COM725-Data Analytics and Visualization Software product and technical report (AE2)

Student ID	Student Name
Q16155807	Gunalan Om Prakash Narayan

Table of Contents

Figure table:

Introduction

This project is about performing EDA on children eyesight data, this data was gathered by a group of researchers from Fudan University in Jinshan Hospital of Fudan University in Shanghai. The scope of this project is to understand if the gathered data can show trends and insights on eyesight data that can help to take steps in preventing the variance in eye power.

Aims/Objectives

To gather data, preprocess the data and perform exploratory data analysis on the cleaned data.

Literature Review

The study's performed by a group of clinical researchers on Chinese eyesight data [2] of children aged 6 to 13 to check if cycloplegia in children develops into pseudo myopia in later stages, they concluded that pseudo myopia is more prevalent in younger, more hyperopic children. Pseudo myopia is not a sole risk factor for myopic progression. In addition to this study, we have another study [3] discover that the number of years of school education, rather than age, was linked to the development of myopia in young individuals. Our findings indicate that factors such as exams, academic pressures, and the use of digital screens may contribute to this. Like East Asia and Southeast Asia, North America and Europe are also witnessing an increasing prevalence of myopia. Considering the significant impact of this condition on public health and the potential vision-threatening complications, further research is needed to better understand the factors that mediate the relationship between school education and myopia. The study [4] made by a group of researchers in a preschool which is in Wuxi City, China found that prevalence of myopia was 1.5%, hyperopia was 37.6% and astigmatism was 36.0% this data is from 889 people. This study concluded that significant prevalence of astigmatism among preschool children in Wuxi City. It also established a strong correlation between astigmatism and various risk factors including premature birth, parental history of astigmatism, maternal age at childbirth, feeding patterns, and the amount of time spent exposed to electronic screens.

Methods

Dataset

This data set is chosen as there are many parameters that can help us to identify the early eyesight problems in children. The factors include gender, weight, height, environmental factors as well. We need to perform many analysts as well as visualisations to obtain the correlation between the parameters.

The various parameters in this project are gender, birthday, examination date, height, weight, spherical refraction of right eye, spherical refraction of left eye, cylinder refraction of right eye, cylinder refraction of left eye, axis of right eye, axis of left eye, AL of right eye and AL of left eye.

```
headings=list(df.columns)
display(headings)
```

```
 'birthday',
 'examination date',
 'height',
 'weight',
 'spherical refraction of right eye',
 'spherical refraction of left eye',
 'cylinder refraction of right eye',
 'cylinder refraction of left eye',
 'axis of right eye',
 'axis of left eye',
 'AL of right eye',
 'AL of left eye']
```

Figure 1 Headings

These parameters are used to determine the focus of the eye, and there are also parameters that are used to analyze the conditions of their environment.

The shape and data type of the parameters is as follows.

```
Shape of the dataset:  (2024, 13)
Data types of each column:
gender                              object
birthday                             int64
examination date                     int64
height                             float64
weight                             float64
spherical refraction of right eye   float64
spherical refraction of left eye    float64
cylinder refraction of right eye     object
cylinder refraction of left eye      object
axis of right eye                    object
axis of left eye                     object
AL of right eye                      object
AL of left eye                       object
dtype: object
```

Figure 2 Data type

Analysis and Results

Dataset Preparation

In our data set, we had many missing values. Hence, we removed the rows with missing values as it did not have any significant impact on the results that were obtained.

```
Missing values:
gender                               0
birthday                             0
examination date                     0
height                               1
weight                               1
spherical refraction of right eye    8
spherical refraction of left eye     7
cylinder refraction of right eye     7
cylinder refraction of left eye      8
axis of right eye                   48
axis of left eye                    60
AL of right eye                      1
AL of left eye                       1
dtype: int64
```

Figure 3 Null values

And also, we had some insignificant symbols that were also included in the data hence we also replaced the symbols with null values and then we remove the null values using the drop null function.

```
# Check for missing values
print("Missing values before data cleaning: ")
print(df.isnull().sum())

# Drop rows with missing values
df.dropna(inplace=True)

# # Drop rows with blank values
# df.dropna(subset=['gender', 'birthday', 'examination date'], how='all',
inplace=True)

# Drop rows with "/" as the value
df.replace("/", float('nan'), inplace=True)
df.dropna(inplace=True)
```

```
Missing values before data cleaning:
gender                              0
birthday                            0
examination date                    0
height                              1
weight                              1
spherical refraction of right eye   8
spherical refraction of left eye    7
cylinder refraction of right eye    7
cylinder refraction of left eye     8
axis of right eye                  48
axis of left eye                   60
AL of right eye                     1
AL of left eye                      1
dtype: int64
Missing values after data cleaning:
gender                             0
birthday                           0
examination date                   0
height                             0
weight                             0
spherical refraction of right eye  0
spherical refraction of left eye   0
cylinder refraction of right eye   0
cylinder refraction of left eye    0
axis of right eye                  1
axis of left eye                   0
AL of right eye                    0
AL of left eye                     0
dtype: int64
```

Figure 4 Null values before and after

In the gender column of a data set, we had to replace the string values with the numerical values, as some visualisation techniques do not. Accept the string-based values. We both from this operation. After these operations were performed, data set is clean, and it can be used to. We can use our dataset to visualise, predict and form clusters using it. The data is exported to xlsx and csv format for further usage as its easy to use xlsx in Tablue and CSV in WEKA.

```
# Replacing string with numerical values
df.replace("male",1,inplace=True)
df.replace("female",0, inplace=True)

# Check the cleaned dataset
print("Cleaned dataset: ")
print(df.head())
df.to_excel("cleaned_data.xlsx")
df.to_csv("cleaned_data.xlsx")
display(df)
```

	gender	birthday	examination date	height	weight	spherical refraction of right eye
0	0	1970-01-01 00:00:00.000002006	1970-01-01 00:00:00.000002014	123.0	28.8	-0.50
1	0	1970-01-01 00:00:00.000002006	1970-01-01 00:00:00.000002014	133.0	37.9	0.12
2	1	1970-01-01 00:00:00.000002006	1970-01-01 00:00:00.000002014	119.5	21.4	-0.25
3	1	1970-01-01 00:00:00.000002006	1970-01-01 00:00:00.000002014	122.5	23.5	0.25
4	0	1970-01-01 00:00:00.000002006	1970-01-01 00:00:00.000002014	128.0	27.7	-0.12
...	...	...	...	...	...	...
2019	0	1970-01-01 00:00:00.000002008	1970-01-01 00:00:00.000002017	124.0	27.5	0.00
2020	1	1970-01-01 00:00:00.000002008	1970-01-01 00:00:00.000002017	132.5	27.6	0.50
2021	1	1970-01-01 00:00:00.000002008	1970-01-01 00:00:00.000002017	129.5	25.2	1.25

Exploratory Data Analysis

The Exploratory Data Analysis is initially checking the skewness of the data to understand the limits and patterns.

```
df.hist(figsize=(18,18))
```

