Feature and Target Engineering

Data Set

h2o

```
ames <- AmesHousing::make_ames()
ames.h2o <- as.h2o(ames)
```

stratified (*Sale_Price*) training sample

```
set.seed(123)

split <- initial_split(ames, prop = 0.7,
                       strata = "Sale_Price")

ames_train <- training(split)
ames_test <- testing(split)
```

log transformation (Sale_Price)

```
ames_recipe <- recipe(Sale_Price ~ ., data = ames_train) %>%
   step_log(all_outcomes())

ames_recipe
```

```
Data Recipe

Inputs:

      role #variables
   outcome          1
 predictor         80

Operations:

Log transformation on all_outcomes
```

Box-Cox transformation (example)

```
lambda <- 3

y <- forecast::BoxCox(10, lambda)

inv_box_cox <- function(x, lambda) {
   # for Box-Cox, lambda = 0 -> log transform
   if(lambda == 0) exp(x) else (lambda*x + 1)^(1/lambda)
```

```
}

inv_box_cox(y, lambda)
```

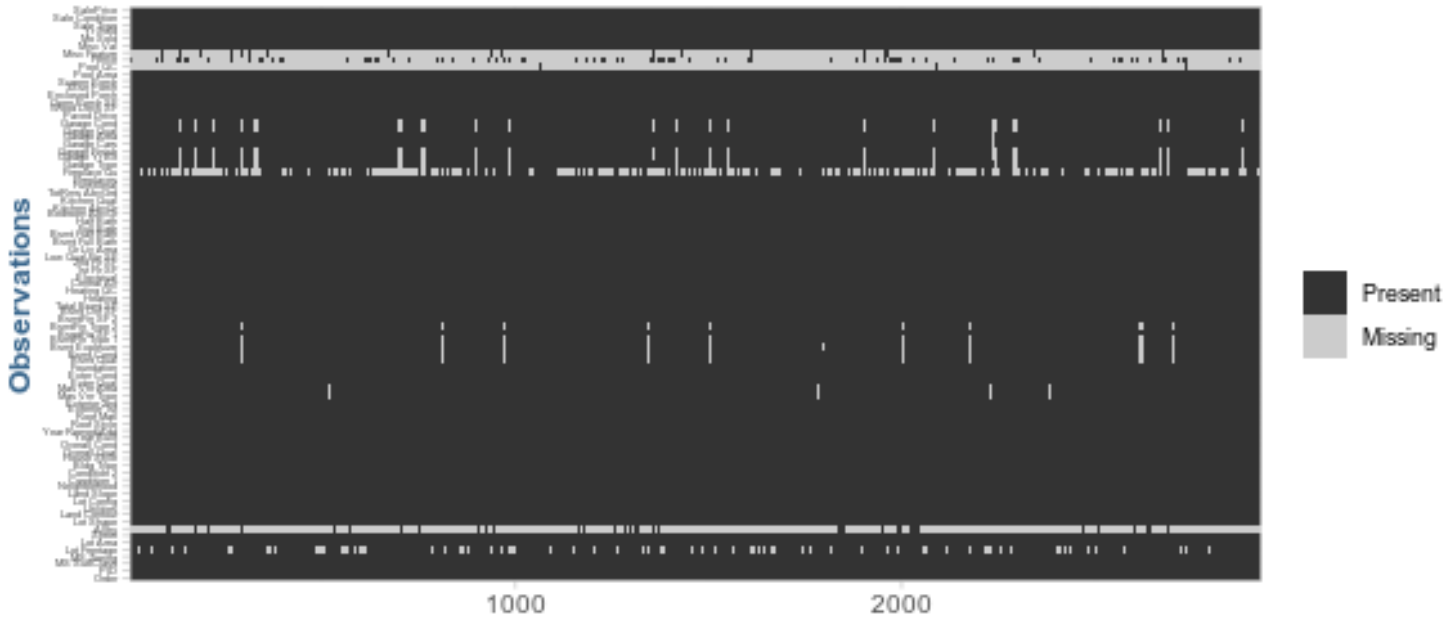
```
[1] 10
attr(,"lambda")
[1] 3
```

Missing Values

```
sum(is.na(AmesHousing::ames_raw))
```

```
[1] 13997
```

```
AmesHousing::ames_raw %>%
   is.na() %>%
   reshape2::melt() %>%
   ggplot(aes(Var2, Var1, fill = value)) +
      geom_raster() +
      coord_flip() +
      scale_y_continuous(NULL, expand = c(0,0)) +
      scale_fill_grey(name = "",
                      labels = c("Present",
                                 "Missing")) +
   xlab("Observations") +
   theme(axis.text.y = element_text(size = 4))
```



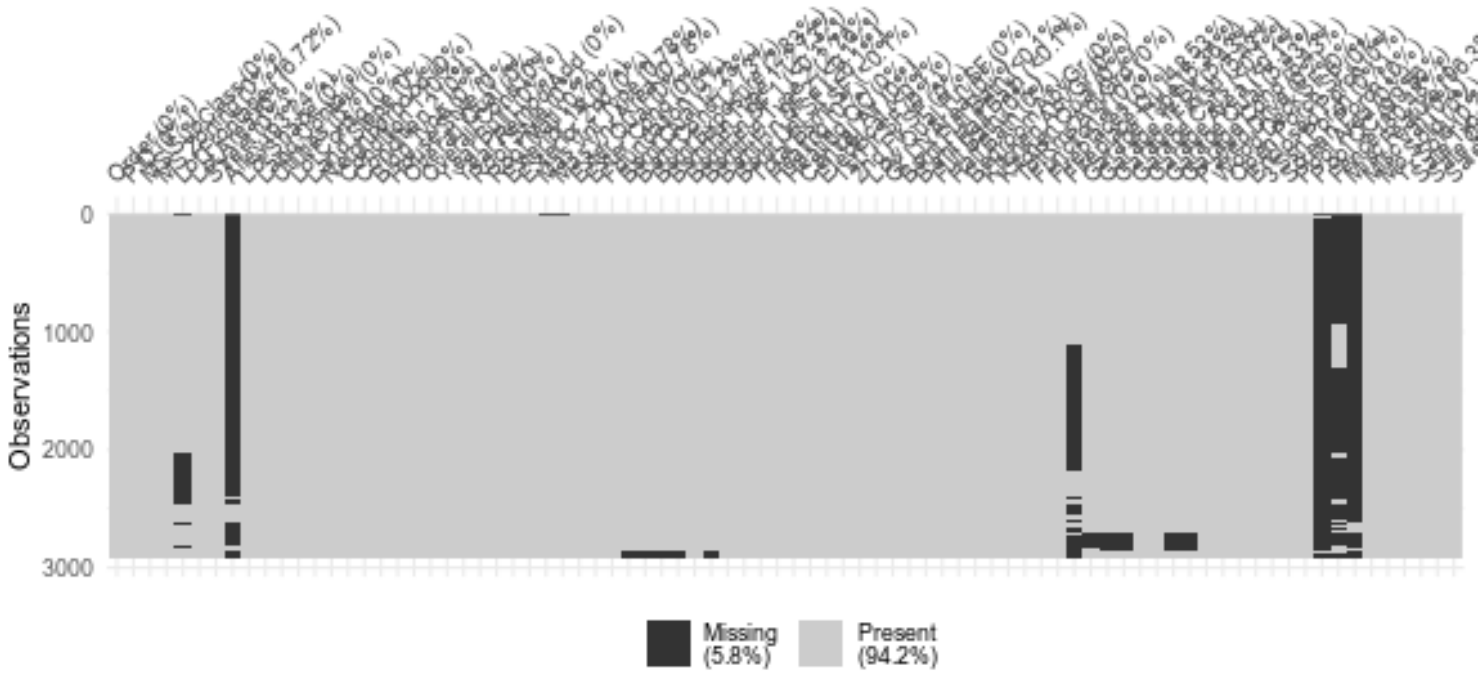
Missing Garage?

```
AmesHousing::ames_raw %>%
   filter(is.na('Garage Type')) %>%
   select(starts_with("Garage"))
```

```
# A tibble: 0 x 7
# … with 7 variables: `Garage Type` <chr>, `Garage Yr Blt` <int>, `Garage
#   Finish` <chr>, `Garage Cars` <int>, `Garage Area` <int>, `Garage
#   Qual` <chr>, `Garage Cond` <chr>
```

Missing values w/cluster (*visdat*)

```
vis_miss(AmesHousing::ames_raw, cluster = T)
```



Missing Value Imputation

basic descriptive statistic

```
ames_recipe %>%
   step_medianimpute(Gr_Liv_Area)
```

```
Data Recipe

Inputs:
```

```
      role #variables
   outcome           1
 predictor          80

Operations:

Log transformation on all_outcomes
Median Imputation for Gr_Liv_Area
```

KNN approach (typical k = 5-10)

```
ames_recipe %>%
   step_knnimpute(all_predictors(), neighbors = 6)
```

```
Data Recipe

Inputs:

      role #variables
   outcome           1
 predictor          80

Operations:

Log transformation on all_outcomes
K-nearest neighbor imputation for all_predictors
```