Overview

As follows is a simple overview on some basic ideas on the improvement to the system that would allow for easy quality checks and to avoid potential issues for the given data set

Introduction

The used datasets here are mainly performance data of assets stored primarily on CSV and Excel files, and bond payments are reported in a PDF document.

Primary Issues

1. Accessing /Downloading loan level data for each month

* The way on how data is reported for ‘Loan-Level Data’ is currently sorted by month. At first glance, this may prove to be easy to spot errors as data is organized on a monthly basis, however further review would prove that running queries for each month will take up a significant amount of time.

1. Issues with data formats on certain columns / datapoints

* Date formats seem to be formatted oddly which may prove harder to draw conclusions from.
* Some columns don’t have data at all
* There are columns that don’t indicate what data they are supposed to be. (Other data 1. etc.)
* Two of the 12 months seem to have additional data columns absent on the rest (March 2007, July 2007)
* Data type mismatch on certain months

Proposed Solution

1. Create a database which compiles most if not all data sets from each month on a yearly basis. This will allow the analyst to run all queries on a single file, created database needs to have less bloat, data not needed for the queries needed to be run should stay separated.
2. In conjunction on a joint database for easier queries, additional data column for year needs to be added.
3. Data formatting / data type should match in all the monthly records