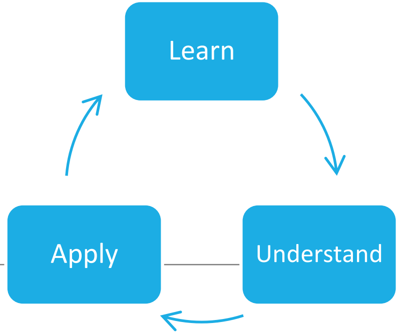
Travel Review Ratings



Subject: Machine Learning

Presented By,

Rajaraman Ganesan – 251056279 (M.Eng)

Vatsal Shah – 251041322 (M.Eng)

Department of Electrical & Computer Engineering (M.Engg)

Western University, Canada

- Objective & Goal
- Dataset Overview
- Approach Design
- Visualization Overview
- Machine Learning Model
- Results
- Amazon SageMaker
- Future Work & *QR Code

Objective & Goal

- Understand the **process** & **difference** by applying algorithms with **custom models** and **pre-build libraries** in Python

- ✓ Apply algorithm in most-demanded languages in the industry - **Python & R**

- **Compare** process, results and time complexity

- Use **Amazon SageMaker** - understand the flow

- Objective & Goal
- Approach Design
- Machine Learning Model
- Amazon SageMaker
- Dataset Overview
- Visualization Overview
- Results
- Future Work & *QR Code

Dataset overview

- Travel planning has become one of important commercial use.
- Many tourists look for some places like malls, restaurants or vacation spots, etc. online in recent times.
- Online review plays a critical role in the tourism industry, which mainly offers services and focuses on customer satisfaction.

No Kaggle Kernels

Oriented: UCI Machine Learning Repository (Link)
Attributes : 25
Tuples: 5456

Attribute 1 : Unique user id

Attribute 2 : Average ratings on churches

Attribute 3 : Average ratings on resorts

Attribute 4 : Average ratings on beaches

Attribute 5 : Average ratings on parks

Attribute 6 : Average ratings on theatres

Attribute 7 : Average ratings on museums

Attribute 8 : Average ratings on malls

Attribute 9 : Average ratings on zoo

Attribute 10 : Average ratings on restaurants

Attribute 11 : Average ratings on pubs/bars

Attribute 12 : Average ratings on local services

Attribute 13 : Average ratings on burger/pizza shops

Attribute 14 : Average ratings on hotels/other lodgings

Attribute 15 : Average ratings on juice bars

Attribute 16 : Average ratings on art galleries

Attribute 17 : Average ratings on dance clubs

Attribute 18 : Average ratings on swimming pools

Attribute 19 : Average ratings on gyms

Attribute 20 : Average ratings on bakeries

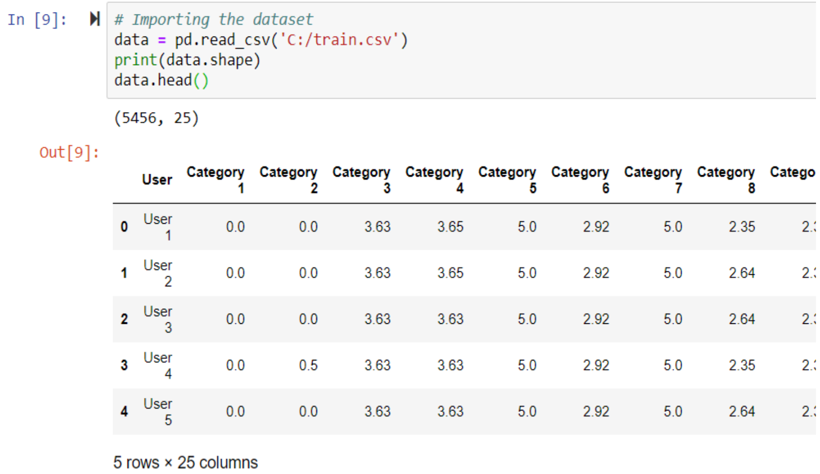
Attribute 21 : Average ratings on beauty & spas

Attribute 22 : Average ratings on cafes

Attribute 23 : Average ratings on view points

Attribute 24 : Average ratings on monuments

Attribute 25 : Average ratings on gardens



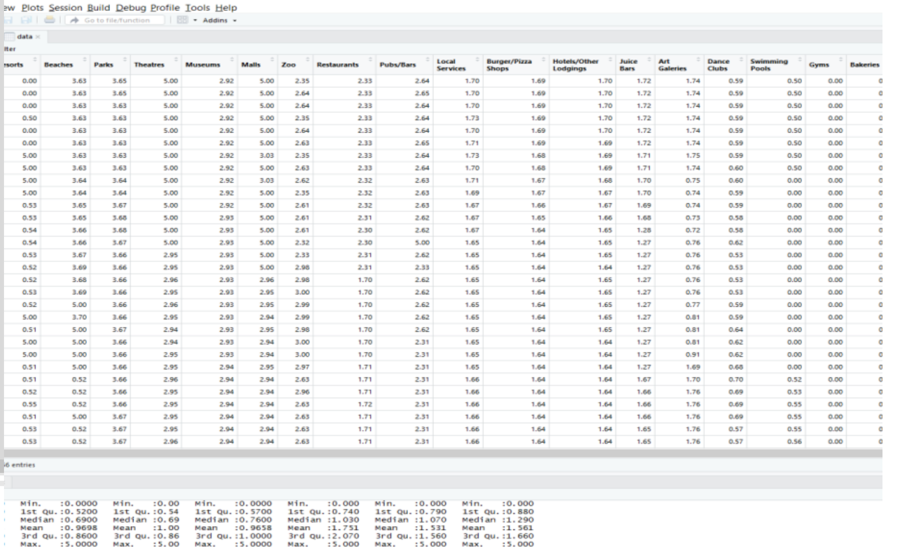
```
In [9]:  # Importing the dataset
         data = pd.read_csv('C:/train.csv')
         print(data.shape)
         data.head()

(5456, 25)

Out[9]:
```

	User	Category 1	Category 2	Category 3	Category 4	Category 5	Category 6	Category 7	Category 8	Catego
0	User 1	0.0	0.0	3.63	3.65	5.0	2.92	5.0	2.35	2.
1	User 2	0.0	0.0	3.63	3.65	5.0	2.92	5.0	2.64	2.
2	User 3	0.0	0.0	3.63	3.63	5.0	2.92	5.0	2.64	2.
3	User 4	0.0	0.5	3.63	3.63	5.0	2.92	5.0	2.35	2.
4	User 5	0.0	0.0	3.63	3.63	5.0	2.92	5.0	2.64	2.

5 rows × 25 columns



esorts	Beaches	Parks	Theatres	Museums	Malls	Zoo	Restaurants	Pubs/Bars	Local Services	Burger/Pizza Shops	Hotels/Other Lodgings	Juice Bars	Art Galeries	Dance Clubs	Swimming Pools	Gyms	Bakeries
0.00	3.63	3.65	5.00	2.92	5.00	2.35	2.33	2.64	1.70	1.69	1.70	1.72	1.74	0.59	0.50	0.00	0
0.00	3.63	3.65	5.00	2.92	5.00	2.64	2.33	2.65	1.70	1.69	1.70	1.72	1.74	0.59	0.50	0.00	0
0.00	3.63	3.63	5.00	2.92	5.00	2.64	2.33	2.64	1.70	1.69	1.70	1.72	1.74	0.59	0.50	0.00	0
0.50	3.63	3.63	5.00	2.92	5.00	2.35	2.33	2.64	1.73	1.69	1.70	1.72	1.74	0.59	0.50	0.00	0
0.00	3.63	3.63	5.00	2.92	5.00	2.64	2.33	2.64	1.70	1.69	1.70	1.72	1.74	0.59	0.50	0.00	0
0.00	3.63	3.63	5.00	2.92	5.00	2.63	2.33	2.65	1.71	1.69	1.69	1.72	1.74	0.59	0.50	0.00	0
5.00	3.63	3.63	5.00	2.92	3.03	2.35	2.33	2.64	1.73	1.68	1.69	1.71	1.75	0.59	0.50	0.00	0
5.00	3.63	3.63	5.00	2.92	5.00	2.63	2.33	2.64	1.70	1.68	1.69	1.71	1.74	0.60	0.50	0.00	0
5.00	3.64	3.64	5.00	2.92	3.03	2.62	2.32	2.63	1.71	1.67	1.68	1.70	0.75	0.60	0.00	0.00	0
5.00	3.64	3.64	5.00	2.92	5.00	2.35	2.32	2.63	1.69	1.67	1.67	1.70	0.74	0.59	0.00	0.00	0
0.53	3.65	3.67	5.00	2.92	5.00	2.61	2.32	2.63	1.67	1.66	1.67	1.69	0.74	0.59	0.00	0.00	0
0.53	3.65	3.68	5.00	2.93	5.00	2.61	2.31	2.62	1.67	1.65	1.66	1.68	0.73	0.58	0.00	0.00	0
0.54	3.66	3.68	5.00	2.93	5.00	2.61	2.30	2.62	1.67	1.64	1.65	1.28	0.72	0.58	0.00	0.00	0
0.54	3.66	3.67	5.00	2.93	5.00	2.32	2.30	5.00	1.65	1.64	1.65	1.27	0.76	0.62	0.00	0.00	0
0.53	3.67	3.66	2.95	2.93	5.00	2.33	2.31	2.62	1.65	1.64	1.65	1.27	0.76	0.53	0.00	0.00	0
0.52	3.69	3.66	2.95	2.93	5.00	2.98	2.31	2.33	1.65	1.64	1.64	1.27	0.76	0.53	0.00	0.00	0
0.52	3.68	3.66	2.96	2.93	2.96	2.98	1.70	2.62	1.65	1.64	1.65	1.27	0.76	0.53	0.00	0.00	0
0.53	3.69	3.66	2.95	2.93	2.95	3.00	1.70	2.62	1.65	1.64	1.65	1.27	0.76	0.53	0.00	0.00	0
0.52	5.00	3.66	2.96	2.93	2.95	2.99	1.70	2.62	1.65	1.64	1.65	1.27	0.77	0.59	0.00	0.00	0
5.00	3.70	3.66	2.95	2.93	2.94	2.99	1.70	2.62	1.65	1.64	1.65	1.27	0.81	0.59	0.00	0.00	0
0.51	5.00	3.67	2.94	2.93	2.95	2.98	1.70	2.62	1.65	1.64	1.65	1.27	0.81	0.64	0.00	0.00	0
5.00	5.00	3.66	2.94	2.93	2.94	3.00	1.70	2.31	1.65	1.64	1.64	1.27	0.81	0.62	0.00	0.00	0
5.00	5.00	3.66	2.95	2.93	2.94	3.00	1.70	2.31	1.65	1.64	1.64	1.27	0.91	0.62	0.00	0.00	0
0.51	5.00	3.66	2.95	2.94	2.95	2.97	1.71	2.31	1.65	1.64	1.64	1.27	1.69	0.68	0.00	0.00	0
0.51	0.52	3.66	2.96	2.94	2.94	2.63	1.71	2.31	1.66	1.64	1.64	1.67	1.70	0.70	0.52	0.00	0
0.52	0.52	3.66	2.95	2.94	2.94	2.96	1.71	2.31	1.66	1.64	1.64	1.66	1.76	0.69	0.53	0.00	0
0.55	0.52	3.66	2.95	2.94	2.94	2.63	1.72	2.31	1.66	1.64	1.64	1.66	1.76	0.69	0.55	0.00	0
0.51	5.00	3.67	2.95	2.94	2.94	2.63	1.71	2.31	1.66	1.64	1.64	1.66	1.76	0.69	0.55	0.00	0
0.53	0.52	3.67	2.95	2.94	2.94	2.63	1.71	2.31	1.66	1.64	1.64	1.65	1.76	0.57	0.55	0.00	0
0.53	0.52	3.67	2.96	2.94	2.94	2.63	1.71	2.31	1.66	1.64	1.64	1.65	1.76	0.57	0.56	0.00	0

6 entries

```
 Min.   :0.0000   Min.   :0.00   Min.   :0.0000   Min.   :0.000   Min.   :0.000   Min.   :0.000
 1st Qu.:0.5200   1st Qu.:0.54   1st Qu.:0.5700   1st Qu.:0.740   1st Qu.:0.790   1st Qu.:0.880
 Median :0.6900   Median :0.69   Median :0.7600   Median :1.030   Median :1.070   Median :1.290
 Mean   :0.9698   Mean   :1.00   Mean   :0.9658   Mean   :1.751   Mean   :1.531   Mean   :1.561
 3rd Qu.:0.8600   3rd Qu.:0.86   3rd Qu.:1.0000   3rd Qu.:2.070   3rd Qu.:1.560   3rd Qu.:1.660
 Max.   :5.0000   Max.   :5.00   Max.   :5.0000   Max.   :5.000   Max.   :5.000   Max.   :5.000
```

