Canadian Podcast Listeners 2018 Data Analyzing - Sprint 2

Aimin Amy Hu

2019-07-24

Sprint_2

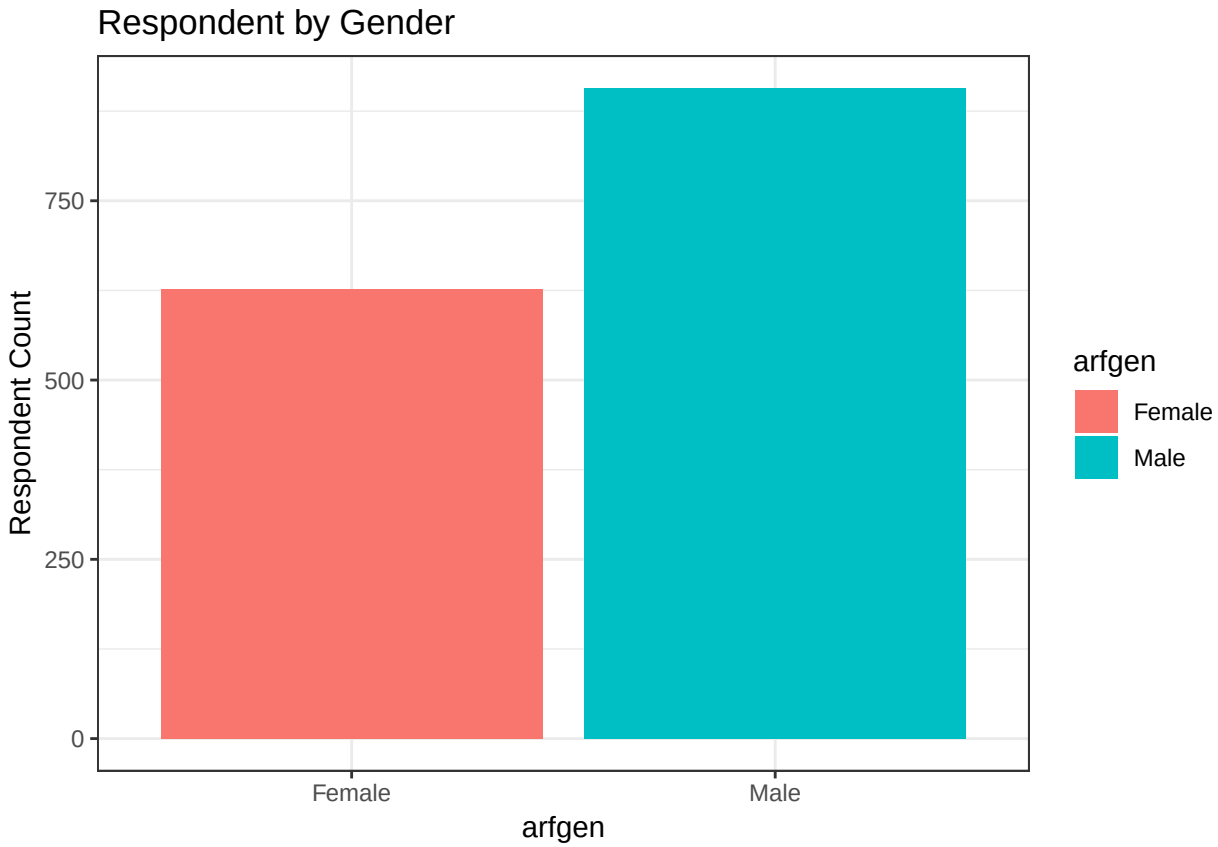
This is continue work from Sprint_1

Part 3: Data Visualization

In this section, I will use ggplots package from R to plot some graphics of this dataset. This will help us to understand the data better through visualzation.

A: Visualization of respondent gender distribution

arfgen: gender

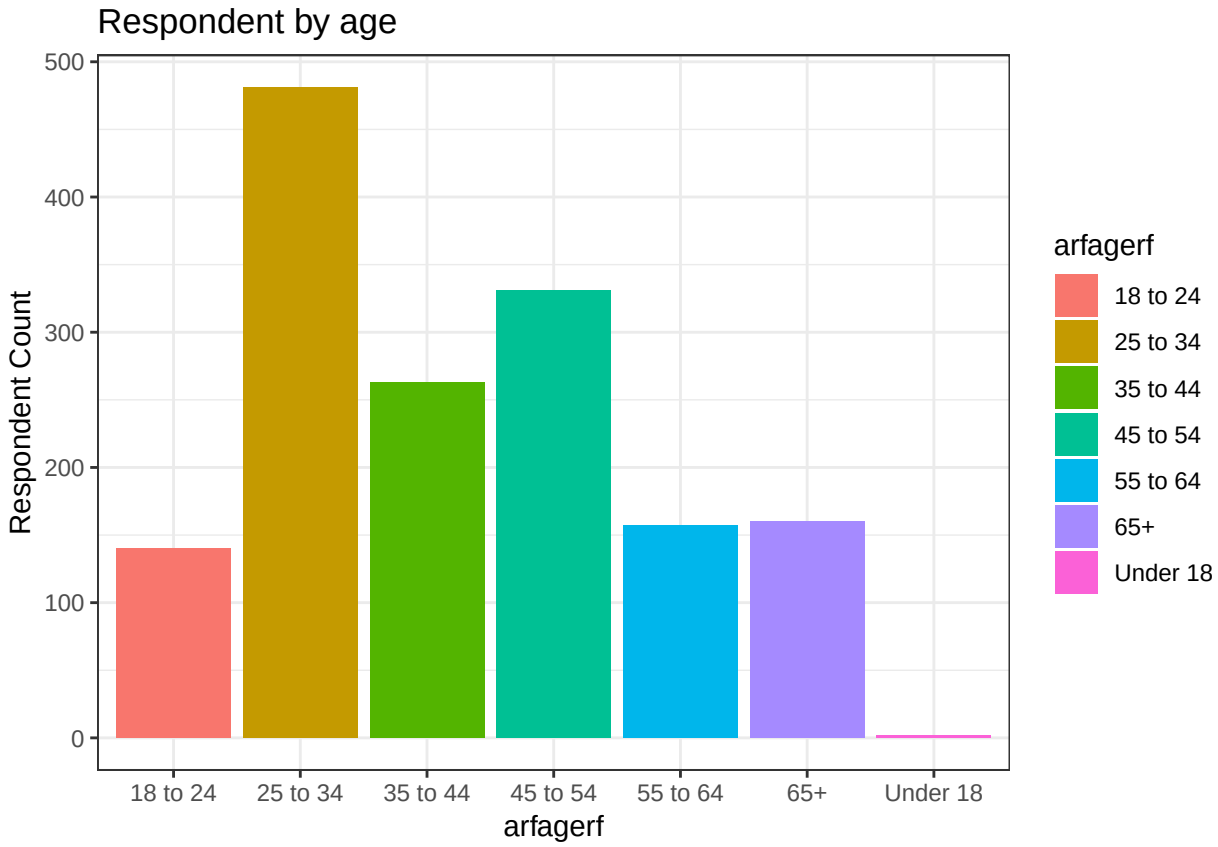


Observation from above plot:

There are about 910 males and 625 females among 1534 listeners.

B: Visualization of respondent age distribution

arfagerf: age



Observation from the plot

1: Age group of 25 to 34 has higher number of listeners among all age groups. This might tell us some insight information such as: + Listeners in this age group, has more leisure time to spend. + Tend to use computer or phone more time.To understand them more, we will need to have futher analysising combine with other variables.

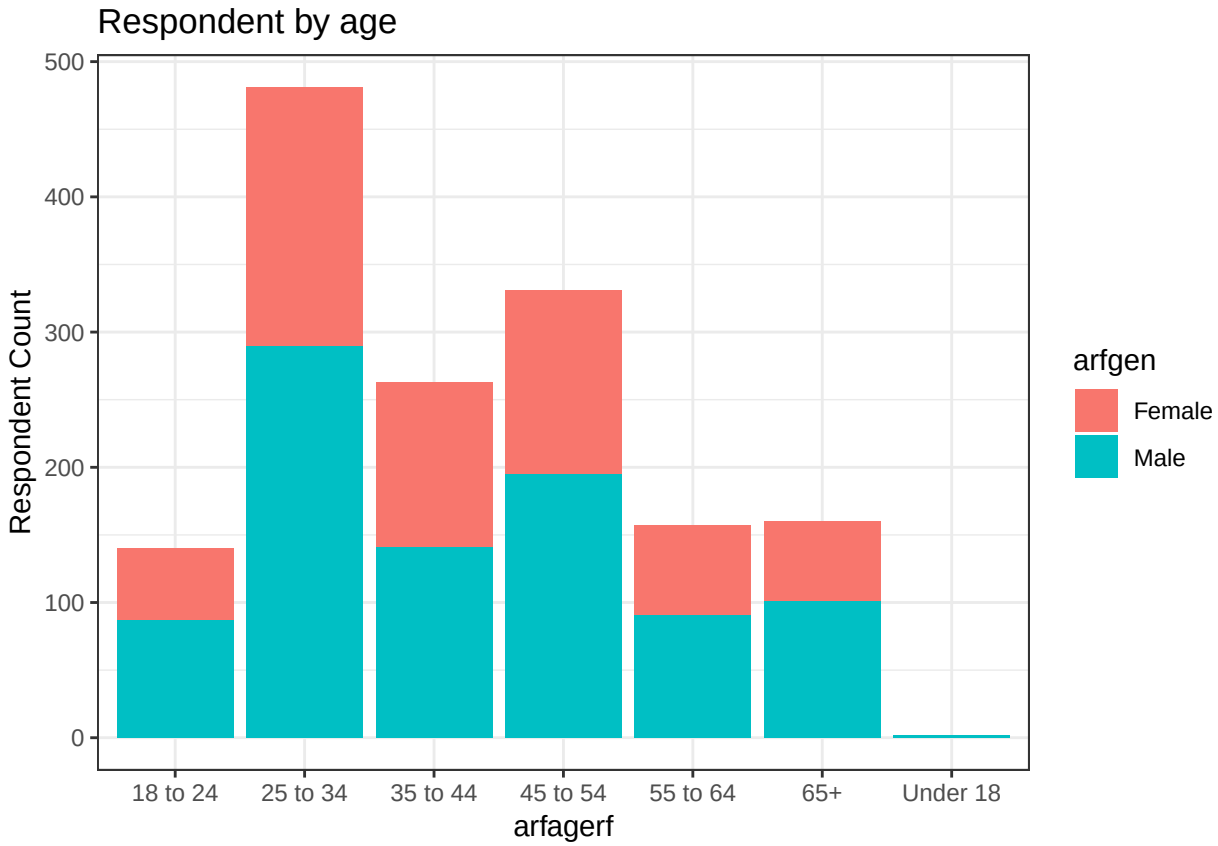
2: There are just few listeners under age 18. This might be: + People under age 18 might not be interested in podcast or there are not much podcast programs are for this group?

- Might not consider this group as a survey group?

3: This plot is clearly telling us age distribution for all 1534 respondents.

C:Visualization of respondent's gender and age distribution. Do a combination of gender and age plot to see the distribution

arfgen: gender arfagerf: age

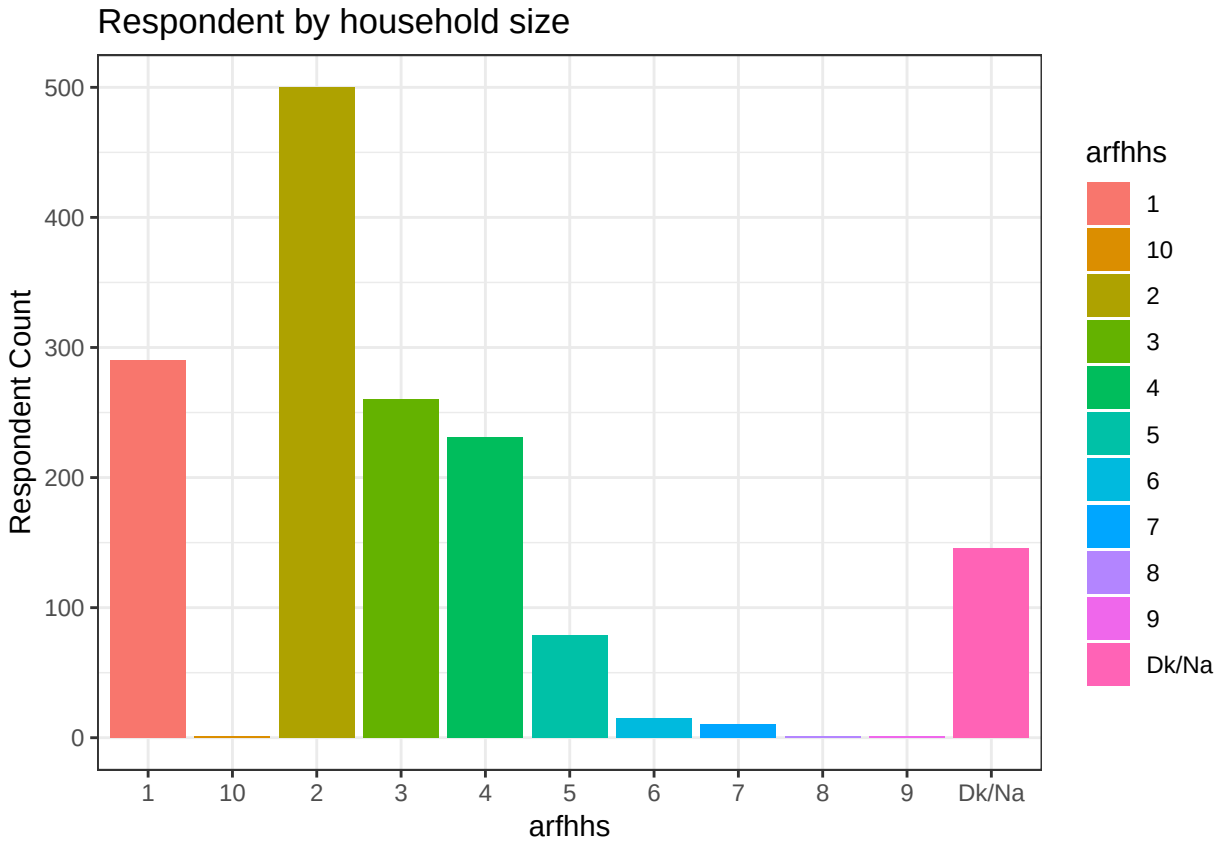


Observation from the plot

1: Respondents in age group of 25 to 34 are the most respondents group in both male and female groups. 2: The second age group with more respondents is age 45 to 54 in both male and female groups. 3: The 3rd group is age 35 to 44 in both male and female groups. 4: In female group, respondents in age group of 18 to 24 are almost same to respondents in age group of 65+. Respondents in age group of 55 to 64 is slightly higher than respondents in age group of 18 to 24. 5: In male group, respondents in age 18 to 24, in age 55 to 64 and in age 65+ are almost same numbers. 6: There are a few respondents under 18 in male group.

