**Investigating the impact of medical factors on the occurrence of myopia**

**Contents**

Introduction

Aim and Objective

Hypothesis

Attribute information

Methods used

Analysis and interpretation

Plots

Acknowledgements

***Introduction:***

Myopia is one of the greatest problems in human being especially among children. The data are a subset of data from the *Orinda Longitudinal Study of Myopia (OLSM)*, a cohort study of ocular component development and risk factors for the onset of myopia in children. Data collection began in the 1989–1990 school year and continued annually through the 2000–2001 school year.

***Summary of the project:***

This project aims to analyse the dataset on the occurrence of myopia and develop a predictive model to identify risk factors associated with myopia. The dataset includes various categorical variables like gender, time spent on computer games and study hours along with continuous variables like spherical equivalent refraction, axial length, anterior chamber depth, lens thickness and others. After preprocessing the data and splitting it based on age group and gender, exploratory data analysis provided insights into the distribution of variables. Bar plots and histograms were utilized to visualize categorical and continuous variables, respectively.

Consequently, a logistic regression model was fitted to check the likelihood of various factors on the occurrence of myopia and the significant predictors were identified. These findings provide valuable insights into the factors contributing to occurrence of myopia and can aid in making informed decisions for prevention and early detection.

***Aim and Objective:***

Our aim will be to find in which age group Myopia appears the most, in which Gender Myopia is more likely to appear or if Myopia is dependent upon Gender. Our objective will be to find the root causes that can trigger Myopia among children by analysing the attributes.

***Hypothesis to be checked:***

**Hypothesis 1**

**H0:** Occurrence of myopia and the factor Gender are independent.

**HA:** Occurrence of myopia and the factor Gender are not independent.

**Hypothesis 2**

**H0:** Occurrence of myopia in child and the occurrence of myopia in mother are independent.

**HA:** Occurrence of myopia in child and the occurrence of myopia in mother are not independent.

**Hypothesis 3**

**H0:** Occurrence of myopia in child and the occurrence of myopia in father are independent.

**HA:** Occurrence of myopia in child and the occurrence of myopia in father are not independent.

***Attribute information:***

1. **ID**: - Index Number.
2. **STUDYYEAR**: - The year in which subject entered the study.
3. **MYOPIC**: - Myopia within the first five years of follow up (0=No, 1=Yes).
4. **AGE**: - Age of subject when visited.
5. **GENDER**: - Gender of subject (0=Female, 1=Male)
6. **SPHEQ**: - Spherical Equivalent Refraction (in diopters) i.e. a calculated value that represents the average refractive power of the eye.
7. **AL**: - Axial Length (in mm) of Eyeball.
8. **ACD**: - Anterior Chamber Depth (in mm) i.e. the depth of the fluid-filled space in the front part of the eye, between the Cornea and the Iris.
9. **LT**: - The thickness of Human lens (in mm).
10. **VCD**: - Vitreous Chamber Depth (in mm) i.e. the depth of the gel-like structure that fills the back two-thirds of the eye, between the Lens and the Retina.
11. **SPORTHR**: - Time spent by the subject engaging in sports/outdoor activities (in Hours/Week).
12. **READHR**: - Time spent by the subject for pleasure (in Hours/Week).
13. **COMPHR**: - Time spent by the subject using Computer (in Hours/Week).
14. **STUDYHR**: - Time spent by the subject reading or studying (in Hours/Week).
15. **TVHR**: - Time spent by the subject watching television (in Hours/Week).
16. **DIOPTERHR**: - Composite of near-work activities by the subject (in Hours/Week).
17. **MOMMY**: - Was the subject's mother myopic? (0=No, 1=Yes)
18. **DADMY**: - Was the subject's father myopic? (0=No, 1=Yes).

***Methods used:***

Methods we used for finding insights from dataset are the followed by:

1. **Chi-square test: -** Chi-square test of independence is a statistical method used to determine if there is a significant association between two categorical variables in a contingency table. It is suitable for qualitative variables because it assesses whether the observed frequencies in the contingency table differ significantly from what would be expected if the variables were independent of each other.

Performing a chi square test of independence will help us find any dependencies between categorical variable ‘MYOPIC’ and other categorical variable such as ‘GENDER’. Identifying the dependencies can help us find reasons that can cause myopia among children.

1. **Summary of continuous variables: -** The summary of continuous variables is a fundamental step.

In analysing and understanding quantitative data. It provides information on key descriptive statistics such as measures of central tendency (mean, median) and dispersion (minimum, maximum, quartiles). By finding out the descriptive statistics of each continuous variable we can have a detailed idea about the values that the variable is having. This step is important to us for finding outliers, recognizing distribution and gaining valuable insights underneath.

1. **Generalized Linear Model (GLM): -** GLM is suitable for this dataset as it enables the analysis of binary outcomes such as presence or absence of myopia using both categorical and continuous variables. With this we can find factors that can make Myopia more likely.

***Analysis and interpretation of outputs:***

**Bar graphs and Histograms**

1. The histogram of Spherical equivalent refraction (SPHEQ) is *positively skewed*.
2. The histogram of Axial Length (AL) is *positively skewed*.
3. The histogram of Anterior chamber depth (ACD) is *negatively skewed*.
4. The histogram of Lens Thickness (LT) is *positively skewed*.
5. The histogram of Vitreous Chamber Depth (VCD) is *positively skewed*.
6. The histogram of time spent in sports (SPORTHR) is *positively skewed*.
7. The histogram of time spent for pleasure (READHR) is *positively skewed*.
8. The histogram of time spent for computer games (COMPHR) is *positively skewed*.
9. The histogram of time spent for study (STUDYHR) is *positively skewed*.
10. The histogram of time spent for television viewing (COMPHR) is *positively skewed*.
11. The histogram of the composite of near work activities (DIOPTERHR) is *positively skewed*.

**Chi-square Tests**

We use chi-square test on attributes MYOPIC and GENDER to see if they are dependent or not.

As output, we get *p-value = 0.1582*, which leads us to accept the null hypotheses i.e. there's no association between occurrence of myopia and gender at diagnosis of the subject.

Again, we use chi-square test on attributes MYOPIC and MOMMY to see if they are dependent or not.

As output, we get *p-value = 0.001314*, which leads us to accept the alternative hypotheses i.e. the association between presence of myopia in the mother and occurrence of myopia in the subject is significant.

Again, we use chi-square test on attributes MYOPIC and DADMY to see if they are dependent or not.

As output, we get *p-value = 0.0003096*. Which leads us to accept the alternative hypotheses i.e. the association between presence of myopia in the father and occurrence of myopia in the subject is significant.