Objective & Goal

Dataset Overview

Approach Design

Visualization Overview

Machine Learning Model

Results

Amazon SageMaker

Future Work & *QR Code

Process Design

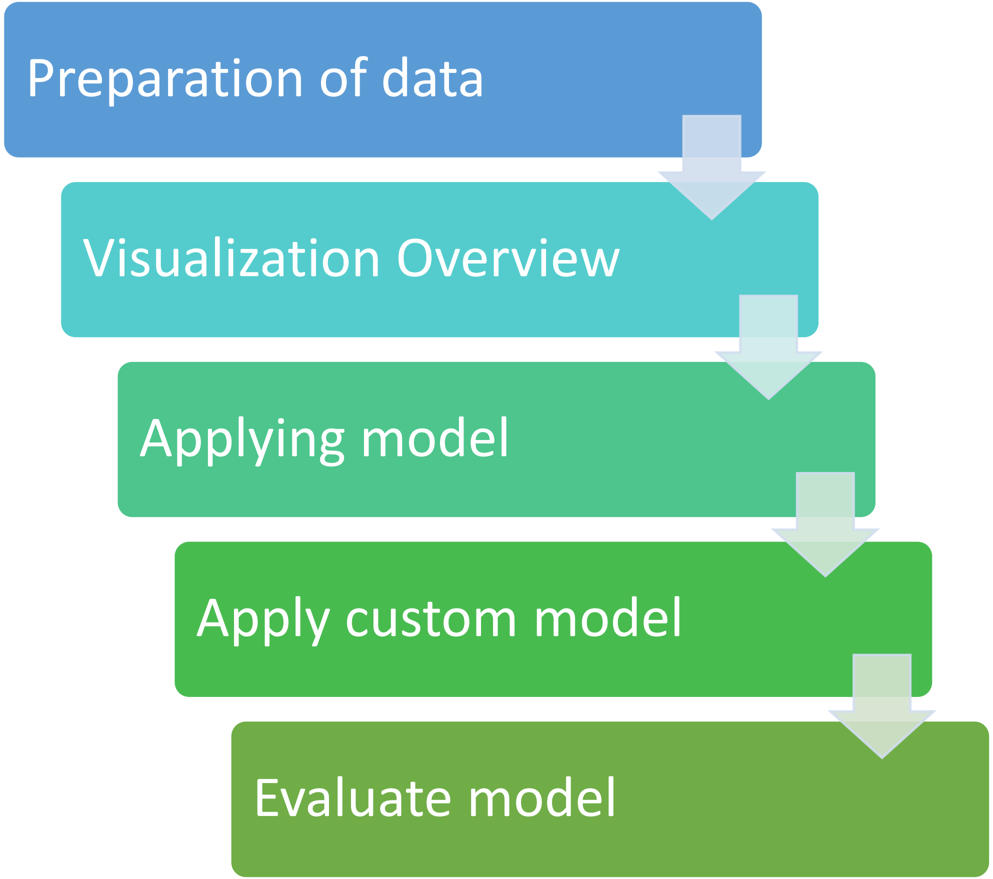
Unsupervised learning approach

Don't have a preliminary info on output values

Explore data to find some intrinsic structures in them

Use of clustering algorithm technique like k-means algorithm.

Approach Design



Preparation of data

Visualization Overview

Applying model

Apply custom model

Evaluate model

Objective & Goal

Dataset Overview

Approach Design

Visualization Overview

Machine Learning Model

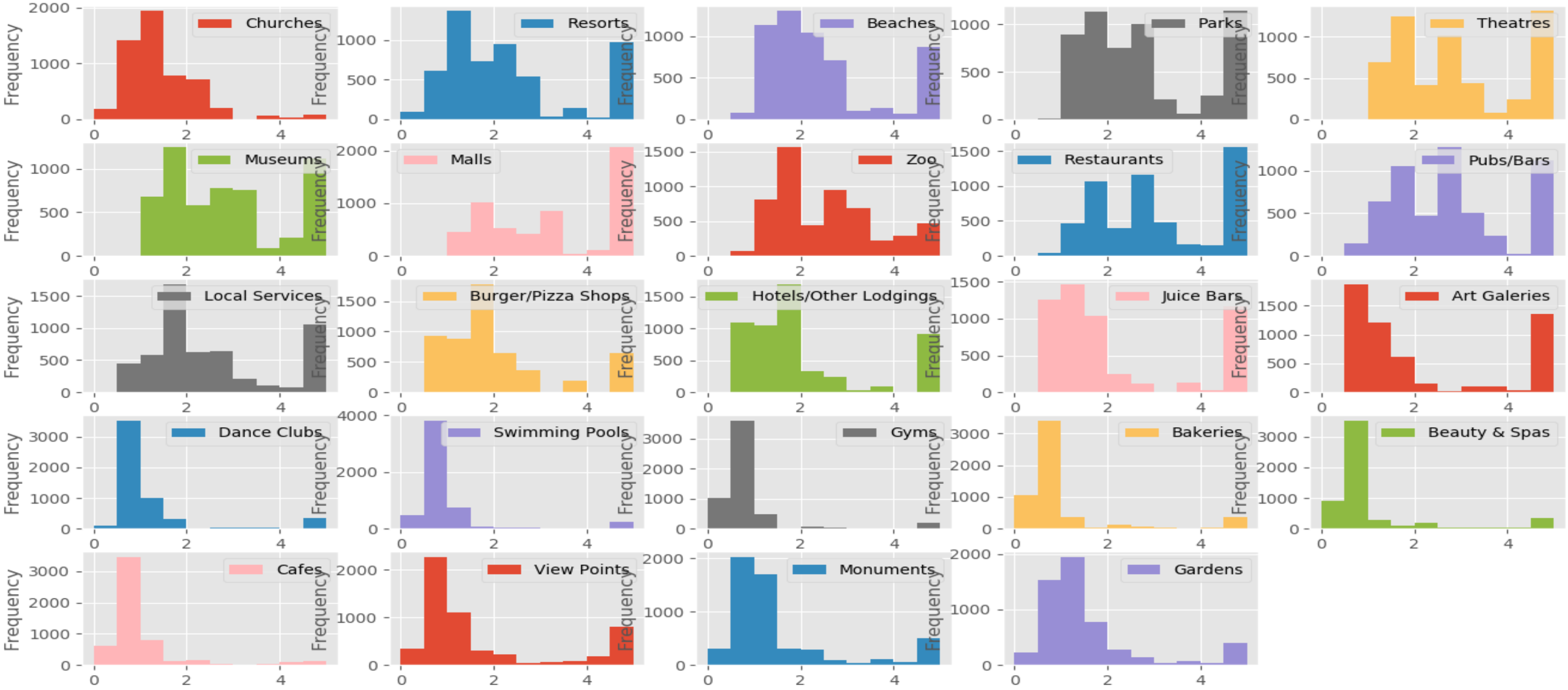
Results

Amazon SageMaker

Future Work & *QR Code

Visualization Overview

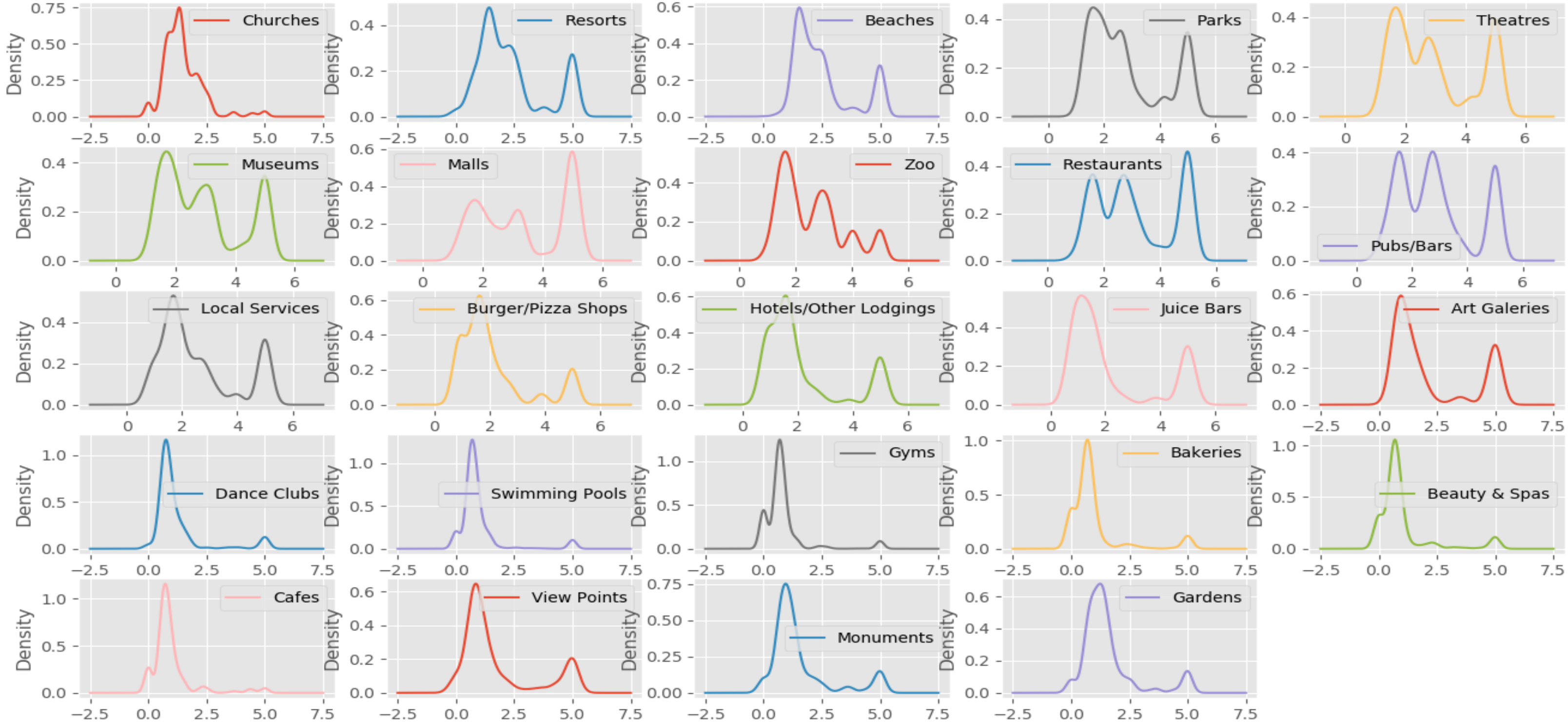
Shows 24 categories reviewed by users with hist diagram



(Screenshot taken during evaluation process)

Continue...

