Canadian Podcast Listeners 2018 Data Analyzing - Sprint 1

Aimin Amy Hu

2019-07-24

Background

How podcasters make money as almost every podcast episode is free to download? One of the most popular ways that podcasters make money through sponsors. This mean podcasters need to find companies are willing to pay them to get their brands in front of the podcast's listeners. Understanding the podcast's listeners is a key for sending relevant and effective ADS. However, there are chellenges to podcasters as they don't have data about the listeners even they saw people did download the episodes. Audience Insights Inc started to analyse Canadian podcast's listeners about 3yrs ago. They collected listener's data by survey questions.

Market Segmentation

Market segmentation analysising will be used for 2018 Canadian Podcast Listeners' data. Listeners segments enable us to understand the patterns that differentiate the listeners.Podcast listeners are comprised of deomgraphics that can be very attractive to sellers of goods and services.

Research Questions:

- To what extent are income, household and education level affecting people listening podcast?
- To what extent are gender and age affecting people listening podcast?
- What are the popular 10 podcasts listed from the survery data?(gender, age)
- How listeners react to the advertising in podcasts? (if I have time)

Part 1: Data Understanding

Data Source

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

Downloaded dataset and saved in local computer.

Use R code to read CSV files from local computer.

```r
#start by loading some libraries
library(data.table)
library(dplyr)
library(tidyr)
library(ggplot2)
library(stringr)
library(DT)
library(knitr)
library(grid)
library(gridExtra)
library(corrplot)
library(methods)
library(Matrix)
library(reshape2)
```

```
#set up working directory - this will set the working directory to the same folder as your R studio RMD
set_wd <- function() {
library(rstudioapi) # make sure you have it installed
current_path <- getActiveDocumentContext()$path
setwd(dirname(current_path ))
print( getwd() )
}
```

```
# Read  CSV files along with header and replace empty values with "NA" when read the CSV file.

listener_df <- fread("podcast18.csv",header = TRUE,na.strings = c("") )
```

Part 2: Data Exploration

Step 1: Data Summary

```
cat("The number of observations are", nrow(listener_df))
```

```
## The number of observations are 1534
```

```
cat("The number of variables are", ncol(listener_df))
```

```
## The number of variables are 584
```

Step 2: Checking duplicated rows

```
#checking if there are duplicated rows and removed it. distinct() keep only unique/distinct rows from a
listener_df <- listener_df %>% distinct(id, .keep_all = TRUE)
```

```
cat('Number of rows after removed duplicated rows (if there are) are: ', nrow(listener_df))
```

```
## Number of rows after removed duplicated rows (if there are) are:  1534
```

There are no duplicated rows in this dataset.

Step 3: Missing values

First: check columns with all values are NA

We know this data came from a single-round survey. Surveys often involve ‘skip’ questions where sections are missed out if irrelevant e.g. details of spouse’s employment do not exist for the unmarried. With saying this, we may see a lot of NA (no response) in this dataset.

