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

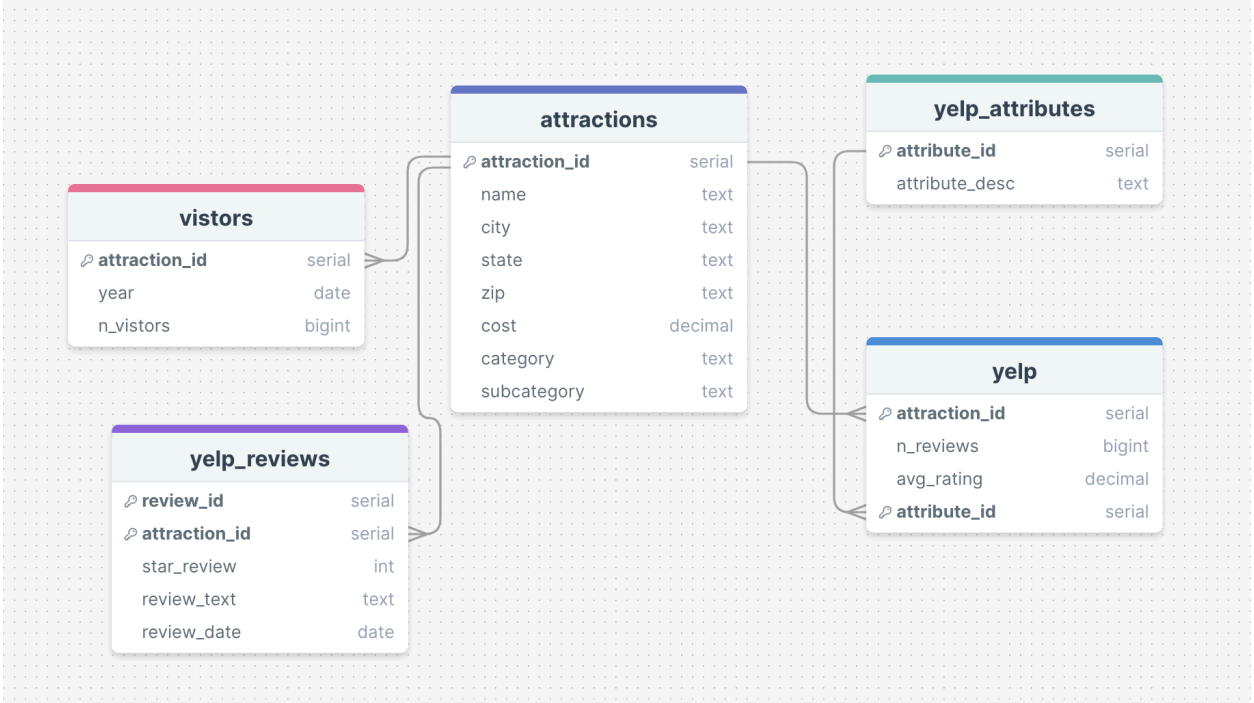
Jordyn Irwin, Joselyn Molina Lagunas, and Brittany Fisher

Data Summary

We have scraped all attraction information such as location, attendance, cost, average Yelp review, number of Yelp reviews, Instagram hashtag information, etc. However, we are still in the process of scraping individual reviews for the top 20 most reviewed attractions in each category. To clean the data, we pivoted from wide to long format, subcategories have been added to National Landmark data to further specify category type (e.g. National Park, National Memorial, National Historic Site, etc.), and the yelp attributes that were combined were separated. Data has been transformed to the correct variable type for each data frame column. Additionally, white space has been trimmed from all column rows. We are currently working on organizing our data to match our ERD. We have begun some EDA on the current data to find trends and patterns that align with our project proposal.

