Brown University

Datathon

Team Member:

Samuel Chan, Yiquan Xu, Zhicong Chu, Yimo Zhang

Content

Introduction

The dataset our group choose is Upserve. We first converted the date to season and weekday and did data analysis to get the basic ideas of the data, like the summary statistics. Then we did regression analysis and tried to find a fitted model for the dataset by using machine learning techniques and linear regression. After that we found the popular items based on different region and season. Finally, we did data visualization to get a better understanding of data.

Data Analysis

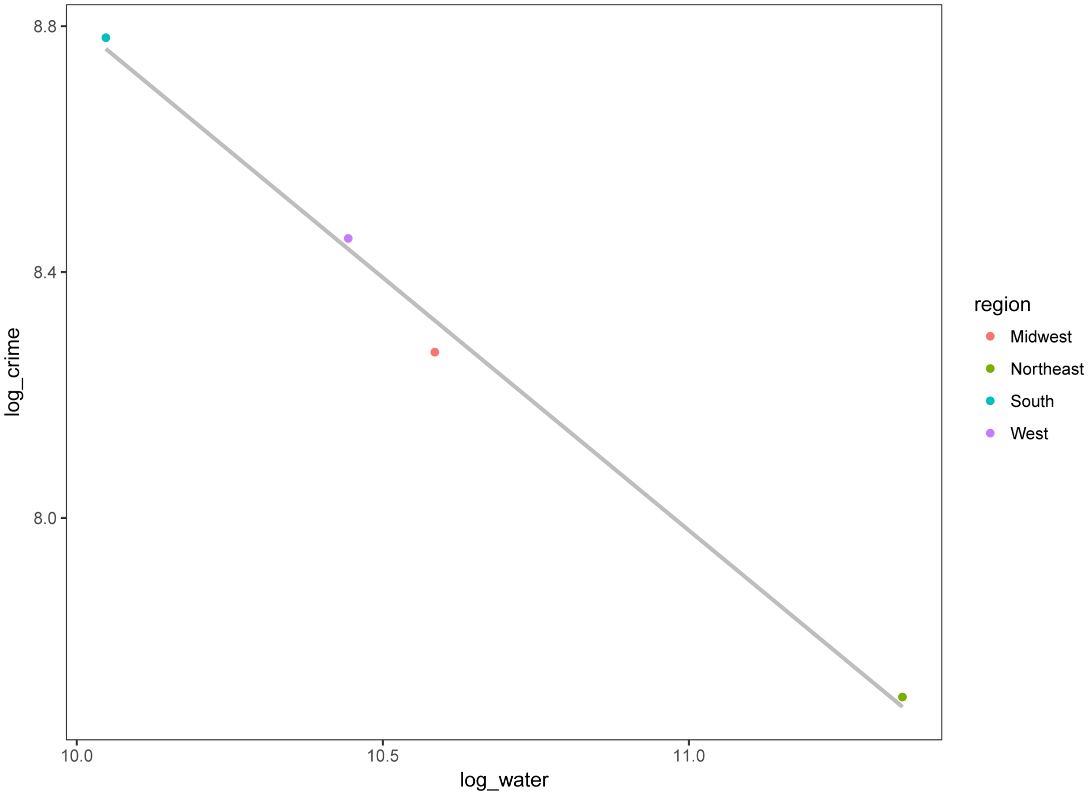
0.1 Summary Statistics

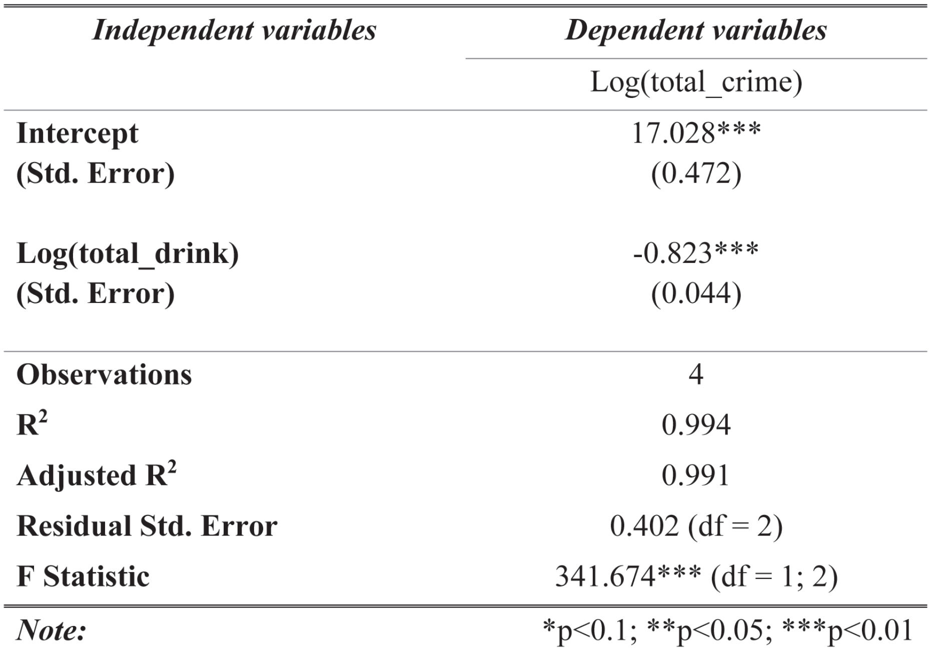
The summary statistics are the table above. From the table we could find the minimum, mean and maximum of different variables. The total sales have negative numbers, which represent compliments and discounts. There are no NAs or Inf among all the variables, so we don't need to deal with the missing value. The maximum number of total_sales are 223662, which far exceeds the 1.5 IQR, so we consider it outliers and delete it here.

0.2 Linear Regression

We fit two linear models to this dataset. Firstly, we use total_sales as response and the baseline information, including region, shift_bin and item_category as predictors. However, the result was not desirable, for not significance was reported and the model exhibits very small R squared value.

Secondly, we use crime rate dataset in U.S as complementary and fit a linear model with sales per person as response and crime rate as predictor. The result shows strong linear relationship between these two variables. However, due to the small sample size, we cannot make valid conclusion based on that.

