Kaitlin's EDA

Kaitlin Maciejewski

12/1/2018

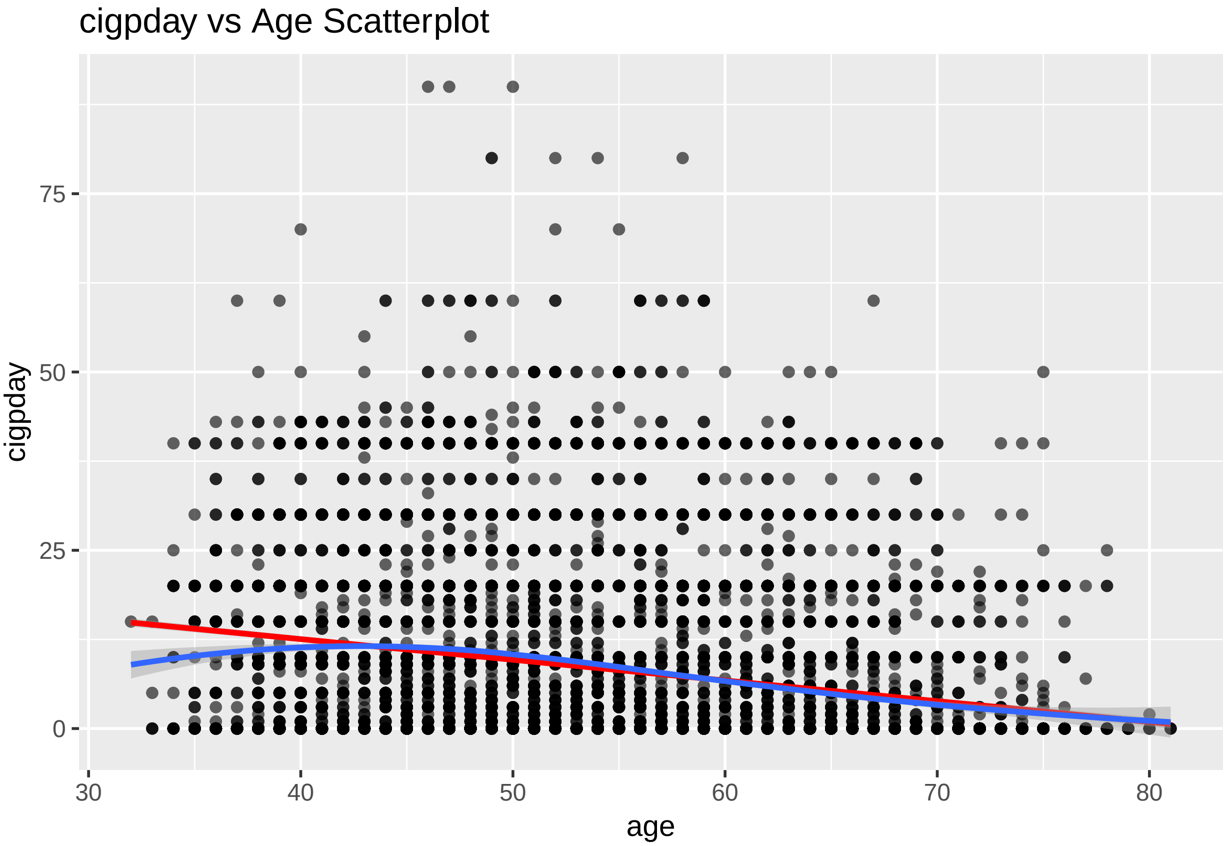
Exploratory

```
# save data in separate folder "final_data" one directory up
library(dplyr)
data <- read.csv("../final_data/frmgham2.csv") %>% janitor::clean_names()
attach(data)

library(psych)
knitr::kable(describe(data)[,c(2,3,4,5,8,9,10,13)], digits = 3)
```

