QM 7103 Analytics

Data Mining and Analysis Project

Dig Inn 'to the' Data

Presented by

Nigar Mutallimova, MSBA

nim3091@utulsa.edu

Jim Sill, PhD Cyber Fellow

jjs2819@utulsa.edu

QM 7103 Analytics

Data Mining and Analysis Project

Introduction

As a part of this class, we were to find, qualify, and select a dataset that was able to be mined and analyzed, yielding specific analytical information, which could be used to form marketing decisions, and / or possible forecasting capacities. This data could be in any format, that could be transitioned into any usable format, within the "R Studio" or "Jupiter Lab" tools utilizing r code, and python, for the analysis.

Nigar and I utilized a dataset that encompassed a health food restaurant chain in New York City. The data was for the calendar year of 2018, and included three csv files, which contained the restaurant identifying information; the restaurant sales; and the restaurant menu items. The name of the restaurant chain's name is / was "Dig Inn", they are still in business and their current menu items, ordering information, and restaurant locations can still be found at DIG | Seasonal, fresh, local food (Dig Inn) . Dig Inn has been in business since 2008, and it has a loyal following, including metropolitan workers, college students, and young people from 20-35 years old. Dig Inn's menu is considered to be pricy in comparison to 'fast food' like McDonalds, or Burger King, but affordable, when compared to the traditional 'sit-down' restaurant, serving healthy options for a center of the plate meal. *(use Pei-Wei vs PF Chang – same owners, same food almost with PW being slightly cheaper*) Dig Inn had eight physical locations, with over 100 combinations possible, from the menu items that can be ordered, for in-house dining, to go, or for delivery.

The raw data consisted of more than 2.5M individual data points, our goal for this data was to mine and analyze it, to find out what the restaurant's gross sales numbers were, by the day, by the hour, and by the location. We also wanted to know what the major sellers were, what food item(s) were the main sales, and if it differentiated between each restaurant location. In the review of the data, most of the menu items were healthy choices, however a couple did not seem to fit, as they were 'comfort foods' and not 'health foods', this gave us another viewpoint, and we wanted to see if their comfort food offerings such as cookies, and mac & cheese, were just trivial menu items, for the occasional purchase or if they were actually substantial to the corporate sales revenue model.

Raw Data Metrics

All of the more than 2.5M raw datapoints were comprised of three of the data files, composed within a *.csv file format. I have included screenshots of raw data files used in for the mining and analysis once the data was curated. In each of the raw csv datasets there were approximately 5% voids / null values or there was also corrupted data in some cells. Each of these datasets had the ability to be joined using outer, or inner joins, by either sales order number, or by restaurant identification number, or restaurant name.

QM 7103 Analytics

Data Mining and Analysis Project

ORDER_ID	DATETIME	RESTAURANT_ID	TYPE	DRINKS	COOKIES	MAIN	BASE	SIDE_1	SIDE_2	MAIN_NAME	BASE_NAME	SIDE_1_NAME	SIDE_2_NAME
O689957	4/12/2018 13:14	R10002	PICKUP	0	2	I1	I7	I11	I8	Spicy Meatballs Marketbowl	Farro with Summer Vegetables	Charred Broccoli with Lemon	Cauliflower with Garlic and Parmesan
O1497863	8/21/2018 18:14	R10007	IN_STORE	1	0	I1	I5	I10	I9	Spicy Meatballs Marketbowl	Classic Brown Rice	Roasted Sweet Potatoes	Jasper Hill Mac & Cheese
O1443303	8/12/2018 18:04	R10006	DELIVERY	0	0	I1	I6	I10	I11	Spicy Meatballs Marketbowl	Farm Greens with Mint	Roasted Sweet Potatoes	Charred Broccoli with Lemon
O2092404	11/19/2018 12:10	R10008	IN_STORE	0	0	I1	I5	I11	I9	Spicy Meatballs Marketbowl	Classic Brown Rice	Charred Broccoli with Lemon	Jasper Hill Mac & Cheese
O1382733	8/1/2018 20:53	R10008	IN_STORE	0	0	I0	I6	I12	I12	Charred Chicken Marketbowl	Farm Greens with Mint	Cashew Kale Caesar	Cashew Kale Caesar
O1642363	9/14/2018 13:59	R10007	IN_STORE	0	0	I3	I6	I12	I15	Herb Roasted Chicken Marketbowl	Farm Greens with Mint	Cashew Kale Caesar	Snap Peas

Figure 1 Orders dataset: orders.csv [2.5M vectors/datapoints (13 Columns)]

location	day	number of orders	percentage of deliveries
Bryant Park	1/1/2018	394	0
Bryant Park	1/2/2018	807	0
Bryant Park	1/3/2018	744	0
Bryant Park	1/4/2018	791	0
Bryant Park	1/5/2018	677	0
Bryant Park	1/8/2018	856	0
Bryant Park	1/9/2018	897	0

Figure 2 Restaurant Location: rest_location.csv [28.5K vectors/datapoints (4 Columns)]

RESTAURANT_ID	NAME	ADDRESS	LAT	LONG	OPENING_DATE	DELIVERY_START
R10001	Columbia	2884 Broadway, New York, NY 10025	40.81147	-73.96123	8/9/2014	1/1/2017
R10002	Midtown	1379 6th Ave, New York, NY 10019	40.76364	-73.97796	3/19/2013	5/1/2018
R10003	Bryant Park	70 W 40th St, New York, NY 10018	40.752911	-73.983498	5/21/2013	5/1/2018
R10005	Flatiron	40 W 25th St, New York, NY 10010	40.7436	-73.99107	11/14/2013	3/5/2016
R10004	NYU	109 Macdougal St, New York, NY 10012	40.72993	-74.00082	1/10/2014	1/1/2017
R10006	Upper East Side	1045 Lexington Ave, New York, NY 10021	40.77201	-73.96078	5/29/2014	8/2/2017
R10007	Upper West Side	2140 Broadway, New York, NY 10023	40.77543	-73.98205	2/2/2015	8/2/2017
R10008	Williamsburg	45 S 3rd St, Brooklyn, NY 11249	40.713749	-73.965782	10/12/2015	1/1/2017

Figure 3 Restaurant Information: restaurant.csv [8 vectors/datapoints (7 Columns)]

RESTAURANT_ID	ORDER_ID	DATETIME	TYPE	DRINKS	COOKIES	MAIN_NAME	BASE_NAME	SIDE_1_NAME	SIDE_2_NAME	NAME	ADDRESS
R10001	O2226026	12/7/2018 19:17	PICKUP	2	2					Columbia	2884 Broadway, New York, NY 10025
R10001	O1734393	9/28/2018 14:41	PICKUP	0	0	Herb Roasted Chicken Marketbowl	Farm Greens with Mint	Snap Peas	Snap Peas	Columbia	2884 Broadway, New York, NY 10025
R10001	O1638120	9/13/2018 18:43	PICKUP	0	0	Herb Roasted Chicken Marketbowl	Farm Greens with Mint	Cashew Kale Caesar	Green Goddess Beans with Sesame	Columbia	2884 Broadway, New York, NY 10025
R10001	O2286140	12/16/2018 17:27	IN_STORE	0	0	Grilled Organic Tofu Marketbowl	Farm Greens with Mint	Roasted Sweet Potatoes	Roasted Sweet Potatoes	Columbia	2884 Broadway, New York, NY 10025
R10001	O1774800	10/4/2018 15:31	IN_STORE	0	1	Charred Chicken Marketbowl	Classic Brown Rice	Roasted Sweet Potatoes	Jasper Hill Mac & Cheese	Columbia	2884 Broadway, New York, NY 10025
R10001	O1722050	9/26/2018 18:04	IN_STORE	0	0	Herb Roasted Chicken Marketbowl	Farm Greens with Mint	Charred Broccoli with Lemon	Cauliflower with Garlic and Parmesan	Columbia	2884 Broadway, New York, NY 10025
R10001	O734903	4/18/2018 20:40	IN_STORE	0	0	Charred Chicken Marketbowl	Classic Brown Rice	Roasted Sweet Potatoes	Jasper Hill Mac & Cheese	Columbia	2884 Broadway, New York, NY 10025
R10001	O1421215	8/8/2018 18:30	IN_STORE	0	1	Spicy Meatballs Marketbowl	Farm Greens with Mint	Cashew Kale Caesar	Blistered Shishitos	Columbia	2884 Broadway, New York, NY 10025
R10001	O499284	3/15/2018 17:08	IN_STORE	0	0	Herb Roasted Chicken Marketbowl	Farm Greens with Mint	Roasted Sweet Potatoes	Charred Broccoli with Lemon	Columbia	2884 Broadway, New York, NY 10025

Figure 4 Final Order Data: orders_final.csv [2.39M vectors/datapoints (12 Columns)]

Data Preparation and Curation

This data in its raw form was compiled within three different comma separated value formatted files. The files would not read into / import into python easily, so multiple manipulations using the R Studio import csv was utilized to clean and curate the raw data. These files were too large to do so within Microsoft Excel, without further corruption occurring. In doing this 'cleaning-up' removing 'voided' or 'empty cells' was done, as was the removal of a leading column which was an alphanumeric index for the data, which also had corrupted data in it. Since this column would not be utilized once read into either R or Python, the choice was to just drop the indices was made.

In the utilization of R Studio, for curation of the data, we did also struggle with the importing of the largest of the dataset. A simple fix was found, but it did cause a bit of frustration until it was found. The use of a single " \ " was being used in either Python or R Studio, however to complete the importation of the larger dataset, a double " \\ " had to be used, this syntax is noted on the R Studio documentation, but not within the Python documentation. Once this simple change was made, the larger dataset was imported with no further issues. Furthermore, after the import was completed, we utilized the dyplr library to continue to manipulate the data, by creating inner joins, and dropping additional data that was not going to be used or could be a source of issues later.

Upon completion the final dataset was exported using the R Studio, back into a 'Cleaned' and 'Manipulated' *.csv dataset. This final dataset was distributed to both myself and

to Nigar. As we had divided the tasks of manipulation and analysis, using the analytics tools evenly between us. Having a unified dataset was paramount in this, or we ran the risk of having a poor representation of the end results or having improper analysis.

Nigar would continue to use the R Studio and the data analytics tools within that framework to analyze and make visualization outputs from the data. Jim would do the same utilizing the Python framework and its data analytics tools. At the end of the processing and manipulation, there were 2.394M data points to analyze. These ~2.4M data points comprised over 30 unique variables, and 14 various columns, outside of any created by Nigar or Jim, in the analysis of the data. This dataset created the ability to have 100's of combinations, and uniqueness within its totality.

Data Mining and Analysis

We have listed simply some of the findings from our data analysis below, each point having been derived from either r code, or python, being visually represented and having the ability to be interpreted as follows:

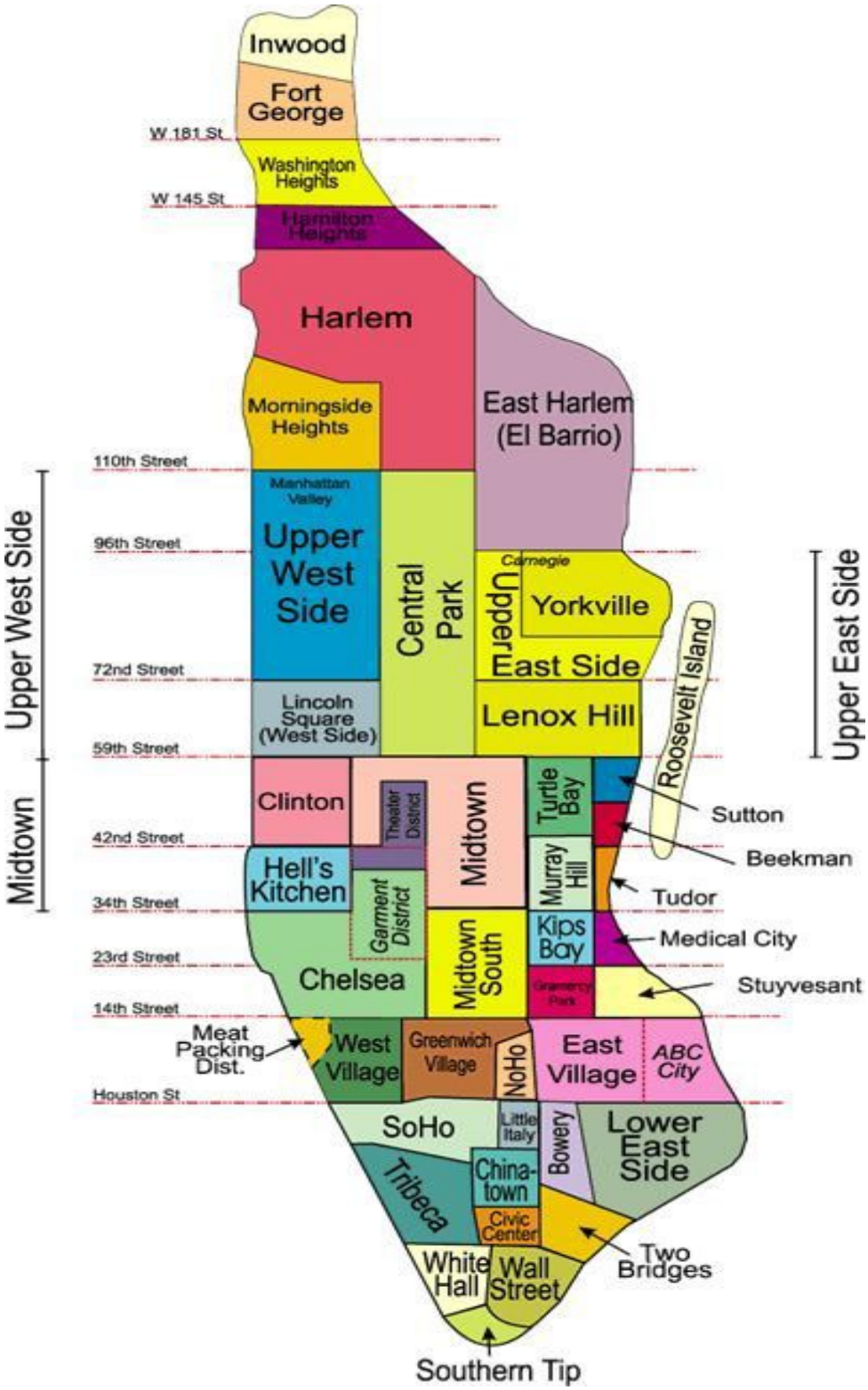
- The least number of orders was made during the early morning 9AM – 10AM and late evening 11PM.
- The highest number of orders was observed during lunch time, from noon to 2PM.
- At 5PM most popular dish was Spicy Meatballs.
- At 3PM it was Herbed Chicken and Grilled Tofu.
- the highest number of deliveries was observed at 6PM, and at 2PM
- Most orders were dine in orders, which comprised 72% of all orders.
- Delivery orders comprised 11% of all orders placed, and most deliveries took place near the NYU Campus store location.
- Pick-up orders comprised 17% of all orders placed, and most pick-up orders took place near Bryan Park and Columbia store locations.
- The number one selling side dish is “Mac & Cheese”, and its market share makes up 25% of all orders, or 1.2M individual orders. The second highest selling side dish is the roasted sweet potatoes, comprising 17% of all side dish sales.

- As a health food restaurant, the base ingredient for the bowl is the farm greens and brown rice, for these top selling healthy bowls, their market share is 43% and 33% respectively.
- The 'Protein' of choice dominating almost 60% of all orders placed was chicken, either charred (blackened) chicken, or herb roasted chicken. The second choice of protein was spicy meatballs, comprising 26% of all orders.
- The highest grossing store location for Dig Inn, was the NYU location, having 18% of all orders, from all ten locations. Bringing in the rear was Bryant Park, having only a 9.6% market share of sales corporately.
- Drinks besides water, constitute only 15% of all orders.
- Cookies, while on the menu, only constitute 20% of all orders. **I guess people who do eat at these 'health' food restaurants do eat healthily.*

As you can see, we focused on the sales aspect of the data, determining the percentages of sales by menu item, store location, type of dish or side, date, and time of day. An interesting data point that was not available for review, would have been the actual customer information, as it would have granted further insight into the types of client base the locations are selling to. Nigar pulled some demographic data from the NYC statistics, and we did presume to ask some interesting questions, based upon this data, but these queries were no more the conjecture, as the hard data was not there, or capable of being combined with the raw data we selected. The factual or hard results are what is reported upon in this document.

Other Data Curated

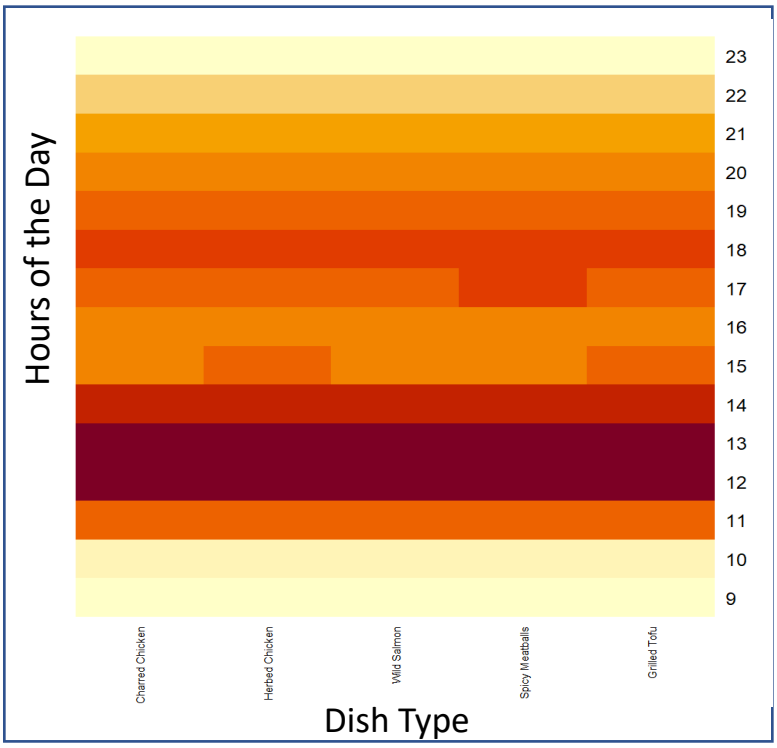
We did review other demographic data, and use it in our own analysis, which will be verbally portrayed in our MS Power Point presentation, and executive overview, but the for the most part, among all NYC neighborhoods, delivery option is the most popular in Upper East and Upper West Side while it's the least used in Midtown and Bryan Park. Interestingly, in the Upper East Side, dine-in is the least preferred option for the guests, which can be explained by posh lifestyle and that area traditionally being home to some of the world's most wealthy, powerful, and influential families. It's possible that elite of that area does not like being noticed. Customers of the restaurant's NYU branch do not prefer Delivery orders and dine-in or pick up their meal. Probably, the reason for that is that the population in that area are mostly students and it's inconvenient to deliver their food to the university. On the other hand, customers in Midtown and Williamsburg, which is known as one of the most traditional neighborhoods with more than 35% of Hasidic Orthodox population, mostly do not opt for delivery option and eat either in-store or pick up their meal. Overall, the map can tell us that most of the orders are made around NYU, in the Upper West and Upper East Side, while the least number of orders are made around Bryan Park and Columbia.

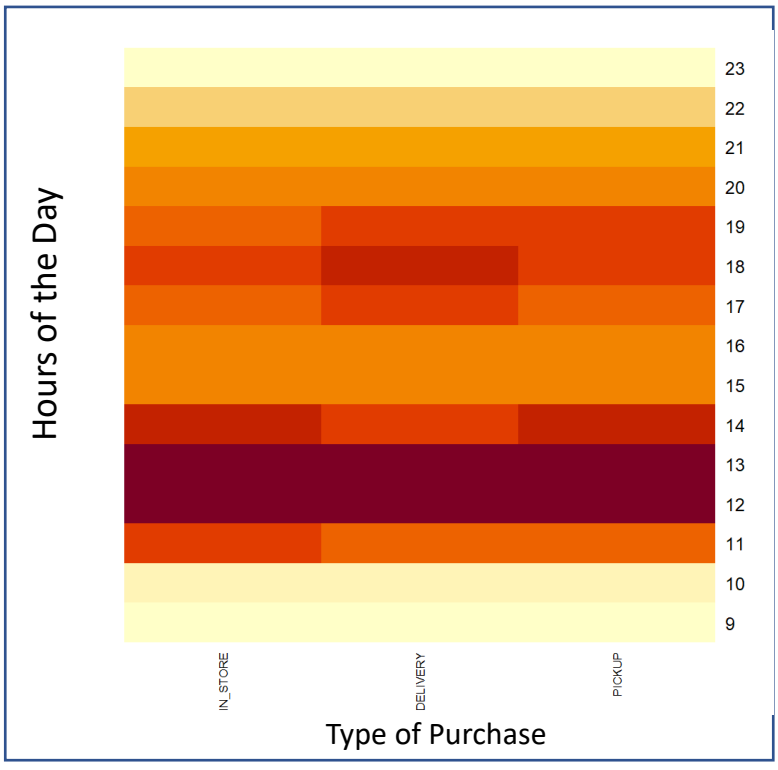

Inwood
Fort George
W 181 St
Washington Heights
W 145 St
Hamilton Heights
Harlem
Morningside Heights
East Harlem (El Barrio)
110th Street
Manhattan Valley
Upper West Side
96th Street
Central Park
Carnegie
Upper East Side
Yorkville
72nd Street
Lincoln Square (West Side)
Lenox Hill
59th Street
Roosevelt Island
Upper West Side
Midtown
Clinton
Theater District
Midtown
Turtle Bay
Sutton
42nd Street
Hell's Kitchen
Garment District
Murray Hill
Beekman
Tudor
34th Street
Kips Bay
Medical City
23rd Street
Chelsea
Midtown South
Gramercy Park
Stuyvesant
14th Street
Meat Packing Dist.
West Village
Greenwich Village
NoHo
East Village
ABC City
Houston St
SoHo
Little Italy
Bowery
Lower East Side
Tribeca
China-town
Civic Center
Two Bridges
White Hall
Wall Street
Southern Tip

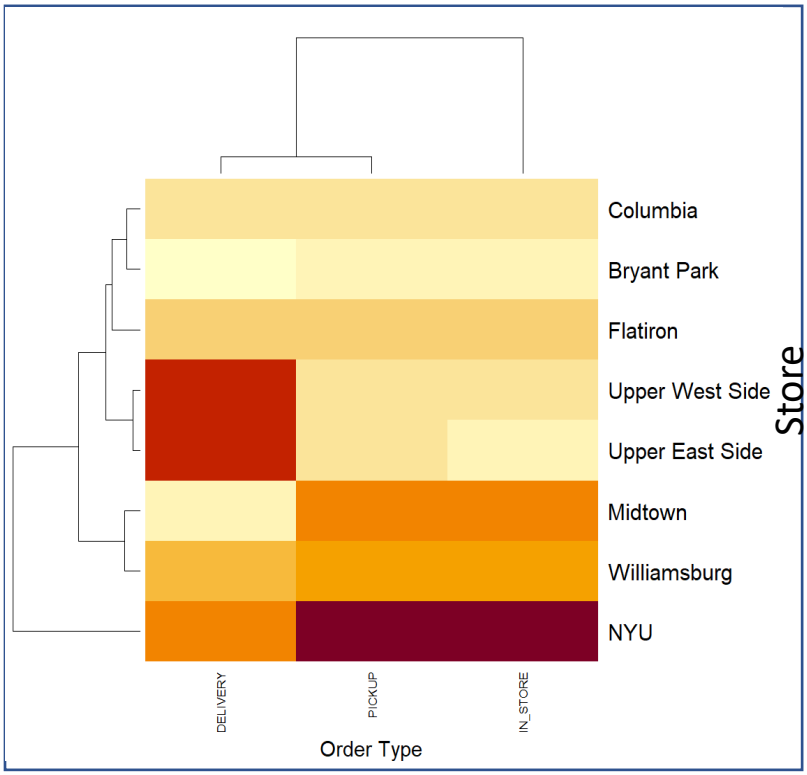
Conclusions

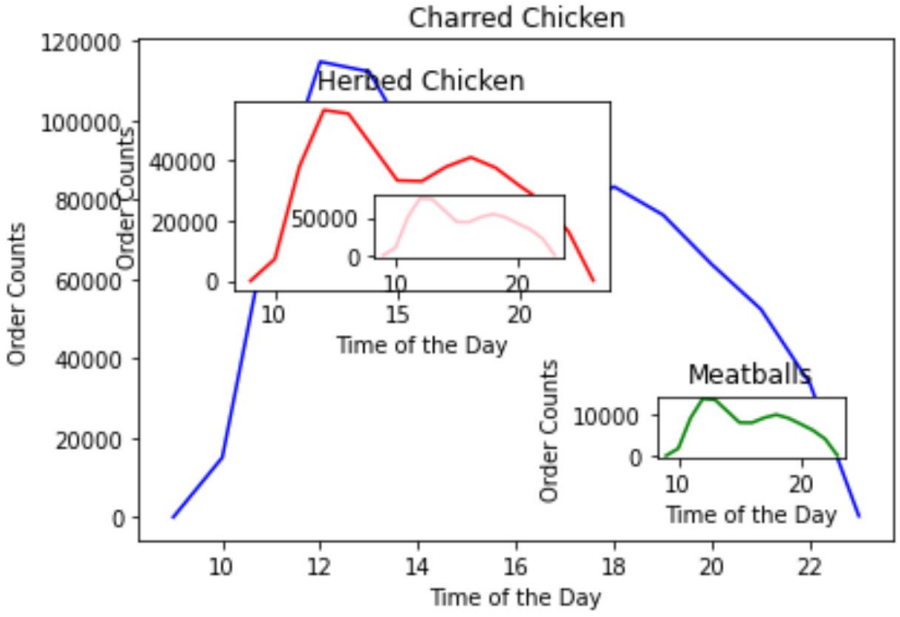
This data was fun to work with, and it did yield some good results, I believe that the use of predictive analysis could be applied to this data, once more of a historical bundle was created. A single year's worth of data is nice to look at, and you could ultimately project, the next month or possibly quarter's expected sales, by location, even by the hour, but to expand this across the horizon to be used as an annual forecasting model, one would need more data. It was apparent that certain items did do better than others and should Dig Inn want to capture or maintain a higher sales volume, items of that flavor profile or texture could be experimented with and substituted into the menu to replace more poorly performing items. Seasonality did not seem to affect this chain's sales, but certain locations did outperform others, during certain times of the day. Furthermore, actual sales numbers would have been a good thing to have. As it was, we only had a 'count' of sales, not a 'sales' number. This specific dollarization data would have yielded more insights, helping locations, and product mix projections developing and projecting net yield or profit and loss by item, by store location. This information along with demographic information would be key in determining the true outcome of the sales projections, and possible menu changes for the restaurant chain. Overall, in our opinion this was nice and insightful data to work with.

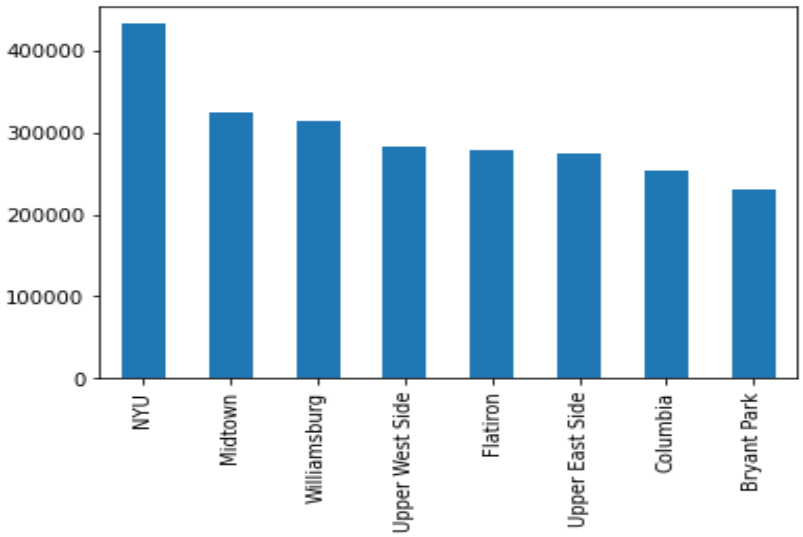
Charts

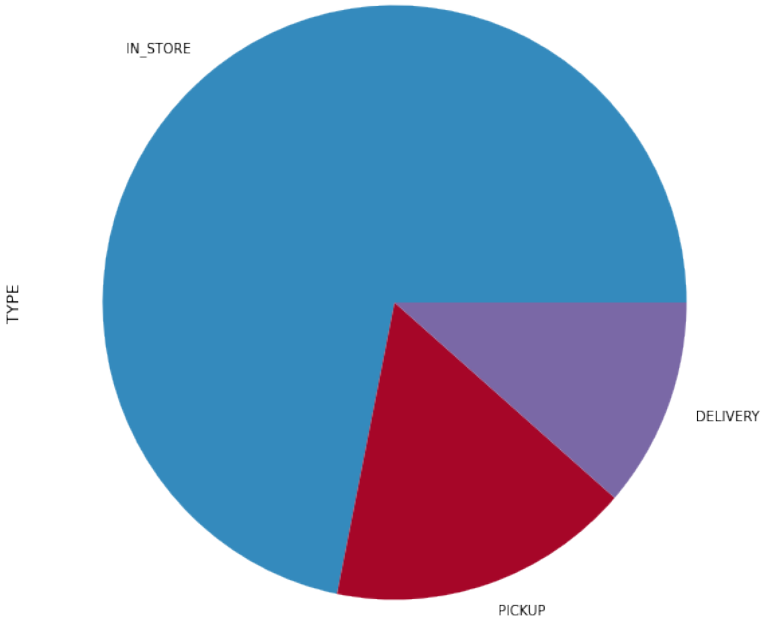

Hours of the Day
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
Charred Chicken
Herbed Chicken
Wild Salmon
Spicy Meatballs
Grilled Tofu
Dish Type

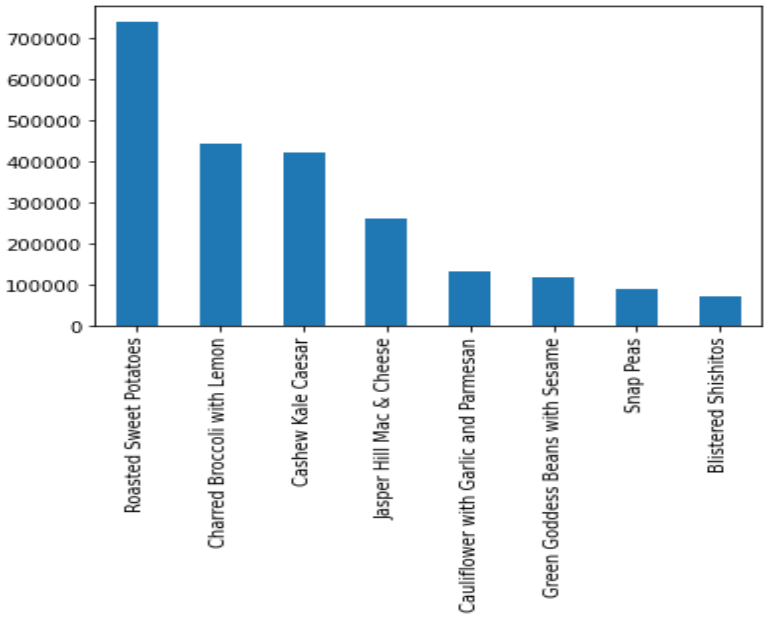

Hours of the Day
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
IN_STORE
DELIVERY
PICKUP
Type of Purchase

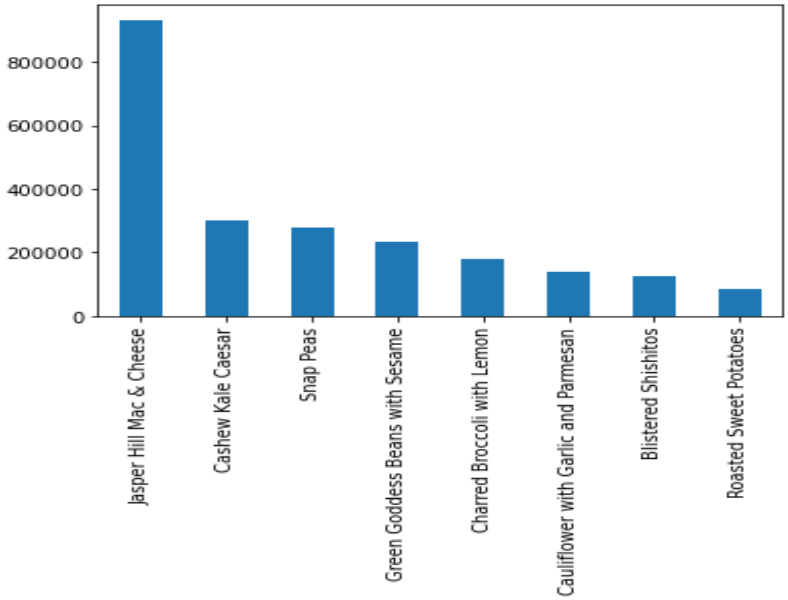

Columbia
Bryant Park
Flatiron
Upper West Side
Upper East Side
Midtown
Williamsburg
NYU
Store
DELIVERY
PICKUP
IN_STORE
Order Type

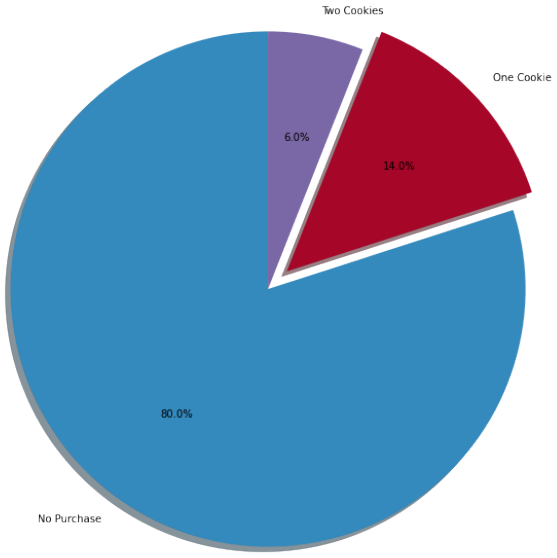

Charred Chicken
Herbed Chicken
Meatballs
Order Counts
Time of the Day

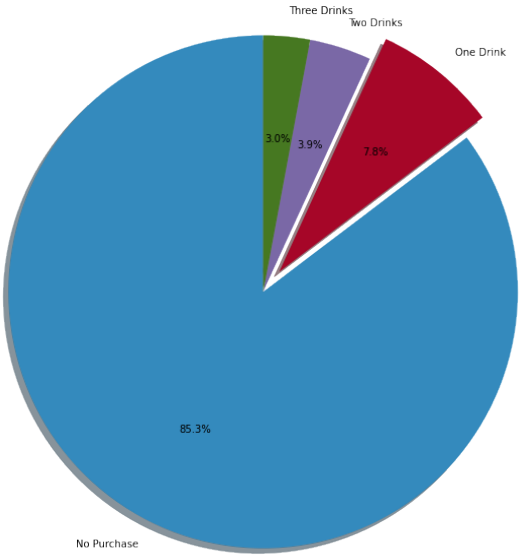

400000
300000
200000
100000
0
NYU
Midtown
Williamsburg
Upper West Side
Flatiron
Upper East Side
Columbia
Bryant Park

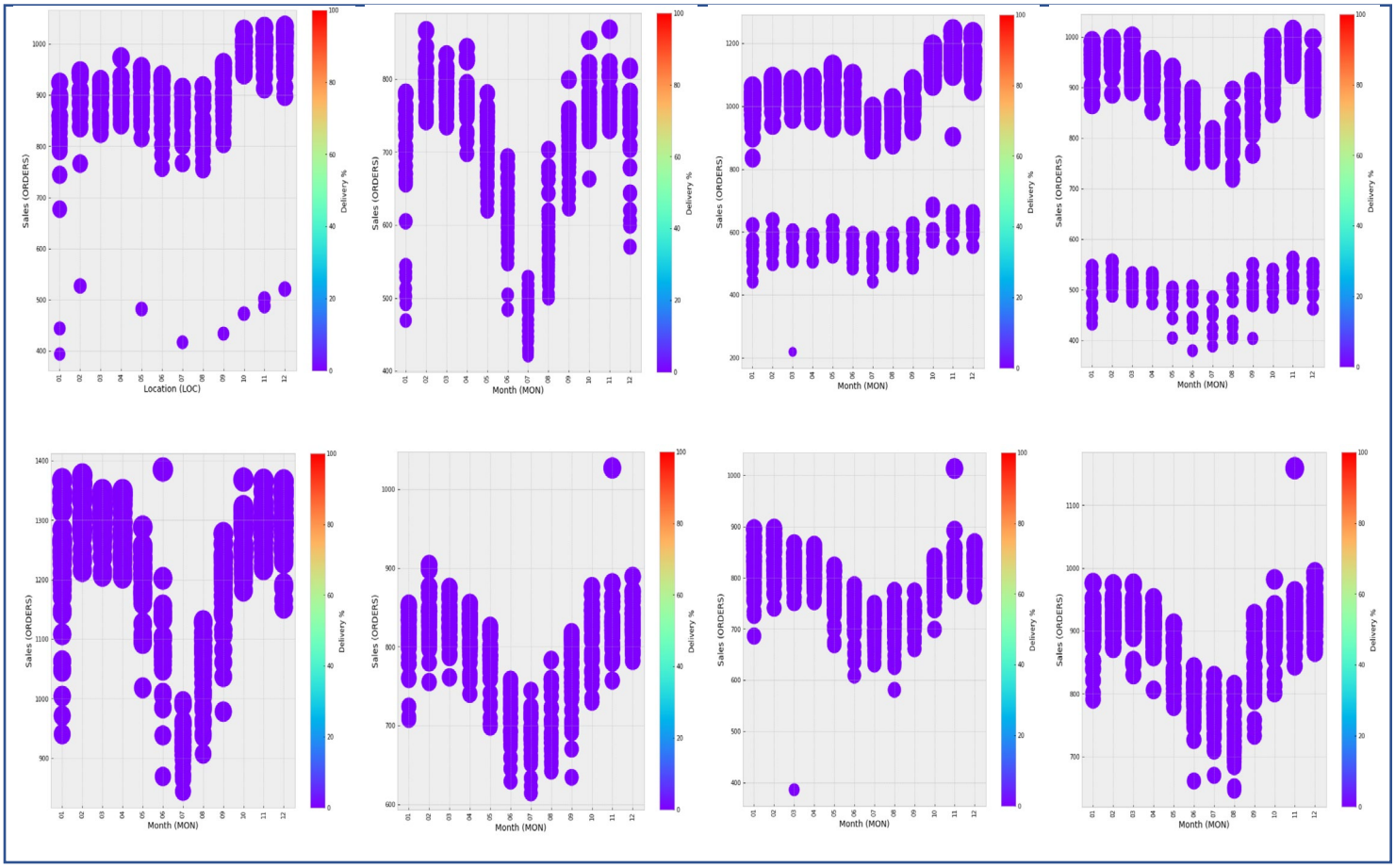

IN_STORE
TYPE
DELIVERY
PICKUP


700000
600000
500000
400000
300000
200000
100000
0
Roasted Sweet Potatoes
Charred Broccoli with Lemon
Cashew Kale Caesar
Jasper Hill Mac & Cheese
Cauliflower with Garlic and Parmesan
Green Goddess Beans with Sesame
Snap Peas
Blistered Shishitos


800000
600000
400000
200000
0
Jasper Hill Mac & Cheese
Cashew Kale Caesar
Snap Peas
Green Goddess Beans with Sesame
Charred Broccoli with Lemon
Cauliflower with Garlic and Parmesan
Blistered Shishitos
Roasted Sweet Potatoes


Two Cookies
One Cookie
6.0%
14.0%
80.0%
No Purchase


Three Drinks
Two Drinks
One Drink
3.0%
3.9%
7.8%
85.3%
No Purchase


Sales (ORDERS)
Location (LOC)
Month (MON)
Delivery %