Data Cleaning and Analysis:

Problem 1:

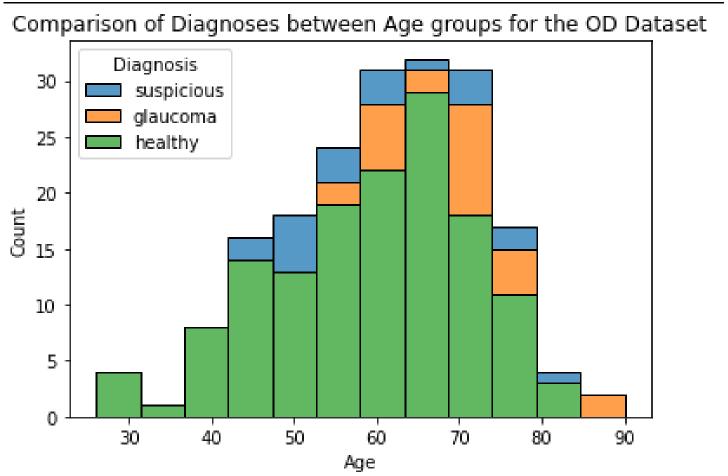
To deal with data merging, I loaded the 2 OD files, merged them based on common columns, sorted by an ascending ID value, and rearranged the columns to match the OS file format. I then saved the result as 'od.xlsx' as required and adjusted the formatting using the openpyxl library to match the OS file's appearance including borders and bold fonts.

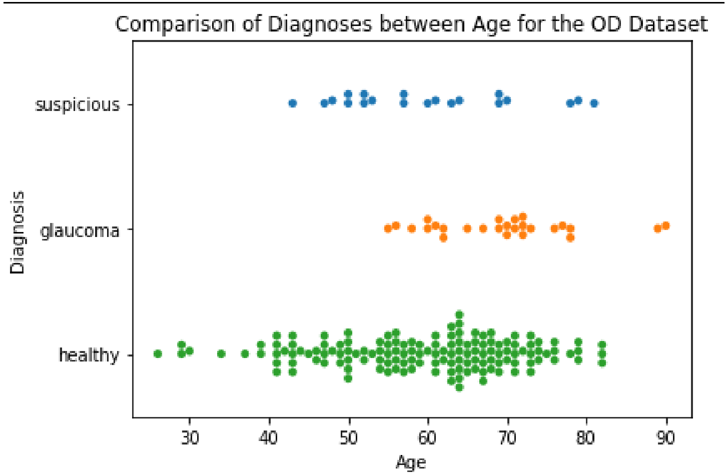
Problem 2:

During data cleaning, I checked for both the diagnosis and gender columns for unique values. I found spelling errors in the diagnosis column. To address this I turned the whole column lowercase and made a spell map to correct remaining spelling errors. I then calculated the average and standard deviation for each column and removed rows with values more than 3 standard deviations from the mean. This was a compromise I chose to make which will have potentially removed some useful data, however it will have also removed outliers that existed in the datafile due to errors. I then also chose to remove rows missing essential data, such as any row missing an IOP value or any of the following : dioptre 1 or 2 , astigmatism, phakic/pseudophakic values. After ensuring data for both eyes was available by ensuring if data was removed from one data set I eliminated the corresponding patient data in the other file as well. The remaining sample size was 188, reduced from 244, with the VF_MD column being the only incomplete one. However this means I have lost nearly a quarter of the entries through my cleaning process and as such may lose some insights.

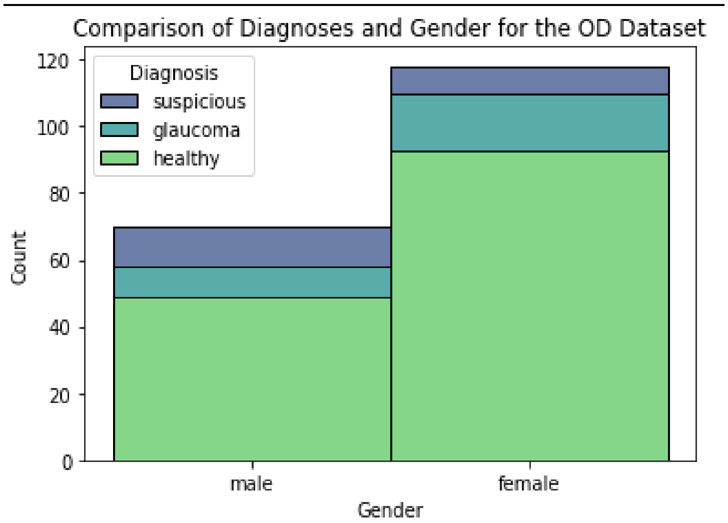
Problem 3:

There are some key relationships that can be observed from the datasets, the most intuitive is the relationship between diagnosis and age. The data displays a tendency for the risk of glaucoma to increase with age. The graphs suggest that glaucoma is present mostly in patients older than 55, however the majority of suspicious diagnoses occur prior to this in middle aged patients aged between 40 and 54. However, there are healthy people across almost the entirety of age ranges, consequently, although there is unquestionably a strong correlation between a patient's age and glaucoma, age doesn’t itself guarantee the onset of the disease.



