Пермский филиал федерального государственного автономного образовательного учреждения высшего образования   
«Национальный исследовательский университет   
«Высшая школа экономики»

*Факультет экономики, менеджмента и бизнес-информатики*

Чепоков Елизар, Ануфриев Павел, Дроздов Андрей, Кроливецкая Арина

**ОПРЕДЕЛЕНИЕ ВЛИЯНИЯ СПОСОБА И РЕЖИМА ЭКСПЛУАТАЦИИ НА ВИДЫ ДАВЛЕНИЯ В СКВАЖИНАХ**

*Проектная работа*

студента образовательной программы «Программная инженерия»

по направлению подготовки *09.03.04 Программная инженерия*

Руководитель:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

А.В. Кычкин

Пермь, 2021 год

**Оглавление**

[**Введение** 3](#_Toc66722758)

[**Постановка задачи** 4](#_Toc66722759)

[**Выборка данных** 5](#_Toc66722760)

[**Очистка данных** 8](#_Toc66722761)

[**Трансформация** 10](#_Toc66722762)

[**Анализ** 14](#_Toc66722763)

[**Заключение** 17](#_Toc66722764)

[**Приложение А. Листинг скрипта для выборки** 18](#_Toc66722765)

[**Приложение Б. Листинг скрипта для экспорта** 20](#_Toc66722766)

# **Введение**

**Данные о проекте:** Технологический процесс добычи нефти. В работе находятся несколько скважин, каждая из которых оснащена насосным оборудованием. Оборудование может работать в различных режимах. Основные технологические параметры о работе оборудования измеряются с дискретностью 1 сутки (раз в 24 часа). В связи с нестабильной связью, особенностями сбора данных (ручной ввод) имеются значительные пропуски в данных.

**Цель работы:** определить влияние способа и режима эксплуатации на виды давления в скважинах.

**Исходные данные:** файл Excel с параметрами скважин, результатами измерений, режимами работы оборудования. Каждая вкладка файла содержит данные по 1 месяцу работы всего массива скважин.

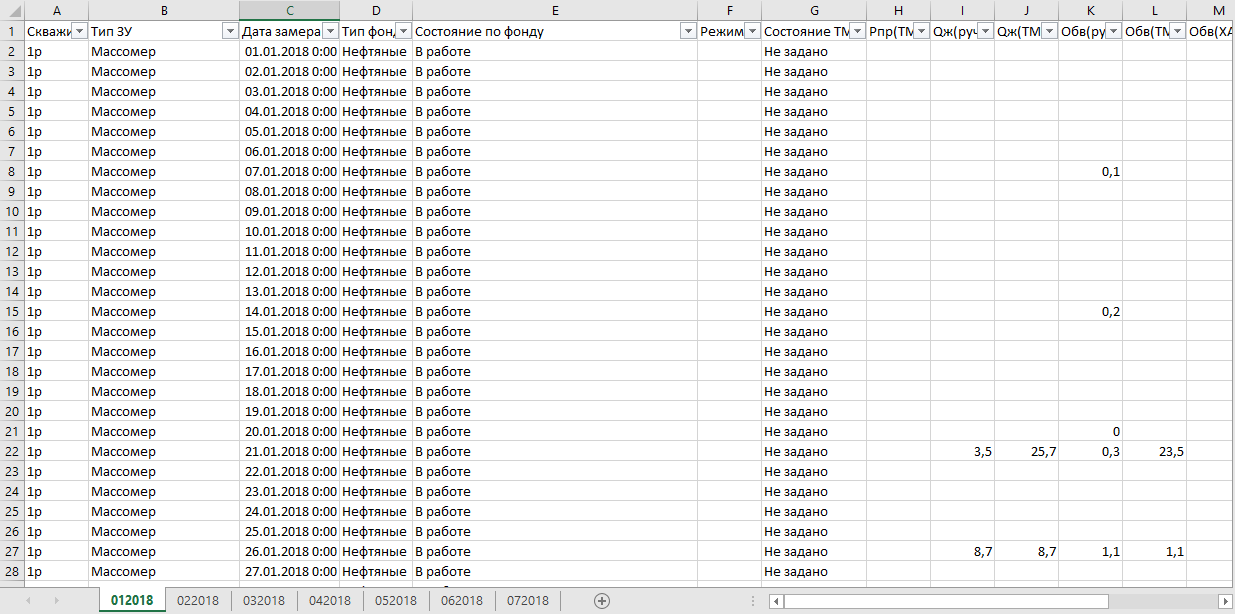
**Подготовительный этап:**

1. Обработка данных для исследований.py - реализует парсинг данных из исходного файла и формирование датафрейма. Формируется новый датафрейм с данными только по одной выбранной скважине. Также в файле реализуется начальная очистка данных (удаление столбцов с пустыми значениями во всех строках) для выбранной скважины.
2. Сохранение\_данных\_в\_CSV\_для\_Influx.py - реализует парсинг данных из исходного файла, подготовку датафрейма для экспорта и сам экспорт в csv файл. Далее можно использоваться скрипт CSV2Influx.py для экспорта в БД InfluxDB.

