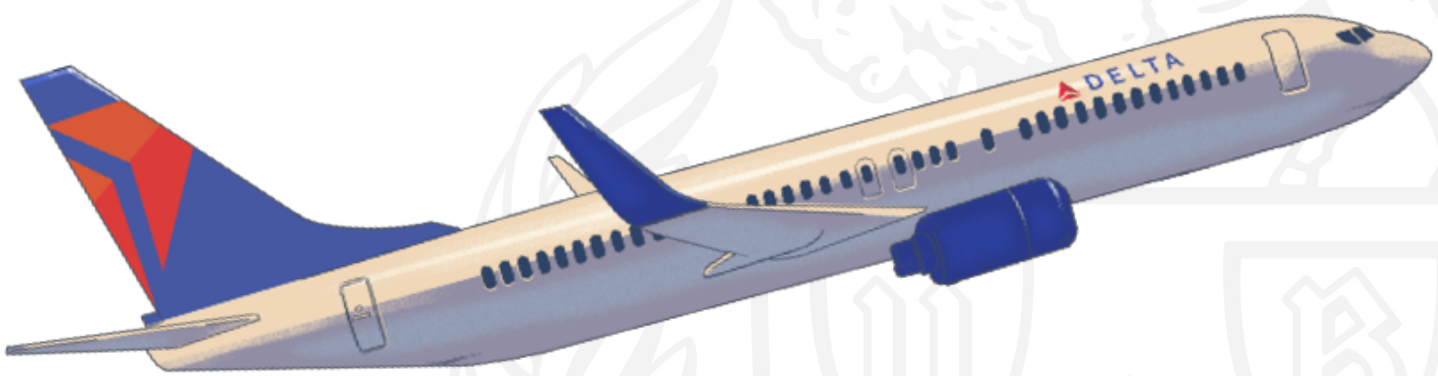
PCA AND K-MEANS CLUSTERING OF HIGH DIMENSIONAL AIRCRAFT DATA

ANISH SHAH

University at Buffalo The State University of New York

Agenda

- General Idea
- About the Dataset
- Analysis
- Principal Components Analysis
- Clustering
- Findings



General Idea

As the data has both quantitative and qualitative information and is relatively complex we look to address these questions:

- What can we say about the different aircraft in Delta's fleet, coming at it with 'fresh eyes'?
- Which planes are similar? And Which are dissimilar?

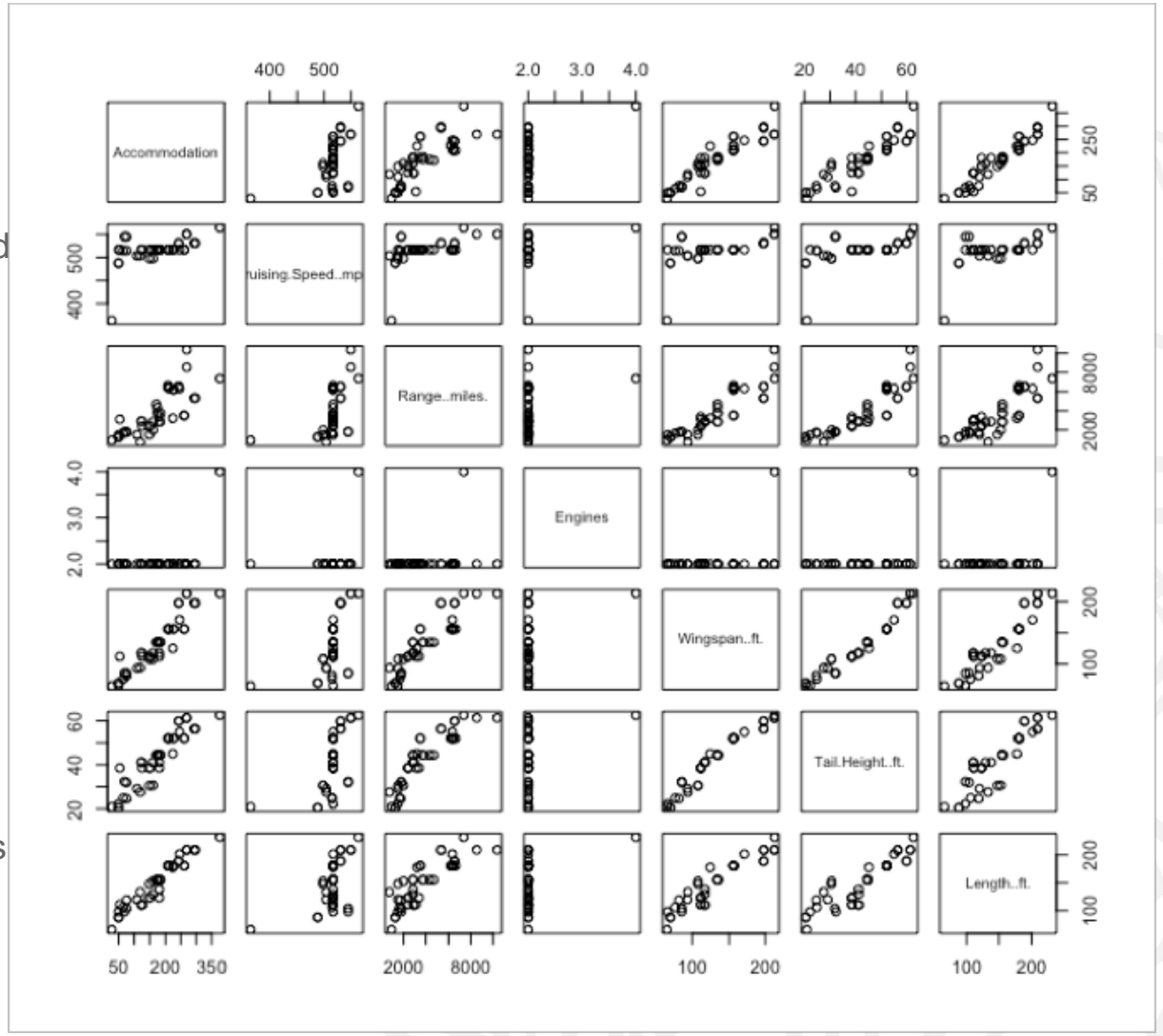
We used PCA and K-Means Clustering for the getting the answers for these questions.

About the Dataset

- The data set comprises 33 variables of 44 aircrafts taken from Delta's website.
- Including both quantitative measures on attributes like cruising speed, accommodation and range in miles, as well as categorical data on, say, whether a particular aircraft has Wi-Fi or video.
- These binary categorical variables were transformed into quantitative variables by assigning them values of either 1 or 0, for yes or no respectively.

