Infrastructure Development of Schools in India

Harsh Talajia

14/03/2021

Objective of the Study

The aim of this study is to determine the state-wise level of infrastructure development of schools in India on the basis of basic facilities provided by the school. Based on this criteria, similar states can be clustered together so that the States belonging to the same cluster can have a common development policy and strategy.

About Data

The data is broken down on state level. The values of variables 'Computer', 'Electricity', 'Water', 'Boys_washroom', and 'Girls_washroom' indicate the percentage of schools in that state having that respective facility. The data of year 2015-16 is used here, since it is the latest one available.

This dataset was compiled from five different files taken from Open Government Data platform: data.gov.in/catalog/school-education-statistics

```r
library(readxl)
schools <- read_excel("C:/Users/Harsh/Desktop/M.Sc ASA/SEM 2/AMDA/ICA 2/schools_2016.xlsx")
schools<-data.frame(schools[,-1], row.names=schools$State_UT)
kable(head(schools))
```

	Computer	Electricity	Water	Boys_washroom	Girls_washroom
Andaman & Nicobar	57.00	90.10	100.00	100.00	100.00
Andhra Pradesh	30.59	93.50	95.37	99.69	99.72
Arunachal Pradesh	24.36	39.54	81.47	95.65	96.56
Assam	10.76	25.55	86.21	82.80	83.94
Bihar	9.37	37.78	94.43	89.16	90.05
Chandigarh	94.53	100.00	100.00	100.00	100.00

```r
dim(schools)
```

```
[1] 36 5
```

The value of first cell 57 indicates that in Andaman & Nicobar, 57% schools have computer facility.

Number of observations is 36, which is the number of States and Union territories in India. The five variables mentioned above are used to measure infrastructure development of schools.

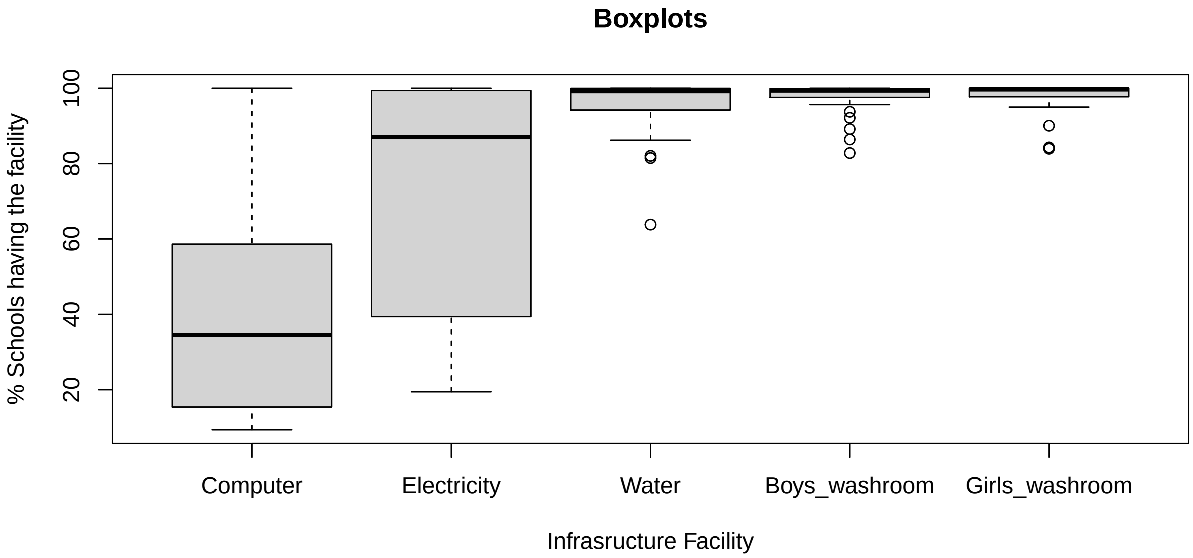
Exploring Data

Checking missing values and outliers

```
kable(sapply(schools, function(x)sum(is.na(x))))
```

