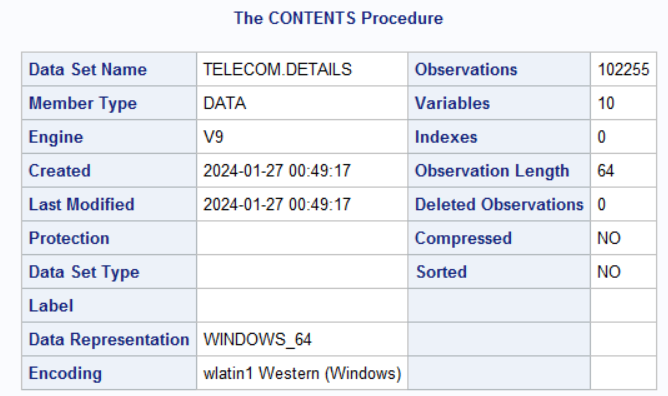
**Analysis Requests:**

**Question 1.1 Explore and describe the dataset briefly. For example, is the acctno unique? What is the number of accounts activated and deactivated? When is the earliest and latest activation/deactivation dates available?**

* 1. **a) Explore and describe the dataset briefly**

**A close-up of a computer screen

Description automatically generated**

**A screenshot of a table

Description automatically generated**

There are 102255 observations and 10 variables in the given dataset

**FINDING HEAD OF DATA FOR 10 OBSERVATION**

A close-up of a computer screen

Description automatically generated

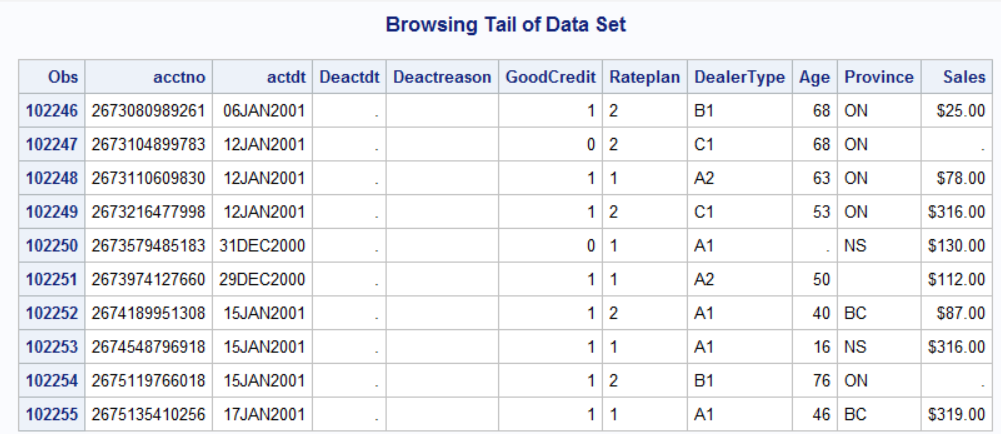
**A screenshot of a data set

Description automatically generated**

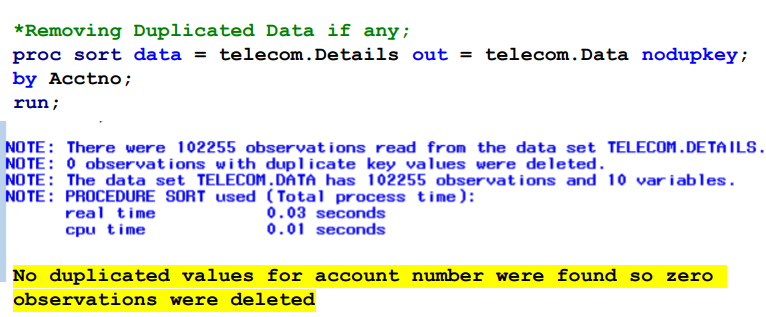
**FINDING TAIL OF DATA FOR 10 OBSERVATION**

**A close-up of a computer screen

Description automatically generated**

****

* 1. **(b) Finding if all the account numbers are unique?**

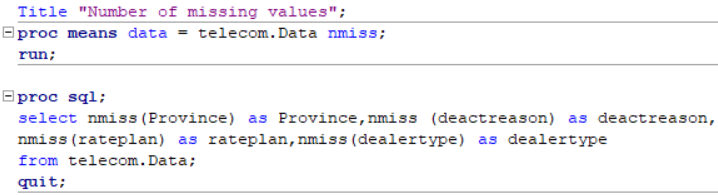


**Number of unique accounts**

**A screenshot of a computer

Description automatically generated**

**Checking For Missing Values in the Dataset**

****

**A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated**

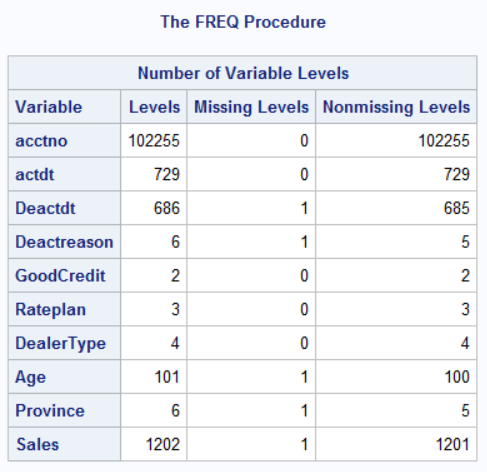
Here missing values in the Deactivation Date is considered that those accounts are still active with the company.

**Finding Number Of Unique levels in each Variables**

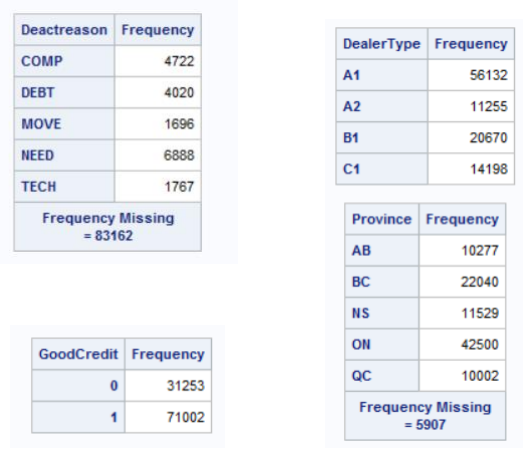
Number Of Unique/Distinct Values In All Variable

A close-up of a text

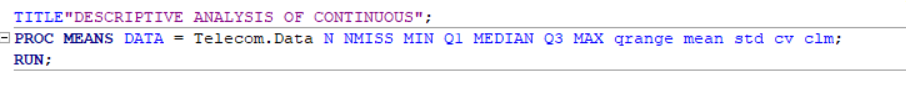
Description automatically generated

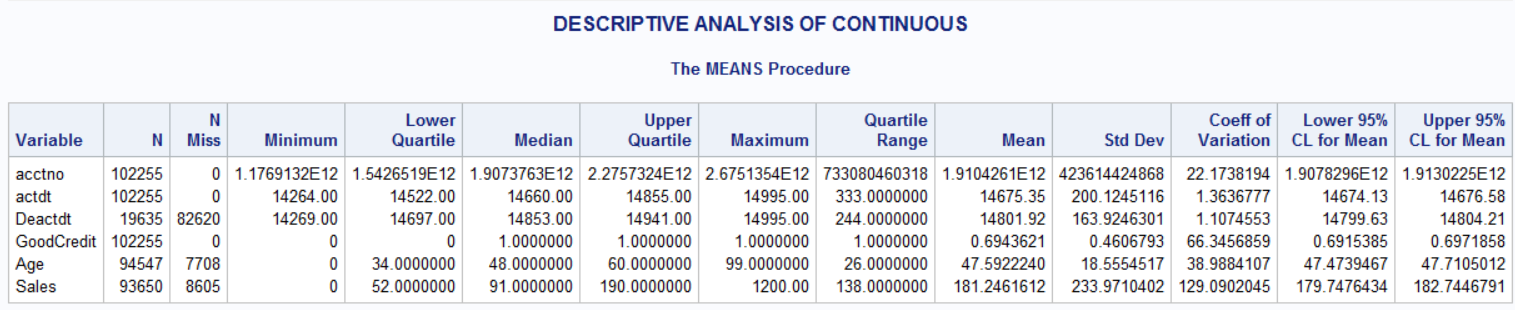
****

**Number Of Unique value in each Variables Group**

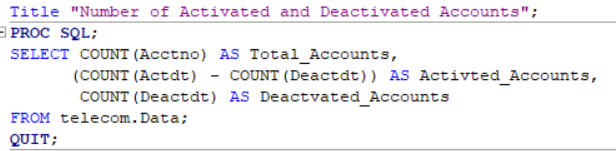
****

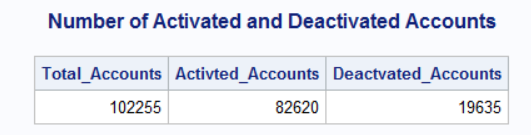
**DESCRIPTIVE ANALYSIS OF CONTINOUS VARIABLES**

