

IMD0905 - Data Science I

Lesson #16 - Data Cleaning Walkthrough

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
```



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

Finding All of the Relevant Datasets

- Investigating the correlations between [SAT scores](#) and demographics might be an interesting angle to take.
- Correlate SAT scores with factors like race, gender, income, and more
 - ap_2010.csv - [Data on AP test results](#)
 - class_size.csv - Data on [class size](#)
 - demographics.csv - Data on [demographics](#)
 - graduation.csv - Data on [graduation outcomes](#)
 - hs_directory.csv - [A directory of high schools](#)
 - sat_results.csv - Data on [SAT scores](#)
 - survey_all.txt - Data on [surveys](#) from all schools
 - survey_d75.txt - Data on surveys from New York City district 75



The test consists of three sections, each of which has 800 possible points.
The combined score is out of 2,400 possible point



SAT Results

The most recent school level results for New York City on the SAT. Results are available at the school level for the graduating seniors of 2012. Records contain 2012

	DBN	SCHOOL NAME	Num of SAT Test	TSAT Critical Reading	SAT Math Avg. Sci	SAT Writing Avg. S	
1	01M292	HENRY STREET S	29	355	404	363	
2	01M448	UNIVERSITY NEIG	91	383	423	366	
3	01M450	EAST SIDE COMM	70	377	402	370	
4	01M458	FORSYTH SATELL	7	414	401	359	
5	01M509	MARTA VALLE HIG	44	390	433	384	
6	01M515	LOWER EAST SIDI	112	332	557	316	
7	01M539	NEW EXPLORATIC	159	522	574	525	
8	01M650	CASCADES HIGH	18	417	418	411	
9	01M696	BARD HIGH SCHO	130	624	604	628	
10	02M047	47 THE AMERICAN	16	395	400	387	
11	02M288	FOOD AND FINAN	62	409	393	392	
12	02M294	ESSEX STREET A	53	394	384	378	
13	02M296	HIGH SCHOOL OF	58	374	375	362	

Finding Background Information

- New York City is made up of five boroughs, which are essentially distinct regions.
- Only high school students take the SAT, so we'll want to focus on high schools.
- Each school in New York City has a unique code called a **DBN**, or district borough number.



Reading in Data (best practices)

```
import pandas as pd
data_files = [
    "ap_2010.csv",
    "class_size.csv",
    "demographics.csv",
    "graduation.csv",
    "hs_directory.csv",
    "sat_results.csv"
]
```

```
data = {}
for f in data_files:
    d = pd.read_csv("datasets/{0}".format(f))
    key_name = f.replace(".csv", "")
    data[key_name] = d
    print(d.shape)
```

Exploring the SAT Data

		Combine				
	DBN	SCHOOL NAME	Num of SAT Test Takers	SAT Critical Reading Avg. Score	SAT Math Avg. Score	SAT Writing Avg. Score
0	01M292	HENRY STREET SCHOOL FOR INTERNATIONAL STUDIES	29	355	404	363
1	01M448	UNIVERSITY NEIGHBORHOOD HIGH SCHOOL	91	383	423	366
2	01M450	EAST SIDE COMMUNITY SCHOOL	70	377	402	370
3	01M458	FORSYTH SATELLITE ACADEMY	7	414	401	359
4	01M509	MARTA VALLE HIGH SCHOOL	44	390	433	384

The DBN appears to be a unique ID for each school

We may eventually want to combine the three columns that contain SAT scores

Exploring Graduation Data

Repeat

	Demographic	DBN	School Name	Cohort	Total Cohort	Total Grads - n	Total Grads - % of cohort	Total Regents - n
0	Total Cohort	01M292	HENRY STREET SCHOOL FOR INTERNATIONAL	2003	5	s	s	s
1	Total Cohort	01M292	HENRY STREET SCHOOL FOR INTERNATIONAL	2004	55	37	67.3	17
2	Total Cohort	01M292	HENRY STREET SCHOOL FOR INTERNATIONAL	2005	64	43	67.2	27
3	Total Cohort	01M292	HENRY STREET SCHOOL FOR INTERNATIONAL	2006	78	43	55.1	36
4	Total Cohort	01M292	HENRY STREET SCHOOL FOR INTERNATIONAL	2006 Aug	78	44	56.4	37

NaN values

Exploring Data on Advanced Placement Exam



Advanced Placement Computer Science Principles (also called AP CSP) is an [AP Computer Science](#) course and examination offered by the [College Board](#) to [high school](#) students as an opportunity to earn college credit for a [college](#)-level [computer science](#) course

	DBN	SchoolName	AP Test Takers	Total Exams Taken	Number of Exams with scores 3 4 or 5
0	01M448	UNIVERSITY NEIGHBORHOOD H.S.	39	49	10
1	01M450	EAST SIDE COMMUNITY HS	19	21	s
2	01M515	LOWER EASTSIDE PREP	24	26	24
3	01M539	NEW EXPLORATIONS SCI,TECH,MATH	255	377	191
4	02M296	High School of Hospitality Management	s	s	s

Exploring Class Size

Padded CSD		School Code		DBN		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)
01		M015		01M015		0	1	M	M015	0K	GEN ED	-	-
19	+	M022	=	19M022		1	1	M	M015	0K	CTT	-	-
02		M016		02M016		2	1	M	M015	01	GEN ED	-	-
99		M025		99M025		3	1	M	M015	01	CTT	-	-
						4	1	M	M015	02	GEN ED	-	-

Exploring Directory of High Schools

