BMS COLLEGE OF ENGINEERING

(Autonomous Institute, Affiliated to VTU, Belagavi)

DEPARTMENT OF MACHINE LEARNING

(UG Program: B.E. in Artificial Intelligence and Machine Learning)

Course : MOOC with Project
Course Code: 22AM6PWMWP

Event Feedback Analysis

Semester End Examination : Project Presentation

Date: 22nd July, 2023

Presented By,

Student Name & USN :

MONESH S 1BM20AI039

PRABHAT G P 1BM20AI043

SANKETH P 1BM20AI048

SIDDARTH A 1BM20AI049

Semester & Section: **6A**

Batch Number: B4

Faculty In-Charge:

Dr. Monika P

Assistant Professor

Department of Machine Learning

BMS College of Engineering

Agenda

- Introduction
- Literature Review
- Open Issues
- Problem Statement
- Proposed Architecture
- Functional & Non-Functional Requirements
- Methodology
- Implementation
- Testing and Validation
- Results and Discussion
- Conclusion
- About MOOC (Details : Title, number of hours)
- References

Introduction

- Understanding the sentiments and preferences of event attendees is crucial for organizers to make data driven decisions.
- Clustering and comparing each class of audience against each feature helps us to give a broad idea about different audience.

- Our projects to aims to extract valuable insights from the event feedback dataset by employing unsupervised learning techniques such as K-means and PCA decomposition.
- Creating a detailed Power BI dashboard which can be used by the stakeholders to properly analyze every class of audience with any feature they want.

⇒

Literature Review

TITLE / YEAR	APPLIED METHODOLOGY / ALGORITHM USED	FINDINGS	RESULTS	LIMITATIONS
Understanding Customer Experience and Satisfaction through Airline Passengers’ Online Review , 2019	CONCOR (CONvergence of iterated CORrelation)	All evaluation factors except ‘entertainment’ factor significantly had impact on customer satisfaction and recommendation	Online review can provide both academic implication and practical implication to develop sustainable strategy in the airline industry	model couldn't handle large datasets
Sentiment Analysis of Events in Social Media 2019	Network Analysis using Natural Language Processing	Focuses on the network features, without an in-depth analysis of the textual content. Natural Language Processing analyses only the textual content, not integrating the graph-based structure of the network.	We can observe that MABED and OLDA manage to detect different emerging events when analyzing the most representative topic keywords using the text preprocessing CT, although some are the same.	model was taking long duration to compute results
Opinion mining on large scale data using sentiment analysis and k-means clustering 2017	K-means clustering	Clear insight of customer preference and behavior to help decision makers for better decision making	Sentiment analysis on the large scale dataset of product (6 categories) reviews given by various customers on the internet	categories were including less insightful information

Literature Review

TITLE / YEAR	APPLIED METHODOLOGY / ALGORITHM USED	FINDINGS	RESULTS	LIMITATIONS
Students feedback analysis model using deep learning-based method 2023	DTLP - Combination of CNNs, Bidirectional LSTMs and Attention Mechanism	Unified feature set, which is representative of word embedding, sentiment knowledge, sentiment shifter rules, linguistic and statistical knowledge.	The results showed that DTLP outperforms the existing systems in the field.	The major limitations related to this work include the pre trained word embedding method, which is a google pretrained model that contains public online data.
Crowd characterization for crowd management using social media data in city events, 2019	DBSCAN Clustering Algorithm	Crowd managers could apply crowd management measures by taking into consideration the semantic and qualitative interpretation of social media posts.	The results show that less than 1% of users performed this operation.	The limited availability of meaningful profile pictures and location information reduces the intrinsic utility of sociodemographic analysis of social media data
Applications of Student Feedback using Machine Learning Model 2022	Naïve bayes and Random forest	These are naive Bayes and random forest classification techniques that use the joint probabilities of classes and words to determine the class probabilities assigned to texts	Naive bayes accuracy gives 95% and random forest accuracy gives 30%	Random forest doesn't give better solution.

Open Issues

- **Data Quality and Quantity:** There may be many incomplete and biased feedbacks by the attendees.
- **Subjectivity and Sentiment Analysis**: Acknowledge the potential for misinterpretation of the user feedback.
- **Ethical Consideration:** Address any privacy concerns related to the collection of the feedback data.

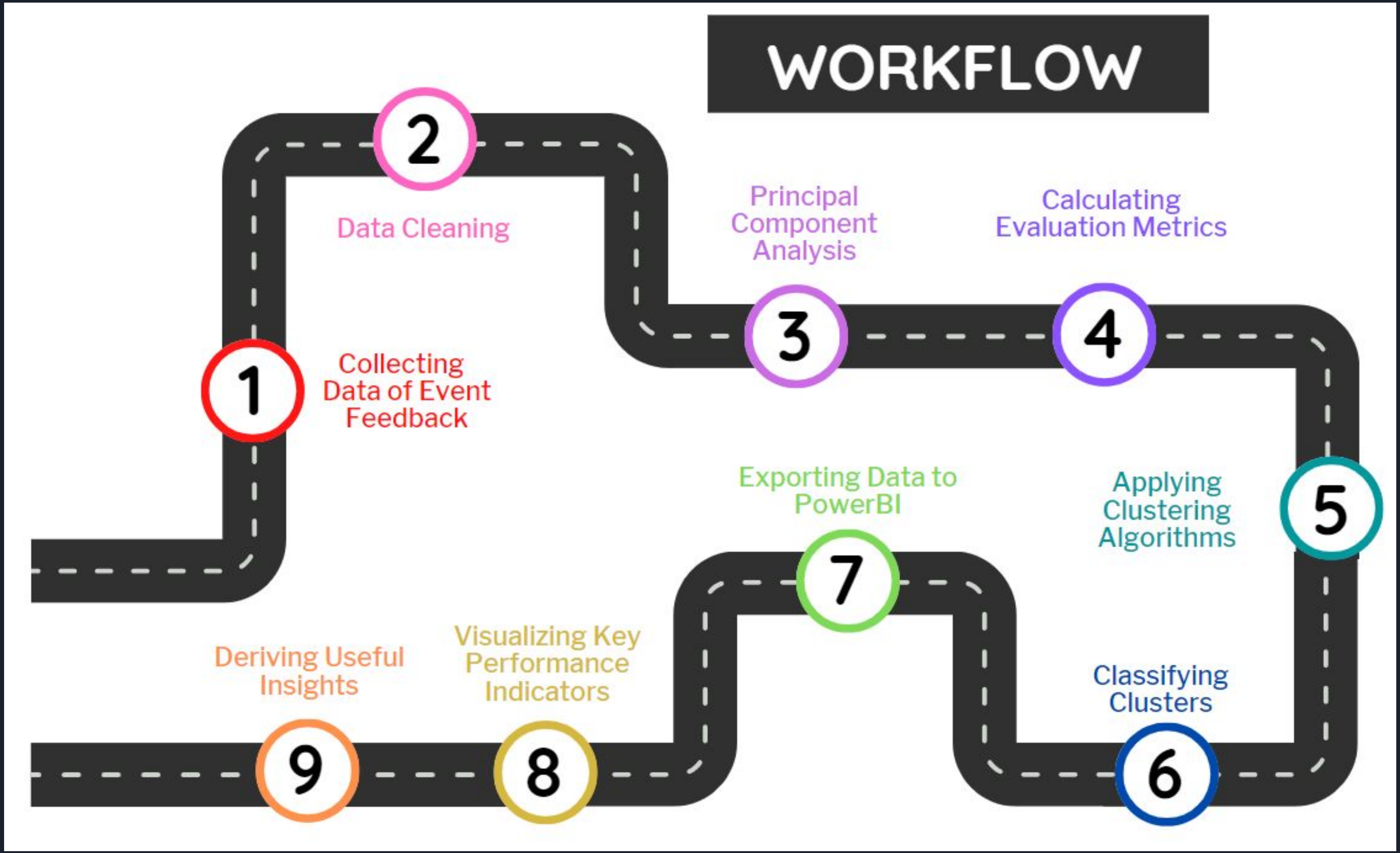
Open Issues

- **Scalability and Performance:** Potential challenges in analyzing large datasets on real-time feedback
- **Dashboard Customization and User Interface:** Limitations regarding the dashboard customization and User interface design.
- **Cluster Interpretation:** Complexity involved in interpreting and labelling the generated clusters

Problem Statement

- Suppose we have the feedback data from most of the attendees
- The event organizer wants to analyze each feedback and group similar users so that it can help the organizer to target important customers.
- Looking at the raw data it is difficult to extract useful insights.

