ADS 503 - Applied Predictive Modeling

Summer 2024 - Week 2

Dave Hurst

University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

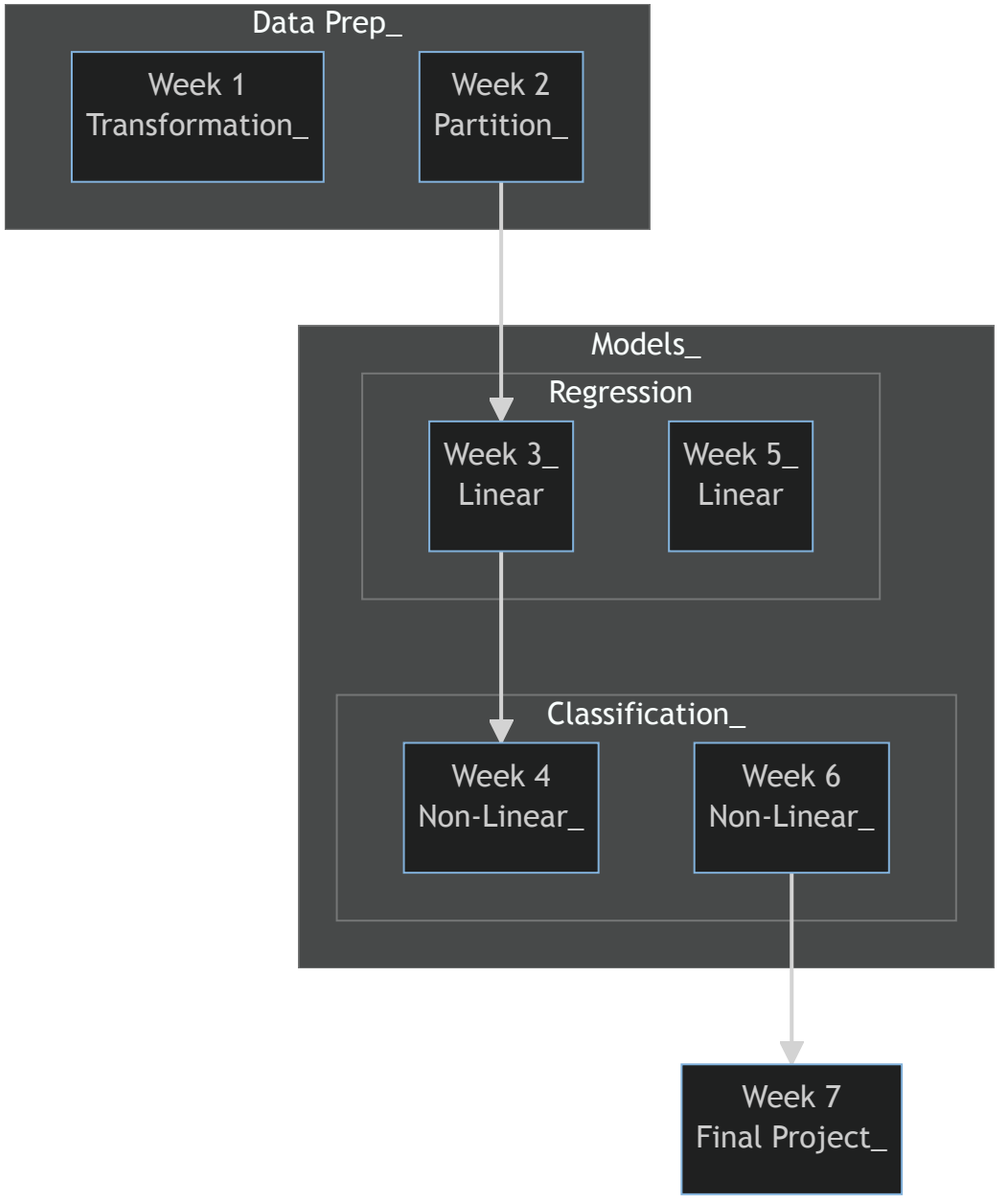
Start Recording! (Link TBD)

University of San Diego®
SHILEY-MARCOS SCHOOL OF ENGINEERING

Agenda

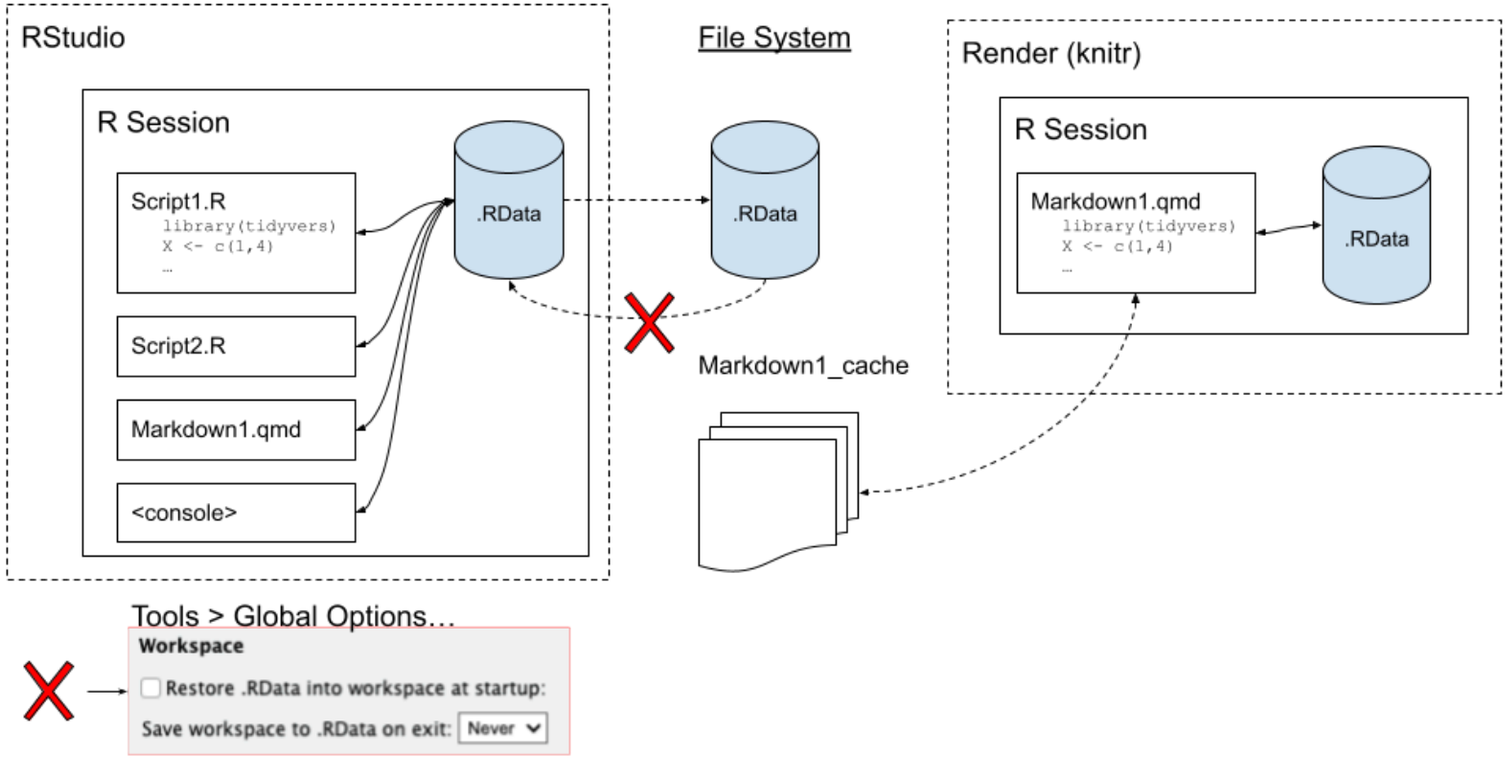
- Course Map
- RStudio and R Session data
- Final Project
- Assignment 1 Review
- QA
  - Binning

Course Map



University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

RStudio and R Session Data



- Remove all variables (not recommended)
- Restart Session

University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

Final Project

- Video Presentation (10-15 min)
  - Problem statement
  - Data used
  - EDA
  - Data preprocessing and splitting
  - Modeling techniques and performance
  - Hyperparameter tuning
  - Final model selection
- Technical Report (10-12 pages, APA 7)

Final Project (Con’t)

- Executive Summary (5 slides/pages)
  - Slides or PDF
  - Non-technical audience focus
  - No presentation required
- Recommendations:
  - GitHub for version control
  - Explore novel datasets
  - Identify areas for improvement
  - Tables / Visualizations are key

Assignment 1 Review

- Warnings
  - `suppressPackageStartupMessages()` or; `#| warnings: false`
- Plots
  - `par(mfrow = c(m,n)` or facet_wrap()`
- Tables
  - `knitr::kable()` or library(gt)`
- Iteration
  - X`apply` or `purrr::map_...`

Assignment 1 - Suppressing Warnings

```
1 suppressPackageStartupMessages(library(tidyverse))
```

… later in notebook …

```
1 library(tidyverse)
```

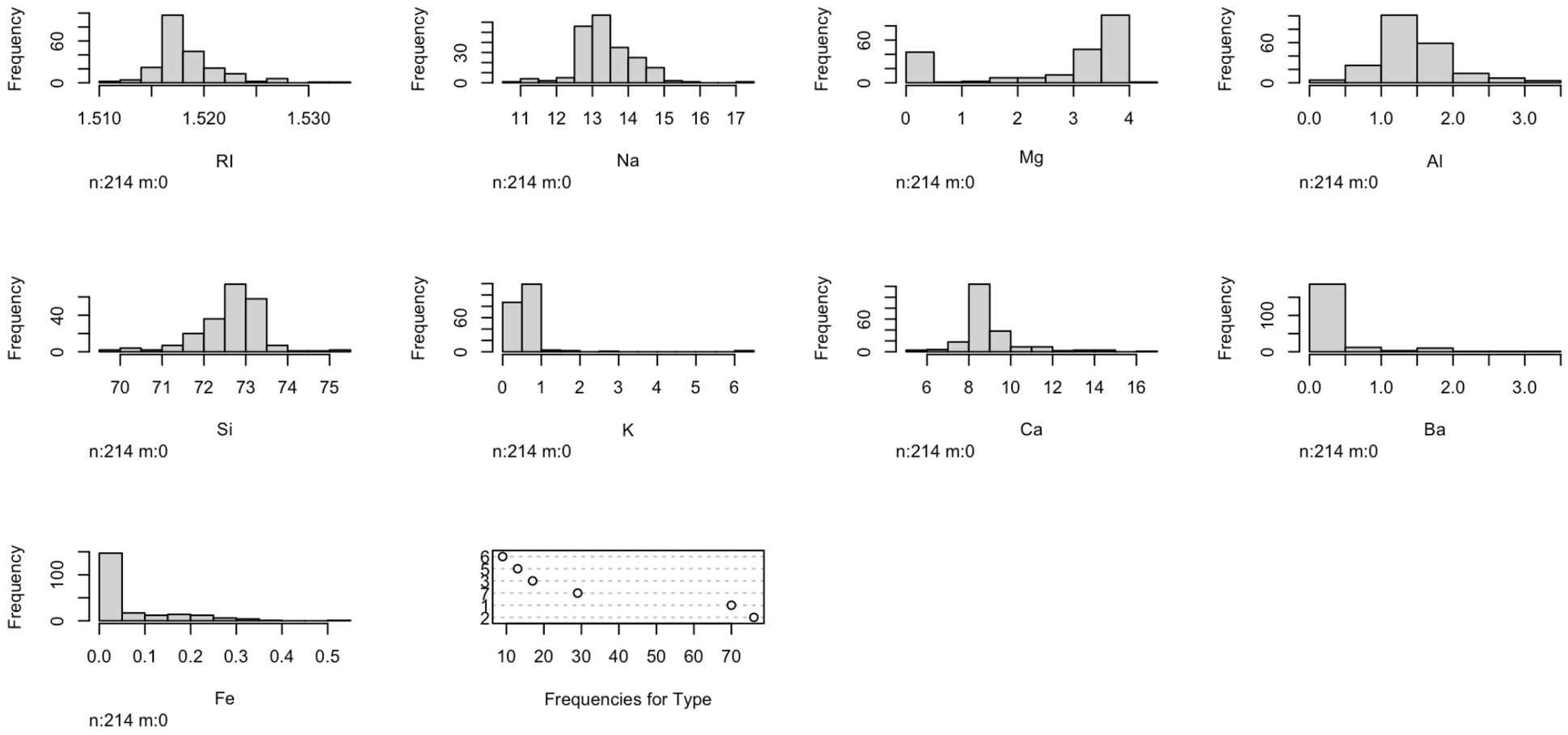
… also handy in targeted chunks …

```
#| warning: false
#| message: false
```

University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

Assignment 1 - Compact Plotting (using `Hmisc()`)

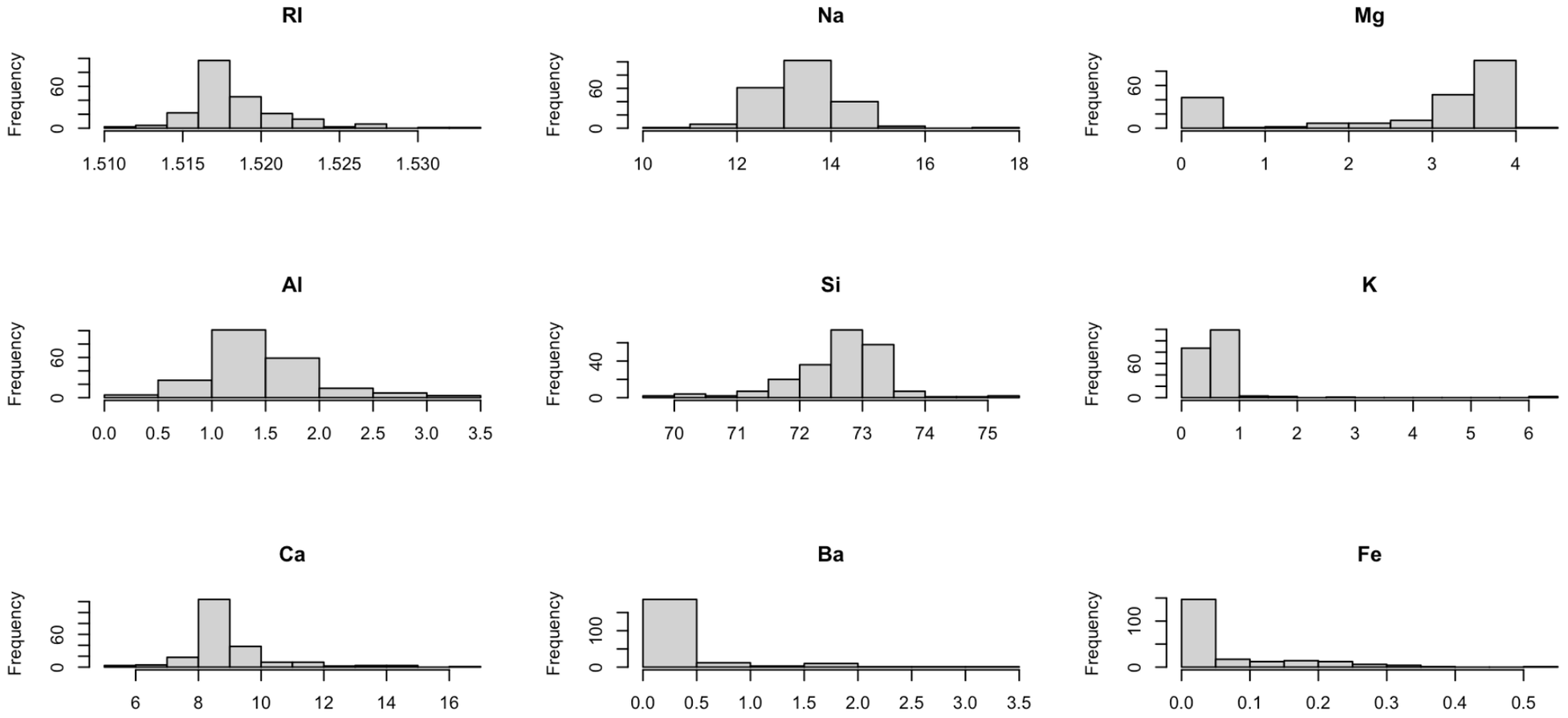
```
1 library(mlbench)
2 data("Glass")
3 library(Hmisc)
4 hist.data.frame(Glass)
```



Assignment 1 - Compact Plotting (using `par(mfrow=...)`)

```
1  par(mfrow = c(3,3))
2  for (i in 1:9) {hist(Glass[ ,i], main = names(Glass)[i], xlab = NULL)}
```

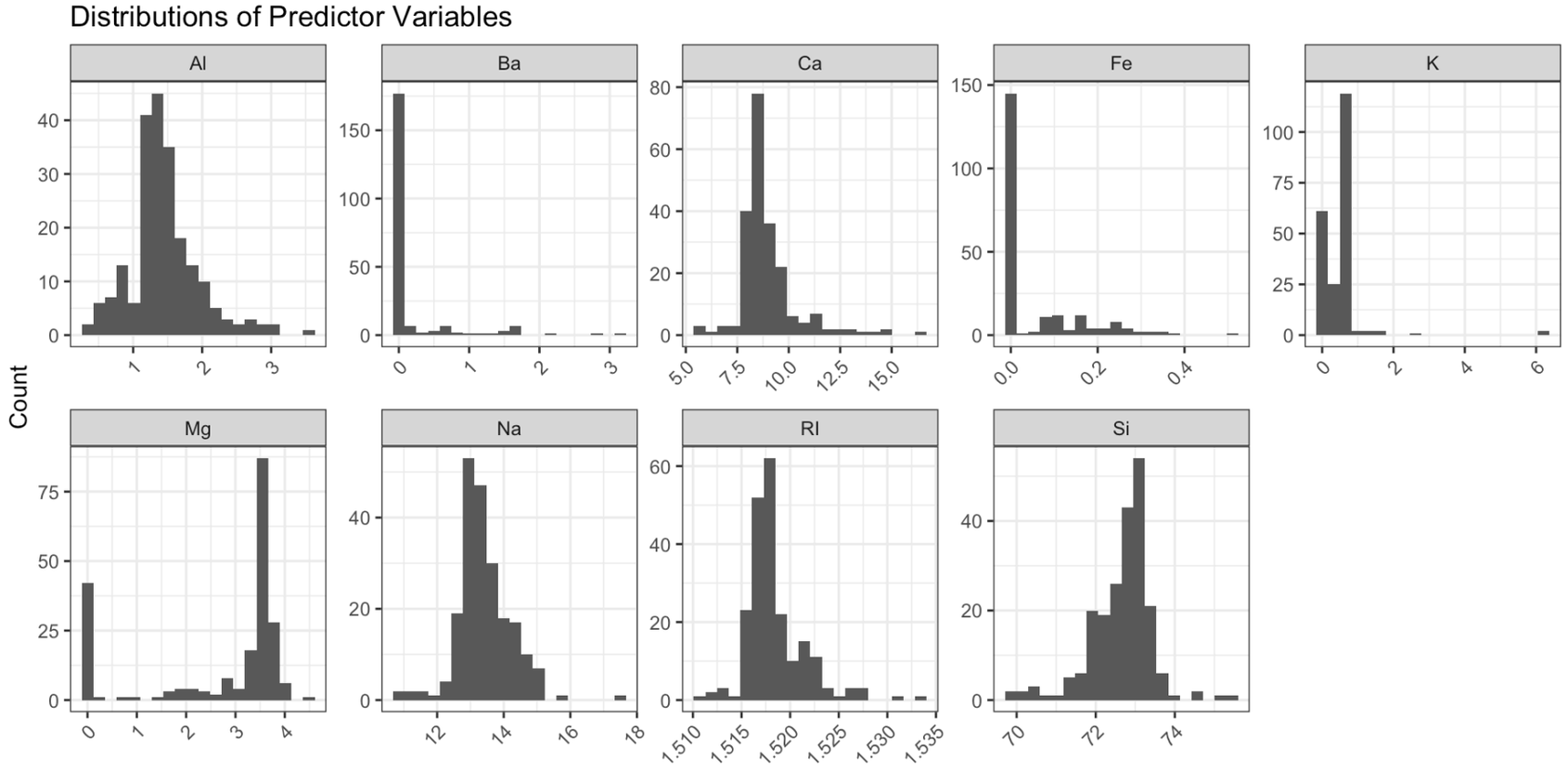
```
1  par(mfrow = c(1,1))
```



University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

Assignment 1 - Compact Plotting (using `facet_wrap...`))

```
Glass |>
    pivot_longer(-Type, names_to = 'Element', values_to = 'value') |>
    ggplot(aes(x=value)) +
    facet_wrap(~Element, scales = "free", ncol = 5) +
    geom_histogram(bins = 20) +
    theme_bw() +
    labs(title = "Distributions of Predictor Variables", x = NULL, y = "Count") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

Assignment 1 - Iteration using `purrr::map_...()`

3.1.c (10 points)

Are there any relevant transformations of one or more predictors that might improve the classification model? Assume the model requires the predictors to have approximately symmetric distribution. Apply relevant transformations to the predictors and observe the changes to the distributions of predictors.

```
suppressPackageStartupMessages(library(e1071))
skew_glass <- Glass |>
    select(-Type) |>
    map_dbl(skewness)
skew_glass |> round(3)
```

```
    RI     Na     Mg     Al     Si      K     Ca     Ba     Fe 
 1.603  0.448 -1.136  0.895 -0.720  6.460  2.018  3.369  1.730 
```

University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

Deep dive : purrr::map()

```
skewness(Glass[ ,1])
```

```
[1] 1.602715
```

```
skewness(Glass[ ,2])
```

```
[1] 0.4478343
```

```
skews <- purrr::map(Glass[ ,1:9], skewness)
class(skews)
```

```
[1] "list"
```

```
skews[1:5]
```

```
$RI
[1] 1.602715

$Na
[1] 0.4478343

$Mg
[1] -1.136452

$Al
[1] 0.8946104

$Si
[1] -0.7202392
```

```
skews <- purrr::map_dbl(Glass[ ,1:9], skewness)
class(skews)
```

```
[1] "numeric"
```

```
skews[1:5]
```

```
        RI         Na         Mg         Al         Si
 1.6027151  0.4478343 -1.1364523  0.8946104 -0.7202392
```

University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

Deep dive : `purrr::map()` with custom functions

```
report_skew <- function(x) {
    skewness(x) |> round(3)
}
purrr::map_dbl(Glass[ ,1:9], report_skew)
```

```
   RI     Na     Mg     Al     Si      K     Ca     Ba     Fe
1.603  0.448 -1.136  0.895 -0.720  6.460  2.018  3.369  1.730
```

… as a formula

```
purrr::map_dbl(Glass[ ,1:9], ~report_skew(.x))
```

```
   RI     Na     Mg     Al     Si      K     Ca     Ba     Fe
1.603  0.448 -1.136  0.895 -0.720  6.460  2.018  3.369  1.730
```

Deep dive : `purrr::map()` with custom functions

```
report_skew <- function(x, xname) {
    tibble(
        predictor = xname,
        skew = skewness(x, na.rm = TRUE),
        min = min(x, na.rm = TRUE)
    )
}
predictors <- Glass[ ,1:9]
purrr::map2_dfr(predictors, names(predictors), ~report_skew(.x, .y))
```

```
# A tibble: 9 × 3
  predictor   skew   min
  <chr>      <dbl> <dbl>
1 RI         1.60   1.51
2 Na         0.448 10.7
3 Mg        -1.14   0
4 Al         0.895  0.29
5 Si        -0.720 69.8
6 K          6.46   0
7 Ca         2.02   5.43
8 Ba         3.37   0
9 Fe         1.73   0
```

University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

Assignment 1 - Apply `BoxCoxTrans()`

… Apply relevant transformations to the predictors and observe the changes to the distributions of predictors.

```
suppressPackageStartupMessages(library(caret))
report_bct_skew <- function(x, xname) {
    if(any(x == 0)) x = x + 0.0001
    bct <- BoxCoxTrans(x)
    trans <- predict(bct, x)
    tibble(
        Predictor = xname,
        `Original Skew` = skewness(x, na.rm = TRUE),
        `Transformed Skew` = skewness(trans, na.rm = TRUE),
        Lambda = bct$lambda
    )
}
predictors <- Glass[ ,1:9]
map2_dfr(predictors, names(predictors), ~report_bct_skew(.x, .y))
```

```
# A tibble: 9 × 4
  Predictor `Original Skew` `Transformed Skew` Lambda
  <chr>               <dbl>              <dbl>  <dbl>
1 RI                  1.60             1.57    -2
2 Na                  0.448            0.0338  -0.100
3 Mg                 -1.14            -1.39     0.400
4 Al                  0.895            0.0911   0.5
5 Si                 -0.720           -0.651    2
6 K                   6.46             0.00954  0.400
7 Ca                  2.02            -0.194   -1.1
8 Ba                  3.37             1.68    -0.6
9 Fe                  1.73             0.735   -0.300
```

University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

Assignment 1 - Table output example

… which predictors would be candidates for a BoxCox transformation?

```
bct_keep <- map2_dfr(predictors, names(predictors), ~report_bct_skew(.x, .y)) |>
        filter(abs(`Original Skew`) > 0.5,
           abs(`Transformed Skew`) < 0.5)
```

```
bct_keep
```

```
# A tibble: 3 × 4
  Predictor `Original Skew` `Transformed Skew` Lambda
  <chr>               <dbl>              <dbl>  <dbl>
1 Al                  0.895             0.0911   0.5
2 K                   6.46              0.00954  0.400
3 Ca                  2.02             -0.194   -1.1
```

```
bct_keep |> knitr::kable()
```

Predictor	Original Skew	Transformed Skew	Lambda
Al	0.8946104	0.0910590	0.5
K	6.4600889	0.0095367	0.4
Ca	2.0184463	-0.1939557	-1.1

University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

Assignment 1 - Table output example w/ `gt::gt()`

```
suppressPackageStartupMessages(library(gt))
bct_keep |> gt()
```

Predictor	Original Skew	Transformed Skew	Lambda
Al	0.8946104	0.091058992	0.5
K	6.4600889	0.009536743	0.4
Ca	2.0184463	-0.193955732	-1.1

```
bct_keep |> gt() |>
    fmt_number(decimals = 3) |>
    fmt_number(columns = 'Lambda', decimals = 1)
```

Predictor	Original Skew	Transformed Skew	Lambda
Al	0.895	0.091	0.5
K	6.460	0.010	0.4
Ca	2.018	−0.194	−1.1

Q&A - Binning

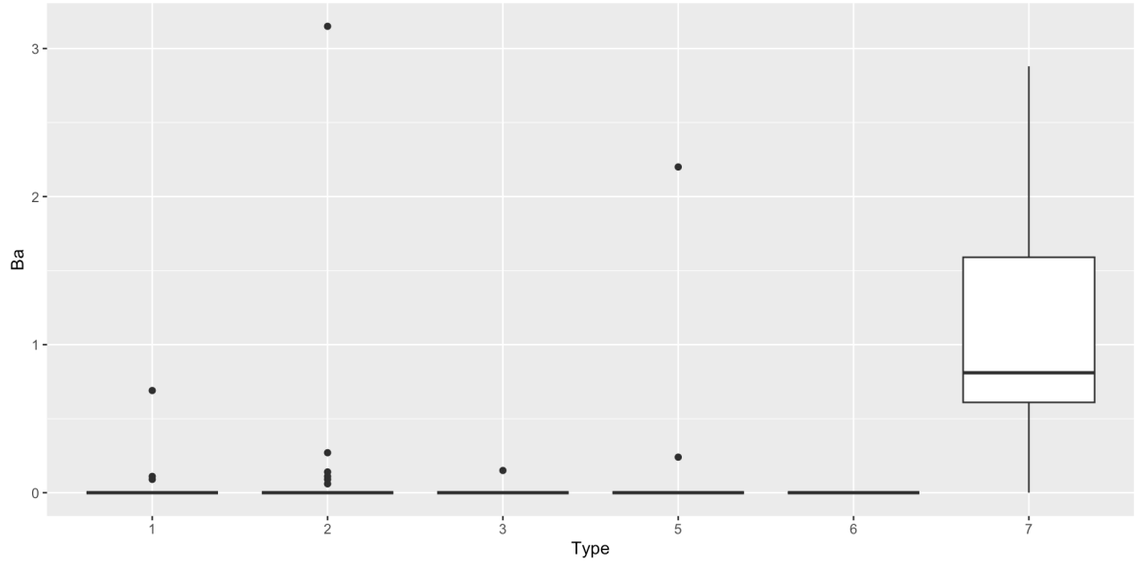
"…could you share some examples of binning numeric predictors and why it should be avoided?"

There are many issues with the manual binning of continuous data. (Kuhn and Johnson, 2013)

- Manual vs Automated
- Econometrics vs Predictive Modeling
- Colinearities

Example

```
1  Glass |> ggplot(aes(Type, Ba)) + geom_boxplot()
```




```
1  glass_aug <- Glass |>
2      mutate(Ba_hi = Ba > 0.05)
3  glass_aug |> head() |> gt()
```

RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	Type	Ba_hi
1.52101	13.64	4.49	1.10	71.78	0.06	8.75	0	0.00	1	FALSE
1.51761	13.89	3.60	1.36	72.73	0.48	7.83	0	0.00	1	FALSE
1.51618	13.53	3.55	1.54	72.99	0.39	7.78	0	0.00	1	FALSE
1.51766	13.21	3.69	1.29	72.61	0.57	8.22	0	0.00	1	
1.51742	13.27	3.62	1.24	73.08	0.55	8.07	0	0.00	1	
1.51596	12.79	3.61	1.62	72.97	0.64	8.07	0	0.26	1	

Q&A

University of San Diego®
SHILEY-MARCOS SCHOOL OF ENGINEERING