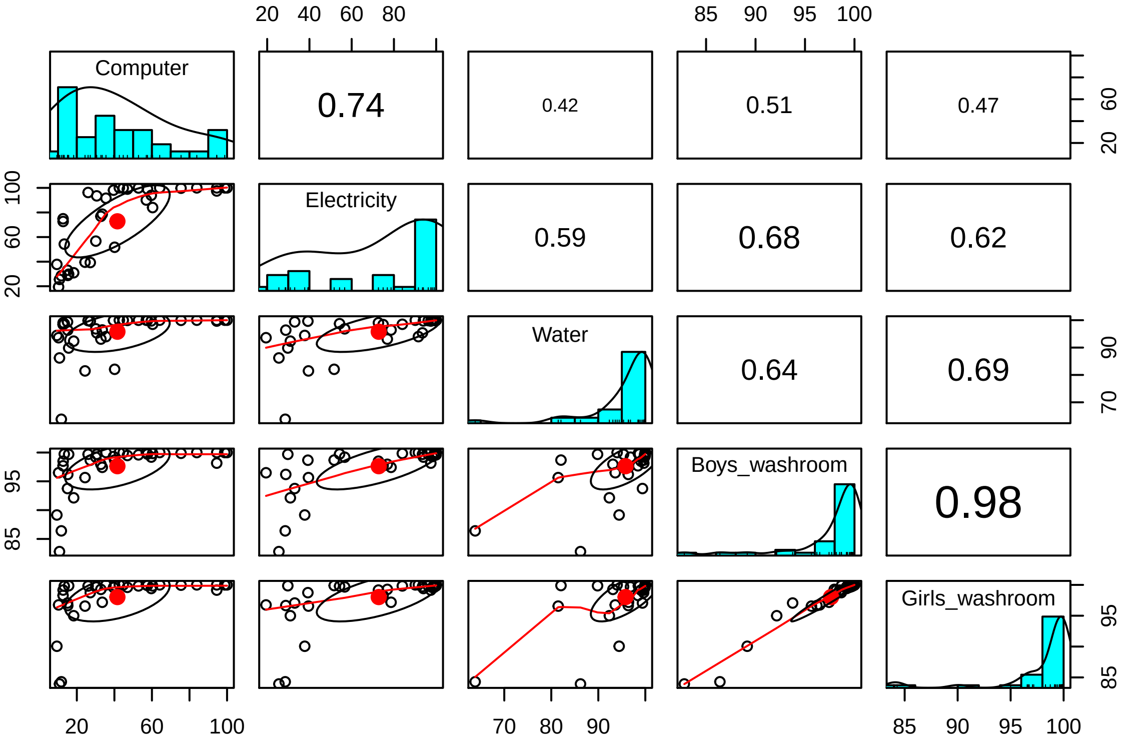
	x
Computer	0
Electricity	0
Water	0
Boys_washroom	0
Girls_washroom	0

No missing observations are present in the data.



In a typical state, very few schools have Computers. Electricity coverage varies from state to state. Overall, drinking water and washroom facilities are available for most of the schools across all states.

Checking the distribution shape and correlation of variables



Boys_washroom and Girls_washroom variables have nearly perfect linear relationship. Therefore, we replace these two variables with a single variable named 'Washroom' whose value is their Geometric Mean.

