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)

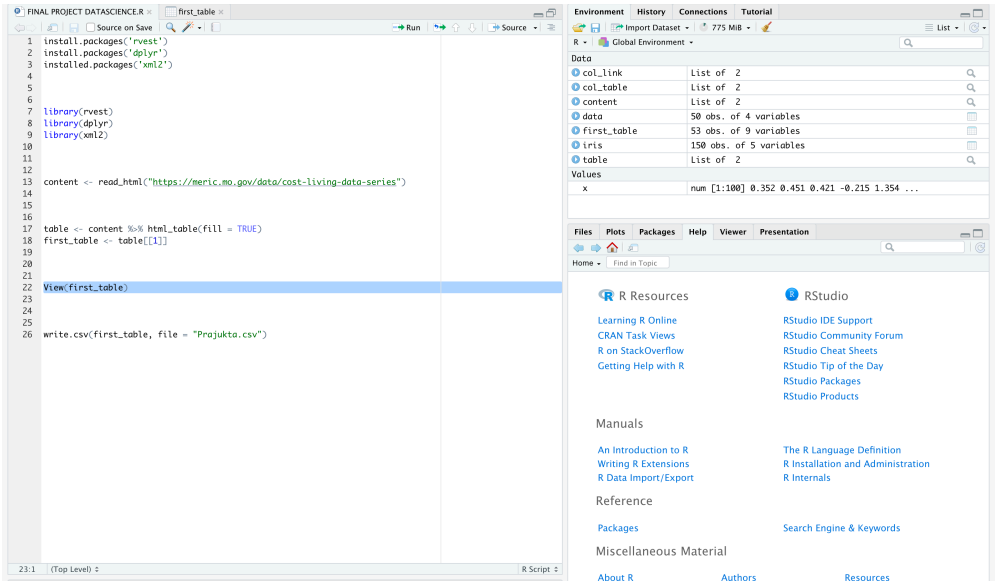
content <- read_html("https://meric.mo.gov/data/cost-living-data-series")

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

first_table <- table[[1]]

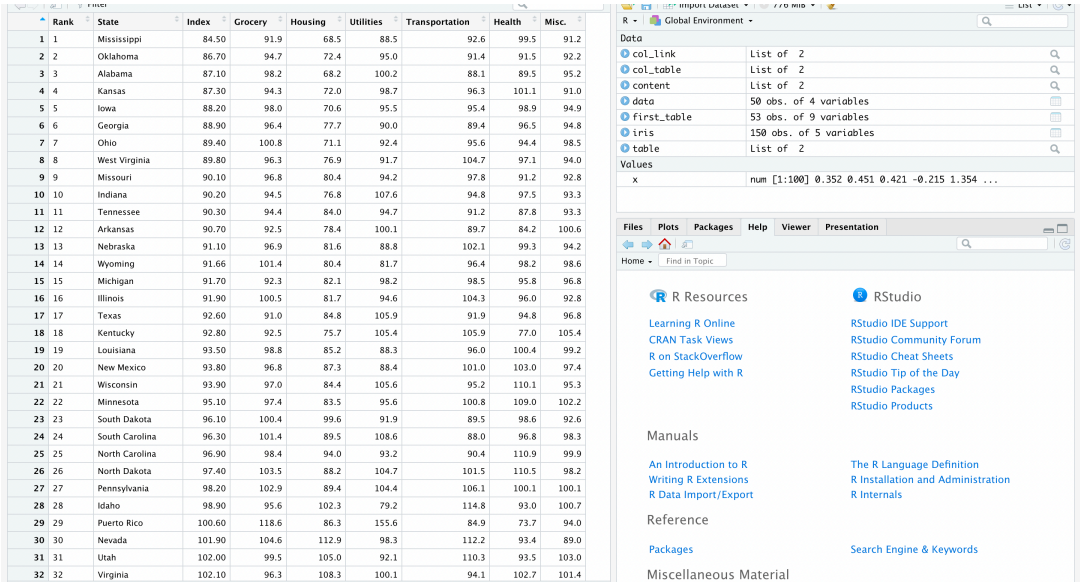
View(first_table)

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



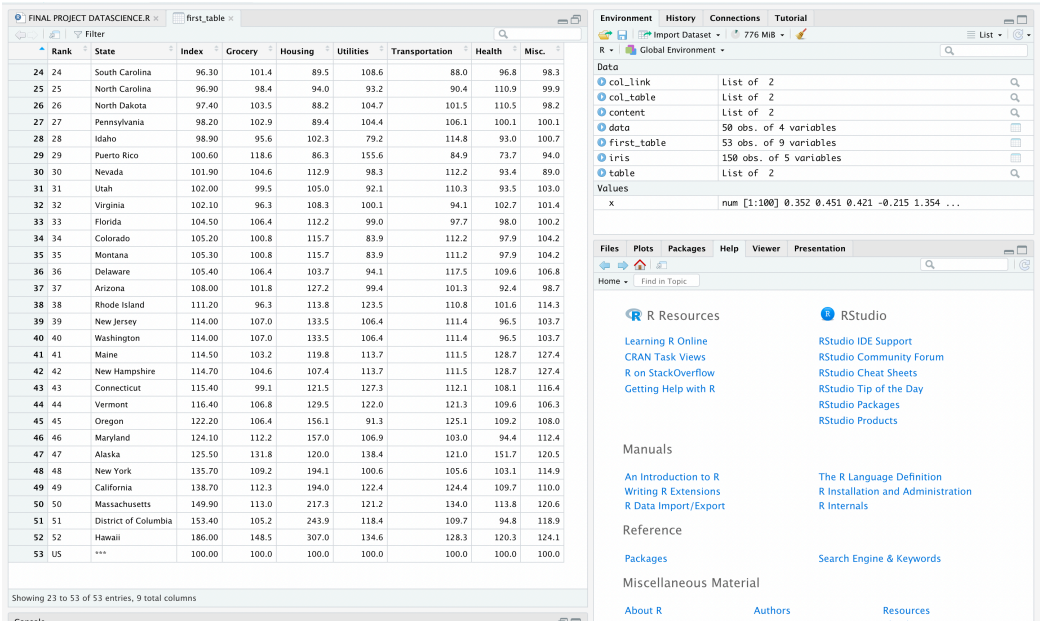
View table

DATA TABLE



	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	96.0	92.8
17	17	Texas	92.60	91.0	84.8	105.9	91.9	94.8	96.8
18	18	Kentucky	92.80	92.5	75.7	105.4	105.9	77.0	105.4
19	19	Louisiana	93.50	98.8	85.2	88.3	96.0	100.4	99.2
20	20	New Mexico	93.80	96.8	87.3	88.4	101.0	103.0	97.4
21	21	Wisconsin	93.90	97.0	84.4	105.6	95.2	110.1	95.3
22	22	Minnesota	95.10	97.4	83.5	95.6	100.8	109.0	102.2
23	23	South Dakota	96.10	100.4	99.6	91.9	89.5	98.6	92.6
24	24	South Carolina	96.30	101.4	89.5	108.6	88.0	96.8	98.3
25	25	North Carolina	96.90	98.4	94.0	93.2	90.4	110.9	99.9
26	26	North Dakota	97.40	103.5	88.2	104.7	101.5	110.5	98.2
27	27	Pennsylvania	98.20	102.9	89.4	104.4	106.1	100.1	100.1
28	28	Idaho	98.90	95.6	102.3	79.2	114.8	93.0	100.7
29	29	Puerto Rico	100.60	118.6	86.3	155.6	84.9	73.7	94.0
30	30	Nevada	101.90	104.6	112.9	98.3	112.2	93.4	89.0
31	31	Utah	102.00	99.5	105.0	92.1	110.3	93.5	103.0
32	32	Virginia	102.10	96.3	108.3	100.1	94.1	102.7	101.4

Data table

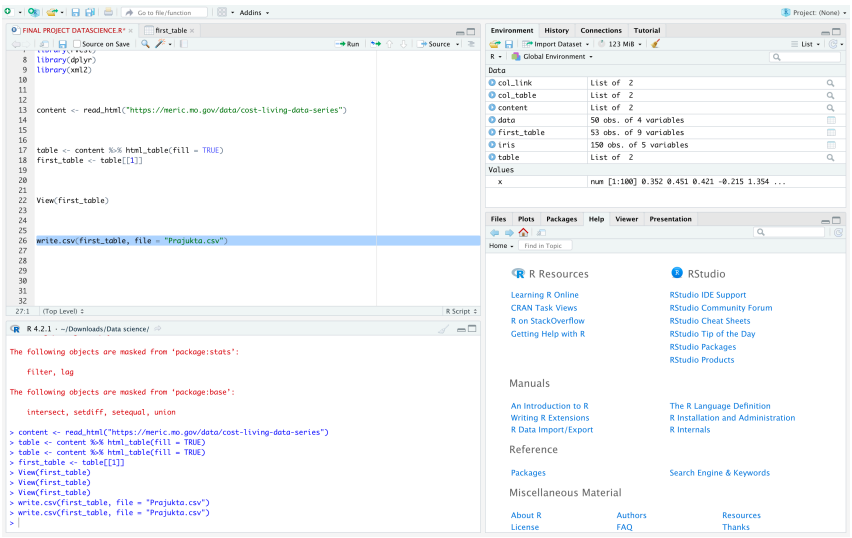


	Rank	State	Index	Grocery	Housing	Utilities	Transportation	Health	Misc.
24	24	South Carolina	96.30	101.4	89.5	108.6	88.0	96.8	98.3
25	25	North Carolina	96.90	98.4	94.0	93.2	90.4	110.9	99.9
26	26	North Dakota	97.40	103.5	88.2	104.7	101.5	110.5	98.2
27	27	Pennsylvania	98.20	102.9	89.4	104.4	106.1	100.1	100.1
28	28	Idaho	98.90	95.6	102.3	79.2	114.8	93.0	100.7
29	29	Puerto Rico	100.60	118.6	86.3	155.6	84.9	73.7	94.0
30	30	Nevada	101.90	104.6	112.9	98.3	112.2	93.4	89.0
31	31	Utah	102.00	99.5	105.0	92.1	110.3	93.5	103.0
32	32	Virginia	102.10	96.3	108.3	100.1	94.1	102.7	101.4
33	33	Florida	104.50	106.4	112.2	99.0	97.7	98.0	100.2
34	34	Colorado	105.20	100.8	115.7	83.9	112.2	97.9	104.2
35	35	Montana	105.30	100.8	115.7	83.9	111.2	97.9	104.2
36	36	Delaware	105.40	106.4	103.7	94.1	117.5	109.6	106.8
37	37	Arizona	108.00	101.8	127.2	99.4	101.3	92.4	98.7
38	38	Rhode Island	111.20	96.3	113.8	123.5	110.8	101.6	114.3
39	39	New Jersey	114.00	107.0	133.5	106.4	111.4	96.5	103.7
40	40	Washington	114.00	107.0	133.5	106.4	111.4	96.5	103.7
41	41	Maine	114.50	103.2	119.8	113.7	111.5	128.7	127.4
42	42	New Hampshire	114.70	104.6	107.4	113.7	111.5	128.7	127.4
43	43	Connecticut	115.40	99.1	121.5	127.3	112.1	108.1	116.4
44	44	Vermont	116.40	106.8	129.5	122.0	121.3	109.6	106.3
45	45	Oregon	122.20	106.4	156.1	91.3	125.1	109.2	108.0
46	46	Maryland	124.10	112.2	157.0	106.9	103.0	94.4	112.4
47	47	Alaska	125.50	131.8	120.0	138.4	121.0	151.7	120.5
48	48	New York	135.70	109.2	194.1	100.6	105.6	103.1	114.9
49	49	California	138.70	112.3	194.0	122.4	124.4	109.7	110.0
50	50	Massachusetts	149.90	113.0	217.3	121.2	134.0	113.8	120.6
51	51	District of Columbia	153.40	105.2	243.9	118.4	109.7	94.8	118.9
52	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	100.0

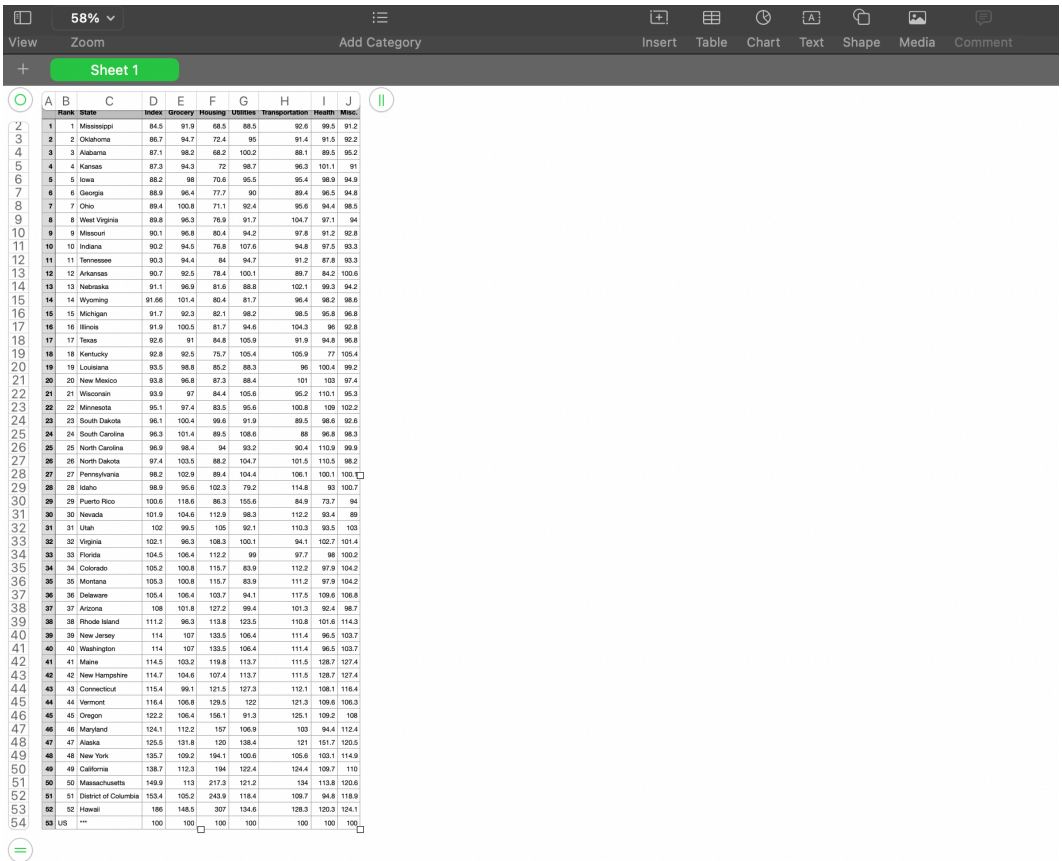
Showing 23 to 53 of 53 entries, 9 total columns

Data table

CSV FILE:



Run csv file

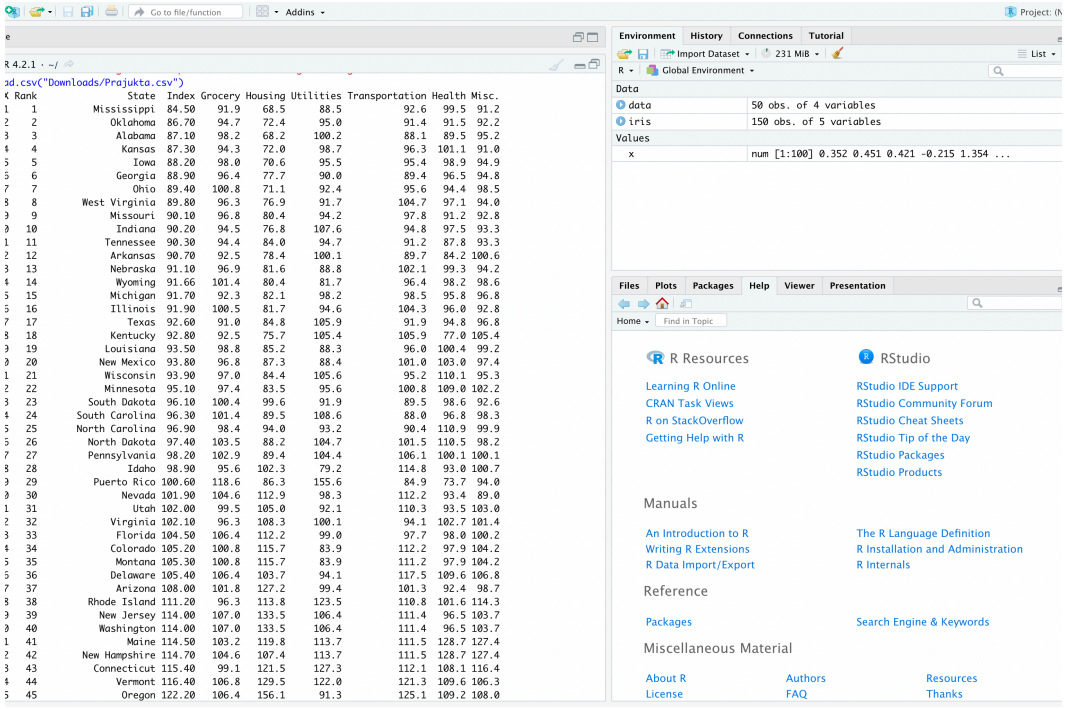


	Rank	State	Index	Grocery	Housing	Utilities	Transportation	Health	Misc.
1	1	Mississippi	84.5	91.9	68.5	88.5	92.6	99.5	91.2
2	2	Oklahoma	86.7	94.7	72.4	95	91.4	91.5	92.2
3	3	Alabama	87.1	98.2	68.2	100.2	88.1	89.5	95.2
4	4	Kansas	87.3	94.3	72	98.7	96.3	101.1	91
5	5	Iowa	88.2	98	70.6	95.5	95.4	98.9	94.9
6	6	Georgia	88.9	96.4	77.7	90	89.4	96.5	94.8
7	7	Ohio	89.4	100.8	71.1	92.4	95.6	94.4	98.5
8	8	West Virginia	89.8	96.3	76.9	91.7	104.7	97.1	94
9	9	Missouri	90.1	96.8	80.4	94.2	97.8	91.2	92.8
10	10	Indiana	90.2	94.5	76.8	107.6	94.8	97.5	93.3
11	11	Tennessee	90.3	94.4	84	94.7	91.2	87.8	93.3
12	12	Arkansas	90.7	92.5	78.4	100.1	89.7	84.2	100.6
13	13	Nebraska	91.1	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	91.7	92.3	82.1	98.2	98.5	95.8	96.8
16	16	Illinois	91.9	100.5	81.7	94.6	104.3	96	92.8
17	17	Texas	92.6	91	84.8	105.9	91.9	94.8	96.8
18	18	Kentucky	92.8	92.5	75.7	105.4	105.9	77	105.4
19	19	Louisiana	93.5	98.8	85.2	88.3	96	100.4	99.2
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	109.2	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
53	US	""	100	100	100	100	100	100	100

CSV File created

Read CSV file:

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



```
ad.csv("Downloads/Prajukta.csv")
  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   96.0  92.8
17  17          Texas  92.60    91.0    84.8     105.9           91.9   94.8  96.8
18  18       Kentucky  92.80    92.5    75.7     105.4          105.9   77.0 105.4
19  19      Louisiana  93.50    98.8    85.2      88.3           96.0  100.4  99.2
20  20     New Mexico  93.80    96.8    87.3      88.4          101.0  103.0  97.4
21  21      Wisconsin  93.90    97.0    84.4     105.6           95.2  110.1  95.3
22  22      Minnesota  95.10    97.4    83.5      95.6          100.8  109.0 102.2
23  23   South Dakota  96.10   100.4    99.6      91.9           89.5   98.6  92.6
24  24 South Carolina  96.30   101.4    89.5     108.6           88.0   96.8  98.3
25  25 North Carolina  96.90    98.4    94.0      93.2           90.4  110.9  99.9
26  26   North Dakota  97.40   103.5    88.2     104.7          101.5  110.5  98.2
27  27   Pennsylvania  98.20   102.9    89.4     104.4          106.1  100.1 100.1
28  28          Idaho  98.90    95.6   102.3      79.2          114.8   93.0 100.7
29  29    Puerto Rico 100.60   118.6    86.3     155.6           84.9   73.7  94.0
30  30         Nevada 101.90   104.6   112.9      98.3          112.2   93.4  89.0
31  31           Utah 102.00    99.5   105.0      92.1          110.3   93.5 103.0
32  32       Virginia 102.10    96.3   108.3     100.1           94.1  102.7 101.4
33  33        Florida 104.50   106.4   112.2      99.0           97.7   98.0 100.2
34  34       Colorado 105.20   100.8   115.7      83.9          112.2   97.9 104.2
35  35        Montana 105.30   100.8   115.7      83.9          111.2   97.9 104.2
36  36       Delaware 105.40   106.4   103.7      94.1          117.5  109.6 106.8
37  37        Arizona 108.00   101.8   127.2      99.4          101.3   92.4  98.7
38  38   Rhode Island 111.20    96.3   113.8     123.5          110.8  101.6 114.3
39  39     New Jersey 114.00   107.0   133.5     106.4          111.4   96.5 103.7
40  40     Washington 114.00   107.0   133.5     106.4          111.4   96.5 103.7
41  41          Maine 114.50   103.2   119.8     113.7          111.5  128.7 127.4
42  42  New Hampshire 114.70   104.6   107.4     113.7          111.5  128.7 127.4
43  43    Connecticut 115.40    99.1   121.5     127.3          112.1  108.1 116.4
44  44        Vermont 116.40   106.8   129.5     122.0          121.3  109.6 106.3
45  45         Oregon 122.20   106.4   156.1      91.3          125.1  109.2 108.0
```

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)

