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
* 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)? **flying**

A Chi-square test should be used due to both variables being categorical data. Assumptions include independent observations and no structural zeros. To remove the NA values, I subset the data. The results of the test yield a significant p-value, meaning that there is an association between gender and whether people think it's rude to bring an unruly child on the plane.

1. Is there a significant difference in tuition (tuition) by type of institution (type)? **college**

The IV is categorical and binary and DV is continuous so a t.test is an appropriate test to use. The assumptions for a t.test are continuous data, a randomly selected and independent sample, normal values or the sample size is sufficiently large, and equal variances. The assumption of normal distribution is not met given the output of the Shapiro test, but the sample size is sufficiently large so we will assume this is okay. The assumption of equal variances is also not met so Welch's t.test must be used. The results show that there is a significant difference in tuition by type of institution. The average tuition of private nonprofit institutions is higher than the average tuition of public institutions.

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

The IV is categorical, but contains more than two groups and the DV is continuous so a one-way ANOVA is the appropriate test to use. Assumptions for ANOVA are normal distribution, independent samples, and equal variances. The data does not have a normal distribution, but the sample size is sufficiently large, and variances are not equal. Results show that there is a significant difference in happiness by region.

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

The three continuous variables I chose were "Freedom" "GDP" and "Generosity" because these variables are not highly correlated with each other. Assumptions checked for the linear model are a linear relationship between variables, homoscedasticity, independence of errors, and normality of error distribution. The DW test shows a linear relationship. The BP test shows homoscedasticity. The model does not pass the Shapiro test for normality, but the qq plot looks normal so we will assume normality. Results show that Freedom is the variable most significantly associated with Corruption. GDP and Generosity have a weaker relationship with Corruption, but still significant at the .05 level.

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? **happy**

Assumptions for the ANCOVA are the same as for the linear model, linear relationship between variables, homoscedasticity, independence of errors, and normality of error distribution. The DW test shows a linear relationship. The BP test shows homoscedasticity. The model does not pass the Shapiro test for normality, but the qq plot looks normal so we will assume normality. Results show that the influence of freedom on corruption does vary by region.

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**

The assumptions checked for the linear model are a linear relationship between variables, homoscedasticity, independence of errors, and normality of error distribution. The DW test shows a linear relationship. The BP test shows homoscedasticity. The model does not pass the Shapiro test, but the qq plot looks normal so we will assume normality. The error distribution is not normal so we will try a log transform of the dependent variable. All of the three factors "radius\_mean", "texture\_mean", and "perimeter\_mean" are significantly associated with whether a breast cancer tumor is malignant or not.

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.**

"Perimeter\_mean" is the most important in explaining whether a breast cancer tumor is malignant or not, closely followed by "radius\_mean", "texture\_mean" is the least important of these three variables.