Минобрнауки России

Федеральное государственное бюджетное образовательное учреждение высшего образования

«Волгоградский государственный технический университет»

Факультет \_\_\_Электроники и вычислительной техники\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Кафедра\_\_\_\_\_Электронно-вычислительные машины и системы\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

**к курсовой работе (проекту)**

по дисциплине \_\_Системы обработки больших данных\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

на тему\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

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

Студент\_Кривабоков Дмитрий Андреевич\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(фамилия, имя, отчество)

Группа\_ЭВМ 1.1\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Руководитель работы (проекта) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_П.Д. Кравченя\_\_\_\_

(подпись и дата подписания) (инициалы и фамилия)

Члены комиссии:

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

(подпись и дата подписания) (инициалы и фамилия)

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

(подпись и дата подписания) (инициалы и фамилия)

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

(подпись и дата подписания) (инициалы и фамилия)

Нормоконтролер \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(подпись, дата подписания) (инициалы и фамилия)

Волгоград 2022 г.

Минобрнауки России

Федеральное государственное бюджетное образовательное учреждение высшего образования

«Волгоградский государственный технический университет»

Факультет \_\_Электроники и вычислительной техники\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Направление (специальность)\_\_09.04.01 — Информатика и вычислительная техника\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
Кафедра \_\_Электронно-вычислительные машины и системы\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дисциплина\_\_Системы обработки больших данных\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |  |
| --- | --- |
|  | Утверждаю  Зав. кафедрой\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
|  | «\_\_\_\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_20\_\_\_ г. |

**Задание**

**на курсовую работу (проект)**

Студент\_Кривабоков Дмитрий Андреевич\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(фамилия, имя, отчество)

Группа\_ЭВМ 1.1\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Тема: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

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

Утверждена приказом от «\_\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_ 20\_\_\_ г. № \_\_\_\_\_\_\_\_\_

2. Срок представления работы (проекта) к защите «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_20\_\_ г.

3. Содержание расчетно-пояснительной записки: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

4. Перечень графического материала: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

5. Дата выдачи задания «\_\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_20 \_\_\_ г.

Руководитель работы (проекта)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_П.Д. Кравченя\_\_\_\_

подпись, дата инициалы и фамилия

Задание принял к исполнению\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

подпись, дата инициалы и фамилия

Оглавление

[1. Разведочный анализ данных с помощью PySpark 4](#_Toc122877036)

[1.1. Постановка задачи 4](#_Toc122877037)

[1.2. Описание датасета 4](#_Toc122877038)

[1.3. Разведочный анализ 6](#_Toc122877039)

[Выводы по первой главе 10](#_Toc122877040)

# Разведочный анализ данных с помощью PySpark

## 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 (таблица 1):

Таблица 1 ­– Типы признаков датасета

|  |  |  |
| --- | --- | --- |
| 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 количество пропущенных значений всех признаков.

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 - 14563
* County - 0
* State - 0
* ZipCode - 59234

Далее было произведено удаление строк с пропущенными значениями.

Для определения выбросов были вычислены 25%, 50% и 75% квартили, а так же получен межквартильный диапазон.

После удаления строк с выбросами признаков в датасете осталось 6341972 строк.

При расчете статистических показателей признаков были получены следующие данные (таблица 2):

Таблица 2 – Статистические показатели

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| summary | Precipitation(in) | LocationLat | LocationLng | ZipCode |
| count | 7479165 | 7479165 | 7479165 | 7419931 |
| mean | 0.09518491703 | 38.733361884 | 91.92212567 | 52478.0940 |
| stddev | 0.918590648760 | 5.46481468189 | 13.47916032643 | 25714.89711421 |
| min | 0.0 | 24.5571 | -124.555 | 1022 |
| 25% | 0.0 | 34.4978 | -97.8283 | 31314 |
| 50% | 0.0 | 39.275 | -89.8222 | 54017 |
| 75% | 0.05 | 42.955 | -81.9565 | 73521 |
| max | 1104.13 | 48.9402 | -67.7928 | 99362 |

Далее были получены визуализации распределения

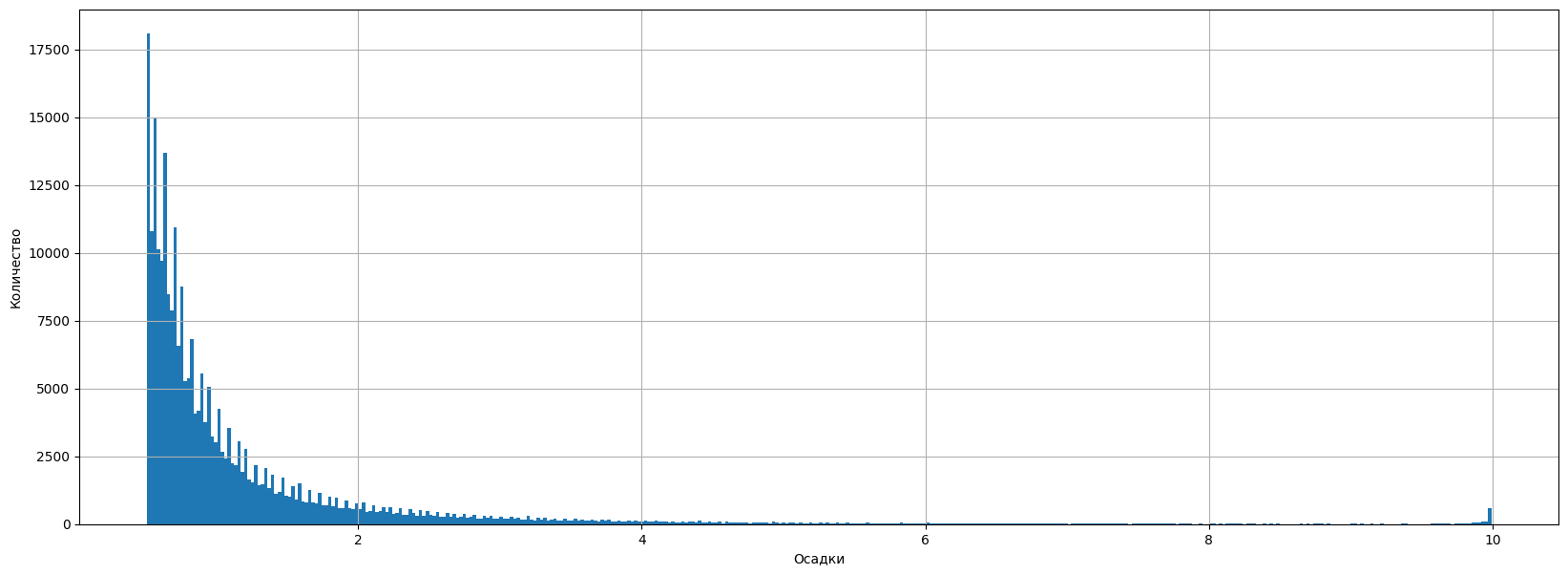


Рисунок 1 – визуализация распределения признака Precipitation(in) до удаления выбросов

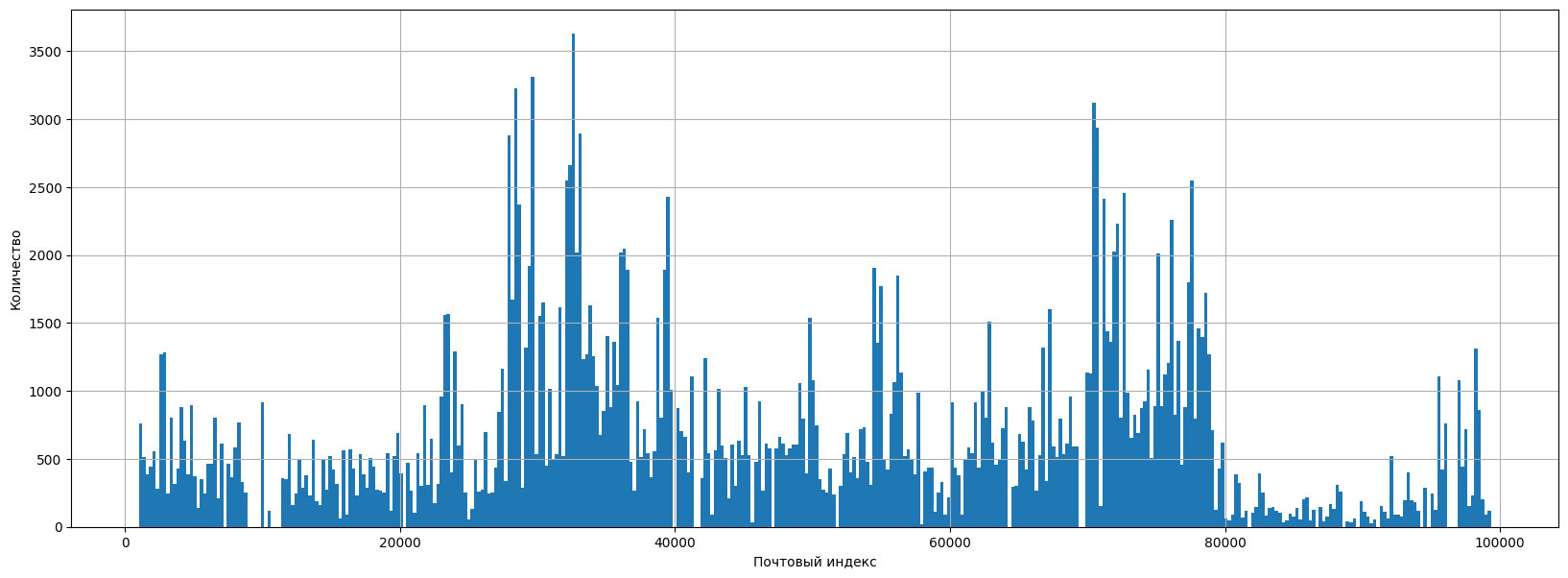


Рисунок 2 – визуализация распределения признака ZipCode

Для определения корреляции между признаками LocationLat и LocationLng (долгота и широта), была построена диаграмма рассеяния и тем самым получены координаты всех метеостанций США.

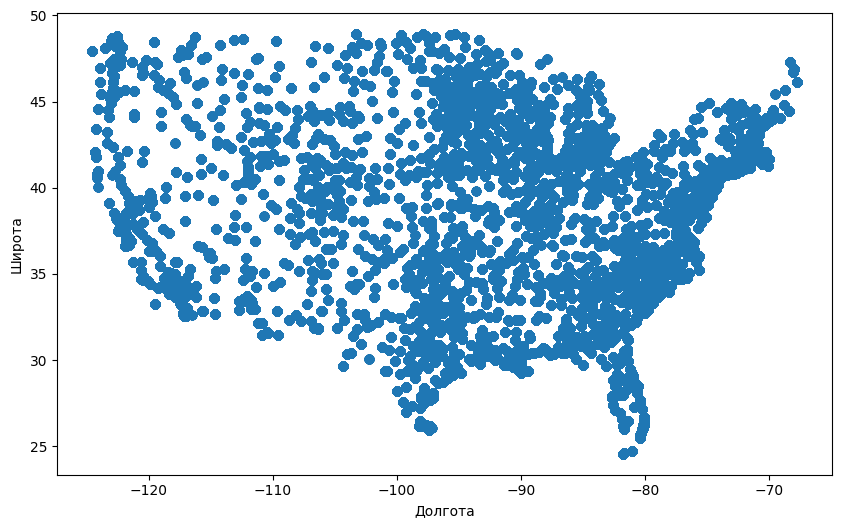


Рисунок 3 – координаты метеостанций

Для определения зависимости количества выпадавшего снега от широты была построена диаграмма рассеяния, показанная на рисунке 4.

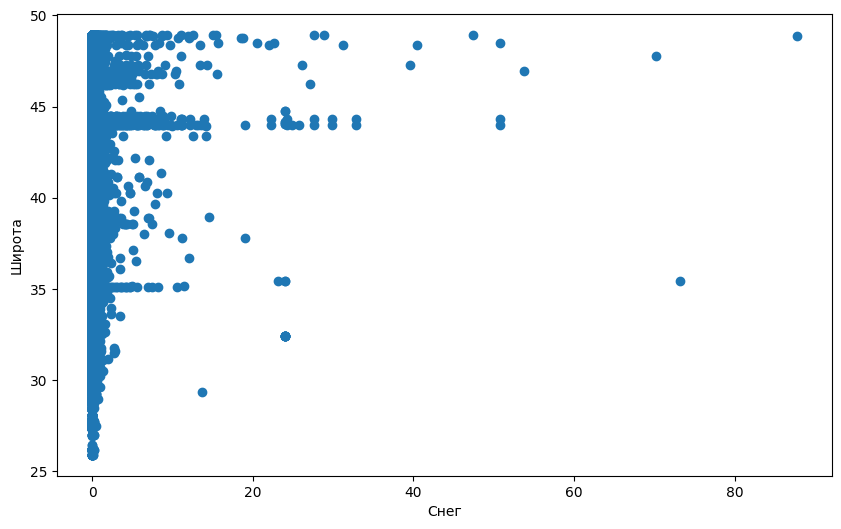


Рисунок 4 – зависимость выпадавшего снега от широты

По полученной зависимости видно, что в северной части страны количество снежных осадков больше, чем в южной.

# Выводы по первой главе

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.