# The Developed and UnderDeveloped Districts in TamilNadu

### 

### Table of Contents

1. Introduction
2. Data acquisition and cleaning
3. Exploratory Data Analysis
4. Model Development
5. Results
6. Discussion
7. Conclusion

**1.Introduction**

This project aims to find the Districts which are developed and underdeveloped in Tamil Nadu. With the help of Four Square Data, The venues in Different districts have been collected. The collected data contains venue category and also give the exact location of the venue. By using this data from Four Square we can do the exploratory analysis of data and group them based on their similarities. based on the venue data the districts have grouped by using a Machine Learning algorithm called K-means Algorithm. By finding the areas which are underdeveloped can help the persons who are trying to open venues in that particular District. And also by finding the areas which are developed can be used to get more insights about the district. We can know more about the district which are underdeveloped and it can give us enough information about the diffrent venues and based on that we can recommand a person to open a venue on that district.

#### 

#### **1.1. Business Problem**

In this project we can find the developed and underdeveloped areas in TamilNadu. this helps to establish a venue in a underdeveloped area so that we can get more profit there and at the same time we don't have competitors. and also by knowing the developed area we can get more insights about the area and it can give us the idea of what type of venue to open in that area.

#### **1.2. Scope**

1. The scope of this project is to create a clustering machine learning model which groups the districts based on there similarities.
2. This project use Datas from the Four Square which gives the information about a location.
3. This project groups the districts based on the venue data from from Four Square

# 

# 2. Data acquisition and cleaning

### 2.1. Data Sources

**2.1.1. List of Districts**

The list of Districts were extracted from wikipedia <https://en.wikipedia.org/wiki/List_of_districts_of_Tamil_Nadu> this lists district and it also has additional information of the districts like area of the district, population density, capital of the district, diffrent Taluks in the district in the district, and Code of the district.

**Example**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| District | Code | Area |  | Population | Taluks |
| Ariyalur | AR | 1949 |  | 754,894 | Ariyalur |
| Chengalpattu | CGL | 2,944.96 |  | 2,556,244 | Chengalpattu |

here only the District name, District code and population are extracted.

**2.1.2. Latitude and Longitude**

The Latitude and Longitude of different district are provided by Python library called **GioPy**. By giving the code and name of the district we can get the Latitude and Longitude of the district. GioPy sometimes gives error during the process so we should be careful by using try-exec in python.

using **GioPy** Latitude and Longitude are extracted

**Example**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| District | Code | Area | Population | Taluks | Latitude | Longitude |
| Ariyalur | AR | 1949 | 754,894 | Ariyalur | 11.076036 | 79.117455 |
| Chengalpattu | CGL | 2,944.96 | 2,556,244 | Chengalpattu | 12.684089 | 79.983637 |

**2.1.3. Four Square Data**

Foursquare is a social location service that allows users to explore the world around them.The Foursquare API allows application developers to interact with the Foursquare platform. The API itself is a RESTful set of addresses to which you can send requests, so there's really nothing to download onto your server. We can explore the venues around us by difining the Latitude and Longitude. It gives the list of Venues. In this project we get the venues of different district and we use it to find similarity between the districts.

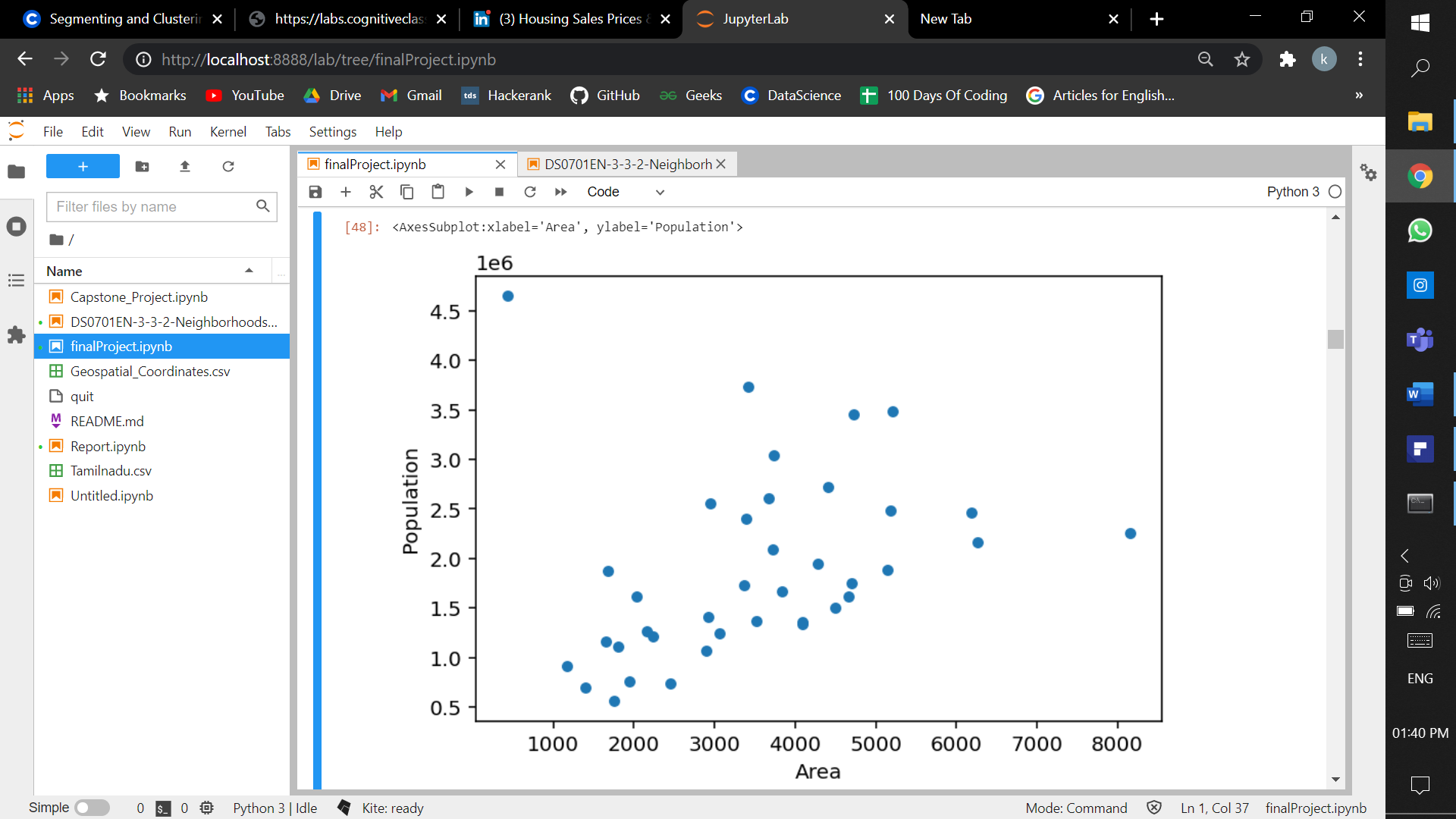
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| No | District | Latitude | Longitude | Venue | venue Latitude | Venue Longitude | Venue Category |
| 0 | Chengalpattu | 12.684089 | 79.983637 | SRK | 12.698709 | 79.970428 | Movie Theater |
| 1 | Chengalpattu | 12.684089 | 79.983637 | Latha Cinemas | 12.680661 | 79.980511 | Movie Theater |
| 2 | Chengalpattu | 12.684089 | 79.983637 | Changalpettu Bus Stand | 12.692468 | 79.979310 | Bus Station |
| 3 | Chengalpattu | 12.684089 | 79.983637 | Chengalpet To Beach Train | 12.693512 | 79.981451 | Light Rail Station |
| 4 | Chengalpattu | 12.684089 | 79.983637 | Kolavai Lake | 12.710869 | 79.980555 | Lake |

