Occupational wage data

CLUSTER ANALYSIS IN R

Dmitriy (Dima) Gorenshteyn

Lead Data Scientist, Memorial Sloan Kettering Cancer Center

Occupational wage data

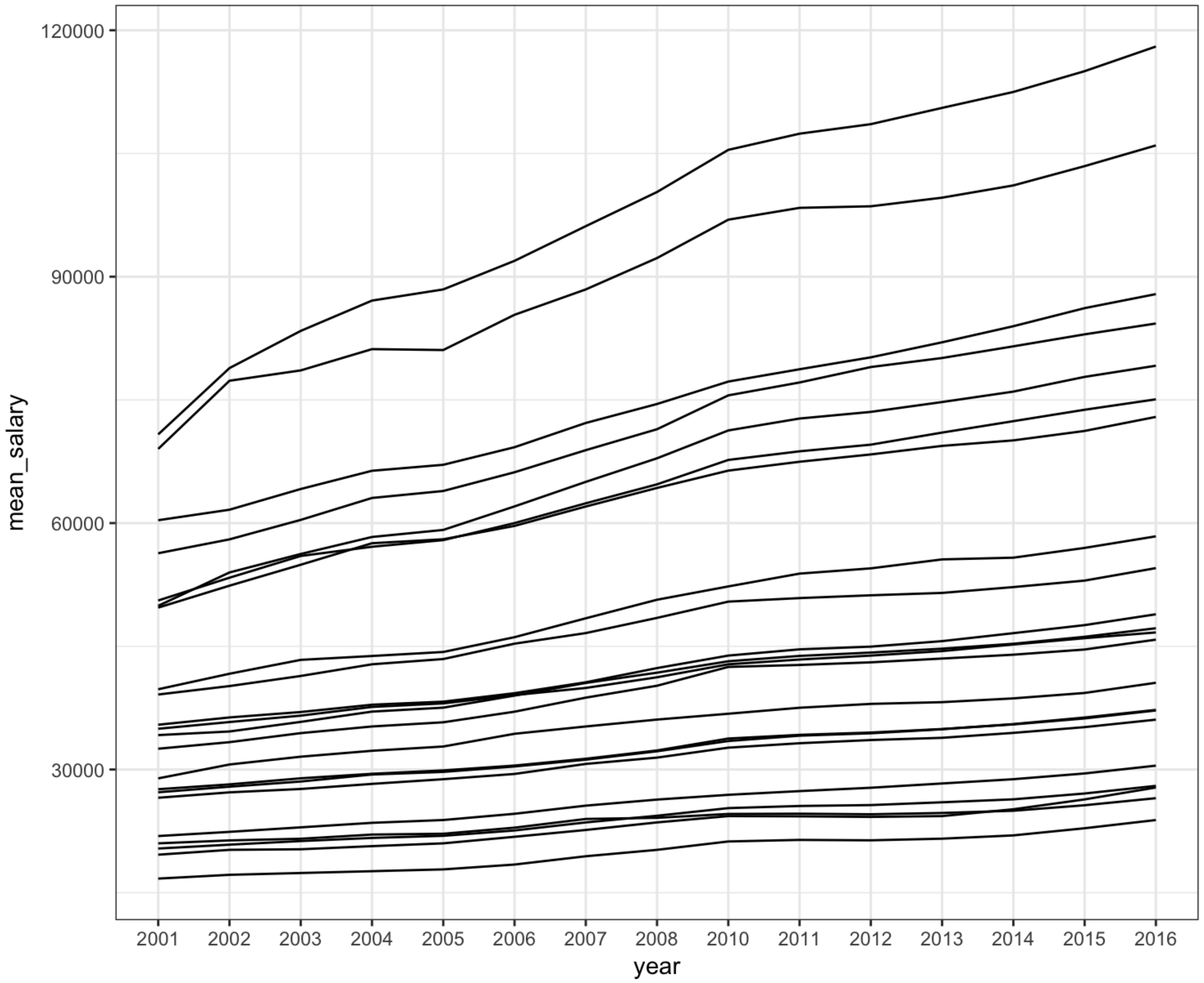
- 22 Occupation Observations
- 15 Measurements of Average Income from 2001-2016

Occupational wage data

```r
print(oes)
```

```
                             2001  2002  2003  2004  2005 ...
Management                  70800 78870 83400 87090 88450 ...
Business Operations         50580 53350 56000 57120 57930 ...
Computer Science            60350 61630 64150 66370 67100 ...
Architecture/Engineering    56330 58020 60390 63060 63910 ...
Life/Physical/Social Sci.   49710 52380 54930 57550 58030 ...
Community Services          34190 34630 35800 37050 37530 ...
...                         ...   ...   ...   ...   ...   ...
```

Occupational wage data



Next steps: hierarchical clustering

- Evaluate whether pre-processing is necessary
- Create a distance matrix
- Build a dendrogram
- Extract clusters from dendrogram
- Explore resulting clusters

Let's practice!

