MISM 6212 GROUP PROJECT
Tourism Industry in China Under the Situation Of COVID

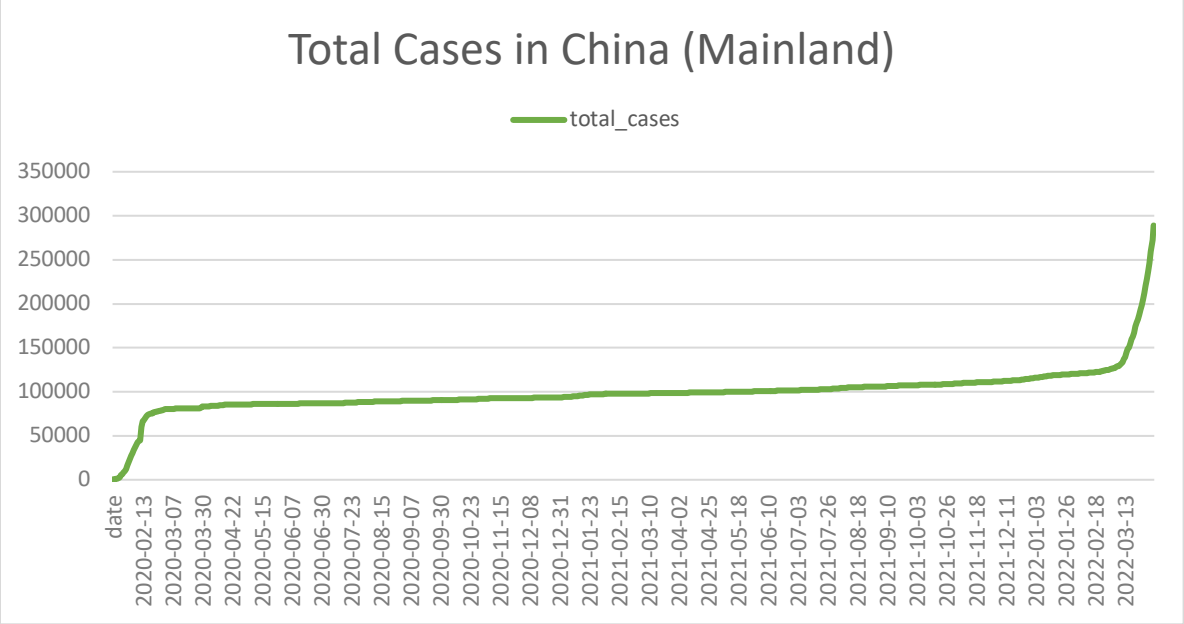
Group 10

Chuwen Feng
Changjiang Song
Zijing Bai
Yanlin Wu
Hongyi Hu

04/29/2022

Part1: Introduction

Since the large-scale outbreak of the coronavirus in January 2020, the number of people infected with the coronavirus has been larger and the spread is wider, and the impact is significantly greater than that of SARS: in two months, there have been more than 70,000 cumulative confirmed cases and 2,000 new cases. More than 5,000 cases of infection and more than 300 deaths were affected by SARS in 2003. In terms of the scope of transmission, the epidemic spreads faster and wider. The picture shows the number of confirmed cases in mainland China from the beginning of 2020 to the present.

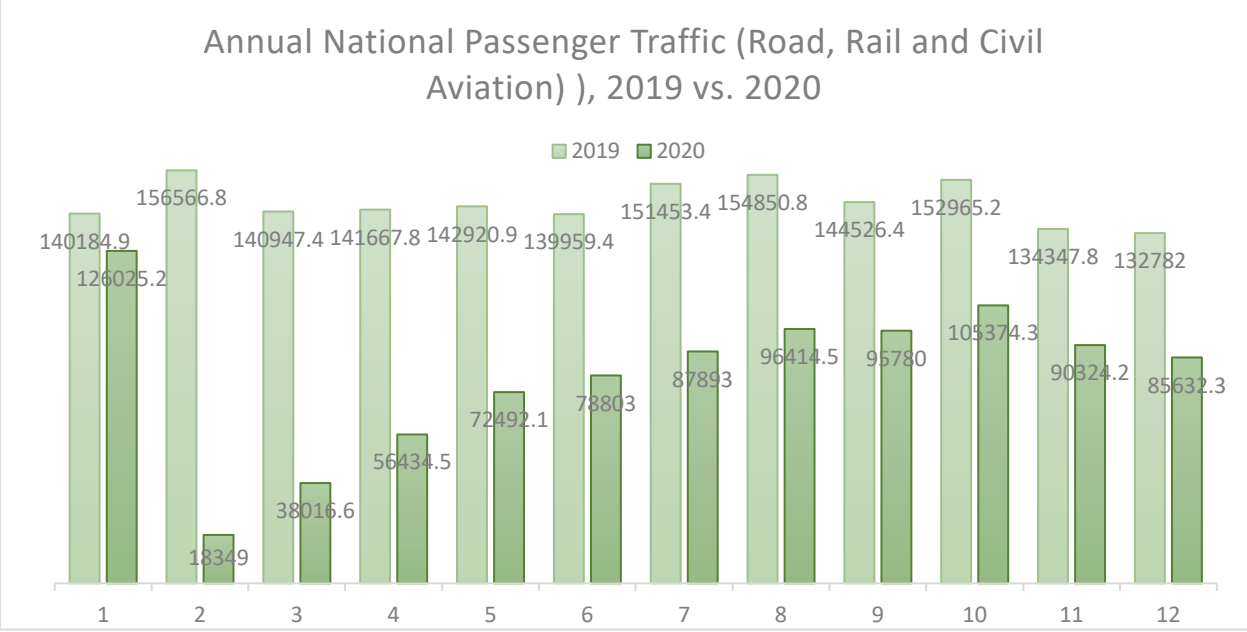


SOURCE: Our World in Data

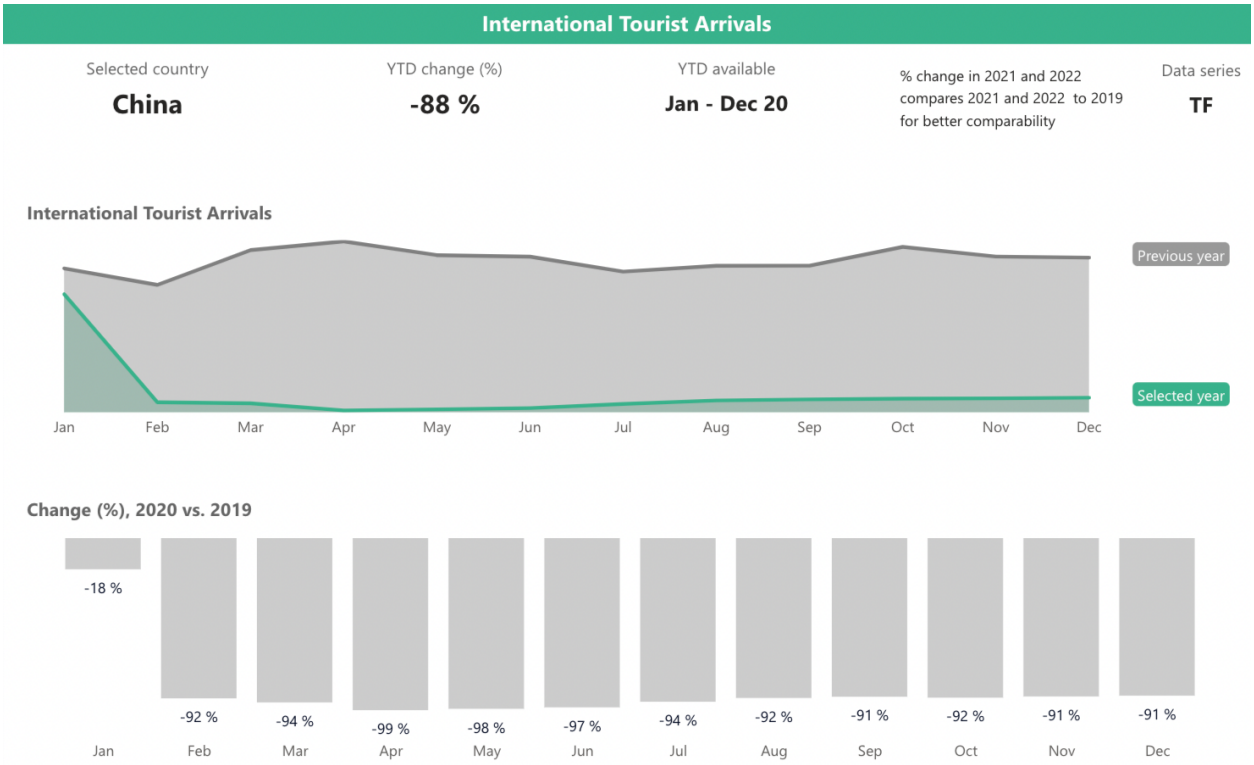
The outbreak of the coronavirus had a huge impact on the tourism industry in China and the world. The most significant impact was the change in the number of tourists. The figure below reflects the change in the number of tourists in mainland China from 1900 to 2021.