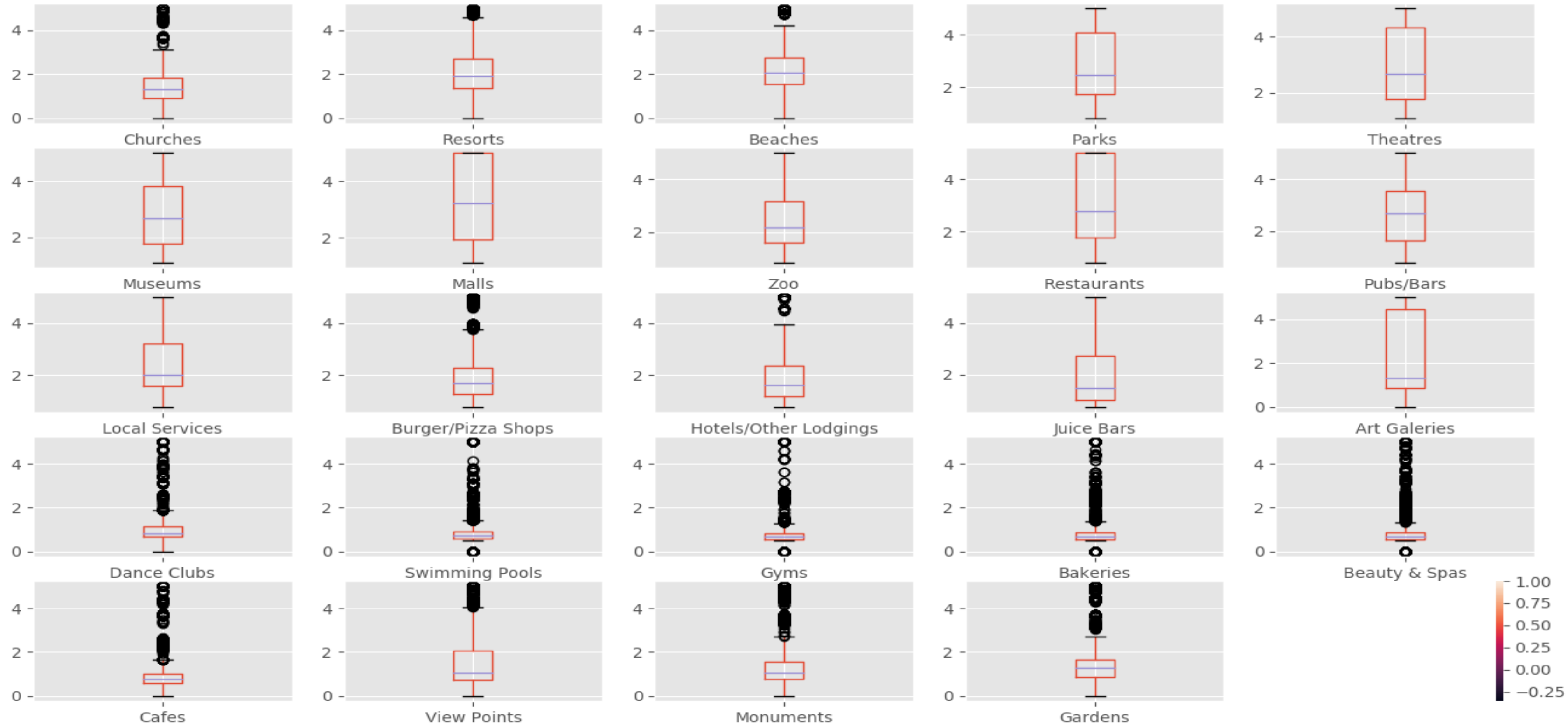
Shows 24 categories reviewed by users with density diagram



(Screenshot taken during evaluation process)

Continue…

Shows 24 categories reviewed by users with box plot diagram



(Screenshot taken during evaluation process)

Objective & Goal

Dataset Overview

Approach Design

Visualization Overview

Machine Learning Model

Results

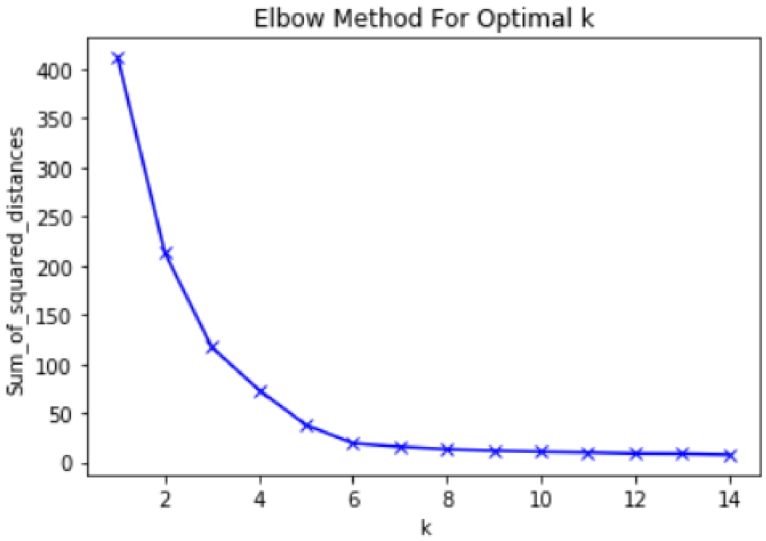
Amazon SageMaker

Future Work & *QR Code

K-Means Algorithm

- **Choose no: of clusters**
- Initialization
- Assign cluster
- Move centroid
- Optimization
- Convergence

- Within-cluster sum of squares is a measure of the variability of the observations within each cluster
- Idea behind using elbow method to choose after which **WSS** is almost constant.

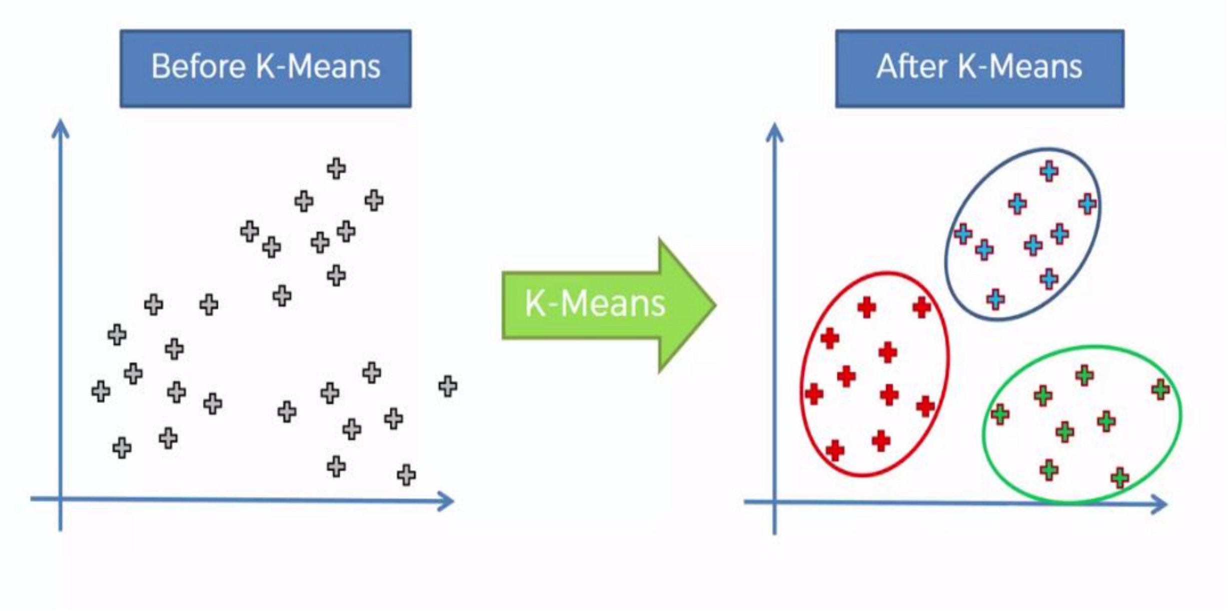


(Screenshot taken during evaluation process)

K-Means Algorithm

- Choose no: of clusters
- **Initialization**
- Assign cluster
- Move centroid
- Optimization
- Convergence

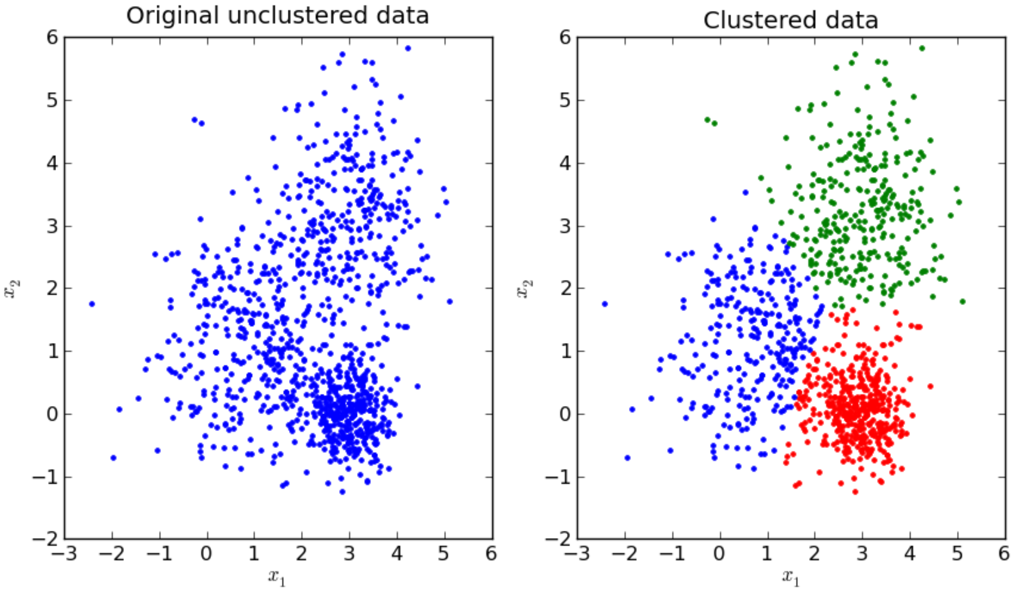
- Initialize k points, randomly
- Value of clusters are determined by elbow curve.



K-Means Algorithm

- Choose no: of clusters
- Initialization
- **Assign cluster**
- Move centroid
- Optimization
- Convergence

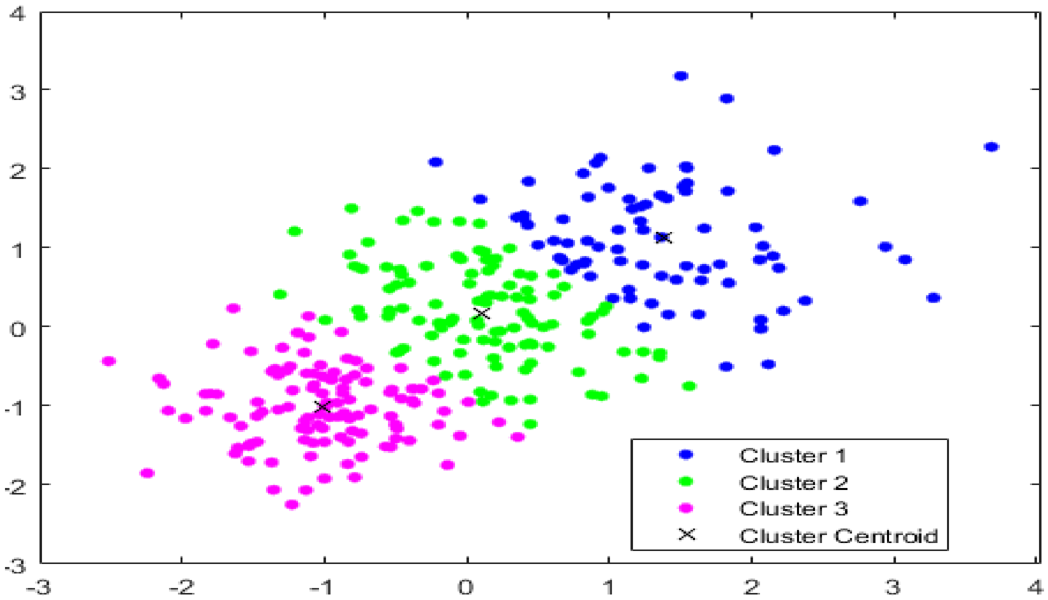
- Dist. between data points and centroid are computed.
- Based on min. Distance, data are divided into groups.



K-Means Algorithm

- Choose no: of clusters
- Initialization
- Assign cluster
- **Move centroid**
- Optimization
- Convergence

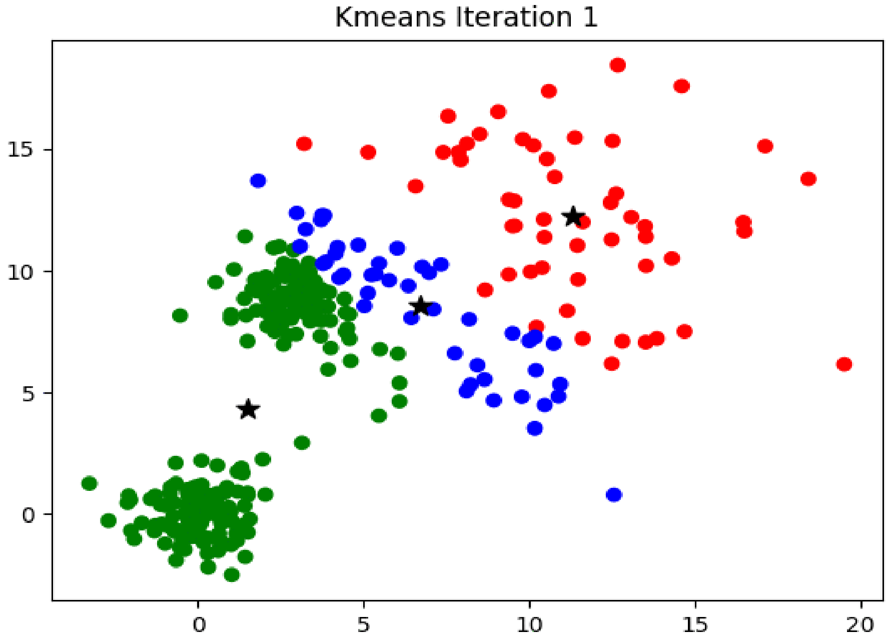
- Compute the mean of all three dots.
- Reposition blue, green, pink cluster centroid to this mean.



