UFRN UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

IMD INSTITUTO METRÓPOLE DIGITAL

IMD0905 - Data Science I

Lesson #17 - Data Cleaning Walkthrough: combining the data

Ivanovitch Silva
November, 2018

Agenda

- Case study: NYC open data (education)
- Data cleaning walkthrough
- Combining data
- **Groupby**
- **Merge (inner, outer, right, left)**

Update the repository

```
git clone https://github.com/ivanovitchm/IMD0905_datascience_one.git
```

Or ….

```
git pull
```

NYC OpenData

Data cleaning vs Storytelling

Controversial issues in the U.S. : educational system is the efficacy of standardized tests, and whether they're unfair to certain groups

Combining the Data

sat_results

DBN	...
01M022	...
05M345	...
02M456	...
99M520	...

+

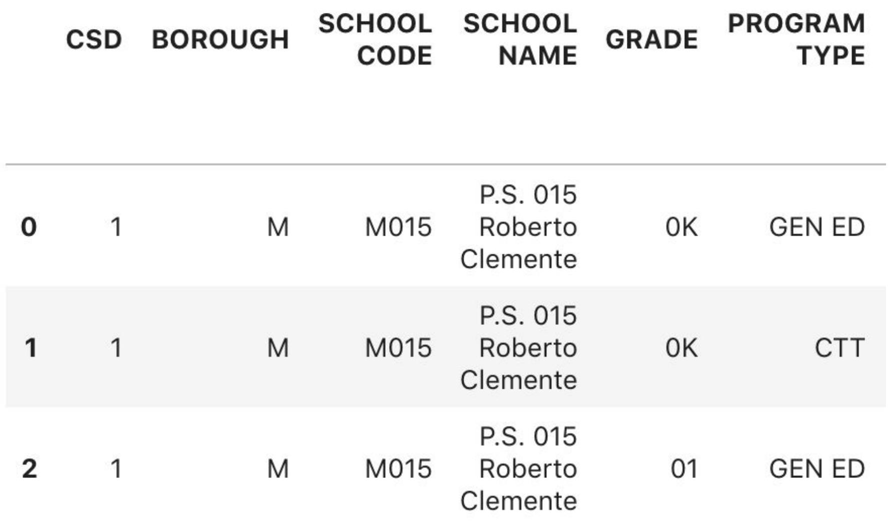
class_size

DBN	...
01M022	...
01M022	...
05M345	...
05M345	...

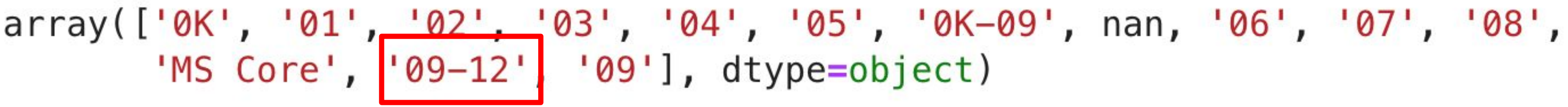
A single row in the **sat_results** data set may match multiple rows in the **class_size** data set. Problem!!!!!

We'll condense the **class_size**, **graduation**, and **demographics** data sets so that each DBN is unique

Condensing the **class_size** dataset

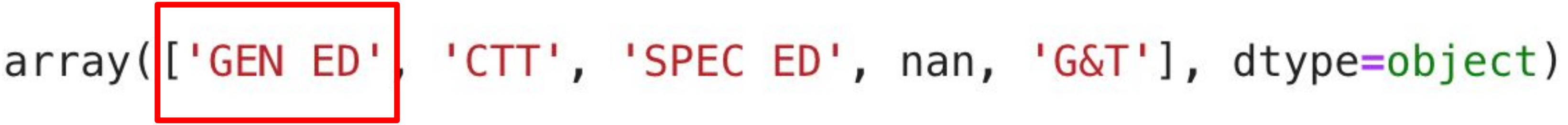


	CSD	BOROUGH	SCHOOL CODE	SCHOOL NAME	GRADE	PROGRAM TYPE
0	1	M	M015	P.S. 015 Roberto Clemente	0K	GEN ED
1	1	M	M015	P.S. 015 Roberto Clemente	0K	CTT
2	1	M	M015	P.S. 015 Roberto Clemente	01	GEN ED