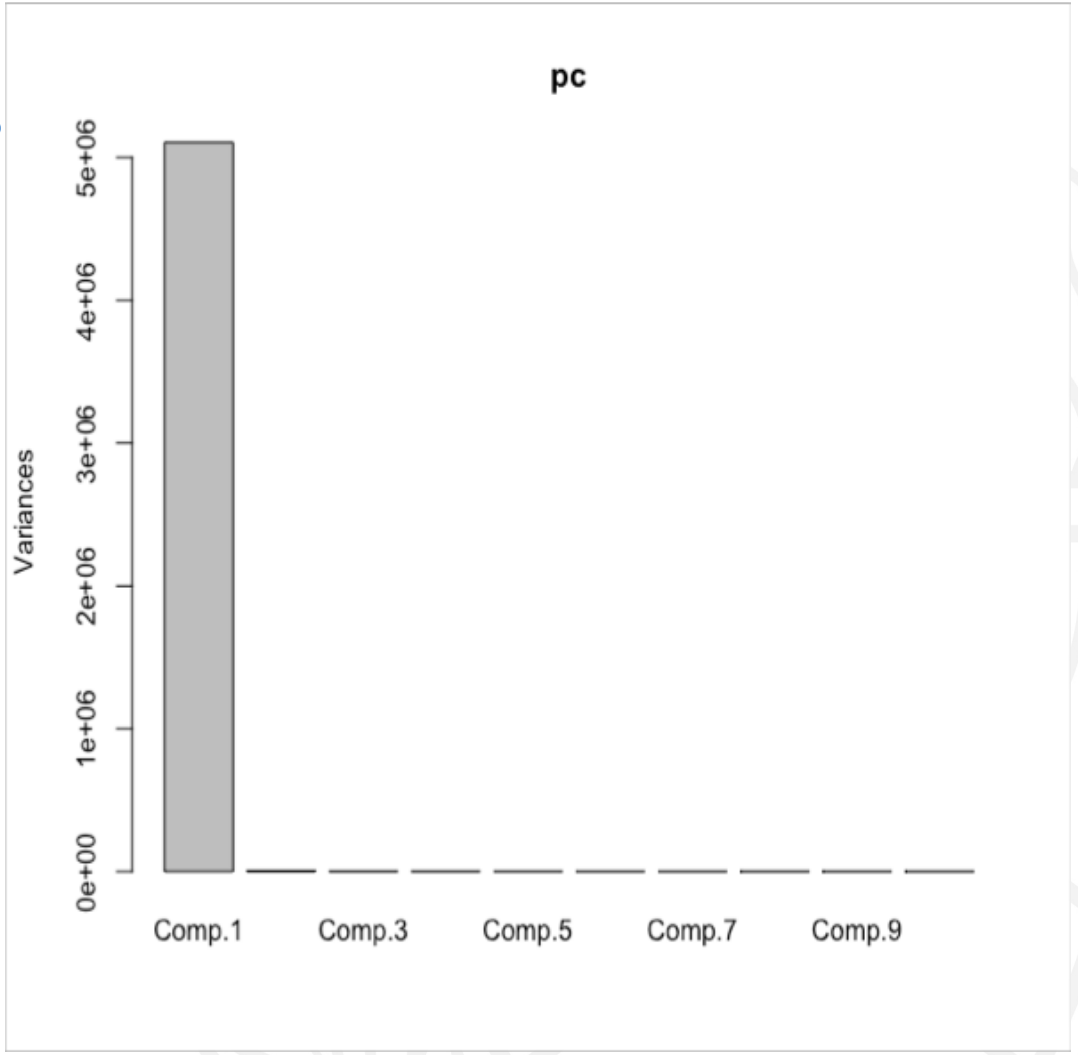
Analysis

- At the start we just took Intermediary quantitative variables related to the aircraft physical characteristics: cruising speed, total accommodation, and other quantities like length and wingspan.
- These variables are about in the middle of the data frame, so we visualized all of them at once using a scatterplot matrix.
- Strong positive correlations between all these variables, as all of them are related to the aircraft's overall size
- The exception here is the variable right in the middle which is the number of engines. There is one lone outlier [Boeing 747-400 (74S)] which has four, while all the other aircraft have two engines



Principal Components Analysis

- So to make visualizing such high dimensional data easier we used dimensionality reduction technique such as PCA.
- We just naively applied principle components to the data to see that the first principal component has a standard deviation of around 2200 and accounts for over 99.8% of the variance in the data.
- Looking at the first column of loadings, we see that the first principle component is just the range in miles.

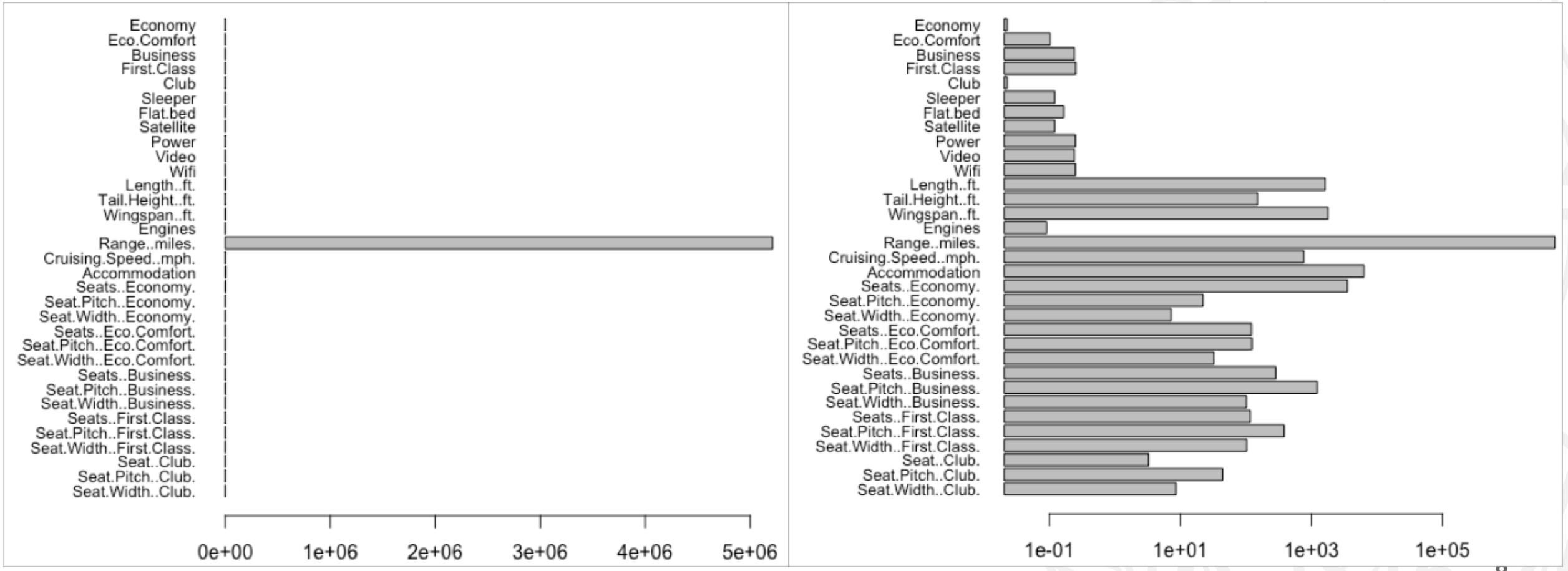


```
> # First component dominates greatly. What are the loadings?
> summary(pc) # 1 component has > 99% variance
Importance of components:
                             Comp.1       Comp.2       Comp.3
Standard deviation     2259.2372556 6.907940e+01 2.871764e+01
Proportion of Variance    0.9987016 9.337038e-04 1.613651e-04
Cumulative Proportion     0.9987016 9.996353e-01 9.997966e-01
```

- This is because the scale of the different variables in the data set is quite variable.

