Mickey Mouse or The Met: How Has Tourism Rebounded After COVID-19?

A Project Proposal

Jordyn Irwin, Joselyn Molina Lagunas, and Brittany Fisher

Research Question

What travel/tourism industries in the United States have rebounded after the COVID-19 pandemic and why are some more successful in their recovery efforts?

Background & Motivation

The COVID-19 pandemic had a marked impact on many businesses in the United States, with some thriving and others being forced to adapt their business practices to maintain their revenue. The travel industry was one of the most heavily impacted by pandemic restrictions, both nationally and globally. While the United States Congress has addressed the pandemic's effect on tourism and the factors that affect tourism in general (Levin, 2023), the factors that influence whether or not an attraction has yet to fully recover from the effects of COVID-19 have yet to be identified. The aim of this research is to determine factors that have impacted industries' revenue, and to determine why some businesses have fully recovered post-pandemic and others have not. Through the utilization of data reporting the number of visits to various U.S. attractions as well as sentiment analysis from travel blogs and review sites, we aim to identify which factors have the most impact on post-pandemic recovery and if these factors are ubiquitous among different categories of leisure attractions. Businesses can use the insights gained from this research to help recoup revenue to pre-pandemic levels, bolstering both the local and national economies.

Data

We intend to obtain and split our data into 4 categories: Paradise (Resorts/Beaches/Ski resorts), Wilderness (Forests/Deserts/Parks), Culture (Museums/Cities/Festivals), and Play (Theme parks). This will allow us to identify any differences in trends between categories that may exist. The goal is to obtain data for the top 20 U.S. attractions for each category. However, we may be limited by the availability of data and be forced to narrow our research focus to fewer categories. All data will be combined/joined together based on attraction name and/or location. The following section outlines the data sources that we have identified thus far in our research.

National Park Data

- We will be utilizing data provided by the National Park Service regarding visitor use statistics, in order to obtain total visitor counts throughout the years (pre and post pandemic). Data is available from 1979 through 2023, and is separated by each individual National Park. These sources are listed below:
    - https://irma.nps.gov/Stats/
    - https://irma.nps.gov/Stats/Reports/National
    - https://irma.nps.gov/Stats/SSRSReports/National%20Reports/Annual%20Park%20Ranking%20Report%20(1979%20-%20Last%20Calendar%20Year)
    - https://www.nps.gov/rlc/researchdatabases.htm
        - Includes extensive details on US national parks that can be combined by name/location
    - https://www.sciencedirect.com/science/article/pii/S2352340919312612
        - Dataset for Indoor and Outdoor attractions
        - Includes official number of visits, social media reviews and environmental data for 76 U.S. National Parks with data from 2000-2020

Theme Park & Museum Data

- Information from the Themed Entertainment Association will be utilized to obtain visitor attendance numbers for various theme parks and museums throughout the country.
    - https://www.teaconnect.org/tea-theme--museum-index.html
    - https://acrobat.adobe.com/link/track?uri=urn%3Aaaid%3Ascds%3AUS%3A11a0b514-be51-3910-a988-6e7131387fd5&viewer%21megaVerb=group-discover
        - PDFs from TEA/AECOM display attendance numbers for the top attractions from 2019-2022 for a variety of categories (theme parks, water parks, museums, etc.). As this information is international, data from attractions located in the U.S. will be extracted for analysis.

Web Scraped Data

- Attraction information will be scraped from a variety of sources in order to obtain location, price, activity type, and other relevant business information. Additionally, attraction reviews will be scraped from a variety of sources in order to perform sentiment analysis. This additional information obtained from internet reviews will allow us to determine factors that contribute to attraction

success/economic recovery, and will allow our research to be unique to other existing research.

- Scrape Wikipedia for specific information about locations/attractions
- Scrape Yelp for reviews and business info
  - https://docs.developer.yelp.com/docs/fusion-intro
  - Scrape data by category
  - This would be very helpful to fulfill our 'novel' data requirements
- https://travel.usnews.com/
  - Gives information about location, price, type of activity
- Travel blogs for location reviews/sentiment analysis

Methodology

I. Statistical Thinking

A. We will analyze our data by testing our hypotheses with appropriate statistical methods (t-tests, Tukey HSD, proportion tests, Chisq tests, etc)
B. In a basic sense, those will be evaluated by statistical significance (e.g p-values < 0.05); however, interpretation will be dependent on the context of those results.
C. All statistical analyses will likely be conducted using R.

II. Data Visualization

A. Depending on if the data/analysis is significant spatially we will use maps.
B. Bar charts, scatter plots, and histograms will all be useful to show trends and reinforce our findings.
C. A mixture of R and Python will likely be used for visualizations.

III. Data Engineering

A. Since we intend to scrape our data, we will need to input that data into SQL tables.
B. Depending on how pre-existing datasets are formed, we may need to clean the data and be able to import that into integrated SQL tables.
C. If scraping live data, we will use Railway and Github to maintain that process and scrape once a day.

IV. Machine Learning

A. We intend to use machine learning to predict which travel industries will continue to grow/falter. Machine learning will be conducted using Python.
B. Additionally, we want to create a vacation/trip recommendation dashboard with a machine learning model. This will likely be presented in the form of a Shiny app/dashboard.

V. Data Ethics

A. In addition to ensuring the protection of personal data, the data we intend to collect/use is publicly available which eliminates most privacy issues. However, since we intend to collect data from travel blogs, we would need to be mindful of discluding personally identifiable information from those blogs.

B. Tourism trends and locations can be influenced by discriminatory practices based in class, race, gender, etc. They greatly impact travel experience and access to leisure attractions. Since we want to include a recommendation model, we are cognizant of our ethical responsibilities and are careful of not reinforcing those issues.

References

Levin, A. G. (2023). U.S. Tourism: Economic Impacts and Pandemic Recovery. *Congressional Research Service*, *R47857*(3). https://crsreports.congress.gov/product/pdf/R/R47857/3