Trend, Seasonality Analysis, and Enhancement Strategies for Japanese Tourists in Korea

: A Regression and Time Series Data Analysis and Forecasting

Name: Jeesoo, Kim Student ID: 20190741
Name: Daehyeon, Kang Student ID: 20190337
Name: Yungu, Lee Student ID: 20191371

Time Series Data Analysis and Forecasting (Capstone Design)
2024 Spring Semester
Professor Myungsuk, Kim

Ⅰ. Problem Motivation

The COVID-19 pandemic has significantly impacted the travel industry globally, with South Korea being no exception. The data from the Ministry of Culture, Sports, and Tourism of South Korea shows a notable decline in the rate of domestic travel within the country post-COVID-19. The annual number of domestic travelers dropped from 344,750,000 in 2019 to 225,199,000 in 2020 as shown in the below chart[1]. Although there has been a gradual recovery, the recovery remains sluggish compared to pre-pandemic levels. This trend emphasizes the need for various strategic measures to rejuvenate tourism in Korea and boost the travel economy within Korea.

*units 1000 times, %

Chart : Number of domestic travel and increasing rate

source: Ministry of Culture, Sports and Tourism, South Korea

Year	Annual	1Q	2Q	3Q	4Q
2019	344,750	84,061	80,915	90,658	89,116
2020	225,199	59,100	55,818	59,319	50,962
2021	245,127	54,277	65,810	64,588	60,452
2022	283,722	65,228	71,903	75,759	70,833
2023(B)	296,977	73,374	73,366	75,893	71,344
Growth rate (B-A)	4.7	17.1	2.0	0.2	0.7

<Chart 1: Number of domestic travel by quarters and recent increasing rate>

In addition to addressing the domestic travel slowdown, it is a possible alternative to develop strategies to attract foreign tourists to Korea as the Korean Wave is gaining popularity that K-POP special concert is going to open at Germany in Oct 2024.[2] Moreover, a special feature by The New York Times has highlighted Okdongsik, a Korean restaurant famous for its pork soup, attracting attention for New York city's top 8 dishes.[3] This indicates a growing interest in Korean cuisine and culture across various countries.

For the purposes of this research, we need to focus on a specific country to effectively analyze the impact of exogenous variables on the number of tourists from that country. Including all countries would make it difficult to identify the specific factors influencing foreign tourist numbers in Korea, as each country may have unique interests in Korea, which could act as confounding effects.

Japan has been identified as a key market due to its substantial contribution to the total number of foreign travelers visiting Korea as seen in the below Chart 2. In specific, over 20% of the foreign tourists in Korea are from Japan in average from Jan 2015 to Apr 2024.[4]

1) Ministry of Culture, Sports, and Tourism, South Korea. "Number of Domestic Travel and Increasing Rate." Accessed [JUNE 20, 2024].
2) 김보경. "한류 알려온 '케이콘' 12년 역사 처음으로 9월 독일서 개최." Naver News. June 21, 2024.
3) Wells, Pete. "The Top Dishes of 2023 in NYC." The New York Times. December 13, 2023, https://www.nytimes.com/2023/12/13/dining/top-dishes-nyc-2023.html
4) Ministry of Culture, Sports, and Tourism, South Korea. "Annual Number of Travelers from Japan and Total." Accessed [June 20, 2024].

Chart : Annual number of travelers from Japan and total

source: Ministry of Culture, Sports and Tourism, South Korea
Unit: 명(person)

	2015Y	2016Y	2017Y	2018Y	2019Y	2020Y	2021Y	2022Y	2023Y	2024Y	Total
JAPAN	1,742,531	2,213,099	2,223,214	2,864,110	3,179,436	406,701	1,047	259,031	2,258,375	874,659	16,022,203
TOTAL	10,135,489	13,932,925	10,415,594	12,414,348	14,432,275	1,653,471	211,846	1,998,937	8,880,899	3,928,048	78,003,832
Percentage	17.19%	15.88%	21.35%	23.07%	22.03%	24.60%	0.49%	12.96%	25.43%	22.27%	20.54%

<Chart 2: Annual number of travelers from Japan and total>

By targeting and implementing effective marketing strategies towards Japanese tourists, Korea can potentially increase its inbound tourism significantly as the following reason. That is, the travel dynamics between Korea and Japan which could provide insights into creating better enhancement strategies. According to the Japan National Tourism Organization (JNTO), the number of Korean tourists visiting Japan has been considerable than the opposite case, even though the Japanese population is more than twice the size of Korea's.

Interestingly, the number of Korean tourists visiting Japan is significantly higher. For example, in January 2024, the number of Koreans visiting Japan reached 857,000, while only 137,713 Japanese visited Korea[5]. Despite the impact of exchange rates, the fact that the number of Korean tourists traveling to Japan far exceeds the number of Japanese tourists visiting Korea, even though Japan's population is more than twice than that of Korea, indicates potential for increasing Japanese tourism to Korea through targeted analysis and strategic measures. As a result, this discrepancy suggests there are untapped opportunities to significantly boost the number of Japanese visitors to Korea through developing tailored promotional activities using our time-series analysis and forecasting model.

	Oct 23	Nov 23	Dec 23	Jan 24	Feb 24
Korea to Japan	631,120	649,880	782,730	857,000	818,500
Japan to Korea	250,235	272,686	191,892	137,713	177,633

* Each data from JNTO and Ministry of Culture, Sports and Tourism, South Korea

<Chart 3: Monthly Number of Korean Tourists Visiting Japan and Japanese Tourists Visiting Korea>

Ⅱ. Literature Review

Several studies have examined various factors influencing tourism in Korea, most focusing either on forecasting tourist numbers or analyzing factors that could affect these numbers. Molemeng (2023) conducted an in-depth study on the factors influencing the number of visits and behavioral intentions of tourists from China, Japan, and the U.S., based on the KTO 2019 International Visitor Survey.[6] From this paper, the japanese tourists prefer individual travel

5) Japan National Tourism Organization (JNTO) and Ministry of Culture, Sports, and Tourism, South Korea. "Monthly Number of Korean Tourists Visiting Japan and Japanese Tourists Visiting Korea." Accessed June 23, 2024.

6) Molemeng. (2023). The Factors Influencing the Number of Visits and Behavioral Intentions of China, Japan and U.S. Tourists - Based on KTO 2019 International Visitor Survey. Master's thesis, Sejong University, Graduate School, Department of Hotel and Tourism Management.

over group tours, though both segments are significant. Furthermore, a higher proportion of Japanese visitors are women, primarily aged between 21-30 years old and majority of Japanese tourists are visiting Korea for their leisure, recreation, relaxation purposes according to the paper. Meanwhile, Japanese tourists tend to visit Korea four or more times, and their intentions to revisit are significantly influenced by travel immigration procedures, food, and group tours.[7)]

In another significant study, Deukhee Park, Sanghoon Kang, and Gyehee Lee (2020) explored the demand forecasting of Japanese tourists to Korea, aiming for sustainable tourism growth. Their research, titled "Demand Forecasting of Japanese Tourists to Korea for Sustainable Tourism Growth: An Application of Time-Series Econometric Models," applied time-series models such as ARIMA, Holt exponential smoothing model, and Winters additive/multiplicative model to predict tourist demand.[8)] This study was mainly focused on the model selection and forecasting the demand of Japanese tourists rather than the factors affecting the number.

Furthermore, Soeon Park, Gunhee Lee, and Inhye Lee (2015) analyzed the factors influencing Japanese and Chinese tourists visiting Korea, focusing on economic variables and media data. The research found that increased negative media exposure about Japan correlated with a decrease in the number of Japanese tourists visiting Korea, particularly affecting those in 50s age group. Furthermore, while the Korean Wave (Hallyu) did not significantly impact the number of Japanese tourists, it had a substantial effect on increasing the number of Chinese tourists.[9)] This study highlighted that not only macroeconomic factors like exchange rates and GDP but also negative media exposure and diplomatic relations influence the number of tourists from Japan and China visiting Korea. Meanwhile, the authors suggested that future research could achieve more accurate measurements by analyzing information exchanged on social media platforms like Twitter, Facebook, and blogs, in addition to traditional news outlets.

