Attempting to Predict Kaggle Dataset Usability

Marc Marquez

University of Florida

Background

- What is Kaggle?
  - Website that features:
    - Data science courses, competitions, datasets
    - 317,983 datasets as of 4/20/2024
- Kaggle datasets have a Usability Rating
  - Unknown formula
  - Maybe there's a way to predict it?

kaggle

Usability

8.13

This score is calculated by Kaggle.

Completeness · 75%

- ✓ Subtitle
- ✓ Tag
- ✓ Description
- ✗ Cover Image

Credibility · 100%

- ✓ Source/Provenance
- ✓ Public Notebook
- ✓ Update Frequency

Compatibility · 67%

- ✓ License
- ✗ File Format
- ✓ File Description

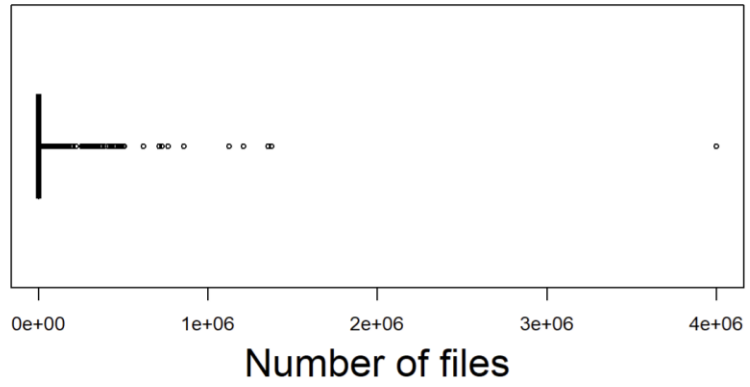
The Data

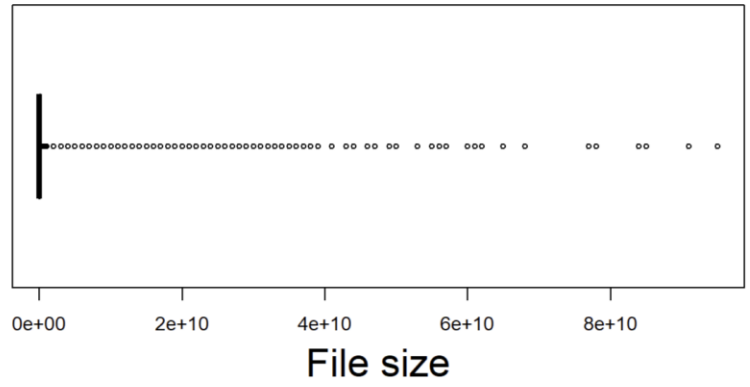
- https://www.kaggle.com/datasets/rajugc/kaggle-dataset
- 9159 entries of dataset information:
  - Number of files
  - File size
  - Filetypes
  - Number of upvotes
  - Medal type (none, bronze, silver, gold)
  - Date (Months since 2015)
  - Day of upload
  - Usability (response)

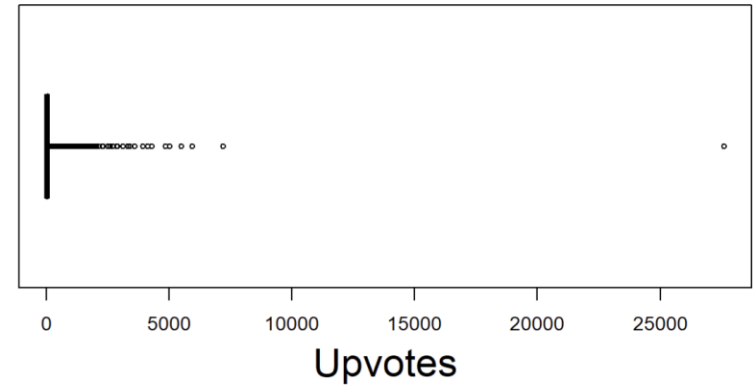
Methods

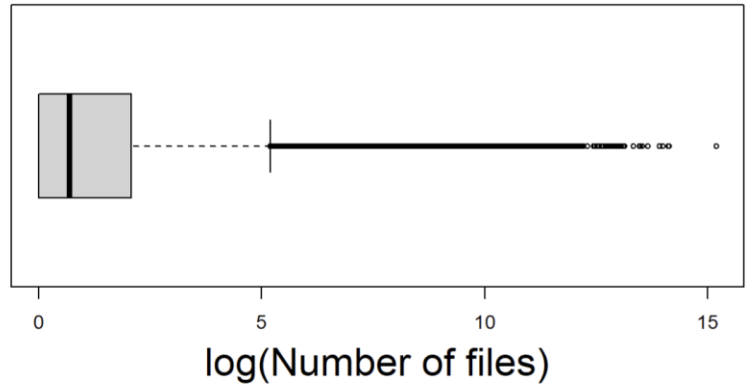
- Distributions of quantitative predictors:
  - Some have too many outliers
    - Log transformation can salvage file size
    - Drop number of files and number of upvotes
- All are skewed, as well as Usability
- Categorical predictors have enough entries per level
  - JSON and SQLITE incorporated into Other