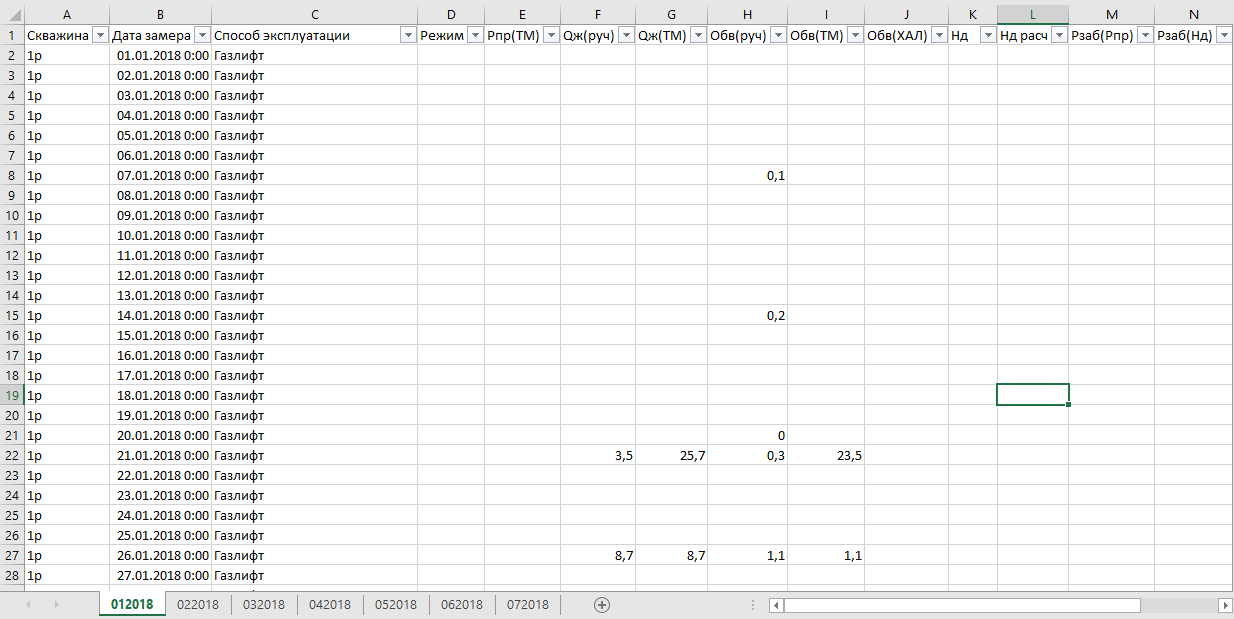
# **Постановка задачи**

1. **Осуществить выборку данных:**
   1. Исследовать файл «Данные для исследований.xlsx» с информацией о скважинах за 7 месяцев;
   2. Выбрать данные в соответствии с номером проекта (100 скважин на проект);
   3. Выполнить фильтрацию данных в соответствии с задачами проекта (выбрать столбцы: Скважина, Дата замера, Способ эксплуатации, Режим, и все столбцы с давлениями);
2. **Выполнить очистку данных:**
   1. Заполнить пропуски;
   2. Удалить аномалии;
3. **Провести трансформацию данных:**
   1. Сгруппировать данные по скважинам;
   2. Выполнить сглаживание в окне скользящего среднего;
   3. Значения давлений привести при помощи нормализации (деления на мах. значение) к диапазону: от 0 до 1;
4. **Анализ данных:**
   1. Минимальные, максимальные, средние значения давлений по скважинам, сгруппировать по месяцам и за 7 месяцев;
   2. Построить линейную регрессию способа и режима добычи на виды давления по скважинам;
5. **Визуализировать полученные данные.**

# **Выборка данных**

Выборка данных производилась из файла “Данные для исследований.xlsx” в котором хранится информация о скважинах за 7 месяцев. В соответствии с 10 номером нашей группы мы рассматривали скважины, с порядковыми номерами с 1 по 100. На рисунке 1.1 представлен исходный файл до фильтрации данных.

***Рисунок 1.1 – файл «Данные для исследований.xlsx»***

Мы провели фильтрацию данных в соответствии с задачами проекта и выбрали столбцы в соответствии с поставленной задачей: Скважина, Дата замера, Способ эксплуатации, Режим, и все столбцы с давлениями. Результат представлен на рисунке 1.2.

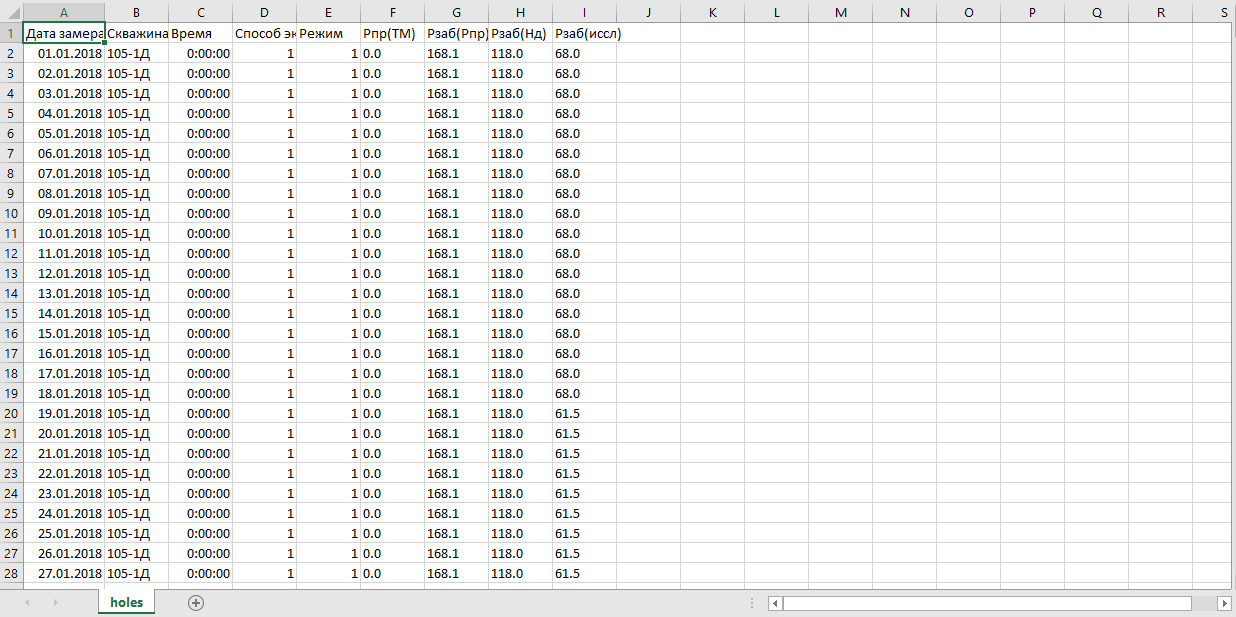
***Рисунок 1.2 – файл «Data.xlsx»***

Данные в файле сгруппированы по скважинам и располагаются в хронологическом порядке выполнения замеров. Для выборки данных из исходного файла был отредактирован представленный в качестве примера преподавателем код на python (см. листинг программы в Приложении А). Так как программное обеспечение InfluxDB не позволяет загрузку отсутствующих значений, файл “Data.xlsx” был дополнен временными данными.

Среди предложенных вариантов заполнения были выявлены следующие:

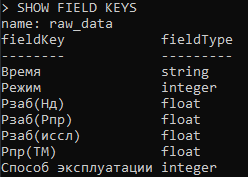
1. Заполнение последним доступным значением;
2. Заполнение следующим доступным значением;
3. Заполнение нулями;
4. Интерполяция (линейная, полином, сплайн).

Из данных вариантов было выбрано заполнение следующими или последними доступными значениями, так как только данные варианты заполнения подходят в нашем случае. Вариант заполнения интерполяцией был отклонен из-за неэффективности для большого массива данных с большим количеством пропущенных значений. Заполнение нулями невозможно так как из-за большого количества пропущенных значений данные пустые значения серьезно повлияют на все возможные статистические метрики.

Результат заполнения представлен на рисунке 1.3.

