DOT Survey Data Analysis

UshaKiran.Kota

May 23, 2016

# Coursolve/Need230 : DOT Data Analysis

## Part 1 -Introduction

Digital Opportunity Trust (DOT)(ref://<https://www.dotrust.org>) is a leading international social enterprise headquartered in Ottawa, Canada with local operations around the globe. DOT mobilizes youth the talent and energy helping them to develop both an entrepreneurial spirit and technology and business skills that will last a lifetime. Young people are encouraged to become leaders of change as they facilitate technology, business, and entrepreneurial learning experiences to people in their own communities.

DOT's unique youth-led programs empower people living in communities that are developing, in transition, or under stress with the confidence to use technology for entrepreneurial, community, educational, and personal development.

\_In September 2015, twenty youth from Kenya, Rwanda, Ethiopia, Uganda, Tanzania, Lebanon, and Indigenous Canada embarked on a project to survey their peers about how they are using technology and social media for work, income, learning, leadership, and employment. 580 youth from urban, peri-urban, and rural areas were surveyed through face-to-face qualitative and quantitative interviews.

The result of this survey is a raw dataset of 580 responses and more than 1600 variables ,

Coursolve/Need230 is a short term project to analyse:

How are youth in Kenya, Rwanda, Ethiopia, Uganda, Tanzania, Lebanon, and Indigenous Canada using technology and social media for work, income, learning, leadership, and employment?

DOT provided the team with a Raw Survey responses.xlsx as an input data for the exploratory Analysis and expects to find answers for some or all of the below questions from the results of Need230

### Document Purpose

The purpose of this document is describe one of the several methods of how to handle dataset of large number of categorical variables to reduce the dimensions of the data set for ease of exploration for research questions

### Need230/DOT Data Analysis - Research Questions

1.How does access to technology vary by gender?

2.How does use of social media for work, income, learning, leadership, and employment vary by gender?

3.In what ways are youth in these countries using technology and social media to support or engage in formal work? Informal work?

4.In what ways are youth in these countries using technology and social media to supplement income, or to support primary incomes?

5.In what ways are youth in these countries using technology and social media for learning?

6.What barriers and incentives to online learning are there among the surveyed youth?

7.How are young people positioning themselves as leaders in their communities or among peer groups using social media and technology?

8.How are young people gaining and/or participating in formal or informal employment using technology and social media?

## Justification

DOT intends to learn more about how our key stakeholder group (youth in the identified countries) are benefitting from technology and social media in the areas of employment, entrepreneurship, learning, and leadership.

DOT will use the results of Need230 to inform program and project design, as well as implementation strategies.

This document is intended to present analysis for Questions 1, 3, 5 from the above list of research questions

### 1.Survey Data Insights

This section summarises the understanding of Survey Data

#### 1.Populations and samples

The target population in the DOT survey report is youth of different age categories who belong to Canada , Sub sahara countries of Africa and Lebanon, well stratifies by Age, Gender and Location The data consists of 568 respondents randomly sampled from the target locations and by Age, Gender forming a well representaion of stratified sampled data. Hence it is reasonable to assume that the estimates on the sample cases will be unbiased

DOT research team conducted a prospective observational study, with a cohort of 568 individuals and responses collected over 1500 features. Hence it is intended to check for any evidence of association between the features over which the data was collected

Ref:<https://onlinecourses.science.psu.edu/stat504/book/export/html/102>

Each row of data represents a case of 1539 features. The total number of cases are 568 DOT Data is a collection of responses over features that are dichotmous ( binary, nominal and categorical (multi-level responses)) type, collected from each respondent. Data also includes variable that elicit open-ended responses to questions from the researcher #### 3. Respondent Demographics View

#### 2.Data Handling

This section describes how the data is organised for exploratory and inferential analysis.

For a majority of categorical multi-level variables, Data is organanized to individual subsets for each set of explanatory and response variables. this ensures a Multinomial sampling with fixed samples sizes, stratified by different variables. Cross tabulation of multi-way (minimum of 3 dimensions) is a chosen as a perferred method to study any association between the chosen features

Data cleaning up process comprised of the foll:

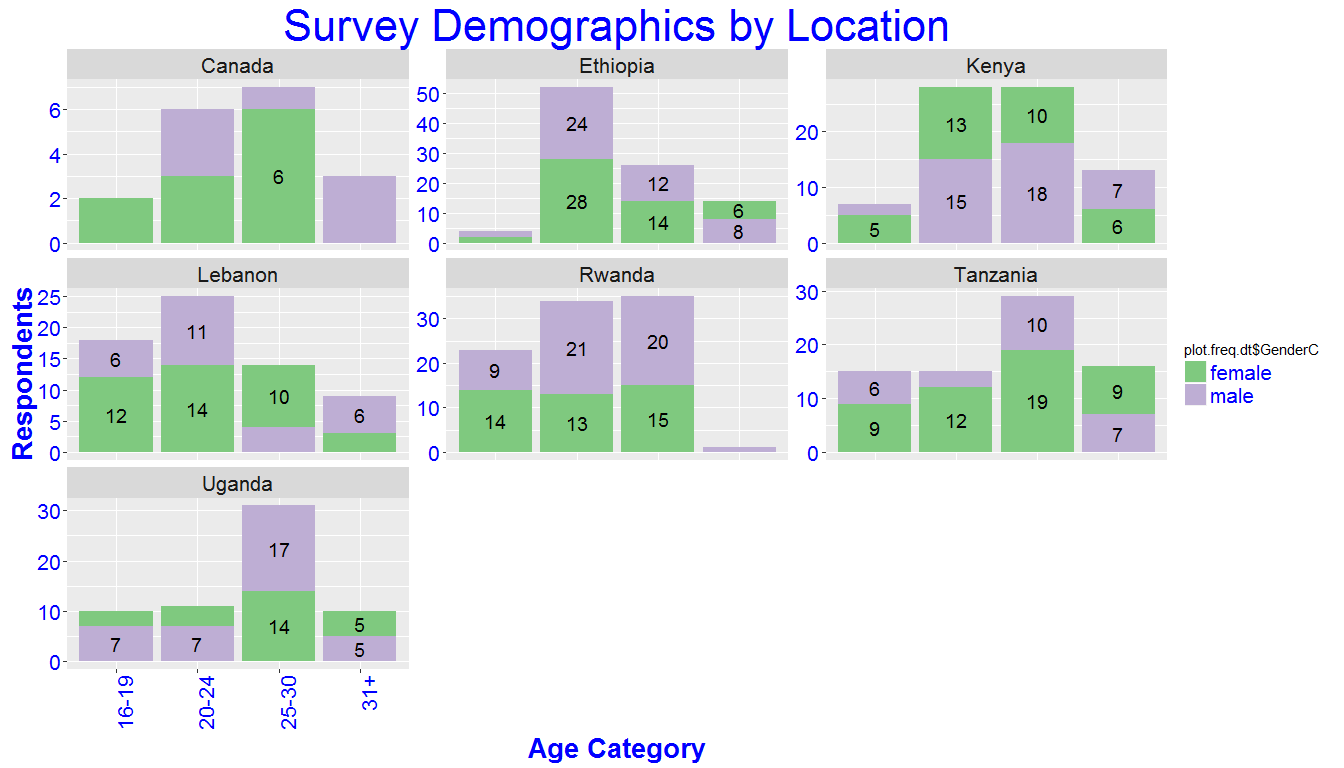
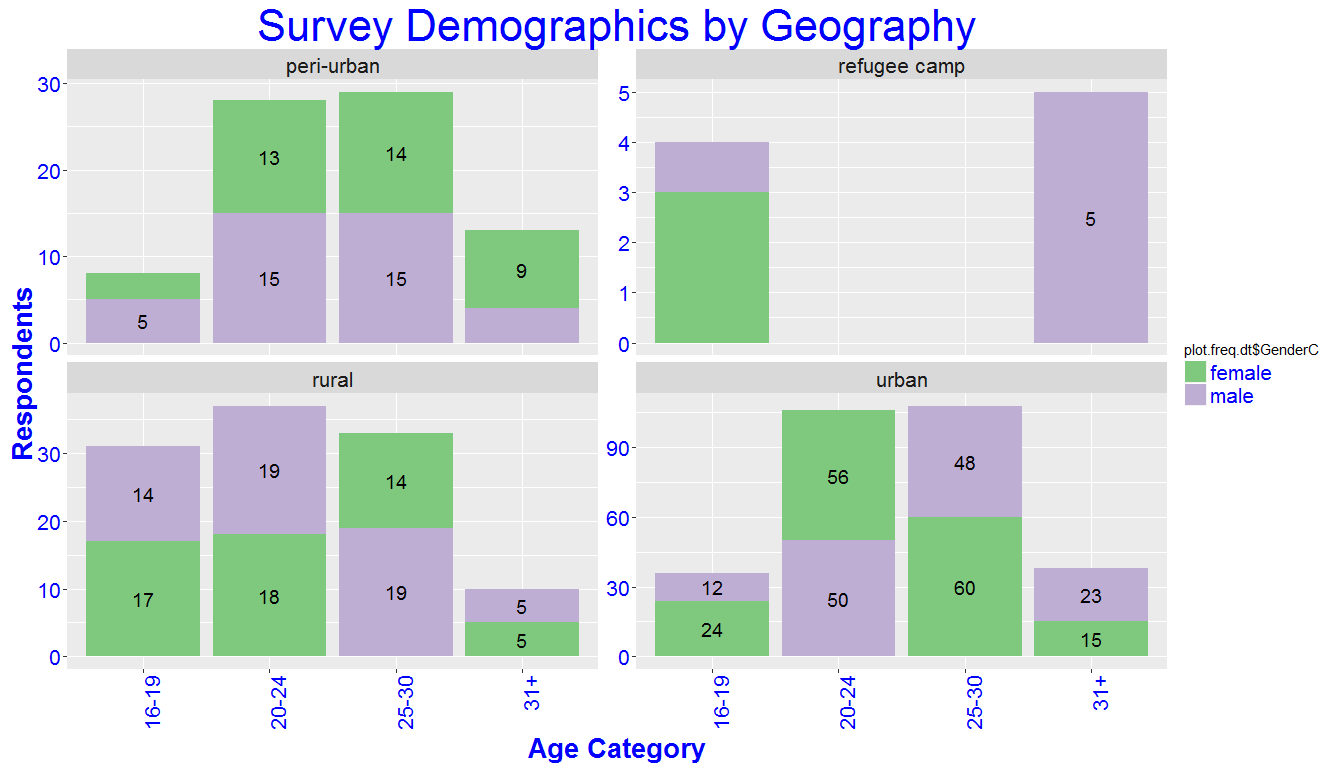
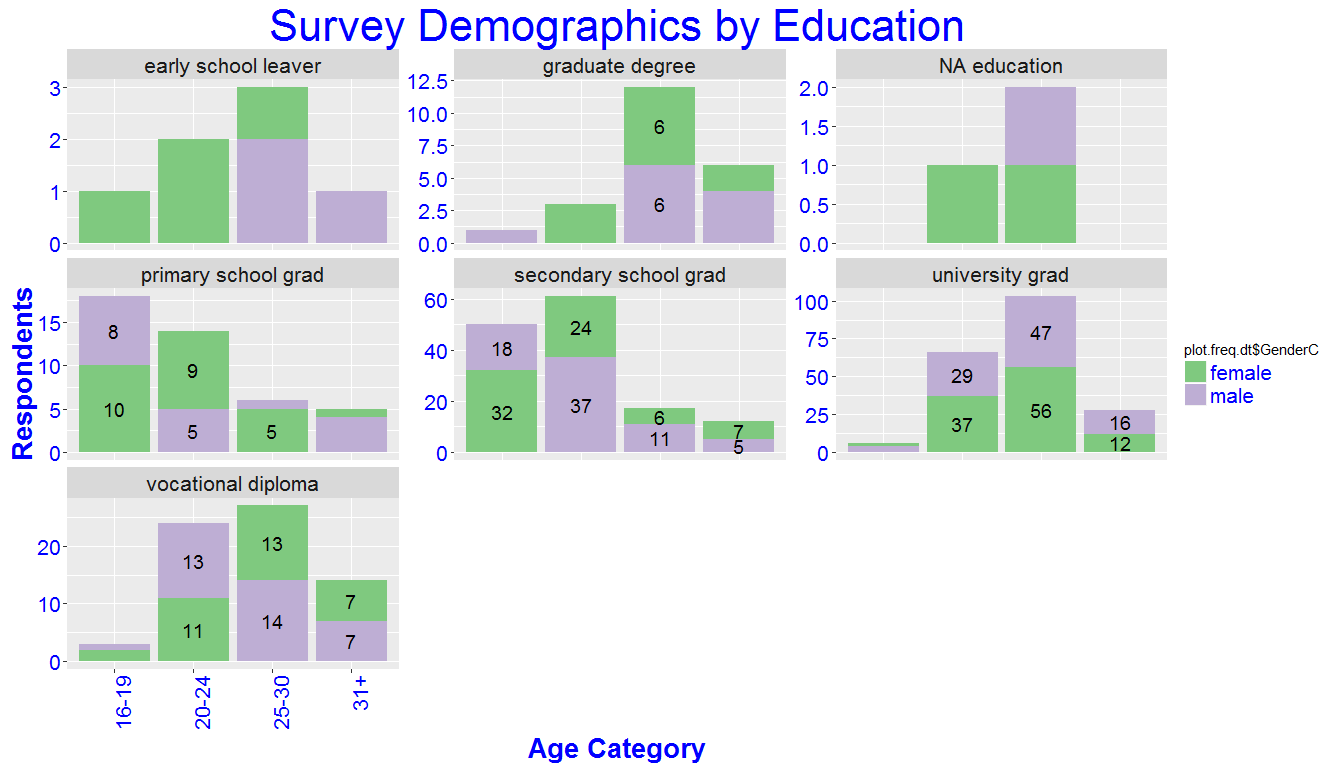
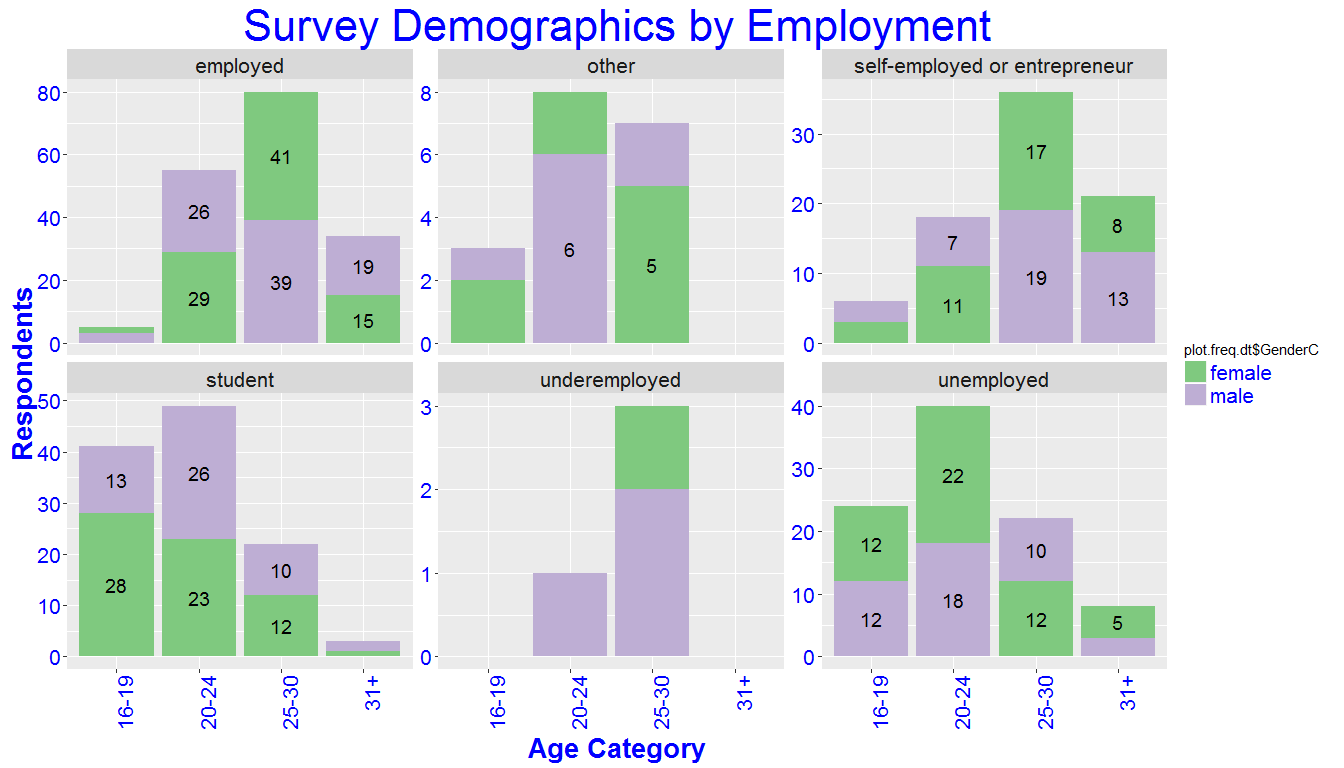
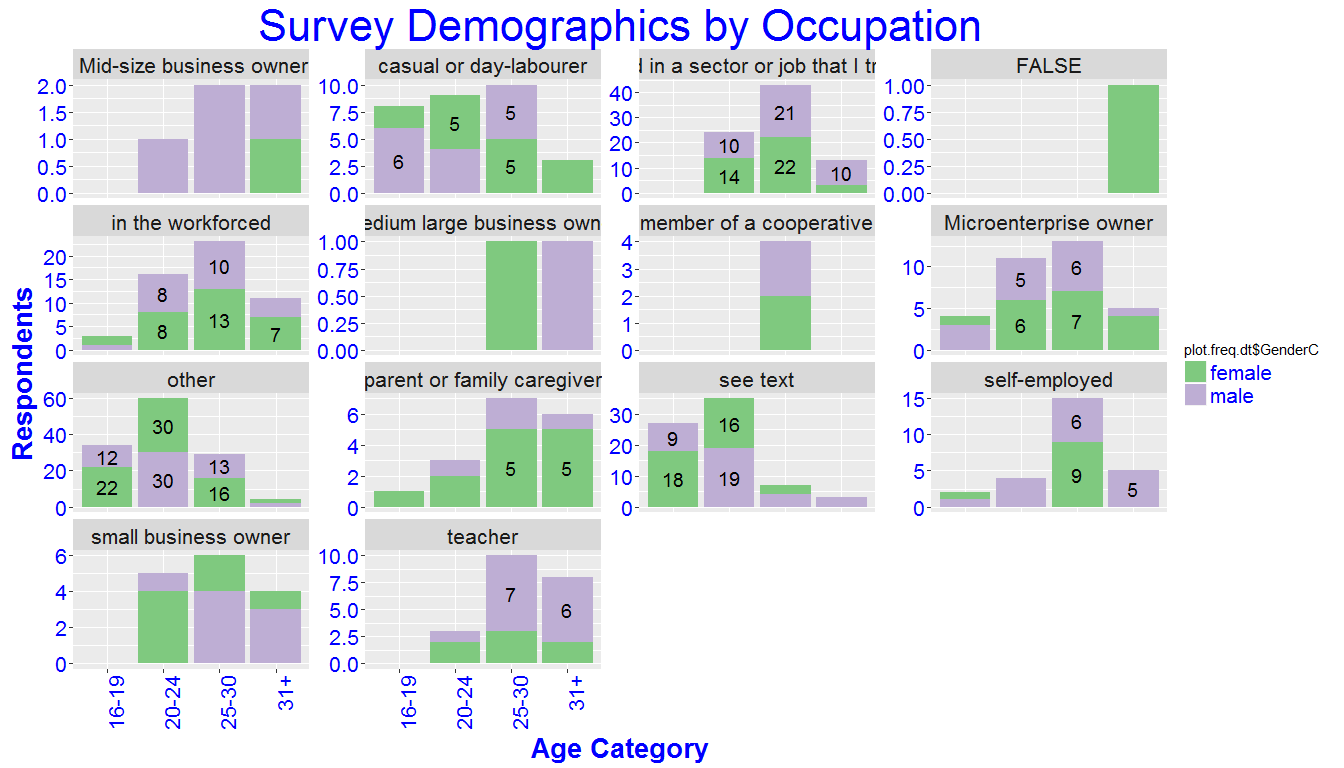
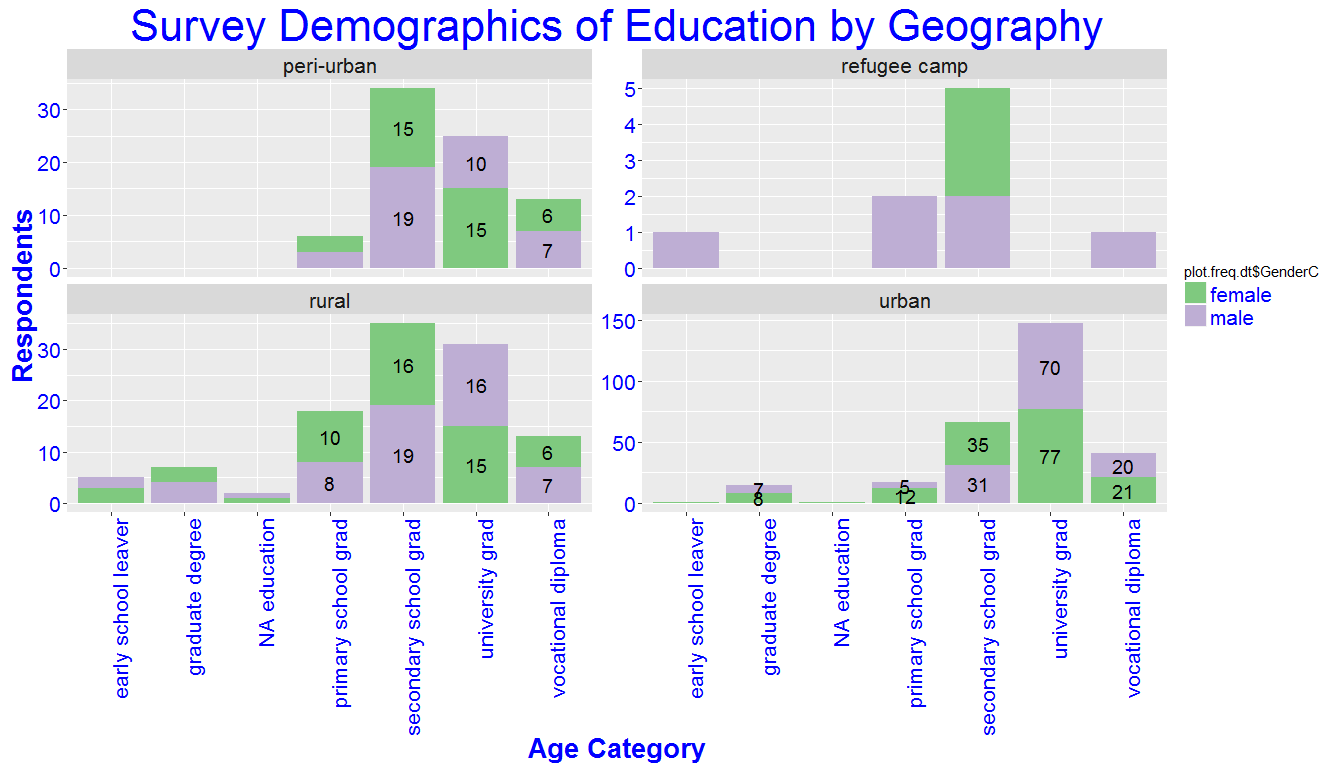
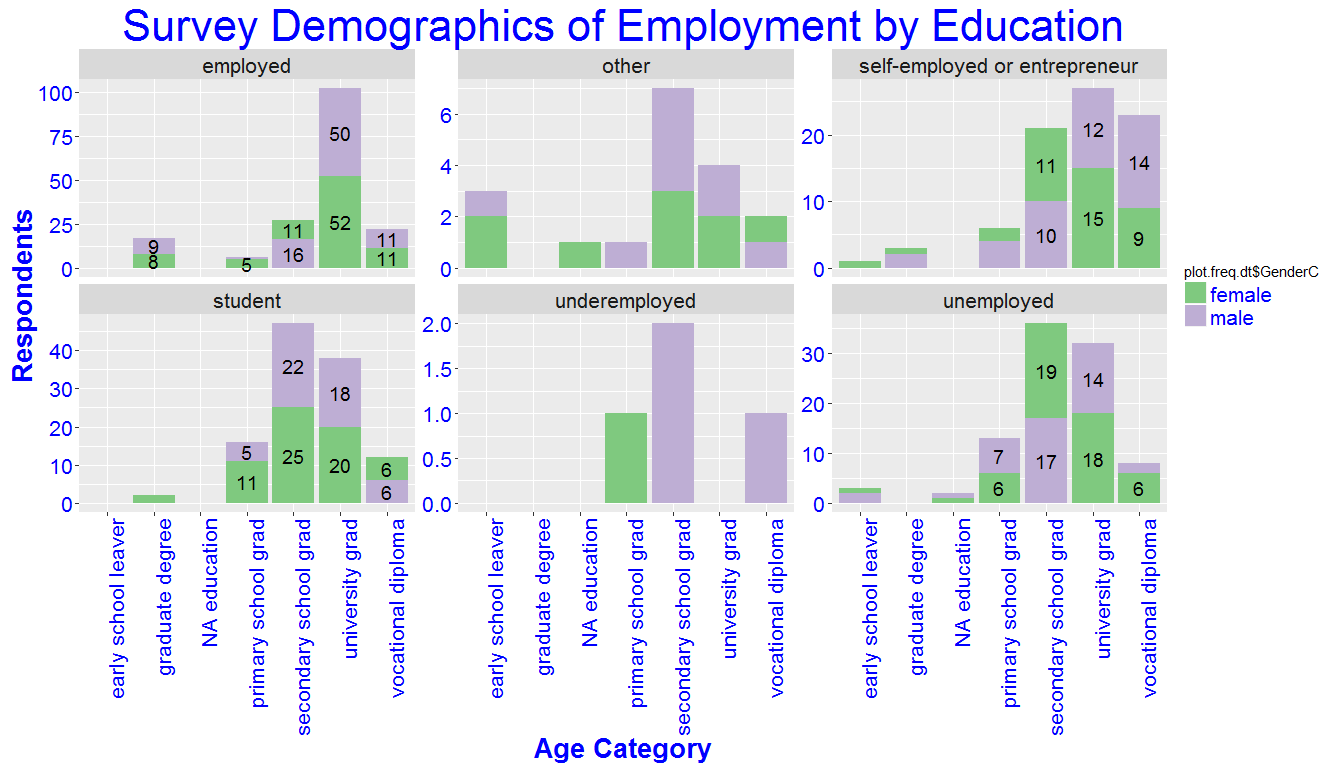
1. creating segmented data sets for variables of analysis
2. shortening variables names by length and or replace for convenient handling with the software
3. reshaping the data to group the appropriate variable for each survey question within the chosen research Questions for analysis

