CUSTOMER CHURN ANALYSIS: TELCO

****

MUWANI ROBSON

May ,29 2020

TABLE OF CONTENTS

Introduction ................................................................................................................... 3

The Study ........................................................................................................ 3

The Variables ................................................................................................... 3

Objectives……………………………………………………………………………………………………………………. 3

Study Framework ............................................................................................................ 4

Hypotheses ..................................................................................................................... 4

Methodology ................................................................................................................... 5

Descriptive Analysis ........................................................................................................... 5

Univariate Analysis ............................................................................................ 5-6

Bivariate Analysis ............................................................................................... 6-9

Outlier Detection .............................................................................................. 9-10

Inferential Analysis .......................................................................................................... 11

Two-Sample T-test ............................................................................................. 11

Chi-square Analysis ............................................................................................ 12

Scatterplot Matrix ............................................................................................................. 13

Logistic Regression ........................................................................................................... 13

Summary of Results ........................................................................................................... 13

Recommendations ........................................................................................................... 13

Appendix .......................................................................................................................... 14

PROC Statements Used ..................................................................................................... 14

SAS Scripts ....................................................................................................................... 14-28

**Executive Summary**

Telecom companies typically spend most of their effort and resources on customer acquisition, even though the cost of retaining an existing customer is five times lower than acquiring a new one. Customer retention is a measure of how many of your customers continue to buy from you over time and are therefore loyal to your brand. Churn, sometimes known as customer attrition, is at the opposite end of the spectrum, i.e. the number of customers that stop buying from your company. The aim of this research project is to provide business analytics, visualize and discover the root causes of churn in Telecoms based on the Telco data set and implore strategies to minimise deactivations and increase customer lifetime value.

**Introduction**

Telecom industry retention surveys show that while price and product/ service are important, most customers leave a service provider because of dissatisfaction with the way they are treated. It costs hundreds of dollars to acquire a new telecom customer. When a customer churns, you not only lose the future revenue from this customer, but also the resources you spent on acquiring the customer in the first place. Researches by Bain company estimates that for a telecom provider with 5 million customers and an average churn of 2 to 2.5%, a reduction in churn by even 50 basis points would be worth $410 million in customer lifetime value over 30 months. When customers leave after poor experiences, they not only will not return, but they often amplify their message of dissatisfaction to others using social media. Key to this churn analysis is that high customer retention means long term customer value for the business, hence the need to manage customer attrition rate in saturated businesses like telecoms.

**Objectives**

* How to increase customer lifetime value (CLV)
* Promote actions that drive customer satisfaction, spend and loyalty
* Analyse customer distribution and interactions across service platforms
* Improve customer service delivery

**The Study**

The study is based on a subset of a large data set from Telco company. The dataset has 71,047 and 58 variables. Only 11 relevant variables were filtered for project analytics. Missing data classification on churn status being our target variable was excluded from the study.

**The Variables:**

The study has 57 selected predictor variables and one target variable. For the purpose of this project we selected the most important predictors and reduced them to 12. The composite variables for the data set are as follows:

CustomerID, Churn, MonthlyRevenue, MonthlyMinutes, TotalRecurringCharge, DirectorAssistedCalls, OverageMinutes, RoamingCalls, PercChangeMinutes, PercChangeRevenues, DroppedCalls, BlockedCalls, UnansweredCalls, CustomerCareCalls, ThreewayCalls, ReceivedCalls, OutboundCalls, InboundCalls, PeakCallsInOut ,OffPeakCallsInOut, DroppedBlockedCalls, CallForwardingCalls, CallWaitingCalls, MonthsInService ,UniqueSubs ActiveSubs ,ServiceArea, Handsets, HandsetModels ,CurrentEquipmentDays, AgeHH1, AgeHH2, ChildrenInHH ,HandsetRefurbished, HandsetWebCapable, TruckOwner, RVOwner ,Homeownership, BuysViaMailOrder, RespondsToMailOffers, OptOutMailings, NonUSTravel ,OwnsComputer, HasCreditCard, RetentionCalls, RetentionOffersAccepted, NewCellphoneUser, NotNewCellphoneUser, ReferralsMadeBySubscriber, IncomeGroup, OwnsMotorcycle ,AdjustmentsToCreditRating, HandsetPrice, MadeCallToRetentionTeam, CreditRating, PrizmCode ,Occupation & MaritalStatus

**The Study Framework**

The study has a total of 10 selected predictor variables, 5 of them grouped as modifiable churn drivers (numeric type), 5 as non modifiable churn drivers(character type).

See the flow chart below:

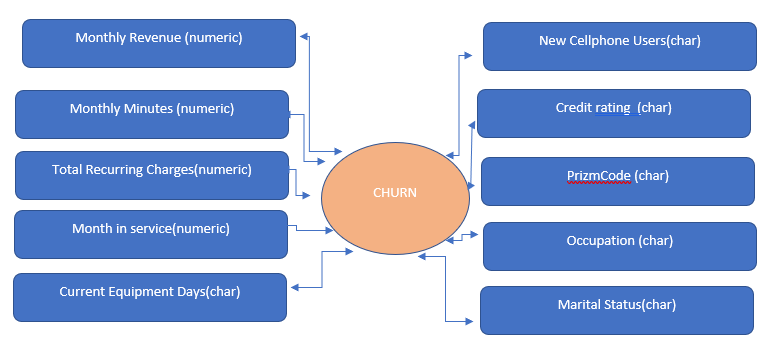


Figure 1, study framework

**Hypotheses**

Using the Telco dataset, the researcher aims to answer the following questions:

* What are the churn drivers in Telco?
* What are the telco churn behaviors?
* Does churn results in significant revenue loss?
* What are the demographic pattens in churn behavior?
* What is the disconnection trend/behavior?
* What is the relationship between credit rating in churn behavior?
* What is the relationship between equipment days in churn behavior?
* What is the relationship between occupation in churn behavior?
* What is the relationship between location in churn behavior?

**Methodology**

SAS 9.4 was used in reading, management, analysis and modelling of the data. A total of 15 SAS PROC statements were used in the analysis of the data (see Appendix).

SAS proc import statement was used to import the dataset for maximum control of the final output. Data types were established at import with nine categorical variables and four numerical variables. Custom formats were created for improved analysis and readability of the data. Missing data values on character variables was excluded in the analysis. Missing values on continuous variables was filled with mean.

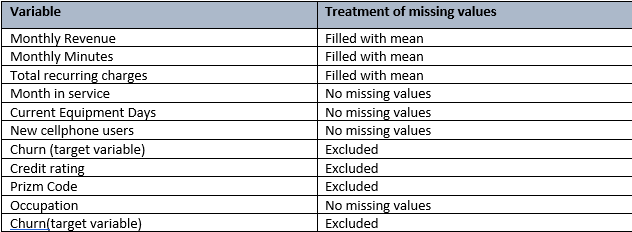


Table 1: Treatment of missing data values

**Descriptive Analysis**

The data reported a total of 36336 active accounts and 14711 deactivated accounts, representing 71.18% and 28.82% respectively. Notably, the deactivation rate is extremely high according to Bain telecommunications survey.

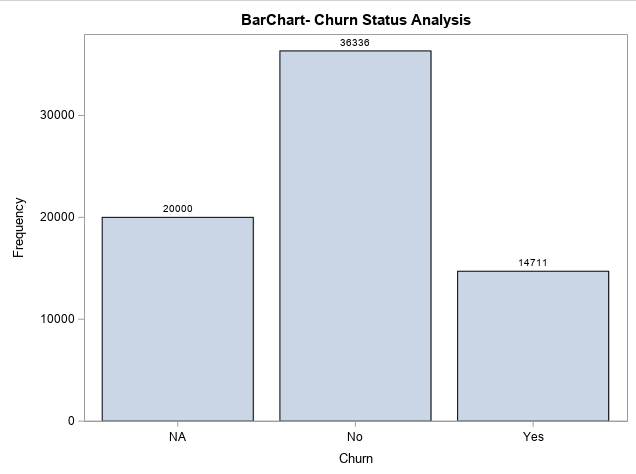
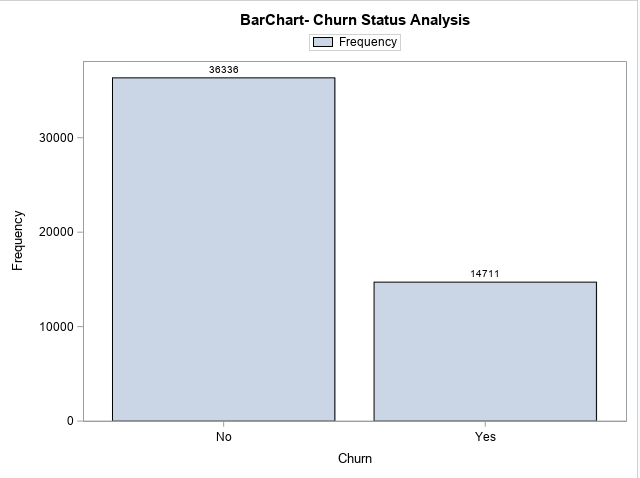


Figure 2 (including unclassified accounts) Figure 3 (excluding unclassified accounts)

**Univariate Analysis**

Univariate analysis can give a better understanding of the distribution of each variable in the framework. The main purpose of univariate analysis was to describe the data and find patterns that exist within it. Additionally, using univariate analysis I was able to observe each of the variables I terms of near normal distribution. This will allow us to use parametric statistics later that require normality as a prerequisite. No transformation was done on the variables as they depicted a near normal distribution, however, monthly revenue had extreme values thus skewed to the right and total recurring charges were more conical in shape.

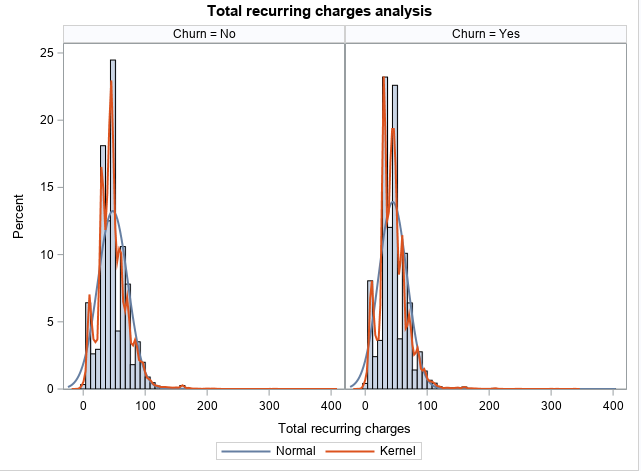
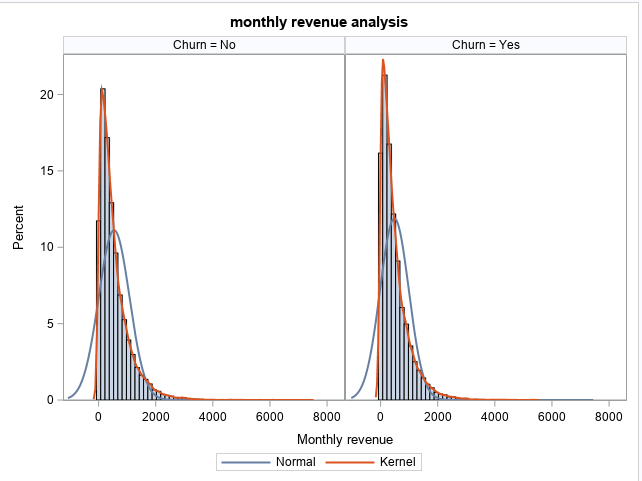


Fig 4: distribution of Monthly charges Fig 5: distribution of monthly recurring charges

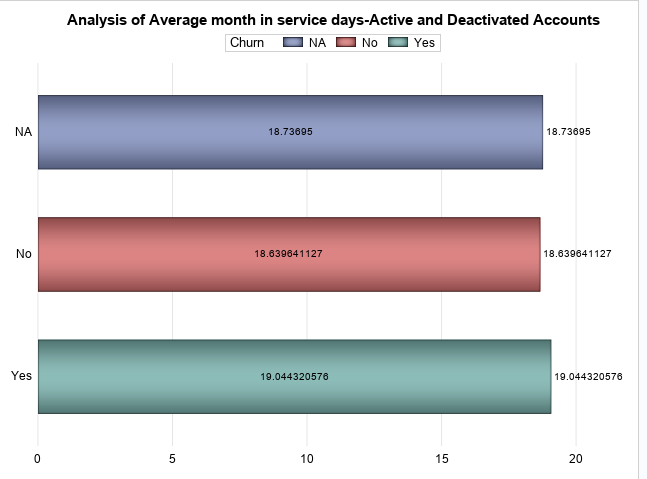
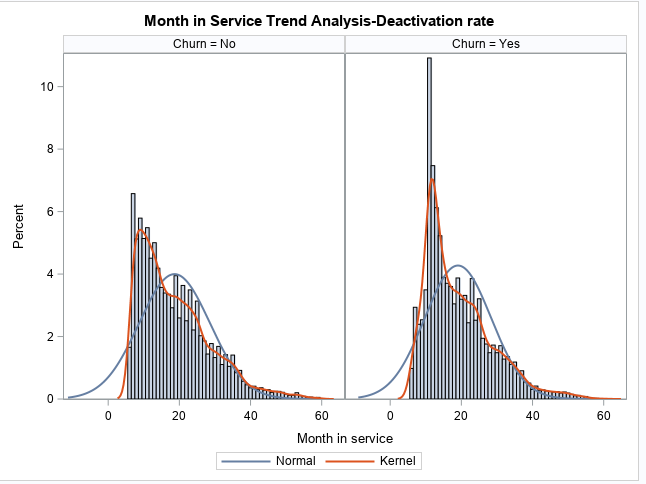
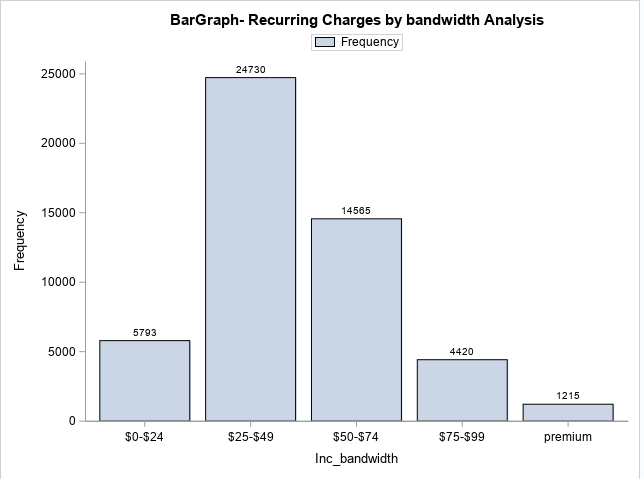


Fig 6: Months in Service Trend Fig 7: Average days of Active Accounts (active and deactivated)

Most customers deactivations happening around 19mnths and the maximum stay is 60months (5 years)

**Bivariate analysis**

**Bivariate analysis** is one of the statistical analysis where two variables are observed. One variable here is dependent while the other is independent. These variables are usually denoted by X and Y. So, here we analyse the changes occurred between the two variables and to what extent. It is important to compare two variables to each other in the analysis. Boxplots can show the distribution of a variable based on a group that it is a part of. In this case, a comparison of each distribution is seen when CAD is present and when it is not.

**Revenue Analysis**

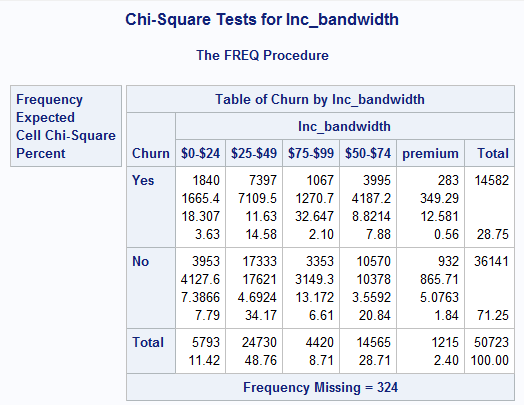
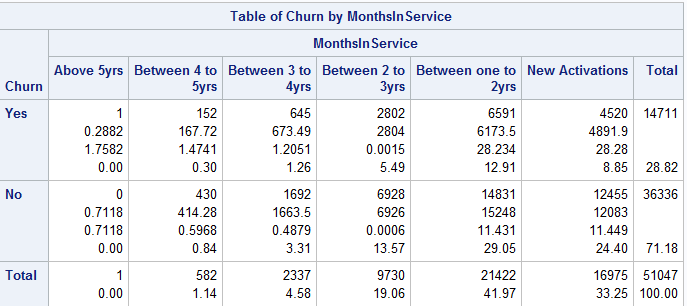


Fig 8: Recurring charges Analysis by bands Table 2: Chi-Square Tests for Recurring charged bands)

**Months in service Analysis**

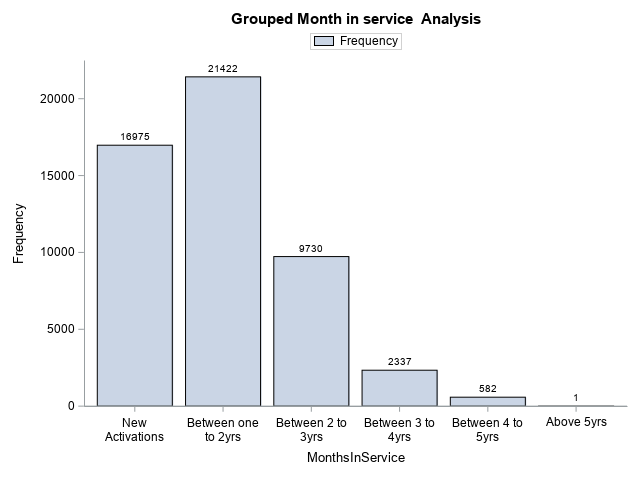


Fig 9 : Month in service Analysis by segments Table 3: Chi-Square Tests for Month in service Analysis

Pick activations and deactivations between 1 and 2 years, about 41% of the total. customers in the data base.

Most Accounts deactivation between 12 and 24months. (12.91%). New activations terminating service too, about 8.85%.

**Equipment Days Analysis**

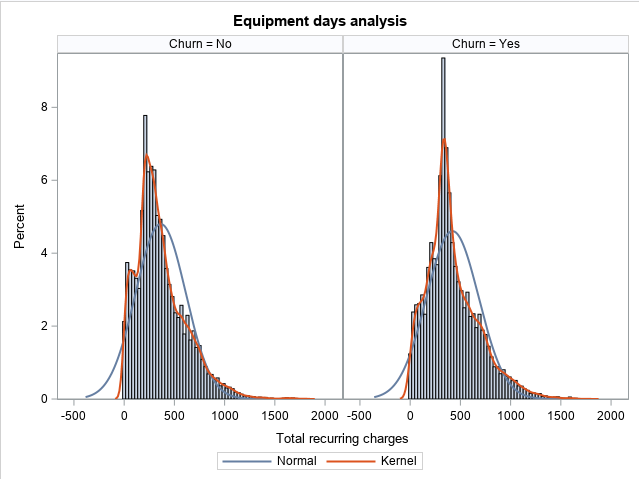
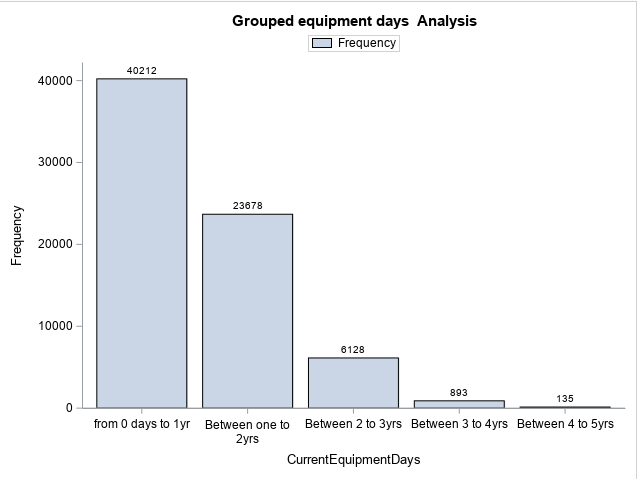
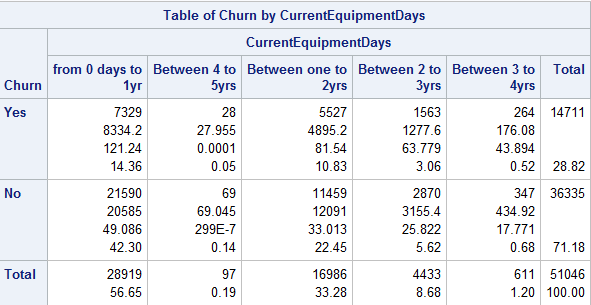
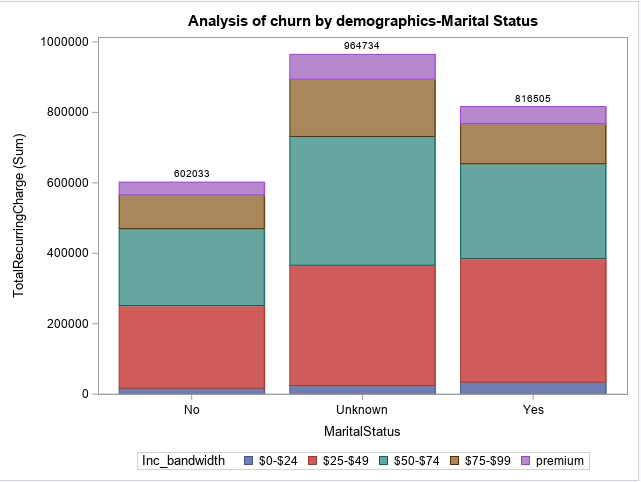
 Univariate analysis Bivarate Analysis Chi-Square Analysis

Fig 10 : equipment days Analysis Table 4: Chi-Square Tests for equipment days Analysis

* **Equipment Days Analysis:** 50% of the churners have equipment days less than one year, Most equipment days up to 2years.

**Demographic Analysis**: Marital Status and churn Behavior



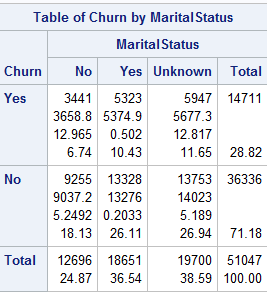
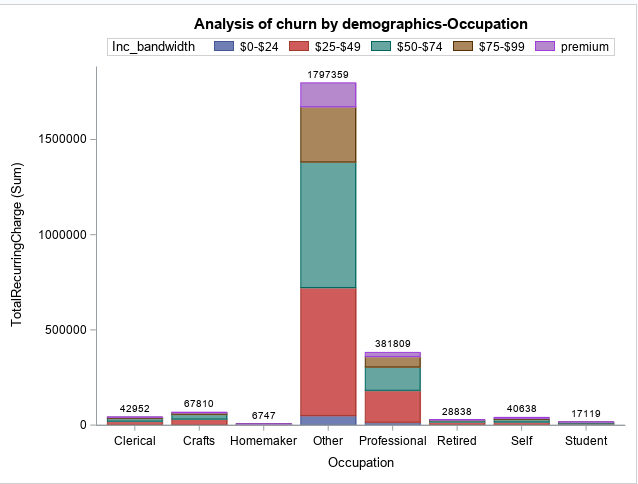


Fig 11 : Marital status in in Churn behavior Table 5: Chi-Square Tests for Marital status in Churn behavior

**Marital Status**: Churn rate is related to marital status, excluding the unknown category, married people are leaving more than those not married(10.43%) compared to 6.74% for those who did not churn. The rate is proportional to the total, as we have more married clients than not.

**Demographic Analysis**: Occupation



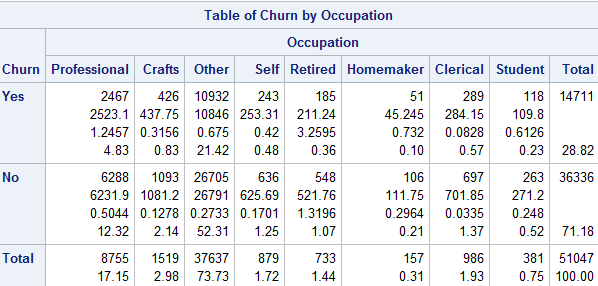


Fig 12 : Occupation in Churn behavior Table 6: Chi-Square Tests for Occupation in Churn behavior

**Occupation Analysis**: Most people leaving in the other category, Professionals leaving at a significantly high proportion. Numbers proportionally related to the totals in the subs data base, Shows company not able to handle numbers, a clear sign of dissatisfaction amongst customers.

MONTHLY CHARGES ANALYSIS OF CHURN AND CREDIT RATING

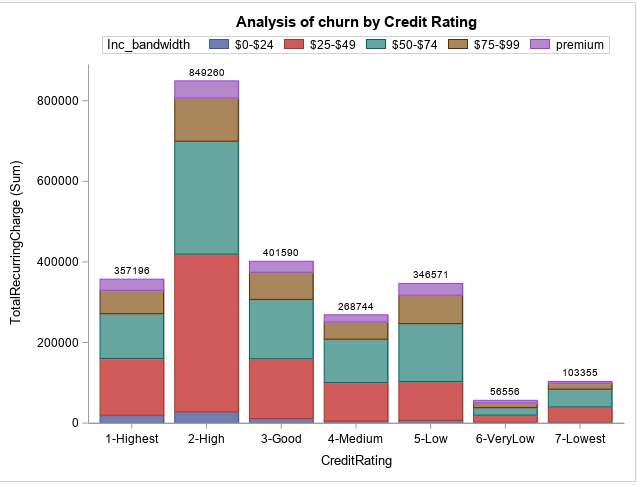
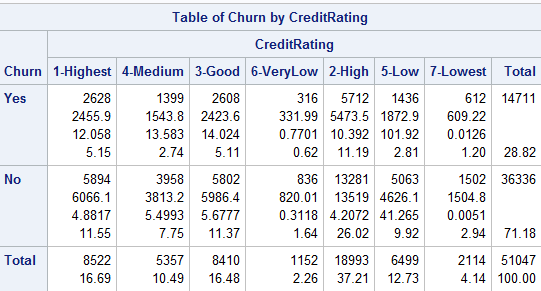


Fig 13 : Credit Rating in Churn behavior Table 7: Chi-Square : Credit Rating Churn behavior

**Credit Rating in Churn behavior**: company has fairly good client acquisition with best credit rating (1 and 2) with over 51%. More deactivation also happening from those with good credit over 16%. Company has significant portion of unhealthy customers (19.13%) with bad credit, deactivation does exits in these groups (5 to 7)

**CHURN &REVENUE ANALYSIS BY AREA CODE**

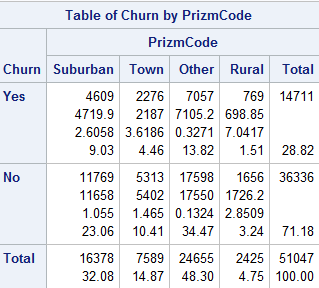
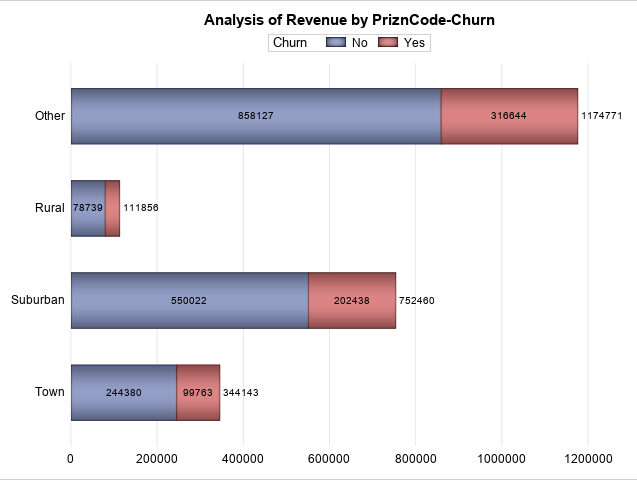


Fig 14 : Location and Churn behavior Table 8: Chi-Square : Location and Churn behavior

**Location and Churn behavior**: More churn in suburban areas excluding the others category. The revenue generated and loss is proportional to the clientele data base, more numbers more revenue and more terminations. Need to investigate the other category

**CELL PHONE USER ANALYSIS**

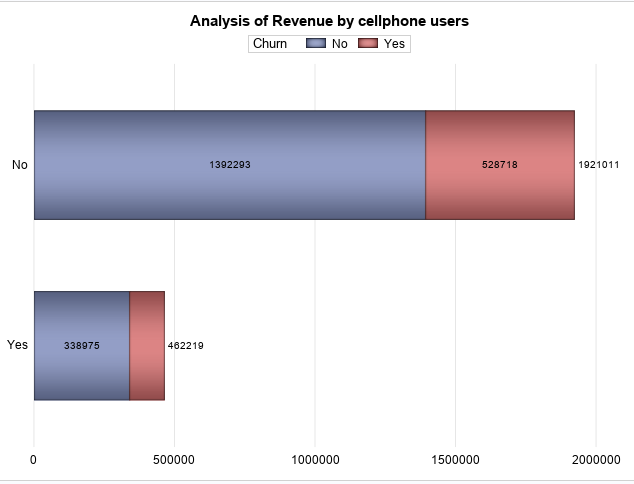
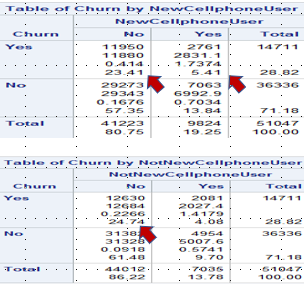
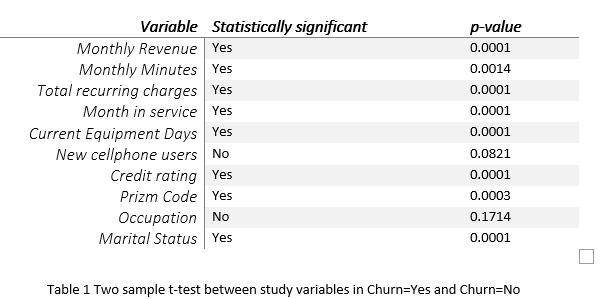
****

Fig 15 : Phone use/gadgets in Churn behavior Table 9: Chi-Square : Phone use/gadgets in Churn behavior

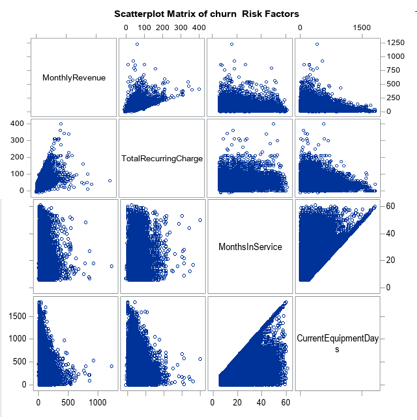
**Phone use/gadgets in Churn behavior**: Even new cellphone user leaving (5.41%). More churners on clients using own cell phones

SAMPLE T-TEST FOR STUDY VARIABLE ANALYSIS



**Inferential Analysis** : Visually, it appears that all variables are key determinants of churn behavior except occupation or whether someone got new cellphone or not.

SCATTERPLOT MATRIX



Plot of continuous data against other continuous variables to assess for multicollinearity between independent variables. High positive correlation between monthly revenue and month in service as well as equipment days was noted from the analysis.

**Outlier Detection**

**Outliers** are extreme values that deviate from other observations on data , they may indicate a variability in a measurement, experimental errors or a novelty. These can be problematic as they can cause a skewness in the data, pulling the mean away from the median in the direction of the outlier.

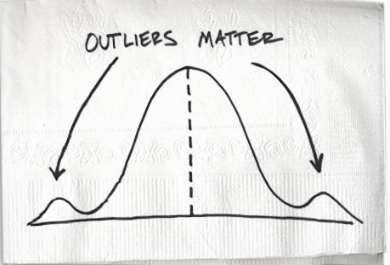


Fig 9: Outlier Analysis by bands

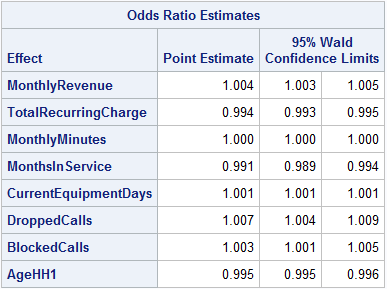
PROC univariate is useful in compiling the five number summary: minimum, Q1, median, Q2, max. Interquartile range can also be reported in this procedures. The outliers have been defined by the following two equations:

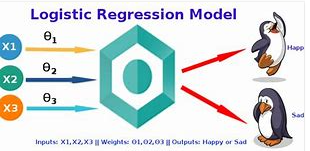
𝒐𝒖𝒕 𝒍𝒊𝒆𝒓> Q3+(**3**\*IQR)(the upper bound) 𝒐𝒖𝒕𝒍𝒊𝒆𝒓 < Q1-(**3**\*IQR)(the lower bound)

The outliers were not removed from the analysis, separate grouped were created upper bound, lower bound and data values within range formulated part of the bivariate analysis.

**Logistic Regression**

A useful solution will be able to predict the most likely causes of churn and flag any customers at risk. For example, how much of your churn is coming from simple recurrent monthly charges, occupation or location or how much of it is coming from that poorly strategized credit rating process. The summary model results from analysis are set as below:





* Backward Feature Elimination
  + Above are Features left (from framework)
  + Model concordance = 58.5%
* **Odds ratio** =

e ^ (0.434 + 1.004(MR) + 0.994(TRC) + 1.000(MM)+ 0.991(MIS) + 1.001(CED) +1.007(DC)+1.003(BC) +0.995(AG))

For every one unit change in **monthly revenue**, the log odds of churn (versus not churn) increases by 1.004 etc.

**Summary findings**

* High customer disconnections at 28.82% against industry standard of 2.5% per 5million customers
* Monthly revenue has obvious correlation with TotalRecurringCharge & monthly in service.
* Bulk of customers leaving around 20months in service, customer retention less than 2yrs, sign of customer frustration.
* Most contributing income bandwidth/segments are those paying between $25-$49(48.76%) and $50-$74(28.71%)
* More disconnections from these segments that are generating more revenue
* Disconnections highly correlated to equipment's days
* More churners from married customers (10.43%), professionals (4.83%)
* Company losing more of customers with good credit ~16%.
* In high revenue generating groups, High and Good Credit Rating are more easily to churn.
* Use of gadgets as a hold or strategy to create loyalty proved ineffective as activations with new gadgets are disconnecting from the service.
* More revenue contributions from suburban areas and towns compared to rural communities

**Recommendations**

* Company to focus on retention policy by conducting various market campaign strategies like running monthly or quarterly promotions, pricing strategies, expand product offering and or improve service delivery.
* To better service the market, the company must identify needs by segments and link product offer to specific segments, (say age groups, geographical locations, marital status or occupation)
* Optimise customer loyalty drivers, company must create loyalty groups and document loyalty indicators, say for stable married people, professionals etc.
* Company must create powerful customer focused value proposition:
* Relevance (solve problems), Quantifiable value (unique benefits) & differentiations.
* Company must devise ways for managing total customer experience by conducting service quality surveys and customer support periodically.
* Continue to focus more on customers with good credit as this will increase sales as revealed by the analysis and thrive to retain them.

APPENDIX

**PROC STATEMENTS USED**

1. PROC FORMAT enables you to define your own informats and formats for variables
2. PROC PRINT prints the observations in a SAS data set
3. PROC CONTENTS shows the contents of a SAS data set and prints the directory of the SAS library
4. PROC MEANS provides data summarization tools to compute descriptive statistics for variables across all observations and within groups of observations
5. PROC FREQ produces one-way to n-way frequency and contingency (crosstabulation) tables
6. PROC SORT orders SAS data set observations by the values of one or more character or numeric variables
7. PROC UNIVARIATE provides a variety of descriptive measures, graphical displays, and statistical methods, which you can use to summarize, visualize, analyze, and model the statistical distributions of numeric variables
8. PROC SGPLOT creates one or more plots and overlays them on a single set of axes
9. PROC SGPANEL creates a panel of graph cells for the values of one or more classification variables
10. PROC SGSCATTER creates a paneled graph of scatter plots for multiple combinations of variables
11. PROC TTEST performs t tests and computes confidence limits for one sample, paired observations, two independent samples, and the AB/BA crossover design
12. PROC LOGISTIC investigate the relationship between these discrete responses and a set of explanatories.

**SAS CODES**

/\*import the data as csv file\*/

LIBNAME MUW "C:\Users\Admin\Desktop\SAS PROJECT";

**PROC** **IMPORT** OUT= MUW.SAS\_PROJECT

DATAFILE= "C:\Users\Admin\Desktop\SAS PROJECT\Project Data F

iles\Telco Churn Data.csv"

DBMS=CSV REPLACE;

GETNAMES=YES;

DATAROW=**2**;

**RUN**;

\*data preparation -view data set..10obs;

**PROC** **PRINT** DATA = MUW.SAS\_PROJECT (OBS = **10**);

TITLE "CHURN DATA SET"; **RUN**;

/\*scan for duplicates and remove them\*/

TITLE "Count of Distinct Customer IDs in RSA1";

**PROC** **SQL**; SELECT COUNT(CustomerID)AS TOTAL\_COUNT,

COUNT(DISTINCT CustomerID) AS UNIQUE\_COUNT FROM MUW.SAS\_PROJECT ;

**QUIT**;

/\*scan for duplicates and remove them\*/

**Proc** **sort** data=MUW.SAS\_PROJECT out=MUW.Telcom\_data nodupkey; \*/sort data and scan for duplicates/\*;

by CustomerID;

**run**;

/\* No duplicates found\*/

/\*summary data\*/

**PROC** **CONTENTS** DATA=MUW.Telcom\_data;

**RUN**;

/\*data type\*/

**PROC** **CONTENTS** DATA=MUW.Telcom\_data varnum short;

**RUN**;

/\*scan for missing data under var CHURN\*/

ods table onewayfreqs=temp;

**proc** **freq** data=MUW.Telcom\_data;

table \_all\_ / missing;

format \_numeric\_ nmissfmt. \_character\_ $missfmt.;

**run**;

/\*DESCRIPTIVE ANALYSIS-CHURN\*/

**PROC** **FORMAT**;

VALUE $CHAR " "="MISSING"

OTHER= "NOT MISSING"

RUN;

/\* 28.15 missing, label as missing\*/

/\*label as missing\*/

**PROC** **FREQ** DATA=MUW.Telcom\_data;

TABLE Churn/MISSING;

FORMAT \_CHARACTER\_$CHAR.;

**RUN**;

/\*delete missing data on key column-Churn\*/

**data** MUW.Telcom;

set MUW.Telcom\_data;

if Churn eq"NA" then delete;

**run**;

/\*descriptive Analysis-Barchart for Churn status; \*/

**proc** **sgplot** data=MUW.Telcom;

Vbar Churn / datalabel colormodel=twocolorramp;

keylegend / location=outside position=top fillheight=**10** fillaspect=**2** ;

title"BarChart- Churn Status Analysis (ex\_missing";

**run**;

/\*analysis of churn with missing data\*/

**proc** **sgplot** data=MUW.Telcom\_data;

vbar Churn / datalabel colormodel=twocolorramp;

title"BarChart- Churn Status Analysis(inc\_missing";

**run**;

\*UNIVARIATE ANALYSIS\*;

/\*scan for missing monthinservice Analysis\*/

**proc** **means** data=MUW.Telcom N NMISS MIN MEAN STD MAX;

var MonthsInService;

**run**;

\*NO MISSING UNDER MONTHINSERVICE;

ods pdf file ="C:\Users\Admin\Desktop\SAS PROJECT\monthinservice\_graph\_1" style= MoonFlower notoc ;

TITLE "Month in Service Tren Analysis-Deactivation rate";

**proc** **sgpanel** data = MUW.Telcom;

panelby Churn / columns = **2**;

histogram MonthsInService;

density MonthsInService; density MonthsInService/ type = kernel;

colaxis label = 'Month in service';

**run**;

ods pdf close;

/\*scan for missing MONTHLY REVENUE\*/

TITLE "Missing month in revenue values";

**proc** **means** data=MUW.Telcom maxdec=**2** N NMISS MIN MEAN STD MAX;

var MonthlyRevenue;

**run**;

/\*imputation of missing: mean=58.83\*/

**proc** **sql**;

create table Telco\_revenue as

select \*, coalesce (MonthlyRevenue,**58.83**) as MonthlyRevenue\_updated

from MUW.Telcom;

