Report: Food Resources in Relation to Demographics

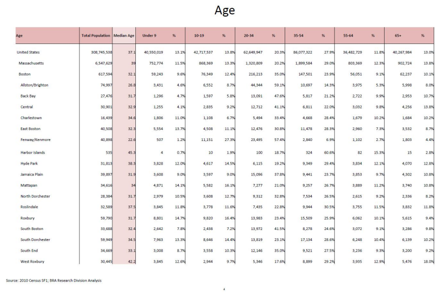
Johnson Lam and Katie McKay

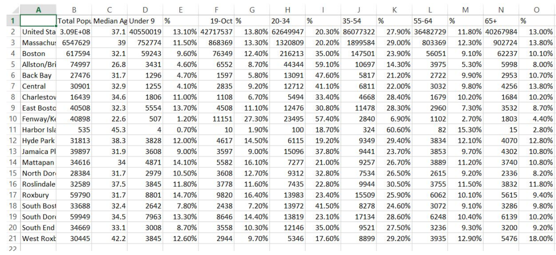
Introduction describing problem

The goal of our project was to find areas in Boston with food desserts, areas that lack food resources, and attempt to find a relationship to the demographics of Boston. The motivation behind our project is that if such a correlation was to found, we could use it as a predictor to determine neighborhoods prone to having food desserts. This could help cities better identify such areas lacking food resources and could aid the city in providing more food resources to locals in this area. Such issues currently exist in other cities where people have to spend long commutes to buy groceries.

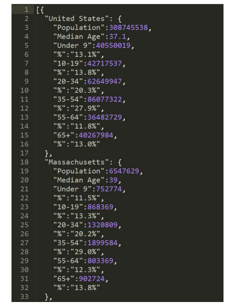
For food resources we used the following datasets: Active Food Establishment Licences: the locations of all the restaurants in Boston, Summer Farmers Markets: the locations and availabilities of all the farmers markets in the Boston Area, Corner Stores: the locations of convenience stores and small markets in Boston, Food Pantries: the locations of all the food banks and food pantries in Boston. This was obtained from the Boston Data Portal. These datasets contained contained location data in various forms and had to be standardized. Once standardized, an API call was made to Google maps to determine the neighborhood in which the food resources was located in. When these were obtained, the relational data paradigm was used to aggregate the data and count the number of food resources located in that area.

For data on the Boston demographics, we obtained the datasets Boston In Context: Neighborhoods. 2007-2011 American Community Survey, 2010 Census and Boston In Context: Planning Districts. 2007-2011 American Community Survey, 2010 Census from the City of Boston Data Portal. These two datasets were extremely identical, except that the Planning Districts dataset had finer granularity and described a few more areas than the Neighborhoods dataset. Otherwise, the datasets were nearly identical in the type of demographics they described, eg. age, income, ethnicity, etc. We decided to use the Neighborhoods dataset in order to more easily categorize the neighborhoods food resources were located at. To better utilize the dataset, I wrote code to transform it from a pdf format, to a text format, to a csv file, to a json file. The following depicts the transformation of the format.





	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		Total Popu	Median Ag	Under 9	%	19-Oct	%	20-34	%	35-54	%	55-64	%	65+	%
2	United Sta	3.09E+08	37.1	40550019	13.10%	42717537	13.80%	62649947	20.30%	86077322	27.90%	36482729	11.80%	40267984	13.00%
3	Massachus	6547629	39	752774	11.50%	868369	13.30%	1320809	20.20%	1899584	29.00%	803369	12.30%	902724	13.80%
4	Boston	617594	32.1	59243	9.60%	76349	12.40%	216213	35.00%	147501	23.90%	56051	9.10%	62237	10.10%
5	Allston/Bri	74997	26.8	3431	4.60%	6552	8.70%	44344	59.10%	10697	14.30%	3975	5.30%	5998	8.00%
6	Back Bay	27476	31.7	1296	4.70%	1597	5.80%	13091	47.60%	5817	21.20%	2722	9.90%	2953	10.70%
7	Central	30901	32.9	1255	4.10%	2835	9.20%	12712	41.10%	6811	22.00%	3032	9.80%	4256	13.80%
8	Charlestov	16439	34.6	1806	11.00%	1108	6.70%	5494	33.40%	4668	28.40%	1679	10.20%	1684	10.20%
9	East Bosto	40508	32.3	5554	13.70%	4508	11.10%	12476	30.80%	11478	28.30%	2960	7.30%	3532	8.70%
10	Fenway/Ke	40898	22.6	507	1.20%	11151	27.30%	23495	57.40%	2840	6.90%	1102	2.70%	1803	4.40%
11	Harbor Isla	535	45.3	4	0.70%	10	1.90%	100	18.70%	324	60.60%	82	15.30%	15	2.80%
12	Hyde Park	31813	38.3	3828	12.00%	4617	14.50%	6115	19.20%	9349	29.40%	3834	12.10%	4070	12.80%
13	Jamaica Pl	39897	31.9	3608	9.00%	3597	9.00%	15096	37.80%	9441	23.70%	3853	9.70%	4302	10.80%
14	Mattapan	34616	34	4871	14.10%	5582	16.10%	7277	21.00%	9257	26.70%	3889	11.20%	3740	10.80%
15	North Dor	28384	31.7	2979	10.50%	3608	12.70%	9312	32.80%	7534	26.50%	2615	9.20%	2336	8.20%
16	Roslindale	32589	37.5	3845	11.80%	3778	11.60%	7435	22.80%	9944	30.50%	3755	11.50%	3832	11.80%
17	Roxbury	59790	31.7	8801	14.70%	9820	16.40%	13983	23.40%	15509	25.90%	6062	10.10%	5615	9.40%
18	South Bos	33688	32.4	2642	7.80%	2438	7.20%	13972	41.50%	8278	24.60%	3072	9.10%	3286	9.80%
19	South Dor	59949	34.5	7963	13.30%	8646	14.40%	13819	23.10%	17134	28.60%	6248	10.40%	6139	10.20%
20	South End	34669	33.1	3008	8.70%	3558	10.30%	12146	35.00%	9521	27.50%	3236	9.30%	3200	9.20%
21	West Roxb	30445	42.2	3845	12.60%	2944	9.70%	5346	17.60%	8899	29.20%	3935	12.90%	5476	18.00%
22															



```
[{
  "United States": {
    "Population":308745538,
    "Median Age":37.1,
    "Under 9":40550019,
    "%":"13.1%",
    "10-19":42717537,
    "%":"13.8%",
    "20-34":62649947,
    "%":"20.3%",
    "35-54":86077322,
    "%":"27.9%",
    "55-64":36482729,
    "%":"11.8%",
    "65+":40267984,
    "%":"13.0%"
  },
  "Massachusetts": {
    "Population":6547629,
    "Median Age":39,
    "Under 9":752774,
    "%":"11.5%",
    "10-19":868369,
    "%":"13.3%",
    "20-34":1320809,
    "%":"20.2%",
    "35-54":1899584,
    "%":"29.0%",
    "55-64":803369,
    "%":"12.3%",
    "65+":902724,
    "%":"13.8%"
  },
```

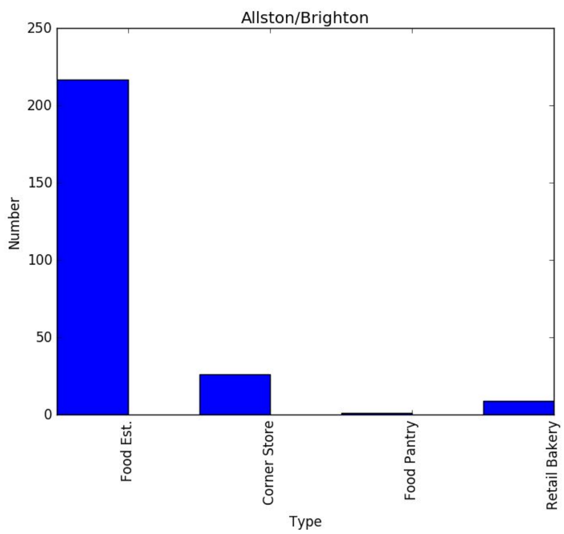

Algorithms/Techniques/Tools used

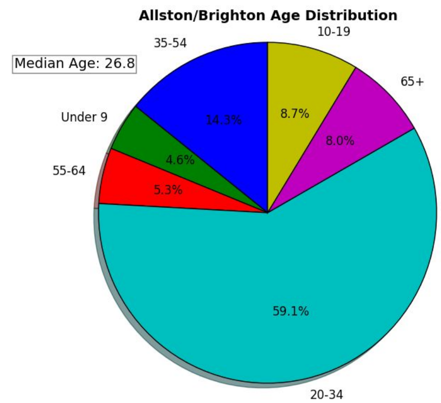
For this project, we used the relational data paradigm to project the multiple datasets, each describing a demographic, into one large dataset. In the resulting dataset, the keys of the dictionary are the neighborhoods and further broken down, each neighborhood consists of another dictionary that is keyed by the demographic.

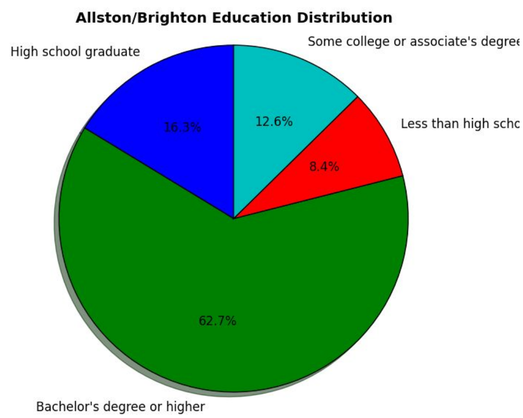
Issues we ran into trying to correlate the two datasets was a lack of data for food resources and a lack of detailed data for the demographics. We could easily find the mean for demographics, eg. average education of Dorchester or average income in Back Bay, however we weren't able to calculate the standard deviation because the data consisted of only categories. For example, there were only six categories for age and therefore only six data points. The central limit theorem states that we need at least a sample of 30 data points to even consider an underlying normal distribution. With only six, we can define a mean and standard deviation, however it would not be a well defined with the lack of data. And so, we decided to not calculate the standard deviation as it would not be a well defined variance among the data. Data collected on food resources were also limited. We only had data on restaurants, convenience stores, food pantries, and bakeries. We could have obtained more data on grocery stores, which are a much bigger indicator of food resources. With such limited data, with the main contributor being food establishments, it was difficult to determine whether such a neighborhood lacked food resources. Due to the limitation of the data, we were not able to calculate a correlation among the two datasets. For neighborhoods, we were only able to generate the averages among the neighborhoods and visualize their "distribution" with pie charts since a normal distribution would not have been appropriate.

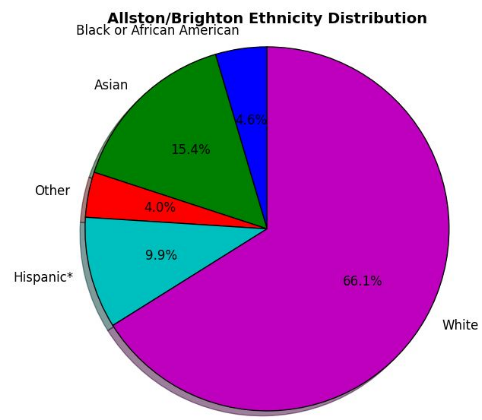
Summary of results and conclusion

Our results were therefore inconclusive as we not perform actual statistical analysis on the data. The best we could do was visualize the food resources alongside the "distribution" of a few demographics. The following is an example for the Allston/Brighton neighborhood.

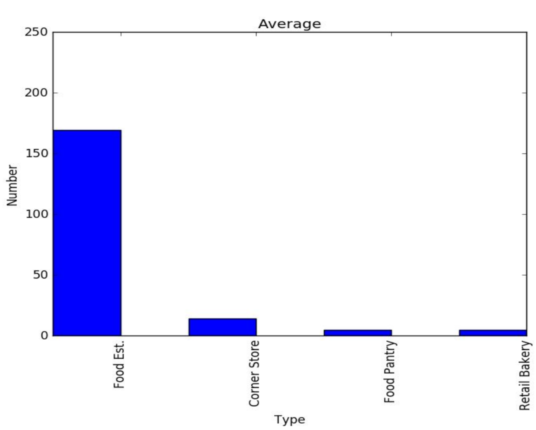


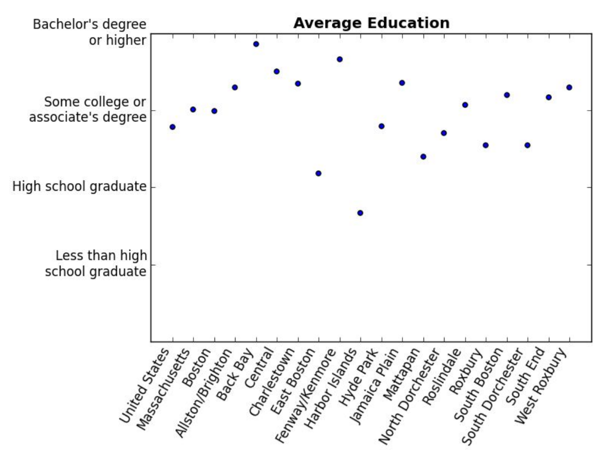


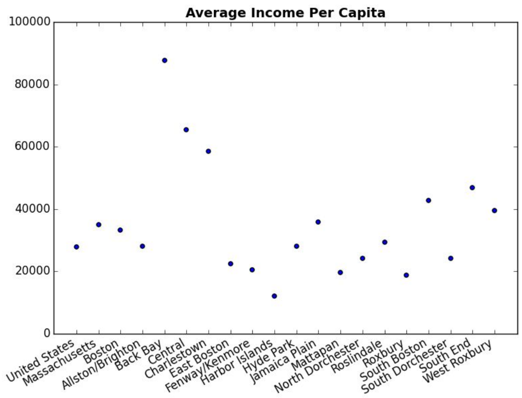


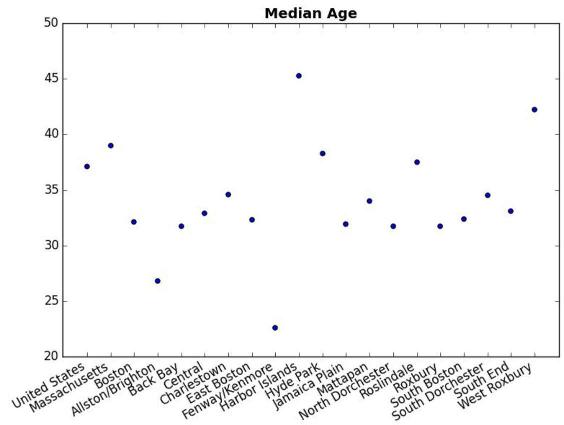


The following depicts the averages among all the neighborhoods in relation to Boston and the United States.









Although we see differences in food availability, we were not able to conclude whether food deserts were present due to the lack of data present from the Boston Data Portal. Open problems still remain with this topic as we were not able to identify predictors of food desserts.

Future Work

For the future, we definitely would need more datasets to better understand whether food desserts exist and more detailed datasets to run better statistical analysis. Future ideas that could be implemented would be to find other resources that could be lacking and trying to find predictors of that. This could be used on current data to determine whether these are current issues. However, it could also be used on previous historical data and compared to the present to predict whether such an issue will occur in the near future. I think this latter idea would be the bigger motivation for future projects. Cities could predict what future issues may lay ahead based on current situations and address them sooner rather than later.