Contents

Figures

Capstone Project 1: Final Report

Chapter 1 Problem statement

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.

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.

Chapter 2 Description and wrangling of the datasets

1. Datasets used for basic data science analysis

The first step I did after downloading the data from the website of New Mexico Oil Conservation Division is to convert the files from XML to CSV format, appling xml.etree.ElementTree.iterparse function. The code written for this purpose will be included in the repository on GitHub.

Then I convert the CSVs to Pandas data frames, applying pd.read_csv function, and there are 16 data frames in total. After reviewing the contents of each data frame with the help of the data dictionary provided by OCD ('OCD Interface v1.1 Data Dictionary.xlsx', will be upload to GitHub), I choose 5 out of the 16 tables for this project to analyze the well completion designs, by ignoring the tables used to track the resource reports, acreage and spacing regulations, and records of a custody transfer off a property for oil and gas. The 5 tables I will use for exploratory data analysis are: *pool, wchistory, wellhistory, wcinjection, wcproduction* (explanation of the tables at the end of this document).

Next step is to evaluate each of the chosen tables: In this section, I combine related columns into one column, reduce the size of big tables by only keeping useful columns for further analysis, remove trailing spaces, replace missing values as numpy NaNs, and fill the NaNs.

- *pool*: producing formation or subdivision of producing formation properties: producing property
    - clear the trailing spaces in pool name column *pool_nam*,
    - cut down the table wchistory size by only keeping the columns useful in further analysis:

        - *pool_idn* - producing formation or subdivision of producing formation as defined by Order Hearing Order
        - *pool_nam* - name assigned to pool when pool is created by an R-Order
        - *pool_typ_cde* - pool type, as determined by the type of fluids in the pool
            - 1: Gas (Prorated)
            - 2: Gas (Non-prorated)
            - 3: Oil
            - 4: Associated
            - 5: Salt Water Disposal
- *wchistory*: well completion data
    - replace missing values with numpy NaNs,
    - combine three columns (*api_st_cde*, *api_cnty_cde*, *api_well_idn*) to be one column *api* to identify wells,
    - cut down the table wchistory size by only keeping the columns useful in further analysis:
        - *api* - well identify number
        - *pool_idn* - producing formation or subdivision of producing formation identifier
        - *eff_dte* - the date which the record is effective
        - *rec_termn_dte* - date record no longer effective
        - *wc_stat_cde* - status of well completion
            - A: Active
            - C: Cancelled
            - D: Dry Hole
            - N: New, Not Drilled
            - X: Never Drilled
            - T: Temporary Abandonment
            - P: Zone Permanently Plugged
            - Z: Zones Temporarily Plugged
        - *well_typ_cde* - well type code
            - C: CO2
            - G: Gas
            - I: Injection
            - M: Miscellaneous
            - O: Oil
            - S: Salt Water Disposal
            - W: Water
        - *prodn_meth_cde* - producing method code
            - P: Pumping
            - F: Flowing
            - G: Gas Lift
    - fill the NaNs in column *prodn_meth_cde* from the value of the same well (same *api*),
    - change production method 'D' and 'I' to 'F' (flowing) according to search on website

(https://wwwapps.emnrd.state.nm.us/ocd/ocdpermitting/Data/Wells.aspx), since the data dictionary doesn't explain the codes.

- *wellhistory*: well data
  - replace the missing values with numpy NaNs,
  - combine three columns (*api_st_cde, api_cnty_cde, api_well_idn*) to be one column *api* to identify wells,
  - fill the NaNs in column *directional_status* from the value of the same well (same *api*),
  - cut down the table wchistory size by only keeping the columns useful in further analysis:
    - *api* - well identify number
    - *directional_status* - differentiate vertical, horizontal, directional
- *wcinjection*: *wcinjection*: injection data
  - remove the trailing spaces in *inj_knd_cde* column of table wcinjection,
  - replace missing values with NaNs,
  - combine three columns (*api_st_cde, api_cnty_cde, api_well_idn*) to be one column *api* to identify wells,
  - delete the three columns to cut down table size, and the useful columns:
    - *api* - well identify number
    - *pool_idn* - identifier code for specific pool
    - *prodn_mth* - month in which injection occurred
    - *prodn_yr* - year that injection occurred
    - *inj_knd_cde* - product kind code
      - W: Water
      - G: Gas
      - C: CO2
      - O: Other
- *wcproduction*: production data
  - cut down the table *wchistory* size by only keeping the columns useful in further analysis,
  - combine three columns (*api_st_cde, api_cnty_cde, api_well_idn*) to be one column *api* to identify wells,
  - further cut down the table wchistory size by deleting the three columns:
    - *api* - well identify number
    - *pool_idn* - identifier code for specific pool
    - *prodn_mth* - month in which production occurred
    - *prodn_yr* - year that production occurred
    - *prd_knd_cde* - product kind code
      - O: Oil
      - G: Gas
      - W: Water
    - *prod_amt* - volume of fluid produced (MCF or BBLS)
  - remove the trailing spaces in *prd_knd_cde* column

Some of the NaNs are ignorable because of the large data size, for example, only 1 NaN out of 338,720 rows in *well_typ_cde* in table *wchistory*. The NaNs in column *directional_status* in

table *wellhistory* are forward filled after grouping by 'api', and the NaNs in column *prodn_meth_cde* are filled from the value of the same well (same *api*). However, there are still 40713 out of 257745 rows missing *direction_status* information, and 54208 out of 338720 rows missing *prodn_meth_cde* information. I will pay attention when use the column combined with other tables. I need to point out that some of the columns may contain thousands of NaNs, but I'm not using those columns to solve the problem of this project.

Even though the boxplot of gas production in Figure 1 suggests the existence of outliers, those 'outliers' are good data points to analyze the correlation between production and well completion method, so they are included in my analysis.

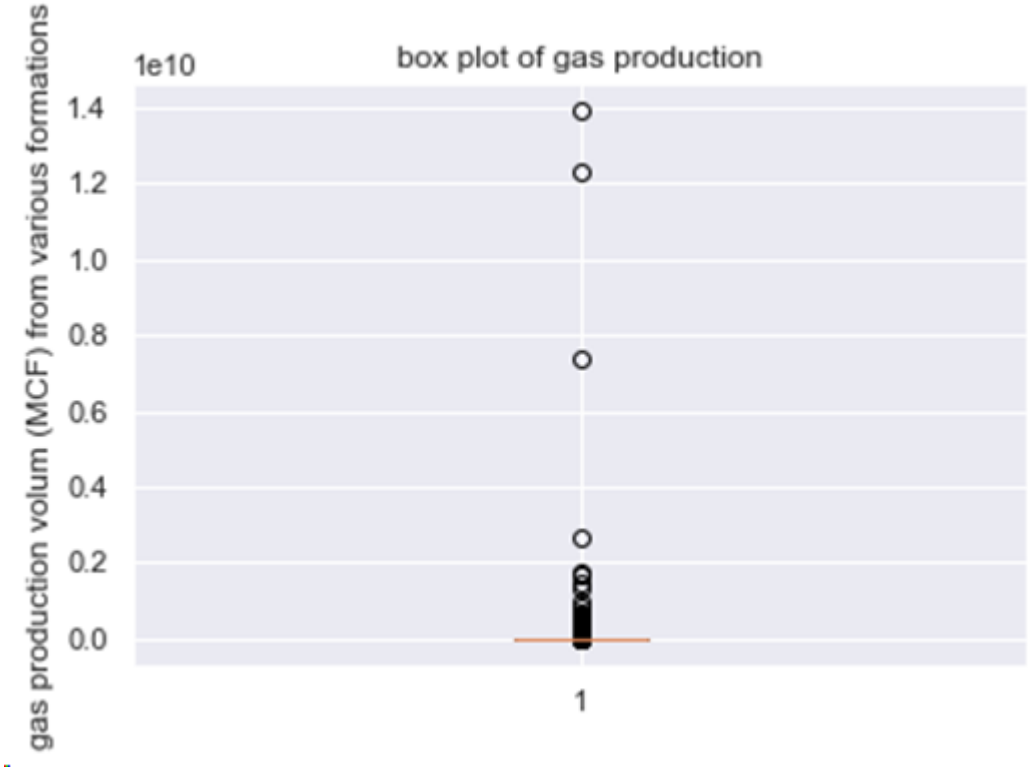

Figure 1 Review gas production in boxplot

2. Datasets used for inferential statistics and machine learning

I use the *spills*.csv file containing oil and gas field spills incidents data in New Mexico, downloaded from the website of New Mexico Oil Conservation to practice inferential statistics and machine learning skills, and the columns are going to be analyzed are:

- *Incident Number*: incident identifier
- *Facility*: facility identifier if the incident happened in a facility
- *API*: well identifier if the incident happened in a well

- *Operator Name*
- *Severity*
- *Incident Type*
- *Incident Date*
- *Material Spilled*
- *Volume Spilled*
- *Volume Recovered*
- *Spill Cause*
- *Spill Source*
- *District*
- *County*
- *Waterway Affected*
- *Groundwater Impact*

Forward and backward fill the NaNs in the table with the information of the same incident, drop duplicated rows, and convert the incident date as pandas datetime and set it as the index of the table.

Chapter 3 Basic data science analysis

Exploratory Data Analysis

In this section, I will explore each table individually or combining them together to get the correlations. Figures along with analysis are in below:

Figure 2 indicates that among the total *4978* formations (pools), *43(1%)* formations are gas(prorated), *1993(40%)* are gas (non-prorated), 2761(56%) are oil formations, *34(1%)* are associated formations, and *147(3%)* are salt water disposal formations.

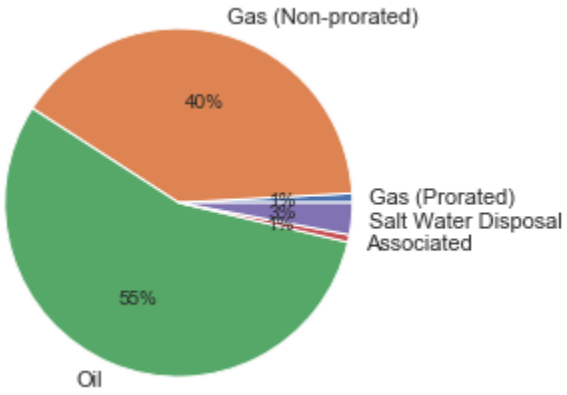

Figure 2 Types of producing formation or subdivision of producing formation

Figure 3 shows that the number of active wells and permanently plugged wells are dominate in the history. The number of active wells has increased linearly from around 22k to 65k, since 1970s until the year around 2010, and the number has been stable since 2010.

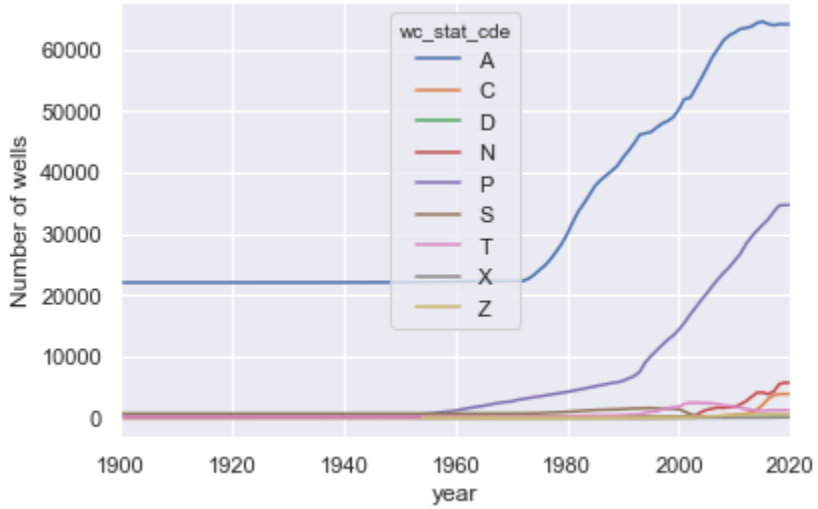

Figure 3 Total numbers of wells varying completion status from 1900 to 2019

The following three pairs of plots compare the number of total wells and the number of active wells, with varying well types, well directional status, and producing methods.

Figure 4 shows the total number of gas wells and total number of oil wells are both increasing linearly since 1980s, but the increase of active oil wells has slowed down since 2000, with current number of active oil wells around 27k. The number of active gas wells has dropped from around 34.3k to 33k since 2010. The number of active injection wells started to reduce in early 1990s, with current number of injection wells around 3k.

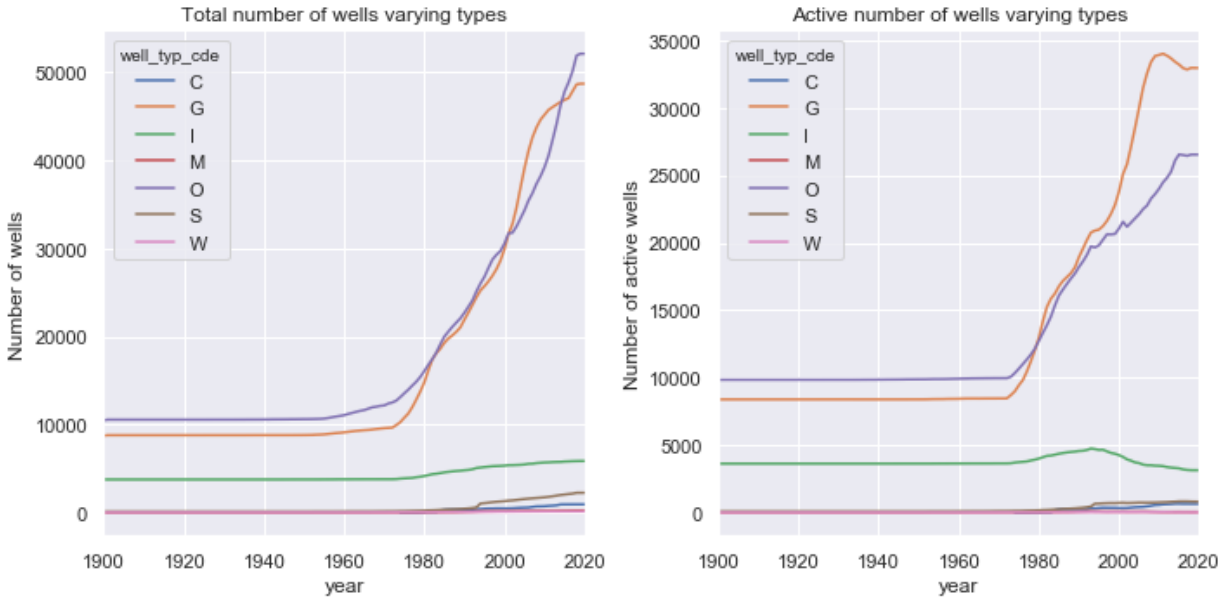

Figure 4 Total and active number of wells varying well types from 1900 to 2019

Figure 5 shows the vertical wells are dominate among both all kinds of wells and active wells in the history. The number of active vertical wells has grown linearly from 22k to 60k since 1970s, and it decreased about 3k since 2010, while the active horizontal and directional wells have increased from 0 to 5.5k and 1.5k since 2010.

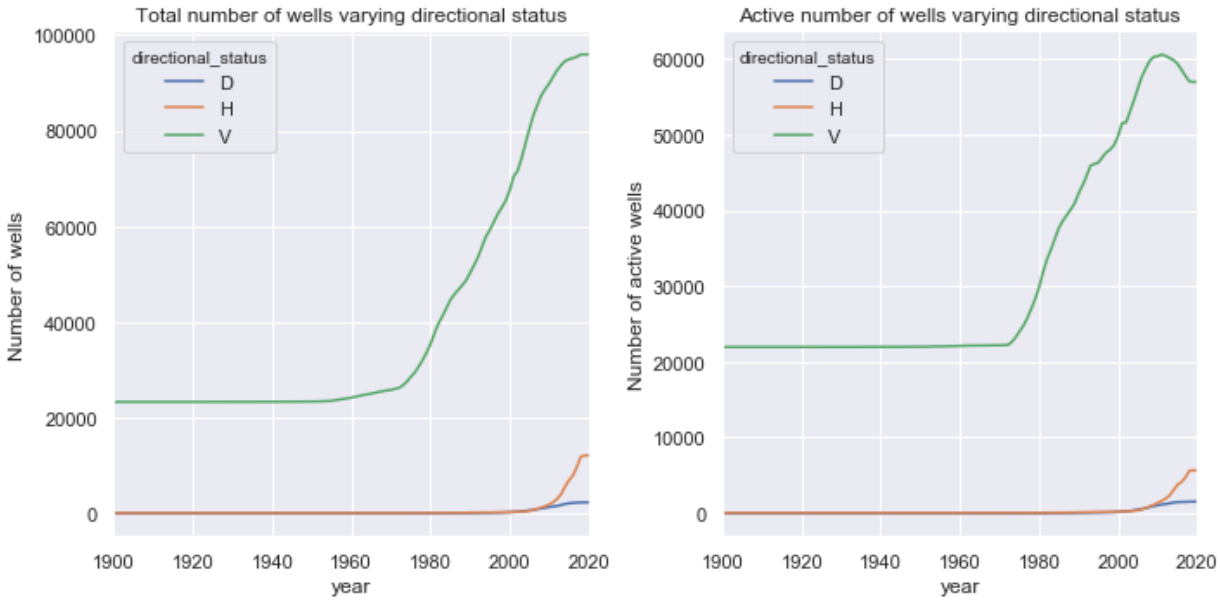

Figure 5 Total and active number of wells varying well directional status from 1900 to 2019

Figure 6 shows that the Flowing and pumping are two dominate producing methods, and flowing has been more popular than pumping since 1970s when they started to increase linearly. The growth of active flowing producing method started to slow down since 2008, with 36k wells currently applying flowing producing method. The active pumping (current 25k well) started to drop in 2012 when gas lift was introduced onsite.

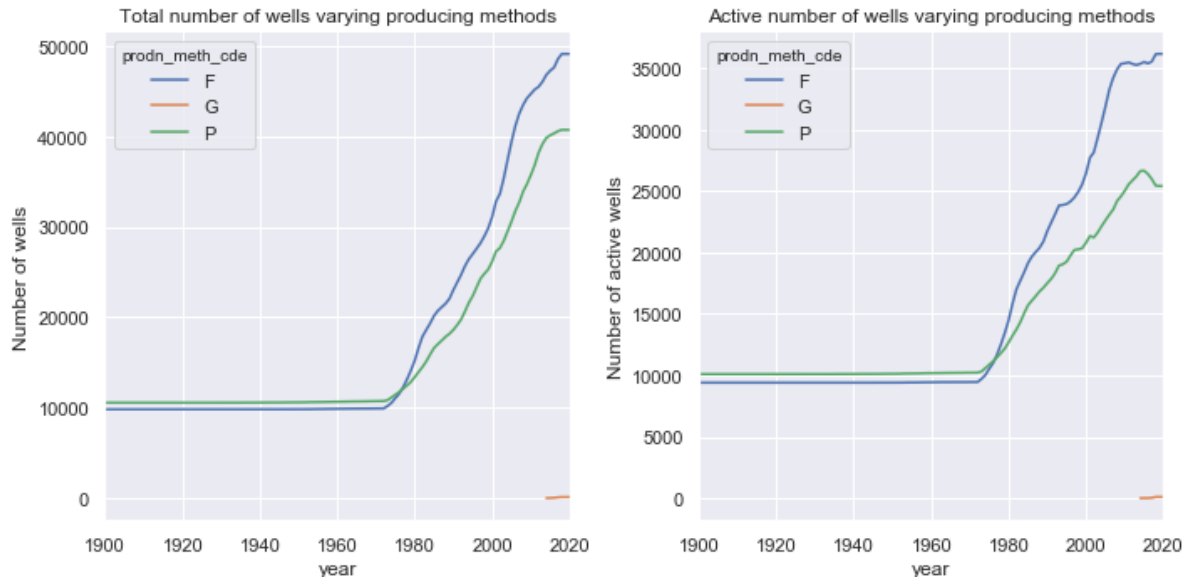

Figure 6 Total and active number of wells with varying producing methods from 1900 to 2019

Figure 7 shows the total water injection volume, the annual water injection volume into each formation, and the top 5 formations with the most total water injection volume. The peak of water injection happened in early 1990s, about *4.7e9* bbls/year. After the peak point, the total water injection volume increases linearly below *1.0e9* bbls/year. The top 5 formations which have the most total water injection volume are:

- HOBBS;GRAYBURG-SANANDRES, with total injection volume 3.42e9 bbls
- SWD;DEVONIAN, with total injection volume 2.16e9 bbls
- VACUUM;GRAYBURG-SANANDRES, with total injection volume 1.61e9 bbls
- EUNICEMONUMENT;GRAYBURG-SANANDRES, with total injection volume 1.30e9 bbls
- LANGLIEMATTIX;7RVRS-Q-GRAYBURG, with total injection volume 1.26 bbls

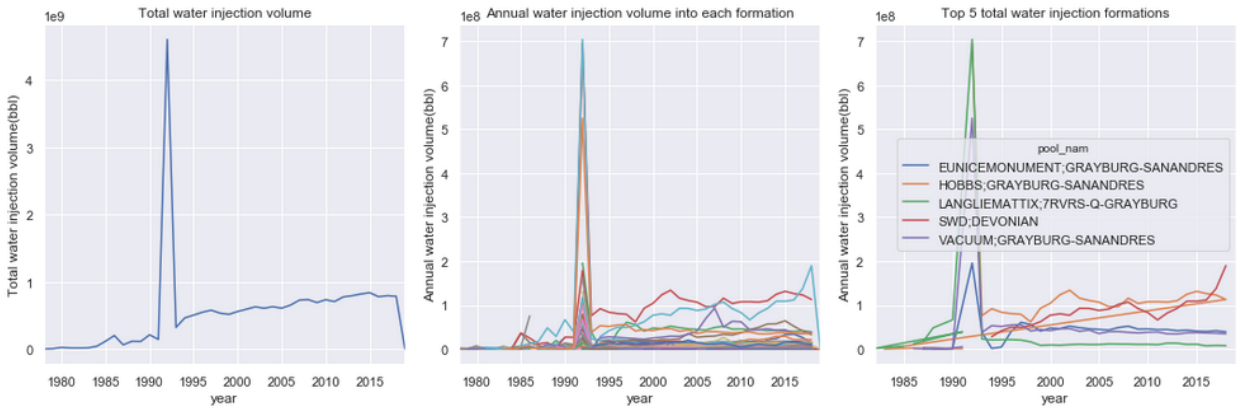

Figure 7 Total and annual water injection in each formation and top 5 water injection formations from 1980s to 2019

Figure 8 shows the total gas injection volume, the annual gas injection volume into each formation, and the top 5 formations with the most total gas injection volume. The peak of gas injection happened in early 1990s, about *2.3e8* MCF/year. After the peak point, the total gas injection volume keeps below *1e7* MCF/year. The top 5 formations which have the most total gas injection volume are:

- EMPIRE;ABO, with total injection volume 2.57e8 MCF
- PUERTOCHIQUITOMANCOS,WEST, with total injection volume 3.56e7 MCF
- GRAMARIDGE;MORROW(GAS), with total injection volume 2.47e7 MCF
- AGI;WOLFCAMP, with total injection volume 1.17e7 MCF
- CAPROCK;QUEEN, with total injection volume 7.07e6 MCF

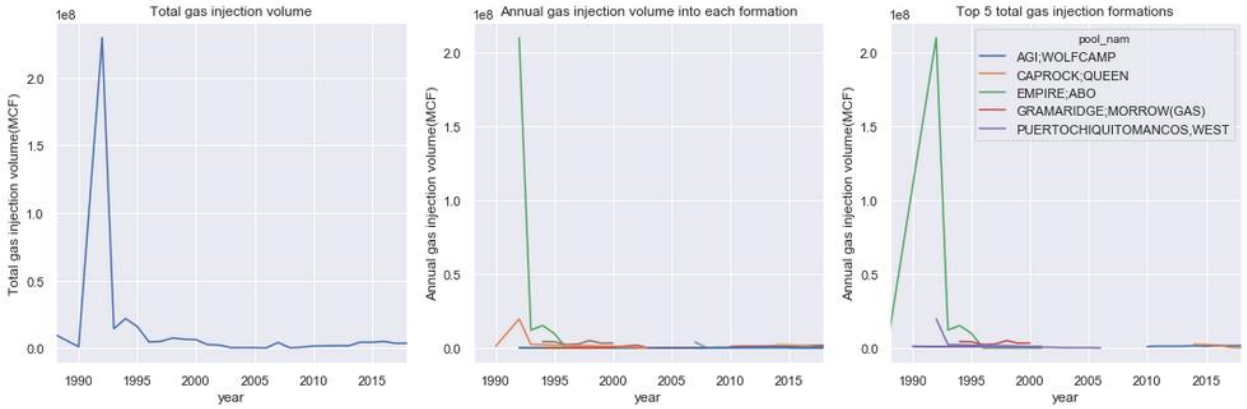

Figure 8 Total and annual gas injection in each formation and top 5 gas injection formations from 1990s to 2019

Figure 9 shows the total CO2 injection volume, the annual CO2 injection volume into each formation, and the top 5 formations with the most total CO2 injection volume. The peak of CO2 injection happens in early 1990s, about 6.7e7 MCF/year. After the peak point, the total CO2 injection volume increases linearly with low peak at 2006, about 3.9e7 MCF/year. The top 5 formations which have the most total CO2 injection volume are:

- VACUUM;GRAYBURG-SANANDRES, with total injection volume 7.98e8 MCF
- HOBBS;GRAYBURG-SANANDRES, with total injection volume 7.92e8 MCF
- MALJAMAR;GRAYBURG-SANANDRES, with total injection volume 1.62e7 MCF
- CAPROCK;QUEEN, with total injection volume 1.30e7 MCF
- AGI;SANANDRES, with total injection volume 5.44e6 MCF

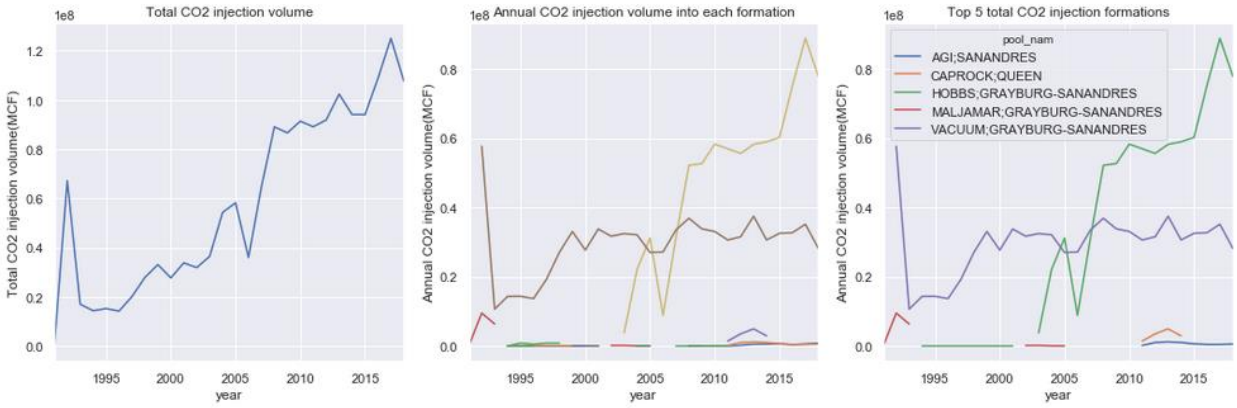

Figure 9 Total and annual CO2 injection in each formation and top 5 CO2 injection formations from 1990s to 2019

