**Задание**

1. Предварительная обработка данных

a. Визуализация значимых признаков (диаграммы рассеяния, ящики с усами, гистограммы)

b. Очистка данных (удаление пропусков, нормализация, удаление дубликатов)

c. Корреляция данных (матрица корреляций)

**Выполнение работы**

В качестве датасета была выбрана таблица данных по погоде в Австралии (рис.1). Информация по набору данных приведена на рисунке 2.

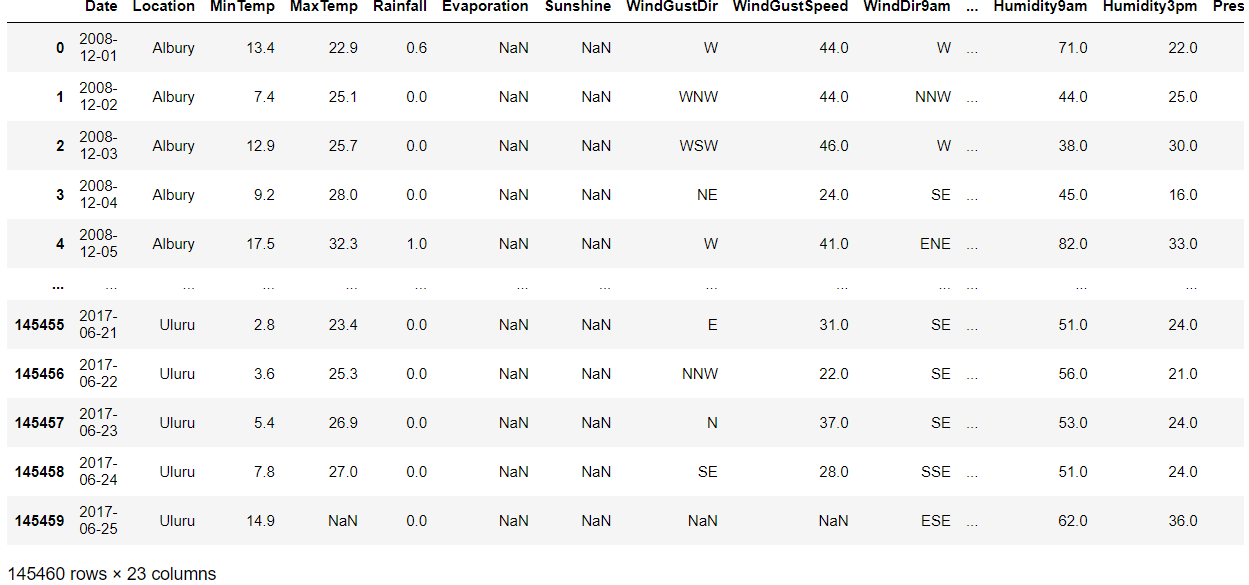


Рисунок 1

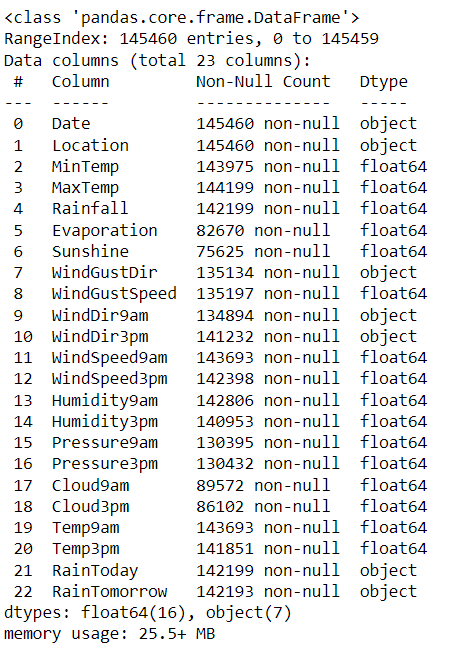


Рисунок 2

Гистограмма по признаку «MaxTemp» приведена на рис.3. По ней видно, что распределение по признаку близко к нормальному с мат.ожиданием между 20 и 25 градусами.