**Descriptive Statistics:**

The descriptive statistics of some attributes are given below: -

1. *SPHEQ: -*
   * + Minimum Value: - -0.6990
     + 1st Quartile: - 0.4562
     + Median: - 0.7290
     + Mean: - 0.8010
     + 3rd Quartile: - 1.0340
     + Maximum Value: - 4.3720
2. *AL: -*
   * + Minimum Value: - 19.90
     + 1st Quartile: - 22.04
     + Median: - 22.46
     + Mean: - 22.50
     + 3rd Quartile: - 22.97
     + Maximum Value: - 24.56
3. *ACD: -*
   * + Minimum Value: - 2.772
     + 1st Quartile: - 3.424
     + Median: - 3.585
     + Mean: - 3.579
     + 3rd Quartile: - 3.730
     + Maximum Value: - 4.250
4. *LT: -*
   * + Minimum Value: - 2.960
     + 1st Quartile: - 3.436
     + Median: - 3.542
     + Mean: - 3.541
     + 3rd Quartile: - 3.640
     + Maximum Value: - 4.112
5. *VCD: -*
   * + Minimum Value: - 13.38
     + 1st Quartile: - 14.93
     + Median: - 15.36
     + Mean: - 15.38
     + 3rd Quartile: - 15.84
     + Maximum Value: - 17.30
6. *SPORTHR: -*
   * + Minimum Value: - 0.00
     + 1st Quartile: - 6.00
     + Median: - 10.00
     + Mean: - 11.95
     + 3rd Quartile: - 16.00
     + Maximum Value: - 45.00
7. *READHR: -*
   * + Minimum Value: - 0.000
     + 1st Quartile: - 0.000
     + Median: - 2.000
     + Mean: - 2.796
     + 3rd Quartile: - 4.000
     + Maximum Value: - 20.000
8. *STUDYHR: -*
   * + Minimum Value: - 0.00
     + 1st Quartile: - 0.00
     + Median: - 1.00
     + Mean: - 1.49
     + 3rd Quartile: - 2.00
     + Maximum Value: - 15.00
9. *TVHR: -*
   * + Minimum Value: - 0.000
     + 1st Quartile: - 4.250
     + Median: - 8.000
     + Mean: - 8.948
     + 3rd Quartile: - 12.000
     + Maximum Value: - 31.000
10. *COMPHR: -*
    * + Minimum Value: - 0.000
      + 1st Quartile: - 0.000
      + Median: - 1.000
      + Mean: - 2.105
      + 3rd Quartile: - 3.000
      + Maximum Value: - 30.000
11. *DIOPTERHR: -*
    * + Minimum Value: - 2.00
      + 1st Quartile: - 15.00
      + Median: - 23.00
      + Mean: - 26.02
      + 3rd Quartile: - 34.00

Maximum Value: - 101.00

**Fitting a generalised linear model for the original data:**

* We get a p-value of <2\*10-16, by forming a model between MYOPIC and SPHEQ, which is less than 0.05 i.e. this factor is significant for the fitted model.
* We get a p-value of 0.0312, by forming a model between MYOPIC and SPORTHR, which is less than 0.05 i.e. this factor is significant for the fitted model.
* We get a p-value of 0.0245, by forming a model between MYOPIC and MOMMY, which is less than 0.05 i.e. this factor is significant for the fitted model.
* We get a p-value of 0.0156, by forming a model between MYOPIC and DADMY, which is less than 0.05 i.e. this factor is significant for the fitted model.

**Fitting a generalised linear model for the male observations:**

* We get a p-value of 2.81\*10-10, by forming a model between MYOPIC and SPHEQ, which is less than 0.05 i.e. this factor is significant for the fitted model.
* We get a p-value of 0.0267, by forming a model between MYOPIC and DADMY, which is less than 0.05 i.e. this factor is significant for the fitted model.

**Fitting a generalised linear model for the female observations:**

* We get a p-value of 2.99\*10-9, by forming a model between MYOPIC and SPHEQ, which is less than 0.05 i.e. this factor is significant for the fitted model.
* We get a p-value of 0.0133, by forming a model between MYOPIC and MOMMY, which is less than 0.05 i.e. this factor is significant for the fitted model.

Next, we divide our dataset *age wise* and try to find a linear model between MYOPIC and other attributes on the divided data. For age = 5 and 9 years we have not found suitable observations that can form a linear model with MYOPIC.

**For age 6:**

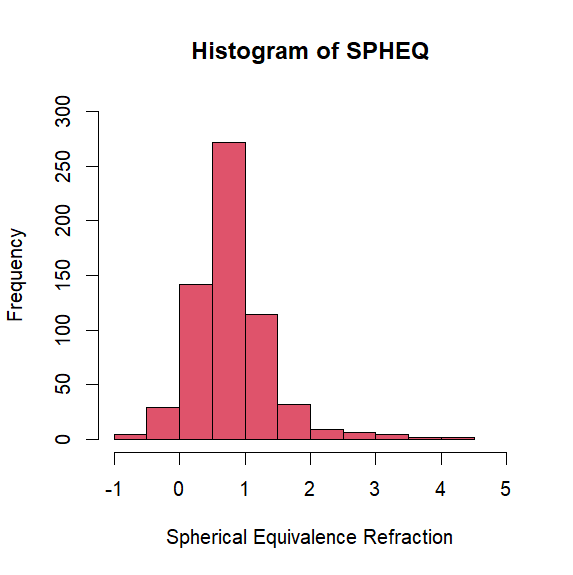
* We get a p-value of 4.44\*10-11, by forming a model between MYOPIC and SPHEQ, which is less than 0.05 i.e. this factor is significant for the fitted model.
* We get a p-value of 0.0235, by forming a model between MYOPIC and SPORTHR, which is less than 0.05 i.e. this factor is significant for the fitted model.
* We get a p-value of 0.0351, by forming a model between MYOPIC and DADMY, which is less than 0.05 i.e. this factor is significant for the fitted model.

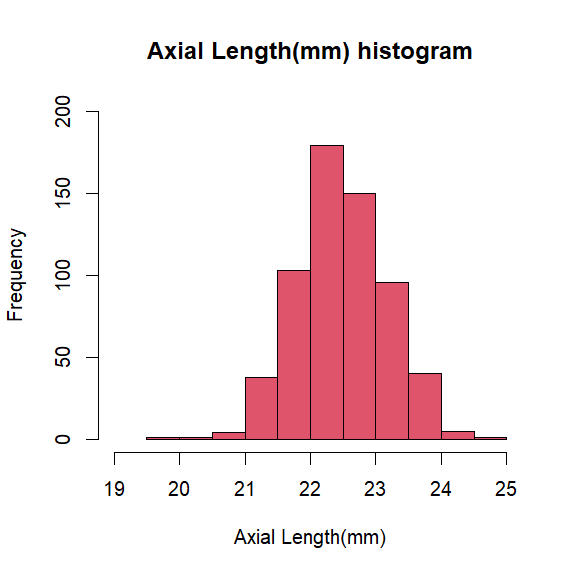
**For age 7:**

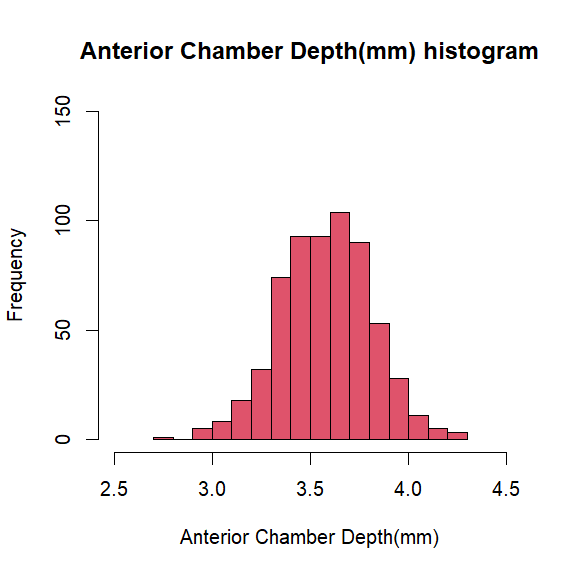
* We get a p-value of 0.000176, by forming a model between MYOPIC and SPHEQ, which is less than 0.05 i.e. this factor is significant for the fitted model.