Figure 10 shows the annual total gas production in New Mexico state, and the annual gas production in each formation, and the top 5 formations which have the most total gas production. The annual gas production reaches peak in early 1990s, around 3e10 MCF/year, and then keeps stable between 2e9 MCF/year and 3e9 MCF/year. The individual formation follows the pattern, and the top 5 gas production formations are :

- BLANCO-MESAVERDE(PRORATEDGAS), with total production volume 1.39e10 MCF
- BASINFRUITLANDCOAL(GAS), with total production volume 1.23e10 MCF
- BASINDAKOTA(PRORATEDGAS), with total production volume 7.39e9 MCF
- BRAVODOMECARBONDIOXIDEGAS640, with total production volume 2.69e9 MCF
- EUMONT;YATES-7RVRS-QUEEN(GAS), with total production volume 1.75e9 MCF

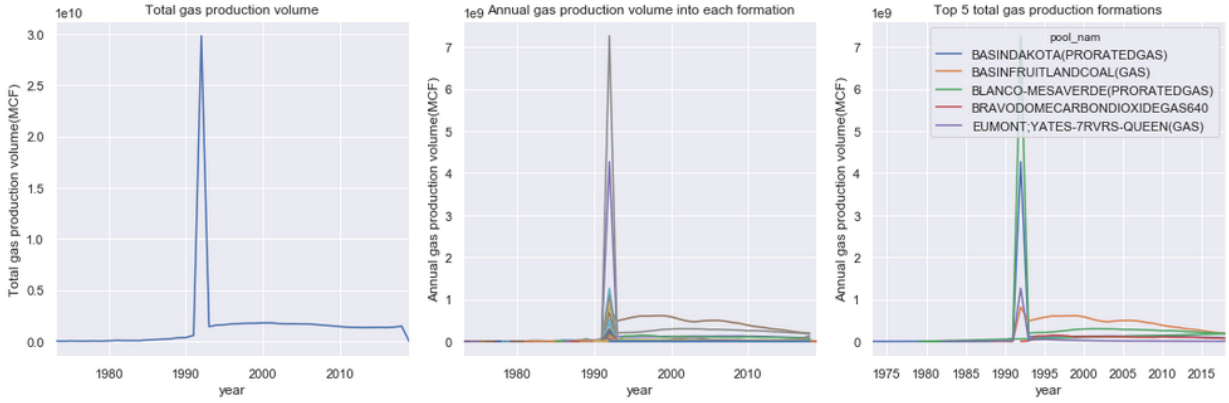

Figure 10 Total and annual gas production in each formation and top 5 gas production formations from 1970s to 2019

Figure 11 shows the annual total oil production, and the annual oil production in each formation, and the top 5 formations with the most total oil production. The annual total oil production reaches peak in early 1990s, around 2.8e9 bbls/year, and then keeps stable around 1e9 bbls/year until 2010, and the total annual production has increased to about 2e9 bbls/year since 2010. The individual formation follows this pattern that the production reaches peak in early 1990s. The top 5 formations which produce the most oil are:

- VACUUM;GRAYBURG-SANANDRES, with total production volume 3.97e8 bbls
- HOBBS;GRAYBURG-SANANDRES, with total production volume 3.89e8 bbls
- EUNICEMONUMENT;GRAYBURG-SANANDRES, with total production volume 3.70e8 bbls
- EMPIRE;ABO, with total production volume 2.15e8 bbls
- MALJAMAR;GRAYBURG-SANANDRES, with total production volume 1.55e8 bbls

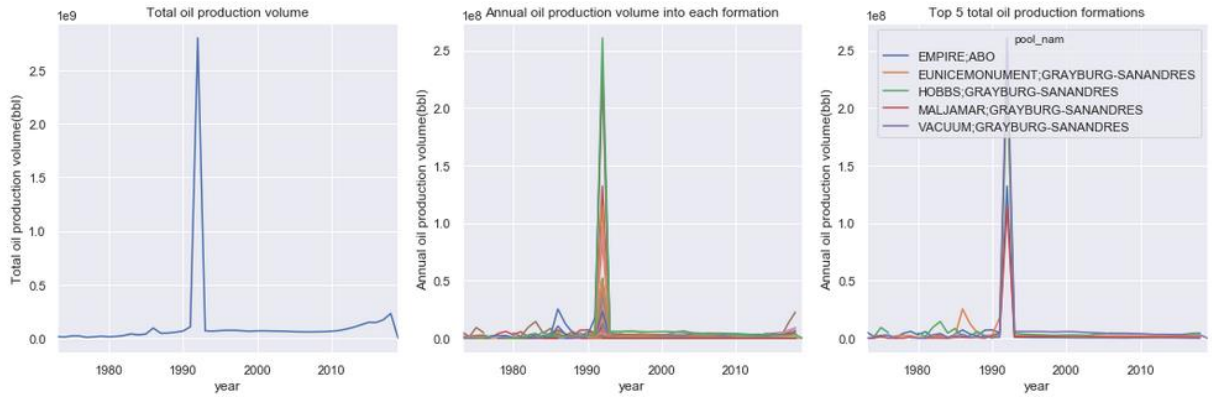

Figure 11 Total and annual oil production in each formation and top 5 oil production formations from 1970s to 2019

Figure 12 shows the annual total water production, and the annual water production in each formation, and the top 5 formations with the most total water production. The peak annual total water production happens in early 1990s, around 5.7e9 bbls/year, after that the water production volume increases linearly from 4e8 bbls/year to 1e9 bbls/year. The individual formation follows similar pattern. The top 5 formations which produce the most water are:

- HOBBS;GRAYBURG-SANANDRES, with total production volume 3.78e9 bbls
- VACUUM;GRAYBURG-SANANDRES, with total production volume 1.47e9 bbls
- EUNICEMONUMENT;GRAYBURG-SANANDRES, with total production volume 1.43e9 bbls
- INDIANBASIN;UPPERPENN(ASSOC), with total production volume 8.19e8 bbls
- JALMAT;TAN-YATES-7RVRS(OIL), with total production volume 8.13e8 bbls

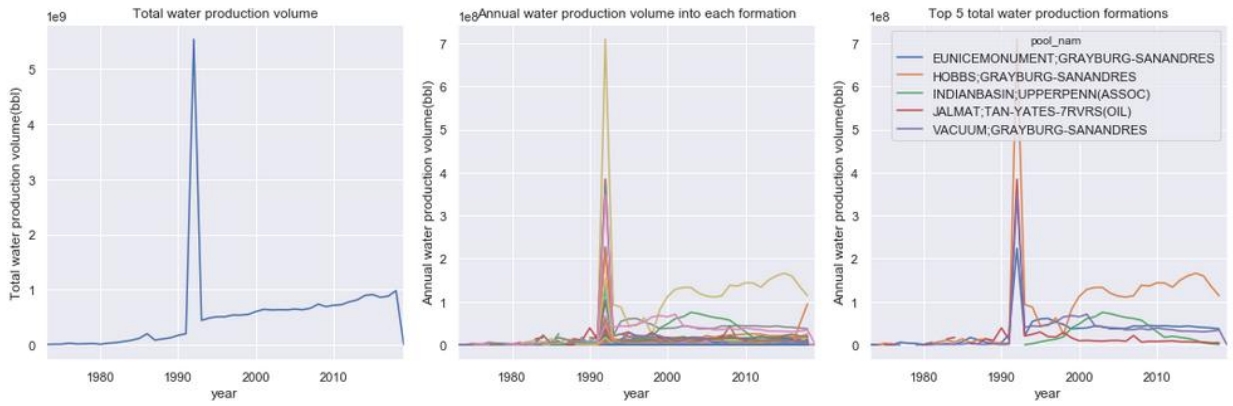

Figure 12 Total and annual water production in each formation and top 5 water production formations from 1970s to 2019

Figure 13 shows the annual total CO2 production in New Mexico state, and the annual CO2 production in each formation, and the top 3 formations which have the most total CO2 production. The annual CO2 production reaches peak in 1992, around 9.5e8 MCF/year. There are three formations produced CO2 from late 1980s to early 1990s, and the total CO2 production from each formation is:

- BRAVODOMECARBONDIOXIDEGAS640, with total production volume 1.06e10 MCF
- WESTBRAVODOMECO2GAS, with total production volume 7.37e6 MCF
- NORTHEASTCOUNTIESUNDES,CO2, with total production volume 5.30e5 MCF

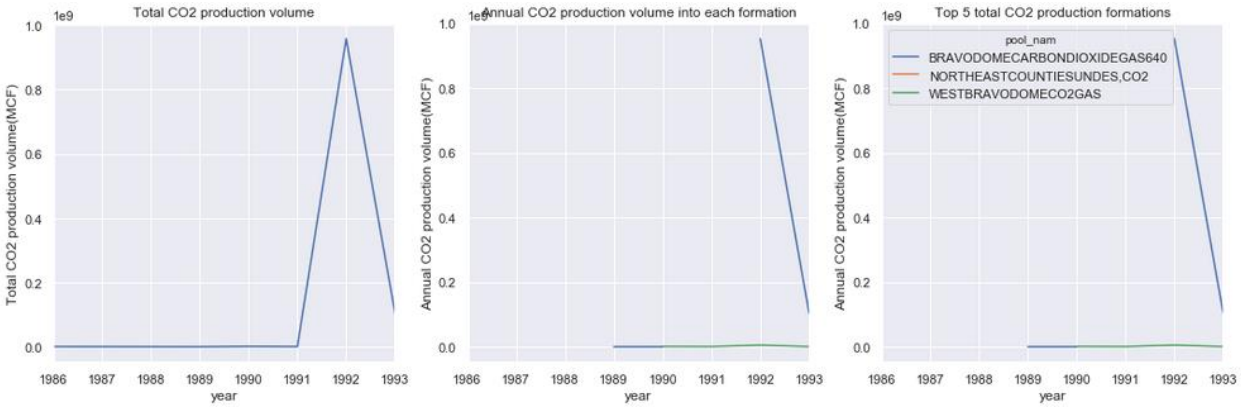

Figure 13 Total and annual CO2 production in each formation and top 5 CO2 production formations from 1986 to 1993

Figure 14 indicates annual total gas, oil and water production volume from vertical, horizontal, and directional wells. It shows that the vertical wells dominate the production in the history until 2010s when horizontal wells started to apply in oil and gas fields. The gas production from horizontal wells is very close to the production from vertical wells in 2018, with the tendency to exceed in coming year. The oil production from horizontal wells has exceeded the production from vertical well in 2013, and the production volume from horizontal wells is about four times of the production from vertical wells. At the

same time, water production from horizontal wells increases and exceeds the production from verti cal wells in 2017.

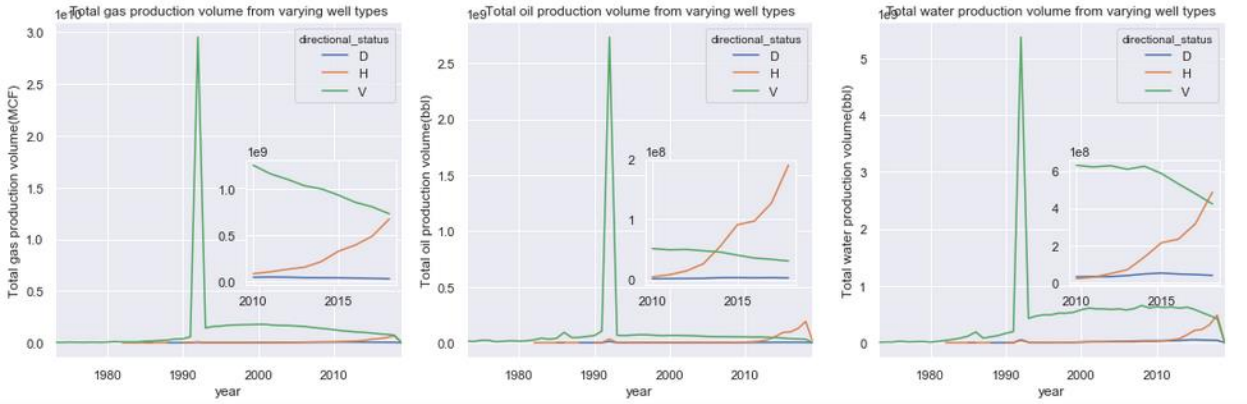

Figure 14 Total gas, oil and water production in each formation and top 5 gas production formations from 1970s to 2019

The following is converting the production gas to barrel of oil equivalent (BOE), and analyzing the relationship between production and injection. The barrel of oil equivalent (BOE) is a unit of energy based on the approximate energy released by burning one barrel (42 U.S. gallons or 158.9873 litres) of crude oil. The USGS gives a figure of 6,000 cubic feet (170 cubic meters) of typical natural gas are equivalent to one BOE.

Figure 15 the logarithmic scale plot shows the total oil/gas production after converting vs. total water production from directional, horizontal, and vertical wells respectively. Vertical wells lead both oil/gas (5e3 MMBL) and water (2e4 MMBL) production; total 7e2 MMBL oil and gas, and 1.8e3 MMBL are produced from horizontal wells; total 56 MMBL oil and gas, and 7e2 MMBL water are produced from directional wells.

