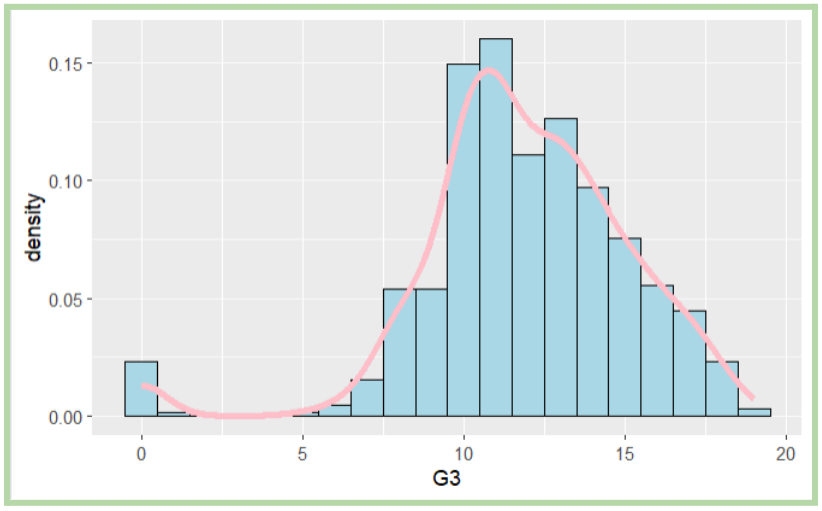
Project 2

1. **Chi-Squared Goodness of Fit Test:**
    a. Null hypothesis (H_0): The distribution of the number of failures in the subject of portuguese among genders is equal in proportion
    b. Alternative Hypothesis (Ha): The distribution of the number of failures in the subject of portuguese among genders is not equal in proportion
    c. Alpha (α)= 0.05
    d. Testing for assumptions:
        i. The data provided is categorical/can be put in categories
        ii. The observed data is independent of one another
        iii. Contingency Table in R-document
    e. Result/Conclusion:
        i.
```
        Chi-squared test for given probabilities

data:  observed_freq
X-squared = 1.0901, df = 1, p-value = 0.2965
```
            1. As observed in the results, the p-value for the Chi-squared goodness of fit test is 0.2965, which is higher than the alpha of 0.05 which was chosen. In the context of this problem, this means that I fail to reject the null hypothesis. As a result, there is no significant evidence to suggest that the distribution of failures among genders is unequal in proportion.

2. **ANOVA:**
    A. Null hypothesis (H_0:) :
        a. H_{01}: There is no statistically significant difference in G3 grades between students who do or do not wants to take higher education
        b. H_{02}: There is no statistically significant difference in G3 grades between students who did or did not have extra paid classes within the course subject
        c. H_{03}: The interaction between extra classes and seeking higher education does not significantly predict G3 grades.
    B. Alternative Hypothesis (Ha:):
        a. H_{a1}: There is a statistically significant difference in G3 grades between students who do or do not wants to take higher education

b. H_{a2}: There is a statistically significant difference in G3 grades between students who did or did not have extra paid classes within the course subject
c. H_{a3}: There is an interaction effect between freetime, studytime, and G3 grades

a. Alpha (α)= 0.05
b. Testing of assumptions:
   i. Independent observations in data
   ii. Normally distributed data: While checking the normality of the G3 data, I utilized histogram to see if any of the data was left or right skewed, but it was normally distributed
      1. 

   iii. In the R-code I also performed a shapiro test along with qq plots
   iv. Finally I ran a levene test to assess the variance of the data in Rstudio

c. Results/Conclusion:

   i. Additive 2-Way ANOVA & TukeyHSD:
      1.

```
> G3_add <- aov(G3 ~ higher + paid, data = student_por)
> summary(G3_add)
             Df Sum Sq Mean Sq F value Pr(>F)
higher        1    746   746.2  80.477 <2e-16 ***
paid          1     27    26.8   2.888 0.0897 .
Residuals   646   5990     9.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

```
Fit: aov(formula = G3 ~ higher + paid, data = student_por)

$higher
           diff      lwr     upr p adj
yes-no 3.478761 2.717292 4.24023     0

