# Table of Contents

Contents

[Table of Contents 1](#_Toc133872711)

[Working with restaurant dataset 2](#_Toc133872712)

[Creating Spark ETL JOB session 2](#_Toc133872713)

[Read restaurant\_csv dataset 2](#_Toc133872714)

[Get Nulls in dataset. Also I defined geocode UDF in order to geocode missing coordinates. 3](#_Toc133872715)

[Applying Geocode UDF 4](#_Toc133872716)

[Save dataframe 4](#_Toc133872717)

[Applying Geohash 5](#_Toc133872718)

[Working with Weather dataset 6](#_Toc133872719)

[Reading Weather dataset and applying geohash UDF 6](#_Toc133872720)

[Deleting duplicated rows 7](#_Toc133872721)

[Saving weather cleaned dataset 7](#_Toc133872722)

[Joining Restaurant and Weather Datasets 8](#_Toc133872723)

[Join Restaurant and Weather datasets 8](#_Toc133872724)

[Cleaning joined dataset from Nulls 8](#_Toc133872725)

[Further Manipulations with joined dataset. 9](#_Toc133872726)

[Leave joined dataset as it is. 9](#_Toc133872727)

[Droping duplicates for geohash and wthr\_date together. 9](#_Toc133872728)

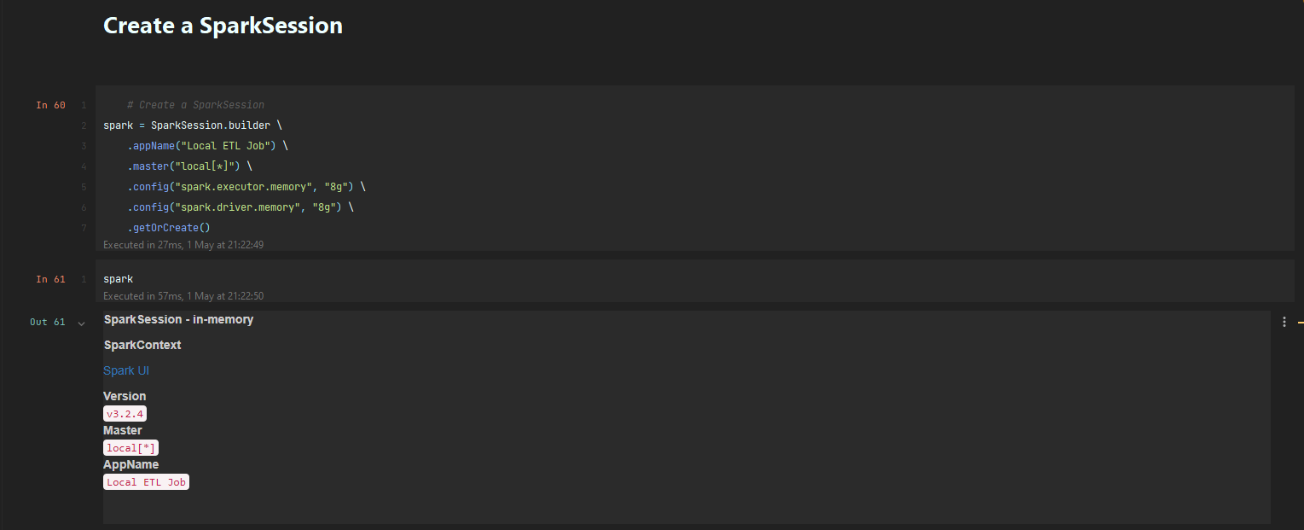
[Saving datasets. 10](#_Toc133872729)

