# Machine Learning (online) Übung 3

## 1 Aufgabe: K-Means Clustering on House Data

### 1.1 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.

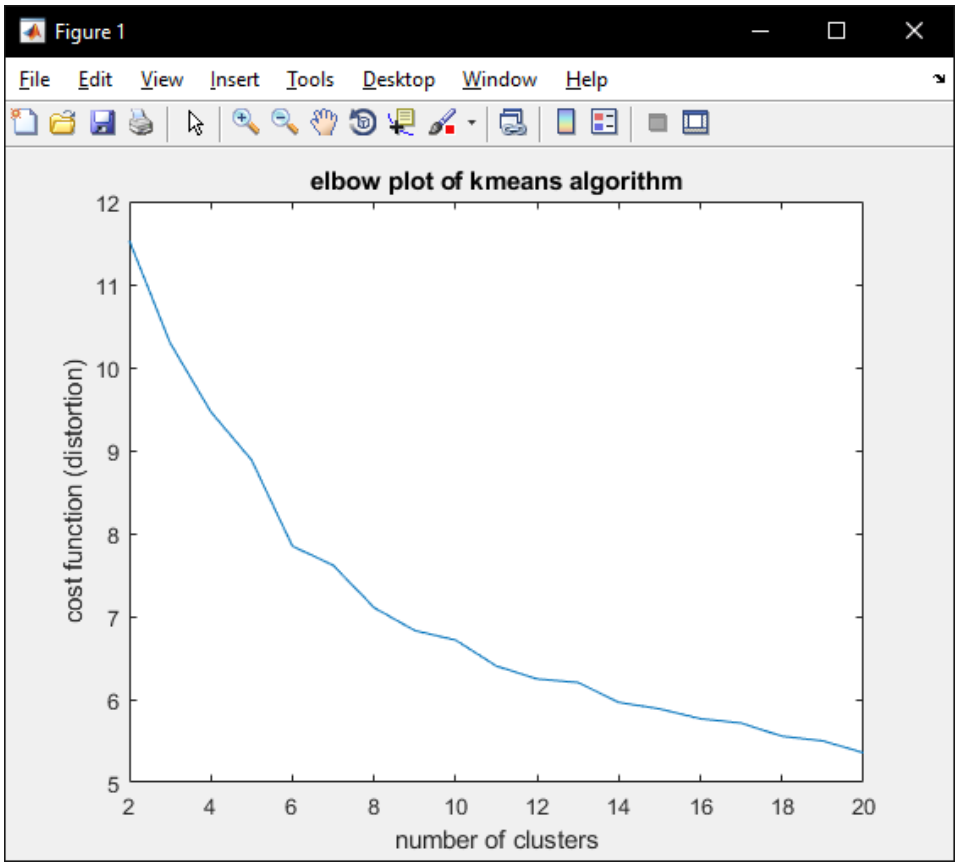
Um das «Problem» mit den kategorischen Features zu umgehen, könnte eine andere Distanzmetrik verwendet werden, wie z.B. die Hamming oder Jaccard Distanz. Dies ist aber hier wohl out of scope und ohnehin nicht nötig, da es ja genügend Features zur Analyse gibt. Eine weitere Möglichkeit wäre natürlich, ein one-hot encoding zu nutzen. Auch darauf verzichte ich aber mit der Begründung, dass auch ohne diese paar kategorischen Features mehr als genügend Features für ein sinnvolles Clusterung vorhanden sind.

### 1.2 Bestimmung der Anzahl Cluster K

#### 1.2.1 Elbow Plot

Eine mögliche Variante zur Wahl der Anzahl Cluster ist der Elbow Plot. Ich verwende den K-Means Clustering Algorithmus mit 20 zufälligen Initialisierungen (s. Code unten), woraus der darauffolgende Plot resultiert.

```
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
```

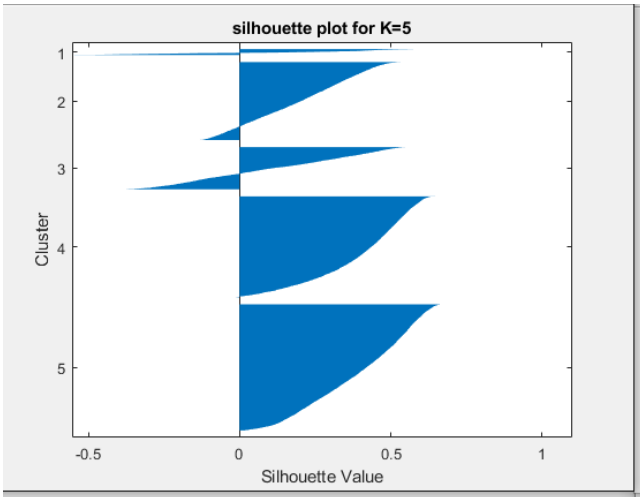


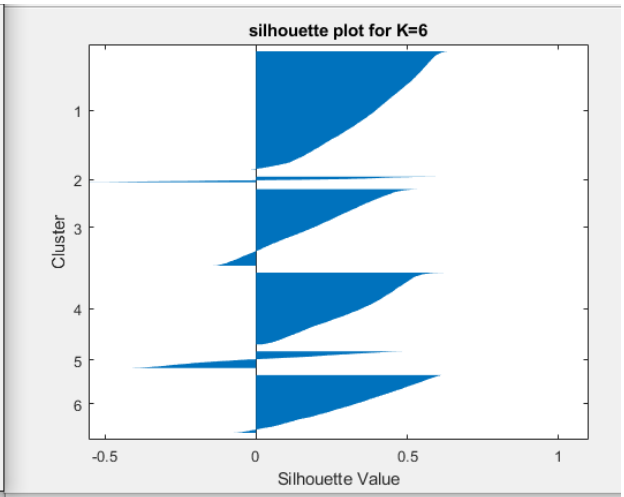

Bei diesem Plot ist nur ein schwacher «Elbow» zu erkennen. Am ehesten kommt wohl der Wert 6 in Frage, da die Kurve dort den grössten Knick macht.

