

---

COLLEGE  
RECOMMENDATION  
SYSTEM USING K-MEANS  
ALGORITHM

HARSHITA CHADHA  
SHRUTI GUPTA  
NEELAM SHARMA  
NITISH PATHAK



---

# TABLE OF CONTENTS

**01**

**OBJECTIVES**

**02**

**METHODOLOGY**

**03**

**RESULTS**

**04**

**CONCLUSIONS**

---





**01**

**OBJECTIVES**

---

A horizontal bar at the bottom of the page with a color gradient from black on the left to light gray on the right.

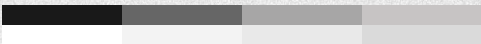
---

---

“74 zettabytes of data will be created in 2021. That's up from 59 zettabytes in 2020 and 41 zettabytes in 2019. (What is a zettabyte? It's a trillion gigabytes.)”

**–Statista**



- 
- Utilizing massive online data and the existing technology provisions to derive useful insights.
  - Collecting and centralizing location data using online datasets and Foursquare API.
  - Applying suitable machine learning technique: Recommendations systems to derive useful insights from created database.
  - Overcoming the disadvantages of existing recommender system frameworks to create a more generalized approach.
  - Creating a rudimentary recommendation system that can prove to be useful for prospective students looking to make decisions about higher education institutions to find universities similar to the ones they already have in mind.
- 
- 

---

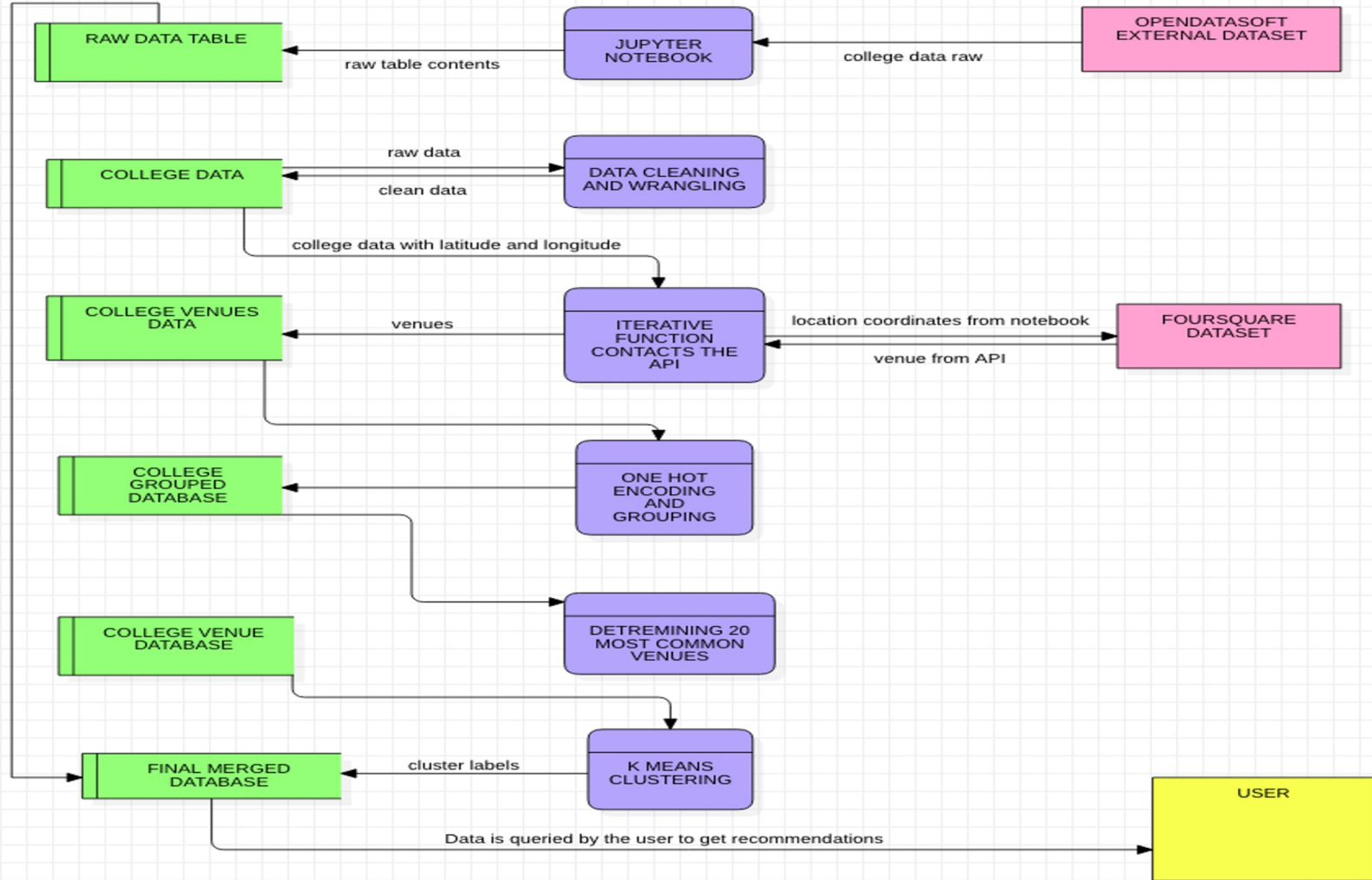
---

# 02

**METHODOLOGY**



college information



---

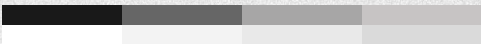
## DATA MINING

- A dataset of US colleges was considered
- A large number of the attributes were irrelevant as a consequence of which the data had to be cleaned and wrangled before it could be used for development purposes.
- From the large data frame, five main attributes were singled out and extracted to a new more relevant database.

