WAYNE STATE UNIVERSITY
COLLEGE OF ENGINEERING

COMPUTER SCIENCE DEPARTMENT

Course Project

CSC5800: Intelligent Systems

Analyzing the Effect of COVID19 on South Korean Tourism by using Data Mining

Supervised by

Dr. Suzan Arslantürk

Submitted by

Sujata Gorai
Access Id – hc6837
&
Alokparna Bandyopadhyay
Access Id - gq6225

ABSTRACT

This course projects includes analysis on Tourism prediction for South Korea which revolves around 3 time series datasets from Kaggle. The major goal of this course project is to do an exploratory data analysis on the entire dataset and use time series forecasting methods to estimate how the Covid-19 pandemic might have affected the South Korean tourism industry.

1. INTRODUCTION: OVERVIEW

The tourism industry contributes to the economic growth of a country. Tourists contribute to sales, profits, jobs, tax revenues, and income in an area. Through secondary effects, tourism affects most sectors of the economy. Economic impact analysis of tourism activity focuses on changes in sales, income, and employment in a region resulting from tourism activity.

In 2019, South Korea saw a record number of around 17.5 million visitors, making it the 20th most visited country in the world, and the 6th most visited in Asia. Most non-Korean tourists come from other parts of East Asia such as Japan, China, Taiwan, and Hong Kong. The recent popularity of Korean popular culture, often known as the "Korean Wave", in these countries has increased tourist arrivals.

We have taken a recent dataset from Kaggle based on the South Korean tourism records from January 2019 to April 2019. Initially we tried to do an exploratory data analysis on the dataset. Then we further proceeded to forecast visitor count from China in the months of January 2020 – April 2020 based on the visitor record in the year 2019. We have then compared this forecasted data with the actual visitor count from January 2020 – April 2020 to identify and analyze the effect of Covid19 on South Korean tourism for Chinese visitors.

2. DATASET OVERVIEW

The entire analysis revolves around three datasets obtained from Kaggle: South Korea Visitors. [Source: https://www.kaggle.com/bappekim/south-korea-visitors]

1. Enter_korea_by_age.csv (Data separated by age for visitors)
2. Enter_korea_by_purpose.csv (Data separated by purpose for visitors)
3. Enter_korea_by_gender.csv (Data separated by gender for visitors)

All these datasets deal with the visitors of foreigners into South Korea. It includes visitors from other countries, overseas Koreans and crew members, except for some of the foreign arrivals who are not considered tourists (diplomats, soldiers, permanent residents, visiting cohabitation and residence).

Following are common column in each dataset:

- **date:** Date (year - month)

- **nation:** Country of departure
- **visitor:** Number of visitors
- **growth:** Growth percentage in the number of visitors compared to the same month last year
- **share:** Percentage of all visitors in the month

All the three datasets contain data for 16 months from January – 2019 to April – 2020. The visitor counts are aggregated for a month and reported on the first of every month. So, we can assume that the record count of visitors on 01 January 2019 is actually the cumulative visitor count for the previous month of December 2018. This assumption is true for all the months in our dataset. Each month's data is further divided into nation according to the origin of the tourists. There is total of 960 entries in each of the files.

ATTRIBUTE INFORMATION

1. Dataset: *Enter_Korea_by_gender.csv*

	date	nation	visitor	growth	share	male	female	crewman
0	2019-1	China	392814	28.737870	35.555117	147511	231722	13581
1	2019-1	Japan	206526	23.606830	18.693468	75070	129029	2427
2	2019-1	Taiwan	87954	16.003693	7.961057	30805	56202	947
3	2019-1	Hong Kong	35896	3.533212	3.249086	12172	22729	995
4	2019-1	Macao	2570	-12.376406	0.232621	748	1787	35

The gender is divided into 3 categories – Male, Female and Crewman. Crewmen are those people who serve in the transport company like the airplane, ship etc.

2. Dataset: Enter_Korea_by_age.csv

	date	nation	visitor	growth	share	age0-20	age21-30	age31-40	age41-50	age51-60	age61
0	2019-1	China	392814	28.737870	35.555117	36520	108591	103657	48574	40893	40998
1	2019-1	Japan	206526	23.606830	18.693468	18015	57921	34165	39811	33857	20330
2	2019-1	Taiwan	87954	16.003693	7.961057	18888	17927	18595	18862	8169	4566
3	2019-1	Hong Kong	35896	3.533212	3.249086	3890	11384	7400	5461	4629	2137
4	2019-1	Macao	2570	-12.376406	0.232621	223	1013	762	264	181	92

The visitors are divided into 6 age groups according to their ages. The age groups are Age 0 -20 years, Age 21 -30 years, Age 31 -40 years, Age 41 -50 years, Age 51 -60 years and Age 61 years and above.

3. Dataset: *Enter_Korea_by_purpose.csv*

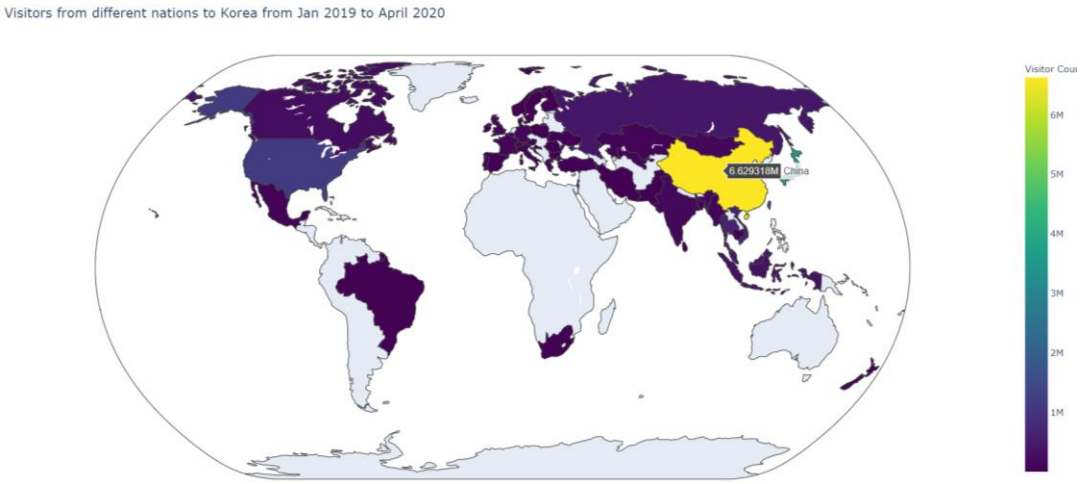
	date	nation	visitor	growth	share	tourism	business	official affairs	studying	others
0	2019-1	China	392814	28.737870	35.555117	320113	2993	138	8793	60777
1	2019-1	Japan	206526	23.606830	18.693468	198805	2233	127	785	4576
2	2019-1	Taiwan	87954	16.003693	7.961057	86393	74	22	180	1285
3	2019-1	Hong Kong	35896	3.533212	3.249086	34653	59	2	90	1092
4	2019-1	Macao	2570	-12.376406	0.232621	2506	2	0	17	45

The visitors are divided into 5 groups according to their purpose of their visit to South Korea as follows:

- Tourism – Tour the nation
- Business – Business or work related
- Official Affairs – Government Affairs
- Studying – Students for education
- Others

3. EXPLORATORY DATA ANALYSIS

We plotted the number of visitors according to origin nation for the whole dataset of 16 months from Jan-2019 to Apr - 2020 months using Choropleth from the plotly module.



From this plot we can observe that the highest number of visitors to South Korea was from China which was more than 6.5 Million and represented as yellow in the world map. Japan holds the 2nd position with more than 3 million visitors and is represented in green on the map.

nation	visitor
China	6629318
Japan	3695581
Taiwan	1424629
USA	1170719
Hong Kong	783159

These are the top 5 nation from which visitors come to South Korea the most.

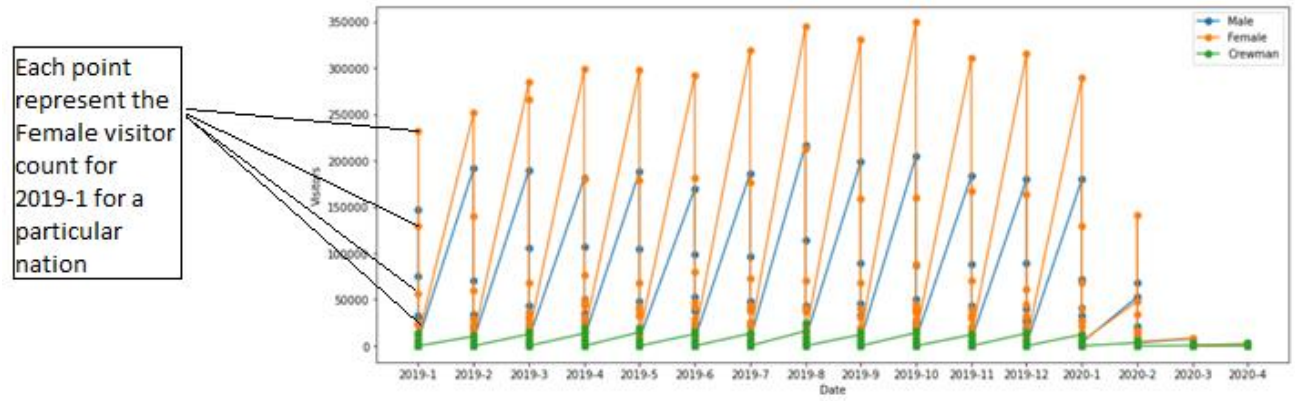
We have done a detailed exploratory data analysis on the three datasets and recorded our observations in the following sub-sections.