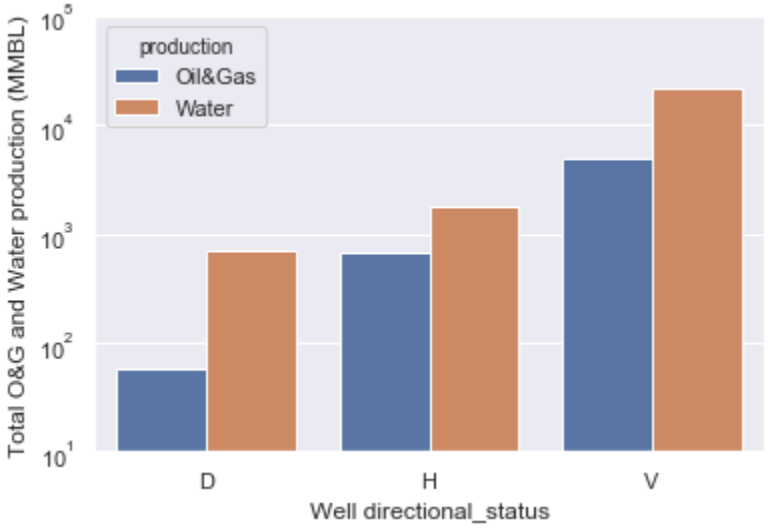

Figure 15 Total oil/gas and water production volume from different wells

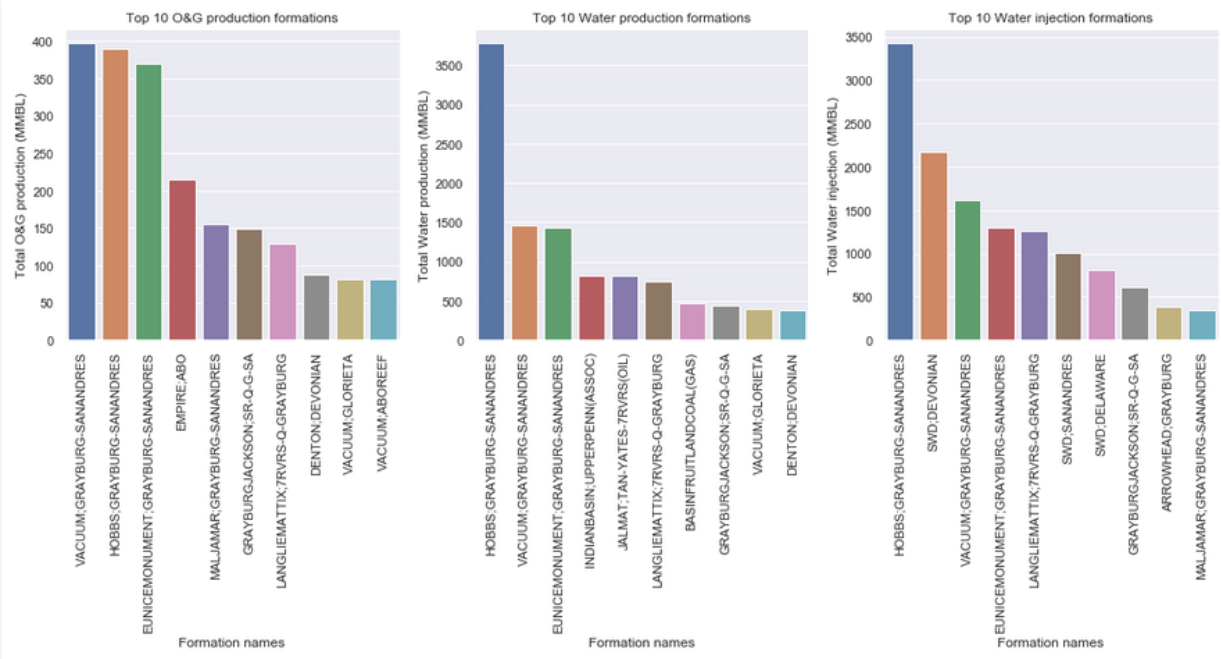

Figure 16 The top 10 oil/gas production, water production and injection formations

The above Figure 16 shows the top 10 oil and gas production formations, top 10 water production formations, and top 10 water injection formation. There is a high correlation between production and injection generally, meaning more injection and more production, for example the No.1 water injection formation *HOBBS;GRAYBURG-SANANDRES* produces the No.1 water and No.2 oil and gas.

In reality, we prefer high oil and gas production and low water production with either high or low water injection, and trying to avoid the opposite. The No.2 water injection formation *SWD;DEVONIAN* doesn't production corresponding oil and gas, which is out of top 10 oil and gas production formations; and the formation is out of top 10 water production either, thus it's interesting to further study on the real formations which water injection into. The No.4 oil and gas production formation *EMPIRE;ABO* is out of the top 10 water production and injection, which is a signal of high water flooding efficiency. Figure 17 indicate positive linear relationships between water production and oil/gas production, water injection and oil/gas production, and water injection and water production from various formations.

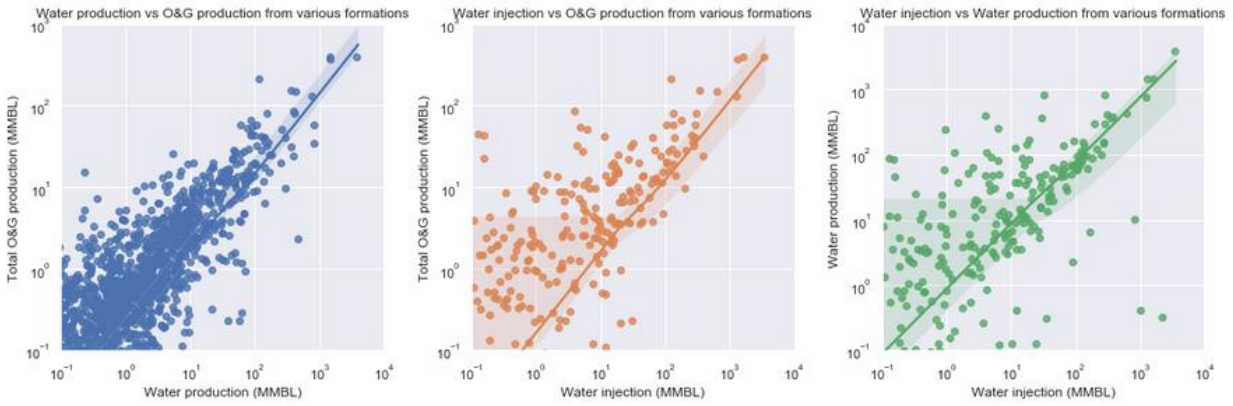

Figure 17 The linear relationships between water production and oil/gas production, water injection and oil/gas production, and water injection and water production from various formations.

Conclusion

- Among the total *4978* formations (pools), *43(1%)* formations are gas(prorated), *1993(40%)* are gas (non-prorated), 2761(56%)* are oil formations, *34(1%)* are associated formations, and *147(3%)* are salt water disposal formations.
- The number of active wells has increased linearly from around 22k to 65k, since 1970s until the year around 2010, and the number has been stable since 2010.
- The total number of gas wells and total number of oil wells are both increasing linearly since 1980s, but the increase of active oil wells has slowed down since 2000, with current number of active oil wells around 27k. The number of active gas wells has dropped from around 34.3k to 33k since 2010. The number of active injection wells started to reduce in early 1990s, with current number of injection wells around 3k.
- The vertical wells are dominate among both all kinds of wells and active wells in the history. The number of active vertical wells has grown linearly from 22k to 60k since 1970s, and it decreased about 3k since 2010, while the active horizontal and directional wells has increased from 0 to 5.5k and 1.5k since 2010.
- Flowing and pumping are two dominate producing methods, and flowing has been more popular than pumping since 1970s when they started to increase linearly. The growth of active flowing producing method started to slow down since 2008, with 36k wells currently applying flowing producing method. The active pumping (current 25k well) started to drop in 2012 when gas lift was introduced onsite.
- Both the injection (water, gas, CO2) and production (gas, oil, water, and CO2) reach peak in 1992:
    - The peak water injection is about *4.7e9* bbls/year. After the peak point, the total water injection volume increases linearly below *1.0e9* bbls/year.
    - The peak gas injection is about *2.3e8* MCF/year. After the peak point, the total gas injection volume keeps constant below *1e7* MCF/year.
    - The peak CO2 injection is about 6.7e7 MCF/year. After the peak point, the total CO2 injection volume increases linearly with a low peak at 2006, about 3.9e7 MCF/year.
    - The peak annual gas production is around 3e10 MCF/year, and then the production keeps stable between 2e9 MCF/year and 3e9 MCF/year.

    - The peak annual oil production is around 2.8e9 bbls/year, and then the production keeps stable around 1e9 bbls/year until 2010, and the total annual production has increased to about 2e9 bbls/year since 2010.
    - The peak annual total water production is around 5.7e9 bbls/year, after that the water production volume increases linearly from 4e8 bbls/year to 1e9 bbls/year.
    - There are three formations produced CO_2 from late 1980s to early 1990s, and peak annual CO_2 production is around 9.5e8 MCF/year.
- The vertical wells dominate the production in the history until 2010s when horizontal wells started to apply in oil and gas fields. The gas production from horizontal wells is very close to the production from vertical wells in 2018, with the tendency to exceed in coming year. The oil production from horizontal wells has exceeded the production from vertical well in 2013, and the production volume from horizontal wells is about four times of the production from vertical wells. At the same time, water production from horizontal wells increases and exceeds the production from vertical wells in 2017.
- Converting gas to barrel of oil equivalent, the vertical wells lead both total oil/gas (5e3 MMBL) and total water (2e4 MMBL) production; total 7e2 MMBL oil and gas, and 1.8e3 MMBL are produced from horizontal wells; total 56 MMBL oil and gas, and 7e2 MMBL water are produced from directional wells.
- There is a positive linear relationships between water production and oil/gas production, water injection and oil/gas production, and water injection and water production from various formations. In reality, we prefer high oil and gas production and low water production with either high or low water injection, and trying to avoid the opposite. The No.2 water injection formation *SWD;DEVONIAN* doesn't production corresponding oil and gas, which is out of top 10 oil and gas production formations; and the formation is out of top 10 water production either, thus it's interesting to further study on the real formations which water injection into. The No.4 oil and gas production formation *EMPIRE;ABO* is out of the top 10 water production and injection, which is a signal of high water flooding efficiency.
- The efficiency of water flooding is good in some formations which have high oil and gas production, such as
    - HOBBS;GRAYBURG-SANANDRES,
    - VACUUM;GRAYBURG-SANANDRES,
    - EUNICEMONUMENT;GRAYBURG-SANANDRES,
    - EMPIRE;ABO,
    - MALJAMAR;GRAYBURG-SANANDRES

Future Work:

1. The influence of production methods on production;
2. Further water, gas and CO_2 flooding efficiency study on certain formations;
3. As the horizontal well starts to dominate production in the past 5 years, it's interesting to further study on horizontal well completion methods, such as hydraulic fracturing.

Chapter 4 Inferential statistics analysis

Exploratory Data Analysis

This section will study the correlations among the monthly number of incidents, volume spilled and recovered, severity, incident type, material spilled, cause, source, waterway and groundwater impacted, location distribution, and operators.

The monthly number of incidents plot Figure 18 shows the incidents since 1980s. There are more major incidents in the history. The monthly number of incidents is stable around 20/month from 1986 to 1995, goes down to below 10/month from 1996 to 1998, goes up to around 20/month in 1999, then goes down to below 10/month in 2000, after that it keeps increasing to around 60/month in 2018, and the maximum number of major and minor incident happens in 2016 which is more than 100/month.

Figure 19 shows the volume spilled in minor incidents is below 50 barrels with few exceptions, and the volume spilled in major incidents ranges widely up to 60,000 barrels. The volume recovered is less or equal to the volume spilled with few exceptions.

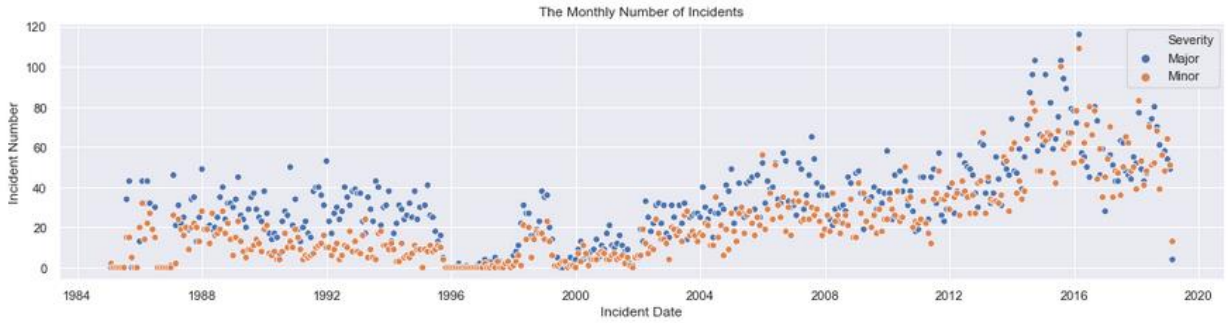

Figure 18 The monthly number of major and minor incidents changes from 1980s to 2019

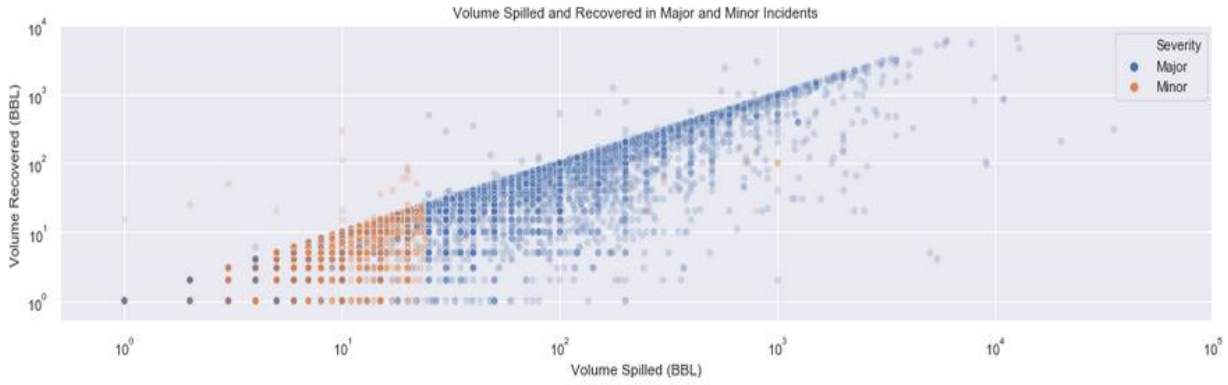

Figure 19 The spilled and recovered volumes in major and minor incidents

The correlation between monthly major and minor incidents plot Figure 20 indicates a positive linearly relationship between major and minor incidents. To reduce the number of incidents, we should try to avoid both of them.