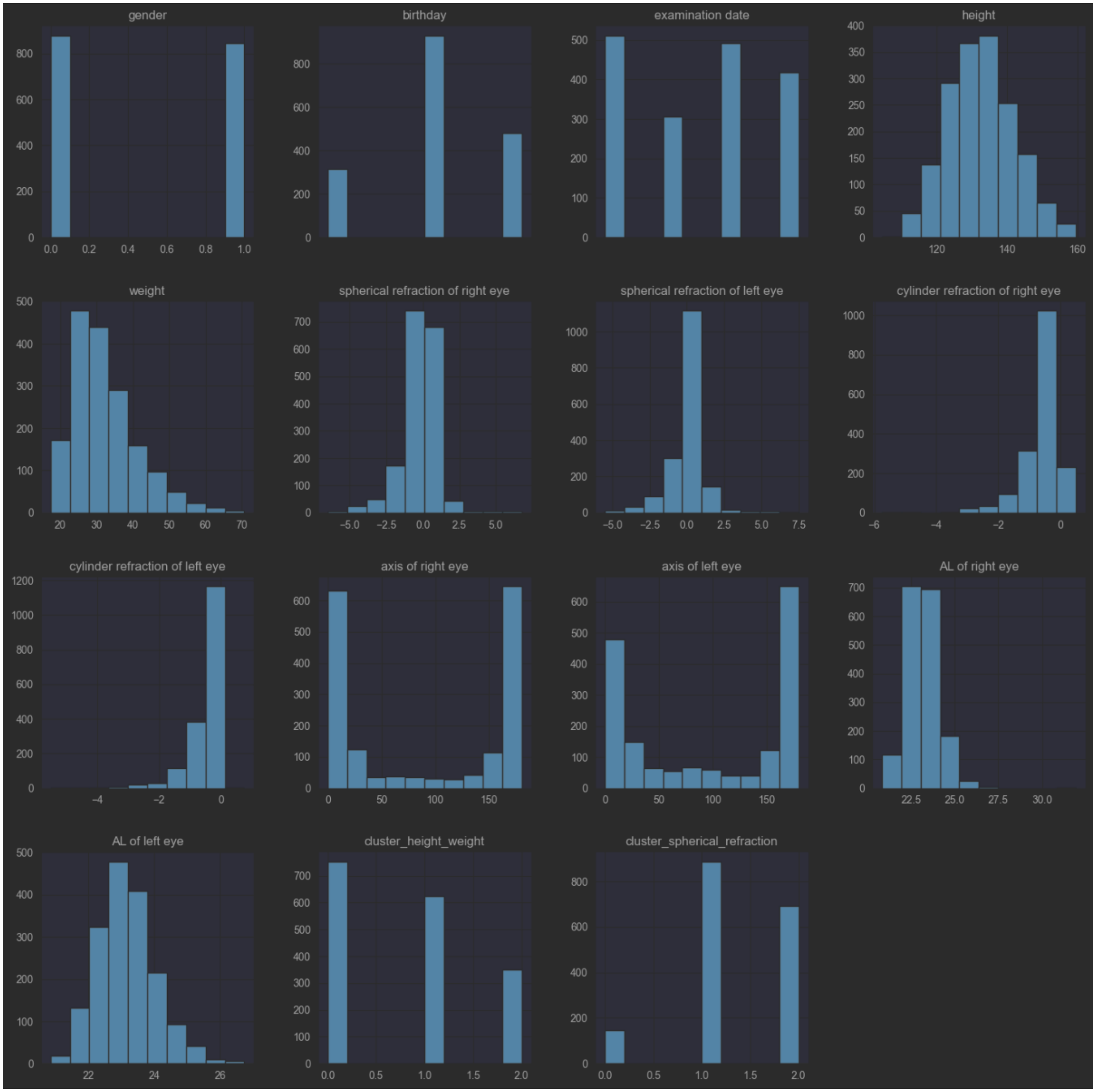


Figure 5 Skewness check

Using the histograms, we came to know that examination date parameter is random based on availability. We can see the height is a bit left skewed that is also related to birthdate and gender. The weight of the children is normal to their age we don't have significant data to say any of them are obese as this is because the data is right skewed. The spherical refraction field is also a little bit left skewed, and most people don't have problems in this area. The cylindrical refraction is also left skewed but here majority of the people have a bit of refraction in there right eye and the left eye is said to be mostly normal for the majority. The Axis of both eyes is distributed in an U-Based distribution. Axial Length of the eyes are both right skewed and they are mostly around 23mm.

```
# Create a scatter plot of weight vs. height with gender as hue
sns.scatterplot(x='weight', y='height', hue='gender', data=df)
plt.title('Weight vs. Height')
plt.xlabel('Weight (kg)')
plt.ylabel('Height (cm)')
plt.show()

# Create a box plot of spherical refraction of right eye by gender
sns.boxplot(x='gender', y='spherical refraction of right eye', data=df)
plt.title('Spherical Refraction (Right Eye) by Gender')
plt.xlabel('Gender')
plt.ylabel('Spherical Refraction (D)')
plt.show()

# Create a box plot of spherical refraction of left eye by gender
sns.boxplot(x='gender', y='spherical refraction of left eye', data=df)
plt.title('Spherical Refraction (Left Eye) by Gender')
plt.xlabel('Gender')
plt.ylabel('Spherical Refraction (D)')
plt.show()

# Create a line chart of age vs. AL of right eye with visual impairment
sns.lineplot(x='age', y='AL of right eye',hue="gender" ,data=df)
plt.title('Age vs. Axial Length (Right Eye) with Visual Impairment')
plt.xlabel('Age (years)')
plt.ylabel('Axial Length (mm)')
plt.show()

# Create a line chart of age vs. AL of left eye with visual impairment
sns.lineplot(x='age', y='AL of left eye',hue="gender" ,data=df)
plt.title('Age vs. Axial Length (Left Eye) with Visual Impairment')
plt.xlabel('Age (years)')
plt.ylabel('Axial Length (mm)')
plt.show()
```

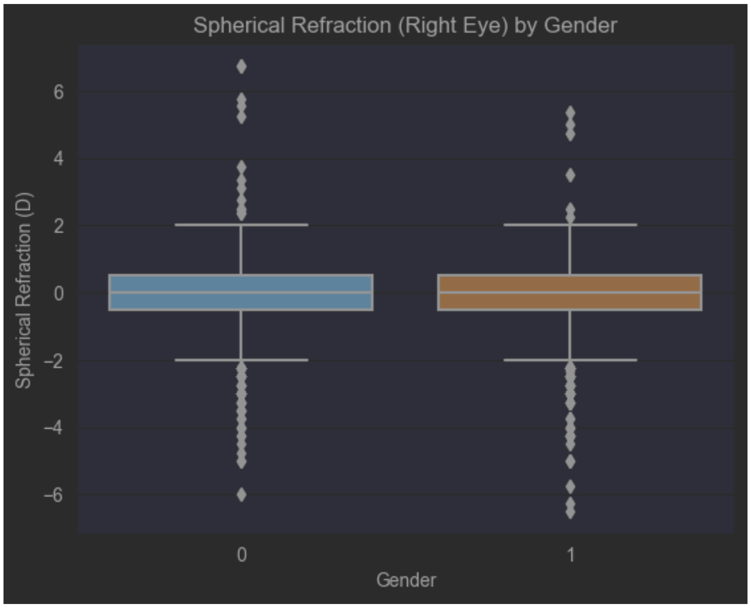


Figure 6 Box plot Spherical refraction of right eye by gender

Figure 7 Box plot Spherical refraction of left eye by gender

We can see that we have range of values around +2 and -2 for spherical refraction of both eyes, and there are also some outliers present in our data.