### 2.2 Data Cleaning

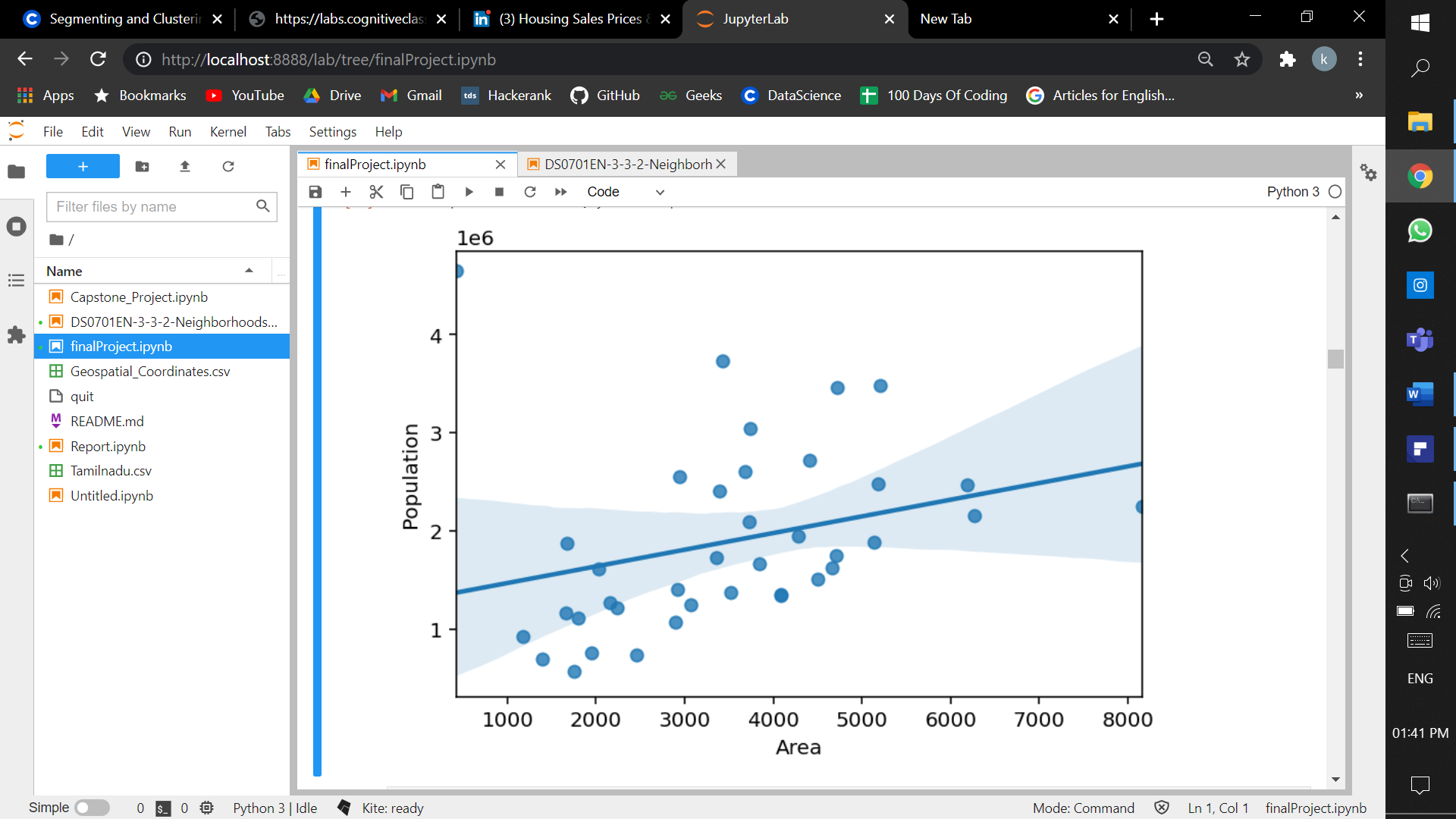
The Data from the wiki-pedia is scraped by using **pandas** library. The scraped data contains various tables. The table which contains the list of Districts is extracted and then converted into pandas **Data Frame**. only the Districts name, Code, population and area features are selected.

