**1. Load the data set and view it using the View()command.**

Siehe Dokumentation

**2. Discuss the dataset based on str() and summary(). What does it tell you about the houses?**

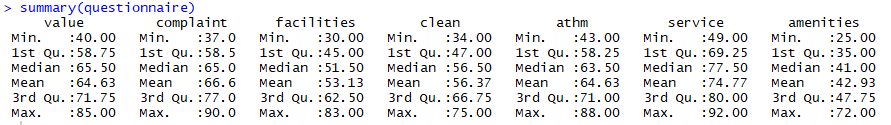
**I.e., what are the most problematic aspects (attributes) the management should focus on?**

Insgesamt fällt auf, dass sehr wenige Daten vorhanden sind, um wirklich sinnvoll clustern zu können!

Basierend auf summary() über alle Häuser gesehen fällt folgendes auf:

1. "Amenities" ist das Kriterium, das insgesamt am wenigsten positive Antworten bekommen hat:

* Tiefster Median
* Tiefster Durschnitt
* Tiefstes 1. Quartil
* 3. Quartil ist tiefer als das 1. Quartil aller anderen Kriterien
* Tiefstes Maximum
* Tiefstes Minimum



1. "Facilities" und "clean" sind die zweitschlechtesten Bereiche, clean ist leicht besser bewertet als "Facilities".
2. Service ist in allen Häusern sehr gut bewertet worden

Management sollte insbesondere die Infrastruktur (Amenities) der Häuser verbessern (z.B. W-LAN anbieten, mehr Sender im TV zeigen usw.) und die Sauberkeit (clean) erhöhen. Auch die Ausstattung der Badezimmer sollte verbessern werden.

* Je nach genauen Kritikpunkten der Kunden an den Badezimmern könnte eine allgemein höhere Sauberkeit bereits auch die Bewertung der Badezimmer verbessern

**3. Apply K-means with 4 clusters.**

Siehe Dokumentation

**4. Evaluate your clustering results using the evaluation measures discussed in class. Interpret them.**

4 Cluster: Hoher Wert für Total withinss (schlecht): Abstände der Datenpunkte in den Clustern zu den Cluster-Zentren ist eher hoch



Tiefer Wert für betweenss (schlecht): Cluster scheinen nah beieinander zu sein, Ratio ist ebenfalls tief mit 58,7%. Scheint kein gutes Clustering zu sein.

Mit mehr Clustern werden bessere Ergebnisse erreicht, bspw. mit 8 Clustern:



Im Vergleich zum k-means mit 4 Clustern ist bei 8 Clustern:

* Total Withinss tiefer (gut)
* betweenss höher (gut)
* Ratio höher (ebenfalls gut)

**5. Interpret your clustering results:**

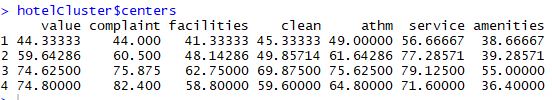
1. What is the biggest, what is the smallest cluster?

Der grösste Cluster ist Nr 1. , der klenste Cluster Nr. 1



1. Inspect the coordinates of the cluster means (i.e. the cluster centers) and interpret them:

What are the groups of houses the management should focus on, and why? For each group of houses, what are the most problematic aspects?



Das Management sollte sich insbesondere stark auf die Häuser der Gruppe 1 konzentrieren, aber auch bei den Häusern der Gruppe 3 sollten diverse Punkte verbessert werden.

Die problematischen Aspekte der 4 Gruppen sind:

**Gruppe 1:** Alle Attribute schlecht bewertet (auch der Service fällt stark ab zu den anderen Gruppen)

**Gruppe 2:** Facilities, Clean und amenities sind schlecht bewertet

**Gruppe 3:** Amenities

**Gruppe 4**: Clean, Amenities (Amenities ist bei Gruppe 4 am schlechtesten bewertet von allen Gruppen)

1. Estimate the average percentage of positive answers per cluster (average over the attribute values of each cluster center-point) to get an overall performance KPI for each cluster of houses. Which cluster has the best performance on average?



Im Durchschnitt gesehen performen die Häuser des 3. Clusters mit Abstand am besten

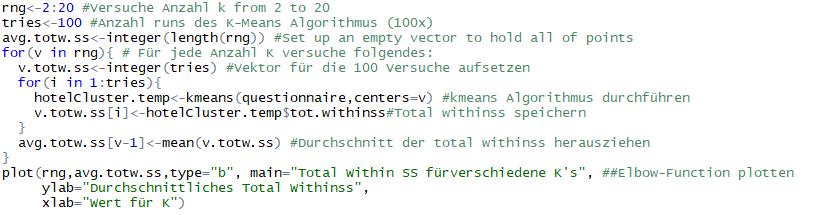
**6. Try some other values for the number of clusters (instead of 4). Write down the resulting**

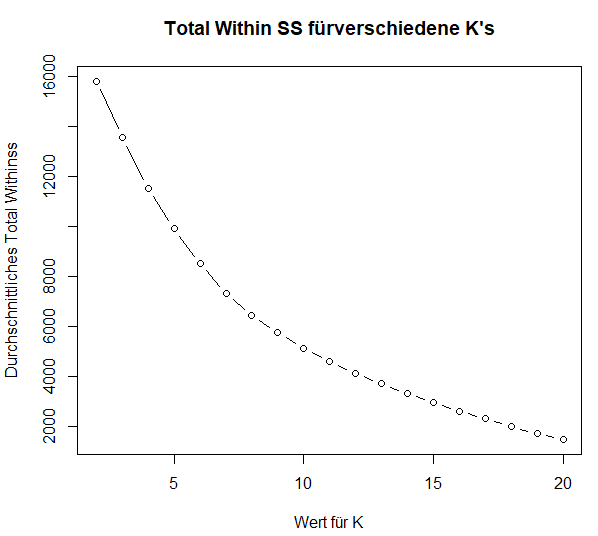
**withinss values and draw the respective graph. What do you observe? Are 4 clusters a good**

**choice?**

4 Clusters sind keine gute Wahl, besser wäre eine Anzahl Cluster zwischen 5 – 7, ab dann sinkt der Wert für "total withinss" immer weniger stark, der zusätzliche Nutzen von mehr Clustern inmmt immer mehr ab.

Schlussendlich ist das Datenset aber nicht optimal/zu klein um gut clustern zu können. Die schlechten Häuser können bei dieser Anzahl schon gut ohne Cluster herausgesehen werden. Clustern würde erst ab einer grösseren Anzahl Häuser auch wirklich Sinn machen





(Siehe Dokumentation für weitere Details)