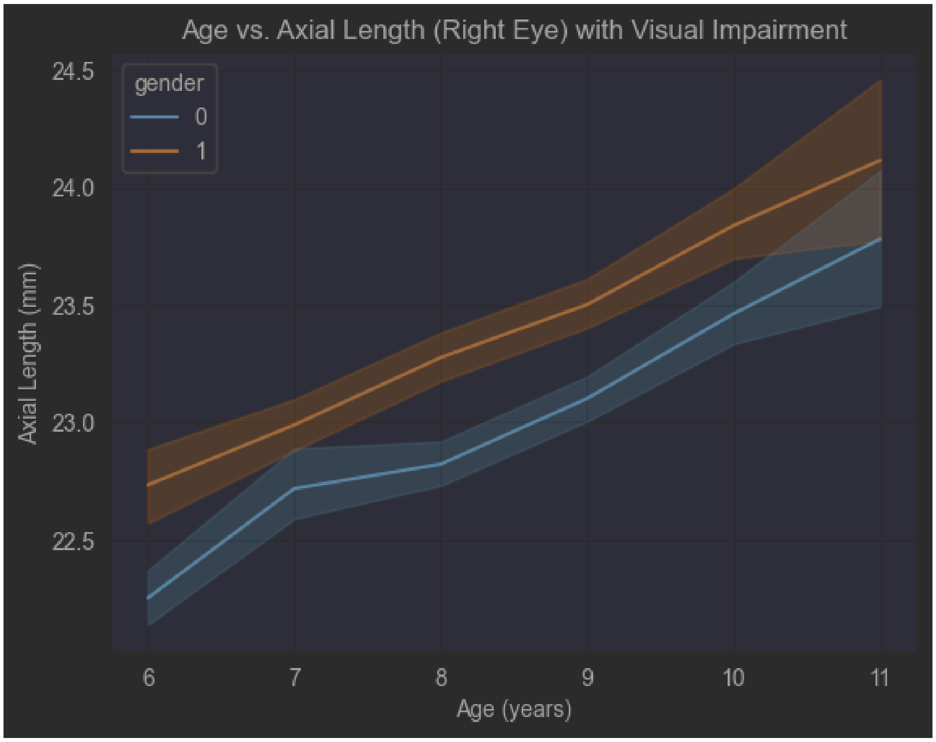


Figure 8 Line plot of Age Vs AL of Right eye with gender

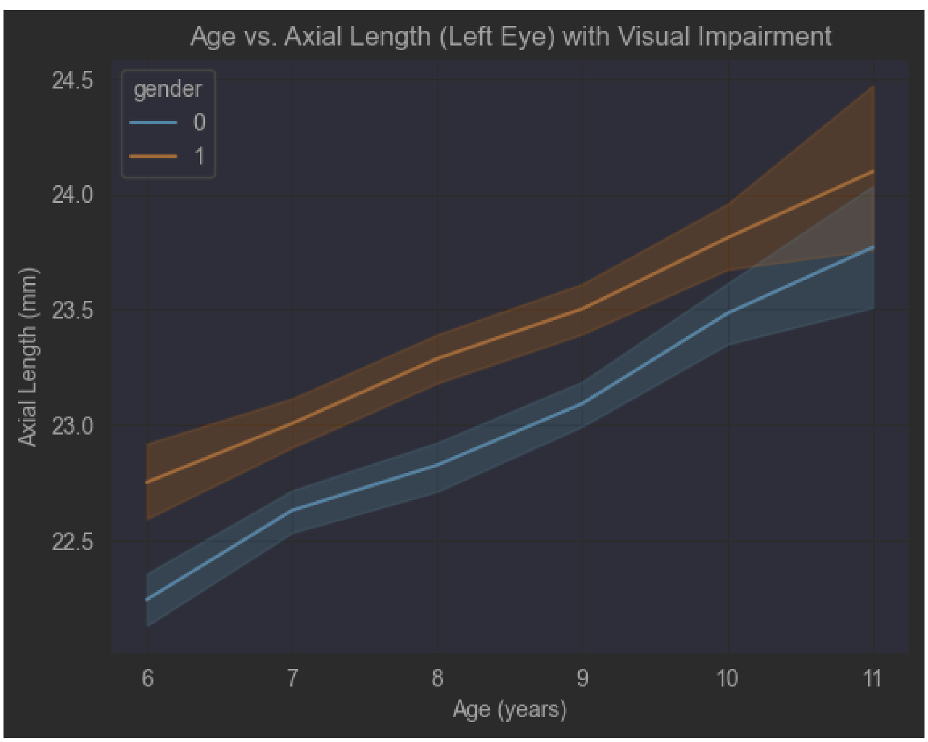


Figure 9 Line plot of Age Vs AL of Left eye with gender

Using a line plot by comparing the age and axial length with gender we can see that we have a clear difference in axial length between male and female children.

```
male_data = df[df['gender'] == 1]
female_data = df[df['gender'] == 0]

# Perform t-test for height between male and female
t_stat, p_value = ttest_ind(male_data['height'], female_data['height'])
print('T-statistic for height between male and female:', t_stat)
print('P-value for height between male and female:', p_value)

# Perform t-test for weight between male and female
t_stat, p_value = ttest_ind(male_data['weight'], female_data['weight'])
print('T-statistic for weight between male and female:', t_stat)
print('P-value for weight between male and female:', p_value)

# Perform t-test for AL of right eye between male and female
t_stat, p_value = ttest_ind((male_data['AL of right eye']),female_data['AL of right
eye'])
print('T-statistic for AL of right eye between male and female:', t_stat)
print('P-value for AL of right eye between male and female:', p_value)

# Perform t-test for AL of left eye between male and female
t_stat, p_value = ttest_ind((male_data['AL of left eye']),female_data['AL of left
eye'])
print('T-statistic for AL of left eye between male and female:', t_stat)
print('P-value for AL of left eye between male and female:', p_value)

# Perform t-test for spherical refraction of right eye between male and female
t_stat, p_value = ttest_ind(male_data['spherical refraction of right eye'],
female_data['spherical refraction of right eye'])
print('T-statistic for spherical refraction of right eye between male and female:',
t_stat)
print('P-value for spherical refraction of right eye between male and female:',
p_value)

# Perform t-test for spherical refraction of left eye between male and female
t_stat, p_value = ttest_ind(male_data['spherical refraction of left eye'],
female_data['spherical refraction of left eye'])
print('T-statistic for spherical refraction of left eye between male and female:',
t_stat)
print('P-value for spherical refraction of left eye between male and female:',
p_value)
```

```
T-statistic for height between male and female: 2.040556630478639
P-value for height between male and female: 0.041447161618858495
T-statistic for weight between male and female: 7.571681459167011
P-value for weight between male and female: 5.972419749818757e-14
T-statistic for AL of right eye between male and female: 9.087386198494409
P-value for AL of right eye between male and female: 2.7106162364844255e-19
T-statistic for AL of left eye between male and female: 9.822514391699105
P-value for AL of left eye between male and female: 3.4059260097458673e-22
T-statistic for spherical refraction of right eye between male and female: -0.29732204306523824
P-value for spherical refraction of right eye between male and female: 0.7662565161012492
T-statistic for spherical refraction of left eye between male and female: 0.17972644857133308
P-value for spherical refraction of left eye between male and female: 0.8573884741310694
```

The P value and T-Statistic values are calculated to understand the occurrence of the data. This helps us to understand the uniquness of the data. The results show that some of the parameters are unique and does not occur randomly.