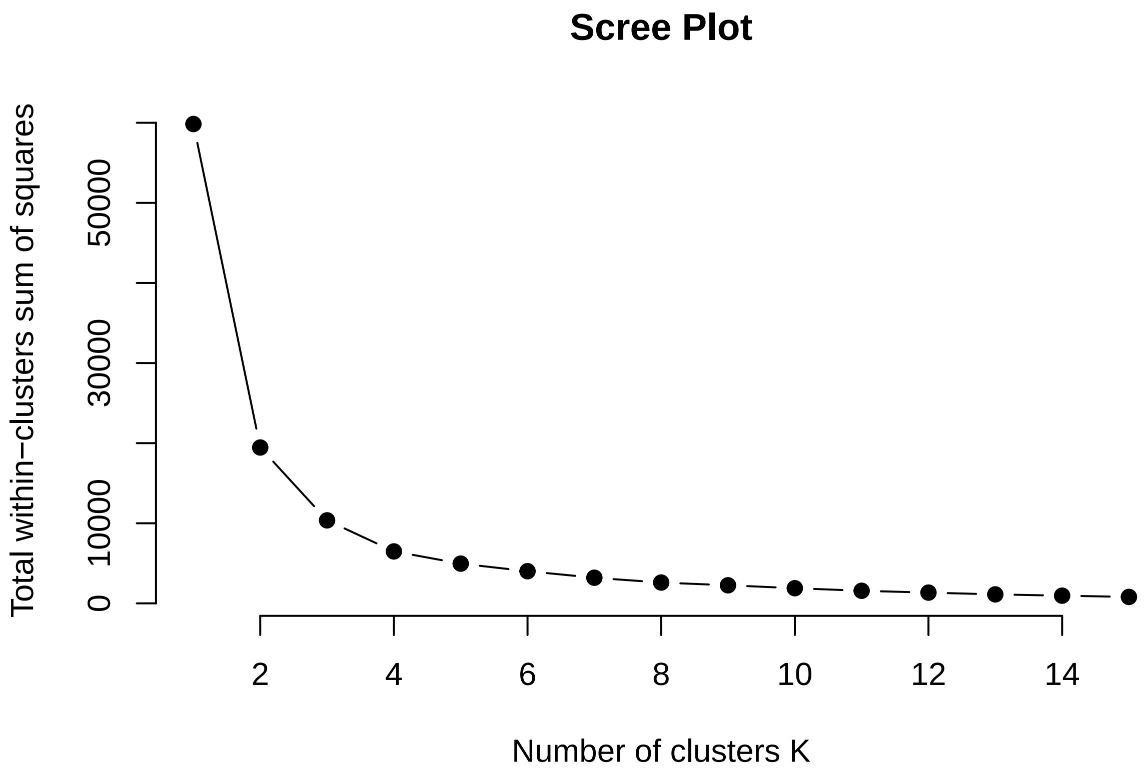
Also, data does not need to be scaled, since each variable is measured in terms of percentage.

Final data that is used for clustering:

	Computer	Electricity	Water	Washroom
Andaman & Nicobar	57.00	90.10	100.00	100.00
Andhra Pradesh	30.59	93.50	95.37	99.70
Arunachal Pradesh	24.36	39.54	81.47	96.10
Assam	10.76	25.55	86.21	83.37
Bihar	9.37	37.78	94.43	89.60
Chandigarh	94.53	100.00	100.00	100.00

Finding optimal number of clusters

While forming clusters, we try to minimize within-cluster sum of squares, and maximize between-cluster sum of squares.



Scree plot suggests that either 3 or 4 clusters should be formed.

The variance in data that can be explained by clusters is the ratio of Between S.S. and Total S.S.

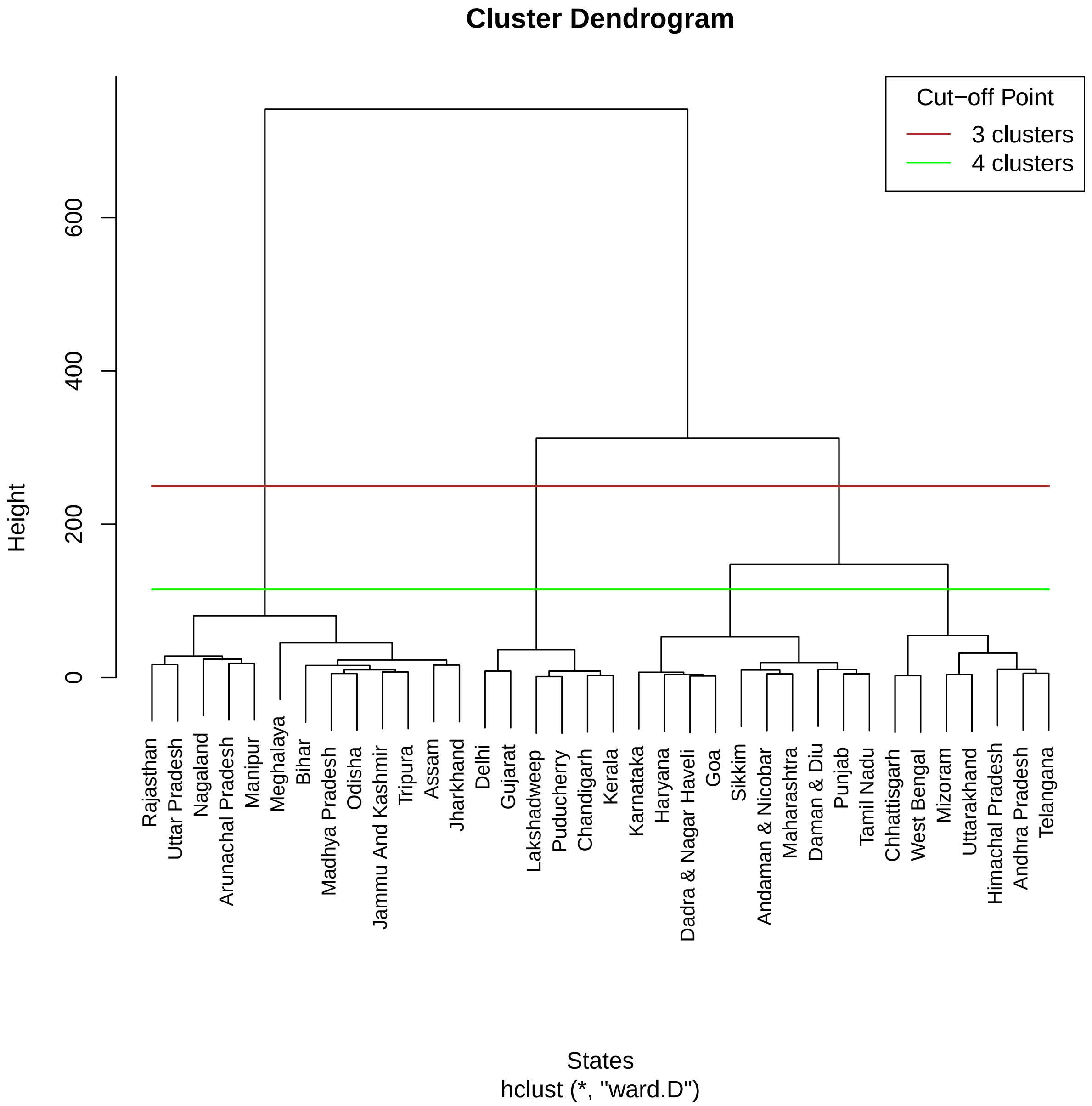
i.e. Variance explained = Between S.S. / Total S.S.

No_of_Clusters	Variance_Explained
1	0%
2	67.48%
3	82.68%
4	89.16%
5	91.71%

From the above table, 3 clusters explain 83% variation whereas 4 clusters explain 89% variation. The difference is of only 6%.

Ward's Heirarchical clustering

```
##
## Call:
## hclust(d = dist_matrix, method = "ward.D")
##
## Cluster method   : ward.D
## Distance         : euclidean
## Number of objects: 36
```



Dendrogram suggests that either 3 or 4 clusters are suitable.

Taking the Sum of Squares Analysis and Ward's Clustering into consideration, we will proceed with 3 clusters.

Applying k-means clustering for k=3

```
set.seed(111)
schools_kmean <- kmeans(schools, 3)
schools_kmean$size
```

```
## [1] 17 13  6
```

3 Clusters of sizes 17, 13, 6 respectively are formed.

The cluster assignments are as follows:

```
##    Andaman & Nicobar       Andhra Pradesh    Arunachal Pradesh
##                    1                    1                    2
##                Assam                Bihar           Chandigarh
##                    2                    2                    3
##         Chhattisgarh Dadra & Nagar Haveli          Daman & Diu
##                    1                    1                    1
##                Delhi                  Goa              Gujarat
##                    3                    1                    3
##              Haryana     Himachal Pradesh    Jammu And Kashmir
##                    1                    1                    2
##            Jharkhand            Karnataka               Kerala
##                    2                    1                    3
##          Lakshadweep       Madhya Pradesh          Maharashtra
##                    3                    2                    1
##              Manipur            Meghalaya              Mizoram
##                    2                    2                    1
##             Nagaland               Odisha           Puducherry
##                    2                    2                    3
##               Punjab            Rajasthan               Sikkim
##                    1                    2                    1
##           Tamil Nadu            Telangana              Tripura
##                    1                    1                    2
##        Uttar Pradesh          Uttarakhand          West Bengal
##                    2                    1                    1
```

The following table contains values of **cluster centres**:

Cluster	Size	Computer	Electricity	Water	Washroom	Tier
1	17	41.72	91.08	98.49	99.40	Tier 2
2	13	18.57	36.56	90.36	94.87	Tier 3
3	6	91.20	99.53	99.94	99.76	Tier 1

Cluster 3 consists of 6 states where almost every school has these infrastructure facilities. Among all variables, 'Computers' has lowest mean of 91%. Lets call these **'Tier 1' states**.

