Coraline Case Study

“M5 Forecasting - Accuracy”

A presentation by Thanak Rattanopastkul, 2022

Tasks

1. **Preview** the dataset
2. **Define** objectives
3. **List** the tools needed
4. **Plan** the process and underlying logic
5. **State** the results from the process using visualization and explanation
6. **Summarize** the case study
7. **Suggest** implementations from the findings
8. **Suggest** additional materials for better analytics
9. **Deliver** Analytic Report **(this presentation)**
10. **Deliver** BI Dashboard (Optional)
11. **Deliver** Additional analytic plan during COVID-19

Task# 1

Task 1: Preview the dataset

Task 1: Preview the dataset

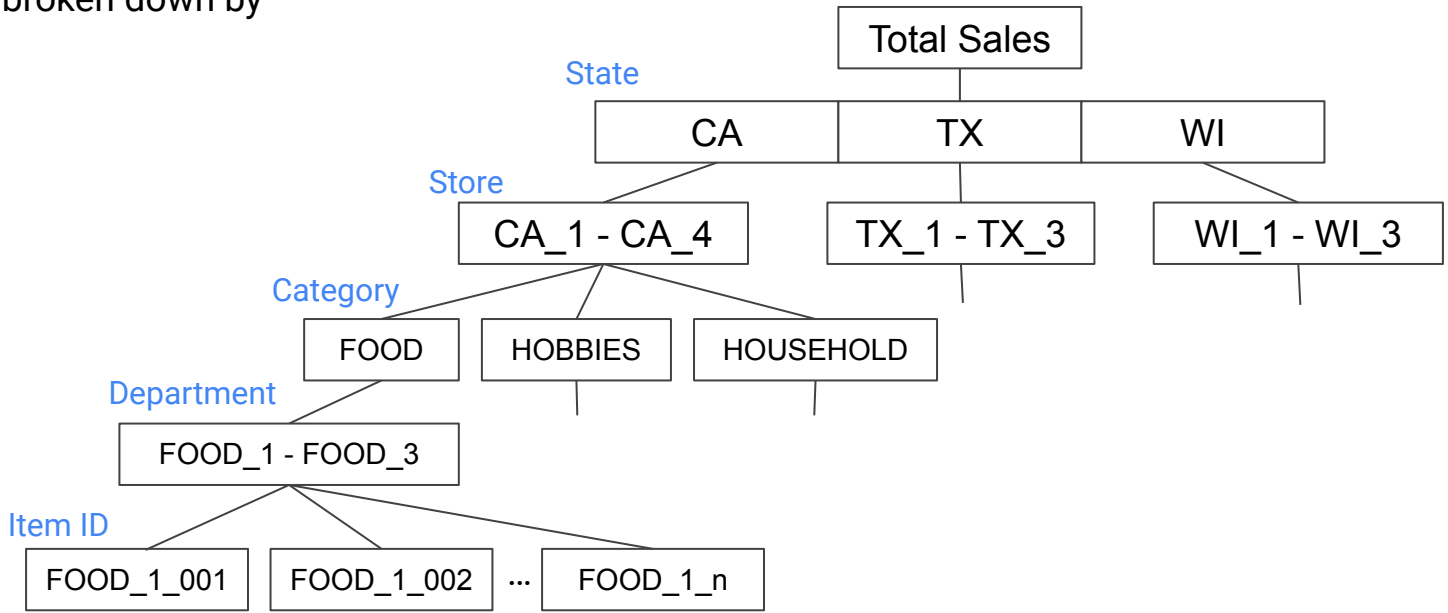
- The data is a hierarchical sales record from Walmart in 3 States
  - California (CA)
  - Texas (TX)
  - Wisconsin (WI)
- The period
  - 2011-01-29 to 2016-05-22

- There are 5 CSV files provided
  - calendar.csv (actual date,events,SNAP[1])
  - sales_price.csv (daily price for every SKU)
  - sales_train_validation.csv (1913 days sales record)
    - Training dataset for the competition
  - sales_train_evaluation.csv (1941 days sales record)
    - Evaluation dataset for Kaggle's public LB
  - sample_submission.csv
    - Example for submission

1 SNAP(Supplemental Nutrition Assistance Program) is a subsidy program to supplement the food budget of needy families. reference

Task 1: Preview the dataset

- Total sales can be broken down by
  - states_id
  - stores_id
  - cat_id
  - dept_id
  - Item_id



Task 1: Preview the dataset

- 3 product categories
  - FOODS
  - HOBBIES
  - HOUSEHOLDS
- Products are specified by product_id
- Product price could be varied from time to time and from one store to another.

- Weekend, Events, and Holiday could affect the sales.
- SNAP period should drive the food sales.
- File to use in the analytics
  - sales_train_evaluation.csv
  - sell_prices.csv
  - calendar.csv
- Not using the other files because this is not a project for ML competition

Task# 2

Define Objectives

Task 2: Define Objectives (1/2)

After a while of of skimming the dataset, theses are what I'm curious about:

1. How does the landscape of this multi State - multi Store - multi Product looks like?
    - How does the sales distributed between:
        - Time
        - States
        - Stores
        - Products
    - Were there any seasonality of the total sales?
2. How big is the effect of Weekend, Holiday, Events on the sales?
3. Are there any interesting patterns in sales?
4. What would be the story of the dataset as whole?

Task 2: Define Objectives (2/2)

What to deliver in this project

1. Deliver the market report
    - Sales distribution summary
    - Revenue summary
    - Sales seasonality summary
    - Holiday effect summary
    - etc..
2. If there are any interesting patterns within the dataset, deliver the report explaining about their characteristics
3. Deliver a BI dashboard
4. Deliver a COVID-19 pandemic special case analytic plan

Task# 3

Tools List

Task 3: Tools Required

Here is the list of tools:

1. Python and Jupyter Lab - Good for multiple medium-big size CSV analysis in a timely manner
    - pandas
    - numpy
    - matplotlib
    - etc…
2. Tableau Public - as an interactive BI dashboard
3. Git - to store the project repository and do the version control

Task# 4

Process Planning

A presentation by Thanak Rattanopastkul, 2022

Task 4: Processes and Underlying Logic (1/2)

Process:

1. Check the data integrity and fix the problems
    a. Duplicated data
    b. Missing data
    c. Dating mistakes
2. Construct a table that contains all the necessary data
    a. sales_train_evaluation sales data are arranged vertically. Need to be formatted.
3. EDA: Top down exploring the sales distribution
    a. To learn the characteristic of the dataset and, at the same time, mark any interesting patterns
        i. How many States, Stores, Category, Department, Product_id
        ii. Total Sales
        iii. Total Sales by State
        iv. Total Sales by State by Store
        v. Total Sales by State by Store by Product Department

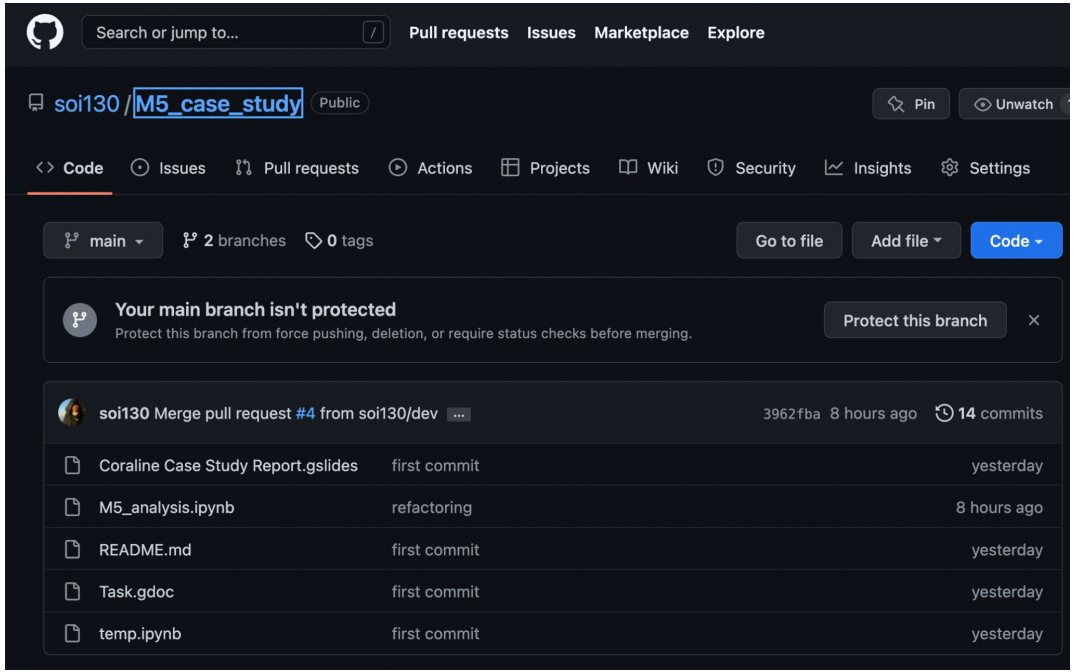
Task 4: Processes and Underlying Logic (2/2)

4. From the EDA these insight are expected:
    a. the seasonal pattern
    b. sales went up during the weekend and events (not sure what to expect during the events)
    c. product champion
    d. store champion
5. Describe and analyze the insight gained during the analysis
6. Visualize the insight using Matplotlib
7. Make some assumptions about the situation during COVID-19 pandemic
8. Derive an analytics plan to to the same analysis during COVID-19
9. Deliver all of the deliverables
    a. Analytics Report (This presentation)
    b. BI Dashboard
    c. COVID-19 plan document

Task# 5

The Analysis Result

Task 5: Results



Project Repository:

The main Jupyter Notebook and other project files used in the analysis can be found in [this GitHub repository](#).

Task 5: Results (1/17)

3 States

CA, TX, WI

10 Stores

4 in CA | 3 in TX | 3 in WI

3 Categories

FOOD| HOUSEHOLD| HOBBIES

3049

Different Products

7 Departments

FOOD 1-3 | HOUSEHOLD 1-2 | HOBBIES 1-2

Interval

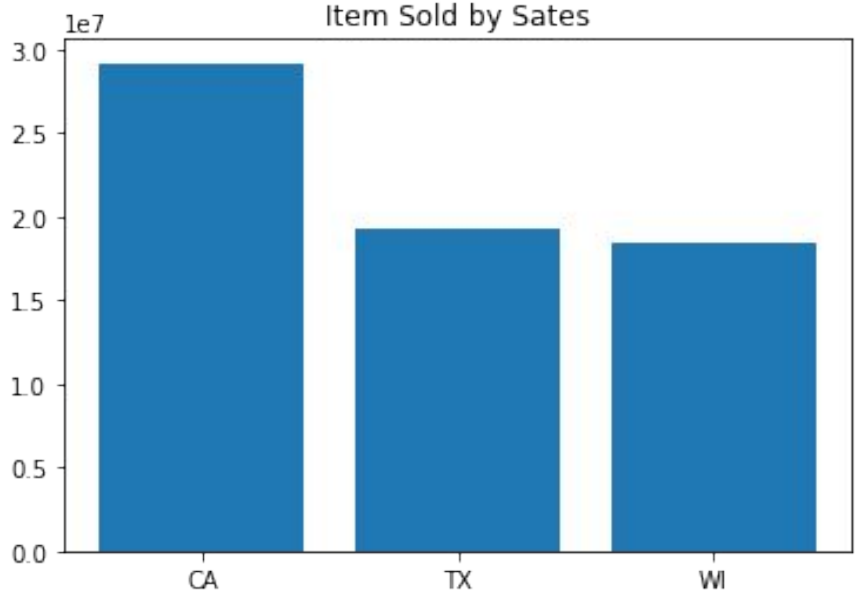
2011-01-29 to 2016-05-22

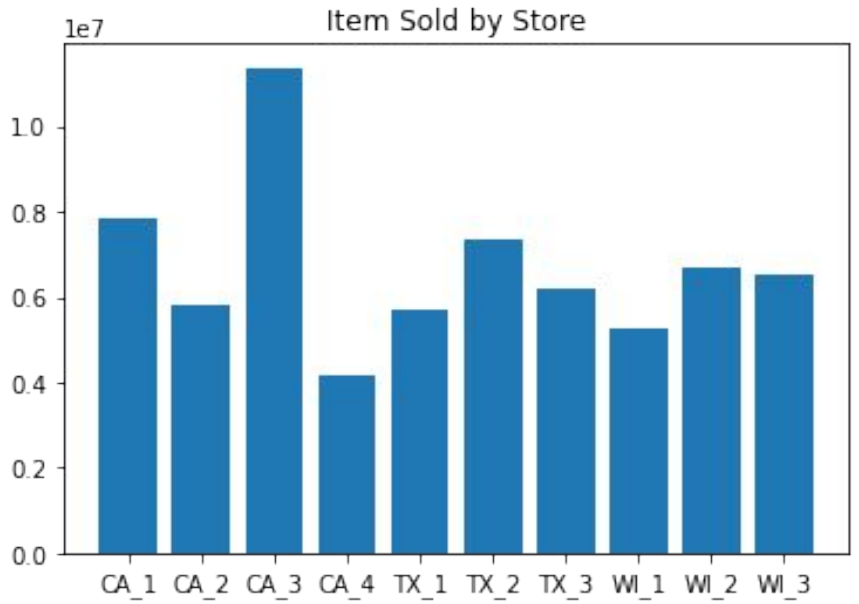
Task 5: Results (2/17)

Uptrend in business:

Considering MA30 (For smoothing the day that all the shops are closed such as New Year Day) of total product sold, average total number of item sold from every stores combined has been increasing constantly.

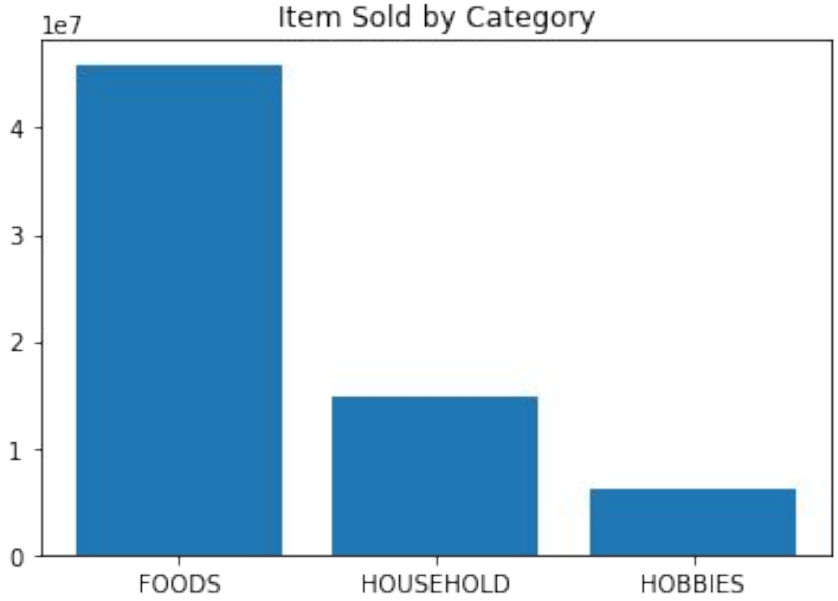
Task 5: Results (3/17)

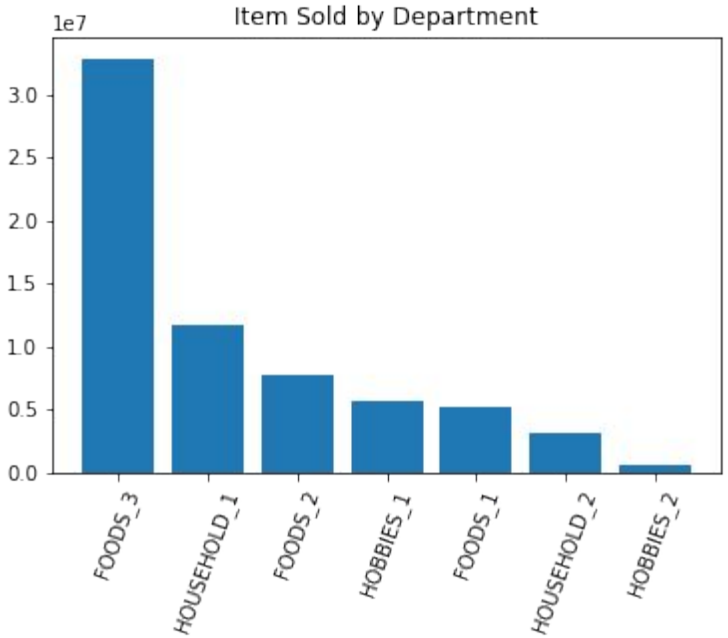




Item sold: CA is the best performer in term of item sold. To be exact, CA_3 is the best

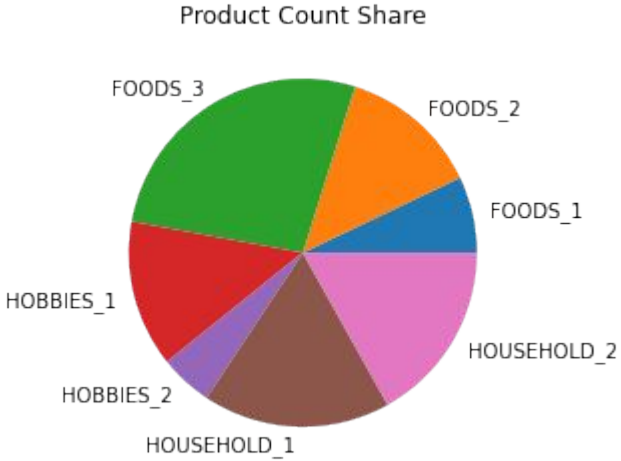
Task 5: Results (4/17)

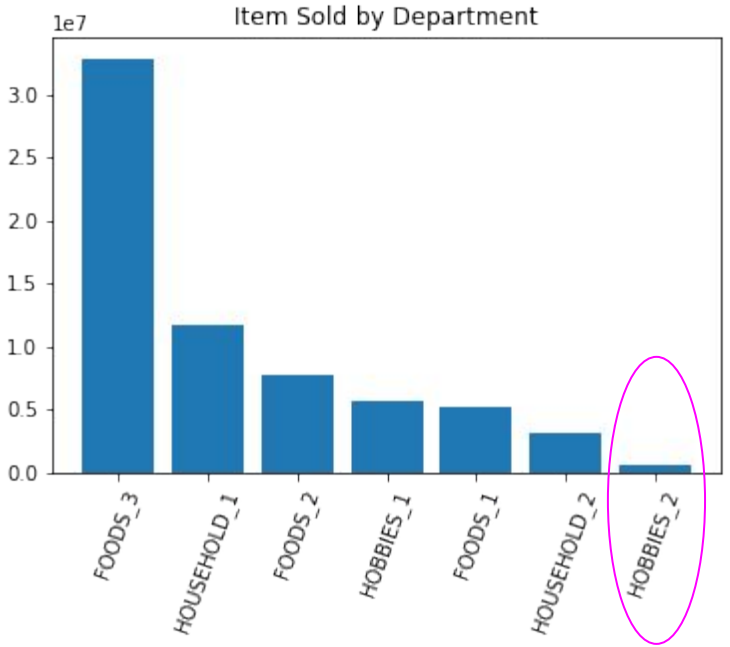




While FOOD is the most selling product category, FOOD_3 and HOUSEHOLDS_1 is the most selling department.

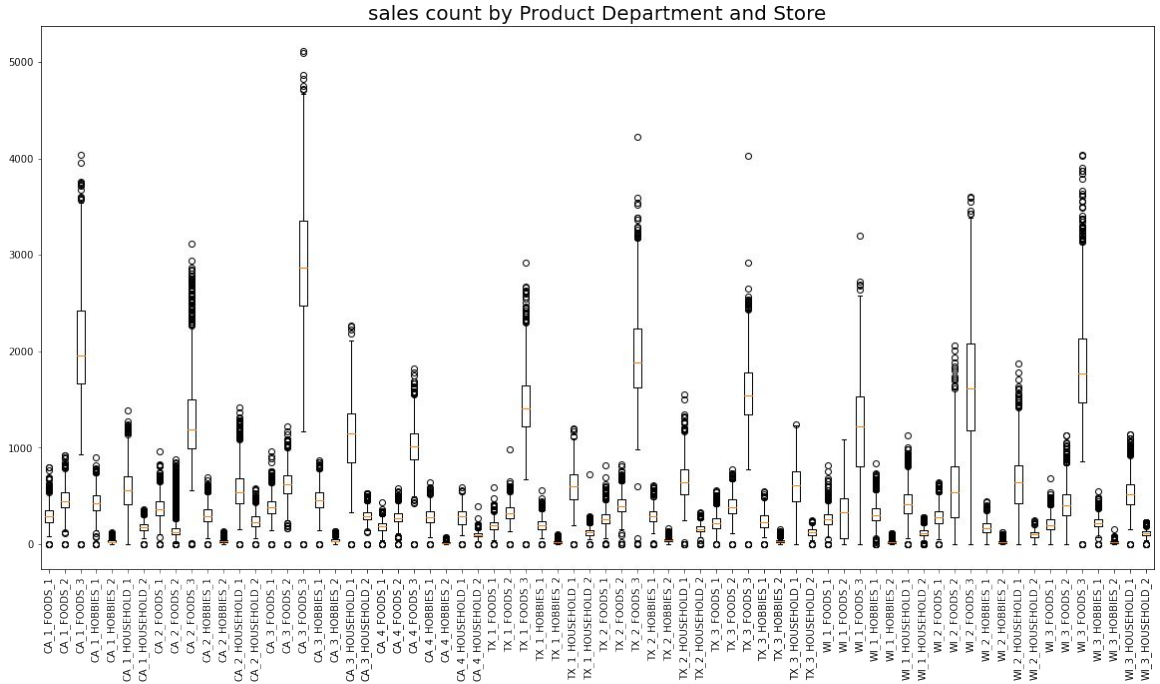
Task 5: Results (4.5/17)





HOBBIES_2 has a lowest item_id count (149). This might be the reason why it didn't generate much sales.

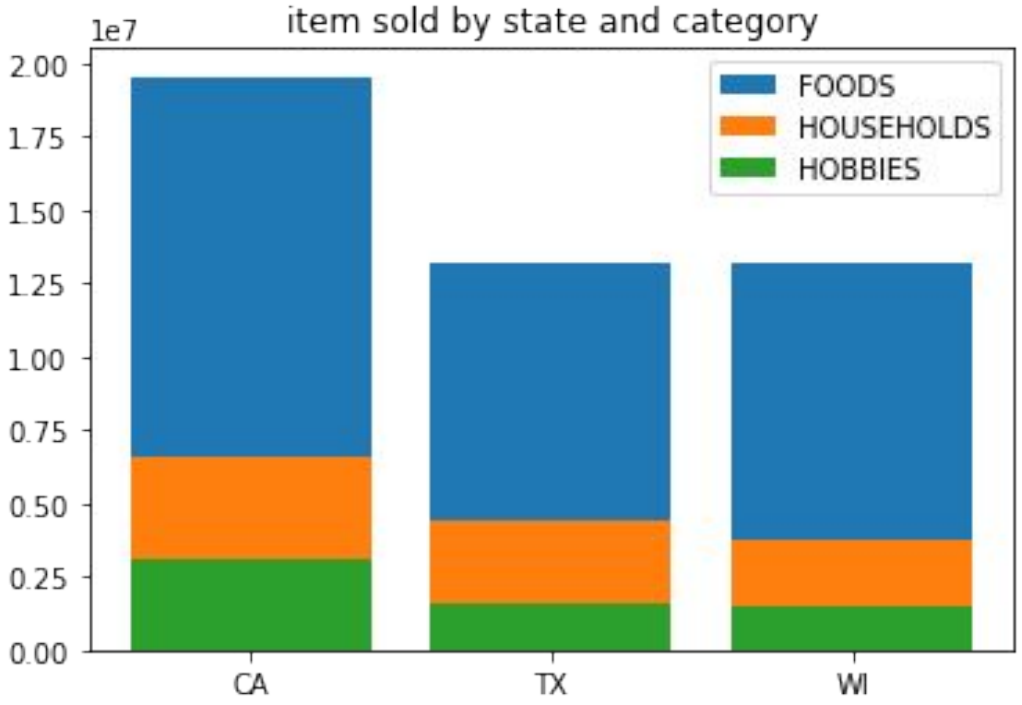
Task 5: Results (5/17)



Sales Distribution

FOOD product (Especially FOODS_3)has the widest range of item counts. With the higher lower bound, FOODS_3 is the most selling product across every States and every stores.

Task 5: Results (6/17)

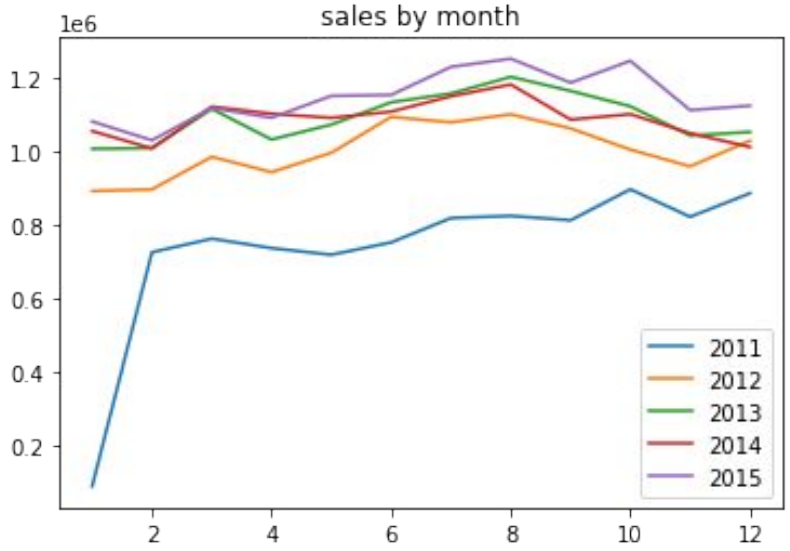


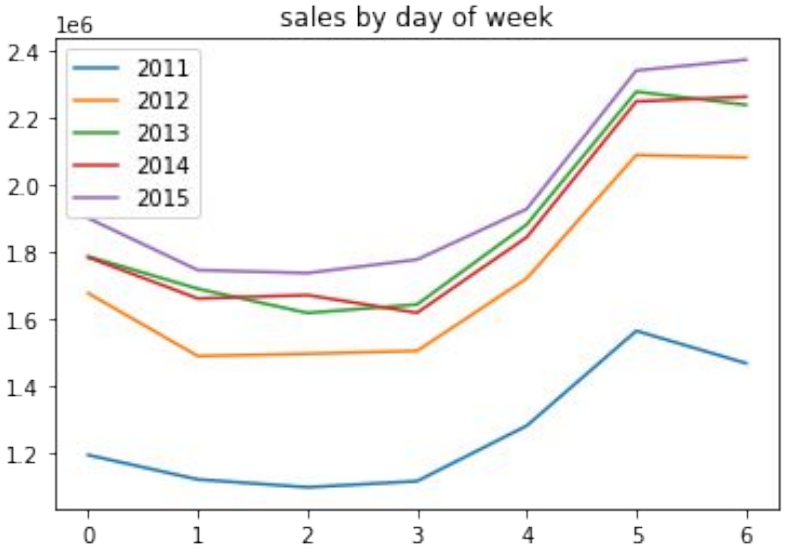
Item sold by States

Foods are the best selling category across all states.

In fact, foods made up to more than half of the total product sold in every States.

Task 5: Results (7/17)



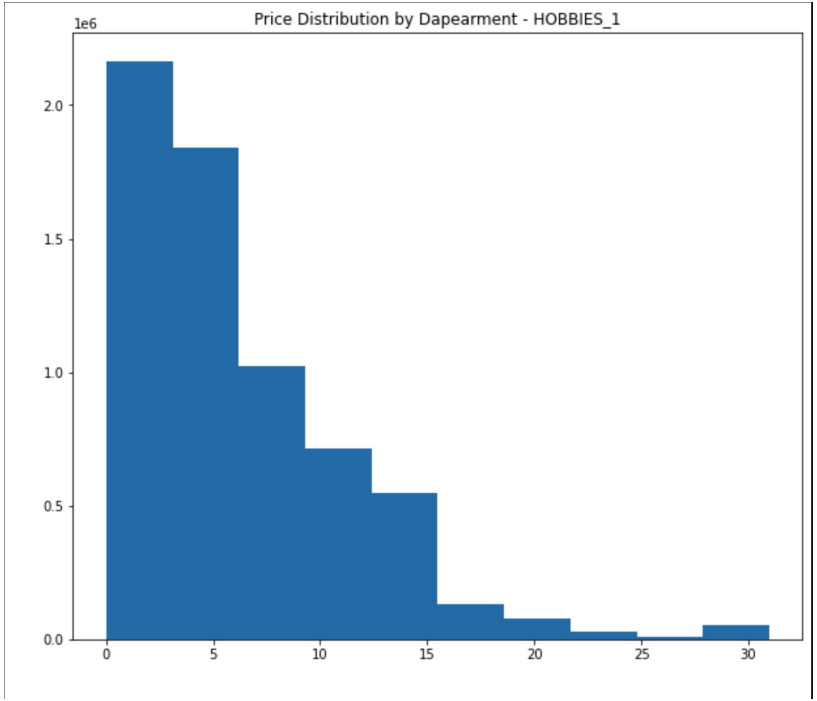


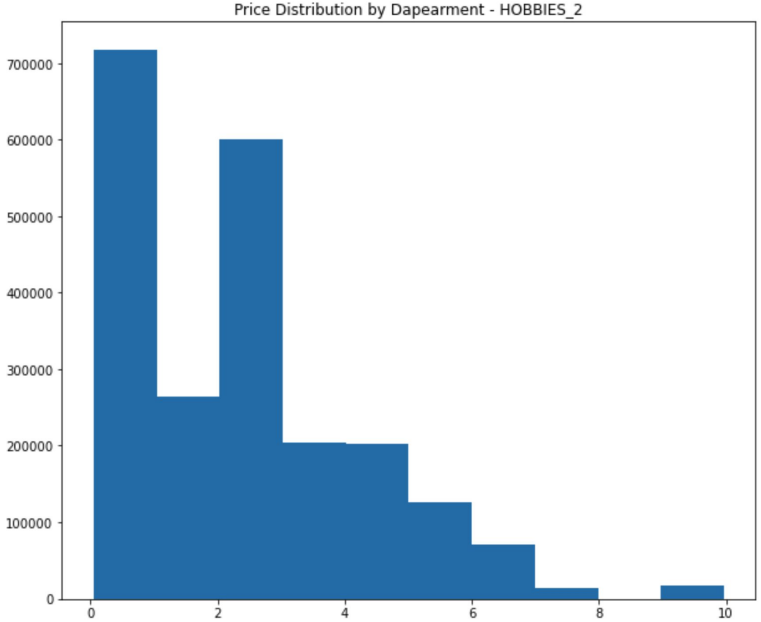
There are some seasonality. Product sells better in Q3 and Weekend.

Note 1: Monday is represented with 0

Note 2: The dataset starts at 2011-01-29. Therefore, Jan 2011 start at a very low level.

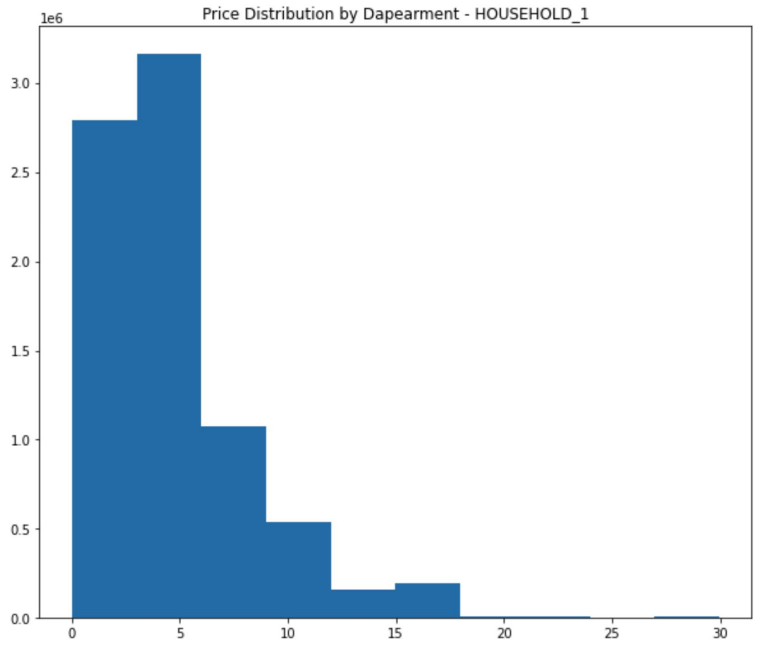
Task 5: Results (8/17)

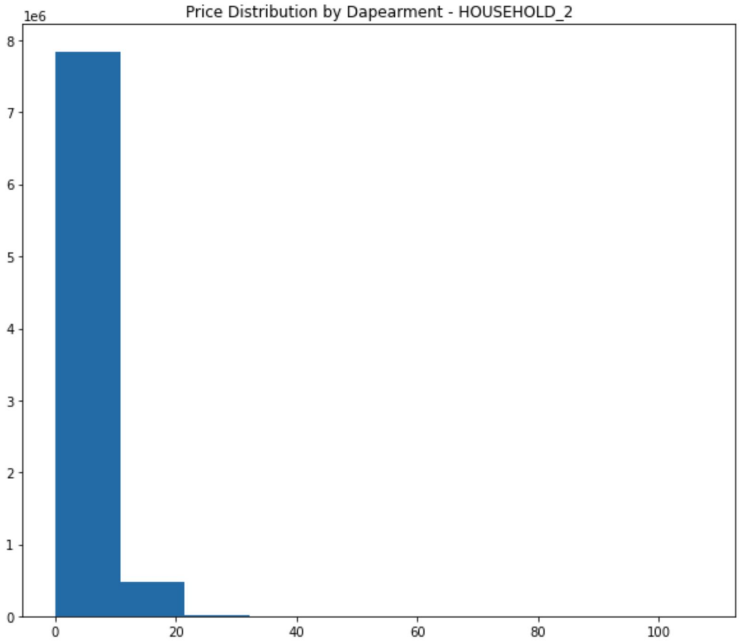




HOBBIES, the least selling products,cost lower than $5 for most of the HOBBIES items. Also most of the HOBBIES item sold are below $5

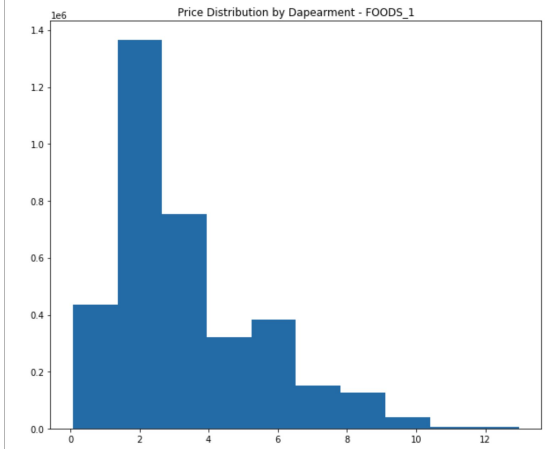
Task 5: Results (9/17)

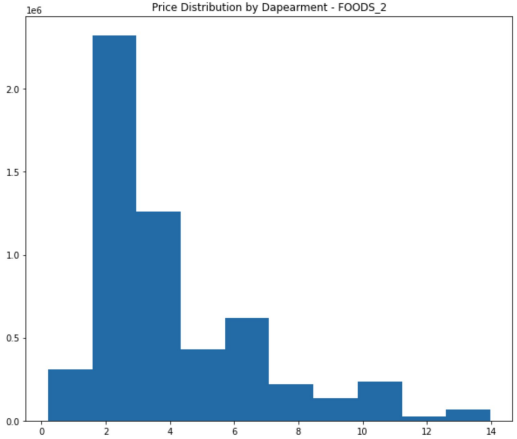


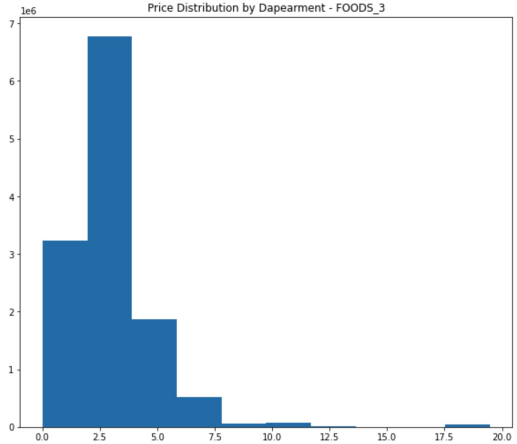


HOUSEHOLDS products are not expensive either. Almost all of the HOUSEHOLDS items cost between $0-$10.

Task 5: Results (10/17)

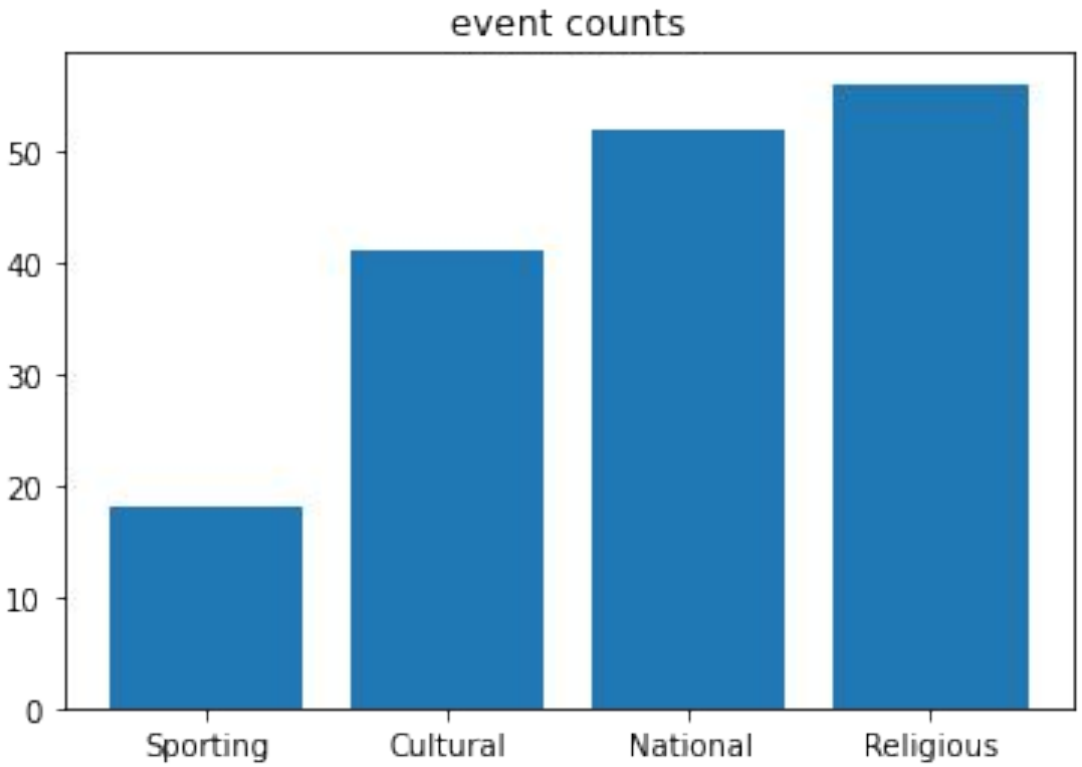






Food is the most selling product and it has an interesting price distribution. FOODS products price concentrated around $2.5-$4. However, more expensive foods has some record too My hypothesis is that Walmart stores are selling both budget FOODS and more premium FOODS. And Customer’s preferences varied.

Task 5: Results (11/17)



Events Distribution

Religious and National are the majority event categories.

There are 162 event days, or 8.23% of the total period.

Religious: 56 days

National: 52 days

Cultural: 41 days

Sporting: 18 days

Task 5: Results (12/17)

Average Weekday Sales	Average Weekend Sales	% difference
$ 90,348	$ 119,506	32.27%

Average Non Events Sales	Average Event Sales	% difference
$ 94,852	$ 99,041	4.41%

It is expected that weekend would boost sales significantly, however, to my surprise, events didn't boost sales that much. My assumption is that whenever it is holiday, most people want to go relax or travel instead of going to grocery store.

Task 5: Results (13/17)

FOODS SALES	Average non SNAP	Average SNAP	% of SNAP Boost
CA	$ 55,307.15	$ 61,227.22	10.70 %
TX	$ 54,707.31	$ 62,446.58	14.15 %
WI	$ 54,590.51	$ 62,684.01	14.83 %

SNAP

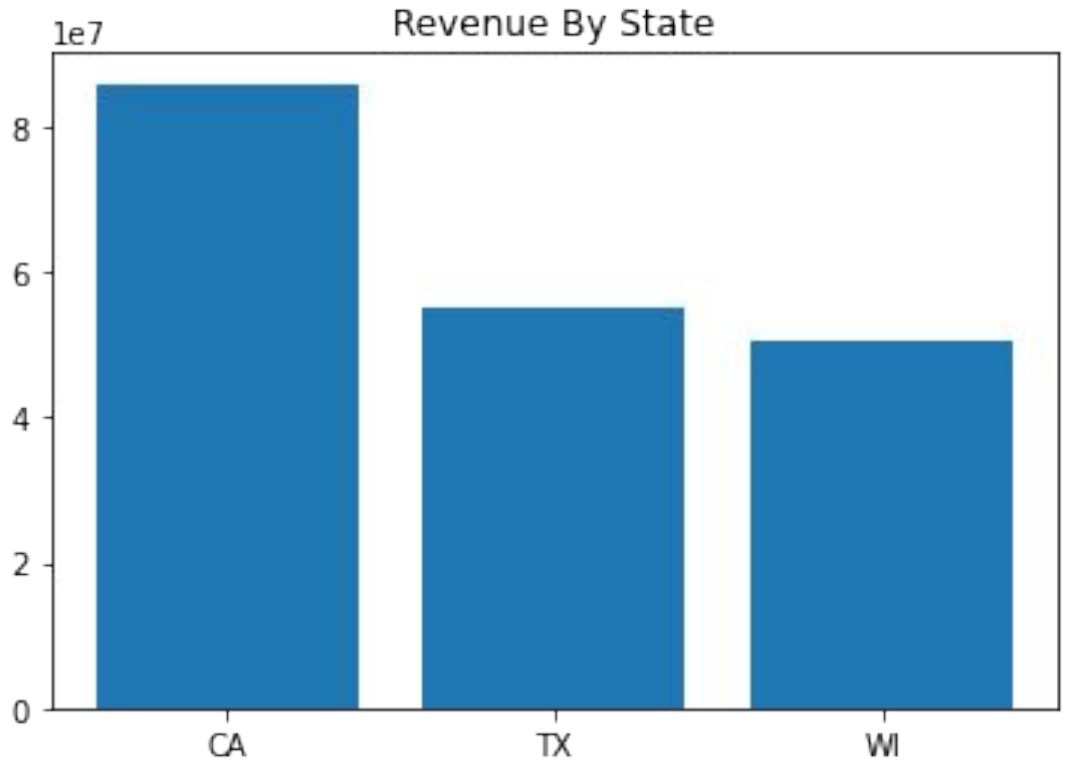
There are 650 SNAP days for each states that doesn't need to be the same day.

SNAP day is a day that the store accept the SNAP subsidy food coupon.

SNAP coupon boost FOODS sales more than 10%.

Considering that SNAP days are 650 days, or, almost 2 years, from the total of around 5 years, SNAP is a really good opportunity.

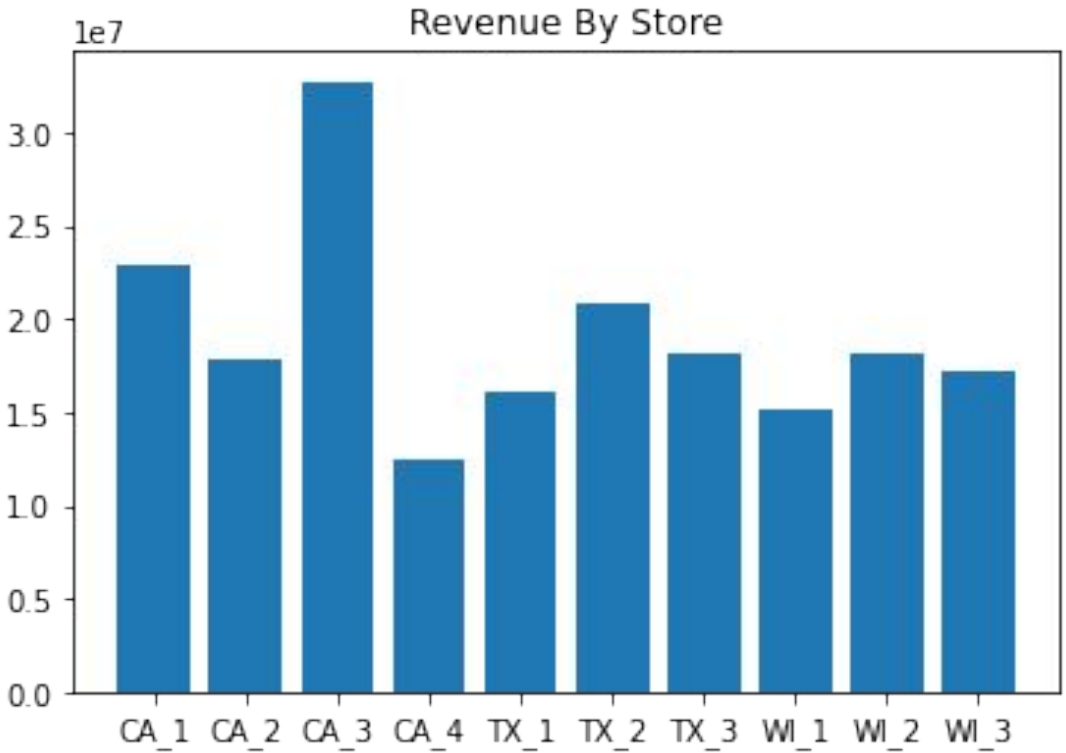
Task 5: Results (14/17)



Revenue By State

California revenue outperformed other States by a considerable amount.

Task 5: Results (15/17)

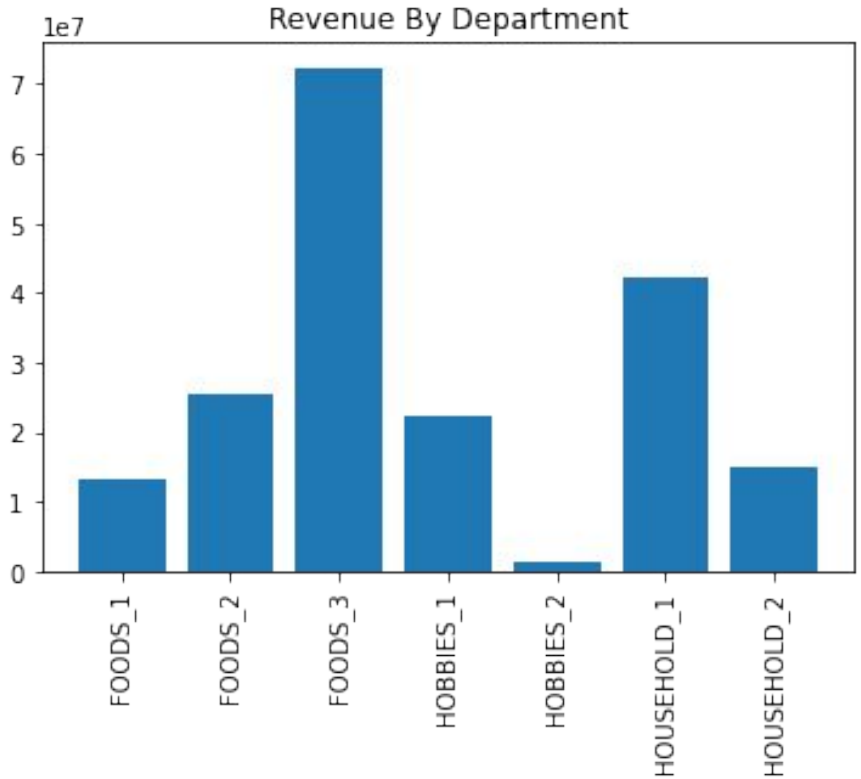


Revenue By Store

CA_3 generated the highest total revenue followed by CA_1, TX_2 and WI_2

One interesting characteristic is that California has the highest revenue variance among all of the States.

Task 5: Results (16/17)

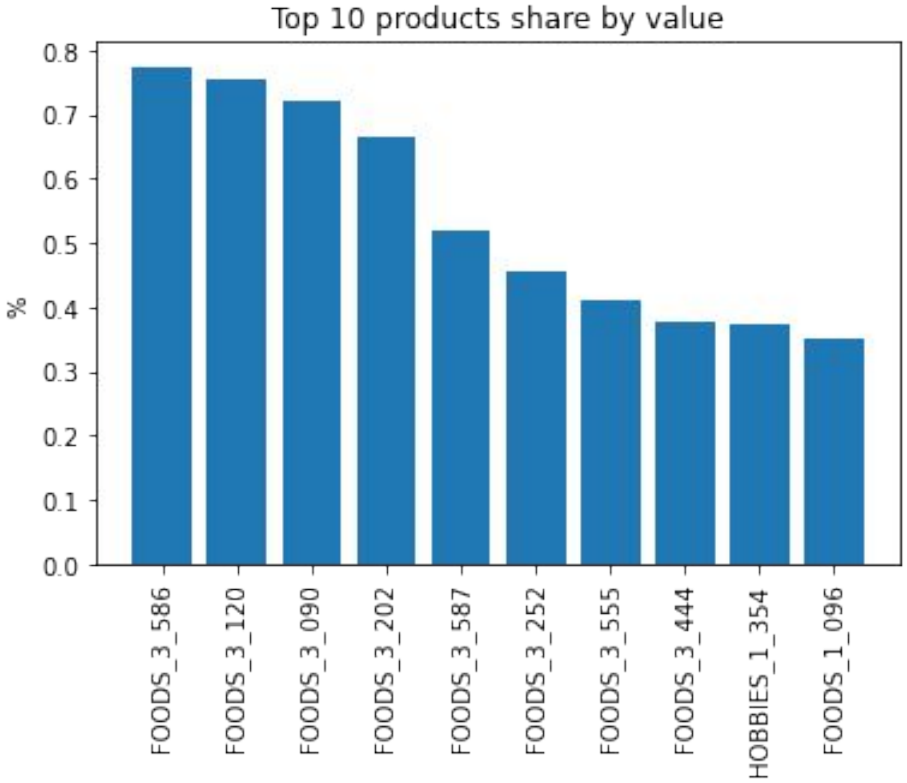


Revenue By Department

FOODS_3 (the cheapest food product on average) sells the most. In fact, almost 50% of the FOODS sold are FOOD_3.

HOUSEHOLDS_1 and FOODS_2 are the second and the third.

Task 5: Results (17/17)



Top 10 products share

FOODS_3 were 8 out of 10 of the top 10 product share

Though there are the top selling products, the share they have in total sales were below 0.8%

FOODS_3_586, FOODS_3_120,FOODS_3_090 were the best revenue generator

Task# 6

The analysis summary

A presentation by Thanak Rattanopastkul, 2022

Task 6: Summary (1/2)

- **Total Sales** had been increasing during the period of 2011-01-29 to 2016-05-22
- As the revenue picking up **California is the best performer** in both number of item sold and revenue
- For **Texas** and **Wisconsin**, even their sales are smaller, however, their sales are more consistent among their stores.
- **FOODS** sales is **over 50%** of total product sales
- The best selling product for every States are **FOODS_3**
    - Top 5 best selling products
        - FOODS_3_586
        - FOODS_3_120
        - FOODS_3_090
        - FOODS_3_202
        - FOODS_3_587
    - FOODS price distribution varied from $0 to $20. However the mode are around $2.5-$4
- **SNAP day** had 10% boosted in FOOD sales.
    - WI and TX had seen almost 15% boost.
    - SNAP record might imply lower personal income in TX and WI compared to CA
    - SNAP days are about 40% of the time
        - Note that SNAP policy and the numbers of SNAP days is subjected to changes.

Task 6: Summary (2/2)

- Compared to other kind of product, customer's varied the amount of food purchasing a lot. Sometimes they buy in bulk. Sometimes, they buy small.
- **HOUSEHOLD and HOUSEHOLD_1** are seconds in sales on category and department level
  - However the amount sold was **far below** of the FOODS
- **HOBBIES** was the least selling category
  - This might due to the very small product variety and customer consider Walmart a "grocery" store
  - Compared to everything else, **HOBBIES_2** had almost no sales.
- On average **Q3 is the best selling Quarters** Also, customers like to shop at **Weekend** especially at **Saturday**.
  - On average, weekends had sales increased for 32.27%
  - On average, days that has **events** had sales increased for 4.41%
- From the data we have, Walmart is considered "Grocery" Store

Task# 7

Insight Implementation

A presentation by Thanak Rattanopastkul, 2022

Task 7: Insight Implementation

- **California**
  - Depended on the constraints, but if we can boost CA_3 and CA_4 revenue, CA will be even better.
- **Strategies**
  - Add more product that compatible with the SNAP program since the SNAP day boost FOODS sales for more than 10% and there are a lot of SNAP days. Especially for TX and WI where SNAP boost FOODS sales for more than 14%
  - Some Up-Sales campaign in Q1-Q2 during Weekend to stabilize and boost revenue
  - HOBBIES_2 had some problem. It almost don't generate revenue and it had a relatively low product count. If it is associated with some considerable marginal cost, we should kill these HOBBIES_2 products.
  - Weekend is the prime time. In-store marketing should be implemented during weekend
  - Make sure that all the best sellings are well stocked
- **New Products**
  - For HOBBIES, emphasize on HOBBIES_1 items which are cheaper than $5 since almost all of the HOBBIES sales are HOBBIES_1 cheaper than $5
  - For HOUSEHOLDS, emphasize on HOUSEHOLDS_1 items which are cheaper than $5 since almost all of the HOUSEHOLDS sales are HOUSEHOLDS_1 cheaper than $5
  - FOODS_3 is the best selling product category. Since we know that SNAP helped a lot in food sales, if we want to add more food product in the store, we should consider FOOD_3 with SNAP compatible.

Task# 8

Further Supplementary Materials

Task 8: Further Supplementary Materials

1. **Data from Walmart Online Delivery Sales** and **Cumulative E-commerce market capitalization**
    a. So we could see the whole picture of Walmart sales. For example, currently we can't confirm that the declining in sales of some product was due to the customer migration to Online Shopping.
2. **Store location (at least at City scale) and opening hours**
    a. Different locations means different customers demographic
    b. Different opening hours mean that the total sales by store could be normalized using the total operation time of a store
    c. With **a.** and **b.**, the analysis would be more robust. The customer behavior could be mined from these informations
3. **Direct Competitors Performance**
    a. So we could see the dynamic in the industry, this data coupled with the information from **2.** could be used to fabricate strategic marketing plan for Walmart in the future.
4. **(If applicable) True products identity and data of other product categories**
    a. Many more strategies could be derived if we know what is what.

Task# 9

Deliver the analytic report **(this presentation)**

Task# 10

Deliver BI Dashboard

A presentation by Thanak Rattanopastkul, 2022

Task 10: BI Dashboard

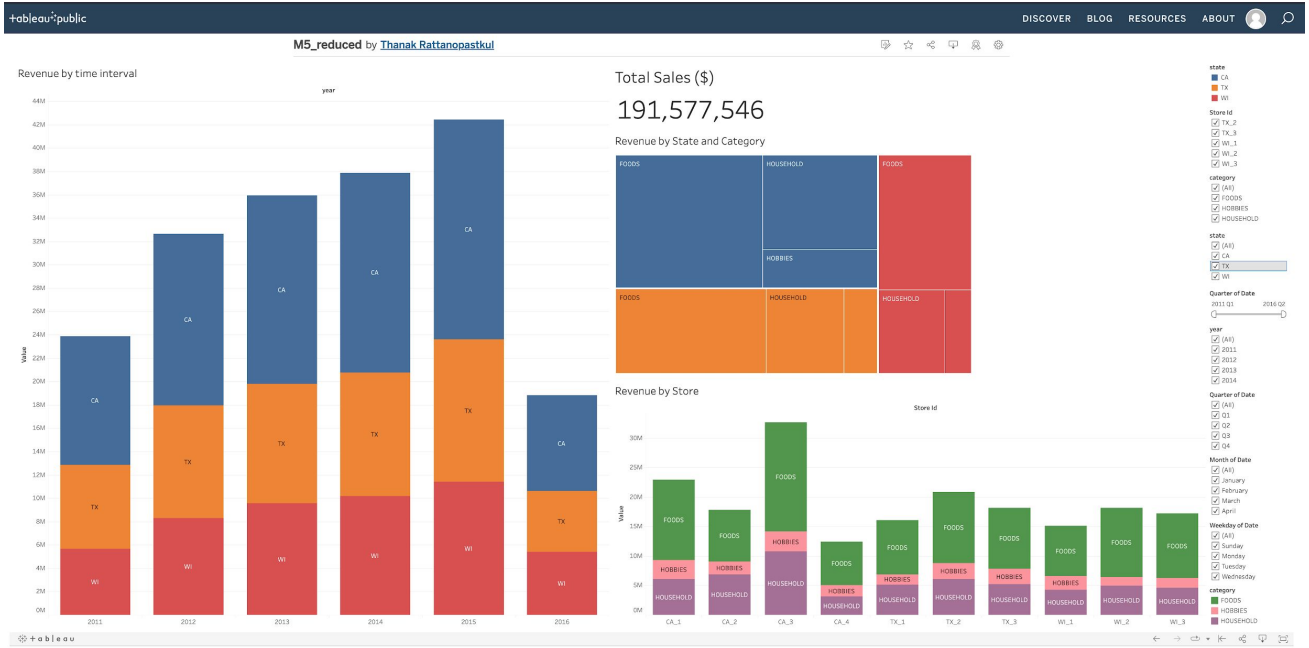


Tableau Public: The interactive BI dashboard for this project can be found here.

Task#11

Analysis plan during COVID-19

A presentation by Thanak Rattanopastkul, 2022

Task 11: Analysis plan during COVID-19

COVID - 19 ASSUMPTIONS:

1. There would be some occasional lock-downs or curfews.
2. Logistics might get shut down.
3. Economy in downturn
    a. Income effect
    b. Discontent workforce
    c. Political conflicts
4. Inflation from economics boosts policy.

Task 11: Analysis plan during COVID-19

Assumption#1: There would be some occasional lock-downs or curfews.

Expectations:

1. frozen product might get temporarily boosted from ppl preparing for lockdown
2. Might see some change in HOBBIES since ppl are stuck at home and need to relax
3. Might see some change in HOUSEHOLD since ppl are stuck at home and use more supplies
4. Store might close earlier during curfews

Changes in Analytics:

1. Take in account the temporarily change in shopping behavior
    - Any irregular changes in product sold patterns during COVID-19 have to be noted as a special situation.
    - Lower product sold and lower revenue are expected from the lockdown and curfews
    - YoY or QoQ analysis has to expect some dramatic change rate.

Task 11: Analysis plan during COVID-19

Assumption#2: Logistics might get shut down

Expectations:

1. Some items might be able to restock

Changes in Analytics:

1. If some item sales vanish during this period, we need to check the stock before concluding that the item just doesn't sell

Strategy Suggestion:

1. Make sure that all of the best selling items are well stocked (see slide #34).

Task 11: Analysis plan during COVID-19

Assumption#3: Economy in downturn

Expectations:

1. Some people at risk might lose their jobs, and, in turn, their income.
2. Coupled with the curfews and lockdown, we might see significantly lower sales.
3. There might be some protest, or dramatic changes in the government policy
4. There should be some subsidy policy form the government.

Changes in Analytics:

1. Expected a significantly lower revenue.
2. Expected a higher utilization of SNAP coupon

Strategy Suggestion:

1. Make sure that SNAP compatible product are well stocked
2. Adding cheaper products to the stores.
3. Reserve some cash in case of any emergency.

Task 11: Analysis plan during COVID-19

Assumption#4: Inflation from economic boosts policy

Expectations:

1. Higher product buying cost
2. Higher product selling price
3. Cost fluctuation in imported products

Changes in Analytics:

1. Some producers might halt their production.
    - If some item sales vanish during this period, we need to check the stock before concluding that the item just doesn't sell
2. Sales value and count will change due to the price change
    - Have to regards product price change from the inflation in the analytics
    - We could use this opportunity to learn a lot about "elasticity" of each product.

Strategy Suggestion:

1. If applicable, using contract buying to stabilize the cost
2. Using Forward or Future contracts to stabilize the cost of applicable products.