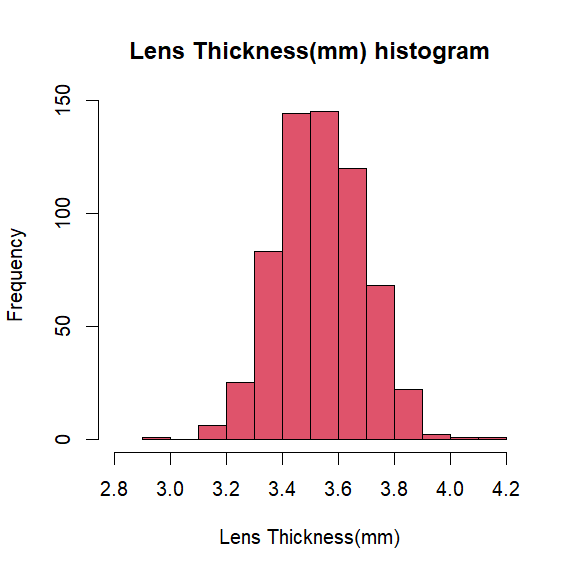
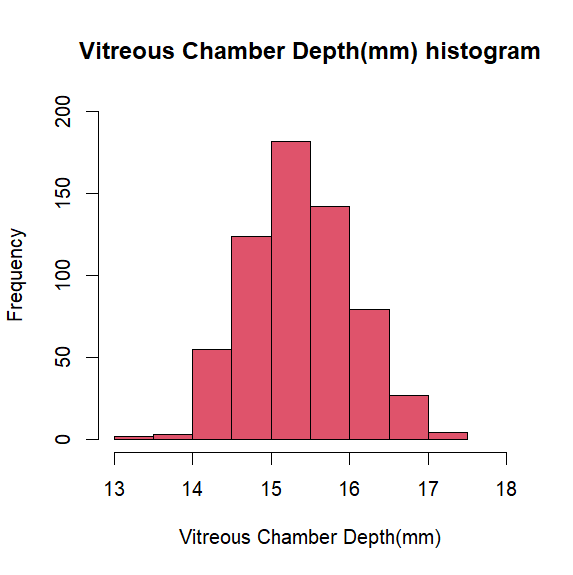
**For age 8:**

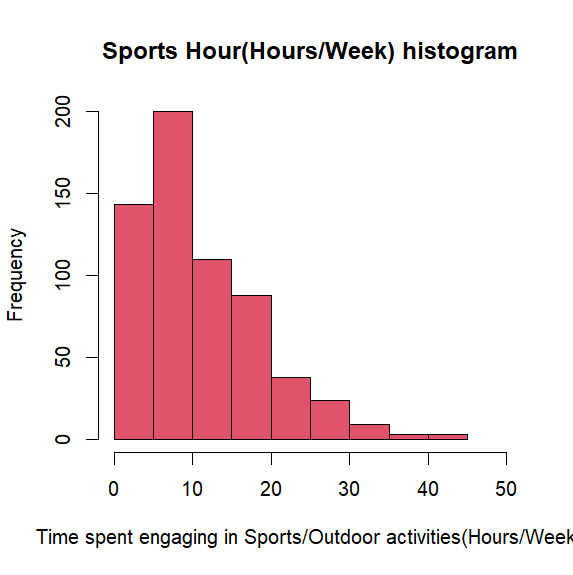
* We get a p-value of 0.0372, by forming a model between MYOPIC and SPHEQ, which is less than 0.05 i.e. this factor is significant for the fitted model.
* **Plots: -**
  + - Histograms: -

