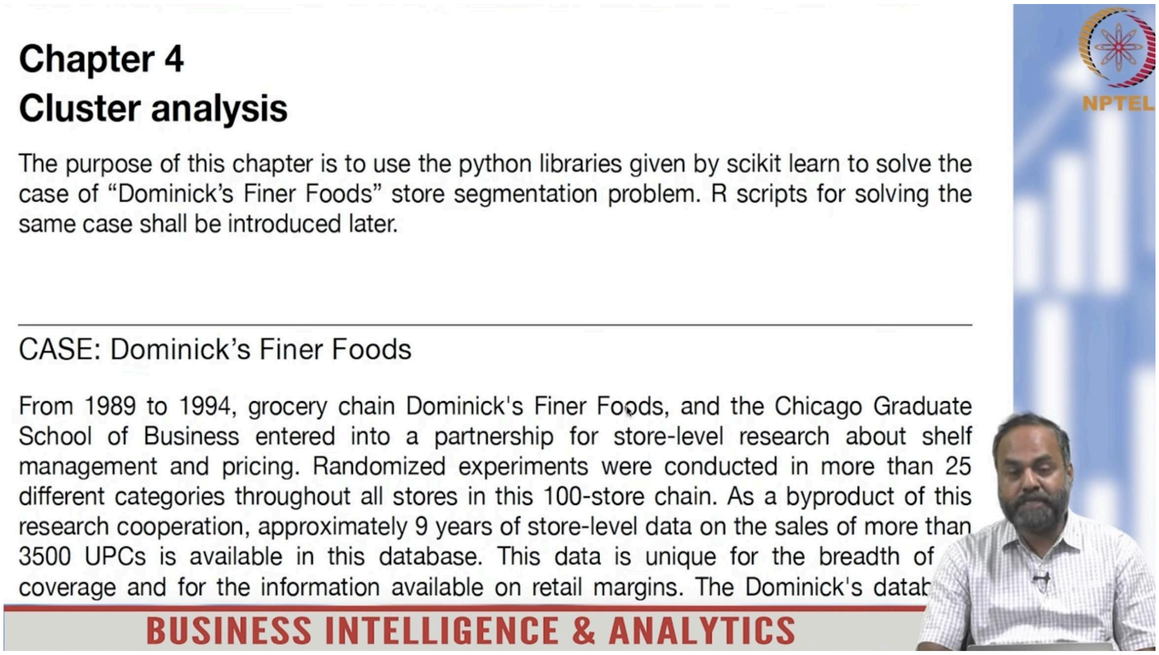
Course Name:Business Intelligence and Analytics
Professor Name:Prof. Saji.K.Mathew
Department Name:Department of Management Studies
Institute Name:Indian Institute of Technology Madras
Week:10
Lecture:36

IMPLEMENTATION IN PYTHON: clustering for segmentation and profiling

Hello and welcome to this class today which is an extension of the concepts we have been learning in the last class on cluster analysis, cluster analysis and also clustering. So these are titled two different ways but we focus on how to apply clustering as a technique, as a modelling technique to model data, particularly when you are looking for patterns of similarity among objects and when that is the purpose. So we want to identify and group objects based on their similarity, clustering is used and clustering is a very powerful and very intuitive method to group objects.



Chapter 4
Cluster analysis

The purpose of this chapter is to use the python libraries given by scikit learn to solve the case of "Dominick's Finer Foods" store segmentation problem. R scripts for solving the same case shall be introduced later.

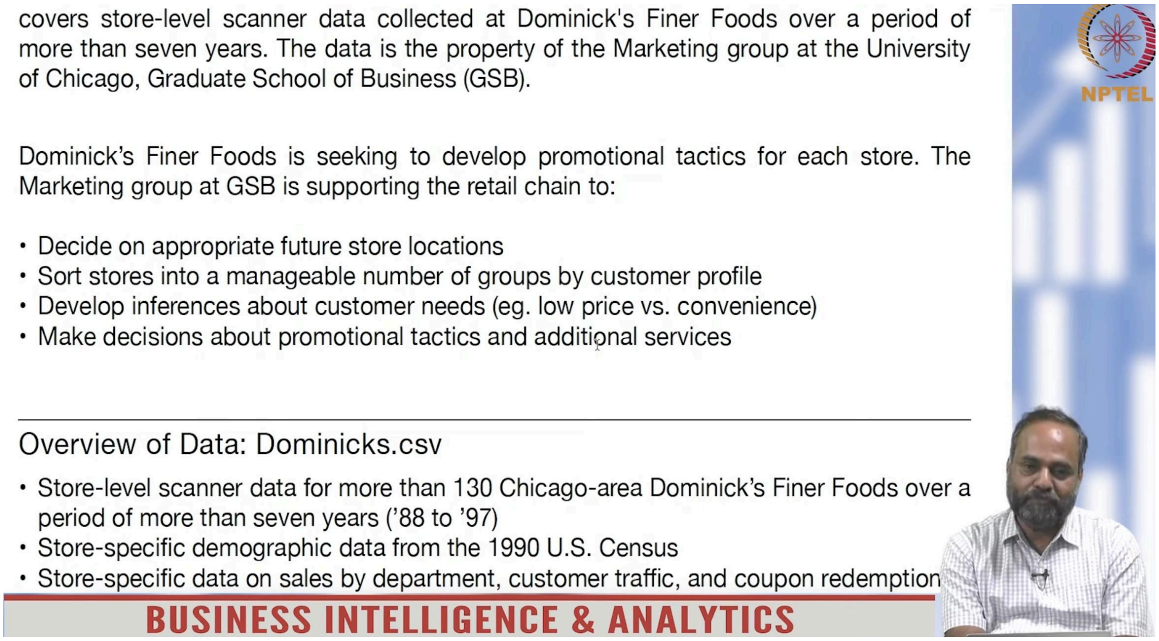
CASE: Dominick's Finer Foods

From 1989 to 1994, grocery chain Dominick's Finer Foods, and the Chicago Graduate School of Business entered into a partnership for store-level research about shelf management and pricing. Randomized experiments were conducted in more than 25 different categories throughout all stores in this 100-store chain. As a byproduct of this research cooperation, approximately 9 years of store-level data on the sales of more than 3500 UPCs is available in this database. This data is unique for the breadth of coverage and for the information available on retail margins. The Dominick's datab

And as we learnt clustering does not require a target variable, clustering only requires variables. And you don't have to group them as dependent variables or independent variables or target variable and explanatory variables, that is not required. Clustering is an algorithm which seeks to bring together objects who are similar, which are similar and create clusters of homogeneity. So as we learnt, clusters are homogeneous within and heterogeneous across.