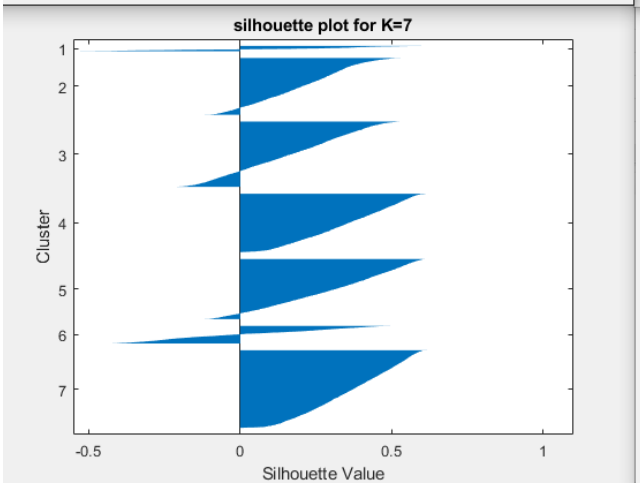
### 1.2.2 Silhouette Plot

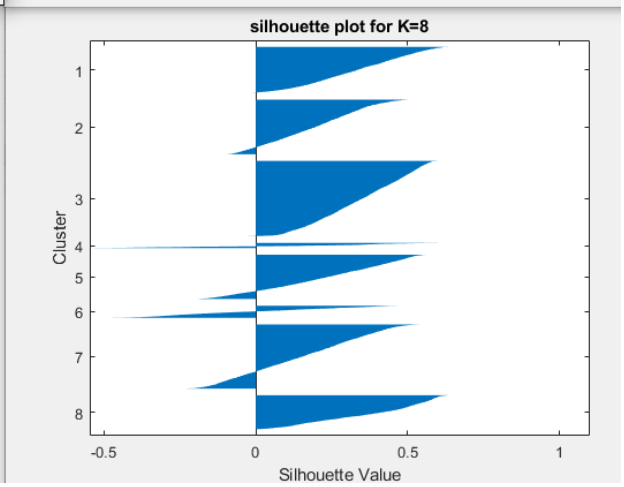
K=6 scheint gemäss Elbow Plot keine schlechte Wahl zu sein. Ich berechne hier die Silhouettenwerte für Werte von K im Bereich von 5 bis 9, zum Testen, ob ev. 6 auch wirklich die beste Wahl ist, oder ob leicht höhere oder tiefere Werte womöglich besser sind.

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.


silhouette plot for K=5
Cluster
Silhouette Value
1
2
3
4
5
-0.5
0
0.5
1


silhouette plot for K=6
Cluster
Silhouette Value
1
2
3
4
5
6
-0.5
0
0.5
1


silhouette plot for K=7
Cluster
Silhouette Value
1
2
3
4
5
6
7
-0.5
0
0.5
1


silhouette plot for K=8
Cluster
Silhouette Value
1
2
3
4
5
6
7
8
-0.5
0
0.5
1

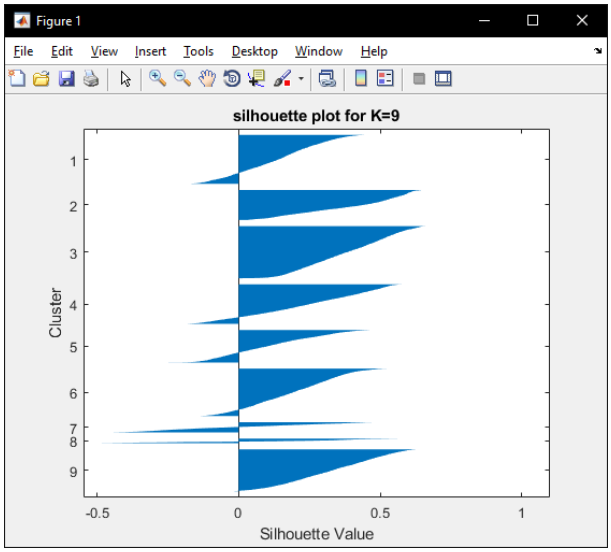
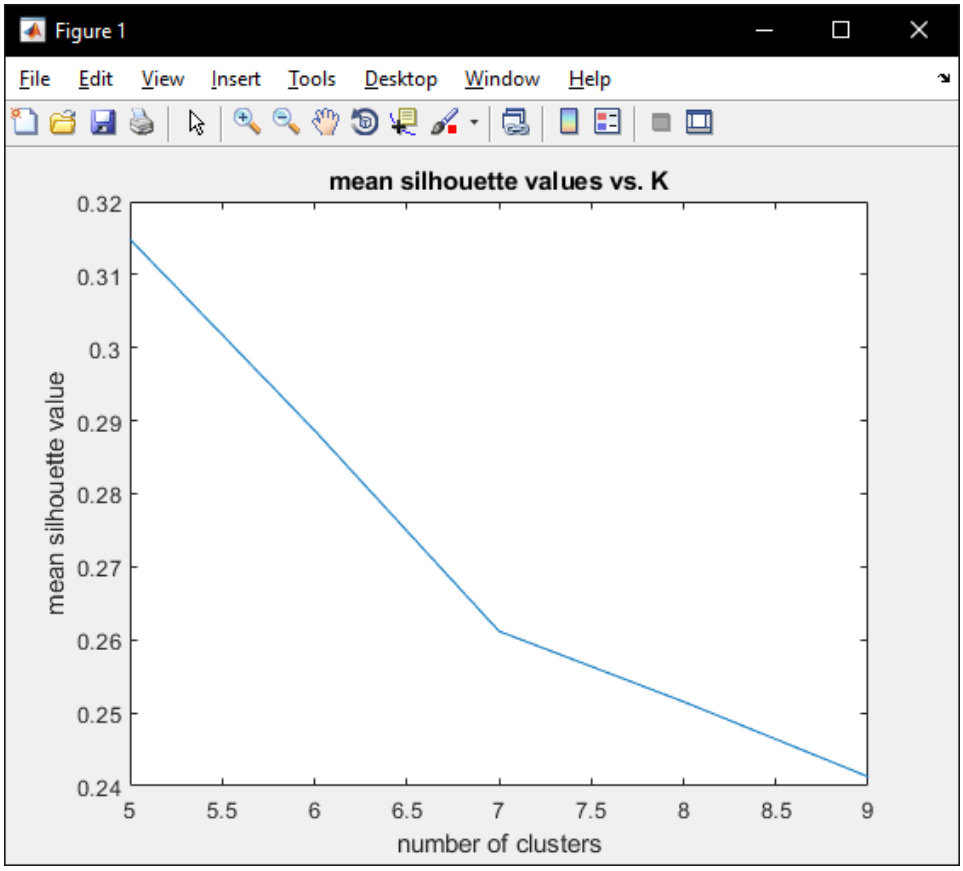


Figure 1
File Edit View Insert Tools Desktop Window Help
silhouette plot for K=9
Cluster
Silhouette Value
1
2
3
4
5
6
7
8
9
-0.5
0
0.5
1

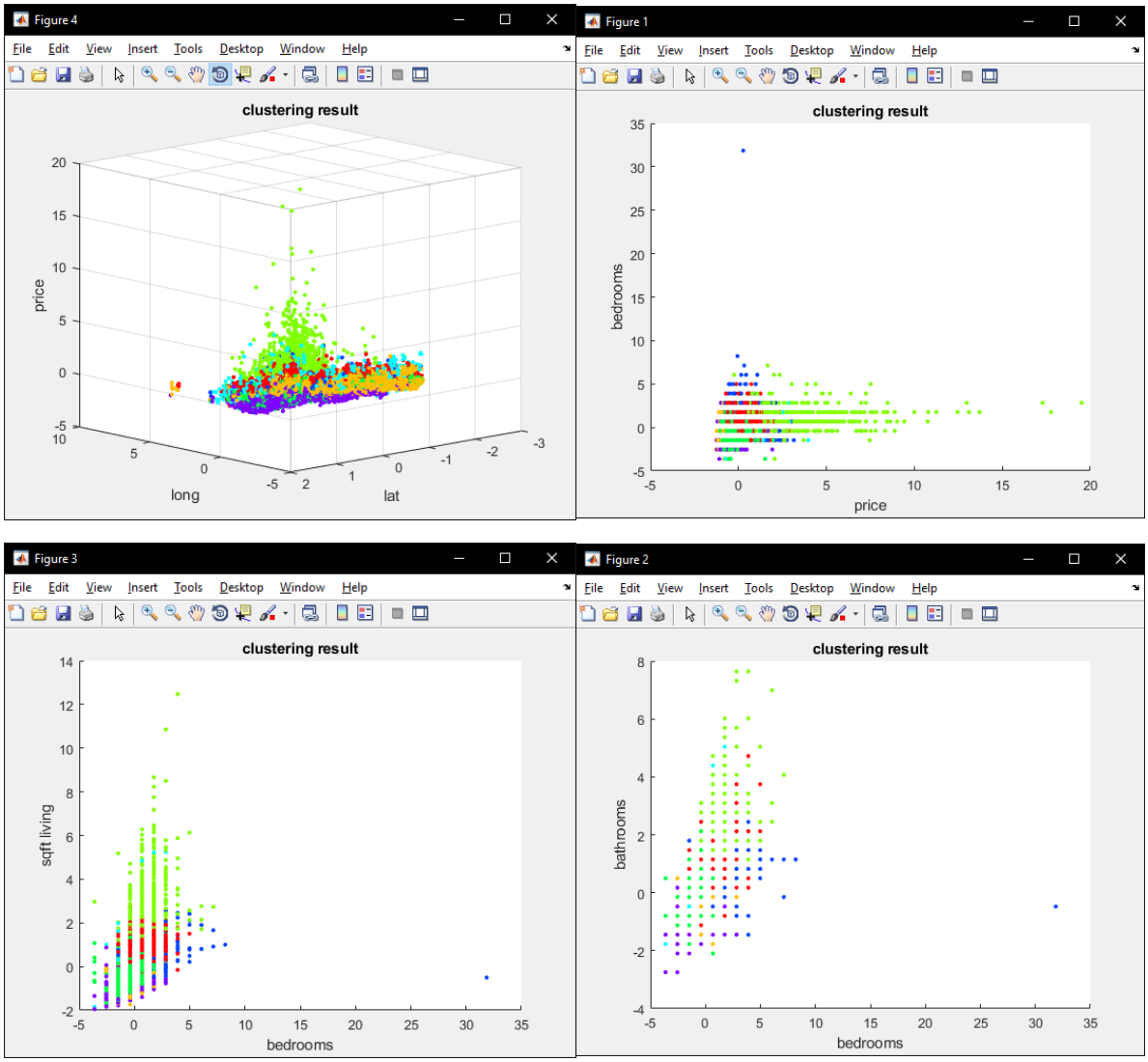



Bei der Betrachtung der Durchschnittlichen Silhouettenwerte gegenübergestellt mit K sieht ist ein leichter Knick bei K=7 zu sehen. Auch bei der Betrachtung der Silhouetten-Plots sieht man bereits, dass aus K=7 eine leicht bessere/ausgeglichenere Verteilung der Punkte resultiert als aus K=6.

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

## 1.3 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 Feature-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 unter anderem zu erkennen, dass grössere Häuser mehr Schlafzimmer haben und auch entsprechend gruppiert werden oder dass die Anzahl Schlaf- und Badezimmer eine gewisse Korrelation aufweisen (was natürlich zu erwarten war).

Es ist zu bemerken, dass die Farben hier nur die Zugehörigkeiten zu den Clustern anzeigen. Die Farbe-Cluster Beziehung ist irrelevant und könnte bei mehrmaligen ausführen des Codes ändern (wegen der Zufälligkeit des K-Means Algorithmus).

### 1.3.1 Genauere Analyse der Cluster

Nachfolgend berechne ich den Mittelwert und die Varianz jedes Features 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. Es gilt zu beachten, dass die Varianzen und Mittelwerte aus den bereits normalisierten Features berechnet wurden.

```
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

# 2 Aufgabe: Clustering nach Dimensionsreduktion mit PCA

Bei dieser Aufgabe verwende ich die genau gleichen (normalisierten) Features wie bei der vorherigen Aufgabe.

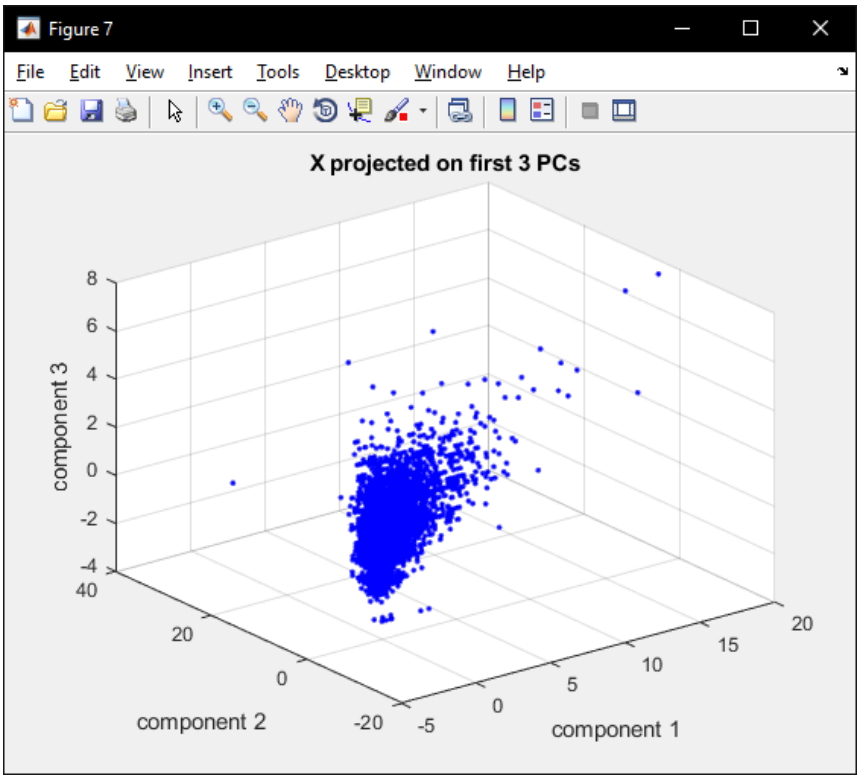
Vor dem Clustering wird aber diesmal der Feature Space auf die ersten drei Principal Components projiziert. Dadurch verkleinert sich natürlich der Feature Space und das Clusterung wird schneller. Die Frage stellt sich, ob sich durch diese Dimensionsreduktion die Qualität des Clusterung verschlechtert oder nicht.

Die neue Feature Matrix Z wird folgendermassen berechnet. Für weitere Details bitte direkt den Code konsultieren.

```
[U, S] = pca(X);
n_components = 3;
Z = projectData(X, U, n_components);
```

## 2.1 Visualisierung und potentiell negative Einflüsse für das Clustering

Jetzt befinden wir uns im 3-dimensionalen Raum und können die ersten drei Principal Components einfach visualisieren:




In dieser Visualisierung sind nicht wirklich verschiedene Cluster zu erkennen, sondern vielmehr einfach eine Ansammlung von Datapoints, die alle relativ eng zusammenhängen, abgesehen von ein paar Ausreissern. Gerade diese Ausreisser könnten das Clustering unter Umständen negativ beeinflussen. Beispielsweise gibt es ganz wenig Häuser mit einer sehr hohen Anzahl Schlafzimmer, wie den

Visualisierungen in Aufgabe 1 entnommen werden kann. Um dennoch ein gutes Clusterung zu erhalten, lasse ich den K-Means Algorithmus aber jeweils mit 20 zufälligen Initialisierungen laufen und verwende das beste Endergebnis. Das dauert natürlich entsprechend länger, ist aber eine effektive Taktik um ungute Resultat von «schlechten» Zufallsinitialisierungen zu vermeiden.

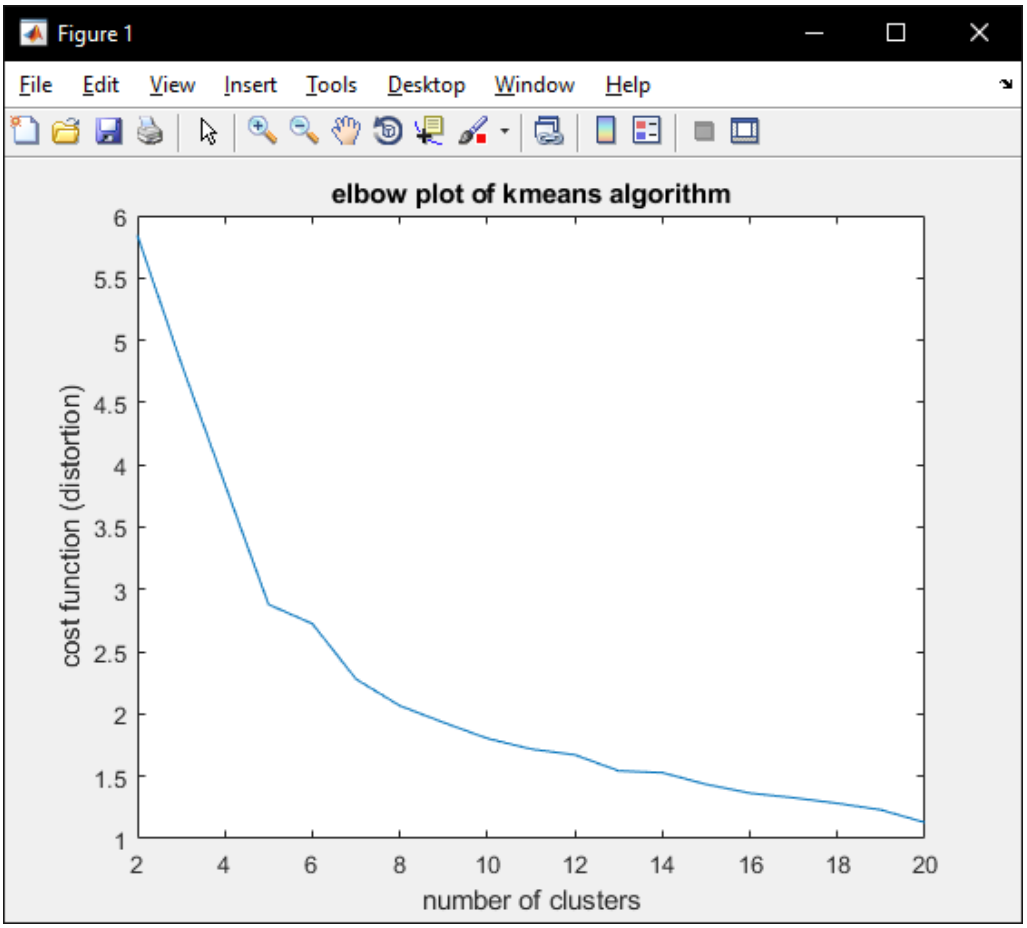
