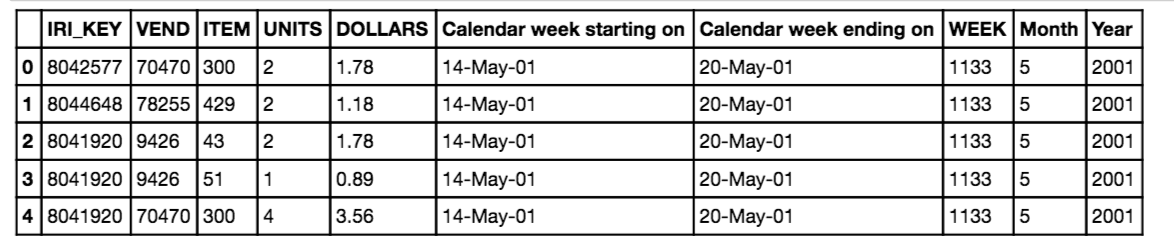
IRI Group Project

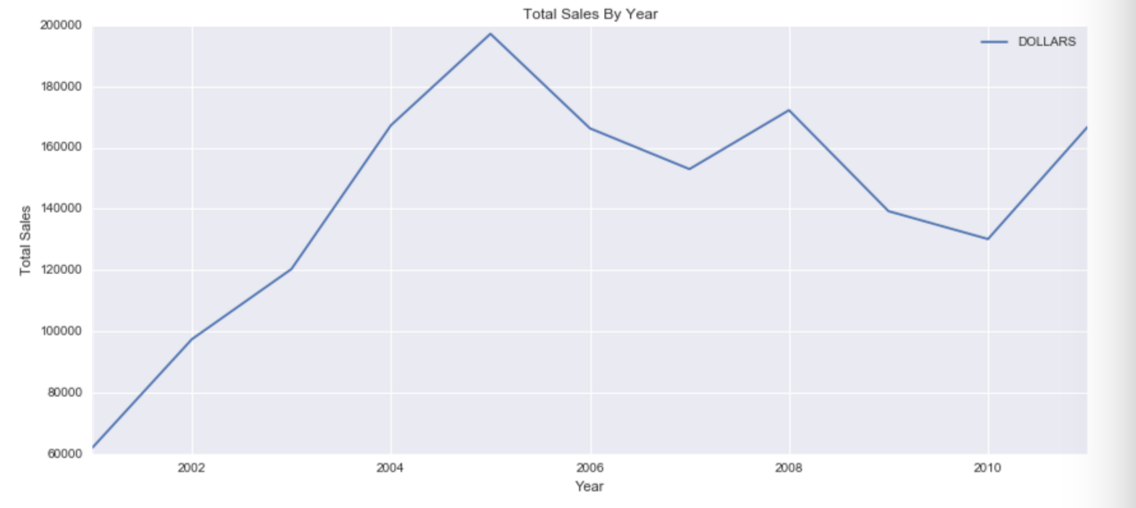
The entire IRI Dataset consists of multiple tables which may play a role in consumer trends. These tables include delivery (delivery\_stores), demographics (demos), drug store (drug), grocery (groc), and panel data. Each of these tables contains important information with data specific to its table name.

Drug Store Information

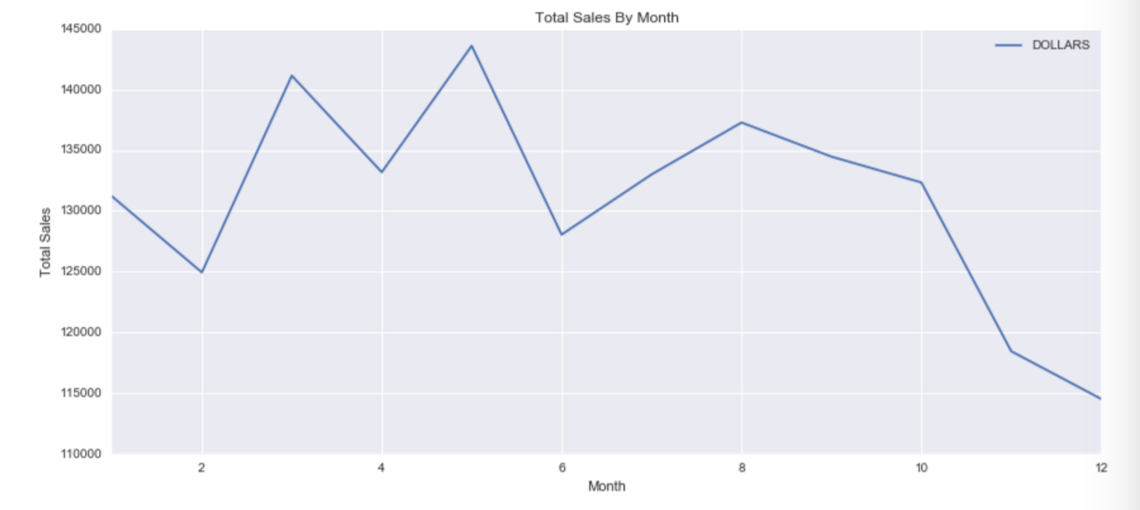
In order to obtain geographic data from the HDF file, we linked the delivery store table to the drug store table using the common ‘IRI\_KEY’, which describes the specific ID for each store in the table.



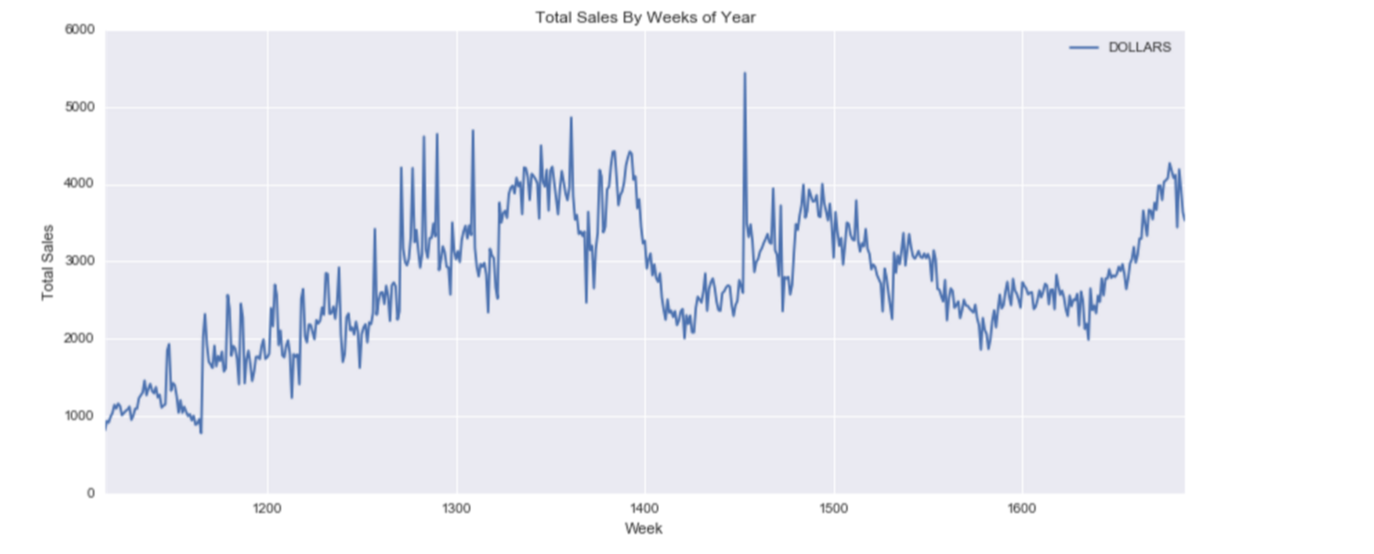
Each IRI\_KEY is listed on multiple rows with specific data pertaining to the units sold, dollars sold, week, month, and year. In order to get summary statistics for the geographic data on drug stores, we summarized total sales by year. This is shown below.



We see that the total peak point for sales was reached in 2005 for the yogurt. Total sales per month was also investigated, this table is provided below.