SOURCE: Ministry of Culture and Tourism

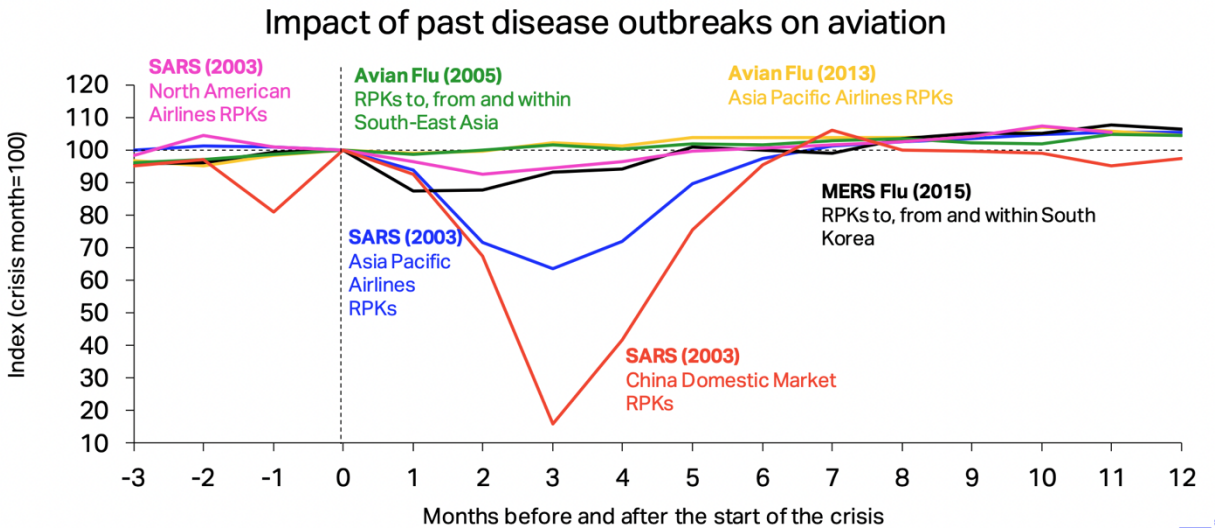
Worries about contracting pneumonia and measures to control population movements, according to the Ministry of Transport, dropped 82.3% in January compared with the same period in 2019. According to the forecast of the China Tourism Academy, domestic tourism revenue in the first quarter will grow by 69%, with a decrease of 1.18 trillion yuan. Only 44% of the civil aviation passenger load factor and 14% of the hotel occupancy rate make the upstream and downstream of the tourism industry chain experience huge operating pressure. In addition, the large number of refunds and changes of orders also brought significant cash flow challenges. From January 15th to January 31st, the total amount of free refunds for domestic and foreign airlines exceeded 20 billion yuan. Under the dual pressure of huge loss of income and cash flow, it can now be regarded as a major challenge moment for China's tourism and transportation industry.



SOURCE: Ministry of Transport



Previous disease outbreaks have peaked after 1-3 months and recovered pre-outbreak levels in 6-7 months. There are reasons to believe that the impact of the epidemic on the tourism and transportation industry is still short-term. The transportation industry will continue to grow in the long term and is expected to recover in a relatively short period of time.



SOURCE: IATA Economics

People and businesses around the world were hurting — including theme parks, which were shut down for months as countries tried to get the COVID-19 pandemic under control. In 2020, theme parks, amusement parks and water parks around the world are struggling to survive. Due to the impact of the spread of the coronavirus, they had to suspend or even close.

Now, a new Global Attractions Attendance report by the Themed Entertainment Association and AECOM is showing just how much attendance dropped at theme parks in 2020. The highest drops were seen at Universal Studios Hollywood and Disneyland Resort — which makes sense, considering that theme parks in California had to stay closed for over a year amid state restrictions that were in place amid the pandemic. Both Universal Hollywood and Disneyland Resort saw attendance declines of just over 80%.

https://insidethemagic.net/2021/09/2020-theme-park-attendance-report-ks1/#:~:text=Overall%2C%20theme%20parks%20across%20the,approximately%2083%20million%20in%202020.

2019 Rank	Park location	Change	Attendance 2020	Attendance 2019
1	Magic Kingdom Theme Park at Walt Disney World Resort, LAKE BUENA VISTA, FL, U.S.	-66.90%	6,941,000	20,963,000
2	Disneyland Park at Disneyland resort, ANAHEIM, CA, U.S.	-80.30%	3,674,000	18,666,000
3	Tokyo Disneyland at Tokyo Disney Resort, TOKYO, JAPAN	-76.80%	4,160,000	17,910,000
4	Tokyo Disney sea at Tokyo Disney Resort, TOKYO, JAPAN	-76.80%	3,400,000	14,650,000
5	Universal Studios Japan, OSAKA, JAPAN	-66.20%	4,901,000	14,500,000
6	Disney's Animal Kingdom at Walt Disney world, LAKE BUENA VISTA, FL, U.S.	-70.00%	4,166,000	13,888,000
7	Epcot at Walt Disney world, LAKE BUENA VISTA, FL, U.S.	-67.50%	4,044,000	12,444,000

8	Chime long Ocean Kingdom, HENGQIN, CHINA	-59.10%	4,797,000	11,736,000
9	Disney's Hollywood Studios at Walt Disney World, LAKE BUENA VISTA, FL, U.S.	-68.00%	3,675,000	11,483,000
10	Shanghai Disneyland at Shanghai Disney Resort, SHANGHAI, CHINA	-50.90%	5,500,000	11,210,000
11	Universal Studios Florida at Universal Orlando, FL, U.S.	-62.50%	4,096,000	10,922,000
12	Universal's Island of Adventure at Universal Orlando, FL, U.S.	-61.40%	4,005,000	10,375,000

Credit: TEA/AECOM

Shanghai Disney, which was also closed for a couple of months and noticed an attendance drop of just over 50%.

Overall, **theme parks across the world saw an average attendance decline of about 67%.** According to the data, over 250 million people visited theme parks in 2019, and that

dropped to approximately 83 million in 2020. This makes sense as theme parks were closed for an extended period, and even when theme parks reopened, attendance was significantly limited for quite a while.
Fortunately, many theme parks are beginning to see crowds return through their gates once again. Many theme parks have reopened with a number of health and safety procedures in place — including mask mandates and attendance caps — to keep Guests and employees safe. Disney Parks alone have seen a huge jump in numbers at both its Florida and Southern California theme parks in 2021.

With the gradual normalization of the epidemic, the tourism industry in mainland China has gradually recovered. In 2020, the number of domestic tourists was 2.879 billion, down 52.1% from the same period of the previous year. Domestic tourism revenue was 2.23 trillion yuan, a year-on-year decrease of 61.1%. Due to proper management and control, although the proportion of overall tourist arrivals will decline in 2020, the tourism industry will gradually pick up as time goes on. From the Spring Festival, Qingming Festival to Labor Day, tourism consumption has become increasingly active, urban leisure day trips, and suburban and surrounding tours have recovered. Most tourists understand the measures taken under the normalization of epidemic prevention and control and are satisfied with the services provided by market players. of. According to the monitoring and comparison of holiday data, during the seven-day Spring Festival holiday (excluding the extended three days), the country received a total of 248 million domestic tourists, and achieved domestic tourism revenue of 278.1 billion yuan, a year-on-year decrease of 40.3% and 45.9% respectively. During the Qingming holiday in 2020, the total number of domestic tourists received was 43.254 million, a year-on-year decrease of 61.4%; tourism revenue was 8.26 billion yuan, a year-on-year decrease of 80.7%. During the five-day Labor Day holiday, the country received a total of 115 million domestic tourists and a total tourism revenue of 47.56 billion yuan, which was 53.5% and 36.7% higher than that of Labor Day in 2019, respectively. During the Qingming Festival, the average radius of urban residents' activities was 3.6 kilometers, an average increase of 36.8% compared with the Spring Festival period; the recreational radius of non-local tourists at the destination was 12.9 kilometers, an average increase of 16.0% compared with the Spring Festival holiday. During the May Day period, the average travel time of tourists has exceeded 40 hours, and the travel distance is 136 kilometers, of which the average travel distance of local tourists is 40.5 kilometers. The average leisure time of tourist destinations is 16.7 kilometers, an increase of 50% compared with the average value of the Spring Festival holiday in 2020. Tourist satisfaction reached 84.8 points, and the proportion of tourists who chose to travel by car reached a record high of 64.1%.

The rapid recovery of China's tourism industry is largely due to China's special isolation and epidemic prevention policy. Because of this policy, China can block the continued spread of the virus in a short period of time. Therefore, all tourism in 2021 will also be in a state of gradual recovery. Having a health code escorts people's travel. At the same time, people have

unknowingly become accustomed to wearing masks to go out, which not only prevents the spread of Covid-19 but also prevents the spread of some epidemic diseases. People are no longer afraid of the epidemic, and they have begun to face the epidemic with a positive, healthy and optimistic attitude.

The global epidemic has gradually improved, and the tourism economy has gradually recovered. People no longer tend to gather indoors for activities such as watching movies, singing, etc. More inclined to outdoor activities: such as playing basketball, picnic in the park, etc. Because the open space is easier to circulate the air, and it is less likely to spread the epidemic. However, tourist attractions are extensive and derivative, so when we investigate people's travel preferences, we also focus on the choice of destinations. To be more specific, we have the following three questions:

1. If we can add covid as a dimension to predict people's choice of travel destination?

By using logistic regression, we can interpret the result of classification in order to predict the consumers' travel willingness for the next 6 months.

2. If the segmentation of the tourists changes slightly due to the impact of the epidemic?

In this part, by using k-prototypes we can perform cluster analysis on tourists in the survey.

3. Under the epidemic, are the representative tourist destinations in people's minds different?

The analysis of tourist attractions selection under the epidemic will be explained through the dimensionality reduction method of Principal Component Analysis (PCA).

The answers to these questions are important because they can help travel companies better design products and better cater to their customers. Based on these questions above, the travel companies are able to understand the travel preferences of different types of tourists after the epidemic.

Part2: Data

1. Data sources:

1.1. Because the goal of our group project is to analyze the change of travel consumers' propensity to travel after the epidemic, it is very important for us to collect the latest information from tourist. So, we designed a questionnaire for the data we needed. All responses to the questions in the questionnaire will be converted into data by **QUALTRICS**

1.2. The question of the questionnaire needs to be designed through three different machine learning models toward different target, which are classification, clustering, dimensionality

reduction.

2. Features:

2.1. The questionnaire questions in the first part are built on the models of Classification. The relevant discriminant data is needed to be collected. The So-called discriminant features can reflect people's characteristics that tell us how to find, reach, identify segments (age, income, education, profession, lifestyles, media habits, use occasions; industry, size, location, organizational structure). All discriminant features in the data set include 'SEX', 'AGE', 'PROFESSION', 'HHINCOME, 'VACCINE', 'EDUCATION', 'TRANSPORTATION','DISTANCE','ANXIETY','ACCOMMODATION','TRAVEL_DURATION', 'BUDGET', 'COMPANION', 'NECESSITY', and 'Questionnaire satisfaction'. Because our entire questionnaire is designed with a total of 26 different questions. of satisfaction, we assume that if people do not agree with our questionnaire or do not take it in seriously, then their data will not be referred. Thus, we designed a One-Vote Veto Question called "Questionaries satisfaction". All the information of people who choose "bad" and "good" options under this question will not be used in the following machine learning model. (See Appendix2.1 for the explanation of each columns/feature).

2.2. The questionnaire questions in the second part are designed from the perspective of applying dimensionality reduction for unsupervised study. Based on each question, people need to prioritize 5 locations, including Natural, Cultural attractions, historical sites, Theme Parks, and Business center(mall). Because the final output format of the questionnaire is wide format, each question that needs to be sorted by location will be divided into 5 features and displayed separately. For example, Q21_1 means "Suppose that the motivation" for traveling is to socialize, where would you rank Nature attractions among all the following attractions? Q21_2 means "Suppose that the motivation for traveling is to socialize, where would you rank Cultural attractions among all the following attractions? Because we have a total of 9 questions that requires people to rank tourist destination, so a total of 45 related features are created in the case of outputting data in wide format. Moreover, the above features are classified as segmentation features. The so-called segmentation features are characteristics that tell us why segments differ (e.g., needs want, benefits, solutions to problems, usage situation, past behavior.) At last, here are all 8 questions. (See appendix2.2 for the explanation of each columns/feature).

1. Suppose that the motivation for traveling is to relax, which of the following travel destinations is your favorite? (Rate from 1 (preferred) to 5 (not preferred)):

a. Nature b. Cultural c. Architecture d. Theme Parks e. Business Center

2. Suppose that the motivation for traveling is to release, which of the following travel destinations is your favorite? (Rate from 1 (preferred) to 5 (not preferred)):

a. Nature b. Cultural c. Architecture d. Theme Parks e. Business Center

3. Suppose that the motivation for traveling is to socialize, which of the following travel destinations is your favorite? (Rate from 1 (preferred) to 5 (not preferred)):

a. Nature b. Cultural c. Architecture d. Theme Parks e. Business Center

4. Suppose that the motivation for traveling is to learn, which of the following travel destinations is your favorite? (Rate from 1 (preferred) to 5 (not preferred)):

a. Nature b. Cultural c. Architecture d. Theme Parks e. Business Center

5. Suppose that the motivation for traveling is to entertain, which of the following travel destinations is your favorite? (Rate from 1 (preferred) to 5 (not preferred)):

a. Nature b. Cultural c. Architecture d. Theme Parks e. Business Center

6. In case you are in a low-budgeting situation, which of the following travel destination is your favorite? (Rate from 1 (preferred) to 5 (not preferred)):

a. Nature b. Cultural c. Architecture d. Theme Parks e. Business Center

7. In case you are in a high-budgeting situation, which of the following travel destination is your favorite? (Rate from 1 (preferred) to 5 (not preferred)):

a. Nature b. Cultural c. Architecture d. Theme Parks e. Business Center

8. Which of the following travel destinations is physically more away from you? (Rate from 1 (Closest) to 5 (furthest)):

a. Nature b. Cultural c. Architecture d. Theme Parks e. Business Center

9. Thinking of the potential risk of COVID, which of the following travel destination is your favorite? (Rate from 1 (preferred) to 5 (not preferred)):

a. Nature b. Cultural c. Architecture d. Theme Parks e. Business Center

2.3. The questions for an unsupervised learning model of clustering will not be designed because all the data extract from above-mentioned questions are enough for clustering machine learning. The data of clustering will be mainly based on discriminant data introduced in Section 2.1.

2.4. After extracting data from the above questionnaire, a raw data set (255,60) which consisting of 255 respondents and 60 features generated finally.

