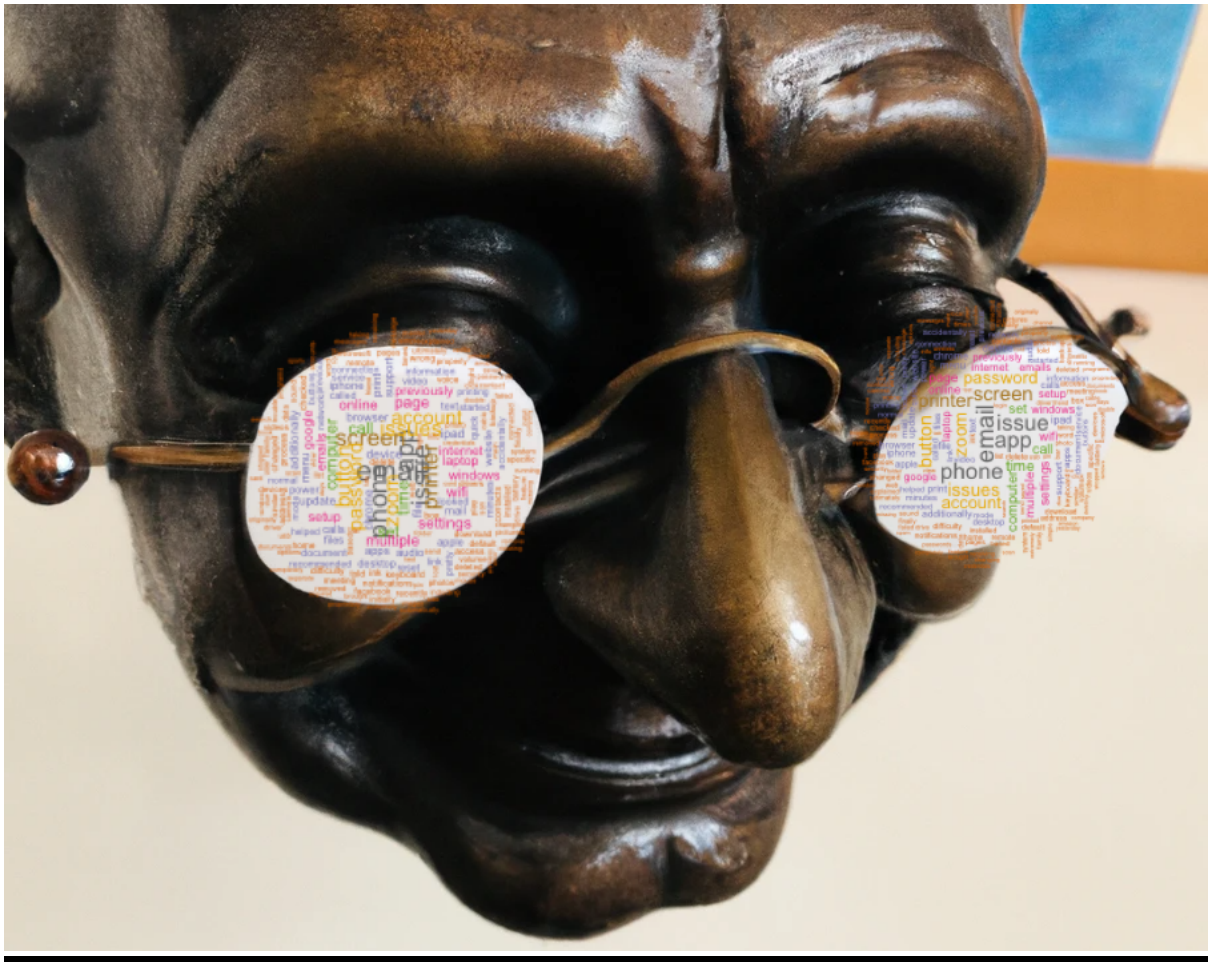
DATA MANIPULATION AND FEATURE ENGINEERING



INTRODUCTION

Technology can be daunting for anyone, particularly for seniors who face multiple challenges dealing with technology. And Covid has only exacerbated this by adding a layer of anxiety that the senior residents can do without. That said, are the problems and challenges of the seniors significantly different from others when it comes to technology? By analysing the answers filled in by the Tech Support team to a questionnaire, I have tried to understand the problems faced by the seniors. It is worthwhile to ask what are the seniors trying to achieve with technology? What is their biggest challenge? Which technology or device is the most problematic? By understanding the way, the seniors interact with technology will help us find solutions to some of the common problems.

In this project, I have focussed on that part of analysis where we prepare the data for further analysis or model building. I have not sought to solve the problems of the seniors with regards to technology, but have taken the necessary step of understanding them.

EXECUTIVE SUMMARY

The TekHub project is part of an active research collaboration between RIT and local senior living community, Jewish Senior Life (JSL). The dataset documenting over 1500 interactions between seniors living at JSL and RIT-provided tech support was analysed to find answers to the following questions:

a. Which technologies are seniors struggling with the most? Why?
b. What tasks are seniors trying to do?
c. How do the above interact?
d. How are these factors changing over time?

To answer the questions above, the questionnaire resulting in this dataset was studied carefully. I have approached the problem by studying each variable by generating as many questions as possible about the variable, it's interactions with other variables, with a view to generate meaningful insights. The study was divided into the following sections:

1. Exploratory Data Analysis
2. Feature Engineering
3. Natural Language Processing
4. Insights/Conclusion

In this study, I have selected each variable separately and have analysed them individually, and how they interact with other variables in the dataset. I have then cleaned them and transformed them in such a way, that they are fit for purpose for answering the questions mentioned above.

Exploratory Data Analysis (EDA)

"Exploratory data analysis (EDA) is the process of analyzing data in order to gain insight into patterns and trends, and to detect outliers and anomalies. In its simplest form, EDA is a way of exploring data to understand its characteristics and structure, and to identify relationships between different variables. The main aim of EDA is to make sense of the data so that it can be used to answer questions and inform decisions. EDA is an iterative process and typically involves visualizing the data, transforming and cleaning it, and then making inferences about it."

- Edward Tufte

As part of the EDA, I have looked into detecting patterns, trends, commonalities and anomalies based on the data. The first step is to look at the structure of the dataset.

The Dataset

```
  Timestamp                      What technology is the client visit for? What are the main goals for technology use?
Min.   :2020-04-06 15:37:48.14   Length:1281                              Length:1281
1st Qu.:2020-10-05 11:02:43.00   Class :character                         Class :character
Median :2021-03-01 10:19:02.00   Mode  :character                         Mode  :character
Mean   :2021-03-03 21:55:56.01
3rd Qu.:2021-07-21 17:09:24.31
Max.   :2022-02-14 16:30:46.65

What are the primary challenges? Were you able  to resolve the problem?
Length:1281                      Length:1281
Class :character                 Class :character
Mode  :character                 Mode  :character



Notes on this visit (include suggestions for UI improvements) The client was satisfied with the visit
Length:1281                                                    Min.   :1.000
Class :character                                               1st Qu.:4.000
Mode  :character                                               Median :5.000
                                                               Mean   :4.478
                                                               3rd Qu.:5.000
                                                               Max.   :5.000
                                                               NA's   :2
```

One of the basic steps in EDA, is to find out some general information about the dataset 'tekhubdata'. The dataset contains 1281 observations or rows, and 7 variables or columns. Each column is actually a question in the questionnaire. For ease of use we can change the names of the columns, so that the columns look like:

```
  timestamp                      whichtech          goal           challenge          resolved           note           satisfaction
Min.   :2020-04-06 15:37:48.14   Length:1281        Length:1281    Length:1281        Length:1281        Length:1281    Min.   :1.000
1st Qu.:2020-10-05 11:02:43.00   Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.:4.000
Median :2021-03-01 10:19:02.00   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :5.000
Mean   :2021-03-03 21:55:56.01                                                                                                Mean   :4.478
3rd Qu.:2021-07-21 17:09:24.31                                                                                                3rd Qu.:5.000
Max.   :2022-02-14 16:30:46.65                                                                                                Max.   :5.000
                                                                                                                              NA's   :2
```

Missing values:

The next step would be to find and handle the missing values. In this dataset, the majority of the missing values are in the 'challenge' variable:

```
#checking for missing values
sapply(tekhubdata, function(x) sum(is.na(x)))
 timestamp    whichtech         goal    challenge     resolved         note satisfaction
         0            2            3           87            2            0            2
```

There are several ways to handle missing values:

1. Deleting the missing values: This can be done by using the na.omit() or complete. cases() functions.
2. Imputing the missing values: This can be done by using the mean, median, or mode of the non-missing values.
3. Replacing missing values with a placeholder value: This can be done by using the is.na() function.

4. Predictive modelling: This involves using machine learning algorithms to predict the missing values.

I have chosen the first option of na.omit(), and checked if the missing values are indeed deleted.

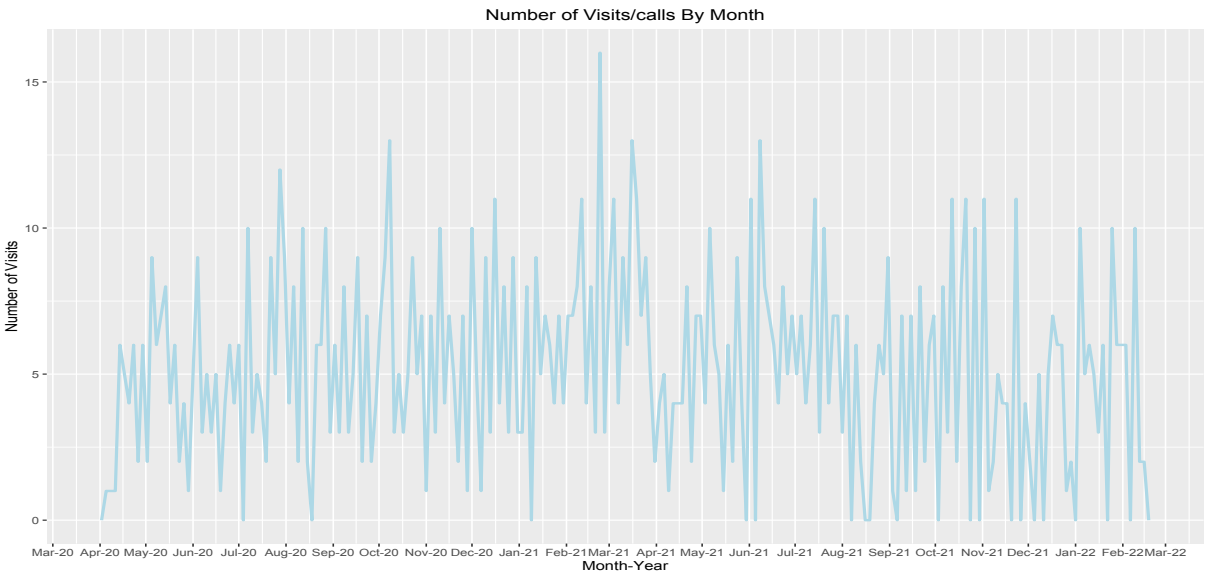
```
> sapply(tekhubdata, function(x) sum(is.na(x)))
   timestamp    whichtech         goal    challenge     resolved         note satisfaction
           0            0            0            0            0            0            0
>
```

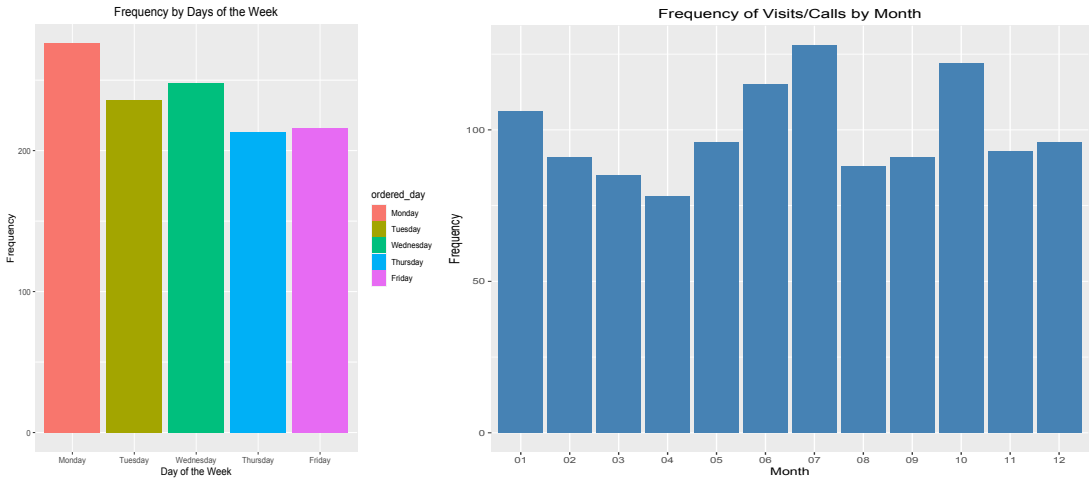
FEATURE ENGINEERING

Feature engineering is the process of transforming raw data into features that can be used to build more effective machine learning models. This process involves selecting relevant data, transforming it into a suitable form, and creating new features that can improve the accuracy and performance of the model. Feature engineering can involve a variety of techniques, including feature selection, feature extraction, and feature construction.

TIMESTAMP

The timestamp variable is the time given in terms of year-month-day hours, minutes, seconds of the visit or call made by the Tech Support person. The survey started during April 2020 and ended during March 2022.
The timestamp variable is of the class POSIXct and of the format “2020-04-06 15:37:48". This needs to be converted to Date with the ‘as.Date()’ function, of the month format (%m). Then monthly visits are aggregated and plotted. The graph below shows the frequency of the calls or visits made by the Tech Support person over a period of two years. The purpose of the graph is to see any unusual spikes or drops in calls. It appears the demand has been fairly steady even throughout.





With the above plots I have tried to see which day of the week, month of the year is the busiest for Tech Support. July, October, and June are busy months for the Tech Support. Similarly Monday is the busiest day.

WHICHTECH

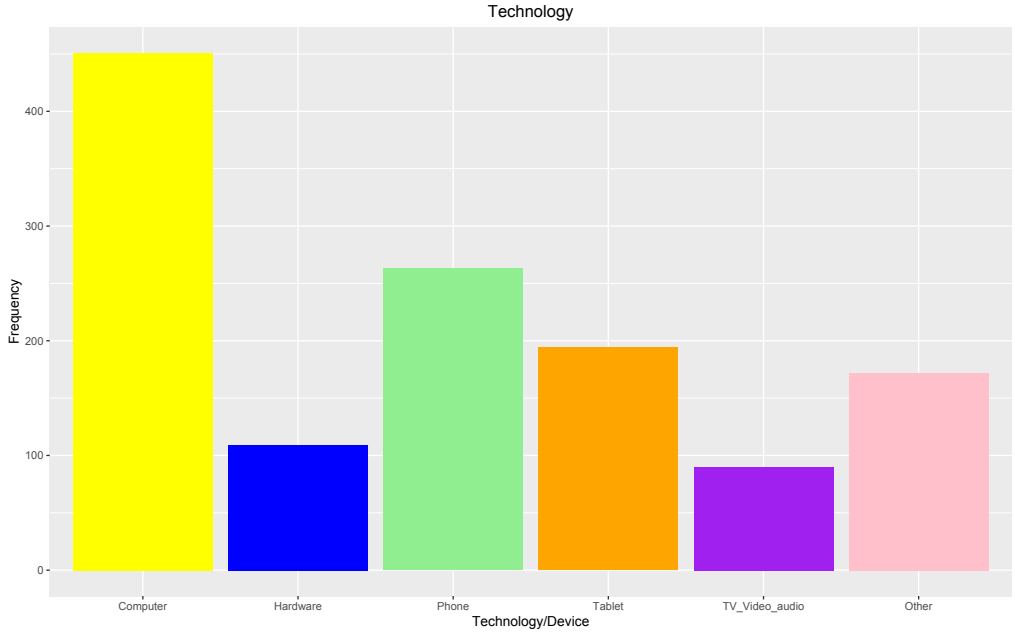
The variable 'whichtech' is the result of answers given to the question "What technology is the customer visit for?" in the questionnaire. Judging by the answers, here the term 'technology' is interchangeably used with a device or an appliance. There are 118 unique items in this list which I have manually categorized them into 10 categories, as follows:
1.Account Management 2. Computer 3. Dotcom 4. Hardware 5. Phone 6. Software 7. Tablet 8. TV/Video/Audio 9. Websites 10. Other

```
[1] "Smartphone: iOS"                          "Laptop"
[3] "Linux Telikin"                            "Tablet:iOS"
[5] "Desktop"                                  "Smart TV"
[7] "Smartphone: Android"                      "Desktop Windows"
[9] "Tablet: Android"                          "Laptop, TV, Smartphone: Android, Printer"
11] NA                                         "Modem + Router"
13] "iPad and iPhone"                          "Wireless Printer"
15] "Roku"                                     "Cloud"
17] "Desktop/Speakers"                         "Google Sheets"
19] "Desktop Mac"                              "Flashdrive"
21] "Printer"                                  "Digital Picture Frame"
23] "Online Shopping"                          "Phones"
25] "Movie"                                    "Account"
27] "Accessories"                              "PowerPoint"
29] "Router"                                   "Nook"
31] "Password Manager"                         "Email"
33] "Bluetooth"                                "Literally everything"
35] "Apple Watch"                              "Amazon"
37] "Captioned Phone"                          "Password"
39] "Internet"                                 "Remote"
41] "Amazon Self Publishing"                   "Flip Phone"
43] "Website"                                  "iCloud"
45] "Webcam"                                   "Zoom"
47] "Antivirus"                                "Phone"
49] "Video"                                    "Landline"
51] "Specific Issue"                           "Flip phone"
53] "Hardware"                                 "Word"
55] "Presentation"                             "Password Reset"
57] "Wireless Keyboard"                        "Keyboard"
59] "Infrastructure"                           "Installation"
61] "Apple TV"                                 "Medical Devices"
63] "Recording"                                "Play"
65] "Voice command device (e.g. Echo)"         "Research"
67] "CD"                                       "DVD"
69] "Security"                                 "TV"
71] "YouTube"                                  "Monitor"
```

Step #1 ‘Unique Tech’: The first step is to find the unique items in the list of technology/device that the residents were having problems with.

Step #2 Combining categories: In order to study the reason why the customer is having a visit or a call, 118 unique items have to be to be bunched together in manageable categories. Here, using the gsub function, I have assigned the items to the following categories: 1.Account Management 2. Computer 3. Dotcom 4. Hardware 5. Phone 6. Software 7. Tablet 8. TV/Video/Audio 9. Websites 10. Other

Step #3 Convert to factors: These categories are then converted to factors and the top 5 are plotted.



