**Time Series Analysis/Forecast on the Revenue of Walmart Stores   
in California**

**DSO522 - Applied Times Series Analysis for Forecasting**

Team 10  
YiHsin (Amy) Chung, Puhsin Huang, John Kim,   
Dong Wook (Sam) Ko, Yura Shakhnazarian

**Table of Contents**

[**Executive summary**](#_3oaw2yt8txo) **2**

[**Introduction and Motivation**](#_arp06glip23n) **3**

[**Data Preprocessing**](#_qrj1ms399z4l) **4**

[**Model Building Procedures & Description**](#_tryd5c3ik6tk) **6**

[Weekly Model](#_4sn2e1orjzrv) 6

[Monthly Models](#_cir802i32mpy) 8

[Forecasting Model](#_ssvw8xwqqw6m) 8

[Linear Regression Model with External Variables](#_3qhq929szftb) 10

[**Model Result & Evaluation**](#_p8zckb86dznr) **11**

[Weekly Model](#_eb0r9yyqcior) 11

[Monthly Models](#_pbz1zucdlm94) 12

[Forecasting Model](#_cxomihgzv0n8) 12

[Linear Regression Model with External Variables](#_8nttnzekqq7o) 14

[**Shiny Application**](#_2t3nuhes7o37) **16**

[**Business Interpretation of the Result**](#_7xygmwsfwgjz) **17**

[Short-Term Projects: Improve digital marketing campaigns](#_jd2k4cu0kszl) 17

[Long-Term Projects](#_rd8gxg70ilzm) 18

[**Conclusion**](#_uk7ajfwfb15i) **20**

[**Appendix**](#_cdgr01lxrte6) **21**

# 

# Executive summary

The purpose of the team’s project is to present time series models and business strategies Walmart can adopt by utilizing the results of these models. The main data the team worked with was the daily sales of the Walmart stores in California from 01/29/2011 to 06/19/2016. As the team was interested in exploring the short and long-term analysis, the data was transformed into weekly and monthly data.

Using the weekly data, the team was able to create a short-term forecasting model with the forecast horizon of 16 weeks. The team identified an SARIMA(1,1,0)x(1,1,0)52 model with training MAPE of 1.92 and testing MAPE of 2.47. With the confidence in our model, the team also recommended possible methods to improve the digital marketing campaigns for Walmart.

The team created both a forecasting model and a linear regression for the monthly data. Similar to the short-term forecast, the best model with the forecast horizon of 12 months was the SARIMA(1, 1, 2)x(1, 1, 0)12 model with training MAPE of 2.12 and testing MAPE of 6.71. Through linear regression that utilized other external variables, such as sales from other states, sales by categories, holidays, and more, the team was able to identify the contributing factors of the sales in California. With the of 0.89, the team is confident to apply the findings to help the marketing, product, and logistics departments in Walmart to plan ahead.

The team hopes that this project will not only report the best models from the initial data but also deliver tangible benefits to Walmart’s business. In order to facilitate better understanding of our models using visualizations, the team has also created a Shiny app. The link for the application can be found [here](https://yuridias.shinyapps.io/shiny/).

# Introduction and Motivation

Forecasting sales is a common task in a real-world setting. Inaccurate business forecasts could lead to actual or opportunity losses. The methods used can be applied in various business areas, including marketing planning, inventory management, and logistic arrangement. We believe that the chosen topic would lead to a rewarding experience of solving a problem that closely mimics the question the actual retail companies ask themselves regularly.

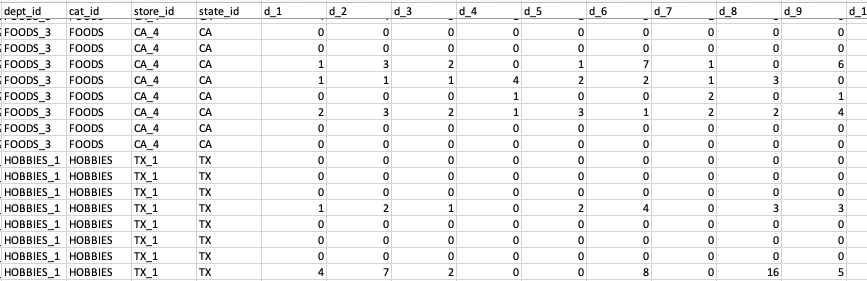
Through this report, the team would like to introduce possible strategies Walmart can launch based on the time series analysis/forecast on its historical sales data. The team has narrowed the scope of the project by considering the sales of Walmart stores in California as the primary variable of interest.

There are 3 main components of the project: Short-Term Analysis, Long-Term Analysis, and business implementation of the results. For the Short and Long-Term Analysis, the team attempted to not only find the best model that led to the most accurate forecasts of the sales but also interpret what factors contribute to the sales the most by also considering external variables. Based on the results of these analyses, the team also devised potential strategies that can directly impact Walmart’s business. The team believes that delivering accurate sales predictions is crucial in actualizing these strategies and seek to create the best possible models.

# Data Preprocessing

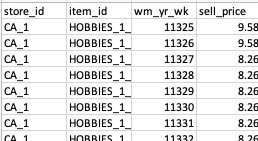
Before performing the actual analysis, the team first prepared the data. The source of the data is a Kaggle competition organized by the University of Nicosia. The competition provided the sales information of the Walmart stores in California, Texas, and Wisconsin over approximately 5.5 years (January 29, 2011 ~ June 19, 2016). There were mainly 2 datasets that were used for the project: Units Sold and Price. For additional details about the data source, please refer to this [link](https://www.kaggle.com/c/m5-forecasting-accuracy/overview).

The “Units Sold” dataset included the number of units sold daily at each store. As seen in the screenshot of the dataset below, each row represented different items under different categories and departments.



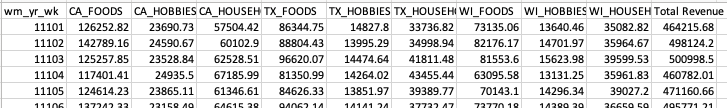
“Units Sold” Dataset from the original data source

Another dataset that was provided included the prices of the same items at different locations. Unlike the quantity dataset, the prices were updated weekly, which made sense for a retailer like Walmart.

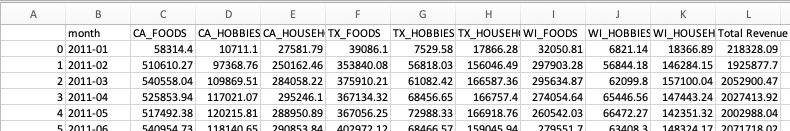


“Prices” Dataset from the original data source

Combining the 2 datasets just mentioned, the team created a new dataset that showed the total daily revenue for all of the locations combined. The process required processing the datasets to be in the same format and matching the period. During this step, the team also decided to convert the daily data into weekly and monthly sales data. This decision not only fits better to reach our goal but also satisfies the technical capacity of the models the team used in the next section. Please refer to the screenshot below for the processed dataset that was used in the analysis. Because the focus of the project was on the stores in California, the team looked at the total revenue in California as the variable of interest.



Weekly Sales Data after preprocessing



Monthly Sales Data after preprocessing

# 

# 

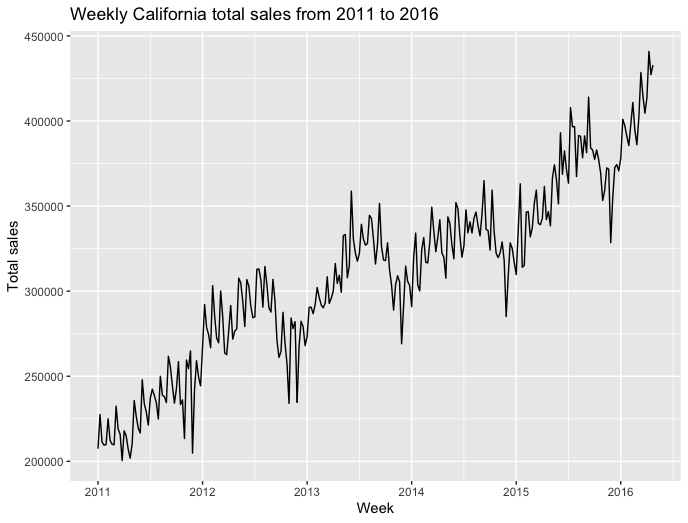
# 

# 

# Model Building Procedures & Description

## Weekly Model

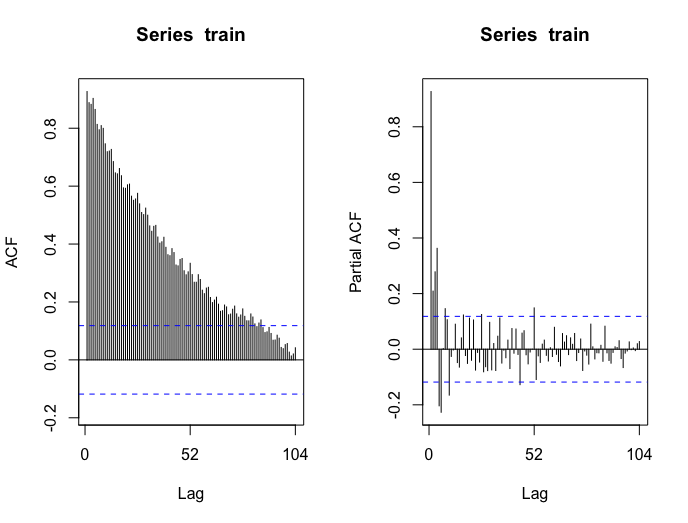
The first step is to look at the trend plot of California’s weekly sales below. In terms of time series components, there is the level, the trend, the seasonal component, and the noise. An obvious increasing trend of the weekly Walmart sales can be observed. For the seasonal component, there seems to be a yearly seasonal cycle as well as a monthly cycle.



After understanding the weekly sales, the next step is to implement different time series forecast models and pick the final model based on the best performance on the testing set. The forecast horizon is set to 16 weeks to reduce the risk of biased data for a smaller time window. The following are the models the team has tried:

* + Naive Forecast
  + Seasonal Naive Forecast
  + Moving Average (trailing), w= [4, 52]
  + Moving Average (trailing & rolling forward), w= [4, 52]
  + Simple Exponential Smoothing
  + Holt-Winters Exponential Smoothing
  + ARIMA models

The team referred to the autocorrelation function (ACF) and partial autocorrelation function (PACF) to decide sets of parameters to explore.



From the plot above, the ACF plot is trailing off whereas the partial ACF plot shows a huge cut off after the first lag. However, lag 2, lag 3, and lag 4 are all showing significant relationships in the partial ACF plot. Also, there is a seasonal pattern every 4 lag in the ACF plot, and the pattern is trailing off. The seasonal part of the partial ACF plot seems to cut off after the first monthly season (lag 4) and yearly season (lag 52). Accordingly, the team selected the following models to implement.

