Exploratory Data Analysis is one of the critical processes of performing initial investigations on data analysis.

Basic idea is to discover the patterns, anomalies, test hypotheses, and check the assumptions with the help of summary statistics and graphical representations.

The main idea about exploratory data analysis are

1. maximize insight into a data set;
2. uncover underlying structure;
3. extract important variables;
4. detect outliers and anomalies;
5. test underlying assumptions;
6. develop parsimonious models;
7. determine optimal factor settings.

**Exploratory Data Analysis**

The dataset we are using for exploratory data analysis is from gss\_cat.

gss\_cat

The datasets contains 21,483 observations and 9 variables.

# A tibble: 21,483 x 9

    year marital      age race  rincome

 1  2000 Never mar~    26 White $8000 to~

 2  2000 Divorced      48 White $8000 to~

 3  2000 Widowed       67 White Not appl~

 4  2000 Never mar~    39 White Not appl~

 5  2000 Divorced      25 White Not appl~

 6  2000 Married       25 White $20000 -~

 7  2000 Never mar~    36 White $25000 o~

 8  2000 Divorced      44 White $7000 to~

 9  2000 Married       44 White $25000 o~

10  2000 Married       47 White $25000 o~

**Dataset Overview**

gss\_cat %>% glimpse()

Above function will provide basic idea about variables in the datasets.

Rows: 21,483

Columns: 9

$ year    2000, 2000, 2000, 2000,~

$ marital Never married, Divorced~

$ age     26, 48, 67, 39, 25, 25,~

$ race    White, White, White, Wh~

$ rincome $8000 to 9999, $8000 to~

$ partyid "Ind,near rep", "Not st~

$ relig   Protestant, Protestant,~

$ denom   "Southern baptist", "Ba~

$ tvhours 12, NA, 2, 4, 1, NA, 3,~

**Introduce**

gss\_cat  %>% introduce()

Let’s introduce the data set based on introduce function.

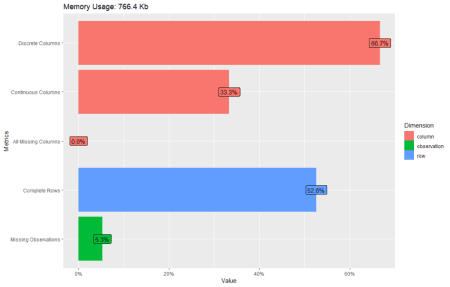
rows columns discrete\_columns continuous\_columns all\_missing\_columns total\_missing\_values complete\_rows total\_observations memory\_usage

1 21483       9                6                  3                   0                10222         11299             193347       784776

Now you can see the dataset contains 21483 observations with 9 variables all other kind of information’s.

**Visualization**

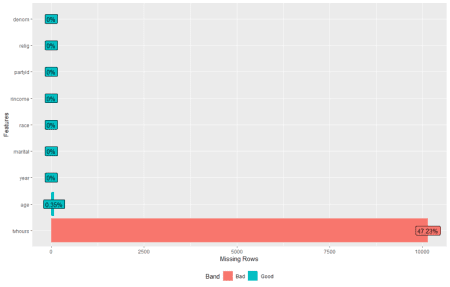
Most EDA techniques are graphical in nature with a few quantitative techniques. Here we are going to explain different kind of visualization techniques to identify the pattern and relationships.

gss\_cat %>% plot\_intro()

66.7% contains Discrete columns, 33.3% contains continuous columns.

52.6% contains complete rows and 5.3% contains missing observations.

All columns with missing values are 0%.

gss\_cat %>% plot\_missing()

gss\_cat %>% profile\_missing()

# A tibble: 9 x 3

  feature num\_missing pct\_missing

1 year              0     0

2 marital           0     0

3 age              76     0.00354

4 race              0     0

5 rincome           0     0

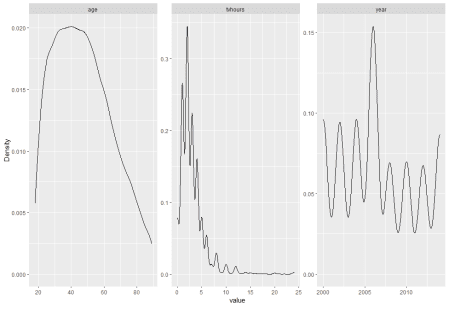
6 partyid           0     0

7 relig             0     0

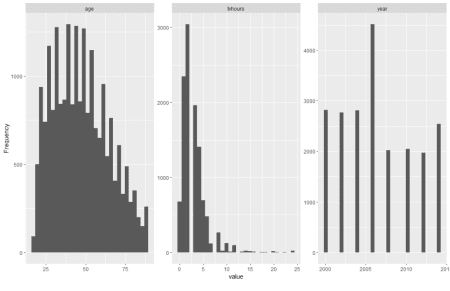
8 denom             0     0

9 tvhours       10146     0.472

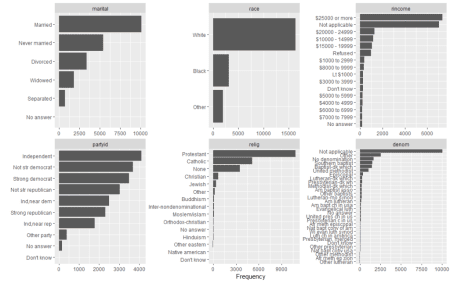
**Density Plot**

gss\_cat  %>% plot\_density()

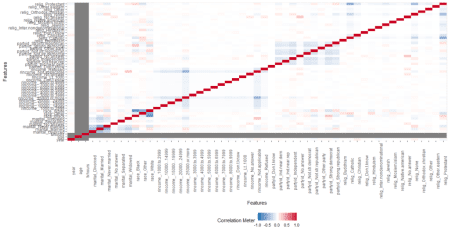
**Histogram**

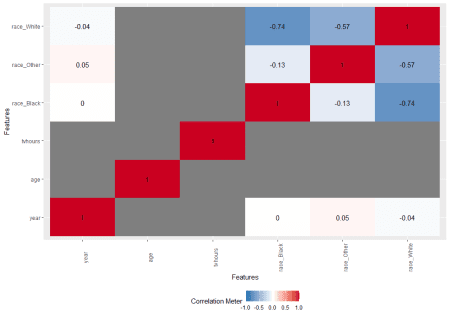
gss\_cat  %>% plot\_histogram()

**Bar Plot**

gss\_cat  %>% plot\_bar()

**Correlation Plot**

gss\_cat  %>% plot\_correlation()

gss\_cat  %>% plot\_correlation(maxcat = 5) 

maxcat can adjust based on your variables and interest. You can make use Data Explorer package for quick checking before proceeding to further analysis.

**Conclusion**

Exploratory Data Analysis is an approach to identify every nuance from the data at the early encounter and its plays a crucial role in the data analysis field.