International Journal of
Advanced Science and Convergence

ISSN : 2619-8150
Volume 2, Number 4, December 2020
https://doi.org/10.22662/IJASC.2020.2.4.025

A Study on Changes in Foreign Tourists Using Big Data Analysis Method

Jeongwon Lee[1] and Choong Ho Lee[2]

[1]Student, Department of Information Communication and Engineering, Hanbat National University, Daejeon, 34158, Republic of Korea

[2]Professor, Department of Information Communication and Engineering, Hanbat National University, Daejeon, 34158, Republic of Korea

Abstract

Background/Objectives: In this study, by analyzing foreign tourists entering the country by country, we tried to study the changes in the inflow of foreign tourists according to the monthly pattern. **Methods/Statistical analysis**: For the purpose of using statistical data on foreign tourists entering the country for the purpose of using it as basic data for establishing tourism policy and marketing strategy by the Korea Tourism Organization, data on monthly foreign tourist statistics are collected, and pandas and matplotlib technologies provided by Python are used. Thus, data preprocessing and time series data were analyzed. **Findings**: The number of foreign tourists steadily increased from 2010 to 2016, but the number of tourists decreased sharply in early 2017. It was analyzed that the number of tourists gradually increased from mid-2017. Every summer, the influx of tourists was analyzed to be very high, and especially in the summer of 2015, the number of tourists was analyzed to have decreased significantly. **Improvements/Application**: In this study, the trend of monthly tourists by country was analyzed, and if the relationship between social issues in the section in which the change trend curve rapidly decreases, it is likely to be possible to analyze the factors that influence the change of tourists.

Index Terms

Big data, Tourism, Tourists, Analytics, Foreigners.

Corresponding author : Choong Ho Lee
chlee@hanbat.ac.kr

• Manuscript received November 24, 2020.
• Revised December 15, 2020 ; Accepted December 20, 2020.
• Date of publication December 31, 2020.

I. INTRODUCTION

More and more people are concerned about how to use data analysis rather than data analysis itself. Data analysis is not a result itself, but a process that is used to make any conclusion or result.

In this study, data on foreign tourist statistics from 2010 to 2019 were collected based on an effective tool for conducting data analysis and an understanding of the industry. Through pre-processing of data, this study analyzes seasonal patterns of foreign tourists by nationality and events in which foreign tourists visit increases or decreases.

II. THEORETICAL BACKGROUND

A. *Data Analysis Tool*

As tools for data analysis, there are various tools such as Excel, Tableau, and R, but in this study, Python, which many developers are currently using, was used [1].

Compared to other tools, Python has a feature that it does not require a separate cost to use Python, and there are various libraries that many developers already use and can utilize. And it has a great advantage that it can be used universally [2].

B. *Understanding of industry*

Data analysis should be able to be used to achieve the goal based on the result of analyzing the data, not the purpose itself, and at this time, understanding of the industrial process is essential.

Even if you analyze and examine the data, if you do not understand the derived numbers or the meaning behind the results, you cannot understand the results of analyzing the data and it is difficult to use them. Understanding the industry is called domain knowledge.

III. RESEARCH ANALYSIS

A. *Analysis Method*

As the basic data for establishing tourism policy and marketing strategy, data of foreign tourists visiting Korea provided by Korea Tourism Organization were collected. This data provides data by gender, purpose, age, means of transportation, and nationality. In this study, data by nationality and purpose among monthly outpatient entrants were used.

In order to classify foreigner data entering Korea for tourism purposes by nationality, data preprocessing and time series analysis were performed using pandas and matplotlib technologies provided by Python [3-4].

For reference, the data used in the analysis are data provided by the Korea Tourism Organization, and because the original analysis was meaningful, the Korean language shown in the figures and tables below was expressed as a result of analysis, so this study has no relation to separate data translation. To inform.

B. *Analysis Tools*

In this study, to analyze foreign tourist data, Anaconda, which provides an effective Python program for data analysis and various libraries, was used as shown in “Fig.1” below.

Fig. 1. Python programming and Anaconda

Python programs are high-level programming languages, platform-independent, interpreted, object-oriented, and dynamic typing interactive languages. Anaconda is managed through conda, a package management system used for large-scale data processing and predictive analysis for the purpose of simplifying package management and deployment.

C. *Analysis Library*

In this paper, foreign immigration data was analyzed using the pandas, matplotlib, and seabon libraries provided by the Python program as shown in “Fig.2”.

Fig. 2. Matplotlib and Seaborn

Pandas is a library for processing Python data and is used as an essential library for tasks such as data analysis using Python. Pandas can easily manipulate data in the form of a table consisting of rows and columns, and provides various functions and usage methods for data processing.

Matplotlib is a Python library used for data visualization and 2D graph plotting. It provides a

number of functions and usage methods to easily draw various types of graphs.

Seaborn is a library for drawing graphs like matplotlib. It provides basic colors that are neatly implemented without special decoration, and provides a palette function that enables more beautiful graphing [5]. Also, due to its high compatibility with pandas data frames, it is used very often as a visualization function with matplotlib.

In this study, the pandas library was used to collect and pre-process a lot of data from 2010 to 2019, and the monthly and seasonal changes that change with the passage of time can be checked through a time series graph using the matplotlib and seaborn libraries [6-7]. In addition, a heat map graph was used to check the occurrence of events by nationality.

D. Analysis Result

One sample data among the year and month foreign inbound travelers data used for this analysis is analyzed in advance, and the algorithm used for the analysis is applied to the entire year and month data.

The following is the property of the sample data. It is data in the form of a data frame as shown in "Fig. 3", and consists of a total of 7 variables: nationality, tourism, commercial, public, study abroad/training, and others, and the minimum value for each variable, Basic statistical data such as the median value and the maximum value are shown in "Table.1" below.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 67 entries, 0 to 66
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Nationality  67 non-null     object
 1   Tourism      67 non-null     int64
 2   Commercial   67 non-null     int64
 3   Public       67 non-null     int64
 4   Study Aboad   67 non-null     int64
 5   Etc          67 non-null     int64
 6   Sum          67 non-null     int64
dtypes: int64(6), object(1)
memory usage: 3.8+ KB
```

Fig. 3. Data attributes of foreigners

TABLE 1. Basic statistical data for foreigners

	Tourism	Commercial	Public	Study Aboad	Etc	Sum
count	67.00000	67.000000	67.000000	67.000000	67.000000	67.000000
mean	26396.80597	408.208955	132.507463	477.462687	5564.208955	32979.194030
std	102954.04969	1416.040302	474.406339	2009.484800	17209.438418	122821.369969
min	0.00000	0.000000	0.000000	0.000000	16.000000	54.000000
25%	505.00000	14.500000	2.500000	17.500000	260.000000	927.000000
50%	1304.00000	45.000000	14.000000	43.000000	912.000000	2695.000000
75%	8365.00000	176.500000	38.000000	182.000000	2824.500000	14905.500000
max	765082.00000	10837.000000	2657.000000	14087.000000	125521.000000	916950.000000

In order to analyze the time series using the data of foreign immigrants, as shown in "Table.2", the base year/month variable was added, and the ratio of tourists by nationality and the total ratio of foreign tourists were analyzed.

TABLE 2. Year/Month and tourist ratio based on foreigners

	Nationality	Tourism	Commercial	Public	Study Aboad	Etc	Sum	Base Year/Month	Continent	Tourists ratio	Overall Ratio
0	일본	198805	2233	127	785	4576	206526	2019-01	아시아	96.3	22.5
1	대만	86393	74	22	180	1285	87954	2019-01	아시아	98.2	9.8
2	홍콩	34653	59	2	90	1092	35896	2019-01	아시아	96.5	3.9
3	마카오	2506	2	0	17	45	2570	2019-01	아시아	97.5	0.3
4	태국	34004	37	199	96	6998	41334	2019-01	아시아	82.3	3.8

Among the total nationalities, the top five nationalities with the largest proportion of foreigners entering Korea for tourism purposes are analyzed as shown in "Table.3". The largest number of foreigners entering Korea for tourism purposes was Chinese, followed by Japanese, Taiwanese, US, and Hong Kong nationalities.

Table. 3. Top 5 countries based on foreign tourists

	Nationality	Tourism	Commercial	Public	Study Aboad	Etc	Sum	Base Year/Month	Continent	Tourists ratio	Overall Ratio
17	중국	320113	2993	138	8793	60777	392814	2019-01	아시아	81.5	36.2
0	일본	198805	2233	127	785	4576	206526	2019-01	아시아	96.3	22.5
1	대만	86393	74	22	180	1285	87954	2019-01	아시아	98.2	9.8
25	미국	42989	418	2578	229	16523	62737	2019-01	아메리카	68.5	4.9
2	홍콩	34653	59	2	90	1092	35896	2019-01	아시아	96.5	3.9

In order to automatically pre-process and integrate foreign tourist data from 2010 to 2019, the repetitive data pre-processing process was programmed as a function and executed repeatedly. The preprocessing and integration results for all data are shown in "Fig.4".

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6960 entries, 0 to 6959
Data columns (total 11 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Nationality     6960 non-null   object
 1   Tourism         6960 non-null   int64
 2   Commercial      6960 non-null   int64
 3   Public          6960 non-null   int64
 4   Study Aboad      6960 non-null   int64
 5   Etc             6960 non-null   int64
 6   Sum             6960 non-null   int64
 7   Base Year/Month  6960 non-null   object
 8   Continent       6960 non-null   object
 9   Tourists ratio(%)  6960 non-null   float64
 10  Overall Ratio(%)  6960 non-null   float64
dtypes: float64(2), int64(6), object(3)
memory usage: 598.2+ KB
```

Fig. 4. Total integrated data of foreign tourists

Time series analysis was performed as shown in "Fig.5~Fig.9" for the top 5 countries with the largest number of inbound tourists using the integrated data of all foreign tourists that have been pre-processed [8-9], and hits as shown in "Fig.10~Fig.14" Map analysis was performed [10].

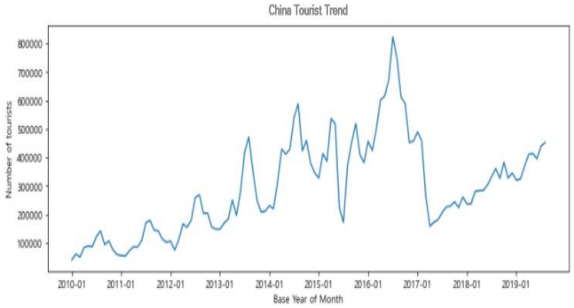

Fig. 5. Analysis on the trend of Chinese tourists entering the country

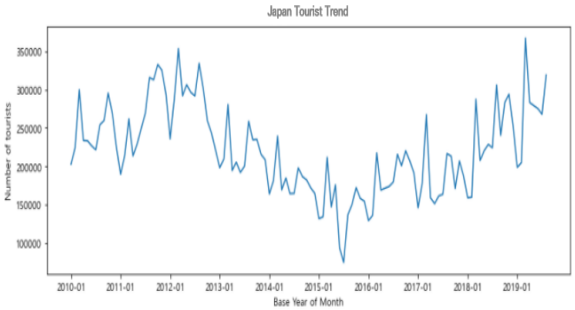

Fig. 6. Analysis of the trend of Japanese nationals entering the country

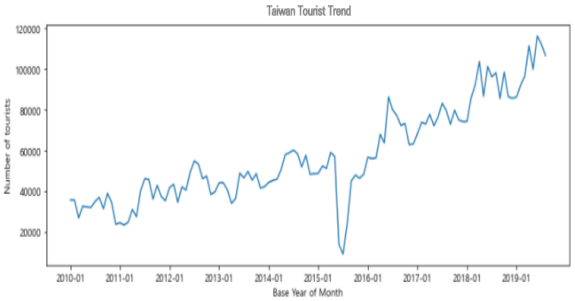

Fig. 7. Analysis of Taiwanese nationalities and tourists entering the country

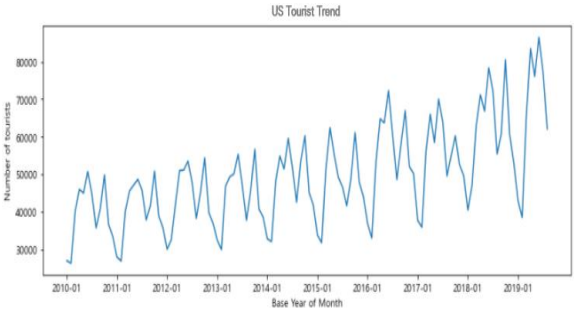

Fig. 8. Analysis of US tourists entering the country

Fig. 9. Analysis of Hong Kong nationalities and tourists entering the country

The following is the heat map analysis result.

Fig. 10. Analysis of heat map of Chinese nationals arriving tourists

Fig. 11. Heat map analysis of Japanese nationals arriving tourists

Fig. 12. Analysis of heat map of Taiwanese tourists entering the country

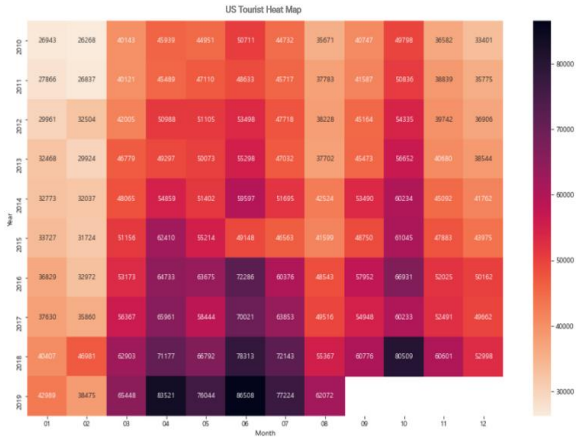

Fig. 13. Analysis of heat map for US citizens

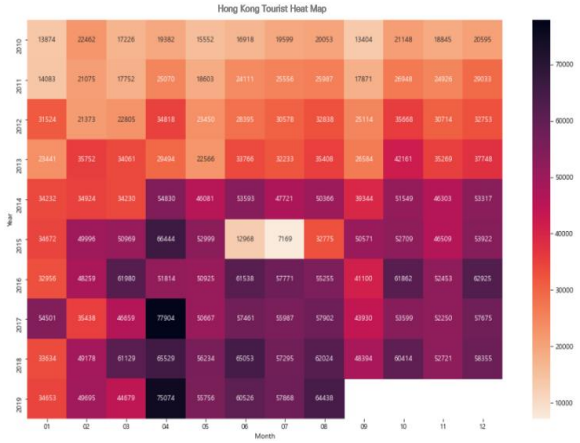

Fig. 14. Analysis of heat map of Hong Kong nationals arriving tourists

IV. CONCLUSION

The country accounting for the majority of foreign tourists visiting Korea was analyzed as China, followed by Japan, Taiwan, the United States, and Hong Kong. Among these countries, time series graphs and heat map graphs were analyzed, focusing on China, which has the most tourists.

As a result of analyzing the time series graph of Chinese tourists, the number of tourists from 2010 to 2018 shows a steadily increasing trend. It was analyzed that the number of tourists decreased significantly in early 2017. Also, from mid-2017, the number of tourists is showing a gradual increase. Seasonally, the number of tourists shows the maximum in summer every year, but it is analyzed that the number of tourists is significantly decreasing in the summer of 2015.

The heat map graph can be analyzed in detail by year and month, and the analysis results are as follows.

As the color progressed from January 2010 to March 2017, it was analyzed that the number of tourists steadily increased. Seasonally, July to August are generally dark colors, showing the pattern that most tourists visit, followed by April and October. The number of tourists is low from June to August 2015, and the number of tourists is very low from March to June 2017. However, it was analyzed that the number of tourists gradually increased from July 2017 to April 2019.

When comprehensively analyzing the time series graph and the heat map graph, Chinese tourists were the highest in summer as seasonal characteristics, followed by spring, autumn, and winter in order. In the summer of 2015 and March 2017, the number of tourists decreased dramatically, but it was analyzed that the number of tourists has continued to increase since.

Among the data analysis results, it is necessary to confirm the events in which the number of foreign tourists rapidly decreased.

From June to August 2015, there was a temporary decrease due to the aftermath of MERS, but it has been showing a normal increase since.

In March 2017, the trend of rapid decline continued due to China's retaliatory measures against THAAD, but it was analyzed that the number of tourists gradually increased afterwards.

REFERENCES

[1] W. S. Ha. (2014). Productive Small-scale Data Processing using Python. *Journal of the Korean Institute of Mineral and Energy Resources Engineers*, 51(5), 705-714.

[2] D. H. Lee. (2015). An Alternative Approach for Implementing Interactive Media Contents - a Case Study of Teaching Computer Game Programming using Python. *Korean Imaging Society*, 13(10), 145-156.

[3] D. K. Kim. (2016). Introduction to Python VII (Matplotlib). *Korean Society of Facility Engineering*, 45(6), 98-100.

[4] H. S. Kang, & H. C. Kim. (2020). A Design on Deep Learning Lecture for Computer Programming

Education. *Korea Digital Contents Society*, 21(10), 1801-1808.

[5] C. J. Lee, & M. S. Hong. (2019). Spatiotemporal Variations of Fine Particulates in and around the Korean Peninsula. *Korean Society for Atmospheric Environment*, 35(6), 675-682.

[6] K. A. Jang, K. S. Lee, & W. J. Kim. (2016). Trend of Research and Industry-Related Analysis in Data Quality Using Time Series Network Analysis. *Korea Information Processing Society*, 5(6), 295-306.

[7] K. S. Choi. (2012). Social Big Data Analysis Service: Unstructured Text Big Data Analysis and Application Service. *Korea Intelligent Information System Society*, 2012(5), 59-76.

[8] H. Jeong, C. H. Lee, H. J. Oh, Y. C. Yoon, H. K. Kim, Y. H. Jo, & C.Y. Ock. (2014). Automatic Generation of Issue Analysis Report Based on Social Big Data Mining. *Transactions on Software and Data Engineering*, 3(12), 553-564.

[9] S. J. Lee, J. S. Lee, H. Cho, & W. S. Han. (2011). A Visualization Tool for Ranked Subsequence Matching in Time-Series Databases. *Journal of KISS: Databases*, 38(2), 92-103.

[10] G. W. Bae, & H. J. Kim. (2016). A Novel Clustering Method for Analyzing Floating Population : Application to Jeju Island. *Korea Tourism and Leisure Association*, 28(1), 25-43.