Podcasts

Angela Zhai

12/14/2018

I. Data Source

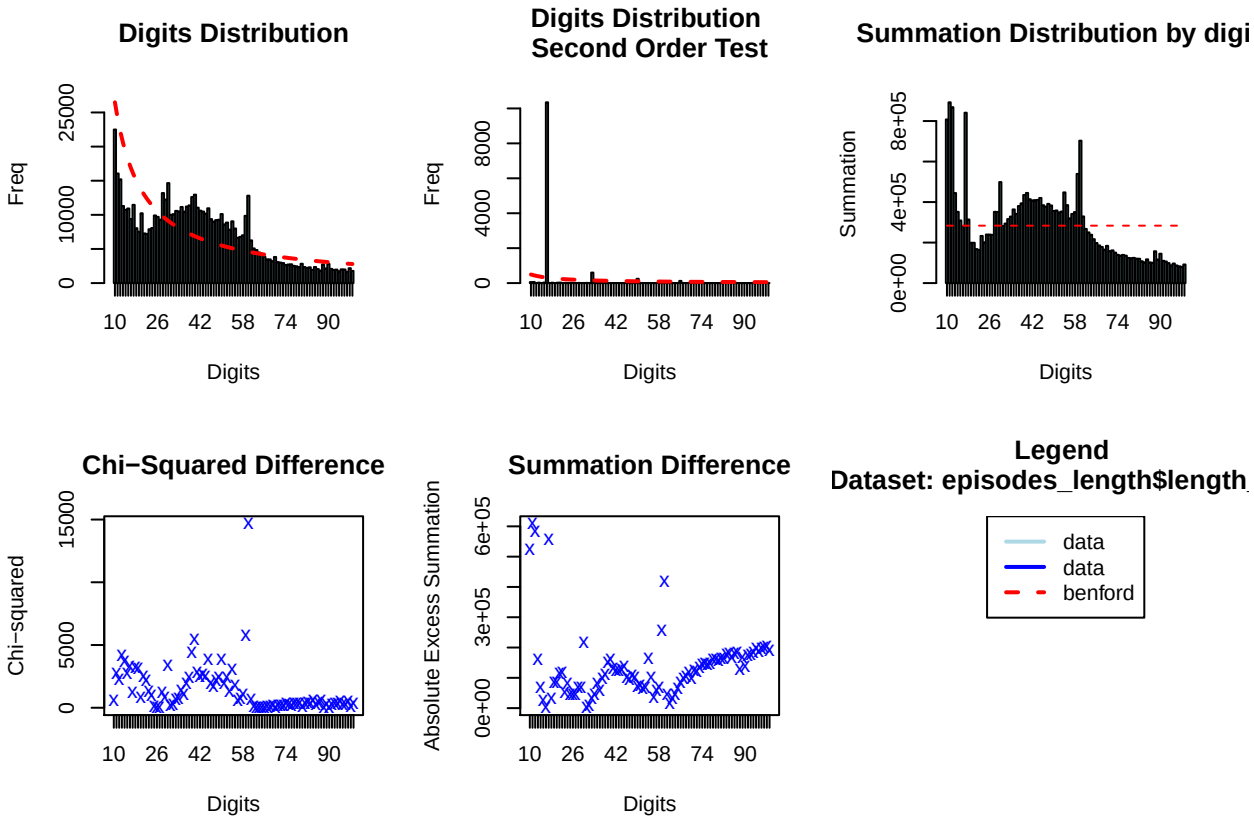
The data comes from Listen Notes (https://www.listennotes.com/), which is a podcast search engine. This dataset includes the meta data of (almost) all podcast episodes that were published in December 2017.

Data *podcasts.csv* contains 121175 diffrent podcasts' information, such as title, language, categories, and description.

Data *episodes.csv* contains 881046 different episodes published in December 2017. The information for those episodes are the audio length, publish date, and episode content description.

II. Benford's Law Test

Following is the Benford's Law test for audio length in minutes of episodes.



The original data (audio length in minutes) is in blue and the expected frequency according to Benford's law is in red.

