CAPSTONE PROJECT-02

General Objective:

- To explore and visualize demographic data (population, literacy rate, graduates) for cities in the state of Uttar Pradesh using R.

Why K-means clustering -

K-means clustering was selected due to its ability to effectively group cities based on key demographic attributes, such as effective literacy rate and total population. By focusing on these two features, K-means simplifies complex data into interpretable clusters, aiding in understanding regional demographic trends within Uttar Pradesh. This approach allows for the identification of cities with similar demographic profiles, facilitating targeted interventions and policy decisions related to education, healthcare, and urban development. Utilizing K-means clustering provides a practical framework for analyzing and visualizing demographic data, enabling policymakers to make informed decisions tailored to the unique needs of different city clusters.

Data Visualization code in R -

```
# Install and load necessary packages
install.packages(c("dplyr", "ggplot2", "cluster"))
library(dplyr)
library(ggplot2)
library(cluster)

# Read the CSV file into a data frame
df <- read.csv("states.csv")

# Subset the data for Uttar Pradesh
jkdf <- df[df$state_name == 'UTTAR PRADESH', ]

# Select relevant features for clustering
cluster_data <- select(jkdf, population_total, total_graduates, effective_literacy_rate_total)

# Standardize the data
cluster_data_scaled <- scale(cluster_data)

# Determine the optimal number of clusters using the elbow method
wss <- numeric(10)
for (i in 1:10) {
  km <- kmeans(cluster_data_scaled, centers = i, nstart = 10)
  wss[i] <- sum(km$withinss)
}
plot(1:10, wss, type = "b", xlab = "Number of Clusters", ylab = "Within-cluster Sum of Squares")
```

```r
# Based on the elbow method, select the optimal number of clusters
k_optimal <- 3  # Adjust as needed based on the plot

# Apply k-means clustering with the optimal number of clusters
km <- kmeans(cluster_data_scaled, centers = k_optimal, nstart = 10)

# Add cluster labels to the data
jkdf$cluster <- as.factor(km$cluster)

# Visualize the clusters
ggplot(jkdf, aes(x = population_total, y = effective_literacy_rate_total, color = cluster)) +
  geom_point() +
  labs(x = "Population Total", y = "Effective Literacy Rate Total", color = "Cluster") +
  theme_minimal()

  # Plotting

# Population
ggplot(jkdf, aes(x = name_of_city, y = population_total)) +
  geom_bar(stat = "identity", fill = "blue") +  # Create the bar graph
  labs(x = "City", y = "Population", title = "Population of Cities") +  # Add labels and title
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) # Rotate x-axis labels if needed


# Summarize population data for each city
city_summary <- aggregate(population_total ~ name_of_city, data = jkdf, FUN = sum)

# Create pie chart
ggplot(city_summary, aes(x = "", y = population_total, fill = name_of_city)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar(theta = "y") +
  labs(fill = "City") +
  theme_void() +
  theme(legend.position = "right") +
  guides(fill = guide_legend(title = "City"))

# Bubble plot
ggplot(jkdf, aes(x = name_of_city, y = population_total, size = total_graduates)) +
  geom_point(color = "green", alpha = 0.6) +
  labs(x = "City", y = "Population", title = "Population of Cities wrt Graduates (Bubble Plot)") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

ggplot(jkdf, aes(x = name_of_city, y = population_total, size = effective_literacy_rate_total)) +
  geom_point(color = "brown", alpha = 0.6) +
  labs(x = "City", y = "Population", title = "Population of Cities wrt Literacy Rates (Bubble Plot)") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

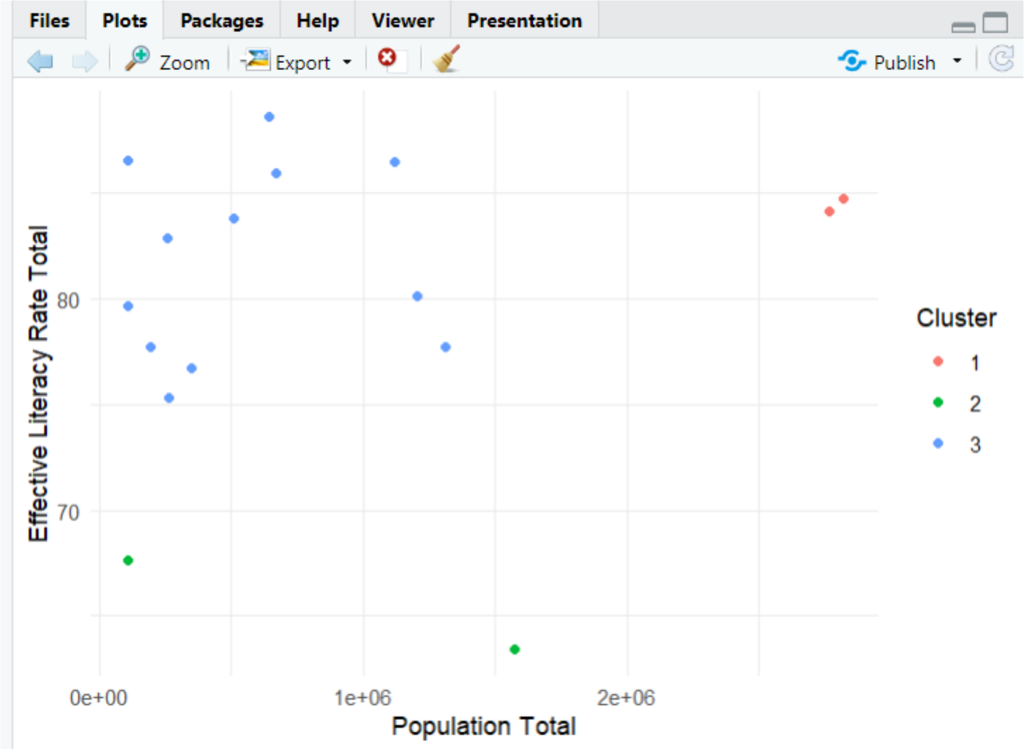
```
# Create the circular plot
ggplot(jkdf, aes(x = name_of_city, y = population_total, fill = name_of_city)) +
  geom_bar(stat = "identity") +
  coord_polar(theta = "y") +
  labs(x = NULL, y = NULL, title = "Population of Cities (Circular Barplot)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5))

