DATA CLEANING REPORT

Index

Consistency

1. unified enrollment casing
2. status value mapping

   status_map = {'ACTIVE': 'A', 'PURGED': 'P', 'PREREG': '17', 'INACTIVE': 'I'}

3. reasoncode value mapping

   df_cleaning.loc[:,'reasoncode'].replace('MAIL-CHECK', 'MAILCHECK')

4. prevcounty == '1 '

   couldn't determine if it meant '10' or '01', since both are present in the data - set all 5 records to nan

Voters Removed

(removed all voters with corrupted data to ensure continuity)

1. empty rows (2)
2. invalid dates   61 voters

   datetime64 only supports year between (1677, 2262)

3. gender == ' ' (space)   250 voters
4. dob < '1900-01-01'   22844 voters
5. register age ('year_diff') < 16   7878 voters
6. reasoncode: 'COURT' - mentally incompetent, felon   69652 voters
7. reasoncode: 'DUPLICATE'   7 voters
8. year_diff > 120 (to filter out as many bad dates as possible)   8 voters
9. wrongfully assigned status = '17': status = '17' when register age > 18   7 voters
10. come back from being dead: record kept getting updated after reasoncode == 'DEATH'   253 voters

Other Data Quality Issues & Concerns

1. changed birthdays: found out a list of voters whose birthdays changed, pickled the list 21375 voters

2. repeating register dates: too many to remove 11437367

3. rzip5 has bad values (only has 4 digits)

concerns:

voter register dates might be too damaged to reflect the switching time

Columns Removed

fname, lname, middle name, suffix, mailadd2, mailadd3, mailadd4, lastvotedyear (empty column)

Features Created

1. enrollment: merged original enrollment and otherparty

2. year_diff: float64, (regdate - dob)/ 365.25

3. switched: binary, if the voter changed their party affiliation in this record, switched = 1, otherwise 0

4. from: if the voter changed their party affiliation in this record, from = the option they switched from, otherwise np.nan

5. parsed dates: dob, regdate, lastvoteddate

   columns created:

dob_year	category
lastvoteddate_year	category
lastvoteddate_month	category

lastvoteddate_day	category
lastvoteddate_dayofweek	category
regdate_year	category
regdate_month	category
regdate_day	category
regdate_dayofweek	category

verified facts:

all reason codes match status

according to nys website:

inactive reasons: MAILCHECK, RETURN-MAIL, NCOA; purged reasons: all of the rest

Match.

Result Data File & Data Types

Converted columns to their proper nullable types then saved as a parquet file (2G) for memory efficiency.

33000000+ rows

Data Types in .parquet file:

sboeid	string[python]
dob	datetime64[ns]
gender	category
regdate	datetime64[ns]
vrsource	category
status	category
reasoncode	category
inact_date	datetime64[ns]

```
purge_date              datetime64[ns]
lastvoteddate           datetime64[ns]
idrequired                   category
idmet                       category
voterhistory            string[python]
year_diff                    float64
enrollment                    category
rhalfcode              string[python]
rpredirection           string[python]
rstreetname             string[python]
rpostdirection          string[python]
rapartmenttype            string[python]
rapartment              string[python]
raddrnonstd             string[python]
rcity                     category
rzip4                      category
mailadd1               string[python]
countycode                     category
ed                        category
ld                        category
towncity               string[python]
ward                        category
cd                        category
sd                        category
ad                        category
prevcounty                    category
prevaddress             string[python]
prevname                string[python]
```

```
countyvrnumber          string[python]
switched                      bool
from                      category
dob_year                    category
lastvoteddate_year            category
lastvoteddate_month             category
lastvoteddate_day             category
lastvoteddate_dayofweek         category
regdate_year                  category
regdate_month                   category
regdate_day                   category
regdate_dayofweek               category
rzip5                     category
dtype: object
```

NYS Voting Law Research Reference

1. ID REQUIREMENT

ID REQUIRED

When registering to vote, the Board of Elections checks the identity of the voter before Election Day, through the voter's provided DMV number (driver's license number or non-driver ID number), or the last four digits of their social security number.

If the voter does not have a DMV or social security number, they may use a valid photo ID, a current utility bill, bank statement, paycheck, government check or some other government document that shows their name and address. They can also submit a copy of one of those types of ID with their voter registration form.

If the Board is unable to verify a voter's identity before Election Day, they will be asked for ID when voting for the first time.

ID MET

If it's not met can still vote affidavit.

2. VOTER STATUS AND REASON CODE

(1) Active. The voter is properly registered and is eligible to vote in elections. There are several categories of active voters, as specified below:

(i) active;

(ii) active miliary;

(iii) active UOCAVA;

(iv) active special presidential; and

(v) active special Federal.

(2) Inactive. The voter is still eligible to vote in elections, but is not included in the poll book. NYSVoter shall allow a county election official to designate a voter as inactive, noting the reason for the designation, such as "election material mailed to registrant returned as undeliverable" or "moved with an out of county forwarding address", "affidavit ballots".

(3) Purged. The voter is no longer eligible to vote in an election, and will not appear in the list of registered voters. This list is to be utilized to prevent deceased voters from overwhelming valid voters when doing voter searches and to allow for voters who later re-register to vote to resurrect and utilize their unique identifier.

(4) Pre-registered. The voter has met all the requirements to be an active voter but has not yet attained the age of 18. Pre-registered voters that will be 18 years old on or before the election date are included in the poll book and are eligible to vote in the election. The voter must be at least 16 years old to pre-register.

(5) NYSVoter store reason codes for inactive and purged voters indicating or explaining the reason for a specific voter's status, as follows:

(i) inactive status - mail check, NCOA, returned mail;

(ii) purged status - death, voter request, felon, ADF (adjudicated) incompetent, NVRA, moved out of country.