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 = [];

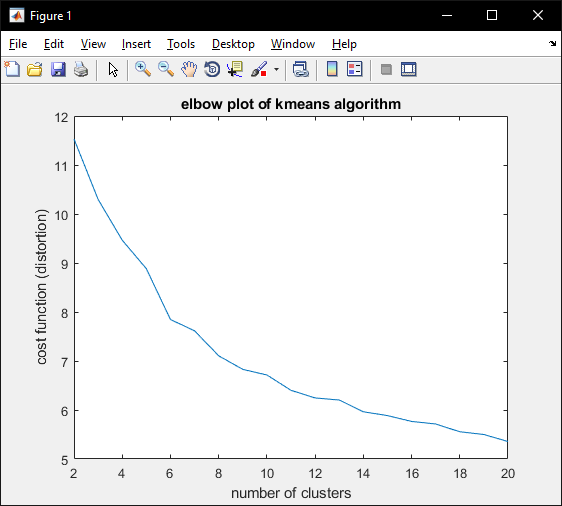
N = 20;

for K = 2:20

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

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

end

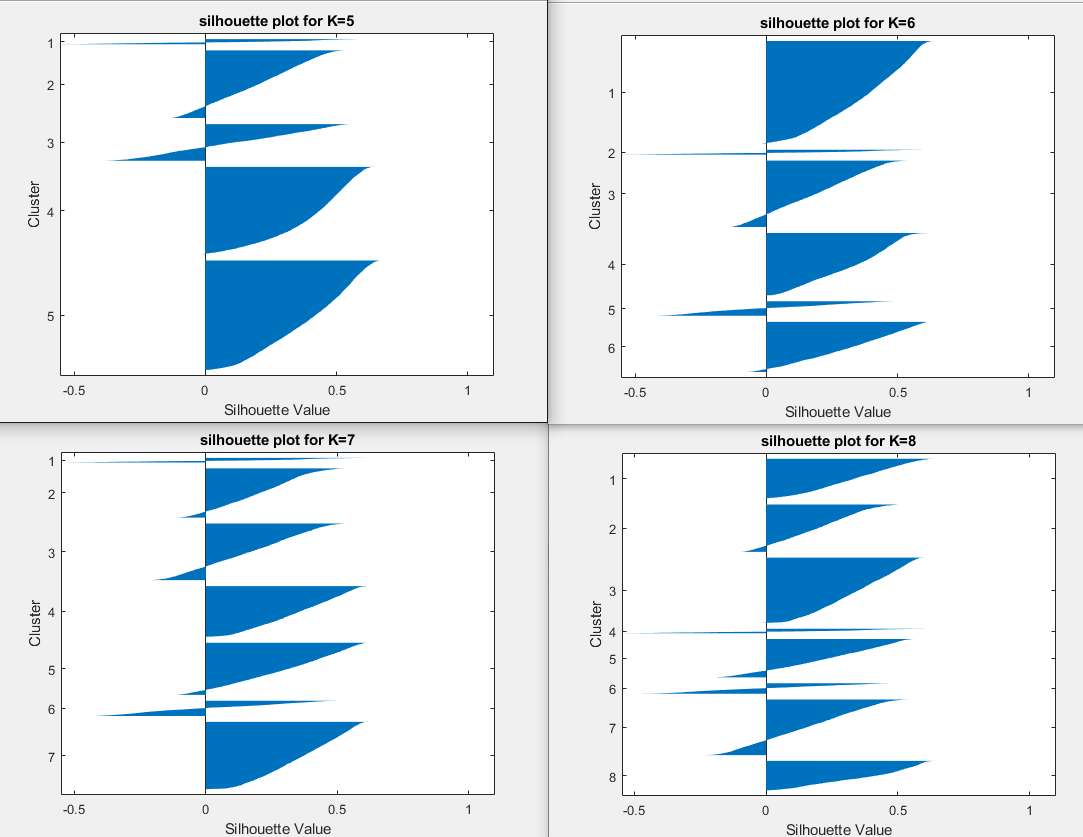


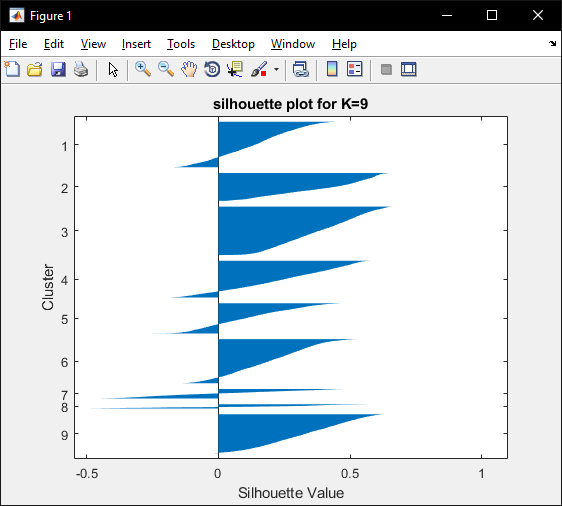
Bei diesem Plot ist nur ein schwacher «Elbow» zu erkennen. Am ehesten kommt wohl der Wert 6 in Frage, da die Kurve dort den grössten Knick macht.

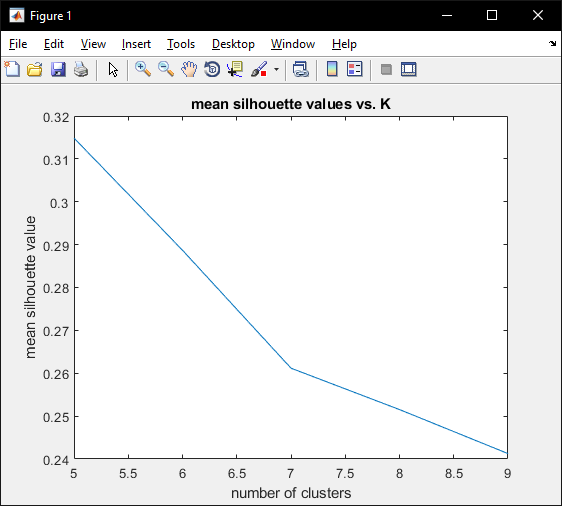
### Silhouette Plot

K=6 scheint gemäss Elbow Plot keine schlechte Wahl zu sein. Ich berechne hier die Silhouettenwerte für Werte von K im Bereich von 5 bis 9, zum Testen, ob ev. 6 auch wirklich die beste Wahl ist, oder ob leicht höhere oder tiefere Werte womöglich besser sind.

Der Silhouette Plot zeigt, wie ähnlich jeder Wert ist zu Werten aus dem gleichen Clustern im Vergleich zu Werten aus fremden Clustern. Werte gehen von -1 bis 1, wobei hohe Werte angeben, dass der Datenpunkt gut in den Cluster passt und schlecht mit Datenpunkten aus anderen Clustern übereinstimmt.





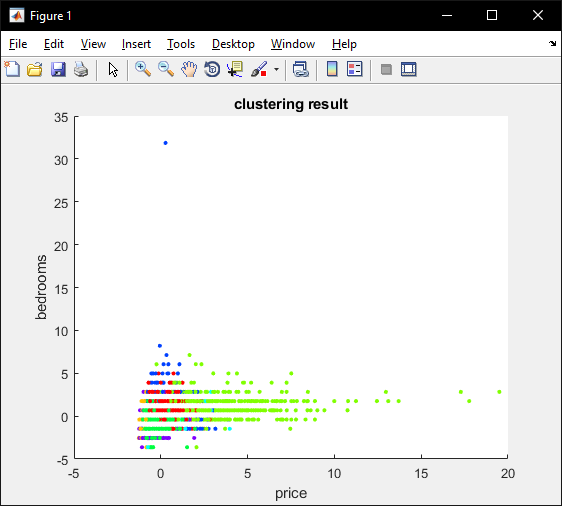
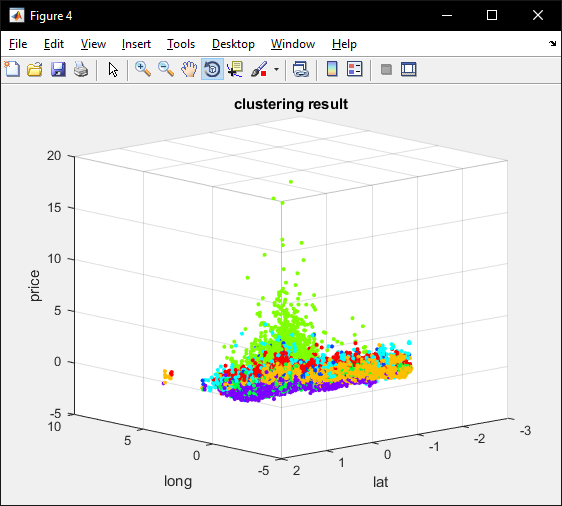


