Introduction 简介:

In this blog, the Exploratory Data analysis for M5 competition data is performed using python and sales for 28 days were forecasted using Xgboost, Catboost, Lightgbm, and Facebook prophet. The best model is chosen by comparing the SMAPE error rate and One standard error rule.

在这个博客中，将会使用 python 对沃尔玛的数据进行探索性分析（EDA），然后会使用 Xgboost, Catboost, Lightgbm 和 Facebook prophet 分别对 M5 进行预测，最后会使用 SMAPE error rate 和 一个标准误的置信区间进行选择最佳模型。

Goal 目标:

In March this year(2020), the fifth iteration named M5 competition was held. Competitors have been challenged to **predict sales data** provided by the retail giant Walmart for the next 28 days i.e., till 22nd May 2016, and to make uncertainty estimates for these forecasts**.**

在 2020 年 5 月份，我参加了 Kaggle 上的 M5 预测比赛。在比赛中，参赛选手被要求利用沃尔玛的数据预测未来 28 天的销售数据（数据是 2016 年的，所以我们被要求预测到 2016 年 5 月 22 号），以及估计销售数据的统计分布。本文侧重在预测模型的建立部分，统计分布模型部分看有没有空再做。

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

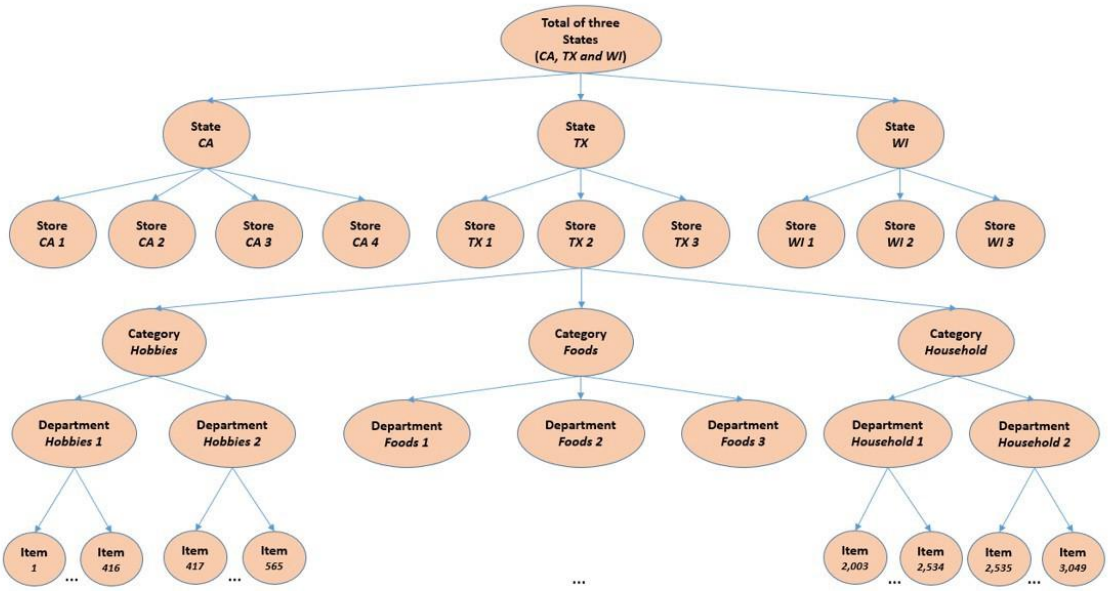
Dataset 数据集:

The data: We are working with **42,840 hierarchical time series**. The data were obtained in the 3 US states of California (CA), Texas (TX), and Wisconsin (WI). “Hierarchical” here means that data can be aggregated on different levels: item level, department level, product category level, and state level. The sales information reaches back from Jan 2011 to April 2016. In addition to the sales numbers, we are also given corresponding data on prices, promotions, and holidays.

The data comprises **3049** individual products from *3 categories* and *7 departments*, sold in *10 stores* in *3 states*. The hierachical aggregation captures the combinations of these factors. For instance, we can create 1 time series for all sales, 3 time series for all sales per state, and so on.

我们总共使用 42840 个时间序列“分层”数据。这些数据来自美国 3 个州：加利福尼亚州（CA）、德克萨斯州（TX）和威斯康星州（WI）。这里的“分层”意味着数据可以在不同的级别上聚合：商品级、部门级、产品类别级和州级。销售信息包含了从 2011 年 1 月至 2016 年 4 月数据。除了销售数据，整体数据集还包含相应的价格、促销和假期数据。

这些数据包括来自 3 个类别和 7 个部门的 3049 个产品，在 3 个州的 10 个商店销售。分层聚合可以很好表现不同因素组合的数据。例如，我们可以为所有销售创建 1 个时间序列，为每个州的所有销售创建 3 个时间序列，依此类推。



Hypothesis 假设:

Some of the factors that may affect sales are:

Weekend: Customers shopping time and spending mostly depends on the weekend. Many customers may like to shop only at weekends.