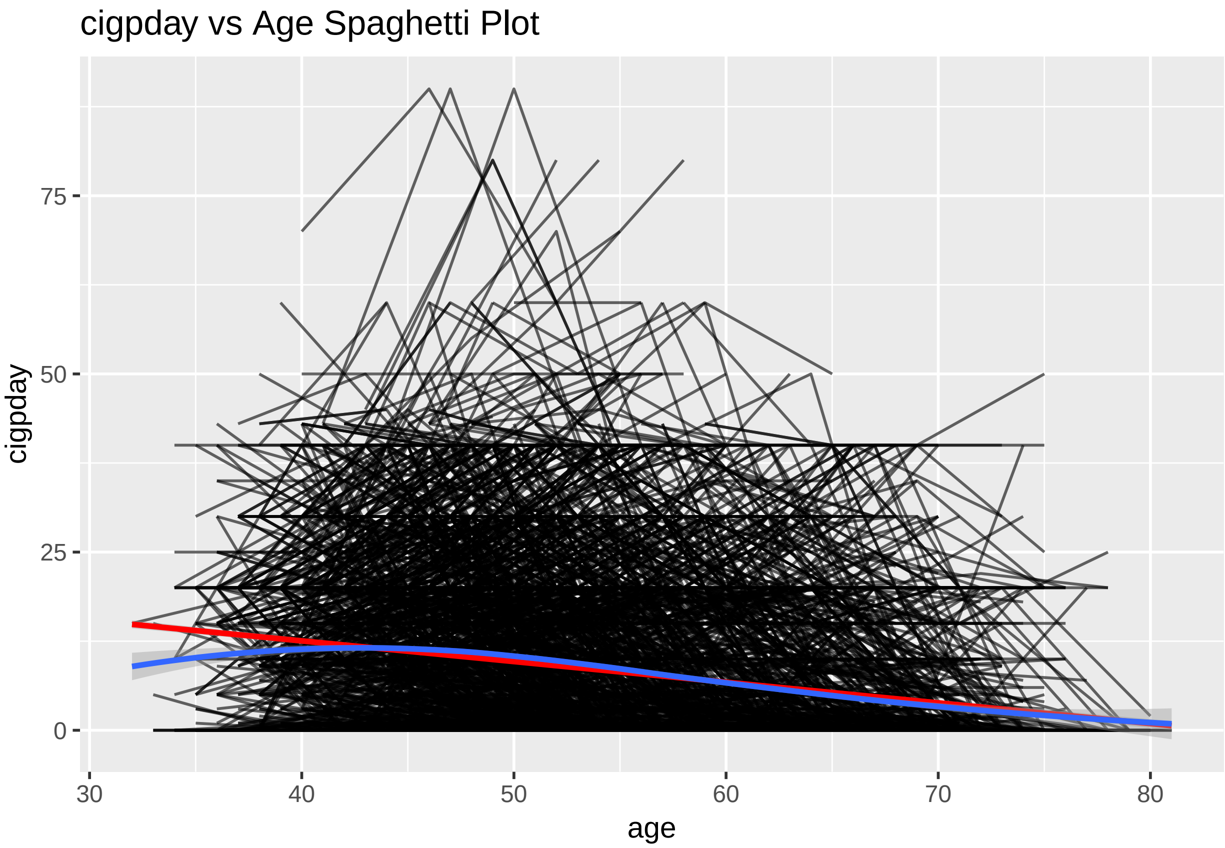
	n	mean	sd	median	min	max	range	se
randid	11627	5004740.917	2900877.440	5006008.00	2448.00	9999312.0	9996864.00	26902.680
sex	11627	1.568	0.495	2.00	1.00	2.0	1.00	0.005
totchol	11218	241.162	45.368	238.00	107.00	696.0	589.00	0.428
age	11627	54.793	9.564	54.00	32.00	81.0	49.00	0.089
sysbp	11627	136.324	22.799	132.00	83.50	295.0	211.50	0.211
diabp	11627	83.038	11.660	82.00	30.00	150.0	120.00	0.108
cursmoke	11627	0.433	0.495	0.00	0.00	1.0	1.00	0.005
cigpday	11548	8.250	12.187	0.00	0.00	90.0	90.00	0.113
bmi	11575	25.877	4.103	25.48	14.43	56.8	42.37	0.038
diabetes	11627	0.046	0.209	0.00	0.00	1.0	1.00	0.002
bpmeds	11034	0.086	0.280	0.00	0.00	1.0	1.00	0.003
heartrte	11621	76.782	12.463	75.00	37.00	220.0	183.00	0.116
glucose	10187	84.125	24.994	80.00	39.00	478.0	439.00	0.248
educ	11332	1.990	1.027	2.00	1.00	4.0	3.00	0.010
prevchd	11627	0.072	0.259	0.00	0.00	1.0	1.00	0.002
prevap	11627	0.054	0.226	0.00	0.00	1.0	1.00	0.002
prevmi	11627	0.032	0.176	0.00	0.00	1.0	1.00	0.002
prevstrk	11627	0.013	0.114	0.00	0.00	1.0	1.00	0.001
prevhyp	11627	0.460	0.498	0.00	0.00	1.0	1.00	0.005
time	11627	1957.019	1758.777	2156.00	0.00	4854.0	4854.00	16.311
period	11627	1.899	0.807	2.00	1.00	3.0	2.00	0.007
hdlc	3027	49.365	15.627	48.00	10.00	189.0	179.00	0.284
ldlc	3026	176.467	46.863	173.00	20.00	565.0	545.00	0.852
death	11627	0.303	0.460	0.00	0.00	1.0	1.00	0.004
angina	11627	0.164	0.370	0.00	0.00	1.0	1.00	0.003
hospmi	11627	0.099	0.299	0.00	0.00	1.0	1.00	0.003
mi_fchd	11627	0.154	0.361	0.00	0.00	1.0	1.00	0.003
anychd	11627	0.272	0.445	0.00	0.00	1.0	1.00	0.004
stroke	11627	0.091	0.288	0.00	0.00	1.0	1.00	0.003
cvd	11627	0.249	0.433	0.00	0.00	1.0	1.00	0.004
hyperten	11627	0.743	0.437	1.00	0.00	1.0	1.00	0.004
timeap	11627	7241.557	2477.780	8766.00	0.00	8766.0	8766.00	22.979
timemi	11627	7593.847	2136.730	8766.00	0.00	8766.0	8766.00	19.816
timemifc	11627	7543.037	2192.120	8766.00	0.00	8766.0	8766.00	20.330

