This exam is open book and open internet but you are NOT allowed to work with anyone else or ask anyone other than Meha or Oscar any questions about the exam. It is due at noon on Sunday, April 23.

Please answer the following questions by analyzing the associated datasets. For all tests, please:

* check whether the data meet the requirements/assumptions of the test you plan to run
* complete any transforms needed to make the data meet the required assumptions
* run the test
* interpret the results (do not include only the R output)
* check model fit in the case of linear regressions and/or glms
* if you have the option between running a linear model with a transformed y variable or a glm, choose the linear model with a transformed y variable. only run a glm when you have to.

Provide all answers in R or R markdown (similar to the take home quiz 4). Use the following scripts to load the datasets. The dataset to be used for each question is provided in bold at the end of the question.

Dataset Please use the following scripts to load in the data from GitHub

flying = read.table(file="https://raw.githubusercontent.com/OscarFHC/NRE538\_2017Fall/master/Final/flying.csv",header=TRUE, sep=",")

college = read.table(file="https://raw.githubusercontent.com/OscarFHC/NRE538\_2017Fall/master/Final/college.csv",header=TRUE, sep=",")

happy = read.table(file="https://raw.githubusercontent.com/OscarFHC/NRE538\_2017Fall/master/Final/happy.csv",header=TRUE, sep=",")

cancer = read.table(file="https://raw.githubusercontent.com/OscarFHC/NRE538\_2017Fall/master/Final/cancer.csv",header=TRUE, sep=",")

1. Is there a significant association between gender (gender) and whether people think it’s rude to bring an unruly child on the plane (unruly\_child)? If yes, which gender tends to think that bringing an unruly child is more rude? **flying**

Test: Chi- square test of independence

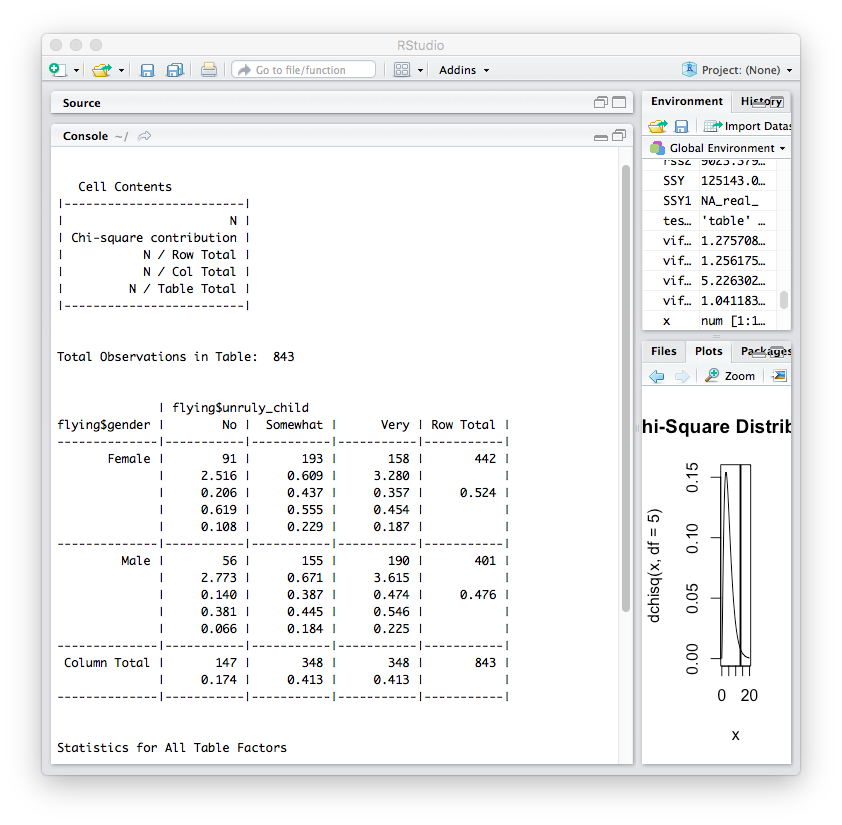
Assumptions:

1. Random sampling
2. Independent observations: Since the variables of interest are not dependent on each other (such as in the case of paired samples with before/ after’ values), I consider the observations to be independent.
3. No structural zeros: there are more than 10 observations under each variable of interest, with very few observations with no data. Therefore, I consider this assumption to be met.

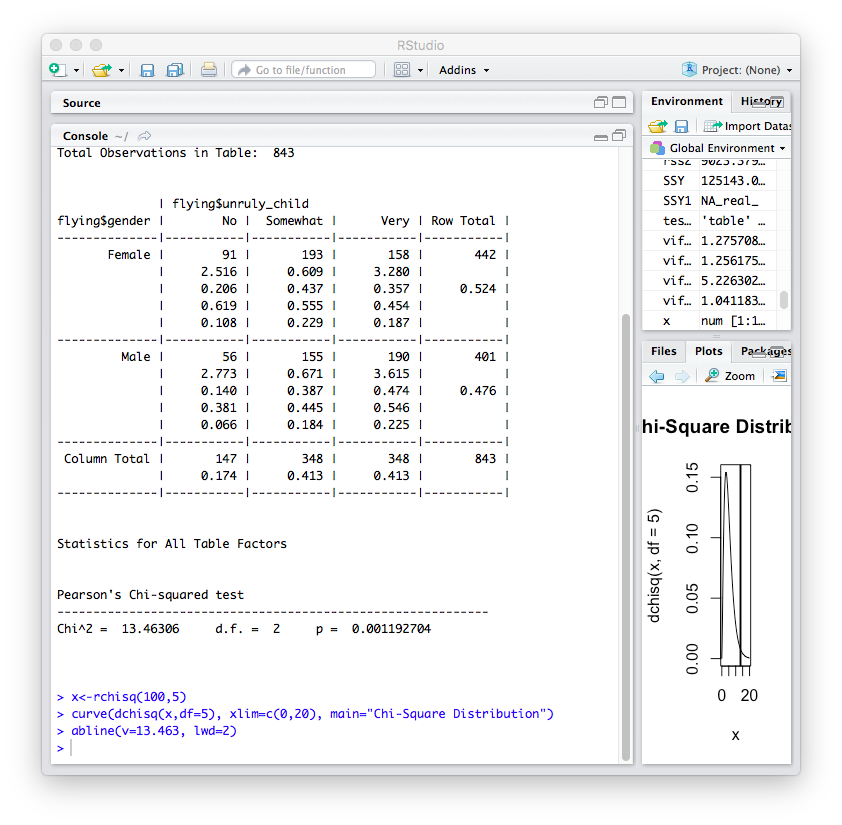
H0: there is no association between gender and whether people think its rude to bring an unruly child on a plane

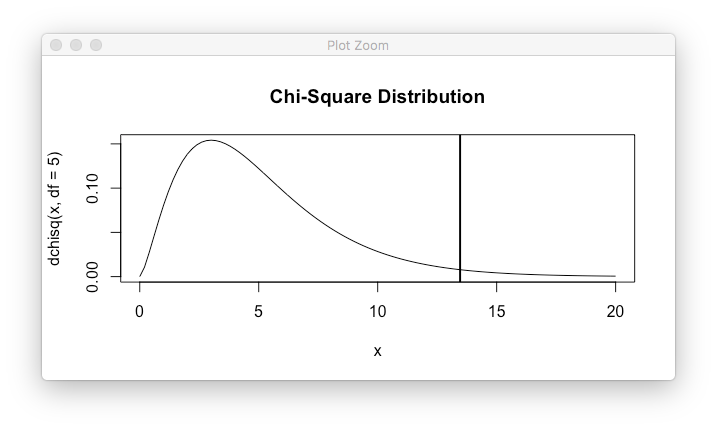
Ha: there is an association between gender and if people think its rude to bring an unruly child on the plane.

R output:



This value tells us that males think it’s ruder to bring an unruly child on a plane.





Interpretation of results:

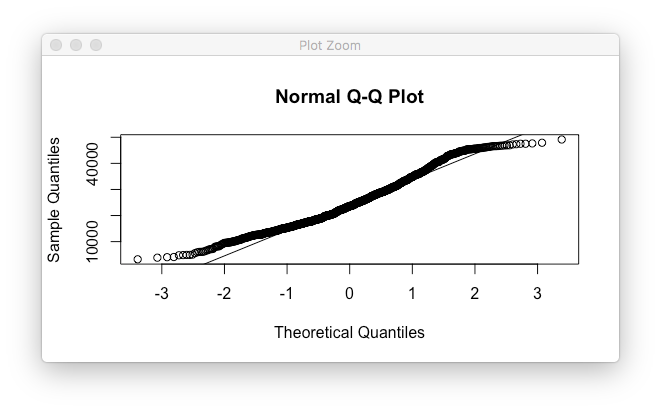
There is a significant association between gender and whether people think it’s rude to bring an unruly child on a plane (p value= 0.0011). According to the data, this association is stronger for the male gender.

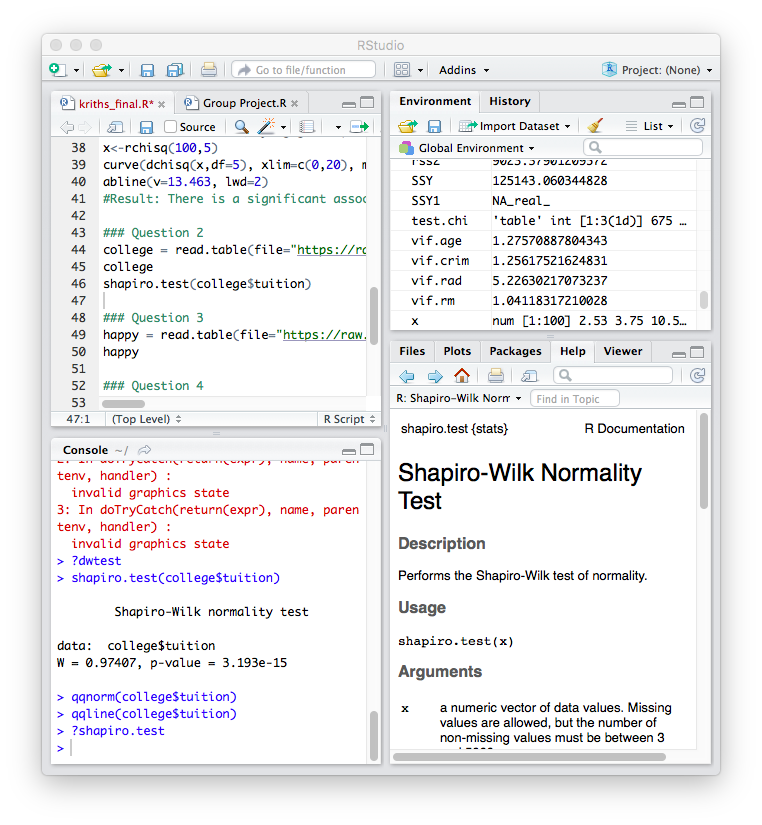
1. Is there a significant difference in tuition (tuition) by type of institution (type)? If yes, which type has a higher tuition? **college**

Test: Two sample t-test

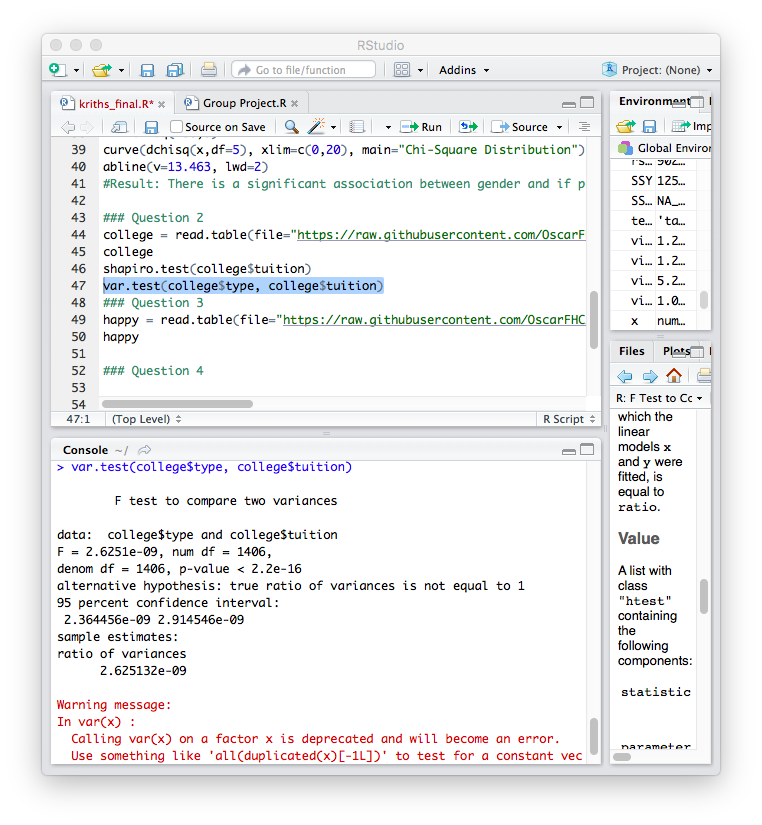
Assumptions:

1. Random sampling: can assume
2. Observations are independent: yes, as the variables measurements are not related to one another, as in the case of “regular interval” or “before/ after measurements”
3. Normal distribution/ large sample size: yes; Shapiro- Wilk test (p-value= 3.193e-15) and qq plot both show normal distributions.