Proposed Architecture

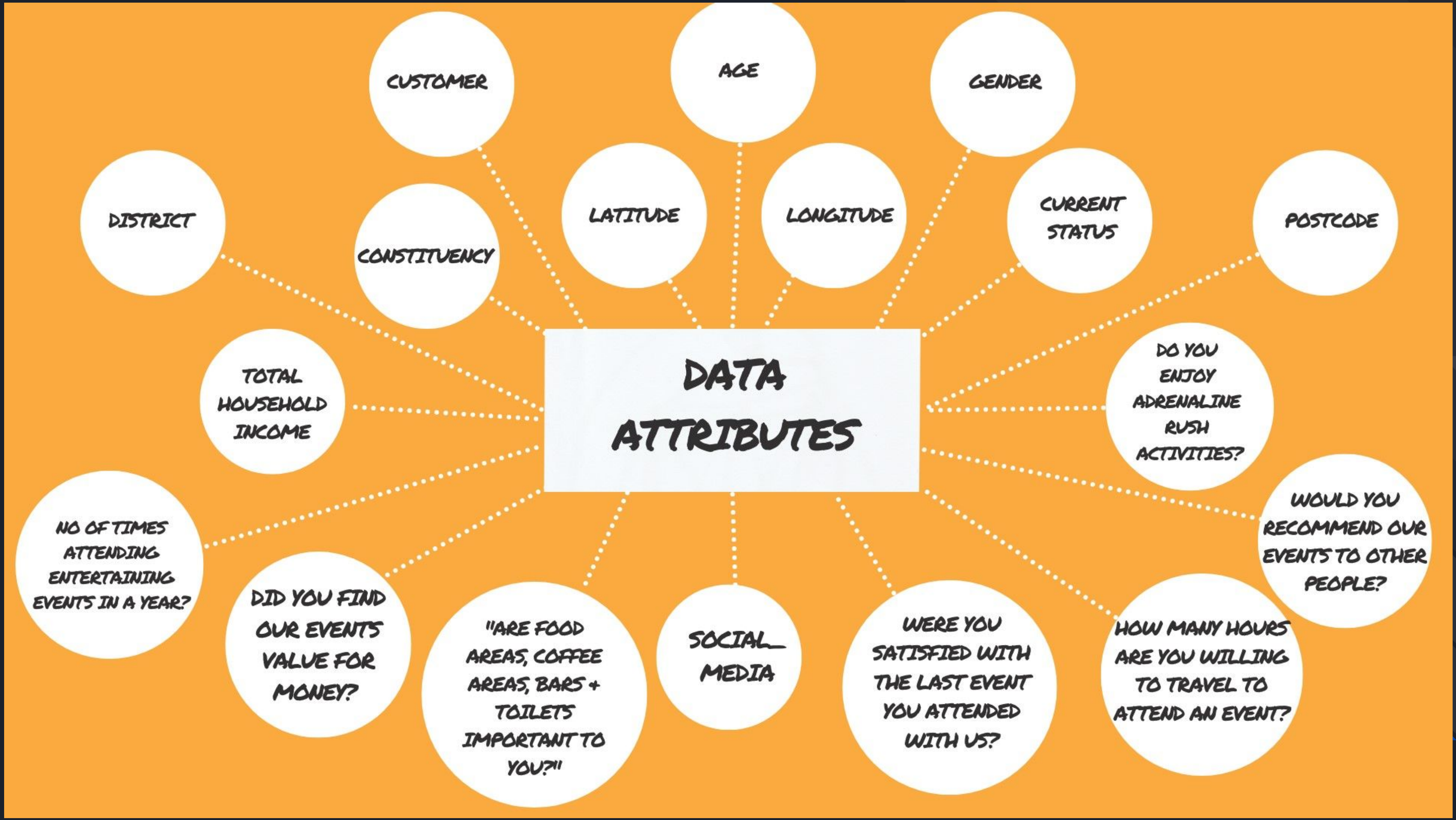


Functional and Non Functional Requirements

Functional	Non Functional
• Data Import and Processing • Machine Learning Model Training • Data Visualization • Model Evaluation and Interpretation • Reporting and Exporting	• Performance(Speed) • Accuracy • User Interface and User interaction • Maintainability • Security and Privacy

Methodology

Event Feedback Data Fields



Methodology

Data Cleaning

Dealing with Null values - Remove the null values or filling the null values with appropriate measure

```
CUSTOMER                                                                0
Age                                                                     0
Gender                                                                  0
Postcode                                                                0
District                                                                0
Constituency                                                            0
latitude                                                                0
longitude                                                               0
Current_Status                                                          0
Total_Household_Income                                                  5
How often you attend Entertaining events in a year?                     5
Social_Media                                                            5
How many hours are you willing to travel to attend an event?            5
Do you enjoy adrenaline-rush activities?                                5
Are food areas, coffee areas, bars & toilets important to you?          5
What is your favourite attraction from below:                           5
Were you satisfied with the last event you attended with us?            5
Would you recommend our events to other people?                         5
Did you find our events value for money?                                5
dtype: int64
```

Methodology

Data Cleaning

Categorical to Numerical Data - There are many categorical variables in our dataset. We convert those into numerical values using One-Hot Encoding technique.

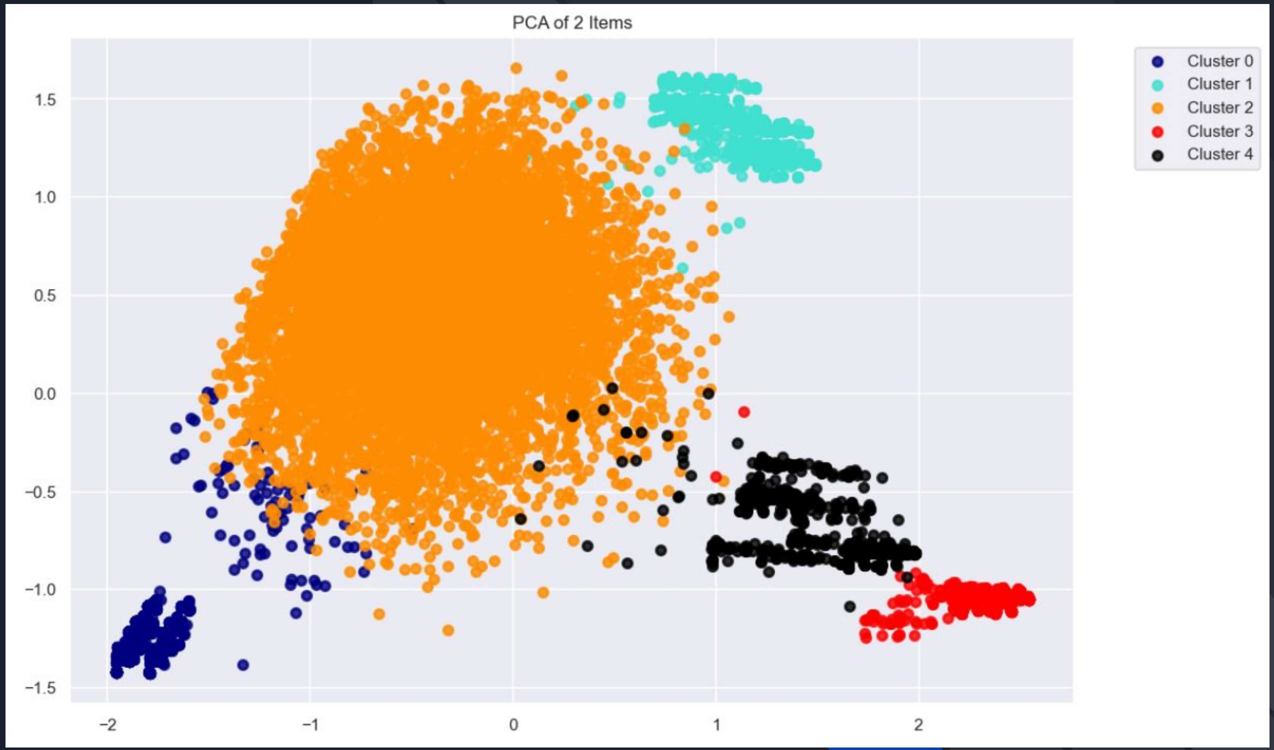
	Age_17 or younger	Age_18-20	Age_21-25	Age_26-32	Age_33-39	Age_40-49	Age_50-59	Age_60-64	Age_65 or older	Gender_Female	...	Would you recommend our events to other people?_Somewhat Unlikely	Would you recommend our events to other people?_Very Likely	Would you recommend our events to other people?_Very Unlikely
0	0	0	0	0	0	1	0	0	0	1	...	1	0	0
1	0	0	0	0	0	0	0	1	0	0	...	0	0	0
2	0	0	0	0	0	0	1	0	0	0	...	0	0	0
3	0	0	0	0	0	0	1	0	0	1	...	0	0	0
4	0	0	0	0	0	0	0	1	0	0	...	0	0	1

