Programming language: R

Data: *cities.csv*

**INTRODUCTION**

The following has the exploratory data analysis and visualization of 500 top cities of India performed using R. Data source: Kaggle. Data wrangling and data analysis performed using dplyr, plyr; data visualization using ggplot.

**READ DATA**

cities<- read.csv(“C:/Career/Study/Data/Kaggle data/500 Indian Cities/cities\_r2”)

view(cities)

library(dplyr)

library(plyr)

library(ggplot2)

library(gridextra)

library(xlsx)

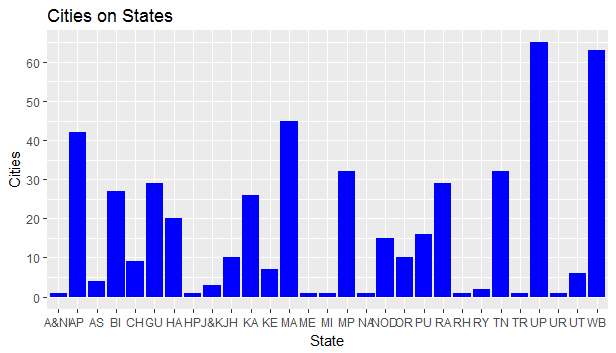
library(DT)

**EXPLORATORY DATA ANALYSIS**

by\_state <- cities %>% group\_by(state\_name) %>% summarise(Total = n(), Population = sum(population\_total), Male\_Population = sum(population\_male), Female\_Population = sum(population\_female), Male\_Percent = Male\_Population/Population \* 100, Female\_Percent = Female\_Population/Population \* 100, Graduates = sum(total\_graduates), Male\_Grads = sum(male\_graduates), Female\_Grads = sum(female\_graduates), Grads\_percent = Graduates/Population \* 100, Male\_Grads\_Percent = Male\_Grads/Male\_Population \* 100, Female\_Grads\_Percent = Female\_Grads/Female\_Population \* 100) %>% arrange(desc(Total))

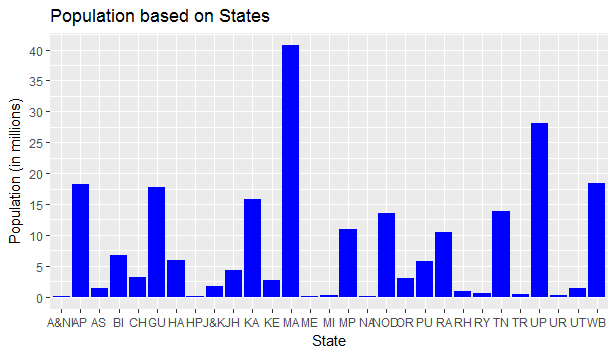
**Cities based on States**

ggplot(by\_state, aes(state,Total)) + geom\_bar(stat = "identity", fill= "blue") + scale\_y\_continuous(breaks= seq(0,45,5)) + ggtitle("Cities on States") + xlab("State") + ylab("Cities")

****

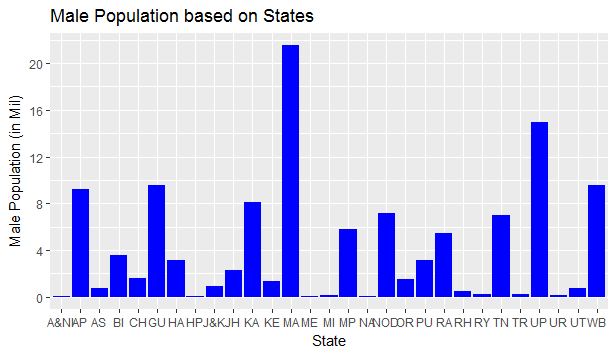
**Population based on States:**

viz1<- ggplot(by\_state, aes(state, population\_in\_mil)) + geom\_bar(stat = "identity", fill= "blue") + scale\_y\_continuous(breaks= seq(0,45,5)) + ggtitle("Population based on States") + xlab("State") + ylab("Population (in millions)")