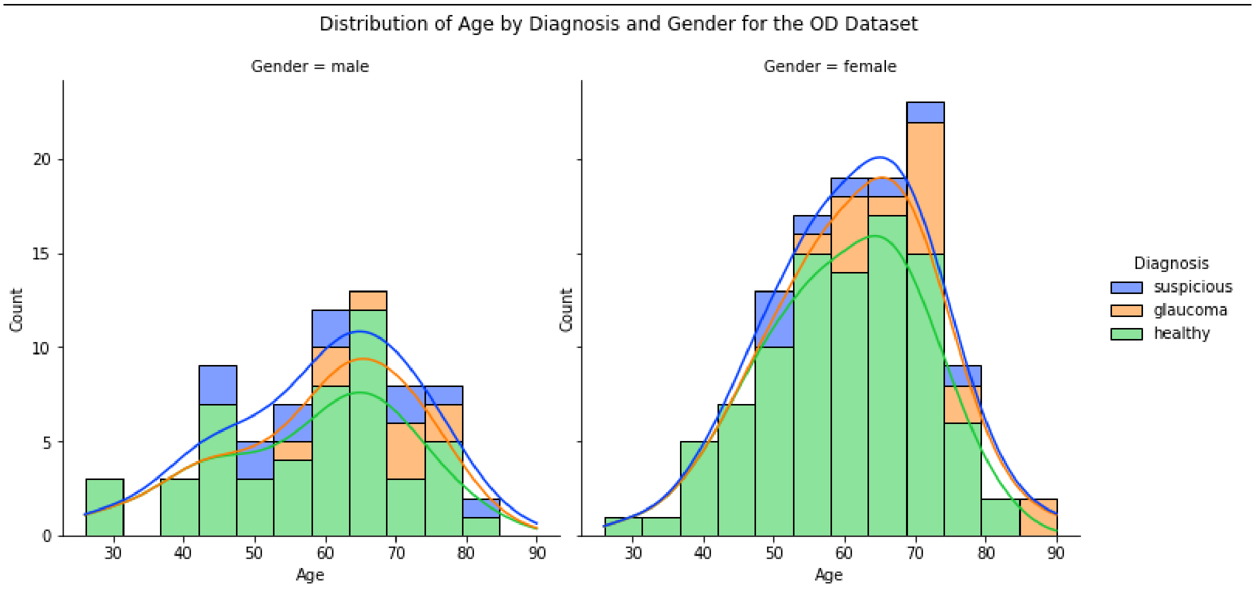


Another relationship observed is between the diagnosis of patients and their gender. It is observed from the histogram below that more female patients were seen in the clinic. It is also observed that proportionally, females have a greater tendency to have glaucoma. However if you include the 'suspicious' data this relation is skewed in favour of males.

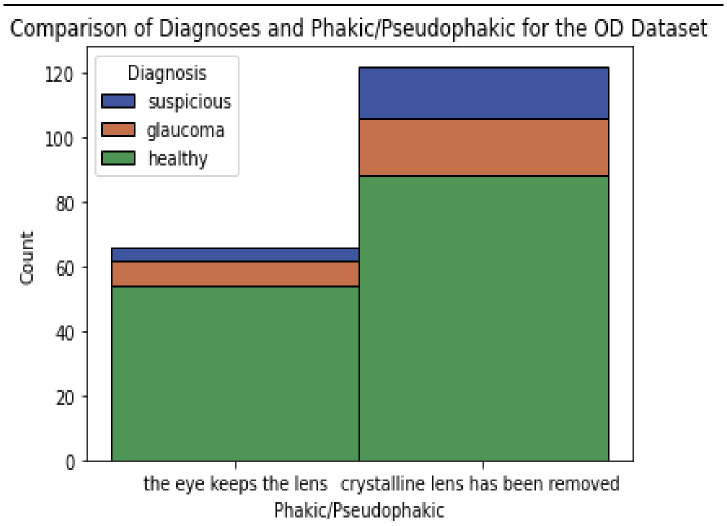


	glaucoma	suspicious	healthy	Total	glaucoma proportion	suspicious proportion	healthy proportion
male	9	12	49	70	0.129	0.171	0.700
female	17	8	93	118	0.144	0.068	0.788

Interestingly, when age and gender are simultaneously investigated, the distributions and trends in diagnosis are similar, however the distribution of the 'male' graph does not follow the expected bell curve as strongly. This may suggest a larger sample size would be of benefit in order to better compare the difference gender is having on the diagnosis. From the graph however it appears that suspicion for glaucoma occurs earlier on in life in males than females.