***Рисунок 1.3 – файл «holes.csv»***

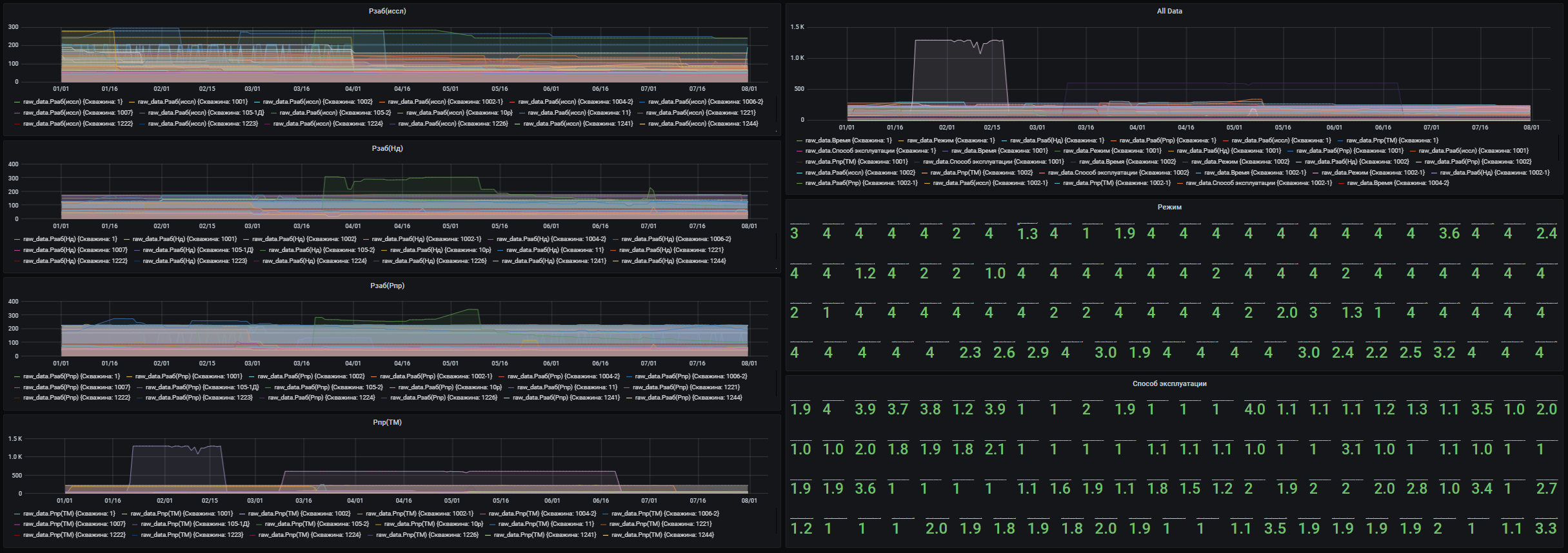
Заполнение осуществлялось функцией python, fillna, где в зависимости от расположения пропущенного значения выбирался вариант заполнения “bfill” или “ffill”, после чего данные записывались в новый документ, который будет загружаться в базу данных influx.

Для полученных “сырых” значений был проведен анализ с использованием веб-приложения Grafana, для чего было дополнительно отфильтрованы и загружены в базу данных значения из файла “holes.csv”. Так же были преобразованы данные с помощью языка python (см. Приложение Б). Данные были загружены с типами столбцов представленными на рисунке 1.4.

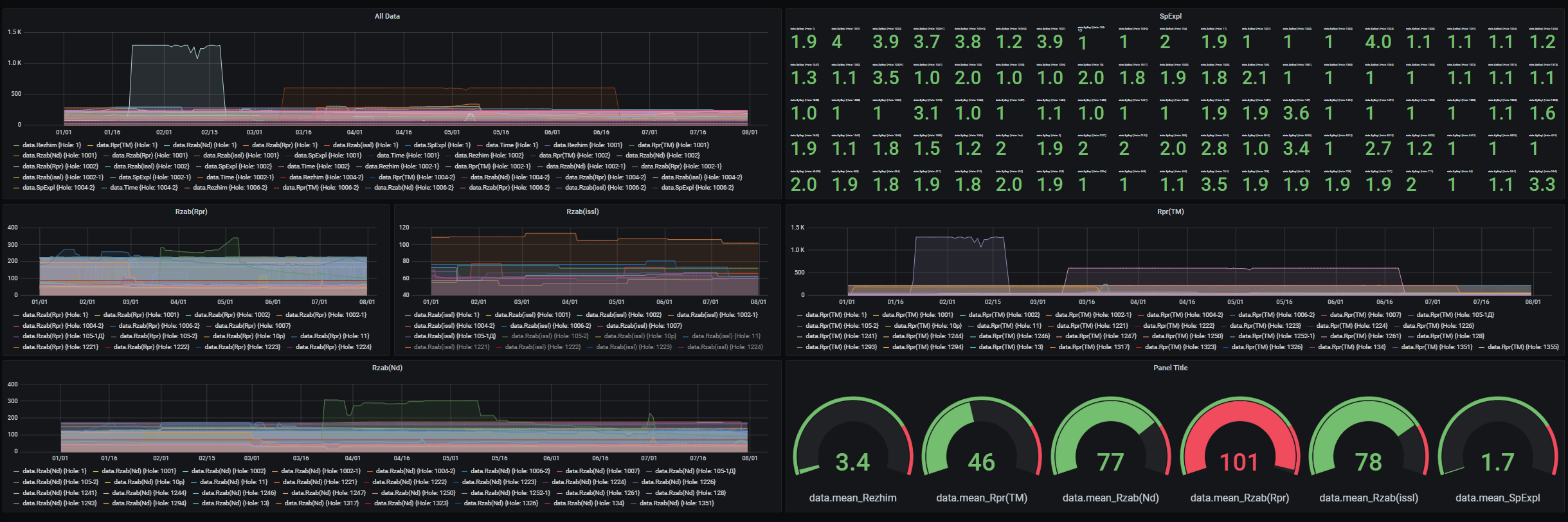
***Рисунок 1.4 – файл «holes.csv»***

Результат действий представлен в виде сырых данных на рисунке 1.5 и рисунке 1.6.

***Рисунок 1.6 – 2 вариант результата***



***Рисунок 1.5 – 1 вариант результата***



# **Очистка данных**

Во время выборки и визуализации сырых данных было выявлено, что часть значений имеет большое отклонение от нормы, в связи с чем был сделан вывод о том, что анализируемые данные содержат аномальные значения.

Существуют различные виды аномалий:

1. Дубли - выбор одного значения
2. Противоречия - числовой результат противоречит реальному состоянию объекта
   1. Типы и виды физических процессов, выбор диапазонов измерений
   2. Резкое изменение параметров, которое невозможно объяснить

Аномалии можно выявить различными способами:

1. Пороговый фильтр - фильтрация всплесков или провалов
2. Фильтр по доверительному интервалу или стандартному отклонению
3. Фильтр по квантилю, как меры оценки распределения величин

Для определения аномалий нами был выбран метод фильтрации по квантилям уровня 0,01 и 0,99. Квантили рассчитывались для каждого из столбцов массива данных. Для вычисления использовалась библиотека pandas, код представлен ниже.

