University of Southampton
Department of Marketing
MANG6329 Data Analytic

Data Analytic R code Report

Lulu Xu
Student Number
Major
lx2n17@soton.ac.uk

Table of Contents

List of Tables and Figures

List of Tables

List of Figures

Introduction

The purpose of this project is to research the outcome of a marketing campaign that performed by a sales company. The target of the market campaign is to boost the sales volume on the company's website. The database is provided for analysis, including eight profile variables and three outcome variables of 48009 contacted customers. In consideration of the amount of data, an R program is implemented for both primary data analysis of each variable and investigating the relationships between profile variables and outcome variables.

The whole R program consists of four separate R script files for from Question 1 to Question 4 respective. The brief introduction for each file follows:

1. Display basic statistical properties and data distribution for all 8 profile variables.

2. Analyze and plot the relationship between each predictive variable and the target variable "visit".

3. For all customers with "visit" == 1, analyze and plot and relationship between each predictive variable and the target variable "conversion".

4. For all customers with "conversion" == 1, analyze and plot and relationship between each predictive variable and the target variable "spend". Plot linear regressions according to "recency" against "spend", "history segment" against "spend" and "history" against "spend".

This report illustrates the design process, code explanation and result for the R program. A conclusion and a recommendation sections are included at the end of this report as well.

Design Process

As mentioned above, there are four separate R files with different functionalities in this project. This section firstly describes shared libraries and mechanisms applied in this project and then demonstrates the design idea in detail for each file.

Shared Libraries and Mechanisms

There are three libraries included in this R program, ggplot2, ggthemes and plyr. The simple introductions for these three libraries are listed below:

- ggplot2: a plotting system for R, based on the grammar of graphics, which tries to take the good parts of base and lattice graphics and none of the bad parts. [3]
- ggthemes: some extra themes, geoms, and scales for ggplot2. [4]
- plyr: a set of tools that solves a common set of problems: you need to break a big problem down into manageable pieces, operate on each piece and then put all the pieces back together. [6]

The library ggplot2 offers flexible and easy-using functionality. There are several advantages of using ggplot2 in this project.

- Easy to extract necessary data from a matrix and only plot for necessary data.
- The layer overlay method allows users to add color, scale, label and other information to the diagram conveniently.
- High flexibility for adjusting plotting.

The library plyr is another magical package for splitting large data set into small subsets and analyzing against those subsets. ddply[5] is the only function applied in this project.

Since there are many duplicated variables are declared in each file, a clear memory command is executed at the beginning of each file in order to keep the clarity and uniqueness of variables.

The given csv data file cannot be read by normal method that provided in our labs. After doing some research, it was found that the file is encoded by UTF-16 LE format and all columns are separated by "\t". It is required to execute command

```
read.csv2("direct_marketing.csv",fileEncoding="UTF-16LE", sep="\t")
```

to read the file specifically.

Design for Question 1

There are eight profile variables needed to be analyzed in Question 1.

The first variable "recency" indicates months since last purchase which is a categorical variable, and therefore the statical measurements such as mean, median, variance do not make sense for "recency" analysis. Hence, a histogram of the volume of customers for each "recency" is sufficient for "recency" analysis.

For categorical variables "history segment", "zip code", "newbie" and "channel", a histogram is inadequate to represent the ratio between the entire data and every single element. Pie charts offer better visualization and are provided for all above four variables. The categorical variables "men" and "women" are considered as a whole. A pie chart is plotted for evaluating the percentage of customers purchased only men's product, women's product or both in past twelve months.

The variable "history" is regarded as a numeric variable. The box plot (a.k.a. box and whisker diagram) is a standardized way of displaying the distribution of data.[1] According to the Box-plot of "history", it not only extracts basic statical properties, but also displays outliers obviously. A further data correction can be operated concerning outliers.

Design for Question 2

Question 2 requires studying the relationship between each predictive variable and the target variable "visit".

Firstly, a histogram about the volume of customers with "visit" == 1 against "recency" is plotted, and a stacked bar chart about customers with "visit" == 1 and 0 against "recency" is plotted right after for better visualization. A line chart is provided for the percentage of customers with "visit" == 1 against "recency" with the intention to show the rate of change for each month. The Same mechanism is applied to analyze "history segment", "men" and "women".

For variables "zip code", "newbie" and "channel", stacked bar charts are maintained for visualizing dimensionality. But bar chart is utilized for percentage display.

The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories.[2] Hence, chi-squred test is the proper statical and it is performed to analyze all variable.

Design for Question 3

Question 3 requires studying the relationship between each predictive variable and the target variable "conversion" when "visit" == 1.

For each variable, a bar chart about the volume of customers with "conversion" == 1 against "recency" is plotted. Afterwards, according to customer population, the program depicts line charts about the percentage of customers who visited the website within three months after being contacted. The same mechanism is applied to plot all variables except "history".

Similar to question 2, chi-square test is applied to analyze whether predictive variables are independent to conversion.

Design for Question 4

Question 4 requires creating a linear regression over the relevant predictive variables and the variable "spend". A simple linear regression is applied to the variable "history" since it is the only numerical variable.

R Program Implementation and Result

Question 1 Implementation and Result

The first histogram is the volume of customers against "recency". The source code and result follows:

```
hist(direct_marketing$recency,breaks=seq(0,12,by=1),
     xlim = c(0,12), ylim = c(0, 7000), labels = TRUE,
     main = "Distribution of Months Since Last Purchase",
     xlab="Months Since Last Purchase",
     ylab = "Population", col=c("red3", "grey17"))
```

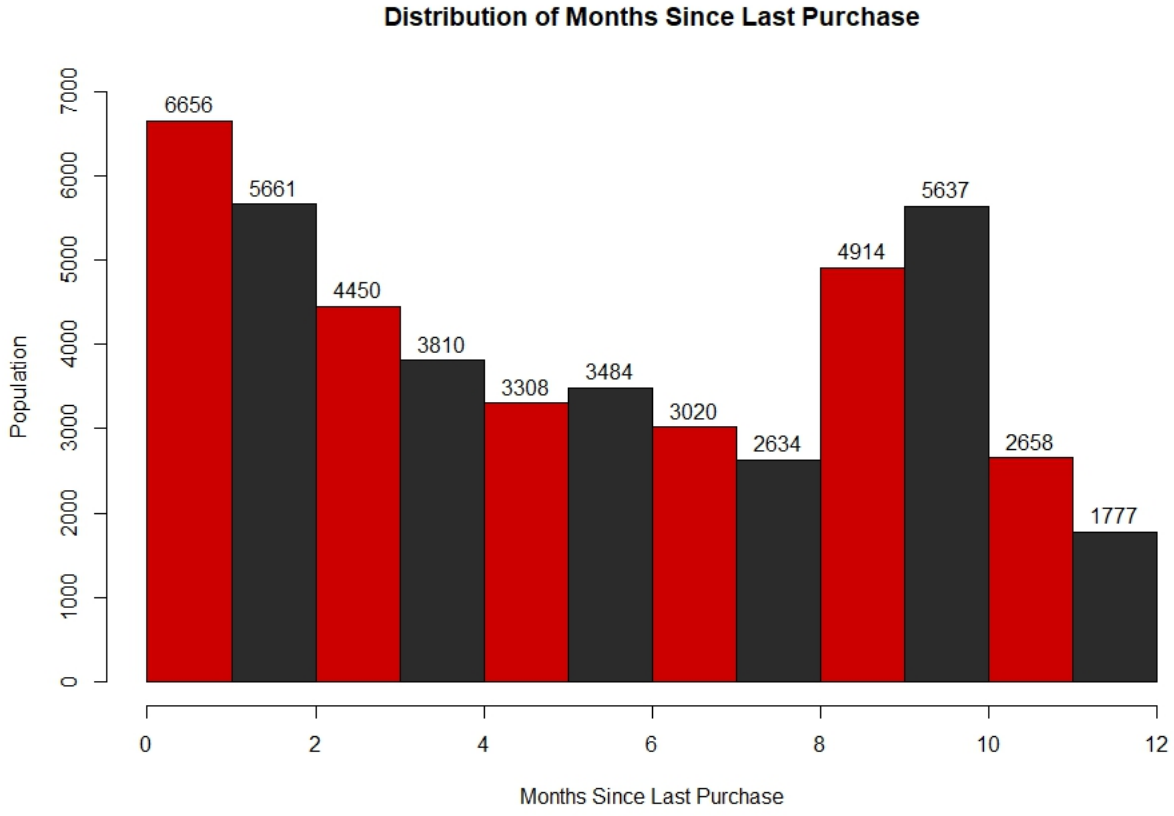

Figure 1: Distribution of Months Since Last Purchase

Since there are five pie charts are plotted, a function is extracted as below:

```
table_draw_pie <- function(table, title){
        slices <- as.vector(table)
```

```
        lbls <- names(table)
        pct <- round(slices/sum(slices)*100)
        lbls <- paste(lbls, pct)
        lbls <- paste(lbls, "%", sep="")
        pie(slices, labels = lbls, main = title)
}
```

Let's take the variable "history segment" as a step-by-step example.

1. Source code:

```
        h_s <- table(direct_marketing$history_segment)
```

   Result:

```
        > h_s
      1) $0 - $100    2) $100 - $200    3) $200 - $350
           17215             10727              9242
      4) $350 - $500    5) $500 - $750 6) $750 - $1.000
           1388              4803              3669
       7) $1.000 +
            965
```

2. Source code:

```
        table_draw_pie(h_s, "Pie Chart of history_segment")
```

3. Inside the function, extract numbers and labels from table obtained above. Calculate percentage and append percentage to label.

   Source code:

```
        slices <- as.vector(table)
        lbls <- names(table)
        pct <- round(slices/sum(slices)*100)
        lbls <- paste(lbls, pct)
        lbls <- paste(lbls, "%", sep="")
```

   Result:

```
> slices
[1] 17215 10727  9242  4803  3669  1388   965
> lbls
[1] "1) $0 - $100 36%"     "2) $100 - $200 22%"
    "3) $200 - $350 19%"
    "4) $350 - $500 10%"   "5) $500 - $750 8%"
[6] "6) $750 - $1.000 3%" "7) $1.000 + 2%"
```

4. Finally Draw pie chart.
   Source code:

```
pie(slices, labels = lbls,
    col=rainbow(length(lbls)), main = title)
```

   Result:

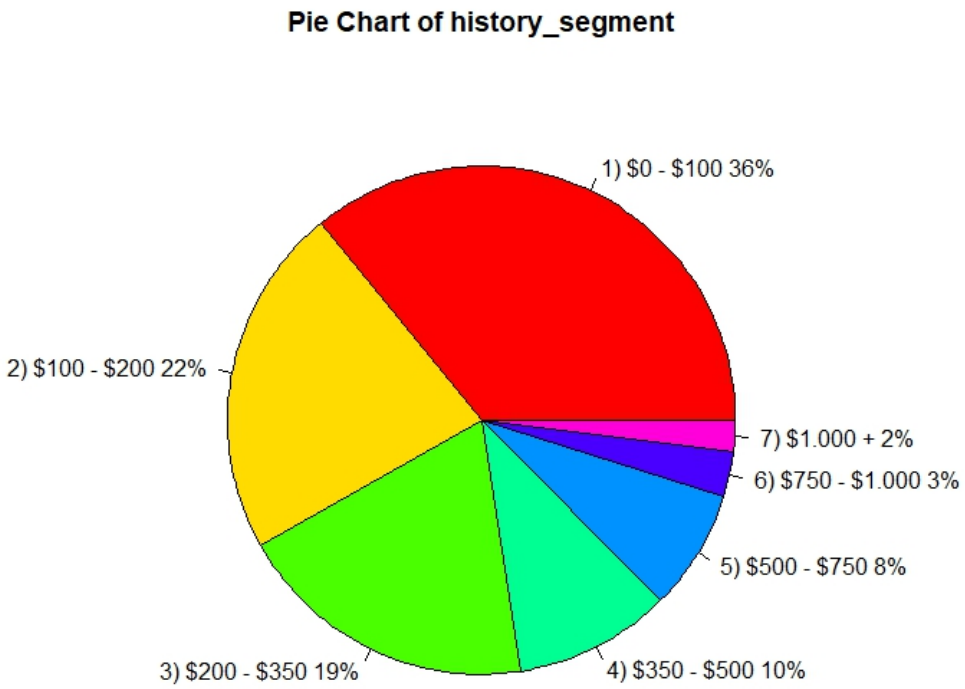

Figure 2: Pie chart of history segment

Other pie charts are listed below.