$paid
            diff       lwr       upr     p adj
yes-no -0.854485 -1.842113 0.1331434 0.0898154
```

2.

a. In the results, the p-value for "paid" is higher than the alpha of 0.05, meaning that we fail to reject the null hypothesis. In the context of this problem this means that there is not a significant difference in means for students. Though, for the "higher" category the p-value is very close to 0, meaning that we reject the null hypothesis. In the context of this problem, this means that there is a statistically significant difference in G3 grades between students who did or did not have extra paid classes within the course subject.

ii. ANOVA with interaction & TukeyHSD:

1.

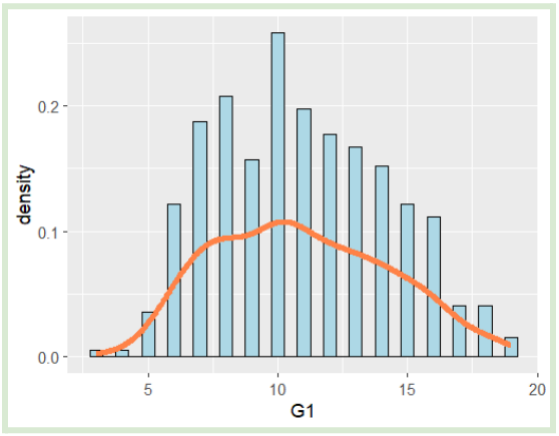
```
> G3_interact <- aov(G3 ~ higher * paid, data = student_por)
> summary(G3_interact)
             Df Sum Sq Mean Sq F value Pr(>F)
higher        1    746   746.2  80.482 <2e-16 ***
paid          1     27    26.8   2.888 0.0897 .
higher:paid   1     10     9.7   1.045 0.3071
Residuals   645   5981     9.3
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.

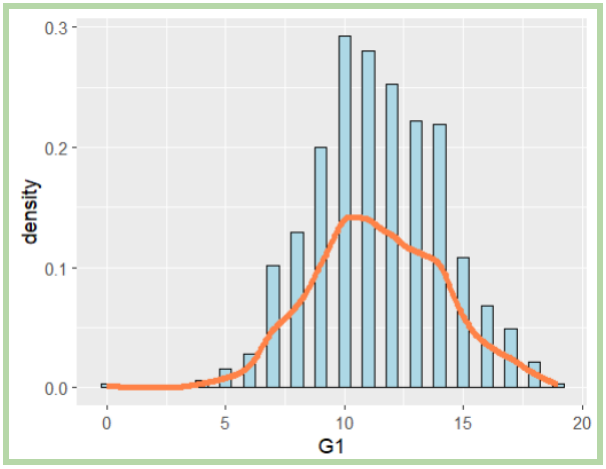
```
$`higher:paid`
                      diff        lwr       upr     p adj
yes:no-no:no     3.5806595  2.5583405 4.6029786 0.0000000
no:yes-no:no     0.9090909 -3.7209510 5.5391328 0.9577161
yes:yes-no:no    2.5757576  0.9506959 4.2008192 0.0002920
no:yes-yes:no   -2.6715686 -7.2123078 1.8691705 0.4287651
yes:yes-yes:no  -1.0049020 -2.3546611 0.3448571 0.2215810
yes:yes-no:yes   1.6666667 -3.0465065 6.3798398 0.7990862
```

a. In the results, the p-value for "paid" is higher than the alpha of 0.05, meaning that we fail to reject the null hypothesis. In the context of this problem this means that there is not a significant difference in means for students. Though, for the "higher" category the p-value is very close to 0, meaning that we reject the null hypothesis. In the context of

this problem, this means that there is not a statistically significant difference in G3 grades between students who did or did not have extra paid classes within the course subject.

3. **Additional test: Independent t-test**
    a. <u>Null Hypothesis:</u>
        i. H0: there is significant difference between the G1 grades in the subjects of mathematics and portuguese
            1. μMathematics = μPortuguese
    b. <u>Alternative hypothesis:</u>
        i. Ha: There is a difference between the mean of the G1 grades in the subjects of mathematics and portuguese is not 0
            1. μMathematics ≠ μPortuguese
    c. <u>Assumptions:</u>
        i. *Independence:*
            1. The two grades are independent of each other as they are from different subjects as opposed to the same subject, this allows for this assumption to be met
        ii. *Normality*:
            1. 

                a. The histogram for the the G1 grades for math are normally distributed and do not seem to have an extreme left or right skew present

2. 

   a. The histogram for the G1 grades for Portuguese seem to be normally distributed as well, without having an extreme right or left skew present.

d. Results/Conclusion:

i.
```
        Welch Two Sample t-test

data:  student_por$G1 and student_math$G1
t = 2.4664, df = 715.02, p-value = 0.01388
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1000007 0.8804288
sample estimates:
mean of x mean of y
 11.39908  10.90886
```

   1. Because the p-value of 0.01388 is less than the alpha of 0.05, I reject the null hypothesis meaning that there is a significant difference between the difference in the mean of G1 for mathematics and G1 for Portuguese, which is not 0. Additionally, the G1 grade for Portuguese is significantly higher than for mathematics.