TECHNICAL REPORT

Mini-Project

“\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_”

Advanced SAS Programming (ASP)

By

<< Kaiyao Qin, 1118-01992: >>

Metro College of Technology

2019

TABLE OF CONTENTS

PROJECT PURPOSES …………………………………………………………………………………………………………..<page3>

IMPLEMENTATION DETAILS …….………………………………………………………………………………….…….. <page4-5>

RESULTS AND DISCUSSIONS ……………………………………………………………………………………………….. <page6-9>

CONCLUSIONS AND RECOMMENDATIONS …………………………………………………………………………. <page10>

APPENDIX A. SAS CODE ……………………………………………………………………………………………………… <page11-19>

PROJECT PURPOSES

The purpose of this project is to investigate the customer distribution and business behaviors of a wireless company for 2 years, and then gain insightful understanding about the customers, and to forecast the deactivation trends for the next 6 months. The followings are some useful analysis propositions designed for this project:

**1.** General descriptive statistics about the dataset

**2.** Age and province distributions of active and deactivated customers

**3.** Analysis report based on the Segmented customers 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.

**4.** Statistical Analysis:

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

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

*3) Forecast the number of accounts that will deactivate for the subsequent 6 months.*

4) 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.

5) Test the general association between the tenure segments and “Good Credit”

“RatePlan ” and “Dealer Type”, respectively.

6) Test the general association between the account status and “Good Credit”

“RatePlan ” and “DealerType” , respectively.

7) Any significant association between the account status and the tenure segments?

8) Any significant sales amount differerence among different account status, GoodCredit, and

customer age segments, respectively?

IMPLEMENTATION DETAILS

The dataset was imported into SAS window as **ds** to do the following analysis:

**1.** For general descriptive statistics, “proc sql” was used to explore the count of unique accounts, exact time span of the dataset, the count of customers that is still active and deactivated, respectively.

**2.** A dataset **ds2a** was created, containing customers’ activation status. Then “proc univariate” histogram was used to illustrate the age (numerical variable) distribution for active and deactivated customers, respectively. “proc freq” frequency table was used to illustrate the province (categorical variable) distribution for active and deactivated customers, respectively.

**3.** A new dataset **ds1**, containing segments of sales, age and activation status was created. Based on the new dataset, two (active, deactivated) 3-ways (province, age \* sale segments) frequency table analysis were implemented using “proc tabulate”

**4.** Statistical Analysis:

1) A new dataset **tenure0** was created by adding new column ‘tenure’ which is the time span between the activation date and the deactivation date for the deactivated customers, and the time span between the activation date and fiscal date of this dataset ('03/31/2001') for the active customers. Then their respective frequency distribution was done by “proc univariate” histogram.

2) A dataset **test** was created, containing only the deactivated customers and their deactivation month. Then a simple frequency distribution table of deactivation by month was created using “proc freq”

*3)* A dataset **forcast1** was created containing the count of deactivation thru out the time span (JAN1999 to JAN2001) of the dataset by monthly interval. Then, a time series analysis procedure “proc forecast” was implemented on this newly created dataset to forecast the number of accounts that will be deactivated for the subsequent6 months (FEB2001 to JUL2001) –result is in **forcast\_6month**.

4) A new dataset **ds2**, containing segments of tenure and activation status was created. Based on the new dataset, a two-way frequency table on tenure segments by activation status was implemented using “proc tabulate”.

5) To test the general association between the tenure segments and “Good Credit”,“RatePlan ” and “DealerType”, respectively, one chi-square analysis test was created for each of the three 2-way frequency tables using chisq statement in “proc freq”

6) To test the general association between the account status and “Good Credit”, “RatePlan ” and “DealerType” , respectively, one chi-square analysis test was created for each of the three 2-way frequency tables using the chisq statement in “proc freq”

