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

Phase - 1 Presentation

Date: 12th June, 2023

Presented By,

Student Name & USN :

MONESH S	1BM20AI039
PRABHAT G P	1BM20AI043
SANKETH P	1BM20AI048
SIDDARTH A	1BM20AI049

Semester & Section: **6A**
Batch Number:5

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 Methodology
- Functional & Non-Functional Requirements
- Expected Outcome
- 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

AUTHOR / TITLE / YEAR	APPLIED METHODOLOGY / ALGORITHM USED	FINDINGS	RESULTS	LIMITATIONS
"Analyzing Event Feedback Data Using Sentiment Analysis and Clustering Techniques" by A. Smith and B. Johnson.	Using SVM with Linear Kernel	Sentiment Analysis	Clusters with optimal inertia	More number of features
"Clustering Event Attendees based on Feedback and Social Media Data" by C. Lee and D. Kim	Crowd Characterization	Influencing pedestrian behavior in Social media	Social Media Proxy	Privacy Concerns
“Event Feedback Analysis: A Machine Learning Approach for Understanding Attendee Preferences”	Segmentation and Clustering	Clustering for large datasets	Combining large number of columns without losing much variance	Majority of the dataset doesn’t have numerical values

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.

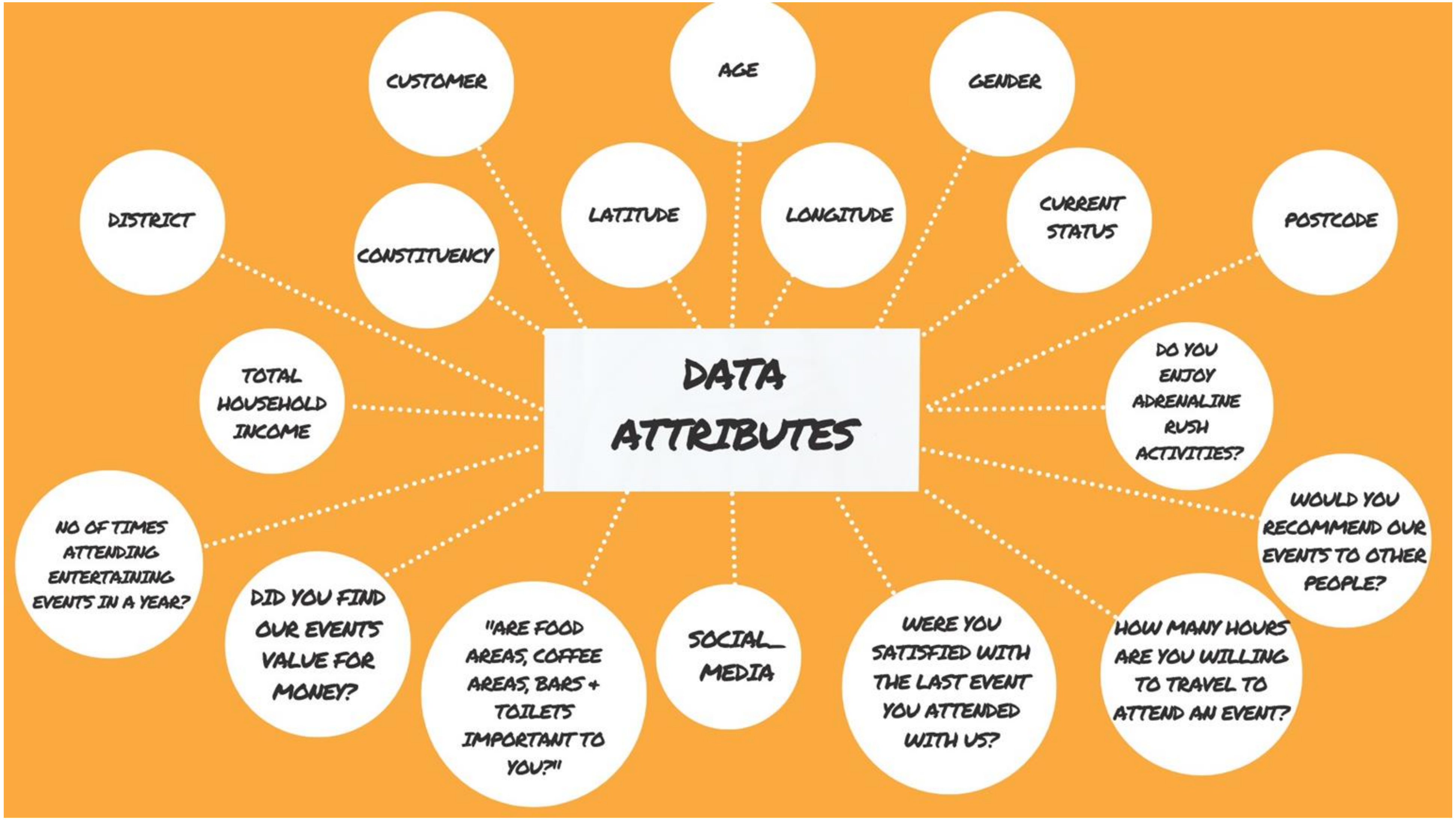
- Looking at the raw data it is difficult to extract useful insights.

Proposed Methodology

Workflow

1. Collected the raw data
2. Cleaned the data by **removing null values**, converting the **categorical data to numerical data**, adding dummies
3. Calculating the **inertia** for different number of clusters
4. Reducing the features using **PCA**
5. Again running **k-means** for different number of clusters to compare the inertia
6. Exporting the data with cluster variables to Power BI
7. Extracting useful **Key Performance Indicators** in **Power BI** using DAX
8. Creating a detailed report in Power BI to better understand the different classes of audience

Proposed Methodology

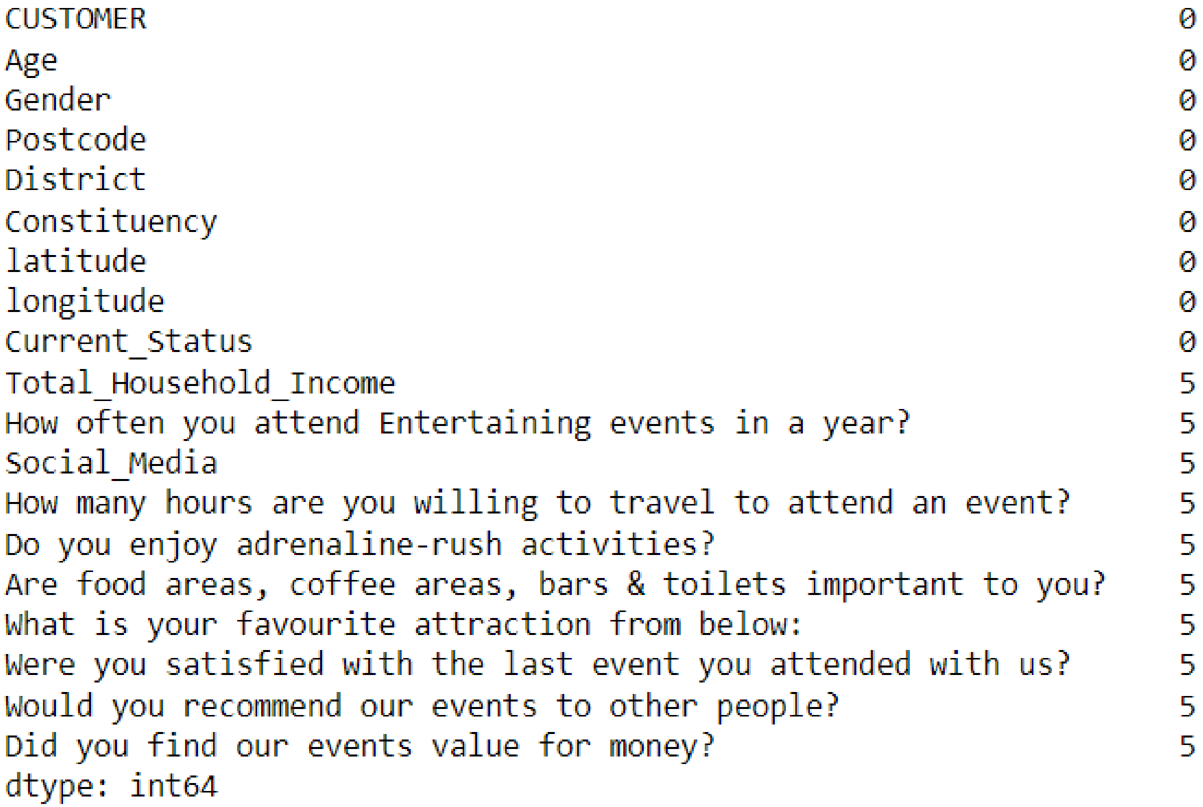


DATA ATTRIBUTES
CUSTOMER
AGE
GENDER
DISTRICT
CONSTITUENCY
LATITUDE
LONGITUDE
CURRENT STATUS
POSTCODE
TOTAL HOUSEHOLD INCOME
DO YOU ENJOY ADRENALINE RUSH ACTIVITIES?
NO OF TIMES ATTENDING ENTERTAINING EVENTS IN A YEAR?
WOULD YOU RECOMMEND OUR EVENTS TO OTHER PEOPLE?
DID YOU FIND OUR EVENTS VALUE FOR MONEY?
"ARE FOOD AREAS, COFFEE AREAS, BARS + TOILETS IMPORTANT TO YOU?"
SOCIAL MEDIA
WERE YOU SATISFIED WITH THE LAST EVENT YOU ATTENDED WITH US?
HOW MANY HOURS ARE YOU WILLING TO TRAVEL TO ATTEND AN EVENT?

Proposed Methodology

Data Cleaning

Dealing with Null values - We remove these rows which has 5 null value rows. As the number of rows are much greater than these missing value rows it is feasible to remove the entire null value rows. Otherwise we need to fill them using mean or any other statical parameters.

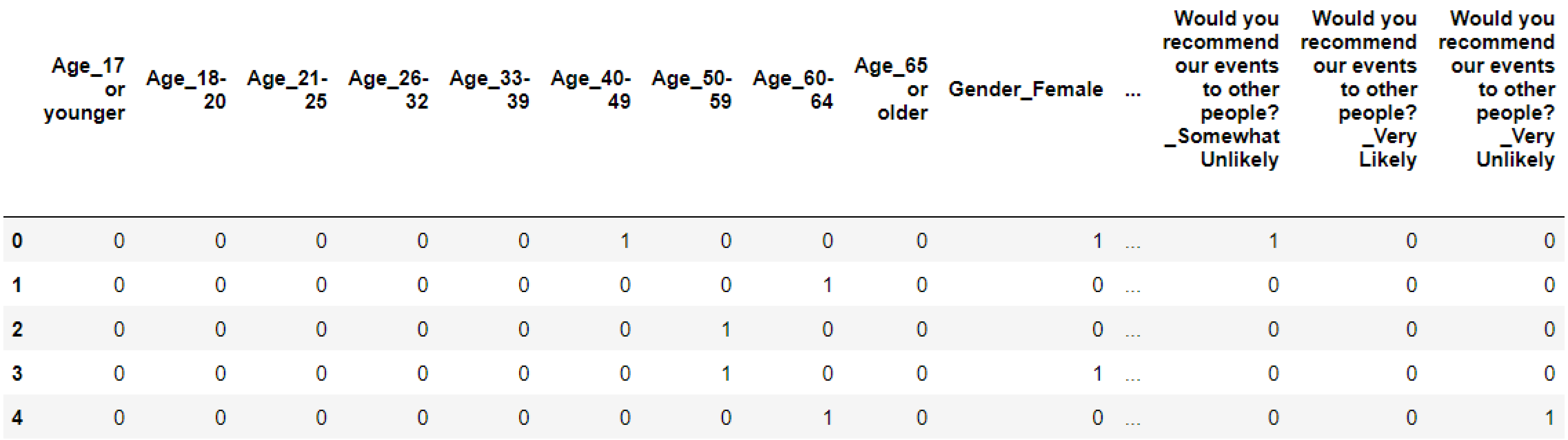


```
CUSTOMER                                                          0
Age                                                               0
Gender                                                            0
Postcode                                                          0
District                                                          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
```

Proposed Methodology

Data Cleaning

Categorical to Numerical Data - There are many categorical variables in our dataset. We convert those into numerical values using the **get_dummies()** function of **pandas** which uses **One-Hot Encoding** technique.

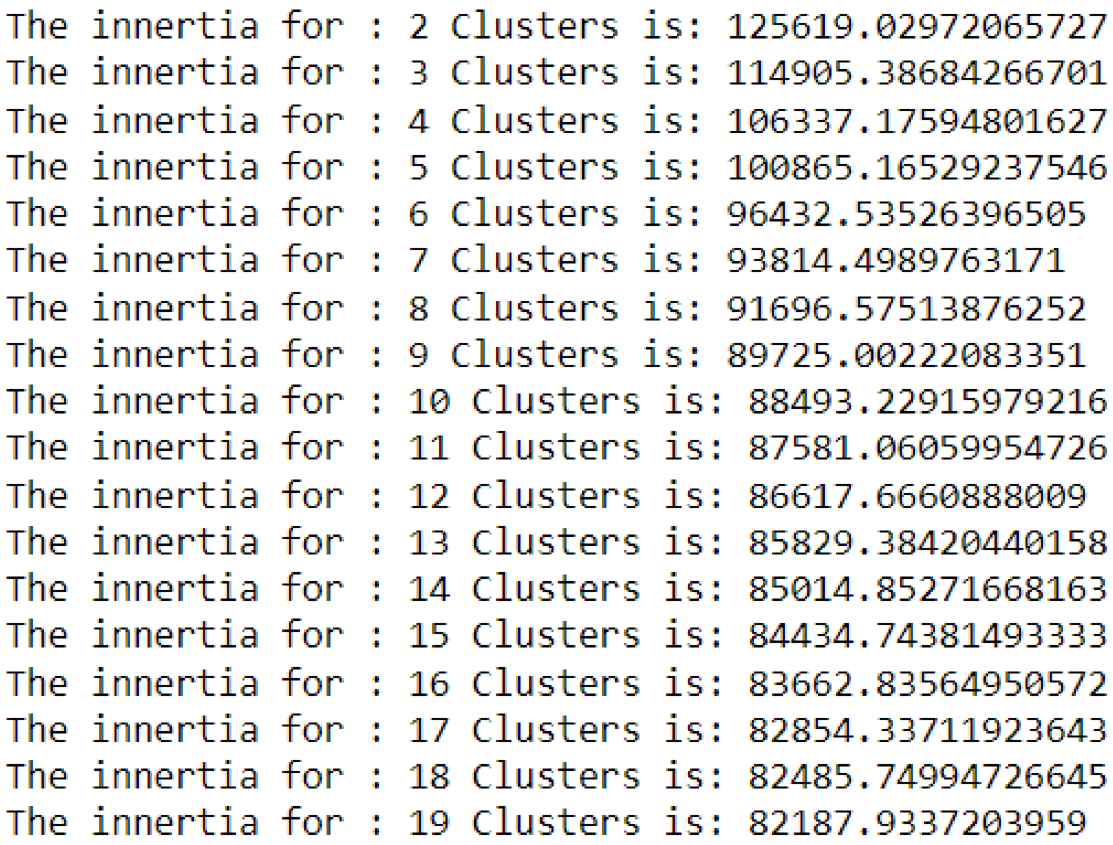


	Age_17 or younger	Age_18-20	Age_21-25	Age_26-32	Age_33-39	Age_40-49	Age_50-59	Age_60-64	Age_65 or older	Gender_Female	...	Would you recommend our events to other people?_Somewhat Unlikely	Would you recommend our events to other people?_Very Likely	Would you recommend our events to other people?_Very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

Proposed Methodology

K-Means

Running K-means clustering algorithm using scikit-learn api from 2 clusters to 19 clusters and check their inertia

