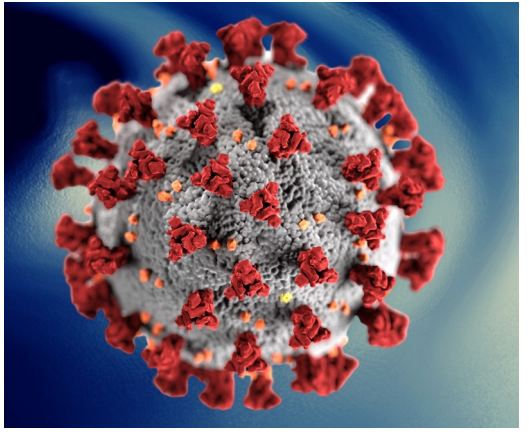
Choreograph interview

Covid dataset example

Presented by

Sarkis Kassounian



Step 1: Acquire the data

- Data aquiring can be either the most dificult of easiest part of process
  - Sale contracts (data providers)
  - Legal issues
  - GDPR
  - Etc …
  - Existence of data
- I'm lucky … You just gave it to me :)
- I chose the ***test_data_covid_attitudes_and_purchase_recency.sav***

Step 2: Exploring the data

- As the data is in .sav format (SPSS friendly), this can give us the option of including the metadata
- The files contains answers of a survey regarding the behavior of people during the covid pandemic period
- It also contains demographic information regarding the participants
- I converted the metadata into a dataframe to easily visualize the questions and available answers and information (summary of survey components)
- The same columns will later be considers as the features
- The data is mostly categorical or discreate integers

df_answers - DataFrame

Index	questions	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10
Market	Market	nan	Argentina	Australia	Austria	Belgium	Brazil	Canada	China	Colombia	Czech Republic	Denmar
Q01_NET	Age ranges NET (6 categories)	nan	Under 18	18-24	25-34	35-44	45-54	55-64	65 Plus	nan	nan	nan
Q02	Gender	nan	Male	Female	nan	nan	nan	nan	nan	nan	nan	nan
DEM10_ALL	GL | Income | Brackets scaled to the average comparable across markets NET	nan	25%	50%	75%	Average 100%	125%	175%	250%	350%	Do not wish to answer	Don't
DEM05_01	GL | Working Status | Full time worker	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
DEM05_02	GL | Working Status | Part time worker	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
DEM05_03	GL | Working Status | Freelancer	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
DEM05_04	GL | Working Status | Self-employed	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
DEM05_05	GL | Working Status | Retired	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
DEM05_06	GL | Working Status | Full time homemaker	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
DEM05_07	GL | Working Status | In full time education	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
DEM05_08	GL | Working Status | In part-time education	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
DEM05_09	GL | Working Status | Unemployed before COVID - 19	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
DEM05_10	GL | Working Status | Other	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
DEM05_11	GL | Working Status | Full time worker - temporary job retention/ wage support scheme	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
DEM05_12	GL | Working Status | Part time worker - temporary job retention/ wage support scheme	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
DEM05_13	GL | Working Status | Freelancer - temporary job retention/ wage support scheme	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
DEM05_14	GL | Working Status | Self-employed - temporary job retention/ wage support scheme	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
DEM05_15	GL | Working Status | Unemployed - since COVID-19	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
DEM05_16	GL | Working Status | In unpaid employment or full time care of another household member	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
Lifestage_01	Lifestage | Young single NET	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
Lifestage_02	Lifestage | child free couples NET	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
Lifestage_03	Lifestage | Parent with young children NET	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
Lifestage_04	Lifestage | Parent with older children NET	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
Lifestage_05	Lifestage | Any parent NET	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
Lifestage_06	Lifestage | Empty nest NET	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
Lifestage_07	Lifestage | Mature singles NET	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
Lifestage_08	Lifestage | Senior singles NET	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
Lifestage_09	Lifestage | Hotel Parents NET	No	Yes	nan	nan	nan	nan	nan	nan	nan	nan	nan
QP07_01	GL | Attitudes | Going to restaurants, pubs or bars does not hold the same appeal/enjoyment as before	nan	Completely disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Completely agree	nan	nan	nan	nan	nan
QP07_02	GL | Attitudes | Supporting my local community is more important than before the crisis	nan	Completely disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Completely agree	nan	nan	nan	nan	nan
QP07_03	GL | Attitudes | I am using more local shops and services	nan	Completely disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Completely agree	nan	nan	nan	nan	nan