Highest number of residents had Computer -related problems. Phones and Tablets have not easy with the seniors, with more than 250 cases of phone related problems, and almost 200 Tablet issues.

CHALLENGE

By far the most challenging variable in terms of prepping the data has been the variable ‘challenge’. The ‘challenge’ variable is the result of the answer to the question “What are the primary challenges?”. But the real challenge was the way to handle the way the answers were recorded with more than one of the following challenges mentioned by one customer:

- Awareness- knowledge of available solutions to address the task
- Confidence- belief in ability to handle the task
- Cognitive- ability to understand the requirements for use
- Memory- ability to remember key requirements (e.g. password)
- Physical- ability to manipulate the device (e.g. advanced swiping required)
- Sensory- ability to detect the state of the device (e.g. text too small)

- Other:

The ‘challenge’ column before modification looks like this:

challenge	resolved
Awareness- knowledge of available solutions to address the task, Confidence- belief in ability to handle the task, Memory- ability to remem...	Yes
areness- knowledge of available solutions to address the task	No
nfidence- belief in ability to handle the task	No
onfidence- belief in ability to handle the task, Physical- ability to manipulate the device (e.g. advanced swiping required)	Yes
Awareness- knowledge of available solutions to address the task, Confidence- belief in ability to handle the task	Yes
Confidence- belief in ability to handle the task	Yes
Awareness- knowledge of available solutions to address the task, Confidence- belief in ability to handle the task	Yes
Awareness- knowledge of available solutions to address the task	Yes
Awareness- knowledge of available solutions to address the task, Cognitive- ability to understand the requirements for use	Yes
Awareness- knowledge of available solutions to address the task, Confidence- belief in ability to handle the task	Yes
Awareness- knowledge of available solutions to address the task, Confidence- belief in ability to handle the task	Yes
Memory- ability to remember key requirements (e.g. password), Sensory- ability to detect the state of the device (e.g. text too small)	Yes
Awareness- knowledge of available solutions to address the task, Confidence- belief in ability to handle the task	Yes
Awareness- knowledge of available solutions to address the task	Yes

Step #1Shorten Strings: The first task is to replace the rather lengthy strings like, for e.g. “Awareness- knowledge of available solutions to address the task”, with “Awareness”. We use the regex function gsub() to accomplish this. The ‘challenge’ column is transformed into this:

challenge	resolved
Awareness, Confidence, Memory- ability to remember key requirements (e.g. password)	Yes
Awareness	column 4: character
Confidence	No
Confidence, Physical- ability to manipulate the device (e.g. advanced swiping required)	Yes
Awareness, Confidence	Yes
Confidence	Yes
Awareness, Confidence	Yes
Awareness	Yes
Awareness, Cognitive	Yes
Awareness, Confidence	Yes
Awareness, Confidence	Yes
Memory- ability to remember key requirements (e.g. password), Sensory- ability to detect the state of the device (e.g. text too small)	Yes
Awareness, Confidence	Yes
Awareness	Yes

We notice that certain values like ‘Awareness’, ‘Confidence’ etc have changed, whereas other values like the ‘Memory- ability to remember key requirements (e.g. password)’, ‘Physical- ability to manipulate the device (e.g. advanced swiping required)’, etc haven’t changed. The reason for this is the presence of punctuations in the strings. When the punctuations are removed with the gsub function, the column looks like this:

challenge
Awareness, Confidence, Memory
Awareness
Confidence
Confidence, Physical
Awareness, Confidence
Confidence

Awareness, Confidence, Memory

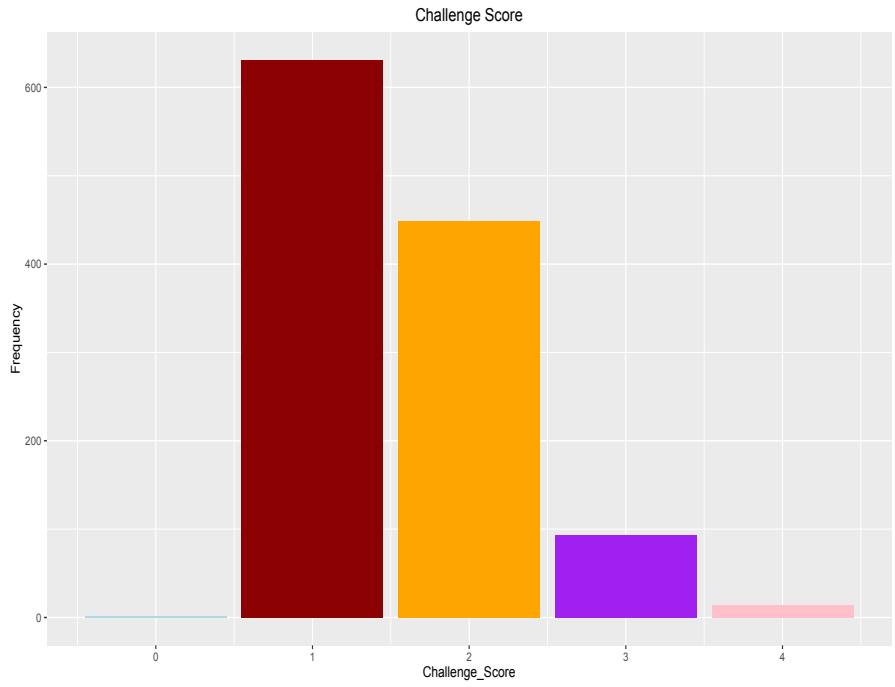
Step #2 Split & Bind: The next step is split the ‘challenge’ column into separate new columns for ‘Awareness’, ‘Confidence’, etc to ensure encoding in the next step. To do this, I have created a new dataframe called ‘challenge_split’ and combined it with ‘tekhubdata’.

