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**

As this question is comparing two categorical variables I will use a chi-square test. I am assuming that no person was surveyed twice so I can assume independence of observations. After making a chi-square table I could see that all cells were filled and that each cell had more than 5. After running the chi-square test I found a p-value of 0.001193 which means that there is a significant association between gender and the perception that bringing an unruly child on a plane is rude. Checking the table I find that it is men who are more likely to find unruly children rude.

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

I am assuming that the data came from a random sample. Based on the nature of the data I can also assume that the observations are independent (I don’t think they surveyed the same school multiple times). The tuition data are continuous. The sample size is 1407 so I can assume normality due to the CLT. Upon running an F-test though I find a p-value of 2.2e-16 which means that I reject the null hypothesis that the variances are equal. As such I run a Welch’s t-test for unequal variances. After running the Welch’s t-test I find a p-value of 2.2e-16 meaning that there is a significant difference in tuition by type of institution. After plotting a boxplot of the two groups I find that Private nonprofit institutions have a higher tuition than Public institutions.

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

As there are four regions I decided to do an ANOVA. As no country in the sample is repeated I can assume that each observation is independent. A qq-plot of Hscore looks just about normal but since the sample size is 157 I can assume normality under the CLT. Each region has more than 15 samples represented. I ran Levene’s Test to test for equal variance and got a p=value of 0.5179 which means that I can assume equal variances. After running the ANOVA I found a p-value of 1.28e-12 which indicates that there is a significant difference in happiness by region. To investigate further I ran a Tukey’s HSD Post Hoc test and found that the Americas-Caribbean and Europe regions had the highest happiness scores (though they were not significantly different from each other). Then comes Asia-Australia and last is Africa-Mideast.

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

As I will be checking for associations between three continuous independent variable and one continuous dependent variable I will be running a multiple linear regression. To begin I checked Corruption for normality using a qq-plot. It was definitely not normally distributed so I tried a log transform, square root transform and a cube root transform. After comparing the qq-plots and histograms of all three transforms I decided that the cube root transform was the closest to a normal distribution. The three dependent variables I chose to test were Freedom, Generosity, and GDP. I tested a correlation coefficients and found none over 0.362 so I can assume there will be no problems with multiple collinearity. I also ran a VIF and found the following values: Freedom = 1.367952, Generosity = 1.189185, and GDP = 1.189712. These VIF values show that they are not particularly correlated. Running the model I find that Freedom is the only dependent variable that is significantly correlated with Corruption. To check for autocorrelation I ran a Durbin-Watson test and found a p-value of 0.1932 which means that autocorrelation is not an issue here. To check for homoscedasticity I ran a studentized Breusch-Pagan test and found a p-value of 0.04066 which means that the residuals are heteroscedastic which is a problem with the model. To check for normality of residuals I made a qq-plot and ran a Shapiro-Wilk normality test. With a p-value of 0.7654 I found that the residuals are normally distributed. The model itself has an adjusted R2 of 0.2341 which means that it only explains 23.4% of the variation in Corruption.

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**

As this model will be comparing a continuous independent variable and a categorical independent variable with a continuous dependent variable, I will run an ANCOVA. I have kept Corruption in its cube-root transformed state to maintain its normality. I will be examining the interaction between Freedom and Region. Upon running the ANCOVA I find that there is a significant interaction between Freedom and two regions, namely Africa-Mideast and Europe, to predict the cube root of corruption. For Africa-Mideast I find that adding an interaction with Freedom produces a beta coefficient of 0.31375 which is actually a weaker association with the cube root of corruption than the Africa-Mideast region on its own (0.40617). For Europe the interaction produces a beta coefficient of 0.42202 which means that it has an association with the cube root of corruption of 0.42202 + 0.31375 = 0.73577 which means that as freedom in Europe increases by 1 the cube root of corruption increases by 0.73577 a stronger effect than Africa-Mideast.

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**

Since I am comparing a categorical dependent variable that is binomial with three continuous independent variables I chose to run a glm. I checked for issues of multicollinearity by getting correlation and VIF values and none of the variables I chose are highly correlated. I ran my first model with radius\_mean, texture\_mean, and smoothness\_mean. After running the glm I find that all three variables are significantly associated with whether a breast cancer tumor is malignant or not. My model has an AIC score of 195.29. I then ran two other glms but replaced radius\_mean with perimeter\_mean in one and then area\_mean in the other. The glm with area\_mean returned an AIC score of 190.53 but the glm with perimeter\_mean returned an AIC of 189.68. As such the glm with perimeter\_mean is the best fit model and the one I ended up using. Perimeter\_mean is 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.**

Based on the glm I ran with perimeter\_mean, texture\_mean, and smoothness\_mean, the independent variable that had the largest beta coefficient was smoothness\_mean. This means that smoothness\_mean increases the log odds ratio of whether the tumor will be malignant by 134.02497 which is a much larger impact than any of the other variables.