EXPLORATORY DATA ANALYSIS OF GENETIC DATA

A SECOND YEAR PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF B.Sc. IN COMPUTATIONAL MATHEMATICS

BY

1. Aayush Shrestha (21)

2. Anmol Jha (31)

DEPARTMENT OF MATHEMATICS

SCHOOL OF SCIENCE

KATHMANDU UNIVERSITY

DHULIKHEL, NEPAL

June 2023

CERTIFICATION

This project entitled Exploratory Data Analysis of Genetic Data is carried out under my supervision for the specified entire period satisfactorily, and is hereby certified as a work done by following students

1. Aayush Shrestha (21)
2. Anmoj Jha (31)

in partial fulfillment of the requirements for the degree of B.Sc. in Computational Mathematics, Department of Natural Sciences, Kathmandu University, Dhulikhel, Nepal.

_____

Mr. Kiran Kumar Shrestha
Department of Natural Sciences (Mathematics),
School of Science, Kathmandu University,
Dhulikhel, Kavre, Nepal
Date:_____

APPROVED BY:

I hereby declare that the candidate qualifies to submit this report of the Math Project (MATH 252) to the Department of Natural Sciences.

_____

Dr. Rabindra Kayastha
Department of Mathematics
School of Science
Kathmandu University
Date:_____

ACKNOWLEDGMENTS

This project was carried out under the supervision of Mr. Kiran Kumar Shrestha. I would like to express my sincere gratitude towards my supervisors for his excellent supervision, guidance and suggestion for accomplishing this work. And to the entire faculty of Department of Natural Sciences (Mathematics) for encouraging, supporting and providing this opportunity.

I am indebted to all my friends for helping and guiding me over the related software.

Special appreciation goes to Mr. Simon K. Shrestha(Department Of Biotechnology) for his help with the analysis of genetic data and his valuable suggestions.

Lastly, we would like to thank everyone who helped us directly and indirectly during the duration of completing our project work.

ABSTRACT

This report summarizes a semester-long effort that looked at medical data through data analysis and the development of an approachable dashboard to display the results.The goal is to extract insightful knowledge from the data and offer decision-supporting tools to healthcare practitioners.

The project entails preparing and cleaning the data, then using statistical analysis methods. Despite obstacles like missing values, a broad medical data set from recognized healthcare organizations is utilized. Strategies are used to deal with these problems.

Finding patterns, trends, and relationships in the medical data is one of the anticipated results. The project's goal is to find significant insights that can improve medical procedures and patient care by utilizing the right visualization libraries and statistical approaches. The results will aid in bettering healthcare decision-making.

Numerous results are expected. First, the study will provide crucial information about demographic trends and correlations between variables. Second, by enabling healthcare practitioners and the concerned patient to make data-driven decisions, the dashboard will streamline procedures and improve productivity in the medical industry.

To sum up, the purpose of this project is to analyze medical data thoroughly and produce an interactive dashboard that will show the results.

CONTENTS

CERTIFICATION ii

ACKNOWLEDGEMENTS iii

ABSTRACT iv

LIST OF FIGURES viii

LIST OF TABLES ix

1 INTRODUCTION 1

1.1 **Background** . . . 1

1.2 **Objectives** . . . 2

1.3 **Significance** . . . 2

1.4 **Limitations** . . . 2

1.5 **Specifications** . . . 3

2 METHODOLOGY 4

2.1 **Theoretical Framework** . . . 4

2.2 **Data Collection And Preprocessing** . . . 5

2.3 **Exploratory Data Analysis** . . . 5

2.3.1 Non-graphical . . . 5

2.3.2 Graphical . . . 5

2.3.3 Domain Based . . . 6

2.4 **Dashboard Creation** . . . 6

3 RESULTS 7

3.1 **Findings** . . . 7

3.2 **EDA** . . . 7

3.2.1 Non - Graphical 7
3.2.2 Graphical 9
3.3 Dashboard 20

4 CONCLUSIONS **23**

LIST OF FIGURES

1.1 Sample Data Of The Project 3

3.1 Columns With No Variance 8
3.2 Correlation Heatmap Of Numeric Attributes 8
3.3 Disorder And Status 9
3.4 Disorder And History Of Maternal Illness 10
3.5 Disorder And Heart - rate 10
3.6 Disorder And Gender 11
3.7 Disorder And Follow up 11
3.8 Disorder And Folic Acid details 12
3.9 Disorder And Blood Test Result 12
3.10 Disorder And Birth Defect 13
3.11 Disorder And Autopsy Report 13
3.12 Disorder And Asphyxia 14
3.13 Disorder And Respiratory Rate 14
3.14 Disorder And IVF details 15
3.15 Disorder And History Of Pregnancy 15
3.16 Relation Between Target Attributes 16
3.17 Disorder And Patient's Age 17
3.18 Disorder And Patient's Mother's Age 17
3.19 Disorder And Patient's Father's Age 18
3.20 Disorder And White Blood Cell Count 18
3.21 Disorder And Blood Cell Count 19
3.22 Disorder And Number Of Previous Abortions 19
3.23 Disorder And Frequency Of Symptoms Being True 20
3.24 Disorder And Frequency Of Symptom Being False 20
3.25 Disorder And Frequency Of Gene Being Yes 21

3.26 Disorder And Frequency Of Gene Being No 21
3.27 Dashboard Sample 22

LIST OF TABLES

3.1 Names Of Discarded Columns 7

4.1 Remaining Columns 23

CHAPTER 1

INTRODUCTION

1.1 Background

The availability of enormous volumes of data in the healthcare industry in recent years has greatly increased the prospects for understanding and enhancing decision-making. A key tool for identifying patterns, trends, and correlations in large medical datasets is exploratory data analysis (EDA). To visualize these insights and support data-driven decision-making in the healthcare industry, interactive dashboard creation has also become crucial.

The necessity to use EDA and dashboard generation to solve issues and enhance healthcare procedures is what spurred this project's inception. The sheer volume and complexity of medical data are often too much for traditional analysis tools to manage, making it difficult to draw useful conclusions.Data scientists and healthcare professionals can find hidden patterns, outliers, and correlations in the data by using EDA's systematic and exploratory methodology.

Additionally, interactive dashboards that visualize these insights are a potent tool for better data communication and decision-making. Healthcare personnel may engage with data visualizations, personalize views, and receive insightful information quickly thanks to dashboards' straightforward and user-friendly interfaces. EDA and dashboard development have the power to transform healthcare procedures, maximize resource use, and improve patient care.

1.2 Objectives

The Objectives of this project were as follows

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

1.3 Significance

Exploratory data analysis (EDA) and the development of interactive dashboards hold the potential to transform healthcare procedures, which is why this initiative is significant. This research intends to identify useful insights, such as correlations, trends, and risk factors that can significantly influence decision-making in the healthcare industry by conducting a thorough EDA on medical data. These insights can aid healthcare practitioners in maximizing resource allocation, enhancing overall health care plans, and improving patient outcomes. A user-friendly platform for visualizing and analyzing these insights will be provided by the creation of an interactive dashboard customized to the requirements of healthcare professionals, enabling users to make defensible decisions based on data-driven evidence.

1.4 Limitations

The data used in this project is synthetic, meaning it has been artificially generated to simulate real-world medical data. Although synthetic data allows for controlled experiments and can provide valuable insights, it may not fully capture the complexity and nuances of actual patient data. The findings and conclusions drawn from the analysis should be interpreted with caution, considering the inherent differences between synthetic and real-world data. Efforts have been made to ensure the quality of synthetic dataset,

but there is still a possibility of errors or limitations in the data generation process. The availability of computational resources and time constraints may also impact the scope and depth of the analysis. Despite these limitations, this project serves as an exploration of the potential benefits and challenges associated with synthetic medical data in the context of exploratory data analysis and dashboard creation. It provides insights into the effectiveness of these methodologies and offers considerations for future research utilizing real-world medical data.

1.5 Specifications

- Language: R (tidy-verse packages, dashboard packages).
- IDE: Jupyter notebook and RStudio
- Dashboard tool: flexboard package and `https://visual.is`
- Data: Kaggle dataset

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, -71.073784042699969)	*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 1.1: Sample Data Of The Project

CHAPTER 2

METHODOLOGY

2.1 Theoretical Framework

Exploratory Data Analysis (EDA) plays a pivotal role in the research process, allowing for a comprehensive examination of data without making any assumptions. Komorowski, Marshall, Salciccioli, and Crutain [3] emphasize the significance of EDA as a vital step in understanding datasets and uncovering valuable insights. The objectives of EDA encompass gaining insight into the dataset, visualizing potential relationships between features and outcome variables, detecting outliers and anomalies, and creating relevant variables. EDA methods can be categorized as graphical or non-graphical, as well as univariate or multivariate. Graphical techniques involve visualizing data through various plots, such as scatter plots, histograms, and box plots, to uncover patterns and trends. Non-graphical methods encompass numerical summaries, such as measures of central tendency, dispersion, and correlation coefficients, to provide a quantitative understanding of the data. Univariate analysis focuses on examining individual variables, while multivariate analysis explores relationships between multiple variables simultaneously.

By adopting an exploratory mindset during EDA, researchers gain a deeper understanding of the dataset and identify potential issues or limitations. EDA serves as a crucial step in formulating research details, enabling researchers to make informed decisions about data selection, variable creation, and subsequent analysis.

2.2 Data Collection And Preprocessing

The Data is a synthetic dataset obtained from `https://www.kaggle.com/datasets/mukund23/predict-the-genetic-disorder?select=train.csv`. And as shown in 1.1, there are 2 target features and 43 parameter attributes. The EDA methodology discussed here will be used as a tool for dimensionality reduction.
The only preprocessing done before EDA is to convert the data file into a csv format and then to remove NA values from the data to ensure proper study of trends and relationships.

2.3 Exploratory Data Analysis

2.3.1 Non-graphical

Uni-variate

For each attribute, their count, sum, maximum, minimum, and standard deviation is computed if its a numeric attribute and unique values is its a non numerical attribute type. Based on this, those attribute that have no variance is removed as, when an attribute has no variance, it means that it has the same value across all instances or observations in the dataset. Such attributes do not contribute any discriminatory or informative value to the model.[2]

Multi-variate

A correlation analysis is done to see how one attribute is dependent/ correlated to another. This is done on the numeric attributes to detect and remove highly correlated attributes as suggested in [5]

2.3.2 Graphical

Uni-variate

For each attribute, their count, spread is measured. Mostly using Bar graphs, pie charts, density charts and stem and leaf charts. This gives us the tentative spread of the data in relation to the attribute.

Multi-variate

The same charts plus where the data is grouped by using a second feature and few more such as Scatter plots, line charts helps us visualize the relation between the attributes. Also we can study the spread with respect to the features.

2.3.3 Domain Based

As the data is to be dealt with as a data to be used in further development of a genetic inheritance classification model, attributes that are of less relevance to the domain are also not considered as a significant attribute.

2.4 Dashboard Creation

To facilitate data visualization and improve accessibility, we develop ouw own interactive dashboard in R as well as using the tool `https://visual.is`. The dashboard presents the insights obtained from the EDA in a user-friendly and visually appealing manner. It includes interactive charts, filters, and intuitive navigation features, enabling healthcare professionals to explore and interact with the data effectively. The dashboard serves as a valuable tool for decision-making and enhances the overall usability of our project.

CHAPTER 3

RESULTS

3.1 Findings

In this chapter, we present the results and visualization of the data set and theory described in CHAPTER-2 and finally discuss the results.

3.2 EDA

3.2.1 Non - Graphical

Based on summary statistics, there were found columns that had no variance. As such, those columns were removed from the dataset as per the reasoning in 2.3.1 Uni-variate.

Further more, the following columns were removed based on the domain of study as these columns were deemed to be less significant for further studies of the data. These are:

Table 3.1: Names Of Discarded Columns

Patient Id	Patient First Name	Family Name
Father's name	Institute Name	Location of Institute
Parental consent	Place of birth	H O radiation exposure x.ray
H O substance abuse		

Finally, the correlation between every pair of numeric attributes shows that there are no highly correlated attributes.

	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 3.1: Columns With No Variance

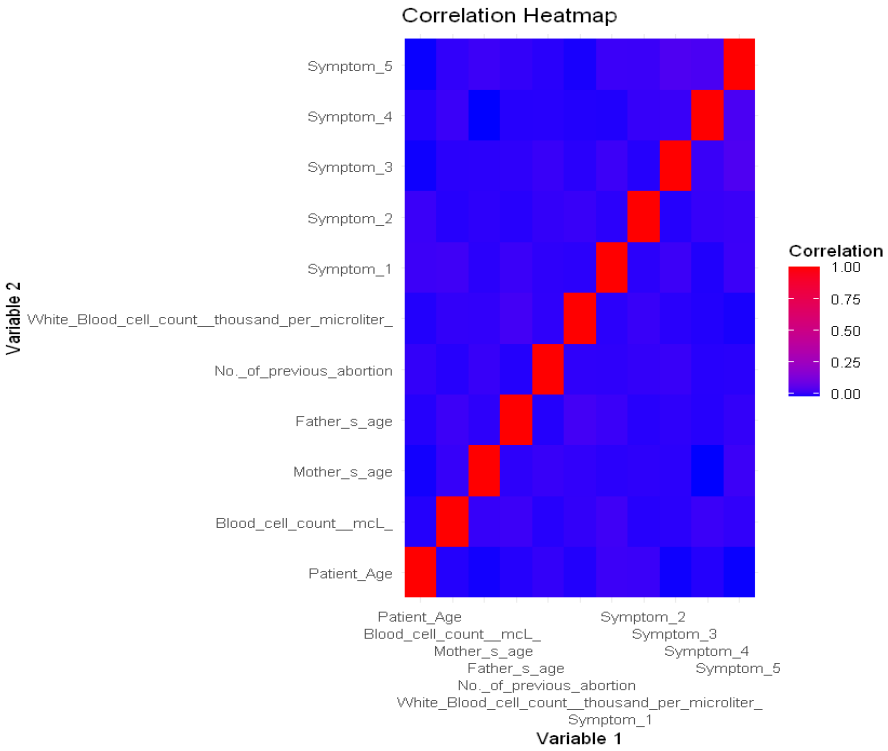

Figure 3.2: Correlation Heatmap Of Numeric Attributes

3.2.2 Graphical

Based on the graphical analysis of the attributes in the dataset, we found many columns that does not provide any predictive value with regards to the target attributes.

According to [1], attribute significance is an important concept in data mining. The evaluation of attribute usefulness is crucial for effective data analysis. If the distribution of data across different attribute categories is equal, it indicates that the attribute does not have a strong discriminatory power to differentiate between the diseases. This lack of discriminatory power suggests that the test 1 attribute is not significantly associated with the diseases in your dataset.

Similarly, [4] discusses the significance of attributes in machine learning. When the distribution of data is symmetric between different attribute categories, it implies that the attribute does not provide valuable predictive information for distinguishing between diseases. In this case, the symmetric distribution of data for the test 1 attribute suggests a lack of significance in relation to the diseases.

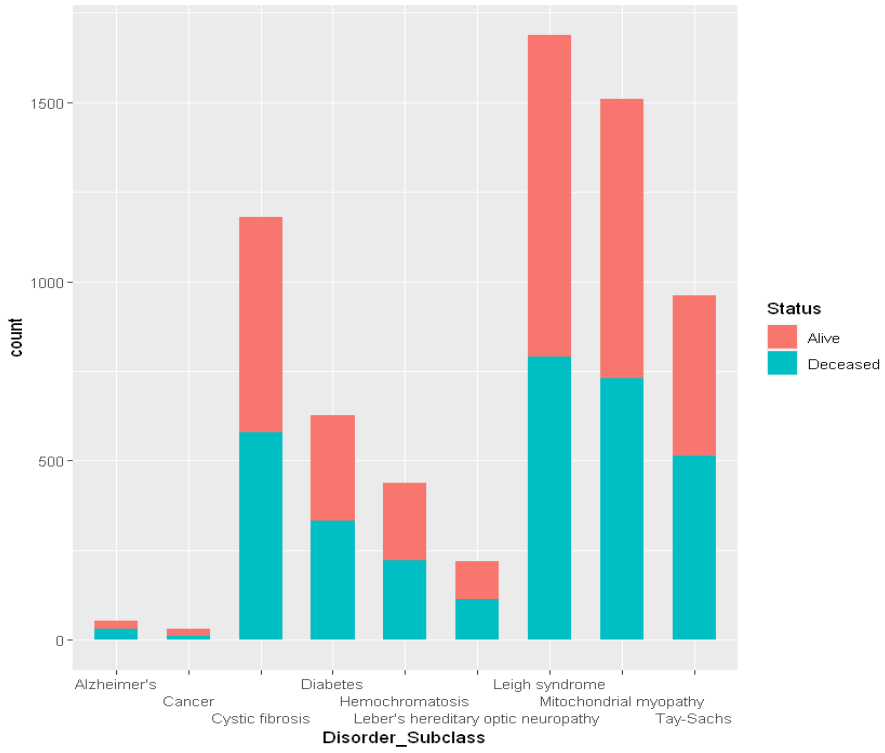

Figure 3.3: Disorder And Status

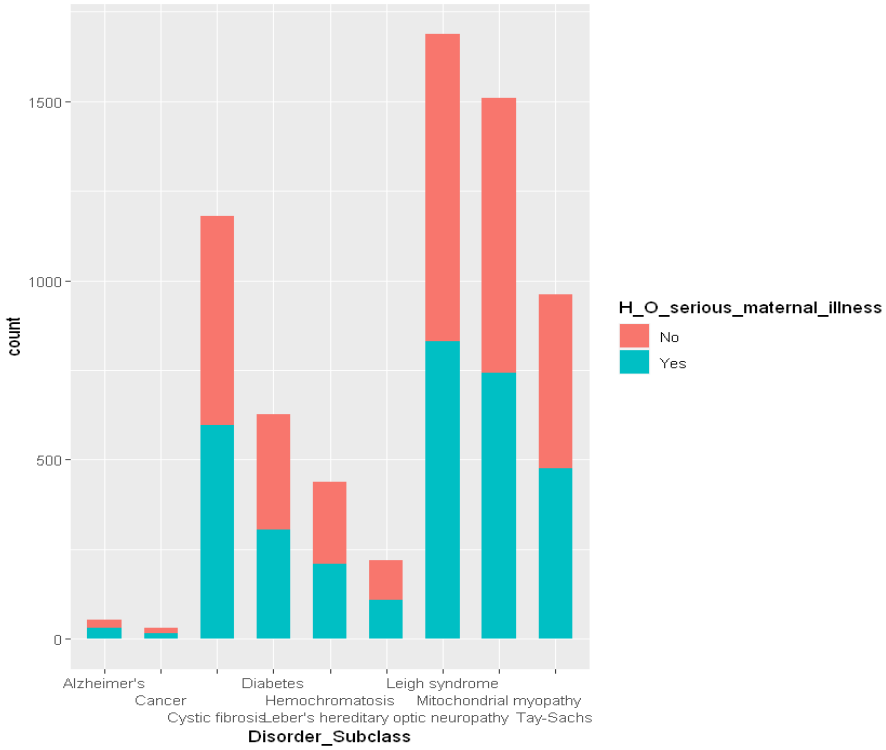

Figure 3.4: Disorder And History Of Maternal Illness

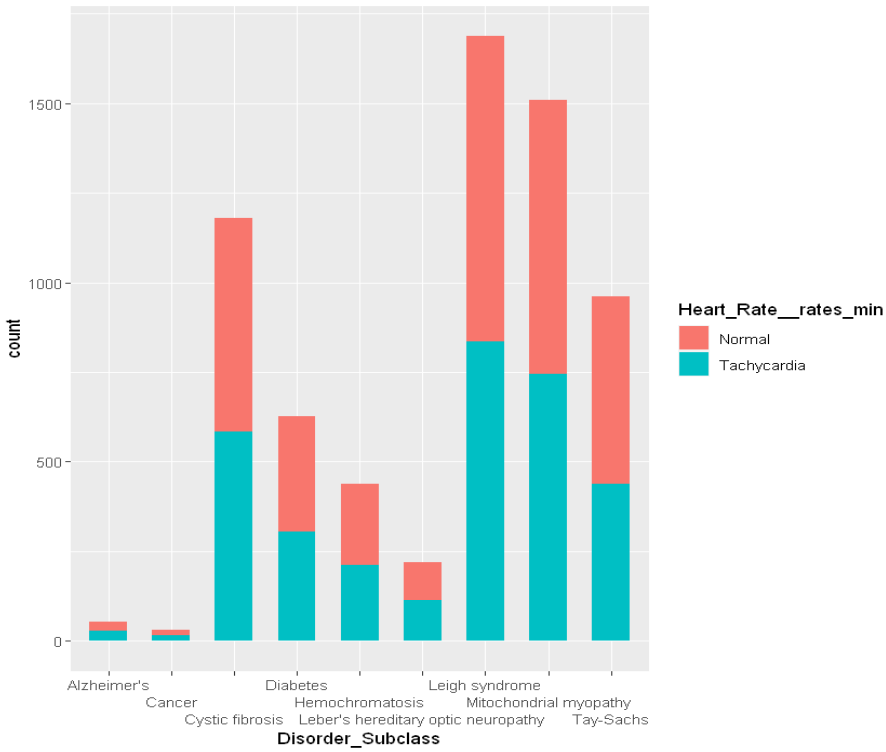

Figure 3.5: Disorder And Heart - rate

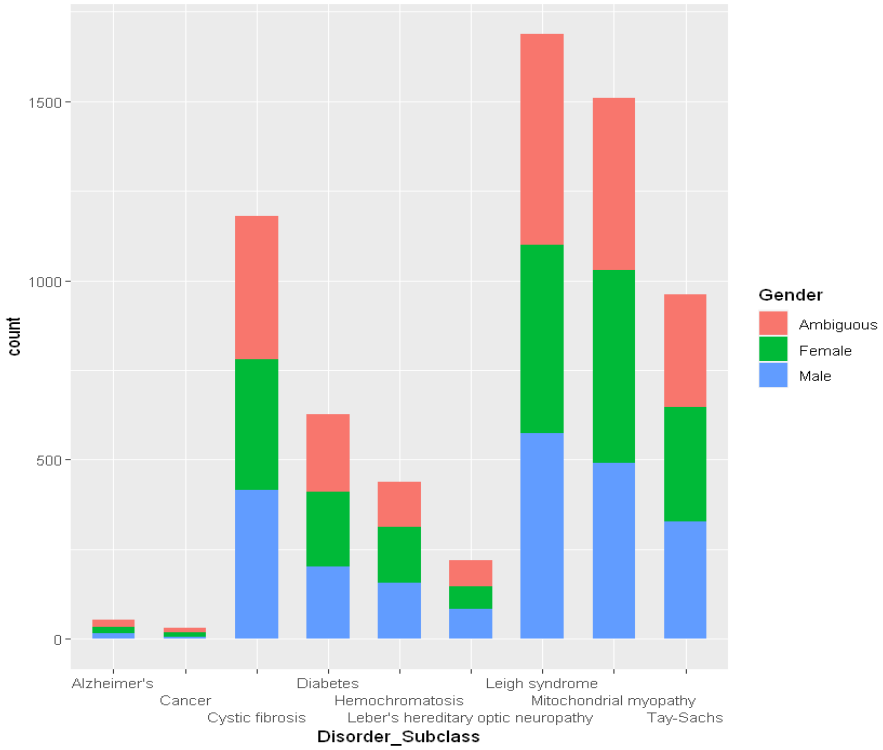

Figure 3.6: Disorder And Gender

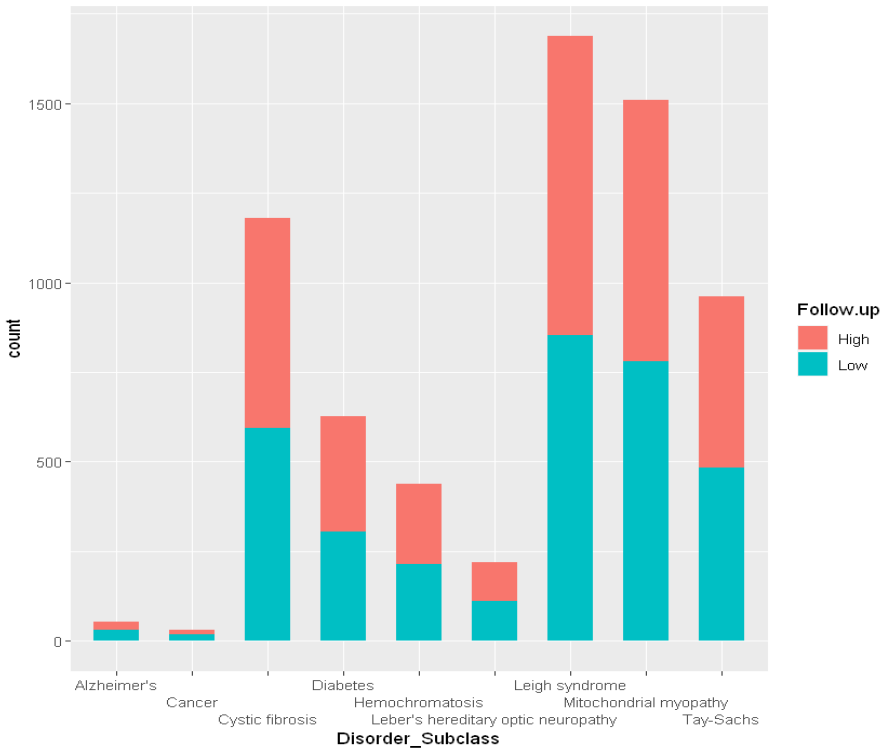

Figure 3.7: Disorder And Follow up

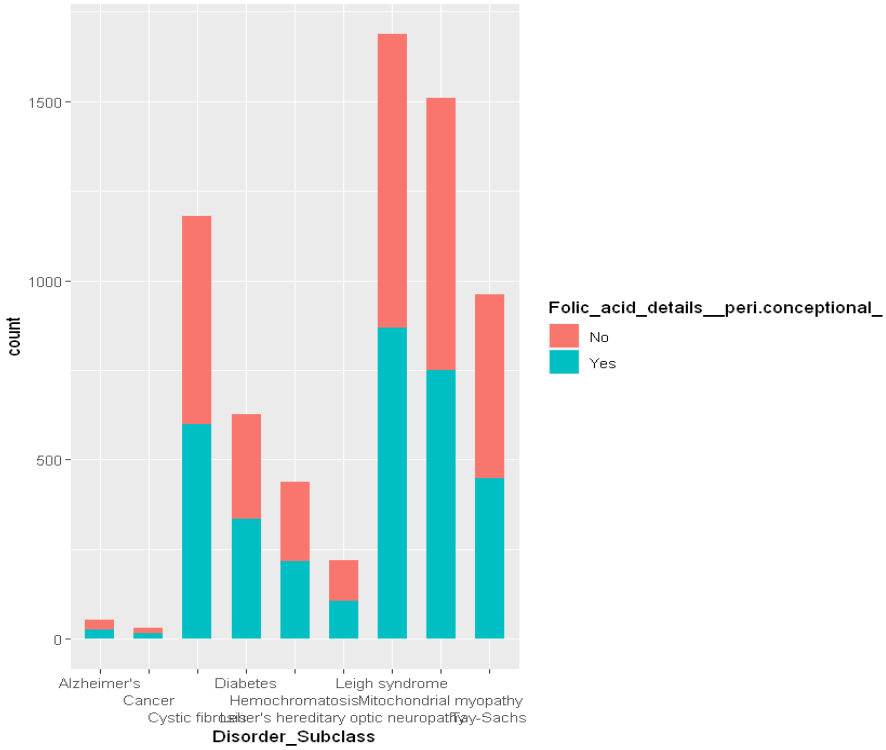

Figure 3.8: Disorder And Folic Acid details

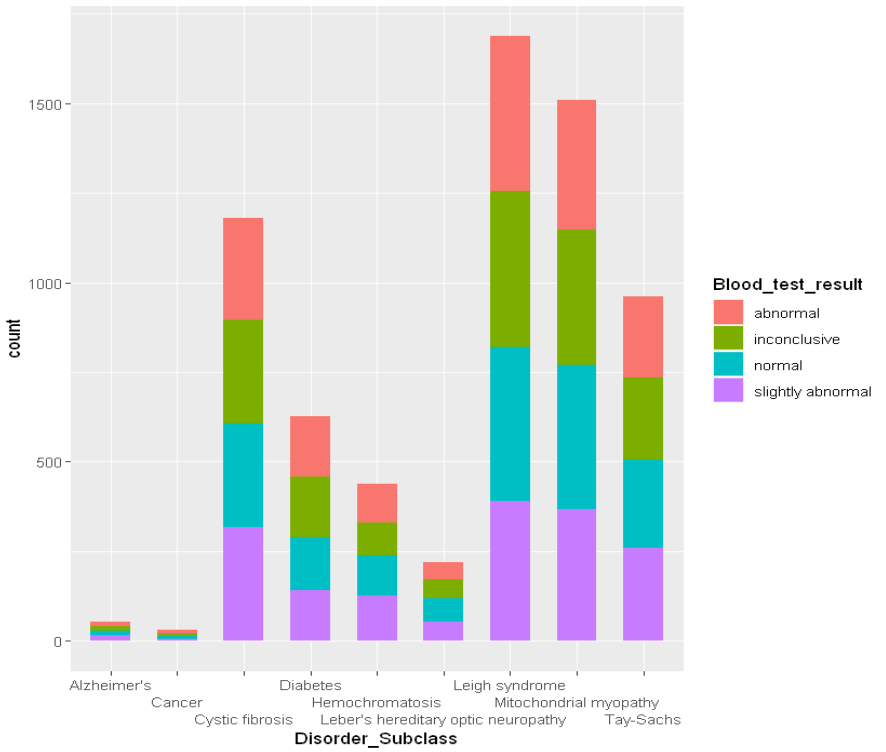

Figure 3.9: Disorder And Blood Test Result

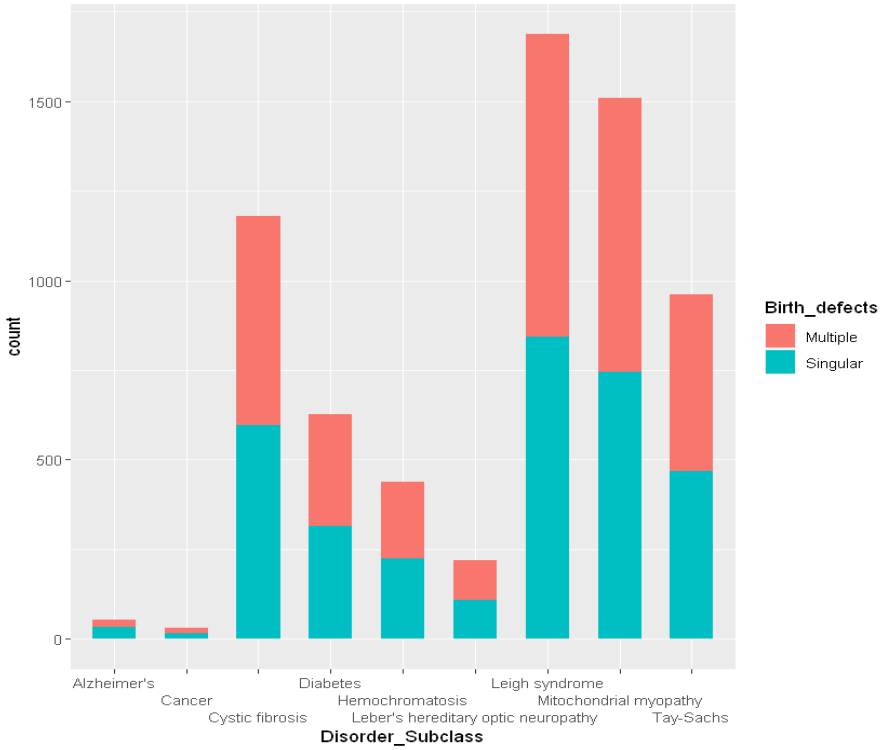

Figure 3.10: Disorder And Birth Defect

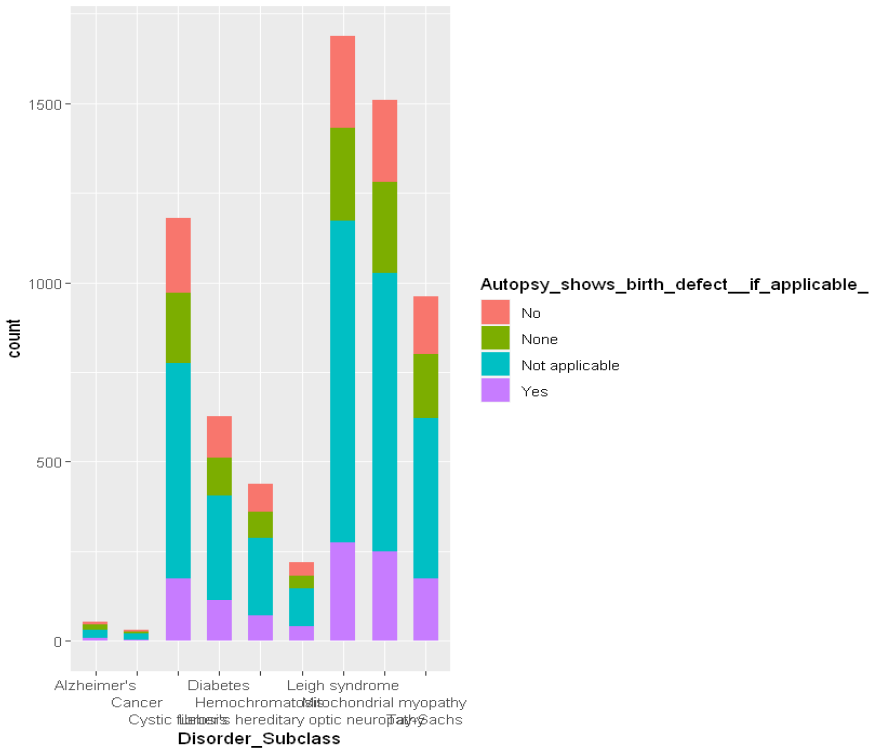

Figure 3.11: Disorder And Autopsy Report

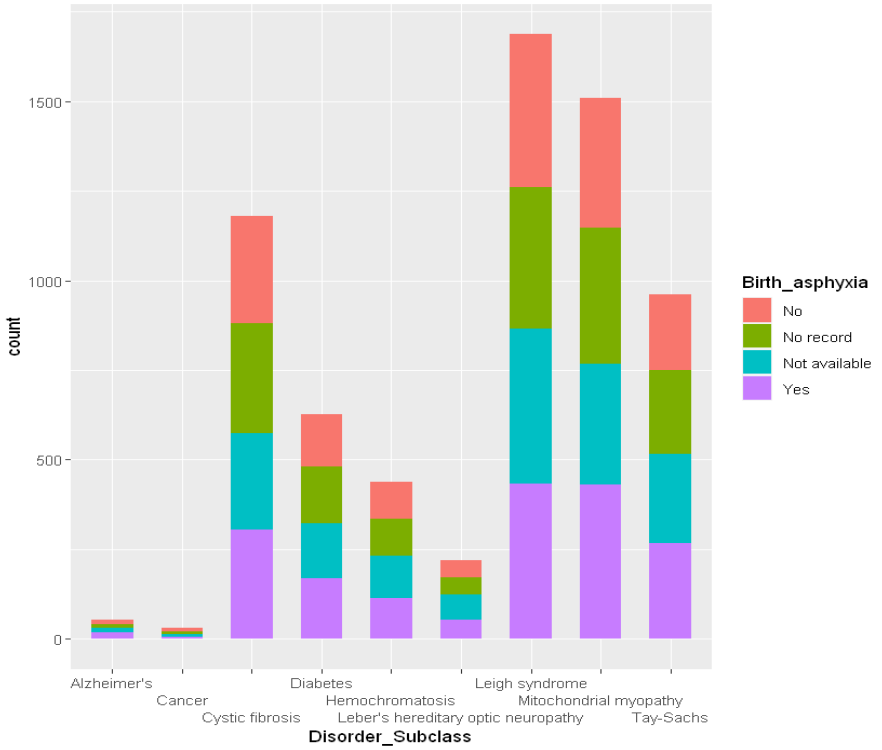

Figure 3.12: Disorder And Asphyxia

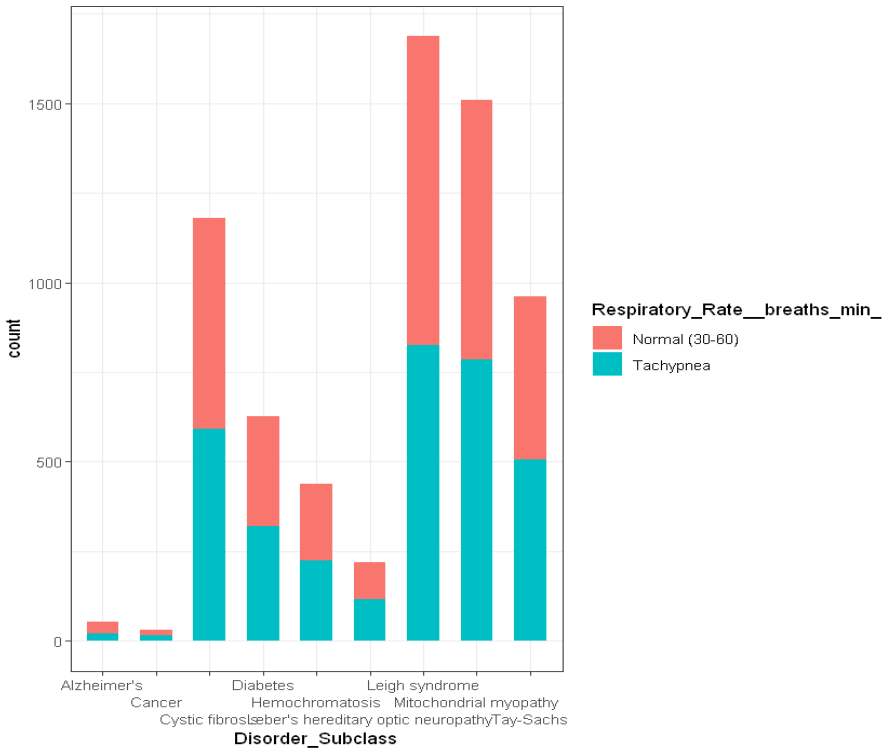

Figure 3.13: Disorder And Respiratory Rate

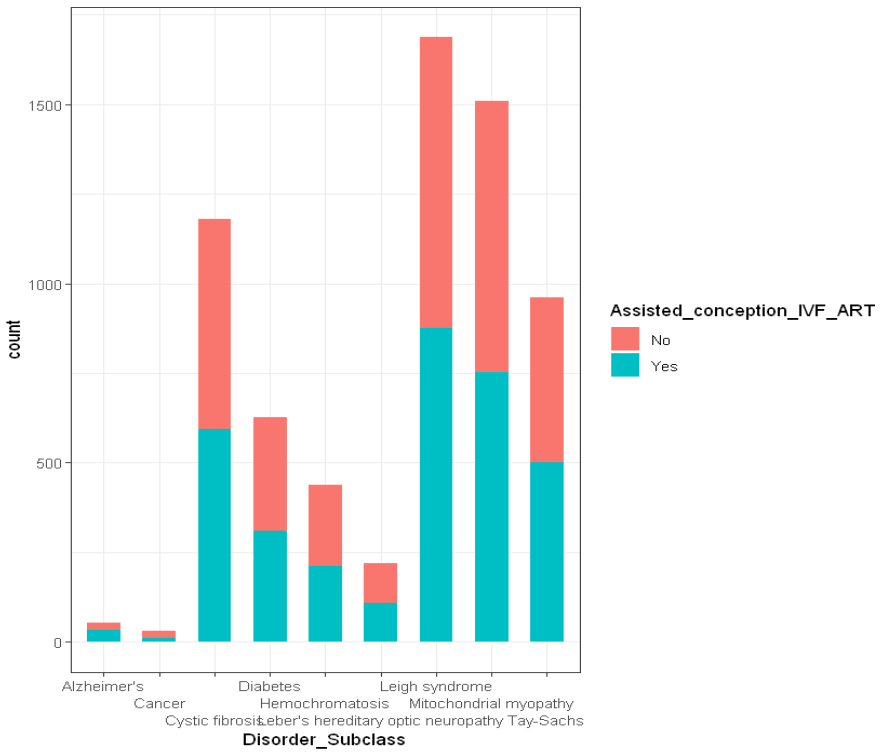

Figure 3.14: Disorder And IVF details

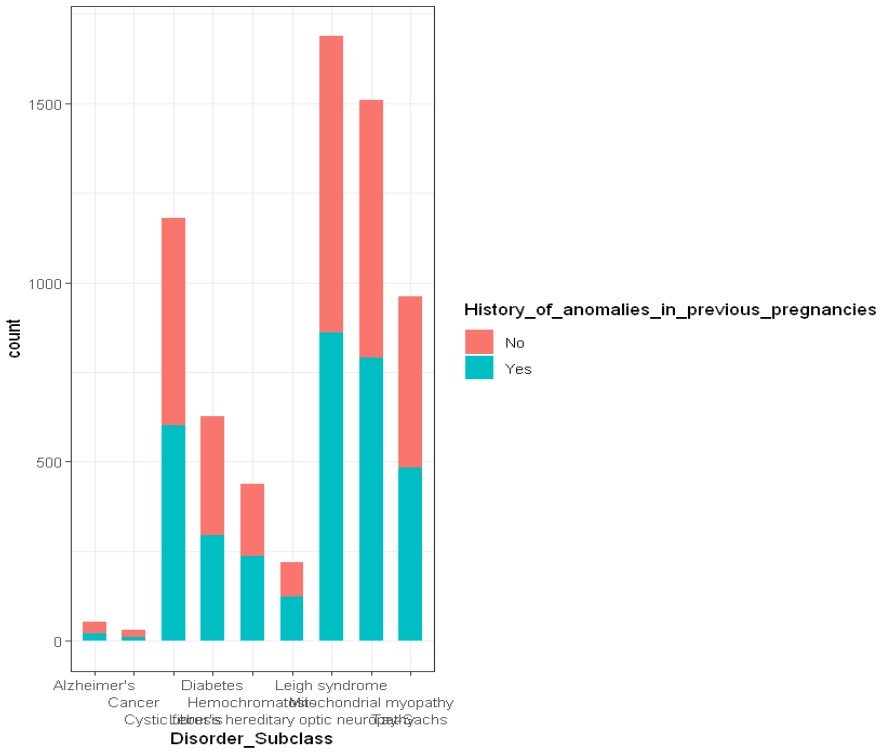