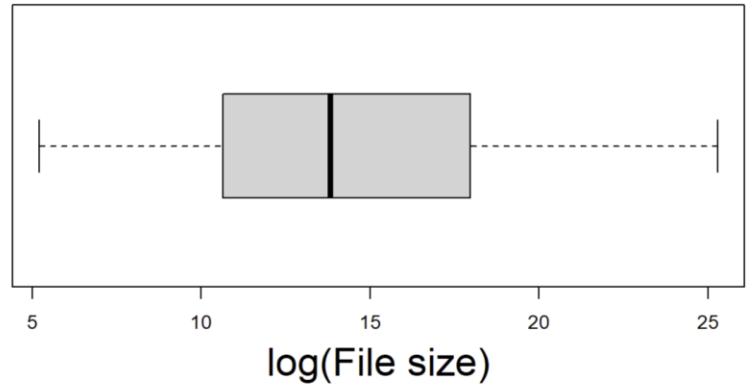
Cont. Predictor Boxplots

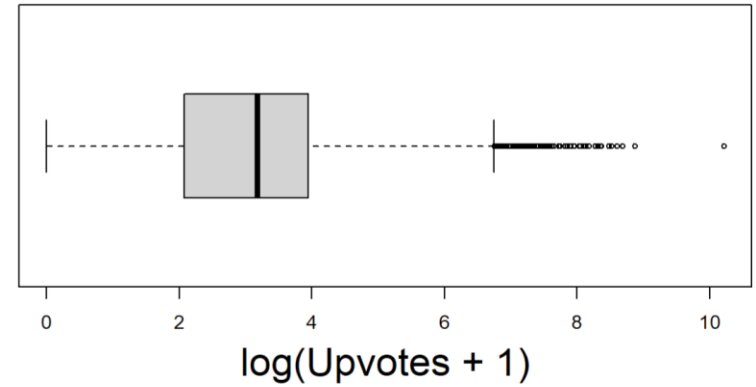




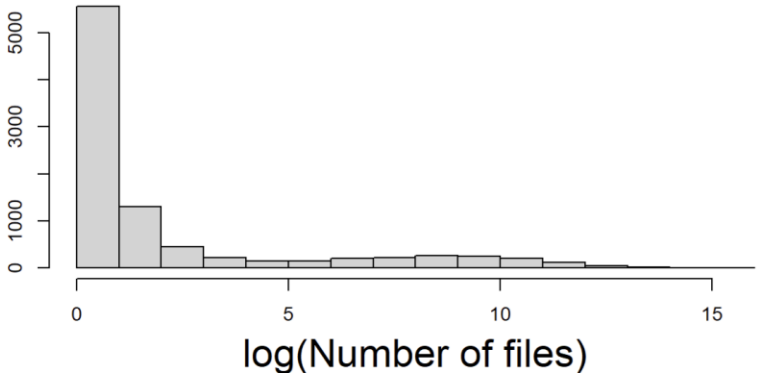


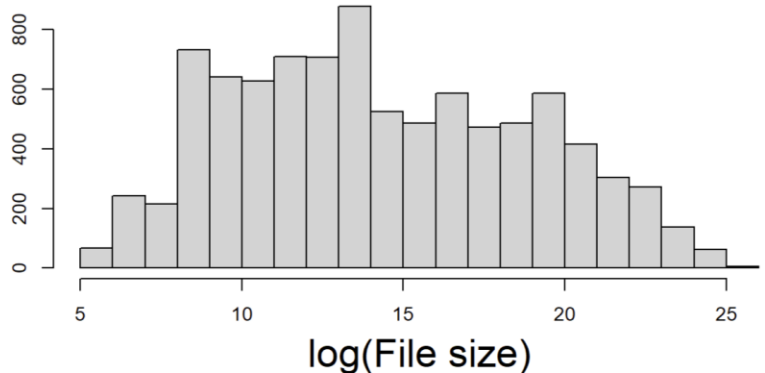


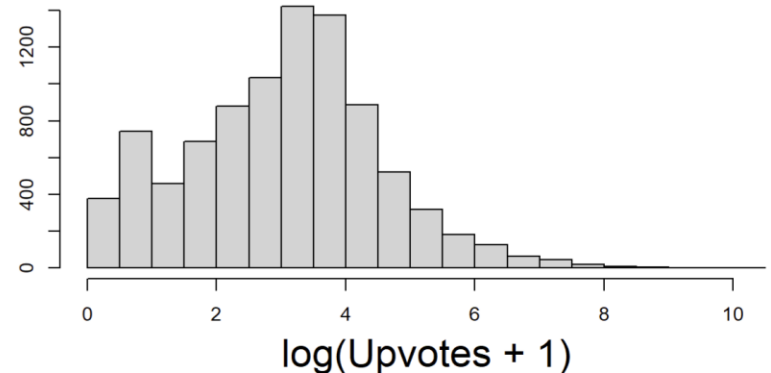


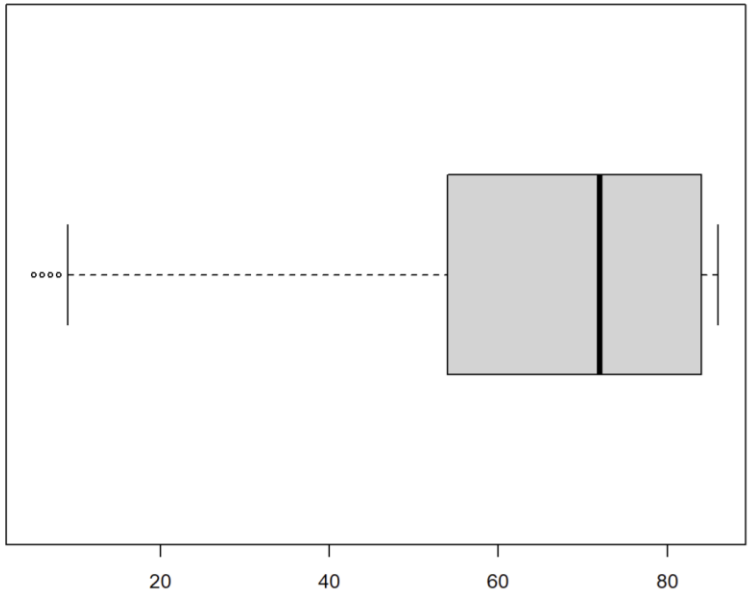


Histograms

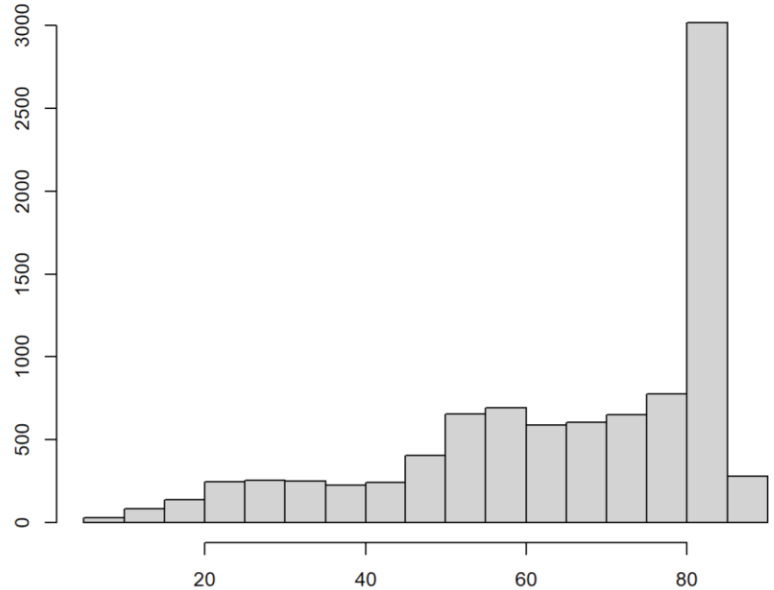