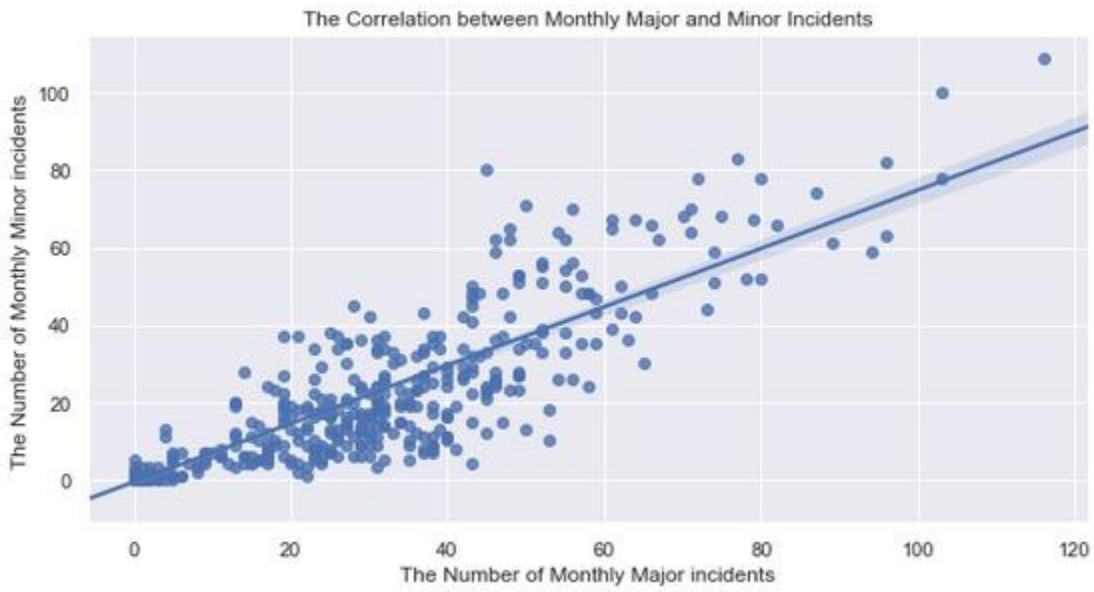

Figure 20 The correlation between monthly major and minor incidents

The following plots from Figure 21 to Figure 24 indicate the correlation between the number of major and minor incidents and incident type, material spilled, spill cause and spill source. The top three incident types are produced water release, oil release and natural gas release which is corresponding to the top three spilled materials; The top three spill causes are equipment failure, corrosion and human error; and the top three spilled sources are tank (any), flow line - production, and pipeline (any). We need to pay special attention to fire, triethylene and sulphuric acid spill, and generator, since they are related to major incident only.

Figure 25 and Figure 26 show the relationship between major and minor incident and groundwater impact and waterway affected, indicate that groundwater is more likely impacted by major incidents, and the waterway are 100% affected by major incidents. There is a loose correlation (0.7%) between groundwater impact and waterway affected.

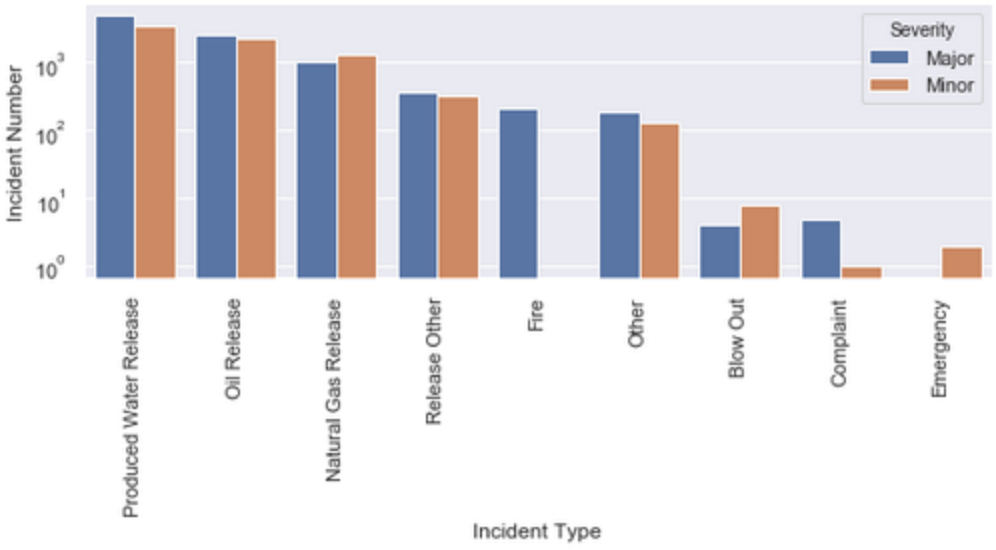

Figure 21 The number of major and minor incidents in various incident type categories

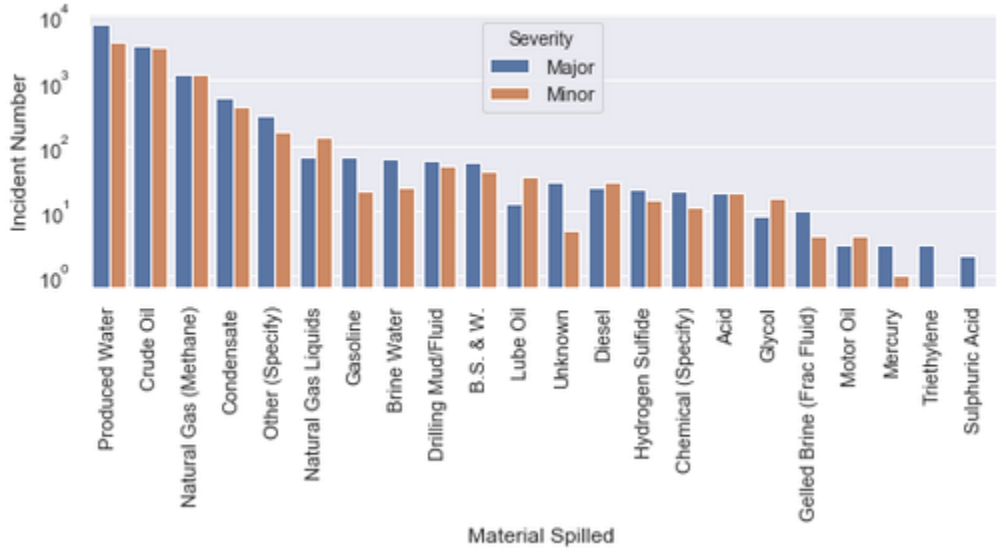

Figure 22 The number of major and minor incidents in various material spilled categories

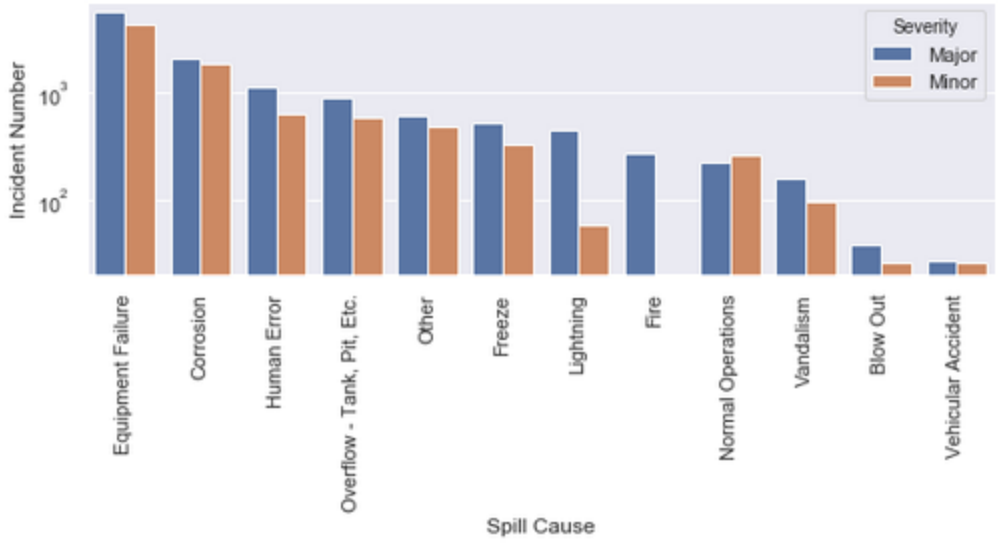

Figure 23 The number of major and minor incidents in various spill cause categories

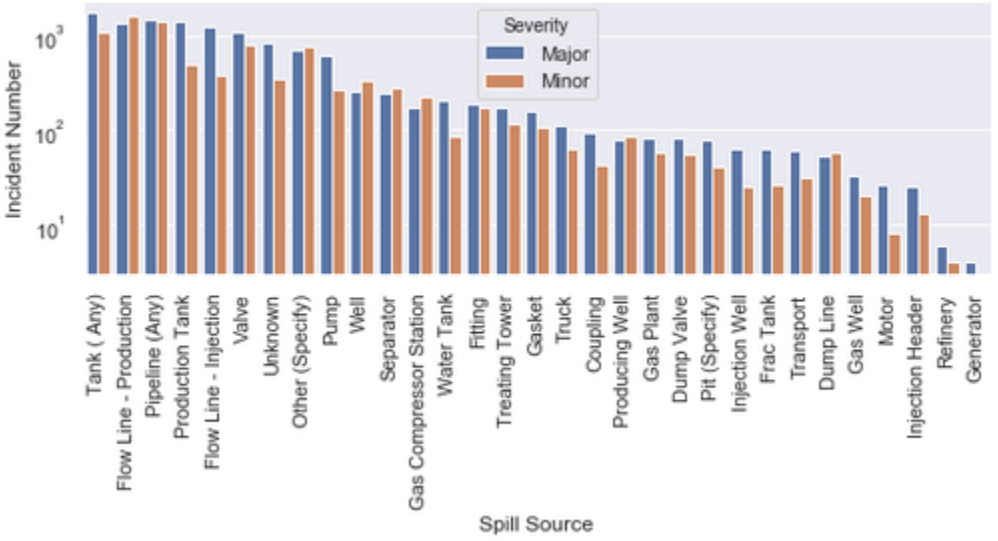

Figure 24 The number of major and minor incidents in various spill source categories

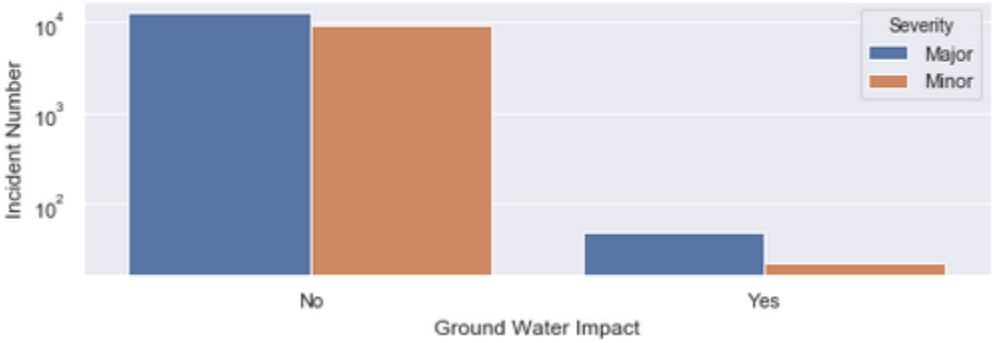

Figure 25 The number of major and minor incidents in ground water impact categories

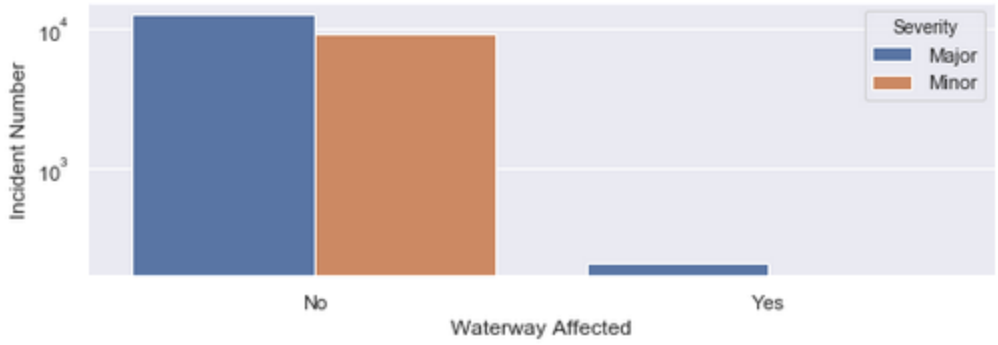

Figure 26 The number of major and minor incidents in waterway affected categories

The following Figure 27 shows the number of incidents in each district and counties in that district. The most of major and minor incidents happens in county Lea in Hobbs, Eddy in Artesia, San Juan in Aztec, and 0 (missing county name probably) in Santa Fe. The following counties have major incidents happened only: Eddy in Aztec, county Rio Arriba, Lea, Socorro, Torrance, Bernalillo, Cibola and Valencia in Santa Fe.