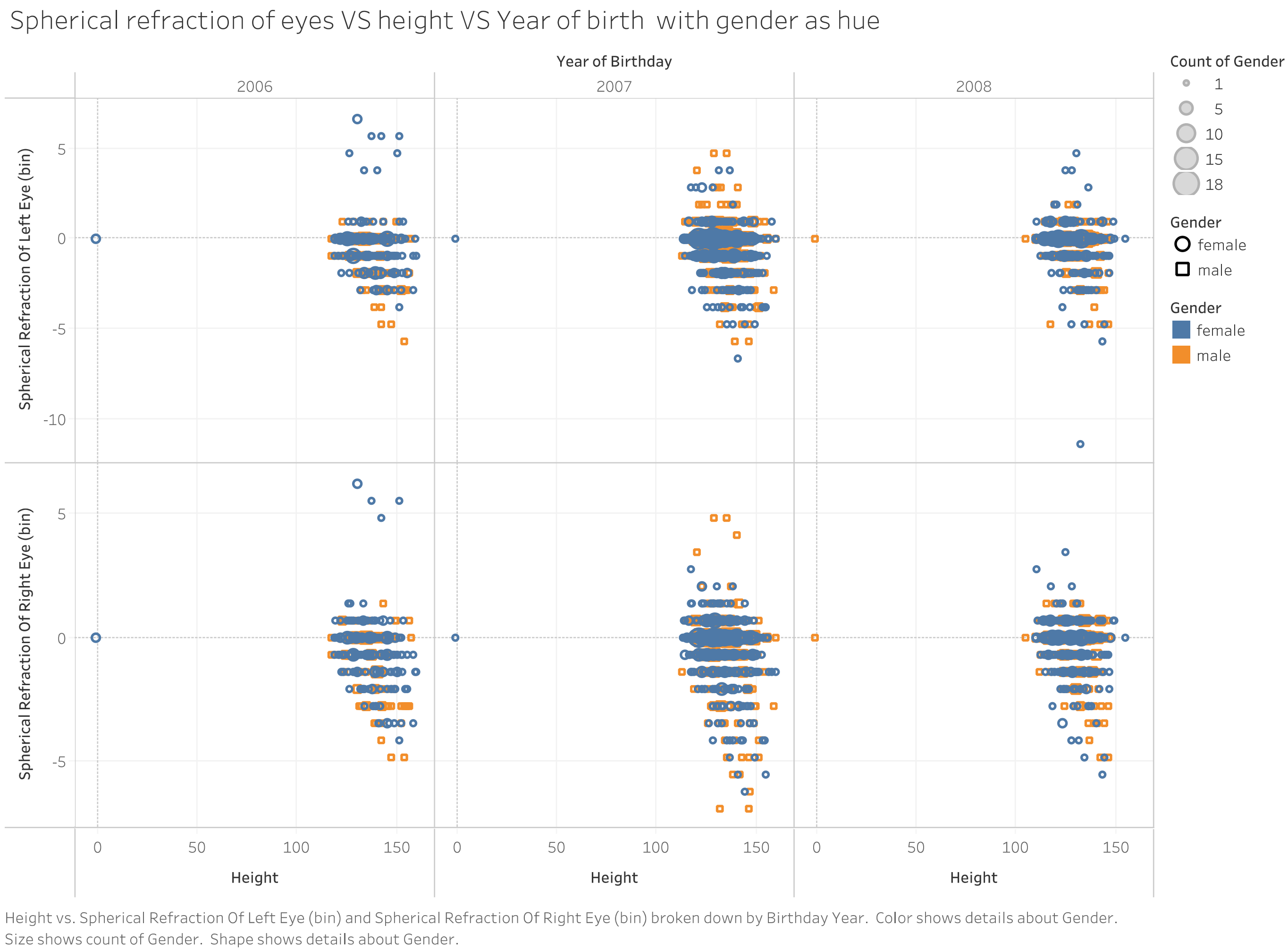
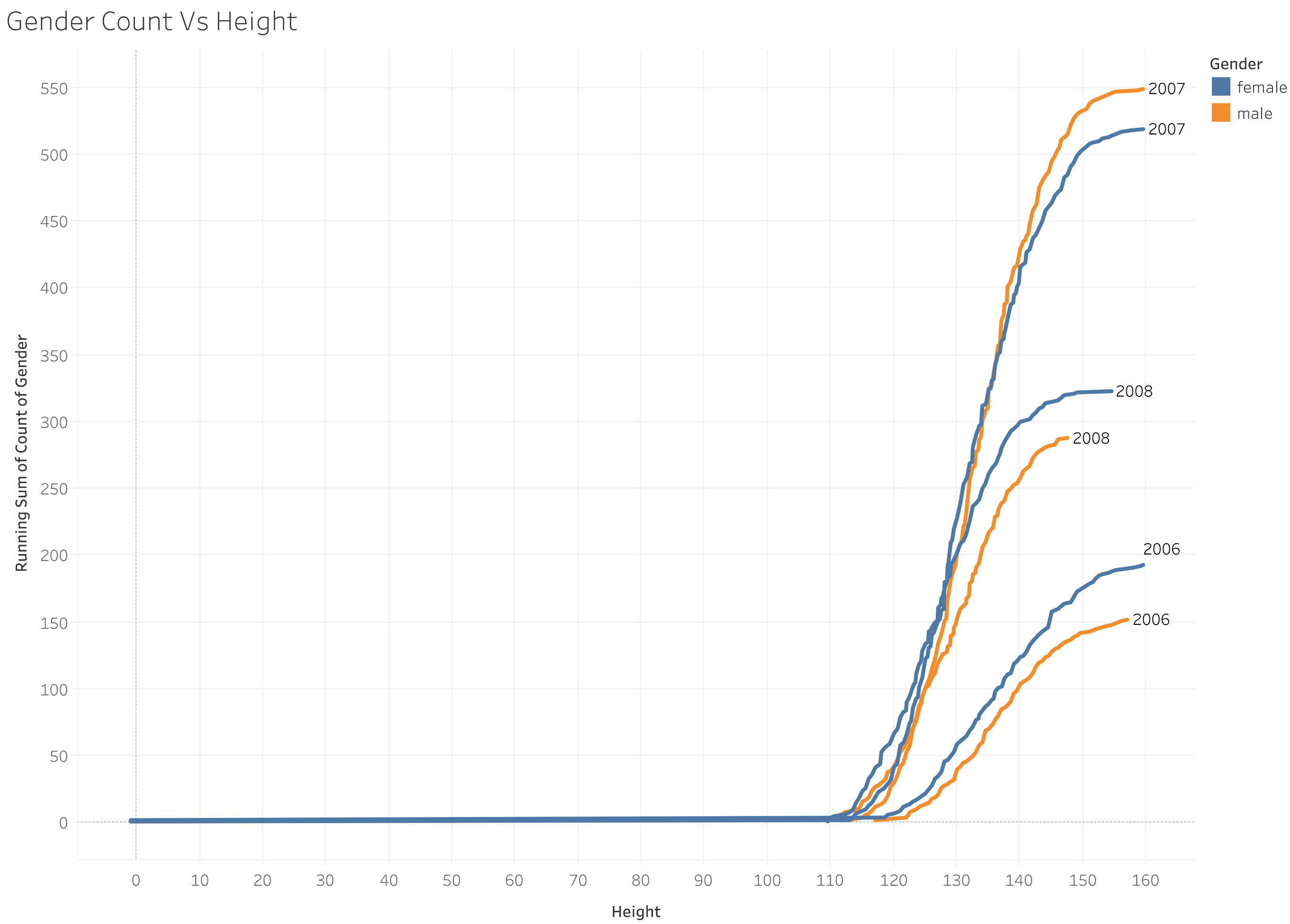


Figure 10 Spherical refraction Vs Height vs Birthday(Year)

From the visualization using tablue we can see that children with height above than the average of their age have some probability to have eyesight variance in negative such as -5 and most children are not affected by any eye variance.



The trend of Running Sum of Count of Gender for Height. Color shows details about Gender. The marks are labeled by Birthday Year. Details are shown for Birthday Year. The data is filtered on count of Gender, which ranges from 1 to 23 and keeps Null values.

Figure 11 Gender count Vs Height

By comparing the gender count based on their height we can see that the height of children born in 2007 is similar in both groups, in the year of 2008 the height of male children's is low compared to female children's this similar trend continues in 2006 year born children's as well.

