Università degli Studi di Milano

Data Science and Economics (LM-91)

Antarctic Penguins

Species Exploration

Shihab Hamati

Nov 3, 2022

Objective

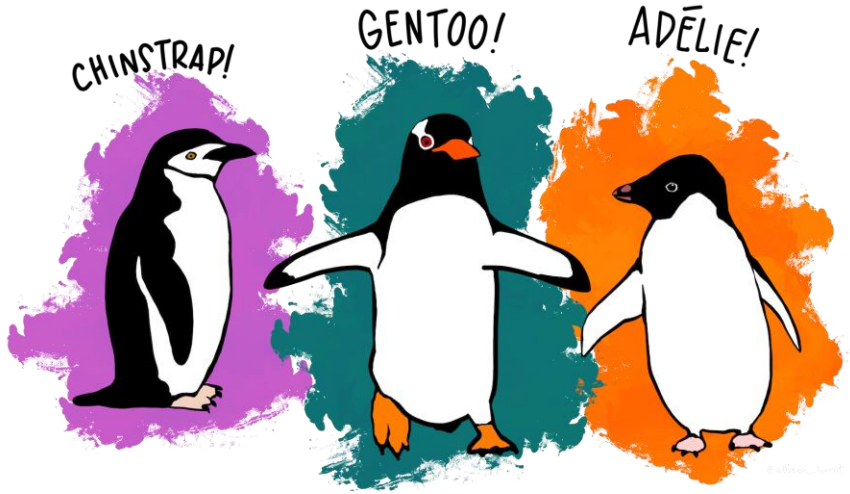
Thousands of new species are found annually

- **Most species on Earth are yet undiscovered!** It is estimated that just under a quarter of the 8.75 million are described (link)
- Scientists **discover around 18,000 new species a year** at a cost of billions USD
- Taxonomists require **expensive and time-consuming resources to make definitive discoveries** of new species
- With **modern unsupervised ML algorithms**, it is possible to identify patterns across the data months prior to the conclusion of a formal genetic analysis

Multiple unsupervised ML techniques are explored to identify groupings of Antarctic penguins

- The analysis explores the **differences and similarities** between observations of the Palmer Archipelago Penguins dataset

- Different clustering techniques aim is to distinguish the penguins based on some traits that are **common to each group but significantly different across groups**



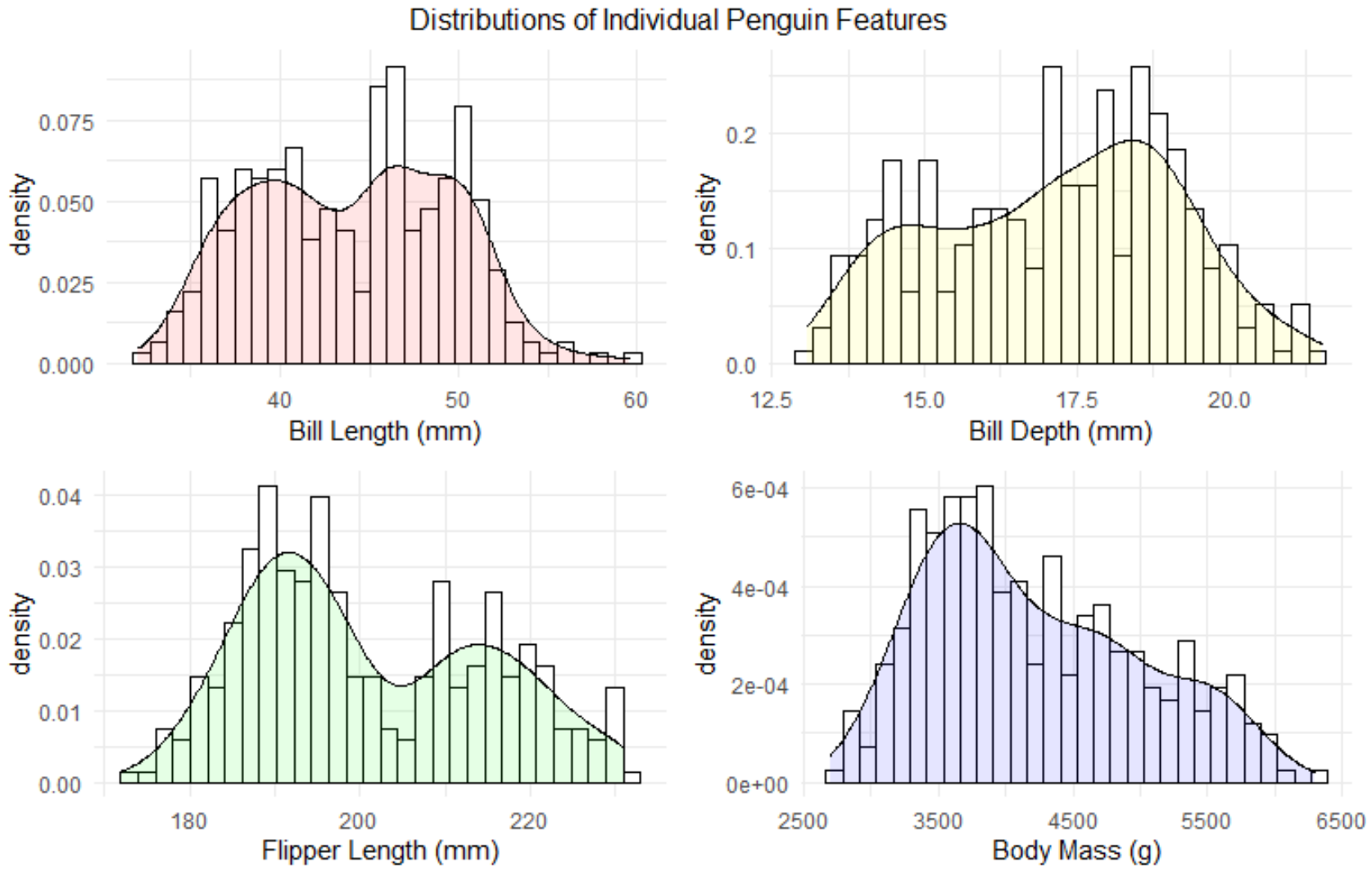
Exploratory Data Analysis

Dataset records the physical measurements, gender, and geography of he studied penguins