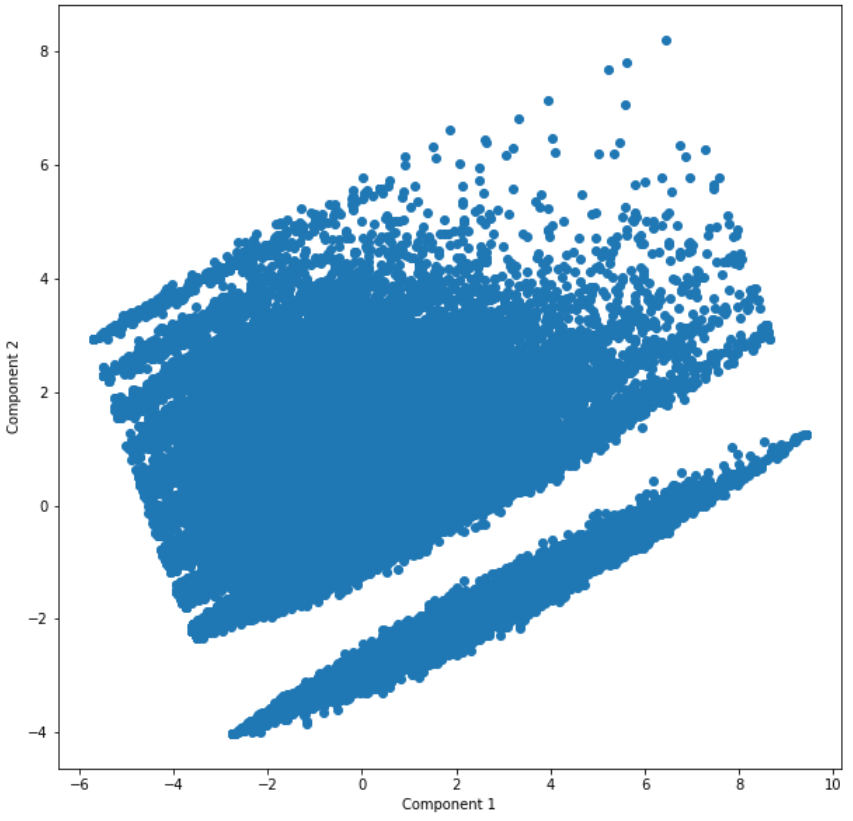
Format | Resize | ☑ Background color | ☑ Column min/max | Save and Close | Close

Step 3: Pre- the data

- As the exercise in question 2 required to apply segmentation only the financial measures and attititudes, I selected the columns which are associated to them
- Small summary to understand the nature of the data
- Later convert to a numpy matrix to implement quicker computations
- Apply standardization of the data
  - Subtract by mean and devide by std

