

CECS 551 ARTIFICIAL INTELLIGENCE

Artificial Intelligence (AI) Approach in Retail Market Analysis and Growth

***Group: 6*** ***Supervisor***

Dr. Mahshid Fardadi

* Ankit Ramrakhyani (029371899)
* Aparna Popade (026735460)
* Sarthak Jariwala (029341388)
* Shruthi Venkatchalam (029395663)
* Devansh Goel (029362552)

~ FALL 2022 ~

**Table of Contents**

[1. Introduction 4](#_Toc118496638)

[2. Dataset 01 5](#_Toc118496639)

[2.1 Problem Statement 5](#_Toc118496640)

[2.1.1 Analyze the dataset for CECS551\_dataset\_01 5](#_Toc118496641)

[2.2 Dataset Description: - 5](#_Toc118496642)

[2.3 Final Dataset Visualization: 5](#_Toc118496643)

[A).Identify the key variables for the model using correlation plots, heatmaps, histograms, feature Importance (SHAP). 5](#_Toc118496644)

[B).For the first 10 stores visualize the weekly and monthly sales patterns for top 35% of the Department sales. 10](#_Toc118496645)

[B.1 Identify the best department across the first ten stores 13](#_Toc118496646)

[C).Investigate the relationship between weekly sales over CPI and unemployment for the first 10 stores. You can explore the what-if scenarios while writing the report. 14](#_Toc118496647)

[D).Investigate the impact of various types of discounts, for example, discount promotional, discount clearance, discount damaged good, discount competitive and discount employee on the overall sales. 16](#_Toc118496648)

[D.1)Which types of discounts is helpful in increasing the sales? Consider top 30% of the best performing stores (sales per 1000 square feet). 18](#_Toc118496649)

[D.2) Does the observed behavior hold true for all the stores? Consider bottom 30% of the least performing store (sales per 1000 square feet). 19](#_Toc118496650)

[E). Identify the products which are highly impacted by external factors: “temperature”, “gas price”, and “holiday”. Is there any correlation between overall sales and holiday? 19](#_Toc118496651)

[2.4 Google Colab URL 24](#_Toc118496652)

[3. Dataset 2 24](#_Toc118496653)

[3.1 Problem Statement 24](#_Toc118496654)

[3.2 Dataset Description 24](#_Toc118496655)

[3.3 Final Dataset Visualization 25](#_Toc118496656)

[3.3.1 Selling price vs holidays based on categories 25](#_Toc118496657)

[3.3.2 Selling price based on holidays across three categories 26](#_Toc118496658)

[3.4 Tableau URL: 29](#_Toc118496659)

[4. Conclusion 29](#_Toc118496660)

**Abstract**

The project entails analyzing inventory data of two datasets of around more than 30 stores of an international retail business. It is designed to implement two-week sprints of the scrum process, mimicking a real tech company software development and machine learning work environment. The purpose of the analysis is to use the inventory data to improve sales, resulting in a more efficient operation.

Dataset 01: The task is to predict the department-wide sales for each store.

Dataset 02: The goal is to predict the unit sales for each product for next 10 days from 10 different stores across various states.

For Sprint1, we have used tableau and python for data visualization. For dataset 01, python is used to discover underlying trends between weekly sales and different explanatory features like CPI, size of store, gas price etc. For dataset 02, tableau is used for data visualization.

# **1. Introduction**

For Sprint1, we have used tableau and python for data visualization. For dataset 01, python is used to discover underlying trends between weekly sales and different explanatory features like CPI, size of store, gas price etc. For dataset 02, tableau is used for data visualization. The goal is to create a Tableau dashboard and communicate the results.

During the relevant sprint, we delivered the dataset to the team members and made an effort to complete the project's first milestone on time. The accompanying table shows what each team member provided as a result.

|  |  |  |
| --- | --- | --- |
| **Name** | **Dataset** | **Role and Contribution** |
| Aparna Popade | 1 | Scrum Master, Python code for dataset 1 and Report |
| Ankit Ramrakhyani | 1 | Report for sprint 1 for dataset1 |
| Shruthi Venkatchalam | 2 | Data Analyst for Dataset 2 |
| Sarthak Jariwala | 2 | Report for sprint 1 for dataset2, tableau publishing |
| Devansh Goel | 2 | Report Editor |

# **2. Dataset 01**

For dataset 1, we have created the Colab Notebook for data visualization.

## **2.1 Problem Statement**

### **2.1.1 Analyze the dataset for CECS551\_dataset\_01**

We performed the exploratory data analysis using Python for dataset\_01. Please use the right data visualization method for specific problem statements (choose the right chart type, for example, boxplot, histogram, scatterplot, pie-chart, etc.)

## **2.2 Dataset** **Description**

The Table below shows features of dataset 1 of a 35 Different Stores.

|  |  |
| --- | --- |
| **Features** | **Description** |
| Store | The Store Number |
| Type | Store Segregated into Three Types A, B And C |
| Size | Size Of the Store |
| Dept | Department ID |
| Date | MM/DD/YYYY Format |
| Is holiday | Yes/no |
| Weekly sales | Sales per week |
| Temperature | Temperature In Fahrenheit |
| Gas price | Price per gallon in $ discount |
| Promotional | Type Of Discount |
| Discount clearance | Type of discount |
| Discount damaged good | Type of discount |
| Discount competitive | Type of discount |
| Discount employee | Type of discount |
| CPI | The Consumer Price Index (CPI) |
| Unemployment present | The unemployment rate in the region where the store is |

## **2.3 Final Dataset Visualization:**

### **Identify the key variables for the model using correlation plots, heatmaps, histograms, feature Importance (SHAP).**

Chart, histogram

Description automatically generated

**Fig**: The visualization of the correlation Matrix. The scale on the Right-hand side of the matrix provides a coefficient of correlation.

***Inference:*** *From the correlation graph, there is very strong correlation between discount\_competetive and discount\_promotional. One of the reasons behind that would be addition of new product in market, either by same company or by the competitor.*

Shape, square

Description automatically generated

**Fig**: The distribution of Sales across all stores

***Inference:*** *we can see highly skewed sales data*

Chart

Description automatically generated with medium confidence

**Fig**: Feature importance using SHAP values

***Inference:*** *Above graph shows important features in determining the sales. Size of store & Type of store are crucial to determine Weekly\_sale.*

Chart

Description automatically generated

**Fig**: Beeswarm plot for SHAP values

***Inference:*** *From Beeswarm plot, when size of store is bigger avg sales goes up. One of the reasons behind that will be availability & variety of products available in store. Smaller store will have space restriction to add more stuff in inventory thereby less foot traffic and sales. One of the important points to note here is only biggest size causes more sales, medium and low store sizes show drop in average sale. Department being a categorical variable is still very much important to determine sales. It is clear from the fact that, one dept will perform better than other (Pharmacy, grocery), getting exact name of department will be useful to pinpoint Dept with highest sales. CPI & Unemployment shows negative trend with Sales. Increase in these two will cause sales to go down*

Timeline

Description automatically generated

**Fig**: Waterfall SHAP values for first record

***Inference:*** *This graph shows how smaller size and less department are negatively impacting sales for one of the stores. CPI & Temperature have positive impact on sales.*

### **For the first 10 stores visualize the weekly and monthly sales patterns for top 35% of the Department sales.**

Chart, histogram

Description automatically generated

**Fig**: Graphical representation of the data that shows the Monthly sales (in Millions) over a period of time for top 35% departments

***Inference:*** *It can be seen from the visualization of the data that the sales were lowest in the end of Jan considering holidays are already over*

Chart, line chart

Description automatically generated

**Fig:** This graph represents the Month wise data of sales (in Millions) for top 35% departments

***Inference:*** *From the visual output that can be seen above, we can infer that the sales were high in month of March because of spring break and low in Jan considering holidays already over*

Chart

Description automatically generated

**Fig:** The graph above depicts the data of weekly sales (in Millions).

***Inference****: It can be seen that there is a frequent fluctuation in the sales data, however it can also be seen that there are two major peaks in November because of Thanksgiving*

Text, whiteboard

Description automatically generated

**Fig:** This data Visualization is a representation of the data of weekly sales in millions in correspondence with week number.

***Inference:*** *We can see less sale during the start of the year*

#### **B.1 Identify the best department across the first ten stores**

Chart, bar chart

Description automatically generated

**Fig:** The above bar graph is a representation of sales patterns for top 35% of the department sales for first 10 Stores

***Inference:*** *This graph, indicates that the Dept 96 has very less sales with $22480259 whereas Dept 38 has maximum sales with $106373651.*

### **Investigate the relationship between weekly sales over CPI and unemployment for the first 10 stores. You can explore the what-if scenarios while writing the report.**

Chart

Description automatically generated

**Fig:**  the data above visually represents the relationship between the Weekly Sales and the CPI.

***Inference:*** *Its can be seen that there Is a moderate level of concentration of weekly sales in the early points of the CPI. However, the concentration of points goes way higher between 200 and 220 and around 220 CPI. This same level also has the lowest level of the CPI. There is no evidence of correlation between CPI and Weekly\_Sales. Lower CPI had higher weekly sales, however because of higher concentration of CPI around 220, its hard to draw conclusion for relationship.*

Chart

Description automatically generated with low confidence

**Fig:** This is a graphical version of the data that portrays the data which shows us the relationship between Weekly\_sales and Unemployment.

***Inference:*** *This graph clearly shows that there is a very uneven relation between both the metrics. The graph starts off at 30000 weekly sales with around 4 unemployment. Contradictory to this, there were nearly 5000 sales with 6 to 7 unemployment factors.*

### **Investigate the impact of various types of discounts, for example, discount promotional, discount clearance, discount damaged good, discount competitive and discount employee on the overall sales.**

Chart

Description automatically generated

**Fig:** Sales Figures differences with or without Promotional Discount.

***Inference:*** *This pictorial representation clearly shows that there are slightly more sales when Promotional Discount are taken into consideration.*

Chart

Description automatically generated

**Fig:** Total Sales with or without Clearance Discount.

***Inference:***  *This is a clear demonstration of the statistics that shows that there are slightly more sales seen when the Clearance discount provided.*

Chart

Description automatically generated

**Fig:** Total Sales when Damaged Good Discount given or not.

***Inference:*** *It can be seen that the Damaged Good Discount is provided; the sales are slightly increased.*

Chart

Description automatically generated

**Fig:** Total Sales when Competitive Discount is given or not.

***Inference****: Another visual representation of the fact that when Competitive Discount are involved into the equation, the sales tend to go up.*

#### **D.1** **Which types of discounts is helpful in increasing the sales? Consider top 30% of the best performing stores (sales per 1000 square feet).**

Chart, bar chart

Description automatically generated

**Fig:**  Graphical representation of the variation of the increment of the sales when different Discounts are taken into consideration for top 30% stores.

***Inference:*** *In Comparison to all the* *discounts given for top 30 percent Stores,* *sales are seen to be maximum when Promotional Discounts are given.*

#### **D.2** **Does the observed behavior hold true for all the stores? Consider bottom 30% of the least performing store (sales per 1000 square feet).**

Chart, bar chart

Description automatically generated

**Fig:** Graphical representation of the variation of the increment of the sales when different Discounts are taken into consideration for bottom 30% stores.

***Inference:*** *Similar to the previous result, it can be seen here that Promotion Discounts perform way better in terms of sale in comparison to other sorts of discounts.*

### **E.** **Identify the products which are highly impacted by external factors: “temperature”, “gas price”, and “holiday”. Is there any correlation between overall sales and holiday?**

A picture containing chart

Description automatically generated

**Fig:** Graphical representation of effect of Temperature on Sales for Dept 38.

Chart, scatter chart

Description automatically generated

**Fig:** Graphical representation of effect of Temperature on Sales for Dept 95

Chart, scatter chart

Description automatically generated

**Fig:** Graphical representation of effect of Temperature on Sales for Dept 92.

***Inference from all the figures above considering the Temperature:*** *Relation between temperature and Sales is not linear. Low and high temperature sees lower sales, considering unwillingness of people to come out it too hot or too cold weather. Drawing the trendline shows how temperature is affecting top 3 Department. With temperature, sales for****Dept =38****goes down.*

Chart, scatter chart

Description automatically generated

**Fig:** Graphical representation of effect of Gas Price on Sales for Dept 38

Chart, scatter chart

Description automatically generated

**Fig:** Graphical representation of effect of Gas Price on Sales for Dept 95

Chart, scatter chart

Description automatically generated

**Fig:** Graphical representation of effect of Gas Price on Sales for Dept 92

***Inference from all the figures above considering the gas prices:*** *General trend for sales with gas prices is negative. With increase in gas prices, sales go down.****Department =38****is worst affected by gas prices. To further analyze it would be important to look at distance of store from nearby neighborhood.****Department=95****is least affected by gas prices.*

Chart, bar chart

Description automatically generated

**Fig:** Graphical representation of the variation of the increment of the sales when different Discounts are taken into consideration for bottom 30% stores.

***Inference: Department = 92****has higher average sales during Holidays. Other two departments have lower sales during holiday.*

Chart

Description automatically generated

**Fig:** Graphical representation of Average Sales with Holiday.

***Inference:*** *On an average, sales during holidays is more than non-holiday seasons.*

Chart, bar chart

Description automatically generated

**Fig:** Graphical representation of Average Sales with Holiday

***Inference:*** *With no surprise, Thanksgiving sales is highest. Customers tend to do most of the holiday shopping during that time. As a result, sales during Christmas is lower than usual. There can be multiple factors for this, one is pre shopping, other can be winter, people tend to not go out and celebrate with family*

## **2.4 Google Colab URL**

*<*[*https://colab.research.google.com/drive/1paontoradBN5KuUtzlcBBbmquBtWpXcH?usp=sharing*](https://colab.research.google.com/drive/1paontoradBN5KuUtzlcBBbmquBtWpXcH?usp=sharing)*>*

# **3. Dataset 2**

For dataset 2 we have deployed the tableau dashboard.

## **3.1 Problem Statement**

1. Visualize the daily, monthly, and total sales in STATE1 and STATE2. Identity the department with the highest and lowest sales across all stores. You may make a comparison of “State” with “Item Categories”.
2. Determine the product sales and availability over time in each store, and average price on each category.
3. Visualize the monthly and total sales for each store and each category. Identify the most sold products in each state in each department.
4. Investigate the average sales of highly sold products for a given price of each month and on weekdays, and average sales on event types, for example, cultural, national, etc.

## **3.2 Dataset Description**

* **calender.csv**

|  |  |
| --- | --- |
| **date** | date |
| **weekday** | categorical |
| **wday** | weekdays in numeric month – month in numeric year – year in numeric |
| **d** | each day assigned in sequential order |
| **event name 1, event name 2** | the name of events |
| **event type 1, event type 2** | the type of events |
| **snap STATE1 and STATE2** | snap is a nutritional program for low-income families. |

* **data\_test.csv and data\_train.csv**

|  |  |
| --- | --- |
| **id** | product id |
| **item id** | items |
| **dept id** | department |
| **cat id** | category |
| **store id** | store id |
| **state id** | state id |
| **d 1 - d 1941** | day 1 to day 1941 |

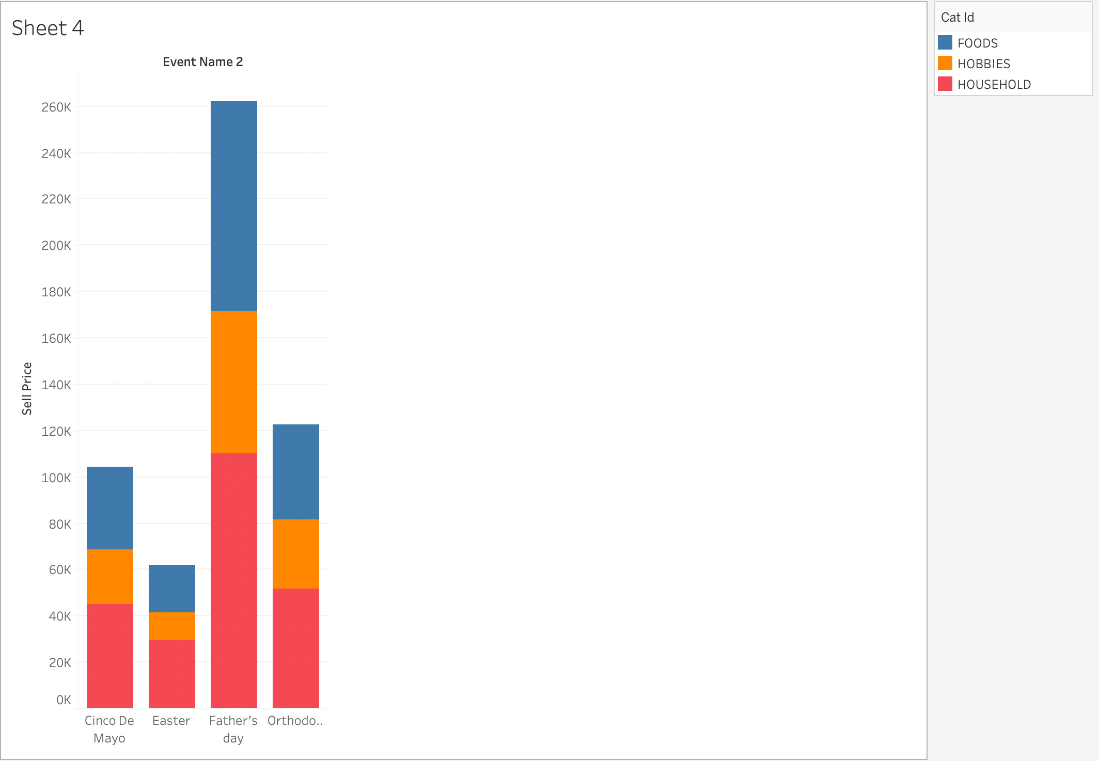
* **price.csv**

|  |  |
| --- | --- |
| **store id** | store id |
| **item id** | item id |
| **sell price** | selling price |

## **3.3 Final Dataset Visualization**

We have analyzed dataset 2 in the following tableau dashboard (link to be displayed) and have described the key findings as follows.

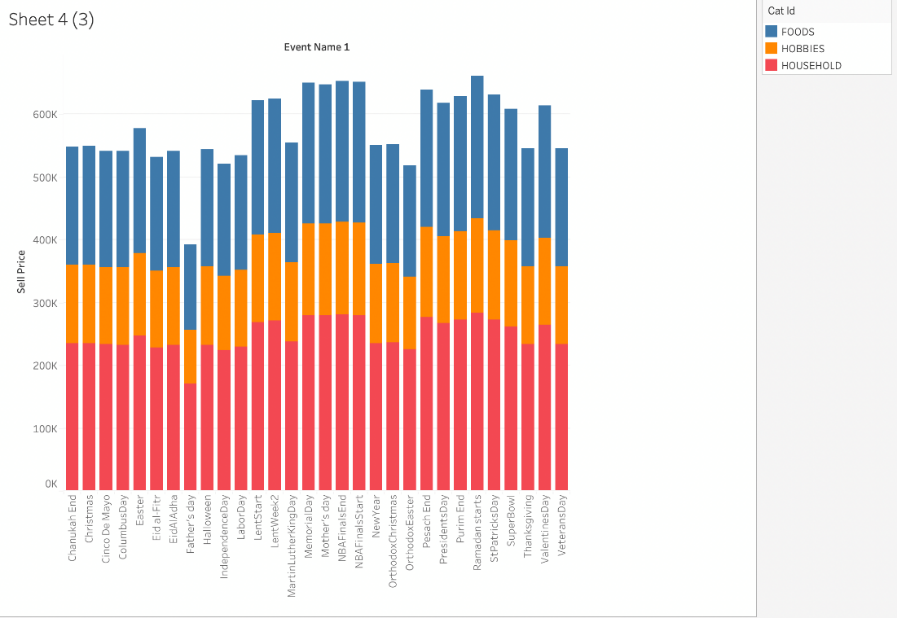
### **3.3.1 Selling price vs holidays based on categories**



**Fig:** This data visualization represents the selling price on the 4 holidays for the categories of Food, Hobbies and Households.

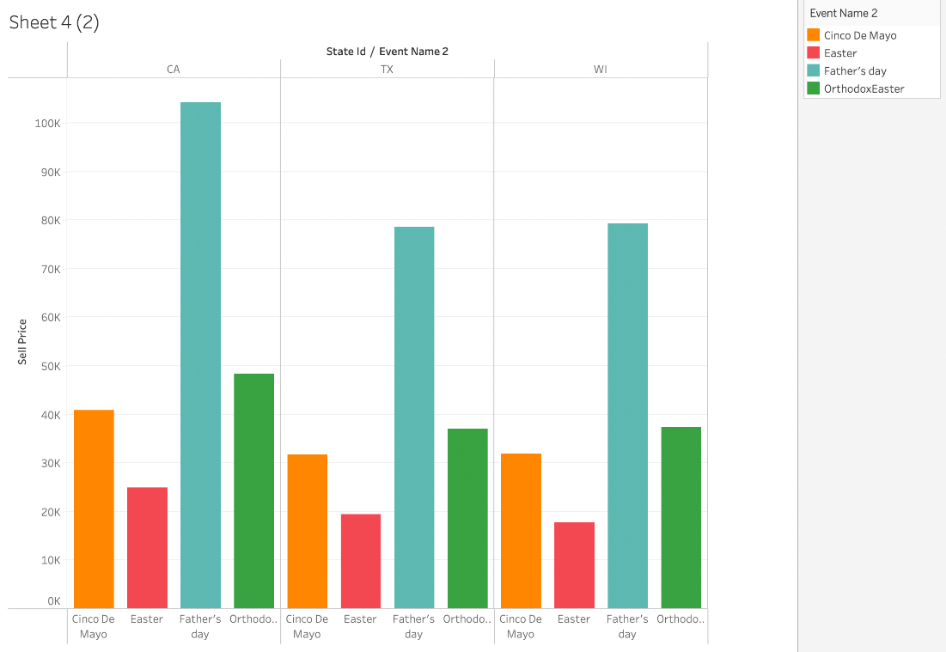
***Inference****: For any category, Easter has the lowest selling price.*

### **3.3.2 Selling price based on holidays across three categories**



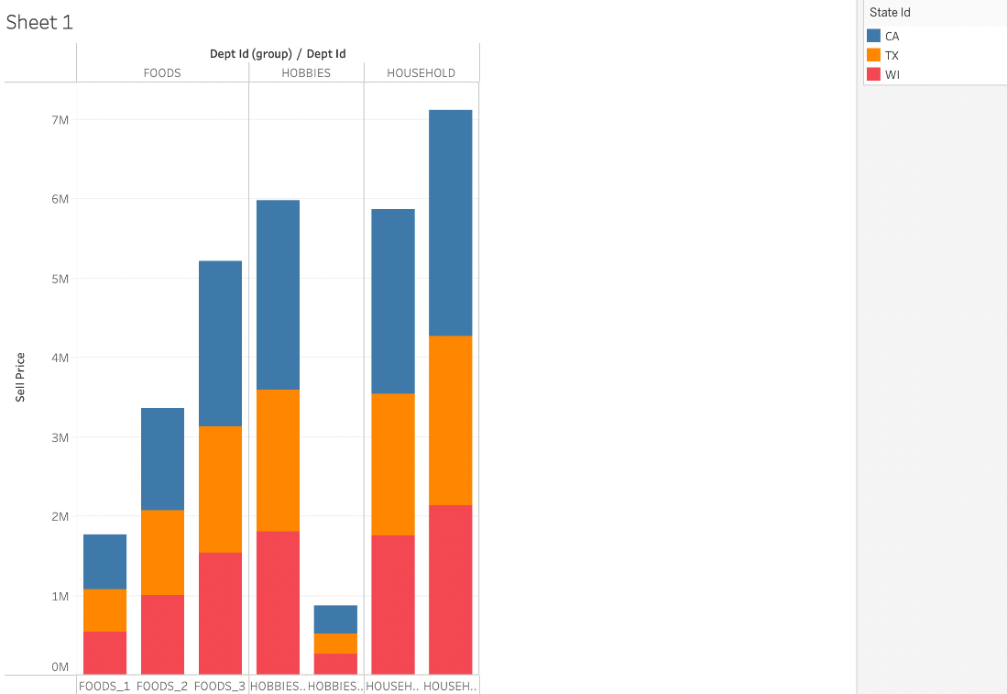
**Fig:** This is a much more detailed visualization of the graphs that represent the selling price of the 3 categories across the multiple holidays (more than 4)

***Inference****: It can be seen that Father's Day has the lowest whereas Ramadan Starts Day has the highest selling price*



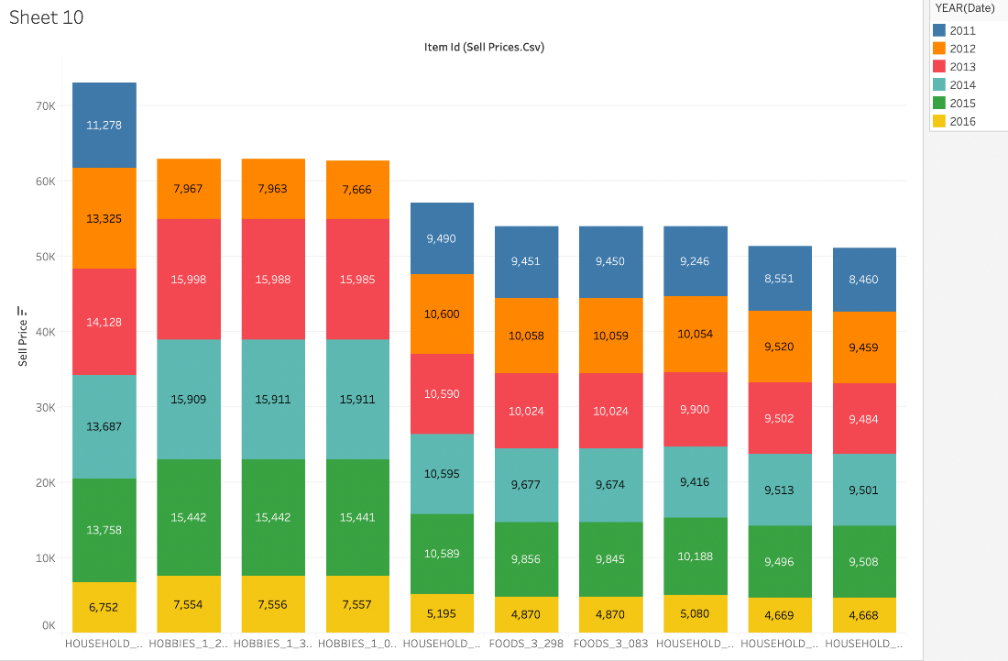
**Fig:** This Graph is a visual representation of the data that shows the selling price on multiple days (in this case 4), in 3 states (California, Texas, Wisconsin)

***Inference****: California has the highest selling price on Father's Day. Overall easter has lowest number of selling price in all the states.*



**Fig:** This visual representation of the data depicts the multiple items in the 3 categories (Food, Hobbies, Households) in 3 states (California, Texas, Wisconsin)

***Inference****: 2nd micro category of Hobbies has the lowest selling price across all the states however 2nd micro category of the Household has the highest selling price across all the states.*



**Fig:** This graph represents the data that shows the selling price of the multiple categories (Household, Hobbies, Foods) that are further divided into micro categories and shows the percentage ratio of the selling price in the multiple years (from 2011 to 2016).

***Inference****: The graph clearly depicts that 1st micro category of Households has the most significant selling price from 2011 to 2016. However, in the last micro category of Households, the selling price spanning 6 years, from 2011 to 2016, is the lowest.*

Chart, scatter chart

Description automatically generated

**Fig:** The plotted graph above displays the sum of the Sell Price for multiple events in Event Name 1.

***Inference:*** *It is clear from the graph that the selling price for each day in the Event Name 1 in 3 States (California, Texas, and Wisconsin) are almost always above 150K with the selling price for the state of California almost always above 200K. It can also be seen that. The selling price for Texas and Wisconsin are nearly the same.*

## **3.4 Tableau URL**

*<*[*https://public.tableau.com/app/profile/sarthak.jariwala/viz/Group\_06\_1\_Popade\_2\_Ramrakhyani\_3\_Goel\_4\_Jariwala\_5\_Venkatachalam/Final\_Dashboard?publish=yes*](https://public.tableau.com/app/profile/sarthak.jariwala/viz/Group_06_1_Popade_2_Ramrakhyani_3_Goel_4_Jariwala_5_Venkatachalam/Final_Dashboard?publish=yes)*>*

# **4. Conclusion**

As for the Sprint 1 we understood the business scenario from the given dataset for an organization.

We observed few key findings for two different datasets.

Data Set 01:

* There are high sales if any type of discount is given to customers and among all of them promotional discount is one to attract Customers.
* Size, Type of store are crucial factors in determining its weekly sales.
* Sales gets decrease when: -
  + Too hot or too freezing weather.
  + Increase in gas price
* But Sales during holiday are more compared to non-holiday seasons, and Thanksgiving holiday sales is highest compared to other holiday sales.

Data Set 02:

Household\_2 item #446 seems to have been the highest selling product amongst all in the given company. Though Father’s Day had higher spending trend in event\_2, they have been one of events on which the spending was done moderately less in event \_1. We could also see a trend that people tend to spend more on household products on occasions when compared to spendings made on hobbies and other activities. Spending trends in TX and WI seem to have similar spending trends on occasions when compared to that of CA.