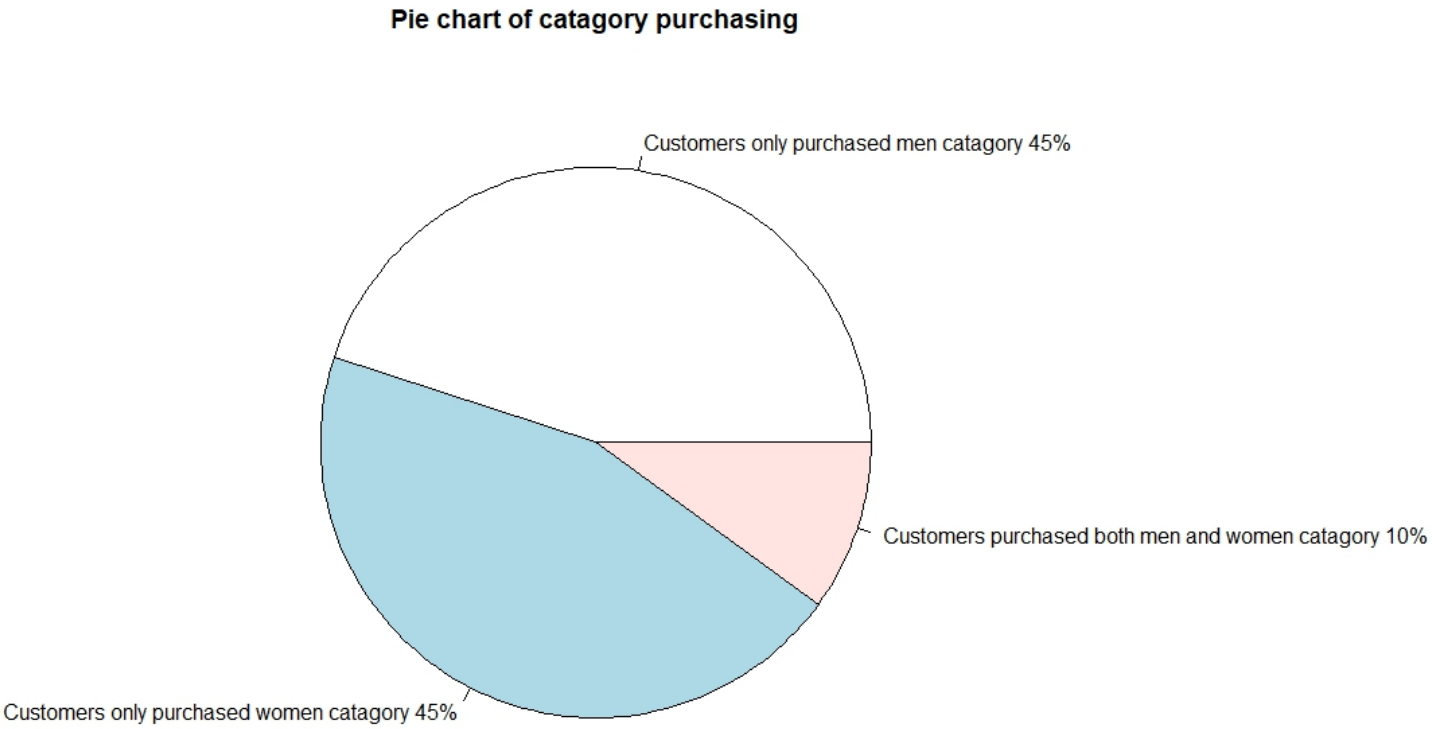

Figure 3: Pie chart of category purchasing

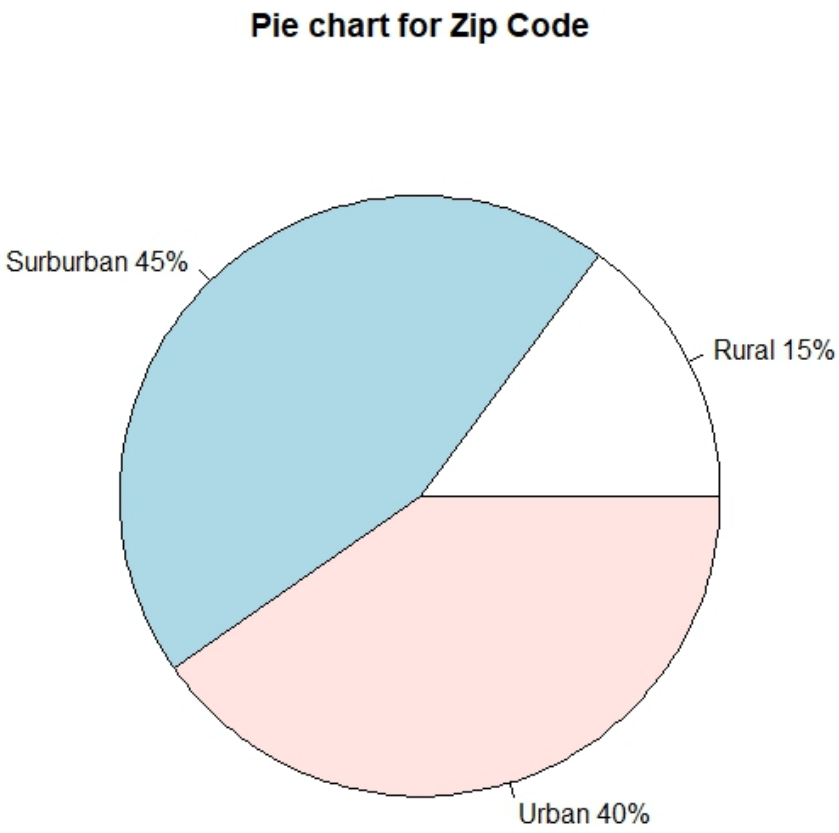

Figure 4: Pie chart of Zip Code

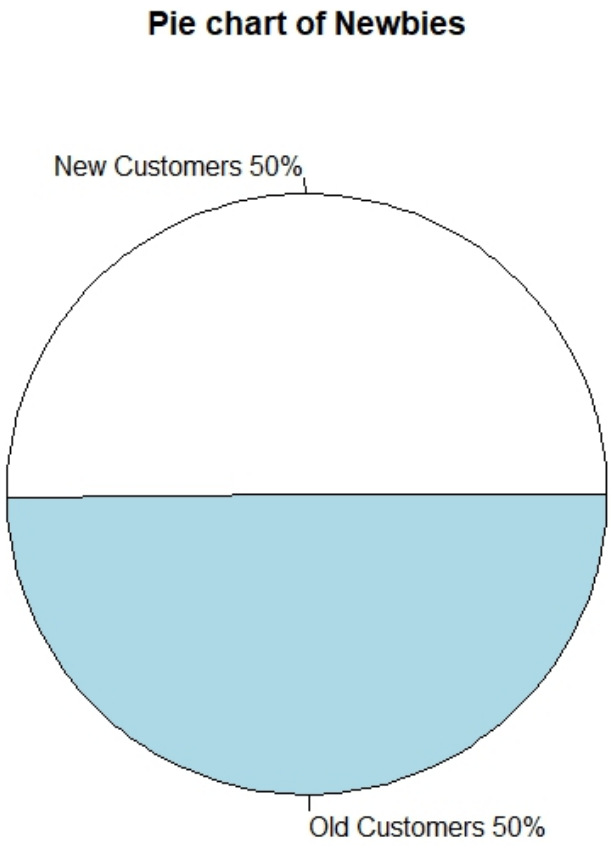

Figure 5: Pie chart of Newbies

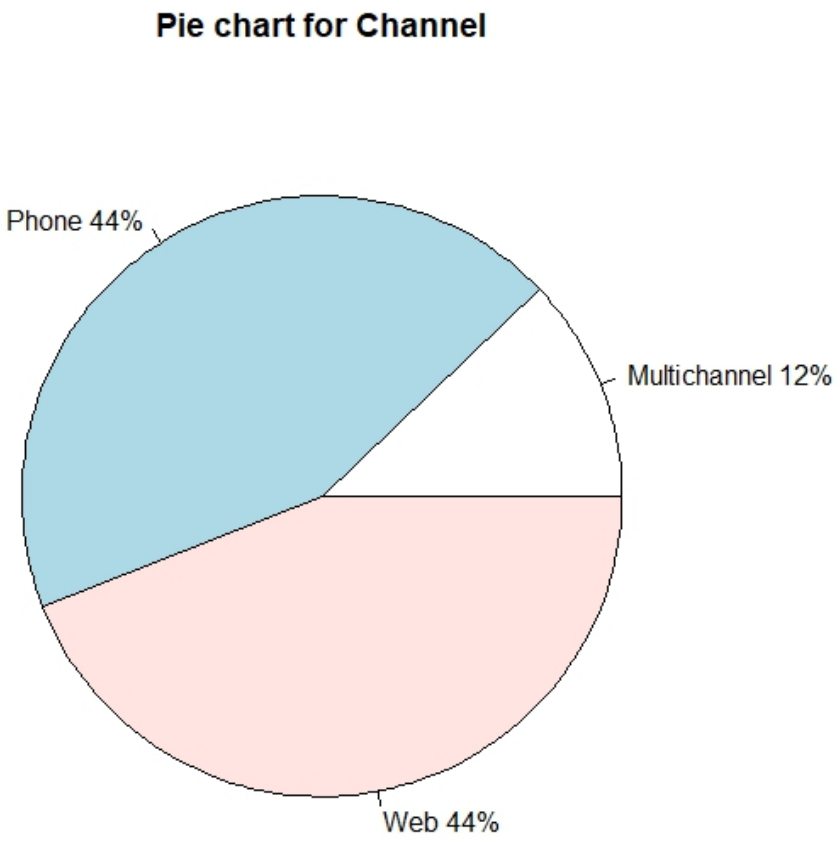

Figure 6: Pie chart of Channel

Steps of generating box plot for the variable "history" are explained as below.

1. Draw box plot for "history".

```
bxp <- boxplot(history_value, varwidth=TRUE,
                main="Boxplot of history")
```

2. Label five number summary to the box plot. Adjust the label position and font size

```
text(x = col(bxp$stats) - .5, y = bxp$stats,
     labels = bxp$stats,cex  = 0.8)
```

3. Result

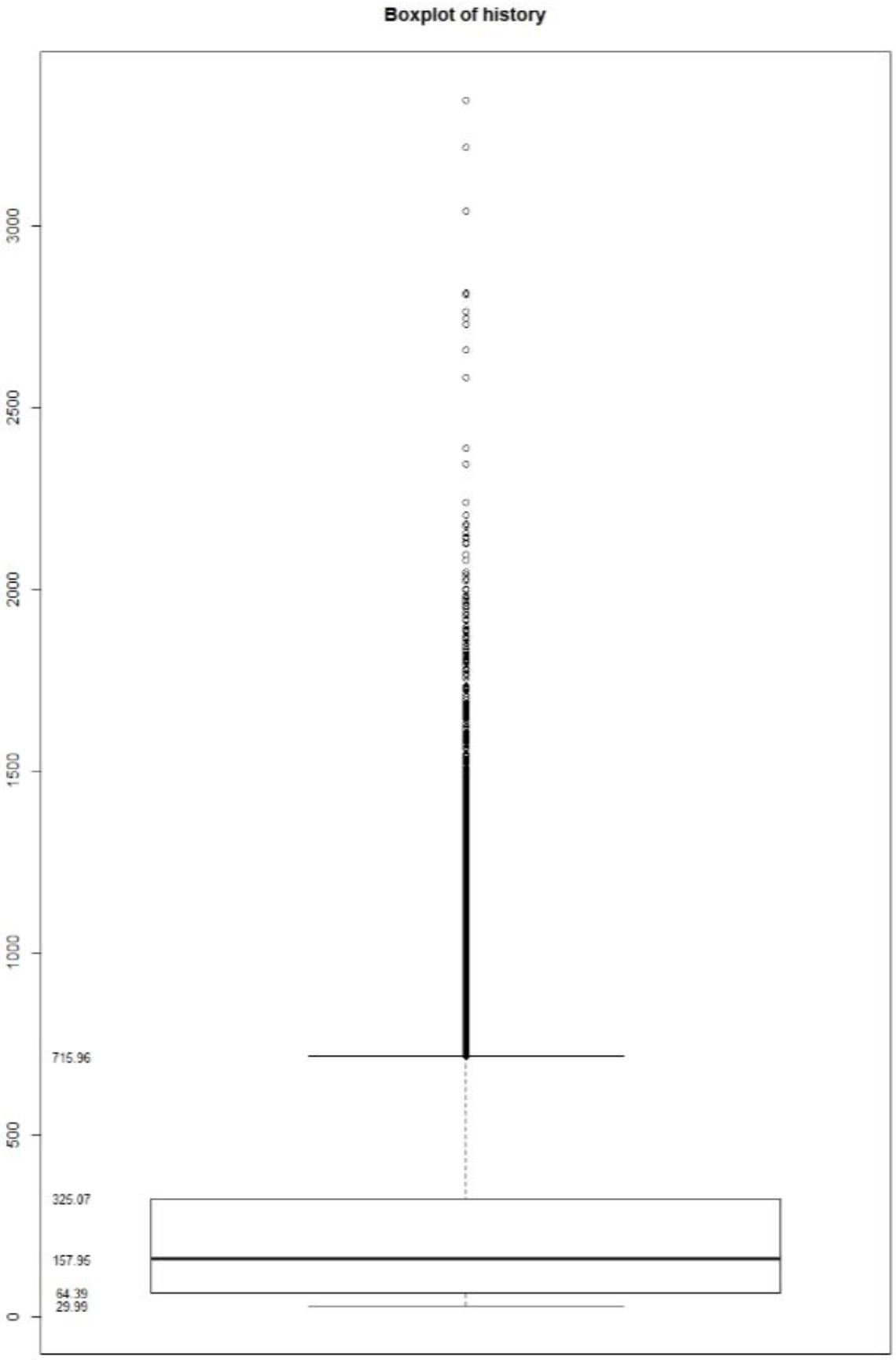

Figure 7: Box plot of history

Since there are many outliers appears on the box plot, when analyzing the variable "history", all outliers need to be eliminated for data correction.

Question 2 Implementation and Result

As mentioned in section 2.3, for all variables, there are stacked bar charts for visualizing the comparison of volume of customers with "visited" == 0 and "visited" == 1, and line charts or bar charts of percentage of customers with "visit" == 1. The analysis by using Chi-Square test is included as well. Let's take "channel" as a step-by-step example.

At the beginning of the illustration, a general hypothesis can be made for all variables:
Ho: The variable is independent of visited.
Ha: The variable is not independent of visited.

1. Extract column "visited"
   Source code :

```
visit_col = direct_marketing$visit
```

2. Construct a matrix with only "channel" and "visited", and substitute "visited" == 1 as "Visited", "visited" == 0 as "Not Visited".
   Source code :

```
dt_channel_and_visit = data.frame(chn = channel,
                          visited = visit_col)
dt_channel_and_visit$visited<-
        replace(dt_channel_and_visit$visited,
                dt_channel_and_visit$visited == 1,
                "Visited")
dt_channel_and_visit$visited<-
        replace(dt_channel_and_visit$visited,
                dt_channel_and_visit$visited == 0,
                "Not Visited")
```

   Result:

```
> head(dt_channel_and_visit)
chn      visited
1 Phone Not Visited
```

```
      2    Web Not Visited
      3    Web Not Visited
      4    Web Not Visited
      5 Phone     Visited
      6 Phone Not Visited
```

3. Apply ddply to count the number of visited and not visited customers.
   Source code :

```
dt_channel_and_visit <- ddply(dt_channel_and_visit,
                              .(chn, visited), nrow)
```

   Result:

```
> dt_channel_and_visit
           chn      visited    V1
1 Multichannel Not Visited  4843
2 Multichannel      Visited   982
3        Phone Not Visited 18365
4        Phone      Visited  2676
5          Web Not Visited 17772
6          Web      Visited  3371
```

4. Apply ggplot to create the stacked bar chart.
   Source code :

```
ggplot(dt_channel_and_visit,
       aes(x = chn, y = V1, fill=visited)) +
   geom_bar(stat = "identity",width = 0.6) +
   geom_text(data = dt_channel_and_visit,
             aes(x = chn, y = V1,label = V1),
             size=4,vjust = 1.5) +
   labs(title = "Comparison between Customers who visited
                 website or not within 3 months after
                 being contacted according to Channel",
        x="Channel types", y="Population")
```

   Result:

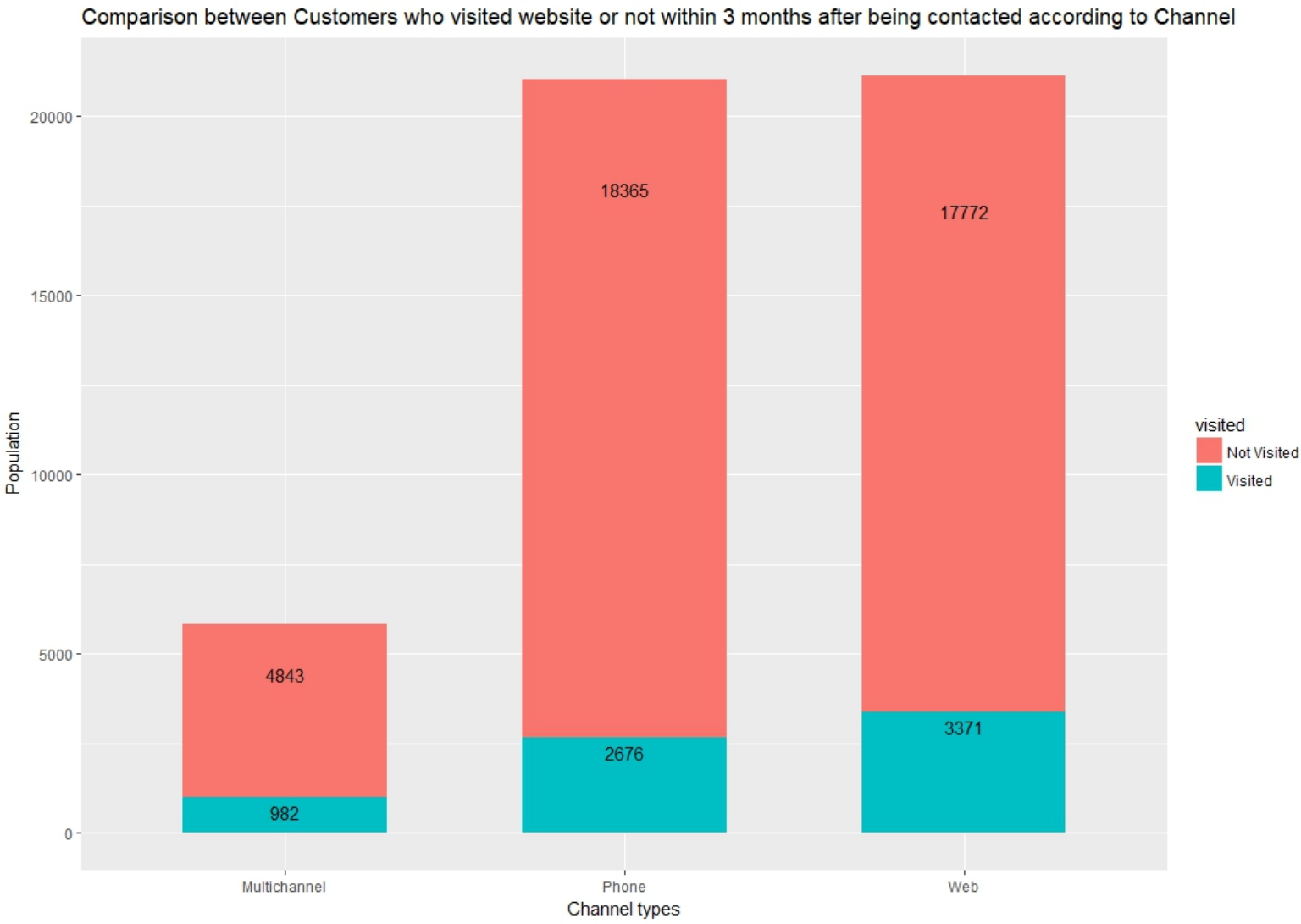

Figure 8: Stacked bar chart of Visit against Channel

5. Obtain the amount of customers for each channel.
   Source code :

```
channel_total_customers = aggregate(V1 ~ chn,
                    dt_channel_and_visit, sum)[,"V1"]
```

   Result:

```
> channel_total_customers
[1]  5825 21041 21143
```

6. Obtain amount of customers who visited website for each channel.
   Source code :

```
channel_visited_counts <- subset(dt_channel_and_visit,
                                 visited == "Visited")[,"V1"]
```

Result:

```
> channel_visited_counts
[1]  982 2676 3371
```

7. Calculate the percentage of customers who visited website for each channel and create a data frame for the percentage.
   Source code :

```
channel_visited_pct <-
        round(channel_visited_counts/channel_total_customers*100)
channel_visited_pct_dt <-
        data.frame(obj = c(0:2), val = channel_visited_pct)
```

Result:

```
> channel_visited_pct_dt
  obj val
1   0  17
2   1  13
3   2  16
```

8. Apply ggplot to create the bar chart of percentage.
   Source code :

```
ggplot(channel_visited_pct_dt, aes(x = obj, y = val), size = 5) +
        geom_bar(stat = "identity",width = 0.6, fill = "bisque") +
        scale_x_continuous(breaks=seq(1,3,by=1)) +
        geom_text(data = channel_visited_pct_dt,
                  aes(x = obj, y =(val),label = paste0(val, "%")),
                  size=4,vjust = -0.5) +
        labs(title = "Percentage of Customers who visited website
             within 3 months after being contacted
             according to Channel",
             x="Chanel types", y="Percentage") +
        scale_x_discrete(limits=c(0:2),
                         labels=c("Multichannel", "Phone", "Web"))
```

Result:

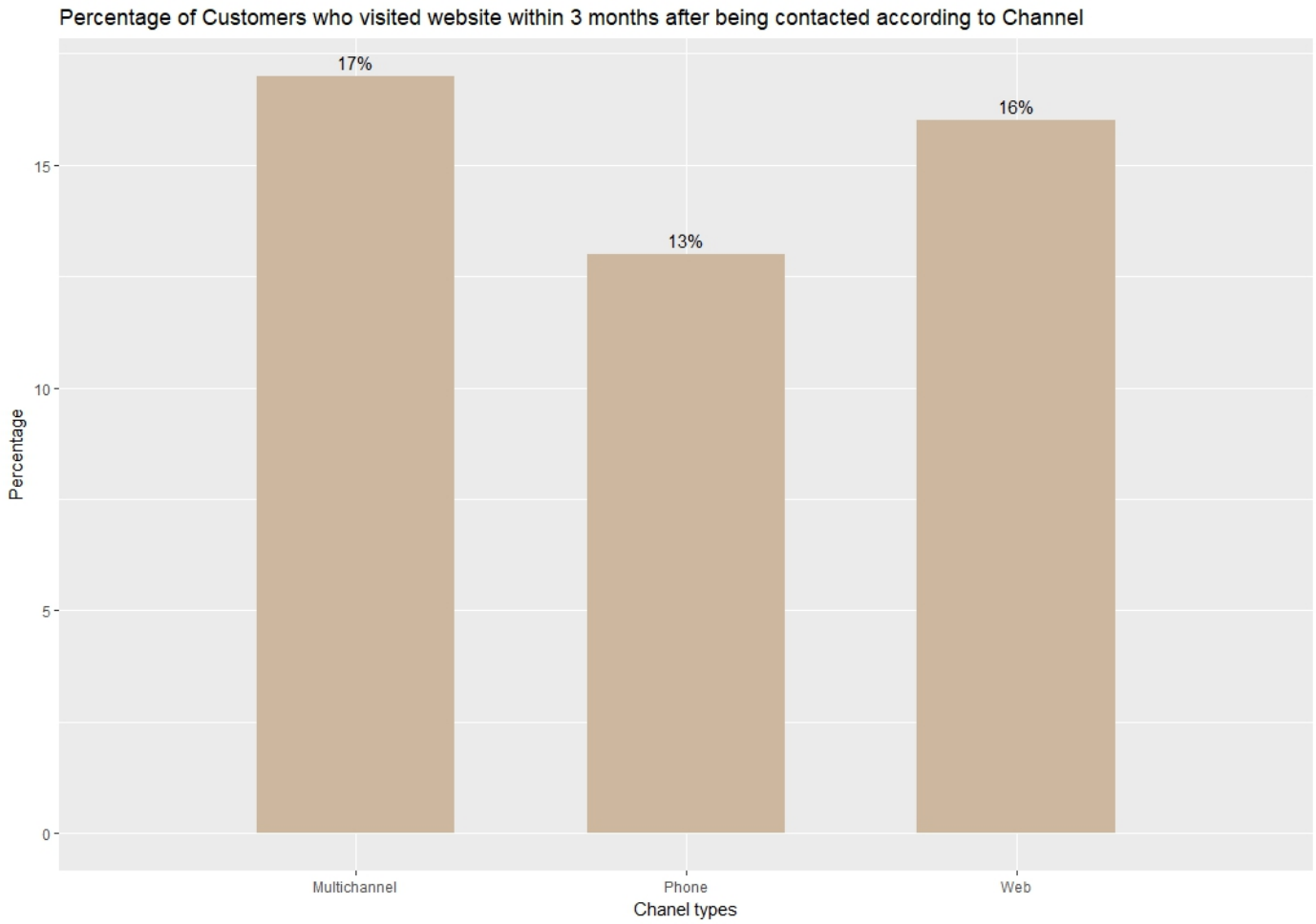

Figure 9: Percentage bar chart of Visit against Channel

9. Extract table for Chi-Square test

Channel	Multichannel	Phone	Web	Row Sum
Visited	982	2676	3371	7029
Not Visited	4843	18365	17772	40980
Col Sum	5825	21041	21143	48009

Figure 10: Table of Channel for Chi-Square testing

10. Apply the build-in Chi-Square function

Source code :

```
channel_chi_table = rbind(channel_visited_counts,
channel_not_visited_counts)
channel_chi_table
```

Result:

```
channel_visited_counts      982  2676  3371
channel_not_visited_counts 4843 18365 17772
```

Source code :

```
chisq.test(channel_chi_table)
```

Result:

```
        Pearson's Chi-squared test

data:  channel_chi_table
X-squared = 113.89, df = 2, p-value < 2.2e-16
```

Since the p-value is smaller than 0.05, Ho is rejected. Therefore, channel is not independent of visit.

Charts and analysis for other variables are listed below.

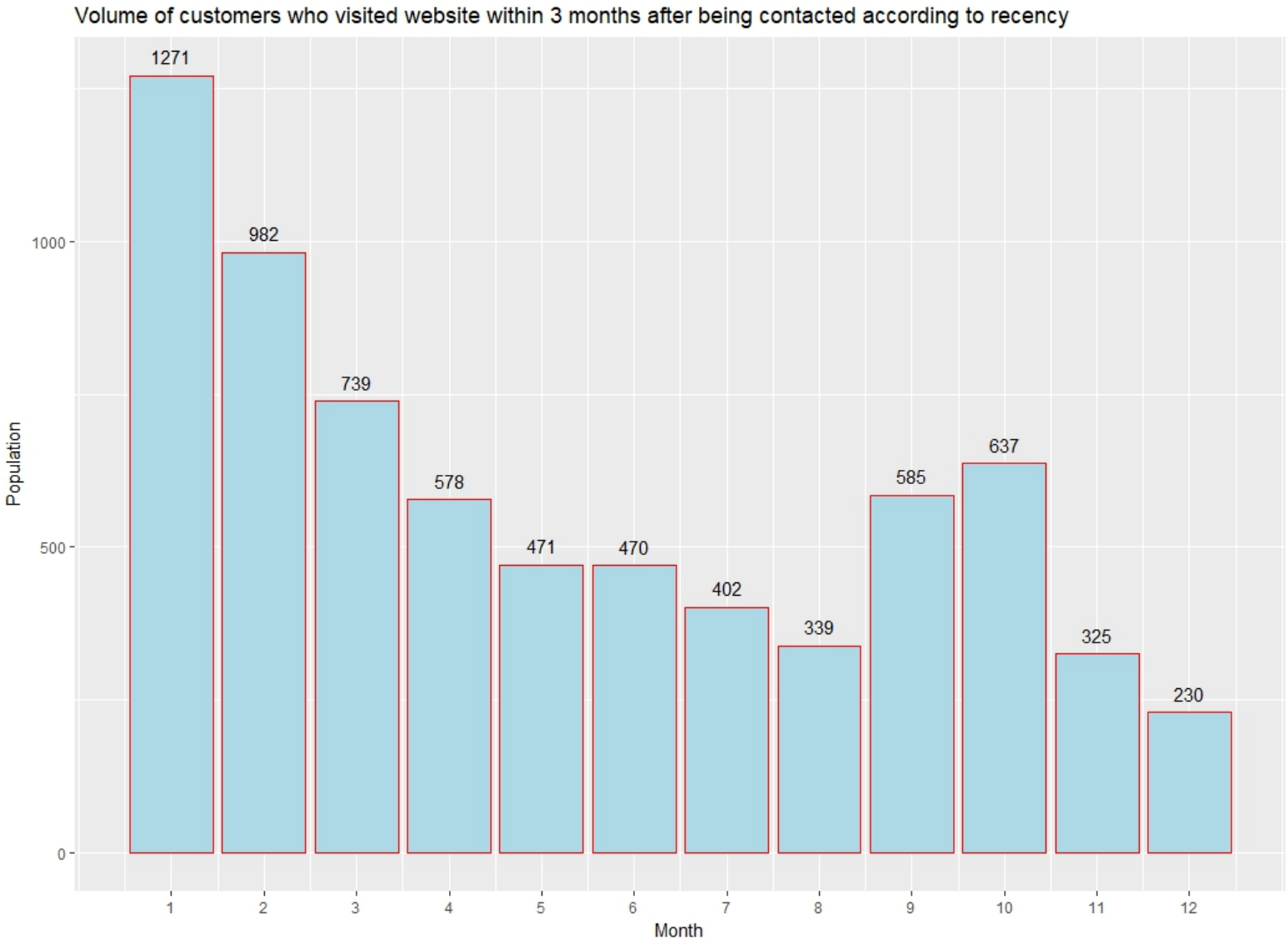

Figure 11: Population bar chart of Visit against Recency

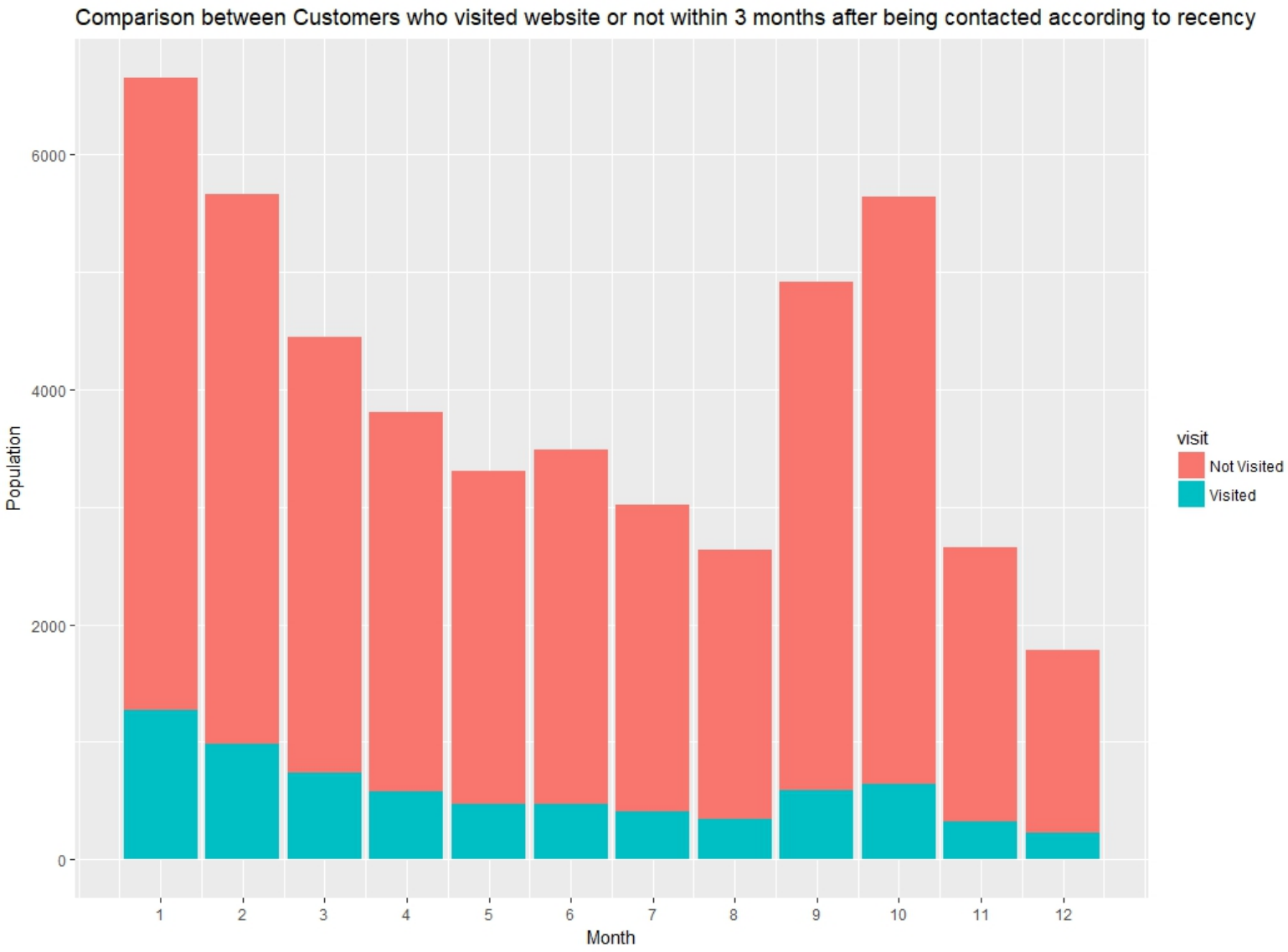

Figure 12: Stacked bar chart of Visit against Recency

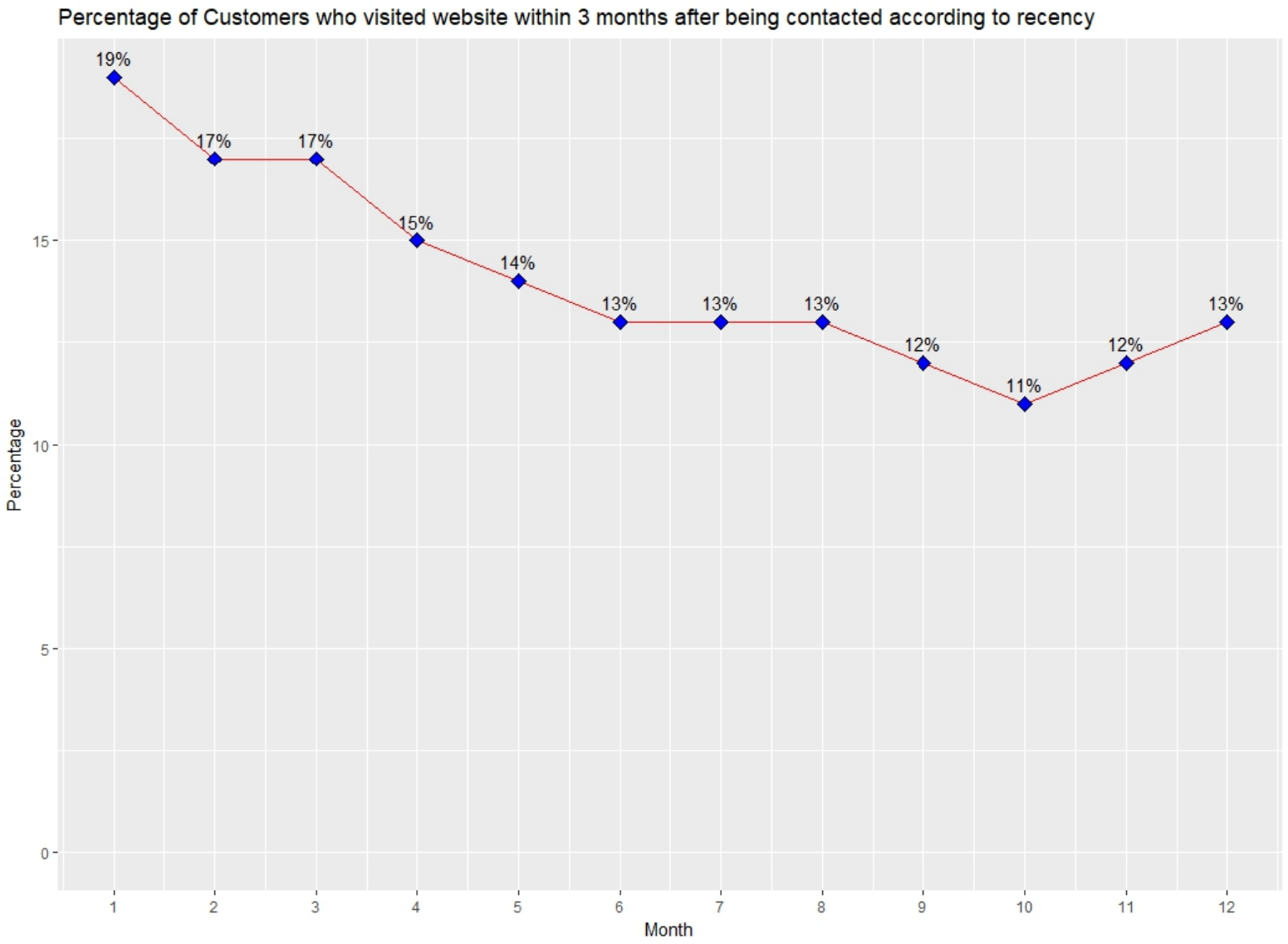

Figure 13: Line chart of Visit against Recency

Recency	1	2	3	4	5	6	7	8	9	10	11	12	Row Sum
Visited	1271	982	739	578	471	470	402	339	585	637	325	230	7029
Not Visited	5385	4679	3711	3232	2837	3014	2618	2295	4329	5000	2333	1547	40980
Col Sum	6656	5661	4450	3810	3308	3484	3020	2634	4914	5637	2658	1777	48009

Figure 14: Table of Recency for Chi-Square testing

Chi-Square Result of Recency:

```
        data:  recency_chi_table
        X-squared = 264.73, df = 11, p-value < 2.2e-16
```

Since the p-value is smaller than 0.05, Ho is rejected. Therefore, recency is not independent of visit.

