FeatureSelection

Anushree Shivarudrappa

Pre-Processing

```
library(data.table)
library(ggplot2)
library(dplyr)
library(scales)
library(RColorBrewer)
library(tidyr)
library(corrplot)
```

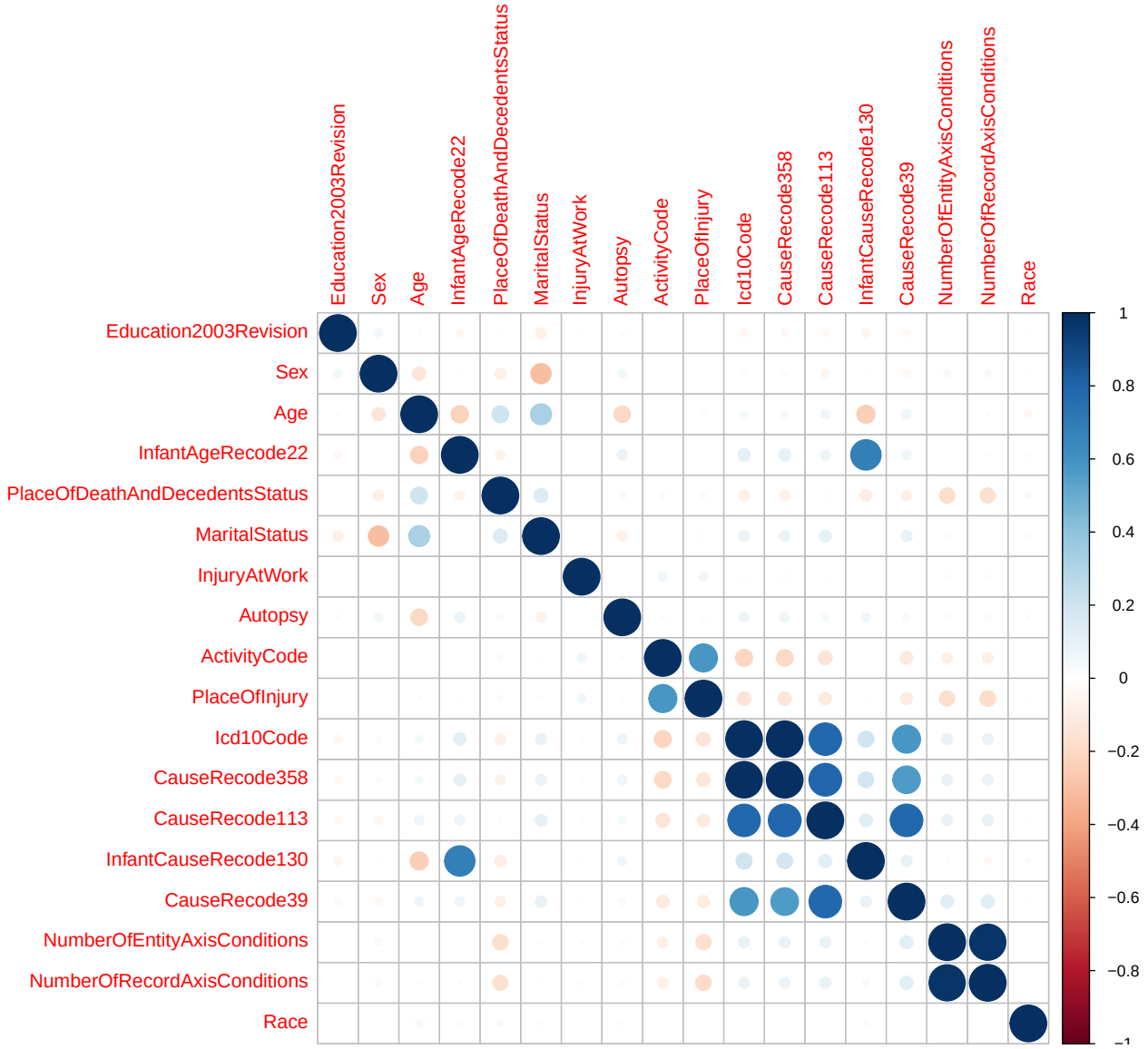
Data Loading

```
Death_US <- fread("DeathRecords.csv", header = T)
```

