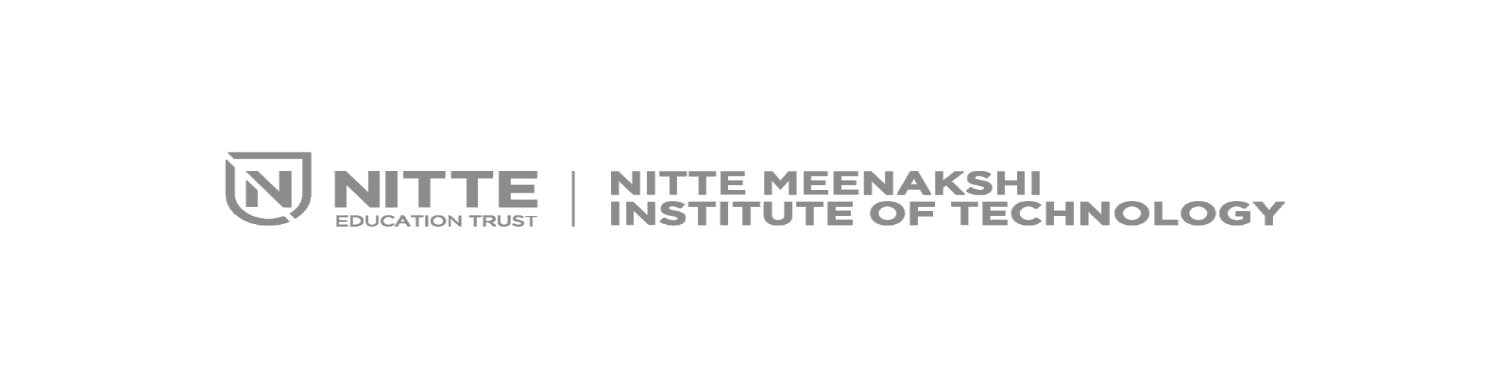
****

**Department of Information Science & Engineering**

**Exploratory Data Analysis**

**Report (LA Assignment)**

**Submitted to**

**Dr. Manoj Kumar M V**

**Associate Professor**

**Department of Information Science and Engineering**

**Team Members**

Pooja Rai (1NT18IS111)

Vivek Kumar (1NT18IS184)

# Exploratory Data Analysis

To begin with, we import the data set into a data frame named my\_data, using read.csv function with sep=”,” to specify comma separated csv file.

my\_data <- read.csv("COVID-19 India Statewise Vaccine Data.csv", header = TRUE, sep = ",")  
show(my\_data)

## State.UTs Dose.1 Dose.2  
## 1 Andaman and Nicobar 156470 19823  
## 2 Andhra Pradesh 12588369 3116201  
## 3 Arunachal Pradesh 509275 84530  
## 4 Assam 5792114 1230867  
## 5 Bihar 13684112 2215545  
## 6 Chandigarh 439301 87576  
## 7 Chhattisgarh 7926420 1600677  
## 8 Dadra and Nagar Haveli and Daman and Diu 387082 40234  
## 9 Delhi 6088881 1804655  
## 10 Goa 817614 116566  
## 11 Gujarat 20070986 5621236  
## 12 Haryana 7433759 1423139  
## 13 Himachal Pradesh 3351501 514786  
## 14 Jammu and Kashmir 3845292 680870  
## 15 Jharkhand 5790247 1064300  
## 16 Karnataka 18973266 3739413  
## 17 Kerala 10842601 3254223  
## 18 Ladakh 170902 56034  
## 19 Lakshadweep 46779 7921  
## 20 Madhya Pradesh 17917804 2396391  
## 21 Maharashtra 26067919 6371846  
## 22 Manipur 581261 74141  
## 23 Meghalaya 619745 79481  
## 24 Mizoram 518686 54369  
## 25 Nagaland 440344 59874  
## 26 Odisha 9868279 2148249  
## 27 Puducherry 440326 67068  
## 28 Punjab 6112362 1002728  
## 29 Rajasthan 20751161 3980702  
## 30 Sikkim 401506 72304  
## 31 Tamil Nadu 13073449 2569324  
## 32 Telangana 9493770 1577710  
## 33 Tripura 1955931 594420  
## 34 Uttar Pradesh 26792830 4488619  
## 35 Uttarakhand 3579840 821974  
## 36 West Bengal 16833415 5056919  
## 37 Miscellaneous 1736281 1521424  
## Total.Vaccination.Doses  
## 1 176293  
## 2 15704570  
## 3 593805  
## 4 7022981  
## 5 15899657  
## 6 526877  
## 7 9527097  
## 8 427316  
## 9 7893536  
## 10 934180  
## 11 25692222  
## 12 8856898  
## 13 3866287  
## 14 4526162  
## 15 6854547  
## 16 22712679  
## 17 14096824  
## 18 226936  
## 19 54700  
## 20 20314195  
## 21 32439765  
## 22 655402  
## 23 699226  
## 24 573055  
## 25 500218  
## 26 12016528  
## 27 507394  
## 28 7115090  
## 29 24731863  
## 30 473810  
## 31 15642773  
## 32 11071480  
## 33 2550351  
## 34 31281449  
## 35 4401814  
## 36 21890334  
## 37 3257705

The data set we are working on is a collection of the vaccination drive for COVID-19. The data includes state-wise distribution of each dose and cumulative doses administered.