# 3. Exploratory Data Analysis

### 3.1. Relationship between Area and Population

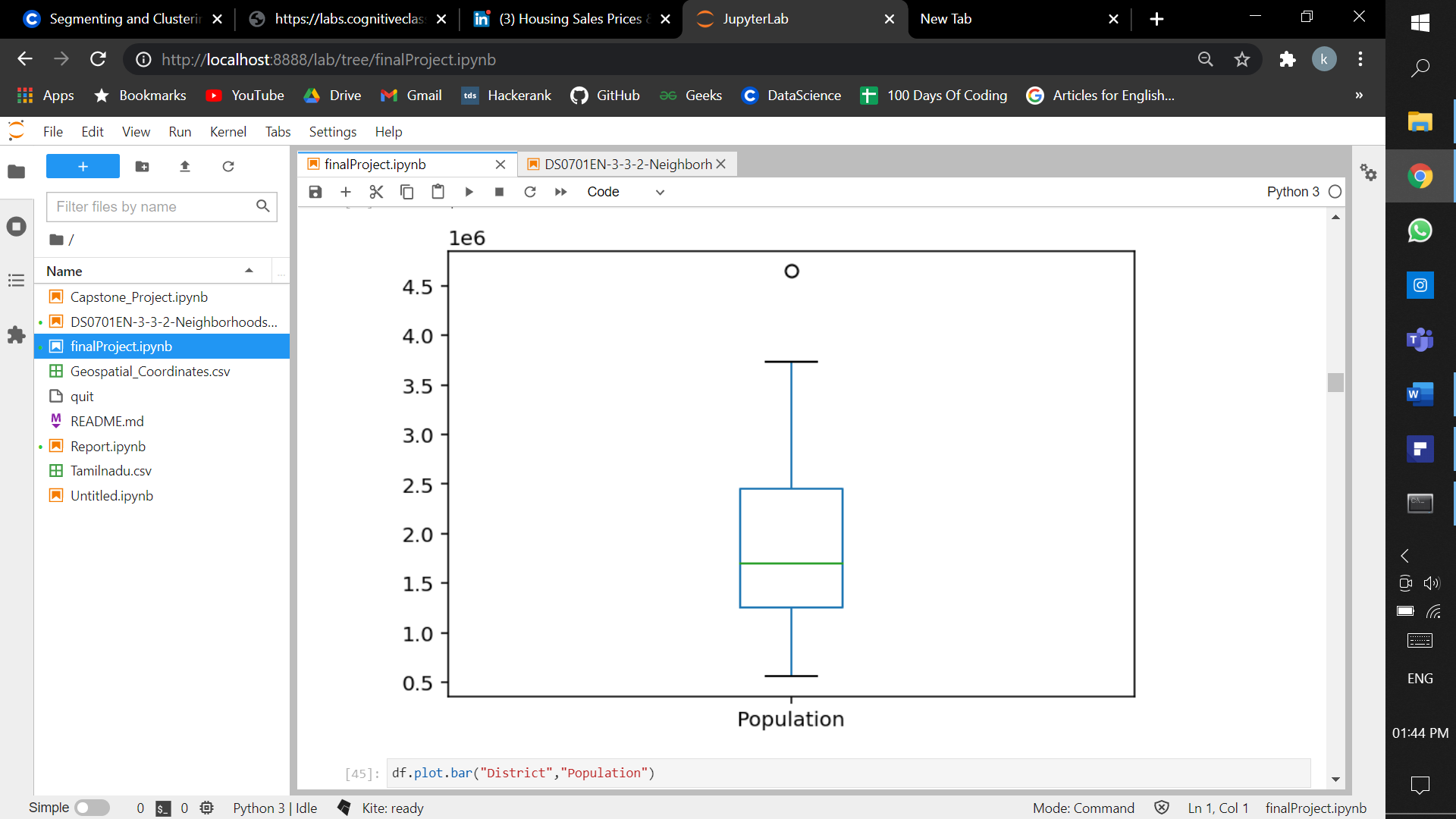


**3.2. Regression between Area and Population**

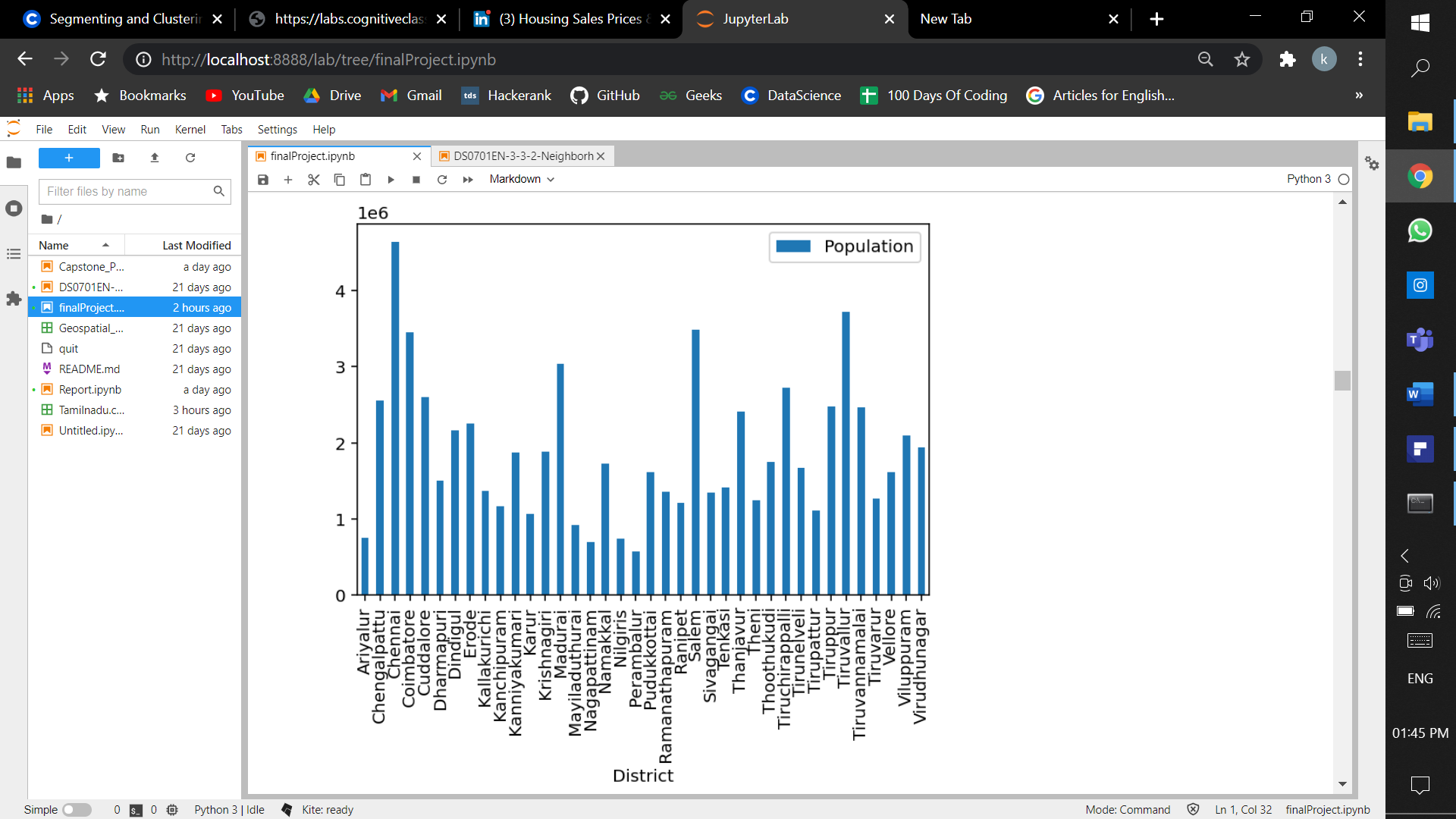
****

### 3.3. Population distribution

Box Plot to find the outliers



Bar Chart to find the distribution of Population between districts



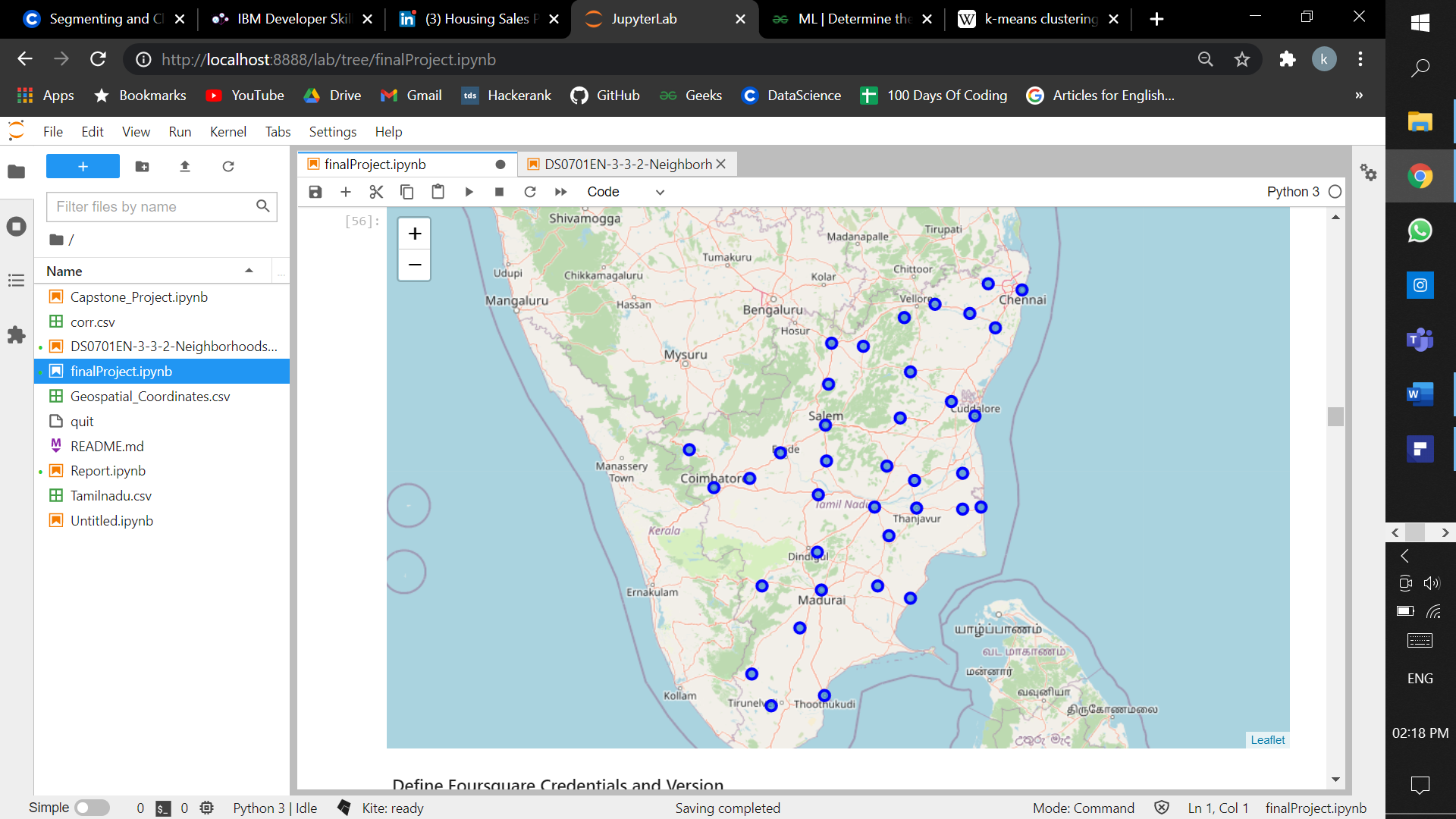
**3.4. Correlation between Features**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Area** | **Population** | **Population\_Density** | **Latitude** | **Longitude** |
| **Area** | 1 | 0.295268 | -0.33469 | -0.13857 | -0.4122 |
| **Population** | 0.295268 | 1 | 0.519458 | 0.222161 | 0.096738 |
| **Population\_Density** | -0.33469 | 0.519458 | 1 | 0.261485 | 0.314549 |
| **Latitude** | -0.13857 | 0.222161 | 0.261485 | 1 | 0.532381 |
| **Longitude** | -0.4122 | 0.096738 | 0.314549 | 0.532381 | 1 |

**3.5. Tamil Nadu Map**

**Folium** is a Python library used for visualizing geospatial data. It is easy to use and yet a powerful library. **Folium** is a Python wrapper for Leaflet.js which is a leading open-source JavaScript library for plotting interactive maps.

It has the power of Leaflet.js and the simplicity of Python, which makes it an excellent tool for plotting maps. Folium is designed with simplicity, performance, and usability in mind. It works efficiently, can be extended with a lot of plugins, has a beautiful and easy-to-use API.

