[Summary report]

Junior DS Test for CARTO

Hyojung Lee

Summary

Test assigned: 3-Nov-2020
Test submitted: 10-Nov-2020

- **Tools**: Docker, Postgis, Jupyter notebook
- **Languages**: Python and SQL
- **Libraries**: Python's geodata libraries (e.g. geopandas, geopy, GeoAlchemy2)
  + Pydata stack (pandas, numpy, sklearn, matplotlib, seaborn)

- **Results**
  - Jupyter notebook - ETL
  - Jupyter notebook - EDA + Modeling

This is a very brief summary report.

Full documentation is available on github repository.

I. ETL - process

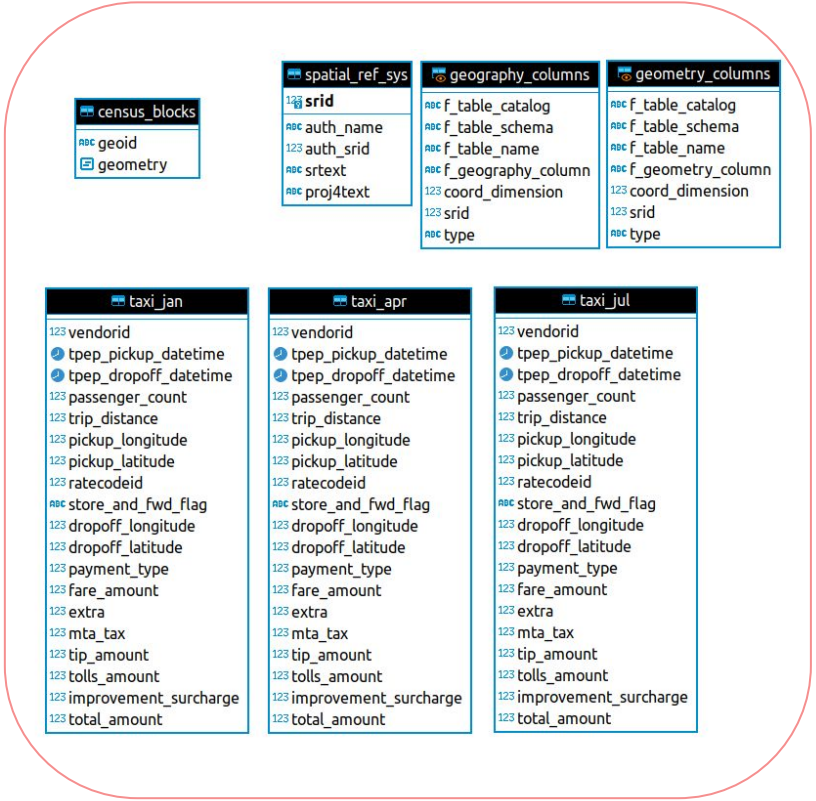
1. Prepared docker containers for a database (Postgis) and Jupyter notebook using docker compose.

2. Created three tables in the db for each month's data. (table name: 'taxi_jan', 'taxi_apr', 'taxi_jul')

3. NY taxi data values were transformed and inserted to the corresponding tables.

4. NYC census block geometries (.geojson file) was also loaded to the database. (table name: census_blocks)

Database schema after completing ETL



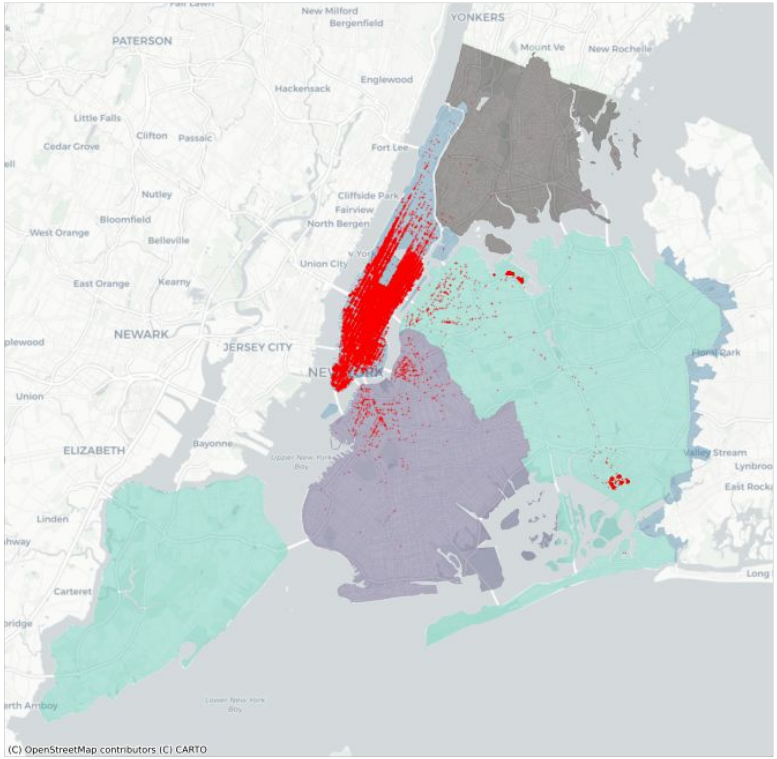
I. ETL - discussion

How to scale up the ETL for larger data?

- **More partitioning of the data**
  During ETL, I partitioned original data files by month.
  To improve the process, the data can be further partitioned by days.

- **Process the incoming data incrementally**
  For example, the function I prepared within the notebook takes data month by month.
  Therefore, in case there is more upcoming data, I don't need to reprocess the previous data.

II. EDA - process



Projection of pickup points (30,000 points) & NY census blocks

1. Pickup points (red dot) only within census blocks (colored polygons) were selected by query and saved as a geodataframe.
2. Only 30,000 pickup records were sampled from the database.
3. Total number of pickups within each census block were counted.
4. ACS data was cleaned and merged with the number of pickup data.
5. Top 10 census blocks with highest number of pickups were chosen and its ACS attribute's summary statistics were created.

II. EDA - discussion

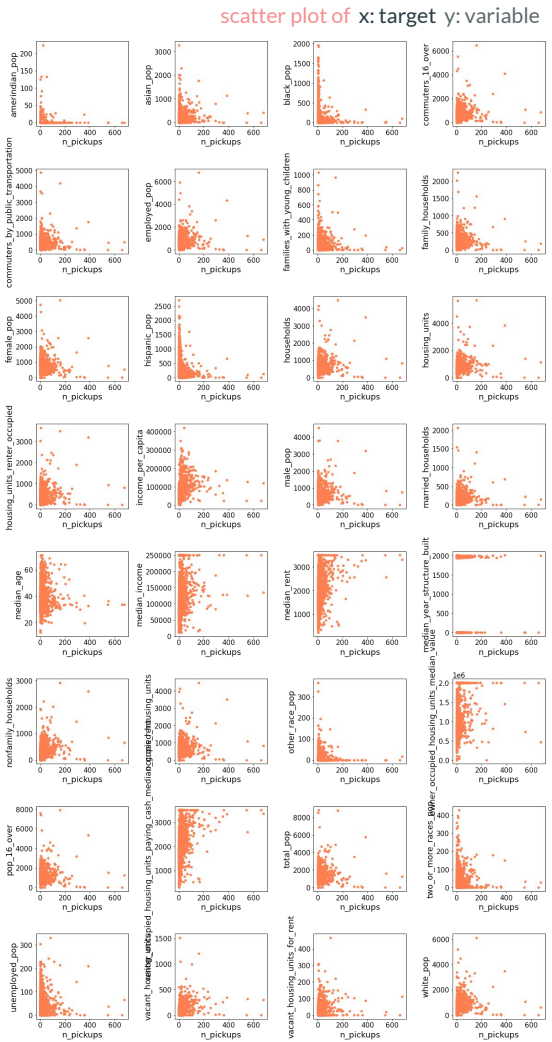
Linear regression was selected as a baseline

Why linear regression?

- Target (N of taxi pickups within each census block) is a continuous variable. Therefore it is a regression problem.

- As a part of EDA, linear relationship between the target and features (ACS data fields) of each census block was examined by
  - scatter plot (figure on the right)
  - correlation coefficient

  And some demographic & socioeconomic attributes demonstrated certain linearity regarding the target values.



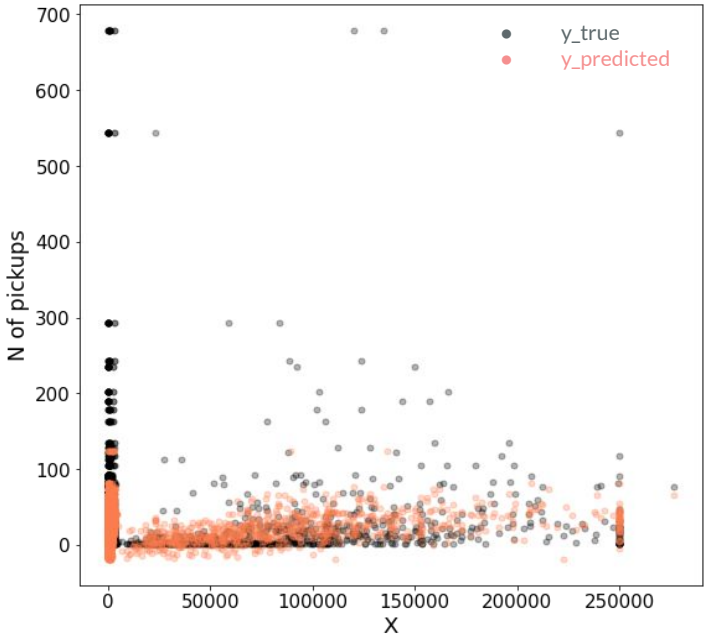
III. Modeling - summary

LinearRegression() from scikit-learn library

- **Target**: N of taxi pickups within each census block

- **Features**: Selected demographic & socioeconomic variables with higher correlation coefficient (r > 0.09)

- **Model performance evaluated by MAE (Mean Absolute Error)**
  Why choose MAE?
  - Intuitively easy to interpret how off out prediction is from a true value.
  - MAE of current baseline: 24
  - Instead, MSE (Mean Sqaured Error) can be also considered, if the model requires higher penalization on errors far off the mark.



III. Modeling - discussion

How can we improve the model performance in the future?

1. Improve the current features
   ex: outlier removal, missing value imputation, normalization or aggregation.

2. Additional features
   ex: pickup dates, more geodata (area, borough boundaries, street info…)

3. Try different algorithms
   ex: Decision tree, Random forest, XGBoost...

If anything, please contact to soyhyoj@gmail.com

Thank you.