In this project, you will work in groups, exploring one or more datasets, and creating a compelling and original product that represents a synthesis of techniques you learned in this course. We have provided several options to choose from in designing your final project, but we welcome original ideas of your own – at our discretion, a project demonstrating multiple skills covered in this course (or beyond) may be approved.

This project is an opportunity to demonstrate your understanding of the course material and create something interesting that you can include in your project portfolio when applying to jobs. Start brainstorming and working on the project early! The best projects are the product of consistent and regular work – not the ones which are done at the last minute.

The final project report is due on Wednesday, April 30th. Lightning talks will be on Monday, April 28th and Wednesday, April 30th.

1 Groups

For this project, you must work in groups of 2 - 4. Your groups may include members from both sections of the course. **No submissions will be accepted from individuals or from larger groups.** As soon as possible, look through the various options for this project, decide on the type of project you are interested in, and start looking for potential group mates. Posting on Ed Discussion is a great way to find a like-minded group!

While this is a group project, you will also be graded on your individual contributions to the project. Ensure that every group member has a voice in decision making and an opportunity to contribute in equal measure to the final product! If you experience any difficult working on the project or with your group, *let the instructors know as soon as possible.* You are responsible for your individual contributions to the project.

2 Project Requirements

The project is worth 110 points. You will be graded on three project components (detailed below), a group final report (100 points) and presentation (10 points), and your individual contribution to the project.

Each project component has options to provide you with some flexibility in designing a project that matches your interests and that makes sense for the data you choose to work with. Read each description carefully to ensure you understand the requirements for each component.

2.1 Component 1: Dataset(s)

To fulfill the requirements for this component, choose a data set meeting **one or more** of the descriptions below. In your report, state which option(s) you selected and how your data set meets the requirement. Provide links to your data sources and describe any limitations on use, i.e., if the data set is in the public domain or is restricted to non-commercial use, etc.

1. **A large dataset**. "Large" is somewhat subjective, and may mean more than one thing. A data set with 100,000 records is definitely large. A data set with 10,000 rows may be "large" if it has many columns or is composed of many tables. A data set may be large if it presents some difficulty in loading, cleaning, or querying simply due to its size.

   If you go this route, please keep it somewhat reasonable – our RDBMS is a shared resource! As a rule of thumb, keep it below 4 GB unless you want to work with it in a PostgreSQL instance on your own machines.

   This option is likely to be a good fit with the **Data science and analysis** or **Advanced technology** options for Component 2, but of course may be used with other options.

2. A data set that is composed of, or may be decomposed into, **3 or more interrelated tables**. A join of these tables must result in 1,000 or more tuples. At least 2 tables must represent entities with significant structure – i.e., entities with multiple attributes of interest as opposed to "lookup tables":

- An example of an entity with significant structure might be historical meteorological statistics for agricultural districts in the state of Colorado recording rain accumulation, snow accumulation, average wind speeds, and so forth by year and month over multiple years.
- An example of a lookup table that would not be considered an entity with significant structure is a table mapping state abbreviations (e.g., "CO") to state names.

This option is likely to be a good fit with the **Software development** option for Component 2, but of course may be used with other options.

3. **A "mash up" data set**. This is a dataset that is composed of 2 or more tables that are not obviously interrelated, but which can be combined by joining along some shared attribute either in the tables or that can be derived from attributes in the tables. The goal with this data set is to discover interesting analyses that offer some potential new insight into a topic. For example, you might combine a data set on alcohol sales by geographic location with a data set on climatic conditions in the same location to determine if factors such as temperature, inclement weather, etc. are driving factors for alcohol consumption.

   This option is likely to be a good fit with the **Data science and analysis** option for Component 2, but of course may be used with other options.