Given that nearly a decade has passed since this study was conducted, this research aims to extend the analysis by incorporating data from Google Trends to examine how often Japanese individuals search for the related key-words selected base on the prior studies. This approach will allow for a more direct and current analysis of the factors influencing Japanese tourists' visits to Korea as well as the forecasting the number of tourists from Japan to Korea. Through this comprehensive approach, we aim to propose specific strategies at the end of this report to significantly increase the number of Japanese travelers, who have a high potential for substantial growth. This increase in Japanese travelers can also stimulate the domestic travel economy. Note that, as the prior studies revealed that the impact of GDP and

---

7) Molemeng. (2023). The Factors Influencing the Number of Visits and Behavioral Intentions of China, Japan and U.S. Tourists - Based on KTO 2019 International Visitor Survey. Master's thesis, Sejong University, Graduate School, Department of Hotel and Tourism Management.

8) Deukhee Park, Sanghoon Kang, and Gyehee Lee. "Demand Forecasting of Japanese Tourists to Korea for Sustainable Tourism Growth: An Application of Time-Series Econometric Models." Journal of Tourism Studies 34, no. 3 (2020): 47-60.

9) Soeon Park, Gunhee Lee, and Inhye Lee. "Analysis of Factors of Visiting Tourists to Korea using Economic Variables and Media Data: A Case Study of Japan and Chinese Tourist." International Journal of Tourism and Hospitality Research 29, no. 7 (2015): 57-72.

exchange rate on the tourist data, our report mainly focused on the key-word data which Japanese people directly searched at Google in Japanese so that we could capture more accurate impact.

Ⅲ. Statement of Research Objectives

The primary objective of this study is to investigate and analyze the factors influencing Japanese tourists' visits to Korea, aiming to develop strategies that can significantly increase the number of these visitors and, consequently, stimulate the domestic travel economy.

This research will utilize regression models to examine the relationship between keyword data, such as search trends from Google Trends, and tourist data. By doing so, we aim to identify the key factors that significantly influence Japanese tourists' decisions to visit Korea.

Additionally, through the application of decomposed models, this research will interpret the trends and seasonality within the Japanese tourist data. Understanding these cyclical patterns and long-term trends is crucial for grasping the dynamics of Japanese tourism to Korea.

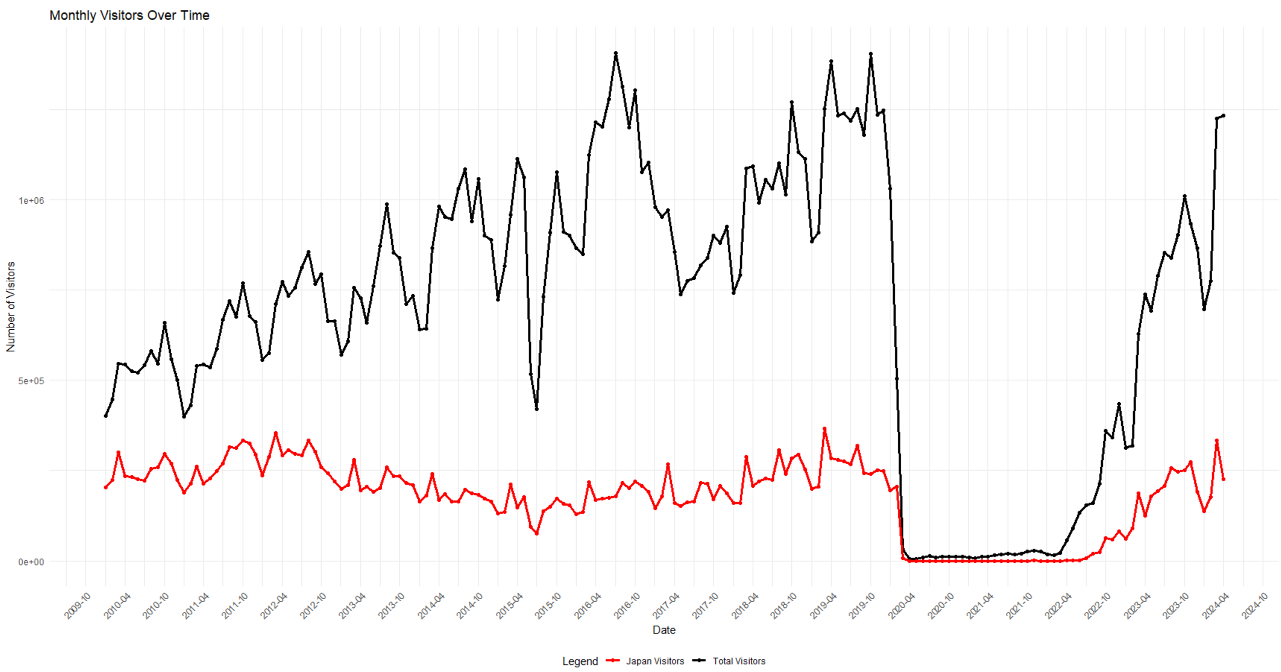
Furthermore, the SARIMAX (Seasonal Auto-regressive Integrated Moving Average with Exogenous Regressors) model will be employed to predict the number of Japanese tourists visiting Korea using the exogenous variables we found through the upper simple regression analysis explained. This model will account for auto-regressive, seasonal variations and external factors, providing robust forecasts as well as the meaningful implications to build our appropriate strategies to boost the total number of Japanese tours to Korea.

Finally, the objective is to properly interpret the analysis results and develop specific strategies to attract more Japanese tourists. These strategies will be informed by the insights gained from the regression models, trend and seasonality analysis, and SARIMAX forecasts. Through this comprehensive approach, we aim to propose targeted strategies that can leverage the potential for substantial growth in Japanese tourism to Korea, thereby contributing to the stimulation of the domestic travel economy.

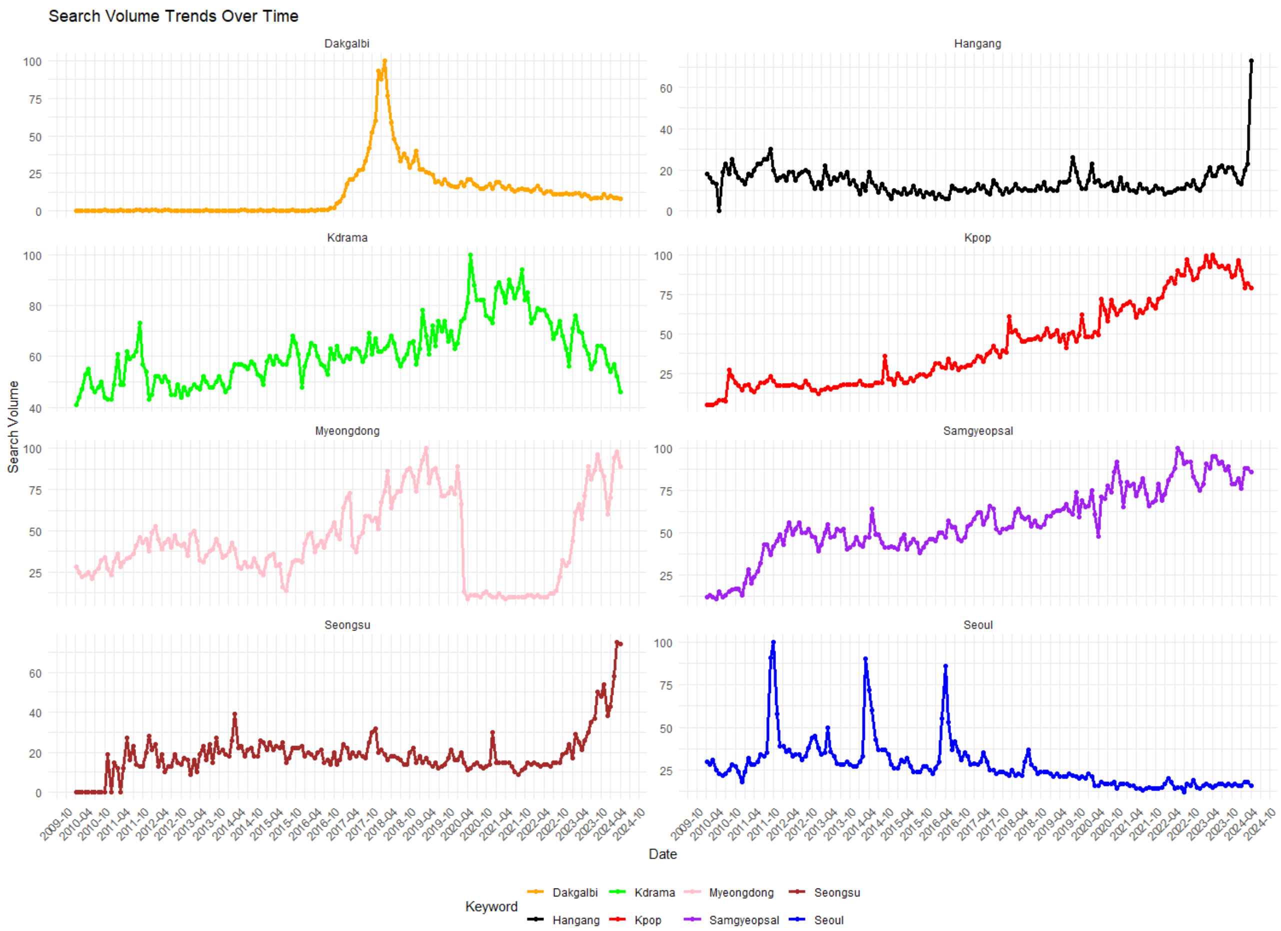
Ⅳ. Description of Data & Applied Methodologies