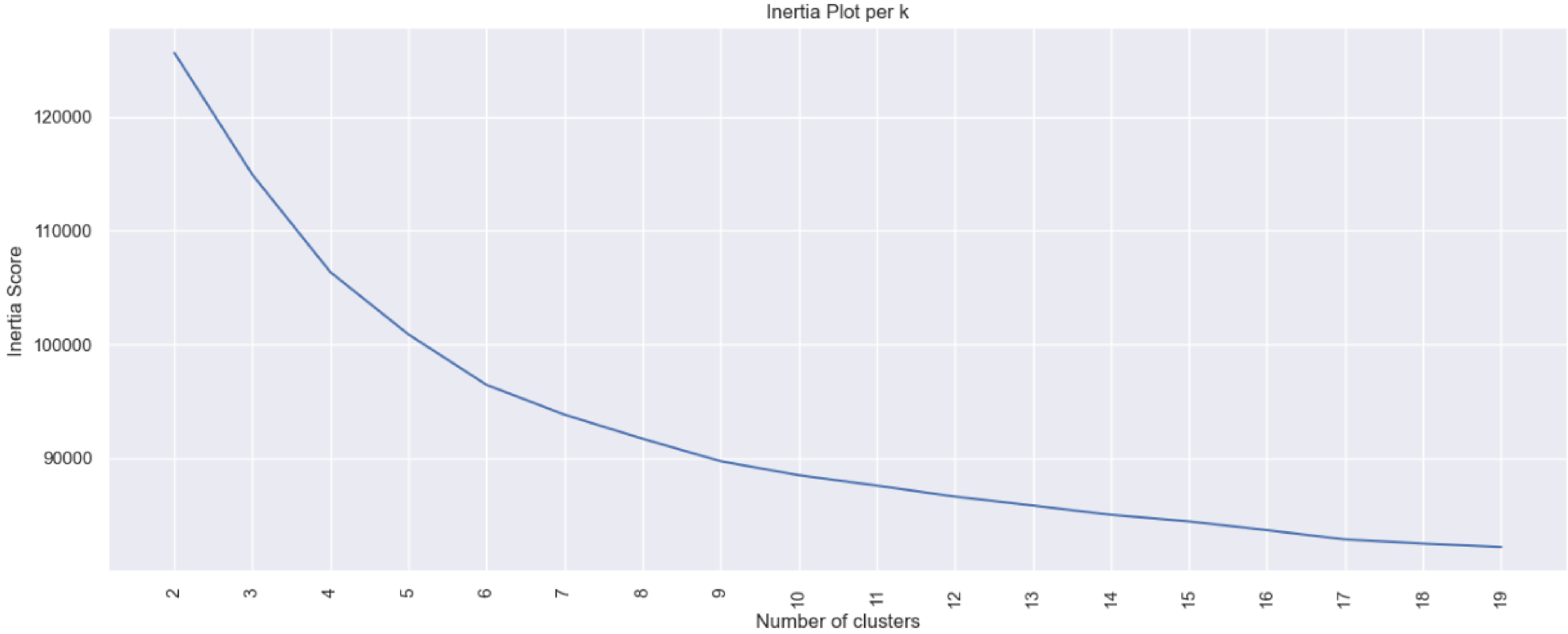


```
The innertia for : 2 Clusters is: 125619.02972065727
The innertia for : 3 Clusters is: 114905.38684266701
The innertia for : 4 Clusters is: 106337.17594801627
The innertia for : 5 Clusters is: 100865.16529237546
The innertia for : 6 Clusters is: 96432.53526396505
The innertia for : 7 Clusters is: 93814.4989763171
The innertia for : 8 Clusters is: 91696.57513876252
The innertia for : 9 Clusters is: 89725.00222083351
The innertia for : 10 Clusters is: 88493.22915979216
The innertia for : 11 Clusters is: 87581.06059954726
The innertia for : 12 Clusters is: 86617.6660888009
The innertia for : 13 Clusters is: 85829.38420440158
The innertia for : 14 Clusters is: 85014.85271668163
The innertia for : 15 Clusters is: 84434.74381493333
The innertia for : 16 Clusters is: 83662.83564950572
The innertia for : 17 Clusters is: 82854.33711923643
The innertia for : 18 Clusters is: 82485.74994726645
The innertia for : 19 Clusters is: 82187.9337203959
```

Proposed Methodology

K-Means

Using elbow method we decide the number of clusters

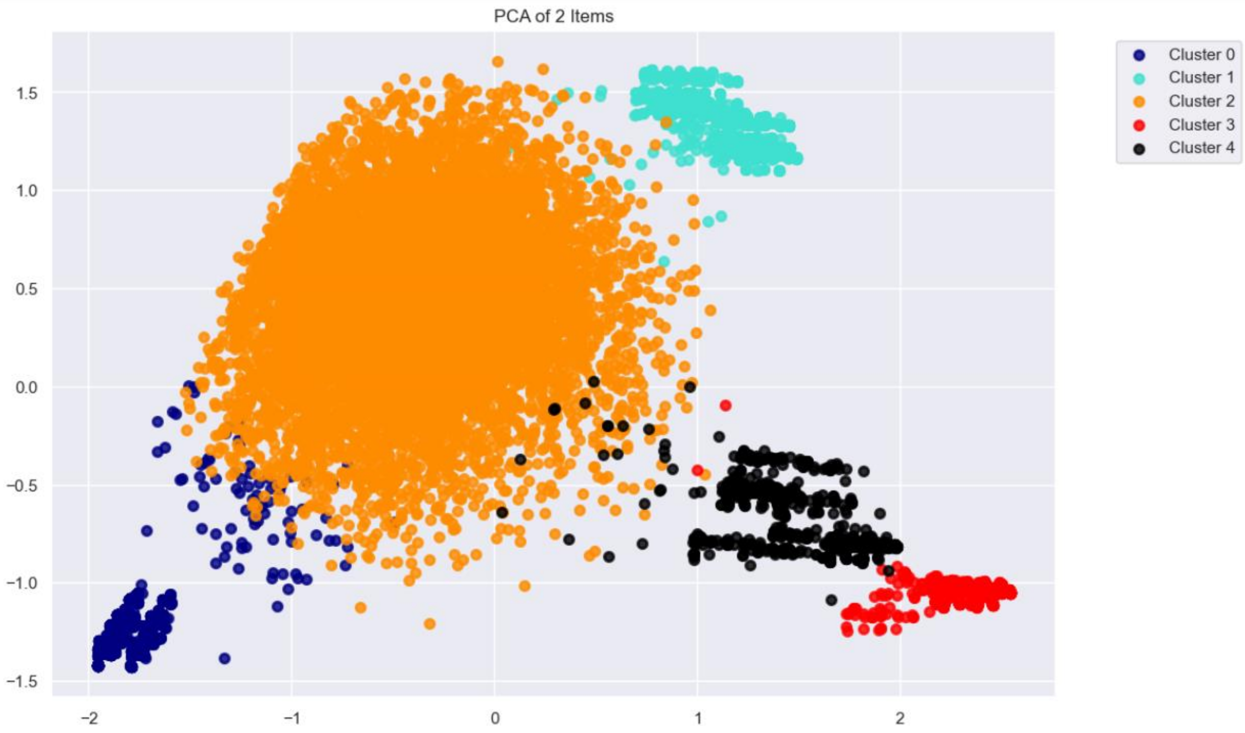


Proposed Methodology

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)

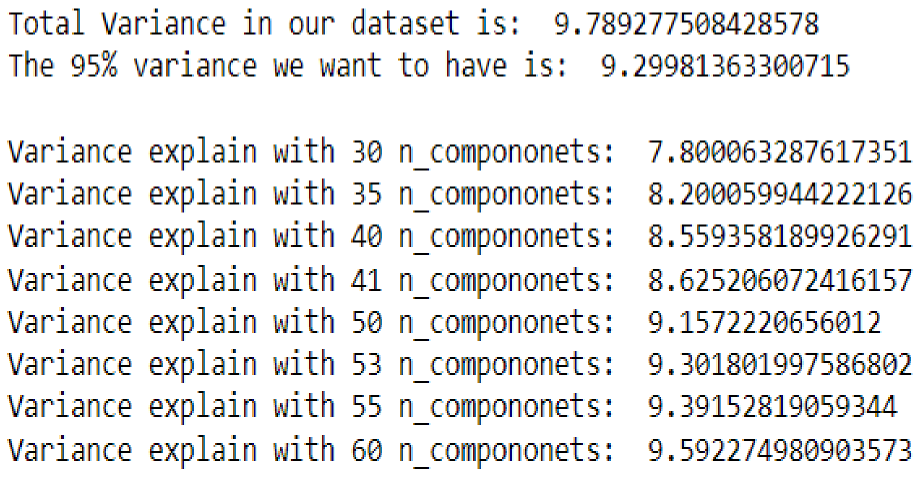
Trying PCA with 2 components we get explained variance as 0.167 and 0.078 respectively.



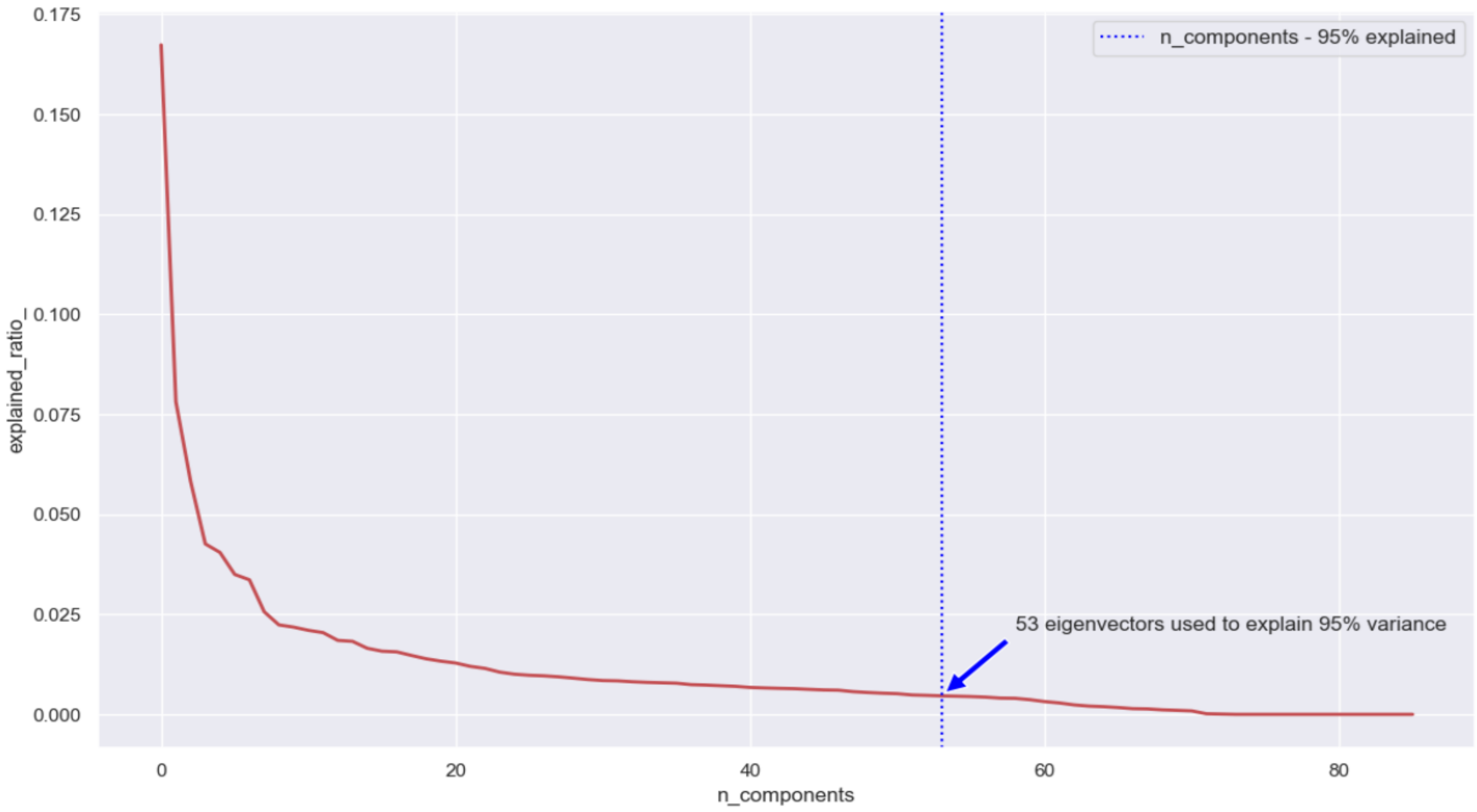
Proposed Methodology

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
```



53 eigenvectors used to explain 95% variance

Proposed 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

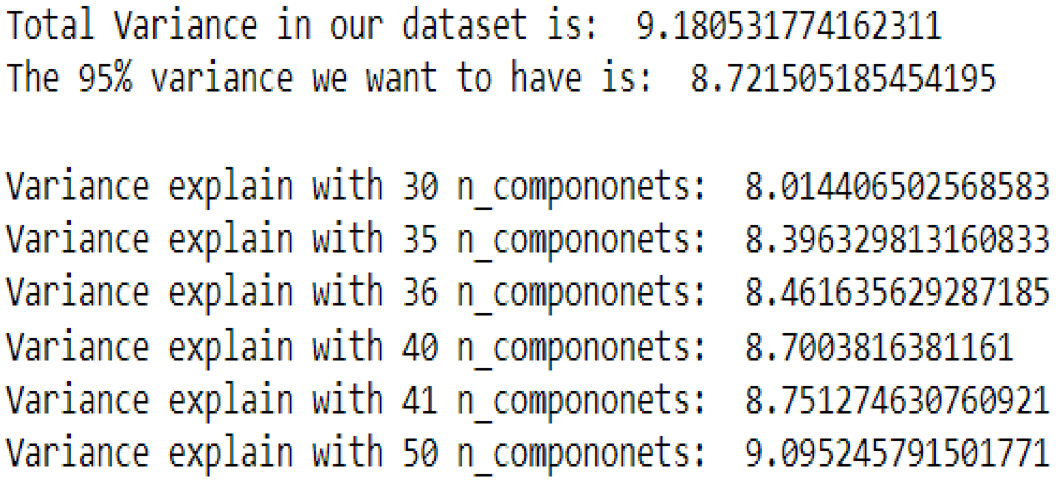
→ attending the events 4 times a year and 5+ times a year are similar

Combining such many similar features to produce new features and removing the old features reduces the columns without any loss in explained variance.

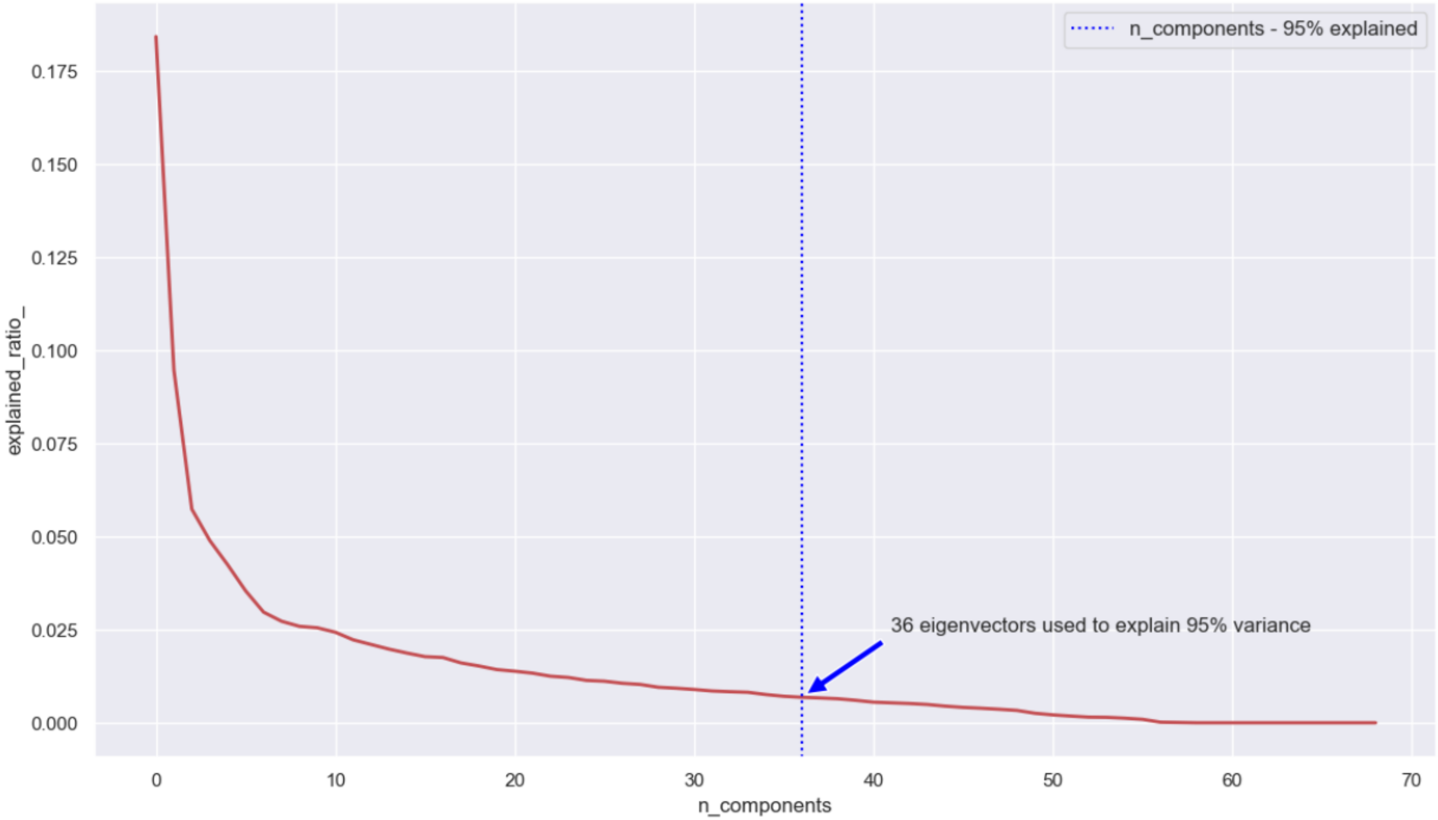
After this step we again run PCA for different principal components to check how many number of components gives approximately 95% of the total variance.

Proposed Methodology

PCA



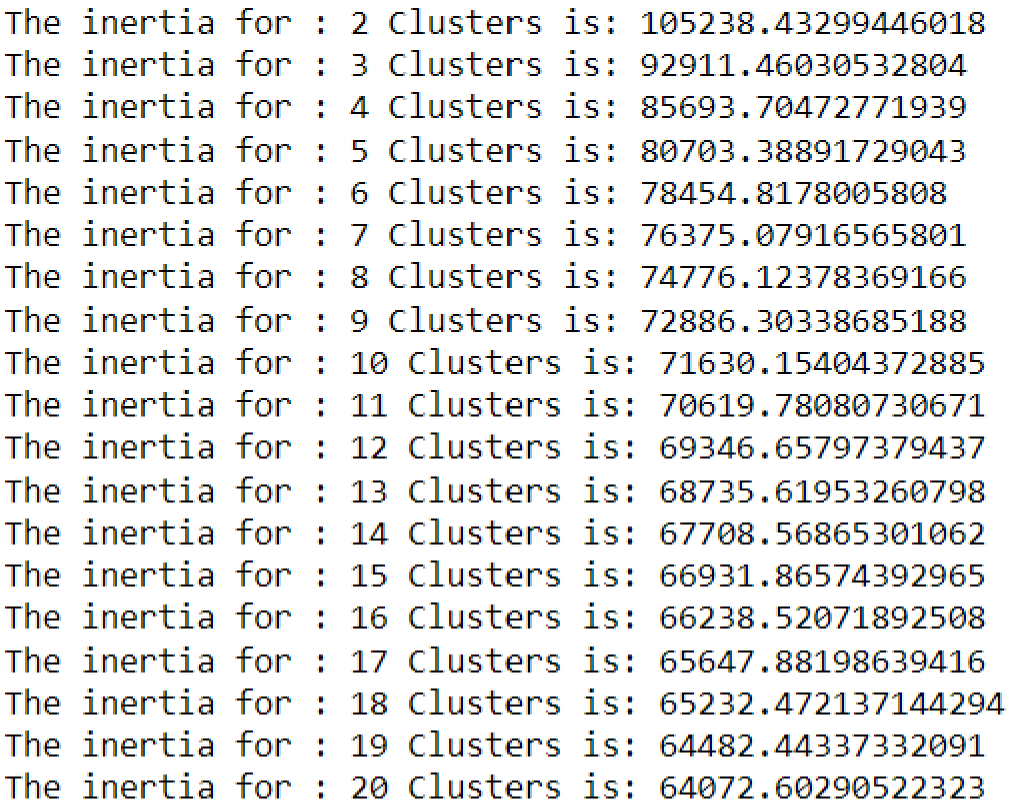
```
Total Variance in our dataset is:  9.180531774162311
The 95% variance we want to have is:  8.721505185454195

