**Customer Distribution and Deactivation Analysis**

Project for Advanced SAS

Zihan Zeng

# **Objectives**

The objective of this project is to investigate the customer distribution and business behaviors, and to explore the statistical relationships between the sales data of the telecommunication company and other customer account features, in order to learn the impact of these features and to forecast the deactivation trend.

The basic customer account information in a two-year period of this telecommunication company were obtained and processed in SAS, then it was transformed and segmented for the following statistical analysis.

# **Key Findings**

* Tenure segment is associated with "Status", "GoodCredit", "RatePlan", and "DealerType".
* There is not enough evidence for a significant difference in deactivation rate among the age groups and provinces.
* There is not enough evidence for significant sales differences among different "Status", "GoodCredit", and customer age groups.
* There are significant deactivation rate differences among different groups of "GoodCredit", "RatePlan", and "DealerType".

# **Data Preparation**

## **Import Data**

Data file "New\_Wireless\_Fixed.txt" was imported to SAS work library table "tc".

**data** tc;

infile "C:\Users\zihan\Documents\Metro College\5. SAS\Project\New\_Wireless\_Fixed.txt"

DLM=" " truncover;

input Acctno$ **1**-**13**

@**15** Actdt mmddyy10.

@**26** Deactdt mmddyy10.

DeactReason$ **41**-**44**

@**53** GoodCredit

@**62** RatePlan

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

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

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

\_Sales $ **83**-**92**;

Sales = input(\_Sales,dollar32.);

format Actdt Deactdt mmddyy10. Sales dollar8.2;

drop \_Sales;

**run**;

**proc** **print** data=tc (obs=**50**);**run**;

**proc** **means** data=tc n nmiss;**run**;

**proc** **sql**;

select count(\*) as accts,count(distinct acctno) as unique\_accts

from tc

;**quit**;

**proc** **means** data=tc n nmiss mean median stddev min max;

var Actdt Deactdt Age Sales;

**run**;

**proc** **freq** data=tc;

table DeactReason GoodCredit RatePlan DealerType Province;

**run**;

There are 102,255 observations in this dataset, with 102,255 unique accounts, meaning there is no duplicates. There are 19,635 accounts that has been deactivated during this period, which is a deactivation rate of 19.20%. The dataset starts with the earliest activation date of 1999-01-20, ends on 2001-01-20, while first deactivation occurs on 1999-01-25, ends on 2001-01-20.

Some missing data has been observed in this dataset, Age column has 7,708, Province has 5,907, and Sales column has 8,605 missing observations.

## **Data Transformation**

***Step 1.*** *Create segment for age and sales.*

Proc format segment was created for each column and applied to table "tc" to create a new table "tc1", with new columns "AgeGrp" and "SalesGrp".

**proc** **format**;

value AgeGroup

low-<**20** ='Under 20'

**20** - **40** ='21-40'

**40** - **60** ='41-60'

**61**-high ='Above 60'

;

value SalesGroup

low -<**100** ='Below 100'

**100** - **500** ='100-500'

**500** - **800** ='500-800'

**800** -high ='Above 800'

;

**run**;

**data** tc1;

set tc;

format AgeGrp agegroup. SalesGrp salesgroup.;

AgeGrp = Age;

SalesGrp = Sales;

**run**;

**proc** **print** data=tc1 (obs=**50**);**run**;

***Step 2.*** *Calculate tenure days.*

TenureDays was created by calculating days interval between the account date and the last account date 2001-01-20, or 14995. For deactivated accounts, tenure days is the interval between account date and the deactivation date.

**data** tc1;

set tc1;

\* last actdt = 14995;

if missing(deactdt) then TenureDays = intck("day",actdt,**14995**);

else TenureDays = intck("day",actdt,deactdt);

**run**;

**proc** **print** data=tc1 (obs=**50**);**run**;

***Step 3.*** *Add account status, and segment tenure days.*

Status column was added by checking if deactivation date is null, if a deactivation date exists then Status is Deactivated, else the account is Active. Tenure column was created based on the number of tenure days calculated previously.

**proc** **format**;

value tenuregrp

low- **30** = '<30 days'

**30** - **60** = '31-60 days'

**60** -**365** = '61 days - one year'

**365**-high= 'Over one year'

;**run**;

**data** tc1;

set tc1;

length Status $**11**;

format Tenure tenuregrp.;

if missing(deactdt) then Status = 'Active';

else Status = 'Deactivated';

Tenure = TenureDays;

**run**;

**proc** **print** data=tc1 (obs=**50**);**run**;

# **Exploratory Analysis**

## **Trends and Distributions**

***Q.*** *Report the number of accounts of percent of all for each tenure segment.*

**proc** **freq** data=tc1;

table Tenure /nocum;

**run**;

|  |  |  |
| --- | --- | --- |
| **Tenure** | **Frequency** | **Percent** |
| **<30 days** | 10092 | 9.87 |
| **31-60 days** | 8138 | 7.96 |
| **61 days - one year** | 45389 | 44.39 |
| **Over one year** | 38636 | 37.78 |
| **Total** | 102255 | 100.00 |

***Q.*** *Give simple statistics of tenure days.*

**proc** **means** data=tc1 n mean std min q1 median q3 max qrange maxdec=**1**;

var TenureDays;

**run**;

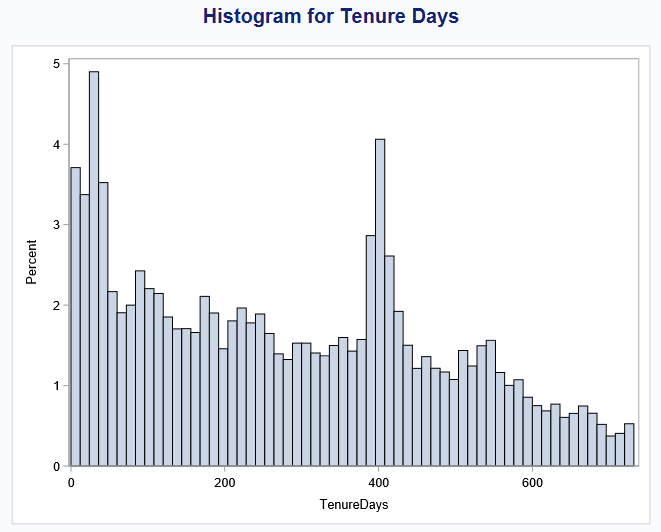
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Analysis Variable : TenureDays** | | | | | | | | |
| **N** | **Mean** | **Std Dev** | **Minimum** | **Lower Quartile** | **Median** | **Upper Quartile** | **Maximum** | **Quartile Range** |
| 102255 | 282.6 | 197.3 | 0.0 | 101.0 | 265.0 | 426.0 | 731.0 | 325.0 |

**proc** **sgplot** data=tc1;

histogram TenureDays;

title "Histogram for Tenure Days";

