DOT Survey Data Analysis

UshaKiran.Kota

May 23, 2016

1.Coursolve/Need230 : DOT Data Analysis Digital Internship

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.

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 567 responses for over 1500 collection of questions. 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 and quantitative Analysis and expects the collaborators to find answers for some or all of the below research questions

1.1 Coursolve Need230/DOT Data Analysis - Research Questions

As part of need230, DOT is interested in the analysis of following research questions using DOT's survey data

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?

1.2 Document Purpose

The purpose of this document is to present a summary of Data analysis and visualizations performed on sample questions (**use cases**) of DOT survey data to provide useful insights for future research.

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

1.3 Justification

DOT intends to learn more about how 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 from Need230 to inclue them as part of program and project design, and as well as in their implementation strategies.

2.Executive Summary

As part of Coursolve 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 Youth in the selected countries on how they leverage on technology for their professional and entreprenueral growth. Survey data is analysed with the help of Exploratory views, Qualitative comments and appropriate quantitative methods

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 social status.

Since the data are nominal categorical in nature with multi-level outcomes that vary across sample sizes, it will be appropriate to assume that "point estimates" (mean , equal variance), and ranges for summary statistics vary a lot and estimates of normal distribution do not hold. Since the study is observational, showing any associations within the response and explanatory variables is considered sufficient.

Exploratory methods such as stacked bar plots, mosaic plots and contingency tables are used visualize the data and any associations within groups or between groups of samples. The methods employed for analysis include useful control variables such as *Gender* , *Age* and *Location* of the respondent wherever they are appropriate

Quanitative analysis is conducted primarily for test of independence from the control variable.Since the sample sizes with multi-level responses are large enough, **Pearson's chi-squared test** is a suitable choice for test for independence between the Gender of respondents and their response for technology access or use.

Comparing row and column proportions, computing contingency ratios are also assumed to be applicable tests and effective incase of response variables with nominal outcomes.

Research Questions 3 and 5 elicit response in the form of "open ended text". Exploratory views such as Wordclouds and Latent Topics are included as part of the text analysis tools in these cases.

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

3. Survey Data Insights

DOT team provided the collaborators on Coursolve need230 a clean data set with the response captured by their research team.This section summarises the understanding of Survey Data

3.1.Nature of Survey

DOT data is assumed to be a propsective observational study conducted with structured sampling strategies.

3.2 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. 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

3.3. Assumptions on Conditions

The following assumptions are made on conditions for independence

3.3.1 Observations are independent.

The survey is conducted a stratified simple random sample and consists of fewer than 10% of the youth population, which establises independence,both within the data and between each subsets of data.

3.3.2 Success-failure condition.

Although the sample size is failry large data of 567 cases, since response data has multiple outcomes, the distribution of outcomes is expected to follow a multinomial distribution.

3.4 Data Handling

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

3.4.1 Missing Data

DOT data set had fewer (19) NA cases w.r.t demographic features. These were dropped from the data because such data would not be useful for analysis, The data among the survey features is treated as a large sparse matrix

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"
```

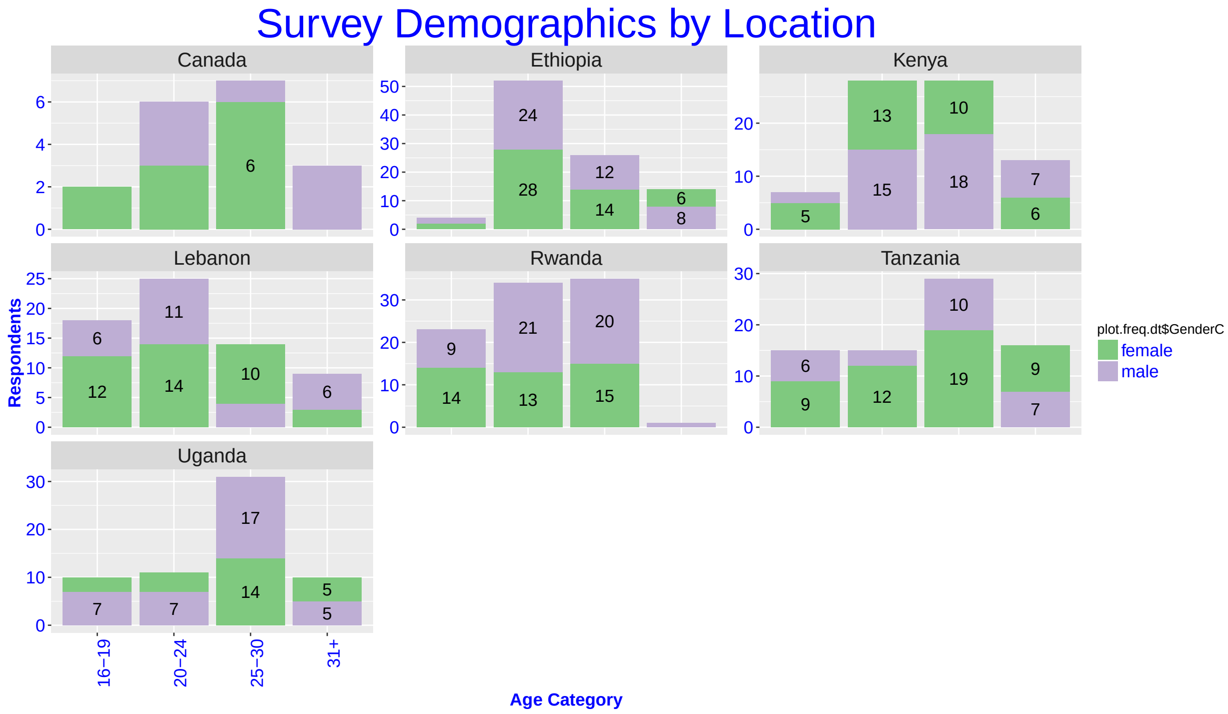
```
##          ID     Status Language Location WeightedScore
## 1: 43413695 Incomplete  English                     35
## 2: 43486387   Complete  English                     26
## 3: 43645338   Complete  English                   1063
## 4: 43675088   Complete  English                    961
## 5: 43749321   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
```

4.Survey Demographics

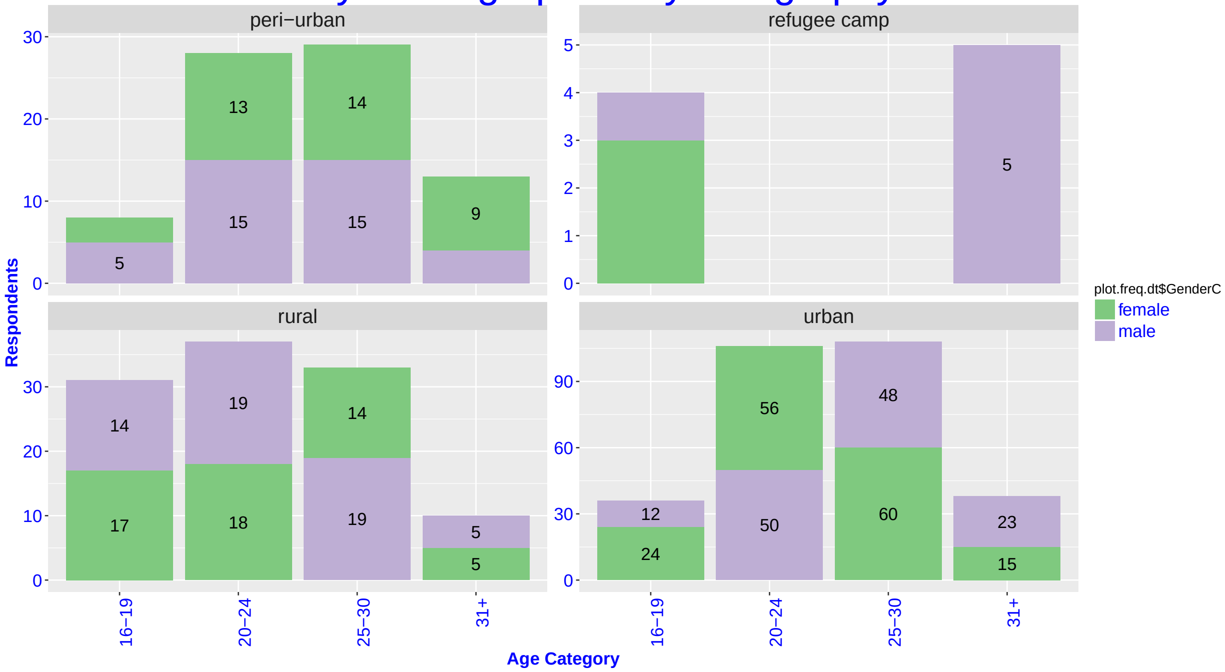
The sections below describe the respondents'demographics with help of appropriate plots.

4.1 Qualitative Analysis

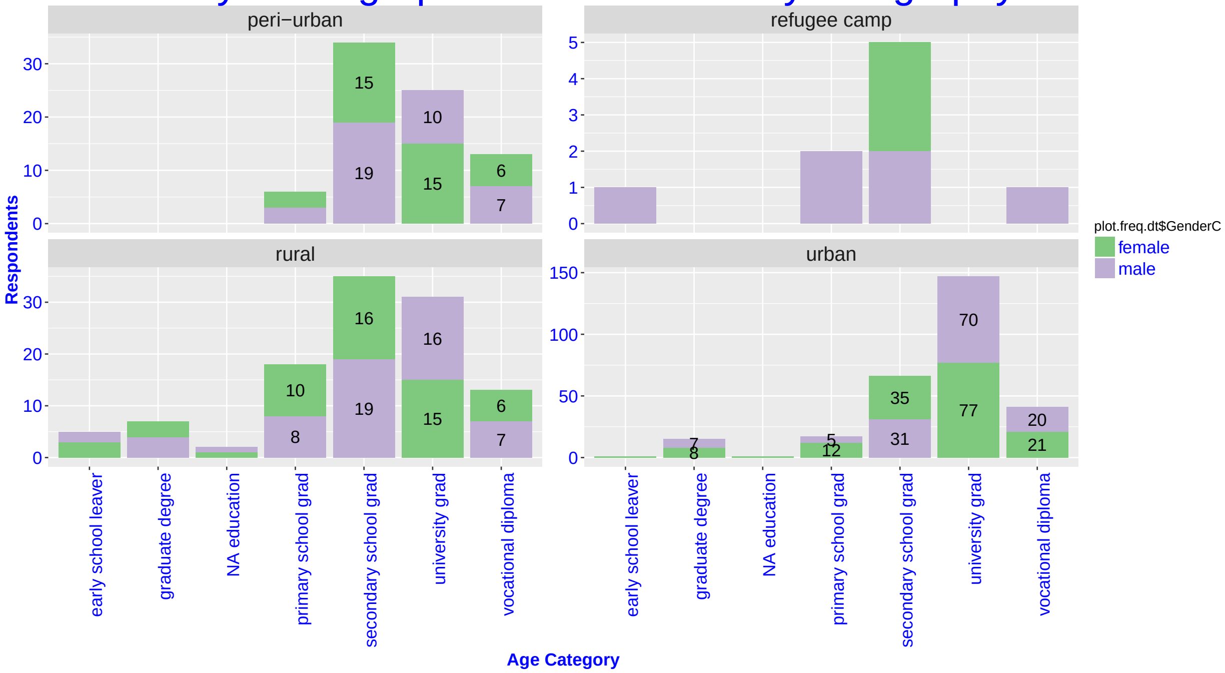
1. The proportion of female respodents is higher than males in most of the sampled locations, with Kenya and Rwanda being exceptions
2. The count of respondents from Canada is lower than rest of the countries
3. Respondents from Urban regions are surveyed more than rural and peri-urban
4. Respondents from Urban regions seem to be mostly university graduates, while from rural are mostly secondary school grads
5. University grad female students are also seen are more in proportions as compared to male stundent in the employed and unemployed segement.

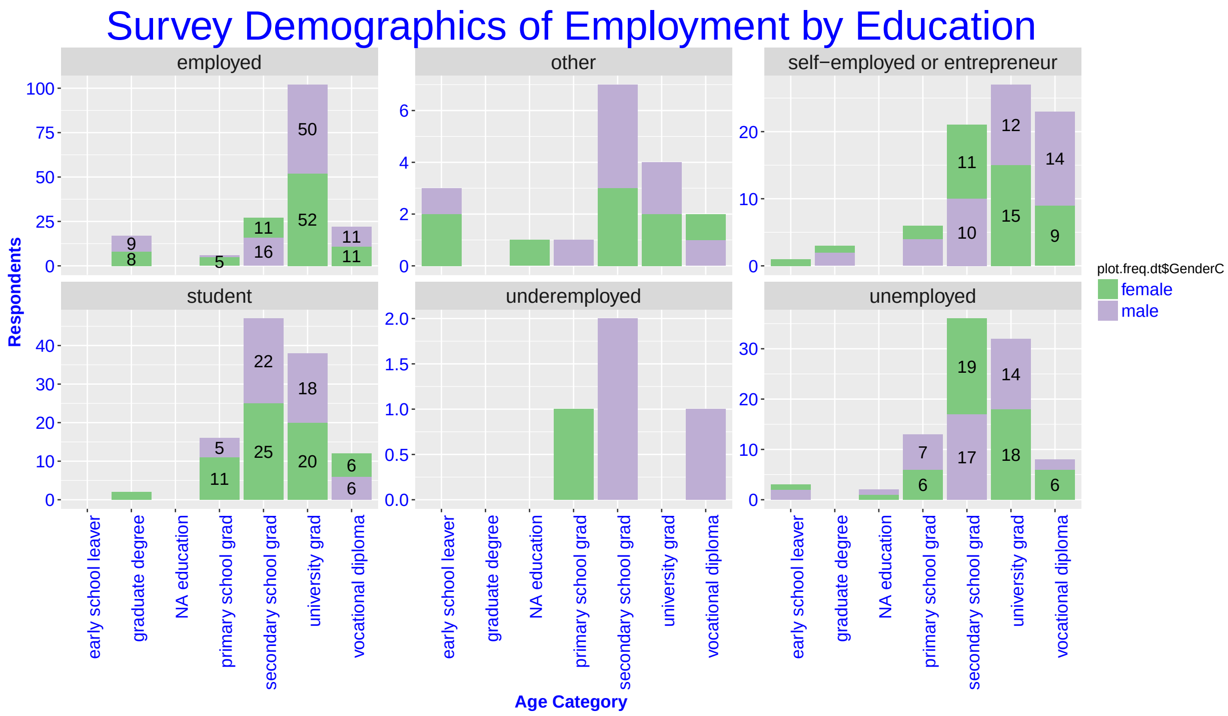


Survey Demographics by Geography



Survey Demographics of Education by Geography