That is how cluster algorithm works to build clusters. And today in order to understand clustering well and also the application of clustering in business and management well, we are going to discuss a problem. That could be solved using clustering technique and then after understanding the problem, after understanding the data we will use clustering techniques which we discussed in the class to solve the problem in Python, in a Python platform. That is what we are going to do first today. So you can turn to chapter 4 of your exercise book and that is already displayed on the screen which is titled cluster analysis.

So the purpose of this exercise is to help you understand clustering well and explain it to you with the help of a case. This case is a real case and the data is also real life data. I obtained from public domain which was posted by the students of Chicago Graduate School of Business, who did a summer internship like you, like what you do, you do summer internship for large organizations. So they will have specific business objectives. So they did one, these students did for Dominick's Finer Foods.



covers store-level scanner data collected at Dominick's Finer Foods over a period of more than seven years. The data is the property of the Marketing group at the University of Chicago, Graduate School of Business (GSB).

Dominick's Finer Foods is seeking to develop promotional tactics for each store. The Marketing group at GSB is supporting the retail chain to:

- Decide on appropriate future store locations
- Sort stores into a manageable number of groups by customer profile
- Develop inferences about customer needs (eg. low price vs. convenience)
- Make decisions about promotional tactics and additional services

Overview of Data: Dominicks.csv

- Store-level scanner data for more than 130 Chicago-area Dominick's Finer Foods over a period of more than seven years ('88 to '97)
- Store-specific demographic data from the 1990 U.S. Census
- Store-specific data on sales by department, customer traffic, and coupon redemption

I am not sure if it is a real name, it looks like it is camouflage or the real name is hidden but the data is shared, was shared in public domain and therefore we use that data. And the case describes, first the context that the Dominick's Finer Foods is a retail chain or a grocery chain in USA and it has about 100 stores or beyond and so these stores are basically selling grocery and grocery means a lot of items. And there are different categories of sales that happen through grocery stores and the stores are located in different parts of the country. It is not in one city or it is not in one region, it is in different parts of the country. So therefore you can see the business problems or the business problem objectives given to the team are four.

So essentially the problem statement is, Dominick's Finer Foods is seeking to develop promotional tactics for each store. The marketing group of GSB is supporting the retail chain to decide on appropriate future store locations. So predicting appropriate locations for stores, that is one objective. And then the second objective is, sort stores into a manageable number of groups by customer profile, second objective. Third is develop interfaces about, inferences sorry, inferences about customer needs, example low price versus convenience.

And make decisions about promotional tactics and additional services. So these objectives look like very unrelated but they are, if you look at all the objectives because you know it is a project and you know the company has multiple problems, so related to these stores, though they are defining different objectives.

But look at the first one, decide on appropriate future store locations. This is not a problem that we are going to solve using cluster analysis or clustering. Because it is a prediction problem and we have to find out what is an objective function, say maximize sales or profit and then what determines sales or profit and model it. Maybe some dependence technique like regression and then try to predict it. So this is a different problem. This we are not, using clustering technique we cannot address this problem. So we just hold it.

We just do not look at the first objective. I am just presenting the whole case to you but we cannot address all problems using clustering. So we leave the first objective. Look at the second, third and fourth and see if we can address them.

Sort stores into manageable number of groups. That sounds like a grouping problem or a clustering problem. We have to group what? Customers or stores? We have to group customers or we have to group stores. Stores. Please pay attention to that statement. It is store segmentation, not customer segmentation.

So we have to look at store as the unit, not customer as the unit. An object here would be a store, not a customer. So far we were talking about customers or respondents as objects or units. So each clustering will have a unit. What are we grouping? That is the question.

So here, what are we grouping or what is the basic unit? It is a store. So it is a store segmentation problem. So data should be given at the aggregate level, at the store level, not at the customer level. Only then we can do this. So that is point number 1.

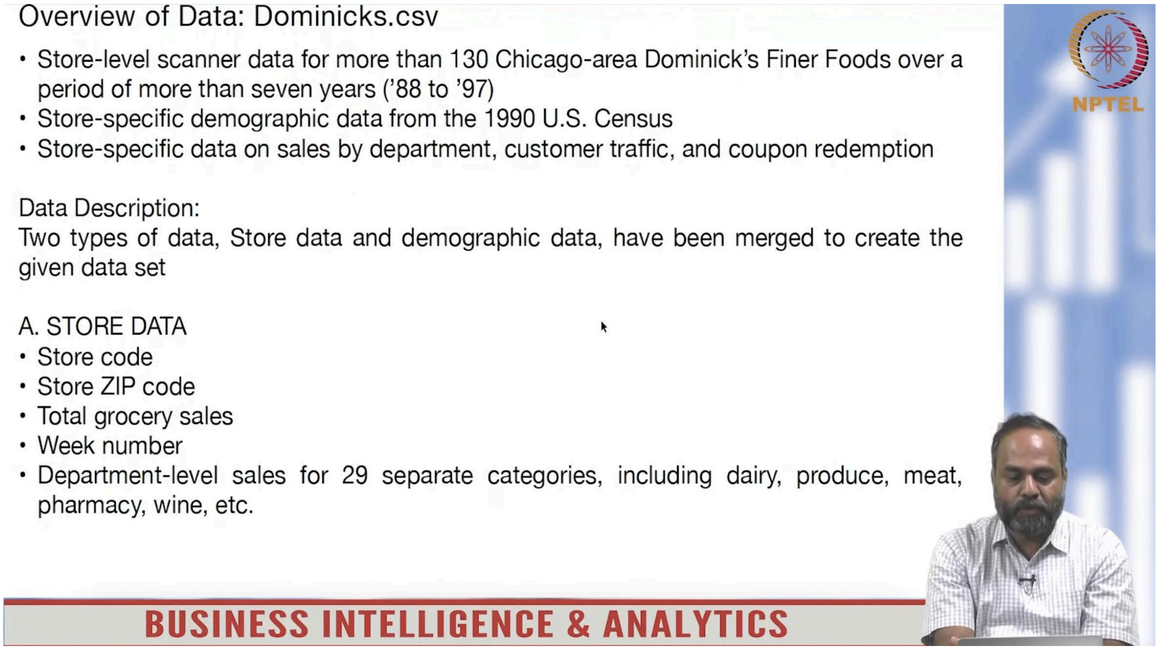
And look at the next objective. Develop inferences about customer needs. Well if you look, if you have done a good segmentation or grouping using cluster analysis of different stores, they have 100 plus stores and then , suppose we have a few manageable groups. That is what they are saying. Few manageable groups. So segmentation is, basically that you convert a large population into a few manageable segments. So look at each segment and look at the characteristics of each segment and understand the groups or segments through their characteristics. So we expect that each segment or each cluster be having certain unique features and they be different from other clusters. That is how clusters are formed. So each cluster is expected to have unique features that define them.

So develop inference about customer needs, looking at the cluster characteristics. And fourth objective, make decisions about promotional tactics and additional services. So essentially the organization is saying, so we cannot personalize, we cannot manage every customer's need. That is personalization. But we can actually group stores into certain clusters and look at what are the common characteristics and then probably develop promotional tactics for a group of stores which are alike.

So that is the plan. So that, you do your promotions in a most efficient way. In a efficient way. So that is the purpose or that is the business problem here. Now how did the students convert this into a analytics problem? Let us look at it. And they asked for data and what data did they get access to? Store level scanner data, scanner data meaning the, yeah, yeah.

So you know the scanner means the barcode or RFID, depending on what is the identifier for items in stores. So each item has identifier and then of course, it is an automated store. There is a POS. And POS data is captured for more than 130 Chicago area, Dominick's Finer Foods over a period of more than 7 years. So what they are receiving is data, sales data of about 130 stores in different parts of that state. For how many years? 7 years data.

Do you feel okay here, to work with this data? It is 7 years data. And look at the next characteristic of the data. Store specific demographic data from the 1990 US Census. So each store has customers and customers byproducts. So there is customer, aggregated customer buying data, because it is store level.



Overview of Data: Dominicks.csv

- Store-level scanner data for more than 130 Chicago-area Dominick's Finer Foods over a period of more than seven years ('88 to '97)
- Store-specific demographic data from the 1990 U.S. Census
- Store-specific data on sales by department, customer traffic, and coupon redemption

Data Description:
Two types of data, Store data and demographic data, have been merged to create the given data set

A. STORE DATA
- Store code
- Store ZIP code
- Total grocery sales
- Week number
- Department-level sales for 29 separate categories, including dairy, produce, meat, pharmacy, wine, etc.

And customers demographic data also, okay. And these are combined or joined into singular records based on store ID, okay. Therefore that data, what is the span of the data of demography of stores? When is the US Census done? It is a census data, 1990 okay. So census data or demographic data corresponds to 1990 but store data or the sales data corresponds to 7 years.

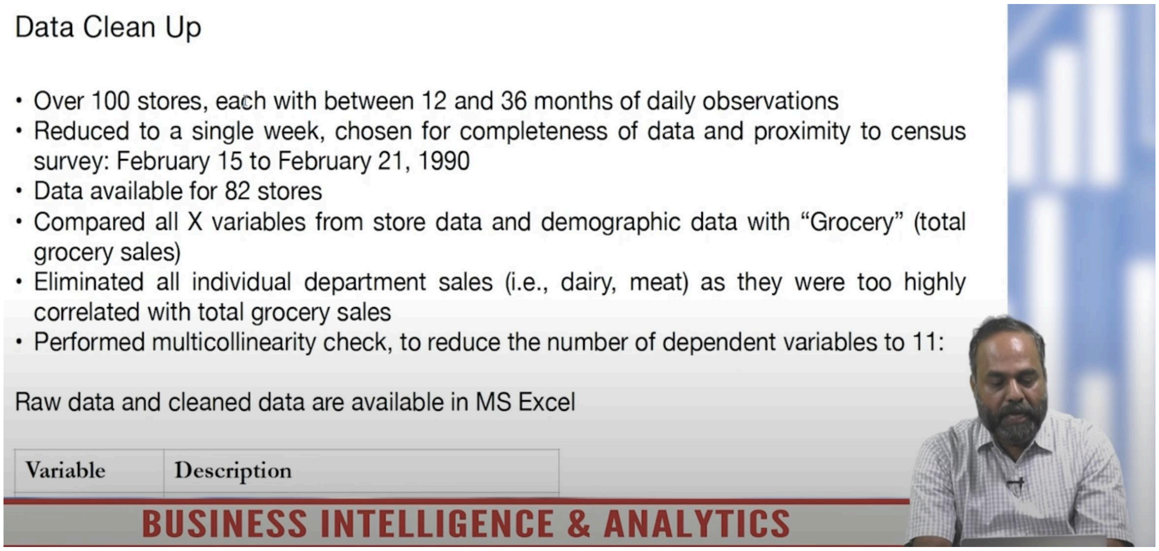
Do you see a problem there? Okay. The data are not synchronized with respect to a timeline. Now here, there is a problem in terms of preparing the data, okay. And that is a challenge you will also face when you work with practical problems. Especially when you work with demographic data which will have a fixed date or a fixed time period when it was collected. But if you are collecting other types of data, like you know sales data or database data, they may be collected over a longer period of time okay.

So these two should match, okay. If they do not match it is not good data, you know it is a data where time will also be available and that is not captured, you know. So it is not time series data that we are working with. And what is this kind of data that we typically use in a exercise like this, like regression or clustering or decision trees etc. We call it cross sectional data, okay. Data collected at the same time, okay from different data points, okay or different sources.

Cross sectional data is what we use. Cross sectional data means from different sections but at the same time. If we collect data from different times, time itself will confound or add variance to the variability to the data. Therefore it is not reliable. Variability in the data should be due to other phenomena with respect to time.

And therefore we have to fix time . Collect data at the same time and then analyze that data. So it should be cross sectional data. But this is not cross sectional data because data is collected over 7 years. Pay attention to this. Always ask a question if you are working on problems like this, what is the duration of data collection.

If it is spanning years you should actually fix it to a shorter time period okay. It does not have to be one day or one hour. But you see how they have prepared the data. Then the third aspect was store specific data on sales by department, customer traffic etc. So when they looked at the data, there were two types of data, store data and demographic data, having been merged to create the given data set.



Data Clean Up

- Over 100 stores, each with between 12 and 36 months of daily observations
- Reduced to a single week, chosen for completeness of data and proximity to census survey: February 15 to February 21, 1990
- Data available for 82 stores
- Compared all X variables from store data and demographic data with “Grocery” (total grocery sales)
- Eliminated all individual department sales (i.e., dairy, meat) as they were too highly correlated with total grocery sales
- Performed multicollinearity check, to reduce the number of dependent variables to 11:

Raw data and cleaned data are available in MS Excel

Variable	Description

BUSINESS INTELLIGENCE & ANALYTICS

The data consists of different attributes like what is mentioned there. And then you can look at how this project group worked and cleaned the data or prepared the data for analysis. The first step was data cleanup over 100 stores each with between 12 to 36 months of daily observations. Reduced to a single week chosen for completeness of data and proximity to census survey February 15 to February 21. They actually followed a very very clear timeline not even the guidance I gave, because in survey data when we work with it typically data collection period is about 6 months.

You cannot collect all the survey data in a day or in a week. So we just let survey data be analyzed, although the time period if it extends for longer period would bring challenges. But these researchers they filtered out data. And in that process, what is the tradeoff that happens? You lose a lot of data, right. You lose a lot of data but that is called editing or scissoring.

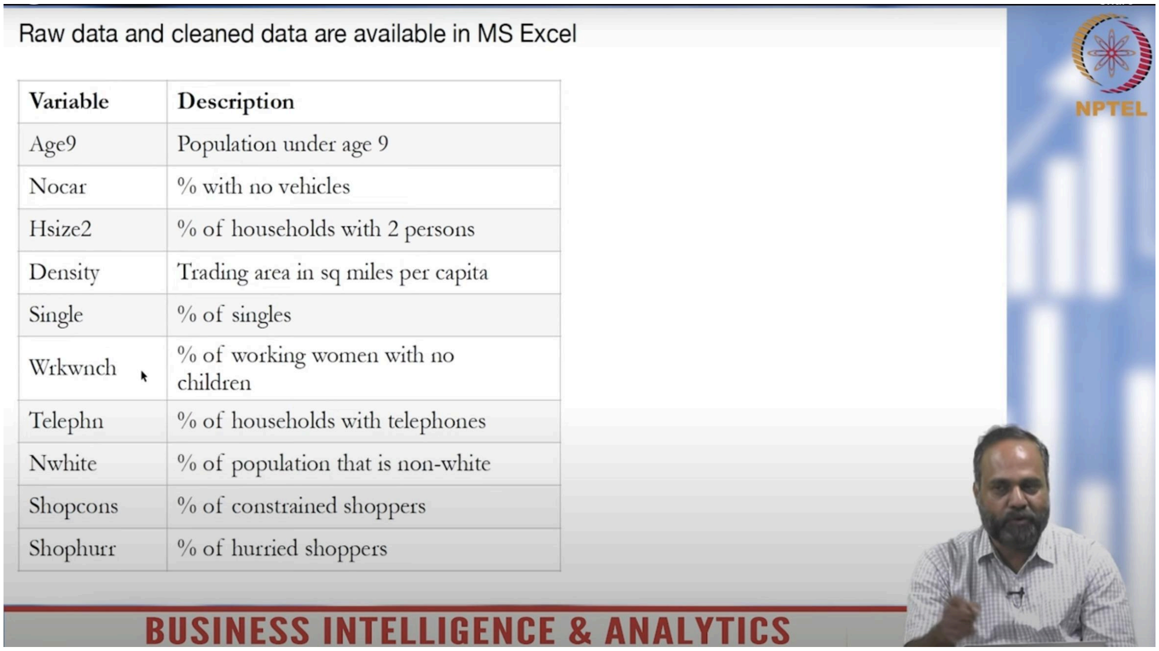
You have to be ruthless in doing that when you work on data analysis. You may be losing a lot of data but that data would only add noise to your analysis because they do not pertain to the same time period. So they aligned the data with a month, sorry a week where of course they looked at completeness of data and other characteristics where what is the best week in 1990 to take the data, okay. So that is an important lesson you learn from this in data preparation for analytics, fixing the timeline of the data, making it strictly cross sectional data when the data is dispersed over long period of time.

Then you can see that they worked on auto correlation of data, compared all x variables from store data and demographic data with grocery and eliminated all individual department sales as they were too highly correlated with total grocery sales. So they check for multicolinearity. So in this kind of sales a lot of sales individual item sales will be correlated. And they may be strongly correlated with the total sales. So they eliminated all those variables and checked multicolinearity.

And finally, you can see the number of variables were reduced to 11. So they reduced the number of variables and they it also led to reduction in the size of data because they made it strictly cross sectional data. So the final variables that emerged are described here in the case, that is there but they have the total sales, right. They have, let us look at each data and see the total sales is also captured, if you look at the original data, I am coming to that. But not individual item wise data because that is correlated with total grocery sales.

And the purpose is to segment stores, not individuals. But aggregate individual characteristics are also captured. So this is aggregate level, store level variables, keep that in mind. So population of customers under age 9, age 9, it is population, say demography. Now percentage with no vehicle, so percentage of customers with no vehicle, the variable is labeled, no car.

Percentage of households with two persons, trading area in square miles per capita, so store concentration in the region. Percentage of singles percentage of working women with no children, percentage of households with telephones, percentage of population that is non-white, sorry, this looks like little racial profiling but they do collect this kind of data. See those companies which are committed to the practice of analytics, they collect a lot of customer level data. You see this may be for programs, loyalty programs through which they actually give questionnaires and collect this data. And then customer IDs are available, so therefore aggregate they, they aggregate the data based on these characteristics as well.



Raw data and cleaned data are available in MS Excel

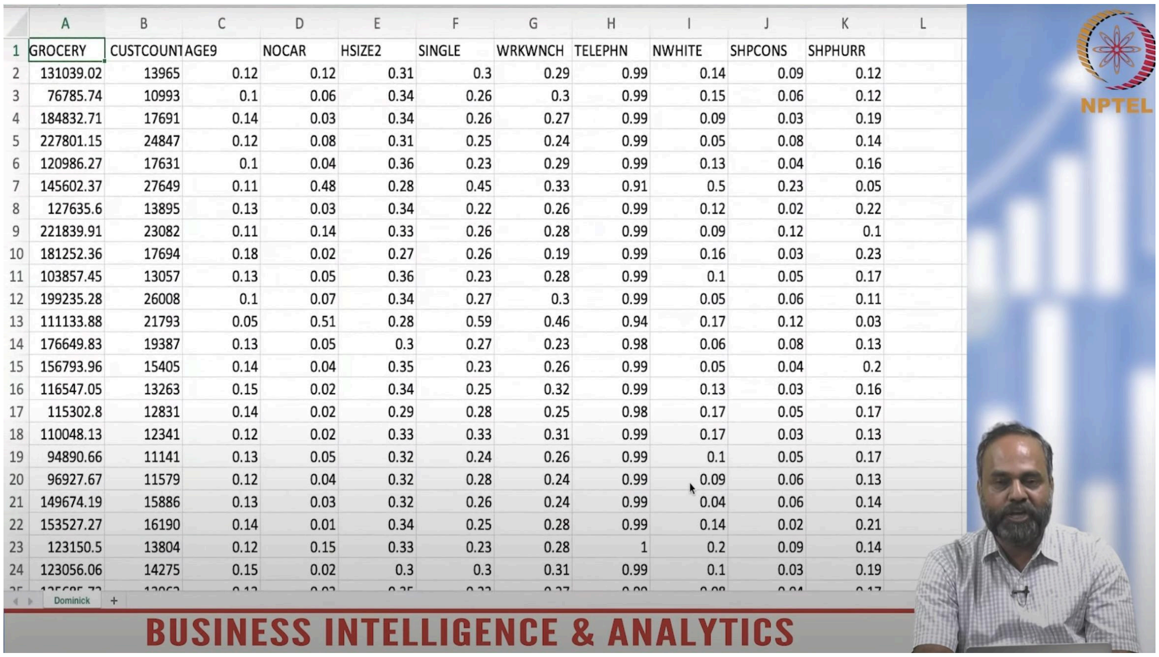
Variable	Description
Age9	Population under age 9
Nocar	% with no vehicles
Hsize2	% of households with 2 persons
Density	Trading area in sq miles per capita
Single	% of singles
Wrkwnch	% of working women with no children
Telephn	% of households with telephones
Nwhite	% of population that is non-white
Shopcons	% of constrained shoppers
Shophurr	% of hurried shoppers

BUSINESS INTELLIGENCE & ANALYTICS

Percentage of population that is non-white. Percentage of constrained shoppers, percentage of hurried shoppers. There is an article called the science of shopping which you can access in the public domain, so to understand the the depth of analytics in retail particularly in North America, where they track a lot of characteristics of customers . A lot of hidden cameras do, are placed in retail stores. They track where you enter, where you turn to, which product you take up and how long you actually look at the different features of the product and how long you spend in a store. These are all sampled and captured for the purpose of analytics.

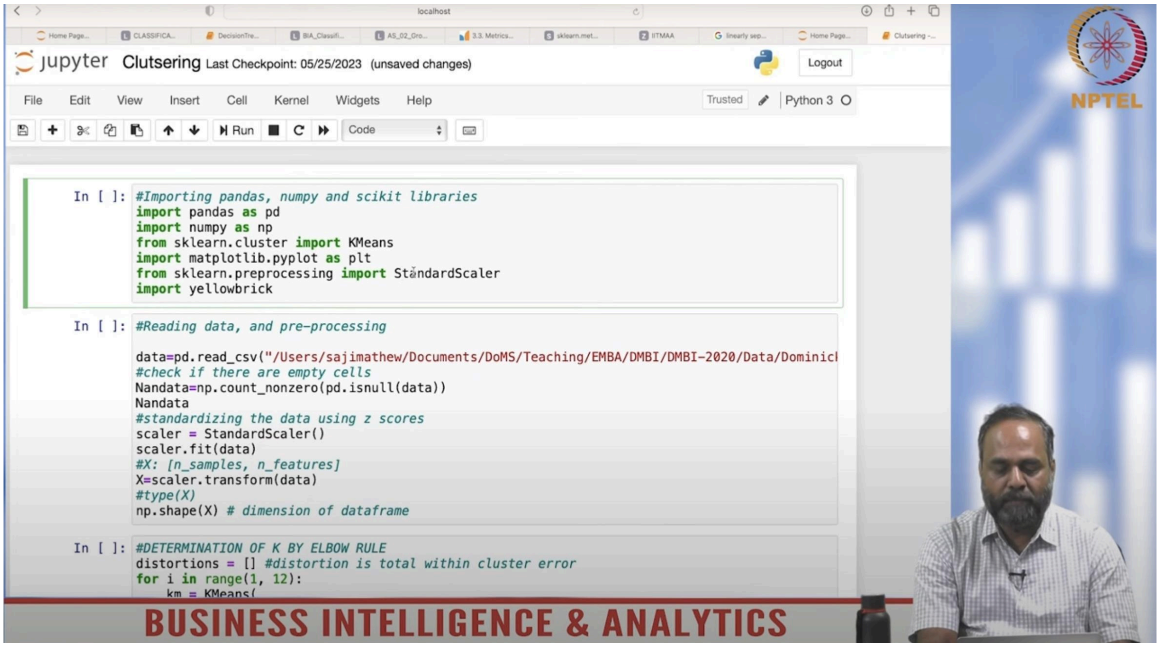
So there is no privacy in retail business, it's up to you whether you want to go there or not. Alright so these are characteristics or attributes they captured and in addition to total sales, which is actually figuring, you can actually open the data that is shared with you which is dominicks.csv, there is a file dominicks.csv and you can see or visualize that data in the screen as in different columns. There is grocery sales which is a total sales of grocery, customer count, these are two things which are not shown in that case file. Then customers under the age of 9, no car, household of size 2, we discussed these variables.

And what are these characteristics of customers, constraint shoppers and hurried shoppers.Have you heard about this? In marketing, these are important characteristics to understand customer behavior or consumer behavior. It is percentage, it's not about one customer, it's about the store level data. So percentage of constrained shoppers, who are constrained shoppers, shoppers who are constrained for money and they will have a particular behavior. So they expect that certain shops or shopping areas will be characterized by a demography where customers are constrained. What is the other attribute, percentage of hurried shoppers who are hurried shoppers? Yeah, who are constrained for time, they do not have probably money problem but they have time problem.



