

## **Project overview:**

This project is based on data Web Scraping ,pre-processing and data Visualization. A database may have missing data for a number of reasons, such as that the column needs a value or that the data was not captured at the time of data collection. The process of cleaning up raw data and turning it into usable information is known as data preparation. In this project ,we will analyse the dataset contains statistics in Index,Grocery,Housing, Utilites, Transportation,Health, Misc in 50 state.

## **project solution design:**

The automated technique of extracting data from webpages is called "web scraping." Web scrapers, a type of software used for web scraping, are used to complete this procedure. According to user needs, they automatically load and extract data from the websites.

The process of cleaning up raw data and turning it into usable information is known as data preparation. The data can then be used for a variety of tasks, including reporting and decision-making.

Data pre-processing techniques are used when the data is inconsistent, which indicates that the data is not recorded in accordance with the constraints on the column, noisy, which may contain numerous errors or outliers, and incomplete, which indicates that some attribute value is missing.

The database may have missing data for a number of reasons, such as that the column needs a value to be entered or that the data was not captured at the time of data collection.

Several methods can be used to fill in the missing values:

The first approach to handling the missing value of Assault in place G ignores the tuple; the second approach fills the Assault value of place G with "unknown."

The data considered to be outliers are those that may differ from all other data or those that may result in problems. The percentage of the urban population in this case must range from 0% to 100%. The value of the urban population being greater than 100% is therefore unusual.

State I, which has 570 people living in metropolitan areas, is an anomaly in the data set, as is state G, which lists assault as "unknown."

Take away the rows to smooth the data and get rid of these two outliers.

The urban population is split into four categories: small for those with less than 50% of the total, medium for those with between 50% and 60%, big for those with between 60% and 70%, and extra-

large for those with more than 70% of the total.

Most serious, less serious, and least serious crimes are categorized similarly. The most serious crimes go into three categories: murder over 10, assault over 250, and assault between 5 and 10, between 110 and 250. The least serious crimes fall into the categories of murder under 5, and assault under 110.

For obtaining summary statistics, R has a large variety of functions. The sapply() function can be used in conjunction with a given summary statistic to produce descriptive statistics. Mean, sd, var, min, max, median, range, and quantile are examples of potential functions that can be utilized in sapply.

Data visualization is a method for presenting insights in data through the use of visual signals like graphs, charts, maps, and many others. This is beneficial because it makes it simple and intuitive to interpret complex data sets, which enables users to make wiser decisions. Data visualization is a technique for displaying data insights by utilizing visual cues like graphs, charts, maps, and more. This helps people make better decisions since it makes it easy and intuitive to analyze complex data sets.

R is a programming language intended for scientific research, graphical data analysis, and statistical computing. It is typically chosen for data visualization since it provides flexibility and requires little coding thanks to its packages.

# Data Collection via Web Scraping :

```
install.packages('rvest')
```

```
install.packages('dplyr')
```

```
installed.packages('xml2')
```

```
library(rvest)
```

```
library(dplyr)
```

```
library(xml2)
```

```
content <- read_html("https://meric.mo.gov/data/cost-living-data-series")
```

```
table <- content %>% html_table(fill = TRUE)
```

```
first_table <- table[[1]]
```

```
View(first_table)
```

```
write.csv(first_table, file = "Prajukta.csv")DATA TABLE
```

The screenshot shows the RStudio interface with the following details:

- Code Editor:** The left pane displays the R script code for data collection and manipulation.
- Environment Pane:** The right pane shows the global environment with objects like `content`, `table`, and `first_table`.
- RStudio Sidebar:** The bottom right contains links to R Resources, Manuals, Reference, Packages, and Miscellaneous Material.

```
1 install.packages('rvest')
2 install.packages('dplyr')
3 installed.packages('xml2')
4
5
6
7 library(rvest)
8 library(dplyr)
9 library(xml2)
10
11
12
13 content <- read_html("https://meric.mo.gov/data/cost-living-data-series")
14
15
16
17 table <- content %>% html_table(fill = TRUE)
18 first_table <- table[[1]]
19
20
21
22 View(first_table)
23
24
25
26 write.csv(first_table, file = "Prajukta.csv")
```

View table

## DATA TABLE

The screenshot shows the RStudio interface. On the left is a data table titled 'first\_table' with columns: Rank, State, Index, Grocery, Housing, Utilities, Transportation, Health, and Misc. The table lists 53 US states from Mississippi at rank 1 to Virginia at rank 53. On the right is the 'Global Environment' pane, which displays a list of objects: col\_link, col\_table, content, data, first\_table, iris, and table. Below the list is a 'Values' section showing a vector of numerical values.

Rank	State	Index	Grocery	Housing	Utilities	Transportation	Health	Misc.
1	Mississippi	84.50	91.9	68.5	88.5	92.6	99.5	91.2
2	Oklahoma	86.70	94.7	72.4	95.0	91.4	91.5	92.2
3	Alabama	87.10	98.2	68.2	100.2	88.1	89.5	95.2
4	Kansas	87.30	94.3	72.0	98.7	96.3	101.1	91.0
5	Iowa	88.20	98.0	70.6	95.5	95.4	98.9	94.9
6	Georgia	88.90	96.4	77.7	90.0	89.4	96.5	94.8
7	Ohio	89.40	100.8	71.1	92.4	95.6	94.4	98.5
8	West Virginia	89.80	96.3	76.9	91.7	104.7	97.1	94.0
9	Missouri	90.10	96.8	80.4	94.2	97.8	91.2	92.8
10	Indiana	90.20	94.5	76.8	107.6	94.8	97.5	93.3
11	Tennessee	90.30	94.4	84.0	94.7	91.2	87.8	93.3
12	Arkansas	90.70	92.5	78.4	100.1	89.7	84.2	100.6
13	Nebraska	91.10	96.9	81.6	88.8	102.1	99.3	94.2
14	Wyoming	91.66	101.4	80.4	81.7	96.4	98.2	98.6
15	Michigan	91.70	92.3	82.1	98.2	98.5	95.8	96.8
16	Illinois	91.90	100.5	81.7	94.6	104.3	96.0	92.8
17	Texas	92.60	91.0	84.8	105.9	91.9	94.8	96.8
18	Kentucky	92.80	92.5	75.7	105.4	105.9	77.0	105.4
19	Louisiana	93.50	98.8	85.2	88.3	96.0	100.4	99.2
20	New Mexico	93.80	96.8	87.3	88.4	101.0	103.0	97.4
21	Wisconsin	93.90	97.0	84.4	105.6	95.2	110.1	95.3
22	Minnesota	95.10	97.4	83.5	95.6	100.8	109.0	102.2
23	South Dakota	96.10	100.4	99.6	91.9	89.5	98.6	92.6
24	South Carolina	96.30	101.4	89.5	108.6	88.0	96.8	98.3
25	North Carolina	96.90	98.4	94.0	93.2	90.4	110.9	99.9
26	North Dakota	97.40	103.5	88.2	104.7	101.5	110.5	98.2
27	Pennsylvania	98.20	102.9	89.4	104.4	106.1	100.1	100.1
28	Idaho	98.90	95.6	102.3	79.2	114.8	93.0	100.7
29	Puerto Rico	100.60	118.6	86.3	155.6	84.9	73.7	94.0
30	Nevada	101.90	104.6	112.9	98.3	112.2	93.4	89.0
31	Utah	102.00	99.5	105.0	92.1	110.3	93.5	103.0
32	Virginia	102.10	96.3	108.3	100.1	94.1	102.7	101.4

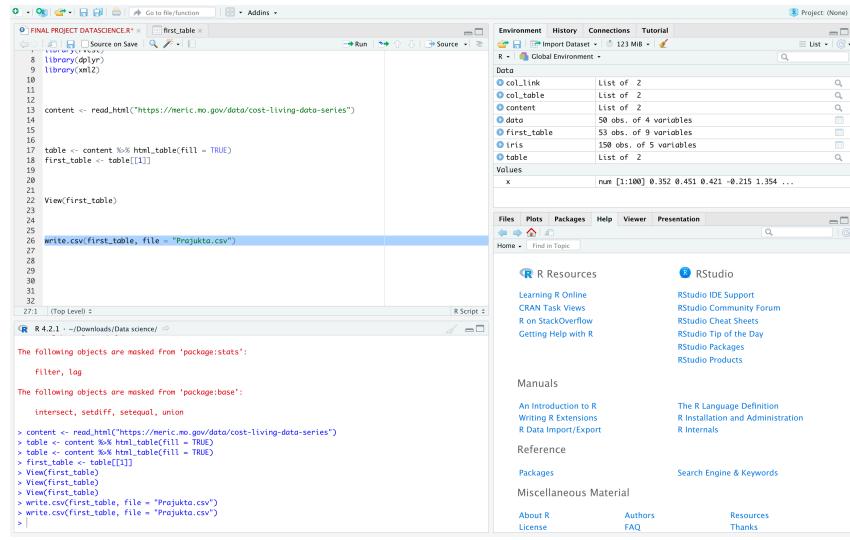
Data table

This screenshot shows the RStudio interface again, but with a different dataset named 'first\_table'. The data table contains 53 entries, each with columns: Rank, State, Index, Grocery, Housing, Utilities, Transportation, Health, and Misc. The values are generally higher than in the previous table, with some outliers like Puerto Rico having very high values in the Utilities and Transportation columns. The Global Environment pane shows the same list of objects as before: col\_link, col\_table, content, data, first\_table, iris, and table.

Rank	State	Index	Grocery	Housing	Utilities	Transportation	Health	Misc.
24	South Carolina	96.30	101.4	89.5	108.6	88.0	96.8	98.3
25	North Carolina	96.90	98.4	94.0	93.2	90.4	110.9	99.9
26	North Dakota	97.40	103.5	88.2	104.7	101.5	110.5	98.2
27	Pennsylvania	98.20	102.9	89.4	104.4	106.1	100.1	100.1
28	Idaho	98.90	95.6	102.3	79.2	114.8	93.0	100.7
29	Puerto Rico	100.60	118.6	86.3	155.6	84.9	73.7	94.0
30	Nevada	101.90	104.6	112.9	98.3	112.2	93.4	89.0
31	Utah	102.00	99.5	105.0	92.1	110.3	93.5	103.0
32	Virginia	102.10	96.3	108.3	100.1	94.1	102.7	101.4
33	Florida	104.50	106.4	112.2	99.0	97.7	98.0	100.2
34	Colorado	105.20	100.8	115.7	83.9	112.2	97.9	104.2
35	Montana	105.30	100.8	115.7	83.9	111.2	97.9	104.2
36	Delaware	105.40	106.4	103.7	94.1	117.5	109.6	106.8
37	Arizona	108.00	101.8	127.2	99.4	101.3	92.4	98.7
38	Rhode Island	111.20	96.3	113.8	123.5	110.8	101.6	114.3
39	New Jersey	114.00	107.0	133.5	106.4	111.4	96.5	103.7
40	Washington	114.00	107.0	133.5	106.4	111.4	96.5	103.7
41	Maine	114.50	103.2	119.8	113.7	111.5	128.7	127.4
42	New Hampshire	114.70	104.6	107.4	113.7	111.5	128.7	127.4
43	Connecticut	115.40	99.1	121.5	127.3	112.1	108.1	116.4
44	Vermont	116.40	106.8	129.5	122.0	121.3	109.6	106.3
45	Oregon	122.20	106.4	156.1	91.3	125.1	109.2	108.0
46	Maryland	124.10	112.2	157.0	106.9	103.0	94.4	112.4
47	Alaska	125.50	131.8	120.0	138.4	121.0	151.7	120.5
48	New York	135.70	109.2	194.1	100.6	105.6	103.1	114.9
49	California	138.70	112.3	194.0	122.4	124.4	109.7	118.9
50	Massachusetts	149.90	113.0	217.3	121.2	134.0	113.8	120.6
51	District of Columbia	153.40	105.2	243.9	118.4	109.7	94.8	118.9
52	Hawaii	186.00	148.5	307.0	134.6	128.3	120.3	124.1
53	US	***	100.00	100.0	100.0	100.0	100.0	100.0

Data table

## CSV FILE:



```

1 #!/usr/bin/Rscript
2
3 # Load required libraries
4 library(dplyr)
5 library(xlsx)
6
7 # Read the HTML content from the URL
8 content <- read_html("https://eric.ed.gov/data/cost-living-data-series")
9
10 # Convert the content to an HTML table
11 table <- content %>% html_table(fill = TRUE)
12 first_table <- table[[1]]
13
14 # View the first table
15 View(first_table)
16
17 # Write the table to a CSV file
18 write.csv(first_table, file = "Projukta.csv")
19
20
21
22
23
24
25
26
27
28
29
30
31
32

```

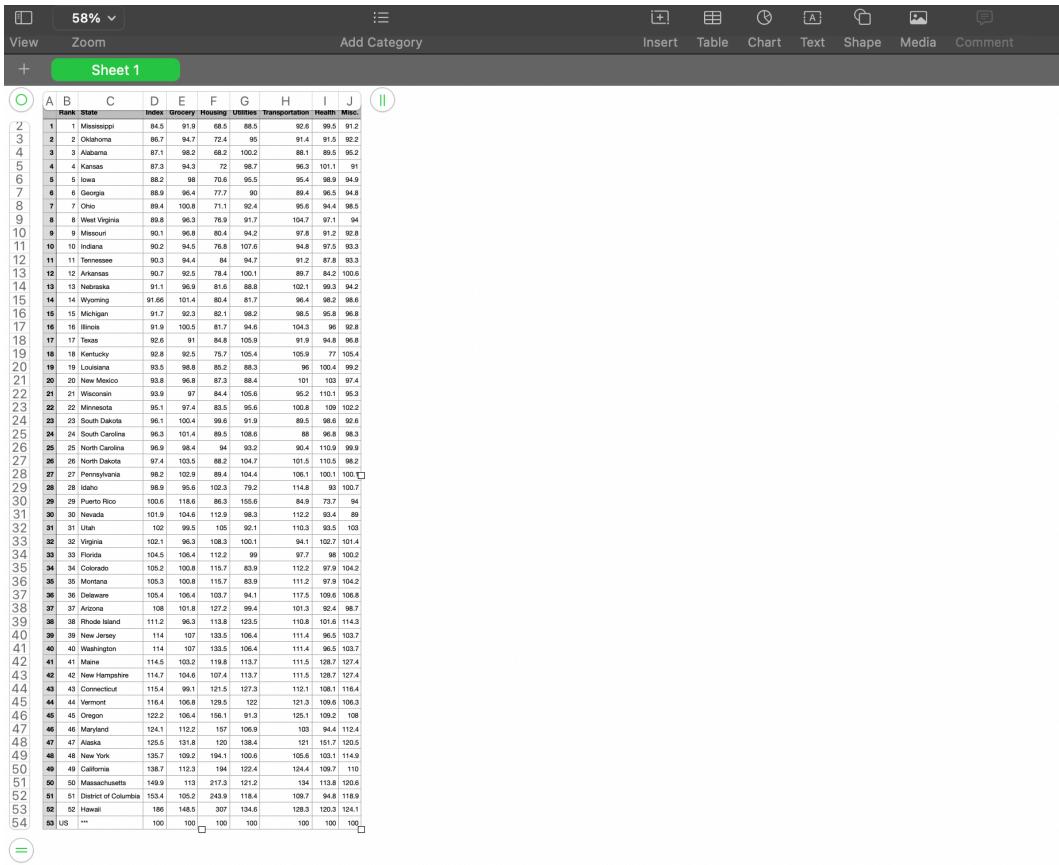
The following objects are masked from 'package:bosix':  
 intersect, setdiff, setequal, union

```

> content <- read_html("https://eric.ed.gov/data/cost-living-data-series")
> table <- content %>% html_table(fill = TRUE)
> table <- content %>% html_table(fill = TRUE)
> first_table <- table[[1]]
> View(first_table)
> write.csv(first_table, file = "Projukta.csv")
> write.csv(first_table, file = "Projukta.csv")
>

```

Run csv file



Rank	State	Index	Grocery	Housing	Utilities	Transportation	Health	Misc.
1	Mississippi	84.5	91.9	68.5	88.5	92.6	99.5	91.2
2	Oklahoma	86.7	94.7	72.4	95	91.4	91.5	92.2
3	Alabama	87.1	98.2	68.2	100.2	88.1	89.5	95.2
4	Kansas	87.3	94.3	72	98.7	96.3	101.1	91
5	Iowa	88.2	98	70.6	95.5	96.4	98.9	94.9
6	Georgia	88.9	96.4	77.7	90	89.4	96.5	94.8
7	Ohio	89.4	100.8	71.1	92.4	95.6	94.4	98.5
8	West Virginia	89.8	95.3	76.9	91.7	104.7	97.1	94
9	Missouri	90.1	95.8	80.4	94.2	97.8	91.2	92.8
10	Indiana	90.2	94.5	76.8	107.6	94.8	97.5	93.3
11	Tennessee	90.3	94.4	84	94.7	91.2	87.8	93.3
12	Arkansas	90.7	92.5	78.4	100.1	89.7	94.2	100.6
13	Nebraska	91.1	96.9	81.6	88.4	102.1	93.3	94.2
14	Wyoming	91.5	99	101.4	84.4	84.7	96.4	98.6
15	Michigan	91.6	96.7	82	93.2	98.5	95.8	94.8
16	Illinois	91.9	100.5	81.7	94.6	104.3	96	93.8
17	Texas	92.6	91	84.8	105.9	91.9	94.8	96.8
18	Kentucky	92.8	92.5	75.7	105.4	105.9	77	105.4
19	Louisiana	93.5	98.8	85.2	88.3	96	100.4	99.2
20	New Mexico	93.8	96.8	87.3	88.4	101	103	97.4
21	Wisconsin	93.9	97	84.4	105.6	95.2	110.1	95.3
22	Minnesota	95.1	97.4	83.5	95.6	100.8	109	102.2
23	South Dakota	96.1	100.4	99.6	91.9	89.5	98.6	92.6
24	North Carolina	96.3	101.4	89.5	108.6	88	96.8	98.3
25	Utah	96.9	98.4	94	93.2	90.4	110.9	99.9
26	North Dakota	97.4	103.5	88.2	104.7	101.5	110.5	98.2
27	Pennsylvania	98.2	102.9	89.4	104.4	106.1	100.1	100
28	Idaho	98.9	95.6	102.3	79.2	114.8	93	100.7
29	Puerto Rico	100.6	118.6	86.3	115.6	84.9	73.7	94
30	Nevada	101.9	104.6	112.9	98.3	112.2	93.4	89
31	Montana	102.1	99.5	105	92.1	110.3	93.5	103
32	Virginia	102.1	96.3	108.3	100.1	94.1	102.7	101.4
33	Florida	104.5	106.4	112.2	99	97.7	98	100.2
34	Colorado	105.2	100.8	115.7	83.9	112.2	97.9	104.2
35	Montana	105.3	100.9	107	93.7	112.3	97.8	104.2
36	Wyoming	105.6	106.6	100.1	94.1	117.3	100.1	105.9
37	Arizona	106	101.8	137.2	99.4	101.3	92.4	98.7
38	Rhode Island	111.2	96.3	113.8	123.5	110.8	101.6	114.3
39	New Jersey	114	107	133.6	106.4	111.4	96.5	103.7
40	Washington	114	107	133.5	106.4	111.4	96.5	103.7
41	Maine	114.5	103.2	119.8	113.7	111.5	128.7	127.4
42	New Hampshire	114.7	104.6	107.4	113.7	111.5	128.7	127.4
43	Connecticut	115.4	99.1	121.5	127.3	112.1	106.1	116.4
44	Vermont	116.4	106.8	129.5	122	121.3	109.6	106.3
45	Oregon	122.2	106.4	156.1	91.3	125.1	109.2	101
46	Maryland	124.1	112.2	157	106.9	103	94.4	112.4
47	Alaska	125.5	131.8	120	138.4	121	151.7	120.5
48	New York	135.7	109.2	194.1	100.6	106.6	103.1	114.9
49	California	136.7	112.3	194	122.4	124.4	109.7	110
50	Massachusetts	140.9	113	217.3	121.2	134	113.8	120.6
51	District of Columbia	155.4	105.2	243.9	118.4	109.7	94.8	118.9
52	Hawaii	156	148.5	307	134.6	128.3	120.3	124.1
53	US	---	100	100	100	100	100	100

CSV File created

## Read CSV file:

```
> read.csv("Downloads/Prajukta.csv")
```

The screenshot shows the RStudio interface with the following details:

- Environment:** Shows the 'data' object as a global environment containing 50 observations and 4 variables.
- Data:** Shows the 'data' dataset with columns: State, Index, Grocery, Housing, Utilities, Transportation, Health, and Misc.
- Code Editor:** Displays the R code used to read the CSV file: `ad.csv("Downloads/Prajukta.csv")`.
- Table:** A large table below the code editor lists the data for all 50 states, including their names and corresponding values for each variable.

Csv

## **Data pre-processing:**

Data Cleaning >Data Integration > Data Transformation >Data Reduction  
>Data Discretization

## **Data Cleaning:**

### **Data Munging:**

No munging or wrangling is necessary for the data set.

### **Handling Missing Data:**

```
>mean(data$Index, na.rm=TRUE)
```

```
>mean(data$Grocery, na.rm=TRUE)
```

```
>mean(data$Housing, na.rm=TRUE)
```

```
>mean(data$Utilites, na.rm=TRUE)
```

```
>mean(data$Transportation, na.rm=TRUE)
```

```
>mean(data$Health, na.rm=TRUE)
```

```
>mean(data$Misc, na.rm=TRUE)
```

125% ▾

Zoom Add Category Insert Table Chart Text Shape Media Comment Collaborate

Sheet 1

Prajukta\_1

Rank	State	Index	Grocery	Housing	Utilities	Transportation	Health	Misc.
1	Mississippi	84.5	91.9	68.5	88.5	92.6	99.5	91.2
2	Oklahoma	86.7	94.7	72.4	95	91.4	91.5	92.2
3	Alabama	87.1	98.2	68.2	100.2	88.1	89.5	95.2
4	Kansas	87.3	94.3	72	98.7	96.3	101.1	91
5	Iowa	88.2	98	70.6	95.5	95.4	98.9	94.9
6	Georgia	88.9	96.4	77.7	90	89.4	96.5	94.8
7	Ohio	89.4	100.8	71.1	92.4	95.6	94.4	98.5
8	West Virginia	89.8	96.3	76.9	91.7	104.7	97.1	94
9	Missouri	90.1	96.8	80.4	94.2	97.8	91.2	92.8
10	Indiana	90.2	94.5	76.8	107.6	94.8	97.5	93.3
11	Tennessee	90.3	94.4	84	94.7	91.2	87.8	93.3
12	Arkansas	90.7	92.5	78.4	100.1	89.7	84.2	100.6
13	Nebraska	91.1	96.9	81.6	88.8	102.1	99.3	94.2
14	Wyoming	91.66	101.4	80.4	81.7	96.4	98.2	98.6
15	Michigan	91.7	92.3	82.1	98.2	98.5	95.8	96.8
16	Illinois	91.9	100.5	81.7	94.6	104.3	96	92.8
17	Texas	92.6	91	84.8	105.9	91.9	94.8	96.8
18	Kentucky	92.8	92.5	75.7	105.4	105.9	77	105.4
19	Louisiana	93.5	98.8	85.2	88.3	96	100.4	99.2
20	New Mexico	93.8	96.8	87.3	88.4	101	103	97.4
21	Wisconsin	93.9	97	84.4	105.6	95.2	110.1	95.3
22	Minnesota	95.1	97.4	83.5	95.6	100.8	109	102.2
23	South Dakota	96.1	100.4	99.6	91.9	89.5	98.6	92.6
24	South Carolina	96.3	101.4	89.5	108.6	88	96.8	98.3
25	North Carolina	96.9	98.4	94	93.2	90.4	110.9	99.9
26	North Dakota	97.4	103.5	88.2	104.7	101.5	110.5	98.2
27	Pennsylvania	98.2	102.9	89.4	104.4	106.1	100.1	100.1
28	Idaho	98.9	95.6	102.3	79.2	114.8	93	100.7
29	Puerto Rico	100.6	118.6	86.3	155.6	84.9	73.7	94
30	Nevada	101.9	104.6	112.9	98.3	112.2	93.4	89

## Handling missing data

125% ▾

Zoom Add Category Insert Table Chart Text Shape Media Comment Collaborate

Sheet 1

20	20 New Mexico	93.8	96.8	87.3	88.4	101	103	97.4
21	21 Wisconsin	93.9	97	84.4	105.6	95.2	110.1	95.3
22	22 Minnesota	95.1	97.4	83.5	95.6	100.8	109	102.2
23	23 South Dakota	96.1	100.4	99.6	91.9	89.5	98.6	92.6
24	24 South Carolina	96.3	101.4	89.5	108.6	88	96.8	98.3
25	25 North Carolina	96.9	98.4	94	93.2	90.4	110.9	99.9
26	26 North Dakota	97.4	103.5	88.2	104.7	101.5	110.5	98.2
27	27 Pennsylvania	98.2	102.9	89.4	104.4	106.1	100.1	100.1
28	28 Idaho	98.9	95.6	102.3	79.2	114.8	93	100.7
29	29 Puerto Rico	100.6	118.6	86.3	155.6	84.9	73.7	94
30	30 Nevada	101.9	104.6	112.9	98.3	112.2	93.4	89
31	31 Utah	102	99.5	105	92.1	110.3	93.5	103
32	32 Virginia	102.1	96.3	108.3	100.1	94.1	102.7	101.4
33	33 Florida	104.5	106.4	112.2	99	97.7	98	100.2
34	34 Colorado	105.2	100.8	115.7	83.9	112.2	97.9	104.2
35	35 Montana	105.3	100.8	115.7	83.9	111.2	97.9	104.2
36	36 Delaware	105.4	106.4	103.7	94.1	117.5	109.6	106.8
37	37 Arizona	108	101.8	127.2	99.4	101.3	92.4	98.7
38	38 Rhode Island	111.2	96.3	113.8	123.5	110.8	101.6	114.3
39	39 New Jersey	114	107	133.5	106.4	111.4	96.5	103.7
40	40 Washington	114	107	133.5	106.4	111.4	96.5	103.7
41	41 Maine	114.5	103.2	119.8	113.7	111.5	128.7	127.4
42	42 New Hampshire	114.7	104.6	107.4	113.7	111.5	128.7	127.4
43	43 Connecticut	115.4	99.1	121.5	127.3	112.1	108.1	116.4
44	44 Vermont	116.4	106.8	129.5	122	121.3	109.6	106.3
45	45 Oregon	122.2	106.4	156.1	91.3	125.1	109.2	108
46	46 Maryland	124.1	112.2	157	106.9	103	94.4	112.4
47	47 Alaska	125.5	131.8	120	138.4	121	151.7	120.5
48	48 New York	135.7	105.2	194.1	100.6	105.6	103.1	114.9
49	49 California	138.7	112.3	194	122.4	124.4	109.7	110
50	50 Massachusetts	149.9	113	217.3	121.2	134	113.8	120.6
51	51 District of Columbia	153.4	105.2	243.9	118.4	109.7	94.8	118.9
52	52 Hawaii	186	148.5	307	134.6	128.3	120.3	124.1

## Handling missing data

# Smooth Noisy Data:

## For Index

```
install.packages('tidyverse')
library(tidyverse)
Data <- read.csv("Prajukta_1.csv");
Data$Index <- gsub("[()]", "", Data$Index)
view(Data)
```

The screenshot shows the RStudio interface. In the top-left pane, there is a code editor with the following R code:

```
1 install.packages('tidyverse')
2 library(tidyverse)
3 
4 
5 
6 
7 
8 Data <- read.csv("Prajukta_1.csv");
9 Data$Index <- gsub("[()]", "", Data$Index)
10 view(Data)
```

In the bottom-left pane, the R console output is displayed:

```
The downloaded binary packages are in
  /var/folders/s0/135xdmz9dgdhmrm4r83qs8000gn/T/Rtmps6vAlv/downloaded_packages
-- Attaching packages --
  ✓ purrr 0.3.5
  ✓ tibble 3.1.8
  ✓ dplyr 1.0.10
  ✓ tidyverse 1.3.2
  ✓ readr 2.1.3
  ✓ forcats 0.5.2
-- Conflicts --
  ▶ dplyr::filter() masks stats::filter()
  ▶ dplyr::lag()  masks stats::lag()
  ▶ tidyverse::dplyr()
  ▶ tidyverse::lag()
> library(tidyverse)
> Data <- read.csv("Prajukta_1.csv");
> Data$Index <- gsub("[()]", "", Data$Index)
> view(Data)
> |
```

In the top-right pane, the Global Environment is shown with the following objects:

- content: List of 2
- data: 50 obs. of 4 variables
- Data: 53 obs. of 10 variables
- first\_table: 53 obs. of 9 variables
- iris: 150 obs. of 5 variables
- table: List of 2

In the bottom-right pane, the RStudio Help center is visible.

## Smooth Noisy Data for Index

The screenshot shows the RStudio interface. In the top-left pane, there is a code editor with the following R code:

```
1 X Rank State Index Grocery Utilities Transportation Health Misc.
2 1 1 Mississippi 84.5 91.9 68.5 88.5 92.6 99.5 91.2
3 2 2 Oklahoma 86.7 94.7 72.4 95.0 91.4 91.5 92.2
4 3 3 Alabama 87.1 98.2 68.2 100.2 88.1 96.3 101.1 91.0
5 4 4 Kansas 87.3 94.3 72.0 98.7 96.3 101.1 94.9 94.9
6 5 5 Iowa 88.2 98.0 70.6 95.5 95.4 98.9 94.9 94.9
7 6 6 Georgia 88.9 96.4 77.7 90.0 89.4 96.5 94.8 94.8
8 7 7 Ohio 89.4 100.8 71.1 92.4 95.6 94.4 98.5 98.5
9 8 8 West Virginia 89.8 96.3 76.9 91.7 104.7 97.1 94.0
10 9 9 Missouri 90.1 96.8 80.4 94.2 97.8 91.2 92.8
11 10 10 Indiana 90.2 94.5 75.8 107.6 94.8 97.5 93.3
12 11 11 Tennessee 90.3 94.4 84.0 94.7 91.2 87.8 93.3
13 12 12 Arkansas 90.7 92.5 78.4 100.1 89.7 84.2 100.6
14 13 13 Nebraska 91.1 96.9 81.6 88.8 102.1 99.3 94.2
15 14 14 Wyoming 91.66 101.4 80.4 81.7 96.4 98.2 98.6
16 15 15 Michigan 91.7 92.3 82.1 98.2 98.5 95.8 96.8
17 16 16 Illinois 91.9 100.5 81.7 94.6 104.3 96.0 92.8
18 17 17 Texas 92.6 91.0 84.8 105.9 91.9 94.8 96.8
```

In the bottom-left pane, the R console output is displayed:

```
The downloaded binary packages are in
  /var/folders/s0/135xdmz9dgdhmrm4r83qs8000gn/T/Rtmps6vAlv/downloaded_packages
> library(tidyverse)
-- Attaching packages --
  ✓ ggplot2 3.4.0   ✓ purrr 0.3.5
  ✓ tibble 3.1.8    ✓ dplyr 1.0.10
  ✓ tidyverse 1.3.2  ✓ stringr 1.5.0
  ✓ readr 2.1.3     ✓ forcats 0.5.2
-- Conflicts --
  ▶ dplyr::filter() masks stats::filter()
  ▶ dplyr::lag()  masks stats::lag()
  ▶ tidyverse::dplyr()
  ▶ tidyverse::lag()
> library(tidyverse)
> Data <- read.csv("Prajukta_1.csv");
> Data$Index <- gsub("[()]", "", Data$Index)
> view(Data)
> |
```

In the top-right pane, the Global Environment is shown with the following objects:

- content: List of 2
- data: 50 obs. of 4 variables
- Data: 53 obs. of 10 variables
- first\_table: 53 obs. of 9 variables
- iris: 150 obs. of 5 variables
- table: List of 2

In the bottom-right pane, the RStudio Help center is visible.

## Smooth Noisy Data table for Index

## For Grocery:

```
install.packages('tidyverse')
library(tidyverse)
Data <- read.csv("Prajukta_1.csv");
Data$Grocery <- gsub("[()]", "", Data$Grocery)
view(Data)
```

The screenshot shows the RStudio interface with the following details:

- Code Editor:** An R script titled "cleaning.R" containing the provided R code.
- Environment View:** Shows the global environment with objects like "content", "data", "first\_table", "iris", and "table".
- Output View:** Displays the output of the R code, including the loading of tidyverse packages and the resulting Data frame.
- RStudio Help:** A sidebar with links to R Resources, RStudio support, manuals, and reference sections.

## Smooth Noisy Data for Grocery

The screenshot shows the RStudio interface with the following details:

- Code Editor:** An R script titled "cleaning.R" containing the provided R code.
- Environment View:** Shows the global environment with objects like "content", "data", "first\_table", "iris", and "table".
- Output View:** Displays the output of the R code, including the loading of tidyverse packages and the resulting Data frame.
- RStudio Help:** A sidebar with links to R Resources, RStudio support, manuals, and reference sections.

## Smooth Noisy Data table for Grocery

## For Housing:

```
install.packages('tidyverse')
```

```
library(tidyverse)
```

```
Data <- read.csv("Prajukta_1.csv");
```

```
Data$Housing <- gsub("[()]", "", Data$Housing)
```

```
view(Data)
```

```

1 install.packages('tidyverse')
2 library(tidyverse)
3
4
5
6
7
8 Data <- read.csv("Prajukta_1.csv");
9 DataHousing <- gsub("[()]","",Data$Housing)
10 view(Data)

The downloaded binary packages are in
  /var/folders/s0/135xdmz90gdgnhmm4r83qs80000gn/T//RtmpDBtxqa/downloaded_packages
> library(tidyverse)
-- Attaching packages --
✓ ggplot2 3.4.0   ✓ purrr  0.3.5
✓ tibble   3.1.8   ✓ dplyr   1.0.10
✓ tidyverse 1.3.2   ✓ stringr 1.5.0
✓ readr    2.1.3   ✓ forcats 0.5.2
-- Conflicts --
✖ dplyr::filter() masks stats::filter()
✖ dplyr::log()   masks stats::log()
> Data <- read.csv("Prajukta_1.csv");
> DataGrocery <- gsub("[()]","",Data$Grocery)
> view(Data)
> Data <- read.csv("Prajukta_1.csv");
  
```

R 4.2.1 - ~/Downloads/

The downloaded binary packages are in  
/var/folders/s0/135xdmz90gdgnhmm4r83qs80000gn/T//RtmpDBtxqa/downloaded\_packages

Attaching packages --  
tidyverse 1.3.2 --

Conflicts --  
dplyr::filter() masks stats::filter()  
dplyr::log() masks stats::log()

Data <- read.csv("Prajukta\_1.csv");  
DataGrocery <- gsub("[()]","",Data\$Grocery)  
view(Data)  
Data <- read.csv("Prajukta\_1.csv");  
DataHousing <- gsub("[()]","",Data\$Housing)  
view(Data)

## Smooth data for Housing

```

1 * cleaning.R* x 2 data x 3 Data x 4 first_table x 5
  Go to file/function | Addins |
  Filter
  X Rank State Index Grocery Housing Utilities Transportation Health Misc.
  1 1 Mississippi 84.50 91.9 68.5 88.5 92.6 99.5 91.2
  2 2 Oklahoma 86.70 94.7 72.4 95.0 91.4 91.5 92.2
  3 3 Alabama 87.10 98.2 68.2 100.2 88.1 89.5 95.2
  4 4 Kansas 87.30 94.3 72 98.7 96.3 101.1 91.0
  5 5 Iowa 88.20 98.0 70.6 95.5 95.4 98.9 94.9
  6 6 Georgia 88.90 96.4 77.7 90.0 89.4 96.5 94.8
  7 7 Ohio 89.40 100.8 71.1 92.4 95.6 94.4 98.5
  8 8 West Virginia 89.80 96.3 76.9 91.7 104.7 97.1 94.0
  9 9 Missouri 90.10 96.8 80.4 94.2 97.8 91.2 92.8
  10 10 Indiana 90.20 94.5 76.8 107.6 94.8 97.5 93.3
  11 11 Tennessee 90.30 94.4 84 94.7 91.2 87.8 93.3
  12 12 Arkansas 90.70 92.5 78.4 100.1 89.7 84.2 100.6
  13 13 Nebraska 91.10 96.9 81.6 88.8 102.1 99.3 94.2
  14 14 Wyoming 91.66 101.4 80.4 81.7 96.4 98.2 98.6
  15 15 Michigan 91.70 92.3 82.1 98.2 98.5 95.8 96.8
  16 16 Illinois 91.90 100.5 81.7 94.6 104.3 96.0 92.8
  17 17 Texas 92.60 91.0 84.8 105.9 91.9 94.8 96.8
  ... ...
  ng to 18 of 53 entries, 10 total columns

R 4.2.1 - ~/Downloads/
```

downloaded binary packages are in  
/var/folders/s0/135xdmz90gdgnhmm4r83qs80000gn/T//RtmpDBtxqa/downloaded\_packages

library(tidyverse)

Attaching packages --  
tidyverse 1.3.2 --

plot2 3.4.0 ✓ purrr 0.3.5
bbl 3.1.8 ✓ dplyr 1.0.10
dfr 1.2.1 ✓ stringr 1.5.0
adr 2.1.3 ✓ forcats 0.5.2

conflicts() --  
by::filter() masks stats::filter()  
lyr::lens() masks stats::log()
ta <- read.csv("Prajukta\_1.csv");
DataGrocery <- gsub("[()]","",Data\$Grocery)
ew(Data)
ta <- read.csv("Prajukta\_1.csv");
taHousing <- gsub("[()]","",Data\$Housing)
ew(Data)

## Smooth Data Housing

## For Utilities:

```
install.packages('tidyverse')
```

```
library(tidyverse)
```

```
Data <- read.csv("Prajukta_1.csv");
```

```
Data$Utilities <- gsub("[()]", "", Data$Utilities)
```

```
view(Data)
```

The screenshot shows the RStudio interface. The code editor on the left contains the following R script:

```
1 install.packages("tidyverse")
2 library(tidyverse)
3 
4 
5 
6 
7 
8 Data <- read.csv("Prajukta_1.csv");
9 Data$Utilities <- gsub("[()]", "", Data$Utilities)
10 view(Data)
```

The global environment on the right lists several objects:

- content
- data
- Data
- first\_table
- iris
- table

The `Data` object is shown in the preview pane as a list of 53 observations with 10 variables.

The screenshot shows the RStudio interface. The code editor on the left contains the same R script as the previous screenshot:

```
1 install.packages("tidyverse")
2 library(tidyverse)
3 
4 
5 
6 
7 
8 Data <- read.csv("Prajukta_1.csv");
9 Data$Utilities <- gsub("[()]", "", Data$Utilities)
10 view(Data)
```

The global environment on the right lists the same objects as before:

- content
- data
- Data
- first\_table
- iris
- table

The `Data` object is shown in the preview pane as a list of 53 observations with 10 variables.

## For Transportation:

```
install.packages('tidyverse')
```

```
library(tidyverse)
```

```
Data <- read.csv("Prajukta_1.csv");
```

```
Data$Transportation <- gsub("[()]", "", Data$Transportation)
```

```
view(Data)
```

The screenshot shows the RStudio interface. In the top-left pane, there is a code editor with the following R code:

```
install.packages('tidyverse')
library(tidyverse)

Data <- read.csv("Prajukta_1.csv");
Data$Transportation <- gsub("[()]", "", Data$Transportation)
view(Data)
```

In the top-right pane, the "Environment" tab is selected, showing the global environment with variables like `content`, `data`, `first\_table`, and `iris`. The `x` variable is also listed.

At the bottom right, there is a sidebar with links to R Resources and RStudio Support, and sections for Manuals, Reference, and Miscellaneous Material.

This screenshot is similar to the one above, but it includes a large data grid overlaid on the bottom half of the screen. The grid displays a table with columns: X, Rank, State, Index, Grocery, Housing, Utilities, Transportation, Health, Misc., Tr, and Data. The data consists of 53 entries from various US states, such as Mississippi, Oklahoma, Alabama, Kansas, Iowa, Georgia, Ohio, West Virginia, Missouri, Indiana, Tennessee, Arkansas, Nebraska, Wyoming, Michigan, Illinois, and Texas. The "Transportation" column contains values like 92.6, 91.4, 88.1, etc.

## For Health:

```
install.packages('tidyverse')
```

```
library(tidyverse)
```

```
Data <- read.csv("Prajukta_1.csv");
```

```
Data$Health <- gsub("[()]", "", Data$Health)
```

```
view(Data)
```

The screenshot shows the RStudio interface. The code editor window displays the following R code:

```
install.packages('tidyverse')
library(tidyverse)

Data <- read.csv("Prajukta_1.csv");
Data$Health <- gsub("[()]", "", Data$Health)
view(Data)

# In `data.frame`(*tmp*, Transportation, value = character(0)):
# placement has 0 rows, data has 53
Data$Transportation <- gsub("[()]", "", Data$Transportation)
ew(Data)
Data <- read.csv("Prajukta_1.csv");
Data$Utilities <- gsub("[()]", "", Data$Utilities)
ew(Data)
Data <- read.csv("Prajukta_1.csv");
Data$Transportation <- gsub("[()]", "", Data$Transportation)
ew(Data)
Data <- read.csv("Prajukta_1.csv");
Data$Health <- gsub("[()]", "", Data$Health)
ew(Data)
```

The global environment pane shows the following variables:

- content: List of 2
- data: 50 obs. of 4 variables
- Data: 53 obs. of 10 variables
- first\_table: 53 obs. of 9 variables
- iris: 150 obs. of 5 variables
- table: List of 2

The viewer pane shows the first few rows of the Data variable:

	Rank	State	Index	Grocery	Housing	Utilities	Transportation	Health	Misc.
1	1	Mississippi	84.50	91.9	68.5	88.5	92.6	99.5	91.2
2	2	Oklahoma	86.70	94.7	72.4	95.0	91.4	91.5	92.2
3	3	Alabama	87.10	98.2	68.2	100.2	88.1	89.5	95.2
4	4	Kansas	87.30	94.3	72.0	98.7	96.3	101.1	91.0
5	5	Iowa	88.20	98.0	70.6	95.5	95.4	98.9	94.9
6	6	Georgia	88.90	96.4	77.7	90.0	89.4	96.5	94.8
7	7	Ohio	89.40	100.8	71.1	92.4	95.6	94.4	98.5
8	8	West Virginia	89.80	96.3	76.9	91.7	104.7	97.1	94.0
9	9	Missouri	90.10	96.8	80.4	94.2	97.8	91.2	92.8
10	10	Indiana	90.20	94.5	76.8	107.6	94.8	97.5	93.3
11	11	Tennessee	90.30	94.4	84.0	94.7	91.2	87.8	93.3
12	12	Arkansas	90.70	92.5	78.4	100.1	89.7	84.2	100.6
13	13	Nebraska	91.10	96.9	81.6	88.8	102.1	99.3	94.2
14	14	Wyoming	91.66	101.4	80.4	81.7	96.4	98.2	98.6
15	15	Michigan	91.70	92.3	82.1	98.2	98.5	95.8	96.8
16	16	Illinois	91.90	100.5	81.7	94.6	104.3	96	92.8
17	17	Texas	92.60	91.0	84.8	105.9	91.9	94.8	96.8

The screenshot shows the RStudio interface. The code editor window displays the same R code as the previous screenshot, including the data frame manipulation and variable assignment. The global environment pane shows the same variables:

- content: List of 2
- data: 50 obs. of 4 variables
- Data: 53 obs. of 10 variables
- first\_table: 53 obs. of 9 variables
- iris: 150 obs. of 5 variables
- table: List of 2

## For Misc:

```
install.packages('tidyverse')
```

```
library(tidyverse)
```

```
Data <- read.csv("Prajukta_1.csv");
```

```
Data$Misc <- gsub("[()]", "", Data$Misc)
```

```
view(Data)
```

The screenshot shows the RStudio interface. The code editor at the top contains the following R code:

```
install.packages('tidyverse')
library(tidyverse)

Data <- read.csv("Prajukta_1.csv");
Data$Misc <- gsub("[()]", "", Data$Misc)
view(Data)
```

The environment pane on the right shows the following variables:

- content: List of 2
- data: 50 obs. of 4 variables
- First\_table: 53 obs. of 11 variables
- iris: 150 obs. of 5 variables
- table: List of 2

The values pane shows the first few rows of the 'x' variable:

	x
1	num [1:100] 0.352 0.451 0.421 -0.215 1.354 ...

The screenshot shows the RStudio interface with a data table overlaid on the code editor. The data table displays the 'Data' variable from the environment pane. The columns are labeled: X, Rank, State, Index, Grocery, Housing, Utilities, Transportation, Health, Misc., and M. The data shows 53 entries for US states.

X	Rank	State	Index	Grocery	Housing	Utilities	Transportation	Health	Misc.	M
1	1	Mississippi	84.50	91.9	68.5	88.5	92.6	99.5	91.2	91
2	2	Oklahoma	86.70	94.7	72.4	95.0	91.4	91.5	92.2	92
3	3	Alabama	87.10	98.2	68.2	100.2	88.1	89.5	95.2	95
4	4	Kansas	87.30	94.3	72.0	98.7	96.3	101.1	91.0	91
5	5	Iowa	88.20	98.0	70.6	95.5	95.4	98.9	94.9	94
6	6	Georgia	88.90	96.4	77.7	90.0	89.4	96.5	94.8	94
7	7	Ohio	89.40	100.8	71.1	92.4	95.6	94.4	98.5	95
8	8	West Virginia	89.80	96.3	76.9	91.7	104.7	97.1	94.0	94
9	9	Missouri	90.10	96.8	80.4	94.2	97.8	91.2	92.8	92
10	10	Indiana	90.20	94.5	76.8	107.6	94.8	97.5	93.3	93
11	11	Tennessee	90.30	94.4	84.0	94.7	91.2	87.8	93.3	93
12	12	Arkansas	90.70	92.5	78.4	100.1	89.7	84.2	100.6	10
13	13	Nebraska	91.10	96.9	81.6	88.8	102.1	99.3	94.2	94
14	14	Wyoming	91.66	101.4	80.4	81.7	96.4	98.2	98.6	95
15	15	Michigan	91.70	92.3	82.1	98.2	98.5	95.8	96.8	96
16	16	Illinois	91.90	100.5	81.7	94.6	104.3	96.0	92.8	92
17	17	Texas	92.60	91.0	84.8	105.9	91.9	94.8	96.8	96
...	...	...	...	...	...	...	...	...	...	...

The environment pane on the right shows the following variables:

- content: List of 2
- data: 50 obs. of 4 variables
- Data: 53 obs. of 11 variables
- First\_table: 53 obs. of 9 variables
- iris: 150 obs. of 5 variables
- table: List of 2

The values pane shows the first few rows of the 'x' variable:

	x
1	num [1:100] 0.352 0.451 0.421 -0.215 1.354 ...

## **Data Integration:**

No Data Integration is necessary for the data set.

## **Data Transformation:**

As is well known, the data transformation process includes one or more of the following steps: normalization, summarization, noise removal, smoothing, and noise removal from data.

We shall use smoothing in this example because it is easier than summarizing and normalizing.

For Index:

```
dataset$Index = as.numeric(format(round(dataset$Index,0)))
glimpse(dataset)
```

For Grocery:

```
dataset$Grocery = as.numeric(format(round(dataset$Grocery,0)))
glimpse(dataset)
```

For Housing:

```
dataset$Housing = as.numeric(format(round(dataset$Housing,0)))
glimpse(dataset)
```

For Utilites:

```
dataset$Utilites = as.numeric(format(round(dataset$Utilites,0)))
glimpse(dataset)
```

For Transportation:

```
dataset$Transportation =
as.numeric(format(round(dataset$Transportation,0))) glimpse(dataset)
```

For Health:

```
dataset$Health = as.numeric(format(round(dataset$Health,0)))
glimpse(dataset)
```

For Misc:

```
dataset$Misc = as.numeric(format(round(dataset$Misc,0)))
glimpse(dataset)
```

## **Data Reduction:**

Data reduction aims to produce a reduced representation of the dataset that may be utilized to produce the same or comparable analytical results is what data reduction aims to produce.

For Index:

```
dataset$Index<-as.numeric(format(round(dataset$Index, 0)))
```

For Grocery:

```
dataset$Grocery<-as.numeric(format(round(dataset$Grocery, 0)))
```

For Housing:

```
dataset$Housing<-as.numeric(format(round(dataset$Housing, 0)))
```

For Utilites:

```
dataset$Utilites<-as.numeric(format(round(dataset$Utilites, 0)))
```

For Transportation:

```
dataset$Transportation<-as.numeric(format(round(dataset$Transportation, 0)))
```

For Health:

```
dataset$Health<-as.numeric(format(round(dataset$Health, 0)))
```

For Misc:

```
dataset$Misc<-as.numeric(format(round(dataset$Misc, 0)))
```

## Data Discretization

All of the attributes included in our dataset are continuous type, as can be seen (values in real numbers). The attribute values may need to be discretized into binary or categorical categories, though, depending on the model you wish to create.

# Descripting statistics

**Compute the mean value of Index:**

```
meanIndex= mean(Prajukta_1$Index)  
print(meanIndex)
```

**Compute the mean value of Grocery:**

```
meanGrocery= mean(Prajukta_1$Grocery)  
print(meanGrocery)
```

**Compute the mean value of Housing:**

```
meanHousing= mean(Prajukta_1$Housing)  
print(meanHousing)
```

**Compute the mean value of Utilites:**

```
meanUtilites= mean(Prajukta_1$Utilites)  
print(meanUtilites)
```

**Compute the mean value of Transportation:**

```
meanTransportation= mean(Prajukta_1$Transportation)  
print(meanTransportation)
```

**Compute the mean value of Health:**

```
meanHealth= mean(Prajukta_1$Health)  
print(meanHealth)
```

**Compute the mean value of Misc:**

```
meanMisc= mean(Prajukta_1$Misc)  
print(meanMisc)
```

# Data visualization:

For index

```
library(ggplot2)
library(mosaicData)

ggplot(data = CPS85, mapping = aes(x = State, y = Index))

library(mosaicData)
ggplot(data = CPS85, mapping = aes(x = State, y = Index)) + geom_point()
```

For Grocery

```
library(ggplot2)
library(mosaicData)

ggplot(data = CPS85, mapping = aes(x = State, y = Grocery))

library(mosaicData)
ggplot(data = CPS85, mapping = aes(x = State, y = Grocery)) + geom_point()
```

For Utilities

```
library(ggplot2)
library(mosaicData)

ggplot(data = CPS85, mapping = aes(x = State, y = Utilities))

library(mosaicData)
ggplot(data = CPS85, mapping = aes(x = State, y = Grocery)) + geom_point()
```

For Housing

```
library(ggplot2)
library(mosaicData)

ggplot(data = CPS85, mapping = aes(x = State, y = Housing))

library(mosaicData)
```

```
ggplot(data = CPS85, mapping = aes(x = State, y = Grocery)) + geom_point()
```

## For Transportation

```
library(ggplot2)
library(mosaicData)

ggplot(data = CPS85, mapping = aes(x = State, y = Transportation)) + geom_point()

library(mosaicData)
ggplot(data = CPS85, mapping = aes(x = State, y = Grocery)) + geom_point()
```

## For Health

```
library(ggplot2)
library(mosaicData)

ggplot(data = CPS85, mapping = aes(x = State, y = Health)) + geom_point()

library(mosaicData)
ggplot(data = CPS85, mapping = aes(x = State, y = Grocery)) + geom_point()
```

## For Misc

```
library(ggplot2)
library(mosaicData)

ggplot(data = CPS85, mapping = aes(x = State, y = Misc)) + geom_point()

library(mosaicData)
ggplot(data = CPS85, mapping = aes(x = State, y = Grocery)) + geom_point()
```



## **Discussion and Conclusion:**

In order to offer the necessary dataset, pre-processing techniques are used, and all missing data and outliers are found. without data pre-processing we can not get a perfect analysis result. We need to do web scraping and then we can do pre-processing , Descripting statistics and Data visualization.

