D: Visualization of respondent's household size distribution

arfhhs: household size

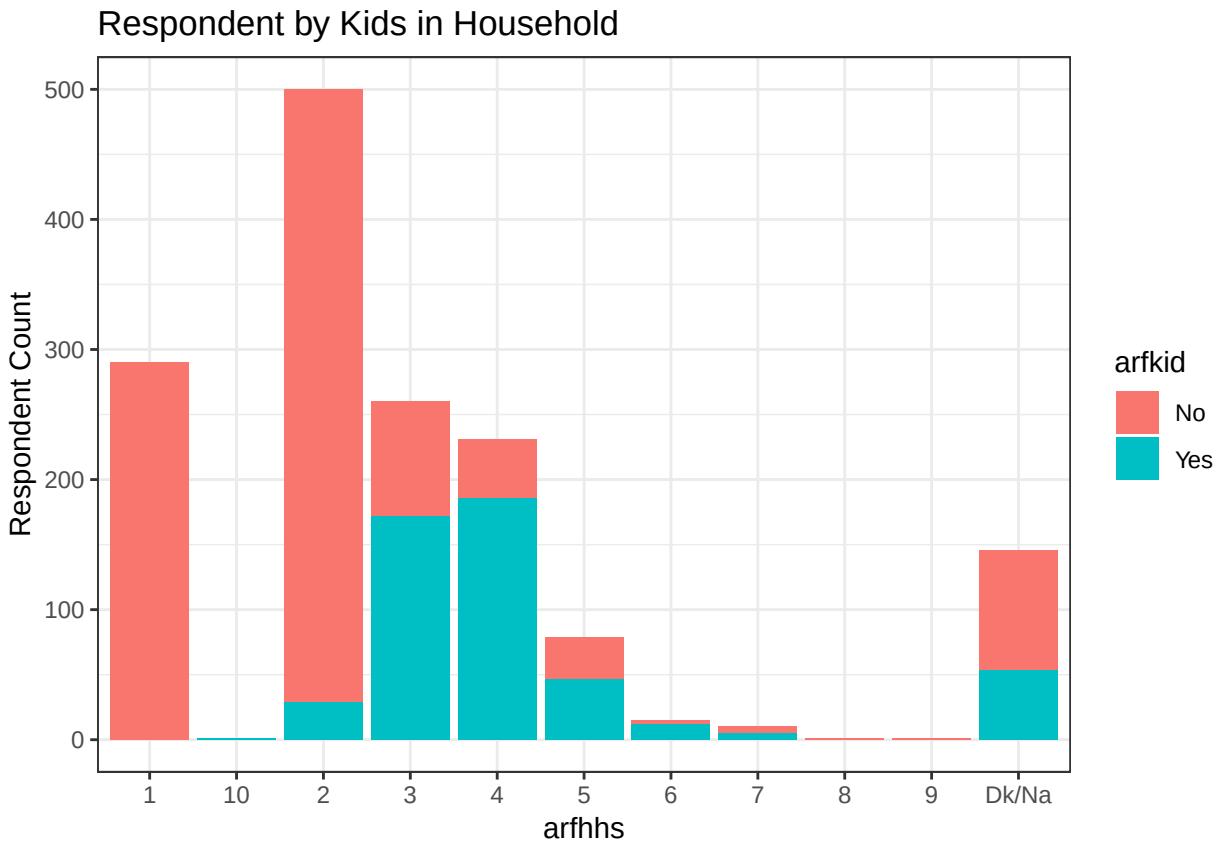


Observation from the plot

1:Among the 1534 respondents, there are 500 respondents with household size 2. They might be married couples without kids. 2: There are about 290 respondents with household size 1. They might be singles. 3: There are about 260 respondents with household size 3. They might be married couples with one kid. 4: There are about 230 respondents with household size 4. They might be married coupples with 2 kids. 5: There are about 190 respondents did not give household size information.

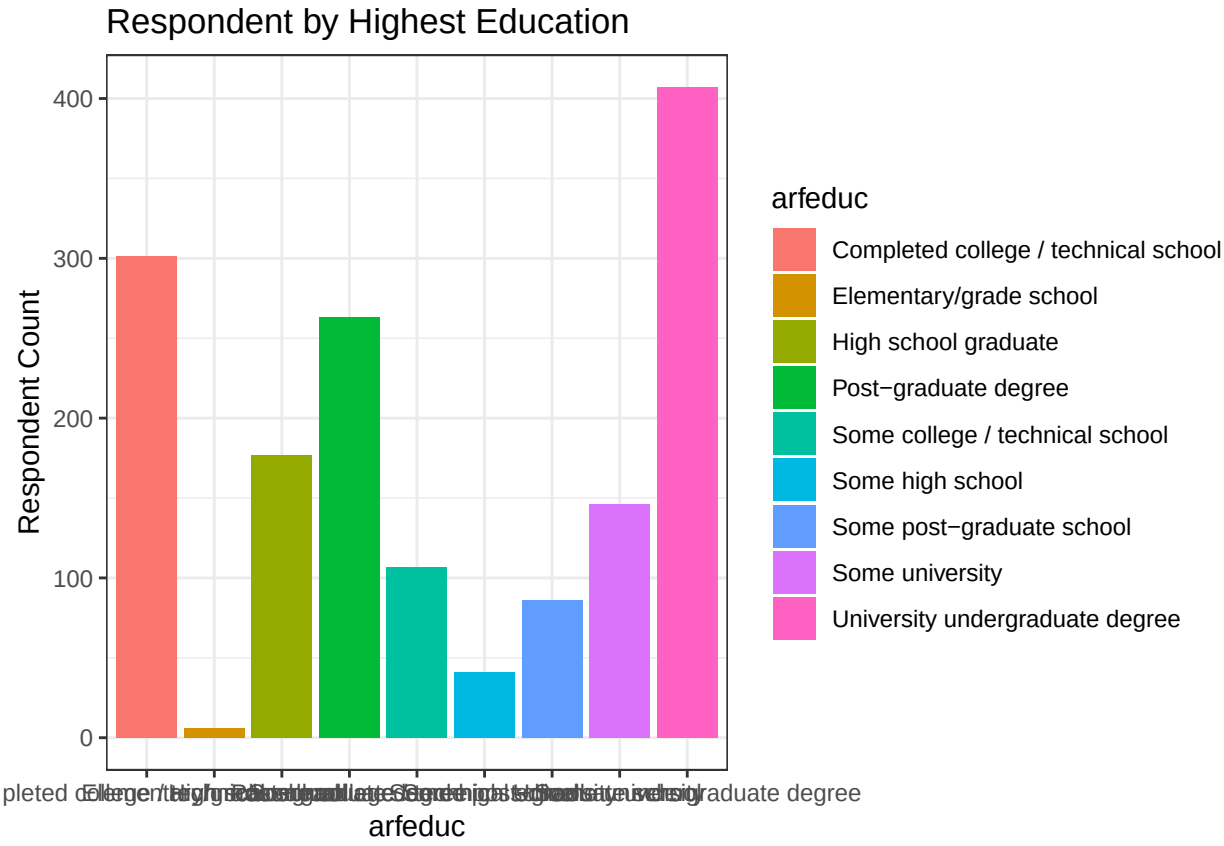
E: Visualzation of Combine Household size abd Kids in Household for Analyzing

arfhhs: household size arfkid: kids in household



F: Visualzation of respondents' edcuation level

arfeduc: highest education

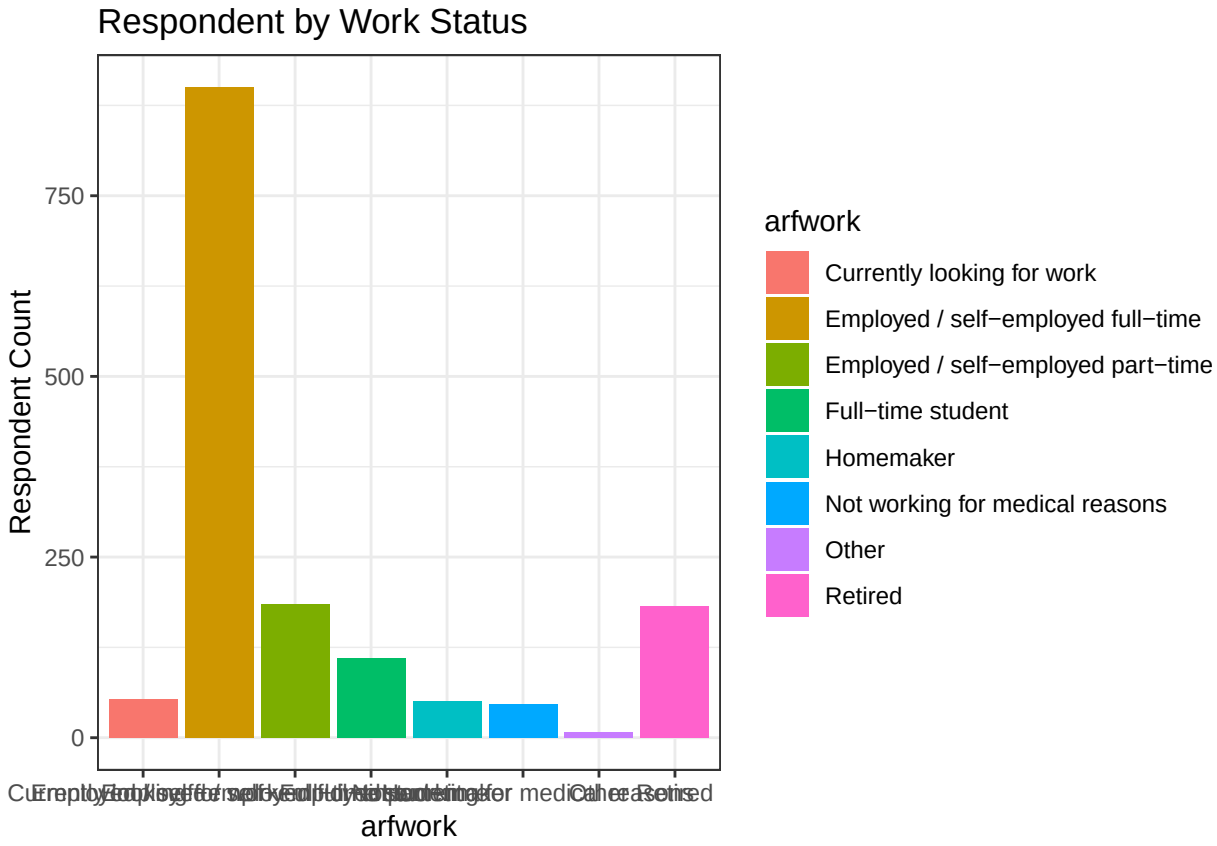


Observation from the plot

1: The hightest number of respondents is 410 with University undergraduate degree. 2: The second hightest number of respondents is 300 with Completed college/technical school. 3: There are about 260 respondents who have Post-graduate degree. 4: There are about 250respondents with High school graduate or Some high school or Elementary/grade school.

