Customer Segmentation

by Ivana Hybenova

Background info

(To get you on the same page with the audience the presentation is built for.)

Projects is done for a mall that has some clients subscribed to the membership card.

When clients subscribe, they provide information like gender, age and annual income.

Because they have this card, they use it to buy all sorts of things in the mall and therefore the mall has the purchase history of each of its client member.

Based on information like annual income, the number of times per week they shop in the mall and the amount of money they spend per year the spending score is computed.

The metric takes values between 1 and 100. The closer is the score to 1, the less the client spends and the closer is the score to 100, the more client spends.

So far there was no analysis done based on these information.

Current situation

- Decisions based on gut-feeling

- General marketing

- General approach of customer experience department

- No way of monitoring „customer success"

Approach

- Input:
- 200 customers
- 4 features: Gender, Age, Annual Income, Spending Score
- No missing/incorect values thanks to quality assurance procedures applied when calculating spending score

- Output:
- Groups of customers with similar values

Clustering – technical part

- Clustering method:
 **Kmeans clustering (K++)**
  based on
  Euclidean Distance

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$
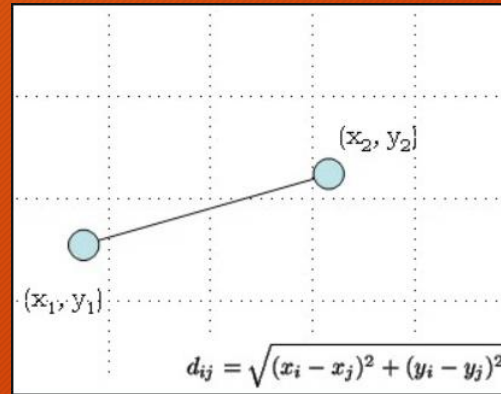
- Asuring equal importance of the features:
  Standard Scaler

age	annual_income_thousands	spending_score	Female	Male
19	15	39	0	1
21	15	81	0	1
20	16	6	1	0
23	16	77	1	0
31	17	40	1	0

| Female	Male	age_scaled	annual_income_thousands_scaled	spending_score_scaled
-1.128152	1.128152	-1.424569	-1.738999	-0.434801
-1.128152	1.128152	-1.281035	-1.738999	1.195704
0.886405	-0.886405	-1.352802	-1.700830	-1.715913
0.886405	-0.886405	-1.137502	-1.700830	1.040418
0.886405	-0.886405	-0.563369	-1.662660	-0.395980

Clusering – technical part

- Selecting optimal number of segments: **Elbow Method**
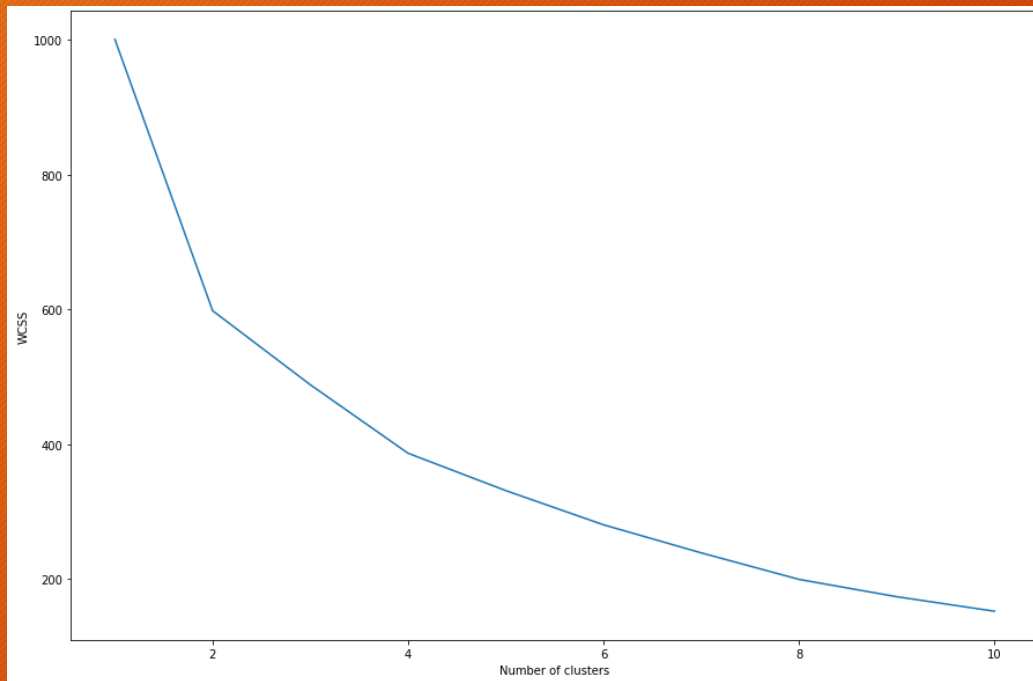
Clustering – technical part

- Visual check: **PCA – 2 components**
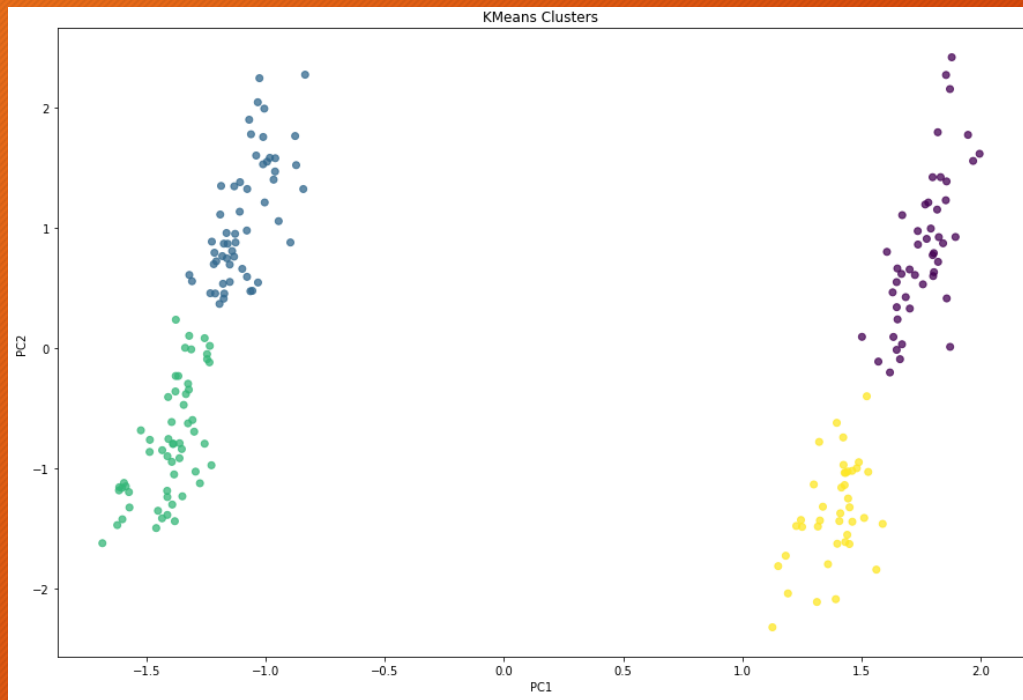
4 segments described

- Found segments

Conclusion

1. 2 clusters has higher spending score than the other 2

2. It is not the gender

3. It is not the annual income

4. It is the age!

Suggestions for future investigation

- Do we have products of their interest?

- If yes, are the prices comparable with our local competitors?

- Are the products reasonably placed in the mall?

- Is our marketing appropriate also for older customers?

Deployment

- Segmenting new customers on daily basis – fully automated (Airflow)
  - Code to segment new customers
  - The DAG file

- Conducting segmentation on regular basis – manually
  - -> again with 30 – 40 new customers (keeping ~15 % ratio)
  - -> several months after applying any changes to spot the improvement, if any
  - Code to perform cluster analysis

Thank you!