Exploratory Data Analysis

Aayush Shrestha Anmol Jha

Kathmandu University
Department of Computational Mathematics

MATH 252 Progress Presentation

Outline

1. Introduction
2. Objective
3. Progressed Work
4. Work to Complete

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a method of analyzing data using statistical summaries and graphical representations.
EDA is a critical step in any Data Analysis or Data Science project as it provides a better understanding of data set variables and their relationships shows and what data can reveal beyond formal modeling or hypothesis testing tasks, and it.
It aids in determining how to best manipulate data sources to obtain the answers required, making it easier for data scientists to discover patterns, detect anomalies, determine test hypotheses, and validate assumptions.

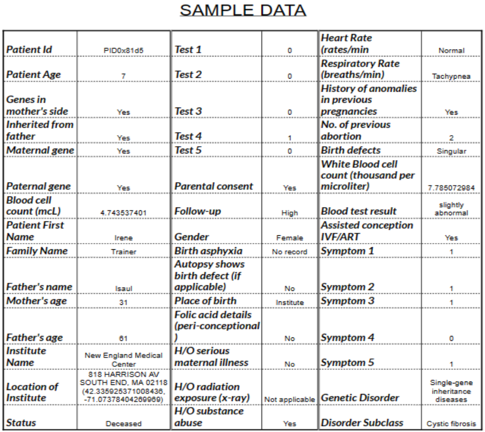
Dashboard

Dashboard is basically a GUI for Data Visualization. A dashboard is a good choice if you need to summarize and present a lot of information on a single window.
Dashboard is one of the major tool for presenting the data as it gives at-a-glance views of key performance indicators(KPI) relevant to a particular objective.
Dashboard helps Data Analysts and Data Scientists perform many data-related tasks, and also provides a visual aid for other stakeholders to understand data, and make accurate data-based decisions.

Goals For the project

- Familiarity with data cleaning, feature extraction, and data conversion.
- Describing the data using statistical method such as mean, median, standard deviation, correlation and so on.
- Implementation of different plots in R
- Implementation of a dashboard .
- Understanding and explaining the trends and outliers in the data based on the domain.

Initial View



SAMPLE DATA

Patient Id	PID0x81d5	*Test 1*	0	*Heart Rate (rates/min*	Normal
Patient Age	7	*Test 2*	0	*Respiratory Rate (breaths/min)*	Tachypnea
Genes in mother's side	Yes	*Test 3*	0	*History of anomalies in previous pregnancies*	Yes
Inherited from father	Yes	*Test 4*	1	*No. of previous abortion*	2
Maternal gene	Yes	*Test 5*	0	*Birth defects*	Singular
Paternal gene	Yes	*Parental consent*	Yes	*White Blood cell count (thousand per microliter)*	7.785072984
Blood cell count (mcL)	4.743537401	*Follow-up*	High	*Blood test result*	slightly abnormal
Patient First Name	Irene	*Gender*	Female	*Assisted conception IVF/ART*	Yes
Family Name	Trainer	*Birth asphyxia*	No record	*Symptom 1*	1
Father's name	Isaul	*Autopsy shows birth defect (if applicable)*	No	*Symptom 2*	1
Mother's age	31	*Place of birth*	Institute	*Symptom 3*	1
Father's age	61	*Folic acid details (peri-conceptional)*	No	*Symptom 4*	0
Institute Name	New England Medical Center	*H/O serious maternal illness*	No	*Symptom 5*	1
Location of Institute	818 HARRISON AV SOUTH END, MA 02118 (42.335925371008436, -71.07378404269969)	*H/O radiation exposure (x-ray)*	Not applicable	*Genetic Disorder*	Single-gene inheritance diseases
Status	Deceased	*H/O substance abuse*	Yes	*Disorder Subclass*	Cystic fibrosis

Figure: Sample raw data.

Removing Columns

The following Columns will not be useful in prediction of the disorder hence removed.
"Patient Id","Patient First Name","Family Name","Father's name","Institute Name","Location of Institute","Parental consent","Place of birth"
Below is the code:

```
drop <- c("Patient Id","Patient First Name","Family Name","Father's name","Institute Name","Location of Institute","Parental consent","Place of birth")
df = df[,!(names(df) in drop)]
```

Dashboard

Exploratory Data Analysis

Dashboard