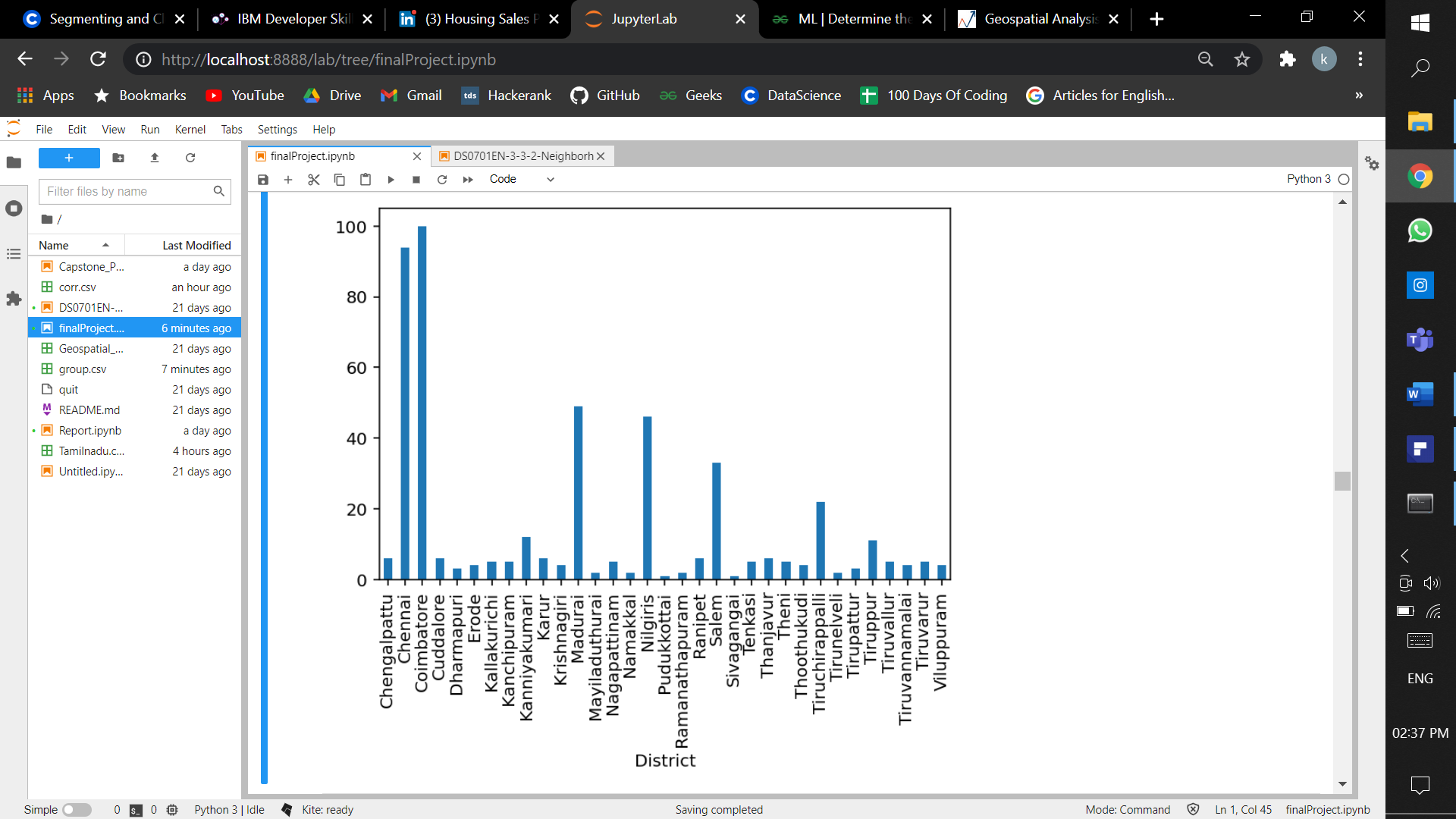


### 3.6. Grouping All the venues by District

Grouping gives the number of venues in the particular district.

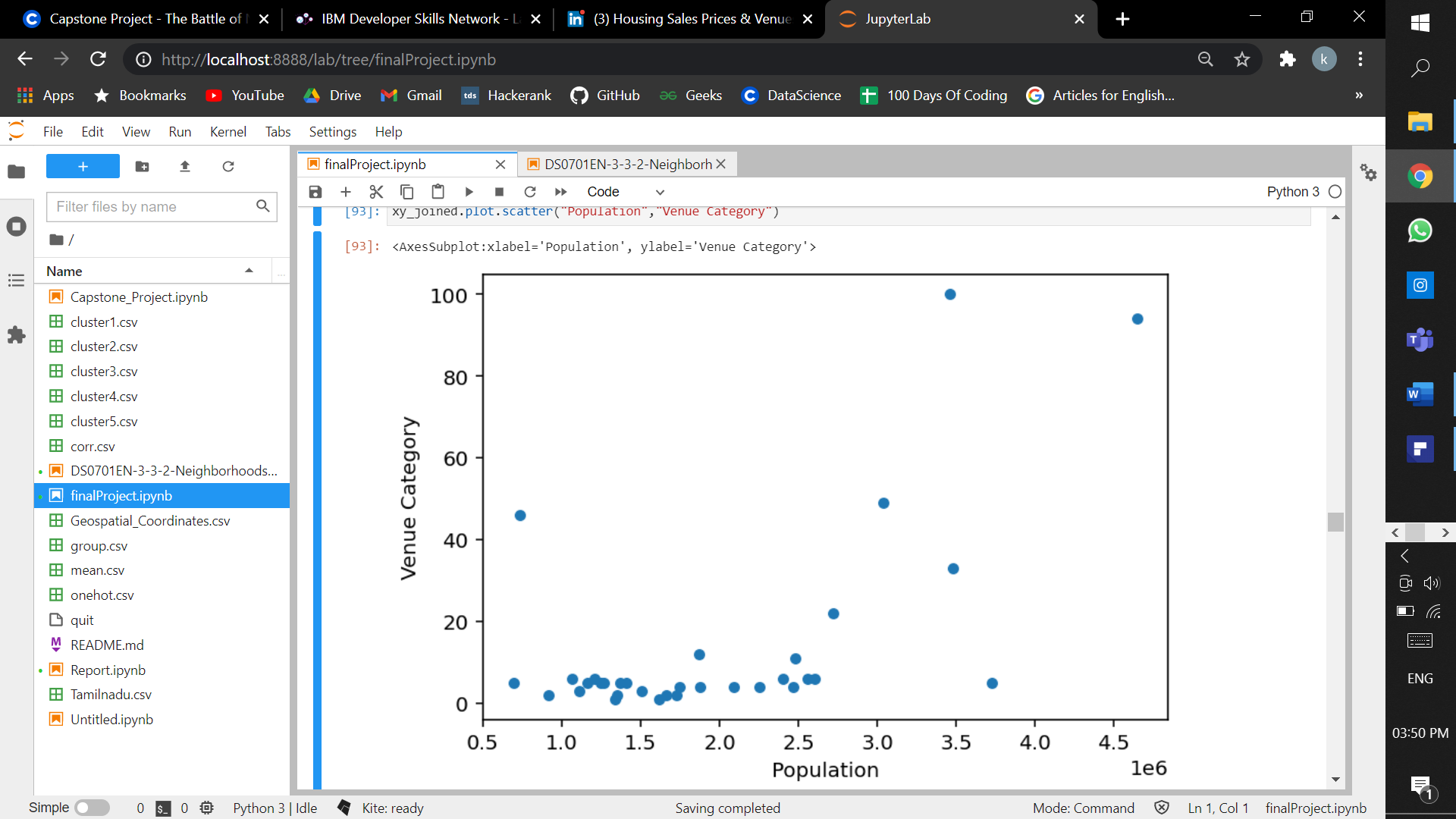
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| District | District Latitude | District Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
| Chengalpattu | 6 | 6 | 6 | 6 | 6 | 6 |
| Chennai | 94 | 94 | 94 | 94 | 94 | 94 |
| Coimbatore | 100 | 100 | 100 | 100 | 100 | 100 |
| Cuddalore | 6 | 6 | 6 | 6 | 6 | 6 |
| Dharmapuri | 3 | 3 | 3 | 3 | 3 | 3 |

**3.7. District vs Number of Venues**

****

### 3.8 Population vs Number of venues

From the below plot we can understand that as the population increases the number of venues in the district also increases hence there is a linier dependecy between this features.