# Male Population

ggplot(jkdf, aes(x = name_of_city, y = population_male)) +
  geom_bar(stat = "identity", fill = "skyblue") +  # Create the bar graph
  labs(x = "City", y = "Population", title = "Population of Cities") +  # Add labels and title
  theme(axis.text.x = element_text(angle = 90, hjust = 1))  # Rotate x-axis labels if needed

# Summarize population data for each city
city_summary <- aggregate(population_male ~ name_of_city, data = jkdf, FUN = sum)

# Create pie chart
ggplot(city_summary, aes(x = "", y = population_male, fill = name_of_city)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar(theta = "y") +
  labs(fill = "City") +
  theme_void() +
  theme(legend.position = "right") +
  guides(fill = guide_legend(title = "City"))

# Female Population

ggplot(jkdf, aes(x = name_of_city, y = population_female)) +
  geom_bar(stat = "identity", fill = "pink") +  # Create the bar graph
  labs(x = "City", y = "Population", title = "Population of Cities") +  # Add labels and title
  theme(axis.text.x = element_text(angle = 90, hjust = 1))  # Rotate x-axis labels if needed

# Summarize population data for each city
city_summary <- aggregate(population_female ~ name_of_city, data = jkdf, FUN = sum)

# Create pie chart
ggplot(city_summary, aes(x = "", y = population_female, fill = name_of_city)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar(theta = "y") +
  labs(fill = "City") +
  theme_void() +
  theme(legend.position = "right") +
```

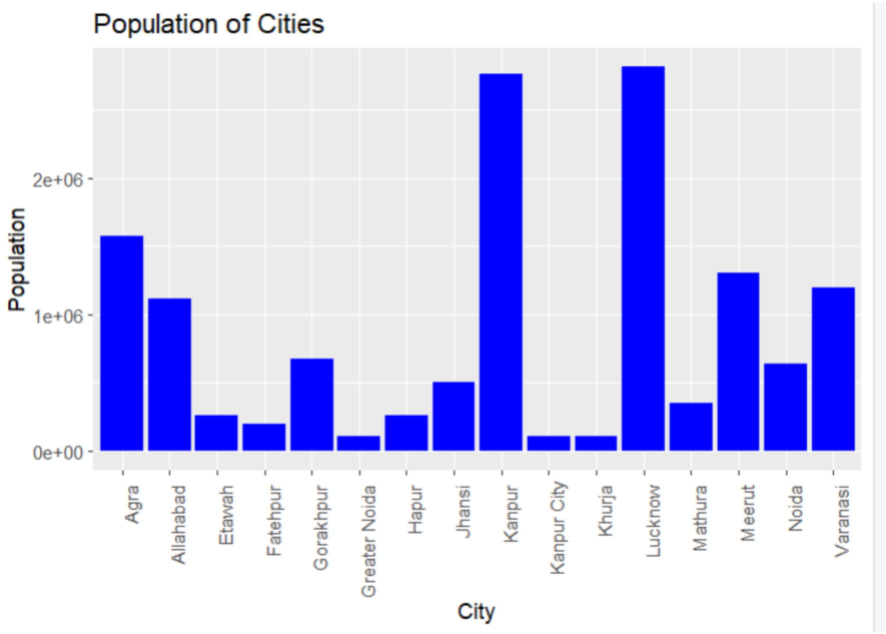
```r
  guides(fill = guide_legend(title = "City"))

# Male vs Female population

males <- sum(jkdf$population_male)
females <- sum(jkdf$population_female)

# Bar graph
maleFemaledf <- data.frame(Gender = c("Females", "Males"),
                           Population = c(females, males))

# Plot for bar graph
ggplot(maleFemaledf, aes(x = Gender, y = Population, fill = Gender)) +
  geom_bar(stat = "identity") +
  labs(x = "Gender", y = "Population", title = "Population by Gender (Bar Graph)")

# Create pie chart
ggplot(maleFemaledf, aes(x = "", y = Population, fill = Gender)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar(theta = "y") +
  labs(fill = "City") +
  theme_void() +
  theme(legend.position = "right") +
  guides(fill = guide_legend(title = "City"))

# literacy

ggplot(jkdf, aes(x = name_of_city, y = effective_literacy_rate_total)) +
  geom_bar(stat = "identity", fill = "red") +  # Create the bar graph
  labs(x = "City", y = "Population", title = "Population of Cities") +  # Add labels and title
  theme(axis.text.x = element_text(angle = 90, hjust = 1))  # Rotate x-axis labels if needed

# Male Population

ggplot(jkdf, aes(x = name_of_city, y = effective_literacy_rate_male)) +
  geom_bar(stat = "identity", fill = "blue") +  # Create the bar graph
  labs(x = "City", y = "Literacy Rate", title = "Population of Cities") +  # Add labels and title
  theme(axis.text.x = element_text(angle = 90, hjust = 1))  # Rotate x-axis labels if needed

# Female Population

ggplot(jkdf, aes(x = name_of_city, y = effective_literacy_rate_female)) +
  geom_bar(stat = "identity", fill = "pink") +  # Create the bar graph
  labs(x = "City", y = "Literacy Rate", title = "Literacy rates") +  # Add labels and title
  theme(axis.text.x = element_text(angle = 90, hjust = 1))  # Rotate x-axis labels if needed
```

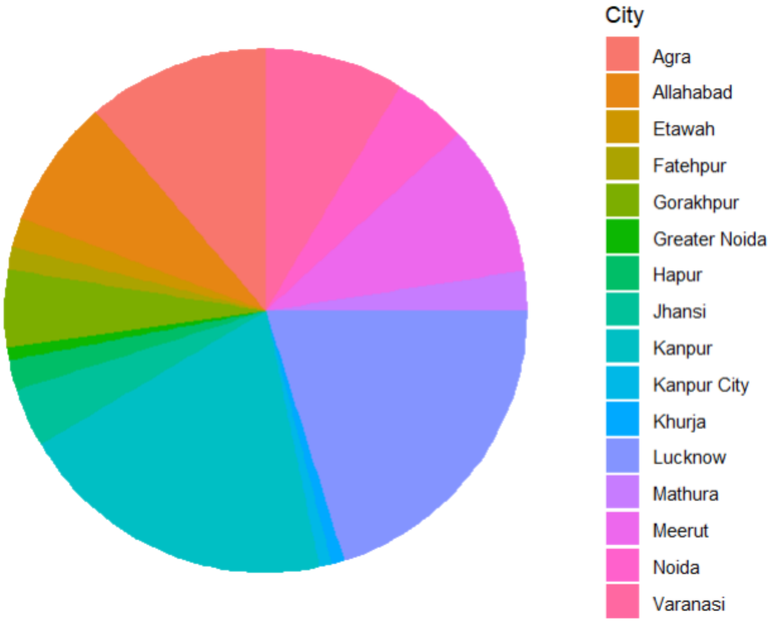
```
# Male vs Female population

males <- sum(jkdf$effective_literacy_rate_male)/nrow(jkdf)
females <- sum(jkdf$effective_literacy_rate_female)/nrow(jkdf)

