BIKE SHARE IN PHILADELPHIA

This final project aims to explore the spatial-temporal patterns of trips of Philadelphia's bike-share system Indego, and their relationship with demographic variables and weather data.

Datasets & Spreadsheets

All datasets and spreadsheets used in this project have been uploaded to Github page and stored in folder named "Data".

- 5-Year American Community Survey (ACS) in Philadelphia (API through https://api.census.gov/)
- Liquor Violations in Philadelphia (API thourgh https://phl.carto.com)
- Bike-Share Trip Counts of Indego (https://www.rideindego.com/about/data/)
    - indego-trips-2018-q1.csv
    - indego-trips-2018-q2.csv
    - indego-trips-2018-q3.csv
    - indego-trips-2018-q4.csv
- Indego Stations (http://www.rideindego.com/stations/json/)
    - stations.json
- Weather Data (Average Daily Temperature & Precipitation) in Philadelphia in 2018 (https://www.wunderground.com/weather/us/pa/philadelphia)
    - weather.csv

Results

This project is divided into two parts, **explanatory analysis** and **regression analysis**. In the first part explanatory analysis, I successively examined the internal characters of each dataset and found some interesting spatial or temporal patterns. The objective of the second part is to analyze the statistical relationship between the trip counts and other explanatory variables. I combined all data by spatial-join and merge functions provided by geopandas and pandas packages in Python, and then regressed trip counts on other predictors using OLS model. Since the main purpose of this project is to gather, store, and analyze datasets using visualization techniques, the regression analysis serves only as an extension, so that I didn't spending too much time improving its goodness of fitting.

Generally, some significant and obvious findings among the data I collected and used in this project

include:

- Census tracts of Philadelphia exhibit some degree of clustering, that is, tracts similar to each other (measured in values of each demographic variables) turn out to close to each other spatially.
- Indego bikes were more frequently used during commuting on weekdays in 2018.
- Indigo trips in 2018 were statistically related to temperature.
- Indego stations at center city yielded more usages than outer zones.
- Indigo trips at each station were statistically related to the station's location (and hence the cenus tract it belonged to), and the time.

PART I - DATA WRANGLING & EXPLANATORY ANALYSIS

1 Demographic Variables of Census Tracts

1.1 Data Collection Through API

Use *census_area* to request 5-year ACS data (in the form of census tract with geometry info)

1.2 Data Wrangling

Convert to GeoDataFrame

Use *geopandas t*o convert requested census tracts to GeoDataFrame based on geometry field.

Rename

Change the column names in terms of variable codes to corresponding descriptive names.

Calculate the Percentages

Reclassify Tracts Based on Income

Use *.loc* function to classify census tracts into 3 groups based on median household income.

1.3 Explanatory Analysis

Numerical Relationship

Use *scatter_matrix* function from *hvplot* package to visualize all of the pairwise relationships of selected demographic variables.

Use *corr* function from *pandas* package to visualize the correlation matrix of all demographic variables.

Spatial Patterns

Use *alt* package to plot choropleth maps of each demographic variable.

1.4 K-Means Clustering

Use *sklearn.cluster* package to do k-means clustering analysis.

Select Variables for Clustering

Select numerical variables.

Standardize Data

For uniform scale and remove the influence of measurement units.

Classify

Plot: Clustering over Spcace

Use basic *pyplot* from *matplotlib* and *plot* from *pandas* to visualize K-means clustering results for census tracts and simpler grouping based on median household income.

2 Indego Trip Data

Import Data

The indigo trip data was directly downloaded from Indego website.

2.1 Temporal Patterns

Use *seaborn* and *hvplot* packages to plot the temporal variance of trip counts

2.2 Spatial Distribution

Import Station Data

The indigo station data was directly downloaded from Indego website.

Convert to GeoDataFrame

Project station data to Mercator system

Join Trip Counts to Station Data

Count trips by start-station and use *merge* function to join this data to GeoDataFrame station data

Plot: Trip Counts at Each Station

Use *folium* package to create interactive plot of trip counts by station

3 3 Weather

3.1 Temperature & Precipation

Import Data

The weather data is collected from web: https://www.wunderground.com/weather/us/pa/philadelphia

Plot: Weather

Use *pyplot* to create overlaid plots of average temperature and precipitation in 2018 in Philly

3.2 Indego Trips vs. Weather

This time we compare Indego trips with average temperature & precipitation by WEEK and examine the relation between Indego trips and weather.

4 Environmental Factor: Liquor Violations

4.1 Request Data

Use *carto.sql* to request data.

Plot: Liquor Violations Map (Large Data)

Use datashader to plot this large dataset that contains more than 1 million observations.

PART II - REGRESSION ANALYSIS

5 Combine All Datasets

5.1 Trip Counts by Temporal Variables

Grouby indigo trip data by station and time-related variables, and count the number.

5.2 Join Weather Data to Trip Counts (by Date)

Merge weather data to trip data.

5.3 Spatial-Join Demographics to Trip Counts (by Location)

Each station will be added with demographic variables of its corresponding census tract where it locates.

6 OLS Regression Model

6.1 Modeling

Use *statsmodels* to regress trip counts on explanatory variables.

6.2 Goodness of Fitting

Examine the performance of this simple OLS model.