MGTA 495 - Final Project Description

For our final project, we want you to be able to use the tools that you've been exposed to to perform some analysis on real-world data. We will post a list of datasets that you may choose from at the end of Week 6.

This project is to be completed in groups of 3-4 people. If you need help finding a group, check out the "Search for Teammates!" feature of Piazza. It is worth a total of 40 points, and will consist of three main parts: working with ETL tools, working with Databases, and performing analytics with the tools presented in the final third of the course. The requirements of each section are listed at the end of this document. The grading breakdown is as follows:

Total: 40 points

- ETL: 10 points
- DB: 10 points
- Analytics: 10 points
- Report + Presentation (problem setup, dataset description/exploration, analysis, etc.): 5 points
- Evaluate team members: 5 points

TritonEd Submission Requirements

Each team member must submit a document evaluating your fellow team members and their contributions. One person from your team will submit a zip file containing:

1. Your code (Jupyter notebooks, ETL/SQL scripts, etc.)
2. A short written report that summarizes your methods and results
3. The presentation that you present in our final two weeks

This will be due on **March 19th, 2019 at 11:59pm**. Additionally, you will be presenting your work during our normal class time in Week 10 and Finals week, in a **12 min (+3 min for questions)** presentation. For overall participation in this course (10 points), you'll also be required to give feedback on other presentations. Towards the end of the course, we'll have sign-up sheets to schedule presentations and which teams you will provide feedback for. You are required to sign up to evaluate totally at least two other teams from the class; these two teams must come from different presentation days to ensure that every group has an audience.

Below are the different section requirements for the project.

ETL Requirements

- Develop a Pentaho Kettle script to load the data on your local Postgres database. This will include any steps you take to clean up/pre-process the data like we did in A2 to make the Databases/Analytics task simpler.

Database Requirements

1. For the World Bank Data: Choose any two countries that have data in at least 3 different domains (e.g., poverty and equity), each with at least 5 indicators, for 4 consecutive years. For each country, construct a set of basis statistics (means, variances, etc.) over the available data. Compute indicator correlations across domains. Compare the performances of the two countries year by year. Every step of the above process should be done by SQL queries.

2. For the Housing Data: Pick any two regions (i.e., data with different region IDs). Compute the histogram, mean and variance of the total house price for every zipcode in each region. Create two groups, the houses that are above the mean and those that are at or below the mean. For each group, find the indicator variables whose average value deviates most between the two groups. Every step of the above process should be done by SQL queries.
3. For the Electronic Commerce Data: Answer the following questions (formulated in SQL):
    a. Which category of products receive most 5 ratings (survey_score)?
    b. What are the top 5 best selling products and what are their categories? What's the average rating of each of these products?
    c. For which products are the buyers and sellers mostly from the same state? (you have to specify what "mostly" means)
    d. Give a regional break-up of total sales of heavier (weighing at least 1 kilogram) products.

Analytics Requirements

- Use Spark MLLib or AWS SageMaker for analysis
- Include one clustering or classification task
- Describe analysis task
    - To include: type of task (e.g., classification), why is it interesting
- Describe data
    - To include: data quality issues, characteristics of the dataset (summary statistics, correlation, outliers, etc.)
- Describe data preparation process
    - To include: features used, train and test datasets
- Describe model
    - To include: input, setup, and output
- Describe analysis results and insights gained
    - To include: discussion of results, actionable insights

MGTA 495 - Datasets

You may choose from the datasets uploaded to TritonEd.

- WorldBank dataset- this is the same dataset we have used earlier, but
- E-Commerce dataset
- Housing dataset