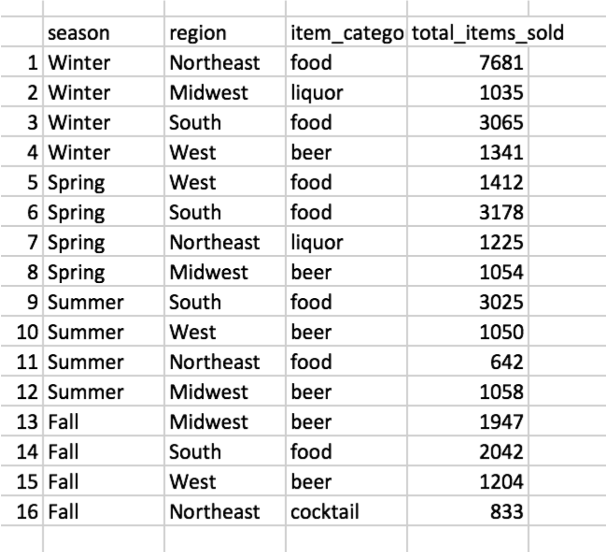




Independent variables	***Dependent variables***
	Log(total_crime)
Intercept	17.028***
(Std. Error)	(0.472)
Log(total_drink)	-0.823***
(Std. Error)	(0.044)
Observations	4
R^2	0.994
Adjusted R^2	0.991
Residual Std. Error	0.402 (df = 2)
F Statistic	341.674*** (df = 1; 2)
Note:	*p<0.1; **p<0.05; ***p<0.01

0.3 Data Manipulation

As the sales were measured based on region and time, we were interested in the most popular item grouped by region and time. We got the following table: We could find the consumption of the categories are highest in winter. The overall consumption of the categories are lowest in summer but the South region has high consumption on food.



	season	region	item_catego	total_items_sold
1	Winter	Northeast	food	7681
2	Winter	Midwest	liquor	1035
3	Winter	South	food	3065
4	Winter	West	beer	1341
5	Spring	West	food	1412
6	Spring	South	food	3178
7	Spring	Northeast	liquor	1225
8	Spring	Midwest	beer	1054
9	Summer	South	food	3025
10	Summer	West	beer	1050
11	Summer	Northeast	food	642
12	Summer	Midwest	beer	1058
13	Fall	Midwest	beer	1947
14	Fall	South	food	2042
15	Fall	West	beer	1204
16	Fall	Northeast	cocktail	833

0.4 Model Building

We tried to classify on the categorical variable shift bin from the region of the store, category of the sold items and the total sales of each observation. Different statistical learning techniques were used to build the classifier including logistical regression, K Nearest Neighbors and Artificial Neural Network. The logistical regression turned out to perform the best on prediction. Accuracy was above 80%. The performance was not very satisfying and we may run into over-fitting without cross validation.

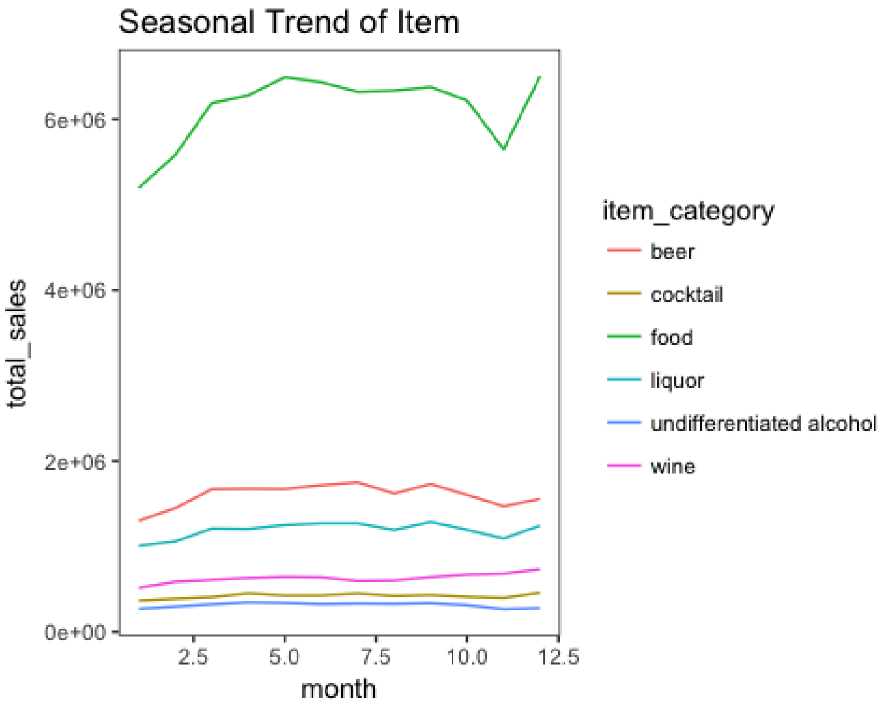
For each store, we also created secondary variables by computing the mean and standard deviation at every specific shift_bin. Those variables were used for clustering. However, the model did not perform well with the vast majority of the stores fell into one big cluster, indicating the stores did not have distinguishable patterns on those secondary features. We also tried use those secondary features to classify the geographic area of the stores but the performance was not good.

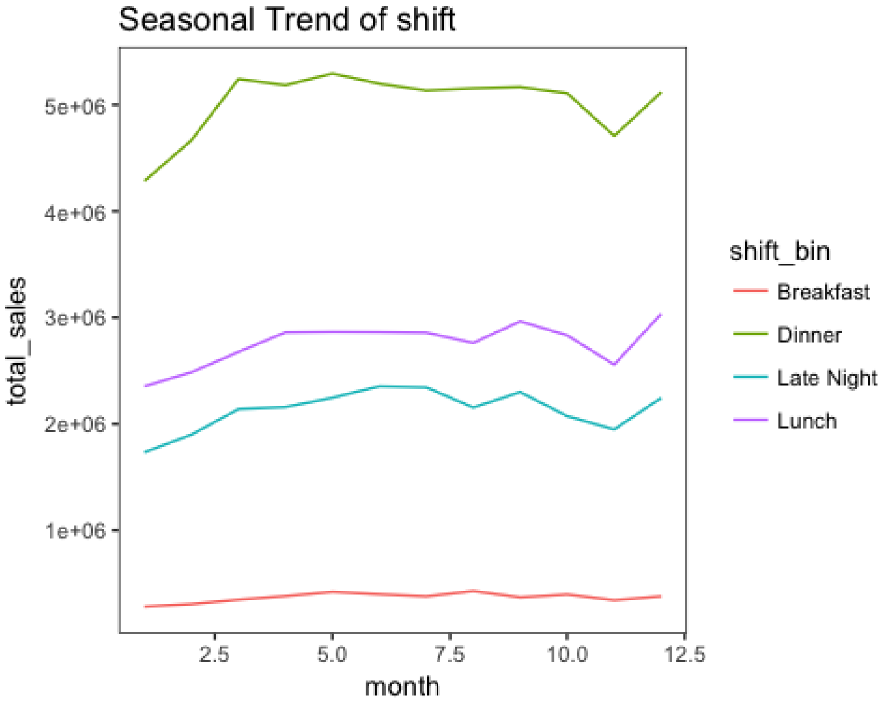
Data Visualization

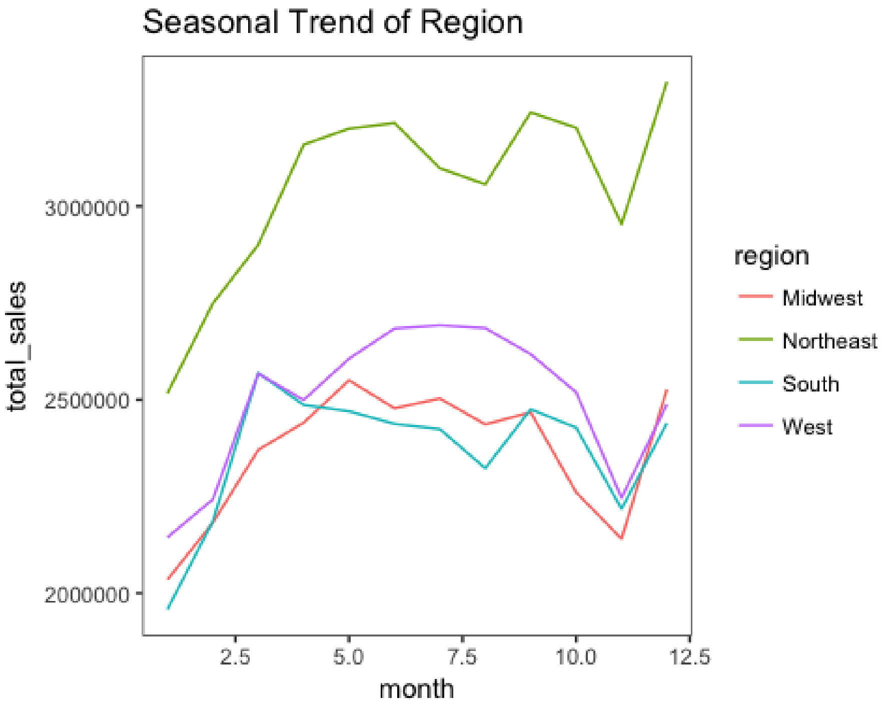
We draw plots about the seasonal trend of region, category and shift of total sales and compare it with each other. We could find the line of the trend of shift was clearly separated. For the item category we could find the consumption on food was a lot greater than consumption on other categories, which makes sense because humans can not live without food and not all people drink alcohol. The seasonal trend of region have intercept on Midwest, West and South.

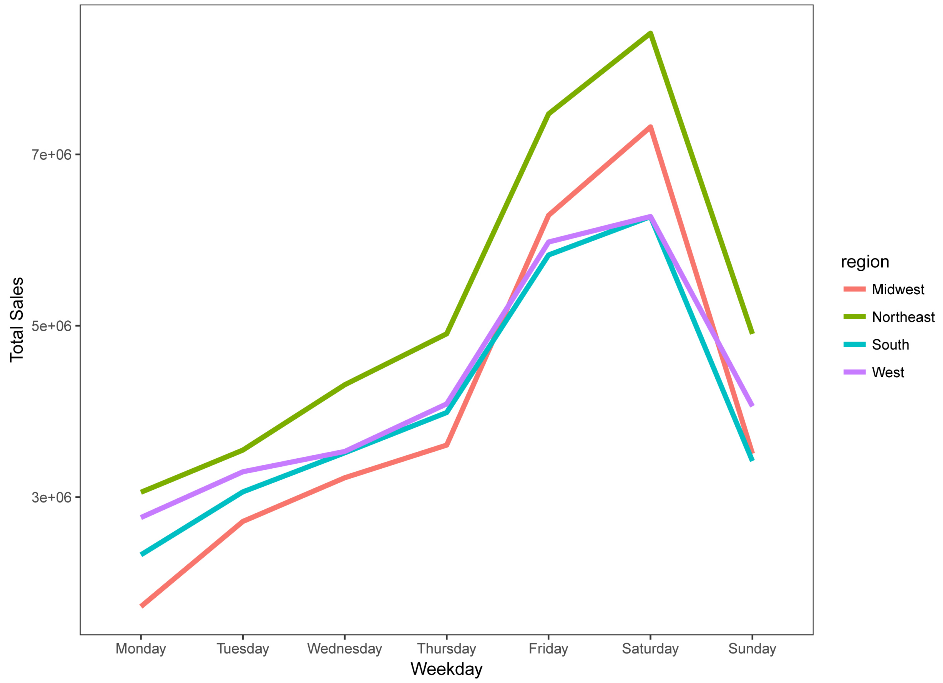
We also did visualization about average beer and cock price versus different regions. In these plot we could find the Northeast region has the highest average beer and cocktail price, while the price of the South region is quite spread out.

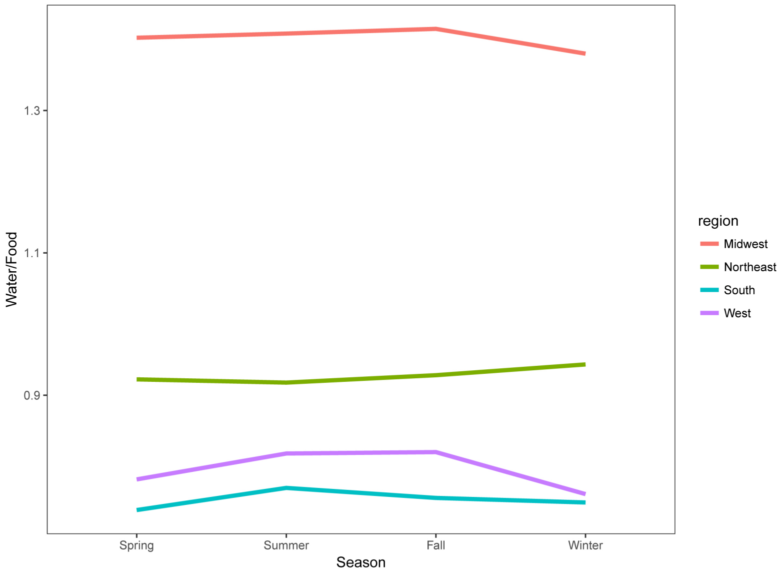
Since we found there were large difference on food consumption and other consumption, we drew a visualization on the (Other/Food)'s trend in different season. We could see in the west and south part the other consumptions were lower than the food consumption.

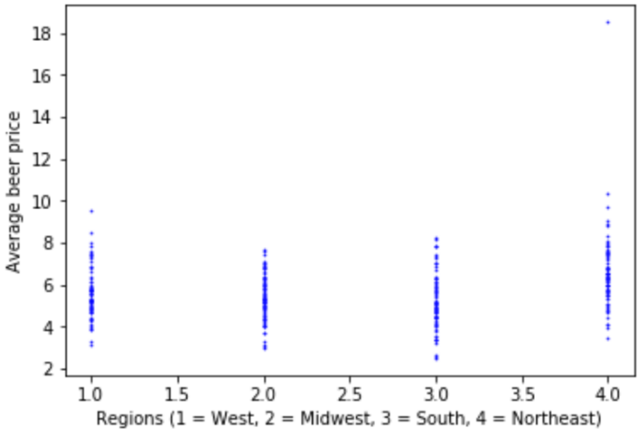

Seasonal Trend of Item
total_sales
6e+06
4e+06
2e+06
0e+00
2.5
5.0
7.5
10.0
12.5
month
item_category
beer
cocktail
food
liquor
undifferentiated alcohol
wine

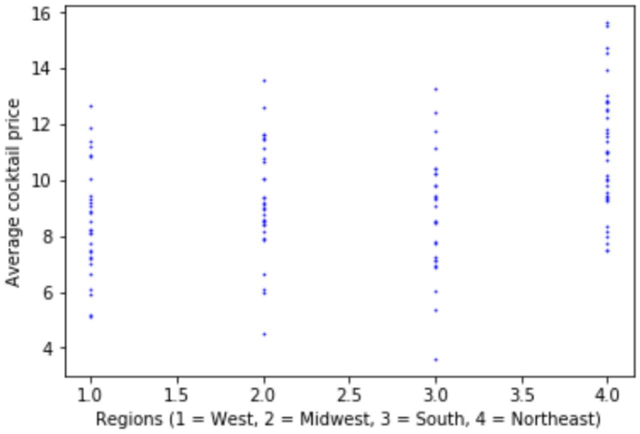

Seasonal Trend of shift
total_sales
5e+06
4e+06
3e+06
2e+06
1e+06
2.5
5.0
7.5
10.0
12.5
month
shift_bin
Breakfast
Dinner
Late Night
Lunch


Seasonal Trend of Region
total_sales
3000000
2500000
2000000
2.5
5.0
7.5
10.0
12.5
month
region
Midwest
Northeast
South
West


Total Sales
7e+06
5e+06
3e+06
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Weekday
region
Midwest
Northeast
South
West







Reference

https://www.ucrdatatool.gov/Search/Crime/State/RunCrimeStatebyState.cfm

–Crime Rate Dataset