Weiter würden kategorische Features das Ergebnis tendenziell verschlechtern, zumindest solange sie nicht vorher bearbeitet werden, z.B. mit einem one-hot Encoding. Dies würde aber wiederum die Dimensionalität und somit auch die Laufzeit erhöhen.

Eine weitere Eigenschaft, die potentielle Schwierigkeit an der Cluster Einteilung aufbringen könnte ist, dass man auch jetzt nach Anwendung von PCA nicht wirklich klar erkennbare Cluster entstanden sind (oder zumindest für uns nicht direkt sichtbar), sondern eben eher nur ein grosser Blob.

## 2.2 Bestimmung der Anzahl Cluster K

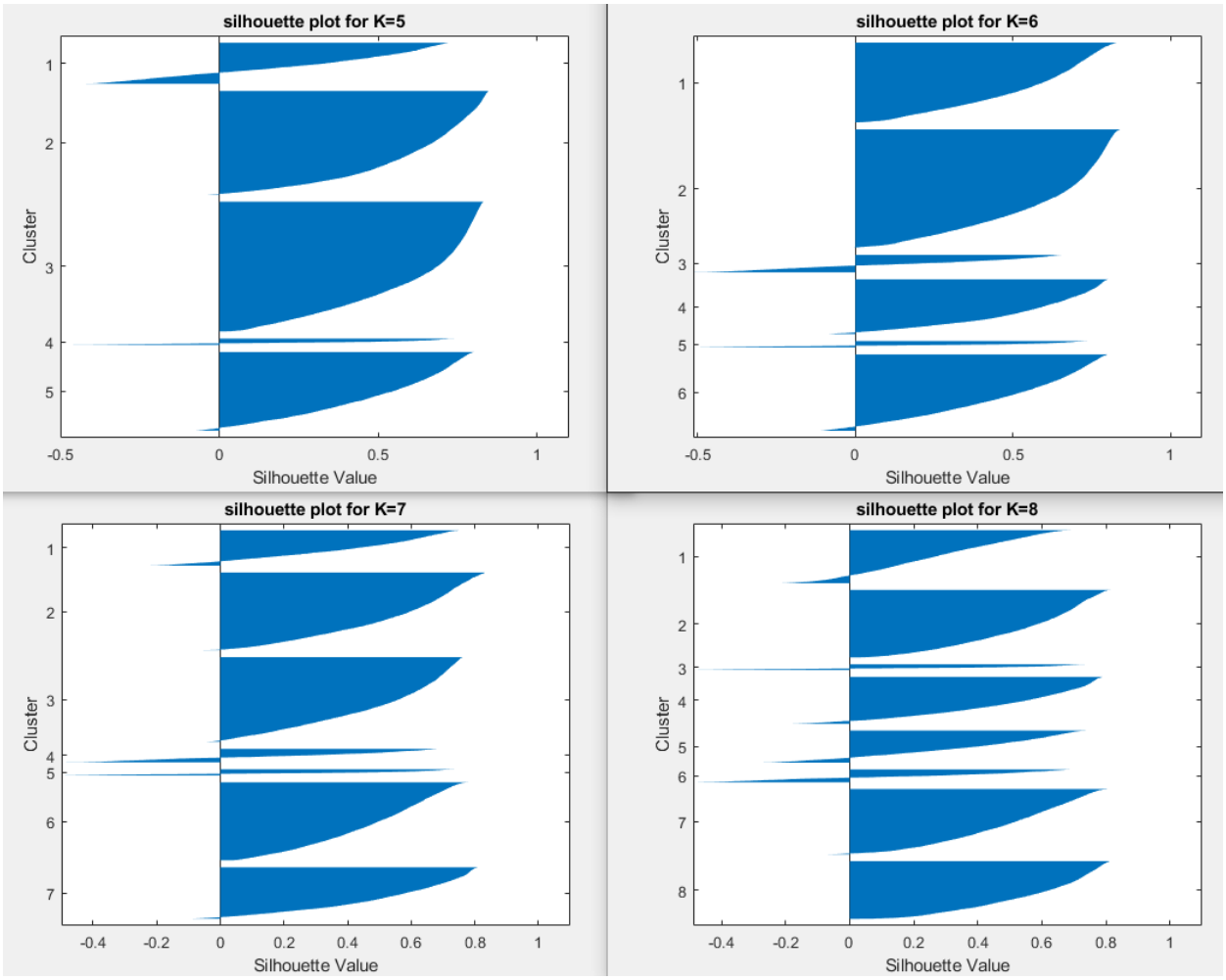
### 2.2.1 Elbow Plot

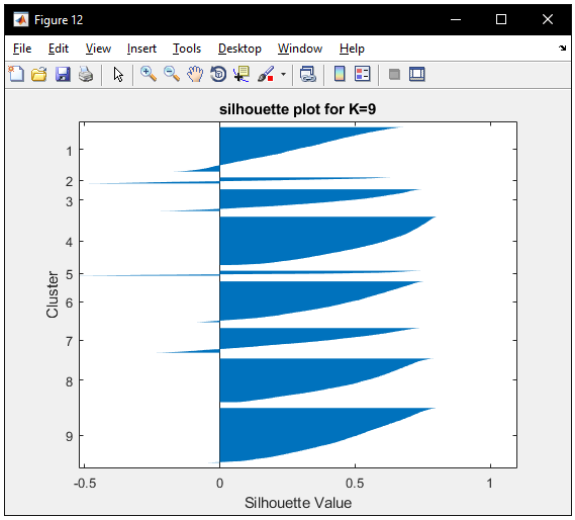
Das Clustering wird auf die genau gleiche Weise mit der gleichen Parametrisierung wie bei der ersten Aufgabe durchgeführt. Der Resultierende Elbow-Plot sieht ähnlich aus, wie beim Clustering ohne vorherige Dimensionsreduktion. Gemäss Plot käme am ehesten K=5 oder K=7 in Frage, dort gibt es jeweils einen kleinen «Elbow».

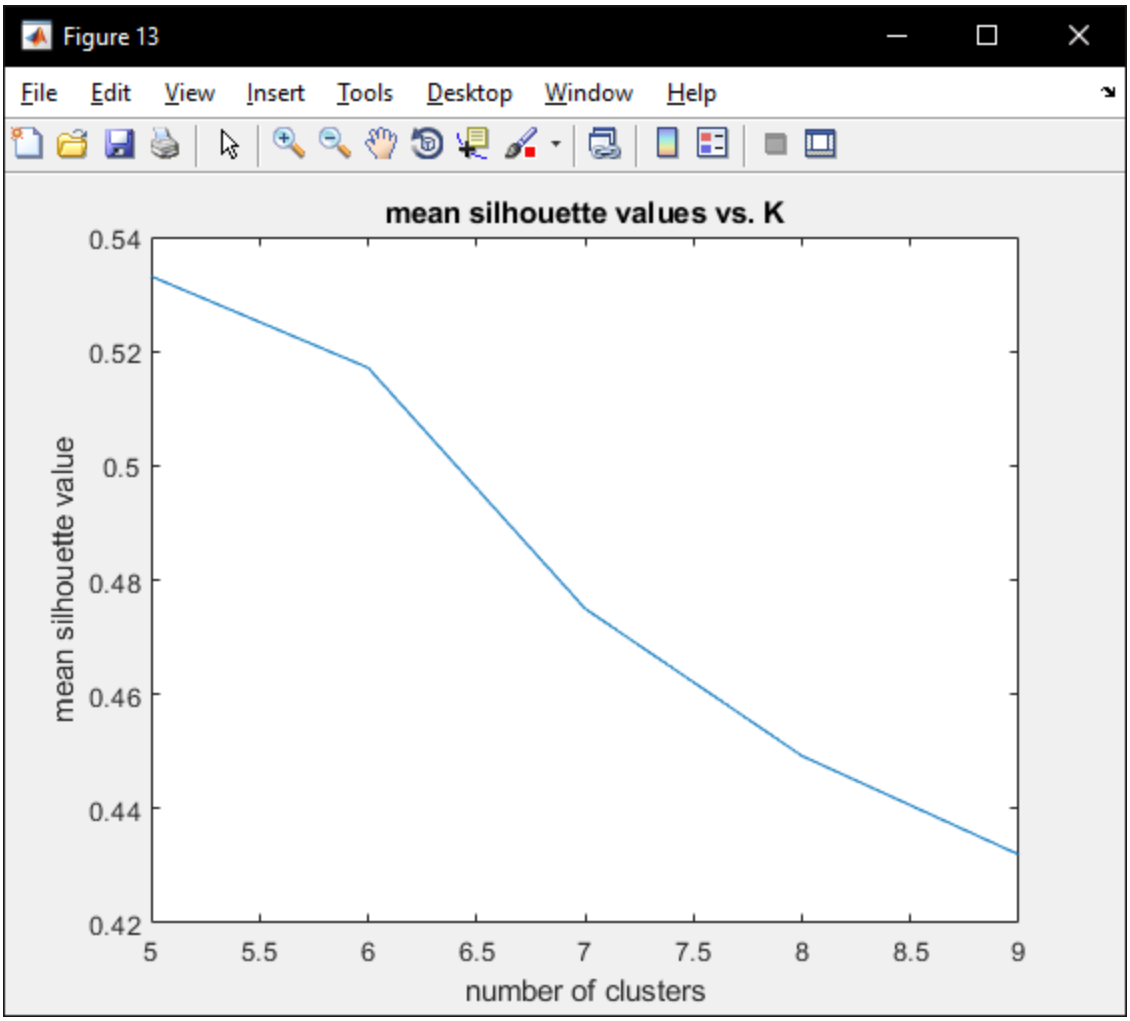


### 2.2.2 Silhouette Plot

Zur genaueren Analyse werden nachfolgend nochmals die Silhouetten berechnet für K-Werte im Bereich von 5 bis 9 und deren Mittelwerte gegenüber K verglichen.



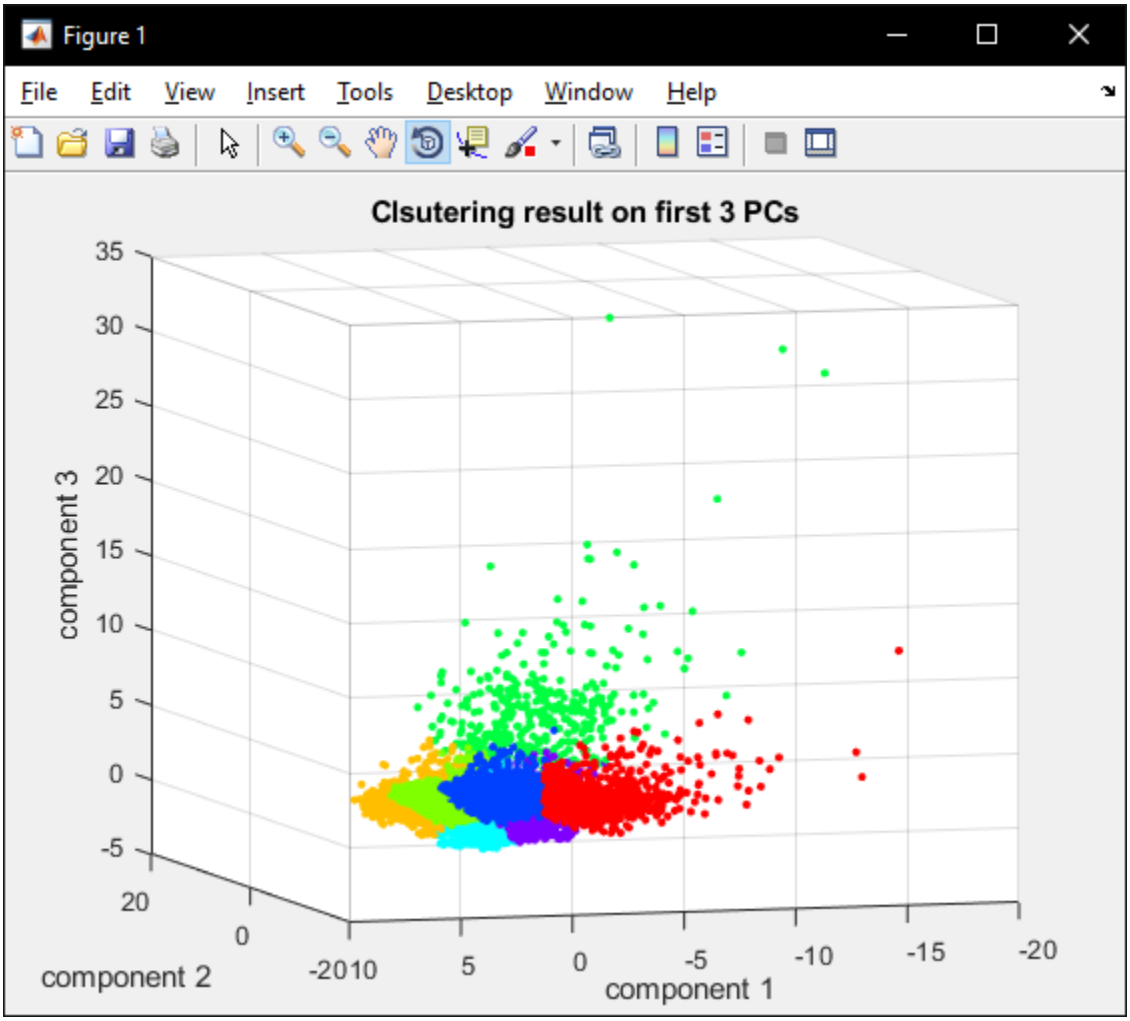





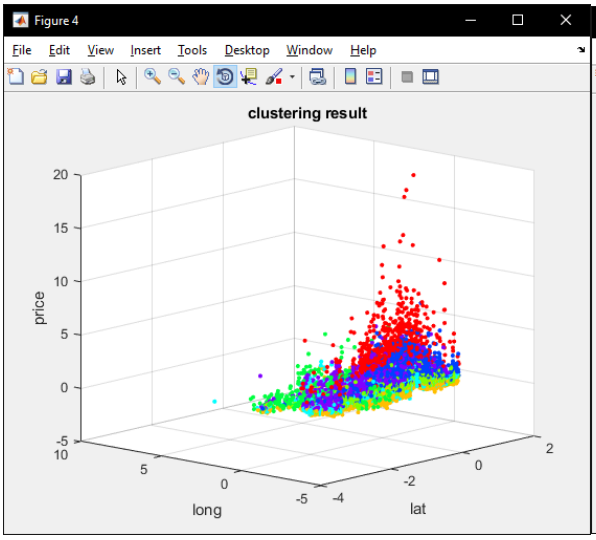