	n	mean	sd	median	min	max	range	se
timechd	11627	7008.154	2641.345	8766.00	0.00	8766.0	8766.00	24.496
timestrk	11627	7660.880	2011.077	8766.00	0.00	8766.0	8766.00	18.651
timecvd	11627	7166.083	2541.668	8766.00	0.00	8766.0	8766.00	23.571
timedth	11627	7854.103	1788.370	8766.00	26.00	8766.0	8740.00	16.585
timehyp	11627	3598.956	3464.165	2429.00	0.00	8766.0	8766.00	32.127

```
# spaghetti
# lorellogram
#

library(ggplot2)

ggplot(data= data, aes(age, cigpday)) +
  geom_point(alpha = .6) +
  geom_smooth(method = 'lm', col = 'red') +
  geom_smooth(method = 'loess') +
  ggtitle("cigpday vs Age Scatterplot")
```

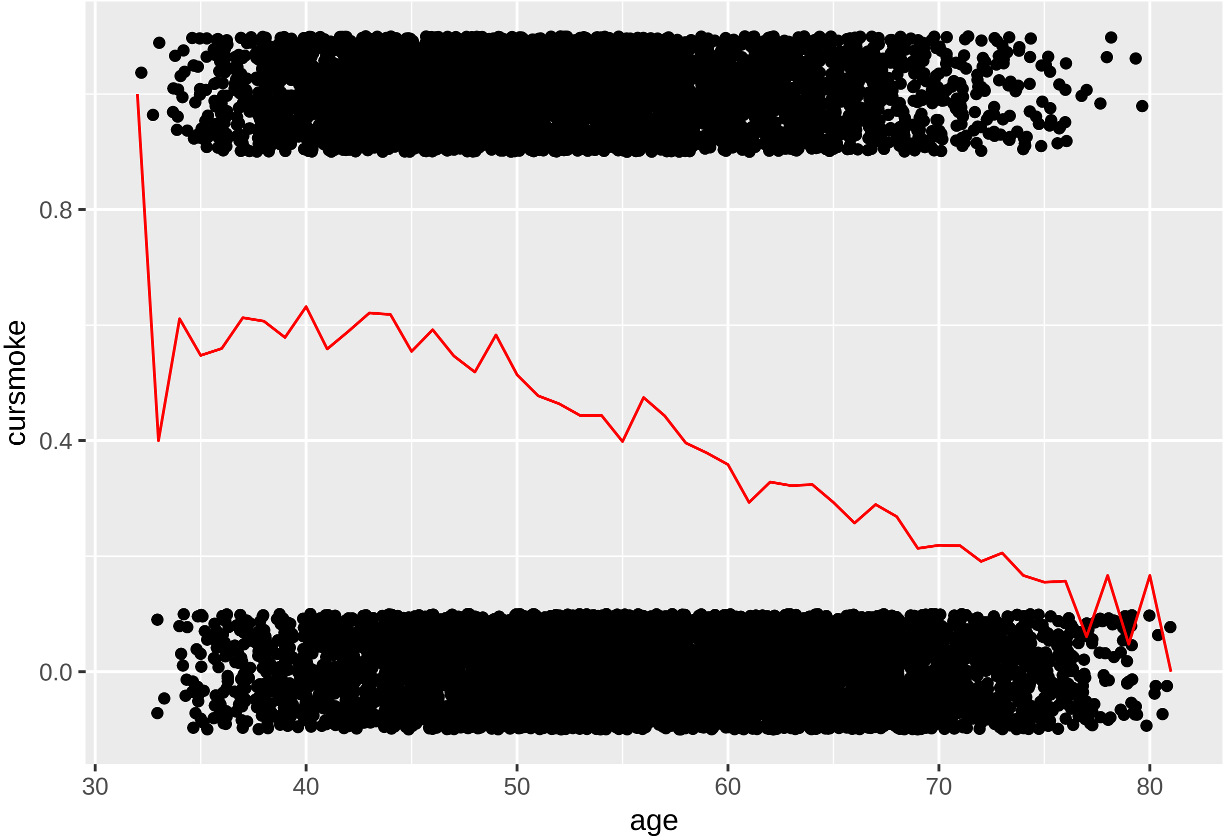


```
ggplot(data=data, aes(age, cigpday, group = randid)) +
  geom_path(alpha = .6) +
  geom_smooth(aes(group = NULL), method = 'lm', col = 'red') +
  geom_smooth(aes(group = NULL), method = 'loess') +
  ggtitle("cigpday vs Age Spaghetti Plot")
```