****

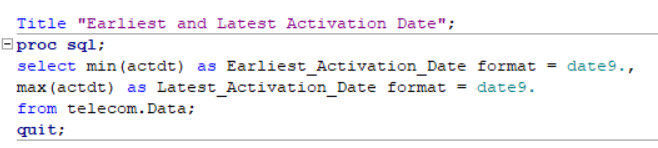
****

* 1. **(c) What is the number of accounts activated and deactivated?**

****

****

**When is the earliest and latest activation dates available?**

****

**A screenshot of a facebook page

Description automatically generated**

**When is the earliest and latest deactivation dates available?**

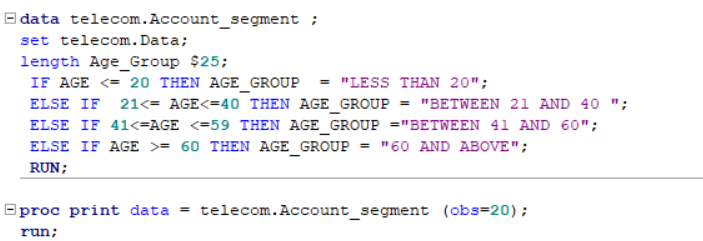
**A screen shot of a computer

Description automatically generated**

**A screenshot of a computer

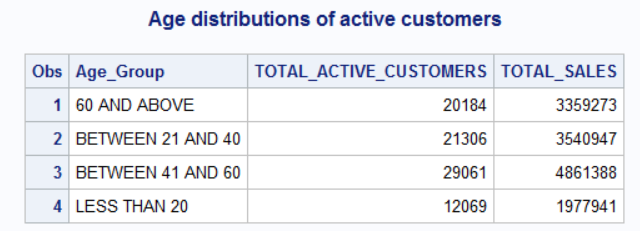
Description automatically generated**

* 1. **What is the age and province distributions of active customers ?**

****

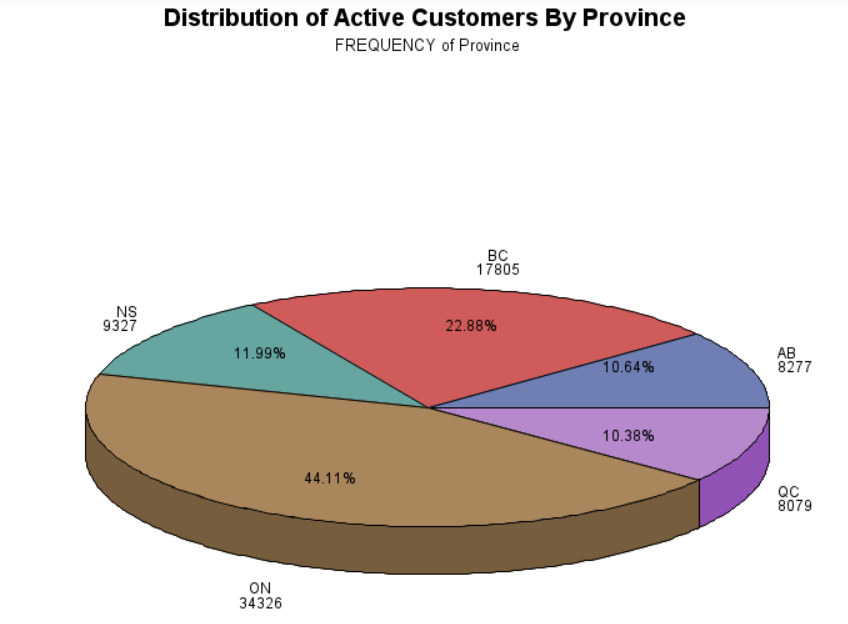
**A screenshot of a computer code

Description automatically generated**

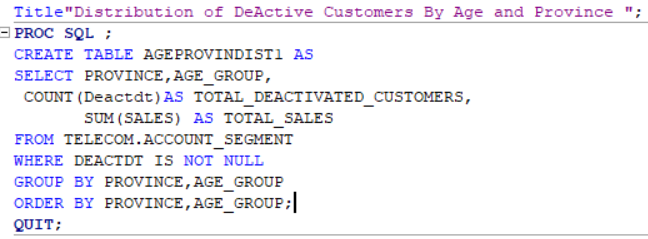
****

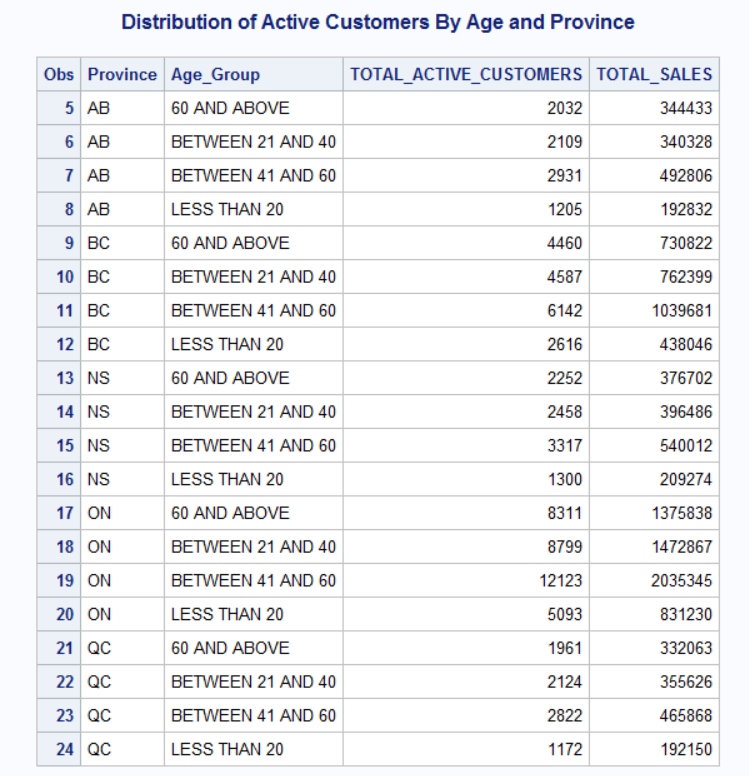
**A pie chart with numbers and a few percentages

Description automatically generated**

****

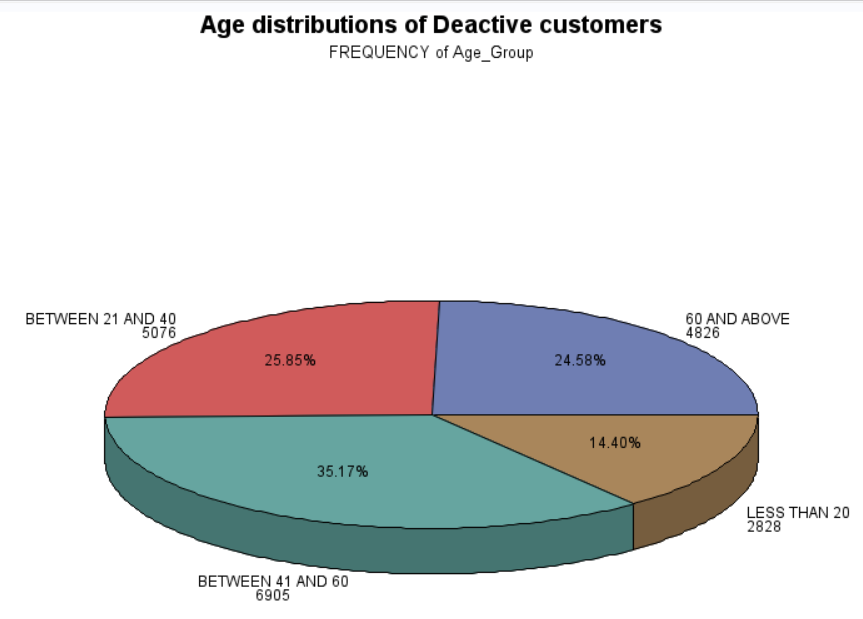
**1.2 What is the age and province distributions of Deactivated customers ?**

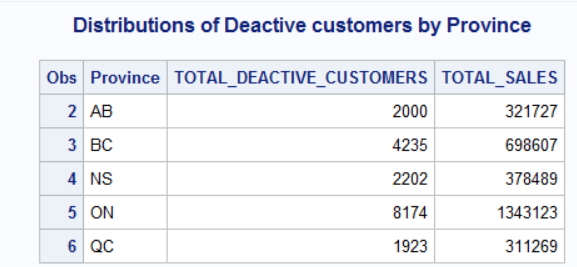
****

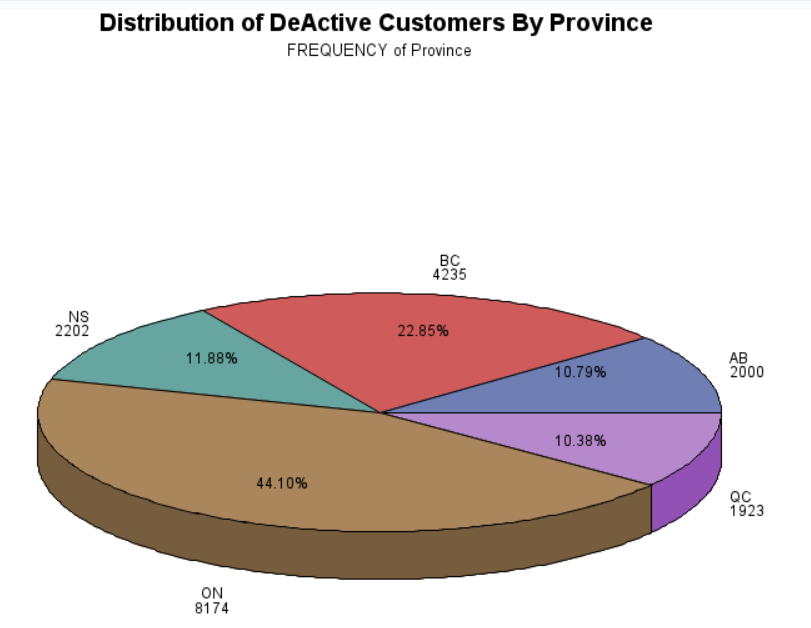
****

**A screenshot of a computer

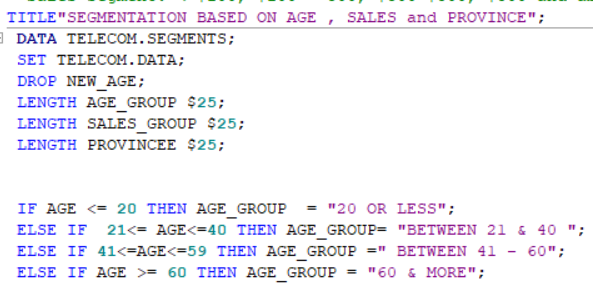
Description automatically generated**

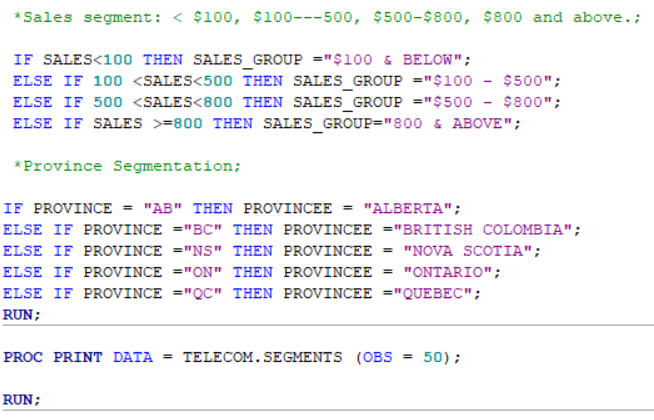
****

****

****

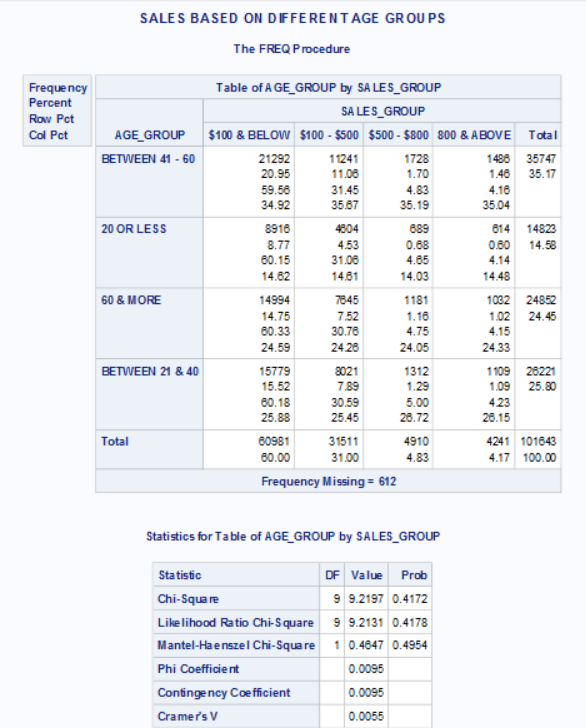
**1.3 Segment the customers based on age, province and sales amount: Sales segment: < $100, $100---500, $500-$800, $800 and above. Age segments: < 20, 21-40, 41-60, 60 and above. Do Analysis of sales based on segmentation**