Hier ist ein ähnliches Bild zu sehen, wie bei der vorherigen Aufgabe, die Silhouetten-Werte sind jedoch um einiges höher. Wiederum scheint K=7 eine geeignete Wahl zu sein.

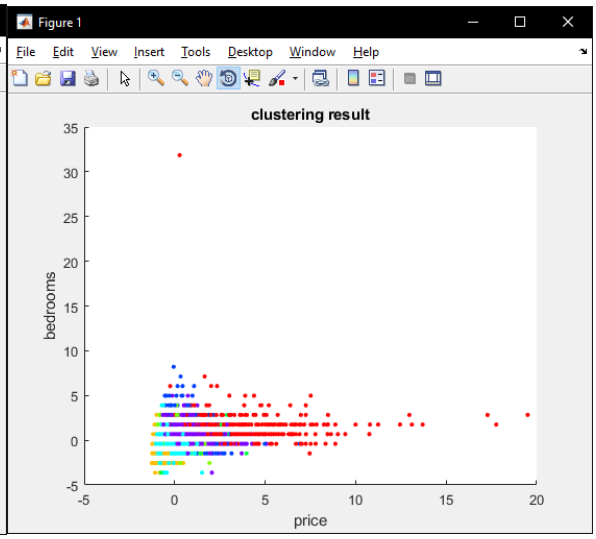
## 2.3 Cluster Visualisierung und Charakterisierung

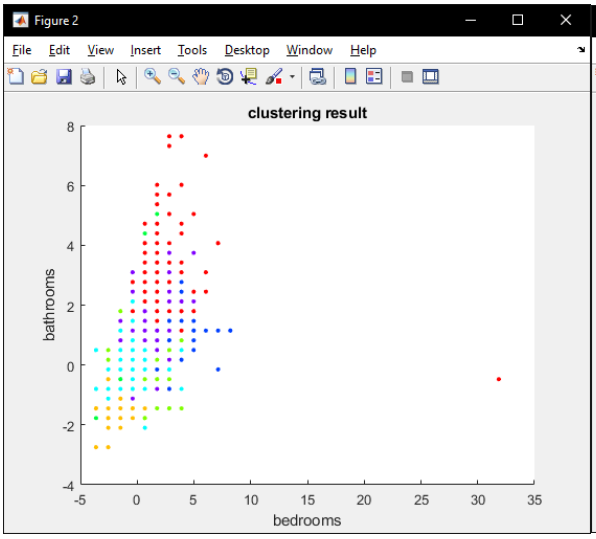
Da eine Dimensionsreduktion stattgefunden hat, können die Cluster hier sehr gut visualisiert werden:

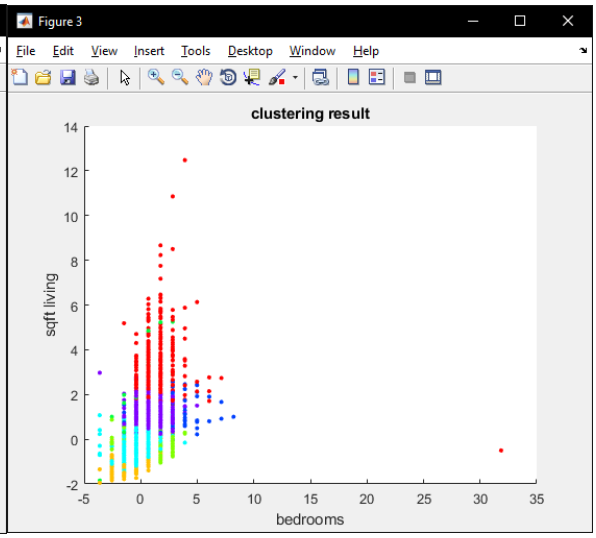



Einfach so zum Vergleich habe ich nachfolgend die Cluster noch anhand der in der vorherigen Aufgabe gewählten Features visualisiert. Die Farben sind zwar anders, aber die sind ohnehin nicht relevant, es geht nur um die Zugehörigkeit. Und diese scheint zumindest visuell sehr ähnlich wie vorher zu sein, man kann die gleichen Korrelationen wie bei Aufgabe 1 feststellen. Bis jetzt schein es also nicht, als hätte die Dimensionsreduktion eine wesentliche Verschlechterung des Resultats gebracht.




### 2.3.1 Genauere Analyse der Cluster

Nachfolgend die Varianzen und Mittelwerte der (neuen) Features pro Cluster. Die Cluster-Nummern entsprechen nicht den vorherigen Cluster-Nummern, werden aber in einem späteren Schritt korreliert.

```
cluster 1 contains 3.9% of the datapoints:
        lat             : var = 0.32,     mean = 0.34
        floors          : var = 0.54,     mean = 0.81
        grade           : var = 0.67,     mean = 2.36
        long            : var = 0.68,     mean = 0.32
        sqft lot15      : var = 0.77,     mean = 0.29
        yr built        : var = 0.78,     mean = 0.58
        condition       : var = 0.82,     mean = -0.18
        sqft lot        : var = 0.85,     mean = 0.27
        bathrooms       : var = 1.05,     mean = 2.04
        sqft living     : var = 1.25,     mean = 2.74
        sqft living15   : var = 1.35,     mean = 2.24
        sqft above      : var = 1.47,     mean = 2.45
        bedrooms        : var = 2.02,     mean = 1.23
        sqft basement   : var = 3.02,     mean = 1.10
        price           : var = 4.86,     mean = 2.94
```

```
cluster 2 contains 24.5% of the datapoints:
        sqft living    : var = 0.10,     mean = -0.97
        price          : var = 0.10,     mean = -0.62
        sqft lot       : var = 0.10,     mean = -0.13
        sqft above     : var = 0.11,     mean = -0.81
        sqft lot15     : var = 0.13,     mean = -0.14
        sqft basement  : var = 0.15,     mean = -0.49
        floors         : var = 0.19,     mean = -0.75
        sqft living15  : var = 0.21,     mean = -0.81
        bathrooms      : var = 0.25,     mean = -1.11
        grade          : var = 0.30,     mean = -0.91
        bedrooms       : var = 0.51,     mean = -0.77
        yr built       : var = 0.58,     mean = -0.65
        long           : var = 0.82,     mean = -0.29
        lat            : var = 1.09,     mean = -0.22
        condition      : var = 1.10,     mean = 0.08