This section is worth **10 points**:

- (3 points) Include a specific citation of the dataset, including the date it was accessed
- (2 points) The selected option is clearly indicated
- (5 points) Includes a description of how the dataset fulfills the requirements

2.2 Component 2: Application

To fulfill the requirements for this component, apply concepts from this course to **one or more** of the following. In your report, state which option(s) you selected and a brief overview of your work and how it met the requirement for your option(s).

- **Data science and analysis**. Your project should dissect and analyze your data set in multiple ways (see **Outputs** section below). Your analysis must do more than scratch the surface with e.g. simple correlations – your goal is to extract insights from the data that are not trivially obvious from an inspection of the data. While your work must include significant applications of SQL for data cleaning, aggregation, or other transformations, you may also utilize tools such as pandas, sciKit-learn, TensorFlow, etc. It is acceptable for your work to be somewhat inconclusive if the methodology is **thorough** and correct.
- **Software development**. You will need to create application software (e.g., a web application, CLI application, or similar) that connects to the database and includes functionality to query, add, modify, and remove data. Your software must follow best practices with regards to database programming (e.g., prevent SQL injection attacks). Your application should connect to the course database unless you are combining this with the next option, but otherwise may use any language or frameworks you wish.
- **Advanced technology**. Use a new, unusual, or cutting-edge database and provide a compelling demonstration of the advantages or trade-offs of using that database versus a classic RDBMS like Postgres, MySQL, SQLite, SQL Server, or other relational technologies we talked about in class. Examples of an "advanced database" include: graph databses like Kuzu, Neo4J, Virtuoso, Blazegraph, XTDB; array databases like SciDB; distributed databases like CockroachDB, FoundationDB or TiDB; vector databases like Chroma, Pinecone, Qdrant or pgvector; or streaming databases like Apache Druid, Flink, Materialize. You are not limited to this list, but you should choose something different enough from out-of-the-box Postgres.

This section is worth **10 points**:

- (2 points) The selected option is clearly indicated

- (8 points) Includes an overview of the project and how it meets the requirements for the indicated option

2.3 Component 3: Outputs

To fulfill the requirements for this component, include at least 4 of these deliverables. Each of the deliverables must be different. In your report, include a section listing which deliverables you have included (in addition to any sections of your report implementing the deliverable).

- **Data loading and cleaning.** Only include this deliverable if your data needed more than trivial cleaning. Devote a section of your report to a concise and complete description of the issues with your data set and the steps you took to correct them. Include a SQL script showing your data cleaning queries as an appendix to your report (properly formatted, not a screenshot!).
- **Database schema design.** Provide a clear and complete schema diagram of your database design, showing all tables, primary and foreign keys, and relationships. Include a section in your report explaining your design choices, such as normalization decisions and any trade-offs made between normalization and performance. Your documentation must include information about constraints implemented to maintain data integrity. If your database design evolved during the project, briefly explain the significant changes and why they were necessary.
- **Interesting or complex queries.** Your query or set of queries should demonstrate multiple SQL techniques beyond simple SELECT...WHERE...: e.g., subqueries, CTEs, window functions, and grouping and aggregation. These queries must not be created simply to demonstrate techniques, but must have a recognizable function in your analysis, software, or evaluation. Include a section in your report for each query, include the SQL (properly formatted, not a screenshot!), and describe the design and purpose of the query.
- **Performance tuning.** Choose 2-3 queries that perform poorly and add indexes or otherwise modify the queries or database to improve performance. Your changes must result in a demonstrable speed up of your queries. Include a section in your report showing relevant "before and after" snippets from the query plan and timings or other evidence of the speed up. (If you chose the **Advanced technology** option, you may provide alternative evidence appropriate to your database.)
- **Performance comparison.** If using the **Advanced technology** option, perform a comparison of performance between your chosen database and Postgres. Provide at least 5 carefully designed benchmarks (data loading, query execution, or other tasks) which you execute on both systems. Include a section in the report in which you compare the results using statistical analysis techniques, including at least one visualization. Comment on why the results are what they are: when is Postgres faster/more efficient and when is the other data system faster/more efficient? Make sure to include an explanation of why this is the case.
- **Statistical and machine learning analyses.** Provide 2-3 careful analyses of your data that reveal something of interest that is not obvious from a casual examination of the data. Simple 2-variable correlations are not interesting analyses, unless accompanied by significant work to transform the data and derive the variables of interest. At least one analysis must be performed in SQL or enabled by SQL (e.g., data transformations). Include properly labeled and formatted visualizations where appropriate.
- **Data visualization.** These are **not** visualizations of your **Statistical and machine learning analyses**, if you are doing them. These are other, creative ways of visualizing your data or summaries of your data, such as Sankey diagrams, dendograms, sunburst diagrams, etc. Include the visualization (properly labeled and formatted) and an discussion of the significant features shown in the visualization.
- **Software.** If you chose the **Software development** application, provide your source code either as a link (in your report) to a public repository or as additional files in your submission. Your report should describe the functionality of your application (at least 3 operations that can be performed using the software) and how that functionality is implemented and interacts with the database. Provide screenshots or transcripts of the application in action to illustrate each function (but keep it concise - ellide long outputs or transcripts).

- **Evaluation of technology.** If you chose the **Advanced technology** application, provide a summary of the key features of the database technology you selected and how it differs from an RDBMS; provide a compelling demonstration of the advantages or trade-offs of using the new technology. This could be descriptions of queries that are made possible or much easier with the new technology, deployment considerations that are possible with the new technology, or other interesting factors which are not related to performance. For performance-related comparsisons, please see the task above.
- **Your choice (with instructor approval).** If you would like to do something non-trivial and interesting with a database that you do not see above, make a detailed post on EdStem with a description of what output you want to create. The instructors will get back to you with an approval. *If you want to choose this option,*

This section is worth **80 points**:

- (20 points) First indicated output is clearly indicated and described in the report using the requirements listed above
- (20 points) Second indicated output is clearly indicated and described in the report using the requirements listed above
- (20 points) Third indicated output is clearly indicated and described in the report using the requirements listed above
- (20 points) Fourth indicated output is clearly indicated and described in the report using the requirements listed above

3 Final report

In addition to the elements discussed for the components above, your report should include an introduction and overview, a conclusion, and references (if relevant). Your report should **prominently** list the names of all group members at the start of the report.

Each group member should submit to Canvas a short text document or PDF describing their contributions to the overall project in some detail – code you wrote, analyses you performed, sections you wrote, etc. Include a list of all group members with *what percentage of the overall project effort they contributed to* (including yourself!). For example, if there are 3 group members who made equal contributions to the project, then this document should have 3 names each with 33% after them. This submission will be made to the Final Project Contribution assignment on Canvas.

One group member should submit your project report and any auxiliary files (such as source code) in a .zip archive to Canvas under the Final Project assignment.

3.1 Checklist

- ☐ List of group members
- ☐ Introduction and overview
- ☐ Component 1
  - ☐ Which option(s) you chose
  - ☐ How your data set meets the requirement
  - ☐ Links to data source(s)
  - ☐ Restrictions on use
- ☐ Component 2
  - ☐ Which option(s) you chose
  - ☐ Overview of work and how it meets the requirements
- ☐ Component 3
  - ☐ A list of all your deliverables
  - ☐ Required sections, figures, and/or supplemental material for each deliverable (details above)
- ☐ Conclusion

4 Presentation

Your group will give a **2 minute "lightning talk"** on one of the final days of class (Friday, April 25th and Monday, April 28th). Signup sheets for each presentation day will be posted soon. The presentation is worth **10 points**. Your talk needs to do the following:

- (2 points) Introduce the dataset(s) and/or technologies you are using ($< 20\%$ of your presentation)
- (8 points) Show off the most interesting outputs of your project. Explain how to interpret them and discuss their significance. ($\geq 80\%$ of your presentation)