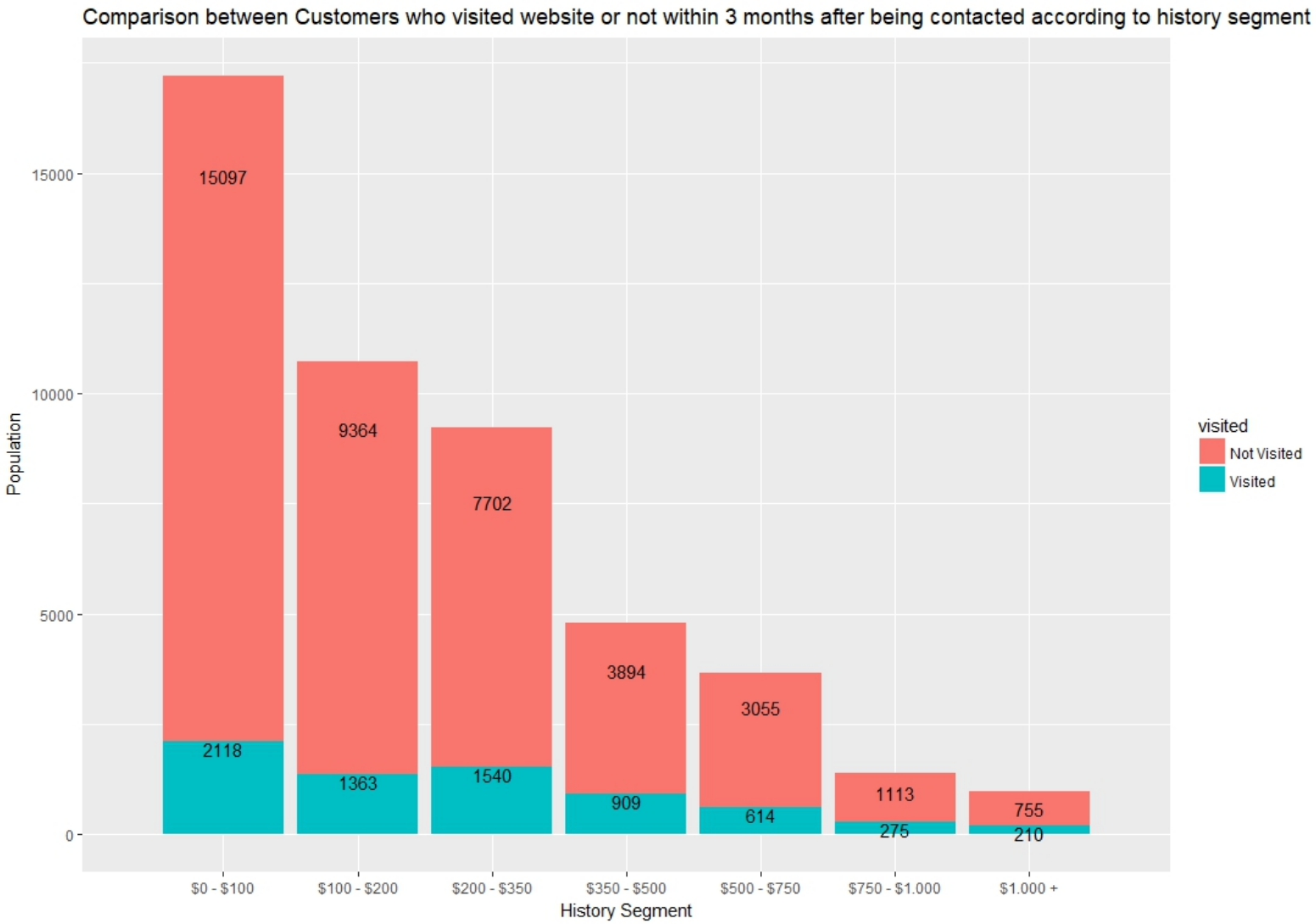

Figure 15: Stacked bar chart of Visit against History Segment

Figure 16: Line chart of Visit against History Segment

History Segment	$0~$100	$100~$200	$200~$350	$350~$500	$500~$750	$750~$1000	$1000+	Row Sum
Visited	2118	1363	1540	909	614	275	210	7029
Not Visited	15097	9364	7702	3894	3055	1113	755	40980
Col Sum	17215	10727	9242	4803	3669	1388	965	48009

Figure 17: Table of History Segment for Chi-Square testing

Chi-Square Result of history segment:

```
data:  hist_seg_chi_table
X-squared = 289.93, df = 6, p-value < 2.2e-16
```

Since the p-value is way smaller than 0.05, Ho is rejected. Therefore, history segment is not independent of visit.

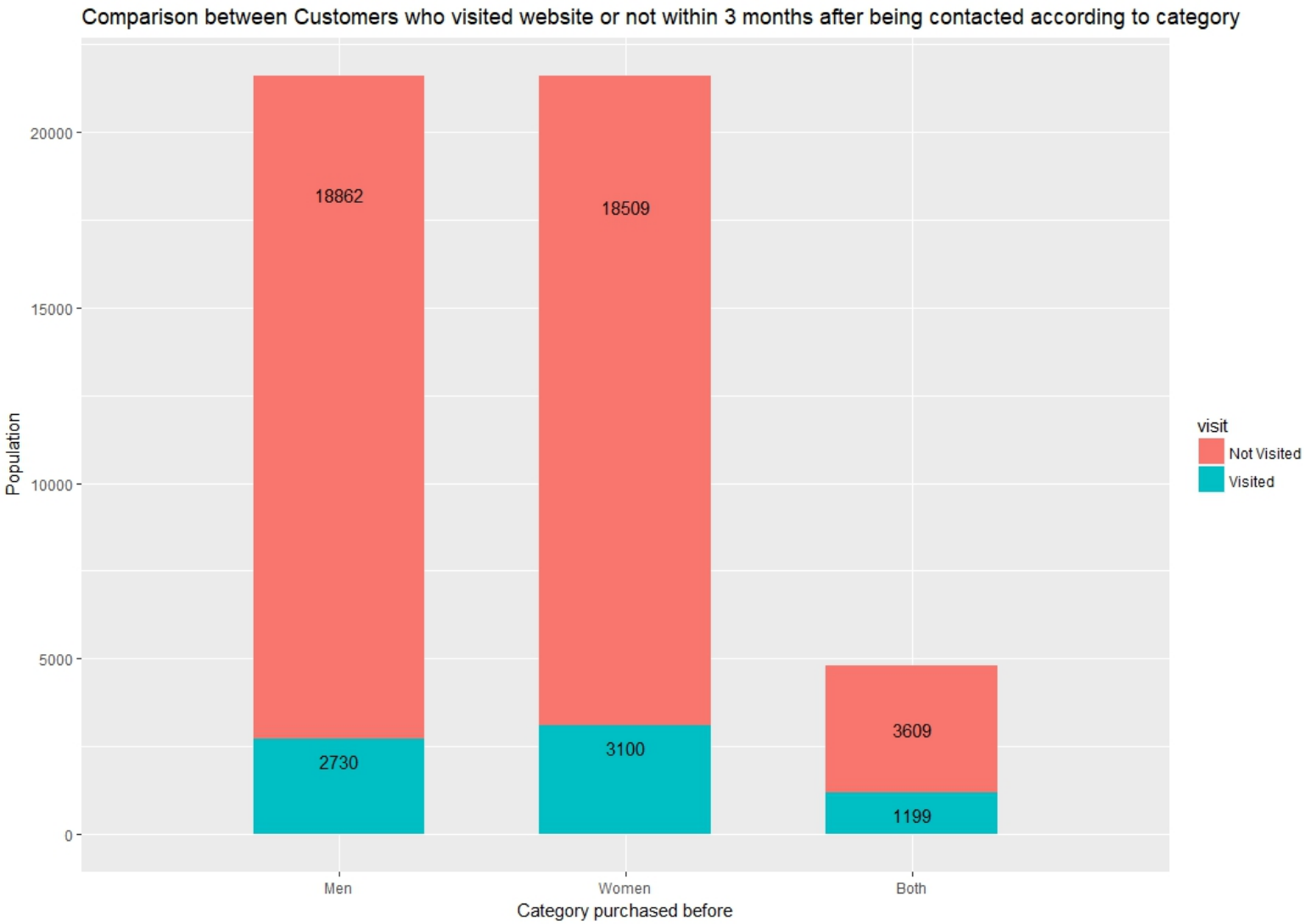

Figure 18: Stacked bar chart of Visit against Category

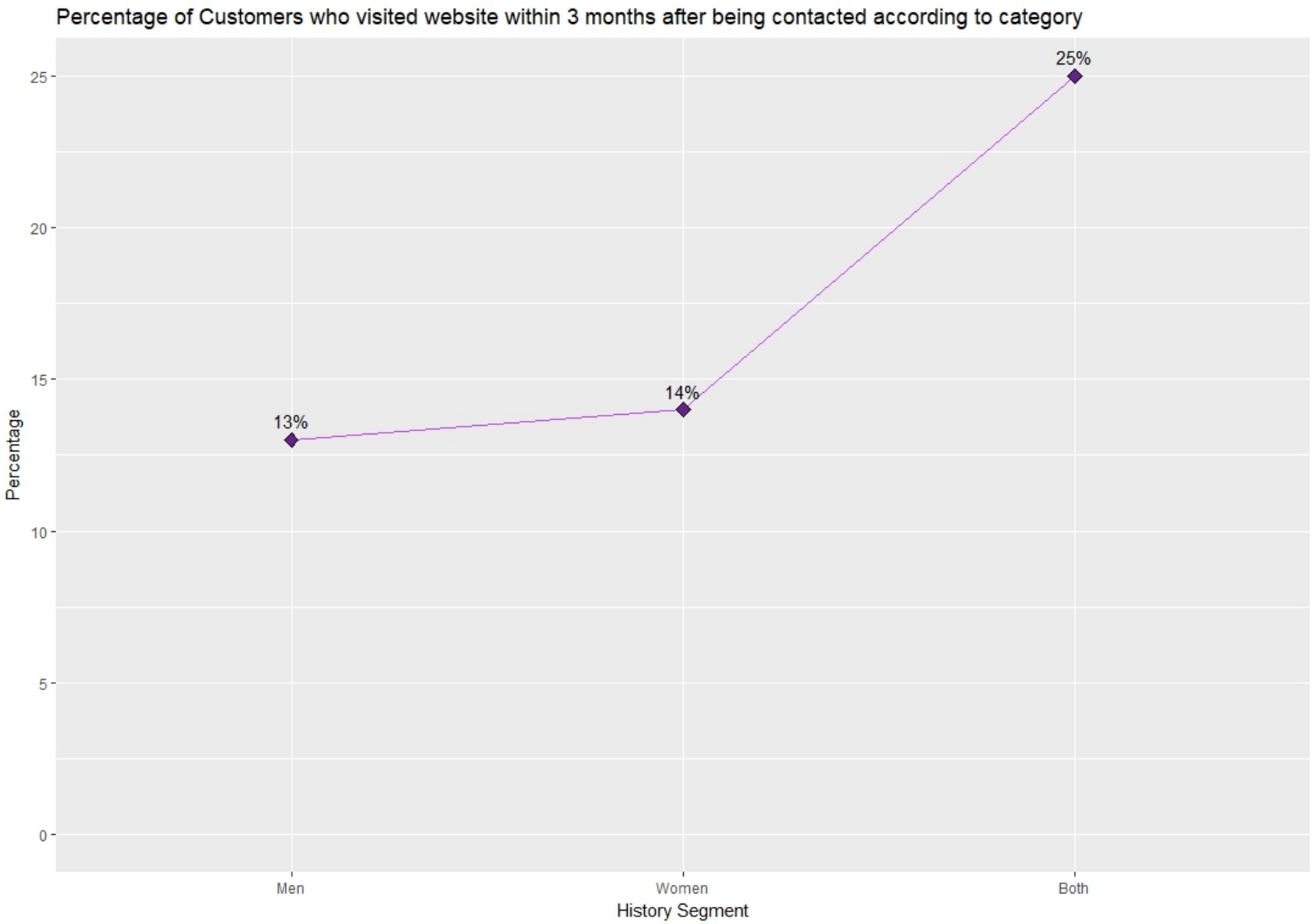

Figure 19: Line chart of Visit against Category

Category	Men	Women	Both	Row Sum
Visited	2730	3100	1199	7029
Not Visited	18862	18509	3609	40980
Col Sum	21592	21609	4808	48009

Figure 20: Table of Category for Chi-Square testing

Chi-Square Result of Category:

```
        data:  category_chi_table
        X-squared = 478.32, df = 2, p-value < 2.2e-16
```

Since the p-value is way smaller than 0.05, Ho is rejected. Therefore, category is not independent of visit.