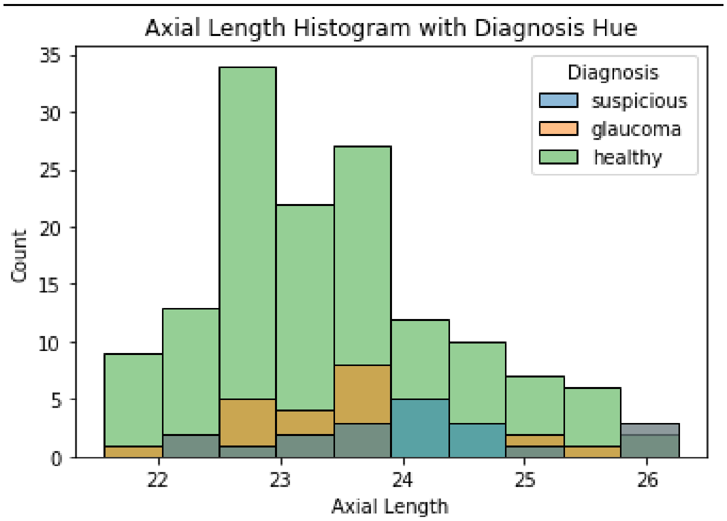


In the same way gender and diagnosis are displayed above, the below graph compares the presence of the crystalline lens and the diagnosis. The proportions show that there is a slightly higher tendency for glaucoma to occur when the lens is removed; this effect is further emphasised if the suspicious values are included. Nearly 30% of the patients missing the lens are suspicious or confirmed to have glaucoma whereas when the lens is kept this value is only 18%.

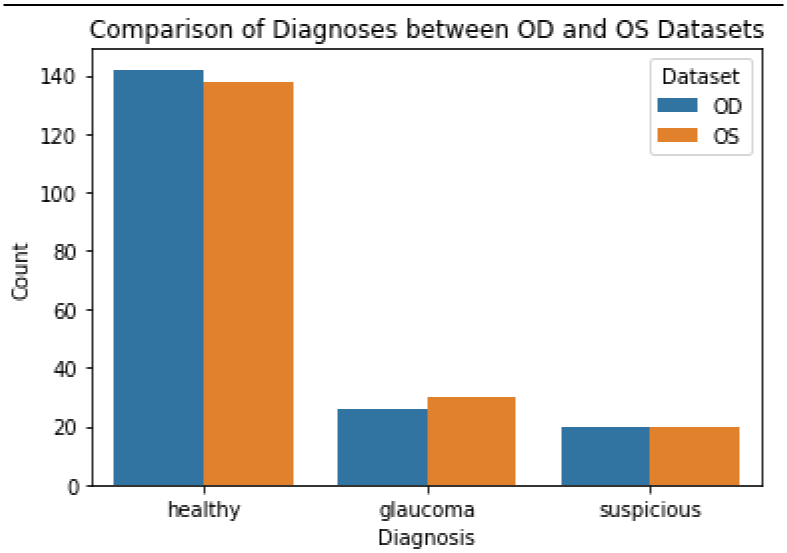


	suspicious	glaucoma	healthy	Total	suspicious proportion	glaucoma proportion	healthy proportion
lens kept	4	8	54	66	0.061	0.121	0.818
lens removed	16	18	88	122	0.148	0.148	0.721

An interesting trend in the more quantified data below shows a strong correlation occurs between the Axial length and diagnosis in which the majority of glaucoma patients have an axial length under 24mm and nearly 95% of all glaucomic and suspicious patients have an axial length under 25mm. This may suggest a causal relationship .



A final observation found was that although incredibly similar, the diagnosis between the left and right eyes are not exactly the same, meaning that although for the majority of patients that have glaucoma in one eye it is very likely they have glaucoma in both, it is not a certainty.



In conclusion, factors like age, gender, presence of crystalline lenses, and axial length contribute to glaucoma diagnosis, but a larger sample size and further investigation are needed to better understand their influence and a deeper understanding of retinal biology is needed to determine causality.