1.1. Постановка задачи

Цели и задачи работы:

- Познакомиться с понятием «большие данные» и способами их обработки;
- Познакомиться с инструментом Apache Spark и возможностями, которые он предоставляет для обработки больших данных.
- Получить навыки выполнения разведочного анализа данных использованием pyspark.

1.2. Описание датасета

Чтобы лучше следить за потреблением энергии, правительство хочет, чтобы поставщики энергии установили интеллектуальные счетчики в каждом доме в Англии, Уэльсе и Шотландии. Поставщикам энергии предстоит добраться до более чем 26 миллионов домов, и цель состоит в том, чтобы к 2020 году в каждом доме был установлен интеллектуальный счетчик.

Это внедрение счетчиков возглавляет Европейский союз, который обратился ко всем правительствам-членам с просьбой рассмотреть интеллектуальные счетчики в рамках мер по модернизации нашего энергоснабжения и борьбе с изменением климата. После первоначального исследования британское правительство решило внедрить интеллектуальные счетчики в рамках своего плана по обновлению нашей стареющей энергетической системы.

В этом наборе данных вы найдете (переработанную версию данных) из лондонского хранилища данных, которая содержит данные о потреблении энергии для выборки из 5567 лондонских домохозяйств, принявших участие в проекте UK Power Networks по снижению выбросов углерода в Лондоне в период с ноября 2011 по февраль 2014

года. Данные с интеллектуальных счетчиков, по-видимому, связаны только с потреблением электроэнергии.

Рассматриваемая в данном анализе часть полученных данных является файл, который содержит файлы блоков с ежедневной информацией, такой как количество измерений, минимум, максимум, среднее значение, медиана, сумма и стандартное отклонение.

Ссылка на датасет: https://www.kaggle.com/datasets/jeanmidev/smart-meters-in-london файл: daily_dataset.csv

Датасет содержит 3 510 433 записи.

1.3. Разведочный анализ выбранного датасета

Допасет состоит из 9 признаков:

LCLid – Идинтификатор пользователя

Day – День замера

energy_median - Медиана энергии

energy_mean – Среднее значении энергии

energy_max – Максимальное значении энергии

energy_count – Количество измерений

energy_std – Стандартное отклонение

energy_sum – Сумма значений энергии

energy_min - Минимальное значении энергии

1.3.1. Типы признаков в датасете

Типы признаков показаны на рисунке 1, из данного рисунка можно увидеть, что датасет состоит из 1 признака типа string, одного признака типа timestamp, 6 признаков типа double и 1 признака типа integer.

Columns overview

	Column Name	Data type
0	LCLid	string
1	day	timestamp
2	energy_median	double
3	energy_mean	double
4	energy_max	double
5	energy_count	int
6	energy_std	double
7	energy_sum	double
8	energy_min	double

Рисунок 1 – Типы признаков в датасете

1.3.2. Пропущенные значения и их устранение

По проведенному анализу было обнаружено 11331 (Столбец energy_std) + 30 строк (По всем остальным показателями энергии) с пропущенными значениями. Существует несколько способов устранить пропущенные значения, однако в данном случает подойдет наипростейший из этих способов – удаление строк с пропущенными значениями, этот способ был выбран исходя из того, что 30 пропущенных значений не имеют вообще никаких значений энергия, отсюда они совершенно никак не повлияют на дальнейший анализ данных. Также если посчитать общее количество строк с пропущенными значениями, то их всего будет 11361, что является очень малым процентом от всех записей в датасете (около 0.32%), исходя из этого данные записи почти никак не повлияют на дальнейши анализ датасета, поэтому их можно просто удалить. Данные выполненные действия можно увидеть на рисунке 2.

```
Количество пропущенных значений до удаления
+-----+---+-------------+-----------+----------+------------+----------+----------+----------+
|LCLid|day|energy_median|energy_mean|energy_max|energy_count|energy_std|energy_sum|energy_min|
+-----+---+-------------+-----------+----------+------------+----------+----------+----------+
|    0|  0|           30|         30|        30|           0|     11331|        30|        30|
+-----+---+-------------+-----------+----------+------------+----------+----------+----------+

Количество пропущенных значений после удаления
+-----+---+-------------+-----------+----------+------------+----------+----------+----------+
|LCLid|day|energy_median|energy_mean|energy_max|energy_count|energy_std|energy_sum|energy_min|
+-----+---+-------------+-----------+----------+------------+----------+----------+----------+
|    0|  0|            0|          0|         0|           0|         0|         0|         0|
+-----+---+-------------+-----------+----------+------------+----------+----------+----------+
```