Let us check if there are any columns with all values are NA(values in entire column are NA).

```
# check columns with all values are NA and display the column's name
variables_with_allNA = sapply(listener_df,function(x) all (is.na(x)))
names(listener_df)[variables_with_allNA]
```

```
##   [1] "arfracep" "arfraceq" "arfracer" "arfraces" "arfracet" "qp4gnia"
##   [7] "qp4gnib"  "qp4gnic"  "qp4gnid"  "qp4gnie"  "qp4gnif"  "qp6aj"
##  [13] "qp6ak"    "qp6al"    "qp6am"    "qp6an"    "qp6ao"    "qp6bf"
##  [19] "qp6bg"    "qp6bh"    "qp6bi"    "qp6bj"    "qp7n"     "qp7o"
##  [25] "qp7p"     "qp7q"     "qp7r"     "qp7s"     "qp7t"     "qe1am"
##  [31] "qe1an"    "qe1ao"    "qe1ap"    "qe1aq"    "qe1ar"    "qe1as"
##  [37] "qe1at"    "qe1au"    "qe1av"    "qe1aw"    "qe1ax"    "qe1ay"
##  [43] "qe9ae"    "qa1d"     "qa1e"     "qa2bf"    "qa2bi"    "qa2bj"
##  [49] "qa2bk"    "qa2bl"    "qa2bm"    "qa2bn"    "qa2bo"    "qa2cc"
##  [55] "qa2cd"    "qa2ce"    "qa2cf"    "qa2cg"    "qa2ch"    "qa2ci"
##  [61] "qa2cj"    "qa2ck"    "qa2cl"    "qa2cm"    "qa2cn"    "qa2co"
##  [67] "qa3e"     "qa3f"     "qa3g"     "qa3h"     "qa3i"     "qa3j"
##  [73] "qm4b"     "qm4c"     "qm4d"     "qm4e"     "qm4f"     "qm4g"
##  [79] "qm4h"     "qm4i"     "qm4j"     "ba01"     "ba02"     "ba03"
##  [85] "ba04"     "ba05"     "ba06"     "ba07"     "ba08"     "ba09"
##  [91] "ba10"     "ba11"     "ba12"     "ba13"     "ba14"     "ba15"
##  [97] "ba16"     "ba17"     "ba18"     "ba19"     "ba20"     "ba21"
## [103] "ba22"     "ba23"     "ba24"     "ba25"     "ba26"     "ba27"
## [109] "ba28"     "ba29"     "ba30"     "ba31"     "ba32"     "ba33"
## [115] "ba34"     "ba35"     "ba36"     "ba37"     "ba38"     "ba39"
## [121] "ba40a"    "ba40"     "ba41"     "ba42"     "ba43"     "ba43a"
## [127] "ba44"     "ba45a"    "ba47"     "ba48"     "ba49"     "ba50"
## [133] "ba51"     "ba52"     "ba52b"    "ba53"     "ba54"     "ba55"
## [139] "ba56"     "ba75"     "ba76"     "ba77"     "ba75b"    "ba76b"
## [145] "ba77b"    "ba78a"    "ba78b"    "ba78c"    "ba79"     "ba80"
## [151] "ba81"     "ba82"     "ba82a"    "ba82b"    "ba82c"    "ba83"
## [157] "ba84"     "ba85"     "ba86"     "ba87"     "ba88"     "ba89"
## [163] "ba90"     "ba91"     "ba91b"    "ba92"     "ba93"     "ba94"
## [169] "bb117"    "bb118"    "bb119"    "bb120"    "bb121"    "bb122"
## [175] "bb123"    "bb124"    "bb123b"   "bb123c"   "bb125"    "bb126"
## [181] "bb127"    "bb128"    "bb129"    "bb130"
```

We see 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.

To keep the data clean for later use, we will subset dataset without these 183 variables.

```
#remove columns with all values are NA
listener_clean <-Filter(function(x) !(all(x=="")), listener_df)
```

```
cat("After removed columns with all values are NA, the number of columns are now", ncol(listener_clean)
```

```
## After removed columns with all values are NA, the number of columns are now 400
```