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
```



Proposed Methodology

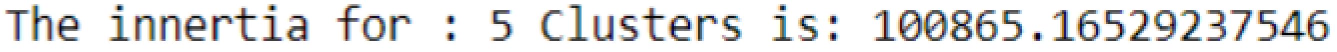
K-Means

We again run K-means from 2 clusters to 20 clusters to check the inertia score for 36 features(reduced by PCA). Our main aim will be to reduce the inertia score for 5 clusters.



```
The inertia for : 2 Clusters is: 105238.43299446018
The inertia for : 3 Clusters is: 92911.46030532804
The inertia for : 4 Clusters is: 85693.70472771939
The inertia for : 5 Clusters is: 80703.38891729043
The inertia for : 6 Clusters is: 78454.8178005808
The inertia for : 7 Clusters is: 76375.07916565801
The inertia for : 8 Clusters is: 74776.12378369166
The inertia for : 9 Clusters is: 72886.30338685188
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: 65232.472137144294
The inertia for : 19 Clusters is: 64482.44337332091
The inertia for : 20 Clusters is: 64072.60290522323
```

Previously obtained inertia for 5 clusters



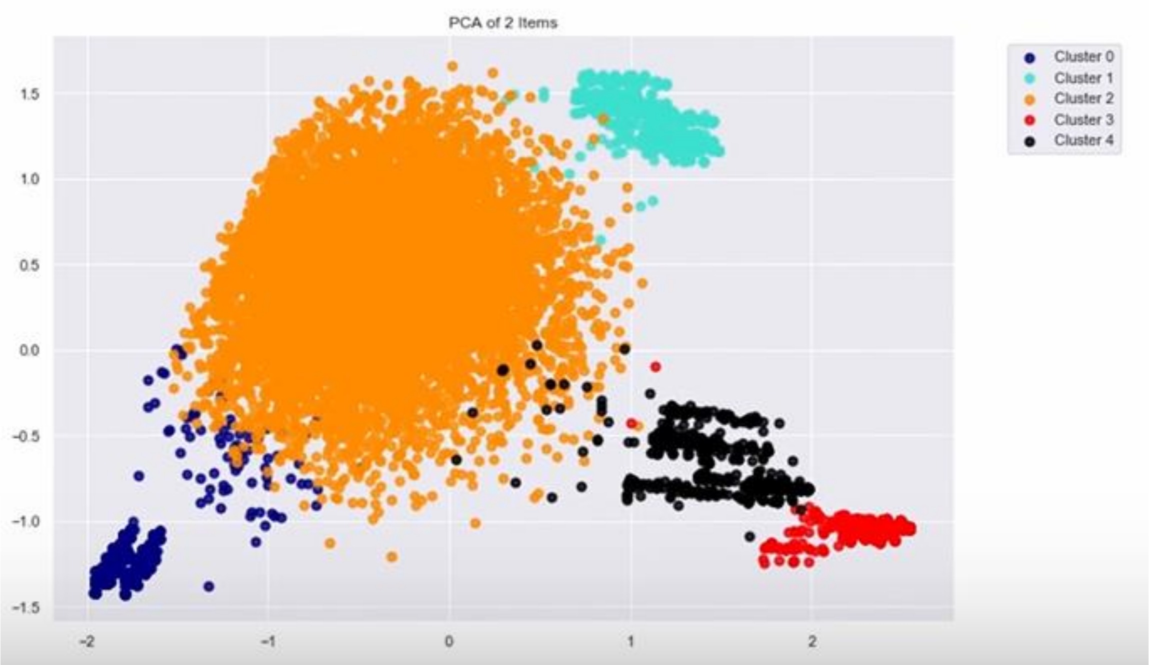
```
The inertia for : 5 Clusters is: 100865.1652923754
```

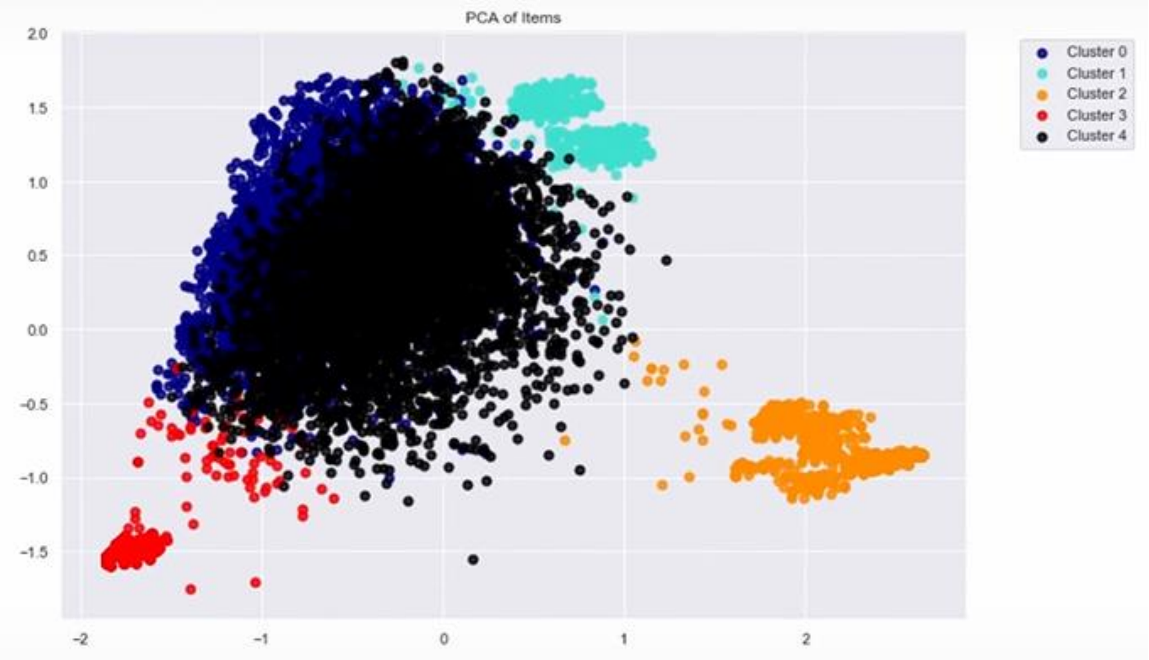
Functional and Non-Functional Requirements

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

Expected Outcome

clusters using K-means clustering unsupervised ML algorithm

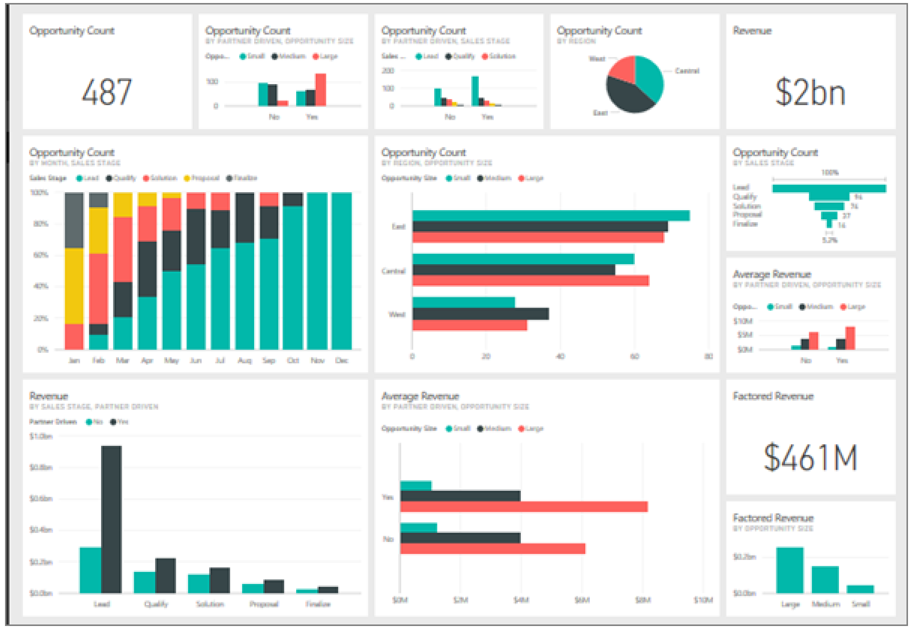


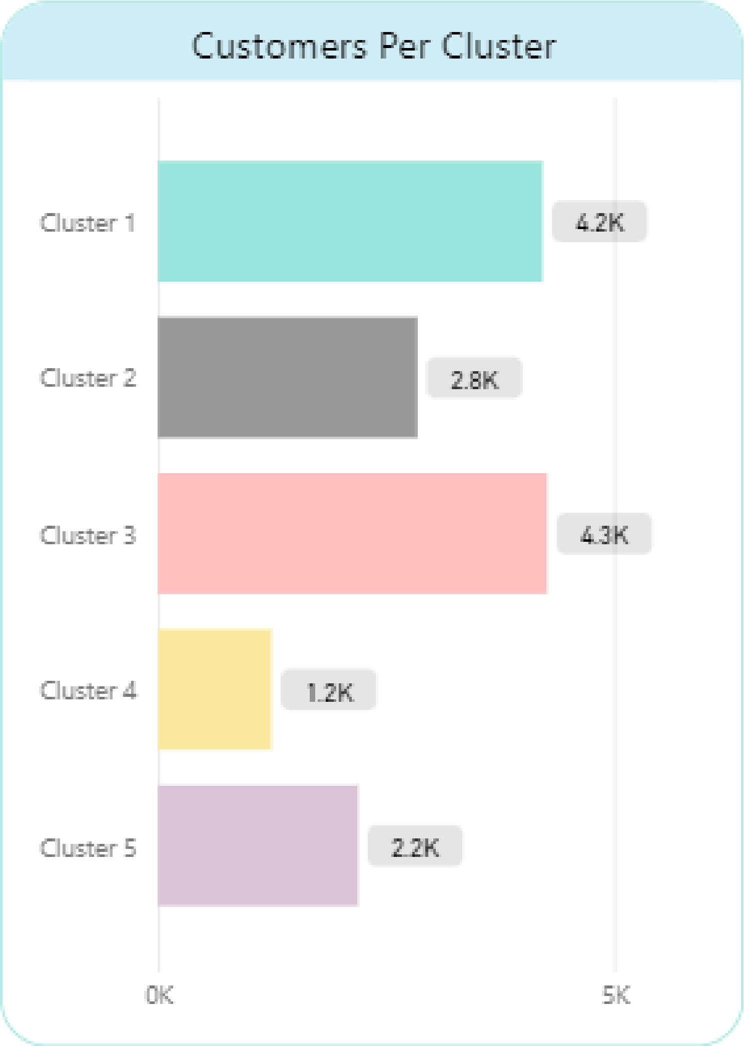


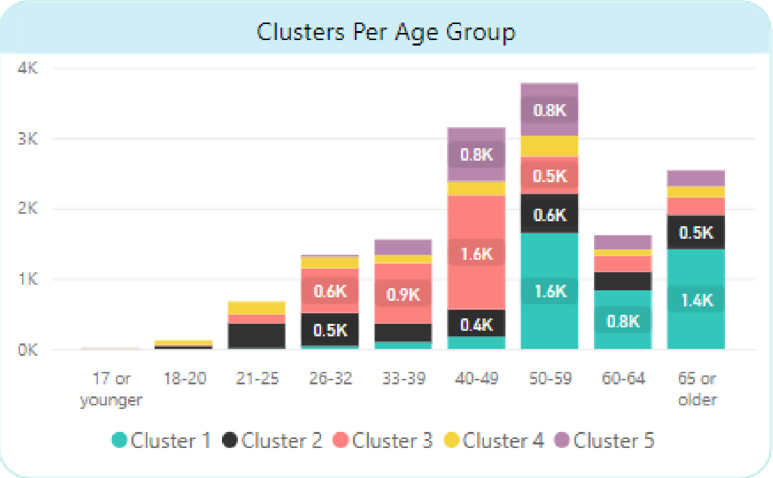
Before PCA

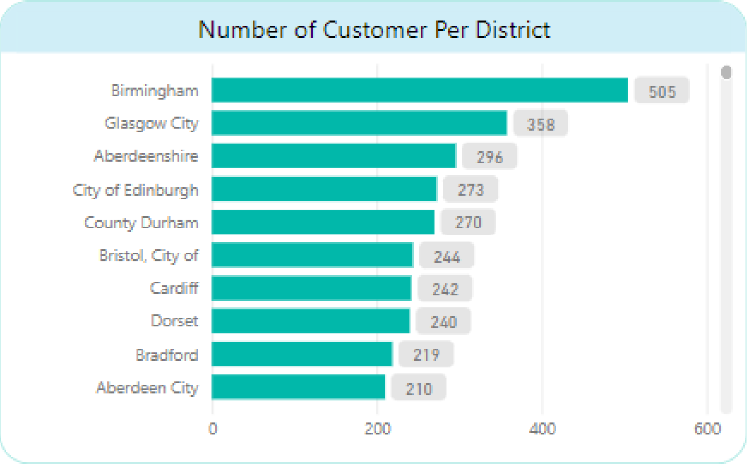
After PCA

Expected Outcome









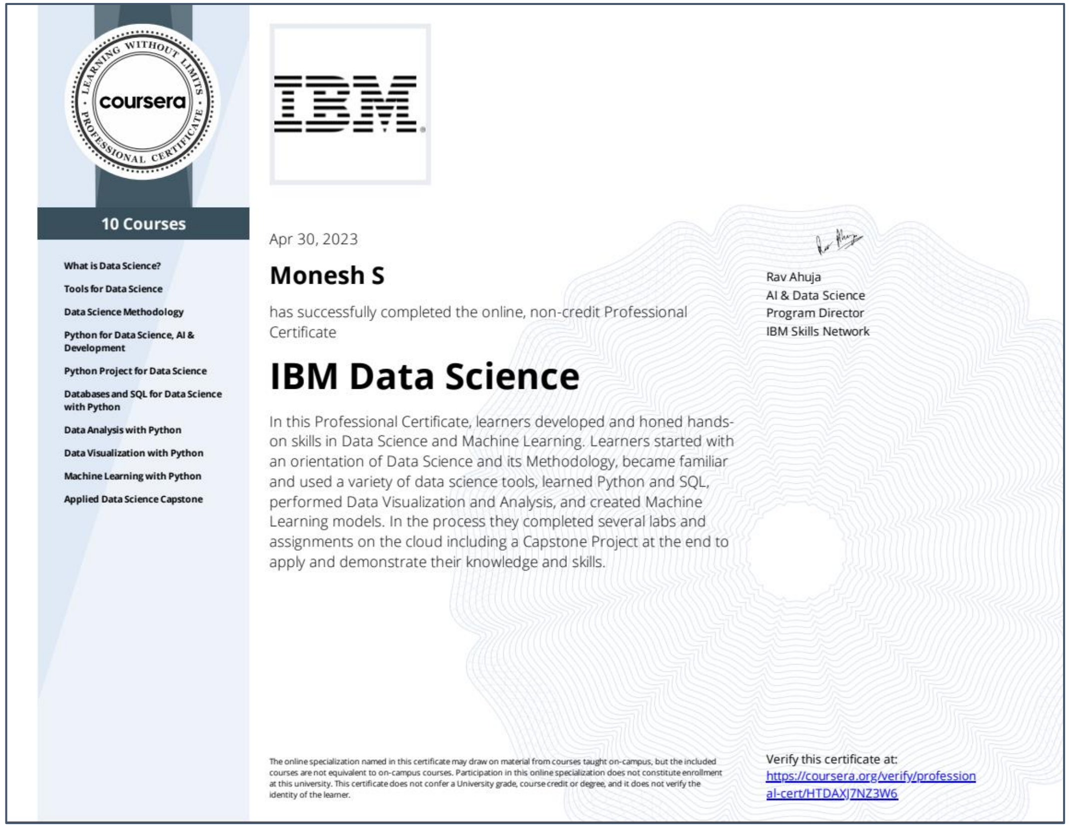
Conclusion

The integration of Excel, Machine Learning (ML), and Power BI offers a powerful suite of tools for data analysis and visualization. This project has demonstrated the capabilities and advantages of leveraging these technologies to extract insights, make data-driven decisions, and drive business growth.

- ➔ **Excel**: Excel is a widely used spreadsheet software that offers powerful data analysis and manipulation capabilities. It allows data storing, cleaning, transformation, and modeling. With its wide range of functions and formulas.
- ➔ **Machine Learning (ML)**: ML algorithms provide the ability to discover patterns and insights from large datasets and automate predictions. By applying **K-means clustering**, we can uncover hidden trends and relationships in reduced data using **Principal Component Analysis(PCA)**, leading to improved decision-making and predictive capabilities.
- ➔ **Power BI**: Power BI is a business intelligence tool that enables the creation of interactive **dashboards and reports**.

By integrating data from Excel spreadsheets, ML models and Power BI enables data-driven storytelling and empowers decision-makers with real-time information.

About MOOC



LEARNING WITHOUT LIMITS · PROFESSIONAL CERTIFICATE

coursera

IBM

10 Courses

What is Data Science?

Tools for Data Science

Data Science Methodology

Python for Data Science, AI & Development

Python Project for Data Science

Databases and SQL for Data Science with Python

Data Analysis with Python

Data Visualization with Python

Machine Learning with Python

Applied Data Science Capstone

Apr 30, 2023

Monesh S

has successfully completed the online, non-credit Professional Certificate

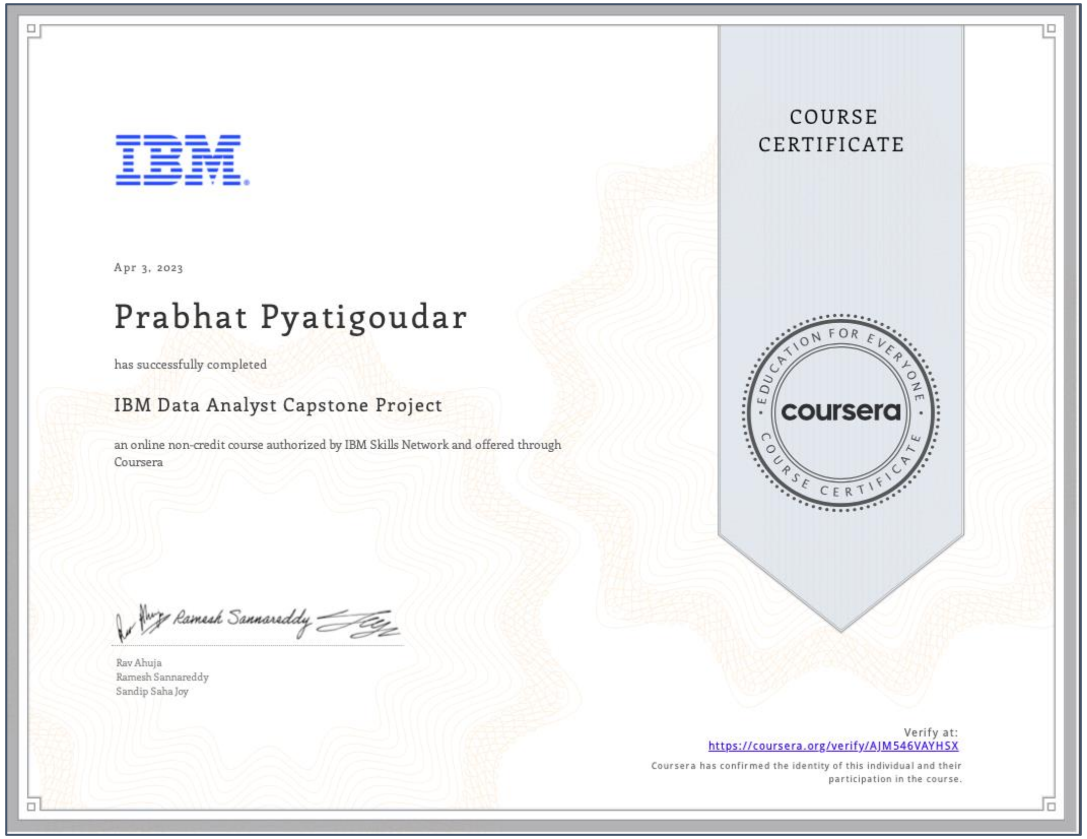
IBM Data Science

In this Professional Certificate, learners developed and honed hands-on skills in Data Science and Machine Learning. Learners started with an orientation of Data Science and its Methodology, became familiar and used a variety of data science tools, learned Python and SQL, performed Data Visualization and Analysis, and created Machine Learning models. In the process they completed several labs and assignments on the cloud including a Capstone Project at the end to apply and demonstrate their knowledge and skills.

Rav Ahuja
AI & Data Science
Program Director
IBM Skills Network

The online specialization named in this certificate may draw on material from courses taught on-campus, but the included courses are not equivalent to on-campus courses. Participation in this online specialization does not constitute enrollment at this university. This certificate does not confer a University grade, course credit or degree, and it does not verify the identity of the learner.

Verify this certificate at:
https://coursera.org/verify/professional-cert/HTDAXJ7NZ3W6



IBM

COURSE CERTIFICATE

EDUCATION FOR EVERYONE · COURSE CERTIFICATE

coursera

Apr 3, 2023

Prabhat Pyatigoudar

has successfully completed

IBM Data Analyst Capstone Project

an online non-credit course authorized by IBM Skills Network and offered through Coursera

Ramesh Sannareddy

Rav Ahuja
Ramesh Sannareddy
Sandip Saha Joy

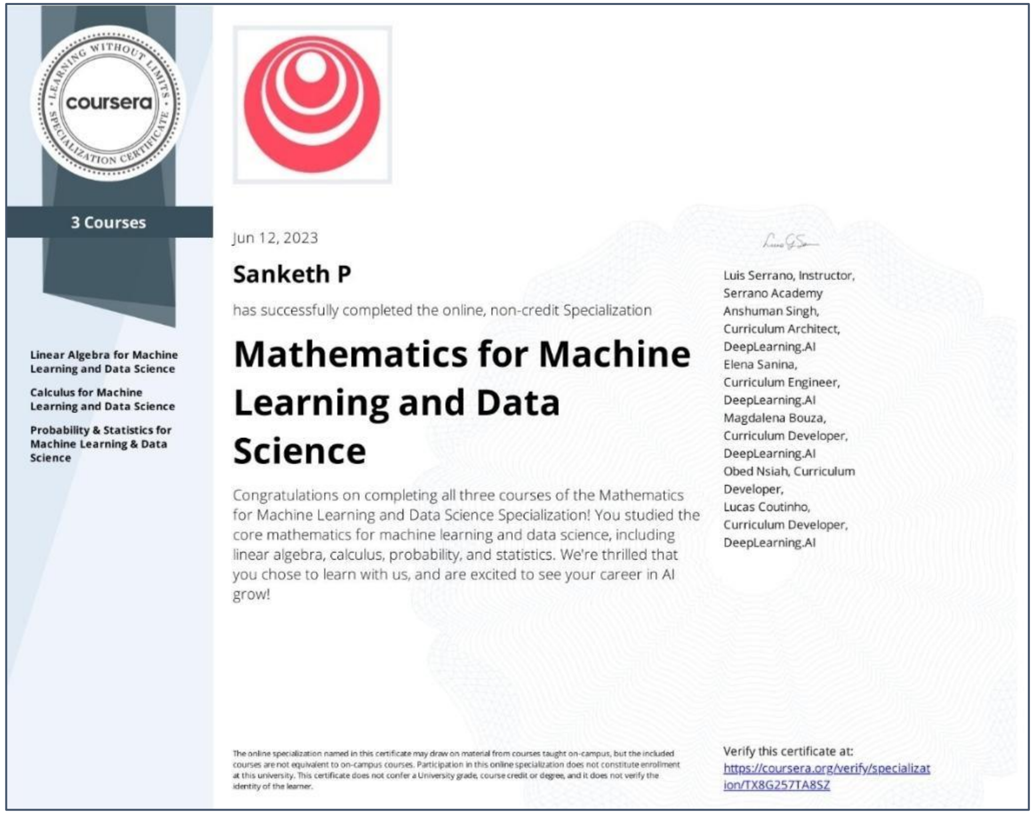
Verify at:
https://coursera.org/verify/AJM546VAYHSX

Coursera has confirmed the identity of this individual and their participation in the course.

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

Prabhat G P
IBM Data Analyst
2 Weeks
Coursera IBM Data Analyst Capstone

About MOOC



LEARNING WITHOUT LIMITS · coursera · SPECIALIZATION CERTIFICATE

3 Courses

Linear Algebra for Machine Learning and Data Science

Calculus for Machine Learning and Data Science

Probability & Statistics for Machine Learning & Data Science

Jun 12, 2023

Sanketh P