Benford's analysis of the first digits indicates the data not follows Benford's Law. The mean value of audio length is 39 minutes, and most episodes' length is between 25 to 50 minutes. Therefore, frequency of 10 minutes to 20 minutes is lower than Benford distribution expected frequency, and frequency of 25 minutes to 50 minutes is higher than Benford distribution expected frequency.

```
##
## Benford object:
##
## Data: episodes_length$length_min
## Number of observations used = 639887
## Number of obs. for second order = 11884
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic  Value
##         Mean  0.549
##          Var  0.065
##  Ex.Kurtosis -0.639
##     Skewness -0.449
##
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     13       9300.56
```

```
## 2     14       8436.08
## 3     60       8210.52
## 4     11       8125.41
## 5     16       7457.55
##
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  episodes_length$length_min
## X-squared = 128600, df = 89, p-value < 2.2e-16
##
##
##  Mantissa Arc Test
##
## data:  episodes_length$length_min
## L2 = 0.073192, df = 2, p-value < 2.2e-16
##
## Mean Absolute Deviation: 0.004289441
## Distortion Factor: 0.5563855
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

Above result shows 5 largest discrepancies. As we can see from the plot, the highest deviation is 13.

From the log mantissa of the data, we can tell that the data does not follow Benford's Law. Because the data which follows Benfords's Law should has Mean closes to 0.5, Variance closes to 0.083, Ex. Kurtosis closes to -1.2, and Skewness closes to 0.

Degree of freedom equals 89 and p-value less than 0.01, so failed to accept Benford's law. X-squared value equals 128600 and far away from the value of degree of freedom. All in all, this dataset does not follow Benford's law.

The distortion factor is 0.556.

III. Exploratory Data Analysis

Fig 1. Bar Chart for Podcasts Languages

Only contains top 10 languages

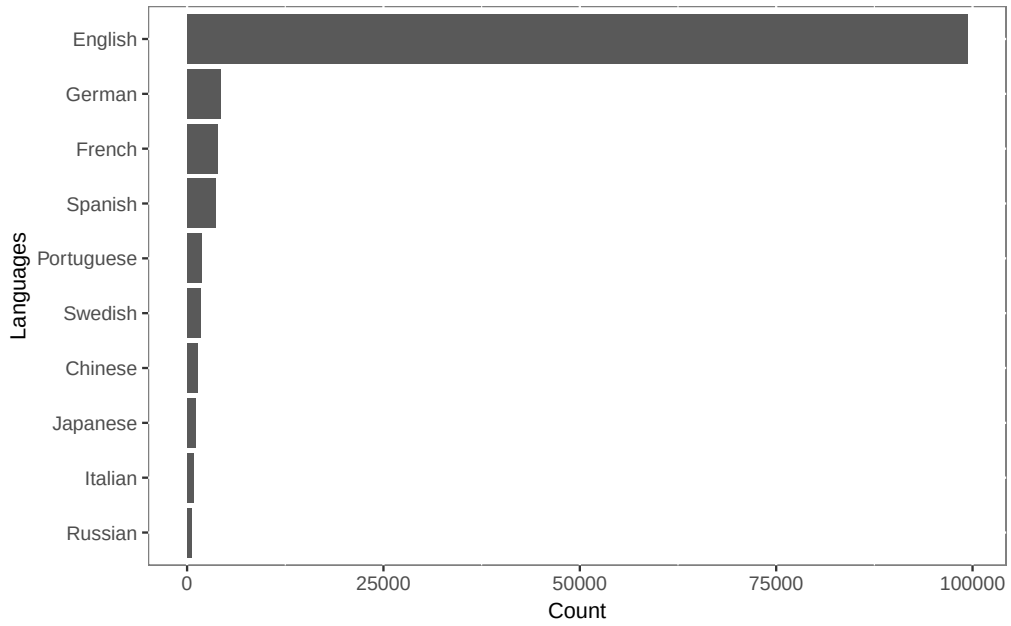
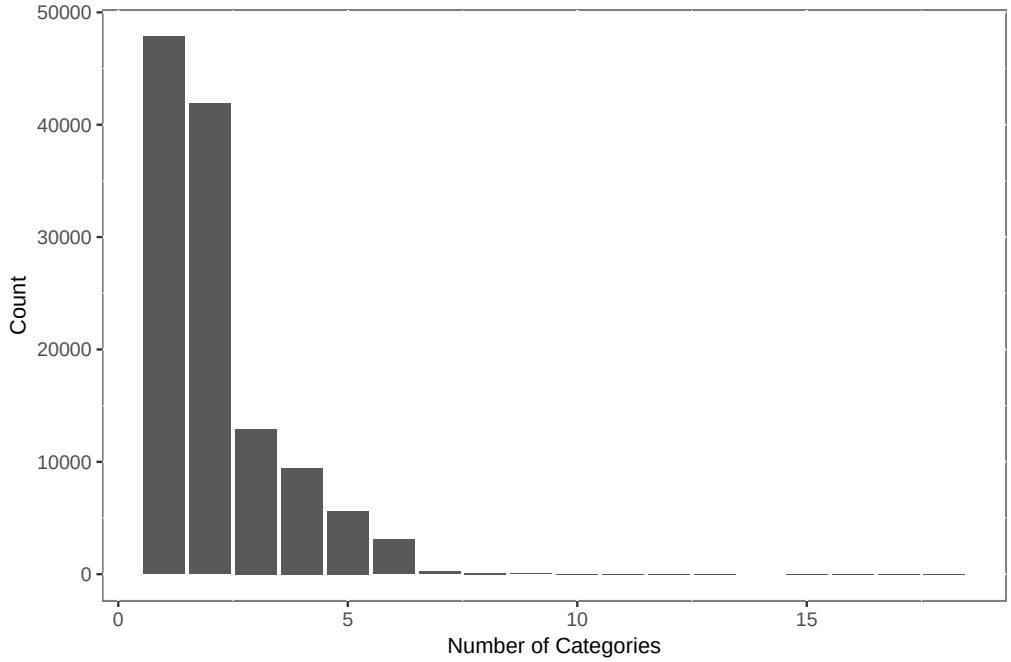


Fig 1 shows top 10 languages used for podcasts. Clearly, most of the podcasts are by English. Only a few used other languages.

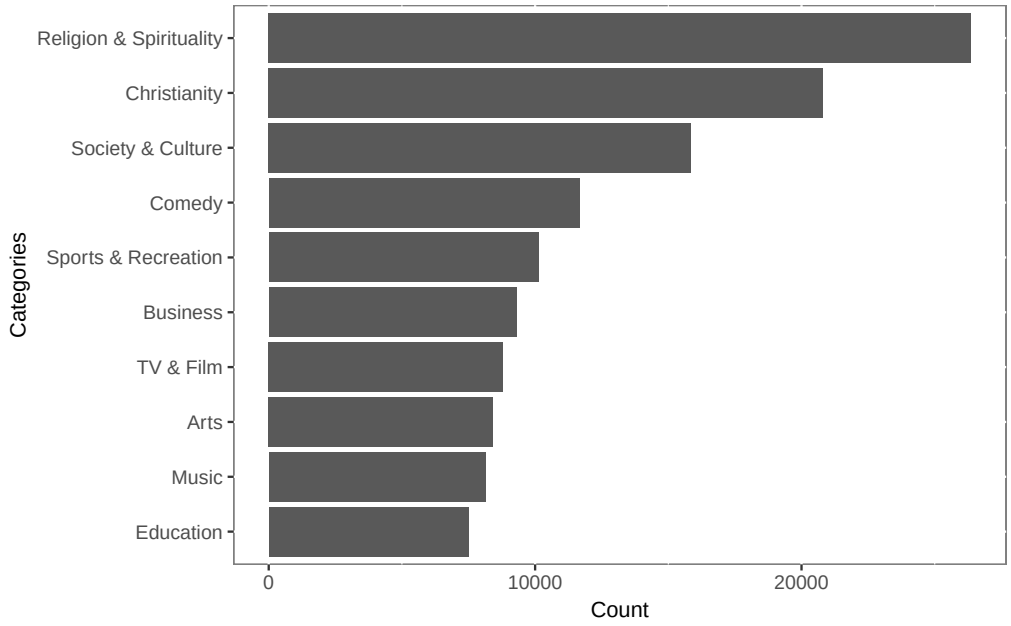
Fig 2. Bar Chart for Number of Categories Assigned to Podcasts



Each podcasts are collected under different categories, and some providers will assign a bunch of categories to their podcasts. Fig 2 tells that most podcasts are tagged only one or two categories, but some are given more than 15 categories.

Fig 3. Bar Chart for Top 10 Categories