## COMBINING WITH FOURSQUARE

- Obtaining college-specific venue information using the foursquare application programming interface (API) to
- About 100 venues per college located within a radius of 500 meters were obtained.
- After the suitable extraction of locations, one hot encoding was applied to the table.

## CLUSTERING

- The K-means clustering algorithm is used to cluster the colleges based on the top 20 most popular venues present in a 500-meter radius around them.
  - The K-means algorithm groups data unsupervised based on the similarity of the data points to each other.
  - The algorithm tries to minimize intra-cluster distances while at the same time attempting to maximize the inter-cluster distances
- 
- 

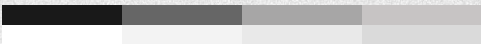


---

## ADDING IDENTIFIERS

- The sum of squared errors (SSE) is used to obtain the optimum value of K.
- The elbow point of this plot is treated as the optimum K value for the given data set. For the present dataset, the value of K was chosen to be 4.
- The cluster labels thus obtained are joined with the data to create comprehensive query point.

## USER INTERFACING

- The joined data is stored as a database which can be queried to get results and evaluate the utility of the undertaking.
  - This database serves as a recommender and based upon the cluster label of the entered location, outputs similarly clustered location data points.
- 
- 



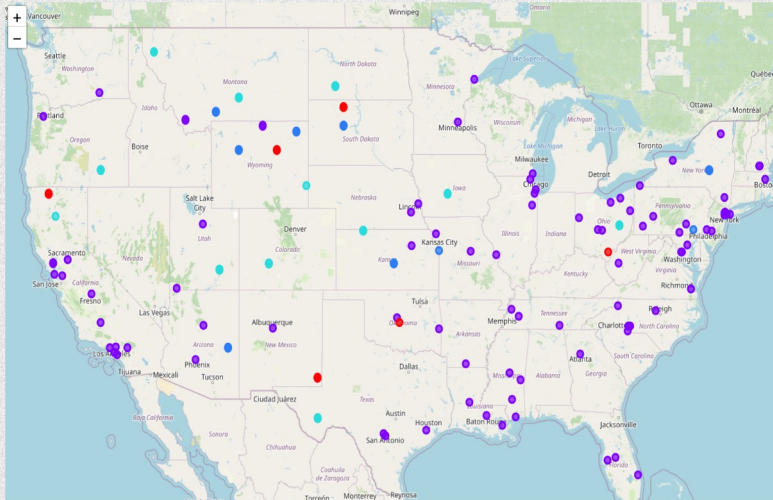
**03**

**RESULTS**

---

A decorative horizontal bar at the bottom of the page, consisting of a black segment on the left, a white segment in the middle, and a grey segment on the right.

# VISUALISATION



The map helps understand the distribution of different types of colleges segmented based on venue categories better. It can be inferred that the purple markers indicate all those data points that have a K label of 1 assigned to them. A deeper examination reveals that hotels and bars are the most popular venue categories for these colleges. This indicates that they are located near a tourist destination.

# USER INTERFACE

Enter the name of the college you are currently interested in...WEST VIRGINIA JUNIOR COLLEGE-CHARLESTON  
The top colleges similar to yours are...

- 1 . FORTIS INSTITUTE-PORT SAINT LUCIE
- 2 . COLUMBIA BASIN COLLEGE
- 3 . YORK TECHNICAL COLLEGE
- 4 . DYERSBURG STATE COMMUNITY COLLEGE
- 5 . SEWANEE-THE UNIVERSITY OF THE SOUTH

The kind of venues you will find around this institution most popularly are:

- 1 . Bar
- 2 . Sports Bar
- 3 . Bank
- 4 . Pizza Place
- 5 . American Restaurant
- 6 . Café

The user is prompted to enter the name of an institution they are already partial towards. The parser looks for the institution in the database and identifies its cluster label. Subsequently, the program outputs a list of the 5 most similar colleges based on cluster segmentation labels. Further, the most common venues around such an institution are also listed out for further information.


---

---

# 04

**CONCLUSIONS**



- 
- 
- The project utilizes an unsupervised learning based, restriction-free methodology to utilize data on the internet that is otherwise seemingly of no proper utility and using external tools and machine learning transforms it into useful information.
  - The results obtained, reinforce our initial assumption: the application of powerful and versatile technology to untapped data resources can help obtain exciting results and new insights about the physical phenomenon that can be generalized and used to make helpful predictions.
  - Furthermore, the end product of the undertaking also helps us demonstrate how deviations from traditional recommender methodologies can help overcome the rigidity of their approach and enhance the scope of applicability.
- 
- 

---

---

# THANKS

Do you have any questions?  
[harshitaachadha@gmail.com](mailto:harshitaachadha@gmail.com)  
[shrutiguptaa25@gmail.com](mailto:shrutiguptaa25@gmail.com)

---

