****

**Department of Computer Engineering**

**CMPE 256 Summer 2020**

**M5 Forecasting – Accuracy, Estimate the unit sales of Walmart retail goods**

**Project Report**

**Team 2**

**Shraddha Bhangale (014636505)**

**Under the guidance of,**

**Prof. Kaikai Liu**

Table of Contents

**Abstract3**

**Chapter 1 : Introduction4**

**Chapter 2 : Exploratory data analysis5**

2.1 Categories of products 5

2.2 Sales in states7

2.3 Sales Distribution8

2.4 Department wise sales 9

2.5 Store wise sales10

2.6 Categorical sales in states12

2.7 Price distribution across categories13

2.8 Daily sales14

2.9 Seasonal Decomposition15

2.10 Monthly sales across the States16

2.11 Weekly sales across the categories 18

2.12 Average sales across week days 20

2.13 Average sales across months 21

2.14 Events in year22

2.15 Event day sales across months 24

2.16 Snap days over the year26

2.17 Impact of Snap days 27

**Forecasting Models28**

3.1 Exponential Smoothing29

3.2 Holt Winter’s Exponential Smoothing30

3.3 Moving Average31

3.4 SARIMAX32

3.5 Prophet33

3.6 Models comparison34

**Conclusion35**

**References** **36**

**Abstract**

Forecasting is a prediction or estimation of a future situation as accurately as possible. The sales forecasting is the logical start of business planning and is particularly important as it is the foundation upon which all company plans are built in terms of markets and revenue. Therefore, it is necessary to predict upcoming sales, cost and profits. Presenting here is a study of several time series forecasting methods applied to sales data of Walmart. The data is from three US States (California, Texas, and Wisconsin) and includes item level, department, product categories, and store details over a period of approximately 5 and a half years. There are some noticeable increase and decrease in the sale of products due to events. The forecasting models implemented here are exponential smoothing, Holt Winters exponential smoothing, moving average, ARIMA and prophet. A comparative analysis of these algorithms is performed to indicate which algorithm works best.

**Chapter 1**

**Introduction**

A world today where competitive margins are becoming increasingly narrower and actions must be decisive yet informed, the ability to accurately make forecasts is of premier importance.

We cannot predict the future of  business, but we can reduce risk by eliminating the guesswork. With accurate forecasting, we can make a systematic attempt to understand future performance. This will allow you to make better informed decisions and become more resistant to unforeseen financial requirements. Without correctly estimating financial requirements and understanding changing market

A more traditional yet still thoroughly compelling application of forecasting is sales prediction, which is the focus of this work. As markets become more and more global and competition is ruthless, optimizing an organization’s operational efficiency

is of premium importance. When companies must spread their resources broadly and consumers have a surfeit of choices, every advantage a company can squeeze out will make a difference. If a company can match the demand of a product with just the right amount of supply, then there will be no lost sales due to a lack of inventory as well as no costs from overstocking.

Sales forecasting uses patterns gleaned from historical data to predict future sales, allowing for informed courses-of-action such as allocating or diverting existing inventory, or increasing or decreasing future production.

Forecasting using machine learning involves EDA (Exploratory Data Analysis), data processing and feature generation, training of data. We use five different models for prediction and see how the model performs. We use root mean square error as evaluation metrics. The model having lowest possible root mean square error will be the best fit model. Then, to increase the accuracy of models, we focus on tuning the parameters.

Going forward with the report we will cover exploratory data analysis, followed by forecasting models, comparison of models and the conclusion.

**Chapter 2**

**Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is the first step in the process data analysis. In exploratory data analysis, we make sense of the data we have and then figure out what questions we want to ask and how to frame them, as well as how best to manipulate the available data sources to get the answers we need. We do this by taking a broad look at patterns, trends, outliers, unexpected results and so on in our existing data, using visual and quantitative methods to get a sense of the story this data tells. We look for clues that suggest our logical next steps, questions or areas of research. The data for M5 forecasting accuracy is huge and to understand how sale varies depending on department, seasonality, events and other factors following EDA is performed.

**2.1 Categories of products**

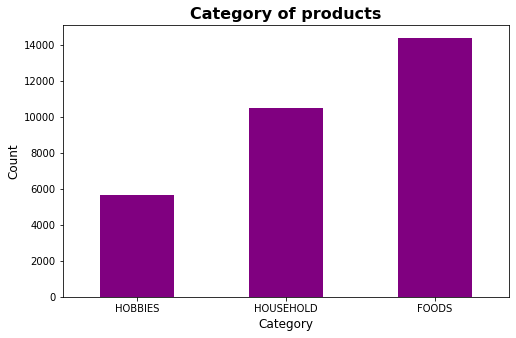
To start the EDA, I started with the exploring the categories of the items we have in our data. 

Figure 1: Categories of product

The Figure 1 shows that there are three categories of products:

1. HOBBIES
2. HOUSEHOLD
3. FOODS

We can also see that the items in FOODS category are the most sold.

These categories are further divided into departments and the above figure shows the combined sales of each categories.

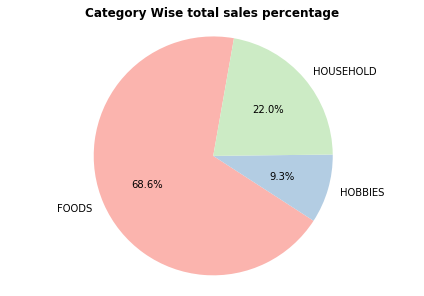


Figure 2: Percentage sales by category

From the above figure, we can see that sale of FOODS category is 68.6% of all over sale, whereas, the sale of HOUSEHOLD and HOBBIES categories is quite less comparatively.

**2.2 Sales in states**

From Figure 3, we can see the sales in California (CA) are the highest. Sales in TX and WI are almost the same.

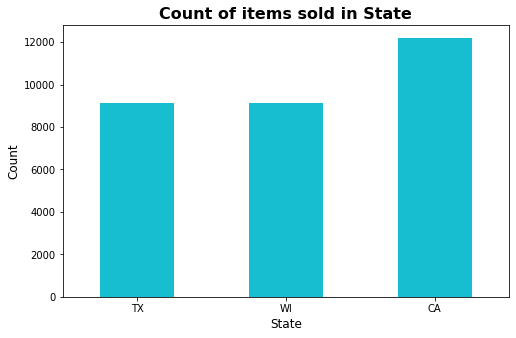
****

Figure 3: Sales in States

We will see clearer view of sales across the states in upcoming section.

**2.3 Sales Distribution**

We can have deeper look on the sales distribution from the figure below.

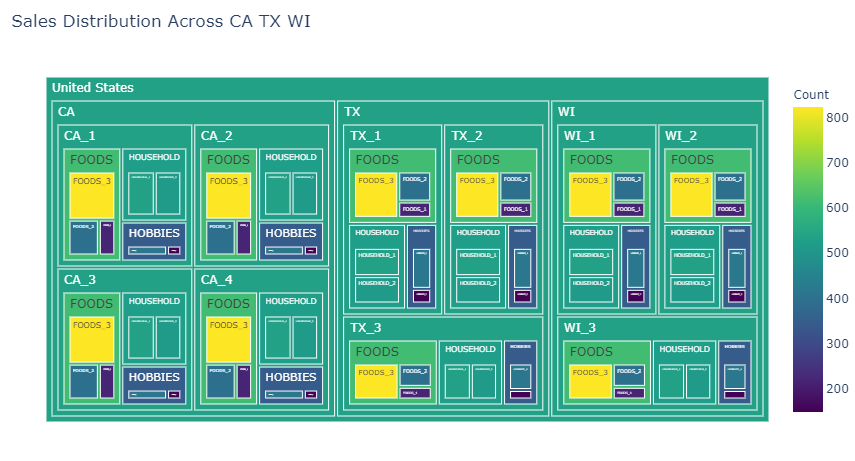
****

Figure 4: Sales distribution across states

From Figure 4, we can see that sales in California (CA) is more as compared to Texas(TX) and Wisconsin (WI). We see that in all the states the FOODS category is most common category of all. And the items in FOODS\_3 department are the highest selling department across all the three states, followed by HOUSEHOLD and HOBBIES categories.

**2.4 Department wise sales**

As we know that there are three categories of items, these categories are further divided into departments. There are three, two and two sub-categories of FOODS, HOUSEHOLD and HOBBIES respectively.

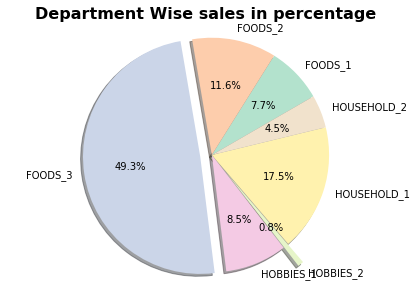


Figure 5: Department wise sales

From Figure 5, we can see that the 49.3% of total sales happened in FOODS\_3 followed by HOUSEHOLD\_1, FOODS\_2, HOBBIES\_1, FOODS\_1, HOUSEHOLD\_2 and HOBBIES\_2. We can see that the sale of HOBBIES\_2 is less than one percent of total sale.

**2.5 Store wise sales**

In this section, we want to see how is the sale across all the stores. From the below table we can see that there are ten stores, four in California, three in both Texas and Wisconsin.



Table 1 : Store Sales

Below we visualize the store wise sales because it is easy to understand this way.

From Figure 6 below, we can see that CA\_3 has the highest number of sales, whereas, CA\_ has the lowest number of sales. We see the both the maximum and minimum sale occurs in California this may be due to several reasons.



Figure 6: Store wise sales

The scale on y-axis is exponential due to the limitation of scale.

**2.6 Categorical sales in states**

In this section, we can see the categorical sales across the three states. We already know that the sales of FOODS category are highest among all.

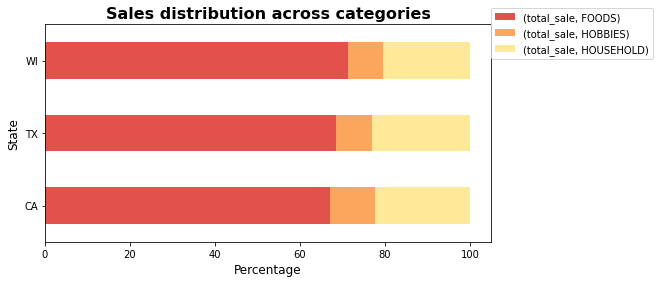


Fig 7: Sales distribution in states

Sales of FOODS can be seen more in Wisconsin (WI), followed by Texas (TX) and California (CA).

We can see that the sales of HOBBIES are the highest in California (CA) as compared to other two states.

The sales in HOUSEHOLD category are far better in comparison to HOBBIES category. HOUSEHOLD items are sold mostly in Texas (TX) followed by and California (CA) and Wisconsin (WI).

**2.7 Price distribution across categories**

In this section, we see how the price varies across the categories of items sold at the stores. Here we plot the graph of density of price versus the cost of items.

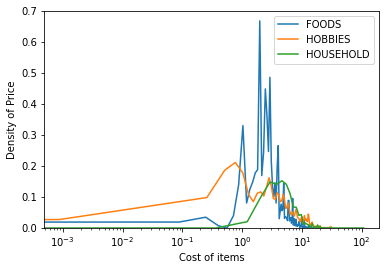


Figure 8: Density vs Cost

In the Figure, we see that most of the items have price range from $0 to $10. The scale on x-axis is logarithmic due to the limitation.

Maximum and minimum prices are 107.32 and 0.01 respectively.

**2.8 Daily sales**

In this section, we can see the sales trend over the years.

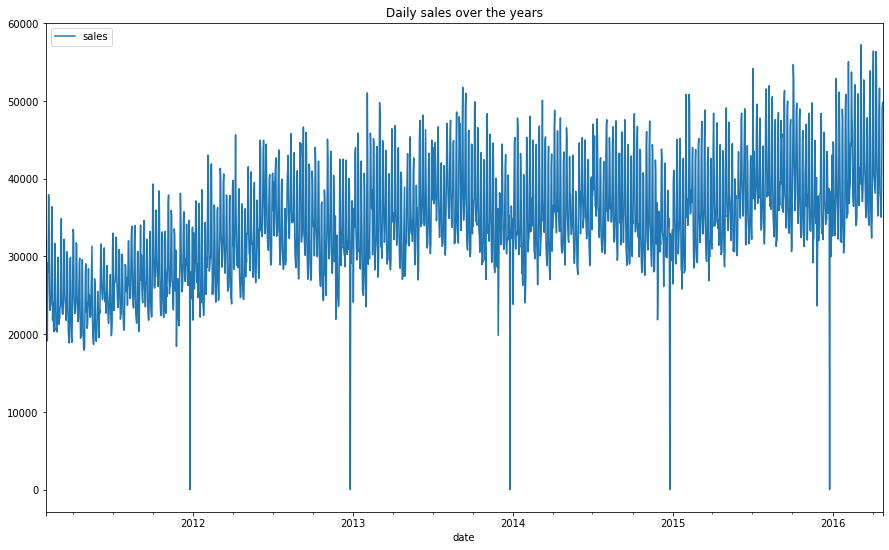


Figure 9: Daily sales over the years

We can see that the overall sales are increasing and the sales goes to zero on the Christmas every year.

**2.9 Seasonal Decomposition**

A seasonal pattern occurs when a time series is affected by seasonal factors such as the time of the year or the day of the week. In other words, seasonality refers to fluctuations in the sales revenue that are caused by external factors and occur on a predictable schedule around the same time(s) every year. We can’t control seasonal shifts in consumer behavior, especially when those shifts are permanently tied to the holiday schedule or growing seasons.  
  
It is necessary to figure out the root causes of seasonality so that business can capitalize on their strong cycles and survive their slow periods*.*

Let’s see the overall seasonality in our data.

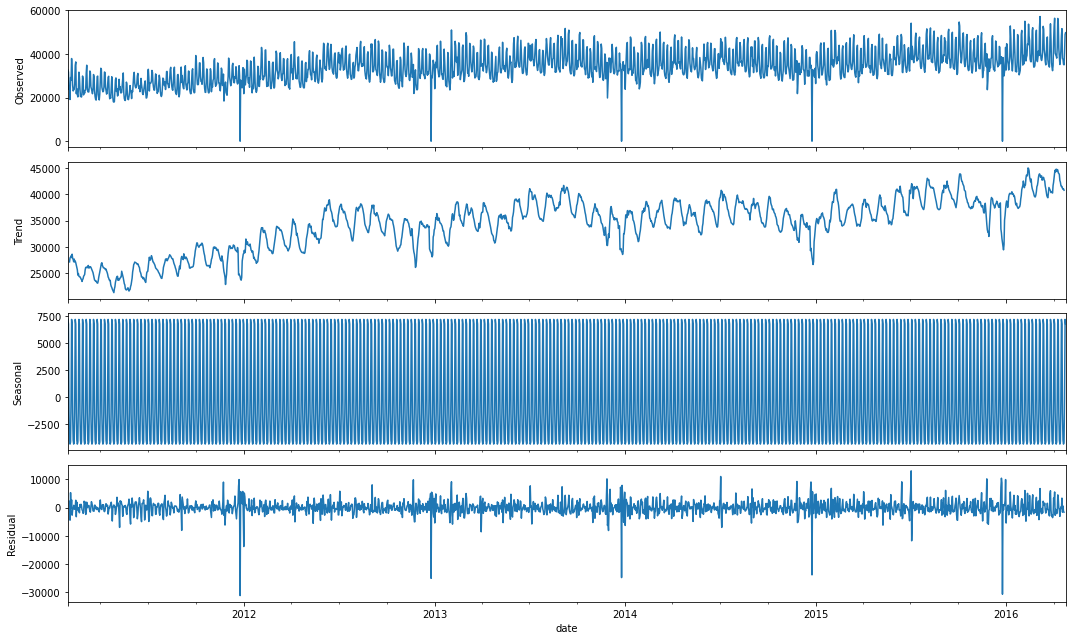


Figure 10: Seasonality of Sales

There are four graphs, which gives below information:

Observed: We see our actual time series (daily sales) across the years

Trend: Our trend seems to be increasing over the years

Seasonality: The graph shows a seasonal aspect. We can see strong weekly seasonality.

Residual: The graph shows are random variations in the series after the trend and seasonal data is removed. It helps in understanding the noise.

**2.10 Monthly sales across the States**

In this section we see the monthly sales across the states.

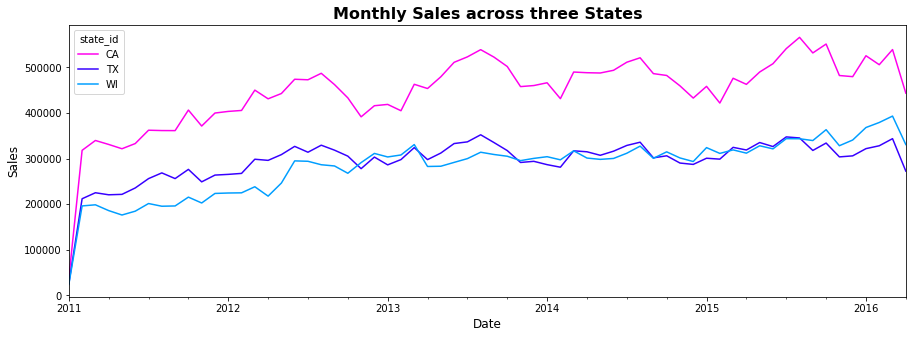
****

Figure 11: Monthly sales across states

From Figure 11 and Figure 12 (below), we can see all the states show different sales over the years. From the graph, we observe that California (CA) shows the increasing trend over the years.

Texas (TX) shows increasing trend until the third quarter of 2013 and then show decrease in sales from then up till first quarter of 2014. Later the sales seems to almost stagnant.

Wisconsin (WI) shows the increasing trend till 2013 and seems stagnant between 2013 to 2015. Later the trend increases.

All the states show different seasonality.

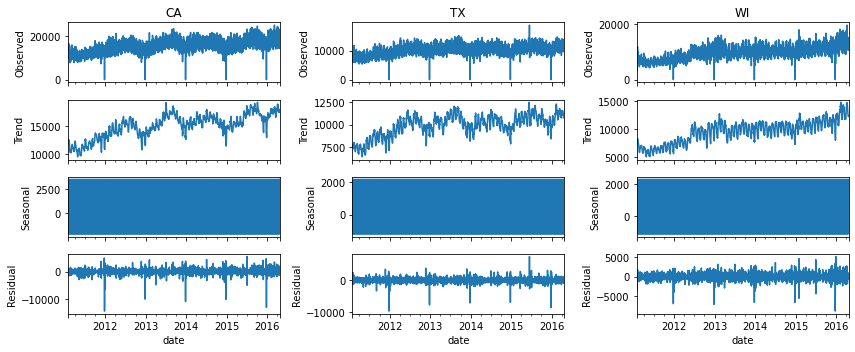


Fig 12: Time series decomposition across states

**2.11 Weekly sales across the categories**

In this section we will see the weekly trend of sale for all categories

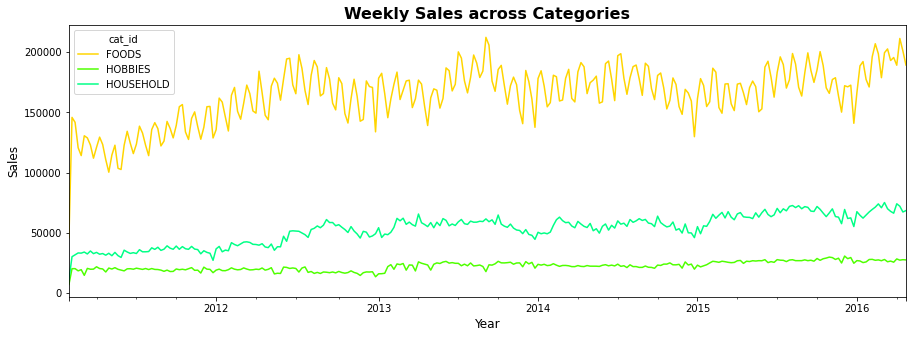


Figure 13: Sales across categories

From figure 13, we can see increase in the sales of FOODS category over the years. Sales for HOBBIES over the years are pretty much flat and show very less seasonality. . The HOUSEHOLD categories shows increase of sales over the years.

FOODS show highest seasonality.

Let’s see the time series decomposition.

From Figure 13, we can see that FOOD sales have increased quite much till the mid of 2012 and then decrease rapidly. We see that sales increase mid of year and decrease towards the end of the every year

HOBBIES sales have stagnant till 2012 and spiked in 2013. In 2014 midyear sales starts to decrease and again increases in 2015 Aug.

HOUSEHOLD sales show good increasing trend and clean seasonality in around the month of March and August.

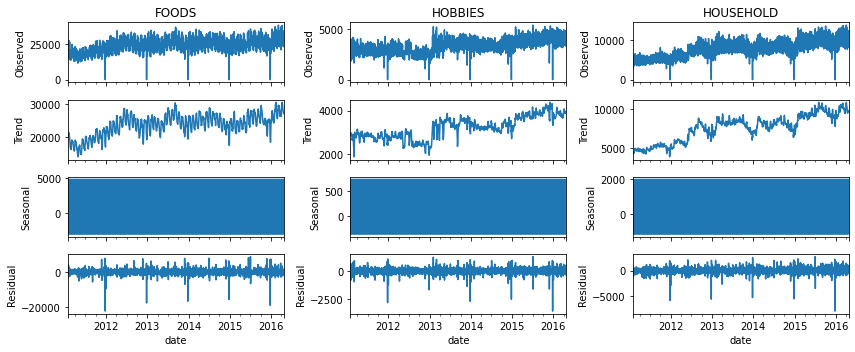


Figure 13: Decomposition across categories

**2.12 Average sales across week days**

In this section, we can see the weekly seasonality of sales.

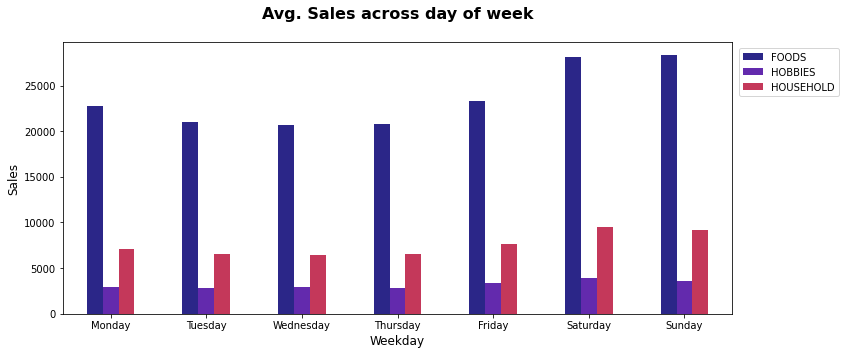


Figure 14: Average sales across week days

Weekly seasonality of sales with highest sales on weekends and lowest sales in Wednesday and Thursday.

Overall sales of all categories on Saturday has more sales than Sunday that may be due to the first day of weekend. In the case of Household and Hobbies although the differences are quite small.

Let’s now see the average sales across months.

**2.13 Average sales across months**

Here again we see that Sales of FOODS category is highest. FOODS sales are far better than other categories.

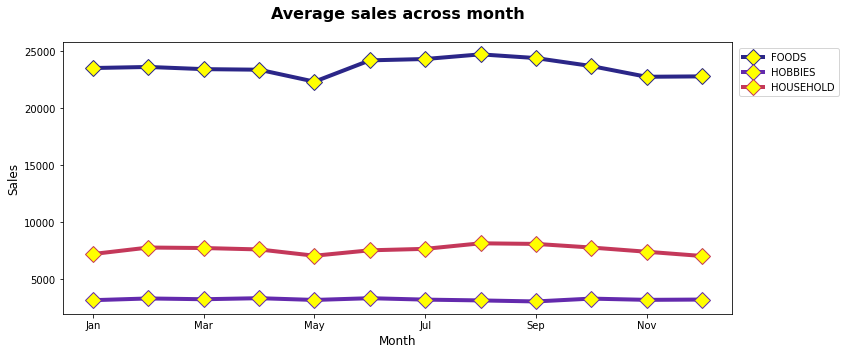


Figure 15: Average sales across months

From the above figure we can see that FOODS sales remain constant till april and decrease in May this can be due to some event. The sales increase again in June and shows increasing trend till September and then the sales decrease.

For HOBBIES, we see that the sales is constant through out the year. It does not show seasonality.

For HOUSEHOLD, We see the sale increases little bit between January to May and decrease in May, this could be owing to events.The sales again see increases after May and decreases in the year end.

**2.14 Events in year**

In our data, we have some events days every year. These days can cause variations in sale of items. In this section we will explore the types of events and in upcoming section we will see how these events affect sales.

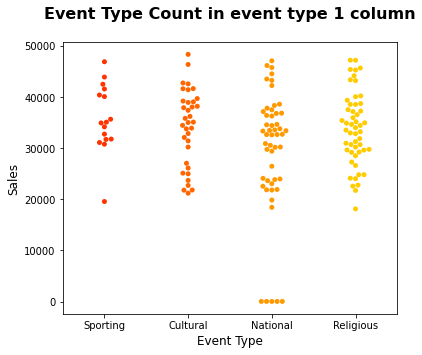


Figure 16 : Event type 1

From Figure 16 and Figure 17, we see that there are two types of events called event\_type\_1 and event\_type\_2.

From figure 16, we can see that event\_type\_1 is further divided into four sub-events called Sporting, Cultural, National and Religious. It is clear from the figure that there are majority or Religious events and Sporting events are the least.

From figure 17, we can see that event\_type\_2 is further divided into two sub-events called Cultural and Religious. It is clear from the figure that there are more Religious events and less Cultural events.



