Study of Macau bus routes analysis using Data Science

by

SI Pui Lam, Bryce, M-B9-5549-1
KU Chon Tong, Don, M-B9-5512-3
NG CEN Andre, M-B9-5532-2

Group H

Supervisor: Dr. U Leong Hou

A group project submitted for
the course of CISC7201 INTRODUCTION TO DATA SCIENCE PROGRAMMING at
the University of Macau

Table of Content

Study of Macau bus routes analysis using Data Science

This study aims to analyze the bus routes in Macau by designing and conducting a model to read and visualize the data since 2016 till today, and writing up the process and results as a Group Project.

1. Problem Statement

The bus transportation system in Macau has been developing frequently over the recent years, so its bus lines and bus stations also have been changing. However, the changes in bus routes doesn't necessarily means improvement, so our objective is to investigate whether the changes in the bus routes before and after can effectively disperse the bus lines and keep the traffic more fluent, and determine what are the standards and outputs of these actions from the government.

2. Data Collection

Data Set Description:

This bus route dataset was collected by the Macau Transport Bureau (DSAT) and given to the public site "https://www.dsat.gov.mo/bus/site/index.aspx". Lucky for us, the data in the page has been Web-crawled by Ng Sio Lei and posted in his Github repository: macau-bus-data[1]. This dataset Includes the routes and stops information, schedule and geographical data. We used the records that starts from February 2016 .

Import the data:

To begin with, we have to preprocessed this data source as they are all stored in multiple json files. We have converted the route schedule to quantitative and readable data from the plain text format in the sourced dataset.

```
if __name__ == '__main__':
    rootDir = os.getcwd()
    now = datetime.now()
    logFilename = '{year}{month}{day}{hour}{minute}{second}'.format(
        year = now.year,
        month = now.month,
        day = now.day,
        hour = now.hour,
        minute = now.minute,
        second = now.second
    )

    models.init_db(rootDir + '/data.db')
```

[1] Github repository: macau-bus-data:https://github.com/ngsiolei/macau-bus-data

```
    print("create & init sqlite db ...")
    models.create_tables()
    print("db created")
    print("----------------------------------------")
...
    print("create bus schedule ...")
    create_bus_schedule('../raw/macau-bus-data')
    os.chdir(rootDir)
...
```

In the above code snippet, we've downloaded the bus route schedule dataset and the bus stop dataset into "data.db".

Python Libraries used:

Shapely
What we have in our data was the coordinates of the bus stops, and bus moving gps location. So, in order to process this data, we use the "shapely" library, which help us to read and analyse geographical data. Moreover, we chose it because it supports many readers and can also do geometric operations.

Peewee
While in dealing with our database, we use the "peewee" library, because it is very user friendly and we can get our data with a clear coding.

3. Data Cleaning

We began the data cleaning by checking if our datasets have any inconsistency or redundant information, and we found the following:

3.1. Inconsistency of bus stop between route stop data and bus stop data. Bus stop M190 was found in the route stops data of the buses, but no information found in the bus stop dataset, so we removed them.

3.2. Inconsistency of bus line between route coordinate data and route schedule data. 2 bus lines were found in the coordinates, but no information in the schedule, so we removed them.

e.g.:71X and 26AT

3.3. Redundant Chinese characters used for explaining announcements were found in the route stop data, so we removed those Chinese characters. (See Appendix 3.0.1)

3.4. Redundant marking for bus 51A as return route, while in the route schedule data is actually a circular route, so we fixed this as a circular route. (See Appendix)

Next, we use SQL to read the database we created and find whether there are "ghost" buses or stops which were not used at all, which means the null values (see Appendix 3.0.2).

Lastly, when we were inserting the information into our database, we found that many of the bus line's codes from route coordinates data and bus stop data doesn't match. It turned out that many bus codes had some suffixes after it, so we made some processing here to match the names using the "re" module (see Appendix 3.0.3).

4. Data Exploration & Analysis

Exploring the bus route changes

In the diagram (Appendix 4.0.1) , we illustrate the structure of our database, where we can see that besides of bus route and bus stop information, we also have introduced the Parish (the categories of different areas in Macau) and Bus Agency information.

Then, we come up with some statistics using Jupyter Notebook, shown as below:

4.1. Number of routes on Bus Agencies (See Appendix 4.1.1);
4.2. Number of bus stop in month (See Appendix 4.1.2);
4.3. Number of bus routes in month (See Appendix 4.1.3);
4.4. Number of bus shift in weekday (See Appendix 4.1.4);
4.5. Number of bus stops on different Parish (See Appendix 4.1.5);
4.6. Top 10 bus stop that pass by most no. of routes (See Appendix 4.1.6);
4.7. Top 10 longest distance of bus route (See Appendix 4.1.7);

5. Data Visualization

In order to have an easier way to understand and explore the data, we create a web-based interactive visualization application for the data. The application is now temporarily deployed at brycesi15.github.io powered by Github Page service.

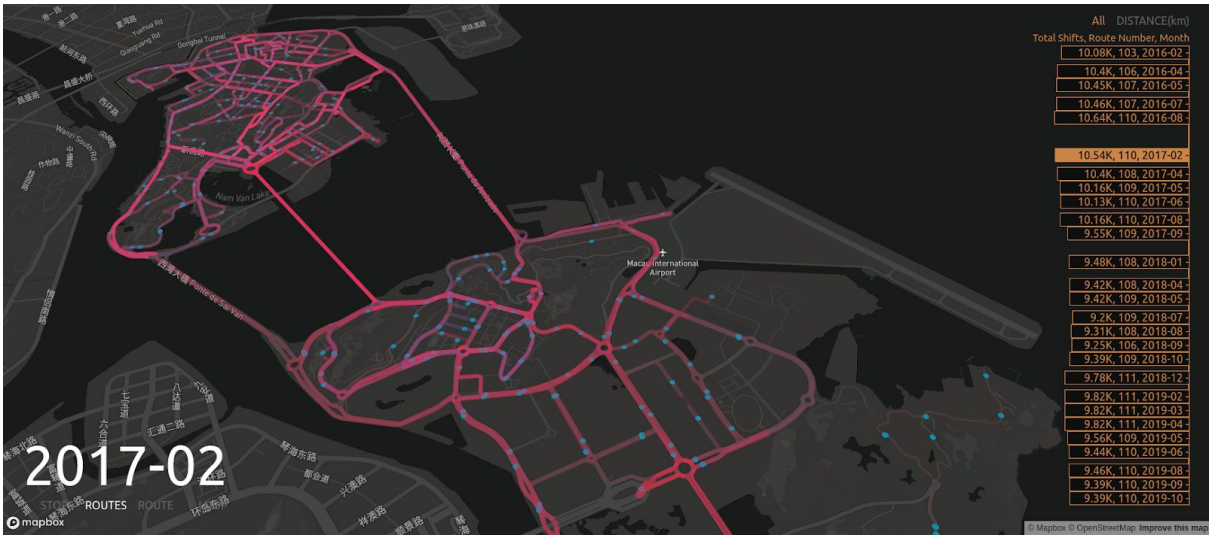
5.1. The visualizations

We have created 3 visualizations based on different point of view on the data:

1. The "STOPS" view - focus on the bus stop distribution and their workload.
   We provide a 3D bar chart on the map to show their distribution and workload. We also provide some statistical data like the workload and stop number.
2. The "ROUTES" view - focus on the overall bus route planning.
   We provide bus flow heatmap based on the bus shift we extract. We also provide some statistical data like the total route number and distance.
3. The "ROUTE" view - focus on single bus route changing.
   We provide a closer view for the change of each bus route based on their driven route and the pass stops. We also provide the their distance information.

5.2. Implement detail

We use the Python with Pandas to extract the key data we need from the clean and processed data, includes the total shift of each route, the total shift pass of each stop etc. We use React with mapbox and deck.gl for our application development.



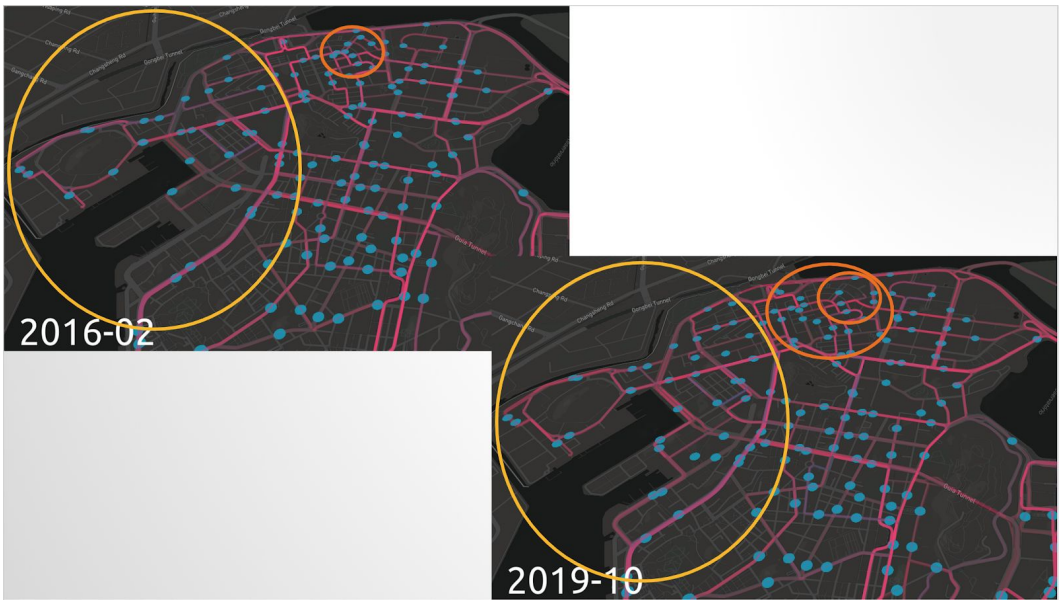
We develop a map view as main view for visualize our geographic data. A control panel at the bottom right corner for switch to different visualizations. An interactive timeline panel to show the statistical data .

5.3. Key finding exploration - Nearby bus planning of border gate after the typhoon Hato

We know that the border gate terminal was destroyed by the typhoon and take one year to revamp. In the figures, we can see the bus workload distribution before (See Appendix 5.3.1) and after the typhoon Hato (See Appendix 5.3.2) the workload of border gate terminal was dispersed to new bus stops set nearby. And this planning is keep after the terminal is reopened until now(See Appendix 5.3.3).

5.4 Key finding exploration - The change of North Macau

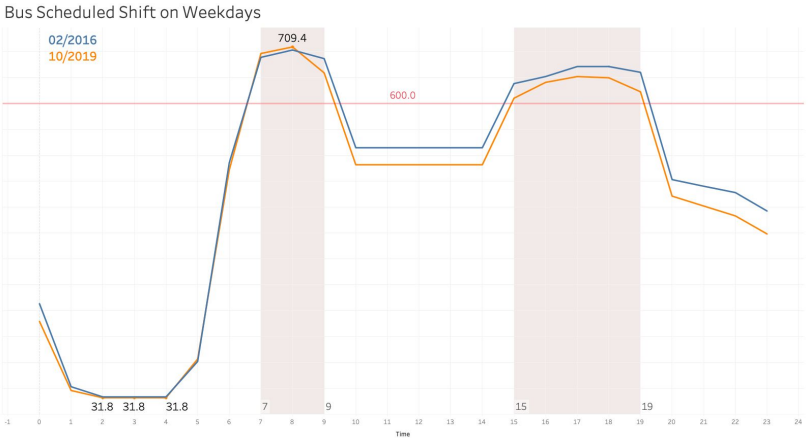
By comparing the bus traffic of the North Macau between 2016-02 and 2019-10, we find out the stop and traffic in the compact area of the orange circle is dispersed to nearby area. And the area inside the yellow circle is well developed and used in the bus planning among 3 year. From those clues we find out one of the planning strategy is to dispersed the bus traffic to different stop and area.



6. Conclusion

Summing up the above findings, we can conclude that the outputs from the new routes and stops strategies of the Macau Traffic Bureau have dispersed the buses while also fulfilling the needs of passengers by expanding the routes. Thus, the data tells us that they are improving.

Last but not least, from our findings we also have a suggestion. We found that the bus schedule shift in a single day haven't increase comparing 2016 and 2019, which is probably due to the full load of the Macau traffic. We suggest the government to keep declining the vehicles on the road, which will contribute in the improvement for Public Transport.



Appendix

See "Appendix.pdf"