	A	B	C	D	E	F	G	H	I	J	K	L
1	GROCERY	CUSTCOUNT	AGE9	NOCAR	HSIZE2	SINGLE	WRKWNCH	TELEPHN	NWHITE	SHPCONS	SHPHURR	
2	131039.02	13965	0.12	0.12	0.31	0.3	0.29	0.99	0.14	0.09	0.12	
3	76785.74	10993	0.1	0.06	0.34	0.26	0.3	0.99	0.15	0.06	0.12	
4	184832.71	17691	0.14	0.03	0.34	0.26	0.27	0.99	0.09	0.03	0.19	
5	227801.15	24847	0.12	0.08	0.31	0.25	0.24	0.99	0.05	0.08	0.14	
6	120986.27	17631	0.1	0.04	0.36	0.23	0.29	0.99	0.13	0.04	0.16	
7	145602.37	27649	0.11	0.48	0.28	0.45	0.33	0.91	0.5	0.23	0.05	
8	127635.6	13895	0.13	0.03	0.34	0.22	0.26	0.99	0.12	0.02	0.22	
9	221839.91	23082	0.11	0.14	0.33	0.26	0.28	0.99	0.09	0.12	0.1	
10	181252.36	17694	0.18	0.02	0.27	0.26	0.19	0.99	0.16	0.03	0.23	
11	103857.45	13057	0.13	0.05	0.36	0.23	0.28	0.99	0.1	0.05	0.17	
12	199235.28	26008	0.1	0.07	0.34	0.27	0.3	0.99	0.05	0.06	0.11	
13	111133.88	21793	0.05	0.51	0.28	0.59	0.46	0.94	0.17	0.12	0.03	
14	176649.83	19387	0.13	0.05	0.3	0.27	0.23	0.98	0.06	0.08	0.13	
15	156793.96	15405	0.14	0.04	0.35	0.23	0.26	0.99	0.05	0.04	0.2	
16	116547.05	13263	0.15	0.02	0.34	0.25	0.32	0.99	0.13	0.03	0.16	
17	115302.8	12831	0.14	0.02	0.29	0.28	0.25	0.98	0.17	0.05	0.17	
18	110048.13	12341	0.12	0.02	0.33	0.33	0.31	0.99	0.17	0.03	0.13	
19	94890.66	11141	0.13	0.05	0.32	0.24	0.26	0.99	0.1	0.05	0.17	
20	96927.67	11579	0.12	0.04	0.32	0.28	0.24	0.99	0.09	0.06	0.13	
21	149674.19	15886	0.13	0.03	0.32	0.26	0.24	0.99	0.04	0.06	0.14	
22	153527.27	16190	0.14	0.01	0.34	0.25	0.28	0.99	0.14	0.02	0.21	
23	123150.5	13804	0.12	0.15	0.33	0.23	0.28	1	0.2	0.09	0.14	
24	123056.06	14275	0.15	0.02	0.3	0.3	0.31	0.99	0.1	0.03	0.19	

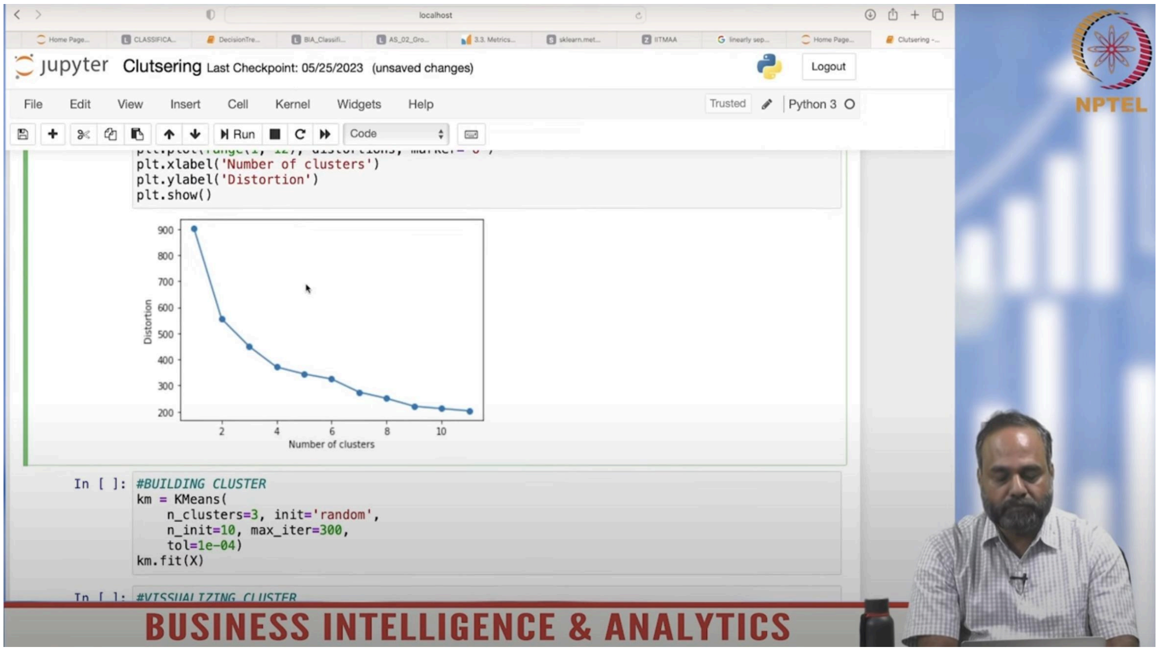
So they want products to be available as soon as they enter the store, a store, want to spend minimum time in the store and go whereas a constrained shopper may spend a lot of time in the store looking for discounts, looking for promotions and so on or coupons,encashing coupons and so on. So these are customer behavior based on demographic characteristics. So what are we going to do, as I said we are going to omit the first problem but we are intently going to focus on a clustering problem which we will derive or we will sort of convert this business problem into a clustering problem, where we must develop a clustering, a cluster model appropriate for the given objectives of Dominick's using python libraries, that is our problem converted into analytics problem. Choose the appropriate number of clusters or segments, here cluster means store segment, store segments for which the company wants to run certain promotions okay and suggest promotional tactics for each segment. So objective one and two, is desk, we will sit here, use data, use our programming skills, develop solutions and then we have to go and make a presentation to the business, business users. So therefore we have to profile the clusters and explain or work with the business group to sort of identify promotional tactics that fit the different store segments. That is the objective .


