CMPE 257 – Arpitha Gurumurthy

Homework 2 – Clustering with K-Means, GMM and Birch algorithms

Link to colab:
https://drive.google.com/drive/u/2/folders/1AsrAq7yPbLna2yuXyzowoAzMnoLrhchO

Link to data:
https://drive.google.com/drive/u/2/folders/1ntQ3EiY6xZu6UfyL_X1ARrcOP1KskDv2

Goal:
To help small scale online shops to maximize income by inventory replacement / hyper-personalization.

Using the below dataset and picking only one month – December 2019:
https://www.kaggle.com/mkechinov/ecommerce-events-history-in-cosmetics-shop

Pre – requisites:

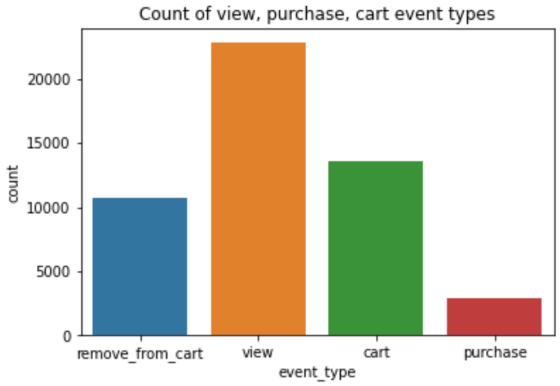
1. Loading the dataset from google sheets using the corresponding link.

```
[2] df_Dec.shape

    (50000, 9)
```

```
df_Dec.isnull().sum()

event_time            0
event_type            0
product_id            0
category_id           0
category_code     49144
brand             21941
price                 0
user_id               0
user_session         15
dtype: int64
```

2. Data wrangling:

- Dropped the column ‘category_code’ since it contains too many null values.
- Replaced all null values in the columns brand and user_session with ‘Not Available’ as part of data cleaning.
- ‘event_time’ – converted the date type to date time type and split the column into 2 separate columns, one containing the date and the other containing the time for each event.
- There are 4 types of event_type in the dataset: ‘view’, ‘cart’, ‘remove_from_cart’ and ‘purchase’.

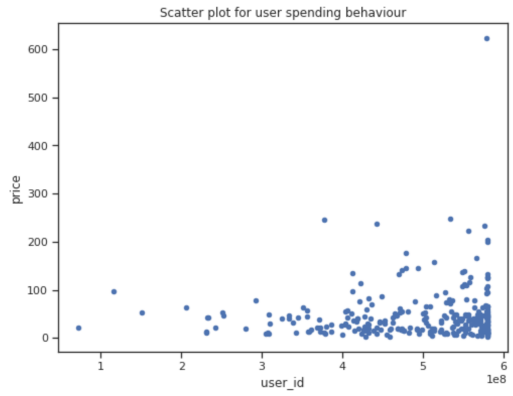


CLUSTERING:

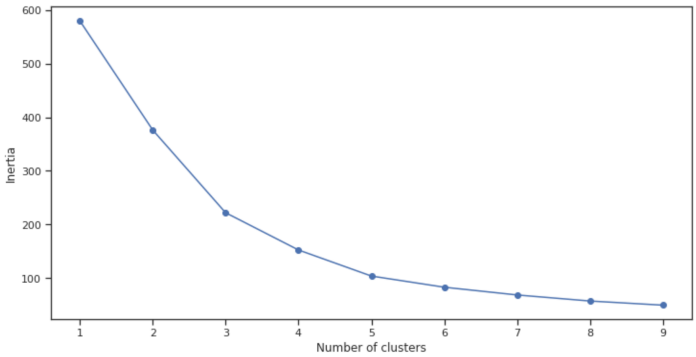
Trying to cluster the customer spending behavior.