Further we add a few columns to make our analysis more relevant. Below we add the data on population of each state and calculate the total percentage of population vaccinated based on that data.

my\_data$populations <- c(380581,49576777,1383727,31205576,104099452,1055450,25545198,586956,16787941,1458545,60439692,25351462,6864602,12258433,32988134,61095297,33406061,274289,64473,72626809,112374333,2855794,2966889,1097206,1978502,41974218,1247953,27743338,68548437,610577,72147030,35004000,3673917,199812341,10086292,91276115,NA)  
  
f\_data <- mutate(my\_data, percent\_vaccinated = (my\_data$Dose.2\*100)/my\_data$populations)  
  
show(f\_data)

## State.UTs Dose.1 Dose.2  
## 1 Andaman and Nicobar 156470 19823  
## 2 Andhra Pradesh 12588369 3116201  
## 3 Arunachal Pradesh 509275 84530  
## 4 Assam 5792114 1230867  
## 5 Bihar 13684112 2215545  
## 6 Chandigarh 439301 87576  
## 7 Chhattisgarh 7926420 1600677  
## 8 Dadra and Nagar Haveli and Daman and Diu 387082 40234  
## 9 Delhi 6088881 1804655  
## 10 Goa 817614 116566  
## 11 Gujarat 20070986 5621236  
## 12 Haryana 7433759 1423139  
## 13 Himachal Pradesh 3351501 514786  
## 14 Jammu and Kashmir 3845292 680870  
## 15 Jharkhand 5790247 1064300  
## 16 Karnataka 18973266 3739413  
## 17 Kerala 10842601 3254223  
## 18 Ladakh 170902 56034  
## 19 Lakshadweep 46779 7921  
## 20 Madhya Pradesh 17917804 2396391  
## 21 Maharashtra 26067919 6371846  
## 22 Manipur 581261 74141  
## 23 Meghalaya 619745 79481  
## 24 Mizoram 518686 54369  
## 25 Nagaland 440344 59874  
## 26 Odisha 9868279 2148249  
## 27 Puducherry 440326 67068  
## 28 Punjab 6112362 1002728  
## 29 Rajasthan 20751161 3980702  
## 30 Sikkim 401506 72304  
## 31 Tamil Nadu 13073449 2569324  
## 32 Telangana 9493770 1577710  
## 33 Tripura 1955931 594420  
## 34 Uttar Pradesh 26792830 4488619  
## 35 Uttarakhand 3579840 821974  
## 36 West Bengal 16833415 5056919  
## 37 Miscellaneous 1736281 1521424  
## Total.Vaccination.Doses populations percent\_vaccinated  
## 1 176293 380581 5.208615  
## 2 15704570 49576777 6.285606  
## 3 593805 1383727 6.108864  
## 4 7022981 31205576 3.944382  
## 5 15899657 104099452 2.128297  
## 6 526877 1055450 8.297503  
## 7 9527097 25545198 6.266058  
## 8 427316 586956 6.854688  
## 9 7893536 16787941 10.749710  
## 10 934180 1458545 7.991937  
## 11 25692222 60439692 9.300570  
## 12 8856898 25351462 5.613637  
## 13 3866287 6864602 7.499138  
## 14 4526162 12258433 5.554299  
## 15 6854547 32988134 3.226312  
## 16 22712679 61095297 6.120623  
## 17 14096824 33406061 9.741415  
## 18 226936 274289 20.428818  
## 19 54700 64473 12.285763  
## 20 20314195 72626809 3.299596  
## 21 32439765 112374333 5.670197  
## 22 655402 2855794 2.596161  
## 23 699226 2966889 2.678934  
## 24 573055 1097206 4.955223  
## 25 500218 1978502 3.026229  
## 26 12016528 41974218 5.118020  
## 27 507394 1247953 5.374241  
## 28 7115090 27743338 3.614302  
## 29 24731863 68548437 5.807138  
## 30 473810 610577 11.841913  
## 31 15642773 72147030 3.561233  
## 32 11071480 35004000 4.507228  
## 33 2550351 3673917 16.179462  
## 34 31281449 199812341 2.246417  
## 35 4401814 10086292 8.149417  
## 36 21890334 91276115 5.540243  
## 37 3257705 NA NA

After this we summarize our data using the summary method.

summary(f\_data)

## State.UTs Dose.1 Dose.2   
## Length:37 Min. : 46779 Min. : 7921   
## Class :character 1st Qu.: 518686 1st Qu.: 79481   
## Mode :character Median : 5790247 Median :1064300   
## Mean : 7462159 Mean :1611247   
## 3rd Qu.:12588369 3rd Qu.:2396391   
## Max. :26792830 Max. :6371846   
##   
## Total.Vaccination.Doses populations percent\_vaccinated  
## Min. : 54700 Min. : 64473 Min. : 2.128   
## 1st Qu.: 593805 1st Qu.: 1439840 1st Qu.: 3.862   
## Median : 6854547 Median : 21069702 Median : 5.642   
## Mean : 9073406 Mean : 33634622 Mean : 6.605   
## 3rd Qu.:15642773 3rd Qu.: 52292506 3rd Qu.: 8.031   
## Max. :32439765 Max. :199812341 Max. :20.429   
## NA's :1 NA's :1

summary(f\_data$Dose.1)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 46779 518686 5790247 7462159 12588369 26792830

summary(f\_data$Dose.2)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 7921 79481 1064300 1611247 2396391 6371846

summary(f\_data$Total.Vaccination.Doses)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 54700 593805 6854547 9073406 15642773 32439765

summary(f\_data$percent\_vaccinated)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 2.128 3.862 5.642 6.605 8.031 20.429 1

From the summary we can observe that on an average 9073406 doses have been administered. Approximately 6.605% population has been vaccinated.

Next, we find the maximum and minimum values in each respective column. We use the max and min functions for this with the first parameter as the variable and rm.na parameter to exclude the NA values.

max(f\_data$Dose.1)

## [1] 26792830

max(f\_data$Dose.2)

## [1] 6371846

max(f\_data$Total.Vaccination.Doses, na.rm = TRUE)

## [1] 32439765

max(f\_data$percent\_vaccinated, na.rm = TRUE)

## [1] 20.42882

min(f\_data$Dose.1)

## [1] 46779

min(f\_data$Dose.2)

## [1] 7921

min(f\_data$Total.Vaccination.Doses, na.rm = TRUE)

## [1] 54700

min(f\_data$percent\_vaccinated, na.rm = TRUE)

## [1] 2.128297

The maximum population vaccinated in a specific state is 20.42%, in the state of Ladakh. On the other hand Bihar has the lowest vaccinated population of 2.12%.

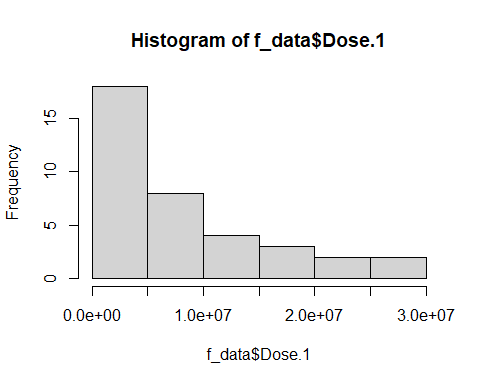
Maharashtra has administered the maximum number of doses and Lakshadweep has the lowest doses administered.

## PLOTS

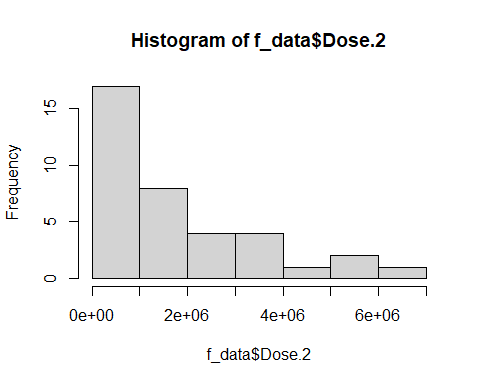
#### Histograms

We plot the histograms of each variable using the hist() function.

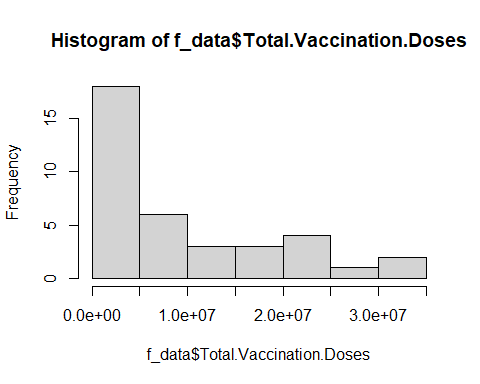
hist(f\_data$Dose.1)



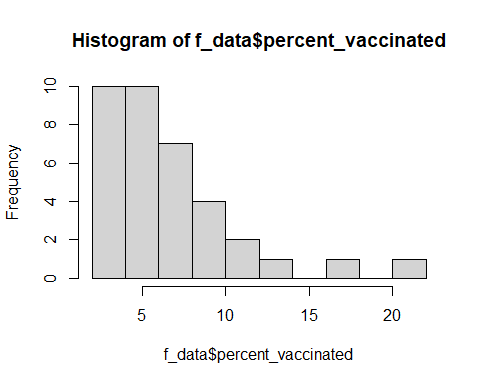
hist(f\_data$Dose.2)



hist(f\_data$Total.Vaccination.Doses)

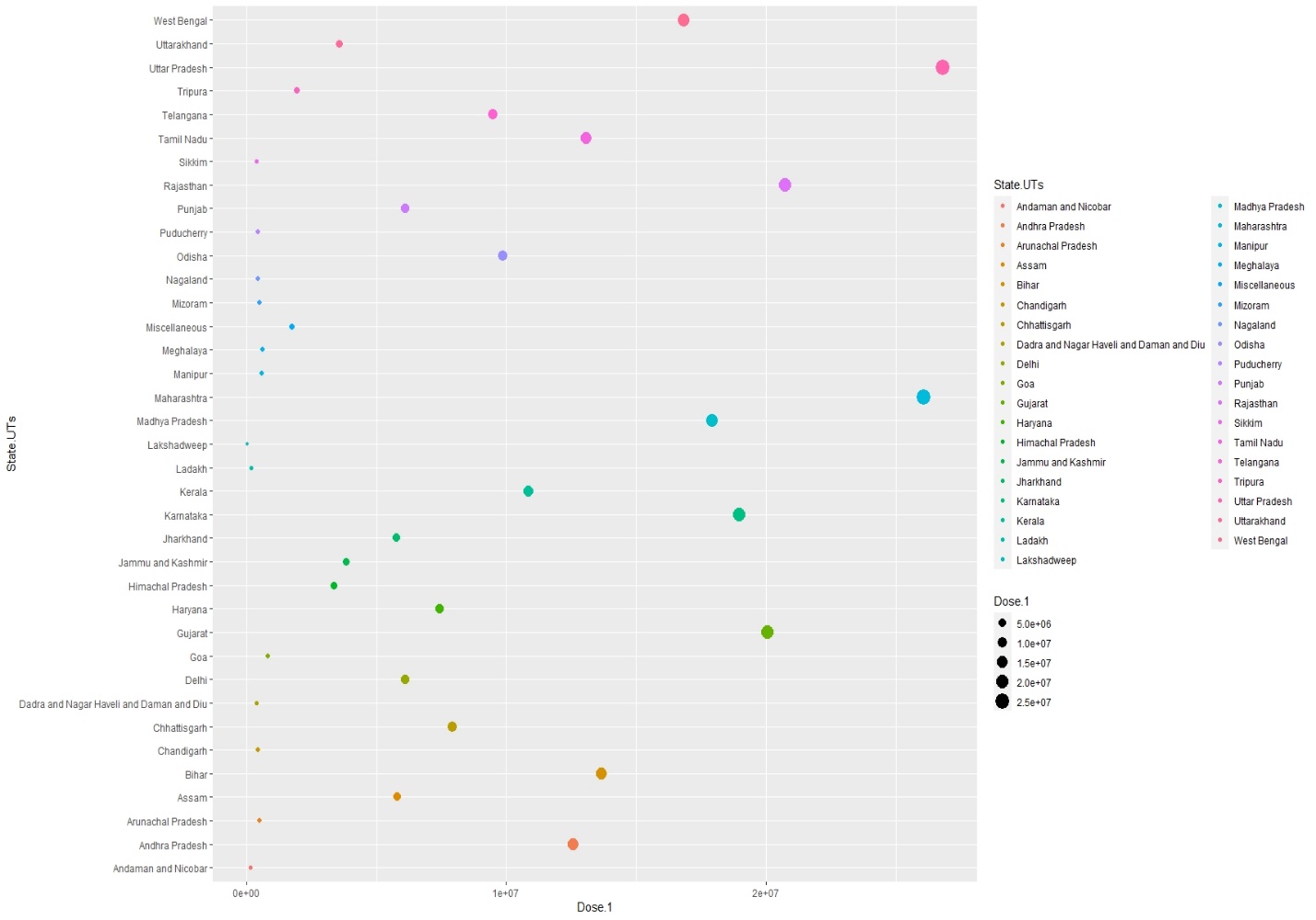


hist(f\_data$percent\_vaccinated)

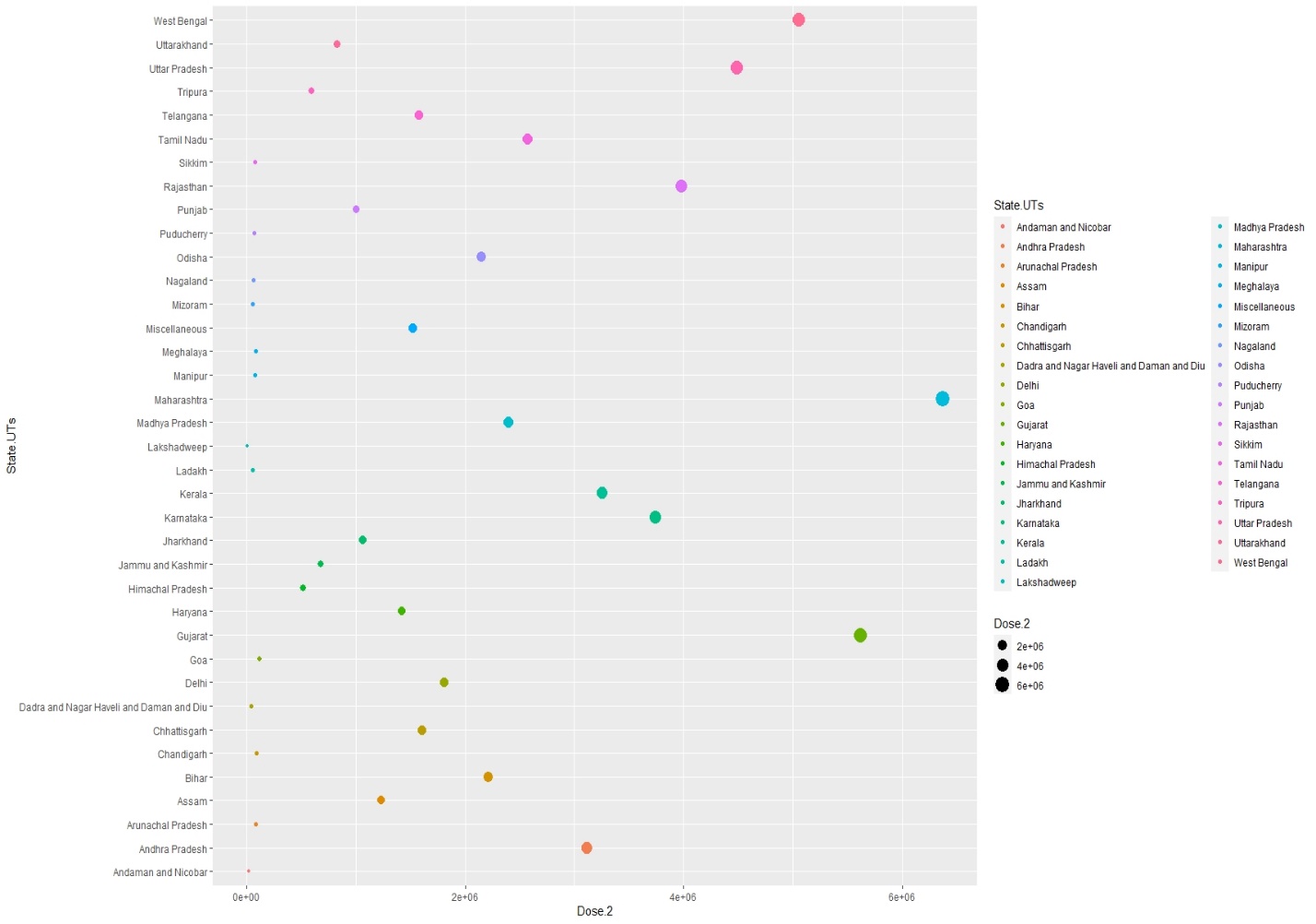


Next we plot graphs for state-wise distribution of doses administered. ggplot is used for this.

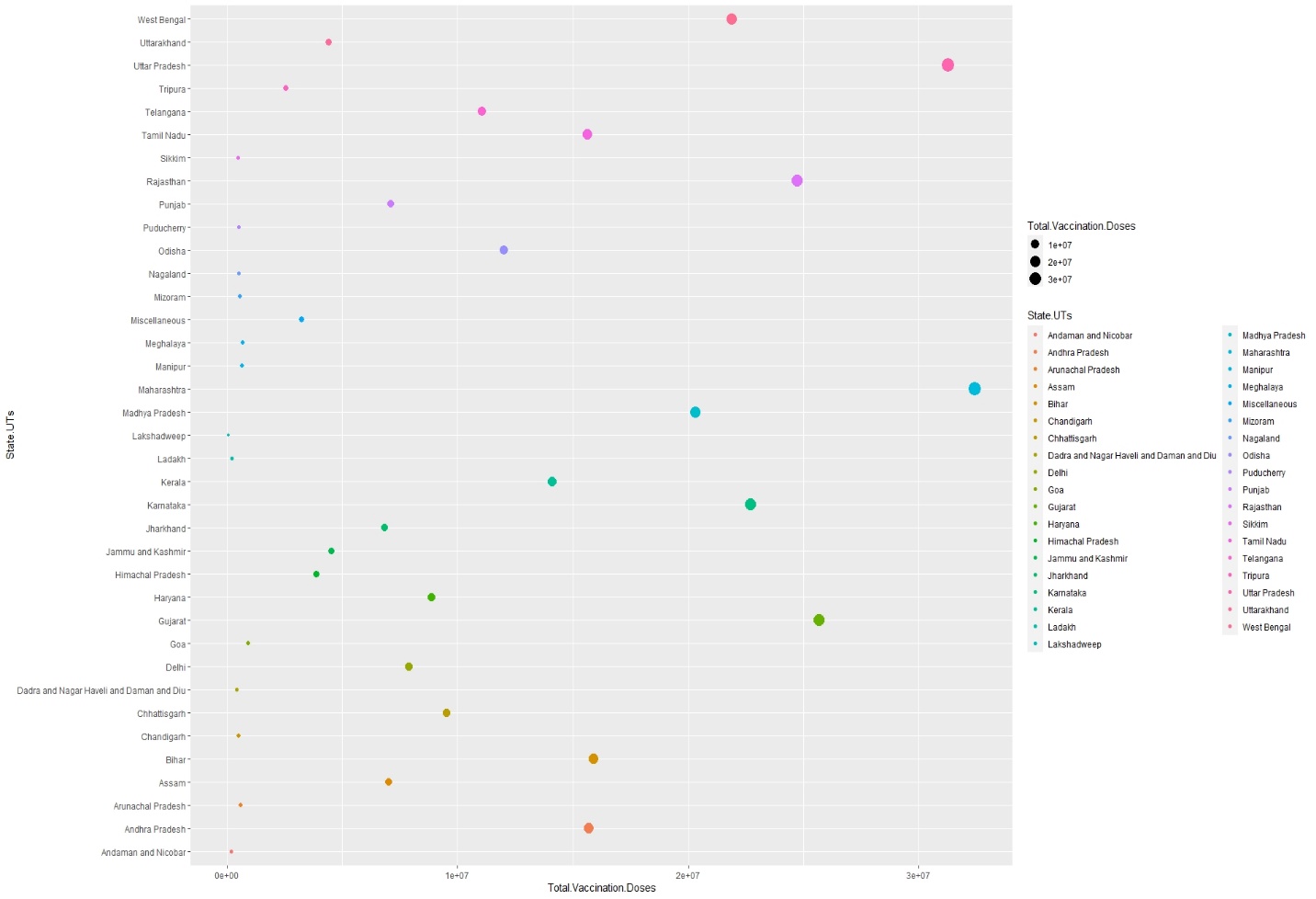
ggplot(f\_data, aes(x = Dose.1, y = State.UTs, col = State.UTs, size = Dose.1)) + geom\_point()



ggplot(f\_data, aes(x = Dose.2, y = State.UTs, col = State.UTs, size = Dose.2)) + geom\_point()

