Data Cleaning and Analysis Report

Module Name: COMP2271

Date: 23/04/23

Submitted as part of the degree of MSc Natural Sciences to the Board of Examiners in the Department of Computer Sciences, Durham University

I. Data Merging

First, we imported the datasets using pandas and performed a left merge using the common columns. Performing a left merge ensured that the 5 values present in both datasets were matched up, and all other columns from both datasets were included in the merge. We sorted the data by ID and formatted the sub-columns to match OS.

II. Data Cleaning

Our first task in preparing the datasets was to modify the format of the OS and OD datasets by merging sub-columns and renaming them using an underscore. After this, we discovered that the data type of the numeric columns in OS were all objects, so we changed them to ‘float64’. Several columns had a considerable number of missing values and there were some unexpected minimum and maximum values that we decided to investigate further using boxplots. Additionally, we noticed some typos in the diagnosis column (‘heal.’ and ‘glau.’), we replaced these with ‘healthy’ and ‘glaucoma’.

When deciding which outliers to remove, we were careful in our selection process. For instance, while RE_dioptre_1 had values ranging from -8 to 11.25, we chose not to remove any, since most online resources [1] suggest that these values could be valid. However, RE_dioptre_2 included outliers of -200 and -70 in OD and OS, respectively (Figures 1 and 2). While we considered removing values outside of a specific threshold, our lack of domain made it challenging to identify the threshold. After their removal, the boxplots looked much better (Figures 3 and 4).

