Goals and Intended Uses

This project envisioned using data on student loan indebtedness, records of legislation introduced on the subject, and election results to explore whether candidates and parties could find cause to adjust their engagement on the issue, including campaign strategy and legislative activity while in office. The ETL process could be easily modified without change in data sources to examine consumer (credit card) debt and mortgage debt. With a change of data sources, other topics could be examined. (Currently, issue-based canvasing and campaining is used only for high-profile state (gubernatorial) and national races (US congressional, presidential).) Measures of priority are proposed to equate more legislative time spent on an issue, as represented by the number of bills introduced during the previous legislature, and how issue-based bills' sponsors fared in a subsequent election.

Extraction: Data sources and formats:

A. Senate Election Results (1976-2018-senate.csv) — **CSV**
[Statistics of the Congressional Election](http://history.house.gov/Institution/Election-Statistics/Election-Statistics/)

The file contains state-granular returns for US Senate elections, by state, from 1976 to 2018. This report is published biennially by the Clerk of the House of Representatives, derived from official state election websites, with an accompanying documentation text file, Codebook for U.S. Senate Returns 1976–2018. We limited analysis to Senate races in 2016.

- Year/office
- State abbreviation;
- Candidate;
- Party;
- candidate votes;
- write-in status.

Parties are as they appear in the House Clerk report. In states that allow candidates to appear on multiple party lines, separate vote totals are indicated for each party.

B. Consumer Debt (ConsDebt2003-2018XStatePerFed.xlsx) – **Excel**
https://www.newyorkfed.org/medialibrary/Interactives/householdcredit/data/xls/area_report_by_year.xlsx

State Level Household Debt Statistics 2003-2018,
Federal Reserve Bank of New York, March, 2019.
Sources: New York Fed Consumer Credit Panel and Equifax.
Tabbed excel spreadsheets by year; Each year's sheet gives values for each state in 1 row.

Mortgage Debt Balance per Capita
% Mortgage Debt Balance 90+ Days Delinquent
Credit Card Debt Balance per Capita
% Credit Card Debt Balance 90+ Days Delinquent
Student Loan Debt Balance per Capita
% Student Loan Debt Balance 90+ Days Delinquent (and in default)**

**sample size of 1% of the population; other delinquency data represents 5% of population.
These data exclude US Territories and are subject to sampling variation such that national and state totals may not match those reported by other sources such as the Quarterly Report on Household Debt and Credit.

C . Senate Congressional Bills — Ibrahim_Web_Scraping_Code.ipynb — **Scrape**
https://www.congress.gov/advanced-search/legislation

Congress.gov is the official website for U.S. federal legislative information, via Library of Congress (LOC), using data from the Office of the Secretary of the Senate. The scope of our project did not justify download of 3000+ pieces of legislation for each of 35 years, when likely analysis would focus on one

2-year legislative period, or 1 "congress". The catalog's interface narrowed our search using search keywords [student, education, loan, debt] that resulted in this QUERY string in the address bar:

collection:(BILLS) AND publishdate:range(,2019-05-04) AND congress:(115)
AND chamber:(Senate or Joint) AND content:(student or education and debt or loan)

The resulting URL was used for scraping, using Beautiful Soup:

https://www.congress.gov/search?searchResultViewType=expanded&q={%22congress%22:[%22115%22],%22source%22:%22legislation%22,%22search%22:%22student+debt+loan+education%22,%22chamber%22:%22Senate%22}&pageSize=250'

Scraping retrieved a dataset of Senate legislation from the 115th Congress (Jan2017 to Jan2019).
Retrieval of HTML tags yielded lists, which were then converted to a dataframe:

date_introduced_list = []	senator_details_list = []	cosponsors_list= []
bill_num_list = []	last_name_list = []	party_list = []
bill_name_list = []	first_name_list = []	state_list = []

congress_bills_on_debt_data = pd.DataFrame({

'Date_Introduced':	'Senator_Details':	'Party':
'Bill_number':	'Last_Name':	'State':
'Bill_Name':	'First_Name':	'Number_of_Cosponsors':

GUI-retrieved data would have looked like this:

Scraped data print like this:

```
11/29/2017
S.2169
Student Right to Know Before You Go Act of 2017
Sen. Wyden, Ron [D-OR]
Wyden,
Ron
3
D
OR
```

Transform: Data cleaning and transformation

A: Election voting data: (1976-2018-senate.xlsx) —> [1976-2018-senate.csv]
discrete element = bill number 756 rows, with 521 discrete candidates.
Differently spelled entries of candidates' names were reconciled.
Only general elections, including special elections, were included.
Write-ins (uncontested races) largely show as N/A in the candidate column, c/w source data.
Where names were available, write-ins were included, for completeness.
The incumbent is named in the "Incumbent" column for all candidates in a race.
The incumbent is noted with a Boolean in the "Incumbent?" column.
Candidates running under multiple parties are noted in the Boolean column "Duplicate?"
A manual Boolean column "Other than N/A?" was used in a manual verification process.
Candidates running for >1 party get a "total votes" line, with "et al" in the party designation.
Any manual changes not derived by Excel formulas are highlighted yellow in the Excel file.
Limited to data after 2005, sorted by year/asc, state/asc, & candidatevotes/descending.
Only included columns "year","state","state_po","candidate","party","candidatevotes","totalvotes".

B. Consumer Debt: (ConsDebt2003-2018XStatePerFed.xlsx) —> [debt.csv]

Keep only columns showing indebtedness: pull each kind of indebtedness (consumer, mortgage and student) into a separate pandas dataframe. Limit to years 2017-2018.

Calculated deliquency of debt as a percentage, rather than absolute value.

Eliminated Puerto Rico and "Total" rows, keeping just the states.

Averaged the values of 2017 and 2018 in all dataframes to a new column.

Merged the 8 debt type dataframes into one dataframe with 8 columns.

Average 2 years' values, for congruence with 2-year congressional period.

Eliminate Puerto Rico and "Total" rows, keep only rows for each state.

Set the state column as the index prior to merging.

Merged the 8 dataframes into one with 8 columns.

C . <u>Senate Congressional Bills</u> — Ibrahim_Web_Scaping_code.ipynb **— Scrape**

discrete element = bill number 69 records, no missing values

Date reconciliations: Federal Reserve data is reported 4th quarter of a calendar year, December 31. Election results are mostly announced November 6. Congress takes their seats January 3.

```
[364]: congress_bills_on_debt_data.head()
```

[364]:

	Date Introduced	Bill number	Bill Name	Senator Details	Last Name	First Name	Party	State	Number of Cosponsors
0	11/29/2017	S.2169	Student Right to Know Before You Go Act of 2017	Sen. Wyden, Ron [D-OR]	Wyden,	Ron	D	OR	3
1	04/07/2017	S.888	Understanding the True Cost of College Act of ...	Sen. Grassley, Chuck [R-IA]	Grassley,	Chuck	R	IA	9
2	03/30/2017	S.799	Dynamic Repayment Act of 2017	Sen. Warner, Mark R. [D-VA]	Warner,	Mark	D	VA	1
3	11/16/2017	S.2155	Economic Growth, Regulatory Relief, and Consum...	Sen. Crapo, Mike [R-ID]	Crapo,	Mike	R	ID	26
4	06/20/2017	S.1384	Joint Consolidation Loan Separation Act	Sen. Warner, Mark R. [D-VA]	Warner,	Mark	D	VA	3

These functionally concur in reference to the 2-year period of activity of the 115th Congress.

* Election data from 2014, 2012, 2010 was included in election data <u>only</u> to verify incumbent candidates.

Load: Create Database, Tables, Indexes

Although smaller files (50 states in 2 of 3 sources, 500 rows in the 3rd) do not require the abilities of a relational database, yet all out data adheres to state-level granuarity, a 1- to 2-year time period, and a limited number of senators (100) and no data types exceeded schema representations of a SQL structure.

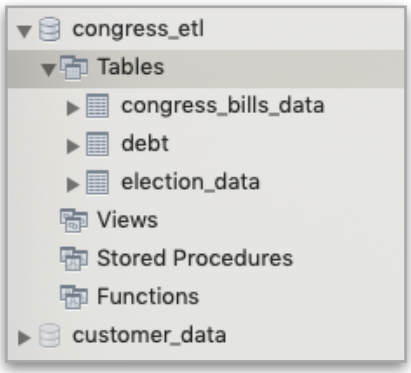
Tables Structure and Datatypes are defined on the Sql Script (Creating_Congress_ETL_DB.sql)

Database name: congress_etl

Tables: congress_bills_data
debt
election_data

Common element between tables is the column "state_po" (2-letter state abbreviations)

Tested by querying "SELECT *"



FYI only, Contextual information:

SUDOC numbers are applied to all congressional public documents.

The système universitaire de documentation or SUDOC is a system used by the libraries of French universities and higher education establishments to identify, track and manage the documents in their possession. The catalog contains more than 10 million references, allowing a search for bibliographical and location information in over 3,400 documentation centers. The Superintendent of Documents (SuDocs) is in charge of the dissemination of information at the GPO. This is accomplished through the Federal Depository Library Program (FDLP), the Cataloging and Indexing Program and the Publication Sales Program. (Wikipedia)

The United States Government Publishing Office (GPO) (formerly the Government Printing Office, until 2014) is a legislative branch agency of the United States federal government that is required by law to produces and distributes information products and services for all three branches of the Federal Government, including U.S. passports for the Department of State as well as the official publications of the Supreme Court, the Congress, the Executive Office of the President, and executive departments, and independent agencies, and to maintain access and integrity of those publications.

Congress.gov supersedes the THOMAS system which was retired on July 5, 2016, to make federal legislative information freely available to the public.

<u>Background information on the gravity of Student Loan Indebtedness:</u>

Institutions of higher education play a critical role in supporting and promoting students' overall financial health and well-being. (Consumer Financial Protection Bureau's, or CFPB's, report on Consumer Financial Wellness, 2016) This is particularly true when students are first time participants in the marketplace for consumer financial products and services. The Credit Card Accountability, Responsibility, and Disclosure Act ("CARD Act") requires that the CFPBureau publish five annual reports annually, including one on just the college credit card market in particular. In a departure from prior practice, the CFPB has begun releasing all current and historical data collected by the Bureau and the Federal Reserve in a single, consolidated dataset alongside this report to facilitate examination of the ollege credit card market.

More than 10 million students (approximately 40 percent) attend a college or university that has an agreement with a financial institution to offer college-sponsored deposit or prepaid accounts. The CARD Act required institutions to disclose these agreements, and to abide by restrictions on the marketing of credit cards to students on or near college campuses via gifts or any tangible inducement toward application for credit cards, and to ensure the best financial interests by conducting periodic reviews of fees.

The Bureau's research into what contributes to financial well-being underscores the critical role that colleges play as trusted sources of information for their students, as young people develop their financial decision-making skills and navigate a marketplace where missteps can drive up the cost of college. (Student banking, Annual report to Congress, Consumer Finanial Protection Bureau, December 2016)