```
## [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          0.01593625      0.04382470  0.007968127
##   male            0.01276596      0.04680851  0.004255319
##
##          primary school grad secondary school grad university grad
##   female          0.09960159             0.2749004       0.4262948
##   male            0.07659574             0.3021277       0.4085106
##
##          vocational diploma
##   female          0.1314741
##   male            0.1489362
## [1] "2-way Contingency table: varying by Gender and Employment status"
##
##           employed      other self-employed or entrepreneur   student
##   female 0.3466135 0.03585657                     0.1553785 0.2549801
##   male   0.3702128 0.03829787                     0.1787234 0.2170213
##
##          underemployed unemployed
##   female   0.003984064  0.2031873
##   male     0.012765957  0.1829787
```

5. Research Question-I Analysis

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 :

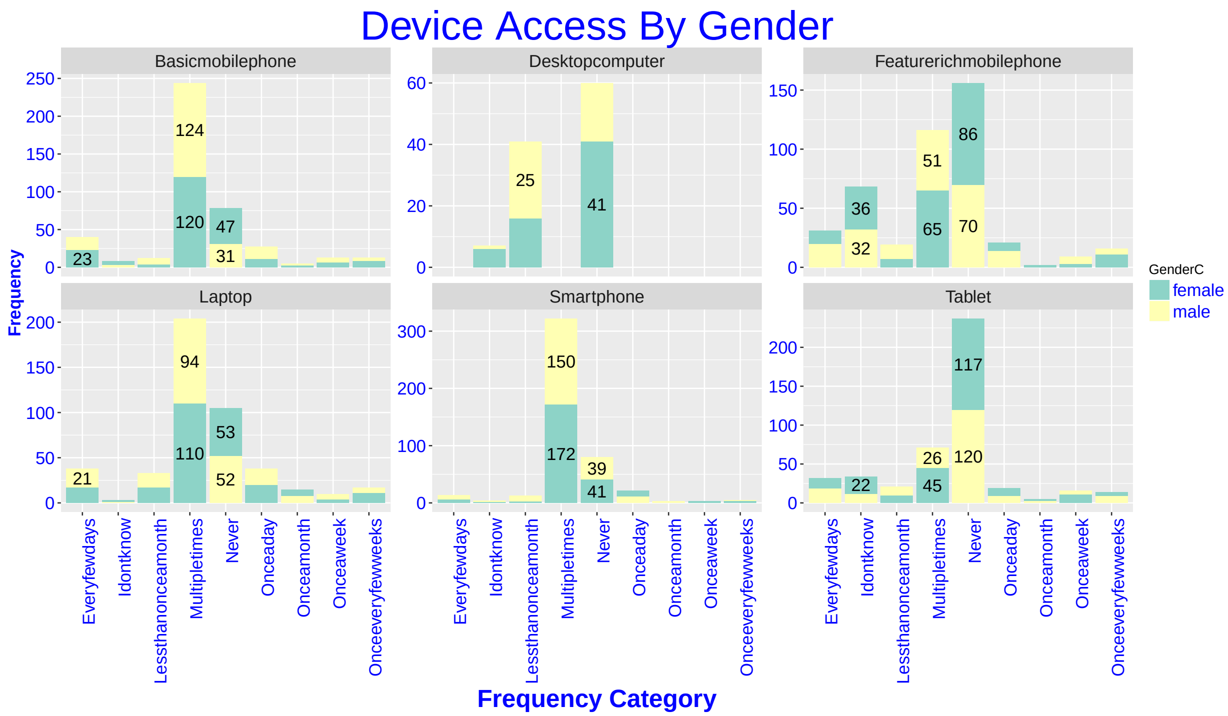
1. Access to electronic devices such as smartphones, tablets, laptops etc, their use frequencies, and location of use
2. Access to Internet, Use frequency of internet by the respondent on a device and location of use
3. Access to any digital services via internet access on the devices that they use.

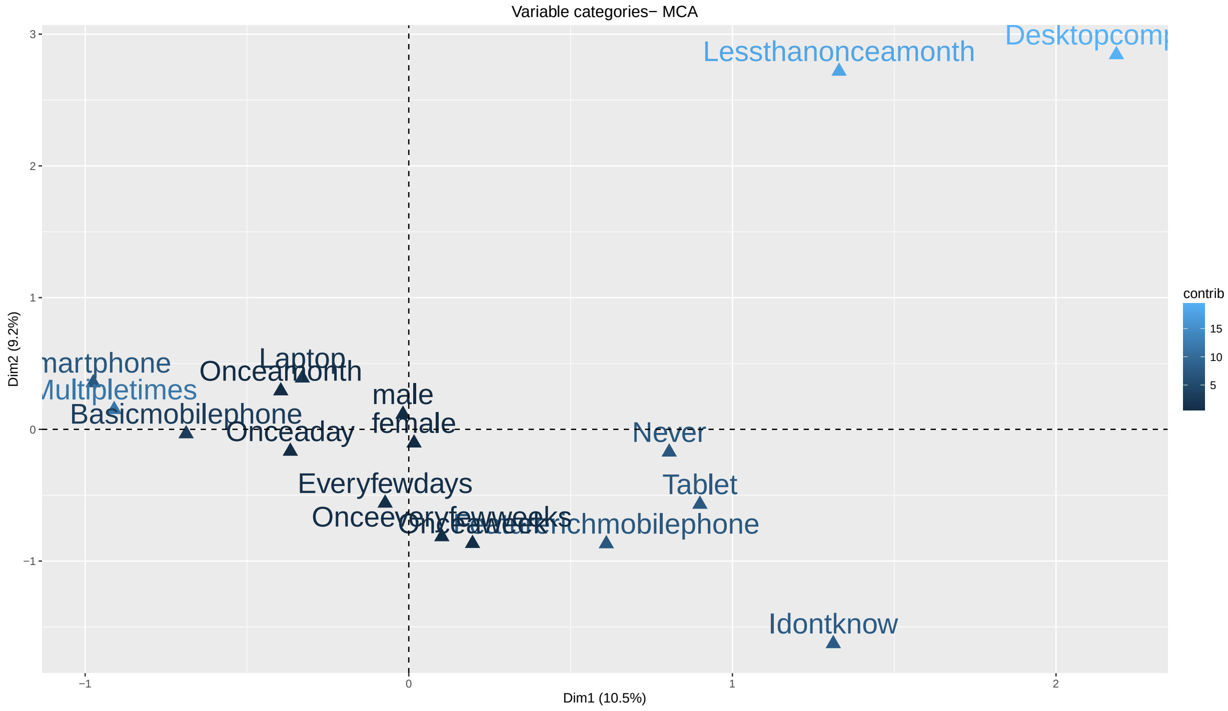
5.1.What devices do you have access to?

In this section we investigate for any assoication between the device that is accessed by the respondent and the respondernt's gender.

The bar plot below illustrates the devices accessed by the respondents of each category(Male/Female)

A correspondence analysis plot below displays Device association with the gender if any. An Independence test is conducted to check for the assoication.





	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

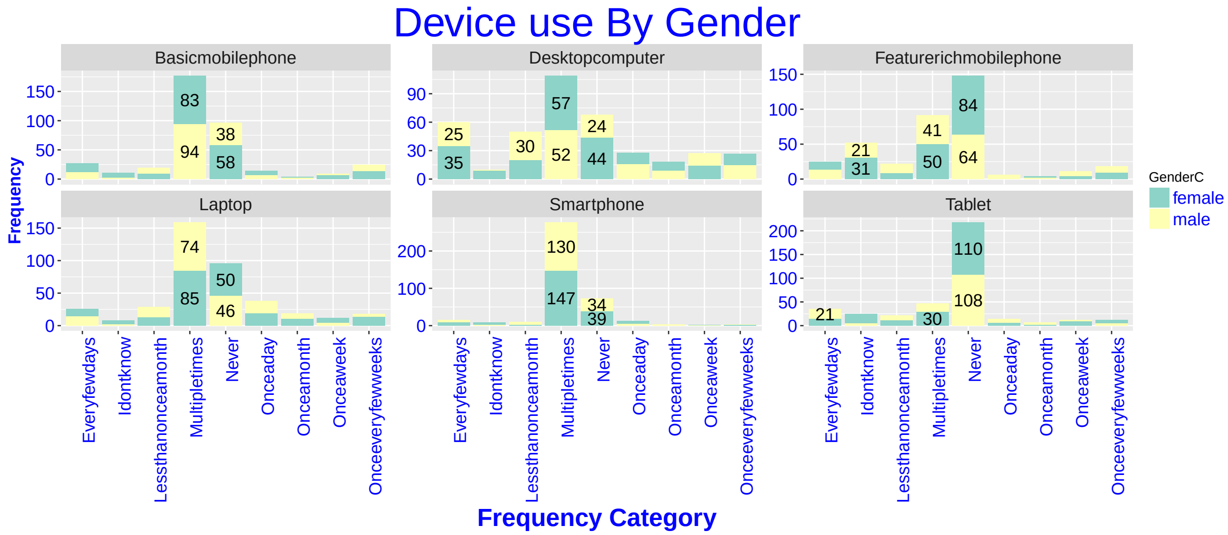
```
## Call: xtabs(formula = ~device + GenderC, data = dot.gen.acc.dev.dt)
## Number of cases in table: 2362
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 1.6292, df = 5, p-value = 0.8977
```

In the contingency table aboce Observe that the female to male proportions along the columns do not vary indicating that device alone may not have any assoication with device user's gender.However the row profiles show a slight variation across the gender category, leaving a scope for exploring for any statistically significant assoication with the access frequencies of devices.

The corresponding analysis plot above shows that both male and female respondents are having an access to a Smartphone multiple times, basic mobile phone every and laptop once a month. Not many respondents not seem to have a frequent access Desktop or a Laptop. The CA factor analysis plot is able to capture about only 19% of association, again leaving a scope for exploring for other contributors, that can be dummy coded into the table. We use this point in the next question where device name is also included into the crosstabulation test.

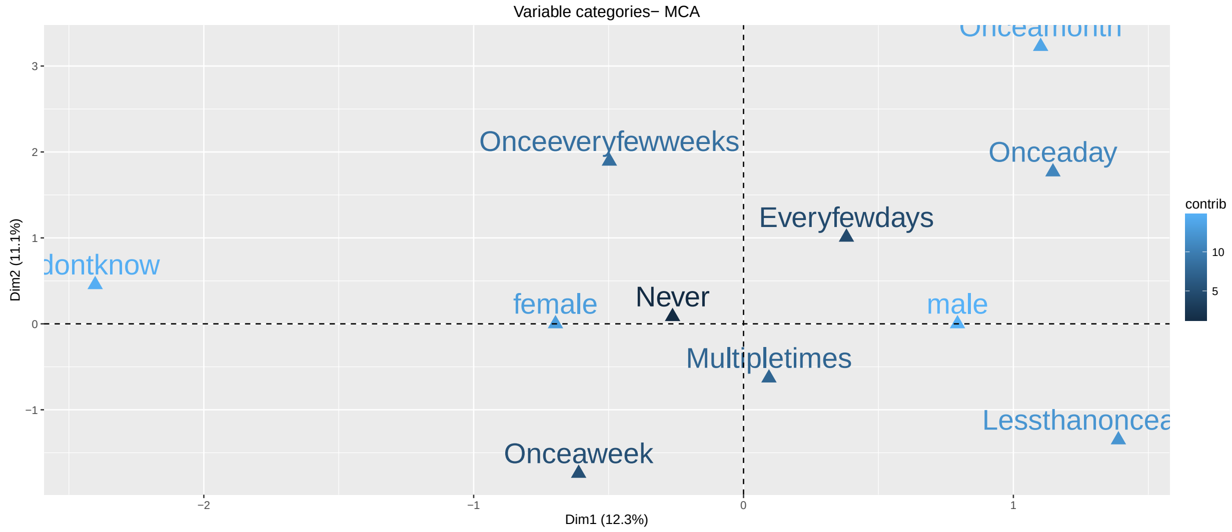
To interpret this plot, the closer two points of the same type are, the more similar those two row / column profiles are. And the closer two points of different types are, the more of their probability mass is in the cell representing their intersection.

5.2.How often do you actually use these devices?



	female	male
Everyfewdays	0.5053191	0.4946809
Idontknow	0.7017544	0.2982456
Lessthanonceamonth	0.4342105	0.5657895
Multipletimes	0.5255814	0.4744186
Never	0.5507868	0.4492132
Onceaday	0.4513274	0.5486726
Onceamonth	0.4545455	0.5454545
Onceaweek	0.5753425	0.4246575
Onceeveryfewweeks	0.5673077	0.4326923