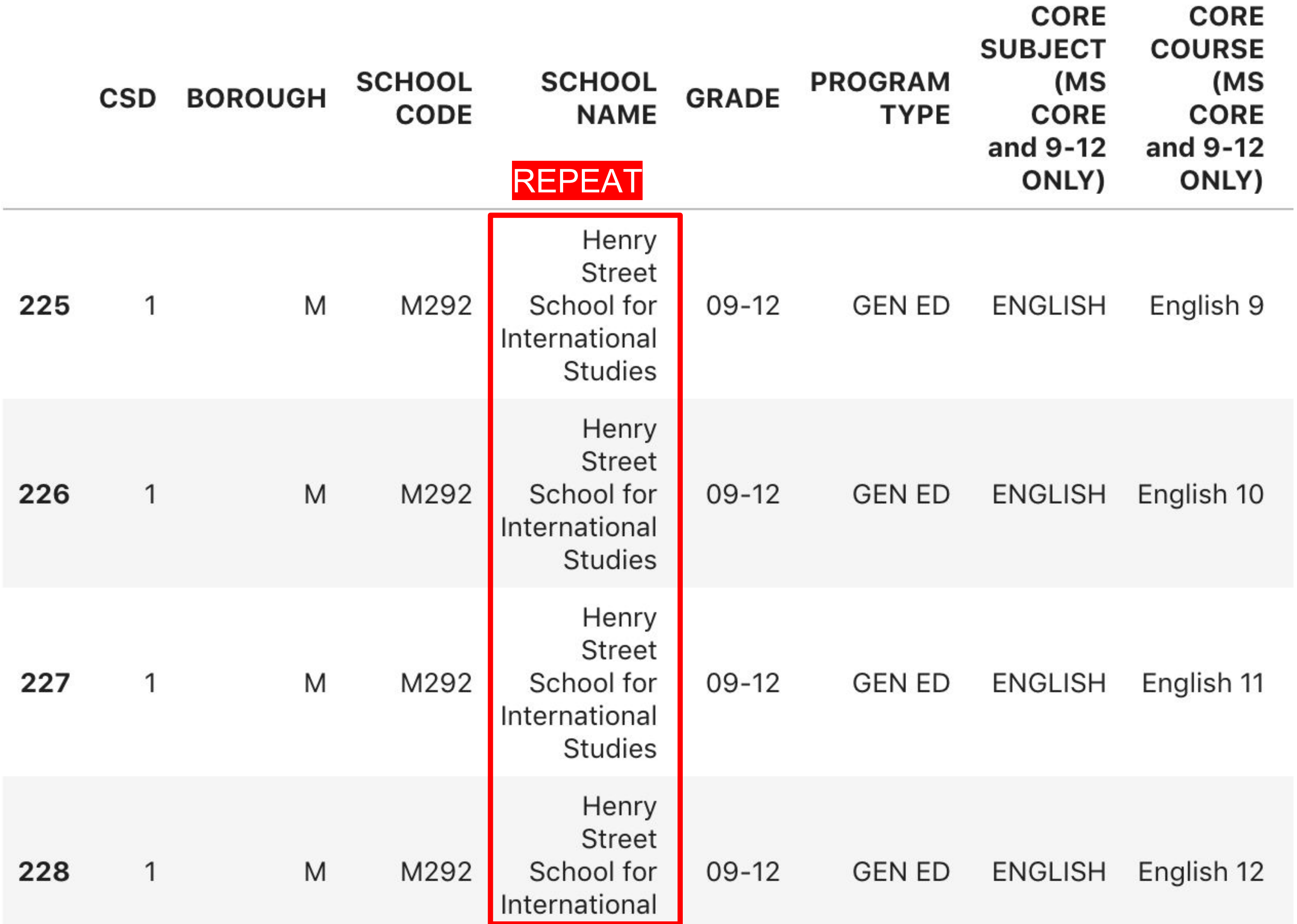


```
array(['0K', '01', '02', '03', '04', '05', '0K-09', nan, '06', '07', '08',
       'MS Core', '09-12', '09'], dtype=object)
```

High-School



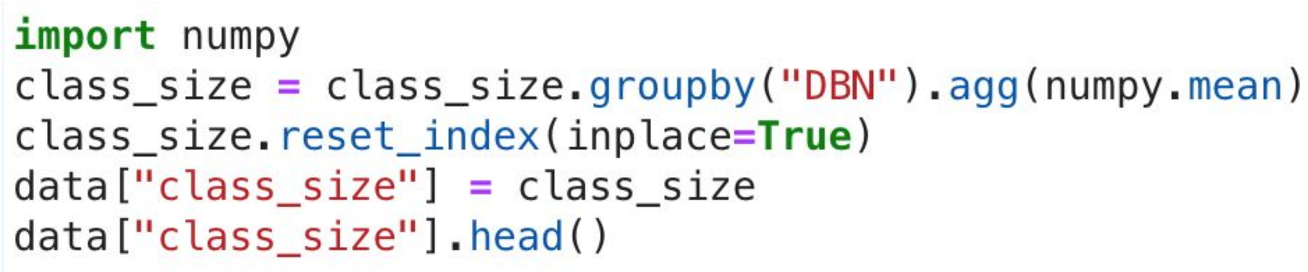
```
array(['GEN ED', 'CTT', 'SPEC ED', nan, 'G&T'], dtype=object)
```



