Podcast Listeners Client Base Demographic Segmentation Using Advanced Analytics

Omar Hassan

August 2019

Contents

1 Introduction

Podcast space is a growing media space. Publishers and advertising agencies have been having a challenge in identifying different segments of listeners within this space. This might be due to the fact the industry is new and data around this industry is scattered and challenging to put together to make a picture of its audience.

Here we present a research study on podcast listeners data using advanced analytical methods to identify listeners segments. Netflix would be a great example; they are able to segment their user base in a way in order to be able to recommend existing content to them or identify what kind of new content to invest in and produce. Our goal is to attempt to provide something similar to Netflix. We will provide meaningful insights to help publishers and advertising agencies reach their target audience.

2 Background

K-Means is one of the most popular "clustering" algorithms. K-means stores k centroids that it uses to define clusters. A point is considered to be in a particular cluster if it is closer to that cluster's centroid than any other centroid.

K-Means finds the best centroids by alternating between (1) assigning data points to clusters based on the current centroids (2) choosing centroids (points which are the center of a cluster) based on the current assignment of data points to clusters. [1]

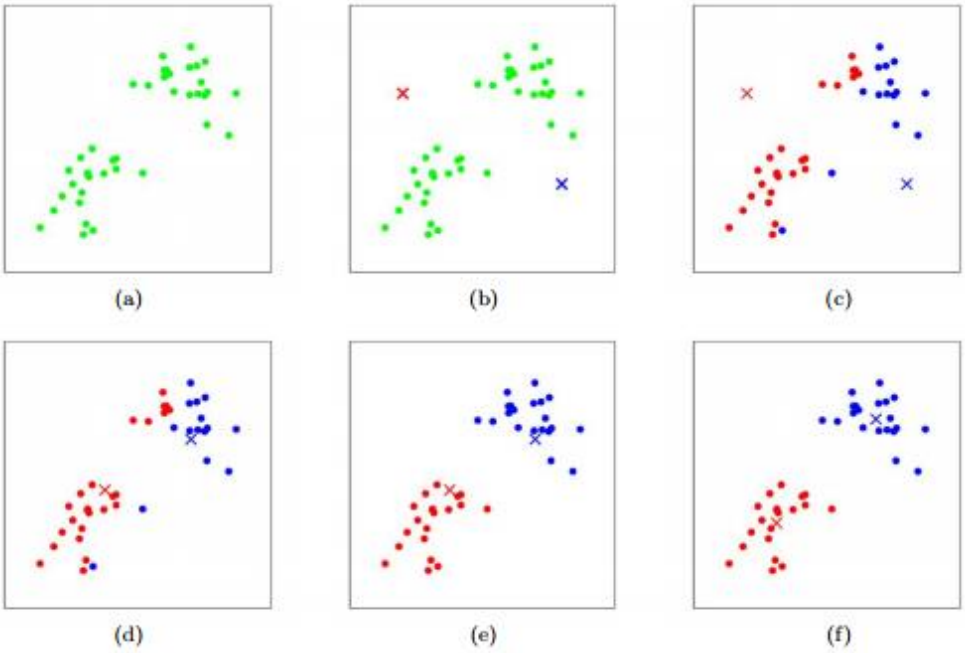


Figure 1: K-means algorithm. Training examples are shown as dots, and cluster centroids are shown as crosses. (a) Original dataset. (b) Random initial cluster centroids. (c-f) Illustration of running two iterations of k-means. In each iteration, we assign each training example to the closest cluster centroid (shown by "painting" the training examples the same color as the cluster centroid to which is assigned); then we move each cluster centroid to the mean of the points assigned to it.

3 Research Question and Objective

What are the various podcast listeners characteristics, and demographic segments in the podcast space?

The objective of this research study is to apply advanced analytics on podcasts and their listeners data to produce valuable insights on different audience segments.

4 Methodology

4.1 Data sources [3]

We are going to be obtaining data on more than 1,500 podcast listeners from an annual Canadian Podcast study. This data contains different publishers, genre and listener's data such as their demographics, their preferred platforms. Potentially we might try to obtain more data about the podcast industry.

4.2 Cleaning

The data is composed of over 500 columns, so we had to do some understanding of the data set and drop columns.

Here is what we did:

We dropped all the columns that contain no data. This was done by replacing white space with NaN values then dropping all columns of NaNs. With that we were left with 400 columns. Still a lot of columns and a lot of potential of unnecessary data.

We looked at % of NaN on the 400 columns left and then we dropped all columns that contain 50% or less NaN values. This left us with 184 columns. This is more manageable.

We then dropped all the columns that contained same data.

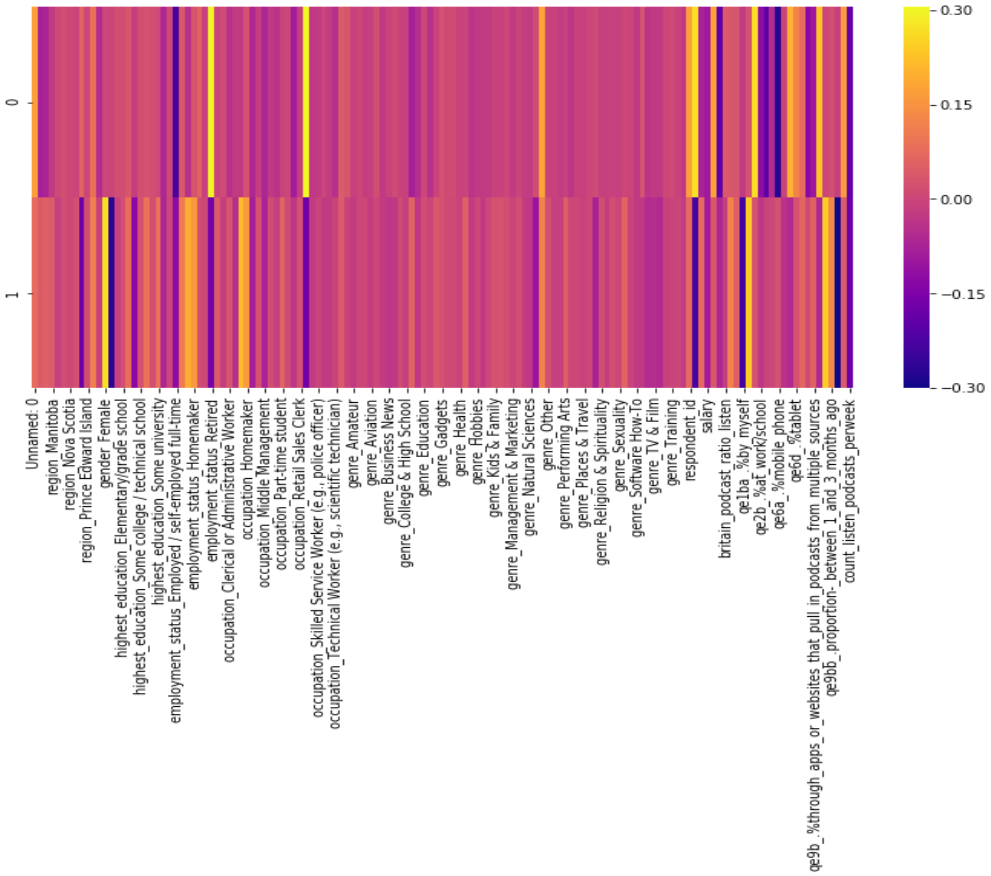
4.3 Data types segregation

We separated our data in to two data frames a numerical and a categorical one.

4.4 Modelling

4.4.1 Categorical data encoding and principal component analysis

We leveraged data encoding method to convert our categorical data into numerical as ones and zeroes. Then, we leveraged Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. [2]



4.4.2 Data Scaling

Since our data contain numerical values with various scales such as ones and zeros, age and income values in thousands. We need to transform our data and make our variable have a mean value of zero and a standard deviation of one.