# Working with restaurant dataset

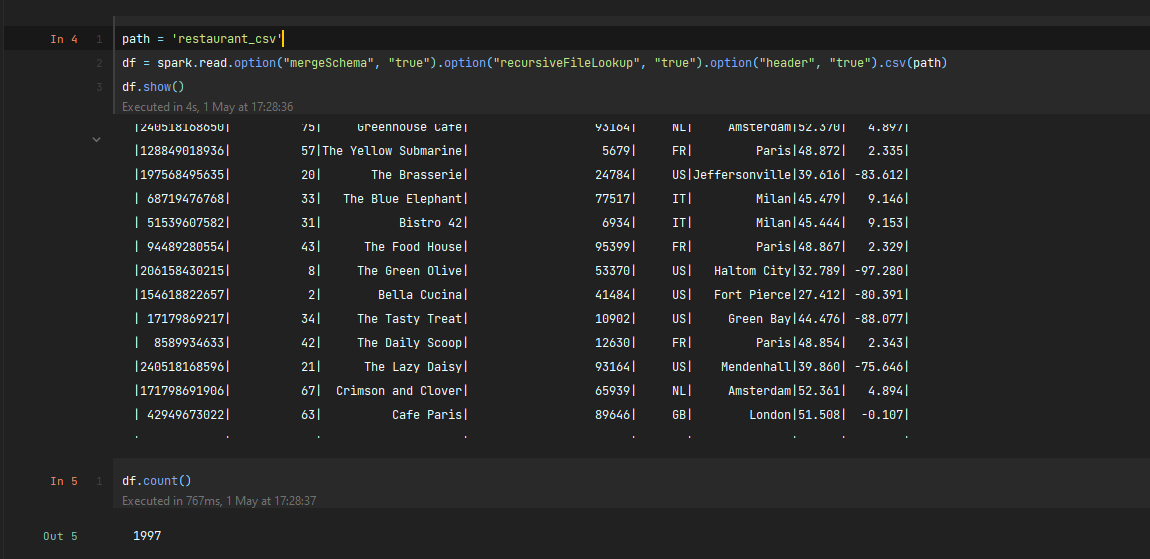
Install Spark locally using one of the methods described [here](https://spark.apache.org/downloads.html) or in Docker.

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.

## Creating Spark ETL JOB session



## Read restaurant\_csv dataset



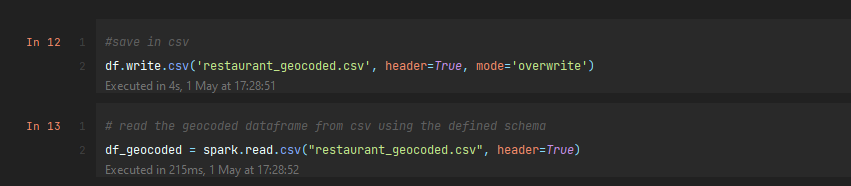
## Get Nulls in dataset. Also I defined geocode UDF in order to geocode missing coordinates.

## Applying Geocode UDF

After applying we can see that there is no Nulls in lat and lng.

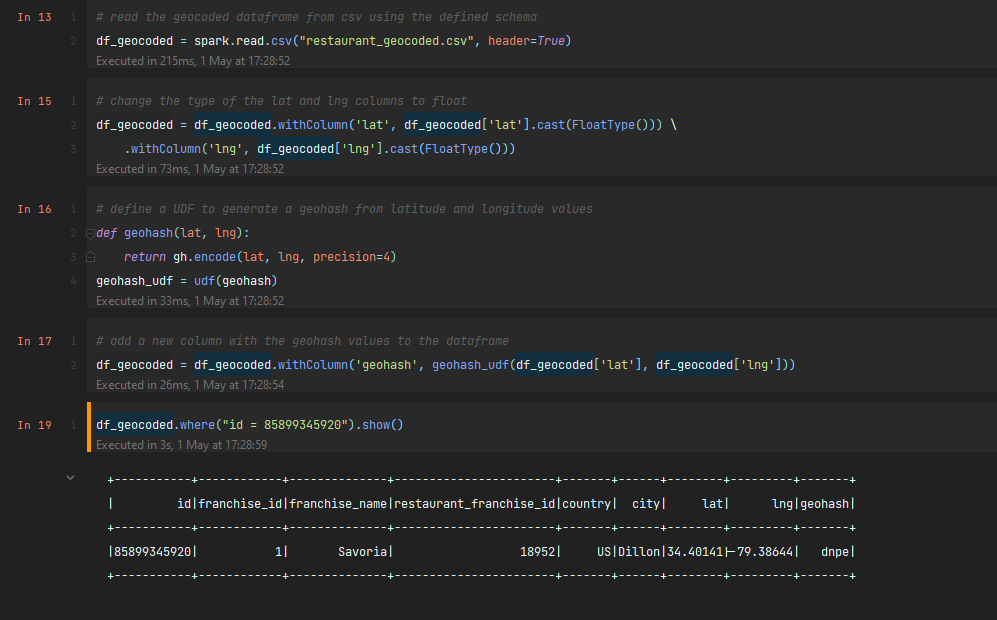
## Save dataframe

I saved dataframe before applying hash, in order to make easier if something goes wrong



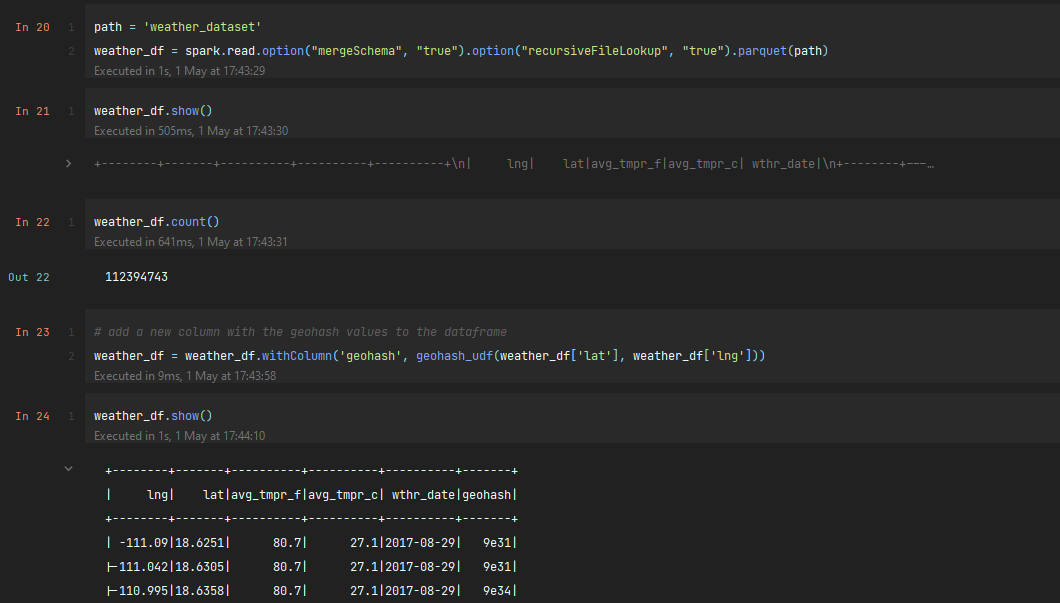
## Applying Geohash

I also created geohash UDF in order to hash my geocode.



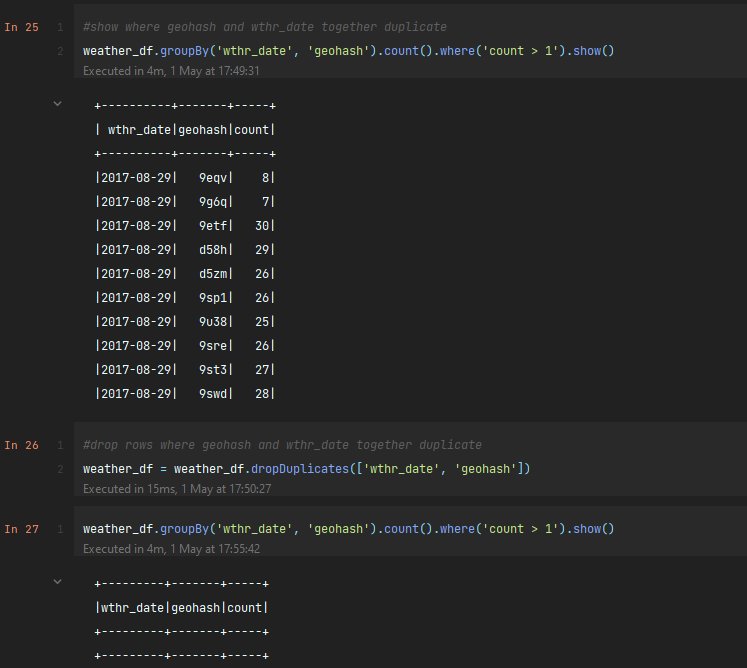
# Working with Weather dataset

## Reading Weather dataset and applying geohash UDF



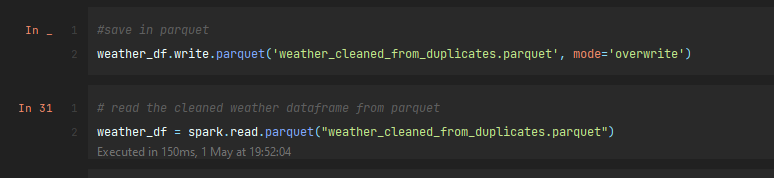
## Deleting duplicated rows

I took wthr\_date and geohash as unique together. After cleaning duplicates we can see that there is only unique value for wthr\_date and geohash together



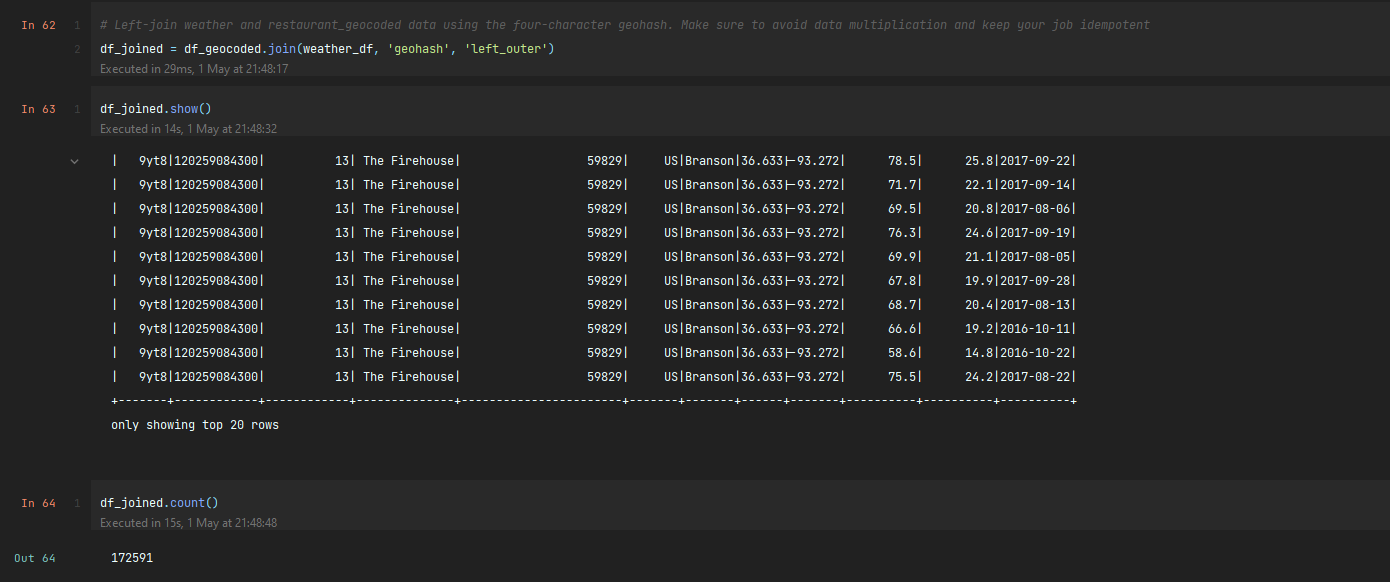
## Saving weather cleaned dataset

I saved cleaned dataset. Then read it again from file in order to make easier further steps