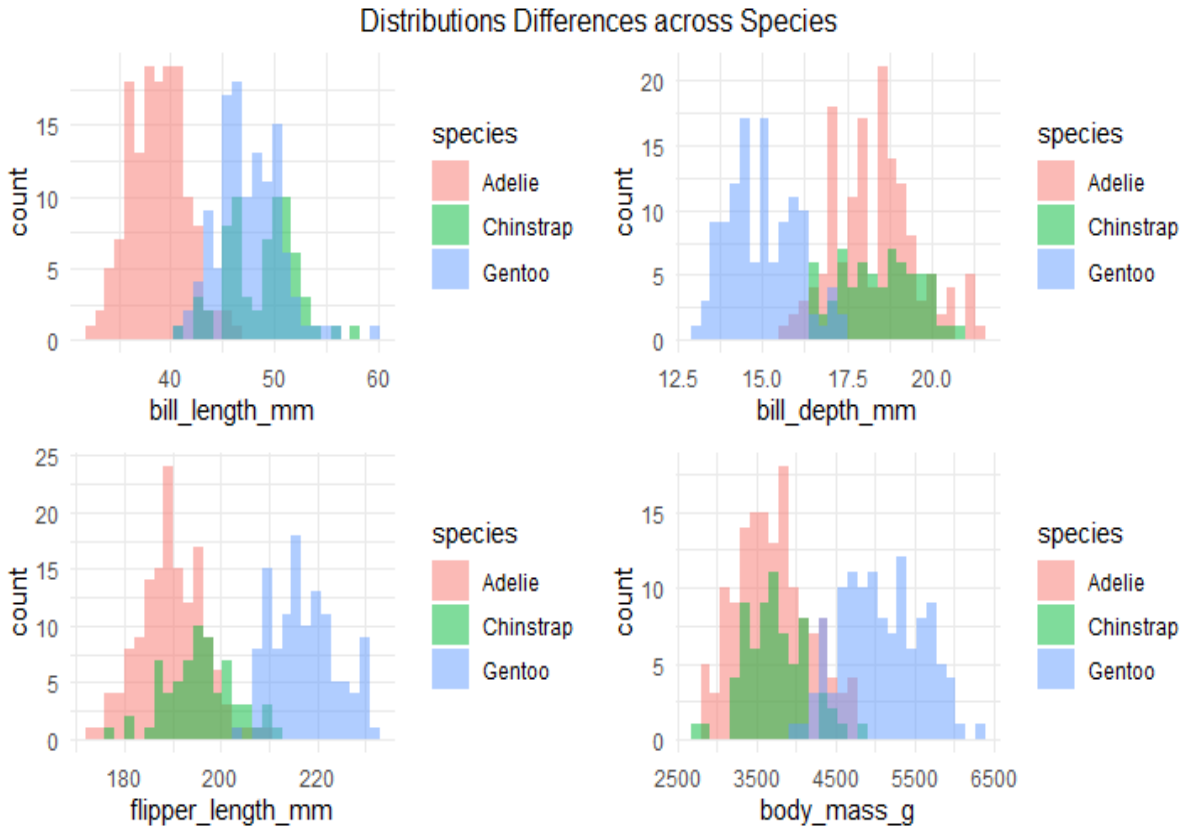
- **Species*:** *multiclass categorical*, describes which of three species a bird belongs to
- **Island:** *multiclass categorical*, the region in which an observation was made
- **Bill Length** (mm): *numerical*, length of peak, from head and towards observer
- **Bill Depth** (mm): *numerical*, dimension of the beak from top to bottom
- **Flipper Length** (mm): *numerical*, the length of the "wing" or "arm"
- **Body Mass** (g): *numerical*
- **Sex**: *binary categorical*, male or female
- **Year**: *numerical*, year of recorded measurement (between 2007-2009)

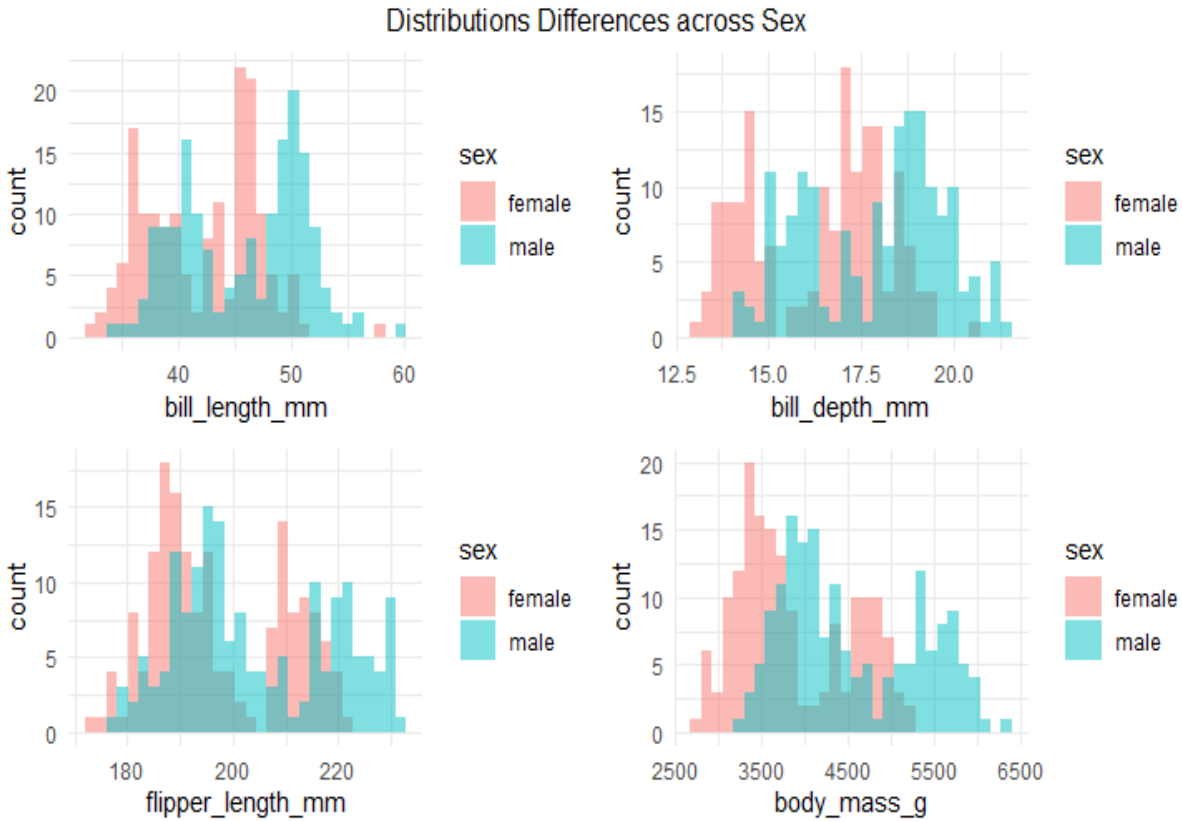
* This column is dropped from all the unsupervised analyses, and used in lieu of subject matter experts only ex post facto to understand the clustering and decomposition results