In this section, we will explain the data used in this report, including its visualization and basic statistical analysis results. First, we visualized the time series data, also the target of estimation, which consists of the monthly number of Japanese tourists visiting Korea. And for reference purposes, the monthly number of all foreign tourists visiting Korea are plotted as well. All time-series related data are between Jan 2010 and Apr 2024. The split point of the train data and test data are explained in detail at the next part of this paper.

On the below figure 1 (shown in page 6), it clearly seems that the trend of the number of Japanese tourists visiting Korea differs from the trend of the total number of foreign tourists visiting Korea. This discrepancy may be due to the fact that the total number of visitors includes not only the 20% contributed by Japanese tourists but also the remaining 80% from other countries, whose influences might outweigh that of the Japanese tourists. Therefore, since the primary goal of this report is to increase the number of Japanese visitors, the time series data analysis was conducted using the data specifically for Japanese tourists.



<Figure 1: Time-series representation of the tourists data>



<Figure 2: Time-series representation of the exogenous data>

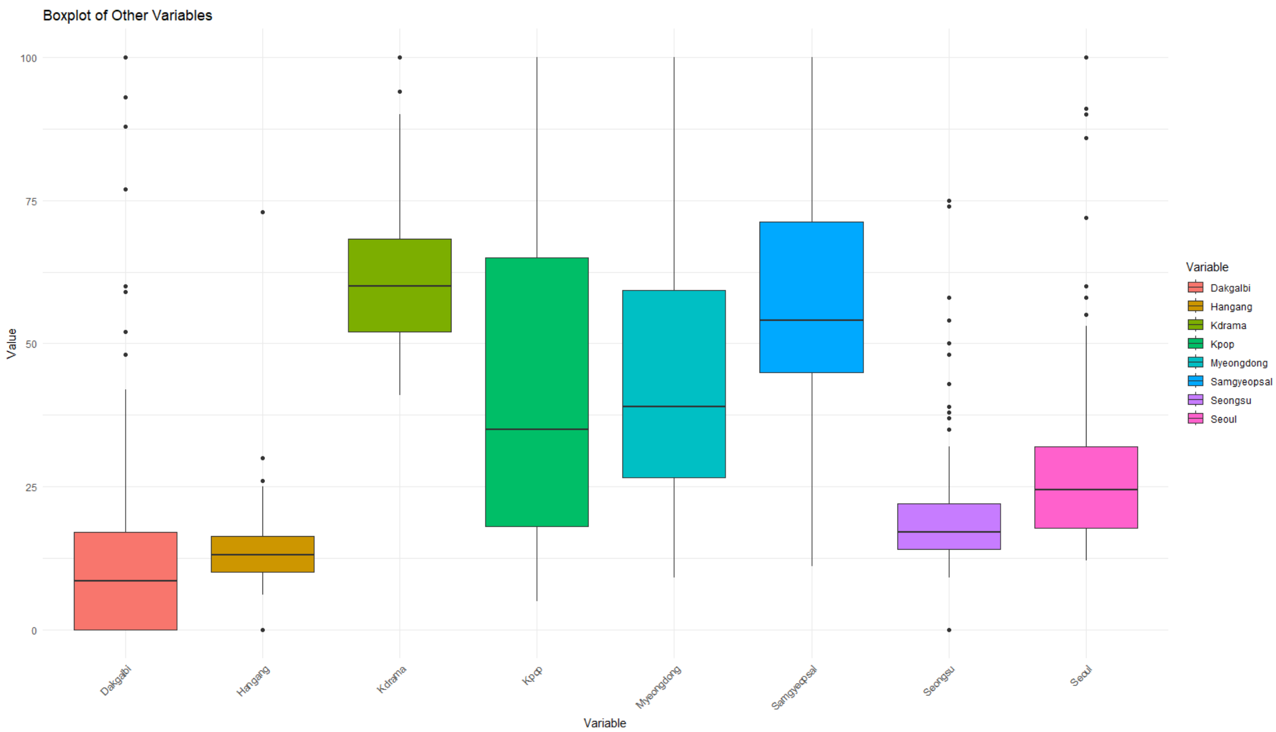
The following section provides an explanation of the related key-words selected based on prior researches in the field. First, these variables were plotted as time series, as shown in Figure 2(shown in page 6). Those variables shown in graphs are first selected with prior studies that revealed significant travel factors that influence the Japanese travel visit to Korea. Then, the specific key-words are selected regarding the above several factors using the domain knowledge.

It is important to note that the scale of each search count differs. Google does not disclose the exact search volumes; therefore, each keyword is displayed on a scale from 0 to 100 based on its own magnitude. Unlike the travel data, Google Trends search data continued to be collected online even during the COVID-19 pandemic, making this data significant. However, this project was performed both with the original data as given and the version with adjustments to account for the travel ban during the COVID-19 period, based on the common sense understanding that travel was forcibly restricted during this time while search data were not.

Next, basic statistic table is given for the each variables used in this paper on the below table 1. The mean, median, max, min, standard deviation values are given for the Japanese visitors to Korea and other key-word data.

Metric	Japan_Visitors	Kpop	Seoul	Kdrama	Samgyeopsal	Dakgalbi	Seongsu	Myeongdong	Hangang
Mean	176772.40	42.28488	27.48837	61.66279	56.59884	11.88372	18.73837	43.00581	13.837209
Median	198194.00	35.00000	24.50000	60.00000	54.00000	8.50000	17.00000	39.00000	13.000000
Max	367157.00	100.00000	100.00000	100.00000	100.00000	100.00000	75.00000	100.00000	73.000000
Min	16.00	5.00000	12.00000	41.00000	11.00000	0.00000	0.00000	9.00000	0.000000
SD	99404.49	26.94667	13.96382	12.06232	20.53571	17.32417	10.95104	24.74371	6.572778

