Capstone Project 1 Proposal

The upstream of oil and gas industry is a combination of geophysics, geology and engineering, and the complicated and varied nature of the underground formations makes the industry a mixture of science and art. Therefore people must be careful when applying previous experience to new projects through data science analysis.

Production and injection are two technical considerations in upstream oil and gas industry. The injection fluid to underground formations is used to push the oil and gas out from production wells. This project will analyze the oil, gas, water and CO2 production and injection patterns from various formations and the active well changes in New Mexico. The spill incidents are analyzed through inferential statistics and machine learning skills. The goal of this project is to provide oil and gas companies with useful reference information on water flooding and well type design to enhance oil and gas recovery, and how to avoid major spill incidents.

This capstone project contains three parts of data science studies: basic data analysis, inferential statistics and machine learning.

To practice the basic data analysis skills, I'm choosing the formation, well, production and injection data (ftp://164.64.106.6/Public/OCD/OCD%20Interface%20v1.1/), collected by the Oil Conservation Division, which regulates oil, gas, and geothermal activity in New Mexico. The basic data analysis will study the well type and injection and production volumes from each formation from 1970s to February 2019.

To practice inferential statistics skills, I'm choosing the spill incidents data from 1980s until now (https://wwwapps.emnrd.state.nm.us/ocd/ocdpermitting/Data/Spills/Spills.aspx), collected by the Oil Conservation Division, which regulates oil, gas, and geothermal activity in New Mexico. The inferential statistics will study whether the wells tends to have a higher probability to cause a major incident than facilities given that an incident occurred.

To practice machine learning skills, I'm using the same spill incidents data as the inferential statistics study. The machine learning will apply supervised k-neighbors classification to predict the missing incident severity values from the known incident severity and incident features. And random forest will apply to calculate the incident feature importance.

These datasets are public and free to download.

My preliminary plan is this:

1. Download the datasets, unzip, and convert these XML files (about 65GB) to flat files (CSVs)

2. Identify the interesting tables and columns for this project, data wrangling and cleaning and deal with missing data
3. Explore data for interesting correlations between production, injection, and well types; and study the spill incident features, impacts
4. Apply z proportion test on spill incident data to prove whether wells are more likely to cause major incidents than facilities give an incident occurred
5. Use supervised classification machine learning to predict the missing values of incident severity
6. Wrap up the project reports in documentation, Jupyter Notebook containing codes, graphics and texts, and a slide deck.