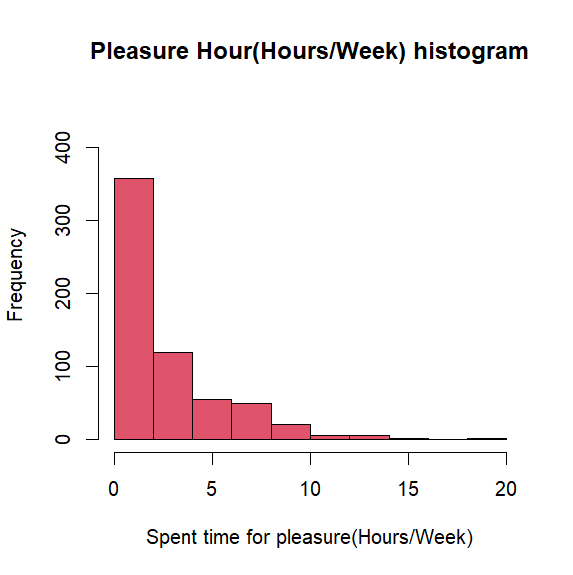


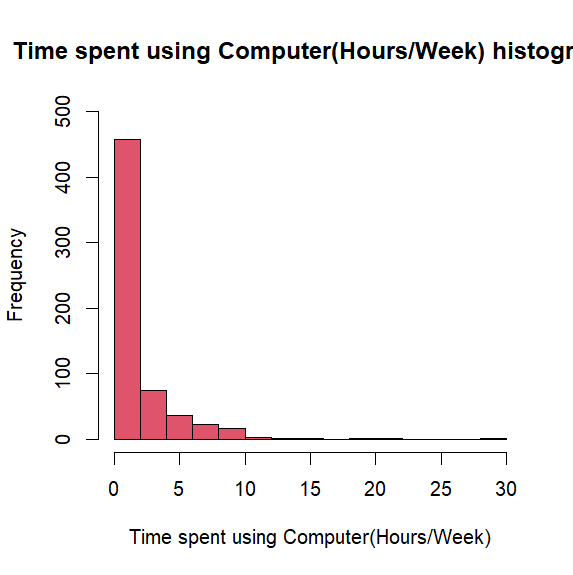


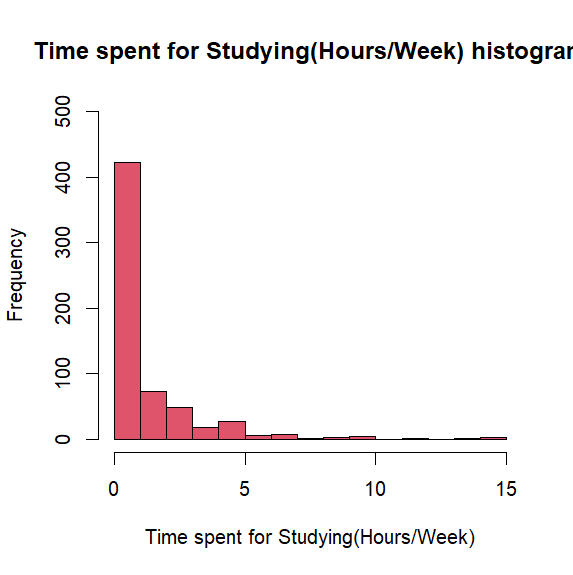


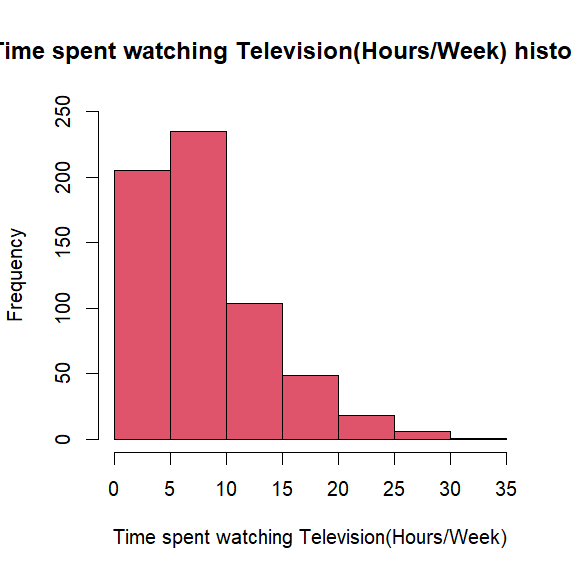
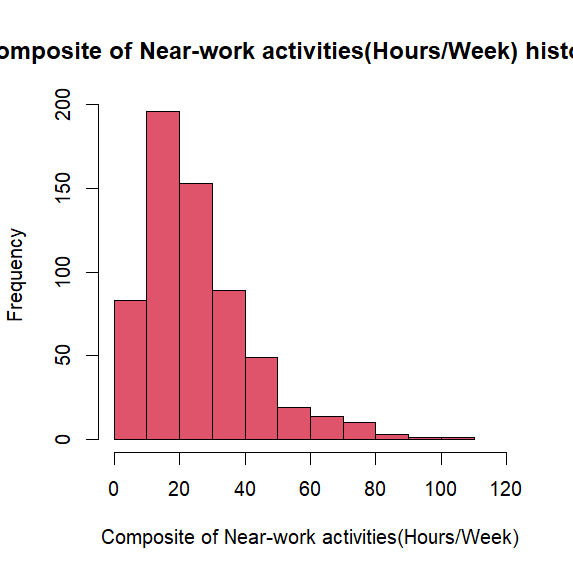