# -\*- coding: utf-8 -\*-

# !pip install influxdb

import pandas as pd

import numpy as np

df = pd.read\_csv('holes.csv', sep=';', engine='python')

q\_low = df.quantile(.01)

q\_high = df.quantile(.99)

print(q\_low)

print(q\_high)

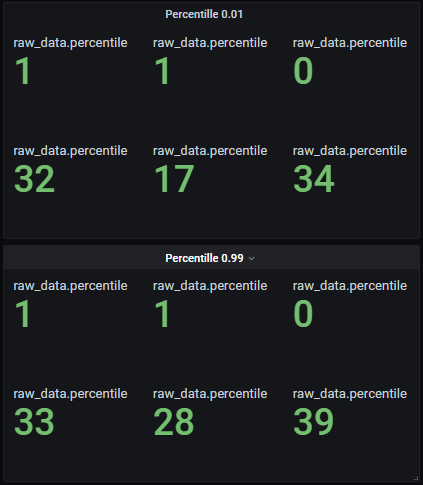
df = df[(df['Режим'] >= q\_low['Режим']) & (df['Режим'] <= q\_high['Режим'])]

df = df[(df['Рпр(ТМ)'] >= q\_low['Рпр(ТМ)']) & (df['Рпр(ТМ)'] <= q\_high['Рпр(ТМ)'])]

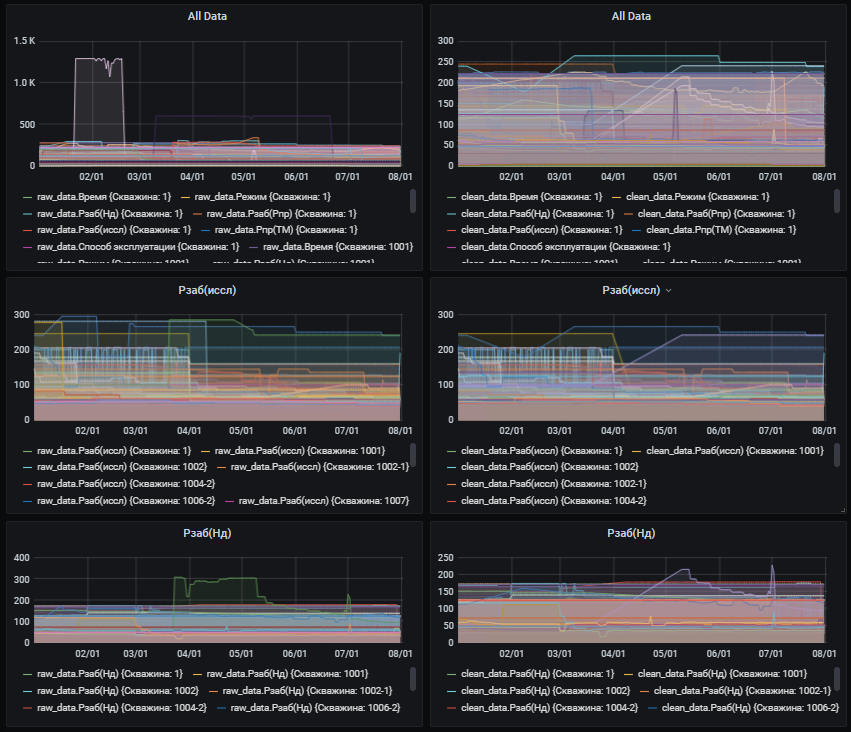
df = df[(df['Рзаб(Рпр)'] >= q\_low['Рзаб(Рпр)']) & (df['Рзаб(Рпр)'] <= q\_high['Рзаб(Рпр)'])]

df = df[(df['Рзаб(иссл)'] >= q\_low['Рзаб(иссл)']) & (df['Рзаб(иссл)'] <= q\_high['Рзаб(иссл)'])]

df.to\_csv('clear\_holes.csv', index=False, sep=';')

На рисунке 2.1 представлены значения квантилей.

***Рисунок 2.1 – Квантили массива данных***

На рисунке 2.2 представлено количество строк до и после срабатывания скрипта для удаления аномалий и соответственно на рисунке 2.3 представлен результат удаления аномалий.

***Рисунок 2.3 – Результат удаления аномалий***

***Рисунок 2.2 – количество строк данных***

# **Трансформация**

Для текущей задачи стоит выделить 3 вида группировки:

1. Группировка по времени - разделение временного ряда по неделям / месяцам
2. Группировка по скважинам - разделение всех метрик по скважинам с различными ID
3. Группировка по способам добычи, по режимам работы и другим метрикам

Так как стоит отделить разные скважины стоит произвести группировку по скважинам.

Фильтрация нужна для того, чтобы:

1. устранить высокочастотные составляющие, всплески и провалы
2. устранить низкочастотные составляющих, сезонность и тренд

Существуют различные виды фильтров:

1. фильтр сглаживающий, простого скользящего среднего с окном 5 значений
2. фильтр экспоненциального сглаживающего среднего, веса значений уменьшаются по мере удаления от центральной точки, взять окно равное 5
3. фильтр медианный
4. фильтр Калмана и другие фильтры с адаптацией к сигналу

Для данных был применен сглаживающий фильтр скользящего среднего. Данный алгоритм формирует набор данных, который служит для построения модели прогнозирования. Код алгоритма представлен ниже.

# -\*- coding: utf-8 -\*-

import numpy as np

import pandas as pd

data = pd.read\_csv('clean\_holes.csv', sep=';', engine='python')

df = pd.DataFrame(data)

print(f"DataFrame:\n{df}\n")

print(f"column types:\n{df.dtypes}")

holes\_list = []

col\_List= df['Скважина'].tolist()

num = 1

for i in range(len(col\_List)):

if (i == 0) or (col\_List[i] != col\_List[i-1]):

holes\_list.append(col\_List[i])

num += 1

columns = ['Рпр(ТМ)', 'Рзаб(Рпр)', 'Рзаб(Нд)', 'Рзаб(иссл)']

for hole in holes\_list:

df1 = df[lambda df: df['Скважина'] == hole]

for col in columns:

col\_List= df1[col].tolist()

for i in range(len(col\_List)):

delta = len(col\_List) - i

left = col\_List[i-1]+col\_List[i-2]+col\_List[i-3]+col\_List[i-4]+col\_List[i-5]

if delta == 1:

right = col\_List[0]+col\_List[1]+col\_List[2]+col\_List[3]+col\_List[4]

if delta == 2:

right = col\_List[i+1]+col\_List[0]+col\_List[1]+col\_List[2]+col\_List[3]

if delta == 3:

right = col\_List[i+1]+col\_List[i+2]+col\_List[0]+col\_List[1]+col\_List[2]

if delta == 4:

right = col\_List[i+1]+col\_List[i+2]+col\_List[i+3]+col\_List[0]+col\_List[1]

if delta == 5:

right = col\_List[i+1]+col\_List[i+2]+col\_List[i+3]+col\_List[i+4]+col\_List[0]

if delta >= 6:

right = col\_List[i+1]+col\_List[i+2]+col\_List[i+3]+col\_List[i+4]+col\_List[i+5]

col\_List[i] = (left + right) / 10

df1[col] = col\_List

df[lambda df: df['Скважина'] == hole] = df1

max = df.max()