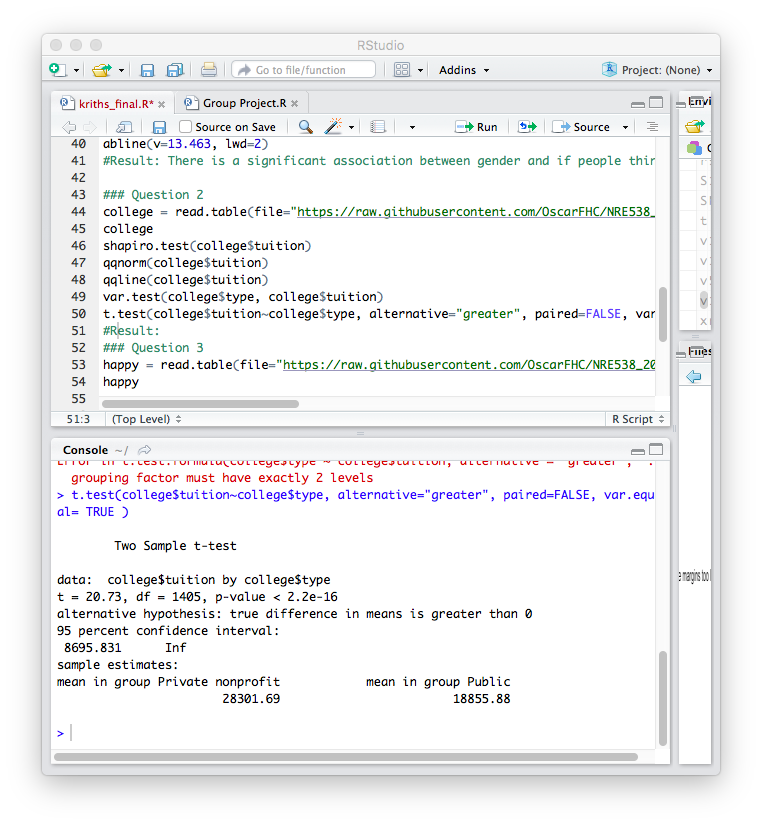
1. Equal variance between the two populations: F test for the variables indicate similar variances (p-value< 2.2e-16)



H0: there is no difference in tuition paid by type of institution

Ha: there is difference in tuition paid by type of institution.

R output:



Interpretation of results:

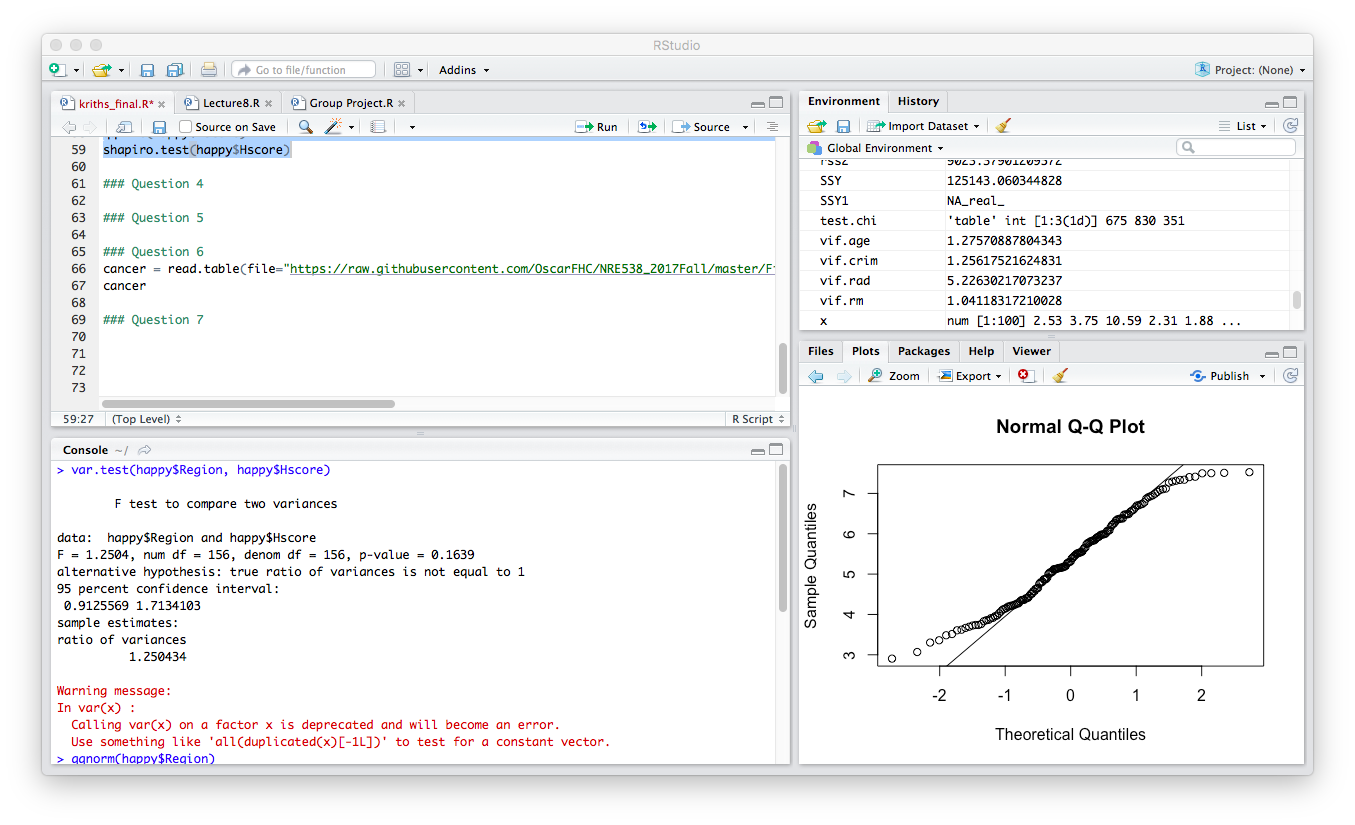
There is a difference in the tuition paid by the type of institution (p-value<2.2e-16). Average tuition in private non-profits is greater (28,301.69) than public institutions (18855.88)