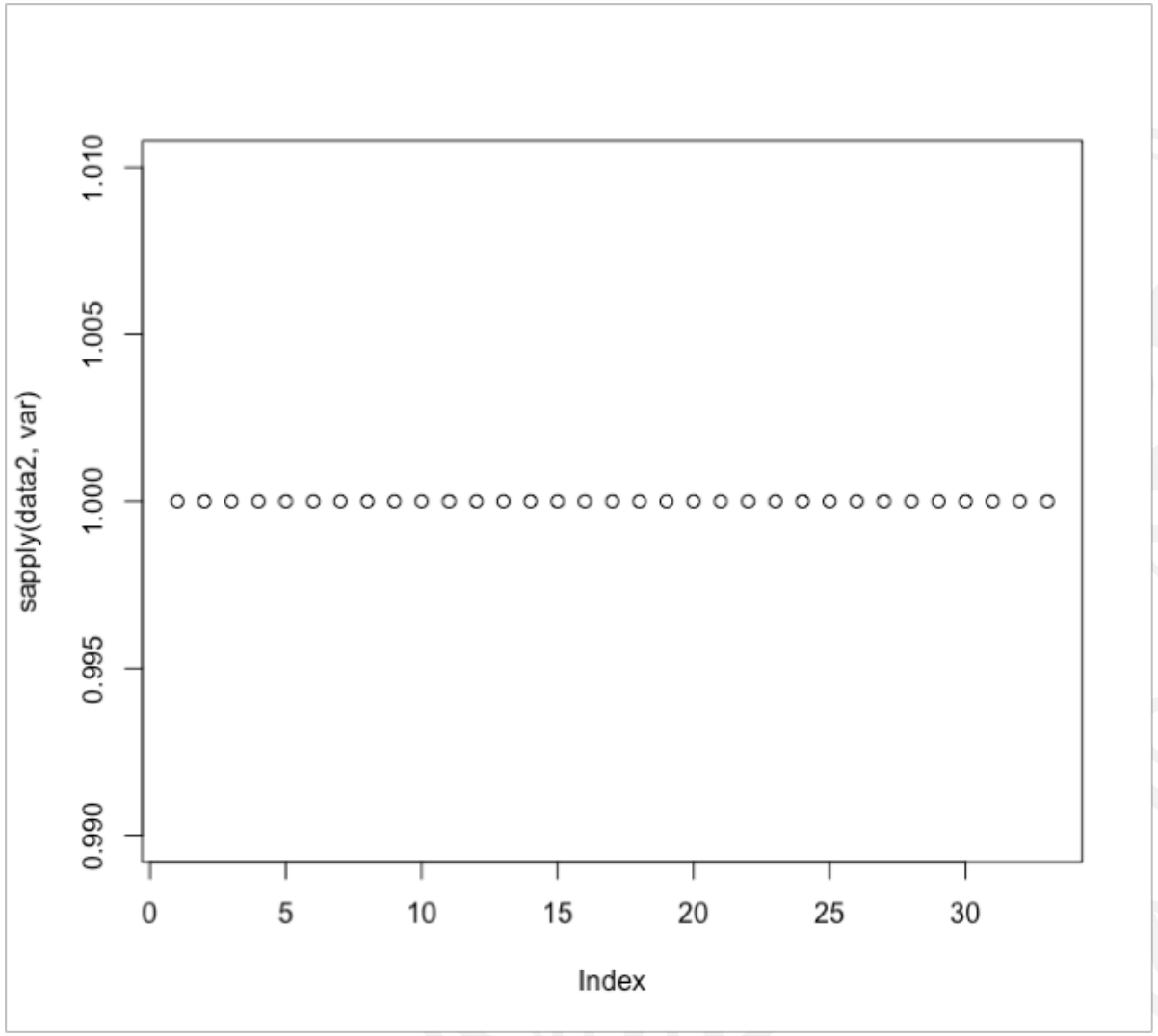
```
> loadings(pc) # Can see all variance is in the range in miles

Loadings:
                      Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
Seat.Width..Club.                                 -0.144 -0.110
Seat.Pitch..Club.                                 -0.327 -0.248
Seat..Club.
Seat.Width..First.Class.               0.250      -0.160
Seat.Pitch..First.Class.               0.515 -0.110 -0.386  0.112
Seats..First.Class.                    0.258 -0.124 -0.307 -0.109
Seat.Width..Business.                 -0.154  0.142 -0.108
Seat.Pitch..Business.                 -0.514  0.446 -0.298  0.154
Seats..Business.                      -0.225  0.187
Seat.Width..Eco.Comfort.                                    0.285
Seat.Pitch..Eco.Comfort.               0.159                0.544
Seats..Eco.Comfort.                                         0.200
Seat.Width..Economy.                                 0.125  0.110
Seat.Pitch..Economy.                                 0.227  0.190
Seats..Economy.               0.597         -0.136  0.345 -0.165
Accommodation                 0.697                -0.104
Cruising.Speed..mph.                   0.463  0.809  0.289 -0.144
Range..miles.          0.999
Engines
Wingspan..ft.                 0.215         0.103 -0.316 -0.357
Tail.Height..ft.                                   -0.100
Length..ft.                   0.275         0.118 -0.318  0.467
Wifi
Video
Power
Satellite
Flat.bed
Sleeper
Club
First.Class
```

We can see this by plotting the variance of the different columns in the data frame (regular scaling on the left, logarithmic on the right):