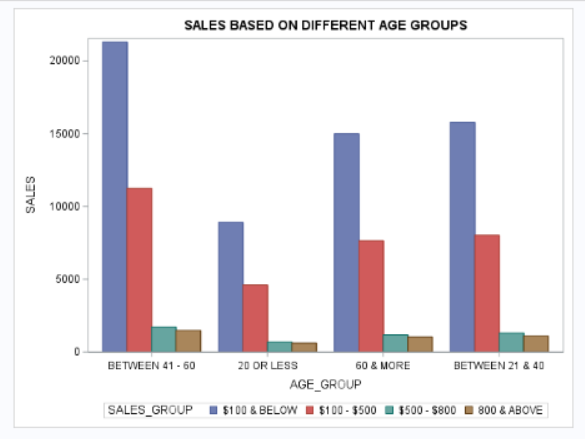
****

****

****

**SALES BASED ON AGE GROUP**

****

****

Irrespective of age the sales is highest for the category $100 and below

**SALES BASED ON PROVINCE**

**A screenshot of a computer screen

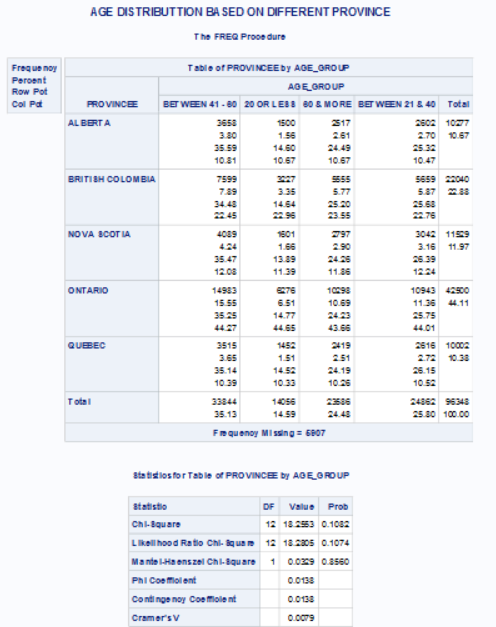
Description automatically generated**

**A graph of sales

Description automatically generated**

Irrespective of location/province the sales is highest for the category $100 and below

**AGE DISTRIBUTION IN PROVINCES**

****

**A graph of different colored bars

Description automatically generated**

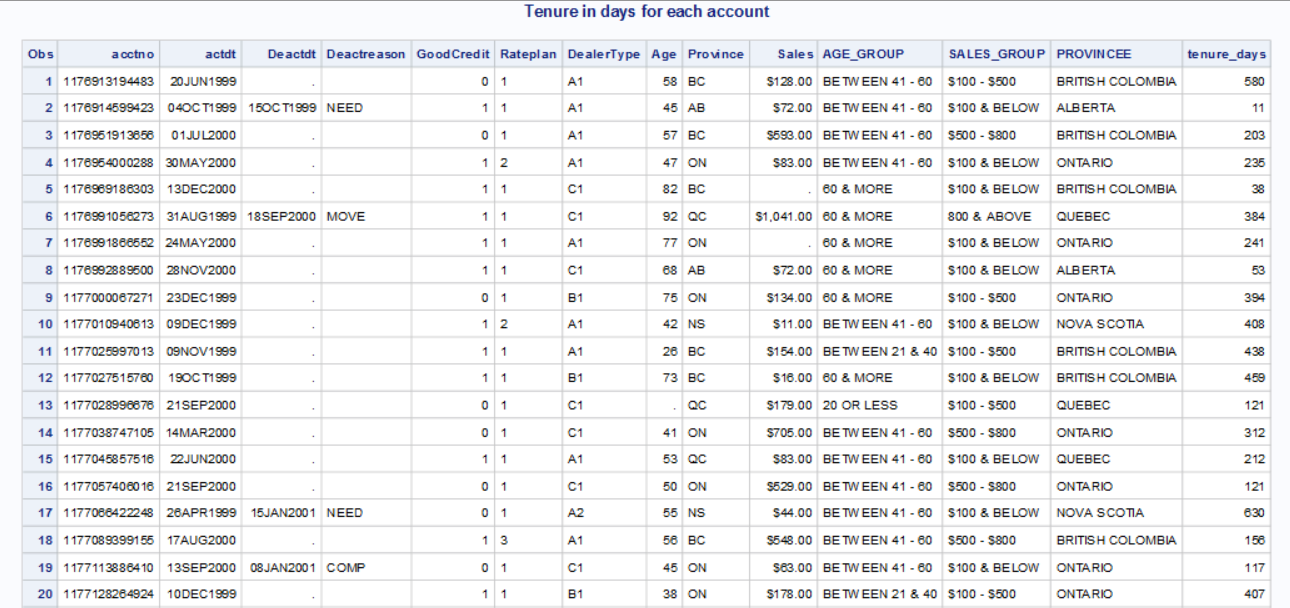
In each province maximum number of customers belong to the age group between 41 and 60

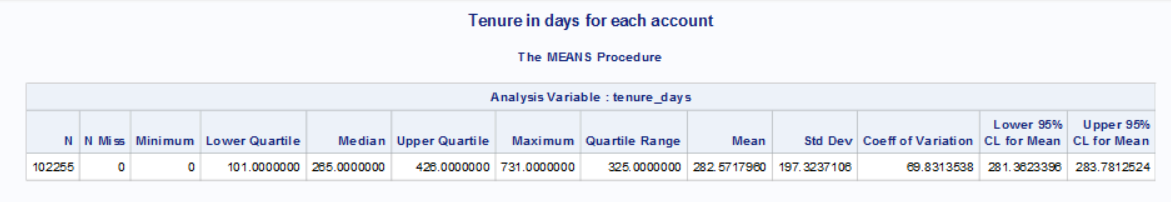
**1.4.Statistical Analysis:**

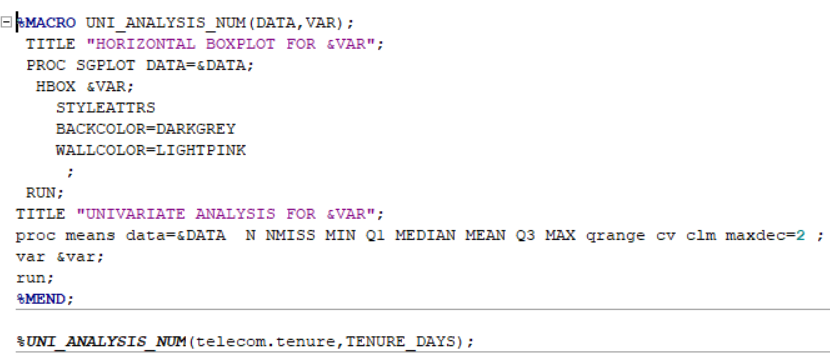
**1.4(a) Calculate the tenure in days for each account and give its simple statistics.**

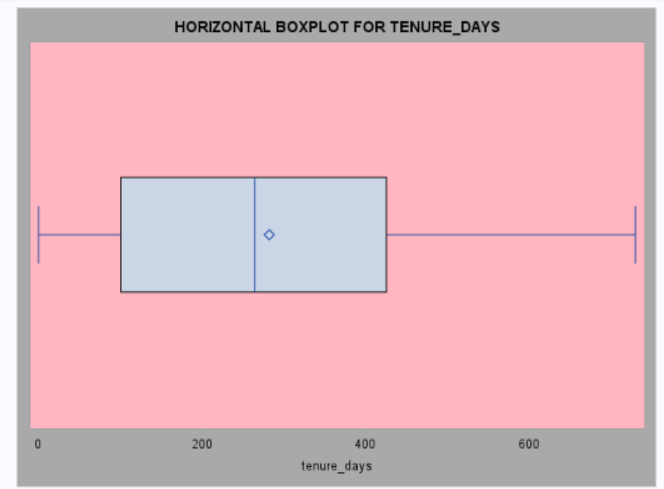
**A computer code with text

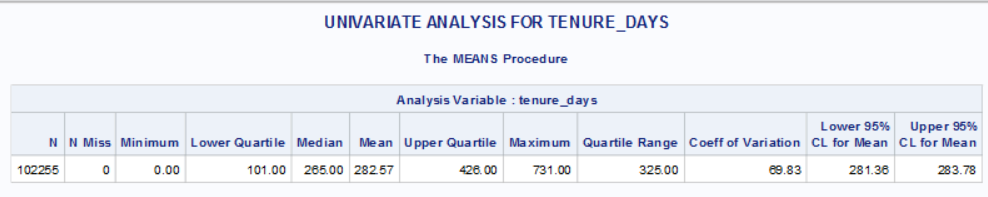
Description automatically generated**

****

****

****

****

****

**2) Calculate the number of accounts deactivated for each month?**

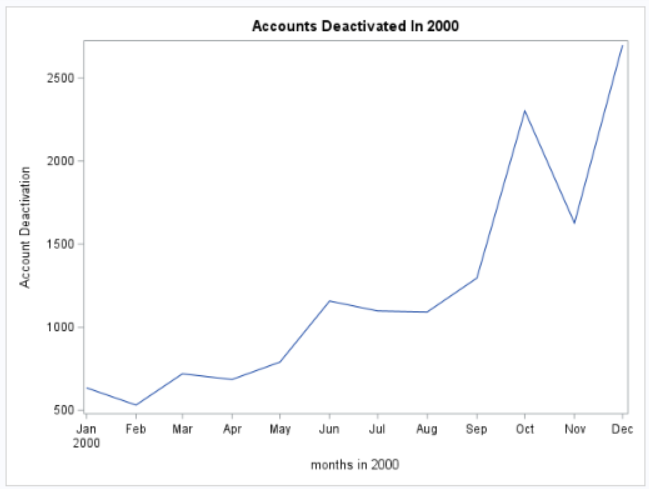
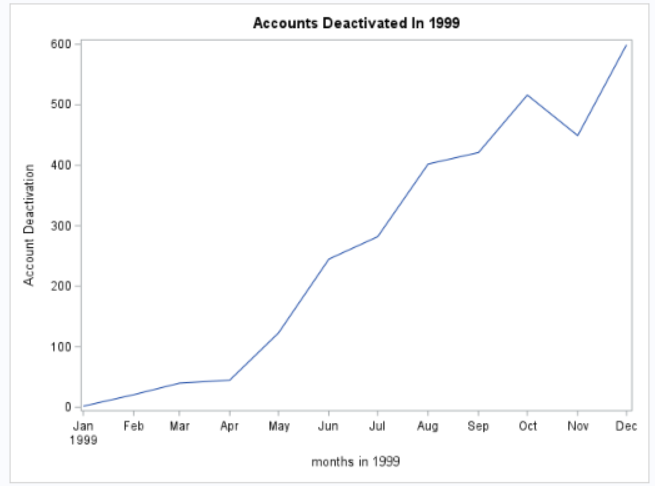
Total accounts deactivated each month from 1999 to 2001

