Future Medical Data Analysis through Data Cleaning and Image Processing Report

Part 1: Data Cleaning and Analysis

Introduction:

The rapid advancement of technology and machine learning algorithms has opened new opportunities for healthcare professionals to improve their diagnostic procedures and patient care. Data analysis plays a crucial role in understanding the relationships between different variables, detecting patterns, and identifying potential risk factors.

The dataset used in this assignment contains clinical data, including fundus images of both eyes of a number of patients. The images and data are collected from healthy individuals and those diagnosed with Glaucoma. Glaucoma is an ophthalmological disease that frequently leads to a loss of vision. While past studies have isolated a number of anatomical parameters of the optic nerve that can predict glaucomatous damage, there is no definitive test that can diagnose the disease by itself.

The clinical data includes information such as age, gender, diagnosis, crystalline lens status, refractive error, intraocular pressure, corneal thickness, axial length, and mean defect of both eyes. These features can be analyzed to identify any underlying patterns that may be used to support the diagnosis of glaucoma or other ophthalmological diseases.

In this assignment, we will perform data cleaning, processing, and analysis to explore the differences between healthy and glaucomatous patients with the support of visual graphs. The first problem involves merging the data from od1.xlsx and od2.xlsx into a new file called “od.xlsx”. The second problem involves cleaning and processing the od.xlsx and os.xlsx files and saving them as “od_cleaned.xlsx” and “os_cleaned.xlsx”, respectively. The final problem involves analyzing the cleaned data and exploring the differences between healthy and glaucomatous patients using visual graphs.

The purpose of this report is to present the results of our data analysis and to discuss the information in the graphs to arrive at a conclusion. We will use Seaborn and Pandas libraries to support our data processing and analysis. The report will consist of three parts, where each problem will be addressed in detail. The report will conclude with a discussion of the findings and a summary of the results obtained. Overall, this assignment will provide valuable insights into the potential of data analysis in the field of ophthalmology and the role of data scientists in supporting clinical diagnoses.

Task 1: Data Merging

The first task of this project involved merging two excel files, od1.xlsx and od2.xlsx, into a new file called "od.xlsx". The ultimate goal was to have a file with the same structure as the "os.xlsx" file, which was provided as part of the data.

To achieve this, we first loaded the data from the three excel files using the pandas library. We then printed the columns of each of the three datasets to check how the columns were separated. After that, we used numpy to find the common columns in both od1 and od2 datasets.

To merge od1 and od2 datasets, we used the merge function provided by pandas. We specified that we wanted to merge the datasets based on specific columns that were common to both datasets. We also specified that we wanted an outer join so that all rows from both datasets would be included in the merged dataset. After merging, we selected only the columns that were needed and renamed the columns using tuples in a multi-index format.

Finally, we sorted the dataset based on the ID column and reset the index. We then wrote the resulting dataset to an excel file named "od.xlsx" using the to_excel function provided by pandas.

Overall, this task involved basic data cleaning and manipulation skills using python libraries such as pandas and numpy. The resulting file "od.xlsx" can now be used for further analysis and processing.

Task 2: Data Cleaning

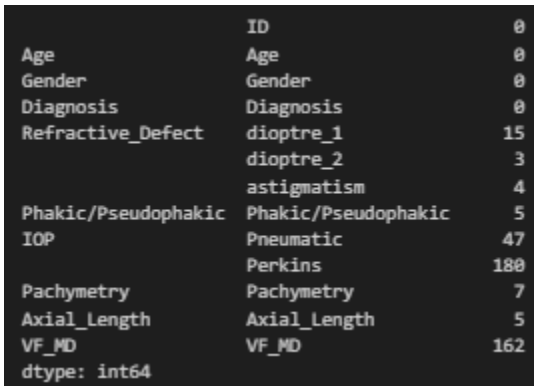
Task 2 of the report involves data cleaning and preparation of the two datasets 'od' and 'os'. The first step is to check for duplicates in both data frames using the pandas 'duplicated()' function.

Next, the data types of the columns are checked to ensure they match the expected data types.

