INSTITUTE OF TRANSPORTATION - TECHNISCHE UNIVERSITÄT MÜNCHEN

M.Sc. in Transportation Systems

Applied Statistics in Transport

Project Report

18th of November, 2012

Author
Chaklader Asfak Arefe

Abstract

The focus of this project report is to grasp the quality to handle large data set for performing statistical analysis and draw a conclusion for the attained question. It doesn't explore any relevant literature as suggested and hence, responsibility and draw backs of the findings solely go to the author. The report encompasses an effort to relate between different economic stratified groups with their usage of non-motorized vehicles (here, cycle). I hope the visualizations and test results will help to understand proper scenario in this particular context. Statistical software R is used for performing the analysis and building the visualizations.

Table of Contents

Introduction

Mobility links different purposes of our daily life and it's a large part our daily time expenditure. It's the back bone of modern economies and societies. In this paper, the data set given to us is quite large (about twenty six thousand households) and there are too many variables, thus giving us the ability to handle huge amounts of information and analyze the traveling patterns which depends on a number of different factors like education, economic conditions, number of people in a household, which part of the country people are living. This paper encompasses an effort to understand the relation of economical stratification with the usage of non-motorized vehicles. It's assumed that cycle is only used by the people who stay middle or bottom of income group levels as high income level people goes to personal motorized vehicles for more comfort and uses lesser public transport. The effort includes statistical analysis and derives a conclusion from them.

1.0 Methodology

1.1 Cycle ownership in Germany

This small research is mainly focused on to analyze peoples spending psychology for non-motorized vehicles. People may think to spend more in present without considering much about their future or May chose reasonable options in context of their financial condition disregarding comfort. So, it come's as important question as what portion of their income they use for travel and it will be nice to know about travelling by cycle in particular for economic stratified groups. The cycle comes with many advantages such as reduce emission of CO_2, reduce Brown smog, reduce traffic on the main traffic streams, improve physical fitness of users etc. Hence, this investigation comes in interest whether there is some relation exists between income levels and cycle usage.

1.2 Method:

For performing this analysis, we separated the data into different parts depending on the household income level and number of cycles per household. We have fifteen different level of income starting from five hundred euro to more than seven thousand euro. First we remove all the false (N.A) values which are not required for our procedure and then to count the number of cycles for particular household's income level for the whole dataset. The amount of counted cycles and households income levels were written into the data variables for the further use for t test and Graphs. Afterwards we use these data variables to plot main graphs and sub graphs per income level. Finally we do some T-Tests with the explanation about results.

1.3 Descriptive Analysis:

Descriptive statistics is the discipline of quantitatively describing the main features of a collection of data. For the convenience of our analysis, here is an effort to visualize number of cycles that belong to different income groups. It considered fifteen different income groups including people whose income is lesser than five hundred euro to more than seven thousand euro.

(a) Income group less than five hundred euro:

Figure (a) is expressing number of cycles which are used against their corresponding number of households. We can able to see that a large number of household for this particular income group (more than eighty) don't belong to any cycle and about one hundred household have at least one cycle.

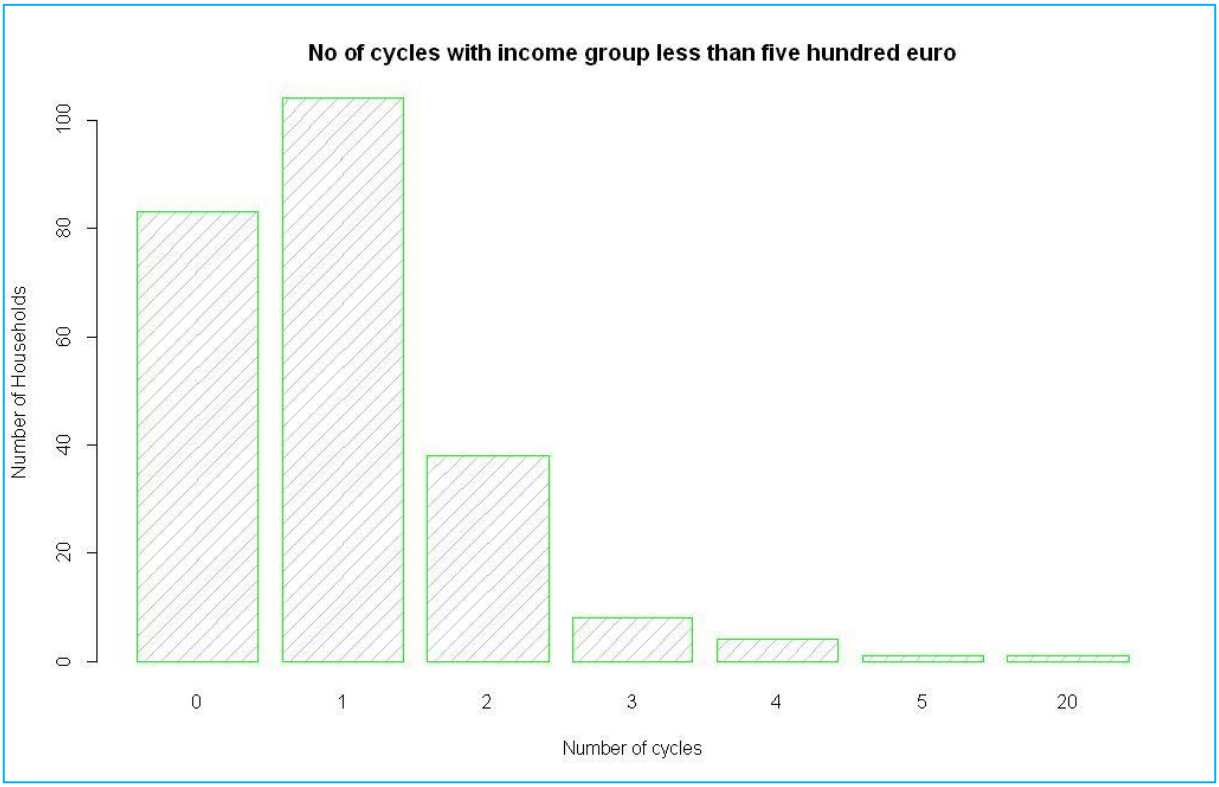


figure:a

(b) Income group less than nine hundred euro:

Figure (b) represents that most of the household have one or two cycles. However, we also can able to observe that around three hundred household have no cycle.

