Student Statistics Report

By

Olivia Rancour
Rishi Reddy
Nikhil Preeth Birra

February 17, 2015

Professor Prince

The University of Akron

EXECUTIVE SUMMARY

The issue is how to boost the customer response rate to the renewal offers that it mails to its magazine subscribers. GCC's relational database contains over a terabyte of data, which encompasses 75 million customers. The data in this database is used for new customer acquisition, customer reactivation, and identifying cross-selling opportunists for products. The industry response rate is about 2%, however, GCC has had a historically higher response rate. The goal is to ensure that GCC maintains its place as one of the top achievers in the targeted market. GCC is considering the development of a targeted marketing strategy for a hobby-based magazine. The Data worksheet contains the data used to train, validate, and test a classification method. The data-mining goal is to explore and clean up the data in the Data worksheet.

I. DATA CLEANING

This dataset represents the data used to train, validate, and test a classification method. The data represents a class of 45,000 customers and provides a variety of variables. Each row of the data table represents all of the characteristics of one of the customers, and each column represents one of the characteristics, also known as variables or fields.

We checked the data type that JMP has assigned to each variable. All variables without any quantitative value can be characterized as nominal. Variables with numeric concepts in clear ordering can be characterized as ordinal. Continuous characterization represents variables that only contain numeric values. Listed below are the changes we made:

- Renewal → Nominal
- Resident Length → Ordinal
- Household Size → Ordinal

a.) MISSING DATA

Some data had no recorded value. Missing data may be due to equipment malfunction, inconsistencies, misunderstandings, and various other reasons. Records containing missing data can be deleted; however, this is dangerous because patterns of missing values can be systematic. Often times, it is more beneficial if missing data is inferred or imputed. Imputation of missing data considers what the likely value is given the record's other attributive values. Imputation requires tools like multiple regression and classification.

First, to find missing data, open our data table → select tables → Missing Data Pattern. We then selected all of the columns in our *Data* worksheet to find patterns of missing data and clicked add columns. We selected the "Count Missing Value Codes" check box to count the missing value codes as missing values and then clicked "Ok". A new JMP table was created that indicated how many rows had missing values for any column or a combination of columns. This helped us to identify the missing data and determine how to best handle the missing values. In the *Data* worksheet, 2,923 rows had missing variables and 9 columns had missing variables.

Count	Number of columns missing	Patterns	Missing Values in Columns
42077	0	"00000000000000000000000000000000000000"	
58	1	"00000000000000000000000000000000000001"	MonthsSinceExpire
152	1	"00000000000000000000000000000000000100"	MonthsSince1stOrder
336	1	"00000000000000000000001000000000000000"	MonthsSinceLastPayment
8	2	"00000000000000000000001000000000000100"	MonthsSinceLastPayment, MonthsSince1stOrder
12	1	"00000000000000000000010000000000000000"	TotalPadOrders
1562	1	"00000000000000000000100000000000000000"	DollarsPerIssue
1	2	"00000000000000000000100000000000000100"	DollarsPerIssue
49	2	"00000000000000000000101000000000000000"	DollarsPerIssue
11	2	"00000000000000000000110000000000000000"	DollarsPerIssue, TotalPaidOrders
10	3	"00000000000000000000111000000000000000"	DollarsPerIssue, TotalPaidOrders, MonthsSinceLastPayment
4	1	"00000000000000100000000000000000000000"	HomeValue
6	1	"00000000100000000000000000000000000000"	HouseholdSize
677	1	"00010000000000000000000000000000000000"	HomeOwner
3	2	"00010000000000000000000000000000000100"	HomeOwner, MonthsSince1stOrder
10	2	"00010000000000000000001000000000000000"	HomeOwner, MonthsSinceLastPayment
17	2	"00010000000000000000010000000000000000"	HomeOwner, DollarsPerIssue
1	3	"00010000000000000000110000000000000000"	HomeOwner, DollarsPerIssue, TotalPaidOrders
3	2	"00010000000000100000000000000000000000"	HomeOwner, HomeValue
3	2	"00010000100000000000000000000000000000"	HomeOwner, HouseholdSize

Figure 1: The above figure gives the count of all the missing columns in the give data.

The above table was created by analyzing the JMP table given for the missing values; we have added the fourth column giving the name of the columns missing in the dataset. The first columns give us the count of the missing values based on the pattern. The second column gives the number of columns missing based on the patterns. The third column is the code described by JMP to easily describe the data.

For example let us consider the first row of the pattern column. It has all zeros which mean in the data set there are 42077 rows with no data missing in it. The second row describe that there are 58 rows with data missing in the "MonthSinceExpire" column. The missing values are described by mentioning "1" in the place of the column. Similarly, the above table also has rows with more than one value missing in it.Let us consider row number five it has eight rows missing in both the columns MonthSinceLastPayment and MonthSince1stOrder.

As there are many missing values in the data set and considering them for the future analysis will not give us the right output we ignore them for now. This is done by selecting all the rows in new missing data pattern table created, then right click and select Exclude/Unexclude. This will exclude all the rows from the original table.

b.) RECODING OF VARIABLES

If we have data which is coded incorrectly (for example, contains incorrect wording or errors), we can recode so that it follows a consistent format. It can also be used to replace missing or empty values. We select the column which we want to recode and select "Cols" and select "recode" option.

We want to recode renewal variable as it is changed to nominal and having '0' or '1' doesn't give much information. So we recode '0' to 'N' and '1' to 'Y'. As the values 0, 1 are numeric and we are changing it to character. First we right click on the Renewal column and select "Column info". Then we change the datatype from Numeric to Character. Next we select the "Renewal" column and select "Cols" and click on "recode" and change the value of 0 to N and 1 to Y. The below figure displays the recode values for renewal variable.

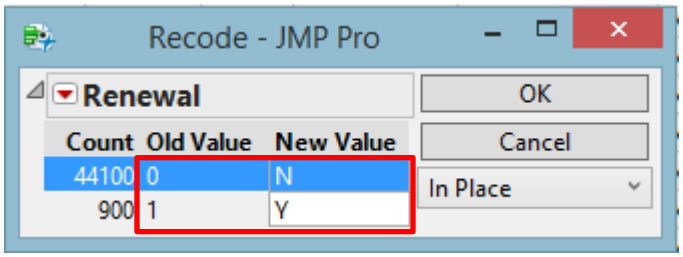


Figure 2: The above figure shows the old values and new values being assigned for Renewal variable.

Analyzing the value code assigned to MagazineStatus the original value 'N' which is gift subscription does not match to what it means. So we want to recode it to 'G'. so we click on MagazineStatus variable and then select cols → recode. Then change the value on N to G as shown below.

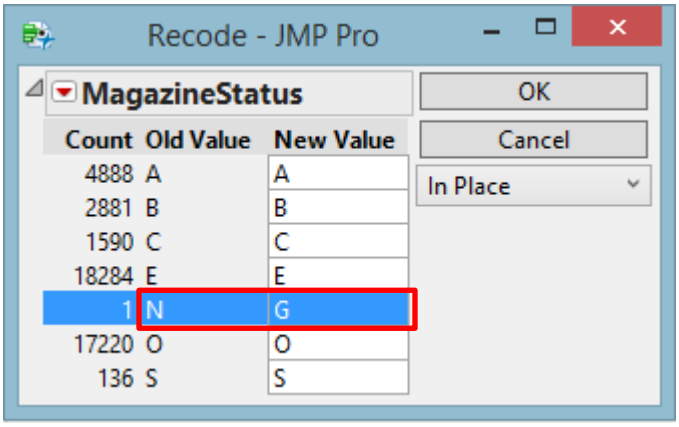


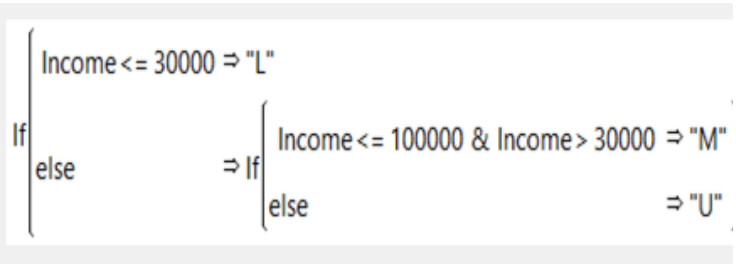
Figure 3: The above figure shows the old value and new value for MagazineStatus variable.

c.) TRANSFORMING VARIABLES

In order to make easier comparisons, some of the dataset needed to be transformed. Transformations allowed us to create new columns for new information. To transform data we created a new column from the table main menu. Each new column is then labeled it appropriately and given the appropriate modeling type. After clicking the column properties button, we selected "Formula" for editing. Through this function we were able to create a cleaner worksheet that better illustrates comparisons, conditions, and statistics. Below is an example of a transformation we performed. Below are the transformations we have performed.

i.) **Navigation:** Column → New Column → "Economic Status" → Data Type: Character →

Modeling Type: Nominal → Formula→



This is creates a new column "EconomicStatus" with the following values.

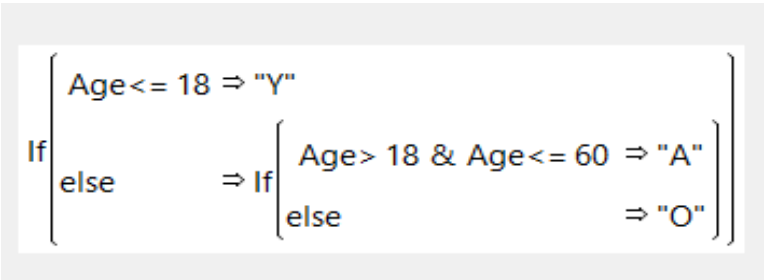
U → "Upper class"
M → "Middle class"
L → "Lower class"

We have created a formula where in if income is less than 30000 we will categorize it has Lower class, income range between 30000 and 100000 we categorize it has Middle class and the rest will fall into Upper class.

ii.) We are categorizing age into three different types. Customers with age less than 18 will fall into Young, age between 18 and 60 will fall into Adult and the rest are categorized as old.

Navigation: Column → New Column → "Age Classification" → Data Type: Character →

Modeling Type: Nominal → Formula →



This is creates a new column "AgeClassification" with the following values.

Y → "Young"
A → "Adult"
O → "Old"

The following shows how the final data set after transforming the columns.

	RequestedCancellations	NoPayCancellations	PaidComplaints	GiftDonor	NumberGiftDonations	MonthsSince1stOrder	MonthsSinceLastOrder	MonthsSinceExpire	EconomicStatus	AgeClassification
1	0	3	0	N	0	118	89	52	L	O
2	0	0	0	Y	2	90	90	52	L	O
3	0	0	0	N	0	91	54	17	M	O
4	0	0	0	N	0	158	60	20	L	A
5	0	1	0	N	0	50	37	34	L	O
6	0	0	0	N	0	59	59	46	L	A
7	0	0	0	N	0	60	60	47	U	A
8	0	0	0	N	0	55	55	42	L	A
9	0	0	0	N	0	50	50	25	L	A
10	0	0	0	N	0	47	47	34	L	A
11	0	0	0	N	0	31	31	13	M	A
12	0	0	0	N	0	25	25	12	M	A
13	0	0	0	N	0	16	16	6	L	A
14	0	0	0	N	0	58	58	33	L	A
15	0	0	0	N	0	360	121	79	L	O
16	0	3	0	N	0	68	68	57	L	O
17	0	0	0	N	0	127	119	104	M	A
18	0	0	0	N	0	35	35	13	M	A
19	0	0	0	N	0	50	29	3	L	O
20	0	0	0	N	0	71	71	11	L	A

Figure 4: The above figure shows the transformed data.

d.) ELIMINATION OF VARIABLES

Listwise deletion excludes cases with missing data on any variable are excluded from the sample for analysis of these variables. The analysis is only run on cases which have a complete set of data. It is important to be careful with listwise deletion because it can lead to bias. Pairwise deletion attempts to create less loss than listwise deletion. In pairwise deletion, only cases relating to each pair of variables with missing data involved in an analysis are deleted. We imputed the missing data with pairwise deletion.

i.) We can eliminate GiftDonor as we have already have another variable NumberGiftDonations. All the values with "N" in GiftDonor will have a value 0. So if we want to consider all the Giftdonors we can use NumberGiftDonations variable with a condition greater than 0.

Navigation: Right click on the variable "GiftDonor" → select Exclude/Unexclude.

ii.) We can eliminate Age as we have already transformed it to AgeClassification by classifying them into young, adult, and old.

Navigation: Right click on the variable "Age" → select Exclude/Unexclude.

II. GRAPHS AND TABLES

a.) CORRELATIONS

i.) Correlation between YearsSinceLastOrder and MonthsSinceLastPayment:

- **Naviagtion:** Analyze → Multivariate Methods → Multivariate → YearsSinceLastOrder, MonthsSinceLastPayment drag them into Y, columns → OK.

The below diagram gives the correlations between YearsSinceLastOrder and MonthsSinceLastPayment. The correlation value r=0.8020 which is a **moderate positive correlation**.

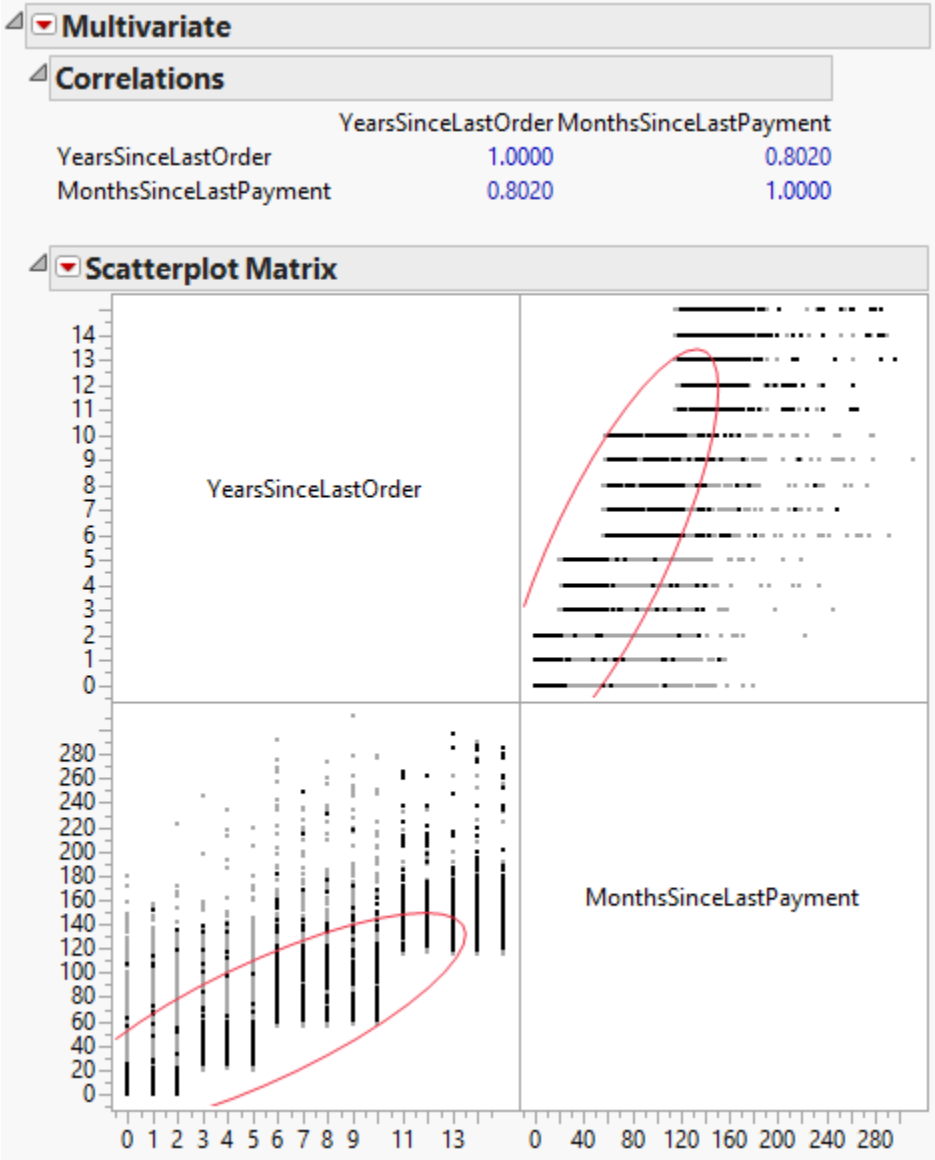


Figure 5: The above image gives the Scatterplot matric between YearsSinceLastOrder and MonthsSinceLastPayment

ii.) Correlation between NumberGiftDonations and Income:

- **Navigation:** Analyze → Multivariate Methods → Multivariate → Select NumberGiftDonations and Income → drag them into Y, columns → OK.

The below diagram gives the correlations between NumberGiftDonations and Income. The correlation value r=0.0338 which slightly tends to **no correlation**.

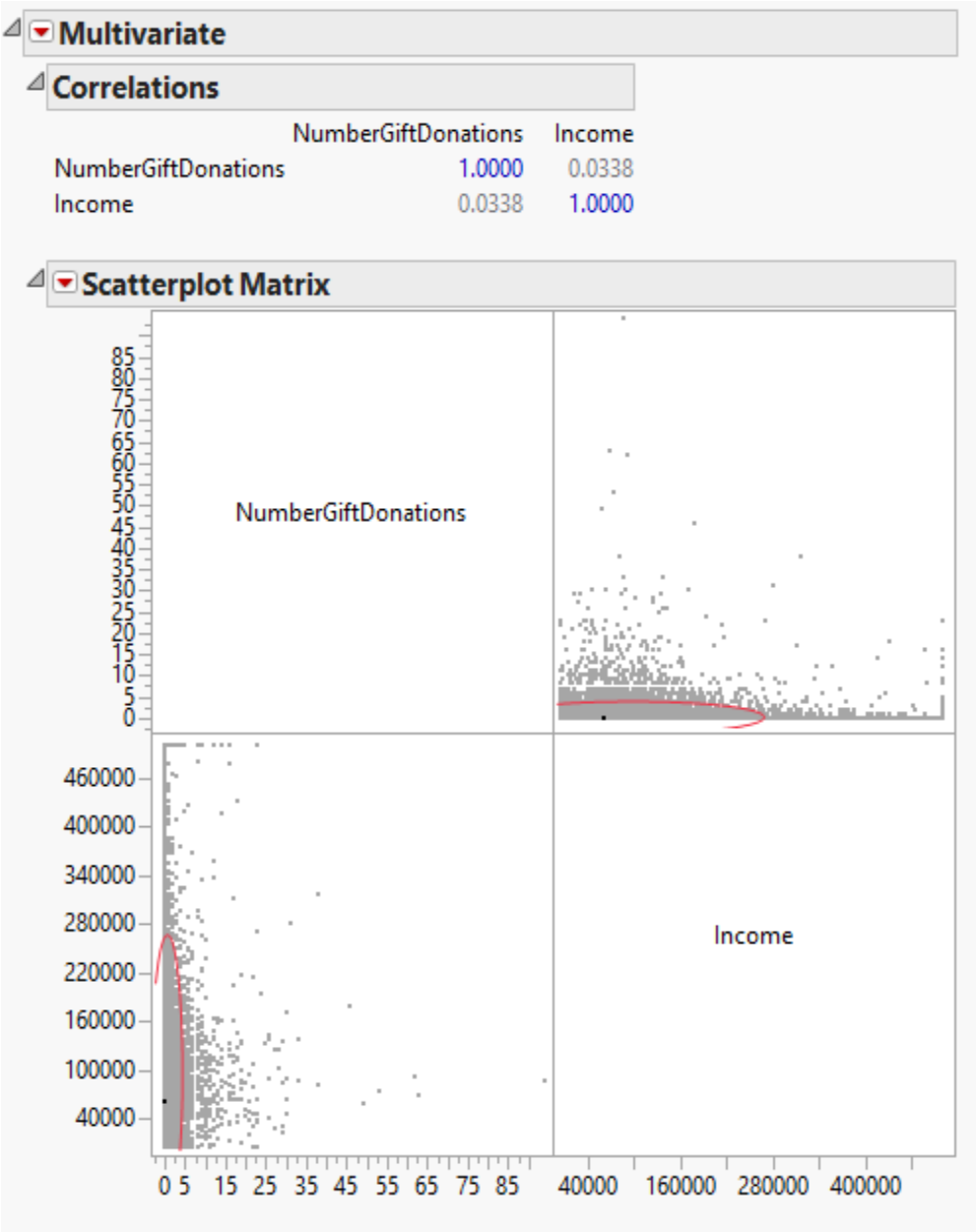
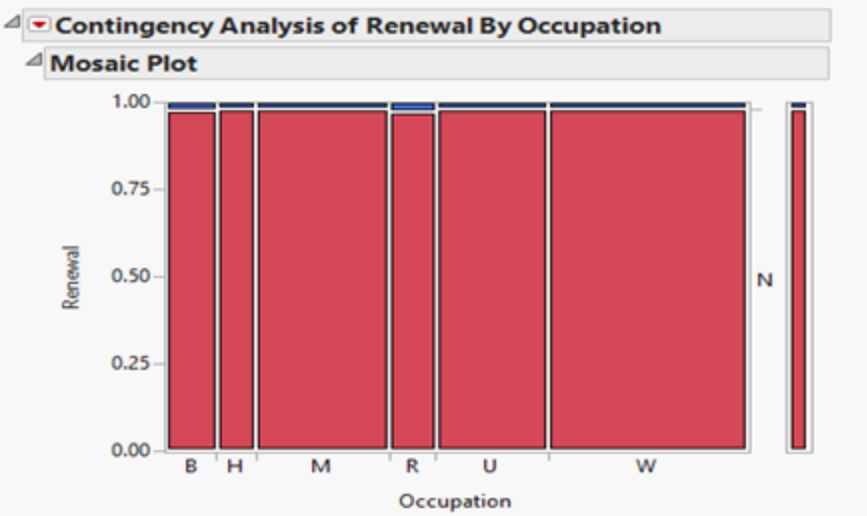


Figure 6 : The above image gives the Scatterplot matric between NumberGiftDonations and Income.

iii.) Contingency Analysis of Renewal by occupation.

- Process: Analyze → Fit Y by X → "Drag renewal to Y" → "Drag Occupation to X" → ok.



Contingency Table

Renewal			
Occupation	N	Y	
Count			
Total %			
Col %			
Row %			
B	3791	89	3880
	8.72	0.20	8.92
	8.90	10.19	
	97.71	2.29	
H	2847	49	2896
	6.55	0.11	6.66
	6.68	5.61	
	98.31	1.69	
M	9834	201	10035
	22.62	0.46	23.08
	23.08	23.02	
	98.00	2.00	
R	3381	93	3474
	7.78	0.21	7.99
	7.93	10.65	
	97.32	2.68	
U	8264	151	8415
	19.00	0.35	19.35
	19.39	17.30	
	98.21	1.79	
W	14494	290	14784
	33.33	0.67	34.00
	34.01	33.22	
	98.04	1.96	
	42611	873	43484
	97.99	2.01	

Tests

N	DF	-LogLike	RSquare (U)
43484	5	6.2211458	0.0015

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	12.442	0.0292*
Pearson	13.100	0.0225*

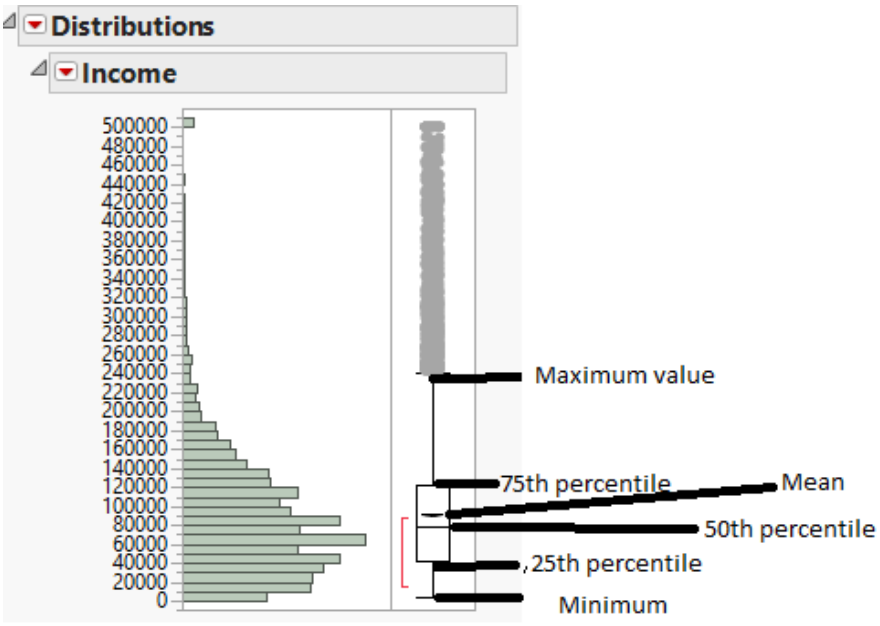
Figure 7: This illustrates the contingency analysis of renewal by occupation.

- People working in professional, executive, sales, marketing, and services, clerical contribute to more than 50% of the renewals being done.
- Professional/executive are the majority of people who renewed the magazines.

III. DESCRIPTIVE STATISTICS

i.) Descriptive statistics for Income.

- Analyze → Distribution → select "Income" → click Ok.



Quantiles

100.0%	maximum	500000
99.5%		500000
97.5%		268000
90.0%		170000
75.0%	quartile	122000
50.0%	median	78000
25.0%	quartile	42500
10.0%		19000
2.5%		5000
0.5%		5000
0.0%	minimum	5000

Summary Statistics

Mean	91222.782
Std Dev	72142.472
Std Err Mean	351.69698
Upper 95% Mean	91912.115
Lower 95% Mean	90533.449
N	42077

Figure 8: The above image illustrates the box plot for income variable.

The above box plot of the income displays the center position of the data and range of the data.It gives the maximum, minimum, 25th, 50th, 75th percentile of the income data. From the quantiles in the figure we can see that the maximum salary is 500000, the 75th percentile is 122000, median is 78000, 25th percentile is 42500, minimum is 5000. We can also find the mean from the summary statistics which is 91222.728$. In above the figure all the dots above the maximum value are the outliers. The outlines in the income are in large numbers, so not considering them in the analysis can be dangerous.The red brackets in the box plot defines the shortest path of the data.

ii.) Descriptive statistics for MonthsSinceLastOrder.

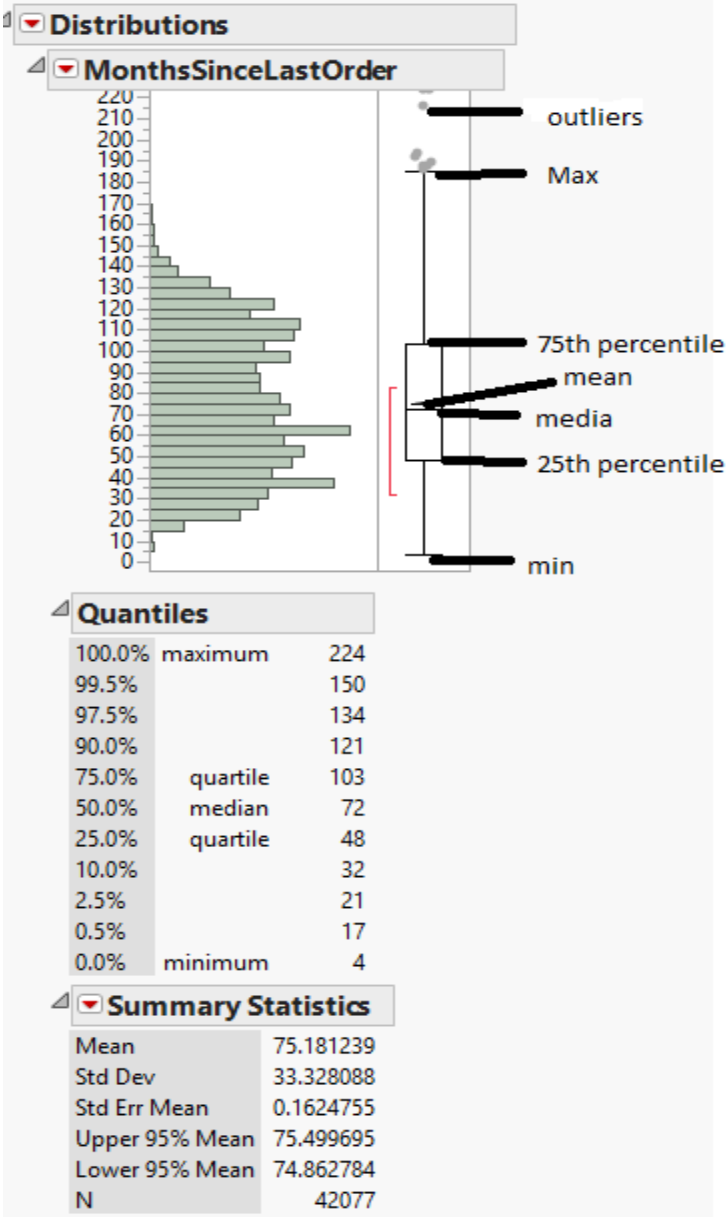


Figure 9: The above image illustrates the box plot for MonthsSinceLastOrder variable.

- Analyze → Distribution → select “MonthsSinceLastOrder” → click Ok.

The above box plot of the MonthSinceLastOrder displays the center position of the data and range of the data. It gives the maximum, minimum, 25th, 50th, 75th percentile of the monthsincelastorder data. From the quantiles in the figure we can see that the maximum value is 224, the 75th percentile is 103, median is 72,25th percentile is 48, minimum is 4. We can also find the mean from the summary statistics which is 75.18 .In above the figure all the dots above the maximum value are the outliers. The outliers in the monthsincelastorder are less in numbers and hence can be ignored. The red bracket in the box plot defines the shortest path of the data.