```
## Call: xtabs(formula = ~use.dev.freq + GenderC, data = dot.gen.use.dev.dt)
## Number of cases in table: 2358
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 26.055, df = 8, p-value = 0.001028
```



The contingency table shows noticeable variations across row and column profiles indicating possible association between rows (device use frequencies) and columns (Gender category) We explore this further with a CA plot , association mosaic plot and conduct an Independence test on the data.

The corresponding analysis plot above shows that both the frequency category "multipletimes" co-occurs with male and female respondents almost equally (the difference in distance between these factors on the plot is very less)

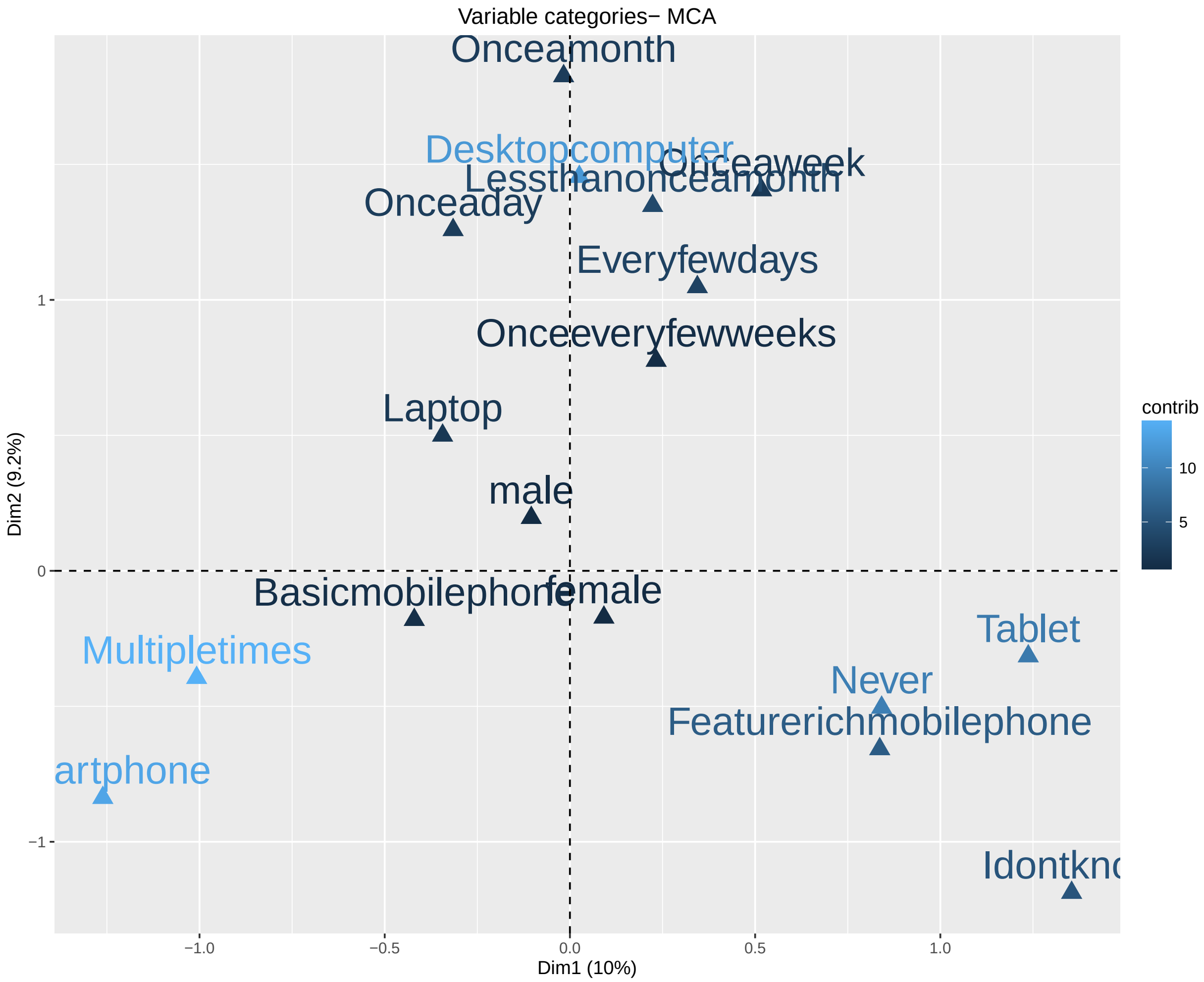
However, the frequency "Idontknow" seems to show a big difference among female and male respondents, in this case slightly far off than the othe frequencies associate with males. The same variation is observed in then contingency table above which shows a large variation among the row profiles.

Overall the inertia (data variation) explained by the CA plot about the co-occurence of frequency and the gender variation is only 23.4% (DIM1 + DIM 2 in the plot), which is not very significant. This only still leaves a scope to include more covariates or the assoication of device use frequencies may be weak.

5.2.1 Explore Association

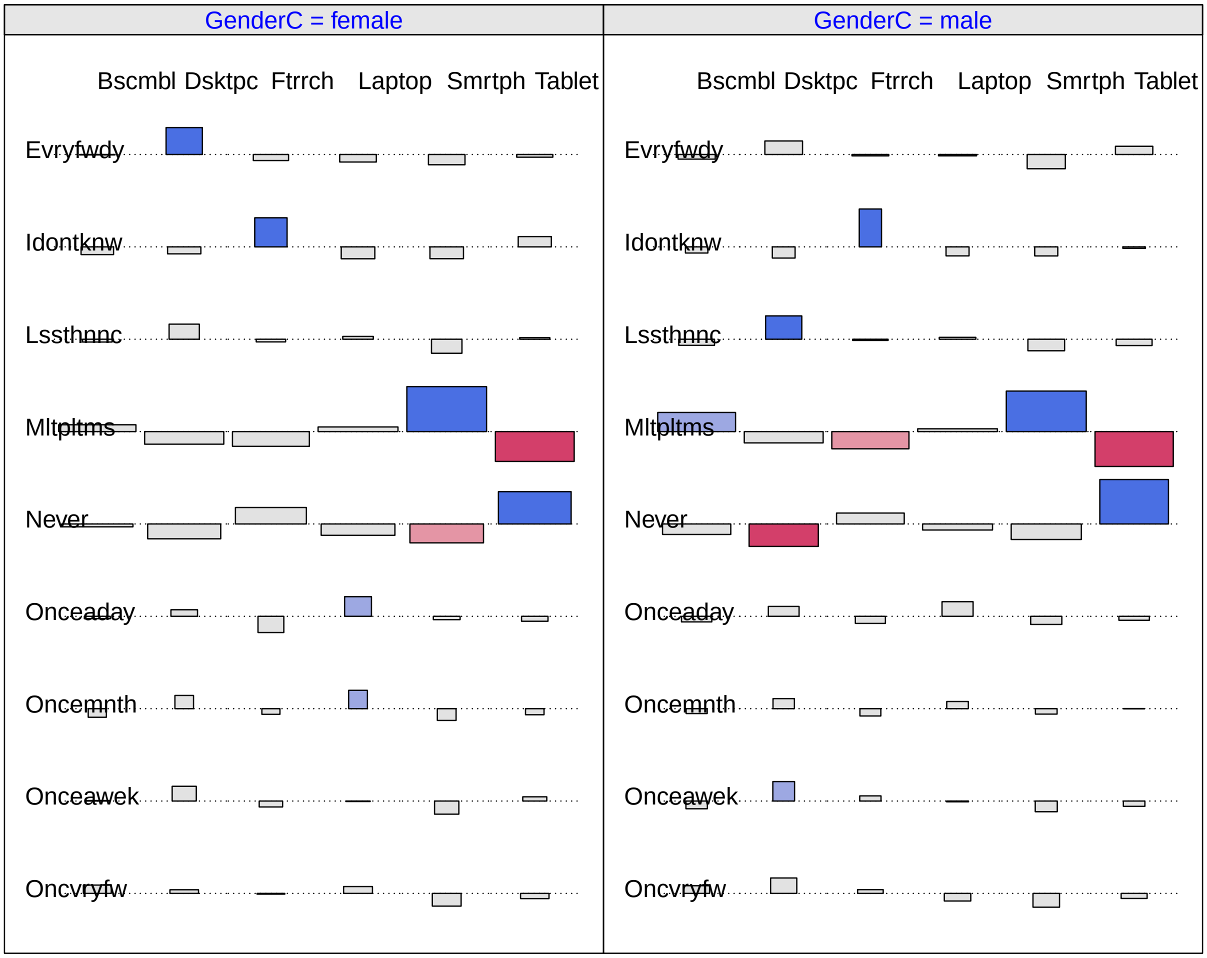
This section explores further possibilites of evidence of association among the two features : device use frequencies and Gender in the presence of a dummy coded variable : **device used**. It is natural to include a feature such as "device used" because, the frequencies are related to using a device. We first check the observed and expected counts for the data , to confirm that the chi square test warning is truly due to a large difference in the observed to expected counts.

If the expected counts are much lower than the observed counts, then a simulation test is performed on the data to assert an association. Mosaic plot is used to check the significant assoications and variation in the observed to expected frequencies.