G: Visualzation of respondents' Work Status

arfwork: work status

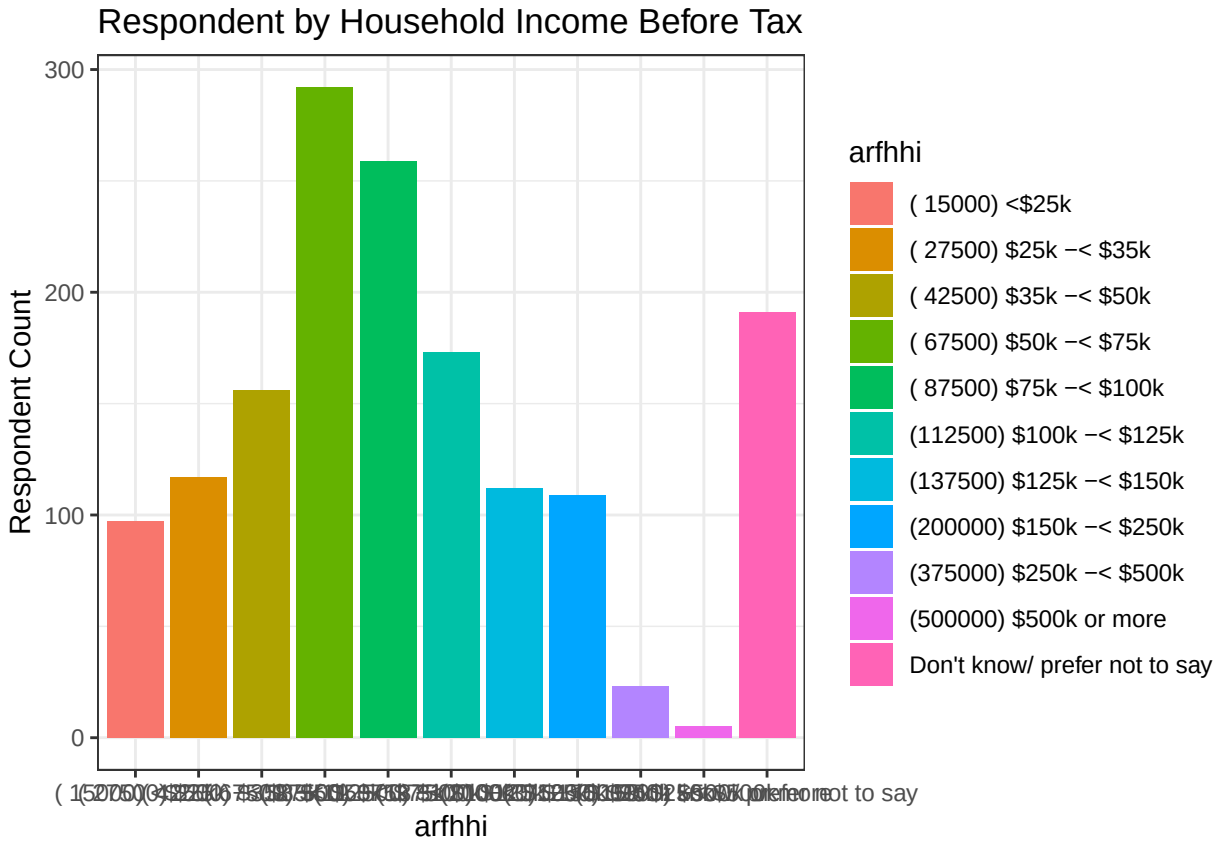


Observation from the plot

1:More than 55% of respondents (is about 890 people) are employed/self-employed full-time. 2:There are about 190 respondents are employed/self-employed part-time. 3: About 190 respondents are retired. 4: About 110 respondents are full-time student.

H: Visualzation of respondents' Household Income Before Tax

arfhhi: household income before tax



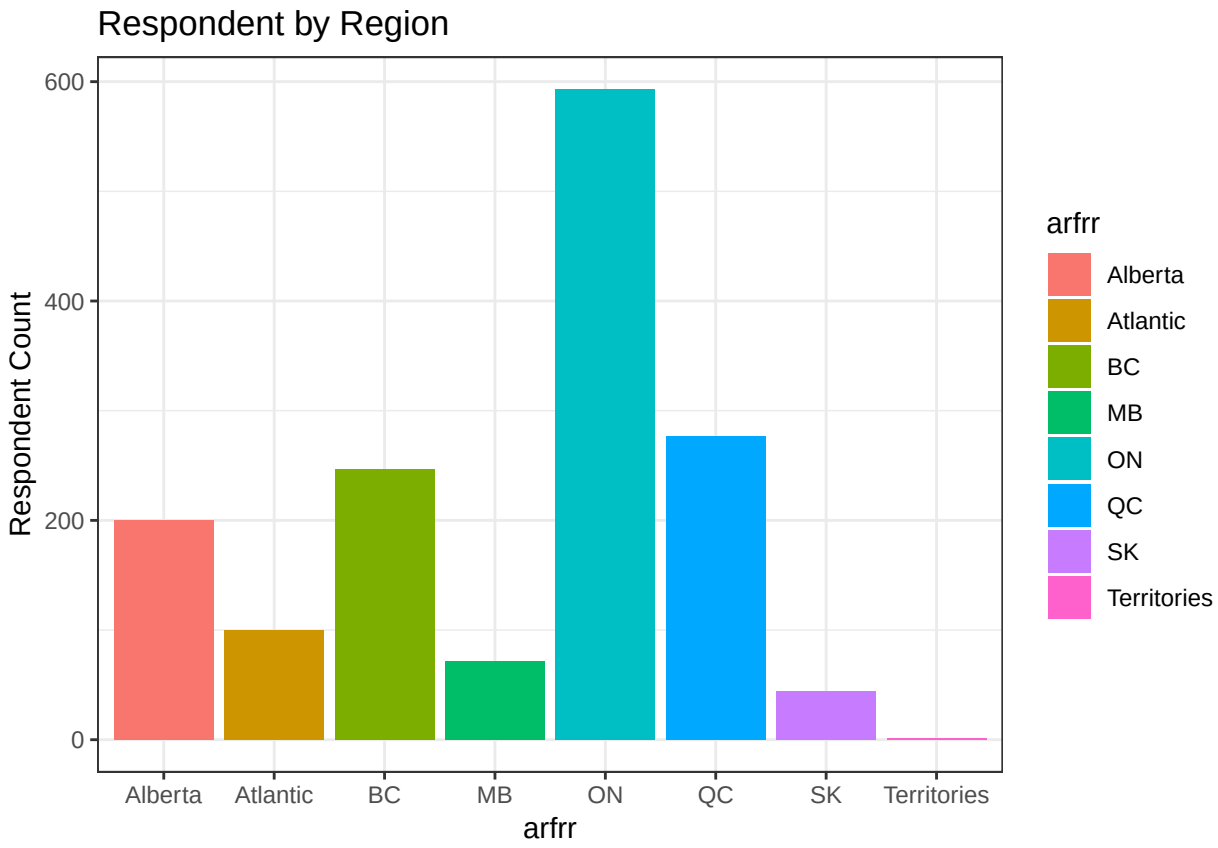
Observation from the plot

1:There are only about 90 respondents has household income less than $25k. 2: There are about 190 respondents who selected "Don't know/prefer not to say". 3: Majority of respondents have household income fall into $50k - $75k, $75k - $100k and $100k - $125k.

I: Visualzation of geographic of Listeners

arfrr: region arfpr: province

There are two attributes but recorded same data. Hence, I will use arfrr(region) for the plot.



Observation from the plot

1: With no surprise, Ontario has more respondents than any other region in Canada. 2: Terrritories has least respondents, just few respondents there.

J: Visualization of Respondents' Recency

I will look at below 3 attributes:

qp4nma: specific podcasts listened to in past month qp4tya : type of podcasts listened to in past month qp4gnaa: genres of podcasts listened to in past month

1: Visualzation of Specific podcasts listened to in past month (qp4nma)

qp4nma: specific podcasts listened to in past month

- All values in this variable are string, we will use "group_by" function to group the podcast name together, then to check how many podcasts were listened by the 1534 listeners in past month.

top 10 podcasts listened among all the listeners

- We see there are top podcasts that respondents selected on the survey. However, the top 3 among the top 10 are not a really podcasts. There are 105 respondents answered with "Other/Unknown Podcast". There 89 respondents answered with" N/A, None" and 70 respondents answered with " Don't Know".

2: Visualization of Type of podcasts listened to in past month (qp4tya)

qp4tya : type of podcasts listened to in past month

- All values in this variable are string, we will use "group_by" function to group the type of podcasts together, then to check how many types of podcasts were listened by the 1534 listeners in past month.

```
## [1] 43
```

There are 43 missing values. These missing values could be no answers from the respondents. We will combine with qp4nma to determine how we are going to deal with these 43 NA's.

```
##                                             qp4nma qp4tya
##  1:                                     ZENandTECH   <NA>
##  2:                           Other/Unknown Podcast   <NA>
##  3:                                       N/A, None   <NA>
##  4:                           Other/Unknown Podcast   <NA>
##  5:                                   CBS Sports NY   <NA>
##  6:                           Other/Unknown Podcast   <NA>
##  7:                                       Triforce!   <NA>
##  8:                           Other/Unknown Podcast   <NA>
##  9:                           Other/Unknown Podcast   <NA>
## 10:                                   NPR (unspec.)   <NA>
## 11:                           Other/Unknown Podcast   <NA>
## 12:                                       N/A, None   <NA>
## 13:                           Other/Unknown Podcast   <NA>
## 14:                           Other/Unknown Podcast   <NA>
## 15:                           Other/Unknown Podcast   <NA>
## 16:                           Other/Unknown Podcast   <NA>
## 17:                                       N/A, None   <NA>
## 18:                           Other/Unknown Podcast   <NA>
## 19:                                       N/A, None   <NA>
## 20:                           Other/Unknown Podcast   <NA>
## 21:              The Crisis of Civilization Podcast   <NA>
## 22:                                       N/A, None   <NA>
## 23:                         Radio-Canada (unspec.)   <NA>
## 24:                           Other/Unknown Podcast   <NA>
## 25:                           Other/Unknown Podcast   <NA>
## 26: The Mike Alkin Show: Talking Stocks Over a Beer   <NA>
## 27:               The Debaters with Steve Patterson   <NA>
## 28:                           Other/Unknown Podcast   <NA>
## 29:                           Other/Unknown Podcast   <NA>
## 30:                           Other/Unknown Podcast   <NA>
## 31:                           Other/Unknown Podcast   <NA>
## 32:                           Other/Unknown Podcast   <NA>
## 33:                           Other/Unknown Podcast   <NA>
## 34:                           Other/Unknown Podcast   <NA>
## 35:                           Other/Unknown Podcast   <NA>
## 36:                           Other/Unknown Podcast   <NA>
## 37:                                       N/A, None   <NA>
## 38:                           Other/Unknown Podcast   <NA>
## 39:                           Other/Unknown Podcast   <NA>
## 40:                 The Next Chapter from CBC Radio   <NA>
## 41:                           Other/Unknown Podcast   <NA>
## 42:                           Other/Unknown Podcast   <NA>
## 43:                           Other/Unknown Podcast   <NA>
##                                             qp4nma qp4tya
```

We combined columns qp4nma and qp4tya to determine the NA's in variable qp4tya. We saw a few records with podcast names in variable qp4nma, but no values in variable qp4tya. We used podcast names to check online and found out which type is and replaced the NA with actual type of podcast.

Replace NA's with actual type in listener_clean data

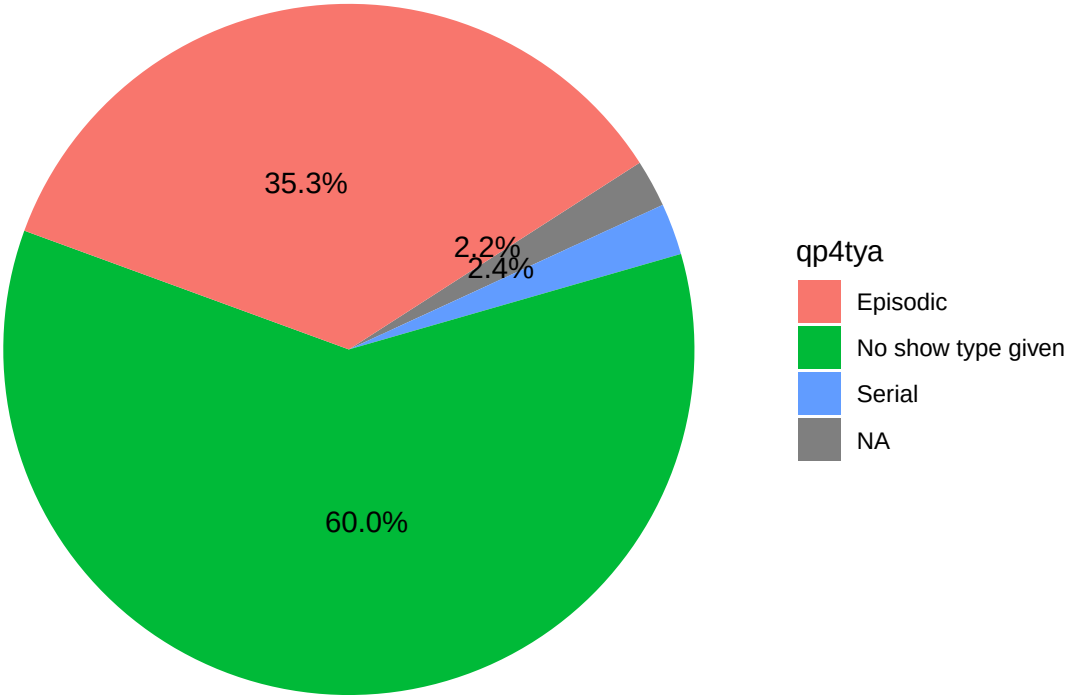
After replaced some NA's, check how many NA's values in variable qp4tya

```
## [1] 34
```

For the rest of NA's in variable qp4tya, we will keep them as NA's as there are no podcast names in variable qp4nma.

```
## # A tibble: 4 x 2
##   qp4tya             type_count
##   <chr>                   <int>
## 1 Episodic                  542
## 2 No show type given        921
## 3 Serial                     37
## 4 <NA>                       34
```

We can use ggplot to create a pie chart to visualize it.



From above pie chart, we see 60% of respondents did not give show type for their listening in the past month on the survey form. 35.3% of respondents listened to Episodic podcasts in the past month. 2.4% of respondents were listening on Serial podcasts in the last month and 2.2% of respondents with NA.

3:Visualization of Genres of podcasts listened to in past month(qp4gnaa)

qp4gnaa: genres of podcasts listened to in past month

```
## [1] 97
```

```
##                                        qp4nma qp4gnaa
##  1:                            CBC (unspec.)    <NA>
##  2:                            CBC (unspec.)    <NA>
##  3:                            CBC (unspec.)    <NA>
##  4:                               ZENandTECH    <NA>
##  5:                            CBC (unspec.)    <NA>
##  6:                      Other/Unknown Podcast    <NA>
##  7:                               Food Psych    <NA>
```

```
##  8:                      La Chapelle - podcast audio    <NA>
##  9:                                   N/A, None    <NA>
## 10:                            Other/Unknown Podcast    <NA>
## 11:                                    CBS Sports NY    <NA>
## 12:                                 The Daily Shoah!    <NA>
## 13:                                    CBC (unspec.)    <NA>
## 14:                            Other/Unknown Podcast    <NA>
## 15:                                       Triforce!    <NA>
## 16:                                    CBC (unspec.)    <NA>
## 17:                            Other/Unknown Podcast    <NA>
## 18:                            Other/Unknown Podcast    <NA>
## 19:                                 Comedy Bang Bang    <NA>
## 20:                      10/3: Canadian News Covered    <NA>
## 21:                                    CBC (unspec.)    <NA>
## 22:                            Other/Unknown Podcast    <NA>
## 23:                                    NPR (unspec.)    <NA>
## 24:                       Dr. Laurie Marbas Podcast    <NA>
## 25:                            Other/Unknown Podcast    <NA>
## 26:                                   N/A, None    <NA>
## 27:                                  As Maple As    <NA>
## 28:                        Corde sensible - Radical    <NA>
## 29:                                          Main    <NA>
## 30:                            Other/Unknown Podcast    <NA>
## 31:                  Big Brother Reviews & After Show    <NA>
## 32:                                    CBC (unspec.)    <NA>
## 33:                                    CBC (unspec.)    <NA>
## 34:                            Other/Unknown Podcast    <NA>
## 35:                                  American Fiasco    <NA>
## 36:                                      BrainStuff    <NA>
## 37:                            Other/Unknown Podcast    <NA>
## 38:                      Casefile True Crime Podcast    <NA>
## 39:                                   À voix haute    <NA>
## 40:                            Other/Unknown Podcast    <NA>
## 41:                                   N/A, None    <NA>
## 42:                        Corde sensible - Radical    <NA>
## 43:                            Other/Unknown Podcast    <NA>
## 44:                                   N/A, None    <NA>
## 45:                                    CBC (unspec.)    <NA>
## 46:                                    CBC (unspec.)    <NA>
## 47:                   Oprah's SuperSoul Conversations    <NA>
## 48:                            Other/Unknown Podcast    <NA>
## 49:                The Crisis of Civilization Podcast    <NA>
## 50:                                   N/A, None    <NA>
## 51:                                  As Maple As    <NA>
## 52:                             Radio-Canada (unspec.)    <NA>
## 53:                            Other/Unknown Podcast    <NA>
## 54:                            Other/Unknown Podcast    <NA>
## 55: The Mike Alkin Show: Talking Stocks Over a Beer    <NA>
## 56:                                  EW Morning Live    <NA>
## 57:                 The Debaters with Steve Patterson    <NA>
## 58:                                    CBC (unspec.)    <NA>
## 59:                                       Dreamland    <NA>
## 60:          James MacDonald - Walk in the Word Audio    <NA>
## 61:                                    CBC (unspec.)    <NA>
```

```
## 62:                            Other/Unknown Podcast    <NA>
## 63:                            Other/Unknown Podcast    <NA>
## 64:                                    CBC (unspec.)    <NA>
## 65:                                       BrainStuff    <NA>
## 66:                                    CBC (unspec.)    <NA>
## 67:                                 Dead Rock Stars    <NA>
## 68:                            Other/Unknown Podcast    <NA>
## 69:                                    CBC (unspec.)    <NA>
## 70:                         The Gossip Garden Podcast    <NA>
## 71:                                        Dekmantel    <NA>
## 72:                            Other/Unknown Podcast    <NA>
## 73:                            Other/Unknown Podcast    <NA>
## 74:                            Other/Unknown Podcast    <NA>
## 75:                            Other/Unknown Podcast    <NA>
## 76:              A New and Ancient Story: The Podcast    <NA>
## 77:                            Other/Unknown Podcast    <NA>
## 78:                            Other/Unknown Podcast    <NA>
## 79:                                     Crazy/Genius    <NA>
## 80:                            Doc Mailloux et Josey    <NA>
## 81:                                        N/A, None    <NA>
## 82:                                    CBC (unspec.)    <NA>
## 83:                              Bright Side Podcast    <NA>
## 84:                                    CBC (unspec.)    <NA>
## 85:                                        CareerJoy    <NA>
## 86:                            Other/Unknown Podcast    <NA>
## 87:                               LCT Platform Talks    <NA>
## 88:                            Other/Unknown Podcast    <NA>
## 89:                          A Prairie Home Companion    <NA>
## 90:                                    CBC (unspec.)    <NA>
## 91:                  The Next Chapter from CBC Radio    <NA>
## 92:                            Other/Unknown Podcast    <NA>
## 93:                                      Baeltestedet    <NA>
## 94:                            Other/Unknown Podcast    <NA>
## 95:                                    CBC (unspec.)    <NA>
## 96:                                    CBC (unspec.)    <NA>
## 97:                            Other/Unknown Podcast    <NA>
##                                                 qp4nma qp4gnaa
```

From above data, we saw there are podcast name in variable qp4nma but no values in qp4gnaa. We will check online to find actual genre and replace the NA's.

Replace NA's with actual type in listener_clean data

We have replaced the actual genre for all NA's which have podcast names in variable qp4nma. Now, we will check how many NA's after the replacement.

```
## [1] 36
```

Run below codes to check if there are still some NA's needed to be replaced by actual genre.

```
##                    qp4nma qp4gnaa
##  1: Other/Unknown Podcast    <NA>
##  2:             N/A, None    <NA>
##  3: Other/Unknown Podcast    <NA>
##  4: Other/Unknown Podcast    <NA>
##  5: Other/Unknown Podcast    <NA>
##  6: Other/Unknown Podcast    <NA>
```

```
##  7: Other/Unknown Podcast    <NA>
##  8: Other/Unknown Podcast    <NA>
##  9:             N/A, None    <NA>
## 10:                  Main    <NA>
## 11: Other/Unknown Podcast    <NA>
## 12: Other/Unknown Podcast    <NA>
## 13: Other/Unknown Podcast    <NA>
## 14: Other/Unknown Podcast    <NA>
## 15:             N/A, None    <NA>
## 16: Other/Unknown Podcast    <NA>
## 17:             N/A, None    <NA>
## 18: Other/Unknown Podcast    <NA>
## 19:             N/A, None    <NA>
## 20: Other/Unknown Podcast    <NA>
## 21: Other/Unknown Podcast    <NA>
## 22: Other/Unknown Podcast    <NA>
## 23: Other/Unknown Podcast    <NA>
## 24: Other/Unknown Podcast    <NA>
## 25: Other/Unknown Podcast    <NA>
## 26: Other/Unknown Podcast    <NA>
## 27: Other/Unknown Podcast    <NA>
## 28: Other/Unknown Podcast    <NA>
## 29: Other/Unknown Podcast    <NA>
## 30: Other/Unknown Podcast    <NA>
## 31:             N/A, None    <NA>
## 32: Other/Unknown Podcast    <NA>
## 33: Other/Unknown Podcast    <NA>
## 34: Other/Unknown Podcast    <NA>
## 35: Other/Unknown Podcast    <NA>
## 36: Other/Unknown Podcast    <NA>
##                    qp4nma qp4gnaa
```

There are still 35 NAs in variable qp4gnaa. Since respondents did not select/list podcast names in variable qp4nma, I will just keep these 35 NAs in the data.

use group_by genres of podcasts to see how many genres of podcasts in the data.

Visualization of Top 10 genres Between Male and Female

arfgen: gender qp4gnaa: genres of podcasts listened to in past month

	arfgen	qp4gnaa
1	Male	News & Politics
2	Male	Comedy
3	Male	Society & Culture
4	Male	Professional
5	Male	Sports & Recreation
6	Male	History
7	Male	TV & Film
8	Male	Music
9	Male	Christianity
10	Male	Personal Journals

	arfgen	qp4gnaa
1	Female	Comedy
2	Female	Society & Culture
3	Female	News & Politics
4	Female	History
5	Female	Christianity
6	Female	Music
7	Female	TV & Film
8	Female	Personal Journals
9	Female	Self−Help
10	Female	Professional

From above tables, we saw the top 10 genres of podcasts are different between male and female.

K: Visualization of Respondents' Frequency

We will look at below 3 attributes: qs4: frequency listening to audio podcasts

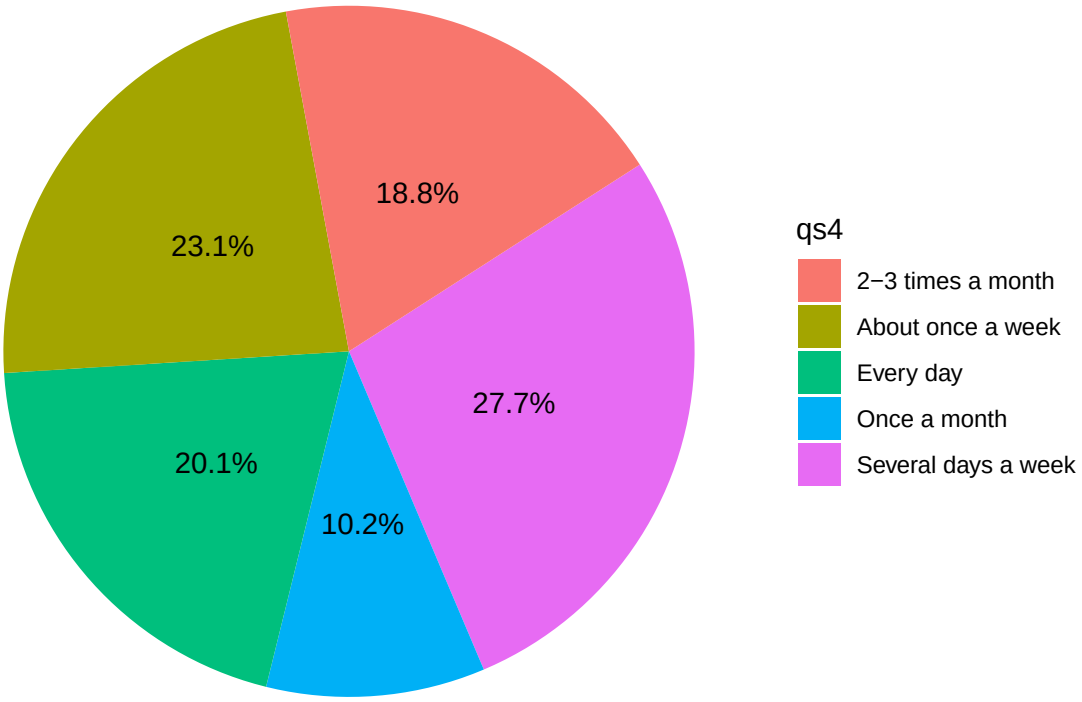
qp1c: combined time listening typical week

qp3: first started listening to podcasts

1:Visualization of Frequency listening to audio podcasts (qs4)