Methodology

PCA

There are total 86 columns present in our dataset. We need to reduce the number of features. This can be done using Principal Component Analysis(PCA)

Trying PCA with 2 components we get explained variance as 0.167 and 0.078 respectively.



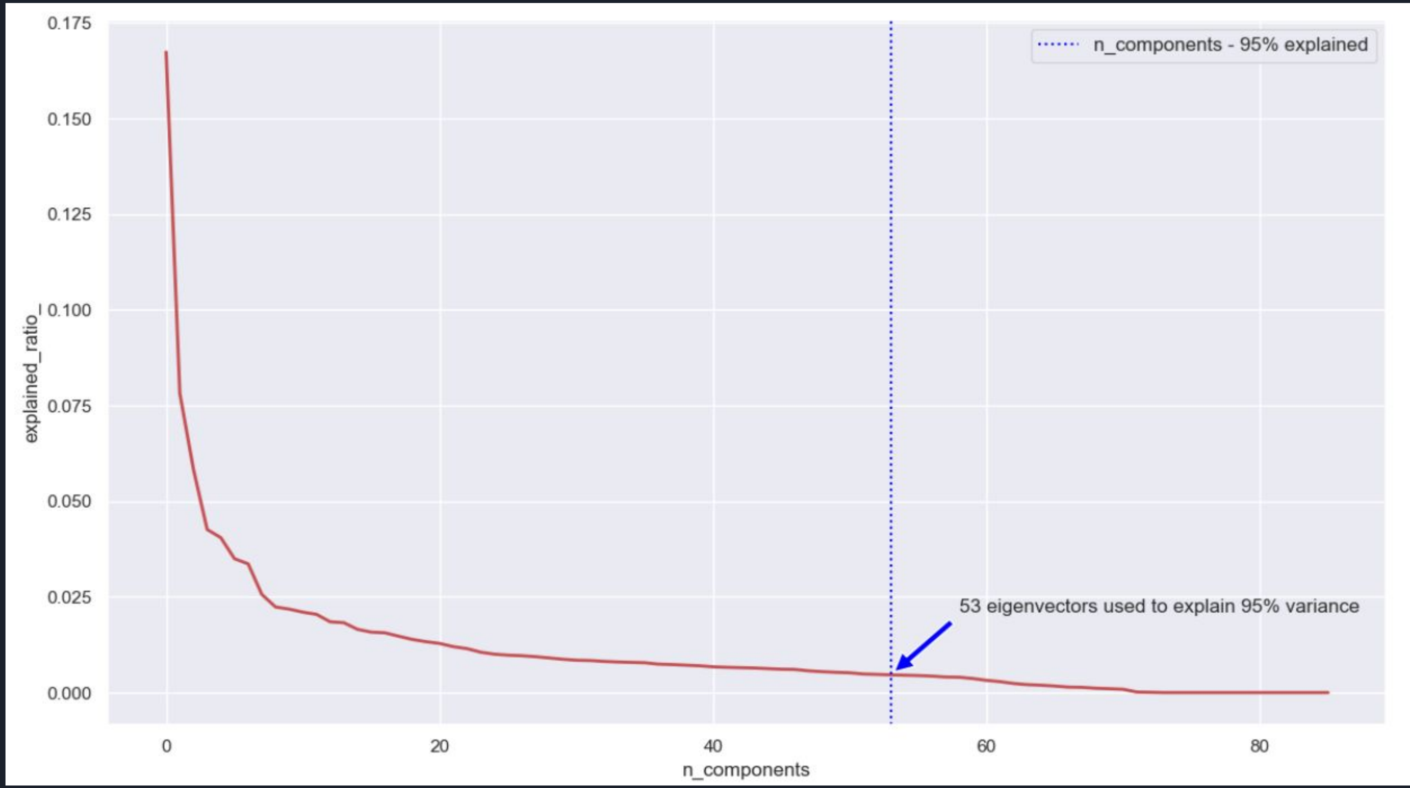
Methodology

PCA

There total variance of our data is 9.78. We want to have 95% of the total explained variance ratio. Checking the explained variance for different number of components

```
Total Variance in our dataset is:  9.789277508428578
The 95% variance we want to have is:  9.29981363300715

Variance explain with 30 n_compononets:  7.800063287617351
Variance explain with 35 n_compononets:  8.200059944222126
Variance explain with 40 n_compononets:  8.559358189926291
Variance explain with 41 n_compononets:  8.625206072416157
Variance explain with 50 n_compononets:  9.1572220656012
Variance explain with 53 n_compononets:  9.301801997586802
Variance explain with 55 n_compononets:  9.39152819059344
Variance explain with 60 n_compononets:  9.592274980903573
```



Methodology

Combining Similar Features

- There are many similar features in our dataset such as
  - attending the events once a year and twice a year are similar
  - attending the events 4 times a year and 5+ times a year are similar