Figure 17: Event type 2

From figure 17, we can see that there are event\_type 2 has less Religious events than cultural events.

**2.15 Event day sales across months**

In this section, lets see how the sales vary during event days and non-event days.

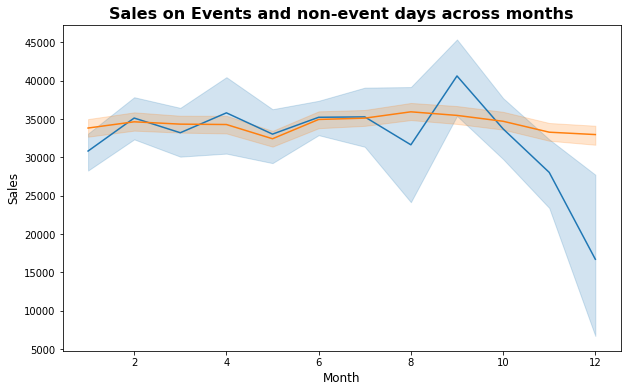


Figure 18: Sales variation on Event days

In the above graph, plot with orange color show the sales on the days of events and plot with blue show the sales on the non-event days.

This graph clearly shows that there is strong effect of events on the sale.

We can see that some events cause increase in the sale while some causes sales to decrease.

To see more clearly identify which event cause increase and decrease of sales, let’s look at Figure 19 below.

We can see that overall sales are less in May and December that might be because of sporting and national events respectively.

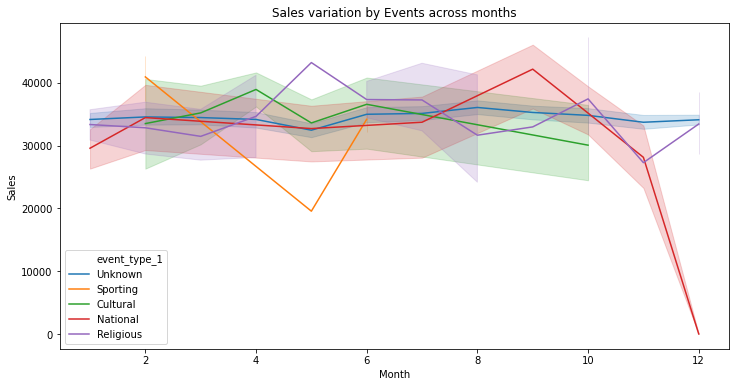


Figure 19: Sales variation on different Events

Different colors show how the events affect on sales.

We see that Sporting events fall between February and June, Cultural events fall between February and December , National and Religious events fall all over the year.

**2.16 Snap days over the year**

In our data, we have some SNAP days every year. SNAP provides nutrition benefits to supplement the food budget of needy families so they can purchase healthy food and move towards self-sufficiency. These days can cause variations in sale of items. SNAP days are supported in Walmart. In this section we will explore the SNAP days and its effects on the sales.

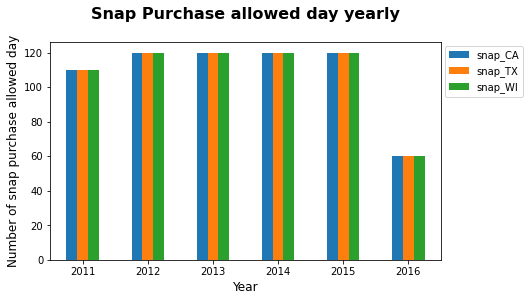


Figure 20 : Snap purchase days

From the above graph we see that the snap days for 2012 to 2016 are 120 and reduced to half in 2016

**2.17 Impact of Snap days**

In this section, let’s see what is the impact of sales on SNAP days.

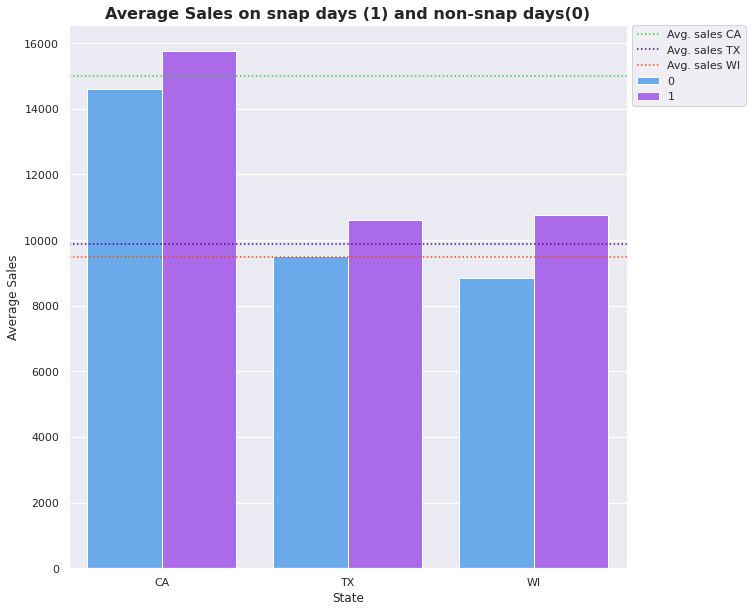


Figure 21: Effect of SNAP days

From Figure 21, we can see that there is increase in sales on the SNAP days in all states. All the states have higher sales on SNAP days. We can see the maximum average increase in WI (around two thousand) while CA and TX have one thousand more sales on each day of SNAP.

The dotted line on graph shows the states average sales.

**Chapter 3**

**Forecasting models**

Making predictions about the future is called extrapolation in the classical statistical handling of time series data. More modern fields focus on the topic and refer to it as time series forecasting. Forecasting involves taking models fit on historical data and using them to predict future observations. Descriptive models can borrow for the future (i.e. to smooth or remove noise), they only seek to best describe the data. An important distinction in forecasting is that the future is completely unavailable and must only be estimated from what has already happened.

In order to do the prediction, we need to training data set and test data set. The M5 forecasting accuracy dataset availlable on kaggle there are two different data for validation and evaluation. While using evaluation table for the purpose of testing, the calculation of root mean square error was not possible. So, to predict the sales the sales\_train\_validation was splited into training and testing data.

In order to, analyse how the the models are trained and how well they are good at forecasting the sales first four rows are used. These rows belongs to HOBBIES category. In this project we have used Exponential Smoothing, Holt-Winters Exponential Smoothing, Moving Average, ARIMA and FBProphet models. We will calculate root mean square error for all these models and will compare them to see which model works best for the M5 data.

**3. 1 Exponential Smoothing**

Exponential smoothing is a time series forecasting method for univariate data that can be extended to support data with a systematic trend or seasonal component. Exponential smoothingof [time series](https://www.statisticshowto.com/timeplot/) data assigns exponentially decreasing weights for newest to oldest [observations](https://www.statisticshowto.com/observation-in-statistics/). In other words, the older the data, the less priority (“weight”) the data is given; newer data is seen as more relevant and is assigned more weight. Smoothing parameters (smoothing constants) determines the weights for observations. This parameter controls the rate at which the influence of the observations at prior time steps decay exponentially. Alpha is often set to a value between 0 and 1. Large values mean that the model pays attention mainly to the most recent past observations, whereas smaller values mean more of the history is taken into account when making a prediction.

Exponential smoothing is usually used to make short term forecasts, as longer term forecasts using this technique can be quite unreliable. There are three types of exponential smoothing:

1. **Simple (single) exponential smoothing** uses a weighted moving average with exponentially decreasing weights.
2. **Holt’s trend-corrected double exponential smoothing** is usually more reliable for handling data that shows [trends](https://www.statisticshowto.com/trend-analysis/), compared to the single procedure.
3. **Triple exponential smoothing**(also called the Multiplicative Holt-Winters) is usually more reliable for parabolic trends or data that shows trends *and*[seasonality](https://www.statisticshowto.com/timeplot/#seasonality).

With exponential smoothing  model, we use the seasonal periods and seasonal additive model. The root mean square error value is 0.19. This value changes when we change seasonal periods and type of seasonal component. By trial and error, we get 0.19 is the lowest RMSE.

**3. 2 Holt Winter’s exponential smoothing**

Holt-Winters is one of the most popular forecasting techniques for time series. Holt-Winters is a model of time series behavior. Forecasting always requires a model, and Holt-Winters is a way to model three aspects of the time series: a typical value (average), a slope (trend) over time, and a cyclical repeating pattern (seasonality). Holt-Winters uses exponential smoothing to encode lots of values from the past and use them to predict “typical” values for the present and future.

The three aspects of the time series behavior—value, trend, and seasonality—are expressed as three types of exponential smoothing, so Holt-Winters is called triple exponential smoothing. The model predicts a current or future value by computing the combined effects of these three influences. The model requires several parameters: one for each smoothing, the length of a season, and the number of periods in a season.

With Holt Winter’s exponential smoothing  model, we use the seasonal periods, trend and seasonal additive model. The root mean square error value is 0.20. This value changes when we change seasonal periods and type of seasonal component. By trial and error, we get 0.20 is the lowest RMSE.

**3.3 Moving Average**

Moving average smoothing is a naive and effective technique in time series forecasting. It can be used for data preparation, feature engineering, and even directly for making predictions.

Smoothing is a technique applied to time series to remove the fine-grained variation between time steps. The hope of smoothing is to remove noise and better expose the signal of the underlying causal processes. Moving averages are a simple and common type of smoothing used in time series analysis and time series forecasting. Calculating a moving average involves creating a new series where the values are comprised of the average of raw observations in the original time series. A moving average requires that we specify a window size called the window width. This defines the number of raw observations used to calculate the moving average value. The “moving” part in the moving average refers to the fact that the window defined by the window width is slid along the time series to calculate the average values in the new series.

With this implementation, the best achievable RMSE value is 0.25.

**3.3 SARIMAX**

ARIMA stands for Autoregressive Integrated Moving Average. It is a generalization of the simpler Autoregressive Moving Average and adds the notion of integration.

An [ARIMA model](https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average) is a class of statistical models for analyzing and forecasting time series data. It explicitly caters to a suite of standard structures in time series data, and as such provides a simple yet powerful method for making skillful time series forecasts.

A linear regression model is constructed including the specified number and type of terms, and the data is prepared by a degree of differencing in order to make it stationary, i.e. to remove trend and seasonal structures that negatively affect the regression model.

Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component .It adds three new hyperparameters to specify the autoregression, differencing and moving average for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

The Seasonal Autoregressive Integrated Moving-Average with Exogenous Regressors ([SARIMAX](https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/)) is an extension of the SARIMA model that also includes the modeling of exogenous variables. Exogenous variables are also called covariates and can be thought of as parallel input sequences that have observations at the same time steps as the original series. The primary series may be referred to as endogenous data to contrast it from the exogenous sequence(s). The observations for exogenous variables are included in the model directly at each time step and are not modeled in the same way as the primary endogenous sequence. The method is suitable for univariate time series with trend and/or seasonal components and exogenous variables.

In this implementation, the best achievable RMSE value is 0.18.

**3.5 Prophet**

Prophet is an open source time series forecasting project by Facebook. It is based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, including holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. It is also supposed to be more robust to missing data and shifts in trend compared to other models.

With this implementation, we achieved RMSE of 0.22 when we set weekly seasonality to true. We can see the drastically increases and RMSE changes to 4 when we set yearly seasonality.

**3.6 Model Comparison**

In this section, we compare the models. Let’s look at the figure below where we have plot RMSE loss for the models we have implemented.

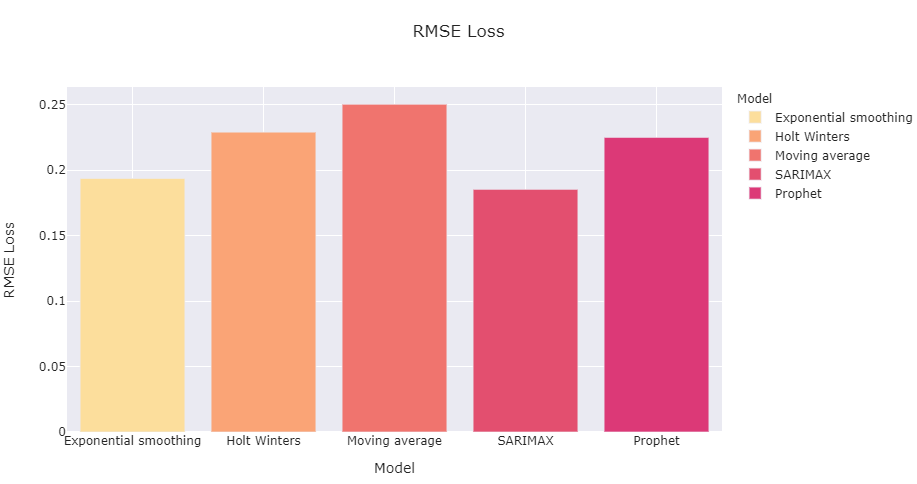


Figure 21: Forecasting model comparisons

From figure 21, we can see that SARIMAX is the best forecasting model for our data with root mean square error of 0.18. Exponential Smoothing is the second best fit 0.19.

**Conclusion**

While working on the project we explored our data, trained models and did predictions. We learnt to see analyze different aspects of data. We understand the seasonality and trends. Let’s say we have a business which is seasonal or we know that there are slow or busy periods at certain times during the year, plan on increasing or reducing staff when necessary. This can help us bring additional workers on early enough to train them or use contractors during busy times instead of hiring employees. Knowing demand can also help us plan the sales levels at which adding another shift or expanding your production capacity is profitable.

We have implemented various forecasting models and we see that SARIMAX and Exponential Smoothing model works best for our data having the least root mean square error of 0.18 and 0.19 respectively.

**References**

1. “The importance of Business Forecasting” available at <https://cashmanager.io/blog/the-importance-of-business-forecasting/>
2. Sam Ashe –Edmunds, “The Advantage of Accurate Forecasting” available at <https://smallbusiness.chron.com/advantages-accurate-forecasting-60830.html>
3. Shelby Blitz, ‘What is Exploratory Data Analysis?’ Available at https://www.sisense.com/blog/exploratory-data-analysis/
4. Ben Goldstein, ‘What does ‘seasonality really means in sales?’ Available at <https://www.nutshell.com/blog/sales-seasonality-definition-and-causes/>
5. FNS, ‘Supplemental nutrition assistance program (snap),’ Available at https://www.fns.usda.gov/snap/supplemental-nutrition-assistance-program.
6. J. Brownlee, What is time series forecasting?" Available at https:// machinelearningmastery.com/time-series-forecasting/.
7. “Exponential Smoothing: Definition of Simple, Double and Triple” , Available at <https://www.statisticshowto.com/exponential-smoothing/>
8. J. Brownlee, “A Gentle Introduction to Exponential Smoothing for Time Series Forecasting in Python” , Available at <https://machinelearningmastery.com/exponential-smoothing-for-time-series-forecasting-in-python/>
9. O. Matter, “Holt-winters forecasting simplified” available at https://orangematter.solarwinds.com/2019/12/15/holt-winters-forecasting-simplified/
10. J. Brownlee, “Moving Average Smoothing for Data Preparation and Time Series Forecasting in Python”, Available at <https://machinelearningmastery.com/moving-average-smoothing-for-time-series-forecasting-python/>
11. J. Brownlee, “Classical time series forecasting methods in python(cheat sheet)," Available at https://machinelearningmastery.com/time-series-forecasting-methods-in-python-cheat-sheet/
12. “Prophet - Forecasting at scale," Available at <https://facebook.github.io/prophet/>