day_1_oct_29_2016_occupancy_eda

I am going to explore this dataset: http://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+#

The occupancy dataset has various environmental readings from an office, as well as whether the office is occupied. The occupancy reading is derived from photos taken by the office camera.

```
df <- read.csv("~/Downloads/databank/occupancy_data/datatest.txt")
df$date <- as.POSIXct(df$date)
head(df)
```

```
##                     date Temperature Humidity    Light      CO2
## 140 2015-02-02 14:19:00     23.7000   26.272 585.2000 749.2000
## 141 2015-02-02 14:19:59     23.7180   26.290 578.4000 760.4000
## 142 2015-02-02 14:21:00     23.7300   26.230 572.6667 769.6667
## 143 2015-02-02 14:22:00     23.7225   26.125 493.7500 774.7500
## 144 2015-02-02 14:23:00     23.7540   26.200 488.6000 779.0000
## 145 2015-02-02 14:23:59     23.7600   26.260 568.6667 790.0000
##     HumidityRatio Occupancy
## 140   0.004764163         1
## 141   0.004772661         1
## 142   0.004765153         1
## 143   0.004743773         1
## 144   0.004766594         1
## 145   0.004779332         1
```

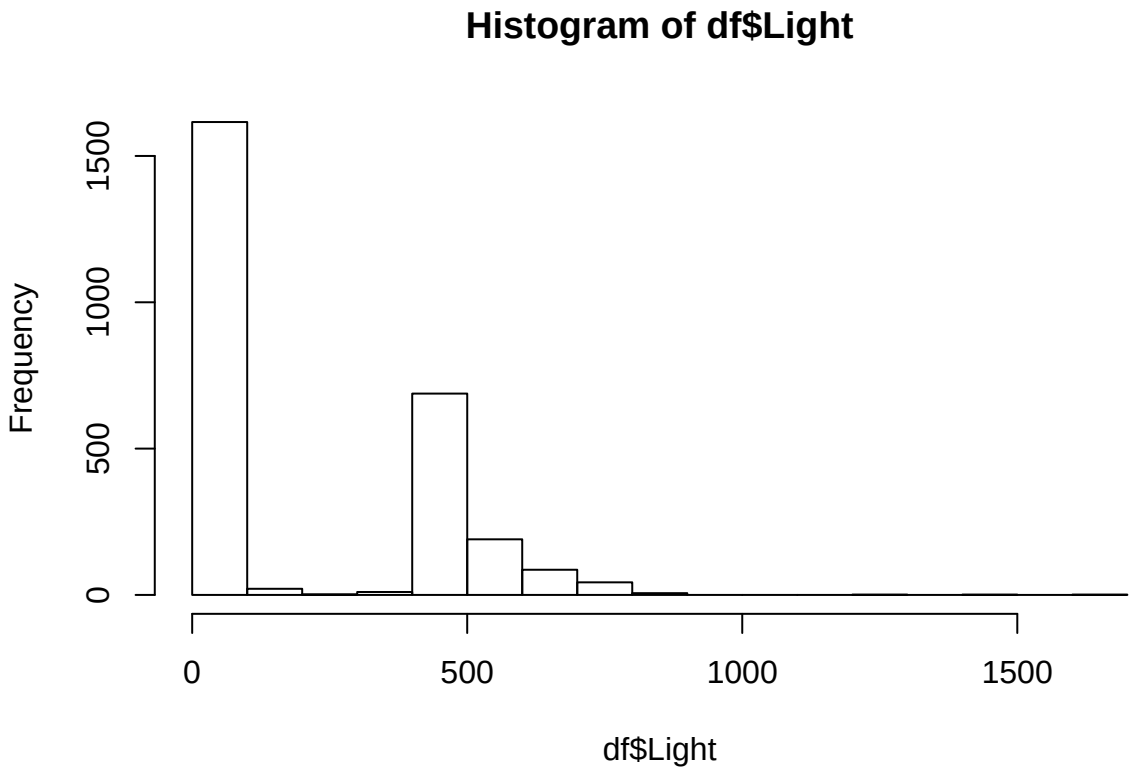
```
summary(df)
```

```
##       date                       Temperature       Humidity
##  Min.   :2015-02-02 14:19:00   Min.   :20.20   Min.   :22.10
##  1st Qu.:2015-02-03 01:25:00   1st Qu.:20.65   1st Qu.:23.26
##  Median :2015-02-03 12:30:59   Median :20.89   Median :25.00
##  Mean   :2015-02-03 12:30:59   Mean   :21.43   Mean   :25.35
##  3rd Qu.:2015-02-03 23:37:00   3rd Qu.:22.36   3rd Qu.:26.86
##  Max.   :2015-02-04 10:43:00   Max.   :24.41   Max.   :31.47
##      Light             CO2          HumidityRatio        Occupancy
##  Min.   :   0.0   Min.   : 427.5   Min.   :0.003303   Min.   :0.0000
##  1st Qu.:   0.0   1st Qu.: 466.0   1st Qu.:0.003529   1st Qu.:0.0000
##  Median :   0.0   Median : 580.5   Median :0.003815   Median :0.0000
##  Mean   : 193.2   Mean   : 717.9   Mean   :0.004027   Mean   :0.3647
##  3rd Qu.: 442.5   3rd Qu.: 956.3   3rd Qu.:0.004532   3rd Qu.:1.0000
##  Max.   :1697.2   Max.   :1402.2   Max.   :0.005378   Max.   :1.0000
```

We see that for 36.5% of the samples, the office was occupied during 2015-02-02 to 2015-02-04.

Lights

```
hist(df$Light)
```



Light is in Lux. The data looks very bimodal. Obviously when the lights are off, there are no one in the office, and vise versa.

```
nrow(df[df$Light != 0,])
```

```
## [1] 1050
```

```
nrow(df[df$Light == 0,])
```

```
## [1] 1615
```

```
cor(df$Light, df$Occupancy)
```

```
## [1] 0.9279491
```

Wow, r=0.93 is a lot... So When it's dark, there is usually no one in the office?

```
summary(df[df$Light == 0,]$Occupancy)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0       0       0       0
```

```
summary(df[df$Light != 0,]$Occupancy)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  1.0000  1.0000  0.9257  1.0000  1.0000
```

This says that when lights are off, there is for sure no one in the office. But when the lights are on, 7% of the time, there are no one in the office. Soo.... we have some energy wasters in the office....