Dataset Details

Table 1 summarizes the data sources used for the project, all of them publicly available, under the following categories:

- Bureau of Economic Analysis Data (GDP, Personal Consumption, Income, and Employment)
- Federal Reserve Bank of St. Louis (Unemployment)
- Annual Retail Trade Survey (Monthly Retail Sales and Inventories)
- Mobility Patterns (Apple Mobility Reports, Descartes Lab Mobility Changes, Google Community, Foursquare Community Mobility Data)
- Household Pulse Survey 2020
- COVID-19 (The COVID-19 Tracking Project)
- Restaurants platforms (Yelp dataset, OpenTable Data)

Every category includes a list of datasets, main features, and topics of interest.

Category	Datasets	Source	Features	Topics	Related topics
Bureau of Economic Analysis Data	1. Domestic Product and Income by Industry and Expanded Detail 2. Personal Consumption Expenditures by Major Type of Product 3. Income and Employment by Industry (until 2019)	https://www.bea.gov/data/by-place-us	Measured per qtr; will be updated Nov 25.	Trends of GDP in food service, employment, incomes(Table 1.5.3. Real Gross Domestic Product, Expanded Detail, Quantity Indexes)	Determine GDP trends in food services vs food purchased during the pandemic. Compare performance with other categories of durable-nondurable goods and services.
				Analysis of expenditures in food purchased and food services (Table 1.5.1. Percent Change From Preceding Period in Real Gross Domestic Product, Expanded Detail and Table 1.5.2. Contributions to Percent Change in Real Gross Domestic Product, Expanded Detail)	Determine the change in personal consumption expenditures in food purchased for consumption vs food services per qrt 2018-2020.
Federal Reserve Bank of St. Louis	Monthly Unemployment Numbers by Industry	https://fred.stlouisfed.org/release/tables?rid=50&eid=4635#snid=4770	Measured Monthly and separated by Industry. Includes both percentages as well as raw numbers	Trends of Unemployment and how it affected each industry.	Determining how this recession compares to that of 2008 and how long it took for things to recover. Month to Month change by industry can determine which industries are

		https://drive.google.com/drive/folders/14cO5A5K0ulc16antLYLwJEjv95So1YLM	Includes monthly data from 2020 alone along with data from 2005 to 2020.		recovering and if their recovery is influencing the restaurant industry. Analysis of whether these trends are shown in our other graphs.
Monthly Retail and Food Services Sales and Inventories (Annual Retail Trade Survey)	1. Estimates of Monthly Retail and Food Services Sales 2. Estimates of Monthly Retail Inventories/Sales Ratios	https://drive.google.c fromom/drive/folders/1HjBamjglzwTK_PZ7VxyPBZDtPe7nWiTk	Data 1992 to 2020. 3 Datasets: First on sales, second on inventory/sale percentage, and third on current data not incorporated into the first dataset. Measured monthly with a cumulative total for each year up to 2019. 2020 is also measured monthly but with a predicted annual total. Each sheet in the excel database contains both adjusted and non-adjusted sales. 2 Dataframes: First - Shape(28x16) contains annual sales totals in relevant industries. Second - Shape(336x16) contains monthly sales in relevant industries.	Compare performance of food/beverage stores vs food services for a duration of three decades. Example entries: Retail sales, total, Restaurants and other eating places, Full-service restaurants	Trend Analysis: Industries affected by covid. Factors: 2008 Recession, DotCom Bust, COVID-19, Seasons, inflation. Graphs showing the evolution of sales in food services over years and on a monthly basis.
Mobility Patterns	1. Apple Mobility Reports	https://drive.google.com/drive/folders/1NB1oEsFE33XczpebAINhuFH828U2fCbk	The relative amount of route requests from every region/date. Information broken into states, counties, date from Jan 2020-October 2020. 3 datasets with information by state and 3 datasets (complete) with information by county (incomplete). States datasets by type of transportation (transit, driving, walking). 50 rows (states), 293 variables, 5 categorical variables and 288 numerical variables representing number of route requests for each state from January to October 2020.	Transit	Can look at transportation type preferences by county: walking, driving, transit. Identify zones with more foot traffic.

	2. Descartes Lab Mobility Change	https://drive.google.com/drive/folders/1oDDY1Vhpaxa29Iy1_Mj6zU1aZw6pEkGr	The distance a typical member of a given population moves in a day (kms). 2 types of datasets: **The median of the max-distance mobility for all samples(m50)** in the specified region and **the percent of normal m50 in the region, with normal m50 defined during 2020-02-17 to 2020-03-07 (m50_index).** Information broken into states, counties, dates. Number of datasets with county divisions: 2 and number of datasets with state division: 2 Datasets with county divisions: 3k rows, 247 variables. 4 categorical variables, 242 numerical m50/m50_index. Dates from 03-01-2020 to 10-30-2020. Datasets with states divisions: 51 rows, 247 variables. 3 categorical variables, 242 numerical m50/m50_index. Dates from 03-01-2020 to 10-30-2020.	Transit	Compare mobility before and after lockdowns in the different states for Milestone 1. Once we choose a state or city, use the mobility by county to identify the areas more affected by lockdowns (Milestone 2).
	3. Google Community Mobility Data	https://www.google.com/covid19/mobility/	Global information is broken down into counties. Keeps track of mobility changes as a percent. Industries tracked (6 different types of places): grocery and pharmacy, parks, parks, transit stations, retail and recreation, workplaces. 2 types of datasets: mobility by county and state. 51 rows, 258 columns, 2 categorical variables, 256 numerical variables	Transit	Can look at public transport changes, mobility for different purposes. The trends of mobility to groceries, parks, workplaces and residential give us indirect information of the potential flow of people to restaurants close to those areas.

			representing visits and duration of visits to different places between February and October 2020 compared to a baseline.		
	4. Foursquare Community Mobility Data	https://drive.google.com/drive/folders/1_IJWAae0MtYr7U_k3RUqBcaDhga3siRv	Visits, average duration in minutes and median visit length in minutes to 25 categories of places. Based on 13 million users from 01-01-2020 to 10-29-2020 by state. 52 rows, 302 columns, 2 categorical variables and 303 numerical variables representing the number of visits, average duration of visits and median duration of visits to Food stores and Food restaurants by state.	Visits, duration of visits to Food stores and Fast Food Restaurant.	Compare visits to food stores and fast food by states before and after the lockdowns (data available from January).
	5. Foursquare COVID-19 National and Regional	https://console.aws.amazon.com/dataexchange/home?region=us-east-1#/subscriptions/prod-hwaqvsrhtj7hm	AWS Data Exchange. Indexed foot traffic to 19 categories of venues. The indexed data is broken out geographically, with included data for National, SF, NYC, LA, and Seattle. The data is normalized against U.S. Census data to remove age, gender and geographical bias. Data is provided daily from 02/19/2020.	Updated daily foot traffic information splitting dining in casual and fast food restaurants (national level) and by city.	Compare visits to food stores and fast food by states during the entire year.
Household Pulse Survey		https://www.census.gov/data/experimental-data-products/household-pulse-survey.html	17 weeks from April 23 to October 26 2020. Surveyed people between 50k-100k per week. Variables between 82 to 188, mostly categorical. Missing data designed as -88 and -99. Require use of a data dictionary to translate the name of columns and categories. Demographics,	Affordability of food, free meals and spending use of the Economic Impact Payment	Recognize groups eligible for the social food programs but not included (insights for the National Association of Restaurants).
				Shopping and purchase preferences.	Shopping modalities, payment modalities, resumed/avoided eating at restaurants. Use of credit cards, apps to buy online. Consumer preferences (prepared food vs ingredients to cook at home)
				Trips and teleworking variables	Fewer transit trips, planned trips, trips to stores (give us information about likely to leave the home to buy meals vs use of delivery)

			spending, food, shopping, teleworking, trip trends variables. Dataset includes sub variables (secondary questions of the survey) which values depend on the answers to primary questions. In consequence, there are missing values in all the secondary variables and they will be removed during their specific analysis.		
COVID-19	The COVID Tracking Project	https://covidtracking.com/data/national/	COVID Data sets per location	States New tests Cases (confirmed plus probable) Negative PCR tests (people) Cumulative hospitalized/Ever hospitalized Currently hospitalized/Now hospitalized Deaths (confirmed and probable) Recovered Total test results. Do not require a cleaning process.	Connecting to Household expenditures and mobility patterns to understand the impact accordingly.
Restaurants	Yelp dataset	https://drive.google.com/drive/folders/1mp2texeym4VJbnPQFFFxnYyNFMjNRInu	Name, location (state and county), status (open, closed), attributes (take-out, outdoor dining, parking), categories (type of food), hours, stars, reviews.	Restaurant current status current services offered, location and popularity	Can determine how the restaurants were faring pre pandemic. Ratings and review count give us clues into how popular/competent these places may have been. We could also potentially find out more into how these restaurants responded to Covid (hours, takeout options)
	OpenTable dataset	https://drive.google.com/drive/folders/1BspmA9jUOuXjVOrTBeGa8-h5DSiZP3On	Sample of +20k restaurants across the country in the OpenTable network (online reservations, phone reservations and walks-in). States and metros with +50 restaurants on the OpenTable platform.	Tracking seated diners related to the same dates in 2019. Do not require a cleaning process.	Overall impact of COVID-19 in the industry showing year over year seated diners at a sample of restaurants.

	Foursquare + Apptopia	https://console.aws.amazon.com/dataexchange/home?region=us-east-1#/subscriptions/prod-wwjytnvaaazfq	National level, 30 weeks	Indexed foot traffic and app usage data for 37 retailers in the dining vertical from January 2019-July 2020. By combining Foursquare's location data and Apptopia's mobile app performance data we are able to understand how people are altering their dining habits since COVID-19. Year-over-Year visits indexed visit to restaurants, usage of app aggregated by category of food and individual dinings.	Overall analysis of usage of mobile app versus foot traffic for dining.

Table 1: Summarize data sources

Data Wrangling

1. Data Cleaning

Table 2 explores the data cleaning steps required in every dataset and how the methodology used assures that the Data is ready for the Exploratory Data Analysis. Data acquisition includes direct download of Excel, CSV and json files from the corresponding websites and web-scraping. Data cleaning incorporates the use of regular expressions, missing values exploratory methods consolidated in a Python script, extraction of the variables of interest per dataset, exploration of particular inconsistencies and development of specific methods according to the nature of the dataset.

Data is mostly numerical (GDP, Personal Consumption Expenditures, Unemployment, Sales, Inventories, Seated dining at restaurants, changes in mobility patterns, number of visits to places, duration of visits) except by the Household Pulse Survey (variables are categories representing answers to the survey) and Yelp Dataset (include categories of food, services, location).

Dataset	Cleaning steps	Why it is required?
Household Pulse Survey	Build sub-dataset of spending: extract EIP and EIPSPND variables over weeks and demographics	Expenditure patterns: Track percent change of people receiving EIP over weeks and its use by demographics

	Build sub-dataset of shopping variables over weeks and demographics: 1. extract CHNGHOW and WHYCHNGD variables. 2. extract FEWRTRIPS, FEWRTRANS variables 3. extract EXPNS_DIF: difficulty with expenses	1. Changes in shopping: purchases modalities, cash/credit card, avoid/resume dining in restaurants and reason. Track percent change over weeks and group by demographics. 2. transit trips and trips to stores: identify groups less likely to leave their homes 3. Relation between EIP and EXPNS_DIF
	Food Sufficiency over weeks and demographics 1. extract FOODSUFRSN (food sufficiency), FREEFOOD, WHEREFREE (free groceries), SNAP_YN, PRIFOODSUF.	The NRA is asking to expand the eligibility to RMP as part of SNAP. EDA related to the use of SNAP and restaurants struggles by state. How many people receive SNAP benefits? People that can't get out to buy or they are afraid. How many delivery services the city needs?
	Methods: 1. Incorporate age of the surveyed people, replace codes with nan values, drop duplicates in weekly analysis and include the dates of the survey. 2. Identification of missing values over rows and columns	Age instead birthday year for age groups analysis. Deal with NaN instead of numerical codes for null responses. Avoid duplicates surveyed people present in more than one week and use of dates for better reference.
Descartes Lab Mobility Change (Traffic)	1. Identification of missing values over rows and columns. Drop nan rows and columns in county and states datasets. 2. Extract STATE, COUNTY and m50/m50 index from 03-01-2020 to 10-30-2020. Drop the rest of the columns.	Data is going to be pivoted to visualize trends over the year by state. We'll identify states more affected for the lockdowns using m50_index and compare general trends using m50.
Foursquare Community Mobility Data (Traffic)	1. Identification of missing values over rows and columns. Drop nan rows (extra-row without values was removed). 2. Concatenation of the 6 datasets related to Food and Fast Food mobility.	Analysis of mobility related to Food stores and Food restaurants. Compare these results with m50 data.
Apple Mobility Reports (Traffic)	1. Identification of missing values over rows and columns. 4 states have missing values. We are going to fill them with the median of the rest of the states instead of dropping them. 2. Concatenation of the 3 datasets related to type of transportation.	Analysis of mobility related to type of transportation by state before and after the lockdowns. Compare these results with m50 data and foursquare trends.
Google Community Mobility Data	1.Concatenation of categories of places to generate one dataset by states and another by counties. 2. Identification of missing values over rows and columns over the two datasets. Drop unuseful columns but keep the gaps(missing values) in mobility by counties.	The dataset by states is going to be used to complete the analysis of mobility by state. The dataset by counties will be used in the Milestone 3, when we need more information by counties looking for restaurants near parks, groceries, transit stations, residential and workplaces.
Yelp Dataset	1. Take the 500k rows from businesses dataset and select ones categorized "restaurants" using string match. 2. Clean the reviews dataset by replacing the missing values 3. Use geographic data to find the state and county for each business in the business dataframe 4. Calculate review count and average stars using the reviews dataframe for each business in business dataframe 5. Look at unique values to get a sense of what our data is and what problems we may run into	1. We need to seperate restaurants from everything else 2. We need clean data to accurately sift through our findings 3. Need to find county and state to get a better idea on our trends based on location 4. Gives us a good idea on how said restaurants are faring or fared in terms of popularity and competency 5. We want to see if there are any issues with our data for later versions
Domestic Product and Income by Industry and Expanded Detail (Table 12)	1. Remove all data under quarters 2016 - 2017 as it's blank and not relevant. 2. Align data points according to universal columns for quarter and summarized year (q1 of 2018 - q3 of 2020)	Analysis of service of foods (i.e. restaurants and bars) vs other categories as part of GDP. Assessing trends q3 2016 - q2 of 2020. Seeing relationship to fishing and farming commodity impact to determine relationship.
Personal Consumption Expenditures by Major Type	1. Remove all data under quarters 2016 - 2017 as it's blank and not relevant. 2. Align data points according to universal columns for	See where personal consumption and capital has been spent over time (in particular before and after the impact of

of Product (Tables 1.5.2, 1.5.1, 1.5.3)	quarter and summarized year (q1 of 2018 - q3 of 2020) 3. Identification of inconsistent rows, relative to the other data sets, which are not relevant to the subject matter. Drop rows: [Percent change at annual rate:] and [Percentage points at annual rates:]	COVID. Correlate quarter results with that of GDP and income per industry. Review impact of imports vs exports as it pertains to business and service over same time series.
Income and Employment by Industry (until 2019)	1. Remove all data under quarters 2016 - 2017 as it's blank and not relevant. 2. Align data points according to universal columns for quarter and summarized year (q1 of 2018 - q3 of 2020)	This encapsulates the impact of COVID as it pertains to employment. Marrying income and employment numbers over the same time series with GDP determines how one affects the other. It can also determine the overall health (or risk) of the food services industry (in relation to others) and show impact for the need of relief for workers.
Unemployment by Month (Federal Reserve Bank of St Louis)	1. Import the data from source and download as an excel sheet 2. Replace all the column names with the name of the category they represent 3. Isolate the year and categories we want (Total Unemployment as well as Leisure and Hospitality for 2020)	The dataset was very clean to begin with but the column names had to be changed as before it was just ID numbers. Changing the names to what the column represented is necessary for readability. For Version One of the project we only need Leisure and Hospitality as well as total unemployment for 2020
Monthly Retail and Food Services Sales 1992-2020 (Annual Trade Survey)	1. Prep excel docs by formatting columns and rows for quick read_excel call by python. This included removing the irrelevant NAICS code, comments and white space. 2. Combine all sheets in the Excel data spread into one dataframe. Each sheet contains data for a year of sales in the various industries. 3. While combining, only add to dataframe relevant industries. Relevance is determined by proximity to the restaurant industry and thus includes grocery sales, etc. 4. Separate df into two new dataframes, one including annual sales, the other including monthly sales. 5. Drop columns containing all NaN values, otherwise ignore. This mostly was for the non-adjusted sales. 6. Reset Indices and transpose dfs to make the Kind of Business the key index and the columns be time.	Most values were already clean as this is a public, comercial dataset. However there was a lot of superfluous information as regards the scope of our project which needed to be cut out. Making two dataframes, one of monthly and the other annually, allows for better analysis of data from different perspectives. Transposing time to rows allows for better plotting of line graphs.
Foursquare + Apptopia	1. Join Datasets with aggregate information by type of food. 2. Join Datasets with information by individual dinings 3. Analysis of missing data	Year over year analysis of restaurant app usage vs foot traffic to analyze correlation between app usage and foot traffic in different dinings and types of food.

Table 2: Data cleaning process by dataset

The following diagrams show the methods applied in every dataset.

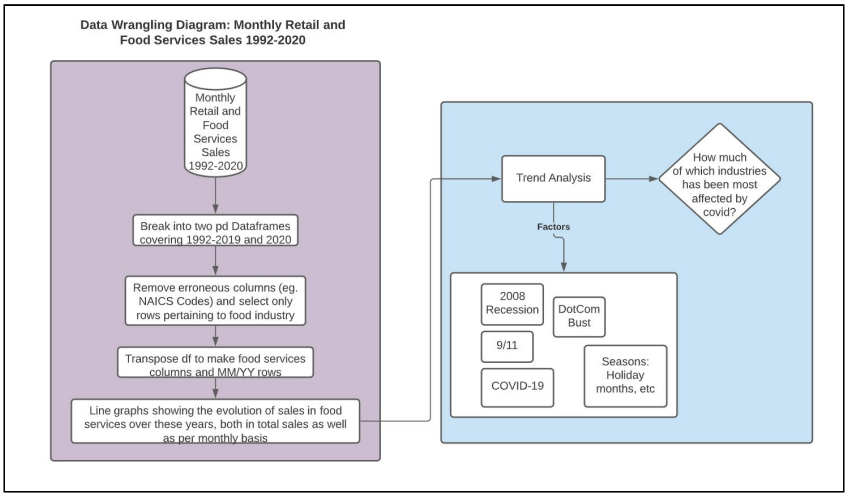

Diagram 1: Data Cleaning Process of Annual Trade Survey

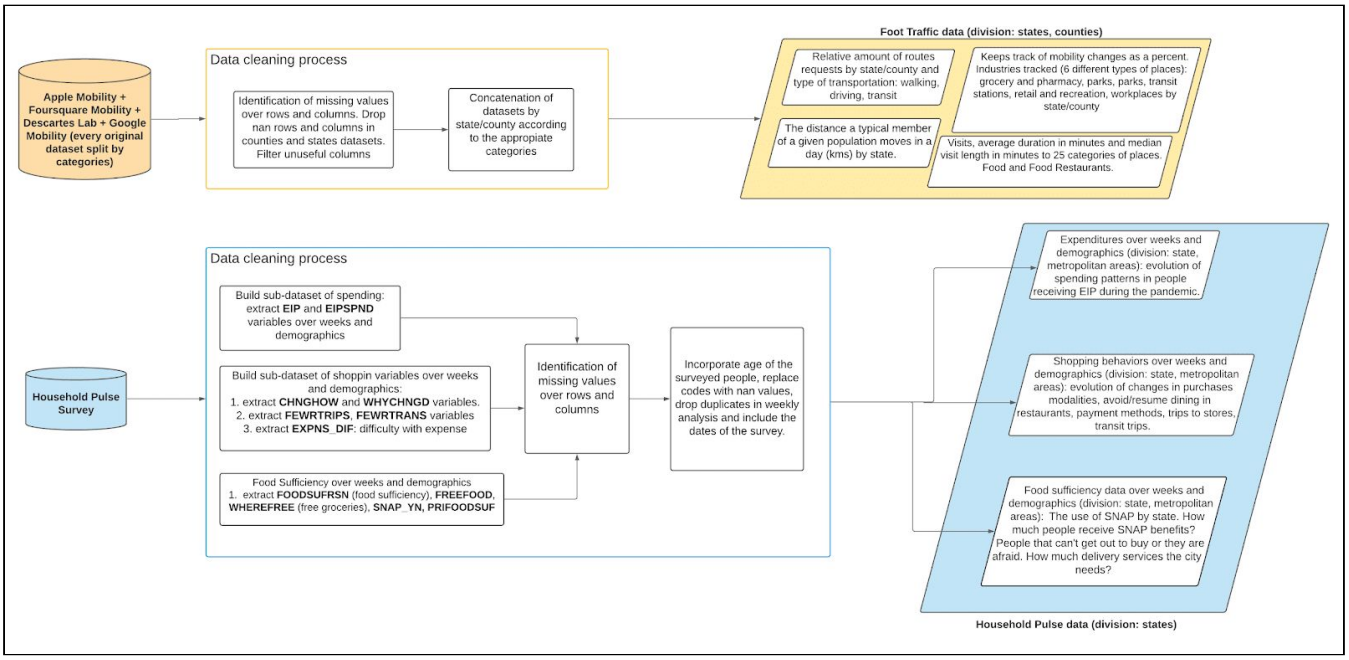

Diagram 2: Data Cleaning Process of Household Survey and Mobility Datasets

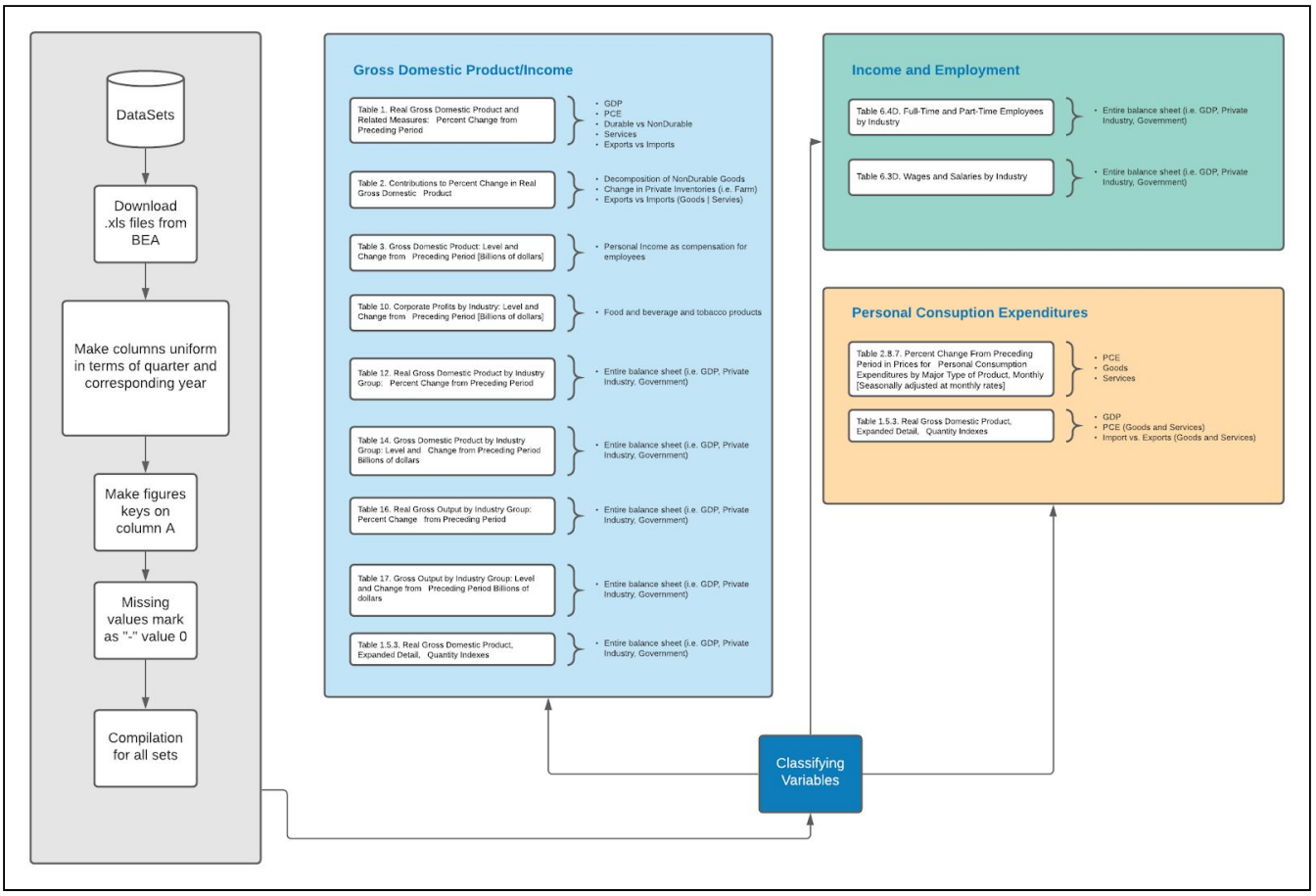

Diagram 3: Data Cleaning Process of BEA Datasets

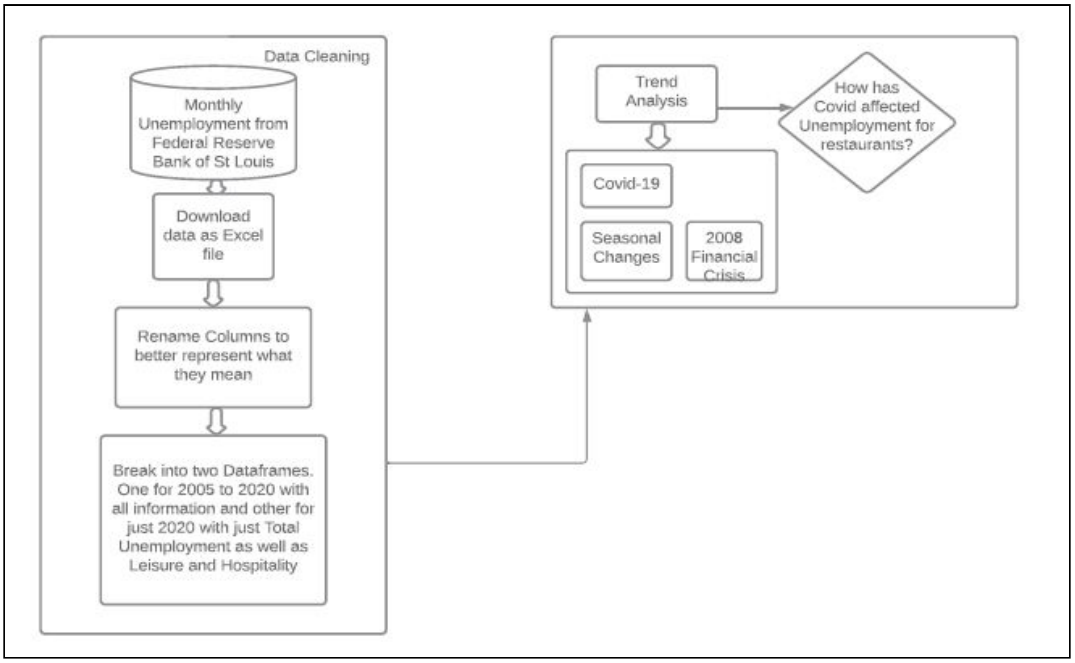

Diagram 4: Data Cleaning Process of Unemployment

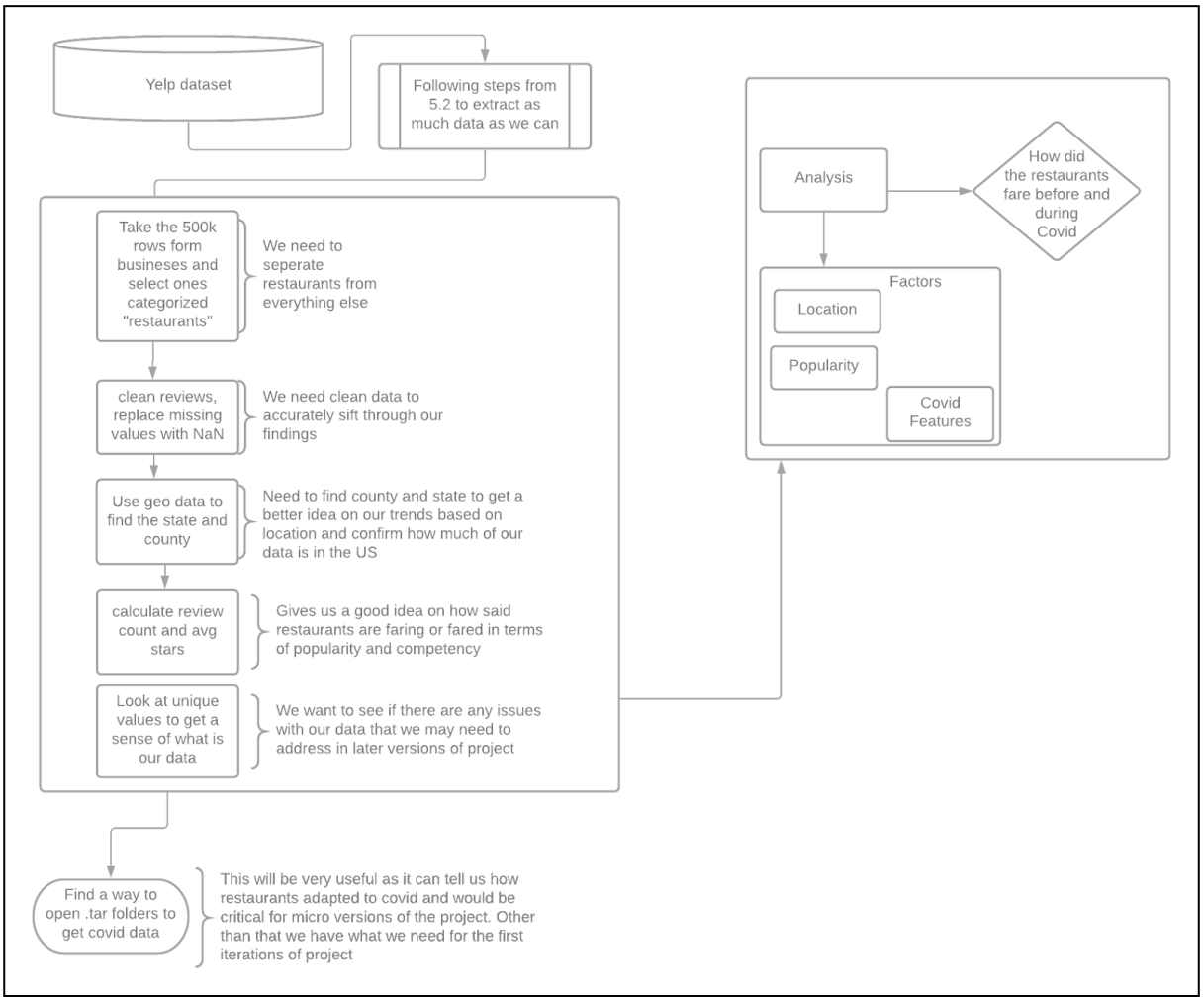

Diagram 5: Data Cleaning Process of Yelp dataset

2. Data Structuring

The diagram 6 explores how datasets are going to be used to achieve the Milestone 1, which is an overall effect of the pandemic over the industry during this year from an economic perspective (tracking GDP variables, sales and inventories from the food industry as goods and services and employment variables) and a consumer perspective, looking for changes in consumer behaviours, mobility patterns related to the acquisition of food (as good and services). This macro analysis pushes us to Milestone 2, where we will choose a specific geographic region to find out the status of restaurants and mobility patterns within specific places in the city and counties.

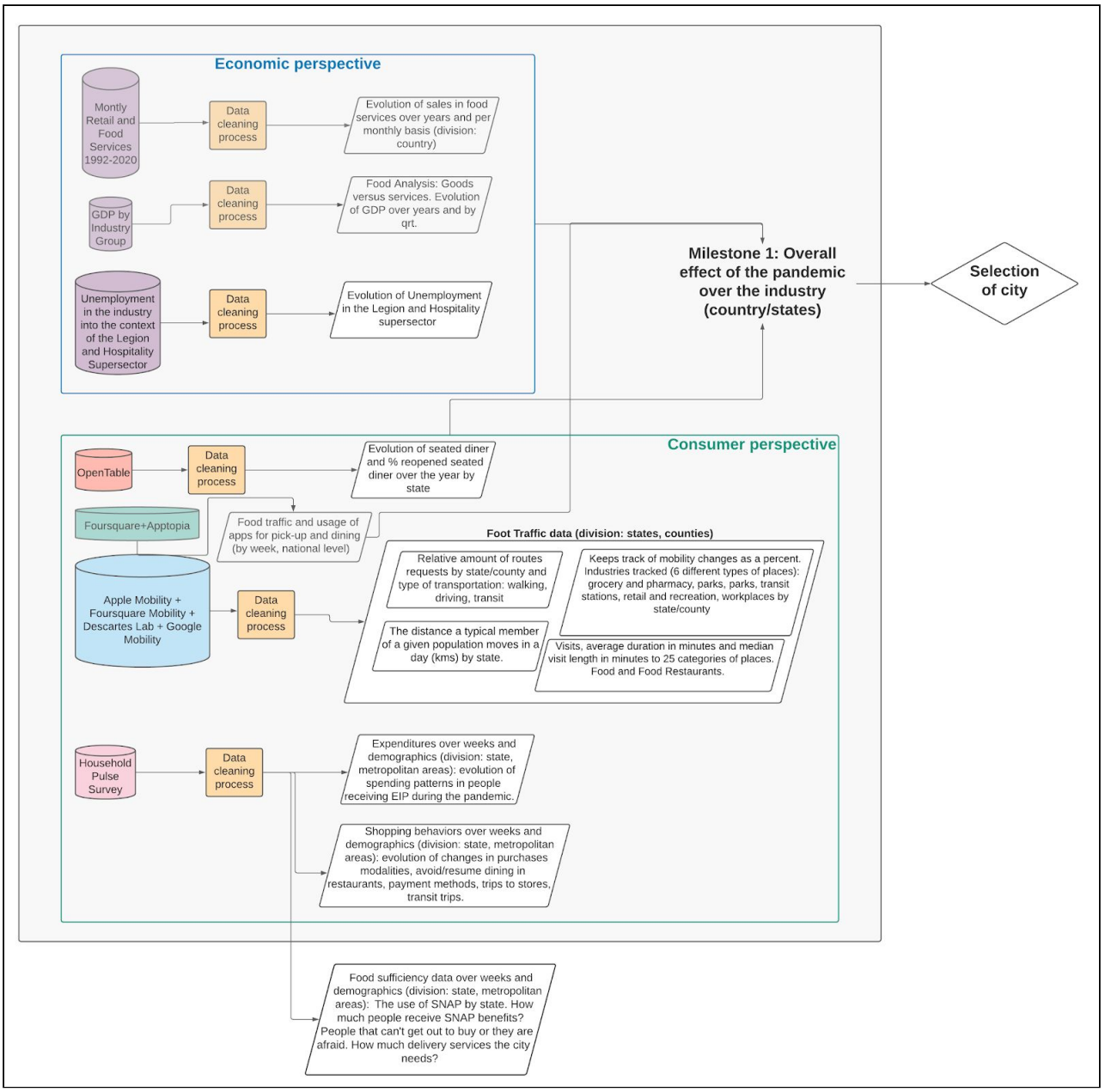

Diagram 6: Structuring Datasets for EDA

The project is available in the following repository: [DS4A2020_Empowerment]. We are using .gitignore to list the large datasets and to avoid exceeding the GitHub file size limit of 100MB. The folder data contains subfolders raw (data previous manipulation) and interim (files after data cleaning steps and ready to the Exploratory Data Analysis). Raw is split into 4 categories of data: economics, restaurants, mobility and census. Every dataset is pre-processed as needed and then a clean version is saved in the folder Interim, which has 4 categories as well as Raw. Mobility and Census (raw and interim) are included in .gitignore because the extension of the files. However, the Jupyter notebooks of census data were created considering this issue and we incorporated a data acquisition notebook to download

directly the census datasets from the website through web scraping, storage locally and after a first glance of cleaning steps, save the useful structured files in the interim folder. In this way, every user can clone the repository and replicate the process. To get access to the full content of datasets, visit our directory in Google Drive. This directory contains exactly the same folders and structure of the GitHub repository, without the storage restrictions.