1. Is there a significant difference in happiness (Hscore) by region (Region)? **happy**

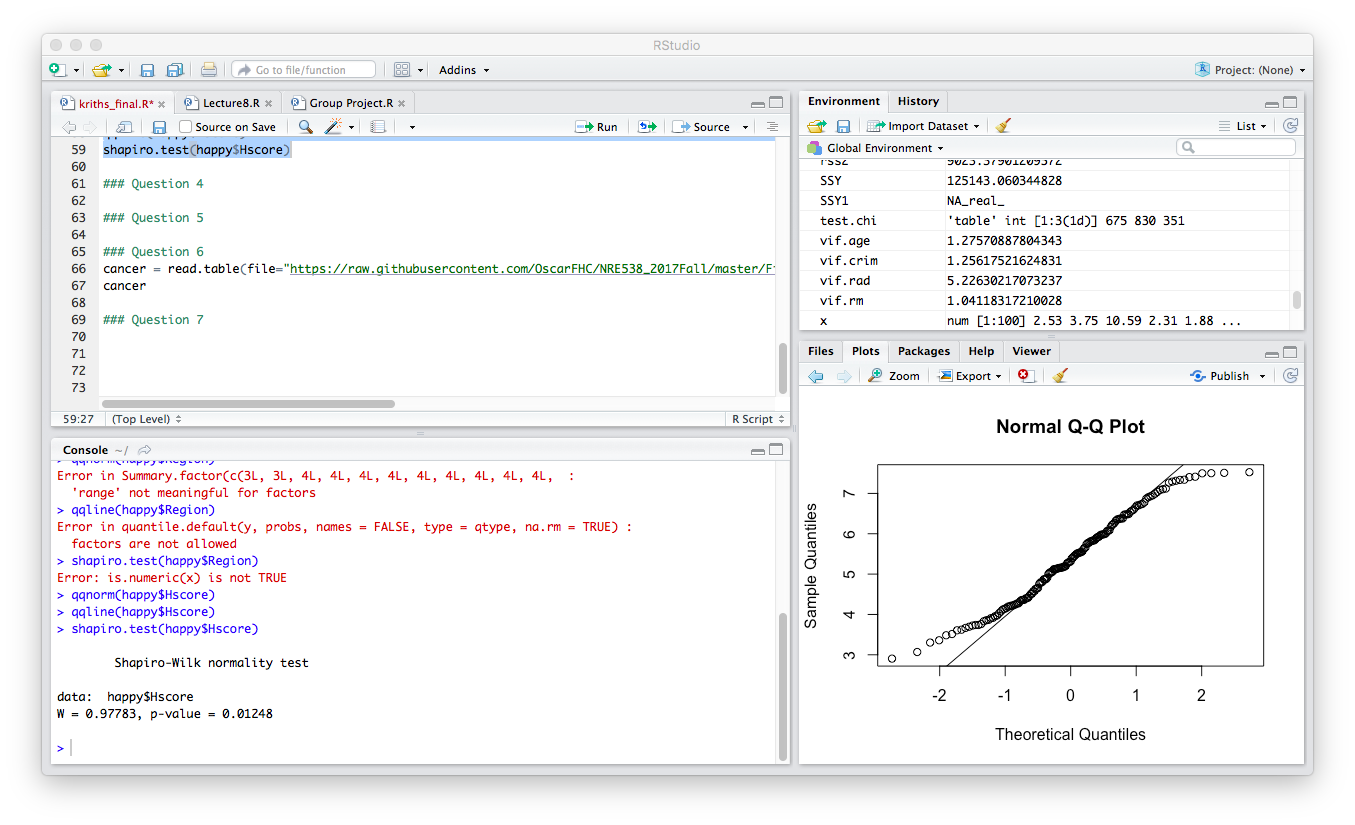
Test: ANOVA

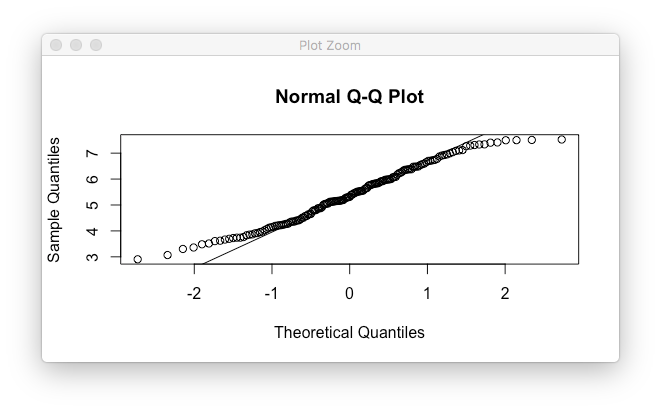
Assumptions:

1. Random sampling: can assume.
2. Independent observations: yes, as the variables measured do not influence each other.
3. Homogeneity of variances: no, as F-test is not significant (p-value= 0.1639).



1. Normal distribution of response variable: data is fairly normal (through Shapiro- Wilk test and qq plots).

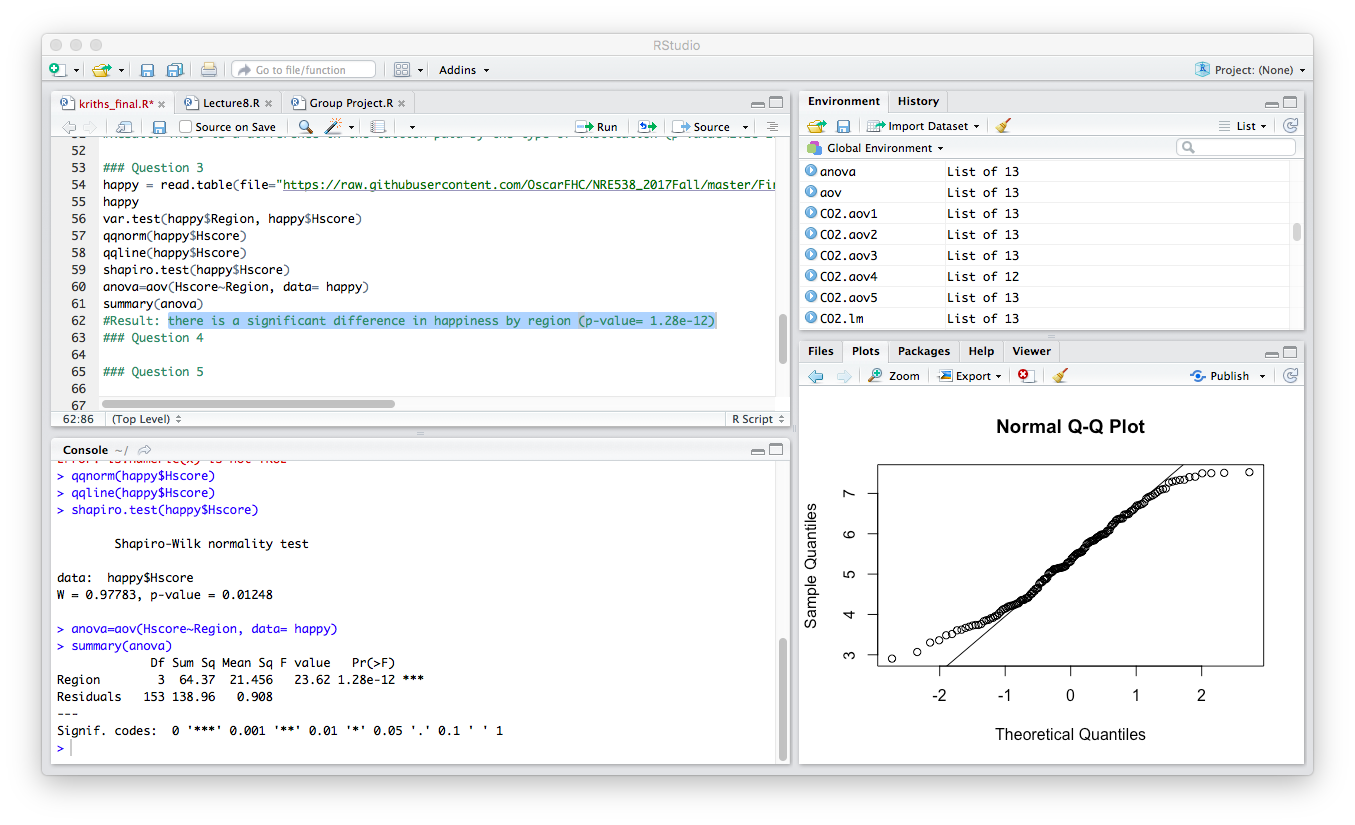




H0: there is no difference in happiness by region

Ha: there is a difference in happiness by region

R output:

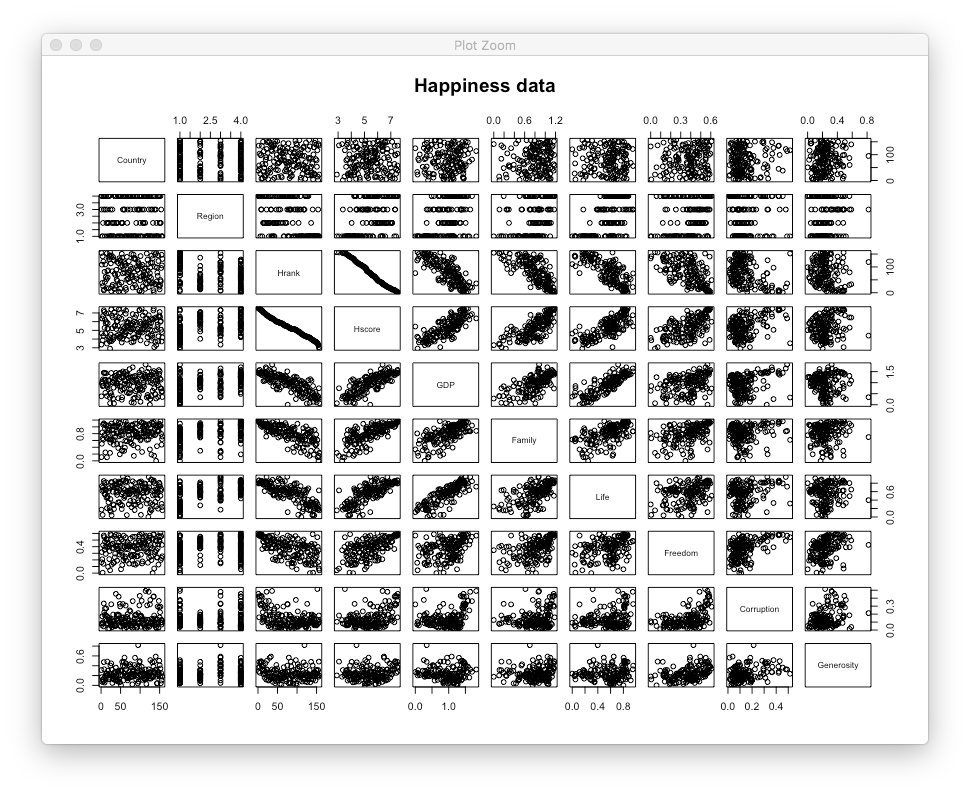


Interpretation of results: there is a significant difference in happiness by region (p-value= 1.28e-12)

1. What factors are significantly associated with a country’s corruption levels (Corruption)? Choose three continuous independent variables to include in your model. **Happy**

Using “pairs” and “cor.test” test, we find the following results:

|  |  |  |
| --- | --- | --- |
| Variable | Correlation values | p-values |
| Hrank | -0.3871016 | 5.492 e-7 |
| Hscore | 0.4020322 | 1.798e-07 |
| GDP | 0.2941848 | 0.0001842 |
| Family | 0.2135609 | 0.00724 |
| Life | 0.2495833 | 0.00162 |
| Freedom | 0.502054 | 2.1e-11 |
| Generosity | 0.3059299 | 9.759e-05 |



Thus, all variables except Country and region (as they are not continuous) have a significant correlation with corruption, although the degree of association varies. Freedom and Hscore have the highest correlation factors.

Independent variables chosen: Freedom, Hscore and GDP

1. Choose one of the continuous independent variables that was significant in the model for Question 4 and interact it with region (Region) to predict corruption (Corruption). This model should only include one continuous independent variable and its interaction with region. Does the influence of your continuous variable on corruption vary by region? If yes, how do you interpret the interaction? **happy**

Test: Univariate linear model

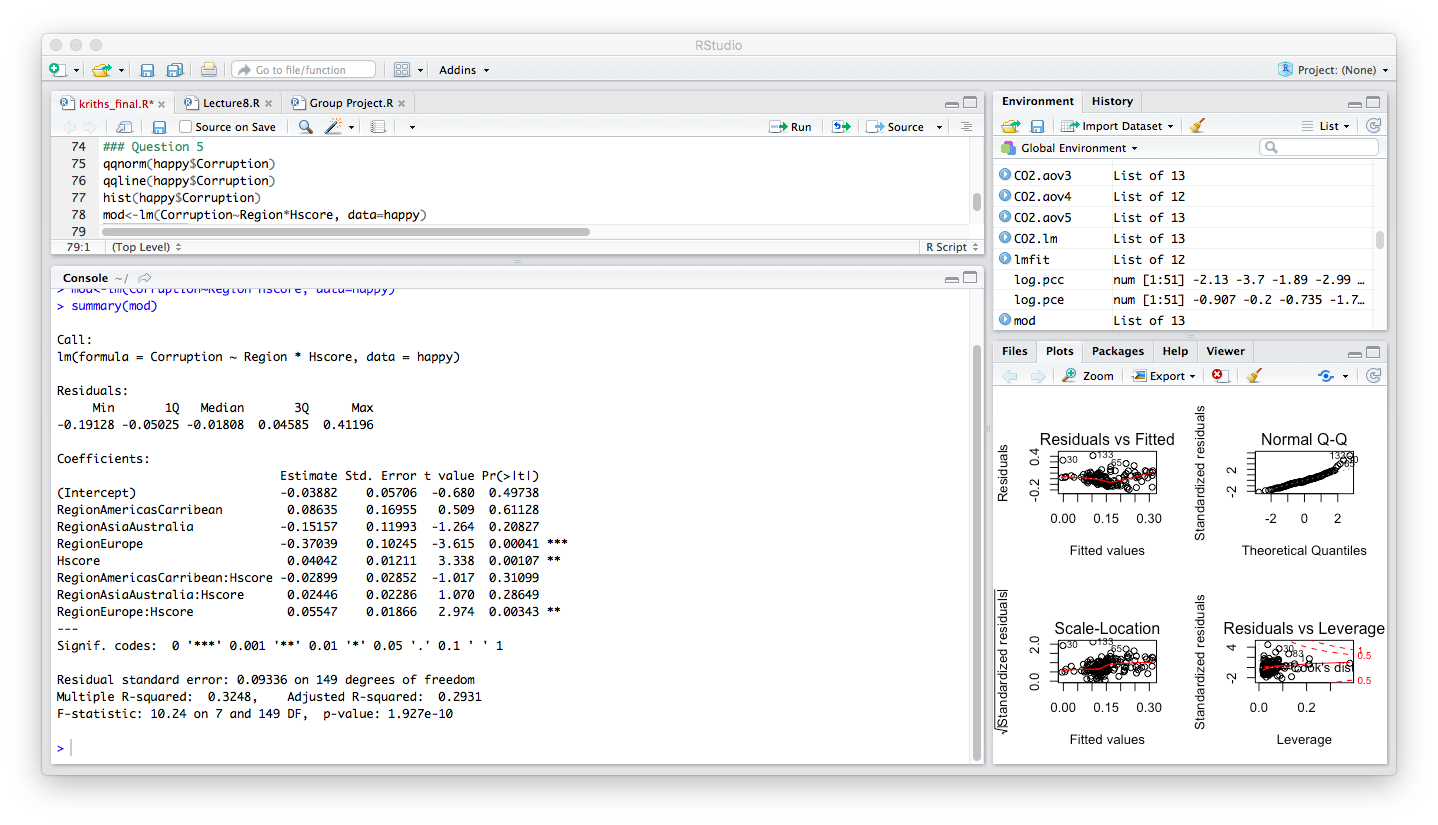
Assumptions:

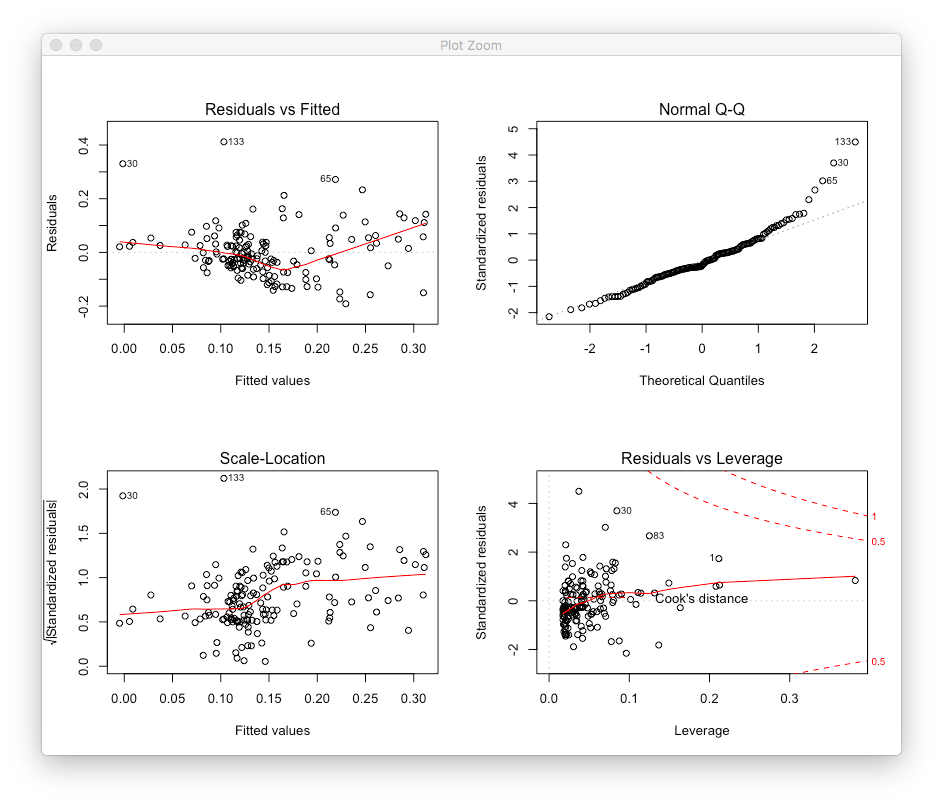
1. Random sampling: can be assumed
2. Normality of the distribution:
   1. qq plot and histogram show non- normal distributions
   2. log transformation shows more normal results than sqrt transformation. Thus, we incorporate this in our model.
3. Linear relationship between the variables of interest: can be assumed

H0: there is no association between corruption, and a region’s Hscore.

Ha: there is an association between corruption, and a region’s Hscore.

R output:





Interpretation of results:

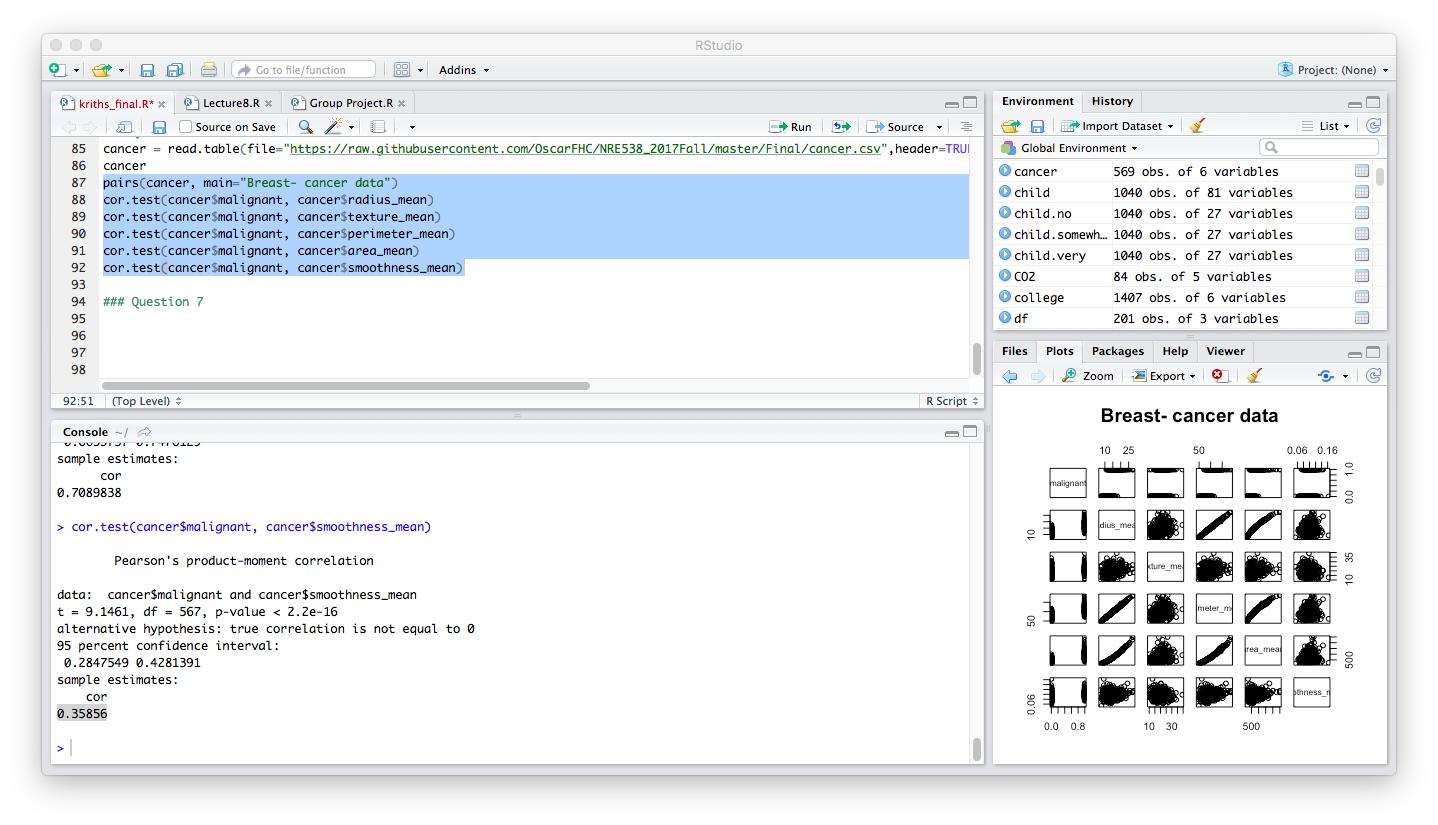
The linear model can be used to predict the log odds of the interaction between corruption and a region’s Hscore (R2= 0.3248, p-value= 1.927 e-10). The region- Hscore interaction term is significant only in Europe.

Thus, *there is a significant association between corruption and a region’s Hscore. However, as the residuals vs. fitted plot does not have a linear, uniform distribution, the model does not give the best fit for our research question. This may be rectified by using a different interaction term, by using a multivariate linear model, or by transforming the explanatory variables to provide a better fit.*

1. Which factors are significantly associated with whether a breast cancer tumor is malignant or not? Choose three continuous independent variables to include in your model. **cancer**

From using “pairs” and “cor.test” tests in R, we can interpret the following:

|  |  |  |
| --- | --- | --- |
| Variable | Correlation factor | P-value |
| Mean radius | 0.7300285 | < 2.2e-16 |
| Mean texture | 0.4151853 | < 2.2e-16 |
| Mean perimeter | 0.7426355 | < 2.2e-16 |
| Mean area | 0.7089838 | < 2.2e-16 |
| Mean smoothness | 0.35856 | < 2.2e-16 |



All factors can be significantly associated with tumor malignancy. Mean area, mean perimeter and mean radius have the highest correlation.

Variables chosen: Mean perimeter, mean texture and mean smoothness.

1. BONUS/EXTRA CREDIT: Which independent variables are the most important in explaining whether a breast cancer tumor is malignant or not? Use the same 3 continuous independent variables you chose for question 6. **cancer.**

Test: Generalized linear model

Assumptions:

1. Random sampling: can be assumed
2. Linear relationship between the variables of interest: can be assumed
3. Normality of data: a plot of the dependent variable (malignancy) shows a binomial distribution.

I assume a linear relationship between the dependent and independent variables and use ‘logit’ function to address the binomial nature of the dependent variable. To decide which variables are most important, I construct 3 different nested GLM models and compare them using anova and AIC values.

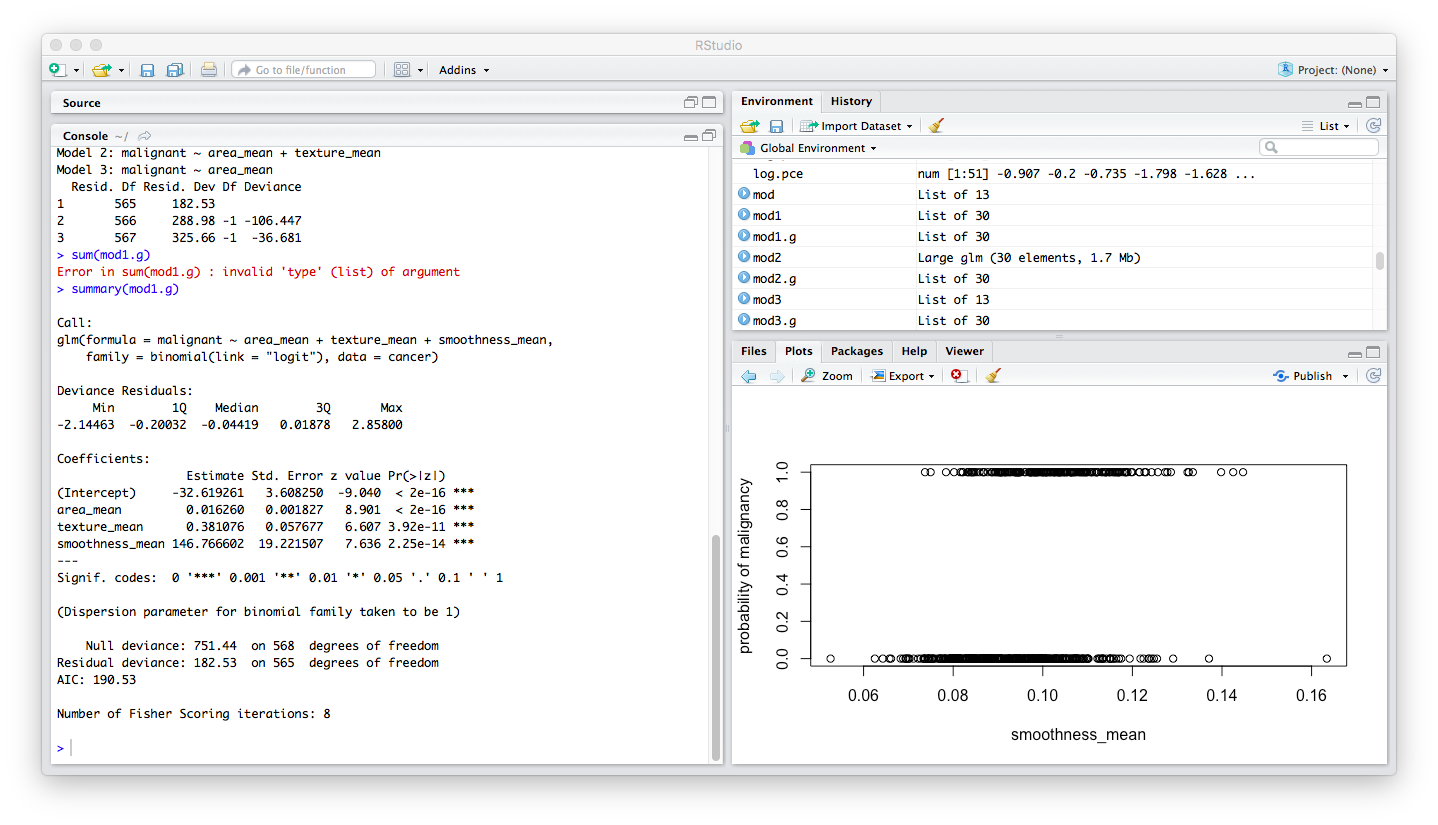
Models used:

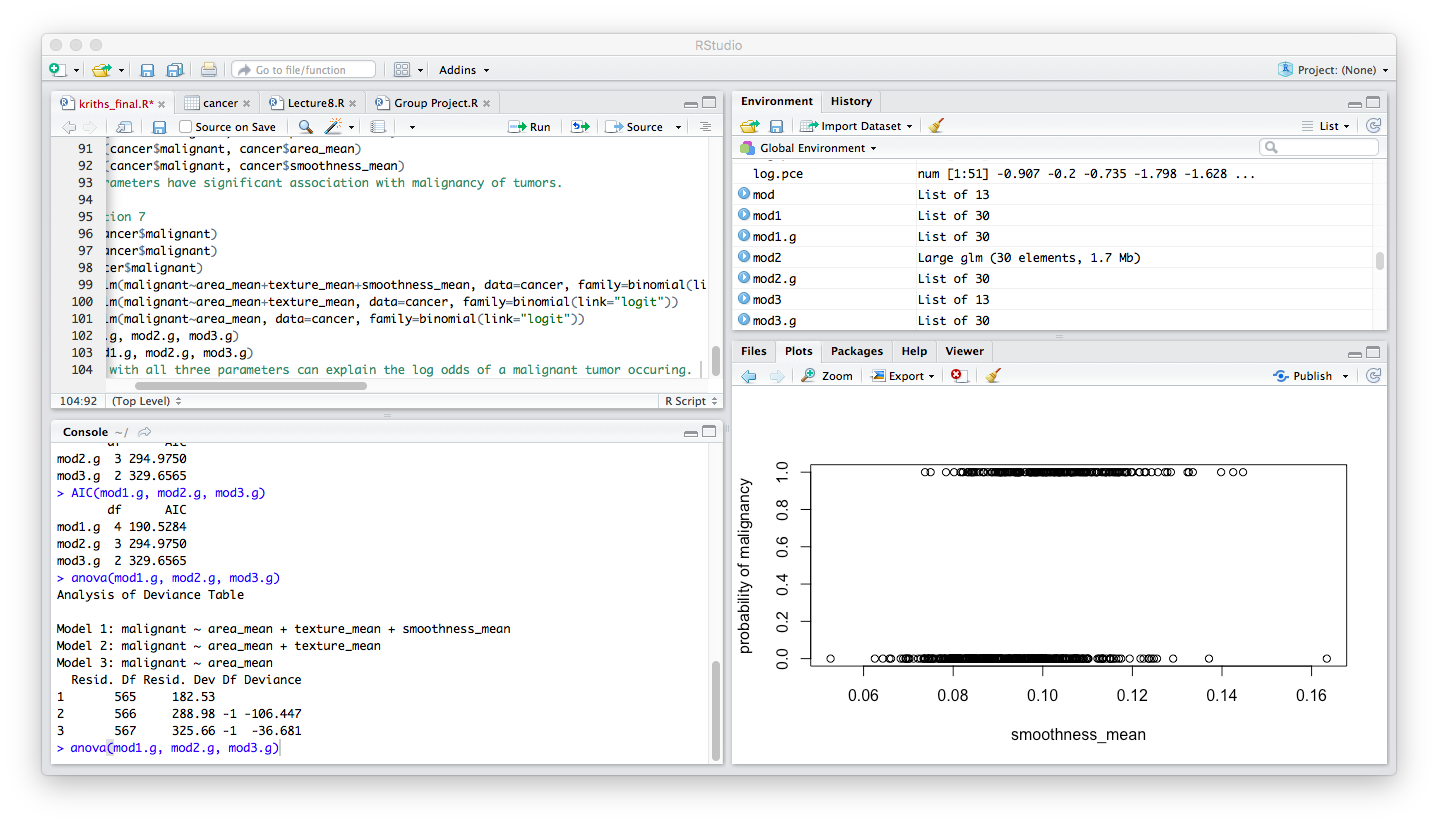
Model1: malignant~area\_mean+texture\_mean+smoothness\_mean

Model2: malignant~area\_mean+texture\_mean

Model3: malignant~area\_mean

R output:





Interpretation of results:

Model 1 has the lowest AIC values and explains the most deviance in the model. Further, a summary of the model 1 shows all 3 variables (area, texture and smoothness) to be equally significant.

Thus, mean area, mean texture and mean smoothness are the most important parameters in explaining the log odds of the occurrence of a malignant tumor.