Figure 3.15: Disorder And History Of Pregnancy

We also found fully dependent relation between the two target attributes. That is, The Genetic Disorder attribute can be derived from the Disorder Subclass attribute. Hence that can also be now removed.

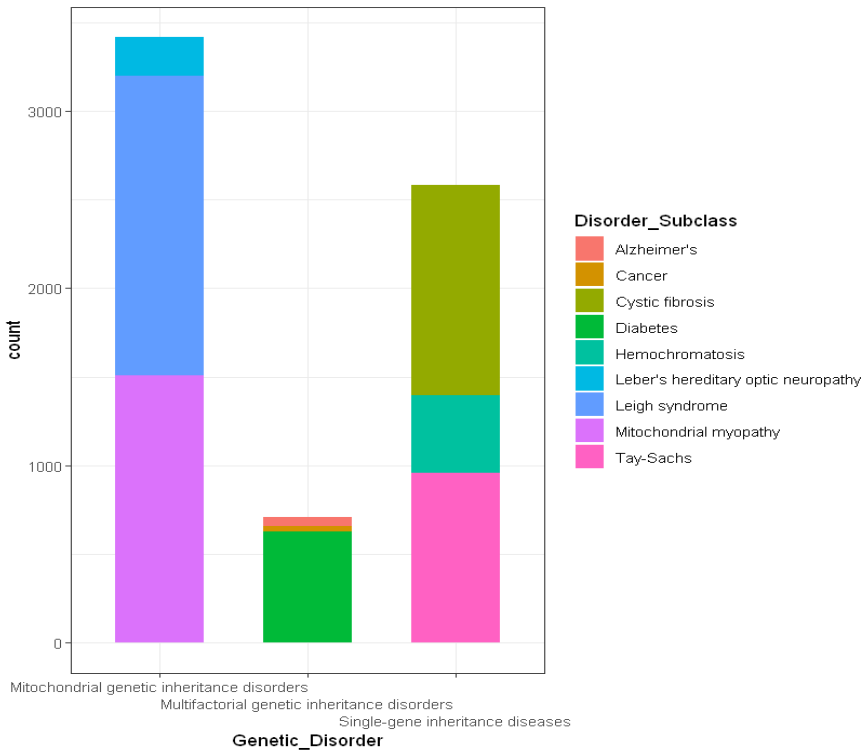

Figure 3.16: Relation Between Target Attributes

For the Numeric attributes, a disorder-wise Box and Whisker plots were created and it displayed relatively significant variance in one or more disorder.

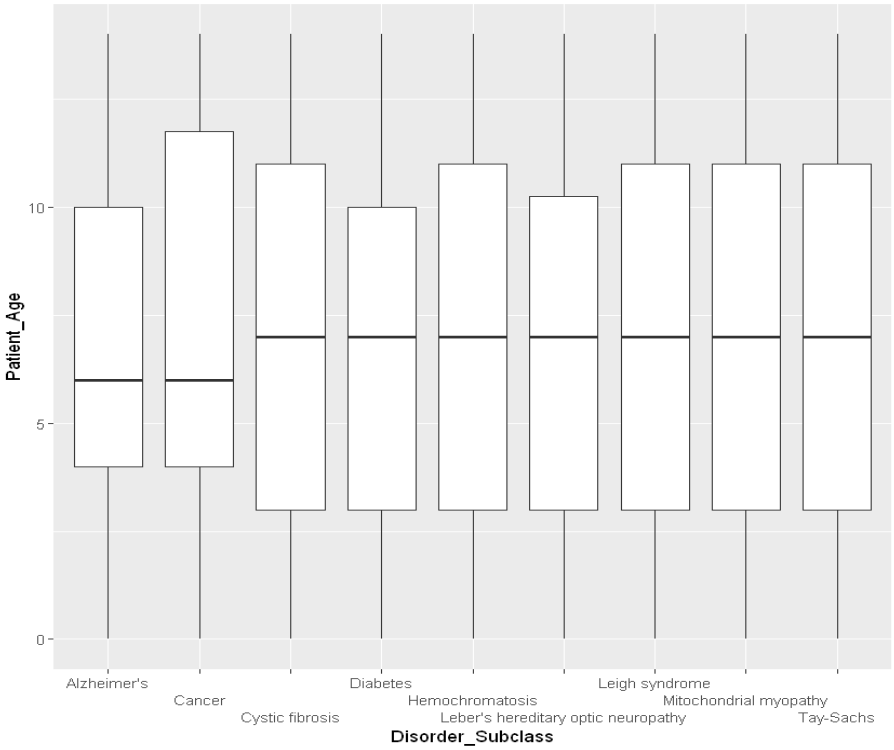

Figure 3.17: Disorder And Patient's Age

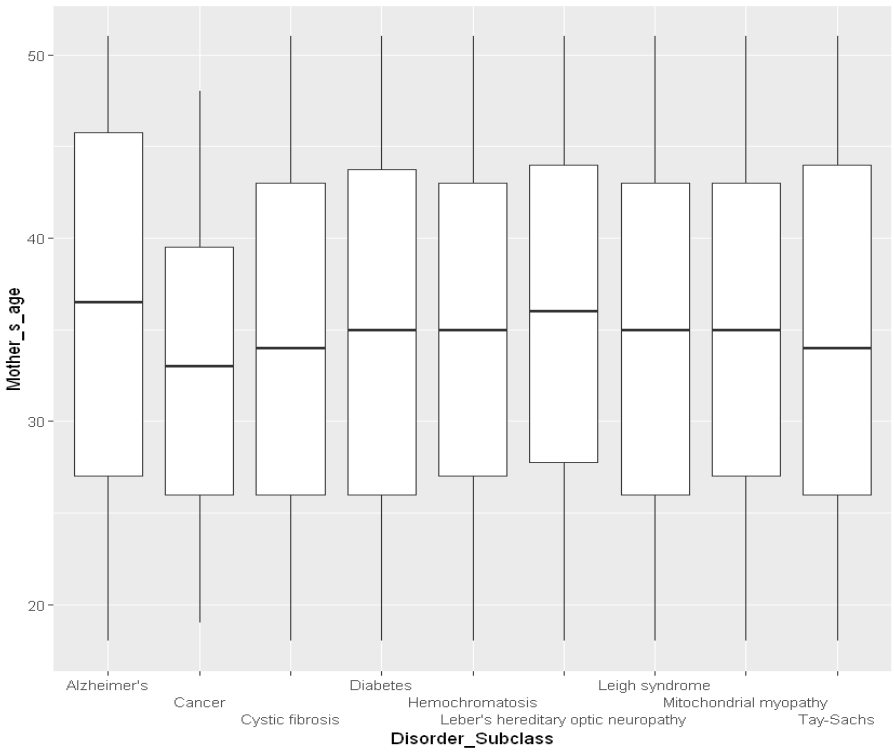

Figure 3.18: Disorder And Patient's Mother's Age

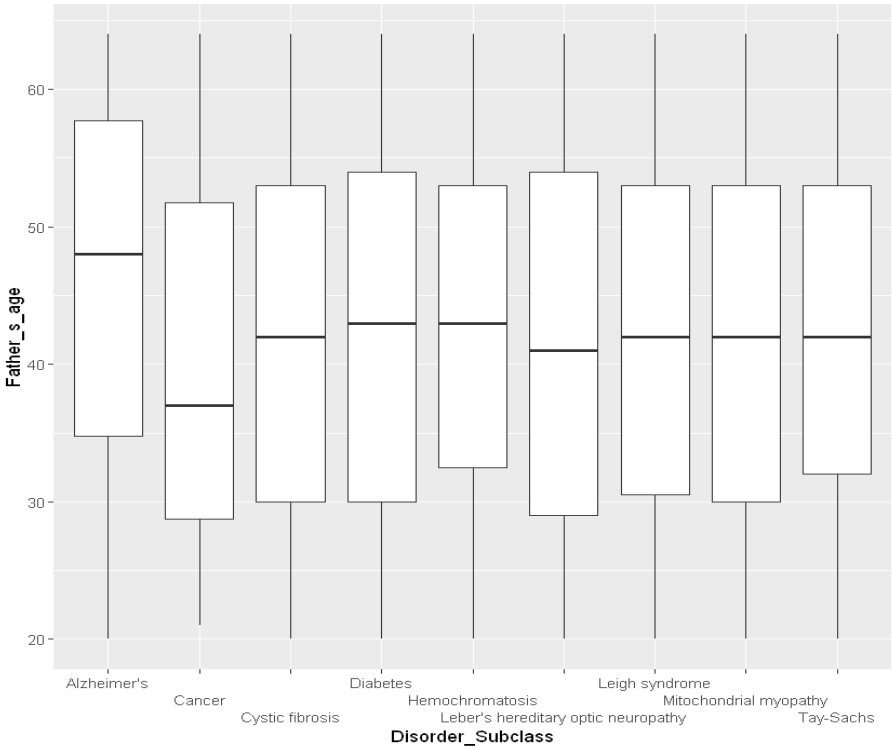

Figure 3.19: Disorder And Patient's Father's Age

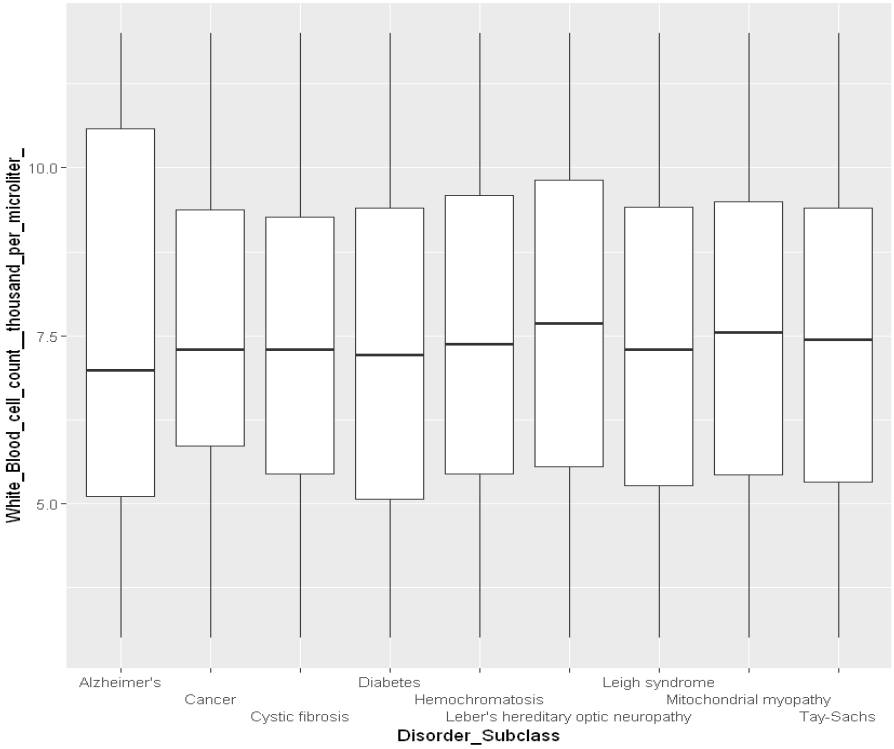

Figure 3.20: Disorder And White Blood Cell Count

