Data Science & ML Course

Lesson #13 Data Cleaning Walkthrough

Ivanovitch Silva

October, 2018

Agenda

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

Update from repository

```
git clone https://github.com/ivanovitchm/datascience2machinelearning.git
```

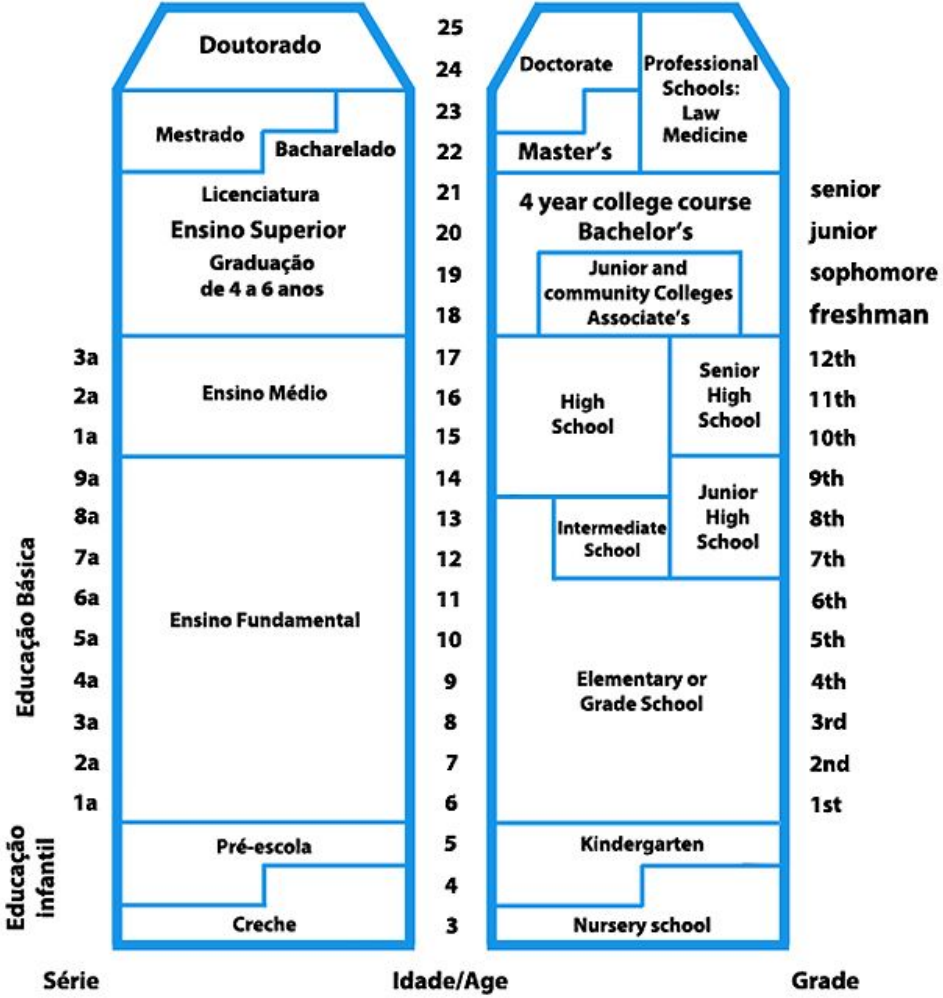
Or ....

```
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


Doutorado
Mestrado
Bacharelado
Licenciatura
Ensino Superior
Graduação
de 4 a 6 anos
Ensino Médio
Ensino Fundamental
Pré-escola
Creche
Educação Básica
Educação Infantil
3a
2a
1a
9a
8a
7a
6a
5a
4a
3a
2a
1a
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
Doctorate
Professional Schools: Law Medicine
Master's
4 year college course
Bachelor's
Junior and community Colleges
Associate's
High School
Senior High School
Junior High School
Intermediate School
Elementary or Grade School
Kindergarten
Nursery school
senior
junior
sophomore
freshman
12th
11th
10th
9th
8th
7th
6th
5th
4th
3rd
2nd
1st
Série
Idade/Age
Grade

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

NYC OpenData

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 T	SAT Critical Readi	SAT Math Avg. Sc	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
01	+	M015	=	01M015
19		M022		19M022
02		M016		02M016
99		M025		99M025

	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)
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	-	-
3	1	M	M015	P.S. 015 Roberto Clemente	01	CTT	-	-
4	1	M	M015	P.S. 015 Roberto Clemente	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

Reading in Survey Data

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

```
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	numberother
0	c	3
1	d	4

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

Lesson 13 - Data Cleaning Section 1

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

	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

Condensing the **class_size** dataset

```
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)
```

	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)
				REPEAT				
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 13 - Data Cleaning Section 1