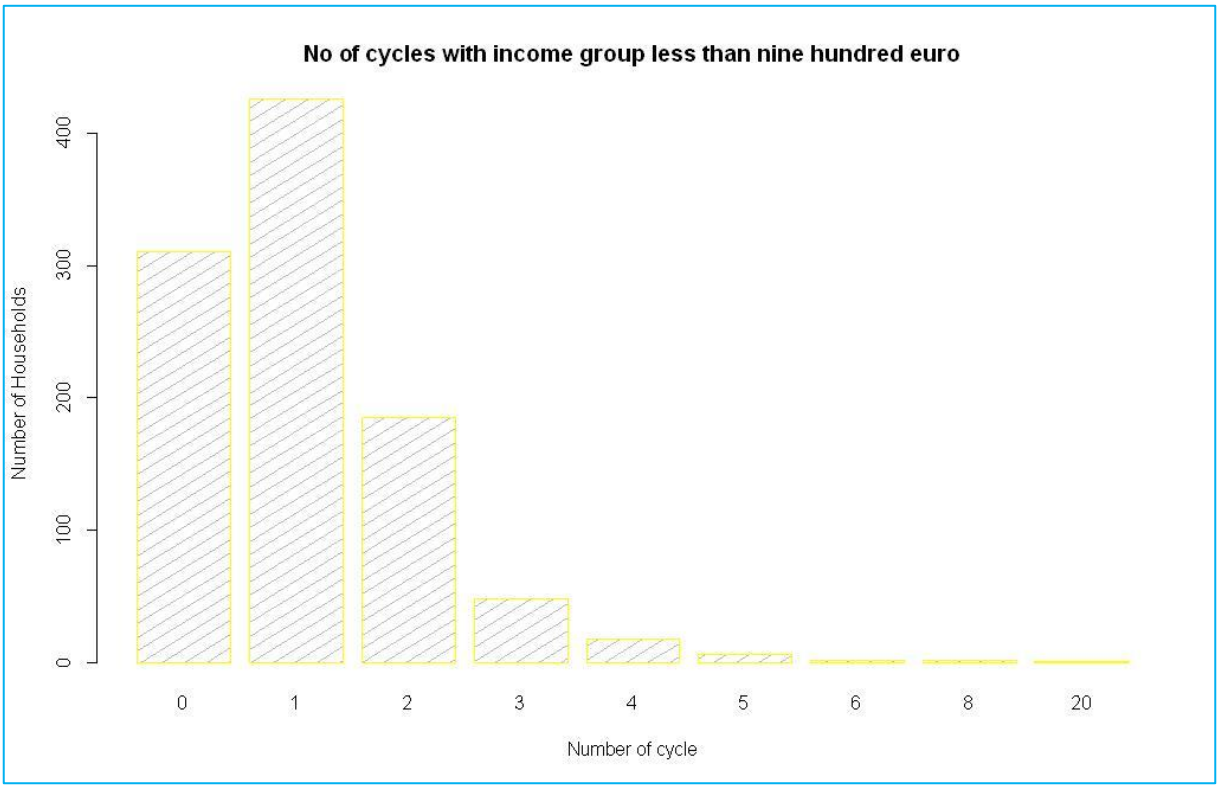
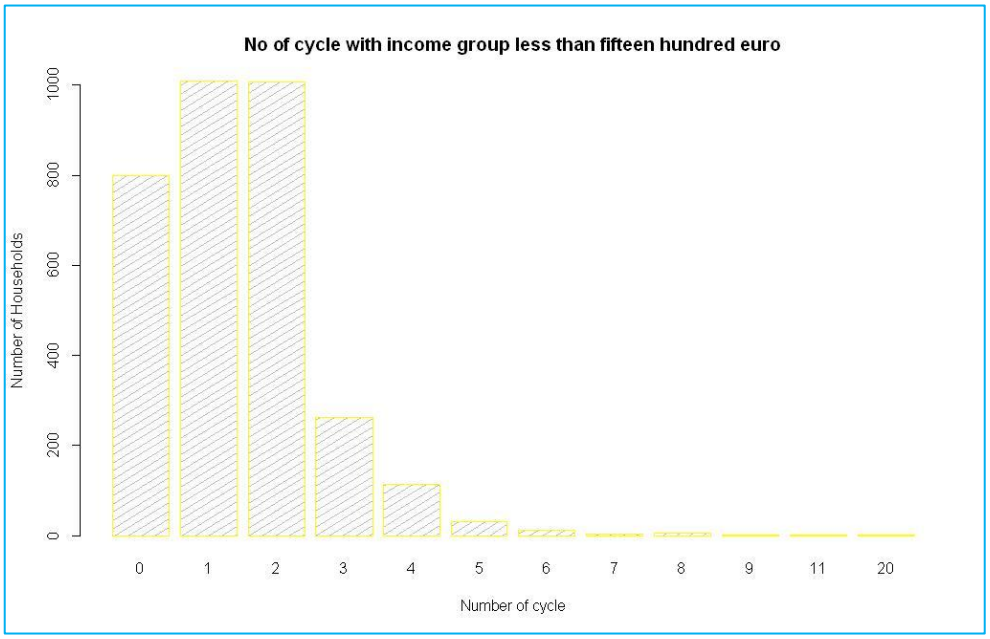


Figure: (b)

(c) Income group less than fifteen hundred euro:

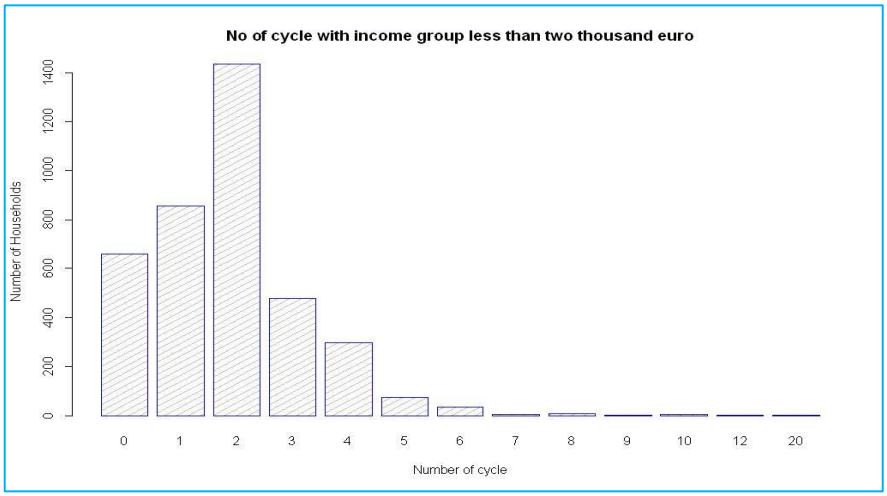
Figure (c) tells us that more than one thousand household have either one or two cycle. Interesting is that, few household have as much as twenty cycles as well.



figure(c)

(d)Income group less than two thousand euro:

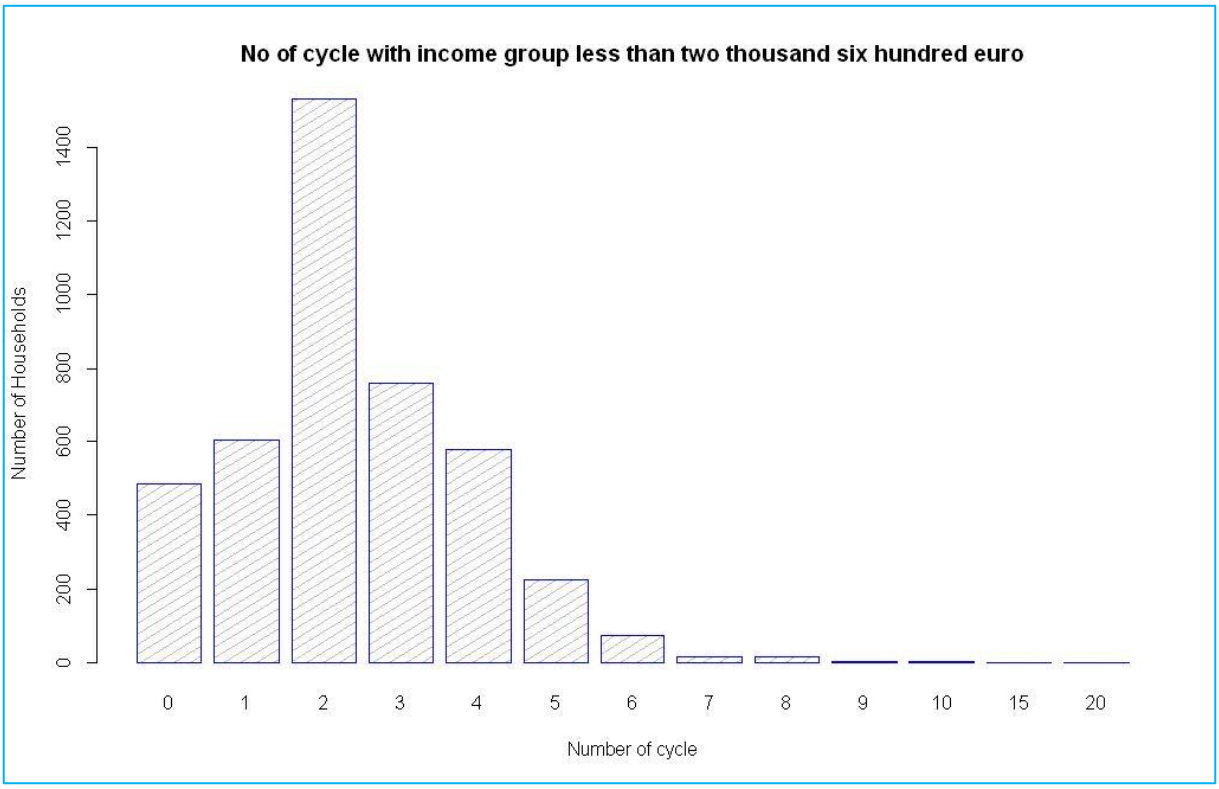
Most household have two cycles and more than six hundred household have no cycle. Few household shares as much as twenty cycles.



Figure(d)

(e)Income group less than two thousand six hundred euro:

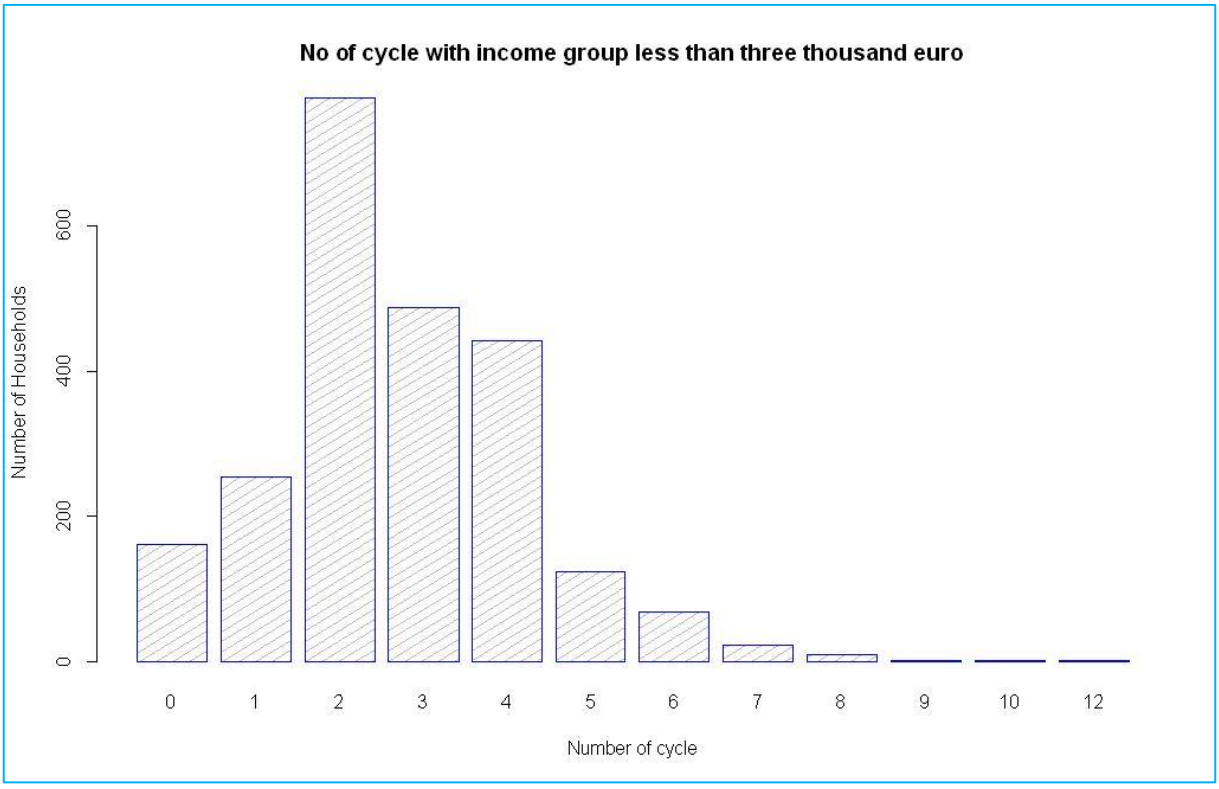
Here as well, most household have two cycles and fewer household have no cycle compare to income group lesser than two thousand euro.



figure(e)

(f)Income group less than three thousand euro:

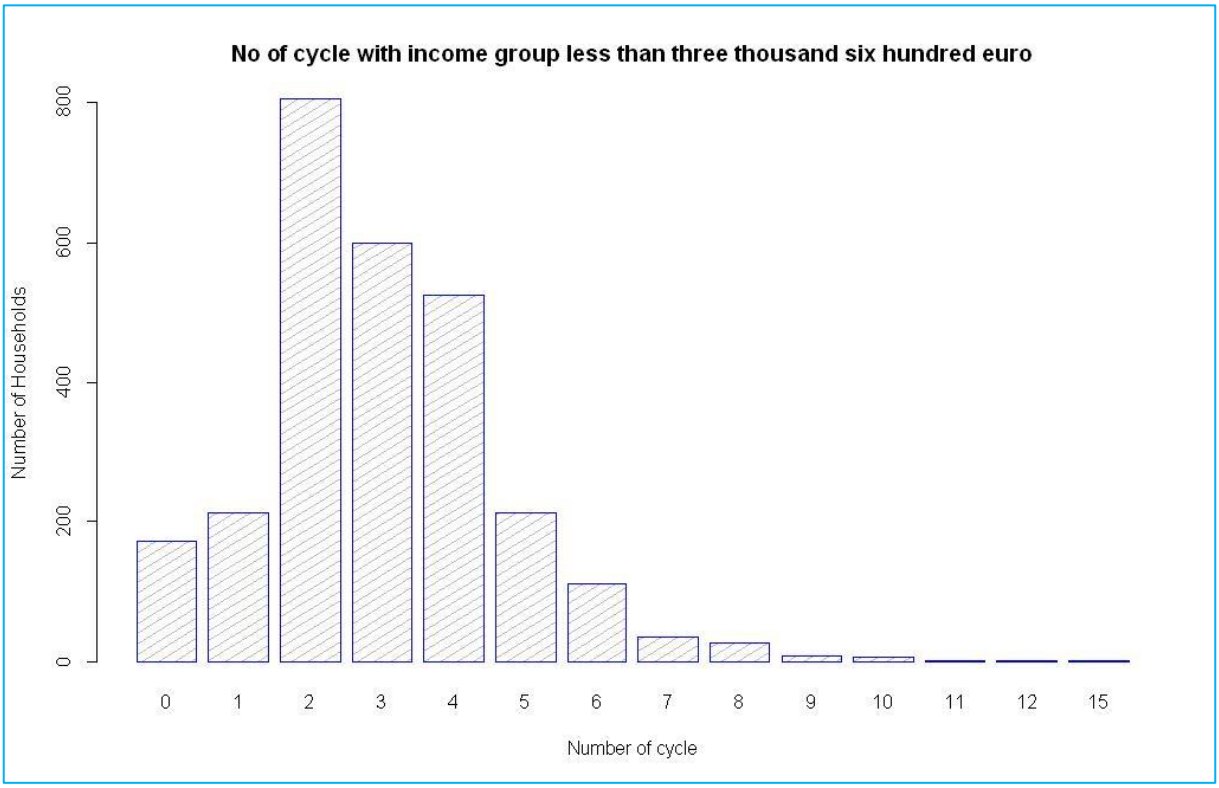
From figure (f) we can see that, most household have at least two cycles and even fewer household s are without a cycle. Maximum number of cycle that belongs to a household is twelve.



figure(f)

(g)Income group less than three thousand six hundred euro:

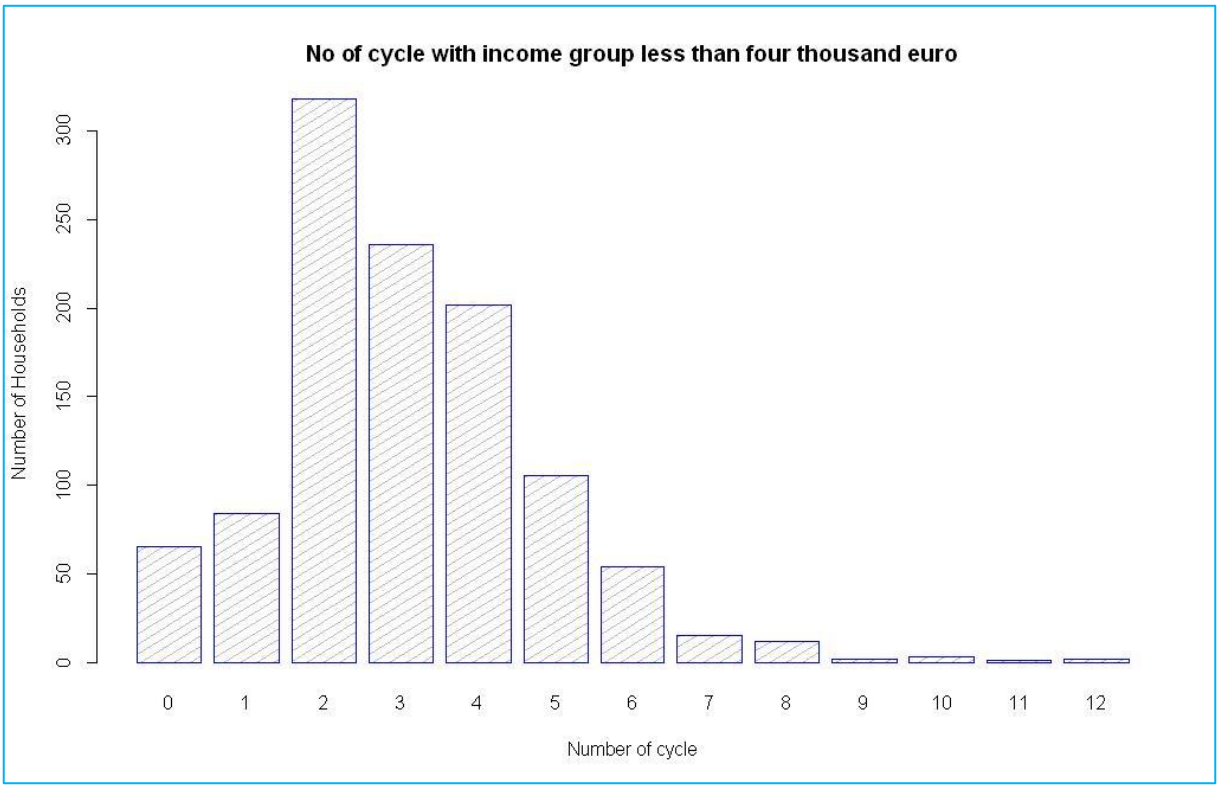
Most of the household have two, three or four cycles and maximum number of cycle that belongs to a single household is fifteen.



Figure(g)

(h)Income group less than four thousand euro:

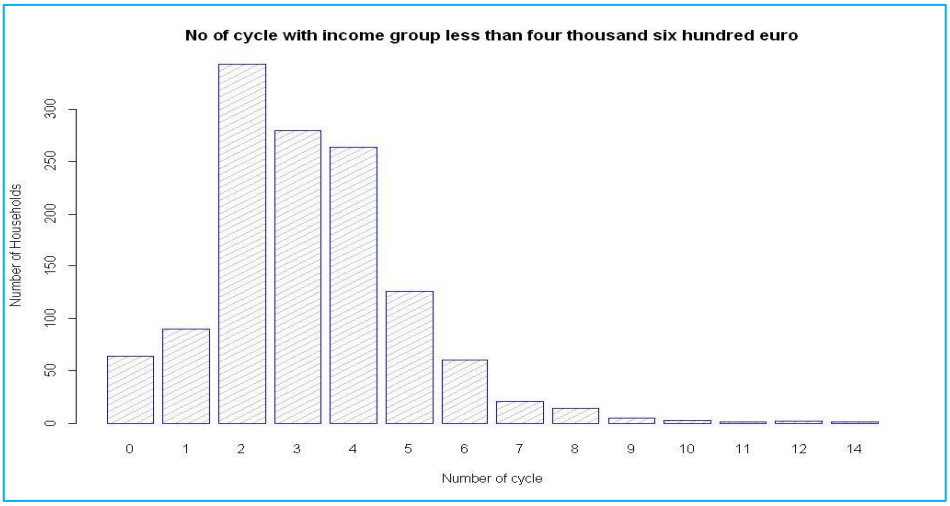
Most of the household have two, three or four cycles and maximum number of cycle that belongs to a single household is twelve. Few household have no cycle (say, more than fifty).



figure(h)

(i)Income group less than four thousand six hundred euro:

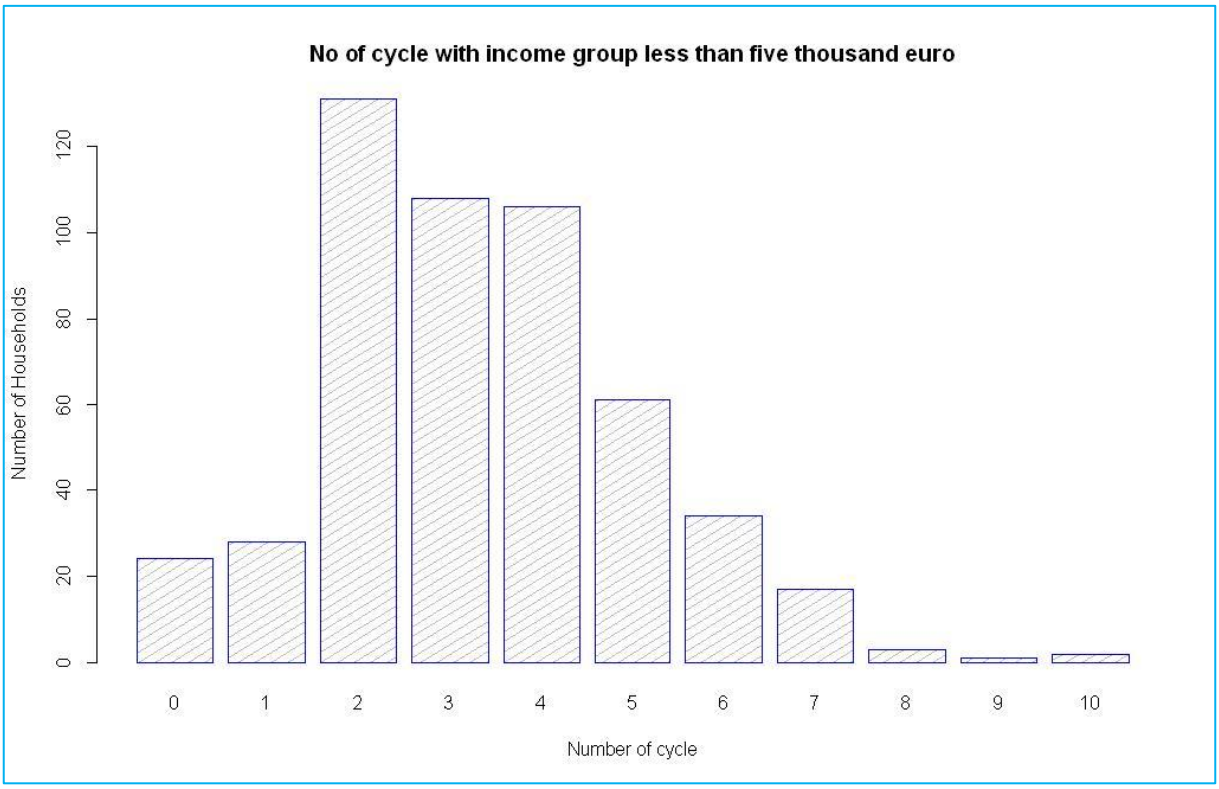
From figure(h), more than two fifty household have either one, two, three or four cycles and maximum number of cycle that belongs to a single household is fourteen.



figure(i)

(j)Income group less than five thousand euro:

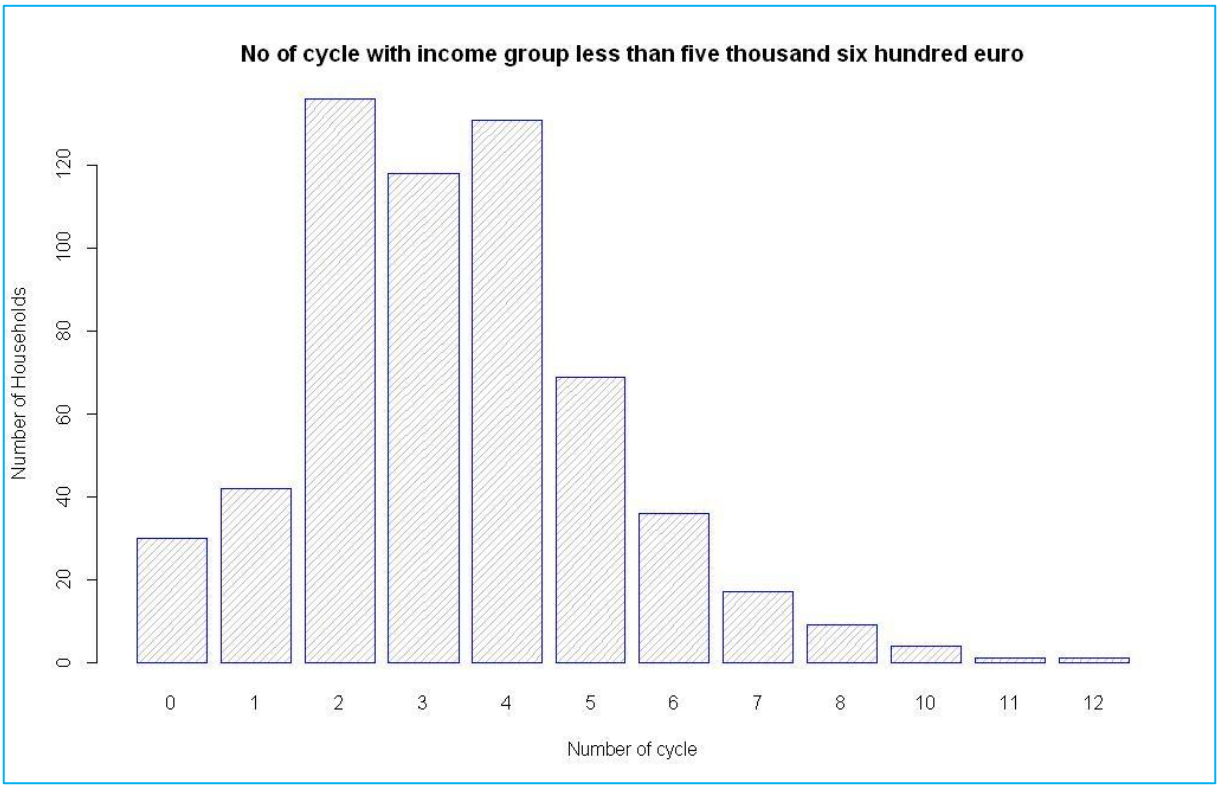
More than one hounded households have at least one, two or three cycles and around sixty household have five cycles. Maximum number of cycle that belongs to a single household is ten.



Figure(j)

(k)Income group less than five thousand six hundred euro:

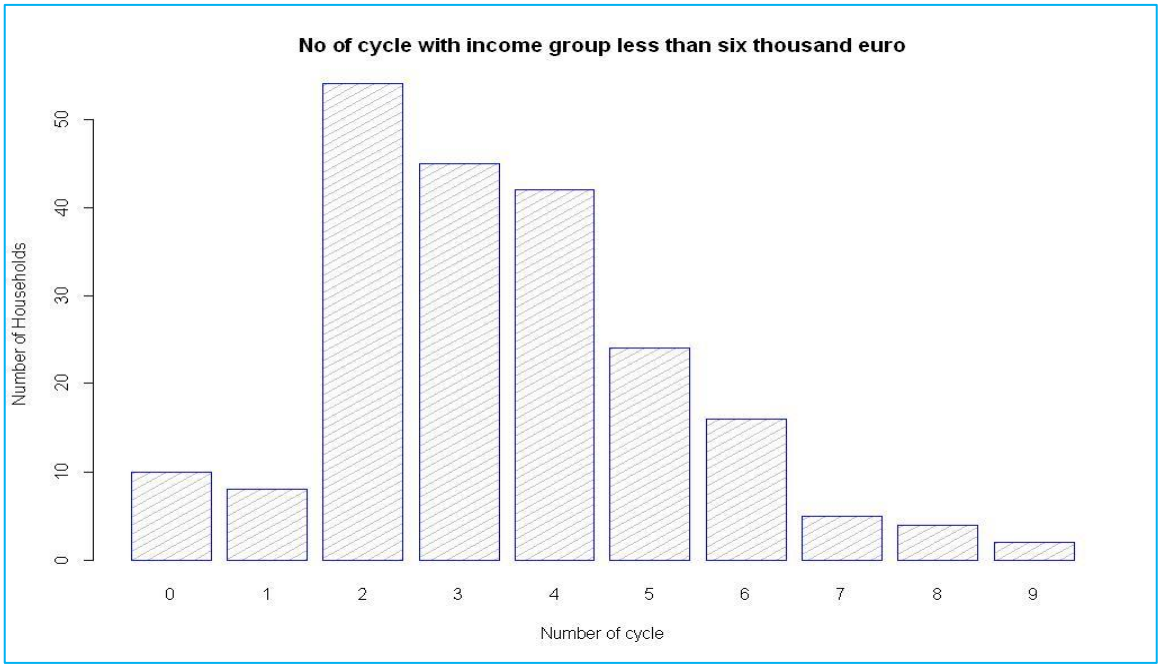
Around one hundred household have at least two, three or four cycles. Maximum number of cycle that belongs to a single household is twelve.



Figure(k)

(l)Income group less than six thousand euro:

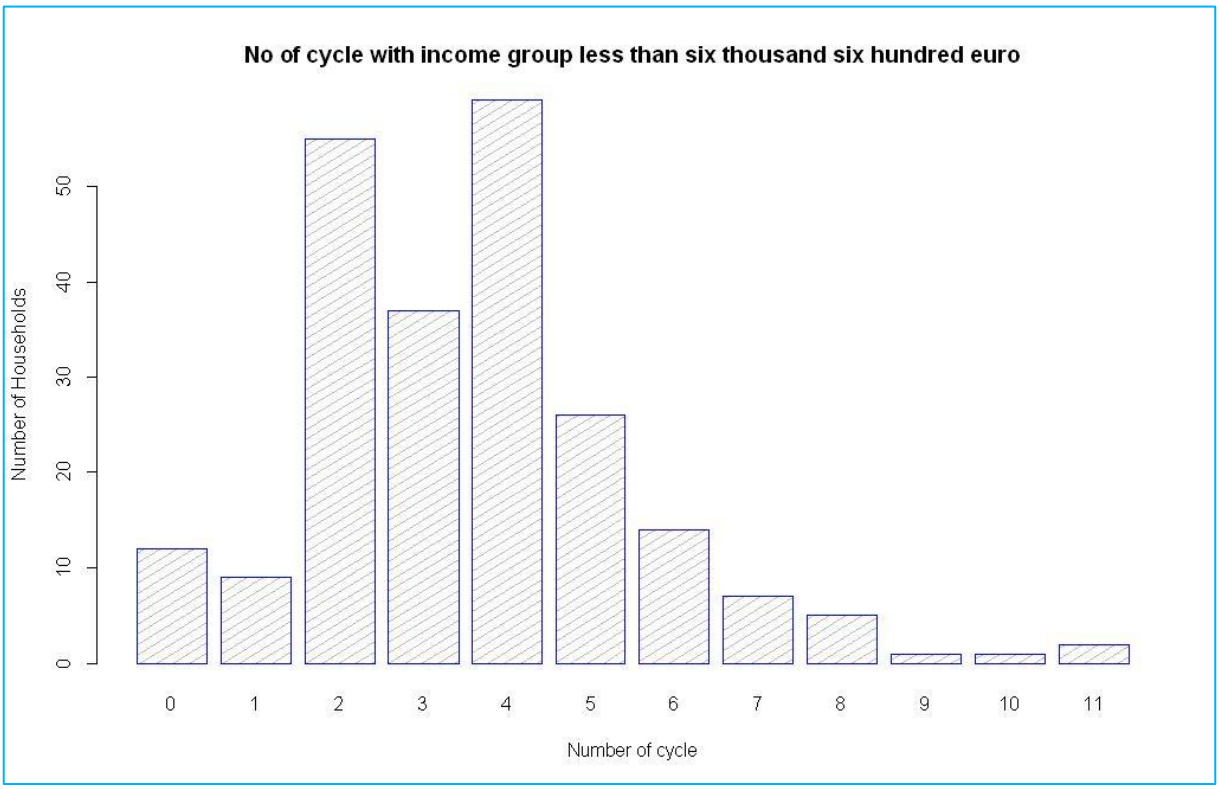
Most household have two cycles and about forty household have three or four cycles. Maximum number of cycle that belongs to a single household is nine.



