Television Viewers Segmentation

1 Problem Statement

To analyze the viewing patterns of television viewers by forming clusters with similar viewers in similar categories to enable us to understand the patterns in viewing channels, genres, duration etc.

2 Explanation

This case aims to discover television viewing patterns of viewers from large scale data collected from televisions. The data record information that when and which channel users have viewed. Users select programs to watch according to their preferences, but in some cases they habitually watch programs such as certain morning news or a series. Such daily habits may be difficult to find since they are not based on program's popularity or strong preferences but rather on unintentional customs. It is desirable to find such patterns because, for example, commercials that require viewers' full attention are unlikely to be effective during the unintentional viewing.

3 Literature Review

Television industry is experiencing a revolution with the rapid development of multimedia and network communication technologies. More abundant television contents and richer consumption experiences are provided to users. For example, they can time-shift their viewing, watch their favorite shows repeatedly, or switch to the Internet at anytime. The viewing habits of users are becoming more diverse, and thus it is more difficult to capture users' viewing patterns. Television advertisers and service providers face challenges in attracting users in the incredibly competitive market. Efficient delivery of advertising and contents plays a key role in cutting cost and promoting competitiveness for them. In order to accomplish that, how to extract the target consumer group from the large community of users becomes an issue to be solved in television industry.

4 Data

The file tv-audience-dataset.csv contains the data relative to TV watching behaviour.

Source: http://recsys.deib.polimi.it/?page_id=76

	channel ID	slot	week	genre ID	subGenre ID	user ID	program ID	event ID	duration
0	46	19	1	5	81	1	202344	50880093	5
1	46	20	1	5	81	1	202344	50880093	15
2	46	20	1	3	28	1	254329	50880094	41
3	1	19	1	6	11	2	109428	51094492	11
4	1	19	1	6	86	2	6017	51094494	5
5	1	19	1	5	98	2	6187	51094496	12
6	4	19	1	5	13	2	142037	51092594	1
7	46	19	1	5	81	2	202344	50880093	1
8	46	19	1	5	81	2	202344	50880093	5
9	46	20	1	5	81	2	202344	50880093	15
10	46	20	1	3	28	2	254329	50880094	41

Fields are:
channel ID: channel id from 1 to 217.
slot: hour inside the week relative to the start of the view, from 1 to 24*7 = 168.
week: week from 1 to 19. Weeks 14 and 19 should not be used because they contain errors.
genre ID: it is the id of the genre, form 1 to 8. Genre/subgenre mapping is attached below.
subGenre ID: it is the id of the subgenre, from 1 to 114. Genre/subgenre mapping is attached below.
user ID: it is the id of the user.
program ID: it is the id of the program. The same program can occur multiple times (e.g. a tv show).
event ID: it is the id of the particular instance of a program. It is unique, but it can span multiple slots.
duration: duration of the view.

genre	genreID	subgenre	subGenreID
movie	3	science fiction	51
movie	3	musical	63
movie	3	animation	17
movie	3	dramatic	29
movie	3	fantastic	46
movie	3	war	67
movie	3	funny	56
movie	3	adventure	4
movie	3	action	34
movie	3	comedy	3
movie	3	thriller	28
movie	3	cinema	7
movie	3	romantic	45
movie	3	biography	54
movie	3	western	19
movie	3	erotic	82
movie	3	short film	66
movie	3	horror	24
movie	3	crime	42
undefined	7	undefined	47

other programs	8 other	112
other programs	8 other programs	50
other programs	8 educational	108
other programs	8 special events	107
other programs	8 shopping	96
entertainment	5 mini series	104
entertainment	5 TV soap	102
entertainment	5 series	92
entertainment	5 science fiction	65
entertainment	5 reality show	61
entertainment	5 soap opera	64
entertainment	5 animation	71
entertainment	5 show	13
entertainment	5 dramatic	74
entertainment	5 docu-fiction	109
entertainment	5 theatre	8
entertainment	5 exhibition	41
entertainment	5 talk show	80
entertainment	5 entertainment	81
entertainment	5 quiz	95
entertainment	5 sit com	36
entertainment	5 fiction	31
entertainment	5 game	98
society	4 cinema magazine	32
society	4 travel magazine	87
society	4 reportage	59
society	4 nature	30
society	4 magazine	60
society	4 topical	10
society	4 music	55
society	4 adventure	75
society	4 fishing	93
society	4 lifestyle	38
society	4 society	27
society	4 technology	84
society	4 religion	52
society	4 history	26
society	4 documentary	5
society	4 economics	111
society	4 hobby	91
society	4 sport	99
society	4 nature magazine	48
society	4 fashion	89
society	4 travels	23

society	4	culture magazine	25
society	4	communities	76
society	4	cinema	53
society	4	culture and society	69
society	4	science magazine	94
society	4	science	15
society	4	politics	62
society	4	cooking	37
society	4	art and culture	6
sport	2	winter sports	79
sport	2	other	20
sport	2	poker	58
sport	2	hockey	22
sport	2	cycling	21
sport	2	wrestling	83
sport	2	sport	43
sport	2	soccer	16
sport	2	equestrian	72
sport	2	rugby	44
sport	2	swimming	78
sport	2	athletics	68
sport	2	baseball	40
sport	2	football usa	39
sport	2	boxe	57
sport	2	basket	2
sport	2	motors	33
sport	2	golf	70
sport	2	sail	97
sport	2	skiing	14
sport	2	tennis	12
sport	2	volley	18
sport	2	handball	103
kids and music	1	series	85
kids and music	1	concert	106
kids and music	1	documentary	110
kids and music	1	animation movie	105
kids and music	1	videoclip	101
kids and music	1	dance	73
kids and music	1	magazine	77
kids and music	1	music	1

kids and music	1 games	100
kids and music	1 cartoons	35
kids and music	1 educational	49
kids and music	1 children	88
kids and music	1 null	114
information	6 news	9
information	6 economics	90
information	6 sport	86
information	6 weather forecast	113
information	6 newscast	11

5 Deliverable

1. Code including all processes from ingestion to output
2. Output i.e. cluster labels
3. Analysis Graphs
4. Output cluster graph

6 Evaluation

Evaluation will be done based on silhouette score.

The Silhouette Coefficient is calculated using the mean intra-cluster distance (`a`) and the mean nearest-cluster distance (`b`) for each sample. The Silhouette Coefficient for a sample is `(b - a) / max(a, b)`. To clarify, `b` is the distance between a sample and the nearest cluster that the sample is not a part of. Note that Silhouette Coefficient is only defined if number of labels is 2 <= n_labels <= n_samples - 1.

7 Data Ingestion

Data is ingested in form of pandas dataframe from a csv file.

```
def Ingest_data():
    """
       Read the data in form of a pandas dataframe.

       Returns:
       -------
       data : (M,9) pd.Dataframe

    """
    data = pd.read_csv( PATH_TO_DATA,
                        names= NAME_COLUMNS,
                        nrows=20000)
    return data
```

8 Data Analysis

Here data is analyzed and visualized to look for patterns and anomalies using graphs. bar graphs, histograms, scatter plots, etc. will be used.

```
def Analysis_data():
    """
       This function analyzes the data and looks for patterns
       to help understand the data more clearly.

       Returns:
       --------
       Charts, graphs for various Features.

    """
    _scatter_plot('user ID','channel ID')
    _histogram_plot('user ID','genre ID')
    _scatter_plot('user ID','duration')
    _histogram_plot('user ID','slot')
    _histogram_plot('channel ID','slot')
    _histogram_plot('genre ID','duration')
    _box_plot('user ID','program ID')
    _bar_chart('user ID')
    _bar_chart('channel ID')
    _bar_chart('slot')
    _bar_chart('genre ID')
    _bar_chart('duration')
    _bar_chart('subGenre ID')
    _bar_chart('week')
```

From small analysis, we take data of first 5000 records. We can see various patterns (which might vary differently as number of records are increased) such as -

=> People usually prefer to see channels between 0 to 25, rarely between 50 to 165 and occasionally channels after 165.