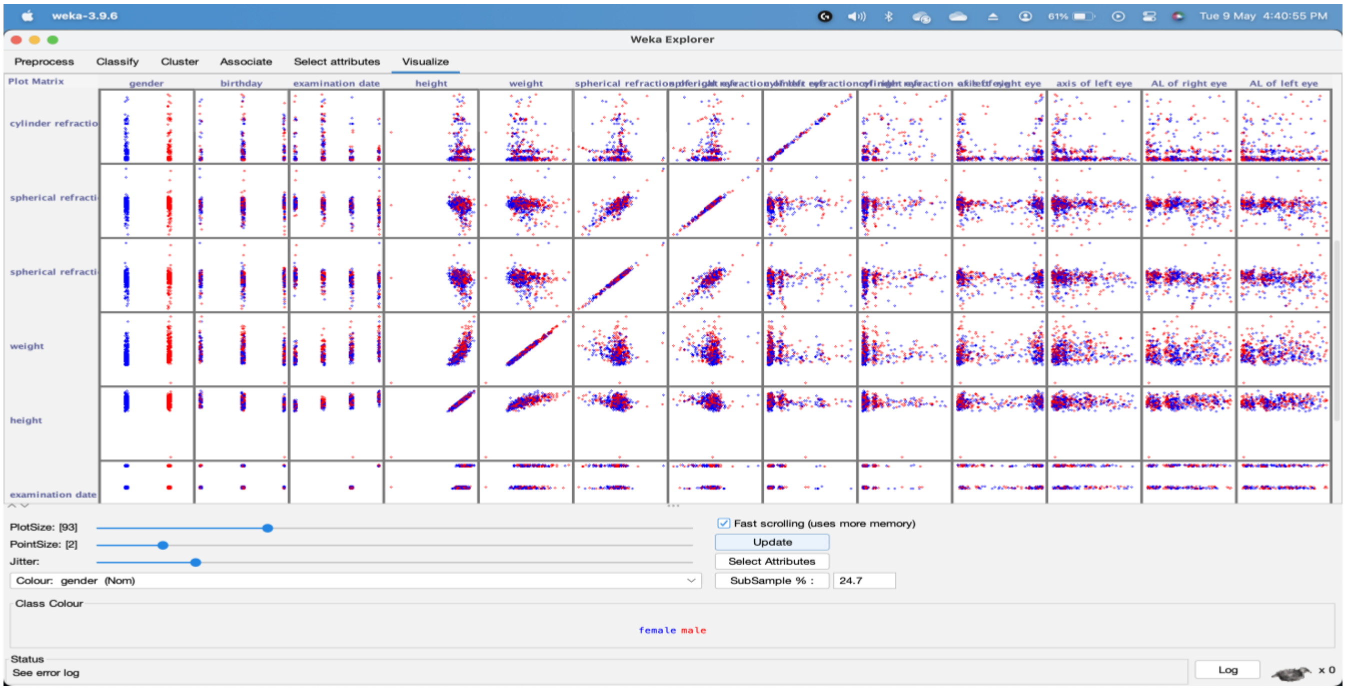


Figure 12 WEKA Visualization 1

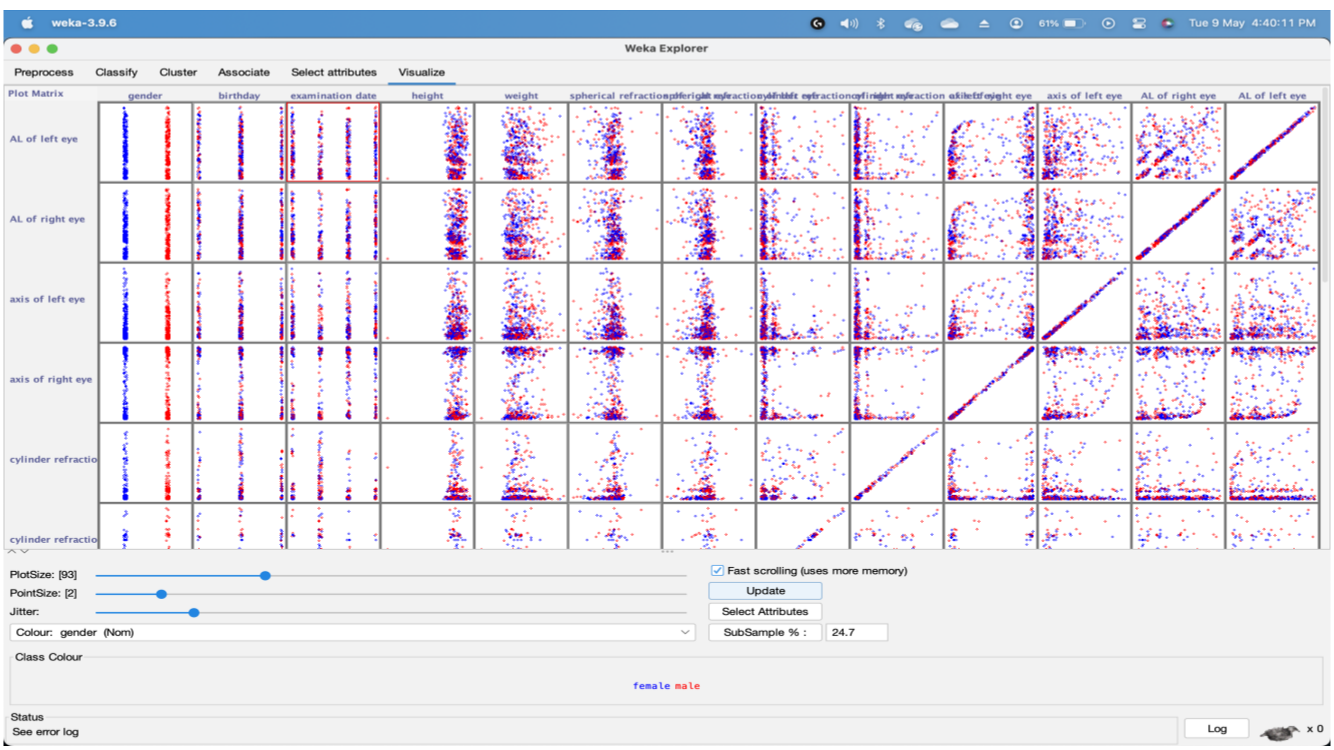


Figure 13 WEKA Visualization 2

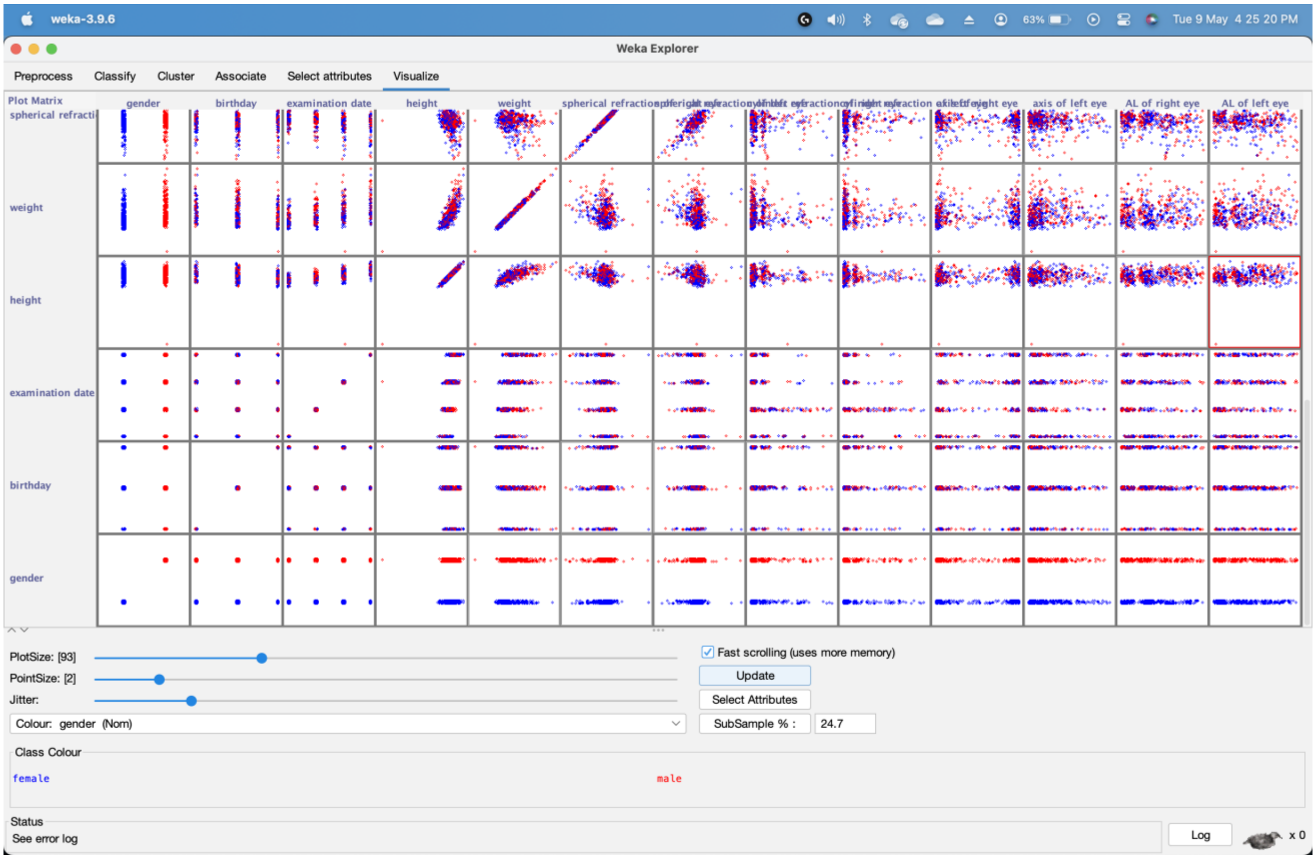


Figure 14 WEKA Visualization 3

Using WEKA we can see all the parameters interlacing as a scatter plot which is helpful to see the scatter plot of each parameter.