**A screenshot of a computer screen

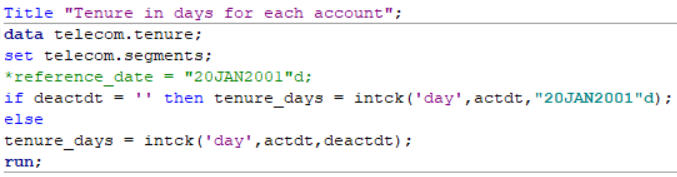
Description automatically generated**

Yearly Account deactivation based on months

****

****

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

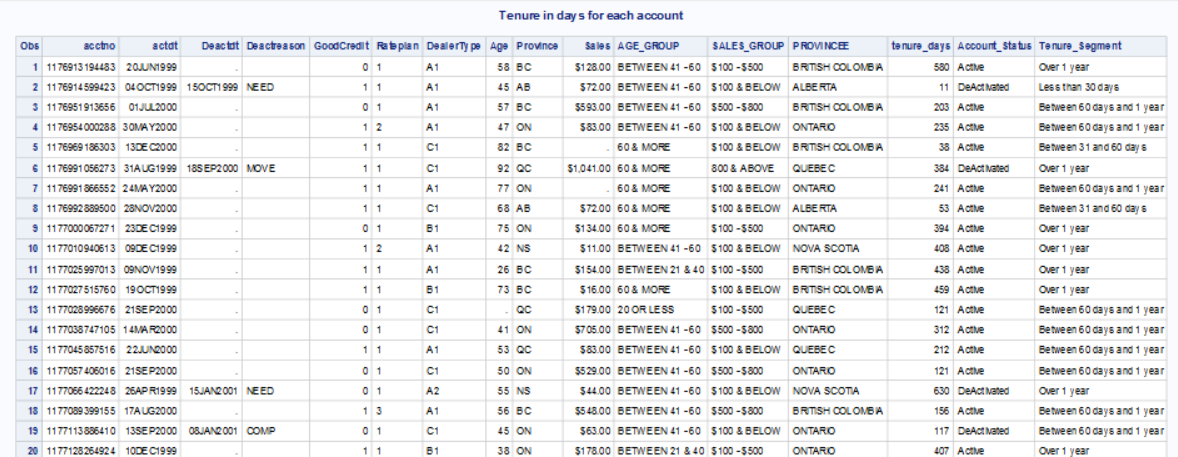
****

**A screenshot of a computer screen

Description automatically generated**

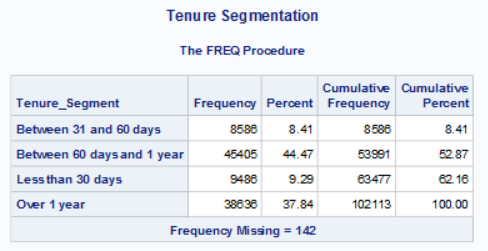
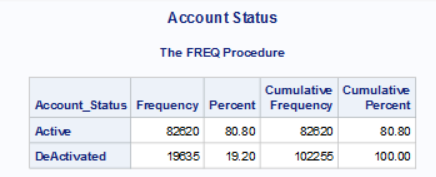
**A screen shot of a computer code

Description automatically generated**

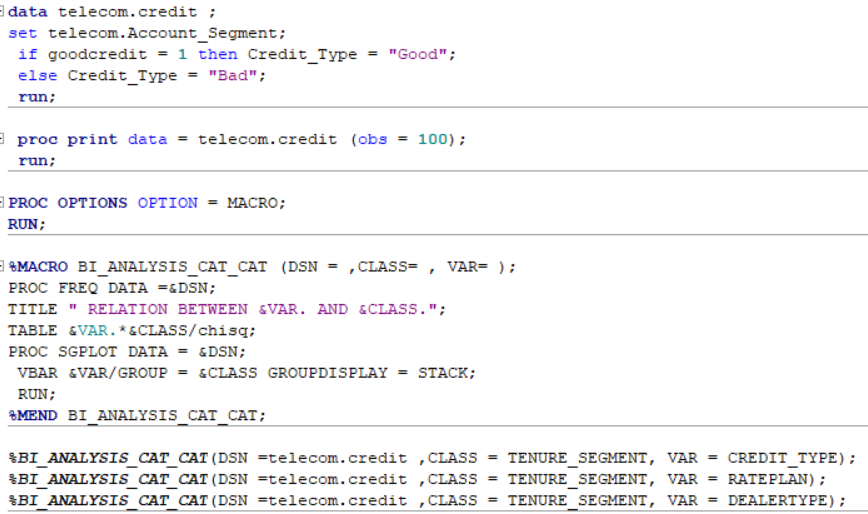
****

**A screenshot of a computer program

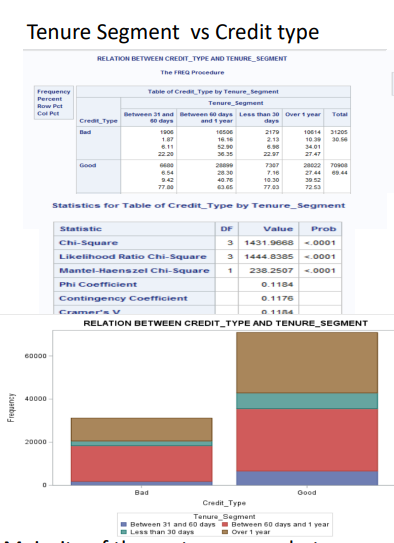
Description automatically generated**

****

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

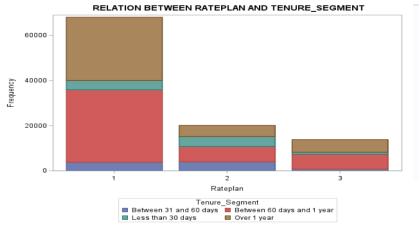
****

****

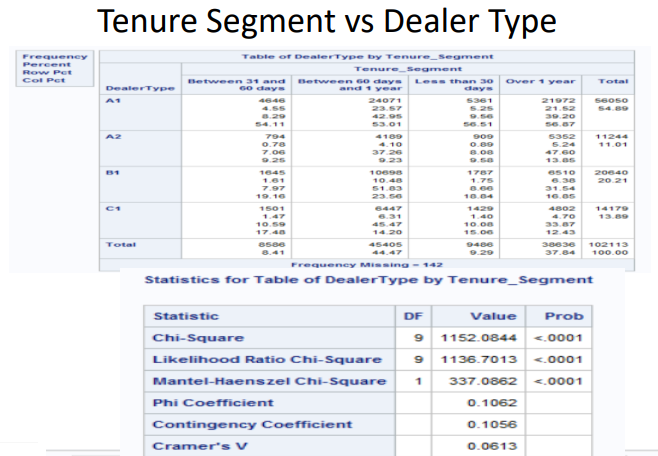
****

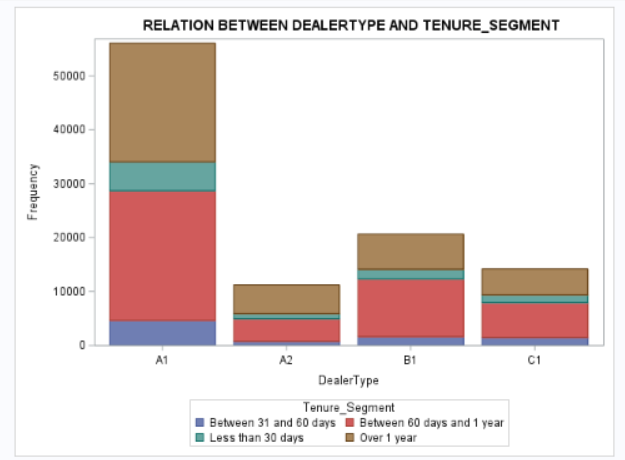
Majority of the customers are between the tenure of 60 days and 1 year, and over 1 year, and they all have good credit.

**A screenshot of a report

Description automatically generated**

Majority of the customers are between the tenure of 60 days and 1 year, and over 1 year, and they are all with rate plan 1.

****

****

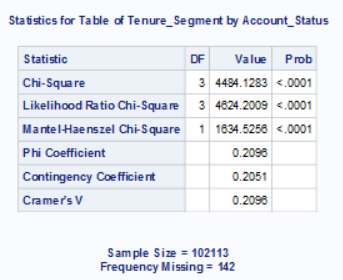
Majority of the customers are between the tenure of 60 days and 1 year, and over 1 year, and they are all with dealer A1.

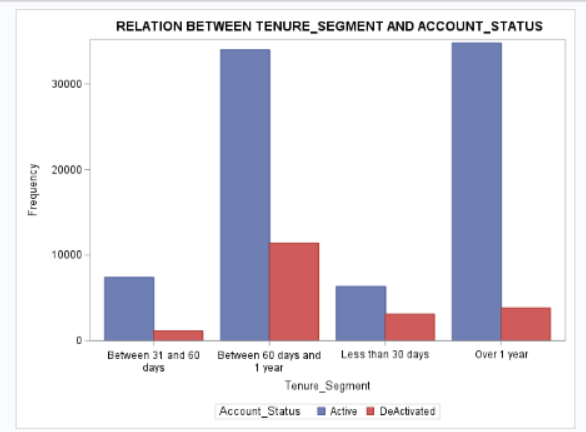
**5) Is there any association between the account status and the tenure segments?**

**A screenshot of a computer code

Description automatically generated**

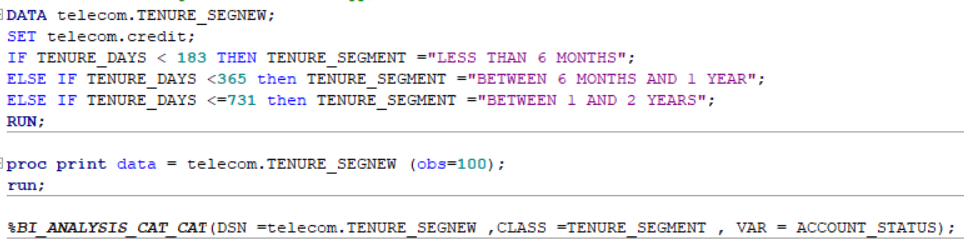
**A screenshot of a computer

Description automatically generated**

****

Number of customers de-activated is highest between tenure of 60 days and 1 year. Deactivated accounts are lower when the tenure is over 1 year.

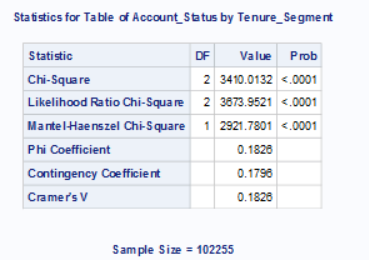
**Could you find out a better tenure segmentation strategy that is more associated with the account status?**

****

**A graph of different colored bars

Description automatically generated**

**A screenshot of a document

Description automatically generated**

Number of active customers is highest for tenure segment between 1 and 2 years. Where as, number of de-activated customers is lowest for the tenure between 1 and 2 years and highest for tenure segment less than 6 months.

