2018 Podcast Listeners Analysis Final Report

Aimin Amy Hu

2019-08-21

1 Abstract

The term of podcasting comes from a combination of the words iPod(a personal digital audio player made by Apple) and broadcasting. Podcasting is a free service that allows internet users to listen to audio files on their computers or personal digital audio players. The podcasting platforms make many people's dreams come true-have their own radio show or writing songs heard by many people. Podcasting is not only used for entainment but also used for informational and educational purposes such as self-guided walking tours, training etc.

We understood that podcasting doesn't have long history. It was developed in 2004. As the personal digital devices are become more popular and enter into almost everyone's life including babies. within 100 million+ podcast listeners in the U.S (according to https://www.adamenfroy.com/podcast-hosting) and between 7 and 10 million Canadians are listening to podcasts(according to a study from late 2017 conducted by Ulster Media and the Globe and Mail).

Problem Statment

A rapdily increasing of updating personal digital devices and internet speed, they are many new things happen every day and also people share their personal experiences on social media. Podcast as a product which belongs to information technology era doesn't have long history, but it definitely affect our life. Podcasting has experienced a huge surge of interest in past several year (especially in the past three to four years). With Google reporting up to 10 million monthly searches for "podcast" via the search engine giant in 2019 based on information from Listennotes website. There are so much potentials in this field but also competition heats up for different podcast platforms.

Per Mr.Jeff Vidler from Audience Insights Inc mentioned in the project meeting, Podcasts reply on ads for their finance support but they have challenge to target the podcast listeners. The one major reason is podcast platforms can see listeners download the podcast episodes but are unable to get much information about the listeners as almost all podcasts are completely free and listeners don't need to put their information such as credit card or names.

Research Questions

This project analyze the podcasing listeners in Canada. In Canada, this is just 3rd year to analysis the podcasting data. I used The Canadian Podcast Listener 2018 Survey Data to focuse on below research questions but will not limite to these:

- To what extent is income affecting people listening podcast?
- To what extent are gender and age affecting people listening podcast?
- To what extent are devices affecting listeners of podcasts?
- What are the popular 10 podcasts listed from the survery data?
- How listeners react to the advertising in podcasts?

About

Audience Insights Inc. is a media research consultancy pulling together teams of media, research and data professionals to meet just about any assignment facing our clients. Whether it's building audience, driving ad revenue, understanding impact of marketing or activating digital opportunities, we're flexible and fluid to fill gaps left in large, consolidated firms.

Objectives

- To help Canadian podcast publishers and advertisers understand more about how to target different segments of podcast listeners.
- To inform business opportunities in this growing media sector.

Method and Language

- Method: Market Segmentation.
- R will be the language for this project.

Analysis

Part 1: Data Understanding

1.1 Dataset

The original 2018 Canadian Podcast Listener data was txt files and SPSS file (.sav) from Audience Insights Inc. Thanks to our teacher Mr. Matthew Tenney to help us and convert to the .sav file to CSV file using PSPP.

R is the language for reading and analyzing the data.

1.2 Data Summary

The number of observations are 1534, the number of variables are 584.There are no duplicated rows in this dataset.