Рисунок 2 – Пропущенные значения и их устранение

1.3.3. Удаление ненужных столбцов

Для дальнейшего анализа было решено удалить два столбца, столбец с уникальным идентификатором и днем замера энергии, так как для анализа данных они особо не имеют смысла и никак не влияют на дальнейший анализ, так как данные из этих двух столбцов невозможно разделить на категории, так как разных значение в этих столбцах очень больше количество. Схемы датасетов до и после удаления столбцов можно увидеть на рисунке 3.

```
root
 |-- LCLid: string (nullable = true)
 |-- day: timestamp (nullable = true)
 |-- energy_median: double (nullable = true)
 |-- energy_mean: double (nullable = true)
 |-- energy_max: double (nullable = true)
 |-- energy_count: integer (nullable = true)
 |-- energy_std: double (nullable = true)
 |-- energy_sum: double (nullable = true)
 |-- energy_min: double (nullable = true)

root
 |-- energy_median: double (nullable = true)
 |-- energy_mean: double (nullable = true)
 |-- energy_max: double (nullable = true)
 |-- energy_count: integer (nullable = true)
 |-- energy_std: double (nullable = true)
 |-- energy_sum: double (nullable = true)
 |-- energy_min: double (nullable = true)
```

Рисунок 3 – Схемы датасетов до и после удаления столбцов

1.3.4. Выбросы и их устранение

Для устранения выбросов, необходимо их сначала определить. Сначала определим их графически, графики выбросов, построить их можно с помощью графика, называемым «Коробка(Ящик) с усами» полученные графики представлены на рисунке 4.

