Exploratory Data Analysis On Genetic Data

Aayush Shrestha Anmol Jha

Kathmandu University
Department of Computational Mathematics

MATH 252 Final Presentation
Supervisor: Mr. Kiran Shrestha
Special Thanks to Mr. Simon Shrestha

Outline

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a method of analyzing data using statistical summaries and graphical representations.

EDA is a critical process that aids in determining how to best manipulate data sources to obtain the required answers, making it easier for data scientists to discover patterns, detect anomalies, determine test hypotheses, and validate assumptions.

Dashboard

A 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.

A 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.

Project

Our project consist of the task to perform EDA on a Kaggle dataset so as to find the significant attributes that contribute to the classification of the target genetic disorder.
Also we were to create a dashboard including all those Key performance indicators.

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.

Non Graphical

Initially we used the summary statistics, count, minimum, maximum, average, standard deviation etc to describe the numeric columns.
Also correlation between every pair of numeric attributes were computed to find highly correlated attributes.

Graphical

We used R programming packages to create visualization related to the distribution of an attribute on its own and also to demonstrate the relation between attributes. Here we mainly used ggplot package to create stacked bar graph and stem and leaf plots.
Also we used a heat-map plot to better visualize the correlation matrix created.

Domain Based

Based on the further continuation of the project as a classification of genetic disorder based on genetic inheritance the attributes were considered for their significane to the project and promptly discarded.

Dashboard

We used `https://visual.is` as a proper dashboard creation tool. This let us interactively choose different plots and data that we want to display in our dashboard.

Initial Pre-processing

The data was changed from an xlsx file to a csv file, and the column names were modified to better suit programming. Also, all the rows with any NA values were removed.The dataset now contains 45 columns and 6706 observations.

```
#importing package
library(readxl)
library(tidyverse)
```

```
#read the excel file
df <- read_excel("train.xlsx")
```

```
#rename the columns eliminating obstacle character
names(df) <- gsub(" ", "_", names(df))
names(df) <- gsub("\\(", "_", names(df))
names(df) <- gsub("\\)", "_", names(df))
names(df) <- gsub("\\/", "_", names(df))
names(df) <- gsub("'", "_", names(df))
```

```
#rows with any na value are dropped
df <-df %>% drop_na()
```

```
#save as csv
write.csv(df, "train.csv", row.names = FALSE)
```

Figure: Preprocessing Code.

Non Graphical EDA

Summary statistics showed that the columns of Test 1, Test 2,, Test 3, Test 4, Test 5 have no variance in them hence they are removed.

	Test_1	Test_2	Test_3	Test_4	Test_5
Min.	0	0	0	1	0
1st Qu.	0	0	0	1	0
Median	0	0	0	1	0
Mean	0	0	0	1	0
3rd Qu.	0	0	0	1	0
Max.	0	0	0	1	0