Distributions of individual physical features



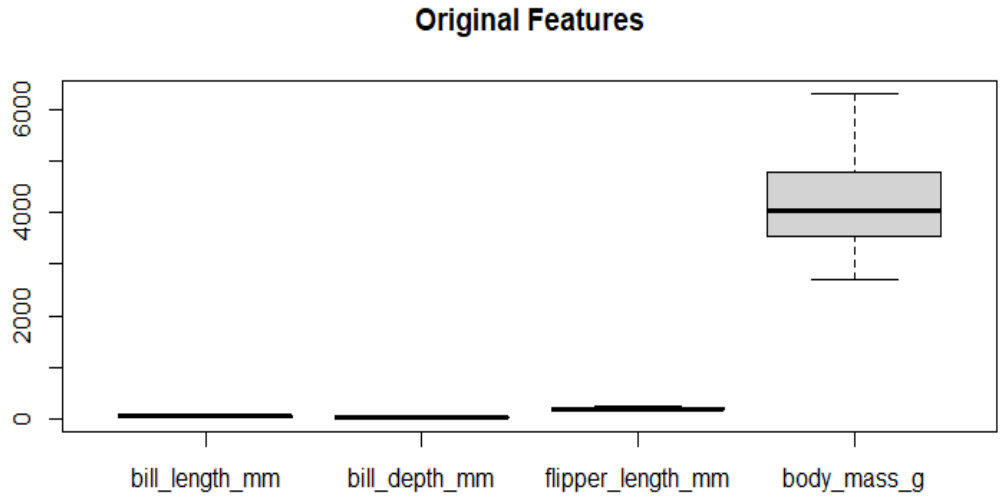
Physical features of grouped by classes

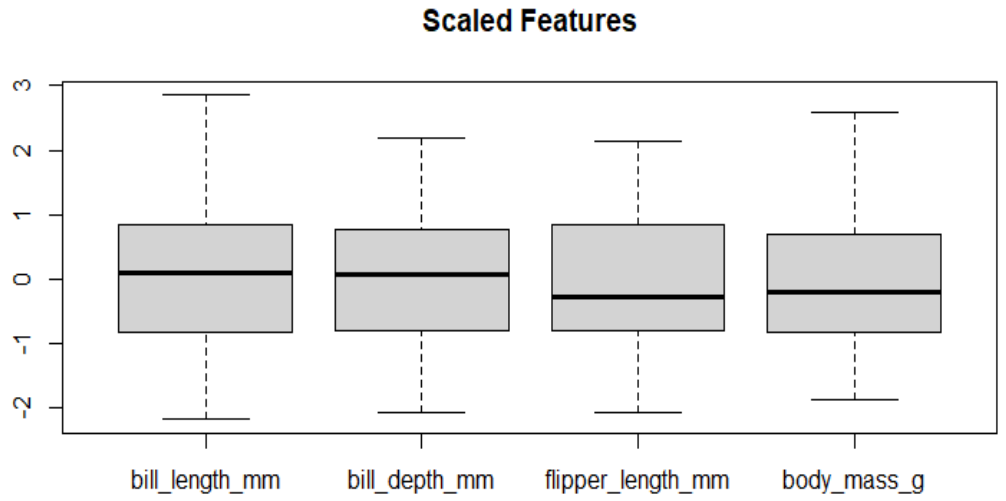




Different physical scales vary widely, so scaling is required for the used unsupervised methods

- Raw **physical measurements differ wildly in scale**, with means ranging from 17.16 to 4207
- Data **must be scaled prior to unsupervised ML** as distances and variances across features are used, and differences in scale will masquerade as false dominant values
- Each feature is **centered by its mean and scaled by its standard deviation**





Correlations of the physical measurements

- The **flipper length and the body mass are highly correlated**:

  This is logical since the flipper length is a good portion of a penguin's height, and consequently the larger the penguin's size the heavier it is expected to weigh

- It appears that **the larger the penguin, the narrower its beak**:

  This reflects the function of the pinguin's beak, which is shaped like a narrow hook to grab and hold on to fish, which is a main source of food