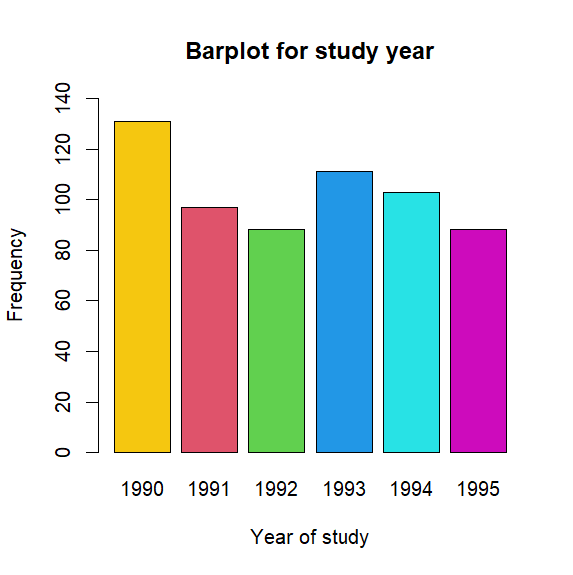


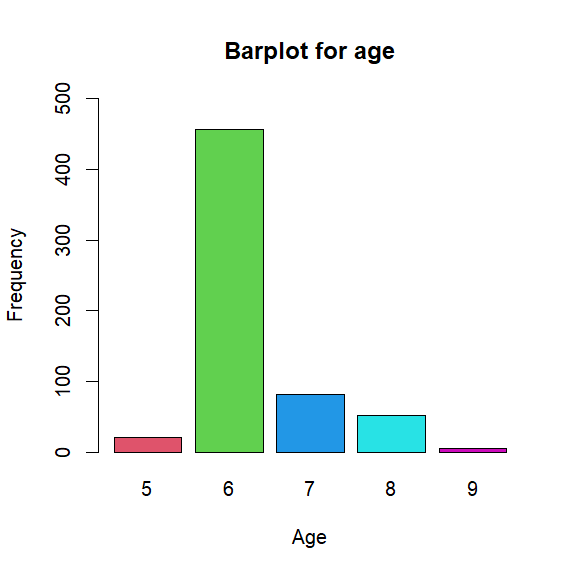


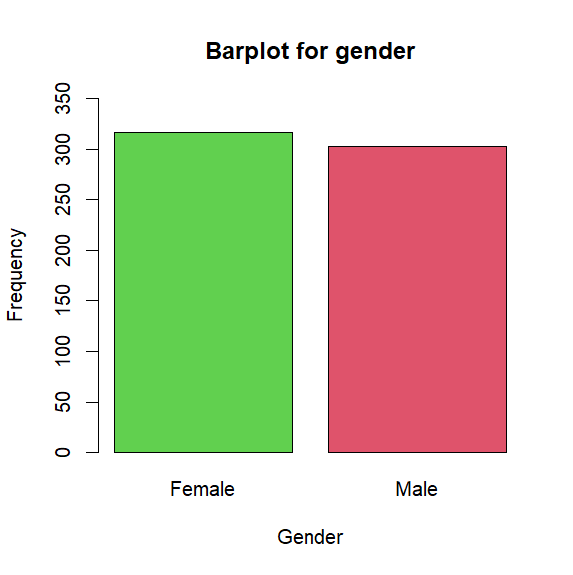


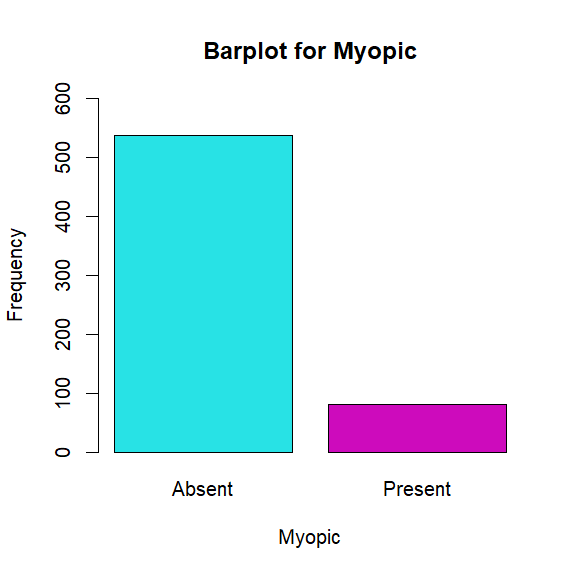
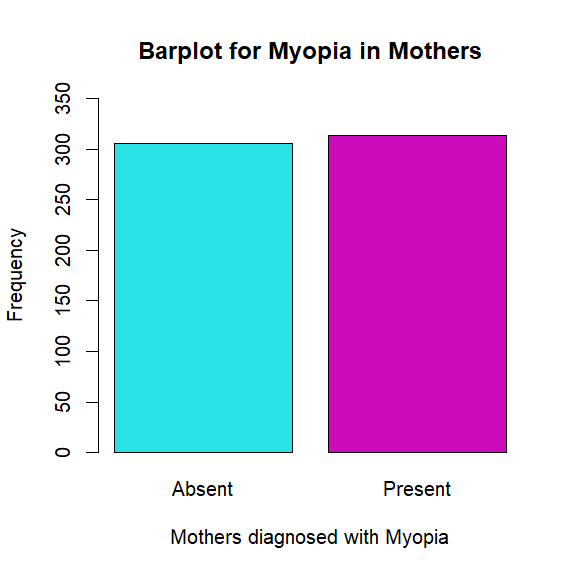


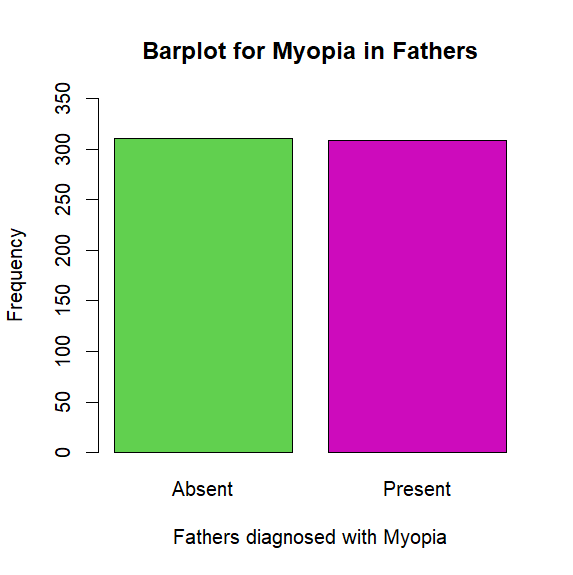
* + - Bar plots: -











* **Acknowledgement: -**

I would like to extend my sincere gratitude to all those who has helped me during the project. Without their valuable assistance, completing the project would not be possible for me.

Firstly, I’ll like to thank Professor *Anjana S.* for her guidance, expertise and continuous motivation. Her feedback and construction of the project was very instrumental in shaping the direction of the project.

I’d like to thank *Mr. Rajib Kumar Pakhira* for his support and guidance throughout the project.

Lastly, I would like to thank my classmates for their helpful attitude and constant support which led me towards the success of the project.

* **References: -**

Advanced Visualization using power BI: - [Power BI Visualization](file:///C:\Users\User\OneDrive\Desktop\Power%20BI%20Course%20resources\Dataset\001.pbix)

The Dataset: - [Myopia data](https://www.kaggle.com/datasets/mscgeorges/myopia-study)