4.4.3 Data Reduction

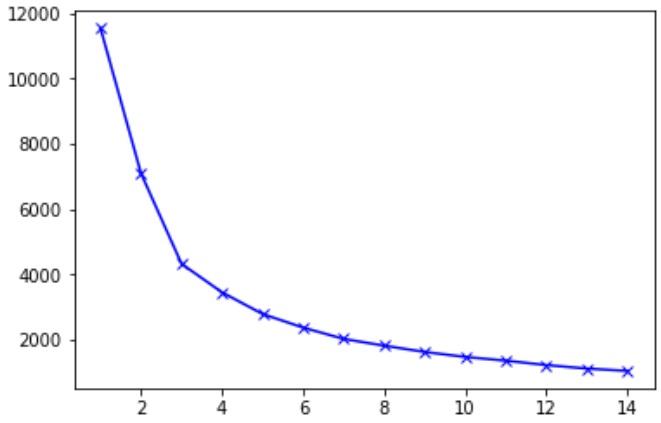
After we used PCA to convert our categorical data into ones and zeros. We leveraged PCA to reduce our large multivariable data set to 2 variables only and retain most of the information.

4.4.4 Clustering

Now since our data is reduced to 2 variables we can now leverage clustering algorithm to measure Euclidean distance between each point and define the groupings of our data points.

4.4.5 Number of centroids of clusters

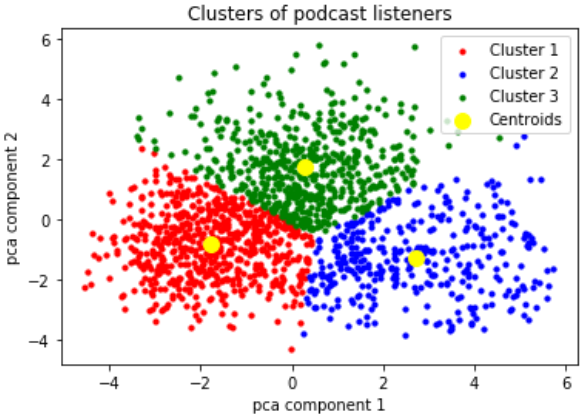
Before we can use Kmeans clustering, we need to determine the inherit optimal number of centroids in our data so we can pass the number of centroids into our Kmeans cluster algorithm. We did that using a method called sum of squared distance and plotted it.



As you can use that after 3 clusters, having more clusters doesn't add much value. So, we decided to go with 3 centroids for our clustering algorithm.

4.4.6 Kmeans clustering

Now, we can pass our data in a Kmean cluster algorithm and it identifies our 3 clusters of data.



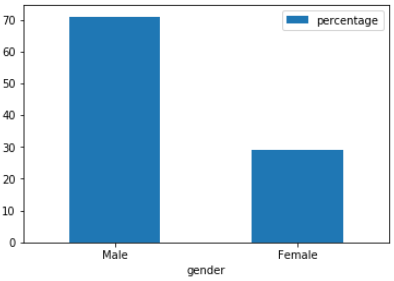
4.4.7 Interpreting Kmeans

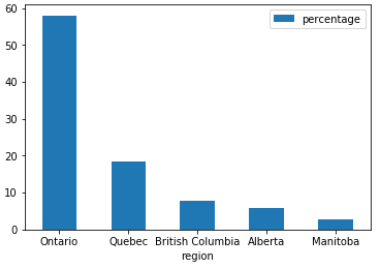
We identified all the responded IDs within each cluster and mapped it back to our categorical data to drive insights on the segmentation.