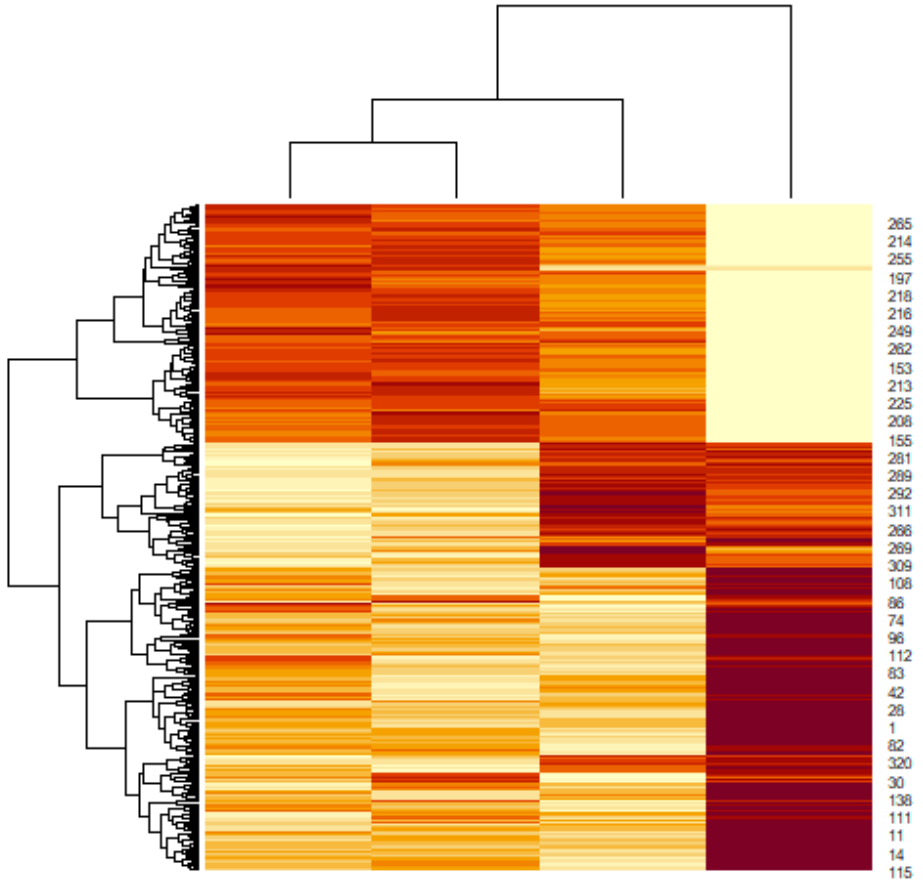
Statistical Models

Clusterability

Hopkins Statistic

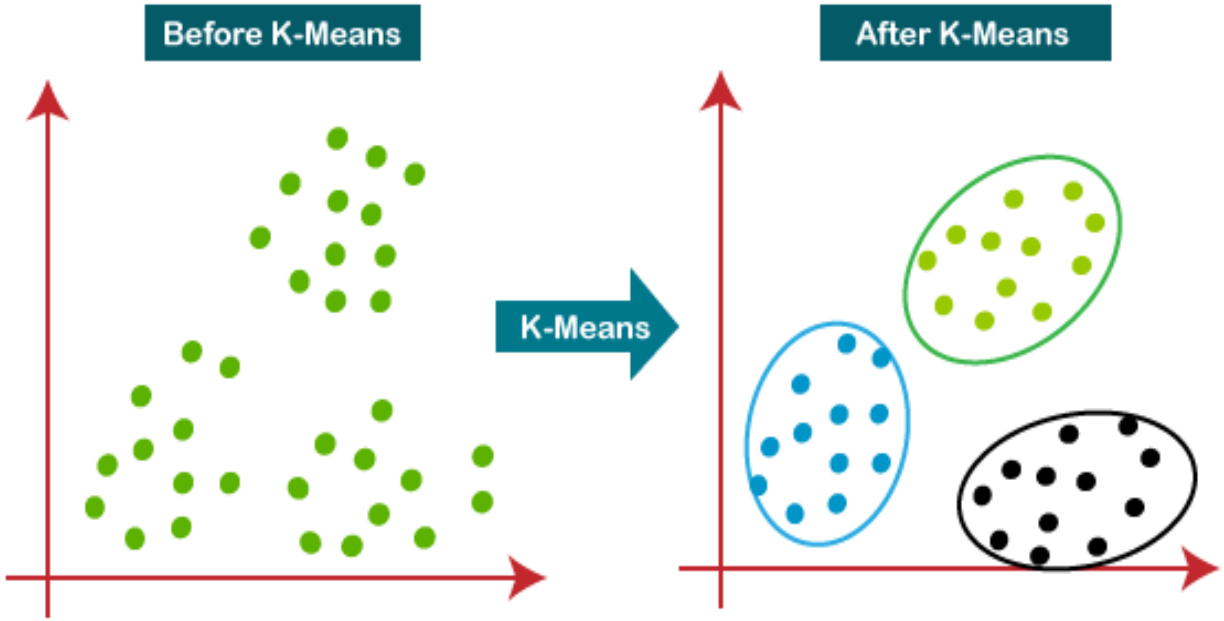
- The Hopkins statistic **measures the clustering tendency** of the dataset
- Values between 0.7 and 1 **indicate a tendency to cluster**
- The dataset has a Hopkins statistic greater than 0.9, thus **exhibiting a clustering tendency**

Visual Heatmap



1 K-Means

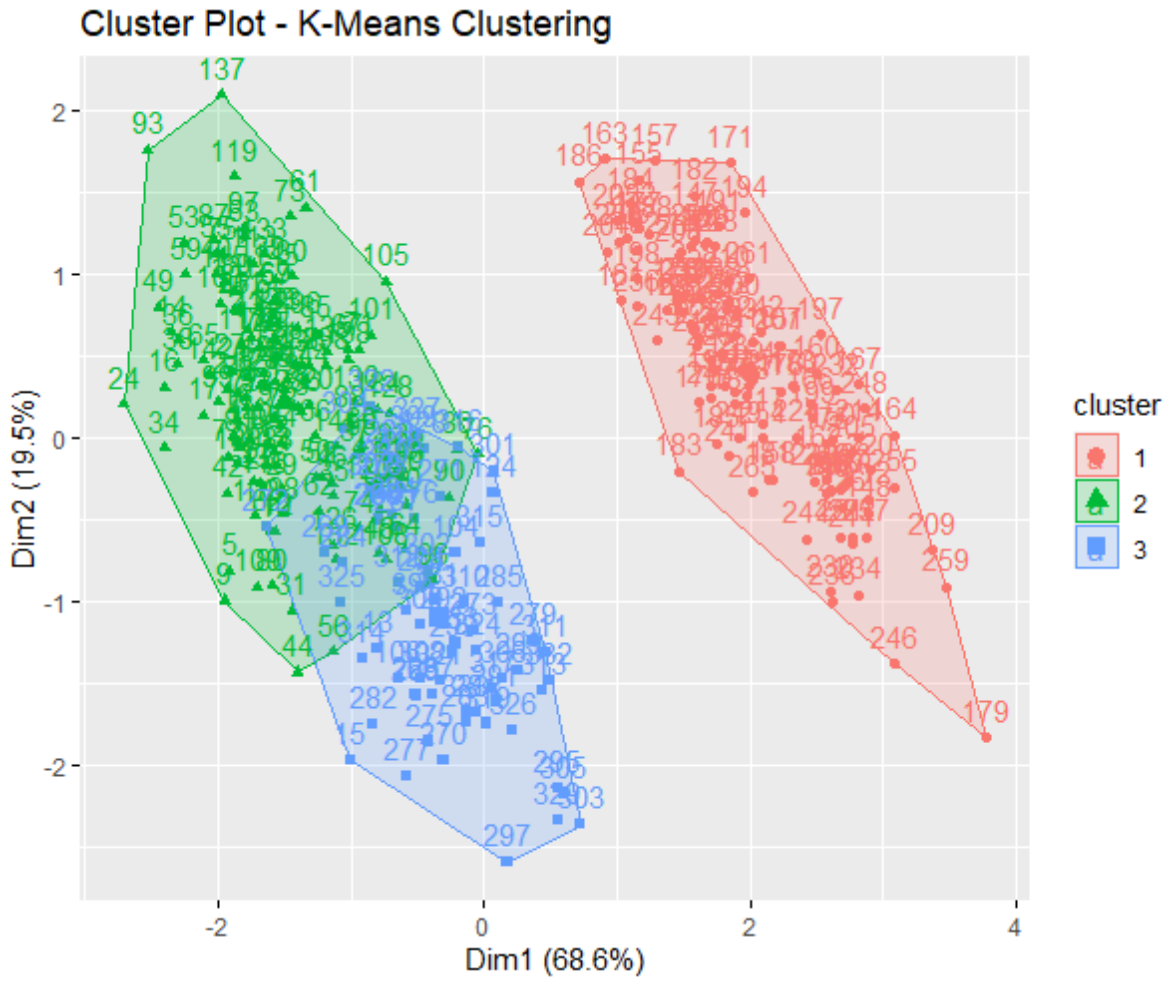
- This method attempts to **partition the observations into a pre-specified number of clusters**
- It achieves this by **iteratively honing on the best pre-specified number of cluster centers** that minimize within-cluster variation
- **Euclidean distances** between scaled observations were computed in 4D space



1 Choosing the number of clusters

Method 1: Elbow Method on WCSS

- **Different number of clusters** were run and the Within Cluster Sum of Squares is plotted
- It is rarely a clear-cut choice, but from the plot below, **k=3 was chosen as the elbow point**



1 Interpreting the clusters

Method 1: Elbow Method on WCSS

- Unsupervised learning is much **more subjective than supervised learning due to the absence of a target label**, which is why **close collaboration** between data scientists and field experts is paramount to interpret or **understand what the uncovered patterns are**

- In this case, the clustering **result is compared to the hidden species features**

- It appears that the K-means algorithm on the scaled numeric dataset **uncovered that there are three distinct penguin species** in the regions explored by the scientists

