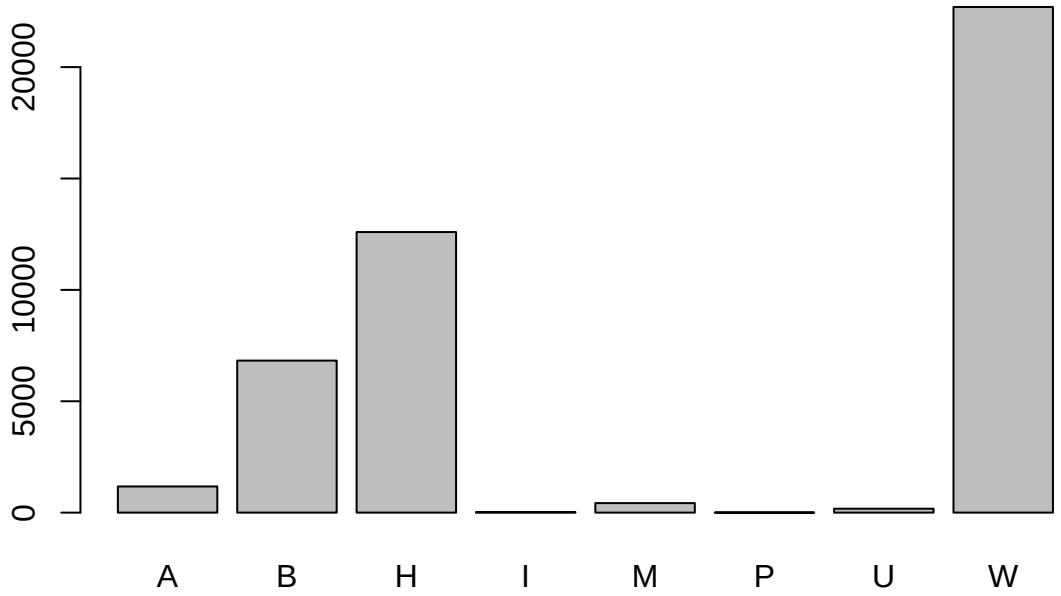
MATH215 - Assignment 2

Pierre Visconti

Problem 1:

```
austin = read.csv("Austin_Profiling.csv")
barplot(table(austin$apd.race))
```




```
austin.new = data.frame(austin)
# deletes rows from austin.new where apd.race != A,B,H,W
austin.new = austin.new[-c(which(austin.new$apd.race=='I'),
                           which(austin.new$apd.race=='M'),
                           which(austin.new$apd.race=='P'),
                           which(austin.new$apd.race=='U')),]
unique(austin.new$apd.race)
```

```
## [1] "W" "H" "B" "A"
```

```
# deletes rows from austin.new where SEX != F,M
unique(austin.new$SEX)
```

```
## [1] "F" "M" "U"
```

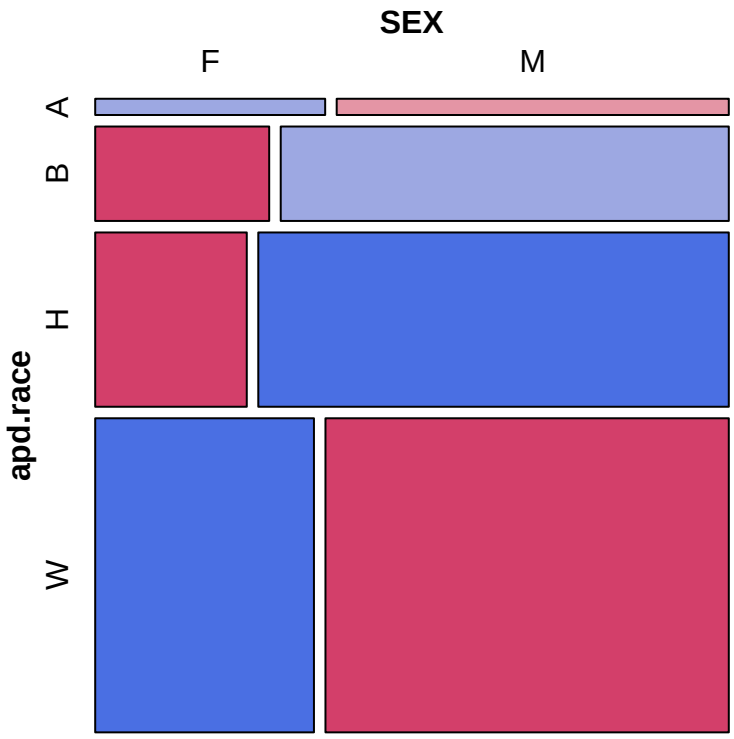
```
austin.new = austin.new[-(which(austin.new$SEX=="U")),]
unique(austin.new$SEX)
```

```
## [1] "F" "M"
```

```
library(vcd)
```