### 

## [1] "Glipmse of Survey Data"

## ID Status Language Location WeightedScore  
## 1: 43413695 Incomplete English 35  
## 2: 43486387 Complete English 26  
## 3: 43645338 Complete English 1063  
## 4: 43675088 Complete English 961  
## 5: 43749321 Complete English Kenya 621  
## ResearcherEmail LocationRespondent Gather Gather2 Gender  
## 1: jkwatampora@dotrust.org Uganda 1 interview 2  
## 2: zfakhry@dotrust.org Lebanon 2 phone 2  
## 3: zfakhry@dotrust.org Lebanon 1 interview 1  
## 4: zfakhry@dotrust.org Lebanon 1 interview 2  
## 5: angendo@dotrust.org Kenya 1 interview 1  
## GenderC Age.Range AgeC Geography GeographyC Education  
## 1: female 3 25-30 2 urban 2  
## 2: female 4 31+ 1 rural 2  
## 3: male 3 25-30 1 rural 1  
## 4: female 2 20-24 1 rural 1  
## 5: male 2 20-24 2 urban 4  
## EducationC Employment.Status EmploymentC  
## 1: university grad 4 student  
## 2: university grad 1 employed  
## 3: graduate degree 3 self-employed or entrepreneur  
## 4: graduate degree 1 employed  
## 5: secondary school grad 2 unemployed  
## Occupation OccupationC  
## 1: 1 self-employed  
## 2: 7 teacher  
## 3: 4 Mid-size business owner  
## 4: 7 teacher  
## 5: 11 casual or day-labourer

#### 3.Exploratory View of Demographics

## [1] "2-way Contingency table: varying by Gender and Age"  
##   
## 16-19 20-24 25-30 31+  
## female 0.1872510 0.3466135 0.3505976 0.1155378  
## male 0.1361702 0.3574468 0.3489362 0.1574468  
## [1] "2-way Contingency table: varying by Gender and Education"  
##   
## early school leaver graduate degree NA education  
## female 0.015936255 0.043824701 0.007968127  
## male 0.012765957 0.046808511 0.004255319  
##   
## primary school grad secondary school grad university grad  
## female 0.099601594 0.274900398 0.426294821  
## male 0.076595745 0.302127660 0.408510638  
##   
## vocational diploma  
## female 0.131474104  
## male 0.148936170  
## [1] "2-way Contingency table: varying by Gender and Employment status"  
##   
## employed other self-employed or entrepreneur student  
## female 0.346613546 0.035856574 0.155378486 0.254980080  
## male 0.370212766 0.038297872 0.178723404 0.217021277  
##   
## underemployed unemployed  
## female 0.003984064 0.203187251  
## male 0.012765957 0.182978723  
## [1] "2-way Contingency table: varying by Gender and OccupationC"  
##   
## Mid-size business owner casual or day-labourer  
## female 0.003984064 0.059760956  
## male 0.017021277 0.063829787  
##   
## employed in a sector or job that I trained for FALSE  
## female 0.155378486 0.003984064  
## male 0.174468085 0.000000000  
##   
## in the workforced medium large business owner  
## female 0.119521912 0.003984064  
## male 0.097872340 0.004255319  
##   
## member of a cooperative Microenterprise owner other  
## female 0.007968127 0.071713147 0.278884462  
## male 0.008510638 0.063829787 0.242553191  
##   
## parent or family caregiver see text self-employed  
## female 0.051792829 0.147410359 0.039840637  
## male 0.017021277 0.148936170 0.068085106  
##   
## small business owner teacher  
## female 0.027888446 0.027888446  
## male 0.034042553 0.059574468  
##   
## early school leaver graduate degree NA education  
## peri-urban 0.00000000 0.00000000 0.00000000  
## refugee camp 0.14285714 0.00000000 0.00000000  
## rural 0.71428571 0.31818182 0.66666667  
## urban 0.14285714 0.68181818 0.33333333  
##   
## primary school grad secondary school grad university grad  
## peri-urban 0.13953488 0.24285714 0.12315271  
## refugee camp 0.04651163 0.03571429 0.00000000  
## rural 0.41860465 0.25000000 0.15270936  
## urban 0.39534884 0.47142857 0.72413793  
##   
## vocational diploma  
## peri-urban 0.19117647  
## refugee camp 0.01470588  
## rural 0.19117647  
## urban 0.60294118  
##   
## employed other self-employed or entrepreneur  
## peri-urban 0.16091954 0.27777778 0.20987654  
## refugee camp 0.01149425 0.00000000 0.03703704  
## rural 0.18390805 0.55555556 0.17283951  
## urban 0.64367816 0.16666667 0.58024691  
##   
## student underemployed unemployed  
## peri-urban 0.14782609 0.25000000 0.10638298  
## refugee camp 0.01739130 0.00000000 0.02127660  
## rural 0.23478261 0.00000000 0.29787234  
## urban 0.60000000 0.75000000 0.57446809

#### 5.Executive Summary

As part of the need 230, DOT sought response to atleast 3 of research questions. Each of these 3 research questions is a group of questions that elicit response over technology access and use related within social environment of target population.

This section describes the approach for data analysis and Exploratory analyses and Statistical analyses that will be peformed on the DOT survey Data.

The purpose of the research is to provide useful insights about the data with the help of Exploratory views, Qualitative comments and results of Inferential analysis on selected data sets Apart from the demographics of the respondents, DOT survey is a collection of responses for survey Questions related to access and use of technology for the activities that the youth are associated with . Response variables are related to the access and use of technology by youth within their living (to change this word) context.

Basic assumptions about conditions of indenpendence and randomness within group are made while analysing the data, Since the data are nominal categorical in nature, it will be appropriate to use "point estimates", and ranges for summary statistics.

Exploratory methods such as stacked bar plots, mosaic plots and contingency tables will be used visualize the data and any associations within groups or between groups of data.

Since the sample sizes with multi-level responses are large enough, we can test for independence between the Gender of respondents and their response for technology choice using a **Pearson's chi-squared test**, which will be performed on selected data sets for the sub group of questions. Regression analysis is also performed on selected data sets to check the linear-log linear variability among the response/explanatory variables

Research Questions 3 and 5 elicit response in the form of "open ended text". Exploratory views such as Wordclouds and Latent Topics will be included as part of the text analysis process in these cases. Document to Topic associations will be illustrated with suitable plots that provide insights into the latent topics discussed by the respondents

All the plots and statistical test results in this document are performed using R statistical Software programming language.

### Research Question I

Research Question I is **"How does access to technology vary by Gender"** THis question has response data within collection of following sub set of questions :

**What devices do you have access to?** **How often do you actually use these devices?**

**Where do you most commonly use the devices you have access to?**

**How much access to do you have to the Internet?**

**How often do you actually use the Internet?**

**Where do you most commonly use the Internet?**

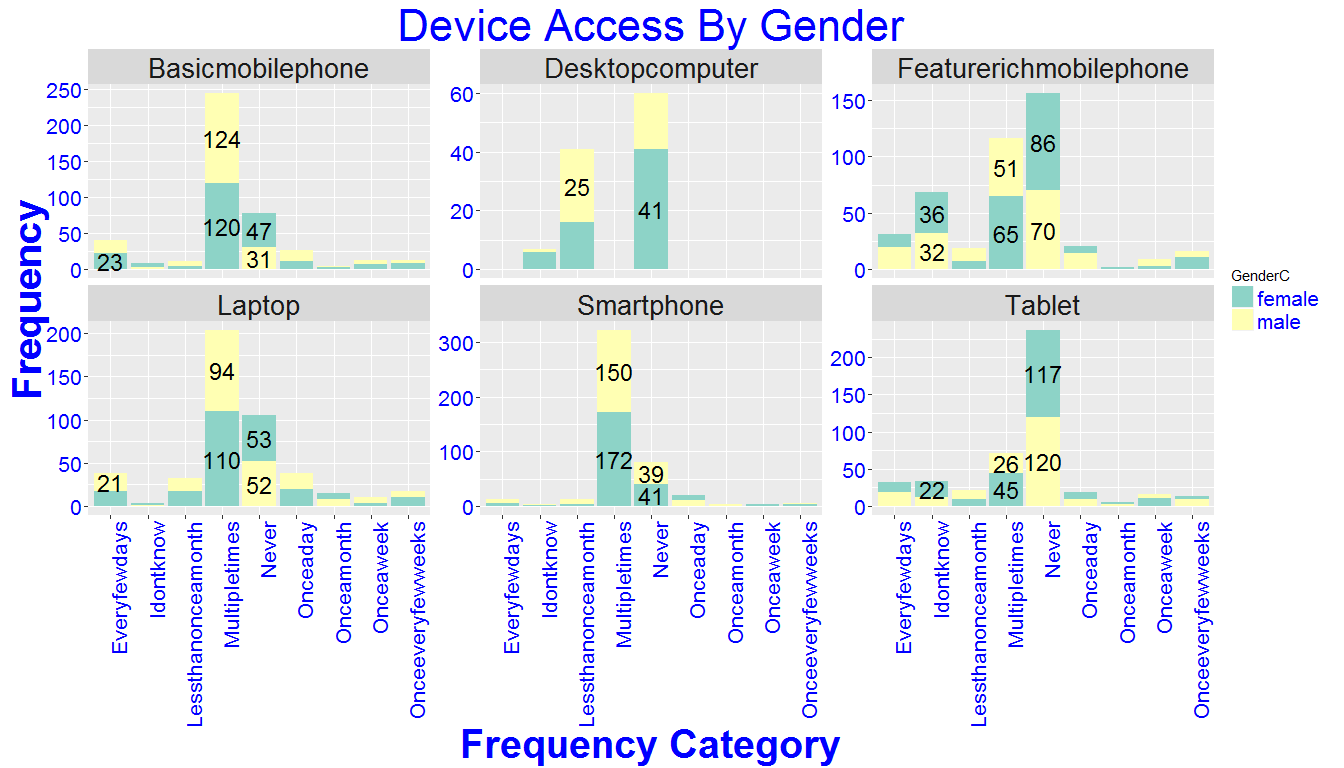
**What digital services/tools do you have access to?**

**Where do you most commonly use digital services?** #

The Question I seeks to answer in the form of investigation for any significant association between Gender of the respondents and access to technology grouped under :

1. Access to electronic devices such as smartphones, tablets, laptops etc, their use frequencies, and location of use
2. Access to Internet, Use frequency of internet by the respondent on a device and location of use
3. Access to any digital services via internet access on the devices that they use.

#### 1. What devices do you have access to?



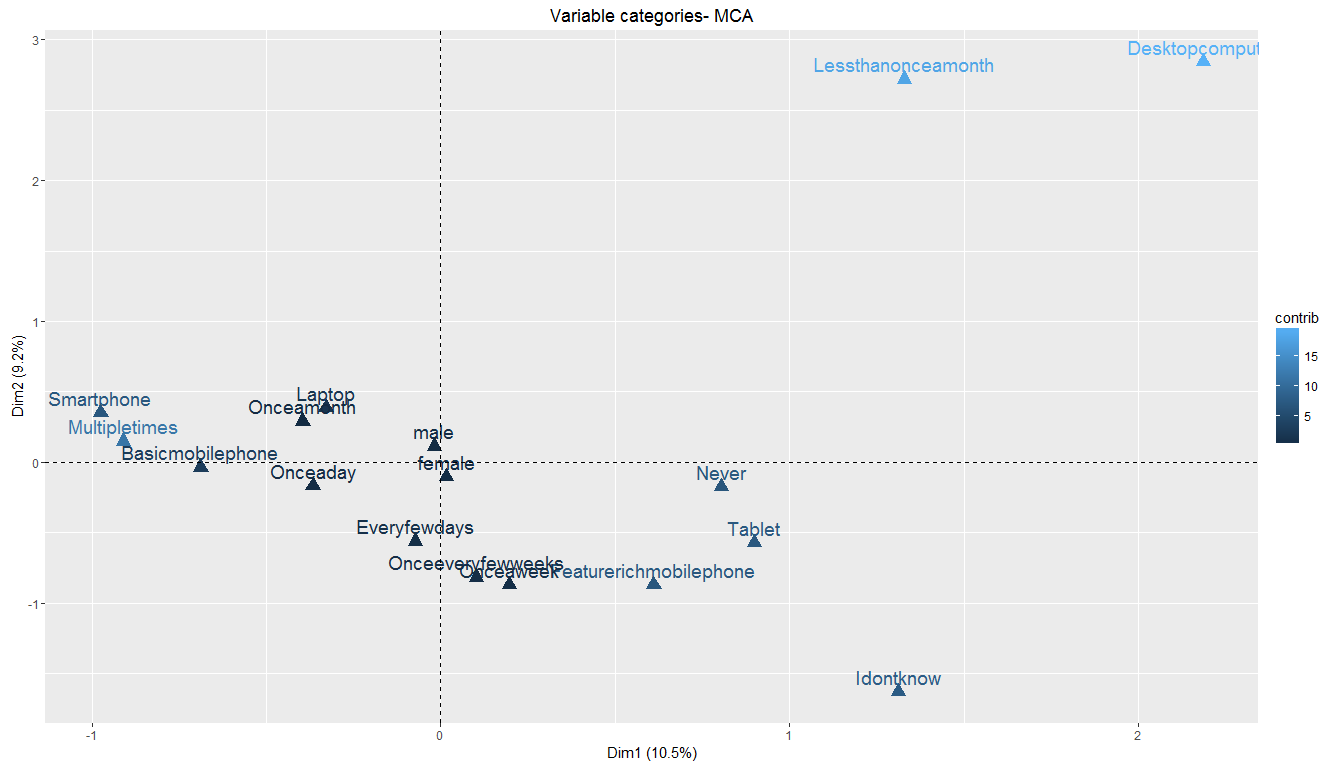
## [1] "Is there an association between respondent's Gender and device accessed ?"

## [1] "The response variable here is the device and the explanatory variable is Gender"

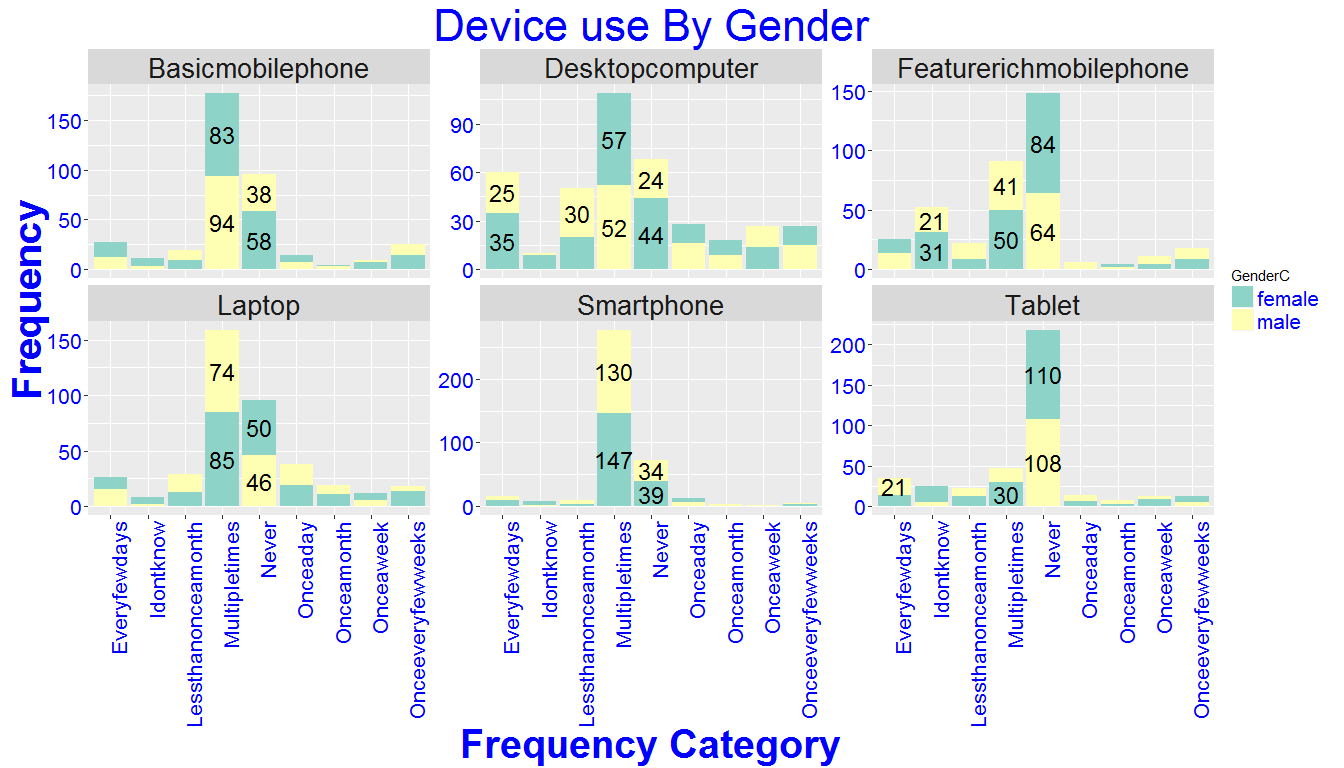
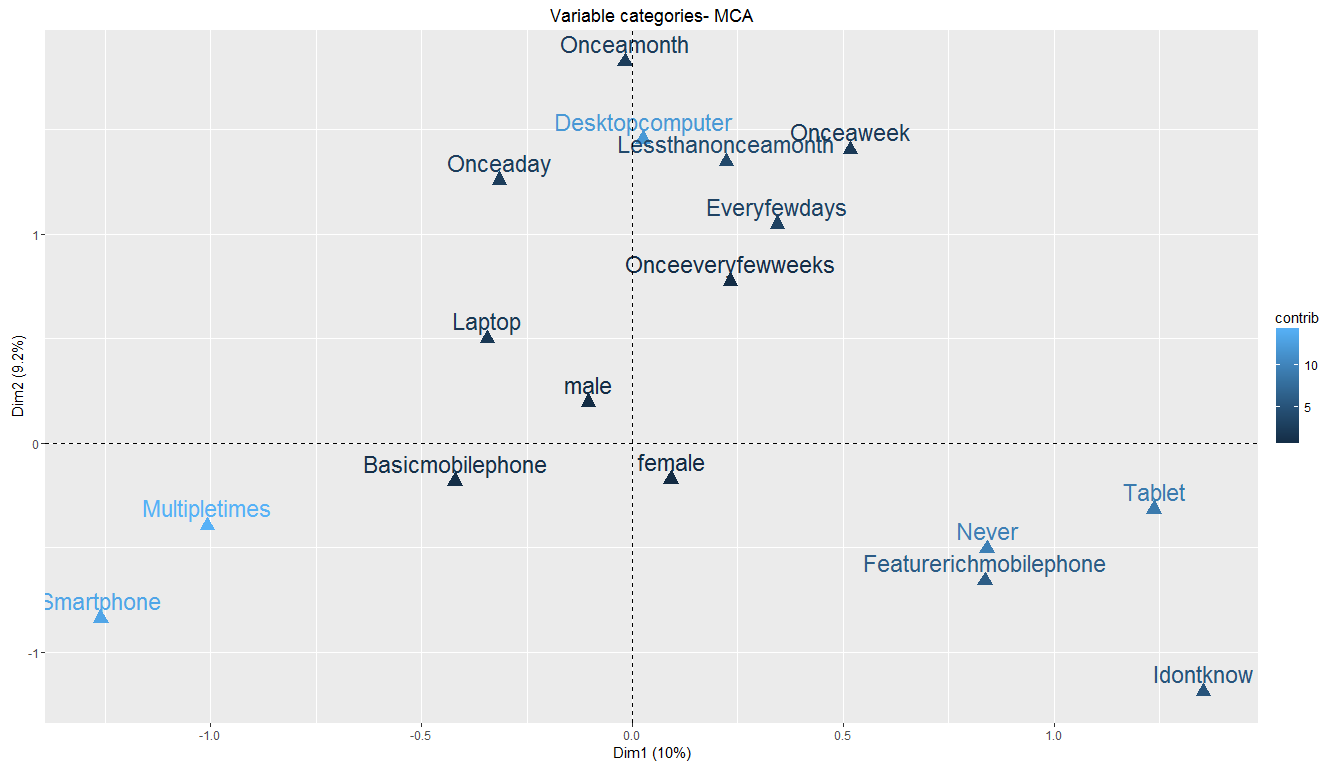
## [1] "A simpler - Chi.Square test should be sufficient to test for Independence"

|  |  |  |
| --- | --- | --- |
|  | female | male |
| Basicmobilephone | 0.5204545 | 0.4795455 |
| Desktopcomputer | 0.5833333 | 0.4166667 |
| Featurerichmobilephone | 0.5205479 | 0.4794521 |
| Laptop | 0.5205184 | 0.4794816 |
| Smartphone | 0.5193966 | 0.4806034 |
| Tablet | 0.5233853 | 0.4766147 |

## Call: xtabs(formula = ~device + GenderC, data = dot.gen.acc.dev.dt)  
## Number of cases in table: 2362   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 1.6292, df = 5, p-value = 0.8977