### 3.8 One-Hot Encoding

One-Hot Encoding makes the venues as feature and make them as zero or one based on the presence of the particular venue in that district.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | District | ATM | Accessories Store | African Restaurant | Arcade |
| 10 | Chennai | 1 | 0 | 0 | 0 |
| 11 | Chennai | 0 | 0 | 0 | 0 |
| 12 | Chennai | 0 | 0 | 0 | 1 |
| 13 | Chennai | 0 | 0 | 0 | 0 |
| 14 | Chennai | 0 | 1 | 0 | 0 |

**3.9 Taking mean of the individual venues in the respective District**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | District | ATM | Accessories Store | African Restaurant | Arcade |
| 0 | Chengalpattu | 0 | 0 | 0 | 0 |
| 1 | Chennai | 0 | 0 | 0.010638 | 0 |
| 2 | Coimbatore | 0 | 0.02 | 0 | 0 |
| 3 | Cuddalore | 0 | 0 | 0 | 0 |
| 4 | Dharmapuri | 0 | 0 | 0 | 0 |

**4.Model Development**

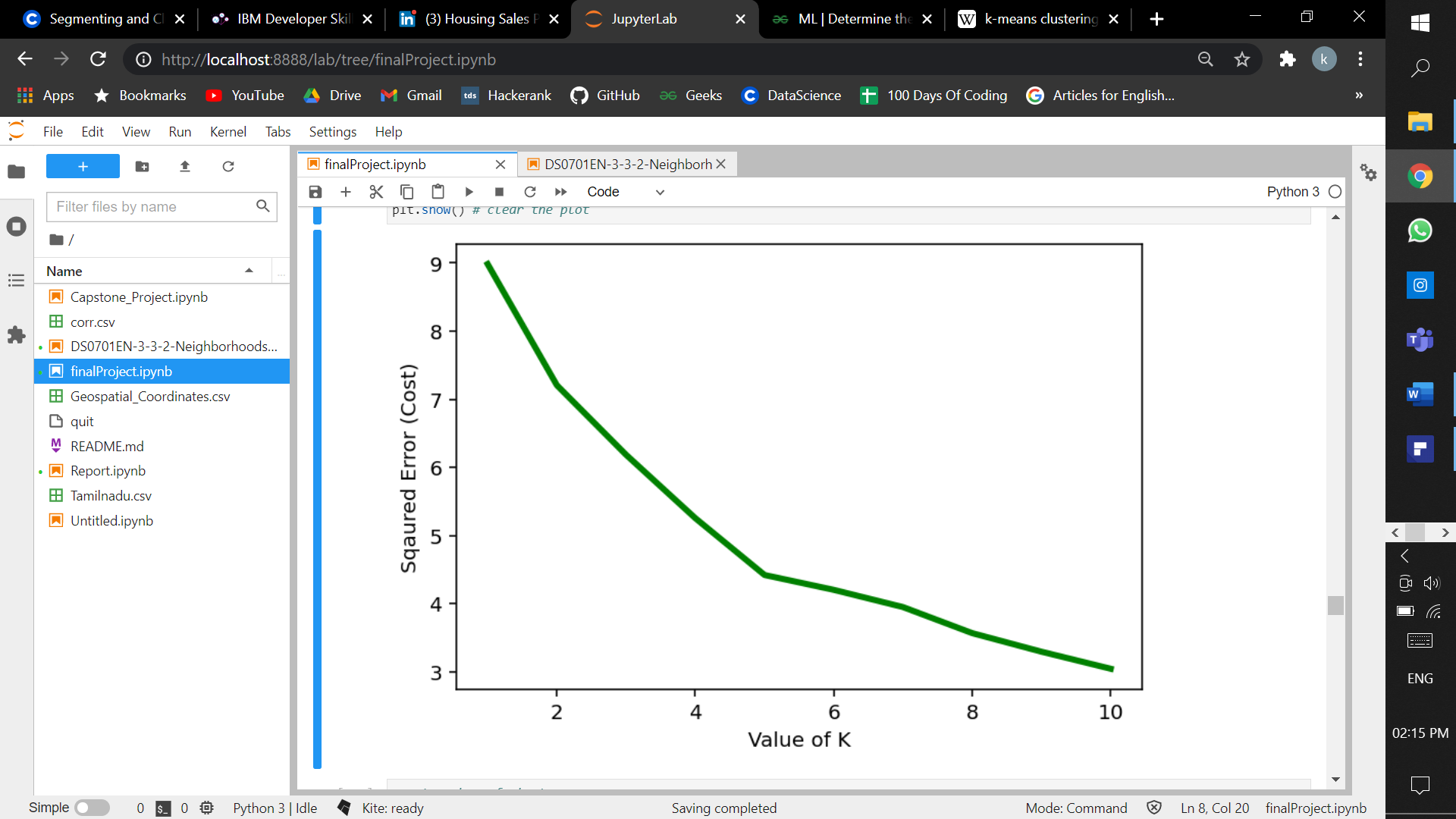
**Clustering model**

**k-means clustering:**

***k*-means clustering** is a method of [vector quantization](https://en.wikipedia.org/wiki/Vector_quantization), originally from [signal processing](https://en.wikipedia.org/wiki/Signal_processing), that aims to [partition](https://en.wikipedia.org/wiki/Partition_of_a_set) *n* observations into *k* clusters in which each observation belongs to the [cluster](https://en.wikipedia.org/wiki/Cluster_(statistics)) with the nearest [mean](https://en.wikipedia.org/wiki/Mean) (cluster centers or cluster [centroid](https://en.wikipedia.org/wiki/Centroid)), serving as a prototype of the cluster. This results in a partitioning of the data space into [Voronoi cells](https://en.wikipedia.org/wiki/Voronoi_cell). *k*-means clustering minimizes within-cluster variances ([squared Euclidean distances](https://en.wikipedia.org/wiki/Squared_Euclidean_distance)), but not regular Euclidean distances, which would be the more difficult [Weber problem](https://en.wikipedia.org/wiki/Weber_problem): the mean optimizes squared errors, whereas only the [geometric median](https://en.wikipedia.org/wiki/Geometric_median) minimizes Euclidean distances. For instance, better Euclidean solutions can be found using [k-medians](https://en.wikipedia.org/wiki/K-medians_clustering) and [k-medoids](https://en.wikipedia.org/wiki/K-medoids).