**run**;title;



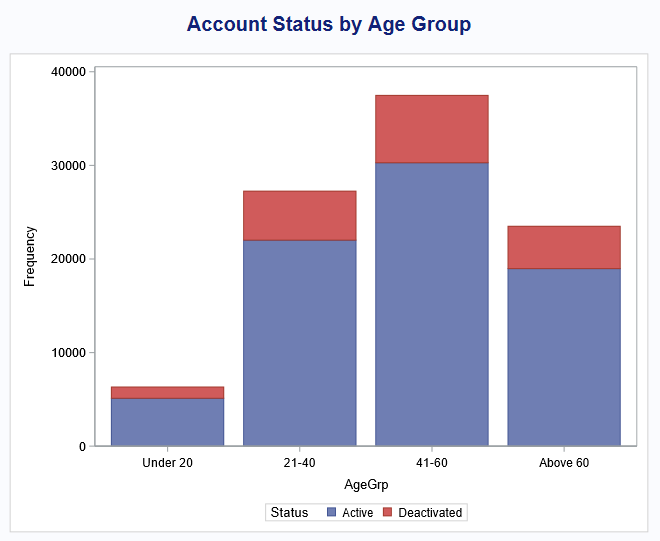
***Q.*** *What is the age and province distributions of active and deactivated customers?*

**proc** **sgplot** data=tc1;

title 'Account Status by Age Group';

vbar AgeGrp /group=Status;

**run**;title;



**proc** **freq** data=tc1;

table AgeGrp\*Status /chisq nopercent;

**run**;

A chi-square test was performed on age group and status features, with a P-value of 0.9057, meaning there is not enough evidence for a significant difference in deactivation rate among the age groups at 5% significance level.

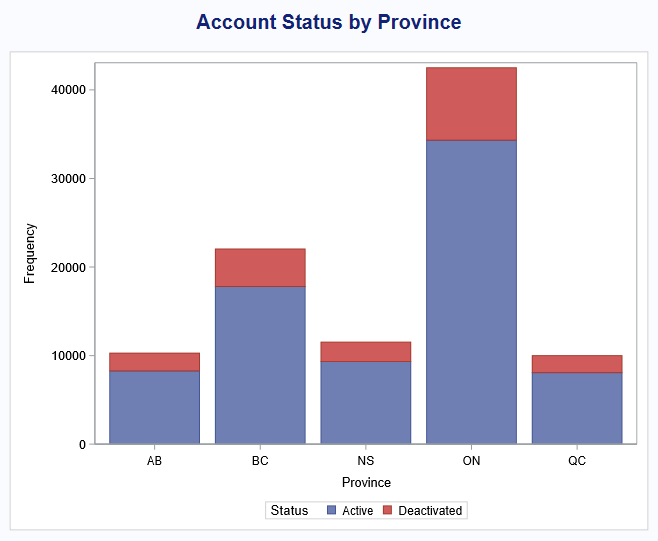
We can see the deactivation rate of each age group is similar to the overall deactivation rate at 19.20%. Observations with missing age feature were omitted by this test as they represent less than 10 percent of overall population.

**proc** **sgplot** data=tc1;

title 'Account Status by Province';

vbar Province /group=Status;

**run**;title;



**proc** **freq** data=tc1;

table Province\*Status /chisq nopercent;

**run**;

A chi-square test was also performed on province and status features, with a P-value of 0.9754, showing that there is also not enough evidence for a significant difference in deactivation rate among the provinces at 5% significance level.

Deactivation rate in each province is similar to the overall deactivation rate. Observations with missing province feature were also not considered in this test.

***Q.*** *Calculate the number of accounts deactivated each month.*

A monthly summary table "tc\_periods" was created by joining the account date by month and the deactivation date by month in a proc sql step, showing the number of accounts created and number of accounts deactivated in each month period.

**proc** **sql**;

create table tc\_periods as

select a.Year, a.Month, a.Period, a.N\_Act, b.N\_Deactivates

from (

select year(actdt) as Year,

month(actdt) as Month,

year(actdt)\***100** + month(actdt) as Period,

count(actdt) as N\_Act

from tc1

group by Year,Month,Period

) as a

left join (

select year(deactdt) as Year,

month(deactdt) as Month,

year(deactdt)\***100** + month(deactdt) as Period,

count(deactdt) as N\_Deactivates

from tc1

where deactdt is not null

group by Year,Month,Period

) b on b.Period = a.Period

order by a.Period

;**quit**;

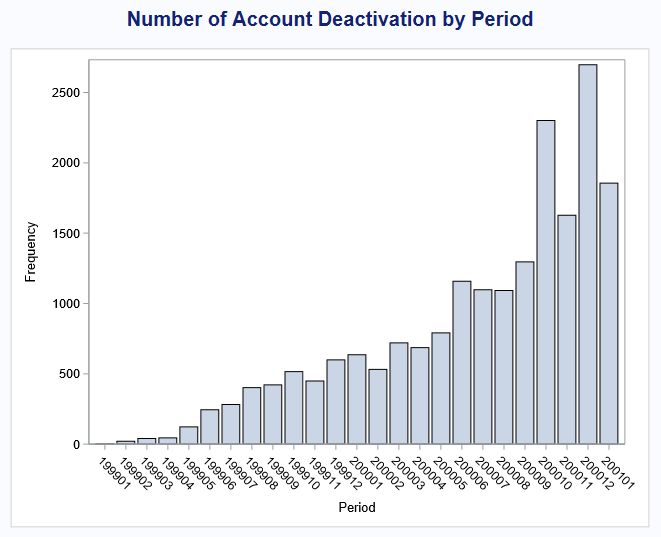
**proc** **print** data=tc\_periods;**run**;

**proc** **sgplot** data=tc\_periods;

title 'Number of Account Deactivation by Period';

vbar period /freq=n\_deactivates;

**run**;title;

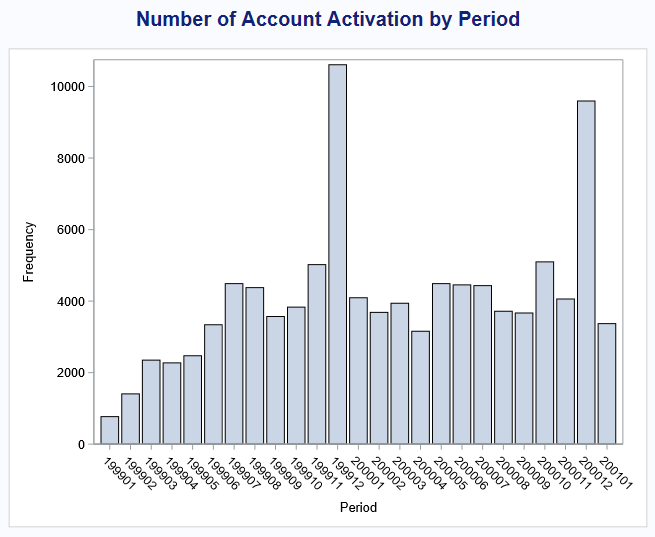


**proc** **sgplot** data=tc\_periods;

title 'Number of Account Activation by Period';

vbar period /freq=n\_act;

**run**;title;



## **Hypothesis Testing**

***Q.*** *Test the general association between the tenure segments and "Good Credit", "Rate Plan", and "Dealer Type".*

Since "Tenure", "GoodCredit", "RatePlan", and "DealerType" are all categorical features, chi-square test is the appropriate test to determine the associations. The null hypothesis for these chi-square tests is that "Tenure" is independent from "GoodCredit", "RatePlan", and "DealerType", respectively.

**proc** **freq** data=tc1;

table Tenure\*GoodCredit /chisq nopercent;

**run**;

Resulting P-value is <0.0001, the null hypothesis that the proportions are equal in all categories is rejected, indicating that Tenure segment is associated with "GoodCredit" at 95% confidence level.