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.

Product Price: The sales are affected the most by the product price. Most customers will check the price tag before making the final purchase.

Product Category: The type of product greatly affects sales. For instance, products in the household like TV will have fewer sales when compared with sales of food products.

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.

And then, I will use exploratory data analysis to test these hypothesis statements.

可能对销售额会进行影响的因素有：

周末：顾客购物时间和消费主要取决于周末。许多顾客可能只喜欢在周末购物。

特殊事件/假日：根据活动和假日，客户的购买行为可能会发生变化。像复活节这样的节日，食品销售可能会上升，而像超级碗决赛这样的体育赛事，家庭用品的销售可能会上升。

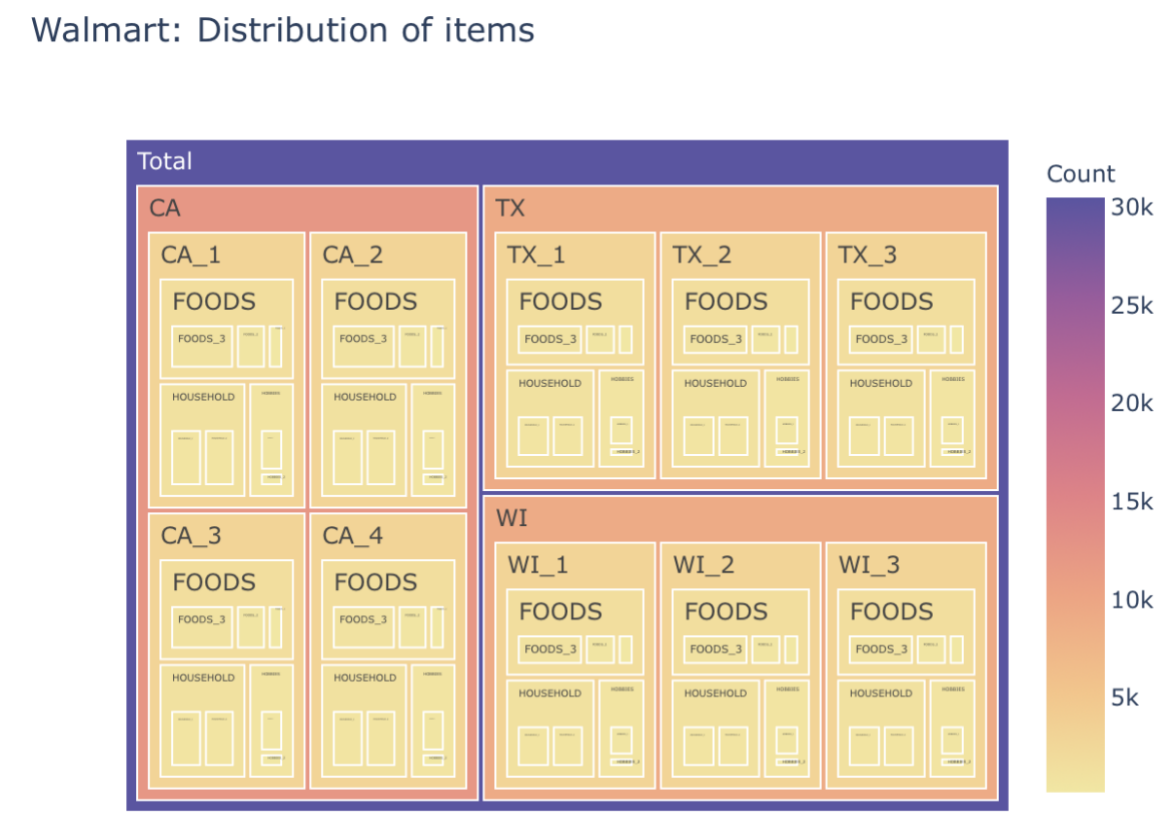
产品价格：产品价格对销售的影响最大。大多数顾客在最后购买前都会查看价格。

产品类别：产品类型对销售影响很大。例如，与食品销售相比，电视等家庭产品的销售额将更少。

位置：位置在销售中也起着重要作用。在加州这样的州，顾客可能会购买他们想要的产品，而不考虑价格，而另一个地区的顾客可能对价格敏感。

在接下来的探索性分析部分，将会用来验证以上假设。

EDA 探索性分析:



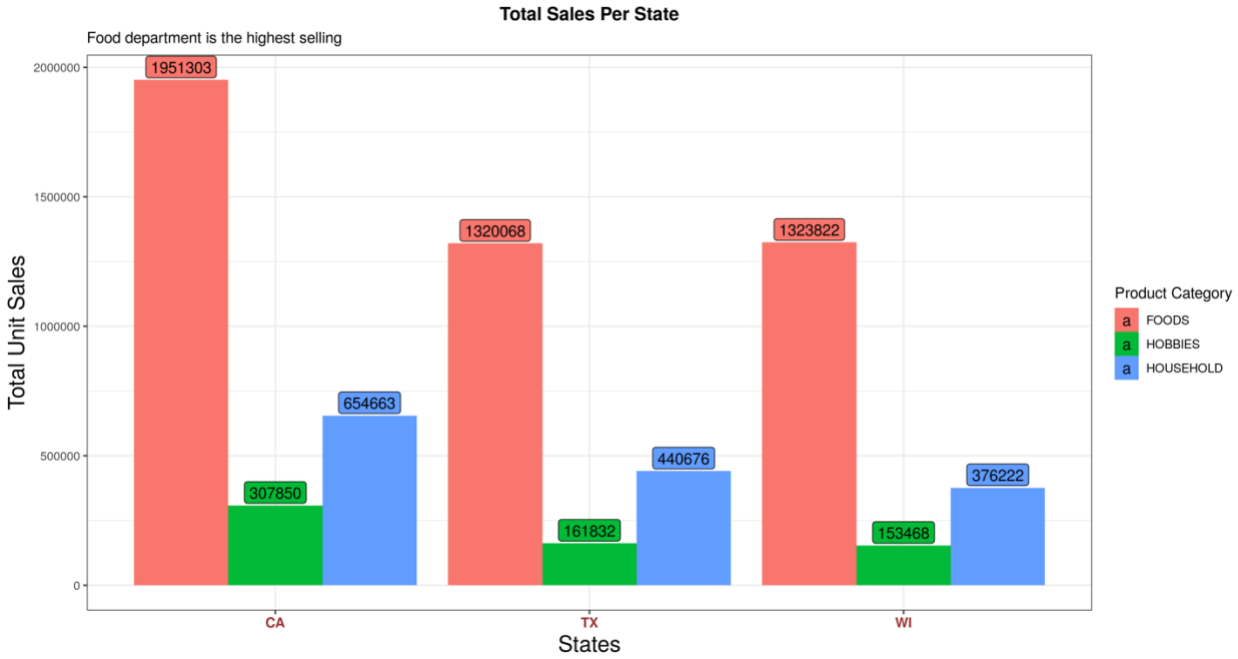
Exploring the location of stores 探索商店地点数据

This section aims to answer:

Which state has the highest sales?

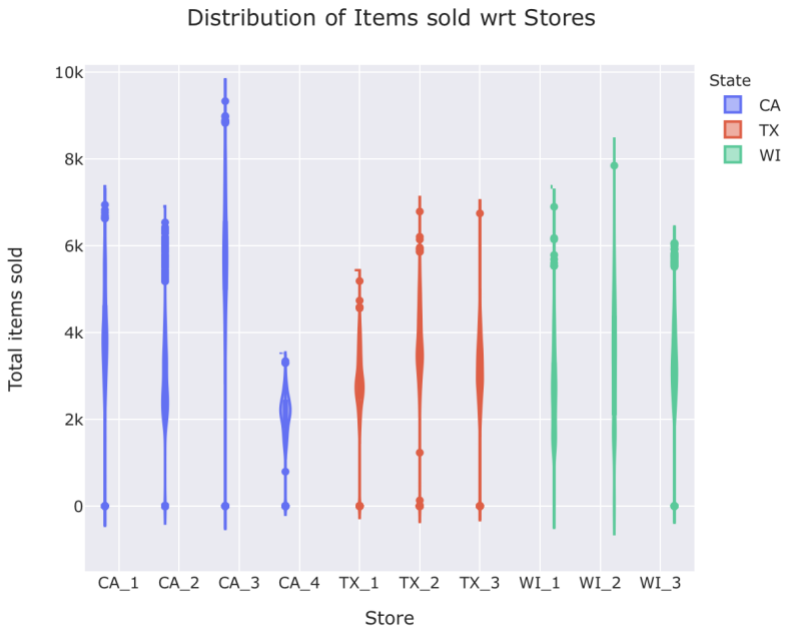
Which department has the highest sales?

The best performing store?



It can be seen from plot that California had the highest sales overall. Having 4 stores and more population might be the reason.

As expected, the Food department recorded the highest sales in all 3 states. Surprisingly, Wisconsin even with low population density when compared to Texas recorded equal sales.



CA_3 in California has sold the most number of items (best performance), while CA_4 has sold the least number of items. The population density and median income also affect these sales.