Data Design

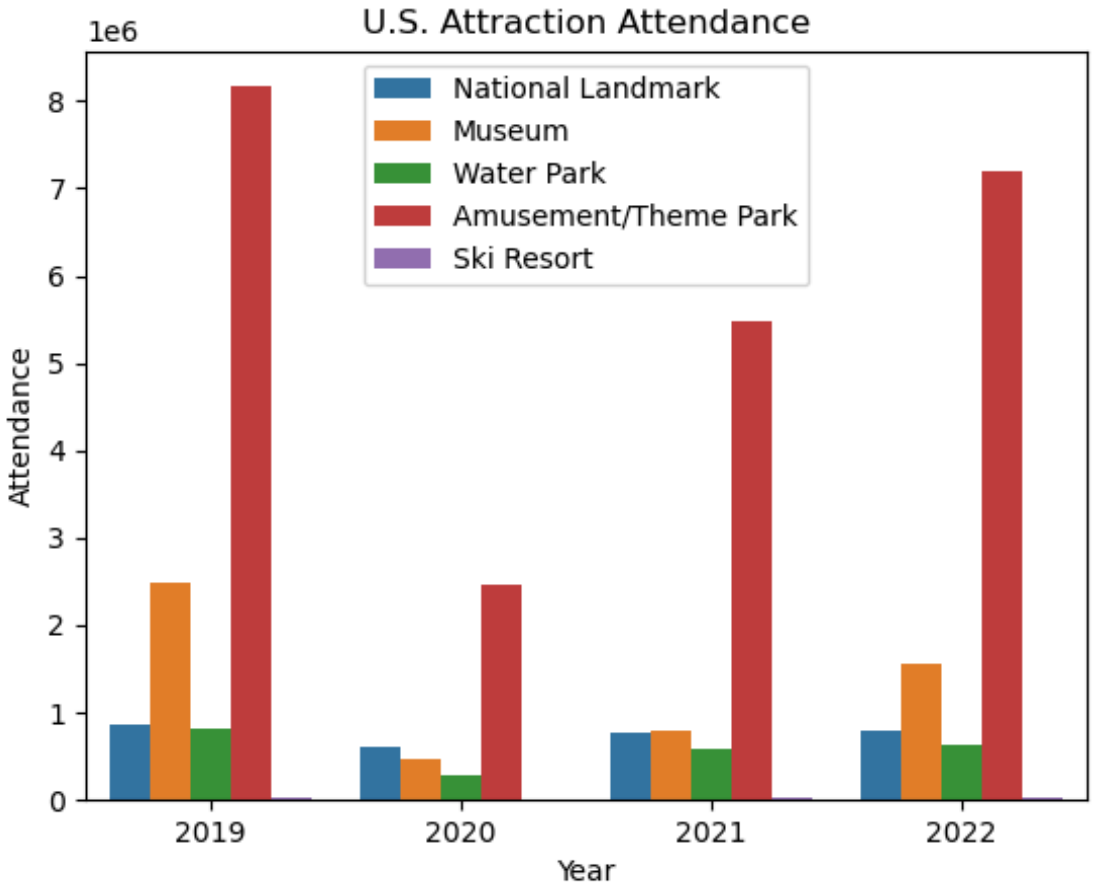
The main table is “attractions” which has an attraction_id as a primary key. For this table, it includes all general data about that attraction. The “visitors” table has the attraction_id as a primary key along with the years and visitors per year. The “yelp_reviews” table has the review_id as the primary key, attraction_id as a foreign key and then the review data. The “yelp” table has the attraction_id as a primary key and attribute_id as a foreign key. It is separate from the reviews table because our reviews table only includes limited years and attractions whereas the ‘yelp’ table has all attractions and the avg_rating for all time. The “yelp_attributes” table includes the descriptions of the attribute_ids. We have worked towards 3NF by separating data to prevent partial dependencies and transitive dependencies.