1. Dataset: *Enter_Korea_by_gender.csv*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 960 entries, 0 to 959
Data columns (total 8 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   date     960 non-null    object
 1   nation   960 non-null    object
 2   visitor  960 non-null    int64
 3   growth   960 non-null    float64
 4   share    960 non-null    float64
 5   male     960 non-null    int64
 6   female   960 non-null    int64
 7   crewman  960 non-null    int64
dtypes: float64(2), int64(4), object(2)
memory usage: 60.1+ KB
```

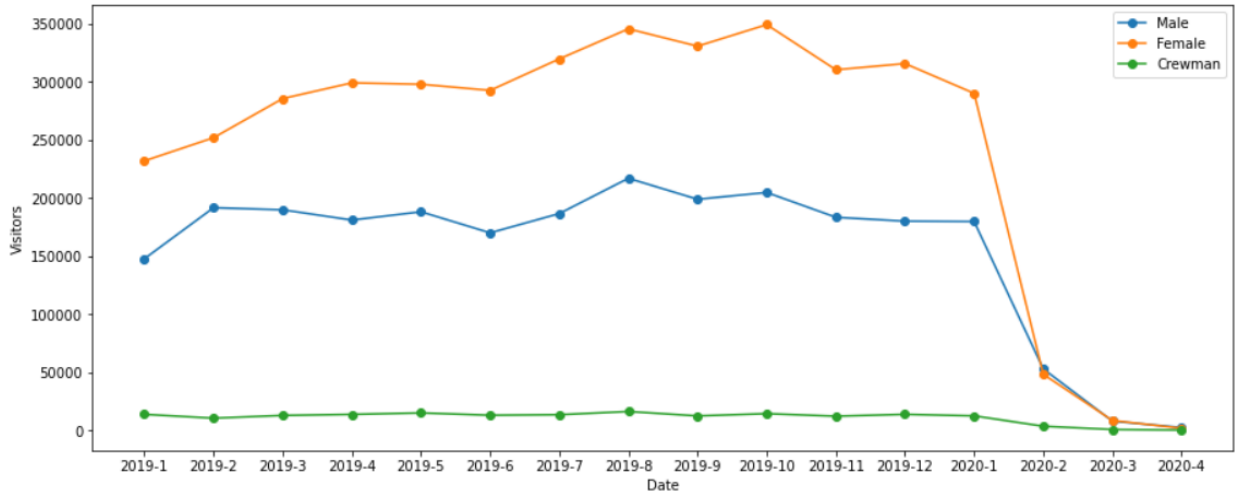
The above figure shows the generic information for each of the 7 attributes for *Enter_Korea_by_gender.csv* file. All the attributes have 960 data points and there are no missing values for this dataset.

We plot date vs visitors from this data for all the nations:



In the above plot we can see that the greatest portion of visitors are female represented as orange line followed by Male and the Crewman. The visitor count for a month for each of the 60 nations are represented as a point on the line in the plot. As the number of nations are quite large and the visitor count so close to each other this plot is difficult to study.

We will take our nation as China for further analysis as it has the highest visitor count and by doing so, we can get a clear view of the data.



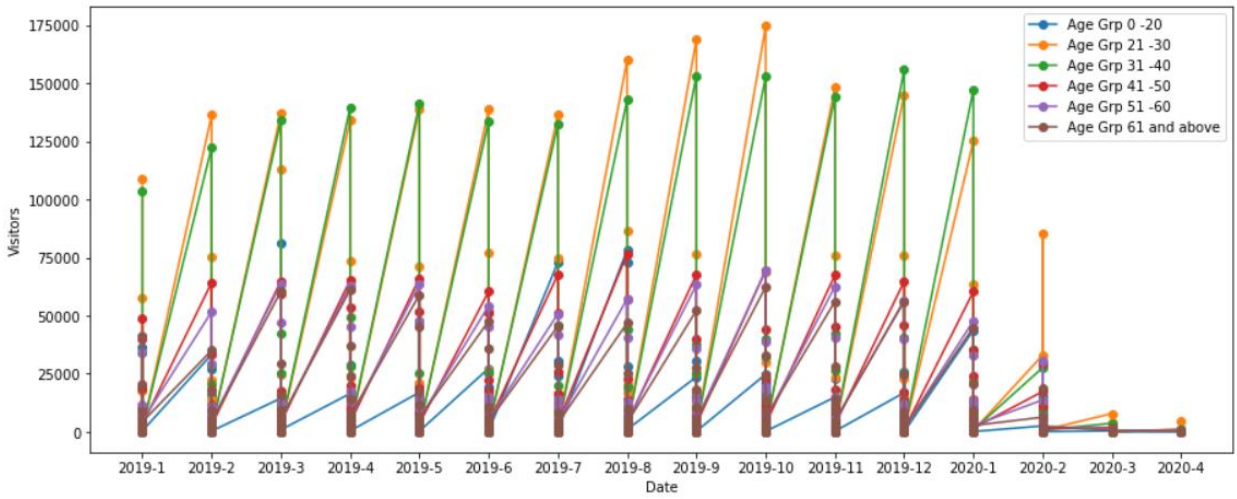
Here we get a clear view of the data for each month. From the above plot we can observe:

- Overall, most of the visitors are female followed by male and crewman.
- We can notice that the Female and the Male visitor follow almost same pattern.
- In 2019-6 we can see that there was sight drop in both and then again in 2019-9.
- From 2019-6 to 2019-8 both of have a steady growth and we see the highest number for female was for the month October in 2019.
- From January 2020 there is a sharp drop in both male and female visitors and it continues till April 2020 where it hits its all-time low count. This drop can be associated to the break of Coronavirus in the nation of China.

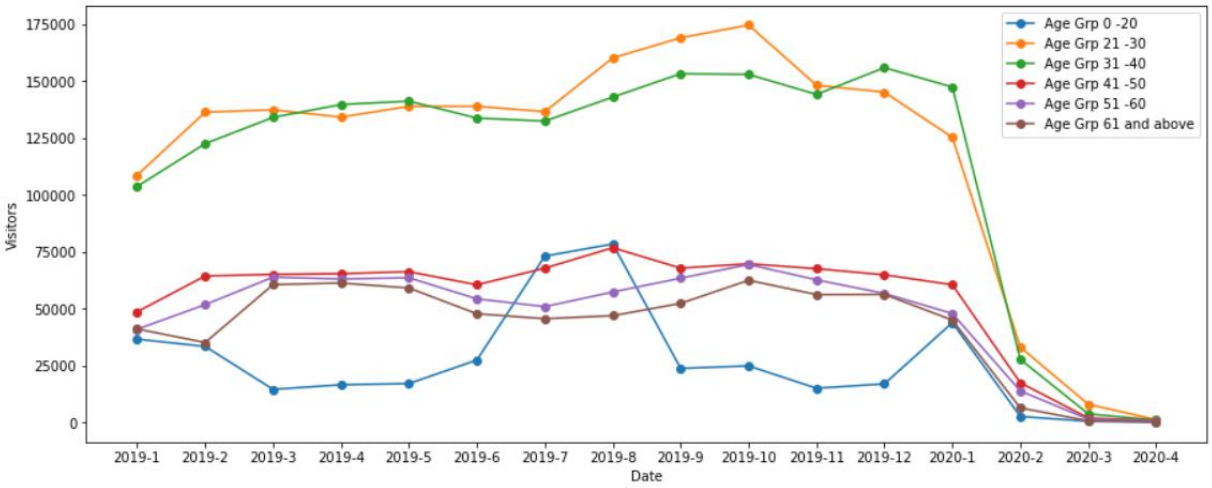
2. Dataset: *Enter_Korea_by_age.csv*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 960 entries, 0 to 959
Data columns (total 11 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   date      960 non-null    object
 1   nation    960 non-null    object
 2   visitor   960 non-null    int64
 3   growth    960 non-null    float64
 4   share     960 non-null    float64
 5   age0-20   960 non-null    int64
 6   age21-30  960 non-null    int64
 7   age31-40  960 non-null    int64
 8   age41-50  960 non-null    int64
 9   age51-60  960 non-null    int64
 10  age61     960 non-null    int64
dtypes: float64(2), int64(7), object(2)
memory usage: 82.6+ KB
```

This shows the generic information for each of the 11 attributes for *Enter_Korea_by_age.csv* file. All the attributes have 960 data points and again there are no missing values for this dataset.



This plot represents the full dataset with all 60 nations divided according to 6 age groups: Age 0-20, Age 21-30, Age 31-40, Age 41-50, Age 51-60 and Age 61 and above. Similarly, this plot is very difficult to interpret for this large number of nations. ***So, we would reduce it to only one nation – China.***



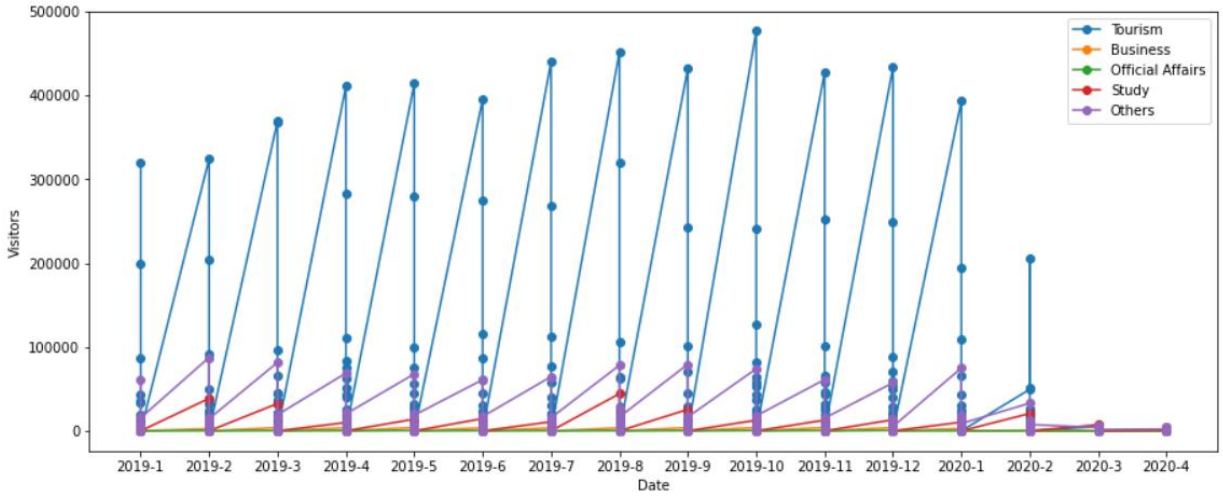
From the plot we can observe that:

- The most visitors from China are from the age group 21-30 and 31-40.
- First age group generally consist of young adults (age 21-30) who come from China maybe to study or to visit South Korea.
- The 2nd age group middle-aged adults (age 30-40) who main purpose of visit maybe work-related or leisure.
- There is also a rise of visitors from the age group 0-20 for the month July to August, this can be due to Summer holidays in school.
- In general, there is rise in visitors from August to October of 2019. After December 2019 we again notice a steep drop in numbers for all the age groups.

3. Dataset: Enter_Korea_by_purpose.csv

```
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   date             960 non-null    object
 1   nation           960 non-null    object
 2   visitor          960 non-null    int64
 3   growth           960 non-null    float64
 4   share            960 non-null    float64
 5   tourism          960 non-null    int64
 6   business         960 non-null    int64
 7   official affairs  960 non-null    int64
 8   studying         960 non-null    int64
 9   others           960 non-null    int64
dtypes: float64(2), int64(6), object(2)
memory usage: 75.1+ KB
```

This shows the generic information for each of the 10 attributes for Enter_Korea_by_purpose.csv file. All the attributes have 960 data points and there are no missing values for this dataset as well.

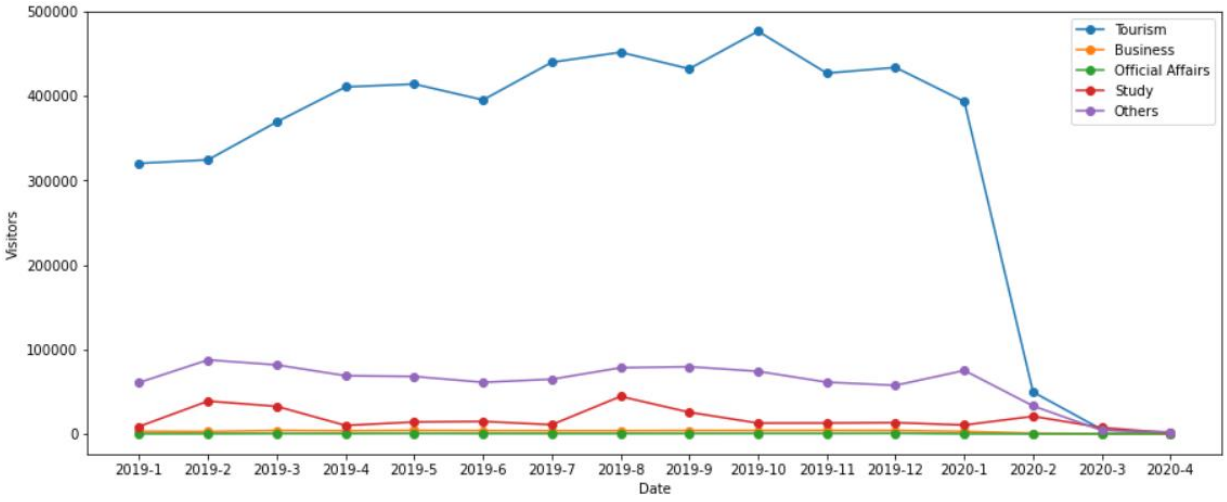


Similar to the above plots this also represents the whole dataset with all the nations but this time the visitors are grouped by the purpose of their visit. The blue line represents the purpose as tourism and is significantly higher than any of the other purpose.

We have separated the nation China to take a closer look at the data.

From the plot below, we can observe that:

- The most visitors who come to South Korea from China are tourists.
- There is rise in tourism form the month March to May as people come to see the Cherry Blossom followed by a slight dip and again picks up from October to November in 2019 as the weather is much cooler and the fall season starts.