Finally, we use python heatmaps to see the correlation of the variables.

```
plt.figure(figsize=(16, 6))
heatmap = sns.heatmap(df.corr(), vmin=-1, vmax=1, annot=True, cmap='BrBG')
heatmap.set_title('Correlation Heatmap', fontdict={'fontsize':18}, pad=12);
```

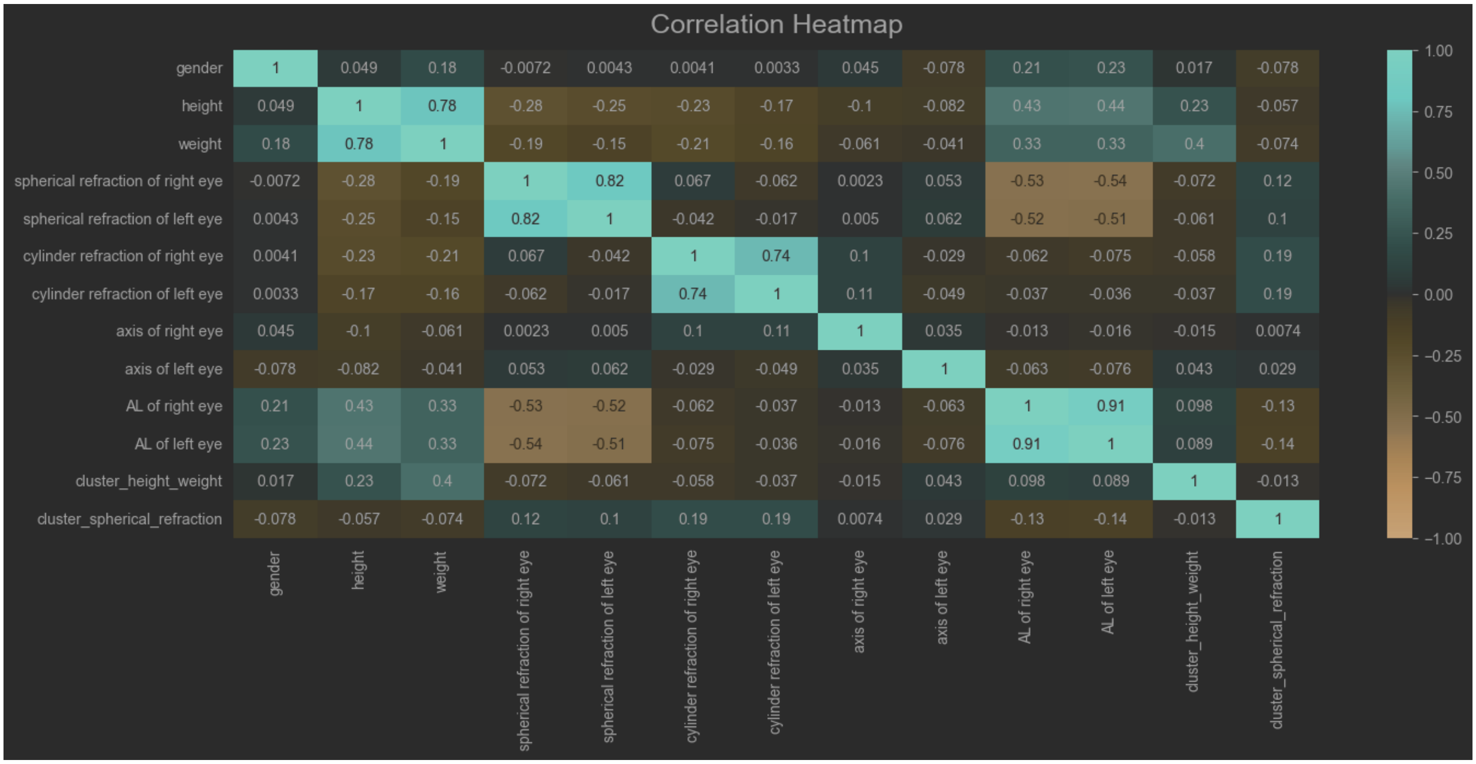


Figure 15 Heatmap for correlation analysis

From the heatmap we conclude that

AL of eyes Vs spherical refraction - Negatively correlated.

height Vs weight – Positively correlated.

spherical refraction left Vs right - Positively correlated.

AL of left Vs right eyes - Positively correlated.

AL of eyes Vs height - Positively correlated.

cylinder refraction left Vs right - Positively correlated.

Data Modelling and Visualization

Using python K-Means clustering the clusters are formed based on Height vs Weight and Spherical refraction of left eye Vs Spherical refraction of right eye.

```
# Perform K-means clustering on height and weight columns
X_height_weight = df[['height', 'weight']]
kmeans_height_weight = KMeans(n_clusters=3, random_state=0).fit(X_height_weight)
df['cluster_height_weight'] = kmeans_height_weight.labels_

# Perform K-means clustering on spherical refraction of right eye and left eye
columns
X_spherical_refraction = df[['spherical refraction of right eye', 'spherical
refraction of left eye']]
kmeans_spherical_refraction = KMeans(n_clusters=3,
random_state=0).fit(X_spherical_refraction)
df['cluster_spherical_refraction'] = kmeans_spherical_refraction.labels_

# Plot the clusters for height and weight
sns.scatterplot(x='height', y='weight', hue='cluster_height_weight', data=df)
plt.xlabel('Height')
plt.ylabel('Weight')
plt.title('K-means Clustering of Height and Weight')
plt.show()

# Plot the clusters for spherical refraction of right eye and left eye
sns.scatterplot(x='spherical refraction of right eye', y='spherical refraction of
left eye', hue='cluster_spherical_refraction', data=df)
plt.xlabel('Spherical Refraction Right Eye')
plt.ylabel('Spherical Refraction Left Eye')
plt.title('K-means Clustering of Spherical Refraction')
plt.show()

# Check the cleaned dataset with clustering labels
print("Cleaned dataset with clustering labels: ")
print(df.head())
```

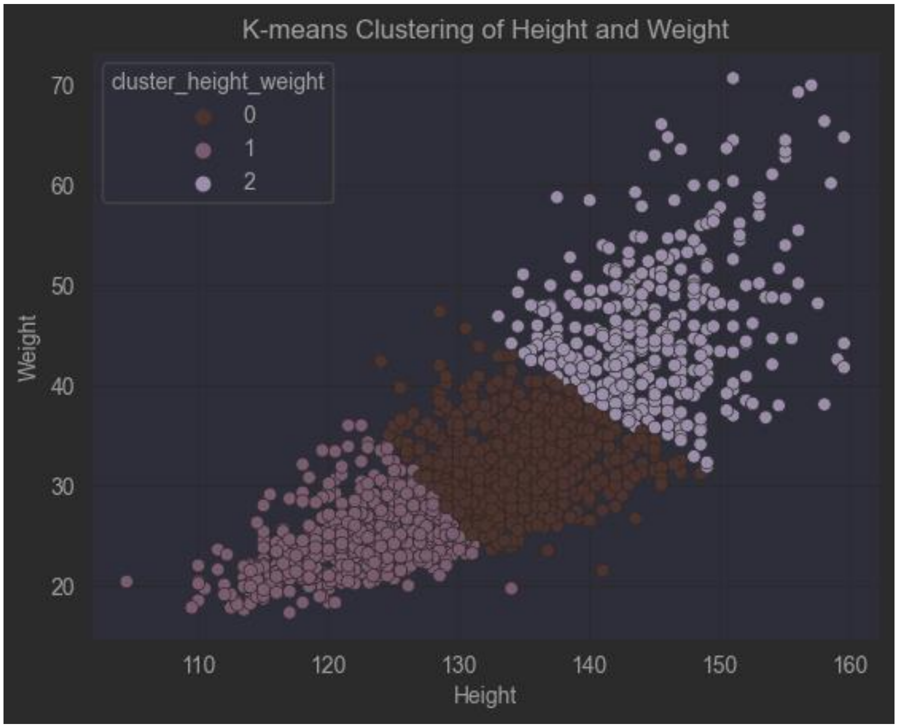


Figure 16 Clustering Height Vs Weight

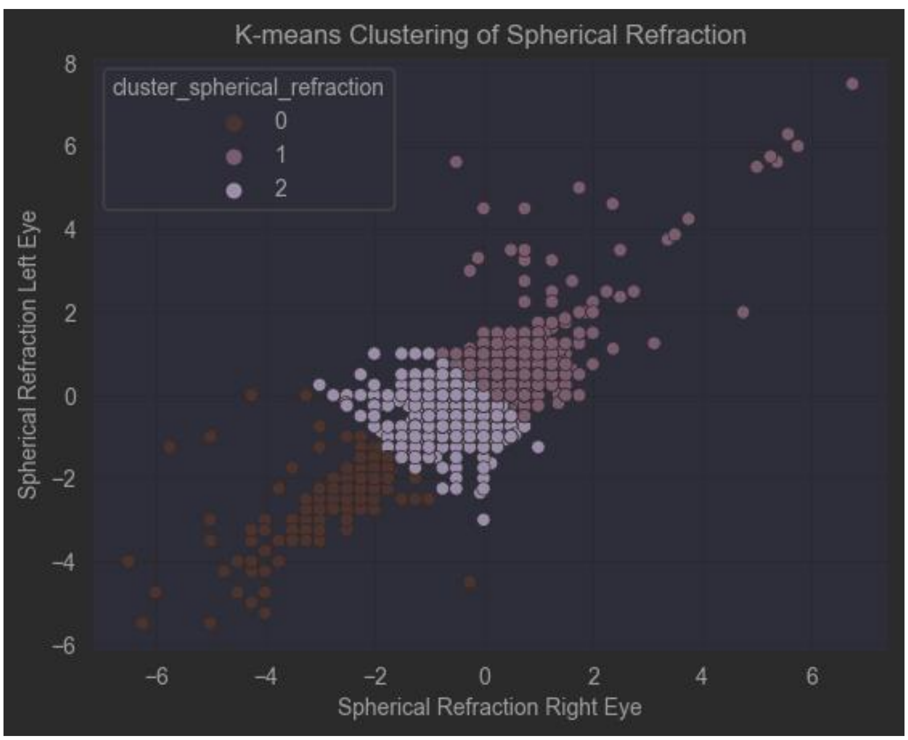


Figure 17 Clustering Spherical refraction

There are 3 clusters formed for both visualizations plots, the first visualization is to form 3 categories of children based on height and weight that helps us to categorize their growth.