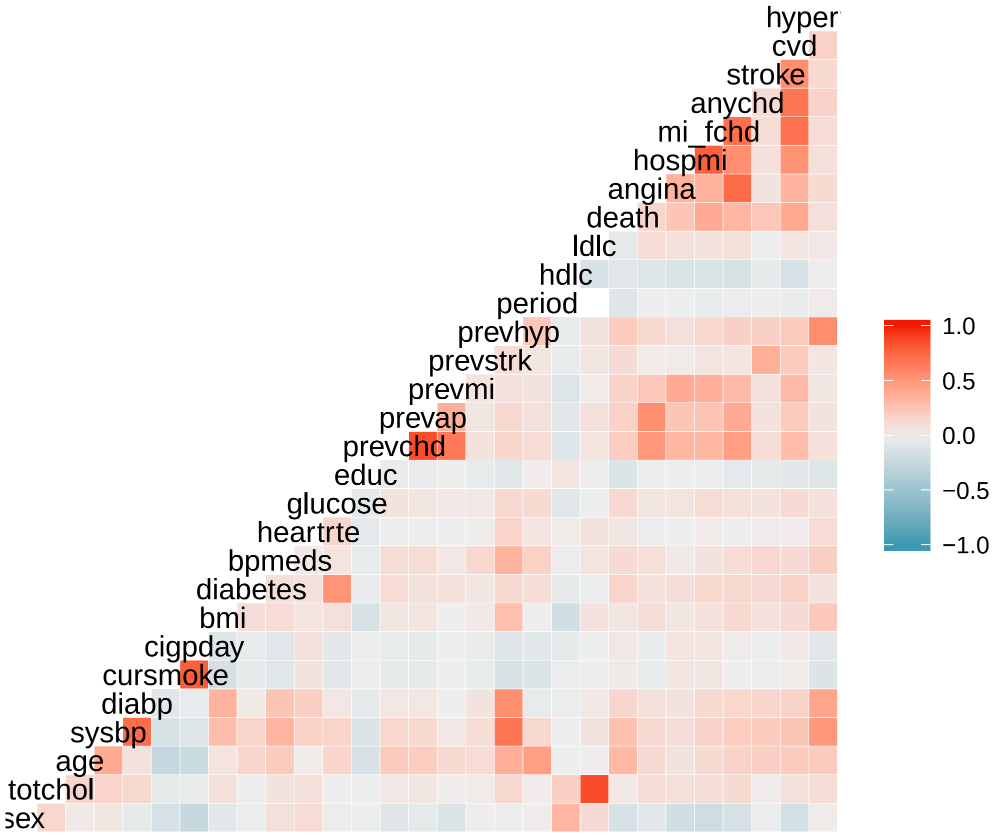


```
###

# data.residual <- data %>% na.omit()  %>%
# group_by(age) %>%
# mutate(mean.cigpday = mean(cigpday)) %>%
# ungroup() %>%
# mutate(residuals = cigpday - mean.cigpday) %>%
# group_by(randid) %>%
# mutate(median.residual = median(residuals)) %>%
# ungroup()
#
# data.stats <- c(min(data.residual$median.residual),
# quantile(data.residual$median.residual,
# c(.25, .5, .75)), max(data.residual$median.residual))
#
# data.id.select <- data.residual %>%
# filter(median.residual %in% data.stats)
#
# data.residual.plot <- ggplot() +
# geom_line(data = data.id.select,
# aes(x = age, y = cigpday, group = randid)) +
# ggtitle("cigpday by time, Selected from Residuals") +
# geom_smooth(data = data,
# aes(x = age, y = cigpday))

###
```