#### 2. How often do you actually use these devices?

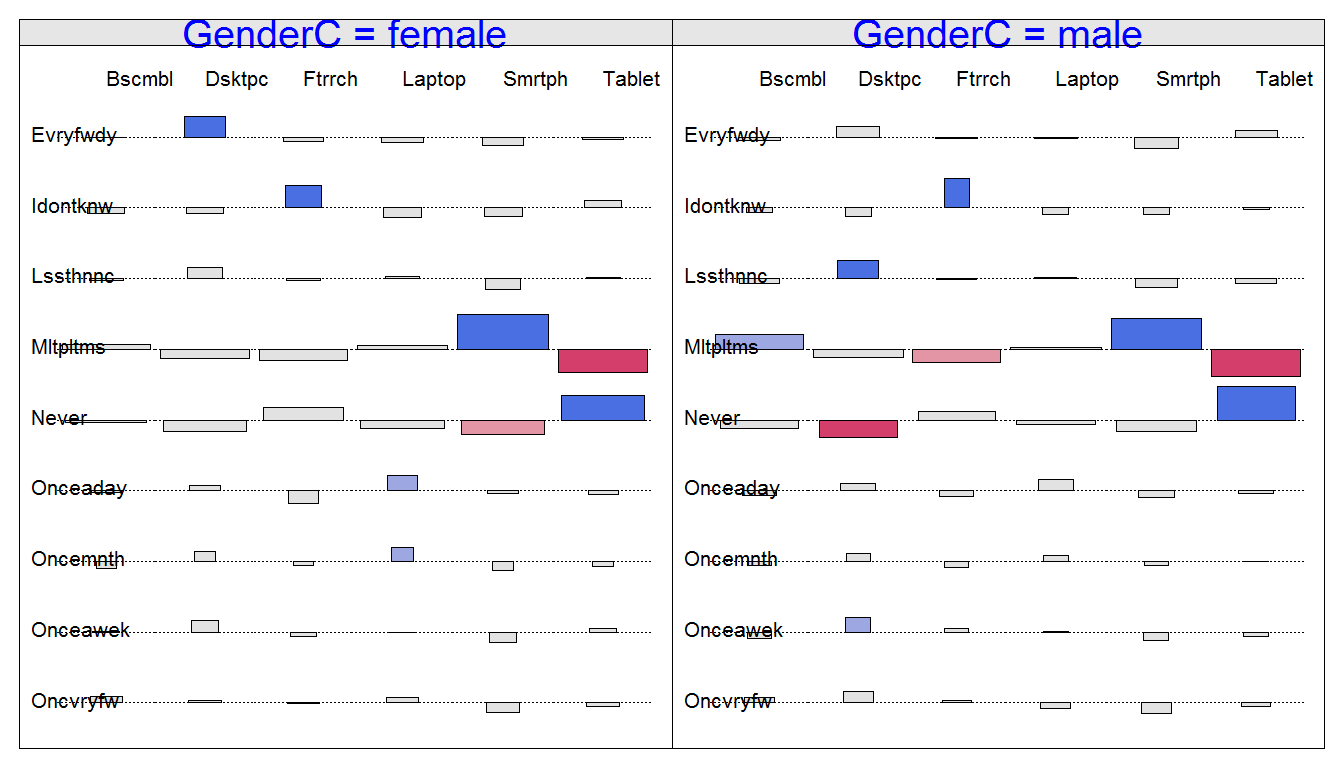
## Call: xtabs(formula = ~use.dev.freq + GenderC + device.use, data = dot.gen.use.dev.dt)  
## Number of cases in table: 2358   
## Number of factors: 3   
## Test for independence of all factors:  
## Chisq = 694.1, df = 93, p-value = 1.044e-92  
## Chi-squared approximation may be incorrect

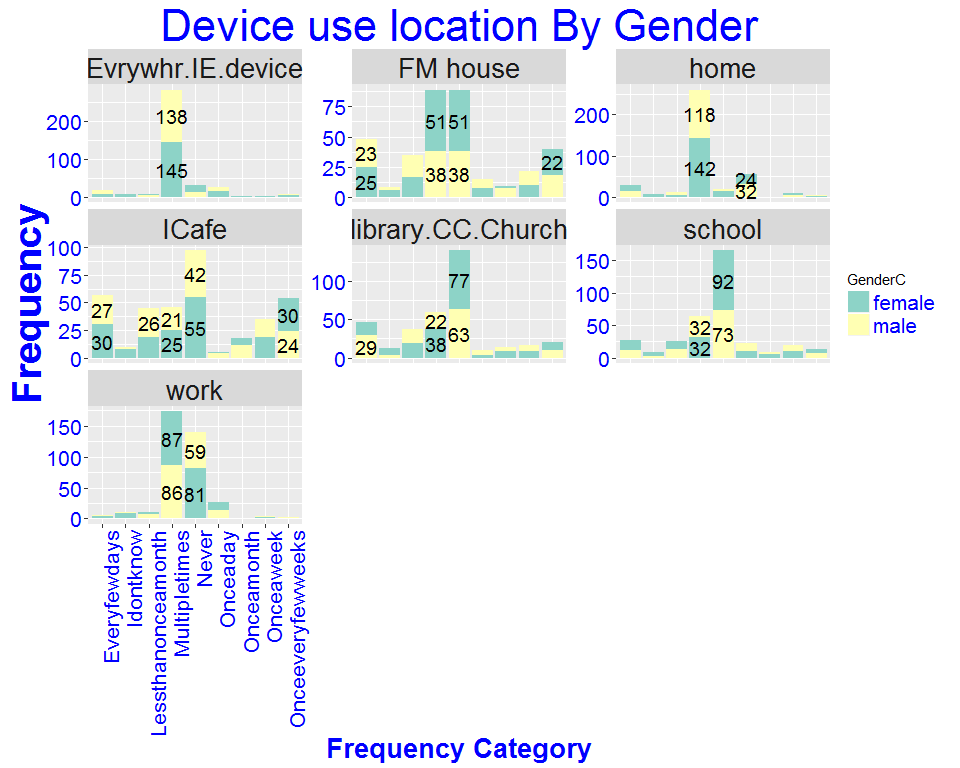
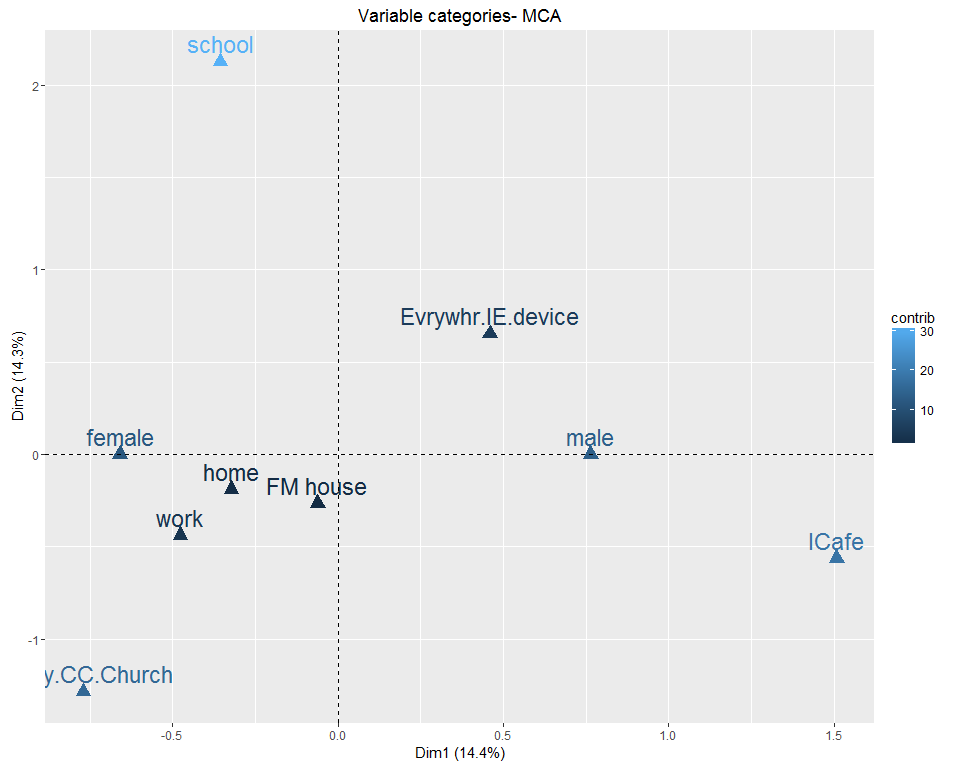
## [1] "the expected counts are much lower and hence the warning"

## [1] "Use Monte-Carlo simulation test"

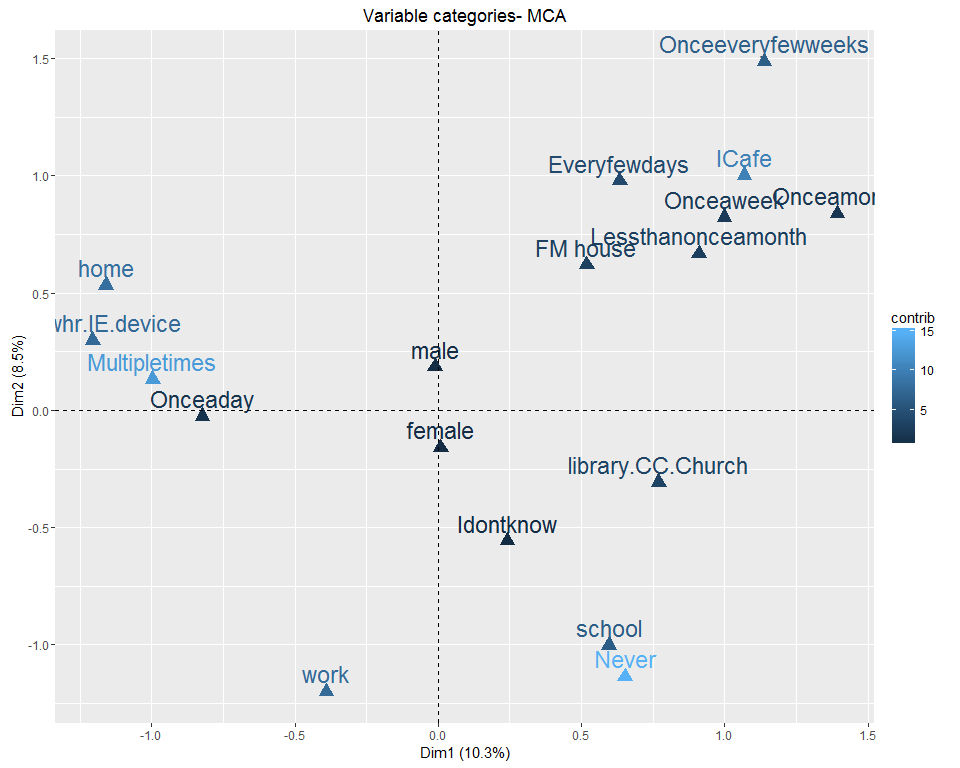
##   
## Chi-squared test for given probabilities with simulated p-value  
## (based on 2000 replicates)  
##   
## data: dev.use.tab  
## X-squared = 4034.977, df = NA, p-value = 0.0004998

## [1] "Check the associations along with the devices"

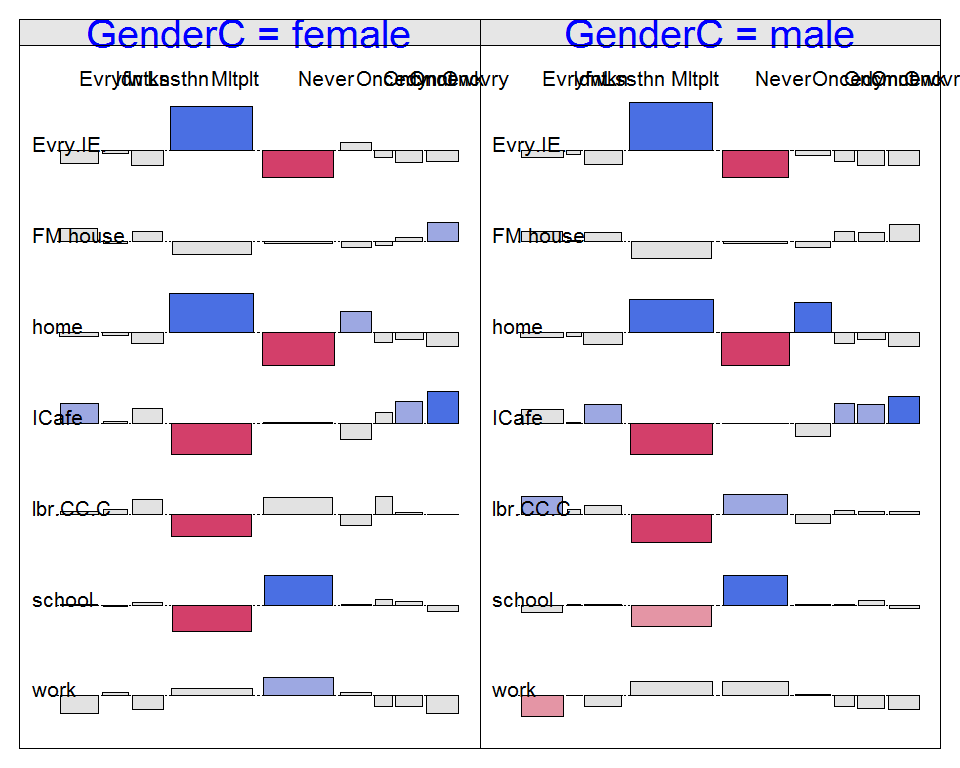
 #### 3. Where do you most commonly use the devices you have access to