Steps:

- Calculated the total amount spent by each user on the online shop by filtering out the rows with only event_type as ‘purchase’.
- Then using ‘.agg’ function, calculated the amount spent per user (using columns – ‘user_id’ and ‘price’)
- On plotting the above calculated sum of price per user using scatter plot, we see the below cluster:



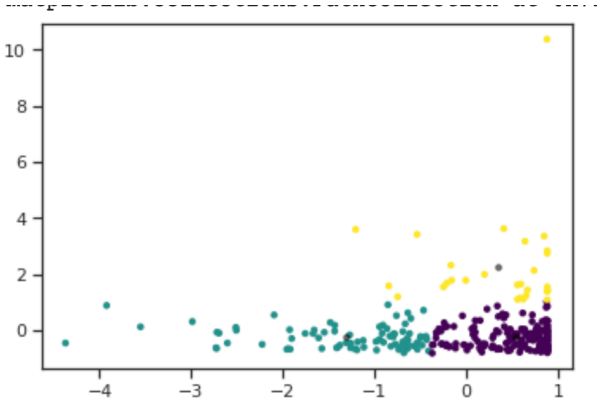
- Using the elbow plot to calculate the ideal number of clusters:



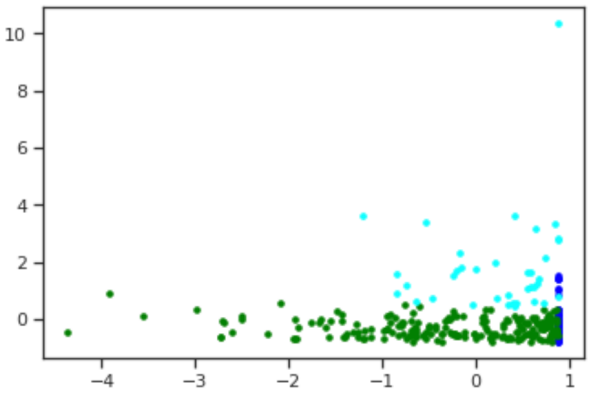
Note : Optimal numbers of clusters : 3 to 5

ALGORITHMS:

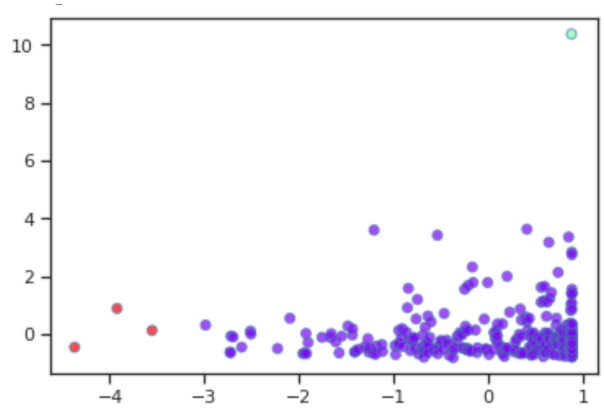
- **Using k-means with number of clusters as 3**:



- **Using GMM with number of clusters as 3:**



- **Using Birch with number of clusters as 3:**



Comparison analysis:

Sl no.	K-Means	GMM	Birch
1	We can see that the clusters are not overlapping, but due to the shape of the data points the clusters are placed very close to one another.	We can see that the clusters are overlapping with another. This may not be suitable to our data points.	The clusters are not equal in size but the distance between the clusters is better defined when compared to the other 2 algorithms.
2	Silhouette score with 3 clusters = 0.523451108850259	Silhouette score with 3 clusters = 0.12111914281419185	Silhouette score with 3 clusters = 0.6123334940423968
3	calinski_harabasz_score = 231.34412204332176	calinski_harabasz_score = 79.43743359495102	calinski_harabasz_score = 52.823586666789744
4	davies_bouldin_score = 0.6908376495519818	davies_bouldin_score = 1.085960353512334	davies_bouldin_score = 0.30980348220091636

From the above evaluation metrics, we can say that Birch is the best clustering technique for our data.