extract_data

October 20, 2016

1 ANES Longitudinal Dataset Extraction Code

The purpose of this script is to take the original anes dataset and extract it into just the desired portions. Currently it simply extracts the thermometer data.

In [13]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

In [19]:
```
# determine which variables are relevant
dfinfo = pd.read_excel('../data/thermometer_vars.xlsx')
dfinfo.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 43 entries, 0 to 42
Data columns (total 2 columns):
VarName        43 non-null object
Description    43 non-null object
dtypes: object(2)
memory usage: 1.0+ KB
```

In [76]:
```
# pull in whole dataset
df = pd.read_stata('../data/anes_timeseries_cdf.dta', columns=dfinfo['VarName'], convert_categ
#df.info()
#df['VCF0201']
```

In [79]:
```
# convert data types to be numeric
print(np.unique(df['VCF0201']))
df.apply(pd.to_numeric).info()
```

```
[  0.   2.   3. ...,  nan  nan  nan]
<class 'pandas.core.frame.DataFrame'>
Int64Index: 55674 entries, 0 to 55673
Data columns (total 43 columns):
VCF0004     55674 non-null float64
VCF0006a    55674 non-null float64
VCF0013     55674 non-null int64
VCF0014     55674 non-null int64
VCF0140a    55012 non-null float64
VCF0301     55012 non-null float64
VCF0201     15486 non-null float64
VCF0202     15486 non-null float64
VCF0203     12690 non-null float64
VCF0204     28920 non-null float64
```

```
VCF0205     20749 non-null float64
VCF0206     43980 non-null float64
VCF0207     39824 non-null float64
VCF0208     15714 non-null float64
VCF0209     35618 non-null float64
VCF0210     41055 non-null float64
VCF0211     43980 non-null float64
VCF0212     43980 non-null float64
VCF0213     34039 non-null float64
VCF0214     13368 non-null float64
VCF0215     14082 non-null float64
VCF0216     13946 non-null float64
VCF0217     26919 non-null float64
VCF0218     32319 non-null float64
VCF0219     19847 non-null float64
VCF0220     31075 non-null float64
VCF0221     5289 non-null float64
VCF0222     14761 non-null float64
VCF0223     36636 non-null float64
VCF0224     32319 non-null float64
VCF0225     23863 non-null float64
VCF0226     9354 non-null float64
VCF0227     15251 non-null float64
VCF0228     21900 non-null float64
VCF0229     18480 non-null float64
VCF0230     6277 non-null float64
VCF0231     20619 non-null float64
VCF0232     24338 non-null float64
VCF0233     15768 non-null float64
VCF0234     20800 non-null float64
VCF0235     9649 non-null float64
VCF0236     5894 non-null float64
VCF0253     17291 non-null float64
dtypes: float64(41), int64(2)
memory usage: 18.7 MB
```

```
In [84]: df.to_hdf('../data/anes_timeseries_thermometer.h5', 'main')
         df.iloc[1:100,:].to_csv('../data/anes_timeseries_thermometer_preview.csv')
```