```
## 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] 15  8  9 83 58  7

## [1] 21.83333 21.83333 21.83333 21.83333 21.83333 21.83333

## [1] "Check again with 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] "The simulation results also indicate a low p-value favoring a association among the row and colu
## [1] "Check the associations along with the devices with a mosaic plot"
```

```
## Call: xtabs(formula = ~use.dev.freq + device.use + GenderC, 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
```



In the mosaic plot above, the female-male "device use" proportions seem to be almost equal by the split. The split w.r.t device-use frequencies vs. devices also looks almost the same. Use frequency of Smartphone seems higher (more than expected) among females than males and similarly, use frequency of tablet among males multiple times is much lower(lower than expected) than the females, or correspondingly there are more males respondents who never use tablet for any purpose than females respondents. Males also tend to use Basicmobile multiple times unlike females who use laptop more than males.

In the mosaic plot above assoication of Females with a laptop once a day and once a month, assoication of males with basicmobile phone multiple times and desktop once a week seem to be statistically significant.

5.2.2. Statistical Significance

This section describes the tests conducted to measure the strength of association and inferences thereafter.

5.2.2.1. Diagnostic

$$(\chi^2)$$

Since the outcome of the response variable is nominal with multiple levels and the datset is considerably large, the diagnostics results will be low yet considered significant. The total inertia and contingency coefficient results are given below.

```
## [1] 0.07978723 0.07017544 0.05921053 0.09651163 0.08297568 0.06194690

## [1] "The Total inertia is"

## [1] 1.711186

## [1] "The trace or Correlation (contingency) Coefficient between the frequencies and gender category

## [1] 1.308123

## [1] 0.687 0.366 0.412 3.802 2.656 0.321
```

The contingeny coefficient is >1 and hence there is a strong association among a few levels of frequencies.

The ratio of observed to expected in some cases , in particular (**device = BasicMobilephone, freq=Multipletimes,gender=Male**) and (**device=Smartphone,freq=Multipletimes, Gender=female**) is very high indicating a strong association between these levels of use of the device.

6. Research Question - Q3 Analysis

Research Question 3 is to extract survey reponses for questions on how the youth in the DOT survey countries are utlizing the technology, primarily social media tools, to their benefit while engaging in formal and informal ways of income generation.

The Question to investigate as part of research analysis is :

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

The respondents were asked the following following subset of questions that evoked a open ended answer which is captured as a text within the response cells

6.1 Open ended questions-answers

An open-ended question in a survey or public opinion poll is an unstructured question in which possible answers are not suggested, and the respondent answers it in his or her own words.

Open ended questions have an advantage to evoke a meaningful qualitative response and latently capture the respondent's knowledge and thoughts that are relevant to the context.

6.2 Content Analysis

The qualitative comments from the respondents can run into multiple lines and often such text is “mined” to investigate “patterns” and then identify “sub-categories” of content. In this context the main category of content is a reponse to “using the technology for work” and the relevant sub-categories can be “devices used for work related activities”, “types of purposes” etc.

The common approach to analysing text responses is adopt a unsupervised machine learning algorithm to automatically “discover” the hidden “sub-categories” within. The hidden sub-categories are referred to as “topics” within the text mining and analysis techniques. Topic modeling is the most common framework adopted to analyse a large Corpus

For the purpose of analysing DOT open-ended response text, Structural Topic Modeling(STM) framework in R statistical software package is adopted because this framework not only processes open ended survey response text as document but also allows covariates to be included as part of topic modeling. The STM framework ouputs a model where each open-ended response is modeled as a mixture of topics. Unlike other topic modeling libraries that require pre-processing of the text, STM makes it simple by reading the response text “as-is” into the framework for further analysis. This makes the analysis less complex and visualizations easier. *Refer to section 9 of this document for details about STM framework.

DOT survey response data is a well formed set of samples that can be thought of (youth in the selected countries) stratified by location of survey and within each strata , the observations form a simple random sample of youth of varying age categories. Hence Age is considered as a covariate that can be included in the text analysis with selected sub-groups(levels). Since STM supports “topic contrast visualizations” between two groups, the topic modeling is performed with selected levels of Age. The intent of this analysis is to study any unique features among the *topics* that vary by Age of the youth. This specific analysis is concluded at the exploratory level with appropriate inferences. The intuition is that words within a topic and also their frequency vary by the covariate (Age), and hence providing valuable information to the inferences. Topic models distributions are structured by influence of covariate.

6.4 Visualization of Text Analyses

Typical Text processing in Open ended survey questions includes visualization of proximities between words and topics, characteristic words that make up for every topic. STM supports visualizations such as wordcloud, plots for words within each topic, Topic proportion over documents, Topic-meta data relations (influence of covariates), topic perspectives, which brings out the contrast in vocabulary used that vary by covariate on a per-topic basis

6.5. Topic Modeling

STM includes document meta data into the text analysis via topic prevelance and topical content coviartes. The text analysis for responses of the below mentioned questions are provided youth of specific age categories and it is interesting to investigate if the responses are in anyway dependent on their Age. It makes good sense to include *Age* as an overlapping covariate in topic prevelance and content parameters of the STM model, so as to investigate the influence on use specific set of words in a topic and the use frequency. In the regression model parameters, topic is used as a response variable and Age as both explanatory and covariate variable.

6.6.Statistical Analysis

The STM model estimates for each document, the proportion of words attributable to each topic,providing a measure of topic *prevalence* and topic *content*, the regression is performed for topic as an outcome as a variation in respondent attributes such as Age, Gender are the explanatory variables.The STM is capable of conducting this regression while also generating the topics. The following sections capture these details.

6.7 In what ways do you use devices for work?

The question in the subject evoked open-ended answers from the respondents and the answers are captured as free text by the DOT researcher

6.7.1 Data handling

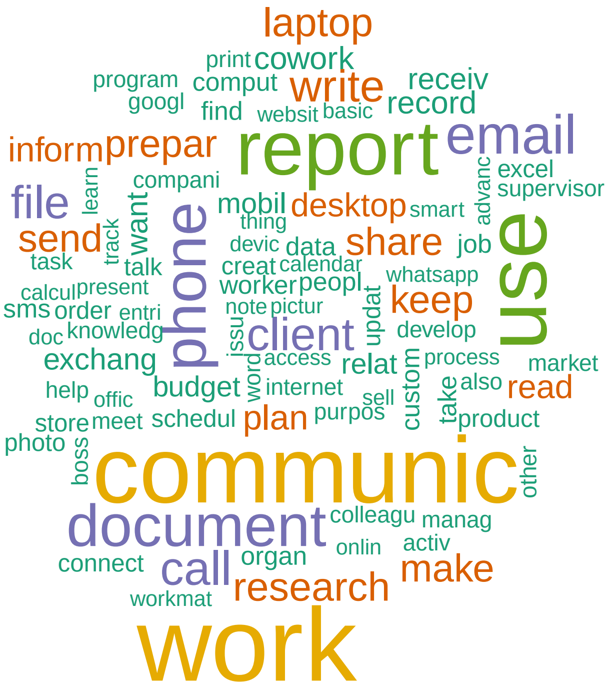
A subset of data with this response variable is extracted from the full data set including the control variables such as Gender, Location and Age for each respondent.

Open-ended response variable is reconstructed as a raw text vector, after filtering the blanks and NULL response cases. Thereafter the text vector is processed using STM library functions in R. The STM model is built using Age as a Topic prevelance and Topic content Covariate and selected topics are re-estimated to investigate any Topic-metadata relations.

A word cloud followed by Topic proportions within Documents is plotted to provide insights into the frequent words and high probability topic-document associations.

Topic-Covariate influence is presented using “Topic perspectives plot” and “Topical differences plot.”

6.7.2 Topics Estimates Visualization



Other Ways of youth using devices for work

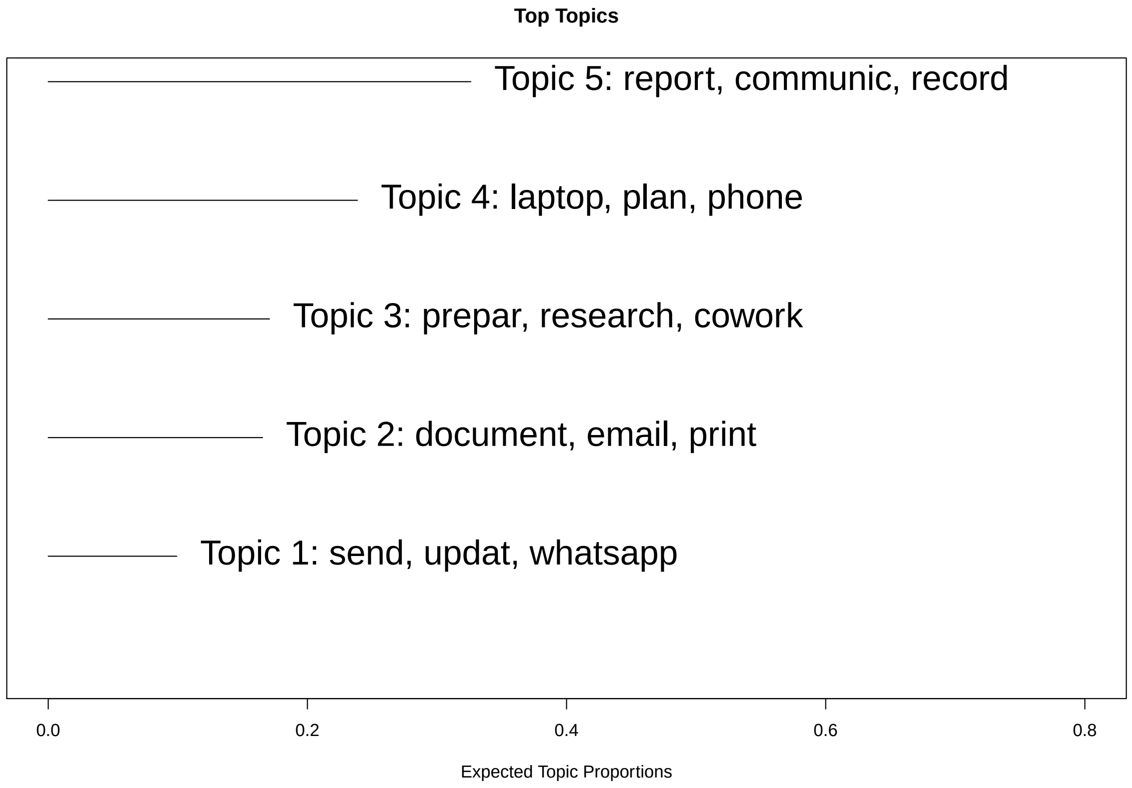
Topic 1:
send, file, task, use, updat, work, call, write, whatsapp, pictur, item, share, make, communic, email, desktop, phone, receiv, want, exchang

Topic 2:
document, email, client, read, use, work, call, write, print, share, make, communic, desktop, phone, file, receiv, want, exchang, keep, budget

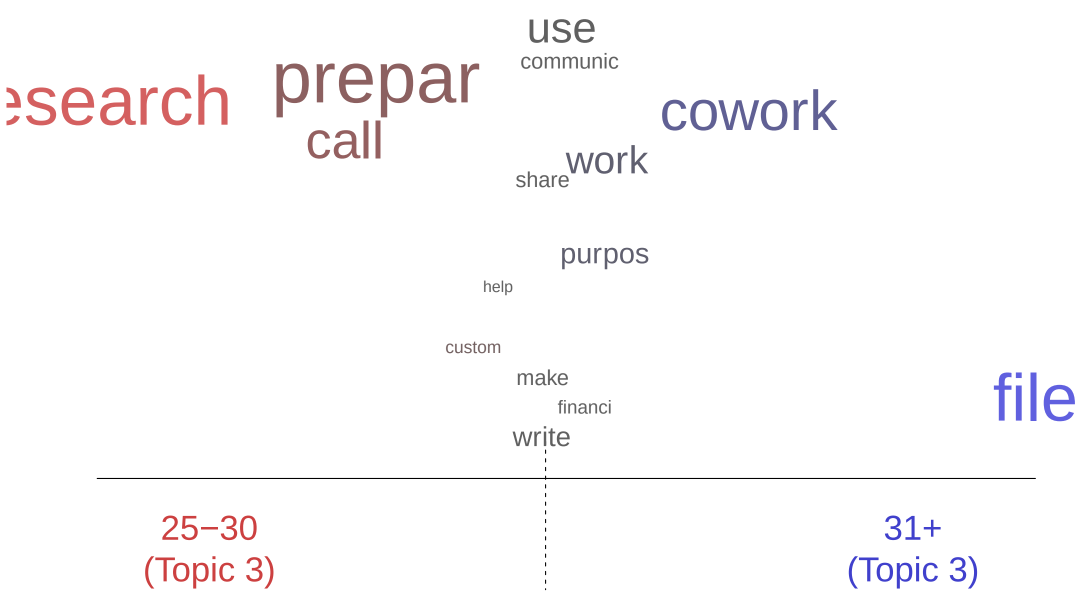
Topic 3:
research, prepar, call, cowork, file, use, work, write, purpos, custom, share, make, help, communic, email, desktop, lesson, financi, want, exchang

Topic 4:
work, phone, laptop, plan, use, organ, call, write, creat, share, make, desktop, receiv, want, exchang, document, communic, keep, email, budget

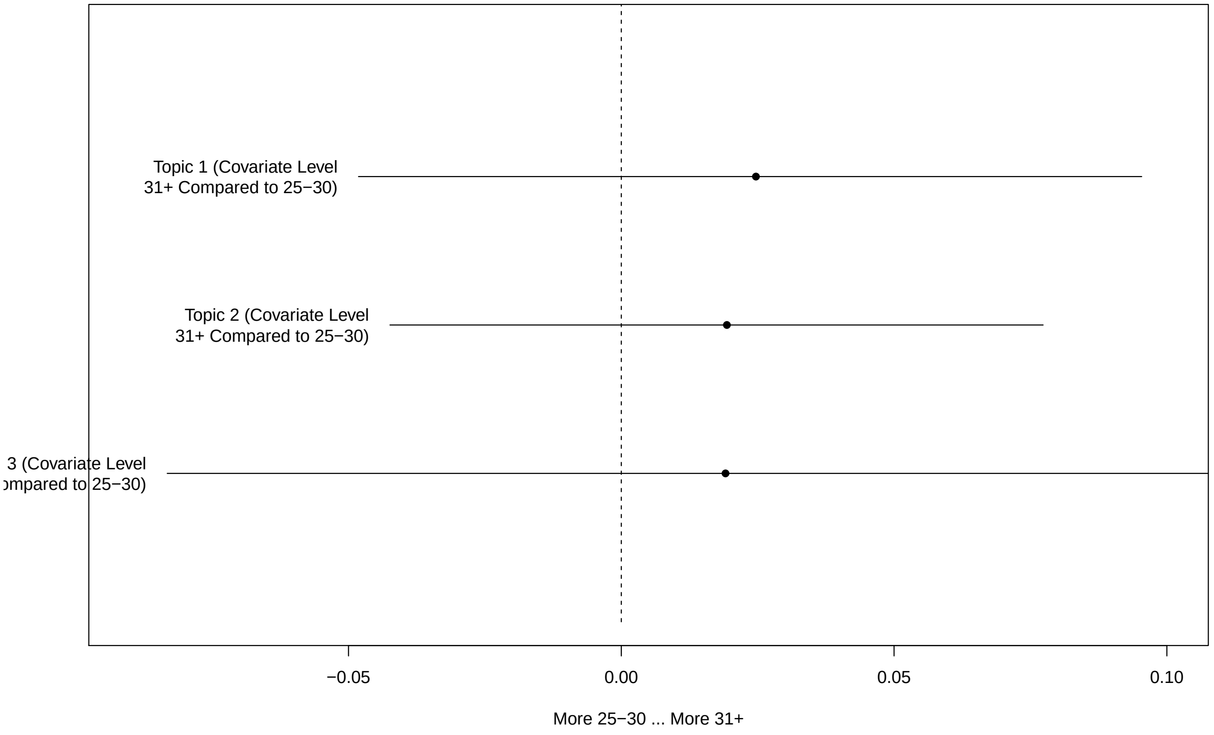
Topic 5:
communic, report, use, keep, inform, record, work, write, relat, job, share, make, call, desktop, receiv, want, exchang, read, budget, mobil

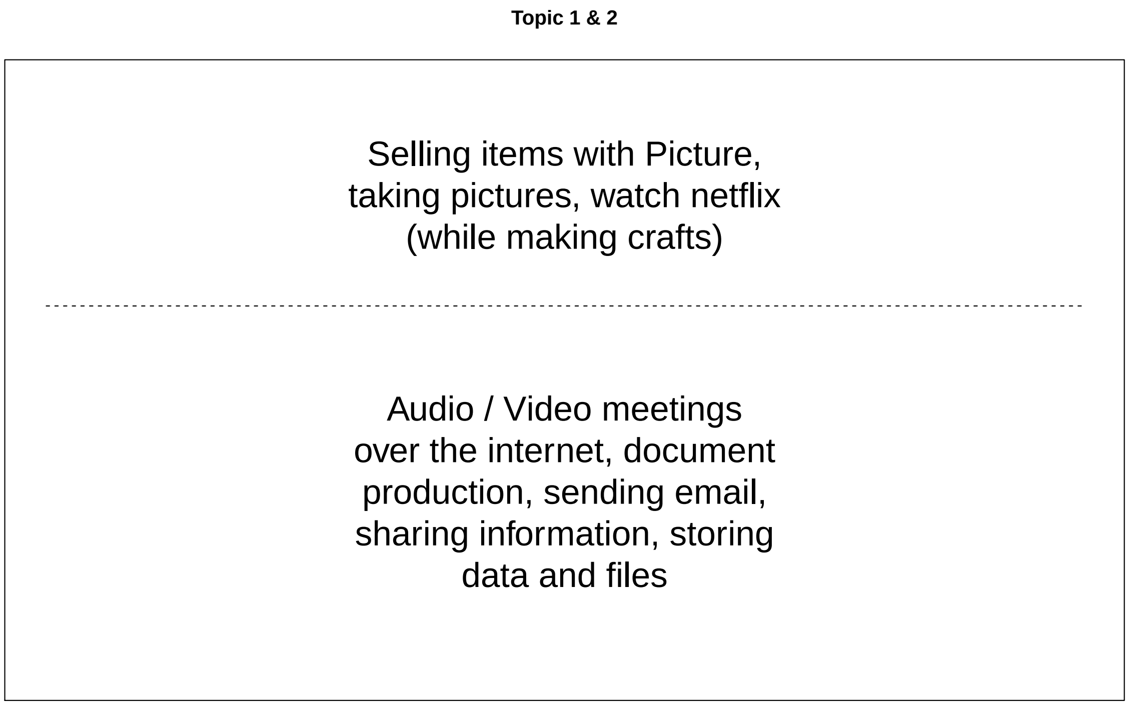


Other Ways of youth using devices for work



Effect of Covariate(Age) on Topic content





Topic 1 & 2

Selling items with Picture, taking pictures, watch netflix (while making crafts)

Audio / Video meetings over the internet, document production, sending email, sharing information, storing data and files

The word cloud plot above indicates that the most frequent words in the responses are “work”, “communicate”, “report”, “use”, “enail” etc.