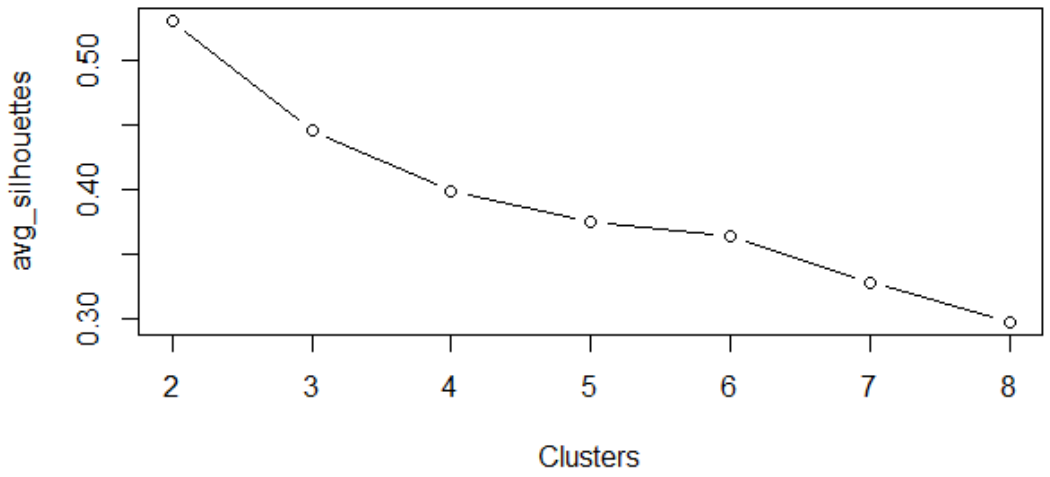
	Adelie	Chinstrap	Gentoo
1	0	0	119
2	139	5	0
3	7	63	0

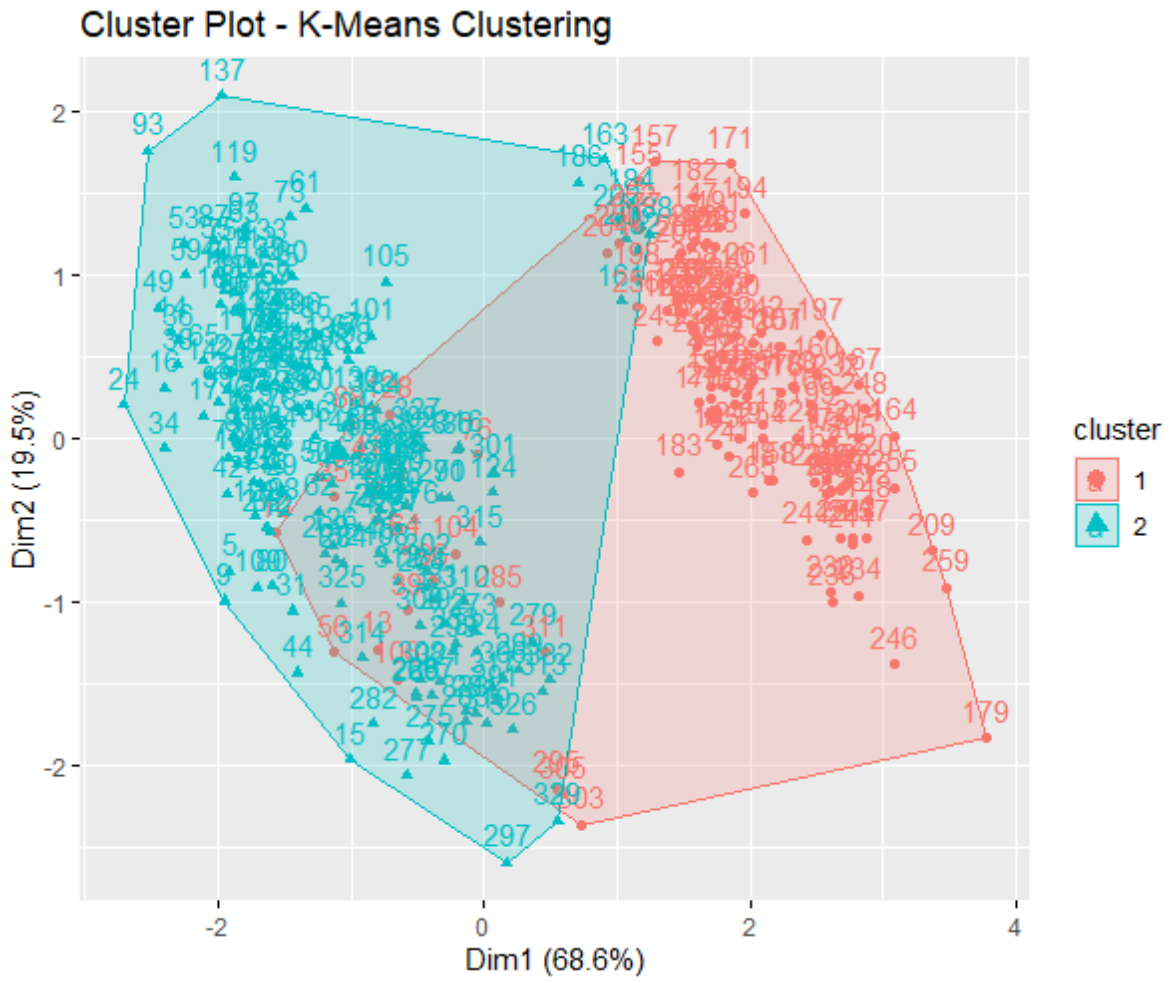
1 Choosing the number of clusters

Method 2: Average Silhouettes

- For different cluster numbers, the silhouette score is computed for each record, **measuring its similarity to its own cluster versus other clusters**

- The average silhouette plot suggests **k=2** clusters (highest average silhouette score)





1 Interpreting the clusters

Method 2: Average Silhouettes

- In this case also, the clustering **result is compared to the hidden species features**
- It appears with two clusters, the algorithm is **distinguishing Gentoo vs non-Gentoo species**
- Most Gentoo live in Biscoe island:

  The two features are highly correlated, which makes sense since members of the same species form social groups that require geographical proximity