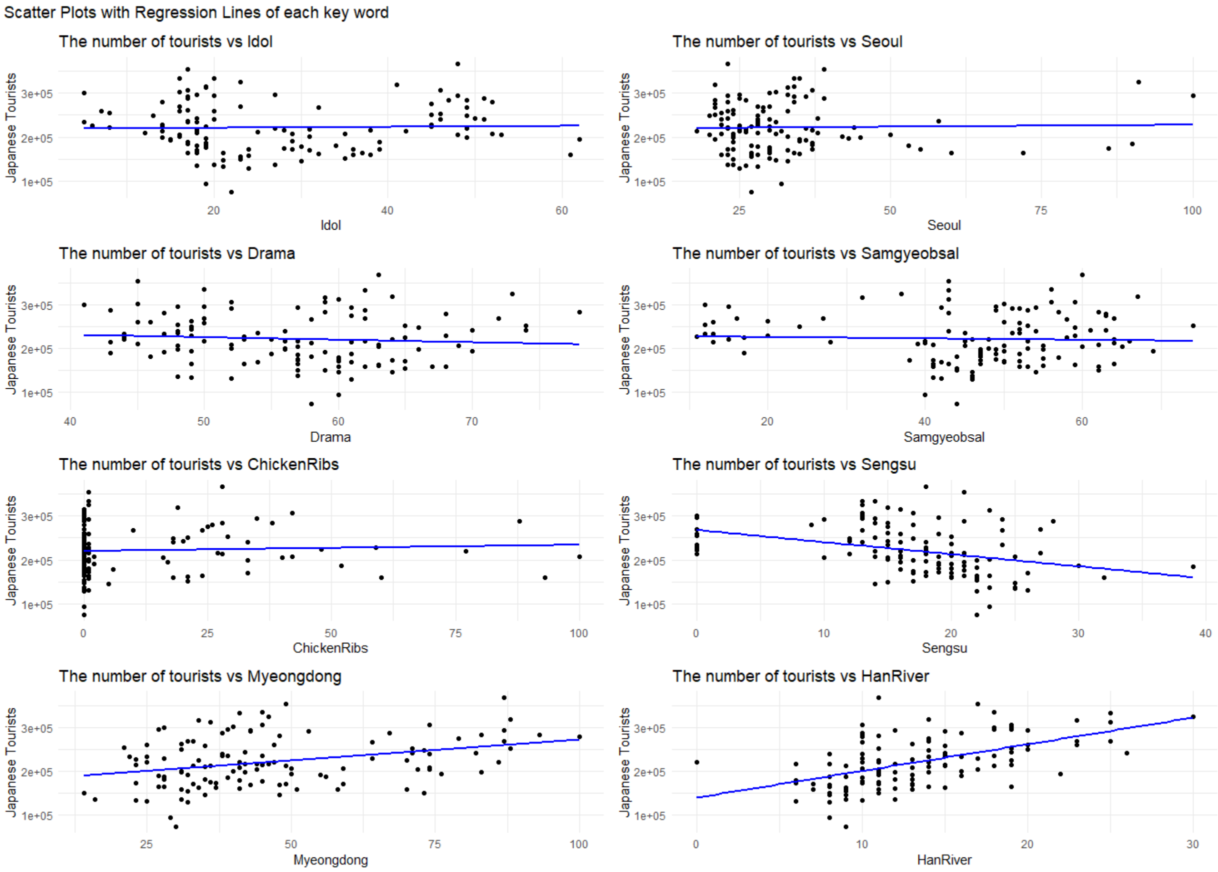
<Table 1: Summary statistics on each attributes>



<Figure 3: Box-plot representation of the key-word data>

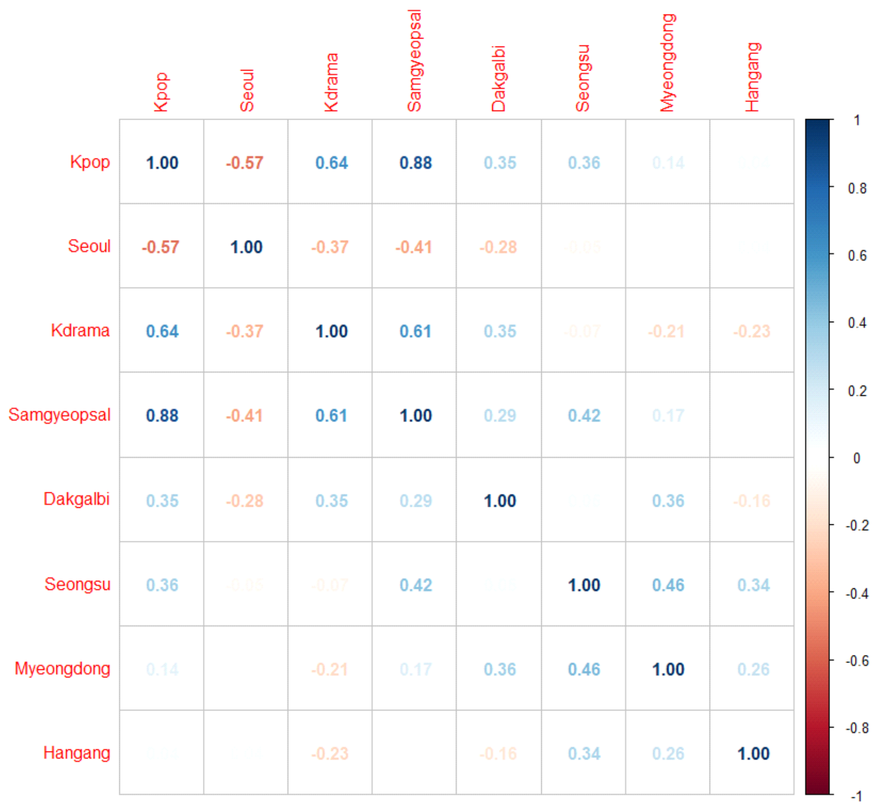
In the remaining part of this section, we performed a preliminary examination to determine whether the selected variables are suitable for use in this report. This examination is included because it intuitively represents the characteristics of the data used. We conducted two types of analyses: one considering the relationship between the variables and Japanese tourist data, and another analyzing the correlations among these exogenous variables using a heat-map visualization.

For each key-word data and Japanese tourist data, we matched them 1 to 1 without considering the time lag and then displayed them as scatter plots with simple regression lines. While the influence between the two could change with the application of time lags, the aim was to preemptively exclude any unrelated keyword data from the model training. This step is crucial to significantly reduce the time required to create final SARIMAX model that determines which variables to include and the optimal lag for each variables.



<Figure 4: Box-plot representation of the key-word data>

As shown in Figure 4, the scatter plot indicates that for the keyword "dakgalbi," the number of Japanese tourists to Korea varies widely even during periods when the search data value is 0. Therefore, it was deemed inappropriate to use this data, and it was excluded from the analysis. Additionally, the correlation analysis among the exogenous variables we conducted is shown in the figure below figure 5(shown in next page).



<Figure 5: Hit-map representation of key-word data>

As can be seen in the heat-map, the two keyword variables "Samgyeopsal" and "Kpop" have a very high correlation of 0.88. In this report, we considered that multi-colinearity issues would arise if the correlation is much higher than 0.6. Therefore, one of the variables was excluded from the full model. In this case, "Kpop" was excluded. However, deleting the “Kpop” is not meant that number of K-pop searches is not related to the dependent variable in this report. If the “Samgyeopsal” key-word data have a high impact on the dependent variable, then we would consider that the “Kpop” key-word data also have a high impact.

In the next section, we aim to include the monthly data values of the analyzed key-words as exogenous variables in the time series data of the number of Japanese tourists visiting Korea. We will train the overall model and evaluate its performance. The full model includes all six keyword variables used, with each having possible lags ranging from 0 (no lag) to 4. This range was chosen based on previous studies, suggesting that searches conducted more than four months prior are not significantly relevant to trip planning.

Specifically, by allowing each variable to have a lag from 0 to 4, we set up a loop to generate a total of **5^6** models. For each model, we calculated the RMSE during the test period and ultimately selected the model with the lowest RMSE as the overall model. It is important to note that, the recent sharp recovery trend in tourist data necessitates the inclusion of as much training data as possible. Therefore, we used data from January 2010 to December 2023 for training and data from January 2024 to April 2024 for testing. After generating the overall model, we created simplified models for variables that had a strong impact on the dependent variable to conduct a more detailed analysis.

Finally, we used the model with the lowest RMSE to predict future tourist numbers. By identifying and targeting the key-words and lags that significantly influenced the number of tourists, we aimed to propose targeted policies. These policies could help improve the situation where there is substantial potential to increase the number of Japanese tourists visiting Korea, but progress has been stagnant.

V. Analysis result

In this part, the analysis results of time-series models and additional tests were executed regarding the models selected based on the evidence explained before. The lag simulation results of the SARIMAX full-model, which includes all the variables explained in the previous section, are shown below. Using the R program, we set up a loop to print the model whenever a lower RMSE was found. To ensure the program was functioning correctly, we also set it to display a check every 5000 models.

```
[1] "Lags - Seoul: 0 Kdrama: 0 Samgyeopsal: 0 Seongsu: 0 Myeongdong: 0 Hangang: 0 RMSE: 79076.925640823"
[1] "Lags - Seoul: 0 Kdrama: 0 Samgyeopsal: 0 Seongsu: 0 Myeongdong: 0 Hangang: 1 RMSE: 60693.9451000964"
[1] "Lags - Seoul: 0 Kdrama: 0 Samgyeopsal: 0 Seongsu: 0 Myeongdong: 1 Hangang: 1 RMSE: 47913.7061307375"
[1] "Lags - Seoul: 0 Kdrama: 0 Samgyeopsal: 0 Seongsu: 1 Myeongdong: 1 Hangang: 1 RMSE: 47846.0603930127"
[1] "Lags - Seoul: 0 Kdrama: 0 Samgyeopsal: 0 Seongsu: 2 Myeongdong: 1 Hangang: 1 RMSE: 44727.6744806252"
[1] "Lags - Seoul: 0 Kdrama: 1 Samgyeopsal: 0 Seongsu: 1 Myeongdong: 0 Hangang: 3 RMSE: 42496.621128885"
[1] "Lags - Seoul: 0 Kdrama: 1 Samgyeopsal: 3 Seongsu: 3 Myeongdong: 0 Hangang: 4 RMSE: 40561.4724101601"
[1] "Lags - Seoul: 0 Kdrama: 1 Samgyeopsal: 3 Seongsu: 4 Myeongdong: 0 Hangang: 4 RMSE: 40484.9237790434"
[1] "Lags - Seoul: 0 Kdrama: 2 Samgyeopsal: 0 Seongsu: 2 Myeongdong: 0 Hangang: 4 RMSE: 38473.492382454"
[1] "5000!"
[1] "Lags - Seoul: 2 Kdrama: 1 Samgyeopsal: 3 Seongsu: 0 Myeongdong: 0 Hangang: 1 RMSE: 36634.8516281763"
[1] "Lags - Seoul: 2 Kdrama: 1 Samgyeopsal: 3 Seongsu: 0 Myeongdong: 0 Hangang: 4 RMSE: 36036.9926211054"
[1] "Lags - Seoul: 2 Kdrama: 1 Samgyeopsal: 3 Seongsu: 1 Myeongdong: 0 Hangang: 1 RMSE: 34595.0947397314"
[1] "Lags - Seoul: 2 Kdrama: 1 Samgyeopsal: 3 Seongsu: 1 Myeongdong: 0 Hangang: 4 RMSE: 34217.9238013849"
[1] "Lags - Seoul: 2 Kdrama: 4 Samgyeopsal: 3 Seongsu: 4 Myeongdong: 0 Hangang: 4 RMSE: 34038.6386693189"
[1] "5000!"
[1] "5000!"
> print(best_lags)

> print(best_lags)
$Seoul
[1] 2

$Kdrama
[1] 4

$Samgyeopsal
[1] 3

$Seongsu
[1] 4

$Myeongdong
[1] 0

$Hangang
[1] 4

> print(paste("Best RMSE:", best_rmse))
[1] "Best RMSE: 34038.6386693189"

> forecast(fit, xreg = as.matrix(exog_test), h = nrow(test_y))
    Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
169       153366.1 117022.4 189709.7  97783.28 208948.8
170       209607.8 170005.7 249209.9 149041.59 270173.9
171       278931.1 238290.6 319571.6 216776.86 341085.4
172       243072.9 202285.8 283860.0 180694.40 305451.4
```

<Table 2: Code result of full model>

As shown in the results above, the optimal lags for each variable are 2, 4, 3, 4, 0, and 4. The RMSE for this model was found to be 34,038. Given the current situation in 2024, where the number of Japanese tourists visiting Korea is rapidly increasing after the impact of COVID-19, the test data was set for a period of four months. This shorter test period and fast recovery situation are likely contributed to the relatively high RMSE result. Indeed, the prediction results attached in the table above show a wide 95% confidence interval. The actual predicted values were 153,366 in January, 209,607 in February, 278,931 in March, and 243,072 in April, while the actual values were 137,713, 177,633, 334,131, and 225,182, respectively.

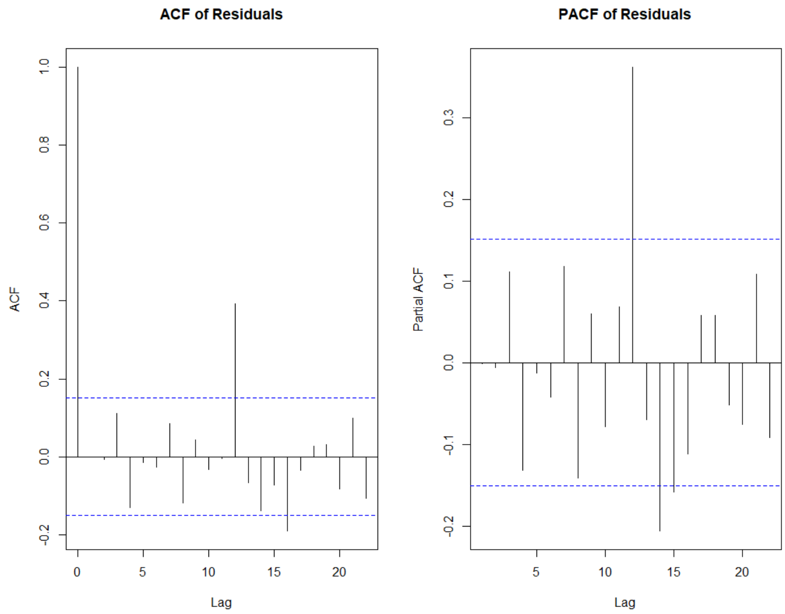
However, the most important objective of this report is not merely to predict future data values but to analyze which keyword factors have more direct influence toward these results. For this purpose, we previously visualized and presented the data using simple regression analysis, and in this section, we aim to interpret the model's summary data to further understand these influences.

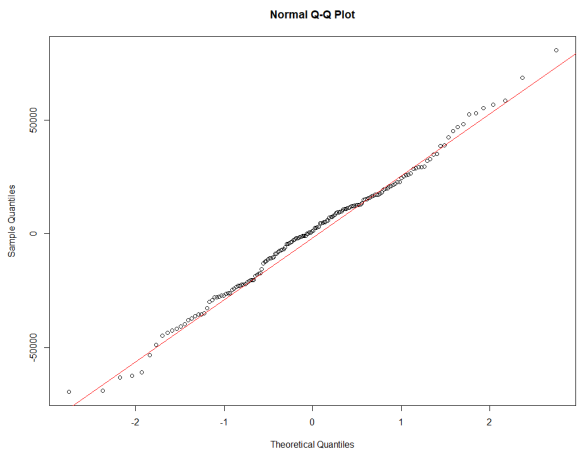
```
> summary(fit)
Series: as.numeric(train_y)
Regression with ARIMA(3,1,3) errors

Coefficients:
          ar1      ar2      ar3      ma1    ma2      ma3       drift  Seoul_lag2  Kdrama_lag4  Samgyeopsal_lag3
      -0.1677  -0.8811  -0.0571  -0.3993  0.605  -0.8200  -2255.4608    113.2918     110.6194         1779.4424
s.e.   0.1050   0.0471   0.0969   0.0779  0.042   0.0917    436.4001    191.7720     399.0480          320.2614
      Seongsu_lag4  Myeongdong_lag0  Hangang_lag4
          650.0689        3279.8770    -1294.0085
s.e.      406.1910         185.5412     495.8653

sigma^2 = 803663858:  log likelihood = -1945.15
AIC=3918.31   AICc=3921.07   BIC=3961.96

Training set error measures:
                   ME     RMSE      MAE       MPE     MAPE      MASE          ACF1
Training set 163.2184 27142.07 21182.66 -461.2268 4297.359 0.7659747 -0.0009930084
> fit
Series: as.numeric(train_y)
Regression with ARIMA(3,1,3) errors

Coefficients:
          ar1      ar2      ar3      ma1    ma2      ma3       drift  Seoul_lag2  Kdrama_lag4  Samgyeopsal_lag3
      -0.1677  -0.8811  -0.0571  -0.3993  0.605  -0.8200  -2255.4608    113.2918     110.6194         1779.4424
s.e.   0.1050   0.0471   0.0969   0.0779  0.042   0.0917    436.4001    191.7720     399.0480          320.2614
      Seongsu_lag4  Myeongdong_lag0  Hangang_lag4
          650.0689        3279.8770    -1294.0085
s.e.      406.1910         185.5412     495.8653

sigma^2 = 803663858:  log likelihood = -1945.15
AIC=3918.31   AICc=3921.07   BIC=3961.96
```