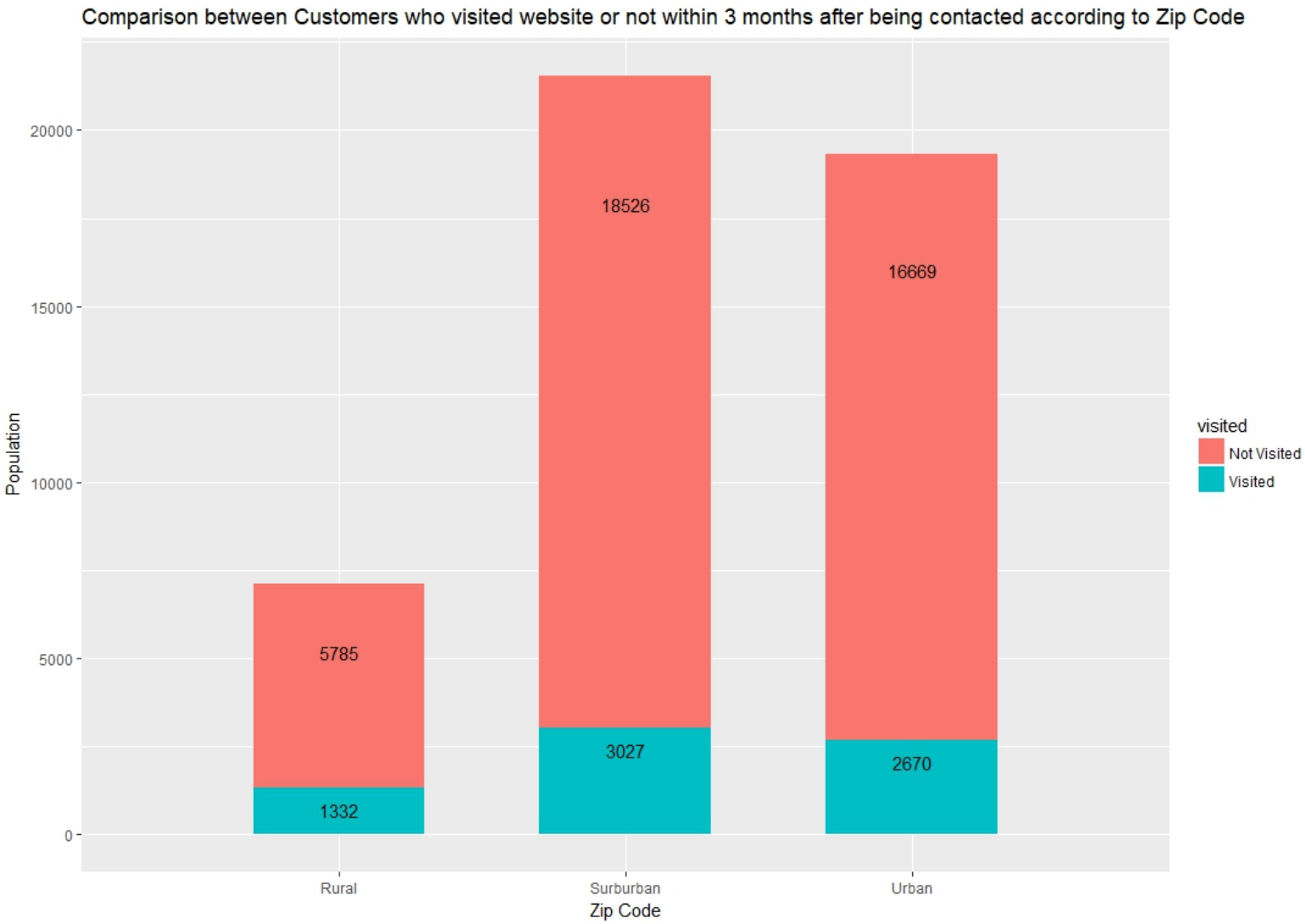

Figure 21: Stacked bar chart of Visit against Zip Code

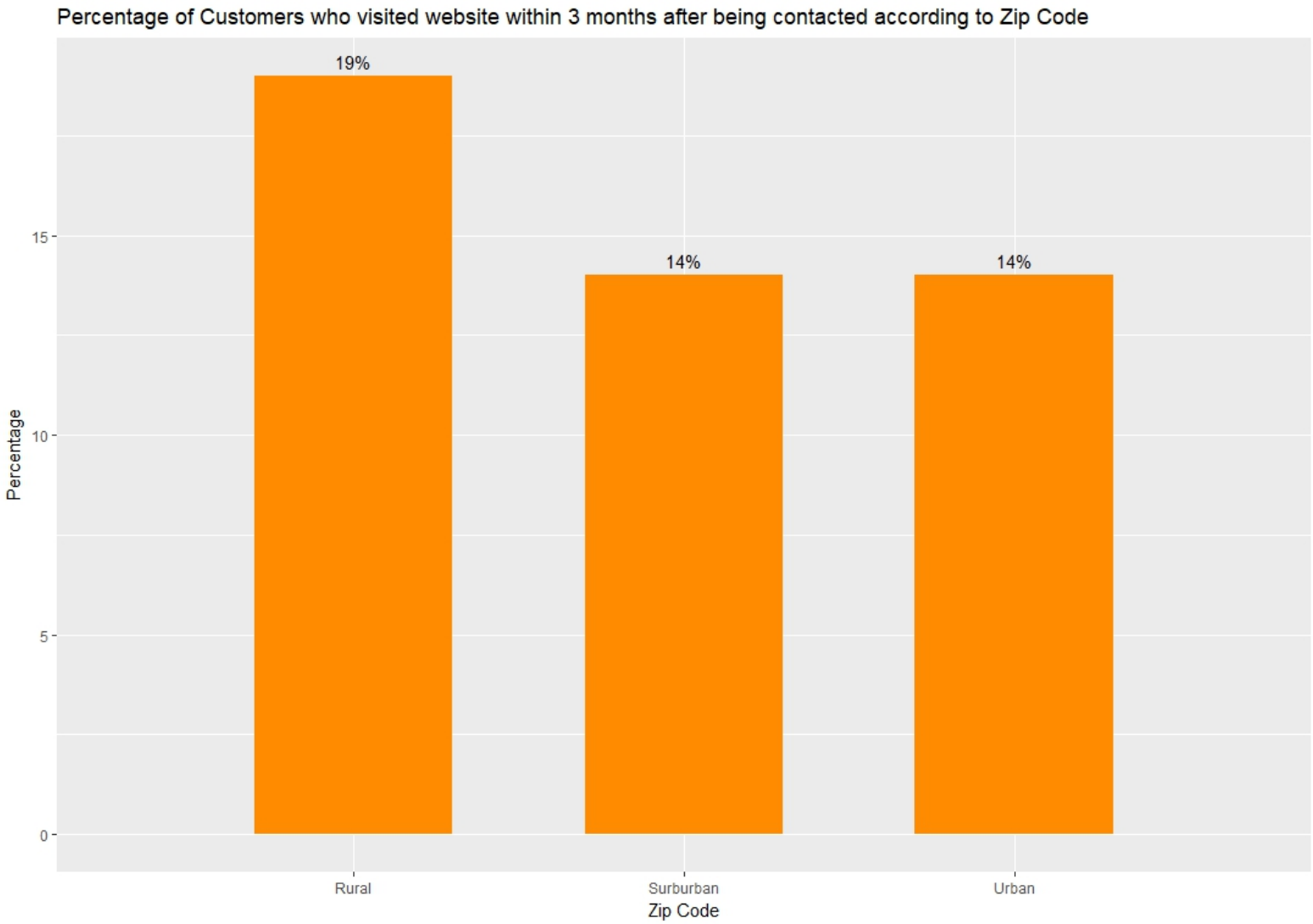

Figure 22: Percentage bar chart of Visit against Zip Code

Zip code	Rural	Surburban	Urban	Row Sum
Visited	1332	3027	2670	7029
Not Visited	5785	18526	16669	40980
Col Sum	7117	21553	19339	48009

Figure 23: Table of Zip Code for Chi-Square testing

Chi-Square Result of Zip Code:

```
    data:  zip_chi_table
    X-squared = 111.47, df = 2, p-value < 2.2e-16
```

Since the p-value is way smaller than 0.05, Ho is rejected. Therefore, zip code is not independent of visit.

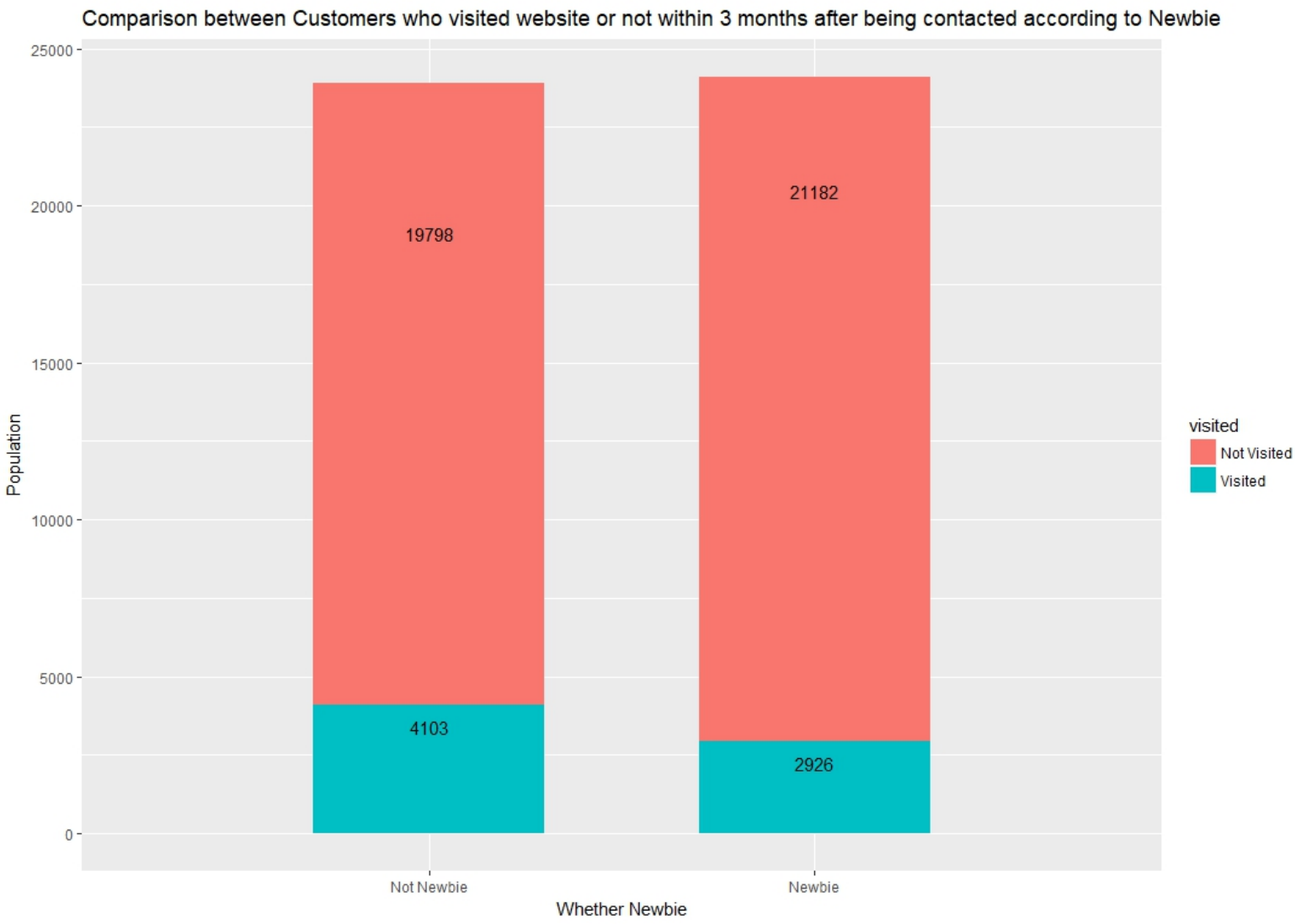

Figure 24: Stacked bar chart of Visit against Newbie

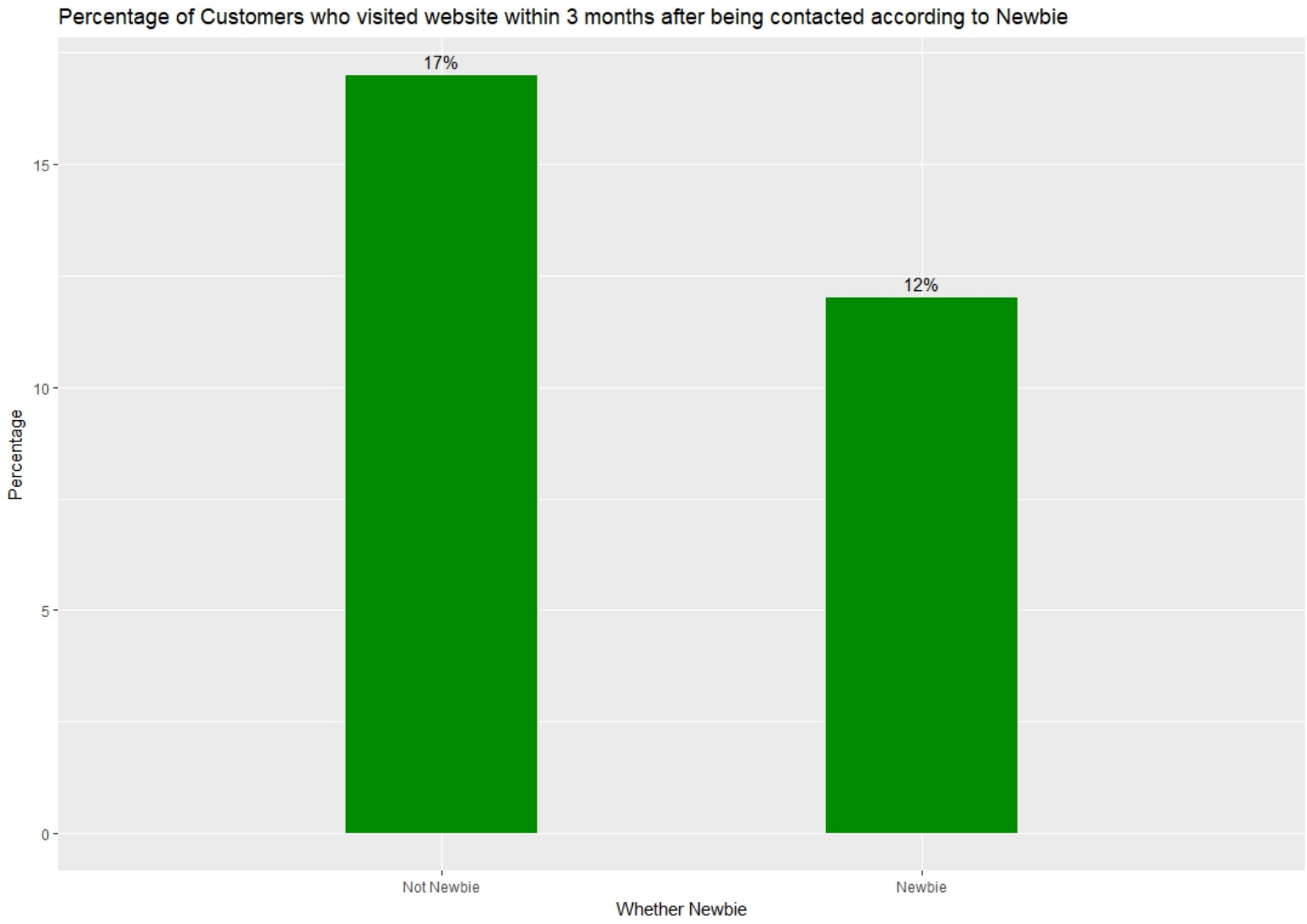

Figure 25: Percentage bar chart of Visit against Newbie

Newbie	Not Newbie	Newbie	Row Sum
Visited	4103	2926	7029
Not Visited	19798	21182	40980
Col Sum	23901	24108	48009

Figure 26: Table of Newbie for Chi-Square testing

Chi-Square Result of Newbie:

```
data:  newbie_chi_table
X-squared = 242.54, df = 1, p-value < 2.2e-16
```

Since the p-value is way smaller than 0.05, Ho is rejected. Therefore, newbie is not independent of visit.

Question 3 Implementation and Result

Same as question 2, for all variables, there are bar charts for visualizing the amount of customers with "conversion" == 1, and line charts or bar charts of percentage of customers with "conversion" == 1. Since comparing with those who with "conversion" == 0, the amount of customers with "conversion" == 1 is so small that it is not necessary to use stacked bar charts. The Chi-Square testing is included for each variable.

The implementation of generating bar charts, line charts and Chi-Square are same as question 2. Therefore, only results are displayed in this section.

Likewise, a general hypothesis can be made for all variables:
Ho: The variable is independent of conversion.
Ha: The variable is not independent of conversion.

