**Final Report**

**Introduction**

The dataset used in this report is collected from a class of math students in secondary school. The data collected varies from a wide range of attributes, specifically 33 attributes. The attributes used in the research are ‘absences’ and ‘famrel’. ‘Absences’ is the number of absences reported and ‘famrel’ is the quality of family relationship a student reports to have ranging from 1 (very bad) to 5 (excellent). The research question extrapolated from this dataset was ‘Is there a difference in the mean number of school absences between students who report different quality of family relationships (1 very bad - 5 excellent)?’.

The null hypothesis of the research question was that ‘there is no difference in the mean number of school absences between students who report different quality of family relationships (1 very bad - 5 excellent)’. The alternative hypothesis was that ‘there is a difference in the mean number of school absences between students who report different quality of family relationships (1 very bad - 5 excellent)’.

**Visualization**

The boxplots shown in plot 1 and plot 2 show how data is distributed. The line through each box shows the median. Those who reported very bad and bad family relationships (1 and 2) showed the highest measurement of the median both being the same as 5.5 number of absences. Those who reported excellent or average family relationships (5 and 3) both had a median of 4.0 as shown on the boxplot and those who reported good family relationships (4) are shown to have had a median of 3.0 as shown on the boxplot.

The lower quartile of the boxplot can be seen as decreasing as students self-report better quality relationships. Those who reported having very bad (1) have a boxplot with the lower quartile of 3.0, those who reported bad (2) have a boxplot with a lower quartile of 2.0 and those who reported average (3) relations showed a lower quartile of 1.5 as shown in the plots. Those reporting good and excellent relations 4 and 5 respectively have boxplots with a lower quartile of 0.

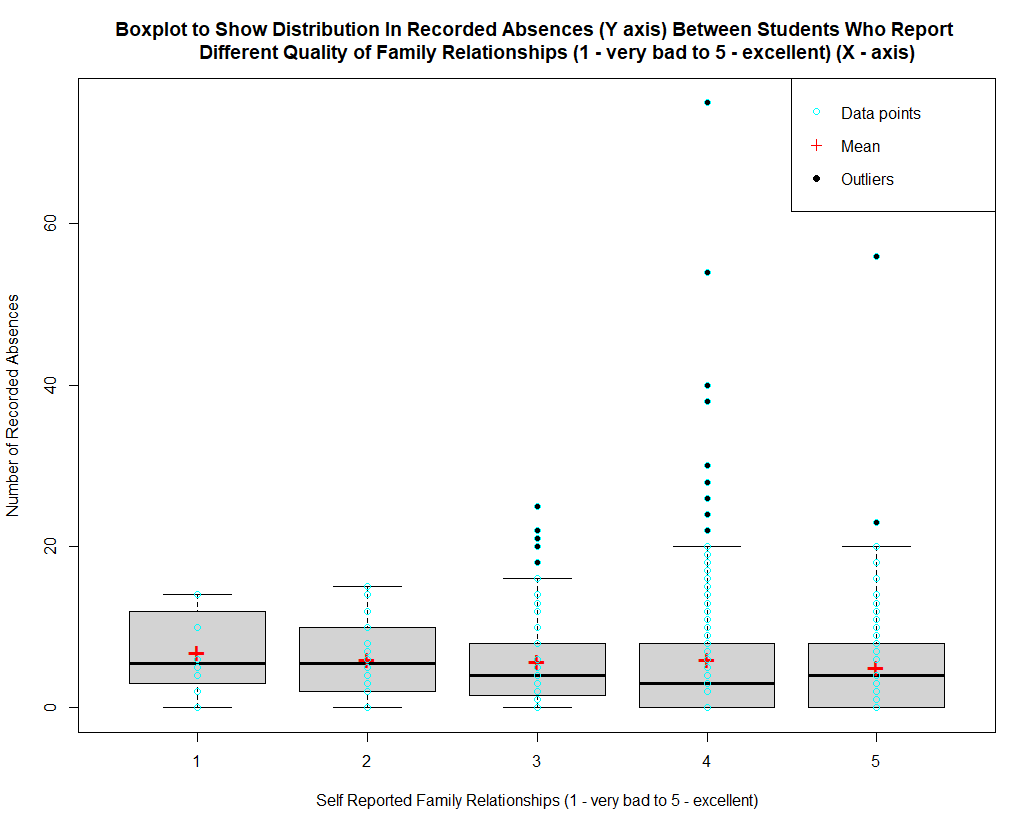
As with the lower quartile a similar pattern is seen with the upper quartile were as students report better quality relations there is then a decrease in the upper quartile of the box. Those reporting very bad and bad (1 and 2) have an upper quartile of 12 and 10 respectively. The rest are shown to have the same upper quartile of 8.

The final aspect of the boxplot are the whiskers that show the lowest and highest values that are still within 1.5 times of the interquartile range. For the lower whisker all reported relationships have the lower whisker of 0. In those who reported good and excellent (4 and 5) do not appear to have whiskers as the lower hinge of the 1st quartile is also 0, so there appears to be no whiskers. However, there are variations in the upper whisker where the value of the higher whisker increases as individuals report better quality family relationships. Those who report very bad have a higher whisker of 14, those who reported bad have a higher whisker of 15, those who report average relations have a upper whisker of 16 followed by an upper whisker of 20 in both those students who report good, and excellent relationship (4 and 5).

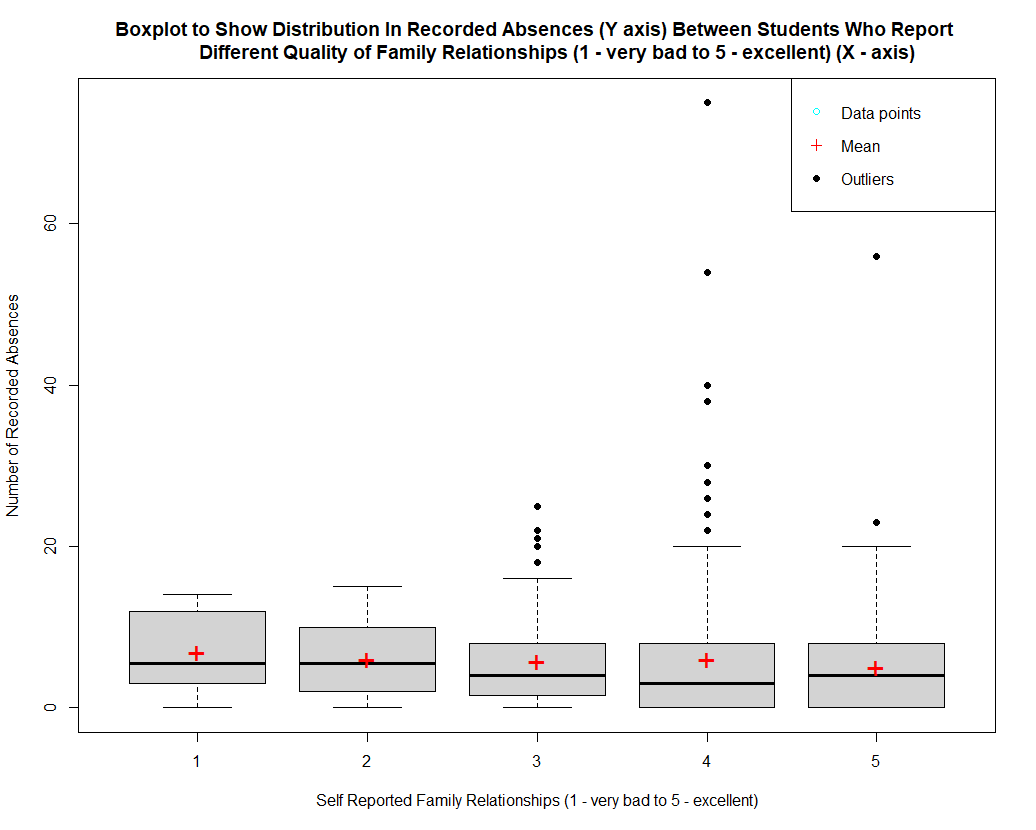
The black points on plots 1 and 2 show the outliers. Those reporting very bad (1) and bad (2) relationships have no outliers. Those reporting average relationships (3) have 5 visible outliers close to the upper whisker. Those reporting good (4) relationships had the most outliers spread the most above the upper whiskers mostly ranging from 20 to 60 with one outlier above 60 number of recorded absences. Those reporting excellent (5) family relationships have two visible outliers in the *Boxplot*, one slightly above 20 the other above 50.

On the boxplots 1 and 2 the mean is indicated with a red ‘+’. In the means there also seems to be a pattern where, as the quality of relations is increased the mean number of absences decreases with those reporting good (4) break the pattern having a higher mean of 6.010256 where those with excellent (5) have mean of 4.952830 and those reporting average (3) have 5.808824 mean. This maybe as a result of those reporting good (3) relations with their family have the highest number of outliers that vary the most along the y-axis.

Upon inspection of these boxplot those who reported poor family relationships (1- very bad and 2 bad) had a less varied boxplot including the box and the whisker. Whereas the rest varied across the y-axis resulting in various outliers. This suggest those reporting poorer relationships have more consistent absences being lower.



***Boxplot 1****: Boxplot to show absences in students of varying family relationships with all data point shown in cyan*



***Boxplot 2****: Boxplot to show absences in students of varying family relationships. Actual plot created in visualization.R due to an error in variable not displaying the data points*

Analysis

Conclusion