Report on Exploratory Data Analysis

Masters in Computer Science (2020-2022),

Department of Computer Science, University of Delhi

Mathematical foundation of Computer Science

Submitted To: Dr. Vasudha Bhatnagar

Submitted By:

- Payal Saha (200008@cs.du.ac.in)
- Tushar Harsh (200013@cs.du.ac.in)
- Riya Swami (200039@cs.du.ac.in)

Table of Contents

Description of data set

Data Set: Statewise_General_Index_july2019_20Aug2020

Source: data.gov.in

Released Under: National Data Sharing and Accessibility Policy (NDSAP)

Contributor: Ministry of Statistics and Programme Implementation

Description: Consumer Price Indices (CPI) measure changes over time in general level of prices of goods and services that households acquire for the purpose of consumption. CPI numbers are widely used as a macroeconomic indicator of inflation, as a tool by governments and central banks for inflation targeting and for monitoring price stability, and as deflators in the national accounts. CPI is also used for indexing dearness allowance to employees for increase in prices. CPI is therefore considered as one of the most important economic indicators. For construction of CPI numbers, two requisite components are weighting diagrams (consumption patterns) and price data collected at regular intervals. The Central Statistics Office (CSO), Ministry of Statistics and Programme Implementation releases Consumer Price Indices (CPI) on base 2010=100 for all-India and States/UTs separately for rural, urban and combined every month with effect from January, 2011. The data is Published by Central Statistical Office and released on 12th of every month.

Attributes:

- **Sector:** Rural/ Urban/ Rural+Urban
- **Year:** 2011 - 2020
- **Name:** name of the month
- **States :** CPI of each state(29) and union territory(7)

NA values: Few attributes have null value like Telangana was separated in 2015 prior to that all values are null, no urban data is available for Arunachal Pradesh.

Understanding Consumer Price Index (CPI)

The Consumer Price Index (CPI) is a measure that examines the weighted average of prices of a basket of consumer goods and services, such as transportation, food, and medical care. It is calculated by taking price changes for each item in the predetermined basket of goods and averaging them.

Changes in the CPI are used to assess price changes associated with the cost of living. The CPI is one of the most frequently used statistics for identifying periods of inflation or deflation.

CPI is an economic indicator. It is the most widely used measure of inflation and, by proxy, of the effectiveness of the government's economic policy. The CPI gives the government, businesses, and citizens an idea about prices changes in the economy, and can act as a guide in order to make informed decisions about the economy.

The index can also be used to adjust people's eligibility levels for certain types of government assistance including Social Security and it automatically provides the cost-of-living wage adjustments to domestic workers.

Calculating CPI

The formula used to calculate the Consumer Price Index for a single item is as follows:

$$CPI_t = \frac{C_t}{C_0} * 100$$

CPI_t = consumer price index in current period

C_t = cost of market basket in current period

C_0 = cost of market basket in base period

Formula for calculating inflation

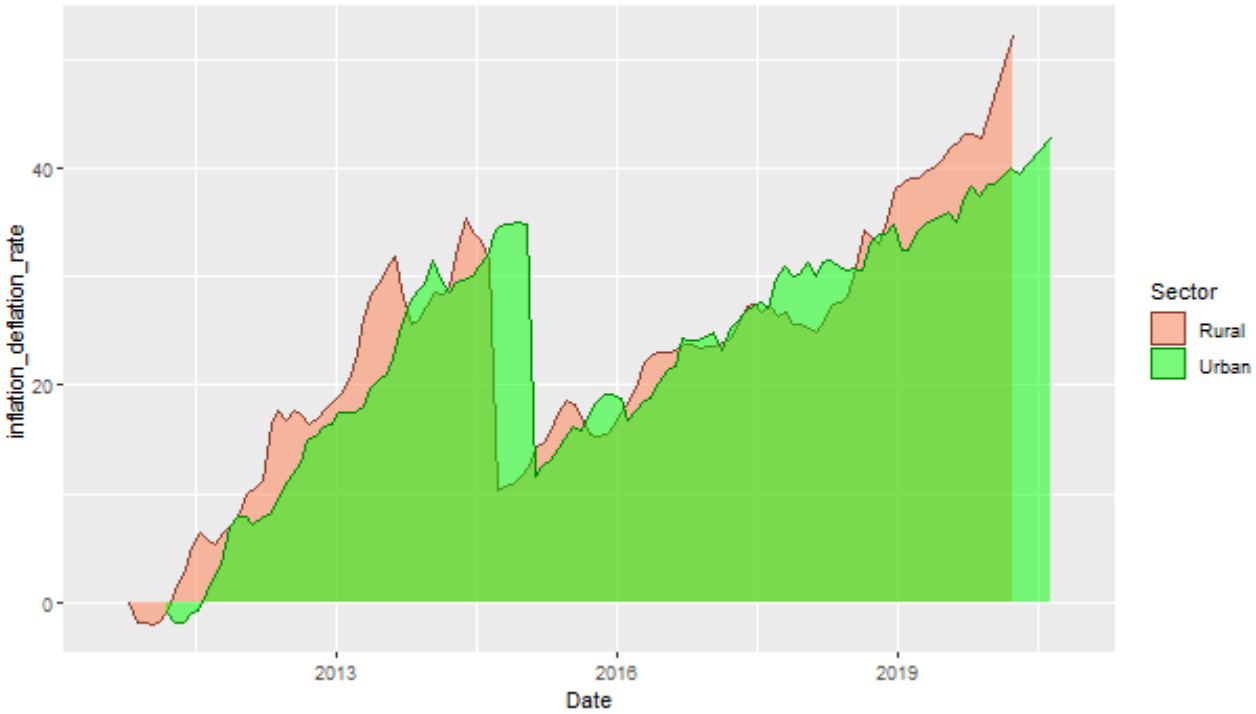
((Current CPI – Past CPI) ÷ Past CPI) x 100 = Inflation Rate or

((B – A)/A) x 100 = Inflation Rate

AREA PLOT

```r
#LIBRARY USED
library(ggplot2)
library(astsa)

data <- read.csv("Statewise_General_Index_july2019_20Aug2020.csv")
print(data)
df_r <- subset(data, Sector=="Rural" | Sector== "Urban")[,c('Sector','Year','Name','Delhi')]
df_r <- df_r[complete.cases(df_r), ]
df_r$Date <- zoo::as.yearmon(paste(df_r$Year, df_r$Name), "%Y %B")

df_r$inflation_deflation_rate <- (df_r$Delhi - 105) / 105 * 100

# Area plot
png("Inflation_Area_Plot.png", width = 612, height = 363)
ggplot( data=df_r, aes(x=Date, y=inflation_deflation_rate))  +
  geom_area(aes(color =  Sector, fill = Sector),
            alpha = 0.5, position = position_dodge(0.8)) +
  scale_color_manual(values = c("coral4", "green4")) +
  scale_fill_manual(values = c("coral", "green"))
dev.off()
```

OUTPUT



INTERPRETATION

The graph shows the inflation and deflation rates over a period of 10 years in Delhi. The plot used is Area Plot which is a variant of Time Series Plot. The inflation rate is calculated using the formula given above.

Inflation rate < 0 signifies Deflation rate.

The Inflation rate of both rural and urban area of Delhi is shown from Jan 2011 to June 2020.

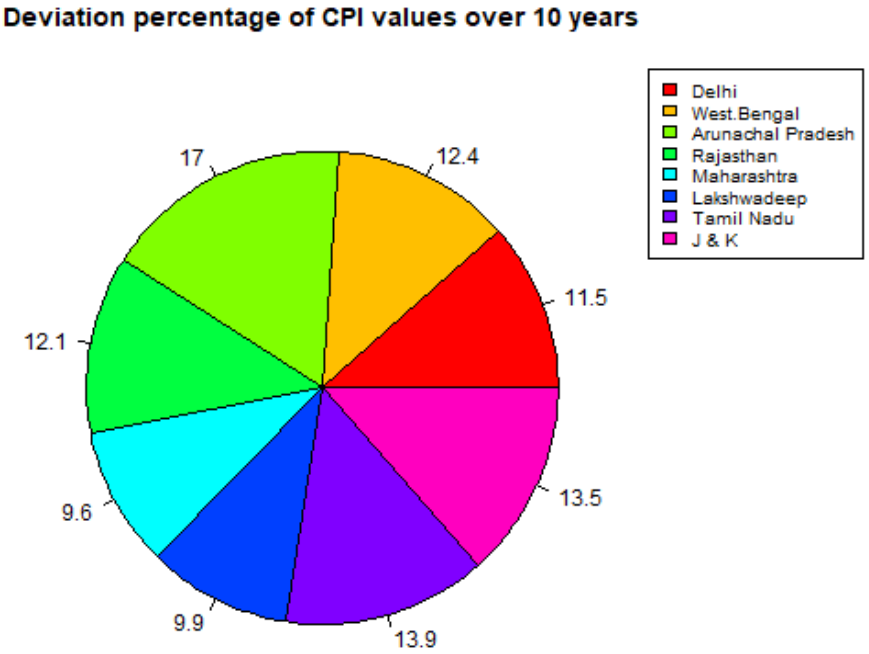
From the graph it can be observed that there occurred a short period of deflation from Jan 2011 to June 2011 in Delhi. It can also be observed that the Inflation rate of rural areas in Delhi increased in a more rapid rate than that of Delhi urban areas from 2019 onwards.

To access the full code refer to the link: code

PIE CHART

```r
#LIBRARY USED
library(matrixStats)

data <- read.csv("Statewise_General_Index_july2019_20Aug2020.csv")

df <- subset(data, Sector=="Rural+Urban")[,c('Delhi','West.Bengal', 'Arunachal.Pradesh',
                                            'Rajasthan','Maharashtra','Lakshwadeep',
                                            'Tamil.Nadu','Jammu.and.Kashmir')]
df <- df[complete.cases(df), ]
mat <- as.matrix(df)
var <- colVars(mat, suma = NULL, std = TRUE)
per <- round(var/sum(var) * 100,1)

# Pie Chart
png("variance_Pie_Chart.png", width = 680, height = 480)
labels <- c('Delhi','West.Bengal', 'Arunachal Pradesh', 'Rajasthan','Maharashtra',
            'Lakshwadeep','Tamil Nadu','J & K')
pie(var,labels= per, main = "Deviation percentage of CPI values over 10 years", col = rainbow(8))
legend("topright", labels, cex = 0.8, fill = rainbow(8))
dev.off()
```

OUTPUT



INTERPRETATION

The graph shows the percentage of deviation of the CPI index values in 8 states – Delhi, West Bengal, Arunachal Pradesh, Rajasthan, Maharashtra, Tamil Nadu, Lakshadweep and Jammu & Kashmir. This graph signifies the fluctuations in CPI indices i.e. the prices of goods and services over the last 10 years in 8 states (8 major market states) of India. These fluctuation patterns can be used for price fluctuation risk of the price index system in India.

It can be seen from the graph that the fluctuation is the largest in Arunachal Pradesh over the past 10 years.

To access the full code refer to the link: code

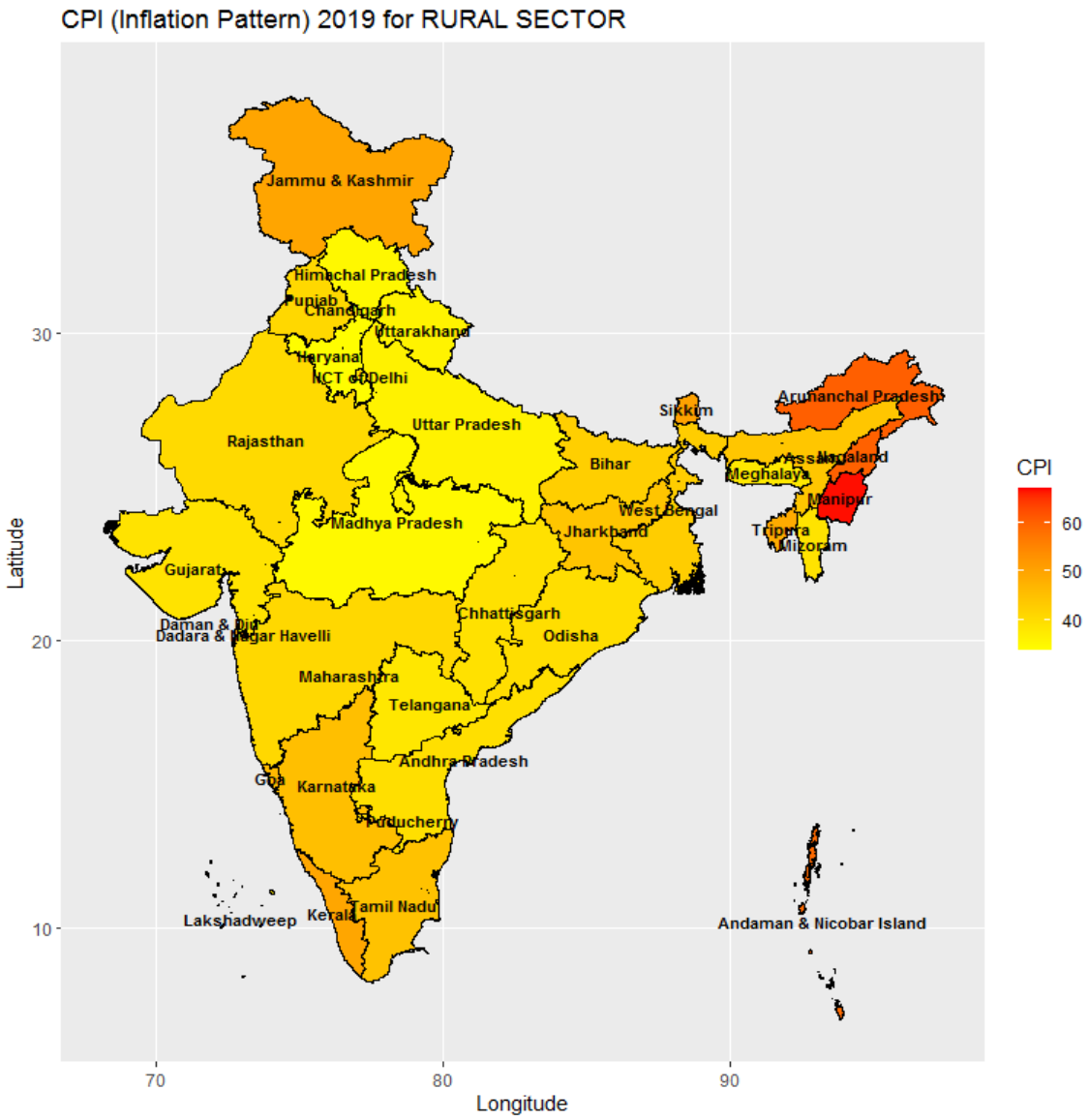
MAP PLOT

```r
#LIBRARIES USED
library(rgdal)     # for reading spatial data i.e. shape file
library(ggplot2)   # for plotting visualizations
library(dplyr)     # for using filter() for selecting rows.

shp <- readOGR('F:/assign1_datasets/Admin2.shp') #reads shape file

file_in = read.csv("F:/assign1_datasets/Statewise_General_Index_july2019_20Aug2020.csv")
shp.f <- fortify(shp, region = "ST_NM")         # converts shape file to dataframe for ggplot2


merge.shp.coef<-merge(shp.f,p, by="id", all.x=TRUE)  #maps data to shape file dataframe
final.plot<-merge.shp.coef[order(merge.shp.coef$order), ]
cnames <- aggregate(cbind(long, lat) ~ id, data=final.plot, FUN=function(x) mean(range(x)))

#2019 has the highest inflation ploting its data for both urban and rural sectors
#(data for 2020 is not complete)

ggplot()+geom_polygon(data = final.plot,aes(x = long, y = lat,
                                            group = group, fill = X9),
                      color = "black", size = 0.25) + coord_map()+
  scale_fill_gradient(name="CPI", limits=c(34,67), low = 'yellow', high = 'red')+
  labs(title="CPI 2019 for RURAL SECTOR")+
  xlab('Longitude')+
  ylab('Latitude')+
  geom_text(data=cnames, aes(long, lat, label = id), size=2, fontface="bold")

ggplot()+geom_polygon(data = final.plot,aes(x = long, y = lat,
                                            group = group, fill = X19),
                      color = "black", size = 0.25) +
  coord_map()+
  scale_fill_gradient(name="CPI", limits=c(30,46), low = 'yellow', high = 'red')+
  labs(title="CPI 2019 for URBAN SECTOR")+
  xlab('Longitude')+
  ylab('Latitude')+
  geom_text(data=cnames, aes(long, lat, label = id), size=2, fontface="bold")
```

Here ‘p’ is the data frame containing the year-wise data from 2011- 2020 for each state and union territory.

To access the full code refer to the link: [code]

OUTPUT



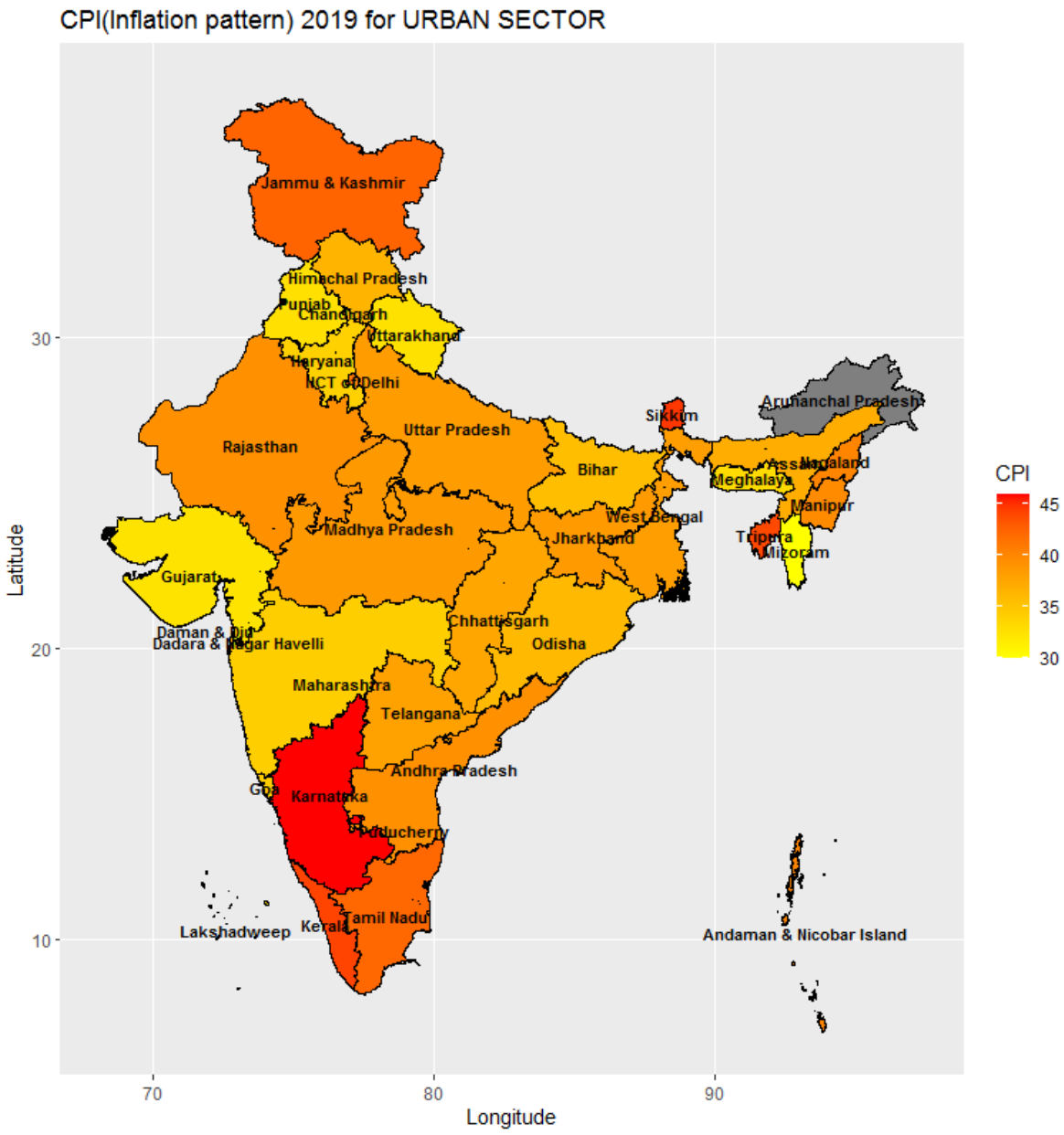
INTERPRETATION

The above plot shows the inflation pattern over different states (rural) of India for the year of 2019.

Inflation is calculated from the base year 2010 considering it as 100, using the formula given above.

It can be interpreted that regions which lie on a certain part of India have similar inflation rate, except for Arunachal Pradesh and Jammu and Kashmir.

OUTPUT



INTERPRETATION

In Urban sector the inflation rates are different and more diverse ranging from 46 being highest in Karnataka to 30 in Mizoram.

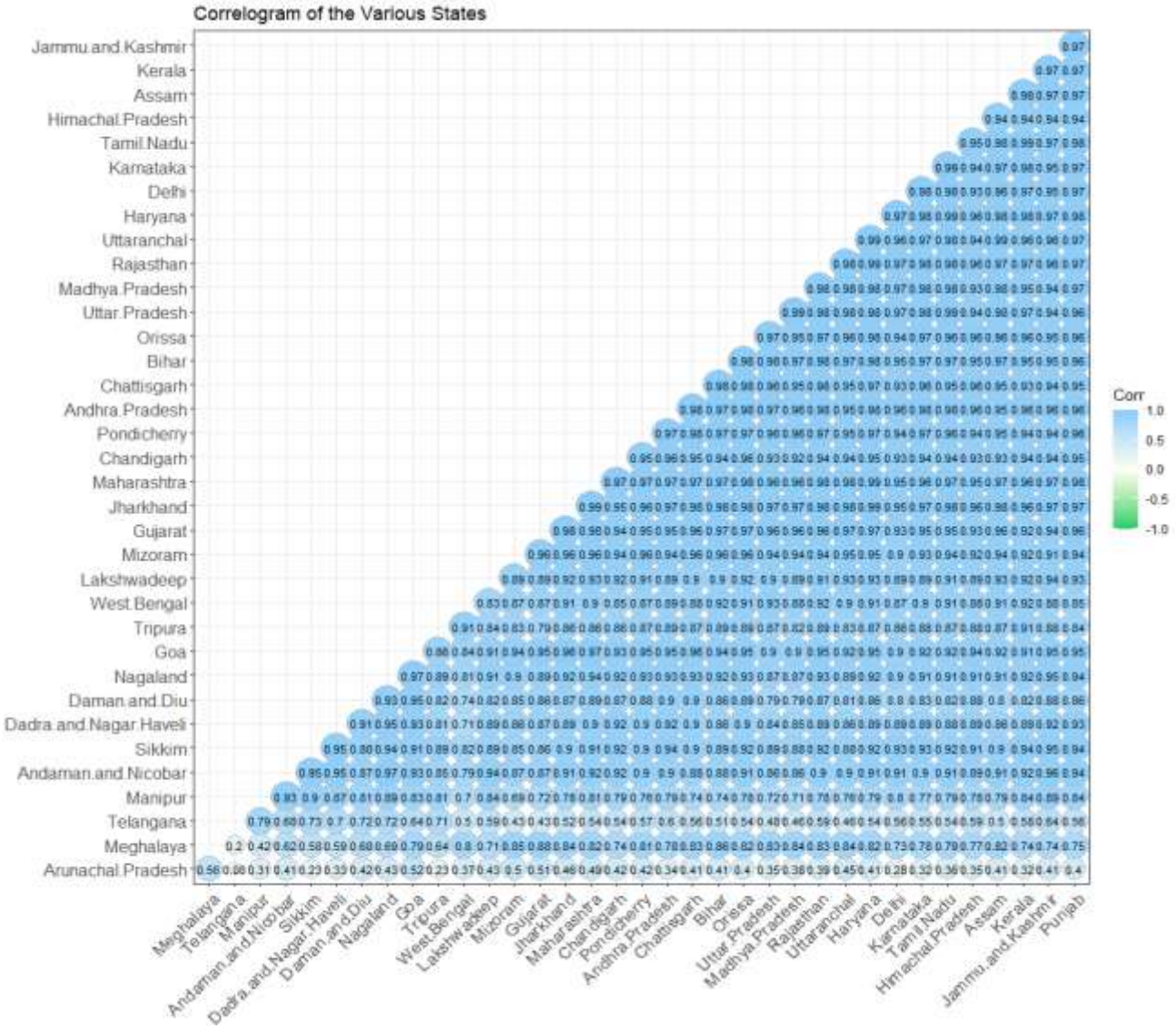
From both the above visualizations it can be interpreted that inflation in rural sector is higher than in urban sector for the year of 2019. Also for previous year it shows almost same inflation pattern.

CORRELOGRAM

```r
#Correlogram
library(ggplot2)
library(tidyverse)
library(ggcorrplot)

df<- read.csv("cpi2020.csv")
options(max.print=1000000)
df
df<-df[-c(1,2:3)]
df[is.na(df)] = 100
df = subset(df, select = -c(X) )
df
cor(df)
round(cor(df),2)
corr=round(cor(df),2)
ggcorrplot(corr,hc.order = TRUE,
           type="lower",
           lab=TRUE,
           lab_size=3,
           method="circle",
           colors=c("springgreen3","ivory","skyblue1"),
           title="Correlogram of the Various States",
           ggtheme=theme_bw)
```

To access the full code refer to the link: [code]

OUTPUT


Correlogram of the Various States
Jammu.and.Kashmir
Kerala
Assam
Himachal.Pradesh
Tamil.Nadu
Karnataka
Delhi
Haryana
Uttaranchal
Rajasthan
Madhya.Pradesh
Uttar.Pradesh
Orissa
Bihar
Chattisgarh
Andhra.Pradesh
Pondicherry
Chandigarh
Maharashtra
Jharkhand
Gujarat
Mizoram
Lakshwadeep
West.Bengal
Tripura
Goa
Nagaland
Daman.and.Diu
Dadra.and.Nagar.Haveli
Sikkim
Andaman.and.Nicobar
Manipur
Telangana
Meghalaya
Arunachal.Pradesh
Corr
1.0
0.5
0.0
-0.5
-1.0

INTERPRETATION

The correlogram is commonly used for checking randomness in a dataset. It gives the

Visual representation of one variable with all other variables in the dataset

Its range lies from -1 to 1

For strong (or good) relation it will give positive i.e. 1(represented by blue in our plot)

0 means no relation (shown by the white color)

-1 means negative relation (shown by the green color)

As shown In the plot we can see that blue is clearly the dominant trend signifying strong relation among the corresponding states apart from Telangana and Arunachal Pradesh where the correlation is not that much prevalent

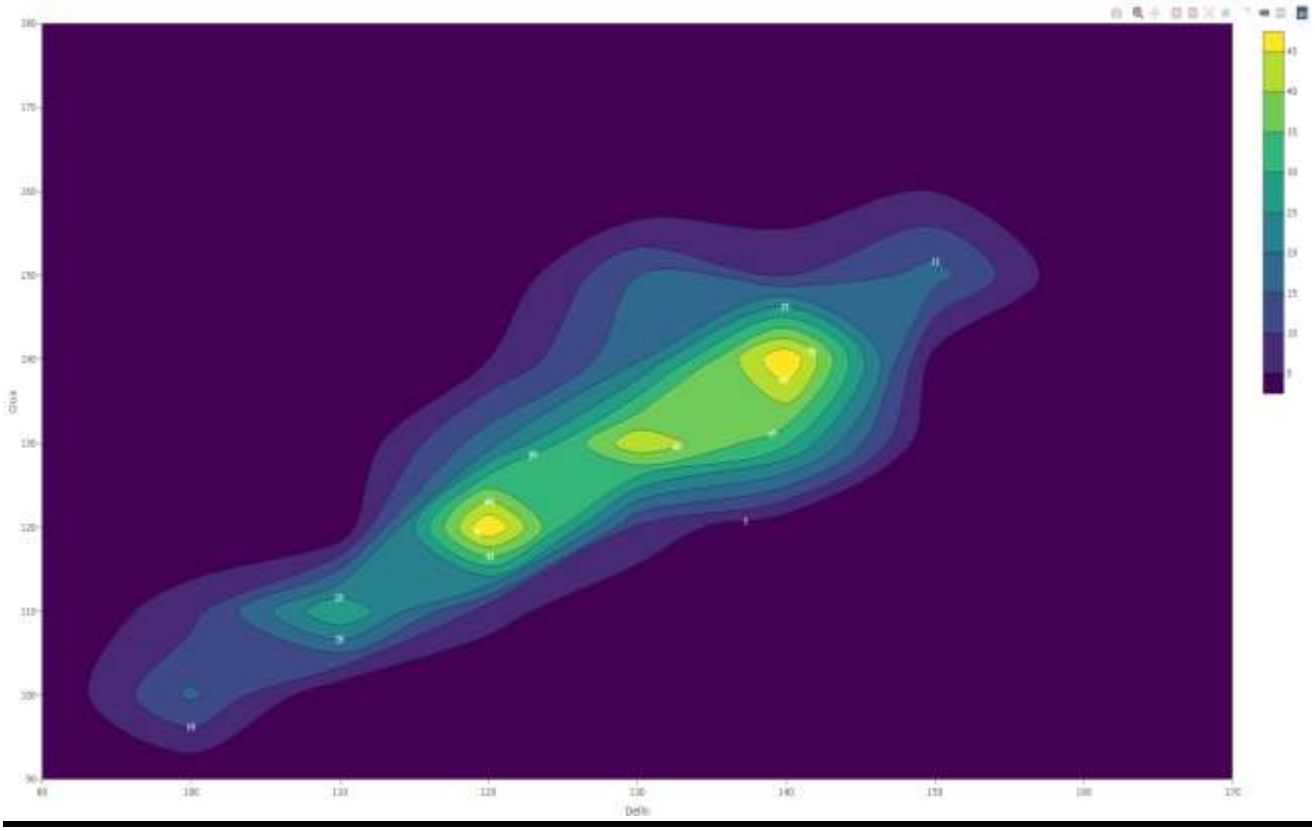
These two states seem to have not so strong relation with other states which is shown by whitish blue color in the plot .Thus this plot summary shows that the states have mostly good relation amongst them.

2D HISTOGRAM CONTOUR

```r
#2D_Histogram_Contour
library(plotly)

cnt <- with(df, table(Delhi, Goa))
figure <- plot_ly(df, x = ~Delhi, y = ~Goa)
figure <- figure %>%
  add_trace(
    type='histogram2dcontour',
    contours = list(
      showlabels = T,
      labelfont = list(
        family = 'Raleway',
        color = 'white'
      )
    ),
    hoverlabel = list(
      bgcolor = 'white',
      bordercolor = 'black',
      font = list(
        family = 'Raleway',
        color = 'black'
      )
    )
  )
figure
```

To access the full code refer to the link: [code]

OUTPUT



INTERPRETATION

It is a type of contour plot that visualizes a bivariate distribution with contour lines. The 2D histogram uses a color-scale to depict distribution across two variables in a contour plot.

The two variables in our case are taken as Delhi and Goa

From the plot it can be interpreted that the denser regions (shown by shades of green up to yellow) have most values lying in them i.e. most of the CPIs values lie in the range 120 to 145 which can be verified in the dataset .Few of the values are in the range less than 110 as compared to other values.

BIBLIOGRAPHY

- For data set : data.gov.in
- For understanding CPI : www.investopedia.com
- For Map plot : rstudio-pubs-static.s3.amazonaws.com
- For histogram contour: plotly.com
- For Correlogram: www.sthda.com
- For Pie chart and Time Series lot: www.tutorialspoint.com