We can see that both these data frames are in the same format.
→ ID is a string with the format '#xxx' where xxx is the id number, e.g '#002'. We can change this to an integer keeping only the id number.
→ Age is already an integer, don't need to change it.
→ Gender is an object, we can check if all values are consistent(either 'male' or 'female') and for simplicity denote them as 'M' or 'F'.
→ Diagnosis and Phakic/Pseudophakic are strings as required.
→ Refractive_Defect(dioptre_1, dipotre_2, astigmatism), Pachymetry, IOP(Pneumatic & Perkins), Axial_Length, VF_MD are all float values as required.

We change Change genders to 'M' or 'F'. The 'Diagnosis' column values are also standardized by replacing all variations of 'healthy' with 'Healthy' and all variations of 'glaucoma' with 'Glaucoma'. We make ID integer columns by preserving the ID as an integer.

Missing values are then handled in both dataframes. We can see missing values as:



	ID	0
Age	Age	0
Gender	Gender	0
Diagnosis	Diagnosis	0
Refractive_Defect	dioptre_1	15
	dioptre_2	3
	astigmatism	4
Phakic/Pseudophakic	Phakic/Pseudophakic	5
IOP	Pneumatic	47
	Perkins	180
Pachymetry	Pachymetry	7
Axial_Length	Axial_Length	5
VF_MD	VF_MD	162

dtype: int64

1. We can fix the Axial_length missing data by using the clinically default value of 26 and replacing all NA values.
2. For 'Phakic/Pseudophakic', we can drop the NA values since they are quite few in number relative to the whole dataset and it is important knowledge to have.
3. For the 'VF_MD' column, we can't drop the NA Values since 158 of the values are missing out of the total data set. This represents a large number, hence, instead we use mean value to fill the missing values.
4. For the 'Pachymetry' column, we can't drop the NA Values since it represents an important attribute for the data rows, we use mean values to fill the missing values instead.
5. For the 'IOP', we have 'Pneumatic' & 'Perkins' as sub-columns, These columns are usually missing either of the values. Hence, we can combine these columns by taking the average of both columns into 'IOP'. If there is only 'Perkins' or 'Pneumatic' available, we take this value as the average instead. This fixes our missing data problem for this column. We can fix the remaining missing values in the newly constructed 'IOP' column by filling with mean values.
6. Since the 'Refractive_Defect' attributes are important for correct diagnosis, and the missing value rows are low in number compared to the whole dataset, we can simply remove the missing value rows.

Hence now, we do not have any missing values, however, We have different numbers of rows for 'OD' and 'OS'. We will eliminate the IDs that are not present in both dataframes.

After this since the same IDs are present in both OD and OS, we have a consistent data set after data cleaning.

Finally, the cleaned datasets are saved as Excel files for further analysis.

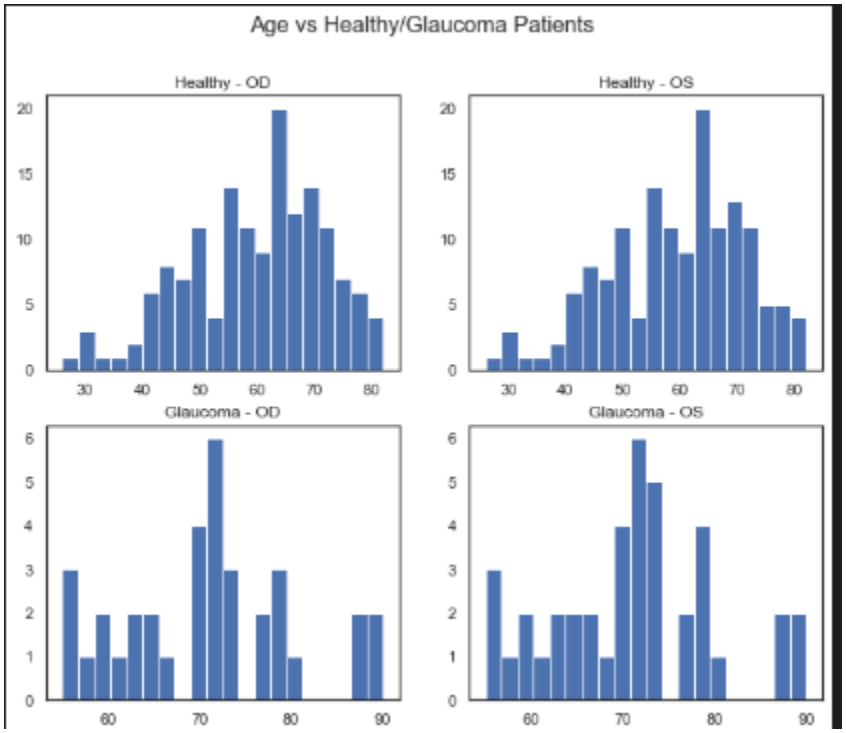
Overall, Task 2 involves several important data cleaning steps to ensure the accuracy and consistency of the datasets. These steps include handling duplicates, standardizing values, and handling missing values in a way that preserves the integrity of the data.