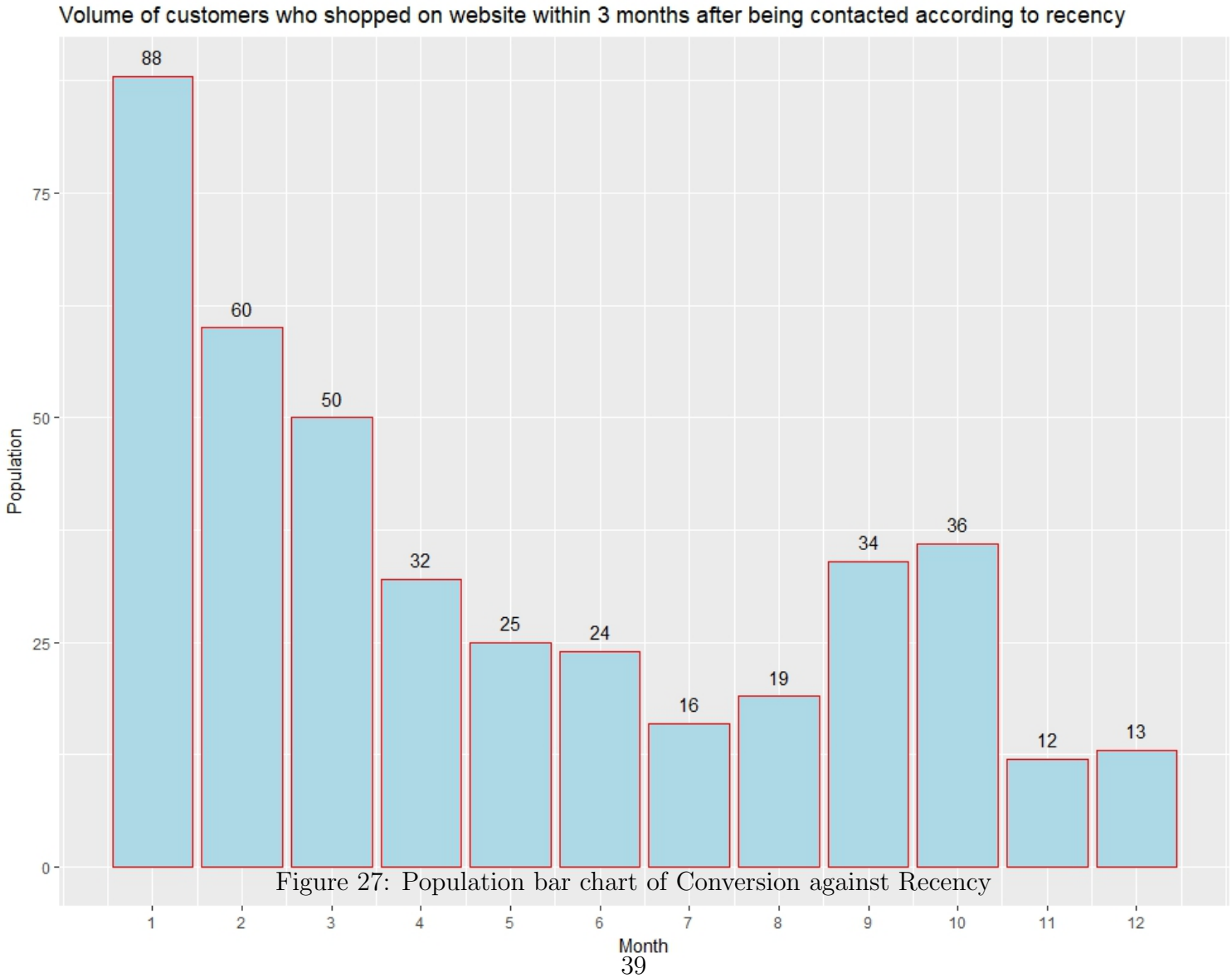

Figure 27: Population bar chart of Conversion against Recency

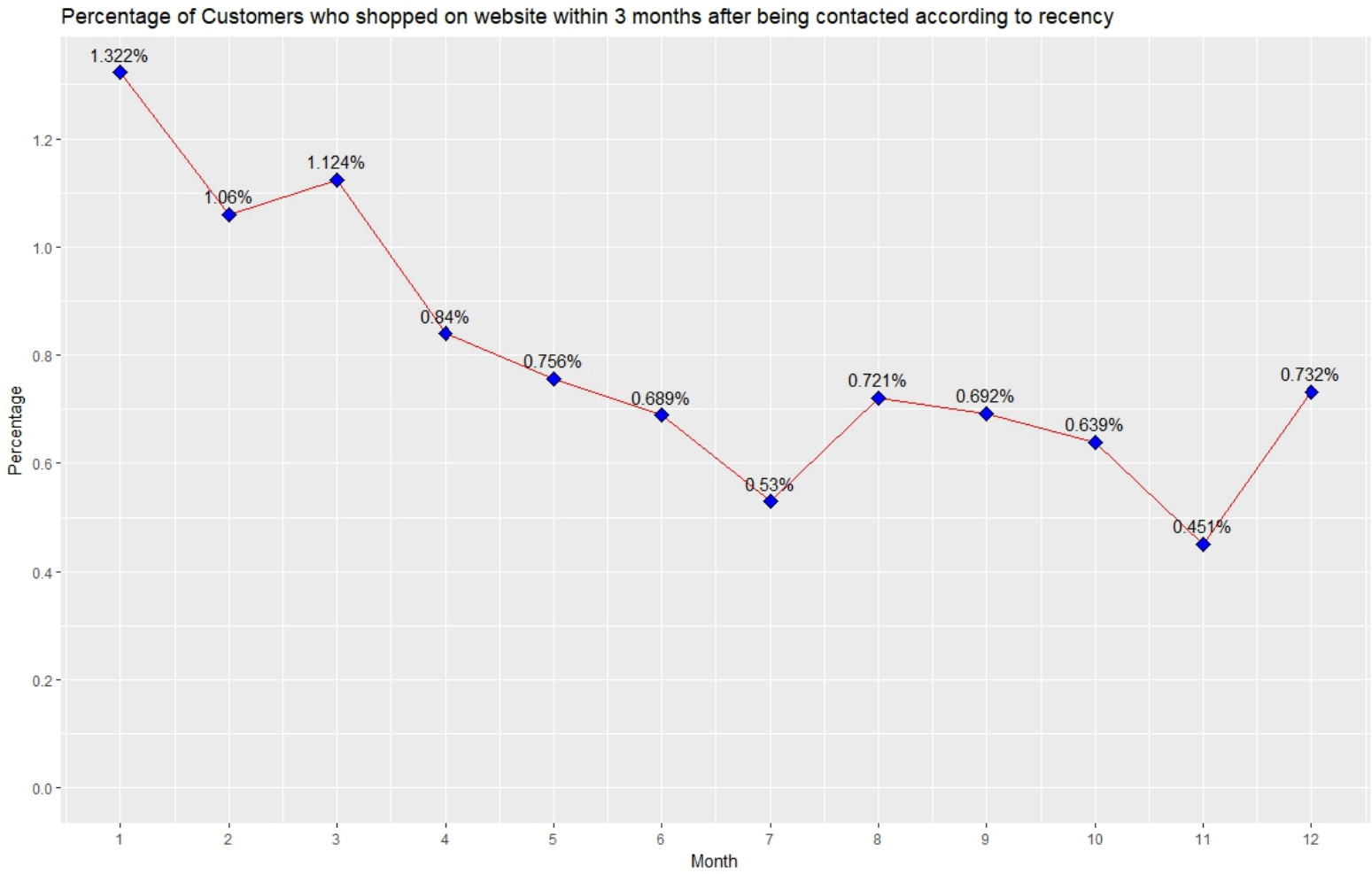

Figure 28: Percentage line chart of Conversion against Recency

Recency	1	2	3	4	5	6	7	8	9	10	11	12	Row Sum
Shopped	88	60	50	32	25	24	16	19	34	36	12	13	409
Not Shopped	6568	5601	4400	3778	3283	3460	3004	2615	4880	5601	2646	1764	47600
Col Sum	6656	5661	4450	3810	3308	3484	3020	2634	4914	5637	2658	1777	48009

Figure 29: Table of Recency for Chi-Square testing

Chi-Square Result of Recency:

```
    data:  recency_chi_table
    X-squared = 39.792, df = 11, p-value = 3.883e-05
```

Since the p-value is smaller than 0.05, Ho is rejected. Therefore, recency is not independent of conversion.

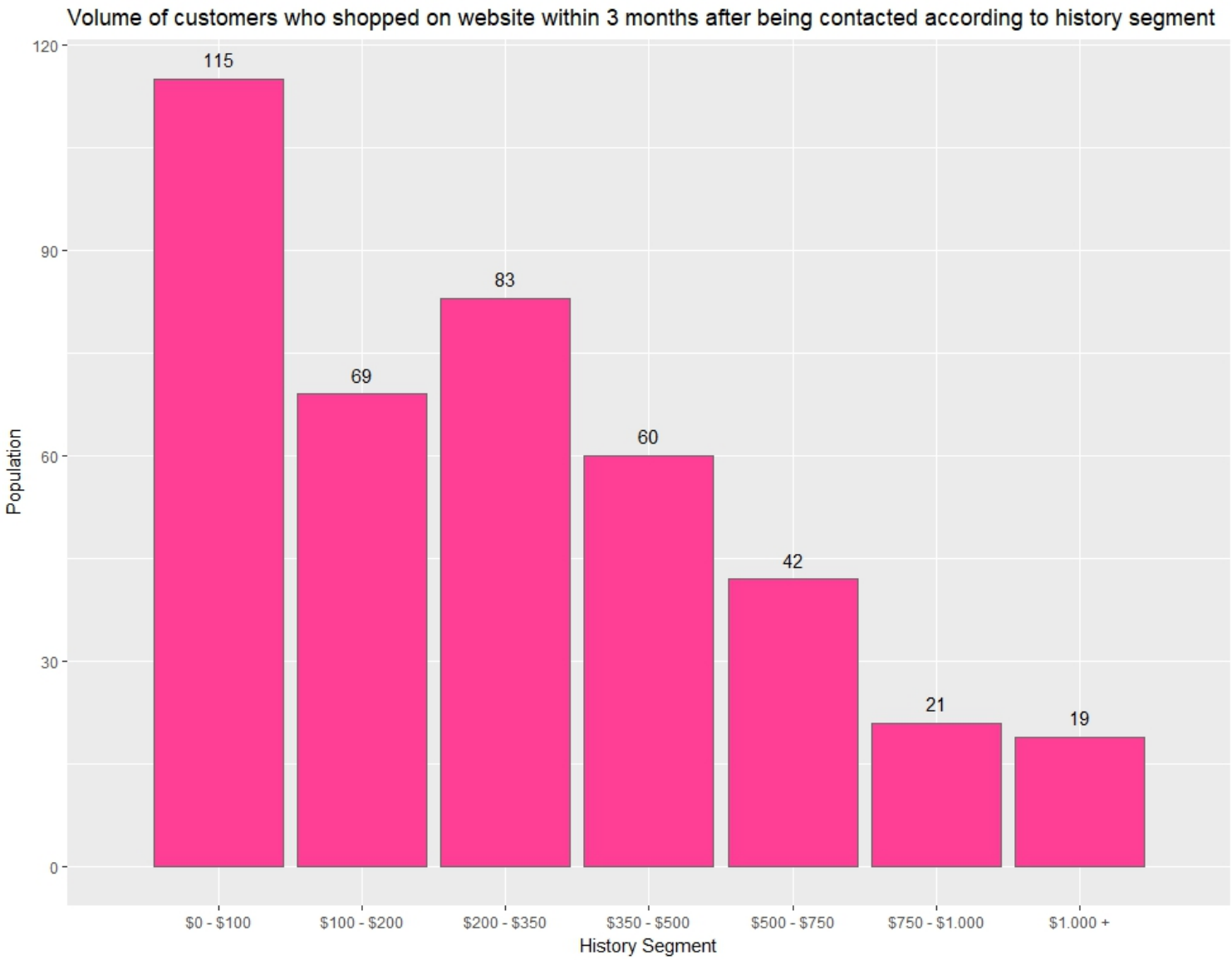

Figure 30: Population bar chart of Conversion against History Segment

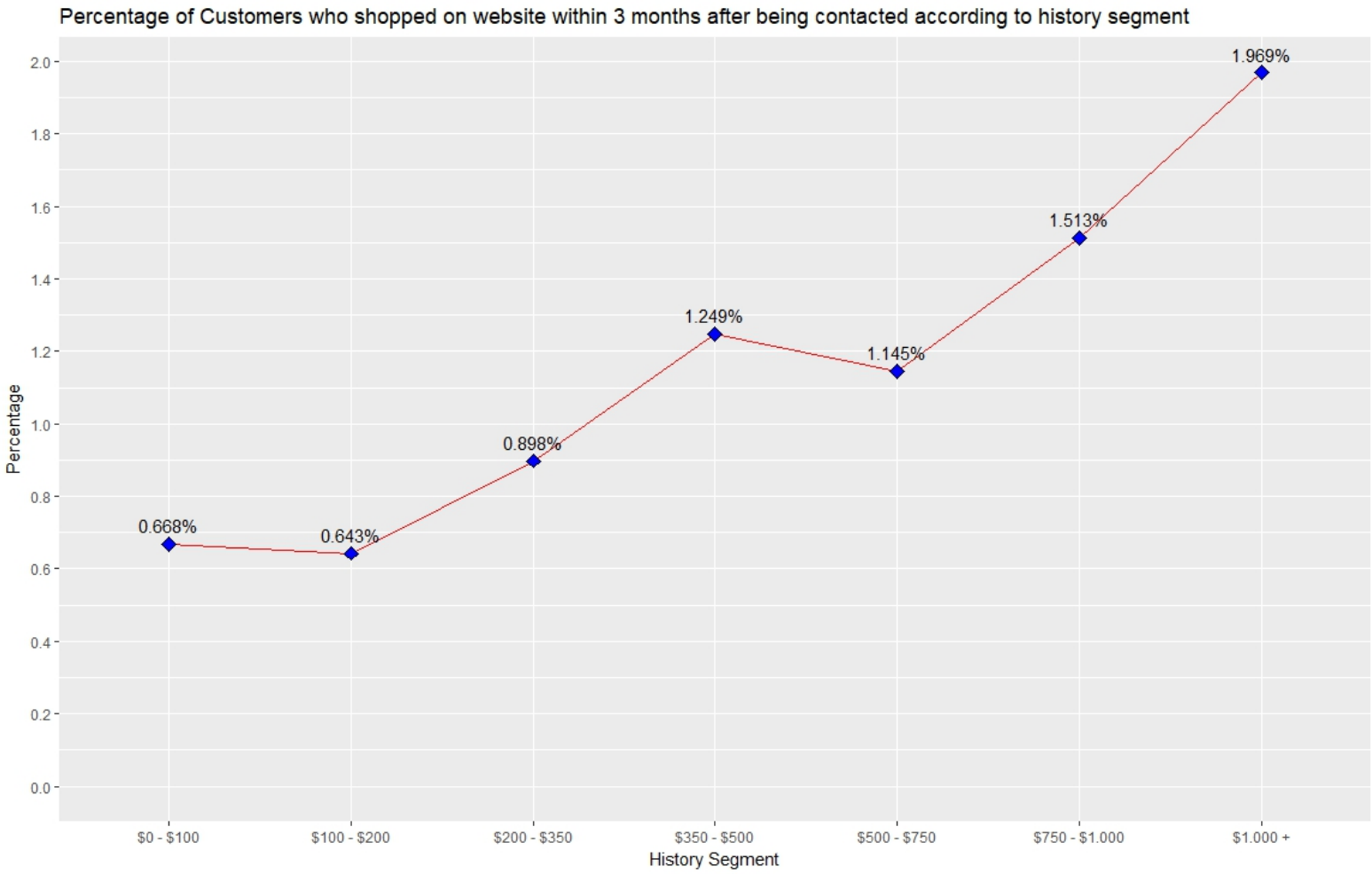

Figure 31: Percentage line chart of Conversion against History Segment

History Segment	$0~$100	$100~$200	$200~$350	$350~$500	$500~$750	$750~$1000	$1000+	Row Sum
Shopped	115	69	83	60	42	21	19	409
Not Shopped	17100	10658	9159	4743	3627	1367	946	47600
Col Sum	17215	10727	9242	4803	3669	1388	965	48009

Figure 32: Table of History Segment for Chi-Square testing

Chi-Square Result of History Segment:

```
    data:  hist_seg_chi_table
    X-squared = 46.791, df = 6, p-value = 2.06e-08
```

Since the p-value is smaller than 0.05, Ho is rejected. Therefore, history segment is not independent of conversion.

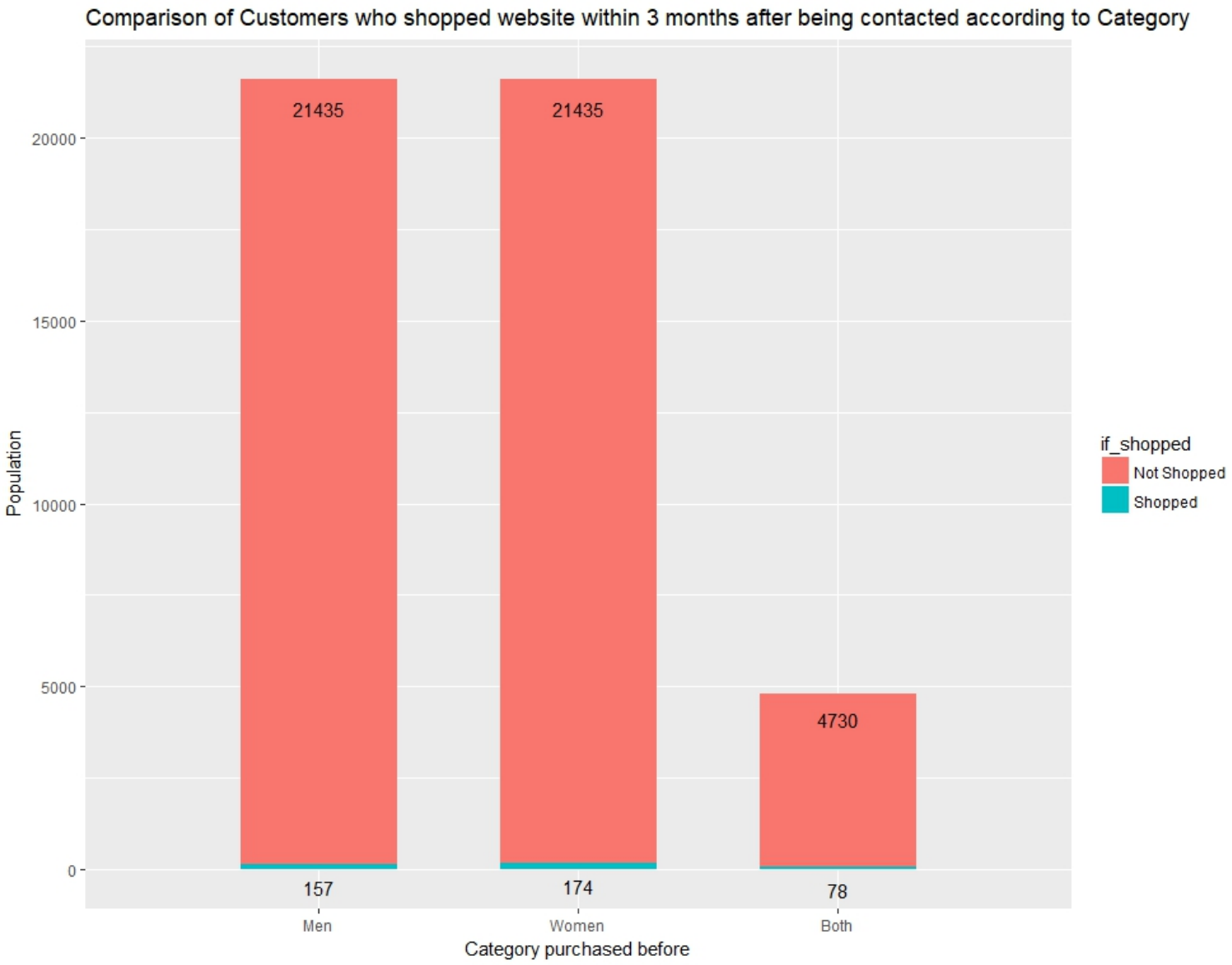

Figure 33: Population bar chart of Conversion against Category

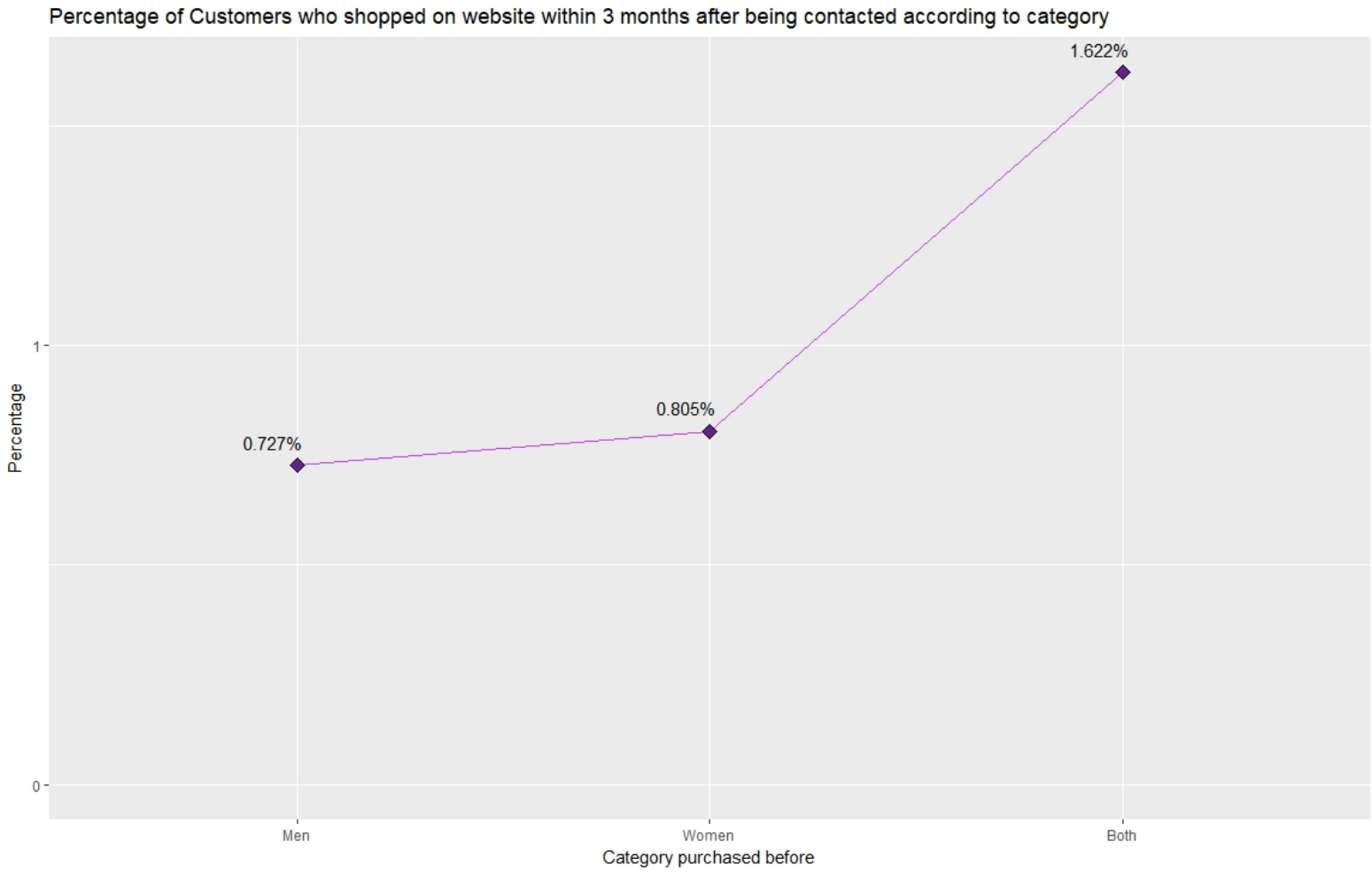

Figure 34: Percentage line chart of Conversion against Category

Category	Men	Women	Both	Row Sum
Shopped	157	174	78	409
Not Shopped	21435	21435	4730	47600
Col Sum	21592	21609	4808	48009

Figure 35: Table of Category for Chi-Square testing

Chi-Square Result of Category:

```
    data:  category_chi_table
    X-squared = 38.321, df = 2, p-value = 4.771e-09
```

Since the p-value is smaller than 0.05, Ho is rejected. Therefore, category is not independent with conversion.

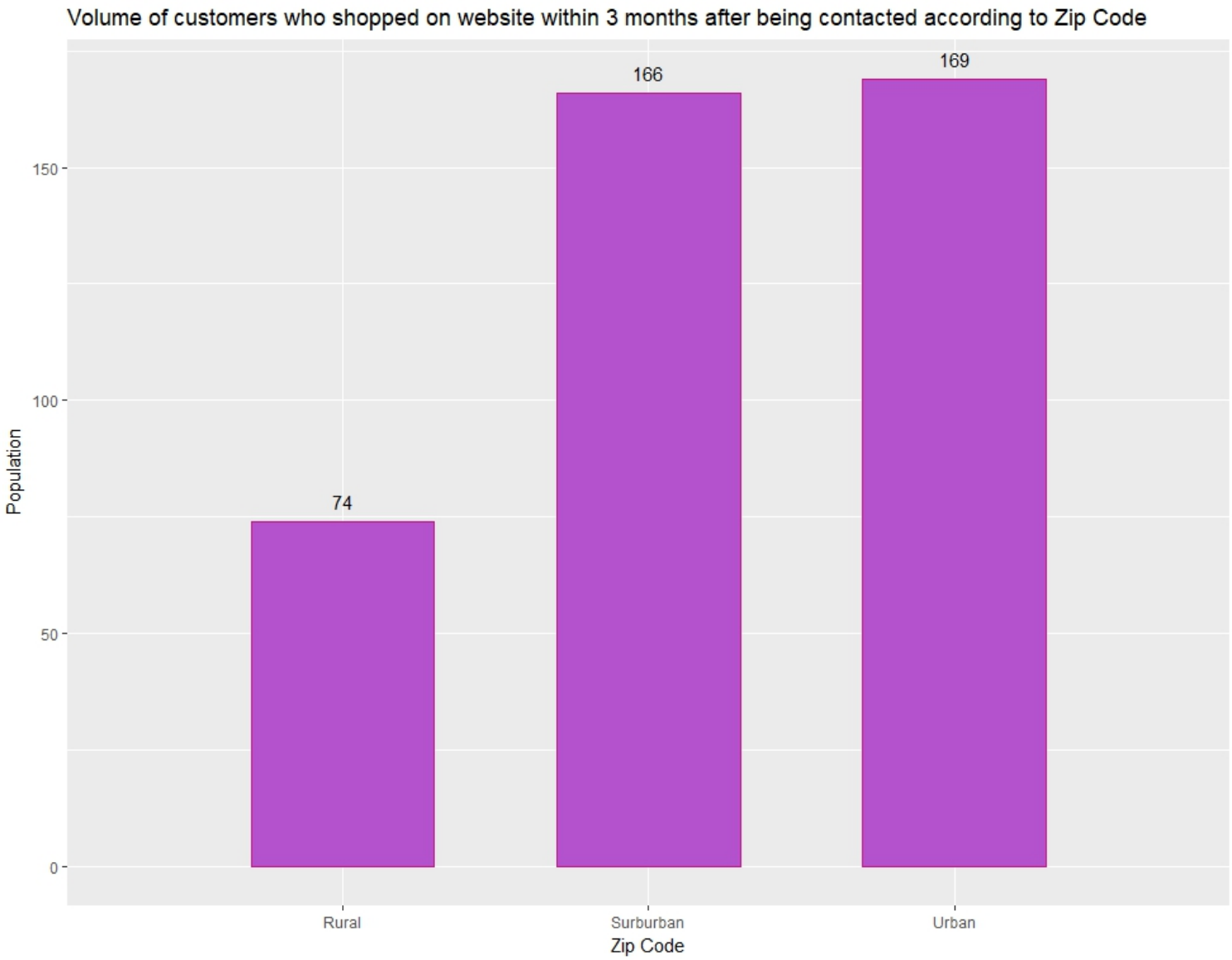

Figure 36: Population bar chart of Conversion against Zip Code

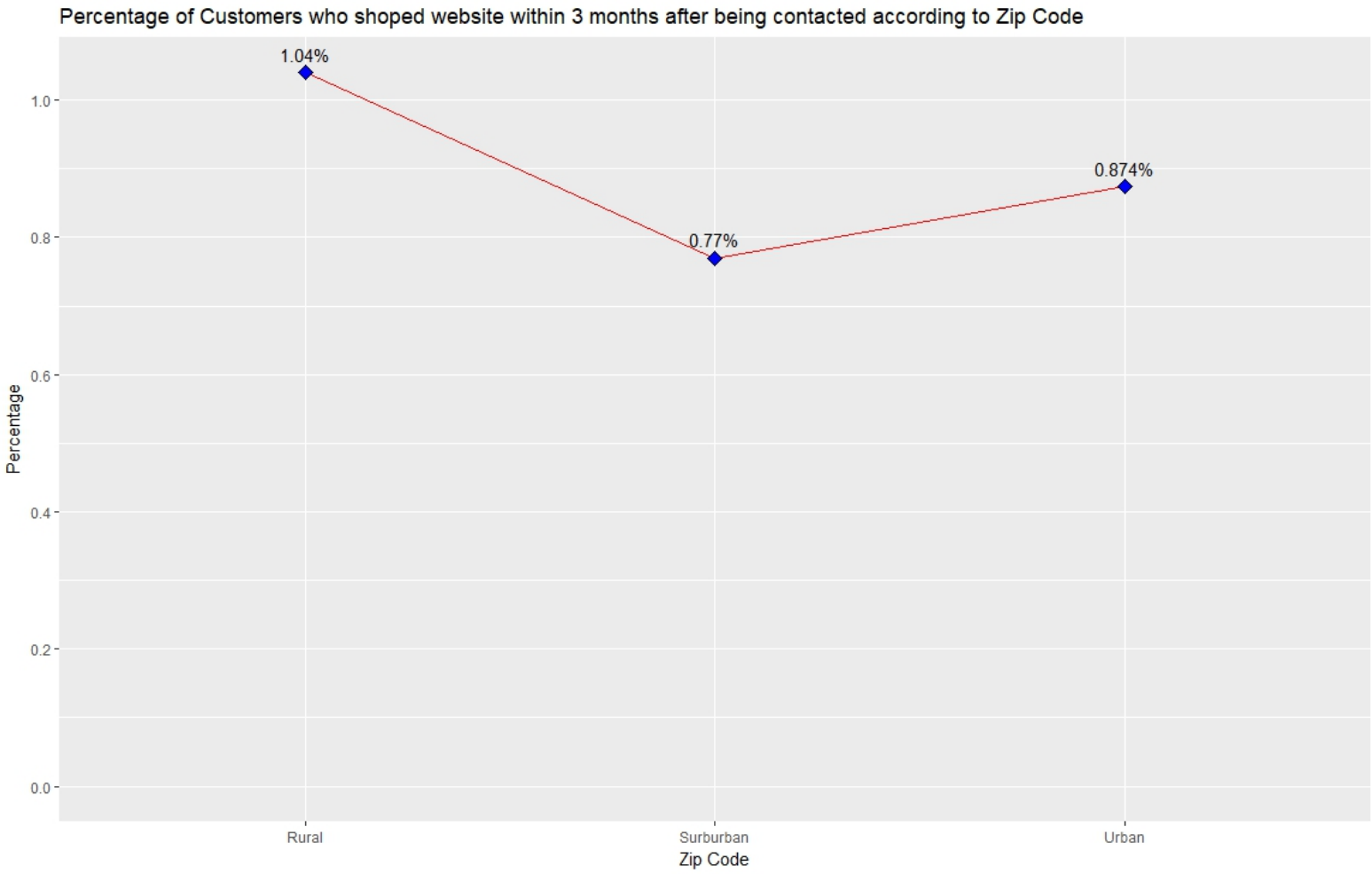

Figure 37: Percentage line chart of Conversion against Zip Code

Zip code	Rural	Surburban	Urban	Row Sum
Shopped	74	166	169	409
Not Shopped	7043	21387	19170	47600
Col Sum	7117	21553	19339	48009

Figure 38: Table of Zip Code for Chi-Square testing

Chi-Square Result of Zip Code:

```
        data:  zip_chi_table
        X-squared = 4.7878, df = 2, p-value = 0.09127
```

Since the p-value is greater than 0.05, Ho is not rejected. Therefore, zip code is independent of conversion.

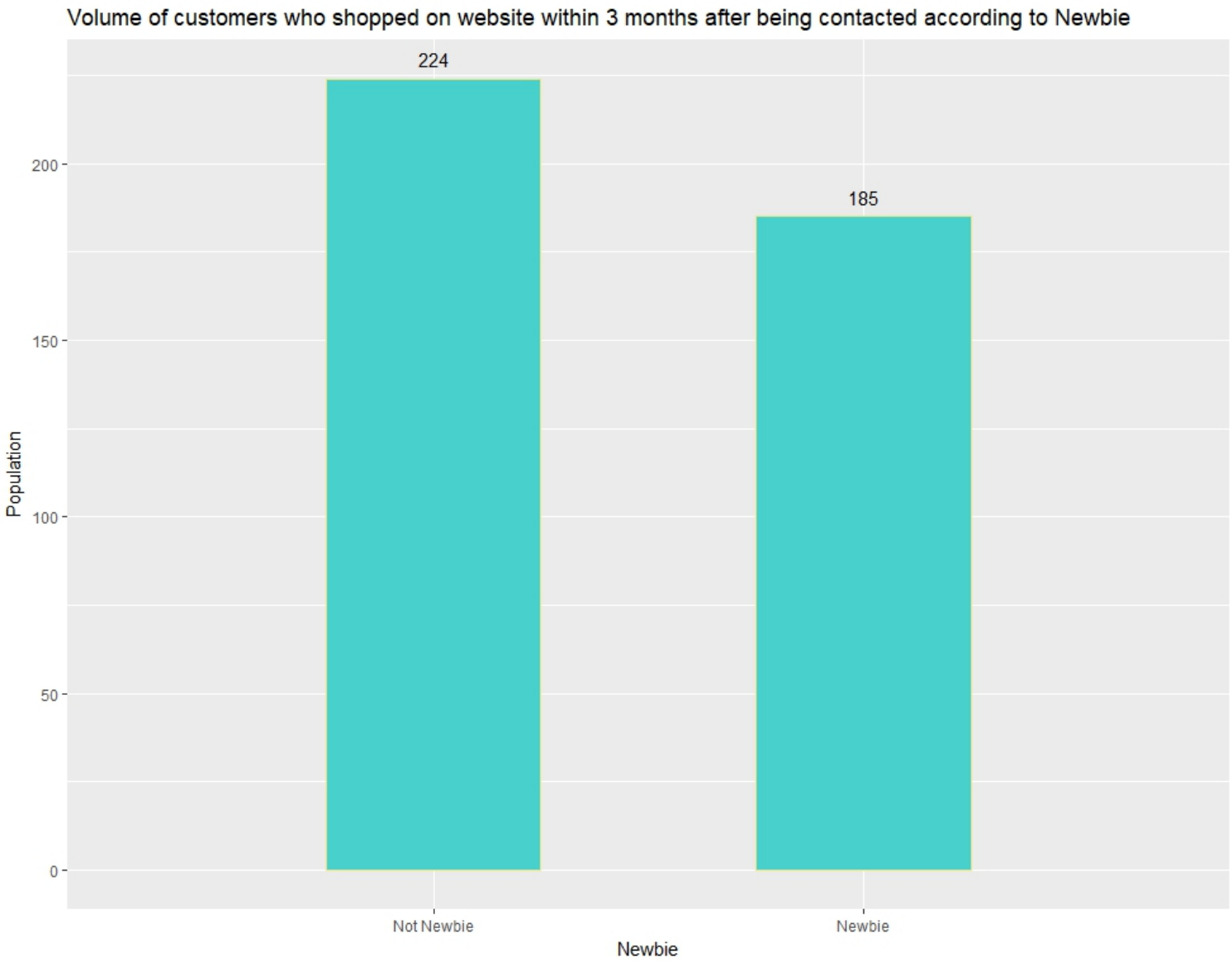

Figure 39: Population bar chart of Conversion against Newbie

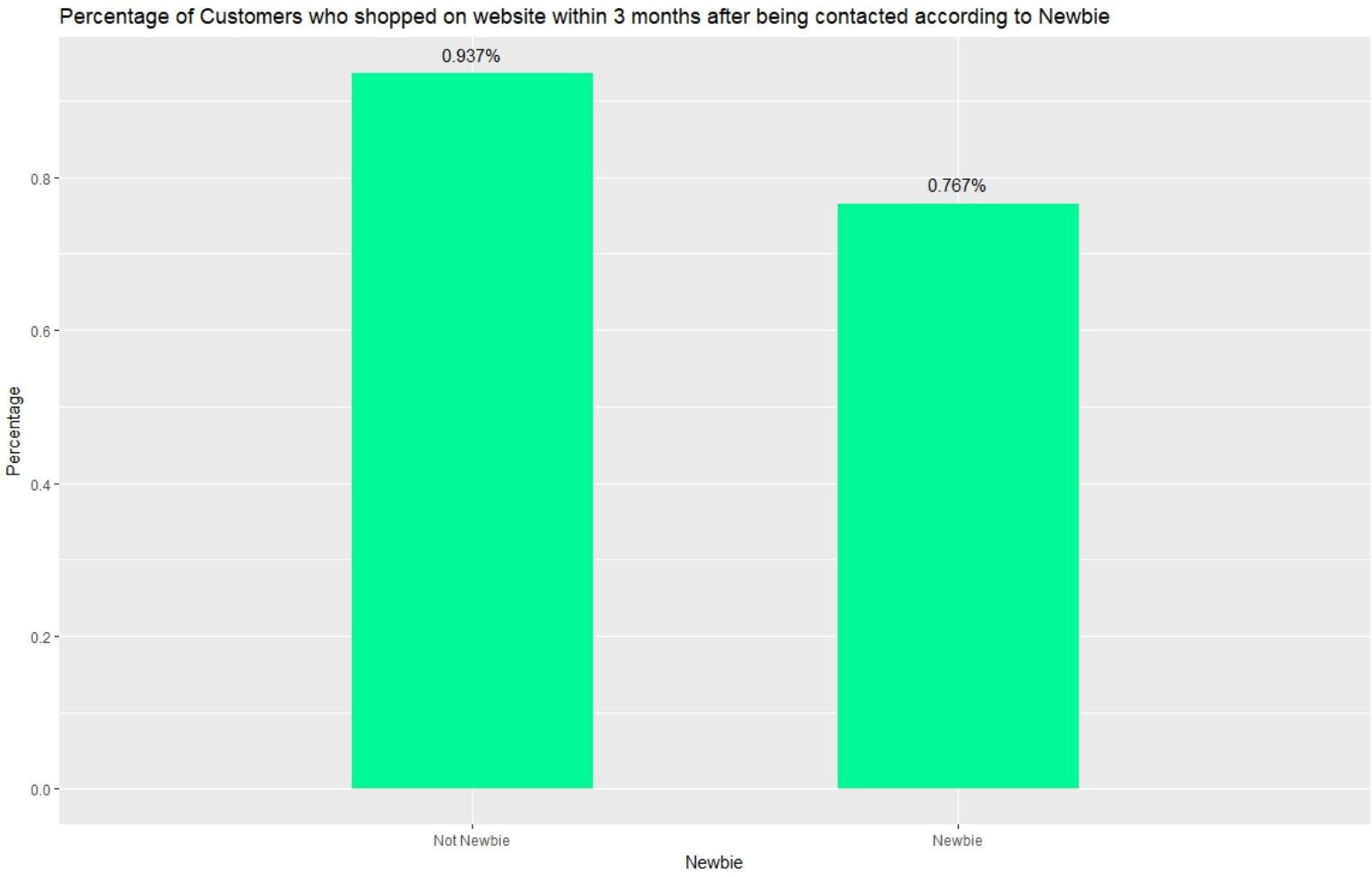

Figure 40: Percentage bar chart of Conversion against Newbie

Newbie	Newbie	Not Newbie	Row Sum
Shopped	224	185	409
Not Shopped	23677	23923	47600
Col Sum	23901	24108	48009

Figure 41: Table of Newbie for Chi-Square testing

Chi-Square Result of Newbie:

```
    data:  newbie_chi_table
    X-squared = 3.8991, df = 1, p-value = 0.04831
```

Since the p-value is smaller than 0.05, Ho is rejected. Therefore newbie is not independent of conversion.