**proc** **freq** data=tc1;

table Tenure\*RatePlan /chisq nopercent;

**run**;

Resulting P-value is <0.0001, the null hypothesis that the proportions are equal in all categories is rejected, indicating that Tenure segment is associated with "RatePlan" at 95% confidence level.

**proc** **freq** data=tc1;

table Tenure\*DealerType /chisq nopercent;

**run**;

Resulting P-value is <0.0001, the null hypothesis that the proportions are equal in all categories is rejected, indicating that Tenure segment is associated with "DealerType" at 95% confidence level.

***Q.*** *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?*

Since both account "Status" and "Tenure" segments are categorical features, a chi-square test was performed to test the association.

**proc** **freq** data=tc1;

table Status\*Tenure Tenure /chisq nopercent;

**run**;

|  |  |  |  |
| --- | --- | --- | --- |
| **Statistic** | **DF** | **Value** | **Prob** |
| **Chi-Square** | 3 | 4340.7439 | <.0001 |

Resulting P-value is <0.0001, the null hypothesis that the proportions are equal in all categories is rejected, meaning that Tenure segment is associated with "Status" at 95% confidence level. From the data we can see that deactivation rate is highest in the "<30 days" segment, and lowest in the "Over one year" segment.

**proc** **freq** data=tc1;

table Status\*Tenure /nofreq norow nopercent;

**run**;

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table of Status by Tenure** | | | | |
|  | **<30 days** | **31-60 days** | **61 days - one year** | **Over one year** |
| **Active** | 68.16 | 86.04 | 74.82 | 90.01 |
| **Deactivated** | 31.84 | 13.96 | 25.18 | 9.99 |

The current way of segmentation for tenure days cuts the dataset into 4 groups by using 30/60/one year as cut points, which results in deactivation rates at 31.84%, 13.96%, 25.18%, and 9.99% for each tenure group.

To analyze the deactivation rate more closely, a new table "tc\_tenure" is created by summarizing the account status by tenure days.

**proc** **sql**;

create table tc\_tenure as

select TenureDays, count(\*) as n, count(deactdt) as n\_deactive

from tc1

group by TenureDays

order by TenureDays

;**quit**;

Cumulative deactivation rate is calculated as "cum\_deact\_rate" starting from tenure days of 0.

**data** tc\_tenure;

set tc\_tenure;

retain nsum n\_deactsum;

nsum = sum(nsum,n);

n\_deactsum = sum(n\_deactsum,n\_deactive);

cum\_deact\_rate = n\_deactsum / nsum;

drop nsum n\_deactsum;

daily\_deact\_rate = n\_deactive / n;

**run**;

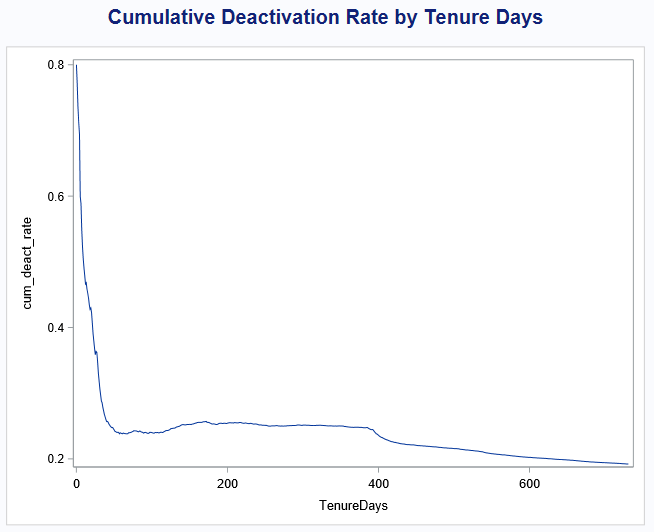
**proc** **print** data=tc\_tenure;**run**;

**proc** **sgplot** data=tc\_tenure;

title 'Cumulative Deactivation Rate by Tenure Days';

series x=TenureDays y=cum\_deact\_rate;

**run**;title;



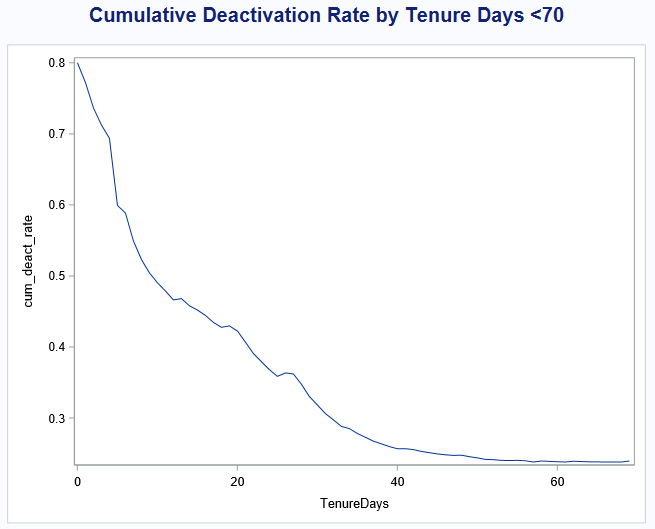
The change of slope in the above plot represents a change in deactivation rate pattern. By reading the detailed table, it was found that the cumulative rate starts at 80% on days 0 and drops to a steady range around 25% near days 60, and remains stable until days 400 where the rate then starts to decline. These two points, 60 days and 400 days, are good cutting points for tenure segments.

**proc** **sgplot** data=tc\_tenure (obs=**70**);

title 'Cumulative Deactivation Rate by Tenure Days <70';

series x=TenureDays y=cum\_deact\_rate;

**run**;title;

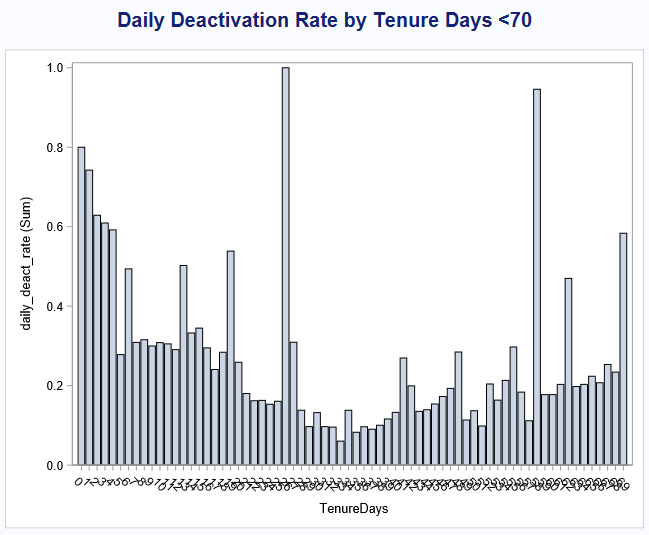


**proc** **sgplot** data=tc\_tenure (obs=**70**);

title 'Daily Deactivation Rate by Tenure Days <70';

vbar TenureDays /response=daily\_deact\_rate;

**run**;title;



If we take a close look at the first 70 days, we can see the deactivation rate for the first 5 days (0-4 days) are significantly higher than the following days. The pattern then changes again at around days 20. These two points are also good cutting points for tenure segment.

The four cutting points (4, 20, 60, 400) is then used to create five new tenure segments, "<5 days", "5-20 days", "21-60 days", "61-400 days", and "Over 400 days".

**data** tc\_tenure1;

set tc\_tenure;

length Tenure $**13**;

if TenureDays < **5** then Tenure = '<5 days';

else if TenureDays <= **20** then Tenure = '5-20 days';

else if TenureDays <= **60** then Tenure = '21-60 days';

else if TenureDays <= **400** then Tenure = '61-400 days';

else Tenure = 'Over 400 days';

**run**;

**proc** **print** data=tc\_tenure1;**run**;

**proc** **sort** data=tc\_tenure1 out=tc\_tenure1;

by Tenure;

**run**;

**data** tc\_tenure1;

set tc\_tenure1;

by Tenure;

retain nsum n\_deactsum;

if first.Tenure then do;

nsum = n;

n\_deactsum = n\_deactive;

cum\_grp\_rate = n\_deactsum/nsum;

end;

else do;

nsum = sum(nsum,n);

n\_deactsum = sum(n\_deactsum,n\_deactive);

cum\_grp\_rate = n\_deactsum/nsum;

end;

drop nsum n\_deactsum;

**run**;

**proc** **sort** data=tc\_tenure1 out=tc\_tenure1;

by TenureDays;

**run**;

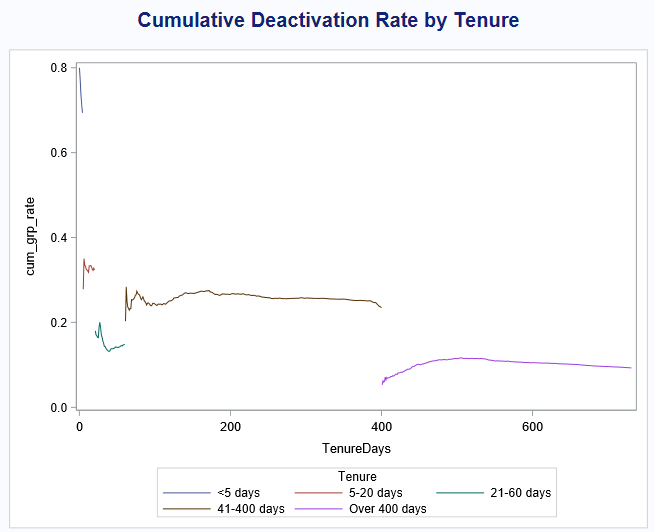
**proc** **print** data=tc\_tenure1;**run**;

**proc** **sgplot** data=tc\_tenure1;

title 'Cumulative Deactivation Rate by Tenure Days';

series x=TenureDays y=cum\_grp\_rate /group=Tenure;

**run**;title;



The cumulative deactivation rate of each new tenure segment is relatively constant.

**proc** **format**;

value newtenuregrp

low-<**5** = '<5 days'

**5** - **20** = '5-20 days'

**20** - **60** = '21-60 days'

**60** -**400** = '61-400 days'

**400**-high= 'Over 400 days'

;**run**;

**data** tc2;

set tc1;

format Tenure newtenuregrp.;

Tenure = TenureDays;

**run**;

**proc** **print** data=tc2 (obs=**50**);**run**;

**proc** **freq** data=tc2;

table Status\*Tenure /chisq nofreq norow nopercent;

**run**;

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table of Status by Tenure** | | | | | |
|  | **<5 days** | **5-20 days** | **21-60 days** | **61-400 days** | **Over 400 days** |
| **Active** | 30.63 | 67.61 | 85.19 | 76.49 | 90.72 |
| **Deactivated** | 69.37 | 32.39 | 14.81 | 23.51 | 9.28 |
| **Total** | 1603 | 4409 | 12218 | 52635 | 31390 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Statistic** | **DF** | **Value** | **Prob** |
| **Chi-Square** | 4 | 5865.4389 | <.0001 |

Applying the chi-square test on the new tenure segment with Status also results in a P-value of <0.0001, indicating that there is an association between the two features at 95% confidence level. This new chi-square value 5865.4 is greater than the chi-square value 4340.7 of the original tenure segments, showing a stronger association between the Status and the new tenure segments.

***Q.*** *Does Sales amount differ among different account status, GoodCredit, and customer age segment?*

The sales amount is a continuous variable, while Status, GoodCredit, and AgeGrp are categorical variables. In order to test the association of sales amount with the other three features, some assumptions on the data have to be verified before using any hypothesis tests.

First, the dataset has the same number of unique accounts as the number of total observations, so the categorical features can be seen as independent.

**proc** **univariate** data=tc1 normal plot;

class Status;

var Sales;

histogram;

**run**;

**proc** **univariate** data=tc1 normal plot;

class GoodCredit;

var Sales;

histogram;

**run**;

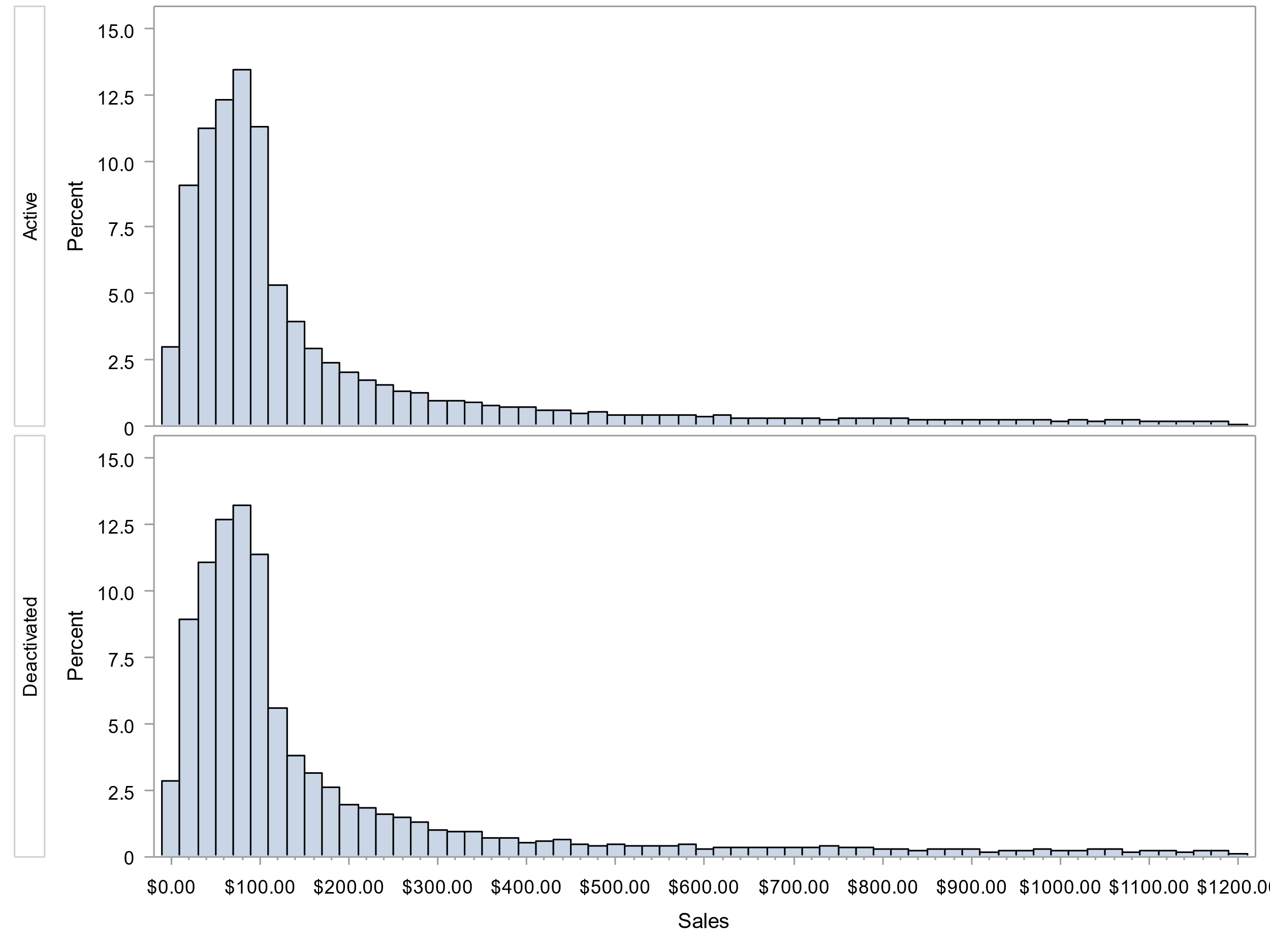
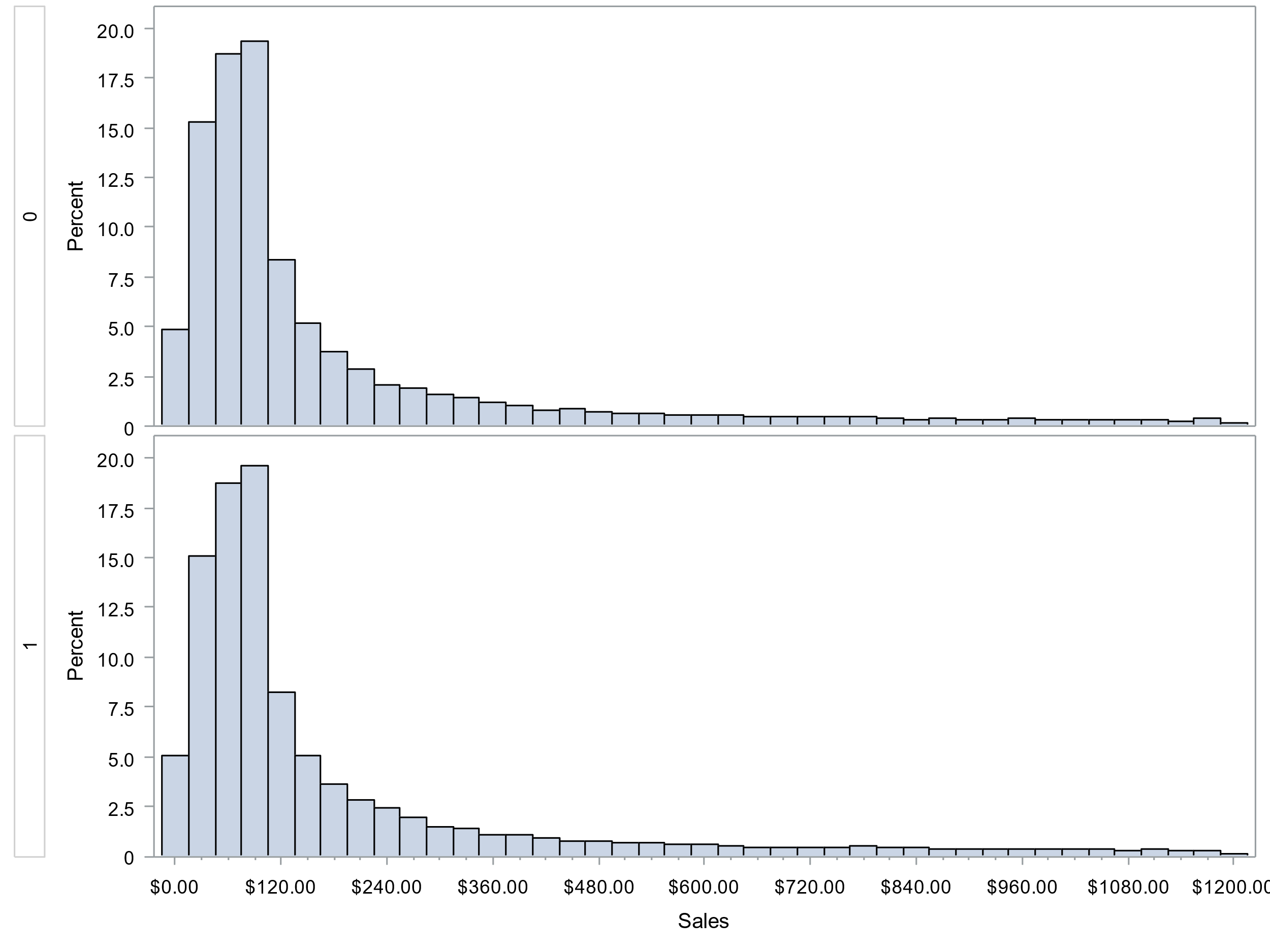
**proc** **univariate** data=tc1 normal plot;

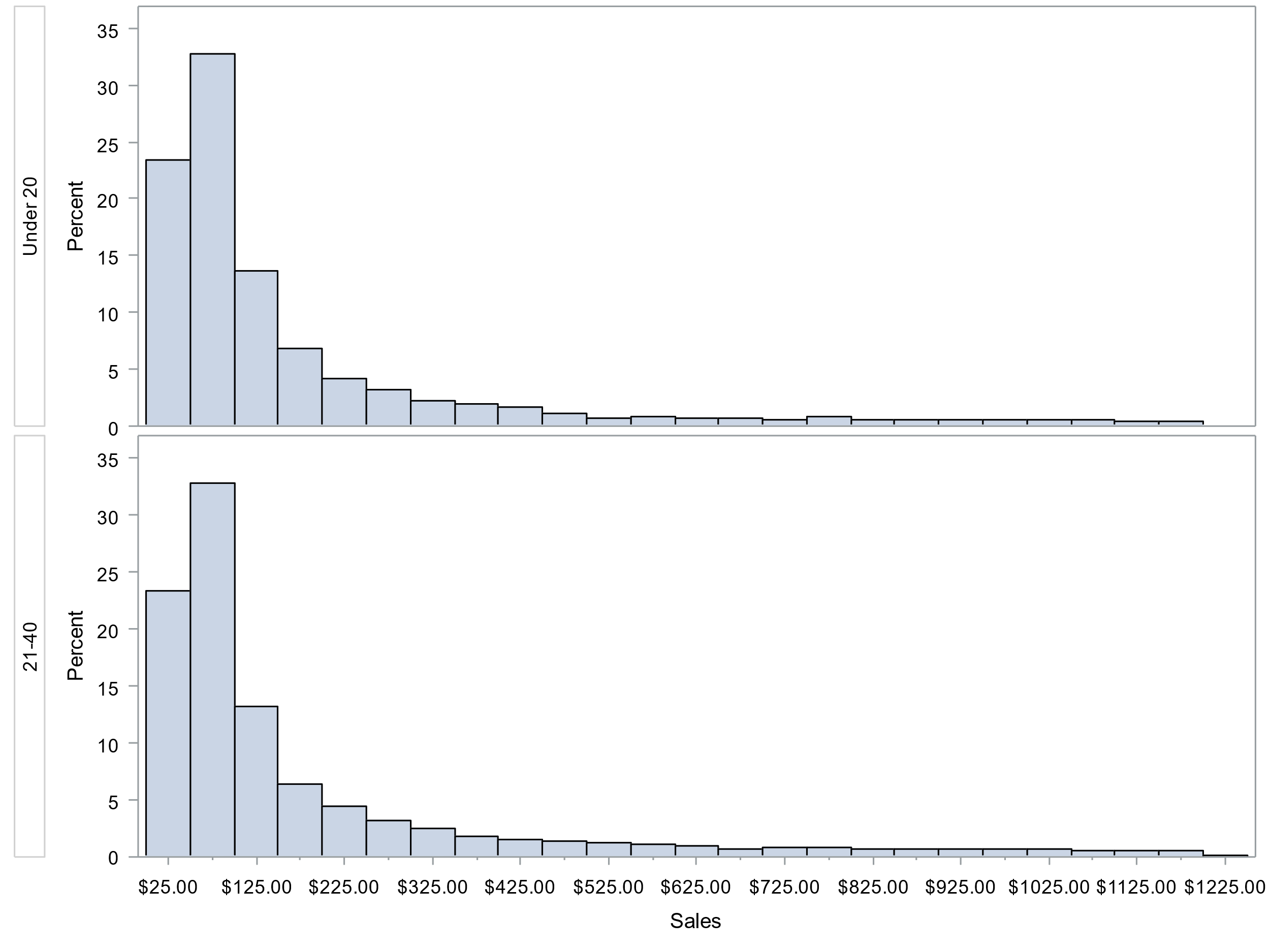
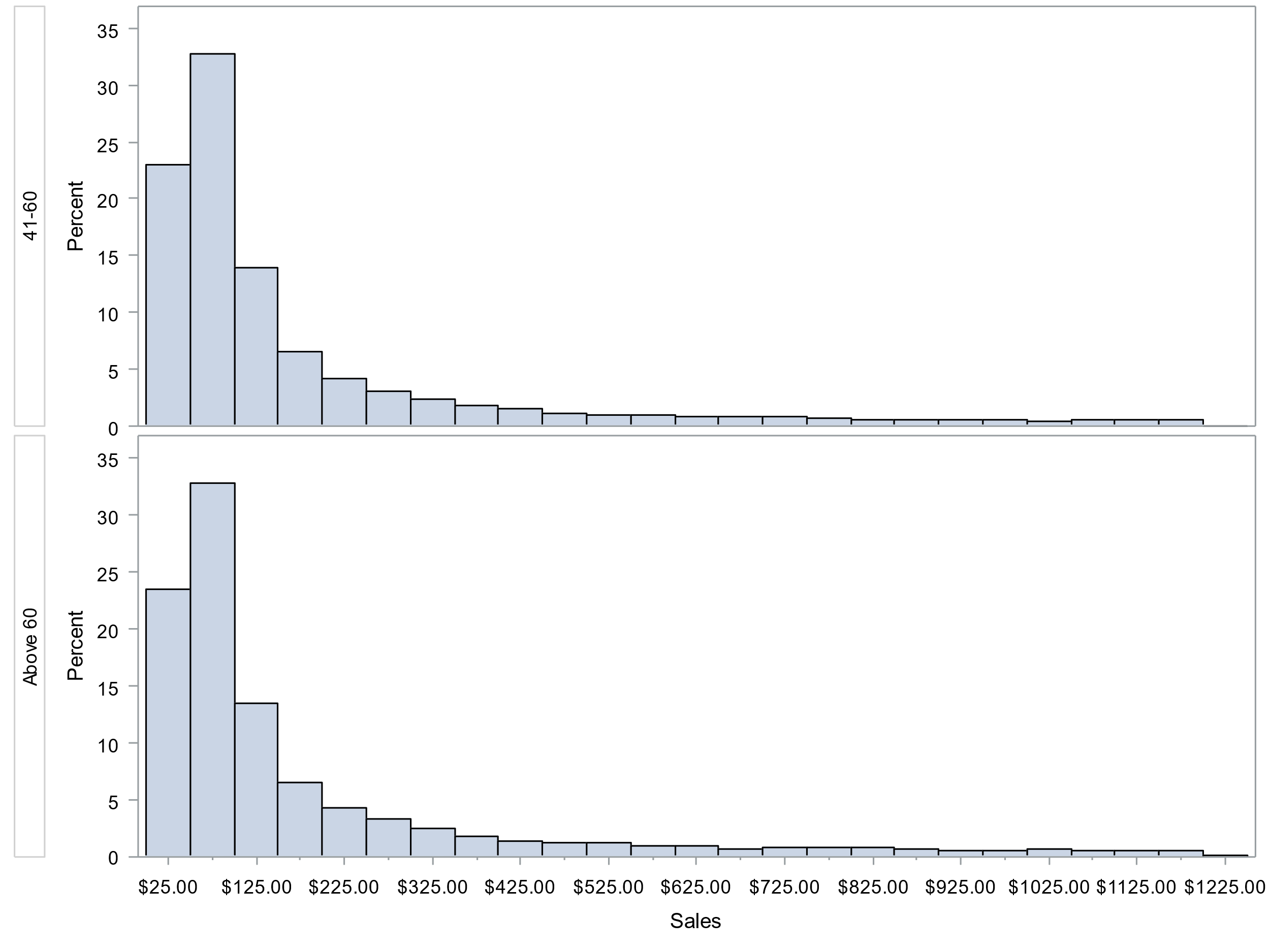
class AgeGrp;

var Sales;

histogram;

**run**;

By checking the number of observations in each category, normality can be assumed by applying central limit theorem, as all categories have more than 30 frequencies.

**proc** **means** data=tc1 std min q1 median q3 max qrange maxdec=**2**;

class Status;

var Sales;

**run**;

**proc** **sql**;

select Status, count(\*) as n\_outlier

from tc1

where (Sales > **191**+**3**\***139** and deactdt is null)

or (Sales > **188**+**3**\***135** and deactdt is not null)

group by Status

;**quit**;

**proc** **sgplot** data=tc1;

vbox Sales /group=Status;

**run**;

The histograms from the above proc univariate step shows the sales data is highly positively skewed, with many outliers above Q3+3\*IQR, so a logarithm transformation is performed.

**data** tc\_log;

set tc1;

SalesLog = log(Sales);

**run**;

**proc** **print** data=tc\_log (obs=**30**);**run**;

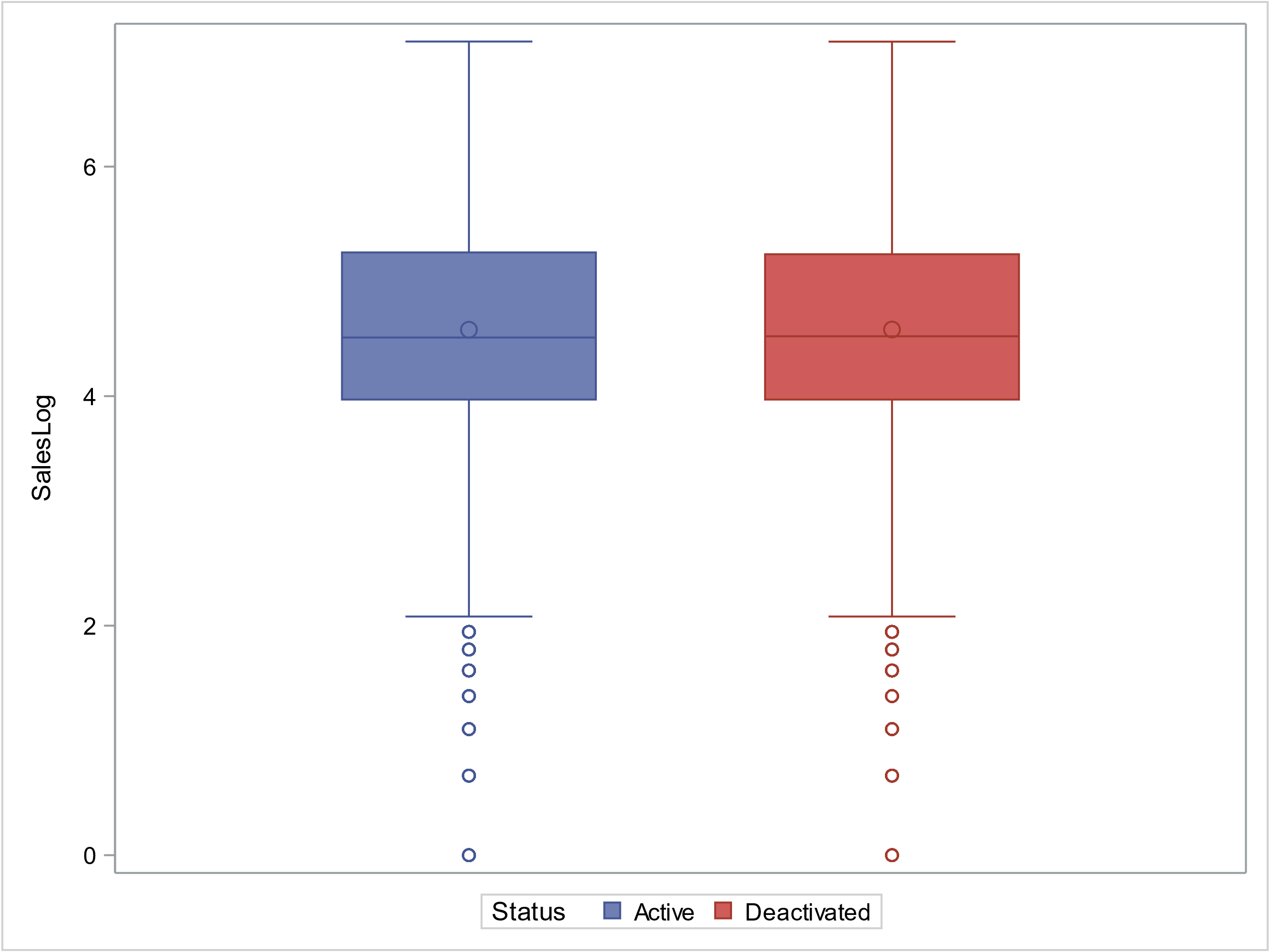
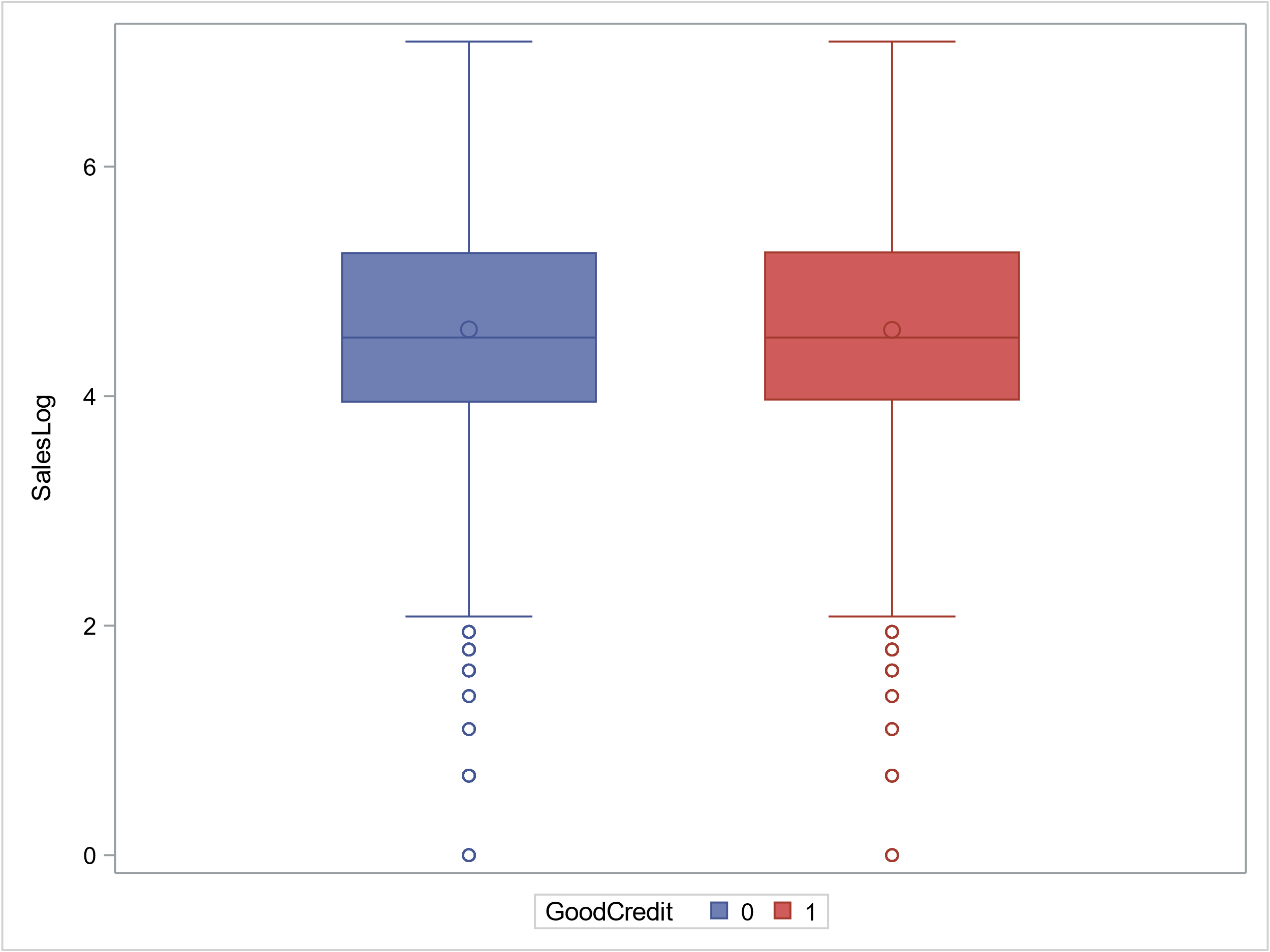
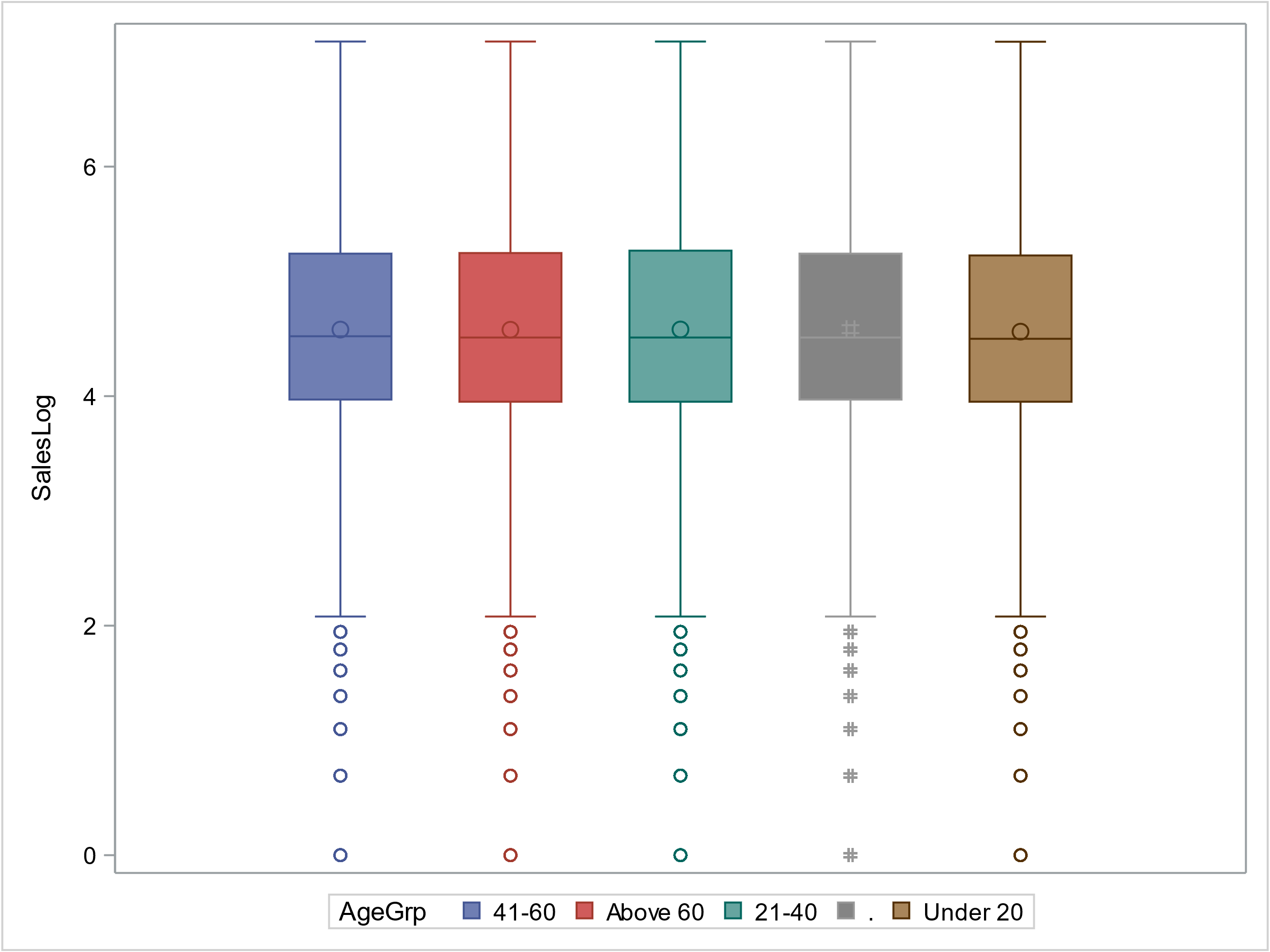
**proc** **means** data=tc\_log std min q1 median q3 max qrange maxdec=**2**;

var SalesLog;**run**;

**proc** **sgplot** data=tc\_log;

vbox SalesLog /group=Status;

**run**;

To check the equality of variances, Leven's test for homoscedasticity is applied to the log transformed sales data.

**proc** **glm** data=tc\_log;

class Status;

model SalesLog=Status;

means Status / hovtest=levene(type=abs);

**run**;

The equality of variances check for Status classification in Levene's test returns a P-value of 0.1012, so the null hypothesis of equal variances is not rejected. Since Status is a two-group feature, a pooled t-test should be used for Status feature.

**proc** **ttest** data=tc\_log;

class Status;

var SalesLog;

**run**;

The pooled t-test returns a P-value of 0.8416, failing to reject the null hypothesis that the mean sales of the two Status groups are the same. We can conclude that there is not enough evidence for sales difference between the two Status groups at 5% significance level.

**proc** **glm** data=tc\_log;

class GoodCredit;

model SalesLog=GoodCredit;

means GoodCredit / hovtest=levene(type=abs);

**run**;

The equality of variances check for GoodCredit classification in Levene's test returns a P-value of 0.7985, so the null hypothesis of equal variances is not rejected. Since GoodCredit is a two-group feature, a pooled t-test should be used for GoodCredit feature.

**proc** **ttest** data=tc1 dist=lognormal;

class GoodCredit;

var Sales;

**run**;

The pooled t-test returns a P-value of 0.6067, failing to reject the null hypothesis that the mean sales of the two GoodCredit groups are the same. We can conclude that there is not enough evidence for sales difference between the two GoodCredit groups at 5% significance level.

**proc** **glm** data=tc\_log;

class AgeGrp;

model SalesLog=AgeGrp;

means AgeGrp / hovtest=levene(type=abs) welch;

**run**;

The equality of variances check for AgeGrp classification in Levene's test returns a P-value of 0.7985, so the null hypothesis of equal variances is not rejected. Since AgeGrp is a multi-group feature, a one-way Anova test should be used for AgeGrp feature.

The one-way Anova test in the above proc glm step returns a P-value of 0.7092, failing to reject the null hypothesis that the mean sales of the AgeGrp groups are the same. We can conclude that there is not enough evidence for sales difference among the customer age groups at 5% significance level.

***Q.*** *Does deactivation rate differ among different "GoodCredit", "RatePlan", and "DealerType" groups?*

Since all these features are categorical, chi-square test is used to test for associations.

**proc** **freq** data=tc1;

table Status\*GoodCredit Status\*RatePlan Status\*DealerType /chisq nopercent;

**run**;

All three chi-square tests returns P-value of <0.0001, the null hypothesis of independence is rejected, indicating that there are significant differences of deactivation rates among the different groups of "GoodCredit", "RatePlan", and "DealerType" at 95% confidence level.

***Q.*** *Which features are associated with deactivation reasons?*

Since all these features are categorical, chi-square test is used to test for associations.

**proc** **freq** data=tc1;

table DeactReason\*Tenure DeactReason\*AgeGrp DeactReason\*GoodCredit

DeactReason\*RatePlan DeactReason\*DealerType /chisq nofreq nopercent;

**run**;

The features that returns a chi-square test P-value less than 0.05 are Tenure, GoodCredit, RatePlan, and DealerType, indicating there are differences in deactivation reasons among these groups at 5% significance level.