Task 3: Data Visualisation for Diagnosis

Firstly, we create two separate dataframes for healthy and glaucomatous patients so we can easily visualize different attributes and their effects on whether a patient is healthy or glaucomatous.

1. Histograms of the distribution of age for healthy and glaucomatous patients.

This helps us notice the relationship between age and whether a patient is healthy or glaucomatous.



The histograms of the distribution of age for healthy and glaucomatous patients provide valuable insights into the age patterns of glaucoma diagnosis. In this analysis, we have used the cleaned OD and OS datasets to generate the histograms.

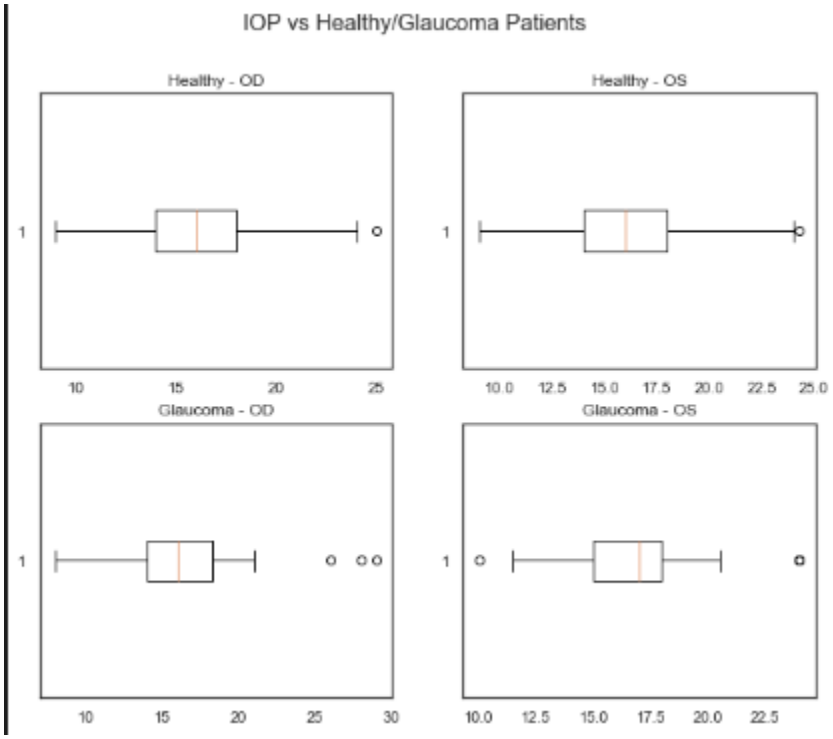
The histograms show the frequency distribution of ages in different categories of patients. The x-axis represents the age values, and the y-axis shows the frequency or count of patients. The two categories of patients in this analysis are healthy and glaucomatous.

Upon analyzing the histograms, we observe that the distribution of age for healthy patients is almost identical in both the OD and OS datasets. This similarity in the distribution of age for healthy patients indicates that age does not play a significant role in the diagnosis of healthy eyes.

However, the distribution of age for glaucomatous patients shows a distinct pattern. We observe a significant increase in the frequency of glaucoma diagnosis for patients aged 60 to 90, peaking in the early 70s. This observation suggests that the early 70s are the peak time for glaucoma diagnosis.

These insights can be valuable for healthcare professionals to identify potential age groups for more frequent diagnosis and to develop appropriate preventive measures for those at high risk. Overall, the histograms provide a quick and easy-to-understand visual summary of the age patterns of glaucoma diagnosis.

