**IMD0905 - Data Science I Lesson #17 - Data Cleaning Walkthrough: combining the data**

**Ivanovitch Silva November, 2018**

2 Agenda

● Case study: NYC open data (education)

● Data cleaning walkthrough

● Combining data

**● Groupby**

**● Merge (inner, outer, right, left)**

3 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

4

5 Combining the Data

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

High-School

6

Condensing the class\_size dataset

REPEAT

7

8 Computing average class size

9 Condensing the Demographics Data set

20112012

10 Left, right, inner and outer joins

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

11 Inner Merge

12 Left Merge

13 Right Merge

14 Outer Merge

15 Performing Left Joins

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

16