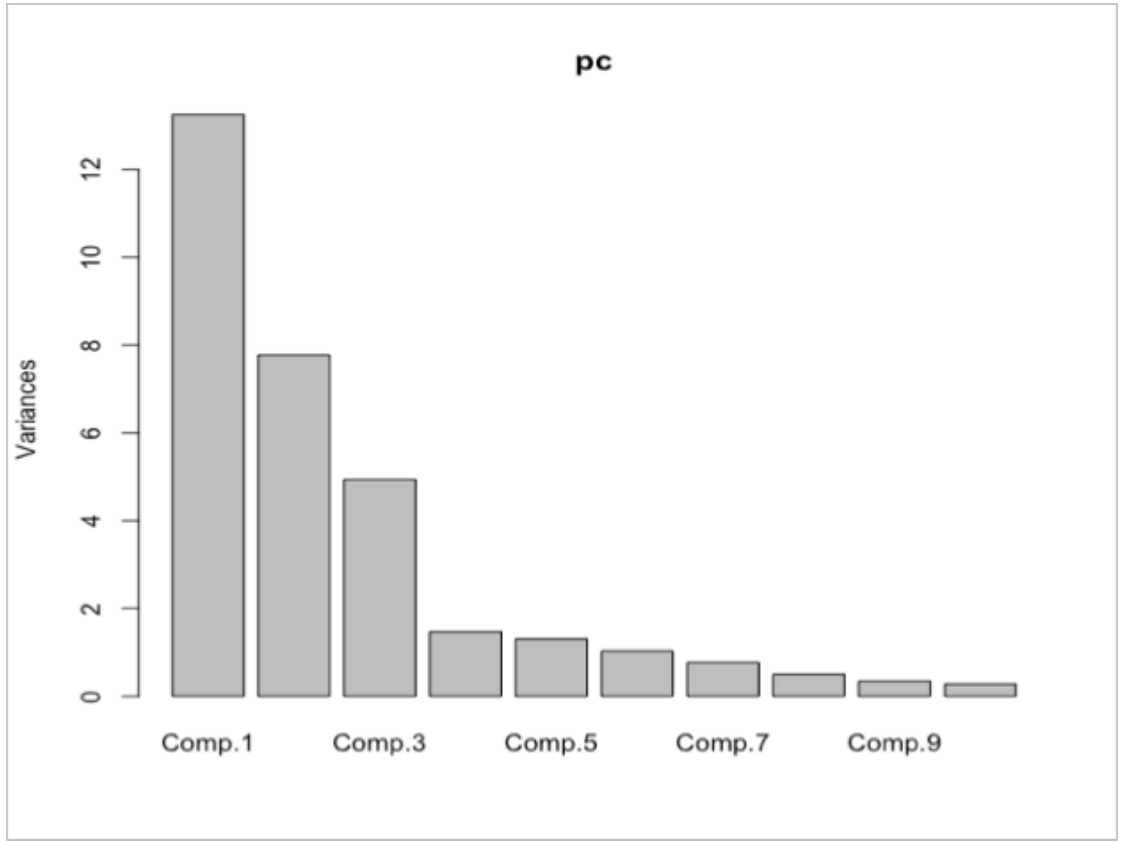


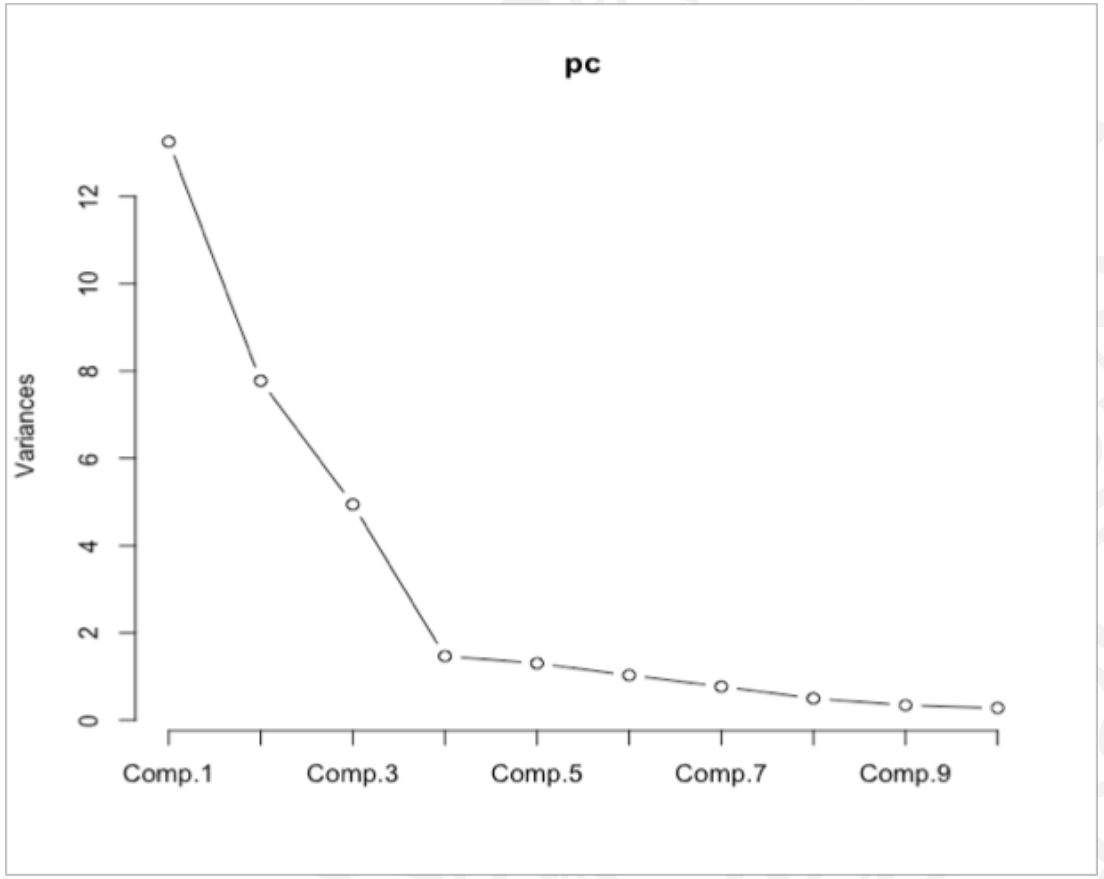
- We correct for this by scaling the data using the scale() function. We can then verify that the variances across the different variables are equal so that when we apply principal components one variable does not dominate.

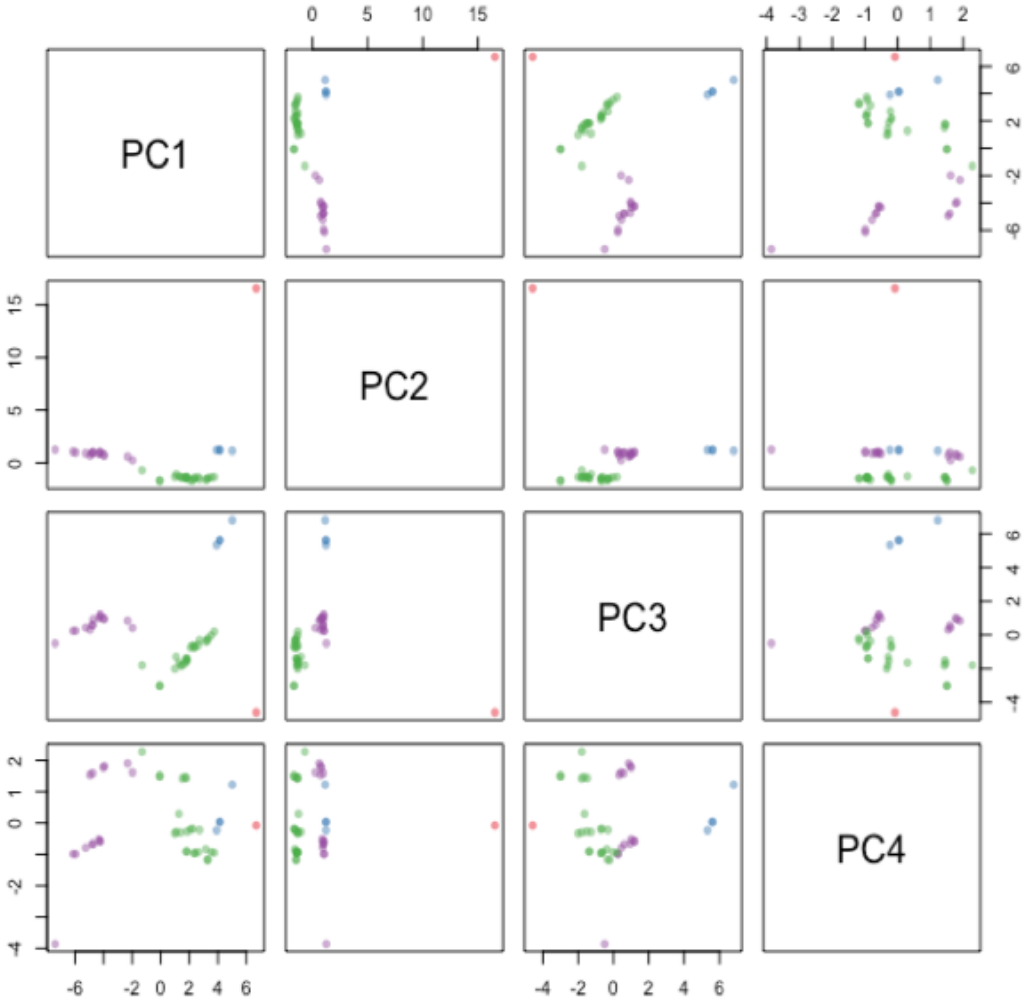
- Now we can apply principal components to the scaled data



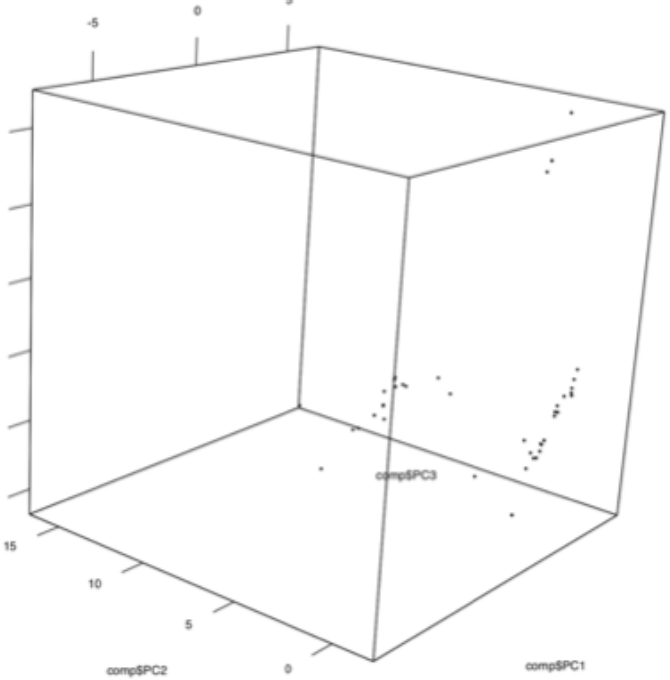
- 4 components is both 'elbow' and explains>85% variance.
- So we we retained the first four principal components.

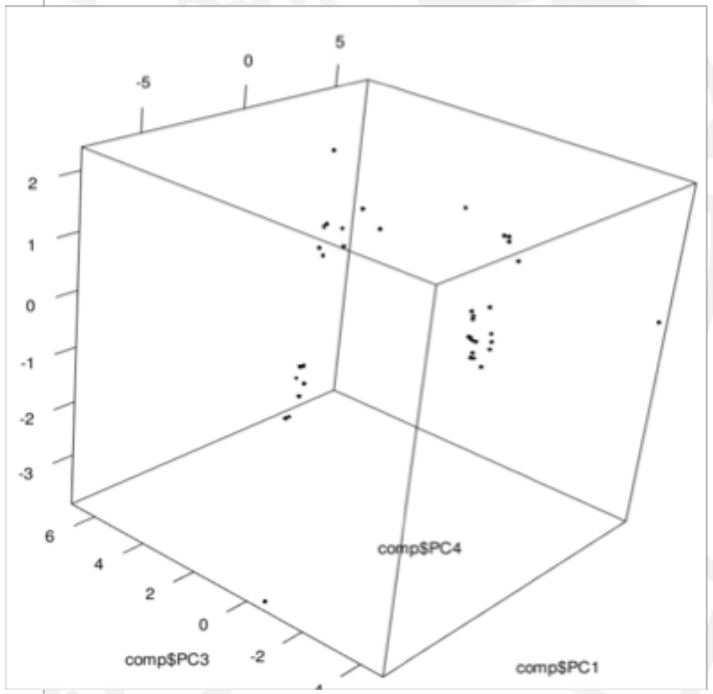






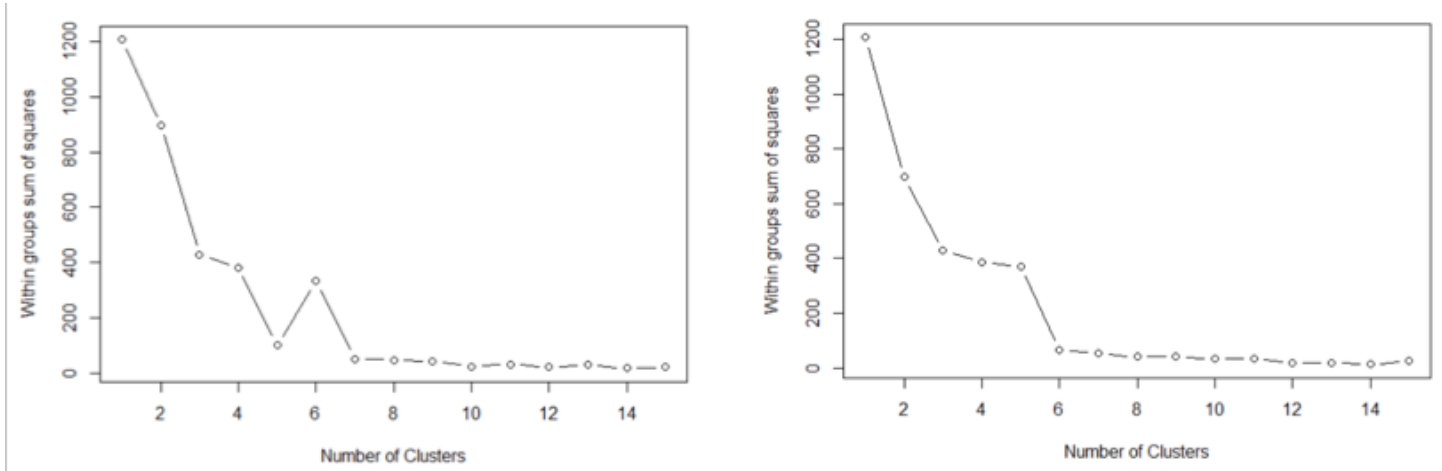
You can better see in the 3D projections that the data are confined mainly to the one plane one the left (components 1-3), with the exception of the outlier, and that there is also bunching in the other dimensions (components 1,3,4 on right).