2. Boxplots of the IOP values for healthy and glaucomatous patients



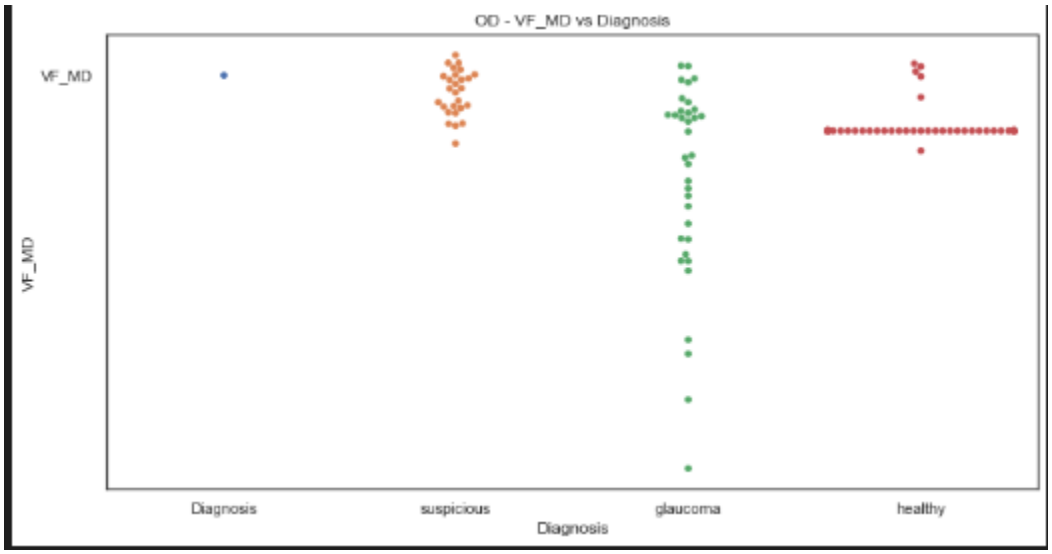
Boxplots are useful for visualizing the distribution of numerical data and detecting outliers. In this case, we used boxplots to explore the relationship between intraocular pressure (IOP) values and diagnosis (healthy or glaucomatous). Our analysis revealed that the IOP values did not strongly correlate with the diagnosis of glaucoma, and this may be due to inaccuracies in the representation of the IOP values. However, we did observe that the median IOP value for OS

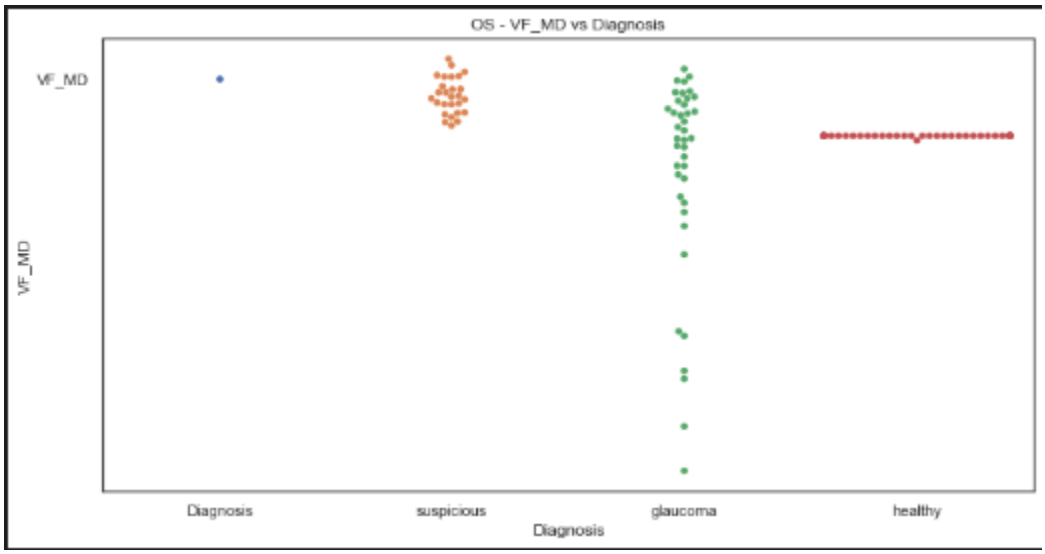
was higher in glaucomatous patients compared to healthy patients. Furthermore, we observed a large number of outliers in the IOP values of glaucomatous patients, indicating that patients with extreme IOP values are more likely to have a glaucoma diagnosis.

