**ETL Project - Columbus City Parking Violations**

**Members:**

Sai, Deepen, Pratikchhya, Pavana

**Introduction:**

The parking violations might seem minor, for example like parking in bike lanes, bus lanes, sidewalks, taking more than a spot or blocking a driveway, etc... But in the end, these minor violations can put pedestrians, cyclists, and other riders in danger and make it harder for city to be safer.

Columbus City Parking Violations and Ticket Status dataset covers the parking violations identified by Parking Enforcement Officer (PEO) and the tickets issued for those violations. Also, the data identifies the status of the ticket (e.g., paid, dismissed, etc.,) This dataset covers years 2013 through 2018. The Meters ID dataset contains identifiable information relating to each parking ticket issued based on Meter Ids. Metadata file provides information on the column names within the Columbus City Parking dataset.

* Size of the data is roughly over 235 MB.
* Data ranges from 2013-2018.
* The data is given in a CSV format and consists of 99 columns all of which add up to more than 745,822 rows of data.
* Most of the columns; however, were not usable as they had null values or values for which did not correlate with meaningful data.
* In the end, we ended up with 16 columns and 19,725 rows with two years of data 2017-2018.

Our group chose to combine Columbus City Parking Violations dataset with Meter ID dataset. Final table has the following columns: Ticket Id (Primary Key), Issue Date, Base fine amount, Unique system-generated number to identify license plate-based accounts, Car make, Issue Time, Latitude and Longitude of the incident, Total number of tickets, Officer badge number, Amount received in payment for ticket, Location of infraction denoted on the ticket, Internal violation code, Meter ID’s denoted on the ticket, Meter Status and Meter rate per hour.

**Process**

Using the ETL processes, the following tasks were done:

**Extract:**

* Extracted Columbus City Parking Violations and Ticket Status 2013-2018 -- COLUEXTRACT

Source:

<https://discovery.smartcolumbusos.com/dataset/conduent/160c98a1_ad56_4658_8553_5ee8e7d0d953>

* Extracted Parking Meters

Source:

<http://opendata.columbus.gov/datasets/parking-meters/data?geometry=-83.245%2C39.931%2C-82.812%2C40.023>

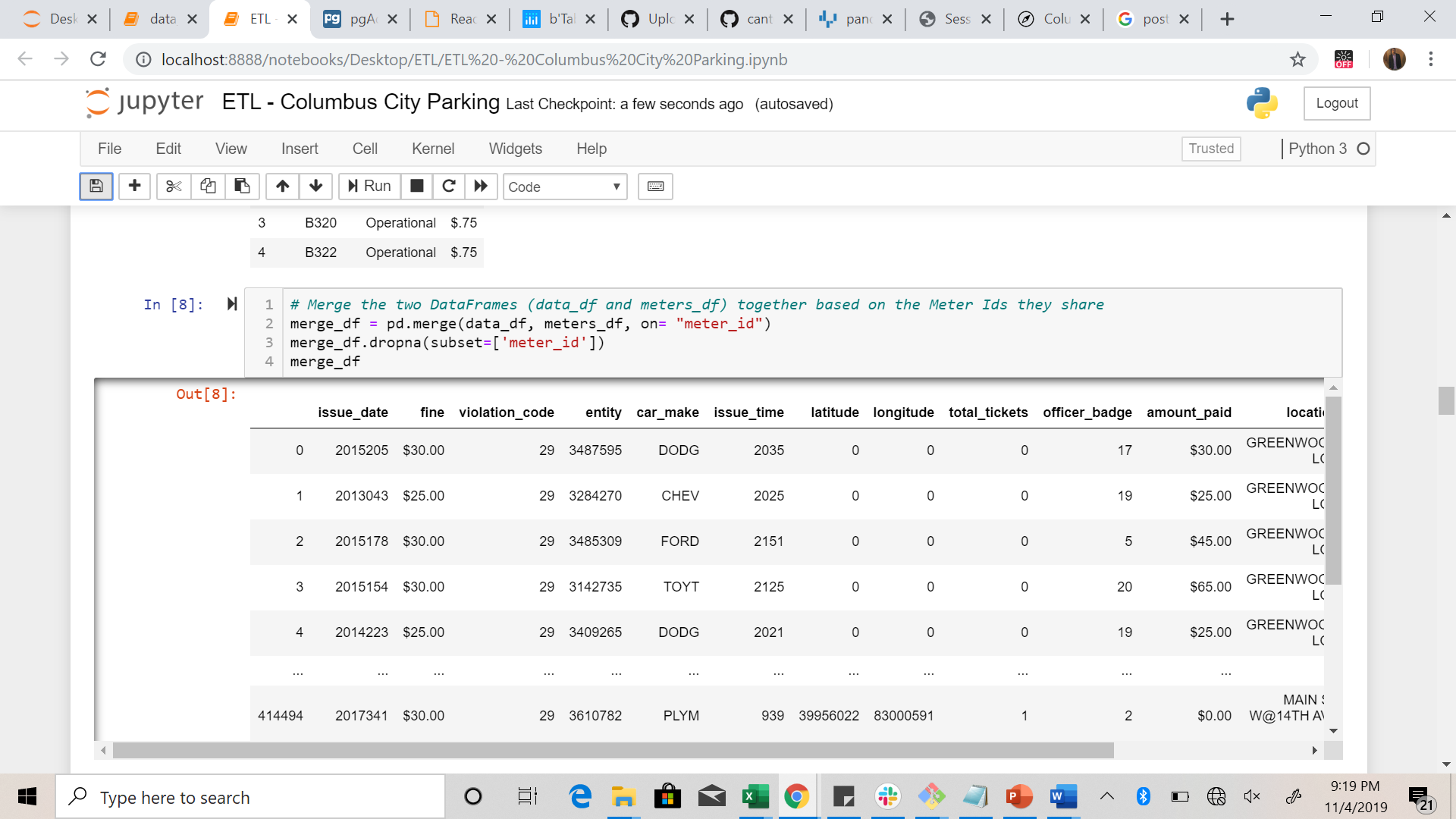
* Extracted the metadata file to decode the column names