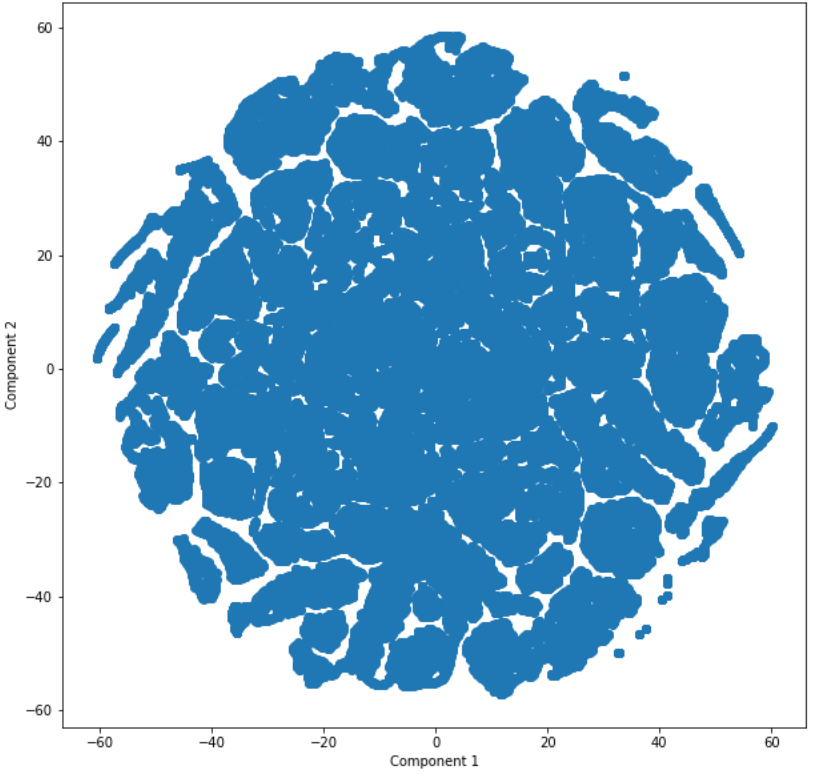
Step 4: Apply segmentation

- We start with applying PCA since it is faster and can reduce the dimension into 2
- We can then visualize the 2 largest components and see if there are any clusters
- An initial visualization gives us two sets of segments as we can clearly see
  - X-axis → PC1
  - Y-axis → PC2



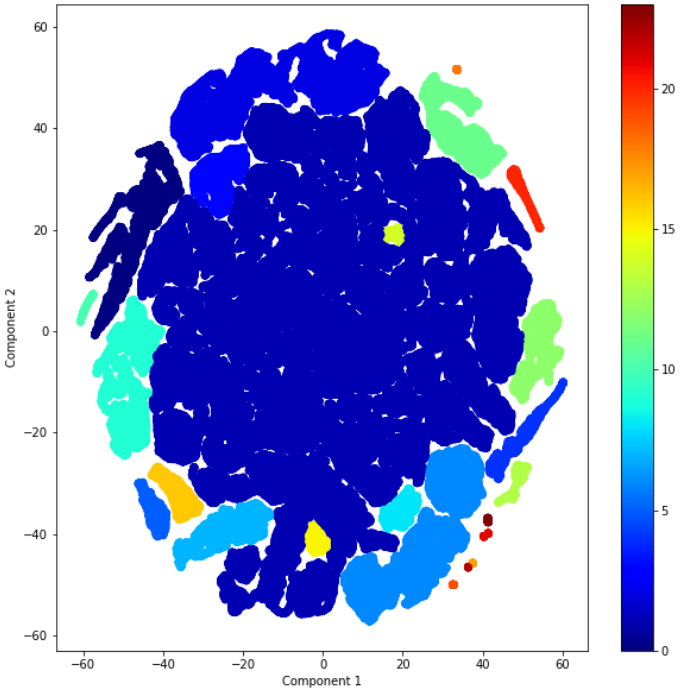
Step 4: Apply segmentation

- We can also apply t-SNE to be able to have a better layer of segmentation since it mantains the neighbourhood distances and assignes a t-distribution weight system
  - X-axis → PC1
  - Y-axis → PC2