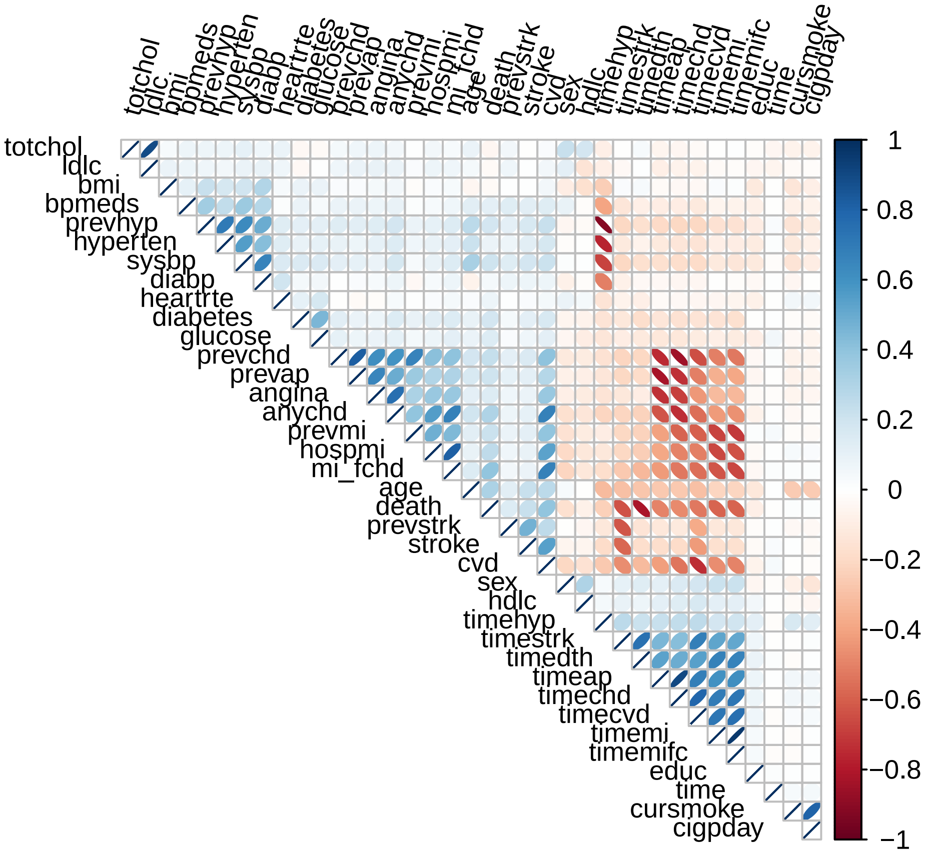
```r
ggplot(data, aes(y = cursmoke, x = age)) + geom_jitter(height = 0.1) +
stat_summary(fun.y = 'mean', geom="line", col = 'red')
```



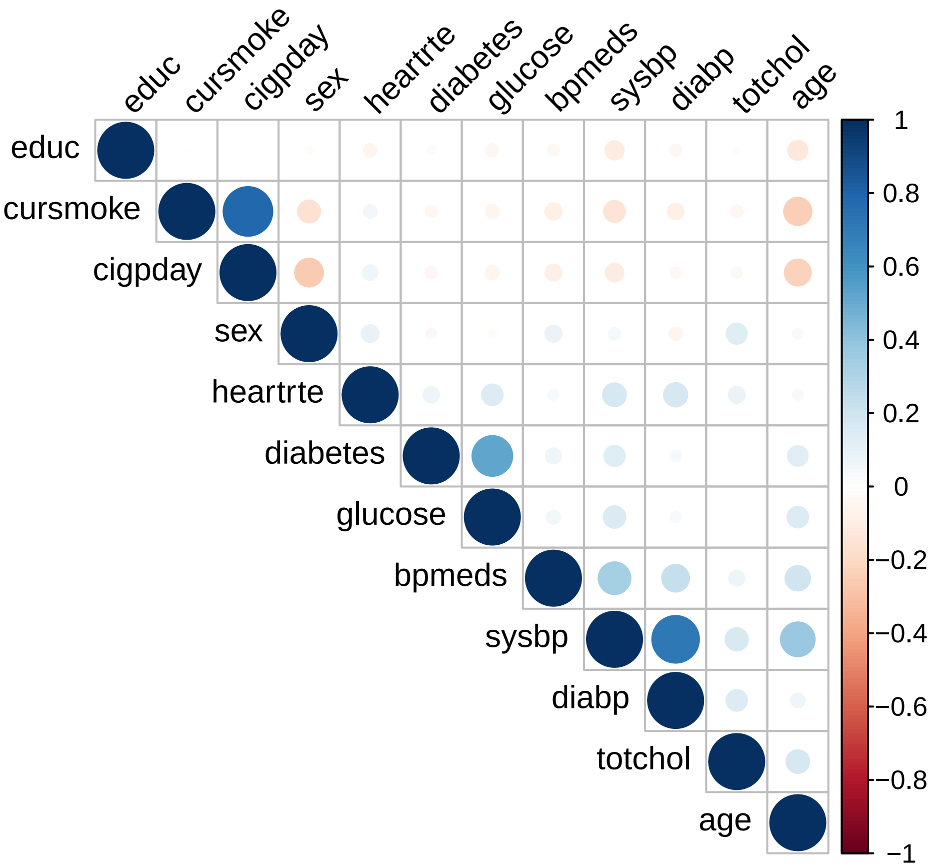
```r
select(data, -c(randid, time,timeap:timehyp)) %>% GGally::ggcorr(.)
```