```
re.findall("\(.+\)",
           data["hs_directory"]["Location 1"][0])
```

	dbn	school_name	borough	Location 1
0	17K548	Brooklyn School for Music & Theatre	Brooklyn	883 Classon Avenue\nBrooklyn, NY 11225\n(40.67029890700047, -73.96164787599963)
1	09X543	High School for Violin and Dance	Bronx	1110 Boston Road\nBronx, NY 10456\n(40.8276026690005, -73.90447525699966)
2	09X327	Comprehensive Model School Project M.S. 327	Bronx	1501 Jerome Avenue\nBronx, NY 10452\n(40.842414068000494, -73.91616158599965)
3	02M280	Manhattan Early College School for Advertising	Manhattan	411 Pearl Street\nNew York, NY 10038\n(40.71067947100045, -74.00080702099967)
4	28Q680	Queens Gateway to Health Sciences Secondary School	Queens	160 20 Goethals Avenue\nJamaica, NY 11432\n(40.718810094000446, -73.80650045499965)

Exploring Demographic Data

Annual school accounts of NYC public school student populations served by grade, special programs, **ethnicity**, **gender**.

	DBN	Name	schoolyear	black_num	hispanic_num	white_num	male_num	female_num
0	01M015	P.S. 015 ROBERTO CLEMENTE	20052006	74	189	5	158.0	123.0
1	01M015	P.S. 015 ROBERTO CLEMENTE	20062007	68	153	4	140.0	103.0
2	01M015	P.S. 015 ROBERTO CLEMENTE	20072008	77	157	7	143.0	118.0
3	01M015	P.S. 015 ROBERTO CLEMENTE	20082009	75	149	7	149.0	103.0
4	01M015	P.S. 015 ROBERTO CLEMENTE	20092010	67	118	6	124.0	84.0

2011 NYC School Survey Data Dictionary

This data dictionary can be used with the school-level data files from the 2011 NYC School Survey. School-level data is available in one file for all community schools (file name: masterfile11_gened_final) and one file for all District 75 schools (file name: masterfile11_D75_final). These files display one line of information for each school, by DBN, that includes the response rate for each school, the number of surveys submitted, the size of the eligible survey population at each school, question scores, the percentage of responses selected, and the count of responses selected. These fields are detailed below.

Field Name	Field Description
dbn	School identification code (district borough number)
sch_type	School type (Elementary, Middle, High, etc)
location	School name
enrollment	Enrollment size
borough	Borough
principal	Principal name
studentsurvey	Only students in grades 6-12 partipate in the student survey. This field indicates whether or not this school serves any students in grades 6-12.
rr_s	Student Response Rate
rr_t	Teacher Response Rate
rr_p	Parent Response Rate
N_s	Number of student respondents
N_t	Number of teacher respondents
N_p	Number of parent respondents
nr_s	Number of eligible students
nr_t	Number of eligible teachers
nr_p	Number of eligible parents
saf_p_10	Safety and Respect score based on parent responses
com_p_10	Communication score based on parent responses
eng_p_10	Engagement score based on parent responses
aca_p_10	Academic expectations score based on parent responses
saf_t_10	Safety and Respect score based on teacher responses
com_t_10	Communication score based on teacher responses
eng_t_10	Engagement score based on teacher responses
aca_t_10	Academic expectations score based on teacher responses
saf_s_10	Safety and Respect score based on student responses
com_s_10	Communication score based on student responses
eng_s_10	Engagement score based on student responses
aca_s_10	Academic expectations score based on student responses
saf_tot_10	Safety and Respect total score
com_tot_10	Communication total score
eng_tot_10	Engagement total score
aca_tot_10	Academic Expectations total score

Reading in Survey Data

Reading in Survey Data

```
all_survey = pd.read_csv("datasets/survey_all.txt",  
                           delimiter="\t",  
                           encoding='windows-1252')  
d75_survey = pd.read_csv("datasets/survey_d75.txt",  
                           delimiter="\t",  
                           encoding='windows-1252')  
survey = pd.concat([all_survey, d75_survey],  
                    axis=0, sort=True)
```


Concatenation

	letter	number
0	a	1
1	b	2

	letter	number	other
0	c		3
1	d		4

	letter	number	number	other
0	a	1.0		NaN
1	b	2.0		NaN
2	c	NaN		3.0
3	d	NaN		4.0

Lesson 16 - Data Cleaning Walkthrough.ipynb