=> People watch genres 4 and 5 the most, 7 after that, 2,3,6 occsionally and rest of them rarely.

=> People usually watch television between 0 to 10 hours and with less frequency beyond 10 hours.

=> The most used slot for watching are 20 and 21 and least are 0 to 5.

=> Genres 0 to 5 are watched for maximum duration of time.

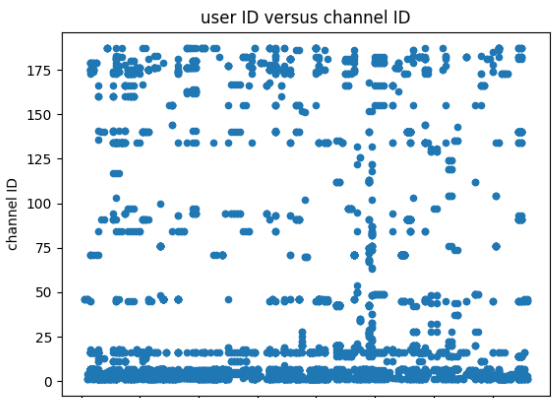
=> There are some users who watch very varied types of channels and genres and very few who watch very less range of programs.

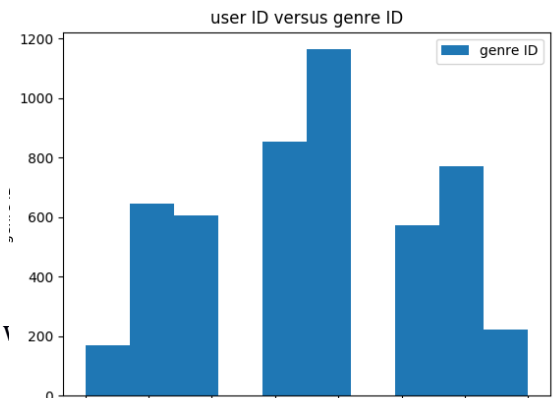
=> Only few channels are being watched very extensively.

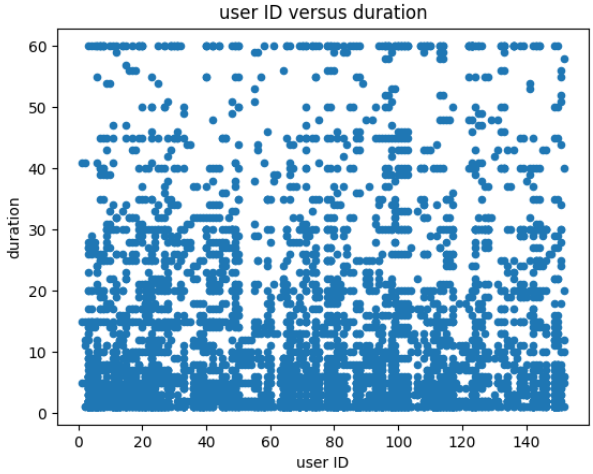
=> Genre 5 is most liked by viewers and genre 1 the least. This shows that viewers watch television for entertainment purposes maximum times.

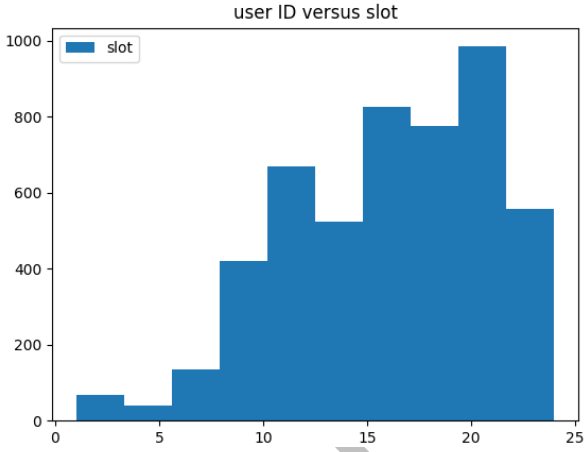
=> Duration for which viewers watch television is maximum for 1 hour and viewing proportion gradually decreases with increase in duration but there is s spike at duration 60 hours which means there are some viewers who watch television all most all the time!