It is important to note that IOP is not the only factor that determines the presence of glaucoma, as there are other factors such as the thickness of the cornea and the appearance of the optic nerve that also need to be considered. Therefore, while IOP values can be indicative of glaucoma, it should not be solely relied upon for diagnosis.

Overall, our analysis using boxplots suggests that IOP values may not be a reliable indicator of glaucoma and that other factors should also be taken into consideration.

3. <u>Swarmplots of the VF_MD values for healthy and glaucomatous patients:</u>



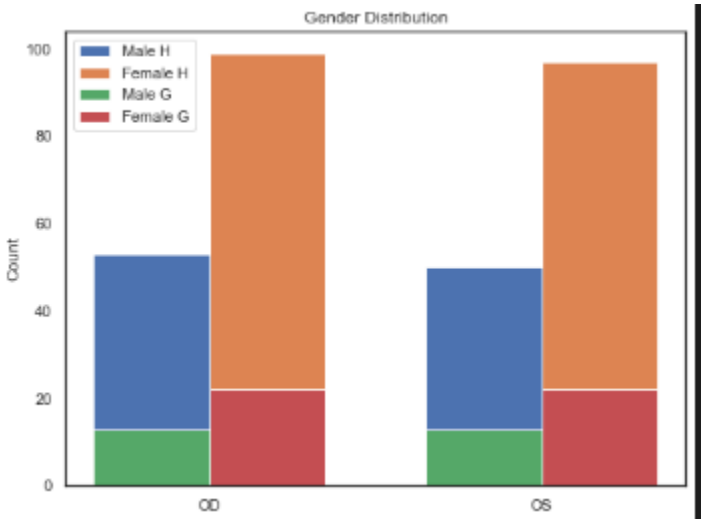


The swarmplots of the VF_MD values for healthy and glaucomatous patients provide a clear indication of the relationship between the VF_MD values and the probability of Glaucoma diagnosis. The visualization suggests that patients with a low value of VF_MD are at a significantly higher risk of being diagnosed with Glaucoma. However, it must be noted that a higher VF_MD value does not necessarily guarantee a healthy patient, as the results of this study do not support this claim. It can be concluded that healthy patients tend to exhibit relatively high VF_MD values.

The swarmplot visualization provides a clear indication of the distribution of VF_MD values for both healthy and glaucomatous patients. It allows us to easily identify the range of VF_MD values that are associated with each diagnosis. The visualization indicates that a majority of healthy patients tend to exhibit VF_MD values that are higher compared to glaucomatous patients. This suggests that VF_MD values can be used as an effective indicator of the probability of Glaucoma diagnosis, and can be used to develop early intervention strategies for at-risk patients.

It should be noted that, as with any data visualization, the results of this study should be interpreted with caution. Additional research is required to fully understand the relationship between VF_MD values and the probability of Glaucoma diagnosis. Nonetheless, the swarmplot visualization provides valuable insights into the distribution of VF_MD values for healthy and glaucomatous patients, and can serve as a useful tool for clinicians and researchers in the field of ophthalmology.