****

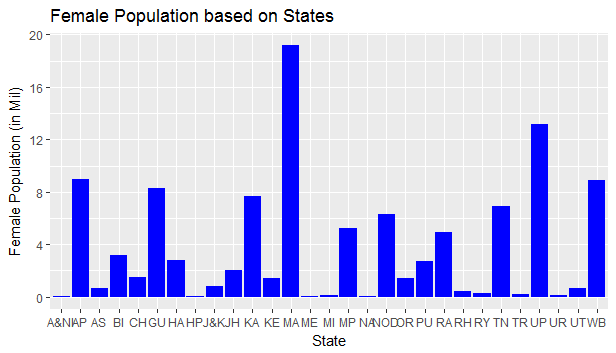
**Male Population based on states:**

viz3<- ggplot(by\_state, aes(state, Male\_Population\_in\_mil)) + geom\_bar(stat = "identity", fill= "blue") + scale\_y\_continuous(breaks= seq(0,22,4)) + ggtitle("Male Population based on States") + xlab("State") + ylab("Male Population (in Mil)")

****

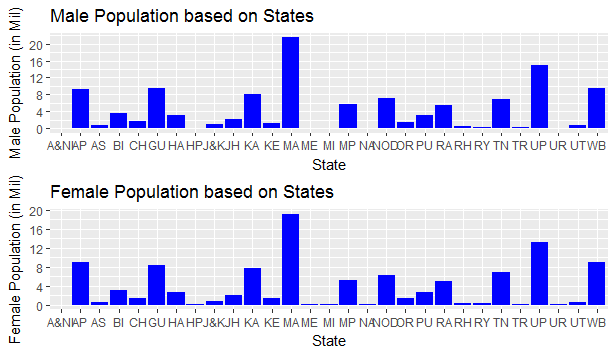
**Female Population based on states:**

viz4<- ggplot(by\_state, aes(state, Female\_Population\_in\_mil)) + geom\_bar(stat = "identity", fill= "blue") + scale\_y\_continuous(breaks= seq(0,20,4)) + ggtitle("Female Population based on States") + xlab("State") + ylab("Female Population (in Mil)")

****

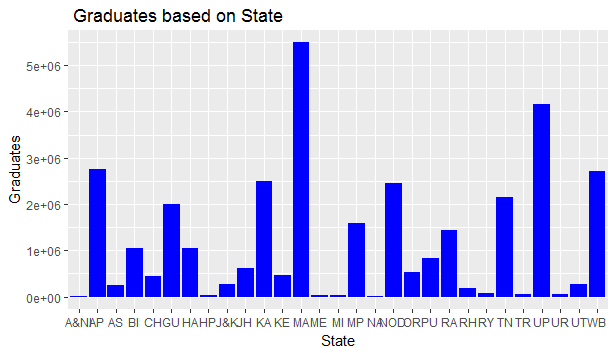
**Male vs Female Population**

grid.arrange(viz3, viz4)

****

**Graduates based on States**

ggplot(by\_state, aes(state, Graduates)) + geom\_bar(stat= "identity", fill= "blue") + scale\_y\_continuous(breaks= seq(0,5494343,1000000))+ xlab("State") + ylab("Graduates") + ggtitle(" Graduates based on State")

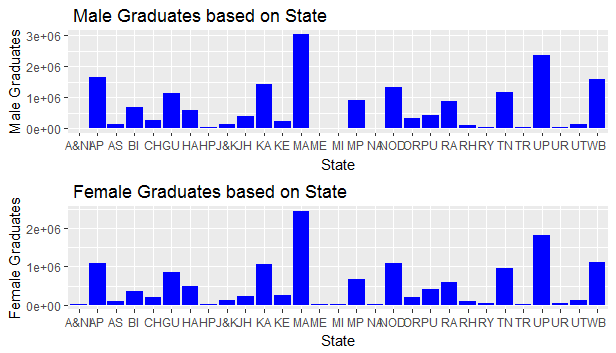
****

**Male Graduates based on States**

viz5<- ggplot(by\_state, aes(state, Male\_Grads)) + geom\_bar(stat= "identity", fill= "blue") + scale\_y\_continuous(breaks= seq(0,5494343,1000000))+ xlab("State") + ylab("Male Graduates") + ggtitle(" Male Graduates based on State")

viz6<- ggplot(by\_state, aes(state, Female\_Grads)) + geom\_bar(stat= "identity", fill= "blue") + scale\_y\_continuous(breaks= seq(0,5494343,1000000))+ xlab("State") + ylab("Female Graduates") + ggtitle(" Female Graduates based on State")

grid.arrange(viz5,viz6)

****