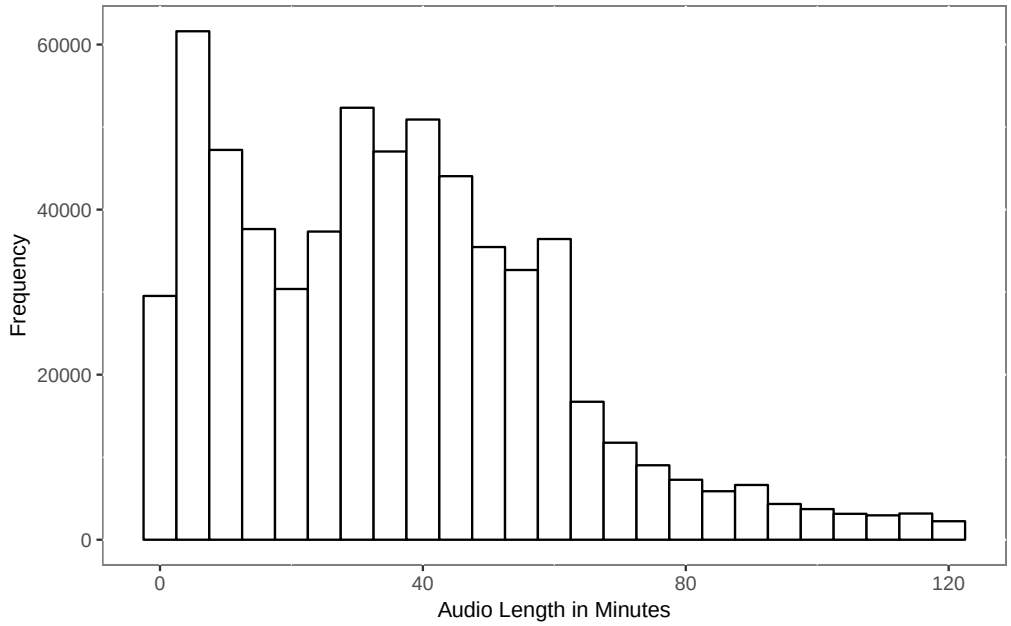
Only contains top 10 categories



Which category has most podcasts? From Fig 3, surprisingly, Religion and Christianity are on the top. Society & Culture is a broad concept for category.

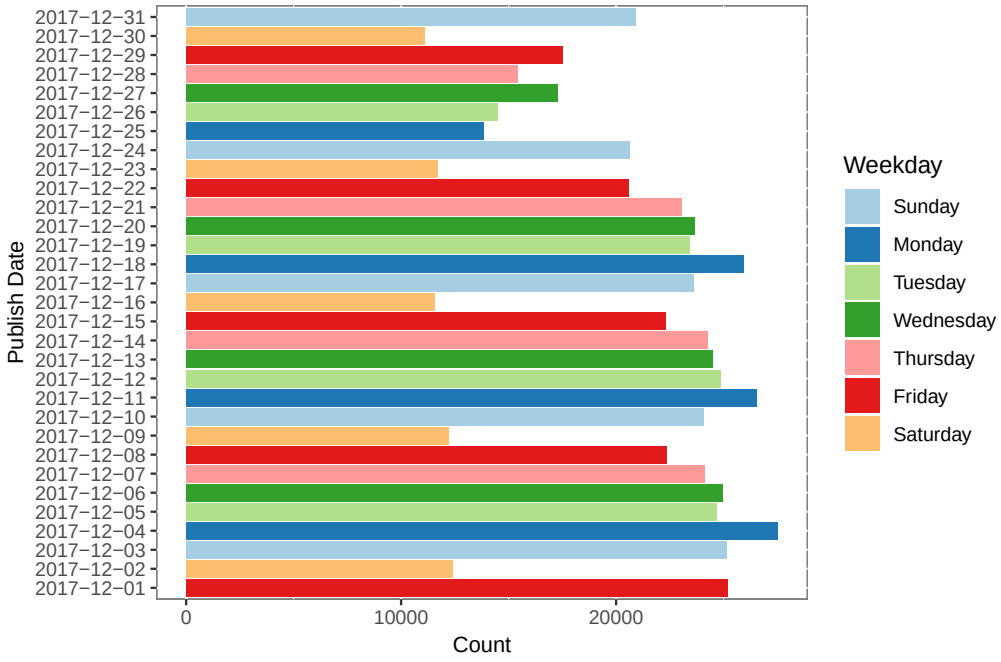
Fig 4. Histogram of Audio Length for Episodes

Only focused on episodes which shorter than two hours



Most episodes are shorter than an hour. From Fig 4, interestingly, a lot of videos are shorter than 10 minutes or between 25 minutes to 45 minutes.

Fig 5. Bar Chart of Publish Date



There is a clear trend in Fig 5 that only a few episodes published on Saturday, and a lot episodes published on Monday. However, holiday is another factor which will influence whether the providers willing to publish their episodes. Like on 25 December, the number of published episodes is different from other Mondays in this month.

IV. Text Mining

Fig 6. 2-grams Wordcloud of Podcasts Description (Music category)

Fig 6 is the wordcloud for music category podcasts description. *Hip Hop* shows a lot. Also there could be many podcasts are talking about electronic music and house music. Most of them are weekly updated and their content is about pop culture.

Fig 7. Words Network of Podcasts Description (Music category)