3. Data Cleansing, Profiling, Transformation Process.

3.1. Pandas in Python will be used to complete the following Data Cleansing, Profiling, and Transformation process. Because the project is targeted on Chinese tourists, all the questions and options in the questionnaire will be presented in Chinese. This will cause all the data information we get to be Chinese version. Therefore, the first task in data profiling is to convert all data and features into English version. Secondly, because we designed all the answers to assign with the meaning of ranking, in the process of profiling data, almost all the data can be converted to numbers 1-5 except the features such 'SEX', 'PROFESSON', 'TRANSPORTATION', 'Questionnaire Satisfaction'. For example, if the variables in the columns of “Anxiety” is ‘insensible’, it will be converted to the number 1, and all variables with an anxiety level of ' very anxious ' will be converted to the number 5. (For a detailed explanation of each variable, please refer to Appendix). SEX will be processed as a category variable which male is equal to 1, and female is equal to 0. For the feature of ‘PROFESSONS’, because there are a large number of unique values, we classify all professions into ten representative profession categories. The ten categories include: Business, Education, Medical, Engineering, Agriculture, Technology, Unemployed, Self-operated business, Government, and Students. Moreover, all the segmentation data that will be used in the dimensionality reduction section also needs to be modified. It is

necessary to know the most preferable travel destination will be ranked as 1 and outputted with the number 1 and the least preferred destination will be outputted as the number 5 in the data set. However, for building a perceptual map based on Principal Component Analysis, the data requires to show that the higher the number, more preferred for the destination. Thus, all the numbers 1 and 2 under all features from Q16_5 to Q24_5 will be replaced with numbers 5, and 4.

3.2. Firstly, all 60 features have Missing value in the dataset. Most of the missing values appear in features Q16_1 to features Q24_5, which means that some respondents did not answer all segmentation questions completely. For instance, if some respondents did not answer the question of Q19, then all variables in Q19_1 to Q19_5 columns corresponding to the respondent is null. However, it is understood by some respondents that they think the orders of the options which are displayed by default is what they have in mind about the order for this destination. In this case, they didn't rank these options position. Because the answer has not changed, QUALTRICS will default that the answer is not filling by respondents, which will make it outputs null values. In the result, in order to remaining more data, all the rows, which missing value is less than or equal to 10 in all Q16_1 to Q24_5 variables, are filled out with numbers.

For example, All the missing values in this case is replaced with the numbers 5 to 1 to indicate that they are sorted in the order of Nature (5), Cultural (4), Architecture (3), Theme Parks (2), and Business Center (1) for each question.

Furthermore, from the perspective of discriminant data, we believe that in the project, the miss value of discriminate features should be profiled as little as possible. The main reason is that the data set we collected has a small volume. If the missing value of discriminant features is filled, the final model output will deviate greatly from reality.

In this end, all the missing value under the discriminant features is deleted.

Last but not the least, the response of One-Vote Veto Question which is “Satisfaction Question” should be taken into account for data cleansing. All rows with a value “Bad” under the column “Satisfaction Question” is deleted.

3.3. To apply the data to the model for classification and clustering, and perceptual map (PCA), the data types of the following features will be unified.

'SEX','PROFESSION','VACCINE','TRANSPORTATION','EDUCATION','COMPANION','Questionnaire_satisfaction','Q25', and 'Q26' will be converted to category data type.

'HHINCOME', 'DISTANCE', 'ANXIETY', 'ACCOMMODATION', 'TRAVEL_DURATION', 'BUDGET', 'NECESSITY' will be converted to Int64 data type.

All the segmentation data, which are 'Q16_1', 'Q16_2', 'Q16_3', 'Q16_4', 'Q16_5', 'Q17_1', 'Q17_2', 'Q17_3','Q17_4', 'Q17_5', 'Q18_1', 'Q18_2', 'Q18_3', 'Q18_4', 'Q18_5', 'Q19_1','Q19_2', 'Q19_3', 'Q19_4', 'Q19_5', 'Q20_1', 'Q20_2', 'Q20_3', 'Q20_4','Q20_5', 'Q21_1', 'Q21_2', 'Q21_3', 'Q21_4', 'Q21_5', 'Q22_1', 'Q22_2','Q22_3', 'Q22_4', 'Q22_5',

'Q23_1', 'Q23_2', 'Q23_3', 'Q23_4', 'Q23_5', 'Q24_1', 'Q24_2', 'Q24_3', 'Q24_4', and 'Q24_5', will be converted to int64 data type

3.4. After the entire Data Cleansing, Profiling, and Transformation Process is completed, our final available data has 144 rows and 60 columns.

4. APPENDIX

1. Tool for outputting the data from Questionary is QUALTRICS https://qualtrics.northeastern.edu/
2. Columns with Description from the Data Set (Sources: The Self-Made Questionary)

2.1 Discriminant Data:

Columns	Description
SEX Male (Dummy Variables)	The Gender of each respondent 1 = Male, 0 = Female
AGE	The Age of each respondent 1 = Age range is 18-34 2 = Age range is 35-47 3 = Age range is 48-60 4 = Age range is 61-74 5 = Age range is 75-88
PROFESSION	The profession of each respondent Business,Education,Medical, Engineering, Agriculture, Technology, Unemployed, self-operated,business, Government, Students.
HHINCOME	Household Revenue of each respondent 1 = $1500-$15000 2 = $15000-$30000 3 = $30000-$80000 4 = $80000-$150000 5 = Above$150000
VACCINE YES	Whether the respondents vaccinated 1 = Yes 0 = No
EDUCATION	Education level of each respondent 1= Master' 2= BACHELOR 3= 3YCollege 4 = VSC' 5 = HighSchool'
TRANSPORTATION'	This is the Chinese version. This column will not be used in this project.

2.2 Segmentation Variables

DISTANCE	What kind of social distancing do you think should be maintained from others when traveling: 1 = insensible 2 = 1_meters 3 = 2_meters' 4 = 5_meters 5 =No_Contact
ANXIETY	Respondents' level of anxiety about the coronavirus 1 = insensible 2 = somewhat insensible 3 = median 4 = somewhat anxious 5 = very anxious
ACCOMMODATION	Do you think it is better to live in a hotel when you are traveling? 1 = not willing to live in hotel 2 = Tends to be a formal hotel 3 = It doesn't matter where I live when I am traveling 4 = If I can stay in a hotel, I won't choose other options' 5 = I must live in a hotel
TRAVEL_DURATION	What is the best travel duration for you 1 = under 3-days 2 = 3-5 days 3 = 5-7 days 4 = 7-14 days 5 = over 14-days
BUDGET	What do you think is the most suitable budget for traveling abroad? 1 = under 300$ 2 = 300-1000 3 = 1000-1500 4 = 1500-3000 5 = over 3000$
COMPANION	How much people do you think to travel together? 2 = 2 people 3-5 = 3-5 people 5-10 = 5-10 people 1 = Travel alone
NECESSITY	Do you think it is an indispensable part of life for traveling?