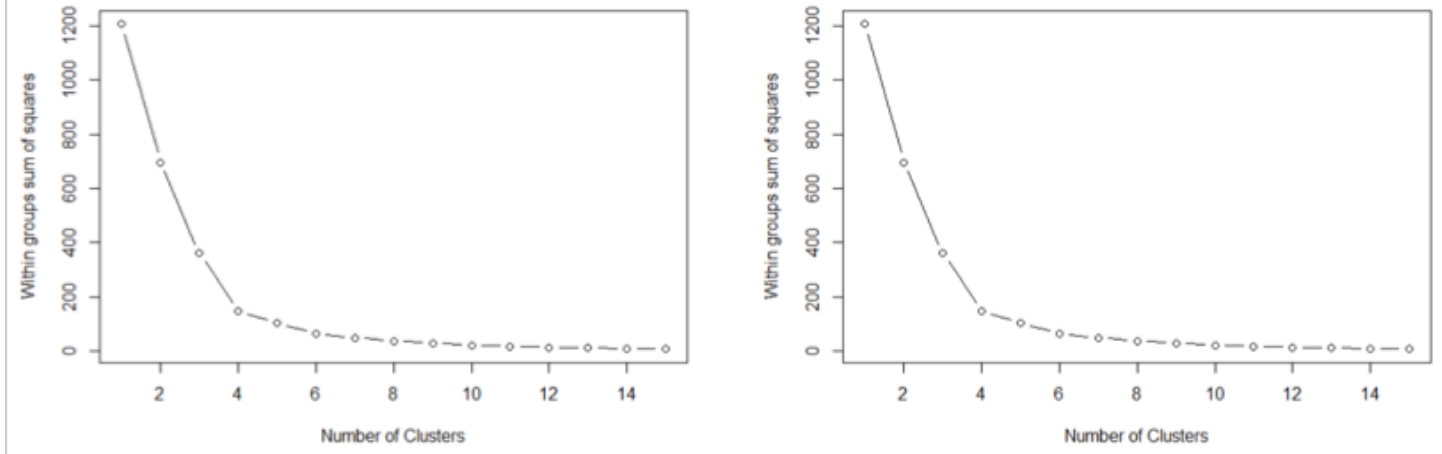


Clustering

- After simplifying the complex data set into a lower dimensional space we can visualize and work with, how do we find patterns in the data, in our case, the aircraft which are most similar.
- For this we used **K-Means Clustering** which is a simple unsupervised machine learning technique.
- We set the nstart parameter and iter.max parameter (otherwise you can get very different results each time you run the algorithm, as below.

