**Массив данных**

В качестве массива данных для выполнения задачи кластеризации был выбран датасет в котором содержатся данные анализов крови здоровых людей и людей, зараженных гепатитом С. В каждой записи содержатся следующие данные:

* Номер пациента
* Диагноз
* Возраст
* Пол
* Значения анализа крови

В качестве целевого параметра выбран диагноз пациента – здоров или заражен. В датасете здоровые пациенты указаны как «blood donors». Зараженные же описаны строками «Hepatitis C», «Fibrosis», «Cirrhosis», то есть включает как сам гепатит С, так и предшествующие ему болезни.

**Дескриптивный анализ**

Для анализа выбираются значения анализа крови. Необходимо провести автоматический анализ как в эксель, чтобы вывелись основные параметры (мат ожидание, отклонение и прочие статистические параметры). После этого построить гистограммы по выбранным данным. Возможно, придется найти минимум и максимум среди всех значений и посчитать количество карманов. Оценить визуальную близость гистограммы к нормальному распределению.

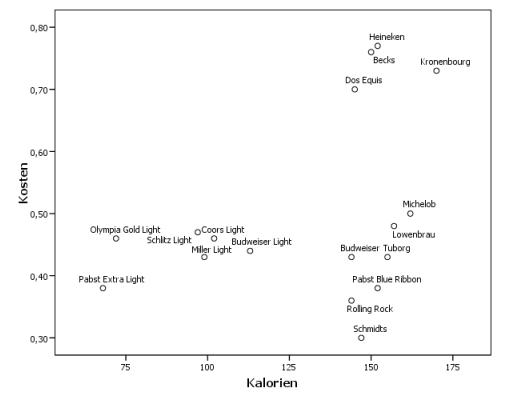
На основе полученных результатов и выводов определить методы для кластеризации.

**Стандартизация данных**

Нужно для перевода всех исследуемых параметров к единому диапазону значений. В SPSS это от -3 до 3. Преобразование происходит с помощью z-преобразования. Более подробно этот процесс расписан в 9 презентации от НАФИ.

**Диаграммы рассеивания**

Это способ графически оценить количество кластеров. Представляют собой диаграммы в 2д/3д, отображающие расположение каждого измерения/записи относительно выбранных 2-х/3-х параметров. По примеру ниже выбран датасет с 17 сортами пива. Диаграммы строятся по калорийности (Kalorien) и стоимости (Kosten). Ниже приведена полученная диаграмма:



По диаграмме видно, что получилось 3 кластера.

**Решение задачи кластеризации двумя методами**

* K-средние и иерархический для SPSS описаны в 9 лекции от НАФИ
* EM, DBSCAN, карта Кохонена – искать дальше