# Bar graph
maleFemaledf <- data.frame(Gender = c("Females", "Males"),
                           effective_literacy_rate = c(females, males))

# Plot for bar graph
ggplot(maleFemaledf, aes(x = Gender, y = effective_literacy_rate, fill = Gender)) +
  geom_bar(stat = "identity") +
  labs(x = "Gender", y = "Literacy Rate", title = "Literates by Gender (Bar Graph)")


# Education

ggplot(jkdf, aes(x = name_of_city, y = total_graduates)) +
  geom_bar(stat = "identity", fill = "red") +  # Create the bar graph
  labs(x = "City", y = "Education", title = "Education of Cities") +  # Add labels and title
  theme(axis.text.x = element_text(angle = 90, hjust = 1))  # Rotate x-axis labels if needed

# Bubble plot
ggplot(jkdf, aes(x = name_of_city, y = total_graduates, size = total_graduates)) +
  geom_point(color = "red", alpha = 0.6) +
  labs(x = "City", y = "Population", title = "Education of Cities (Bubble Plot)") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Male Population

ggplot(jkdf, aes(x = name_of_city, y = male_graduates)) +
  geom_bar(stat = "identity", fill = "skyblue") +  # Create the bar graph
  labs(x = "City", y = "Graduates", title = "Graduates in Cities") +  # Add labels and title
  theme(axis.text.x = element_text(angle = 90, hjust = 1))  # Rotate x-axis labels if needed

# Summarize population data for each city
city_summary <- aggregate(male_graduates ~ name_of_city, data = jkdf, FUN = sum)

# Female Population

ggplot(jkdf, aes(x = name_of_city, y = female_graduates)) +
  geom_bar(stat = "identity", fill = "pink") +  # Create the bar graph
  labs(x = "City", y = "Graduates", title = "Female Graduates in Cities") +  # Add labels and title
  theme(axis.text.x = element_text(angle = 90, hjust = 1))  # Rotate x-axis labels if needed

# Summarize population data for each city
city_summary <- aggregate(female_graduates ~ name_of_city, data = jkdf, FUN = sum)
```

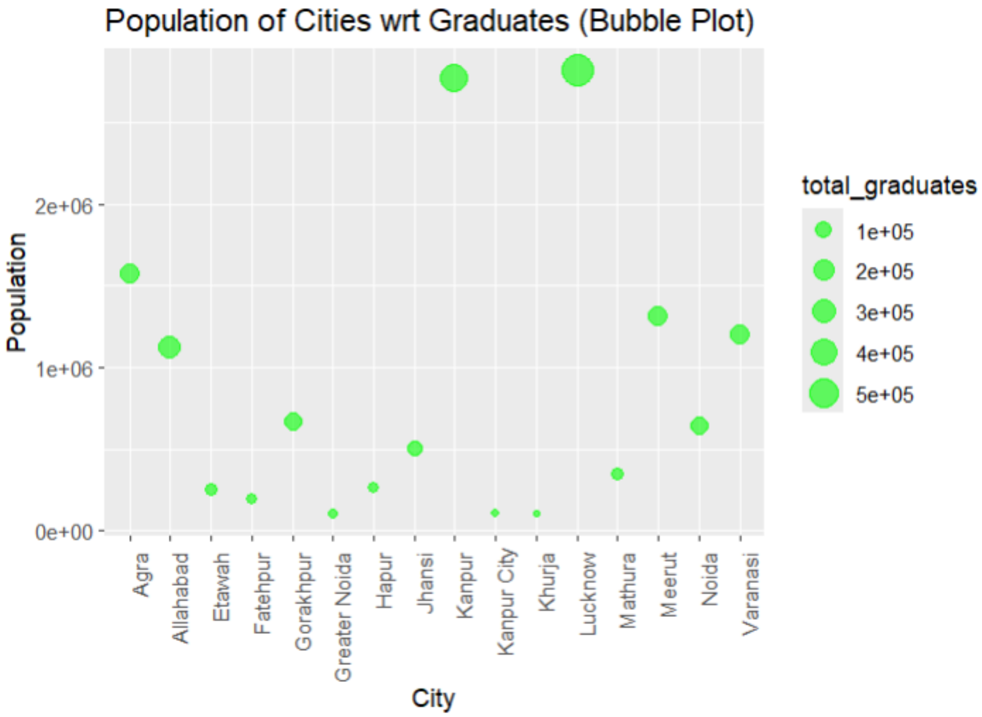
```r
# Create pie chart
ggplot(city_summary, aes(x = "", y = female_graduates, fill = name_of_city)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar(theta = "y") +
  labs(fill = "City") +
  theme_void() +
  theme(legend.position = "right") +
  guides(fill = guide_legend(title = "City"))

# Male vs Female population

males <- sum(jkdf$male_graduates)
females <- sum(jkdf$female_graduates)

# Bar graph
maleFemaledf <- data.frame(Gender = c("Females", "Males"),
                           Population = c(females, males))