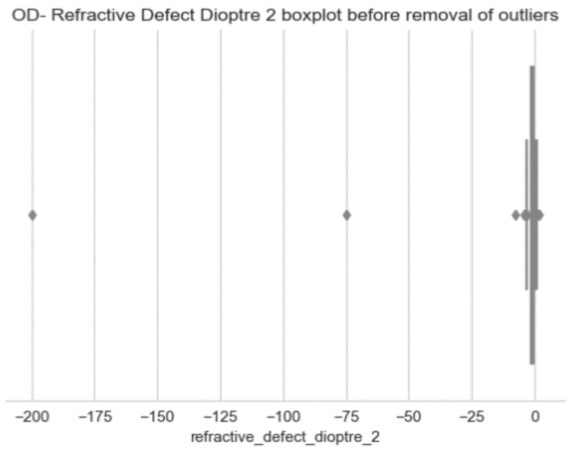


Figure 1- OD- Refractive Defect Dioptre 2 boxplot before outlier removal

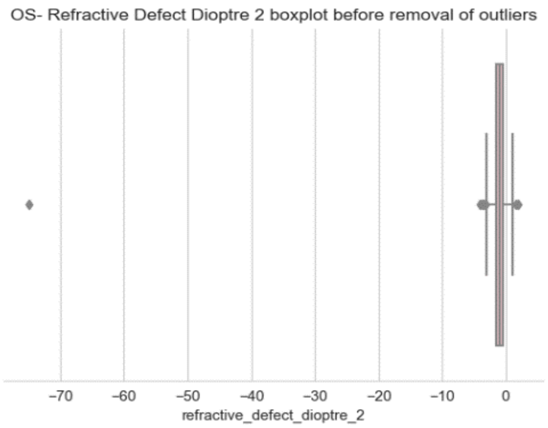


Figure 2- OS- Refractive Defect Dioptre 2 boxplot before outlier removal

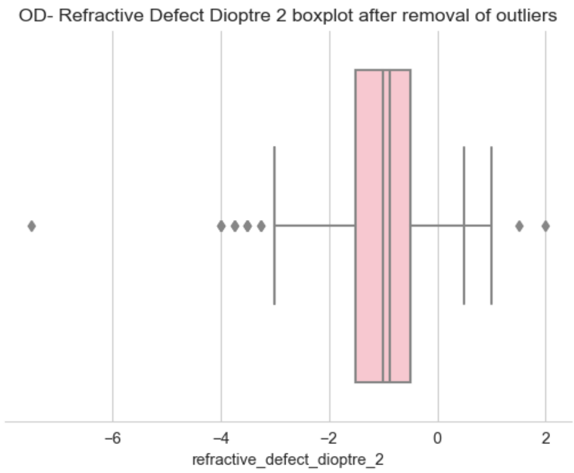


Figure 3-OD- Refractive Defect Dioptre 2 boxplot after outlier removal

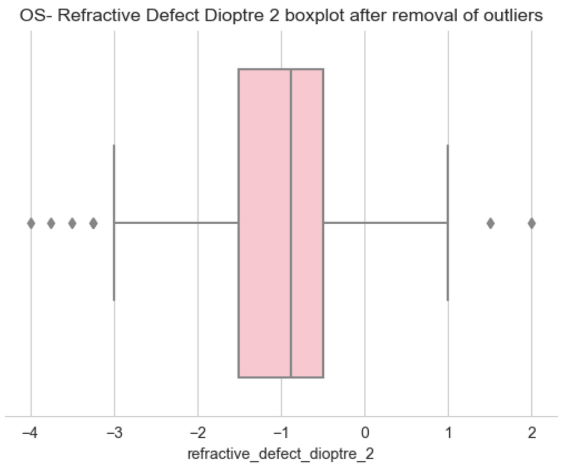


Figure 4-OS- Refractive Defect Dioptre 2 boxplot before outlier removal

Our next step was to replace the missing values. We replaced Dioptre_1, Dioptre_2, and Astigmatism with 0 since it implies no bias towards the negative or positive values. IOP_perkins and VF_MD had 180 and 162 missing values, respectively. We considered several options for dealing with these high numbers of missing values and concluded that replacing them with any value would significantly affect data visualization. Thus, we decided to leave these values alone.

The final step was to replace the values in the phakic/pseudophakic column with ‘phakic’ or ‘pseudophakic’ since we discovered which values they matched up to [2]. While we considered binarizing our categorical data, we decided against it since it would make our visualization less clear.

III. DATA VISUALISATION

Before creating our visualisations, we plotted the distributions of each column for both datasets on the same axis and found them to be very similar. As the eye’s side had little effect on the data, we decided to merge our datasets. This made the visualisation process more efficient, providing a larger sample size and allowing us to present our observations in a single graph. We also added a column to our dataset called ‘dioptres_mean’ after the discovery that ‘dioptre_1’ and ‘dioptre_2’ were repeated measurements of the same value [3].

We began our analysis by plotting the distribution of diagnoses (Figure 5), which revealed that the proportion of ‘healthy’ patients was significantly higher than that of ‘glaucoma’ and ‘suspicious’ diagnoses combined. This was important to consider when interpreting the rest of our visualisations.

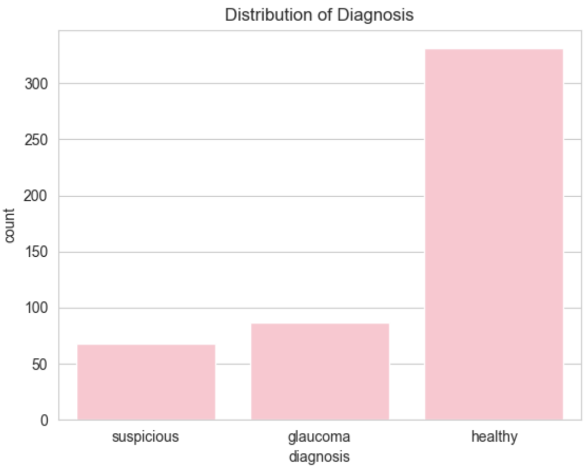


Figure 5- Distribution of diagnoses

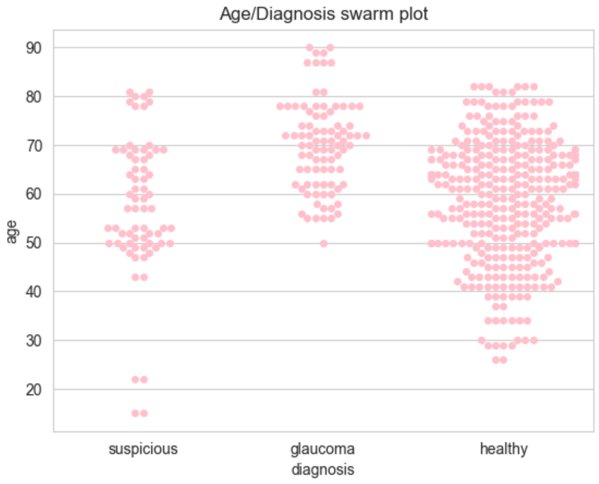


Figure 6- Age/ Diagnosis swarm plot

Our first visualisation involved plotting each column's histograms and swarm plots against diagnosis. The swarm plot of age (Figure 6) indicated that the youngest age at which someone received a glaucoma diagnosis was 55, suggesting that age significantly affected the likelihood of a glaucoma diagnosis. Gender didn’t appear to have a substantial impact on diagnosis.