qs4: frequency listening to audio podcasts

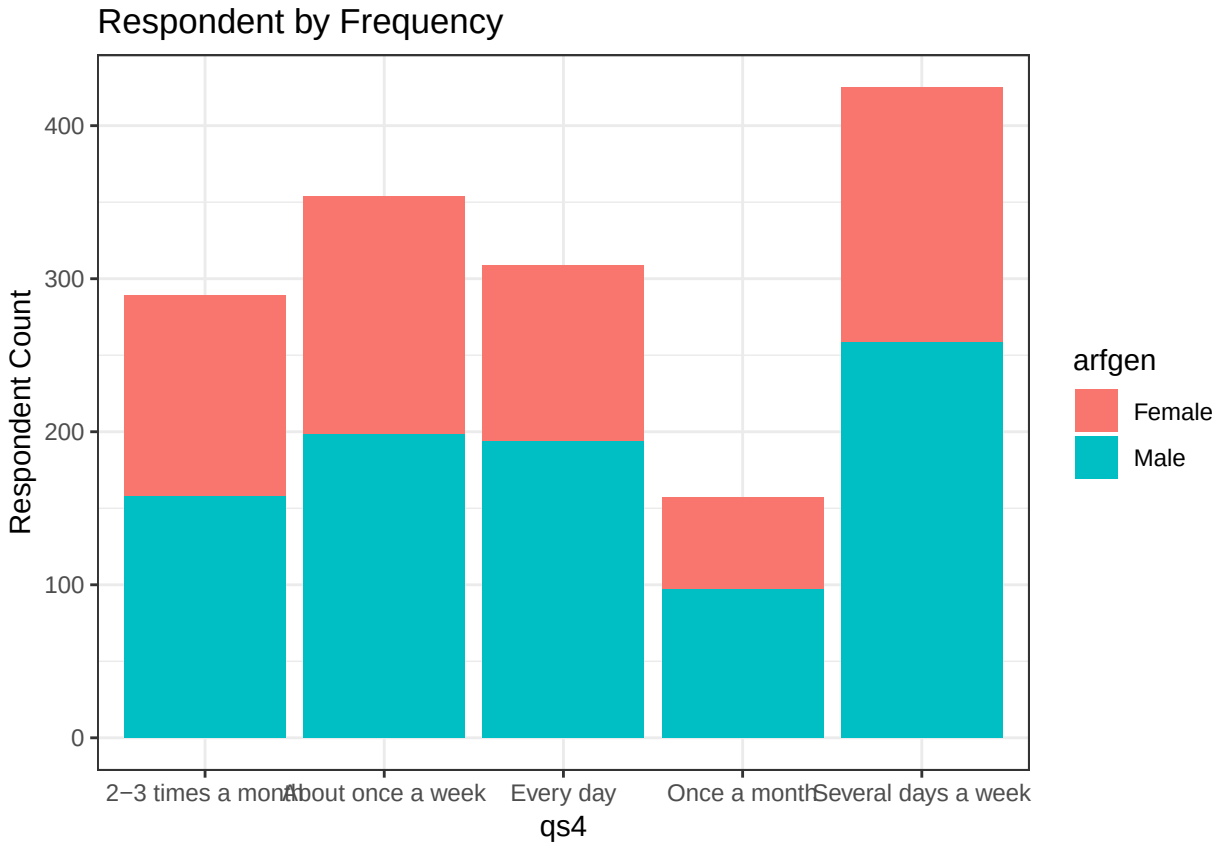
Use a pie chart to see percentage of respondents in each type of frenquency listening to audio podcasts.



We can clearly see the distribution of listeners's frenquency.Every day listeners are only 20.1%. Once a month listeners are 10.2% and 2-3 times a month listeners are 18.8%.

2: Visualization of frequency between male and female

arfgen: gender qs4: frequency listening to audio podcasts



More male listeners are in all frenquency catergories.

3:Visualization of combined time listening typical week (qp1c)

qp1c: combined time listening typical week

```
## [1] "840"                                "DNQ Listen once per week or more"
## [3] "DNQ Listen once per week or more" "360"
## [5] "150"                                "870"
```

variable qp1c recorded listeners' combined time listning typical week. The numbers in this variable are in minutes. There also have values "DNQ Listen once per week or more". This "DNQ" refers to listeners who were not asked that question because they indicated in the earlier question that they don't listen to podcasts on a weekly basis so were not asked how many minutes they listen to per week. Hence, I will remove these DNQ values from this variable and only analyze listeners who listen to podcasts on a weekly basis.

Get a subset dataset without "DNQ" in variable qp1c

Get min., mean, and max. for variable qp1c

```
##    Length     Class      Mode
##      1088 character character
```

There are 1088 rows in this subset data (446 rows removed). Will need to convert the variable to numeric for getting min., mean, and max.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0   120.0   210.0   360.7   420.0  2820.0
```

The minimum is 1. I consider this as an error. It could be wrong recorded value as people will not listen to a podcast only for 1 minute.

```
## [1] 1
```

There are only one record with qp1c==“1”, hence I will remove this one record and get min., mean and max. again.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       4     120     210     361     420    2820
```

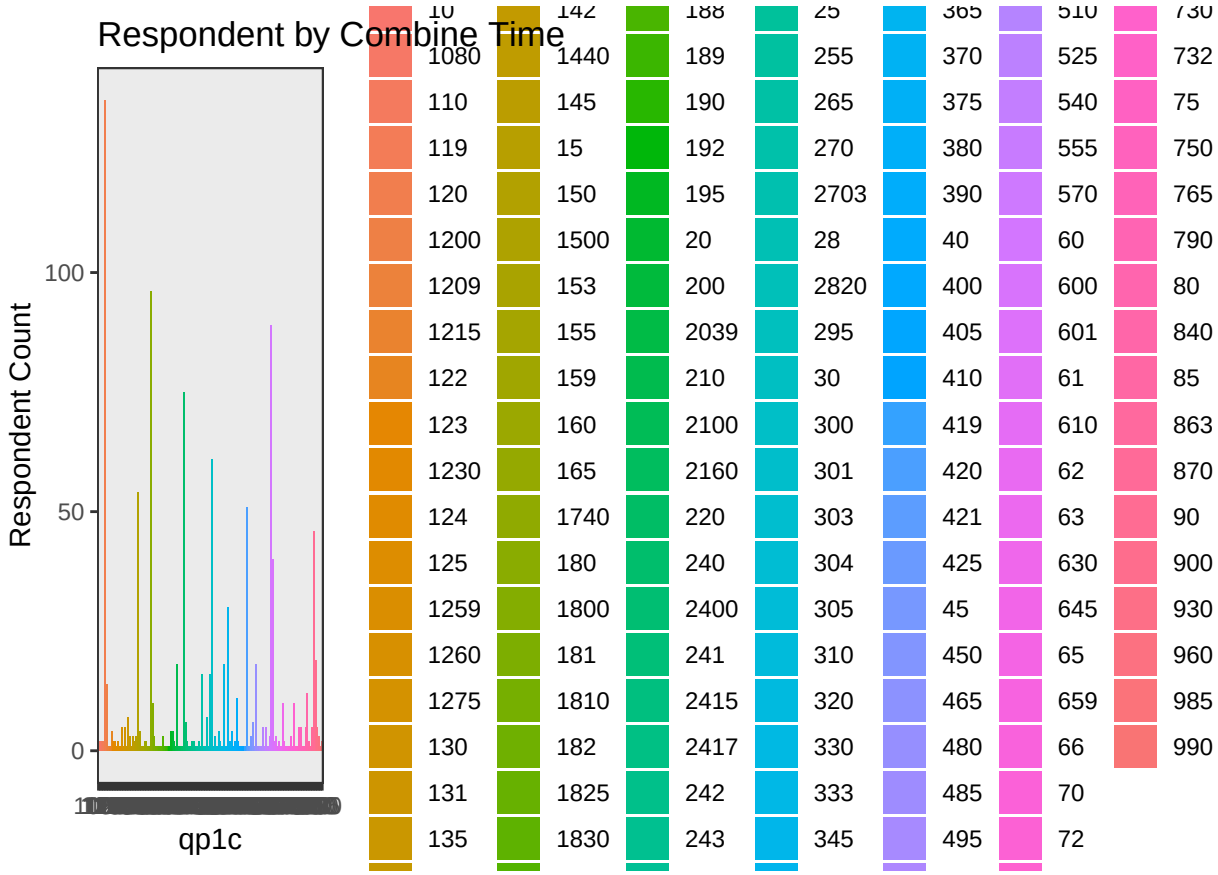
Now, the min. is 4 minute in this variable. It might be an error too as people don’t use 4 minute to indicate their approximate time. Hence, I will remove this record too from the subset data and get summary again.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.0   120.0   210.0   361.4   420.0  2820.0
```

Check how many records for qp1c ==“2820”

```
## [1] 7
```

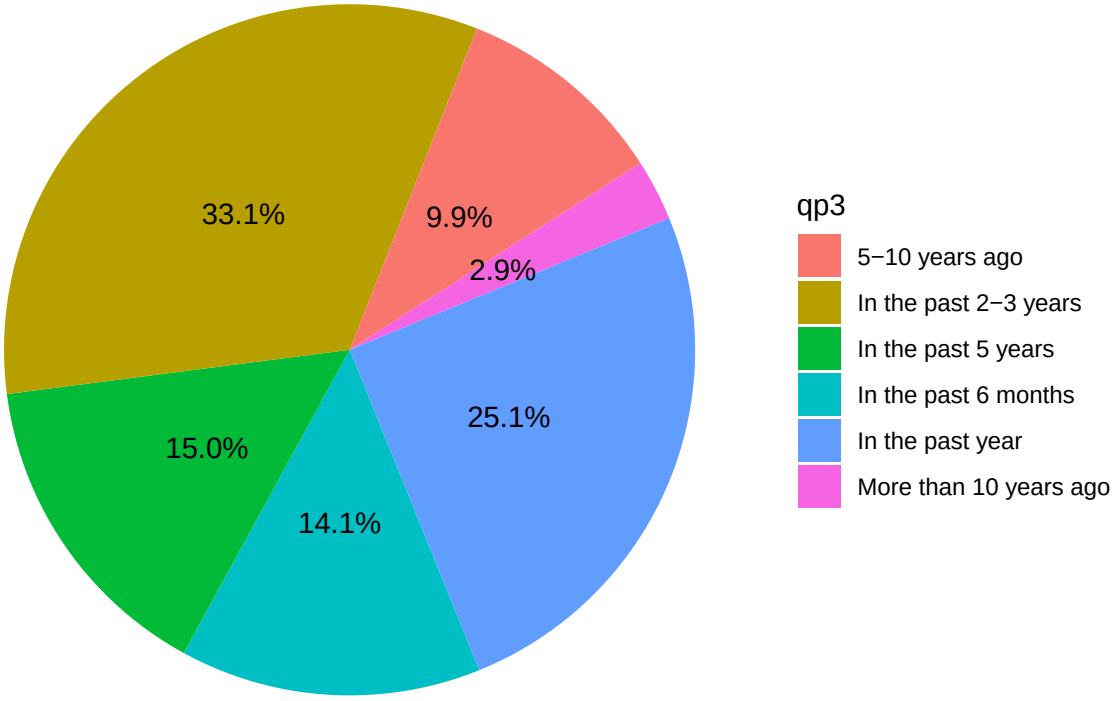
There are 7 records with this number. I am assuming these are correct numbers.



3: Visualization of first started listening to podcasts(qp3)

qp3: first started listening to podcasts

Use a pie chart to see distribution of listeners in which time to start listening to podcasts.



Part 4: Subset a dataset with selected variables and id

In the dataset, there are 584 variables. Based on my research questions, I will only focuse on below variables which will be using for th segmentation analyzing.

Selecting Variables

1:Variables of Demographics of Listeners

arfgen: gender arfagerf: age arfhhs: household size arfkid: kids in household arfeduc:highest education arfwork: work status arfhhi: household income before tax

2:Variables of listeners' Recency and Frequency

In thinking about the "what," we should think about the past, present, and future. What have listeners done, what are they doing, what are they thinking, and what are they likely to do? From the answers to these questions, we learn how our listeners interacted with podcasts in the past. We can comparing this data with the data from other W's, we will be able to indentify the least and most profitable listeners.

Attributes for Recency

qp4nma: specific podcasts listened to in past month qp4tya : type of podcasts listened to in past month qp4gnaa: genres of podcasts listened to in past month

Attributes for Frequency

qs4: frequency listening to audio podcasts qp1c: combined time listening typical week qp3: first started listening to podcasts

3: Variables of how listeners' reacting to ADS.

qa1a: tom1 brands podcst advertising qa2aa: attention paid TP podcast ADS qza3a: tried to get more info on podcast AD

Part 5: Data Analyzing - Segmentation Analyzing

Segmentation provides the knowledge that companies need to tailor their products and services to maximize their profits within each segment.

The analysis of this dataset is expecting to identify more profitable segments for ads.Podcasters can focus their efforts on keeping these listeners happy while increasing their purchases via advertising on the podcasts.

I will use k-Modes Clustering and Hierarchical clustering for the segmentation analyzing.

A: K-Modes Clustering Categorical Data: Partitioning Around Medoids (PAM) algorithm

1: Calculating Distance -Gower distance

In order for a clustering algorithm to yield sensible results, we have to use a distance metric that can handle mixed data types. In this case, we will use something called Gower distance.

Now that the distance matrix has been calculated, it is time to select an algorithm for clustering. While many algorithms that can handle a custom distance matrix exist, partitioning around medoids (PAM) will be used here.

Check attributes to ensure the correct methods are being used

```
## 1119756 dissimilarities, summarized :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.6875  0.7500  0.7358  0.8125  1.0000
## Metric :  mixed ;  Types = N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N
## Number of objects : 1497
```

Print most similar clients

```
##         id arfgen arfagerf arfhhs arfkid              arfeduc
## 74 5693756   Male 45 to 54      3    Yes Post-graduate degree
## 23 5692665   Male 45 to 54      3    Yes Post-graduate degree
##                               arfwork                  arfhhi    qp4nma
## 74 Employed / self-employed full-time (112500) $100k -< $125k N/A, None
## 23 Employed / self-employed full-time   ( 67500) $50k -< $75k N/A, None
##                qp4tya      qp4gnaa               qs4
## 74 No show type given Not recorded 2-3 times a month
## 23 No show type given Not recorded 2-3 times a month
##                                qp1c                  qp3             qa1a
## 74 DNQ Listen once per week or more In the past 2-3 years Don't Know, None
## 23 DNQ Listen once per week or more In the past 2-3 years           google
##                               qa2aa                                qza3a
## 74 (2) Pay a little less attention DNQ Looked to get more information
## 23 (2) Pay a little less attention DNQ Looked to get more information
```

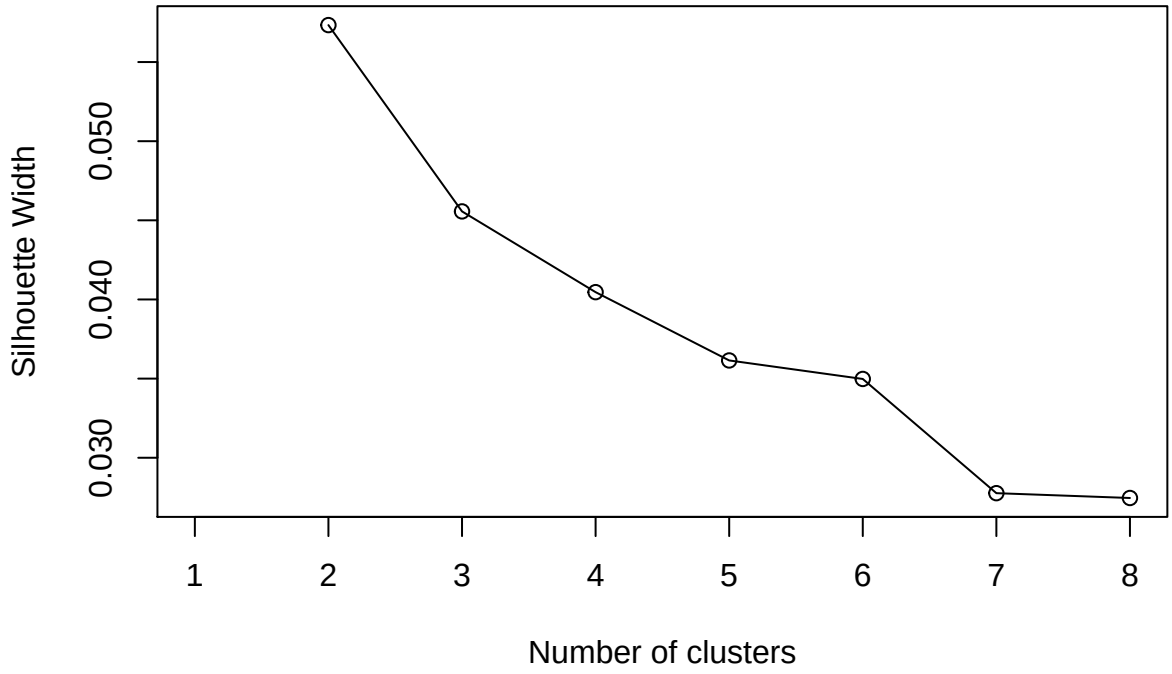
Print most dissimilar clients

```
##        id arfgen arfagerf arfhhs arfkid                          arfeduc
## 7 5692114   Male 35 to 44      5     No University undergraduate degree
## 1 5691168 Female 45 to 54      4    Yes University undergraduate degree
```

```
##                               arfwork                  arfhhi
## 7 Employed / self-employed full-time (200000) $150k -< $250k
## 1                           Homemaker  ( 87500) $75k -< $100k
##             qp4nma             qp4tya         qp4gnaa                 qs4
## 7        N/A, None No show type given    Not recorded Several days a week
## 1 Citations Needed           Episodic News & Politics           Every day
##   qp1c                   qp3        qa1a
## 7  303  In the past 6 months         N/A
## 1  840 In the past 2-3 years Squarespace
##                                 qa2aa                                qza3a
## 7 (3) Neither more nor less attention                                  Yes
## 1     (4) Pay a little more attention DNQ Looked to get more information
```

2: Determining Optimal Clusters:Silhouette coefficient

A variety of metrics exist to help choose the number of clusters to be extracted in a cluster analysis. I will use silhouette width, an internal validation metric which is an aggregated measure of how similar an observation is to its own cluster compared its closest neighboring cluster. The metric can range from -1 to 1, where higher values are better. After calculating silhouette width for clusters ranging from 2 to 8 for the PAM algorithm, I see that 3 clusters yields the highest value.



3: Applying K-Modes(PAM model) and summary of each cluster

```
## [[1]]
##     arfgen        arfagerf       arfhhs    arfkid
##  Female:186   18 to 24: 62   2      :278   No :503
##  Male  :428   25 to 34:298   1      : 91   Yes:111
##               35 to 44: 97   Dk/Na  : 73
##               45 to 54: 68   3      : 68
##               55 to 64: 47   4      : 66
##               65+     : 40   5      : 28
##               Under 18:  2   (Other): 10
##                                  arfeduc
```

```
##  University undergraduate degree     :201
##  Post-graduate degree                :107
##  Completed college / technical school: 92
##  Some university                     : 64
##  High school graduate                : 62
##  Some college / technical school     : 38
##  (Other)                             : 50
##                                arfwork
##  Employed / self-employed full-time:389
##  Employed / self-employed part-time: 74
##  Full-time student                 : 60
##  Retired                           : 47
##  Currently looking for work        : 20
##  Not working for medical reasons   : 14
##  (Other)                           : 10
##                             arfhhi                            qp4nma
##  ( 67500) $50k -< $75k        : 96   The Joe Rogan Experience: 18
##  (112500) $100k -< $125k      : 93   Freakonomics Radio      : 17
##  ( 87500) $75k -< $100k       : 92   99% Invisible           : 16
##  Don't know/ prefer not to say: 76   CANADALAND              : 15
##  ( 42500) $35k -< $50k        : 72   Adam Carolla Show       : 10
##  ( 27500) $25k -< $35k        : 53   Other/Unknown Podcast   :  9
##  (Other)                      :132   (Other)                 :529
##                 qp4tya                  qp4gnaa
##  Episodic          :387   Comedy           :150
##  No show type given:211   News & Politics  : 79
##  Serial            : 16   Society & Culture: 55
##                           Not recorded     : 38
##                           History          : 32
##                           Professional     : 29
##                           (Other)          :231
##                   qs4                                      qp1c
##  2-3 times a month  : 32   120                               : 64
##  About once a week  :140   240                               : 56
##  Every day          :154   DNQ Listen once per week or more: 55
##  Once a month       : 23   180                               : 52
##  Several days a week:265   300                               : 38
##                            150                               : 34
##                            (Other)                           :315
##                      qp3                    qa1a
##  5-10 years ago        : 82   Don't Know, None:134
##  In the past 2-3 years :228   Other           : 79
##  In the past 5 years   : 98   Casper          : 38
##  In the past 6 months  : 80   N/A             : 38
##  In the past year      :109   Squarespace     : 28
##  More than 10 years ago: 17   Audible         : 24
##                               (Other)         :273
##                                                        qa2aa
##  (1) Pay much less attention than I do to other ads    :117
##  (2) Pay a little less attention                       : 88
##  (3) Neither more nor less attention                   :207
##  (4) Pay a little more attention                       :158
##  (5) Pay a lot more attention to ads I hear on podcasts: 44
##
```

```
##
##                                 qza3a        cluster
##  DNQ Looked to get more information:341   Min.   :1
##  No                                 : 91   1st Qu.:1
##  Yes                                :182   Median :1
##                                            Mean   :1
##                                            3rd Qu.:1
##                                            Max.   :1
##
##
## [[2]]
##     arfgen        arfagerf       arfhhs    arfkid
##  Female:120   18 to 24: 33   3      :158   No :126
##  Male  :327   25 to 34:101   4      :109   Yes:321
##               35 to 44: 87   Dk/Na  : 57
##               45 to 54:147   2      : 53
##               55 to 64: 44   1      : 30
##               65+     : 35   5      : 30
##               Under 18:  0   (Other): 10
##                                  arfeduc
##  Completed college / technical school:143
##  Post-graduate degree                : 87
##  University undergraduate degree     : 63
##  High school graduate                : 50
##  Some university                     : 39
##  Some college / technical school     : 31
##  (Other)                             : 34
##                                arfwork
##  Employed / self-employed full-time:295
##  Employed / self-employed part-time: 41
##  Retired                           : 40
##  Homemaker                         : 22
##  Full-time student                 : 17
##  Currently looking for work        : 16
##  (Other)                           : 16
##                             arfhhi
##  ( 87500) $75k -< $100k       :126
##  ( 67500) $50k -< $75k        : 66
##  (112500) $100k -< $125k      : 46
##  Don't know/ prefer not to say: 40
##  (200000) $150k -< $250k      : 38
##  (137500) $125k -< $150k      : 36
##  (Other)                      : 95
##                                          qp4nma
##  N/A, None                                  : 44
##  Other/Unknown Podcast                      : 32
##  CBC (unspec.)                              : 30
##  Don't know                                 : 25
##  Radio Stations / Call Letters / Music Shows: 19
##  Platforms/Apps                             :  7
##  (Other)                                    :290
##                 qp4tya                 qp4gnaa                      qs4
##  Episodic          : 80   Not recorded     :194   2-3 times a month  :90
##  No show type given:360   News & Politics  : 36   About once a week  :99
```

```
##  Serial            :  7   Society & Culture: 30   Every day          :90
##                           Comedy           : 26   Once a month       :88
##                           Professional     : 22   Several days a week:80
##                           History          : 14
##                           (Other)          :125
##                                qp1c                          qp3
##  DNQ Listen once per week or more:178   5-10 years ago        : 42
##  120                              : 34   In the past 2-3 years : 81
##  60                               : 27   In the past 5 years   : 61
##  180                              : 19   In the past 6 months  : 66
##  300                              : 13   In the past year      :185
##  150                              : 12   More than 10 years ago: 12
##  (Other)                          :164
##                               qa1a
##  Don't Know, None                :154
##  Other                           : 59
##  N/A                             : 29
##  Apple / Apple Products / iTunes: 27
##  CBC                             : 13
##  Amazon                          :  9
##  (Other)                         :156
##                                                        qa2aa
##  (1) Pay much less attention than I do to other ads    :114
##  (2) Pay a little less attention                       : 71
##  (3) Neither more nor less attention                   :212
##  (4) Pay a little more attention                       : 39
##  (5) Pay a lot more attention to ads I hear on podcasts: 11
##
##
##                                  qza3a        cluster
##  DNQ Looked to get more information:290   Min.   :2
##  No                                : 54   1st Qu.:2
##  Yes                               :103   Median :2
##                                           Mean   :2
##                                           3rd Qu.:2
##                                           Max.   :2
##
##
## [[3]]
##     arfgen        arfagerf       arfhhs    arfkid
##  Female:307   18 to 24: 43   1      :165   No :380
##  Male  :129   25 to 34: 75   2      :155   Yes: 56
##               35 to 44: 73   4      : 49
##               45 to 54:103   3      : 29
##               55 to 64: 61   5      : 17
##               65+     : 81   Dk/Na  : 13
##               Under 18:  0   (Other):  8
##                                  arfeduc
##  University undergraduate degree     :135
##  Post-graduate degree                : 65
##  High school graduate                : 64
##  Completed college / technical school: 55
##  Some university                     : 41
##  Some college / technical school     : 29
```

```
##  (Other)                              : 47
##                                arfwork
##  Employed / self-employed full-time:197
##  Retired                            : 90
##  Employed / self-employed part-time: 63
##  Full-time student                  : 33
##  Homemaker                          : 17
##  Currently looking for work         : 16
##  (Other)                            : 20
##                             arfhhi                                qp4nma
##  ( 67500) $50k -< $75k         :127   Don't know                    : 45
##  Don't know/ prefer not to say: 72   Other/Unknown Podcast        : 34
##  ( 15000) <$25k                : 44   N/A, None                    : 31
##  ( 42500) $35k -< $50k         : 43   CBC (unspec.)                : 15
##  ( 87500) $75k -< $100k        : 34   Because News from CBC Radio :  6
##  (112500) $100k -< $125k       : 30   As It Happens from CBC Radio:  5
##  (Other)                       : 86   (Other)                      :300
##                 qp4tya                  qp4gnaa
##  Episodic          : 75   Not recorded     :156
##  No show type given:347   News & Politics  : 46
##  Serial            : 14   Comedy           : 36
##                           Society & Culture: 31
##                           History          : 15
##                           Professional     : 13
##                           (Other)          :139
##                   qs4                                      qp1c
##  2-3 times a month  :158   DNQ Listen once per week or more:201
##  About once a week  :105   120                              : 32
##  Every day          : 62   60                               : 31
##  Once a month       : 43   180                              : 24
##  Several days a week: 68   90                               : 15
##                            420                              : 10
##                            (Other)                          :123
##                      qp3                                    qa1a
##  5-10 years ago        : 26   Don't Know, None                :195
##  In the past 2-3 years :183   Other                           : 55
##  In the past 5 years   : 70   N/A                             : 42
##  In the past 6 months  : 60   Apple / Apple Products / iTunes: 15
##  In the past year      : 82   CBC                             : 14
##  More than 10 years ago: 15   Squarespace                     :  9
##                               (Other)                         :106
##                                                    qa2aa
##  (1) Pay much less attention than I do to other ads    :119
##  (2) Pay a little less attention                       : 43
##  (3) Neither more nor less attention                   :235
##  (4) Pay a little more attention                       : 27
##  (5) Pay a lot more attention to ads I hear on podcasts: 12
##
##
##                                qza3a        cluster
##  DNQ Looked to get more information:307   Min.   :3
##  No                                : 48   1st Qu.:3
##  Yes                               : 81   Median :3
##                                           Mean   :3
```

```
##                                          3rd Qu.:3
##                                          Max.   :3
##
```

4: Example of summary of each cluster

Here one can attempt to derive some common patterns for clients within a cluster.

As an example:

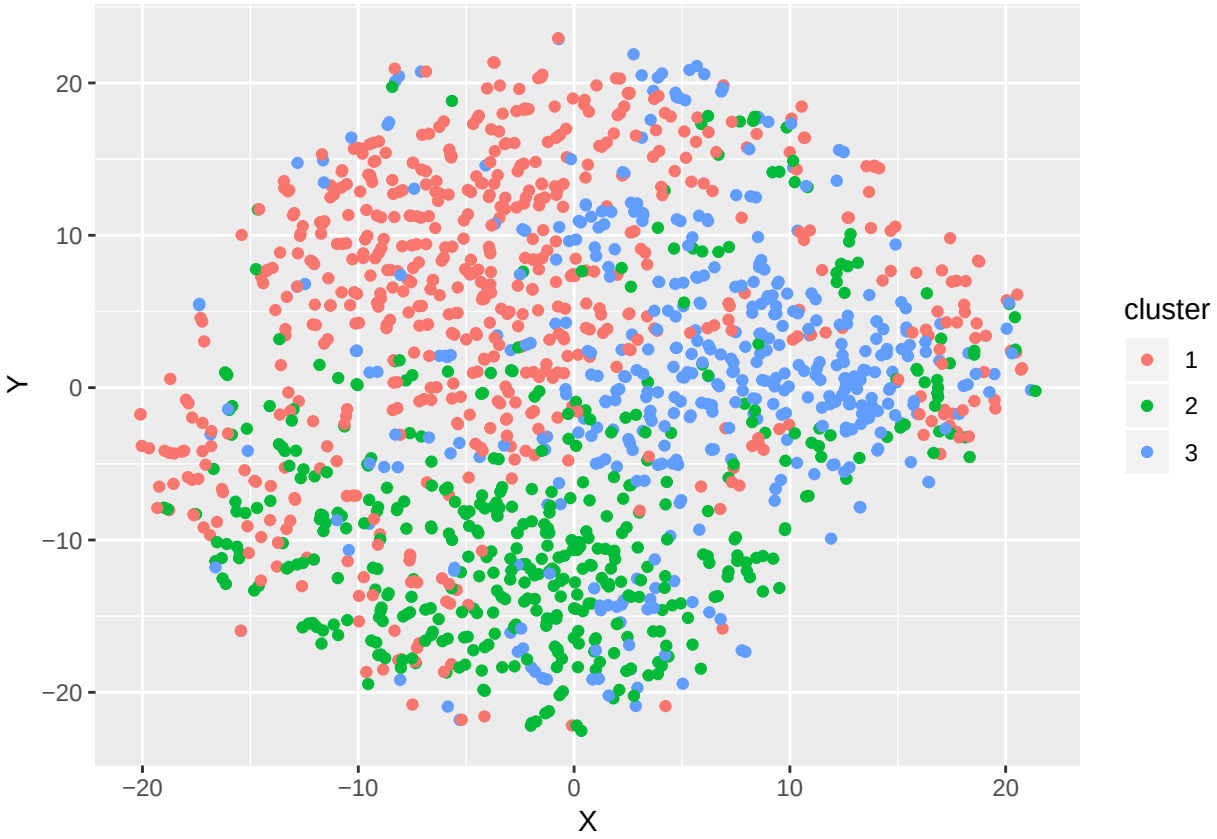
1: Cluster 1 is made of Male x age 25 to 34 x household size 2 x no kids x university udergraduate degree x employed/self-employed full-time x house income $50k -< $100k.

2: Cluster 2 is made of Male x age 45 to 54 and age 25 to 34 x household size 3 and 4 x yes kids x Completed college / technical school x Employed / self-employed full-time x house income $75k -< $100k and other.

3: Cluster 3 is made of Female x age 45 to 54 and 65+ x household size 1 and 2 x no kids x University undergraduate degree x Employed / self-employed full-time and retired x house income 50 -< 75k and (Other).

5: Visualization

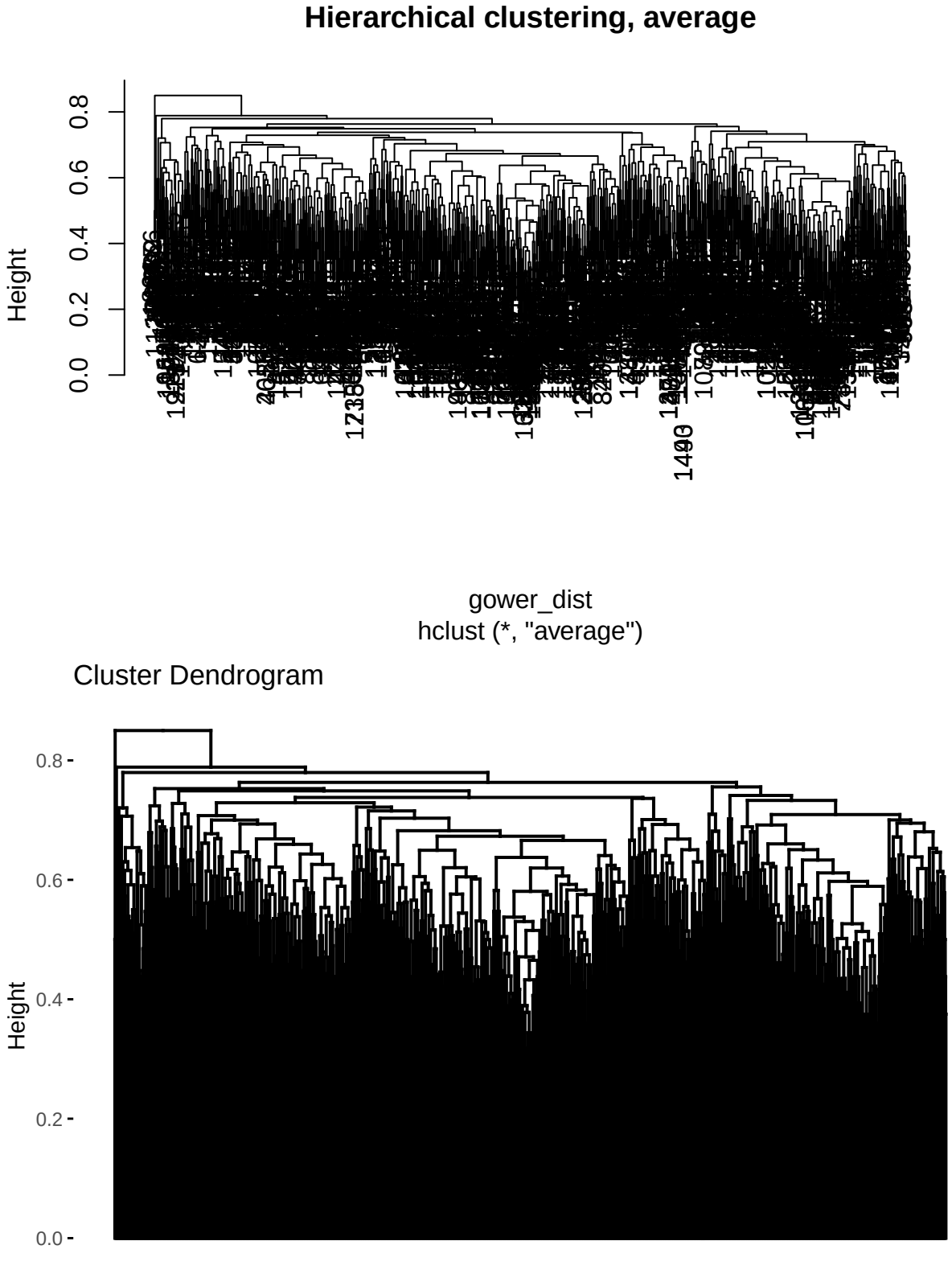
One way to visualize many variables in a lower dimensional space is with t-distributed stochastic neighborhood embedding, or t-SNE. This method is a dimension reduction technique that tries to preserve local structure so as to make clusters visible in a 2D or 3D visualization. While it typically utilizes Euclidean distance, it has the ability to handle a custom distance metric like the one we created above. In this case, the plot shows the three well-separated clusters that PAM was able to detect.



Although not perfect (especially cluster 1), colors are mostly located in similar areas, confirming the relevancy of the segmentation.

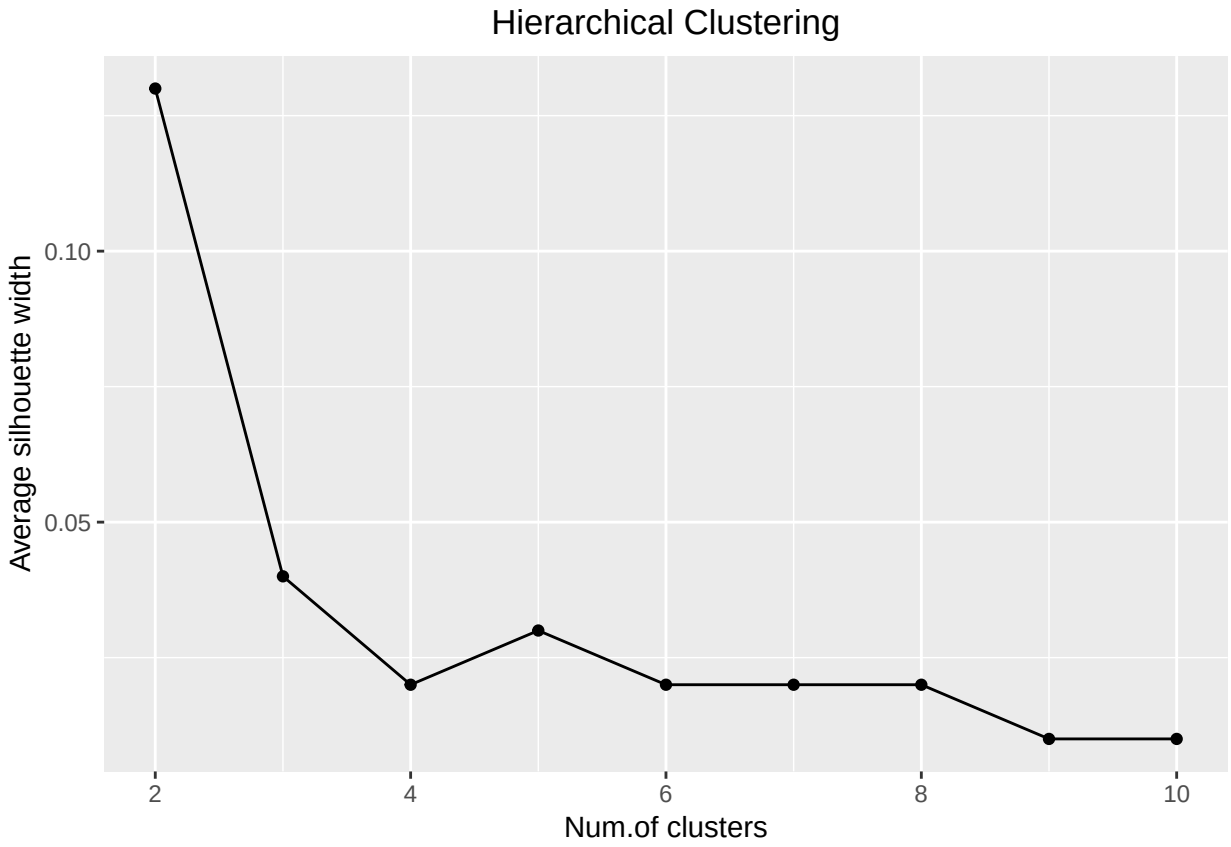
B:Hierarchical Clustering

1: Plot of Hierarchical Clustering



2: Determining Optimal Clusters: Silhouette method

I will use same method as K-Modes for this optimal clusters. I see that 3 clusters yields the highest value. Hence, select k = 3



I see k=3 is an appropriate number of clusters. Let's try k=3

3: Applying hierarchical clustering and summary of each cluster