# Plot for bar graph
ggplot(maleFemaledf, aes(x = Gender, y = Population, fill = Gender)) +
  geom_bar(stat = "identity") +
  labs(x = "Gender", y = "Graduates", title = "Graduates by Gender (Bar Graph)")

library(ggplot2)
library(forecast)
```

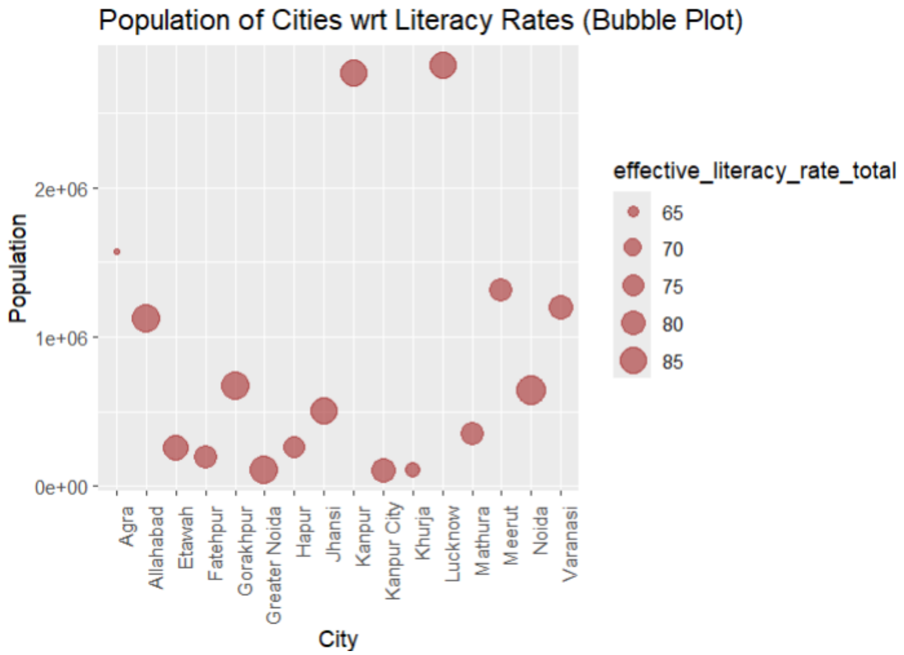
Graphs Plotted-



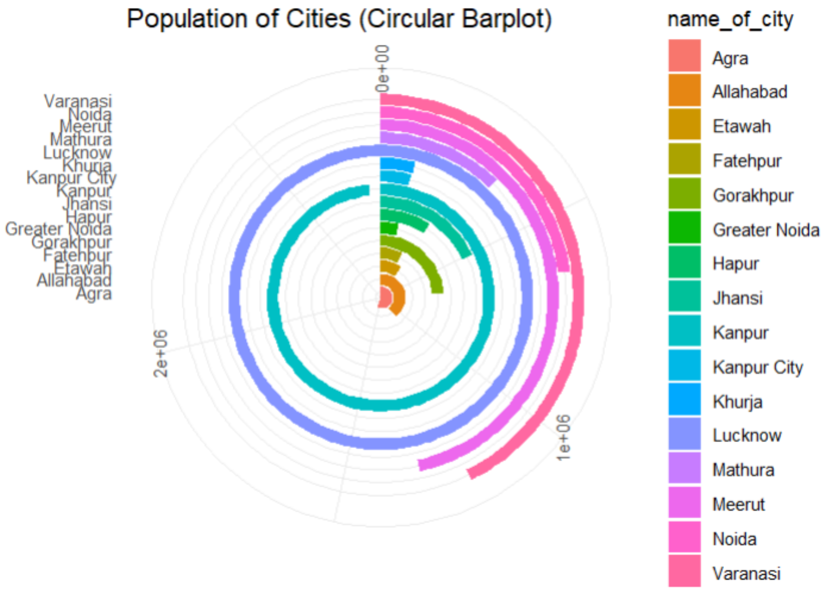



City
Agra
Allahabad
Etawah
Fatehpur
Gorakhpur
Greater Noida
Hapur
Jhansi
Kanpur
Kanpur City
Khurja
Lucknow
Mathura
Meerut
Noida
Varanasi

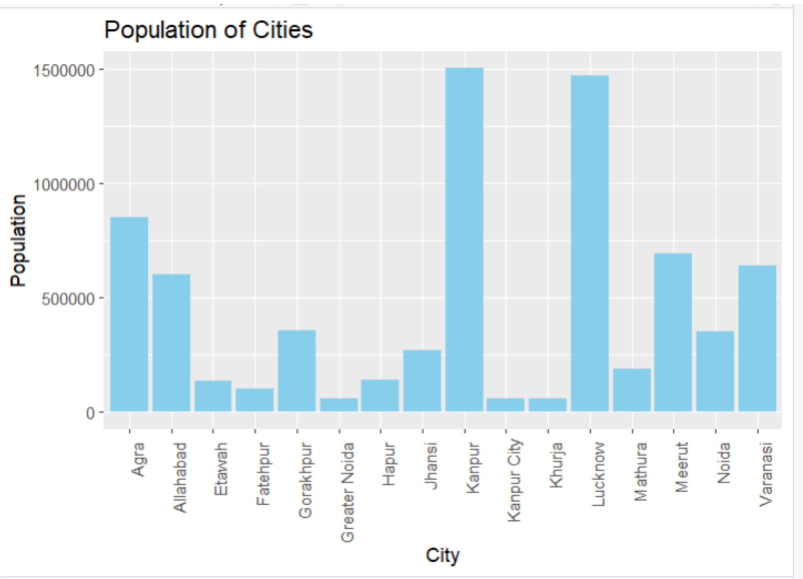

Population of Cities wrt Graduates (Bubble Plot)
Population
2e+06
1e+06
0e+00
Agra
Allahabad
Etawah
Fatehpur
Gorakhpur
Greater Noida
Hapur
Jhansi
Kanpur
Kanpur City
Khurja
Lucknow
Mathura
Meerut
Noida
Varanasi
City