- Combining similar features to reduce dimensionality

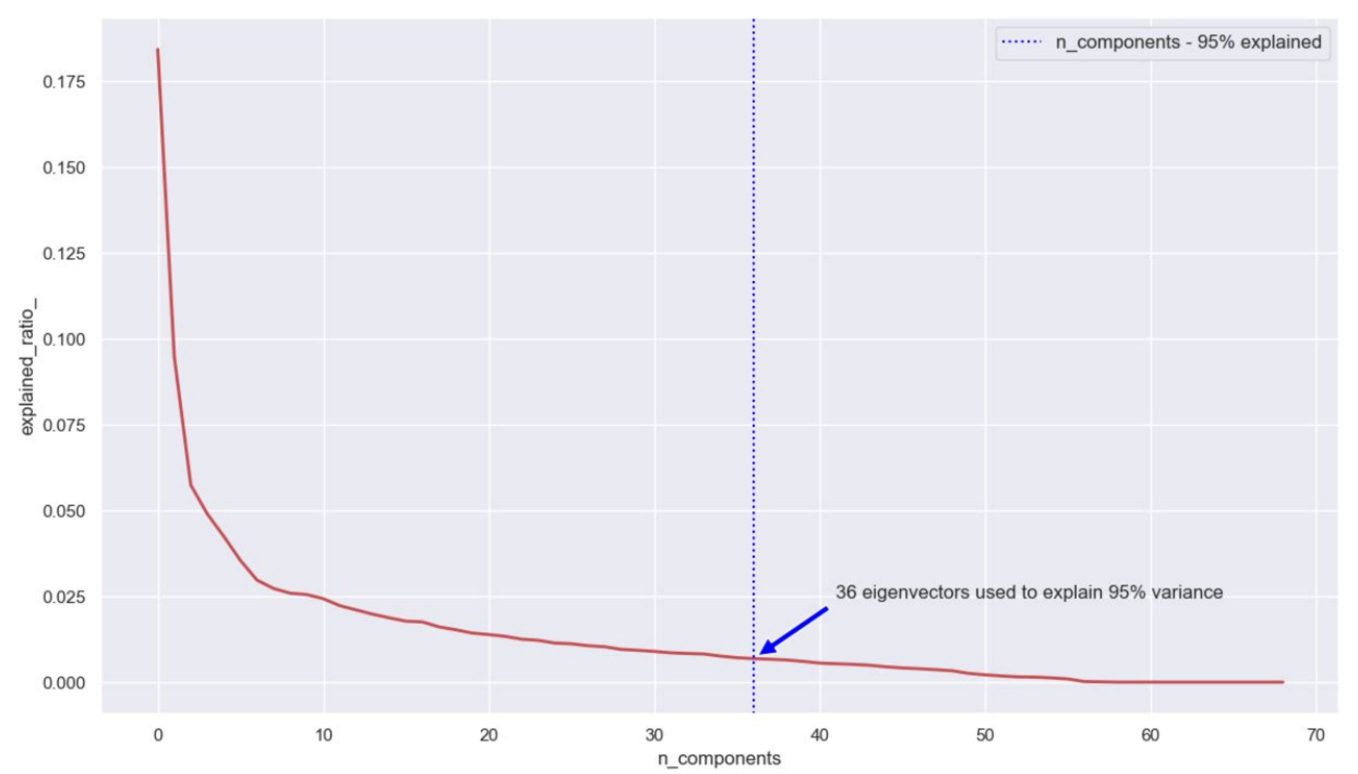
- No loss in expected variance

- Again run PCA to retain 95% of expected variance

Methodology

```
Total Variance in our dataset is:  9.180531774162311
The 95% variance we want to have is:  8.721505185454195

Variance explain with 30 n_compononets:  8.014406502568583
Variance explain with 35 n_compononets:  8.396329813160833
Variance explain with 36 n_compononets:  8.461635629287185
Variance explain with 40 n_compononets:  8.7003816381161
Variance explain with 41 n_compononets:  8.751274630760921
Variance explain with 50 n_compononets:  9.095245791501771
```



Methodology

K-Means

Checking Inertia on different number of clusters before and after PCA

```
The inertia for : 2 Clusters is: 125619.02972065727
The inertia for : 3 Clusters is: 114905.386842667
The inertia for : 4 Clusters is: 106337.17594801627
The inertia for : 5 Clusters is: 100865.16529237546
The inertia for : 6 Clusters is: 96432.53526396505
The inertia for : 7 Clusters is: 93814.4989763171
The inertia for : 8 Clusters is: 91696.57513876252
The inertia for : 9 Clusters is: 89725.00222083348
The inertia for : 10 Clusters is: 88493.22915979216
The inertia for : 11 Clusters is: 87581.06059954726
The inertia for : 12 Clusters is: 86617.6660888009
The inertia for : 13 Clusters is: 85829.38420440158
The inertia for : 14 Clusters is: 85014.85271668163
The inertia for : 15 Clusters is: 84434.74381493333
The inertia for : 16 Clusters is: 83662.83564950572
The inertia for : 17 Clusters is: 82854.33711923643
The inertia for : 18 Clusters is: 82485.74994726645
The inertia for : 19 Clusters is: 82187.9337203959
```

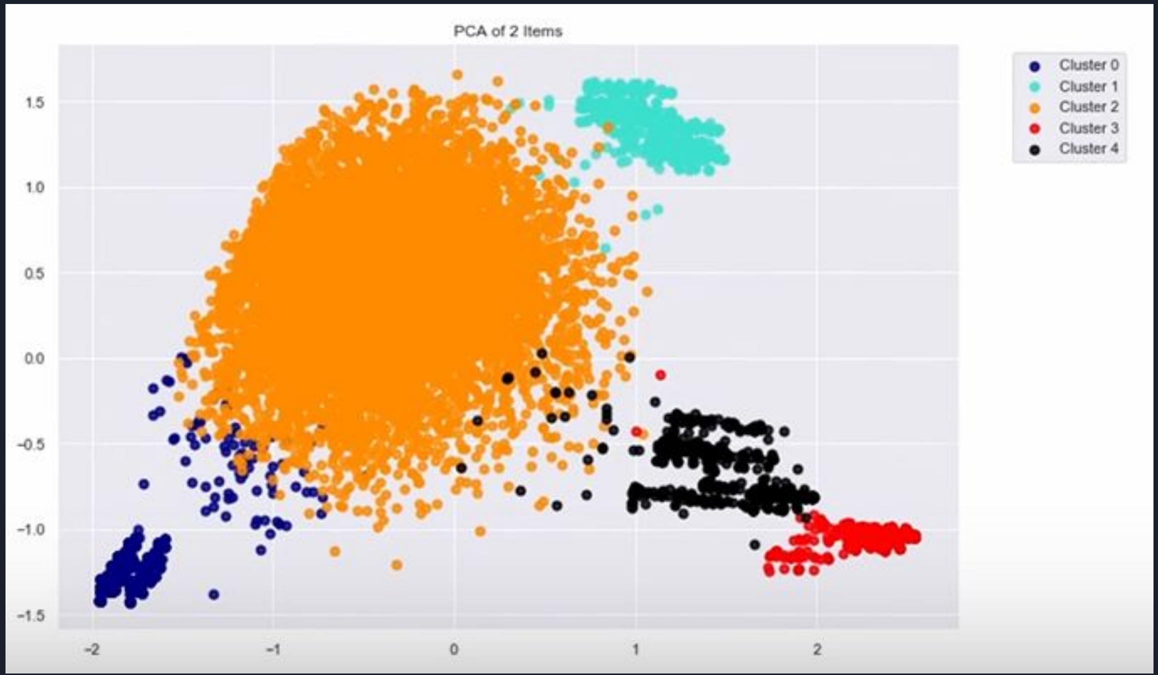
Before PCA

```
The inertia for : 2 Clusters is: 105238.4329946018
The inertia for : 3 Clusters is: 92911.4603053280
The inertia for : 4 Clusters is: 85693.7047277193
The inertia for : 5 Clusters is: 80703.3889172904
The inertia for : 6 Clusters is: 78454.8178005808
The inertia for : 7 Clusters is: 76375.0791656580
The inertia for : 8 Clusters is: 74776.1237836916
The inertia for : 9 Clusters is: 72886.3033868518
The inertia for : 10 Clusters is: 71630.15404372885
The inertia for : 11 Clusters is: 70619.78080730671
The inertia for : 12 Clusters is: 69346.65797379437
The inertia for : 13 Clusters is: 68735.61953260798
The inertia for : 14 Clusters is: 67708.56865301062
The inertia for : 15 Clusters is: 66931.86574392965
The inertia for : 16 Clusters is: 66238.52071892508
The inertia for : 17 Clusters is: 65647.88198639416
The inertia for : 18 Clusters is: 65232.47213714429
The inertia for : 19 Clusters is: 64482.44337332091
The inertia for : 20 Clusters is: 64072.60290522323
```