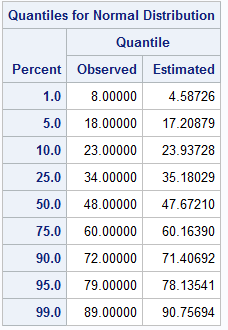
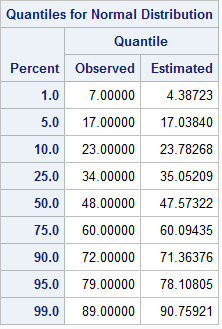
7) To test the significance of the association between the account status and the tenure segments, again a chi square test was used by using the chisq statement in “proc freq”

8) A new dataset **ds3**, containing segments of age was created. To test the significance of Sales amount difference among different account status, and different credit status, a non-parametric T-test (Wilcoxon rank-sum test) was created for each of them using “proc NPAR1WAY”. (Since only two groups to be compare at a time)

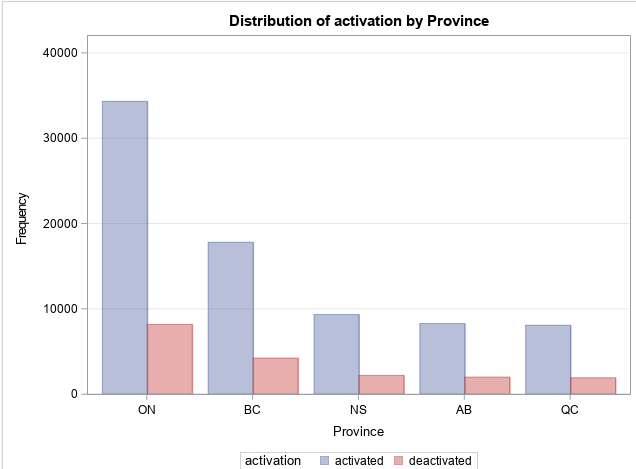
To test the significance of Sales amount difference among different age segments, a non-parametric ANOVA test (Wilcoxon rank-sum test) and the subsequent post-hoc test were implemented using “proc NPAR1WAY”. (Since more than two groups to be compared)

RESULTS AND DISCUSSIONS

**1.** There is a total of 102255 observations with 102255 unique accounts, which means each customer only holds one account. The time span of this dataset is approximately 2 years from 01/20/1999 (earliest activation date) to 03/31/2001(fiscal date), and the latest deactivation date is 01/20/2001. The number of deactivated customers is 19635 (19.2%), while the number of active customers is 82620 (80.8%).

**2.** The mean age for active customers is 48, and the distribution of it is (left). The mean age for deactivated customers is also 48, and the distribution of it is (right):

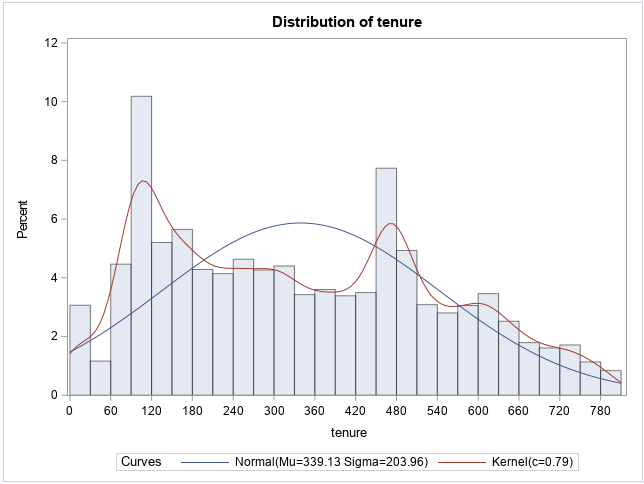
According to the Goodness-of-Fit Tests for Normal Distribution, the age for both active and deactivated customers is NOT normally distributed (p-value <0.05).

 The province distribution for active and deactivated customers is the following:

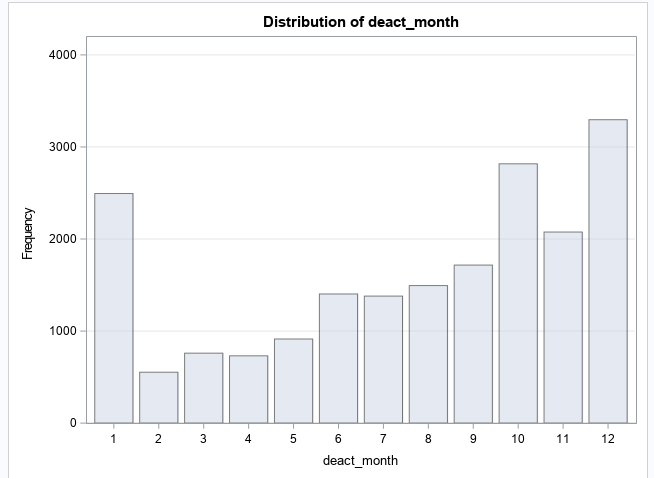
According to the chi-square p value (close to 1; >0.05) of the frequency table, province and activation status are not associated.

**3.** Please refer to “Q3.xls”

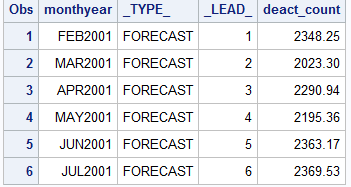
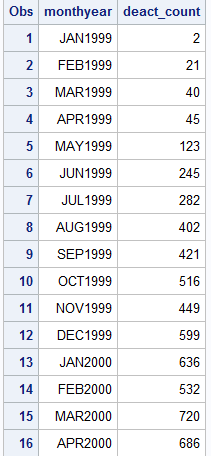
**4.** Statistical Analysis:

1)

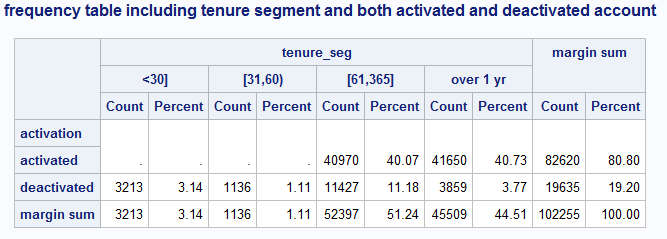
The minimum tenure is 0, the max tenure is 801, the median is 339, and the mean is 319. From the histogram, it is quite obvious that most of the customers prefer to deactivate in their 4th month, or in the 4th month after a year.

2)

It is quite obvious that customers often deactivate the plan in the winter time (Oct to Jan).

3)the following one on the (left) is the frequency count of deactivations by month interval for the existing data. The one on the (right) is the forecast (with time series analysis: proc forecast) of deactivation count for the subsequent 6 months (FEB2001 to JUL2001).

4) A two-way frequency table of tenure segment and activation status:



5) There is a significant association between (tenure segments) and (“Good Credit”, “RatePlan” and “DealerType”), respectively. (the chi-square p value for each of the three association tests is close to zero; <0.05 )

6) There is a significant association between (account status) and (“Good Credit”, “RatePlan” and “DealerType”), respectively. (the chi-square p value for each of the three association tests is close to zero; <0.05)

7) There is a significant association between account status and tenure segments. (the chi-square p value for the association test is close to zero; <0.05)

8) according to p-value of Wilcoxon (0.9008) and median (0.4823) t-test (b/c the assumption of normality and equal variance are both not met- thus non-parametric (standard t test works only for parametric data)), there is no significant sales amount difference among different account status.

according to p-value of Wilcoxon (0.8938) and median (0.8762) t-test (b/c the assumption of equal variance is met, but the assumption of normality is not met- thus non-parametric (standard t test works only for parametric data)), there is no significant sales amount difference among different credit groups.

according to p-value of Wilcoxon (0.7656) and median (0.3784) ANOVA test (b/c the assumption of equal variance and normality are not met- thus non-parametric (standard ANOVA works only for parametric data)), there is no significant sales amount difference among different age segments.

