Final Project Assignment

BDAT1000: Data Manipulation Techniques
Winter 2021
Due: Friday, April 23, 8:00 AM (NOTE CHANGE IN TIME FROM ORIGINAL SYLLABUS)

Project Goal: The goal of this project is to test your ability to: 1. Formulate an engaging research question, 2. Identify and source appropriate datasets to answer the question, 3. Clean, shape and model the data to meet your analytical needs, and then 4. Design a dashboard to generate insights that directly relate to your initial question.

Project Overview: This is a self-directed project, intended to allow you to explore datasets that speak to your individual interests. You will choose a set of interrelated datasets on a topic of interest to you, formulate some research questions that you wish to answer using the datasets, and develop an interactive Power BI dashboard that provides useful insights directly relevant to the question(s).

As indicated, the choice of topic is left to your discretion. Depending on your interests, you could choose projects related to:

- Environmental issues
- Crime and justice
- Public health
- Economic development/poverty
- Education
- Demographic change
- Local/regional/national politics
- Sports
- Etc.

Research questions should explicitly look at the relationship between two or more different types of data. Some general examples of the type of question we are looking for:

- Do a region's political preferences affect/reflect particular features of the region, e.g. economic or demographic characteristics?
- How does economic development relate to environmental impact/health outcomes?
- How does education level relate to income?
- How do specific demographic characteristics (age/race/gender/socioeconomic status/etc.) relate to educational outcomes/crime levels/health outcomes/etc.?
- How have particular events or developments (e.g. COVID, Brexit, anti-racism protests in the US, 9/11 attacks and the U.S. "War on Terror", electoral results, etc.) been reflected in changing data patterns in other aspects of life?
- How do regional characteristics and/or specific team characteristics relate to outcomes in sporting events?

These are generalized examples. Your question(s) should be specific, isolating a particular issue and exploring it from various angles to produce insights.

Specifications and requirements: The goal of this assignment is to demonstrate your ability to bring data from different sources together in a single data model that produces useful analytics/insights. As such, the project should fulfill the following requirements in terms of the methods and materials used:

1. **Multiple data sources** – Your project must incorporate data from at least two different original data sources/tables.
   NOTE: You may use data from almost any public data source (governmental data, open data repositories, websites, etc.). Exceptions are data from Kaggle.com and datasets available in R.
2. **Minimum fact table size** – The data model should include at least one fact table that, after cleaning and reshaping, contains at least 1000 rows of data.
3. **Multi-table dimensional data model in Power BI** – Your data model must be in Power BI and must be a fully-functional data model with both fact table(s) and dimensional table(s) that are used appropriately in the analysis. **NOTE: A single table is NOT a data model.**
4. **Dashboard interactivity** – Your visualizations must be presented as a dashboard and must be dynamic – i.e. users must be able to interact with them by using slicers, filters, drill-downs, drill-throughs, etc. to generate different levels of insight.

   EXAMPLE #1: A pie chart that shows a country's population distribution by age group but which can also show the same distribution at the state/province or smaller regional level.

   EXAMPLE #2: A bar chart of average income by educational level with a drill-down option that allows users to click on one of the bars and see the average income for that educational level broken down by broad area of study.

Once you have completed your dashboard and identified your key insights, you must design and record a 12-15 minute presentation on Screencast-O-Matic (https://screencast-o-matic.com) that fulfills the requirements outlined below.

Deliverable Format: The deliverable for this project will be simply a link to your video presentation on Screencast-O-Matic (https://screencast-o-matic.com), which will be shared with the class.

The video itself should be 12-15 minutes in length and must include the following elements:

- An introduction to your topic/research questions (1-2 minutes)
- An explanation of the data sources used and what they were used for. You will need to **show** the contents of each of the tables used (1-2 minutes)
- A review of the data cleaning and shaping procedures used and the data model used (3-4 minutes)
- A discussion of your findings and their relevance to your initial research question, including demonstrations of any interactive elements in your visualizations (6-8 minutes)

REMINDER: Screencast-O-Matic limits the length of videos recorded using its free accounts to 15 minutes. So be sure to be under that maximum time limit.

Due Date: Friday, April 23, 8:00 am EDT **(NOTE CHANGE FROM SYLLABUS)**