figure(l)

(m)Income group less than four thousand six hundred euro:

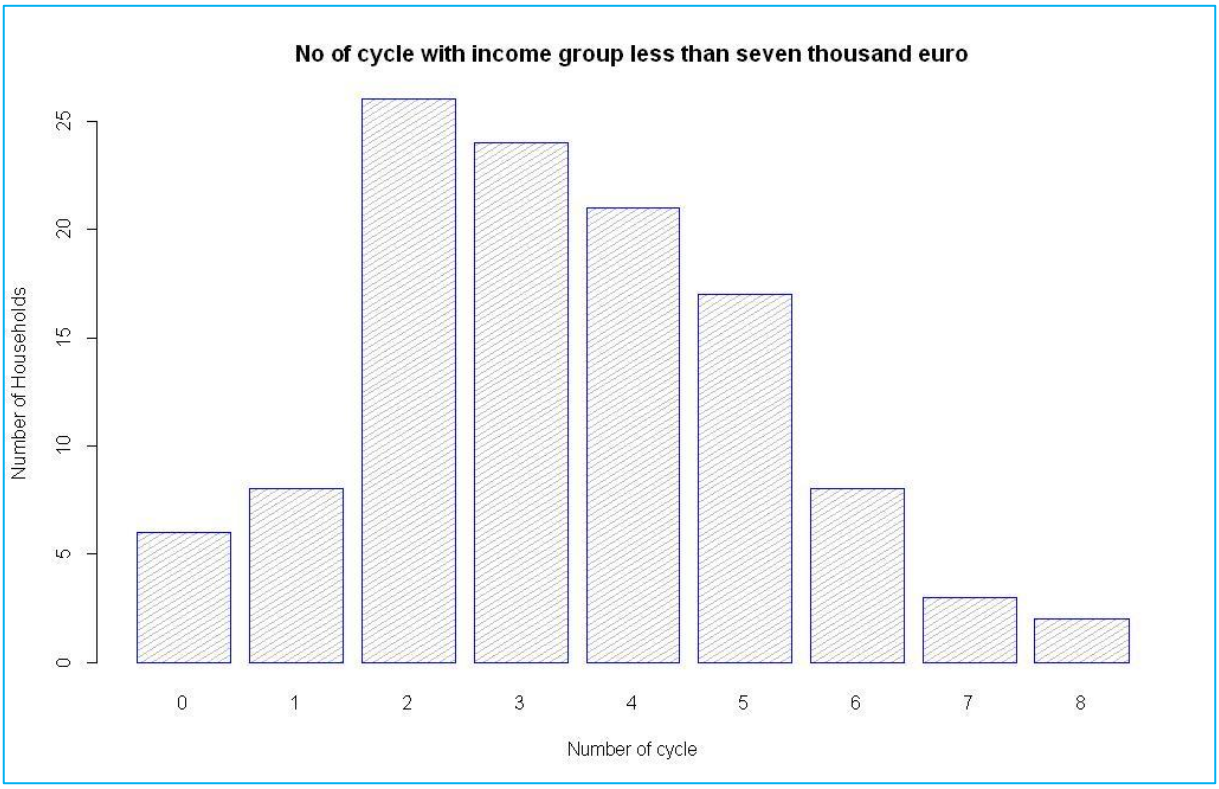
Most household have either two or four cycles and around thirty five household have three cycles. Maximum number of cycle that belongs to a single household is eleven.



figure(m)

(n)Income group less than four thousand six hundred euro:

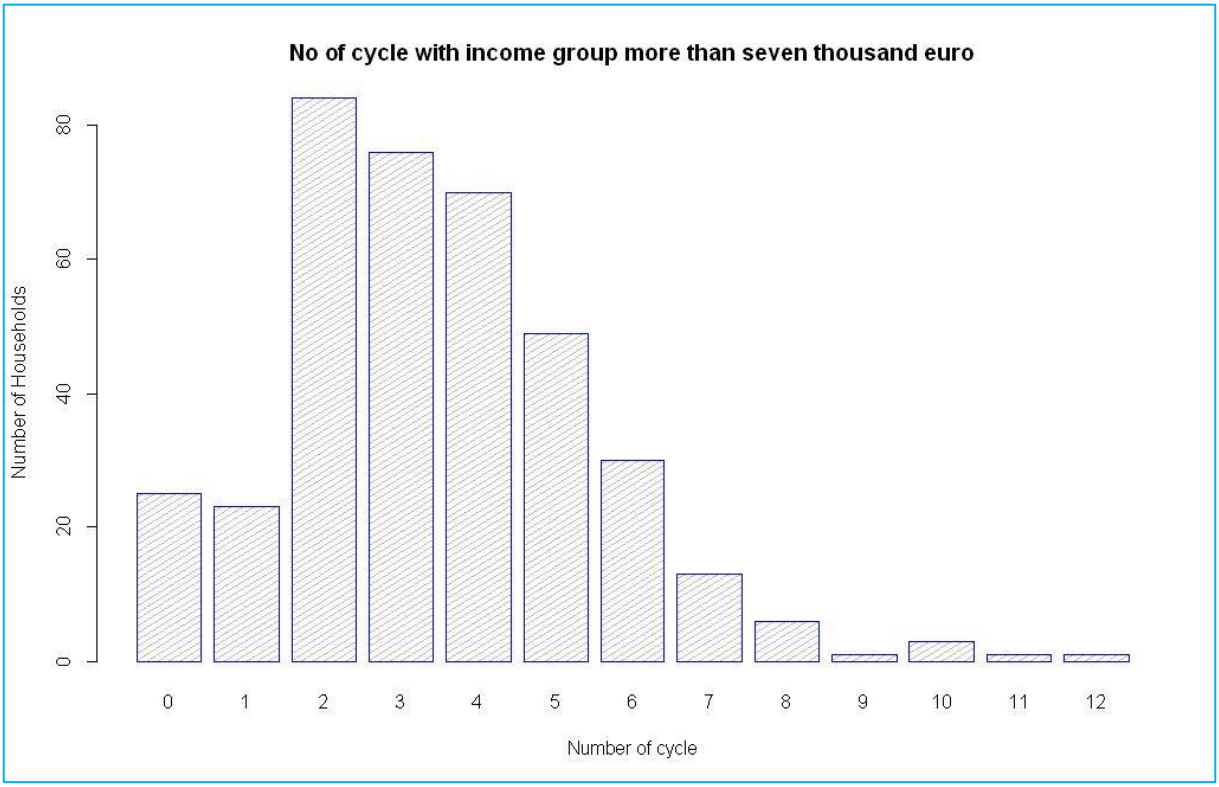
Using of cycles is being lesser from this income group. Around fifteen or more households use two to five cycles. Maximum number of cycle that belongs to a single household is eleven and being used by less than five household.



Figure(n)

(o) Income group less than four thousand six hundred euro:

If the income is more than seven thousand euro, then, the usages of cycles are getting more than previous income level.



Figure(o)

Main graph:

The main graph (figure-p) is plotted through using h2008$hheink and h2008$h04_1 variables, this graph shows number of households against income group level and different colours are showing the how it's shared of various number of cycles that belongs to a single household.