total_graduates
1e+05
2e+05
3e+05
4e+05
5e+05

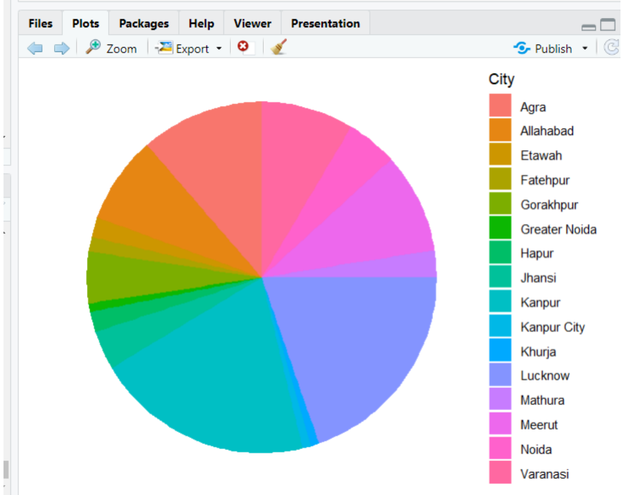


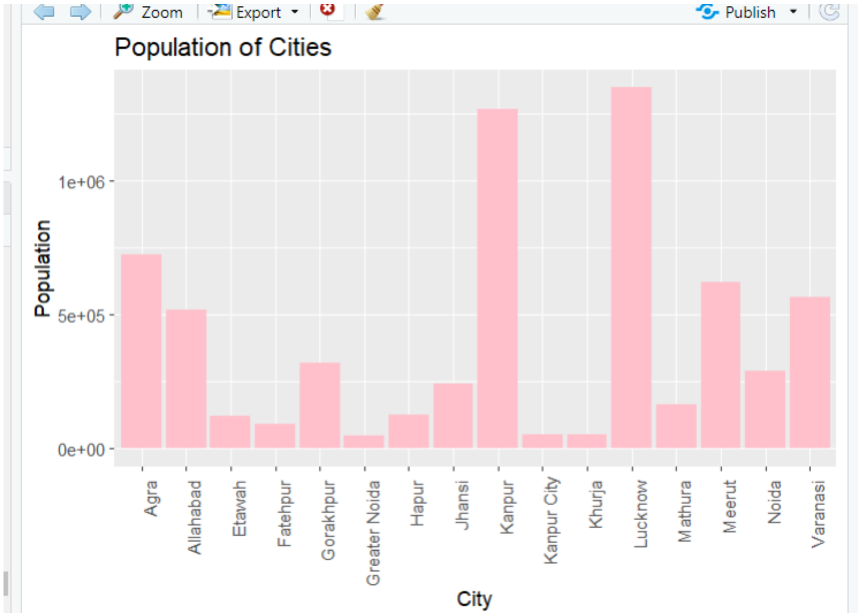
Population of Cities wrt Literacy Rates (Bubble Plot)

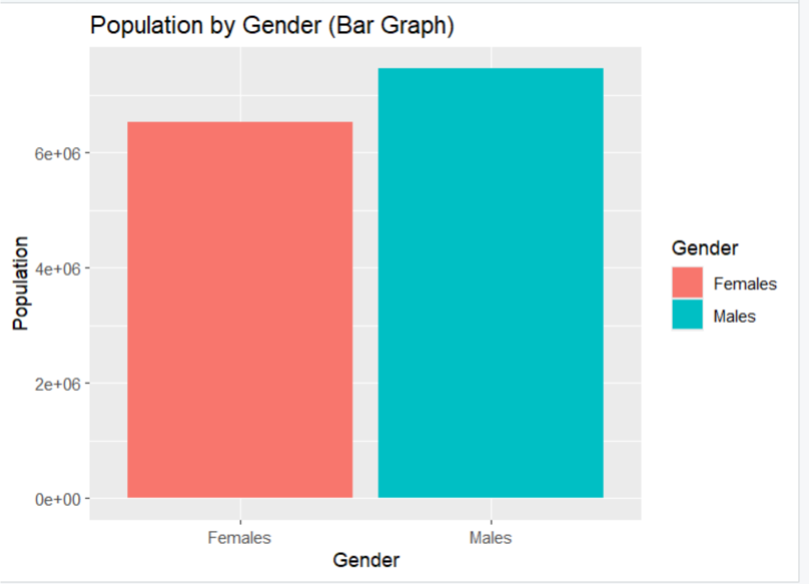


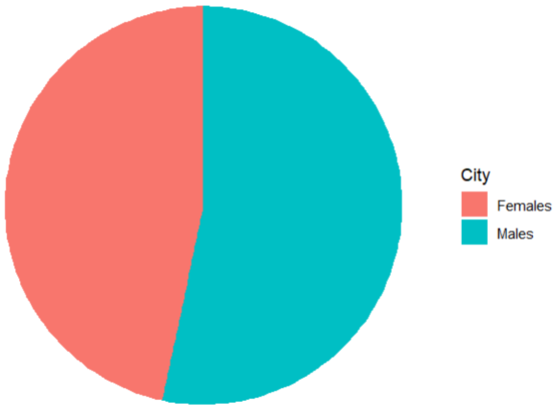
Population of Cities (Circular Barplot)

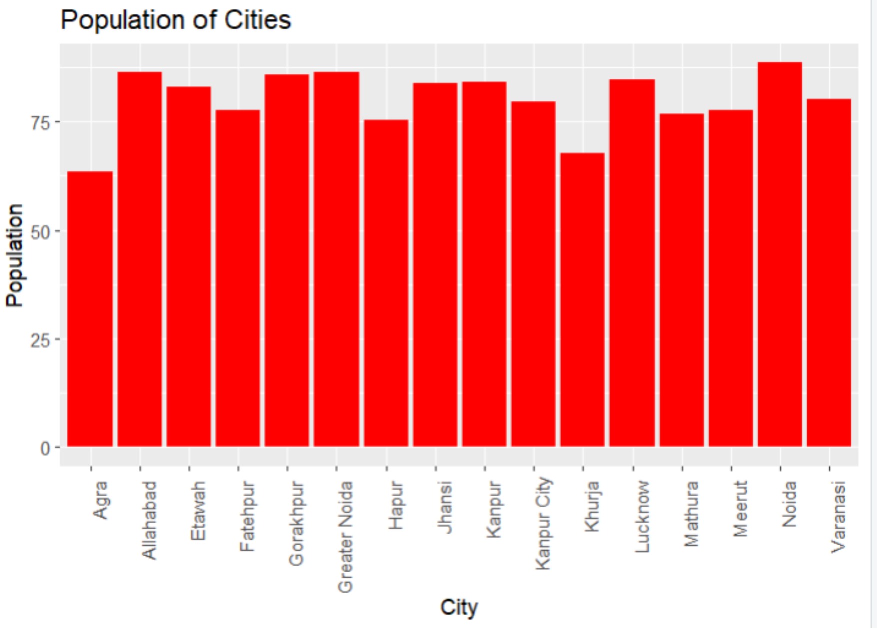

Population of Cities
1500000
1000000
500000
0
Population
Agra
Allahabad
Etawah
Fatehpur
Gorakhpur
Greater Noida
Hapur
Jhansi
Kanpur
Kanpur City
Khurja
Lucknow
Mathura
Meerut
Noida
Varanasi
City

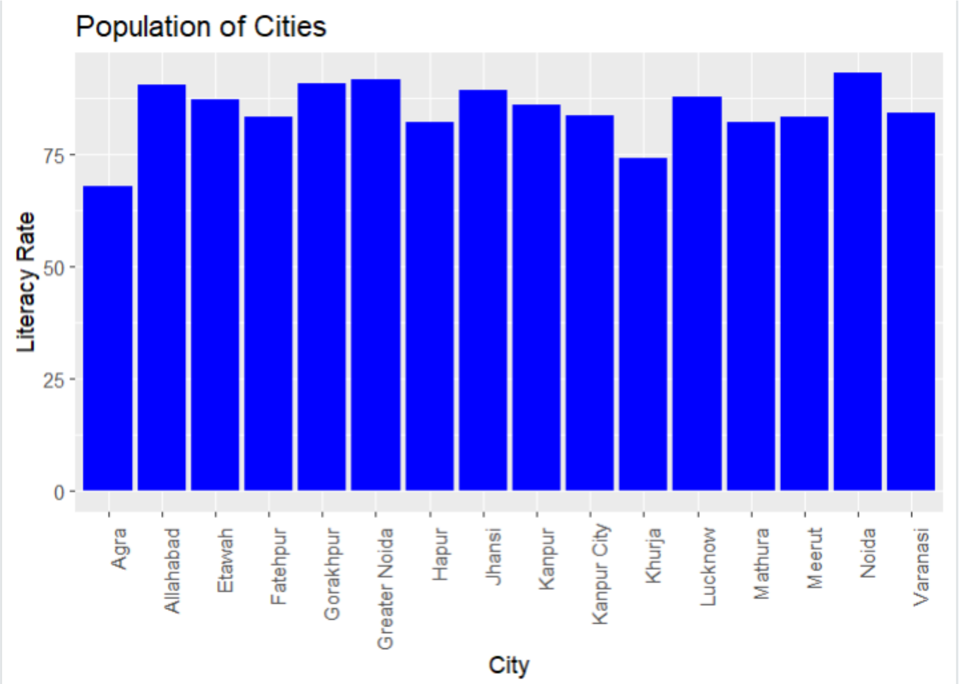

Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
City
Agra
Allahabad
Etawah
Fatehpur
Gorakhpur
Greater Noida
Hapur
Jhansi
Kanpur
Kanpur City
Khurja
Lucknow