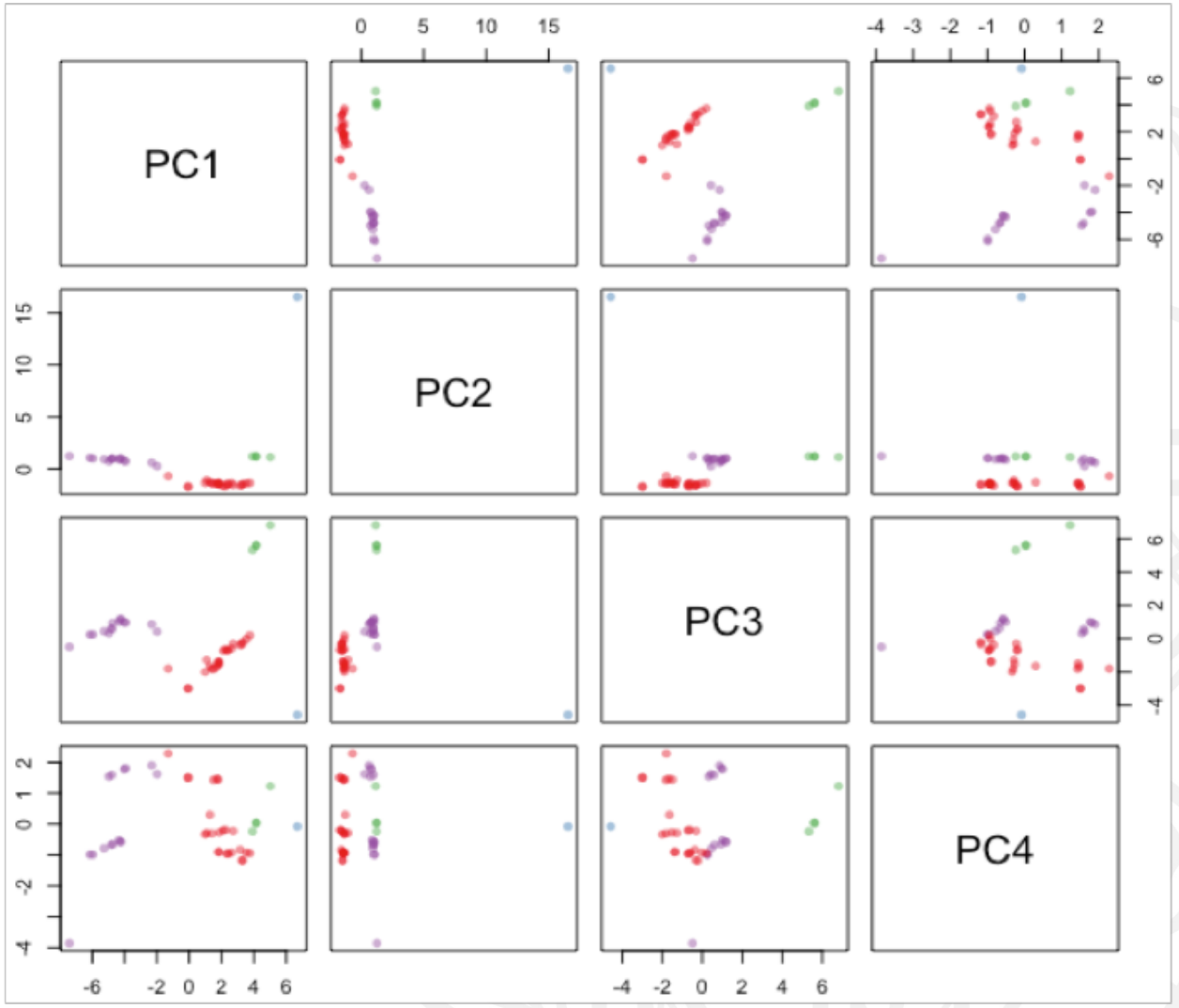


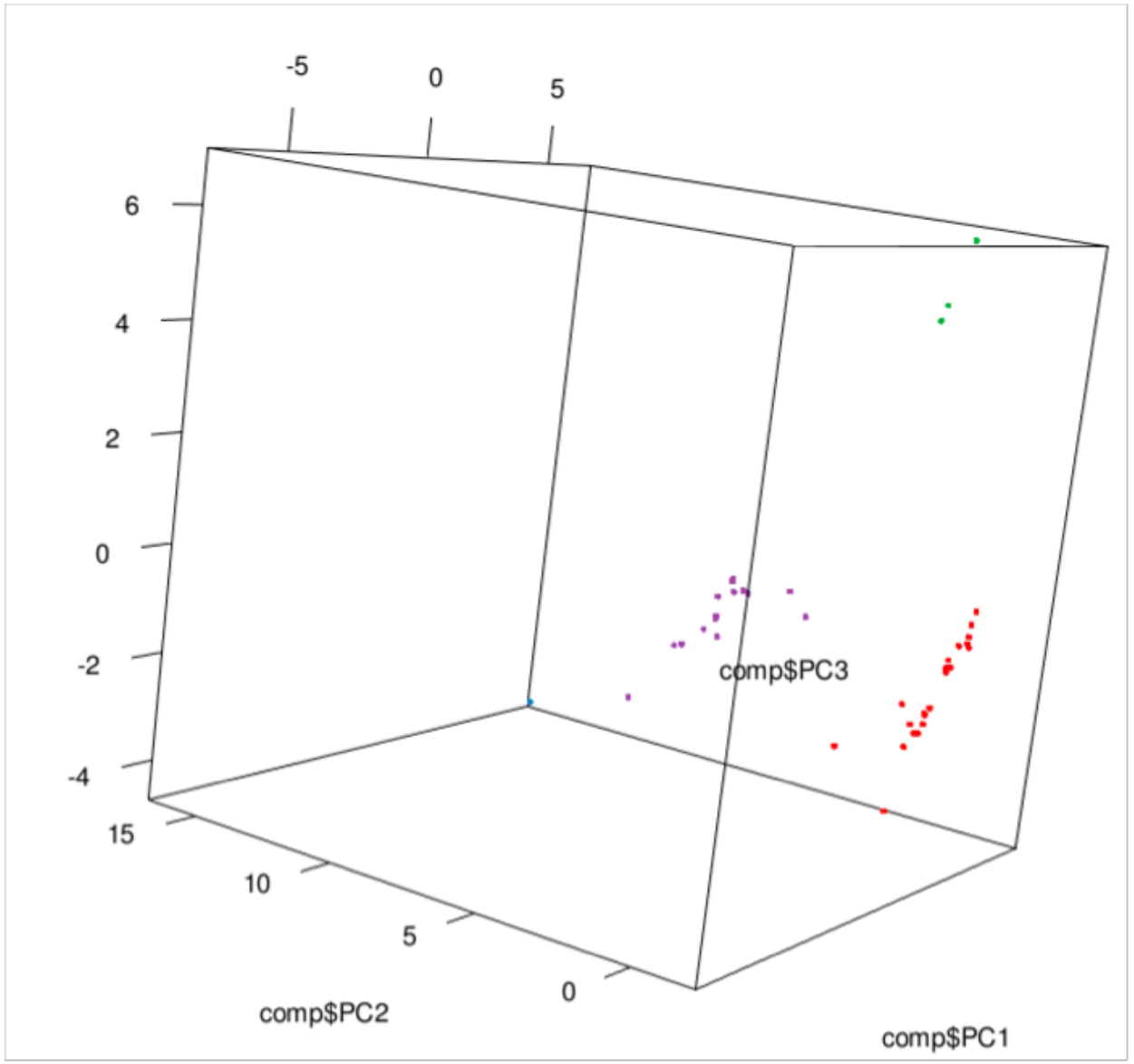
Clustering without the nstart parameter can lead to variable results for each run

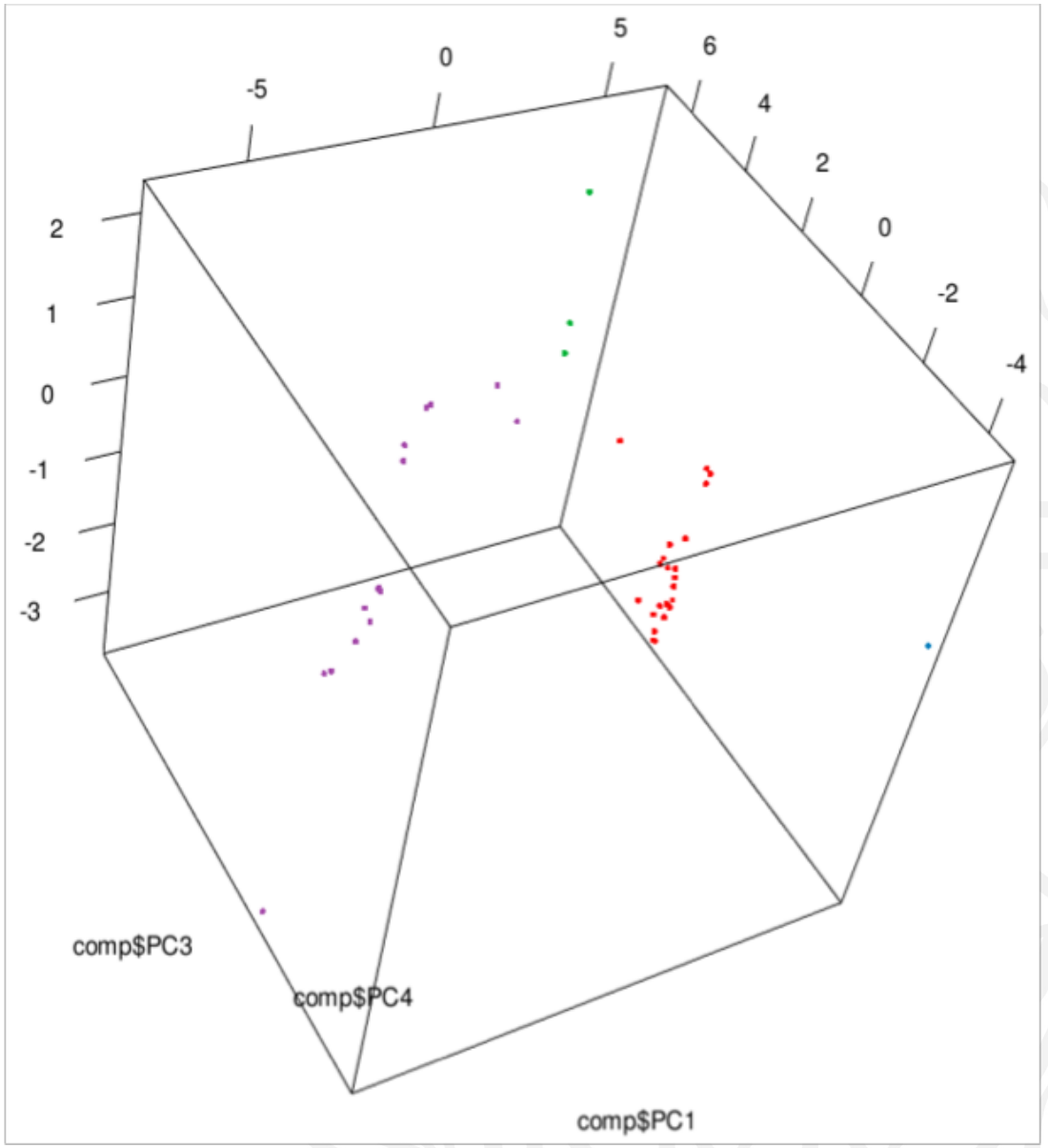


Clustering with the nstart and iter.max parameters leads to consistent results, allowing proper interpretation of the scree plot

- As we can see that "elbow" in the scree plot is at k=4, so we apply the k-means clustering function with k = 4.
- We can see that the one outlier is in its own cluster, there's 3 or 4 in the other and the remainder are split into two clusters of greater size.
- We plotted 3d plots again to explore the space and not lose meaning due to three dimensions being collapsed into a 2D image.