Sample EDAs

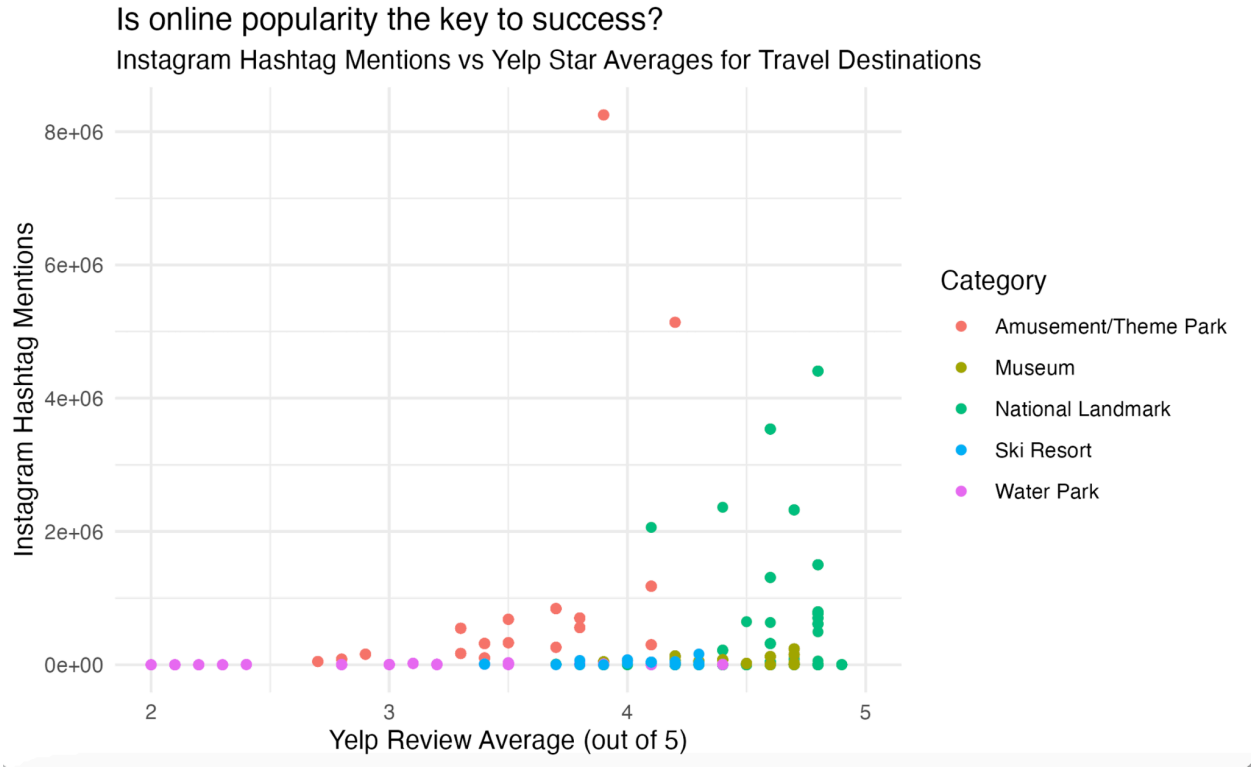
EDA #1

Graph of attendance throughout the years (pre, during, and post pandemic) by category allows us to see which, if any, attraction categories have recovered their attendance to pre-pandemic numbers. Additionally, it gives us an idea of how far attendance fell during the height of the pandemic, and the rate of recovery.



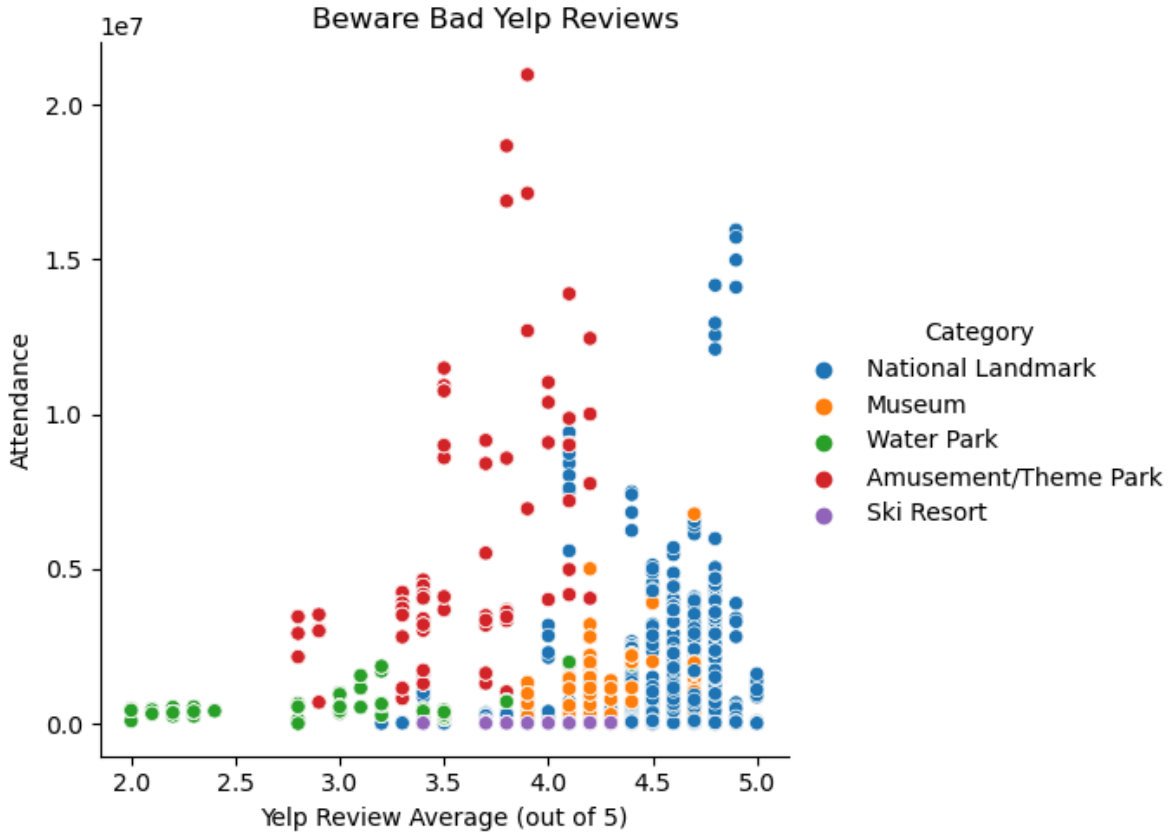
EDA #2

This graph shows the interaction between uses of an instagram hashtag for an attraction and the average Yelp stars an attraction has. It has been colored by category to highlight the differences between specific travel industries. It is giving us an insight into if there is a correlation between the social media popularity of an attraction and the reviews of that particular attraction. This leads to our theory that social media has effected the post-pandemic success of a travel destination.



EDA #3:

Graphing the relationship between star reviews on Yelp and attraction attendance can help us understand if bad reviews are a potential factor that is inhibiting an attraction's recovery from the effects of the covid-19 pandemic.



EDA #4:

Graphing the relation between the cost per person and category activity type allows us to see the variety of price for different activities. This is helpful when considering what activities better fit certain budgets and also allowed us to consider possible changes of price pre, during, and post covid.