Report #4

GROUP 9

Contents

1 Multivariate Normality Check on All Numerical Variables

Let's start looking at the normality of numerical variables without splitting data. 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 Mardia's Test (Mardia 1970) for multivariate normality where they tested:

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

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

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

p-values were close to zero so, we have very strong evidence against the null hypothesis that the variables follow multivariate normal distribution. The results were expected since we did not split the data by factors. Next chapter will explore multivariate normal distribution of data by sex.

Table 1: Maldia's Test

	Beta-hat	kappa	p-val
Skewness	29.74	1957.72	0.00
Kurtosis	71.08	42.86	0.00

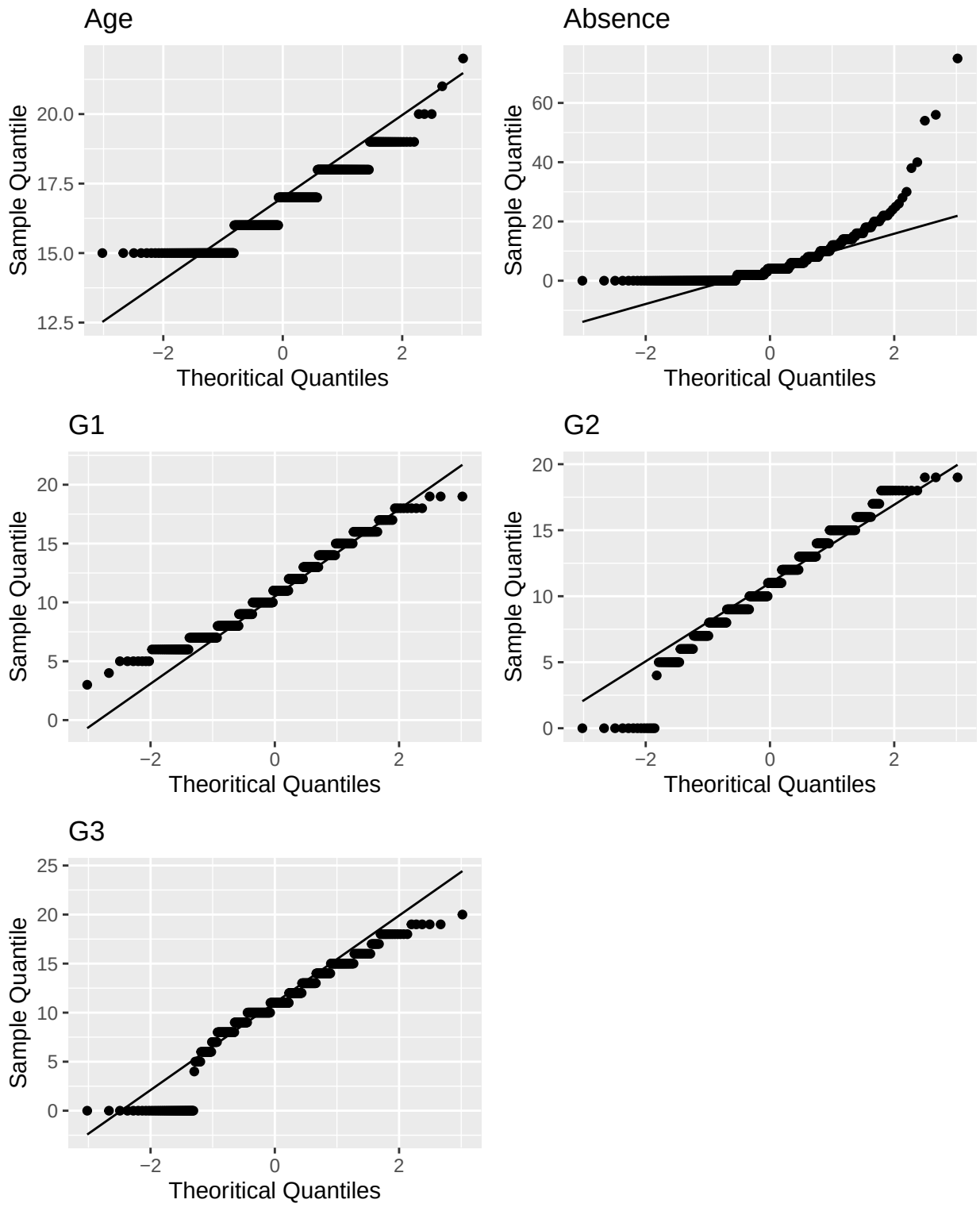

Figure 1: Quantile-Quantile Normality Plot

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.

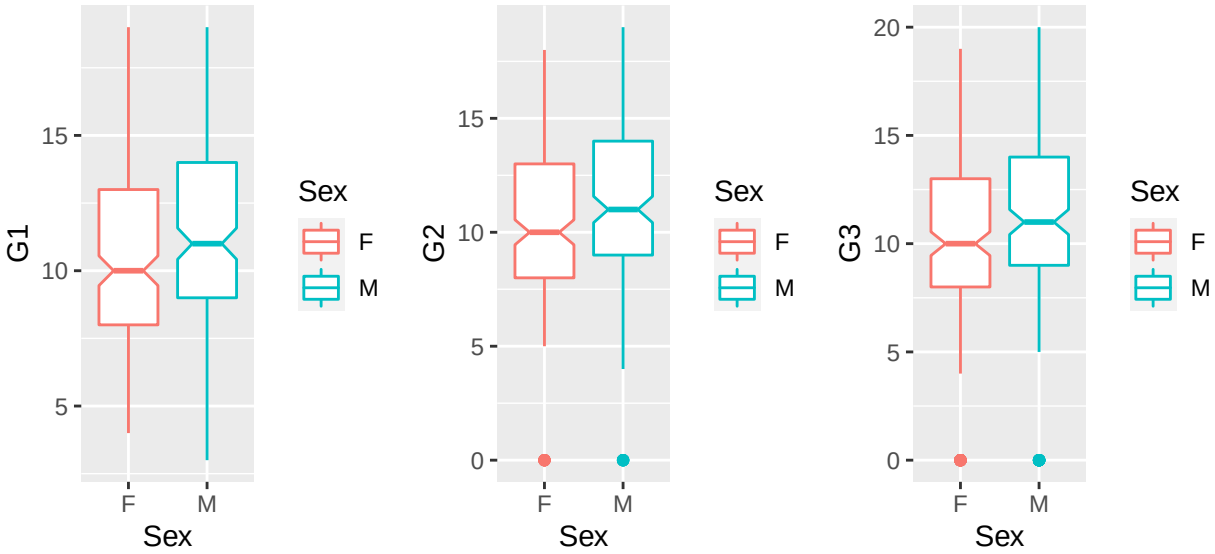

Figure 2: Boxplot

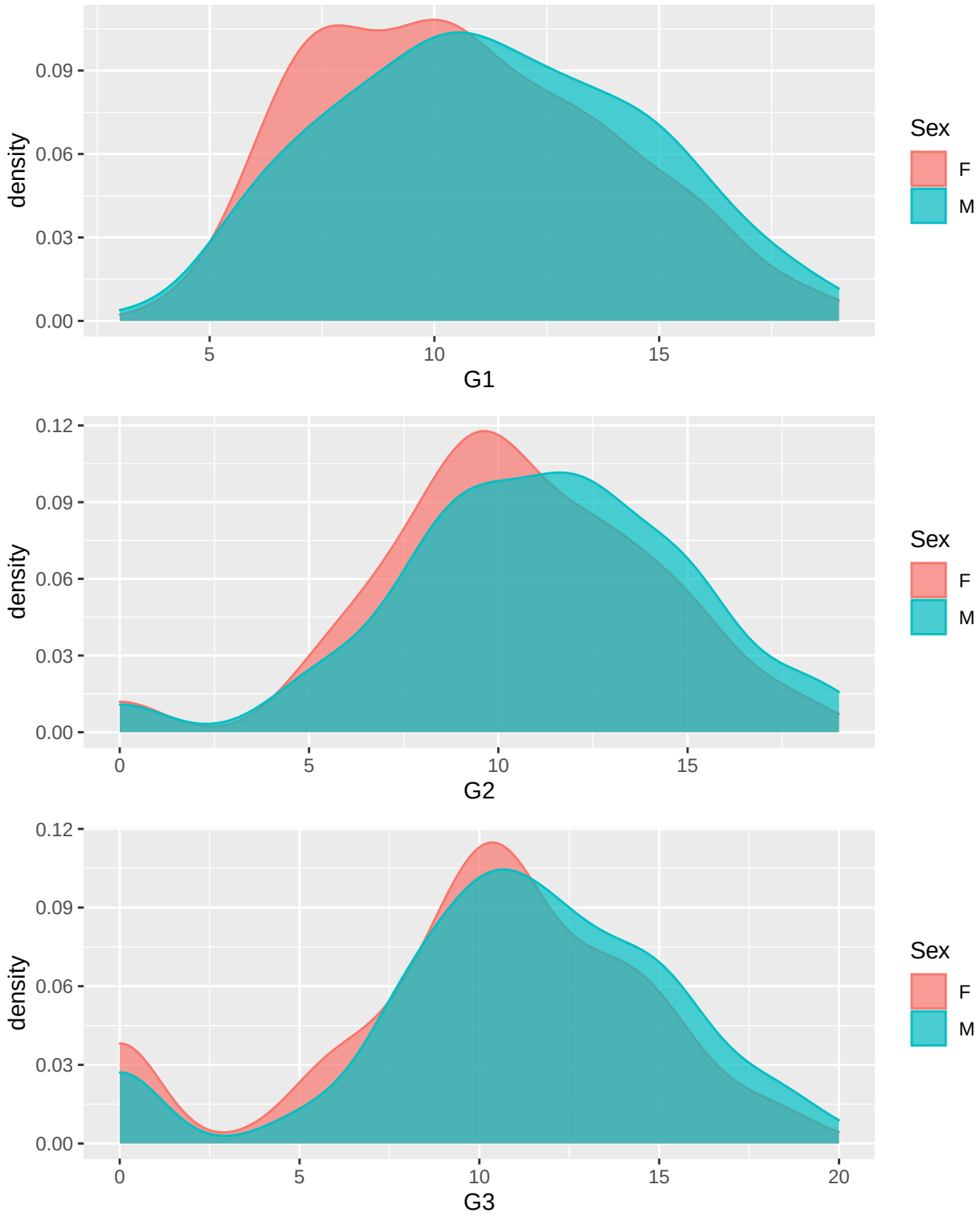

Figure 3: Density plot

2.1 Test for multivariate normality

We will conduct the hypotheses test to check multivariate normal distribution by using Mardia's test. The hypotheses are:

$$H_0 : X_i \ follows \ multivariate \ normal \ distribution$$

$$H_0 : X_i \ does \ not \ follow \ multivariate \ normal \ distribution$$

$$where, \ i = \{Female, Male\} \ and \ X = (X_{G1}, X_{G2}, X_{G3})$$

The table 2 and 3 show the result of multivariate normal test. Since p-values are very small we have very strong evidence against the null hypothesis that the variables for each sex do not have multivariate normal distribution. Since the normality assumptions are violated we should not conduct Hotelling's T^2 test.

Table 2: Maldia's Test for Female data

	Beta-hat	kappa	p-val
Skewness	13.47	466.82	0.00
Kurtosis	29.22	18.72	0.00

Table 3: Maldia's Test for Male data

	Beta-hat	kappa	p-val
Skewness	12.22	380.89	0.00
Kurtosis	30.69	19.58	0.00

2.2 Test of equality of covariance matrices

We will also 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 4 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 4: 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

2.3 Significance test on G3 scores by sex (Kruskal-Wallis test)

Since we can not conduct Hotelling's T^2 test (because normality assumption was violated!) 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):

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

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

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 More multivariate normal test by splitting data by Sex and Failures

Since we could not detect multivariate normal distribution by sex. We will try to split data by Sex and whether the students have failed the other course (there are 4 datasets in total). The Mardia's test was conducted with the hypotheses that:

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

$$H_0 : X_{i,j}\ does\ not\ follow\ multivariate\ normal\ distribution$$

$$where,\ i = \{Female, Male\}\ ,\ j = \{Never\ failed,\ Have\ failed\}$$

$$and\ X = (X_{age}, X_{absence}, X_{G1}, X_{G2}, X_{G3})$$

The table 6,7,8 and 9 show the result of Mardia's test for each dataset. P-values are very small for all tests (except the Kurtosis of data of Male students and who have failed the other course). We are concluding that the numerical variables do not follow multivariate normal distribution even though we split data by sex and failures.

Table 6: Female and has failed the other course

	Beta-hat	kappa	p-val
Skewness	15.87	105.8	4.901e-09
Kurtosis	41.48	2.448	0.01437

Table 7: Male and has failed the other course

	Beta-hat	kappa	p-val
Skewness	10.48	75.14	9.406e-05
Kurtosis	35.42	0.1643	0.8695

Table 8: Female and has NEVER failed the other course

	Beta-hat	kappa	p-val
Skewness	40.14	1124	0
Kurtosis	76.29	31.98	0

Table 9: Male and has NEVER failed the other course

	Beta-hat	kappa	p-val
Skewness	24.02	576.5	0
Kurtosis	66.68	22.72	0

4 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.

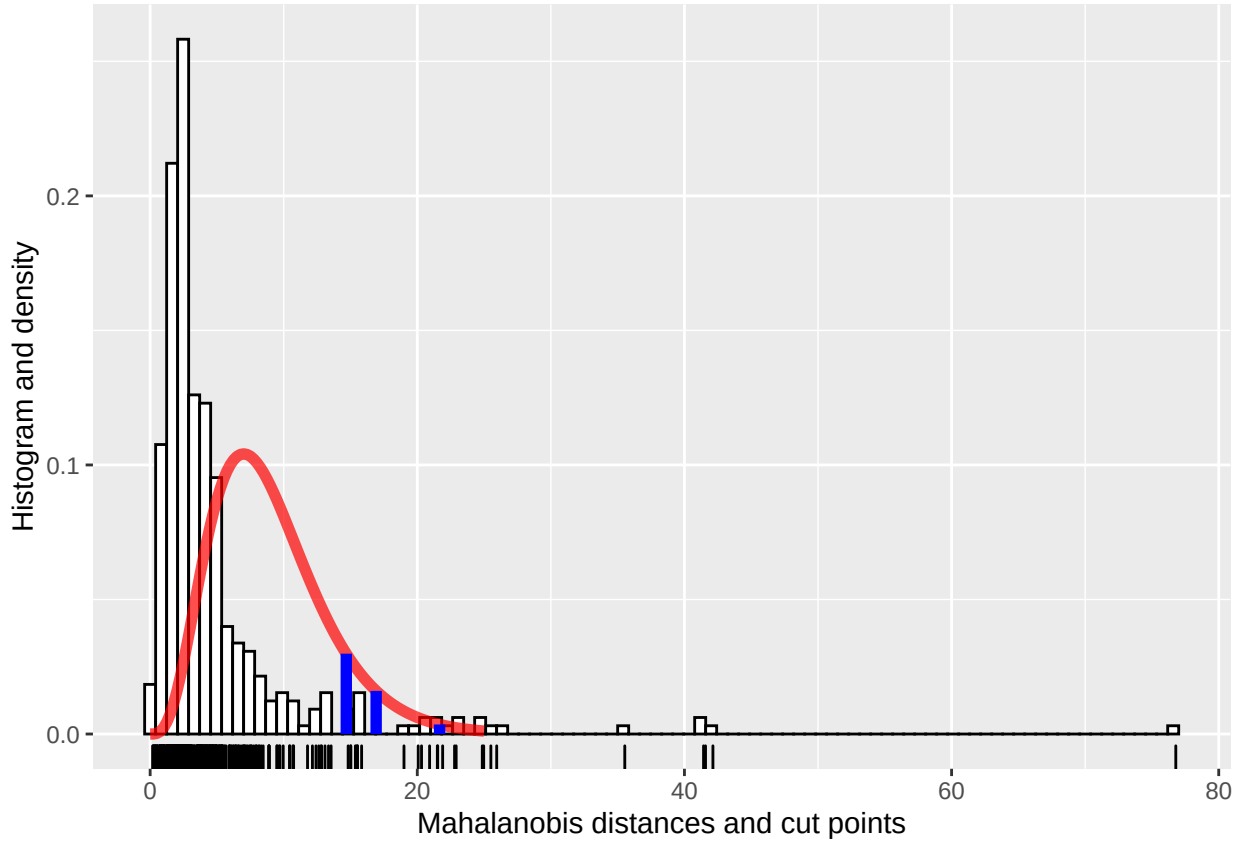

Figure 4: Mahalanobis distances Plot

4.1 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.

Figure 5 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.

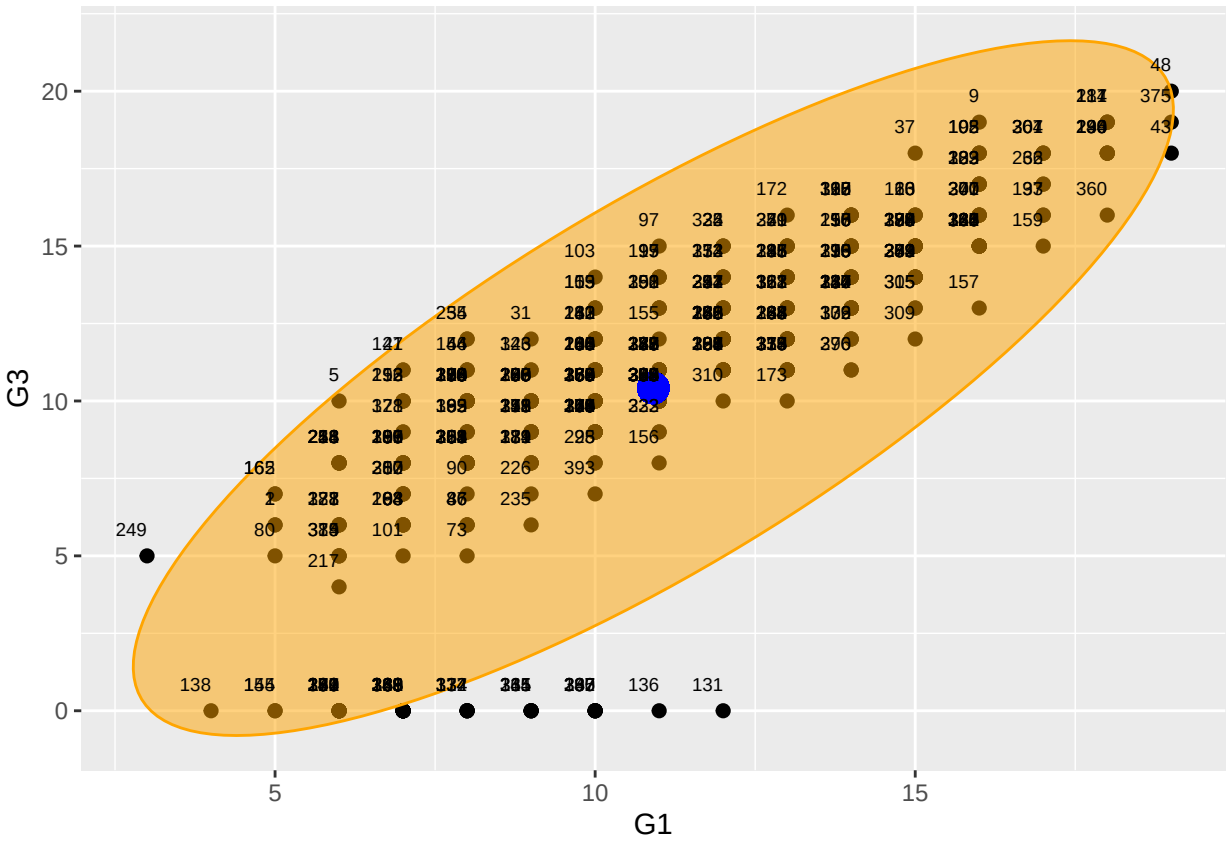

Figure 5: Scatter plot of G1 and G3 variables

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.

5 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.

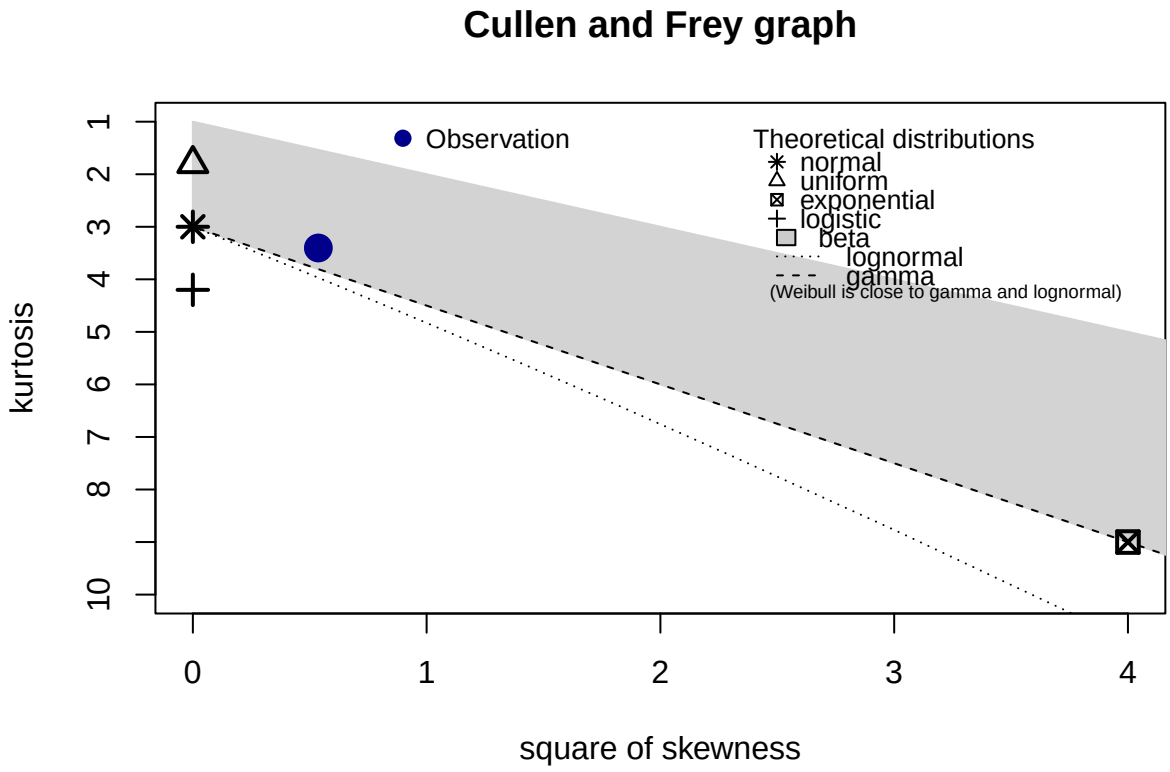

Figure 6: Cullen and Frey graph of G3 variables

```
## 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 in Figure 6 and the R output, 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.

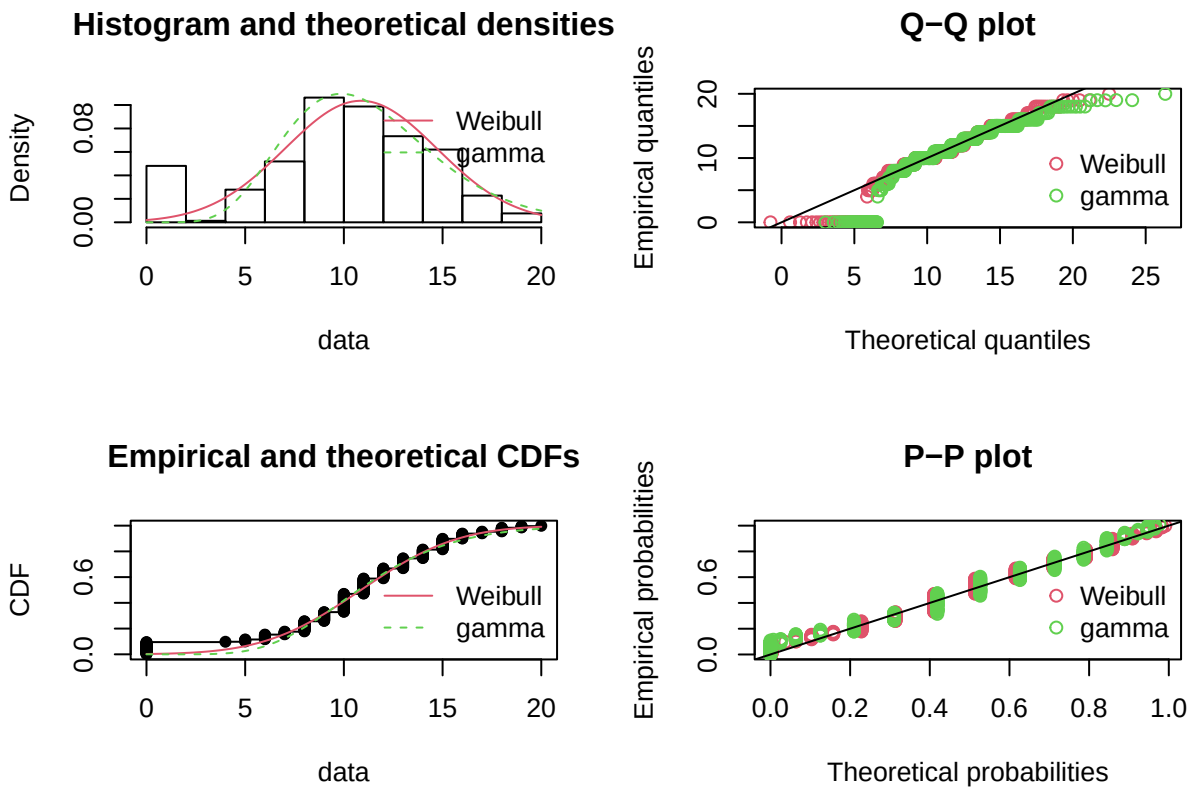

Figure 7: Goodness-of-fit plots for Weibull and gamma distributions fitted to G3 variable

In Figure 7, 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.(Delignette-Muller and Dutang 2014)

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 .

Delignette-Muller, Marie Laure, and Christophe Dutang. 2014. "Fitdistrplus: An r Package for Fitting Distributions." https://cran.r-project.org/web/packages/fitdistrplus/vignettes/paper2JSS.p
.

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

Mardia, K. V. 1970. "Measures of Multivariate Skewness and Kurtosis with Applications." *Biometrika* 57 (3): 519–30. http://www.jstor.org/stable/2334770.