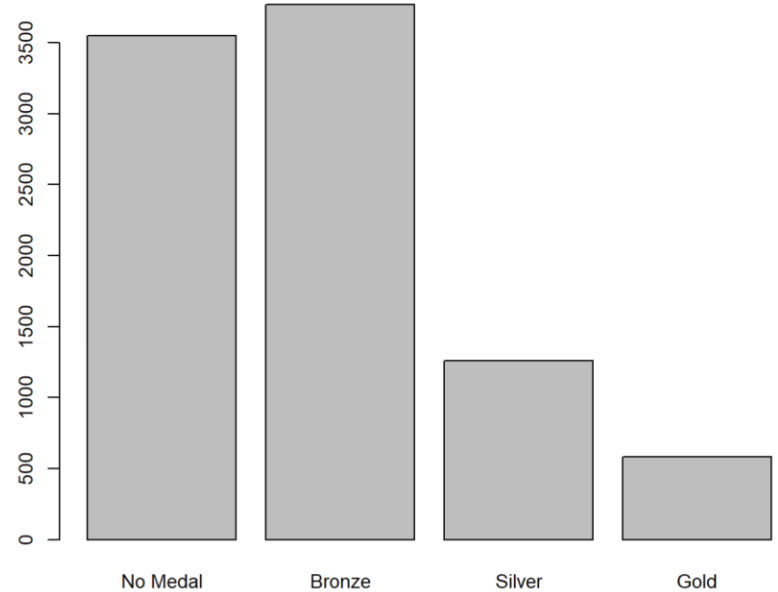




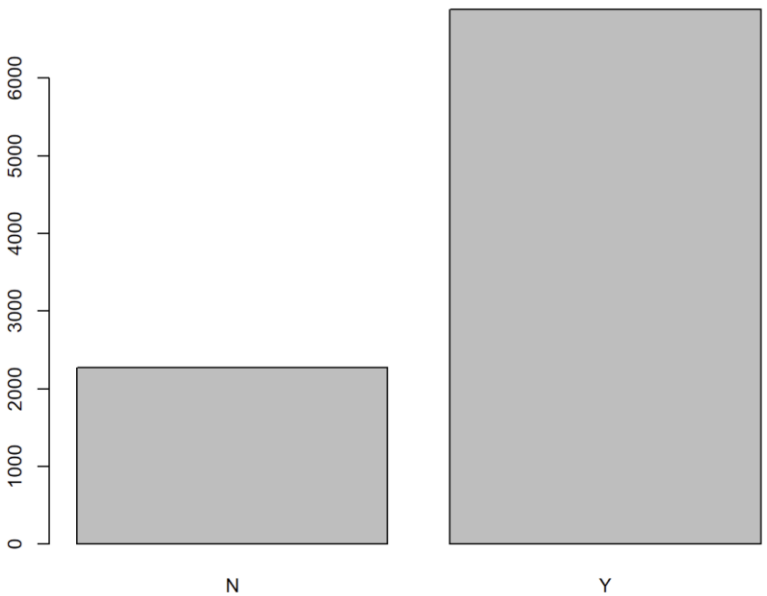
Months since 2015



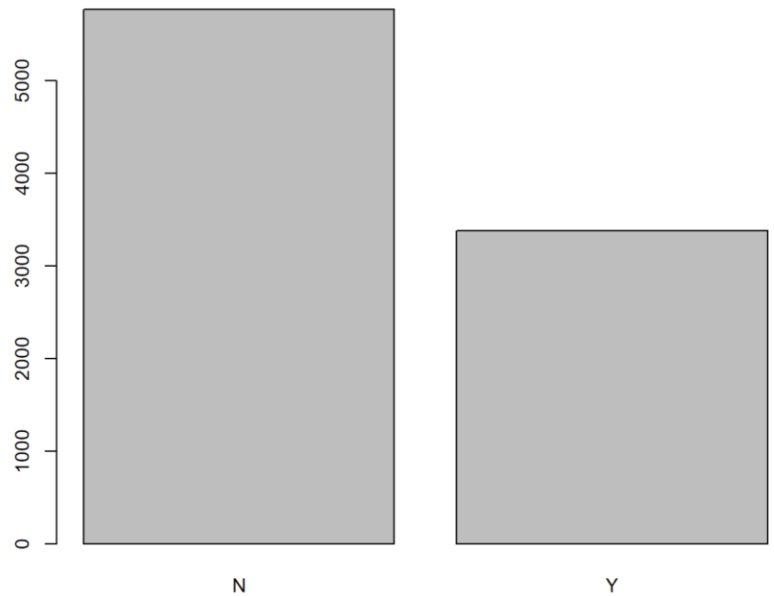
Months since 2015



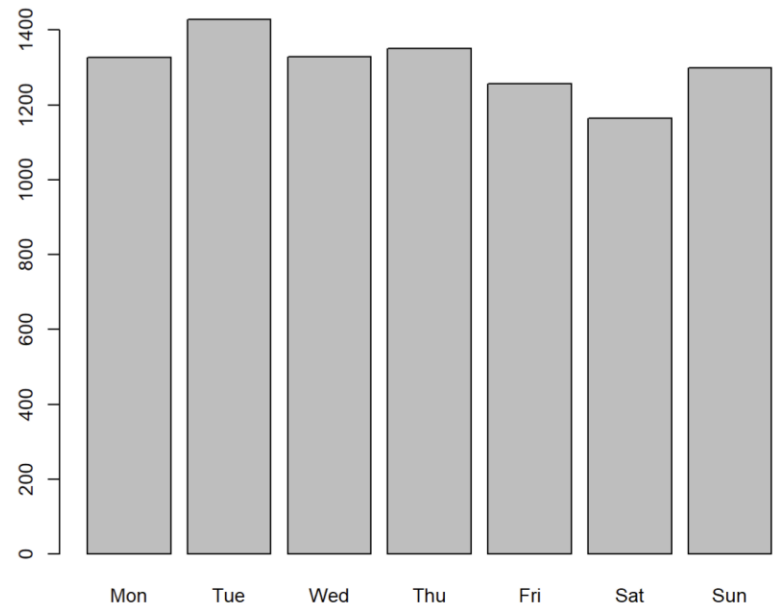
Medal



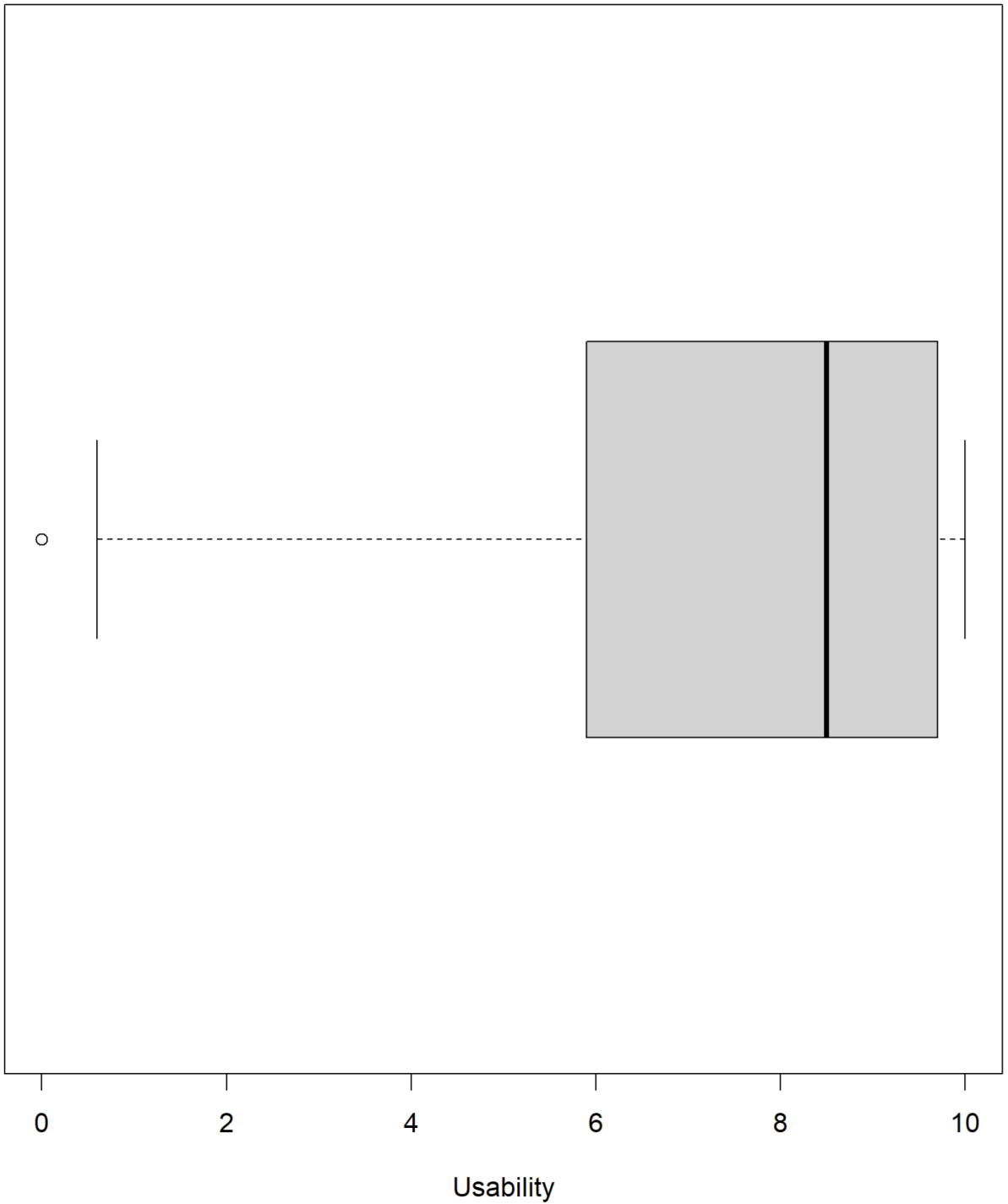
Contains CSV

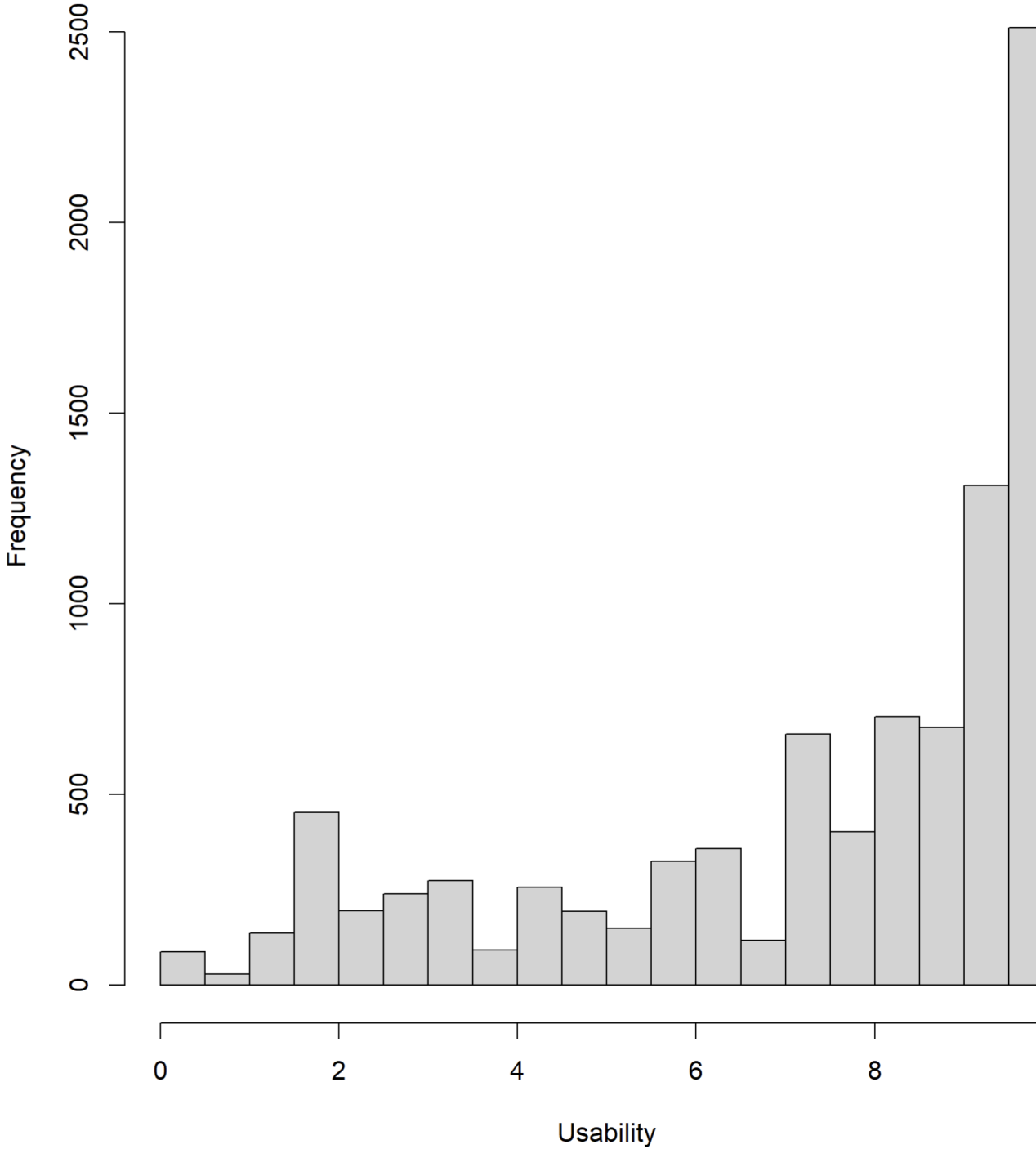


Contains Other



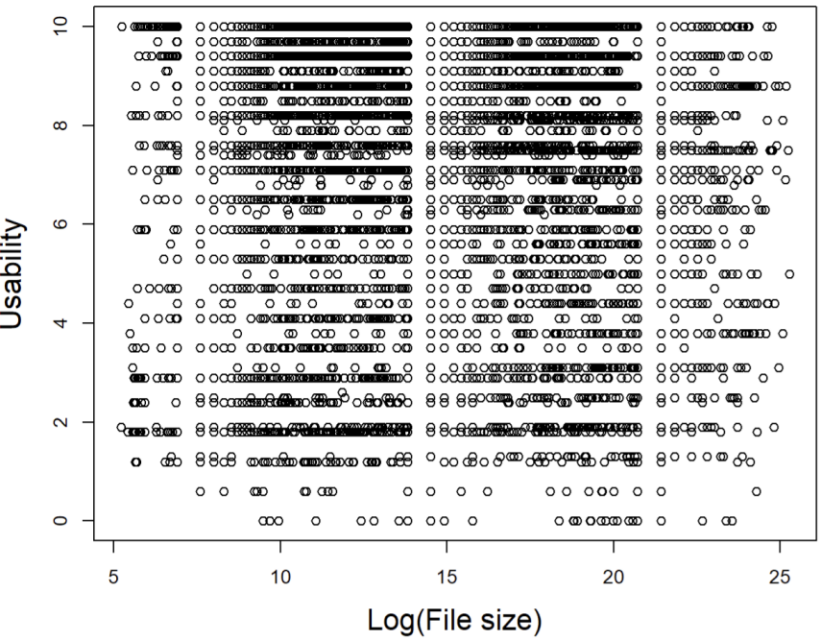
Day of Upload

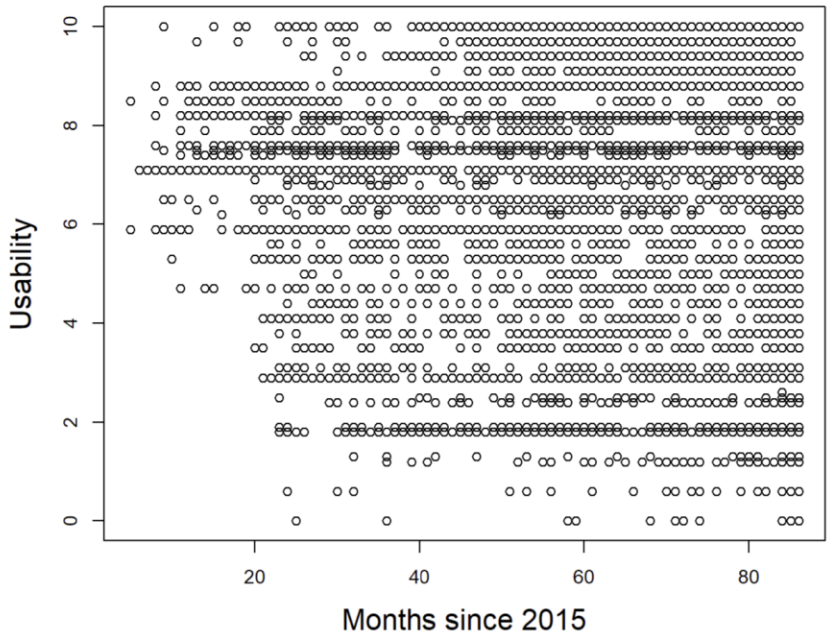

0
2
4
6
8
10
Usability

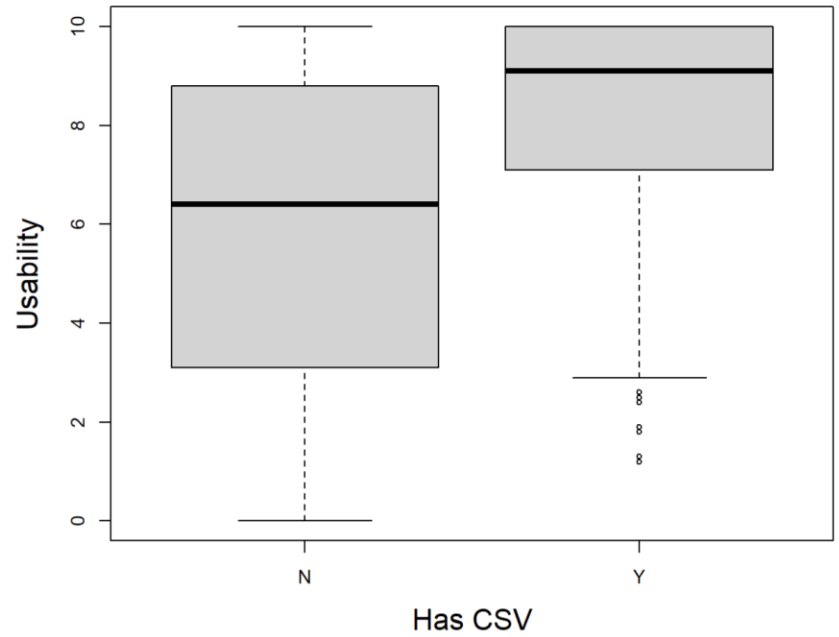

Frequency
0
500
1000
1500
2000
2500
0
2
4
6
8
Usability

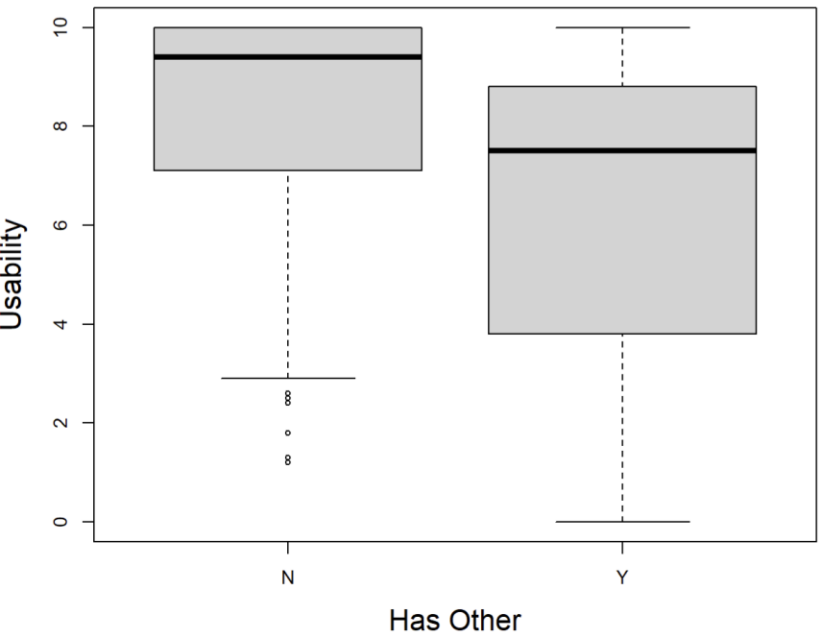
- Usability plotted against each factor individually
  - log(File size) and Months since 2015 had no clear pattern
  - CSV, Other, and Medal appeared to have 1 significant difference in level
- 1 Factor Linear Models then fitted
  - Clear linear trend on all residual plots except Day
  - Extremely significant p-values on Levene and Breusch-Pagan tests
  - Likely due to Usability limits
  - Box-Cox and IWLS did not help