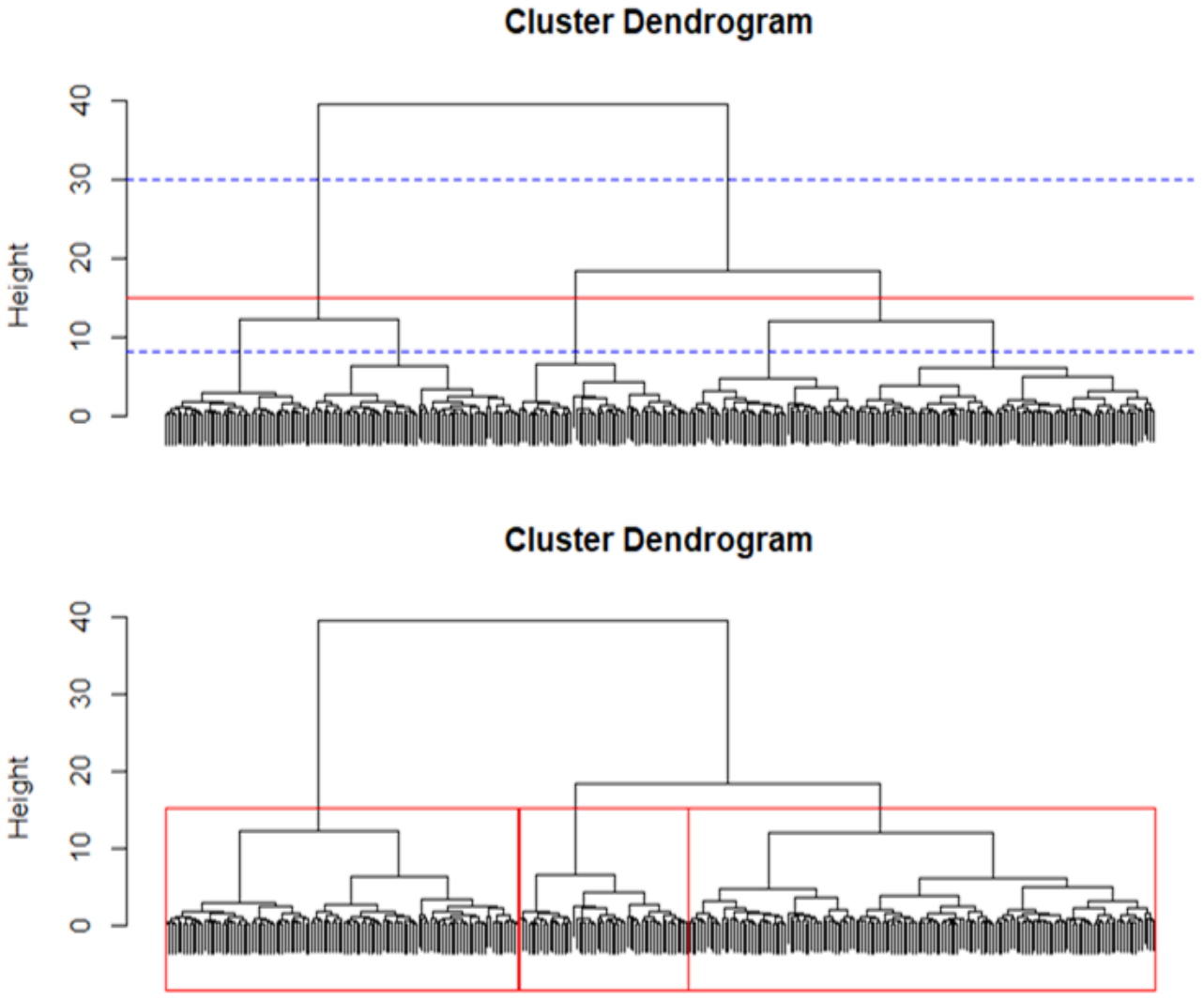
	female	male
1	50	80
2	115	88

	Adelie	Chinstrap	Gentoo
1	14	5	111
2	132	63	8

	Biscoe	Dream	Torgersen
1	115	10	5
2	48	113	42

2 Hierarchical Clustering

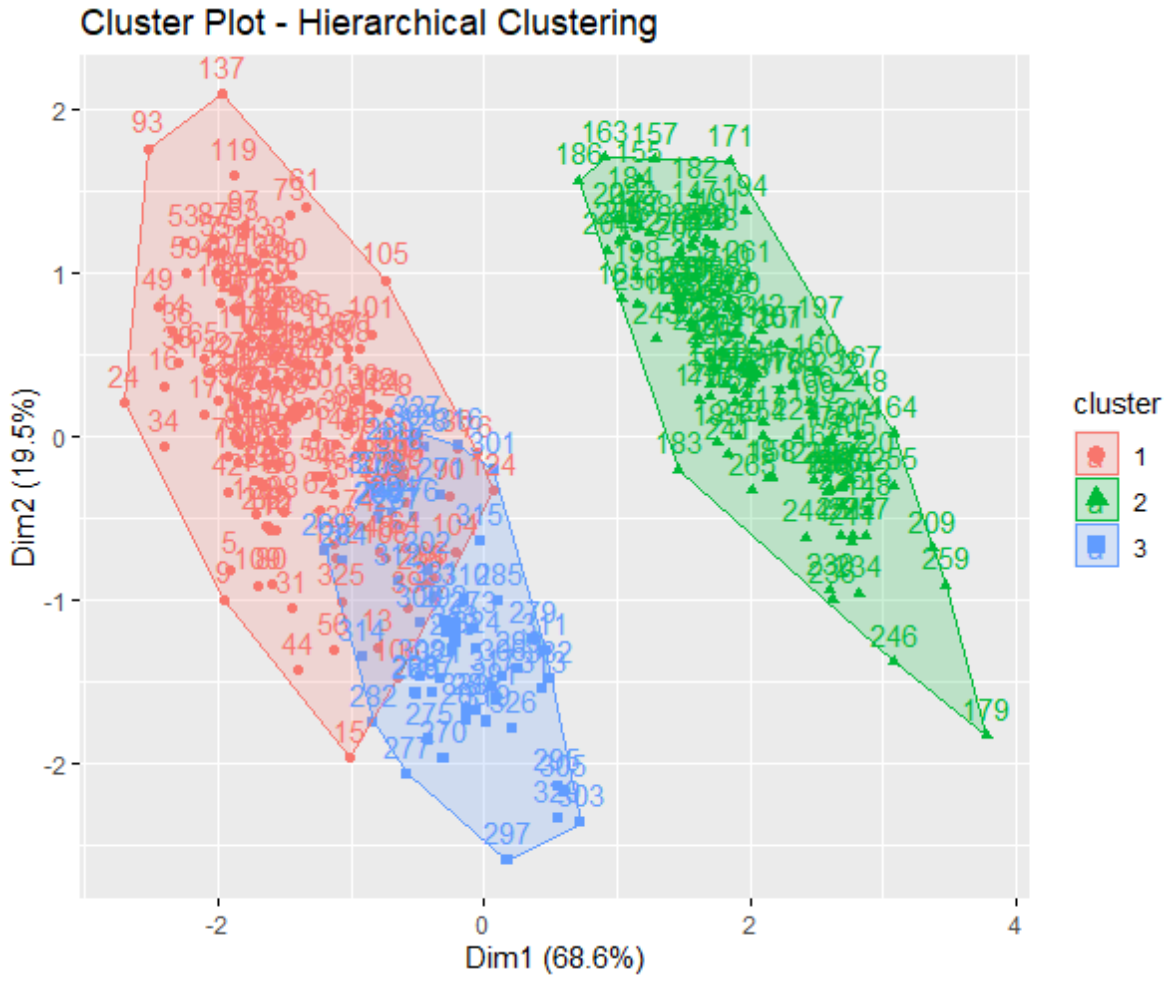
- An alternative to K-means, with the advantage that it **does not require a prior decision** on the number of clusters
- A dendrogram is built **starting from the leaves and combining** clusters all the way up to the root
- The dendrogram displays all cluster sizes, and the **analyst can visually choose an appropriate level** to draw a horizontal line and count how many clusters there are at that level