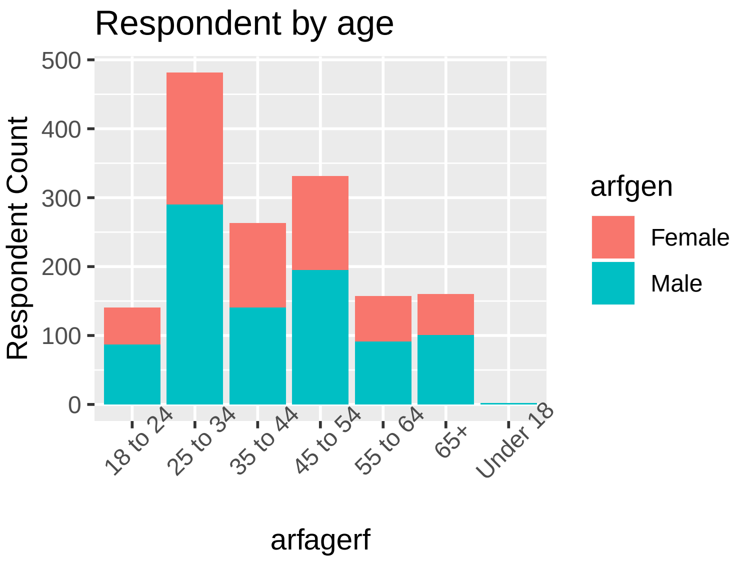
There are 184 variables (columns) in the dataset have NA for all values. This tells us that there is no any responses to these survey questions. As a company to rely on survey to get data, lack of response will definitely affect the data quality. Hence, the company should review these survey questions, make changes or remove these questions for the future survey.

Part 2 Data Visualization

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

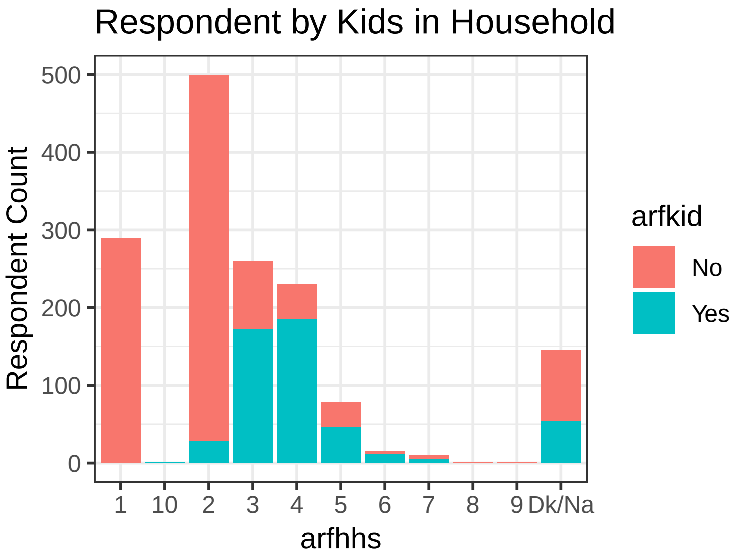
2.1 Visualization of respondent's gender and age distribution

arfgen: gender arfagerf: age



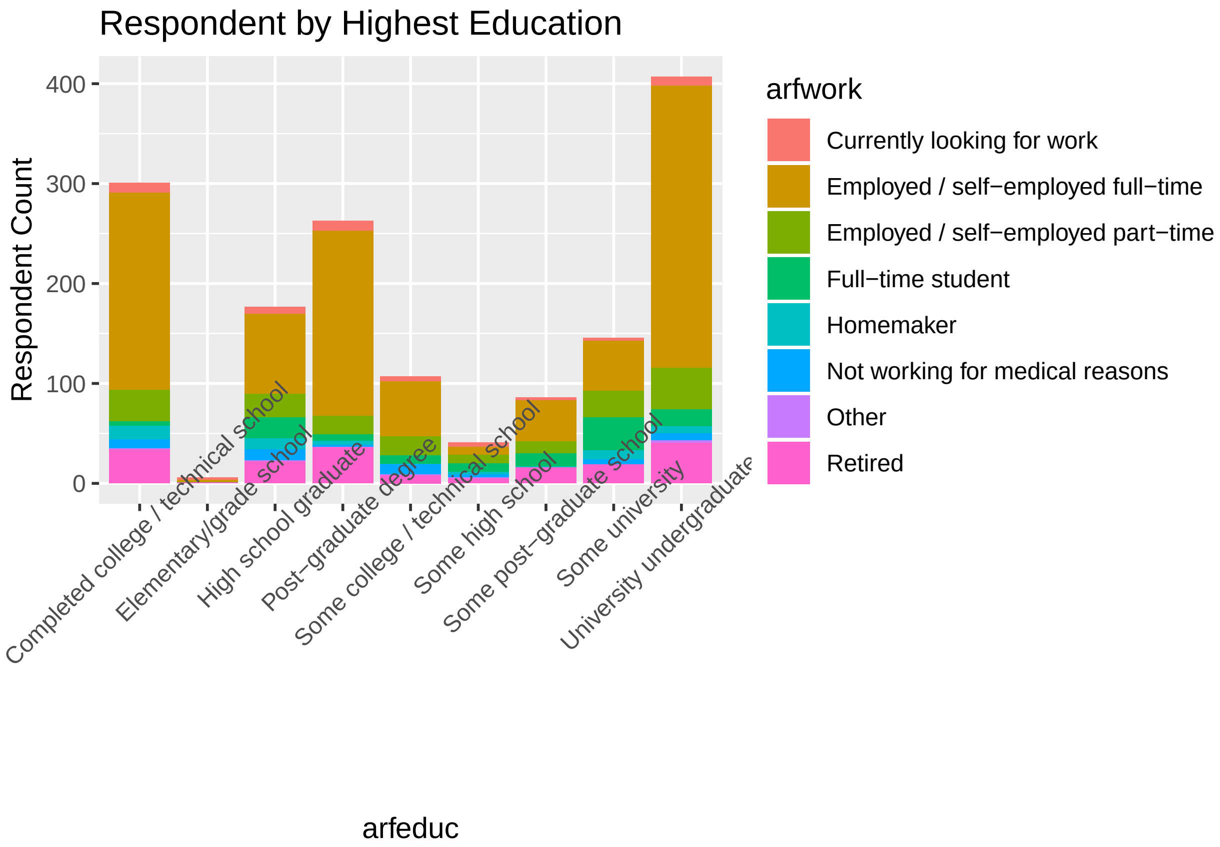
2.2 Visualzation of Combine Household size and Kids in Household for Analyzing

arfhhs: household size arfkid: kids in household



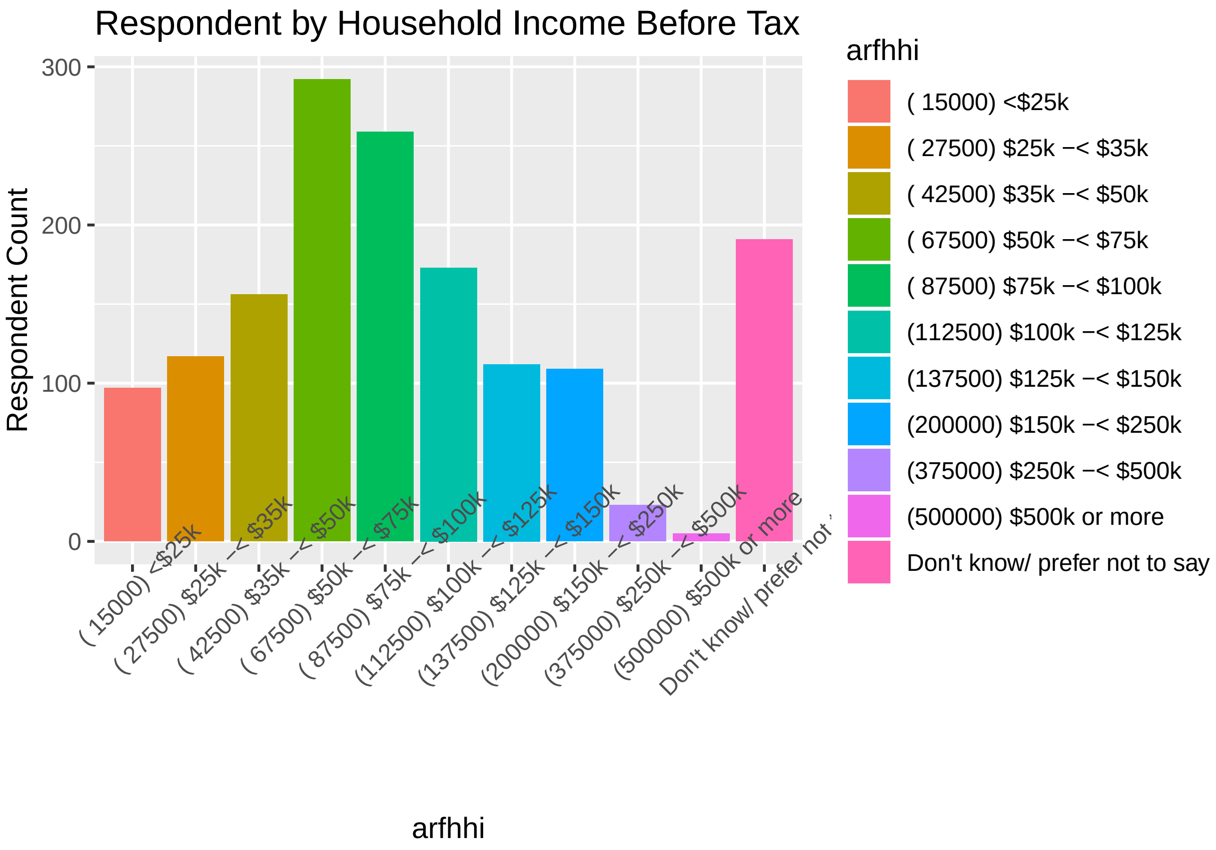
2.3 Visualzation of respondents' education level and Work Status