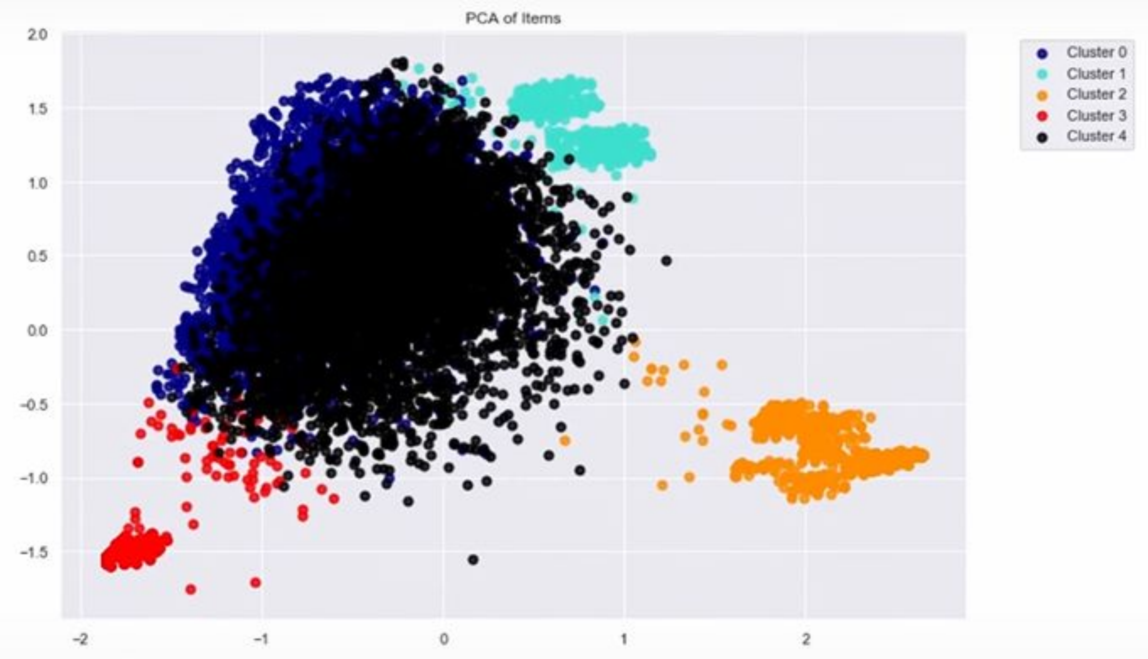
After PCA

Methodology

Clusters using K-means clustering unsupervised ML algorithm



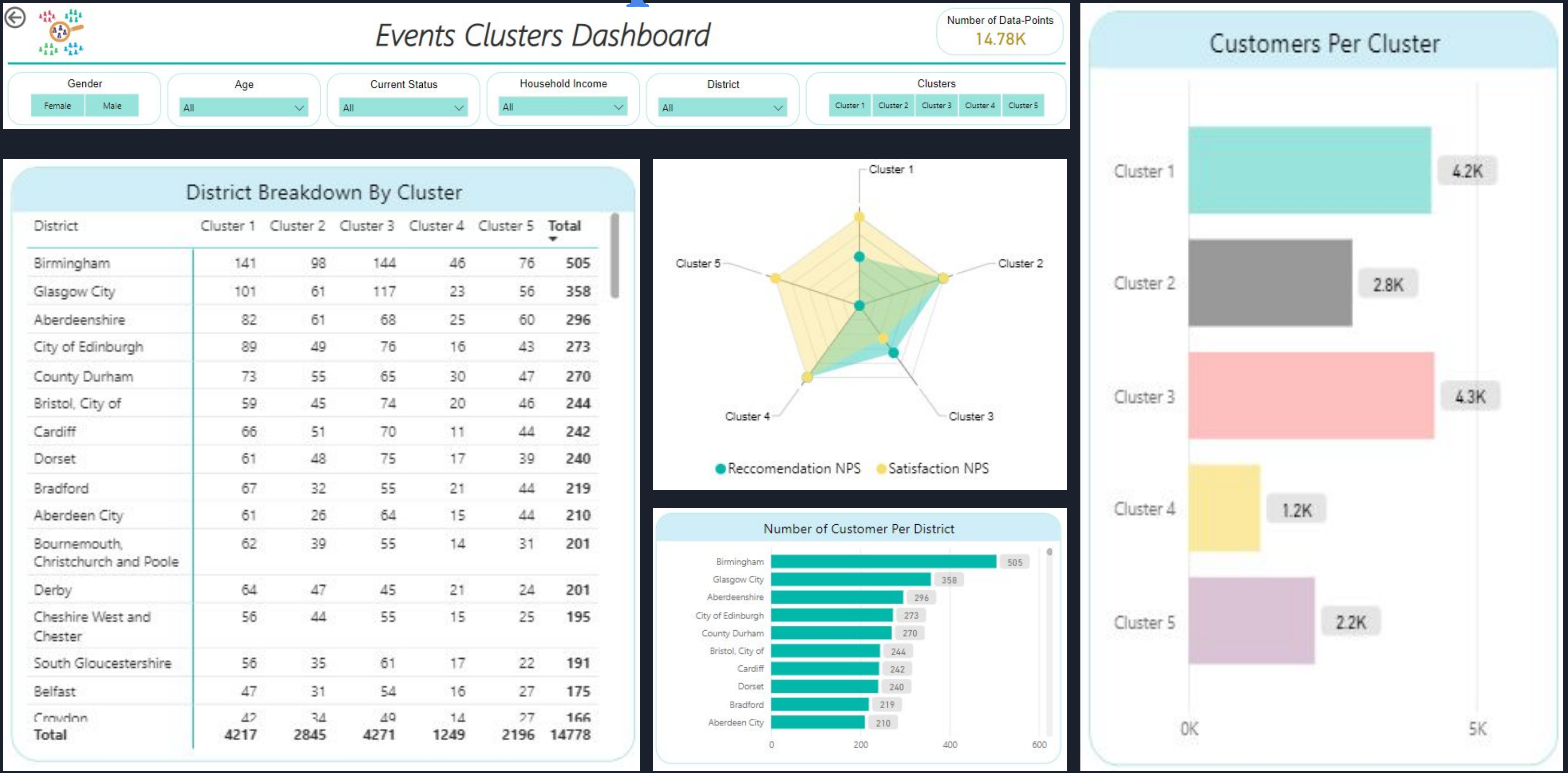
Before PCA



After PCA

Implementation

Implementation



Testing and Validation

Comparison of Agglomerative, DBScan and K means using Silhouette Score

```
Silhouette Score for 3 Clusters (Agglomerative) : 0.5853871181788206
Silhouette Score for 4 Clusters (Agglomerative) : 0.5329035998127906
Silhouette Score for 5 Clusters (Agglomerative) : 0.5316711894131084
Silhouette Score for 6 Clusters (Agglomerative) : 0.5361967616584268
```

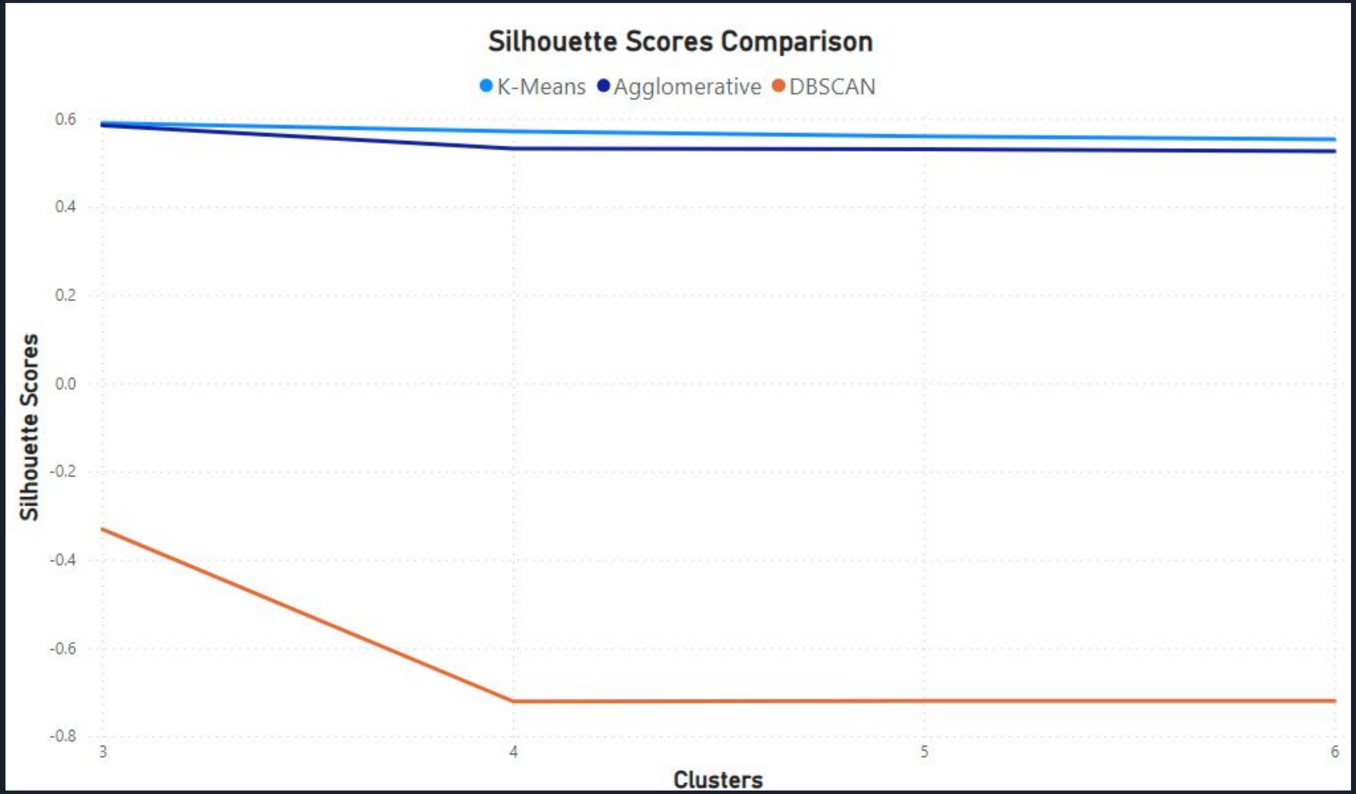
Silhouette Score for Agglomerative Clustering

```
Silhouette Score for 3 Clusters (DBSCAN) : -0.33173096609437486
Silhouette Score for 4 Clusters (DBSCAN) : -0.7210413243280295
Silhouette Score for 5 Clusters (DBSCAN) : -0.7208468069375299
Silhouette Score for 6 Clusters (DBSCAN) : -0.7206349572000377
```

Silhouette Score for DBSCAN Clustering

```
Silhouette Score for 3 Clusters (K-Means) : 0.5901082892403513
Silhouette Score for 4 Clusters (K-Means) : 0.5718013573550145
Silhouette Score for 5 Clusters (K-Means) : 0.5607762864253566
Silhouette Score for 6 Clusters (K-Means) : 0.5534035001687794
```

Silhouette Score for K-Means Clustering



By This, We can conclude that K-Means is better than DBScan and Agglomerative

Testing and Validation

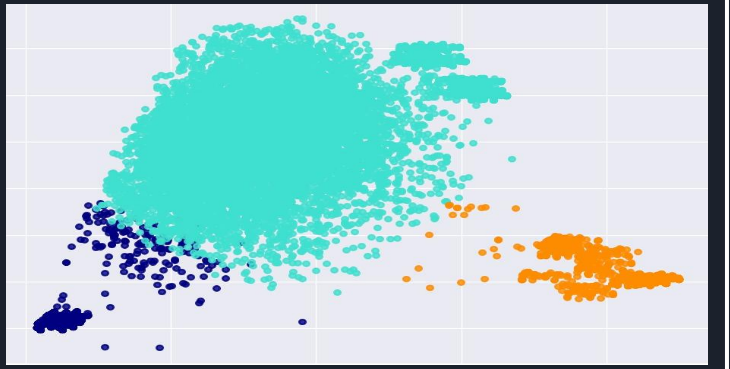
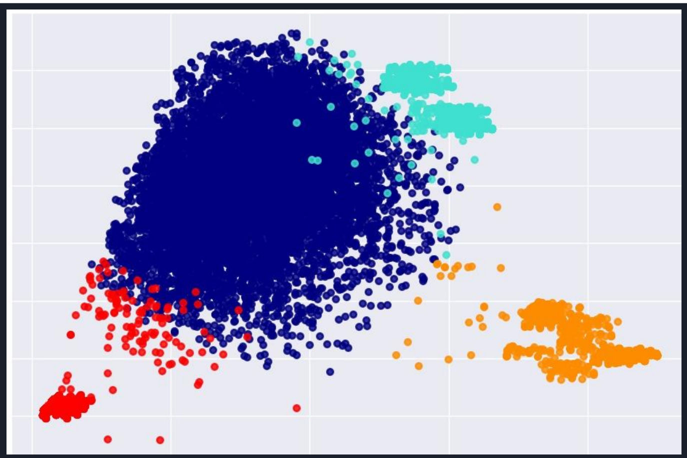
Advantages of K-means over Hierarchical Clustering:

- Efficiency
- Scalability
- Flexibility

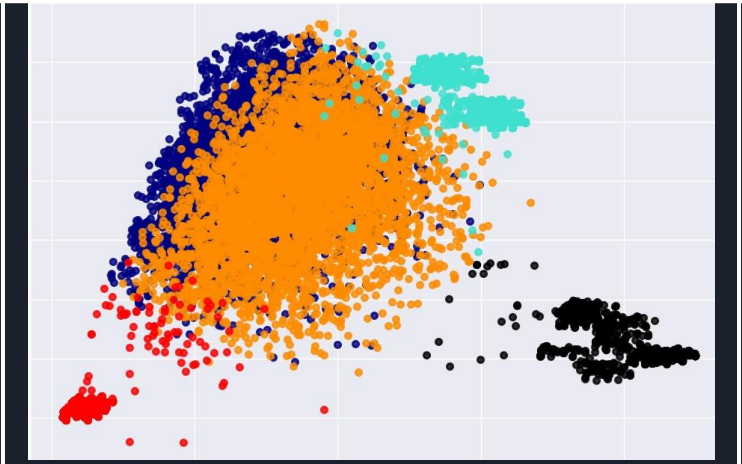
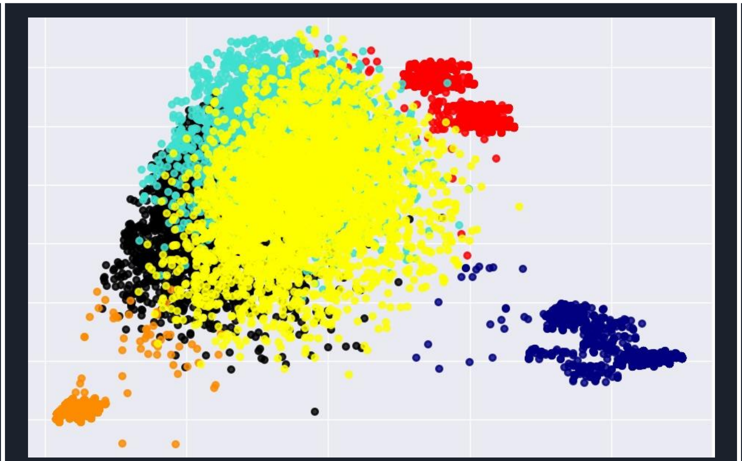
Advantages of K-means over DBSCAN:

- Simplicity
- Handling different density clusters
- Outlier detection

Testing and Validation

Number Of Clusters	Visualization	Inertia	Silhouette Score
3		92911.46	0.5901
4		85693.70	0.5718

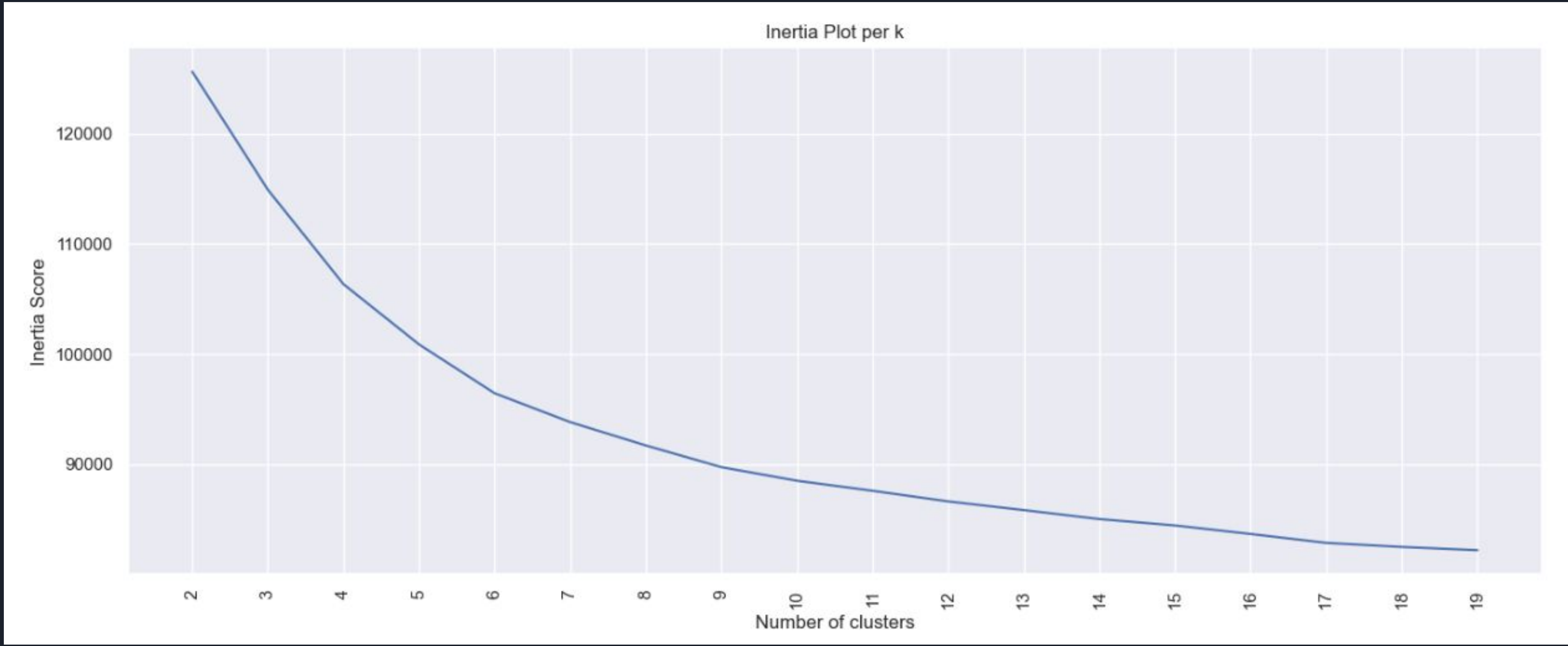
Testing and Validation

Number Of Clusters	Visualization	Inertia	Silhouette Score
5		80703.38	0.5607
6		78454.81	0.5534

Testing and Validation

K-Means

Using elbow method we decide the number of clusters



Results and Discussions

Cluster 1 Traits

- Mostly people with age being 50+
- Mostly Married with Children
- Household Income ranges from 25k to 100k
- Attend Events 3 to 4 times a year
- Don't spend too much time on Social Media (< 1 hour)
- Willing to travel 4 - 6 hours
- Kids Playgrounds is their favourite attraction
- Very satisfied with last event

Cluster 2 Traits

- People who don't have kids - mostly single
- Earn between 20k to 50k
- Attend events mostly once or twice a year
- Spend a lot of time in Social Media; half a day +
- Willing to travel 1 to 2 hours
- Love adrenaline rush activities
- Not bothered with food/coffee/bars/toilet areas
- Somewhat satisfied with last event
- Somewhat likely to recommend it to others
- Event was not value for money

Cluster 3 Traits

- Mostly people with age range between 26 to 50
- Married people who have kids or living with their partners (2+)
- Earn between 50k to 150k
- Attend events 3 to 4 times a year
- Spend mostly 1 to 2 hours in social media
- Mostly willing to travel 4 to 6 hours
- Like a bit of everything in the attractions
- Very likely to recommend their last event
- Very "general" group of people; maybe willing to try new things

Cluster 4 Traits

- People who single, separated, divorced or widowed
- Household income ranges between 50k to 100k or less than 20k
- Attend a lot of events per year; 5 plus
- Spend half a day in social media
- Willing to travel up to 6 hours and they love adrenaline rush activities
- Not bothered with food/bars/coffee/toilet areas
- Mostly satisfied with their last event and willing to recommend
- They do not think the last event was value for money

Cluster 5 Traits

- Mostly people between 40 to 60 age
- Married with children
- High earners; making 100k +
- Attend events 3 times a year
- Do not spend much time on social media; 1 hour or less
- Willing to travel 4-6 hours for the event
- Not adrenaline people
- Food/Coffee/bars/toilets are very importance
- Kids playgrounds are essential
- Very satisfied with last event BUT* Unlikely to recommend (dummy data)
- Last event was value for money

Conclusion

The integration of Excel, Machine Learning (ML), and Power BI offers a powerful suite of tools for data storing, analysis and visualization.

- **Machine Learning (ML)**

  K-means clustering

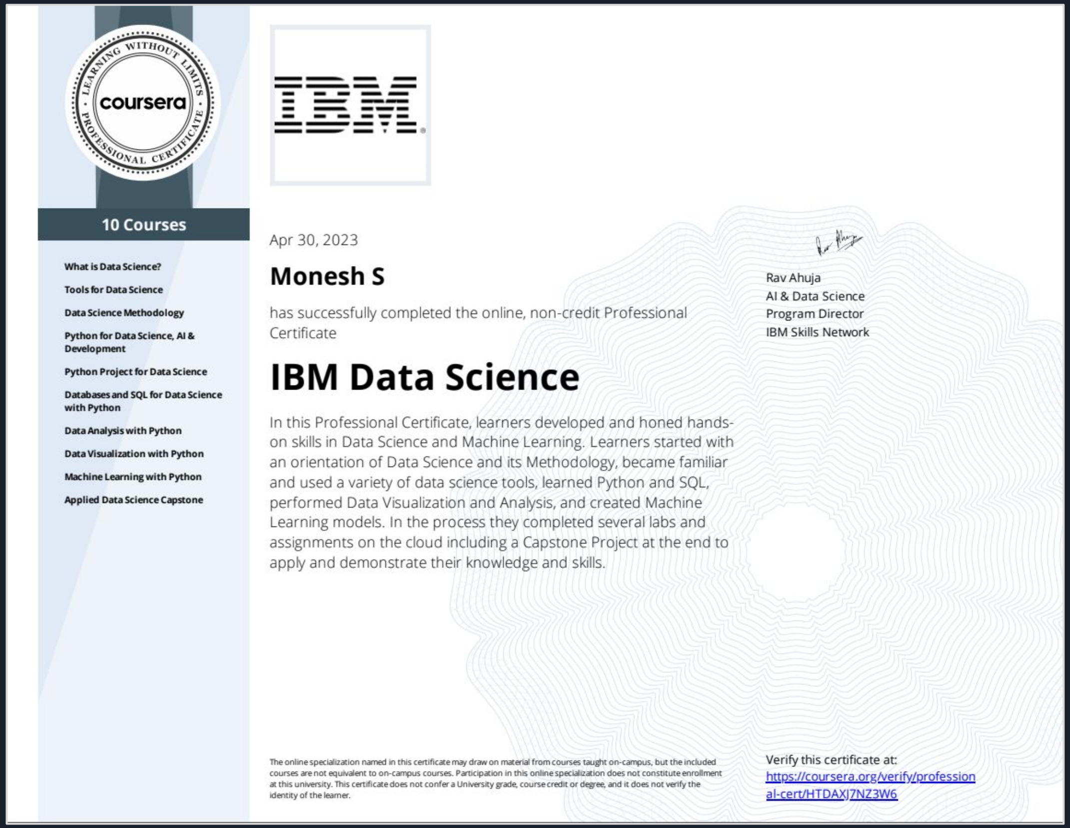
  Principal Component Analysis(PCA)

