# Project 1

*Aaron Niskin, Charlie Edelson, Chris Leonard*

*October 26, 2016*

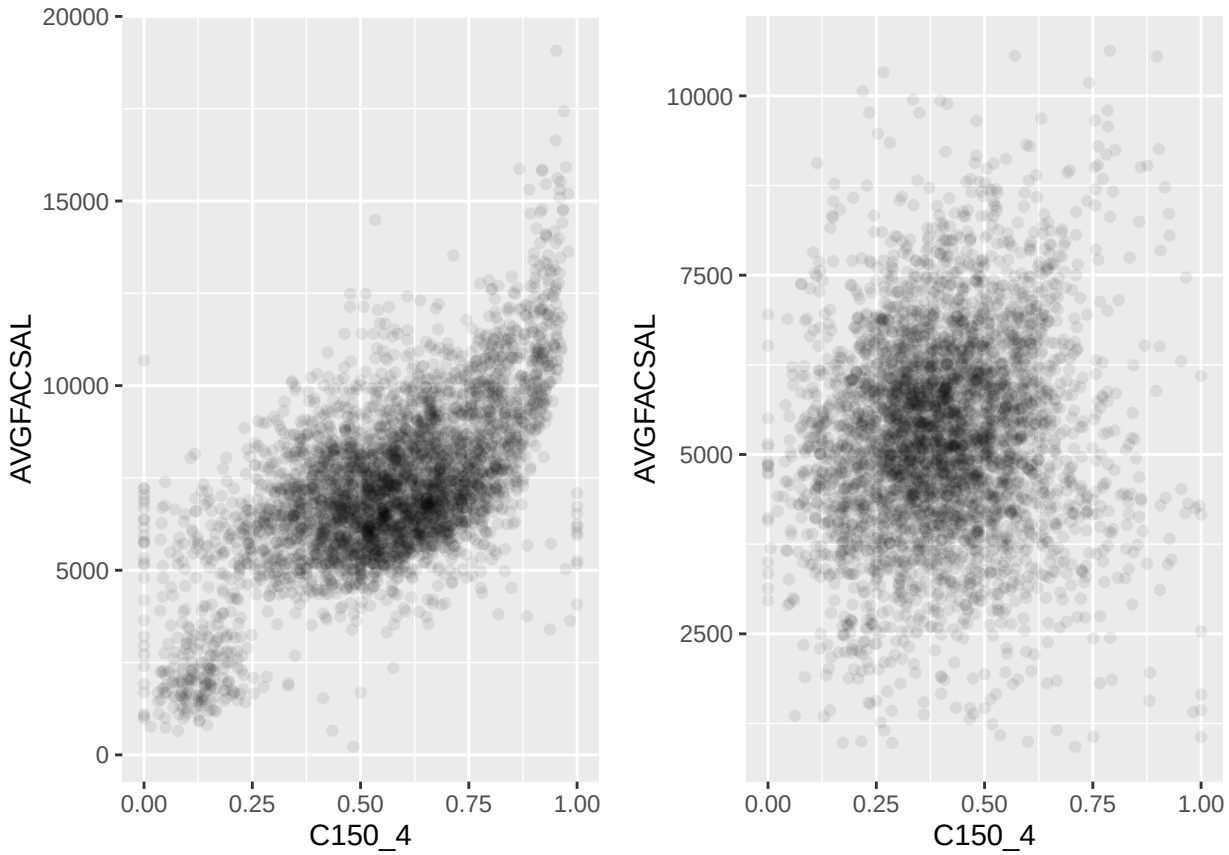
```
library(ggplot2)
library(gridExtra)
setwd("~/Documents/courses/newCollege/current/eda/projects/project1")
```

```
getFilesColumns <- function(years, cols, nullVals) {
  r_formatting_sucks <- function(n) {
    if(n < 10) {
      return(paste("0", n, sep=""))
    }
    else {
      return(n)
    }
  }
  getFileName <- function(startYear) {
    R_is_stupid <- paste("ignore/MERGED", startYear, "_", r_formatting_sucks((startYear + 1) %% 100), "
    return(R_is_stupid)
  }
  getThisFile <- function(year) {
    tmp <- read.csv(getFileName(year), na=nullVals, stringsAsFactors = FALSE)
    tmp <- tmp[,cols]
    tmp$startYear = year
    return(tmp)
  }
  tmpDf <- getThisFile(years[1])
  for (year in years[2:length(years)]){
    tmpDf <- rbind(tmpDf, getThisFile(year))
  }
  tmpDf$startYear <- as.factor(tmpDf$startYear)
  return(tmpDf)
}
```

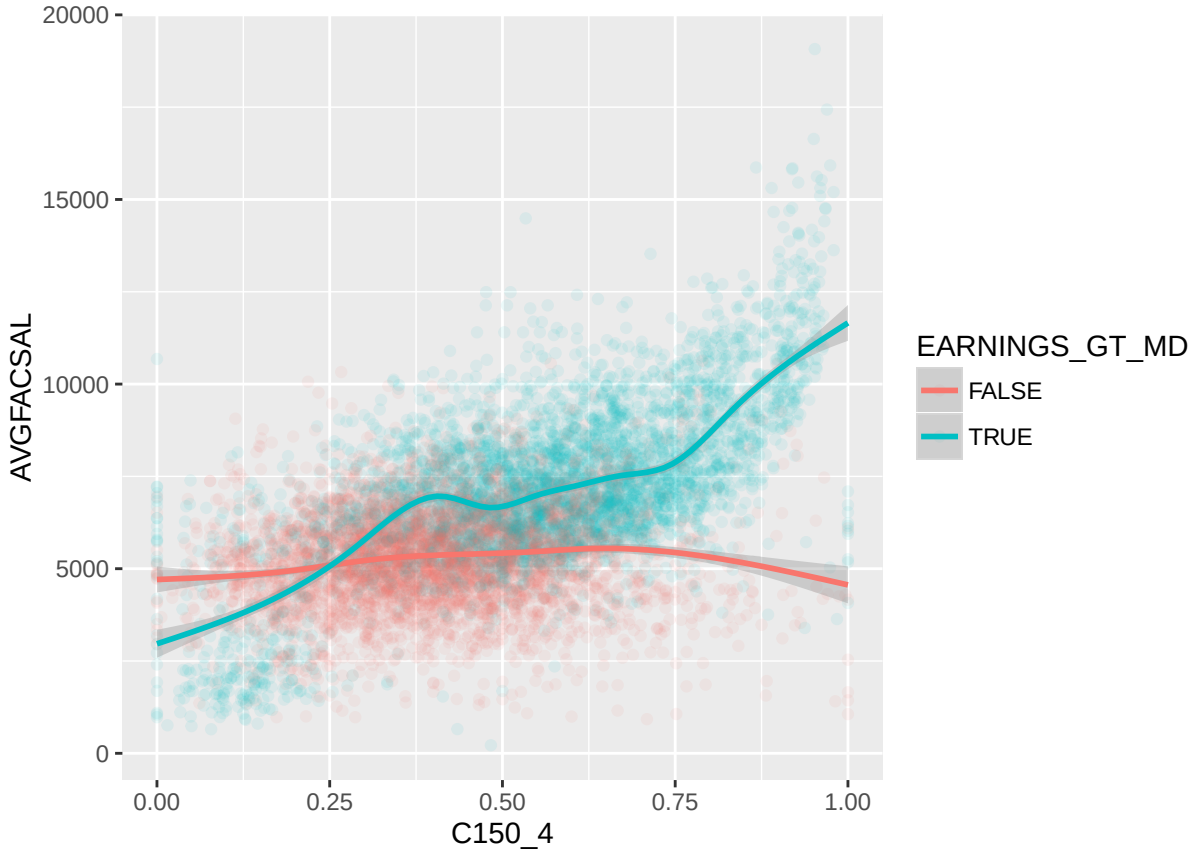
```
dataStartYears <- 1996:2014
csvNullVals = c("NULL", "PrivacySuppressed")
desiredCols = c("FAMINC", "AVGFACSAL", "C150_4", "MD_EARN_WNE_P10")

DF = getFilesColumns(dataStartYears, desiredCols, csvNullVals)
rm(getFilesColumns)
rm(dataStartYears)
rm(csvNullVals)
rm(desiredCols)
```

```
adf = DF[, c("C150_4", "AVGFACSAL", "MD_EARN_WNE_P10")]
adf = adf[complete.cases(adf),]
tmpMD <- median(adf[complete.cases(adf$C150_4),"MD_EARN_WNE_P10"], na.rm=TRUE)
adf$EARNINGS_GT_MD = adf$MD_EARN_WNE_P10 > tmpMD
rm(tmpMD)
```

```
plt1 <- ggplot(data=adf[adf$EARNINGS_GT_MD == TRUE,], aes(x=C150_4, y=AVGFACSAL), main="More than median
plt2 <- ggplot(data=adf[adf$EARNINGS_GT_MD == FALSE,], aes(x=C150_4, y=AVGFACSAL), main="Less than media
grid.arrange(plt1, plt2, ncol=2)
```

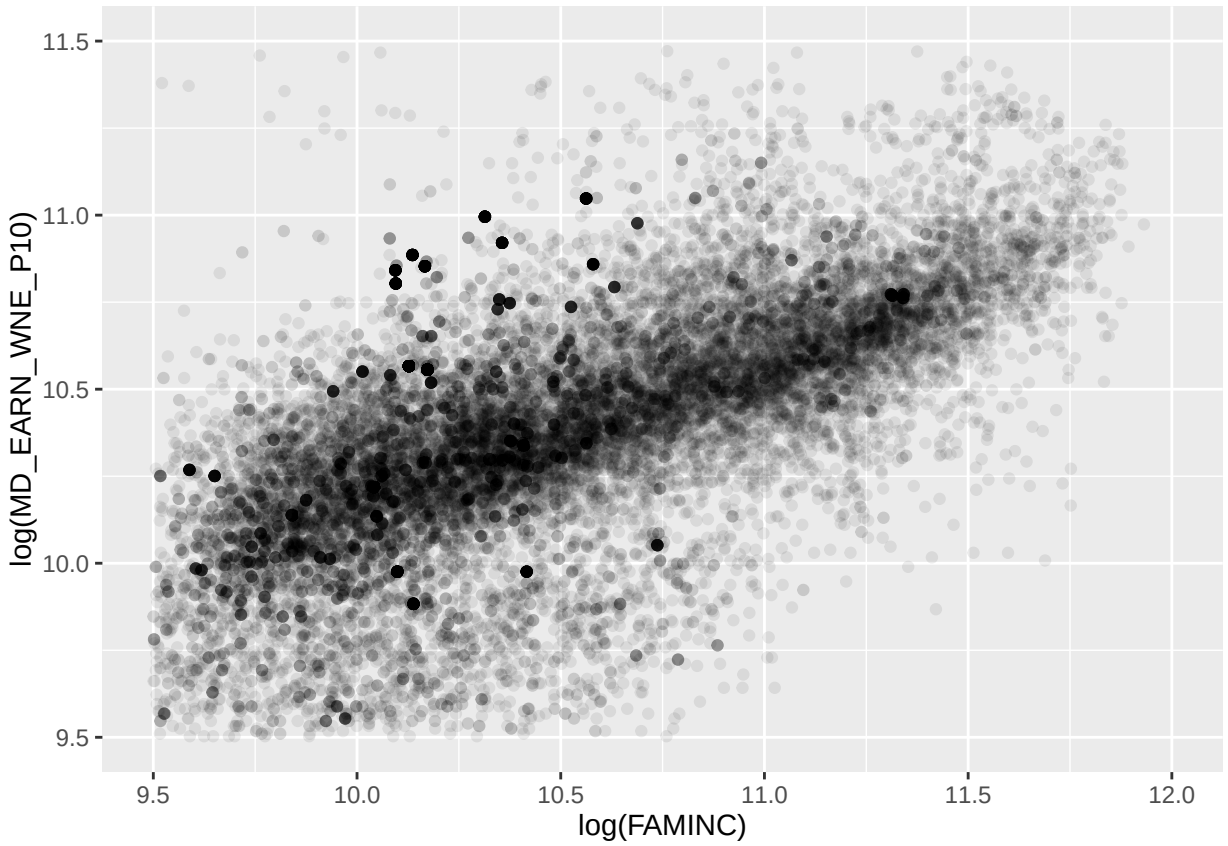



```
ggplot(data=adf, aes(x=C150_4, color=EARNINGS_GT_MD, y=AVGFACSAL), ylim=c(0,10^5)) + geom_point(alpha=I
```




```
adf = DF[,c("FAMINC", "MD_EARN_WNE_P10")]
adf = adf[complete.cases(adf),]
ggplot(data=adf, aes(x=log(FAMINC), y=log(MD_EARN_WNE_P10))) + geom_point(alpha=0.08) + xlim(9.5,12) + y
```

```
## Warning: Removed 937 rows containing missing values (geom_point).
```


log(MD_EARN_WNE_P10)
11.5
11.0
10.5
10.0
9.5
9.5
10.0
10.5
11.0
11.5
12.0
log(FAMINC)

```
rm(adf)
```