Mathura
Meerut
Noida
Varanasi

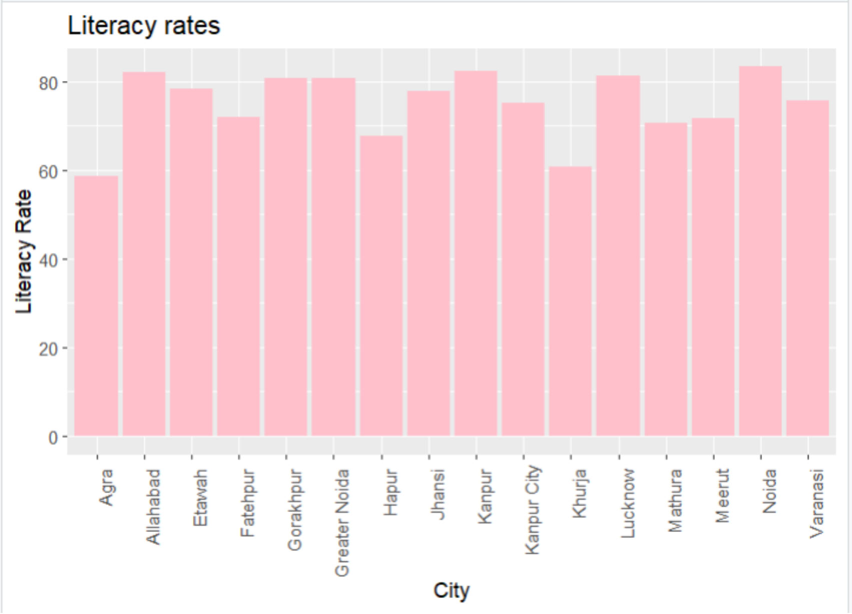

Zoom
Export
Publish
Population of Cities
1e+06
5e+05
0e+00
Population
Agra
Allahabad
Etawah
Fatehpur
Gorakhpur
Greater Noida
Hapur
Jhansi
Kanpur
Kanpur City
Khurja
Lucknow
Mathura
Meerut
Noida
Varanasi
City

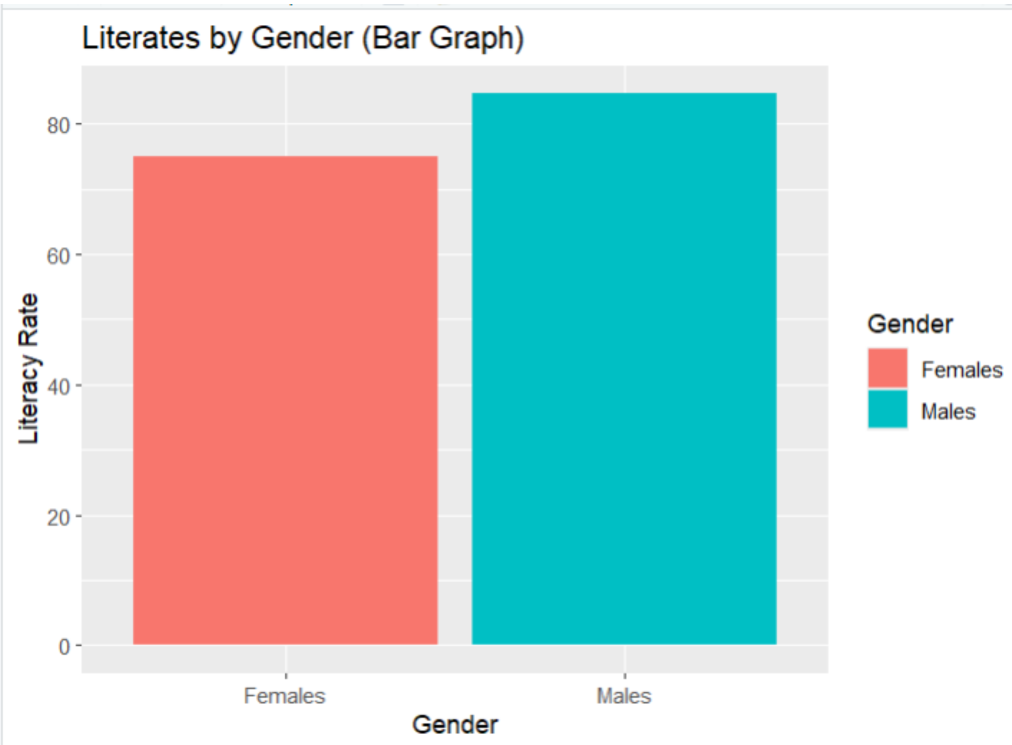

Population by Gender (Bar Graph)
Population
6e+06
4e+06
2e+06
0e+00
Females
Males
Gender
Gender
Females
Males

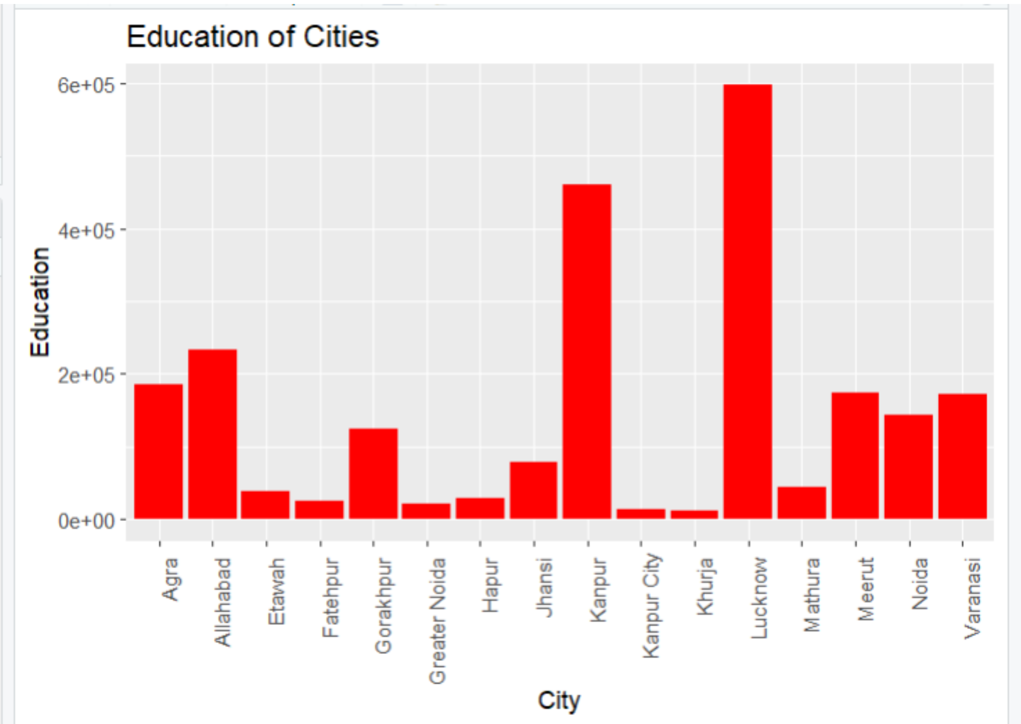

City
Females
Males

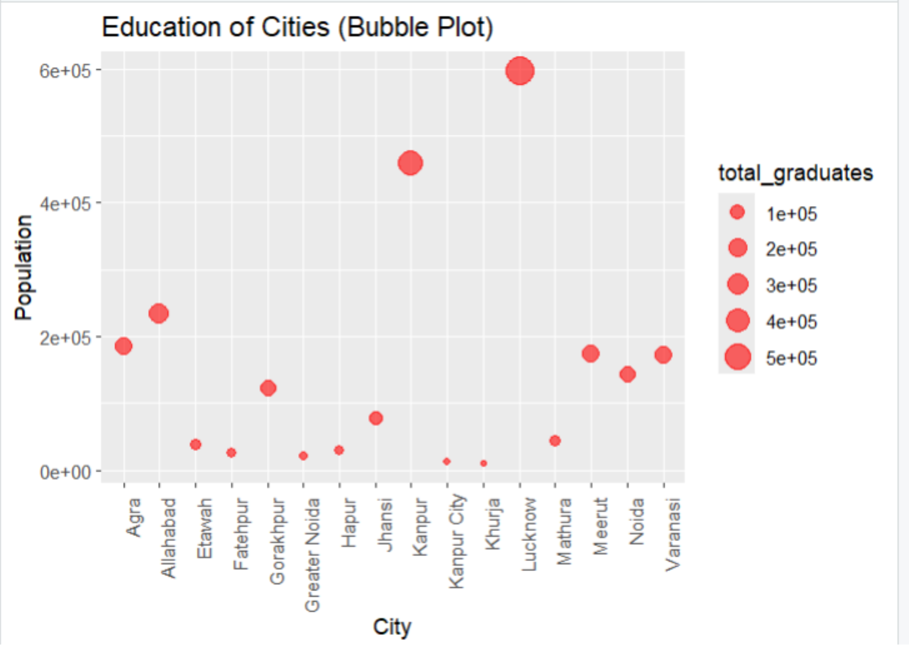

Population of Cities
Population
75
50
25
0
Agra
Allahabad
Etawah
Fatehpur
Gorakhpur
Greater Noida
Hapur
Jhansi
Kanpur
Kanpur City
Khurja
Lucknow
Mathura
Meerut
Noida
Varanasi
City

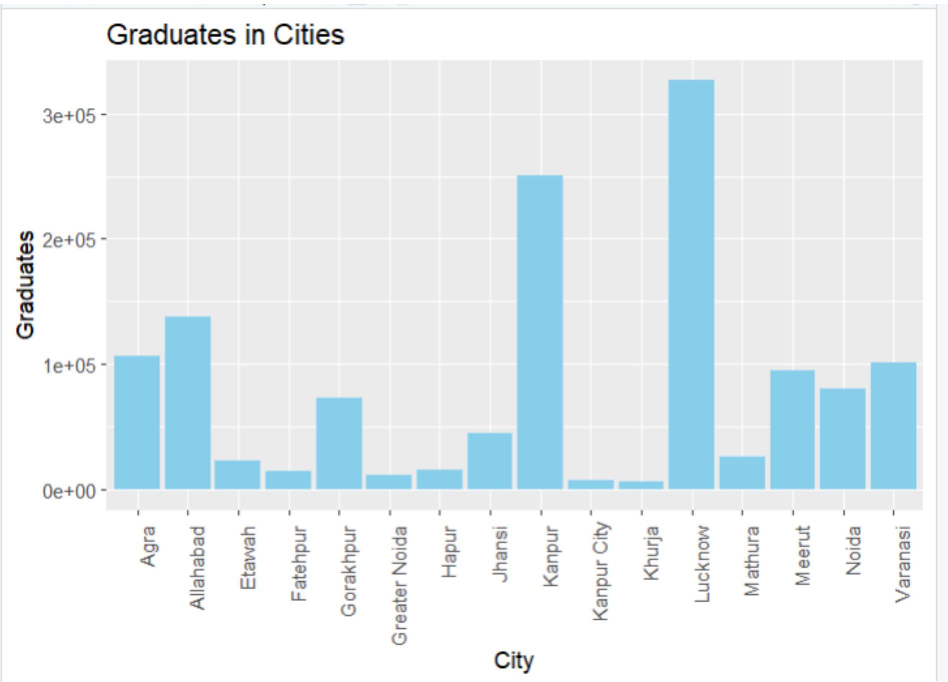