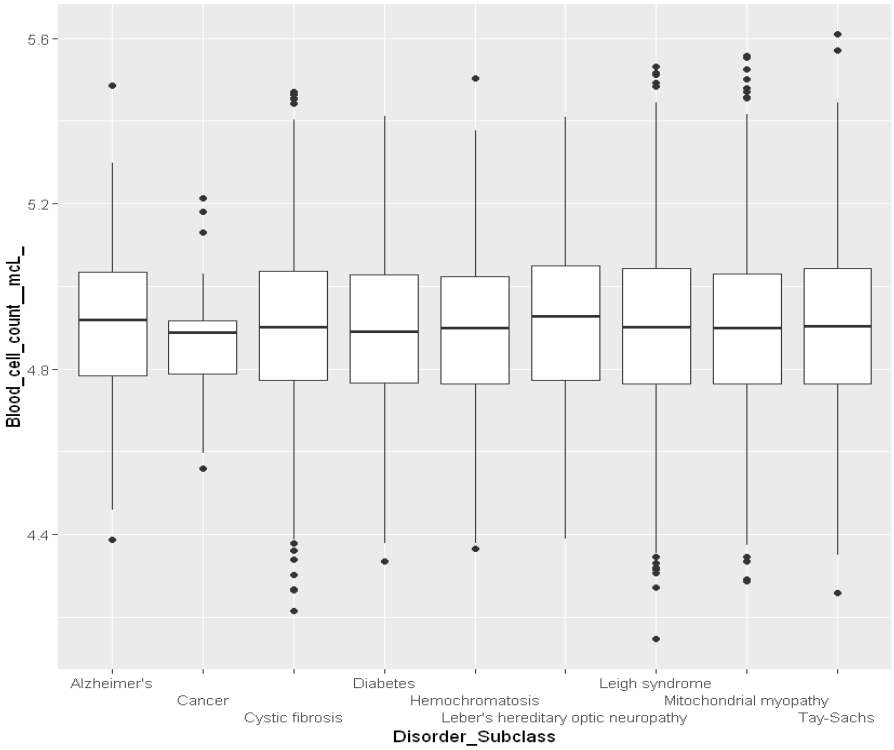

Figure 3.21: Disorder And Blood Cell Count

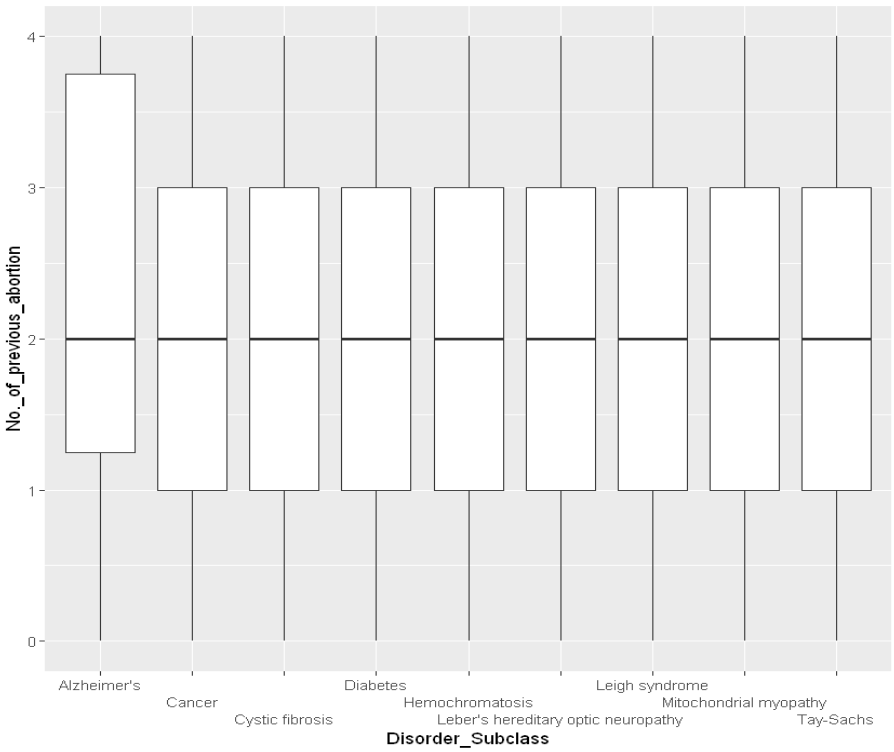

Figure 3.22: Disorder And Number Of Previous Abortions

Based on the domain expertise the following columns are considered significant and have shown a good variance as such can prove proper information to classify the disorder on future data.

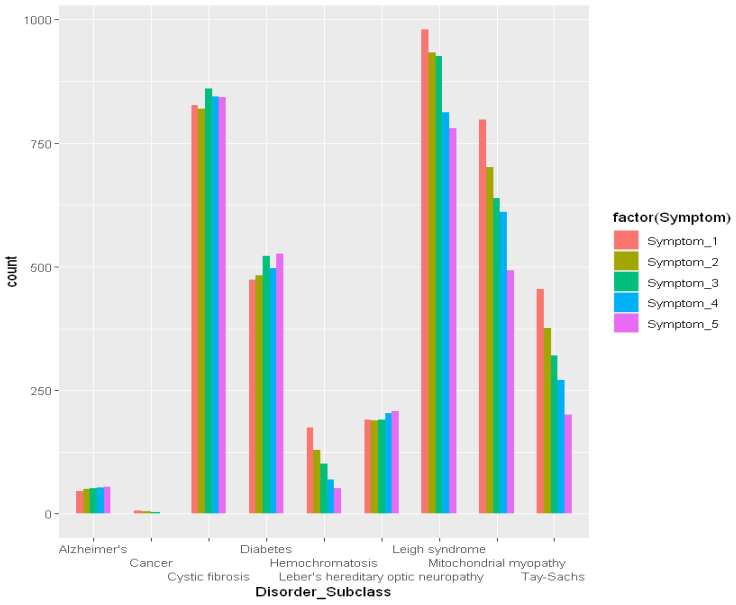

Figure 3.23: Disorder And Frequency Of Symptoms Being True

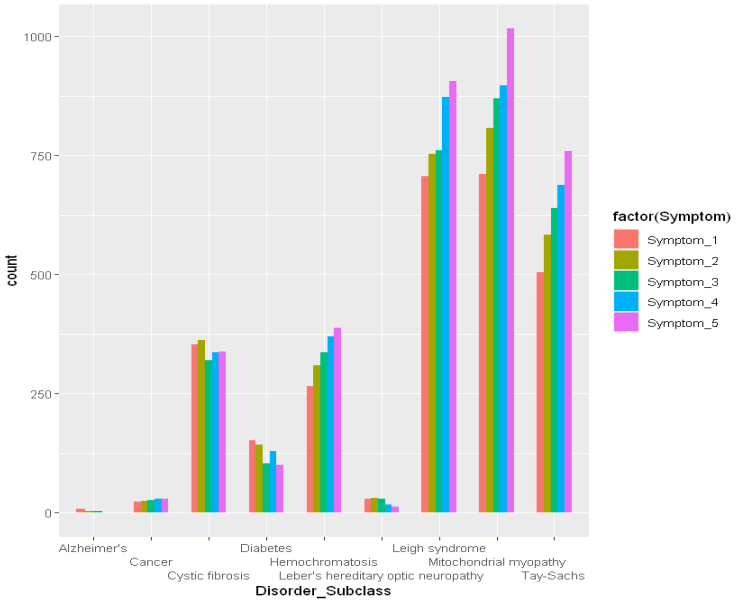

Figure 3.24: Disorder And Frequency Of Symptom Being False

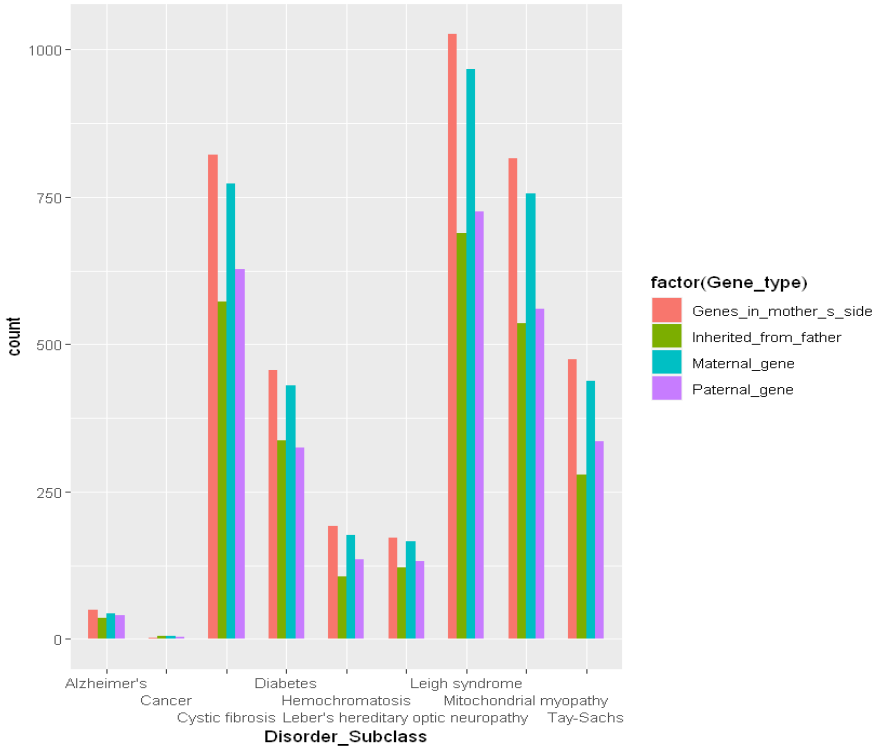

Figure 3.25: Disorder And Frequency Of Gene Being Yes

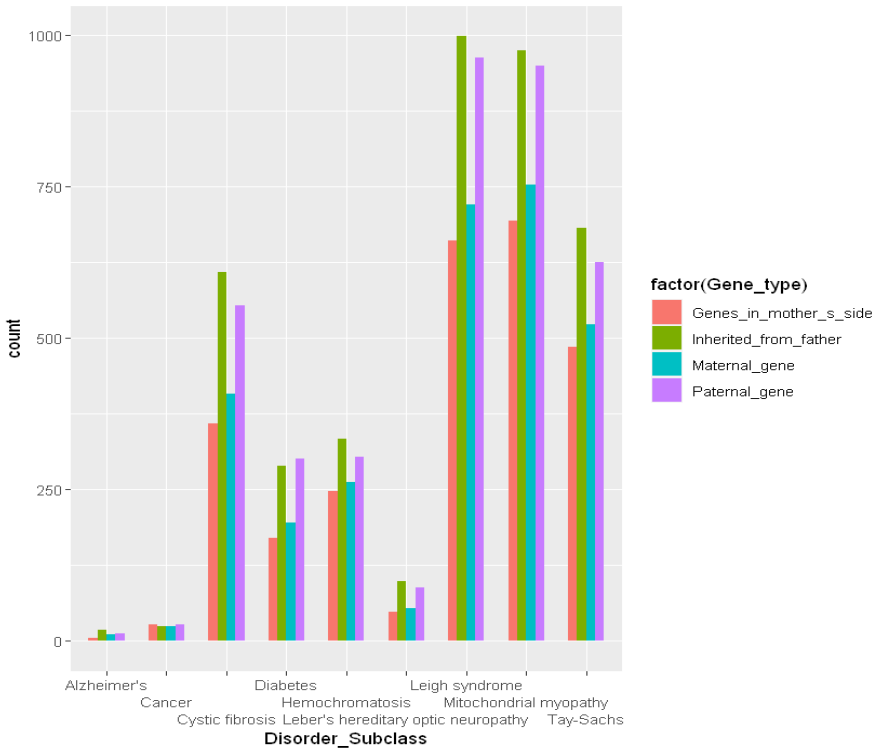

Figure 3.26: Disorder And Frequency Of Gene Being No

3.3 Dashboard

Also an interactive dashboard is implemented using `https://visual.is` which can be found in `https://visual.is/visualizations/new-visualization/bpqD6MZM9CzQb8UTKZUThdx4`.

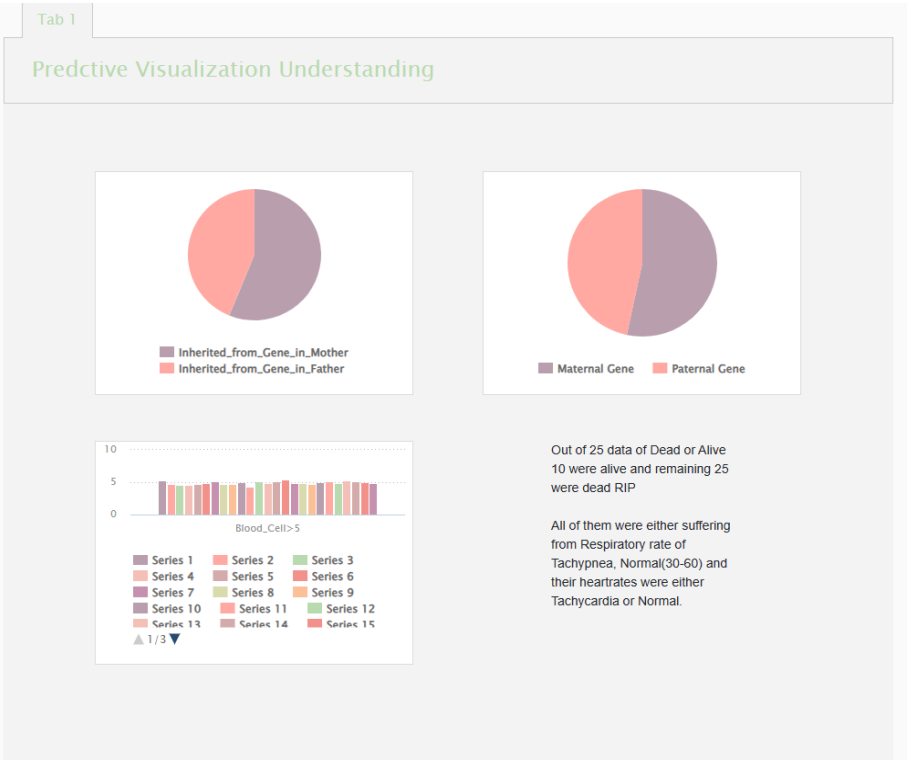

Figure 3.27: Dashboard Sample

CHAPTER 4

CONCLUSIONS

In this EDA project, we aimed to gain insights and streamline the dataset by applying various techniques including statistical analysis, graphical visualization, and domain knowledge. Our objective was to reduce the dimensionality of the dataset while retaining important and informative features. Through a comprehensive analysis, we successfully reduced the dataset from 45 columns to 16 columns.

Table 4.1: Remaining 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 thousand per microliter
Symptom 1
Symptom 2
Symptom 3
Symptom 4
Symptom 5
Disorder Subclass

Statistical analysis played a crucial role in identifying key features that significantly contributed to the dataset. By evaluating statistical measures such as mean, standard deviation, and correlation coefficients, we were able to quantify the relationships between variables and make informed decisions on feature selection.

Graphical visualization techniques were employed to visualize patterns, trends, and outliers in the data. Box plots, scatter plots, and histograms provided valuable insights into the distribution and relationships of variables. By visually examining the data, we gained a deeper understanding of its characteristics, which guided us in selecting the most relevant features.

Domain knowledge also played a significant role in feature selection. By leveraging our understanding of the subject matter, we were able to identify domain-specific attributes that were essential for analysis and decision-making. This domain-based information added an additional layer of context and relevance to the feature selection process.

Our approach was further supported by citable sources, including research papers that discuss the importance of feature selection, dimensionality reduction, and the impact of domain knowledge in data analysis. These sources provided a theoretical foundation and best practices for our EDA project.

Additionally, we implemented a dashboard to present the analyzed data in a visually appealing and interactive manner. The dashboard allowed users to explore the reduced dataset, visualize trends, and gain actionable insights efficiently. By incorporating user-friendly features and interactive elements, the dashboard facilitated data-driven decision-making and enhanced the overall user experience.

In conclusion, our EDA project successfully reduced the dataset from 45 columns to 16 columns through the application of statistical analysis, graphical visualization, and domain-based information. The feature selection process was supported by citable sources, ensuring the validity and reliability of our approach. The implementation of a dashboard provided an effective means of visualizing the analyzed data and empowered users to make informed decisions based on the insights gained.

Future work could involve further refinement of the feature selection process, exploration of advanced visualization techniques, and integration of machine learning algorithms for predictive modeling based on the reduced dataset.

REFERENCES

[1] Jiawei Han, Jian Pei, and Micheline Kamber, *Data mining: Concepts and techniques*, Morgan Kaufmann, 2011.

[2] G. Hughes and A. Likhoded, *Feature selection when attributes are missing completely at random*, Data Mining and Knowledge Discovery **32** (2018), no. 3, 623–655.

[3] Marshall D. C. Salciccioli J. D. & Crutain Y. Komorowski, M., *Exploratory data analysis*, In MIT Critical Data (Ed.), Secondary Analysis of Electronic Health Records (2016), 185–203.

[4] Tom Mitchell, *Machine learning*, McGraw-Hill, 1997.

[5] H. Peng, F. Long, and C. Ding, *Feature selection for high-dimensional data: A fast correlation-based filter solution*, Proceedings of the Twentieth International Conference on Machine Learning, 2005, pp. 856–863.