cluster 3 contains 22.5% of the datapoints:
        sqft lot       : var = 0.06,     mean = -0.14
        sqft lot15     : var = 0.08,     mean = -0.15
        sqft living    : var = 0.13,     mean = -0.21
        sqft above     : var = 0.17,     mean = -0.48
        price          : var = 0.20,     mean = -0.14
        grade          : var = 0.22,     mean = -0.37
        sqft living15  : var = 0.28,     mean = -0.28
        floors         : var = 0.30,     mean = -0.63
        bathrooms      : var = 0.35,     mean = -0.28
        long           : var = 0.48,     mean = -0.42
        yr built       : var = 0.63,     mean = -0.65
        bedrooms       : var = 0.64,     mean = 0.09
        sqft basement  : var = 0.74,     mean = 0.45
        lat            : var = 0.83,     mean = 0.27
        condition      : var = 1.23,     mean = 0.47

cluster 4 contains 1.8% of the datapoints:
        yr built       : var = 0.52,     mean = 0.40
        price          : var = 0.55,     mean = 0.15
        sqft living15  : var = 0.79,     mean = 0.52
        floors         : var = 0.81,     mean = 0.03
        bedrooms       : var = 0.82,     mean = -0.08
        condition      : var = 0.93,     mean = -0.10
        bathrooms      : var = 1.01,     mean = 0.30
        grade          : var = 1.24,     mean = 0.31
        sqft living    : var = 1.26,     mean = 0.61
        sqft basement  : var = 1.27,     mean = -0.05
        sqft above     : var = 1.31,     mean = 0.71
        long           : var = 1.36,     mean = 1.40
        lat            : var = 1.37,     mean = -0.69
        sqft lot15     : var = 12.48,    mean = 5.82
        sqft lot       : var = 18.55,    mean = 5.39

cluster 5 contains 22.4% of the datapoints:
        sqft lot       : var = 0.08,     mean = -0.18
        sqft lot15     : var = 0.10,     mean = -0.20
        sqft basement  : var = 0.12,     mean = -0.52
        price          : var = 0.14,     mean = -0.34
        sqft living    : var = 0.20,     mean = -0.26
        yr built       : var = 0.23,     mean = 0.86
        condition      : var = 0.24,     mean = -0.50
        bathrooms      : var = 0.26,     mean = 0.31
        grade          : var = 0.28,     mean = 0.02
        sqft above     : var = 0.29,     mean = -0.01
        sqft living15  : var = 0.35,     mean = -0.22
        bedrooms       : var = 0.52,     mean = -0.21
        floors         : var = 0.86,     mean = 0.87
        lat            : var = 1.15,     mean = -0.27
        long           : var = 1.22,     mean = 0.26
```

```
cluster 6 contains 10.0% of the datapoints:
        sqft lot15     : var = 0.14,     mean = -0.05
        sqft lot       : var = 0.14,     mean = -0.04
        sqft above     : var = 0.29,     mean = 0.07
        sqft living    : var = 0.29,     mean = 0.78
        bathrooms      : var = 0.45,     mean = 0.58
        grade          : var = 0.45,     mean = 0.43
        long           : var = 0.53,     mean = -0.28
        lat            : var = 0.54,     mean = 0.39
        sqft living15  : var = 0.59,     mean = 0.57
        floors         : var = 0.66,     mean = -0.37
        yr built       : var = 0.69,     mean = -0.50
        price          : var = 0.82,     mean = 0.72
        bedrooms       : var = 1.02,     mean = 0.83
        sqft basement  : var = 1.06,     mean = 1.48
        condition      : var = 1.34,     mean = 0.63

cluster 7 contains 14.9% of the datapoints:
        yr built       : var = 0.20,     mean = 0.90
        floors         : var = 0.21,     mean = 0.92
        sqft lot       : var = 0.23,     mean = 0.00
        sqft living    : var = 0.26,     mean = 0.99
        condition      : var = 0.26,     mean = -0.46
        sqft lot15     : var = 0.28,     mean = 0.03
        sqft above     : var = 0.34,     mean = 1.30
        bathrooms      : var = 0.36,     mean = 0.83
        price          : var = 0.45,     mean = 0.47
        sqft basement  : var = 0.45,     mean = -0.38
        grade          : var = 0.51,     mean = 1.08
        bedrooms       : var = 0.55,     mean = 0.59
        sqft living15  : var = 0.61,     mean = 1.06
        lat            : var = 0.85,     mean = 0.08
        long           : var = 0.85,     mean = 0.67
```

Aus den obigen Daten können die Cluster wie folgt charakterisiert und nach einem manuellen Vergleich mit einem der vorherigen Clustern (in Klammern) korreliert werden:

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

## 2.4 Unterschied zum vorherigen Clustering Ergebnis (ohne PCA)

Die Cluster dieser Aufgabe können mit einem einfachen Vergleich der Varianzen, Mittelwerte und Grössen relativ gut den Clustern der vorherigen Aufgabe zugeordnet werden (es wurde ja auch wieder K=7 gewählt).

| **Cluster Aufgabe 1** | **Cluster Aufgabe 2** | **Unterschied (von Aufgabe 1 zu Aufgabe 2)** |
|---|---|---|
| 1 | 7 | Diese zwei Cluster sind enorm ähnlich, es gibt keine Signifikanten Unterschiede, sowohl bei der Grösse als auch bei den Varianzen und Mittelwerten der Features. |
| 2 | 2 | Der zweite Cluster ist um etwa das 1.5 fache grösser und die condition ist etwas höher. Ansonsten sehr ähnlich. |
| 3 | 1 | Sehr ähnlich, wobei der erste Cluster noch ein wenig grösser ist. Beide Cluster enthalten sehr teure Häuser, wobei in der zweiten Variante die wohl noch ein wenig «besseren» Häuser sind. |
| 4 | 5 | Auch sehr ähnlich, nur ist der erste Cluster hier leicht kleiner und ist ein bisschen mehr fixiert auf neue, zentraler gelegene Häuser. |
| 5 | 4 | Die beiden kleinsten Cluster, beide enthalten preiswerte Häuser, der erste ist wirklich sehr auf preiswerte und eher Häuser in schlechten Zustand fokussiert während der zweite leicht weniger restriktiv ist. |
| 6 | 6 | Hier gibt es eine signifikante Differenz der Grösse, der zweite Cluster enthält nur fast halb so viele Daten wie der erste. Die Häuser im zweiten sind leicht grösser, und die fehlenden befinden sich dafür wohl mehrheitlich im Cluster 2, welcher ohnehin eine grosse Ähnlichkeit zu diesem aufweist (eher billigere/schlechtere Häuser). |
| 7 | 3 | Erstaunlich identisch, nur das Baujahr ist im zweiten leicht höher und er enthält auch noch einige mehrstöckige Häuser. |

## 2.5 Fazit

Da viele der Features eine gewisse Korrelation aufweisen (zum Beispiel Anzahl Schlafzimmer und Badezimmer oder die ganzen Wohnflächen-Angaben (sqft) untereinander) ist es gut möglich, auch mit einer Reduktion des Featurespaces auf die ersten drei Principal Components noch ein relativ gutes Clustering Ergebnis zu erhalten.

In diesem Sinne genügt es für diesen Anwendungsfall sehr wohl, nur die drei ersten Principal Components für das Clustering zu verwenden. Der Featurespace und somit der Ressourcenbedarf des Algorithmus wird reduziert, ohne dass ein signifikanter Qualitätsverlust auftritt – Im Gegenteil, wie schon weiter oben genannt weisen viele der Features eine hohe Korrelation auf, wodurch es auf jeden Fall sinnvoll ist, die Dimensionalität zu reduzieren und so einiges an Performance zu gewinnen. Es resultieren trotzdem noch sehr aussagekräftige Cluster. Würde man zusätzlich noch die one-hot codierten kategorischen Features verwenden, müsste man wohl auf mehr als 3 Komponenten zurückgreifen, um immer noch ein nutzbringendes Resultat zu erhalten.

Weiter wird durch die Reduktion auf 3 Dimensionen natürlich auch die Visualisierung stark vereinfacht.