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**

Because both the independent and dependent variables are categorical, we choose a chi-square test for independence. We assume that the samples are random and observations are independent. After constructing the contingency table, we see no structural zeros and that all cells are filled >5. Our p-value is .0012, which is <.05, so we can reject the null hypothesis that there is no significant association. Based on the contingency table, we see that the male gender tends to think that bringing an unruly child is more rude.

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

Because the independent variable is categorical (and there are only two levels of that category) and dependent variable is continuous, we use a two-sample t-test (because we’re comparing two populations). The data is not quite normal, but we can relax this assumption because our sample size is >30. We assume that the samples are independent of each other, and the sample is randomly selected. Variance is not equal, but we can get around this by using the Welch’s t-test. Our p-value is much less than .05, so we can reject the null that there is no significant difference in tuition by type of institution. Based on a boxplot of the data, we see that the private nonprofit has a higher tuition.

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

Because Hscore is continuous and Region is categorical, a one-way ANOVA is the appropriate test for this question. The data are not normal, but we can proceed anyway because we have >30 samples with >15 samples per group. We assume that the samples are independent of one another, but are violating the assumption of equal variance. Our p-value is <.05, so we can reject the null that there is no 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**

To assess which variables are associated with corruption levels, we build a multiple linear regression of Corruption ~ Freedom + GDP + Life. We achieve normality of the dependent variable by log-transforming Corruption. After running the linear model, we observe independence of the errors and homoscedasticity. However, the residuals fail the Shapiro-Wilk test. The qqplot looks somewhat decent, so we’ll proceed with using this linear model anyway. Based on the model’s results, Freedom is the only variably significantly associated with a country’s corruption levels (or, in this case, the log of the country’s corruption levels). The p-values for GDP and Life are both very high, indicating that they are not significant factors. However, because GDP and Life are highly correlated, this may be due to multi-collinearity.

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**

Yes, the influence of Freedom does vary by Region. This is indicated by the interaction terms in the output lines describing Freedom:[Region], with multiple regions listed. The only significant interaction is for Freedom:RegionEurope, meaning that the influence of Freedom remains constant (i.e., slope remains the same) for the other regions. For Europe, the significant interaction term indicates that the slope is different for this region compared to the other regions. Because the coefficient is positive, it means that the interaction is amplified with increasing freedom.

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**

Because you have a categorical dependent variable, and continuous independent variables, conduct a logistic regression (i.e., a GLM using the binomial family and a logit function). This GLM will use the continuous independent variables radius\_mean, texture\_mean, and smoothness\_mean, as these are not highly correlated with each other. All three have p-values <.05, meaning that they are all significantly associated with the log odds of 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 an AIC comparison of three different GLMS, one for each variable, the tumor’s radius is most important in explaining whether a breast cancer tumor is malignant or not.