Published in Towards Data Science

Jaswanth Badvelu Follow

Jun 19, 2020 · 12 min read · Listen



# M5 Forecasting- Accuracy

## Forecasting comparison using Xgboost, Catboost, Lightgbm

Photo by Jamie Street on Unsplash

**Background of Competition:**

The **Makridakis Competitions** (also known as the *M Competitions*) are series of open competitions organized by teams led by forecasting researcher Spyros Makridakis and intended to evaluate and compare the accuracy of different forecasting methods. he first competition named M-Competition was held way back in 1982 with only 1001 data points, the complexity of the model and data scale increased with every successive iteration.

Link to competition:https://www.kaggle.com/c/m5-forecasting-accuracy

**Aim:**

In March this year(2020), the fifth iteration named M5 competition was held. This m5 competition aims to forecast daily sales for the next 28 days i.e., till 22nd May 2016, and to make uncertainty estimates for these forecasts. In this blog, I am just going to do forecasting and uncertainty will be performed in my next blog with the best-chosen model.

**Dataset:**

The dataset provided contains 42,840 hierarchical sales data from Walmart. The dataset covers stores in three US states (California, Texas, and Wisconsin) and includes item level, department, product categories, and store details for 5 years starting from 29th Jan 2011 to 24th April 2016. Also, it has explanatory variables such as price, snap events, day of the week, and special events and festivals.

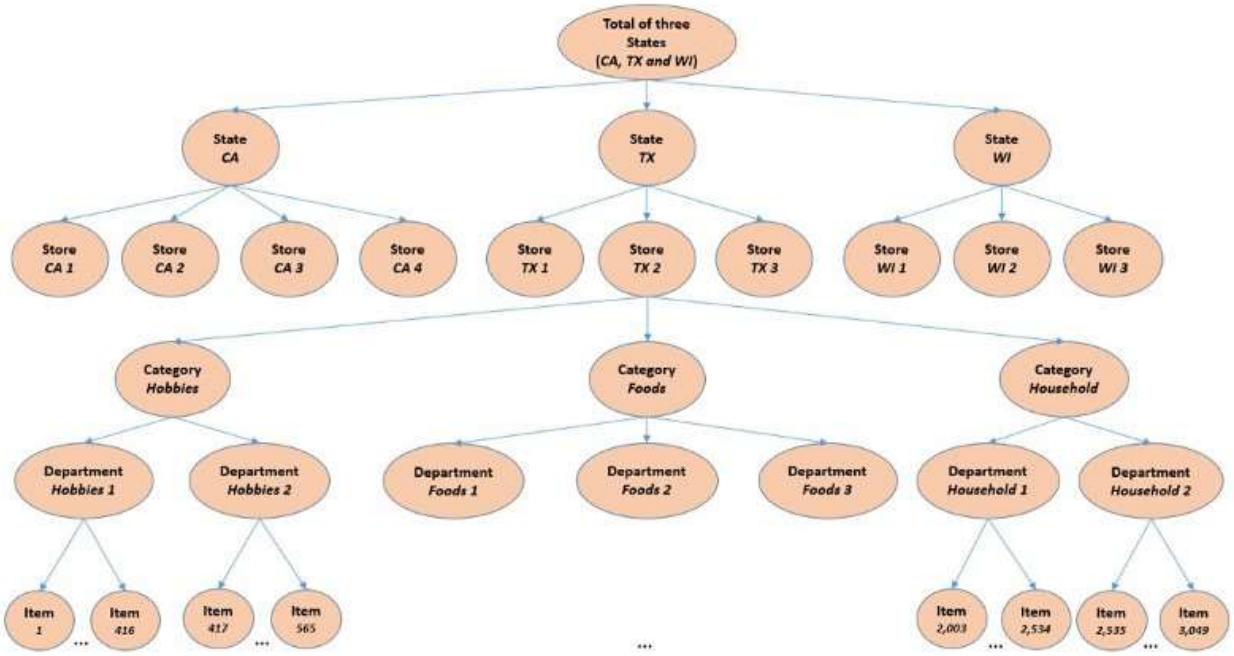



**Figure 1: An overview of how the M5 series data is organized**

The data comprises **3049** individual products from 3 categories and 7 departments, sold in 10 stores in 3 states. The hierarchical aggregation captures the combinations of these factors which makes it feasible to perform a bottom-up approach or top-down approach. For instance, we can create 1 time series for all sales or perform for each state separately and so on.

## Hypothesis

Based on the data given some of the factors that may affect sales are:

1. **Day-** Customers shopping time and spending mostly depends on the weekend. Many customers may like to shop only at weekends.
2. **Special Events/Holidays:** Depending on the events and holidays customers purchasing behavior may change. For holidays like Easter, food sales may go up and for sporting events like Superbowl finals Household item sales may go up.
3. **Product Price:** The sales are affected the most by the product price. Most customers

products.

5. **Location:** The location also plays an important role in sales. In states like California, the customers might buy products they want irrespective of price, and customers in another region may be price sensitive.

Before diving deep into data exploration, A quick overview of population & Median Income for each state:

#### California

Population: 39.51 Million

Median Household Annual Income: $75,277

#### Texas

Population: 29 Million

Median Household Annual Income: $59,570

#### Wisconsin

Population 5.822 Million

Median Income: $60,733

The exploratory data analysis was done to test these hypothesis statements.

## Exploratory Data Analysis

Let's start data analysis by knowing which state recorded the highest sales and also the individual department sales in each of these three states.

#### Exploring the location of stores

This section aims to answer:

1. Which state has the highest sales?

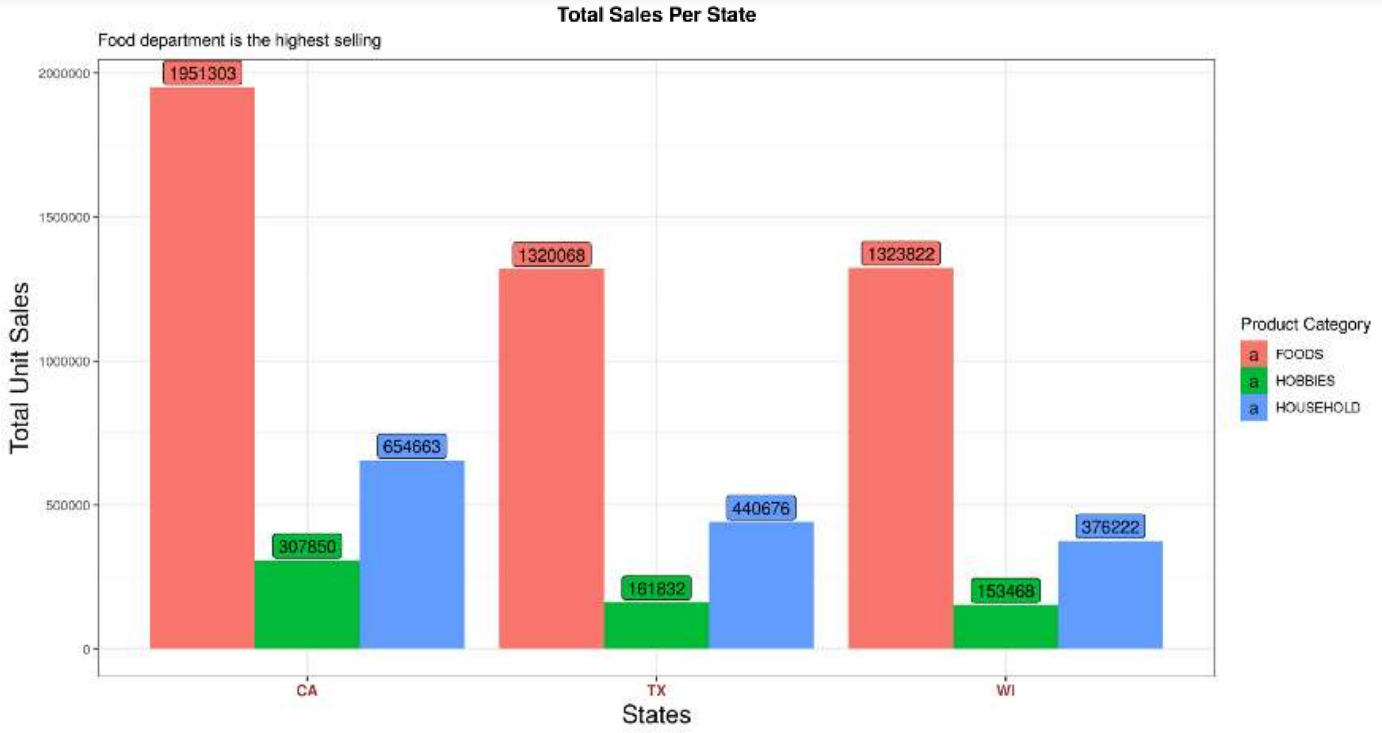



Fig 2. Total sales per state

As expected, the Food department recorded the highest sales in all 3 states. Also, It can be seen from fig 2 that California had the highest sales overall. Having 4 stores and more population might be the reason. Surprisingly, Wisconsin even with low population density when compared to Texas recorded equal sales. To get better understanding sales for each store are plotted.

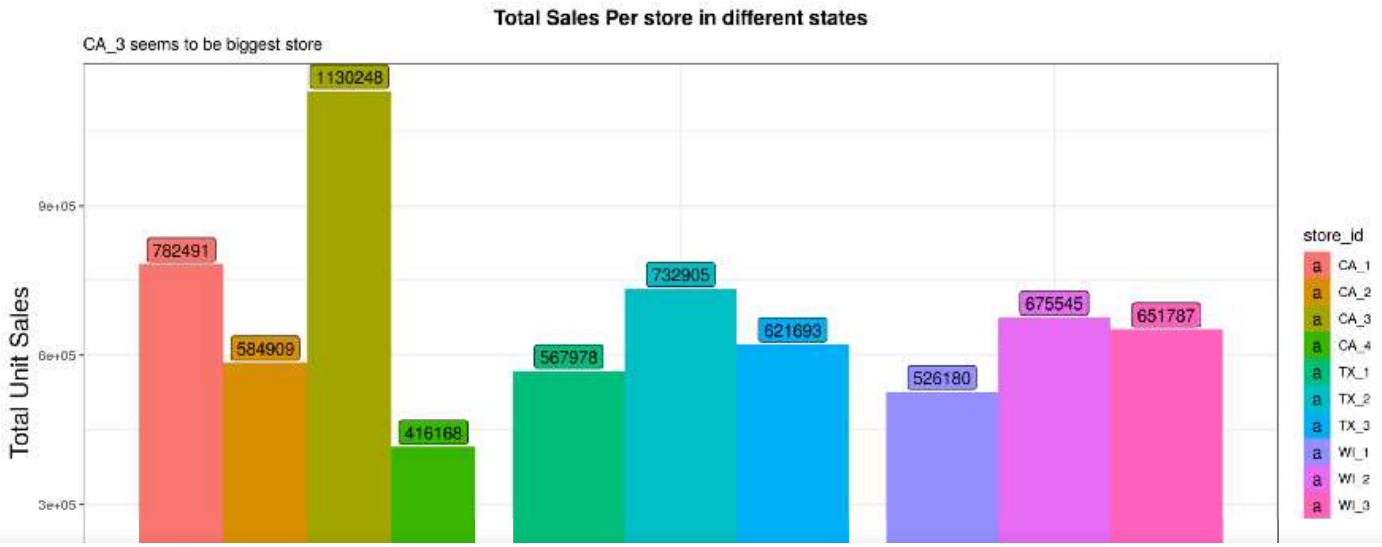


Fig 3.Store wise total sales

Since sales are almost double for CA_3 when compared with other stores. The CA_3 may be a bigger store. The population density and median income also affect these sales.

Now we have an idea about how sales are impacted by a different location, now let's explore the individual department for insights

**Exploring price & product category**

This section aims to answer:

1. How many different products are available in each department?
2. What is the mean price of all the available products across different states?
3. Which department has the highest and least sales?

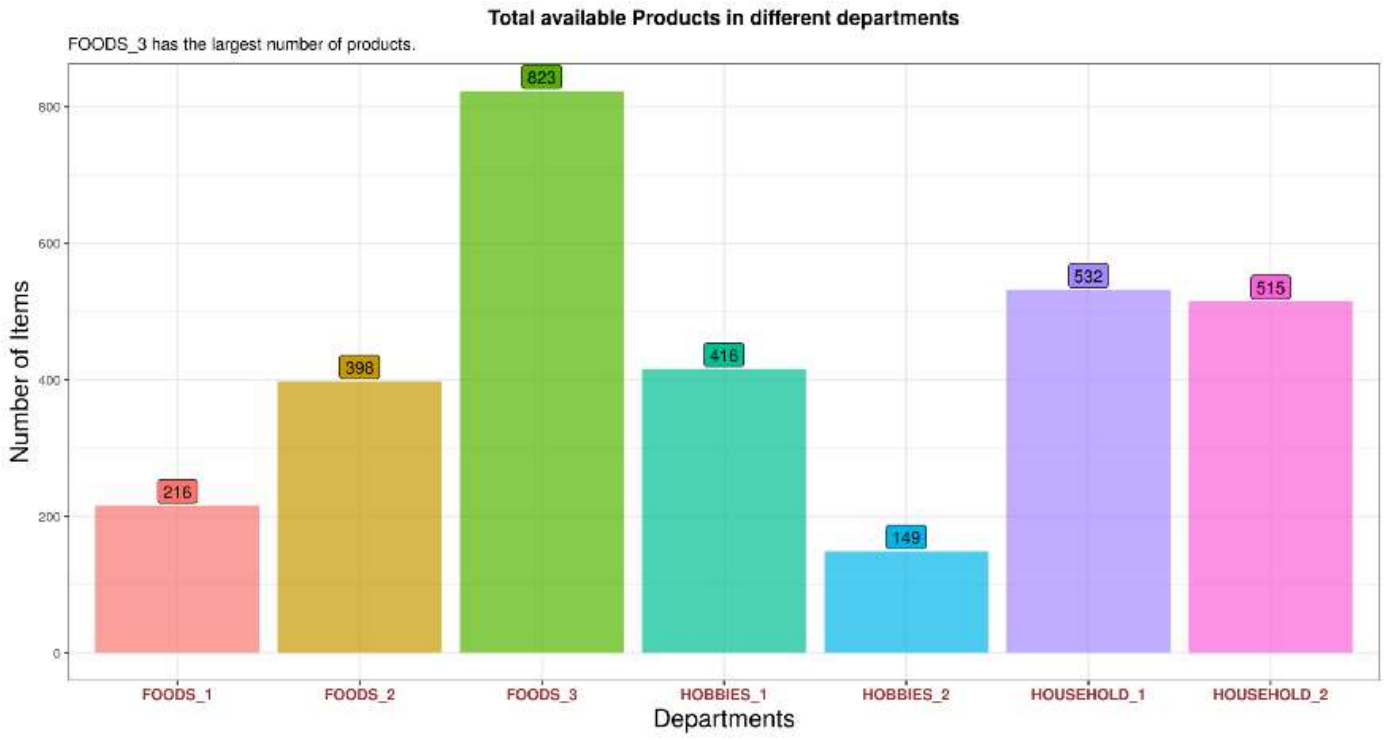



Fig.4 Count of total available products

The number of available products is more in the Food_3 department. So Food_3

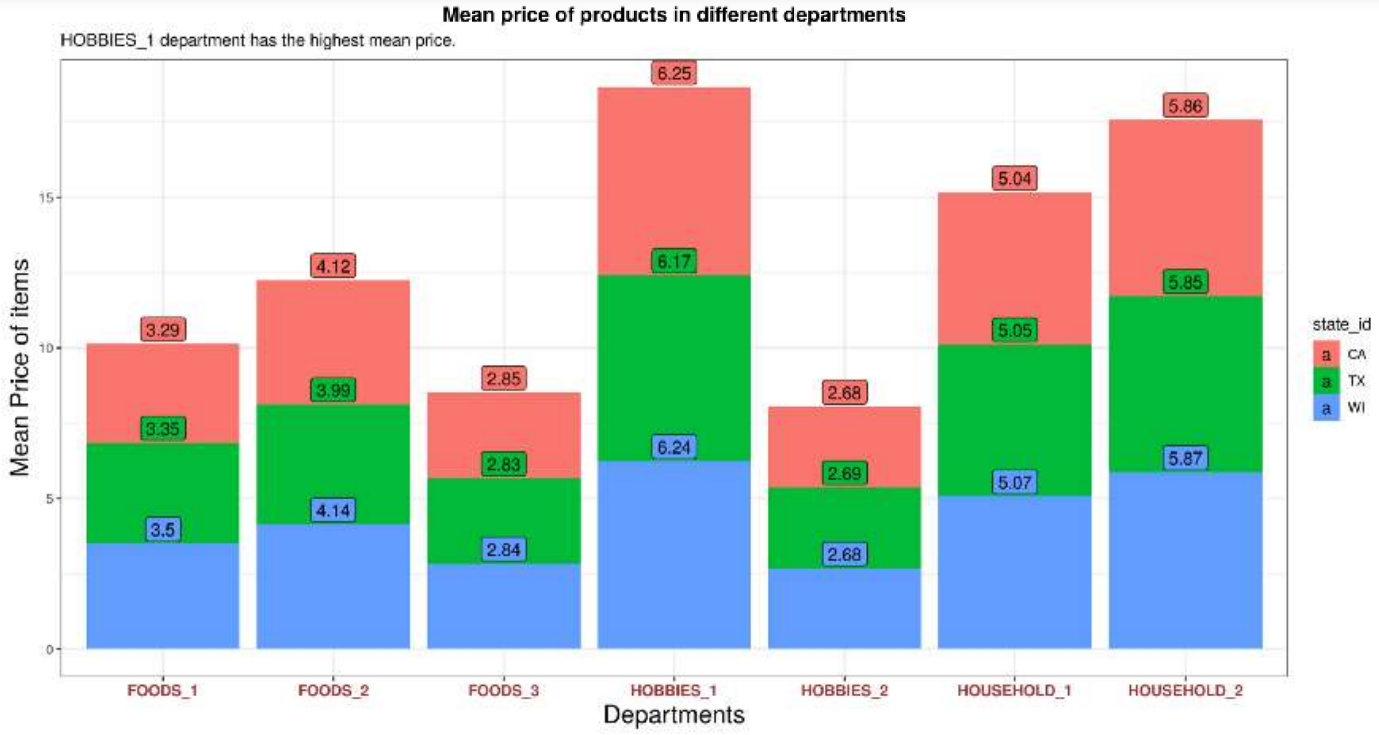



Fig 5. Mean price of all products

It can be seen that the Hobbies_1 department has the highest mean price and Food 3 being the lowest. Despite, California state population having more mean annual household income when compared to texas and Wisconsin, the mean price is almost similar for 3 states which makes the products more affordable for the California state population. Now let's see state wise sales for each product.

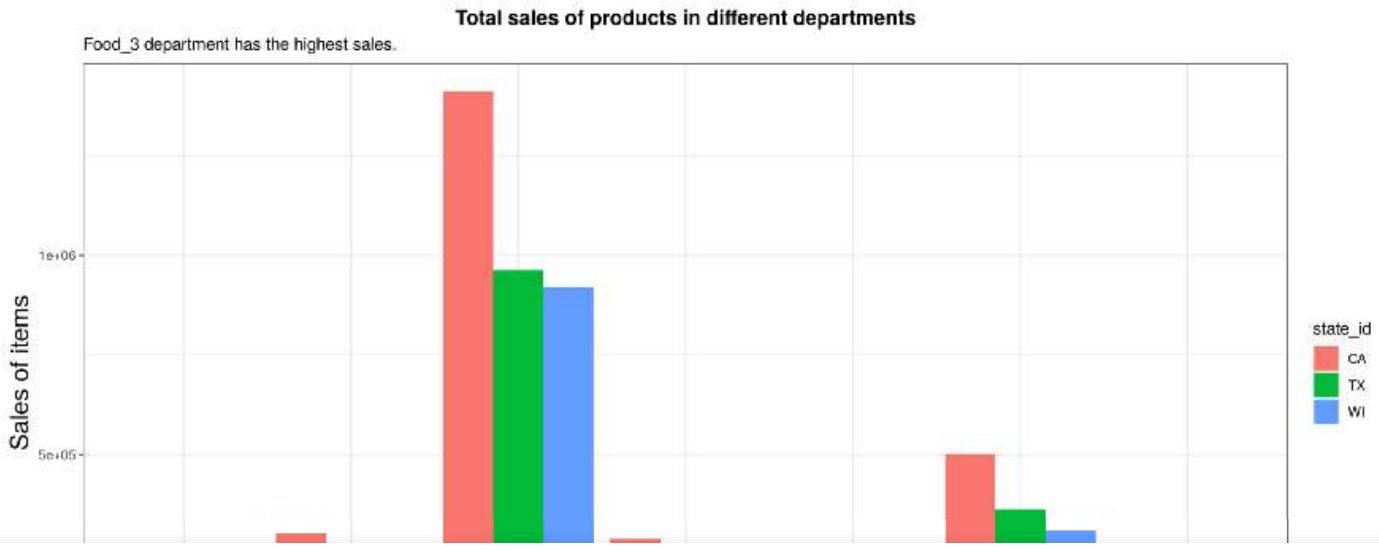


Fig 6. Total sales of all products

Here, the Food 3 department with the lowest mean price had the highest sales. One more interesting thing to note here is despite, Hobbies 1 having the highest mean price and almost double when compared with Hobbies 2, the sales are high for Hobbies 1. HouseHold 1 sales are high. This may indicate that this product department holds the everyday essential items like soaps and detergents.

As observed earlier California state is having more sales followed by Texas and Wisconsin. The expectation being the Food 1and Food 2 categories where Wisconsin sales are higher when compared with texas. So, it can be assumed Wisconsin state population had a liking towards Food 1 and Food 2 departments.

We were able to prove a few thesis statements related to product price, location, and product category. Now, let's jump into Time series analysis to see how different weekdays, months, and events are affecting sales.

## Time series analysis

This section aims to answer:

1. The daily seasonality trend of total sales
2. Which month had the highest and lowest sales?
3. Which weekday do people prefer to grocery shopping in different states?

The time series for all years is plotted to observe the seasonality trend for all 3 states for different departments.

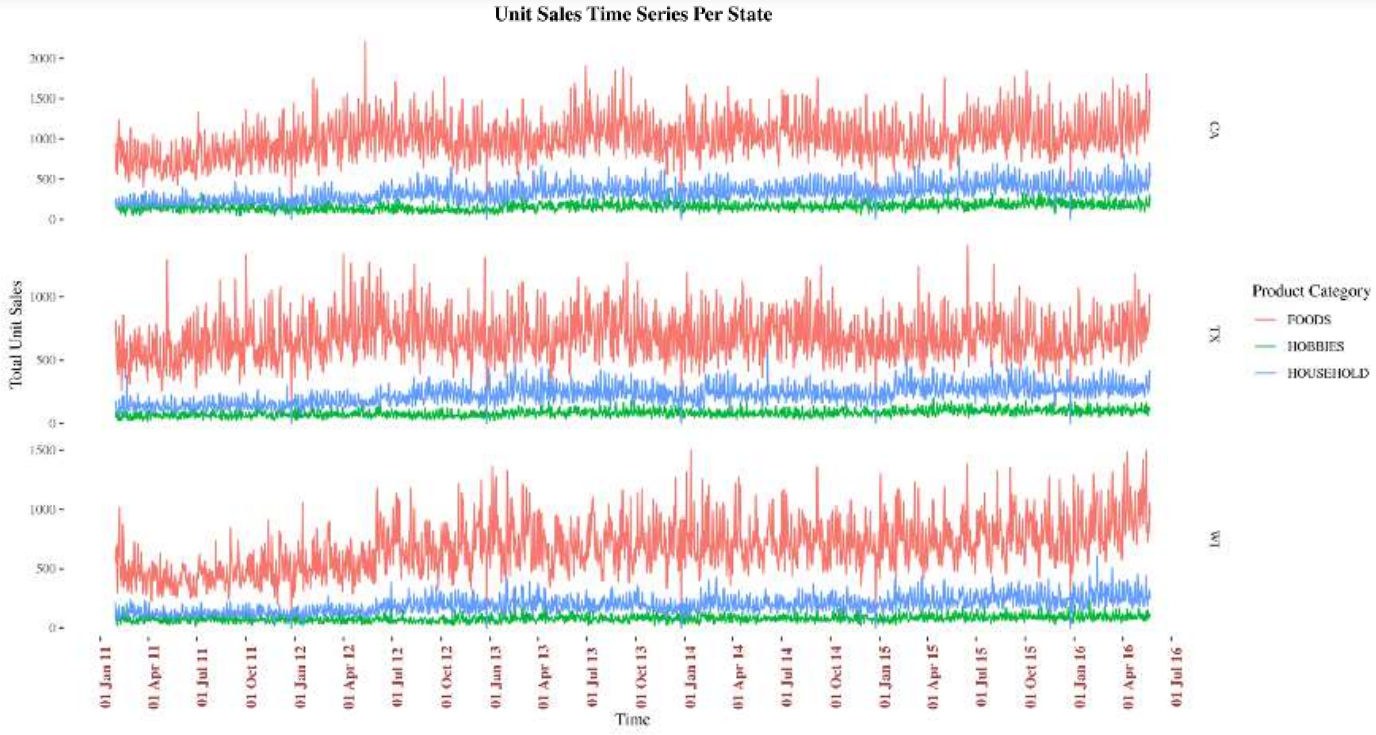



Fig 7. Daily sales trend for departments

The seasonality trend follows the same pattern and is parallel for all 3 states. The highest Food being the department with the highest sales followed by Hobbies and Household. To better understand daily trends a heat map was plotted for the year 2015.

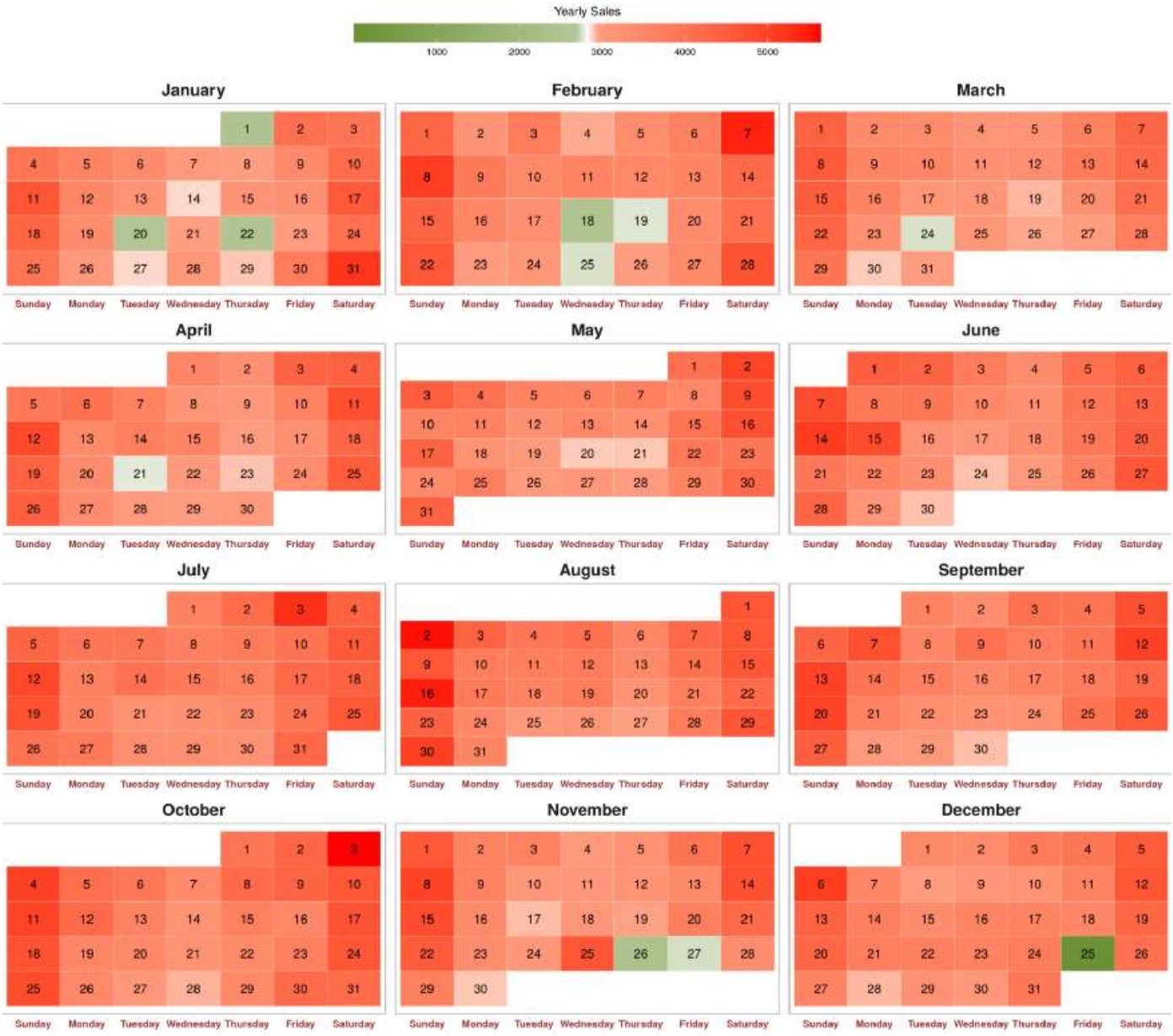



Fig 8. Calendar Heat Map for the year 2015

It appears that Walmart is closed on Christmas. It can be seen that sales are very less on some days like New year and Thanksgiving days. This is due to reduced working hours on Festival days. Also, sales are relatively high on weekends in comparison to normal days.

## Monthly Sales Trend

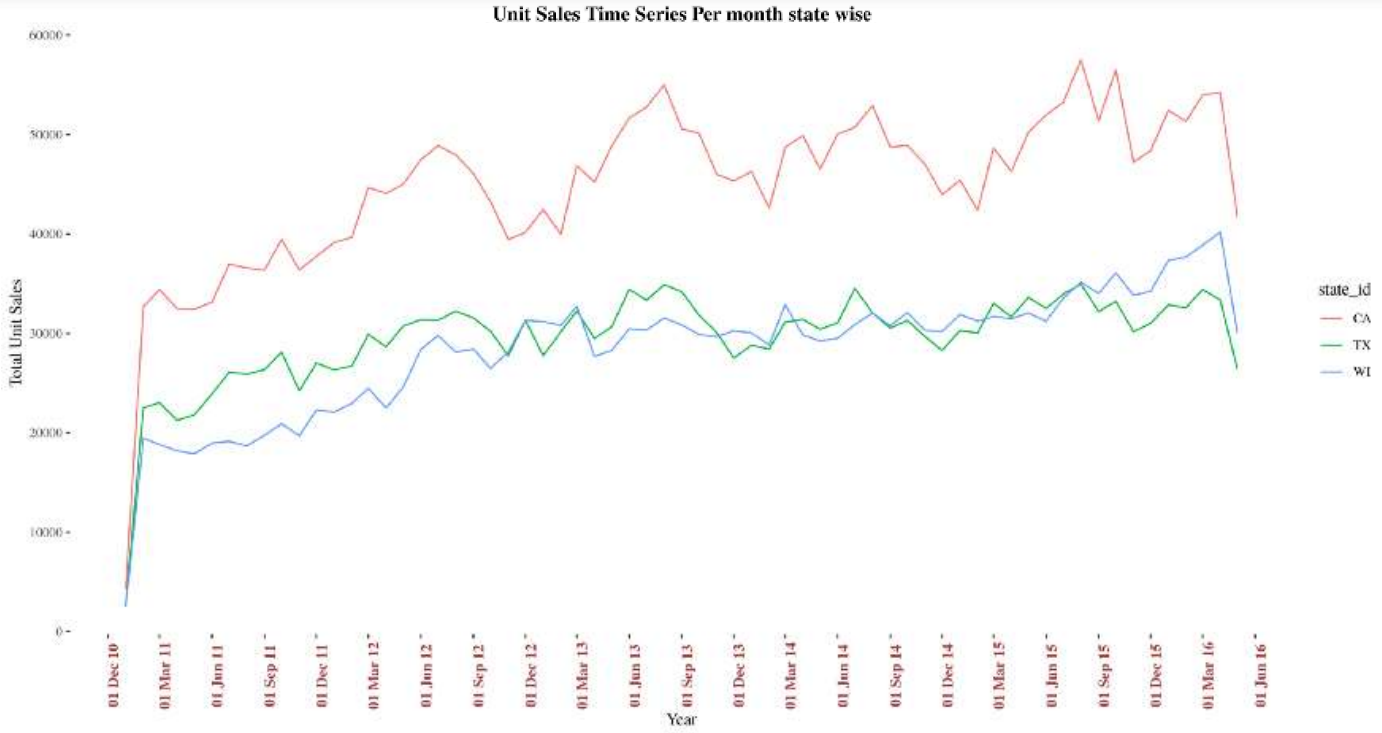



Fig 9. Statewise Monthly sales trend

It is surprising to see the same trend across all 3 states for 5 years. It can be seen that total sales are increasing every year. This trend is due to the introduction of new products every year at Walmart. Also, the trend pattern for increase or decrease is almost similar for every year. To better understand the monthly trends all monthly sales are grouped for a year and the graph is plotted.

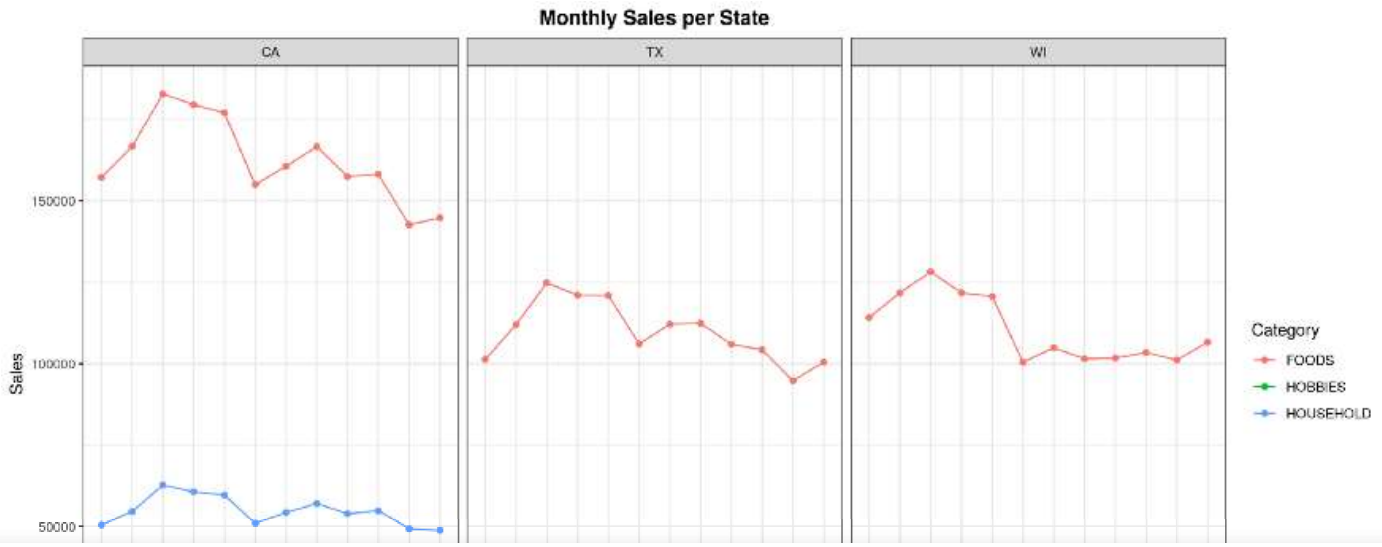


Fig 10. Category wise Monthly sales trend

It can be observed that the sales were increasing every year and are at a peak in March. After March, there is a decrease in sales till May and plummeted in June recording the lowest sales every year. After, June there is a gradual increase in sales for two months, before dropping further until November.

### Weekly Trend

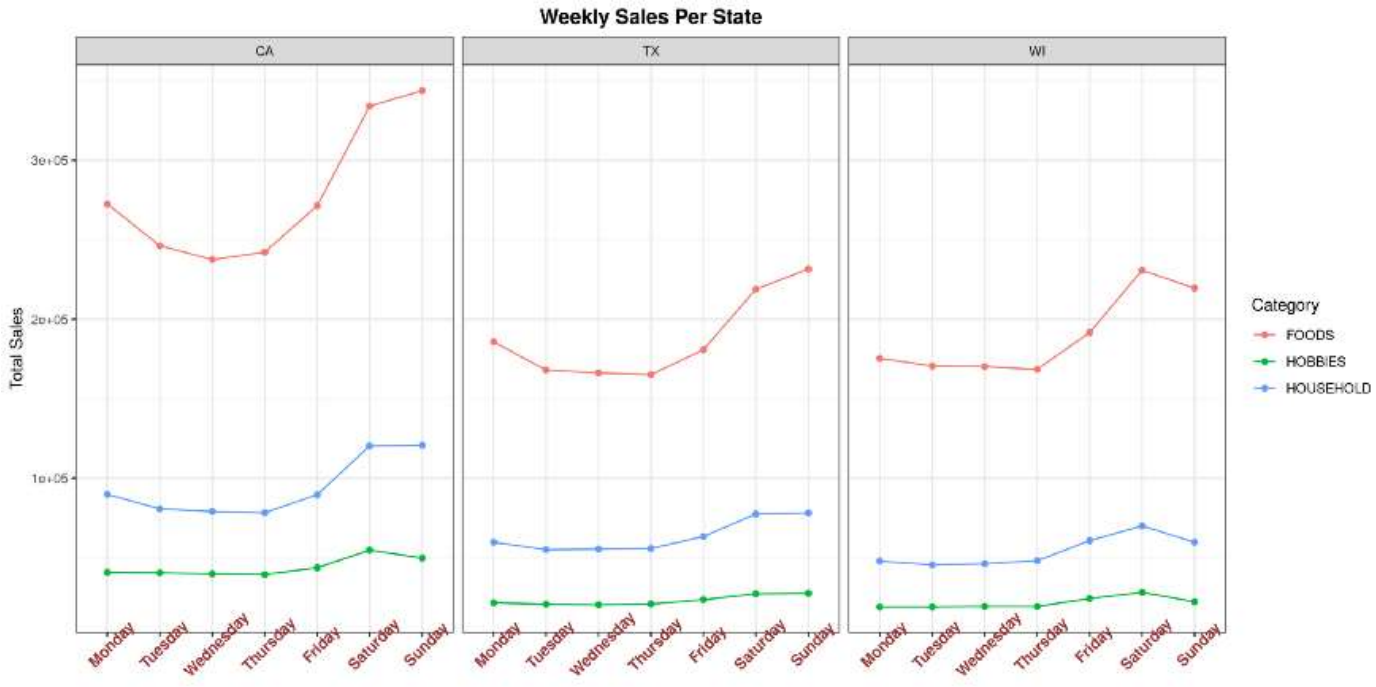



Fig 11. Statewise weekly sales trend

As expected the total sales are more during Saturday and Sunday when compared to normal weekdays. Even here, the Wisconsin state is an exception where peak sales at observed on Saturday, whereas it is Sunday for Calfornia and Texas state. So, maybe the Wisconsin state population prefers to do grocery shopping on Saturday.

To better observe trends for weekdays and month a heat map with total states for weekday vs month is plotted.

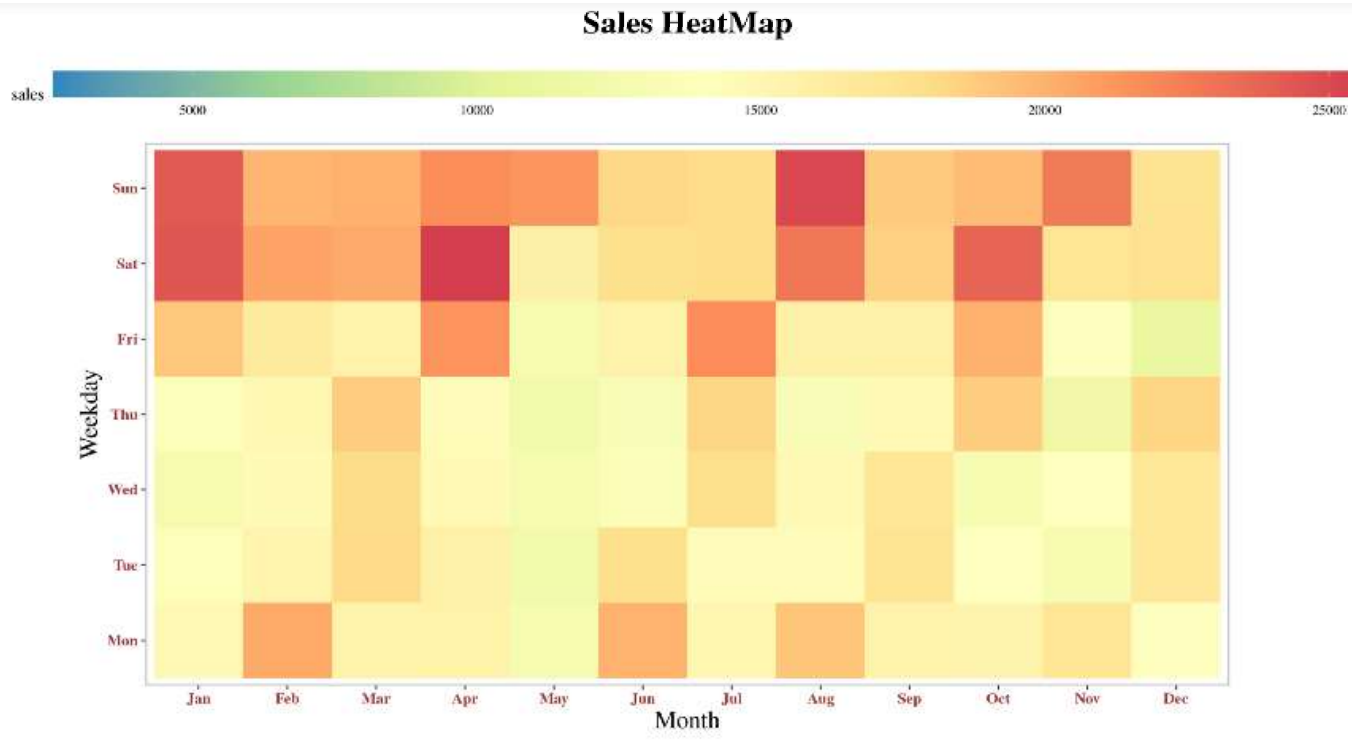



Fig 12. Weekday Vs Month sales heatmap

There is an interesting pattern in this sales heatmap. It can be observed that there is a shift in more number of sales recorded every month. i.e., if the highest sales are recorded on Monday in February, we can see that in March there are more sales on Tuesday.

**Sales trend on Holiday and Special Events:**

This section aims to answer:

1. How festival events and holidays are affecting sales trend?
2. Which holiday recorded the highest sales?

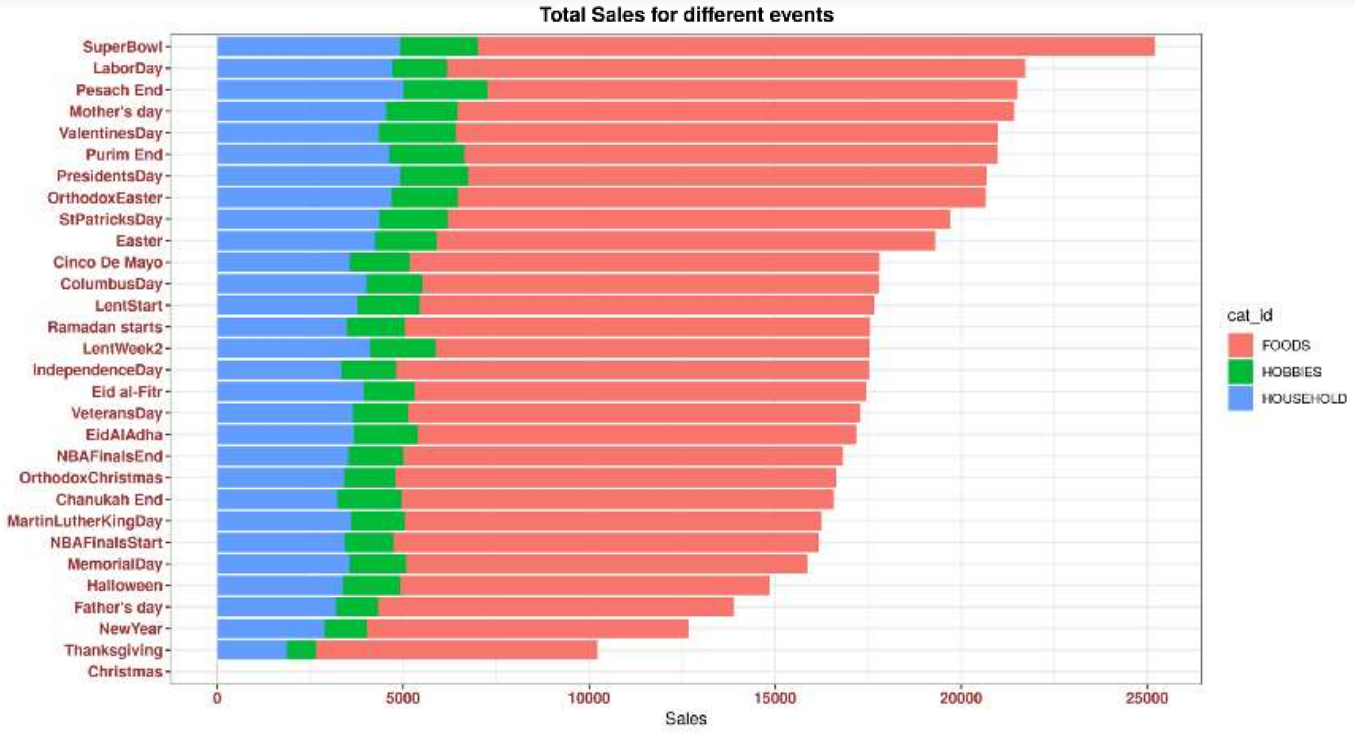
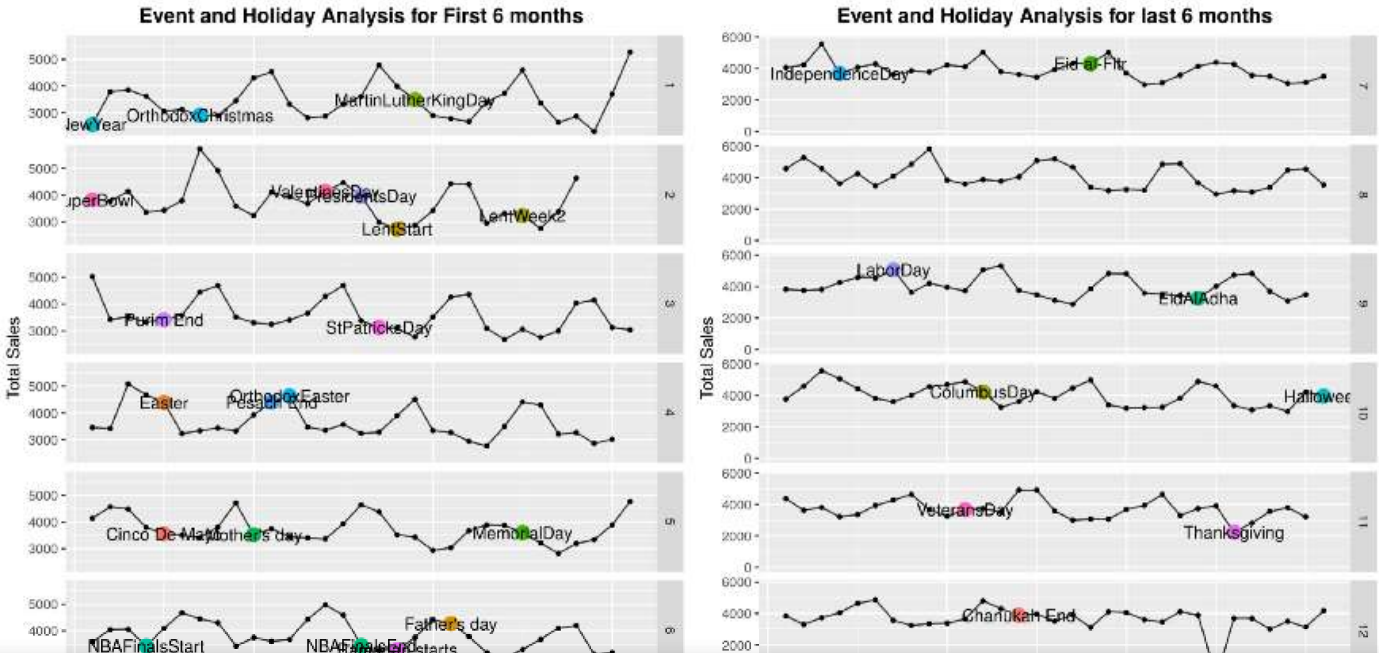



Fig 13. Total Sales on special events

The sales were highest on SuperBowl sporting events. On the day of the National holidays, sales were low. And sales were consistent on the day of the religious festivals. To observe the sales trend, the seasonality trend was plotted for the year 2015.



Observations:

1. On Festival like New year and Easter, the sales were low because of reduced hours.
2. The sales are 0 for Christmas maybe because Walmart is closed.
3. The Sporting events like NBA Final shows an interesting insight, the sales were high the day before the event, and sales dropped on event days.
4. The sales are dropped on special days like father day, mother's day.
5. National holidays and Religious holidays also tend to have a similar effect to sporting events.

## Modeling:

This competition aims to predict sales for 28 days.

Steps Involved:

1. **Introducing Lag and roll mean**
2. **Train/Test Split**
3. **Numerical encoding**
4. **Converting the data into the required format(**data.matrix, LGB.Dataset, Cat.loadpool**)**
5. **Parameter selection**
6. **Model Training/Cross-Validation**
7. **Prediction**
8. **SMAPE Error for Comparison of Models**
9. **One-Standard error rule**


because it is stratifying the entire population before applying random sampling methods. In short, it ensures each subgroup within the population receives proper representation within the sample.

Since the machine learning model, we are using is time-series having lag days and roll mean helps to improve the model so new lag variables are introduced for time-varifying effect variables 1 week,2 weeks,1 month, 2 months respectively. With better computation power 1 year also lag can be introduced since yearly sales patterns are similar. The rolling mean and rolling standard deviation were introduced for 1 week and 1-month lag variables.

For people who don't know about lag, A **lag** is a fixed amount of passing time; One set of observations in a time series is plotted (lagged) against a second, later set of data. The kth lag is the period that happened "k" time points before time

**Train/Test Split**

Since we need to forecast for 28 days, with 5 years of data. All the data with dates less than or equal to March 27th,2016 is considered as training data. And the 28 days data with dates greater than March 27th,2016, and less than April 24th, 2016 is taken as test data. The last 28 days are kept for validation.

**Numerical Encoding:**

As many machine learning models can't read character type data, all the columns should be converted into the numeric format. I used a simple command in R

```
data%>%mutate_if(is.factor, as.integer)
```

**Why Ensembling Models?**

Ensemble methods help improve machine learning results by combining multiple models.


**Why sMape?**

The sMAPE error rate is used because it is a prescribed evaluation metric in the M3 forecasting. The sMape error rate or symmetrical mean absolute percent error is listed as one of the significant, but uncommon forecast error measurements. However, its complexity in calculation and difficulty in explanation makes it a distant third to the fat more common MAD and MAPE forecast error calculations.

The sMape error is calculated as follows

$$SMAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{F_t + A_t}$$

For anyone who wants to learn more about the models used and the advantages of one model over others here is a link to a great article comparing Xgboost vs catboost vs Lightgbm.

**Xgboost:**

The Xgboost requires data in xgb.DMatrix format for prediction so both train and test sets are converted to xgb.Dmatrix matrix using the following command.

```
train_set_xgb = xgb.DMatrix(data = data.matrix(train_data[,features]),
label = data.matrix(train_labels))
test_set_xgb = xgb.DMatrix(data = data.matrix(test_data[,features]),
label = data.matrix(test_labels))
```

The parameters chosen are for Xgboost are as follows

```
params <- list(booster = "gbtree",
               tree_method='gpu_hist', gpu_id=0,task_type = "GPU",
          objective = "reg:tweedie", eta=0.4, gamma=0 nrounds = 20,
           nthreads = 10,early_stopping_round = 10)
```

best model for non-negative data with lots of zero's. So the **Tweedie** objective is used for training the model.

Since the latest Xgboost version supports using GPU the model is trained using GPU. The RMSE evaluation metric was chosen for training the model. An early stopping round of 10 is given so if the model RMSE didn't improve for 10 iterations model will stop. And the best RMSE value will be returned.

The best RMSE value was **2.48**

The **3 fold cross-validation** was done to check model consistency. The best RMSE value returned for cross-validation was **2.637.**

A function was defined to calculate sMAPE value as follows:

```
smape_cal <- function(outsample, forecasts){
  outsample <- as.numeric(outsample)
  forecasts<-as.numeric(forecasts)
  smape <- (abs(outsample-
forecasts))/((abs(outsample)+abs(forecasts))/2)
  return(smape)
}
```

The value of SMAPE for Xgboost is **1.897968**

**Catboost:**

The catboost requires data in load_pool format for prediction so both train and test sets are converted to load_pool format using the following command.

```
train_cat <- catboost.load_pool(data =
data.matrix(train_data_cat[,features]), label =
data.matrix(train_labels))
test_cat <- catboost.load_pool(data =
data.matrix(test_data_cat[,features]), label =
```

```
params_cat <- list(iterations = 1500,
                   metric_period = 100,
                  tree_method='gpu_hist',task_type = "GPU",
                  loss_function = "RMSE",
                  eval_metric = "RMSE",
                  random_strength = 0.5,
                  depth = 7,
                  early_stopping_rounds = 100,
                  learning_rate = 0.18,
                  l2_leaf_reg = 0.1,
                  random_seed = 93)
```

The RMSE is used as a loss function and as an evaluation metric for training the model. The computation time with Kaggle GPU was around 5 mins for 1500 iterations. The early round is given as 100 rounds.

The best RMSE value was **2.36541**

The **3 fold cross-validation** was performed to check model consistency. The best RMSE value returned for cross-validation was **2.39741.**

The value of sMAPE for Catboost is **1.34523,** which seems to better than xgboost.

#### Lightgbm

The lighgbm requires data in lgb_dataset format for prediction so both train and test sets are converted to LGB.Dataset format using the following command.

```
train_set_lgb <- lgb.Dataset(data=as.matrix(train_data[,features]),
label = as.matrix(train_labels))
test_set_lgb <- lgb.Dataset(data=as.matrix(test_data[,features]),
label =as.matrix(test_labels))
valids=list(train=train_set_lgb,test = test_set_lgb)
```

The parameters used are as follows

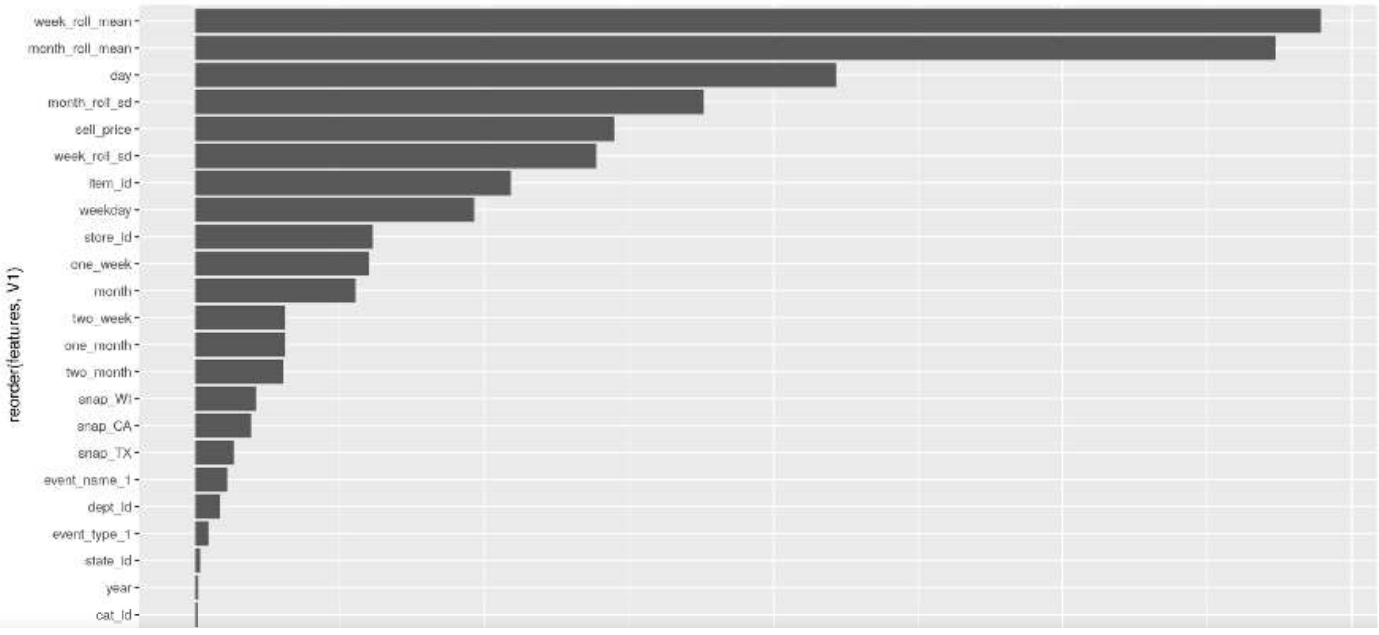

```
            force_row_wise = TRUE,
                    num_leaves=90,
              learning_rate = 0.03,
              feature_fraction= 0.5,
              bagging_fraction= 0.5,
              max_bin=100,
              bagging_freq = 1,
              boost_from_average=FALSE,
            lambda_l1 = 0,
            lambda_l2 = 0,
            nthread = 4)

freeram()
```

The Tweedie is used as an objective function and RMSE as an evaluation metric for training the model.

The best RMSE value was **2.1967701**

The **3 fold cross-validation** was performed to check model consistency. The best RMSE value returned for cross-validation was **2.21.**

The value of sMAPE for the Lgboost model is **1.14,** which is the best out of all 3 models.



Based on the sMape error Lgb is the best model. Just for confirmation one standard error is applied with the cross-validated RMSE values of all 3 models.

**One standard error rule:**

For those who don't know about one standard error rule, one standard error rule is used in cross-validation, in which we take the simplest model whose error is within one standard error of the best model (The model with Least error).

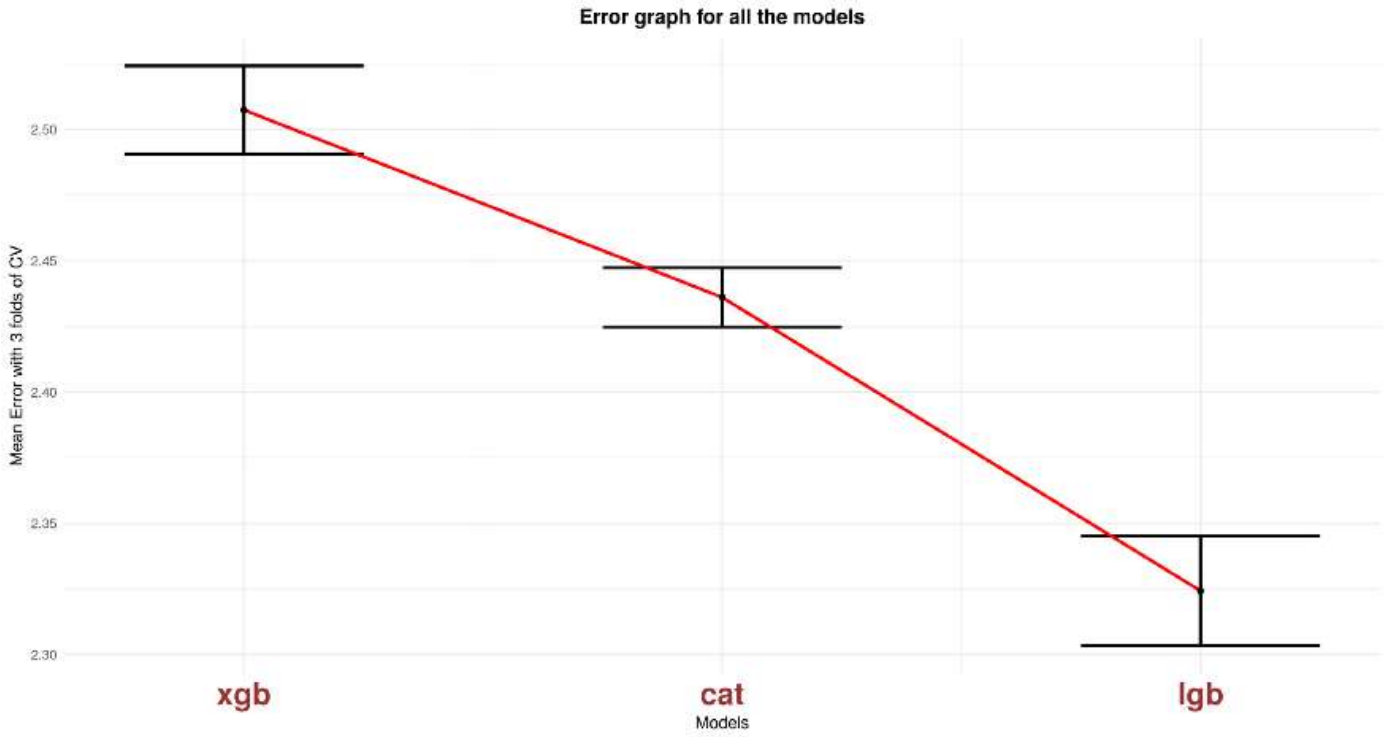



Fig16. one standard error graph

**Conclusion:**

Even here, the LGB is the best performing model with the lowest RMSE value. Since other model mean error points are not in the standard deviation range of the Lihtgbm model. According to one standard error rule, Lightgbm is chosen as the best model.

The submission score with the best performing model is around 0.46.

Kaggle:https://www.kaggle.com/jaswanthhbadvelu/cat-xgb-lgboost-prophet

Dataset: https://www.kaggle.com/c/m5-forecasting-accuracy/data

The sample dataset after stratified sampling looks like this

| dept_id | cat_id | store_id | state_id | day | Unit_Sales | date | wm_yr_wk | ··· | snap_WI | sell_price | one_week |
|---|---|---|---|---|---|---|---|---|---|---|---|
| <fct> | <fct> | <fct> | <fct> | <fct> | <int> | <date> | <int> | ··· | <int> | <dbl> | <int> |
| HOBBIES_1 | HOBBIES | CA_1 | CA | d_724 | 0 | 2013-01-21 | 11252 | ··· | 0 | NA | 0 |
| HOBBIES_1 | HOBBIES | CA_3 | CA | d_1816 | 0 | 2016-01-18 | 11551 | ··· | 0 | 8.26 | 2 |
| HOBBIES_1 | HOBBIES | CA_3 | CA | d_704 | 0 | 2013-01-01 | 11249 | ··· | 0 | NA | 0 |
| HOBBIES_1 | HOBBIES | CA_2 | CA | d_1088 | 1 | 2014-01-20 | 11352 | ··· | 0 | 8.26 | 0 |
| HOBBIES_1 | HOBBIES | WI_3 | WI | d_1805 | 0 | 2016-01-07 | 11549 | ··· | 0 | 8.26 | 0 |
| HOBBIES_1 | HOBBIES | WI_3 | WI | d_1452 | 0 | 2015-01-19 | 11451 | ··· | 0 | 8.26 | 0 |
| HOBBIES_1 | HOBBIES | CA_4 | CA | d_1452 | 0 | 2015-01-19 | 11451 | ··· | 0 | 8.26 | 1 |
| HOBBIES_1 | HOBBIES | TX_1 | TX | d_344 | 0 | 2012-01-07 | 11150 | ··· | 0 | NA | 0 |
| HOBBIES_1 | HOBBIES | CA_3 | CA | d_1088 | 0 | 2014-01-20 | 11352 | ··· | 0 | 8.26 | 0 |
| HOBBIES_1 | HOBBIES | CA_2 | CA | d_724 | 0 | 2013-01-21 | 11252 | ··· | 0 | NA | 0 |
| HOBBIES_1 | HOBBIES | WI_2 | WI | d_1434 | 1 | 2015-01-01 | 11448 | ··· | 0 | 3.97 | NA |