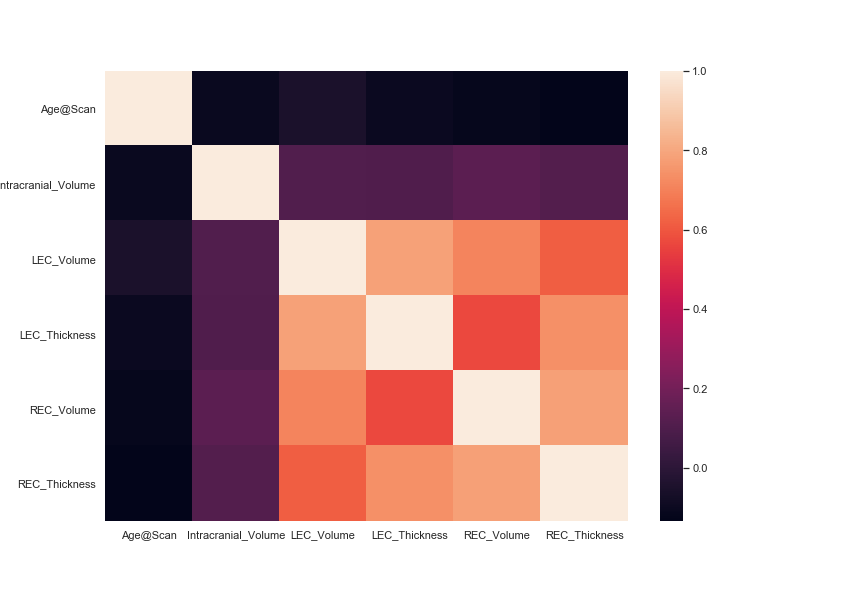
1. Data Preprocessing (going over data preprocessing in a less bs manner)

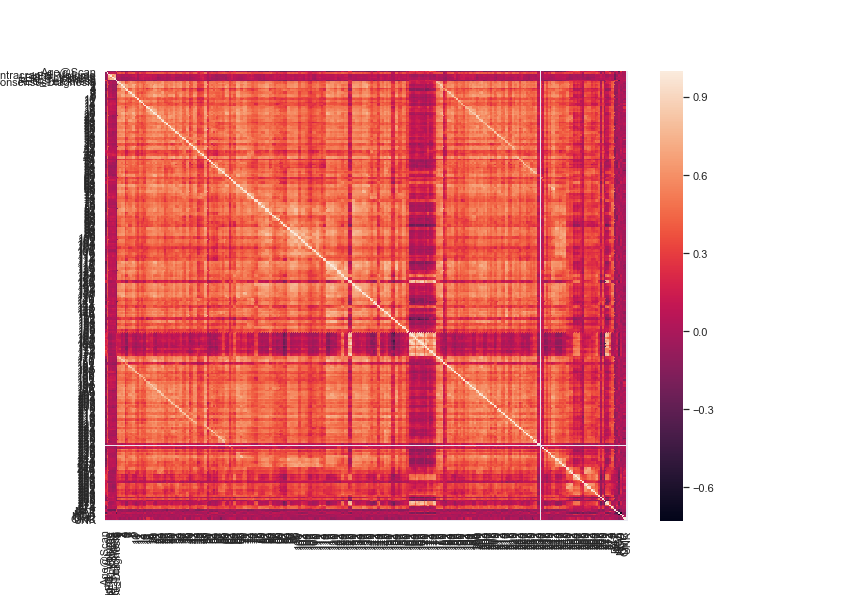
Key Notes

* Last row in BIOCARD\_Entorhinal contains “UNK” so simply removed it
* Removed QMC and OTH – underrepresented
* Removed scan date – shouldn’t have an effect (I think??? Ask Hubert)
* Adjusted class mix for 50/50, as it is binary classification problem (downsampled randomly)

Correlation Plots (checking for multicoliniarity)



Seems ok. Didn’t really check between all the files because of plot interpretability.



Here is figure for reference… two major areas of attention. But hard to parse through.

CURRENT PROBLEMS…

QDA (multicollinearity)

PCA

* Up to 95% explained variance… which is 100 variables. Should probably test with different PCs

TRY TO IMPROVE DATA

* Standardize?
  + Should I standardize one-hot-encoded variables

Standardizing Improved testing on validation set, which is all I rly care about

**Model Improvement**

**Logistic Regression**

Hyperparameter Tuning

* Penalties (used both ridge and lasso regression)
  + Expect lasso shrinking to work better on non-transformed data but tested on both sets nonetheless
* Tolerance
  + Went for lower tolerances in increments of .1
* Intercept Scaling
  + Seems like it wants it as low as possible