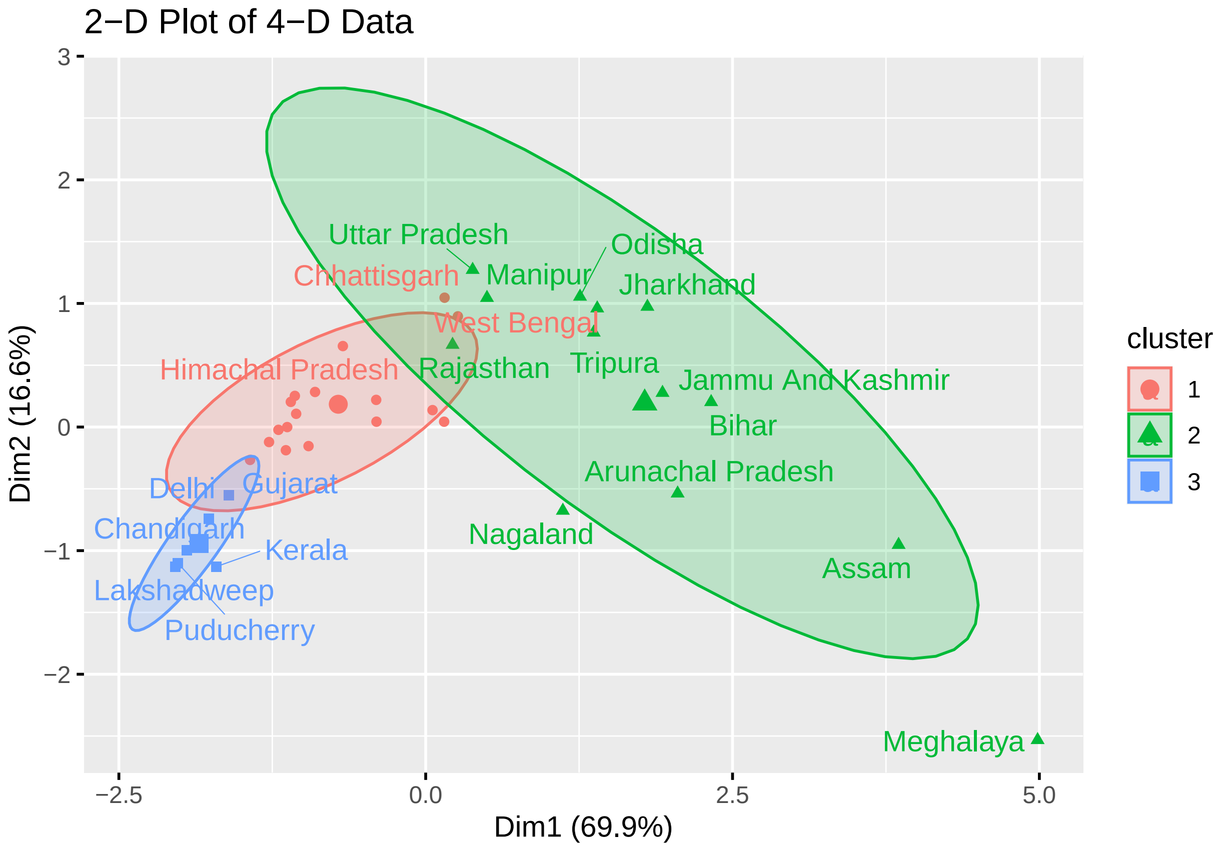
Cluster 1 consists of 17 states where all facilities, except computers, are available in most of the schools. Although Electricity mean is 93%, the mean for Computer is only 42%. These are **'Tier 2' states**.

Cluster 2 consists of 13 States which are under-performing in all 4 areas. These are **'Tier 3' states**.