CLUSTER ANALYSIS IN R

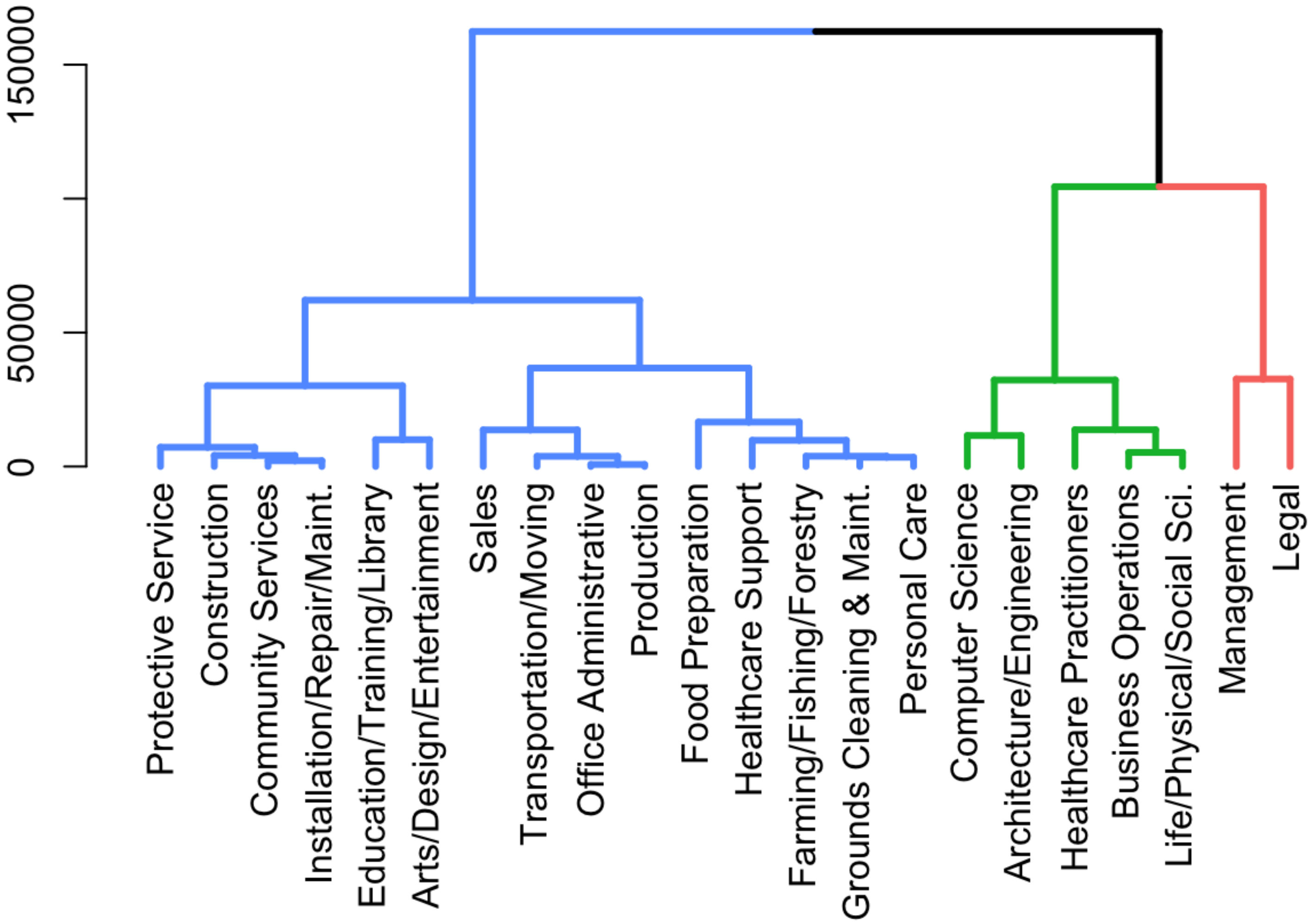
Reviewing the HC results

CLUSTER ANALYSIS IN R

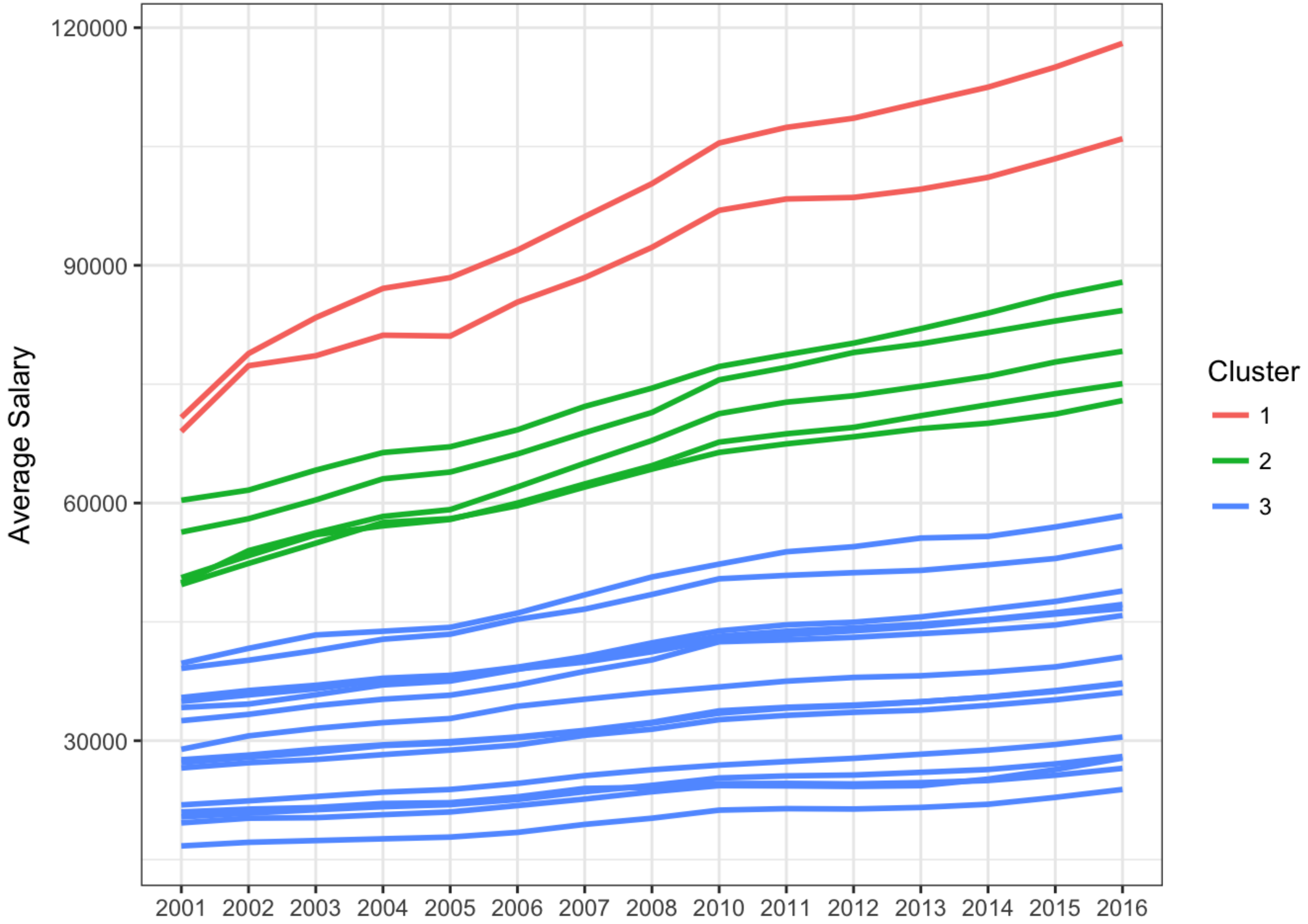
Dmitriy (Dima) Gorenshteyn

Lead Data Scientist, Memorial Sloan Kettering Cancer Center