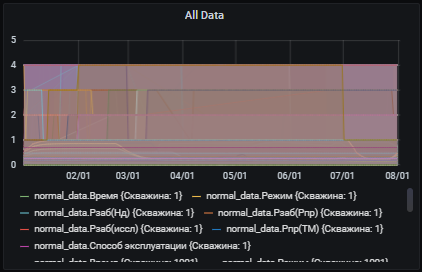
df['Рпр(ТМ)'] = df['Рпр(ТМ)'].apply(lambda x: round(x / max['Рпр(ТМ)'], 3))

df['Рзаб(Рпр)'] = df['Рзаб(Рпр)'].apply(lambda x: round(x / max['Рзаб(Рпр)'], 3))

df['Рзаб(Нд)'] = df['Рзаб(Нд)'].apply(lambda x: round(x / max['Рзаб(Нд)'], 3))

df['Рзаб(иссл)'] = df['Рзаб(иссл)'].apply(lambda x: round(x / max['Рзаб(иссл)'], 3))

df.to\_csv('normal\_holes.csv', index=False, sep=';', encoding='cp1251')

После фильтрации данные были нормализованы делением каждого значения на максимальное к диапазону от 0 до 1. (рис. 3.1)

***Рисунок 3.1 – Нормализованные данные***

На рисунках 3.2 и 3.3 показаны результаты сглаживания по некоторым скважинам.

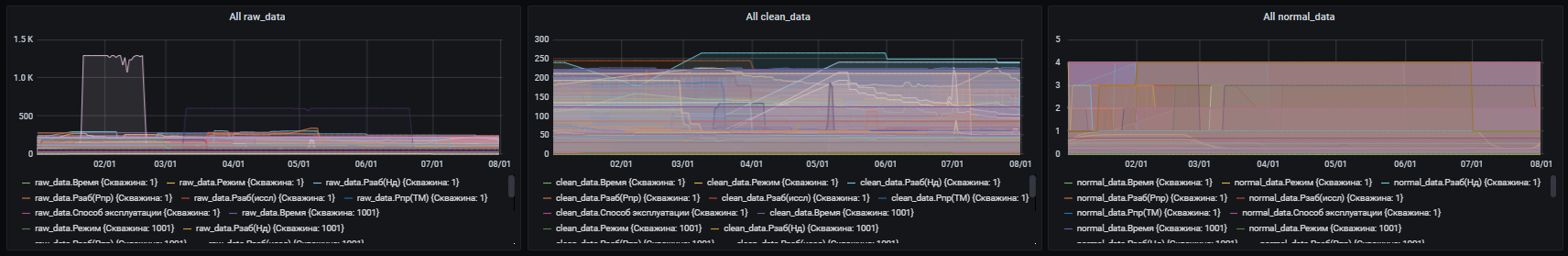
***Рисунок 3.3 – Результат***

***Рисунок 3.2 – Результат***

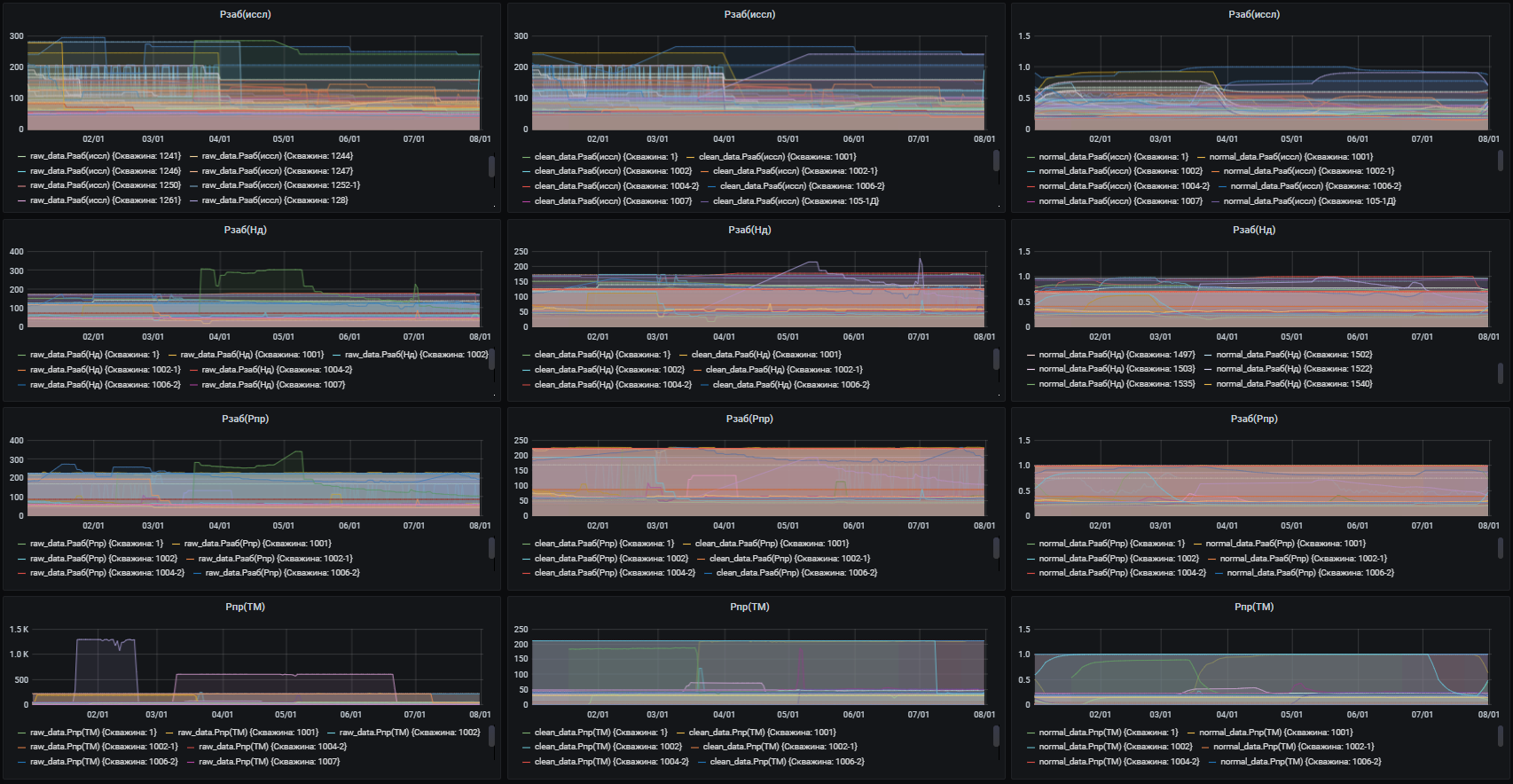
По представленным данным видны резкие перепады во время замера датчиков без сглаживания (слева) и нормализованные данные со сглаживанием (справа). Сглаживание дополняет замеры и помогает получить примерные данные при отсутствии реальных замеров, а так же помогает в визуализации значений на более коротком промежутке, без сглаживания на панели отображалось отсутствие данных, которое неподготовленный человек воспринял бы за ошибку.

# **Анализ**

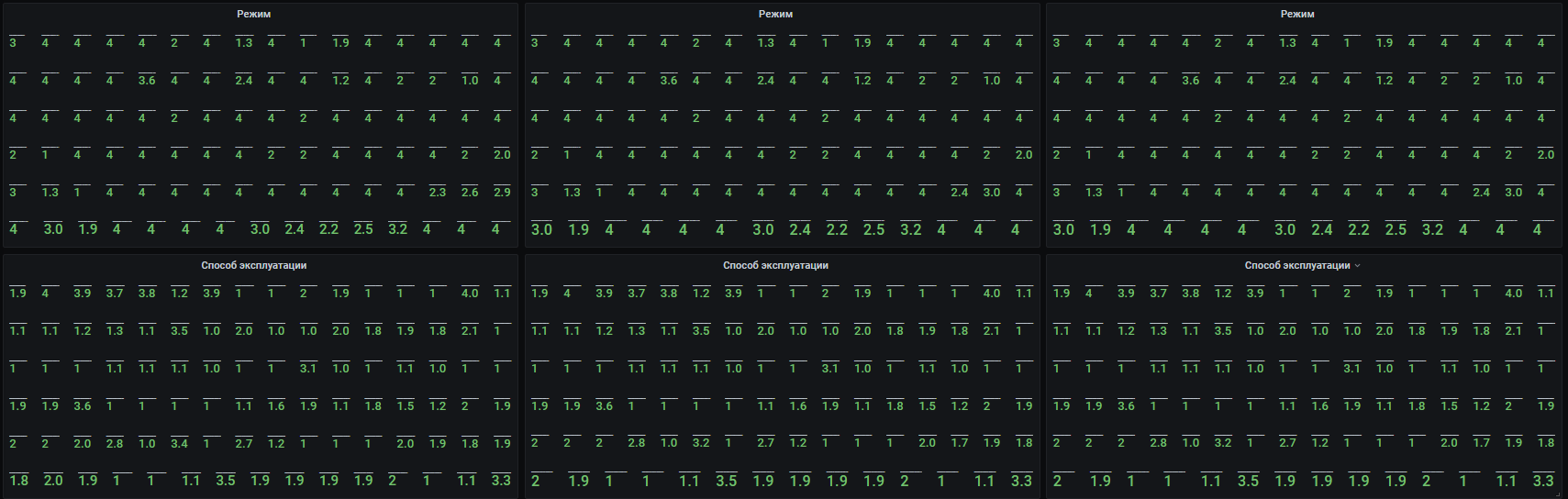
На данном этапе представлен анализ данных с наглядной демонстрацией. В ходе выполнения заданий было создано 3 скрипта для заполнения, очистки и сглаживания значений, а также был модифицирован скрипт для загрузки данных в базу данных Influx. Для выполнения задания потребовалось создать одну базу данных и внести 3 измерения «raw\_data» - список неотфильтрованных значений для визуализыции сырых данных «clean\_data» - список с удалением всех аномалий повторяющихся данных «normal\_data» - нормализованный список с фильтрацией данных.

На рисунке 4.1 представлен формат размещения всех 3 типов данных на панелях Grafana.

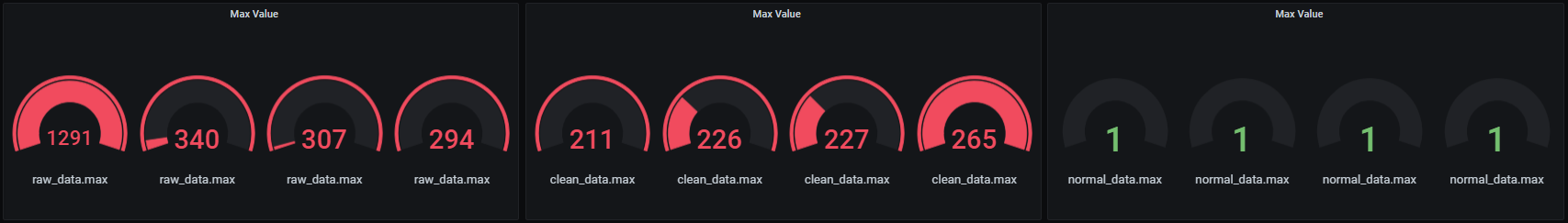
***Рисунок 4.1 – Данные***

На рисунке 4.2 представлены панели графиков с датчиками.

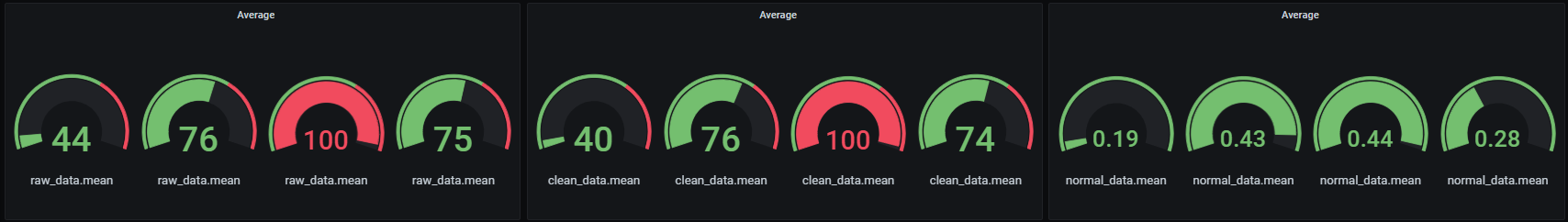
***Рисунок 4.2 – Данные***

На рисунке 4.3 представлены значения режимов и способов эксплуатации

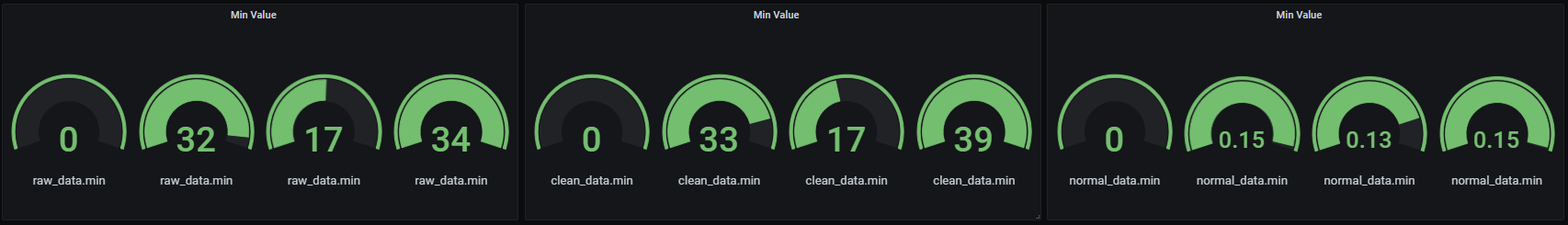
***Рисунок 4.3 – Данные***

На рисунке 4.4 представлены максимальные значения датчиков

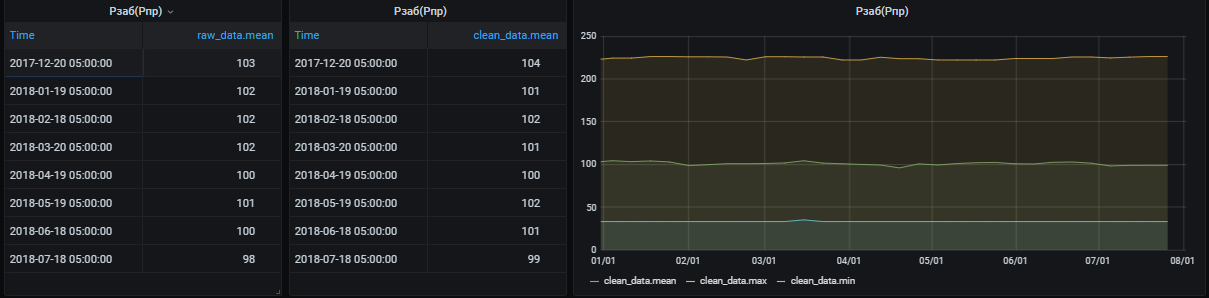
***Рисунок 4.4 – Данные***

На рисунке 4.5 представлены средние значения датчиков

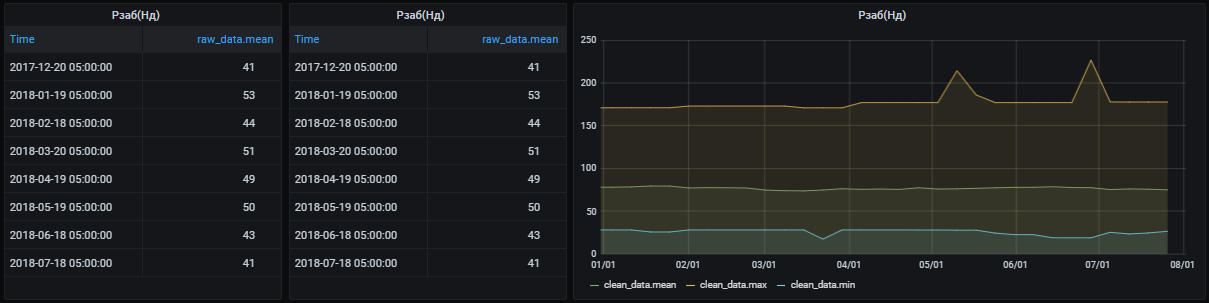
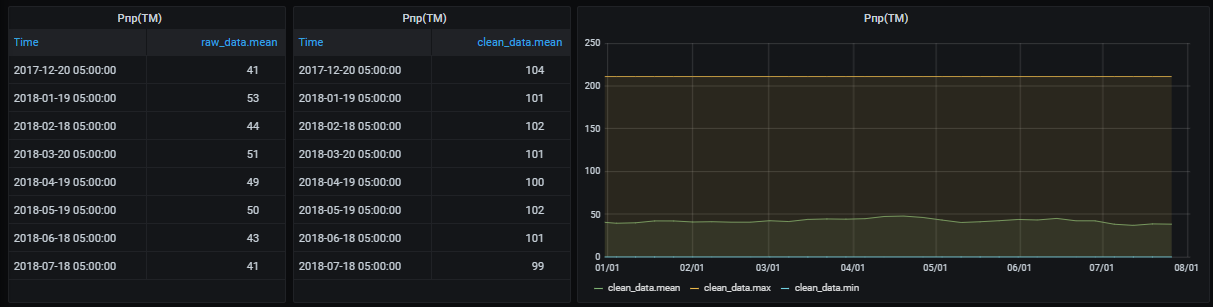
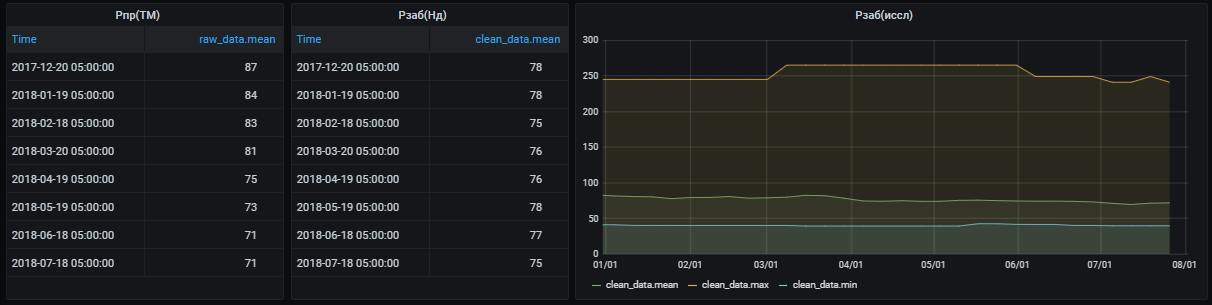
***Рисунок 4.5 – Данные***

На рисунке 4.6 представлены минимальные значения датчиков

***Рисунок 4.6 – Данные***

На рисунках 4.7, 4.8, 4.9, 4.10 представлены графики изменений среднего, максимального и минимального значений, сгруппированных по неделям.

***Рисунок 4.7 – Данные***



***Рисунок 4.10 – Данные***

***Рисунок 4.9 – Данные***

***Рисунок 4.8 – Данные***

# **Заключение**

В ходе работы на основе данных по 500 скважинам были выбраны те, которые соответствовали нашему варианту. Для выбранных данных были произведены очистка (заполнение пропусков и удаление аномальных значений), трансформация (сглаживание методом скользящего окна) и нормализация. Были построены графики, отражающие влияние способа и режима эксплуатации на различные виды давления в скважинах. Мы столкнулись с такими проблемами как некорректное автоматическое определение типа данных в Influx, слетевшая кодировка у текстовых файлов, из-за которой пришлось изменить все названия столбцов, изначально включающих в себя буквы кириллицы, на латинский алфавит.

На основе проведенного в данной работе анализа можно сделать вывод, что способ эксплуатации влияет на давление в скважинах. Основываясь на графиках можно заметить, что давления Рпр(Тм), Рзаб(Рпр) и Рзаб(Нд) обратно пропорционально способу эксплуатации, т.е. с увеличением номера способа эксплуатации значения давлений уменьшаются. Что касается Рзаб(иссл) - они имеют график прямой пропорциональности относительно способа эксплуатации, что означает,

Если посмотреть на графики зависимости от режима давления, то график давления Рпр(Тм) обратно пропорциональны, т.е с увеличением режима значение давления падает. Что касается остальных давлений, их графики имеют параболистический вид, из-за чего сложно характеризовать зависимость данными параметрами. Давления на 1 и 4 режиме находятся примерно на одном уровне, в то время как к 3 режиму значение давлений падает.

# **Приложение А. Листинг скрипта для выборки**

# -\*- coding: utf-8 cp1251 -\*-

"""

Created on 0 0:00:00 0000

@author: Bloodies

"""

# !pip install influxdb

from influxdb import InfluxDBClient

import pandas as pd

import numpy as np

data = 'Дата замера'

hole = 'Скважина'

x1 = 'Способ эксплуатации'

x2 = 'Режим'

y1 = 'Рпр(ТМ)'

y2 = 'Рзаб(Рпр)'

y3 = 'Рзаб(Нд)'

y4 = 'Рзаб(иссл)'

file\_name = 'data.xlsx'

def read\_all\_sheets(file\_name\_excel):

df = pd.DataFrame()

xls = pd.ExcelFile(file\_name\_excel)

for list\_excel in xls.sheet\_names:

df = df.append(pd.read\_excel(xls, list\_excel, parse\_dates=[data], index\_col=data))

return df

df = read\_all\_sheets(file\_name)

df.sort\_index(inplace=True)

def replace\_text\_values\_in\_x(df, nameX):

dict\_changes = {}

\_list = pd.unique(df[nameX]).tolist()

i = 1

for value in \_list:

if (str(value) != str(np.NaN)):

df.loc[df[nameX] == value, nameX] = i

dict\_changes[i] = value

i += 1

df[nameX] = df[nameX].fillna(len(\_list))

dict\_changes[len(\_list)] = np.NaN

return dict\_changes

all\_data = df.copy()

cleaning\_map = lambda x: str(x).strip()

all\_data[hole] = all\_data[hole].map(cleaning\_map)

all\_df\_to\_influx = all\_data.copy()[[hole, x1, x2, y1, y2, y3, y4]]

list\_of\_holes = pd.unique(all\_data[hole]).tolist()

count\_empty\_data = 0

fill\_method = 'bfill'

list\_empty\_data = []

cleaned\_data = pd.DataFrame()

for \_hole in list\_of\_holes[1:100]:

if count\_empty\_data > 0:

fill\_method = 'ffill'

df\_to\_influx = all\_df\_to\_influx[all\_df\_to\_influx[hole] == \_hole][[x1, x2, y1, y2, y3, y4]]

df\_to\_influx.insert(loc=0, column='Время', value=df\_to\_influx.index.time[0])

df\_to\_influx.insert(loc=0, column='Скважина', value=\_hole)

temp\_df = df\_to\_influx[[y1, y2, y3, y4]].dropna(axis=1, how='all')

if (not temp\_df.empty):

cleaned\_data = cleaned\_data.append(df\_to\_influx)

cleaned\_data[y1].fillna(method=fill\_method, inplace=True)

cleaned\_data[y2].fillna(method=fill\_method, inplace=True)

cleaned\_data[y2].fillna(method=fill\_method, inplace=True)

cleaned\_data[y3].fillna(method=fill\_method, inplace=True)

cleaned\_data[y4].fillna(method=fill\_method, inplace=True)

cleaned\_data.to\_csv('output\\' + 'holes'+'.csv', encoding='cp1251', sep=';')

else:

list\_empty\_data.append(\_hole)

count\_empty\_data += 1

print('Сохранение в файл ".csv" выполнено!')

# **Приложение Б. Листинг скрипта для экспорта**

# -\*- coding: utf-8 -\*-

"""

Created on 0 0:00:00 0000

@author: Bloodies

"""

# !pip install influxdb

from influxdb import InfluxDBClient

import csv

import datetime

from pytz import timezone

import pandas as pd

import numpy as np

\_host = 'localhost'

\_port = 8086

index = 0

\_databasename = 'holes'

inputfilename = 'holes.csv' #'holes.csv' #'clear\_holes.csv' #'normal\_holes.csv'

timecolumn = 'Время'

datecolumn = 'Дата замера'

timeformat = '%Y-%m-%d %H:%M:%S'

datatimezone = 'UTC'

\_delimiter=';'

batchsize = 5000

metric = ['raw\_data'] #['raw\_data'] #['clean\_data'] #['normal\_data']

tagcolumns = ['Скважина']

fieldcolumns = []

client = InfluxDBClient(host=\_host, port=\_port)

client.create\_database(\_databasename)

db\_list = client.get\_list\_database()

client.drop\_database(\_databasename) #client.drop\_database(\_databasename)

client.create\_database(\_databasename) #client.create\_database(\_databasename)

client.switch\_database(\_databasename)

epoch\_naive = datetime.datetime.utcfromtimestamp(0)

epoch = timezone('UTC').localize(epoch\_naive)

def unix\_time\_millis(dt):

return int((dt - epoch).total\_seconds() \* 1000)

""" Check if data type of field is float """

def isfloat(value):

try:

float(value)

return True

except:

return False

""" Check if data type of field is int """

def isinteger(value):

try:

if int(value):

return True

else:

return False

except:

return False

datapoints1 = []

count = 0

with open(inputfilename, 'r') as csvfile:

reader = csv.DictReader(csvfile, delimiter=\_delimiter)

fieldcolumns = reader.fieldnames[2:]

for row in reader:

datetime\_naive = datetime.datetime.strptime(row[datecolumn] + ' ' + row[timecolumn], timeformat)

datetime\_local = timezone(datatimezone).localize(datetime\_naive)

timestamp = unix\_time\_millis(datetime\_local) \* 1000000 # in nanoseconds

tags = {}

for t in tagcolumns:

if t in row:

v = row[t]

pass

tags[t] = v

fieldNames = {'Рзаб(Нд)', 'Рзаб(Рпр)'}

fields = {}

for f in fieldcolumns:

v = 0

if f in row:

if isinteger(row[f]):

v = int(row[f])

else:

v = float(row[f]) if isfloat(row[f]) else row[f]

fields[f] = v

datapoints1.append({"measurement": metric[0], "time": timestamp, "fields": fields, "tags": tags})

count += 1

if len(datapoints1) % batchsize == 0:

index += 1

print('Read %d lines'%count)

print('Inserting %d datapoints...'%(len(datapoints1)))

response = client.write\_points(datapoints1)

if response == False:

print('Problem inserting points, exiting...')

exit(1)

print("Wrote %d, response: %s" % (len(datapoints1), response))

datapoints1 = []

print(index)

if len(datapoints1) > 0:

print('Read %d lines'%count)

print('Inserting %d datapoints...'%(len(datapoints1)))

response = client.write\_points(datapoints1)

if response == False:

print('Problem inserting points, exiting...')

exit(1)

print("Wrote %d, response: %s" % (len(datapoints1), response))

print('Done')