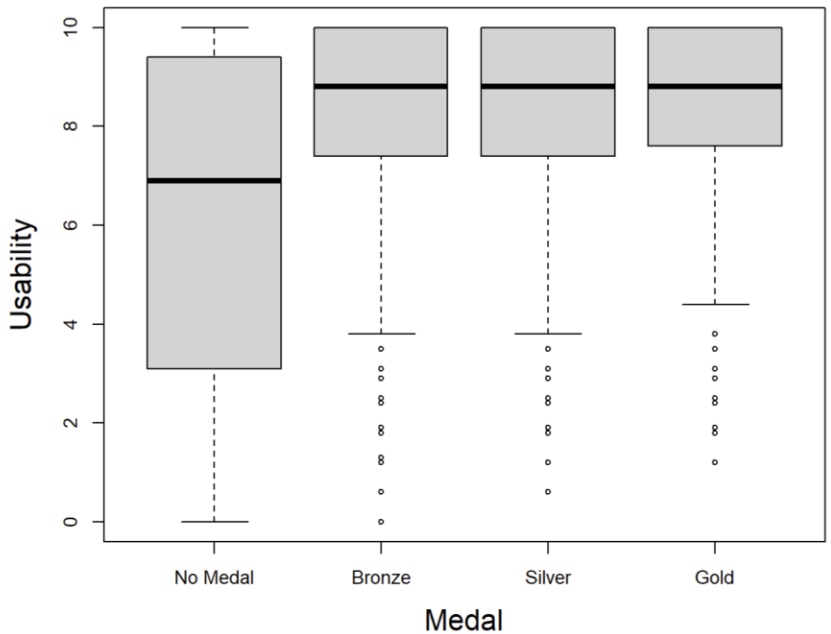
1 Factor Models

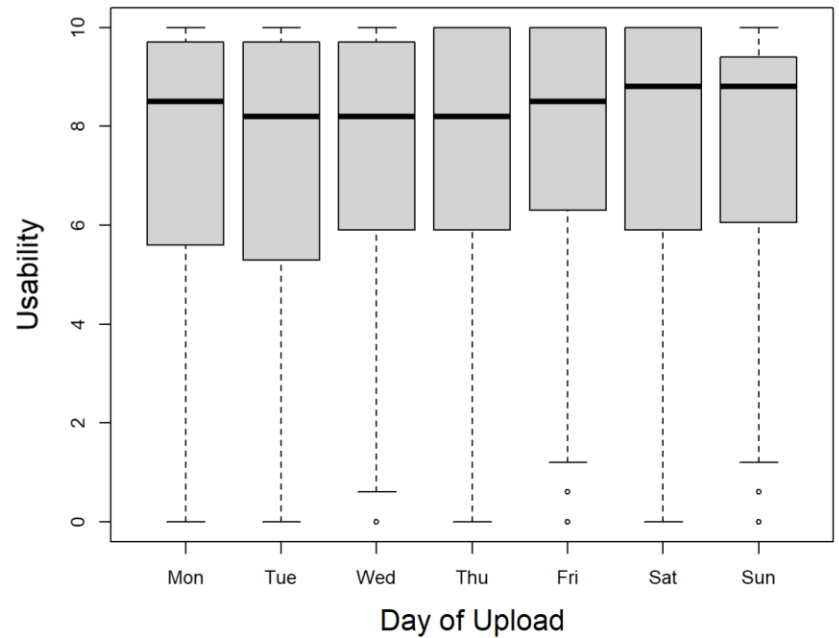


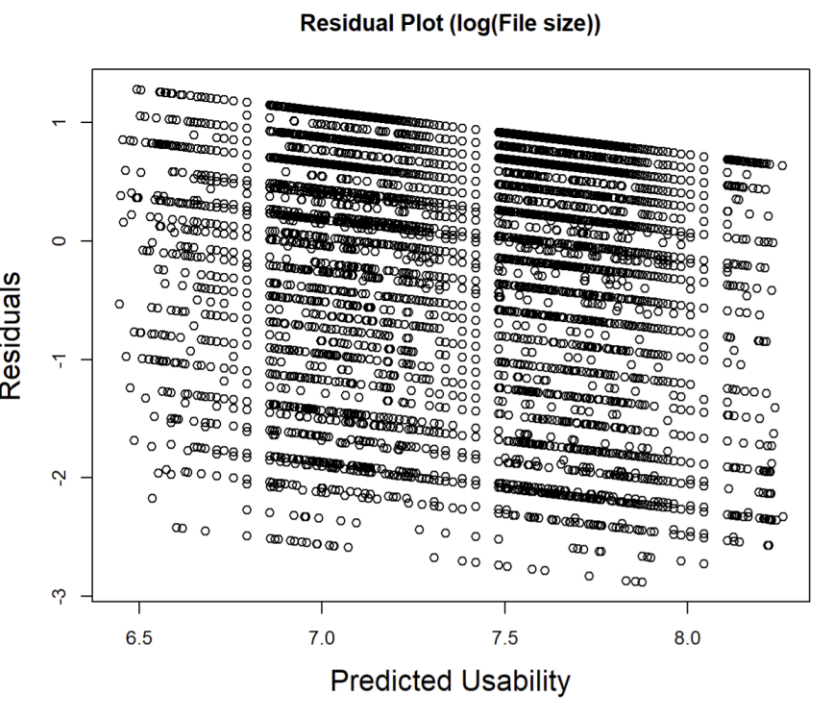


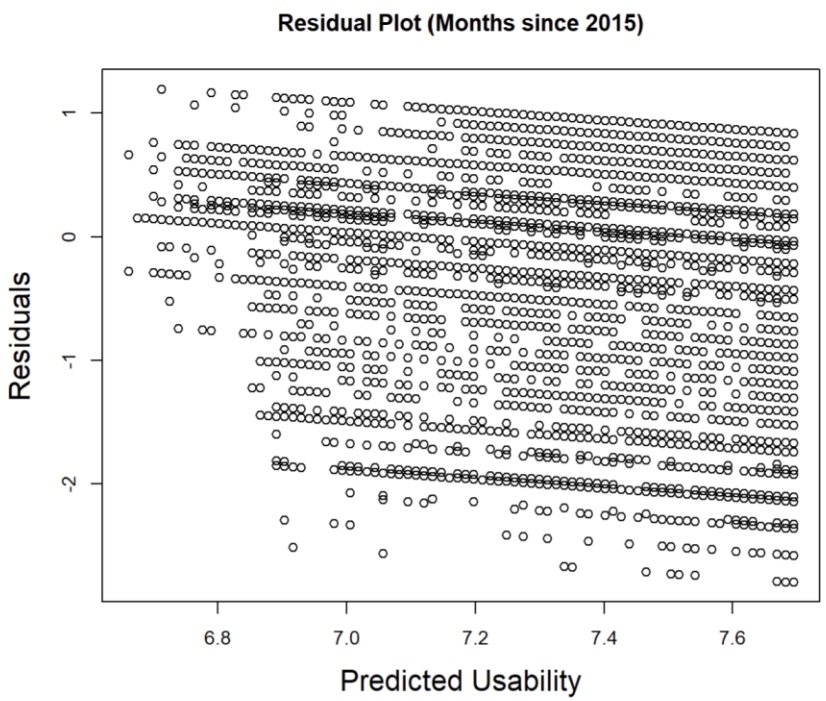


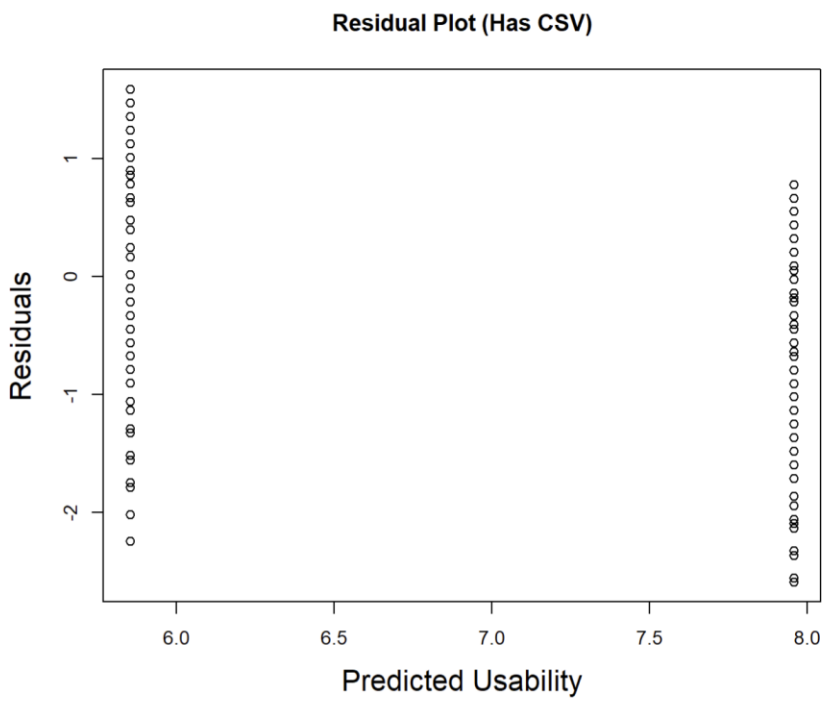


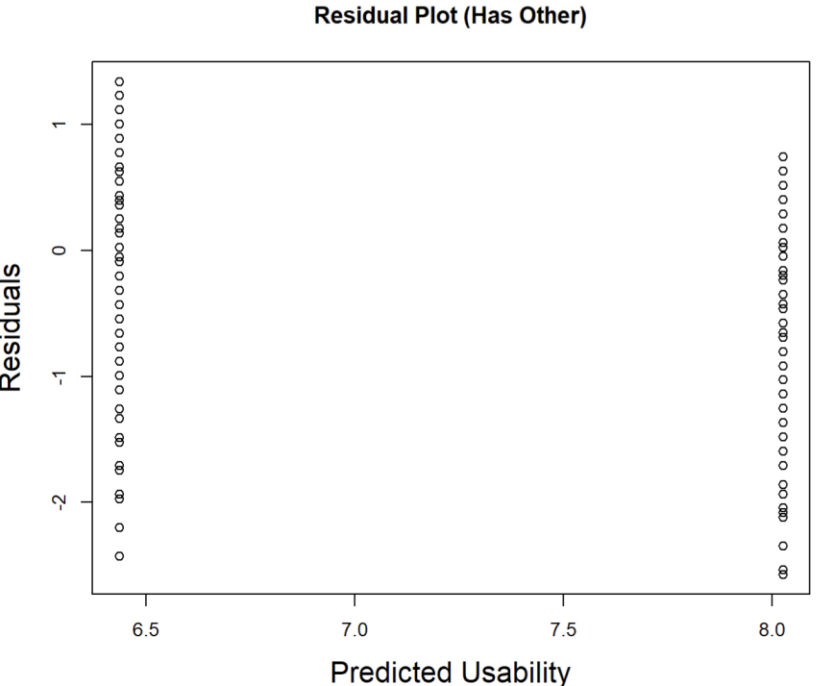


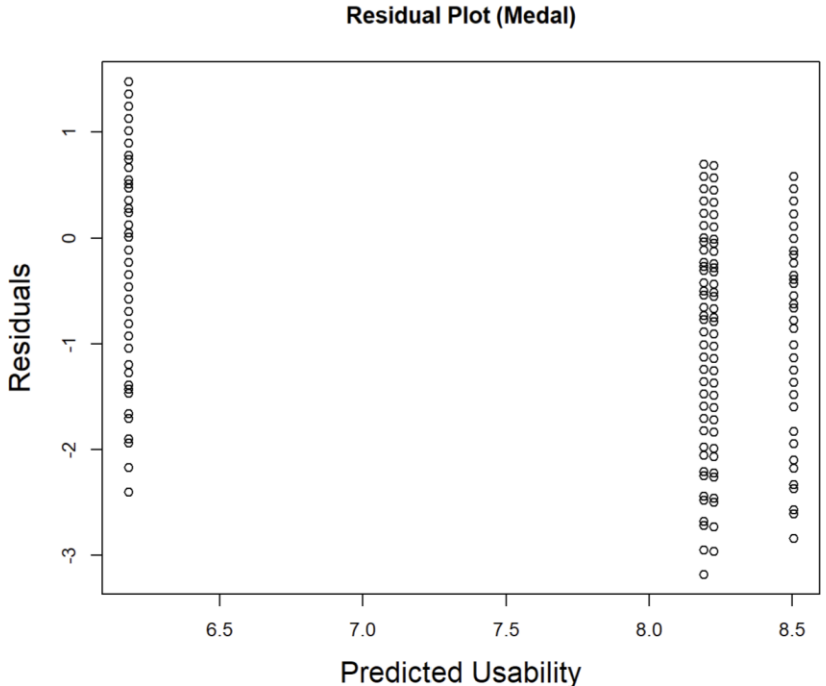


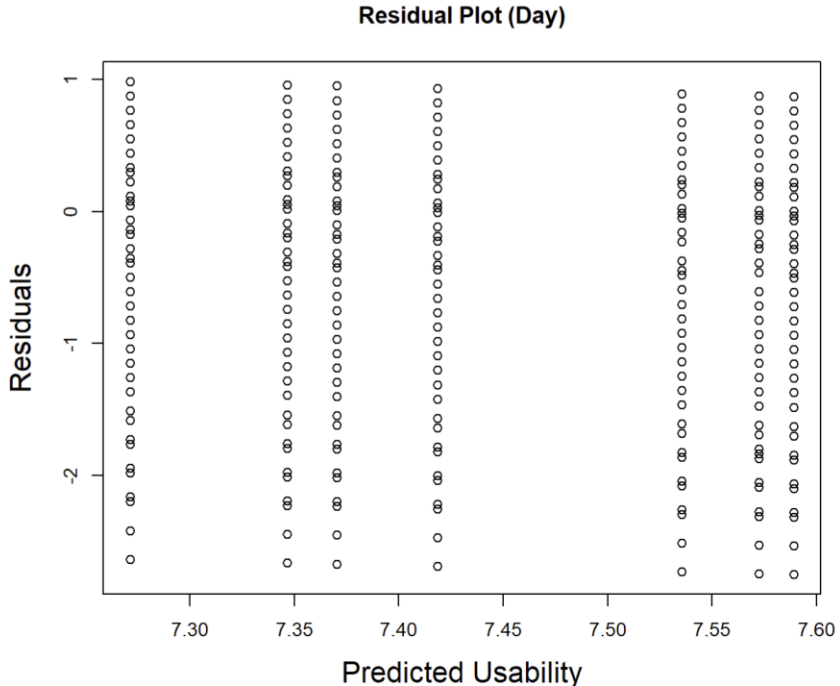

Residual Plot (log(File size))
Residuals
Predicted Usability


Residual Plot (Months since 2015)
Residuals
Predicted Usability


Residual Plot (Has CSV)
Residuals
Predicted Usability


Residual Plot (Has Other)
Residuals
Predicted Usability


Residual Plot (Medal)
Residuals
Predicted Usability


Residual Plot (Day)
Residuals
Predicted Usability

- Parametric methods were insufficient
- Use nonparametric methods instead
  - Tree-based models
  - Why trees?
    - Interpretable
    - Can capture complicated relationships
  - However:
    - Weak to outliers (transformation may fix)
    - Poor predictive performance

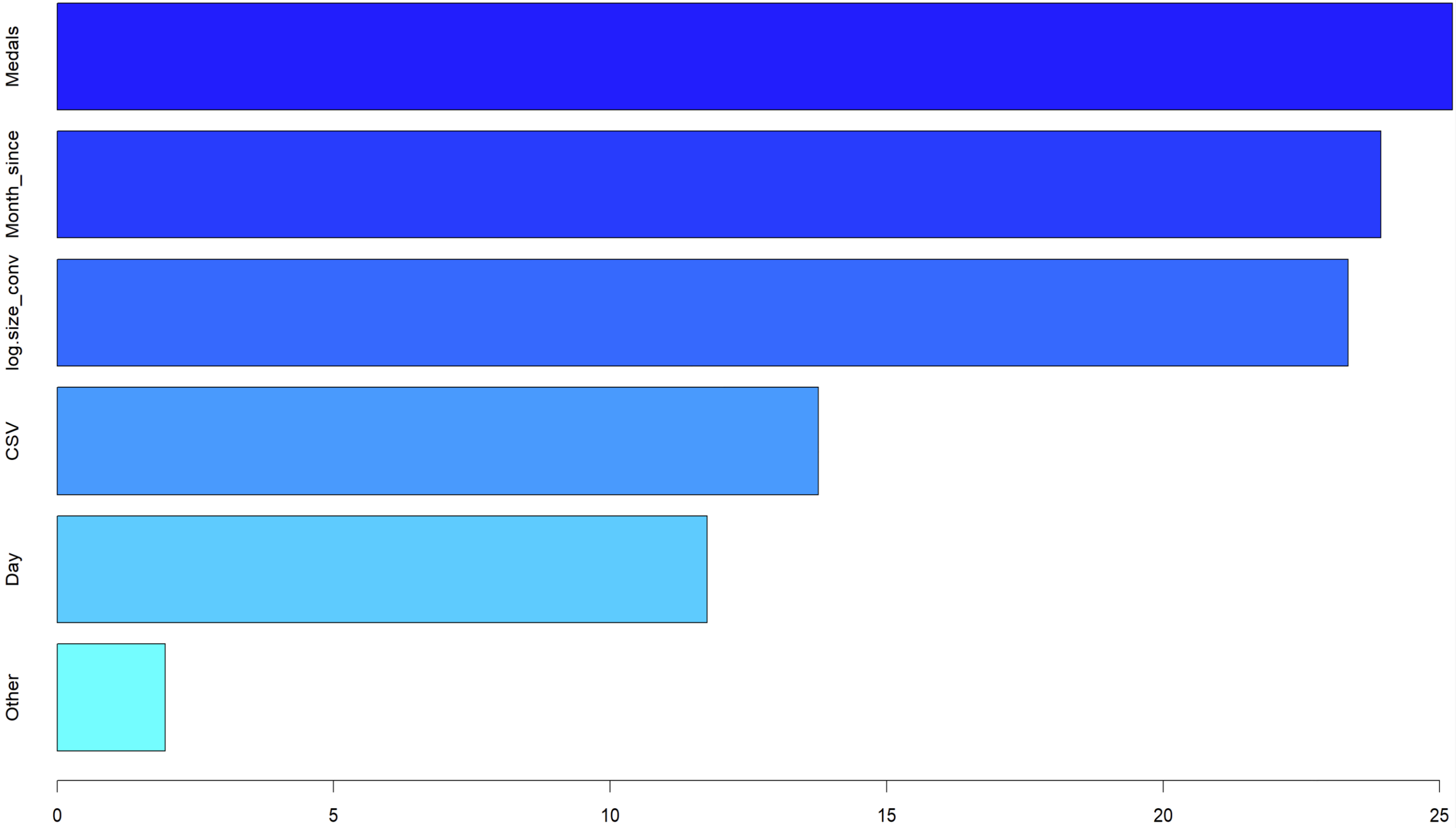
- Multiple tree models were trained and tested
  - Basic Decision tree
  - Bagging tree
  - Random forest
  - Several boosted trees
  - Bayesian Additive Regression Trees
- Model with smallest test error was selected

The Model

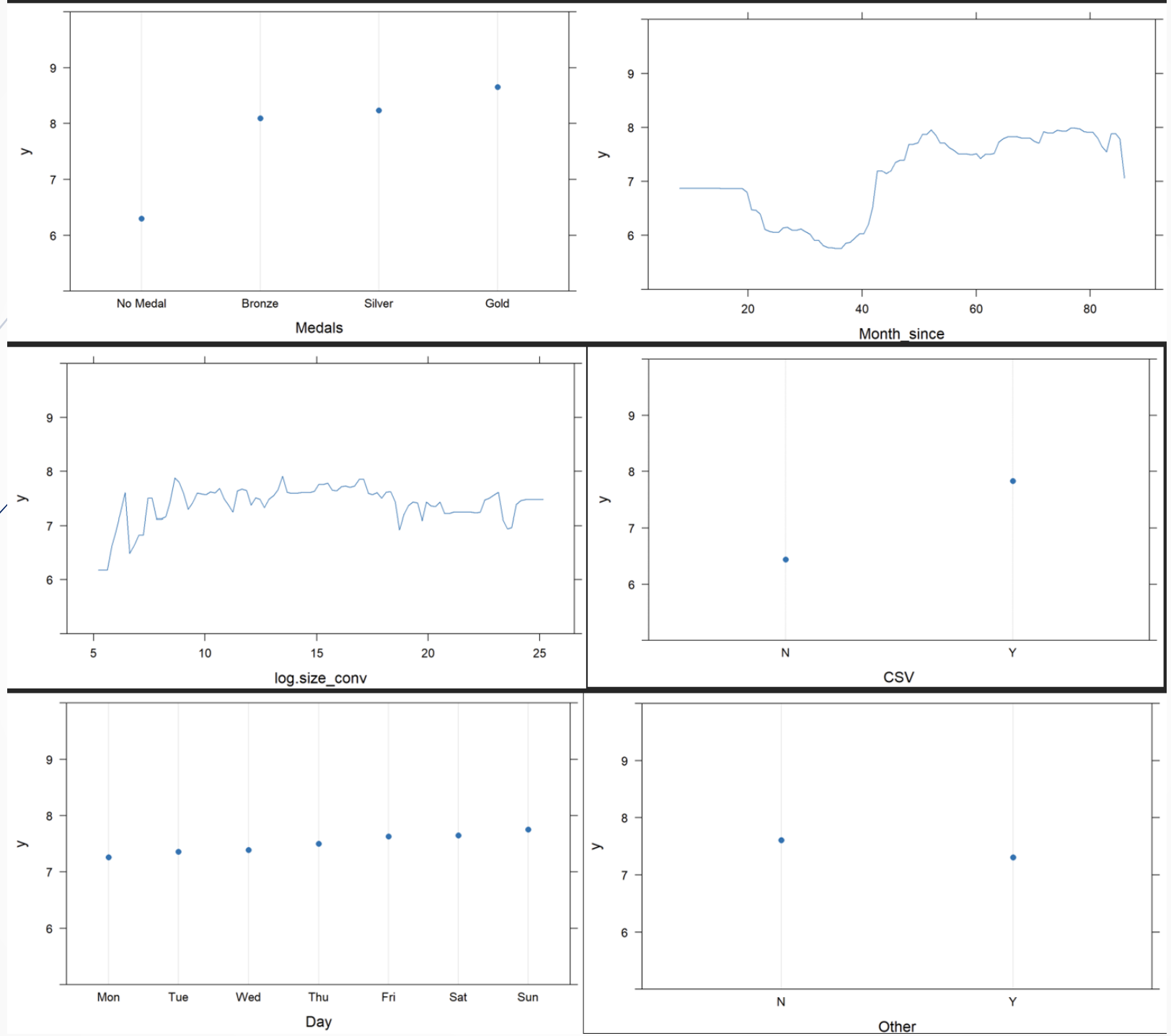
- Boosted regression tree
  - 2000 trees
  - Shrinking parameter of 0.01
  - Interaction depth of 4
  - Predictors:
    - log(File size), Months since 2015
    - Contains CSV file, Contains Other file, Medal, Day of upload

Results

- Relative influence plot shows influence of each predictor:
  - Medal: 25.2352%
  - Months since 2015: 23.9388%
  - Log(File size): 23.3470%
  - CSV filetype: 13.7616%
  - Day of upload: 11.7600%
  - Other: 1.9575%


Medals
Month_since
log.size_conv
CSV
Day
Other
0
5
10
15
20
25

- Partial dependence plots:
  - Datasets with no medals appeared to have less Usability than those that did
  - Datasets that included a CSV file appeared to have more Usability than those that did not
- Mean test error of 5.1983
  - Rather poor for a scale of 0 to 10


y
9
8
7
6
No Medal
Bronze
Silver
Gold
Medals
20
40
60
80
Month_since
5
10
15
20
25
log.size_conv
N
Y
CSV
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Day
Other

Conclusion

- The variables are not as good of predictors of Usability as expected
- What could help?
  - More/better data
  - Different predictors
- "Usability" is subjective
  - Numerical scaling may be inadequate
  - The exact formula is unknown