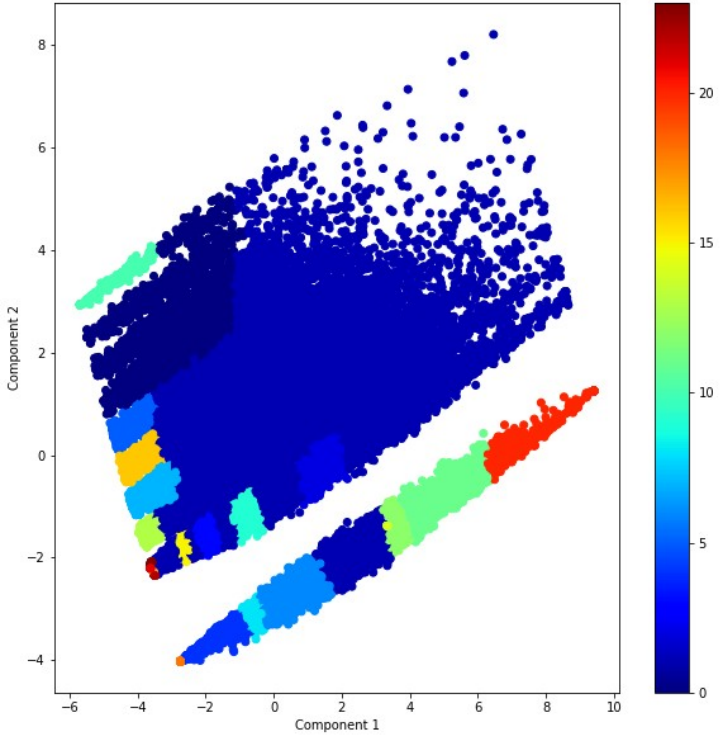
Step 4: Apply segmentation

- To automate the grouping we can apply a clustering algorithm such as DBSCAN and it can create the segments for us
  - X-axis → PC1
  - Y-axis → PC2
  - Colorbar → classes



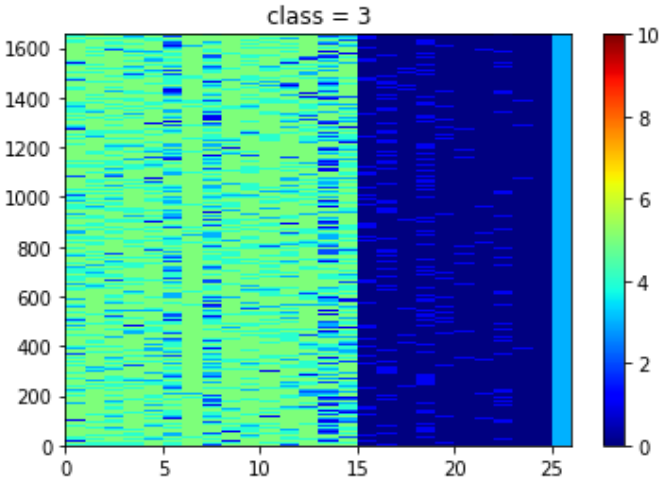
Step 4: Apply segmentation

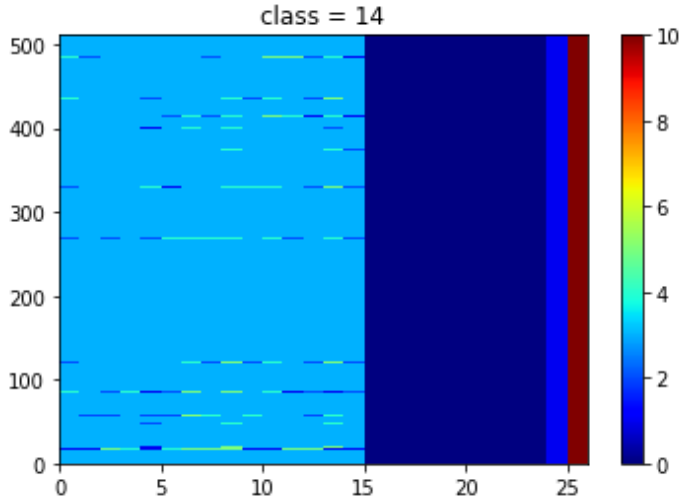
- To automate the grouping we can apply a clustering algorithm such as DBSCAN and it can create the segments for us
  - X-axis → PC1
  - Y-axis → PC2
  - Colorbar → classes



Step 5: Visualize segmentation

- We can have a look at some the classes and see the differences of features and colormesh plots





Step 6: concluding thoughts

- We can see how class 3 has relatively higher QPs than class 14. It also has a complete set of 10s for the last CAT. By visualizing different classes we can understand their behaviour and find a reason to it