arfeduc: highest education arfwork: work status



2.4 Visualzation of respondents' Household Income Before Tax

arfhhi: household income before tax



Part 3: Top 10 List of Podcasts

3.1 Top 10 Specific podcasts listened to in past month

qp4nma: specific podcasts listened to in past month

By identify the top 10 specific podcasts listened to in past month, we can understand listeners' tasts of podcasts. It will help companies to promot their ads on podcasts.

	qp4nma	podcast_count
1	Other/Unknown Podcast	105
2	CBC (unspec.)	51
3	Radio Stations / Call Letters / Music Shows	23
4	The Joe Rogan Experience	22
5	99% Invisible	21
6	Freakonomics Radio	21
7	CANADALAND	18
8	As It Happens from CBC Radio	14
9	Because News from CBC Radio	12
10	Platforms/Apps	12

3.2 Top 10 genres of podcasts listened to in past month between male and female

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

By identify top 10 genres of podcasts listened to in past month between male and female will help companies to understand listeners' category needs and desires.

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

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

3.3 Top 10 favorite podcasts among all listeners

qp5cnma: faviorite podcast enjoyed to listen the most in past 6 month

```
## [1] 0
```

	qp5cnma	faviorite_count
1	Other/Unknown Podcast	119
2	The Joe Rogan Experience	43
3	CBC (unspec.)	31
4	Radio Stations / Call Letters / Music Shows	29
5	This American Life	19
6	Under the Influence from CBC Radio	17
7	Stuff You Should Know	14
8	TED Talks (unspec.)	13
9	Freakonomics Radio	12
10	La soirée est (encore) jeune	11
11	Platforms/Apps	11
12	WTF with Marc Maron Podcast	11

Part 4: Segmentation Analytics

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

4.1 Subset a dataset with selected variables and id

Variables are selected based on below criteria for the segmentation analyzing.

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

Variables of listeners' Recency and Frequency

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

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

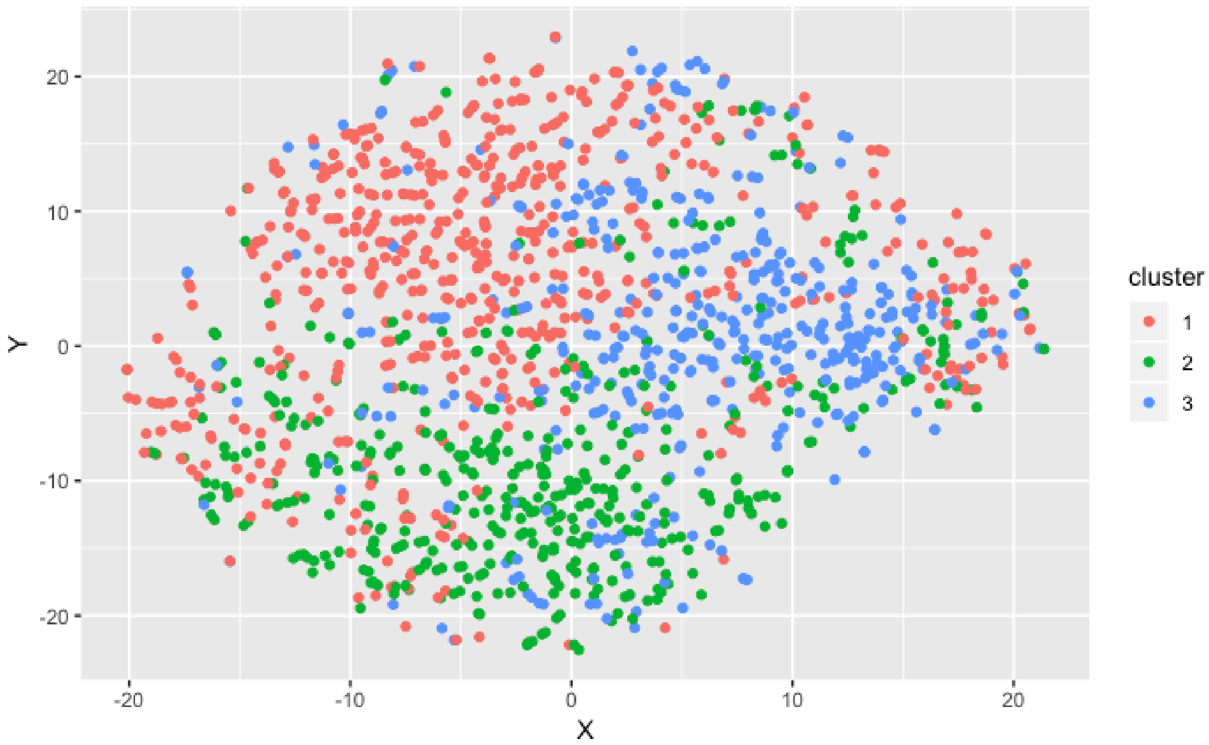
4.2 K-Modes Clustering

Clustering analysis is the unsupervised process of partitioning a group of data objects into clusters, with the objective to grouping objects of high similarity into the same cluster, while separating dissimilar objects into different clusters.

The k-modes algorithm extends the k-means paradigm to cluster categorical data by using: (1) a simple matching dissimilarity measure for categorical objects; (2) modes instead of means for clusters; (3) a frequency-based method to update modes in the k-means fashion to minimize the cost function of clustering. The k-modes algorithm is widely used in real world applications due to its efficiency in dealing with large categorical database.

4.2.1 K-Modes Clustering Visualization

In sprint 2, we have used Silhouette coefficient to determin optimal clusters. K = 3 yields the highest value, hence, the optimal clusters is 3.



4.2.2 K-Modes clustering summary

I summary the cluster results based on Demographics Segmentation, Behavior Segmentation,Occasion Segmentation, Channel Segmentation

Demographics Segmentation

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

Cluster 1 is made of as below:

- 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.

Cluster 2 is made of as below:

- 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.

Cluster 3 is made of as below:

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).