K-Means Algorithm

- Choose no: of clusters
- Initialization
- Assign cluster
- Move centroid
- **Optimization**
- Convergence

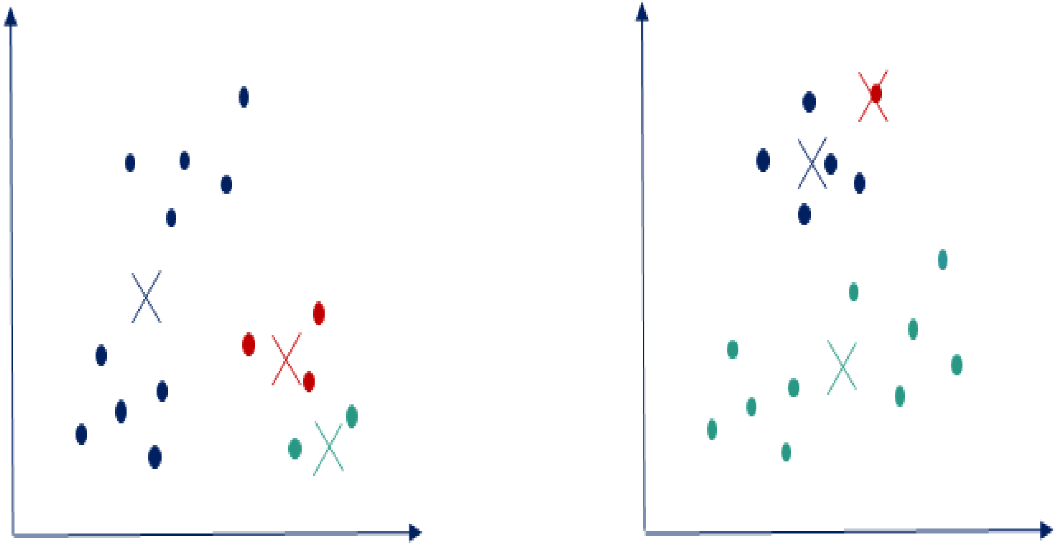
- Redo the above steps till cluster stop changing their positions.



K-Means Algorithm

- Choose no: of clusters
- Initialization
- Assign cluster
- Move centroid
- Optimization
- **Convergence**

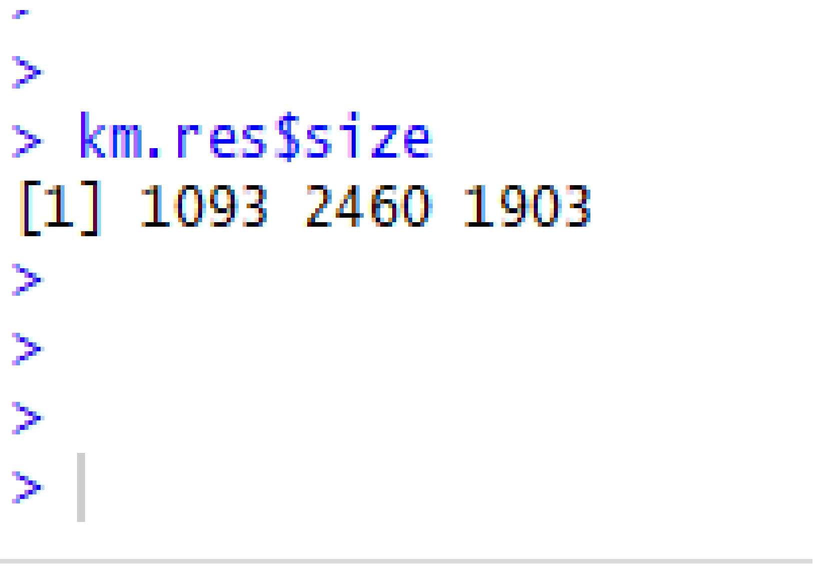
- Divide the data points to clusters, once the algorithm converges.



- Objective & Goal
- Dataset Overview
- Approach Design
- Visualization Overview
- Machine Learning Model
- Results
- Amazon SageMaker
- Future Work & *QR Code

R - Library

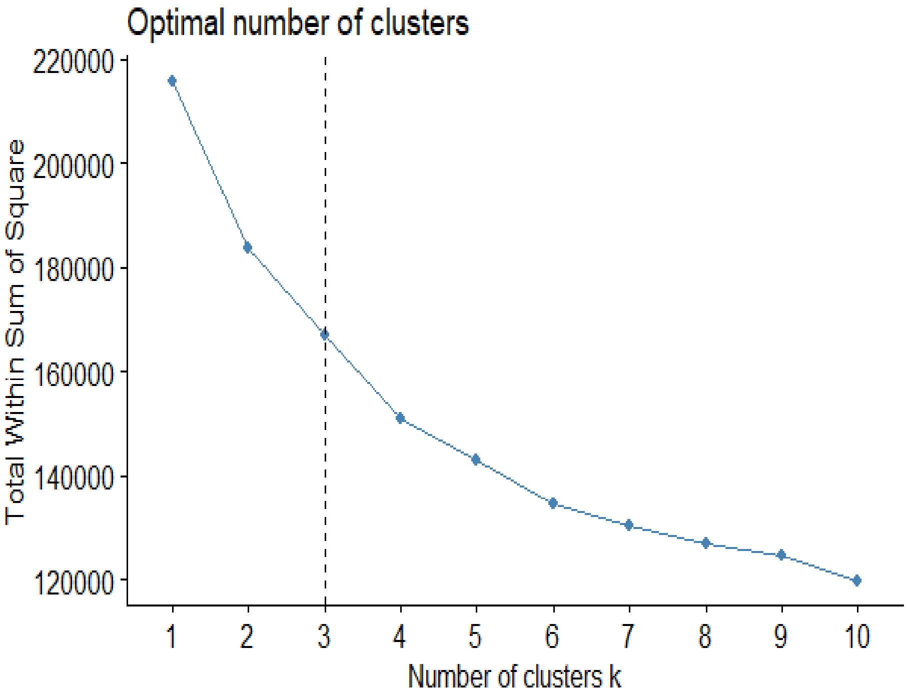
- library (cluster) –
  - Finding groups in data
- library (ggplot2) –
  - system for declaratively creating graphs
- library (factoextra) –
  - Used to visualize clusters



```
>
> km.res$size
[1] 1093 2460 1903
>
>
>
>
```

(Screenshot taken during evaluation process)

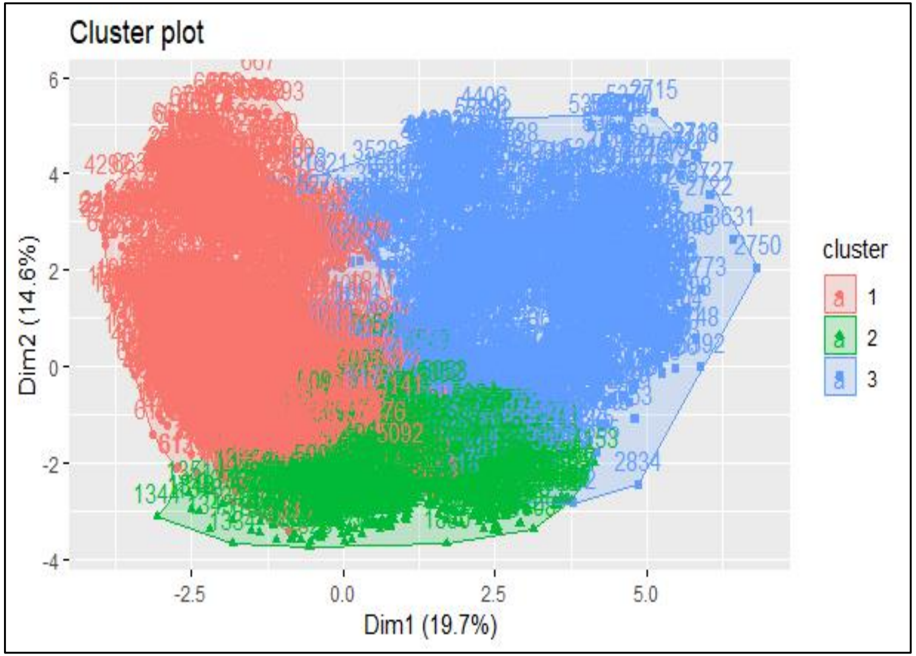
Results - R

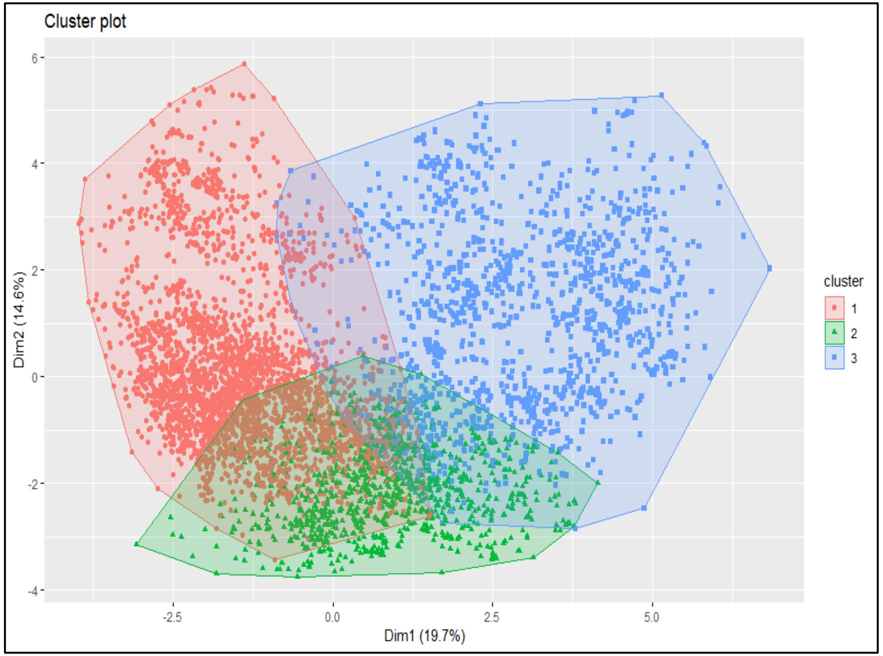


- Method to validate the number of clusters is the elbow method.
- Help finding the appropriate number of clusters in a dataset.
- Here, 3 clusters are suggested.

(Screenshot taken during evaluation process)

Continue...

