Case Study 1: Working with Edgar datasets: Wrangling, Pre-processing and exploratory data analysis

Team 4: Emily Strong and Raksha Kaverappa

Main Code Repository: <https://github.com/erstrong/INFO-7390-ADS-Fall-17-TeamNo.4/tree/master/Assignment%202>

**Part 1A: Data Wrangling**

*Code:*

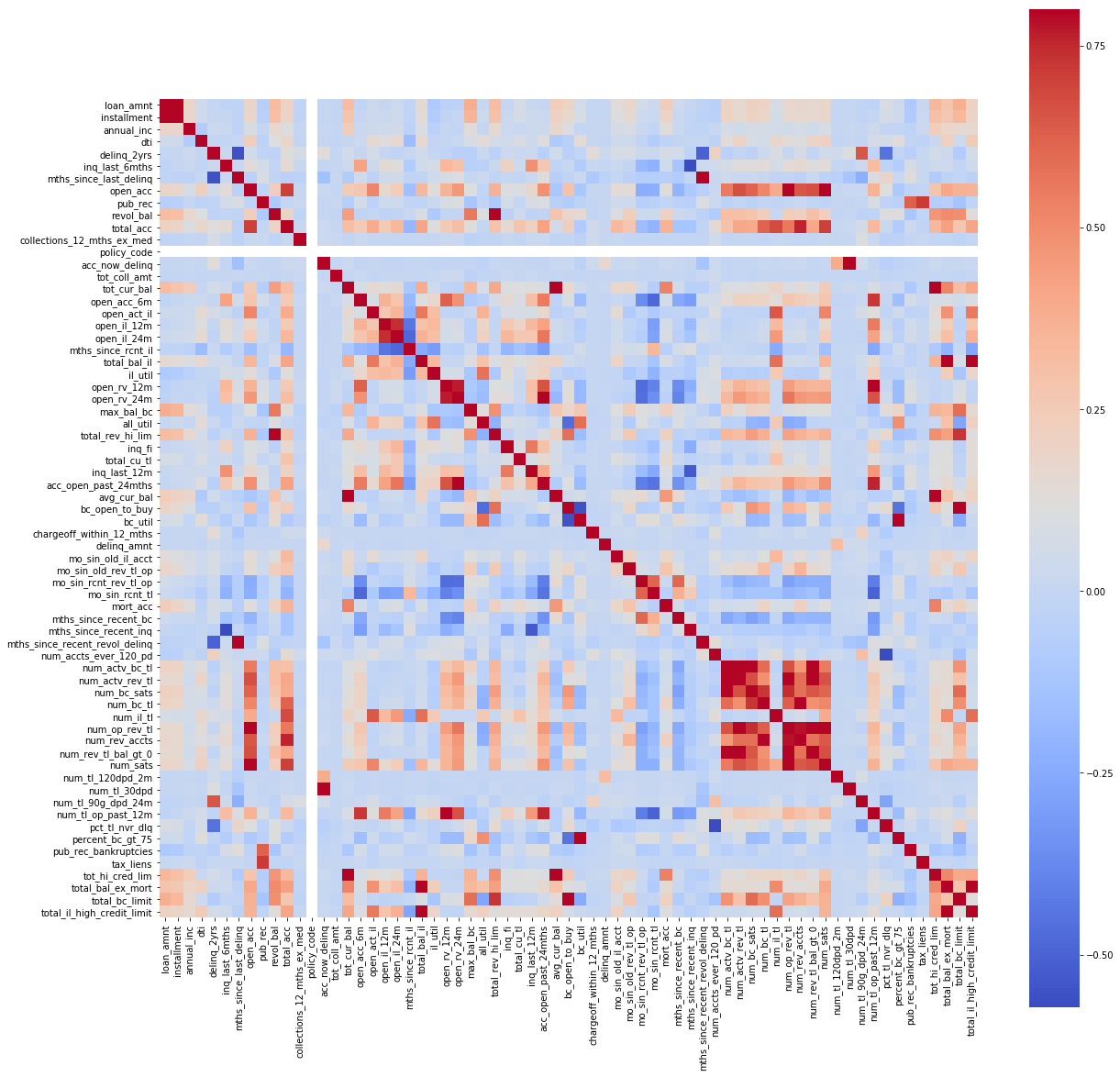
[ADD LUIGI NOTES]

To scrape the approved loan data:

* We created a session with Lending Club investor credentials.
* We scraped from a hidden div on the page the names of the zip files that need to be accessed, and then read the csvs in each zip file into a data frame.
* We then merged the data frames and added a timestamp and the set name.
* We dropped all columns with >50% missing data.
* We removed columns that are computed by Lending Club after the loan application based on the data dictionary.
* We used a correlation heat map (shown below) and the data dictionary to eliminate redundant columns.
* We did missing value substitutions using mean, median, mode and max for:
* We dropped any rows with missing values for the remaining features, cumulatively removing 6% of the data.
* We reformatted numeric features present as strings, and created dummy variables for the categorical features.

To scrape the declined data:

* We followed a similar process but without the authenticated session.
* The only feature with >50% missing data was Risk Score, with 58%. We chose to keep it because of the critical
* For Risk Scores, we converted Vantage scores (applications dated on or after November 6, 2013 ) to use the same scale as FICO scores.
* We dropped rows with missing location information, and used mean and mode substitutions for the remaining features.

****

**Correlation Heat Map for Features with <=50% missing data**

**Part 1B: Exploratory Data Analysis**

**Part 2A: Classification**

**Part 2B: Clustering**

**Part 2C: Prediction**

**Part 2D: Deployment**