Figure: Summary Statistics

Non Graphical EDA

Also a correlation heatmap showed that there are no highly correlated attribute

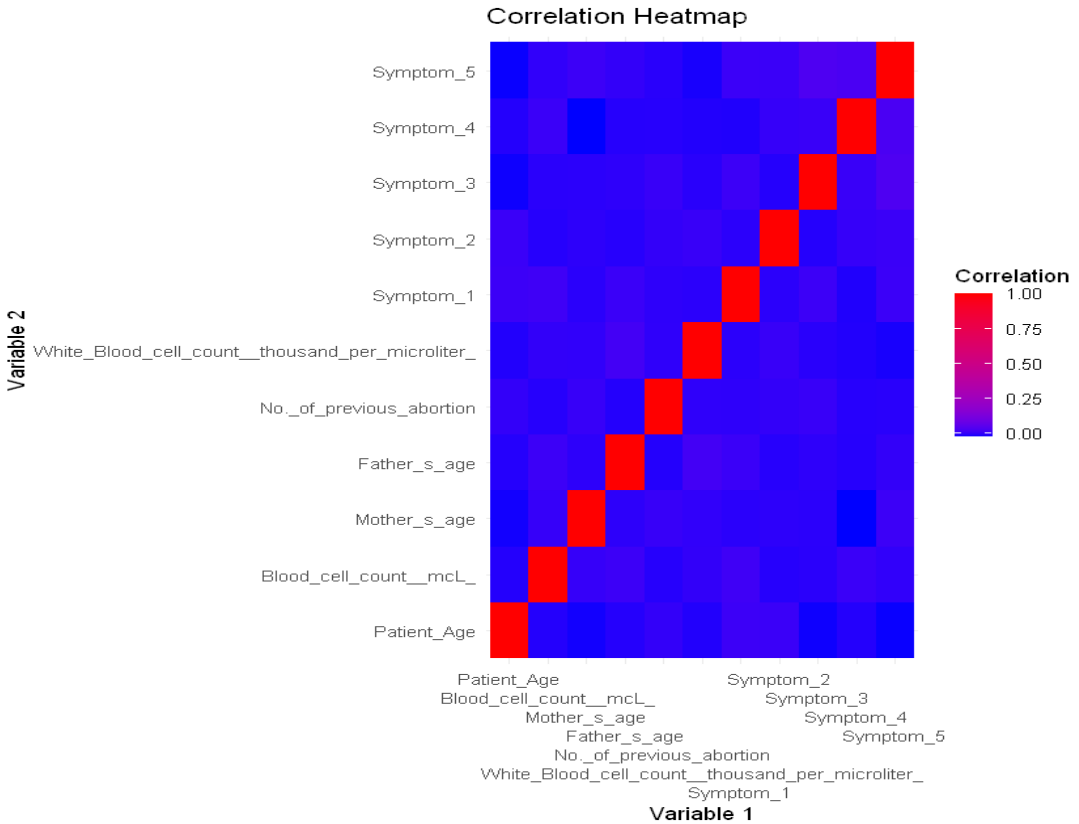

Figure: Correlation Heatmap

Graphical EDA

We find that the Genetic disorder attribute is dependent on the Disorder Subclass attribute. Hence The Genetic Disorder attribute is removed.

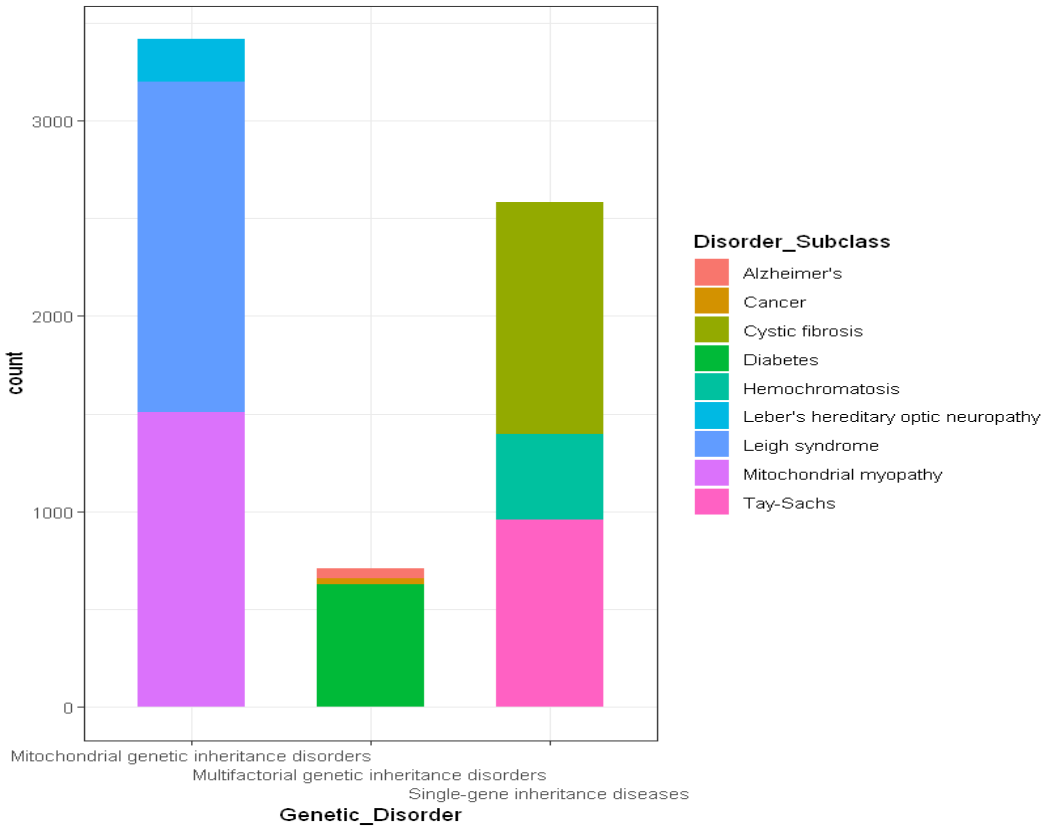

Figure: Dependency between Target Attributes

Graphical EDA

For categorical attributes, a stacked bar graph is created with Disorder subclass attribute on x-axis. By this we found that most of such attributes divided the dataset almost symmetrically along the Disorder Subclass and hence is removed.

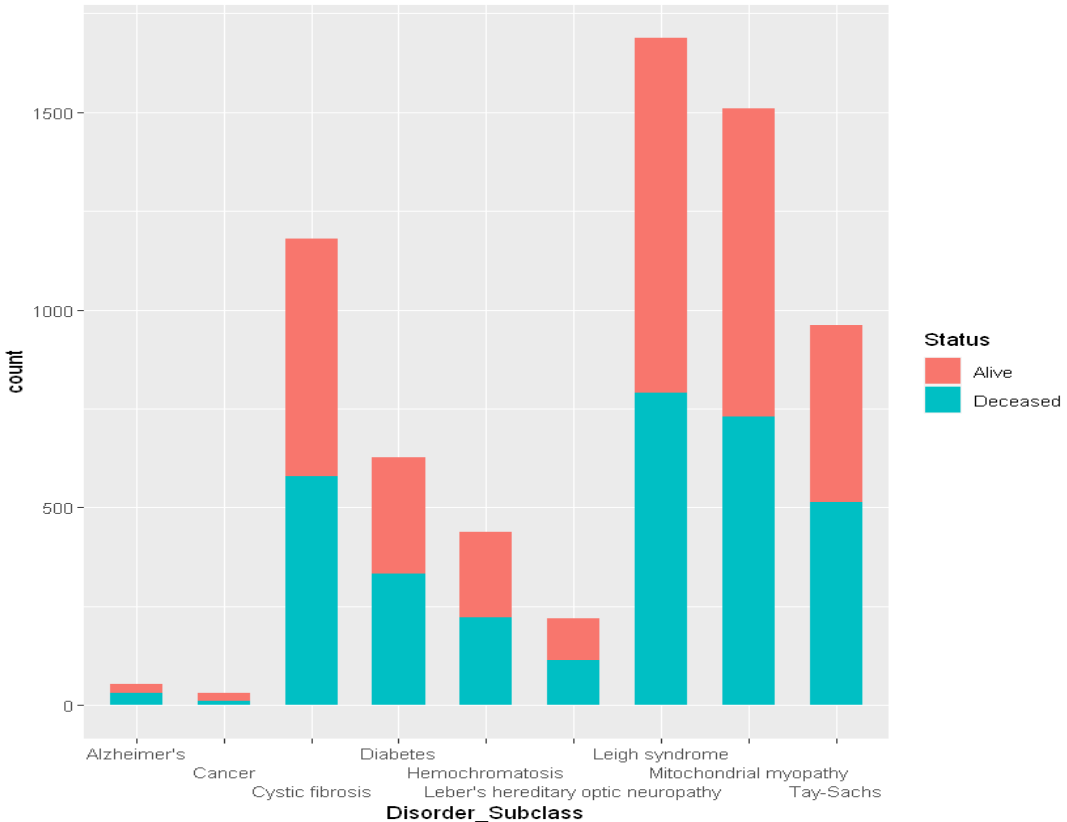

Figure: Disorder Subclass and Status

Graphical EDA

For continuous attributes, a box and whisker plot is created with Disorder subclass attribute on x-axis. Here the data were not as symmetrical.

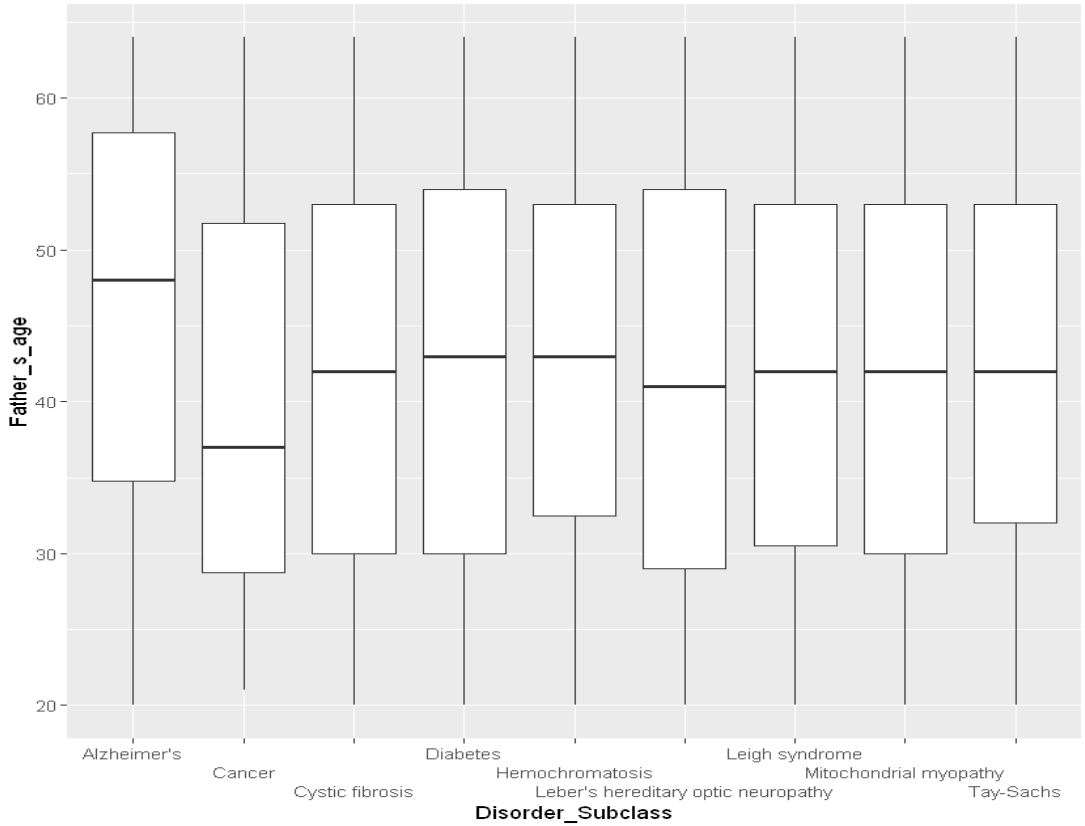

Figure: Disorder Subclass and Patient's Father's age

Graphical EDA + Domain Information

Given the domain of study to be classification of disorder due to genetic inheritance, the following attribute were considered significant and have shown reliable variance.