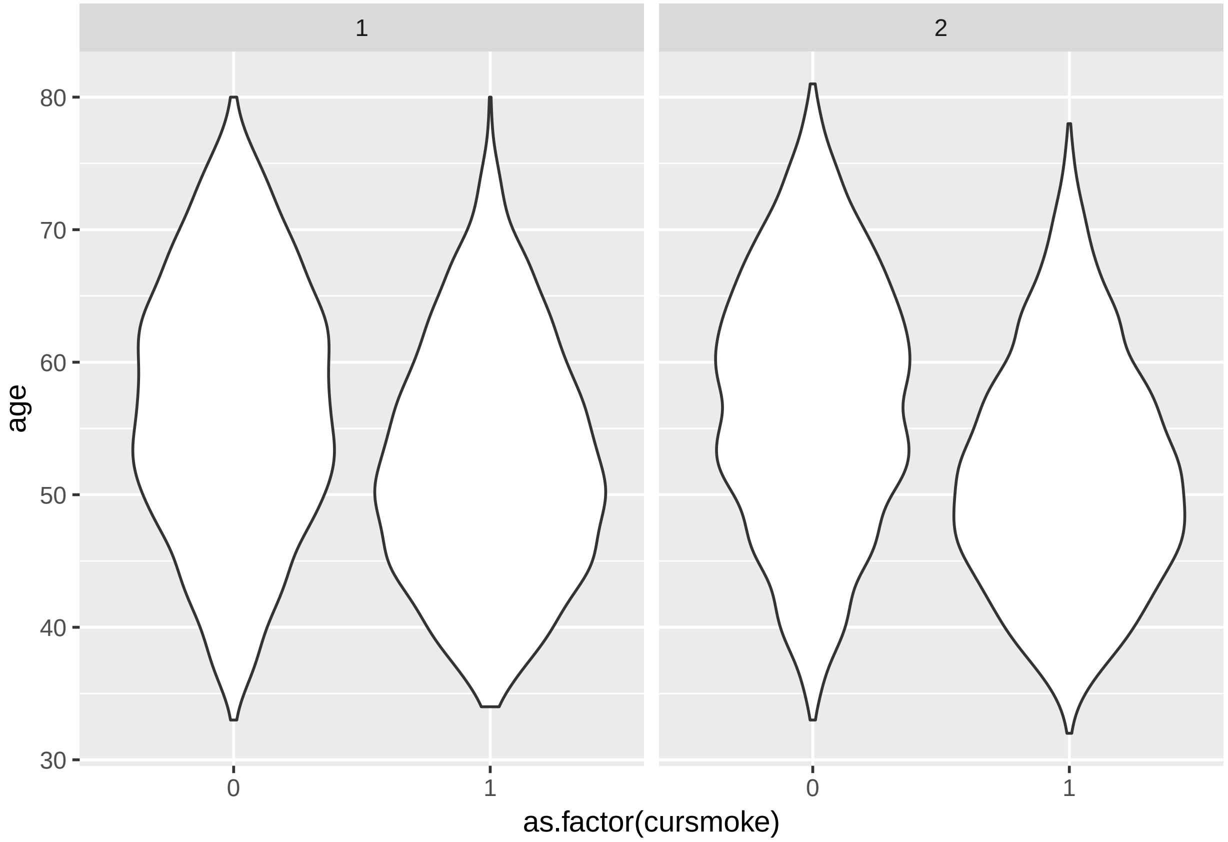
```
corr <- data[,c(-1,-21)] %>% cor(., use = "complete.obs")
library(corrplot)
corrplot(corr, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 75, tl.offset = 1, tl.cex = .8, method = "ellipse")
```