Population of Cities
Literacy Rate
0
25
50
75
Agra
Allahabad
Etawah
Fatehpur
Gorakhpur
Greater Noida
Hapur
Jhansi
Kanpur
Kanpur City
Khurja
Lucknow
Mathura
Meerut
Noida
Varanasi
City

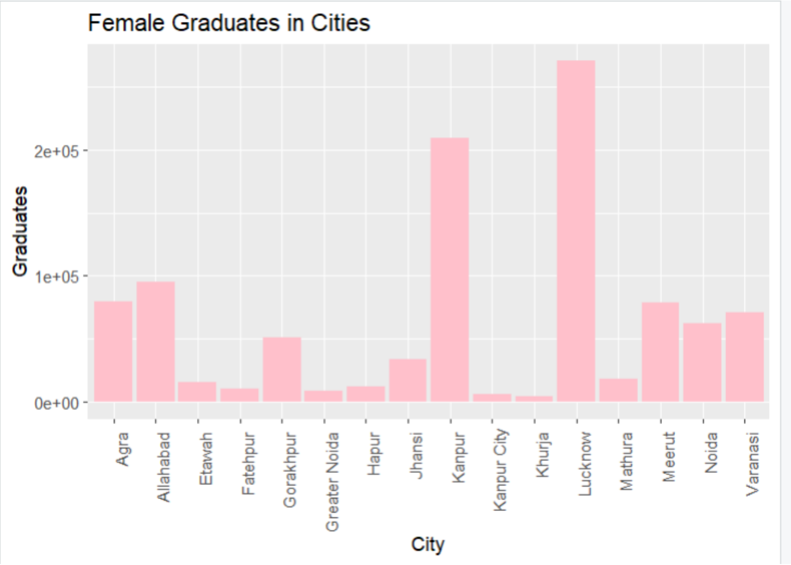

Literacy rates
Literacy Rate
0
20
40
60
80
Agra
Allahabad
Etawah
Fatehpur
Gorakhpur
Greater Noida
Hapur
Jhansi
Kanpur
Kanpur City
Khurja
Lucknow
Mathura
Meerut
Noida
Varanasi
City

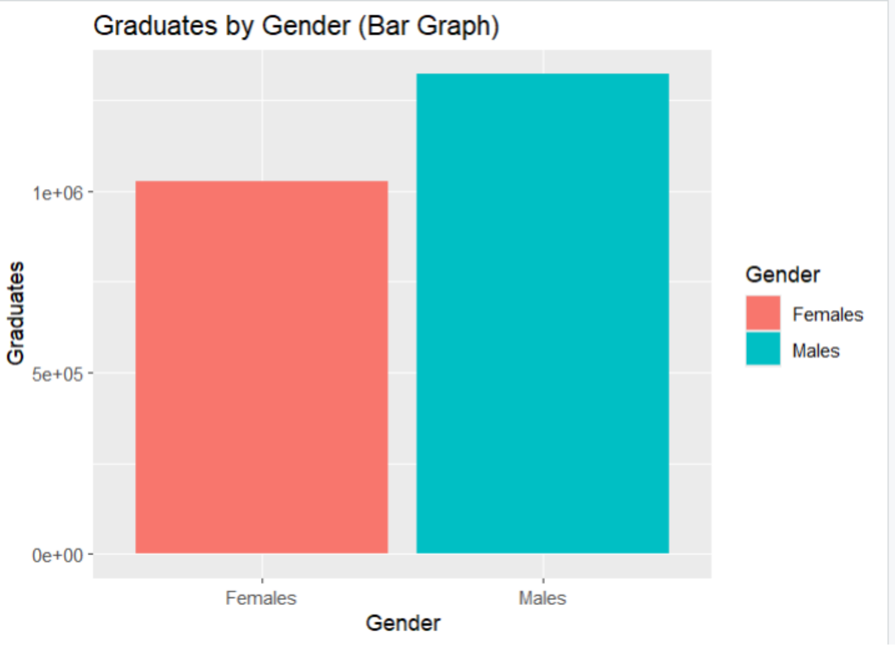

Literates by Gender (Bar Graph)
Literacy Rate
0
20
40
60
80
Females
Males
Gender
Gender
Females
Males


Education of Cities
Education
0e+00
2e+05
4e+05
6e+05
Agra
Allahabad
Etawah
Fatehpur
Gorakhpur
Greater Noida
Hapur
Jhansi
Kanpur
Kanpur City
Khurja
Lucknow
Mathura
Meerut
Noida
Varanasi
City


Education of Cities (Bubble Plot)
Population
6e+05
4e+05
2e+05
0e+00
Agra
Allahabad
Etawah
Fatehpur
Gorakhpur
Greater Noida
Hapur
Jhansi
Kanpur
Kanpur City
Khurja
Lucknow
Mathura
Meerut
Noida
Varanasi
City
total_graduates
1e+05
2e+05
3e+05
4e+05
5e+05


Graduates in Cities
Graduates
3e+05
2e+05
1e+05
0e+00
Agra
Allahabad
Etawah
Fatehpur
Gorakhpur
Greater Noida
Hapur
Jhansi
Kanpur
Kanpur City
Khurja
Lucknow
Mathura
Meerut
Noida
Varanasi
City


Female Graduates in Cities
Graduates
2e+05
1e+05
0e+00
Agra
Allahabad
Etawah
Fatehpur
Gorakhpur
Greater Noida
Hapur
Jhansi
Kanpur
Kanpur City
Khurja
Lucknow
Mathura
Meerut
Noida
Varanasi
City


Graduates by Gender (Bar Graph)
Graduates
1e+06
5e+05
0e+00
Females
Males
Gender
Gender
Females
Males