stat139_project

Evan Arnold and Caleb Ren

11/13/2019

EDA

```
vsrr <- read.csv(url("https://data.cdc.gov/api/views/xkb8-kh2a/rows.csv"))
deaths <- subset(vsrr, Indicator == "Number of Deaths")
View(vsrr)
```

```
## Warning in system2("/usr/bin/otool", c("-L", shQuote(DSO)), stdout = TRUE):
## running command ''/usr/bin/otool' -L '/Library/Frameworks/R.framework/
## Resources/modules/R_de.so'' had status 69
```

Download and clean drug overdose dataset.

```
# overdose dataset
overdose <- read.csv("data/overdose.csv")

# remove columns
bad.cols <- c("Period", "Percent.Complete", "Percent.Pending.Investigation",
              "State.Name", "Footnote", "Footnote.Symbol", "Predicted.Value")
overdose <- overdose[,!(colnames(overdose) %in% bad.cols)]

# reshape dataframe to wide format
overdose <- reshape(overdose, idvar = c("State", "Year", "Month"),
                    timevar = "Indicator", direction = "wide")

# proper column naames
names <- c("state", "year", "month", "overdoseDeaths",
           "natural.semiSynthetic.synthetic.methadone",
           "opioids", "cocaine", "stimulants", "deaths",
           "synthetic.noMethadone", "heroin",
           "natural.semiSynthetic.methadone",
           "natural.semiSynthetic", "percentSpecified",
           "methadone")
colnames(overdose) <- names

# remove aggregate statistics
overdose <- overdose[!(overdose$state %in% c("US", "YC")),]
overdose$state <- droplevels(overdose$state)

# reformat month to ordered factor
months.levels <- c("January", "February", "March", "April", "May", "June",
                   "July", "August", "September", "October", "November", "December")
months.labels <- unname(sapply(tolower(months.levels), function(x) substr(x, 1, 3)))
overdose$month <- ordered(overdose$month, levels = months.levels, labels = months.labels)
```

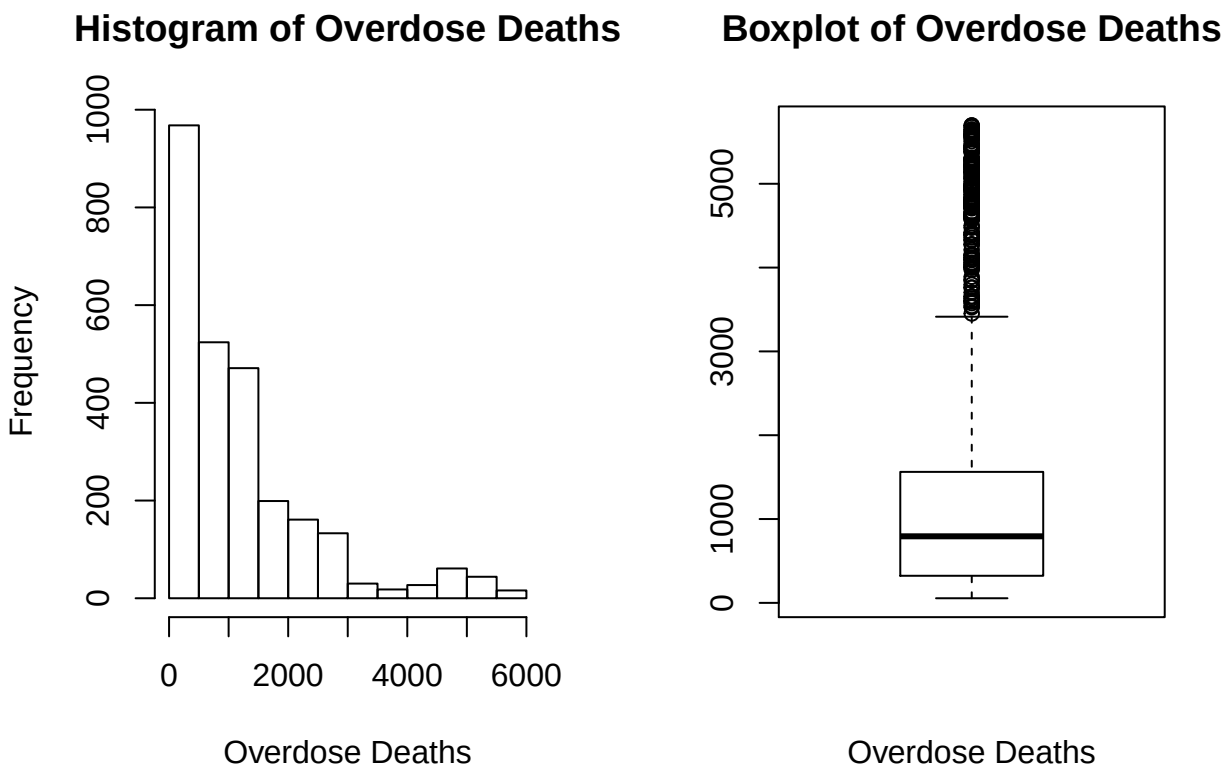
Import and clean unemployment data.

```
# iterate through state data files
unemployment <- data.frame()
```

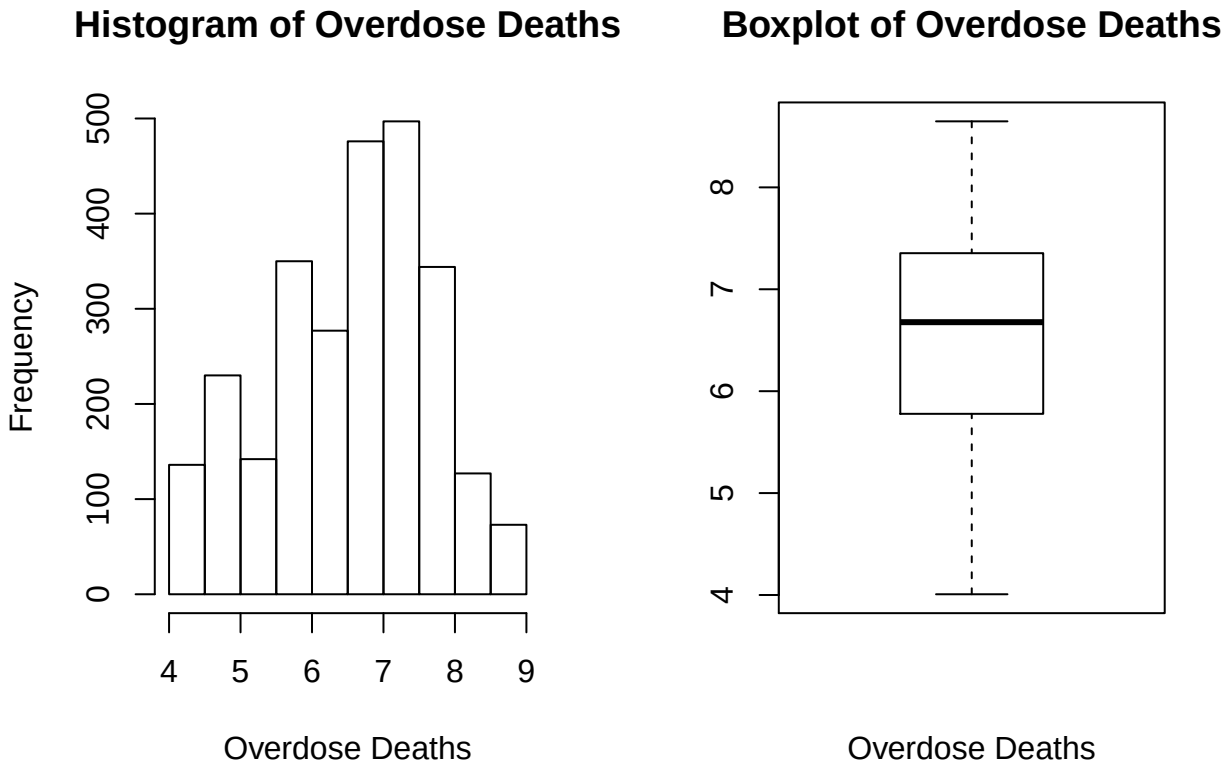
```r
for (file in list.files("data/state", full.names = T)) {

  # state name and data
  state <- substr(basename(file), 1, 2)
  data <- read.csv(file)
  data$state <- rep(state, nrow(data))

  # year and month
  data$year <- sapply(data$DATE, function(x) as.numeric(substr(x, 1, 4)))
  data <- data[data$year >= 2015,]
  month <- sapply(data$DATE, function(x) as.numeric(substr(x, 6, 7)))
  month <- ordered(month, labels = months.labels)
  data$month <- month
  colnames(data)[2] <- "unemployment"

  # record state data
  unemployment <- rbind(unemployment, data)
}
colnames(unemployment)[2] <- "unemployment"
unemployment <- unemployment[,-1] # drop default date column

# merge datasets
overdose <- merge(overdose, unemployment, by = c("state", "year", "month"))

# order
overdose <- overdose[order(overdose$year, overdose$state, overdose$month),]
```

EDA

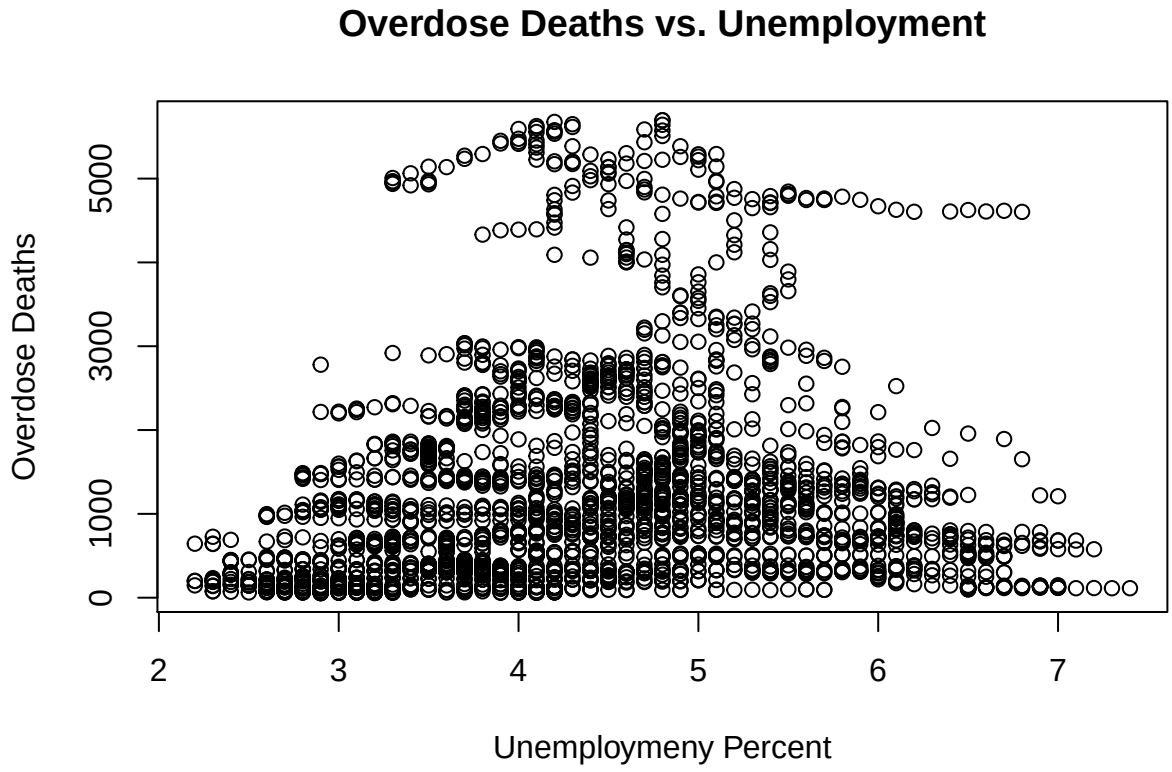
```r
# histogram and boxplot of response
par(mfrow = c(1, 2))
hist(overdose$overdoseDeaths, main = "Histogram of Overdose Deaths",
     xlab = "Overdose Deaths")
boxplot(overdose$overdoseDeaths, main = "Boxplot of Overdose Deaths",
        xlab = "Overdose Deaths")
```




```
# histogram and boxplot of log-response
par(mfrow = c(1, 2))
hist(log(overdose$overdoseDeaths), main = "Histogram of Overdose Deaths",
     xlab = "Overdose Deaths")
boxplot(log(overdose$overdoseDeaths), main = "Boxplot of Overdose Deaths",
        xlab = "Overdose Deaths")
```




```r
# response vs. unemployment
par(mfrow = c(1, 1))
plot(overdoseDeaths ~ unemployment, data = overdose,
     main = "Overdose Deaths vs. Unemployment",
     xlab = "Unemploymeny Percent", ylab = "Overdose Deaths")
```



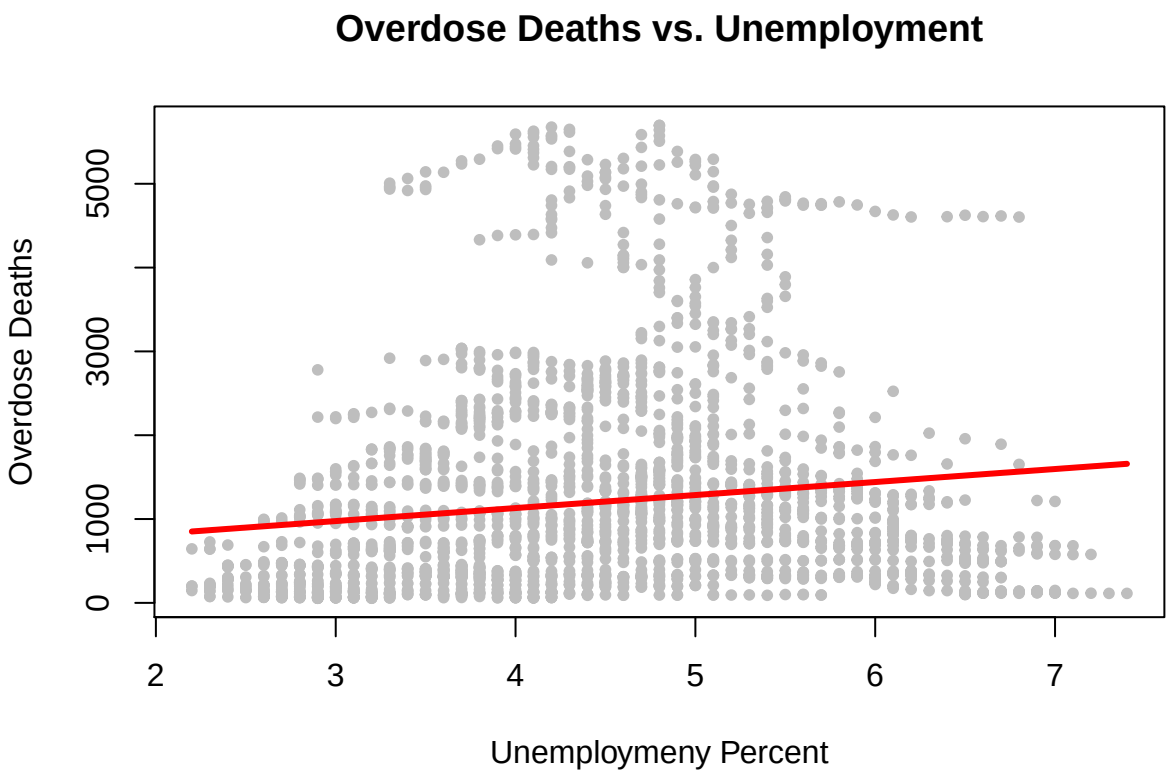
Build baseline model.

```
# simple linear model
summary(lm1 <- lm(overdoseDeaths ~ unemployment, data = overdose))
```

```
##
## Call:
## lm(formula = overdoseDeaths ~ unemployment, data = overdose)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1545.9  -814.3  -374.3   366.3  4513.9
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    510.00      96.31   5.295 1.28e-07 ***
## unemployment   155.25      21.36   7.267 4.81e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1200 on 2650 degrees of freedom
## Multiple R-squared:  0.01954,    Adjusted R-squared:  0.01917
## F-statistic: 52.81 on 1 and 2650 DF,  p-value: 4.807e-13
```

The simple regression model has a positive coefficient for unemployment (155.25). With a t-statistic of 7.267 (p-value < 0.0001), this coefficient is very significant. The model has a positive association between unemployment and overdose deaths.

```
# observations with simple regression line
plot(overdoseDeaths ~ unemployment, data = overdose,
     main = "Overdose Deaths vs. Unemployment",
     xlab = "Unemploymeny Percent", ylab = "Overdose Deaths",
     col = "grey", pch = 20)
x <- seq(min(overdose$unemployment), max(overdose$unemployment), 0.01)
y <- predict(lm1, newdata = data.frame(unemployment = x))
lines(y ~ x, col = "red", lwd = 3)
```


Overdose Deaths vs. Unemployment
Overdose Deaths
0
1000
3000
5000
2
3
4
5
6
7
Unemploymeny Percent