The dendrogram

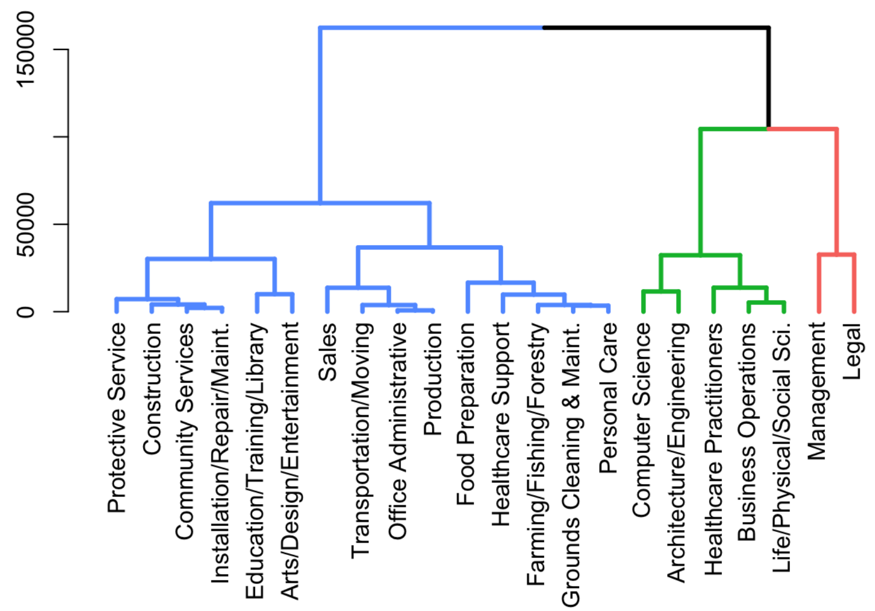


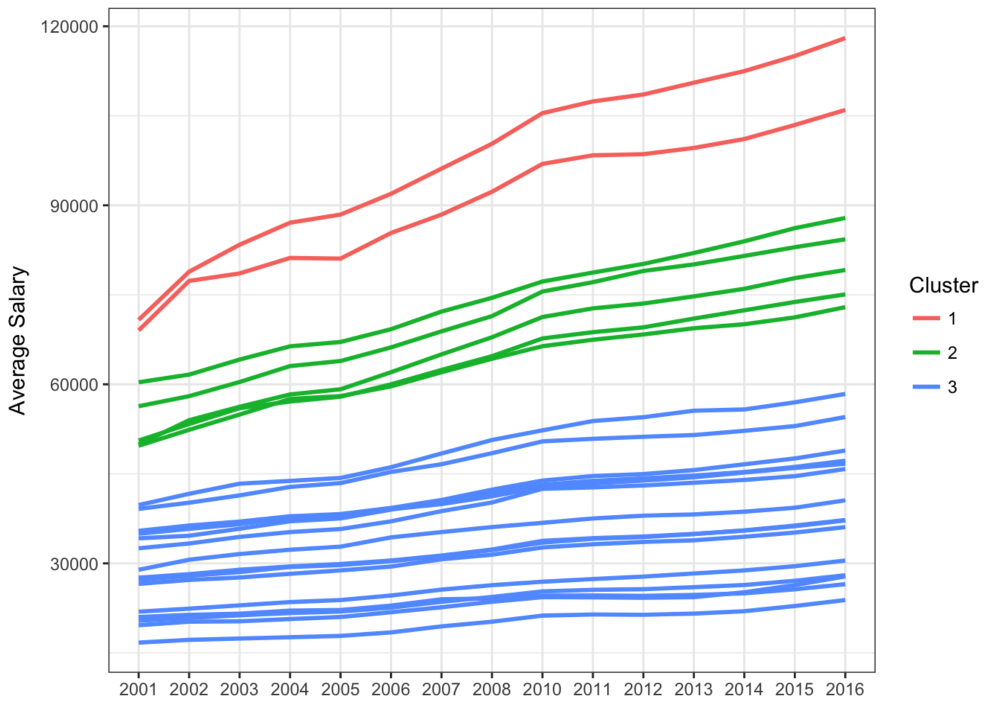
The trends



Average Salary
120000
90000
60000
30000
2001
2002
2003
2004
2005
2006
2007
2008
2010
2011
2012
2013
2014
2015
2016
Cluster
1
2
3

Connecting the two





Next steps: k-means clustering

- Evaluate whether pre-processing is necessary
- Estimate the "best" k using the elbow plot
- Estimate the "best" k using the maximum average silhouette width
- Explore resulting clusters

Let's cluster!