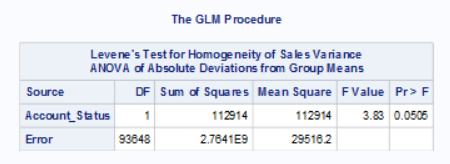
**6.Does Sales amount differ among different account status?**



**Test of Equality of variance**

Levene’s Test for Homogeneity of Variances is produced by PROC GLM to test if variances are considered equal across all groups .

From Levene’s test result we find that p value is >0.0505. As the P value is greater than 5%, we accept null hypothesis and conclude variances are equal.

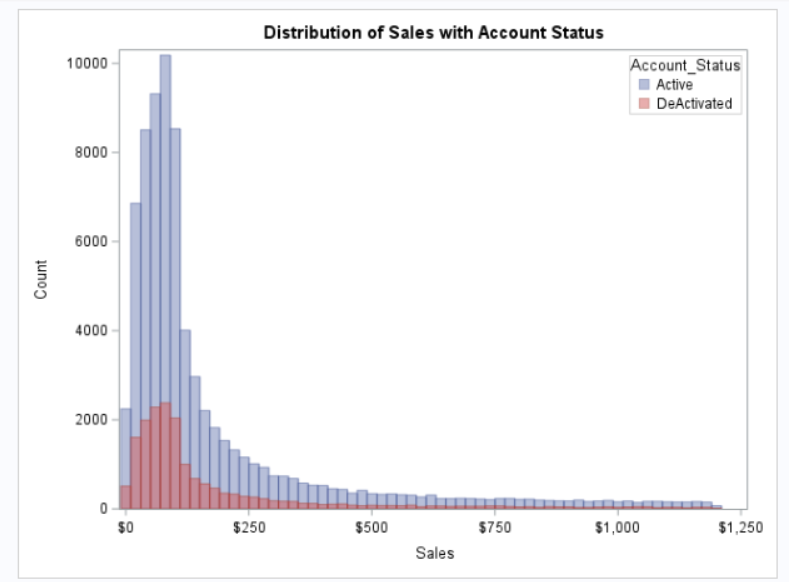


Since p value <0.05. As the p value is less than 5% we conclude that data is not normally distributed

**Test of Independency**

**A screenshot of a data analysis

Description automatically generated**

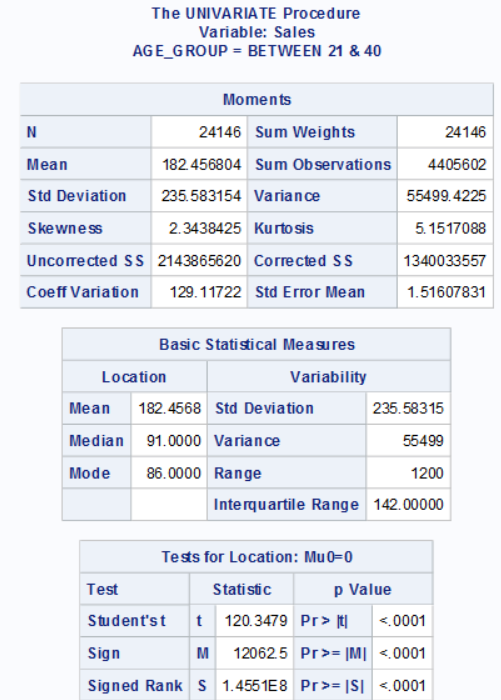
****

Visualization is done using grouped histogram. Both active and de-activated accounts follow the same pattern in terms of sales, where majority of the sales are happening under $250

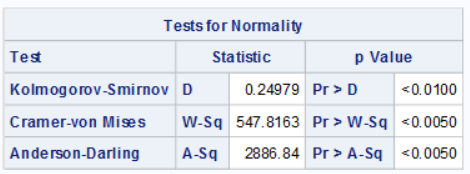
As our categorical variable Account status has 2 levels ,we perform T test for test of independency. Since the Levene’s test p-value > 0.05 and variance are found to be equal pooled T test is appropriate for testing independency

Based on the result from pooled T test , the p value >0.05, hence we fail to reject null hypothesis and find that there is no statistical relationship between sales and Account status.

**Does Sales amount differ among age group?**

****

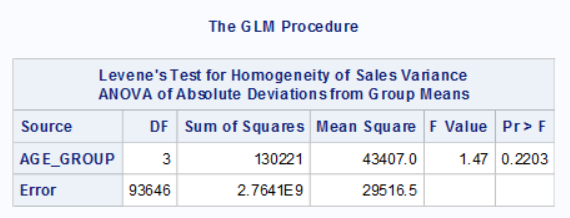
**Test of Normality**

****

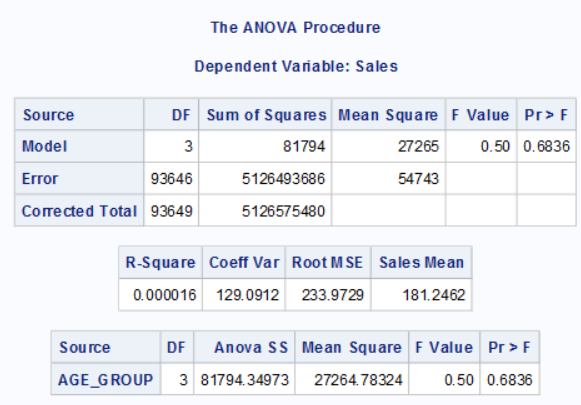
p value is found to be <0.05. As the p value is less than 5% we conclude that data is not normally distributed

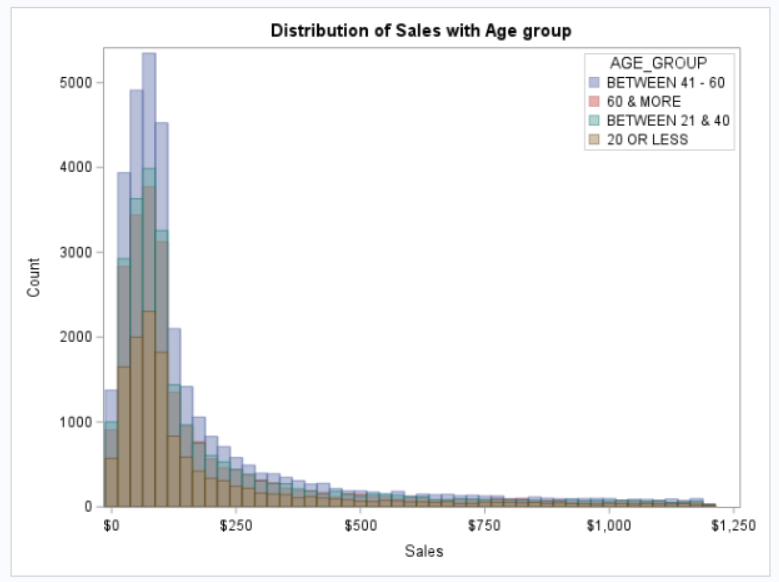
**Test of Equality of variance**

Levene’s Test for Homogeneity of Variances is produced by PROC GLM to test if variances are considered equal across all groups . From Levene’s test result we find that p value is <0.2203. As the P value is more than 5%, we fail to reject null hypothesis and conclude variances are equal.

****

**Test of Independency using One Way Anova**

****

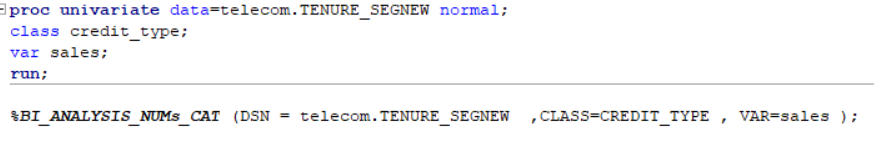
****

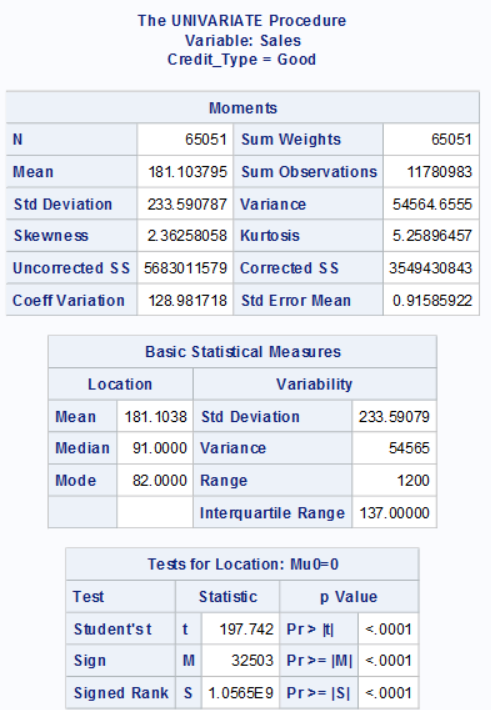
Similarly, all the age groups follow the same pattern in terms of sales, where majority of the sales are happening under $250

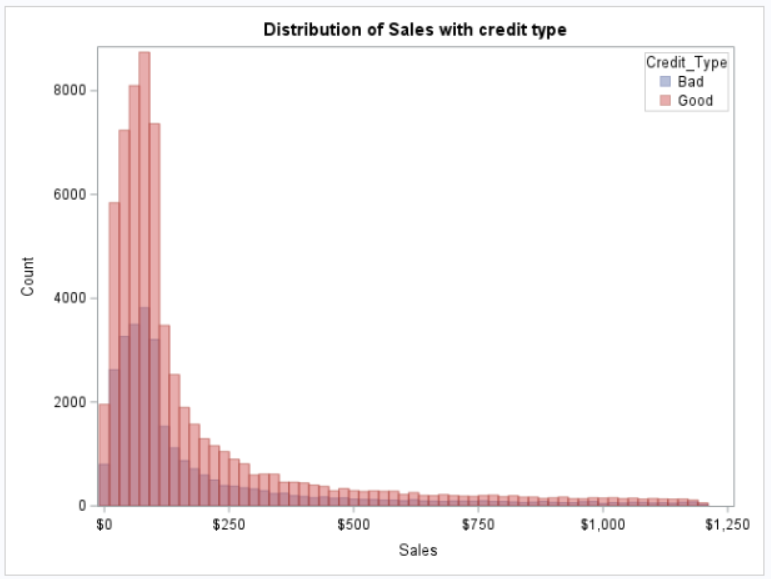
As groups have equal variance we consider p-value of the standard one-way ANOVA results for test of independency

Based on the result from Anova, the p value is 0.683 , as it is more than 5 % we fail to reject null hypothesis there is a no statistical relationship between sales and Age group.

**Does Sales amount differ among credit type**

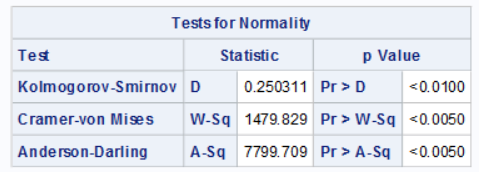
****

****

****

Similarly, both bad and good credit types follow the same pattern in terms of sales, where majority of the sales are happening under $250

**Test of Normality**

****

Since p value is less than 0.0001 which is <0.05. As the p value is less than 5% we conclude that data is not normally distributed

**Test of Equality of Variance**

Levene’s Test for Homogeneity of Variances is produced by PROC GLM to test if variances are considered equal across all groups . From Levene’s test result we find that p value is >0.6795. As the P value is greater than 5%, we accept null hypothesis and conclude variances are equal.

**A screenshot of a computer screen

Description automatically generated**

**Test of Independency – T test**

****

As our categorical variable credit type has 2 levels ,we perform T test for test of independency. Since the Levene’s test p-value > 0.05 and variance are found to be equal, pooled T test is appropriate for testing independency

Based on the result from pooled T test , the p value >0.05, hence we accept null hypothesis and that there is no statistical relationship between sales and Account status.

**CONCLUSIONS**

* Based on the data, it can be noticed that majority of the customers are between age group of 21- 40 and 41-60 years, and they are from Ontario and British Colombia. This is the same case for both active and de-activated customers.
* Irrespective of age and province, sales was highest for category $100 and below. In each province, maximum number of customers belong to the age group between 41 and 60.
* Number of de-activated accounts increased over the months and, year on year.
* Majority of the customers are between the tenure of 60 days and 1 year, and over 1 year, and they all have good credit. Also, majority of the customers had taken rate plan 1 and with dealer A1.
* Number of active customers is highest for tenure segment between 1 and 2 years. Where as, number of deactivated customers is lowest for the tenure between 1 and 2 years and highest for tenure segment less than 6 months.

**Recommendations**

* The middle-aged population makes up the majority of the population overall. Once the middle-aged group has stabilized, we can cultivate new users within the 20–40 year old demographic.
* 30.56% of accounts have bad credit, while 69.44% of accounts have good credit. Too many people have bad credit accounts. In the future, a thorough credit check should be performed on the customer when opening an account.
* RatePlan #1's percentage of the total data is twice the total of RatePlans #2 and #3. This indicates that RatePlan#1 is more well-liked, so we might think about releasing additional plans similar to it in the future.
* Number of deactivated accounts can be reduced by offering promotions and discounts for customers who are less than 6 months tenure.
* By providing continued excellent customer service, it may possible to retain the active customers.

**Appendix – SAS code**

\*Small Project for the Telecom company

Customer Distribution and Deactivation Analyses

Objective:Analyzing and performing EDA on a CRM data of a wireless company for 2 years.

investigating the customer distribution and business behaviors, and then gaining insightful understanding about the customers, and to forecast the deactivation

trends for the next 6 months.

Data:

Acctno: account number.

Actdt: account activation date

Deactdt: account deactivation date

DeactReason: reason for deactivation.

GoodCredit: customer’s credit is good or not.

RatePlan: rate plan for the customer.

DealerType: dealer type.

Age: customer age.

Province: province.

Sales: the amount of sales to a customer;

LIBNAME TELECOM "C:\DSA\Advanced SAS\Final Project";

\*==============================================================================

IMPORTING DATA SET IN TO SAS

==============================================================================;

Title " Importing Data Set in to SAS";

**data** telecom.Details;

infile "C:\DSA\Advanced SAS\Final Project\New\_Wireless\_Fixed.txt";

input acctno **1**-**14**

@**15** actdt mmddyy10.

@**26** Deactdt mmddyy10.

Deactreason $ **41**-**45**

GoodCredit **53**

Rateplan $**62**-**63**

DealerType $**65**-**66**

Age **74** -**75**

Province $ **80**-**81**

Sales dollar11.2

;

format acctno **14.0** actdt date9. Deactdt date9. ;

format Sales DOLLAR11.2;

**run**;

**proc** **print** data = telecom.Details(obs=**10**);

**run**;

title;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Analyzing the data set and Performing EDA

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*==================================================================================

Browsing Descriptive Portion

===================================================================================;

**proc** **contents** data = telecom.Details order = varnum;

**run**;

\*===================================================================================

Browsing Head of data set

====================================================================================;

Title "Browsing Head of Data Set";

**proc** **print** data = telecom.Details (obs = **10**);

**run**;

\*===================================================================================

Browsing tail of data set

====================================================================================;

Title "Browsing Tail of Data Set";

**proc** **print** data = telecom.Details (obs = **102255** firstobs =**102246**);

**run**;

\*====================================================================================

\*Finding number of unique distinct values

=====================================================================================;

Title "Number of unique distinct values in each variables";

**proc** **freq** data =telecom.Details nlevels;

ods exclude onewayfreqs;

**run**;

**proc** **freq** data = telecom.Details;

table deactreason Dealertype province goodcredit/ nopercent nocum;

**run**;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*EDA AND DATA CLEANING;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*Removing Duplicated Data if any;

**proc** **sort** data = telecom.Details out = telecom.Data nodupkey;

by Acctno;

**run**;

\*No duplicated account numbers were found so zero observations were deleted ;

\*=======================================================================================

Checking for missing values

========================================================================================;

Title "Number of missing values";

**proc** **means** data = telecom.Data nmiss;

**run**;

**proc** **sql**;

select nmiss(Province) as Province,nmiss (deactreason) as deactreason,

nmiss(rateplan) as rateplan,nmiss(dealertype) as dealertype

from telecom.Data;

**quit**;

\* Replacing missing values in age with mean age in new column as New\_Age ;

**PROC** **SQL**;

CREATE TABLE TELECOM.WIRELESS\_NOMISS AS

SELECT\*,

COALESCE(AGE,MEDIAN(AGE)) as New\_age

FROM TELECOM.Details

;

**QUIT**;

**PROC** **PRINT** DATA = TELECOM.WIRELESS\_NOMISS (OBS=**50**);

**RUN**;

\*========================================================================================

Descriptive Analysis of continous variables

=========================================================================================;

TITLE"DESCRIPTIVE ANALYSIS OF CONTINUOUS";

**PROC** **MEANS** DATA = Telecom.Data N NMISS MIN Q1 MEDIAN Q3 MAX qrange mean std cv clm;

**RUN**;

**PROC** **UNIVARIATE** DATA = Telecom.Data;**RUN**;

\*---------------------------------------------------------------------------------------------------

QUESTION 1

1.1 Explore and describe the dataset briefly. For example, is the acctno unique? What

is the number of accounts activated and deactivated? When is the earliest and

latest activation/deactivation dates available?

-------------------------------------------------------------------------------------------------;

\*1.1a Explore and describe the dataset briefly;

TITLE "DESCRIPTION OF DATASET";

**PROC** **CONTENTS** DATA = TELECOM.DATA;

**RUN**;

\*Checking if All Account Numbers are Unique ?;

\*Removing Duplicated Data if any;

**proc** **sort** data = telecom.Details out = telecom.Data nodupkey;

by Acctno;

**run**;

\*Number of unique accounts;

**PROC** **SQL** OUTOBS=**20**;

SELECT COUNT(\*) AS TOTAL\_COUNT,

COUNT(DISTINCT Acctno) AS UNIQUE\_ACCOUNTS

FROM TELECOM.DATA

;

**QUIT**;

\*No duplicated account numbers were found so zero observations were deleted ;

\*what is the number of accounts activated and deactivated ?;

Title "Number of Activated and Deactivated Accounts";

**PROC** **SQL**;

SELECT COUNT(Acctno) AS Total\_Accounts,

(COUNT(Actdt) - COUNT(Deactdt)) AS Activted\_Accounts,

COUNT(Deactdt) AS Deactvated\_Accounts

FROM telecom.Data;

**QUIT**;

\*When is the earliest and latest activation/deactivation dates available?;

\*Earliest and Latest Activation Date;

Title "Earliest and Latest Activation Date";

**proc** **sql**;

select min(actdt) as Earliest\_Activation\_Date format = date9.,

max(actdt) as Latest\_Activation\_Date format = date9.

from telecom.Data;

**quit**;

\*Earliest and Latest DeActivation Date;

Title "Earliest and Latest DeActivation Date";

**proc** **sql**;

select min(deactdt) as Earliest\_DeActivation\_Date format = date9.,

max(deactdt) as Latest\_DeActivation\_Date format = date9.

from telecom.Data;

**quit**;

title;

\*QUESTION 2;

\*1.2 What is the age and province distributions of active customers?;

**data** telecom.Account\_segment ;

set telecom.Data;

length Age\_Group $**25**;

IF AGE <= **20** THEN AGE\_GROUP = "LESS THAN 20";

ELSE IF **21**<= AGE<=**40** THEN AGE\_GROUP = "BETWEEN 21 AND 40 ";

ELSE IF **41**<=AGE <=**59** THEN AGE\_GROUP ="BETWEEN 41 AND 60";

ELSE IF AGE >= **60** THEN AGE\_GROUP = "60 AND ABOVE";

**RUN**;

**proc** **print** data = telecom.Account\_segment (obs=**20**);

**run**;

Title"Age distributions of active customers";

**PROC** **SQL** ;

CREATE TABLE AGEDIST AS

SELECT AGE\_GROUP,

(COUNT(Actdt) - COUNT(Deactdt))AS TOTAL\_ACTIVE\_CUSTOMERS,

SUM(SALES) AS TOTAL\_SALES

FROM TELECOM.ACCOUNT\_SEGMENT

WHERE DEACTDT IS NULL

GROUP BY AGE\_GROUP

ORDER BY AGE\_GROUP;

**QUIT**;

**PROC** **PRINT** DATA = AGEDIST;

**RUN**;

**PROC** **GCHART** DATA = TELECOM.ACCOUNT\_SEGMENT;

PIE3D AGE\_GROUP/ PERCENT =INSIDE;

WHERE DEACTDT IS NULL;

**RUN**;

\*=====================================================================================;

**PROC** **SQL** ;

CREATE TABLE PROVINDIST AS

SELECT PROVINCE,

(COUNT(Actdt) - COUNT(Deactdt))AS TOTAL\_ACTIVE\_CUSTOMERS,

SUM(SALES) AS TOTAL\_SALES

FROM TELECOM.ACCOUNT\_SEGMENT

WHERE DEACTDT IS NULL ,PROVINCE IS NOT NULL

GROUP BY PROVINCE

ORDER BY PROVINCE;

**QUIT**;

\*PROC PRINT DATA = PROVINDIST;

\*WHERE PROVINCE IS NOT NULL;

\*RUN;

Title"Distribution of Active Customers By Province ";

**PROC** **GCHART** DATA = TELECOM.ACCOUNT\_SEGMENT;

PIE3D PROVINCE/ PERCENT =INSIDE;

WHERE DEACTDT IS NULL;

**RUN**;

\*=======================================================================================;

**PROC** **SQL** ;

CREATE TABLE AGEPROVINDIST AS

SELECT PROVINCE,AGE\_GROUP,

(COUNT(Actdt) - COUNT(Deactdt))AS TOTAL\_ACTIVE\_CUSTOMERS,

SUM(SALES) AS TOTAL\_SALES

FROM TELECOM.ACCOUNT\_SEGMENT

WHERE DEACTDT IS NULL

GROUP BY PROVINCE,AGE\_GROUP

ORDER BY PROVINCE,AGE\_GROUP;

**QUIT**;

Title"Distribution of Active Customers By Age and Province ";

**PROC** **PRINT** DATA = AGEPROVINDIST;

WHERE PROVINCE IS NOT NULL;

**RUN**;

\*What is the age and province distributions of deactivated customers?;

Title"Distribution of DeActive Customers By Age and Province ";

**PROC** **SQL** ;

CREATE TABLE AGEPROVINDIST1 AS

SELECT PROVINCE,AGE\_GROUP,

COUNT(Deactdt)AS TOTAL\_DEACTIVATED\_CUSTOMERS,

SUM(SALES) AS TOTAL\_SALES

FROM TELECOM.ACCOUNT\_SEGMENT

WHERE DEACTDT IS NOT NULL

GROUP BY PROVINCE,AGE\_GROUP

ORDER BY PROVINCE,AGE\_GROUP;

**QUIT**;

**PROC** **PRINT** DATA = AGEPROVINDIST1;

WHERE PROVINCE IS NOT NULL;

**RUN**;

\*VISUALIZATION;

Title"Age distributions of Deactive customers";

**PROC** **SQL** ;

CREATE TABLE AGEDIST1 AS

SELECT AGE\_GROUP,

COUNT(DEACTDT) AS TOTAL\_DEACTIVE\_CUSTOMERS,

SUM(SALES) AS TOTAL\_SALES

FROM TELECOM.ACCOUNT\_SEGMENT

WHERE DEACTDT IS NOT NULL

GROUP BY AGE\_GROUP

ORDER BY AGE\_GROUP;

**QUIT**;

**PROC** **PRINT** DATA = AGEDIST1;

**RUN**;

**PROC** **GCHART** DATA = TELECOM.ACCOUNT\_SEGMENT;

PIE3D AGE\_GROUP/ PERCENT =INSIDE;

WHERE DEACTDT IS NOT NULL;

**RUN**;

\*=====================================================================================;

Title" Distributions of Deactive customers by Province";

**PROC** **SQL** ;

CREATE TABLE PROVINDIST AS

SELECT PROVINCE,

COUNT(DEACTDT) AS TOTAL\_DEACTIVE\_CUSTOMERS,

SUM(SALES) AS TOTAL\_SALES

FROM TELECOM.ACCOUNT\_SEGMENT

WHERE DEACTDT IS NOT NULL

GROUP BY PROVINCE

ORDER BY PROVINCE;

**QUIT**;

**PROC** **PRINT** DATA = PROVINDIST;

WHERE PROVINCE IS NOT NULL;

**RUN**;

Title"Distribution of DeActive Customers By Province ";

**PROC** **GCHART** DATA = TELECOM.ACCOUNT\_SEGMENT;

PIE3D PROVINCE/ PERCENT =INSIDE;

WHERE DEACTDT IS NOT NULL;

**RUN**;

\*QUESTION- 3;

\*Segment the customers based on age, province and sales amount:;

\*Sales segment: < $100, $100---500, $500-$800, $800 and above.;

TITLE"SEGMENTATION BASED ON AGE , SALES and PROVINCE";

**DATA** TELECOM.SEGMENTS;

SET TELECOM.DATA;

DROP NEW\_AGE;

LENGTH AGE\_GROUP $**25**;

LENGTH SALES\_GROUP $**25**;

LENGTH PROVINCEE $**25**;

IF AGE <= **20** THEN AGE\_GROUP = "20 OR LESS";

ELSE IF **21**<= AGE<=**40** THEN AGE\_GROUP= "BETWEEN 21 & 40 ";

ELSE IF **41**<=AGE<=**59** THEN AGE\_GROUP =" BETWEEN 41 - 60";

ELSE IF AGE >= **60** THEN AGE\_GROUP = "60 & MORE";

\*Sales segment: < $100, $100---500, $500-$800, $800 and above.;

IF SALES<**100** THEN SALES\_GROUP ="$100 & BELOW";

ELSE IF **100** <SALES<**500** THEN SALES\_GROUP ="$100 - $500";

ELSE IF **500** <SALES<**800** THEN SALES\_GROUP ="$500 - $800";

ELSE IF SALES >=**800** THEN SALES\_GROUP="800 & ABOVE";

\*Province Segmentation;

IF PROVINCE = "AB" THEN PROVINCEE = "ALBERTA";

ELSE IF PROVINCE ="BC" THEN PROVINCEE ="BRITISH COLOMBIA";

ELSE IF PROVINCE ="NS" THEN PROVINCEE = "NOVA SCOTIA";

ELSE IF PROVINCE ="ON" THEN PROVINCEE = "ONTARIO";

ELSE IF PROVINCE ="QC" THEN PROVINCEE ="QUEBEC";

**RUN**;

**PROC** **PRINT** DATA = TELECOM.SEGMENTS (OBS = **50**);

**RUN**;

\*Create Analysis Report Based on Segmentation;

TITLE "SALES BASED ON DIFFERENT AGE GROUPS";

**PROC** **FREQ** DATA = TELECOM.SEGMENTS;

TABLE AGE\_GROUP\*SALES\_GROUP/CHISQ;

**run**;

**PROC** **SGPLOT** DATA =TELECOM.SEGMENTS;

VBAR AGE\_GROUP/ group = SALES\_GROUP groupdisplay = CLUSTER;

YAXIS LABEL = "SALES";

**RUN**;

TITLE "SALES BASED ON DIFFERENT PROVINCE";

**PROC** **FREQ** DATA = TELECOM.SEGMENTS;

TABLE PROVINCEE\*SALES\_GROUP/ chisq;

**run**;

**PROC** **SGPLOT** DATA =TELECOM.SEGMENTS;

VBAR PROVINCEE/ group =SALES\_GROUP groupdisplay = CLUSTER;

YAXIS LABEL = "SALES";

**RUN**;

TITLE "AGE DISTRIBUTTION BASED ON DIFFERENT PROVINCE";

**PROC** **FREQ** DATA = TELECOM.SEGMENTS;

TABLE PROVINCEE\*AGE\_GROUP/ chisq;

**run**;

**PROC** **SGPLOT** DATA =TELECOM.SEGMENTS;

VBAR PROVINCEE/ group =AGE\_GROUP groupdisplay = CLUSTER;

YAXIS LABEL = "SALES";

**RUN**;

**PROC** **OPTIONS** OPTION = MACRO;

**RUN**;

\*METHOD USING MACRO;

**%MACRO** BI\_ANALYSIS\_CAT\_CAT (DSN = ,VAR1= , VAR2= );

PROC FREQ DATA =&DSN;

TITLE " RELATION BETWEEN &VAR1. AND &VAR2.";

TABLE &VAR1.\*&VAR2/chisq;

PROC SGPLOT DATA = &DSN;

VBAR &VAR1/GROUP = &VAR2 GROUPDISPLAY = CLUSTER;

RUN;

**%MEND** BI\_ANALYSIS\_CAT\_CAT;

%***BI\_ANALYSIS\_CAT\_CAT***(DSN =TELECOM.SEGMENTS,VAR1= AGE\_GROUP, VAR2 = SALES\_GROUP);

%***BI\_ANALYSIS\_CAT\_CAT***(DSN =TELECOM.SEGMENTS ,VAR1 =SALES\_GROUP , VAR2 =PROVINCE );

%***BI\_ANALYSIS\_CAT\_CAT***(DSN =TELECOM.SEGMENTS ,VAR1 AGE\_GROUP , VAR2 =PROVINCE );

\*1.4.Statistical Analysis:;

\*1) Calculate the tenure in days for each account and give its simple statistics;

**proc** **sort** data = telecom.segments out=telecom.sort ;

by descending actdt ;

**run**;

**proc** **print** data = telecom.sort (obs=**50**);

**run**;

Title "Tenure in days for each account";

**data** telecom.tenure;

set telecom.segments;

\*reference\_date = "20JAN2001"d;

if deactdt = '' then tenure\_days = intck('day',actdt,**"20JAN2001"d**);

else

tenure\_days = intck('day',actdt,deactdt);

**run**;

**proc** **print** data = telecom.tenure(obs = **50**);

**run**;

**proc** **means** data = telecom.tenure N NMISS MIN Q1 MEDIAN Q3 MAX qrange mean std cv clm;

var tenure\_days;

**run**;

**%MACRO** UNI\_ANALYSIS\_NUM(DATA,VAR);

TITLE "HORIZONTAL BOXPLOT FOR &VAR";

PROC SGPLOT DATA=&DATA;

HBOX &VAR;

STYLEATTRS

BACKCOLOR=DARKGREY

WALLCOLOR=LIGHTPINK

;

RUN;

TITLE "UNIVARIATE ANALYSIS FOR &VAR";

proc means data=&DATA N NMISS MIN Q1 MEDIAN MEAN Q3 MAX qrange cv clm maxdec=**2** ;

var &var;

run;

**%MEND**;

%***UNI\_ANALYSIS\_NUM***(telecom.tenure,TENURE\_DAYS);

\*2) Calculate the number of accounts deactivated for each month;

**data** telecom.deact\_yrmonth;

set telecom.segments;

deact\_year = year(deactdt);

deact\_month = intnx('month',deactdt,**0**,'b');

**run**;

**proc** **print** data = telecom.deact\_yrmonth (obs=**50**);

format deact\_month monname3.;

**run**;

\*To find monthly deactivation in all the years from 1999 - 2001;

Title "Number of monthly deactivation from 1999 - 2001";

**proc** **freq** data = telecom.deact\_yrmonth;

table deact\_month/NOCUM;

format deact\_month monname3.;

**run**;

\*for year 1999;

**DATA** telecom.deact\_1999;

SET telecom.deact\_yrmonth;

KEEP acctno deactdt deact\_year deact\_month;

where deact\_year = **1999**;

**RUN**;

**proc** **print** data = telecom.deact\_1999 (obs=**100**);

format deact\_month monname3.;

**run**;

**proc** **sort** data = telecom.deact\_1999 out = telecom.deact1\_1999;

by deact\_month;

**run**;

**proc** **print** data = telecom.deact1\_1999 (obs=**100**);

format deact\_month monname3.;

**run**;

Title"Number Of Accounts Deactivated For Each Month In 1999";

**proc** **sql**;

create table telecom.deactNo\_1999 as

select deact\_month,

count(\*) as Deact\_Number

from telecom.deact1\_1999

group by deact\_month;

**quit**;

**proc** **print** data = telecom.deactNo\_1999;

format deact\_month monname3.;

**run**;

\*for year 2000;

**DATA** telecom.deact\_2000;

SET telecom.deact\_yrmonth;

KEEP acctno deactdt deact\_year deact\_month;

where deact\_year = **2000**;

**RUN**;

**proc** **print** data = telecom.deact\_2000 (obs=**100**);

format deact\_month monname3.;

**run**;

**proc** **sort** data = telecom.deact\_2000 out = telecom.deact1\_2000;

by deact\_month;

**run**;

**proc** **print** data = telecom.deact1\_2000 (obs=**100**);

format deact\_month monname3.;

**run**;

Title"Number Of Accounts Deactivated For Each Month In 2000";

**proc** **sql**;

create table telecom.deactNo\_2000 as

select deact\_month,

count(\*) as Deact\_Number

from telecom.deact1\_2000

group by deact\_month;