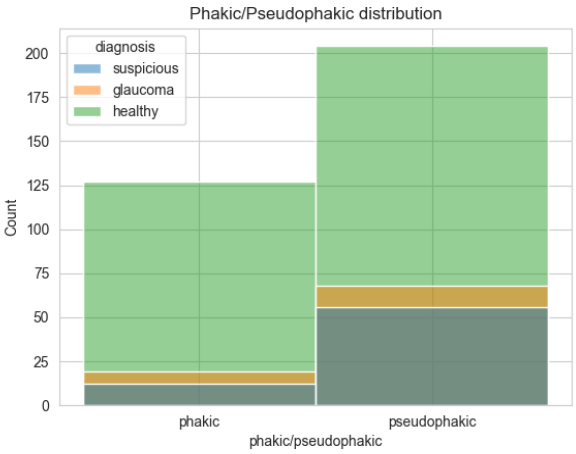


Figure 7- Distribution of Phakic/Pseudophakic, showing diagnosis proportions.

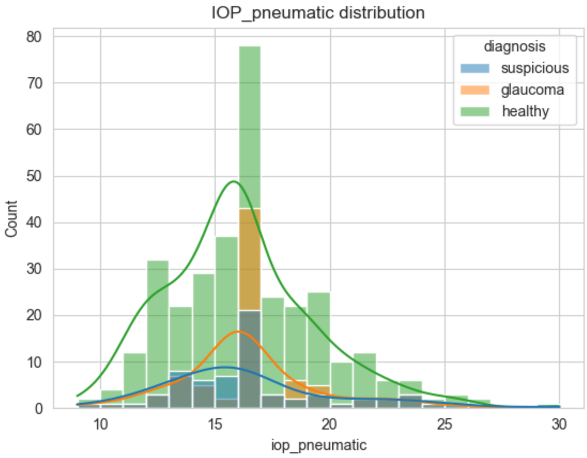


Figure 8- Distribution of IOP_pneumatic, showing diagnosis proportions.

Figure 7 revealed a slight disparity between phakic and pseudophakic patients, with approximately 30% of pseudophakic patients receiving a glaucoma diagnosis compared to roughly 15% of phakic patients.

The value of 16 in Figure 8 appeared to have a high proportion of glaucoma diagnoses, likely since we replaced missing values with 16. This observation suggested the missing values may contain a large proportion of glaucoma patients. The 'iop_perkins' graph revealed a high proportion of glaucoma patients, suggesting that this was only measured if glaucoma was expected. The mean defect had a substantial impact on diagnosis, with all values lower than -6 indicating a glaucoma diagnosis (Figure 9).

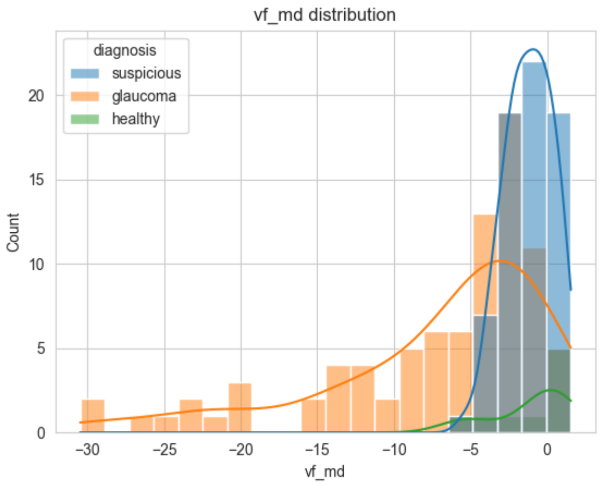


Figure 9- Mean Defect distribution, showing diagnosis proportions.