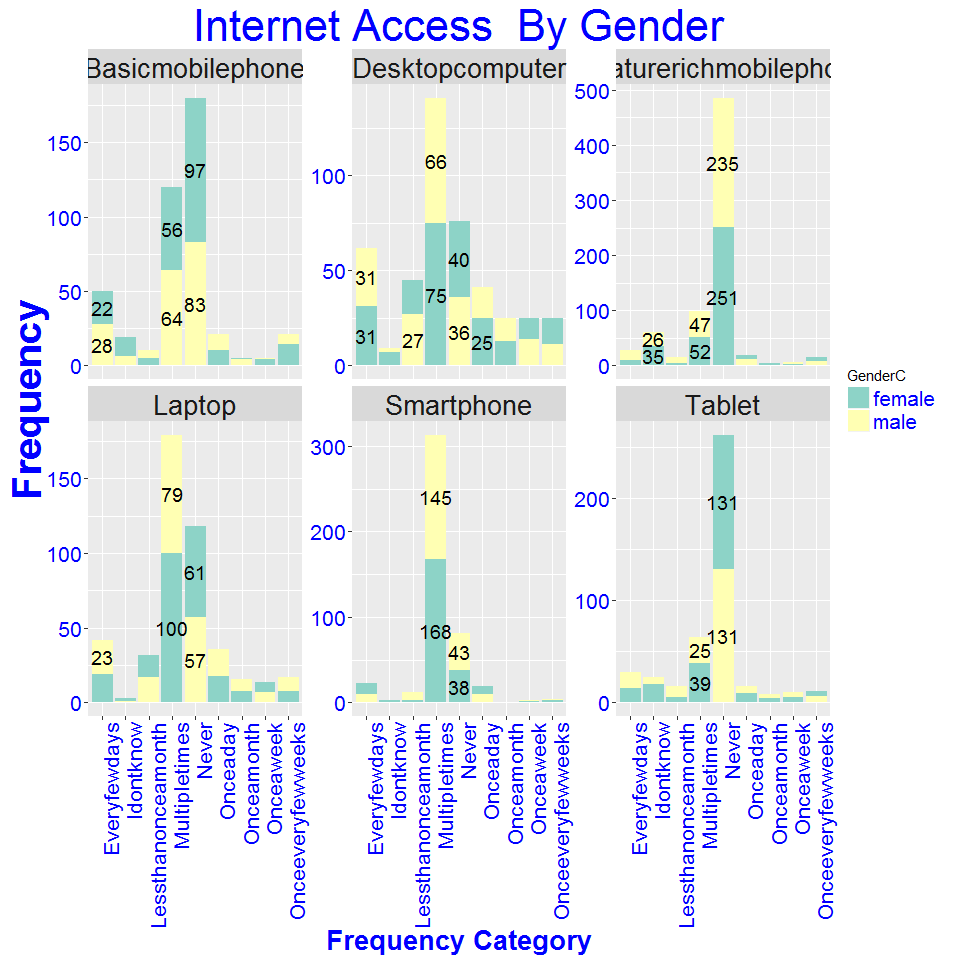
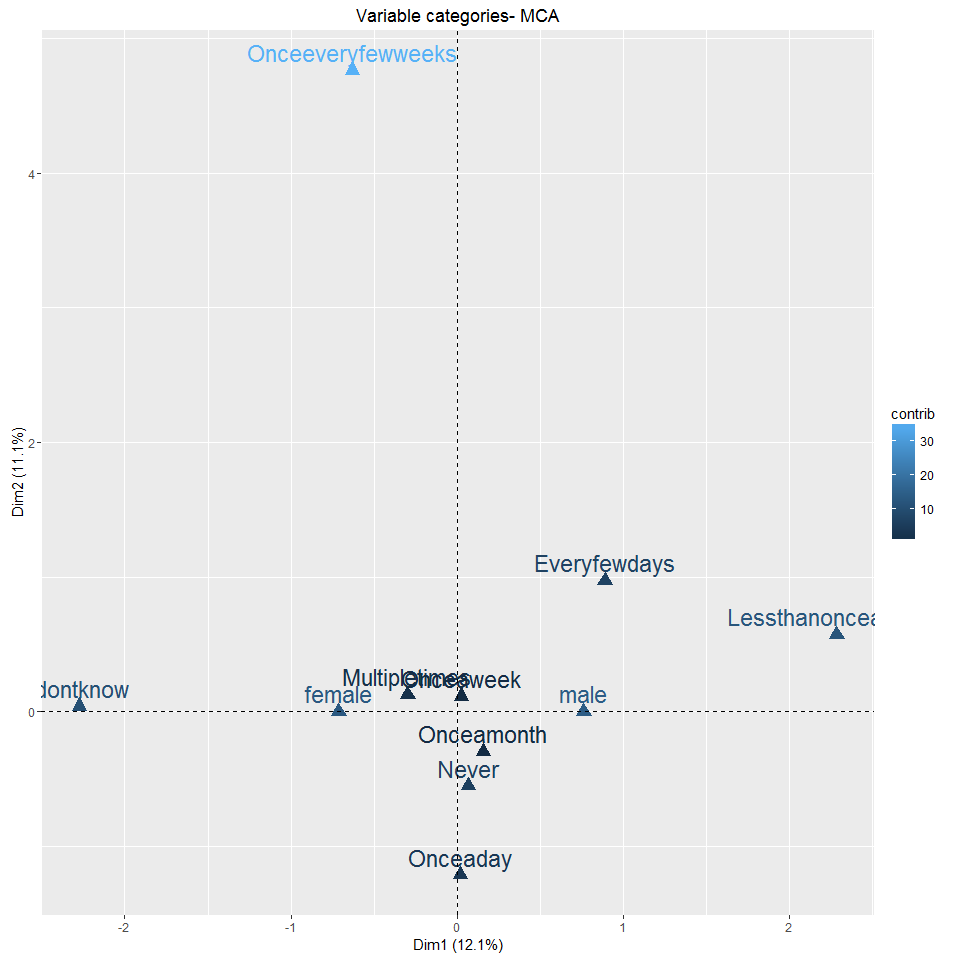
## Call: xtabs(formula = ~use.at + GenderC, data = dot.gen.dev.loc.dt)  
## Number of cases in table: 2582   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 0.13973, df = 6, p-value = 0.9999



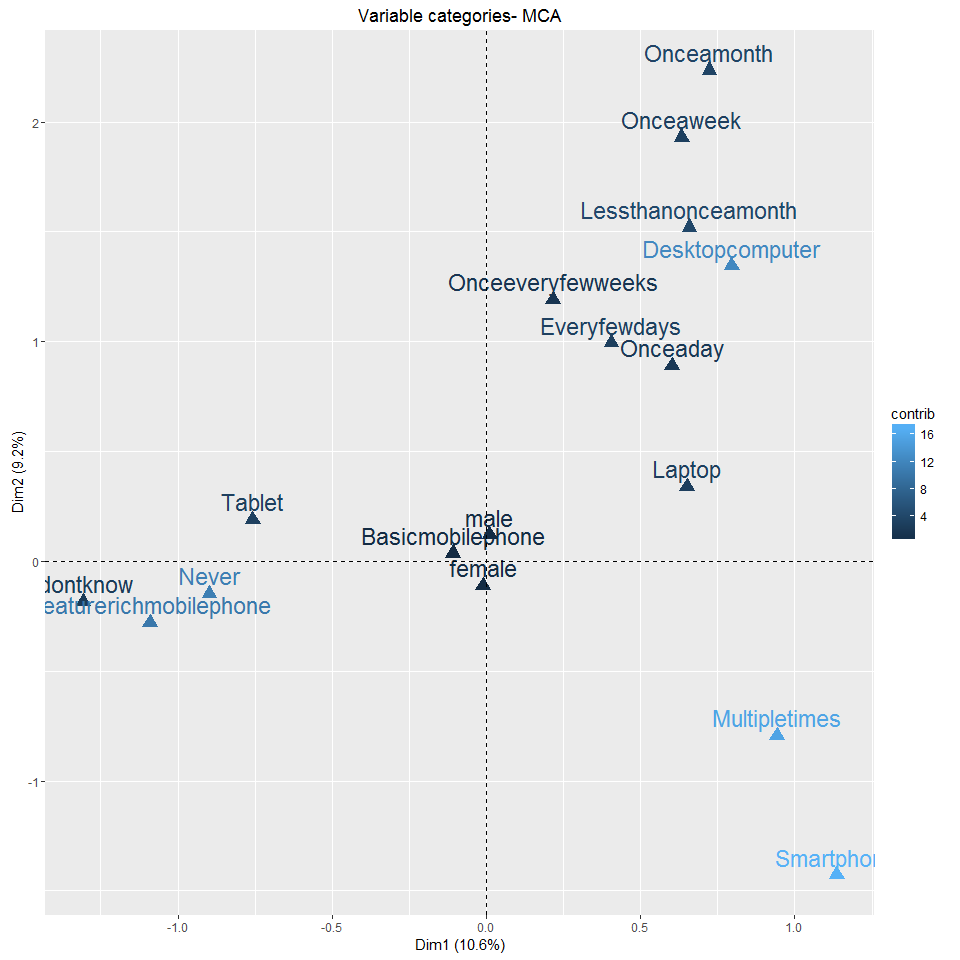
## Call: xtabs(formula = ~use.at + location.use.freq + GenderC, data = dot.gen.dev.loc.dt)  
## Number of cases in table: 2582   
## Number of factors: 3   
## Test for independence of all factors:  
## Chisq = 1064.4, df = 110, p-value = 5.864e-156  
## Chi-squared approximation may be incorrect



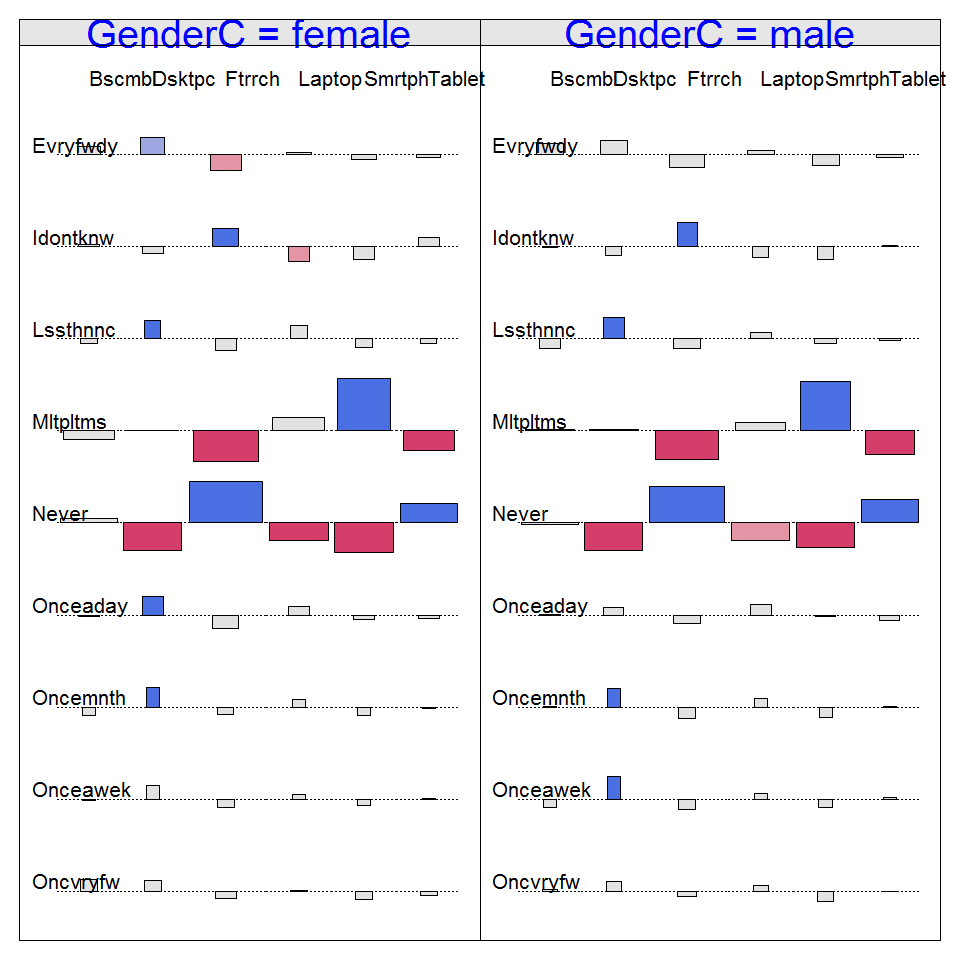
#### 5.How much access to do you have to the Internet?

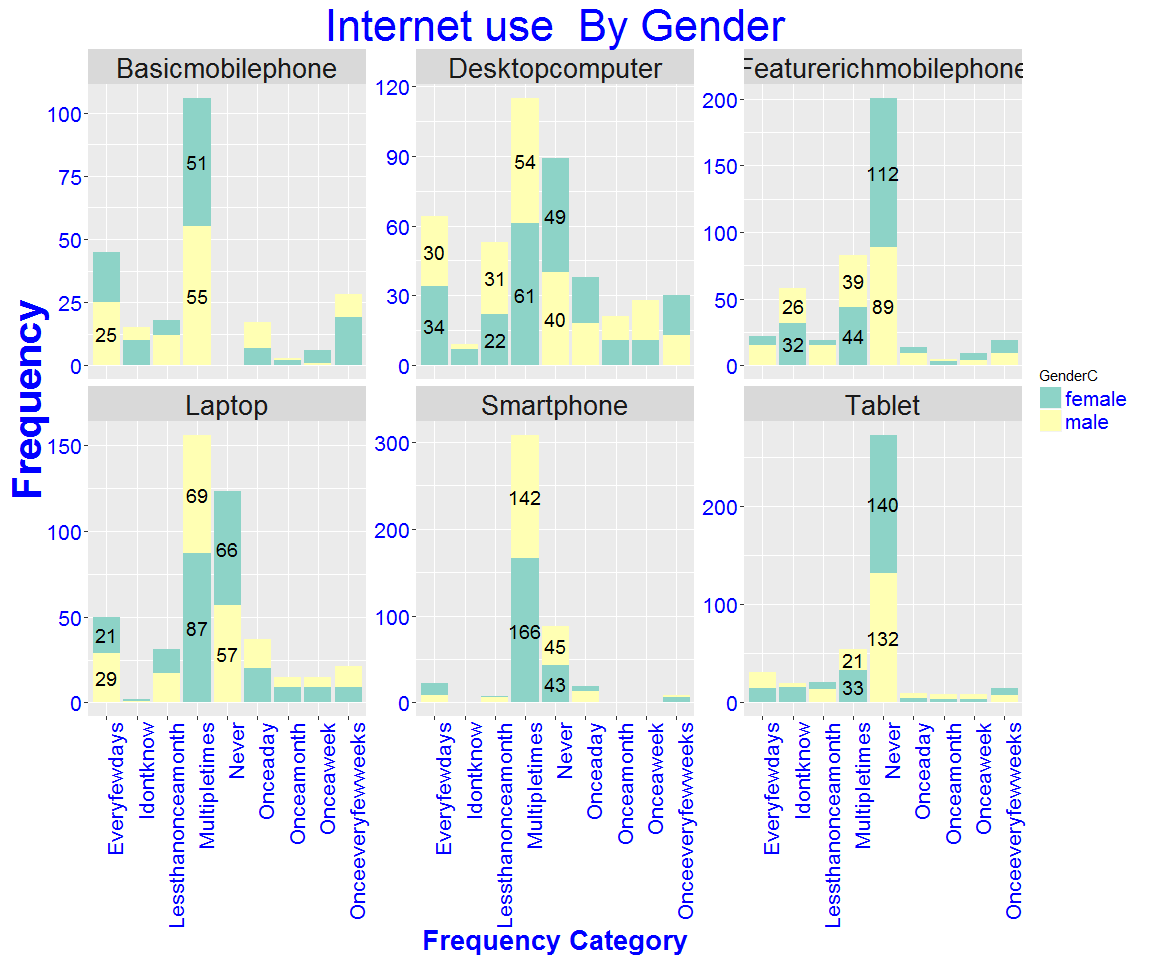
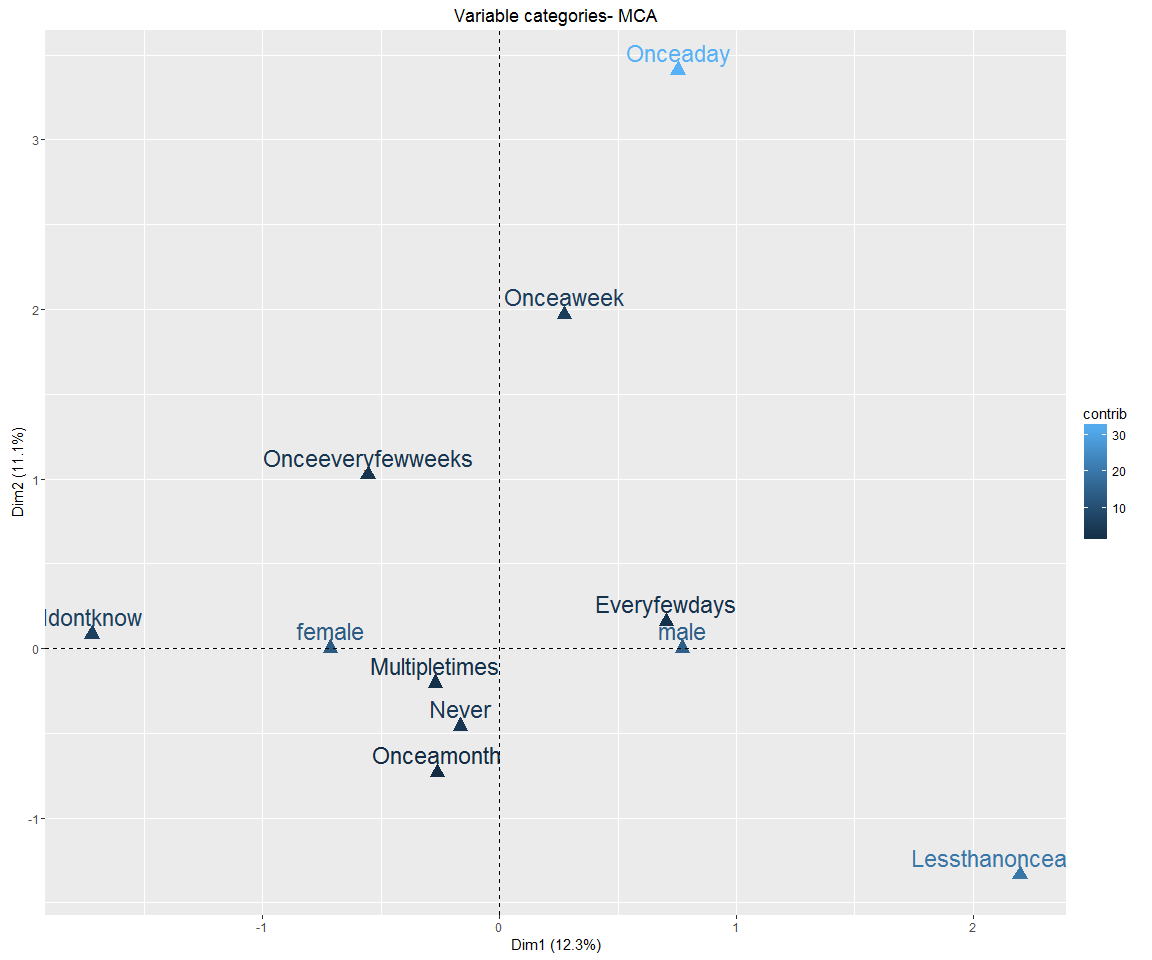
## Call: xtabs(formula = ~inrnt.dev.freq + GenderC, data = dot.gen.acc.intrnt.dt)  
## Number of cases in table: 2967   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 21.839, df = 8, p-value = 0.005222