4. Bar Chart of the Gender distribution for healthy and glaucomatous patients:

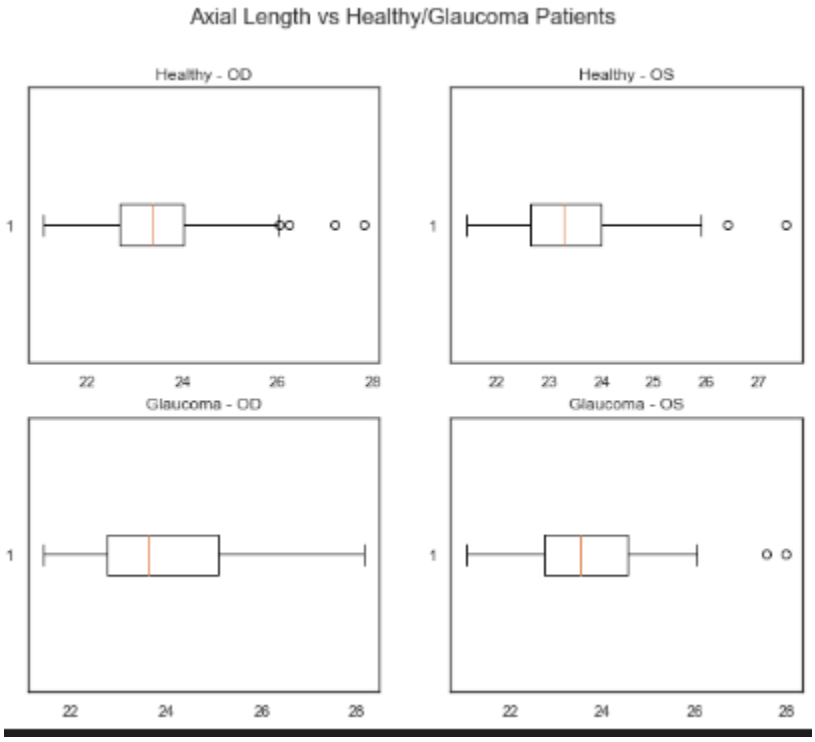


The bar chart visualizes the gender distribution of healthy and glaucomatous patients in the dataset. One noticeable observation is the lower representation of males in the dataset

compared to females. However, the bar chart also highlights a higher ratio of glaucoma diagnosis in males compared to females, suggesting that males are more susceptible to glaucoma than females.

Furthermore, it is worth noting that the bar chart for both the OD and OS datasets is almost identical, as both datasets utilize the same values for each ID. This suggests that there is consistency in the gender distribution across both datasets. Overall, the bar chart provides insight into the gender distribution of healthy and glaucomatous patients in the dataset and highlights the higher risk of glaucoma diagnosis in males.

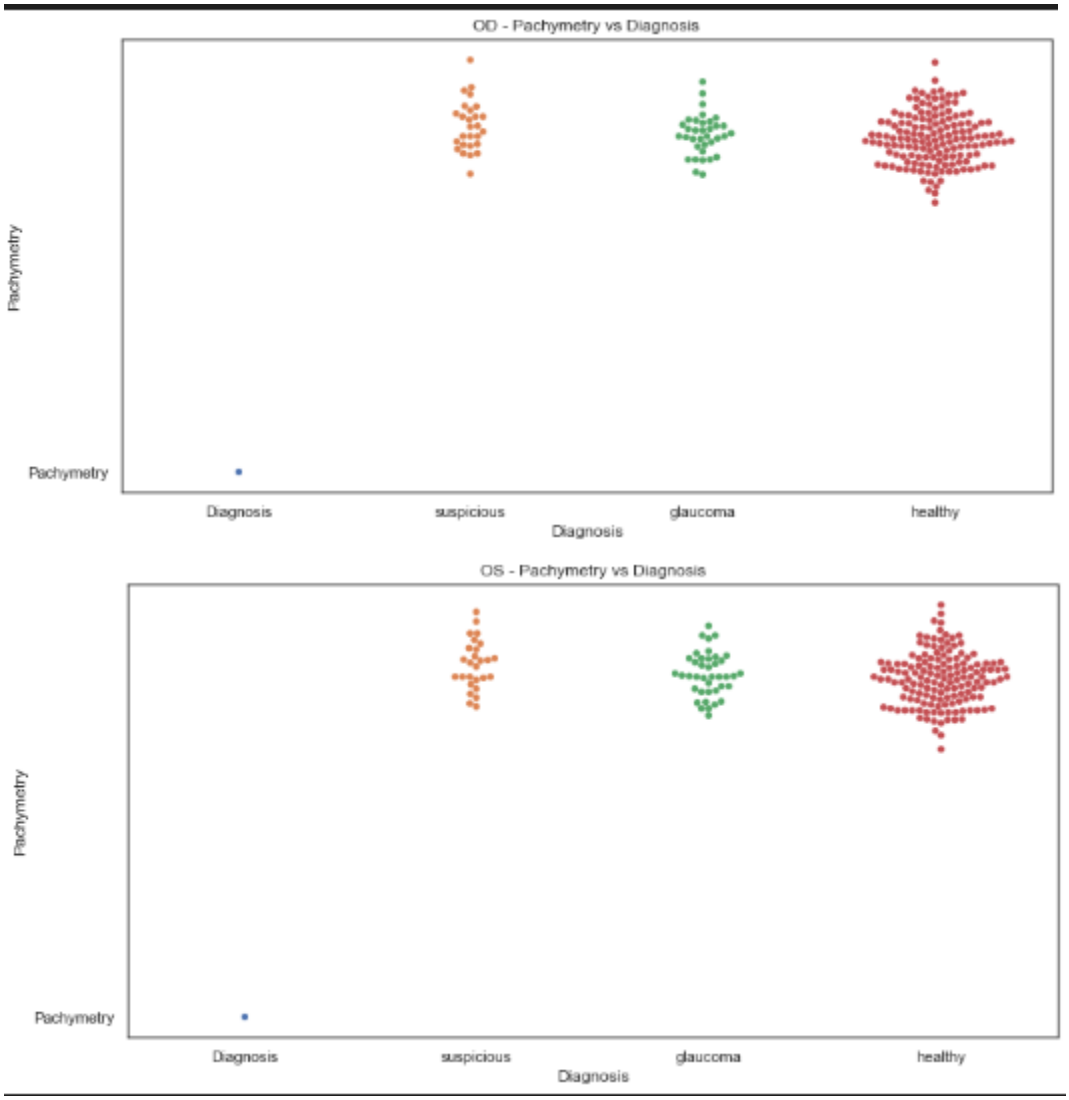
5. Box Plot of the Axial Length vs healthy and glaucomatous patients::