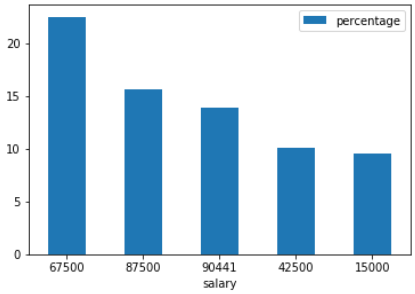
5 Results

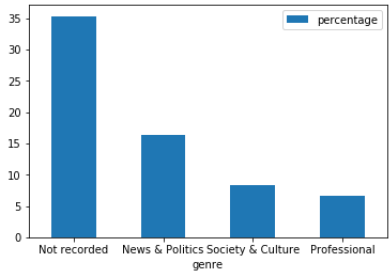
We are now able to identify some of the characteristics of each cluster.

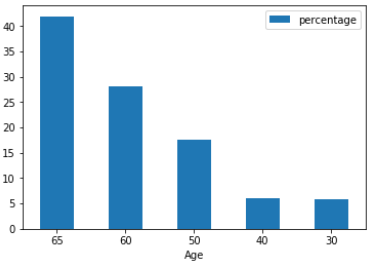
5.1 Cluster 1



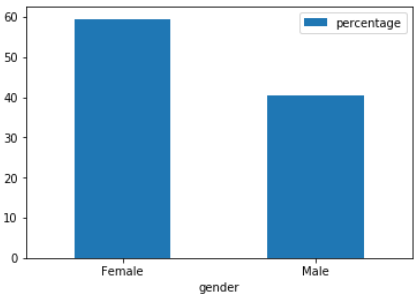


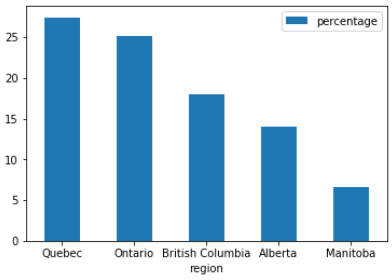


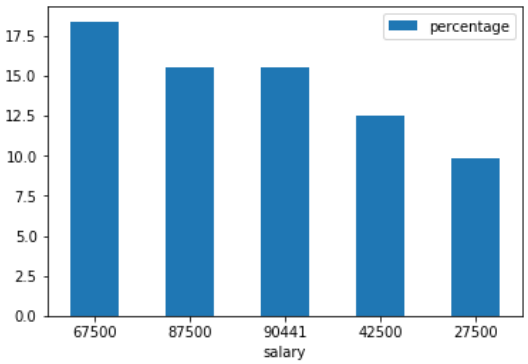


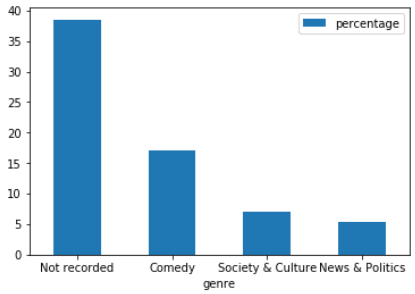


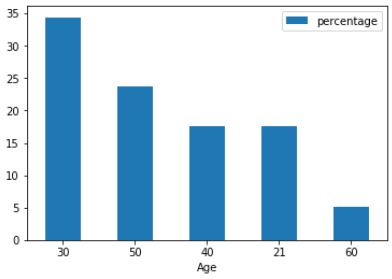
5.2 Cluster 2


percentage
60
50
40
30
20
10
0
Female
Male
gender

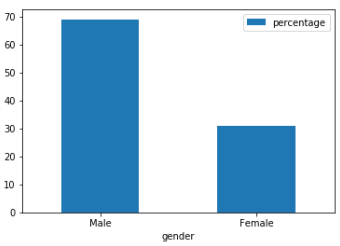

percentage
25
20
15
10
5
0
Quebec
Ontario
British Columbia
Alberta
Manitoba
region