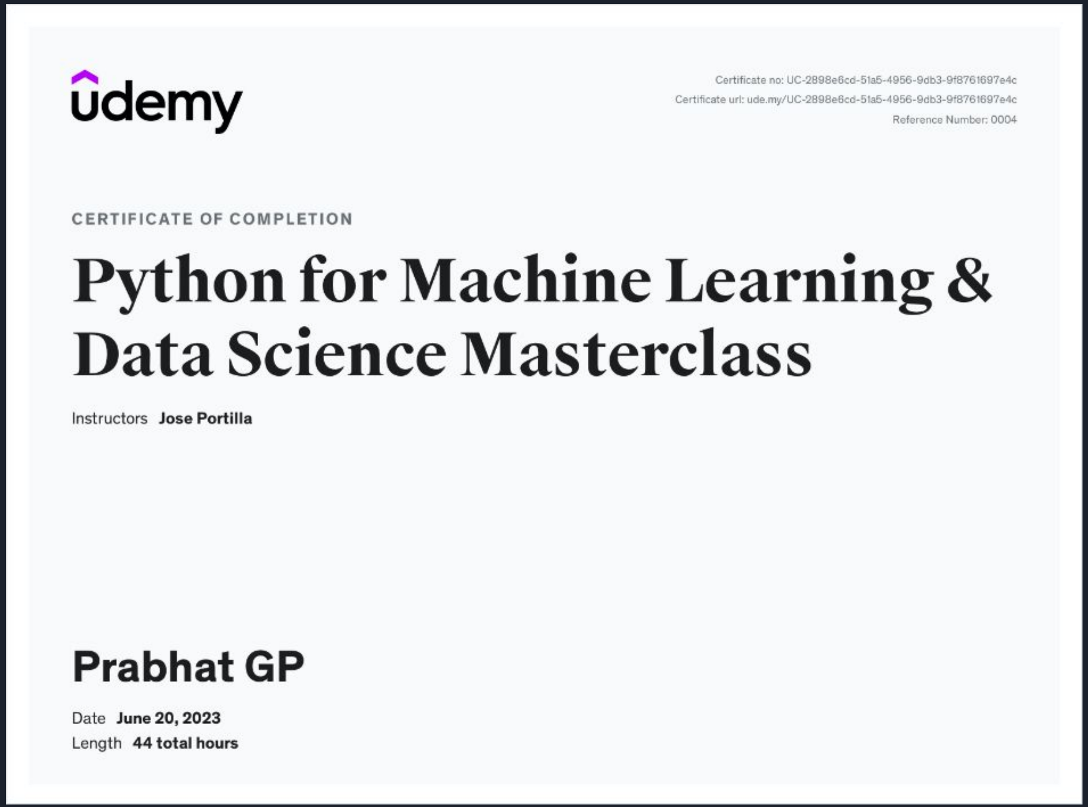
- **Power BI**

  Visually appealing plots

  Insightful Graphs

Certificate of MOOC

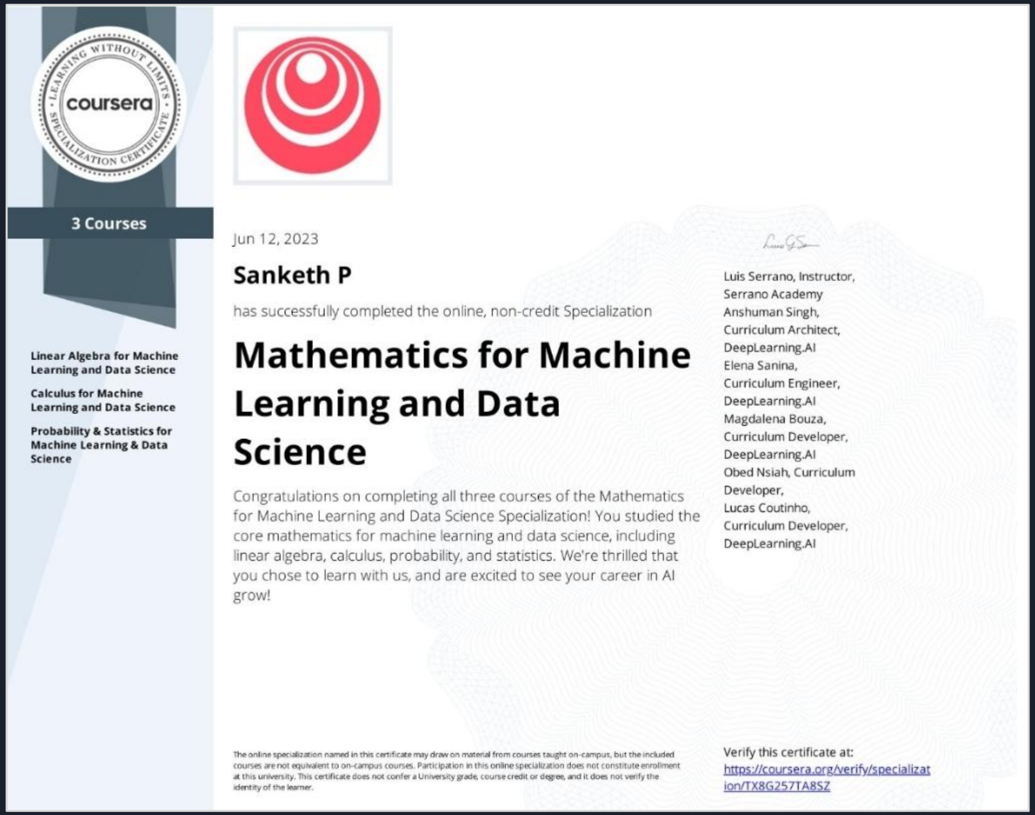




Monesh S
IBM Data Science Specialization
3 Months
Coursera IBM Data Science

Prabhat G P
Python for ML and Data Science
44 Hours
Python for Machine Learning & Data Science

About MOOC

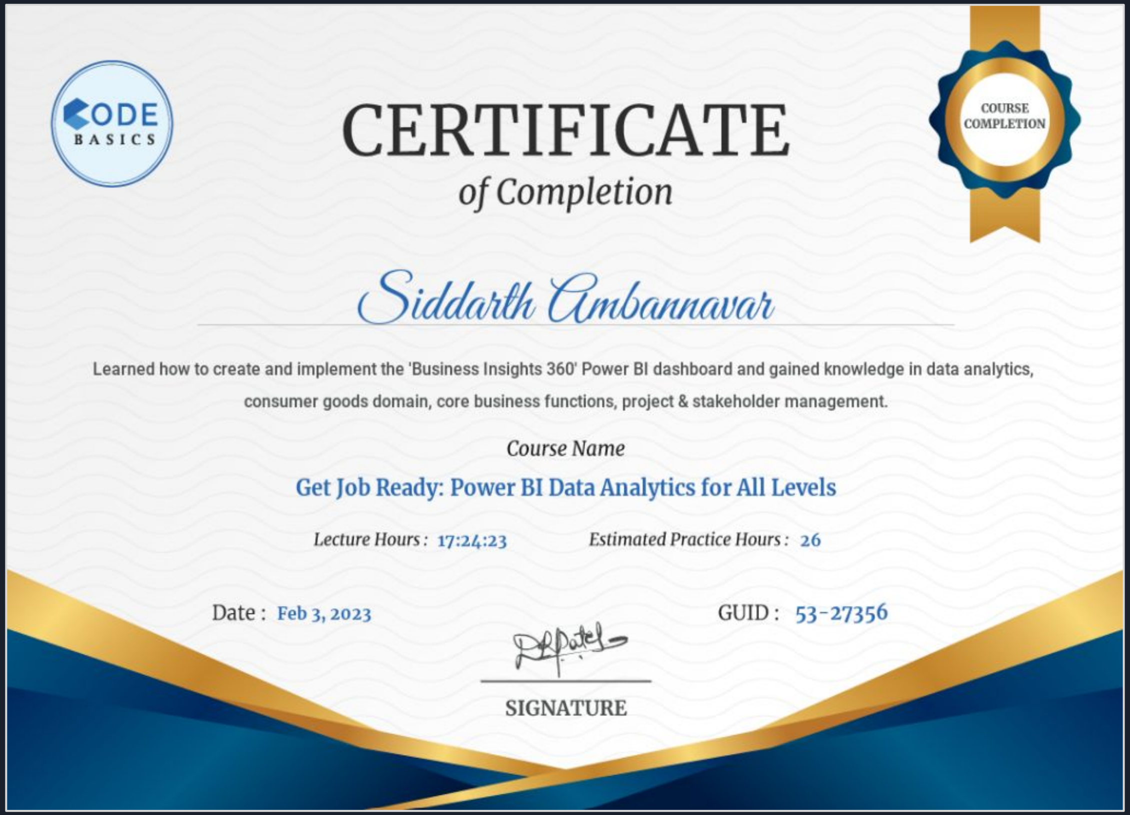


Sanketh P

Mathematics for ML and Data Science Specialization

3 Months

Coursera Maths for ML and Data Science



Siddarth A

CodeBasics Power BI

17 Hours

Power BI Data Analytics

Suggestions / Questions
Please ...

Thank you !