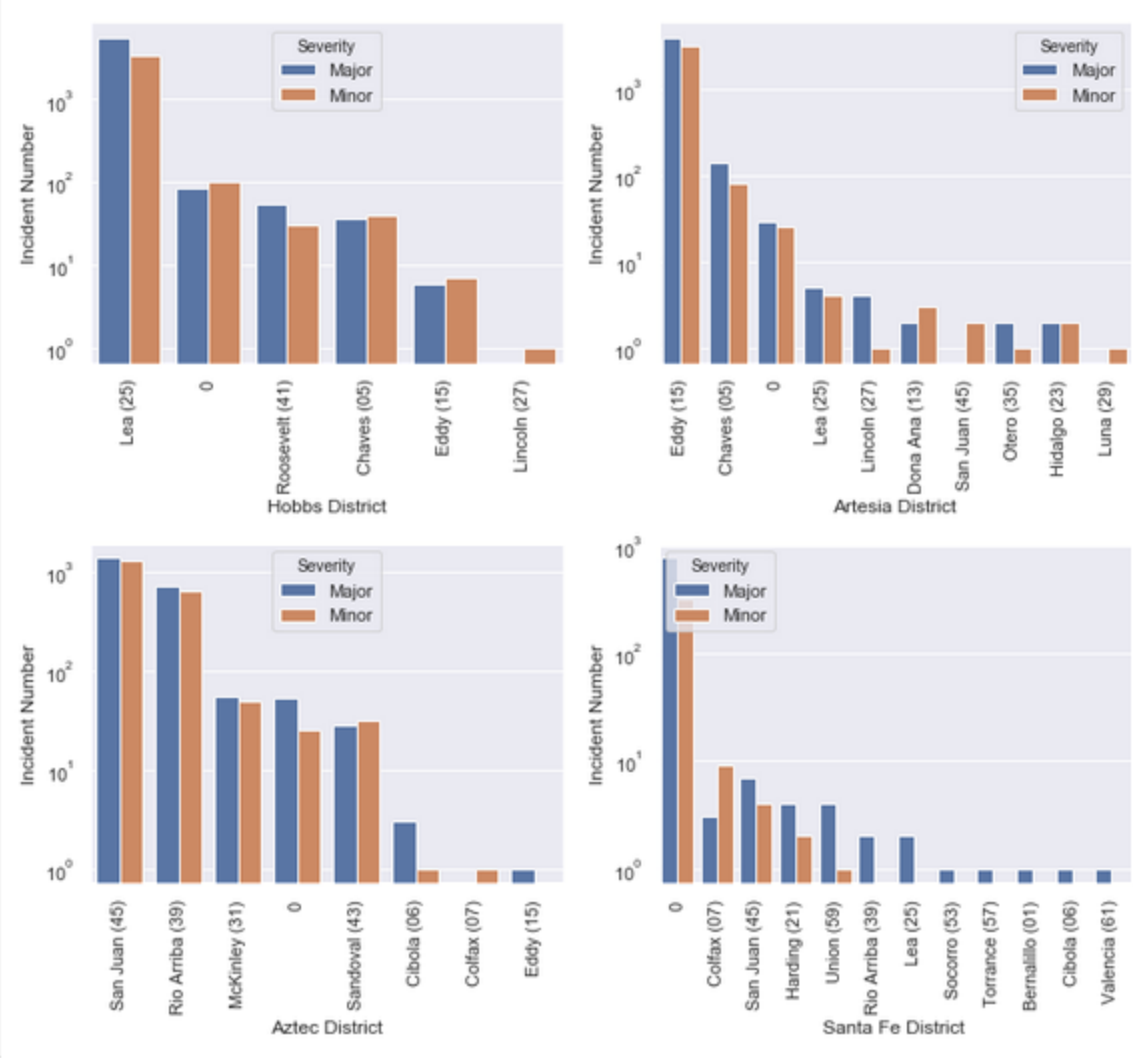

Figure 27 The number of major and minor incidents in each district and county

Figure 28 indicates a positive linear relationship between the total number of major and minor incidents from each operator. The top 10 operators which have the most major and minor incidents are marked out in the plot. COG OPERATING LLC and EOG Y RESOURCES,INC. are the first two operators cause most of incidents in New Mexico.

Figure 28 The relationship of major and minor incidents caused by each operator

Inferential Statistics

From above data exploratory analysis, that the incidents are caused by various reasons, and there is a positive linearly relationship between major and minor incidents. The major incidents have large spilled volume, and tend to impact waterway and groundwater. To further study whether the facility or well has larger probability to cause major incident when an incident occurred, bootstrap and *z* proportion test are applied.

Assumption:

- Null hypothesis: the probability of an incident is a major incident is equal in facilities and wells;
- Alternative hypothesis: the probability of an incident is a major incident in wells is larger than in facilities.

Assume significance level *alpha*= 5%.

In the bootstrap test, the differences between the number of major incidents and minor incidents in the sample is calculated; then conduct the bootstrap test 10000 times, reorganize the severity data, calculate the difference between the number of major incidents and minor incidents each time. The p value of the bootstrap difference is large than the sample difference is 0.5% assuming 5% significance. So we reject the null hypothesis that since p value is less than 5%, and accept the alternative hypothesis that the probability that an incident is a major incident in wells is larger than in facilities.

In the z proportion test, the probabilities of major incidents happen in facility and well are calculated, and the probability of major incidents of the sample is *p*, and then the standard error, *z* score, and *p* value are calculated. *p* value 0.2% which is less than 5%, so the null hypothesis is rejected and alternative hypothesis is accepted.

Conclusion

- The probability that an incident is a major incident in wells is larger than in facilities.
- There are more major incidents in the history since 1980s. The monthly number of incidents has increased to around 60/month in 2018.
- The volume spilled in minor incidents (below 50 barrels) is much less than the volume spilled in major incidents (up to 100,000 barrels).
- The correlation between monthly major and minor incidents indicates a positive linearly relationship between major and minor incidents.
- The top three incident types are produced water release, oil release and natural gas release which is corresponding to the top three spilled materials; The top three spill causes are equipment failure, corrosion and human error; and the top three spilled sources are tank (any), flow line - production, and pipeline (any). Fire, triethylene and sulphuric acid spill, and generator have 100% probability to cause major incident.
- The groundwater is more likely impacted by major incidents than minor incidents, and the waterway are 100% affected by major incidents.
- The most of major and minor incidents happens in county Lea in Hobbs, Eddy in Artesia, San Juan in Aztec, and 0 (missing county name probably) in Santa Fe.
- There is a positive linear relationship between the total number of major and minor incidents from each operator. COG OPERATING LLC and EOG Y RESOURCES,INC. are the first two operators cause most of incidents in New Mexico.

Chapter 5 Machine Learning

Machine Learning

From the EDA in the previous inferential statistics study, it's found that there are 4183 the missing values of incident severity out of 26454 rows.

In the following section, I'm going to apply supervised machine learning to get the missing data of incident severity, and regenerate the correlation plots to compare with the plots in previous section.

To predict the missing incident severity (major or minor) based on the known incident severity and its features, first I reset the table index to update the missing values once they are predicted. The incident features applied for machine learning are: whether the incident happened in a facility or well, the incident type, material spilled, volume spilled, spill cause, spill source, waterway affected or not, groundwater impacted or not.

Then convert the columns of data type from string categories to integer identifiers. From the seaborn pairplot plot in Figure 29 (partial plot, see the whole plot in Jupyter notebook), I can

assume that the features are independent and qualified to do machine learning analysis, as there is no clear correlation pattern between each other.

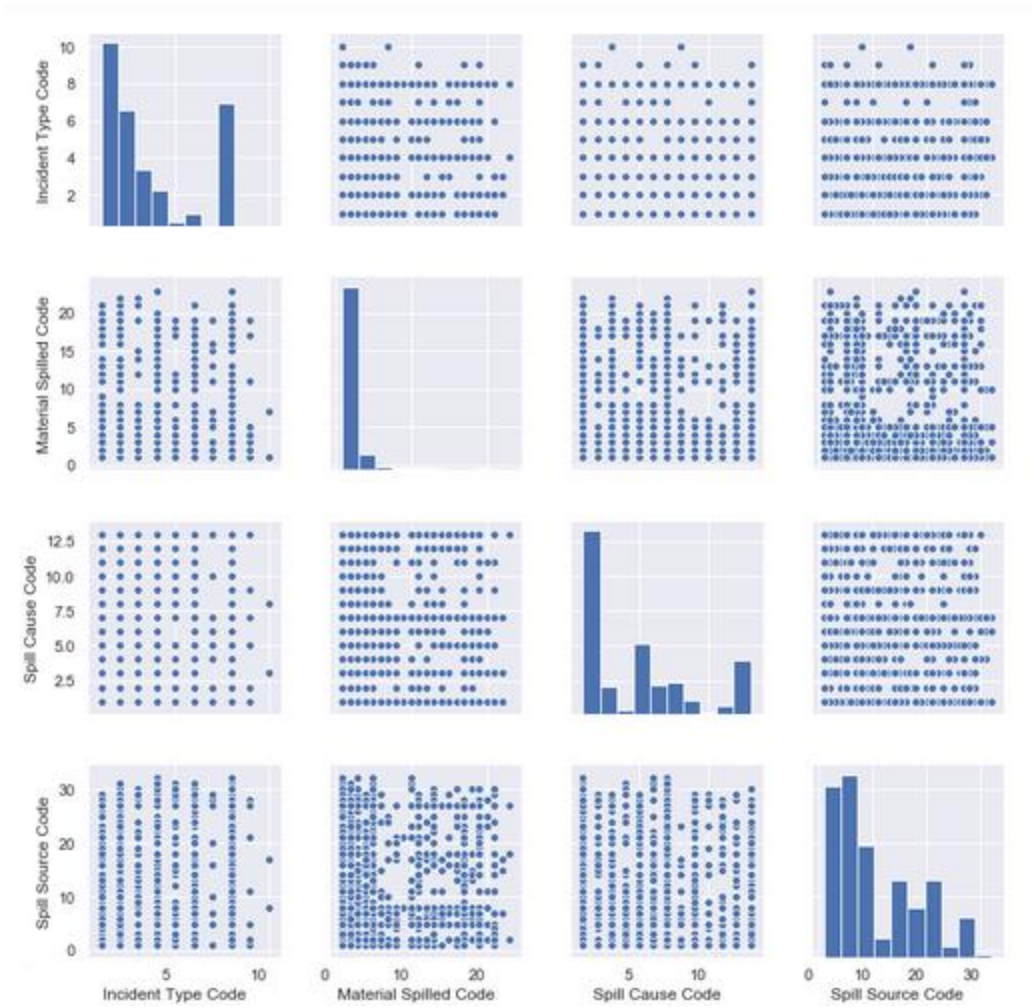

Figure 29 Pairplot of incident features

The data is split into two parts, that the known incident severity and features are used to train the model. I'm choosing the KNeighborsClassifier to conduct the supervised classification machine learning. *X_severity_given* contains the features and *y_severity_given* contains the results. `train_test_split` function is used to split the train the train and test data with *20%* test size, and use `GridSearchCV` to tune and cross validate the parameter `n_neighbors`.

After training the model, I get 86.6% accuracy score for train data and 79.9% accuracy score for test data, with the best `n_neighbors`=5. From the residual plot of the train test data in Figure 30, we can see the residual distribution uniformly in 1 and -1 around 0, which further validates the model.

Then apply the trained model to predict the missing incident severity data, and the initial spills table is updated, and the correlations are re-evaluated among the monthly number of incidents, volume spilled and recovered severity, incident type, material spilled, cause, source, waterway and groundwater impacted, location distribution, and operators.