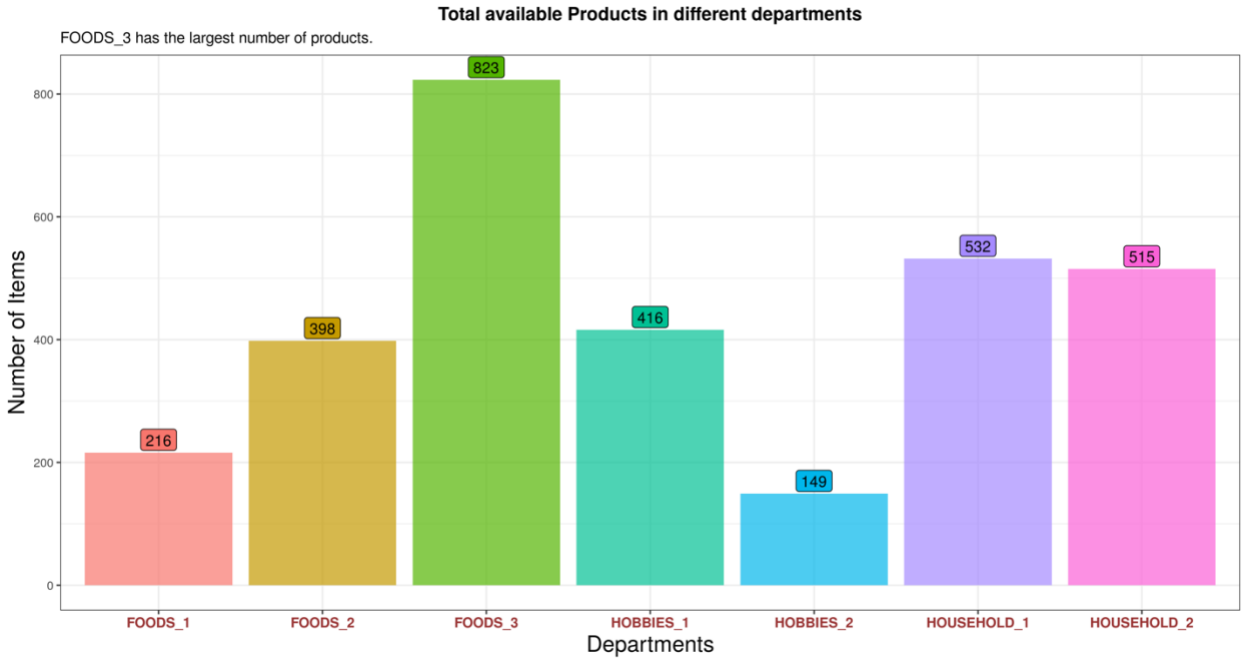
Exploring price & product category 探索价格与产品种类的数据

This section aims to answer:

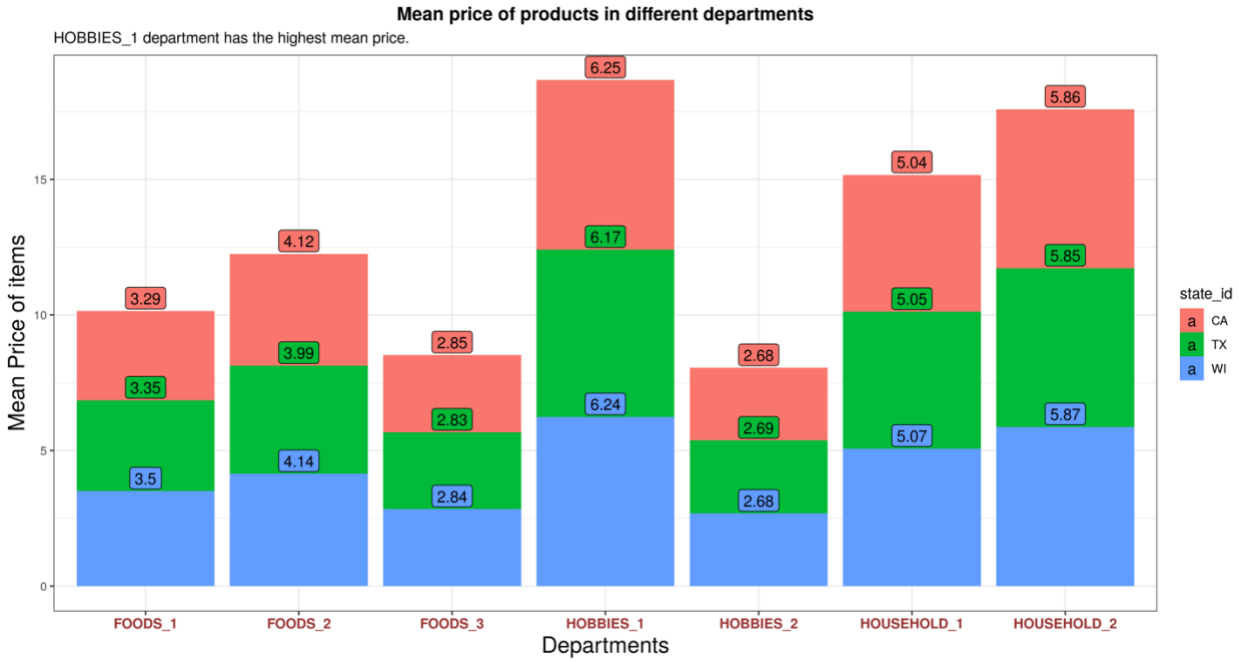
How many different products are available in each department?

