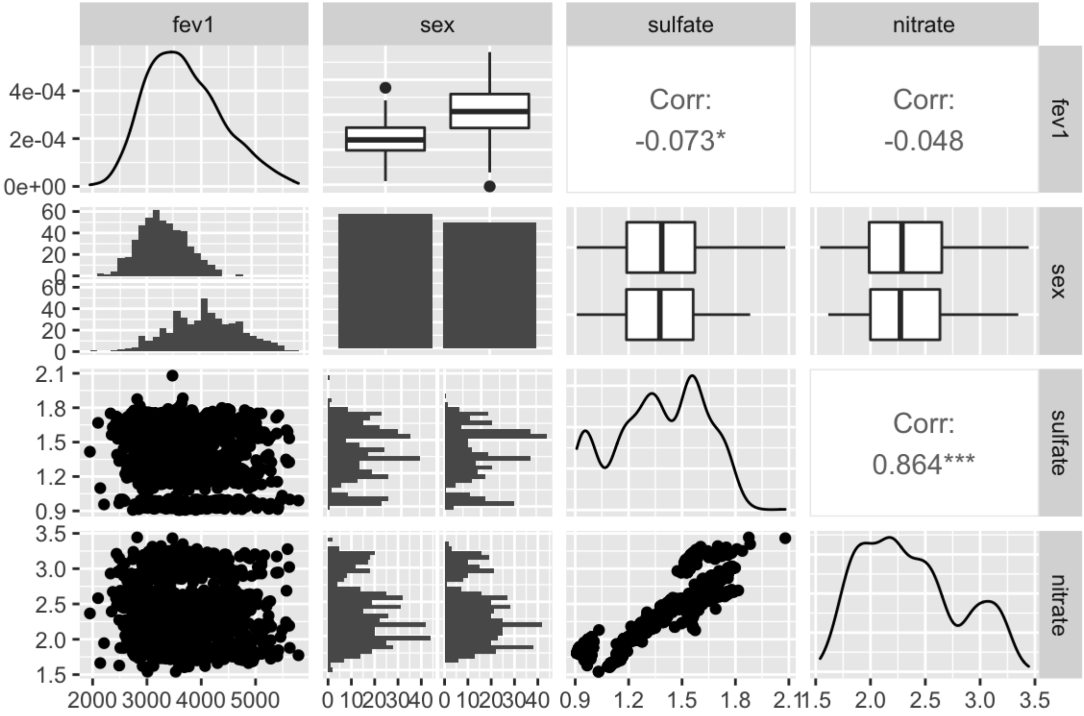
Project tips

Juan Pablo Lewinger

6/28/2021

Scatterplot matrix

```
library(ggplot2)
library(GGally)
setwd("~/LA's best")
chs = read.csv('CHS_cohortE_final_subset.csv')
chs$sex = factor(chs$male, levels=c(0,1), labels=c('F', 'M'))
ggpairs(chs[, c('fev1', 'sex', 'sulfate', 'nitrate')])
```



Project

- Compute descriptive statistics first, particularly graphs
- ggpairs does a great job with both univariate and bivariate plots
- Regression models
- Adjust for demographic and personal characteristics: age, male, race, height, and bmi. Include these variables regardless of significance.
- Run simple models first (i.e. outcome against each risk factor adjusting for variables above)
- OK to include race even if low counts for some categories but probably not ok to stratify by race
- Adjust by parental education (educ), Home age (HomeBuilt), gas stove in residence, (BaseGasstove), pet in residence (BasePets), and second hand smoke (ETS_base) if significant - Explore the association between outcomes and pollution variables (one at a time and together)
- Use anova() to assess significance of categorical variables with more than 2 levels

Anova example

```
fev1_model <- lm(fev1 ~  bmi + race, data=chs)
summary(fev1_model)
```

```
## 
## Call:
## lm(formula = fev1 ~ bmi + race, data = chs)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1602.84  -502.67   -84.31   444.01  2013.24 
## 
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             2719.240    184.479  14.740  < 2e-16 ***
## bmi                       23.221      4.363   5.322 1.27e-07 ***
## raceAsian                266.442    182.528   1.460 0.144685    
## raceCaucasian            456.086    155.801   2.927 0.003498 ** 
## raceMixed                434.195    163.612   2.654 0.008088 ** 
## raceOthers               396.927    158.854   2.499 0.012629 *  
## raceUnknown or Missing   550.968    163.419   3.372 0.000777 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 680.7 on 976 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared:  0.04366,    Adjusted R-squared:  0.03778 
## F-statistic: 7.425 on 6 and 976 DF,  p-value: 8.654e-08
```

Anova example

```
anova(fev1_model)
```

```
## Analysis of Variance Table
##
## Response: fev1
##            Df    Sum Sq  Mean Sq F value    Pr(>F)
## bmi         1  13274682 13274682 28.6490 1.082e-07 ***
## race        5   7369032  1473806  3.1807   0.00744 **
## Residuals 976 452235022   463356
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```