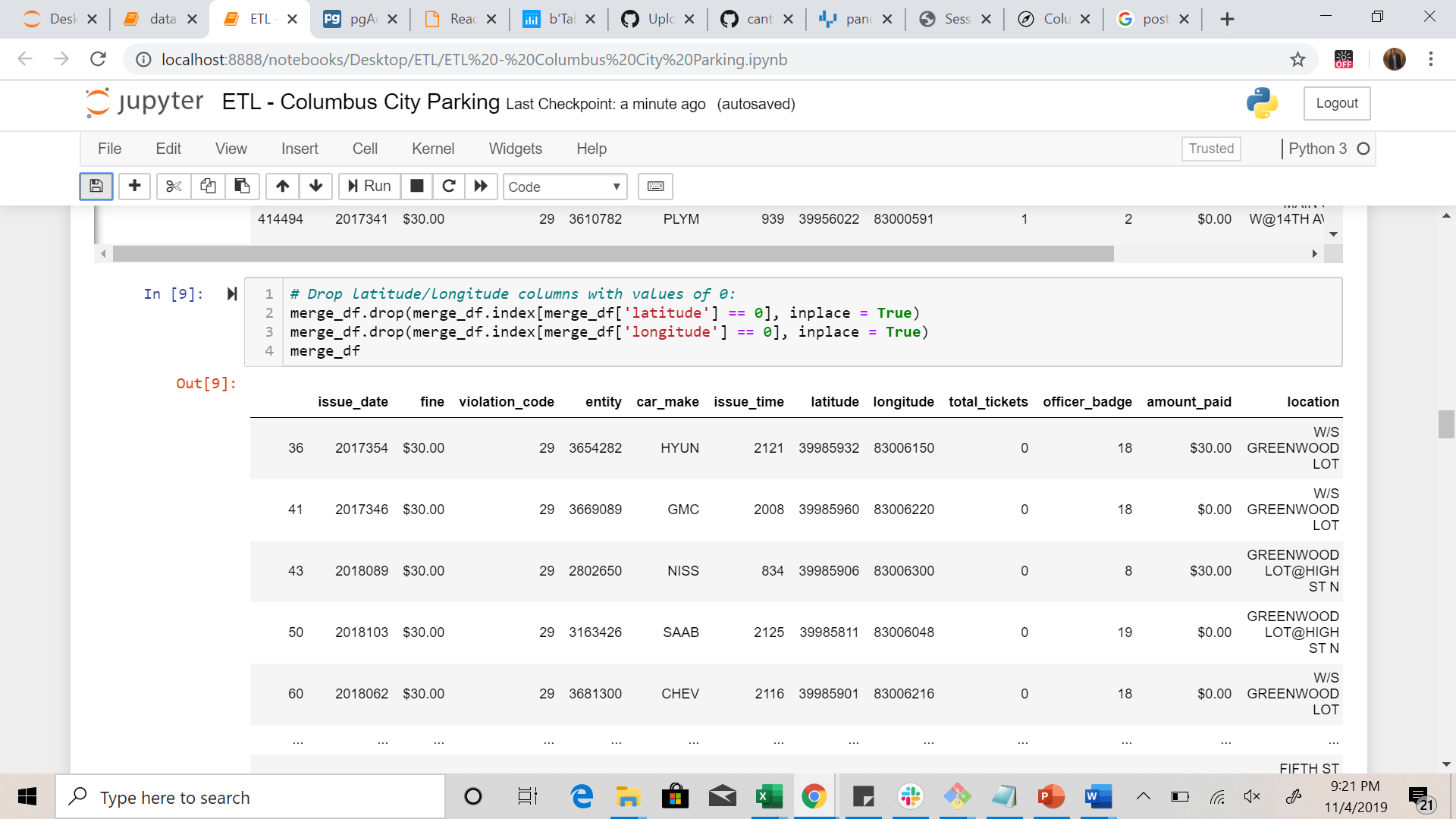
Source:

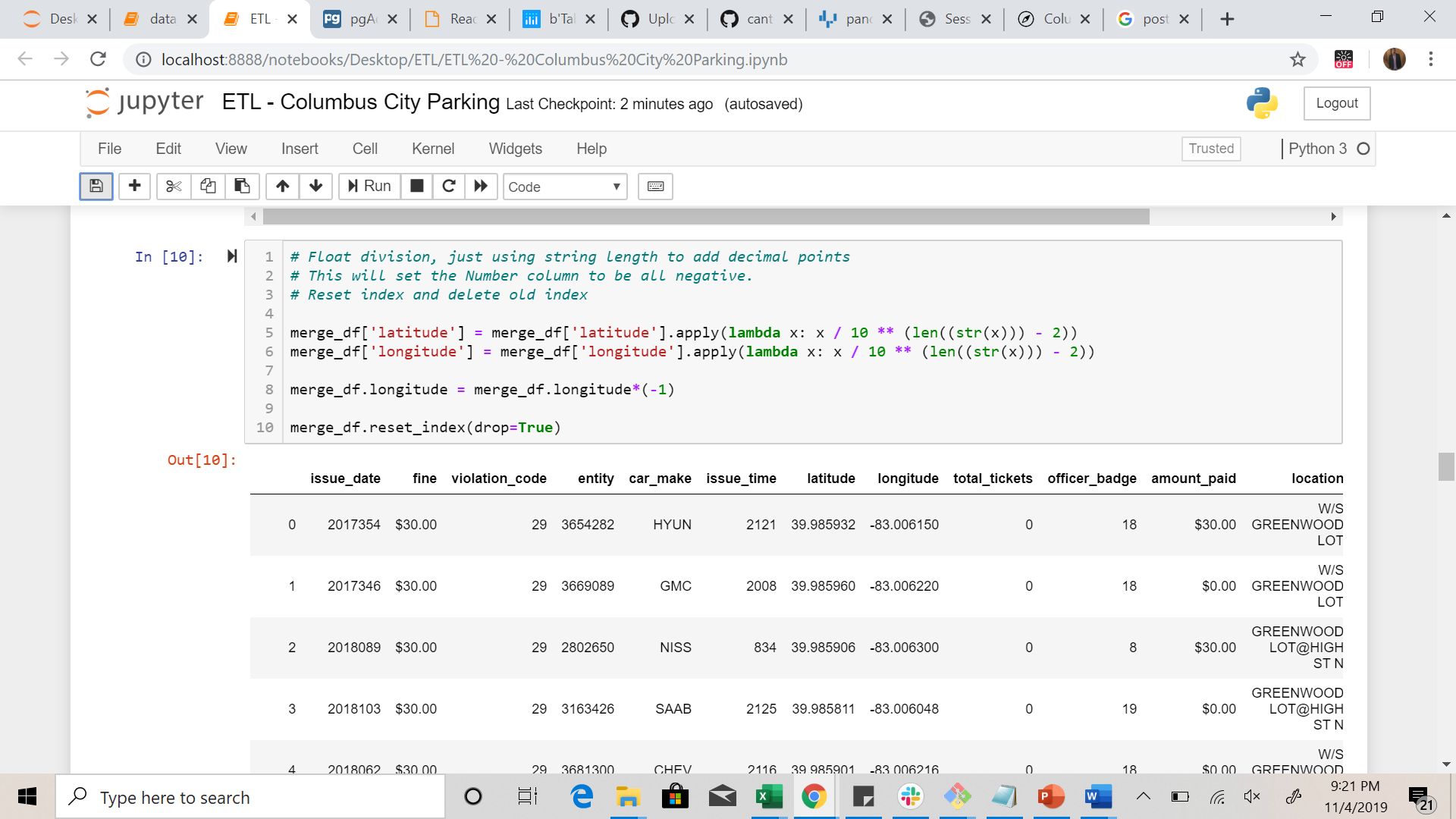
<https://data.world/smartcolumbusos/040b5929-db26-4453-920a-ceb282c4359f/workspace/file?filename=geocoded-parking-violations-csv-5.csv>