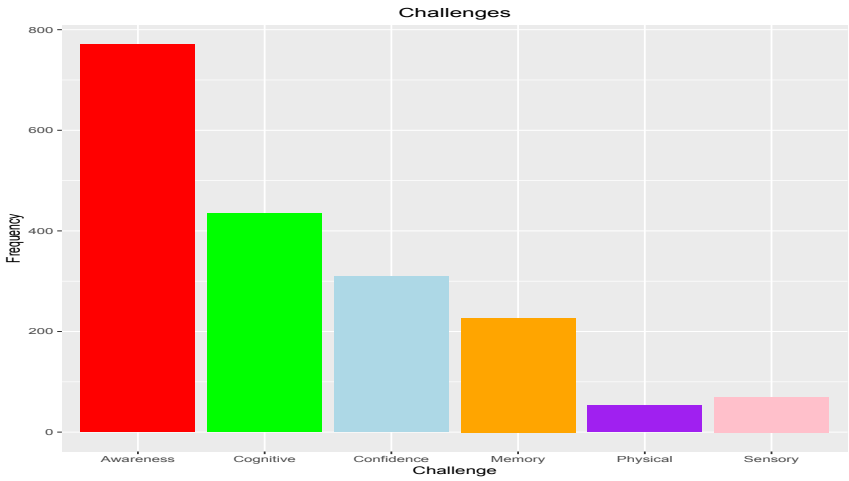
Step #3: Encoding: Using RegEx function grepl() the column ‘challenge’ is encoded, so that if the string 'Awareness' is found in the 'challenge' column, a value of 1 is added to the ‘Awareness’ column, if not a value of 0, for every row. The dataframe will look like this:

Awareness	Confidence	Cognitive	Memory	Physical	Sensory
1	1	0	1	0	0
1	0	0	0	0	0
0	1	0	0	0	0
0	1	0	0	1	0
1	1	0	0	0	0
0	1	0	0	0	0
1	1	0	0	0	0
1	0	0	0	0	0
1	0	1	0	0	0
1	1	0	0	0	0
1	1	0	0	0	0
0	0	0	1	0	0

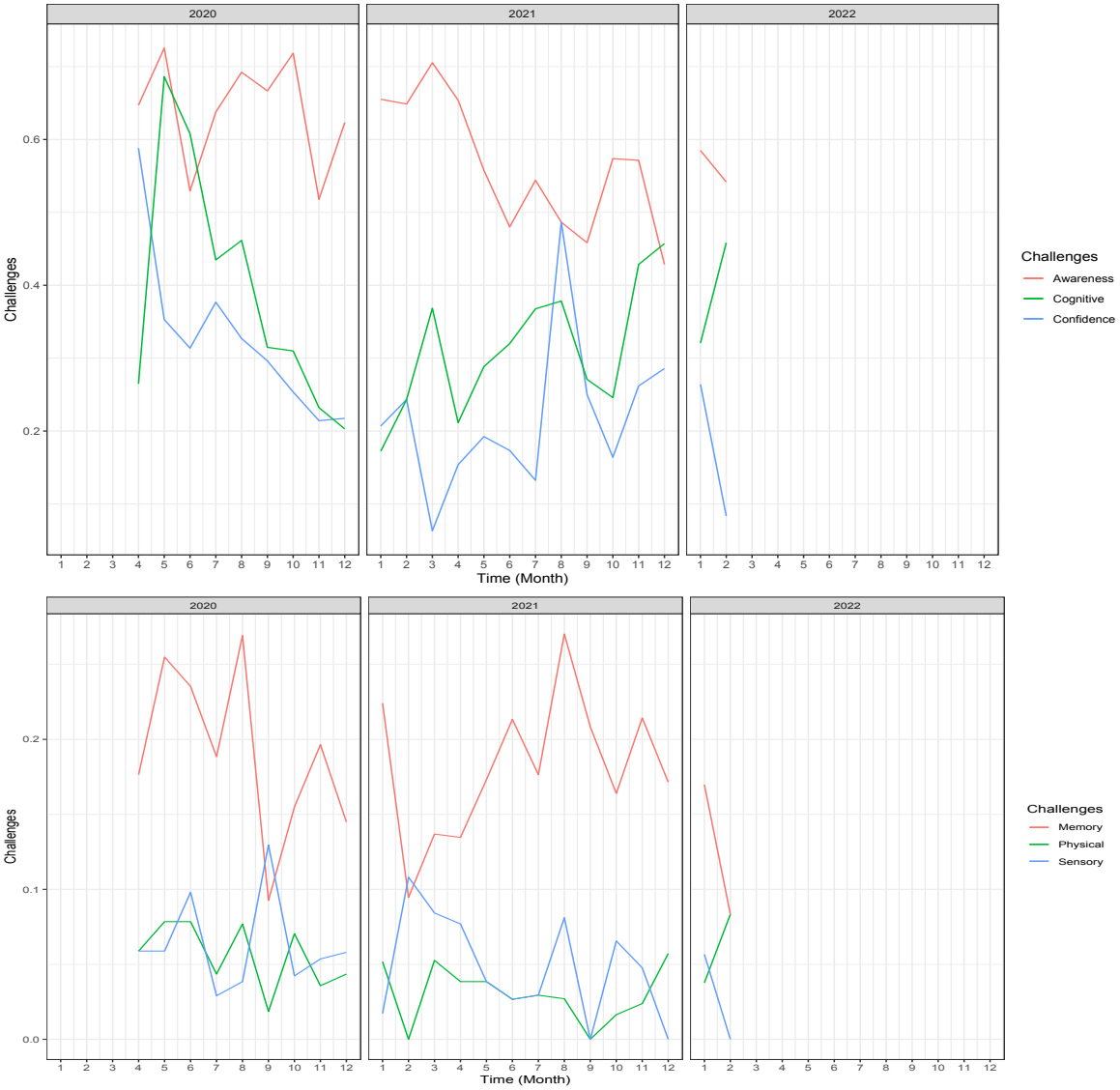
Step # 4 Column Totals: One of the primary objectives is to visualise the challenges faced by the residents, for example, which is the most common challenge. To do this we use the apply(sum) function for columns ‘Awareness’ to ‘Sensory’, to produce the plots below.

Step #5 ‘challenge_score’: Next is to create a column called ‘challenge_score’ to total all the challenges faced by a single resident(observation). For example, if a resident has challenges relating to ‘Awareness’, ‘Confidence’, ‘Memory’, he or she will get a score of 3 (1 for each challenge). The plot below shows the common values of the challenge_score. More than half of the residents had just one challenge, and a little over a quarter of the residents had more than one challenge to deal with.