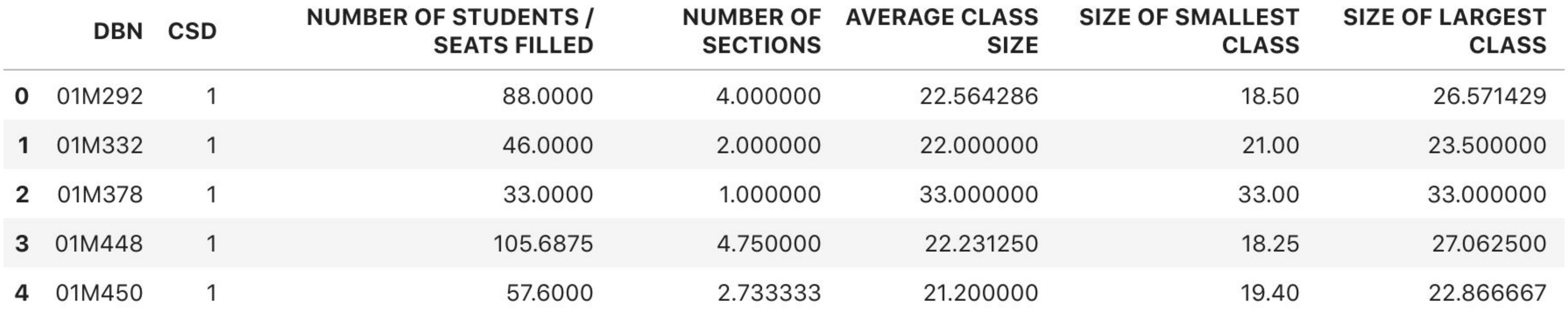
REPEAT

	CSD	BOROUGH	SCHOOL CODE	SCHOOL NAME	GRADE	PROGRAM TYPE	CORE SUBJECT (MS CORE and 9-12 ONLY)	CORE COURSE (MS CORE and 9-12 ONLY)
225	1	M	M292	Henry Street School for International Studies	09-12	GEN ED	ENGLISH	English 9
226	1	M	M292	Henry Street School for International Studies	09-12	GEN ED	ENGLISH	English 10
227	1	M	M292	Henry Street School for International Studies	09-12	GEN ED	ENGLISH	English 11
228	1	M	M292	Henry Street School for International	09-12	GEN ED	ENGLISH	English 12

Computing average class size



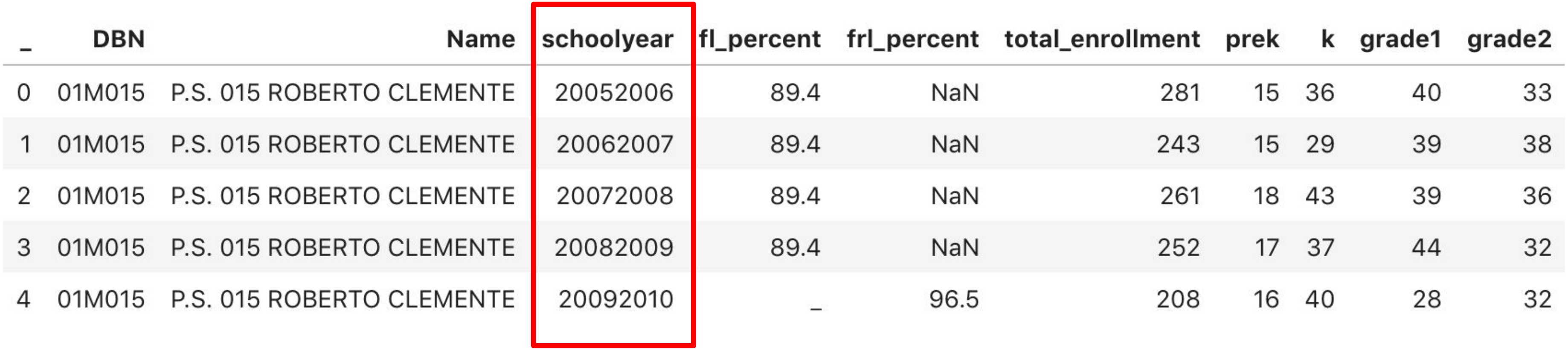
```
import numpy
class_size = class_size.groupby("DBN").agg(numpy.mean)
class_size.reset_index(inplace=True)
data["class_size"] = class_size
data["class_size"].head()
```



	DBN	CSD	NUMBER OF STUDENTS / SEATS FILLED	NUMBER OF SECTIONS	AVERAGE CLASS SIZE	SIZE OF SMALLEST CLASS	SIZE OF LARGEST CLASS
0	01M292	1	88.0000	4.000000	22.564286	18.50	26.571429
1	01M332	1	46.0000	2.000000	22.000000	21.00	23.500000
2	01M378	1	33.0000	1.000000	33.000000	33.00	33.000000
3	01M448	1	105.6875	4.750000	22.231250	18.25	27.062500
4	01M450	1	57.6000	2.733333	21.200000	19.40	22.866667

Condensing the Demographics Data set

20112012



_	DBN	Name	schoolyear	fl_percent	frl_percent	total_enrollment	prek	k	grade1	grade2
0	01M015	P.S. 015 ROBERTO CLEMENTE	20052006	89.4	NaN	281	15	36	40	33
1	01M015	P.S. 015 ROBERTO CLEMENTE	20062007	89.4	NaN	243	15	29	39	38
2	01M015	P.S. 015 ROBERTO CLEMENTE	20072008	89.4	NaN	261	18	43	39	36
3	01M015	P.S. 015 ROBERTO CLEMENTE	20082009	89.4	NaN	252	17	37	44	32
4	01M015	P.S. 015 ROBERTO CLEMENTE	20092010	_	96.5	208	16	40	28	32

Left, right, inner and outer joins

sat_results

DBN	sat_score
01	1800
03	2200
99	1600
101	2300

class_size

DBN	avg_class_size
01	20
03	30
55	50
101	30

Let's say we're merging the following two data sets.

Inner Merge

sat_results

DBN	sat_score
01	1800
03	2200
99	1600
101	2300

+

class_size

DBN	avg_class_size
01	20
03	30
55	50
101	30

=

combined

DBN	sat_score	avg_class_size
01	1800	20
03	2200	30
101	2300	30

Left Merge

sat_results

DBN	sat_score
01	1800
03	2200
99	1600
101	2300

+

class_size

DBN	avg_class_size
01	20
03	30
55	50
101	30

=

combined

DBN	sat_score	avg_class_size
01	1800	20
03	2200	30
99	1600	null
101	2300	30

Right Merge

sat_results

DBN	sat_score
01	1800
03	2200
99	1600
101	2300

+

class_size

DBN	avg_class_size
01	20
03	30
55	50
101	30

=

combined

DBN	sat_score	avg_class_size
01	1800	20
03	2200	30
55	null	50
101	2300	30

Outer Merge

sat_results

DBN	sat_score
01	1800
03	2200
99	1600
101	2300

+

class_size

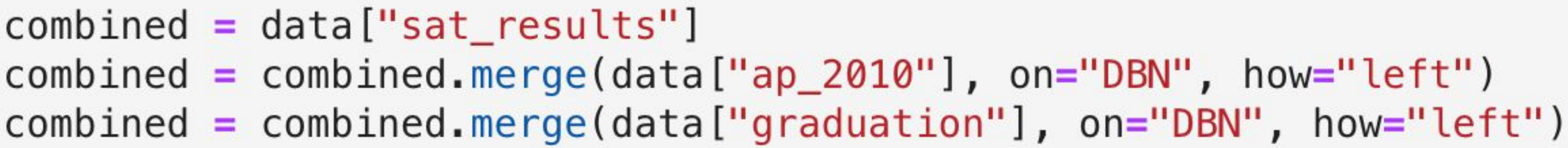
DBN	avg_class_size
01	20
03	30
55	50
101	30

=

combined

DBN	sat_score	avg_class_size
01	1800	20
03	2200	30
99	1600	null
55	null	50
101	2300	30

Performing Left Joins



```
combined = data["sat_results"]
combined = combined.merge(data["ap_2010"], on="DBN", how="left")
combined = combined.merge(data["graduation"], on="DBN", how="left")
```

Lesson 17 - Data Cleaning Walkthrough: combining the data.ipynb