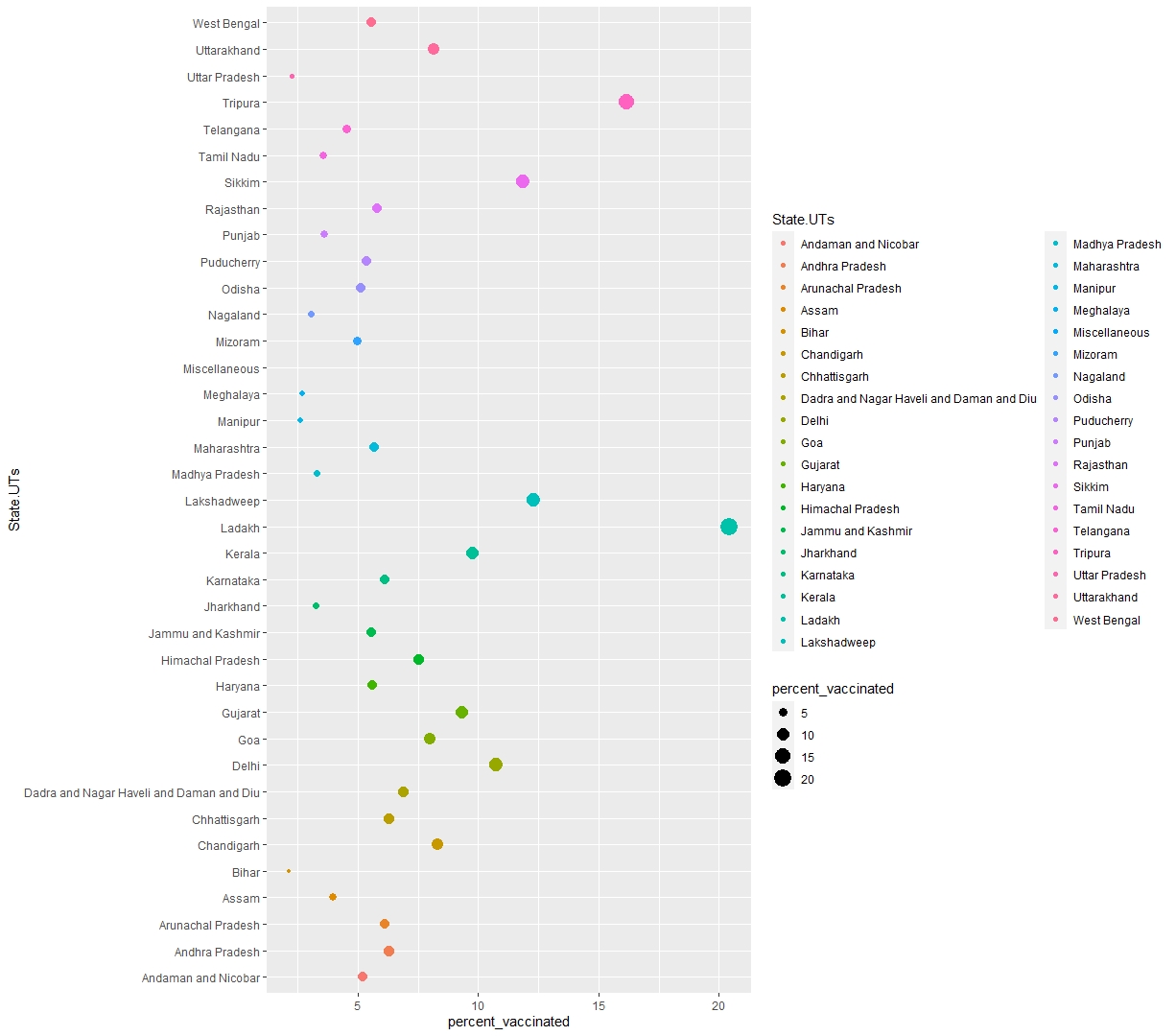


ggplot(f\_data, aes(x = Total.Vaccination.Doses, y = State.UTs, col = State.UTs, size = Total.Vaccination.Doses)) + geom\_point()



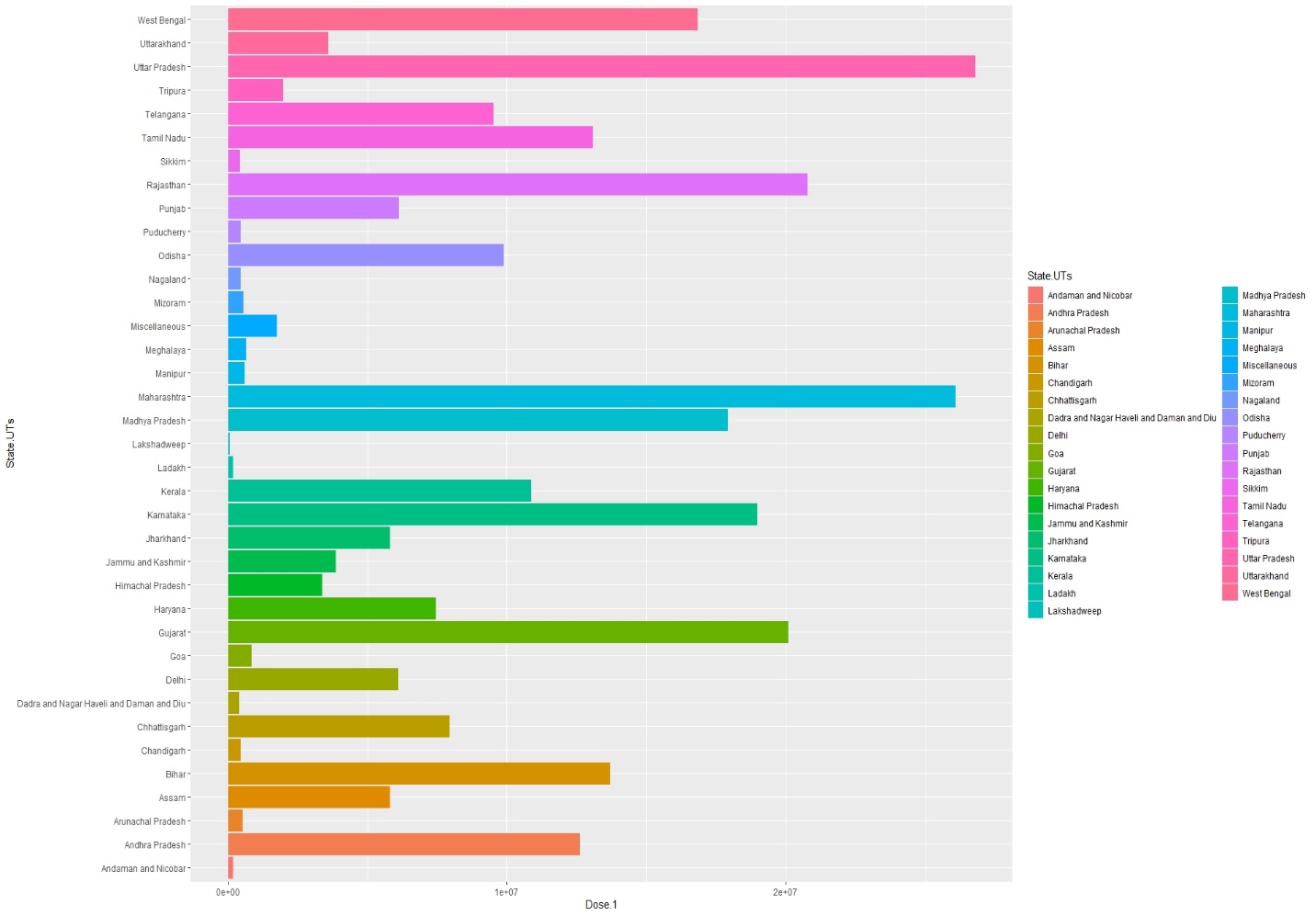
Next we plot percentage vaccinated graph for all states/UTs. The size parameter helps us display the density of the variables.

ggplot(f\_data, aes(x = percent\_vaccinated, y = State.UTs, col = State.UTs, size = percent\_vaccinated)) + geom\_point()