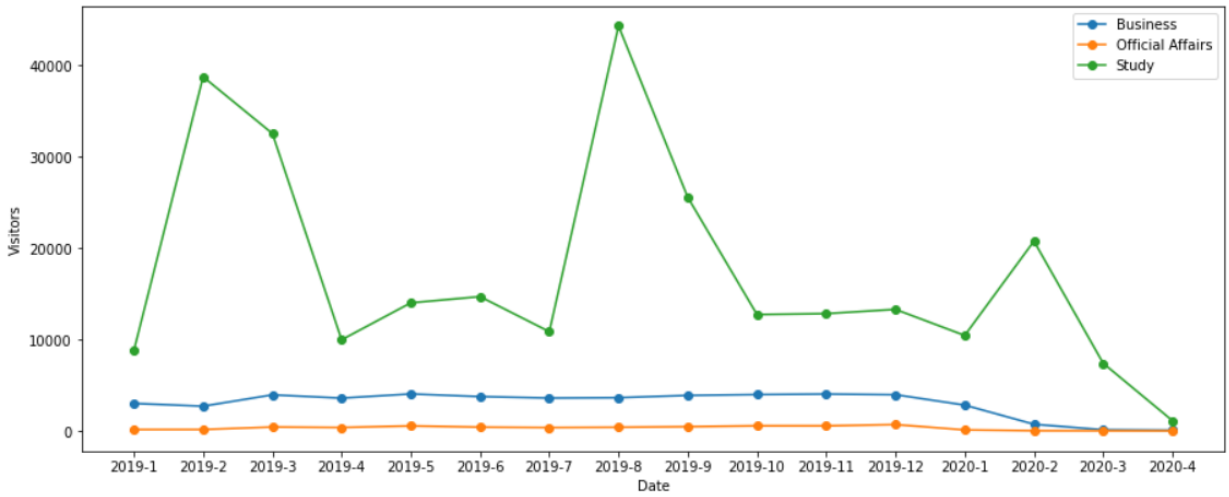
From the above plot we can observe that:

- The most visitors who come to South Korea from China are tourists.
- There is rise in tourism form the month March to May as people come to see the Cherry

Blossom followed by a slight dip and again picks up from October to November in 2019 as the weather is much cooler and the fall season starts.
[Source: https://www.audleytravel.com/south-korea/best-time-to-visit#dec]

- The number of tourists again sees a sharp decline from the month of February 2020 and hits the low at April of 2020. Again, we can associate this with the recent pandemic and the fact that it hit China the most in this time-period.

With the visitor count for tourism being so large we cannot observe the changes for purpose of study, business and Official Affairs. So, we plot them separately.



Now we can observe the changes in incoming visitors for studies more clearly.

- There is a sharp rise for the month of February and March of 2019 and then again it rises from August. This due to semester beginning at the start of March and again at the start of August. [Source: https://www.edarabia.com/school-holidays-south-korea).]
- As the pandemic hit China, the students that come to Korea for studies also take a hit from the month of February in 2020.
- The visitor count for people whose purpose for visit are Business and Official Affairs are almost constant throughout the 16 months.
- There is drop in number for Business from January of 2020 but visitors for Official Affairs remains the same.

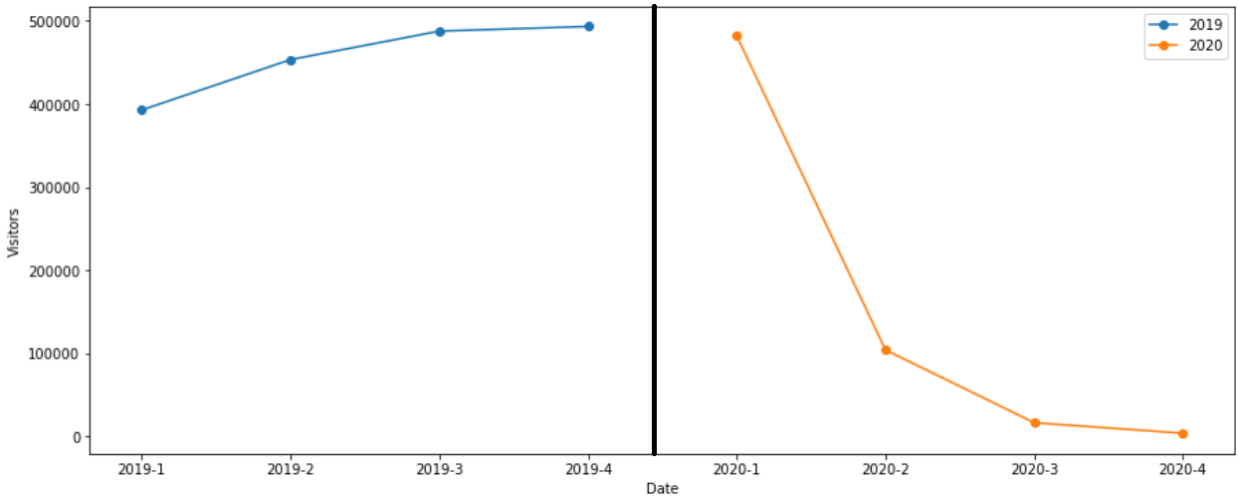
4. EDA on Merged Dataset

We have further combined the 3 datasets and applied some preprocessing to the merged dataset to group the purpose of visit and visitor age-groups into broader categories for easier study.

	date	visitor	tourism	non_tourism	male	female	crewman	ageBelow_30	age30_50	age50_above
0	2019-1	392814	320113	72701	147511	231722	13581	145111	152231	81891
1	2019-2	453379	324291	129088	191410	251668	10301	169605	186727	86746
2	2019-3	487623	369165	118458	189613	285340	12670	151730	198946	124277
3	2019-4	493250	410542	82708	180816	298870	13564	150594	204919	124173
4	2019-5	500413	413949	86464	187922	297661	14830	155808	207275	122500

In 16 months, we only have 4 months in common – January to April. We have compared the visitor records on 01 January – 01 April to see if there is any change in the trend between 2019 and 2020.

- **Comparing tourists according to visitor count**



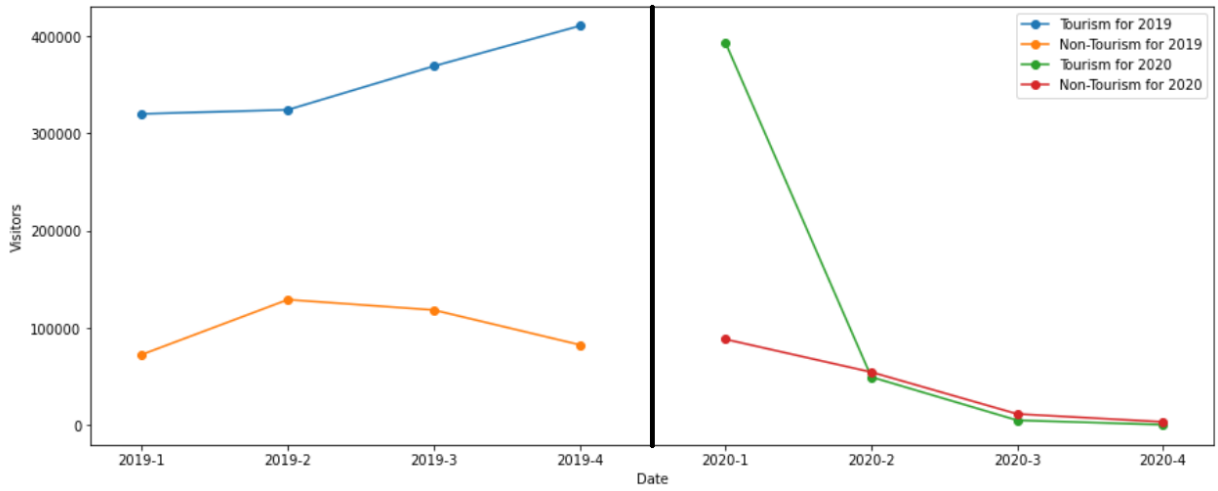
In the above figure, we have plotted the visitor count for both years for months from January to April in both 2019 and 2020.

From the plot we can observe that:

- For 2019 (Blue Line) from Jan to April there was a steady growth in the influx of visitor from China
- For 2020 (Orange Line) from Jan to April there is a sharp drop which indicates that very few visitors are visiting Korea from China during this time frame.

Again, we can associate this fluctuation due to the start of pandemic whose origin point was China during these 4 months.

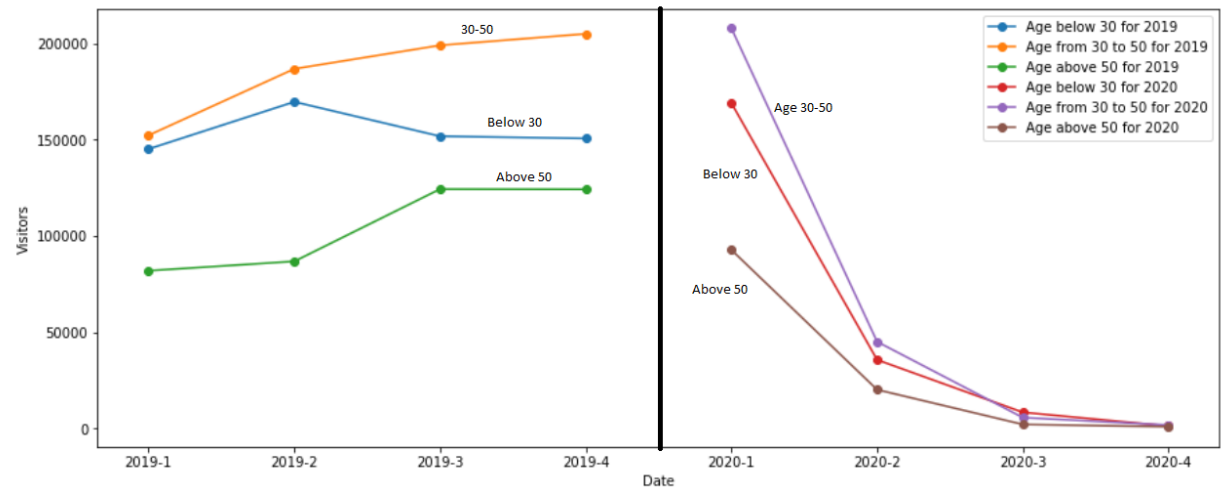
- **Comparing tourists according to Purpose of Visit – Tourism and Non-Tourism**



From the above plot we can observe:

- In 2019 people coming from china for tourism is way higher than people coming for other reasons. There is also a constant growth in number of people from Jan to April.
- In 2020 we see a very different picture the visitor for tourism has a sharp fall and it has gone below the non-tourism for the first time. From this we can conclude that people are still coming to visit Korea but that number has fallen considerably but visitors who are coming for other purpose has reduced but not as for tourism.

- **Comparing tourists according visitor Age-Groups – Below 30, Age 30-50 and Above 50**



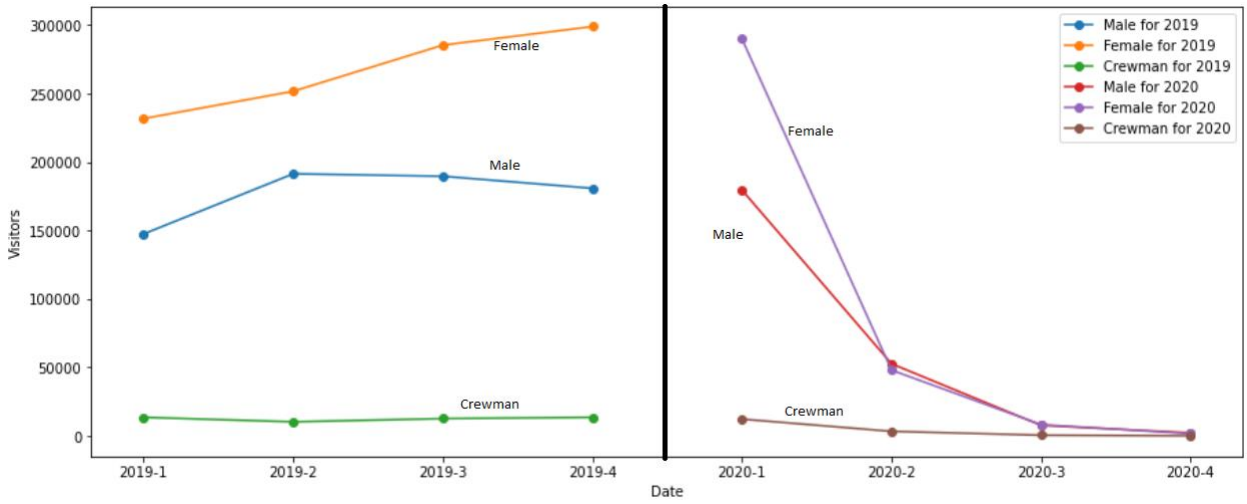
From the above plot we can observe:

- For age-group 30 to 50 there is significant drop in numbers from 2019 to 2020. In March

2020 it can be noticed that the count went below the age group 30 but in 2019 for this same month there was growth in visitors.

- In March, the drop of people in age group 30 -50 below age group 30 -50 maybe because of the reason that young people were getting less affected by the virus than the older ones.
- For Age group Above 50 had the least visitors in 2019 and this did not change in 2020 but as for every other category the total number has dropped.

- **Comparing tourists according to 3 gender groups – Male, Female and Crewman**

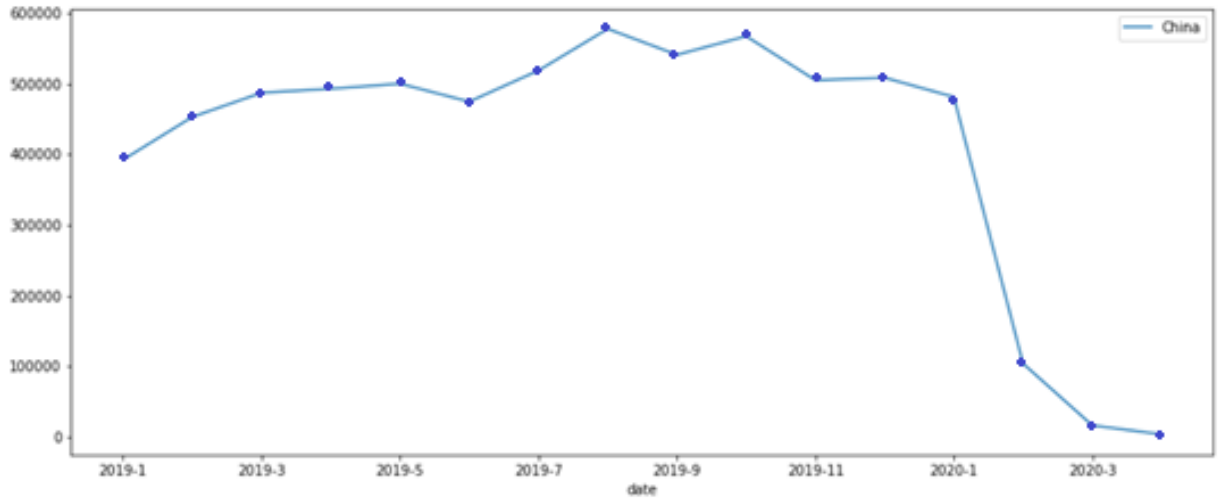


From the above plot we can observe:

- Both male and female gender groups have seen a huge drop in number but what is interesting is that number of females for month February 2020 has gone below the male count. The female visitors in 2019 was always higher than the male but same cannot be said for 2020.
- For both the years the crewman numbers have remained almost constant with maybe a slight drop from March 2020.

4. DATA PREPROCESSING

Among the 960 data points in the datasets we have only 16 unique date data and it is repeated for each of the 60 nation. For doing a prediction we can only use those 16 data that belongs to a certain nation. So, we chose to use the nation as China as it has the highest number of visitors count so observing the trend will be easier.



In the above plot we can observe that that it shows the count of visitors per month for China through the 16 months' time-period. Each point on the line represents the count for that month.

1. Dataset: *Enter_Korea_by_purpose.csv*

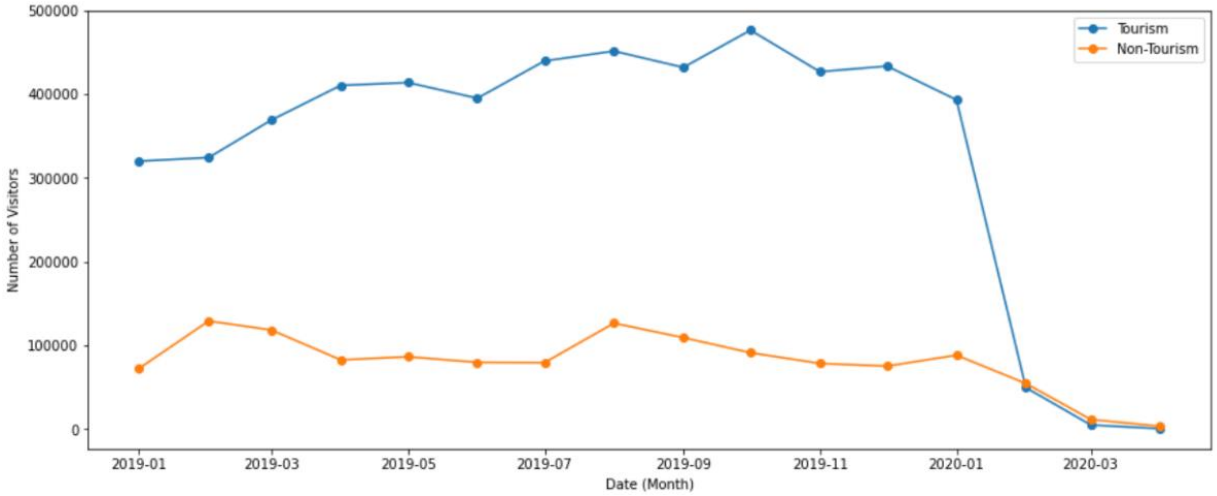
We filter the nation as China for this dataset. There are total of 16 data points with unique dates.

```
df_purpose_china = df_purpose.loc[df_purpose['nation'] == 'China']
```

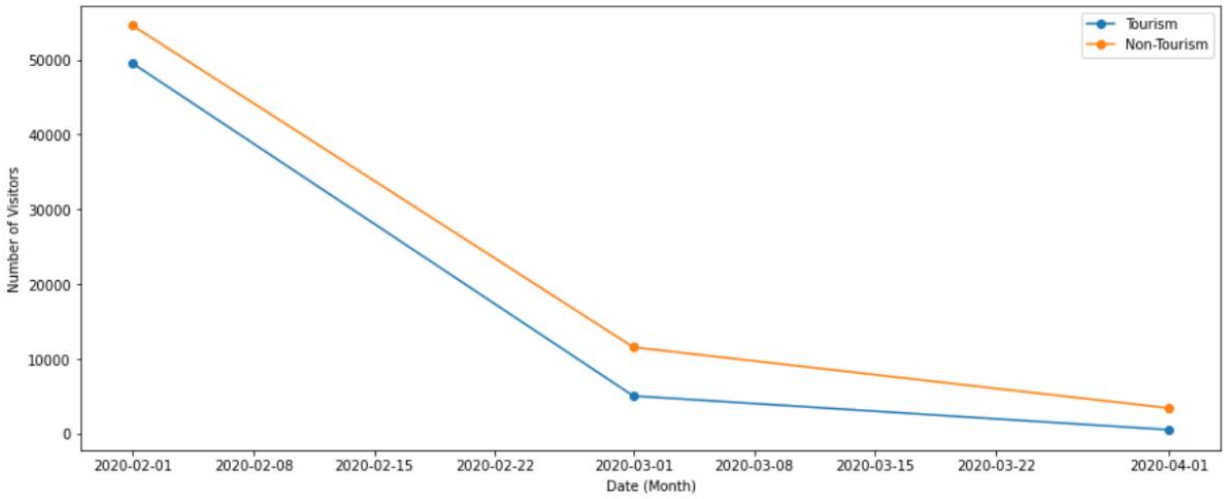
```
df_purpose_china
```

	date	nation	visitor	growth	share	tourism	business	official_affairs	studying	others
0	2019-1	China	392814	28.737870	35.555117	320113	2993	138	8793	60777
60	2019-2	China	453379	31.284441	37.724933	324291	2694	143	38731	87520
120	2019-3	China	487623	20.874389	31.753711	369165	3933	414	32532	81579
180	2019-4	China	493250	34.545722	30.166978	410542	3575	362	9959	68812
240	2019-5	China	500413	35.165657	33.682331	413949	4034	534	14003	67893
300	2019-6	China	475007	25.037708	32.177294	395196	3743	399	14680	60989
360	2019-7	China	519132	26.513573	35.849999	439699	3587	356	10847	64643
420	2019-8	China	578112	20.908521	36.444075	451570	3625	388	44291	78238
480	2019-9	China	541350	24.564249	37.087302	432018	3873	447	25545	79467
540	2019-10	China	567695	19.437542	34.277063	476460	3967	547	12722	73999
600	2019-11	China	505369	25.012121	34.699185	426849	4020	545	12822	61133
660	2019-12	China	508877	22.244216	34.929041	433577	3951	682	13284	57383
720	2020-1	China	481681	22.623175	37.846937	393336	2813	99	10433	75000
780	2020-2	China	104086	-77.042166	15.190335	49520	715	11	20753	33087
840	2020-3	China	16595	-96.596756	19.874966	5040	115	2	7388	4050
900	2020-4	China	3935	-99.202230	13.377528	522	71	0	1112	2230

- The column name for official-affairs is discontinuous so transforming it to a continuous value: official_affairs.
- For the purpose of visitor forcasting in 2020, have divided the purpose of visit of Chinese tourists into two groups, namely ***Tourism*** and ***Non_Tourism***. We have consolidated the visitors who come to South Korea for the purpose of Business, Study, Official Affairs and Others into one column and classified them as Non_Tourism.



As we can see from the above plot, more people from China came to South Korea for the purpose of Tourism in the year 2019. However, in the plot below we can see that the statistics changed in 2020, where we can see that people from China arrived mostly for Non-Tourism purpose like Business, Study, Official Affairs and Others. Since the Covid19 pandemic hit China in December 2019, we can safely assume that people from China travelled to South Korea mainly for non-tourism purpose like business, official travels, study, etc.



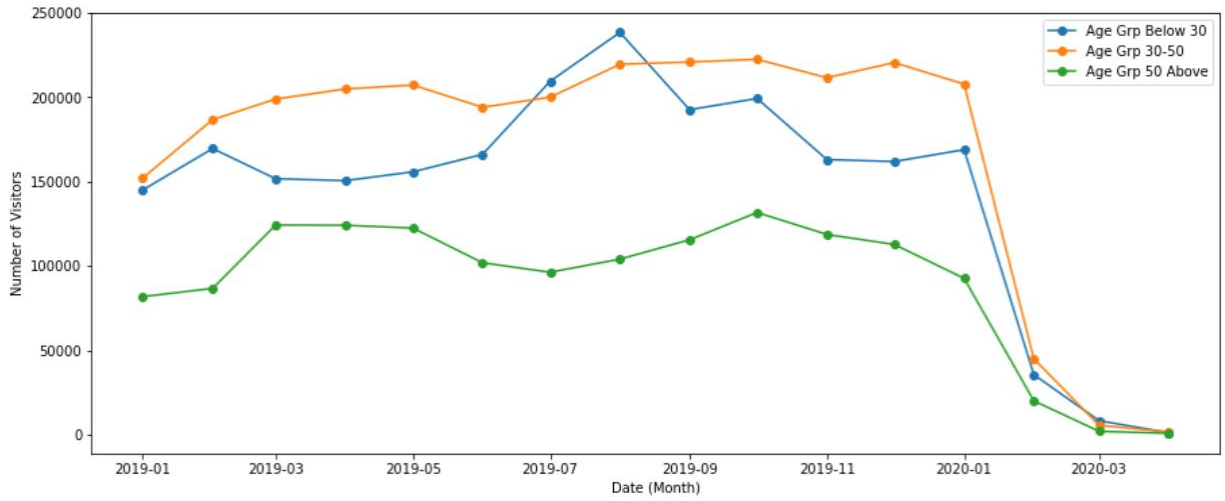
2. Dataset: *Enter_Korea_by_age.csv*

We filter the nation as China for this dataset. There are total of 16 data points with unique dates.

- The column name for the age groups is discontinuous so transforming it to a continuous value
- We have grouped the visitors based on their age into 3 broad age-groups:
  - **ageBelow30** – young people

- **age30_50** – middle aged people
- **age51_above** – senior people

- We also dropped the unnecessary column of nation, visitor, growth and share as they are not needed for the forecasting purpose.



3. Dataset: *Enter_Korea_by_gender.csv*

In the data exploration we noticed that the trend for female and the male were nearly identical and the crewman did not show much fluctuation. Therefore, doing any preprocessing and the applying forecasting model will not yield any new trend or predict anything that is not already shown for the other 2 datasets. Thus, we did not further explore this dataset.

4. Merged Dataset

Combined all the 3 dataset that were preprocessed above which was divided by age, purpose and gender into one based on the common columns present in each dataset – nation and date. For forecasting purposes, we used the individual data sets mentioned above. For Data Exploration this combined data was used (All the results are listed in the that section).

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 16 entries, 0 to 15
Data columns (total 13 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   date          16 non-null     object
 1   visitor       16 non-null     int64
 2   tourism       16 non-null     int64
 3   non_tourism   16 non-null     int64
 4   nation        16 non-null     object
 5   growth        16 non-null     float64
 6   share         16 non-null     float64
 7   male          16 non-null     int64
 8   female        16 non-null     int64
 9   crewman       16 non-null     int64
 10  ageBelow_30   16 non-null     int64
 11  age30_50      16 non-null     int64
 12  age50_above   16 non-null     int64
dtypes: float64(2), int64(9), object(2)
memory usage: 1.8+ KB
```

5. TIME SERIES FORECASTING

We have tried to forecast visitor count from China in the months of January 2020 – April 2020 based on the visitor record in the year 2019 using ARIMA and PROPHET. We have then compared this forecasted data with the actual visitor count from January 2020 – April 2020 to identify the effect of Covid19 on Chinese visitors and South Korean tourism.

5.1 ARIMA Forecasting Model

ARIMA stands for Auto Regressive Integrated Moving Average model, which is a class of statistical models for analyzing and forecasting time series data.

It explicitly caters to a suite of standard structures in time series data and provides a simple yet powerful method of making skillful time series forecasts.

Each of these components AR (Autoregression), I (Integrated) and MA (Moving Average) are explicitly specified in the model as a parameter. A standard notation is used for ARIMA (p, d, q) where the parameters AR, I and MA are substituted with integer values p, d and d respectively to quickly indicate the specific ARIMA model being used.

These ARIMA model parameters can be defined as follows:

p: The number of lag observations included in the model, also called the lag order.
d: The number of times that the raw observations are differenced, also called the degree of differencing.
q: The size of the moving average window, also called the order of moving average.

A linear regression model is constructed including the specified number and type of terms, and the data is prepared removing trend and seasonal structures that negatively affect the regression model.

Akaike Information Criterion (AIC) Score:

The AIC score helps us to test how well our model fits the data set without over-fitting it by rewarding models that achieve a high goodness-of-fit score and penalizing them if they become overly complex.

The model with the lower AIC score is expected to strike a superior balance between its ability to fit the data set and its ability to avoid over-fitting the data set.

We have used AIC scores of competing ARIMA models to find the values of p, q, d parameters that work best for our ARIMA forecasting model.

Steps taken for time series forecasting using ARIMA model:

- Find p ,q, d parameter combinations for Seasonal ARIMA

```
# Find p ,q, d parameter combinations for Seasonal ARIMA
p = d = q = range(0, 2)
pdq = list(itertools.product(p, d, q))
seasonal_pdq = [(x[0], x[1], x[2], 12) for x in list(itertools.product(p, d, q))]
```

- *For purpose Tourism:*
  - *Divide the dataset into train and test data:*
    Since the visitor counts are aggregated for a month and reported on the first of every month, we assume that the record count of visitors on 01 January 2020 is actually the visitor count for December 2019. This assumption is true for all the months in our dataset. We divided the 2019 data for training (January 2019 – January 2020) and 2020 data (February 2020 – April 2020) for testing the ARIMA forecasting model.
  - Find the optimal value of p, d, q for the best ARIMA model based on the lowest AIC score.

```
# Visitor data for purpose 'tourism'
dfTourism = df_purpose_china['tourism']

# Training Dataset: Tourist data for the year 2019
train = dfTourism['2019-01-01':'2020-01-01']

# Find the optimal value of p, d, q for best ARIMA model
for param in pdq:
    for param_seasonal in seasonal_pdq:
        try:
            mod = sm.tsa.statespace.SARIMAX(train,
                                            order=param,
                                            seasonal_order=param_seasonal,
                                            enforce_stationarity=False,
                                            enforce_invertibility=False)
            results = mod.fit()
            print('ARIMA{}x{}12 - AIC:{}'.format(param, param_seasonal, results.aic))
        except:
            continue
```

- The AIC scores of the models are manually compared and p, q, d values of the model with lowest AIC score are chosen as parameters for our ARIMA forecasting model.

```
ARIMA(0, 0, 0)x(0, 0, 0, 12)12 - AIC:346.56391768865797
ARIMA(0, 0, 0)x(0, 0, 1, 12)12 - AIC:4.0
ARIMA(0, 0, 0)x(0, 1, 0, 12)12 - AIC:2.0
ARIMA(0, 0, 0)x(1, 0, 0, 12)12 - AIC:27.24710013942077
ARIMA(0, 0, 0)x(1, 0, 1, 12)12 - AIC:6.0
ARIMA(0, 0, 1)x(0, 0, 0, 12)12 - AIC:312.84992578720943
ARIMA(0, 0, 1)x(0, 0, 1, 12)12 - AIC:6.0
ARIMA(0, 0, 1)x(1, 0, 0, 12)12 - AIC:34.90027025587474
ARIMA(0, 0, 1)x(1, 0, 1, 12)12 - AIC:8.0
ARIMA(0, 1, 0)x(0, 0, 0, 12)12 - AIC:262.9075392159796
ARIMA(0, 1, 0)x(0, 0, 1, 12)12 - AIC:4.0
```

  As we can see the lowest AIC score of the competing ARIMA models is 2.0. Hence the parameter values are chosen as p=0, q=0, d=0 i.e., order=(0,0,0) and seasonal_order=(0,1,0,12)

- The training dataset for purpose – Tourism is fitted using ARIMA model with parameters order=(0,0,0), seasonal_order=(0,1,0,12), enforse_stationarity=False and enforce_invertibility=False.

```
# Fitting the ARIMA model on training data (2019 data) based on lowest AIC score
mod = sm.tsa.statespace.SARIMAX(train,
                                order=(0, 0, 0),
                                seasonal_order=(0, 1, 0, 12),
                                enforce_stationarity=False,
                                enforce_invertibility=False)
results = mod.fit()
```

- Future tourist prediction/forecasting is done a period of 3 months i.e., 1 February 2020 (accumulated visitor count in the month of January 2020) to 1 April 2020 (accumulated visitor count in the month of March 2020).

```
# Forcast 'tourist' visitor count for 2020
pred_uc = results.get_forecast(steps=3)
pred_ci = pred_uc.conf_int()
```

- The same steps are repeated for purpose Non_Tourism, and the three age groups, namely young visitors – ageBelow30, middle aged visitors – age30_50 and senior visitors – age51_above and the forecasting results are plotted for further comparison.

5.2 PROPHET Forecasting Model

Prophet is an open source software released by Facebook's Core Data Science team for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly,

and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data.

Prophet is fully automatic and gets a reasonable forecast on messy data with no manual effort. It is robust to outliers, missing data, and dramatic changes in the time series.

Prophet uses a very flexible regression model (somewhat like curve-fitting) instead of a traditional time series model for accurate forecasting because it gives more modeling flexibility, makes it easier to fit the model, and handles missing data or outliers more gracefully.

At its core, the Prophet procedure is an additive regression model with four main components:

- A piecewise linear or logistic growth curve trend. Prophet automatically detects changes in trends by selecting changepoints from the data.
- A yearly seasonal component modeled using Fourier series.
- A weekly seasonal component using dummy variables.
- A user-provided list of important holidays.

Steps taken for time series forecasting using PROPHET:

- Create a virtual environment for Prophet and install the Prophet package from PyPI (Python)
- *For purpose Tourism:*
  - Reset the date index and flatten the dataset so that date forms a new column in our dataset
  - *Divide the dataset into train and test data:*
    Since the visitor counts are aggregated for a month and reported on the first of every month, we assume that the record count of visitors on 01 January 2020 is actually the visitor count for December 2019. This assumption is true for all the months in our dataset. We divided the 2019 data for training (January 2019 – January 2020) and 2020 data (February 2020 – April 2020) for testing the PROPHET forecasting model.

```
# Reset index and flatten dataset
df_purpose_china = df_purpose_china.reset_index()

# Training Dataset: Visitor data for the year 2019
df_purpose_china2019 = df_purpose_china[:13]
print(df_purpose_china2019)
```

    The figure below shows the training dataset with date column and data for the year 2019.

	date	tourism	non_tourism
0	2019-01-01	320113	72701
1	2019-02-01	324291	129088
2	2019-03-01	369165	118458
3	2019-04-01	410542	82708
4	2019-05-01	413949	86464
5	2019-06-01	395196	79811
6	2019-07-01	439699	79433
7	2019-08-01	451570	126542
8	2019-09-01	432018	109332
9	2019-10-01	476460	91235
10	2019-11-01	426849	78520
11	2019-12-01	433577	75300
12	2020-01-01	393336	88345

- *Prepare the training dataset for forecasting with PROPHET:*

  - Divide the training dataset into 2 separate datasets for Tourism (dfTourism) and Non_Tourism (dfNonTourism)

  - The PROPHET model uses the date 'ds' column to train the forecasting model using data in the 'y' column. So, pre-process the dataset dfTourism by renaming the date column as 'ds' and tourism column as 'y'. Similarly pre-process the dataset dfNonTourism by renaming the date column as 'ds' and non_tourism column as 'y'

```
# Prepare the dataset for modeling with Prophet
dfTourism = df_purpose_china2019.drop(columns=['non_tourism'])
dfTourism = dfTourism.rename(columns={'date': 'ds', 'tourism': 'y'})
print(dfTourism)

dfNonTourism = df_purpose_china2019.drop(columns=['tourism'])
dfNonTourism = dfNonTourism.rename(columns={'date': 'ds', 'non_tourism': 'y'})
print(dfNonTourism)
```

The figure below shows the training dataset dfTourism for forecasting.

	ds	y
0	2019-01-01	320113
1	2019-02-01	324291
2	2019-03-01	369165
3	2019-04-01	410542
4	2019-05-01	413949
5	2019-06-01	395196
6	2019-07-01	439699
7	2019-08-01	451570
8	2019-09-01	432018
9	2019-10-01	476460
10	2019-11-01	426849
11	2019-12-01	433577
12	2020-01-01	393336

The figure below shows the training data dfNonTourism for forecasting.

	ds	y
0	2019-01-01	72701
1	2019-02-01	129088
2	2019-03-01	118458
3	2019-04-01	82708
4	2019-05-01	86464
5	2019-06-01	79811
6	2019-07-01	79433
7	2019-08-01	126542
8	2019-09-01	109332
9	2019-10-01	91235
10	2019-11-01	78520
11	2019-12-01	75300
12	2020-01-01	88345

- The training datasets are fitted using the PROPHET forecasting model with parameter interval_width=0.95.

- Future tourist prediction/forecasting is done a period of 3 months i.e., 1 February 2020 (accumulated visitor count in the month of January 2020) to 1 April 2020 ((accumulated visitor count in the month of March 2020).

