Report #4

GROUP 9

Contents

1 Normality Check on Numerical Variables

From previous EDA, we have found that some of the numerical variables might be normally distributed. We will closely look at the normality. Figure 1 shows Quantile-Quantile plot of each numerical variables. For all of the plots, the points are off from the diagonal line, indicating non-normality.

The table 1 shows the results of Shapiro Wilk test (González-Estrada and Cosmes 2019) for normality where they tested:

$$H_0 : X_i \; follows \; Normal \; distribution$$

$$H_0 : X_i \; does \; not \; follow \; Normal \; distribution$$

$$where, \; i = \{Age, Absences, G1, G2, G3\}$$

All of p-values are very small (when rounded at the 3 decimal points they are zeros) so, we have strong evidence against the null hypotheses that the variables are normally distributed.

We can not apply log transformation to the variables since the variables contain zero value (e.g., some student scored zero at the test). Since the variables are not normally distributed and we can not apply log transformation, we need to find some analysis method that is robust to non-normality.

Table 1: Shapiro Wilk Test Result

		W test statistics	P.value
1	age	0.91	0.000
2	absences	0.67	0.000
3	G1	0.97	0.000
4	G2	0.97	0.000
5	G3	0.93	0.000

2 Assumption checks to compare three test scores (G1, G2, and G3) by sex

Figure 2 shows the box plot of G1, G2 and G3 by sex. It seems that male student tend to score higher than female student. We want to find whether the score differ by sex. Figure 3 shows the density plot of tests by sex. It seems variances do not differ much by sex but they might not have normal distribution. To conduct the multivariate analysis on means of vectors of students' test score(Hotelling's T^2 test), we will check equal variance and normality assumptions.

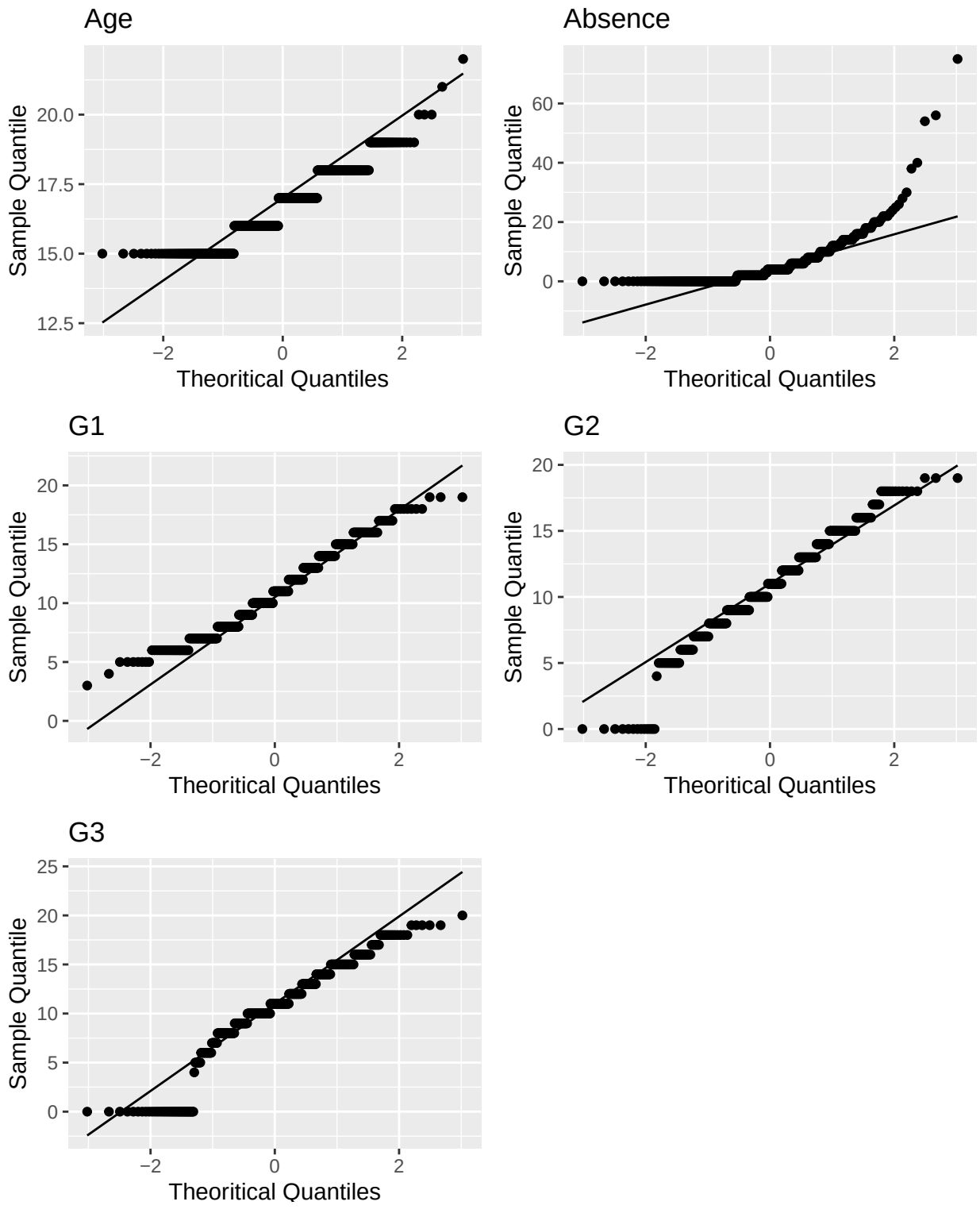


Figure 1: Quantile-Quantile Normality Plot

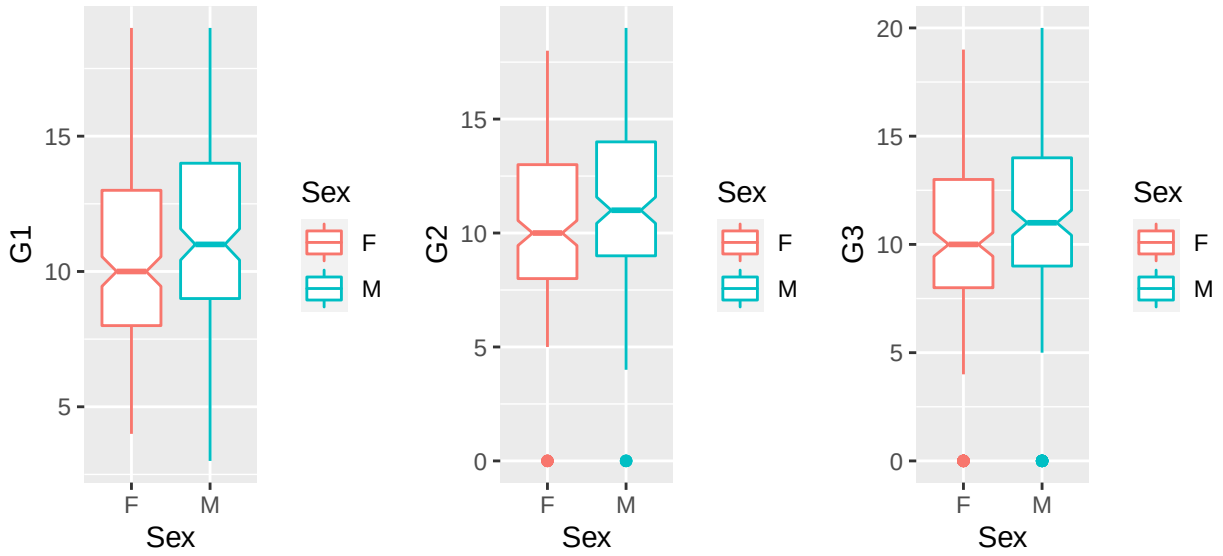


Figure 2: Boxplot

2.1 Test of equality of covariance matrices

We will test equality of covariance matrices using the test discussed by (Cai, Liu, and Xia 2013). The hypotheses are:

$$H_0 : \Sigma_{Female} = \Sigma_{Male}$$
$$H_0 : \Sigma_{Female} \neq \Sigma_{Male}$$

The table 2 shows the results from 4 tests about covariance matrix. They all returned high p-values indicating that we have very small evidence against the null hypothesis that the covariance matrices are equal. Therefore, we can assume equal covariance matrix.

Table 2: Multivariate covariance matrix test

	Tests	Test Statistics	P.values
1	HD	1.01	0.63
2	CLX	1.01	0.64
3	Scott	0.66	0.51
4	LC	-0.66	0.75

2.2 Test for normality of each distribution

We will conduct the hypotheses test to check normality of each distributions by using Shapiro Wilk Test (González-Estrada and Cosmes 2019). The hypotheses are:

$$H_0 : X_{i,j} \ follows \ Normal \ distribution$$

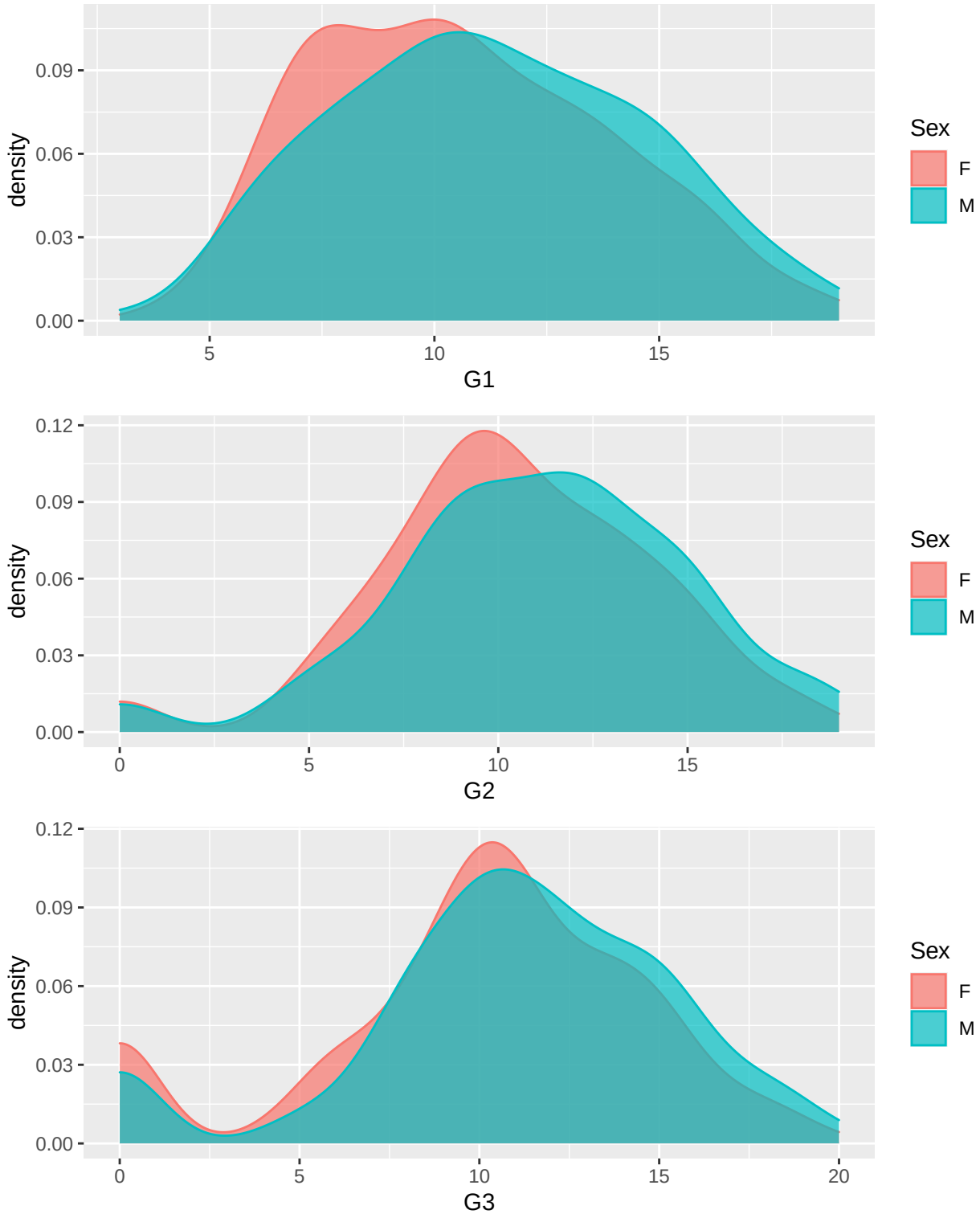


Figure 3: Density plot

$$H_0 : X_{i,j} \; does \; not \; follow \; Normal \; distribution$$
$$where, \; i = \{Female, Male\} \; and \; j = \{G1, G2, G3\}$$

So, in total, we have conducted six hypothesis tests.

The table 3 and the table 4 show the results of Shapiro Wilk test on each tests score(G1, G2, and G3) by sex respectively. Since all p-values are very small we have very strong evidence against that the variables have normal distribution, indicating the violation of normality assumptions. Since the normality assumptions are violated we should not conduct Hotelling's T^2 test. We need to find some alternative methods to find whether the tests score differ among sex.

Table 3: Normality test on Female tests score

	Female	W test statistics	P.value
1	G1	0.97	0.00
2	G2	0.97	0.00
3	G3	0.92	0.00

Table 4: Normality test on Male tests score

	Male	W test statistics	P.value
1	G1	0.98	0.01
2	G2	0.97	0.00
3	G3	0.93	0.00

2.3 Kruskal-Wallis test

Since we can not conduct Hotelling's T^2 test we will just test the significant difference in G3 scores by sex (we won't do the test for G1 and G2 because we want to control for family wise error). We used Kruskal Wallis test (Hollander and Wolfe 1973):

$$H_0 : Medians\ of\ G3\ test\ scores\ do\ not\ differ\ among\ sex$$

$$H_1 : Medians\ of\ G3\ test\ scores\ differ\ among\ sex$$

Kruskal-Wallis test is nonparametric test so, it is robust to non-normality. The table 5 shows the result of Kruskal-Wallis test. Since p-value is around 0.05, we conclude that we have some evidence against the null hypothesis that the medians of G3 scores are same among sex.

Table 5: Kruskal-Wallis rank sum test: `G1` by `Sex`

Test statistic	df	P value
3.551	1	0.05951

3 The Mahalanobis distance

Mahalanobis Distance is a measure of the distance between a point and a distribution. It works very well for multivariate data because it uses covariance between variables to find the distance between two points. And it works well when variables are highly correlated, even if their scales are not the same.

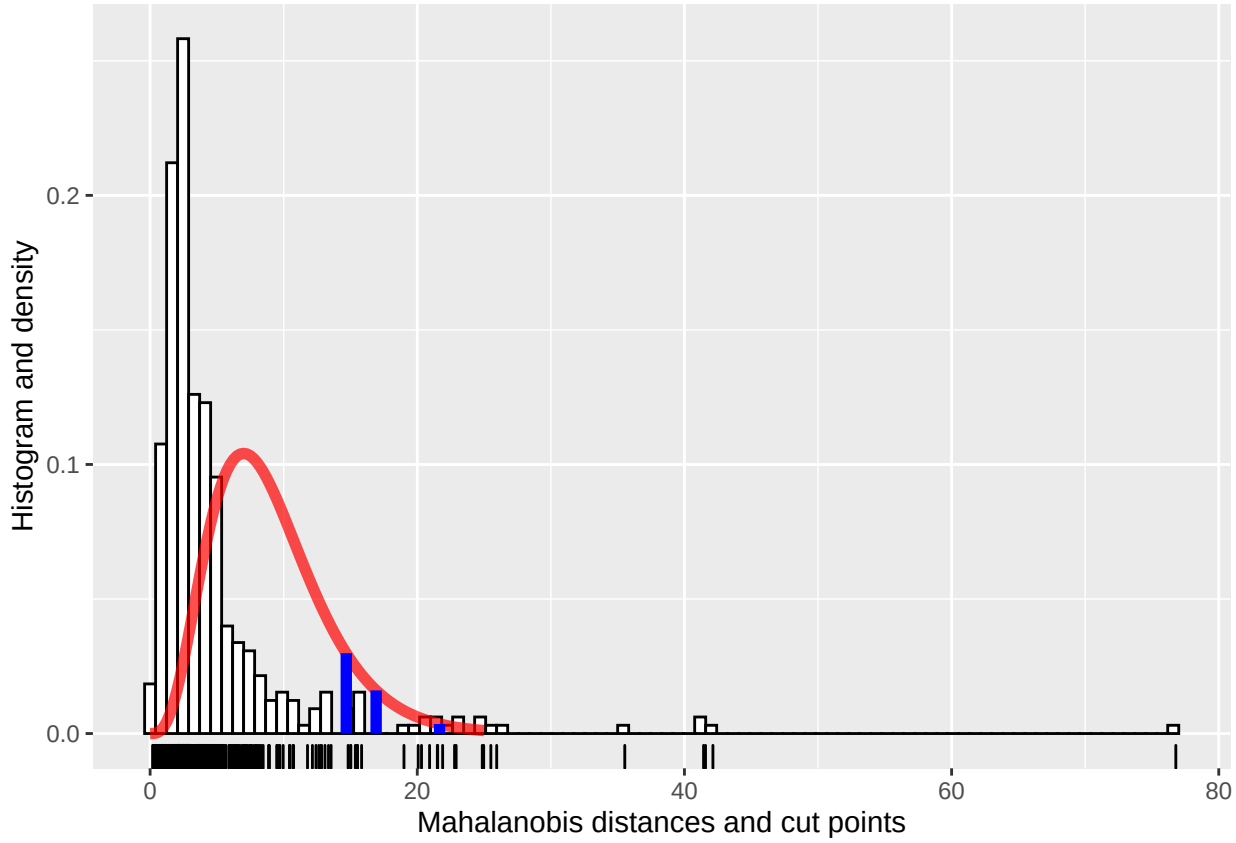
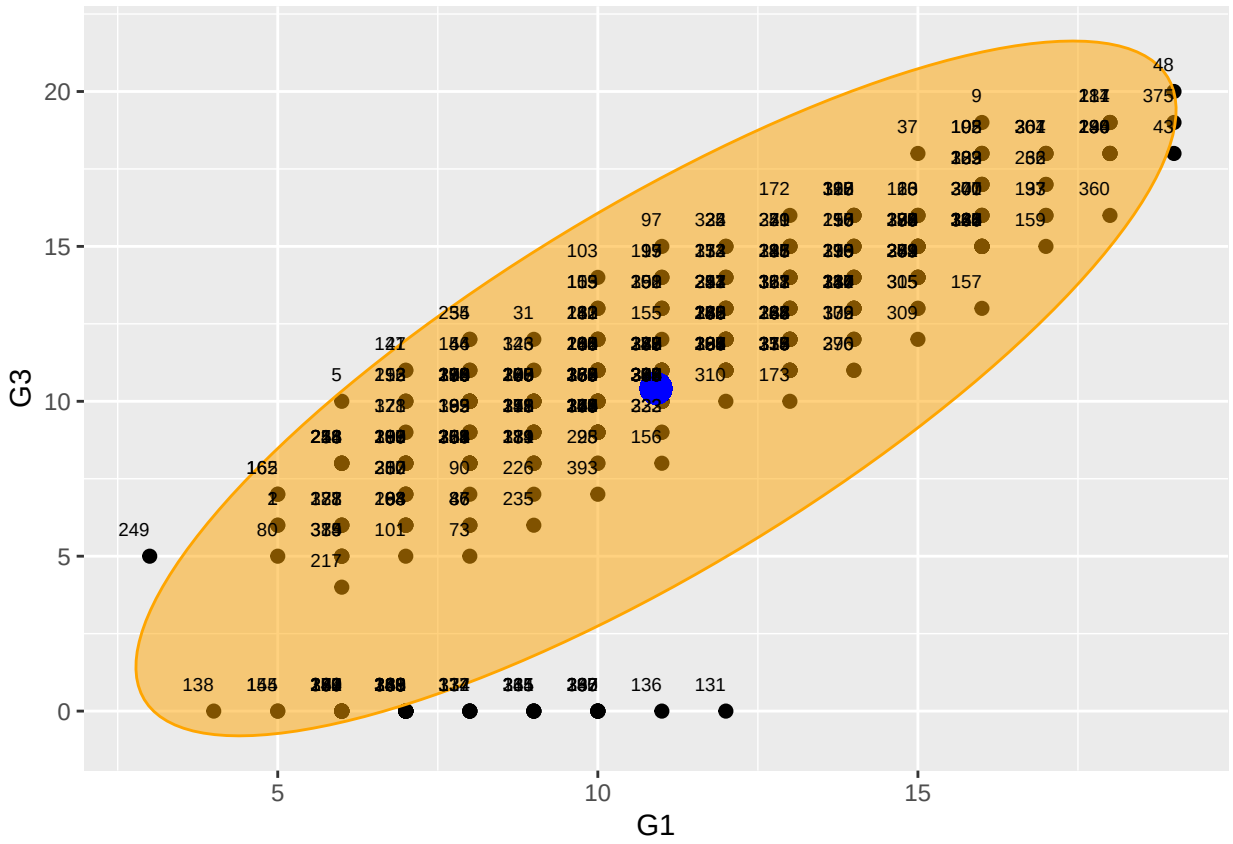


Figure 4: Mahalanobis distances Plot

3.1 Multivariate Outlier Detection

Mahalanobis Distance also gives reliable results when outliers are considered as multivariate.(Cansiz 2020) To find outliers, the distance between every point and centre in the multi-dimension data is calculated, and outliers are found by considering these distances.



Above is the plot of our data and an ellipse from considering center point and covariance matrix. Blue point on the plot shows the center point. Black points are the observations for G1 — G3 variables. As you can see, there are points outside the orange ellipse. It means that these points might be the outliers. If we consider that this ellipse has been drawn over covariance, center and radius, we can say we might have found the same points as the outlier for Mahalanobis Distance.

Typical	Somewhat	Surprising	Very
369	8	6	12

Finally, we have identified the outliers as Very in the above table in our multivariate data. They total number of outliers look pretty similar as the points outside of the ellipse in the scatter plot.

4 Identify the distribution of interested variables

We are motivated by this dataset since we are interested in what features affect overall math grades. We will mainly focus on exploring the distribution of the G3 variable, which is students' final grades.

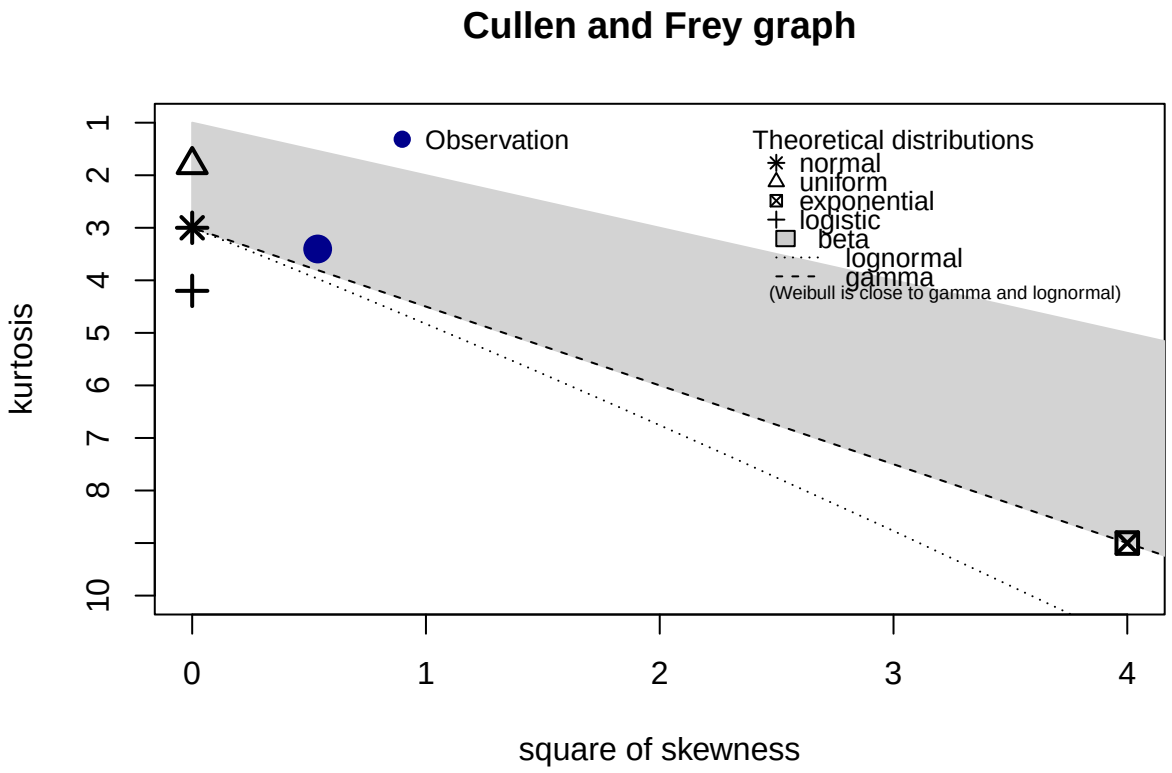


Figure 5: Cullen and Frey graph of G3 variables

```
## summary statistics
## ------
## min:  0   max:  20
## median:  11
## mean:  10.41519
## estimated sd:  4.581443
## estimated skewness:  -0.7326724
## estimated kurtosis:  3.403421
```

Looking at the results on the above graph, the data has a negative skewness and a kurtosis not far from 3, the fit of two common left-skewed distributions could be considered, Weibull and gamma distributions.

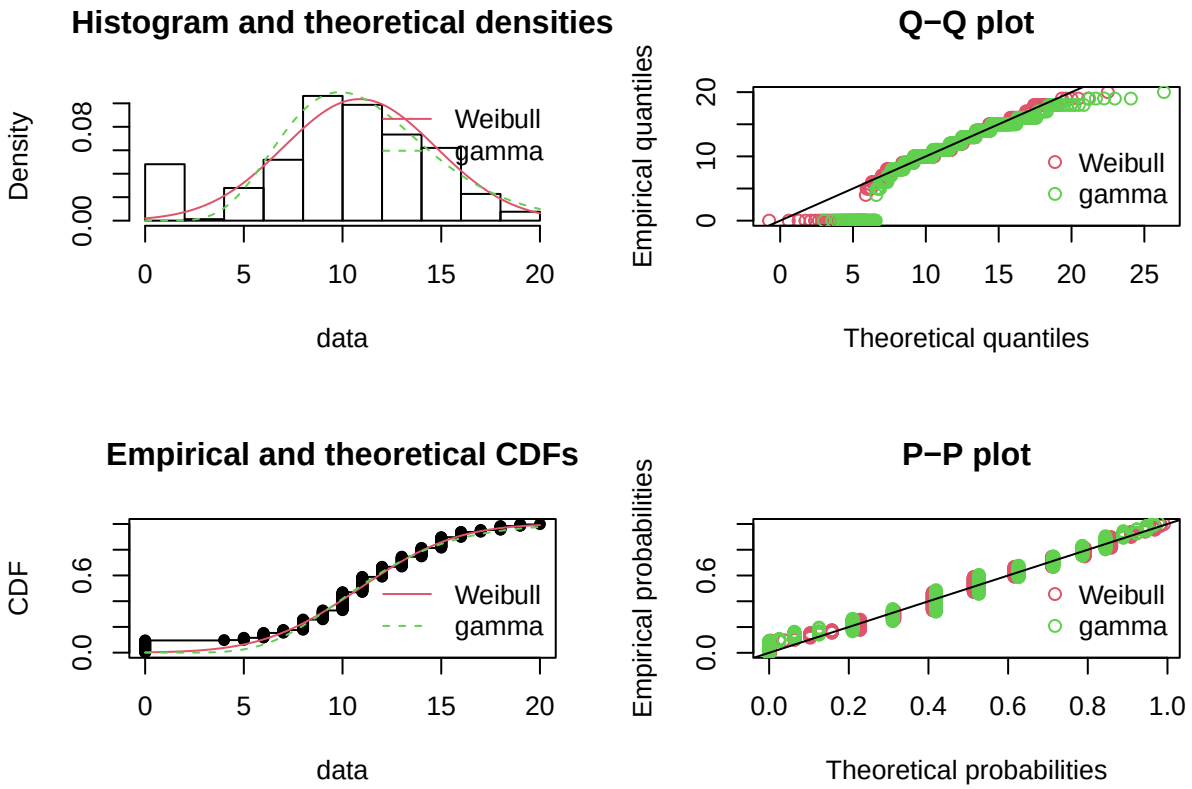


Figure 6: Goodness-of-fit plots for Weibull and gamma distributions fitted to G3 variable)

None of the fitted distributions correctly describes the center of the distribution, but the Weibull distribution could be prefered for their better description of the right tail of the empirical distribution, especially if this tail is important in the use of the fitted distribution, as it is in the context of high grades in Maths.

References

Cai, Tony, Weidong Liu, and Yin Xia. 2013. "Two-Sample Covariance Matrix Testing and Support Recovery in High-Dimensional and Sparse Settings." *Journal of the American Statistical Association* 108 (501): 265–77. https://doi.org/10.1080/01621459.2012.758041.

Cansiz, Sergen. 2020. "Mahalanobis Distance and Multivariate Outlier Detection in r." https://towardsdatascience.com/mahalonobis-distance-and-outlier-detection-in-r-cb9c37576d7d .

González-Estrada, Elizabeth, and Waldenia Cosmes. 2019. "Shapiro–Wilk Test for Skew Normal Distributions Based on Data Transformations." *Journal of Statistical Computation and Simulation* 89 (17): 3258–72. https://doi.org/10.1080/00949655.2019.1658763.

Hollander, M., and D. A. Wolfe. 1973. "Nonparametric Statistical Methods," 115–20.