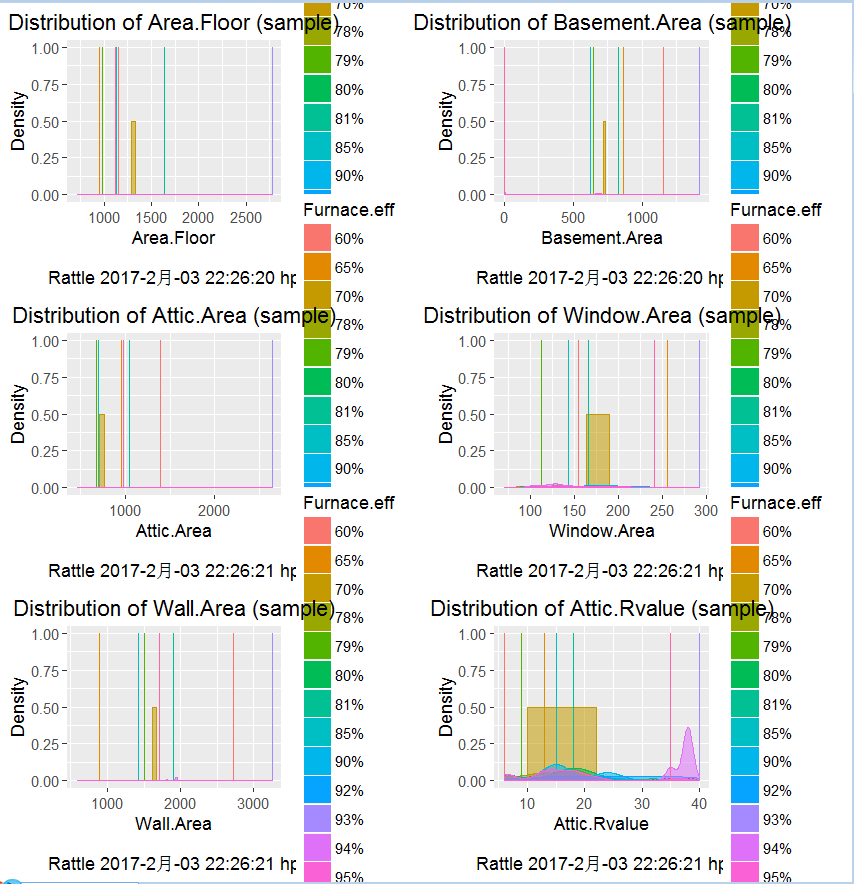
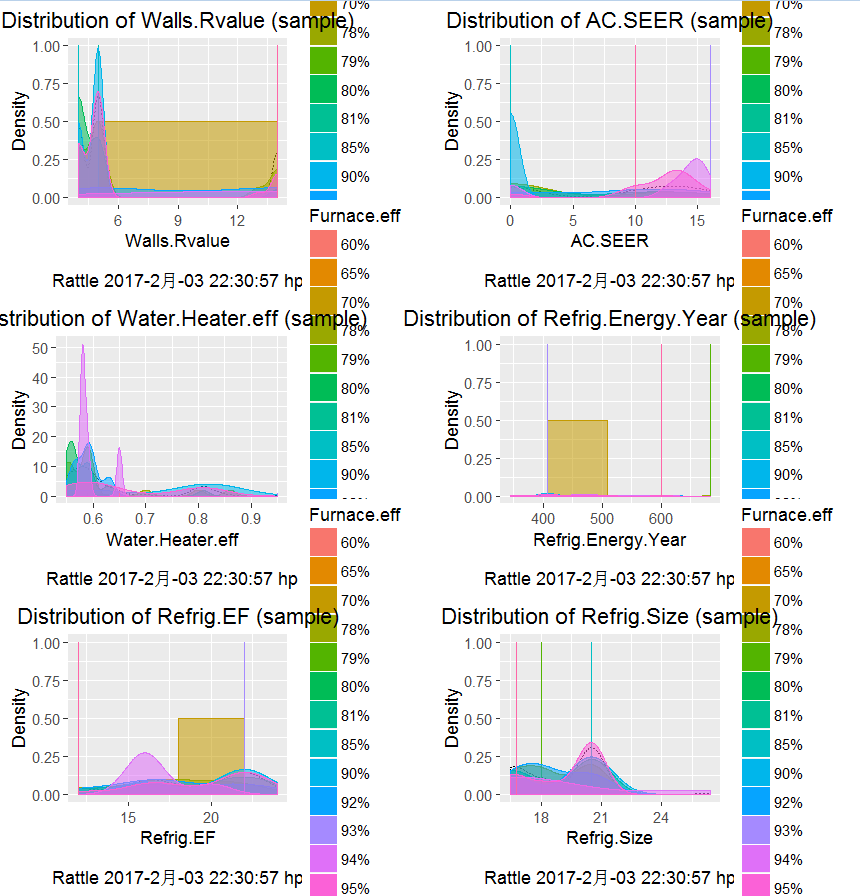
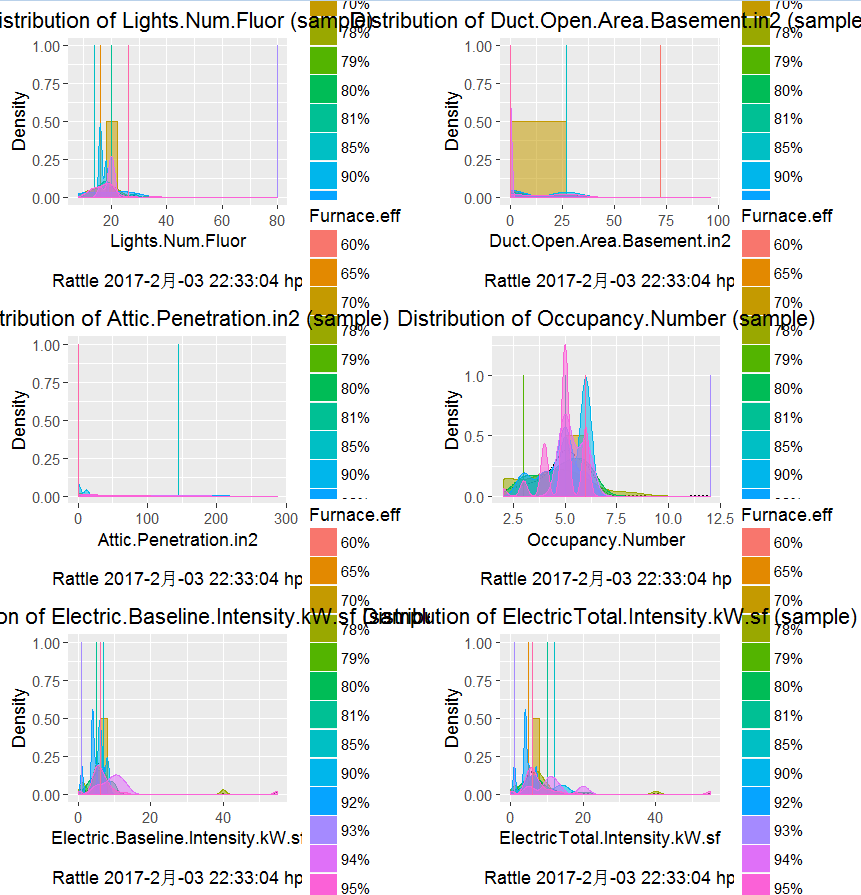
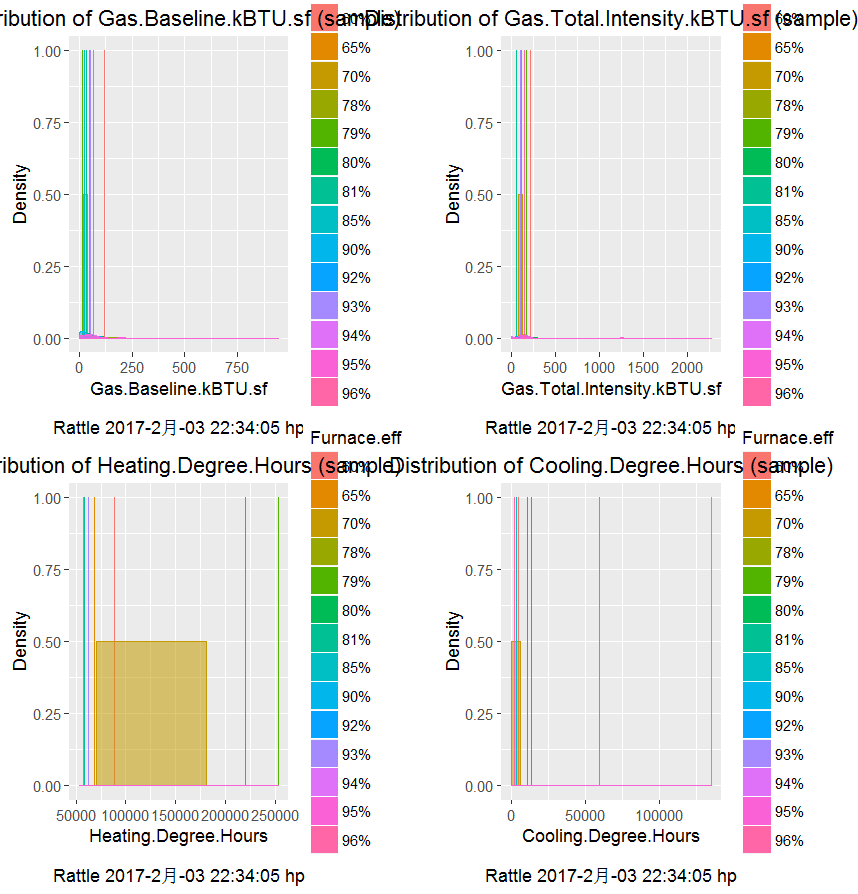
1. **Histograms**
2. Develop Histograms for all of the UD housing characteristics.
3. Comment based upon the histogram which of the factors should be excluded if developing a model to predict the Gas Heating Intensity factor. Your judgement should be based upon factors that don't have great variance. These factors ultimately shouldn't be included in the model.





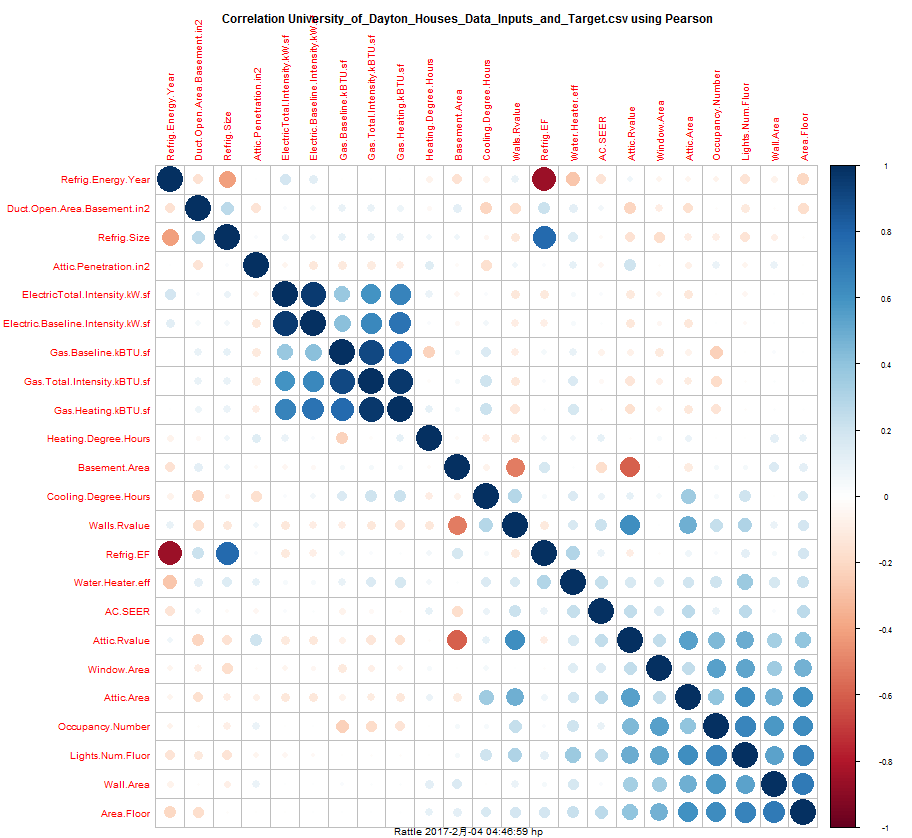
**Data with only 1 or 2 values are removed.**

**2.  Input Factor Correlation**

Ignoring the factors: House Number and Address, and the factors you’ve chosen to eliminate in Problem 1, develop a Pearson pairwise correlation for all input factors. Note: you could use Rattle to do this more easily. To use Rattle you have to install the Rattle package.  See recording attached.

Note: Variables with no variation should be removed.

Identify the factors which have a high degree of correlation (>0.9 and 0.95).



**3.Removing Highly Correlated Factors**

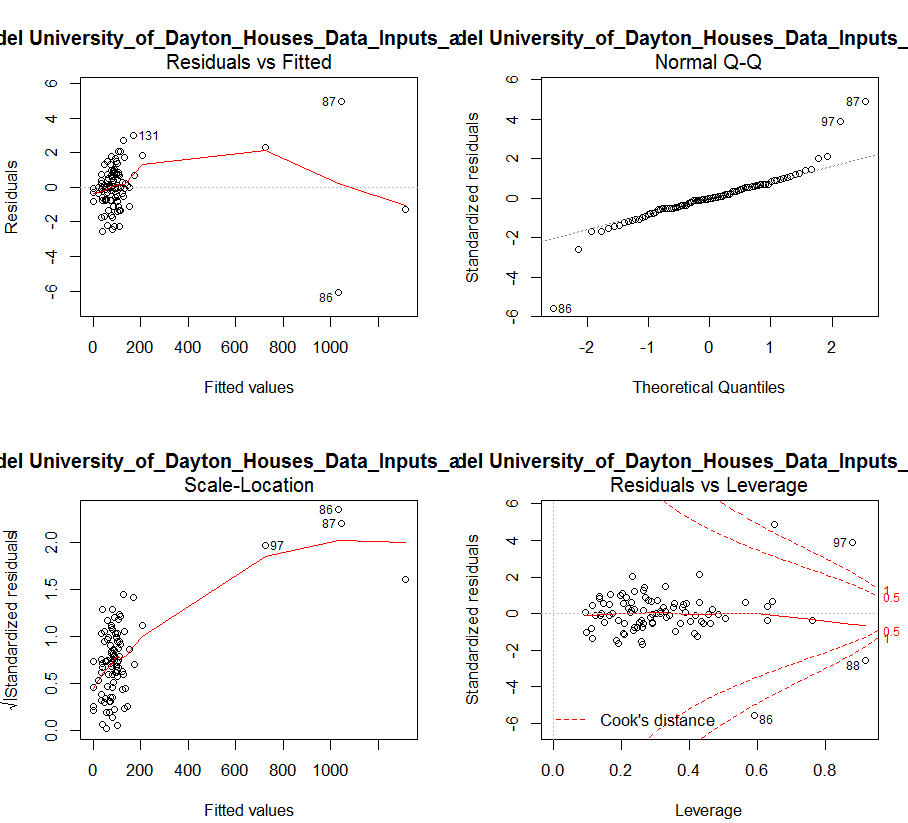
Create data frame that includes the input factors considered in #2 (e.g., with the noted excluded factors). Create new data frames with highly correlated factors removed (for correlation > 0.9, and correlation > 0.95). Print the header rows of each of the created data frames. Copy into Excel, problem 3 tab.

**4.Removing Multi-Colinearity**

Using the remaining input factors from #3, with factors having correlation > 0.95, Identify the factors (if any) that should be retained after removal for co-linearity. Create a new data frame based on these factors. Note: that these should be the factors used to develop a model from.

**5.incipal Component Regression**

Using the remaining input factors from #4, along with the target Gas. Heat. kBTU. sf, first identify how many principal components you need to account for 95% of the variation. Also, develop a principal component regression. Plot the R2 as a function of number of principal components. Lastly, comment upon how many principal components you should use to develop a decent linear model.



*Only 3 factors affect the prediction evidently.*

Summary of the Linear Regression model (built using lm):

**From the model of Rattle:**

**Call:**

lm(formula = Gas.Heating.kBTU.sf ~ ., data = crs$dataset[crs$train,

c(crs$input, crs$target)])

Residuals:

Min 1Q Median 3Q Max

-6.0971 -0.7170 0.0000 0.7257 4.9482

Coefficients:

Estimate Std. Error t value

(Intercept) 4.871542454 10.609314213 0.459

Area.Floor 0.004103349 0.001358372 3.021

Basement.Area 0.000102456 0.001705927 0.060

Attic.Area -0.000634479 0.001253281 -0.506

Window.Area -0.002662491 0.006400096 -0.416

Wall.Area -0.001206731 0.000848644 -1.422

Attic.Rvalue 0.008265327 0.042929776 0.193

Walls.Rvalue 0.149178393 0.076304232 1.955

Furnace.eff65% 2.374915176 3.006390203 0.790

Furnace.eff70% -0.937120481 2.667610542 -0.351

Furnace.eff78% 0.555231362 2.457760596 0.226

Furnace.eff79% 3.380292776 3.431191434 0.985

Furnace.eff80% 0.546254709 2.410054895 0.227

Furnace.eff81% 1.437024980 4.735175829 0.303

Furnace.eff85% 0.577714768 3.005394629 0.192

Furnace.eff90% 0.421569187 2.598674967 0.162

Furnace.eff92% 0.002511592 2.532366329 0.001

Furnace.eff93% 4.092656258 4.379878150 0.934

Furnace.eff94% -0.706052224 2.940897839 -0.240

Furnace.eff95% 2.365763474 2.426163974 0.975

Furnace.eff96% 1.916521079 3.097913003 0.619

AC.SEER -0.144091692 0.044446988 -3.242

Water.Heater.eff -4.524041441 2.580927420 -1.753

Refrig.Energy.Year -0.011163848 0.019446763 -0.574

Refrig.EF -0.411850575 0.559794577 -0.736

Refrig.Size 0.354955750 0.536432172 0.662

Lights.Num.Fluor -0.002608492 0.057422676 -0.045

Duct.Open.Area.Basement.in2 -0.005633179 0.011991649 -0.470

Attic.Penetration.in2 0.004718984 0.003770999 1.251

Occupancy.Number -0.163871720 0.272577925 -0.601

Electric.Baseline.Intensity.kW.sf 0.205994444 0.115123298 1.789

ElectricTotal.Intensity.kW.sf -0.043301668 0.097167937 -0.446

Gas.Baseline.kBTU.sf -0.947667577 0.006054398 -156.525

Gas.Total.Intensity.kBTU.sf 0.974046659 0.002836206 343.433

Heating.Degree.Hours 0.000009534 0.000005241 1.819

Cooling.Degree.Hours -0.000001041 0.000005014 -0.208

Pr(>|t|)

(Intercept) 0.64769

Area.Floor 0.00364 \*\*

Basement.Area 0.95230

Attic.Area 0.61445

Window.Area 0.67882

Wall.Area 0.15997

Attic.Rvalue 0.84795

Walls.Rvalue 0.05502 .

Furnace.eff65% 0.43252

Furnace.eff70% 0.72654

Furnace.eff78% 0.82200

Furnace.eff79% 0.32831

Furnace.eff80% 0.82142

Furnace.eff81% 0.76253

Furnace.eff85% 0.84818

Furnace.eff90% 0.87165

Furnace.eff92% 0.99921

Furnace.eff93% 0.35365

Furnace.eff94% 0.81105

Furnace.eff95% 0.33324

Furnace.eff96% 0.53838

AC.SEER 0.00190 \*\*

Water.Heater.eff 0.08449 .

Refrig.Energy.Year 0.56796

Refrig.EF 0.46463

Refrig.Size 0.51058

Lights.Num.Fluor 0.96391

Duct.Open.Area.Basement.in2 0.64015

Attic.Penetration.in2 0.21542

Occupancy.Number 0.54987

Electric.Baseline.Intensity.kW.sf 0.07837 .

ElectricTotal.Intensity.kW.sf 0.65739

Gas.Baseline.kBTU.sf < 2e-16 \*\*\*

Gas.Total.Intensity.kBTU.sf < 2e-16 \*\*\*

Heating.Degree.Hours 0.07364 .

Cooling.Degree.Hours 0.83624

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.709 on 63 degrees of freedom

Multiple R-squared: 1, Adjusted R-squared: 0.9999

F-statistic: 3.673e+04 on 35 and 63 DF, p-value: < 2.2e-16

==== ANOVA ====

Analysis of Variance Table

Response: Gas.Heating.kBTU.sf

Df Sum Sq Mean Sq F value Pr(>F)

Area.Floor 1 330 330 113.1646 1.061e-15

Basement.Area 1 4602 4602 1576.0636 < 2.2e-16

Attic.Area 1 89591 89591 30680.2432 < 2.2e-16

Window.Area 1 22200 22200 7602.4210 < 2.2e-16

Wall.Area 1 3223 3223 1103.7558 < 2.2e-16

Attic.Rvalue 1 114335 114335 39153.7094 < 2.2e-16

Walls.Rvalue 1 9835 9835 3368.1202 < 2.2e-16

Furnace.eff 13 191364 14720 5040.9218 < 2.2e-16

AC.SEER 1 9371 9371 3209.1986 < 2.2e-16

Water.Heater.eff 1 181111 181111 62020.8634 < 2.2e-16

Refrig.Energy.Year 1 24596 24596 8422.6610 < 2.2e-16

Refrig.EF 1 2380 2380 814.9756 < 2.2e-16

Refrig.Size 1 42795 42795 14655.0795 < 2.2e-16

Lights.Num.Fluor 1 26487 26487 9070.4543 < 2.2e-16

Duct.Open.Area.Basement.in2 1 2865 2865 980.9838 < 2.2e-16

Attic.Penetration.in2 1 18997 18997 6505.4990 < 2.2e-16

Occupancy.Number 1 113569 113569 38891.4719 < 2.2e-16

Electric.Baseline.Intensity.kW.sf 1 1634182 1634182 559620.5838 < 2.2e-16

ElectricTotal.Intensity.kW.sf 1 97006 97006 33219.4349 < 2.2e-16

Gas.Baseline.kBTU.sf 1 579019 579019 198283.1794 < 2.2e-16

Gas.Total.Intensity.kBTU.sf 1 585718 585718 200577.2955 < 2.2e-16

Heating.Degree.Hours 1 11 11 3.7317 0.05789

Cooling.Degree.Hours 1 0 0 0.0431 0.83624

Residuals 63 184 3

set.seed(crv$seed)

crs$nobs <- nrow(crs$dataset) # 142 observations

crs$sample <- crs$train <- sample(nrow(crs$dataset), 0.7\*crs$nobs) # 99 observations

crs$validate <- sample(setdiff(seq\_len(nrow(crs$dataset)), crs$train), 0.15\*crs$nobs) # 21 observations

crs$test <- setdiff(setdiff(seq\_len(nrow(crs$dataset)), crs$train), crs$validate) # 22 observations

# The following variable selections have been noted.

crs$input <- c("House.Number", "Address", "Area.Floor", "Basement.Area",

"Attic.Area", "Window.Area", "Wall.Area", "Attic.Rvalue",

"Walls.Rvalue", "Furnace.eff", "AC.SEER", "Water.Heater.eff",

"Refrig.Energy.Year", "Refrig.EF", "Refrig.Size", "Lights.Num.Fluor",

"Water.Heater.On.Summer", "AC.On.Summer", "Furnace.Pilot.On.Summer", "Refrig.On.Summer",

"Duct.Open.Area.Basement.in2", "Attic.Penetration.in2", "Occupancy.Number", "Electric.Baseline.Intensity.kW.sf",

"ElectricTotal.Intensity.kW.sf", "Gas.Baseline.kBTU.sf", "Gas.Heating.kBTU.sf", "Gas.Total.Intensity.kBTU.sf",

"Heating.Degree.Hours", "Cooling.Degree.Hours")

crs$numeric <- c("House.Number", "Area.Floor", "Basement.Area", "Attic.Area",

"Window.Area", "Wall.Area", "Attic.Rvalue", "Walls.Rvalue",

"AC.SEER", "Water.Heater.eff", "Refrig.Energy.Year", "Refrig.EF",

"Refrig.Size", "Lights.Num.Fluor", "Water.Heater.On.Summer", "AC.On.Summer",

"Furnace.Pilot.On.Summer", "Refrig.On.Summer", "Duct.Open.Area.Basement.in2", "Attic.Penetration.in2",

"Occupancy.Number", "Electric.Baseline.Intensity.kW.sf", "ElectricTotal.Intensity.kW.sf", "Gas.Baseline.kBTU.sf",

"Gas.Heating.kBTU.sf", "Gas.Total.Intensity.kBTU.sf", "Heating.Degree.Hours", "Cooling.Degree.Hours")

crs$categoric <- c("Address", "Furnace.eff")

crs$target <- "Windows.Rvalue"

crs$risk <- NULL

crs$ident <- NULL

crs$ignore <- c("Basement.Rvalue", "Lights.Watts.bulb", "Lights.Operating.Hours")

crs$weights <- NULL

#============================================================

# Rattle timestamp: 2017-02-02 17:45:34 x86\_64-w64-mingw32

# Note the user selections.

# Build the training/validate/test datasets.

**Rattle log export comment**

set.seed(crv$seed)

crs$nobs <- nrow(crs$dataset) # 142 observations

crs$sample <- crs$train <- sample(nrow(crs$dataset), 0.7\*crs$nobs) # 99 observations

crs$validate <- sample(setdiff(seq\_len(nrow(crs$dataset)), crs$train), 0.15\*crs$nobs) # 21 observations

crs$test <- setdiff(setdiff(seq\_len(nrow(crs$dataset)), crs$train), crs$validate) # 22 observations

# The following variable selections have been noted.

crs$input <- c("Area.Floor", "Basement.Area", "Attic.Area", "Window.Area",

"Wall.Area", "Attic.Rvalue", "Walls.Rvalue", "Furnace.eff",

"AC.SEER", "Water.Heater.eff", "Refrig.Energy.Year", "Refrig.EF",

"Refrig.Size", "Lights.Num.Fluor", "Duct.Open.Area.Basement.in2", "Attic.Penetration.in2",

"Occupancy.Number", "Electric.Baseline.Intensity.kW.sf", "ElectricTotal.Intensity.kW.sf", "Gas.Baseline.kBTU.sf",

"Gas.Total.Intensity.kBTU.sf", "Heating.Degree.Hours", "Cooling.Degree.Hours")

crs$numeric <- c("Area.Floor", "Basement.Area", "Attic.Area", "Window.Area",

"Wall.Area", "Attic.Rvalue", "Walls.Rvalue", "AC.SEER",

"Water.Heater.eff", "Refrig.Energy.Year", "Refrig.EF", "Refrig.Size",

"Lights.Num.Fluor", "Duct.Open.Area.Basement.in2", "Attic.Penetration.in2", "Occupancy.Number",

"Electric.Baseline.Intensity.kW.sf", "ElectricTotal.Intensity.kW.sf", "Gas.Baseline.kBTU.sf", "Gas.Total.Intensity.kBTU.sf",

"Heating.Degree.Hours", "Cooling.Degree.Hours")

crs$categoric <- "Furnace.eff"

crs$target <- "Gas.Heating.kBTU.sf"

crs$risk <- NULL

crs$ident <- NULL

crs$ignore <- c("House.Number", "Address", "Basement.Rvalue", "Windows.Rvalue", "Lights.Watts.bulb", "Lights.Operating.Hours", "Water.Heater.On.Summer", "AC.On.Summer", "Furnace.Pilot.On.Summer", "Refrig.On.Summer")

crs$weights <- NULL

# Generate a correlation plot for the variables.

# The 'corrplot' package provides the 'corrplot' function.

library(corrplot, quietly=TRUE)

# Correlations work for numeric variables only.

crs$cor <- cor(crs$dataset[crs$sample, crs$numeric], use="pairwise", method="pearson")

# Order the correlations by their strength.

crs$ord <- order(crs$cor[1,])

crs$cor <- crs$cor[crs$ord, crs$ord]

# Display the actual correlations.

print(crs$cor)

# Graphically display the correlations.

corrplot(crs$cor, mar=c(0,0,1,0))

title(main="Correlation University\_of\_Dayton\_Houses\_Data\_Inputs\_and\_Target.csv using Pearson",

sub=paste("Rattle", format(Sys.time(), "%Y-%b-%d %H:%M:%S"), Sys.info()["user"]))

# Regression model

# Build a Regression model.

crs$glm <- lm(Gas.Heating.kBTU.sf ~ ., data=crs$dataset[crs$train,c(crs$input, crs$target)])