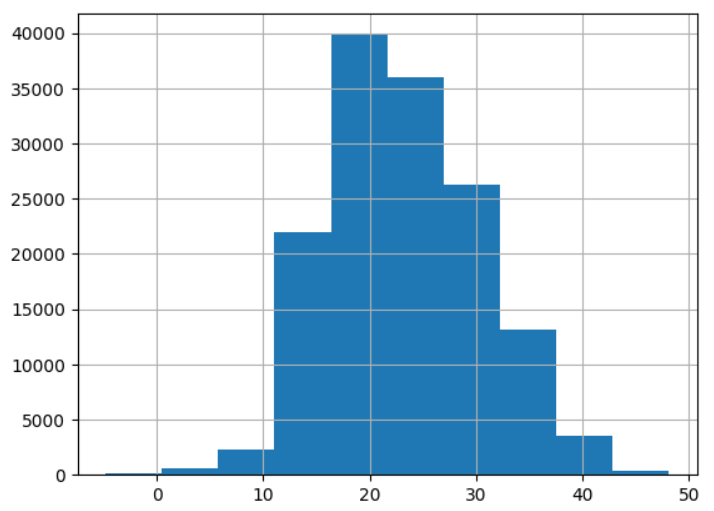


Рисунок 3

На рисунке 4 приведена гистограмма по признаку «Sunshine», показывающая распределение величины.

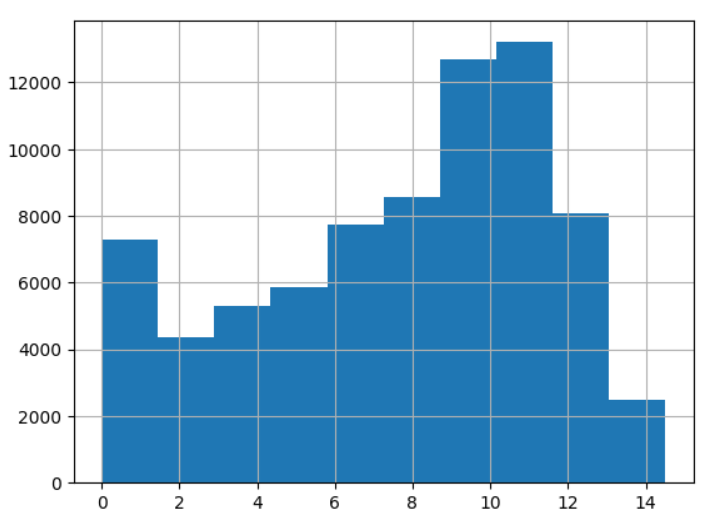


Рисунок 4

На рис. 5 приведена гистограмма по признаку «RainToday». По ней видно, что бόльшая часть дней, зафиксированных в датасете, прошла без осадков.

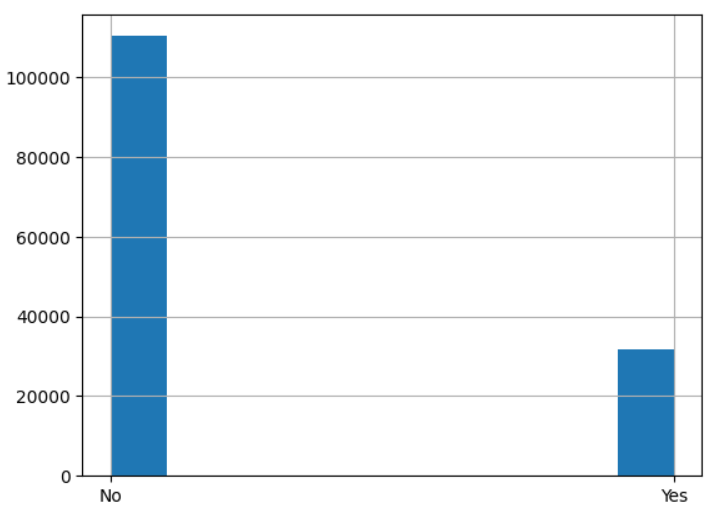


Рисунок 5

Таблица корреляции приведена на рис.6. По ней можно выбрать интересующие пары признаков.

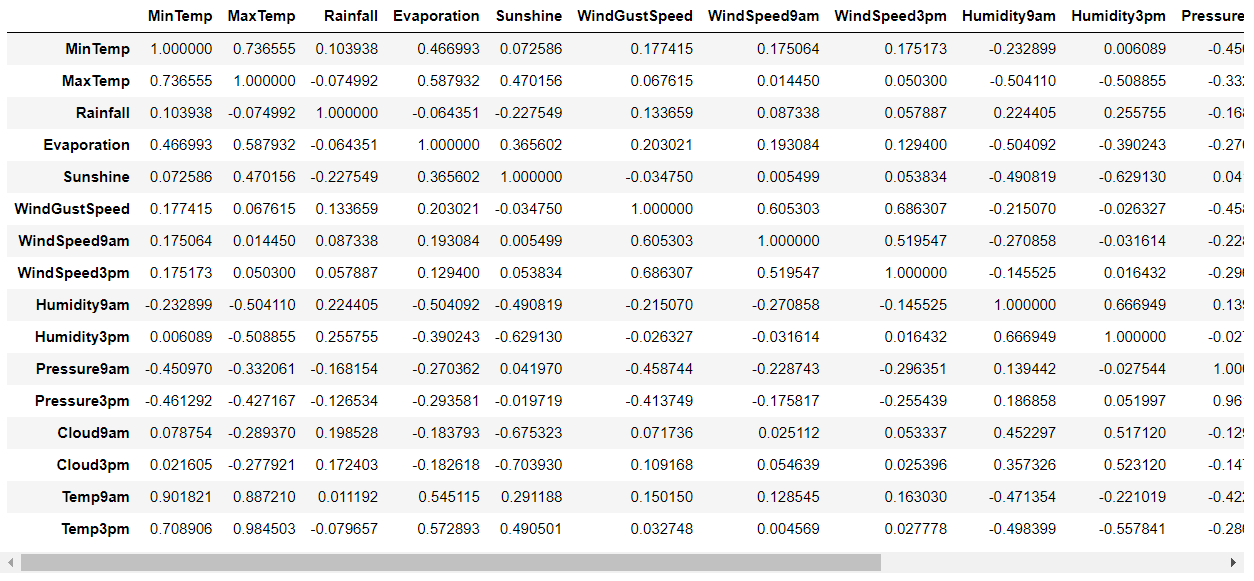


Рисунок 6

Диаграмма рассеяния на рис.7 показывает корреляцию между признаками «MaxTemp» и «Rainfall». Наблюдается некоторая зависимость между признаками: наибольшее количество осадков выпадает при температуре, близкой к средней, с некоторым отклонением. Также заметны выбросы (значения Rainfall, превышающие 350мм), они показывают случаи погодных отклонений от нормы.

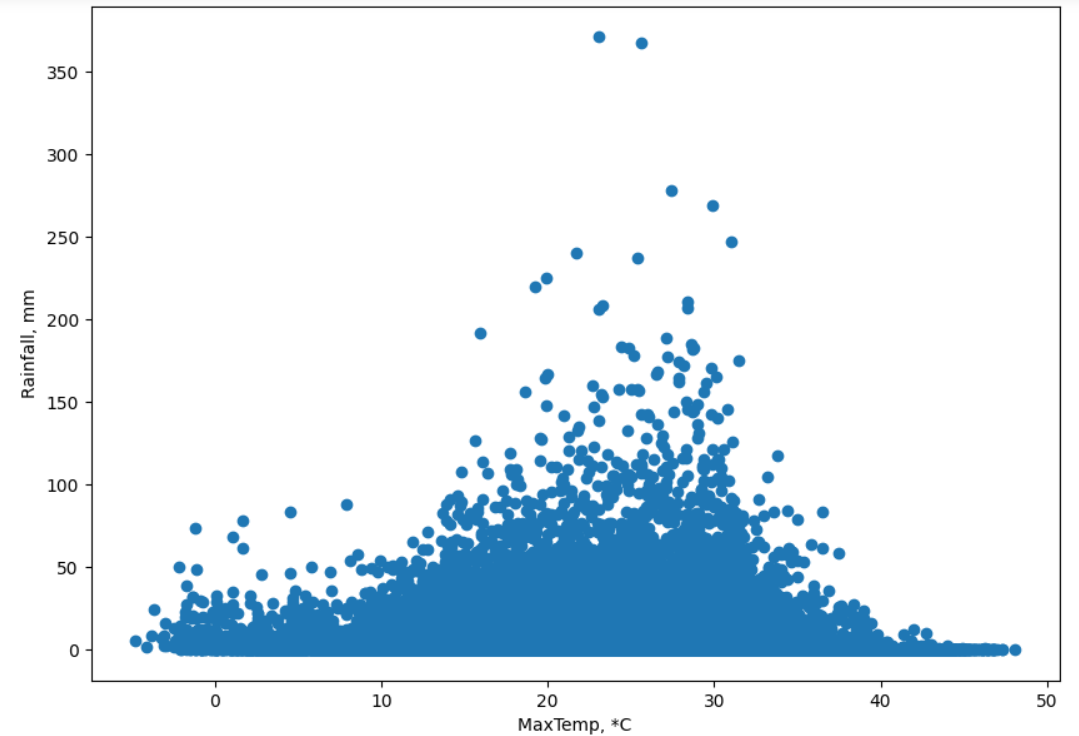


Рисунок 7

Диаграмма рассеяния на рис.8 показывает корреляцию между признаками «MaxTemp» и «Cloud9am». Видно, что признаки мало коррелируют. В целом, при низких температурах облачность выше среднего.

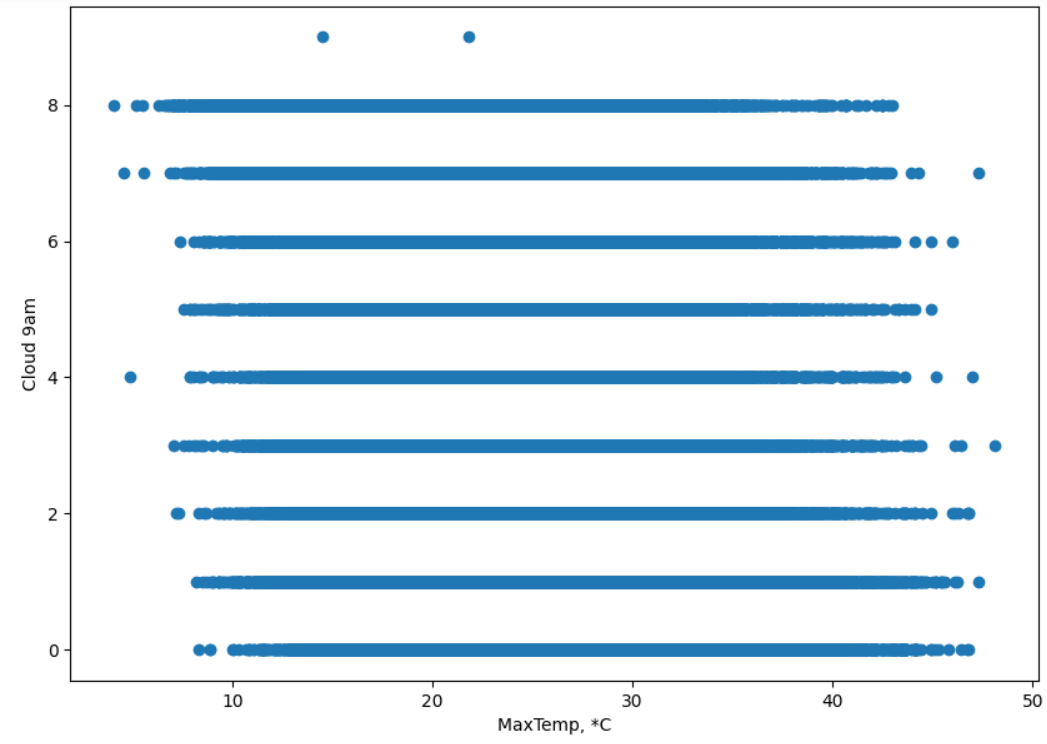


Рисунок 8

Диаграмма рассеяния на рис.9 показывает корреляцию между признаками «MaxTemp» и «Evaporation». Корреляция похожа на положительную. При температурах ниже 20\* коэффициент испарения меньше, затем повышеается.

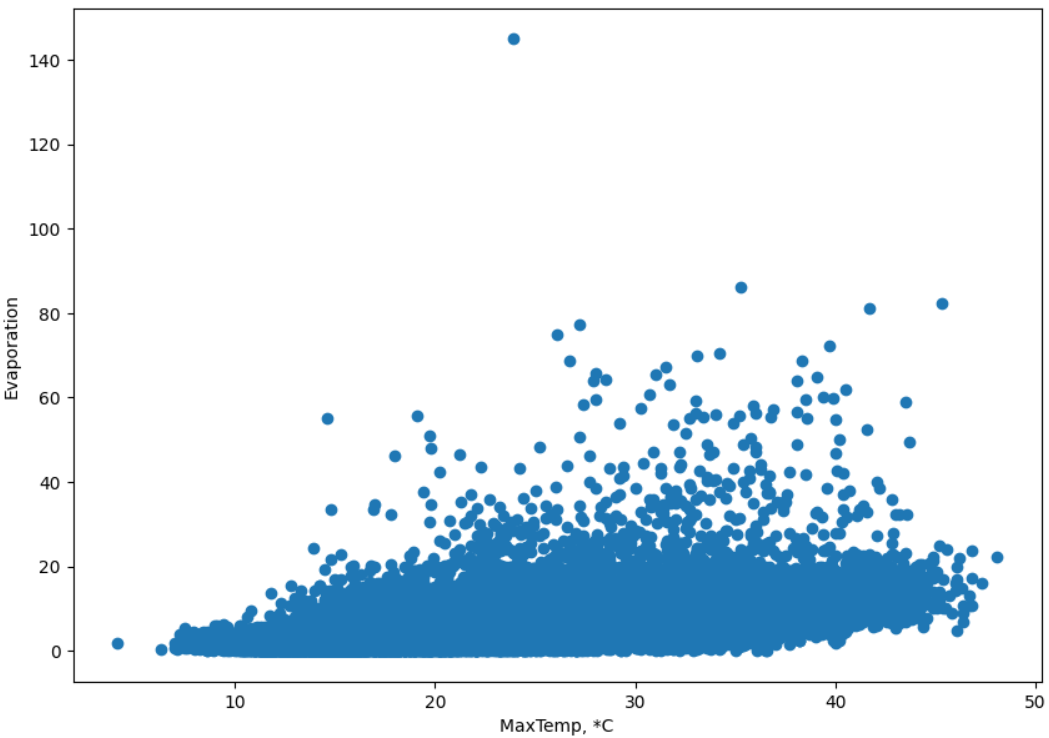


Рисунок 9

Также были построены графики «Ящики с усами», показывающие медиану, нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы. На рис.10 приведена диаграмма, показывающая значения статистических показателей для дней без осадков и с ними. По диаграмме видно, что средняя температура в ясные дни выше, чем в дождливые.

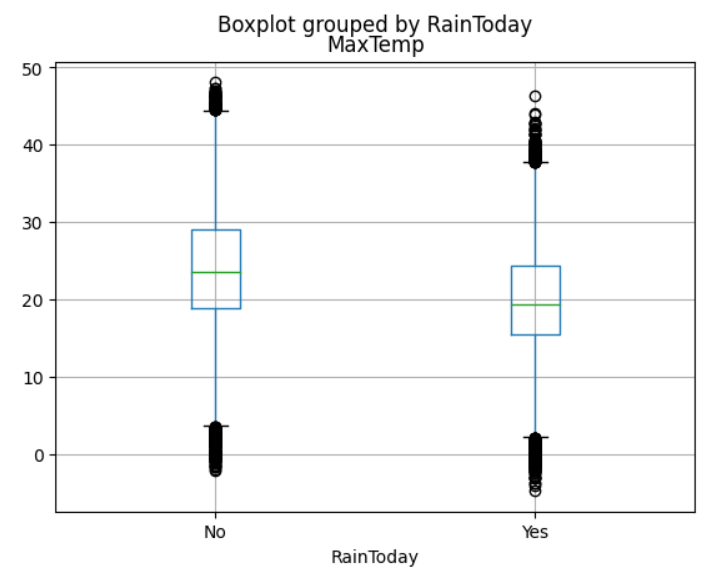


Рисунок 10

На рис.11 приведен ящик с усами, показывающий значения статистических показателей MaxTemp, группированных по локациям (городам Австралии). График сложно назвать наглядным ввиду большого количества локаций. Заметим, что есть локации, чье мат.ожидание сильно отлично от среднего по локациям. Найдем их (рис.12, рис.13) и построим ящик с усами для них (рис.14).

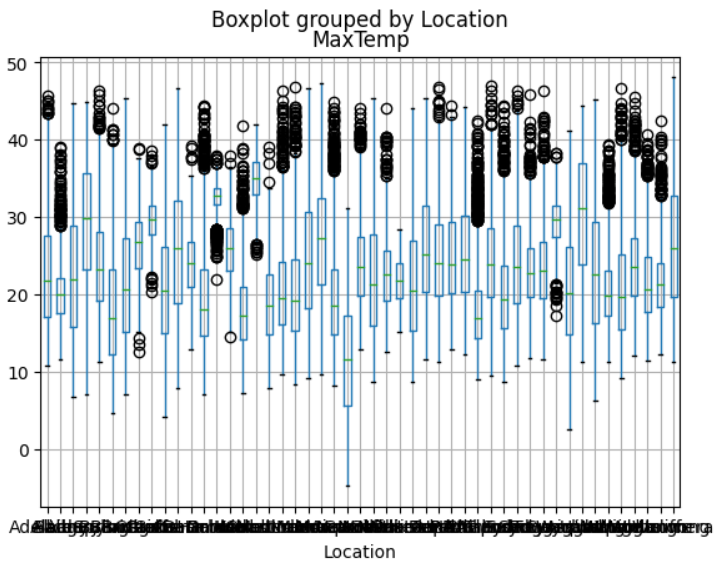


Рисунок 11

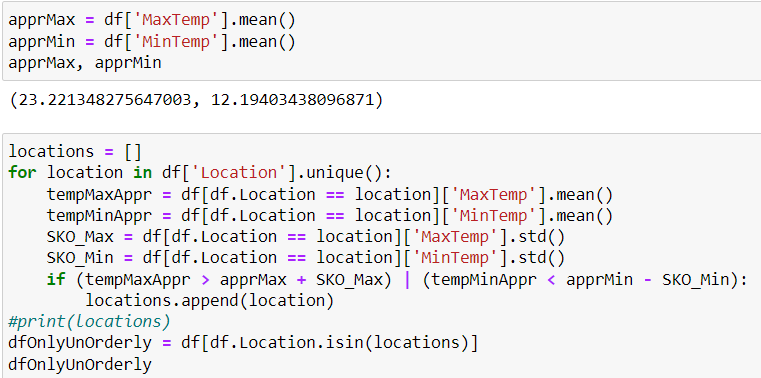


Рисунок 12

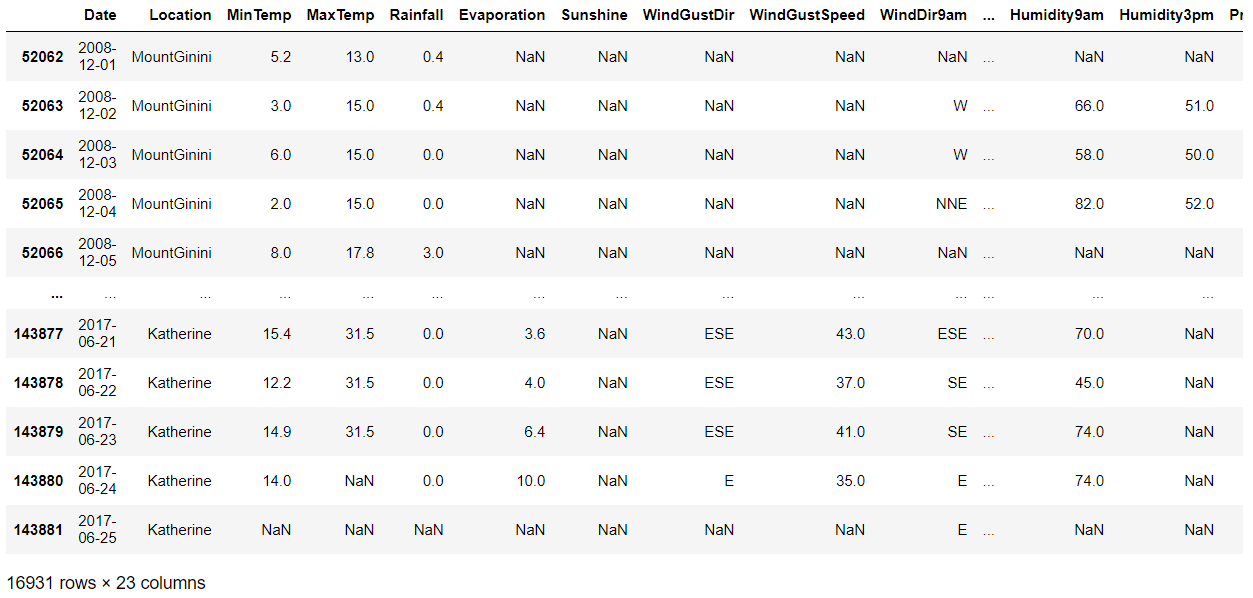


Рисунок 13

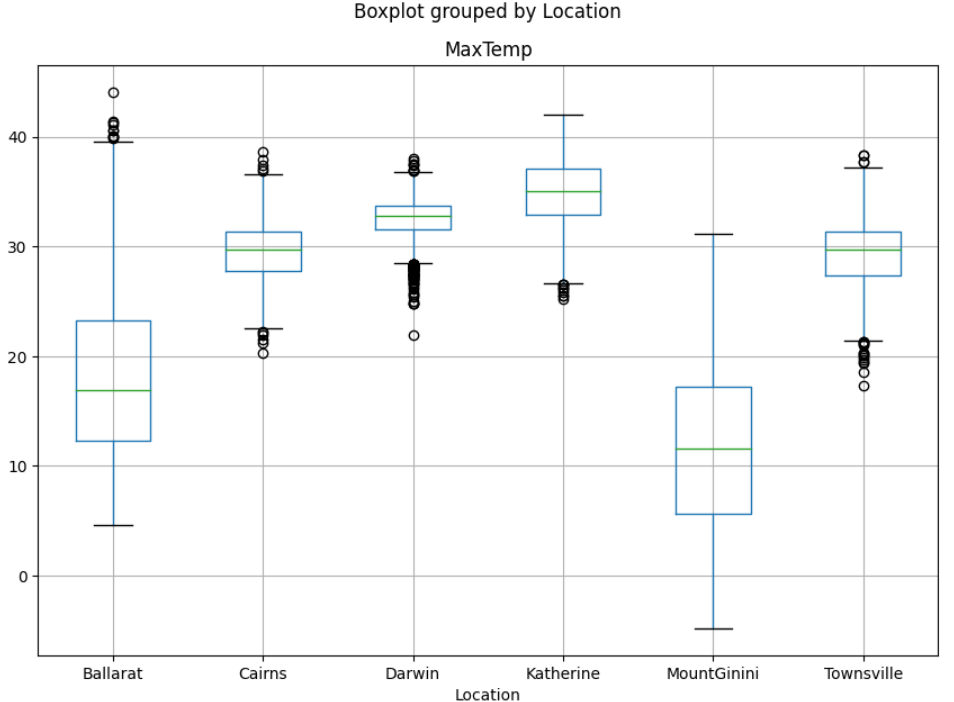


Рисунок 14

Узнаем, сколько ненулевых строк в массиве (рис.15).

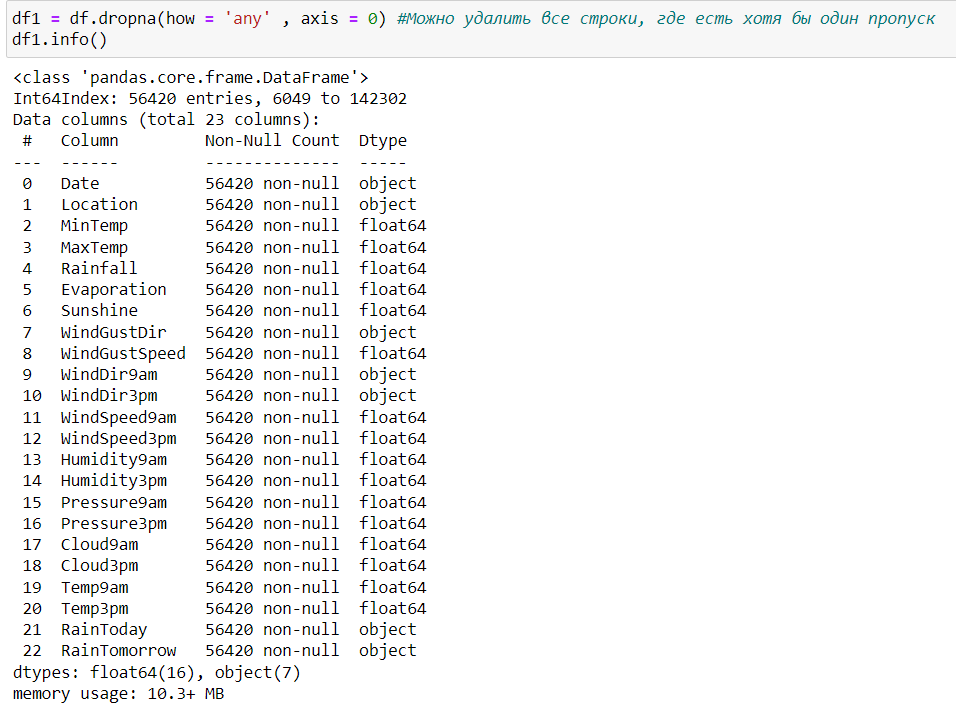
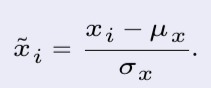
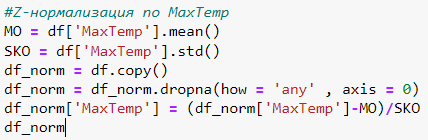


Рисунок 15

Z-нормализация

Для признаков MaxTemp и MinTemp проведем Z-нормализацию по формуле:  (рис.16, 17).



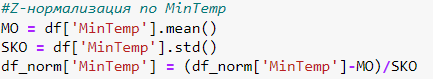


Рисунок 16



Рисунок 17

**Выводы**

В результате выполнения работы была проведена подготовка данных для последующего анализа; изучены способы визуализации данных и отобраны важные для анализа признаки.