Visualization and Clustering of Passenger Flow in Taipei MRT Stations

A05227102 Jana Wilbert

B01106047 Yayam Su

B02902071 PoYao Chen

2017.01.10

Introduction

Motivation Design Data Source

Introduction - Motivation

Motivation

Why?

* Many Large Cities
* large number of people quickly
* Crowded

Introduction - Motivation

Motivation

Issues

To care about...

- Where is the largest passenger flow?
- When is the largest passenger flow?
- How to avoid the crowded flow?

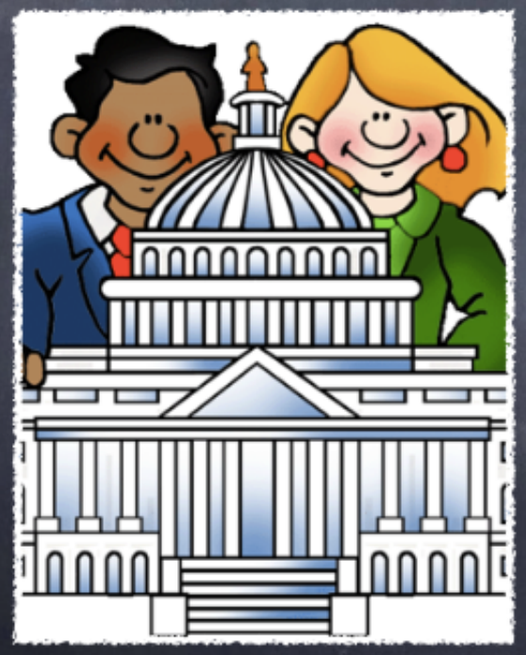
Introduction - Motivation

Motivation

Issues

Who Cares?

I'll make a plan.



Government/
Metro Company

Passengers

Off-peak hour to go

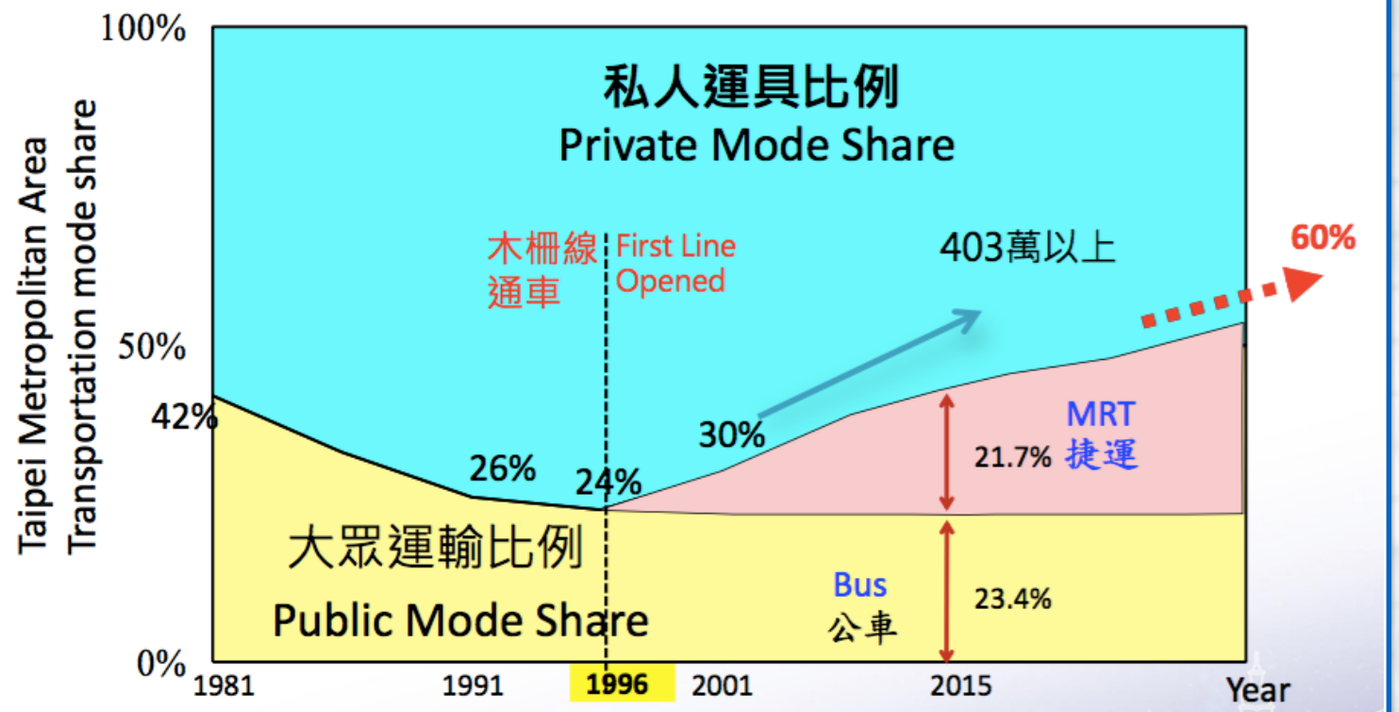
Introduction - Taipei Metro

717 500 000
in 2015

131.1 km

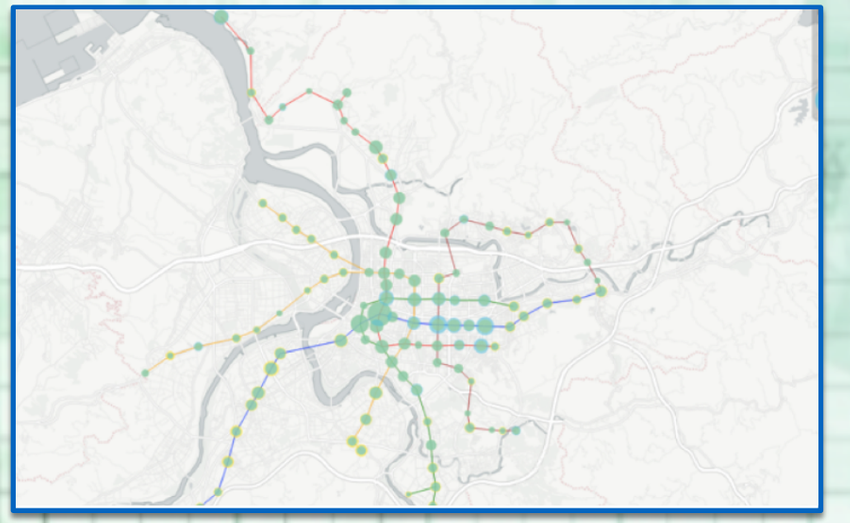
(Wang, 2016)

Development of Mass Transportation in the Taipei Area

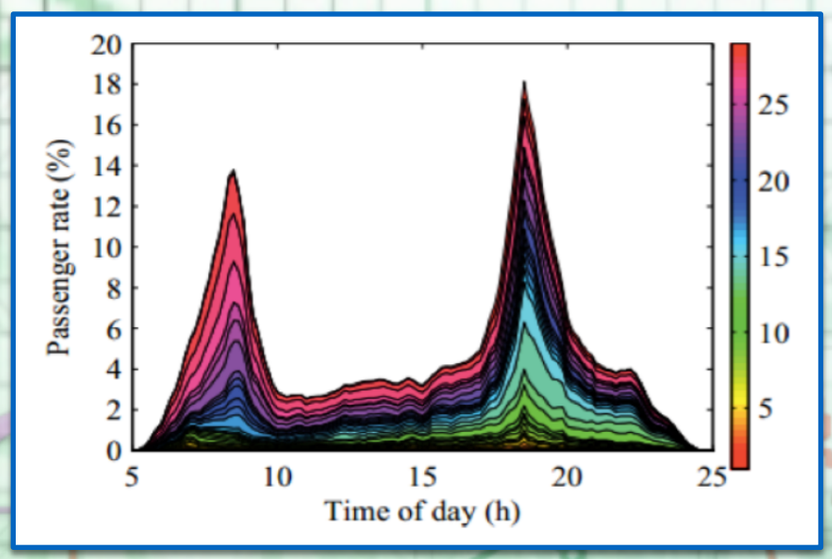


(Wang, 2016)

Introduction – Our Project



(Tsao, 2015)



(Sun et al., 2012)

Classification of stations based on passenger flow

- passenger flow volume
- combination static and dynamic informations

(Wei et al., 2016)

日期	時段	松山機場	中山國中	南京復興	… …
2016/4/1	5	11	42	65	35
2016/4/1	6	70	349	346	409
2016/4/1	7	301	1,201	1,045	1,007
2016/4/1	8	393	1,766	1,600	1,430
2016/4/1	9	291	823	970	1,131

日期	松山機場	中山國中	南京復興	… …
2016/4/1	7,265	17,662	38,162	54,106
2016/4/2	5,213	10,992	20,439	50,926
2016/4/3	4,609	9,088	16,880	45,574
2016/4/4	5,020	8,170	14,966	42,329
2016/4/5	5,809	8,788	16,669	43,172

(from: http://data.taipei/opendata/datalist/datasetMeta?oid=1d71c478-205f-42c5-8386-35f86d74fdd1)

Method

Tools Processing Algorithm

Tools - Data Processing

- Python:
  - Programming Language, Similar to R.
- Pandas:
  - Python Toolkit to load/save csv
  - DataFrame Operation

Advantages: Write One, Process Every time.

```
for prefix in ['in', 'out']:

    df = pd.read_csv("./Data/date_" +prefix+"_201604.csv", encoding="BIG5")

    new_columns = []
    for index, c in enumerate(df.columns):
        new_columns.append(c.replace("/", ""))
    df.columns = new_columns
```

Tools - Algorithm

- Scikit-learn:
  - Python's Machine Learning Toolkit to Cluster Data

Advantages: Easy to use, detailed documentation.

```
from sklearn.cluster import KMeans
import numpy as np

X = np.array([[1, 2], [1, 4], [1, 0],
              [4, 2], [4, 4], [4, 0]])

kmeans = KMeans(n_clusters=2, random_state=0).fit(X)
print kmeans.labels_

kmeans.predict([[0, 0], [4, 4]])
print kmeans.cluster_centers_
```

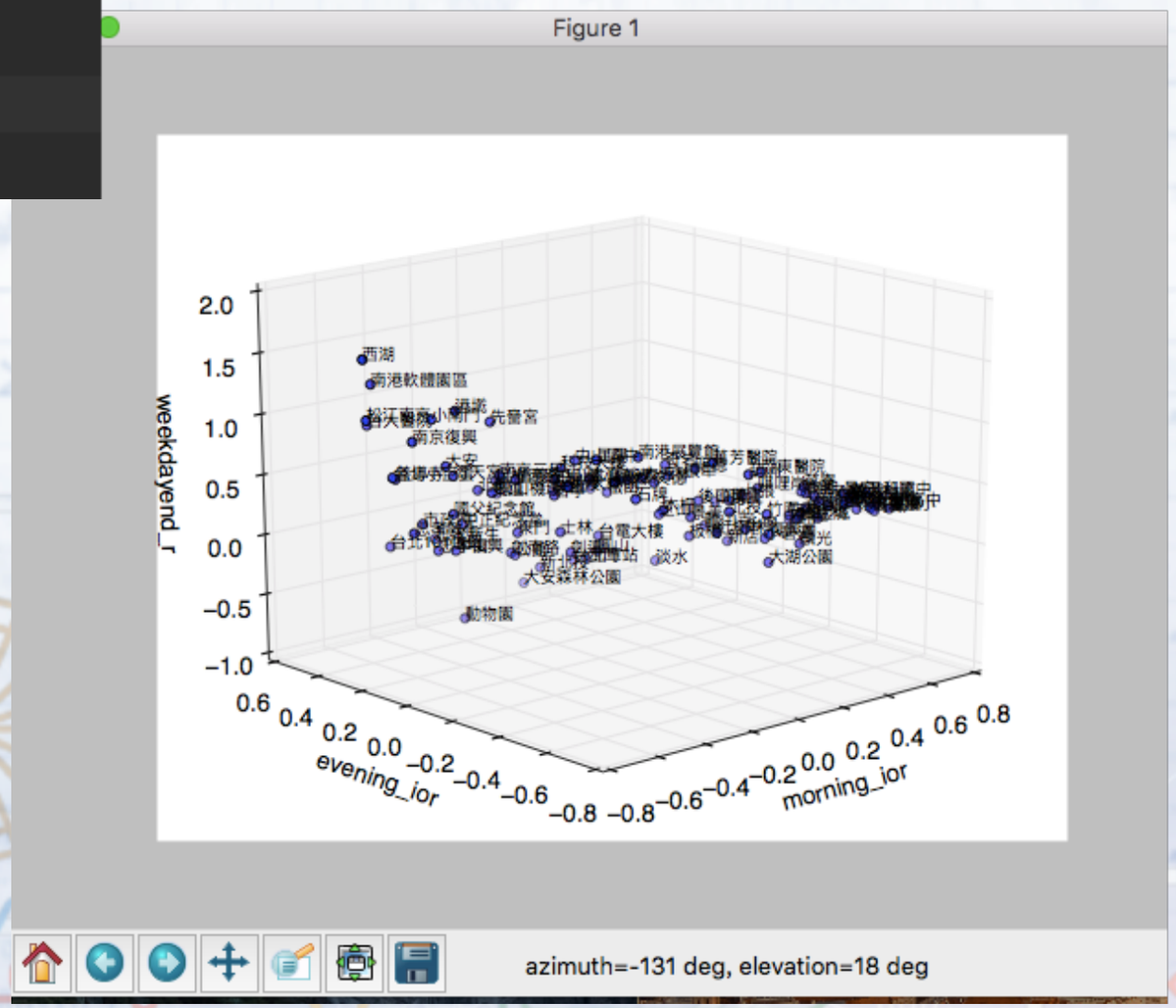
Tools - Visualization

- Matplotlib:
  - Python visualization toolkit
  - 3D Plot, do anything you want
  - Write One, Visualize Every time. (difficult to learn)

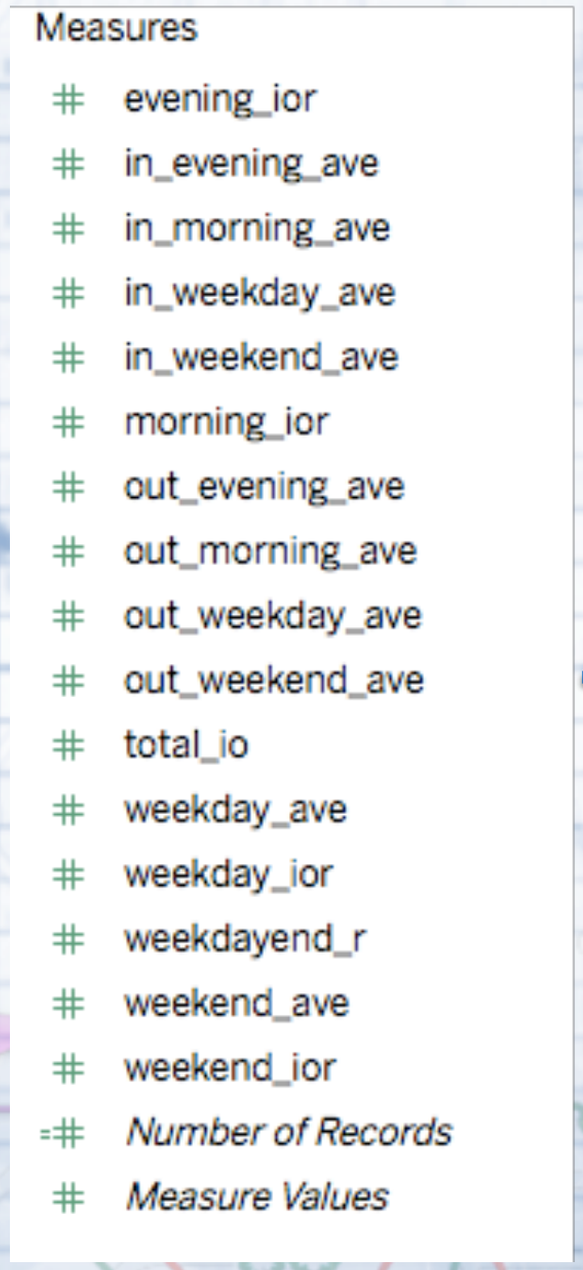
- Tableau:
  - Easy Use, Observation(try features)
  - lots of restriction, data should be cleaned

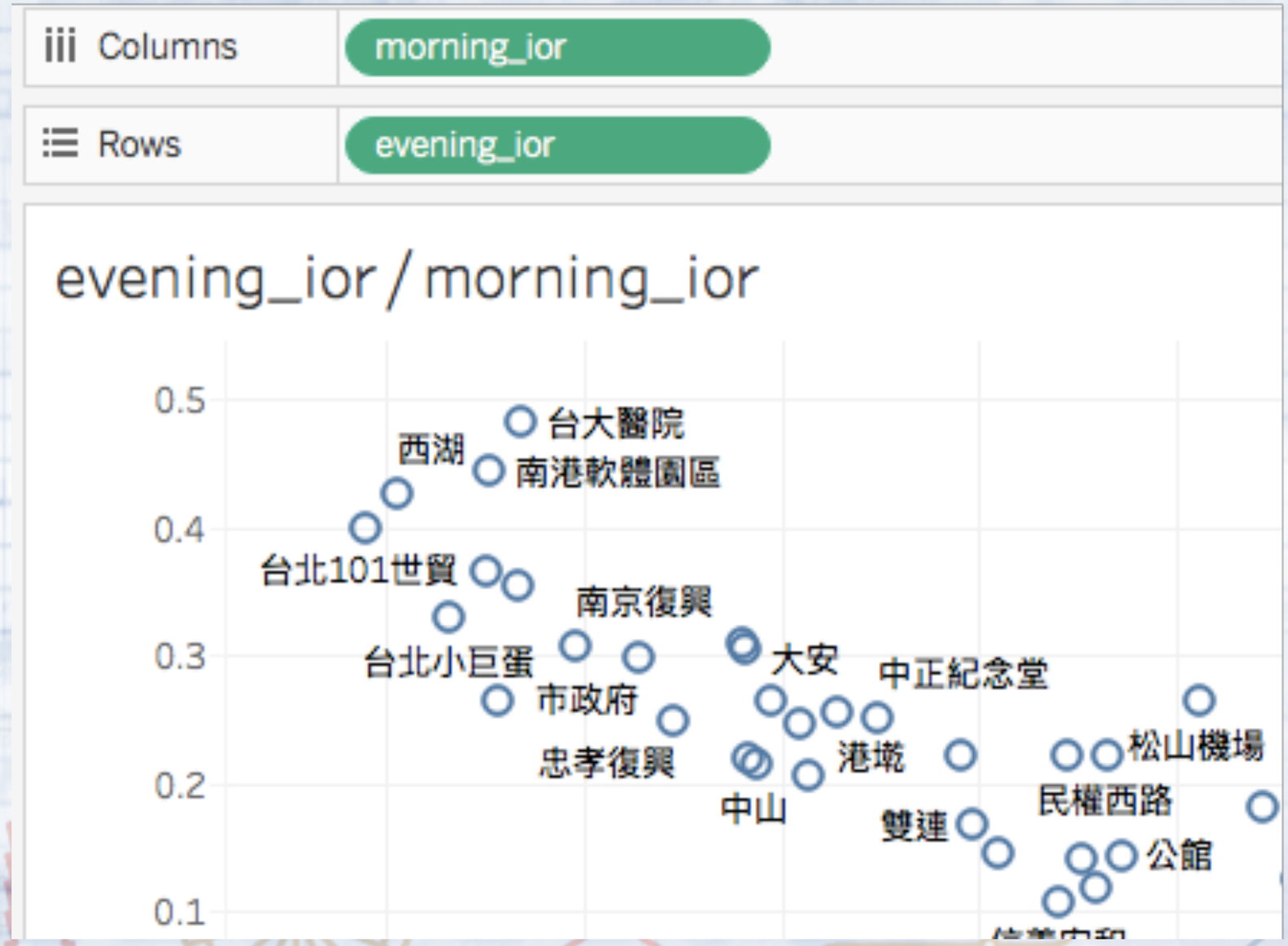
Tools – Visualization (Matplotlib)

```
fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')

x = 'morning_ior'
y = 'evening_ior'
z = 'weekdayend_r'

xs = df[x]
ys = df[y]
zs = df[z]
ax.scatter(xs, ys, zs, c='b', marker='o')
```



Tools – Visualization (Tableau)

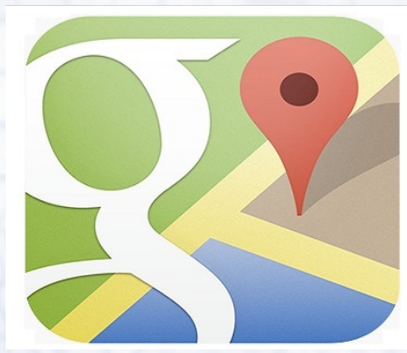




Tools – Visualization (Web)

- html5/css/javascript:
  - Interactive visualization
  - Easy to Share

- Google Map API:
  - Geographic Visualization

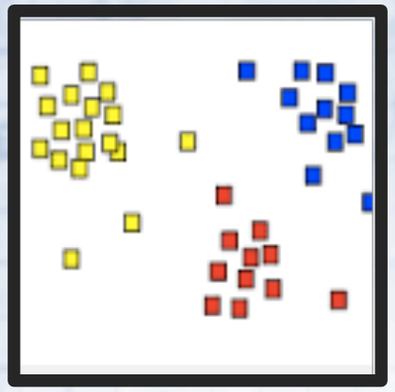


- Plotly:
  - Tool similar to PlotDB, Echart, etc.

Processing

Data Cleaning

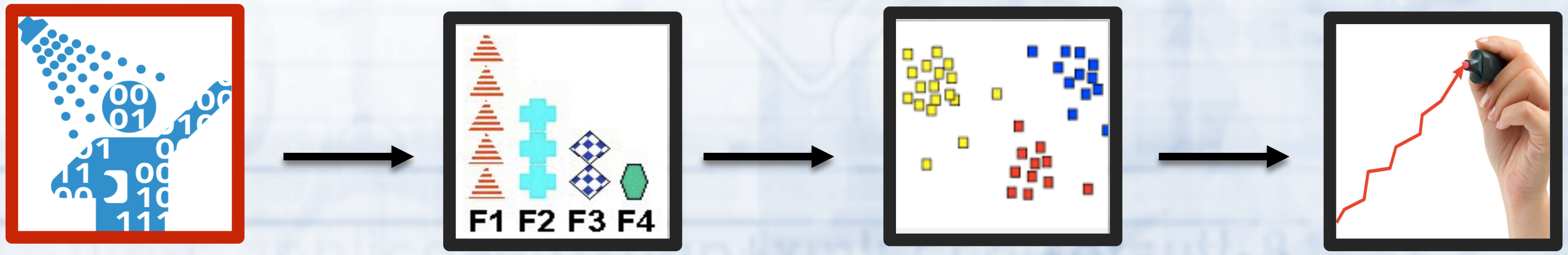
Feature Extraction



Clustering

Visualization

Processing: Data Cleaning



date

¤é´Á	ªQ¤s¾÷³õ	¤¤¤s°ê¤¤	«n¨Ê´_¿³	©¾§µ´_¿³	¤j¦w	¬ì§Þ¤j¼Ó	¤»±i²p	Åï¥ú	¨¯¥è	¸UªÚÂå°|
2016/4/1	7,265	17,662	38,162	54,106	26,245	16,285	12,597	4,981	3,473	17,555
2016/4/2	5,213	10,992	20,439	50,926	14,039	9,939	8,837	6,762	2,641	9,943
2016/4/3	4,609	9,088	16,880	45,574	11,138	8,051	7			
2016/4/4	5,020	8,170	14,966	42,329	10,177	7,046	6			
2016/4/5	5,809	8,788	16,669	43,172	11,328	8,095	7			
2016/4/6	6,272	15,831	35,350	44,246	25,008	14,681	1			

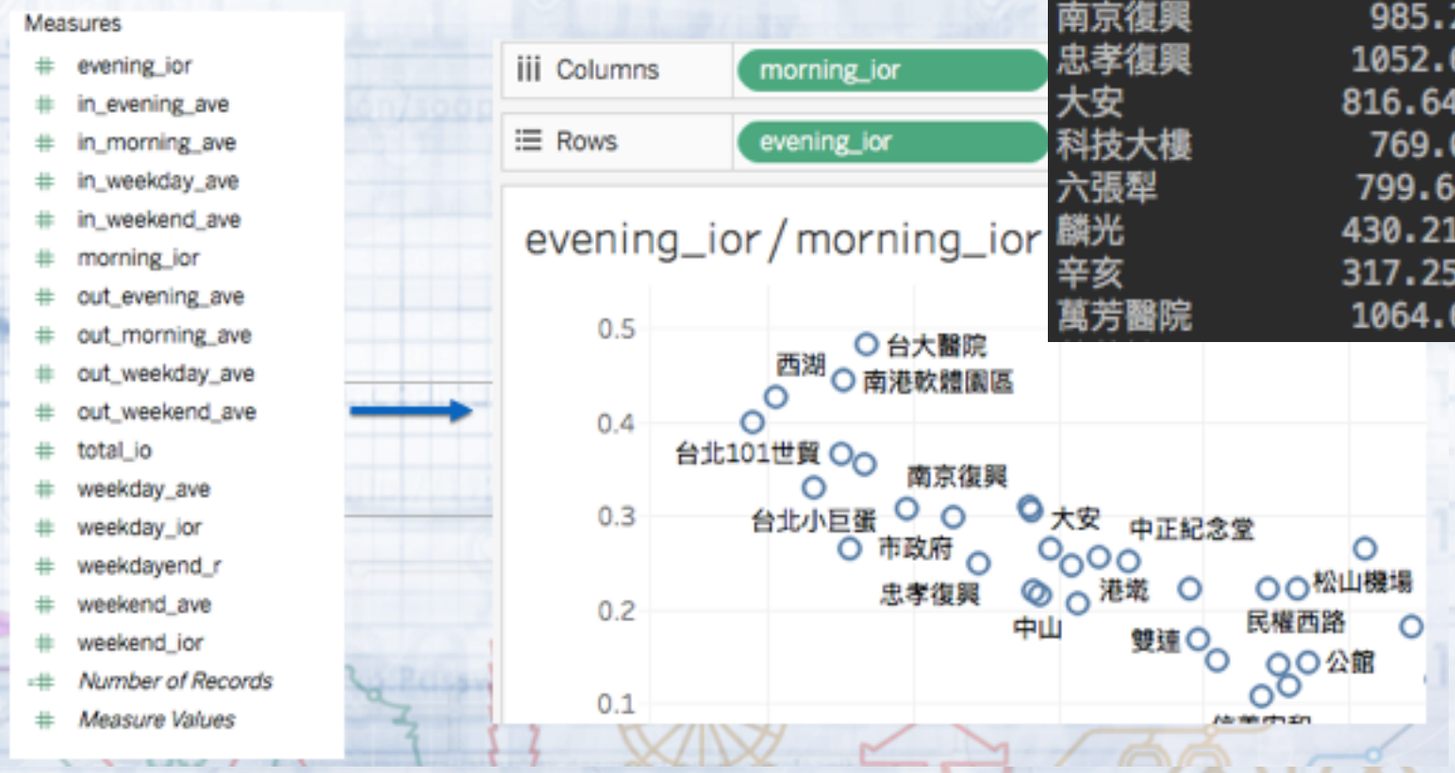
松山機場　中山國中　南京復興　忠孝復興　大安　科技大樓　六張犁

0	7265	17662	38162	54106	26245	16285	12597	4981	3473	17555
1	5213	10992	20439	50926	14039	9939	8837	6762	2641	9943
2	4609	9088	16880	45574	11138	8051	7593	6537	2380	7553
3	5020	8170	14966	42329	10177	7046	6660	5471	1949	7276
4	5809	8788	16669	43172	11328	8095	7230	4140	1965	8382
5	6272	15831	35350	44246	25008	14681	11825	4075	3122	15231
6	6042	16838	36998	47632	25608	15813	11985	4221	3278	15208
7	6787	17533	38334	53141	26542	16921	12747	4505	3460	17525
8	4706	12275	22961	54856	19695	12710	9545	4270	2629	11124

- ➢ Decoding from BIG5 to UTF8
- ➢ Load into Dataframe in Program

Processing: Feature Extraction

	in_morning_ave	out_morning_ave	in_evening_ave	out_evening_ave
松山機場	275.293333	403.613333	350.280952	203.785714
中山國中	855.913333	957.313333	922.152381	824.942857
南京復興	985.220000	2985.460000	2616.590476	1385.023810
忠孝復興	1052.693333	2810.026667	4107.347619	2466.023810
大安	816.640000	1932.720000	1649.957143	961.438095
科技大樓	769.046667	939.693333	901.647619	781.461905
六張犁	799.646667	571.026667	574.533333	738.004762
麟光	430.213333	167.233333	130.666667	323.400000
辛亥	317.253333	80.673333	81.076190	247.938095
萬芳醫院	1064.673333	720.613333	562.466667	1022.300000

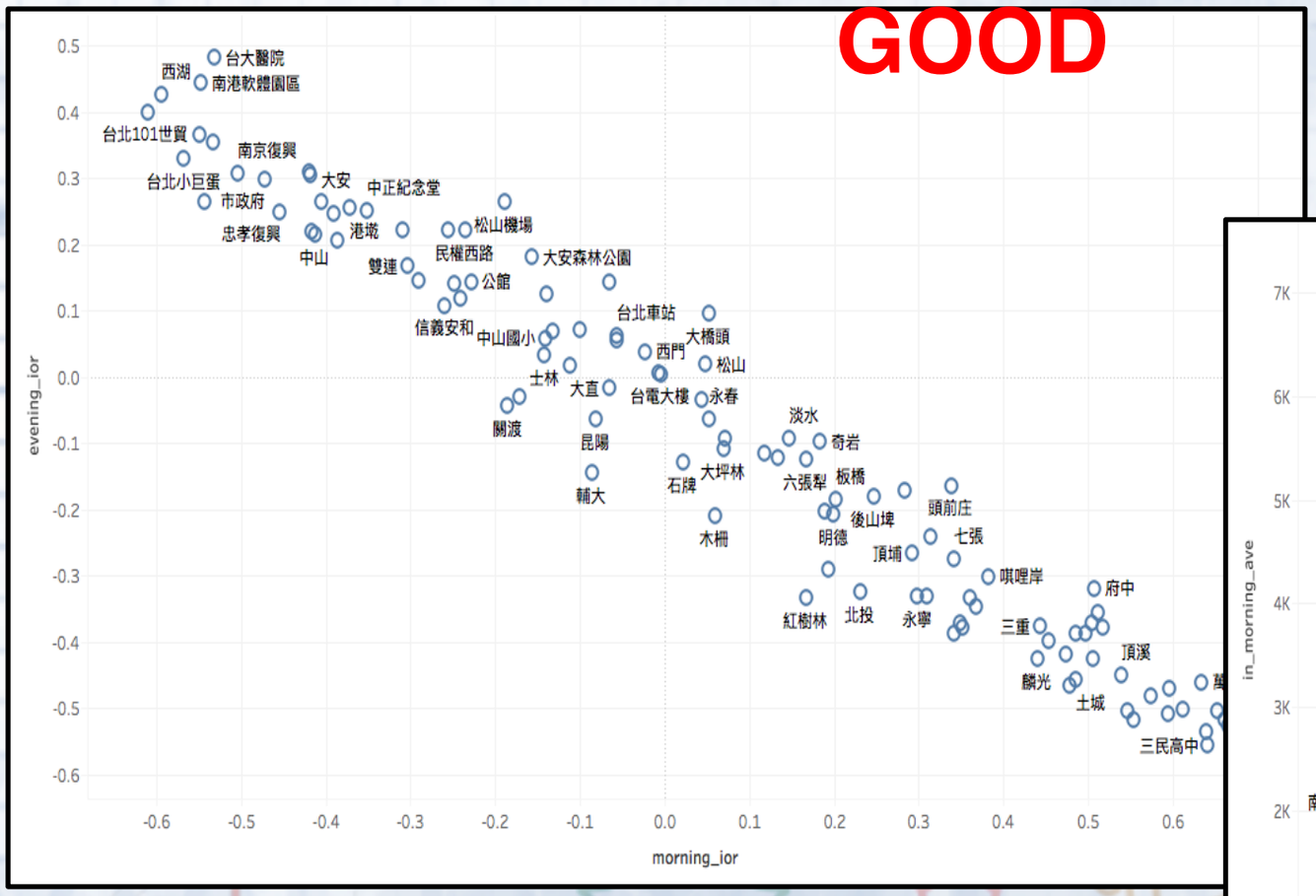


- What's the factor to influence the flow size in different MRT stations?
- What's the relation between fatctors?

Processing: Feature Extraction

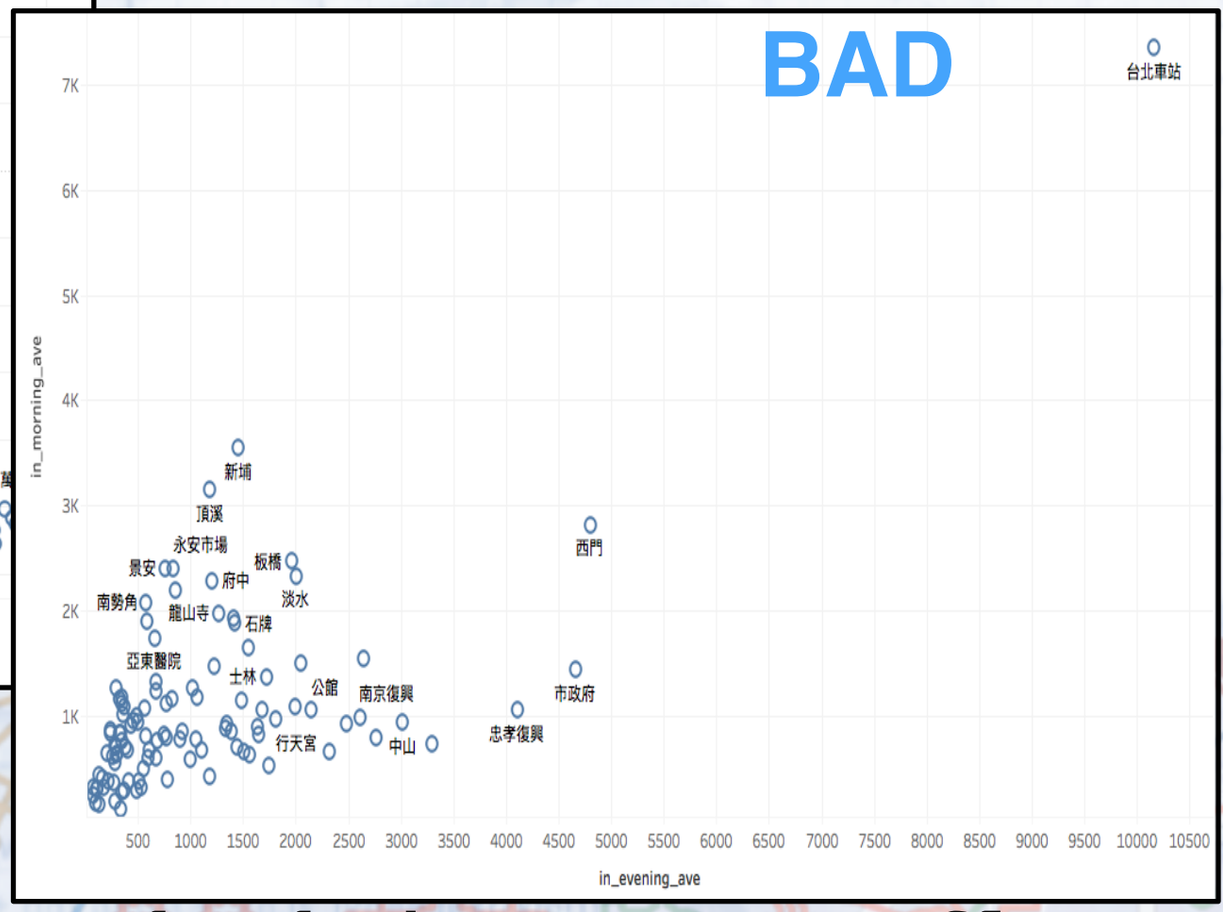
GOOD

$$Ratio(X_1, X_2) = \frac{X_1 - X_2}{X_1 + X_2}$$



evening-check-in-out ratio & morning-check-in-out ratio

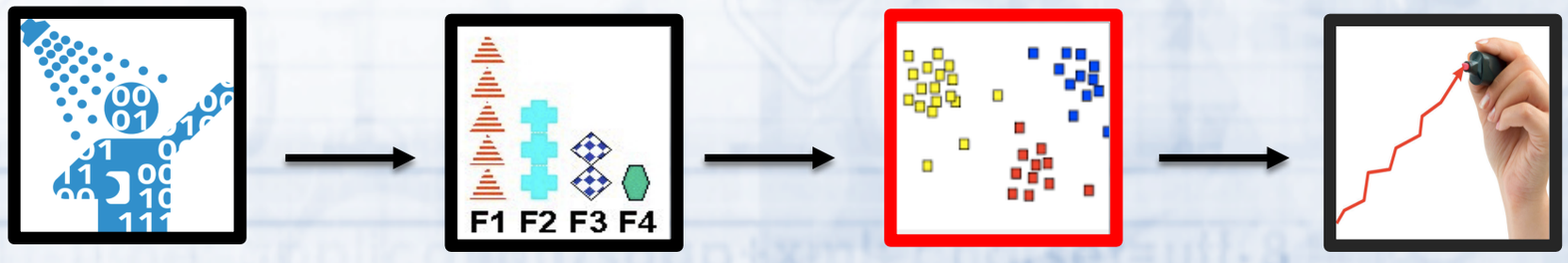
BAD



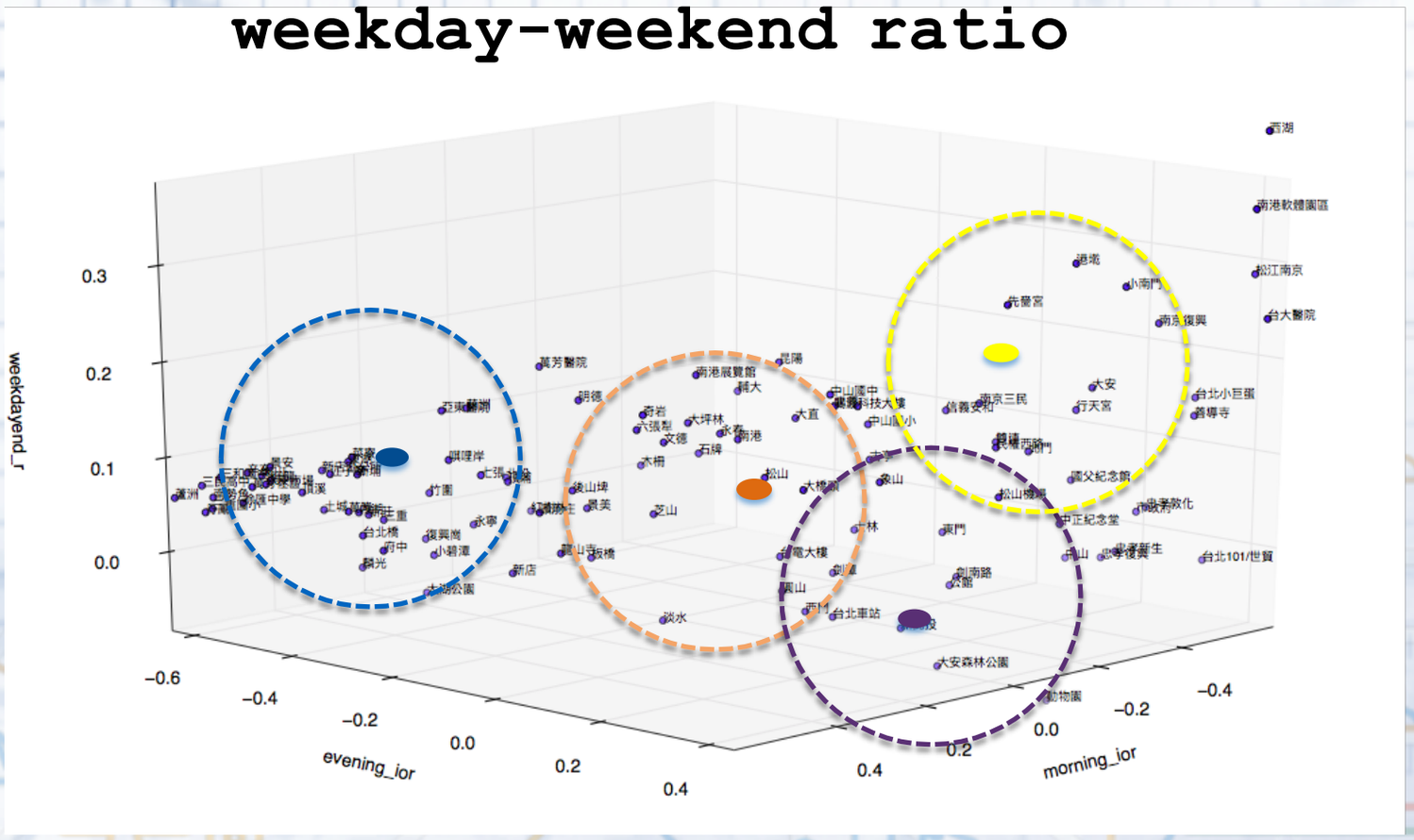
check-in average flow & check-in evening flow

Processing: Clustering

$$Ratio(X_1, X_2) = \frac{X_1 - X_2}{X_1 + X_2}$$



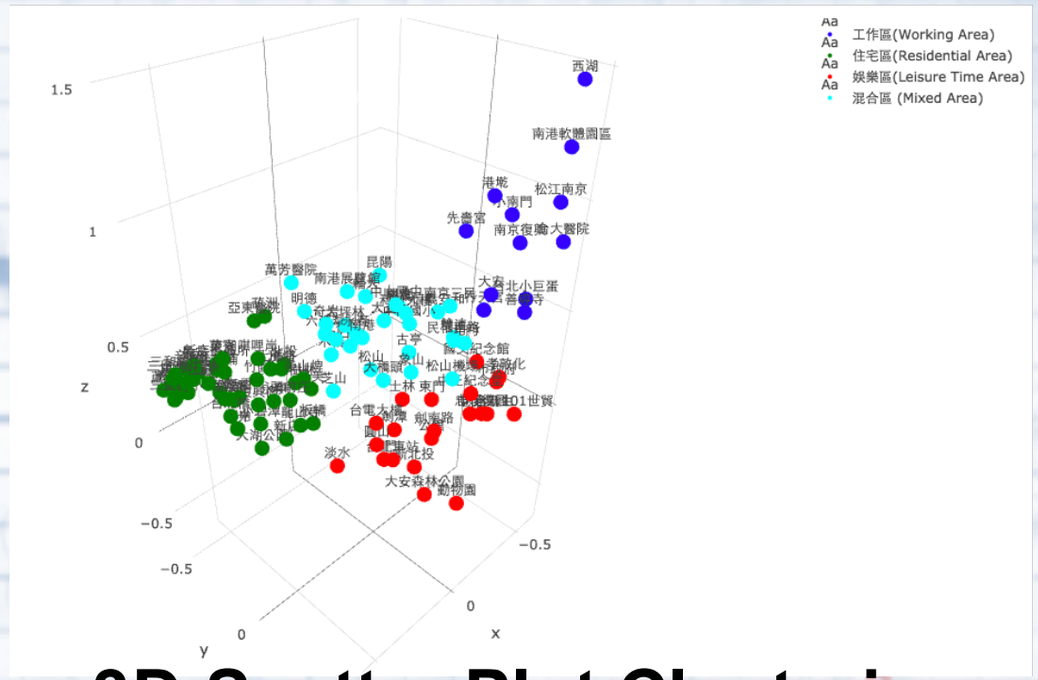
`evening-check-in-out ratio,`
`morning-check-in-out ratio,`
`weekday-weekend ratio`



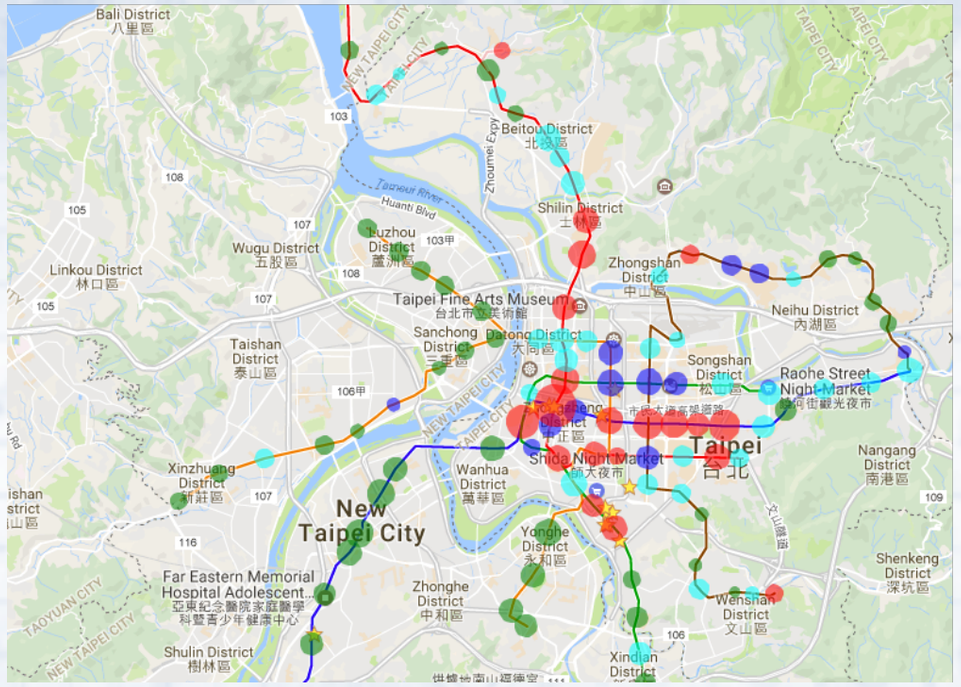
K-means Clustering Algorithm

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_j - \mu_i||^2)$$

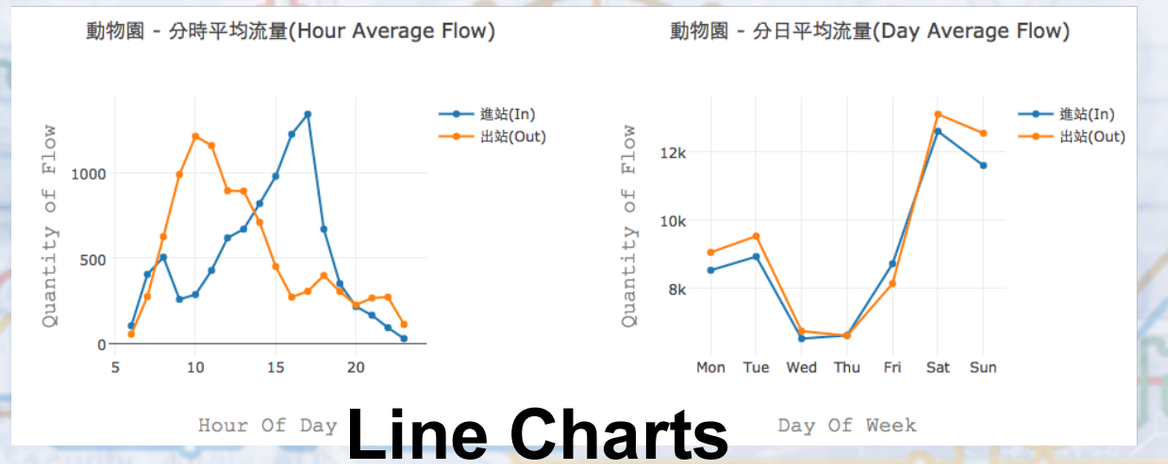
Processing: Visualization



3D-Scatter Plot Clustering



Geographical Network Flow Map



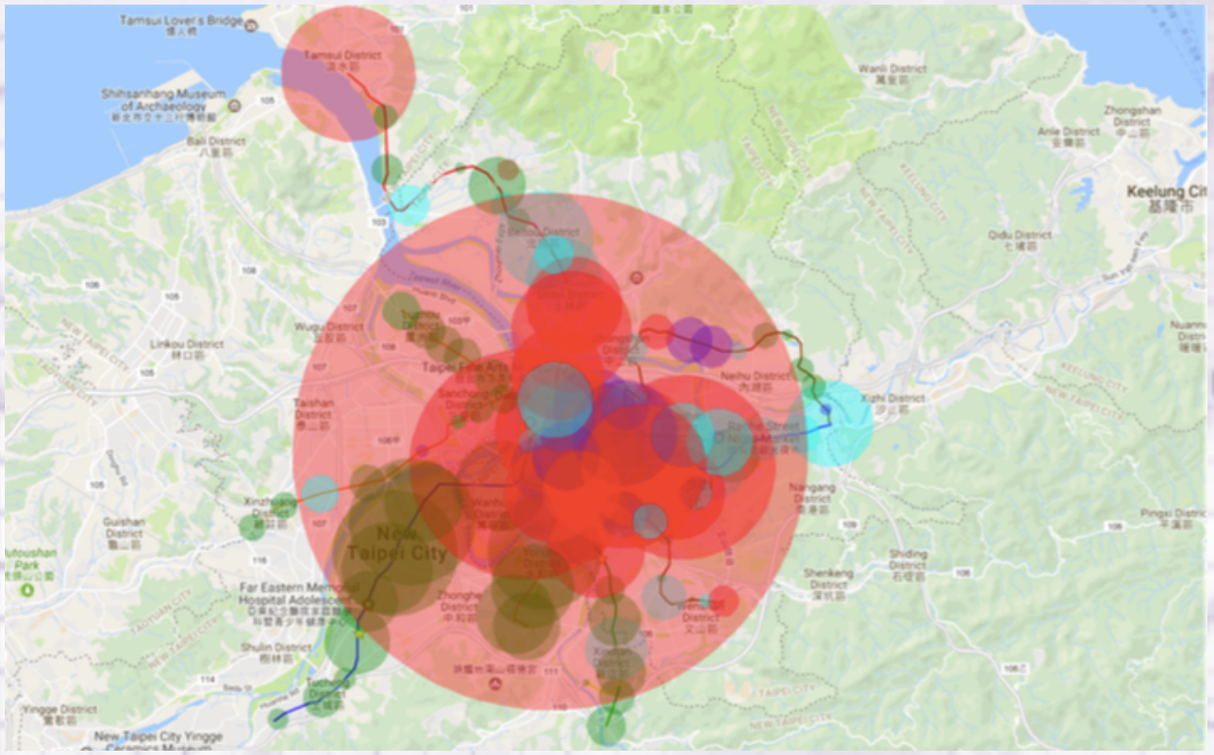
Line Charts

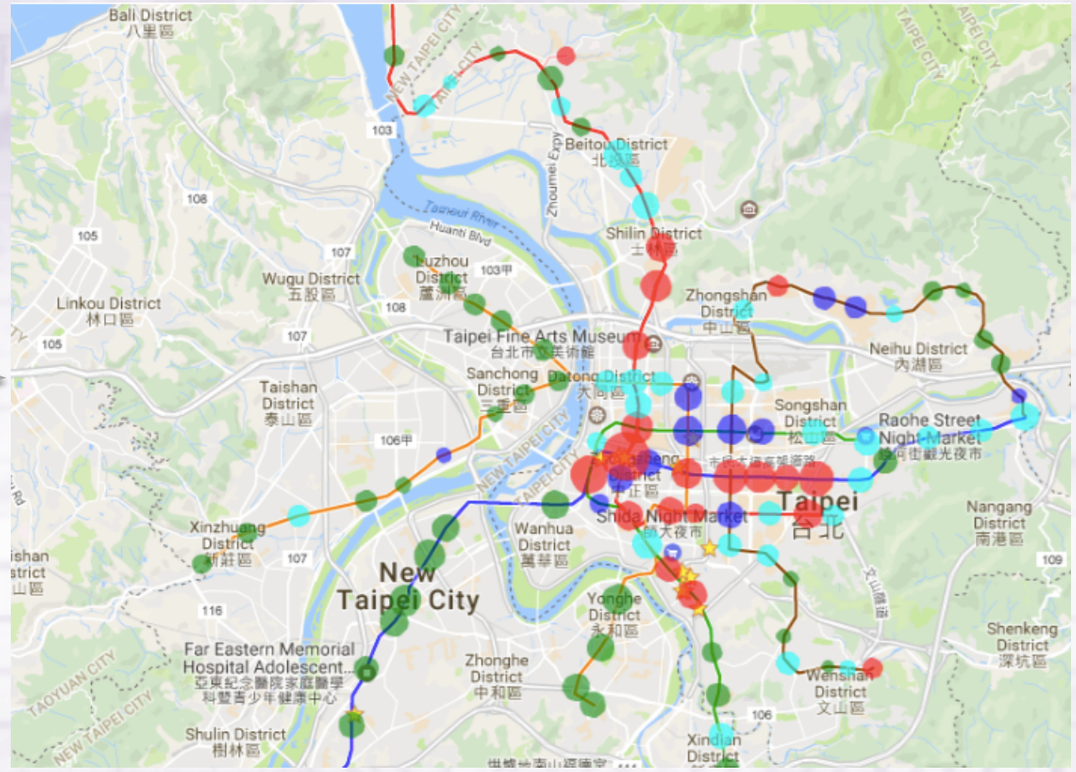
Result

Observation Result Demo

Processing: Observation

- **Extrem Value**



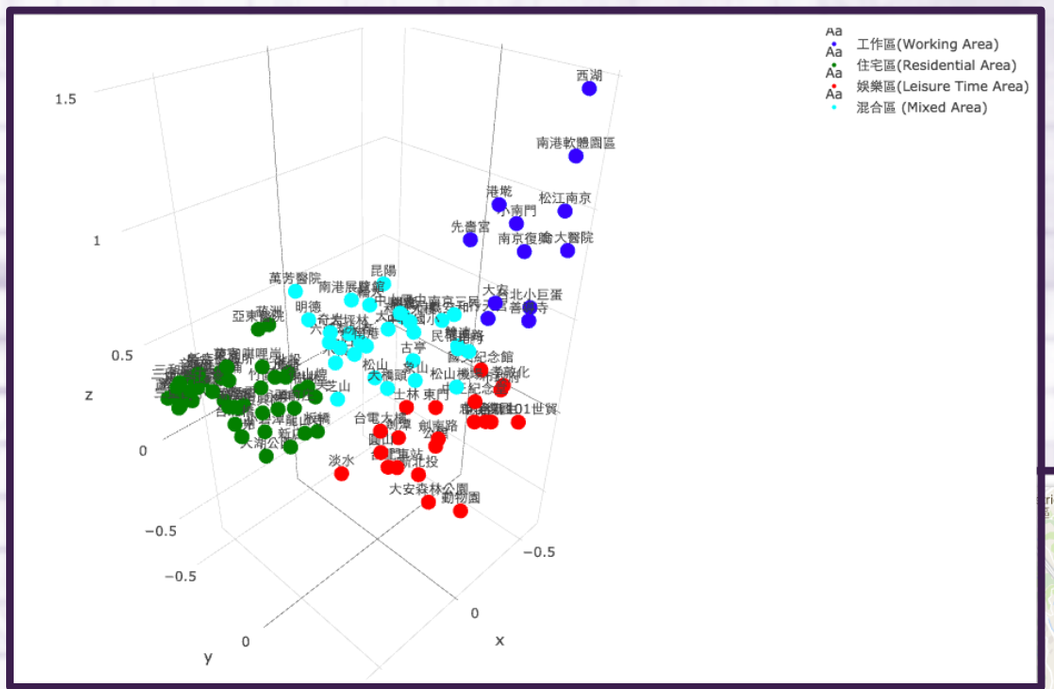


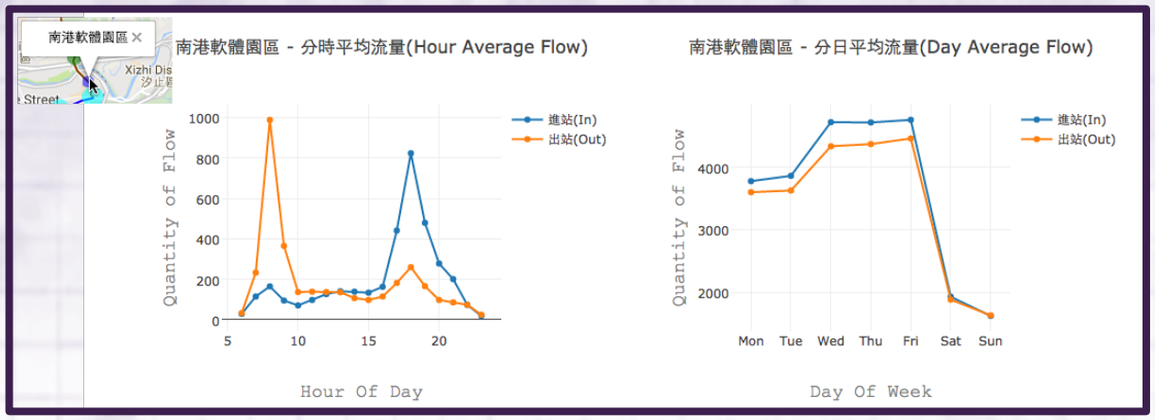
- Taipei Main Station: BIG! BIG! BIG!
- Normalize: Make their size closer

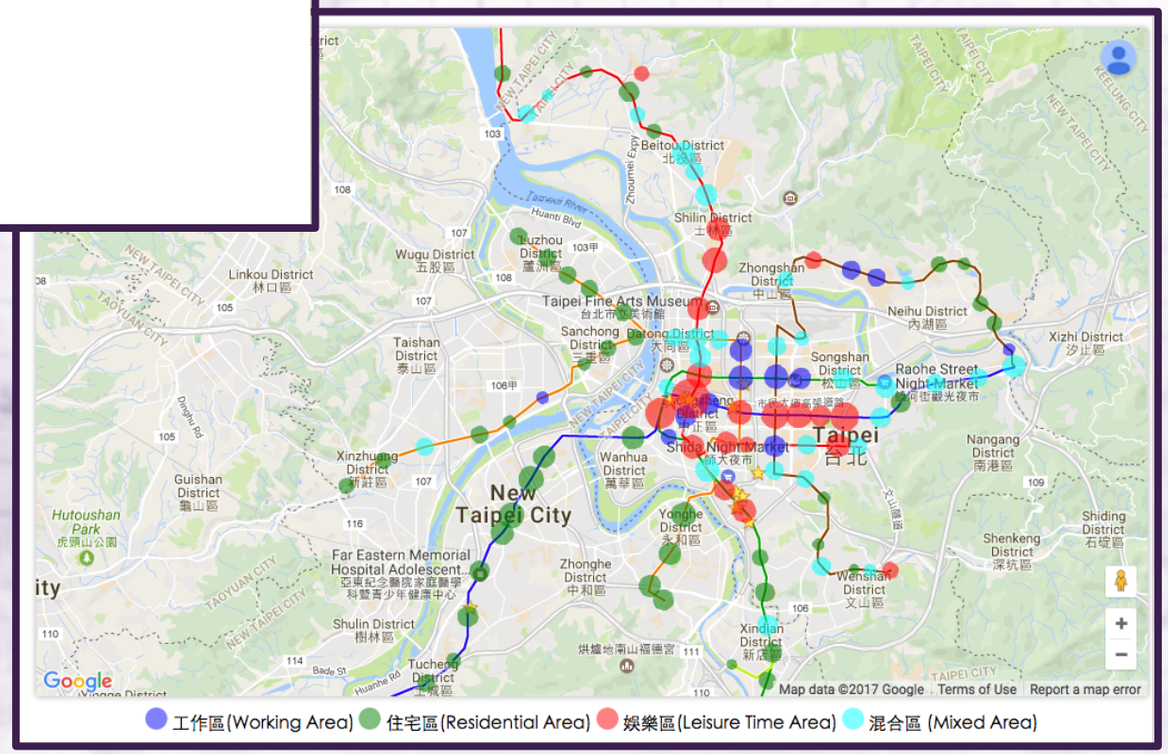
$$X' = 5 * X^{1/3}$$

Demo

https://goo.gl/H71dxp



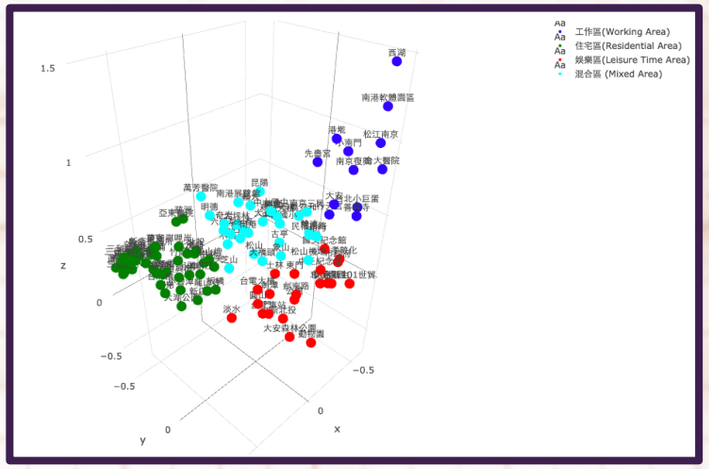




Conclusion

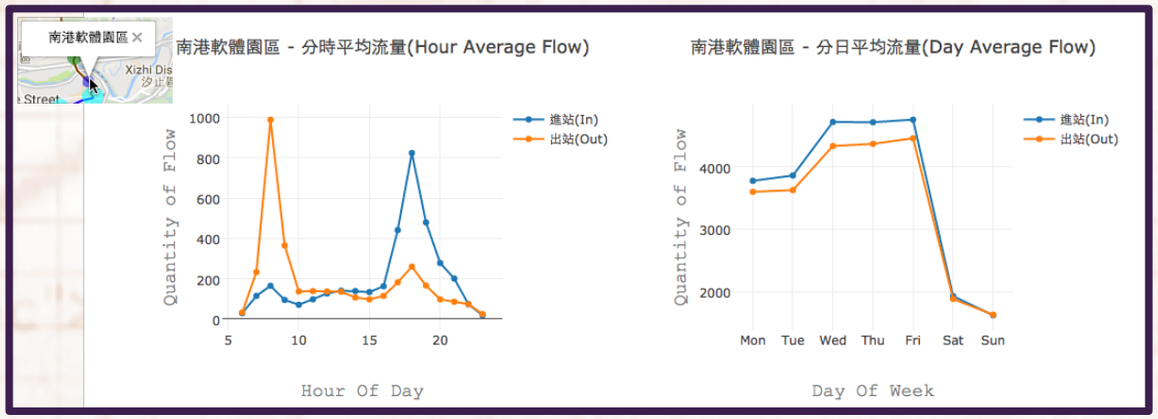
Discussion Future_Work

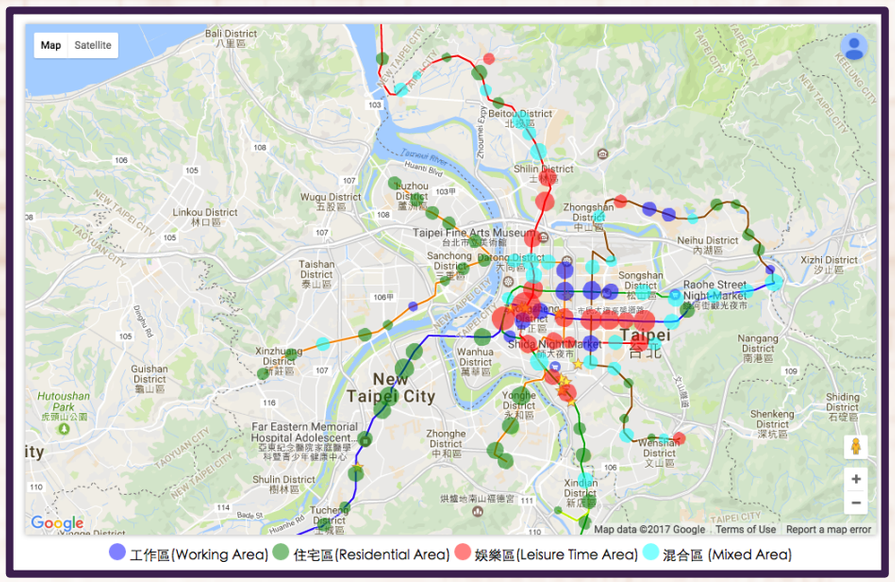
Design of Visualization



To show correlation between stations.

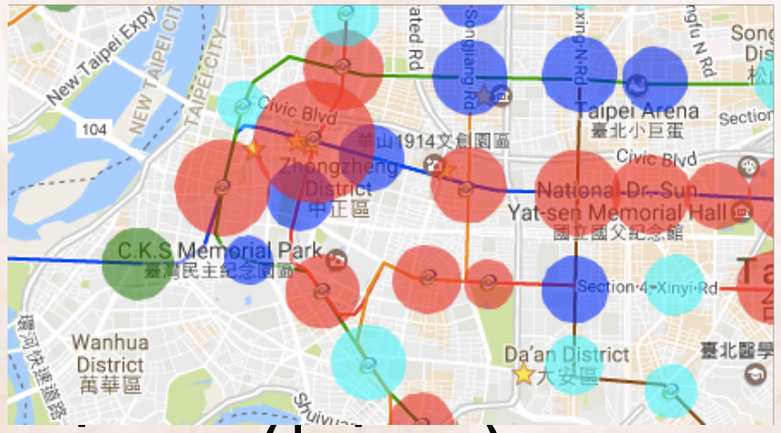
The daily/hourly flow of in/out-checking at each station.





The flow size and type of station on a geographical map.

Discussion

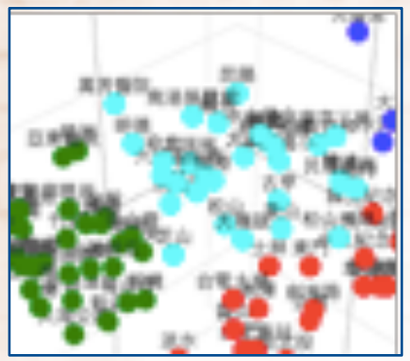


- Visualization
  - Adjustment of the flow circles size. (bias)
  - Use of Mixed-Category not precise enough. (What are they?)
  - Need better features on 3d-plot (parameters). (More distinction in grouping)

- Conclusions
  - Urban and rural development is not balanced
  - The distinction between residential area and working area is not very large.(many mixed area)

Future Work: (Blow a COW)

- **Use More data**
  - Passenger Behavior over seasons
  - Flow on holidays
  - Crowd flow changes over years
- **System Connect to Taipei MRT Company**
  - Automatic update data through API
  - Route Recommend system for passengers
- **Improvement of Clustering:**
  - More features to make the group more precise
  - Improve the category in Mix Area.



References

Sun, L., Lee, D.H., Erath, A. Huang, X. (2012) Using Smart Card Data to Extract Passenger's Spatiotemporal Density and Train's Trajectory of MRT System. Retrieved from:https://pdfs.semanticscholar.org/b692/1f2bdb54e6a476e0ee0730bb957a6bd249a3.p df (03 November 2016)

Tsao, C. (2015). A Day in Taipei Metro. Retrieved from: http://missmoss.github.io/taipeimrtviz/ (07 November 2016)

Wang, J. (2016).台北捷運路網與階段性通車規劃.Taipei Metro network and staging operation planning (Presentation). Retrieved from: http://www.metrotaipei20.com/index.php/en/programataglance/ (31 December 2016).

Wei, Y., Lin, S., Chu, R., Tian, Q. & Fei, W. (2016). A method of grading subway stations. *Procedia Engineering, 137,* 806810.

Thanks for your attention.

Q&A