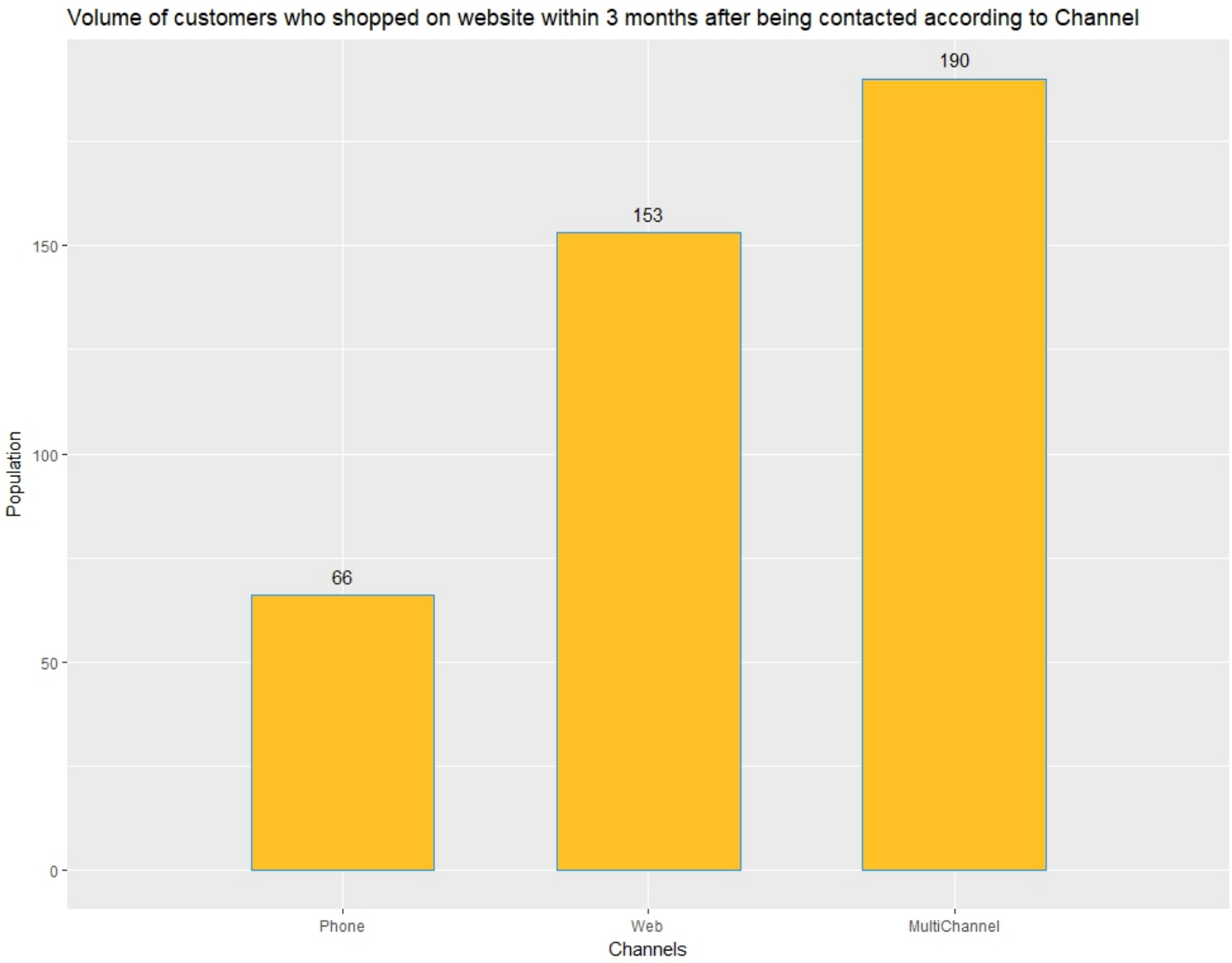

Figure 42: Population bar chart of Conversion against Channel

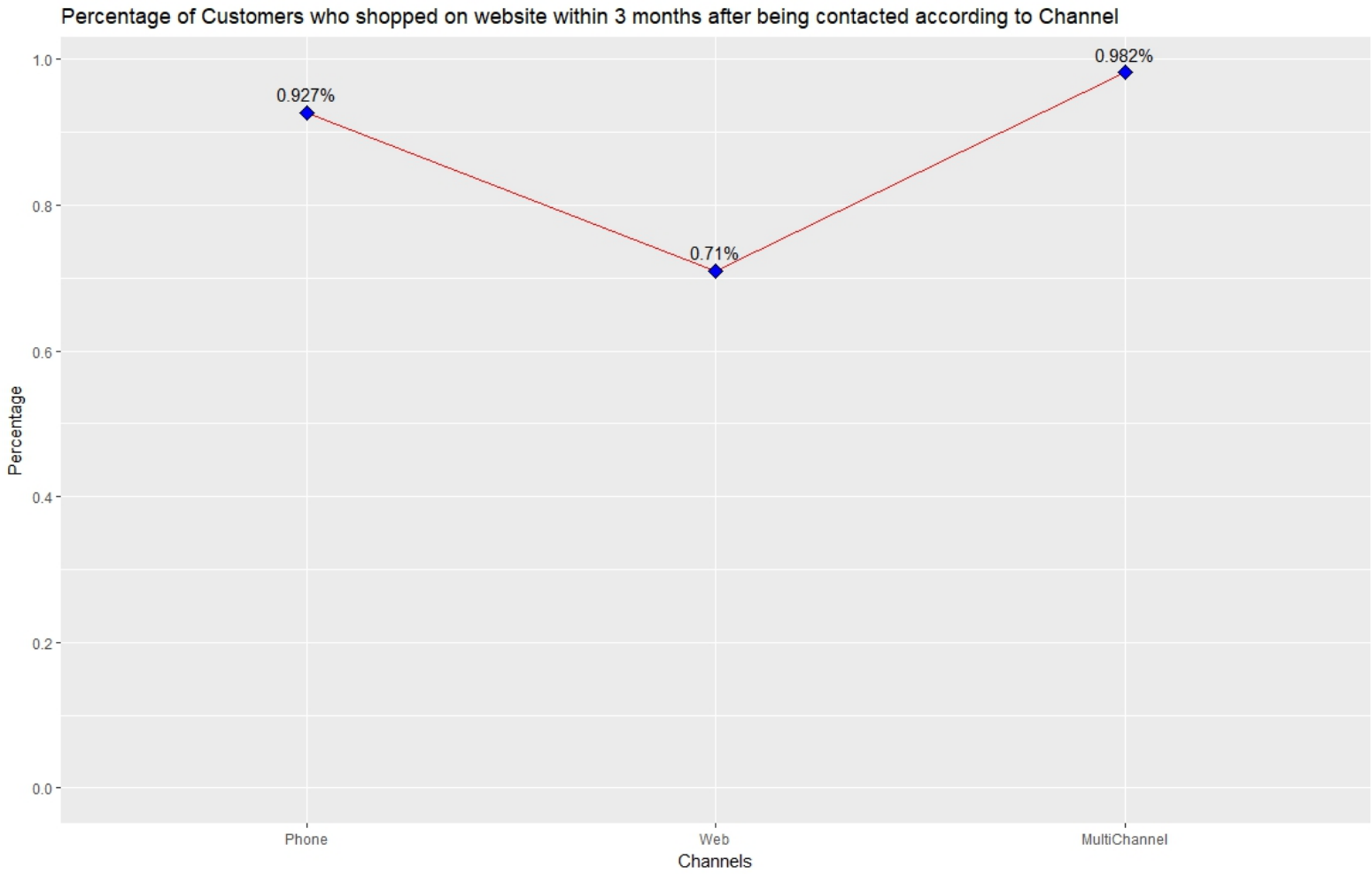

Figure 43: Percentage Line chart of Conversion against Channel

Channel	Multichannel	Phone	Web	Row Sum
Shopped	66	153	190	409
Not Shopped	5759	20888	20953	47600
Col Sum	5825	21041	21143	48009

Figure 44: Table of Channel for Chi-Square testing

Chi-Square Result of Channel:

```
    data:  channel_chi_table
        X-squared = 9.8745, df = 2, p-value = 0.007174
```

Since the p-value is smaller than 0.05, Ho is rejected. Therefore, zip code is not independent of conversion.

Question 4 Implementation and Result

In all predictive variables, only history is a numeric variable so that simple linear regression is proper for examining the relationship between history segment and spend. The process of constructing a linear regression between history and spend is described as below.

1. A shared function is created for extracting a table including only desired variables and spend when conversion == 1.

```
get_spend <- function(colnames){
    return(subset(direct_marketing, conversion == 1)
           [,append(colnames,"spend")])
}
```

For example:

```
hist_spend = get_spend(c("history"))
```

Result:

```
> head(hist_spend, 5)
     history  spend
204   297.80 264.66
340   101.99  29.99
485    61.08  70.85
586   194.81 184.71
649   287.06  29.99
```

2. Perform data correction according to the box-plot of history in question 2.

```
hist_spend <- subset(hist_spend,
                     history < 715.96 & history > 29.99)
```

3. Fit linear model and print out the summary.

```
lm_hist_seg <- lm(hist_spend$spend~1+hist_spend$history)
summary(lm_hist_seg)
```

Result:

```
Residuals:
   Min      1Q  Median      3Q     Max
```

```
-84.73  -77.12  -36.14   38.37  395.35

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       115.35747    9.84783  11.714   <2e-16 ***
hist_spend$history  -0.01872    0.03155  -0.593    0.553
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 100.5 on 329 degrees of freedom
Multiple R-squared:  0.001069,  Adjusted R-squared:  -0.001967
F-statistic: 0.3522 on 1 and 329 DF,  p-value: 0.5533
```

4. Draw scatterplot and linear regression for history and spend.

```
plot(hist_spend$history, hist_spend$spend)
abline(lm_hist_seg, col=2, lwd=3)
```

Result:

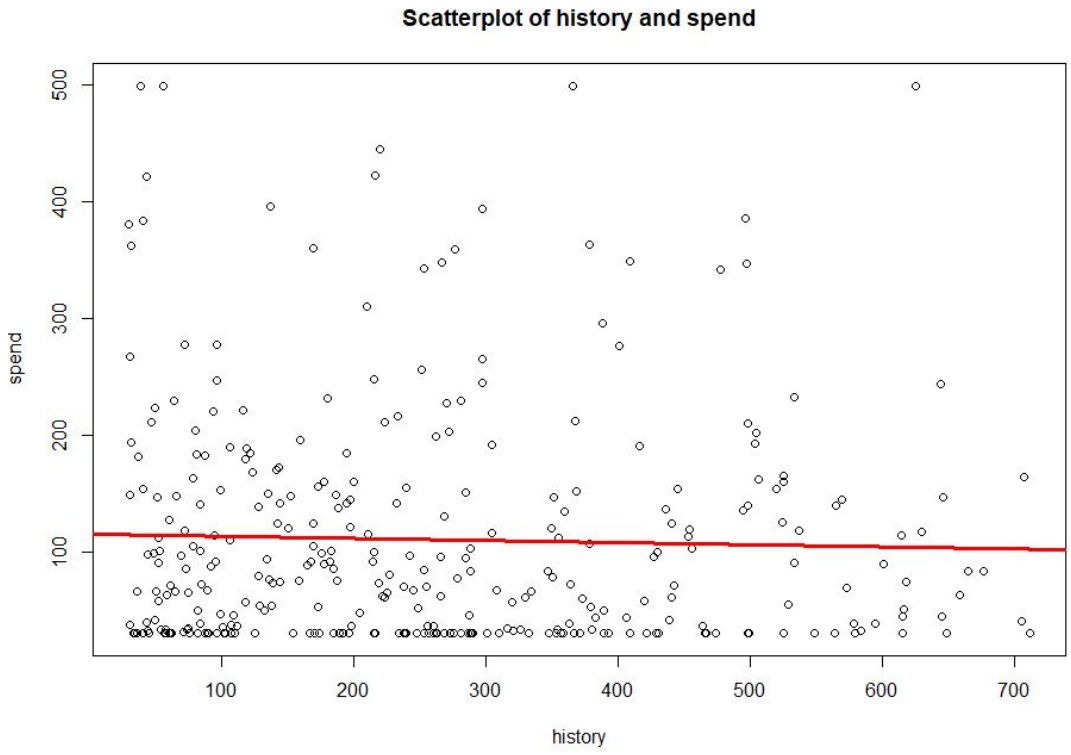

Figure 45: Scatterplot of history and spend

According to the linear regression summary, the p-values of t-test and F-statistic for history are too large and the R-squared value is way smaller than 1. Therefore, it is not a desirable linear regression model.

All other categorical variables will be combined to history for analysis. Let's take recency as an example.

1. Extract table for recency, history and spend

```
recency_hist_spend = get_spend(c("recency", "history"))
```

Result:

```
> head(recency_hist_spend, 5)
    recency history  spend
204       4   297.8 264.66
340       1   101.99  29.99
447       5 1079.62  29.99
485       1   61.08  70.85
586       8   194.81 184.71
```

2.

Conclusion and Recommendation

This section summaries findings and provides recommendations based on the result for question 2, 3, 4. The analysis part in question 3 is covered as well. For the clarity and clearness, conclusions and recommendations are introduced by each variable.

- Recency
  Admittedly, recency is related to both visit and conversion. According to Figure 13 and Figure 28, it can be roughed concluded that customers who made last purchase more recently, the more they visited and shopped on the website after being contacted. The effectiveness of this marketing campaign is more significant to those who shopped recently.

- History Segment
  Apparently, according to Figure 16 and Figure 23, customers who have history segment value are more willing to visit the website and shop online after being contact. The result is very obvious in Figure 23. Special marketing campaigns may be organized to target customers who have high history segment value.

- History

- Category According to Figure 19 and Figure 34, it is clear that customers who purchased both men and women category items made the most significant contribution in visiting website and shopping online after being contacted. Customers who only shopped women's item have higher value in both visit and conversion than those who only shopped men's category.

- Zip Code According to Figure 22 and Figure 37, it is notable that customers who live in rural area responded stronger in this marketing campaign than those who are from suburban and urban area.

- Newbie According to Figure 25 and Figure 40, old customers had a better reaction in this marketing campaign than new customers.

- Channel According to Figure 9 and Figure 43, customers shopped via multichannel and web are more willing to visit website and shop online than those who purchased by phone after the marketing campaign

References

[1] Box Plot:Display of Distribution. http://www.physics.csbsju.edu/stats/box2.html. Accessed: 2017-12-01.

[2] Chi-Squre test. https://en.wikipedia.org/wiki/Chi-squared_test. Accessed: 2017-12-01.

[3] ggplot2 Official website. http://ggplot2.org/. Accessed: 2017-12-01.

[4] Jeffrey B. Arnold. ggthemes v3.4.0 R Documentation. https://www.rdocumentation.org/packages/ggthemes/versions/3.4.0. Accessed: 2017-12-01.

[5] Hadley Wickham. ddply R Documentation. https://www.rdocumentation.org/packages/plyr/versions/1.8.4/topics/ddply. Accessed: 2017-12-01.

[6] Hadley Wickham. plyr v1.8.4 R Documentation. https://www.rdocumentation.org/packages/plyr/versions/1.8.4. Accessed: 2017-12-01.