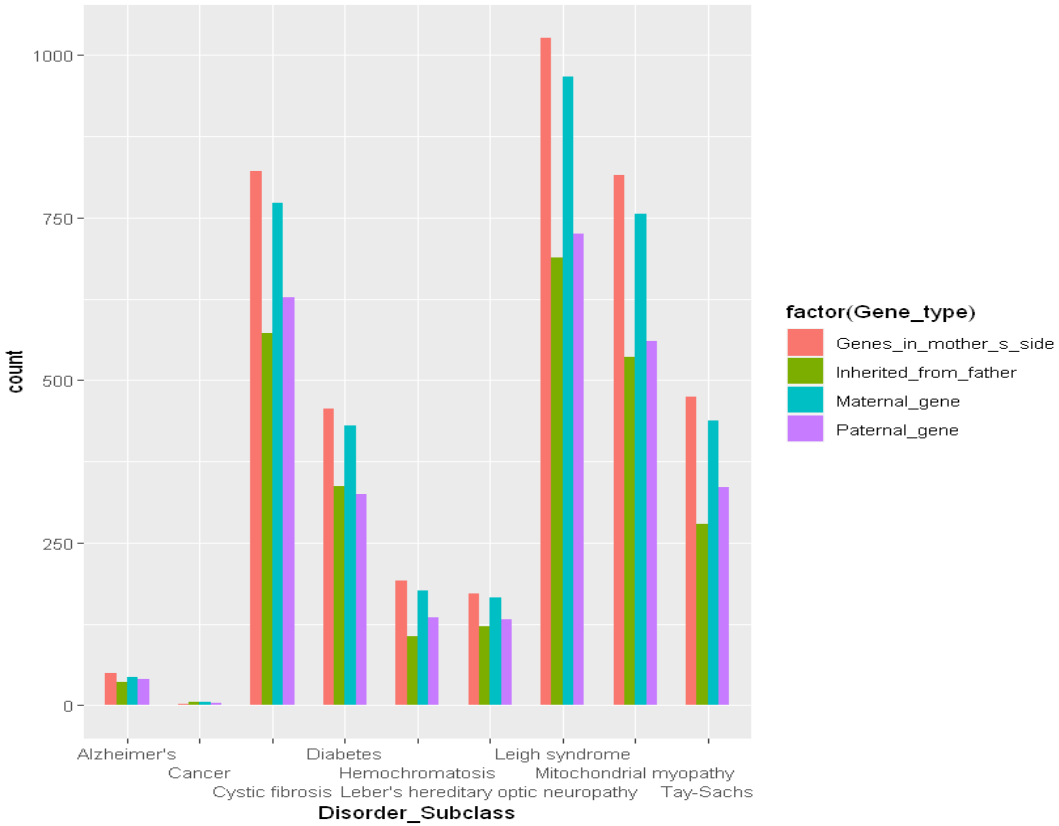

Figure: Disorder Subclass and Genetic Presence (YES)

Graphical EDA + Domain Information

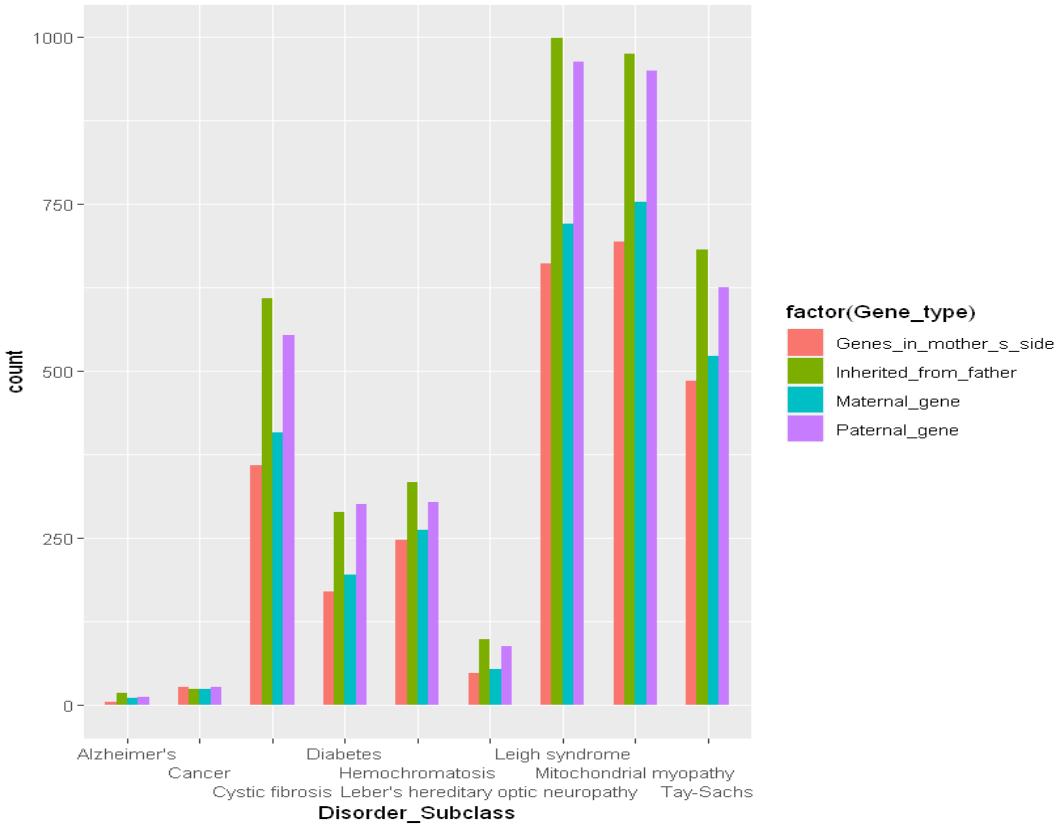

Figure: Disorder Subclass and Genetic Presence (NO)

Graphical EDA + Domain Information

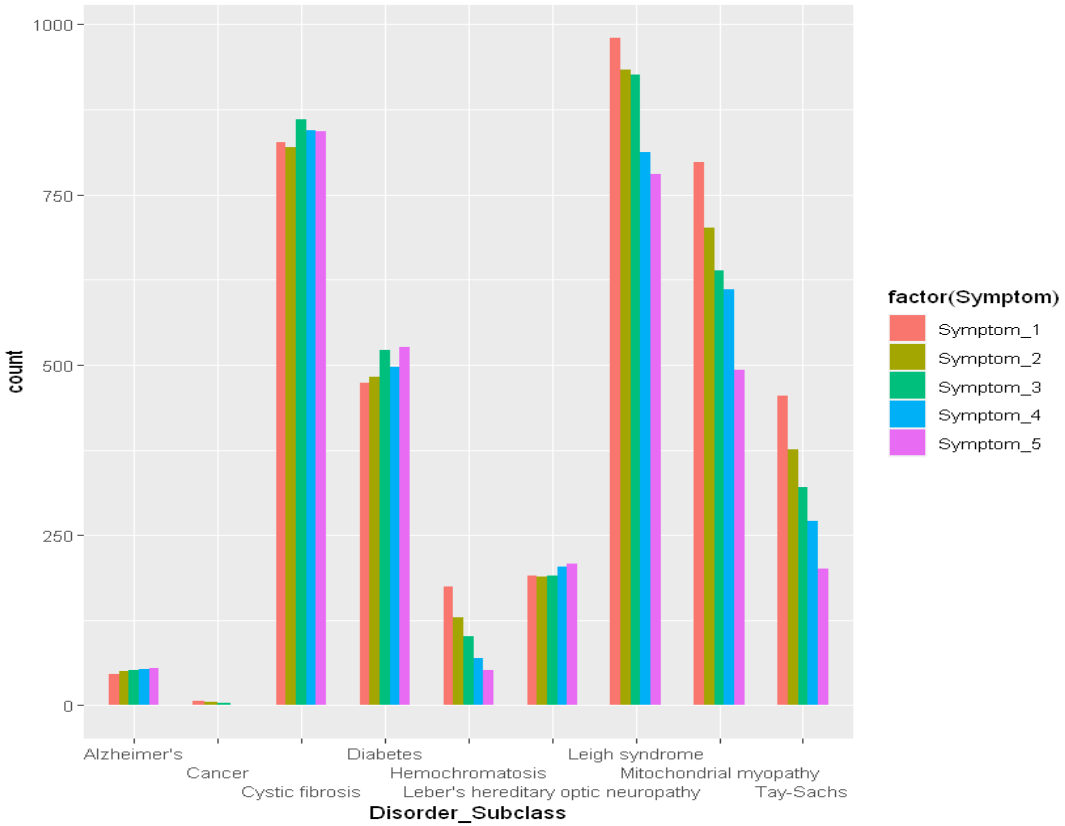

Figure: Disorder Subclass and Symptom Detected (True)

Graphical EDA + Domain Information

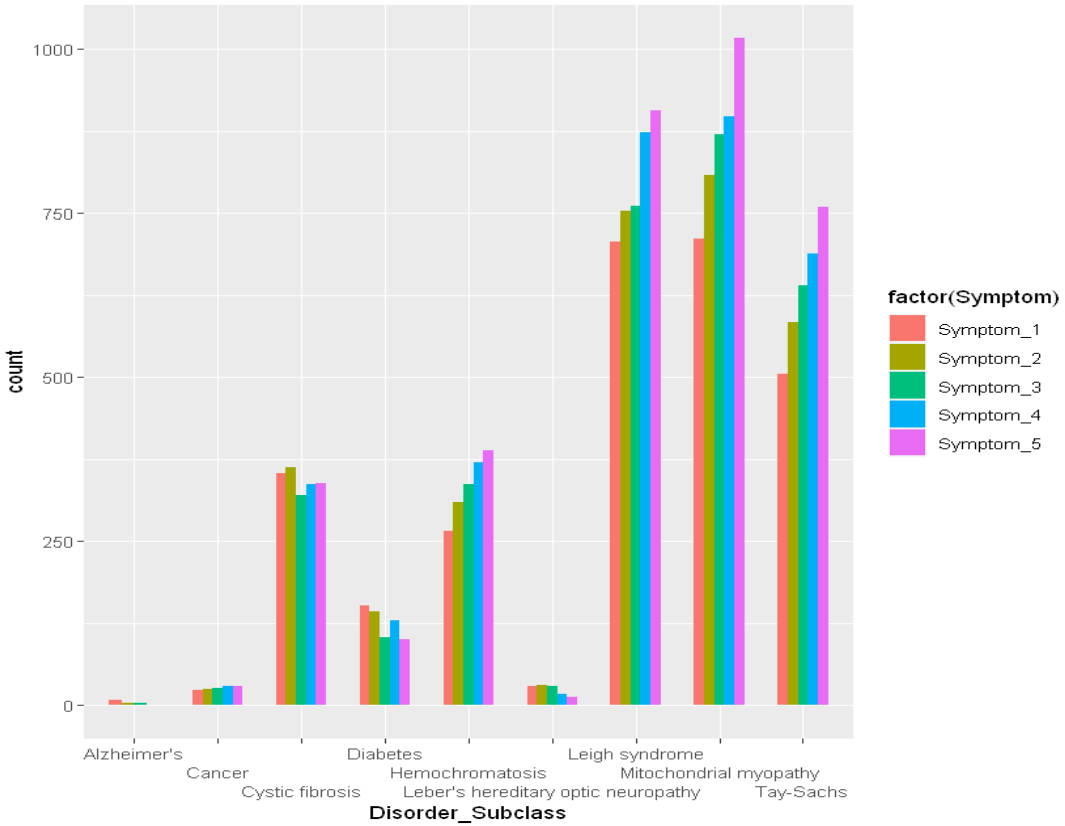

Figure: Disorder Subclass and Symptom Detected (False)

Domain Based

Attributes like Patients Name, birth address, History of alcohol which doesnt have much significance to the domain of genetic inheritance were also removed.
All this left us with 16 columns with one of them being the target column.

Dashboard

We have a basic dashboard framework ready that gives us basic visualization of the data.

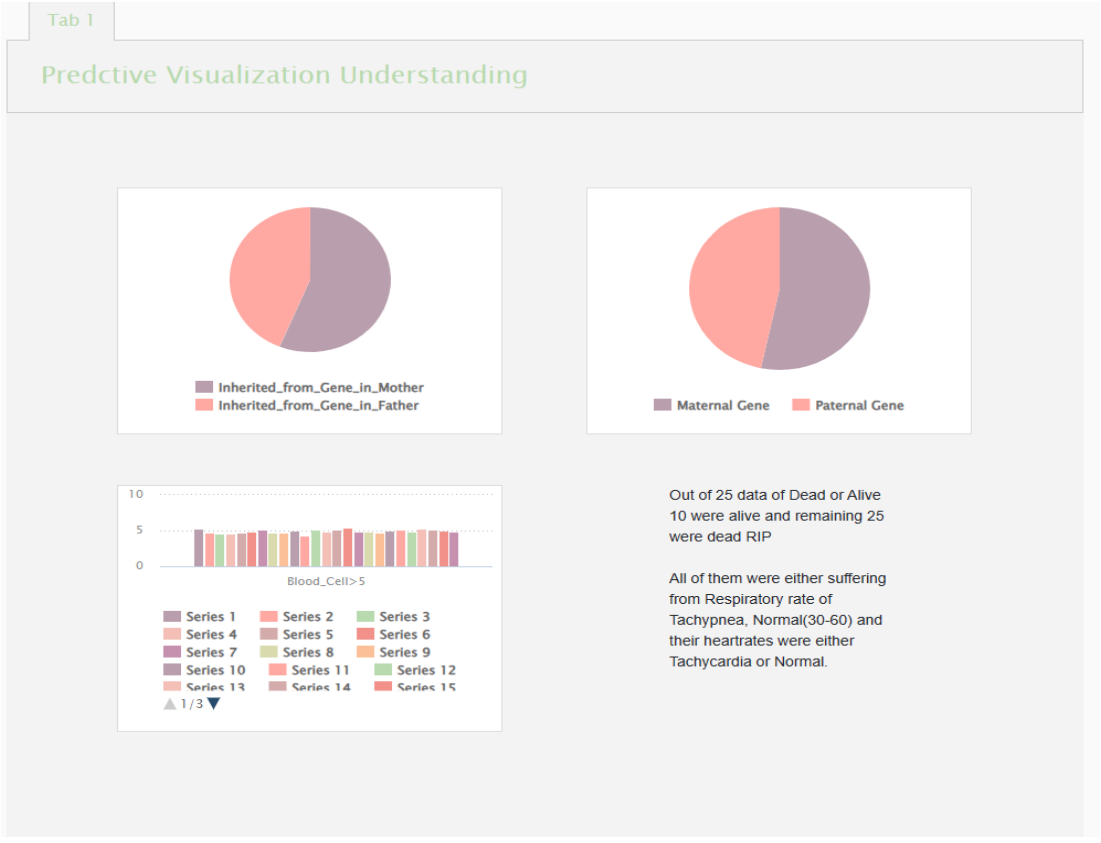

Figure: Dashboard Sample

Conclusion

In this EDA project, we applied various statistical analysis, graphical visualization, and domain knowledge to reduced the dataset from 45 to 16 columns.

Patient Age	Genes in mother's side
Inherited from father	Maternal gene
Paternal gene	Blood cell count mcL
Mother's age	Father's age
Number of previous abortion	White Blood cell count
Symptom 1	Symptom 2
Symptom 3	Symptom 4
Symptom 5	Disorder Subclass

Discussion and Limitations

The dashboard created is entirely dependent on the capability of the tool used.
Given that the data is from a Kaggle dataset, there is a possibility of it being synthetic, as such there are certain trends and outliers that isnt explainable from a real world perspective.
The methodology discussed here is still useful and backed by proper citable sources. Hence there are still proper real world application of this project and or its workflow.

Demonstration

Now we would like to demonstrate the Project

THANK YOU!

Please share if you have any queries.

We would like to show the code now.