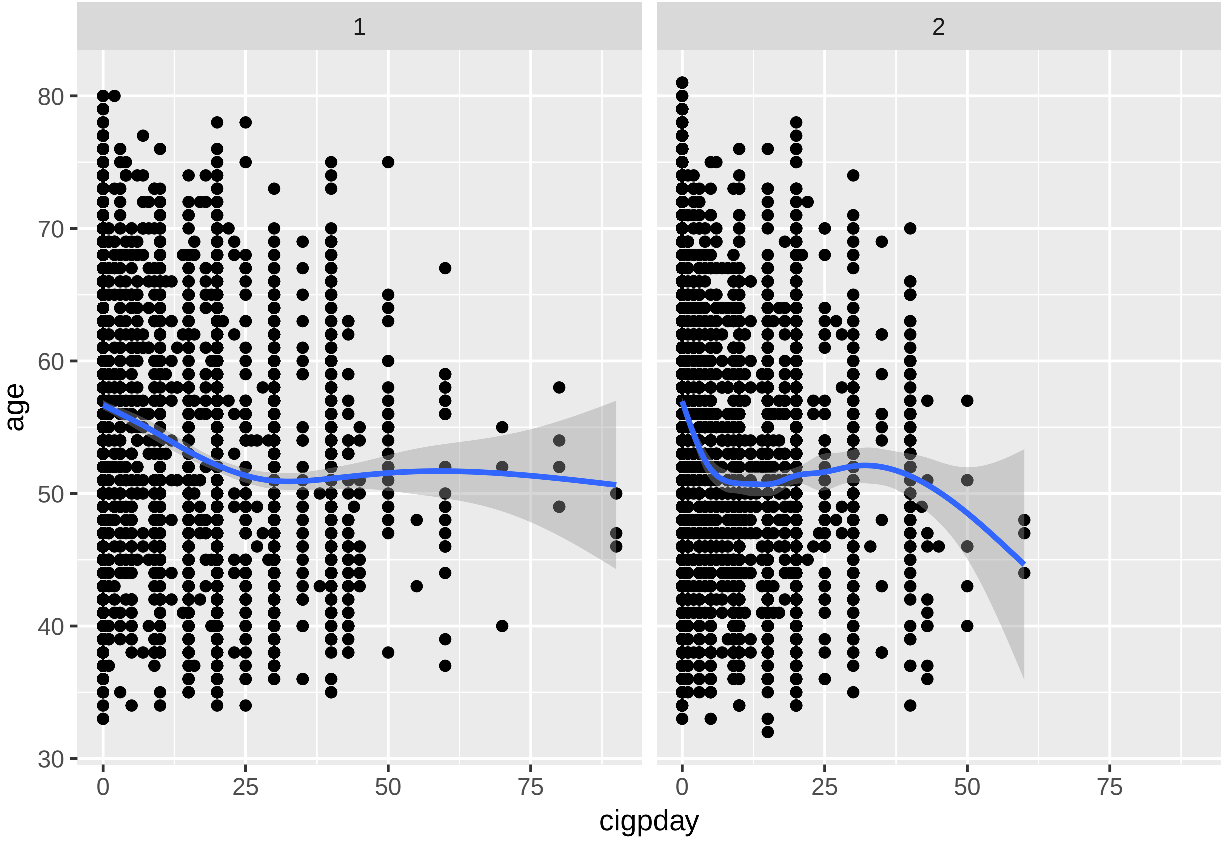
```
corr2 <- data[,c(2,3,4,5,6,7,8,10,11,12,13,14)] %>% cor(., use = "complete.obs")
corrplot(corr2, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```



```
ggplot(data = data,
aes(x = as.factor(cursmoke), y = age)) +
geom_violin() +
  facet_wrap(~sex)
```



```r
ggplot(data = data,
aes(x = cigpday, y = age)) +
geom_point() +
  facet_wrap(~sex) +
  geom_smooth()
```



Models

(1) Is there a relationship between age and smoking status? Does this relationship differ by sex?

```
gee::gee(cursmoke~age, id =randid, family=binomial, corstr = "unstructured", na.action = "na.omit")
```

```
## (Intercept)         age
##  2.80332932 -0.05649341
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Logit
##  Variance to Mean Relation: Binomial
##  Correlation Structure:     Unstructured
##
## Call:
## gee::gee(formula = cursmoke ~ age, id = randid, na.action = "na.omit",
##     family = binomial, corstr = "unstructured")
##
## Number of observations :  11627
##
## Maximum cluster size   :  3
##
```

```
##
## Coefficients:
## (Intercept)         age
##  2.44880724 -0.04948391
##
## Estimated Scale Parameter:  0.9808748
## Number of Iterations:  3
##
## Working Correlation[1:4,1:4]
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.7699785 0.5181921
## [2,] 0.7699785 1.0000000 0.5694565
## [3,] 0.5181921 0.5694565 1.0000000
##
##
## Returned Error Value:
## [1] 0
```

```r
gee::gee(cursmoke~age+sex, id =randid, family=binomial, corstr = "unstructured", na.action = "na.omit")
```

```
## (Intercept)         age         sex
##  3.78337733 -0.05686335 -0.61580832
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Logit
##  Variance to Mean Relation: Binomial
##  Correlation Structure:     Unstructured
##
## Call:
## gee::gee(formula = cursmoke ~ age + sex, id = randid, na.action = "na.omit",
##     family = binomial, corstr = "unstructured")
##
## Number of observations :  11627
##
## Maximum cluster size   :  3
##
##
## Coefficients:
## (Intercept)         age         sex
##  3.47963345 -0.05036642 -0.63551432
##
## Estimated Scale Parameter:  0.9829436
## Number of Iterations:  3
##
## Working Correlation[1:4,1:4]
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.7630374 0.5197662
## [2,] 0.7630374 1.0000000 0.5797515
## [3,] 0.5197662 0.5797515 1.0000000
##
##
```

```
## Returned Error Value:
## [1] 0
```

(2) Is there a relationship between the number of cigarettes smoked per day and age? Does this relationship differ by sex?

(3) The relationship between current smoking status and systolic blood pressure.