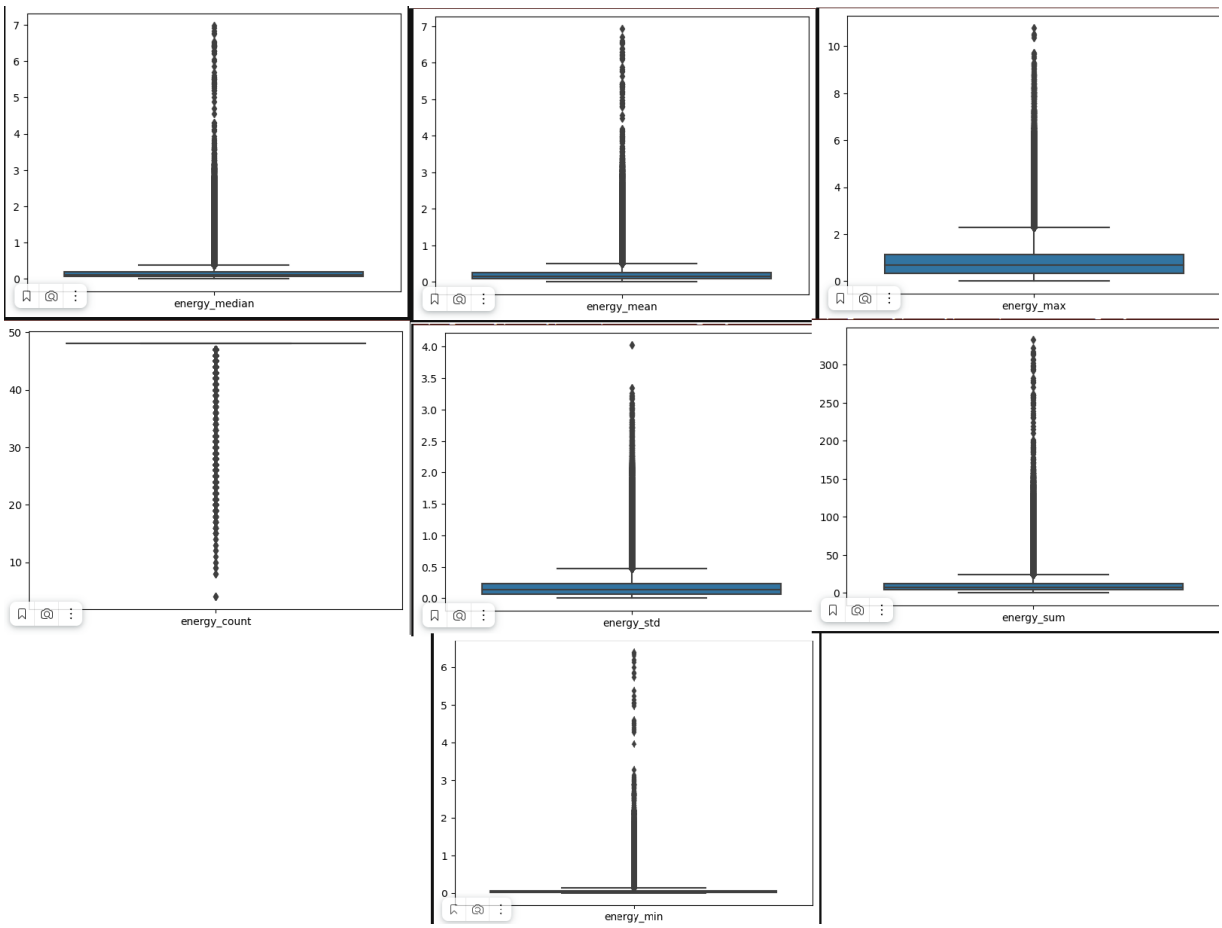

Рисунок 4 – Графики выбросов

По графикам явно видно, что в данных имеются выбросы (Визуально можно увидеть, что есть много точек входящих в зону выбросов). В виде цифр средствами Spark также можно определить выбросы, для данной задачи были определены границы(первый и третий квантиль) и относительная ошибка, которая определяет точность квантилей (было выбрано 0.01 так как при более низком значении ресурсы компьютера не позволяли увеличить точность). По итогам в зависимости от столбцов было найдено от 29 тыс до 230 тыс выбросов. После

обнаружения выбросов они были удалены. Количество выбросов до и после их удаления приведены на рисунке 5.

```
+----------------+---------------+---------------+----------------+--------------+---------------+---------------+
|energy_median_out|energy_mean_out|energy_max_out|energy_count_out|energy_std_out|energy_sum_out|energy_min_out|
+----------------+---------------+---------------+----------------+--------------+---------------+---------------+
|          228286|         201106|         125877|           29750|        164196|         206561|         227676|
+----------------+---------------+---------------+----------------+--------------+---------------+---------------+

+----------------+---------------+---------------+----------------+--------------+---------------+---------------+
|energy_median_out|energy_mean_out|energy_max_out|energy_count_out|energy_std_out|energy_sum_out|energy_min_out|
+----------------+---------------+---------------+----------------+--------------+---------------+---------------+
|               0|              0|              0|               0|             0|              0|              0|
+----------------+---------------+---------------+----------------+--------------+---------------+---------------+
```

Рисунок 5 – Количество выбросов до и после их удаления

Также приведем графики после удаления выбросов, они приведены на рисунке 6. Как можно видеть из рисунка, графики стали выглядеть лучше.