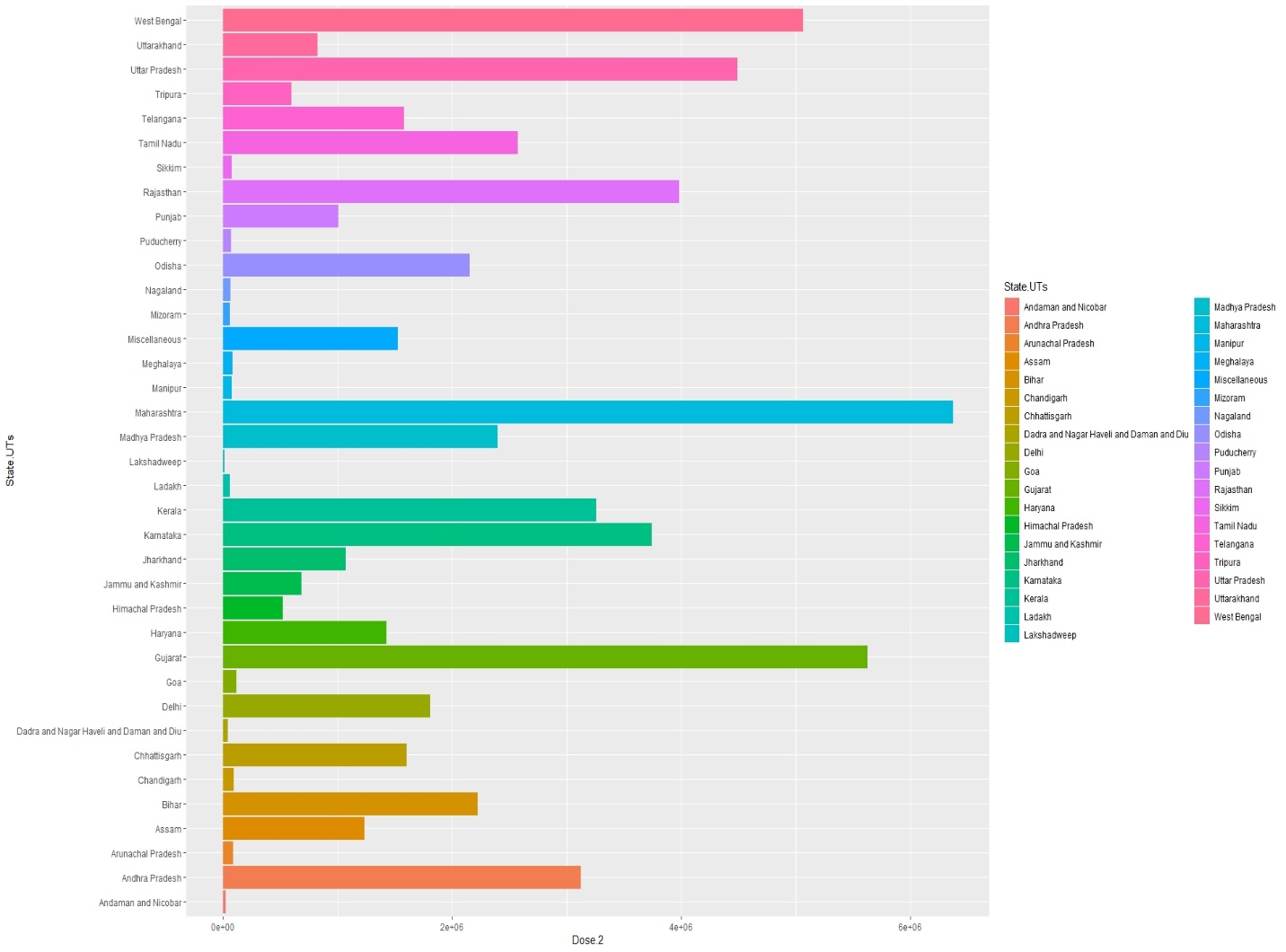


A different style of representation using bar graphs can also be used as shown below.

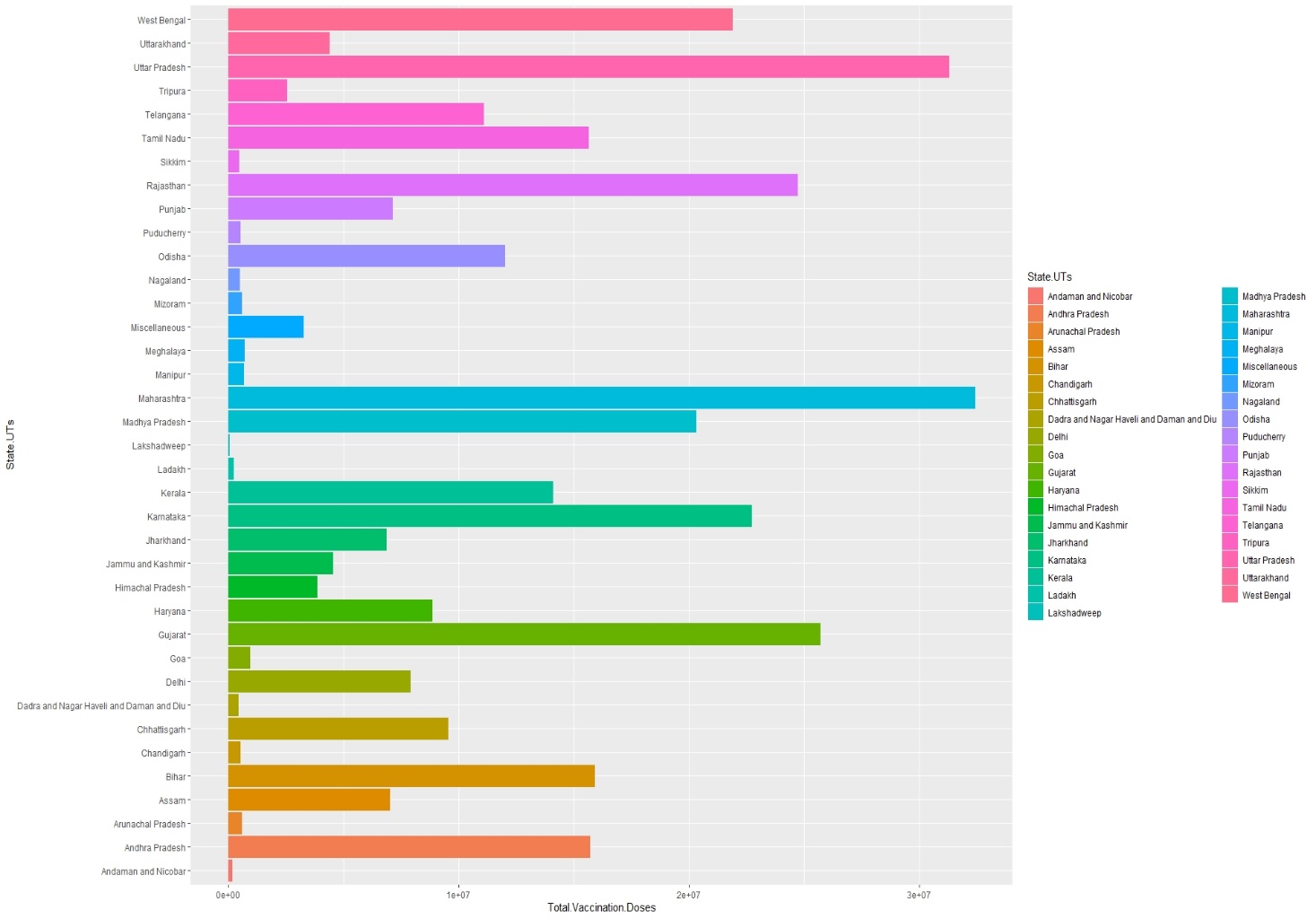
ggplot(f\_data, aes(x = Dose.1, y = State.UTs, col = State.UTs, fill = State.UTs)) + geom\_bar(stat = "identity")



ggplot(f\_data, aes(x = Dose.2, y = State.UTs, col = State.UTs, fill = State.UTs)) + geom\_bar(stat = "identity")



ggplot(f\_data, aes(x = Total.Vaccination.Doses, y = State.UTs, col = State.UTs, fill = State.UTs)) + geom\_bar(stat = "identity")



ggplot(f\_data, aes(x = percent\_vaccinated, y = State.UTs, col = State.UTs, fill = State.UTs)) + geom\_bar(stat = "identity")