CLUSTER ANALYSIS IN R

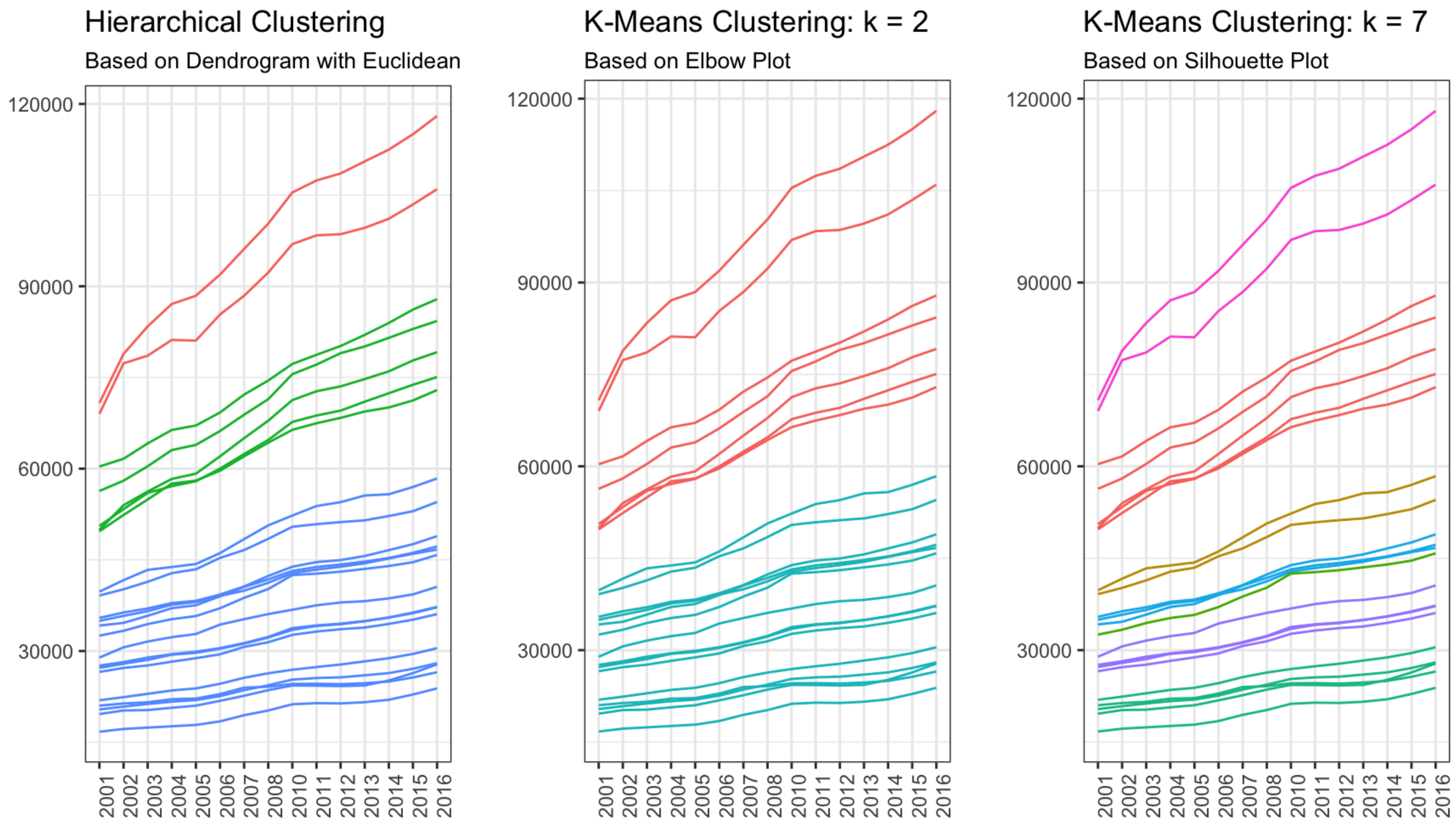
Review K-means results

CLUSTER ANALYSIS IN R

Dmitriy (Dima) Gorenshteyn

Lead Data Scientist, Memorial Sloan Kettering Cancer Center

Three clustering results



Hierarchical Clustering
Based on Dendrogram with Euclidean

K-Means Clustering: k = 2
Based on Elbow Plot

K-Means Clustering: k = 7
Based on Silhouette Plot

Comparing the two clustering methods

	Hierarchical Clustering	k-means
Distance Used:	virtually any	euclidean only
Results Stable:	Yes	No
Evaluating # of Clusters:	dendrogram, silhouette, elbow	silhouette, elbow
Computation Complexity:	Relatively Higher	Relatively Lower

What have you learned?

- **Chapter 1:**
  - What is distance
  - Why is scale important
- **Chapter 2:**
  - How linkage works
  - How the dendrogram is formed
  - How to analyze your clusters
- **Chapter 3:**
  - How k-means works
  - How to estimate k
  - How to analyze how well an observation fits in a cluster

A lot more to learn

- k-mediods
- DBSCAN
- Optics

Congratulations!

CLUSTER ANALYSIS IN R