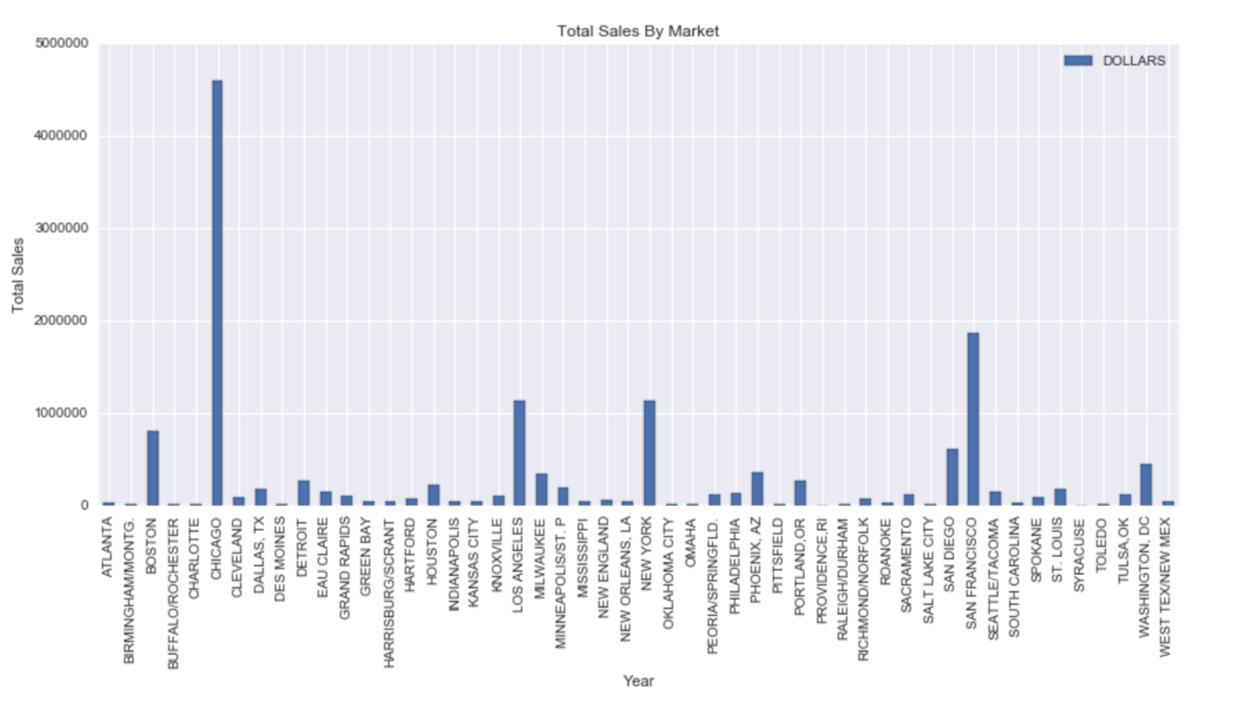


The data looks as though it is pretty consistent across all of the months with sales experiencing a slight low during months 10-12. Total sales by week were also investigated for weekly trends. This graph has a great deal of noise due to the number of samples across a year. Later we will investigate how PCA can be utilized to reduce noise in the data.

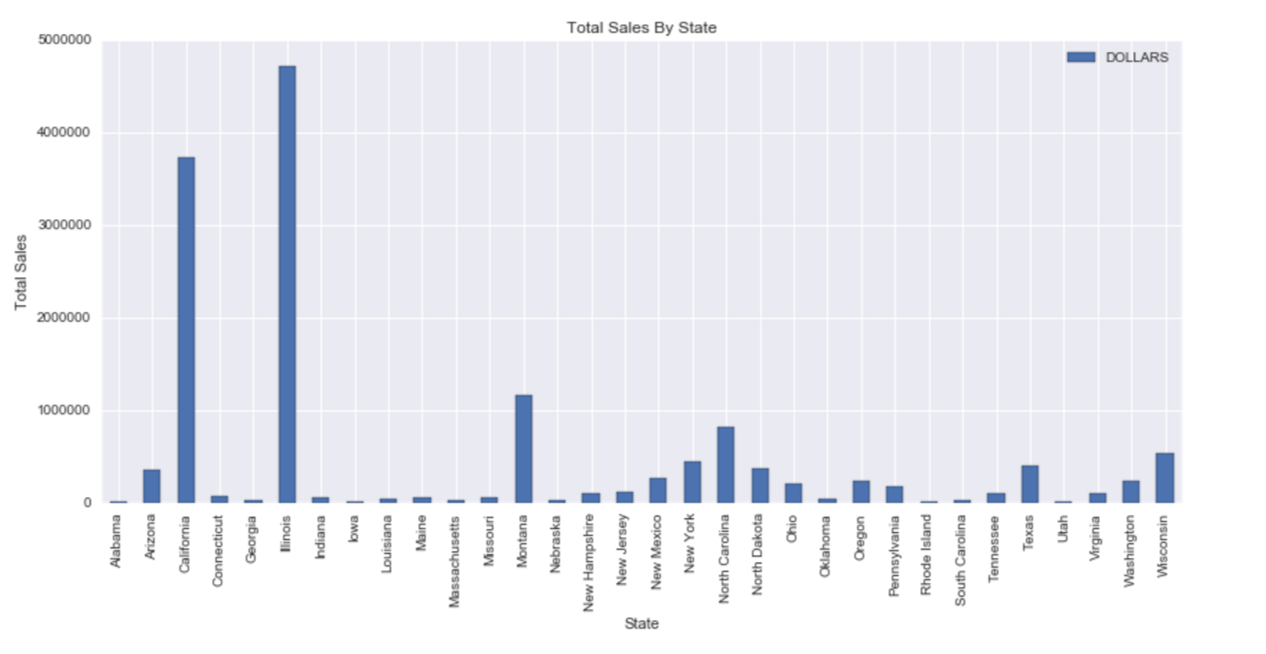


After investigating the trends in the drug store data across time, we joined the drug store sales with the store information table. The resulting graph and columns are shown below. The specific columns that were kept are listed here: ('IRI\_KEY', 'VEND', 'ITEM', 'UNITS', 'DOLLARS', u'EST\_ACV', u'Market\_Name',u'Month', u'Year','WEEK', u'name','Calendar week starting on', 'Calendar week ending on').

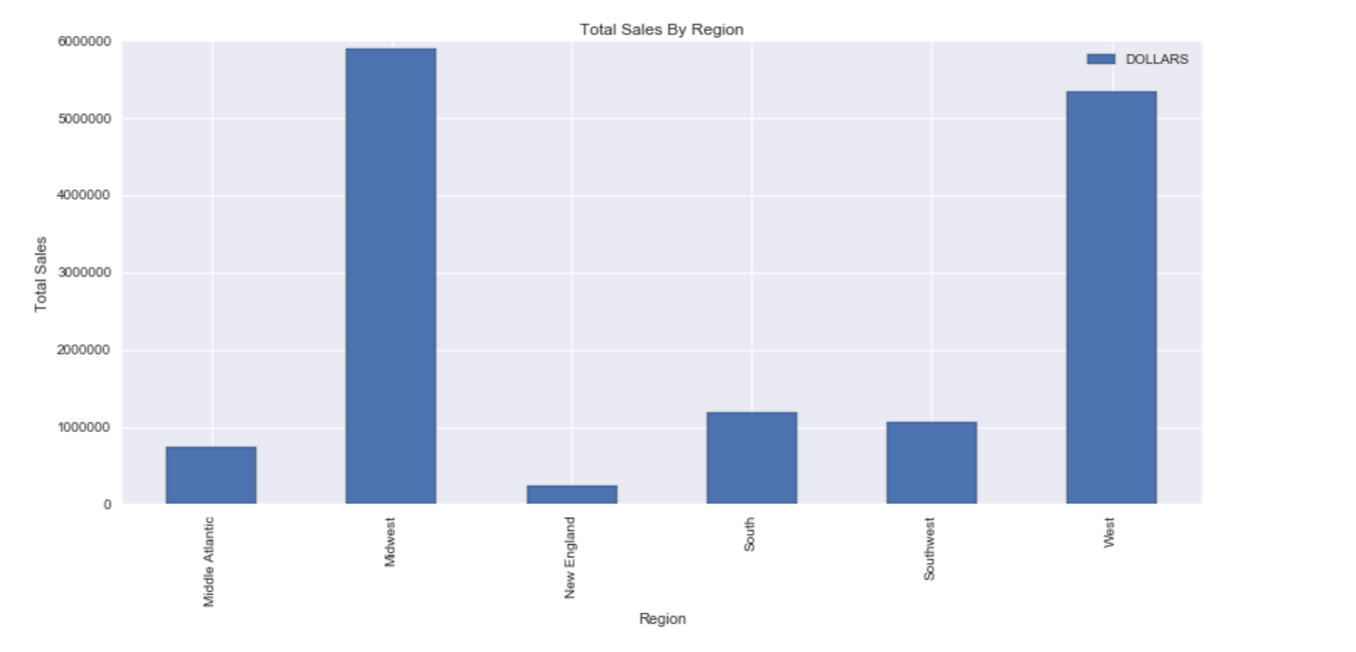
Sales by market graph shows the total sales per each market area aggregated using a group by. This bar graph helps give a general sense for where most of the sales are being made for yogurt across drug stores. We can see that Chicago has the highest sales for yogurt by a large margin compared to all other areas.



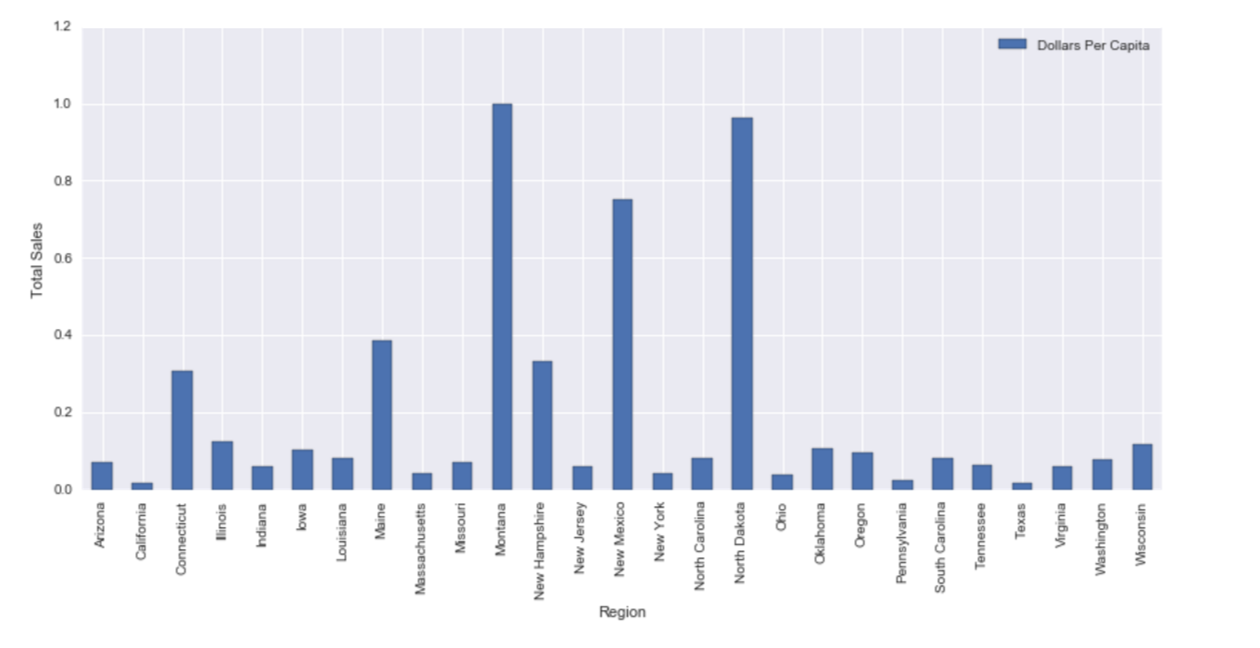
We wanted to further investigate this trend, by graphing the trends across states and regions of the country. In order to do this we mapped the geographic locations to each state and grouped based on states. The following graph shows the sales by state. We can see that this graph is almost identical to the graph above, as the data tended to be distributed equally across all states, i.e. each state has only one or two cities that were listed within in it.



Form here we chose to group the sales trends by region of the country. In order to do this we looked the US census data 2010 for information on how to group the states by region.



In the table for sales by region we can see that the top two spenders in yogurt are located within the west and Midwest. This is a total amount and is not based on population for a per capita score. After noticing the large spikes in large cities we noticed that sales data alone does not give a good representation of the data, because the data may be biased due to the large cities. If we want to know the per capita ratio of sales per person in each state we can divide by the state population. In the following table we graph the normalized values of sales per population (population provided was from US Census 2010).



**Errors**

Multiple assumptions we made which resulted in the regional per capita sales listed above. Many of these assumptions may not accurately portray the data. The first is the population disparity between the actual census population and the specific cities that have data available. This could greatly throw of the data. Another factor is the census data from 2010, this time frame does not match the IRI data timeframe which spans 2001 to 2011. This makes dividing by the total population extremely inaccurate, so the per capita table listed above should most likely be disregarded accept for general inquiries about the data.

**Principal Component Analysis (PCA)**

Next we wanted to