Machine Learning (online) Übung 3

# Aufgabe: K-Means Clustering on House Data

## Feature Engineering

K-Means Clustering verwendet normalerweise die euklidische Distanz zur Berechnung der Cluster Centroiden und Cluster Zugehörigkeiten. Es ist daher nicht sinnvoll, kategorische Features wie z.B. *waterfront* oder *zipcode* zu benutzen, da die euklidische Distanz zwischen zwei Zipcodes nichts sinnvolles aussagt.

Die euklidische Distanz zwischen kontinuierlichen sowie diskreten Features wie *grade* oder *condition* ist aber sehr wohl aussagekräftig. Auch wenn im Dataset nur ganze Zahlen zwischen 1 und 5 für *condition* vorkommen, ist das trotzdem ein sinnvolles Feature.

Ich verwende daher folgende Features (alle normalisiert) für das Clustering:

*price, bedrooms, bathrooms, sqft\_living, sqft\_lot, floors, condition, grade, sqft\_above, sqft\_basement, yr\_built, zipcode, lat, long, sqft living15, sqft\_lot15*

*yr\_renovated* lasse ich aus, da es sehr viele 0 Werte enthält.

Um das «Problem» mit den kategorischen Features zu umgehen, könnte eine andere Distanzmetrik verwendet werden, wie z.B. die Hamming oder Jaccard Distanz. Dies ist aber hier wohl out of scope und ohnehin nicht nötig, da es ja genügend Features zur Analyse gibt.

## Bestimmung der Anzahl Cluster K

### Elbow Plot

Eine mögliche Variante zur Wahl der Anzahl Cluster ist der Elbow plot. Für K-Means mit folgender Parametrisierung kommt der darauffolgende Plot heraus:

max\_iters = 20;

cost\_history = [];

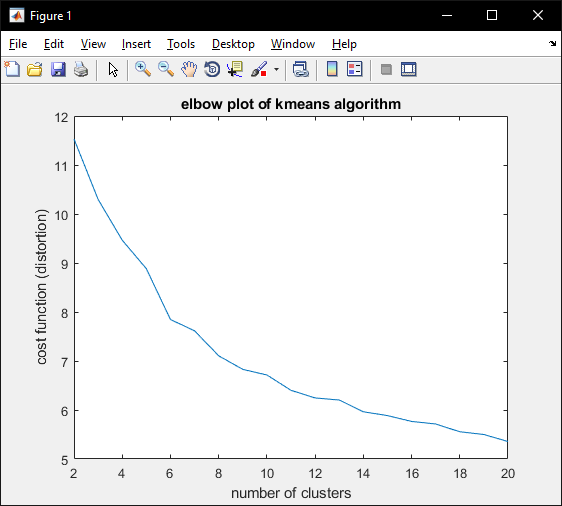
N = 20;

for K = 2:20

[centroids, y, cost] = runkMeansNtimes(X, K, N, max\_iters, false);

cost\_history = [cost\_history, cost];

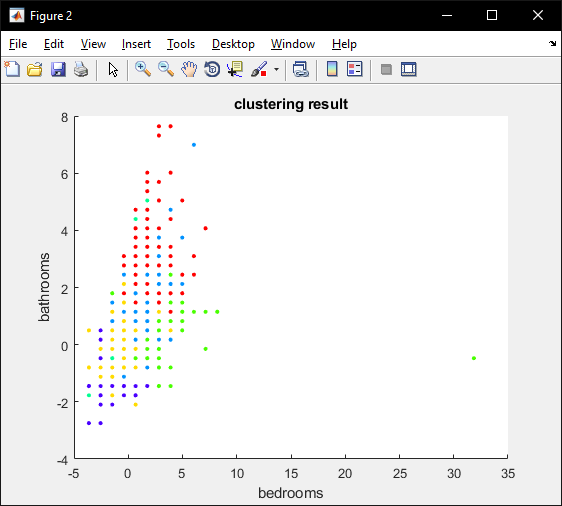
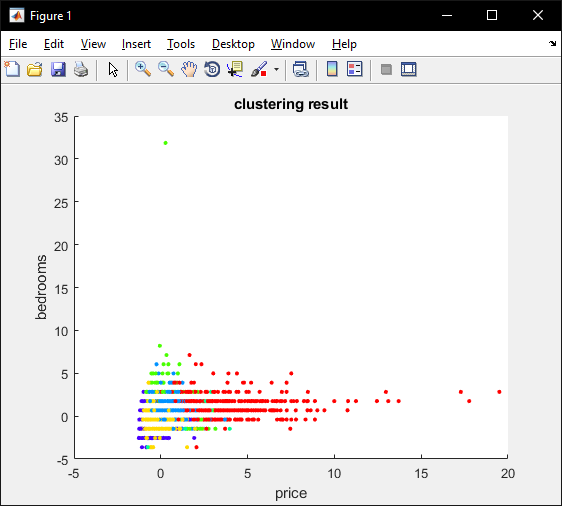
end

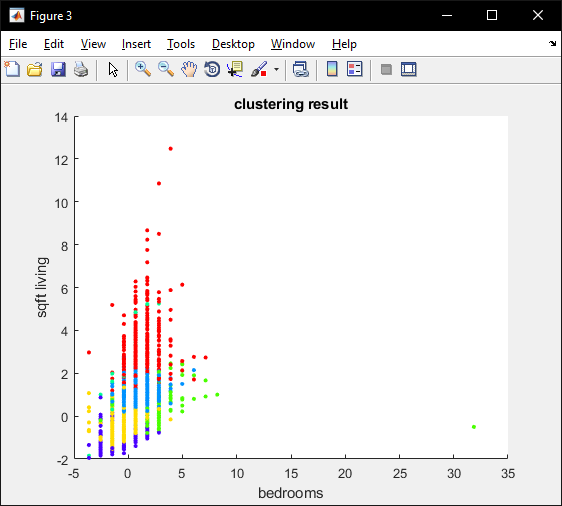


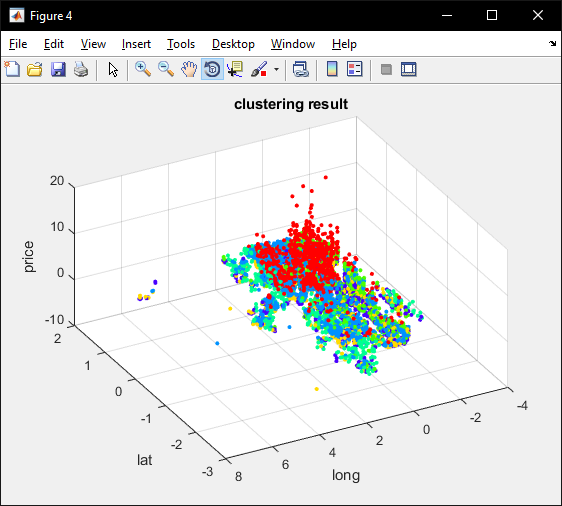
Bei diesem Plot ist nur ein schwacher «Elbow» zu erkennen. Am ehesten kommt wohl der Wert 6 in Frage, da die Kurve dort den grössten Knick macht.

### TODO Silhouette berechnen

## Cluster Visualization







## Charakterisieren der Cluster

Anhand der Varianz der Features pro Cluster ist folgendes ersichtlich:

cluster 1 variances:

lat : 0.31

sqft lot15 : 0.51

floors : 0.65

long : 0.70

sqft lot : 0.71

grade : 0.71

bedrooms : 0.94

yr built : 0.94

condition : 0.94

bathrooms : 1.06

sqft living : 1.27

sqft living15 : 1.38

sqft above : 1.62

sqft basement : 2.77

price : 4.08

cluster 2 variances:

sqft lot : 0.06

sqft lot15 : 0.09

sqft basement : 0.14

price : 0.15

sqft living : 0.19

yr built : 0.21

condition : 0.24

sqft above : 0.27

grade : 0.29

bathrooms : 0.29

sqft living15 : 0.32

bedrooms : 0.51

floors : 0.77

lat : 1.15

long : 1.16

cluster 3 variances:

sqft lot : 0.11

sqft lot15 : 0.15

sqft above : 0.23

sqft living : 0.28

grade : 0.37

floors : 0.39

bathrooms : 0.44

price : 0.44

sqft living15 : 0.46

long : 0.56

yr built : 0.63

lat : 0.78

sqft basement : 0.81

bedrooms : 1.05

condition : 1.30

cluster 4 variances:

yr built : 0.56

price : 0.63

floors : 0.81

condition : 0.86

bedrooms : 0.88

sqft living15 : 0.90

bathrooms : 1.12

sqft basement : 1.22

grade : 1.37

sqft living : 1.44

sqft above : 1.45

lat : 1.46

long : 1.46

sqft lot15 : 12.40

sqft lot : 19.51

cluster 5 variances:

floors : 0.25

sqft living : 0.28

sqft lot : 0.29

yr built : 0.29

sqft lot15 : 0.30

sqft above : 0.34

condition : 0.34

bathrooms : 0.35

sqft basement : 0.38

price : 0.38

grade : 0.54

bedrooms : 0.57

sqft living15 : 0.57

long : 0.84

lat : 0.89

cluster 6 variances:

sqft lot : 0.10

sqft lot15 : 0.11

sqft living : 0.14

sqft above : 0.15

price : 0.17

floors : 0.20

sqft living15 : 0.26

sqft basement : 0.27

bathrooms : 0.31

grade : 0.31

bedrooms : 0.55

yr built : 0.58

long : 0.77

lat : 1.07

condition : 1.13