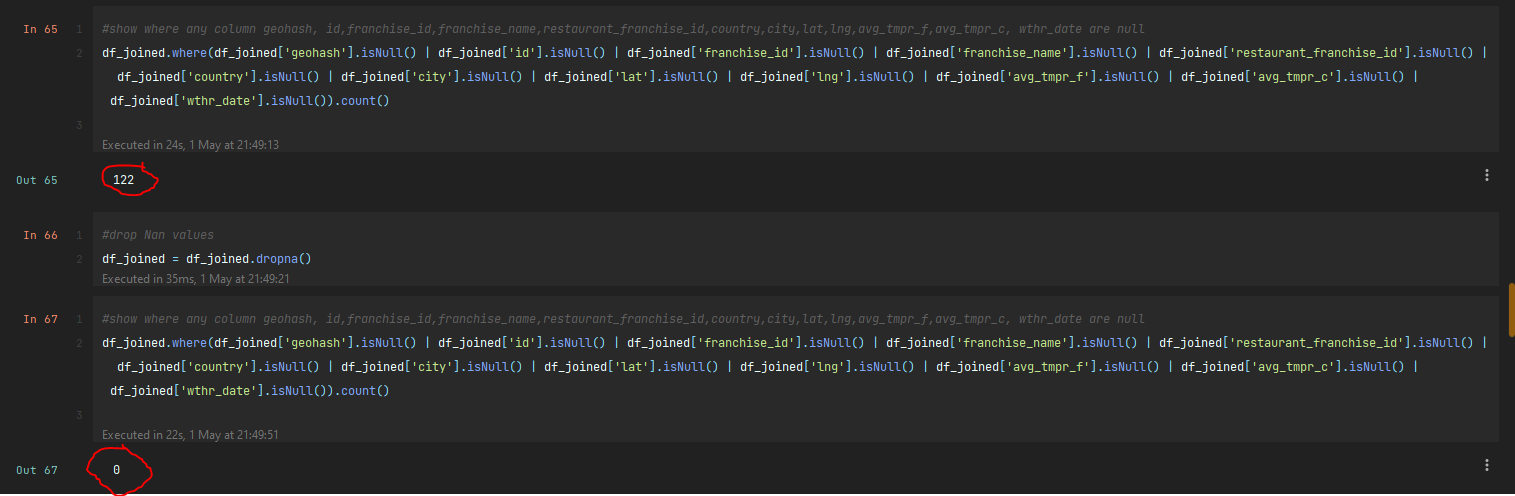
# Joining Restaurant and Weather Datasets

## Join Restaurant and Weather datasets



## Cleaning joined dataset from Nulls

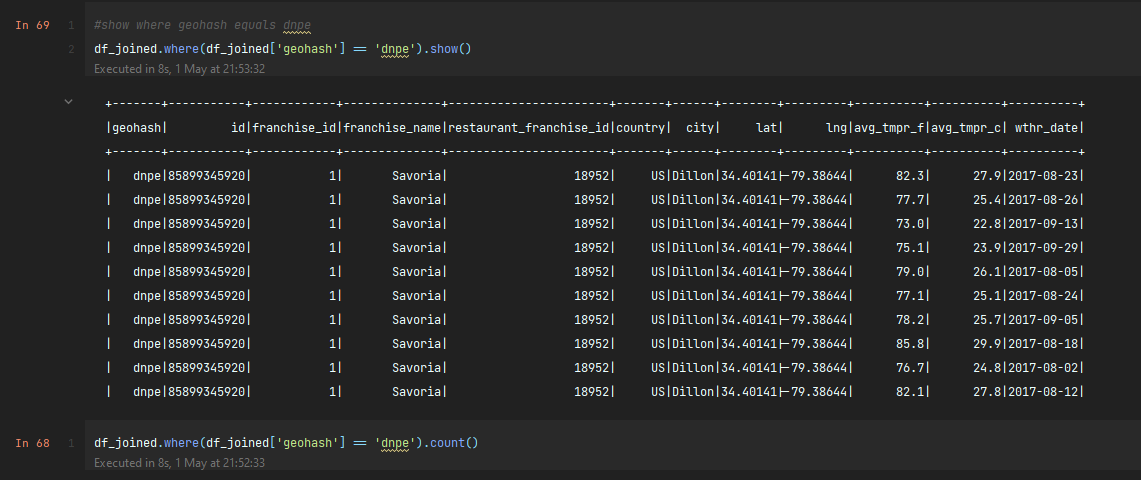
After dropping null are NaN counts changed from 122 to 0