The plot below the wordcloud is for estimated topics (merely as collection of words). With subject understanding and on observation, **Topic1** can be estimated as mostly about “**exchange of content such as pictures, files etc via applications on the device**”, in **Topic2** , the content is about the “**information sharing via emails for job related activities**”, **Topic3** is discussed for “writing related acitivities for work such as research and report”, Topic4 is more focussed on “**how the devices are used for work planning, keeping a track**” “**Topic5**” is discussed for “**personal and professional interests carried out with the devices**.” These inferences about topic *subjects* is purely intuitive. Since the research is an observational study, these inferences can be verified against the documents to topics association as below

The plot below the *Top Topics listing* is a comparison of **effect of Age** on the content (words/vocabulary) used in the topics. It can be observed that Age group 25-30, use vocabulary such as “research document”, “report”,“prepare”, while respondents of Age 31 and above use vocabulary for work related writing as “file”

The Topic metadata(here the AgeC variable is the metadata) relationship is plotted next, to check the variation in topic proportions under the influence of Age. Topic2 seems to be more aligned to a document from Age 31 compared to Topics3 and Topic4. We cross check that against the acutal document content with the next plot that shows the content for topics 2 and 4.When these documents(text vectors for respondent’s answer) are cross verified against the actual responses in the data set, it can be concluded that the model has performed fairly well.

In the above open-ended survey response text analysis, the choice of number of topics is purely arbitrary.

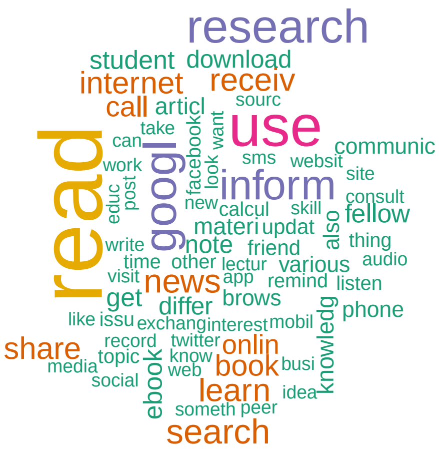
7. Research Question - Q5 Analysis

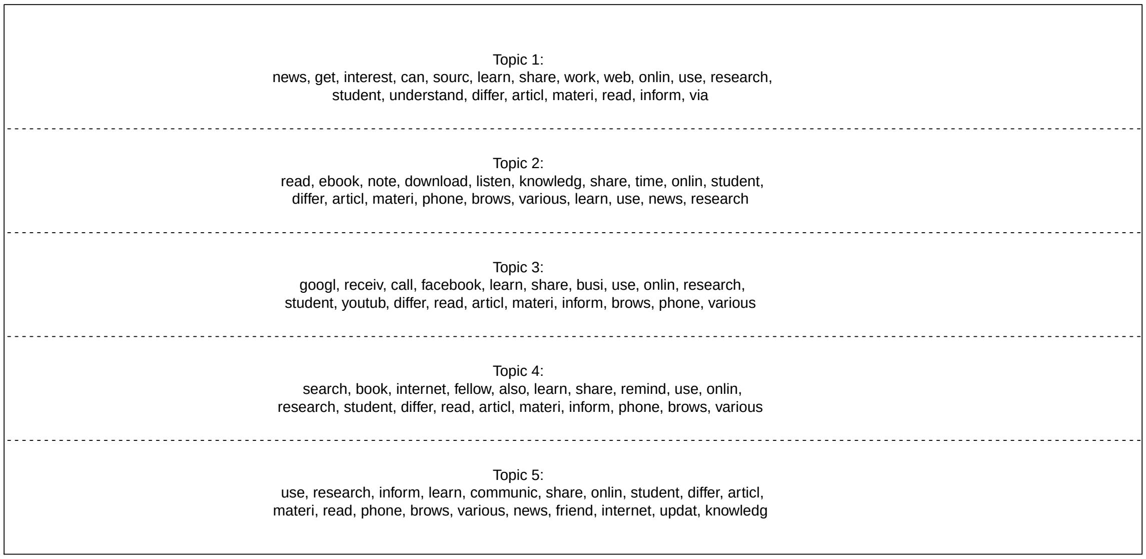
Analysis to Q5 follows the similar approach as in *Question 3* above as Question 5 is also evoking open-end answers from the youth about **In what ways are youth in these countries using technology and social media for learning**

The question in the subject is an extension to DOT's study to understand how youth in these countries are benefiting from technology and social media when used for their learning.

The question is again a collection of following questions posed to the respondents on devices,Internet and digital services used for learning or self-development.The answers from the respondents are again captured as text variable by the DOT researcher.In the sections below, text analysis is conducted for answers for **In what ways do you use devices for learning**

7.1 Topic Estimates and Modeling





Ways of youth using devices for Learning

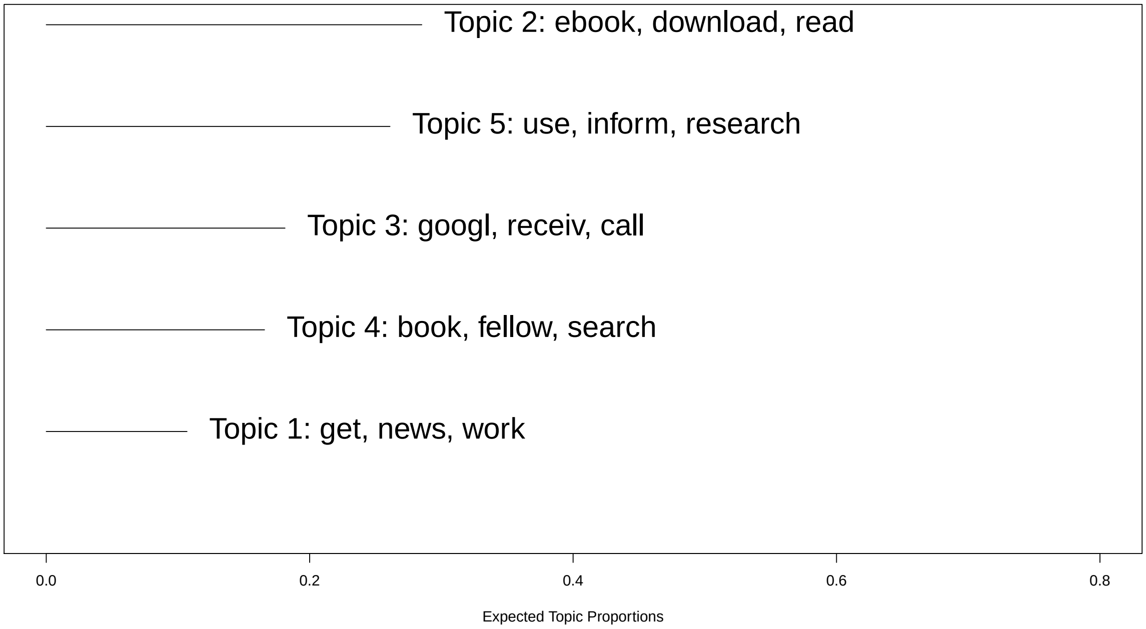
Topic 1:
news, get, interest, can, sourc, learn, share, work, web, onlin, use, research, student, understand, differ, articl, materi, read, inform, via

Topic 2:
read, ebook, note, download, listen, knowledg, share, time, onlin, student, differ, articl, materi, phone, brows, various, learn, use, news, research

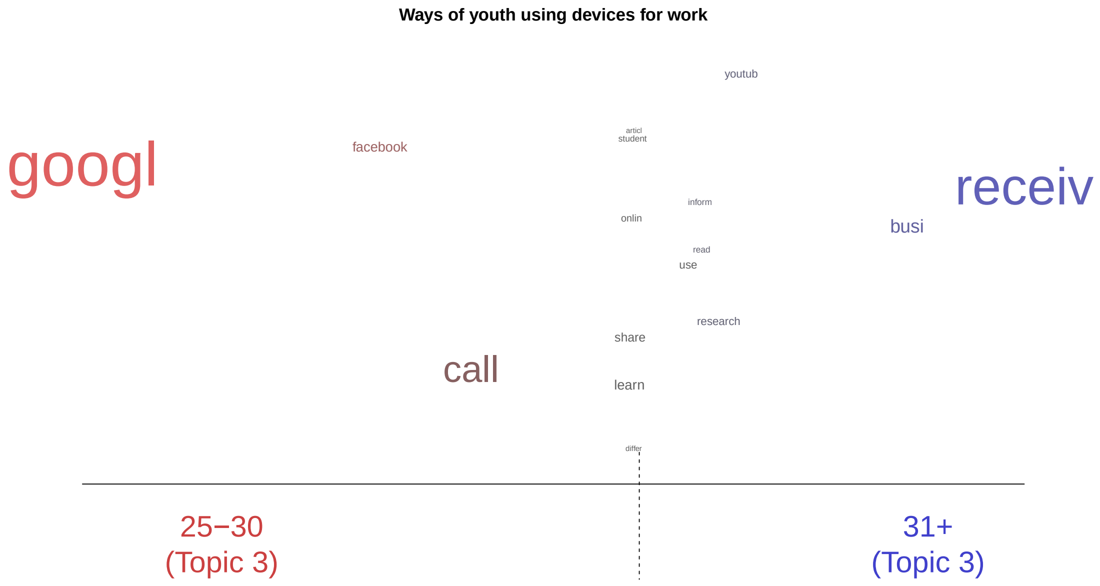
Topic 3:
googl, receiv, call, facebook, learn, share, busi, use, onlin, research, student, youtub, differ, read, articl, materi, inform, brows, phone, various

Topic 4:
search, book, internet, fellow, also, learn, share, remind, use, onlin, research, student, differ, read, articl, materi, inform, phone, brows, various

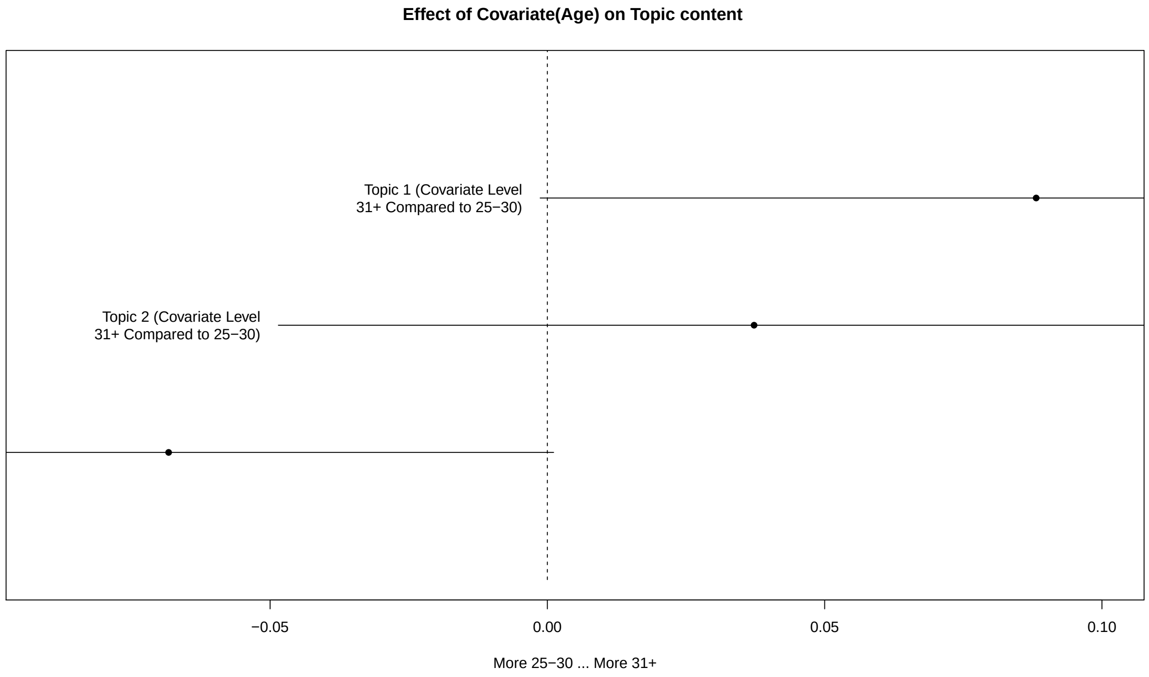
Topic 5:
use, research, inform, learn, communic, share, onlin, student, differ, articl, materi, read, phone, brows, various, news, friend, internet, updat, knowledg



Topic Estimates for devices used for learning



Ways of youth using devices for work



Effect of Covariate(Age) on Topic content

The WordCloud plot above indicates that the most highly used features of mobile device for learning are *read*, *search google*, *use for research*, *access internet for material* ,*getting news updates* etc.

The words distribution among Topics predicted by the text analysis model above almost coincide with the information in Wordcloud for response vocabulary

Topic1 can be estimated to be discussing *information or useful material download for learning*, **Topic2** can be about **reading material online**, **Topic3** can be approximated to be about **tools or applications for learning** used on the device that aide the learning, **Topic4** is primarily about **searching or browsing the internet** for reading material **Topic5** can be expected to be about **information sharing , gain**

knowledge

Further, the **Topic proportions** plot below indicates that the learning with a Mobile device happens primarily via, e-book reading, downloading useful material , followed by accessing useful material for research or other activities, searching and sharing.

Again an observation of the topical contrasts plot below among Age group 25-30 and 31+ indicates that Age group 25-30 primarily use google search and facebook as their learning modes while Age group 31+ look at learning with the device as send and receive updates from peers, while both age groups look at learning as "reasearching for useful material, reading online etc."

A fair conclusion from the above visualization is that Mobile device is used to access internet and thereon access or download reading material such as books for learning purpose. There seems to be a large inclination for learning by search for relevant research material and exchange or share information among freinds for learning.

Mapping it back to the actual data, it can be understood that the estimated topic model has performed reasonably well. Similar analysis can be extended to remaining collection of responses in the learning catergory of DOT data.

8. References

http://www.stat.cmu.edu/~hseltman/309/Book/chapter16.pdf

http://www.sthda.com/english/wiki/correspondence-analysis-basics-r-software-and-data-mining

https://cran.r-project.org/web/packages/vcdExtra/vignettes/vcd-tutorial.pdf Introduction to SAS. UCLA: Statistical Consulting Group. from http://www.ats.ucla.edu/stat/sas/notes2/ (accessed November 24, 2007).

http://www-users.cs.umn.edu/~ludford/stat_overview.htm

http://www.agrocampus-ouest.fr/math/sensometrics2012/Slides/Lebart.pdf

http://scholar.harvard.edu/dtingley/files/topicmodelsopenendedexperiments.pdf

http://www.electionstudies.org/conferences/2008Methods/Popping.pdf

http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/

https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/ https://scholar.princeton.edu/sites/default/files/bstewart/files/stmnips2013.pdf

9. Acknowledgements

I'm thankful to Coursolve and DOT organizations for having given me an opportunity to work on this research internship, which has an immense potential to explore latest tools and techniques in the field of Survey Data Analysis,

My sincere thanks to DOT Proect management and Admin. team for having provided with all the necessary support needed on the project. Special thanks to AnnePatterson and Sharon Nyangweso for being so responsive over the questions posted

This research analysis is a good amount of hardwork combined with challenge and vast learning. This work would not have reached this stage without the motivation from co-collaborators in the team, my special thanks to Peperomia for sharing his valuable insights and talent across comments on the project threads.

I express my sincere thanks to Coursolve devlopment team and experts for their thoughtful effort in bringing up such challenging digital internships to the online learning platform

10. Extensions

This project , although has been analysed to extent of prescribed methods for Survey Data analysis and text analysis, there is a vast potential for further exploration.

Some of the explorations which are assumed to be appropriate for a further data analysis are :

1. Multi-nomial regression Analysis for multi-level response outcome
2. Application of visualizations and statistical tools for discrete data analysis
3. Approach to sruvey data with the help of specific packages for survey data analysis as **survey** in R software suite of libraries

4.Rigorous analysis of Data by other Covariates such as Age, Location , Employment and Education.