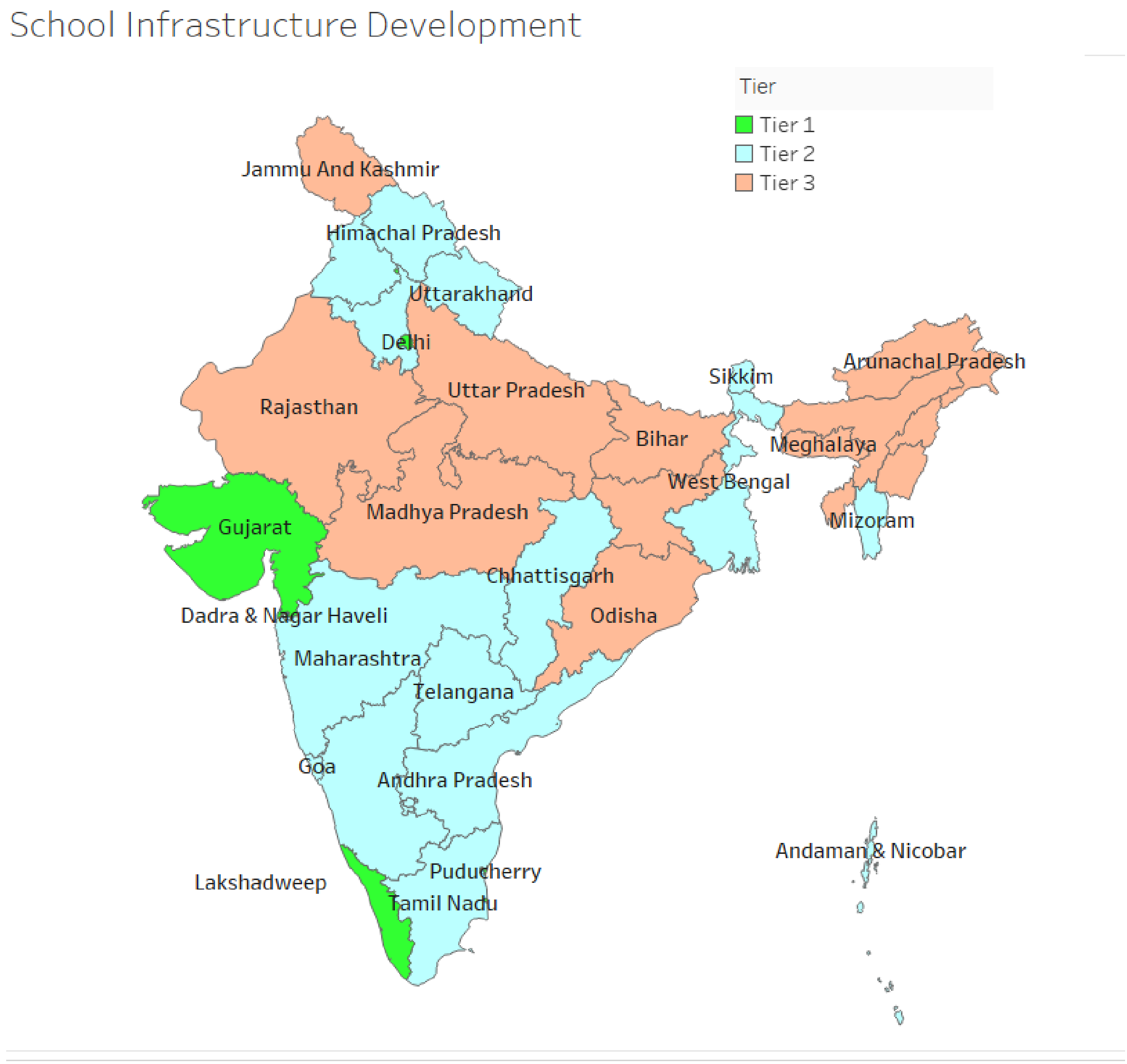
Segregating the states into three tiers as per their cluster membership.

State	Cluster	Tier
Andaman & Nicobar	1	Tier 2
Andhra Pradesh	1	Tier 2
Arunachal Pradesh	2	Tier 3
Assam	2	Tier 3
Bihar	2	Tier 3
Chandigarh	3	Tier 1

```
## Warning: ggrepel: 15 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Both the dimensions collectively explain 69.9% + 16.6% = **86.5% variance** in the data.



Conclusion

The following conclusions are made on the basis of cluster centre values.

Nearly all the schools within the **Tier 1 states** have all infrastructure facilities. Very few schools do not have computers.

Most of the schools within **Tier 2 states** have access to basic facilities like Water, Electricity, and Washrooms. Despite of having electricity, many schools do not have computers. These states should focus on providing students with computers and other relevant technology.

Tier 3 states are under-performing in every aspect. A considerable proportion of Schools do not have access to even basic facilities like Drinking Water and Washroom. On average, only 90% schools have access to Water, and 95% schools have washrooms. This implies that 10% schools do not have access to water and 5% schools do not have washrooms. These States foremost ought to identify such schools and make sure that these basic human needs are fulfilled.