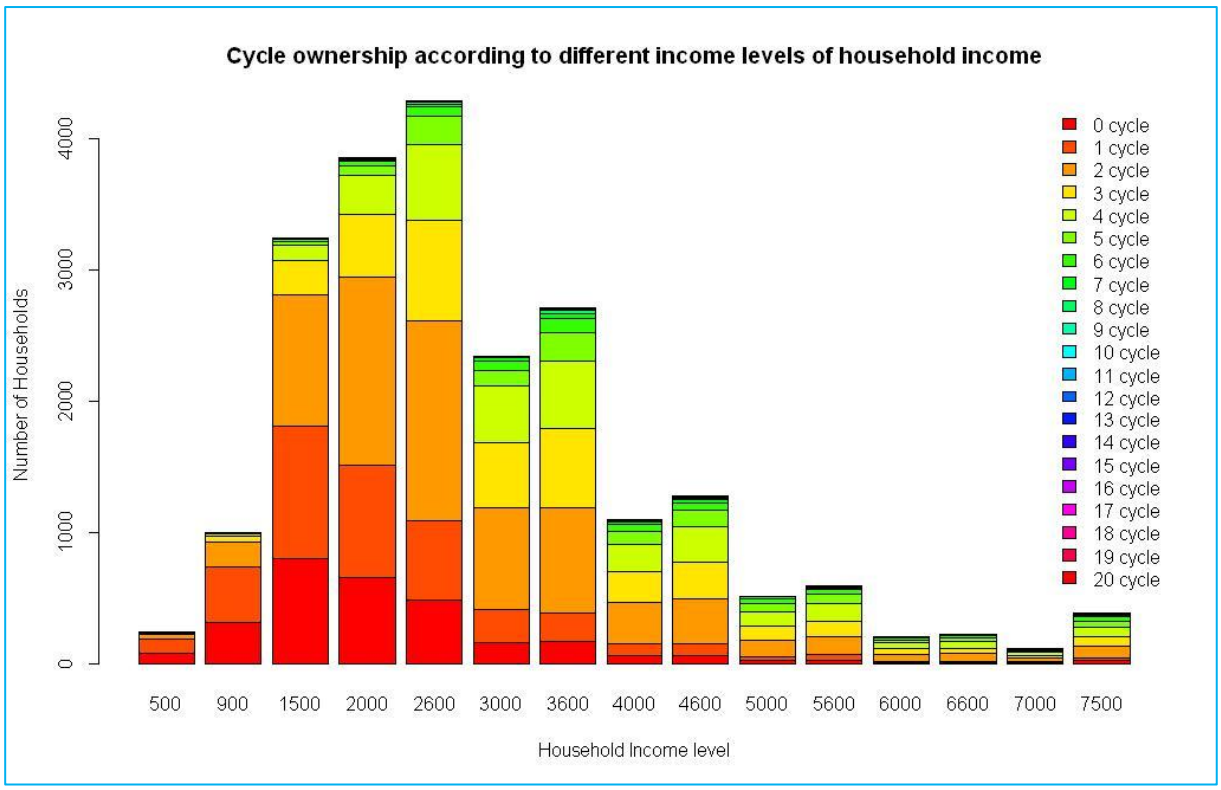


Figure (p)

1.4. T-tests:

Welch's *t* test is an adaptation of Student's *t*-test intended for use with two samples having possibly unequal variances. Here, we go for two tailed t-test where it's possible to test the null hypothesis whether the two population means are equal. All the tests were performed at a 95 percent confidence interval because it's the default for Project-R. A difference between means only becomes insignificant, i.e. the null hypothesis is accepted, if the P-value output is 2.5% or higher. In the following, all income values are per month and in Euro.

First test:

For first case, we will go for t-test between income levels less than five thousand Euros and less than five thousand six hundred Euros. The null hypothesis assumed is that the cycle ownership per household is the same for the mentioned income levels and alternative hypothesis describes its counterpart.

Null hypothesis, Ho = the cycle ownership per household is the same for the mentioned income levels and difference between them equal to zero.

Alternative hypothesis, H1 = the cycle ownership per household is not the same for the mentioned income levels.

Test result from R:

```
-----------------------------------------------------------------

t.test(incomeLevelTen,incomeLevelEleven,"two.sided")

Welch Two Sample t-test

data: incomeLevelTen and incomeLevelEleven
t = -0.3792, df = 1101.822, p-value = 0.7046

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:
-0.2526971 0.1708495

sample estimates:
mean of x mean of y
3.337864 3.378788

-----------------------------------------------------------------
```

The result shows that the null hypothesis is to be accepted as the difference between the means of the two is not significant (3.337864, 3.378788) at 95% confidence interval and the p-value is 0.7046 (greater than 2.5%).In the main graph, both the bars showed somewhat similar pattern of cycle ownership as well.

Second test:

For second case, we will go for t-test between income levels less than nine hundred euro and less than four thousand euro. The null hypothesis (Ho) assumed is that the cycle ownership per household is the same for the mentioned income level and difference between them equal to zero. Alternative hypothesis (H1) describes its counterpart.

Test result from R:

```
-----------------------------------------------------------------

t.test(incomeLevelTwo,incomeLevelEight,"two.sided")

Welch Two Sample t-test

data: incomeLevelTwo and incomeLevelEight

t = -30.4356, df = 1960.414, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.101649 -1.847197
sample estimates:
mean of x mean of y
1.091091 3.065514

-----------------------------------------------------------------
```

The result shows that the null hypothesis is to be rejected as the difference between the means of the two is significant (1.091091, 3.065514) at 95% confidence interval and the p-value is 2.2e-16 (lesser than 2.5%).In the main graph, both the bars showed different patterns of cycle ownership as well.

Third test:

Here, we will go for t-test between income levels less than two thousand six hundred euro and less than five thousand six hundred euro. The null hypothesis (Ho) assumed that the cycle ownership per household is the same for the mentioned income level and difference between them equal to zero. Alternative hypothesis (H1) describes its counterpart.

Test result from R:

```
-------------------------------------------------------------
t.test(incomeLevelFive,incomeLevelEleven,"two.sided")

Welch Two Sample t-test

data: incomeLevelFive and incomeLevelEleven
t = -12.7251, df = 707.847, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.1738313 -0.8600327
sample estimates:
mean of x mean of y
2.361856 3.378788
-------------------------------------------------------------
```

The result shows that the null hypothesis is to be rejected as the difference between the means of the two is significant (2.361856, 3.378788) at 95% confidence interval and the p-value is 2.2e-16 (lesser than 2.5%).In the main graph, both the bars showed different patterns of cycle ownership as well.

2.0 Conclusion:

1. Maximum number of cycle usage for most of the income groups is two. There is one interesting exception when income is more than six thousand euro and lesser than six thousand six hundred euro. In this particular income group, the maximum number of cycle usage is four. Another exception is when income is lesser than fifteen hundred euro, the maximum number of cycle usage is one. So, it seems that quantity of cycle usage in certain household doesn't matter where it stands in economic stratification, rather habit about cycling.
2. Usage of cycle is getting increased with increment of income levels and after a point it starts to get decreased. Probably people whose income is around two thousand or two thousand five hundred euro use more cycles than other.
3. It's interesting that few household in certain economic groups use as many as twenty cycles. The reason could be intense passion of cycle usage in a certain family that goes beyond rationality and may not come in our interest.

3.0 References

1. Walter, S.: July, 2007 (seminar talk in TUM): plotting with R
2. Venables, W. N., Smith, D. M., R Development Core Team: An introduction to R