What is the mean price of all the available products across different states?

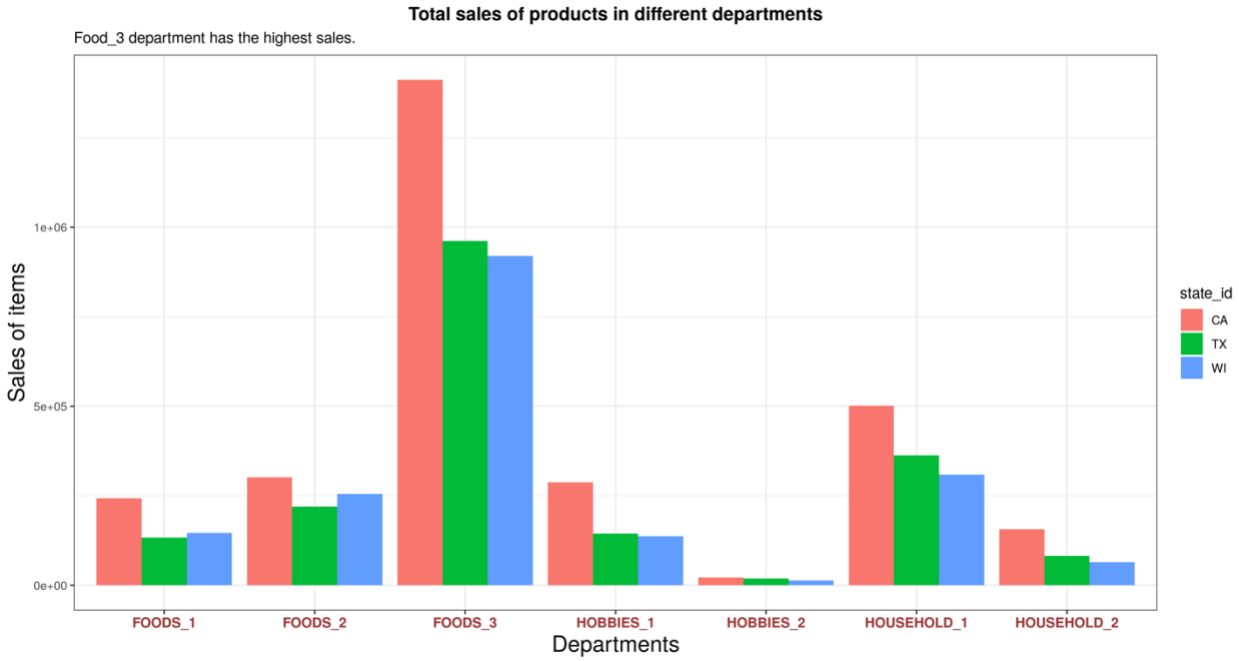
Which department has the highest and least sales?



The number of available products is more in the Food_3 department. So Food 3 department may consist of daily consumed food products like milk etc.



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.



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.

Time Series Analysis 时间序列数据分析

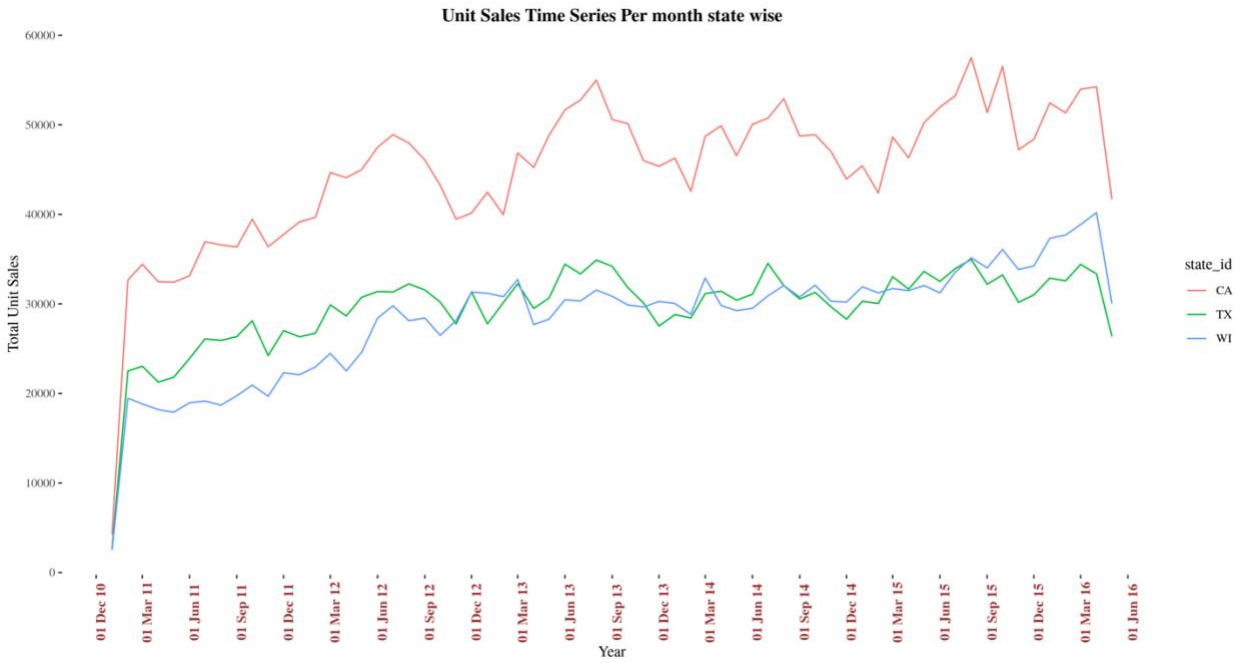
This section aims to answer:

The daily seasonality trend of total sales

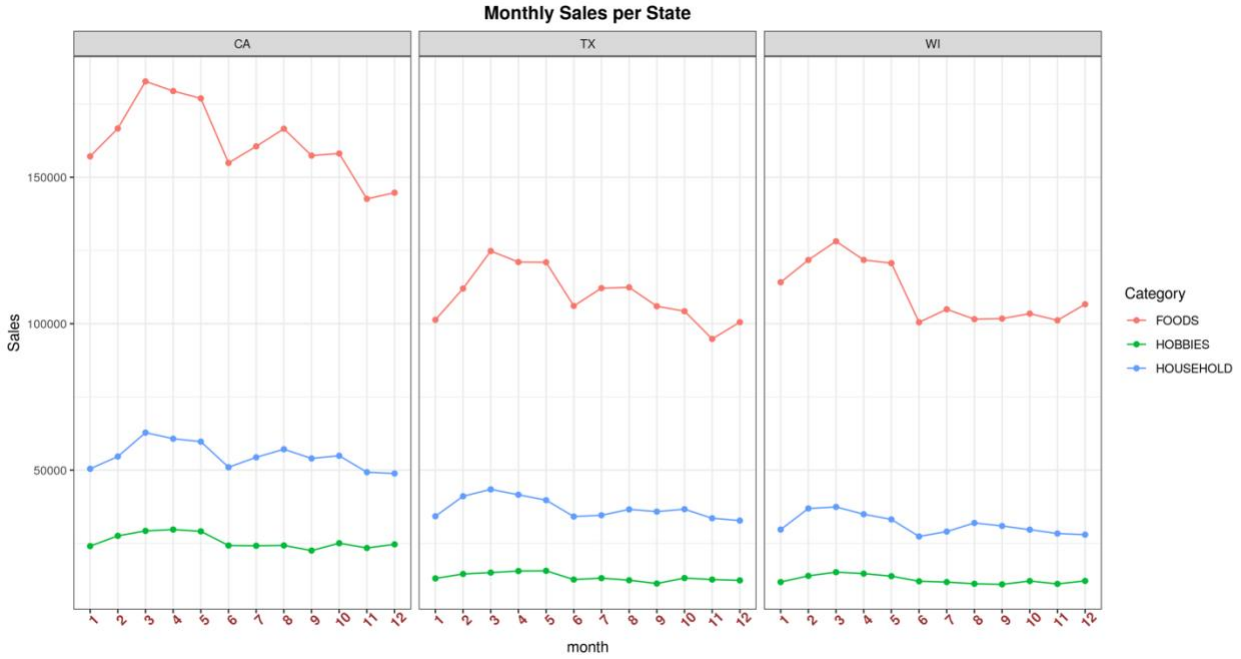
Which month had the highest and lowest sales?

Which weekday do people prefer to grocery shopping in different states?