Behavior Segmentation

- 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

Cluster 1 is made of as below:

- qp4nma (specific podcasts listened to in past month): Other
- qp4tya (type of podcasts listened to in past month): Episodic and No show type given
- qp4gnaa (genres of podcasts listened to in past month): Comedy and Other

Cluster 2 is made of as below:

- qp4nma (specific podcasts listened to in past month): Other
- qp4tya (type of podcasts listened to in past month): No show type given
- qp4gnaa (genres of podcasts listened to in past month): Not recorded and Other

Cluster 3 is made of as below:

- qp4nma (specific podcasts listened to in past month): Other
- qp4tya (type of podcasts listened to in past month): No show type given
- qp4gnaa (genres of podcasts listened to in past month): Not recorded and Other

Occasion Segmentation

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

Cluster 1 is made of as below:

- qs4 (frequency listening to audio podcasts): Several days a week and Every day
- qp1c (combined time listening typical week): Other
- qp3 (first started listening to podcasts): In the past 2-3 years and In the past year

Cluster 2 is made of as below:

- qs4 (frequency listening to audio podcasts): Alomst Evenly every frequency category
- qp1c (combined time listening typical week): DNQ Listen once per week or more and Other
- qp3 (first started listening to podcasts): In the past year

Cluster 3 is made of as below:

- qs4 (frequency listening to audio podcasts): 2-3 times a month and about once a week
- qp1c (combined time listening typical week): DNQ Listen once per week or more and Other
- qp3 (first started listening to podcasts): In the past 2-3 years

Channel Segmentation

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

Cluster 1 is made of as below:

- qa1a (tom1 brands podcst advertising): Other and Don't Know, None
- qa2aa (attention paid TP podcast ADS): Neither more nor less attention, Pay a little more attention and Pay much less attention than I do to other ads.
- qza3a (tried to get more info on podcast AD): DNQ Looked to get more information and Yes.

Cluster 2 is made of as below:

- qa1a (tom1 brands podcst advertising): Don't Know, None and Other
- qa2aa (attention paid TP podcast ADS): Neither more nor less attention and Pay much less attention than I do to other ads.
- qza3a (tried to get more info on podcast AD): DNQ Looked to get more information and Yes.

Cluster 3 is made of as below:

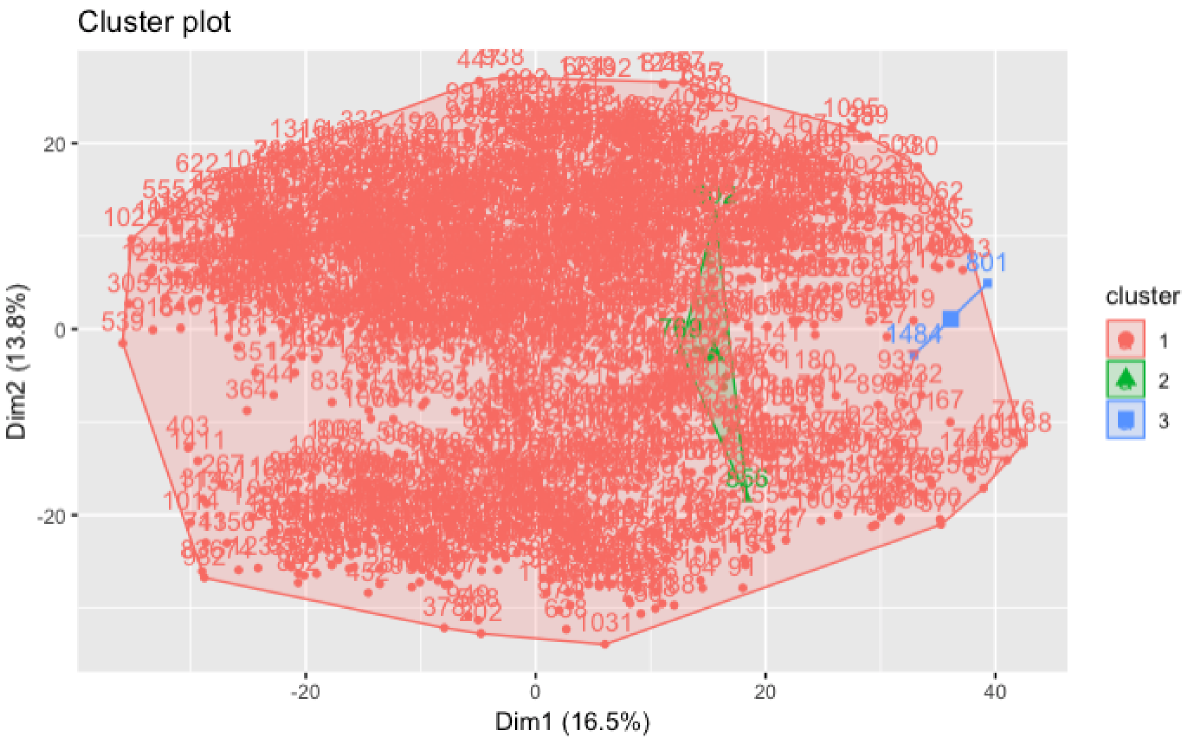
- qa1a (tom1 brands podcst advertising): Other and Don't Know, None
- qa2aa (attention paid TP podcast ADS): Neither more nor less attention and Pay much less attention than I do to other ads.
- qza3a (tried to get more info on podcast AD): DNQ Looked to get more information

4.3 Hierarchical Clustering

In order to select a fitted model for this data, Hierarchical Clustering is another approach. keep cluster k = 3 which is same as K-Modes.

4.3.1 Hierarchical Clustering Cluster Plot

The HCLUST cluster plot showed there are only a few observations in cluster 2 and 3. This might not be a fitted model for this dataset.



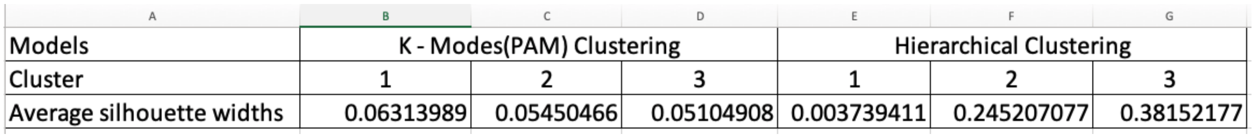
4.4 Clustering validation

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

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

4.4.1 Comparison of cluster average silhouette widths

- K-Modes, average silhouette for each cluster is greater than 0.05.
- HCLUST, average silhouette for cluster 1 is very low just about 0.0037, but from summary of cluster, we saw majority of observations(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.



Models	K - Modes(PAM) Clustering			Hierarchical Clustering		
Cluster	1	2	3	1	2	3
Average silhouette widths	0.06313989	0.05450466	0.05104908	0.003739411	0.245207077	0.38152177

4.4.2 Comparison of average distance within clusters

- K- Modes Clustering 5.516989
- Hierarchical Clustering 5.873582
- Hierarchical clustering has higher number than K-Modes. That means hierarchical clustering is not good as K-Modes clustering

4.5 Conclusion

K-Modes(PAM) clustering is selected for the segmentation analytics. 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.

4.6 Discussion

The conventional k-modes algorithm is efficient and effective in clustering large categorical data. However,the pam cluster average silhouette widths is not high in our case, it is only just above 0.5 for each cluster. There is several possible research dirctions may be worked on in the future to enhance the results. Convert all categorical data into numerical data according to dictionary of variables might be an option, but it will take more time for the coverting.

References

- 1: K-modes Clustering Algorithm for Categorical Data https://pdfs.semanticscholar.org/1069/2c9b80be922903526682f8fae5ad6ffb68f6.pdf
- 2: Using K-modes for clustering categorical data https://analyticsdefined.com/using-k-modes-clustering-categorical-data/
- 3:Clustering on mixed type data https://towardsdatascience.com/clustering-on-mixed-type-data-8bbd0a2569c3
- 4:About segments https://support.google.com/analytics/answer/3123951?hl=en
- 5:How we do Customer segmentation — Customer science & Analytics https://towardsdatascience.com/how-we-do-customer-segmentation-customer-science-analytics-e237f3db32bb