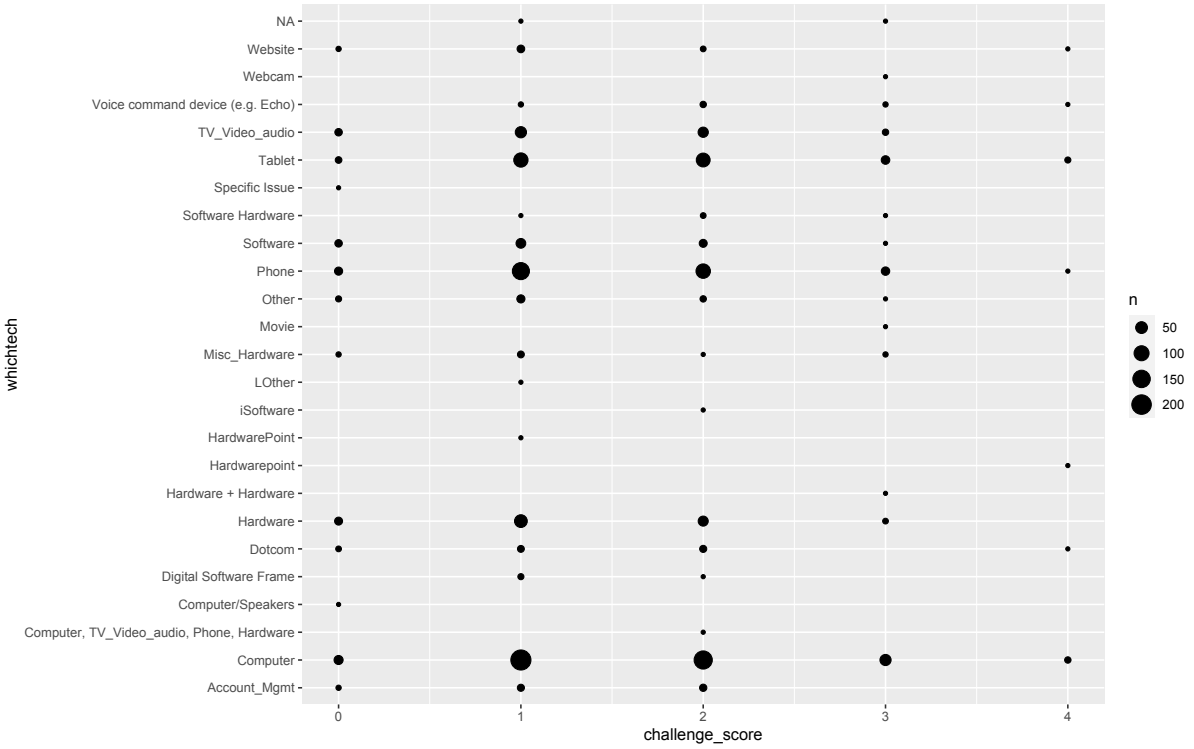


'Awareness- knowledge of available solutions to address the task' ranks as the top challenge facing seniors in this study, followed by cognitive challenges.



The plots above this, shows how the challenges have changed over time. It is important to note how most of the challenges(except Cognitive') show a downward trend over the two years. This could be an important metric in support of the Tech Support team as an overall success of the project.

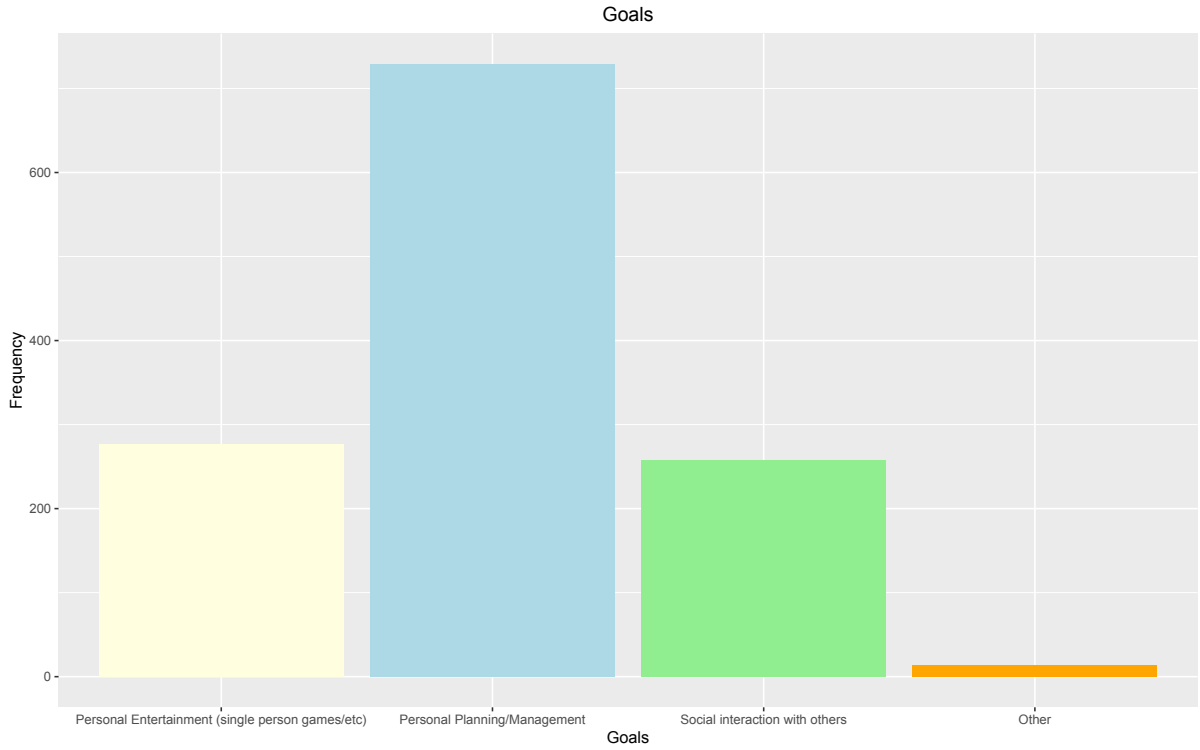
The plot below shows the relationship between the type of technology or device and the challenge score.



GOALS

Answers to the question "What are the main goals for technology use?" is recorded under these categories.

Social interaction with others
Personal Entertainment (single person games/etc)
Personal Planning/Management
Other



NOTE

The ‘note’ variable in this dataset is the result of the notes made by the Tech Support person. Though it comprises of unstructured data, several Text Mining methods were applied to understand and organise the data.

Step#1Unnest: The ‘note’ column is broken down into a collection of words (called ‘Tokens’)

word	n
<chr>	<int>
the	6874
to	5123
and	3230
it	2762
was	2657
a	2181
i	2081
that	1688
on	1584
she	1580

Step #2 Stopwords: A pre-stored data of stop words are retrieved.
Step#3 A list of words after filtering the stopwords is produced.

word	n
resident	1284
phone	497
email	440
issue	377
computer	347
set	303
printer	297
screen	275
password	271
app	251
time	249
ipad	242
account	239
zoom	237

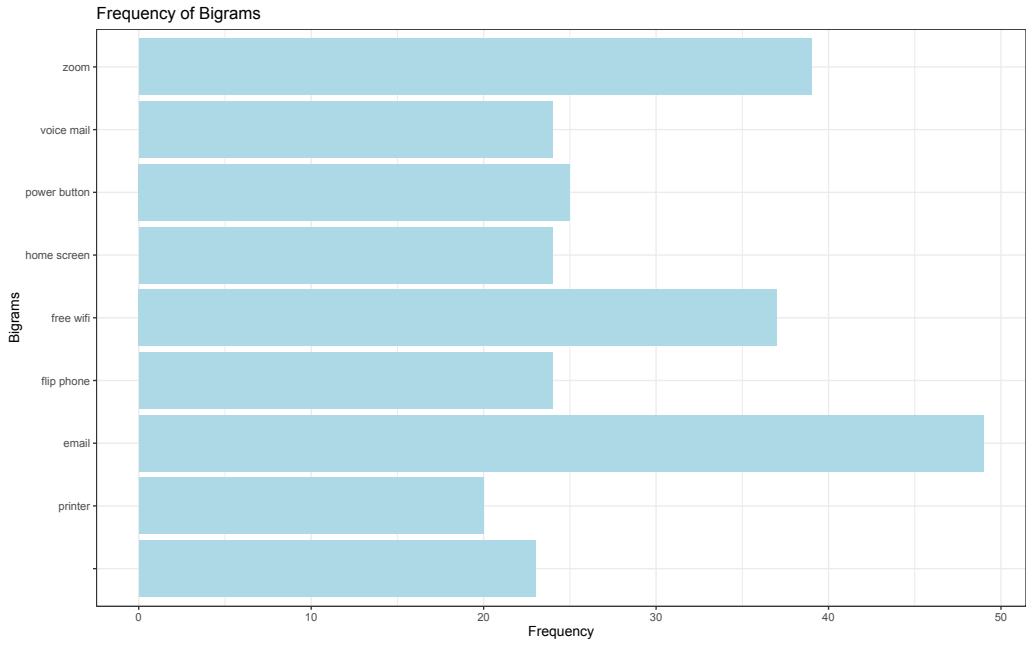
Step#4 ngrams: a list of 2-word phrases called bigrams is extracted from the ‘note’ variable using the ngram(n=2) function .

ngrams	n
<chr>	<int>
how to	528
on the	482
it was	478
to the	476
of the	437
able to	384
wanted to	321
in the	296
so i	283
she was	282

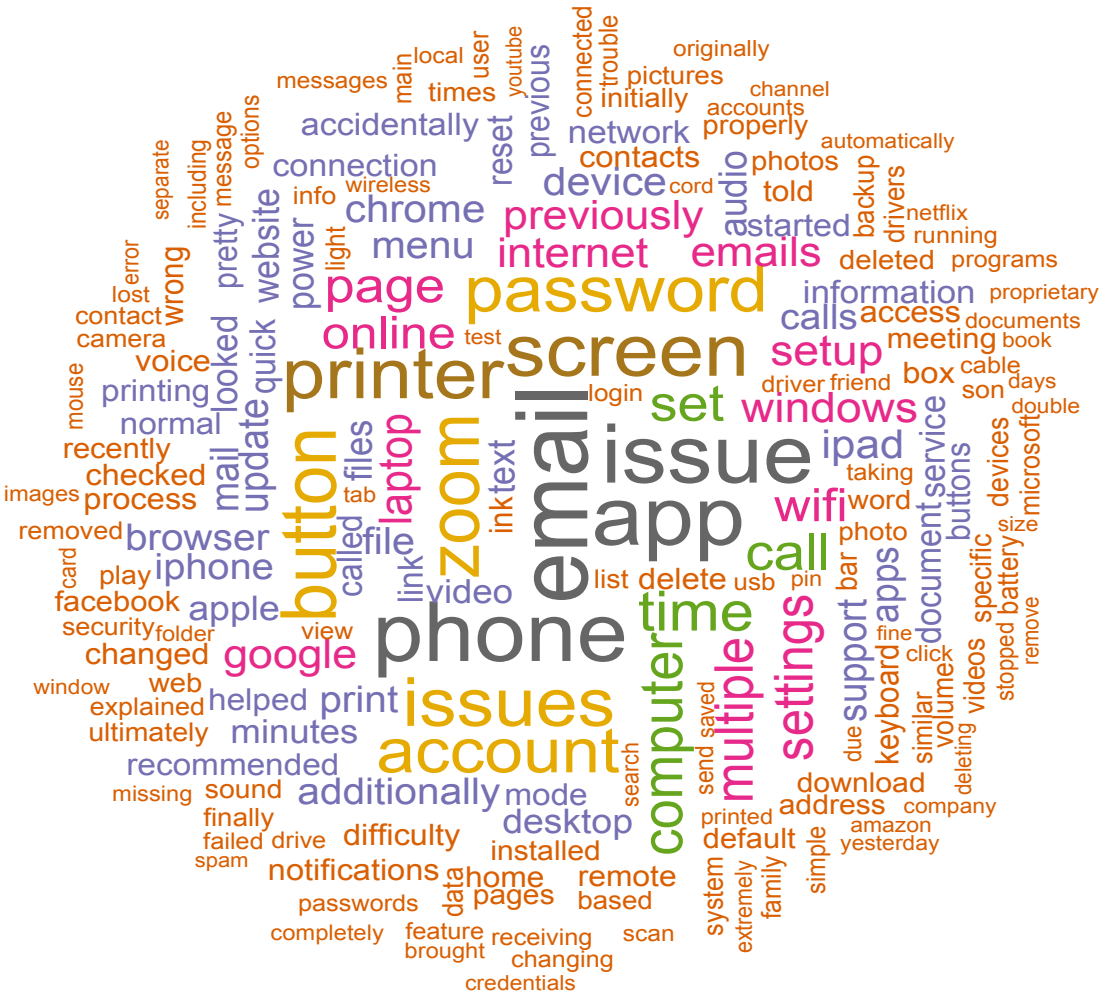
Step #5: Then the bigrams are separated into 2- words ‘word1’ & ‘word2’
Step#6: This list is filtered for the stopwords.
Step#7: The filtered list is sorted in descending order and stored under ‘bigrams_counts’.

