Capstone Project B: Data cleaning:

Data cleaning was completed in ipython notebooks.
3 clean data files were created plus two lookup tables so data could be re-personalized as needed.

Voter data:

The ipython notebook can be found here:
LocalVoters/05_Cleaning Voter.ipynb

Each Column of my data file was considered

Original Data Column	Description of action	output column(s)	Type
'VoterID'	Rows of table were randomly shuffled, the index reset and the new index used as new UID.	'vid'	Num
'Status'	Dropped, all entries are the same 'A'.		
'Abbr'	Kept as is, although it's not understood it is a clean number.	'Abbr'	Num
'Affidavit'	Dropped as PII.		
'LastVoted'	Only 6 entries column dropped as too sparse.		
'Salutation'	Dropped PII, 6388 missing values.		
'LastName'	Dropped as PII.		
'FirstName'	Used to help fill missing gender data then dropped as PII.		
'MiddleName'	4020 NaN, and a lot of initials only, dropped as PII.		
'Suffix'	13057 NaN, dropped as PII.		
'HouseNumber'	Dropped as PII.		
'HouseNumberSuffix'	Dropped empty.		
'StreetPrefix'	Dropped empty.		
'Street'	Dropped as PII.		
'StreetType'	Populate missing values using 'Street': 'Common' => 'CMN' 'GREEN' => 'GRN' and two cross streets => 'UKN'.	'StreetType'	Cat
'BuildingNumber'	Only 3 entries dropped.		

'ApartmentNumber'	Converted to a True/False field.	'isApt'	Bool
'City'	Dropped, all entries are the same.		
'State'	6 missing rows, dropped as all should be the same.		
'Zip'	Cleaned all to 5 digit numerical zip code entries.	'Zip'	Num
'Precinct'	Converted to number and kept.	'Precinct'	Num
'PrecinctSub'	Converted to number and kept.	'PrecinctSub'	Num
'Party'	Converted to category and kept.	'Party'	Cat
'RegDate'	Converted to a dateTime and kept.	'RegDate'	Date
'ImageID	An int between 0 and 48277945 meaning unknown, dropped as Images are often PII.		
'Phone1'	5266 NaN's 8041 values, converted to True/False.	'havePhone'	Bool
'Phone2'	Only 2 values, dropped as PII.		
'Military'	Only 9 'Y' dropped too sparse.		
'Gender'	5223 NaN's 1743 ‘F’ and 1717 ‘M’ were added by comparing to a database of name genders (https://github.com/organisciak/names), remaining missing data was set to ‘UKN’.	'Gender'	Cat
'PAV'	Is voter a Permanent Absentee Voter, converted to category and kept.	'PAV'	Cat
'BirthPlace '	This mixed two and three letter code was assumed to be a two USA state code, and only if that failed to match assumed to be a two or three letter country code. Output was 2 clean columns, state and country code data gathered from wikipedia, 'UNK' added for the 1296 NaN's.	'BirthPlaceState' 'BirthPlaceCountry'	Cat, Cat
'BirthDate'	Cleaned full birthday into 'BirthYear', rest dropped as PII.	'BirthYear'	Int
Mailing Address columns	Compared with main address to create a True/False, Country kept as a category.	'sameMailAddress’ 'MailCountry'	Bool Cat
'LTDate'	An internal column to registration office, dropped.		
'email'	9009 NaN's, Cleaned to keep the service provider with UKN for NaNs.	'EmailProvider'	Str
'RegDateOriginal'	Converted to a dateTime and kept.	'RegDateOriginal'	Date
'PermCategory'	An internal column to registration office, dropped.		
'PrecinctName'	Shown to be formatted combination of Precinct and PrecinctSub so dropped.		
'ResAddrLine1'	Dropped empty.		

'ResAddrLine2'	Dropped empty.		
'E1_110816'	Code indicated vote, converted to category and kept.	'E6_110816'	Cat
'E2_060716'	Code indicated vote and ballot used, Cleaned into 'Vote' and 'BallotType' and kept.	'E5_060716' 'E5_060716BT'	Cat
'E3_110414'	Code indicated vote, converted to category and kept.	'E4_110414'	Cat
'E4_060314'	Code indicated vote, converted to category and kept.	'E3_060314'	Cat
'E5_110612'	Code indicated vote, converted to category and kept.	'E2_110612'	Cat
'E6_060512'	Code indicated vote and ballot used, Cleaned into 'Vote' and 'BallotType' and kept.	'E1_060512' 'E1_060512BT'	Cat
	Added column to indicate number of elections voter has been registered for.	'Tot_Possible_Votes'	Num
	Added column to indicate number of elections voter actually voted in.	'Act_Votes'	Num
	Added column to indicated % of possible elections actually voted in.	'Pct_Possible_Votes'	Num
'District'	Kept as is in case we need to add in other district data.	'District'	Num
'VoterScore'	Score assigned by my friend based on which election someone has reported data for and voted (A or V)	'VoterScore'	Num
'VoterScorePossible'	Score assigned by my friend assuming all reported data was 'vote' (A or V)	'VoterScorePossible'	Num
'VoterScorePctOfPoss'	'VoterScore'/'VoterScorePctOfPoss'	'VoterScorePctOfPoss'	Num
working cols	Other columns related to my friends modeling will be dropped, 'CityArea' will be brought back via our own cleaning and merging later. Columns dropped include 'UpdateStatus', 'cc_2016_email', 'cc_2016_lists', 'cc_2016_tags', 'cc_2018_active', 'age', 'oldest_in_household', 'IsApartment', 'CityArea'		
Household	Unique key linking each voter to a household, looked up and converted to anonymized Hid.	'Hid'	Num

After the cleaning and data anonymizing steps were taken the clean data and a lookup table new ‘VID’ to old ‘VoterID’ were saved for further analysis and modeling

Household data:

The ipython notebook can be found here:
LocalVoters/05_Cleaning HouseHold.ipynb

Each Column of my data file was considered:

Original Data Column	Description of action	output column(s)	Type
'Household_Id'	Rows of table were randomly shuffled, the index reset and the new index used as new UID.	'hid'	Num
'FullAddress'	Shown to be concatenation of 'HouseNumber','Street','StreetType', 'BuildingNumber & 'ApartmentNumber' in all but 4 cases where the Apt numbers appeared to be missing or include typo's and then dropped as PII.		
'HouseNumber'	Dropped as PII.		
'HouseNumberSuffix'	Dropped empty.		
'StreetPrefix'	Dropped empty.		
'Street'	Used to clean 'StreetType' and then Dropped as PII.		
'StreetType'	‘CMN’ ‘GREEN’ => ‘GRN’ and two cross streets => ‘UKN’.	'StreetType'	Cat
'BuildingNumber'	Only 3 entries dropped.		
'ApartmentNumber'	Converted to a True/False field.	'isApt'	Bool
'City'	Dropped, all entries are the same.		
'State'	6 missing rows, dropped as all should be the same.		
'Zip'	Cleaned all to 5 digit numerical zip code entries.	'Zip'	Num
'Precinct'	Converted to number and kept.	'Precinct'	Num
	The Precinct was also used to create a 'CityArea' column	'CityArea'	Cat
'PrecinctSub'	Converted to number and kept.	'PrecinctSub'	Num
'District'	Kept as is in case we need to add in other district data.	'District'	Num

After the cleaning and data anonymizing steps were taken the clean data and a lookup table new ‘HID’ to old 'Household_Id' were saved for further analysis and modeling

Election data:

The ipython notebook can be found here:
LocalVoters/05_Election information.ipynb

Some thought was put into what attributes of the each election could affect a voters’ likelihood to vote and these were gathered into a 3rd data file. Most of the data came from wikipedia.

output column(s)	Description of data	Type

'election'	Unique ID for each election held between 2012 and 2018, including this years target for predicting voting.	str
'dates'	Actual date of the election.	dt
'cycle '	Is this a Congressional only election year or a Presidential election year.	cat
'etype'	Is this a 'Primary' or 'General' election.	cat
'president'	What is the party of the president in power at the time of the election.	cat
'us_senate'	Which party has control of the US Senate at the time of the election.	cat
'us_senate_maj'	How big is the controlling margin in the US Senate.	Num
'us_repre'	Which party has control of the US House of Representatives at the time of the election.	cat
'us_repre_maj'	How big is the controlling margin in the US House of Representatives.	Num
'ca_governor'	What party did the Governor of CA belong too at the election.	cat
'ca_lt_govnor'	Who was the Lieutenant Governor of CA belong too at the election.	cat
'ca_senate'	Which party has control of the CA Senate at the time of the election.	cat
'ca_senate_maj'	How big is the controlling margin in the CA Senate.	Num
'ca_assembly'	Which party has control of the CA Assembly at the time of the election.	cat
'ca_assembly_maj'	How big is the controlling margin in the CA Assembly.	Num