FLIGHT CUSTOMER CLUSTERING

Milka Vincentiya

Background

The objective is to see customer segmentation from an airline company through all transactions (the customers must have purchased tickets and/or exchanged points).

Related Works

The *Flight Customer* dataset is mostly used for LRFMC models (one of them could be accessed here) with the detail of meaning and calculation as stated in the table below.

	Indicators System	Meaning	Symbol	Units	Value change	Definition	Weight
LRFMC model	LOAD_TIME - FFP_DATE	Number of months from the end of the observation window for membership	L	Month	↑	Length of passenger relationship	0.039
	DAYS_FROM_LAST_TO_END	The last flight time to the end of the observation window	R	Day	↓	the length of the passenger's last consumption	0.088
	FLIGHT_COUNT	Number of flights	F	Time	↑	consumption frequency within a certain period of time,	0.239
	SEG_KM_SUM	Total flight kilometers in observation window	M	Kilometre	↑	Upgrade mileage within a certain period of time	0.123
	avg_discount	Average discount rate	C	Percentage %	↑	The average space discount coefficient during a certain period of time	0.511

Note: The symbol "↑" indicates that the larger, the better, "↓" indicates that the smaller, the better.

In this research, there are only 5 columns that match the meaning of LRFMC and proceed into the modeling. The machine learning used is K-Means Clustering, with the number of clusters is 5 (using Elbow Methods and Silhouette Score).

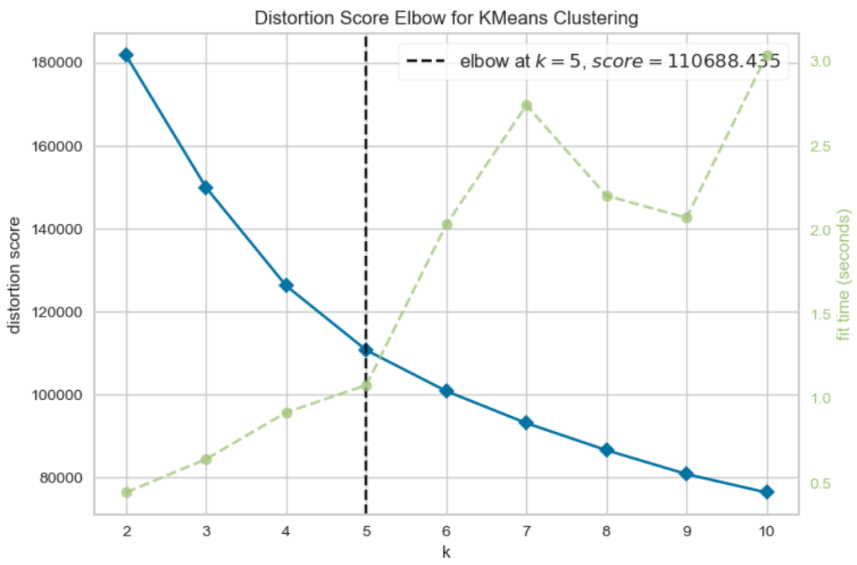


Chart of Elbow Methods and Silhouette score to determine the number of clusters.

The description of 5 customer clusters: important development passengers, important retention passengers, important maintenance passengers, general passengers, and low-value passengers based on a large number of references.
The clustering result.

Result	Number	Average score	Category	level
4	2571	1.918	important maintain passengers	Smaller
1	8670	0.725	important development passengers	medium
3	20980	0.153	important retention passengers	Biggest
2	20191	-0.306	general passengers	Bigger
5	9472	-0.871	low value passengers	medium

The explanation and suggestion of the result.

- Passengers that should be maintained.
  C (the average discount rate) is higher, R is lower, the F(frequency)and M (mileage)are also higher. They are the value travelers of airlines, which is the most ideal type of passengers. They contribute the most to airlines but their proportion is relatively small. Airlines should give priority to their resources, increase the loyalty and satisfaction of such passengers, and maximize the high level of consumption of such passengers.

- Passengers that should be cultivated.
  C (the average discount rate)is higher, and the cabin level is high with a lower, but the F (frequency) and M (mileage) are lower. Some of these passengers have just become members of the company (L is low), and some of them are old members but do not often travel on the company. This paper believes that such passengers are the potential value passengers of airlines, occupying a high proportion of passengers, and have a good development space in the future. Airlines must strive to strengthen the satisfaction of such passengers, making them gradually become loyal passengers of the company.
- Passengers that should be detained.
  C (the average discount rate), the F (frequency) and M (mileage) are higher, but R is lower. The value of their passengers varies greatly and the reasons for the changes vary. Airlines should take certain marketing measures based on the relevant consumption indicators of these passengers to prevent the loss of passengers.
- General passengers and low-value passengers.
  The current value and growth potential are low. Compared with other tourists, they cannot create more value for the company. The company should not waste too much resources on these passengers. More manpower and resources should be invested in more valuable travelers.

Dataset and Features

There are 22 columns.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62988 entries, 0 to 62987
Data columns (total 23 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   MEMBER_NO          62988 non-null  int64
 1   FFP_DATE           62988 non-null  object
 2   FIRST_FLIGHT_DATE  62988 non-null  object
 3   GENDER             62985 non-null  object
 4   FFP_TIER           62988 non-null  int64
 5   WORK_CITY          60719 non-null  object
 6   WORK_PROVINCE      59740 non-null  object
 7   WORK_COUNTRY       62962 non-null  object
 8   AGE                62568 non-null  float64
 9   LOAD_TIME          62988 non-null  object
 10  FLIGHT_COUNT       62988 non-null  int64
 11  BP_SUM             62988 non-null  int64
 12  SUM_YR_1           62437 non-null  float64
 13  SUM_YR_2           62850 non-null  float64
 14  SEG_KM_SUM         62988 non-null  int64
 15  LAST_FLIGHT_DATE   62988 non-null  object
 16  LAST_TO_END        62988 non-null  int64
 17  AVG_INTERVAL       62988 non-null  float64
 18  MAX_INTERVAL       62988 non-null  int64
 19  EXCHANGE_COUNT     62988 non-null  int64
 20  avg_discount       62988 non-null  float64
 21  Points_Sum         62988 non-null  int64
 22  Point_NotFlight    62988 non-null  int64
dtypes: float64(5), int64(10), object(8)
```

From the info of the dataset, we can see that:

- There are 8 columns with object data type (not numeric) and 15 columns with numeric data type (int & float).
- From all those columns, some columns are supposed to be in another data-type (e.g: *MEMBER_NO* not supposed to be *int* because it is not something calculated, and other datetime columns).
- There are around 1-3% of missing values in 7 columns, those columns are: *GENDER*, *WORK_CITY*, *WORK_PROVINCE*, *WORK_COUNTRY*, *SUM_YR_1*, *SUM_YR_2*, and *AGE*.

Statistical summary

	FFP_TIER	AGE	FLIGHT_COUNT	BP_SUM	SUM_YR_1	SUM_YR_2	SEG_KM_SUM	LAST_TO_END	AVG_INTERVAL	MAX_INTERVAL	EXCHANGE_COUNT	avg_discount	Points_Sum	Point_NotFlight
count	62988.000000	62568.000000	62988.000000	62988.000000	62437.000000	62850.000000	62988.000000	62988.000000	62988.000000	62988.000000	62988.000000	62988.000000	62988.0000	62988.000000
mean	4.102162	42.476346	11.839414	10925.081254	5355.376064	5604.026014	17123.878691	176.120102	67.749788	166.033895	0.319775	0.721558	12545.7771	2.728155
std	0.373856	9.885915	14.049471	16339.486151	8109.450147	8703.364247	20960.844623	183.822223	77.517866	123.397180	1.136004	0.185427	20507.8167	7.364164
min	4.000000	6.000000	2.000000	0.000000	0.000000	0.000000	368.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000
25%	4.000000	35.000000	3.000000	2518.000000	1003.000000	780.000000	4747.000000	29.000000	23.370370	79.000000	0.000000	0.611997	2775.0000	0.000000
50%	4.000000	41.000000	7.000000	5700.000000	2800.000000	2773.000000	9994.000000	108.000000	44.666667	143.000000	0.000000	0.711856	6328.5000	0.000000
75%	4.000000	48.000000	15.000000	12831.000000	6574.000000	6845.750000	21271.250000	268.000000	82.000000	228.000000	0.000000	0.809476	14302.5000	1.000000
max	6.000000	110.000000	213.000000	505308.000000	239560.000000	234188.000000	580717.000000	731.000000	728.000000	728.000000	46.000000	1.500000	985572.0000	140.000000

The modus of *EXCHANGE_COUNT* is 0, meanwhile the goal is to see the customer segmentation where the transaction is already done. The data is filtered to all

EXCHANGE_COUNT that is more than 0.

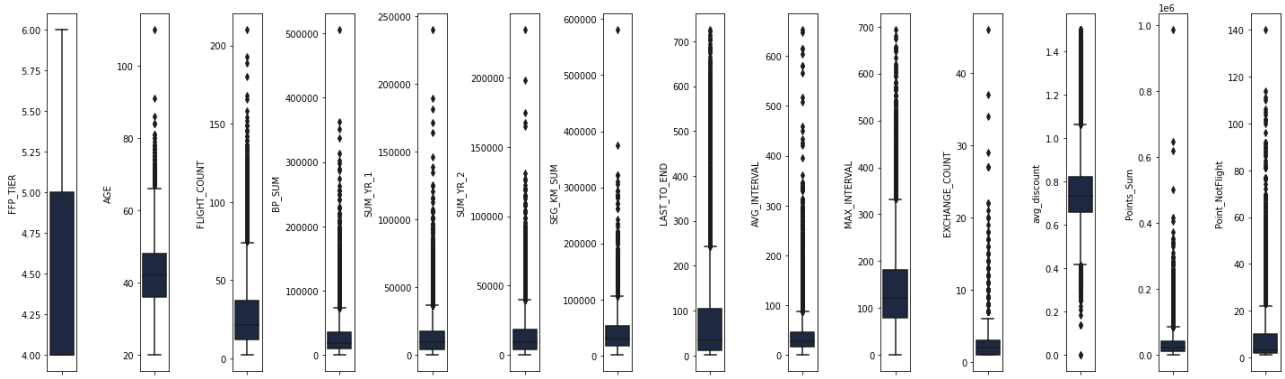
Null-values

```
GENDER              1
FFP_TIER            0
WORK_CITY         244
WORK_PROVINCE     326
WORK_COUNTRY        3
AGE                 0
FLIGHT_COUNT        0
BP_SUM              0
SUM_YR_1            0
SUM_YR_2            0
SEG_KM_SUM          0
LAST_TO_END         0
AVG_INTERVAL        0
MAX_INTERVAL        0
EXCHANGE_COUNT      0
avg_discount        0
Points_Sum          0
Point_NotFlight     0
dtype: int64
```

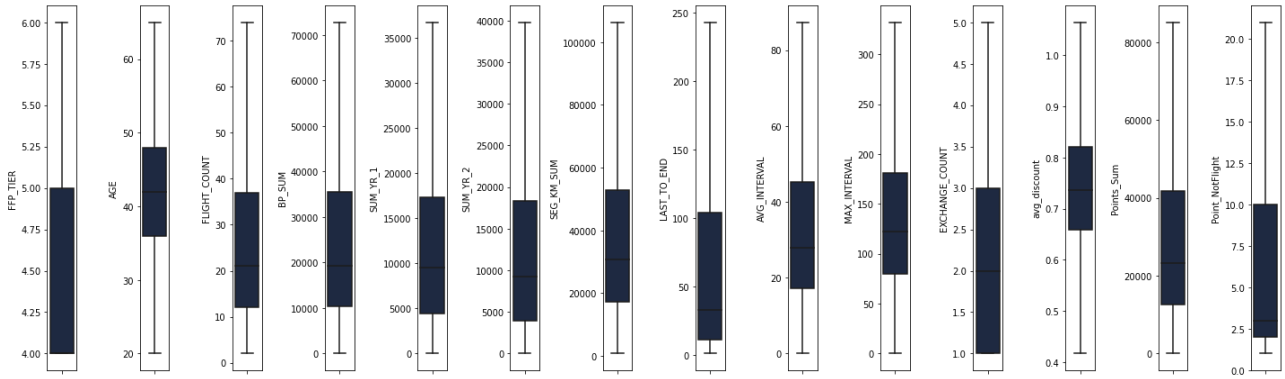
There are null-values in some columns which contain string data. The nulls are filled with 'Not Applicable' as it is not possible to determine the values. Another thing to consider is that the columns with null-values are those that contain anomaly data and it is not handled for this project.

Outliers

The project will also try to make 2 models run with handling outliers and without handling outliers. In consideration that it is normal when there are some customers that traveled more kilos and purchased expensive tickets than the rest of the customers.



The Boxplot Without Handling the Outliers



The Boxplot After Handling the Outliers

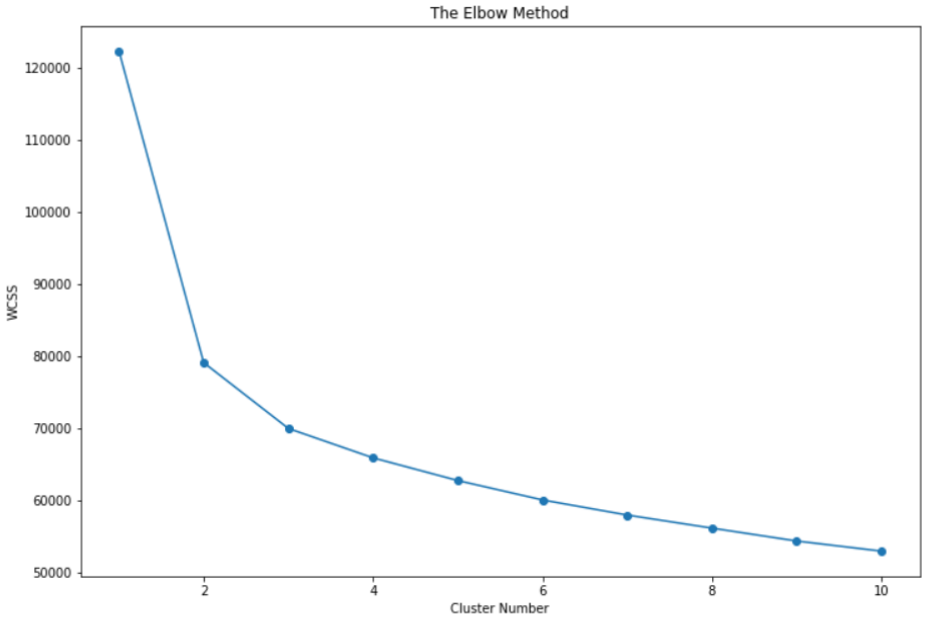
The outliers are handled by replacing the data with the score of Q1 and Q3.

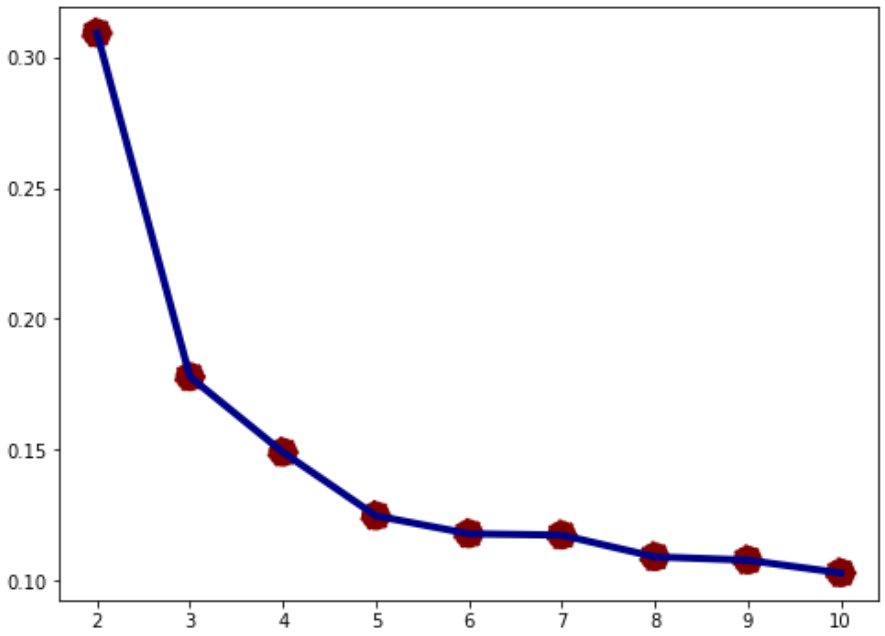
Methods

In this project, the output is clusters, so the focus is on the column with integer (not a string data).

1. Create a new dataset with all numerical columns.
2. Using K-Means to create clusters.
3. After the number of clusters is decided, use PCA to divide the data.
4. Turn the PCA-ed data into the real dataset to see the summary.

The Machine Learning that is used is PCA and K-Means Clustering where the number of clusters is decided based on the *Elbow Method* and *Silhouette Score*.





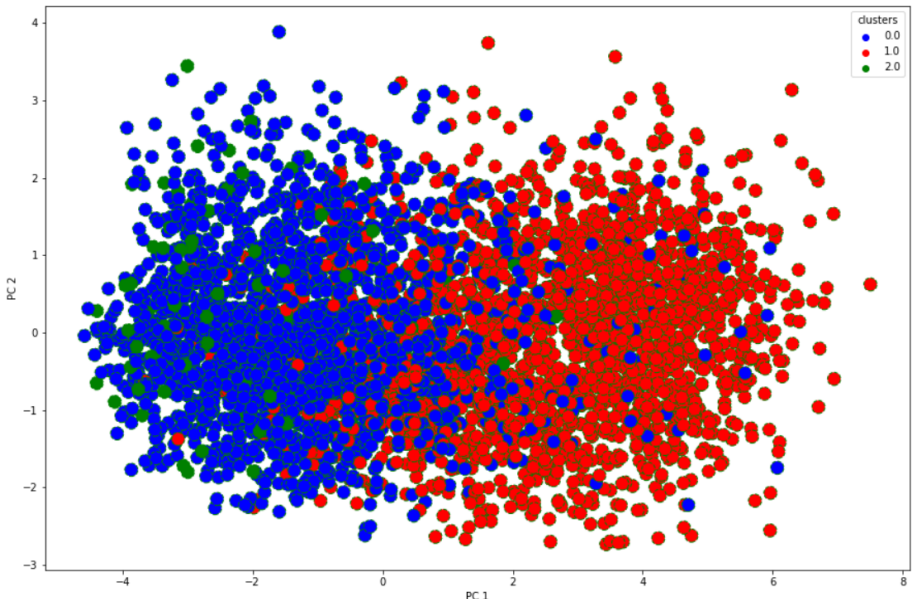
Silhouette Score Chart

As seen on the charts, the steepness of the Elbow Methods is stopped on the 3 and the highest silhouette number is 2. So, I did 2 cases where the cluster is 2 and where the cluster is 3.

Experiment and Discussion

Here is the result of the clustering.

Cluster = 3



The three clusters are not so well divided, only cluster number 0 and 1 that are obviously separated (blue and red). With the total number per cluster:

- Cluster 2: 3542.
- Cluster 1: 2119.
- Cluster 0: 3073.

In the final result and conclusion, 6 features are selected to be concluded:

- **Regular Customer & Medium Spender (Cluster 0, total: 35,2%)**: book domestic and/or international flight (once a month) and might use points for discount. Could be someone who needs to fly for a business meeting or visit family.
  median of booked flights = 26,
  median of km of flights = 38.383 km,
  median of interval between flights = 28,
  median score of exchange = 2 times.
  median score of total points = 26.072,
  average of discount = 73%.
- **Loyal Customer & High Spender (Cluster 1, total: 24,3%)**: book domestic but mostly international flights frequently (once per 10 days) and would use points for discount. Could be a business person who needs to fly to see their clients.
  median of booked flights = 44,
  median of km of flights = 46.070 km,

median of interval between flights = 10,
median score of exchange = 3 times.
median score of total points = 112.843,
average of discount = 77%.

- **Occasional Customer (Cluster 1, total: 40,5%)**: only book domestic flight if necessary (once per 3 months), even without exchanging points. Maybe for holiday only.
  median of booked flights = 11,
  median of km of flights = 15.672 km,
  median of interval between flights = 95,
  median score of exchange = 1 times.
  median score of total points = 11.091,
  average of discount = 70%.

The **number of loyal customer's number is the lowest**. The loyal customers already spent more tickets and flight more kilos, but received almost the same average of discount with the other 2 clusters.

Here is another conclusion for the dataset with outliers.

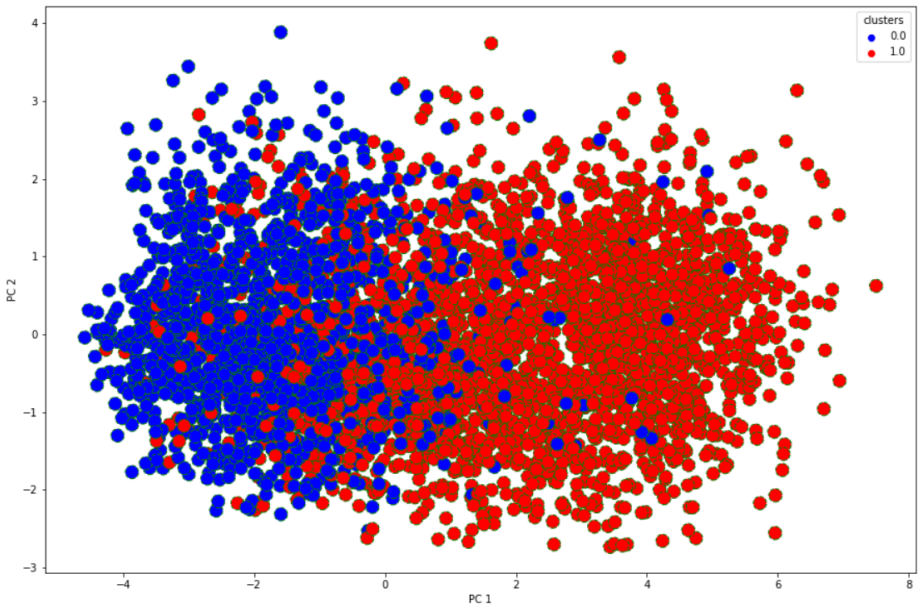
- **Loyal Customer & High Spender (Cluster 0, total: 7%)**
  median of booked flights = 77,
  median of km of flights = 115.113,5 km,
  median of interval between flights = 7,
  median score of total points = 112.843,
  median score of exchange = 5 times.
- **Occasional Customer (Cluster 1, total: 58,6%)**
  median of booked flights = 14,
  median of km of flights = 19.541 km,
  median of interval between flights = 73,
  median score of total points = 14.283,
  median score of exchange = 1 times.
- **Regular Customer & Medium Spender (Cluster 2, total: 34,4%)**
  median of booked flights = 38,
  median of km of flights = 53.558 km,
  median of interval between flights = 14,
  median score of total points = 40.899,
  median score of exchange = 2 times.

The number of loyal customers is still the lowest, it is even more obvious with outliers.

Suggestion:

- The airlines could push a program for those customers in occasional customers during a high season.
- A special program for the loyal customers to stay loyal to the airlines (e.g: reward program, silver/platinum member). Because the average discount is almost the same for 3 clusters. It could also encourage the non-customer or frequent ones to also fly more with the airlines.

Cluster = 2.



The cluster separation is more obvious than in 3 clusters.
The conclusion:

- **Regular Customer (Cluster 0, total: 68,7%)**: book domestic and/or international flight (once in 2 months) and might use points for discount. Could be someone who needs to fly for a business meeting or visit family.
  median of booked flights = 16,
  median of km of flights = 22.169 km,
  median of interval between flights = 59,
  median score of exchange = 1 time,
  average of discount = 71%,
  median score of total points = 16.193.
- **Loyal Customer (Cluster 1, total: 21.3%)**: book domestic but mostly international flights frequently (once per 12 days) and would use points for discounts. Could be a business person who needs to fly to see their clients.
  median of booked flights = 47,
  median of km of flights = 65.989 km,
  median of interval between flights = 12,
  median score of exchange = 3 times,
  average of discount = 77%,
  median score of total points = 54.275.

There are some changes in median values that make both the regular customer and the loyal customer more obvious and segmented. Even though the total number of loyal customers is getting lower while the regular one is getting higher. This could be

because other features are not calculated. Otherwise, it is still obvious that the discount received by both groups is almost the same.

Future Works

There are some things can be future researched:

- Use the datetime data to know when the peak of customers and to know whether it is a first-time customer or customers that already did a transaction before.
- The details of each cluster by their gender, city, province, or country.From the distribution, the airlines could find which city to focus on and what kind of promotion to apply to this customer segment.
- The relation between customer segmentation and airlines' revenue (timeseries).

Reference

Zeyu Zhou et al 2019 J. Phys.: Conf. Ser. 1168 032086