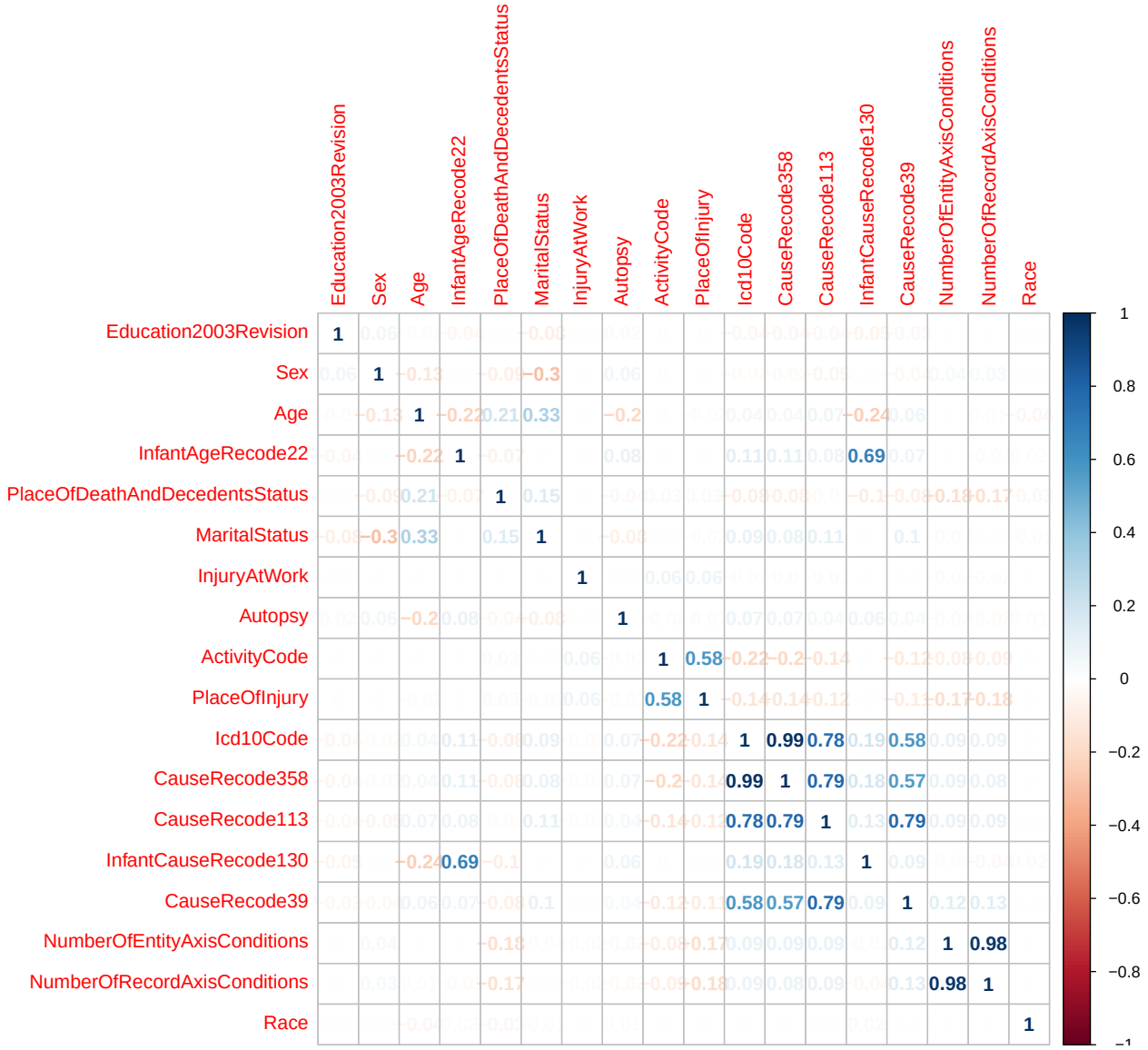
Feature selection for modeling

Predicting the CauseRecode39 from the Natural death dataset

```
# separates natural death
Death_US_natural <- Death_US[Death_US$MannerOfDeath == 7, ]

# Select required variable
require(MASS)
require(dplyr)
natural_sub <- Death_US_natural %>% dplyr::select(Education2003Revision, Sex, Age,
                      InfantAgeRecode22,
                      PlaceOfDeathAndDecedentsStatus, MaritalStatus, InjuryAtWork,
                      Autopsy, ActivityCode, PlaceOfInjury, Icd10Code,CauseRecode358,
                      CauseRecode113, InfantCauseRecode130,CauseRecode39,
                      NumberOfEntityAxisConditions,NumberOfRecordAxisConditions,Race)

# Converting Character variable into Integer variable
natural_sub$Sex <- as.integer(as.factor(natural_sub$Sex))
natural_sub$MaritalStatus <- as.integer(as.factor(natural_sub$MaritalStatus))
natural_sub$InjuryAtWork <- as.integer(as.factor(natural_sub$InjuryAtWork))
natural_sub$Autopsy <- gsub("n", "N", natural_sub$Autopsy)
natural_sub$Autopsy <- as.integer(as.factor(natural_sub$Autopsy))
natural_sub$Icd10Code <- as.integer(as.factor(natural_sub$Icd10Code))

# find the correlation
corr <- cor(natural_sub)
# Plot the correlation
corrplot(corr, method="circle")
```




```
corrplot(corr, method="number")
```



As we know "CauseRecord358", "CauseRecord113", "CauseRecord39" are the recoded bind of ICD10 code. If we use these variable it will cause multicollinearity. So we can not use "CauseRecord358", "CauseRecord113" and "ICD10Code" as the model variable.

From the above visual correlation matrix we can say "Age, InfantAgeRecode22, PlaceOfDeathAndDecedentsStatus, MaritalStatus, ActivityCode, PlaceOfInjury, NumberOfRecordAxisConditions and NumberOfEntityAxisConditions" variables has a correlation with "CauseRecord39", therefore selecting these variables for the model.