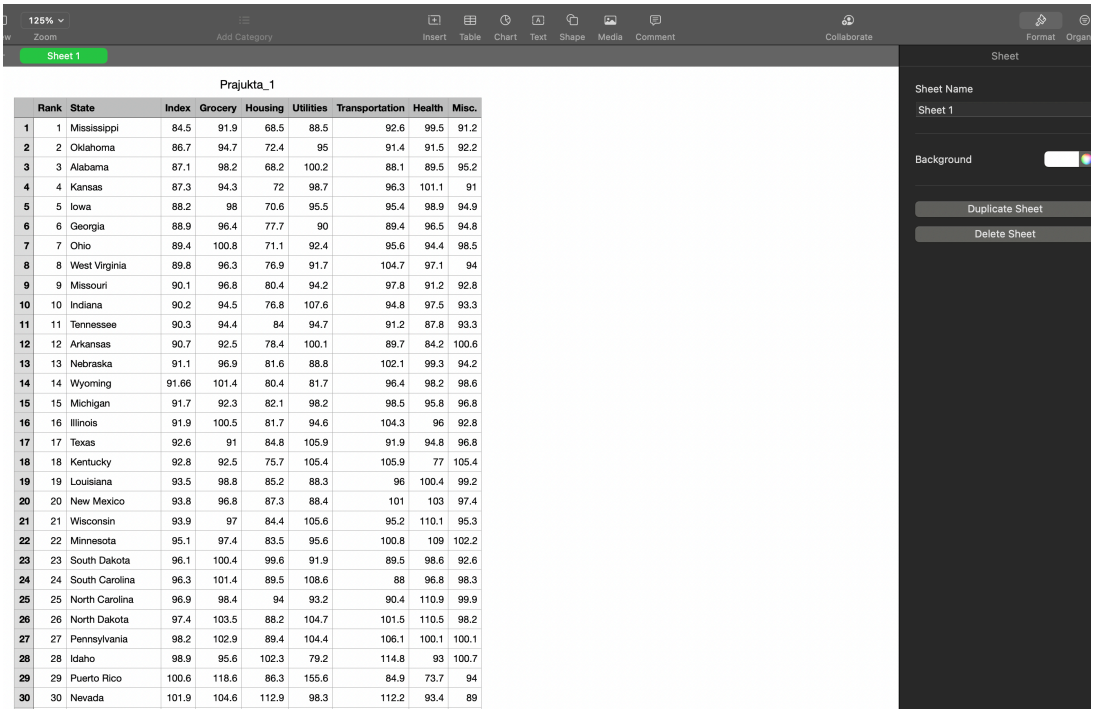
>mean(data$Housing, na.rm=TRUE)

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

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

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

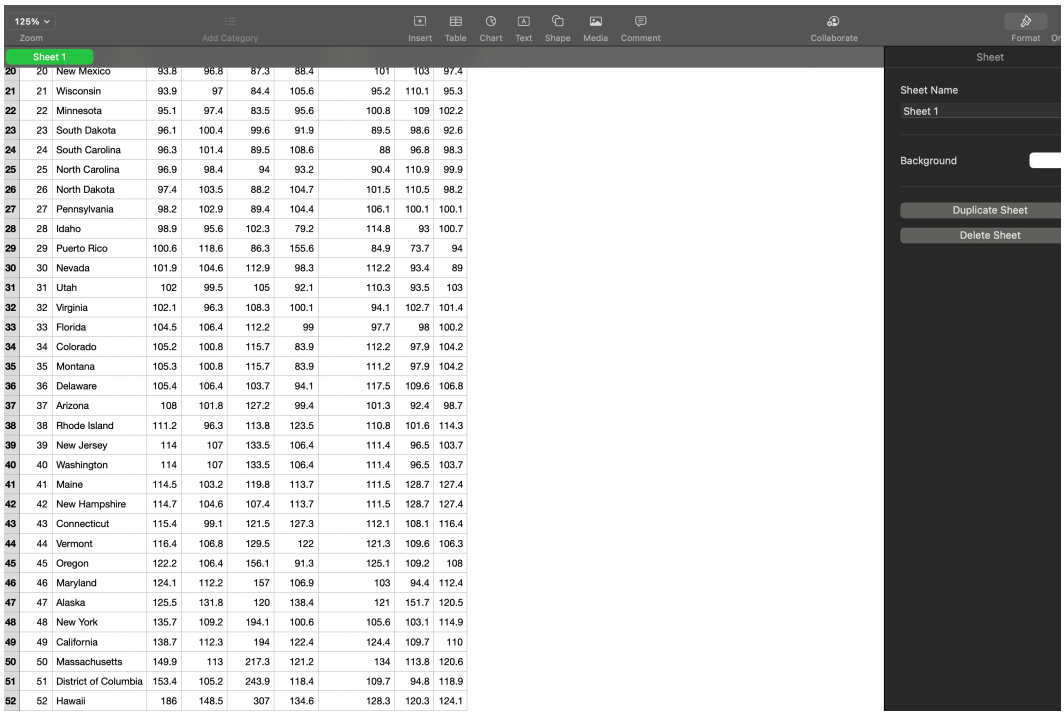
>mean(data$Misc, na.rm=TRUE)



Prajukta_1

	Rank	State	Index	Grocery	Housing	Utilities	Transportation	Health	Misc.
1	1	Mississippi	84.5	91.9	68.5	88.5	92.6	99.5	91.2
2	2	Oklahoma	86.7	94.7	72.4	95	91.4	91.5	92.2
3	3	Alabama	87.1	98.2	68.2	100.2	88.1	89.5	95.2
4	4	Kansas	87.3	94.3	72	98.7	96.3	101.1	91
5	5	Iowa	88.2	98	70.6	95.5	95.4	98.9	94.9
6	6	Georgia	88.9	96.4	77.7	90	89.4	96.5	94.8
7	7	Ohio	89.4	100.8	71.1	92.4	95.6	94.4	98.5
8	8	West Virginia	89.8	96.3	76.9	91.7	104.7	97.1	94
9	9	Missouri	90.1	96.8	80.4	94.2	97.8	91.2	92.8
10	10	Indiana	90.2	94.5	76.8	107.6	94.8	97.5	93.3
11	11	Tennessee	90.3	94.4	84	94.7	91.2	87.8	93.3
12	12	Arkansas	90.7	92.5	78.4	100.1	89.7	84.2	100.6
13	13	Nebraska	91.1	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	91.7	92.3	82.1	98.2	98.5	95.8	96.8
16	16	Illinois	91.9	100.5	81.7	94.6	104.3	96	92.8
17	17	Texas	92.6	91	84.8	105.9	91.9	94.8	96.8
18	18	Kentucky	92.8	92.5	75.7	105.4	105.9	77	105.4
19	19	Louisiana	93.5	98.8	85.2	88.3	96	100.4	99.2
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

Handling missing data



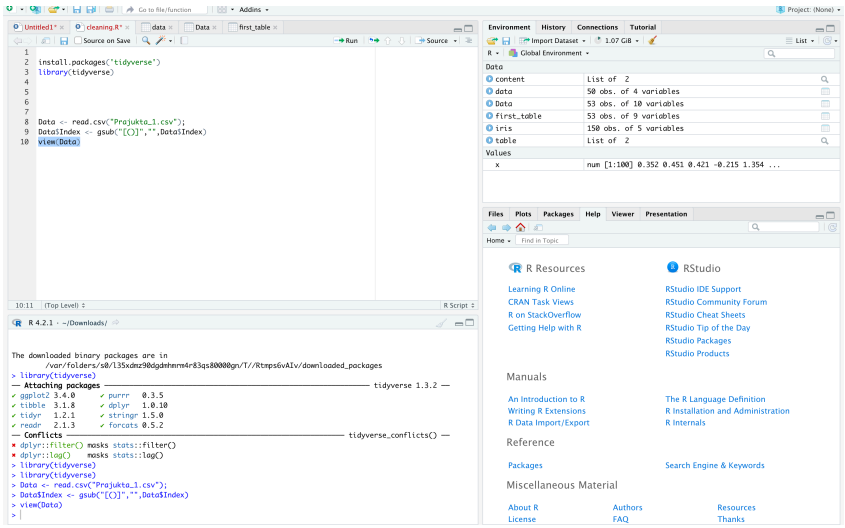
20	20	New Mexico	93.8	96.8	87.3	88.4	101	103	97.4
21	21	Wisconsin	93.9	97	84.4	105.6	95.2	110.1	95.3
22	22	Minnesota	95.1	97.4	83.5	95.6	100.8	109	102.2
23	23	South Dakota	96.1	100.4	99.6	91.9	89.5	98.6	92.6
24	24	South Carolina	96.3	101.4	89.5	108.6	88	96.8	98.3
25	25	North Carolina	96.9	98.4	94	93.2	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	109.2	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

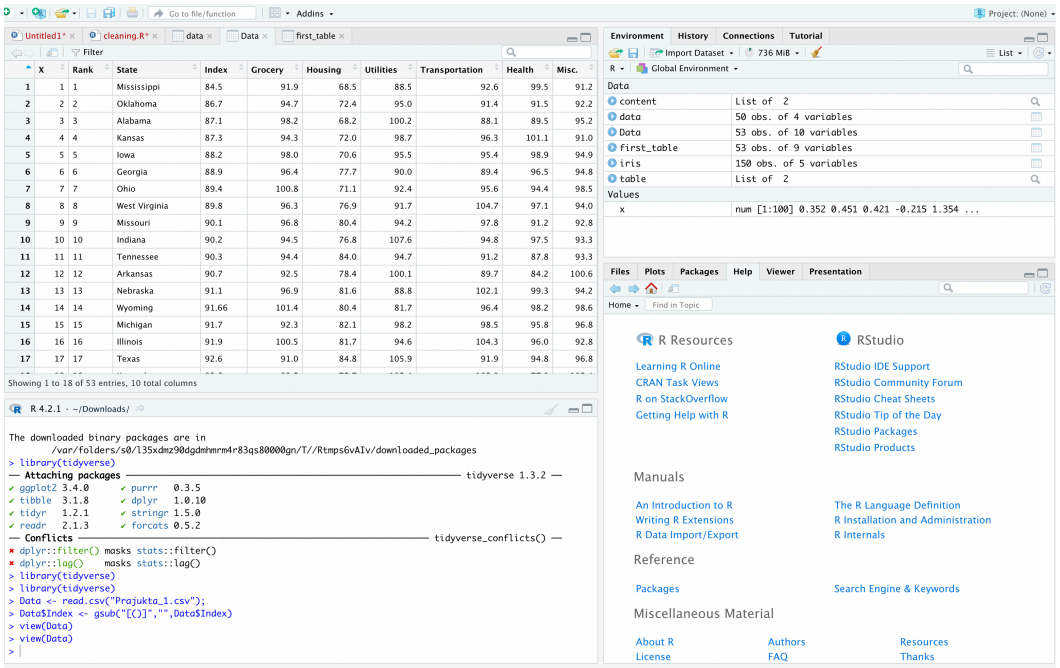
Smooth Noisy Data:

For Index

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



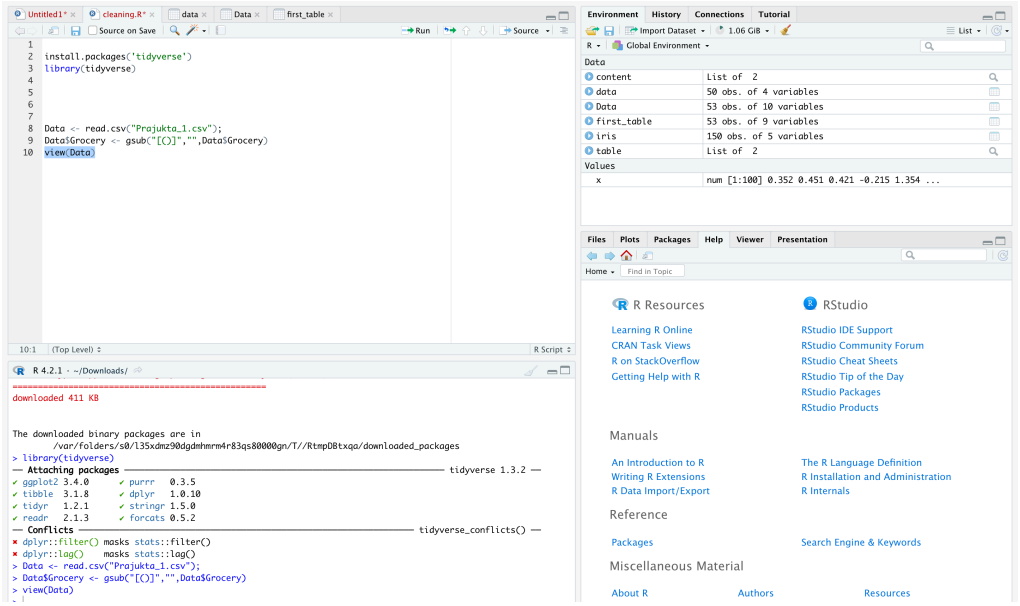
Smooth Noisy Data for Index



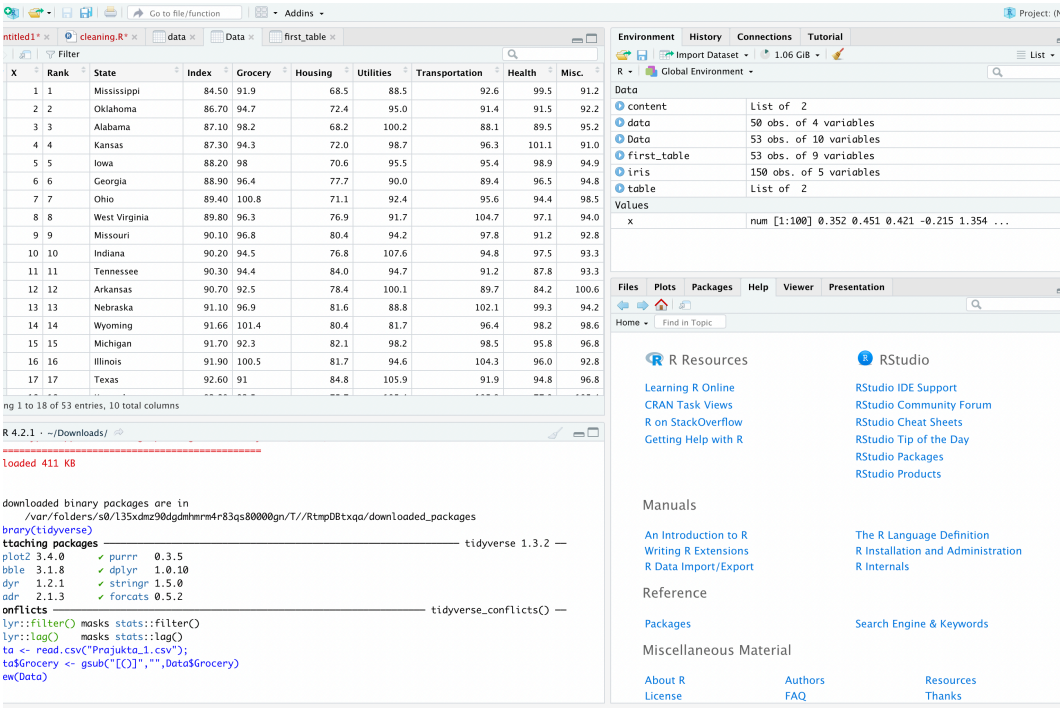
Smooth Noisy Data table for Index

For Grocery:

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



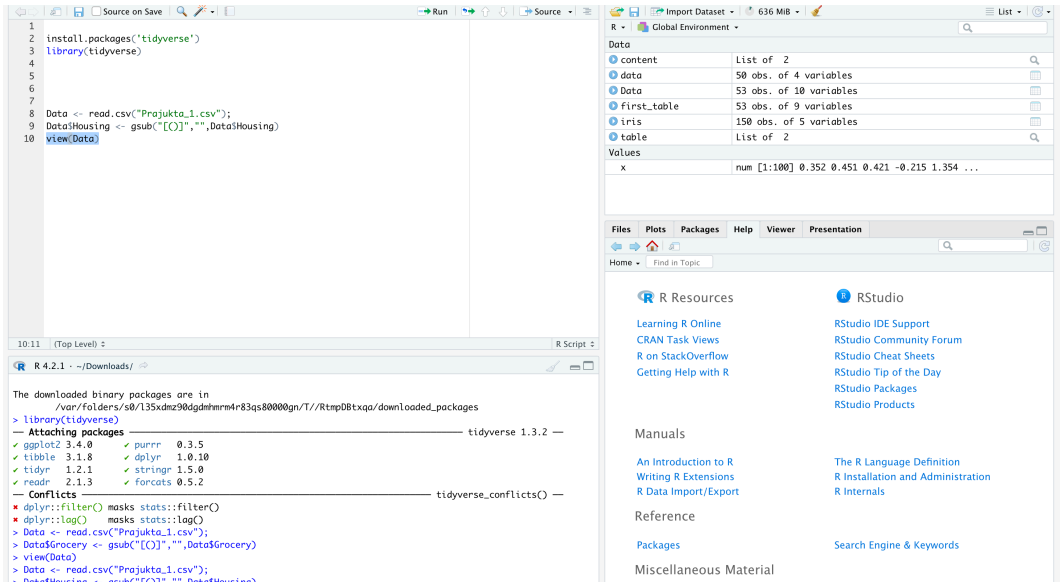
Smooth Noisy Data for Grocery



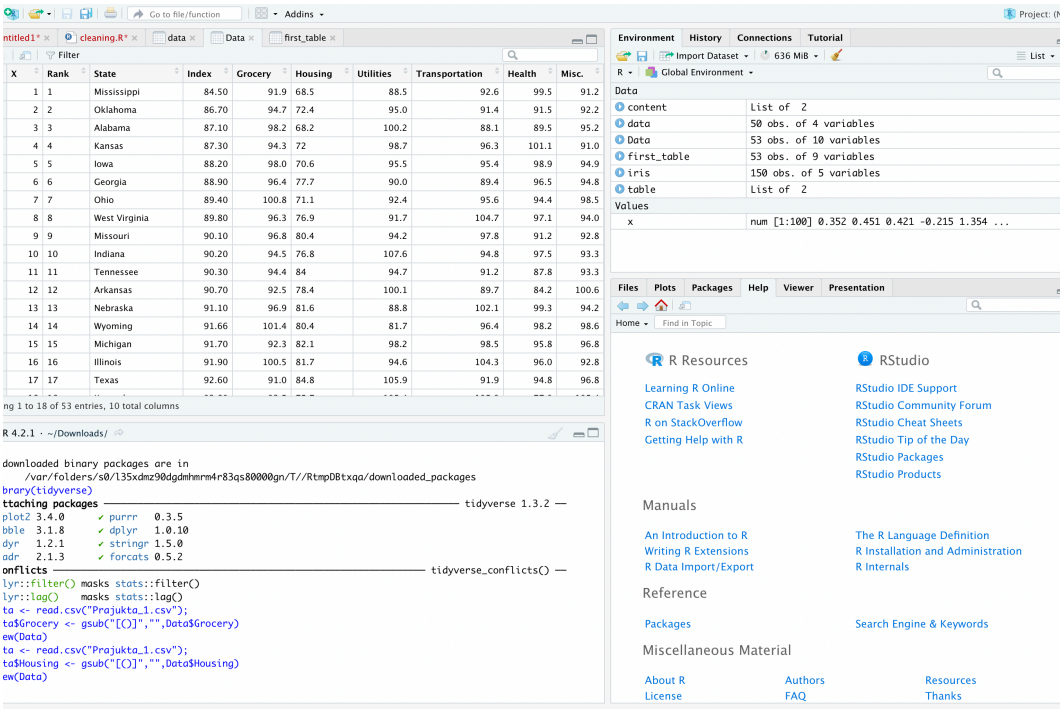
Smooth Noisy Data table for Grocery

For Housing:

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



Smooth data for Housing



Smooth Data Housing

For Utilities:

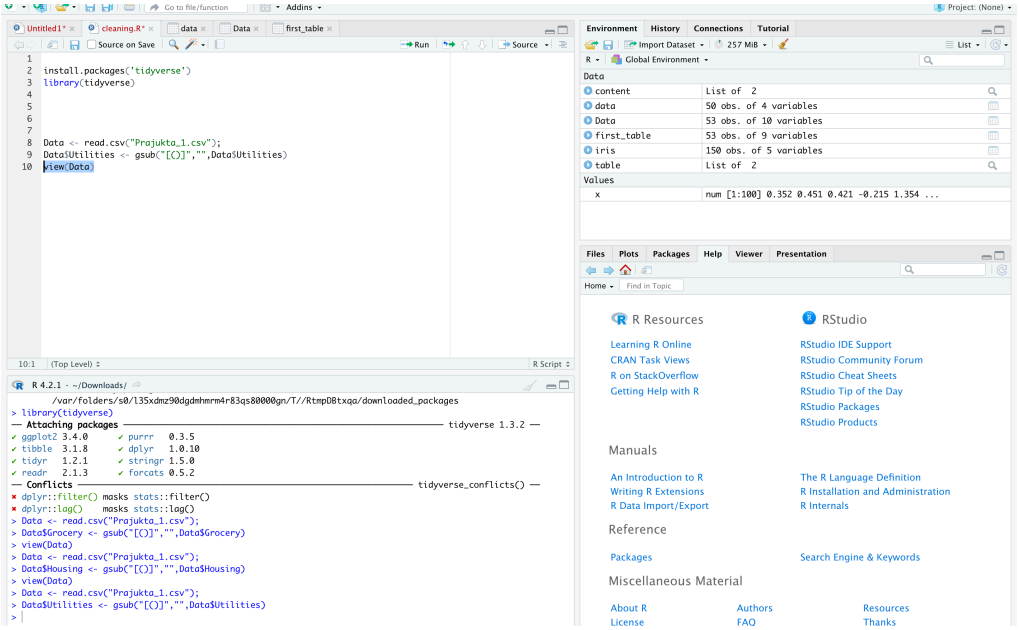
install.packages('tidyverse')

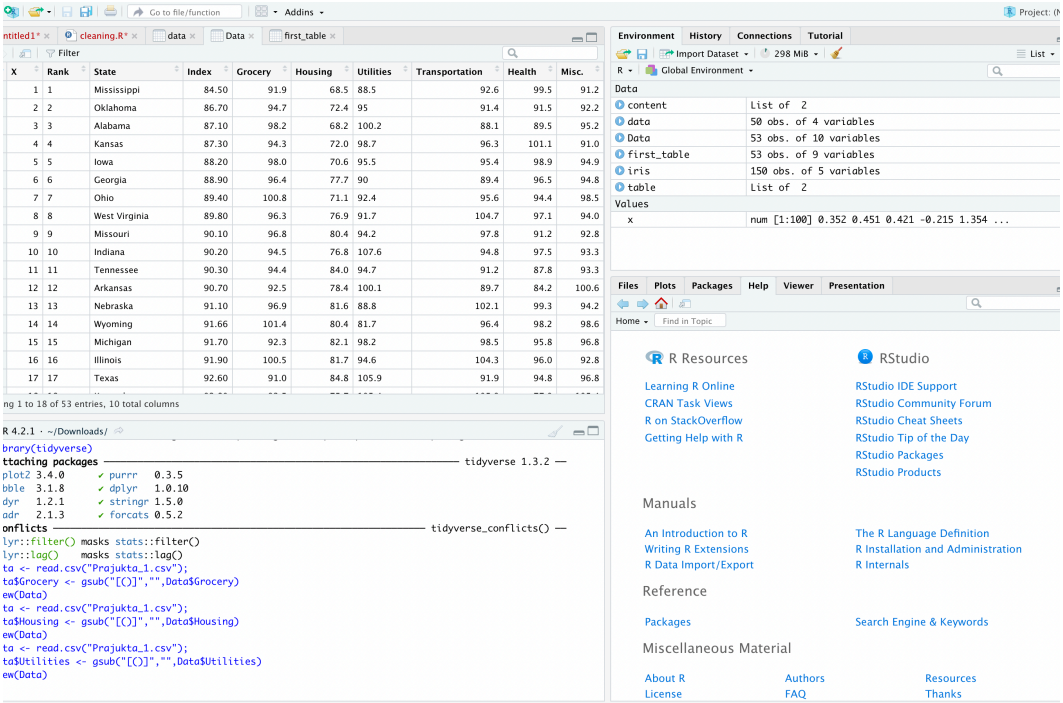
library(tidyverse)

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

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

view(Data)





For Transportation:

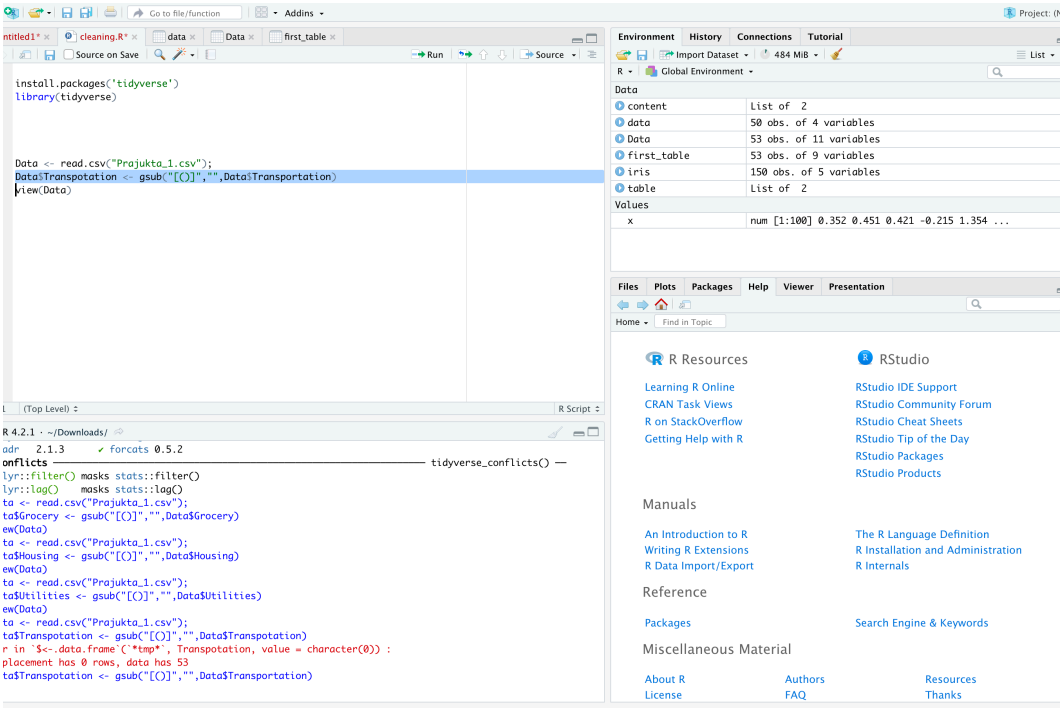
install.packages('tidyverse')

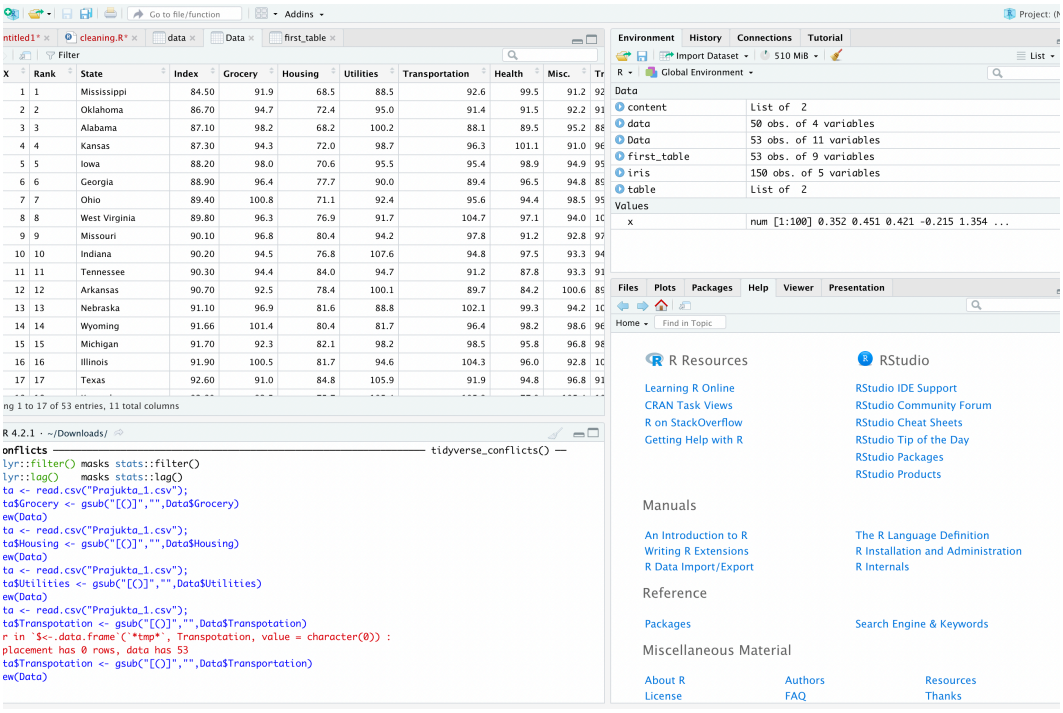
library(tidyverse)

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

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

view(Data)





For Health:

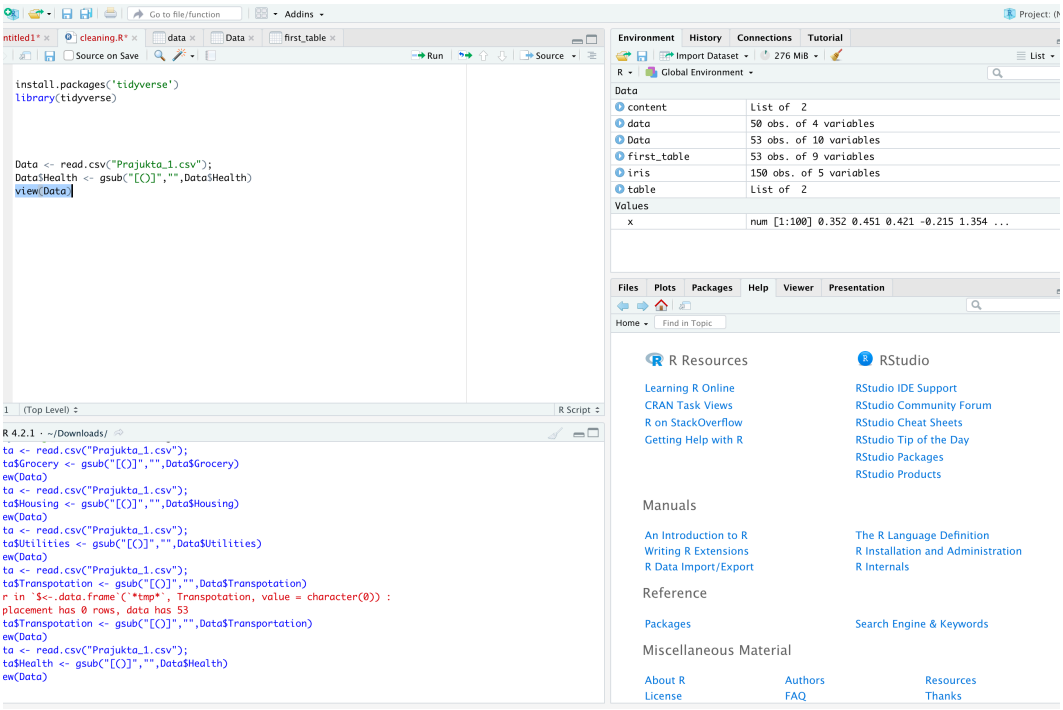
install.packages('tidyverse')

