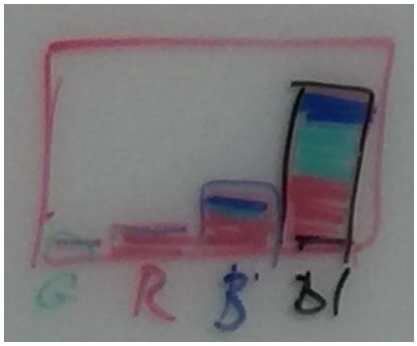
Sane Binning (or Something)

1. Premise
    a. Curse of dimensionality leads to
        i. Continuous to categorical/binning vals for classes ?????
2. Intro to Album review dataset
    a. Where to get
    b. How to preprocess
3. The preliminary pipeline with classifier led to:
    a. Bad results
    b. Naive binning which results in:
        i. Selection bias
4. Redo with ClassBalance report with the hope that:
    a. Using YB results in:
        i. Better binning, thus
            1. Better Results
5. Teaser:
    a. Heatmap version of ClassBalance



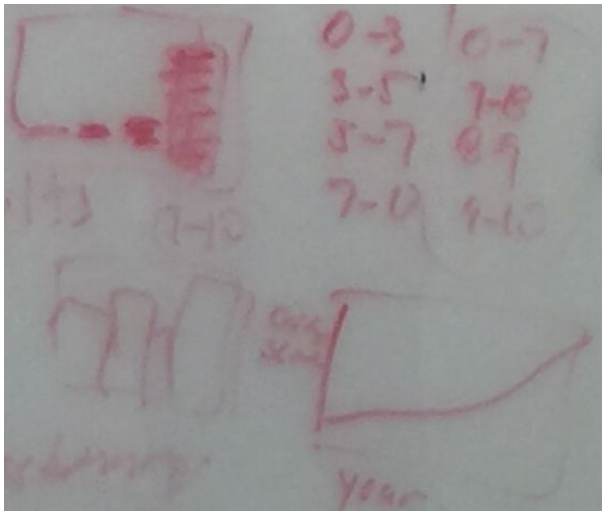
Example

Album Bin Example:
Initially classes are broken down into 4 ranges

0 to 3 - Bad
3 to 5 - Ok
5 to 7 - Good
7 to 10 - Great

However, the data is heavily skewed toward higher scores.
To better balance the classes, we could break down them into different ranges, with the hope that the classes will be balanced

0 to 7-Bad
7 to 8 - OK
8 to 9 - Good
9 to 10 - Great



Workflow

```
Pipeline = Pipeline((
    ('norm', TextNormalizer()),
    ('vect', TfidfVectorizer()),
    ('class', estimator)
    ))
pipeline.fit_predict(docs)
```

TextNormalizer is for doc preprocessing and can consists of:

1. Keyphrase extraction
2. Entity extraction
3. Stemming/Lemmazation
4. Freq Distribution

TfidfVectorizer has additional parameters than can be tuned such as:

1. N-components

How to create binning:

```
y = [0.2, 5.7, 9.1, ….]
X = np.digitize(y, [0.0, 3.0])
```

How feature Unions Work.