	The higher the number, the more willing people are to travel 1 = level 1 2 = level 2 3 = level 3 4 = level 4 5 = level 5
Questionnaire_satisaction	In order to ensure the reliability of the questionnaire, please select "very good" for this question. “Very Good” “Good” “Bad”
Q25 Yes	Have you had any itinerary in the past six months that you consider to be tourism (Tourism is considered to be a trip that requires an overnight stay in a non-resident place or a certain distance from your usual place of residence) 1 = Yes 0 = no
Q26 Yes	Do you have travel plans for the next six months? 1 = Yes 0 = no

2.3 Perspectives of Travel Destination (Designed for PCA positioning)

Q16_1	Thinking of the potential risk of COVID, which of the following travel destination is your favorite? **(Nature)** (Rate from 1 (preferred) to 5 (not preferred)):
Q16_2	Thinking of the potential risk of COVID, which of the following travel destination is your favorite? **(Cultural)** (Rate from 1 (preferred) to 5 (not preferred)):
Q16_3	Thinking of the potential risk of COVID, which of the following travel destination is your favorite? **(Architecture)** (Rate from 1 (preferred) to 5 (not preferred)):

Q16_4	Thinking of the potential risk of COVID, which of the following travel destination is your favorite? **(Theme Parks)** (Rate from 1 (preferred) to 5 (not preferred)):
Q16_5	Thinking of the potential risk of COVID, which of the following travel destination is your favorite? **(Business Center)** (Rate from 1 (preferred) to 5 (not preferred)):
Q17_1	Which of the following travel destinations is physically more away from you? **(Nature)** (Rate from 1 (Closest) to 5 (furthest)):
Q17_2	Which of the following travel destinations is physically more away from you? **(Cultural)** (Rate from 1 (Closest) to 5 (furthest)):
Q17_3	Which of the following travel destinations is physically more away from you? **(Architecture)** (Rate from 1 (Closest) to 5 (furthest)):
Q17_4	Which of the following travel destinations is physically more away from you? **(Theme Parks)** (Rate from 1 (Closest) to 5 (furthest)):
Q17_5	Which of the following travel destinations is physically more away from you? **(Business Center)** (Rate from 1 (Closest) to 5 (furthest)):
Q18_1	Suppose that the motivation for traveling is to learn, which of the following travel destinations is your favorite? **(Nature)** (Rate from 1 (preferred) to 5 (not preferred)):
Q18_2	Suppose that the motivation for traveling is to learn, which of the following travel destinations is your favorite? **(Cultural)** (Rate from 1 (preferred) to 5 (not preferred)):
Q18_3	Suppose that the motivation for traveling is to learn, which of the following travel destinations is your favorite? **(Architecture)** (Rate from 1 (preferred) to 5 (not preferred)):

Q18_4	Suppose that the motivation for traveling is to learn, which of the following travel destinations is your favorite? **(Theme Parks)** (Rate from 1 (preferred) to 5 (not preferred)):
Q18_5	Suppose that the motivation for traveling is to learn, which of the following travel destinations is your favorite? **(Business Center)** (Rate from 1 (preferred) to 5 (not preferred)):
Q19_1	Suppose that the motivation for traveling is to entertain, which of the following travel destinations is your favorite? **(Nature)** (Rate from 1 (preferred) to 5 (not preferred)):
Q19_2	Suppose that the motivation for traveling is to entertain, which of the following travel destinations is your favorite? **(Cultural)** (Rate from 1 (preferred) to 5 (not preferred)):
Q19_3	Suppose that the motivation for traveling is to entertain, which of the following travel destinations is your favorite? **(Architecture)** (Rate from 1 (preferred) to 5 (not preferred)):
Q19_4	Suppose that the motivation for traveling is to entertain, which of the following travel destinations is your favorite? **(Theme Parks)** (Rate from 1 (preferred) to 5 (not preferred)):
Q19_5	Suppose that the motivation for traveling is to entertain, which of the following travel destinations is your favorite? **(Business Center)** (Rate from 1 (preferred) to 5 (not preferred)):
Q20_1	Suppose that the motivation for traveling is to relax, which of the following travel destinations is your favorite? **(Nature)** (Rate from 1 (preferred) to 5 (not preferred)):

Q20_2	Suppose that the motivation for traveling is to relax, which of the following travel destinations is your favorite? **(Cultural)** (Rate from 1 (preferred) to 5 (not preferred)):
Q20_3	Suppose that the motivation for traveling is to relax, which of the following travel destinations is your favorite?**(Architecture)** (Rate from 1 (preferred) to 5 (not preferred)):
Q20_4	Suppose that the motivation for traveling is to relax, which of the following travel destinations is your favorite? **(Theme Parks)** (Rate from 1 (preferred) to 5 (not preferred)):
Q20_5	Suppose that the motivation for traveling is to relax, which of the following travel destinations is your favorite? **(Business Center)** (Rate from 1 (preferred) to 5 (not preferred)):
Q21_1	Suppose that the motivation for traveling is to socialize, which of the following travel destinations is your favorite? **(Nature)** (Rate from 1 (preferred) to 5 (not preferred)):
Q21_2	Suppose that the motivation for traveling is to socialize, which of the following travel destinations is your favorite? **(Cultural)** (Rate from 1 (preferred) to 5 (not preferred)):
Q21_3	Suppose that the motivation for traveling is to socialize, which of the following travel destinations is your favorite?**(Architecture)** (Rate from 1 (preferred) to 5 (not preferred)):
Q21_4	Suppose that the motivation for traveling is to socialize, which of the following travel destinations is your favorite? **(Theme Parks)** (Rate from 1 (preferred) to 5 (not preferred)):
Q21_5	Suppose that the motivation for traveling is to socialize, which of the following travel destinations is your favorite? **(Business**

	Center) (Rate from 1 (preferred) to 5 (not preferred)):
Q22_1	Suppose that the motivation for traveling is to release, which of the following travel destinations is your favorite? **(Nature)** (Rate from 1 (preferred) to 5 (not preferred)):
Q22_2	Suppose that the motivation for traveling is to release, which of the following travel destinations is your favorite? **(Cultural)** (Rate from 1 (preferred) to 5 (not preferred)):
Q22_3	Suppose that the motivation for traveling is to release, which of the following travel destinations is your favorite?**(Architecture)** (Rate from 1 (preferred) to 5 (not preferred)):
Q22_4	Suppose that the motivation for traveling is to release, which of the following travel destinations is your favorite? **(Theme Parks)** (Rate from 1 (preferred) to 5 (not preferred)):
Q22_5	Suppose that the motivation for traveling is to release, which of the following travel destinations is your favorite? **(Business Center)** (Rate from 1 (preferred) to 5 (not preferred)):
Q23_1	In case you are in a low-budgeting situation, which of the following travel destination is your favorite? **(Nature)** (Rate from 1 (preferred) to 5 (not preferred)):
Q23_2	In case you are in a low-budgeting situation, which of the following travel destination is your favorite? **(Cultural)** (Rate from 1 (preferred) to 5 (not preferred)):
Q23_3	In case you are in a low-budgeting situation, which of the following travel destination is your favorite? **(Architecture)** (Rate from 1 (preferred) to 5 (not preferred)):

Q23_4	In case you are in a low-budgeting situation, which of the following travel destination is your favorite? **(Theme Parks)** (Rate from 1 (preferred) to 5 (not preferred)):
Q23_5	In case you are in a low-budgeting situation, which of the following travel destination is your favorite? **(Business Center)** (Rate from 1 (preferred) to 5 (not preferred)):
Q24_1	In case you are in a high-budgeting situation, which of the following travel destination is your favorite? **(Nature)** (Rate from 1 (preferred) to 5 (not preferred)):
Q24_2	In case you are in a high-budgeting situation, which of the following travel destination is your favorite? **(Cultural)** (Rate from 1 (preferred) to 5 (not preferred)):
Q24_3	In case you are in a high-budgeting situation, which of the following travel destination is your favorite? **(Architecture)** (Rate from 1 (preferred) to 5 (not preferred)):
Q24_4	In case you are in a high-budgeting situation, which of the following travel destination is your favorite? **(Theme Parks)** Rate from 1 (preferred) to 5 (not preferred)):
Q24_5	In case you are in a high-budgeting situation, which of the following travel destination is your favorite? **(Business Center)** (Rate from 1 (preferred) to 5 (not preferred)):

Part 3: Analysis

1. Supervised Learning: Classification to classify whether a tourist have travel plans for the next six months.

1.1 Data Profiling and Cleansing for the classification model

- The columns of 'TRANSPORTATION', 'Unnamed: 0','PROFESSION' are deleted

- Dummy Variables such as 'SEX_Male', 'VACCINE_Yes', 'COMPANION_2', 'COMPANION_3-5', 'COMPANION_5-10', 'Q25_Yes' has been transformed

1.2 Test/Train Variables:

- Y-variable: Q_26
- X-variable: 'AGE', 'HHINCOME', 'EDUCATION', 'DISTANCE', 'ANXIETY', 'ACCOMMODATION', 'TRAVEL_DURATION', 'BUDGET', 'NECESSITY', 'SEX_Male', 'VACCINE_Yes', 'COMPANION_2', 'COMPANION_3-5', 'COMPANION_5-10', 'Q25_Yes'.
- 30 percent of x dataset will be set as x-test data set. 30 percent of y dataset will be set as y-test data set. The random state equals to 1

For building a classification model with the best performance to classify whether a tourist have travel plans for the next six months, firstly our group built 10 different classification models. Based on the accuracy score, f1 score, overfitting issues, and the interpretability of the model, we will choose the best one among all of classification models.

Model	F1_score	Accuracy score
Support Vector Machines	**0.814815**	**0.772727**
K-nearest neighbors	**0.830189**	**0.795455**
Logistic Regression	**0.830189**	**0.795455**
Logistic Regression SFS backward	**0.840000**	**0.818182**
Logistic Regression SFS forward	**0.807692**	**0.772727**
Logistic Regression EFS	**0.807692**	**0.772727**
Logistic Regression SelectKBest	**0.830189**	**0.795455**
Random Forest	**0.857143**	**0.818182**
Naive Bayes	**0.814815**	**0.750000**
Stochastic Gradient Decent	**0.814815**	**0.772727**
Linear SVC	**0.814815**	**0.795455**
Voting	**0.780031**	**0.744286**
Decision Tree	**0.761905**	**0.659091**

1.3 Classification Models summary and selection:

Totally we used four features' selections method on Logistic regression.

Based on the F1_score and Accuracy score of each logistic regression model with different feature selection method, we found that the logistic regression model based on SequentialFeatureSelector in Backward performed best among these four models.

The following model will use grid search method to find the best parameters.

- For Support Vector Machines, we used grid search method to find its hyperparameters, which C=1 Gamma = 1, Kernel = 'linear.
- For Decision Tree model, based on grid search method we found the best parameters are that max_depth=3 , min_samples_split= 2, random_state=1
- For random forest, based on grid search method we found that the best parameters are that random_state =1, n_estimators = 200, max_depth= 1, min_samples_split= 8.
- For K-nearest neighbors, based on grid search method we found that n_neighbors should set to 19 and p should set to 1.

The voting method is build based on the combination of the models which has the top four highest score among all the classification models. These models are logistic regression model, Stochastic Gradient Decent, random forest, and Support Vector Machines. The other models such as Linear SVC, Stochastic Gradient Decent, Naive Bayes were training without any tuning hyperparameter methods. Finally, based on the table of model performance, we selected logistic regression model with SequentialFeatureSelector in Backward as the final classification model to predict the Q26. Although Random Forest has equal accuracy score as logistic model with SFS Backward and even has a higher F1 score, Random Forest is harder to be explained than Logistic model. Thus, we decided to give up the Random Forest.

1.4 Logistic Regression with SequentialFeatureSelector in Backward.

- **Confusion Matrix:**

	Prediction for Q_26 =0	**Prediction for Q_26 =1**
Actual for Q_26 = 0	**15**	**4**
Actual for Q_26 = 1	**4**	**21**

- **Performance of Logistic Regression with SequentialFeatureSelector in Backward:**

	Accuracy	**Precision**	**Recall**	**F1**
score	**0.82**	**0.84**	**0.84**	**0.84**

- **The ROC Curve (see the graph in Appendix)**

When we use the y training data set and x training set to calculate F1 score, f1 score is equal to 0.8392857142857144 which is very close to the f1 score calculated with y test set and x test set. This is proved that Logistic Regression with SequentialFeatureSelector in Backward does not have the overfitting problem.

- **Coefficient explanation:**

Using SequentialFeatureSelector in Backward method on Logistic model, we found that the selected features are Q25_Yes, TRAVEL_DURATION, EDUCATION, DISTANCE, and ANXIETY. However, the coefficient of each feature in the below bar graph cannot be directly explained because they are the logit numbers. One thing has also to be mentioned is that it is meaningless to transform each logit number to see the marginal effects. The reason will be showed in the limitation part for Logistic model.

What we can see from the graph is the correlation between each variable 'coefficient and Q26. Education is inversely related to people's propensity to travel in the next six months. In the features of education, the higher the educational level, the smaller the number. Combining the log curve, we can find that the higher the educational level, the more willing people are to travel. People with lower educational backgrounds in China may indeed be more sensitive to the epidemic. They may not believe in the so-called scientific methods of epidemic prevention. On the contrary, people with higher education think that if they protect themselves through scientific methods, they can let themselves be infected by germs. Therefore, highly educated people feel that as long as they wear masks and pay attention to distance, they can go to live or travel normally. Social distance is positively correlated with Q26_no and negatively correlated with Q26_Yes. This can be explained by the fact that the more people care about social distancing, the less willing they are to travel during the epidemic. Anxiety and Q26_Yes were negatively correlated. This is very realistic. The more anxious people are, the less willing they are to travel. Q25_Yes means people had travel experience six months ago. This feature is positively correlated with Q26_Yes. It seems that people who have traveled in the first six months can also consider planning to travel in the next six months, indicating that this group of people does not care much about their travel plans for the epidemic. Travel Duration is also positively correlated with Q26_Yes. This means that the more willing people are to travel for a long time, the more willing they are to plan travel in the future. People who tend to travel for a long time are fonder of travel, so the epidemic does not seem to affect such people. (See Appendix for coefficient of each feature in the bar chart)

- **Application**

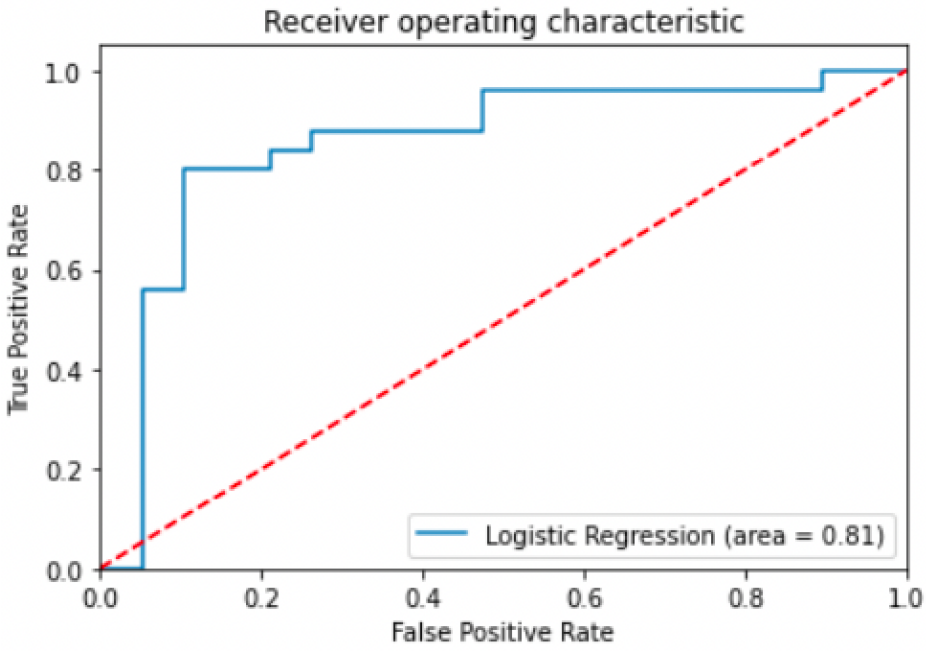
Since the model has an 84% accuracy in predicting the possibility of people planning to travel in the future and the fact that it can effectively avoid the overfitting problem, the

model can be used by travel companies for classify the potential customers. More importantly, the model reflects the relationship between the five selected features and people's traveling demand tendency. For example, travel companies can pay more attention to people who have traveled in the previous six months because these groups have a higher probability to make future travel plans. When targeting the tourist, education and epidemic anxiety level should also be considered. People with higher education are more likely to make travel plans in the future, and the more anxious people are about the epidemic, the less willing they are to travel in the future.

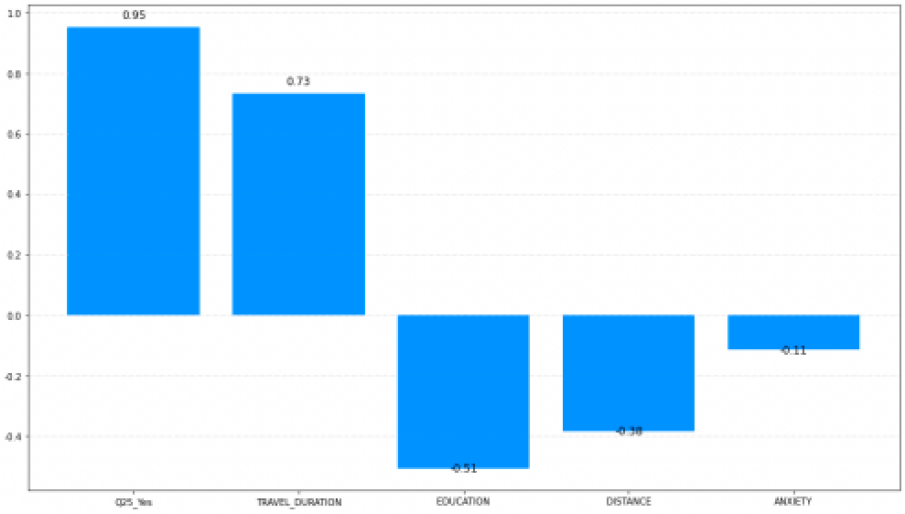
- **Limitation**

Firstly, the marginal coefficient of each feature is not a constant value, so we have no way to accurately explain how much changing a feature alone can affect the probability of Q26_Yes when other features remain unchanged. The reason for this may be that there is a large correlation between our features. So, in future, we should pay more attention to how to design independent features as much as possible. In this way, the interpretability of the coefficient of each feature can be greatly increased. Moreover, the information collection of the questionnaire limits the performance of the model. Due to our limited ability to obtain data, we cannot collect all the information we want. For example, if we can collect an accurate and continuous set of traveling budget data, instead of simply surveying the budget interval, the final model may perform better. In short, we hope to be able to obtain all feature information with continuous numerical significance, and we believe that such data can improve the model greatly. Finally, small data volume is also a problem in this logistic regression model. Larger data volumes would make the classification of the model more convincing.

1.5 Appendix



The ROC Curve for logistic regression model.



The bar chart for the coefficient of each selected features for Logistic regression model.

2. Unsupervised Learning: Segmentation of Current Tourists

2.1 Purpose of learning

As mentioned before, instead of looking for a specific aspect of current tourists, we want to conduct our research from an overall level. One thing that is always important is the way we classify tourists although segments could be vary according to change of selected attributes. We want to know the clusters based on the data we have and try to locate the group of people who would be willing to travel still.

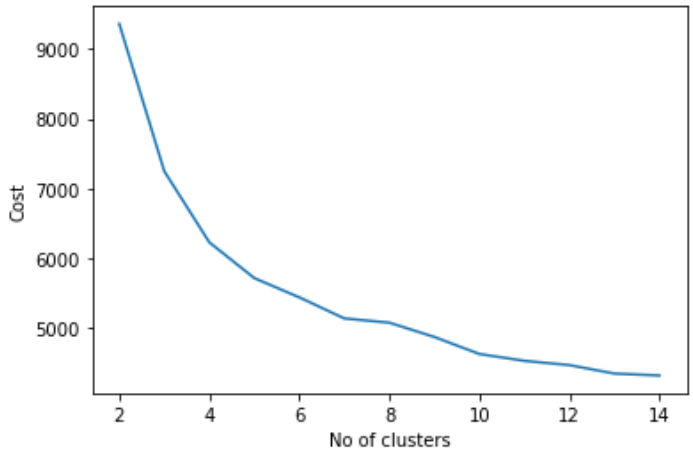
2.2 Data used

Discriminate variables and segmentation variables were used, all the questions designed for principal component analysis and transportation column were dropped at this point.

2.3 Pre-processing

As our data included both numerical and categorical variables, Kmeans or Kmodes are not proper tools here since they only work for dealing with one type of variable. The substitution clustering model to use is called "Kprototype" which can handle both numerical variables and categorical variables simultaneously, as long as we highlight the column index whom the model needs to treat categorically beforehand. Furthermore, we do not need to standardize the data value since it has been standardized from the very beginning when respondents input their feedbacks which has been scaled between integer 1 to 5.

Finding the right number of clusters



Cost means the sum distance each observation has to its own cluster centroid; the best choice is where the elbow appeared. In our case, from the above figure, the number of clusters could be 4, 5 or 6. We decided to run a four clusters model.

2.4 Identify the clusters

The idea is to compare each cluster with the population mean and population mode.

Population Mean/Mode:

	AGE	HHINCOME	DISTANCE	ANXIETY	ACCOMMODATION	TRAVEL_DURATION	BUDGET	NECESSITY
count	129	129	129	129	129	129	129	129
mean	26.72868217	1.976744186	2.790697674	2.596899225	3.085271318	2.906976744	2.891472868	2

	SEX	PROFESSION	VACCINE	EDUCATION	COMPANION	Q25	Q26
top	Male	business	Yes	Bachelor	3	Yes	Yes
freq	104	91	129	65	103	116	103

Cluster 1: "The Lady Force"

	AGE	HHINCOME	DISTANCE	ANXIETY	ACCOMMODATION	TRAVEL_DURATION	BUDGET	NECESSITY
count	26	26	26	26	26	26	26	26
mean	49.80769231	1.692307692	2.884615385	2.653846154	2.307692308	2.615384615	2.5	3.769230769

	SEX	PROFESSION	VACCINE	EDUCATION	COMPANION	Q25	Q26
top	Famale	education	Yes	Bachelor	3	No	Yes
freq	22	9	26	15	20	15	14

Characteristics to this cluster:

- Median age female
- Haven't traveled for the past six months but desire one trip. (Q25-Q26)
- Not very anxious about COVID
- Think traveling is a necessary activity in human life
- Could be defined as the target cluster

Cluster 2: "The Fuerdais"

	AGE	HHINCOME	DISTANCE	ANXIETY	ACCOMMODATION	TRAVEL_DURATION	BUDGET	NECESSITY
count	65	65	65	65	65	65	65	65
mean	25.36923077	2.707692308	3.215384615	2.784615385	2.723076923	2.584615385	3.030769231	3.446153846

	SEX	PROFESSION	VACCINE	EDUCATION	COMPANION	Q25	Q26
top	Male	students	Yes	Master	3	Yes	Yes
freq	38	29	63	32	32	38	44

Characteristics to this cluster:

- Generation Z
- Many of them are students
- Claimed that they are not very anxious about COVID but looking for a further social distance.
- Think traveling is a necessary activity in human life
- Could be defined as the target cluster

Cluster 3: "Travel-Aversion"

	AGE	HHINCOME	DISTANCE	ANXIETY	ACCOMMODATION	TRAVEL_DURATION	BUDGET	NECESSITY
count	25	25	25	25	25	25	25	25
mean	36.16	1.64	4.4	3.16	2.24	2.36	2.52	2.72

	SEX	PROFESSION	VACCINE	EDUCATION	COMPANION	Q25	Q26
top	Famale	self-operated business	Yes	HighSchool	3	No	No
freq	23	10	24	13	12	15	19

Characteristics to this cluster:

- They don't like to travel
- They haven't traveled and not planned to
- Being terrified by COVID
- Consider not to put effort on this cluster

Cluster 4: "Retires"

	AGE	HHINCOME	DISTANCE	ANXIETY	ACCOMMODATION	TRAVEL_DURATION	BUDGET	NECESSITY
count	13	13	13	13	13	13	13	13
mean	60.23076923	1.769230769	2.769230769	2.461538462	2.153846154	2.692307692	1.923076923	2.769230769

	SEX	PROFESSION	VACCINE	EDUCATION	COMPANION	Q25	Q26
top	Famale	unemployed	Yes	Bachelor	3	No	No
freq	8	6	12	5	8	8	7

Characteristics to this cluster:

- Not afraid of COVID
- Not interested in traveling
- Elders
- Consider not to put effort on this cluster

3. Unsupervised Learning: Positioning of Travel Destination

3.1 Purpose of learning

Logically, there is no way that COVID hasn't changed people's perspective toward tourism industry. Yet the question is how things got changed? Obviously, the likelihood of having a trip would be decreased; but is this a universal answer no matter what travel destinations are under consideration? Each travel destination has its own natural attributes such as tourists population density and indoor-outdoor platform applied; We believe those attribute would also be some hidden factors that affect tourist to make travel decisions.