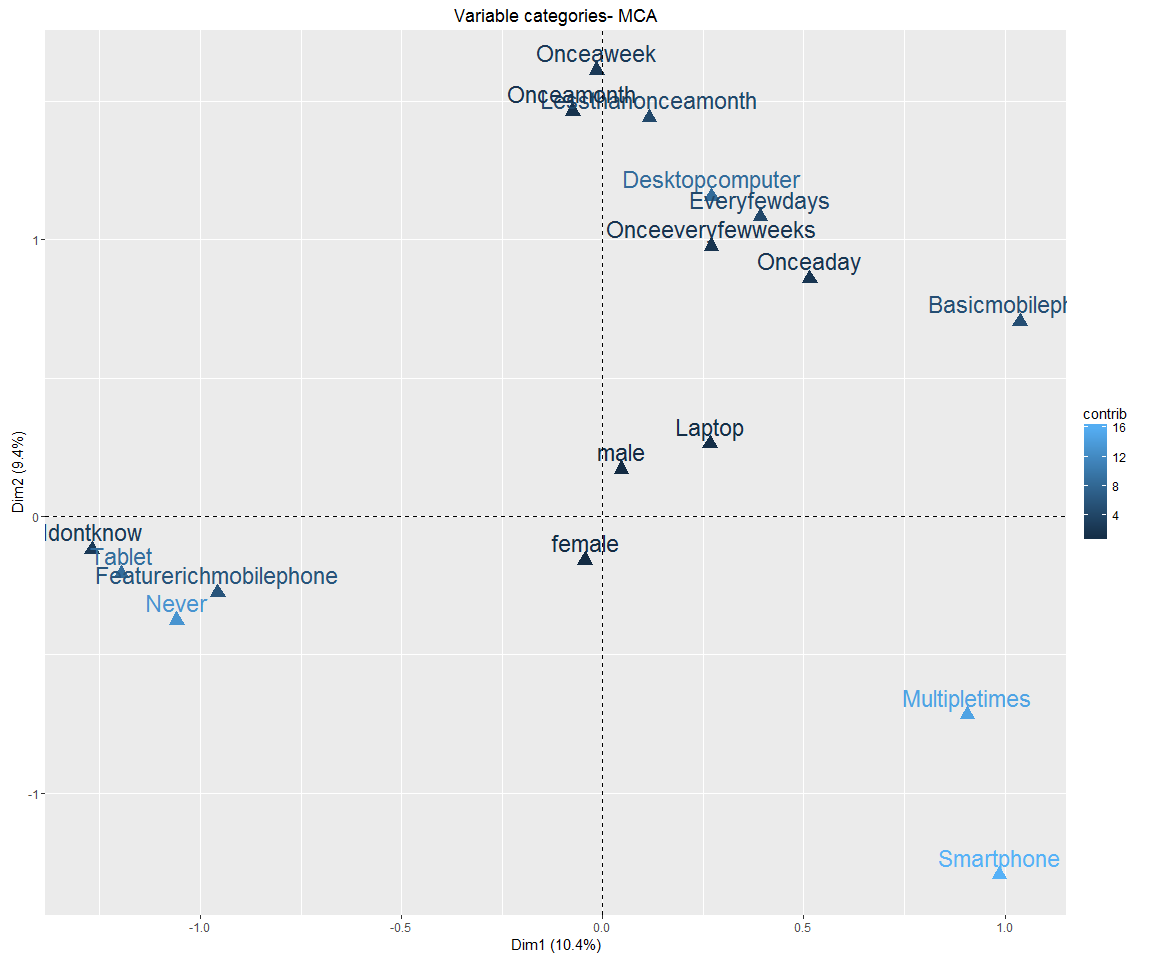
## Call: xtabs(formula = ~inrnt.dev.freq + inrnt.on.device + GenderC,   
## data = dot.gen.acc.intrnt.dt)  
## Number of cases in table: 2967   
## Number of factors: 3   
## Test for independence of all factors:  
## Chisq = 1004.7, df = 93, p-value = 7.321e-153  
## Chi-squared approximation may be incorrect



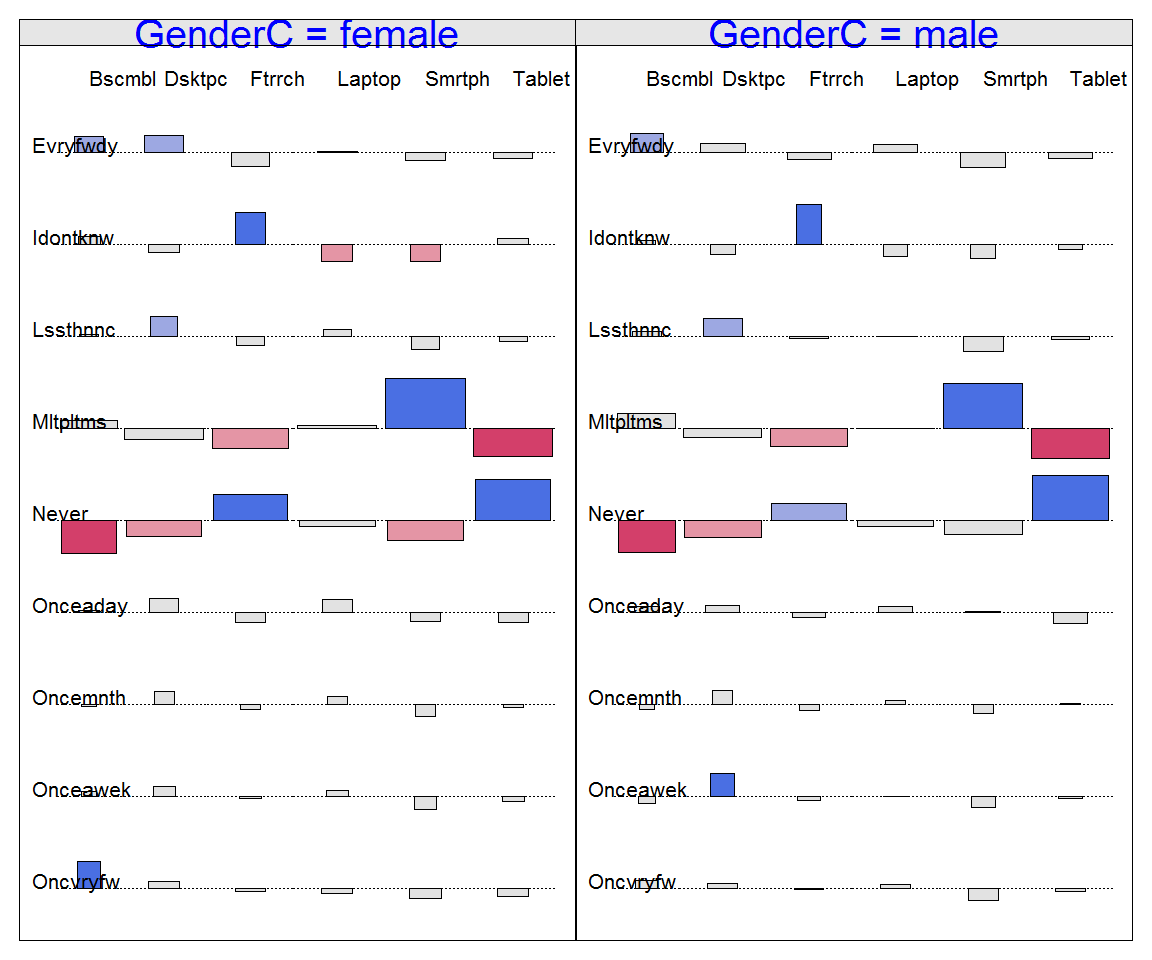
#### 6.How often do you actually use the Internet?

## Call: xtabs(formula = ~intrnt.on.dev.use.freq + GenderC, data = dot.gen.use.intrnt.dt)  
## Number of cases in table: 2463   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 26.076, df = 8, p-value = 0.001019

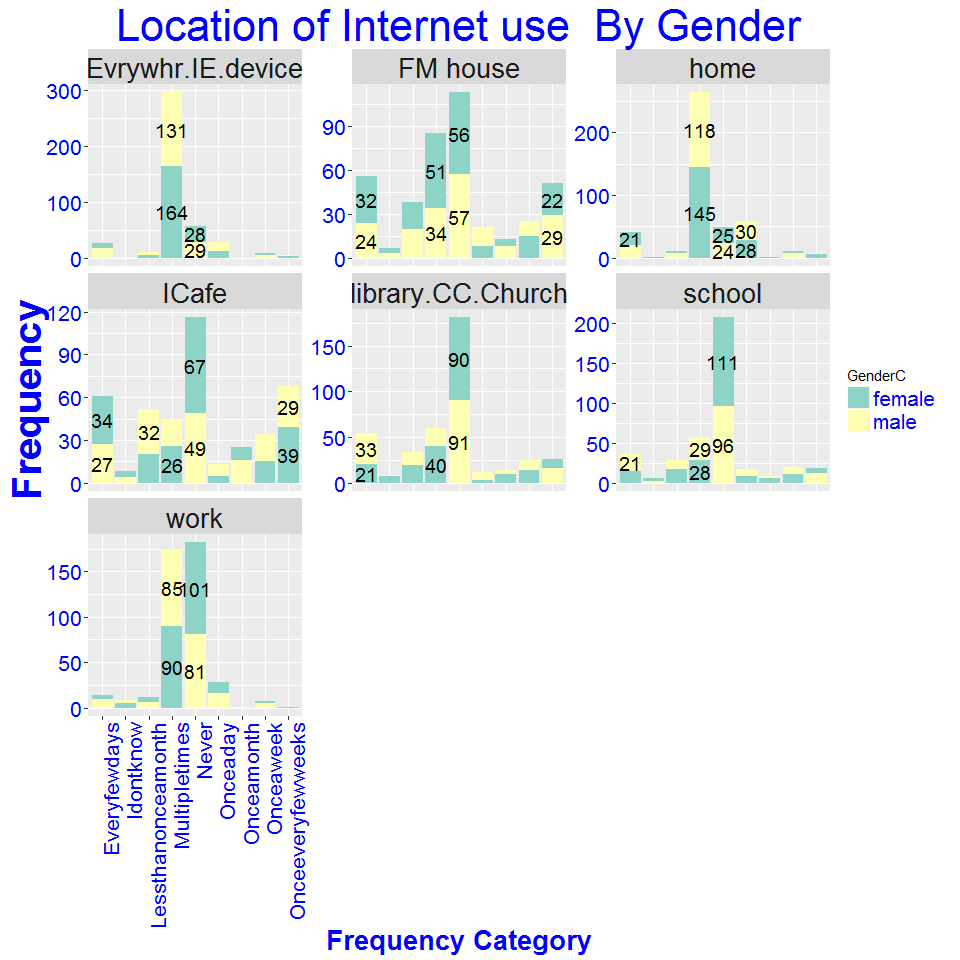


## Call: xtabs(formula = ~intrnt.on.dev.use.freq + intrnt.use.on.dev +   
## GenderC, data = dot.gen.use.intrnt.dt)  
## Number of cases in table: 2463   
## Number of factors: 3   
## Test for independence of all factors:  
## Chisq = 948.2, df = 93, p-value = 9.728e-142  
## Chi-squared approximation may be incorrect

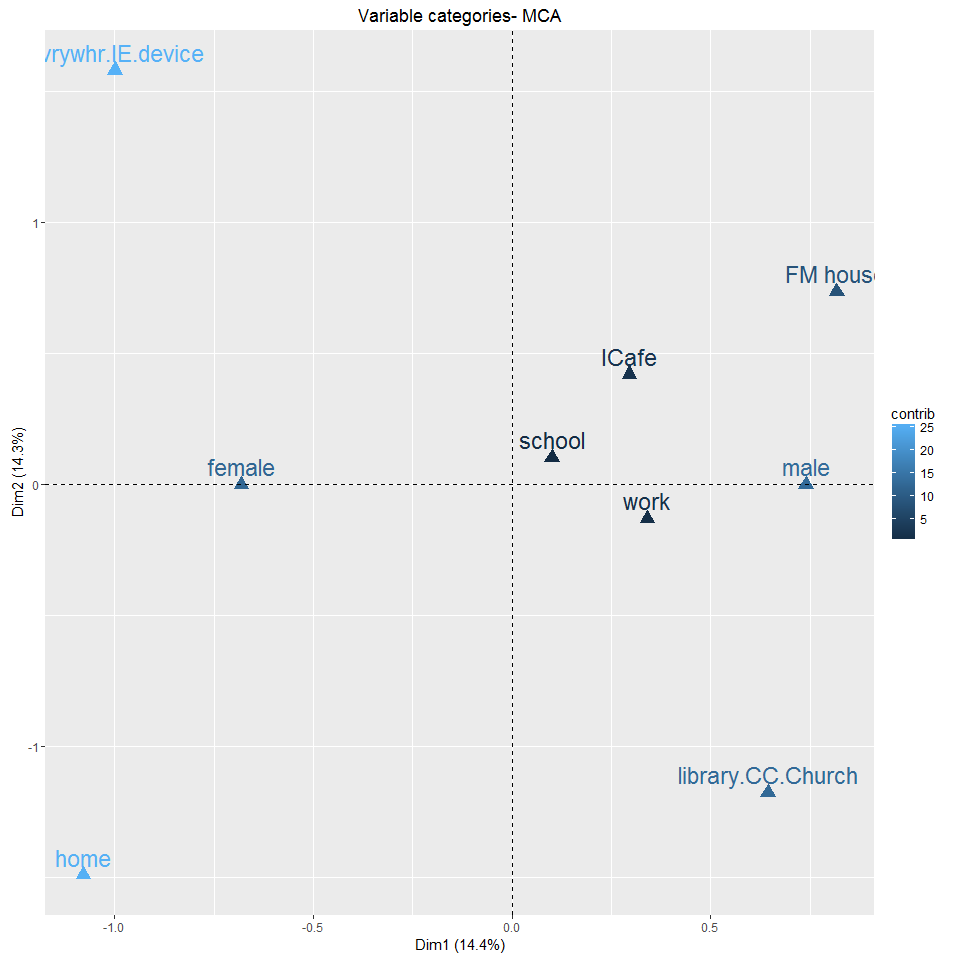


#### 7.Where do you most commonly use the Internet?

plot.tech.data(dot.gen.use.intrnt.loc.dt, "intrnt.use.loc", "intrnt.use.loc.freq", "Location of Internet use By Gender")



