A Big Data Set -- DVC's Scheduling History, Reading

DVC's Admissions and Records office has an electronic database of all the sections of all the courses offered since the year 2000. It has over 70,000 entries, and grows every semester. The database is used to answer queries like these:

- How many MATH courses offered in a specified time period?
- Who's taught COMSC-210 in the last 5 years?
- When was ART-107 last taught?
- How many times has Prof. Burns taught COMSC-210?
- What's the room schedule for ATC-115 this semester?

...and so on. It's queries such as these that led to the development of web pages like these (active links):

- **The DVC room schedule** **(http://web.dvc.edu/roomschedule)**
- **DVC course offering history** **(http://web.dvc.edu/handytools/coursehistory)**
- **DVC's searchable fall schedule** **(http://web.dvc.edu/onlineSchedule/fall/)**

Programmers had to write the PHP script for these web applications, and while it was a lot less challenging than what Google has to do, it's still big enough data that simple for-loops were not the solution.

A Downloadable Database

There are many formats for storing and accessing databases, like SQL, JSON, and AJAX in web applications like the ones listed above. Those require user authentication to access, and we don't have that, so we'll be using an old, basic format for our dataset -- the "flat file". We'll be working with it for some of our studies going forward.

The flat file has one record per line (over 70,000 of them) with tab-separated values. It's suitable for opening in Excel, and for our C++ programs to open and read. Each record has these values:

1. The term, like Fall 2016
2. The section number, 8375
3. The course, like COMSC-210
4. The instructor, like Burns
5. the day(s), time(s), and room(s), like W 7:00-8:30pm L-142

Each combination of term and section number *uniquely identifies* a course offering. For example, our course is offered in Fall 2016 as section number 8375. There is no other section on campus in Fall 2016 that has the section number 8375 except for ours. The data pair {Fall 2016, 8375} uniquely identifies our course.

If the database contains *another* entry for {Fall 2016, 8375}, that's a problem. Just how big of a problem depends -- is it a double-entry for the same class, or are two different classes sharing the same section number. The first is not so bad -- it just messes up counting. The latter *is* a problem, because somebody in the Scheduling office assigned the same unique code more than once! We would not be mentioning this at all unless ***it really happened*** in our data set. It did, and it's going to mess us up, and yes, we'll have to deal with it.

Here's a link for you to download the flat file:

https://drive.google.com/file/d/0B14YTZL55whCSGE0VjcyajZxSmc/view?usp=sharing
(https://drive.google.com/file/d/0B14YTZL55whCSGE0VjcyajZxSmc/view?usp=sharing)

To use it with the code that will be supplied later in this module, name the file **dvc-schedule.txt**. It's *big* -- over 4MB -- so whatever you do, do *not* submit it with your lab assignments that use it! The professor already has a copy.