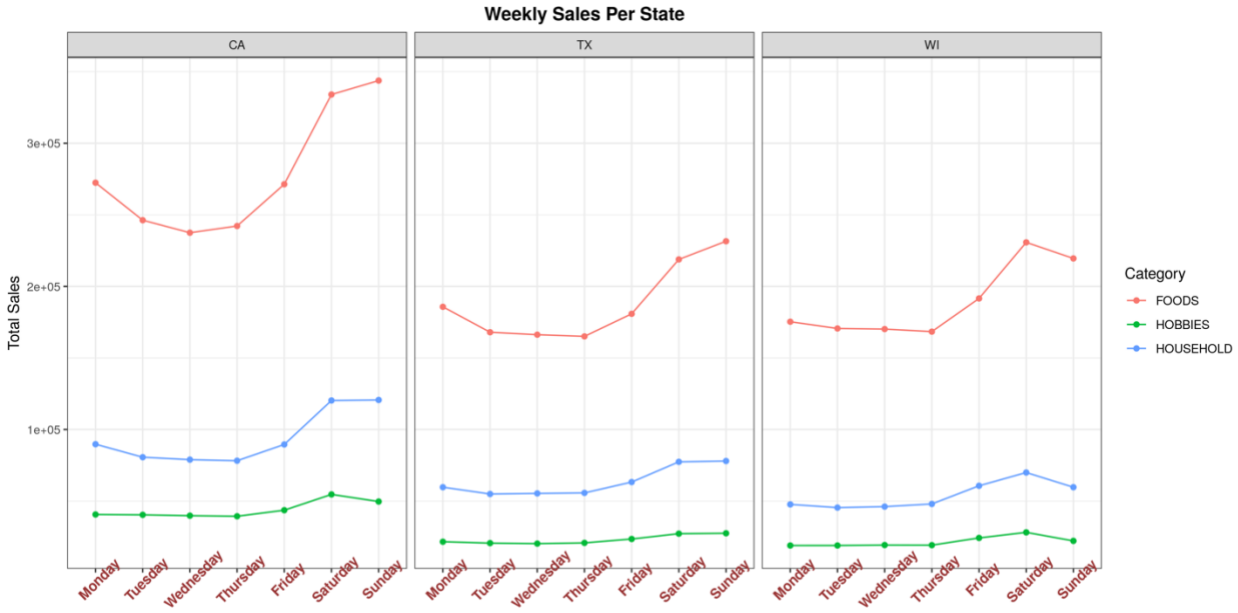
The time series for all years is plotted to observe the seasonality trend for all 3 states for different stores. The seasonality trend follows the same pattern and is parallel for all 3 states.



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 yea

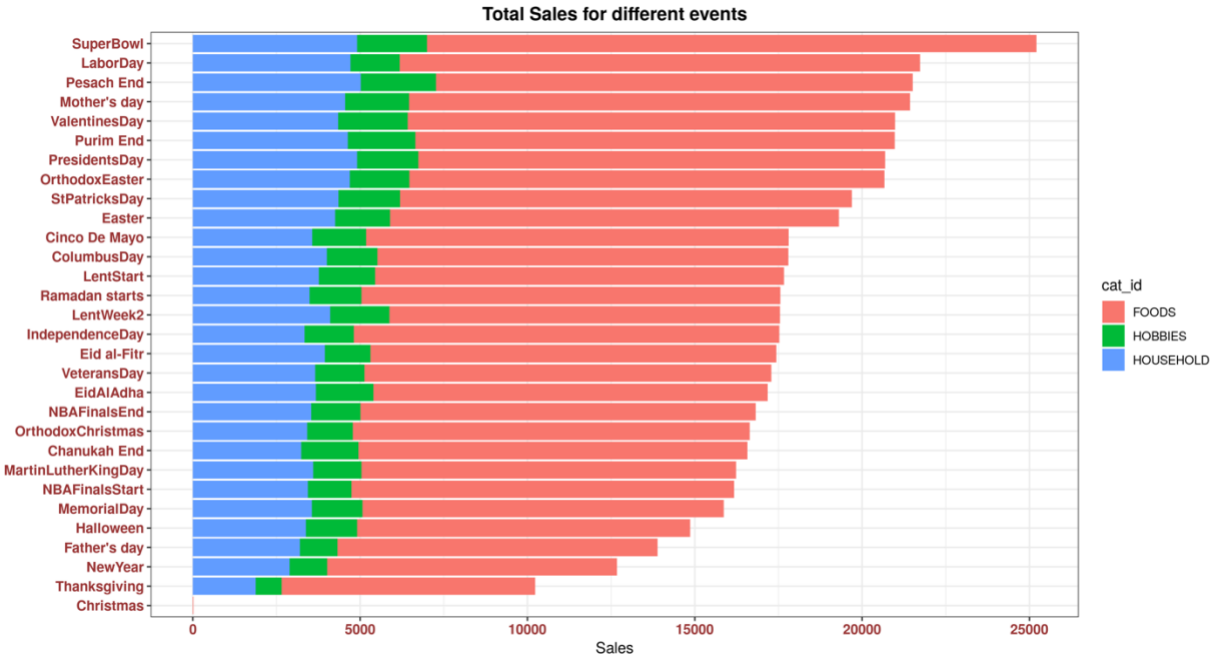


It can be observed that the sales were increasing every year and are at a peak in March. After March, there is a decrease in sales till May and plummeted in June recording the lowest sales every year. After, June there is a gradual increase in sales for two months, before dropping further until November.

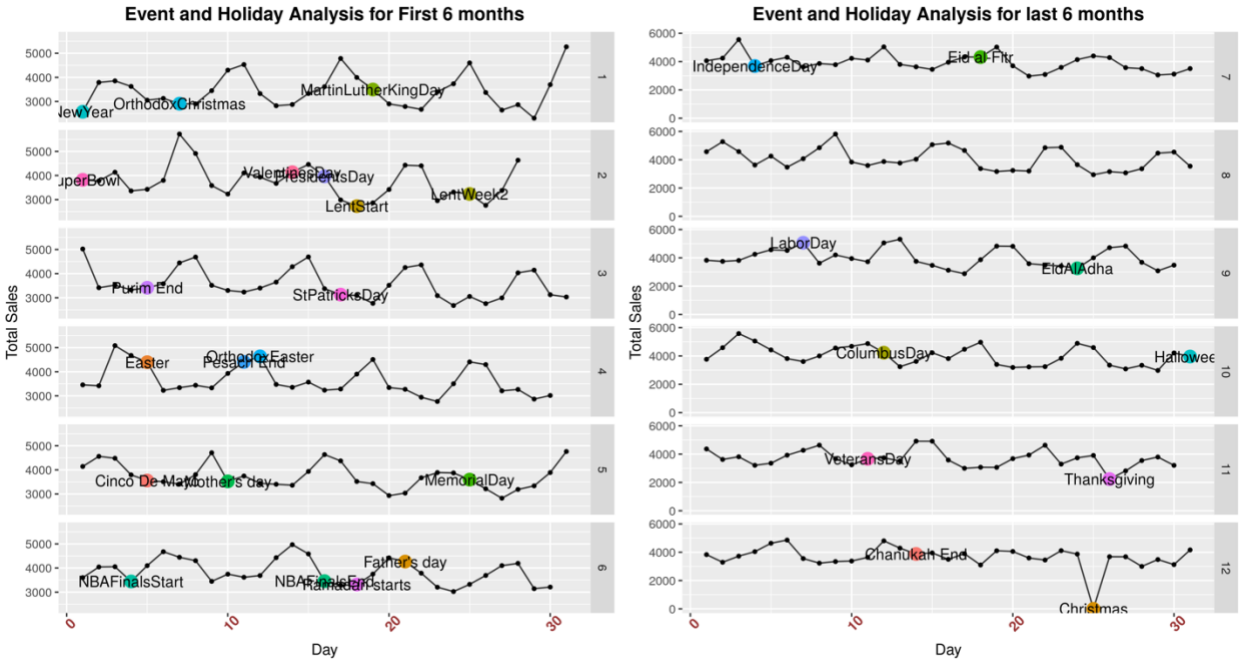


As expected the total sales are more during Saturday and Sunday when compared to normal weekdays. Even here, the Wisconsin state is an exception where peak sales at observed on Saturday, whereas it is Sunday for Calfornia and Texas state.

Sales trend on holiday and special events 在节日和特殊活动期间的销售趋势



The sales were highest on SuperBowl sporting events. On the day of National holidays, sales were low. And sales were consistent on the day of the religious festivals.



On Festival like New year and Easter, the sales were low maybe because of reduced hours.

The sales are 0 for Christmas maybe because Walmart is closed.

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.National holidays and Religious holidays also tend to have a similar effect like sporting events.

Modeling 数学建模:

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.

Why Ensembling Models?

Ensemble methods help improve machine learning results by combining multiple models. Using ensemble methods allows us to produce better predictions compared to a single model. Therefore, the ensemble methods placed first in many prestigious machine learning competitions, so different ensemble are compared using the sMAPE error rate.

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.

Xgboost

Since our data contains a lot of zero values using objective as regression didn't give expected results. After going through some research articles it is found that Tweedie is the 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 RMSSE evaluation metric was chosen for training the model. Early stopping round of 10 is given so if the model RMSE didn't improve for 10 iterations model will stop. And the best RMSSE value will be returned.