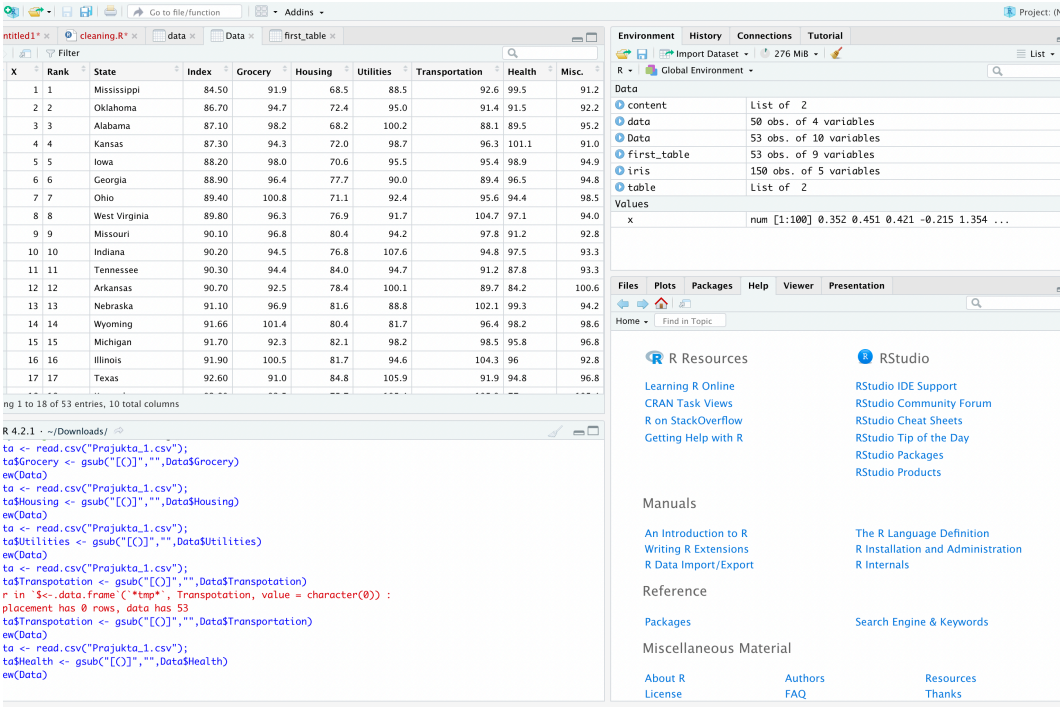
library(tidyverse)

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

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

view(Data)





For Misc:

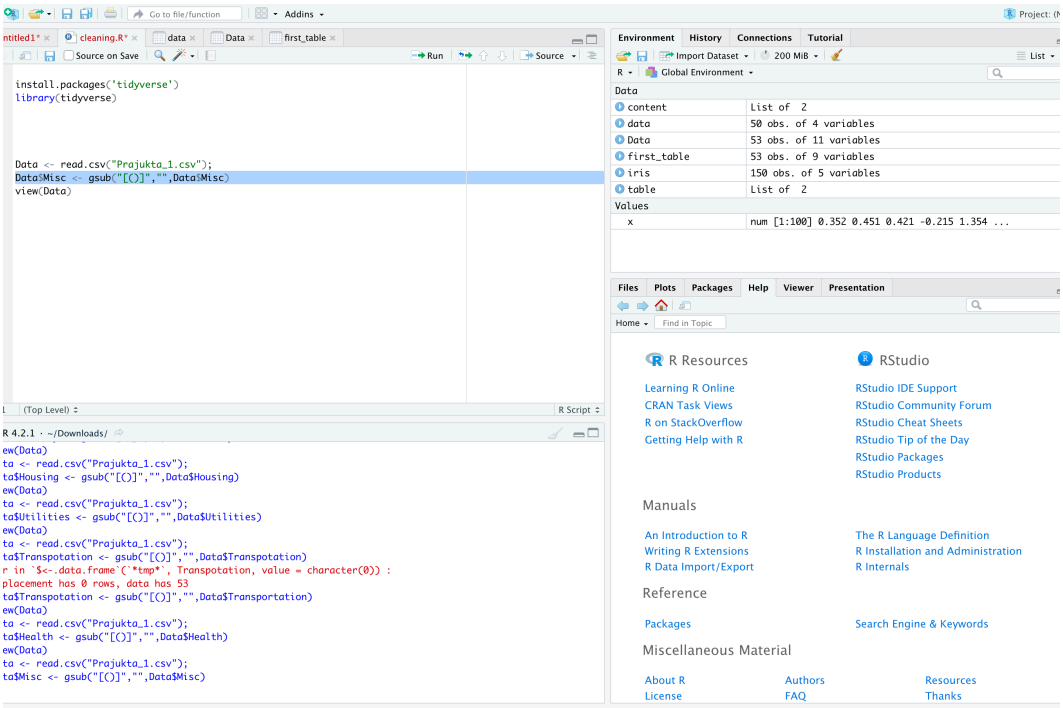
install.packages('tidyverse')

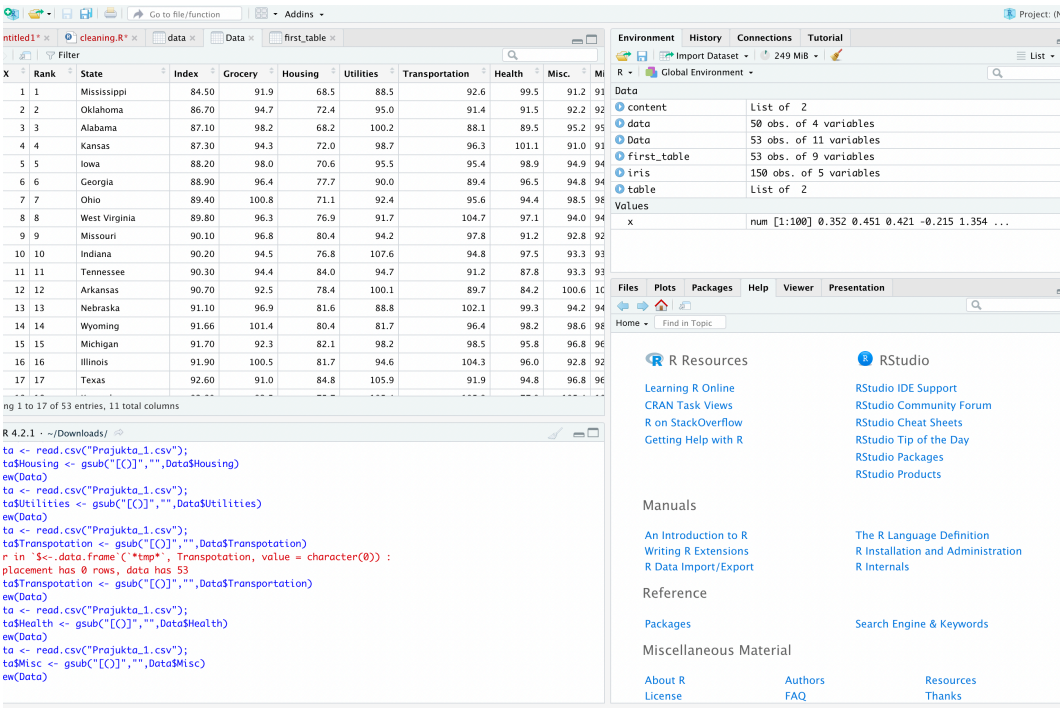
library(tidyverse)

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

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

view(Data)





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))
```

```
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))
```

```
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))
```

```
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.