has successfully completed the online, non-credit Specialization

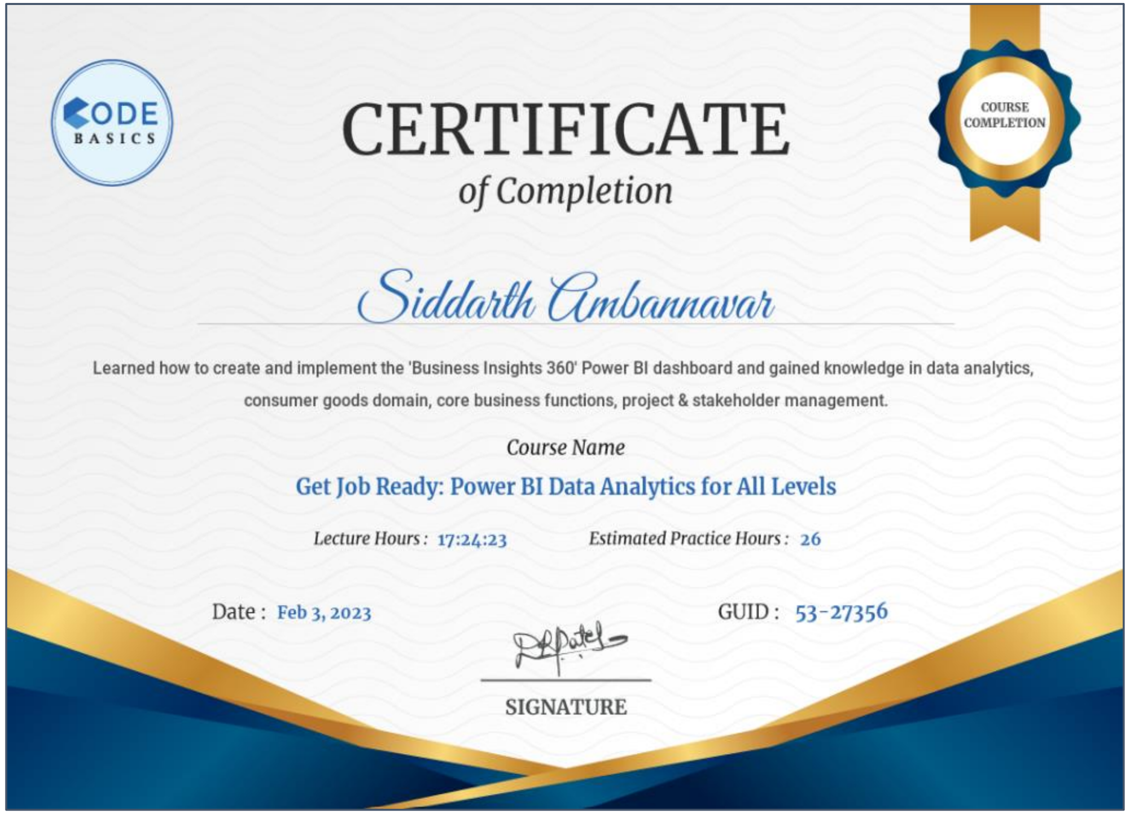
Mathematics for Machine Learning and Data Science

Congratulations on completing all three courses of the Mathematics for Machine Learning and Data Science Specialization! You studied the core mathematics for machine learning and data science, including linear algebra, calculus, probability, and statistics. We're thrilled that you chose to learn with us, and are excited to see your career in AI grow!

Luis Serrano, Instructor, Serrano Academy
Anshuman Singh, Curriculum Architect, DeepLearning.AI
Elena Sanina, Curriculum Engineer, DeepLearning.AI
Magdalena Bouza, Curriculum Developer, DeepLearning.AI
Obed Nsiah, Curriculum Developer,
Lucas Coutinho, Curriculum Developer, DeepLearning.AI

The online specialization named in this certificate may draw on material from courses taught on-campus, but the included courses are not equivalent to on-campus courses. Participation in this online specialization does not constitute enrollment at this university. This certificate does not confer a University grade, course credit or degree, and it does not verify the identity of the learner.

Verify this certificate at:
https://coursera.org/verify/specialization/TX8G257TA8SZ



CODE BASICS

COURSE COMPLETION

CERTIFICATE
of Completion

Siddarth Ambannavar

Learned how to create and implement the 'Business Insights 360' Power BI dashboard and gained knowledge in data analytics, consumer goods domain, core business functions, project & stakeholder management.

Course Name

Get Job Ready: Power BI Data Analytics for All Levels

Lecture Hours : 17:24:23 *Estimated Practice Hours :* 26

Date : Feb 3, 2023

GUID : 53-27356

SIGNATURE

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 !