Лабораторная работа №4
Анализ датасета

Пояснение к заданию:

1. Задание требуется выполнять jupyter notebook.
2. Для подключения к базе данных clickhouse можно использовать библиотеку clickhouse_driver.
3. Для подключения к базе можно использовать следующий код

```
from clickhouse_driver import Client

client = Client('oleg.orbita.work', port = 9000, user=student_ml,
password=CGRV8zemLGgsfdz7Uc6ZJeisGTcPQa, settings={'use_numpy':
True})
```

Задание

1. Загрузите данные из таблицы weather_history, которая находится в базе students_datas.
   Расшифровка:
   - idx – индекс ВМО,
   - year – год,
   - month – месяц,
   - day – день,
   - min_t – минимальная температура воздуха,
   - average_t – средняя температура воздуха,
   - max_t – максимальная температура воздуха,
   - rainfall – количество осадков.
2. Удалите столбец idx.
3. Используя метод info(), ответьте на вопросы:
   3.1. Есть ли в данных пропущенные значения?
   3.2. В каком столбце данных больше всего пропущенных значений?
4. В данных за какой год больше всего пропусков?
5. Объедините столбцы «Год», «Месяц» и «День» в один столбец «Дата» в формате гггг-мм-дд (2000-01-20). Данные в новом столбце должны иметь формат datetime;
6. Для каждого наблюдения рассчитайте размах температур (разность максимальной и минимальной суточных температур) и количество предшествующих ему дней без осадков (используйте циклы Python и условный оператор):
7. Определите самый длинный период засухи.
8. Для каждого года вычислите среднегодовую температуру и общее количество осадков. Запишите результаты в объекты Series.
   8.1. Какой год можно считать самым теплым? Какой самым холодным?
   8.2. В какой год выпало больше всего осадков? В какой меньше всего? Используя запись имя_серии.plot() вы можете построить график и

посмотреть как изменялась температура. С помощью имя_серии. .plot.bar() можно отобразить на столбиковой диаграмме количество осадков, выпавших в каждый год.

9. Выведете наблюдения, удовлетворяющие условиям:
   9.1. Средняя температура воздуха ниже -30 oC .
   9.2. Средняя температура воздуха выше 27 oC и количество дней без осадков больше 3. Полезные функции и методы

Одну и ту же задачу можно решить несколькими способами. Эти функции и методы могут вам понадобиться:

➢ .head() – отобразить несколько первых строк DataFrame;
➢ .info() – информация о DataFrame;
➢ .drop() – удалить строки или столбцы;
➢ .dtypes – узнать тип данных в столбце;
➢ .astype(), to_datetime(), .to_numeric() – изменить тип данных;
➢ .isnull().sum() – вычислить количество пропущенных значений в каждом столбце;
➢ .max(), .min(), .mean() – максимум, минимум, среднее значение;
➢ pd.Grouper(), .groupby() – группировка наблюдений;
➢ .agg() – агрегирование наблюдений;
➢ .tuncate() – логическая индексация (можно использовать даты!);
➢ Уже знакомые вам операторы тоже работают с pandas. Действие (или условие) выполняется (или проверяется) для каждого наблюдения. Так, например, чтобы найти разность между двумя числовыми характеристиками (столбцами) по всему набору данных, используйте оператор «-»:
имя_ DF [«новый_столбец»] = имя_ DF [«столбец_1»] - имя_ DF [«столбец_2»] Если необходимо найти сумму двух числовых характеристика или склеить строковые значения двух столбцов – используйте «+». Если вы хотите получить значения, которые больше заданного числа, например 5, используйте запись имя_ DataFrame [«имя_столбца»] > 5
конструкция: ➢ Шпаргалка: https://smysl.io/blog/pandas/.