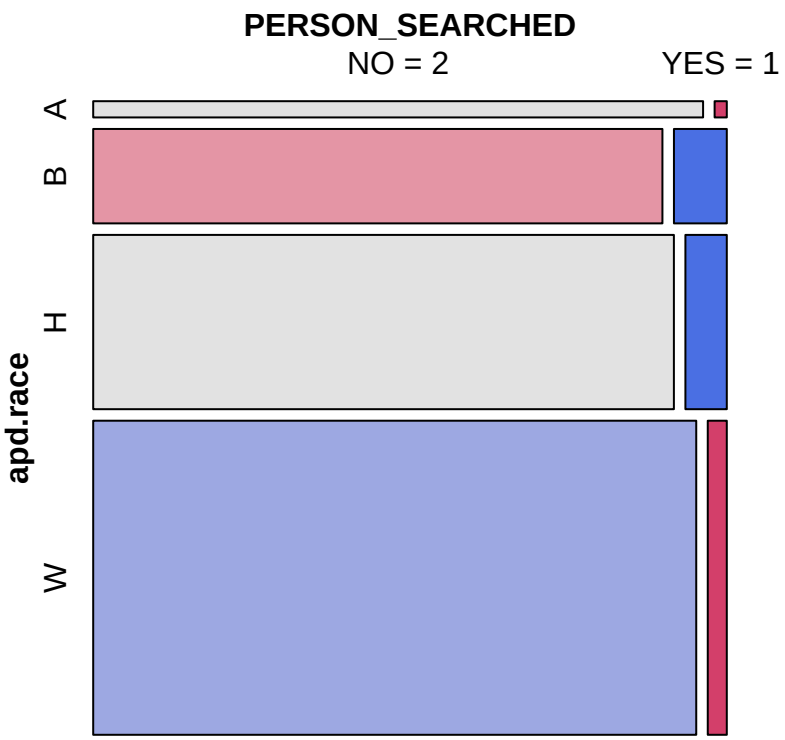
```
## Loading required package: grid
```

```
mosaic(~apd.race+SEX, data=austin.new, shade=T, legend=F)
```



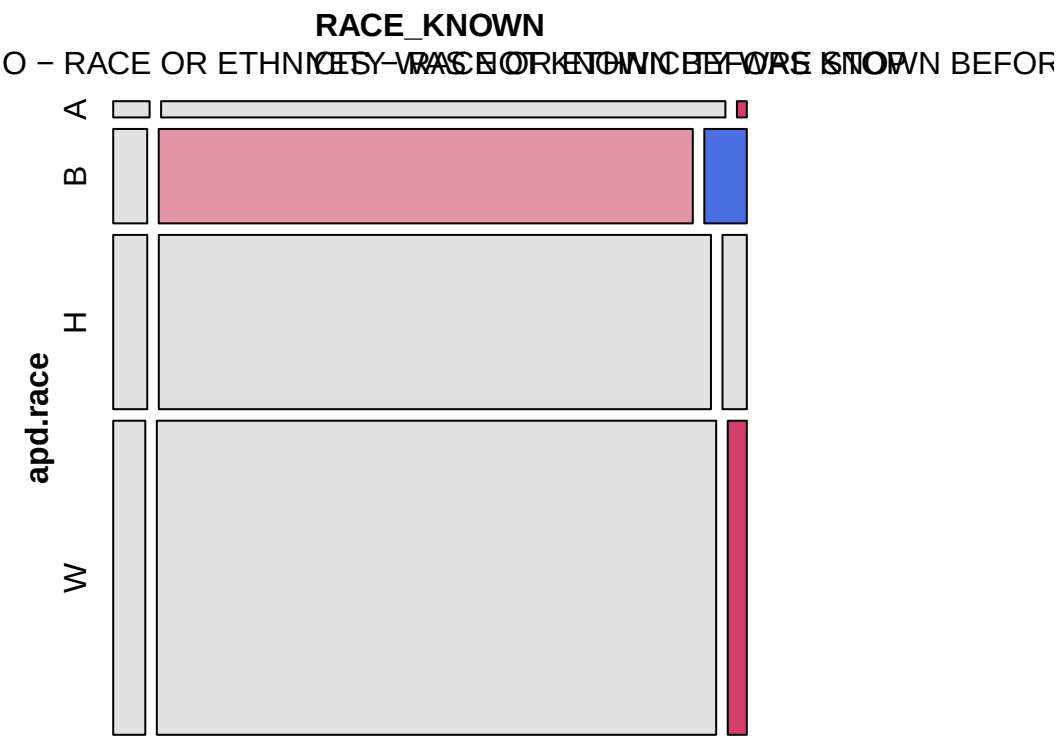
There appears to be a relationship between race and sex that exists. For each race category there is a larger percentage of males than female regardless of the total number of people associated with that race.

```
mosaic(~apd.race+PERSON_SEARCHED, data=austin.new, shade=T, legend=F, na.rm=T)
```



