**CoveryMyMeds Dataset**

The dataset consists of 4 tables:

* ***dim\_claims*** contains information on the claims (# of claims = 1335576) submitted to each Payer (“bin”), including the drug type, and whether it was approved or rejected, along with the relevant reject\_code. This information will be useful for predicting whether or not a claim will be approved or rejected.
* ***dim\_pa*** contains information on prior authorizations (PA) submitted for rejected claims (# of PA = 555951), and whether or not the PA was approved, and includes physician supplied information on whether the diagnosis is correct (correct\_diagnosis), whether other therapies have been tried and failed (tried\_and\_failed), and whether or not there is a contraindication (reason to not take the drug).
* ***dim\_date*** contains calendar date information, broken down into month, day, year, day of the week, weekday vs workday, and whether or not it is a holiday.
* ***bridge*** relates the 3 other tables to each other by matching up each claim to the relevant PA (if a PA was needed) and the date the claim was submitted.

There are several problems or questions that can be answered using this dataset:

* Can we predict if a claim will be approved based on the Payer and its formulary (ie. will a PA need to be issued)?
* If the claim is rejected and a PA issued, can we predict if the PA will be approved based on the Payer and its formulary, and the contents of the PA?
* Does a Payer’s formulary change over time, and if so, how does it change and can we predict a Payer’s future formulary?

**Project Stakeholders**

The analytical and predictive results derived from this dataset will be relevant to business stakeholders, where an understanding of which claims will require PA will help to predict PA volume, which is a predictor of revenue (per CoverMyMeds project description).

**Planned Approach**

Most of the data presented in the tables is categorical and/or binary in nature, and several of the questions are classifying whether or not we can predict an outcome (for example, approve vs. reject), so I think an initial approach will be to construct a Decision Tree or Random Forest classifier model. To answer the question of if/how a formulary changes with time, Time Series Analysis will be used.

* Initial analysis of the data shows that there are a total of 4 unique Payers in the claims data, with 48% of all claims going to Payer#999001, 23% to Payer#417614, 16% to Payer#417740, and 13% to Payer#417380.
* Out of all of the claims, 51% are for Drug A, 26% are for Drug B, and 23% are for Drug C. When looking at each individual Payer, the relative number of claims for each drug submitted to that Payer follows this same split.
* Among all Payers, 58% of all claims are approved, 19% are rejected with Code 70, 16% are rejected with Code 75, and 7% are rejected with Code 76.