* order=(1, 1, 0), seasonal order=(1, 1, 0), seasonal period=4
* order=(1, 1, 0), seasonal order=(1, 1, 0), seasonal period=52
* order=(2, 1, 0), seasonal order=(1, 1, 0), seasonal period=4
* order=(2, 1, 0), seasonal order=(1, 1, 0), seasonal period=52
* order=(3, 1, 0), seasonal order=(1, 1, 0), seasonal period=4
* order=(3, 1, 0), seasonal order=(1, 1, 0), seasonal period=52
  + Time Series Linear Regression Model (tslm with trend & season)
  + Linear Regression Model with external variables

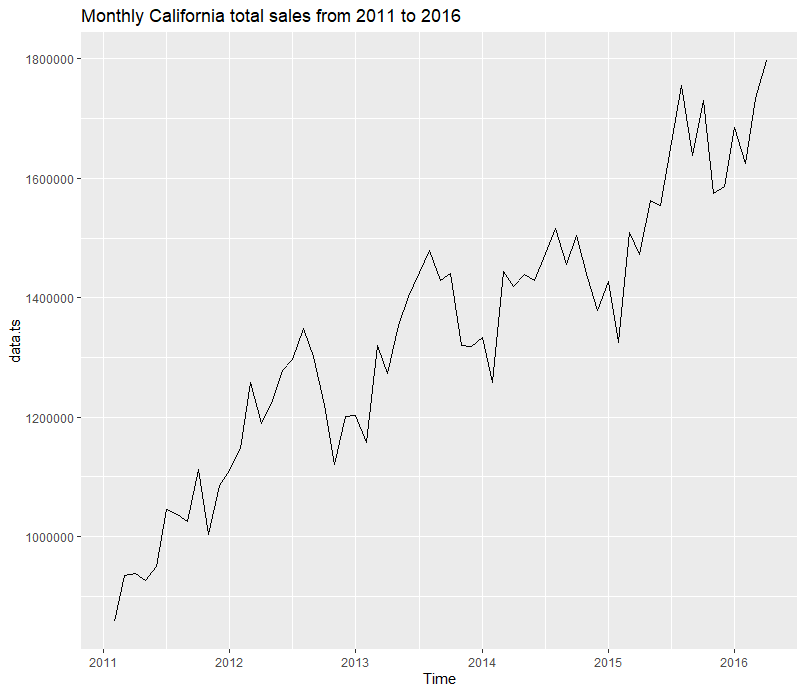
Since sales data of different categories and different states are available, we would like to know whether we could use historical sales of different categories and different states to forecast total sales in California. Additionally, the number of promotion events per week is also available to us. The following are the independent variables that we put into the linear regression model.

* Lag 1 total sales in California
* Lag1 total sales in Texas
* Lag1 total sales in Wisconsin
* Lag 1 sales of food category in California
* Lag1 sales of hobbies category in California
* Lag1 sales of household category in California
* Number of total promotion events

## Monthly Models

### Forecasting Model

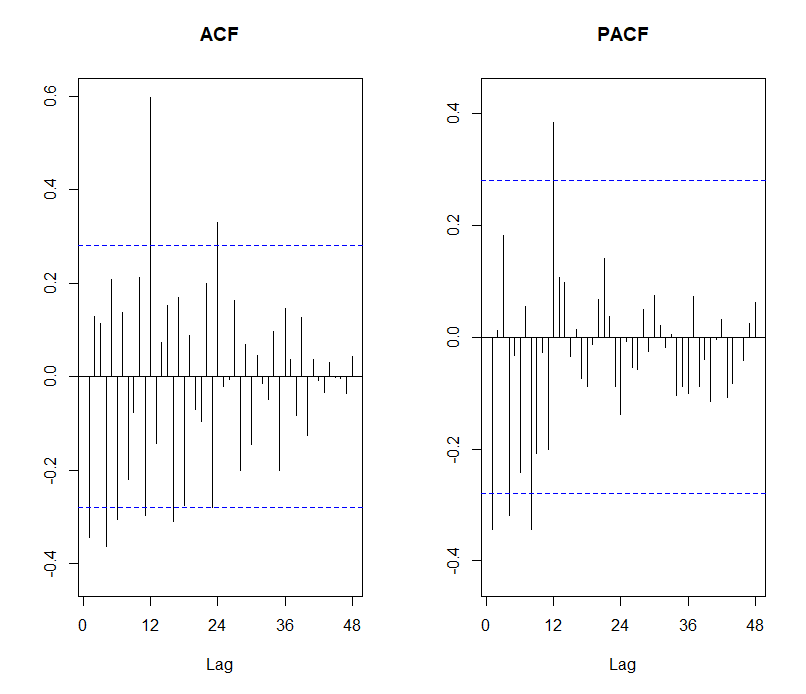
Based on the monthly sales plot for California, it seems that there is a level, trend, noise and a little bit of seasonal component. An increasing trend of the monthly Walmart sales seems obvious but it is difficult to observe a clear seasonal component.



After understanding the monthly sales, our team tried different time series forecast models and picked the final model based on the best performance on the testing set. The forecast horizon was set to 12 months. The following are the models the team tried:

* Naive Forecast
* Seasonal Naive Forecast
* Moving Average (trailing & rolling forward), w=12
* Holt-Winters Exponential Smoothing
* ARIMA models

For the ARIMA model, we differenced the dataset by 1 lag to make it stationary and then looked at the autocorrelation function (ACF) and partial autocorrelation function (PACF) to decide sets of parameters to explore.



From the plot above, both the ACF and PACF are trailing off but it is difficult to see significant spikes. In addition, there seems to be a little bit of seasonal pattern every 12 lag in the ACF, and the pattern is trailing off while the seasonal part of the PACF seems to cut off after the first monthly season (lag 12). Accordingly, the team selected the following models to implement.

* order=(1, 1, 1)
* order=(2, 1, 2)
* order=(3, 1, 3)
* order=(1, 1, 1), seasonal order=(1, 1, 0), seasonal period=12
* order=(2, 1, 1), seasonal order=(1, 1, 0), seasonal period=12
* order=(1, 1, 2), seasonal order=(1, 1, 0), seasonal period=12

### Linear Regression Model with External Variables

Here, we tried to find the relationship between external variables and sales. For the external variables, we looked at historical sales of different categories and different states, the number of promotions per month, and economic factors. We looked at the cross-correlation of each variable with respect to the sales to determine the most correlated lag time. However, we decided to use the same time as the sales for the promotional variables since promotions are planned by Walmart in advance. In the end, the following independent variables were put into the linear regression model:

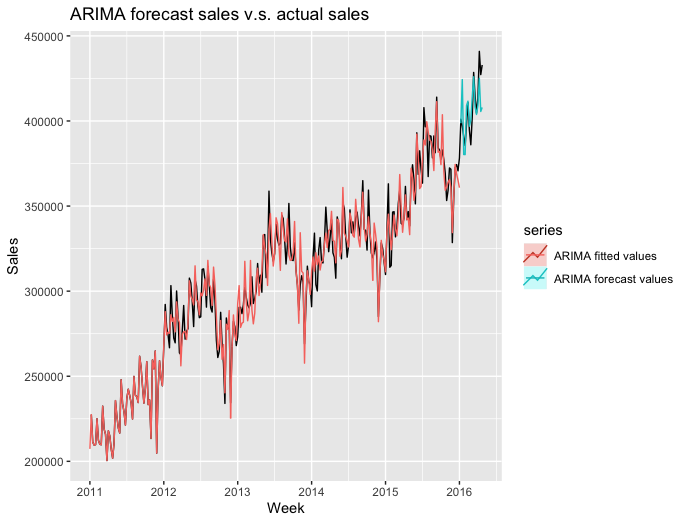
* total sales in Texas (lag 1)
* total sales in Wisconsin (lag 1)
* % of sales for food category in California (lag 7)
* % of sales for hobbies category in California (lag 5)
* % of sales for household category in California (lag 8)
* Number of total promotion events
* Number of total promotion events on weekends
* Number of total promotion events for cultural type
* Number of total promotion events for religious type
* Number of total promotion events for sporting type
* Monthly unemployment rate (lag 1)
* Quarterly real GDP (lag 1)
* Quarterly personal income (lag 1)

Then, we removed the correlated variables and insignificant variables one step at a time by VIF score and p-values.

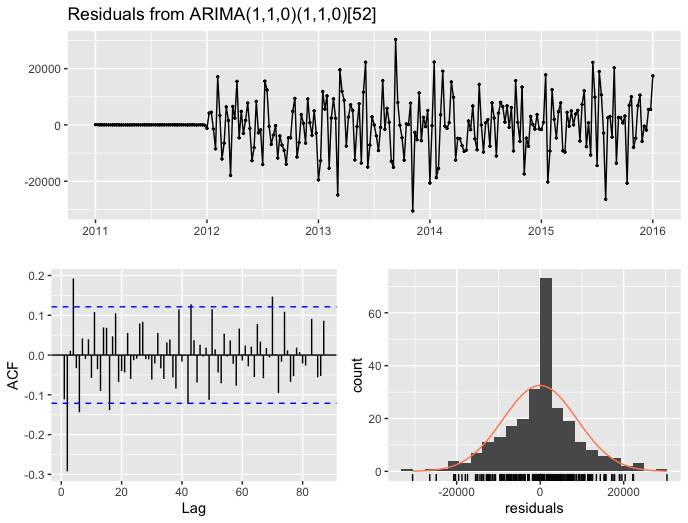
# Model Result & Evaluation

## Weekly Model

Among all the models we tried, the best one is an ARIMA model with order=(1, 1, 0), seasonal order=(1, 1, 0), and seasonal period=52. The model’s MAPE is 1.92 on the training set and 2.47 on the testing set. The error of the testing set is slightly higher than that of the training set, which is considered reasonable. From the line plot below, the forecasted values overlap with the actual values pretty well, and the model does a good job in capturing the trend and seasonality.



We evaluated the model’s robustness by checking whether the residuals meet the assumptions of random spread, independence, and normal distribution. As you can see, the residuals are pretty random and did not have any obvious pattern. The histogram of the residuals is normal distributions. However, the ACF plot indicates that the residuals are not completely independent. It looks like the residuals are correlated at lag 2 and lag 4. To sum up, the model’s residuals meet the assumption of normality and random spread but do not meet the assumption of independence.

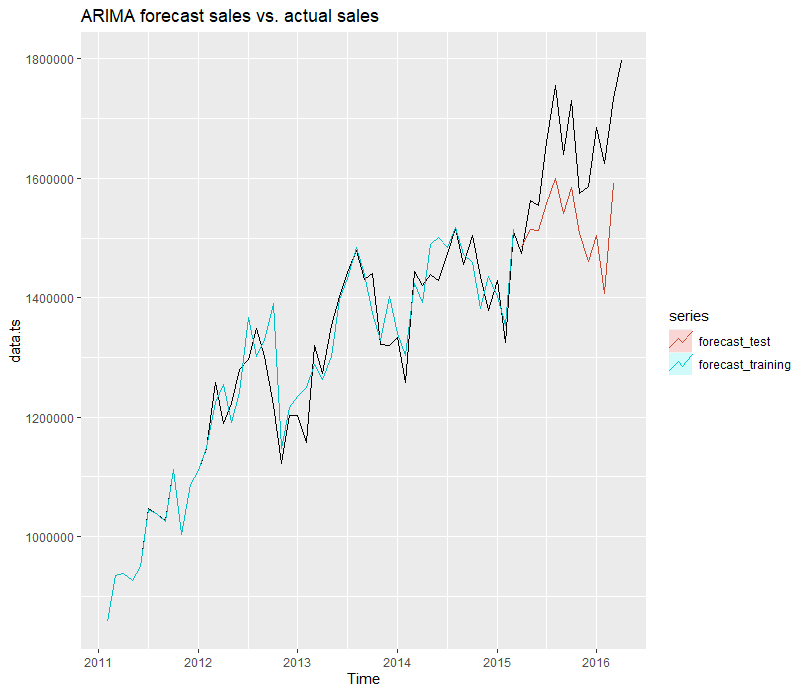


Results of all the models are summarized in a table in the Appendix.

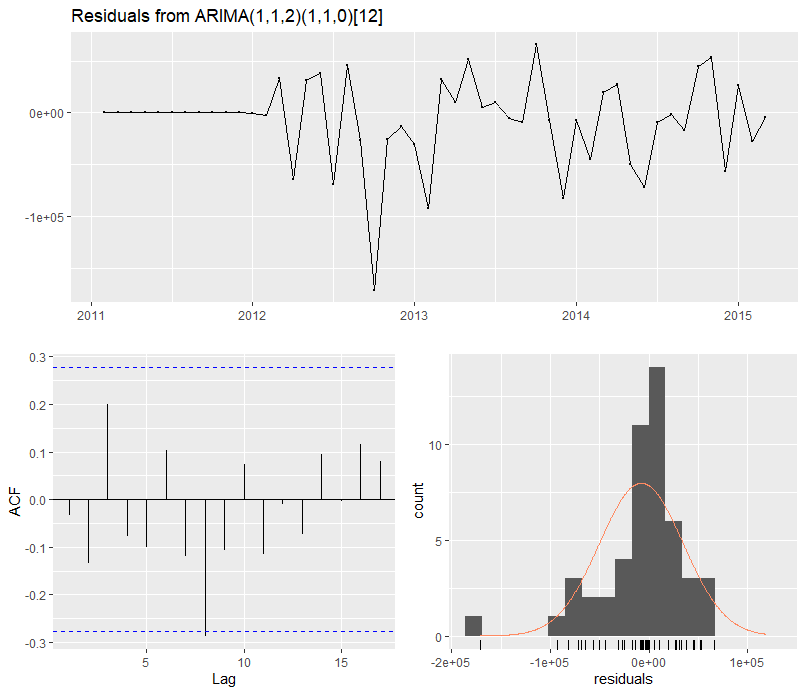
## Monthly Models

### Forecasting Model

Among all the models we tried, the best one is an ARIMA model with order=(1, 1, 2), seasonal order=(1, 1, 0), and seasonal period=12. The model’s MAPE is 2.12 on the training set and 6.71 on the testing set. However, the model seems to be not really robust since the difference between the error in the training and test set is quite high.



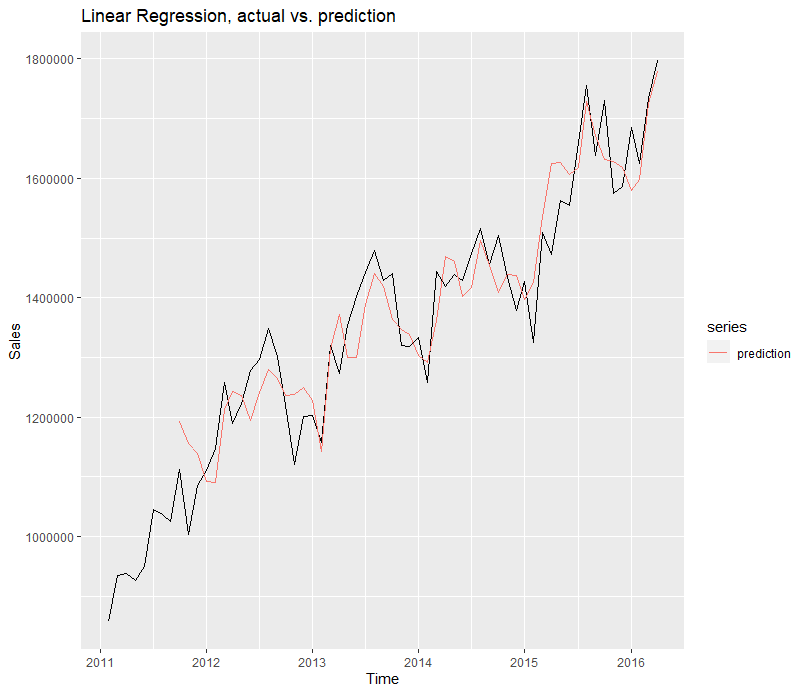
Next, we checked whether the residuals meet the assumptions of constant spread, independence, and normal distribution. As you can see below, the residuals seem to have a constant spread except for one drop between 2012 and 2013. The ACF indicates that the residuals are independent but the histogram suggests that the residuals are not normally distributed.



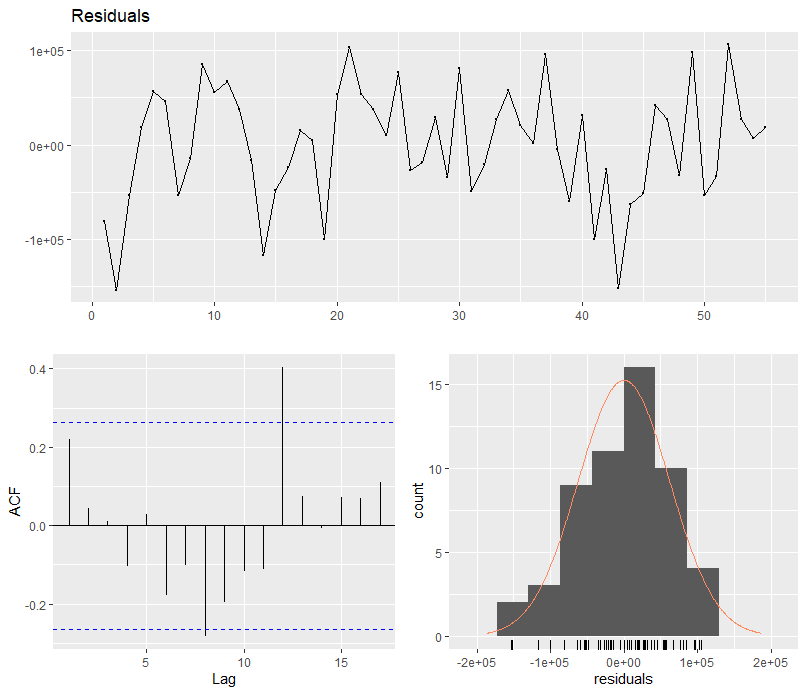
Results of all the models are summarized in a table in the Appendix.