localhost
jupyter Clutsering Last Checkpoint: 05/25/2023 (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code
NPTEL
BUSINESS INTELLIGENCE & ANALYTICS

```
In [ ]: #Importing pandas, numpy and scikit libraries
        import pandas as pd
        import numpy as np
        from sklearn.cluster import KMeans
        import matplotlib.pyplot as plt
        from sklearn.preprocessing import StandardScaler
        import yellowbrick

In [ ]: #Reading data, and pre-processing

        data=pd.read_csv("/Users/sajimathew/Documents/DoMS/Teaching/EMBA/DMBI/DMBI-2020/Data/Dominicl
        #check if there are empty cells
        Nandata=np.count_nonzero(pd.isnull(data))
        Nandata
        #standardizing the data using z scores
        scaler = StandardScaler()
        scaler.fit(data)
        #X: [n_samples, n_features]
        X=scaler.transform(data)
        #type(X)
        np.shape(X) # dimension of dataframe

In [ ]: #DETERMINATION OF K BY ELBOW RULE
        distortions = [] #distortion is total within cluster error
        for i in range(1, 12):
            km = KMeans(
```

So we have looked at the problem, we have looked at the data, now we can actually open the python interface. I am going to use jupyter notebook. So here are a few cells wherein I have put together related codes, of course the first part of the script is to import the relevant libraries, both python libraries and scikit libraries, scikit-learn libraries. So I am importing pandas as pd, numpy as np and scikit-learn.cluster. I am importing K-means as a function, and matplotlib.pyplot, I am using the plot function plt, as plt , that's the short name I give here and for data pre-processing or standardization, I am using standard scalar, which actually does z-score normalization and for graphical display of clusters, I am using yellow brick, which is not a scikit-learn function, it's not in the scikit-learn library. You must install that separately, the yellow brick which has very rich features for graphical displays both for clustering as well as decision trees .

So I run that cell I got, my goodness, my yellow brick has a problem, maybe I updated. I am not going to use it. Maybe I need to update that. So don't worry about it, it's not mandatory for this exercise and in the second cell, I am going to import the data which we just visualized in a csv file in a spreadsheet and we are going to import that data, check if there are, there are null values or missing data, that is what the known data does and then look at the shape of the data.So this is 82 into 11 pandas data frame or a table with 82 records and 11 columns,11 columns and that is what it is, if you want to see whether it's whether it has, sorry, so just want to check whether it has missing data. There is no missing data, there is no missing data and so that is what is the output now. I am now scaling the data, so I am using the standard scalar. That is the next exercise, that is done and the rest of the things, the shape of the data we already had an appreciation, so we go to the next cell to prepare this data for K-means clustering, I am going to use K-means and then during the class we discussed a method or methods to determine number of clusters or the value of K . This is something that we can practically assume, one thing that we notice here is, it is store level data and therefore it is not large data, large the size of the data is just 82 and therefore we do not have to have too many clusters, like 10 store clusters may not make sense, but should be a reasonable number like 3, 4 etc. But at the same time, we should have some method. So we discussed the ELBOW method. We discussed the ELBOW method where we run the K-means for different number of clusters and then look at the total within cluster error and then determine what is a suitable, my machine is taking a little while because I think I am running too many things in the background and you can see the, the plot of distortion which is error, the y-axis is distortion or error, error versus number of clusters and you can see that the error is decreasing or almost becoming steady without much change after the number of clusters reaches a particular number.

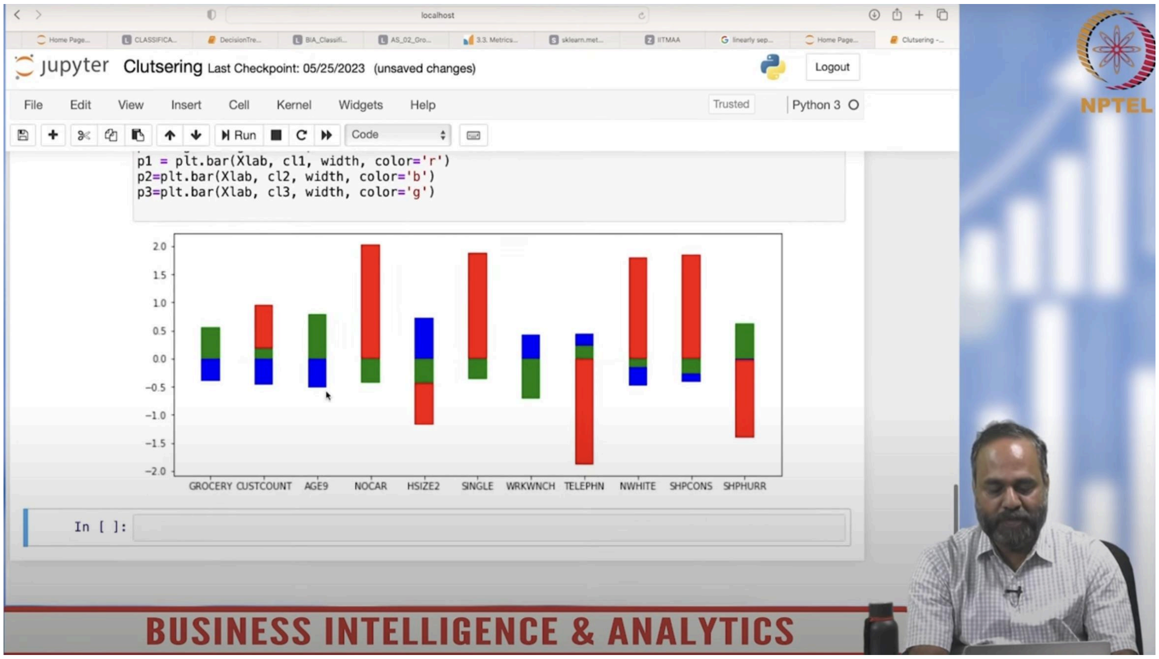


```
plt.plot(range(1, 12), distortions, marker='o')
plt.xlabel('Number of clusters')
plt.ylabel('Distortion')
plt.show()
```

```
In [ ]: #BUILDING CLUSTER
        km = KMeans(
            n_clusters=3, init='random',
            n_init=10, max_iter=300,
            tol=1e-04)
        km.fit(X)
```

```
In [ ]: #VISSUALIZING CLUSTER
```