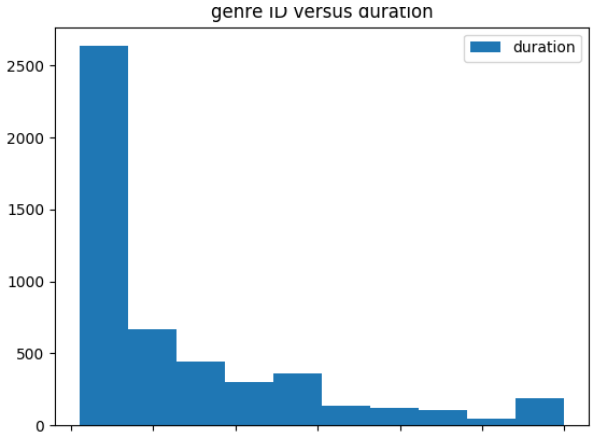
=> From sub-genres, it can be seen that viewers like to watch undefined random things most of the time, then fiction the most, followed by entertainment shows and newscast. The least watched sub-genre are in sports such as cycling, swimming, basketball and also weather forecast, talk shows.

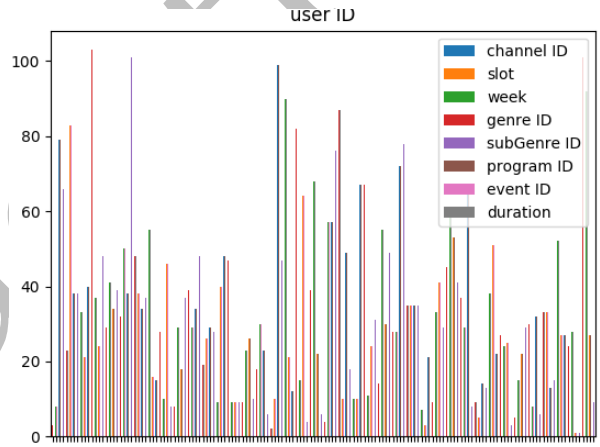


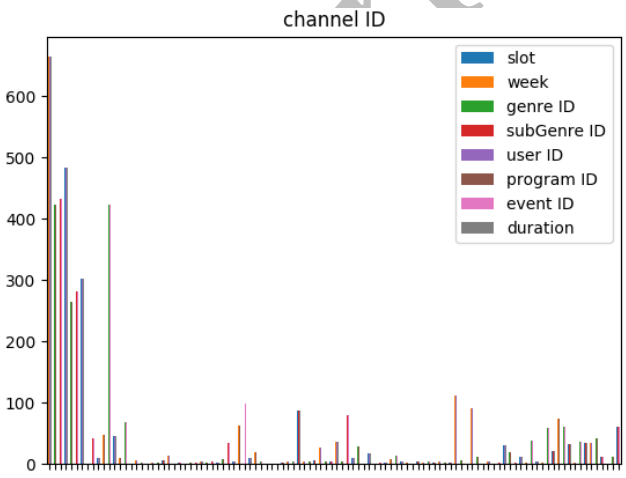


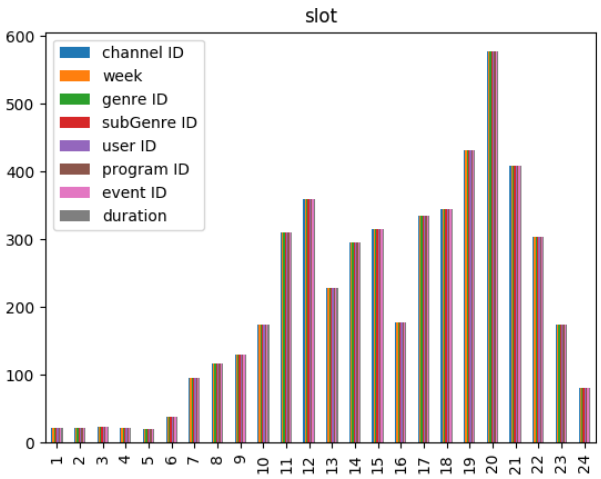

user ID versus duration
duration
user ID

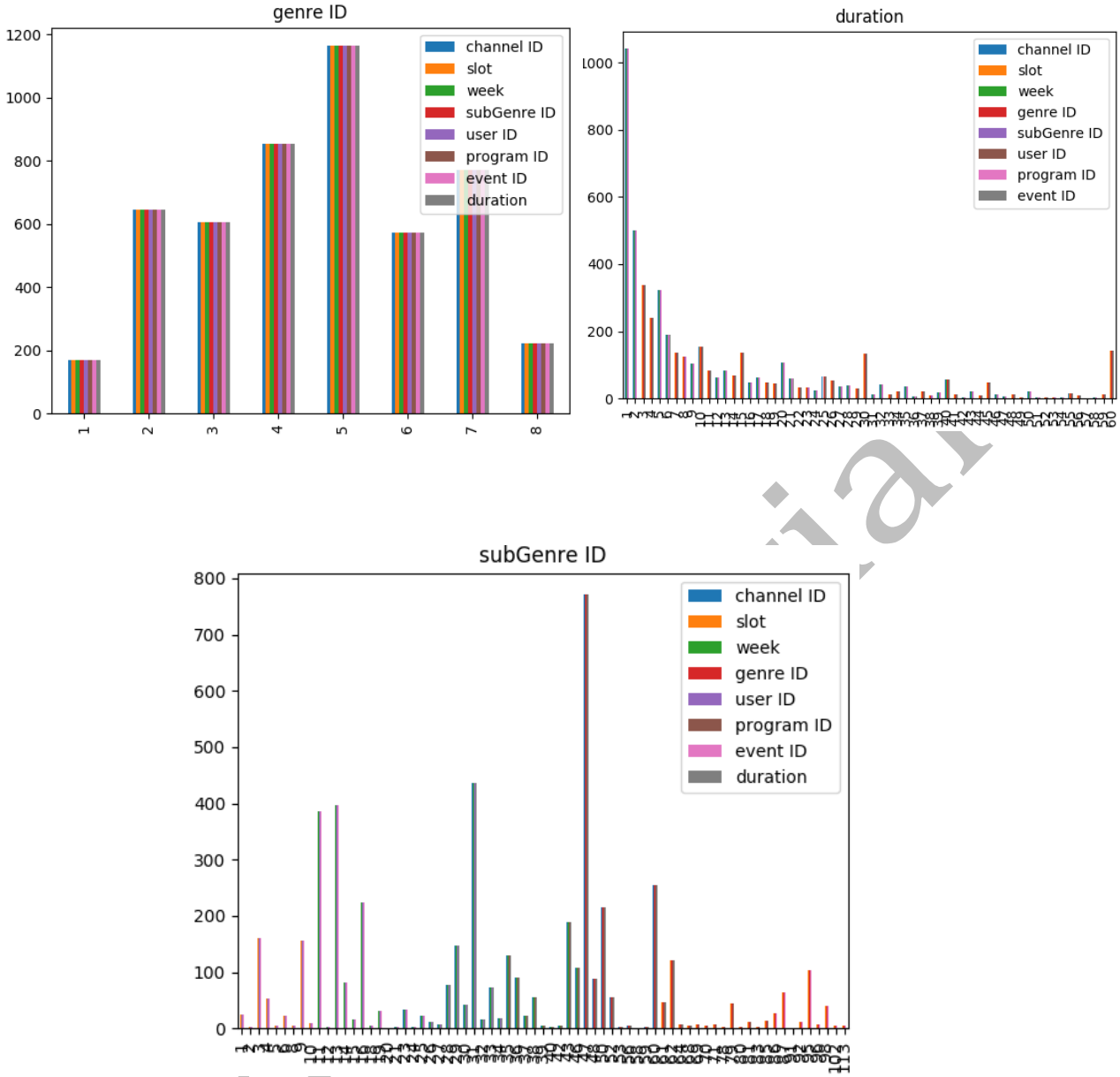

user ID versus slot
slot


genre ID versus duration
duration


user ID
channel ID
slot
week
genre ID
subGenre ID
program ID
event ID
duration


channel ID
slot
week
genre ID
subGenre ID
user ID
program ID
event ID
duration


slot
channel ID
week
genre ID
subGenre ID
user ID
program ID
event ID
duration



As we see from above plots and their deduction, features are varied and less dependent on each other.

