Machine Learning group project

Michal Heydel

27 November 2017

```
# input Stata file
library(haven)
ghs<- read_dta("C:/Users/Michal/Documents/01- Master Degree/GitHub/ST443-Project-group9/Stata/stata8/ghs
```

```
setwd("C:/Users/Michal/Documents/01- Master Degree/GitHub/ST443-Project-group9")
```

```
df = ghs
```

```
##Missing values
df[df== -9]<-NA
df[df== -6 ]<-NA
df[df== -7 ]<-NA
df[df== -8 ]<-NA
```

```
# df[df== ""] <-NA # DOESNT WORK
```

```
df1 = df
```

```
df1 = df1[-which(is.na(df1$grearn)), ]
```

```
df1 = df1[, -which(colMeans(is.na(df1)) > 0.3)]
```

```
write.table(df1, file = "data_reduced_416.csv", col.names = TRUE, na = "NA", sep = "," , row.names = FA
```

```
#df1$se1
```

```
df2 <- read.table(file= "data_reduced_416.csv", na.strings=c("", "NA"), sep=",", header = TRUE)
```

```
df2 = df2[, -which(colMeans(is.na(df2)) > 0.3)]
```

```
write.table(df2, file = "data_reduced_411.csv", col.names = TRUE, na = "NA", sep = "," , row.names = FA
```

```
head(df2)
```

```
dim(df2)
```

```
## [1] 20208   411
```