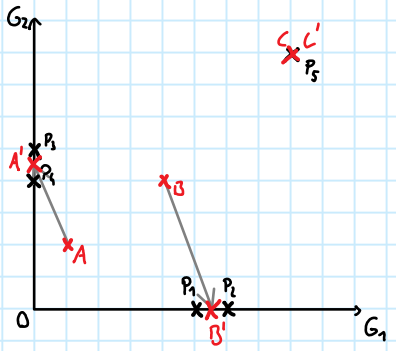
Hausaufgabe 8

	P_1	P_2	P_3	P_4	P_5
G_1	5	6	0	0	8
G_2	0	0	5	4	8

Die initiale Zentren liegen bei $A(1,2)$; $B(4,4)$, $C(8,8)$



Euklidische Distanz $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$

$d_{1A} \approx 4{,}472$ $d_{1B} \approx 4{,}123$ } Cluster um B
$d_{2A} \approx 5{,}385$ $d_{2B} \approx 5{,}099$
$d_{3A} \approx 3{,}162$ $d_{3B} \approx 4{,}123$ } Cluster um A
$d_{4A} \approx 2{,}236$ $d_{4B} \approx 4$
$d_{5C} = 0$ } Cluster um C

Neue Positionen: $B'\left(\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2}\right) = (5{,}5\,;\,0)$

$A'\left(\frac{x_3+x_4}{2}, \frac{y_3+y_4}{2}\right) = (0\,;\,4{,}5)$

$C'(8;8)$

Visuell lässt sich schnell erkennen, dass diese Konfiguration sich nicht mehr ändern wird.

Nun: $WSS_A = 0{,}5^2 + 0{,}5^2 = 0{,}5$
$WSS_B = 0{,}5^2 + 0{,}5^2 = 0{,}5$
$WSS_C = 0$
} $WSS = 1$

Aufgabe 7.2

Das Gene Ontology Enrichment mit den gegebenen differentiell exprimierten Genen ergab einen klaren Trend in funktionellen Gemeinsamkeiten dieser Gene: Die Gene zeigen unerwartet viel Aktivität in der Regulation der Zelldifferentiation. So sind 31 von 32 Genen in diesem sample in der Regulation von biologischen Prozessen beteiligt und davon 16 Gene in der Regulation der Zelldifferentiation. Genauer sind die Funktionen beschrieben als:

1. Negative Regulation der Zelldifferentiation von Innenohr Akustik-Rezeptorzellen (Haarzellen)
2. Negative Regulation der Mechanorezeptor-Differentiation
3. Regulation der Zelldifferentiation der Haarzellen
4. Negative Regulation der Epidermiszellen-Differentiation
5. Negative Regulation der pro-B-Zellen Differentiation

Der p-Wert beschreibt die Wahrscheinlichkeit, die beobachteten Unterschiede (von differenziell exprimierten Genen) zu erhalten, ohne dass ein tatsächlicher Unterschied vorliegt. Verglichen mit dem Signifikanzniveau stellt er die Grenz dar, ab welcher eine Abweichung nicht mehr nur zufällig begründet werden kann.
Die False Discovery Rate FDR kommt zur Anwendung, wenn es aufgrund mehrerer Messungen zu immer mehr Falsch-Positiven Ergebnissen kommt, da aufgrund der größeren Stichprobengröße hierfür die Wahrscheinlichkeit steigt. Sie stellt also eine Multiple Testing Korrektur zum p-value dar, welche die Wahrscheinlichkeit für Falsch-Positive Ergebnisse stark vermindert und die Daten somit aussagekräftiger macht.

Aufgabe 7.3 Gene-set Enrichment (20 %)

Sie haben eine Liste von 50 Genen als differentiell expressiert identifiziert. Von diesen haben 40 die biologische Funktion *Apoptose*. Der Gesamte Datensatz besteht aus 500 Genen von denen 100 die biologische Funktion *Apoptose* haben.

Die Teststatistik lautet: $X^2 = \frac{(O-E)^2}{E}$

Die Nullhypothese H_0 lautet, dass signifikant viele Apoptosegene differentiell exprimiert sind

$O(d.e., A) = 40$; $E(d.e., A) = \frac{50}{500} \cdot \frac{100}{500} \cdot 500 = \frac{5000}{500} = 10$
$O(\neg d.e., A) = 60$; $E(\neg d.e., A) = 90$
$O(d.e., \neg A) = 10$; $E(d.e., \neg A) = 40$
$O(\neg d.e., \neg A) = 390$; $E(\neg d.e., \neg A) = 360$

Damit: $X^2 = \frac{(40-10)^2}{10} + \frac{(60-90)^2}{90} + \frac{(10-40)^2}{40} + \frac{(390-360)^2}{360} = 125$. Nun ist $df = (2-1)(2-1) = 1$ und mit $\alpha = 0{,}05$ ist

$\chi^2(0{,}05;\,1) = 3{,}8415 \ll X^2$

Damit wird H_0 angenommen, es sind signifikant mehr Apoptosegene exprimiert

Aufgabe 7.4

b) Das dataset "palmerpenguins" enthält Daten über 344 Pinguine dreier verschiedener Spezies - Adelie-, Zügel- (Chinstrap) und Eselspinguin (Gento). Die Daten wurden auf den Inseln Biscoe, Dream und Torgersen erfasst. Gemessen wurden die Länge und Tiefe der Schnäbel, die Länge der Flossen sowie das Geschlecht und das Gewicht der Pinguine. Das Jahr der Messung wurde ebenso notiert.
Eine Ergänzung zu diesen Daten finden sich darüber hinaus im dataset "penguins_raw", für welches Eier von Pinguinen der drei Inseln gesammelt wurden. Hier wurden neben ähnlichen geografischen Daten (ergänzt um Region, die in allen Messungen die Region "Anvers" war) und Speziesnamen auch jedem Ei eine Sample Nummer und den Eltern eine individuelle ID zugewiesen. Außerdem wurde zusätzlich zu den Daten des oben beschriebenen Datensatzes festgehalten, in welchem Entwicklungszustand die Eltern waren, an welchem Datum die Eier gesammelt wurden und wie die Promillewerte der Isotope C-13 und N-15 in der Probe waren.

c) In der Datei uebung8.R wurde eine PCA für den Datensatz "penguins" durchgeführt. Eine Visualisierung mittels eines Biplots und mit Vernachlässigung der Datenreihen mit NAs als Einträgen ist in der Datei "biplot_PCA.pdf" zu finden

d) Für das k-mean Clustering wurde zunächst die optimale Anzahl an Zentren mittels der Elbow-method bestimmt. Die erstellte Grafik (siehe "WSS_plot.pdf") lieferte einen Wert von k=3, was nicht überrascht, da in dem Datensatz 3 Pinguin-Arten betrachtet wurden. Den drei Clustern wurden dabei 132, 123 und 87 Datenpunkte zugeordnet.