<Table 3: Full-model summary results>

First, the model was created by incorporating exogenous variables (Seoul_lag2, Kdrama_lag4, Samgyeopsal_lag3, Seongsu_lag4, Myeongdong_lag0, Hangang_lag4) into the existing ARIMA model. According to the summary data of the model above, the effect of Seoul search volume after 2 months is 113, the effect of Korean drama search volume after 4 months is 110, the effect of samgyeopsal search volume after 3 months is 1779, the effect of Seongsu search volume after 4 months is 650, the current effect of Myeongdong search volume is 3279, and the effect of Hangang search volume after 4 months is -1294.

Therefore, it is noteworthy that the effect of samgyeopsal search volume after 3 months is significantly high at 1779, and the current effect of Myeongdong search volume is significantly high at 3280. This implies that searches for samgyeopsal before a trip significantly influence the number of tourists in the subsequent period, whereas searches for Myeongdong during the current month influence the number of tourists in that same month. This suggests that tourists may search for Myeongdong just before their trip or while they are already traveling. The influence of other key-words was considered to be less significant.






```
> print(shapiro_test)

        Shapiro-Wilk normality test

data:  residuals
W = 0.99225, p-value = 0.5058
```

```
> print(ljung_box_test)

        Box-Ljung test

data:  residuals
X-squared = 51.79, df = 20, p-value = 0.0001222
```

<Table 4: Full-model test results>

Next, we conducted tests on the overall model and analyzed the results. We performed residual analysis to check if the residuals followed a normal distribution using a Q-Q plot and analyzed the autocorrelation of the residuals using the Ljung-Box test. The results of these tests are attached below.

From the Q-Q plot and the Shapiro test, we confirmed that the residuals followed a normal distribution. However, based on the attached ACF and PACF results and the Ljung-Box test, we determined that the residuals exhibit autocorrelation, suggesting that the model's RMSE could be further reduced. In particular, we observed that the residuals had an autocorrelation lag of around 12 in the graph. We found that seasonality was not accounted for when using auto.arima. To address this issue, we manually fitted a SARIMAX model and conducted the same analysis below.

```
> fit_seasonal <- Arima(train_y, order = c(3, 1, 3), seasonal = list(order = c(1, 1, 3), period = 12), xreg =
as.matrix(exog_train_ts))
> summary(fit_seasonal)
Series: train_y
Regression with ARIMA(3,1,3)(1,1,3)[12] errors

Coefficients:
          ar1      ar2      ar3      ma1     ma2      ma3      sar1     sma1     sma2     sma3      Seoul
Kdrama  Samgyeopsal  Seongsu
      -0.3776  -0.9226  -0.1476  -0.1705  0.6588  -0.4847  -0.1698  -0.5424  -0.2158  -0.1032    53.1783
-705.3013    952.8058  249.3340
s.e.   0.1717   0.0696   0.1587   0.1627  0.0857   0.1538   0.5717   0.5929   0.4123   0.1397  163.6108   338.6328
   305.3494  329.6895
      Myeongdong    Hangang
       3233.2863  -402.9109
s.e.    183.8627   487.3961

sigma^2 = 458717199:  log likelihood = -1765.09
AIC=3564.18   AICc=3568.65   BIC=3615.92

Training set error measures:
                   ME     RMSE      MAE      MPE     MAPE      MASE          ACF1
Training set 1152.745 19481.63 14320.72 2046.522 6410.242 0.5178436 -0.0007753541
> forecast_values <- forecast(fit_seasonal, xreg = as.matrix(exog_test_ts), h = length(test_y_ts))
> predicted_values <- as.numeric(forecast_values$mean)
> rmse_value_seasonal <- rmse(as.numeric(test_y_ts), predicted_values)
> print(paste("Test Data RMSE:", rmse_value_seasonal))
[1] "Test Data RMSE: 21798.8658018887"
```

```
> forecast_values$mean
Time Series:
Start = 169
End = 172
Frequency = 1
[1] 140679.8 205918.3 306263.6 242939.0
```

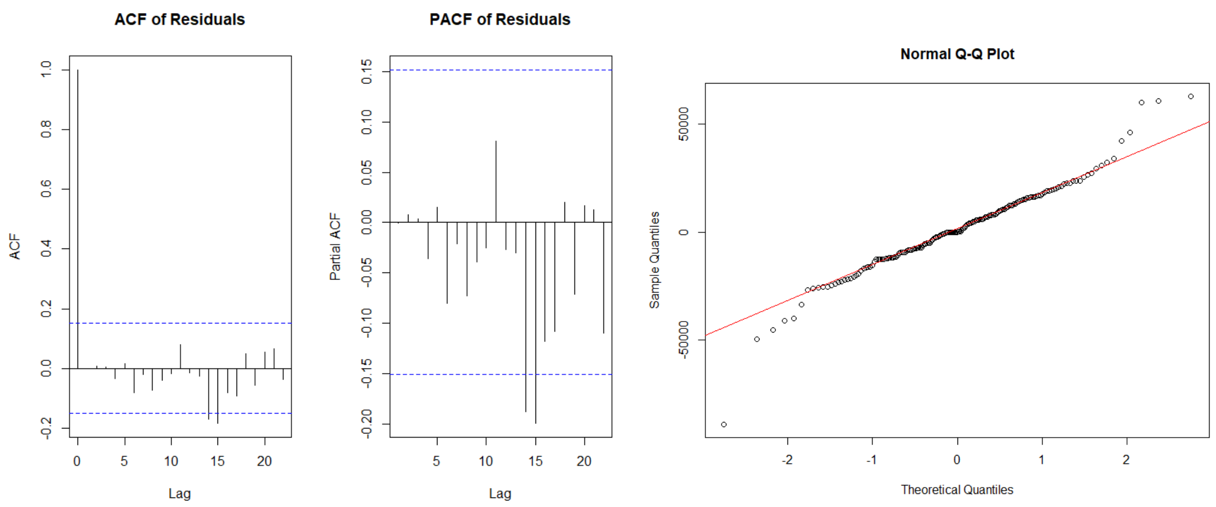
<Table 5: Revised Full-model summary results>

The model generated based on these results is shown in the table above, and it has an RMSE of 21,798. The predicted values for January to April 2024 are 140,680, 205,918, 306,264, and 242,939, respectively, indicating much more improved accuracy.

The regression coefficients from the model provide insights into how different exogenous variables influence the number of Japanese tourists visiting Korea. An increase of one unit in the search volume(% scaled data) for "Seoul" corresponds to an increase of approximately 53.18 Japanese tourists, indicating a positive impact of interest in Seoul on tourist numbers. Conversely, an increase in the search volume for "Kdrama" corresponds to a decrease of approximately 705.30 Japanese tourists. This unexpected result might be due to interactions with other variables or specific conditions in the data. However, through the regression analysis part and graph visualization part, we thought that searching K-drama does not lead to travel to Korea directly.

The search volume for "Samgyeopsal" shows a strong positive influence, with a one-unit increase corresponding to approximately 952.81 additional tourists after the lag 3. Similarly, the search volume for "Seongsu" and "Myeongdong" corresponds to increases of approximately 249.33 and 3233.29 tourists, respectively, indicating significant positive impacts. However, an increase in the search volume for "Hangang" corresponds to a decrease of approximately 402.91 tourists, another unexpected result that may require further analysis. In summary, while positive regression coefficients indicate that increased interest in certain variables (e.g., Samgyeopsal, Seoul, Seongsu, Myeongdong) leads to an increase in tourist numbers, negative coefficients (e.g., Kdrama, Hangang) suggest a decrease, highlighting the complex interactions and the need for further review to understand these dynamics fully. Nonetheless, above model has much more improved prediction RMSE results than the both prior full model and the each model using just one key-word data. Therefore, our team used this revised full model to predict the future number of Japanese tourists visiting Korea.

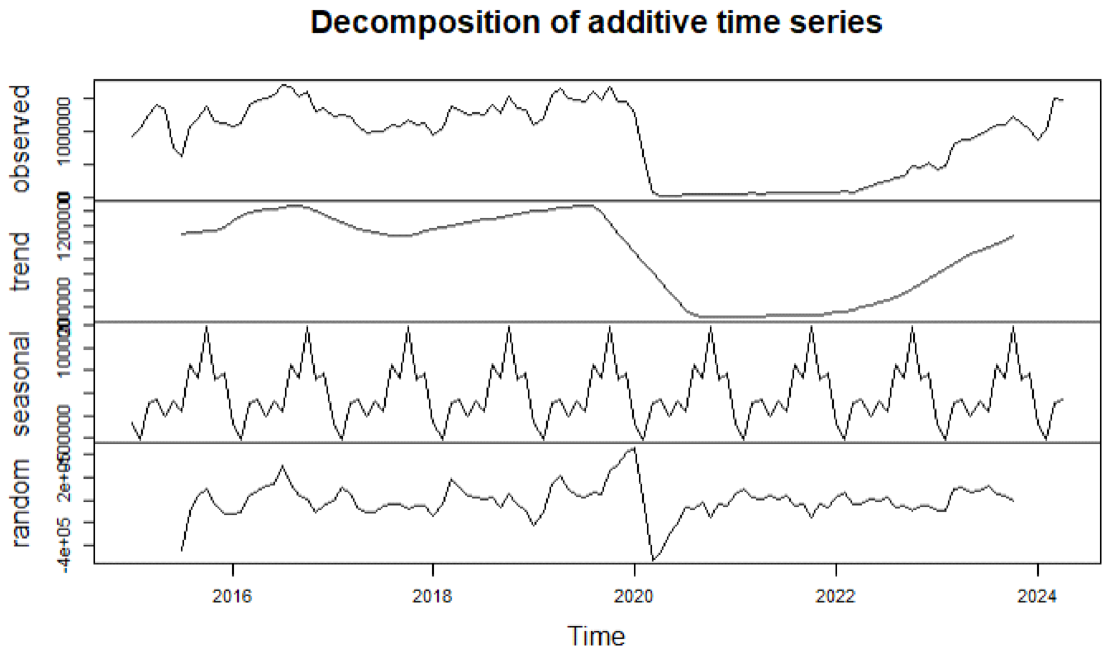
Residual analysis was conducted to check if the residuals followed a normal distribution using a Q-Q plot, and the Ljung-Box test was used to analyze the autocorrelation of the residuals. The results of these tests are attached below(shown in next page). From the Q-Q plot, it was seen that the residuals followed a normal distribution although there was some outliers, and the autocorrelation issue of the residuals was also resolved.




```
> print(ljung_box_test)

        Box-Ljung test

data:  residuals
X-squared = 18.509, df = 17, p-value = 0.3574
```

<Table 6: Revised Full-model test results>

Next, decomposed model of the dependent variable, monthly number of Japanese tourist to Korea, is analyzed to understand the seasonality and the trend.



<Figure 6: Decomposed model of dependent variable>

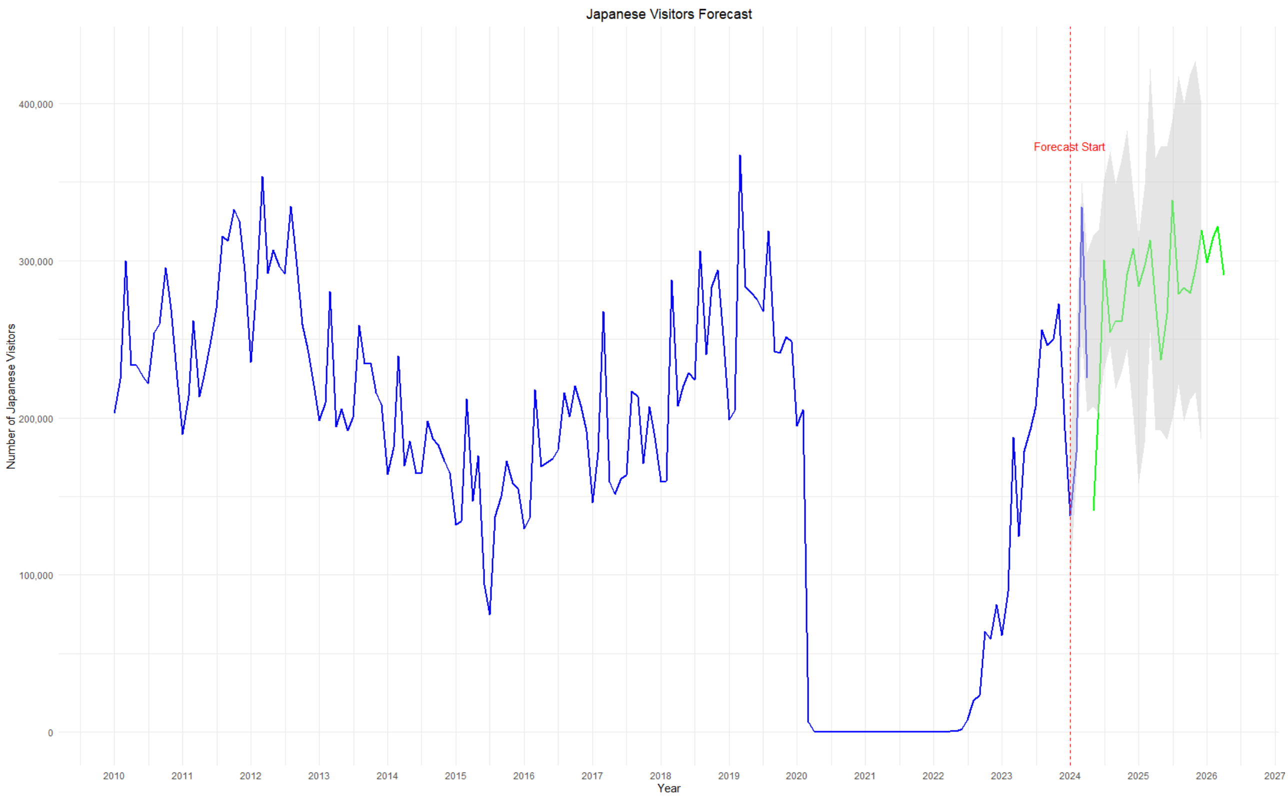
Through the decomposition model, it was seen that there was no clear trend of increase or decrease in the data using Jan 2010 to Apr 2024. According to previous studies, this is often attributed to significant decreases due to pandemics such as COVID-19 or MERS, or specific political events, resulting in no clear trend.

On the other hand, seasonality is more evident. In particular, there is a sharp decrease in visitors at the beginning of the year, followed by a sharp increase in the fall after summer. Generally, there are more monthly visitors around the relatively cool months of March and October, while there is a significant decrease in January and February, the start of the year and winter.

Throughout the analysis given in this part, we are going to show the prediction results of Japanese tourists data(in the next page) and give some strategic suggestion regarding the way to increase the Japanese visitors at the next part. The prediction result of the Japanese visitors during May 2024 to Dec 2025 are shown below chart and the graph is shown at the next page of this report.

Y-M	24-05	24-06	24-07	24-08	24-09	24-10	24-11	24-12	25-01	25-02
unit	261743	261166	291340	307600	283644	296923	313303	272992	236685	266208
Y-M	25-03	25-04	25-05	25-06	25-07	25-08	25-09	25-10	25-11	25-12
unit	338451	278730	282611	279437	294538	319504	298741	314916	321939	290615

Note that regarding the RMSE result of the model that are built only with one key-word data, if we use each variables with no lag, there was a high RMSE results over 50,000.



<Figure 6: Forecasted values graph representation>

VI. Expected Original Contribution

1. Analyzing the factors that make Japanese visit Korea using keyword data and tourist data

Through this report, as discussed in the previous analysis section, we understood that Japanese keyword searches for samgyeopsal, Myeongdong, and k-idols (which show a high correlation with samgyeopsal) influence the number of Japanese tourists. We created a suitable model to interpret each factor through optimized lags using trial and error method. we tried every 5^6 models and found the optimal one.

2. Interpreting the trend and seasonality of Japanese tourist data using a decomposed model

Although there was no overall trend in Japanese tourist data, we identified a clear seasonality where the number of visitors sharply decreases in January, the beginning of the new year and winter, and significantly increases in March and October, which have pleasant weather conditions.

3. Predicting the number of Japanese tourists in Korea using the revised full model

As mentioned in the previous section, we created a final model by including keyword data as exogenous variables in a SARIMAX model and determining their optimal lags through iterations. We identified and corrected issues related to seasonality through model validation, resulting in a final model that recorded an RMSE of 21,798 on the test data. We then used this model to forecast data for 2024 and 2025 and visualized the predictions.

This analysis is significant in integrating factor studies and predictions, which were previously conducted separately. It attempts to provide more accurate predictions and interpretations based on direct search data from Japanese people, addressing the limitations of media data used in prior research.

4. Interpreting the meaning of analysis results and finding strategies accordingly

Based on this analysis, the remaining part of the report aims to propose strategic policies and implications to increase the number of Japanese tourists to boost travel economy in Korea.

First, I would like to summarize the descriptive statistics we analyzed. The trend of Japanese tourists visiting Korea has experienced notable fluctuations over recent years. Specifically, there was a decline in the number of Japanese visitors to Korea during the late 2010s, which further plummeted around the onset of the COVID-19 pandemic. This decrease can be attributed to various factors. However, post-pandemic recovery efforts have shown promising results, with the number of Japanese tourists returning to Korea now nearing its pre-pandemic peak. This resurgence indicates a renewed interest and confidence in travel between the two nations.

We also examined several time series forecasting models like SARIMA and ARIMAX, and these models are crucial for tourism practitioners and policymakers, as they provide a framework to anticipate trends and prepare accordingly. By evaluating the accuracy and reliability of these models, we can refine strategies to enhance tourist experiences and optimize resources. This approach allows for the development of targeted marketing campaigns, improved infrastructure, and tailored services that cater to the preferences and needs of Japanese tourists.

In our study, we utilized Google Trends to analyze the factors influencing Japanese tourists' decisions to visit Korea. This tool provided valuable insights into the search behaviors and interests of potential travelers. Various factors were considered, such as cultural events, economic conditions, bilateral relations, and social trends. The analysis revealed that certain factors had a more significant impact on travel decisions than others.

Interestingly, our research highlighted that the search volume of Korean dramas in Japan, while popular, exhibited a weaker relationship with the actual visits of Japanese tourists to Korea compared to other factors. This finding suggests that while Korean pop culture, including dramas, plays a role in generating interest and soft power, it may not directly translate to travel decisions. Other elements, such as foods like samgyeopsal or places like myeong-dong or influencer like k-idol, appear to have a more substantial influence on the choice to visit Korea.

So, we proposed three regarding policy measures each of which regarding to the place, food, and influencer of Korea. As shown in our analysis, the increase in keyword searches related to Myeongdong and samgyeopsal has a significant impact on the number of Japanese tourists. Specifically, for Myeongdong, it was found that tourists tend to search for this specific area in advance of their visit, with a very significant impact observed within a one-month lag. In contrast, the effect for samgyeopsal was observed with a three-month lag. Based on these results, for Myeongdong, it is clear that tourists do not plan their visits far in advance. Therefore, continuously promoting this area through events could be effective. For instance, marketing campaigns targeting Myeongdong should not be executed three to four months in advance, as such long lead times do not significantly influence travel decisions. Given the short lag, we suggest small-scale, continuous promotions.

Next, for samgyeopsal, the analysis shows a significant impact with a relatively long lag. Typically, tourists make specific plans for their trips closer to the departure date, but they consider what to eat much earlier. This is further supported by the high correlation between samgyeopsal and k-idol searches. The longer lag can be easily understood when considering that planning to attend an idol concert or fan meeting might be reflected in the search data well in advance of the trip for about 3 months.

Therefore, we propose conducting promotional events centered around well-established franchise restaurants or reputable individual eateries that properly represent Korean samgyeopsal culture, focusing on tourist hotspots in Seoul popular among Japanese visitors. Additionally, such promotions can be integrated into group tour packages, as suggested by previous research.

Furthermore, considering idol concerts, combining them with trendy Korean festivals such as Water Bomb (usually held in summer) could attract Japanese tourists. Based on research on Japanese preferences for specific idols, hosting these festivals during the summer, when tourist numbers typically decline, and starting promotion three months in advance could help mitigate the low tourist season.

In the following section, we have detailed the limitations of this report and outlined areas for future research.

Ⅶ. Limitation of this Paper

In this report, we selected the model by considering all variables deemed significant. However, even though the correlations between these variables were low, they could still influence each other to some extent. Therefore, additional follow-up studies are necessary to understand the direct impact of each variable more accurately as the precise interpretation of each variable could determine the success of policies.

Furthermore, since Google Trends data are not exact search volumes but are recorded on a 0-100 scale, the analysis conducted in this report may require entirely new data collection over time as the 0-100 scale updates. Additionally, it is necessary to investigate the search engines primarily used by Japanese people. In this report, we used Google Trends data based on the Japanese region for convenience of data collecting purposes, but data from search services predominantly used by people in each country, such as South Korea's Naver, could lead to more accurate analyses.

Lastly, while this report uses time series data models, various advanced AI models such as LSTM and GRU and much more recently are being used to predict time series data. When the focus is on manufacturing and production decisions through accurate demand forecasting rather than interpretation, using models that can predict precise results would be more appropriate.

Ⅷ. References

1. Ministry of Culture, Sports, and Tourism, South Korea. "Number of Domestic Travel and Increasing Rate." Accessed [JUNE 20, 2024].

2. 김보경. "한류 알려온 '케이콘' 12년 역사 처음으로 9월 독일서 개최." Naver News. June 21, 2024.

3. Wells, Pete. "The Top Dishes of 2023 in NYC." The New York Times. December 13, 2023, https://www.nytimes.com/2023/12/13/dining/top-dishes-nyc-2023.html

4. Ministry of Culture, Sports, and Tourism, South Korea. "Annual Number of Travelers from Japan and Total." Accessed [June 20, 2024].

5. Japan National Tourism Organization (JNTO) and Ministry of Culture, Sports, and Tourism, South Korea. "Monthly Number of Korean Tourists Visiting Japan and Japanese Tourists Visiting Korea." Accessed June 23, 2024.

6. Molemeng. (2023). The Factors Influencing the Number of Visits and Behavioral Intentions of China, Japan and U.S. Tourists - Based on KTO 2019 International Visitor Survey. Master's thesis, Sejong University, Graduate School, Department of Hotel and Tourism Management.

7. Deukhee Park, Sanghoon Kang, and Gyehee Lee. "Demand Forecasting of Japanese Tourists to Korea for Sustainable Tourism Growth: An Application of Time-Series Econometric Models." Journal of Tourism Studies 34, no. 3 (2020): 47-60.

8. Soeon Park, Gunhee Lee, and Inhye Lee. "Analysis of Factors of Visiting Tourists to Korea using Economic Variables and Media Data: A Case Study of Japan and Chinese Tourist." International Journal of Tourism and Hospitality Research 29, no. 7 (2015): 57-72.