## Further Manipulations with joined dataset.

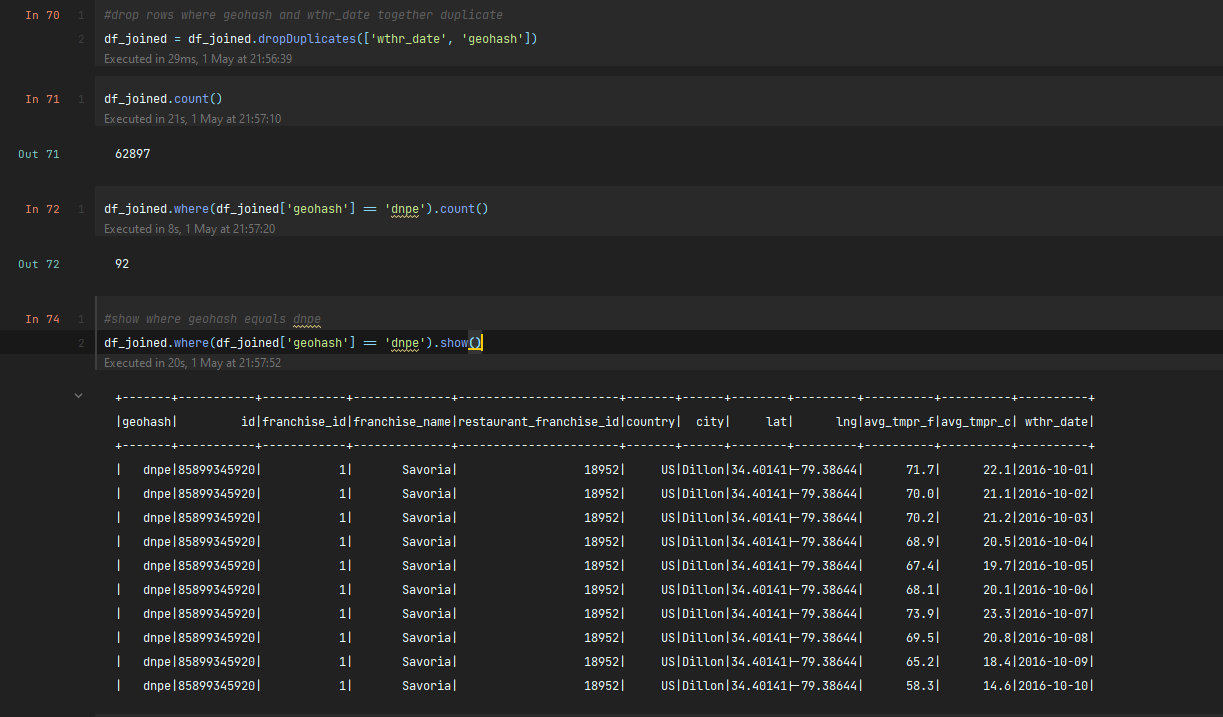
### Leave joined dataset as it is.

Then we will have same geohash and weather but different franchise.



### Droping duplicates for geohash and wthr\_date together.

In that case I will have 92 positions for each geohash



## Saving datasets.

I saved datasets in two variants:

1. Where I DID NOT DROPPED duplicates for wthr\_date and geohash.
2. Where I DROPPED duplicates for wthr\_date and geohash.

