eda

May 13, 2024

1 Load Data and Preprocessing

[27]:
```
import pandas as pd
import os
import plotly.graph_objects as go
import numpy as np
import matplotlib.pyplot as plt

# Set the default DPI
plt.rcParams['figure.dpi'] = 100

well_info = pd.read_csv('well-loc.tsv', sep='\t')

# Path to the sensor data directory
sensor_data_path = 'sensor-data'

# List all TSV files in the directory
sensor_data_files = [f for f in os.listdir(sensor_data_path) if f.endswith('.
 ↪tsv')]

# Sort the sensor_data_files list
sensor_data_files.sort(key= lambda x: int(x.split('.')[0]))

# Load and concatenate all sensor data files into one DataFrame
sensor_data_list = [pd.read_csv(os.path.join(sensor_data_path, file), sep='\t',
                                na_values="-9999") for file in␣
 ↪sensor_data_files]


# Reset the index of the well_loc DataFrame to Well, X, Y
well_info.rename(columns={' ': 'Well'}, inplace=True)

# Reset the index of the sensor data DataFrame to Depth, Porosity, Hydrate␣
 ↪Saturation
for idx, _ in enumerate(sensor_data_list):
        sensor_data_list[idx].columns = ['Depth', 'Porosity', 'Hydrate␣
 ↪Saturation']
```

```
print(well_info.head())  # Display the first few rows to verify it's loaded␣
 ↪correctly
print(sensor_data_list[0].head())  # Display the first few rows to verify it's␣
 ↪loaded correctly
```

```
  Well      X      Y
0  w01  34500  45000
1  w02  36000  45050
2  w03  37050  45020
3  w04  37880  46000
4  w05  35000  46030
       Depth  Porosity  Hydrate Saturation
0  1814.9316       NaN                 NaN
1  1815.0840       NaN                 NaN
2  1815.2364       NaN                 NaN
3  1815.3888       NaN                 NaN
4  1815.5412       NaN                 NaN
```

[28]:
```
# Show the summary of the dataset
import pandas as pd
import os
import matplotlib.pyplot as plt
import seaborn as sns  # For enhanced visualizations

# Assuming previous code has already been run and data is loaded

# 1. Descriptive Statistics
print("Descriptive Statistics for Well Information:")
print(well_info.describe())
print("\nDescriptive Statistics for Sensor Data:")
for idx, sensor_data in enumerate(sensor_data_list):
    print(f"\nSensor Data File {idx + 1}:")
    print(sensor_data.describe())

# 2. Data Quality Checks
print("\nChecking for missing values in Well Information:")
print(well_info.isnull().sum())
print("\nChecking for missing values in Sensor Data:")
for idx, sensor_data in enumerate(sensor_data_list):
    print(f"\nSensor Data File {idx + 1} Missing Values:")
    print(sensor_data.isnull().sum())

print("\nChecking for duplicate rows in Well Information:")
print(well_info.duplicated().sum())
for idx, sensor_data in enumerate(sensor_data_list):
    print(f"\nSensor Data File {idx + 1} Duplicate Rows:")
```

```
    print(sensor_data.duplicated().sum())

# 3. Correlation Analysis
# Assuming sensor data has common columns that can be concatenated for␣
 ↪correlation analysis
if len(sensor_data_list) > 0:
    combined_sensor_data = pd.concat(sensor_data_list, ignore_index=True)
    plt.figure(figsize=(8, 6))
    sns.heatmap(combined_sensor_data.corr(), annot=True, fmt=".2f",␣
 ↪cmap='coolwarm')
    plt.title('Correlation Matrix of Combined Sensor Data')
    plt.show()

# Note: Ensure all plots are properly displayed
plt.show()
```

```
Descriptive Statistics for Well Information:
                  X             Y
count     14.000000     14.000000
mean   35729.285714  47391.428571
std     1281.072053   1859.341724
min    34000.000000  45000.000000
25%    34670.000000  46007.500000
50%    35750.000000  47215.000000
75%    36150.000000  49107.500000
max    38000.000000  50000.000000

Descriptive Statistics for Sensor Data:

Sensor Data File 1:
             Depth     Porosity  Hydrate Saturation
count  1923.00000  1801.000000         1609.000000
mean   1961.38800     0.488843            0.001296
std      84.62263     0.044011            0.006527
min    1814.93160     0.391800            0.000000
25%    1888.15980     0.457700            0.000000
50%    1961.38800     0.488900            0.000000
75%    2034.61620     0.517700            0.000000
max    2107.84440     0.748600            0.088900

Sensor Data File 2:
              Depth     Porosity  Hydrate Saturation
count  1787.000000  1655.000000         1408.000000
mean   1604.300000     0.508038            0.172149
std      51.600678     0.085347            0.242259
min    1515.000000     0.372943           -0.095347
25%    1559.650000     0.452836            0.017835
```

```
50%    1604.300000     0.487566            0.073318
75%    1648.950000     0.541292            0.216235
max    1693.600000     0.900575            0.889902

Sensor Data File 3:
             Depth     Porosity  Hydrate Saturation
count  1654.000000  1487.000000         1264.000000
mean   1782.650000     0.456307            0.004407
std      47.761299     0.103604            0.011378
min    1700.000000     0.299100            0.000000
25%    1741.325000     0.388400            0.000000
50%    1782.650000     0.430400            0.000000
75%    1823.975000     0.488900            0.000000
max    1865.300000     0.926100            0.074600

Sensor Data File 4:
             Depth     Porosity  Hydrate Saturation
count  1829.000000  1690.000000         1474.000000
mean   1636.400000     0.510441            0.173410
std      52.813114     0.099198            0.174286
min    1545.000000     0.292291           -0.673800
25%    1590.700000     0.429363            0.000000
50%    1636.400000     0.509030            0.185950
75%    1682.100000     0.585026            0.317739
max    1727.800000     0.827095            0.651820

Sensor Data File 5:
           Depth     Porosity  Hydrate Saturation
count  2048.0000  1835.000000         1599.000000
mean   1658.3500     0.467805           -0.020579
std      59.1351     0.094342            0.076715
min    1556.0000     0.276610           -0.662127
25%    1607.1750     0.396600           -0.019038
50%    1658.3500     0.457644            0.000000
75%    1709.5250     0.532162            0.000000
max    1760.7000     0.820117            0.077287

Sensor Data File 6:
             Depth     Porosity  Hydrate Saturation
count  1901.000000  1780.000000         1480.000000
mean   1836.000000     0.462810           -0.024724
std      54.891575     0.091795            0.047725
min    1741.000000     0.309091           -0.179033
25%    1788.500000     0.401818           -0.058011
50%    1836.000000     0.454413           -0.019236
75%    1883.500000     0.523589            0.004879
max    1931.000000     0.903613            0.105548
```

Sensor Data File 7:

	Depth	Porosity	Hydrate Saturation
count	942.000000	756.000000	609.000000
mean	1836.648600	0.530543	0.086495
std	41.464424	0.043353	0.066015
min	1764.944400	0.407900	0.000000
25%	1800.796500	0.499200	0.043800
50%	1836.648600	0.532600	0.082300
75%	1872.500700	0.557225	0.116800
max	1908.352800	0.645800	0.346700

Sensor Data File 8:

	Depth	Porosity	Hydrate Saturation
count	2197.000000	1772.000000	1624.000000
mean	1753.819200	0.471518	-0.103858
std	96.677011	0.062733	0.332528
min	1586.484000	0.000100	-4.469665
25%	1670.151600	0.432000	-0.067706
50%	1753.819200	0.471600	0.000000
75%	1837.486800	0.507100	0.000000
max	1921.154400	0.676700	0.155016

Sensor Data File 9:

	Depth	Porosity	Hydrate Saturation
count	1430.000000	1280.000000	1118.000000
mean	1838.782200	0.495177	0.178540
std	62.933543	0.080942	0.135718
min	1729.892400	0.260700	0.000000
25%	1784.337300	0.436475	0.072500
50%	1838.782200	0.487500	0.157700
75%	1893.227100	0.535400	0.261650
max	1947.672000	0.923200	0.702800

Sensor Data File 10:

	Depth	Porosity	Hydrate Saturation
count	674.000000	567.000000	420.000000
mean	1771.269000	0.523384	0.116739
std	29.674006	0.063430	0.108913
min	1719.986400	0.374900	0.000000
25%	1745.627700	0.481850	0.036050
50%	1771.269000	0.518600	0.088450
75%	1796.910300	0.551700	0.159675
max	1822.551600	0.767700	0.434500

Sensor Data File 11:

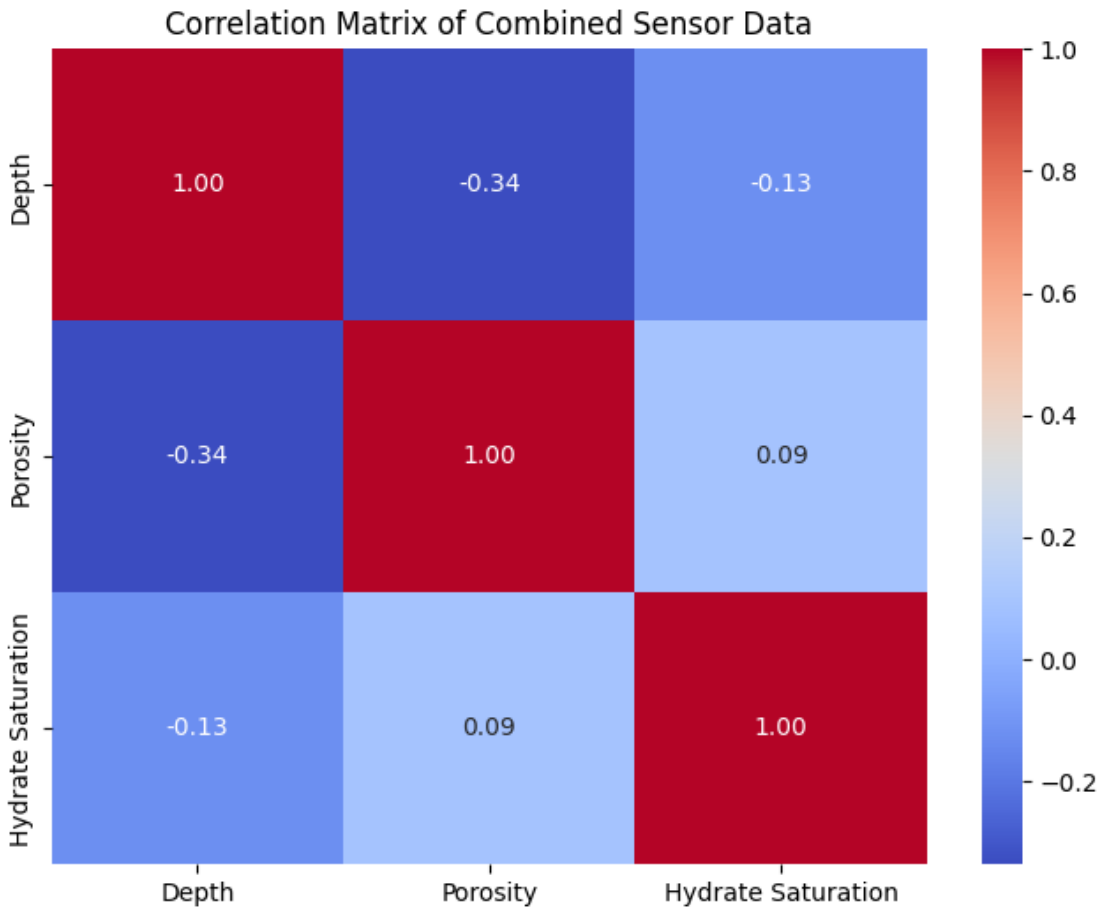
	Depth	Porosity	Hydrate Saturation
count	2121.000000	1992.000000	1390.000000
mean	1844.000000	0.446785	-0.012340

```
std      61.242428     0.079919            0.047527
min    1738.000000     0.287744           -0.211102
25%    1791.000000     0.377917           -0.035446
50%    1844.000000     0.448905            0.000000
75%    1897.000000     0.500168            0.002847
max    1950.000000     0.764798            0.216866

Sensor Data File 12:
             Depth    Porosity  Hydrate Saturation
count   672.000000  550.000000          402.000000
mean   1781.022600    0.516368            0.001983
std      29.586018    0.060835            0.006803
min    1729.892400    0.401800            0.000000
25%    1755.457500    0.457750            0.000000
50%    1781.022600    0.521300            0.000000
75%    1806.587700    0.546975            0.000100
max    1832.152800    0.695500            0.069300

Sensor Data File 13:
             Depth     Porosity  Hydrate Saturation
count  1846.000000  1655.000000         1409.000000
mean   1602.250000     0.526135           -0.022507
std      53.303862     0.091029            0.110989
min    1510.000000     0.392442           -0.583767
25%    1556.125000     0.446003           -0.038089
50%    1602.250000     0.509349           -0.003593
75%    1648.375000     0.598565            0.013295
max    1694.500000     0.886159            0.620332

Sensor Data File 14:
            Depth     Porosity  Hydrate Saturation
count  1879.00000  1733.000000         1478.000000
mean   1913.07720     0.491331            0.008666
std      82.68689     0.070255            0.015802
min    1769.97360     0.363200            0.000000
25%    1841.52540     0.438600            0.000000
50%    1913.07720     0.475700            0.000000
75%    1984.62900     0.547100            0.008575
max    2056.18080     0.740200            0.108300

Checking for missing values in Well Information:
Well    0
X       0
Y       0
dtype: int64

Checking for missing values in Sensor Data:
```

```
Sensor Data File 1 Missing Values:
Depth                   0
Porosity              122
Hydrate Saturation    314
dtype: int64

Sensor Data File 2 Missing Values:
Depth                   0
Porosity              132
Hydrate Saturation    379
dtype: int64

Sensor Data File 3 Missing Values:
Depth                   0
Porosity              167
Hydrate Saturation    390
dtype: int64

Sensor Data File 4 Missing Values:
Depth                   0
Porosity              139
Hydrate Saturation    355
dtype: int64

Sensor Data File 5 Missing Values:
Depth                   0
Porosity              213
Hydrate Saturation    449
dtype: int64

Sensor Data File 6 Missing Values:
Depth                   0
Porosity              121
Hydrate Saturation    421
dtype: int64

Sensor Data File 7 Missing Values:
Depth                   0
Porosity              186
Hydrate Saturation    333
dtype: int64

Sensor Data File 8 Missing Values:
Depth                   0
Porosity              425
Hydrate Saturation    573
dtype: int64
```

```
Sensor Data File 9 Missing Values:
Depth                   0
Porosity              150
Hydrate Saturation    312
dtype: int64

Sensor Data File 10 Missing Values:
Depth                   0
Porosity              107
Hydrate Saturation    254
dtype: int64

Sensor Data File 11 Missing Values:
Depth                   0
Porosity              129
Hydrate Saturation    731
dtype: int64

Sensor Data File 12 Missing Values:
Depth                   0
Porosity              122
Hydrate Saturation    270
dtype: int64

Sensor Data File 13 Missing Values:
Depth                   0
Porosity              191
Hydrate Saturation    437
dtype: int64

Sensor Data File 14 Missing Values:
Depth                   0
Porosity              146
Hydrate Saturation    401
dtype: int64

Checking for duplicate rows in Well Information:
0

Sensor Data File 1 Duplicate Rows:
0

Sensor Data File 2 Duplicate Rows:
0

Sensor Data File 3 Duplicate Rows:
0
```

```
Sensor Data File 4 Duplicate Rows:
0

Sensor Data File 5 Duplicate Rows:
0

Sensor Data File 6 Duplicate Rows:
0

Sensor Data File 7 Duplicate Rows:
0

Sensor Data File 8 Duplicate Rows:
0

Sensor Data File 9 Duplicate Rows:
0

Sensor Data File 10 Duplicate Rows:
0

Sensor Data File 11 Duplicate Rows:
0

Sensor Data File 12 Duplicate Rows:
0

Sensor Data File 13 Duplicate Rows:
0

Sensor Data File 14 Duplicate Rows:
0
```



[29]:
```
negative_data = []
for idx, sensor_data in enumerate(sensor_data_list):
    condition = (sensor_data['Hydrate Saturation'] < 0)
    negative_data.append(len(sensor_data[condition]) / len(sensor_data))

# Plot the number of negative data points
fig = go.Figure(data=[go.Bar(x=sensor_data_files, y=negative_data)])
fig.update_layout(title_text='Percentage of negative Hydrate Saturation points␣
 ↪in each sensor data file')
fig.show()

# Claculate total data point amount in the sensor data
total_data_points = 0
for idx, sensor_data in enumerate(sensor_data_list):
    total_data_points += len(sensor_data)

total_data_points
```

[29]: 22903

[30]:
```
negative_data = []
for idx, sensor_data in enumerate(sensor_data_list):
    condition = (sensor_data['Porosity'] < 0)
    negative_data.append(len(sensor_data[condition]) / len(sensor_data))

# Plot the number of negative data points
fig = go.Figure(data=[go.Bar(x=sensor_data_files, y=negative_data)])
fig.update_layout(title_text='Percentage of negative Porosity data points in␣
 ↪each sensor data file')
fig.show()
```

[31]:
```
# Test how many data points left if we drop the negative data points
total_data_points_after_drop = 0
for idx, sensor_data in enumerate(sensor_data_list):
    condition = (sensor_data['Hydrate Saturation'] >= 0)
    sensor_data = sensor_data[condition]
    total_data_points_after_drop += len(sensor_data)
total_data_points_after_drop
```

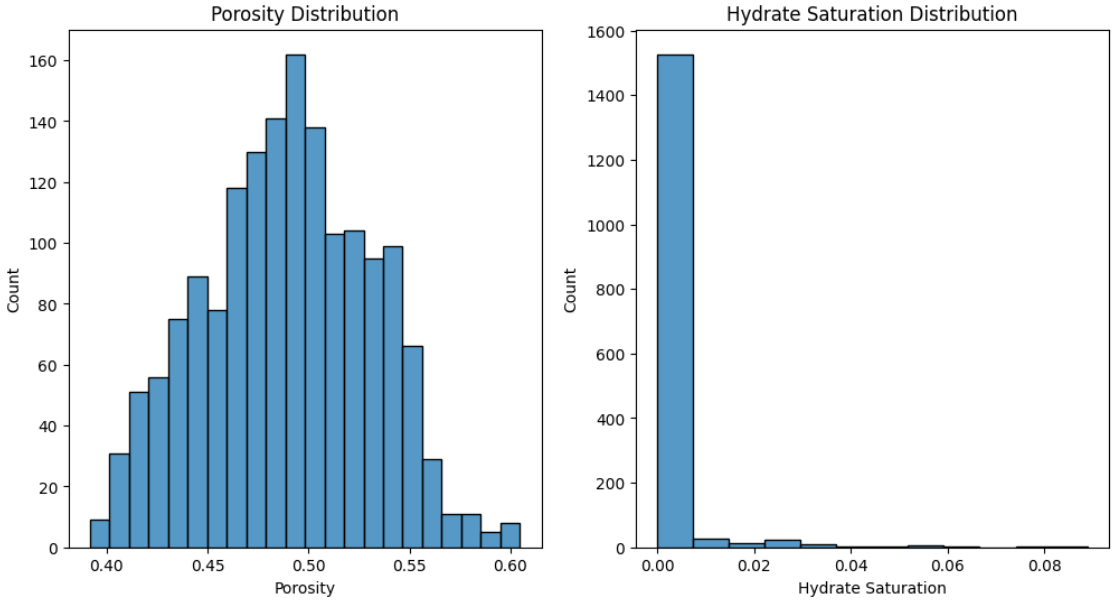
[31]: 13449

[32]:
```
# Check the not a NaN count of the other columns in the sensor data which has␣
 ↪negative hydrate saturation

# Count rows contains negative hydrate saturation and no NaN in total
not_nan_count = 0
for idx, sensor_data in enumerate(sensor_data_list):
    condition = sensor_data['Hydrate Saturation'] < 0
    not_nan_count += len(sensor_data[condition].dropna())
print(not_nan_count)
```

```
3821
```

[33]:
```
for idx, sensor_data in enumerate(sensor_data_list):
    sensor_data = sensor_data[sensor_data['Porosity'] <= 1.0]
    sensor_data = sensor_data[sensor_data['Hydrate Saturation'] <= 1.0]
    sensor_data_list[idx] = sensor_data

# Drop the rows containing values smaller than 0.0 except for the Depth column
for idx, sensor_data in enumerate(sensor_data_list):
    sensor_data = sensor_data[sensor_data['Porosity'] >= 0]
    sensor_data = sensor_data[sensor_data['Hydrate Saturation'] >= 0]
    sensor_data_list[idx] = sensor_data
```

[34]:
```
# Plot the distribution of the cleaned data
fig, ax = plt.subplots(1, 2, figsize=(12, 6))
sns.histplot(sensor_data_list[0]['Porosity'], kde=False, ax=ax[0])
ax[0].set_title('Porosity Distribution')
sns.histplot(sensor_data_list[0]['Hydrate Saturation'], kde=False, ax=ax[1])
ax[1].set_title('Hydrate Saturation Distribution')
plt.show()
```



[ ]: