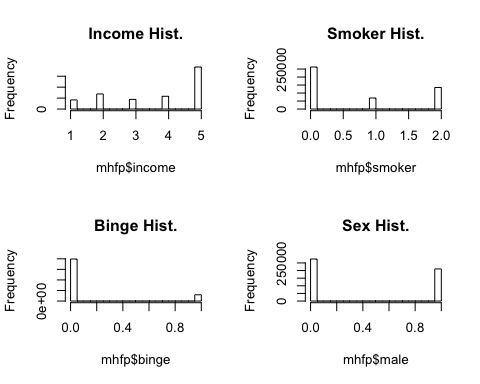
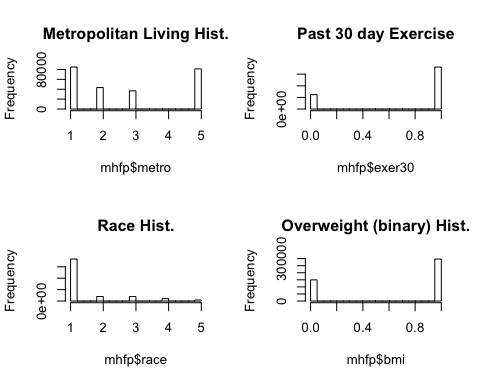
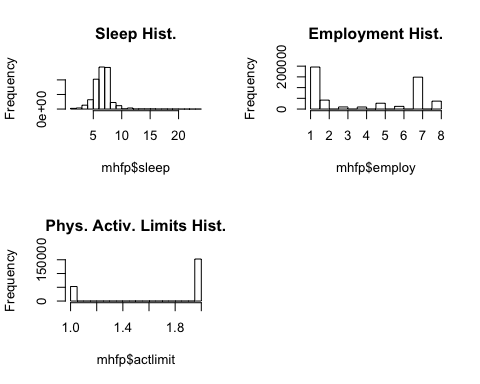
Final Project Outline

Coco Kusiak, Joshua Freeman, and Luke Toomey

11/14/2017

Cleaning the dataset for analyses.

Basic histograms of the data and numbers (n) per variable levels including missingness. Please note we had issues graphing the historgrams for sexorient and transgender variables. However, (n) per level of the variable is available in the tables. 

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 1.000 6.000 7.000 7.054 8.000 24.000 5726

##   
## 1 2 3 4 5 6 7 8 <NA>   
## 195970 41648 9828 9560 27106 12493 148750 36909 4039

##   
## bisexual gay other straight <NA>   
## 3433 3057 902 192445 286466

##   
## 0 1 <NA>   
## 202307 786 283210

##   
## 1 2 <NA>   
## 52769 152698 280836

##   
## 1 2 3 5 <NA>   
## 84893 43390 36904 81088 240028

##   
## 0 1 <NA>   
## 123860 361649 794

##   
## 1 2 3 4 5 <NA>   
## 368048 39555 39224 21351 9442 8683

##   
## 0 1 <NA>   
## 149640 297047 39616

##   
## 1 2 3 4 5 <NA>   
## 41693 68856 44076 58349 192028 81301

##   
## 0 1 2 <NA>   
## 262670 68942 134863 19828

##   
## 0 1 <NA>   
## 397997 59205 29101

##   
## 0 1 <NA>   
## 275631 210606 66

# The Data

-A subset of BRFSS data

# The Variables

## The Outcome

QLMENTL2: Felt depression symptoms in the last 30 days

## The Covariates

-binge: binary for binge drinking in last 30 days -exer30: binary for exercising in the last 30 days, 794 missing -metro: 4 levels of metropolitan status, 240,000 missing -race: 5 level race variable -income: 5 levels of income, 81,000 missing -employ: 9 levels employment status, 73 missing -sex: gender -sexorient: sexual orientation, 280,000 missing -trans: 6 level, 281,000 missing -smoker: 4 level smoking status, 19,000 missing -sleep: continuous number of hours of sleep within 24 hour period, 2 missing -bmi: Overweight or obese calculated variable -actlimit: Activity Limitation Due to Health Problems

# Individual Project: Josh

I plan on analyzing the BRFSS data using complete case analysis and MCMC multiple imputation analysis using the MICE R package (<https://cran.r-project.org/web/packages/mice/mice.pdf>). In order to examine possible differences in findings due to missingness assuming Missing at Random or Missing Completely at Random, imputation will be carried out imputing multiple datasets and using a weighted dataset based on all imputed sub-datasets. This will allow inferences about the population distribution within the dataset and will conserve power in the multiple imputation analysis.

# Individual Project: Coco

I plan on comparing LASSO, ridge regression, and step-wise regression methods for model selection. For LASSO and ridge regression, I plan on using the glmnet package. For step-wise selection, I'll use the stepAIC() function in the MASS package. I plan on comparing the performance between these three models using 10-fold cross validation.

# Individual Project: Luke

I plan to fit a general linear mixed multilevel model with a hierarchical random term to the data. I will consider this as a two level model with with county at level one and state at level two, as the random effects and other covariates listed above as the fixed effects. This will be done using the lme4 package. Diagnostics, including covariate selection will then be done using the MMS package for fixed effects selection in linear mixed models, followed by inference. We will compare the GLM model with the GLMM model.