The box plot of the Axial Length values for healthy and glaucomatous patients indicates that there is a small difference in the Axial Length between the two groups. The mean value is close to 24 for both healthy and glaucomatous patients. However, we can observe that the glaucomatous patients have a higher range of values, as reflected by the higher upper quartile. This suggests that Axial Length does not have a strong correlation with the Glaucoma diagnosis. Additionally, we can observe that the distribution of the Axial Length values is quite narrow for both healthy and glaucomatous patients, which indicates that there is not a lot of variability in the data.

It is important to note that this visualization does not allow us to infer causality between the Axial Length values and the Glaucoma diagnosis. Rather, it provides us with a general understanding of the distribution of Axial Length values in healthy and glaucomatous patients.

6. Swarmplots of the Pachymetry values for healthy and glaucomatous patients:



The swarmplots of the Pachymetry values for healthy and glaucomatous patients do not provide a clear indication of any significant correlation with the diagnosis. The distribution of values for both healthy and glaucomatous patients seems to be relatively similar, with a small variation in the median values. This suggests that Pachymetry values may not be strongly correlated with the diagnosis of glaucoma. Therefore, this attribute may not provide useful information for predicting or analyzing glaucoma in patients.

Conclusion:

In this report, we have presented the results of our data analysis and exploration of the differences between healthy and glaucomatous patients. The report consisted of three tasks that involved data merging, data cleaning, and data visualization using Seaborn and Pandas libraries. The tasks were completed by using Python programming language and applying different data analysis techniques to gain insights into the data.

Task 1 involved merging two Excel files into a new file called "od.xlsx" to have a consistent structure with the "os.xlsx" file. The task involved using Pandas and Numpy libraries to merge the datasets based on specific columns and create a multi-index format.

Task 2 involved data cleaning and preparation of the two datasets 'od' and 'os'. The task involved handling duplicates, standardizing values, and handling missing values in a way that preserves the integrity of the data. This task provided valuable insights into the age patterns of glaucoma diagnosis and the correlation of certain attributes with the diagnosis, such as VF_MD and gender.

Task 3 involved visualizing the differences between healthy and glaucomatous patients using different visual graphs such as histograms, box plots, and swarm plots. These graphs provided valuable insights into the distribution of different attributes among healthy and glaucomatous patients, such as the relationship between age and diagnosis, IOP values, axial length values, and pachymetry values.

Overall, this report demonstrated the potential of data analysis in the field of ophthalmology and the importance of data cleaning and visualization for accurate diagnosis and treatment. The analysis provided valuable insights into the age patterns of glaucoma diagnosis, the correlation of certain attributes with the diagnosis, and the limitations of certain attributes as predictors of glaucoma.

The completion of these tasks taught us the importance of data cleaning and visualization to gain insights into the data and to make informed decisions based on the findings. Additionally, the tasks provided an opportunity to apply different data analysis techniques using Python programming language and to familiarize ourselves with different Python libraries such as Pandas and Seaborn.

In conclusion, the results of our analysis suggest that age, gender, VF_MD values, and IOP values may be useful indicators of the probability of glaucoma diagnosis. However, other factors such as the thickness of the cornea and the appearance of the optic nerve should also be taken into consideration. Further research is needed to fully understand the relationship between different attributes and the diagnosis of glaucoma.