```
## [[1]]
##     arfgen         arfagerf       arfhhs    arfkid
##  Female:609   18 to 24:137   2      :484   No :1006
##  Male  :883   25 to 34:474   1      :286   Yes: 486
##               35 to 44:257   3      :255
##               45 to 54:315   4      :223
##               55 to 64:152   Dk/Na  :142
##               65+     :156   5      : 75
##               Under 18:  1   (Other): 27
##                                   arfeduc
##  University undergraduate degree     :398
##  Completed college / technical school:290
##  Post-graduate degree                :257
##  High school graduate                :176
##  Some university                     :143
##  Some college / technical school     : 98
##  (Other)                             :130
##                                arfwork
##  Employed / self-employed full-time:880
##  Retired                           :177
##  Employed / self-employed part-time:175
```

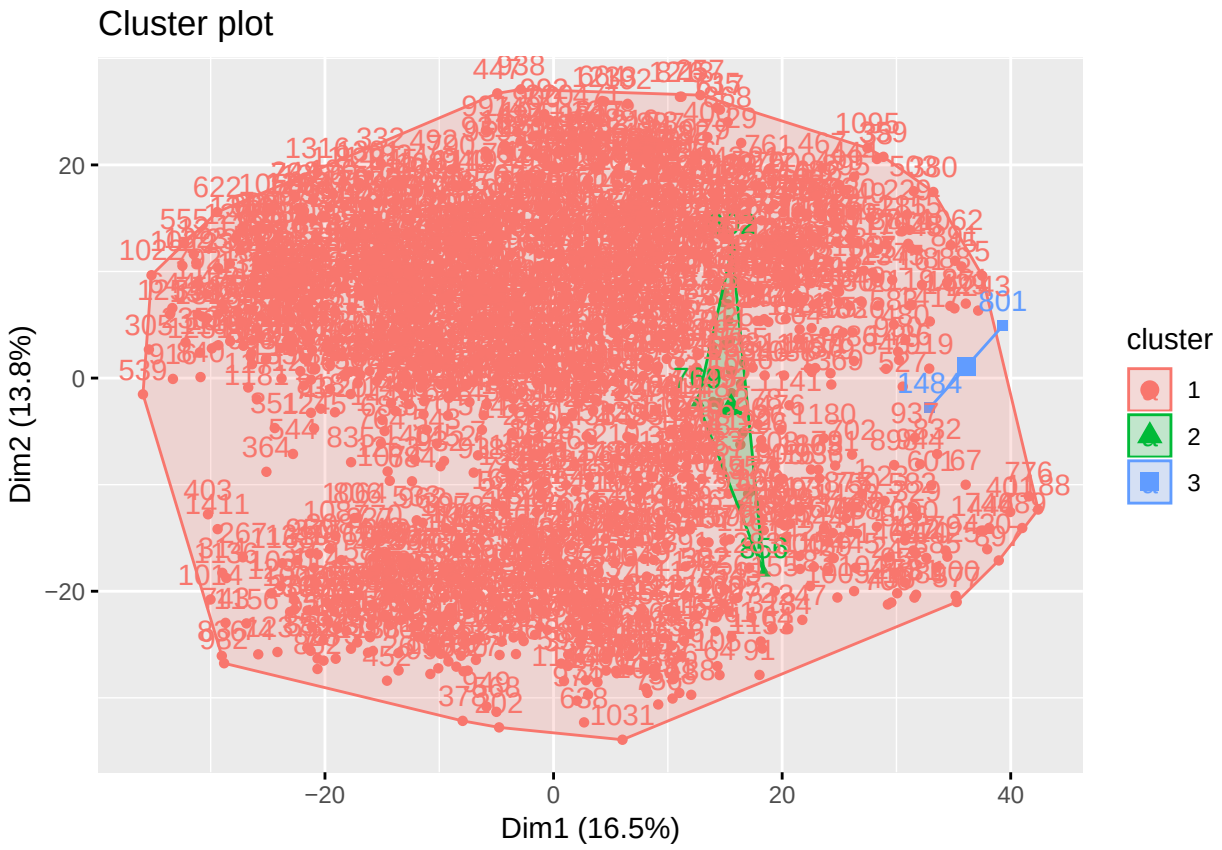
```
##  Full-time student                 :110
##  Currently looking for work        : 51
##  Homemaker                         : 48
##  (Other)                           : 51
##                            arfhhi
##  ( 67500) $50k -< $75k        :288
##  ( 87500) $75k -< $100k       :252
##  Don't know/ prefer not to say:185
##  (112500) $100k -< $125k      :169
##  ( 42500) $35k -< $50k        :148
##  ( 27500) $25k -< $35k        :115
##  (Other)                      :335
##                                        qp4nma
##  N/A, None                                :  83
##  Other/Unknown Podcast                    :  75
##  Don't know                               :  70
##  CBC (unspec.)                            :  51
##  Radio Stations / Call Letters / Music Shows:  23
##  The Joe Rogan Experience                 :  22
##  (Other)                                  :1168
##                qp4tya                 qp4gnaa
##  Episodic          :542   Not recorded     :388
##  No show type given:918   Comedy           :212
##  Serial            : 32   News & Politics  :160
##                           Society & Culture:116
##                           Professional     : 64
##                           History          : 61
##                           (Other)          :491
##                  qs4                                      qp1c
##  2-3 times a month  :280   DNQ Listen once per week or more:434
##  About once a week  :340   120                                :130
##  Every day          :305   180                                : 94
##  Once a month       :154   60                                 : 88
##  Several days a week:413   240                                : 75
##                            300                                : 59
##                            (Other)                            :612
##                    qp3                                      qa1a
##  5-10 years ago        :149   Don't Know, None                 :483
##  In the past 2-3 years :490   Other                            :192
##  In the past 5 years   :227   N/A                              :109
##  In the past 6 months  :206   Apple / Apple Products / iTunes: 64
##  In the past year      :376   Casper                           : 47
##  More than 10 years ago: 44   CBC                              : 39
##                               (Other)                          :558
##                                                      qa2aa
##  (1) Pay much less attention than I do to other ads    :349
##  (2) Pay a little less attention                       :202
##  (3) Neither more nor less attention                   :653
##  (4) Pay a little more attention                       :223
##  (5) Pay a lot more attention to ads I hear on podcasts: 65
##
##
##                                qza3a        cluster
##  DNQ Looked to get more information:935   Min.   :1
```

```
##  No                                :193   1st Qu.:1
##  Yes                               :364   Median :1
##                                           Mean   :1
##                                           3rd Qu.:1
##                                           Max.   :1
##
##
## [[2]]
##     arfgen      arfagerf     arfhhs  arfkid
##  Female:3   18 to 24:0   2      :2   No :1
##  Male  :0   25 to 34:0   4      :1   Yes:2
##             35 to 44:0   1      :0
##             45 to 54:3   10     :0
##             55 to 64:0   3      :0
##             65+     :0   5      :0
##             Under 18:0   (Other):0
##                                  arfeduc
##  Post-graduate degree                :2
##  University undergraduate degree     :1
##  Completed college / technical school:0
##  Elementary/grade school             :0
##  High school graduate                :0
##  Some college / technical school     :0
##  (Other)                             :0
##                                arfwork                              arfhhi
##  Currently looking for work        :1   Don't know/ prefer not to say:2
##  Employed / self-employed full-time:1   (137500) $125k -< $150k      :1
##  Employed / self-employed part-time:1   ( 15000) <$25k               :0
##  Full-time student                 :0   ( 27500) $25k -< $35k        :0
##  Homemaker                         :0   ( 42500) $35k -< $50k        :0
##  Not working for medical reasons   :0   ( 67500) $50k -< $75k        :0
##  (Other)                           :0   (Other)                      :0
##                                                    qp4nma
##  30 For 30 Podcasts                                   :2
##  Caliphate                                            :1
##  ...These Are Their Stories: The Law & Order Podcast:0
##  .NET Rocks!                                          :0
##  @SkatingPj Podcast                                   :0
##  /Film Daily                                          :0
##  (Other)                                              :0
##                 qp4tya                 qp4gnaa                    qs4
##  Episodic          :0   Sports & Recreation:2   2-3 times a month  :0
##  No show type given:0   News & Politics    :1   About once a week  :2
##  Serial            :3   Alternative Health :0   Every day          :1
##                         Amateur            :0   Once a month       :0
##                         Arts               :0   Several days a week:0
##                         Automotive         :0
##                         (Other)            :0
##       qp1c                        qp3
##  110    :1   5-10 years ago        :1
##  150    :1   In the past 2-3 years :2
##  300    :1   In the past 5 years   :0
##  1      :0   In the past 6 months  :0
##  10     :0   In the past year      :0
```

```
##  1080   :0   More than 10 years ago:0
##  (Other):0
##                                qa1a
##  Apple / Apple Products / iTunes:1
##  Blue Apron                     :1
##  Mail Chimp                     :1
##  .ca                            :0
##  ABC                            :0
##  Air Canada                     :0
##  (Other)                        :0
##                                                        qa2aa
##  (1) Pay much less attention than I do to other ads    :1
##  (2) Pay a little less attention                       :0
##  (3) Neither more nor less attention                   :1
##  (4) Pay a little more attention                       :1
##  (5) Pay a lot more attention to ads I hear on podcasts:0
##
##
##                                  qza3a      cluster
##  DNQ Looked to get more information:3   Min.   :2
##  No                                :0   1st Qu.:2
##  Yes                               :0   Median :2
##                                         Mean   :2
##                                         3rd Qu.:2
##                                         Max.   :2
##
##
## [[3]]
##     arfgen      arfagerf     arfhhs  arfkid
##  Female:1   18 to 24:1   8      :1   No :2
##  Male  :1   25 to 34:0   Dk/Na  :1   Yes:0
##             35 to 44:0   1      :0
##             45 to 54:0   10     :0
##             55 to 64:0   2      :0
##             65+     :0   3      :0
##             Under 18:1   (Other):0
##                                   arfeduc
##  Elementary/grade school              :1
##  Some university                      :1
##  Completed college / technical school:0
##  High school graduate                 :0
##  Post-graduate degree                 :0
##  Some college / technical school      :0
##  (Other)                              :0
##                                 arfwork                              arfhhi
##  Employed / self-employed part-time:2   ( 67500) $50k -< $75k        :1
##  Currently looking for work        :0   Don't know/ prefer not to say:1
##  Employed / self-employed full-time:0   ( 15000) <$25k               :0
##  Full-time student                 :0   ( 27500) $25k -< $35k        :0
##  Homemaker                         :0   ( 42500) $35k -< $50k        :0
##  Not working for medical reasons   :0   ( 87500) $75k -< $100k       :0
##  (Other)                           :0   (Other)                      :0
##                                                    qp4nma
##  Alice Isn't Dead                                     :1
```

```
##  RABBITS                                                :1
##  ...These Are Their Stories: The Law & Order Podcast:0
##  .NET Rocks!                                            :0
##  @SkatingPj Podcast                                     :0
##  /Film Daily                                            :0
##  (Other)                                                :0
##                 qp4tya                qp4gnaa                    qs4
##  Episodic          :0   Performing Arts   :2   2-3 times a month  :0
##  No show type given:0   Alternative Health:0   About once a week  :2
##  Serial            :2   Amateur           :0   Every day          :0
##                         Arts              :0   Once a month       :0
##                         Automotive        :0   Several days a week:0
##                         Aviation          :0
##                         (Other)           :0
##       qp1c                         qp3             qa1a
##  180    :1   5-10 years ago        :0   Other      :1
##  555    :1   In the past 2-3 years :0   Squarespace:1
##  1      :0   In the past 5 years   :2   .ca        :0
##  10     :0   In the past 6 months  :0   ABC        :0
##  1080   :0   In the past year      :0   Air Canada :0
##  110    :0   More than 10 years ago:0   Allstate   :0
##  (Other):0                              (Other)    :0
##                                                     qa2aa
##  (1) Pay much less attention than I do to other ads    :0
##  (2) Pay a little less attention                       :0
##  (3) Neither more nor less attention                   :0
##  (4) Pay a little more attention                       :0
##  (5) Pay a lot more attention to ads I hear on podcasts:2
##
##
##                                  qza3a      cluster
##  DNQ Looked to get more information:0   Min.   :3
##  No                                :0   1st Qu.:3
##  Yes                               :2   Median :3
##                                         Mean   :3
##                                         3rd Qu.:3
##                                         Max.   :3
##
```

4 Visualize the result in a scatter plot



Clustering validation

In this section, we'll use cluster.stats() [in fpc package] for comparison between two clusterings.

The cluster.stats() computing a number of distance based statistics which can be used either for cluster validation, comparison between clustering and decision about the number of clusters.

1: Cluster statistics for PAM clustering

pam within clusters sum of squares

```
## [1] 23372.53
```

pam cluster average silhouette widths

```
##           1           2           3
## 0.06313989 0.05450466 0.05104908
```

pam average distance within clusters

```
## [1] 5.516989
```

Number of points in each cluster

```
## [1] 614 447 436
```

Cluster statistics for hierarchical clustering

(HCLUST) within clusters sum of squares

```
## [1] 26501.77
```

(HCLUST) cluster average silhouette widths

```
##           1           2           3
## 0.003739411 0.245207077 0.381521770
```

(HCLUST) average distance within clusters

```
## [1] 5.873582
```

number of points in each cluste

```
## [1] 1492    3    2
```

Comparison of two clusterings

I use some values from running cluster.stats() for the two clusterings comparison.

Comparison of cluster average silhouette widths

K-Modes (PAM) Clustering	Hierarchical Clustering				
1	2	3	1	2	3
0.06313989	0.05450466	0.05104908	0.003739411	0.245207077	0.381521770

- For K-Modes, average silhouette for each cluster is greater than 0.05.
- For HCLUST, average silhouette for cluster 1 is very low just about 0.0037, but from summary of cluster, we saw majority of oberservations(1492) are in cluster 1. This means that many points have a very low value, then the clustering configuration may have too many or too few clusters.

Comparison of average distance within clusters

K-Modes (PAM) Clustering	Hierarchical Clustering
5.513666	5.879817

Hierarchical clustering has higher number than K-Modes. That means hierarchical clustering is not good as K-Modes clustering.

Conclusion of Comparison

After comparison, I decided to use K-Modes(PAM) clustering. Due to many categorical variables in the data, the clustering result are not as good as we expect it, but it did give us useful values for the podcast industry. We can combine the clustering analyzing results with some other further analyzing, it will definitely help the company to project their advertisments to the target listeners.

PAM has the following caracteristics:

- Pros: it's intuitive, more robust to noise and outliers compared to k-means (due to the properties of distances being used), and it produces a "typical individual" for each cluster (useful for interpretation).
- Cons: it's time consuming and computer intensive (run time and memory are quadratic).