Kaggle Project

Bridget, Eva, and Annie

November 18, 2017

```r
# read in the data
train_data=read.csv(file="/home/bridget/Dropbox/STATS202/kaggle/train.csv",header=T)
test_data=read.csv(file="/home/bridget/Dropbox/STATS202/kaggle/test.csv",header=T)

# check correlations
cor(train_data)
```

```
##                         ID  Rel.Compact Surface.Area    Wall.Area
## ID             1.000000000 -0.079452294  0.079556731  0.050111914
## Rel.Compact   -0.079452294  1.000000000 -0.991920018 -0.188572478
## Surface.Area   0.079556731 -0.991920018  1.000000000  0.177694681
## Wall.Area      0.050111914 -0.188572478  0.177694681  1.000000000
## Roof.Area      0.052340236 -0.864831757  0.877916558 -0.315192553
## Height        -0.050350469  0.824310797 -0.855434592  0.304627802
## Orientation   -0.063890529  0.024387804 -0.028384160 -0.018812858
## Glazing.Area   0.035542350 -0.005116515  0.002566042 -0.002268867
## Glazing.Distr -0.004501353 -0.006858788  0.007691487 -0.017346230
## Outcome       -0.037014518  0.613645457 -0.651401903  0.477015505
##                 Roof.Area       Height  Orientation Glazing.Area
## ID             0.05234024 -0.050350469 -0.063890529  0.035542350
## Rel.Compact   -0.86483176  0.824310797  0.024387804 -0.005116515
## Surface.Area   0.87791656 -0.855434592 -0.028384160  0.002566042
## Wall.Area     -0.31519255  0.304627802 -0.018812858 -0.002268867
## Roof.Area      1.00000000 -0.973178598 -0.018219452  0.003578560
## Height        -0.97317860  1.000000000  0.015269106 -0.001719509
## Orientation   -0.01821945  0.015269106  1.000000000  0.005121389
## Glazing.Area   0.00357856 -0.001719509  0.005121389  1.000000000
## Glazing.Distr  0.01585741 -0.017802456  0.009653998  0.218478751
## Outcome       -0.86029097  0.888969806  0.008817138  0.269249436
##               Glazing.Distr      Outcome
## ID             -0.004501353 -0.037014518
## Rel.Compact    -0.006858788  0.613645457
## Surface.Area    0.007691487 -0.651401903
## Wall.Area      -0.017346230  0.477015505
## Roof.Area       0.015857406 -0.860290971
## Height         -0.017802456  0.888969806
## Orientation     0.009653998  0.008817138
## Glazing.Area    0.218478751  0.269249436
## Glazing.Distr   1.000000000  0.071155433
## Outcome         0.071155433  1.000000000
```

```r
# change height and orientation variables to categorical variables
train_data$Height=as.factor(train_data$Height)
train_data$Orientation=as.factor(train_data$Orientation)

# remove ID variable since it just labels the rows
# remove relative compactness because it is linearly correlated with surface area
# remove surface area because it is equal to wall area + 2 *(roof area)
```

```r
summary(lm(Surface.Area~Wall.Area+Roof.Area,data=train_data))
```

```
##
## Call:
## lm(formula = Surface.Area ~ Wall.Area + Roof.Area, data = train_data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.348e-11 -3.090e-13 -9.900e-14 -1.800e-14  1.003e-10
##
## Coefficients:
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 6.671e-12  1.528e-12 4.366e+00 1.47e-05 ***
## Wall.Area   1.000e+00  3.736e-15 2.677e+14  < 2e-16 ***
## Roof.Area   2.000e+00  3.635e-15 5.502e+14  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.993e-12 on 647 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.563e+29 on 2 and 647 DF,  p-value: < 2.2e-16
```

```r
names(train_data)
```

```
##  [1] "ID"            "Rel.Compact"   "Surface.Area"  "Wall.Area"
##  [5] "Roof.Area"     "Height"        "Orientation"   "Glazing.Area"
##  [9] "Glazing.Distr" "Outcome"
```

```r
train_data=train_data[,-c(1,2,3)]

# linear regression on remaining data
lm.fit=lm(Outcome~.,data=train_data)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = Outcome ~ ., data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7083 -1.5954  0.2048  1.5287  7.6019
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -16.390057   2.748369  -5.964 4.08e-09 ***
## Wall.Area       0.053243   0.002819  18.889  < 2e-16 ***
## Roof.Area       0.036890   0.011353   3.249  0.00122 **
## Height7        19.896499   1.026926  19.375  < 2e-16 ***
## Orientation3    0.230632   0.333258   0.692  0.48916
## Orientation4   -0.123776   0.333211  -0.371  0.71041
## Orientation5    0.109721   0.337915   0.325  0.74552
## Glazing.Area   19.932986   0.902111  22.096  < 2e-16 ***
## Glazing.Distr   0.211373   0.077826   2.716  0.00679 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3.011 on 641 degrees of freedom
## Multiple R-squared:  0.9132, Adjusted R-squared:  0.9122
## F-statistic: 843.5 on 8 and 641 DF,  p-value: < 2.2e-16
```