```
gee::gee(cursmoke~sysbp, id =randid, family=binomial, corstr = "unstructured", na.action = "na.omit")
```

```
## (Intercept)       sysbp
##  1.50037060 -0.01305331
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Logit
##  Variance to Mean Relation: Binomial
##  Correlation Structure:     Unstructured
##
## Call:
## gee::gee(formula = cursmoke ~ sysbp, id = randid, na.action = "na.omit",
##     family = binomial, corstr = "unstructured")
##
## Number of observations :  11627
##
## Maximum cluster size   :  3
##
##
## Coefficients:
##  (Intercept)        sysbp
##  0.901709618 -0.008586543
##
## Estimated Scale Parameter:  0.9902469
## Number of Iterations:  3
##
## Working Correlation[1:4,1:4]
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.7729372 0.5089701
## [2,] 0.7729372 1.0000000 0.5703181
## [3,] 0.5089701 0.5703181 1.0000000
##
##
## Returned Error Value:
## [1] 0
```

(2) The relationship between current smoking status and diastolic blood pressure.

```
gee::gee(cursmoke~diabp, id =randid, family=binomial, corstr = "unstructured", na.action = "na.omit")
```

```
## (Intercept)       diabp
##  0.83258871 -0.01331811
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
```

```
##
## Model:
##  Link:                      Logit
##  Variance to Mean Relation: Binomial
##  Correlation Structure:     Unstructured
##
## Call:
## gee::gee(formula = cursmoke ~ diabp, id = randid, na.action = "na.omit",
##     family = binomial, corstr = "unstructured")
##
## Number of observations :  11627
##
## Maximum cluster size   :  3
##
##
## Coefficients:
##  (Intercept)        diabp
##  0.200588474 -0.005679433
##
## Estimated Scale Parameter:  0.9971305
## Number of Iterations:  3
##
## Working Correlation[1:4,1:4]
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.7791796 0.5092309
## [2,] 0.7791796 1.0000000 0.5716509
## [3,] 0.5092309 0.5716509 1.0000000
##
##
## Returned Error Value:
## [1] 0
```

(3) The relationship between current smoking status and serum total cholesterol.

```
gee::gee(cursmoke~totchol, id =randid, family=binomial, corstr = "unstructured", na.action = "na.omit")
```

```
##  (Intercept)      totchol
##  0.180474997 -0.001869103
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Logit
##  Variance to Mean Relation: Binomial
##  Correlation Structure:     Unstructured
##
## Call:
## gee::gee(formula = cursmoke ~ totchol, id = randid, na.action = "na.omit",
##     family = binomial, corstr = "unstructured")
##
## Number of observations :  11218
##
## Maximum cluster size   :  3
##
```

```
##
## Coefficients:
##   (Intercept)       totchol
## -0.0335972229 -0.0009948755
##
## Estimated Scale Parameter:  0.999776
## Number of Iterations:  3
##
## Working Correlation[1:4,1:4]
##            [,1]      [,2]      [,3]
## [1,] 1.0000000 0.7790456 0.4633692
## [2,] 0.7790456 1.0000000 0.5223622
## [3,] 0.4633692 0.5223622 1.0000000
##
##
## Returned Error Value:
## [1] 0
```

What to Turn In: For this assignment you will turn in a single pdf document with (a) your summary of findings and (b) an Appendix with figures and (c) an Appendix with all R code (in the form of a knited pdf).

OBJECTIVE:

An objective or description of the goals of the analysis

We are interested to describe the smoking habits of the participants in the Framingham Heart study as they age and the impact of smoking on certain health outcomes. The Framingham heart study asks participants about their smoking habits at each visit. In particular, participants are asked if they are currently smoking at this visit (0 = Not a current smoker, 1 = Current smoker), which we will refer to as current smoking status. In addition, participants also report the number of cigarettes they are smoking per day. A more complete description of each of variables in the Framingham Heart study can be found in the Framingham Heart Study Longitudinal Data Documentation.

We are interested to answer the following questions:

(1) Is there a relationship between age and smoking status? Does this relationship differ by sex?

(2) Is there a relationship between the number of cigarettes smoked per day and age? Does this relationship differ by sex?

While answering these questions, please account for any confounders that you have evidence may impact the relationship between age and sex with smoking.

Next you are interested in the relationship between certain health outcomes and smoking status. In particular you are interested in :

(1) The relationship between current smoking status and systolic blood pressure.

(2) The relationship between current smoking status and diastolic blood pressure.

(3) The relationship between current smoking status and serum total cholesterol.

Again, while answering these questions, please account for any confounders that you have evidence may impact these relationships.

STUDY DESIGN:

A brief description of the study design and the data

METHODS:

A methods section describing your statistical analysis (please justify all modeling choices that were made with evidence).

RESULTS:

A results section that includes a) descriptive statistics for the data b) a summary of your key findings including supporting numerical summaries (i.e. confidence intervals, pvalues, etc.) c) interpretations of your key findings (i.e. interpretations of coefficients).

CONCLUSION:

A conclusion specifically answering the objective of the analysis.

APPENDIX: