**Spark Practice**

CONTENTS

[1. Prerequisites 2](#_Toc126675179)

[2. Task 2](#_Toc126675180)

Github: https://github.com/alikhan990612/Alikhan\_Zhaksylykov\_spark\_hw

PREREQUISITES



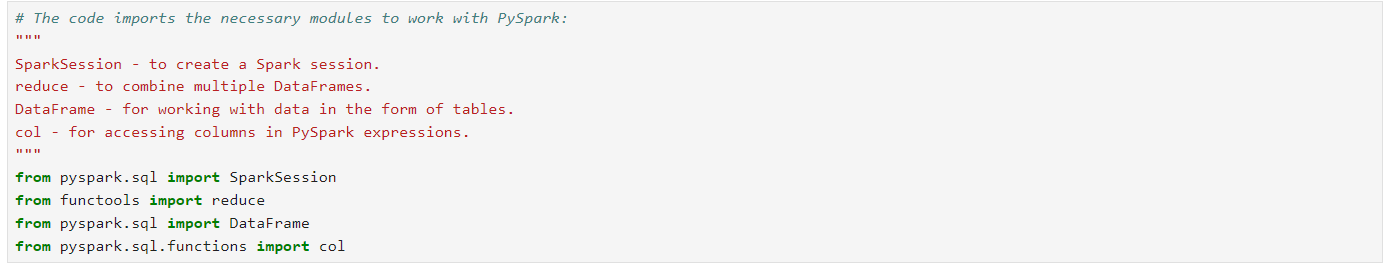
LEARN.

GROW. BE EXCELLENT.

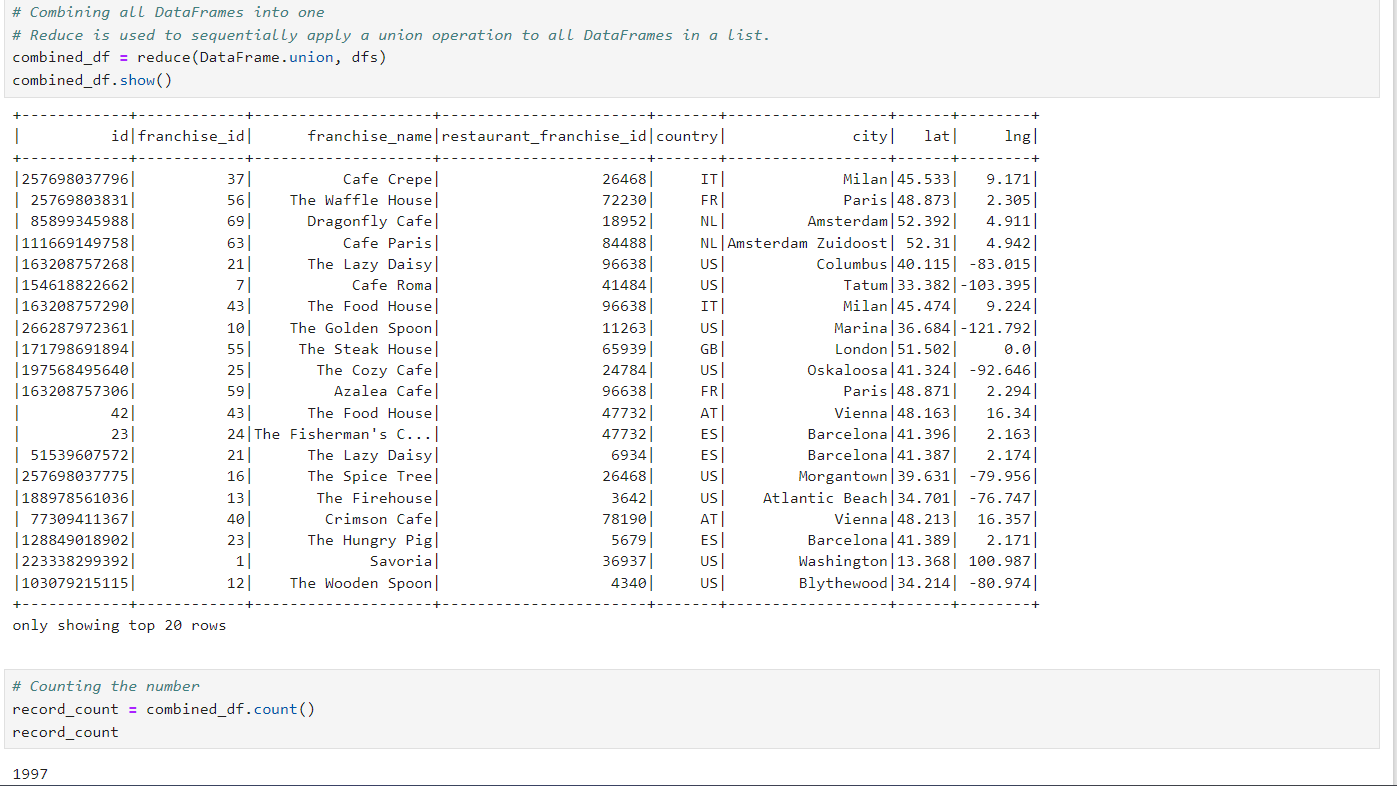
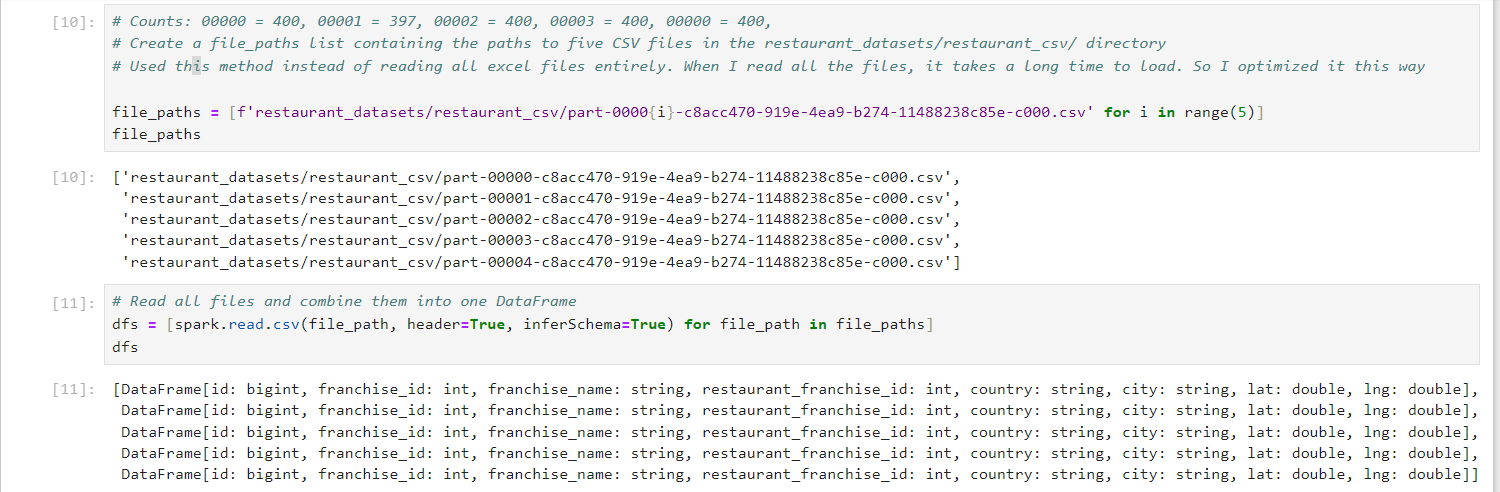
* Install Spark locally using one of the methods described [here](https://spark.apache.org/downloads.html) or in Docker.
* Create a Spark ETL job to read data from a local storage. You can find the data in the Spark Practice—Dataset file on the page with the task description.

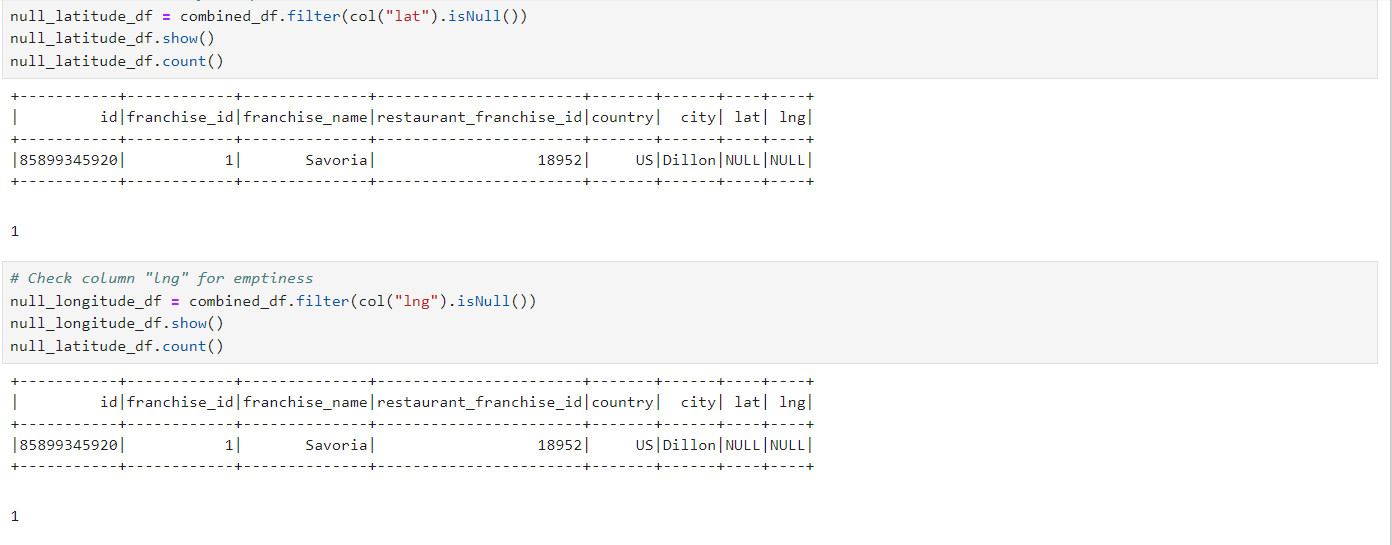
TASK

* Check restaurant data for incorrect (null) values (latitude and longitude). For incorrect values, map latitude and longitude from the [OpenCage Geocoding API](https://opencagedata.com/api" \t "_blank) in a job via the REST API.

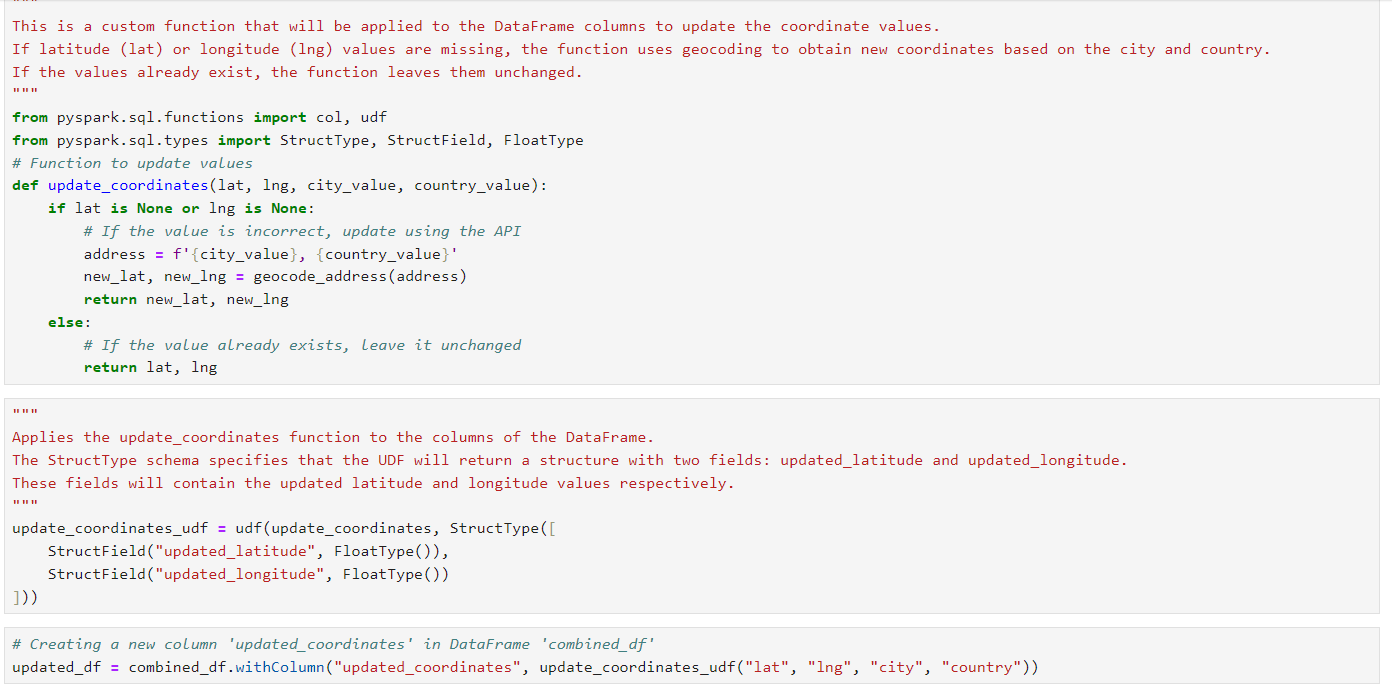


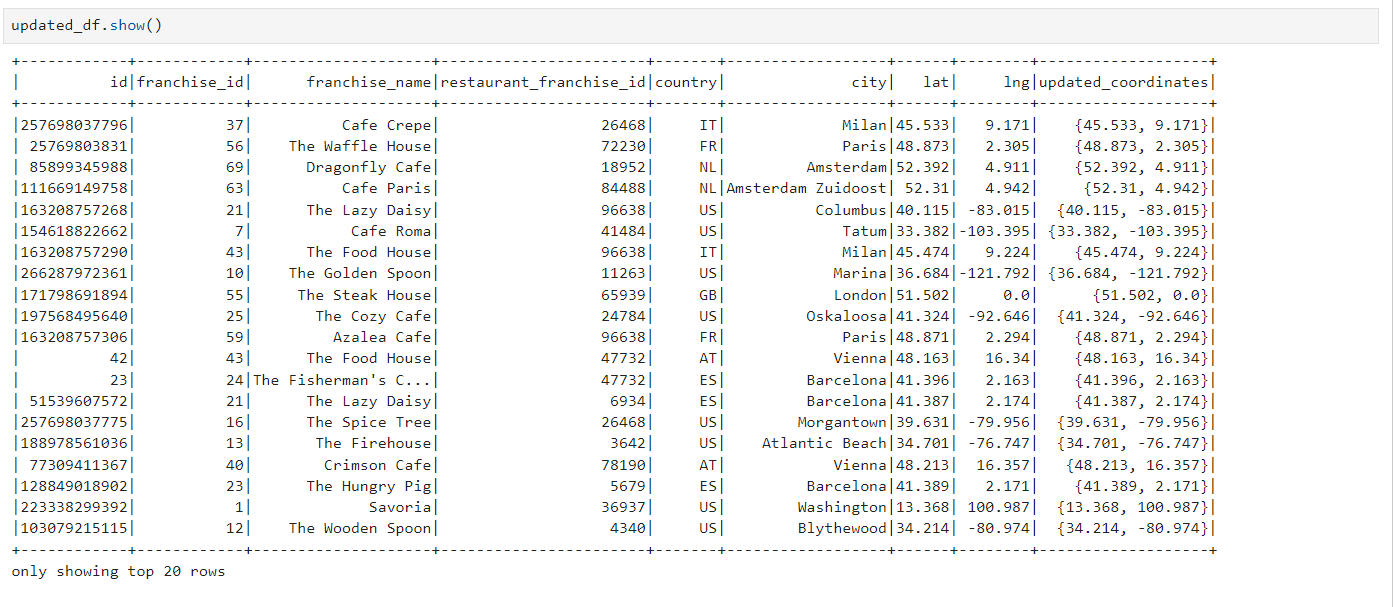


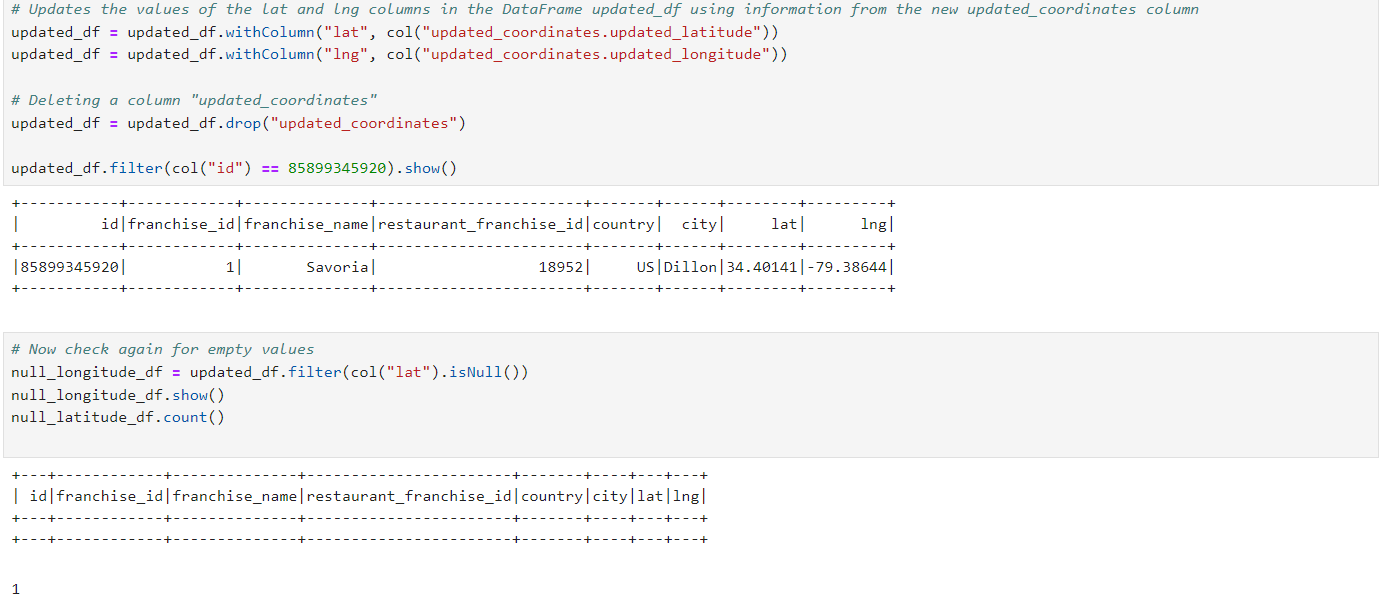




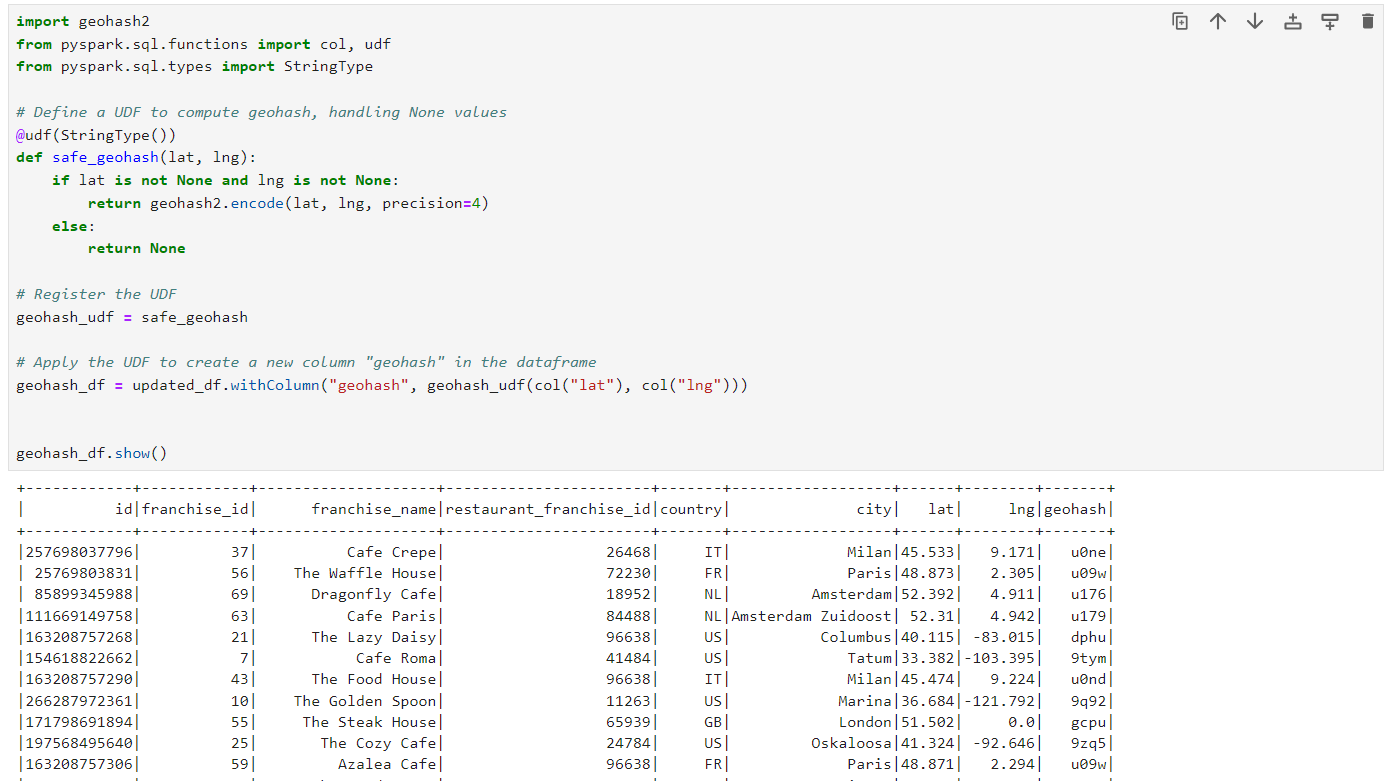


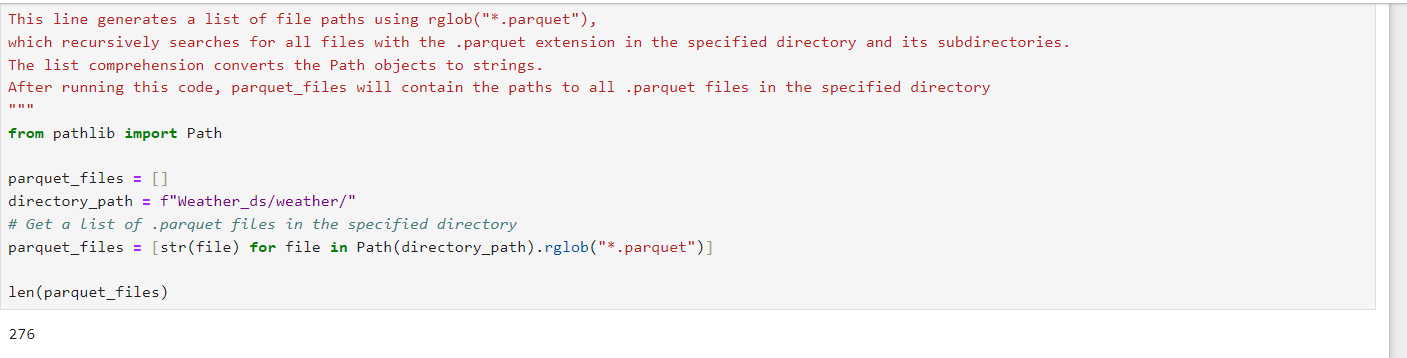


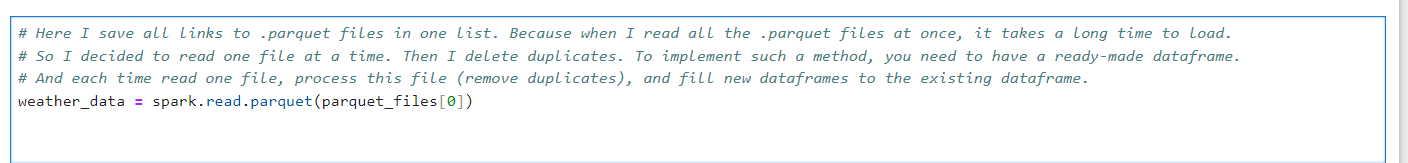


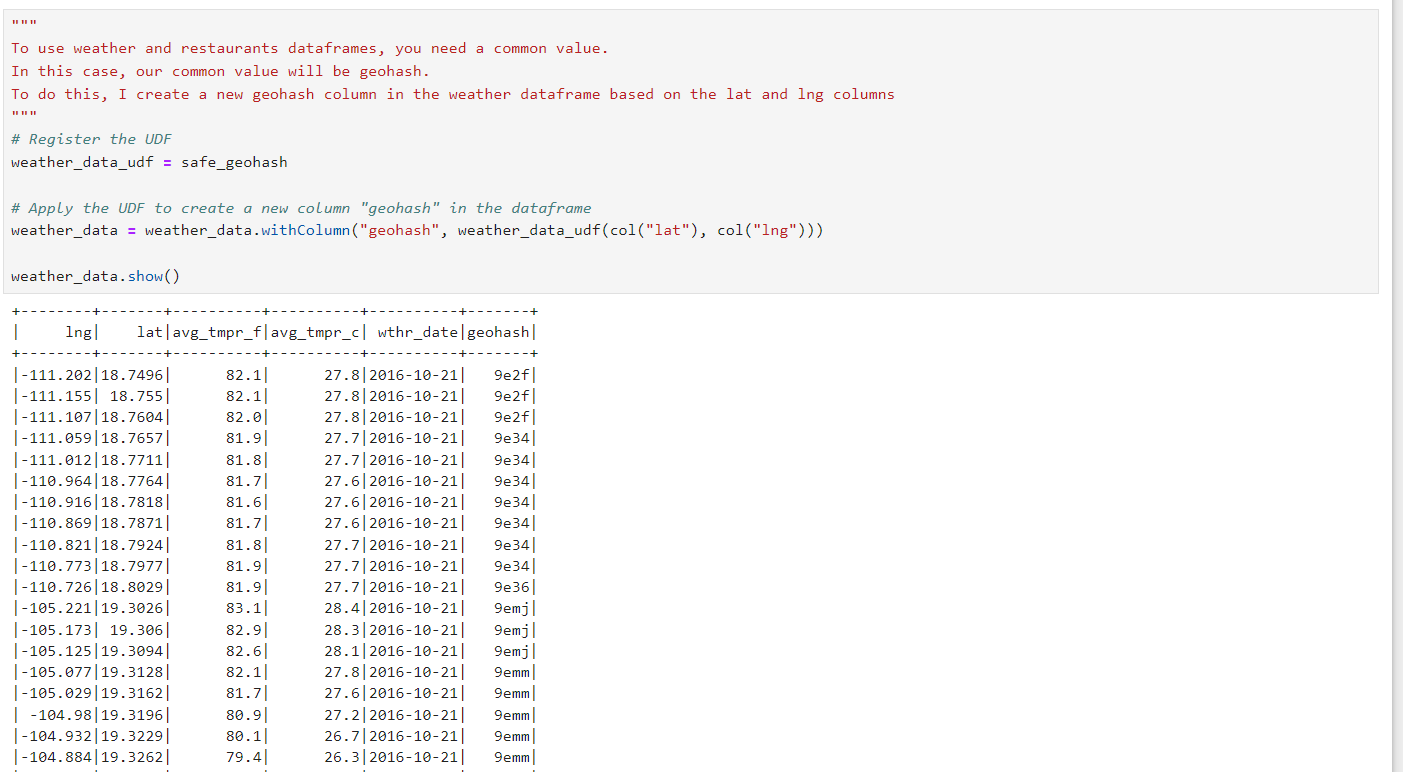


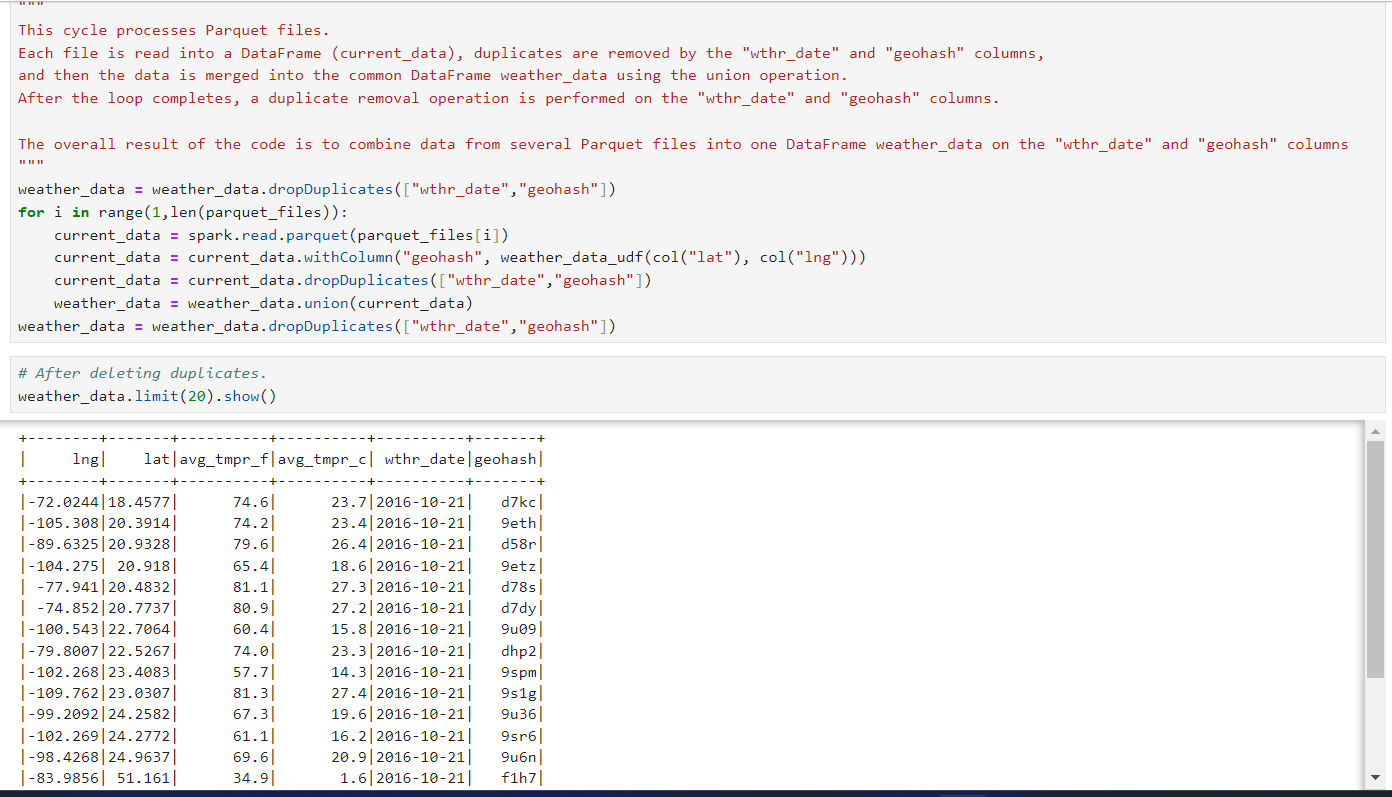
* Generate a geohash by latitude and longitude using a geohash library like geohash-java. Your geohash should be four characters long and placed in an extra column.



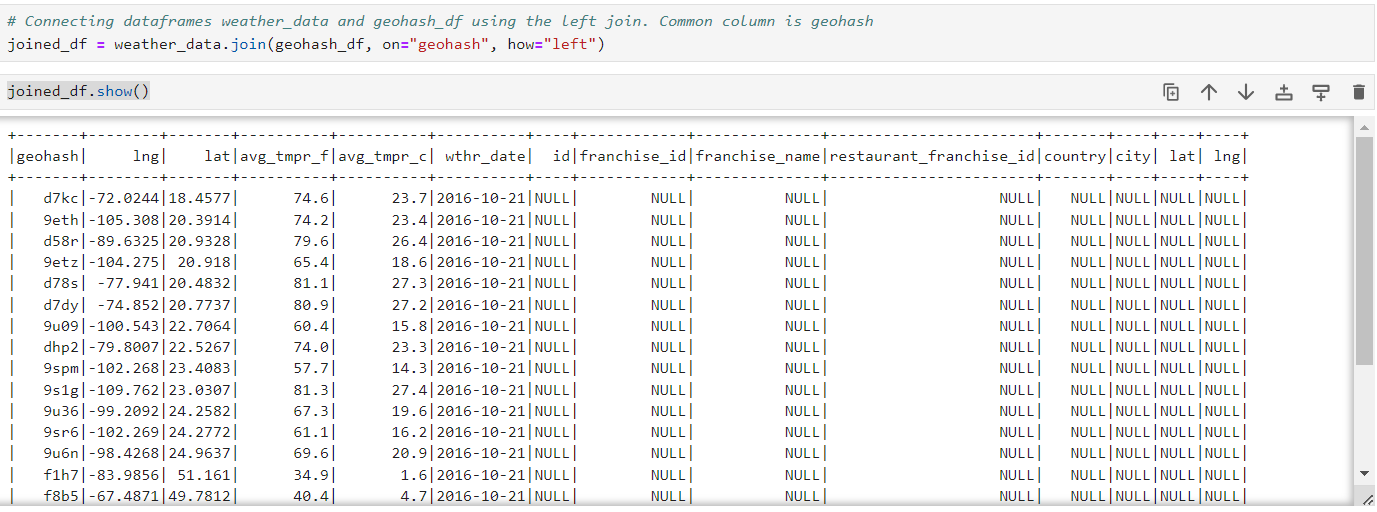


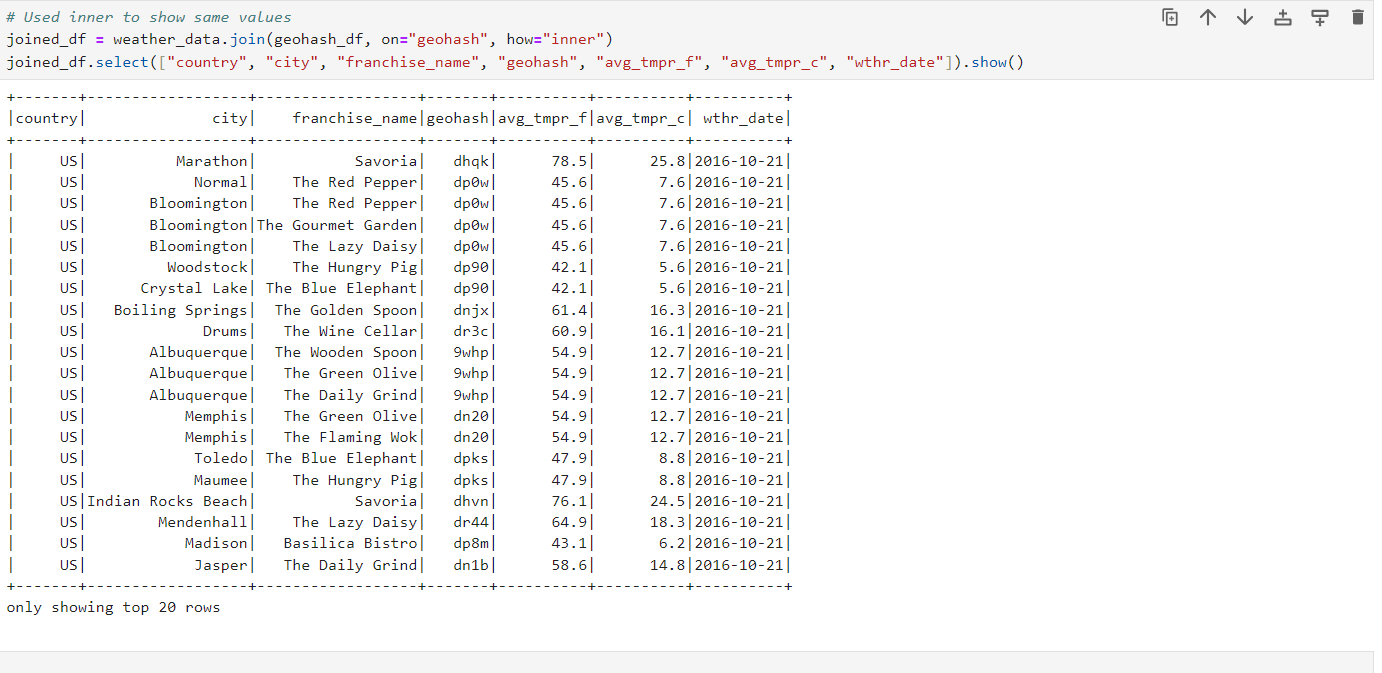






* Left-join weather and restaurant data using the four-character geohash. Make sure to avoid data multiplication and keep your job idempotent.





**Note**: Development and testing should be done locally in your IDE environment.

* Store the enriched data (i.e., the joined data with all the fields from both datasets) in the local file system, preserving data partitioning in the parquet format.

**You are expected to:**

* Upload the source code and implement tests
* Upload your fully documented homework with screenshots and comments in the task Readme MD file with the repo link.