CONCLUSION AND RECOMMENDATIONS

First of all, each customer has only one unique account, which make it more direct and easier to track customers’ behavior. In the 2 years span, 20% of the customers canceled/deactivated their services, which is something that can be improved on. Since the forecast analysis showed an increasing trend of deactivation in the subsequent 6 months (from February to July 2001), It is better to have some strategies to cop with it, so that customer loss can be prevented in some extent.

According to the result of analysis 2, province and activation status are not associated. However, we still need to focus on customers from Ontario, because it is where most of our customers come from. If we hold onto the market in Ontario, and reduce the customer loss by giving special offers or reasonable discount to the regular Ontario customers, we can definitely increase our revenue,

According to the result of analysis 4(1), deactivations mostly occur in the 4th month of customers’ tenure, and the 4th months in the 2nd year of their tenure. It might be a good idea to give the customers some special offer like a gift or an extra tv channel package before these periods.

From the result of analysis 4(2), most frequent deactivation occurs in winter time (Oct to Jan). Therefore, special offers or an extra investment in advertisement in winter period is necessary.

According the result of analysis 4(5) and 4(6), “Good Credit”, “RatePlan” and “DealerType” are associated with tenure segments and activation status. If we use tenure of over 1 year or active account as a guideline, customers with a good credit, 1st rateplan, and A1 dealertype will likely to not deactivate the service, or have a tenure over 1 year. Therefore, do take these attributes into account when assessing or predicting the tenure and activation status of our customers in the future.

According to the result of analysis 4(7), tenure segments and activation status are related. For the active customers, all of them have a tenure at least 2 months, and approximately half of them have a tenure over 1 year. For deactivated customers, most of them have a tenure between 2 month and 1 year. Therefore, it is really important to provide a good experience to the customers, especially between the 2nd month and first year of their tenure.

At the end, according to sales analysis in 4(8), there is no significant sales amount difference among different activation status, different credit groups, or different age segments. Therefore, the service quality and attitude from our employees to a new customer should not be discriminated by his/her age, activation status and credit.

APPENDIX A: SAS CODE

/\*version 1 (non-delimited ds)\*/

filename df 'C:\Users\Jesus kid\Desktop\kqylib\New\_Wireless\_Fixed.txt';

**data** ds;

infile df;

format acctno $13. actdt mmddyy10. deactdt mmddyy10. deactreason $4. goodcredit **1.** rateplan **1.** dealertype $2.

AGE **2.** Province $2. sales dollar10.2;

input @**1** acctno $13. @**15** actdt mmddyy10. @**26** deactdt mmddyy10. @**41** deactreason $4. @**53** goodcredit **1.** @**62** rateplan **1.** @**65** dealertype $2.

@**74** AGE **2.** @**80** Province $2. sales dollar10.2;

**run**;

**proc** **print** data= ds (obs=**10**);

**run**;

**proc** **contents** data=ds;

**run**;

/\*Q1: number of unique values in acctno variable\*/

**proc** **sql**;

select count(distinct acctno) as unique\_account\_count

from ds

;

**quit**;

/\* When is the earliest andv latest activation and deactivation dates available, respectively?\*/

/\*sort by actdt\*/

**proc** **sort** data= ds out=sort;

by actdt;

**run**;

/\*first and last 5 observation (sorted by actdt)\*/

**data** first\_last/view= first\_last;

set sort nobs=\_\_nobs;

if \_n\_ le **5** or \_n\_ gt \_\_nobs-**5**;

**run**;

**proc** **print** data=first\_last;

TITLE 'first and last 5 observation (sorted by actdt)';

**run**;

/\*sort by deactdt (exclude missing value)\*/

**proc** **sort** data= ds out=sort1;

by deactdt;

Where not missing(deactdt);

**run**;

/\* first and last 5 observation (sorted by deactdt, exclude missing value)\*/

**data** first\_last1/view= first\_last1;

set sort1 nobs=\_\_nobs;

if \_n\_ le **5** or \_n\_ gt \_\_nobs-**5**;

**run**;

**proc** **print** data=first\_last1;

TITLE 'first and last 5 observation (sorted by deactdt, exclude missing value)';

**run**;

/\*count of customers with both activation and deactivation\*/

**proc** **sql**;

TITLE 'count for deactivated customers';

select count(\*) as act\_deact\_count

from ds

where actdt IS NOT NULL and deactdt is not null;

**quit**;

**proc** **sql**;

TITLE 'count for active customers';

select count(\*) as act\_deact\_count

from ds

where actdt IS NOT NULL and deactdt is null;

**quit**;

/\*Q2: distribution of age grouped by province for customers with both act and deact\*/

**proc** **format**;

value act\_status

**.** = 'activated'

other = 'deactivated';

/\*dataset with activation and deactivation \*/

**data** ;

set ds;

activation = put(deactdt,act\_status.);

**run**;

**proc** **sql**;

TITLE 'max age';

select max (age) from ds;**quit**;

**proc** **univariate** data= ds2a;

TITLE 'age distribution for active and deactivated customers';

class activation;

var age;

histogram age/ endpoints=(**0** to **99** by **5**) normal kernel;

**run**;

**proc** **freq** data=ds2a order=freq;

TITLE 'province distribution for active and deactivated customers';

tables activation\*province / expected norow nocol cellchi2 chisq nopercent

plots=freqplot(twoway= cluster orient=vertical);

**run**;

/\*http://support.sas.com/documentation/cdl/en/proc/61895/HTML/default/viewer.htm#a002473474.htm\*/

/\*Q3: Segment the customers based on age, province and sales amount:\*/

**proc** **format**;

value saleFmt

low -< **100** = "<100)"

**100** -< **500** = "[100,500)"

**500** -< **800** = "[500,800)"

**800** - high = ">800]"

;

value AgeFmt

low - **20** = "<20]"

**21** - **40** = "[21,40]"

**41** - **60** = "[41,60]"

**61** - high = ">61]"

;

**run**;

/\*dataset containning sale\* age segment, and province \*/

**data** ds1;

set ds;

format sales saleFmt. age AgeFmt. ;

**run**;

ods MSOffice2K body='C:\Users\Jesus kid\Desktop\kqylib\Q3.xls';

**Proc** **Tabulate** data=ds;

TITLE '(active) freq table of province , age \* sale segments ';

Class province age sales;

Table (province ALL), (age ALL)\*(sales ALL);

where deactdt is null and actdt is not null;

Keylabel ALL= 'Total';

format sales saleFmt. age AgeFmt. ;

**Run**;

**Proc** **Tabulate** data=ds ;

TITLE '(deactivated) freq table of province , age \* sale segments ';

Class province age sales;

Table (province ALL), (age ALL)\*(sales ALL);

where deactdt is not null and actdt is not null;

Keylabel ALL= 'Total';

format sales saleFmt. age AgeFmt. ;

**Run**;

ods MSOffice2K close;

/\*Q4. Statistical Analysis:\*/

/\*1) Calculate the tenure in days for each account \*/

/\*dataset containning tenure\*/

**data** tenure0 (drop=fiscal\_date) ;

set ds ;

format end\_date MMDDYY10.;

fiscal\_date = input('03/31/2001',MMDDYY10.);

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

if deactdt=**.** then tenure= intck('day', actdt, end\_date);

**run**;

/\*min and max tenure\*/

**proc** **sql**;

TITLE 'min and max tenure ';

select min(tenure) as min, max(tenure) as max

from tenure0;

**quit**;

/\*tenure distribution\*/

**proc** **univariate** data= tenure0 ;

TITLE 'tenure distribution ';

var tenure;

histogram tenure/ endpoints=(**0** to **801** by **30**) normal kernel;

**run**;

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

/\*create a table only containing deactivated cutomers with deactivated month\*/

**proc** **sql**;

create table act\_deact as

select \*

from ds

where actdt IS NOT NULL and deactdt is not null;

**quit**;

**data** test;

set act\_deact;

deact\_month = put(deactdt, month.); \*convert datetime to month format, save it into new category variable;