**Elbow method**

There is a popular method known as **elbow method** which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. The basic idea behind this method is that it plots the various values of cost with changing k. As the value of K increases, there will be fewer elements in the cluster. So average distortion will decrease. The lesser number of elements means closer to the centroid. So, the point where this distortion declines the most is the **elbow point**.



In the above figure there is a decline in the line which corresponds to k-value 5. So for k means algorithm we should take 5 as the k value.

Fit the Model and predict the labels.

There will be 5 clusters.

**5.Results**

There are five clusters. Each clusters has numbers of Districts and based on the Venues it has the districts are clustered. Within a Cluster all District has similar proporties. Cluster 1 has 10 districts and within that Districts the 1st common venue is Indian Restaurants. So this districts has some similarities and hence they are grouped together. The cluster 3 has more number of venues in the District and they are considerd as developed region. On the otherhand Cluster 5 has No venues in the Districts hence the cluster 5 is a under developed region. Cluster 2 has no similarities between the other clusters hence it is grouped into separate cluster.

**Cluster 1**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **District** | **Population** | **1st Most Common Venue** | **2nd Most Common Venue** | **3rd Most Common Venue** | **4th Most Common Venue** | **5th Most Common Venue** |
| 5 | Dharmapuri | 1506843 | Optical Shop | Indian Restaurant | Bank | Electronics Store | Concert Hall |
| 9 | Kanchipuram | 1166401 | Indian Restaurant | Pizza Place | CafÃ© | Bus Station | Women's Store |
| 12 | Krishnagiri | 1879809 | Indian Restaurant | CafÃ© | Women's Store | Fast Food Restaurant | Convenience Store |
| 13 | Madurai | 3038252 | Indian Restaurant | Hotel | Movie Theater | Shopping Mall | Ice Cream Shop |
| 16 | Namakkal | 1726601 | Coffee Shop | Indian Restaurant | Fast Food Restaurant | Convenience Store | Cosmetics Shop |
| 21 | Ranipet | 1210277 | Indian Restaurant | Hotel | Asian Restaurant | Movie Theater | Women's Store |
| 22 | Salem | 3482056 | Indian Restaurant | Ice Cream Shop | Multiplex | Bakery | Shopping Mall |
| 26 | Theni | 1245899 | Indian Restaurant | Bus Station | Waterfall | Fast Food Restaurant | Convenience Store |
| 29 | Tirunelveli | 1665253 | Train Station | Indian Restaurant | Women's Store | Electronics Store | Concert Hall |
| 33 | Tiruvannamalai | 2464875 | Vegetarian / Vegan Restaurant | Resort | Indian Restaurant | CafÃ© | Women's Store |

**Cluster 2**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | District | Population | 1st Most Common Venue | 2nd Most Common Venue | | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
| 19 | Pudukkottai | 1618345 | Sculpture Garden | Women's Store | | Electronics Store | Concert Hall | Convenience Store |
|  |  |  |  | |  | | | | |  |  |  |  |  |  |  |  |

**Cluster 3**

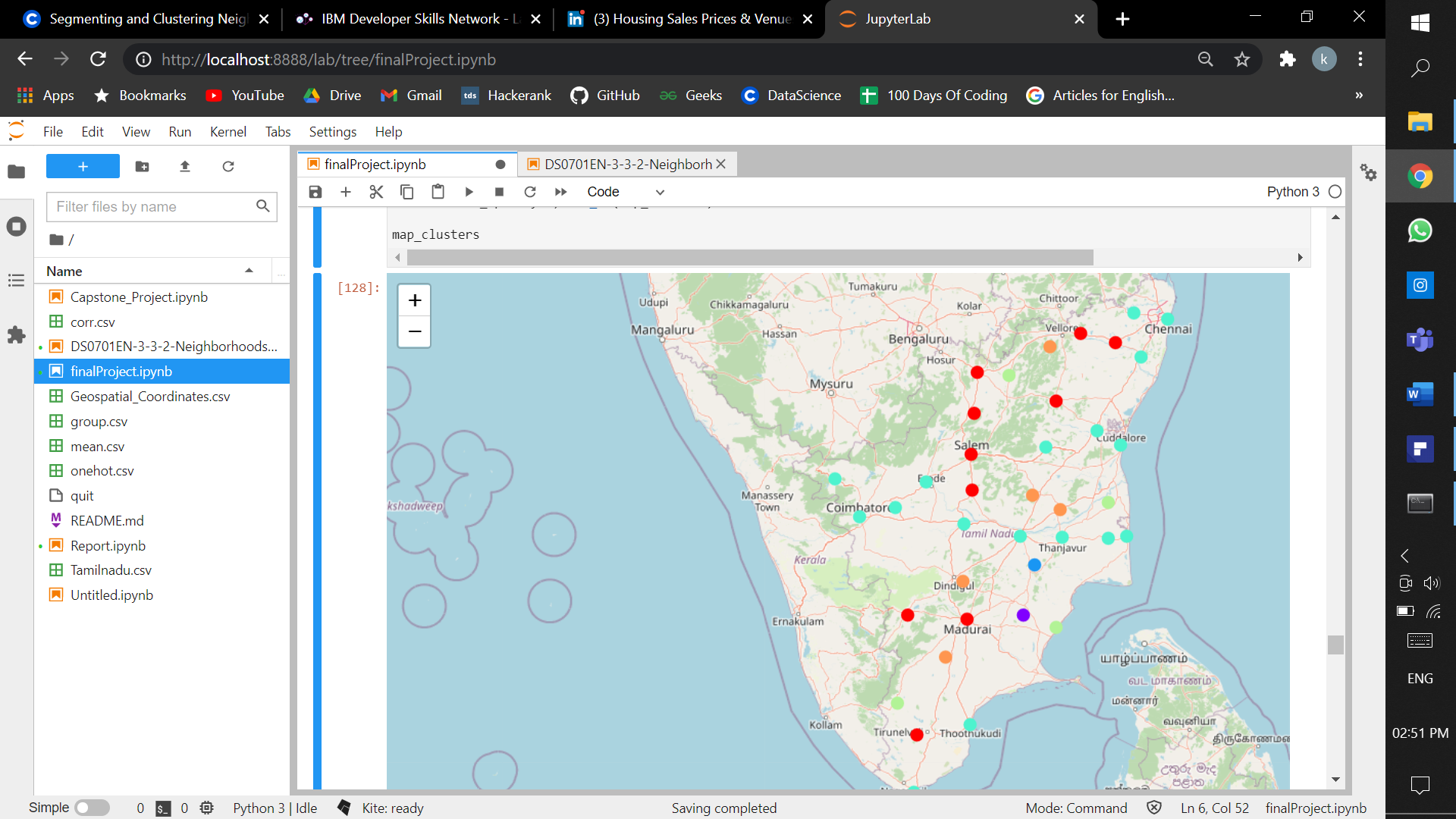
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | District | Population | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
| 1 | Chengalpattu | 2556244 | Light Rail Station | Train Station | Movie Theater | Toll Booth | Bus Station |
| 2 | Chennai | 4646732 | Indian Restaurant | Hotel | CafÃ© | Ice Cream Shop | Fast Food Restaurant |
| 3 | Coimbatore | 3458045 | Indian Restaurant | Clothing Store | Ice Cream Shop | Asian Restaurant | Multiplex |
| 4 | Cuddalore | 2605914 | Movie Theater | Beach | Vegetarian / Vegan Restaurant | Department Store | Women's Store |
| 7 | Erode | 2251744 | Clothing Store | Pizza Place | Food & Drink Shop | Outdoors & Recreation | Diner |
| 8 | Kallakurichi | 1370281 | ATM | Toll Plaza | Indian Restaurant | Restaurant | Bus Station |
| 10 | Kanniyakumari | 1870374 | Historic Site | Beach | Sculpture Garden | Resort | Castle |
| 11 | Karur | 1064493 | Train Station | Hotel | Ice Cream Shop | Food | Bus Station |
| 15 | Nagapattinam | 697069 | ATM | Pharmacy | Cosmetics Shop | Bus Station | Skating Rink |
| 17 | Nilgiris | 735394 | Hotel | Resort | Indian Restaurant | Vegetarian / Vegan Restaurant | CafÃ© |
| 25 | Thanjavur | 2405890 | Historic Site | Museum | Asian Restaurant | Ice Cream Shop | Bus Station |
| 27 | Thoothukudi | 1750176 | Shopping Mall | Harbor / Marina | CafÃ© | Women's Store | Electronics Store |
| 28 | Tiruchirappalli | 2722290 | Indian Restaurant | Hotel | Restaurant | Multiplex | Ice Cream Shop |
| 31 | Tiruppur | 2479052 | Clothing Store | Movie Theater | Bed & Breakfast | Food | Indian Restaurant |
| 32 | Tiruvallur | 3728104 | Historic Site | Train Station | Hotel | Indian Restaurant | Motorcycle Shop |
| 34 | Tiruvarur | 1264277 | Boarding House | Convenience Store | Train Station | Motorcycle Shop | Indie Movie Theater |
| 36 | Viluppuram | 2093003 | Vegetarian / Vegan Restaurant | Costume Shop | Asian Restaurant | Bus Station | Women's Store |