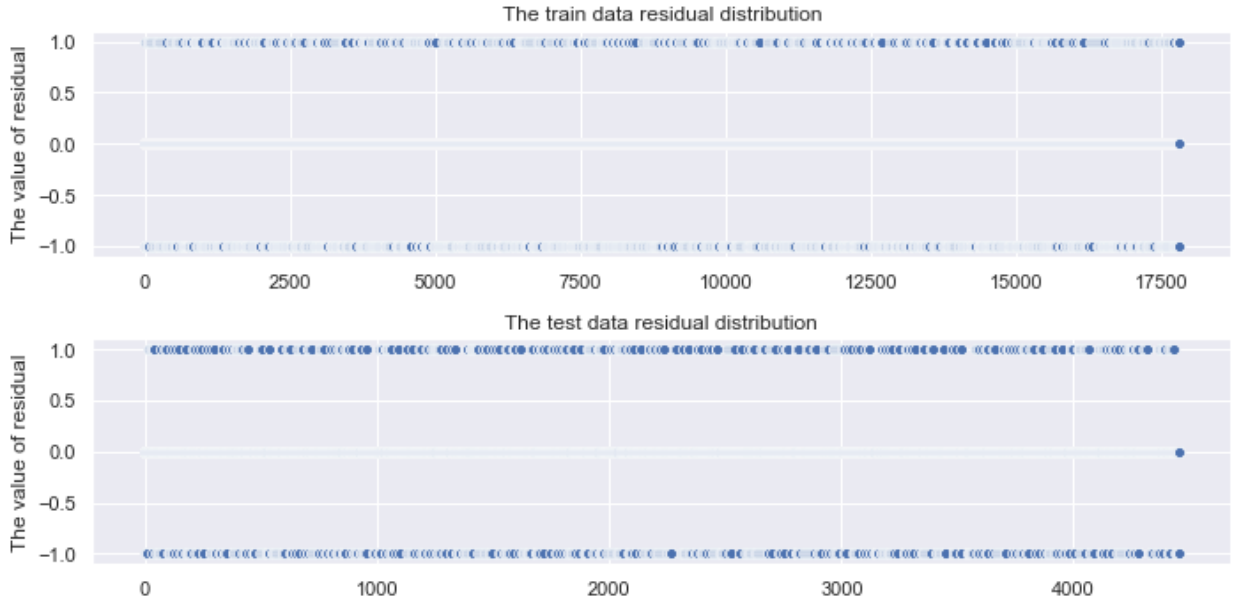

Figure 30 The residual plot of the train and test data

Update and compare EDA results

After updating, there monthly number of incidents increases by 10 to 20/month after 2008, and maximum number of major and minor incidents is more than 120/month in 2016 with more minor incidents than major incidents.

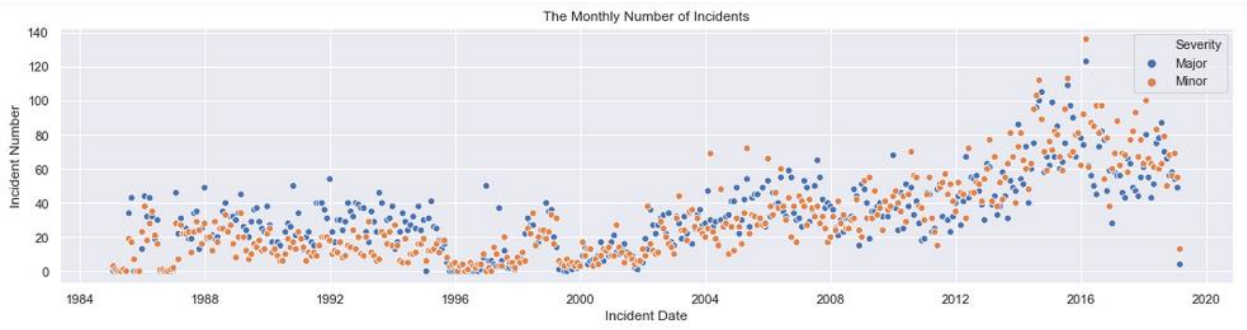

Figure 31 The updated monthly number of major and minor incidents changes from 1980s to 2019

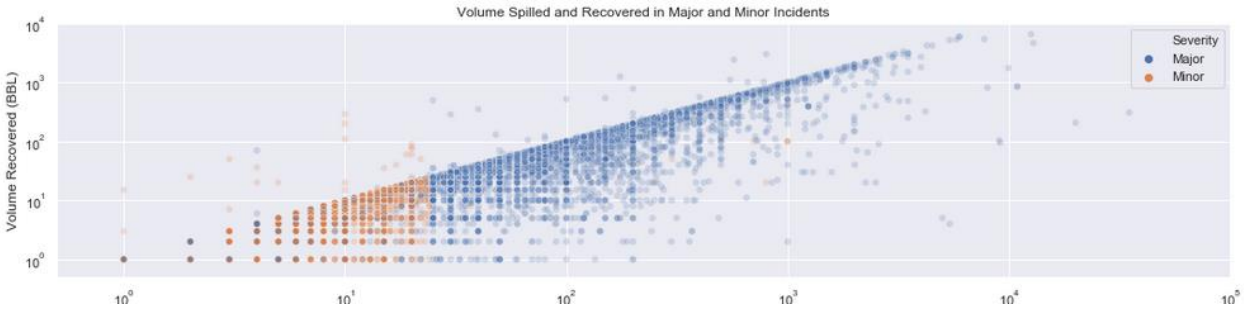

Figure 32 The updated spilled and recovered volumes in major and minor incidents

From the correlation between monthly major and minor incidents, we can see the maximum monthly number of minor incidents is more than major incidents after updating the table.

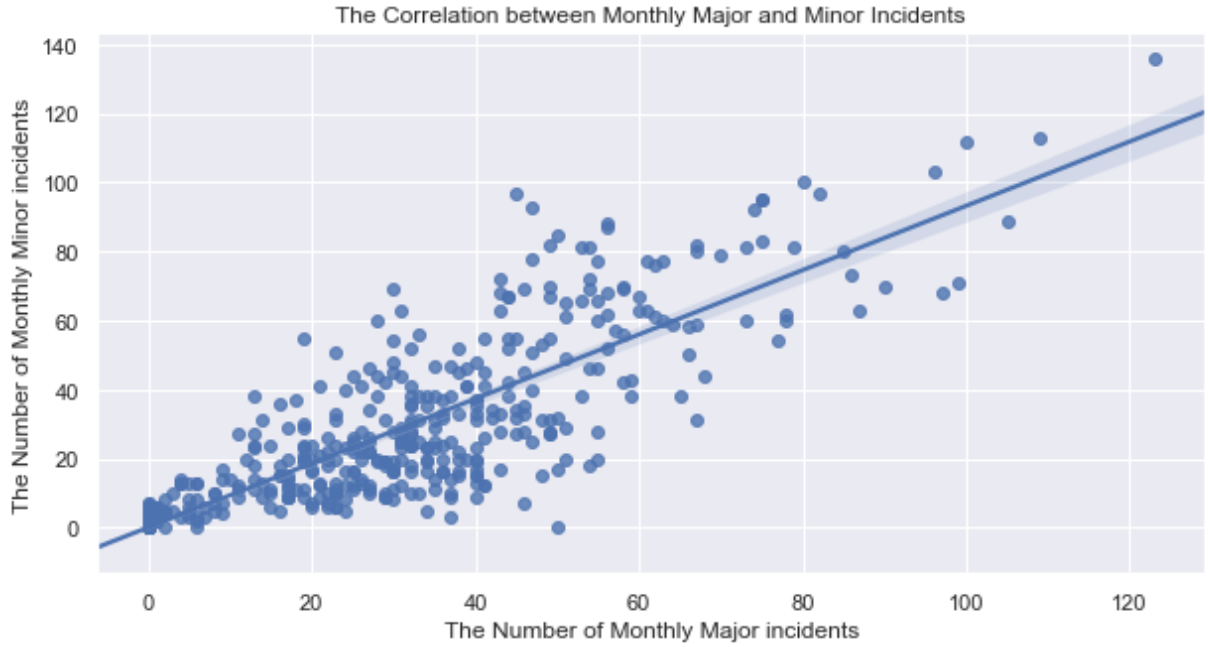

Figure 33 The updated correlation between monthly major and minor incidents

The following are the updated plots of the correlation between the number of major and minor incidents and incident type, material spilled, spill cause and spill source, and it's showing that the number of minor and major incidents of each category is adjusted slightly and reordered.

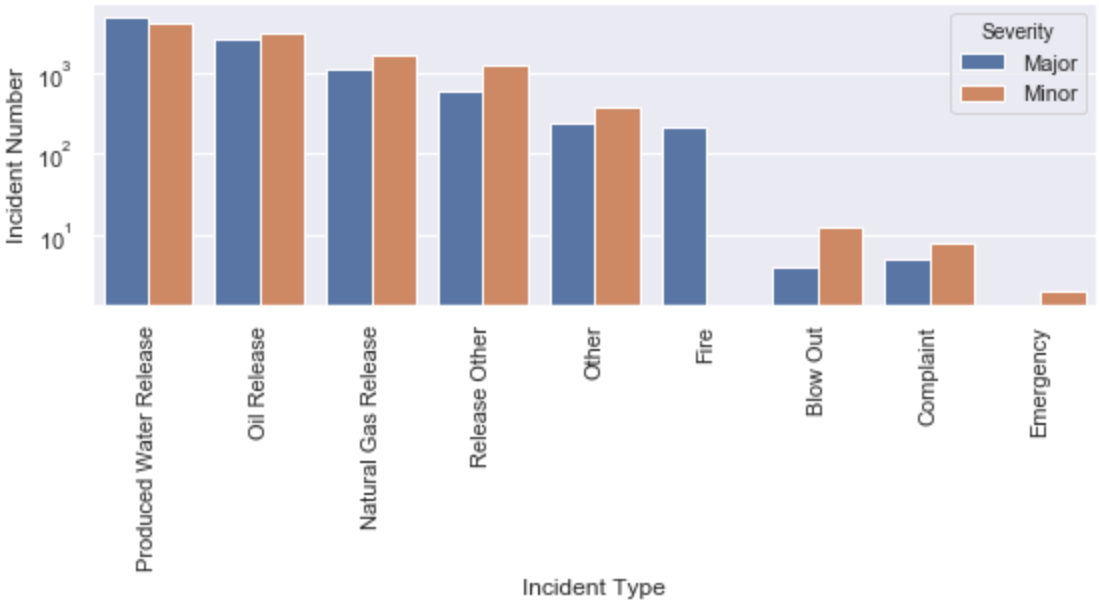

Figure 34 The updated number of major and minor incidents in various incident type categories

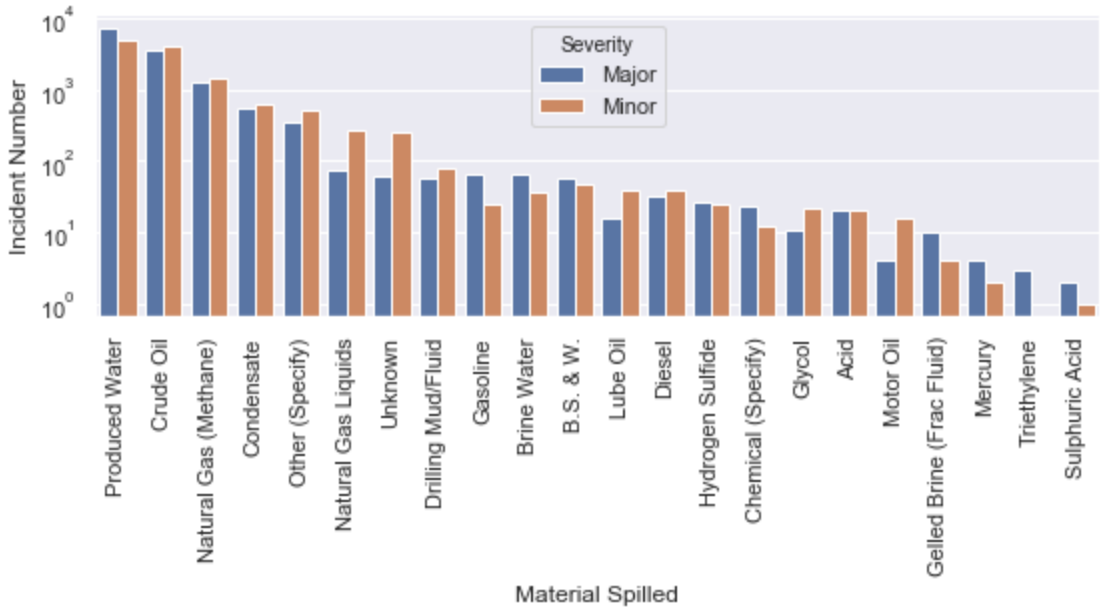

Figure 35 The updated number of major and minor incidents in various material spilled categories

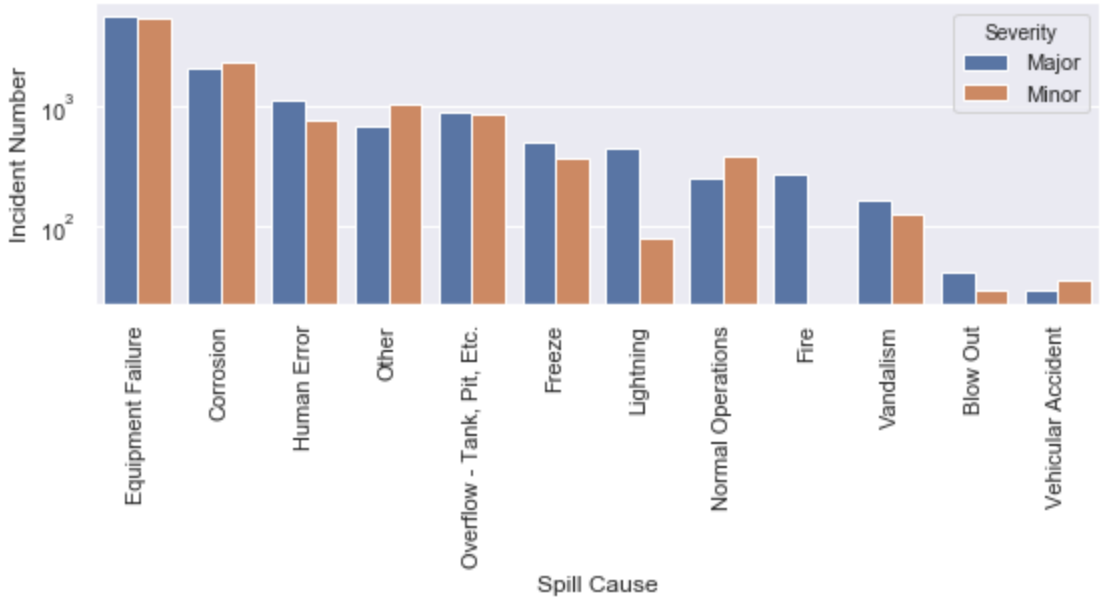

Figure 36 The updated number of major and minor incidents in various spill cause categories