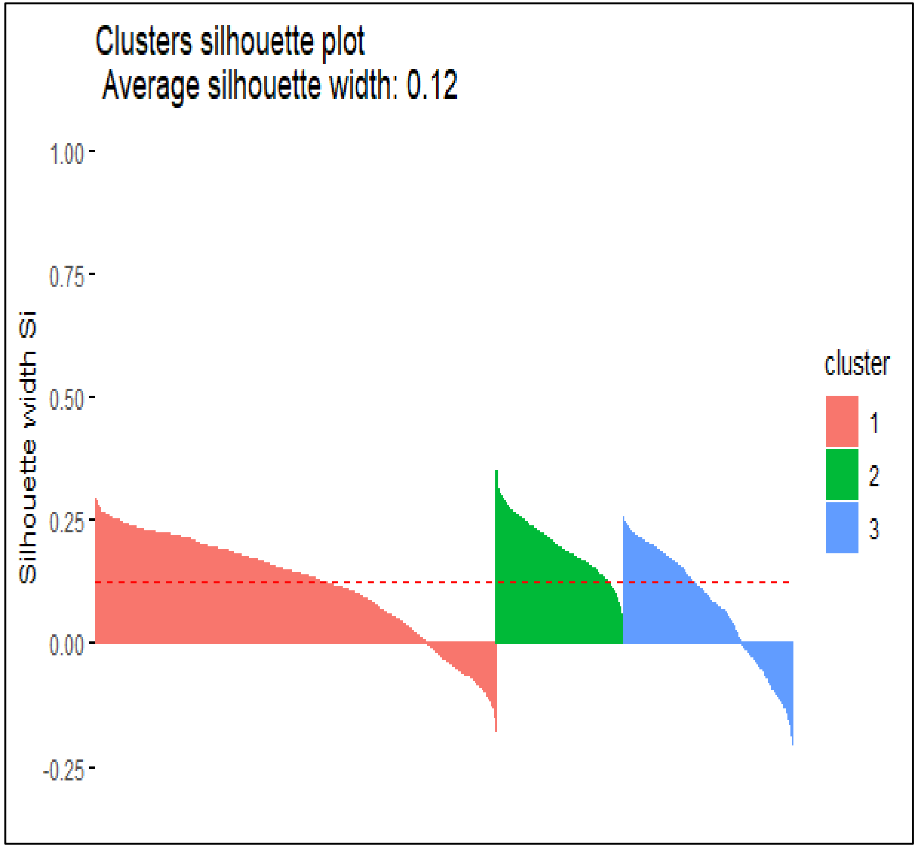




- As it a multidimensional dataset, to visualize cluster plot PCA is used.
- Reduce the dimensionality of a data set consisting of many variables correlated with each other

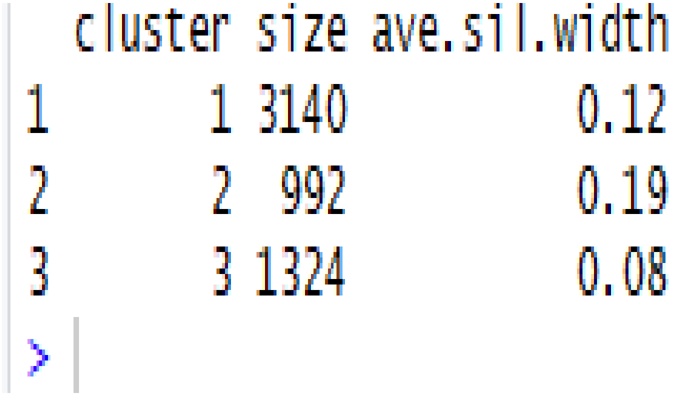
(Screenshot taken during evaluation process)

Continue…



Silhouette Plot shows for each cluster:

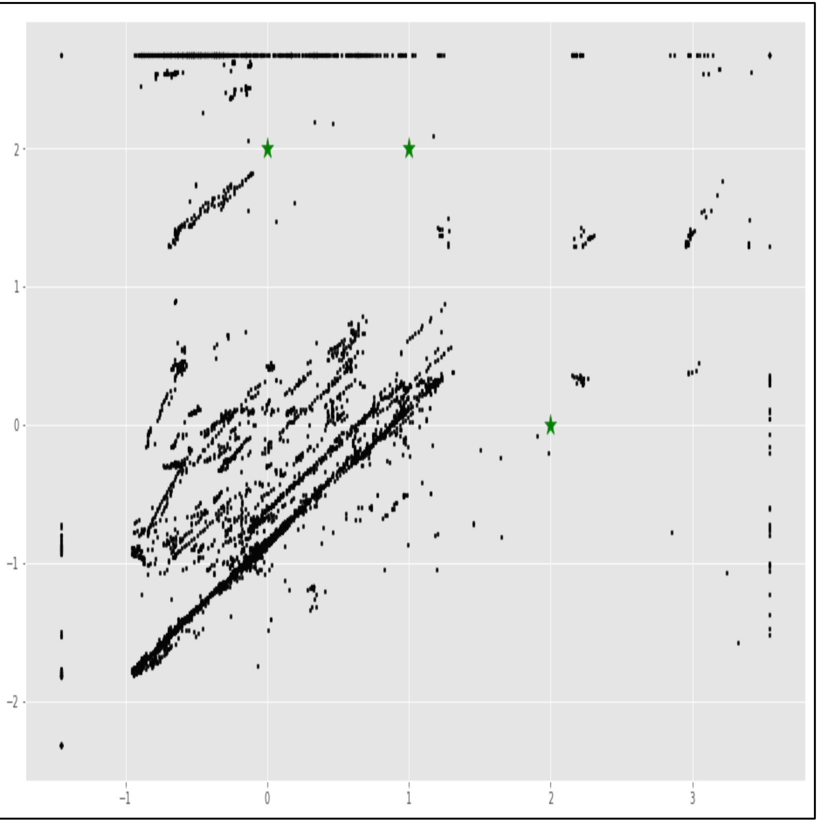
- Each vertical line corresponds to an element.
- The average silhouette width



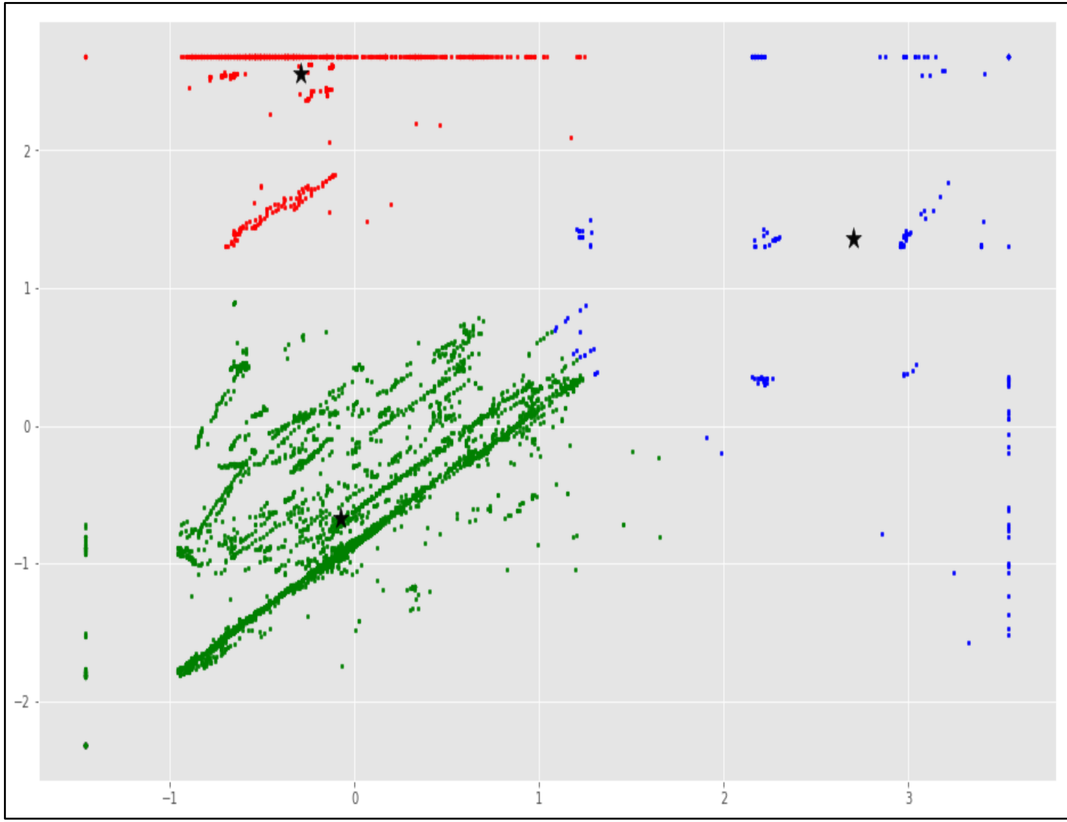
```
  cluster size ave.sil.width
1       1 3140          0.12
2       2  992          0.19
3       3 1324          0.08
>
```

(Screenshot taken during evaluation process)

Results - Python



Plot along with random centroid



Assigning new centroid

(Screenshot taken during evaluation process)

Objective & Goal

Dataset Overview

Approach Design

Visualization Overview

Machine Learning Model

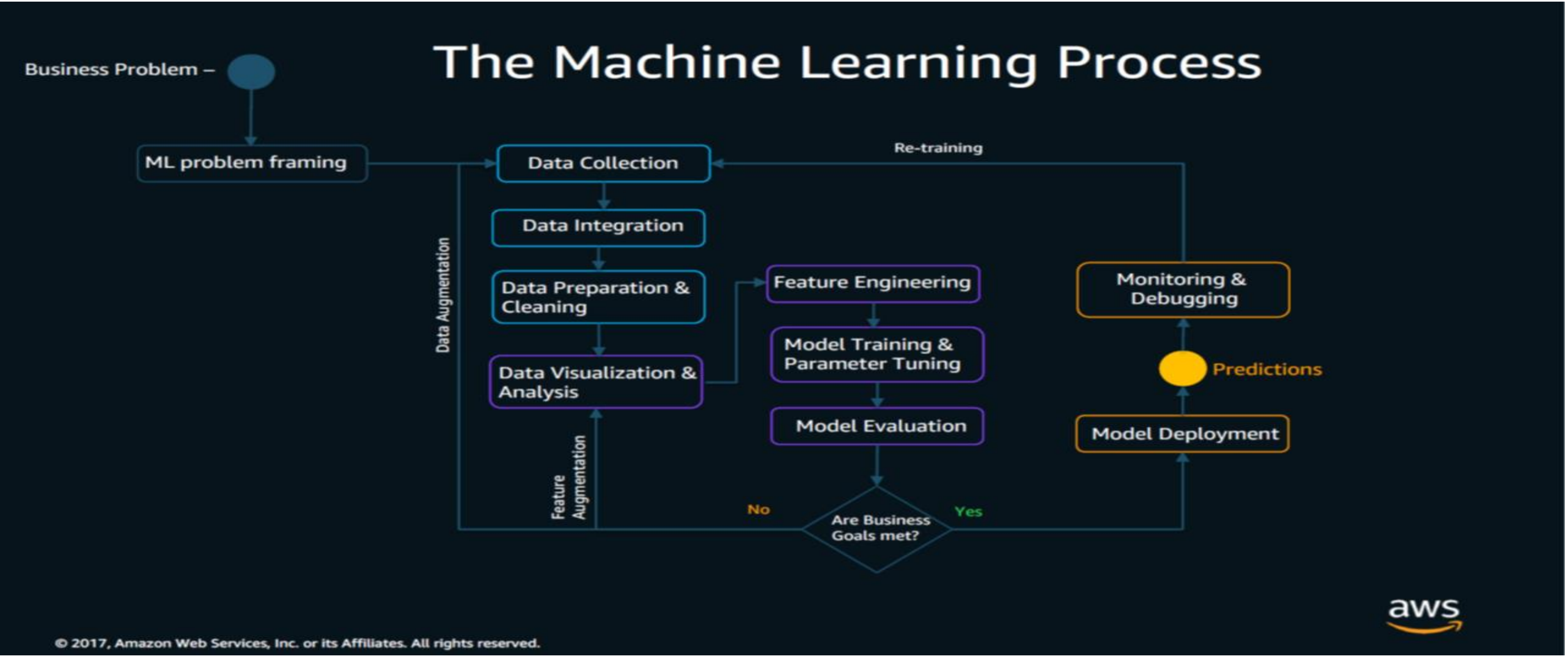
Results

Amazon SageMaker

Future Work & *QR Code

Amazon SageMaker

ML Process on AWS



The Machine Learning Process

Business Problem –

ML problem framing

Data Collection

Data Integration

Data Preparation & Cleaning

Data Visualization & Analysis

Feature Engineering

Model Training & Parameter Tuning

Model Evaluation

Are Business Goals met?

No

Yes

Data Augmentation

Feature Augmentation

Model Deployment

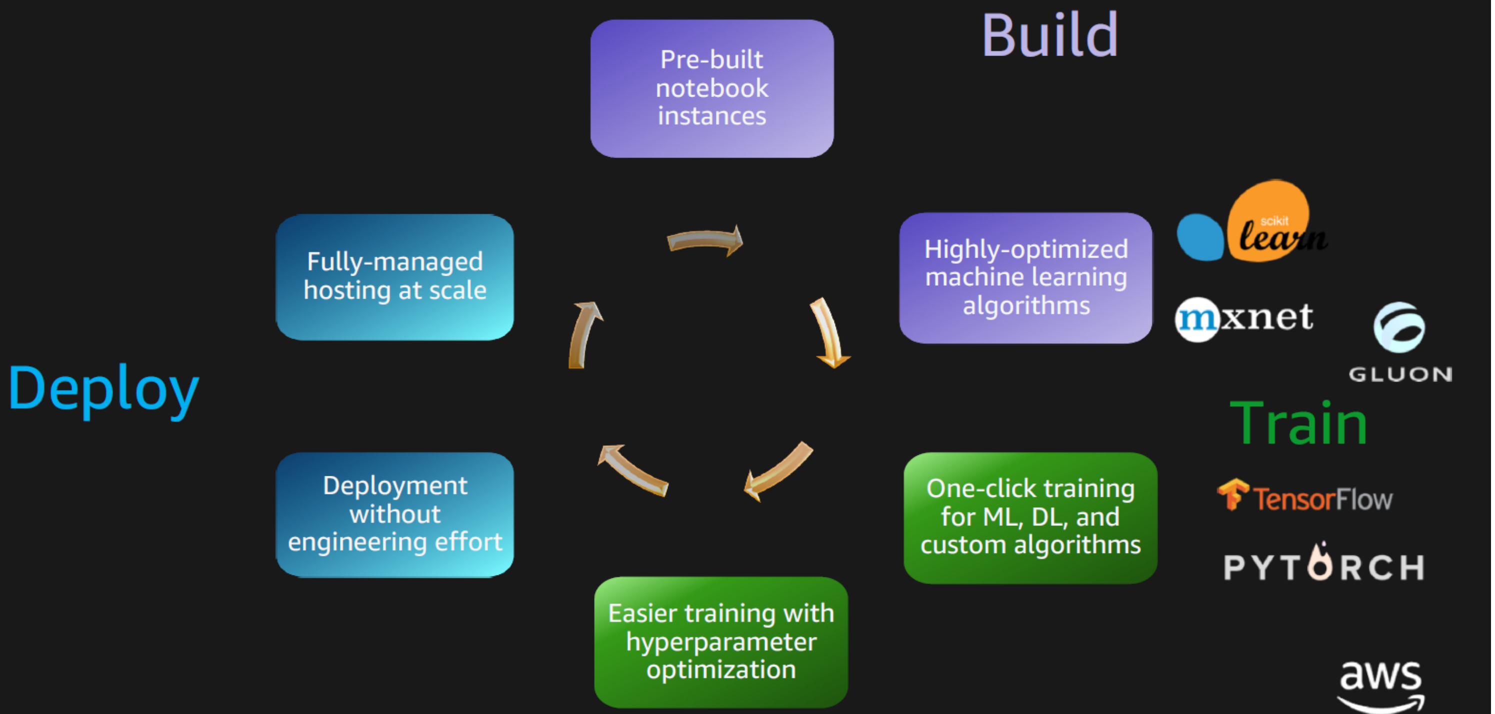
Predictions

Monitoring & Debugging

Re-training

aws

Amazon SageMaker



Build

Pre-built notebook instances

Highly-optimized machine learning algorithms

Train

One-click training for ML, DL, and custom algorithms

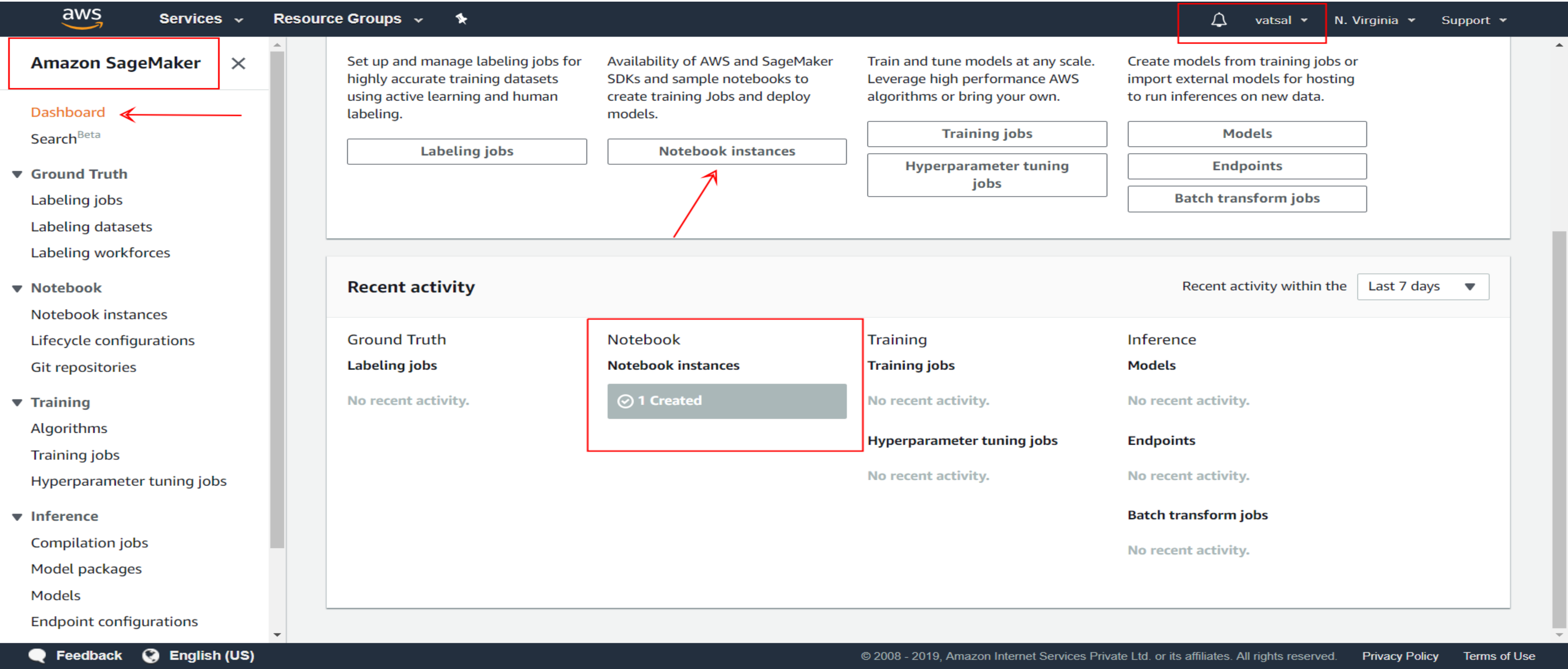
Easier training with hyperparameter optimization

Deploy

Deployment without engineering effort

Fully-managed hosting at scale

Dashboard



Set up and manage labeling jobs for highly accurate training datasets using active learning and human labeling.

Availability of AWS and SageMaker SDKs and sample notebooks to create training Jobs and deploy models.

Train and tune models at any scale. Leverage high performance AWS algorithms or bring your own.

Create models from training jobs or import external models for hosting to run inferences on new data.

Recent activity

Recent activity within the Last 7 days

Ground Truth

Labeling jobs

No recent activity.

Notebook

Notebook instances

1 Created

Training

Training jobs

No recent activity.

Hyperparameter tuning jobs

No recent activity.

Inference

Models

No recent activity.

Endpoints

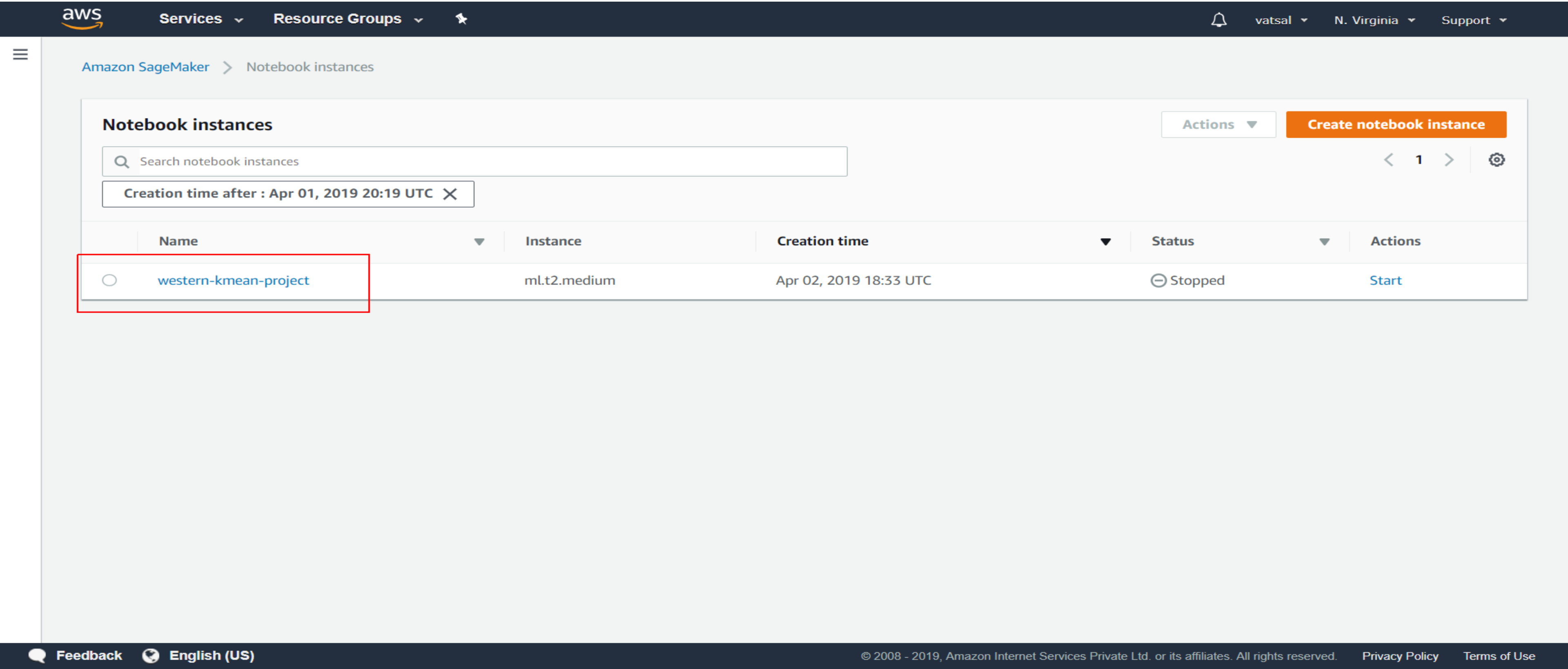
No recent activity.

Batch transform jobs

No recent activity.

© 2008 - 2019, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.

Notebook instances



aws
Services
Resource Groups
vatsal
N. Virginia
Support
Amazon SageMaker > Notebook instances

Notebook instances

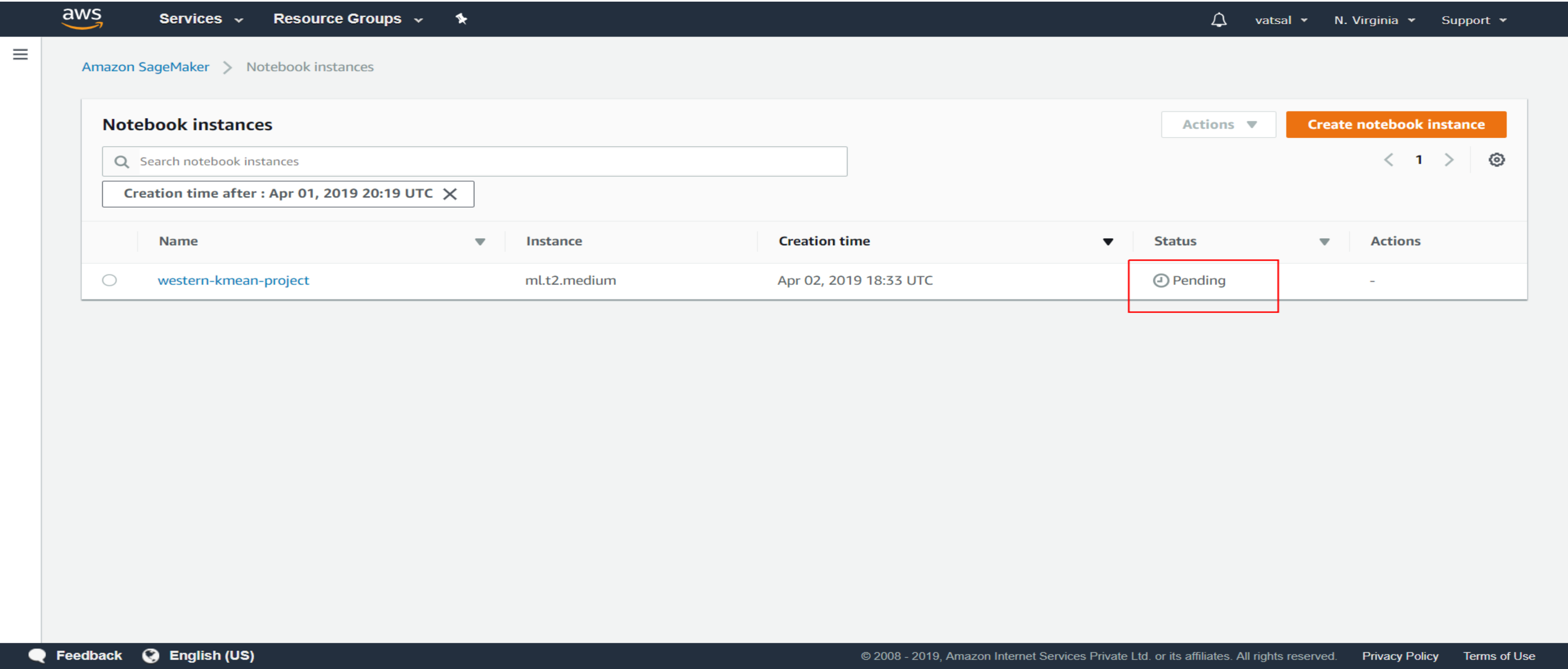
Actions
Create notebook instance
Search notebook instances
1

Creation time after : Apr 01, 2019 20:19 UTC ✕

	Name	Instance	Creation time	Status	Actions
○	western-kmean-project	ml.t2.medium	Apr 02, 2019 18:33 UTC	⊖ Stopped	Start

Feedback
English (US)
© 2008 - 2019, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

Instance in pending status

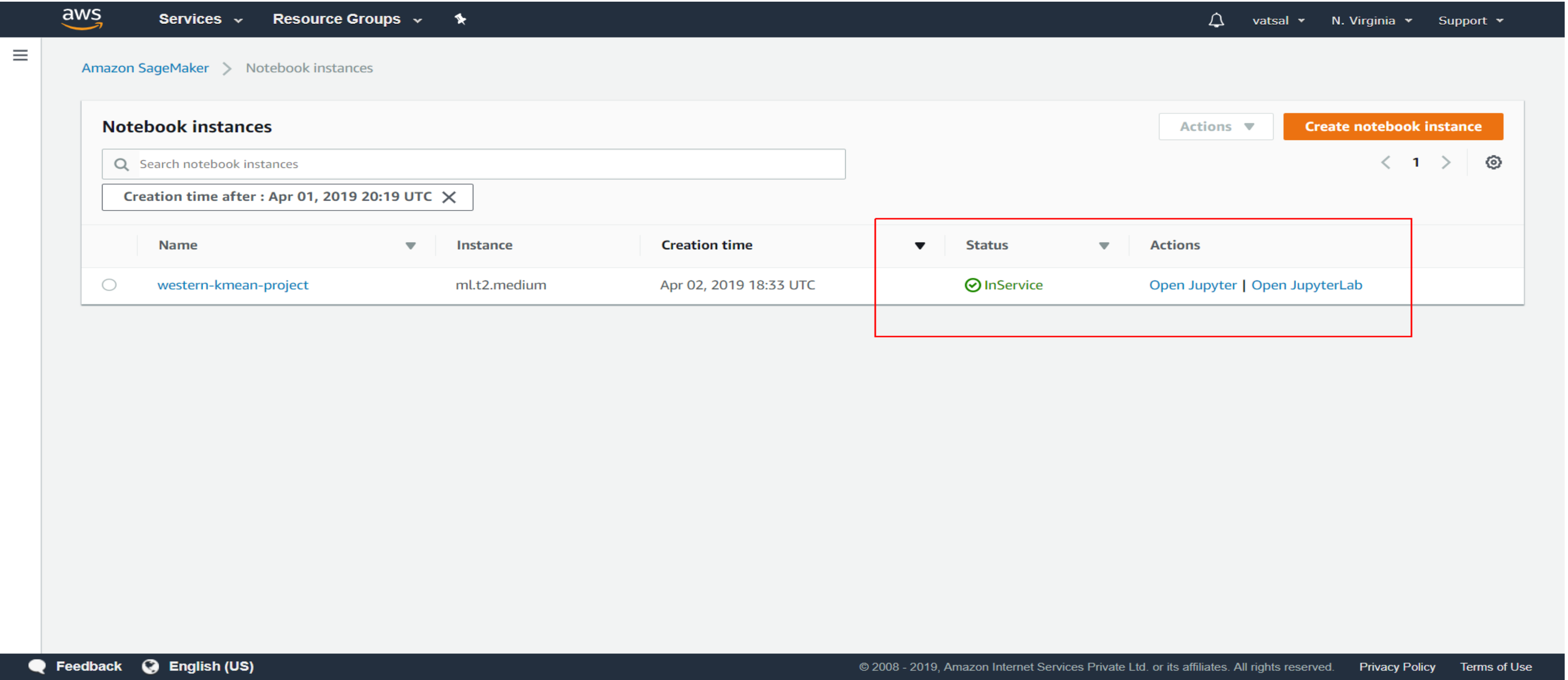


Notebook instances

Creation time after : Apr 01, 2019 20:19 UTC ✕

	Name	Instance	Creation time	Status	Actions
○	western-kmean-project	ml.t2.medium	Apr 02, 2019 18:33 UTC	Pending	-

In Service



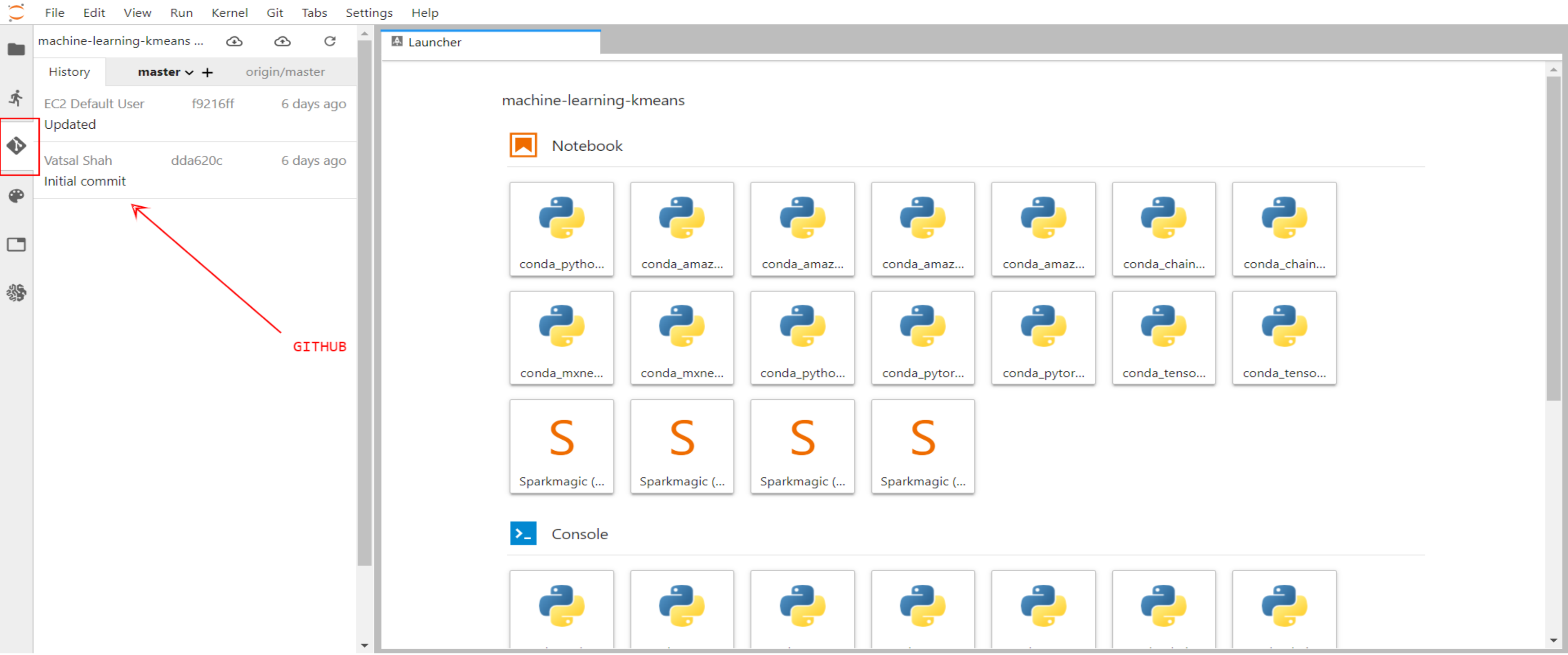
aws
Services
Resource Groups
vatsal
N. Virginia
Support
Amazon SageMaker > Notebook instances
Actions
Create notebook instance
Search notebook instances
1
Feedback
English (US)
© 2008 - 2019, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

Notebook instances

Creation time after : Apr 01, 2019 20:19 UTC

	Name	Instance	Creation time	Status	Actions
○	western-kmean-project	ml.t2.medium	Apr 02, 2019 18:33 UTC	✓ InService	Open Jupyter | Open JupyterLab

Jupyter Lab



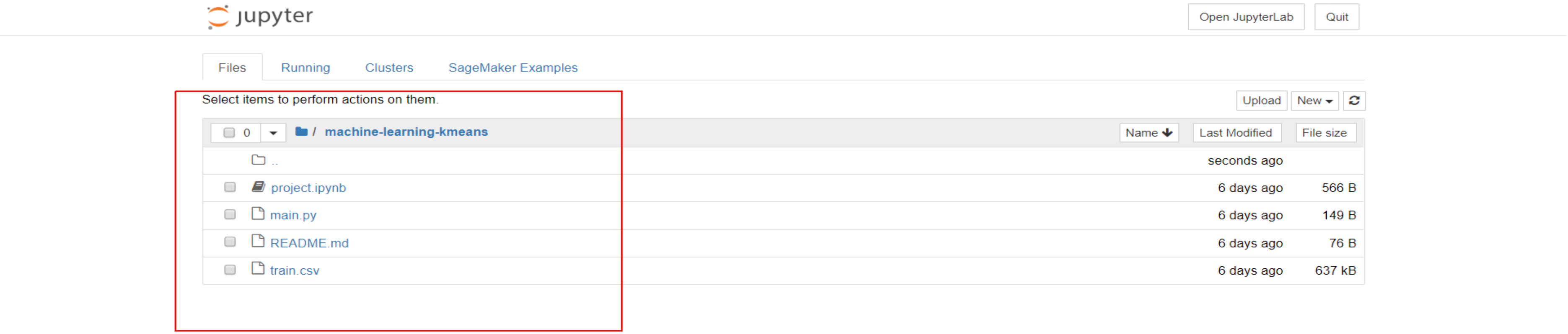
EC2 Default User f9216ff 6 days ago
Updated

Vatsal Shah dda620c 6 days ago
Initial commit

GITHUB

machine-learning-kmeans

Jupyter

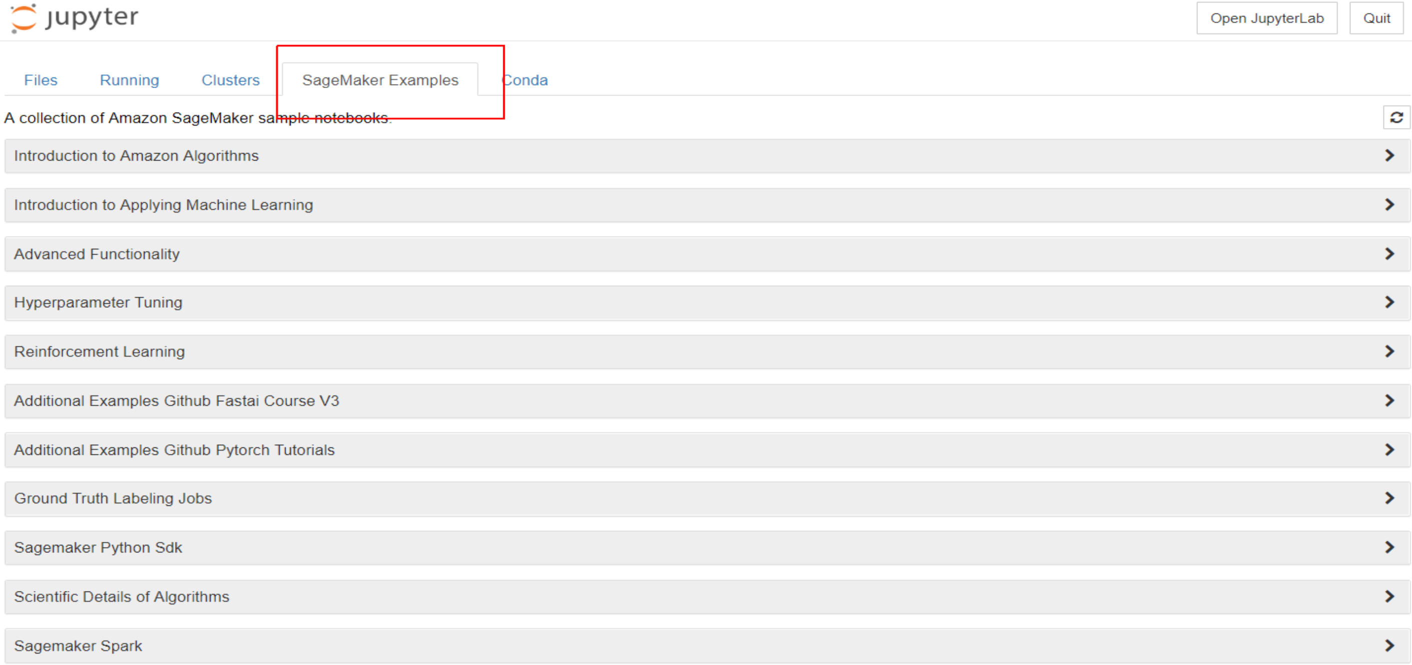


Select items to perform actions on them.