```
# Forcast 'Tourist' visitor count
dfTourism_model = Prophet(interval_width=0.95)
dfTourism_model.fit(dfTourism)
dfTourism_forecast = dfTourism_model.make_future_dataframe(periods=3, freq='MS')
dfTourism_forecast = dfTourism_model.predict(dfTourism_forecast)

# Forcast 'Non-Tourist' visitor count
dfNonTourism_model = Prophet(interval_width=0.95)
dfNonTourism_model.fit(dfNonTourism)
dfNonTourism_forecast = dfNonTourism_model.make_future_dataframe(periods=3, freq='MS')
dfNonTourism_forecast = dfNonTourism_model.predict(dfNonTourism_forecast)
```

- The same steps are repeated for forecasting visitor count from China based on the three age groups, namely young visitors – **ageBelow30**, middle aged visitors – **age30_50** and senior visitors – **age51_above** and the forecasting results are plotted for further comparison.

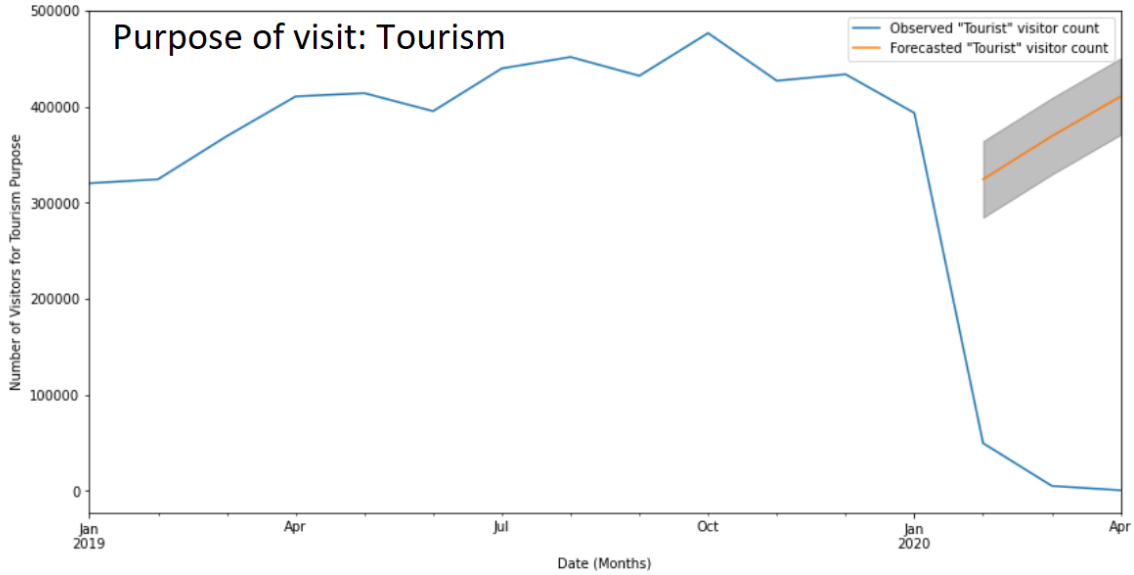
6. FORECASTING RESULTS AND COMPARISON

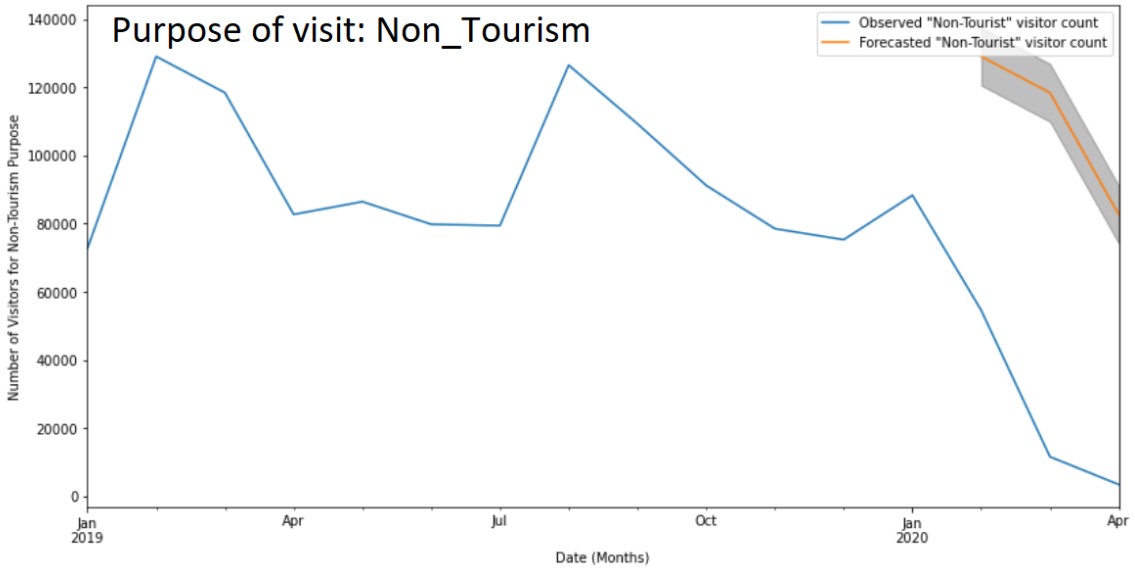
We have used ARIMA and PROPHET forecasting models to predict the number of Chinese visitors in South Korea. We have trained our models using tourism data from China to South Korea from 01 January 2019 to 01 January 2020. We have forecasted the tourist count for the duration of February 2020 to April 2020 and compared our result with the actual tourist data during the same period of time

6.1 ARIMA Forecasting Model

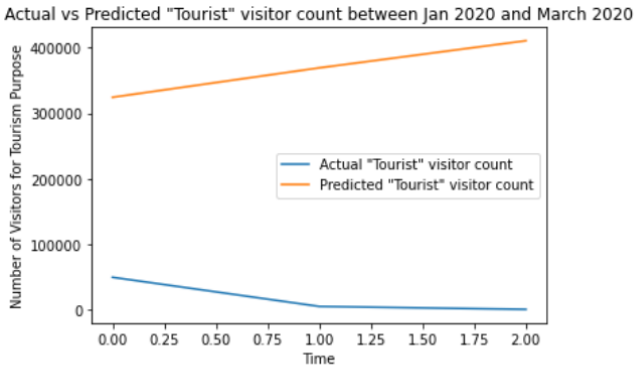
Forecasting Based on Purpose of Visit

The ARIMA forecasting plot on the Purpose of Visit is shown in the figures below.

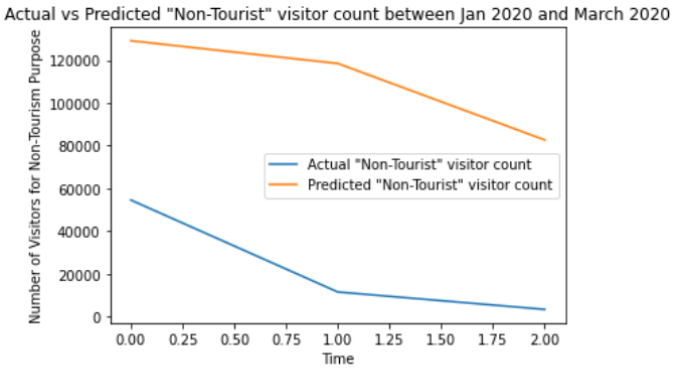




We compared the actual verses predicted per-month cumulated Chinese visitor data based on their Purpose of Visit for the months of January 2020 – March 2020 in the following figures.



Mean Absolute Error: 349638.667



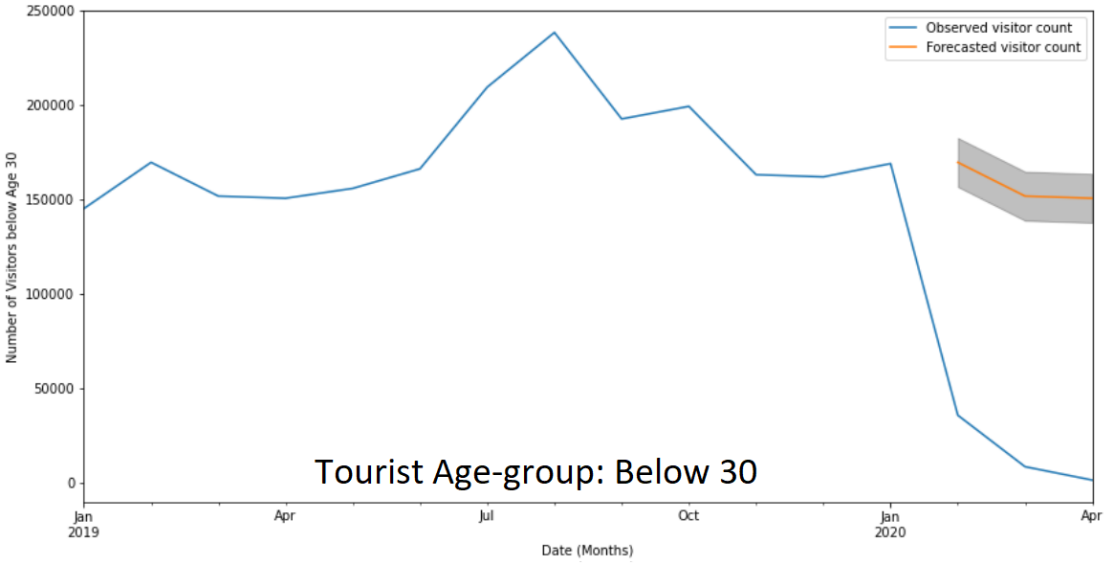
Mean Absolute Error: 86906.667

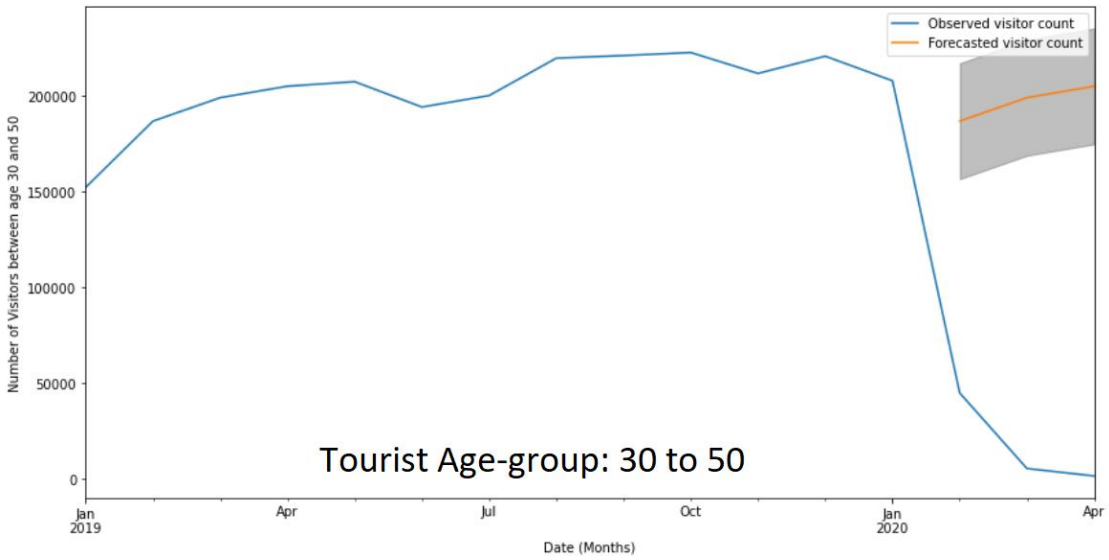
As we can observe from the plots:

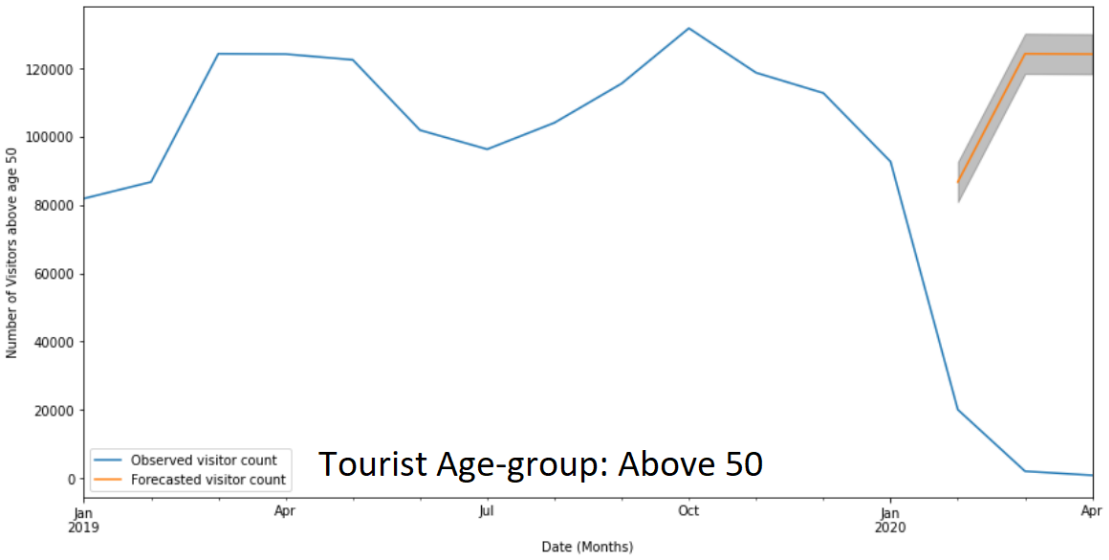
- The actual 'Tourist' visitor count from China is extremely low compared to the forecasted 'Tourist' visitor count, with the Mean Absolute Error (MAE) = 349638.667
- The actual 'Non-Tourist' visitor count from China is also extremely low compared to the forecasted 'Non-Tourist' visitor count, with the Mean Absolute Error (MAE) = 86906.667

Forecasting Based on Visitor Age-Group

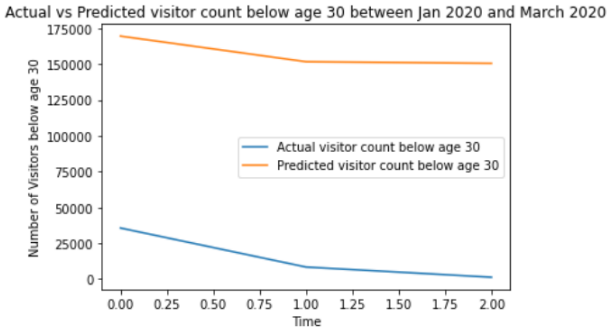
The ARIMA forecasting plot on Age-Group of visitors is shown in the figures below.



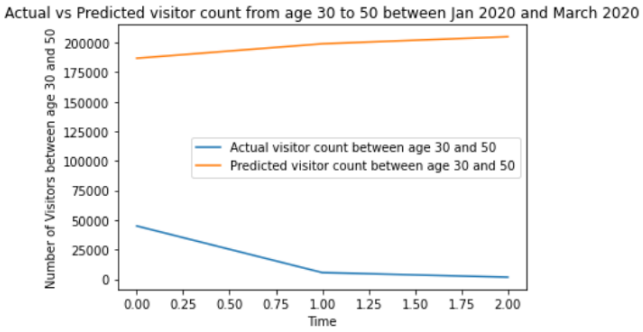




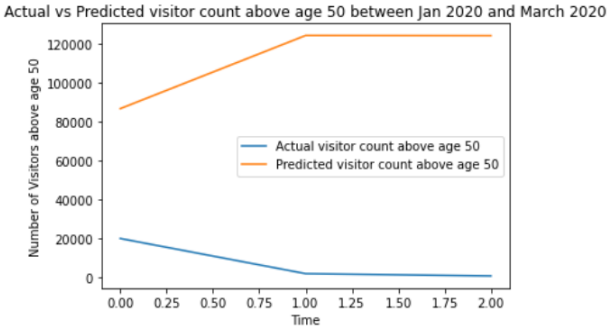
We compared the actual verses predicted per-month cumulated Chinese visitor data based on their Age Groups for the months of January 2020 – March 2020 in the following figures.



Mean Absolute Error: 142245.667



Mean Absolute Error: 179405.333



Mean Absolute Error: 104047.333

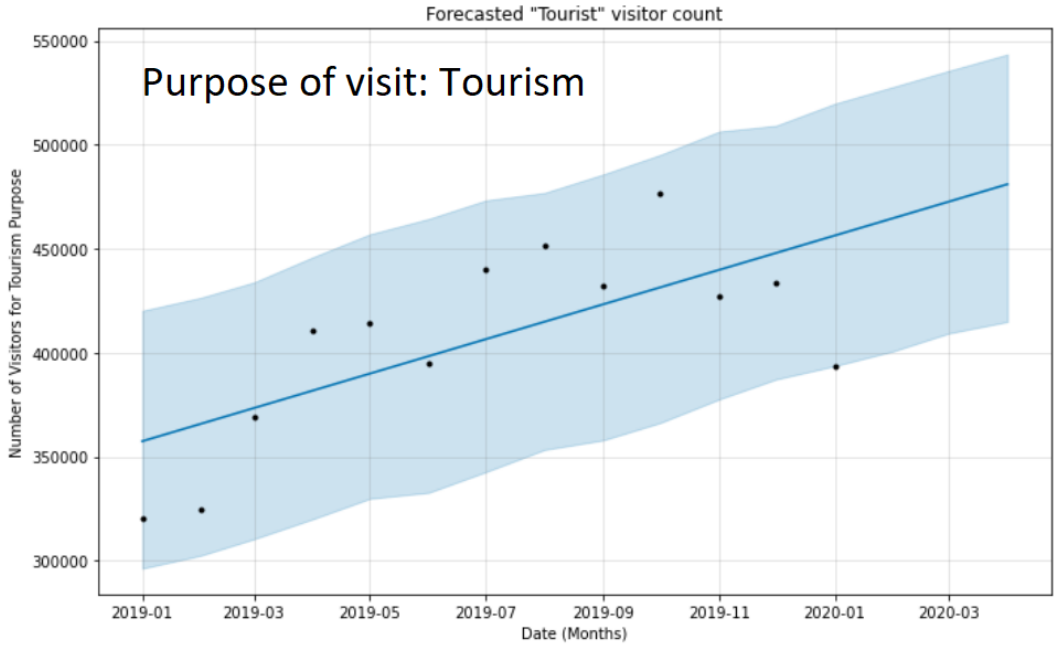
As we can observe from the plots:

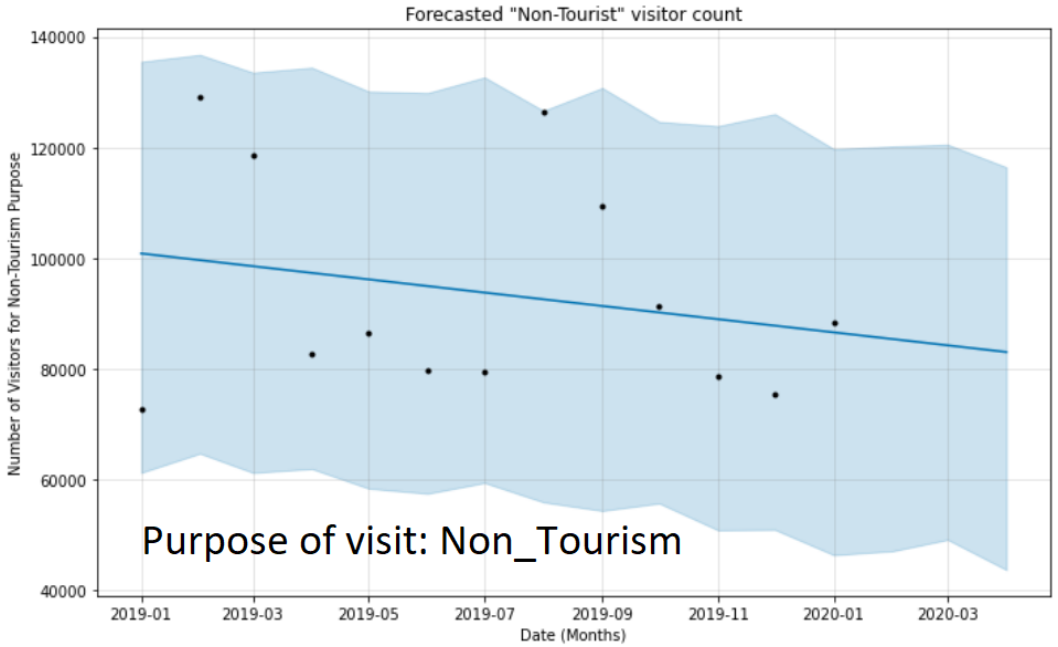
- The actual visitor count below age 30 from China is extremely low compared to the forecasted visitor count below age 30, with the Mean Absolute Error (MAE) = 142245.667
- The actual visitor count between age 30 and 50 from China is extremely low compared to the forecasted visitor count between age 30 and 50, with the Mean Absolute Error (MAE) = 179405.333
- The actual visitor count above age 50 from China is extremely low compared to the forecasted visitor count above age 50, with the Mean Absolute Error (MAE) = 104047.33

6.2 PROPHET Forecasting Model

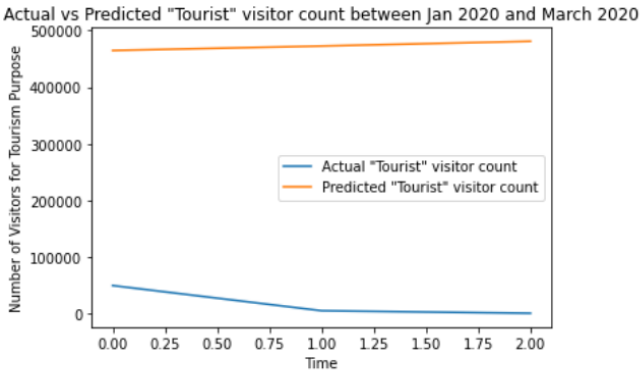
Forecasting Based on Purpose of Visit

The PROPHET forecasting plot on the Purpose of Visit is shown in the figures below.

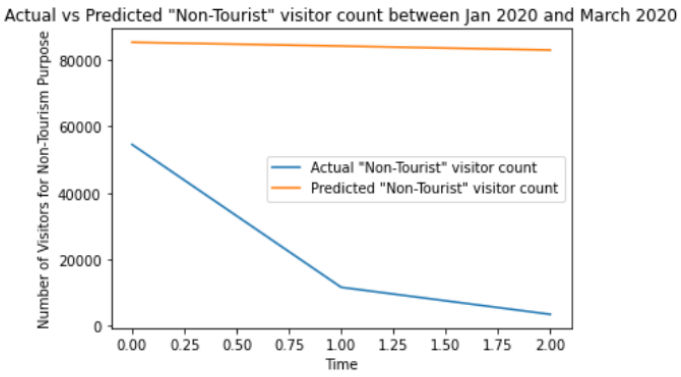




We compared the actual verses predicted per-month cumulated Chinese visitor data based on their Purpose of Visit for the months of January 2020 – March 2020 in the following figures.



Mean Absolute Error: 454415.307



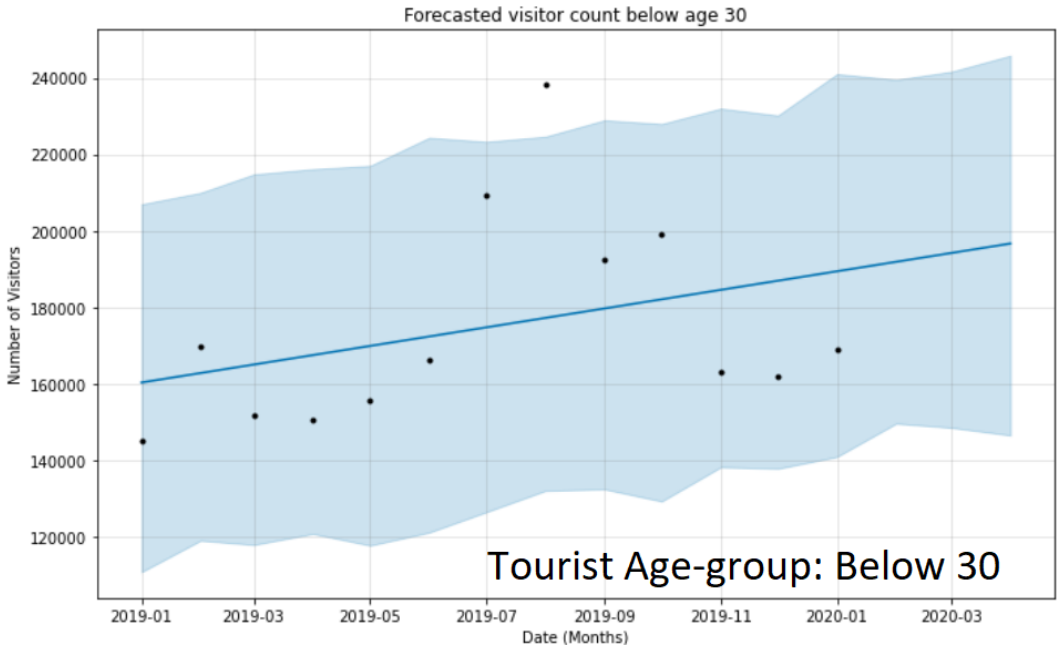
Mean Absolute Error: 61045.460

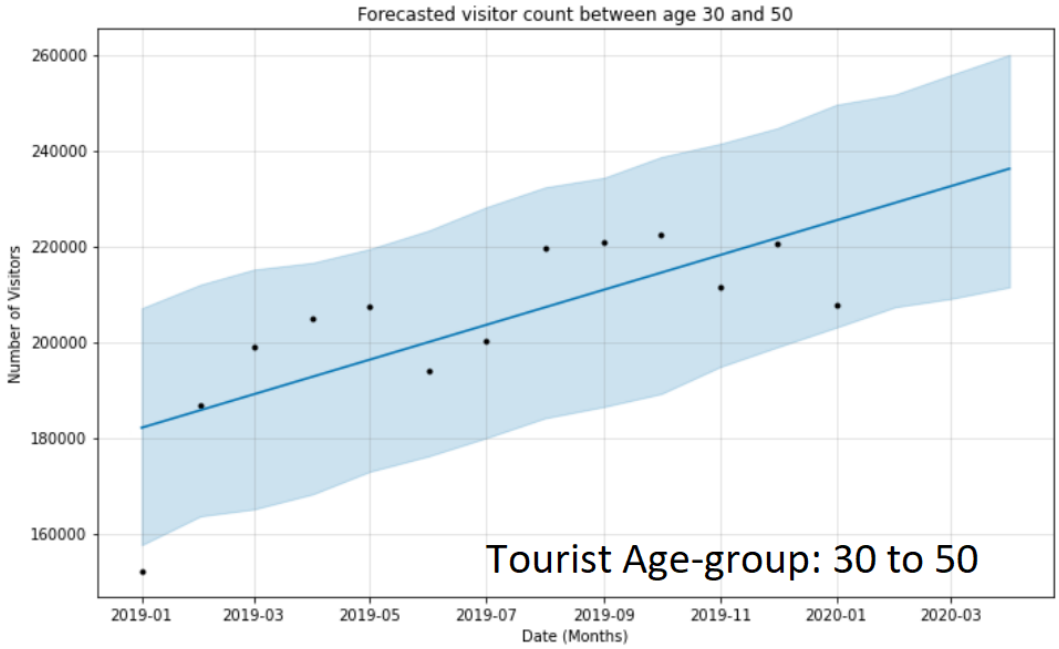
As we can observe from the plots:

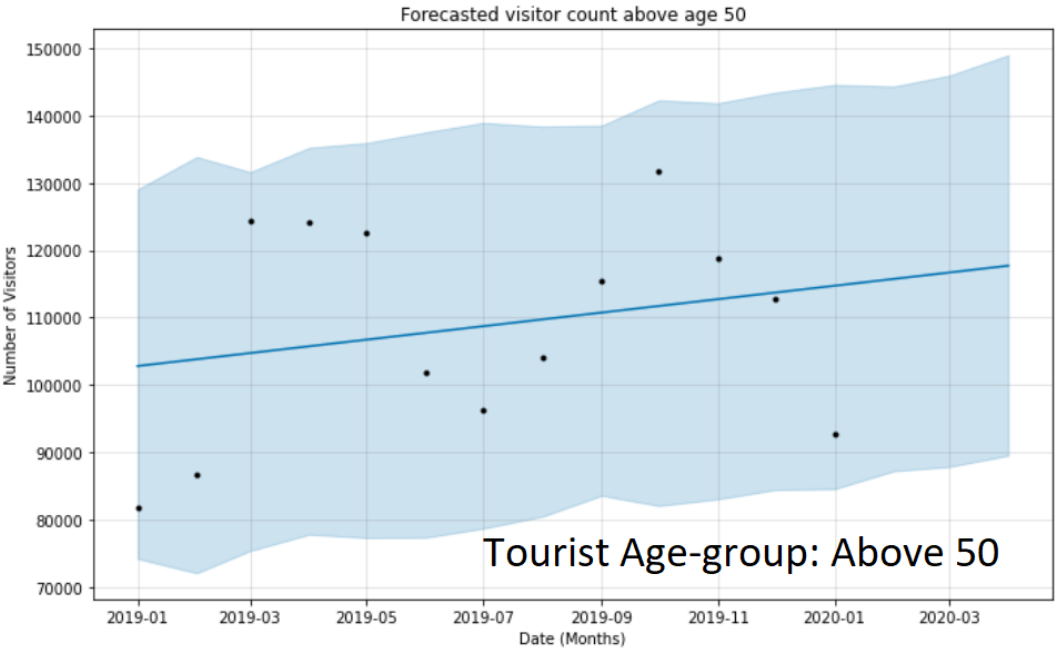
- The actual 'Tourist' visitor count from China is extremely low compared to the forecasted 'Tourist' visitor count, with the Mean Absolute Error (MAE) = 454415.307
- The actual 'Non-Tourist' visitor count from China is also extremely low compared to the forecasted 'Non-Tourist' visitor count, with the Mean Absolute Error (MAE) = 61045.460

Forecasting Based on Visitor Age-Group

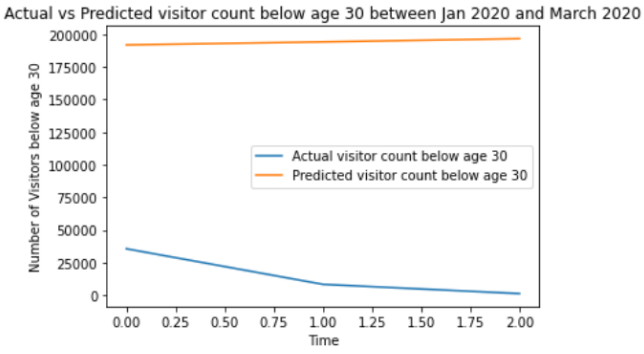
The PROPHET forecasting plot on Age-Group of visitors is shown in the figures below.



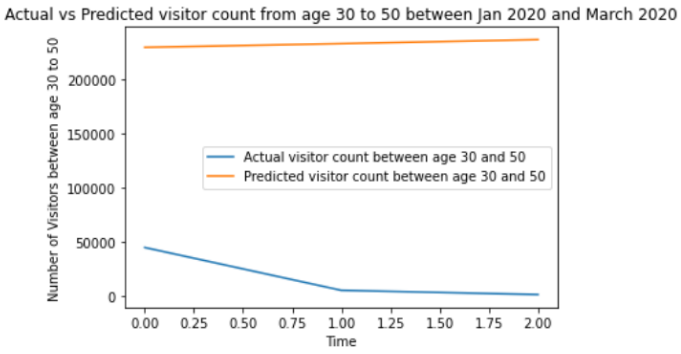

Forecasted visitor count between age 30 and 50
Number of Visitors
260000
240000
220000
200000
180000
160000
2019-01
2019-03
2019-05
2019-07
2019-09
2019-11
2020-01
2020-03
Date (Months)
Tourist Age-group: 30 to 50


Forecasted visitor count above age 50
Number of Visitors
150000
140000
130000
120000
110000
100000
90000
80000
70000
2019-01
2019-03
2019-05
2019-07
2019-09
2019-11
2020-01
2020-03
Date (Months)
Tourist Age-group: Above 50

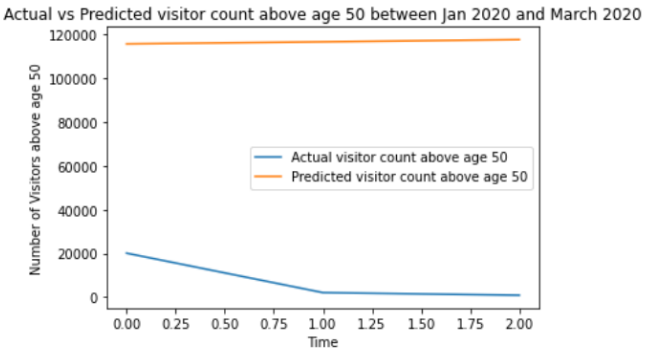
We compared the actual verses predicted per-month cumulated Chinese visitor data based on their Age Groups for the months of January 2020 – March 2020 in the following figures.



Mean Absolute Error: 179187.768



Mean Absolute Error: 215109.925



Mean Absolute Error: 109028.874

As we can observe from the plots:

- The actual visitor count below age 30 from China is extremely low compared to the forecasted visitor count below age 30, with the Mean Absolute Error (MAE) = 179187.768
- The actual visitor count between age 30 and 50 from China is extremely low compared to the forecasted visitor count between age 30 and 50, with the Mean Absolute Error (MAE) = 215109.925
- The actual visitor count above age 50 from China is extremely low compared to the forecasted visitor count above age 50, with the Mean Absolute Error (MAE) = 109028.874

6.3 Comparison between ARIMA and PROPHET Forecasting Models

We have compared the Mean Absolute Errors (MAE) of all the forecasting models done by ARIMA and PROPHET based on Purpose of Visit and Tourist Age–Groups in the following table.

	Mean Absolute Error (MAE) of Time-Series Models				
Time-Series Forecasting Method	Purpose of Visit: Tourism	Purpose of Visit: Non_Tourism	Age–Group: Below 30 yrs	Age–Group: 30–50 yrs	Age–Group: Above 50 yrs
ARIMA	349638.667	86906.667	142245.667	179405.333	104047.333
PROPHET	454415.307	61045.460	179187.768	215109.925	109028.874

As we know, ARIMA assumes that there is a sort of causal relationship between past values and past errors and future values of the time series.

But Prophet does not look for any such causal relationships between past and future. Instead, it simply tries to find the best curve to fit to the data, using a linear or logistic curve and Fourier coefficients for the seasonal components and is recommended only for time series where the only informative signals are trend and seasonality, and the residuals are just noise.

As we have seen, our data consists of monthly cumulative visitor records for 16 months, and hence there is no seasonality, trend and noise in our data. Hence, Prophet was not able to forecast the future well, as we can clearly see from the forecasting plots of the PROPHET model.

ARIMA on the other hand works well even for non-seasonal data with adjusted parameter (p, q, d) values. Also, we have chosen lowest AIC score to identify the best combination of p, q, d values and trained the data with the best ARIMA model possible. ARIMA performed well on this data to learn the trend of the past data and predict accordingly. Hence, the forecasting plots of the ARIMA model looks more promising than those of the PROPHET model.

As we can see from the Mean Absolute Error Table above, the MAE values of each five models forecasted by PROPHET are higher than those by ARIMA, further proving that ARIMA worked better than PROPHET for our South Korean tourism dataset.

The MAE for ARIMA and PROPHET models are both abnormally high, which indicates that the forecasted data for 2020 does not match the actual visitor record of 2020. This is because South Korean tourism was heavily affected by the recent COVID19 pandemic which first broke in China in December 2019 and tourism industry got hit by the blow in the following months of 2020.

7. CHALLENGES AND FUTURE WORK

The main challenge in this project was that the dataset is too small for doing a robust time-series analysis. With only 16 data points applying a forecasting model was hard as there is no trend, seasonality, and noise in the data. PROPHET is strong forecasting method but heavily depends on seasonality and hence was unable to do a correct prediction for our data.

For future work, it will be interesting to work on a dataset consisting of official South Korean tourism

records with daily tourist count data. Then we can look at the seasonality, trend and noise of that time series data and create a robust forecasting model with ARIMA and PROPHET and compare their prediction results.

8. CONCLUSION

From January 2020 there is sharp drop in number of visitors from China to South Korea. This is because South Korean tourism was heavily affected by the recent COVID19 pandemic which first broke in China in December 2019 and tourism industry got hit by the blow in the following months of 2020.

The actual number of Chinese visitors who travelled to South Korea for vacation is absurdly low compared to the forecasted number of visitors in 2020.

Mainly people travelled to South Korea for Non-Tourism purpose like Studies, Business, Official Affairs, etc. and people travelling for the purpose of Tourism was very low when compared to 2019 data and 2020 visitor forecast.

We can back this finding from the large MAE error that we saw for ARIMA and PROPHET forecasting models. as well as from our detailed exploratory data analysis.

9. MEMBER CONTRIBUTION

- **Sujata Gorai** – Data Exploration and Data Pre-Processing
- **Alokparna Bandyopadhyay** – Time Series forecasting using ARIMA and Prophet models and result comparison