Project Group -

Members: 5

Student numbers: Zhenlin Xu(5721679), Luyang Cao(5685915), Yuchen Qi(5695392), Xinghao Lou(5698715), Yimin Xu(5696925)

Research Objective

Requires data modeling and quantitative research in Transport, Infrastructure & Logistics

- Investigate the distribution and hotspot of the origin and destination of the travel by taxi in New York City.
- Detect the hotspot of the origin and destination of travel by taxi in New York City

  Is there any difference of origins and destinations of taxis in New York City between weekdays and weekends? Which region is the hotspot of taxi travel origin and destination? What are the travel characteristics of the hotspots? The main reason why there is a cluster center.
- Analyze the characteristics of the data from different perspectives, such as:

  1.Trip counts per day by hour

  2.The total amount of fare per day by hour

  3.Average trip distance per day by hour

  4.Average travel speed per day by hour

  5.etc.

Contribution Statement

Be specific. Some of the tasks can be coding (expect everyone to do this), background research, conceptualisation, visualisation, data analysis, data modelling

Author 1:

Author 2:

Author 3:

Data Used

The data used in this group project were collected and provided to the NYC Taxi and Limousine Commission (TLC).

Raw data is available at: TLC Trip Record Data - TLC (nyc.gov).

Data Pipeline

1. Background research:

    - Collect and read taxi travel-related literature about the study of the Origin-Destination pair.

2. Data collecting:

    - Download taxi travel data collected from 2014 provided by TLC Trip Record Data - TLC (nyc.gov).
    - Read dataset instructions about the explanation of each column.

3. Data splitting and cleaning:

    - The total size of the dataset for the whole year 2014 is around 27G. We decided to split the data and only use one week of data for the research.
    - Preprocessing steps including datetime parsing, and choosing specific columns are done while reading the csv file.
    - We clean the data with NaN value.
    - We filter the data that have abnormal values.

4. Origin and destination clustering: using the clustering algorithms from Scikit-learn to calculate cluster centers of taxi travel origins and destinations.

    - The k-means algorithm from Scikit-learn is used.
    - An elbow algorithm is implemented to choose the best hyperparameter for k.
    - A collection of hyperparameters are selected to fine-tune the results.

5. Statistical explanatory: using Pandas to calculate several traffic parameters and Matplotlib to visualize the results.

    - To be done after the mid-term check
    - Matplotlib and Seaborn to visualize 2D plots.

6. Origin and destination visualization: using Plotly to visualize the spatial-temporal evolution of the New York City taxi travel origin and destination hotspots.

    - To be done after the mid-term check
    - Geopandas to visualize the geo-scatter plots.

7. Visualization of the different characteristics of the data set by Matplotlib and Plotly express.

    - The characteristics mentioned in the objectives are visualized.

In [ ]:

```

```