MISM6205 – DATA WRANGLING FOR BUSINESS
PROJECT REPORT

Group Number	6
Group Members	Ayushi Anantwar (002751798) Pooja Pendharkar (002653116) Ramya Kumaresan (002990437)
Project Goal	To recommend must-visit destinations in proximity to major attractions in New York City based on the type of tourists.

TABLE OF CONTENTS

TOPIC AND BUSINESS QUESTIONS

This project aims to recommend must-visit destinations in proximity to some major attractions in New York City based on the type of tourists. The business-related question that we are planning to explore through this project is - "How to plan a short yet fulfilling trip for the incoming tourists". Our project aims to maximize the number of locations tourists can visit within a short period. It is assumed that tourists will visit important attractions in the city such as the Statue of Liberty, Central Park, Times Square, and the Empire State Building. So, the results of this project will aid tourists in identifying some interesting places such as eateries, museums, art galleries, beaches, etc. around the four major attractions in New York City.

New York City is taken as a use case in this project as it is popularly known as the world's busiest tourist destination. Over 60 million tourists visit New York City every year, of this number 12 million come from outside the United States. Most of these tourists depend on travel blogs or different websites to gather information on the best places to visit.

The trip planners available in the market are based on travel dates or the number of days for the trip and showcase a list of top places to visit as a recommendation. However, the restrictions of travel time between these places are not considered. The algorithms do not recommend a list that would allow the tourist to visit multiple locations in a short amount of time. For example, Wanderlog recommends a generic list of places to visit. Roadtrippers shows a list of tourist attractions based on route and considers the travel time, however, it is focused specifically on road trips i.e., intercity travel instead of intracity. Most tourists plan their trips during weekends, office leaves, or holidays. They wish to plan a trip that helps them make the most of the available time i.e., being able to cover as many attractions as possible. This project aims to plan such a trip for tourists by recommending must-visit places around the top attractions that one can cover according to their interests.

What makes our project unique is that it not only suggests the must-visit places based on the location proximity but also considers the type of tourists visiting New York City (e.g.: Individuals, Friends, Family, and Couples). This feature of our project helps tourists to choose what is most likely to suit and accommodate their travel preferences from a range of options.

Through our project, we wish to take a step in this direction to help tourists plan a trip with an optimized list of attractions they can visit. The recommendations are shortlisted by traversing multiple datasets to showcase the places which are closest to a specific tourist attraction.

DATASET AND DATA SOURCE DETAILS

The data collection process required us to collect the below-listed datasets that contain most of the required information about the tourist attractions and eateries in New York City. The "***ZIPCODE***" would be used to recommend the best attractions in and around the specific area of choice entered by the user. We have added a separate column called "***CATEGORY***" in each of the datasets to specify the category for each tourist attraction.

Dataset Name	**Dataset URL**	**Variables used for our analysis**	**Dataset Provided By**
Museums	https://data.cityofnewyork.us/Recreation/New-York-City-Museums/ekax-ky3z	***Name*** - name of museum, ***Tel*** - phone no, ***URL*** - website, ***Address1***- Street name, ***Address2*** - building name or floor, ***Zip*** – zipcode	Department of Information Technology & Telecommunications (DoITT)
Art Galleries	https://data.world/city-of-ny/tgyc-r5jh	***Name*** - Name of the art gallery, ***Tel*** - Contact Number of the art gallery, ***URL*** - Link to the art gallery's website, ***Address1*** of the art gallery, ***Address2*** - *Additional details* about address of the art gallery, ***Zip*** - Zipcode of the area where the art gallery is located	Data World
Beaches	https://data.world/city-of-ny/zyf6-z3xt	***Name***-Name of the property, ***location***- gives the location details of the beach, ***Phone***-Contact number, ***Lat and long*** - shows the latitudes and longitudes of beaches, ***Description*** – Contains a few words about the beaches	Data World
Botanical Gardens	https://data.world/city-of-ny/hrii-hezj/workspace/file?filename=botanical-gardens-1.csv	***Name***- different names of the botanical garden, ***Tel***- phone number of different botanical gardens, ***URL***- website of the different botanical gardens, ***Address1*** – it shows the address of different botanical gardens in NY,	Data World

		Zipcode-different area zip codes of different botanical gardens	
Water Trails	https://data.world/city-of-ny/hxay-3qcw/workspace/file?filename=new-york-city-water-trail-kayak-and-canoe-launch-sites-1.json	***Name***- names of the water trails and sport, ***Location***- locations of water trails and sport, ***Description***- more information about water trails and sport, ***Lat and long*** - shows the latitudes and longitudes of water trails and sport	Data World
Hiking Trails	https://data.world/city-of-ny/i8f4-bu5r/workspace/file?filename=directory-of-hiking-trails-1.json	***Name-*** Name of the hiking trail, ***location***- location of the hiking trail, ***Other_Details***-Additional notes	Data World
Restaurants	https://data.cityofnewyork.us/Health/Restaurants-rolled-up-/59dk-tdhz	*DBA*-Name of restaurants, *BORO*-boroughs of New York City, *Street*-Street where the restaurant is located, *Zipcode*- Zipcode	Department of Health and Mental Hygiene (DOHMH)
Zip code Proximity Data - Manually gathered from the website	https://ny.postcodebase.com/zipcode_radius#myarticle	*ZIP Code_Statue* - Zip code of location close to Statue of Liberty, *Distance_Statue* - distance of location zip code from the statue of liberty, *ZIP Code_Empire* - Zip code of location close to Empire State Building, *Distance_Empire* - distance of location zip code from Empire state building, *ZIP Code_Times* - Zip code of location close to Times Square, *Distance_Times* - distance of location zip code from Times Square, *ZIP Code_Central* - Zip code of location close to Central Park, *Distance_Central* -distance of location zip code from Central Park	Manually collected from website - https://ny.postcodebase.com/zipcode_radius#myarticle

INFORMATION QUALITY

- The addresses for a few datasets were not separated as Address and Zip code, this made it difficult for us to extract the data and add it to the appropriate columns. This issue was resolved during the data-wrangling process.
- Since we did not have any readily available dataset to determine the zip code proximity for the four major tourist attractions (Statue of Liberty, Empire State Building, Central

Park, New York Times Square) in New York, the zip code proximity dataset for each of the four major tourist attractions had to be manually created to further enrich the data used for our analysis.

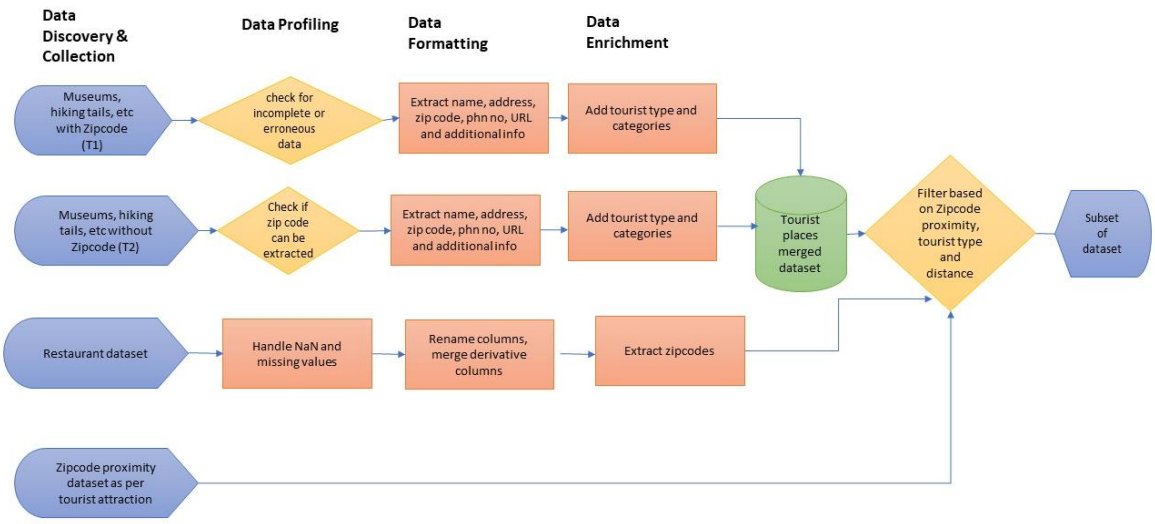
- The process or methodology of data collection is not provided for the datasets from the website "Data World" which makes the reputability of the dataset questionable. However, we went ahead with using these datasets by manually cross-verifying some of the information from the datasets that would be used for our analysis.

METHODS AND TOOLS

Listed below are the common methods used for wrangling the data post data collection -

- **Data Discovery and Collection** – The datasets required to perform analysis and solve the business problem had to be collected and cleaned to suit the business requirements.
- **Data Profiling** – The datasets had to be analyzed for understanding the purpose and validity of different values and attributes.
- **Data Formatting** – The datasets were formatted to suit the requirements of the analysis.
- **Data Enrichment** – New columns were added to the existing datasets and a new dataset was manually created to help enrich the information provided in the originally collected datasets.

The below diagram illustrates the data-wrangling concepts used for our project -



Tools -

The tools that were used for our project include:

- Microsoft Excel

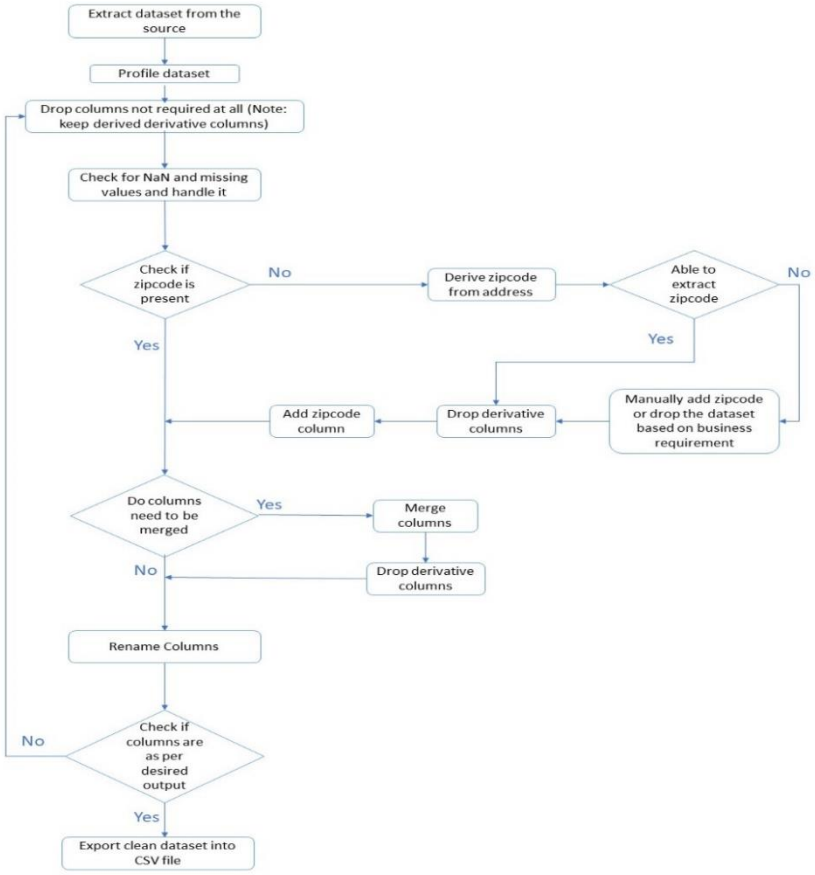
- Python - Pandas, Geopy, PySimpleGUI
- Tableau

Challenges -

- The process of fetching the data on zip code proximity for the four major tourist attractions in New York was a challenge during the data collection phase.
- Generating the zip codes from the "***ADDRESS***" field for those tourist attraction datasets that did not have a value for zip code was a challenge. To ease this process, we used the Geopy library in python. However, we were unable to generate the zip code values for certain locations as Geopy could not detect certain tourist locations based on the address parameter that was passed. Instead of returning the correct zip code values, it returned NaN values. To avoid this issue, we had to manually enter the zip codes for a few datasets.
- This project required us to use the PySimpleGUI library for the first time, it took a while for us to learn about the library and implement its functionalities in our project. Making use of this library allowed us to establish a better user-system interaction for the project to provide personalized recommendations.

DATA WRANGLING PROCESS

The diagram below depicts the data-wrangling process used for our project -



Some of the validation rules and checks implemented for our project are as follows -

- Checked the datatype of each column in the individual datasets to ensure uniformity in the datatypes of the column values before merging the datasets to perform the required analysis.
- Formatted the columns (renamed, arranged them in the correct sequence, dropped the derivative columns after merging them to get the required data value) in each of the datasets to ensure overall uniformity in columns before merging the datasets.
- Ensured that each of the files was in the right format before merging them to obtain the final dataset for tourist attractions.
- Checked if the "***ZIPCODE***" column existed for each of the individual datasets. If no, manually added the zip code values.
- Handled the missing and the null values in each of the individual datasets before merging.

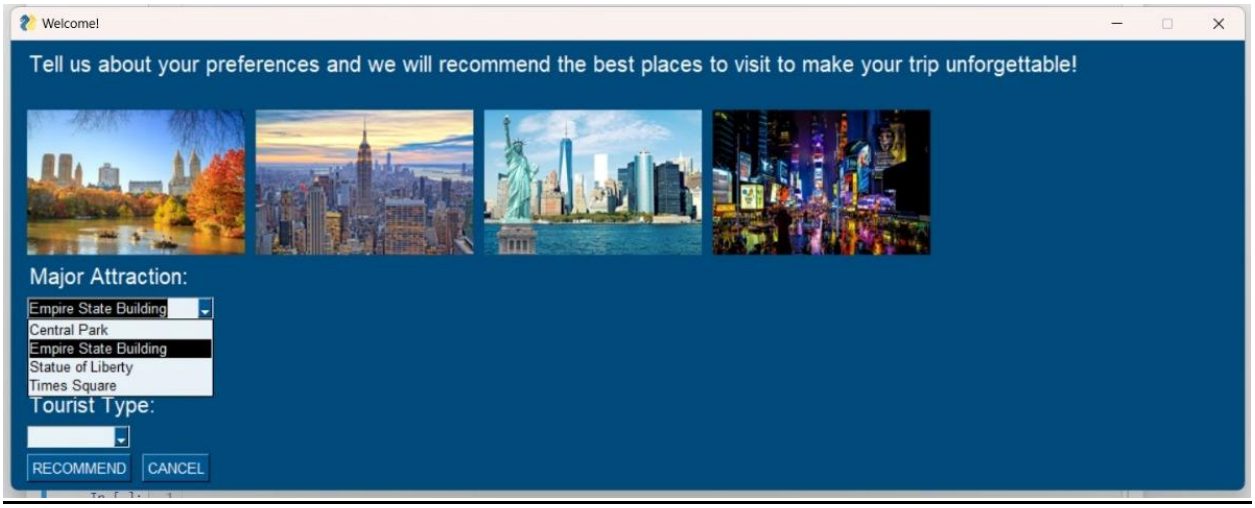
ANALYSIS AND RESULTS

Final Output -

Post fetching the required datasets and wrangling them to obtain the required information. We performed an analysis to determine the secondary tourist attractions and restaurants situated close to the four most popular and primary tourist spots in New York which include the Statue of Liberty, Central Park, New York Times Square, and the Empire State Building based on two primary filters: Distance in miles and tourist type.

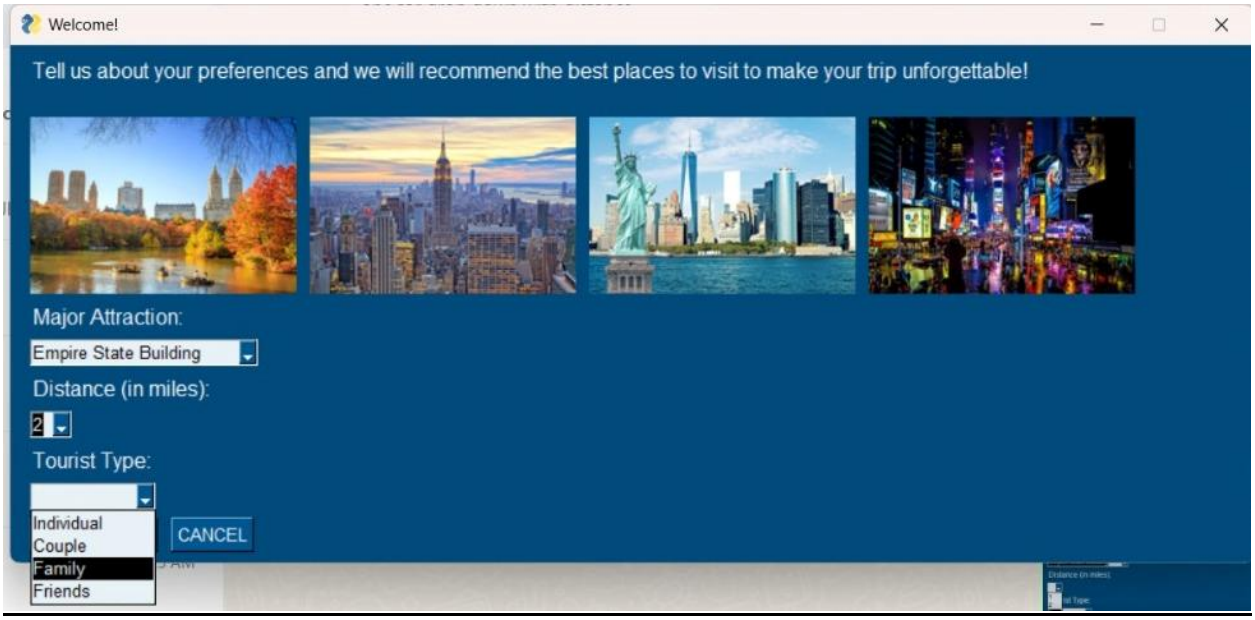
Attached below are the output screenshots –

Step 1: The user chooses one of the primary tourist attractions (Major Tourist Attraction) from the drop-down list



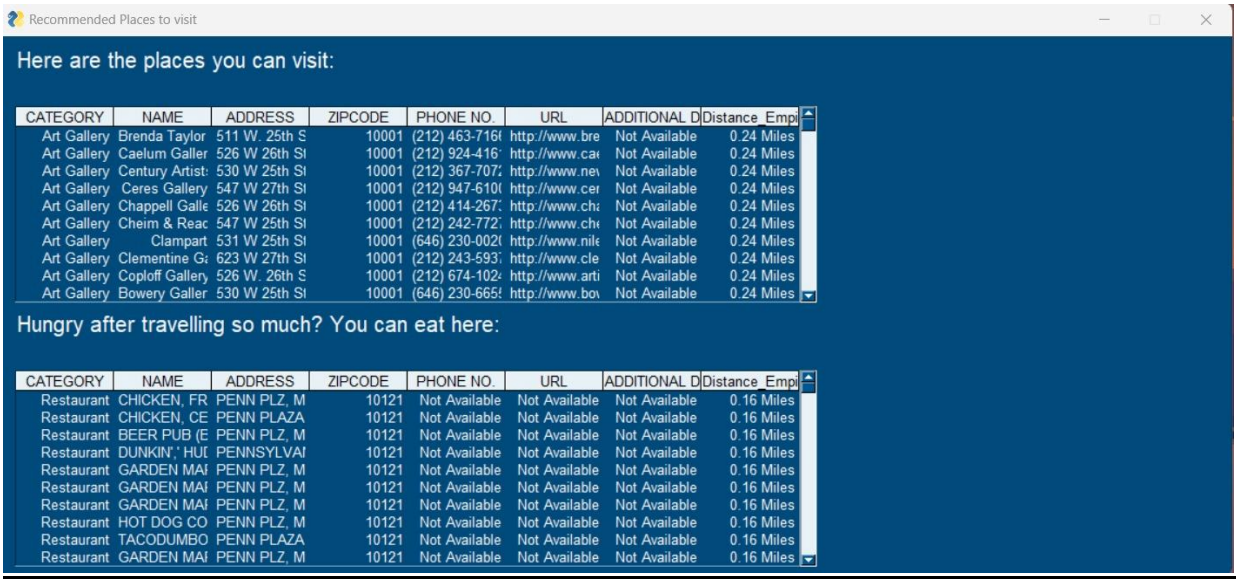
Step 2: The user chooses the distance (in miles) between 1 to 5 from the drop-down list to determine the secondary tourist attractions situated within the specified distance (in miles) from the primary tourist attraction/Major Tourist Attraction

Step 3: The user chooses the Tourist Type from the drop-down list to get the personalized recommendation



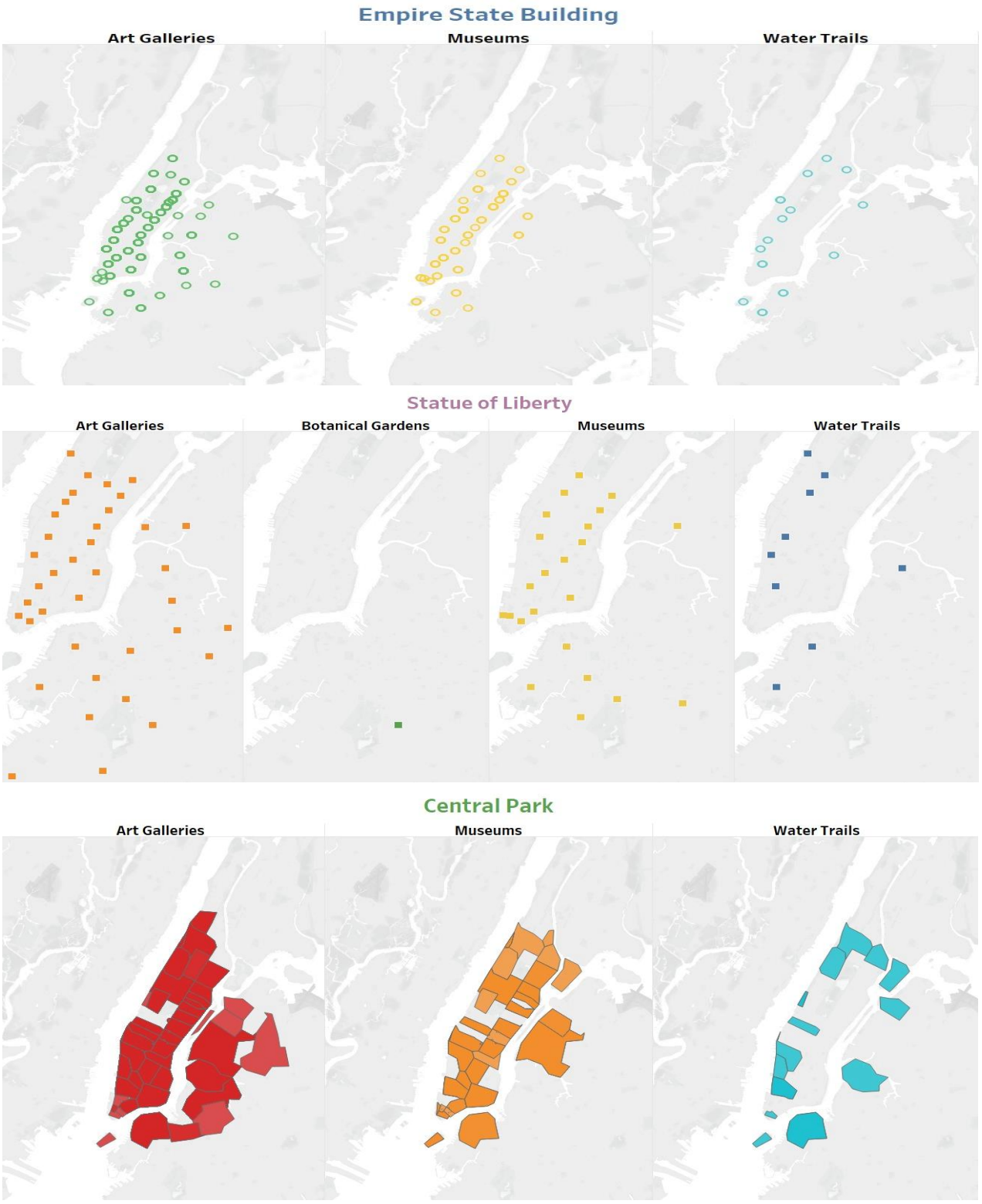
Step 4: The final output is printed in the form of two tables:

- The secondary tourist attractions within the specified distance range for the chosen tourist type
- The restaurants present within the specified distance range
- Each of the above-mentioned tables contains the following details: "***CATEGORY***" of the tourist attraction/ the "***CATEGORY***"- Restaurant, "***NAME***" of the tourist attraction, Exact "***ADDRESS***" of the tourist attraction/Restaurant, Exact "***ZIPCODE***" of the tourist attraction/Restaurant, "***PHONE NO.***" (If available), "***URL***" (if available) and some "***ADDITIONAL DETAILS***" about that tourist spot/restaurant.



Visualizations –

The visualizations for our final project were achieved with the help of Tableau Software. The below visualization represents the concentration of the secondary tourist attractions close to the four primary tourist attractions in NYC (Empire State Building, New York Times Square, Central Park, and Statue of Liberty)


Empire State Building
Art Galleries
Museums
Water Trails
Statue of Liberty
Art Galleries
Botanical Gardens
Museums
Water Trails
Central Park
Art Galleries
Museums
Water Trails

EXTERNAL ARTICLES

Mentioned below are some of the external supporting materials that were used to provide additional insights into our business question and for our project:

- *https://www.planetware.com/tourist-attractions-/new-york-city-us-ny-nyc.htm*
- *https://www.tripadvisor.com/Attractions-g60763-Activities New_York_City_New_York.html*
- *https://www.osc.state.ny.us/reports/osdc/tourism-industry-new-york-city*
- *https://www.tripsavvy.com/planning-a-new-york-city-trip-travel-guide-4178708*
- We could not find any other website or platform which performs a similar analysis. A trip planner called "Inspirock" *(Inspirock)* is quite close to what we are planning to achieve in this project where an itinerary is built for the user. However, it does not consider the type of tourists visiting New York City or the distance between different places.

Mentioned below are a few articles based on which we have allocated the tourist type for each of the secondary tourist locations:

- *https://www.timeout.com/new-york-kids/attractions/family-attractions*
- *https://secretnyc.co/things-to-do-with-friends-nyc/*
- *https://coupletraveltheworld.com/romantic-things-to-do-in-nyc-for-couples/*
- *https://www.theworldandthensome.com/15-fun-things-to-do-alone-in-nyc/*

Mentioned below is a link that helped us get the required zip codes for our project:

- *https://ny.postcodebase.com/zipcode_radius#myarticle*

ADDITIONAL DATA AND ANALYSIS

- Our primary goal was to internally sort the restaurants based on the reviews fetched from the Yelp Datasets to recommend the best ones for the tourists. However, due to the huge file size, we faced technical difficulties in accessing the dataset using Python and converting it to CSV. Due to a shortage of time, we were unable to achieve this task.
- From the GUI perspective, we wanted to further customize and enhance the output to provide recommendations to the tourists in a more refined and attractive way. This can be considered one of the areas of improvement in the project.
- Post the completion of our project, we felt that including a few more secondary tourist attraction datasets around the four major ones could have resulted in a wider range of attraction spots for the tourists to choose from.