To know a certain pattern of clustering, we can take different features two at a time and create clusters on those features to obtain a certain pattern in television viewing.

For now, we cluster the entire data with all features together to get a certain pattern with all information included.

9 Data Munging

On the basis of information of data provided and on based on Data Analysis, data needs to be rearranged, reorganized and made into a proper structure to make it ready for Data modeling process.

Weeks 14 and 19 should not be used because they contain errors. Hence those records were deleted.

```
def Munging_data():
    """
       Rearrange and reorganise the  data  according  to  the
       problems needed to solve  so  the  modelling occurs on
       proper datasets.

       Returns:
       --------
       df : pd.Dataframe
            Dataset for further exploration

    """
    df_ = df[~df['week'].isin([14,19])]
    return df_
```

10 Data Exploration

The data might contain outliers, missing values and various other things which might cause problems in modeling or might lead to wrong results. These anomalies need to be corrected and data must be cleaned before passing it to modeling.

```
def Exploration_data():
    """
       Function to apply data exploration techniques.

       Returns:
       --------
       df_explored : (M,N) pandas DataFrame
                     Data

    """
    df_ = df.drop('program ID',axis=1).drop('event ID',axis=1).astype(float)
    arr = np.array(df_)
    outliers = np.apply_along_axis(_outlier_detection, 0, arr)
    arr[outliers] = np.nan
    arr = _missing_value_treatment(arr)
    df_explored = pd.concat([pd.DataFrame(arr, columns=COLS_REMAINING), df[['program ID', 'event ID']]], axis=1)
    return df_explored
```

11 Feature Engineering

Not Required

12 Modeling

The data finally pre-processed and cleaned is now used for clustering and obtaining the results. We use K-Means model and HAC model for prediction and trace as our evaluation metric .

The silhoutte score will be calculated for both. The model performing better will be chosen.