2 Hierarchical Clustering

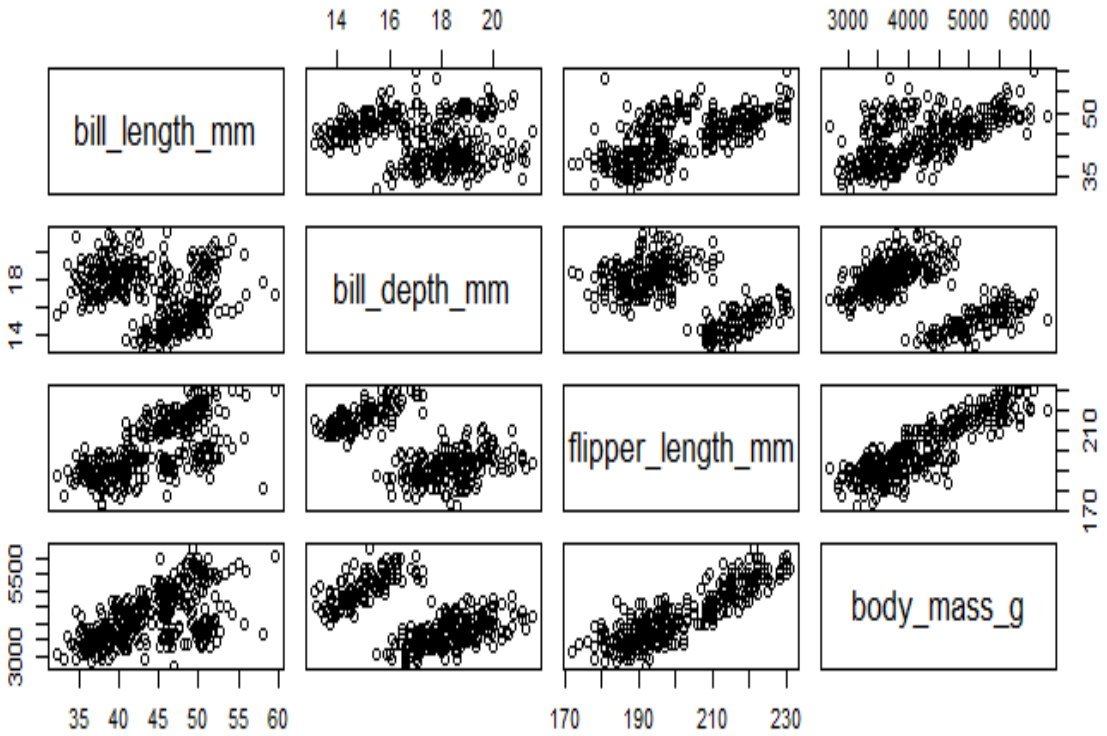
Also as before, hierarchical clustering (k=3) **appears to have identified the three difference species** of penguins

	Adelie	Chinstrap	Gentoo
1	146	11	0
2	0	0	119
3	0	57	0

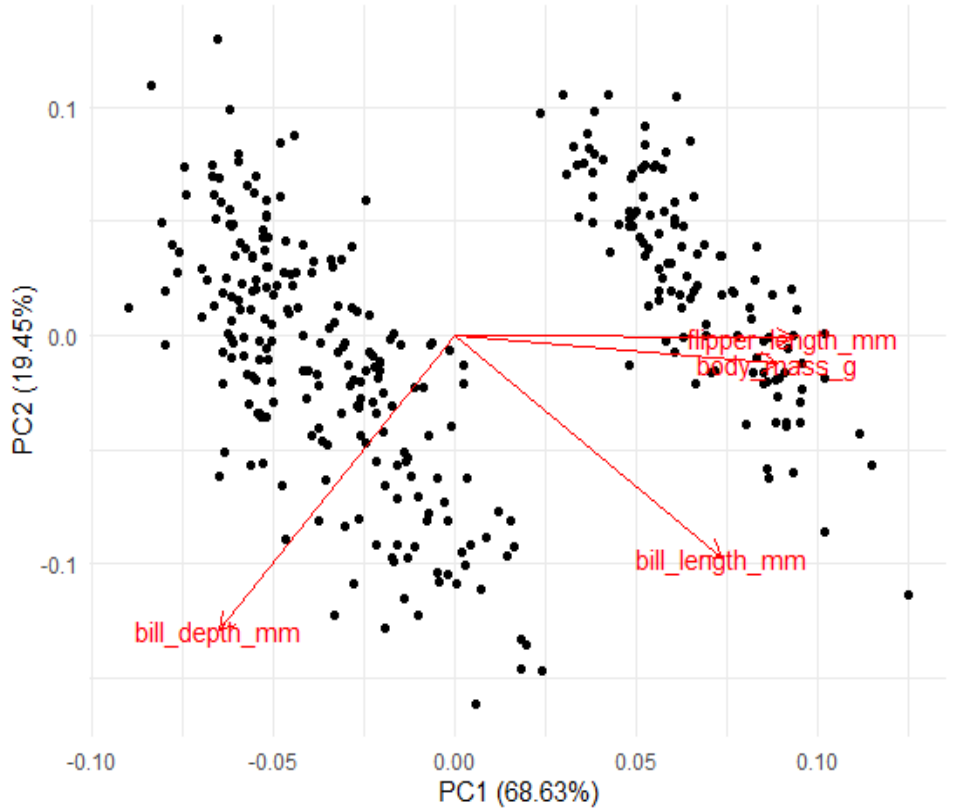


3 Principal Component Analysis (PCA)

There is not one dominant physical features to describe variations across penguins sufficiently