There appear to be some differences between whether a person was searched or not, depending on their race. H and B were searched more than W or A yet the differences are not substantial. I would not say for certain that there exists a relationship.

```
mosaic(~apd.race+RACE_KNOWN, data=austin.new, shade=T, legend=F, na.rm=T)
```



This relationship is extremely similar to the one from before. There is the same trend where B>H>W>A yet the differences are quiet small and it is hard to say whether there exists a relationship.

Problem 2:

```
ed = read.csv("county_education.csv")
names(ed) = c("FIPS", "state", "county", "inc_19", "inc_20", "inc_21",
              "asc_deg_n", "bac_deg_n", "asc_deg_p", "bac_deg_p")
```

```
length(unique(ed$state))
```

```
## [1] 49
```

```
unique(ed$state)
```

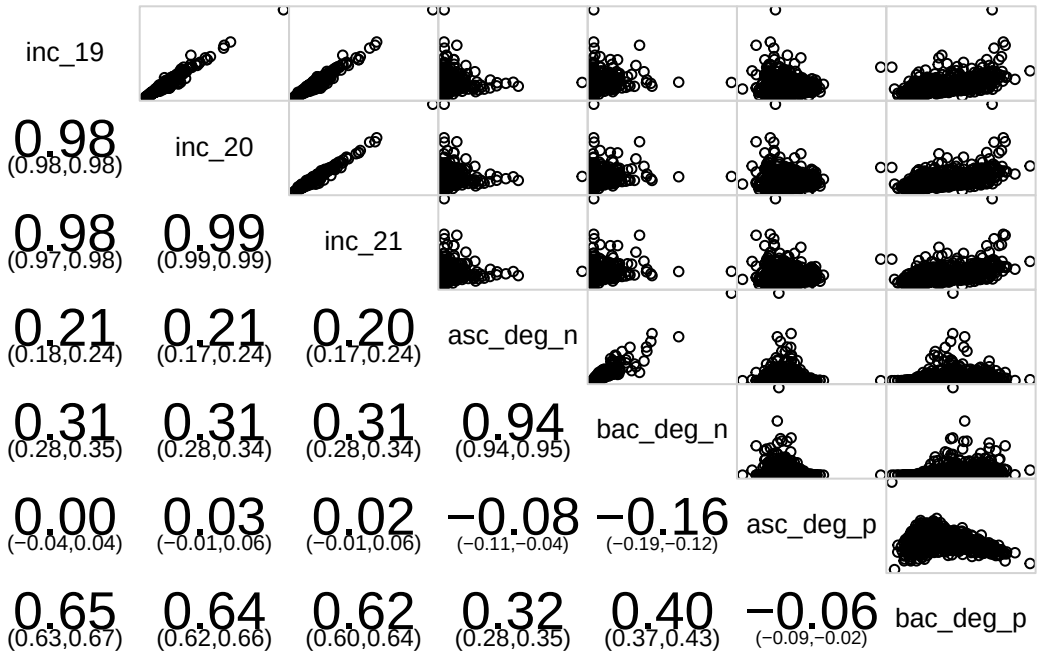
```
##  [1] "VA" "NM" "CO" "MD" "NY" "NC" "TN" "CA" "IN" "MA" "MI" "UT" "KS" "NJ" "OH"
## [16] "WY" "GA" "IA" "PA" "WA" "OR" "TX" "VT" "WI" "MT" "MN" "IL" "MS" "RI" "CT"
## [31] "SD" "ME" "MO" "FL" "SC" "KY" "ID" "AL" "WV" "NH" "ND" "NE" "AK" "OK" "AZ"
## [46] "DE" "HI" "AR" "NV"
```

```
length(unique(ed$county))
```

```
## [1] 3006
```

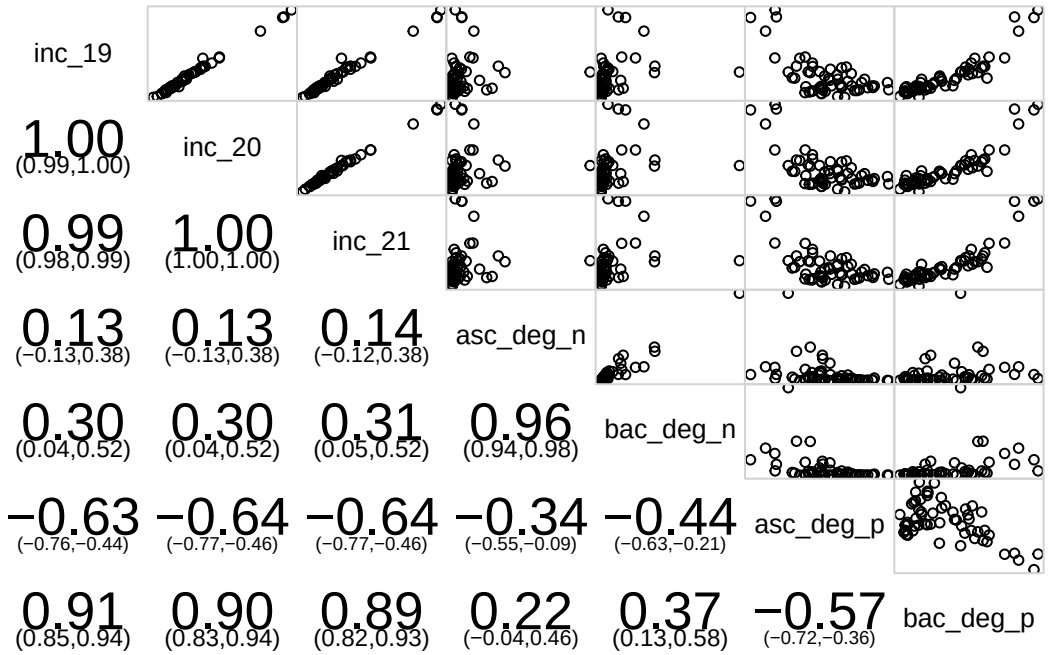
There are 49 states represented in this dataset, the state of Louisiana (LA) is missing. According to Wikipedia, as of 2020 there are 3,143 counties in the US, so this dataset is missing 137 counties.

```
library(corrgram)
suppressWarnings(corrgram(ed[4:10], lower.panel = panel.conf,
                          upper.panel = panel.pts))
```



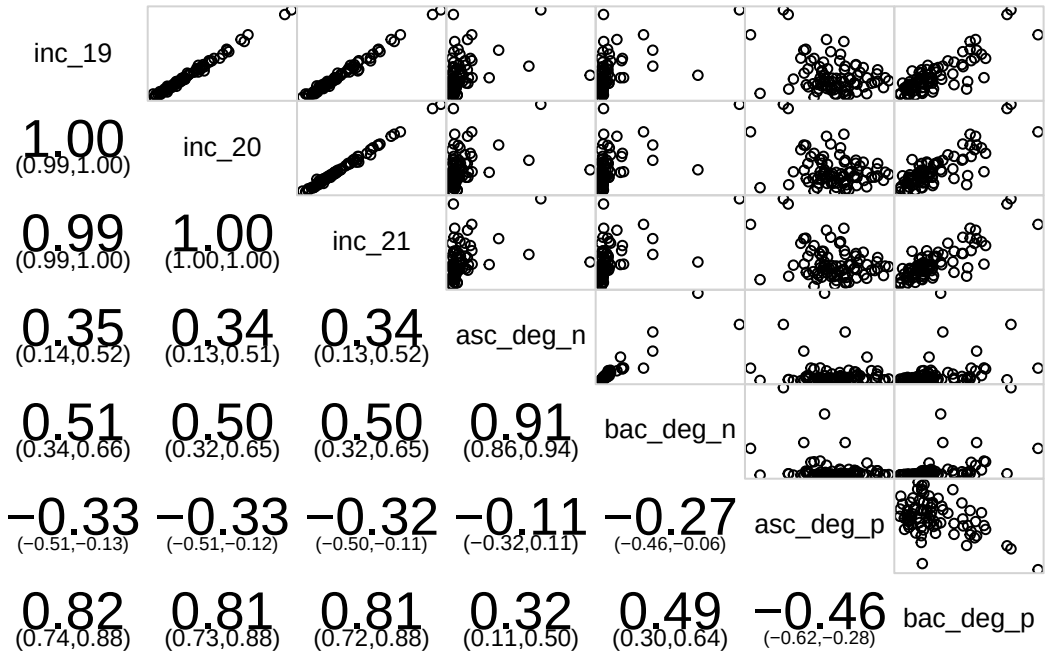
There aren't many meaningful relationships, based off this correlation matrix. The only relationship that appears to exist is between income and the percent of people who have a Bachelor's degree. The other relationships that exist are between variables that have similar data, like between income in 2019 and 2020 for example.

```
ed.CA = data.frame(ed[c(which(ed$state=="CA")),])
suppressWarnings(corrgram(ed.CA[4:10], lower.panel = panel.conf,
                          upper.panel = panel.pts))
```



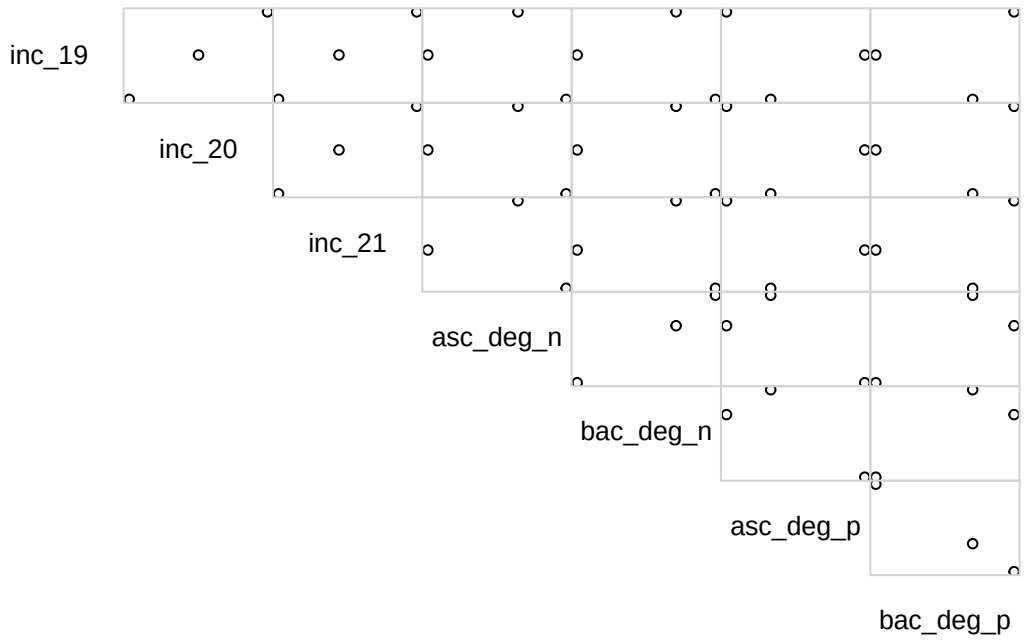
This dataset shows much stronger relationships between income and percentage of people who have either an associate's or bachelor's degree.Interestingly the bachelor's have a positive relationship with income and associate's have a negative relationship. There also appears to be some inverse relations between percent with a bachelor and percent with an associate.

```
ed.MI = data.frame(ed[c(which(ed$state=="MI")),])
suppressWarnings(corrgram(ed.MI[4:10], lower.panel = panel.conf,
                          upper.panel = panel.pts))
```



For this dataset the bachelor percent to income relationship is still present yet the negative relationship with the associate to income is less significant. It is important to note that the income to bachelor percent relationship is less significant here than with CA.

```
ed.HI = data.frame(ed[c(which(ed$state=="HI")),])
suppressWarnings(corrgram(ed.HI[4:10], lower.panel = panel.conf,
                          upper.panel = panel.pts))
```




```
length(ed.HI$FIPS)
```

```
## [1] 3
```

This particular analysis is not meaningful.There appear to only be 3 counties that are included in the dataset for HI, which could be why this is happening.

Problem 3:

```
set.seed(215)
pop = rep(0, 1500)
pop[sample(1:1500, 15, replace=T)]=1
pos.cnt=0
user.cnt=0
for (i in 1:10000) {
  test = sample(1:1500,1)
  temp = runif(1)
  if (pop[test]==1 & temp<=0.9) {
    pos.cnt = pos.cnt + 1
    user.cnt = user.cnt + 1
  } else if (pop[test]==0 & temp<=0.02) {
    pos.cnt = pos.cnt + 1
```

```
  }
}

user.cnt / pos.cnt
```

```
## [1] 0.3039216
```