/\*put statement can convert datetime to number of days from 1960 1 1:

num\_days= PUT(deactdt, $comma6.);

\*/

**run**;

/\* frequency distribution of accounts deactivated for each month.\*/

/\*method 1\*/

**Proc** **Tabulate** data=test;

TITLE 'frequency distribution of accounts deactivated for each month (1) ';

Class deact\_month;

Table (deact\_month all);

where deactdt is not null and actdt is not null;

Keylabel ALL= 'Total';

**Run**;

/\*method 2\*/

**proc** **freq** data=test ;

TITLE 'frequency distribution of accounts deactivated for each month (2) ';

tables deact\_month/

plots=freqplot(twoway= cluster orient=vertical);

**run**;

/\*3)Forecast the number of accounts that will deactivate for the subsequent 6 months.\*/

**data** forcast;

set ds;

format mmyy monyy7.;

mmyy = input(put(deactdt,monyy7.), monyy7.);

**run**;

**proc** **sql**;

create table forcast1 as

select distinct mmyy as monthyear, count(\*) as deact\_count

from forcast

where monthyear <> **.**

group by monthyear

order by monthyear

;

**run**;

**proc** **print** data= forcast1;

title"count of deactivation thru out the time span of the dataset (monthly interval)";

**run**;

**proc** **forecast** data=forcast1 interval=month lead=**6** out=forcast\_6month;

id monthyear;

var deact\_count;

**run**;

**proc** **print** data= forcast\_6month;

title"forecast of the subsequent 6 months after the end date of the original dataset";

**run**;

/\*4) 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.\*/

**proc** **format**;

value deact

**.** = 'activated'

other = 'deactivated' ;

value tenure

low - **30** = "<30]"

**31** - **60** = "[31,60)"

**61** - **365** = "[61,365]"

**366** - high = "over 1 yr"

;

**run**;

/\*dataset containning activation status and tenure segment\*/

**data** ds2;

set tenure0;

activation = put(deactdt,deact.);

tenure\_seg= put(tenure,tenure.);

**run**;

/\*frequency table including tenure segment and both activated and deactivated account\*/

**Proc** **Tabulate** data=ds2 ;

title'frequency table including tenure segment and both activated and deactivated account';

Class activation tenure\_seg;

Table (activation all),(tenure\_seg all)\*(n pctn)/row=float;

Keylabel ALL='margin sum'

n = 'Count'

pctn = 'Percent';

**Run**;

/\*frequency table including tenure and only deactivated account\*/

**Proc** **Tabulate** data=ds2 ;

title'frequency table including tenure and only deactivated account';

Class activation tenure\_seg;

Table (activation all),(tenure\_seg all)\*(n pctn)/row=float;

where activation= 'deactivated';

Keylabel ALL='margin sum'

n = 'Count'

pctn = 'Percent';

**Run**;

/\*frequency table including tenure and only active account\*/

**Proc** **Tabulate** data=ds2 ;

title'frequency table including tenure and only deactivated account';

Class activation tenure\_seg;

Table (activation all),(tenure\_seg all)\*(n pctn)/row=float;

where activation= 'activated';

Keylabel ALL='margin sum'

n = 'Count'

pctn = 'Percent';

**Run**;

/\*5)Test the general association between the tenure segments and “Good Credit”,“RatePlan ” and “DealerType.”\*/

**proc** **freq** data=ds2 ;

title 'chisq freq table between the tenure segments and “Good Credit”,“RatePlan ” and “DealerType';

table tenure\_seg \* goodcredit/ expected norow nocol chisq ;

table tenure\_seg \* rateplan/ expected norow nocol chisq ;

table tenure\_seg \* dealertype/ expected norow nocol chisq;

**run**;

/\*6) Test the general association between the activation status and “Good Credit”,“RatePlan ” and “DealerType.”\*/