### Linear Regression Model with External Variables

The final model has R squared value of 0.89 with 6 variables: total sales in Texas (lag 1), % of sales for hobbies category in California (lag 5), number of total promotion events, number of total promotion events for cultural type, number of total promotion events for religious type, Quarterly personal income (lag 1).



Next, we checked whether the residuals meet the assumptions of constant spread, independence, and normal distribution. As you can see below, it seems that the residuals are homoscedastic, independent except one point around lag 12, and normally distributed approximately.



# Shiny Application

[Link to the Shiny App (https://yuridias.shinyapps.io/shiny)](https://yuridias.shinyapps.io/shiny/)

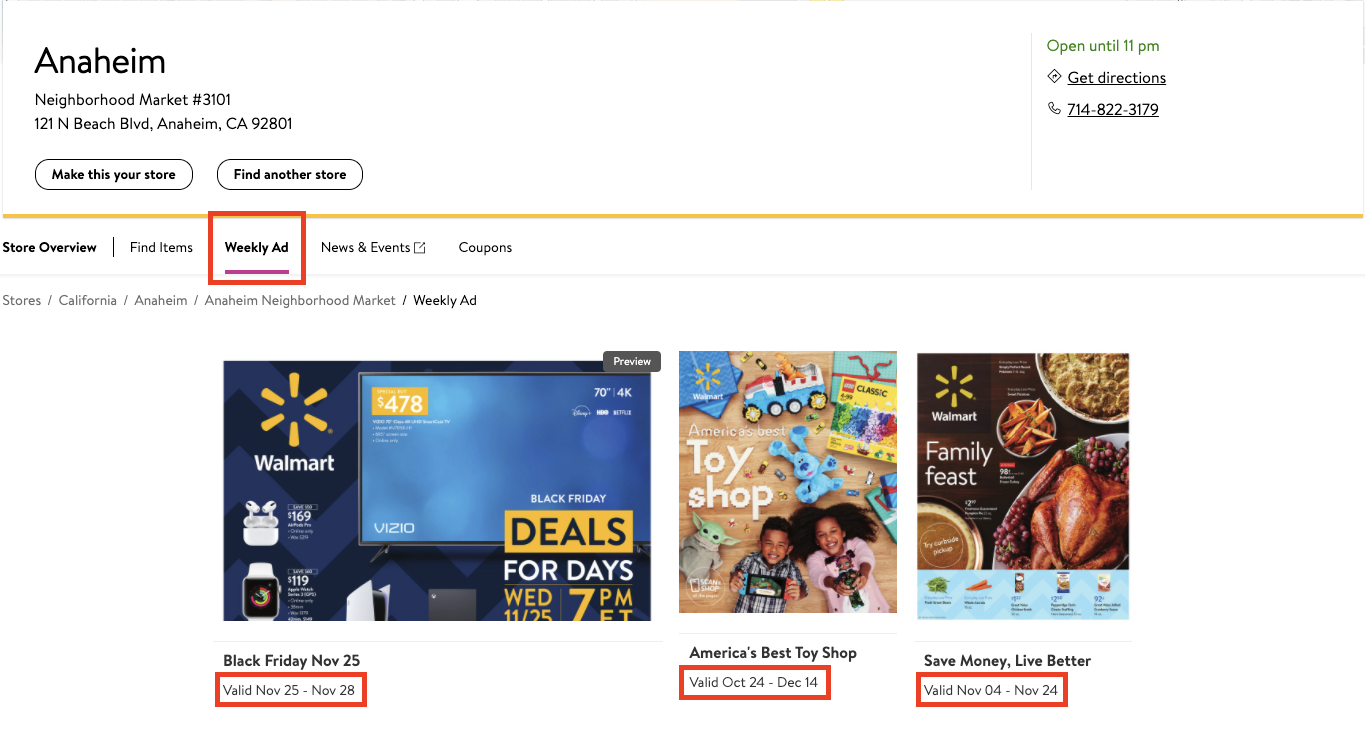
The team has also created a Shiny application to demonstrate our data and models more visually. The application contains the historical sales data along with the modelling implications discussed above. Note that the application is fully interactive and allows the users to manipulate different variables, including frequency, time coverage, forecast horizon, and more. The team hopes that the stakeholders would be able to have a better grasp of our models and explore different scenarios using our application. Refer to the Appendix for a screenshot of the Shiny App.

# Business Interpretation of the Result

As seen above, the team had created different models based on weekly and monthly revenue data. Accordingly, these different analyses will have different impacts on Walmart’s business. For example, the forecast from the weekly data will be used to generate a short-term strategy regarding the digital marketing campaigns whereas the results from the monthly data will be used as a tool to check the contributing factors to the revenue and yearly prediction.

## Short-Term Projects:Improve digital marketing campaigns

As seen in the screenshot below, there is a *Weekly Ad* section on Walmart’s official website for all of the stores not just in California but the entire country. Unfortunately, this feature is not utilized as efficiently as possible. As seen in the screenshot below, despite being under the *Weekly Ad* section, the actual advertisements last for about a month. The team believes that the marketing department can utilize our results from the weekly data to improve their digital marketing efforts.



