In [1]:

```
import pandas as pd
import warnings
import pandas as pd
from pandas.core.common import SettingWithCopyWarning
warnings.simplefilter(action="ignore", category=SettingWithCopyWarning)
```

In [2]:

```
list_excels = ['datasets/rollingsales_queens.xls', 'datasets/rollingsales_bronx.
```

In [3]:

```
for excel in list_excels:
    excel_df = pd.read_excel(excel, skiprows=4, header=[0])
    excel_df_cleaned = excel_df[['TAX CLASS AT PRESENT','ZIP CODE', 'RESIDENTIAL
    excel_df_cleaned.dropna(inplace = True)
    excel_df_cleaned.reset_index(drop=True)
    excel_df_cleaned.to_csv((str(excel) + '_prepped' + '.csv'),index=False)
    print('--------------', excel, '-----------------------')
    print('before: ', len(excel_df.index))
    print('after: ', len(excel_df_cleaned.index))
```

```
-------------- rollingsales_queens.xls -----------------------
before:  20945
after:  14325
-------------- rollingsales_bronx.xls -----------------------
before:  6139
after:  4181
-------------- rollingsales_brooklyn.xls -----------------------
before:  19244
after:  11778
-------------- rollingsales_manhattan.xls -----------------------
before:  12190
after:  1009
-------------- rollingsales_statenisland.xls -----------------------
before:  6483
after:  5470
```

In [4]:

```
#For some reason, no one likes reporting square footage in manhattan....I wonder
# In that case, for manhattan specifically, we won't analyze square footage
# At least this time around, onle 2K rows were lost rather than 11k rows
list_excels=['datasets/rollingsales_manhattan.xls']
for excel in list_excels:
    excel_df = pd.read_excel(excel, skiprows=4, header=[0])
    excel_df_cleaned = excel_df[['TAX CLASS AT PRESENT','ZIP CODE', 'YEAR BUILT',
    excel_df_cleaned.dropna(inplace = True)
    excel_df_cleaned.reset_index(drop=True)
    excel_df_cleaned.to_csv((str(excel) + '_prepped' + '.csv'),index=False)
    print('--------------', excel, '-----------------------')
    print('before: ', len(excel_df.index))
    print('after: ', len(excel_df_cleaned.index))
```

```
-------------- rollingsales_manhattan.xls -----------------------
before:  12190
after:  10761
```

In [5]:

```
excel_df_cleaned.isna().sum()
```

Out[5]:

```
TAX CLASS AT PRESENT    0
ZIP CODE                0
YEAR BUILT              0
SALE PRICE              0
SALE DATE               0
dtype: int64
```

In [6]:

```
excel_df.isna().sum()
#Either they are lazy at filling out values or they don't care, but I have less
```

Out[6]:

```
BOROUGH                                0
NEIGHBORHOOD                           0
BUILDING CLASS CATEGORY                0
TAX CLASS AT PRESENT                  30
BLOCK                                  0
LOT                                    0
EASE-MENT                          12190
BUILDING CLASS AT PRESENT             30
ADDRESS                                0
APARTMENT NUMBER                    6459
ZIP CODE                               0
RESIDENTIAL UNITS                   5832
COMMERCIAL UNITS                   10716
TOTAL UNITS                         5418
LAND SQUARE FEET                   11130
GROSS SQUARE FEET                  11130
YEAR BUILT                          1426
TAX CLASS AT TIME OF SALE              0
BUILDING CLASS AT TIME OF SALE         0
SALE PRICE                             0
```

In [7]:

```
#I will revisit these datasets with just time and value amounts and tax class to
# It seems tax class, sale price, sale date
list_excels=['datasets/rollingsales_queens.xls', 'datasets/rollingsales_bronx.xl
for excel in list_excels:
    excel_df = pd.read_excel(excel, skiprows=4, header=[0])
    excel_df_cleaned = excel_df[['TAX CLASS AT PRESENT', 'ZIP CODE', 'SALE PRICE
    excel_df_cleaned.dropna(inplace = True)
    excel_df_cleaned.reset_index(drop=True)
    excel_df_cleaned1 = excel_df_cleaned[excel_df_cleaned['SALE PRICE'] > 10]
    excel_df_cleaned1.to_csv((str(excel) + '_prepped_bare' + '.csv'),index=False
    print('--------------', excel, '-----------------------')
    print('before: ', len(excel_df.index))
    print('after: ', len(excel_df_cleaned1.index))
```

```
-------------- rollingsales_queens.xls -----------------------
before:  20945
after:  13171
-------------- rollingsales_bronx.xls -----------------------
before:  6139
after:  3982
-------------- rollingsales_brooklyn.xls -----------------------
before:  19244
after:  11624
-------------- rollingsales_manhattan.xls -----------------------
before:  12190
after:  9234
-------------- rollingsales_statenisland.xls -----------------------
before:  6483
after:  4515
```

In [ ]: