Capstone Project Ideas

Fundamentals of Data Science

Kevin Tajkowski

Idea # 1

Identify trends in the leading causes of death in the 50 United States and the District of Columbia for the 17 year period of 1999 - 2015. The data set is available at this CDC (Centers for Disease Control and Prevention) web page: https://data.cdc.gov/NCHS/NCHS-Leading-Causes-of-Death-United-States/bi63-dtpu and can be downloaded as a .csv file. The data set is provided by the National Center for Health Statistics. Note: You may have to copy and paste the link into a web browser.

The CDC and other health organizations would be interested in this information in order to identify the causes of death (not related to physical violence), where more research is needed in order to understand, develop treatments for and possibly cure the most common life-threatening diseases.

Idea #2

Determine the effect of several variables such as humidity, pressure and wind speed on particulate matter at several locations in the city of Beijing, China for the period of January 01, 2010 to Dec 31, 2015. The data is available at http://archive.ics.uci.edu/ml/datasets.html. For easy access to the data choose 'Physical Sciences'. The data set is titled 'PM2.5 Data of Five Chinese Cities' and downloads as a .rar file, which can be extracted to .csv files.

Liang, X., S. Li, S. Zhang, H. Huang, and S. X. Chen (2016), PM2.5 data reliability, consistency, and air quality assessment in five Chinese cities, J. Geophys. Res. Atmos., 121, 10220â€“10236.

Various health organizations would be interested in this data in order to identify the effects of weather on pollutants that may cause health problems.

Idea #3

Identify the countries or regions with the greatest and smallest increases in population from 1950 to 2015 based on estimated population data. The data is available from the United Nations website at https://esa.un.org/unpd/wpp/Download/Standard/Population. The data set is titled 'Total Population - Both Sexes' and is available as an Excel file.

Any individual or organization interested in world population could use these results combined with other data sets to study the effects of violent conflict, poverty and large-scale natural disasters on local populations.