**run**;

/\*confirm missing MONTHLY REVENUE values\*/

TITLE "missing month in revenue data";

**proc** **means** data=Telco\_revenue maxdec=**2** N NMISS MIN MEAN STD MAX;

var MonthlyRevenue\_updated;

**run**;

ods pdf file ="C:\Users\Admin\Desktop\SAS PROJECT\monthinservice\_graph\_1" style= MoonFlower notoc; ;

TITLE "Monthly Revenue Distribution Analysis ";

**proc** **sgpanel** data = Telco\_revenue;

panelby Churn / columns = **2**;

histogram MonthlyRevenue\_updated;

density MonthlyRevenue\_updated / type = kernel;

colaxis label = 'Monthly revenue';

**run**;

ods pdf close;

Title 'Analysis of monthly revenue-Active and Deactivated Accounts';

**proc** **sgplot** data=Telco\_revenue noborder;

hbar Churn / response=MonthlyRevenue\_updated stat=sum group=Churn displaybaseline=auto barwidth=**0.6**

seglabel datalabel dataskin=pressed;

yaxis display=(noline noticks nolabel);

xaxis display=(noline noticks nolabel) grid;

keylegend / location=outside position=top fillheight=**10** fillaspect=**2** ;

/\*format MonthsInService 8.0;\*/

**run**;

/\*scan for missing MONTHLY REVENUE\*/

TITLE "missing month in minutes data";

**proc** **means** data=MUW.Telcom maxdec=**2** N NMISS MIN MEAN STD MAX;

var MonthlyMinutes;

**run**;

/\*imputation of missing: mean=525.65\*/

**proc** **sql**;

create table Telco\_minutes as

select \*, coalesce (MonthlyMinutes,**525.65**) as MonthlyMinutes\_updated

from MUW.Telcom;

**run**;

/\*confirm missing MONTHLY minutes values\*/

TITLE "missing month in revenue data";

**proc** **means** data=Telco\_minutes maxdec=**2** N NMISS MIN MEAN STD MAX;

var MonthlyMinutes\_updated;

**run**;

ods pdf file ="C:\Users\Admin\Desktop\SAS PROJECT\monthinservice\_graph\_1" style= MoonFlower notoc; ;

TITLE "monthly revenue analysis";

**proc** **sgpanel** data = Telco\_minutes;

panelby Churn / columns = **2**;

histogram MonthlyMinutes\_updated;

density MonthlyMinutes\_updated; density MonthlyMinutes\_updated/ type = kernel;

colaxis label = 'Monthly revenue';

**run**;

ods pdf close;

/\*scan for missing MONTHLY RECURRING CHARGES\*/

TITLE "missing TotalRecurringCharge values";

**proc** **means** data=MUW.Telcom maxdec=**2** N NMISS MIN MEAN STD MAX;

var TotalRecurringCharge;

**run**;

/\*imputation of missing: mean=46.83\*/

**proc** **sql**;

create table Telco\_mcharges as

select \*, coalesce (TotalRecurringCharge,**46.83**) as TotalRecurringCharge\_updated

from MUW.Telcom;

**run**;

/\*confirm missing MONTHLY CHARGES values\*/

TITLE "missing total recurring charges values";

**proc** **means** data=Telco\_mcharges maxdec=**2** N NMISS MIN MEAN STD MAX;

var TotalRecurringCharge\_updated;

**run**;

TITLE "Total recurring charges analysis-consolidated";

**proc** **sgplot** data=MUW.Telcom;

histogram TotalRecurringCharge;

density TotalRecurringCharge;

title"Monthly total revenue distribution";

**run**;

ods pdf file ="C:\Users\Admin\Desktop\SAS PROJECT\monthinservice\_graph\_1" style= MoonFlower notoc; ;

TITLE "Total recurring charges analysis";

**proc** **sgpanel** data = Telco\_mcharges;

panelby Churn / columns = **2**;

histogram TotalRecurringCharge\_updated ;

density TotalRecurringCharge\_updated;density TotalRecurringCharge\_updated/ type = kernel;

colaxis label = 'Total recurring charges';

**run**;

ods pdf close;

/\* "Total recurring charges analysis\*/

TITLE "Total recurring charges analysis";

**proc** **sgplot** data = Telco\_mcharges ;

histogram TotalRecurringCharge\_updated / showbins;

density TotalRecurringCharge\_updated;

density TotalRecurringCharge\_updated / type = kernel;

yaxis grid; xaxis label = 'total charges';

keylegend / location = inside position = topright;

title 'Distribution of total charges';

**run**;

/\*scan for missing dropped calls\*/

TITLE "missing TotalRecurringCharge values";

**proc** **means** data=MUW.Telcom maxdec=**2** N NMISS MIN MEAN STD MAX;

var DroppedCalls

;

**run**;

/\* no missing for dropped calls\*/

/\*distribution of monthly revenue\*/

**proc** **sgplot** data=MUW.Telcom;

histogram TotalRecurringCharge;

density TotalRecurringCharge;

title"Monthly total revenue distribution";

**run**;

/\*scan for missing under equipment days\*/

TITLE "missing equipment days values";

**proc** **means** data=MUW.Telcom maxdec=**2** N NMISS MIN MEAN STD MAX;

var CurrentEquipmentDays;

**run**;

/\*delete missing data values\*/

**data** MUW.Equip\_days;

set MUW.Telcom\_data;

if CurrentEquipmentDays eq **.** then delete;

**run**;

/\*confirm deletion\*/

**proc** **means** data=MUW.Equip\_days maxdec=**2** N NMISS MIN MEAN STD MAX;

var CurrentEquipmentDays;

**run**;

/\*Equipment days analysis\*/

ods pdf file ="C:\Users\Admin\Desktop\SAS PROJECT\monthinservice\_graph\_1" style= MoonFlower notoc; ;

TITLE "Equipment days analysis";

**proc** **sgpanel** data = Telco\_mcharges;

panelby Churn / columns = **2**;

histogram CurrentEquipmentDays ;

density CurrentEquipmentDays;density CurrentEquipmentDays/ type = kernel;

colaxis label = 'Total recurring charges';

**run**;

ods pdf close;

/\*Analysis month in equipments days \*/

**proc** **format**;

value equip\_days low-**365** ="from 0 days to 1yr"

**365**-**730** ="Between one to 2yrs"

**730**-**1095** ="Between 2 to 3yrs"

**1095**-**1460** ="Between 3 to 4yrs"

**1460**-**1825** ="Between 4 to 5yrs"

**1825**-high ="Above 5yrs";

**run**;

/\*Bar graph Analysis for Equipment days\*/

**proc** **sgplot** data=MUW.Equip\_days noborder;

vbar CurrentEquipmentDays /datalabel;

keylegend / location=outside position=top fillheight=**10** fillaspect=**2** ;

format CurrentEquipmentDays equip\_days.;

title "Grouped equipment days Analysis";

**run**;

/\*Chi square Analysis for equipment days\*/

**proc** **freq** data=MUW.Telcom order=data;

tables Churn\*CurrentEquipmentDays / expected cellchi2 norow nocol chisq;

output out=ChiSqData n nmiss pchi lrchi;

format CurrentEquipmentDays equip\_days.;

title 'Chi-Square Tests for equipment days';

**run**;

/\*Analysis of equipment days\*/

Title 'Analysis of equipment days';

**proc** **sgplot** data=MUW.Telcom noborder;

hbar CurrentEquipmentDays / response=CurrentEquipmentDays stat=freq group=CurrentEquipmentDays displaybaseline=auto barwidth=**0.6**

seglabel datalabel dataskin=pressed;

yaxis display=(noline noticks nolabel);

xaxis display=(noline noticks nolabel) grid;

/\*keylegend / location=outside position=top fillheight=10 fillaspect=2 ;\*/

format CurrentEquipmentDays equip\_days.;

**run**;

/\*BIVARIATE & CHI-SQUARE ANALYSIS\*/

/\*INCOME BANDWITH ANALYSIS\*/

**data** MUW.Telcom\_Income; /\*formating dates-use this method\*/

length Inc\_bandwidth $**15**;

set Telco\_mcharges;

if **0.1**< TotalRecurringCharge <**25** then Inc\_bandwidth ='$0-$24';

else if **24**<TotalRecurringCharge<**50** then Inc\_bandwidth ='$25-$49';

else if **49**< TotalRecurringCharge<**75** then Inc\_bandwidth ='$50-$74';

else if **74**<TotalRecurringCharge<**100** then Inc\_bandwidth ='$75-$99';

else if TotalRecurringCharge >**99** then Inc\_bandwidth='premium';