4.0 Appendices

R code for project:

```
#Setting the working directorie

setwd("D:\\Lectures\\1st Semester\\Applied statistics
(Transport)\\Data\\Rdata files")

#reading household data

load("H2008.Rdata")

attach(h2008)

# Creating variables for further use

incomeLevelOne=0;incomeLevelTwo=0;incomeLevelThree=0;incomeLevelFour=0;
incomeLevelFive=0;incomeLevelSix=0;incomeLevelSeven=0;incomeLevelEight=0;
incomeLevelNine=0;incomeLevelTen=0;incomeLevelEleven=0; incomeLevelT-
welve=0;incomeLevelThirteen=0;incomeLevelFourteen=0; incomeLevelFifteen=0



#Dividing data of cycle ownage by each income level

numOfEntryLines<-dim(h2008)[1]

lineNum=0

for (lineNum in 1:numOfEntryLines) { if (hheink[lineNum]=="<500") {in-
comeLevelOne[lineNum]=h04_1[lineNum]}



if (hheink[lineNum]=="<900") {incomeLevelTwo[lineNum]=h04_1[lineNum]}

if (hheink[lineNum]=="<1500") {incomeLevelThree[lineNum]=h04_1[lineNum]}

if (hheink[lineNum]=="<2000") {incomeLevelFour[lineNum]=h04_1[lineNum]}

if (hheink[lineNum]=="<2600") {incomeLevelFive[lineNum]=h04_1[lineNum]}

if (hheink[lineNum]=="<3000") {incomeLevelSix[lineNum]=h04_1[lineNum]}

if (hheink[lineNum]=="<3600"){incomeLevelSeven[lineNum]=h04_1[lineNum]}
```

```
if (hheink[lineNum]=="<4000"){incomeLevelEight[lineNum]=h04_1[lineNum]}

if(hheink[lineNum]=="<4600"){incomeLevelNine[lineNum]=h04_1[lineNum]}

if(hheink[lineNum]=="<5000"){incomeLevelTen[lineNum]=h04_1[lineNum]}

 if(hheink[lineNum]=="<5600"){incomeLevelEleven[lineNum]=h04_1[lineNum]}

if(hheink[lineNum]=="<6000"){incomeLevelTwelve[lineNum]=h04_1[lineNum]}

if(hheink[lineNum]=="<6600"){incomeLevelThirteen[lineNum]=h04_1[lineNum]}

if(hheink[lineNum]=="<7000"){incomeLevelFourteen[lineNum]=h04_1[lineNum]}

if(hheink[lineNum]==">7000"){incomeLevelFifteen[lineNum]=h04_1[lineNum]}

 }

#Replacing data which we do not need with the NA

incomeLevelOne<-
re-
place(incomeLevelOne,incomeLevelOne=="98"|incomeLevelOne=="99"|incomeLeve
lOne=="97",NA)

incomeLevelTwo<-
re-
place(incomeLevelTwo,incomeLevelTwo=="98"|incomeLevelTwo=="99"|incomeLeve
lTwo=="97",NA)

incomeLevelThree<-
re-
place(incomeLevelThree,incomeLevelThree=="98"|incomeLevelThree=="99"|inco
meLevelThree=="97",NA)

incomeLevelFour<-
re-
place(incomeLevelFour,incomeLevelFour=="98"|incomeLevelFour=="99"|incomeL
evelFour=="97",NA)

incomeLevelFive<-
re-
place(incomeLevelFive,incomeLevelFive=="98"|incomeLevelFive=="99"|incomeL
evelFive=="97",NA)

incomeLevelSix<-
re-
place(incomeLevelSix,incomeLevelSix=="98"|incomeLevelSix=="99"|incomeLeve
lSix=="97",NA)

incomeLevelSeven<-
re-
place(incomeLevelSeven,incomeLevelSeven=="98"|incomeLevelSeven=="99"|inco
meLevelSeven=="97",NA)
```

```
incomeLevelEight<-
re-
place(incomeLevelEight,incomeLevelEight=="98"|incomeLevelEight=="99"|inco
meLevelEight=="97",NA)

incomeLevelNine<-replace                                    (incomeLevelN-
elN-
ine,incomeLevelNine=="98"|incomeLevelNine=="99"|incomeLevelNine=="97",NA)

incomeLevelTen<-
re-
place(incomeLevelTen,incomeLevelTen=="98"|incomeLevelTen=="99"|incomeLeve
lTen=="97",NA)

incomeLevelEleven<-
re-
place(incomeLevelEleven,incomeLevelEleven=="98"|incomeLevelEleven=="99"|i
ncomeLevelEleven=="97",NA)

incomeLevelTwelve<-
re-
place(incomeLevelTwelve,incomeLevelTwelve=="98"|incomeLevelTwelve=="99"|i
ncomeLevelTwelve=="97",NA)

incomeLevelThirteen<-
re-
place(incomeLevelThirteen,incomeLevelThirteen=="98"|incomeLevelThirteen==
"99"|incomeLevelThirteen=="97",NA)

incomeLevelFourteen<-
re-
place(incomeLevelFourteen,incomeLevelFourteen=="98"|incomeLevelFourteen==
"99"|incomeLevelFourteen=="97",NA)

incomeLevelFifteen<-
re-
place(incomeLevelFifteen,incomeLevelFifteen=="98"|incomeLevelFifteen=="99
"|incomeLevelFifteen=="97",NA)



#Ploting the graphs for the results



barplot(table(incomeLevelOne), main="No of cycles with income group less
than five hundred euro", xlab="Number of cycles", ylab="Number of House-
holds", border="green",density=c(10))

barplot(table(incomeLevelTwo), main="No of cycles with income group less
than nine hundred euro", xlab="Number of cycle", ylab="Number of House-
holds", border="yellow",density=c(15))

barplot(table(incomeLevelThree), main="No of cycle with income group less
than fifteen hundred euro", xlab="Number of cycle", ylab="Number of
Households", border="yellow",density=c(20))
```

```
barplot(table(incomeLevelFour), main="No of cycle with income group less than two thousand euro", xlab="Number of cycle", ylab="Number of Households", border="blue",density=c(20))

barplot(table(incomeLevelFive), main="No of cycle with income group less than two thousand six hundred euro", xlab="Number of cycle", ylab="Number of Households", border="blue",density=c(20))

barplot(table(incomeLevelSix), main="No of cycle with income group less than three thousand euro", xlab="Number of cycle", ylab="Number of Households", border="blue",density=c(20))

barplot(table(incomeLevelSeven), main="No of cycle with income group less than three thousand six hundred euro", xlab="Number of cycle", ylab="Number of Households", border="blue",density=c(20))


barplot(table(incomeLevelEight), main="No of cycle with income group less than four thousand euro", xlab="Number of cycle", ylab="Number of Households", border="blue",density=c(20))


barplot(table(incomeLevelNine), main="No of cycle with income group less than four thousand six hundred euro", xlab="Number of cycle", ylab="Number of Households", border="blue", density=c(20))

barplot(table(incomeLevelTen), main="No of cycle with income group less than five thousand euro", xlab="Number of cycle", ylab="Number of Households", border="blue",density=c(20))

barplot(table(incomeLevelEleven), main="No of cycle with income group less than five thousand six hundred euro", xlab="Number of cycle", ylab="Number of Households", border="blue",density=c(20))

barplot(table(incomeLevelTwelve), main="No of cycle with income group less than six thousand euro", xlab="Number of cycle", ylab="Number of Households", border="blue",density=c(20))

barplot(table(incomeLevelThirteen), main="No of cycle with income group less than six thousand six hundred euro", xlab="Number of cycle", ylab="Number of Households", border="blue",density=c(20))

barplot(table(incomeLevelFourteen), main="No of cycle with income group less than seven thousand euro", xlab="Number of cycle", ylab="Number of Households", border="blue",density=c(30))

barplot(table(incomeLevelFifteen), main="No of cycle with income group more than seven thousand euro", xlab="Number of cycle", ylab="Number of Households", border="blue",density=c(30))


#The levels "Nein", "Weiß nicht", "k.A." replaced by missing values for No of Cycles
```

```
h2008$h04_1<-replace(h2008$h04_1,h2008$h04_1=="98"|h2008$h04_1=="99"
|h2008$h04_1== "97",NA)


#The levels "Nein", "Weiß nicht", "k.A." replaced by missing val-
ues for household income incomelevel


h2008$hheink<-replace(h2008$hheink,h2008$hheink=="WeißNicht"|
h2008$hheink=="k.A."|h2008$hheink=="Verweigert",NA)


#Conversion of the income level in to numeric data to plot

mainplot = c(h2008$hheink)

fdata = factor(mainplot)

rda-
ta=factor(fdata,labels=c("500","900","1500","2000","2600","3000","3600","
4000", "4600", "5000","5600","6000","6600","7000","7500"))

#Plotting of the main graph "Cycle ownership according to differ-
ent income levels of household income"

mainplot2 <- table(h2008$h04_1,rdata )

barplot(mainplot2, main="Cycle ownership according to different income
levels of household income",xlab="Household Income level", ylab="Number
of Households", col=rainbow(20))

legend("topright", c("0 cycle","1 cycle","2 cycle","3 cycle","4 cycle","5
cycle","6 cycle","7 cycle","8 cycle","9 cycle","10 cycle","11 cycle","12
cycle","13 cycle","14 cycle","15 cycle","16 cycle","17 cycle","18 cy-
cle","19 cycle", "20 cycle"), cex= 1.0, bty="n",fill=rainbow(20));

#T-test from various Income Levels compared to the whole Sample

t.test(incomeLevelTen,incomeLevelEleven,"two.sided")

t.test(incomeLevelTwo,incomeLevelEight,"two.sided")

t.test(incomeLevelTwelve,incomeLevelThirteen,"two.sided")
```