Second: check if there is any NA in a column

```
names(which(sapply(listener_clean, anyNA)))
```

```
##   [1] "arfracea"                         "arfraceb"
##   [3] "arfracec"                         "arfraced"
##   [5] "arfracee"                         "arfracef"
##   [7] "arfraceg"                         "arfraceh"
##   [9] "arfracei"                         "arfracej"
##  [11] "arfracek"                         "arfracel"
##  [13] "arfracem"                         "arfracen"
##  [15] "arfraceo"                         "qp4tya"
##  [17] "qp4tyb"                           "qp4tyc"
##  [19] "qp4tyd"                           "qp4tye"
##  [21] "qp4tyf"                           "qp4tyg"
##  [23] "qp4tyh"                         "qp4tyi"
##  [25] "qp4tyj"                         "qp4nmb"
##  [27] "qp4nmc"                         "qp4nmd"
##  [29] "qp4nme"                         "qp4nmf"
##  [31] "qp4nmg"                         "qp4nmh"
##  [33] "qp4nmi"                         "qp4nmj"
##  [35] "qp4lna"                         "qp4lnb"
##  [37] "qp4lnc"                         "qp4lnd"
##  [39] "qp4lne"                         "qp4lnf"
##  [41] "qp4lng"                         "qp4lnh"
##  [43] "qp4lni"                         "qp4lnj"
##  [45] "qp4cta"                         "qp4ctb"
##  [47] "qp4ctc"                         "qp4ctd"
##  [49] "qp4cte"                         "qp4ctf"
##  [51] "qp4ctg"                         "qp4cth"
##  [53] "qp4cti"                         "qp4ctj"
##  [55] "qp4exa"                         "qp4exb"
##  [57] "qp4exc"                         "qp4exd"
##  [59] "qp4exe"                         "qp4exf"
##  [61] "qp4exg"                         "qp4exh"
##  [63] "qp4exi"                         "qp4exj"
##  [65] "qp4ava"                         "qp4avb"
##  [67] "qp4avc"                         "qp4avd"
##  [69] "qp4ave"                         "qp4avf"
##  [71] "qp4avg"                         "qp4avh"
##  [73] "qp4avi"                         "qp4avj"
##  [75] "qp4gnaa"                        "qp4gnab"
##  [77] "qp4gnac"                        "qp4gnad"
##  [79] "qp4gnae"                        "qp4gnaf"
##  [81] "qp4gnba"                        "qp4gnbb"
##  [83] "qp4gnbc"                        "qp4gnbd"
##  [85] "qp4gnbe"                        "qp4gnbf"
##  [87] "qp4gnca"                        "qp4gncb"
##  [89] "qp4gncc"                        "qp4gncd"
##  [91] "qp4gnce"                        "qp4gncf"
##  [93] "qp4gnda"                        "qp4gndb"
##  [95] "qp4gndc"                        "qp4gndd"
##  [97] "qp4gnde"                        "qp4gndf"
##  [99] "qp4gnea"                        "qp4gneb"
## [101] "qp4gnec"                        "qp4gned"
## [103] "qp4gnee"                        "qp4gnef"
## [105] "qp4gnfa"                        "qp4gnfb"
## [107] "qp4gnfc"                        "qp4gnfd"
## [109] "qp4gnfe"                        "qp4gnff"
## [111] "qp4gnga"                        "qp4gngb"
## [113] "qp4gngc"                        "qp4gngd"
## [115] "qp4gnge"                        "qp4gngf"
## [117] "qp4gnha"                        "qp4gnhb"
## [119] "qp4gnhc"                        "qp4gnhd"
## [121] "qp4gnhe"                        "qp4gnhf"
## [123] "qp4gnja"                        "qp4gnjb"
## [125] "qp4gnjc"                        "qp4gnjd"
## [127] "qp4gnje"                        "qp4gnjf"
## [129] "qp4sta"                         "qp4stb"
## [131] "qp4stc"                         "qp4std"
## [133] "qp4ste"                         "qp4stf"
## [135] "qp4stg"                         "qp4sth"
## [137] "qp4sti"                         "qp4stj"
## [139] "qp4mra"                         "qp4mrb"
## [141] "qp4mrc"                         "qp4mrd"
## [143] "qp4mre"                         "qp4mrf"
## [145] "qp4mrg"                         "qp4mrh"
## [147] "qp4mri"                         "qp4mrj"
## [149] "qp4pba"                         "qp4pbb"
## [151] "qp4pbc"                         "qp4pbd"
## [153] "qp4pbe"                         "qp4pbf"
## [155] "qp4pbg"                         "qp4pbh"
## [157] "qp4pbi"                         "qp4pbj"
## [159] "qp5cgnab"                       "qp5cgnac"
## [161] "qp5cgnad"                       "qp5cgnae"
## [163] "qp5cgnaf"                       "qp5dgnab"
## [165] "qp5dgnac"                       "qp5dgnad"
## [167] "qp5dgnae"                       "qp5dgnaf"
## [169] "qp6aa"                          "qp6ab"
## [171] "qp6ac"                          "qp6ad"
## [173] "qp6ae"                          "qp6af"
## [175] "qp6ag"                          "qp6ah"
## [177] "qp6ai"                          "qp6ba"
## [179] "qp6bb"                          "qp6bc"
## [181] "qp6bd"                          "qp6be"
## [183] "qp7a"                           "qp7b"
## [185] "qp7c"                           "qp7d"
## [187] "qp7e"                           "qp7f"
## [189] "qp7g"                           "qp7h"
## [191] "qp7i"                           "qp7j"
## [193] "qp7k"                           "qp7l"
## [195] "qp7m"                           "qe1aa"
## [197] "qe1ab"                          "qe1ac"
## [199] "qe1ad"                          "qe1ae"
## [201] "qe1af"                          "qe1ag"
## [203] "qe1ah"                          "qe1ai"
## [205] "qe1aj"                          "qe1ak"
## [207] "qe1al"                          "qe5aa"
## [209] "qe5ab"                          "qe5ac"
## [211] "qe5ad"                          "qe5ae"
## [213] "qe9aa"                          "qe9ab"
## [215] "qe9ac"                          "qe9ad"
## [217] "qa1b"                           "qa1c"
## [219] "qa2ba"                          "qa2bb"
## [221] "qa2bc"                          "qa2bd"
## [223] "qa2be"                          "qa2bg"
## [225] "qa2bh"                          "qa2cb"
## [227] "qa3a"                           "qa3b"
## [229] "qa3c"                           "qa3d"
## [231] "HH_SIZE"                        "PRIMARY"
## [233] "V455_A"                         "QA1_Advertising_Podcast_2_Code"
## [235] "QA1_Advertising_Podcast_3_Code" "QA2c_code_1"
## [237] "QA2c_code_2"                    "qp4nm_neth"
## [239] "qp5cnm_neth"                        "qp5dnm_neth"
## [241] "qp5cdnm_neth"                       "s_ba83"
## [243] "s_ba84"                             "s_ba85"
## [245] "s_ba86"
```

There are 245 variables(columns) with at least one NA value in this data. Since there are many variables with NA values, we will deal with NA values in analysis stage.For now, we will keep them as it is in the dataset.