**proc** **freq** data=ds2 ;

title 'chisq freq table between the activation status and “Good Credit”,“RatePlan ” and “DealerType';

table activation \* goodcredit/ expected norow nocol chisq relrisk;

table activation \* rateplan/ expected norow nocol chisq ;

table activation \* dealertype/ expected norow nocol chisq ;

**run**;

/\* 7)association between the account status and the tenure segments\*/

**proc** **freq** data=ds2;

title 'chisq freq table between the activation status and tenure segments';

table tenure\_seg \* activation/ expected cellchi2 norow nocol chisq;

**run**;

/\*8) Does Sales amount differ among different account status, GoodCredit, and customer age segments?\*/

/\*Does Sales amount differ among different account status\*/

**Proc** **Ttest** Data=ds2 COCHRAN side=**2** alpha=**0.05** h0=**0** test=diff; \* t-p is lower than the 0.05, reject H0;

title'method1:t test (paramatric): compare activation status';

class activation;

var sales;

**Run**;

**proc** **NPAR1WAY** data=ds2 wilcoxon median;

title'method2: Wilcoxon rank-sum test(same as the Mann-Whitney U test)(non-paramatric)';

class activation;

var sales;

**run**;

/\*Does Sales amount differ among different gredit\*/

**Proc** **Ttest** Data=ds2 COCHRAN side=**2** alpha=**0.05** h0=**0** test=diff; \* t-p is lower than the 0.05, reject H0;

title'method1:t test (paramatric): compare credit groups';

class GoodCredit;

var sales;

**Run**;

**proc** **NPAR1WAY** data=ds2 wilcoxon median;

title'method2: Wilcoxon rank-sum test(same as the Mann-Whitney U test)(non-paramatric)';

class GoodCredit;

var sales;

**run**;

/\*Does Sales amount differ among age segment (ANOVA)\*/

/\*assumption for anova:

Independence of cases ·this is an assumption of the model that simplifies the statistical analysis.

Normality ·the distributions of the residuals are normal (NOT MET, THUS NON\_PARAMETRIC ASSUMED, AND ANOVA NOT THE BEST OPTION).

Equality (or "homogeneity") of variances, called homoscedasticity.

\*/

**data** ds3;

set ds;

format age AgeFmt. ;

**run**;

**proc** **univariate** data=ds3 normal mu0=**0**;

title'normality check';

ods select TestsForNormality;

class age;

var sales;

**run**;

**Proc** **Anova** Data=ds3 ;

title'parametric model';

Class age;

Model sales = age; \* need to define y= numeric, and x=class in model ;

Means age/hovtest=levene SCHEFFE; \* south and west are different

\*Scheffe's multiple-comparison procedure on all main-effect means in the MEANS statement;

\* there is also tukey,LSD, GT2;

**Run**;

\*NON-PARAMETRIC: method 1: find the kruskal-wallis ranks of the whole data,

then do the regular ANOVA and Posthoc test on it (abandoned the original values);

**proc** **rank** data=ds3 out=Ranked\_ds3 ties=mean ;

var sales; \*\*there is also 'by' option in the 'rank' statement;

ranks salesRank;

**run**;

**proc** **glm** data=Ranked\_ds3 ;

title'NON-PARAMETRIC: method 1: anova based on rank, and test equal sales by least\_sq mean';

class age;

model salesRank=age;

means age / hovtest=levene SCHEFFE welch bon tukey alpha=**.05** plots=none;

lsmeans age / adjust=tukey pdiff alpha=**.05** plots=meanplot;

**run**;

**quit**;

\*NON-PARAMETRIC: Method 2: use Kruskal-Wallis Test (based on ranking of the data, and test the difference in sale by rank mean);

**proc** **npar1way** data=ds3 wilcoxon median dscf plots(only)=(wilcoxonboxplot medianplot);

\*\*dscf is the pairwise comparsion based on Kruskal-Wallis ranking (Dwass-Steel-Critchlow-Fligner);

title'NON-PARAMETRIC: Method 2: use Kruskal-Wallis Test (based on ranking of the data, and test the difference in sale by rank mean';

class age;

var sales;

**run**;