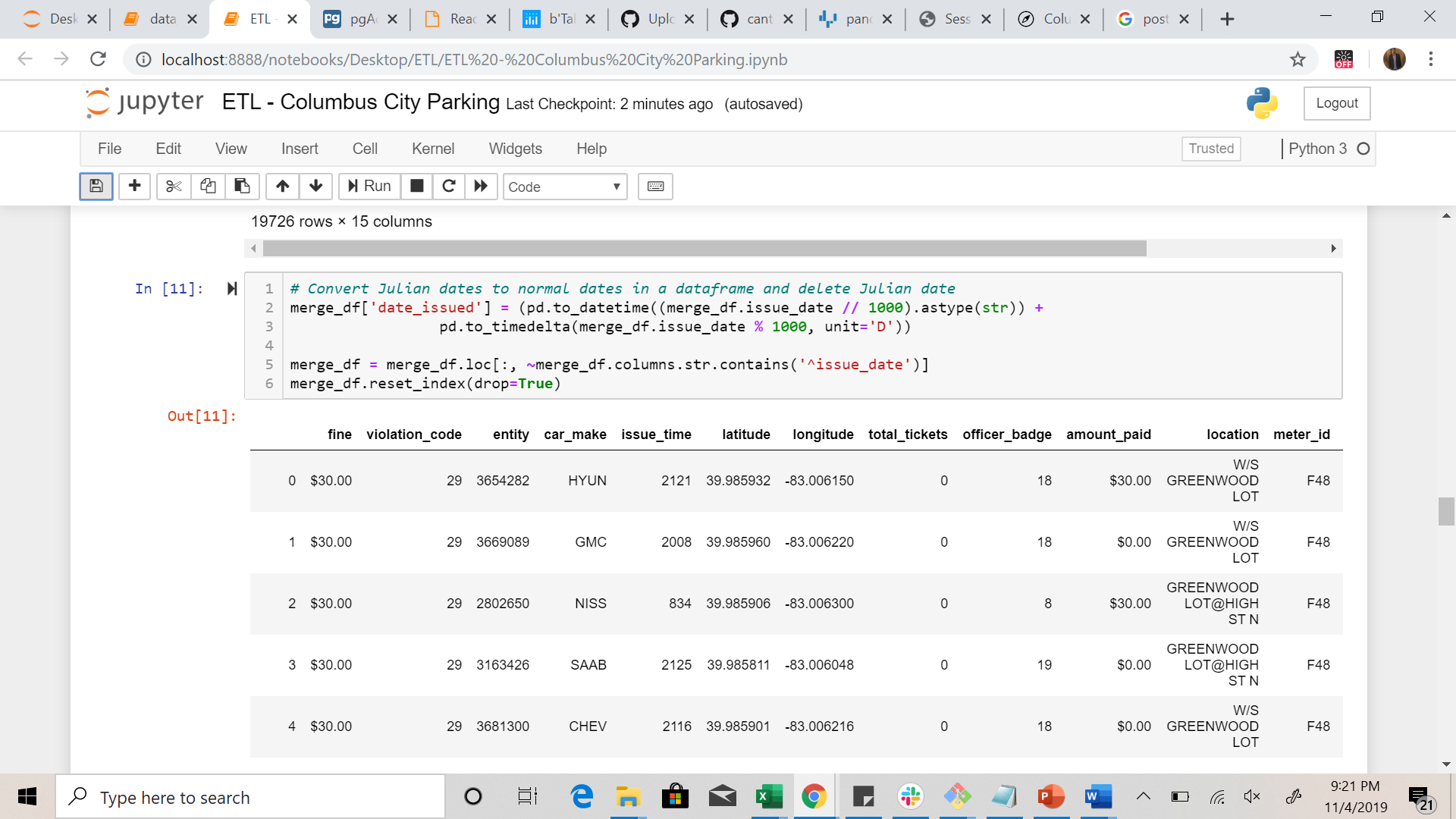
**Transformation:**

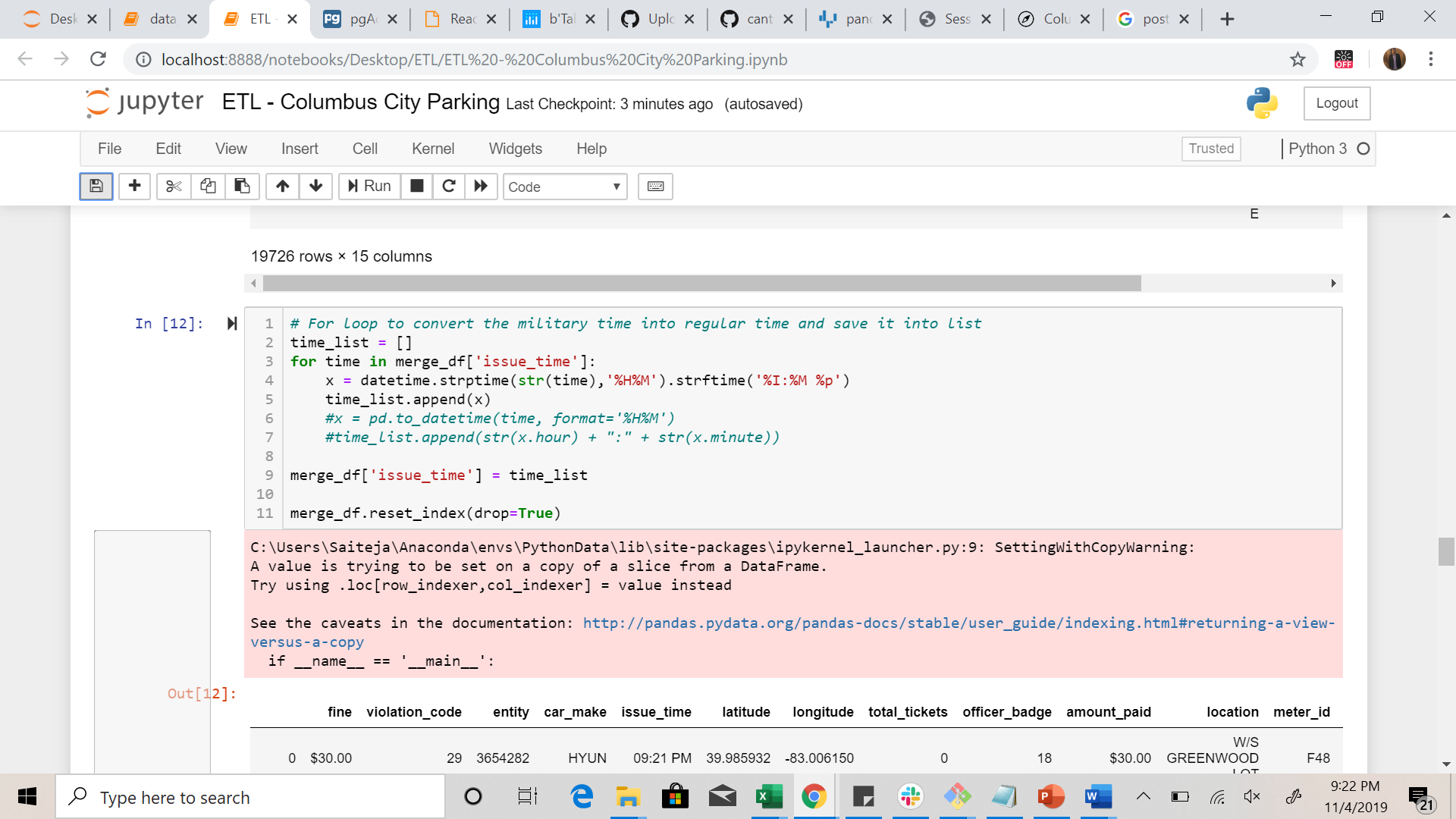
* The Columbus City Parking Violation dataset had to undergo multiple cleaning stages as shown here:

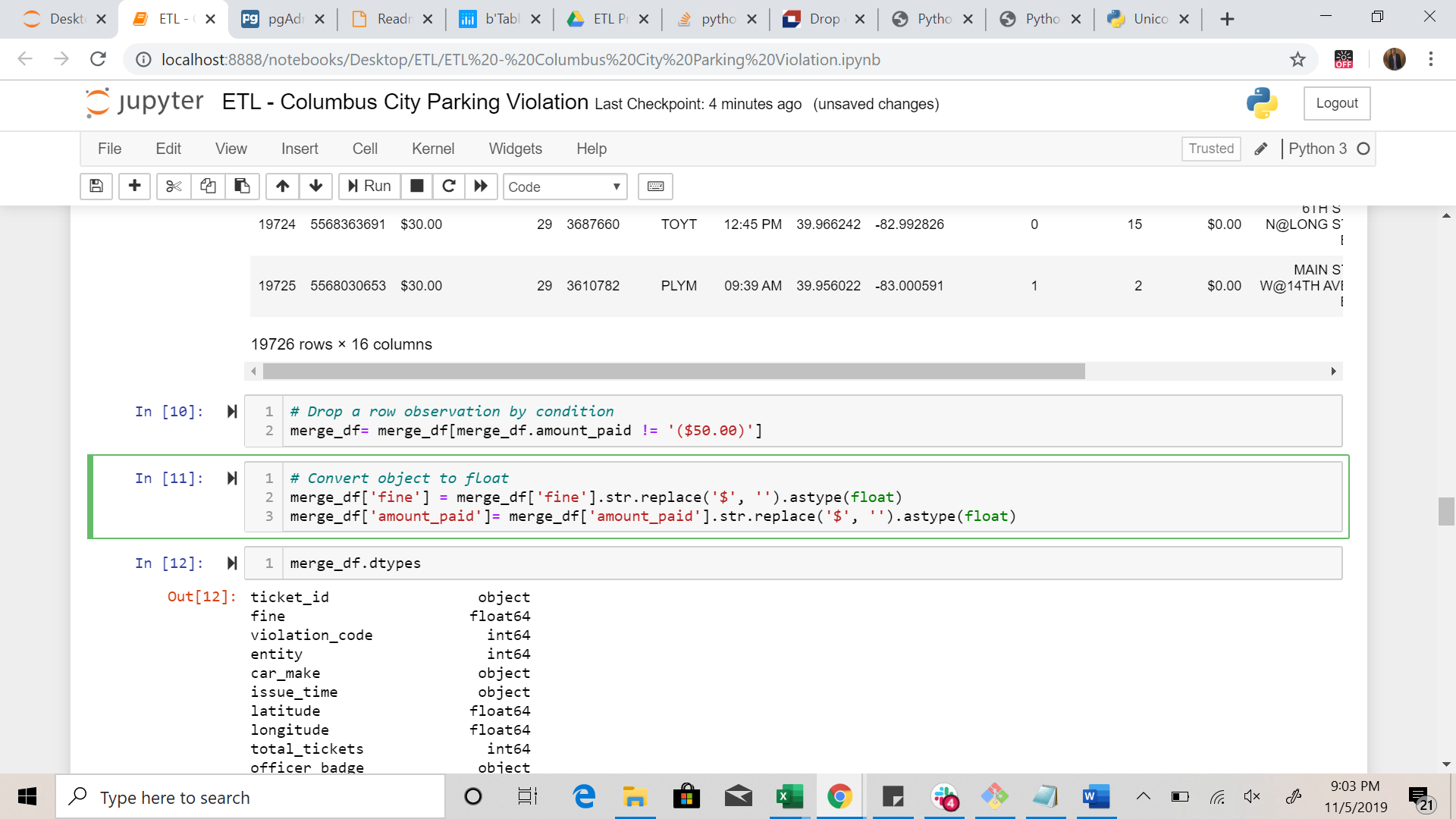


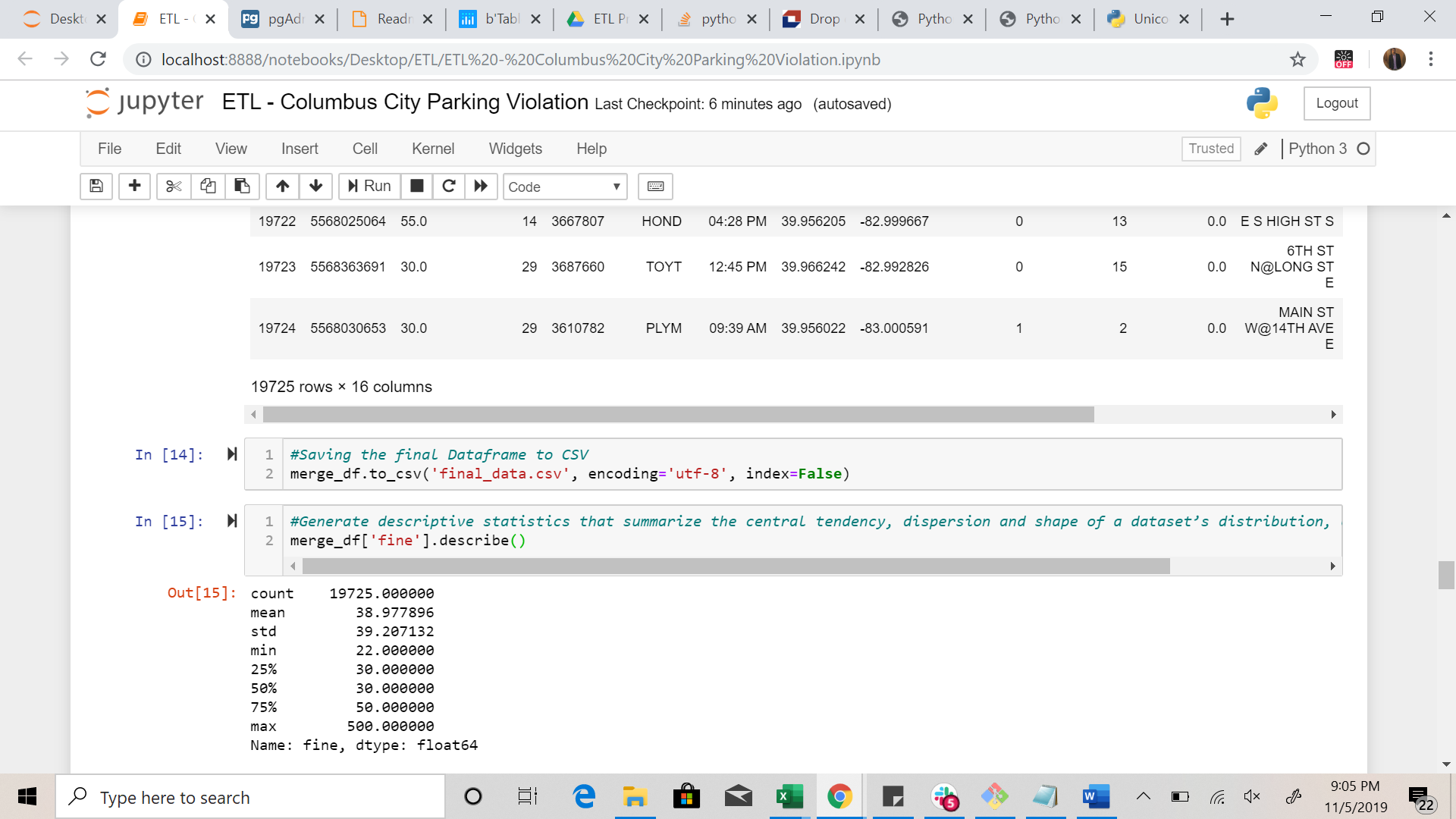












* Merged data was loaded from a CSV file to a DataFrame, and finally the data was loaded into a SQL database.

**Load:**

* Connect to local database
* Exported the data to Postgres tables. Since the final output is a DataFrame, we decided to load the data into a relational database.
* Reason why these columns were selected is to bring awareness to different factors that contribute to parking violations.
  1. Data Exploration. Some of the questions to look into based on the final dataset:
     1. How many tickets were there as a total?
     2. How many tickets were per year during 2017-2018 and if there is any trend?
     3. Popular Car make which received the tickets?
     4. When are tickets most likely to be issued?
     5. Which number plate received the most ticket violations from 2017-2018?