|  |
| --- |
| walmart sales forecast project |
| Capstone project 2 |
|  |
| Yoheita Yoshimura | capstone project2 | March 9, 2024 |

# Introduction

Walmart, the largest company by revenue in the world, operates more than 10,500 stores in 19 countries selling more than 75 million products. By organizing enormous amounts of many items and stores, prediction of the sales can be one of the important factors managing the direction of its business. It enables the company to plan or make informed decisions on the financing, business plans, inventory organization, human resources management and many more. The higher the accuracy of the model the higher the contribution to the company.

In this project, we will be using the data provided by The Makridakis Open Forecasting Center (MOFC) at the Universtiy of Niscosia. The data contains 3 tables. First data contains number of items sold for each 3000 items in each 10 stores located in TX, WI and CA. 3 stores in TX, 3 stores in WI, and 4 stores in CA. Second data contains the price of the each items in the each stores, and last data contains the calendar data providing information such as major events occurred in USA.

Given the data with time series data, we will be using time series forecasting algorithm to forecast the data.

# Approach

Data Acquisition and Wrangling

Evaluate to understand the provided data is crucial fundamental steps before feeding data to train the algorithm. The initial approach will be to understand the structure of the data. The structure will allow us to decide the modification to the data and to delete or modify the data for any missing data, such as time skip or null data.

Sales\_train\_evaluation.csv

A screenshot of a computer

Description automatically generated

The data consists of product detail such as, id, item id, store id, state\_id and rest of date data has format of d\_X where X is the number of date. The date continues from d\_1 to d\_1941 with the number of quantities sold for each product.

A screen shot of a computer code

Description automatically generated

The id includes multiple information including: category ID, Product number, Store ID and State ID.

Refer to below format:

State ID

HOBBIES\_1\_001\_CA\_1\_evaluation

Category ID

Store ID

Product Number

The id column data are in object format and d\_X data are in int64.

The table consists of 30490 rows and 1947 columns, and no missing data.

Fig.1

A screenshot of a black screen

Description automatically generated

Fig.2

A graph of a number of items

Description automatically generated with medium confidence

There are 4 stores in CA, 3 stores in TX and WI, and 3049 items are sold in each store (Fig.1). Fig.2 displays the distribution of item categories are equal through all the stores.

30490 products

7 Department: HOBBIES\_1 … 2, HOUSEHOLD\_1 … 2, FOOD\_1 … 3

3 Category: HOBBIES, HOUSEHOLD, FOOD

10 Stores: CA\_1… 4, WI\_1 … 3, TX\_1 … 3

3 States: CA, TX, WI

The target data will be non-categorical since we are forecasting the numerical value of sales.

Calender.csv evaluation

A screenshot of a computer

Description automatically generated

This table represents date related information and event and SNAP on the related states.

What is SNA?

The Supplemental Nutrition Assistance Program (SNAP) is the largest federal nutrition assistance program. SNAP provides benefits to eligible low-income individuals and families via an Electronic Benefits Transfer card. This card can be used like a debit card to purchase eligible food in authorized retail food stores.[1]

1: <https://www.benefits.gov/benefit/361>

The table columns have the types below. Table has 13 columns with 1969 rows of data. By comparing the number of rows and number of non-null values, we can see that event related columns have many null values. The null values are created for the date when there is no event by comparing the number of data and null values in below.

A screen shot of a computer

Description automatically generated A screen shot of a computer

Description automatically generated

The event data has been replaced to “no\_event” for all the null values.

sell\_prices.csv evaluation

The table contains data of the price of each product in each store by a weekly bases.

A screenshot of a computer

Description automatically generated

Here is the data type for the each column data

A screen shot of a computer

Description automatically generated

There is no missing values in this table.

A screen shot of a computer

Description automatically generated

Exploratory Data Analysis (EDA)

From observing the table, the null values are representing the days without an event and “date” column has a format of “Year-Month-Day” . Therefore, the null values in event columns have been replaced with “No\_Event” and data type of “date” column has been modified to datetime64.

Below table is to represent the number date when SNAP is available in each state to observe any differences in the values. It seems the number of SNAP is almost equal between 3 states.

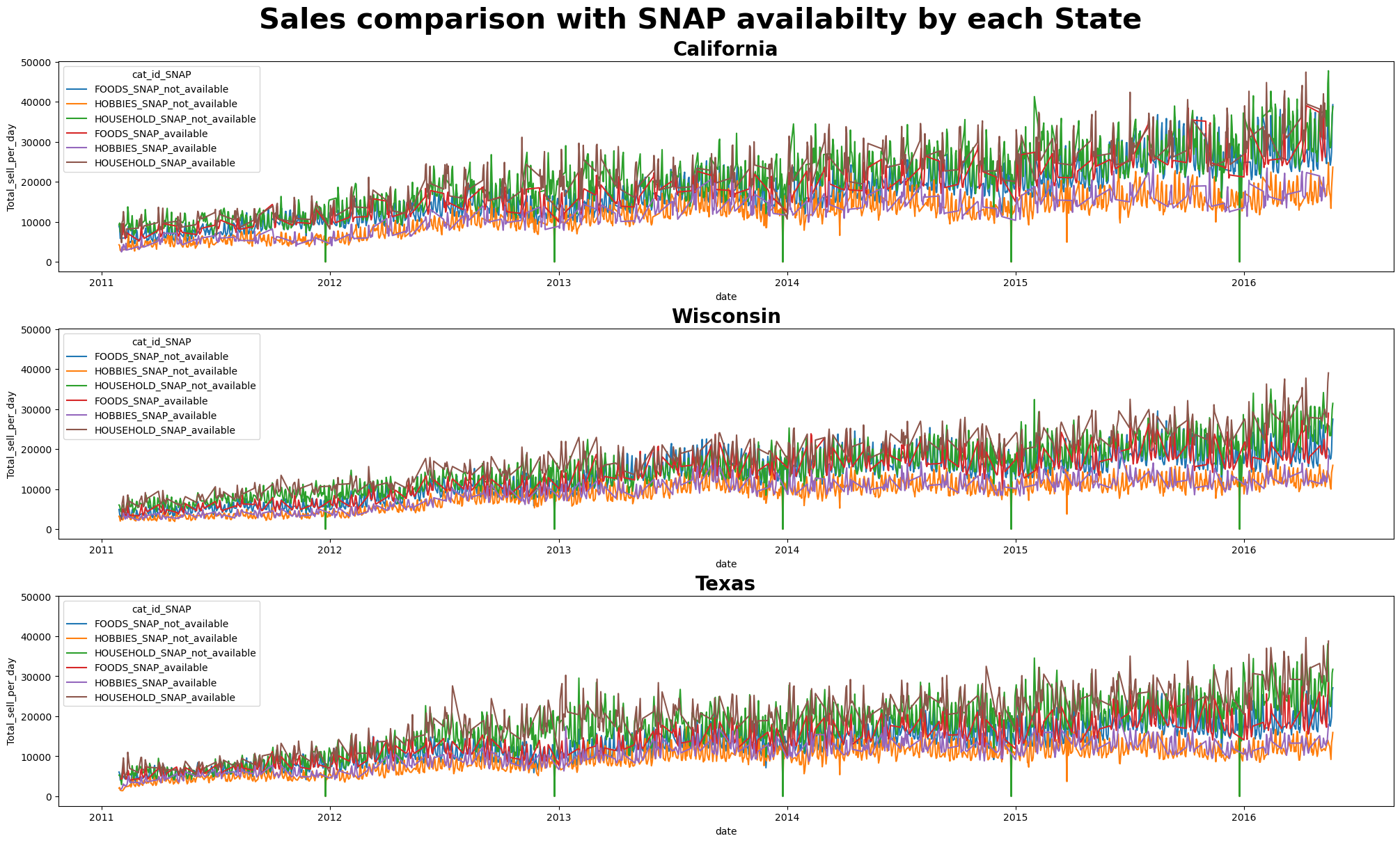
A graph of different colored lines

Description automatically generated

A graph of blue and orange lines

Description automatically generated

SNAP does make sales difference by availability dates. The date with SNAP has higher sales than the day without SNAP.



A graph of different colored lines

Description automatically generated

The SNAP availability does make a difference in the sales as presented in the table. The high difference in FOOD category explains SNAP is indeed only apply for grocery items. It also visualizes the wealthiness of each state. Wisconsin having the most difference in the SNAP availability explaining Wisconsin has lowest economy within the 3 states. The sales difference with SNAP availability explains the data is Exogenous Variable.

Here we have a graph of sales and sales quantity time series graph. There is a sales difference per day. Clearly Saturday and Sunday have the highest sales in either quantity or sales. The interesting point is that the sales gradually increases from 2011, but the quantity of sales have not changed as much since 2011. The reason may be that the average item of price has increased over this 5 years. The sales difference of the weekday can conclude weekday data is Exogenous Variable as well.

A graph of colorful dots

Description automatically generated with medium confidence

A colorful lines and dots

Description automatically generated with medium confidence

Higher sale and sales quantity are observed during weekend, and other weekday seems to have similar pattern on sales.

A graph of a sales report

Description automatically generated with medium confidence

By extracting random item from the table, the graph represents many zero values.

Baseline Modeling