**run**;

**proc** **sgplot** data=MUW.Telcom\_Income noborder;

vbar Inc\_bandwidth /datalabel;

keylegend / location=outside position=top fillheight=**10** fillaspect=**2** ;

title "BarGraph- Recurring Charges by bandwidth Analysis";

**run**;

**proc** **freq** data=MUW.Telcom\_Income order=data;

tables Churn\*Inc\_bandwidth / expected cellchi2 norow nocol chisq;

output out=ChiSqData n nmiss pchi lrchi;

/\*weight Count;\*/

title 'Chi-Square Tests for Inc\_bandwidth';

**run**;

/\*Analysis month in service groups\*/

**proc** **format**;

value months\_grp low-**12** ="New Activations"

**12**-**24** ="Between one to 2yrs"

**24**-**36** ="Between 2 to 3yrs"

**36**-**48** ="Between 3 to 4yrs"

**48**-**60** ="Between 4 to 5yrs"

**60**-high ="Above 5yrs";

**run**;

**proc** **sgplot** data=MUW.Telcom noborder;

vbar MonthsInService /datalabel;

keylegend / location=outside position=top fillheight=**10** fillaspect=**2** ;

format MonthsInService months\_grp.;

title "Grouped Month in service Analysis";

**run**;

**proc** **freq** data=MUW.Telcom\_Income order=data;

tables Churn\*MonthsInService / expected cellchi2 norow nocol chisq;

output out=ChiSqData n nmiss pchi lrchi;

format MonthsInService months\_grp.;

title 'Chi-Square Tests for month in service grps';

**run**;

Title 'Analysis of Average month in service days-Active and Deactivated Accounts';

**proc** **sgplot** data=MUW.Telcom\_data noborder;

hbar Churn / response=MonthsInService stat=mean group=Churn displaybaseline=auto barwidth=**0.6**

seglabel datalabel dataskin=pressed;

yaxis display=(noline noticks nolabel);

xaxis display=(noline noticks nolabel) grid;

keylegend / location=outside position=top fillheight=**10** fillaspect=**2** ;

/\*format MonthsInService 8.0;\*/

**run**;

ods graphics on;

**proc** **anova** data = MUW.Telcom;

class Churn;

model MonthlyRevenue= Churn;

means Churn/scheffe;

title "customer monthly revenue analysis";

**run**;

ods graphics off;

Title 'Analysis of dropped calls-Active and Deactivated Accounts';

**proc** **sgplot** data=MUW.Telcom noborder;

hbar Churn / response=DroppedCalls stat=sum group=Churn displaybaseline=auto barwidth=**0.6**

seglabel datalabel dataskin=pressed;

yaxis display=(noline noticks nolabel);

xaxis display=(noline noticks nolabel) grid;

keylegend / location=outside position=top fillheight=**10** fillaspect=**2** ;

/\*format MonthsInService 8.0;\*/

**run**;

ods graphics on;

**proc** **anova** data = MUW.Telcom;

class Churn;

model DroppedCalls= Churn;

means Churn/scheffe;

title "customer monthly revenue analysis";

**run**;

ods graphics off;

**proc** **sgplot** data=MUW.Telcom;

hbar MaritalStatus / colormodel=twocolorramp;

keylegend / location=outside position=top fillheight=**10** fillaspect=**2** ;

**run**;

Title 'Analysis of churn by demographics-Marital Status';

**proc** **sgplot** data=MUW.Telcom\_Income;

vbar MaritalStatus / datalabel response=TotalRecurringCharge Group=Inc\_bandwidth;

**run**;

**proc** **freq** data=MUW.Telcom order=data;

tables Churn\*MaritalStatus / expected cellchi2 norow nocol chisq;

output out=ChiSqData n nmiss pchi lrchi;

/\*weight Count;\*/

title 'Chi-Square Tests for Marital status';

**run**;

/\*analysis of occupation vs churn\*/

Title 'Analysis of churn by demographics-Occupation';

**proc** **sgplot** data=MUW.Telcom\_Income noborder;

vbar Occupation /datalabel response=TotalRecurringCharge Group=Inc\_bandwidth;;

keylegend / location=outside position=top fillheight=**10** fillaspect=**2** ;

/\*format MonthsInService 8.0;\*/

**run**;

**proc** **freq** data=MUW.Telcom order=data;

tables Churn\*Occupation / expected cellchi2 norow nocol chisq;

output out=ChiSqData n nmiss pchi lrchi;

/\*weight Count;\*/

title 'Chi-Square Tests for Occupation';

**run**;

/\*analysis of CreditRating vs churn\*/

Title 'Analysis of churn by Credit Rating';

**proc** **sgplot** data=MUW.Telcom\_Income noborder;

vbar CreditRating /datalabel response=TotalRecurringCharge Group=Inc\_bandwidth;;;

keylegend / location=outside position=top fillheight=**10** fillaspect=**2** ;

/\*format MonthsInService 8.0;\*/

**run**;

**proc** **freq** data=MUW.Telcom order=data;

tables Churn\*CreditRating / expected cellchi2 norow nocol chisq;

output out=ChiSqData n nmiss pchi lrchi;

/\*weight Count;\*/

title 'Chi-Square Tests for Credit Rating';

**run**;

/\*primcode area Analysis\*/

title "Analysis of churn by PriznCode";

**proc** **freq** data=MUW.Telcom order=data;

tables Churn\*PrizmCode

/ expected cellchi2 norow nocol chisq;

output out=ChiSqData n nmiss pchi lrchi;

/\*weight Count;\*/

**run**;

title "Analysis of Revenue by PriznCode-Churn";

**proc** **sgplot** data=MUW.Telcom noborder;

hbar PrizmCode / response=TotalRecurringCharge stat=sum group=Churn displaybaseline=auto barwidth=**0.6**

seglabel datalabel dataskin=pressed;

yaxis display=(noline noticks nolabel);

xaxis display=(noline noticks nolabel) grid;

keylegend / location=outside position=top fillheight=**10** fillaspect=**2** ;

format MonthsInService **8.0**;

**run**;

/\*New/Not cellphone users Analysis\*/

title "Analysis of Not New cellphone users";

**proc** **freq** data=MUW.Telcom order=data;

tables Churn\*NotNewCellphoneUser / expected cellchi2 norow nocol chisq;

output out=ChiSqData n nmiss pchi lrchi;

/\*weight Count;\*/

**run**;

title "Analysis of New cellphone users";

**proc** **freq** data=MUW.Telcom order=data;

tables Churn\*NewCellphoneUser / expected cellchi2 norow nocol chisq;

output out=ChiSqData n nmiss pchi lrchi;

/\*weight Count;\*/

**run**;

title "Analysis of Revenue by cellphone users";

**proc** **sgplot** data=MUW.Telcom noborder;

hbar NewCellphoneUser / response=TotalRecurringCharge stat=sum group=Churn displaybaseline=auto barwidth=**0.6**

seglabel datalabel dataskin=pressed;

yaxis display=(noline noticks nolabel);

xaxis display=(noline noticks nolabel) grid;

keylegend / location=outside position=top fillheight=**10** fillaspect=**2** ;

format MonthsInService **8.0**;

**run**;

/\*Revenue Analysis by Credit Rating, grouped by Churn\*/

title "Analysis of Revenue by Credit Rating-Churn";

**proc** **sgplot** data=MUW.Telcom noborder;

hbar CreditRating / response=TotalRecurringCharge stat=sum group=Churn displaybaseline=auto barwidth=**0.6**

seglabel datalabel dataskin=pressed;

yaxis display=(noline noticks nolabel);

xaxis display=(noline noticks nolabel) grid;

keylegend / location=outside position=top fillheight=**10** fillaspect=**2** ;

format MonthsInService **8.0**;

**run**;

/\*Revenue Analysis by IncomeGroup, grouped by Churn status\*/

title "Analysis of Revenue by IncomeGroup-Churn";

**proc** **sgplot** data=MUW.Telcom noborder;

hbar IncomeGroup / response=TotalRecurringCharge stat=sum group=Churn displaybaseline=auto barwidth=**0.6**

seglabel datalabel dataskin=pressed;

yaxis display=(noline noticks nolabel);

xaxis display=(noline noticks nolabel) grid;

keylegend / location=outside position=top fillheight=**10** fillaspect=**2** ;

format MonthsInService **8.0**;

**run**;

/\*Revenue Analysis by IncomeGroup, grouped by Churn status\*/

title "Analysis of Revenue by marital status-Churn";

**proc** **sgplot** data=MUW.Telcom noborder;

hbar MaritalStatus / response=TotalRecurringCharge stat=sum group=Churn displaybaseline=auto barwidth=**0.6**

seglabel datalabel dataskin=pressed;

yaxis display=(noline noticks nolabel);

xaxis display=(noline noticks nolabel) grid;

keylegend / location=outside position=top fillheight=**10** fillaspect=**2** ;

format MonthsInService **8.0**;

**run**;

title "Analysis of blocked calls by Churn";

**proc** **sgplot** data=MUW.Telcom noborder;

hbar Churn / response= BlockedCalls stat=sum group=Churn displaybaseline=auto barwidth=**0.6**

seglabel datalabel dataskin=pressed;

yaxis display=(noline noticks nolabel);

xaxis display=(noline noticks nolabel) grid;

keylegend / location=outside position=top fillheight=**10** fillaspect=**2** ;

format MonthsInService **8.0**;

**run**;

ods graphics on;

**proc** **anova** data = MUW.Telcom;

class Churn;

model BlockedCalls= Churn;

means Churn/scheffe;

title "customer BlockedCalls analysis";

**run**;

ods graphics off;

/\*OUTLIER DETECTION\*/

\*monthly revenue;

\*total recurring charges

\*monthly minutes;

\*dropped calls;

\*blocked calls;

/\*calculate quartiles and inter quartiles for Monthly revenue\*/

**proc** **sgplot** data= Telco\_revenue;

vbox MonthlyRevenue\_updated;

**run**;

**proc** **means** data=Telco\_revenue maxdec=**2**;

var MonthlyRevenue\_updated;

output out =revenue p25=Q1 p75=Q3 qrange =IQR;

**run**;

**data** revenue\_01;

set revenue;

lower\_limit =Q1-(**3**\*IQR);

upper\_limit=Q3+(**3**\*IQR);

drop \_TYPE\_ \_FREQ\_;

**run**;

**proc** **print** data=revenue\_01; **run**;

/\*create catesian product\*/

**proc** **sql**;

create table revenue\_02 as

select A.\*,B.\*

from Telco\_revenue as A, revenue\_01 as B

;

**quit**;

**data** revenue\_03;

set revenue\_02;

if MonthlyRevenue\_updated le lower\_limit then range ="below lower limit";

else if MonthlyRevenue\_updated ge upper\_limit then range ="above upper limit";

else range ="within range"

;

**run**;

/\*bivariate analysis for monthly revenue range\*/

**proc** **freq** data=revenue\_03 order=data;

tables Churn\*range / expected cellchi2 norow nocol chisq;

output out=ChiSqData n nmiss pchi lrchi;

/\*weight Count;\*/

title 'Chi-Square Tests for monthly revenue range';

**run**;

/\*calculate quartiles and inter quartiles for monthly recurring charges \*/

**proc** **means** data=Telco\_mcharges maxdec=**2**;

var TotalRecurringCharge\_updated;

output out =charges p25=Q1 p75=Q3 qrange =IQR;

**run**;

**data** charges\_01;

set charges;

lower\_limit =Q1-(**3**\*IQR);

upper\_limit=Q3+(**3**\*IQR);

drop \_TYPE\_ \_FREQ\_;

**run**;

**proc** **print** data=charges\_01; **run**;

/\*create catesian product\*/

**proc** **sql**;

create table charges\_02 as

select A.\*,B.\*

from Telco\_mcharges as A, charges\_01 as B

;

**quit**;

**data** charges\_03;

set charges\_02;

if TotalRecurringCharge\_updated le lower\_limit then charge\_range ="below lower limit";

else if TotalRecurringCharge\_updated ge upper\_limit then charge\_range ="above upper limit";

else charge\_range ="within range"

;

**run**;

**proc** **freq** data=charges\_03 order=data;

tables Churn\*charge\_range / expected cellchi2 norow nocol chisq;

output out=ChiSqData n nmiss pchi lrchi;

/\*weight Count;\*/

title 'Chi-Square Tests for monthly charge range';

**run**;

/\*calculate quartiles and inter quartiles for monthly recurring charges \*/

**proc** **means** data=Telco\_mcharges maxdec=**2**;

var TotalRecurringCharge\_updated;

output out =charges p25=Q1 p75=Q3 qrange =IQR;

**run**;

**data** charges\_01;

set charges;

lower\_limit =Q1-(**3**\*IQR);

upper\_limit=Q3+(**3**\*IQR);

drop \_TYPE\_ \_FREQ\_;

**run**;

**proc** **print** data=charges\_01; **run**;

/\*create catesian product\*/

**proc** **sql**;

create table charges\_02 as

select A.\*,B.\*

from Telco\_mcharges as A, charges\_01 as B

;

**quit**;

**data** charges\_03;

set charges\_02;

if TotalRecurringCharge\_updated le lower\_limit then charge\_range ="below lower limit";

else if TotalRecurringCharge\_updated ge upper\_limit then charge\_range ="above upper limit";

else charge\_range ="within range"

;

**run**;

**proc** **freq** data=charges\_03 order=data;

tables Churn\*charge\_range / expected cellchi2 norow nocol chisq;

output out=ChiSqData n nmiss pchi lrchi;

/\*weight Count;\*/

title 'Chi-Square Tests for monthly charge range';

**run**;

/\*calculate range for monthly minutes \*/

**proc** **univariate** data=Telco\_minutes ;

var MonthlyMinutes\_updated;

**run**;

**proc** **means** data=Telco\_minutes maxdec=**2**;

var MonthlyMinutes\_updated;

output out =call\_minutes p25=Q1 p75=Q3 qrange =IQR;

**run**;

**data** minutes\_01;

set call\_minutes;

lower\_limit =Q1-(**3**\*IQR);

upper\_limit=Q3+(**3**\*IQR);

drop \_TYPE\_ \_FREQ\_;

**run**;

**proc** **print** data=minutes\_01; **run**;

/\*create catesian product\*/

**proc** **sql**;

create table minutes\_02 as

select A.\*,B.\*

from Telco\_minutes as A, minutes\_01 as B

;

**quit**;

**data** minutes\_03;

set minutes\_02;

if MonthlyMinutes\_updated le lower\_limit then call\_range ="below lower limit";

else if MonthlyMinutes\_updated ge upper\_limit then call\_range ="above upper limit";

else call\_range ="within range"

;

**run**;

**proc** **freq** data=minutes\_03 order=data;

tables Churn\*call\_range / expected cellchi2 norow nocol chisq;

output out=ChiSqData n nmiss pchi lrchi;

/\*weight Count;\*/

title 'Chi-Square Tests for call range in minutes';

**run**;

\*Scatterplot Matrix;

ods pdf file ="C:\Users\Admin\Desktop\SAS PROJECT\matrix\_graph\_1" style= MoonFlower notoc;

**proc** **sgscatter** data = MUW.Telcom;

matrix MonthlyRevenue TotalRecurringCharge MonthsInService CurrentEquipmentDays;

/\*label MonthsInService = 'transformed';\*/

title 'Scatterplot Matrix of churn Risk Factors';

**run**;

ods pdf close

/\*Logisic Regression model for prediction, odds ratio\*/

ods pdf file ="C:\Users\Admin\Desktop\SAS PROJECT\matrix\_graph\_1" style= MoonFlower notoc;

**proc** **logistic** data = MUW.Telcom desc plots(only) = oddsratio plots(only) = roc;

class MaritalStatus CreditRating PrizmCode Occupation;

model Churn = MonthlyRevenue TotalRecurringCharge MonthlyMinutes MonthsInService

CurrentEquipmentDays MaritalStatus CreditRating PrizmCode Occupation

/ expb selection = backward;

output out = outdata p = pred\_prob lower = low upper = up;

title 'Logistic Regression for Churn';

**run**;

ods pdf close;

/\*\*\*\*\*\*\*\*\*\*\*\*\*end of code\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/