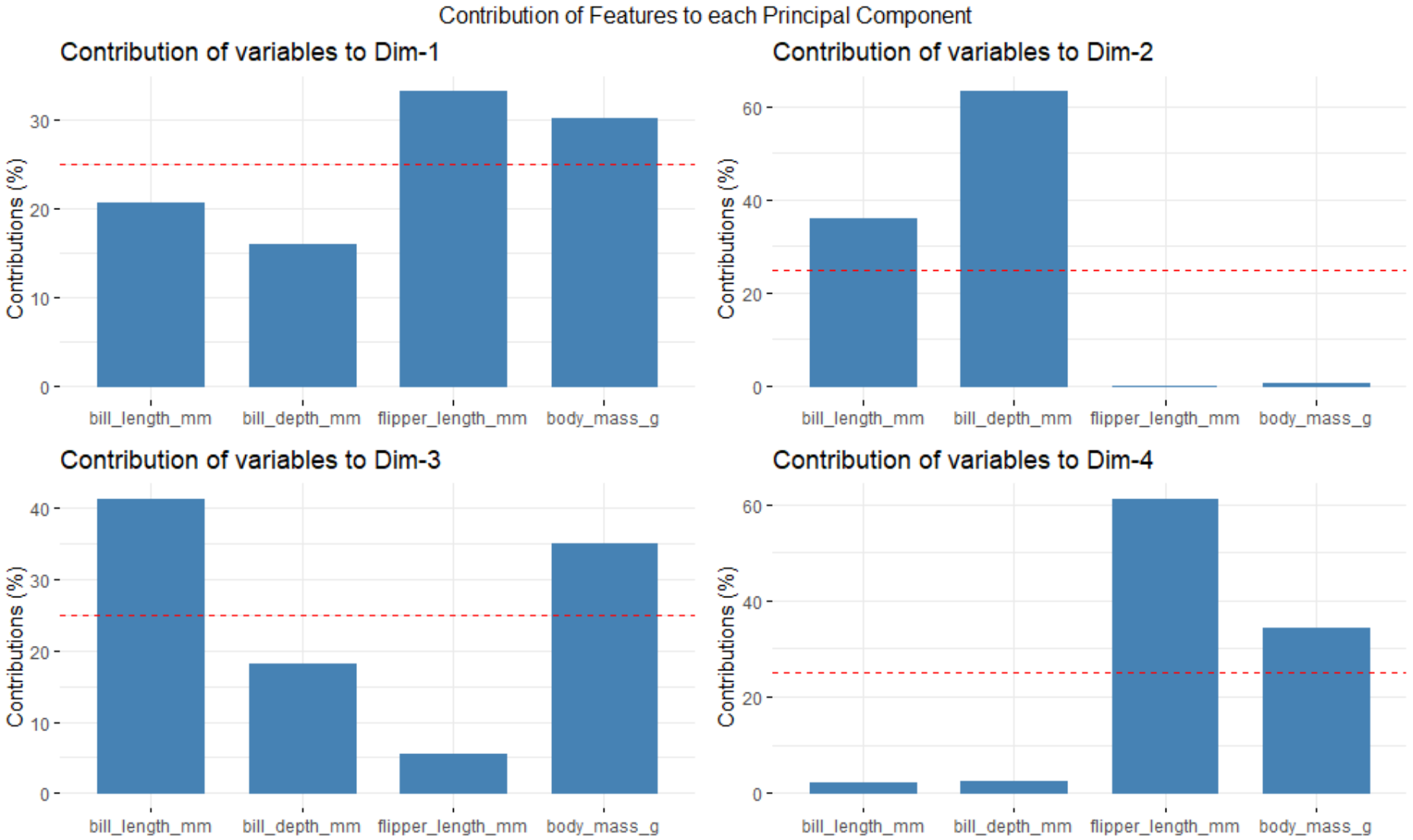


PCA is an unsupervised ML tool that can produce a lower dimensional representation* of the dataset



* When the strongest PCs are kept while the weaker ones dropped (ordered PCs explain more variability than the original ordered features (in order of variance explained)

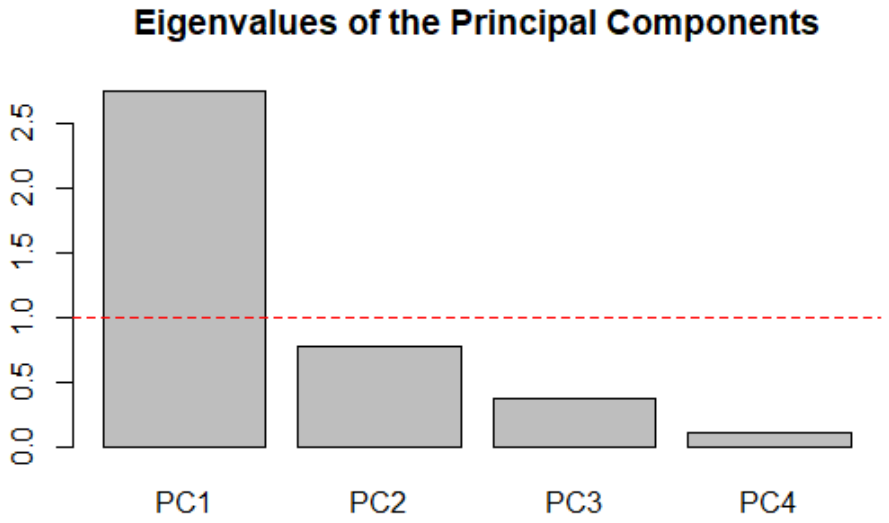
3 Principal Component Analysis (PCA)



3 Clustering after PCA decomposition

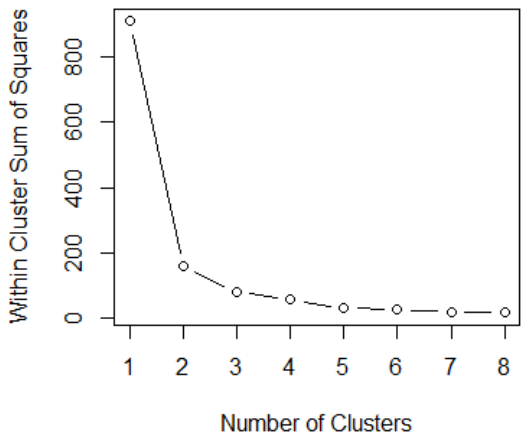
Kaiser Criterion

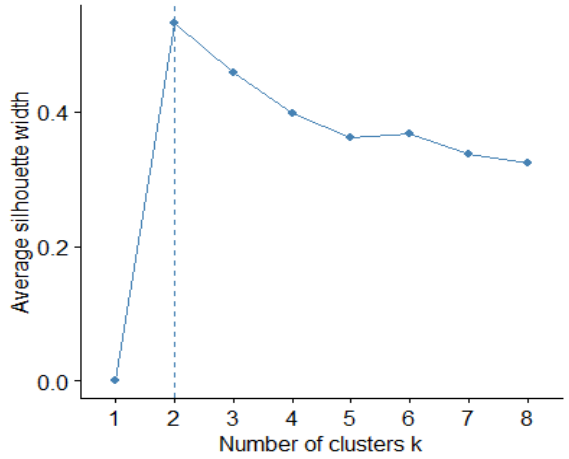
- One way **to determine how many PCs to keep** is to apply the Kaiser criterion
- It suggests keeping only PCs whose **eigenvalues is greater than 1**



Cluster After PCA

- Clustering after decomposition is a popular approach, especially for a **large set of features**
- The optimal number of clusters based on PC1 appears to be **k=2**, based on both the elbow WCSS and silhouette methods





3 PCA achieves comparable clustering to the full data set performance with one quarter of the components

k=2

		Adelie	Chinstrap	Gentoo
K-means (full data)	**1**	14	5	111
	2	132	63	8
PCA (only PC1)	**1**	0	7	119
	2	146	61	0

k=3

	Adelie	Chinstrap	Gentoo
1	0	0	119
2	139	5	0
3	7	63	0

	Adelie	Chinstrap	Gentoo
1	37	59	1
2	109	9	0
3	0	0	118

Key Findings

Key Results

- Clustering and PCA are powerful tools in taxonomy: **scientists can use these techniques** prior to any expensive genetic testing to guide them in the right direction for their **species classification task**
- Methods **differ on the optimal number of clusters** (needs subject matter expertise)
- For k=3, both k-means and hierarchical clustering are **able to split the observed penguins into meaningful groupings, according to species**
- At k=2, the clustering methods split the observations **into larger penguins vs smaller penguins**
- PCA is **able to replicate the clustering performance** at both k=2 and k=3 to a satisfactory level **with only one component**, i.e., 75% less columns