□ 0 / machine-learning-kmeans

Name	Last Modified	File size
..	seconds ago	
□ project.ipynb	6 days ago	566 B
□ main.py	6 days ago	149 B
□ README.md	6 days ago	76 B
□ train.csv	6 days ago	637 kB

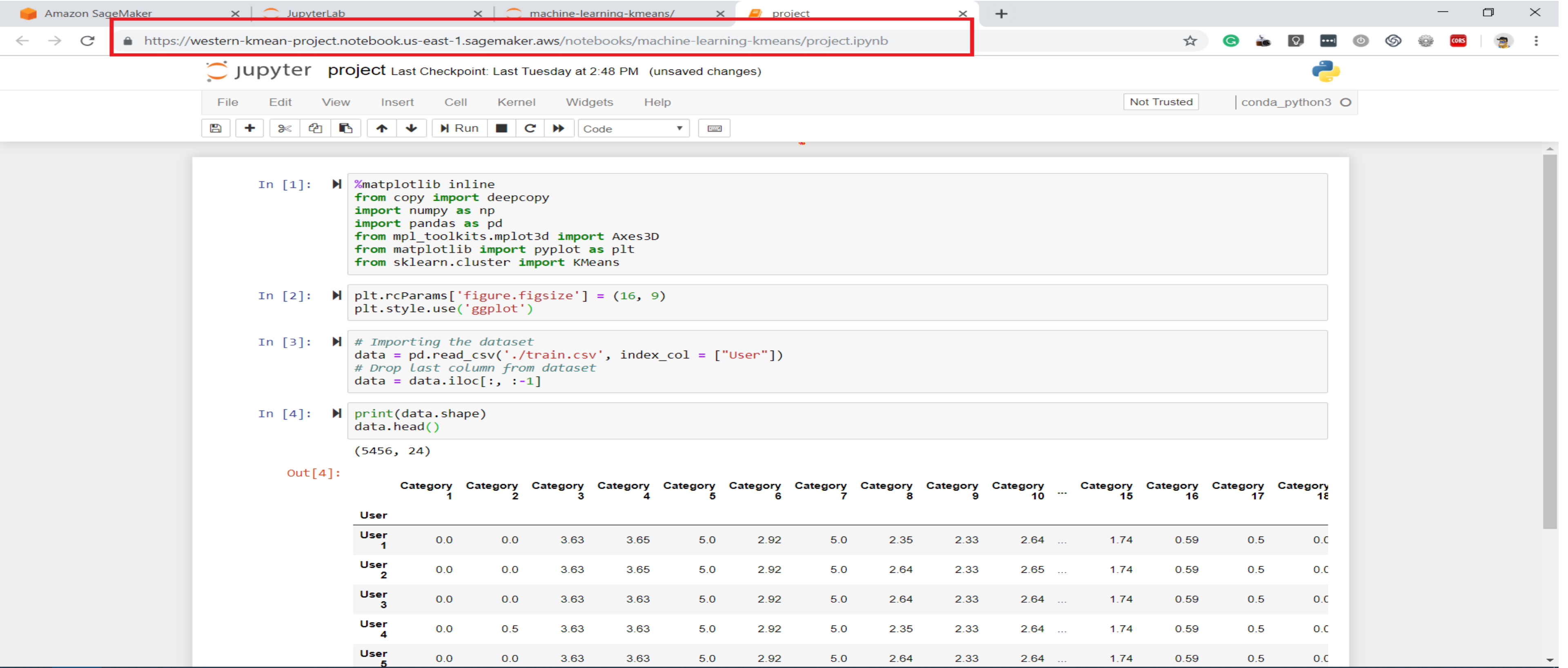
Examples



A collection of Amazon SageMaker sample notebooks.

- Introduction to Amazon Algorithms
- Introduction to Applying Machine Learning
- Advanced Functionality
- Hyperparameter Tuning
- Reinforcement Learning
- Additional Examples Github Fastai Course V3
- Additional Examples Github Pytorch Tutorials
- Ground Truth Labeling Jobs
- Sagemaker Python Sdk
- Scientific Details of Algorithms
- Sagemaker Spark

On AWS



https://western-kmean-project.notebook.us-east-1.sagemaker.aws/notebooks/machine-learning-kmeans/project.ipynb

project Last Checkpoint: Last Tuesday at 2:48 PM (unsaved changes)

In [1]:
```
%matplotlib inline
from copy import deepcopy
import numpy as np
import pandas as pd
from mpl_toolkits.mplot3d import Axes3D
from matplotlib import pyplot as plt
from sklearn.cluster import KMeans
```

In [2]:
```
plt.rcParams['figure.figsize'] = (16, 9)
plt.style.use('ggplot')
```

In [3]:
```
# Importing the dataset
data = pd.read_csv('./train.csv', index_col = ["User"])
# Drop last column from dataset
data = data.iloc[:, :-1]
```

In [4]:
```
print(data.shape)
data.head()
```

```
(5456, 24)
```

Out[4]:

	Category 1	Category 2	Category 3	Category 4	Category 5	Category 6	Category 7	Category 8	Category 9	Category 10	...	Category 15	Category 16	Category 17	Category 18
User															
User 1	0.0	0.0	3.63	3.65	5.0	2.92	5.0	2.35	2.33	2.64	...	1.74	0.59	0.5	0.0
User 2	0.0	0.0	3.63	3.65	5.0	2.92	5.0	2.64	2.33	2.65	...	1.74	0.59	0.5	0.0
User 3	0.0	0.0	3.63	3.63	5.0	2.92	5.0	2.64	2.33	2.64	...	1.74	0.59	0.5	0.0
User 4	0.0	0.5	3.63	3.63	5.0	2.92	5.0	2.35	2.33	2.64	...	1.74	0.59	0.5	0.0
User 5	0.0	0.0	3.63	3.63	5.0	2.92	5.0	2.64	2.33	2.64	...	1.74	0.59	0.5	0.0

Objective & Goal

Approach Design

Machine Learning Model

Amazon SageMaker

Dataset Overview

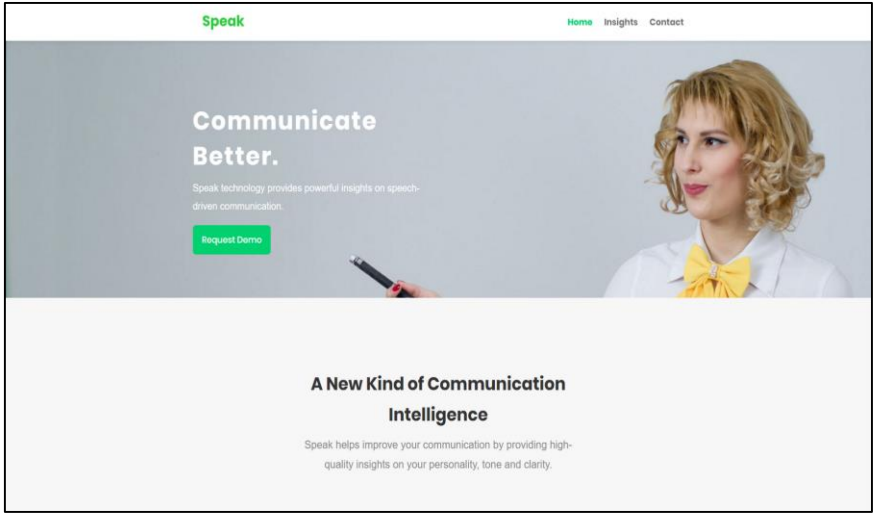
Visualization Overview

Results

Future Work & *QR Code

Technologies Used

Future work



Speak

Communicate Better.

Speak technology provides powerful insights on speech-driven communication.

A New Kind of Communication Intelligence

Speak helps improve your communication by providing high-quality insights on your personality, tone and clarity.

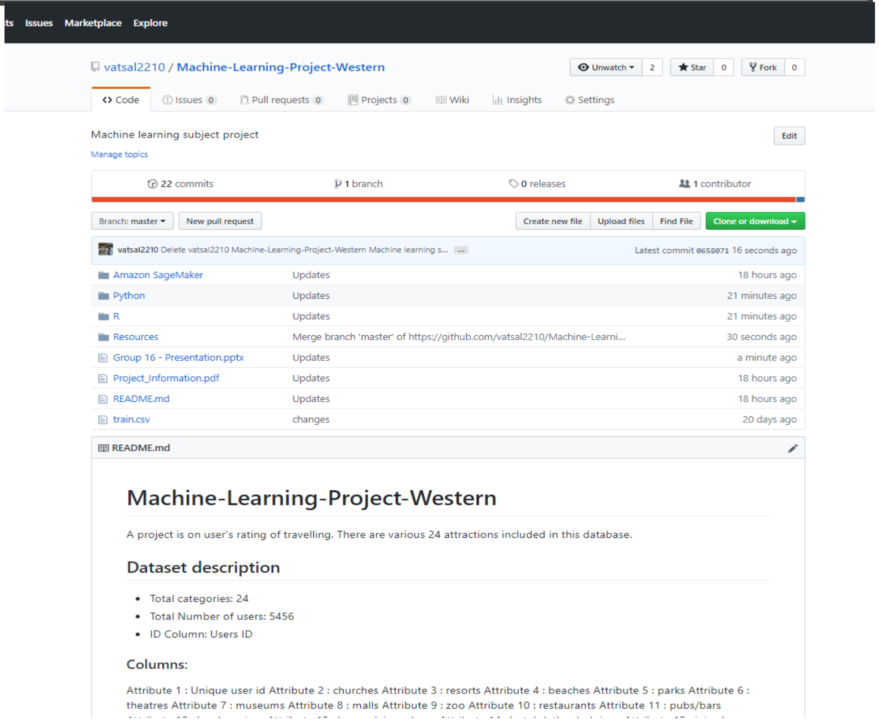
Find — Find more data related to text, audio & video

Explore — Explore more with AWS SageMaker

Apply — Implement for real world problems

QR Code

Access resources and source code



vatsal2210 / Machine-Learning-Project-Western

Machine learning subject project

22 commits | 1 branch | 0 releases | 1 contributor

vatsal2210 Delete vatsal2210 Machine-Learning-Project-Western Machine learning s... … Latest commit 0658071 16 seconds ago

Amazon SageMaker	Updates	18 hours ago
Python	Updates	21 minutes ago
R	Updates	21 minutes ago
Resources	Merge branch 'master' of https://github.com/vatsal2210/Machine-Learni...	30 seconds ago
Group 16 - Presentation.pptx	Updates	a minute ago
Project_Information.pdf	Updates	18 hours ago
README.md	Updates	18 hours ago
train.csv	changes	20 days ago

README.md

Machine-Learning-Project-Western

A project is on user's rating of travelling. There are various 24 attractions included in this database.

Dataset description

- Total categories: 24
- Total Number of users: 5456
- ID Column: Users ID

Columns:

Attribute 1 : Unique user id Attribute 2 : churches Attribute 3 : resorts Attribute 4 : beaches Attribute 5 : parks Attribute 6 : theatres Attribute 7 : museums Attribute 8 : malls Attribute 9 : zoo Attribute 10 : restaurants Attribute 11 : pubs/bars

Resources:

- Getting started with Amazon SageMaker: Link
- Use the Amazon SageMaker SDK: Python: Link
- ‘Evaluation of Partitioning Clustering Algorithms for Processing Social Media Data in Tourism domain’, Shini Renjith, A. Sreekumar, M. Jathavedan, 2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS) | December 06 - 08, 2018 , Trivandrum Link
- ‘Use and Impact of Online Travel Reviews’, Markus Schuckert , Liu XianweiRob Law Link
- ‘Hospitality and Tourism Online Reviews: Recent Trends and Future Directions’, Ulrike Gretzel , Kyung-Hyan Yoo Link
- SageMaker Examples: Link
- Python vs R Comparison: Link
- Python Libraries: Link
- R Libraries : Link
- K-means Algorithm: Link

time for questions

Western
Engineering

THANK YOU