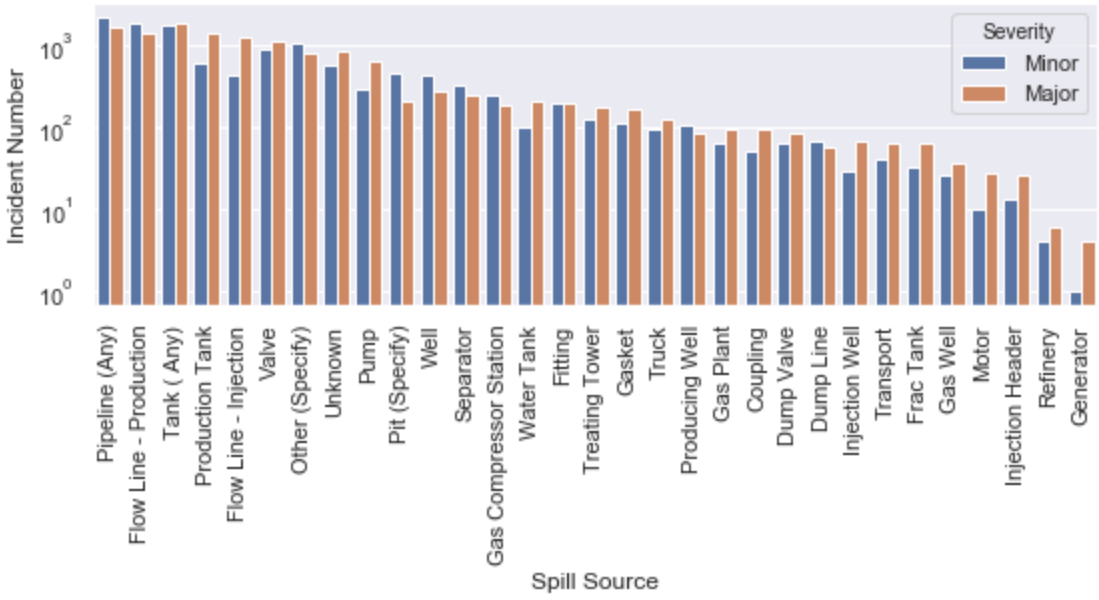

Figure 37 The updated number of major and minor incidents in various spill source categories

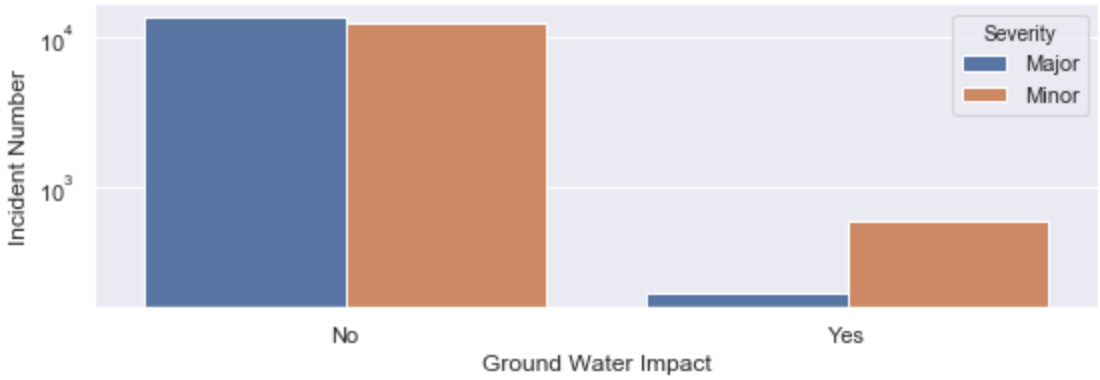

Figure 38 The updated number of major and minor incidents in ground water impact categories

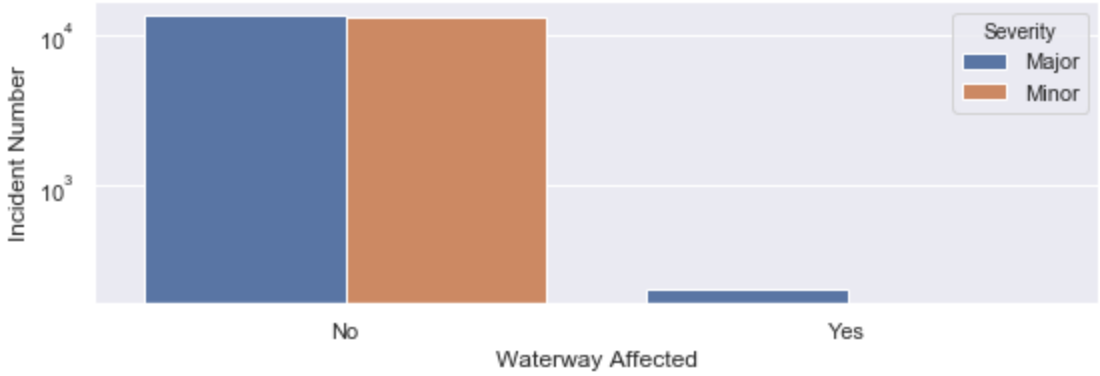

Figure 39 The updated number of major and minor incidents in waterway affected categories

The number of incidents in each district and county is adjusted slightly and reordered too.

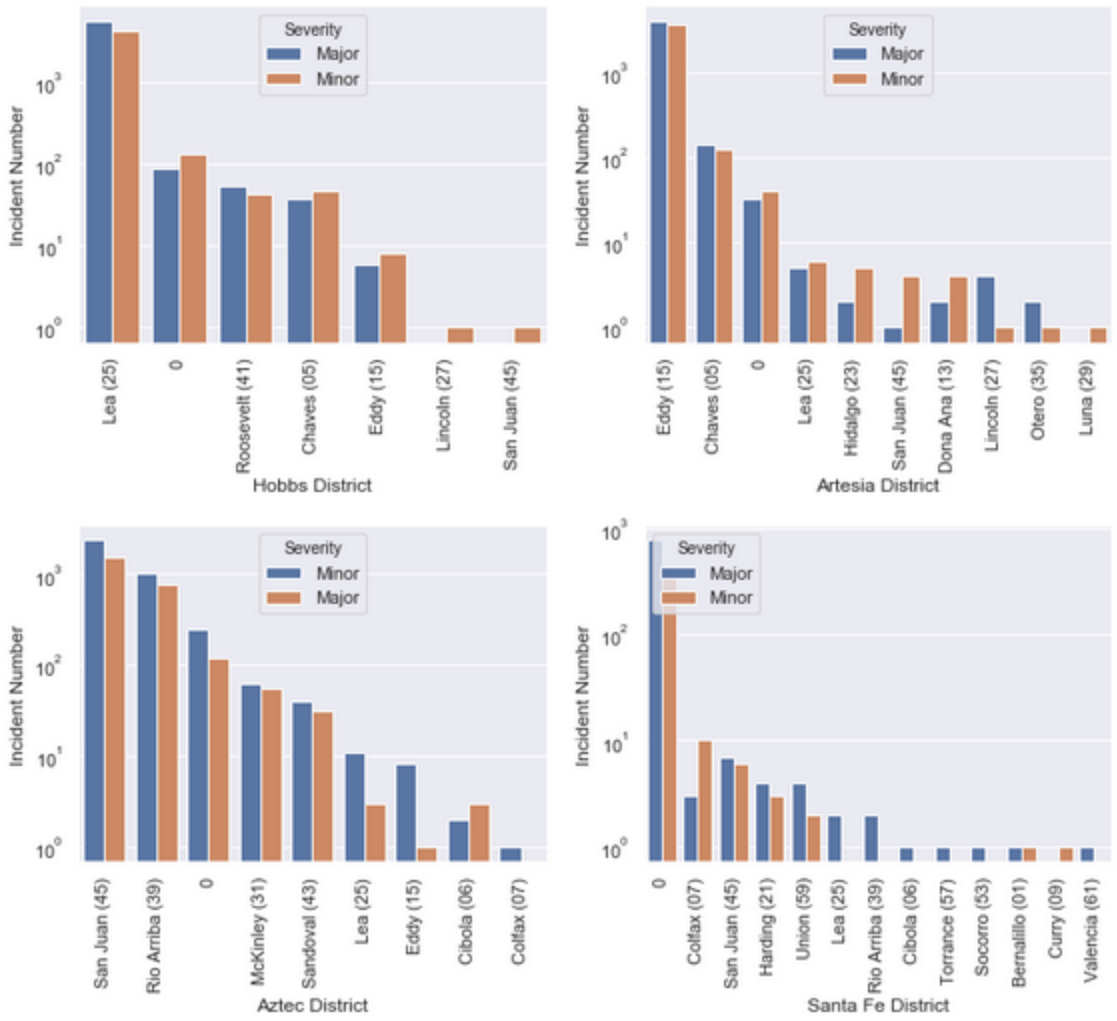

Figure 40 The updated number of major and minor incidents in each district and county

The total number of minor incidents caused by operators increases after updating the data, for example, COG OPERATING LLC, ENTERPRISE PRODUCTS OPERATING LLC, BP AMERICA PRODUCTION COMPANY

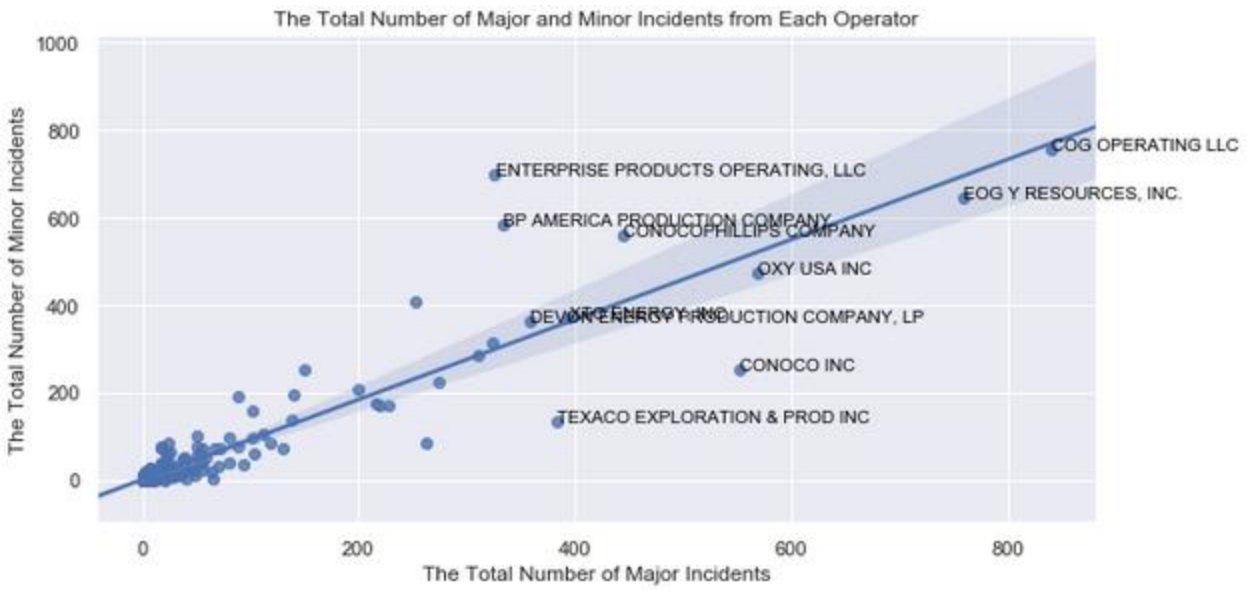

Figure 41 The updated relationship of major and minor incidents caused by each operator

Conclusion

- Assuming independent features of the incidents, the KNeighborsClassifier model is applied to conduct the supervised classification machine learning. After tuning and cross validating the parameter `n_neighbors`, I get 86.6% accuracy score for train data and 79.9% accuracy score for test data, with the best `n_neighbors`=5.
- From the plots of the train data residual distribution and the test data residual distribution, we can see the residual distribution uniformly in 1 and -1 around 0, which further validates the model.
- Then apply the trained model to predict the missing incident severity data, and update the initial spills table, and the correlations are re-evaluated among the monthly number of incidents, volume spilled and recovered, severity, incident type, material spilled, cause, source, waterway and groundwater impacted, location distribution, and operators.
- After updating, there monthly number of incidents increases by 10 to 20/month after 2008, and maximum number of major and minor incidents is more than 120/month in 2016 with more minor incidents than major incidents.
- The maximum monthly number of minor incidents is more than major incidents after updating the table.
- The correlation between the number of major and minor incidents and incident type, material spilled, spill cause and spill source is adjusted slightly and reordered on each category.
- The number of incidents in each district and county is adjusted slightly.
- The total number of minor incidents caused by operators increases after updating the data, for example, COG OPERATING LLC, ENTERPRISE PRODUCTS OPERATING LLC, BP AMERICA PRODUCTION COMPANY.

Appendix

1.New_Mexico_Well_History_and_Oil_Gas_Water_Production_and_Injection_Data_Analysis.ipynb

2.New_Mexico_Oil_and_Gas_Field_Spill_Incidents_Inferential_Statistics_Analysis.ipynb

3.New_Mexico_Oil_and_Gas_Field_Spill_Incidents_Machine_Learning.ipynb