Постановка задачи

Цель и задачи работы:

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

В ходе работы необходимо найти произвольный датасет, в котором представлены табличные данные с объемом от нескольких сотен мегабайт.

Далее необходимо провести разведочный анализ выбранного датасета с определением:

* типов признаков в датасете;
* пропущенных значений и их устранением;
* выбросов и их устранением;
* расчетом статистических показателей признаков (средних, квартилей и т.д.);
* визуализацией распределения наиболее важных признаков;
* корреляций между признаками.

Описание датасета

В качестве датасета был выбран файл формата CSV под названием «WeatherEvents\_Jan2016-Dec2021.csv». Размер файла составляет 897 Мб.

Ссылка – https://www.kaggle.com/datasets/sobhanmoosavi/us-weather-events

Данный датасет содержит общенациональный набор данных о погодных явлениях, который включает в себя 7,5 миллионов событий и охватывает 49 штатов США. Примерами погодных явлений являются дождь, снег, шторм и заморозки. Некоторые события в этом наборе данных являются экстремальными явлениями (например, шторм), а некоторые можно рассматривать как обычные события (например, дождь и снег). Данные собраны с января 2016 года по декабрь 2021 года с использованием исторических сводок погоды, собранных с 2071 метеостанции, расположенных в аэропортах по всей стране.

Описание погодных явлений

Погодное событие - это пространственно-временная сущность, где такая сущность связана с местоположением и временем. Ниже приводится описание доступных типов погодных явлений в этом наборе данных:

* Сильный мороз: случай чрезвычайно низкой температуры с температурой ниже -23,7 градусов по Цельсию.
* Туман: случай, когда из-за тумана или дымки наблюдается плохая видимость.
* Град: случай выпадения твердых осадков, включая ледяные гранулы и град.
* Дождь: случай наличия дождя, варьирующегося от легкого до сильного.
* Снег: случай наличия снега, варьирующегося от легкого до тяжелого.
* Шторм: чрезвычайно ветреное состояние, при котором скорость ветра составляет не менее 60 км / ч.
* Другие осадки: Любой другой тип осадков, который не может быть отнесен к ранее описанным типам событий.

Датасет состоит из табличных данных и содержит такие признаки, как:

1. Идентификатор записи
2. Тип события
3. Серьезность события
4. Время начала события в часовом поясе UTC
5. Время окончания события в часовом поясе UTC
6. Общее количество осадков в дюймах
7. Часовой пояс в США, основанный на местоположении события
8. Станция в аэропорту, на которой сообщается о погодном явлении
9. Широта в координатах GPS метеостанции в аэропорту
10. Долгота в координатах GPS метеостанции в аэропорту
11. Город в адресной записи метеостанции, расположенной в аэропорту
12. Округ в адресной записи метеостанции, расположенной в аэропорту
13. Штат в адресной записи метеостанции в аэропорту
14. Почтовый индекс в адресной записи метеостанции в аэропорту

Разведочный анализ

В ходе анализа были выявлены следующие типы признаков в датасете:

|  |  |  |
| --- | --- | --- |
| Признак | тип данных | nullable |
| EventId | string | True |
| Type | string | True |
| Severity | string | True |
| StartTime(UTC) | timestamp | True |
| EndTime(UTC) | timestamp | True |
| Precipitation(in) | double | True |
| TimeZone | string | True |
| AirportCode | string | True |
| LocationLat | double | True |
| LocationLng | double | True |
| City | string | True |
| County | string | True |
| State | string | True |
| ZipCode | integer | True |

Далее в ходе выявления пропущенных значений было определено процентное соотношение пропущенных значений от числа всех значений.

В итоге была получена следующая статистика:

* EventId - 0%
* Type - 0%
* Severity - 0%
* StartTime(UTC) - 0%
* EndTime(UTC) - 0%
* Precipitation(in) - 0%
* TimeZone - 0%
* AirportCode - 0%
* LocationLat - 0%
* LocationLng - 0%
* City - 0%
* County - 0%
* State - 0%
* ZipCode - 1%

Далее было произведено удаление строк с пропущенными значениями.

Для определения выбросов были построены коробчатые диаграммы признаков, вычислены 25% и 75% квартили, а так же получен межквартильный диапазон.

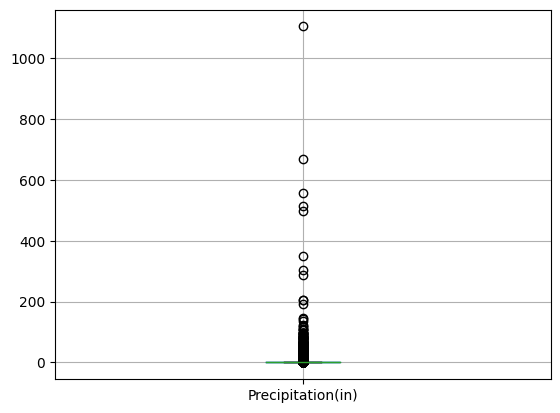


Рисунок 1 – коробчатая диаграмма признака Precipitation(in)

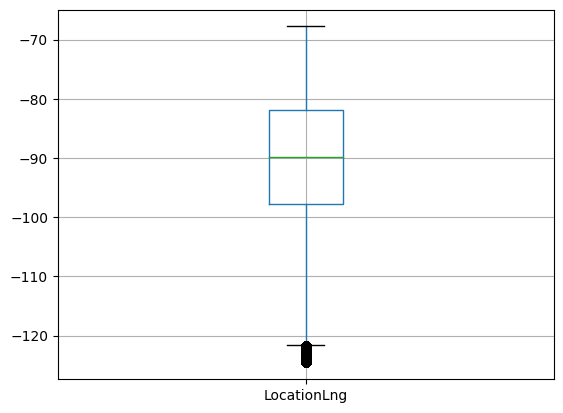


Рисунок 2 – коробчатая диаграмма признака LocationLng

После удаления строк с выбросами признака Precipitation(in) в датасете осталось 6341972 строк.

При расчете статистических показателей признаков были получены следующие данные:

Общее число строк – 7479165

Число указанных городов – 7464602

Число указанных почтовых индексов – 7419931

Среднее число количества осадков – 0.0951 дюймов

Среднеквадратическое отклонение количества осадков – 0.918

Так же были найдены минимальные и максимальные значения всех признаков и их квартили.

Далее были получены визуализации распределения

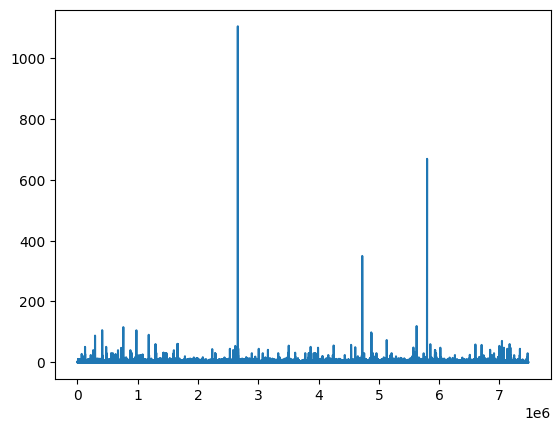


Рисунок 3 – визуализация распределения признака Precipitation(in) до удаления выбросов

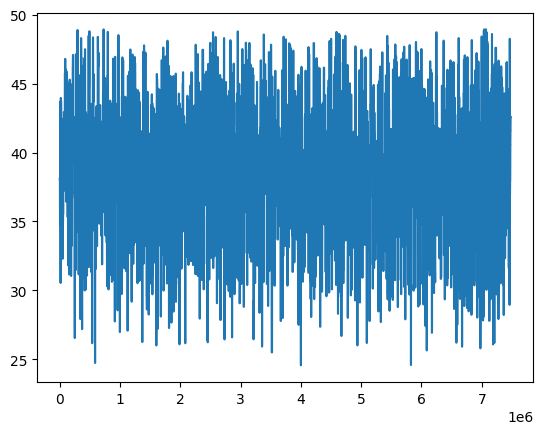


Рисунок 4 – визуализация распределения признака LocationLat

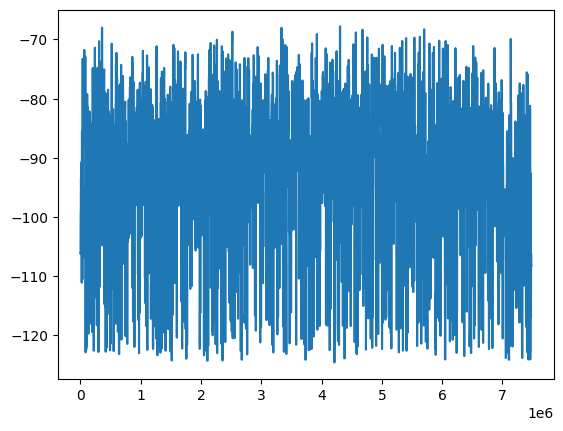


Рисунок 5 – визуализация распределения признака LocationLng

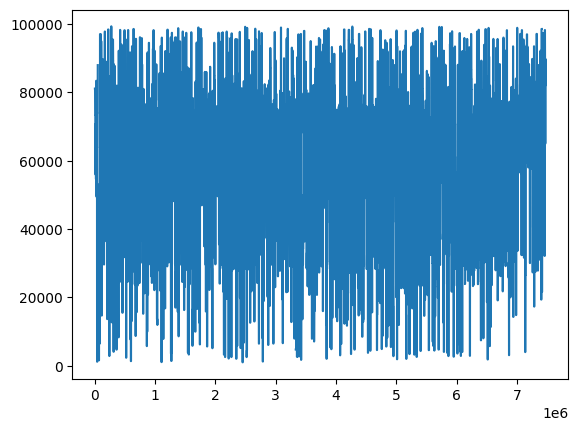


Рисунок 6 – визуализация распределения признака ZipCode

На заключительном шаге была получена корреляция между признаками LocationLat и LocationLng (долгота и широта) и тем самым получены координаты всех метеостанций США.

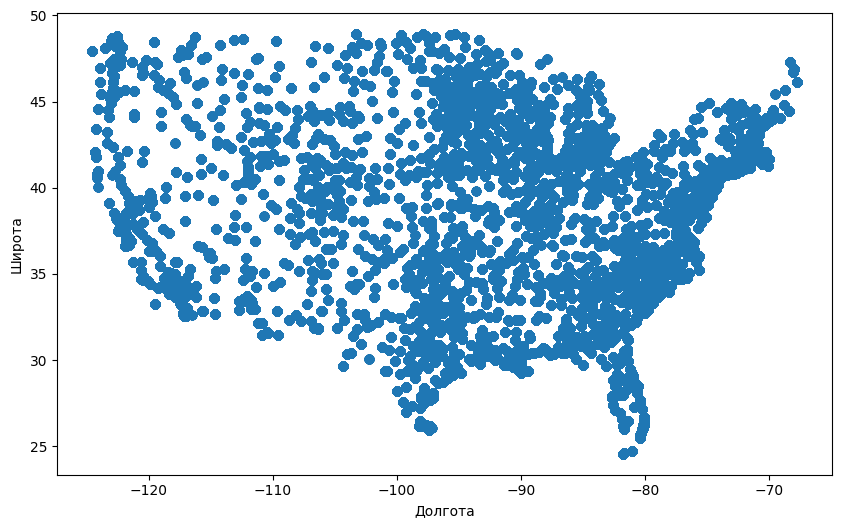


Рисунок 7 – координаты метеостанций

Выводы

В ходе работы был проведен разведочный анализ данных о погодных явлениях в США за 2016 – 2020 годы. В результате были выявлены типы признаков датасета, определены и удалены пропущенные значения и выбросы, а также были построены графики распределения признаков и корреляции между ними.