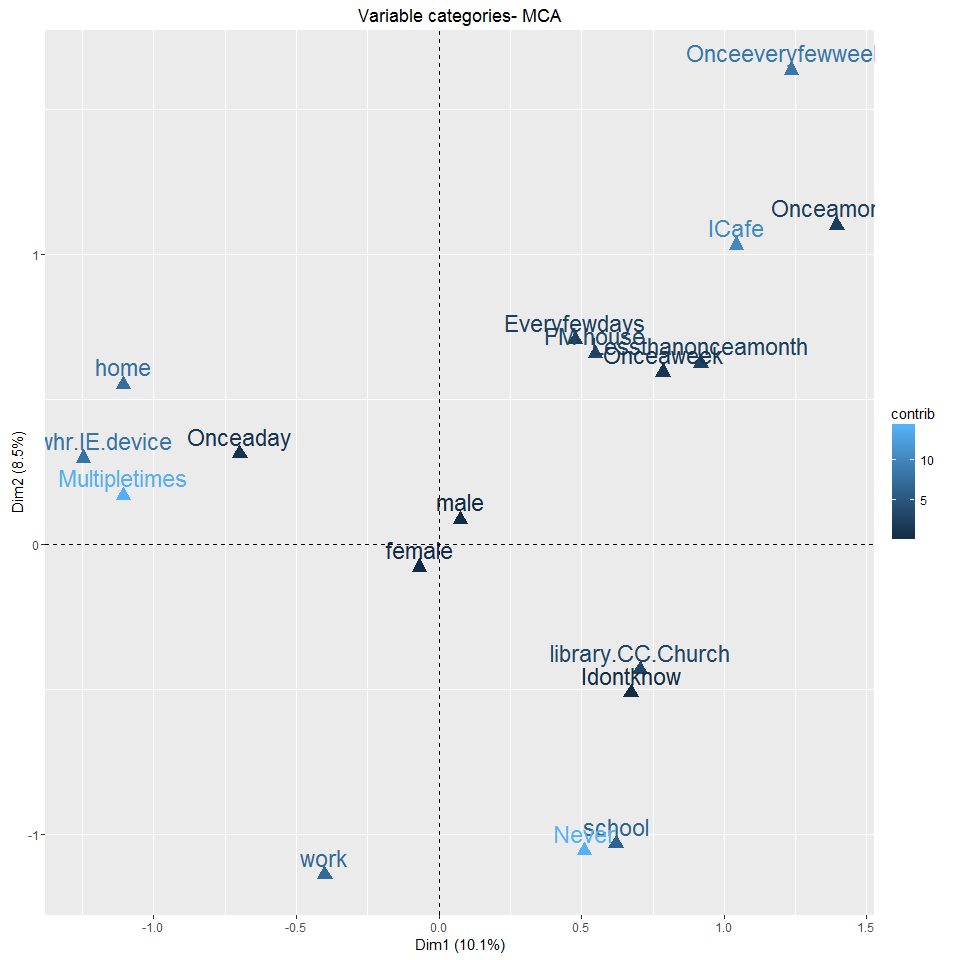
mca.plot(dot.gen.use.intrnt.loc.dt[,list(as.factor(GenderC), as.factor(intrnt.use.loc)),])



dev.use.intrnt.loc.tab<-xtabs(~intrnt.use.loc + GenderC , data = dot.gen.use.intrnt.loc.dt)  
#   
summary(dev.use.intrnt.loc.tab)

## Call: xtabs(formula = ~intrnt.use.loc + GenderC, data = dot.gen.use.intrnt.loc.dt)  
## Number of cases in table: 2948   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 0.20513, df = 6, p-value = 0.9998

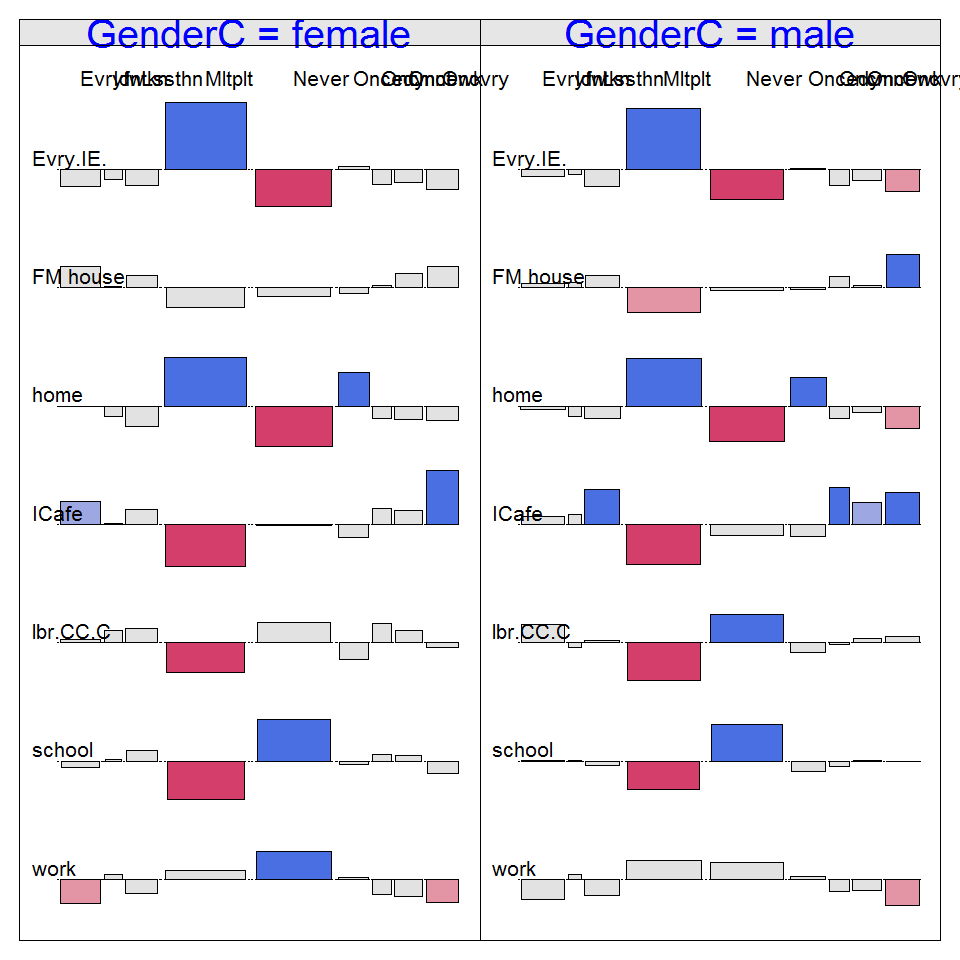
mca.plot(dot.gen.use.intrnt.loc.dt[,list(as.factor(GenderC), as.factor(intrnt.use.loc),as.factor(intrnt.use.loc.freq)) ,])



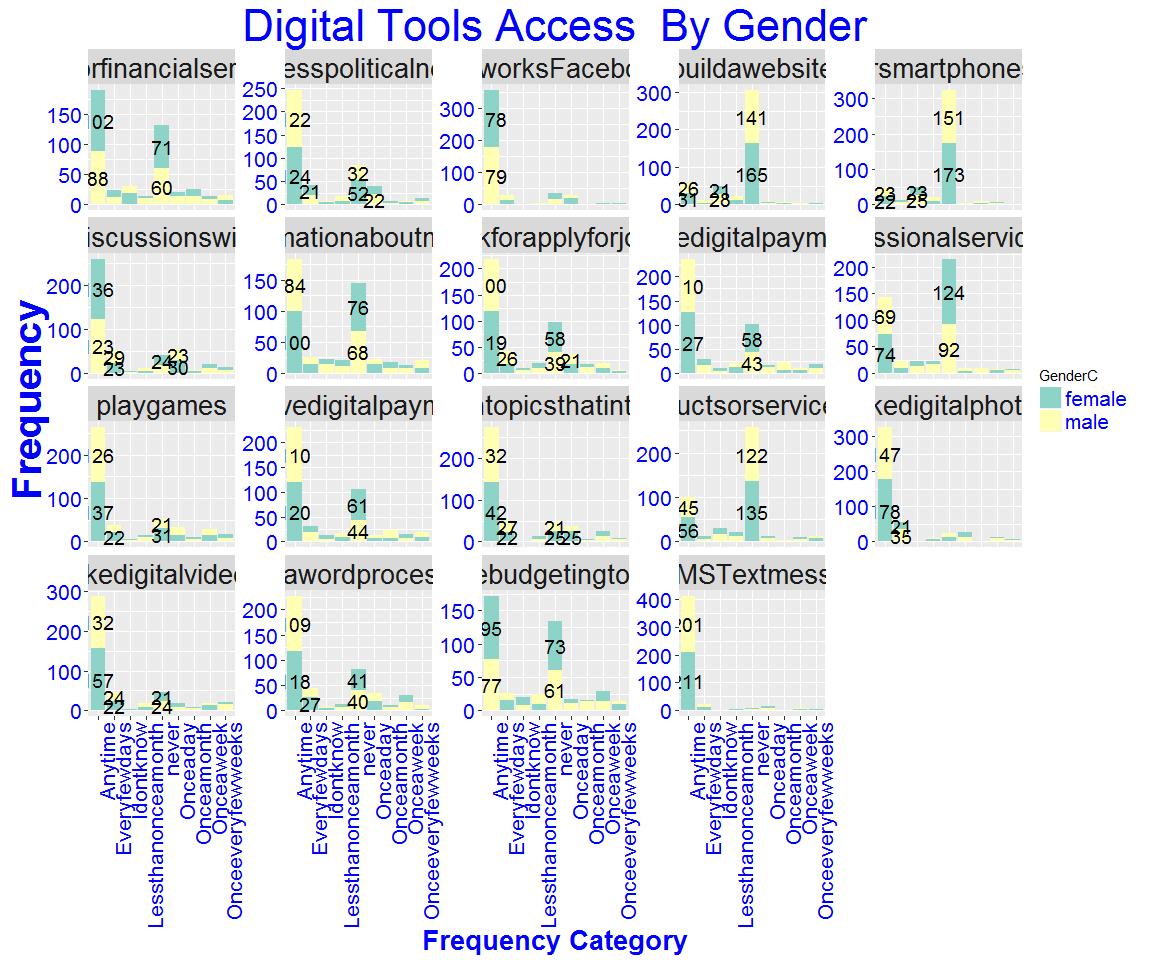
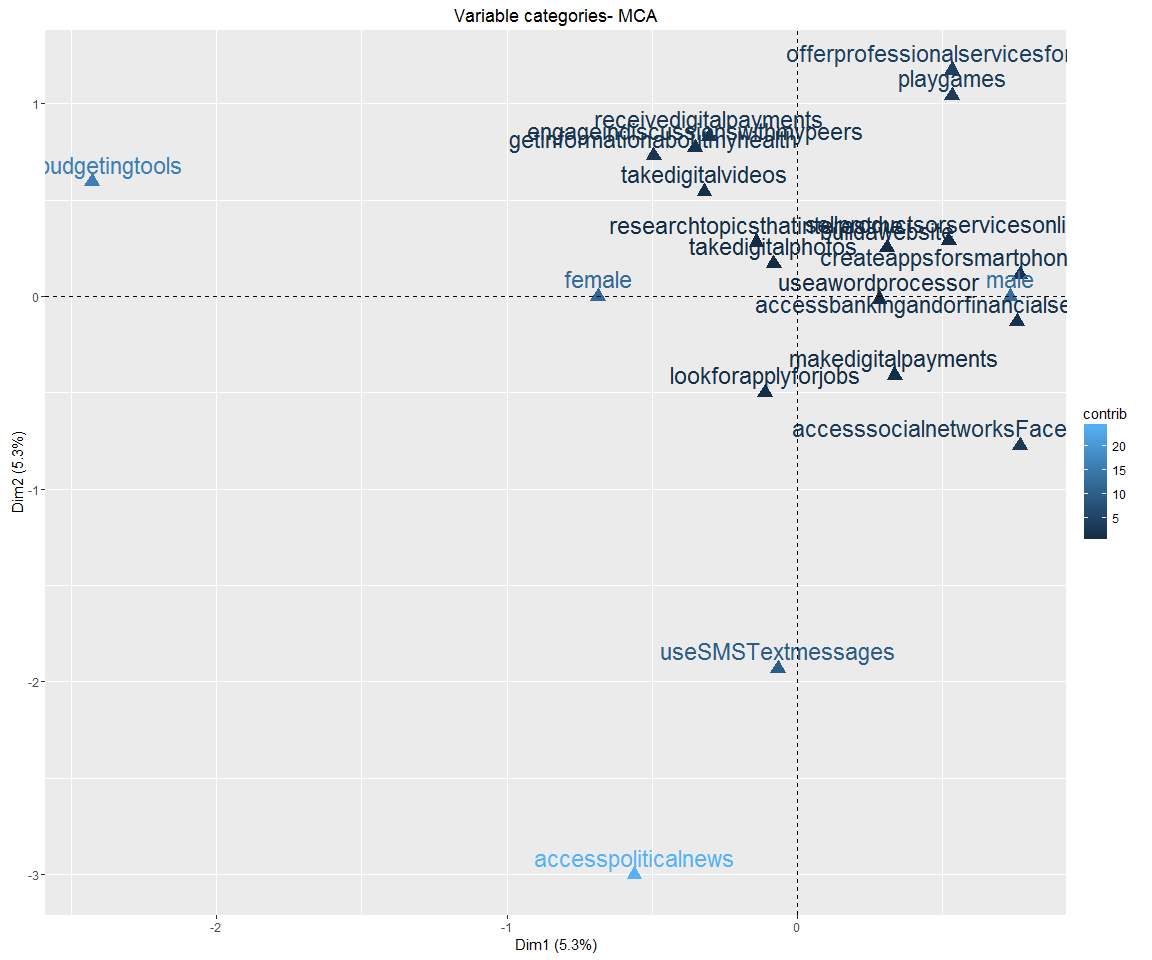
set.seed(2879)  
  
pLocInt<-xtabs(~intrnt.use.loc + intrnt.use.loc.freq + GenderC,data = dot.gen.use.intrnt.loc.dt)  
  
summary(pUInt)

## Call: xtabs(formula = ~intrnt.on.dev.use.freq + intrnt.use.on.dev +   
## GenderC, data = dot.gen.use.intrnt.dt)  
## Number of cases in table: 2463   
## Number of factors: 3   
## Test for independence of all factors:  
## Chisq = 948.2, df = 93, p-value = 9.728e-142  
## Chi-squared approximation may be incorrect