-5
0
5
6
4
2
0
-2
-4
15
10
5
0
comp$PC2
comp$PC3
comp$PC1


-5
0
5
6
4
2
0
-2
-4
2
1
0
-1
-2
-3
comp$PC3
comp$PC4
comp$PC1

- We look at exact clusters below, in order of their increasing size:

```
[1] "Airbus A319 VIP"
```

```
[1] "CRJ 100/200 Pinnacle/SkyWest" "CRJ 100/200 ExpressJet"
[3] "E120"                         "ERJ-145"
```

```
[1] "Airbus A330-200"          "Airbus A330-200 (3L2)"
[3] "Airbus A330-200 (3L3)"    "Airbus A330-300"
[5] "Boeing 747-400 (74S)"     "Boeing 757-200 (75E)"
[7] "Boeing 757-200 (75X)"     "Boeing 767-300 (76G)"
[9] "Boeing 767-300 (76L)"     "Boeing 767-300 (76T)"
[11] "Boeing 767-300 (76Z V.1)" "Boeing 767-300 (76Z V.2)"
[13] "Boeing 767-400 (76D)"     "Boeing 777-200ER"
[15] "Boeing 777-200LR"
```

```
[1] "Airbus A319"           "Airbus A320"           "Airbus A320 32-R"
[4] "Boeing 717"            "Boeing 737-700 (73W)"  "Boeing 737-800 (738)"
[7] "Boeing 737-800 (73H)"  "Boeing 737-900ER (739)" "Boeing 757-200 (75A)"
[10] "Boeing 757-200 (75M)" "Boeing 757-200 (75N)"  "Boeing 757-200 (757)"
[13] "Boeing 757-200 (75V)" "Boeing 757-300"        "Boeing 767-300 (76P)"
[16] "Boeing 767-300 (76Q)" "Boeing 767-300 (76U)"  "CRJ 700"
[19] "CRJ 900"              "E170"                  "E175"
[22] "MD-88"                "MD-90"                 "MD-DC9-50"
```

First Cluster

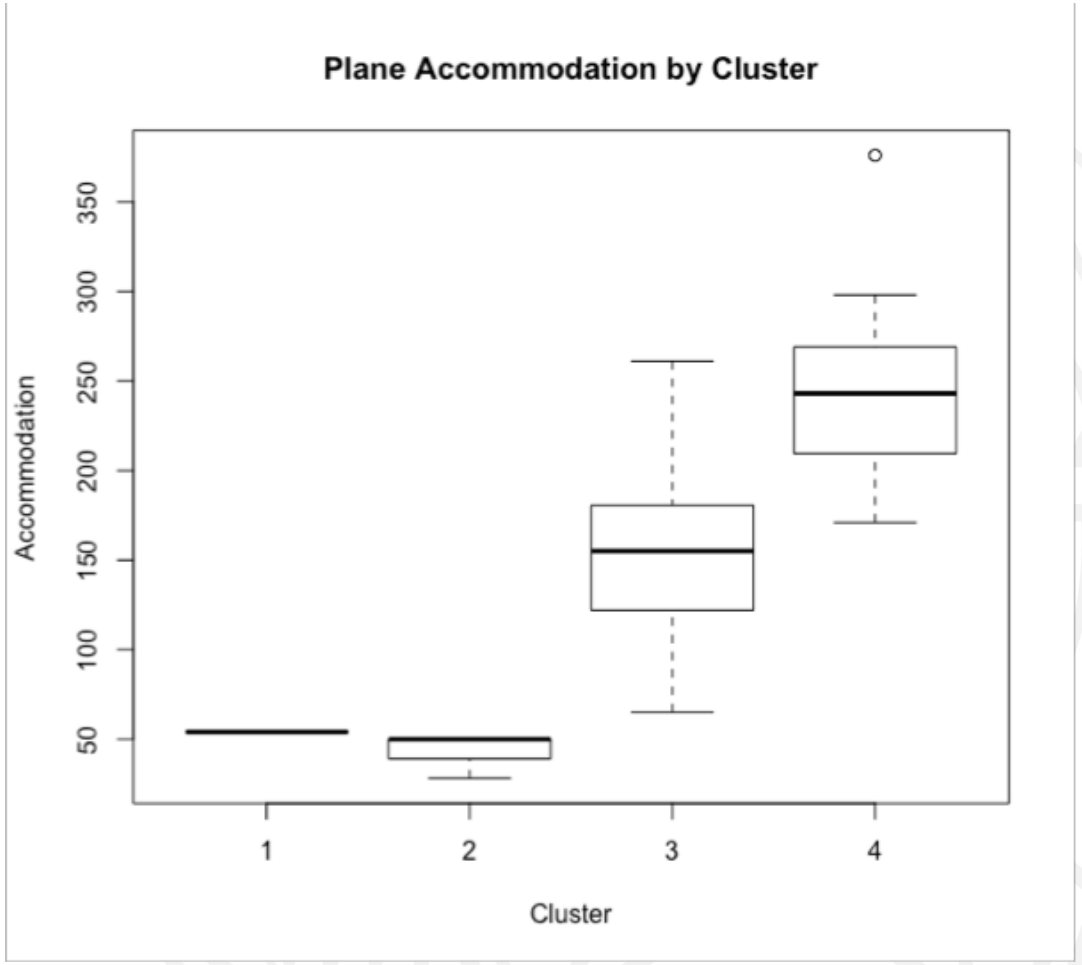
- The first cluster contains a single aircraft, the Airbus A319 VIP. This plane is on its own and rightly so - it is not part of Delta's regular fleet but one of Airbus' corporate jets. This is a plane for people with money, for private charter. It includes "club seats" around tables for working (or not).
- This is apparently the plane professional sports teams and the American military often charter to fly.

Second Cluster

- The second cluster contains four aircraft - the two CRJ 100/200's and the Embraer E120 and ERJ-145.
- These are the smallest passenger aircraft, with the smallest accommodations - 28 for the E120 and 50 for the remaining craft.
- As such, there is only economy seating in these planes which is what distinguishes them from the remainder of the fleet. The E120 also has the distinction of being the only plane in the fleet with turboprops.

Third and Fourth Cluster

- The other two clusters comprise the remainder of the fleet, the planes with which most commercial air travelers are familiar - your Boeing 7-whatever-7's and other Airbus and McDonnell-Douglas planes.
- These are split into two clusters, which seem to again divide the planes approximately by size (both physical and accommodation), though there is crossover in the Boeing craft.




Thank You!