For example, if the revenue for upcoming weeks is expected to decrease, the analytics team can advise the marketing team to boost the promotion. On the other hand, if the sales are expected to increase, Walmart can also try to push the sales further to surpass the predicted values by advertising through additional channels besides their own website:

1. Banner Ads:

Banner ads are just another approach to attract customers. If the time series analysis suggests that the sales will increase at a certain holiday, it is likely that sales of our competitors will also increase. Hence, we can further compete with our competitors by utilizing banner ads. Later, we can analyze how many percent of purchases are sourced from this channel, and if this channel helps us reach beyond our predicted sales.

1. Email marketing:

Walmart can send out personalized emails based on the predictions of the model. With the data that Walmart already has on each customer, such as past purchase behavior or customer segmentation using the rfm(recency, frequency, monetary value) model, Walmart can send out promotional emails at the ideal timing when the stores are predicted to need a boost of sales.

There are also future possibilities regarding this marketing strategy. The data we worked with for this project is the total revenue for all of the stores in California. However, once the team has more detailed data, we can expand our short-term analysis to all of Walmart’s locations. As a result, each store will have its unique prediction for upcoming weeks and therefore have different advertising strategies. We can develop this model even further if we can incorporate what the customers buy from each store to identify the exact products to be on sale for the optimal price level.

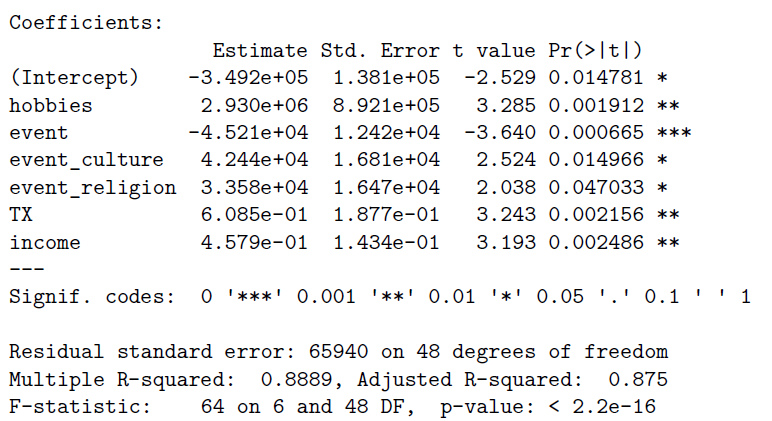
## Long-Term Projects

Knowing the long-term sales trend by forecasting for the next 12 months allows Walmart to plan ahead. For example:

* For the product/vendor managers, If the sales are decreasing, it can consider launching more new products or removing ones that are not so popular.
* For the in-stock managers, the analysis helps them better plan on how to manage the warehouse space in a more efficient way.
* For the marketing managers, the result will help them plan on offline marketing campaigns, which usually take more time to prepare than digital campaigns.

The forecast above was only based on past sales data. However, we would like to further understand the data to identify important factors that have an impact on future sales so we can prepare for the change and build strategies in advance. Therefore, our team built a linear regression model with external variables as mentioned in section 4 and 5. With this model, we are able to interpret how the change of an individual variable will impact sales by inspecting the coefficients.

There could be mainly two business applications from this model. Firstly, a sudden change in the lagged variables like percentage sales for the hobbies category in California or quarterly personal income could be the indicators or signals for Walmart to prepare for the upcoming changes in their revenue stream. Also, the promotional variables can be used to measure the success of the promotions in terms of revenue or to optimize the mix of the promotion categories. For example, we can see that the current impact of the events on the revenue is actually negative regardless of the event type.

****

# 

# Conclusion

This report presents our team’s investigation of Walmart sales data between 2011 and 2016 through the prism of time series analysis and forecasting. In particular, we narrow down our focus to revenues of Walmart stores in California - with two major goals of predicting them accurately and robustly and extracting valuable business insights that could help the company.

Having touched on the motivations behind our topic and these goals as well as the data preprocessing steps, we decided to break down our work into Short-Term Analysis, Long-Term Analysis, and Business Implications. For Short-Term Analysis, we first inspected the weekly series visually and then fitted several statistical models to forecast future sales, out of which a SARIMA(1,1,0)x(1,1,0)52 gave the best results while being quite robust at the same time. For Long-Term Analysis which covered monthly data, the overall methodology was similar; however, the obtained best model was now a SARIMA(1, 1, 2)x(1, 1, 0)12 that unfortunately was not as robust. Independent of this, we also explored the effect of exogenous factors on monthly sales via a linear regression model, obtaining a handful of significant predictors. Finally, for Business Implications, we identified potential short-term projects related to digital marketing campaigns and long-term projects related to the product, inventory, and marketing management.

While we are certain of the practical value that our models and recommendations have, there is definitely room for improvement and further experimentation. Specifically, with more time and knowledge on our hands, we would like to include various exogenous factors in our final models of choice, collect more data, and try other, non-time series algorithms for forecasting. That being said, we hope that this read was worth the time and that our findings can find their utilization by Walmart while incentivizing further research and analyses on the subject matter.