**quit**;

**proc** **print** data = telecom.deactNo\_2000;

format deact\_month monname3.;

**run**;

\*for year 2001;

**DATA** telecom.deact\_2001;

SET telecom.deact\_yrmonth;

KEEP acctno deactdt deact\_year deact\_month;

where deact\_year = **2001**;

**RUN**;

**proc** **print** data = telecom.deact\_2001 (obs=**100**);

format deact\_month monname3.;

**run**;

**proc** **sort** data = telecom.deact\_2001 out = telecom.deact1\_2001;

by deact\_month;

**run**;

**proc** **print** data = telecom.deact1\_2001 (obs=**100**);

format deact\_month monname3.;

**run**;

Title"Number Of Accounts Deactivated For Each Month In 2001";

**proc** **sql**;

create table telecom.deactNo\_2001 as

select deact\_month,

count(\*) as Deact\_Number

from telecom.deact1\_2001

group by deact\_month;

**quit**;

**proc** **print** data = telecom.deactNo\_2001;

format deact\_month monname3.;

**run**;

\*visualization;

Title"Accounts Deactivated In 1999";

**proc** **sgplot** data=telecom.deactno\_1999;

series x=deact\_month y=deact\_number;

xaxis label = "months in 1999";

yaxis label ="Account Deactivation";

format deact\_month monname3.;

**run**;

Title"Accounts Deactivated In 2000";

**proc** **sgplot** data=telecom.deactno\_2000;

series x=deact\_month y=deact\_number;

xaxis label = "months in 2000";

yaxis label ="Account Deactivation";

format deact\_month monname3.;

**run**;

\*3) Segment the account, first by account status “Active” and “Deactivated”, then by

Tenure: < 30 days, 31---60 days, 61 days--- one year, over one year. Report the

number of accounts of percent of all for each segment.;

Title "Tenure in days for each account";

**data** telecom.tenure;

set telecom.segments;

\*reference\_date = "20JAN2001"d;

if deactdt = '' then tenure\_days = intck('day',actdt,**"20JAN2001"d**);

else

tenure\_days = intck('day',actdt,deactdt);

**run**;

**proc** **print** data = telecom.tenure(obs = **50**);

**run**;

**data** telecom.Account\_segment ;

set telecom.tenure;

length Account\_Status $**25**;

length Tenure\_Segment $**30**;

if deactdt = '' then Account\_Status = "Active";

else Account\_Status = "DeActivated";

if tenure\_days <**30** then Tenure\_Segment ="Less than 30 days";

else if tenure\_days <**60** then Tenure\_Segment = "Between 31 and 60 days";

else if tenure\_days <**365** then Tenure\_Segment = "Between 60 days and 1 year";

else if tenure\_days > **365** then Tenure\_Segment = "Over 1 year";

**run**;

**proc** **print** data = telecom.Account\_Segment(obs = **100**);

**run**;

title"Account Status";

**proc** **freq** data = telecom.Account\_Segment;

table Account\_Status;

**run**;

title"Tenure Segmentation";

**proc** **freq** data = telecom.Account\_Segment;

table Tenure\_Segment;

**run**;

\*4) Test the general association between the tenure segments and “Good Credit”

“RatePlan ” and “DealerType.”;

**data** telecom.credit ;

set telecom.Account\_Segment;

if goodcredit = **1** then Credit\_Type = "Good";

else Credit\_Type = "Bad";

**run**;

**proc** **print** data = telecom.credit (obs = **100**);

**run**;

**PROC** **OPTIONS** OPTION = MACRO;

**RUN**;

**%MACRO** BI\_ANALYSIS\_CAT\_CAT (DSN = ,CLASS= , VAR= );

PROC FREQ DATA =&DSN;

TITLE " RELATION BETWEEN &VAR. AND &CLASS.";

TABLE &VAR.\*&CLASS/chisq;

PROC SGPLOT DATA = &DSN;

VBAR &VAR/GROUP = &CLASS GROUPDISPLAY = STACK;

RUN;

**%MEND** BI\_ANALYSIS\_CAT\_CAT;

%***BI\_ANALYSIS\_CAT\_CAT***(DSN =telecom.credit ,CLASS = TENURE\_SEGMENT, VAR = CREDIT\_TYPE);

%***BI\_ANALYSIS\_CAT\_CAT***(DSN =telecom.credit ,CLASS = TENURE\_SEGMENT, VAR = RATEPLAN);

%***BI\_ANALYSIS\_CAT\_CAT***(DSN =telecom.credit ,CLASS = TENURE\_SEGMENT, VAR = DEALERTYPE);

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

Could you find out a better tenure segmentation strategy that is more associated

with the account status?;

**%MACRO** BI\_ANALYSIS\_CAT\_CAT (DSN = ,CLASS= , VAR= );

PROC FREQ DATA =&DSN;

TITLE " RELATION BETWEEN &VAR. AND &CLASS.";

TABLE &VAR.\*&CLASS/chisq;

PROC SGPLOT DATA = &DSN;

VBAR &VAR/GROUP = &CLASS GROUPDISPLAY = CLUSTER;

RUN;

**%MEND** BI\_ANALYSIS\_CAT\_CAT;

%***BI\_ANALYSIS\_CAT\_CAT***(DSN =telecom.credit ,CLASS =ACCOUNT\_STATUS , VAR = TENURE\_SEGMENT);

\*better tenure segmentation strategy;

**DATA** telecom.TENURE\_SEGNEW;

SET telecom.credit;

IF TENURE\_DAYS < **183** THEN TENURE\_SEGMENT ="LESS THAN 6 MONTHS";

ELSE IF TENURE\_DAYS <**365** then TENURE\_SEGMENT ="BETWEEN 6 MONTHS AND 1 YEAR";

ELSE IF TENURE\_DAYS <=**731** then TENURE\_SEGMENT ="BETWEEN 1 AND 2 YEARS";

**RUN**;

**proc** **print** data = telecom.TENURE\_SEGNEW (obs=**100**);

**run**;

%***BI\_ANALYSIS\_CAT\_CAT***(DSN =telecom.TENURE\_SEGNEW ,CLASS =TENURE\_SEGMENT , VAR = ACCOUNT\_STATUS);

\*6 Does Sales amount differ among different account status, GoodCredit, and

customer age segments?;

\*SALES VS ACCOUNT STATUS

\*=======================

\*summarization using proc univariate and test of normality;

**proc** **univariate** data=telecom.TENURE\_SEGNEW normal;

class account\_status;

var sales;

**run**;

**%MACRO** BI\_ANALYSIS\_NUMs\_CAT (DSN = ,CLASS= , VAR=,VAR1= );

%LET N = %SYSFUNC(COUNTW(&VAR));

%DO I = **1** %TO &N;

%LET X = %SCAN(&VAR,&I);

PROC MEANS DATA = &DSN. N NMISS MIN Q1 MEDIAN MEAN Q3 MAX qrange cv clm maxdec=**2** ;

TITLE " RELATION BETWEEN &X. AND &CLASS.";

CLASS &CLASS. ;

VAR &X.;

OUTPUT OUT= OUT\_&CLASS.\_&X. MIN = MEAN= STD = MAX = /AUTONAME ;

RUN;

%END;

**%MEND** BI\_ANALYSIS\_NUMs\_CAT;

%***BI\_ANALYSIS\_NUMs\_CAT*** (DSN = telecom.TENURE\_SEGNEW ,CLASS=ACCOUNT\_STATUS , VAR=sales,VAR1=PROVINCEE );

%***BI\_ANALYSIS\_NUMs\_CAT*** (DSN = telecom.TENURE\_SEGNEW ,CLASS=CREDIT\_TYPE , VAR=sales );

%***BI\_ANALYSIS\_NUMs\_CAT*** (DSN = telecom.TENURE\_SEGNEW ,CLASS=AGE\_GROUP, VAR=sales );

\*Visualization using histogram;

Title "Distribution of Sales with Account Status ";

**proc** **sgplot** data=telecom.TENURE\_SEGNEW;

histogram sales / group=account\_status transparency=**0.5** scale=count;

keylegend / location=inside position=topright across=**1**;

**run**;

title;

\*for test of equality of variance;

**proc** **glm** data=telecom.TENURE\_SEGNEW;

class account\_status;

model sales = account\_status;

means account\_status / hovtest=levene(type=abs);

**run**;

\*T test for test of independancy;

**%MACRO** BI\_ANALYSIS\_NUMs\_CAT\_TTEST (DSN = ,CLASS= , VAR= );

TITLE "TTEST for &VAR. grouped by &CLASS. in &DSN.";

proc ttest data=&DSN.;

var &VAR. ;

class &CLASS.;

run;

QUIT;

**%MEND** BI\_ANALYSIS\_NUMs\_CAT\_TTEST;

%***BI\_ANALYSIS\_NUMs\_CAT\_TTEST*** (DSN =telecom.TENURE\_SEGNEW ,CLASS=account\_status , VAR=sales);

\*SALES VS GOOD CREDIT

\*summarization using proc univariate and test of normality;

**proc** **univariate** data=telecom.TENURE\_SEGNEW normal;

class credit\_type;

var sales;

**run**;

%***BI\_ANALYSIS\_NUMs\_CAT*** (DSN = telecom.TENURE\_SEGNEW ,CLASS=CREDIT\_TYPE , VAR=sales );

\*Visualization using histogram;

Title "Distribution of Sales with credit type ";

**proc** **sgplot** data=telecom.TENURE\_SEGNEW;

histogram sales / group=credit\_type transparency=**0.5** scale=count;

keylegend / location=inside position=topright across=**1**;

**run**;

title;

\*for test of equality of variance;

**proc** **glm** data=telecom.TENURE\_SEGNEW;

class credit\_type;

model sales = credit\_type;

means credit\_type / hovtest=levene(type=abs);

**run**;

%***BI\_ANALYSIS\_NUMs\_CAT\_TTEST*** (DSN =telecom.TENURE\_SEGNEW ,CLASS=credit\_type , VAR=sales);

\*Does Sales amount differ among age group?;

**proc** **univariate** data=telecom.TENURE\_SEGNEW normal;

class Age\_group;

var sales;

**run**;

%***BI\_ANALYSIS\_NUMs\_CAT*** (DSN = telecom.TENURE\_SEGNEW ,CLASS=CREDIT\_TYPE , VAR=sales );

\*Visualization using histogram;

Title "Distribution of Sales with Age group ";

**proc** **sgplot** data=telecom.TENURE\_SEGNEW;

histogram sales / group=age\_group transparency=**0.5** scale=count;

keylegend / location=inside position=topright across=**1**;

**run**;

title;

\*for test of equality of variance;

**proc** **glm** data=telecom.TENURE\_SEGNEW;

class age\_group;

model sales = age\_group;

means age\_group / hovtest=levene(type=abs);

**run**;

\*Test of independancy using one way anova ;

**PROC** **ANOVA** DATA = TELECOM.TENURE\_SEGNEW;

CLASS AGE\_GROUP;

MODEL SALES = AGE\_GROUP;

**RUN**;

\*==============================================================================================================;