Data Management

OR

GETTING PUNCHED IN THE FACE BY SQL AND PANDAS

BY

NICHOLAS ROSS, PHD

PROFESSOR OF DATA SCIENCE
UNIVERSITY OF CHICAGO

Cover Art By: Megan Carlsen

Version: 2023-08-21 02:33:21

Contents

Relational Databases

Pandas

Appendix

Introduction and Errata

Introduction & Errata

Thank you for your interest in learning Data Management via SQL and Python. The material in these lecture notes covers the vocational aspects of learning these tools in a systematic and consistent manner.

Thank you for your interest in learning SQL! At the end of this course, you will be familiar with SQL and comfortable using it in a variety of real-world situations. While we directly use PostgreSQL in the notes, nearly all of the syntax presented is compatible with alternative SQL implementations. In those cases where there are compatibility issues, we try to call them out and address them.

Neither Pandas or SQL is *difficult*, but like learning any other language it requires time and practice. The purpose of these notes and problems are not to be a readable book, but instead a set of notes which are both a reference and guide. The majority of the learning that occurs is not within the text, but within the problem sets and their solutions.

Each module within this text is designed to be a (roughly) one hour lecture. At different levels and experience it is possible for some to run short and others to run long. This course has been taught in as little as 7 weeks to masters level students and taken as long as an entire semester at the undergraduate level. Dependencies between different modules are fairly obvious and quite a bit of the more technical material can be treated as an extension (specifically Modules 4, 10, 12 and 13 and easily skip-able).

The course material is designed to be amenable to a few different environments. It has been taught at the undergraduate level, undergraduate level as well as a free-standing executive certificate. While the primary learning objectives are the same in each of these environments there are (obviously) different expectations around this course at each of these levels.

Undergraduate

At the undergraduate level this has been taught as a semester long course which was paired with an applied machine learning section. When teaching at the undergraduate level students were provided with access to a cloud-based relational database, with limited permissions, that contained the databases used in this course.

The coursework in these notes was paired with a group project and research paper write-up. Students were required to load their data into a database and then created a set of jupyter notebooks and Python libraries to access the data and execute on their own research plan and agenda. Groups of students then did final presentations and long-form write-ups.

Graduate

At the graduate level this has been taught as both a once-a-week, 7-week long introduction to SQL as well as a twice a week quarter length course covering both SQL and Pandas. In the former situation only the core modules regarding SQL syntax were covered while the later included all information presented in these notes.

When teaching at the graduate level, the raw data was provided to the students with the expectations that they would be able to load it into their own SQL (local) instances and work from there. Homework problems are lightly graded and quizzes are given each week in order to assess current knowledge retention.

Executive Certificate

This course material was also taught as an Executive Certificate in a once-a-week, 3 hours per week, 7 week long format. During each week, save the first and last, students were given a short self-assessed quiz. An hour long lecture was then done followed by working on problems from the assignments with the goal to finish the "first five" for the sections covered.

When teaching in this format, only the core SQL syntax was covered and students were provided access and credentials to a cloud-based server which contained the data for the course.

Errata and WIPs

This document is a work in progress and contains quite a few known issues. This preface contains known issues and places where improvements are required.

Overall

1. Remove every reference to Module and change to chapter.
2. Fix the interview notes.
3. DDL
4. Categorical Data
5. Add a section to the start of the Pandas regarding “state” and how, unlike SQL, there is a current “state” of a DataFrame. E.g. row numbers matter a lot.
6. Vector DB Discussion: `https://www.ethanrosenthal.com/2023/04/10/nn-vs-ann/`

SQL

1. Rewrite `cast` section. Currently confusing.
2. Queries with out a from clause and discussion of `select 1` put at start mod 2.
3. Add to the start of the book `https://www.amazingcto.com/postgres-for-everything/`
4. Simple correlated subquery example. Current example is far too complex.
5. Rewrite NoSQL section adding information about a vector and graph databases:
   - Look at: `https://www.theregister.com/2023/03/08/great_graph_debate_wednesday/`
6. Add a resources section to the introduction which contains information on different books to consider.
   - `https://postgrespro.com/community/books/internals`
7. Online PostgresSQL explainer: `https://explain.dalibo.com/`
8. Add to performance consideration section. Discussion on the extreme case of super wide tables and how it effects performance: `https://www.cybertec-postgresql.com/en/column-order-in-postgr`
9. Add more examples of aggregation with case statements.
10. Add more formality to the discussion on what is returned and how it can be used in table vs. scalar.
11. `https://carlineng.com/?postid=sql-critique#blog`
12. Use MTA data and add hour to create timestamp in the date/time section. Add more date/time examples to date and time section, including intervals.
13. Language around analytic functions and LTV incorrect and needs to be fixed.
14. Check GSN and Zynga Dates. Add photos from Zynga as well as their MySQL solution.

Pandas

1. Add `rank` aggregation.
2. In section #1 the way that value_counts and column selection occurs is awkward. Maybe change the ordering to move value counts to after column selection.

3. Add reindex to index discussion in pandas. Overall discussion of muulti-index and reindex needs to be updated.
4. Move `loc` not accepting NaN to first module
5. Add example for `duplicated`
6. Date stuff needs to be redone, both as index (additional section) and as regular/type discussion.
7. MTADF comes out of nowhere in the module 2. When is it first introduced?
8. Other resources to look into:
    - Effective Pandas `https://store.metasnake.com/effective-pandas-book`
    - `https://betterprogramming.pub/pandas-illustrated-the-definitive-visual-guide-`
    - copy warning: `https://stackoverflow.com/questions/32573452/settingwithcopywarning`
9. day_name vs. weekday_name
10. `https://www.practicaldatascience.org/html/views_and_copies_in_pandas.html`
11. Add a short section on creating simple dataframes with dictionaries or lists
12. Move applymap map and apply to the next module and then take all the slice stuff and move it to the module with the groupby object.
13. Cuts / bins
14. iterrows
15. Any / All
16. Pivot
17. Add a section on analyzing the transaction data using Window Functions. Specifically, mimic the functions in the advanced joins.
18. HW #8B needs work, add more.
19. Loading and Saving Data
20. Time Series stuff
21. There are some questions in the HW about correlation, just go over this.