**Overview**

Back Blaze is a Cloud Storage provider for developers and IT department. As such they rely on good hard drives to read, write, and store data. The “goodness” of each hard drive is determined by S.M.A.R.T statistics ([source](https://www.backblaze.com/b2/hard-drive-test-data.html)) as reported by the drive.

This dataset for Jan to Dec 2018 contains all the S.M.A.R.T statistics for each hard drive and if it fails.

We would like to find out which SMART stat contributes to failure or success rates. Ultimately it culminates into a predictive model for hard drive failure.

*Qn: Can we calculate the failure rate for hard drives?*

**EDAs**

1. Meaning of each stat?
2. Sample proportion (Fail vs Others: 98% : 2 %)
3. Means, Variances
4. Outliers
5. Correlations
6. NaNs or Nulls

https://datascienceplus.com/blazing-fast-eda-in-r-with-dataexplorer/

**Prep**

1. Over sample? Balance?
2. Smoothing averages needed?
3. Train-Test-Validation split

**Models**

1. Random Forest
2. Decision Tree
3. Logistic regression
4. Neural Network (which one though?)