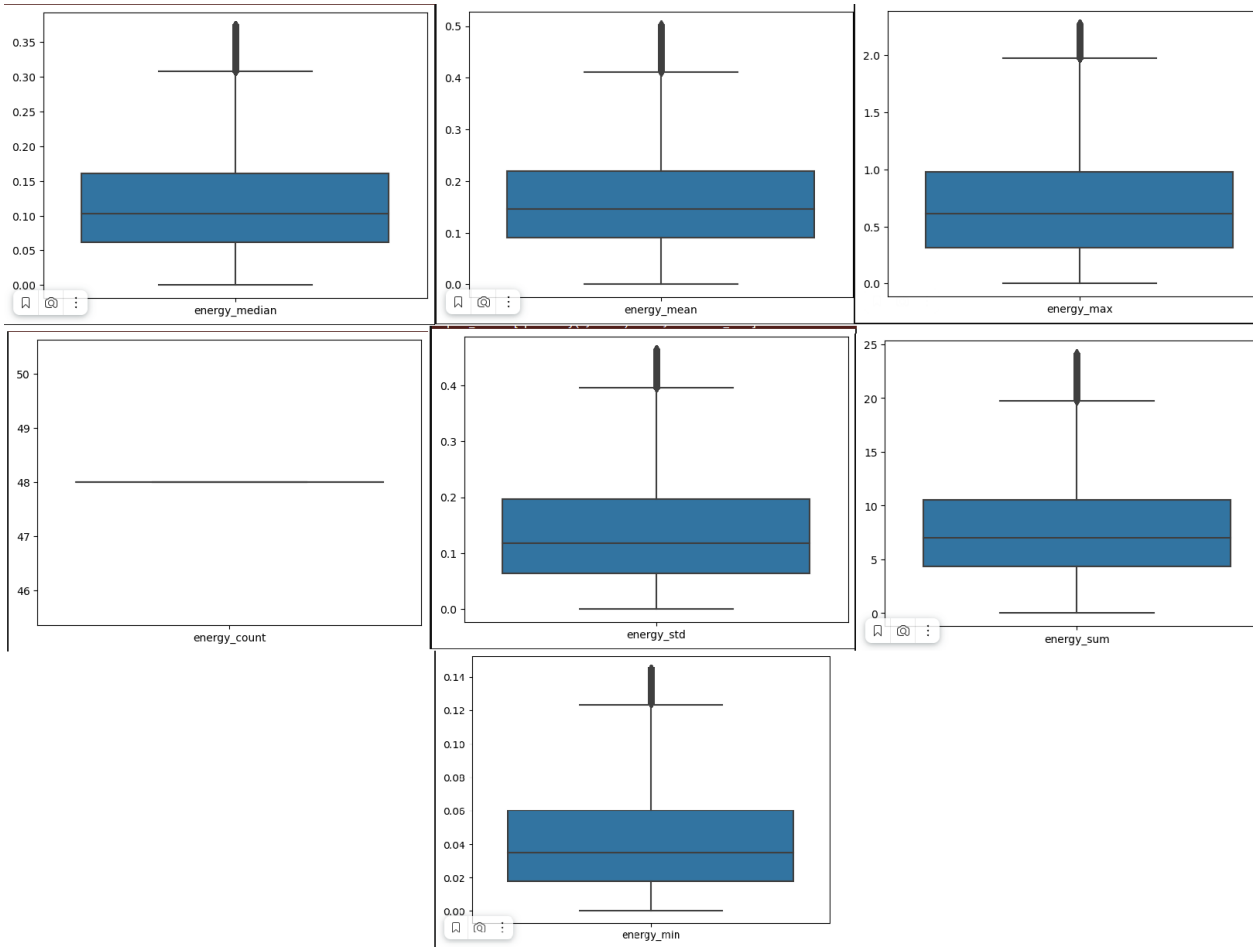

Рисунок 6 – Графики после удаления выбросов

1.3.5. Расчет статистических показателей признаков

В анализе после отчистки также был проведен расчет статистических показателей признаков (количество, среднее, стандартное отклонение, максимальное и квантили). Из полученных результатов можно прийти к следующим выводам: было удалено 467 817 (13.33%) записей с очистки данных, стандартное отклонение во многом низка, кроме energy_max. Мксимальные и минимальные значения не выходят за пределы 50% процентов, кроме столбцов energy_max, energy_sum. Полученные результаты приведены на рисунке 7.

	count	mean	std	min	25%	50%	75%	max
energy_median	3042616.0	0.118319	0.074065	0.0	0.062000	0.102500	0.160500	0.374500
energy_mean	3042616.0	0.162498	0.093654	0.0	0.090792	0.146083	0.219208	0.503229
energy_max	3042616.0	0.692577	0.462645	0.0	0.316000	0.610000	0.980000	2.277000
energy_count	3042616.0	48.000000	0.000000	48.0	48.000000	48.000000	48.000000	48.000000
energy_std	3042616.0	0.139730	0.097191	0.0	0.063433	0.118410	0.196322	0.463973
energy_sum	3042616.0	7.799899	4.495386	0.0	4.358000	7.012000	10.522000	24.155000
energy_min	3042616.0	0.042689	0.031929	0.0	0.018000	0.035000	0.060000	0.145000

Рисунок 7 – Статистические показатели признаков

1.3.6. Визуализация распределения наиболее важных признаков

В данной работе были построены и визуализированы графики распределения признаков. Полученные графики показаны на рисунке 8. Из рисунка видно, что все признаки, кроме count имеют нормальное распределение.

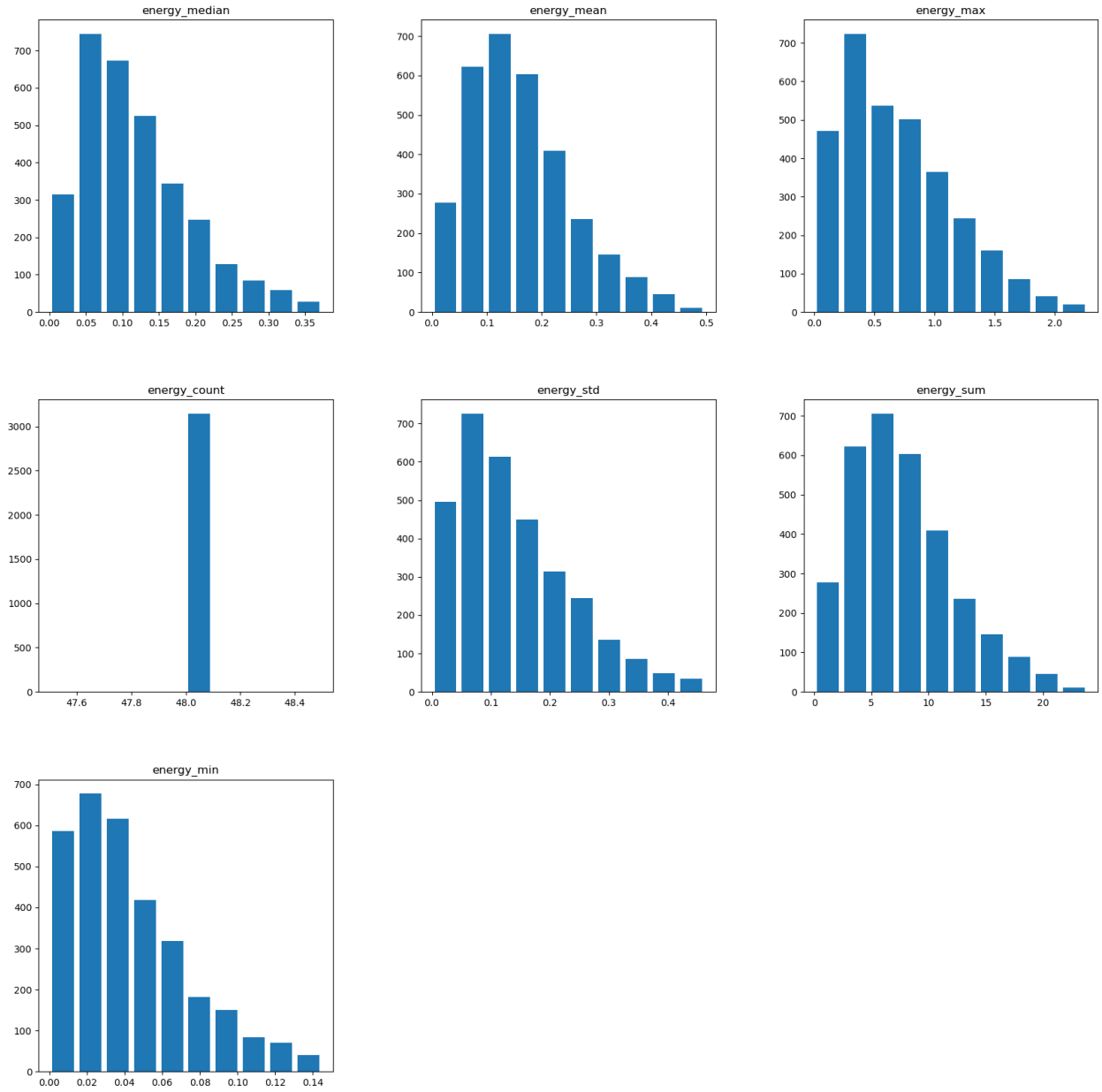

Рисунок 8 – Графики распределение признаков

1.3.7. Корреляция между признаками

В данной работе также была построена корреляционная матрица, с корреляцией между признаками. Данная матрица была построенная двумя способами с использованием Spark и с переводом spark датафрейма в датафрейм Pandas. Оба варианта дали примерно одинаковые значения. Spark может высчитать корреляцию между признаками только, если ему поступают данные в векторном виде, отсюда данные исходного датасета были представлены в виде вектора,

результат перевода можно увидеть на рисунке 9. После преобразования в векторный вид была построена матрица корреляции, она представлена на рисунке 10.

```
После перевода в векторный вид:
+--------------------------------------------------------------------------------------------------------+
|features                                                                                                |
+--------------------------------------------------------------------------------------------------------+
|[0.1415,0.2961666687500003,1.1160001,48.0,0.2814713178628203,14.216000100000002,0.031]                  |
|[0.1015,0.1898125,0.685,48.0,0.1884046862418033,9.111,0.064]                                            |
|[0.114,0.2189791666666666,0.6759999999999999,48.0,0.20291927853038208,10.510999999999996,0.065]        |
|[0.191,0.32597916666666665,0.7879999999999999,48.0,0.2592049619947409,15.646999999999998,0.066]        |
|[0.21800000000000005,0.3575,1.077,48.0,0.28759657027517305,17.16,0.066]                                 |
|[0.1305,0.2350833333333333,0.705,48.0,0.2220696491599295,11.284,0.066]                                  |
|[0.08900000000000001,0.22135416666666666,1.094,48.0,0.26723887549908265,10.625,0.062]                   |
|[0.16049999999999998,0.291125,0.7490000000000001,48.0,0.24907604794434665,13.973999999999998,0.065]     |
|[0.2175,0.3391875000000003,0.866,48.0,0.26310119857478675,16.281000000000002,0.069]                     |
|[0.14950000000000002,0.2617083333333333,0.838,48.0,0.2447927441503373,12.562000000000001,0.066]         |
|[0.14300000000000002,0.2740000000000001,0.778,48.0,0.25212745847913703,13.152000000000005,0.068]        |
|[0.14550000000000002,0.3005208333333333,1.207,48.0,0.29868028801773083,14.425,0.066]                    |
|[0.152,0.3070416666666667,0.888,48.0,0.2644546341928976,14.738,0.066]                                    |
|[0.135,0.27685416666666673,0.782,48.0,0.261185756802965,13.289000000000005,0.064]                       |
|[0.1515,0.32572916666666674,1.252,48.0,0.3098882941898363,15.635000000000005,0.066]                     |
|[0.151,0.25602083333333336,0.812,48.0,0.2252494116065079,12.289,0.068]                                  |
|[0.134,0.2520833333333333,0.851,48.0,0.23721296951853504,12.1,0.068]                                    |
|[0.14750000000000002,0.23550000000000004,0.674,48.0,0.2099953393606427,11.304000000000002,0.068]       |
|[0.10099999999999999,0.21627083333333327,0.731,48.0,0.21520576394077676,10.380999999999997,0.065]      |
|[0.146,0.3310208333333334,0.7859999999999999,48.0,0.2733733702406611,15.889000000000005,0.063]         |
+--------------------------------------------------------------------------------------------------------+
```

Рисунок 9 – Датасет в векторном виде

	energy_median	energy_mean	energy_max	energy_count	energy_std	energy_sum	energy_min
energy_median	1.000000	0.886252	0.493438	NaN	0.514547	0.886252	0.682119
energy_mean	0.886252	1.000000	0.746958	NaN	0.814137	1.000000	0.621188
energy_max	0.493438	0.746958	1.000000	NaN	0.945440	0.746958	0.284913
energy_count	NaN	NaN	NaN	1.0	NaN	NaN	NaN
energy_std	0.514547	0.814137	0.945440	NaN	1.000000	0.814137	0.243589
energy_sum	0.886252	1.000000	0.746958	NaN	0.814137	1.000000	0.621188
energy_min	0.682119	0.621188	0.284913	NaN	0.243589	0.621188	1.000000

Рисунок 10 – Матрица корреляции между признаками

Для вывода матрицы корреляции в Pandas нужно преобразовать датафрейм Spark в датафрейм Pandas, сделать это можно функцией Spark sample, после преобразования можно средствами Pandas вывести матрицу корреляции, она представлена на рисунке 11.

	energy_median	energy_mean	energy_max	energy_count	energy_std	energy_sum	energy_min
energy_median	1.000000	0.889477	0.495739	NaN	0.514701	0.889477	0.671720
energy_mean	0.889477	1.000000	0.748211	NaN	0.809783	1.000000	0.609583
energy_max	0.495739	0.748211	1.000000	NaN	0.948220	0.748211	0.259083
energy_count	NaN	NaN	NaN	NaN	NaN	NaN	NaN
energy_std	0.514701	0.809783	0.948220	NaN	1.000000	0.809783	0.215683
energy_sum	0.889477	1.000000	0.748211	NaN	0.809783	1.000000	0.609583
energy_min	0.671720	0.609583	0.259083	NaN	0.215683	0.609583	1.000000

Рисунок 11 – Матрица корреляции между признаками

1.4. Выводы

В данной работе я познакомился с понятием «большие данные» и выполнил разведочный анализ данных на найденном датасете. В процессе работы я познакомился с инструментом Apache Spark и возможностями, которые он предоставляет для обработки больших данных. В данной работе были найдены типы признаков в датасете, были найдены и устранены пропущенные значения и выбросы, были рассчитаны статистические показатели признаков, также были провизуализированы распределения признаков и были найдены корреляции между признаками. В результате данной главы были подготовлены данные для дальнейшего машинного обучения.