**Note for the group**

I added couple of ipynb files with Linear Regression in sckit-learn. One is for CA only another for all cities. I did some data cleaning and data preprocessing. Code is based on the new datasets. I was able to import data directly from the pgAdmin and joined the table all\_prices and all\_production in pgAdmin. I had to do some tweaks on RAW data (I did not change the originals – copies of a RAW data are on my branch in resources folder). Below is the description what I did.

Altering RAW data (copies):

* Checked matching/non-matching dates in excel with the VLOOKUP function
* 2 unmatching dates found (off by 1 day) -> fixed manually in PRODUCTION\_TABLE:
  + WAS 1/6/2019 NOW: 1/7/2019
  + WAS 1/5/2020 NOW: 1/6/2020
* Fixed headers in ALL\_PRICES table for columns that start with number:
  + WAS: 4046\_units NOW: units\_4046
  + WAS: 4225\_units NOW: units\_4225
  + WAS: 4770\_units NOW: units\_4770\_
* Changed duplicate headers in PRODUCTION\_TABLE (pdAdmin was complaining about duplicates when joining)
  + WAS: total\_volume NOW: total\_volume\_produces

Next, I Imported data in pgAdmin with the same code that we have, used same UNION queries that we have and add new query to join tables (query can be found in queries.txt). Code for importing tables directly form the pgAdmin is in the ipynbs.

Questions from the rubric on ML, some of the EDA and statistics are in the separate file (04\_EDA\_and\_statistics).