# 

# 

# 

# Appendix

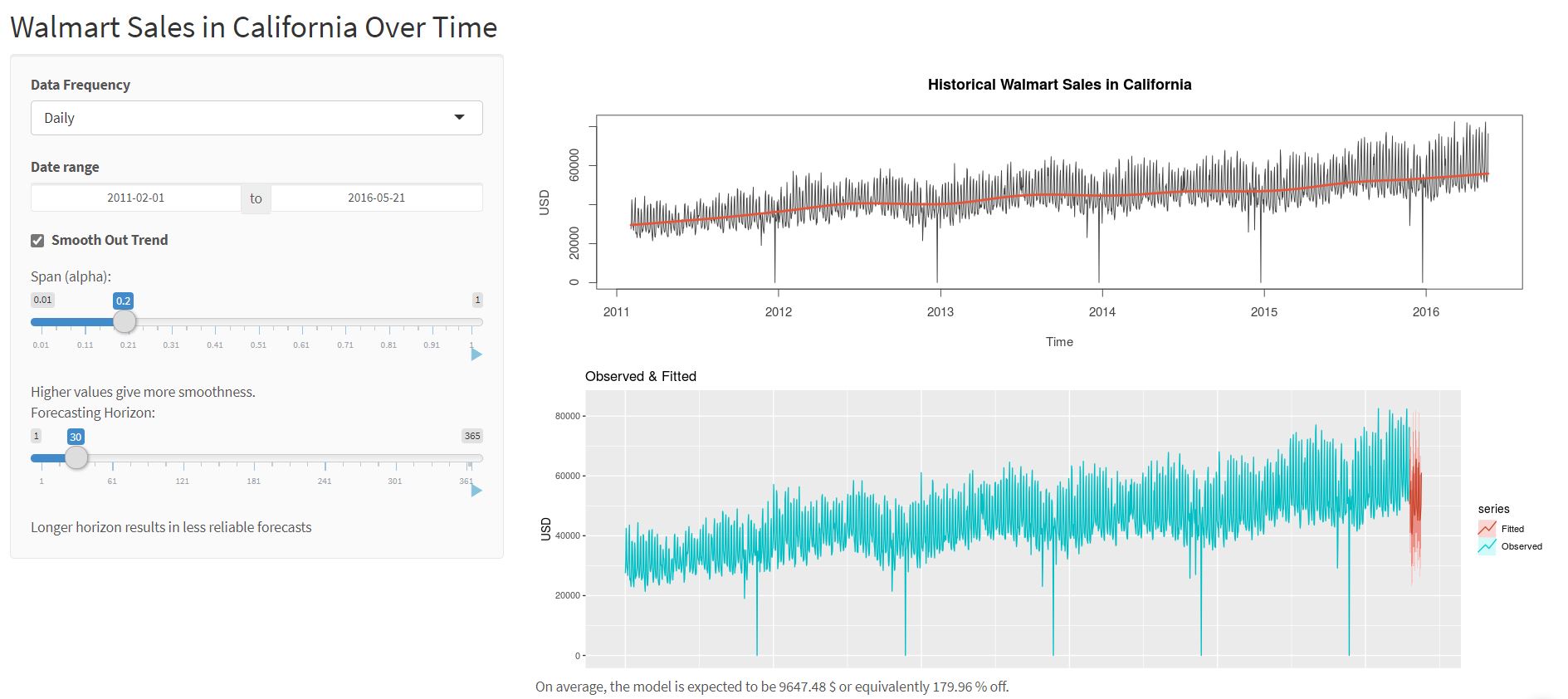
Appendix 1 - Weekly Model Performance

|  |  |  |
| --- | --- | --- |
| Model | Train MAPE | Test MAPE |
| Naive Forecast | 4.59 | 7.16 |
| Seasonal Naive Forecast | 10.64 | 16.11 |
| Moving Average, w=4 | 3.36 | 8.20 |
| Moving Average, w=52 | 6.75 | 10.44 |
| Moving Average rolling forward, w=4 | 3.36 | 2.81 |
| Moving Average rolling forward, w=52 | 6.75 | 8.32 |
| Simple Exponential Smoothing | 3.99 | 8.91 |
| Holt-Winters Exponential Smoothing | 4.03 | 7.18 |
| ARIMA, order=(1, 1, 0), seasonal order=(1, 1, 0), seasonal period=4 | 4.00 | 6.75 |
| ARIMA, order=(1, 1, 0), seasonal order=(1, 1, 0), seasonal period=52 | 1.92 | 2.47 |
| ARIMA, order=(2, 1, 0), seasonal order=(1, 1, 0), seasonal period=4 | 3.91 | 5.48 |
| ARIMA, order=(2, 1, 0), seasonal order=(1, 1, 0), seasonal period=52 | 1.88 | 2.70 |
| ARIMA, order=(3, 1, 0), seasonal order=(1, 1, 0), seasonal period=4 | 3.77 | 2.32 |
| ARIMA, order=(3, 1, 0), seasonal order=(1, 1, 0), seasonal period=52 | 1.82 | 3.01 |
| Time Series Linear Regression Model | 3.81 | 5.39 |
| Linear Regression Model with external variables | 4.25 | 2.90 |

Appendix 2 - Monthly Model Performance

|  |  |  |
| --- | --- | --- |
| Model | Train MAPE | Test MAPE |
| Naive Forecast | 4.59 | 7.63 |
| Seasonal Naive Forecast | 10.61 | 11.44 |
| Moving Average rolling forward, w=12 | 6.51 | 7.01 |
| Holt-Winters Exponential Smoothing | 2.22 | 8.20 |
| ARIMA, order=(1, 1, 1) | 4.38 | 10.22 |
| ARIMA, order=(2, 1, 2) | 4.40 | 8.95 |
| ARIMA, order=(3, 1, 3) | 3.18 | 9.07 |
| ARIMA, order=(1, 1, 1), seasonal order=(1, 1, 0), seasonal period=12 | 2.14 | 6.93 |
| ARIMA, order=(2, 1, 1), seasonal order=(1, 1, 0), seasonal period=12 | 2.15 | 6.95 |
| ARIMA, order=(1, 1, 2), seasonal order=(1, 1, 0), seasonal period=12 | 2.12 | 6.71 |

Appendix 3 - Screenshot of the Shiny App

****