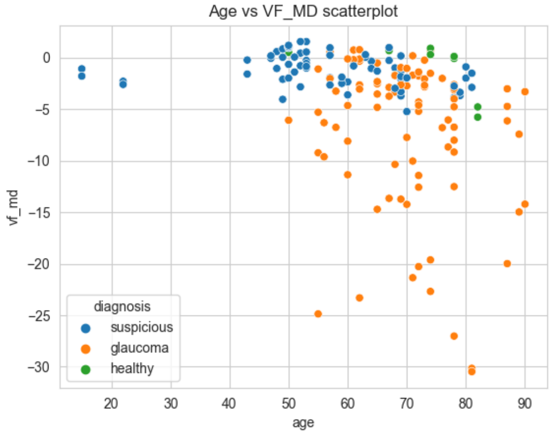
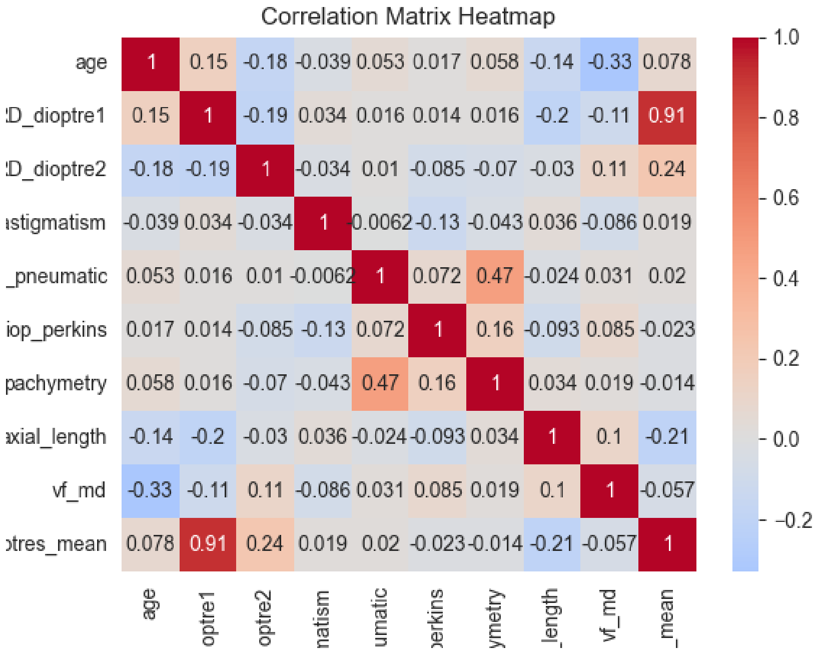


Figure 11- Ages/ Mean Defect scatterplot, with different colours representing different diagnoses.

The mean defect is calculated using other values [4] although the calculation process is unknown to us. We explored factors that could potentially affect this value. We plotted scatterplots of 'vf_md' against other variables and a heatmap (Figure 10). The scatterplot against age (Figure 11) confirmed that older ages and lower mean defect values were associated with a higher likelihood of glaucoma diagnosis. It didn't appear that other values affected the mean defect.



	age	optre1	optre2	natism	umatic	erkins	ymetry	_length	vf_md	_mean
age	1	0.15	-0.18	-0.039	0.053	0.017	0.058	-0.14	-0.33	0.078
D_dioptre1	0.15	1	-0.19	0.034	0.016	0.014	0.016	-0.2	-0.11	0.91
D_dioptre2	-0.18	-0.19	1	-0.034	0.01	-0.085	-0.07	-0.03	0.11	0.24
istigmatism	-0.039	0.034	-0.034	1	-0.0062	-0.13	-0.043	0.036	-0.086	0.019
_pneumatic	0.053	0.016	0.01	-0.0062	1	0.072	0.47	-0.024	0.031	0.02
iop_perkins	0.017	0.014	-0.085	-0.13	0.072	1	0.16	-0.093	0.085	-0.023
pachymetry	0.058	0.016	-0.07	-0.043	0.47	0.16	1	0.034	0.019	-0.014
ixial_length	-0.14	-0.2	-0.03	0.036	-0.024	-0.093	0.034	1	0.1	-0.21
vf_md	-0.33	-0.11	0.11	-0.086	0.031	0.085	0.019	0.1	1	-0.057
otres_mean	0.078	0.91	0.24	0.019	0.02	-0.023	-0.014	-0.21	-0.057	1

Figure 12- Heatmap showing correlation between tables.

In summary, our analysis suggests that age and mean defect are strongly associated with diagnoses. In addition, factors such as pseudophakic/phakic status and IOP values appear to have some impact on diagnosis as well. However, since over half of the values were missing for mean defect and IOP_perkins, it is challenging to make any conclusive statements about their relationships with diagnosis.

Word Count (Not including references and title)- 798.

REFERENCES

[1] Althomali, T.A. (2018) *Relative proportion of different types of refractive errors in subjects seeking laser vision correction, The open ophthalmology journal. U.S. National Library of Medicine.* Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5958297/ (Accessed: April 23, 2023).

[2] Center for Devices and Radiological Health (no date) *What are phakic lenses? U.S. Food and Drug Administration*. FDA. Available at: https://www.fda.gov/medical-devices/phakic-intraocular-lenses/what-are-phakic-lenses (Accessed: April 23, 2023).

[3] de Jong, P.T.V.M. (2021) *The diopter, Eye (London, England)*. U.S. National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8225652/ (Accessed: April 23, 2023).

[4] Doyle, C.K. et al. (2011) The repeatability of mean deviation with size III and size V standard automated perimetry, Investigative Ophthalmology & Visual Science. The Association for Research in Vision and Ophthalmology. Available at: https://iovs.arvojournals.org/article.aspx?articleid=2361579#:~:text=The%20mean%20deviation%20(MD)%20of,visual%20field%20change%20over%20time (Accessed: April 23, 2023).