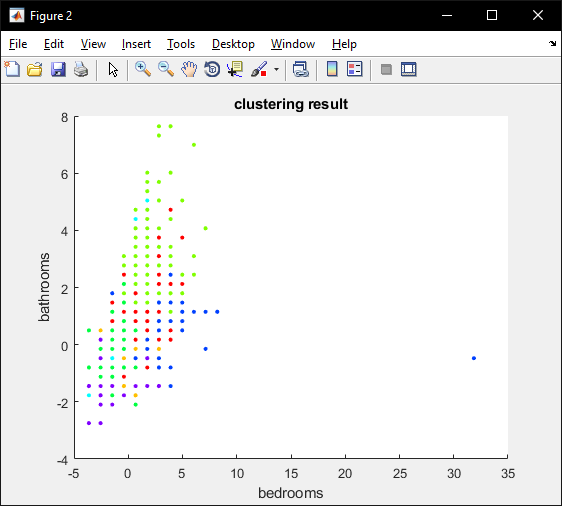
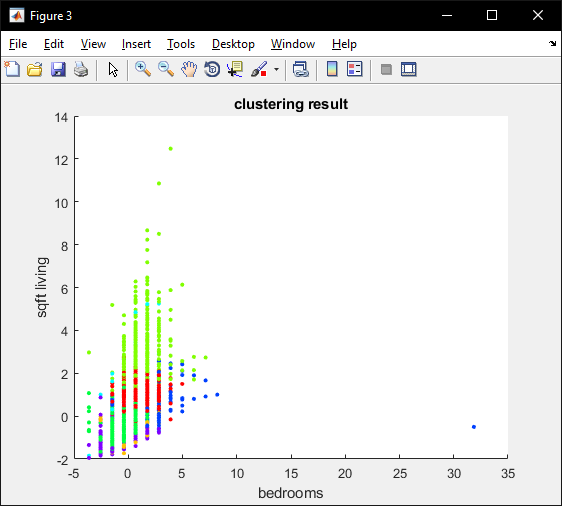
Bei der Betrachtung der Durchschnittlichen Silhouettenwerte gegenübergestellt mit K sieht ist ein leichter Knick bei K=7 zu sehen. Auch bei der Betrachtung der Silhouetten plots sieht man bereits, das aus K=7 eine leicht bessere Verteilung der Punkte resultiert als aus K=6.

Ich entscheide mich darum für **K=7** als geeignete Anzahl Cluster für die weiteren Aufgaben.

## Cluster Visualisierung und Charakterisierung

Dinensionsreduktion ist erst Teil der nächsten Aufgabe, ich verzichte daher darauf, einen Algorithmus wie PCA oder t-SNE zur Visualisierung der entstandenen Cluster zu verwenden. Stattdessen visualisiere ich die Cluster in 2- und 3-Dimensionalen Räumen unter Verwendung von ein paar intuitiv interessanten Beispiel Features-Kombinationen:





In den oberen beiden Plots sieht man bereits, dass z.B. teure Häuser tendenziell im selben Cluster sind. In den unteren beiden Plots ist zu z.B. erkennen, dass grössere Häuser mehr Schlafzimmer haben und auch entsprechend gruppiert werden oder dass die Anzahl Schlaf- und Badezimmer eine gewisse Korrelation aufweiset (was natürlich zu erwarten war).

### Genauere Analyse der Cluster

Nachfolgend berechne ich den Mittelwert und die Varianz jedes Feautres in jedem Cluster. Features mit geringer Varianz charakterisieren in gewisser Weise einen Cluster. Anhand der Mittelwerte kann dann geschlossen werden, welche Werte eines Features einen bestimmten Cluster ausmachen.

cluster 1 contains 17.8% of the datapoints:

floors : var = 0.23, mean = 0.85

sqft lot : var = 0.26, mean = 0.01

yr built : var = 0.27, mean = 0.83

sqft living : var = 0.28, mean = 0.84

sqft lot15 : var = 0.31, mean = 0.03

bathrooms : var = 0.31, mean = 0.71

sqft basement : var = 0.32, mean = -0.47

sqft above : var = 0.33, mean = 1.18

condition : var = 0.34, mean = -0.42

price : var = 0.36, mean = 0.30

sqft living15 : var = 0.54, mean = 0.96

grade : var = 0.54, mean = 0.88

bedrooms : var = 0.56, mean = 0.54

long : var = 0.84, mean = 0.76

lat : var = 0.93, mean = -0.02

cluster 2 contains 16.5% of the datapoints:

price : var = 0.08, mean = -0.67

sqft living : var = 0.16, mean = -0.59

sqft above : var = 0.18, mean = -0.49

floors : var = 0.20, mean = -0.76

sqft lot : var = 0.21, mean = -0.02

sqft lot15 : var = 0.26, mean = -0.01

sqft living15 : var = 0.29, mean = -0.52

grade : var = 0.31, mean = -0.60

yr built : var = 0.35, mean = -0.09

sqft basement : var = 0.38, mean = -0.29

bathrooms : var = 0.40, mean = -0.56

bedrooms : var = 0.44, mean = -0.23

lat : var = 0.61, mean = -1.09

condition : var = 1.11, mean = 0.41

long : var = 1.12, mean = 0.31

cluster 3 contains 5.2% of the datapoints:

lat : var = 0.33, mean = 0.33

sqft lot15 : var = 0.51, mean = 0.20

floors : var = 0.60, mean = 0.75

long : var = 0.69, mean = 0.17

grade : var = 0.69, mean = 2.20

sqft lot : var = 0.71, mean = 0.20

condition : var = 0.89, mean = -0.13

yr built : var = 0.91, mean = 0.46

bedrooms : var = 0.94, mean = 1.02

bathrooms : var = 1.05, mean = 1.83

sqft living : var = 1.24, mean = 2.45

sqft living15 : var = 1.39, mean = 2.01

sqft above : var = 1.52, mean = 2.14

sqft basement : var = 2.71, mean = 1.09

price : var = 4.03, mean = 2.78

cluster 4 contains 17.0% of the datapoints:

sqft lot : var = 0.03, mean = -0.24

sqft lot15 : var = 0.04, mean = -0.27

sqft basement : var = 0.14, mean = -0.49

price : var = 0.16, mean = -0.32

condition : var = 0.17, mean = -0.53

sqft living : var = 0.19, mean = -0.36

yr built : var = 0.19, mean = 0.95

sqft above : var = 0.26, mean = -0.13

sqft living15 : var = 0.29, mean = -0.39

grade : var = 0.29, mean = 0.03

bathrooms : var = 0.31, mean = 0.35

bedrooms : var = 0.51, mean = -0.34

floors : var = 0.58, mean = 1.16

long : var = 0.99, mean = -0.06

lat : var = 1.04, mean = -0.06

cluster 5 contains 1.5% of the datapoints:

yr built : var = 0.57, mean = 0.40

price : var = 0.64, mean = 0.21

floors : var = 0.81, mean = 0.11

condition : var = 0.87, mean = -0.14

bedrooms : var = 0.91, mean = -0.05

sqft living15 : var = 0.92, mean = 0.53

bathrooms : var = 1.14, mean = 0.35

sqft basement : var = 1.25, mean = -0.09

grade : var = 1.38, mean = 0.36

lat : var = 1.46, mean = -0.61

sqft living : var = 1.46, mean = 0.65

sqft above : var = 1.48, mean = 0.77

long : var = 1.50, mean = 1.33

sqft lot15 : var = 12.61, mean = 6.29

sqft lot : var = 19.71, mean = 5.81

cluster 6 contains 19.0% of the datapoints:

sqft lot : var = 0.10, mean = -0.09

sqft lot15 : var = 0.12, mean = -0.10

sqft above : var = 0.23, mean = -0.28

sqft living : var = 0.30, mean = 0.35

grade : var = 0.40, mean = 0.00

floors : var = 0.41, mean = -0.58

bathrooms : var = 0.45, mean = 0.25

price : var = 0.45, mean = 0.23

sqft living15 : var = 0.50, mean = 0.13

long : var = 0.52, mean = -0.32

lat : var = 0.61, mean = 0.38

yr built : var = 0.63, mean = -0.51

sqft basement : var = 0.78, mean = 1.24

bedrooms : var = 1.11, mean = 0.60

condition : var = 1.31, mean = 0.58

cluster 7 contains 23.0% of the datapoints:

sqft lot : var = 0.04, mean = -0.18

sqft lot15 : var = 0.06, mean = -0.20

sqft above : var = 0.16, mean = -0.76

sqft living : var = 0.16, mean = -0.85

price : var = 0.19, mean = -0.35

floors : var = 0.25, mean = -0.66

sqft living15 : var = 0.27, mean = -0.69

bathrooms : var = 0.29, mean = -1.04

sqft basement : var = 0.32, mean = -0.34

grade : var = 0.34, mean = -0.80

long : var = 0.34, mean = -0.64

lat : var = 0.47, mean = 0.50

yr built : var = 0.50, mean = -1.00

bedrooms : var = 0.62, mean = -0.72

condition : var = 1.07, mean = -0.02

Aus den obigen Daten können die Cluster wie folgt charakterisiert werden:

* **Cluster 1**: Grosse, mehrstöckige und tendenziell neuere Häuser
* **Cluster 2**: Billige, sehr kleine Häuser, die aber noch in guten Zustand sind, womöglich in unbeliebten Gegenden (gegen Nordosten)
* **Cluster 3**: Kleinere wohl beliebte und teure Siedlung im südlichen Stadtteil
* **Cluster 4**: Kleine und billige Häuser, die tendenziell neu aber wahrscheinlich qualitativ eher schlecht sind (condition)
* **Cluster 5**: Wenige und eher neuere, in der ganzen Stadt verstreute grössere und eher preiswerte Häuser, Villas.
* **Cluster 6:** Einstöckige, durchschnittliche Häuser, die aber schon etwas älter sind
* **Cluster 7:** Sehr kleine, alte, nicht allzu teure und wohl ein wenig heruntergekommene Häuser