**Cluster 4**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | District | Population | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
| 14 | Mayiladuthurai | 918356 | ATM | History Museum | Concert Hall | Convenience Store | Cosmetics Shop |
| 20 | Ramanathapuram | 1353445 | ATM | Restaurant | Electronics Store | Concert Hall | Convenience Store |
| 24 | Tenkasi | 1407627 | ATM | Indian Restaurant | CafÃ© | Fast Food Restaurant | Convenience Store |
| 30 | Tirupattur | 1111812 | ATM | Train Station | Fast Food Restaurant | Concert Hall | Convenience Store |

**Cluster 5**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | District | Population | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
| 0 | Ariyalur | 754894 | NaN | Nan | Nan | Nan | Nan |
| 6 | Dindigul | 2159775 | NaN | Nan | Nan | Nan | Nan |
| 18 | Perambalur | 565223 | NaN | Nan | Nan | Nan | Nan |
| 35 | Vellore | 1614242 | NaN | Nan | Nan | Nan | Nan |
| 37 | Virudhunagar | 1942288 | NaN | Nan | Nan | Nan | Nan |



**6.Discussion**

Cluster 3 and 5 are Developed and under developed Districts respectively. Hence in cluster 3 there are number of venues and if a person want to open any restaurant or a shop in the Districts of Cluster 3 than there will be less profit, on the other hand Cluster 5 has under developed Districts in it so a person can open any venue there and there will be a great profit.

**Developed Districts:**

|  |
| --- |
| 1. Chengalpattu |
| 1. Chennai |
| 1. Coimbatore |
| 1. Cuddalore |
| 1. Erode |
| 1. Kallakurichi |
| 1. Kanniyakumari |
| 1. Karur |
| 1. Nagapattinam |
| 1. Nilgiris |
| 1. Thanjavur |
| 1. Thoothukudi |
| 1. Tiruchirappalli |
| 1. Tiruppur |
| 1. Tiruvallur |
| 1. Tiruvarur |
| 1. Viluppuram |

**Under Developed Districts:**

|  |
| --- |
| 1. Ariyalur |
| 1. Dindigul |
| 1. Perambalur |
| 1. Vellore |
| 1. Virudhunagar |

1. **Conclusion**

**Tamil Nadu** is a [state](https://en.wikipedia.org/wiki/States_and_union_territories_of_India) in southern India. Its capital and largest city is [Chennai](https://en.wikipedia.org/wiki/Chennai). Tamil Nadu lies in the southernmost part of the [Indian subcontinent](https://en.wikipedia.org/wiki/Indian_subcontinent) and is bordered by the [union territory](https://en.wikipedia.org/wiki/Union_territory) of [Puducherry](https://en.wikipedia.org/wiki/Puducherry_(union_territory)) and the [South Indian](https://en.wikipedia.org/wiki/South_India) states of [Kerala](https://en.wikipedia.org/wiki/Kerala), [Karnataka](https://en.wikipedia.org/wiki/Karnataka), and [Andhra Pradesh](https://en.wikipedia.org/wiki/Andhra_Pradesh). It is bounded by the [Eastern Ghats](https://en.wikipedia.org/wiki/Eastern_Ghats) on the north, by the [Nilgiri Mountains](https://en.wikipedia.org/wiki/Nilgiri_Mountains), the [Meghamalai](https://en.wikipedia.org/wiki/Meghamalai) Hills, and [Kerala](https://en.wikipedia.org/wiki/Kerala) on the west, by the [Bay of Bengal](https://en.wikipedia.org/wiki/Bay_of_Bengal) in the east, by the [Gulf of Mannar](https://en.wikipedia.org/wiki/Gulf_of_Mannar) and the [Palk Strait](https://en.wikipedia.org/wiki/Palk_Strait) on the southeast, and by the [Indian Ocean](https://en.wikipedia.org/wiki/Indian_Ocean) on the south. The state shares a [maritime border](https://en.wikipedia.org/wiki/Maritime_border) with the nation of [Sri Lanka](https://en.wikipedia.org/wiki/Sri_Lanka).

In this project the data from the four square is used to cluster the Districts based on there similarities and it helped us to classify them as Developed and Under Developed Districts. This Classification helps people to know about a particular District and also helps to develop a business in a district.