13 Optimization

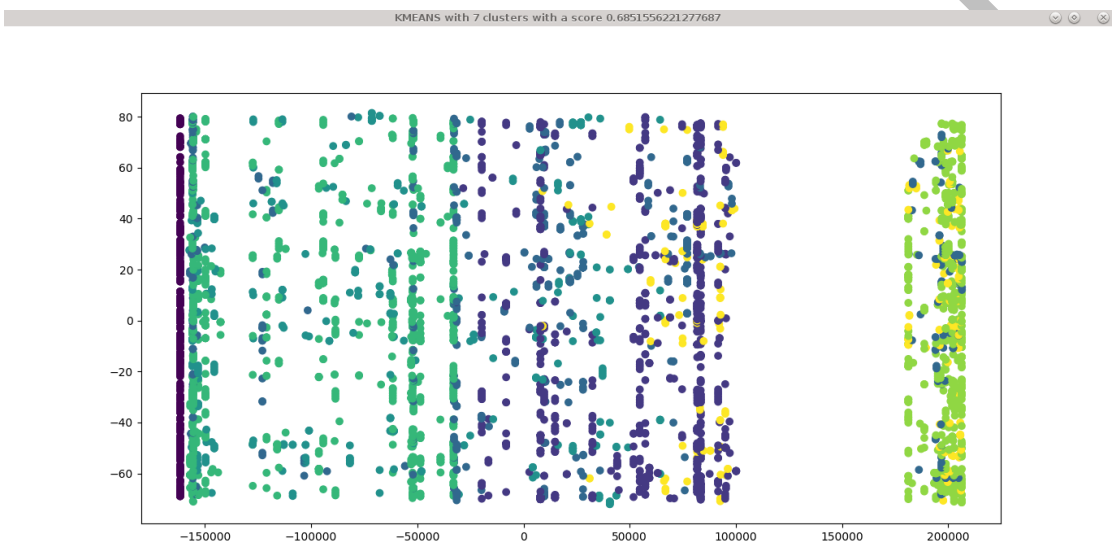
The hyper-parameters of a model need to be optimized to get best possible score of prediction. For this purpose, we use Random Grid Search.

50 Iterations are performed random combination of parameters for each Kmeans and HAC. The model having best score is chosen from all as our final model for clustering. The silhoutte score for best model will come around 0.70 to 0.75.

14 Prediction

From best model obtained from hyper-parameter optimization, we predict cluster labels for the data, which will be our prediction for the problem.

15 Visual Analysis



As we can see from the scatter plot, the clusters are well formed according to the data space. This leads to conclusion that our model performed good and the clusters can be used for further business purposes.

16 Results

K-means outperformed HAC and scored a best silhouette score of 0.68 with 7 clusters.

	channel ID	slot	week	genre ID	subGenre ID	user ID	duration	program ID	event ID	Cluster
0	46.0	19.0	1.0	5.0	81.0	1.0	5.0	202344	50880093	3
1	46.0	20.0	1.0	5.0	81.0	1.0	15.0	202344	50880093	3
2	46.0	20.0	1.0	3.0	28.0	1.0	41.0	254329	50880094	6
3	1.0	19.0	1.0	6.0	11.0	2.0	11.0	109428	51094492	4
4	1.0	19.0	1.0	6.0	86.0	2.0	5.0	6017	51094494	4
5	1.0	19.0	1.0	5.0	98.0	2.0	12.0	6187	51094496	4
6	4.0	19.0	1.0	5.0	13.0	2.0	1.0	142037	51092594	1
7	46.0	19.0	1.0	5.0	81.0	2.0	1.0	202344	50880093	3
8	46.0	19.0	1.0	5.0	81.0	2.0	5.0	202344	50880093	3
9	46.0	20.0	1.0	5.0	81.0	2.0	15.0	202344	50880093	3
10	46.0	20.0	1.0	3.0	28.0	2.0	41.0	254329	50880094	6
11	7.0	6.0	1.0	6.0	11.0	3.0	2.0	109509	51094300	4
12	4.0	6.0	1.0	6.0	11.0	3.0	2.0	128701	51092527	4
13	7.0	6.0	1.0	6.0	9.0	3.0	9.0	171429	51094301	1
14	7.0	7.0	1.0	6.0	9.0	3.0	17.0	171429	51094301	1
15	6.0	7.0	1.0	6.0	9.0	3.0	1.0	244916	51097384	1
16	1.0	7.0	1.0	5.0	31.0	3.0	2.0	53001	50824698	3
17	1.0	7.0	1.0	5.0	31.0	3.0	1.0	197319	50824700	3
18	16.0	7.0	1.0	7.0	47.0	3.0	1.0	1	1	0
19	1.0	7.0	1.0	6.0	11.0	3.0	6.0	109509	51122125	4
20	7.0	7.0	1.0	6.0	9.0	3.0	20.0	171429	51094301	1