word1	word2	n
zoom	call	39
free	wifi	37
email	address	27
power	button	25
flip	phone	24
home	screen	24
voice	mail	24
multiple	times	23
email	account	22
resident's	printer	20

Step#7: The above list is combined again under ‘bigrams’ to produce the plot below

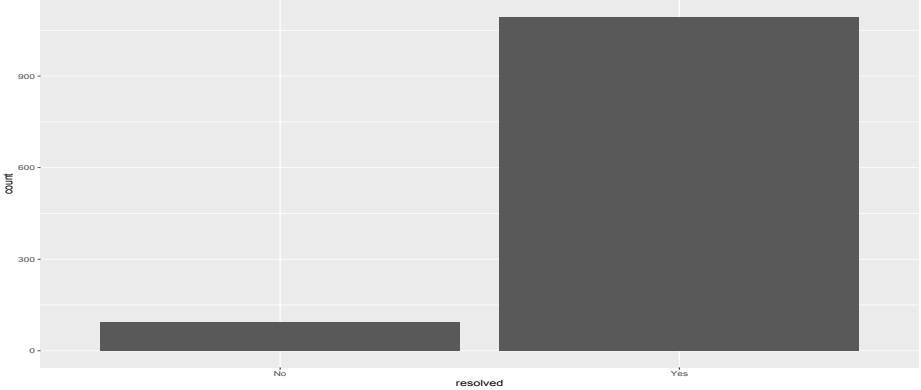


Emails, Zoom (calls, meetings, accounts), and free Wifi are the most-mentioned topics from the analysis of the unstructured data of the ‘note’ column. Zoom appearing prominently shows the frustration faced by the seniors in not able to connect with their families, due to Covid. It is worth noting here, that even though this variable has provided some great insights, these words are not the words of the senior residents themselves, instead of the Tech Person recording this on behalf of the residents. Based on the words/phrases mentioned a Wordcloud is a great way to visualize the problems facing the seniors.

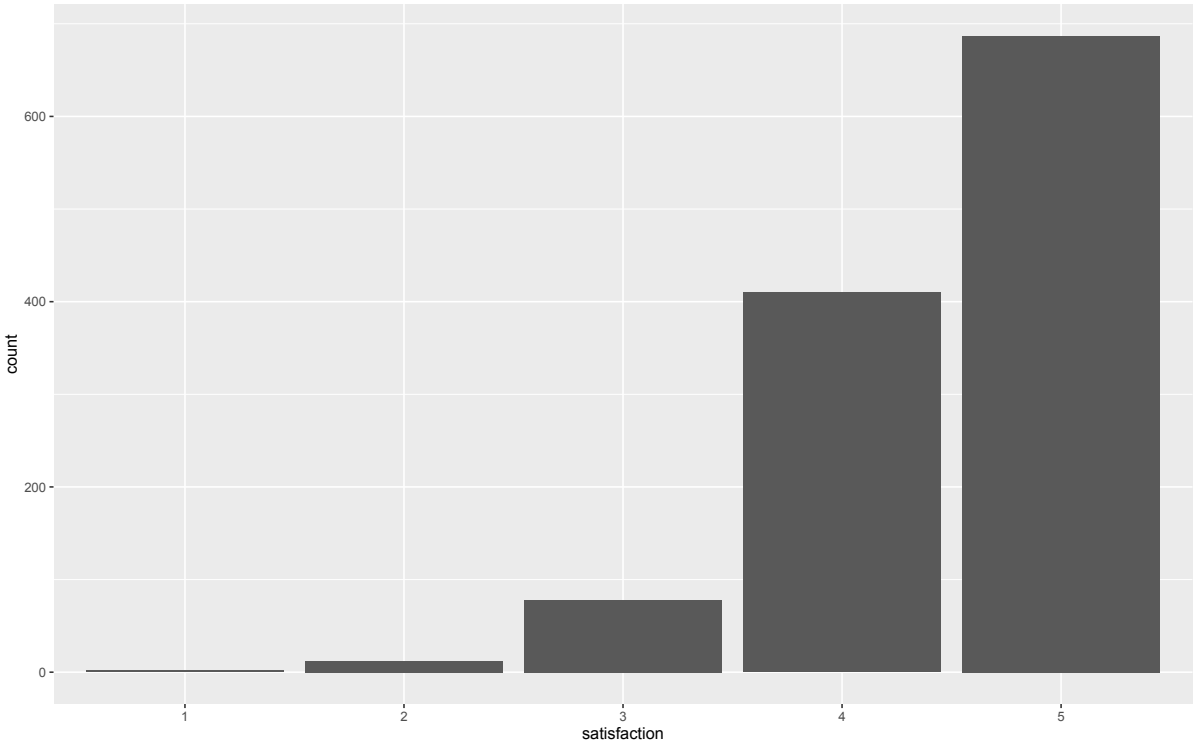


RESOLVED

The variable 'Resolved' shows the success rate of the Tech Support. This variable could be used as a dependent variable for further analysis like Random Forest or Logistic Regression.



SATISFACTION



Residents were almost always satisfied with the Tech Support, with over 90% giving a high satisfaction score.

CONCLUSION

One of the things we can discern from this study is that staying connected with their families, friends is extremely important for the senior residents of the JSL. But it is the micro-problems of not able to find a button in their phone, or not able to get their favourite show on their TV etc can be really demoralizing for the elderly. But the biggest challenge was shown to be ‘Awareness’- knowledge of solutions available. Clearly Tech Support is doing a yeoman’s service in bringing these challenges down. And we live in hope, that companies that make these devices are sensitive to these challenges faced by the elderly. And that the senior citizens across the world are as progressive and enthusiastic about adopting new technology.