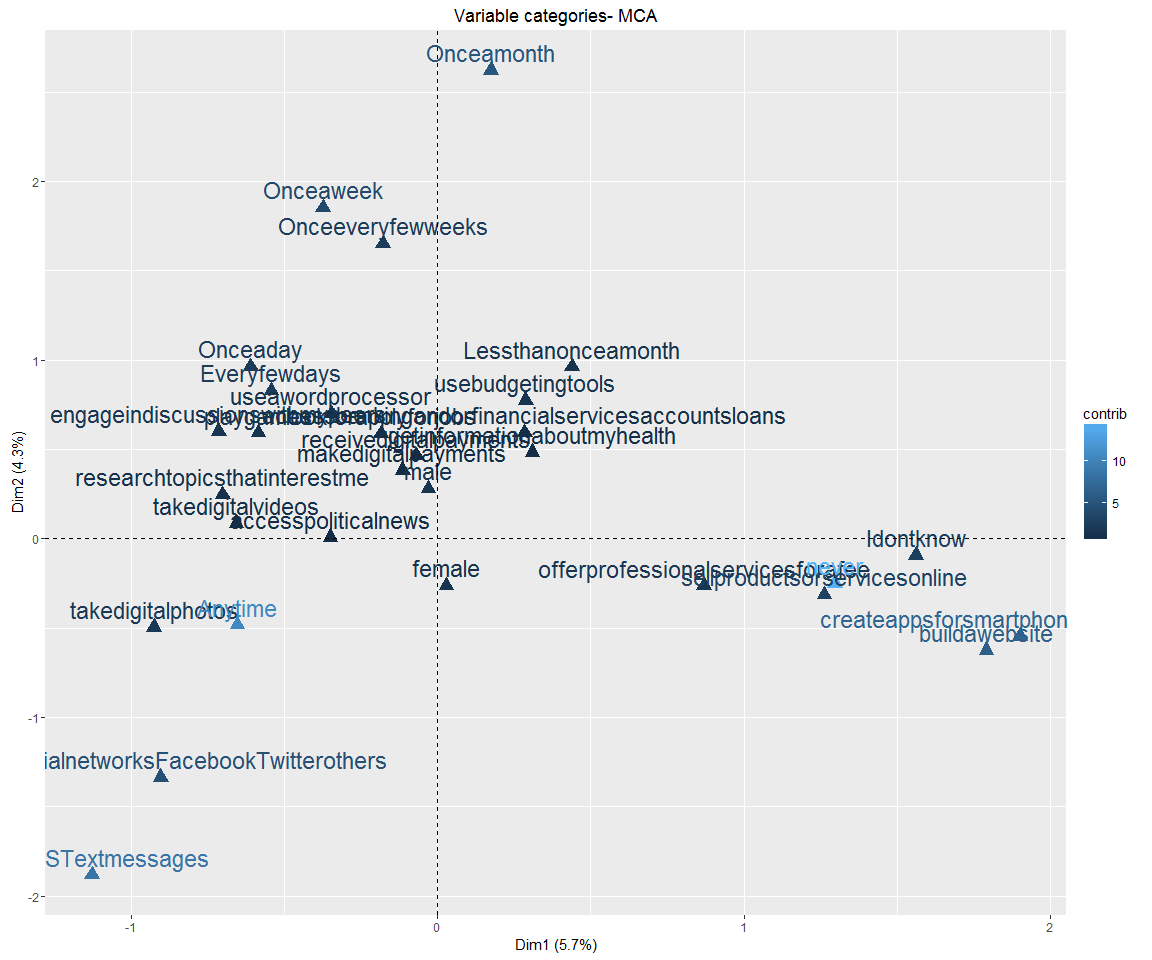
cotabplot(pLocInt, panel = cotab\_coindep, shade = TRUE,legend = FALSE,   
 panel\_args = list(type = "assoc", margins = c(3,1,1,2),   
 varnames = FALSE),las = 2,  
 text\_gp = gpar(col = "blue",fontsize = 30),#increase title font  
 labeling\_args = list(abbreviate\_labs = c(8,6,1,1),  
 rot\_labels = c(0,0,90,0),  
 just\_labels=c("left","left","left","left")),  
 gp\_labels = gpar( fontsize = 16),pop = T)



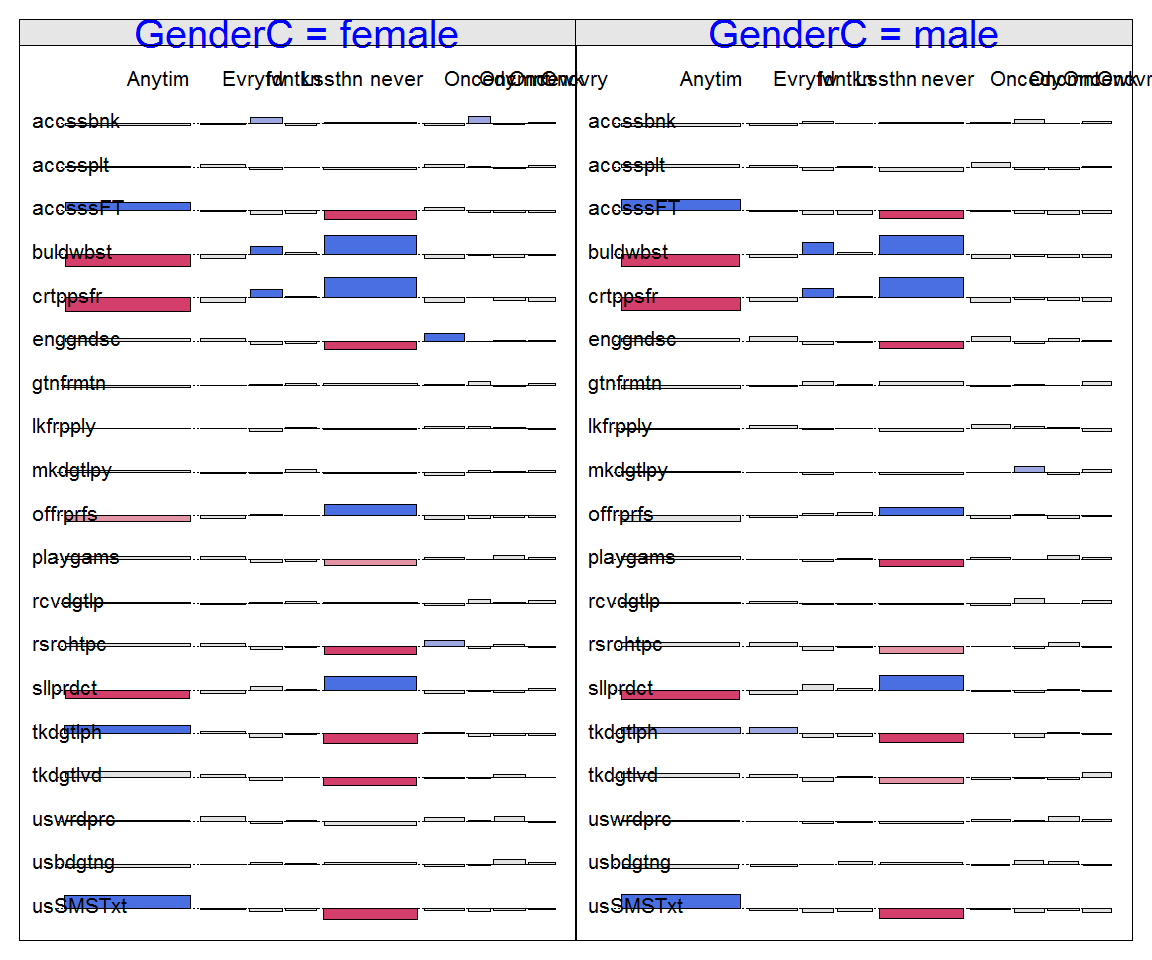
#### 8.What digital services/tools do you have access to?

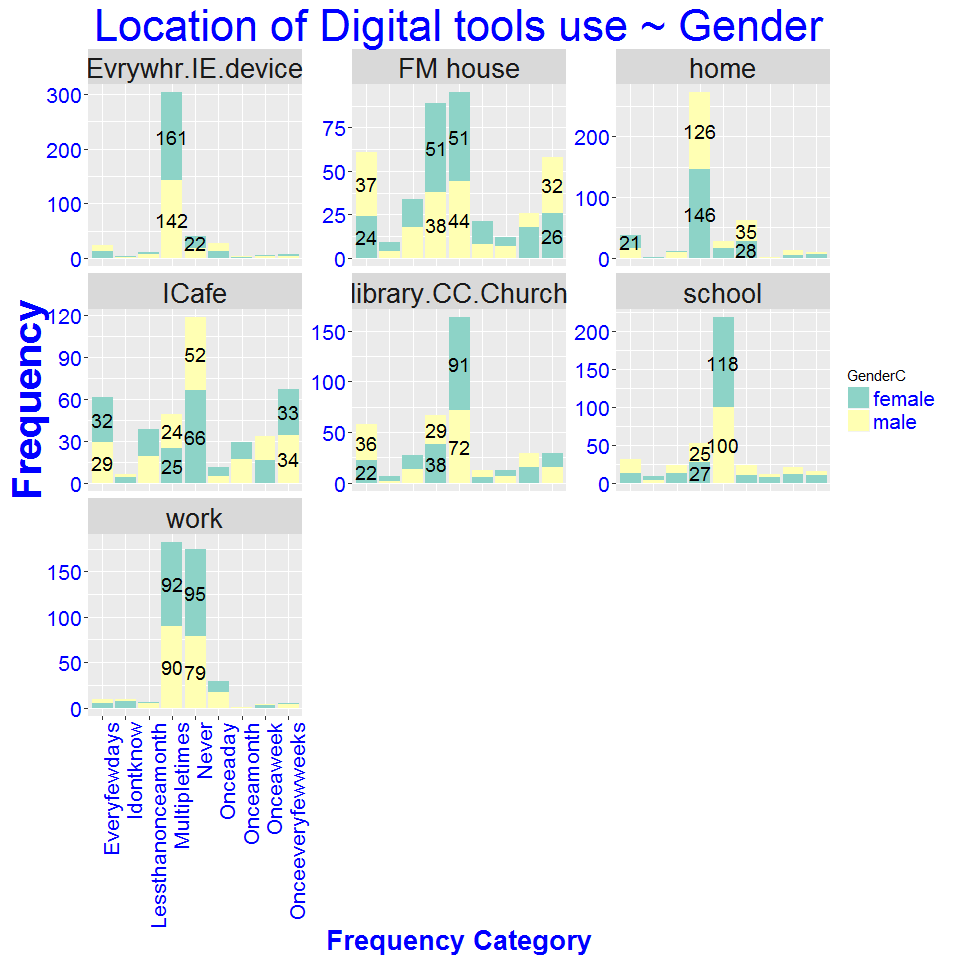
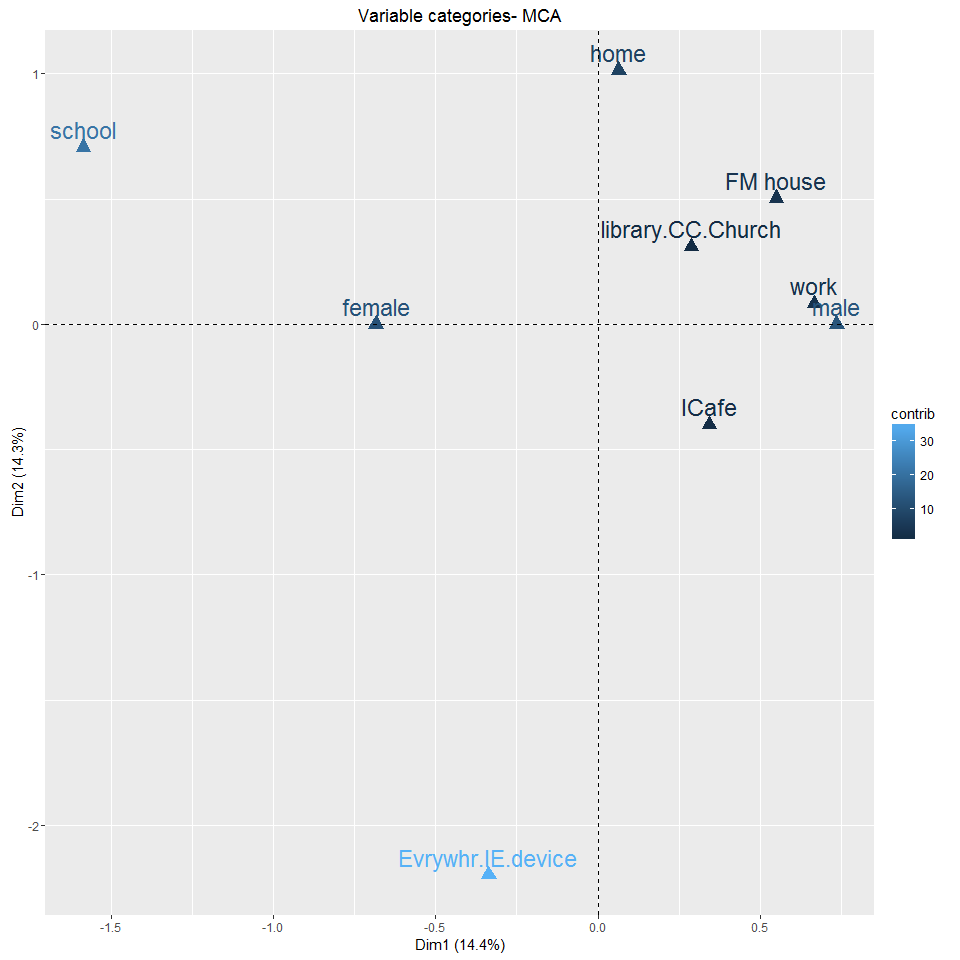
## Call: xtabs(formula = ~tool.ser.dig + GenderC, data = dot.gen.acc.digi.dt)  
## Number of cases in table: 8808   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 0.4383, df = 18, p-value = 1



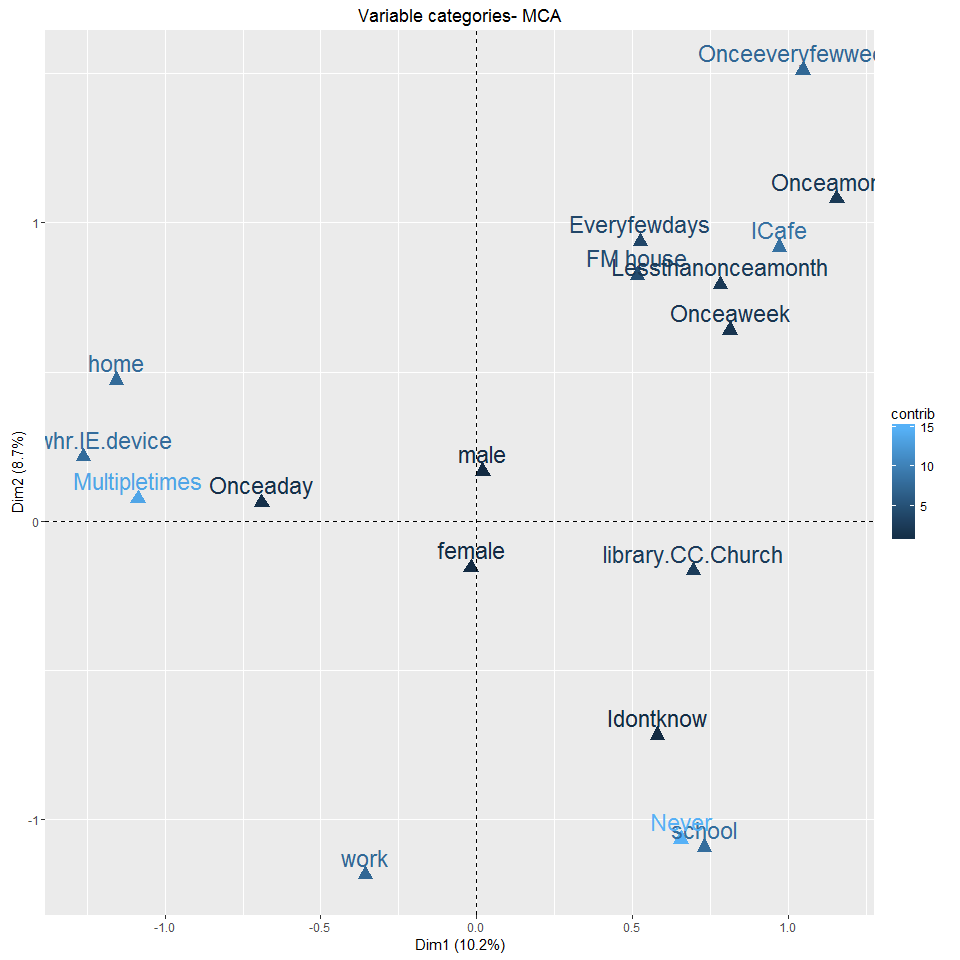
## Call: xtabs(formula = ~tool.ser.dig + frq.ac.dig + GenderC, data = dot.gen.acc.digi.dt)  
## Number of cases in table: 8808   
## Number of factors: 3   
## Test for independence of all factors:  
## Chisq = 3083.9, df = 314, p-value = 0  
## Chi-squared approximation may be incorrect



#### 9.Where do you most commonly use digital services?\*\*

## Call: xtabs(formula = ~digi.loc.of.use + GenderC, data = dot.gen.loc.digi.dt)  
## Number of cases in table: 2910   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 0.12226, df = 6, p-value = 1



## Call: xtabs(formula = ~digi.loc.of.use + at.loc.digi.use.freq + GenderC,   
## data = dot.gen.loc.digi.dt)  
## Number of cases in table: 2910   
## Number of factors: 3   
## Test for independence of all factors:  
## Chisq = 1192.9, df = 110, p-value = 3.36e-181  
## Chi-squared approximation may be incorrect