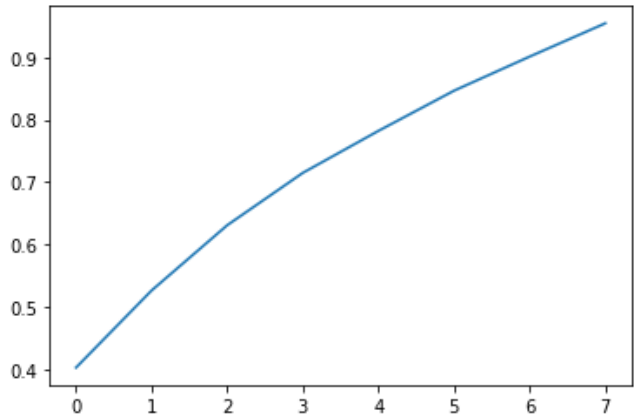
3.2 Data used

Perspectives of Travel Destination (Designed for PCA positioning).

3.3 Pre-processing

The format of data collected by "XM Survey" could not be used directly in python through sklearn.pca. Python required each respondent to answer each question based on every single observed unit, whereas the original one is some sort of short formats without appropriate labels. Thus, we would have to manually transfer data from the raw version to the required version. A few lines of codes then showed us the cumulative variance explained by increasing dimensions.

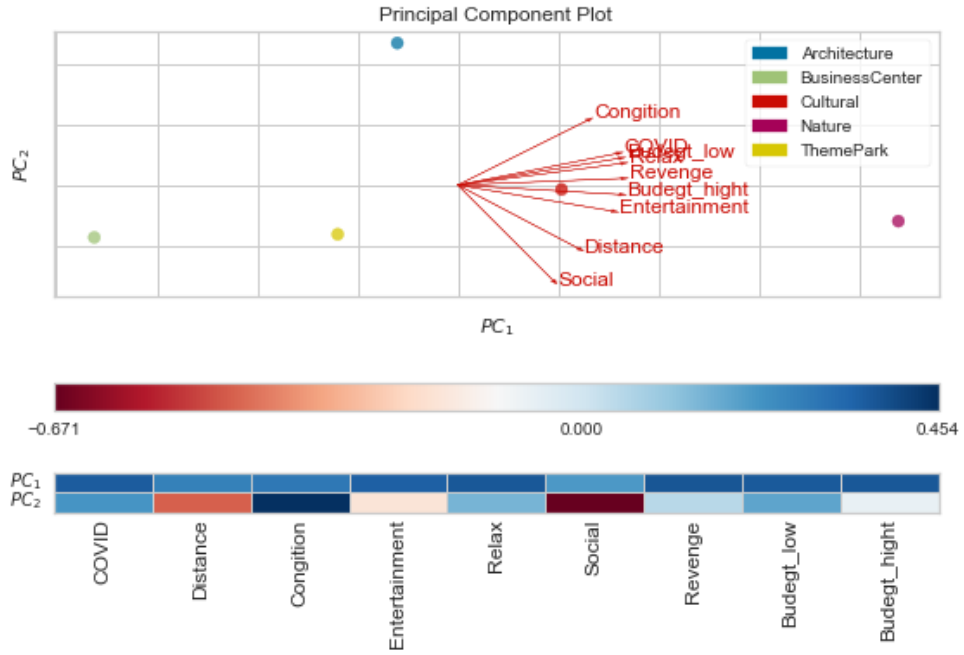
3.4 Dimensions required to explain variances



From the minimum (1 dimension) to the maximum (8 dimensions), showing on the graph is 0 to 7; We have to apply 4 dimensions of principle components to explain more than 70% of the variance, but 4-D cannot be visualized.

So, the idea now would be ignoring the capability of explaining variance but focusing on the pattern it showed on 2D visualizer.

The visualization on 2D-PCA perceptual map



Based on the correlations between each PC to variables, the PC1 and PC2 could be described as:
PC1: “absolute Index”(The location with a higher PC1 is where people are more willing to go no matter what is their travel motivation, yet the level of anxiety about COVID increased)
PC2: “Solo Index” (The location with a higher PC2 is not where people want to go for increasing relationship with others and relatively further away from their home)

3.5 Interpretation of the perceptual map

What we can see from this plot cannot go very specific because the variables set is not very comprehensive. All the "arrows" are pointing to one direction. We can say that our survey might not cover all the possible aspects regarding all the travel destination we selected; still the actual reason could be controversial. One of the persuasive conjectures is that not only does COVID change people's behavior and way of thinking but also the subconscious when making decisions. As the existence of COVID is an established fact, people would have their preference toward each travel destination based on the stereotypes they have coming with that travel destination, even before they fill out the survey. As a conclusion, Business centers, Architecture, and Theme Park type of travel destinations might indeed not be the preference of tourists nowadays no matter what questions we designed.

Part4: Implication

No matter which machine learning methods to discuss, the outcomes, to a certain degree, have answered the question if COVID is affecting tourists’ behavior and decision making. From the classification modeling, using feature selection to end up with only five variables where two of the variables are COVID related is proving that people are making travel decisions considering COVID-related topics.

Likewise, in clustering modeling, a cluster called “Travel-Aversion,” shows a preferred social distance of 4.4, which refers to a distance between 8 to 10 meters away from each other. In contrast, the population mean is only about 2.8. Such a cluster would definitely not be the one that any travel agency wants to put effort into.

Last but not least, even though the positioning modeling couldn’t explain variance to an acceptable level, there are stories we can read from the outcome. Since we have listed almost all the possible motivations for a human being to travel, the reason why travel destinations perform this unevenly is hard to explain in the circumstance that we assume the survey is well designed. The other explanation would be the Nature type and the Cultural type of travel destination are the two growing in popularity no matter for what reason a traveler travels. Suppose this information was exposed for facilities like travel agencies, the advantage in terms of designing a tourism product would be financially beneficial.

Reference

Swensen, K. (2021, September 24). *New report reveals just how much theme park attendance dropped in 2020*. Inside the Magic. Retrieved April 29, 2022, from: https://insidethemagic.net/2021/09/2020-theme-park-attendance-report-ks1/#:~:text=Overall%2C%20theme%20parks%20across%20the,approximately%2083%20million%20in%202020

Ritchie, H., Mathieu, E., Rodés-Guirao, L., Appel, C., Giattino, C., Ortiz-Ospina, E., Hasell, J., Macdonald, B., Beltekian, D., & Roser, M. (2020, March 5). *Coronavirus pandemic (COVID-19)*. Our World in Data. Retrieved April 29, 2022, from: https://ourworldindata.org/coronavirus

Data-Transport. Data- Ministry of Transport of People's Republic of China. (n.d.). Retrieved April 29, 2022, from: https://www.mot.gov.cn/shuju/

China domestic tourist. CEIC. (n.d.). Retrieved April 29, 2022, from: https://www.ceicdata.com/en/china/tourism-industry-overview/cn-domestic-tourist

Rubin, J. (Ed.). (n.d.). *Global Attractions Attendance Report - Aecom.com*. Theme Index Museum Index 2020. Retrieved April 30, 2022, from: https://aecom.com/wp-content/uploads/documents/reports/AECOM-Theme-Index-2020.pdf

Pearce, B. (2020, March 5). *COVID-19 Updated impact* assessment of the novel Coronavirus*. Retrieved April 30, 2022, from https://www.iata.org/en/iata-repository/publications/economic-reports/coronavirus-initial-impact-assessment/