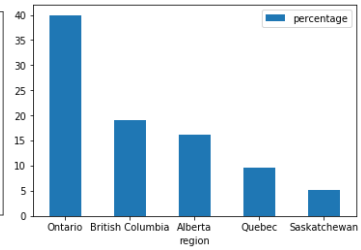

percentage
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
67500
87500
90441
42500
27500
salary

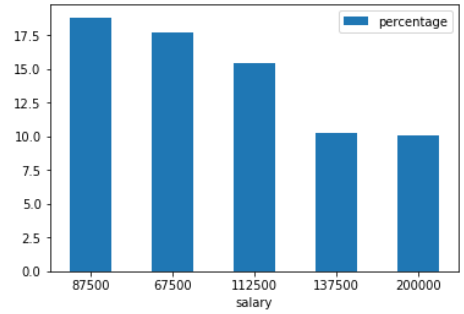

percentage
40
35
30
25
20
15
10
5
0
Not recorded
Comedy
Society & Culture
News & Politics
genre

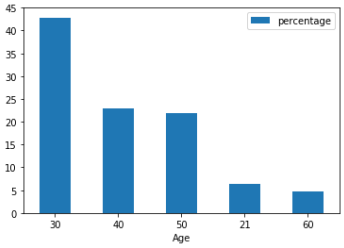

percentage
35
30
25
20
15
10
5
0
30
50
40
21
60
Age

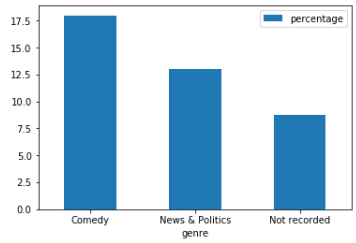
5.3 Cluster 3


percentage
70
60
50
40
30
20
10
0
Male
Female
gender


percentage
40
35
30
25
20
15
10
5
0
Ontario
British Columbia
Alberta
Quebec
Saskatchewan
region


percentage
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
87500
67500
112500
137500
200000
salary


percentage
45
40
35
30
25
20
15
10
5
0
30
40
50
21
60
Age


percentage
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Comedy
News & Politics
Not recorded
genre

6 Discussion

As you can see in the bar plot charts in Result section we have 3 distinct clusters with different characteristics. We can leverage these unique characteristics to advertising agencies to help target their desired audience more precisely.

We can tell that for Cluster One it is characterized by older age listeners in which it is mostly composed of listeners who are 60 years old or above. They mostly reside in Ontario and Quebec City and they are male dominant. The top category they listen to is News/Politics and Society/Culture and they are characterized by average income of 80K per year.

This cluster fits for agencies that want to advertise products or services that are related to men who are close to retirement age within Ontario.

As for cluster two it is characterized by young again but it is very female dominant. They mostly reside in Quebec City and their top listening category is comedy similar to Cluster one.

This cluster fits for agencies that want to advertise products or services that are related to females in their 30s and live in Quebec City or Ontario.

As for cluster three is characterized by young age, male dominant and their top listening category is comedy and they mostly live in Ontario and British Columbia and they have average high income of 106K per year.

This cluster fits for agencies that want to advertise products or services related to men in their 30s within Ontario region.

7 Conclusion

I am glad that we were able to identify various characteristics in our data through unsupervised algorithm models. We were able to drive insights about podcast listeners that will help advertising agencies connect their products with the right audience.

Although we were able to drive insights from our data set, there is an opportunity to improve the results by obtaining more data about podcast listeners so we identify more unique characteristics about them.

Also, there is an opportunity to deploy this model into an app that help advertisers to interact with the data in a more user friendly way.

References

[1] Chris Piech. Based on a handout by Andrew Ng. K Means. https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

[2] Lorraine Li. Principal Component Analysis for Dimensionality Reduction. https://towardsdatascience.com/principal-component-analysis-for-dimensionality-reduction-115a3d157bad

[3] data sources provided by a third party client who is interested in driving insights about podcast listeners. As per our non-disclosure agreement we are not able to share the data itself.