So what is the ELBOW point of this cluster. It is maybe, you can take it as 4 or 3 or 5. So I would suggest that instead of going for a point, go for a range, so try cluster numbers 3, 4 and 5 and see which solution is most useful for application and, you know when you go for a presentation, for example you can, of course talk to domain experts or who are going to use this solution and see what solution is most appropriate as well. So we got a sense, so what I am going to do. I am going to create a solution of number of clusters is equal to 3, number of clusters is equal to 3. So K-means cluster is formed and the name of the object that is formed is km, km is the name of the cluster object that is created and I am not visualizing it using yellow brick here, but what is more interesting for me to see is, well in order to apply this solution, I need to understand what are the distinct characteristics of each cluster. What defines the uniqueness of each cluster. So as we know, a cluster is unique or similar within and they are different across. So we need to look through each cluster and understand how they are unique or discriminating from one to the other. So this is a very custom plot or a custom profiling that I have created, you could try other means also. I am very conventional and or follow very traditional approach of bar graphs to do this, but you can feature it differently.



```
p1 = plt.bar(Xlab, cl1, width, color='r')
p2=plt.bar(Xlab, cl2, width, color='b')
p3=plt.bar(Xlab, cl3, width, color='g')
```

localhost
jupyter Clutsering Last Checkpoint: 05/25/2023 (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
GROCERY CUSTCOUNT AGE9 NOCAR HSIZE2 SINGLE WRKWNCH TELEPHN NWHITE SHPCONS SHPHURR
In [ ]:
NPTEL
BUSINESS INTELLIGENCE & ANALYTICS

So what is done here is that on the x-axis, I have the different attributes and on the y-axis, the values of the attributes. Since the attributes have been standardized, they are in a scale of, say you know, it is normal, normalized and therefore it may be in a scale of, say most of the data will fall between a minus 2 to plus 2, a normalized, standard normal distribution. After the normalization, the data points fall in that range and these are basically z scores or z scores. So the range is that and therefore what you get to see in the bars is the value or the average value of each attribute in different clusters.

So the clusters have been given three different colors. They have been given red, red, blue and green, three colors. So you can see. These colors were assigned by me, red blue and green. So three colors. So there are three clusters red cluster, blue cluster and green cluster. Now we talked about cluster profiling. What is cluster profiling ? That is what you see here. What is the, what are the unique features of each cluster ? For example, look at the red cluster.

Let us go by cluster by cluster. We can also compare clusters because of this plot here but let us look at one cluster like the red cluster. We notice that the value of, the average value of hurried customers is very low, right. It’s a negative side, it's a minus 2 to plus 2.

That is the range, right. The value is very low or hurried customers are very low in this segment. But that makes us ask, what about constrained customers? Constrained customers are very high in the red cluster. Do you notice that? Red has actually discriminating features, between hurried and constrained customers. There is a distinct characteristic or profile that is emerging. That red cluster has very low concentration of hurried customers but very high concentration of constrained shoppers, constrained shoppers meaning time constraint. They have money but they have, sorry, opposite. They don't have money but they have time. Constrained means they, they are constrained for money. So constrained shoppers are very high in red, hurried customers or rich customers are very low. Keep that in mind.

Now look at the demographic features of the red cluster. Percentage of customers with not tele, with telephone very low, singles very high, customers with no car, no automobiles very high. So you are looking at other characteristics, other demographic characteristics of that cluster. So we are understanding these cluster of stores, the customers don't have car, don't have automobiles. Most people don't have automobiles. They have large concentration of singles but household of size 2 is very low but singles are very high. Those having telephone is also low. So, and they are constrained shoppers, right. They are constrained shoppers and there is also a racial profile there, that the non-whites are very high .

So as someone running the organization or the marketing manager marketing head of the organization, do you see any interesting profile in this cluster? And if you are thinking of promotional tactics, is there anything useful here? So important question for a manager is, how is it useful? How do I run my promotional tactics? How do I design promotions for clusters.

Now look at the green cluster. You see almost an opposite characteristic, right. This is almost opposite in nature and they look like an affluent class of stores or located in a affluent area. These are stores in different parts of the state. So these stores are located where affluent class of people are living, by glancing the characteristics. So keep in mind, what is profiled is the average value or the mean value or some central tendency. Some, depending on the variable type, some central tendency of that variable in each cluster. So then you can compare the clusters in terms of the mean values of those attributes . That is what we are doing.

For example, look at something like percentage of household under size 2. That particular attribute, you see that the red cluster has the lowest concentration or the lowest average. The green has the next. It is in the middle but the blue cluster has a lot, their average value is very high. So you can compare the three clusters in terms of the average value of the attributes and then see what is the characteristic of each cluster as a whole. In profiling what you do is, you look at the cluster as a whole.

You remember I described Claritas PRIZMS cluster segmentation, how they actually describe the clusters. You know movers and shakers. They are the affluent class. They actually travel a lot and they are educated. How do you do that? From this. Take a cluster and describe them in terms of what uniquely characterizes them in terms of high value or low value, comparatively, comparatively and that enables you to plan action.

This is an exercise for action, for taking action. When you look at a cluster like the red cluster or the green cluster, you immediately think, they don't have cars. Deliver at home or something like that. So, so not that it is very practical, whether it is practical. So that is useful for discussing actions, not that you act by simply looking at this diagram.

But it is useful for designing programs, but now I will close this session here because my job is to help you understand how to do clustering, how to profile clusters and how to interpret them. Now the action, that is where actually you take it to a practicing manager and see if they find it useful. They do not find it useful, create a lot more solutions. For example, create clusters of size four , four clusters and see if the solution changes. Apply different clustering techniques, hierarchical, a Wards method followed by a K-means and see if the usefulness of the solution actually improves and also one thing to keep in mind is that the cluster sizes should be reasonable. See if you have clusters of reasonable sizes etc you have to output that and see that okay.

So yellow brick could be useful in actually appreciating the sizes of clusters and how they are actually close or distinct and that is your exercise, to run it, update the application and run it . Any questions here? So this is very simple to do and very useful to apply. So I close this exercise here, yeah.

There will be slight variations between solutions, it is K-means okay you start. So between two people, the solutions will be slightly different because you start with random seeds. The starting point is different, so the final solutions can be different. So no, it will not exactly match but as I said, if you fix the initial seeds using a initial solution like the Wards method, determine the centroids of the solution given by Wards method, use those centroids as starting seeds for K-means. Then you have, everyone uses the same seed, then your solutions will match. And that may be one approach suggested in literature, all right.