0 – Undergrowth

1 – Normal growth

2 – High growth

Then for the second plot its similare to first plot but this is used to categorize the eyesight variation between

0 - Negative variance

1- Normal to little variance

2- Positive variance

To further this we use Tablue to preform regression on the data and plot the trend line. This is basically Linear regression performed based on AL axial length of both eyes to find AL of Right eye for both genders.

Trend line for AL of Right eye (Female)

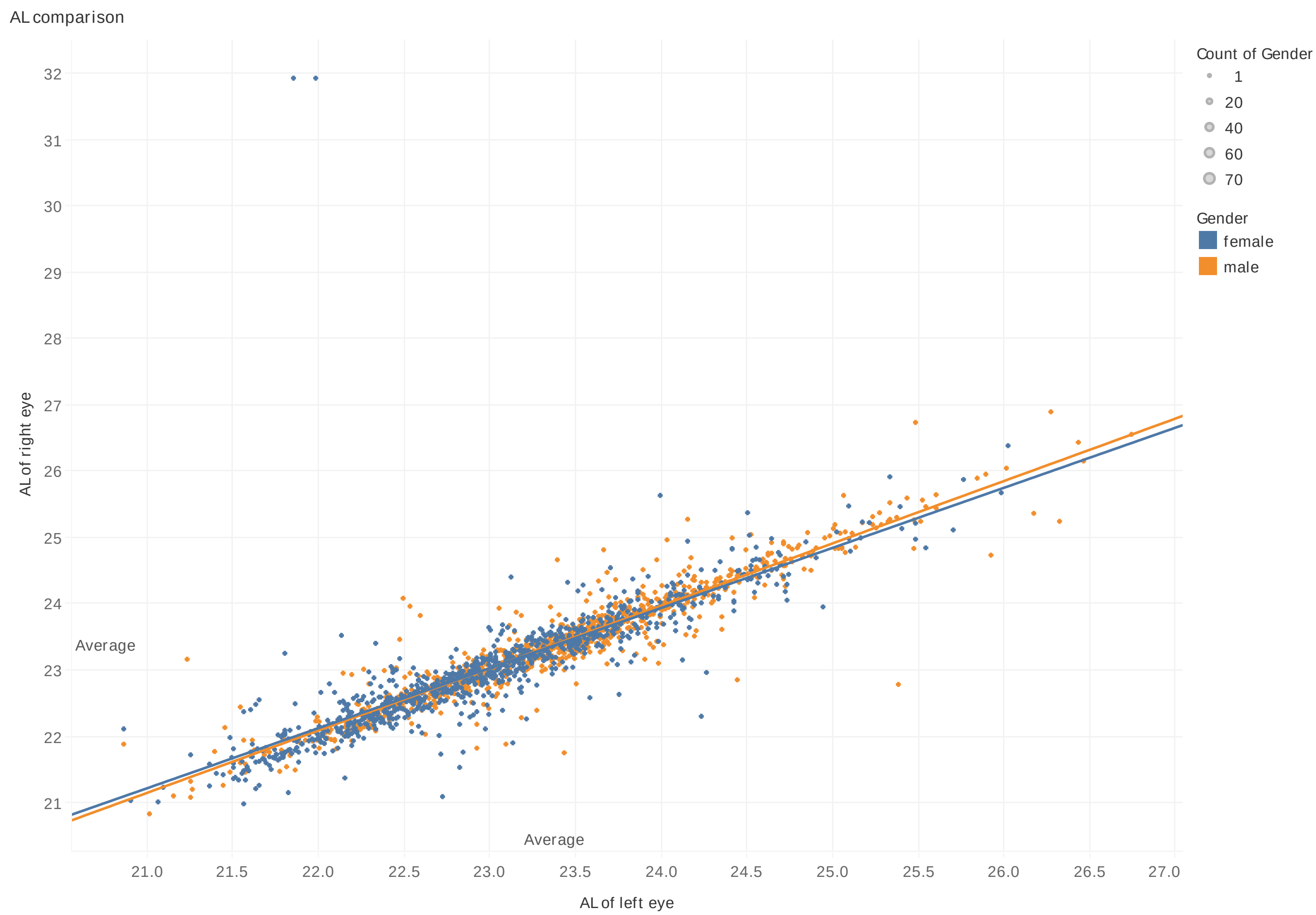
P-value:	< 0.0001
Equation:	AL of right eye = 0.905826*AL of left eye + 2.19348

Coefficients				
Term	**Value**	**StdErr**	**t-value**	**p-value**
AL of left eye	0.905826	0.0214547	42.2203	< 0.0001
intercept	2.19348	0.493536	4.44443	< 0.0001

Trend line for AL of Right eye (Male)

P-value:	< 0.0001
Equation:	AL of right eye = 0.940197*AL of left eye + 1.40205

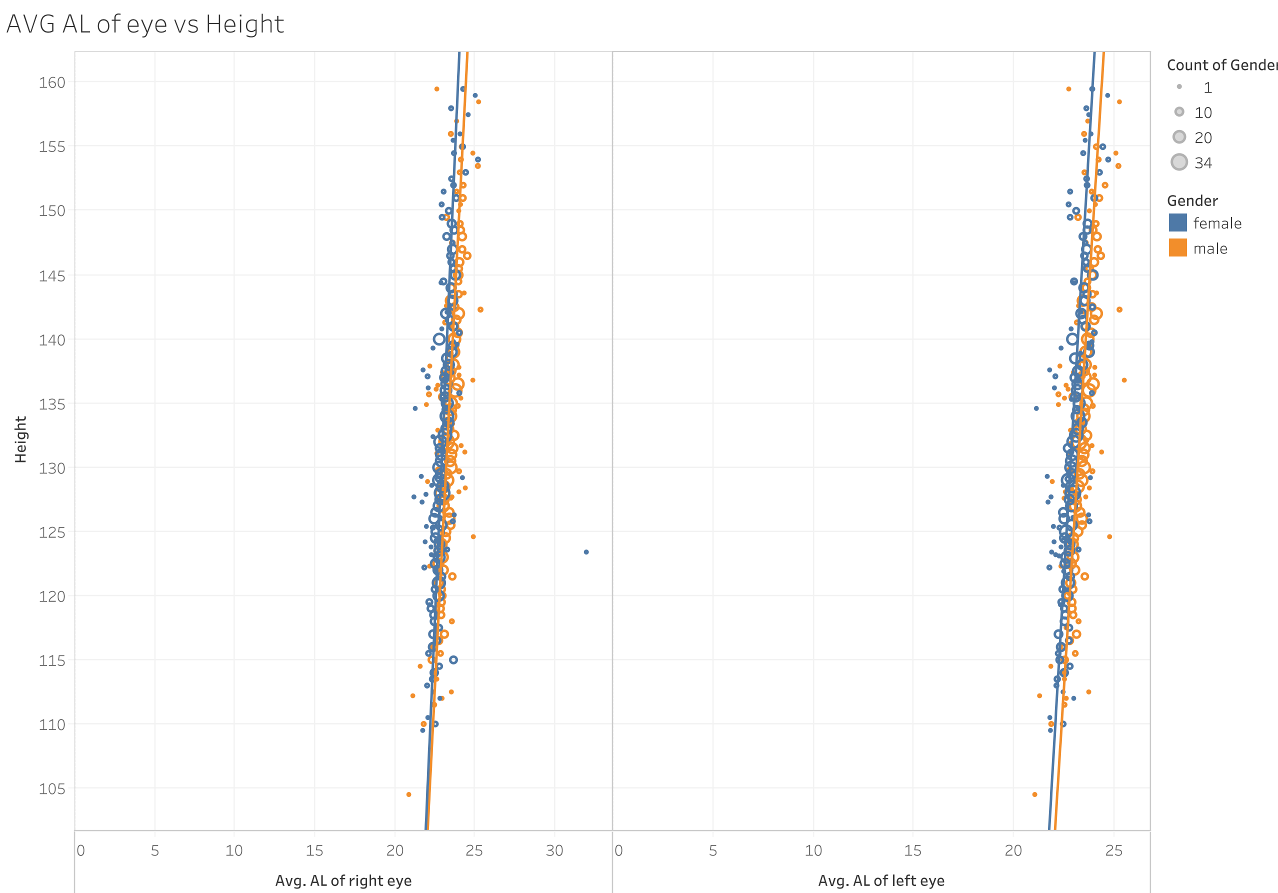
Coefficients				
Term	**Value**	**StdErr**	**t-value**	**p-value**
AL of left eye	0.940197	0.0112344	83.689	< 0.0001
intercept	1.40205	0.262827	5.33448	< 0.0001



AL of left eye vs. AL of right eye. Color shows details about Gender. Size shows count of Gender.

Figure 18 Trend line axial lenght

Then, the same values are deduced with height as an parameter. But here we deduce both eyes AL. Here the trend line shows us that the eye sight variance increases if the height increases.



The plots of average of AL of right eye and average of AL of left eye for Height. Color shows details about Gender. Size shows count of Gender. The data is filtered on Clusters (1), which keeps Cluster 1 and Cluster 2.

Figure 19 Trend line Avg. Axial length VS Height

Trendline for AL of right eye (Female)

P-value:	< 0.0001
Equation:	Avg. AL of right eye = 0.0346213*Height + 18.418

Coefficients				
Term	**Value**	**StdErr**	**t-value**	**p-value**
Height	0.0346213	0.0066306	5.22141	< 0.0001

intercept	18.418	0.885542	20.7986	< 0.0001

Trendline for AL of right eye (Male)

P-value:	< 0.0001
Equation:	Avg. AL of right eye = 0.0410235*Height + 17.8827

Coefficients				
Term	**Value**	**StdErr**	**t-value**	**p-value**
Height	0.0410235	0.0042193	9.72277	< 0.0001
intercept	17.8827	0.565249	31.6368	< 0.0001

Trendline for AL of left eye (Female)

P-value:	< 0.0001
Equation:	Avg. AL of left eye = 0.0373296*Height + 17.9618

Coefficients				
Term	**Value**	**StdErr**	**t-value**	**p-value**
Height	0.0373296	0.0033091	11.2808	< 0.0001
intercept	17.9618	0.441942	40.6429	< 0.0001

Trendline for AL of left eye (Male)

P-value:	< 0.0001
Equation:	Avg. AL of left eye = 0.040068*Height + 17.976

Coefficients				
Term	**Value**	**StdErr**	**t-value**	**p-value**
Height	0.040068	0.0040886	9.79992	< 0.0001
intercept	17.976	0.547736	32.8188	< 0.0001

Finally, we are trying to create a linear trend line to predict cylinder refraction using height as an additional parameter.

Figure 20 Trend line Height Vs Avg. Cylinder Refraction

Trend line for Cylinder refraction of Right eye (Male)

P-value:	< 0.0001
Equation:	Avg. Cylinder Refraction Of Right Eye = -0.0136838*Height + 1.16081

Coefficients				
Term	**Value**	**StdErr**	**t-value**	**p-value**
Height	-0.0136838	0.0027948	-4.8962	< 0.0001
intercept	1.16081	0.373695	3.1063	0.0022796

Trend line for Cylinder refraction of Right eye (Female)

P-value:	< 0.0001
Equation:	Avg. Cylinder Refraction Of Right Eye = -0.0106268*Height + 0.741918

Coefficients				
Term	**Value**	**StdErr**	**t-value**	**p-value**
Height	-0.0106268	0.0026053	-4.07892	< 0.0001
intercept	0.741918	0.346146	2.14337	0.0336961

Trend line for Cylinder refraction of Left eye (Male)

P-value:	0.0003526
Equation:	Avg. Cylinder Refraction Of Left Eye = -0.00939886*Height + 0.692949

Coefficients				
Term	**Value**	**StdErr**	**t-value**	**p-value**
Height	-0.0093989	0.0025681	-3.65989	0.0003526
intercept	0.692949	0.343382	2.01801	0.0454345

Trend line for Cylinder refraction of Left eye (Female)

P-value:	0.359619
Equation:	Avg. Cylinder Refraction Of Left Eye = -0.00241231*Height + -0.244567

Coefficients				
Term	**Value**	**StdErr**	**t-value**	**p-value**
Height	-0.0024123	0.0026252	-0.918909	0.359619
intercept	-0.244567	0.348788	-0.701193	0.484269

Evaluation

From the work performed using various methods the results obtained are that height is associated with eyesight power and it keeps on increasing as they age this is proved by the study conducted in India [5]. There is also difference in AL (Axial length) between male and female children which is confirmed by the study conducted in China by researchers [1]. Finally, the Spherical refraction of the eyes are negatively corelated with Axial length this relationship can be found in many studies [6]. These are the some finding that are done in this EDA.

Limitations and Challenges

Data that we had is broad in terms of medical perspective but if we had some lifestyle data predictions can be done more conclusively.

Conclusion

There are many factors that affect eyesight but here we found that height is also a factor. This finding may be genetically connected but further study and data can help us to make a definitive stand in this fact. The regressions preformed can be used to predict the eyesight power but that will not be accurate in a broad perspective data as the environment and age group changes. The results we found helps us to understand the eye sight problem in children we might need to conduct more studies to understand the same in adults.

Reference List (Harvard Style)

1.Li, T., Jiang, B. and Zhou, X. (no date) *Data from: Axial length elongation and myopia incidence increase in primary school-age children: 3-year follow-up study, Dryad*. Dryad. Available at: https://datadryad.org/stash/dataset/doi:10.5061/dryad.d4mb3pn (Accessed: February 27, 2023).

2. Kang, M.-T. *et al.* (2021) *Prevalence and risk factors of pseudomyopia in a Chinese children population: The Anyang Childhood Eye Study, British Journal of Ophthalmology*. BMJ Publishing

Group Ltd. Available at: https://bjo.bmj.com/content/105/9/1216.abstract?casa_token=8erZXPbeyGoAAAAA%3A4LGtyoj_V-TFrLiyycYyCqJxfHEG_kLsFhPs0bWgpf0iMuzN2nup1NK2uCKO4vspJxyJZMfFV3c (Accessed: April 27, 2023).

3. Chunfeng Zhang, B.A. (2022) *Association of School Education with eyesight among children and adolescents, JAMA Network Open*. JAMA Network. Available at: https://jamanetwork.com/journals/jamanetworkopen/article-abstract/2791731 (Accessed: February 27, 2023).

4. Yang, Z. *et al.* (2022) *Prevalence of and factors associated with astigmatism in preschool children in Wuxi City, china - BMC ophthalmology, BioMed Central*. BioMed Central. Available at: https://bmcophthalmol.biomedcentral.com/articles/10.1186/s12886-022-02358-2 (Accessed: February 27, 2023).

5. Nangia V;Jonas JB;Matin A;Kulkarni M;Sinha A;Gupta R; *Body height and ocular dimensions in the adult population in rural Central India. The Central India Eye and Medical Study, Graefe's archive for clinical and experimental ophthalmology = Albrecht von Graefes Archiv fur klinische und experimentelle Ophthalmologie*. Available at: https://pubmed.ncbi.nlm.nih.gov/20652306/ (Accessed: 09 March 2023).

6. Z;, F.Q.H. *Axial length and its relationship to refractive error in Chinese university students, Contact lens & anterior eye : the journal of the British Contact Lens Association*. Available at: https://pubmed.ncbi.nlm.nih.gov/34030907/ (Accessed: 09 April 2023).