

# Bellevue University

## DSC 630 Predictive Analytics

### **Walmart Sales Forecasting**

Nitin Mahajan | Ganesh Kale

# Project Introduction:

- **Intro - Walmart** operates Walmart, Walmart Neighborhood Market, Wal-Mart, Walmart.com, and Sam's Club and retail companies like this commonly having issues with predicting sales accurately throughout the days, months, and years ahead. There are many varying factors that can cause issues with predicting sales.
- **Goal** – Build the Machine Learning model that would learn from past records, events and Predict the Sales forecast for Store and its departments on specific week of the year.
- **Model - Sales Forecasting** is the process of using a company's sales records over the past few years to predict the short-term or long-term sales performance of the company in the future. This is one of the pillars of proper financial planning. **Sales Forecast Model** that would learn from the past sales records, events and predict the accurate sales so company will be ready to source appropriate resources before the actual event happens.
- **Benefits** - Business Sales Executives often find themselves scrambling for answers when it comes to sales forecasting during business reviews with their leaderships team. The Sales Forecast Model will help sales executives to find such answers upfront and be ready with numbers and predictions to share with leaderships team. This model would help individual stores to upscale their customer satisfaction by stocking the right products at right time and decrease overstocking and wastage of food products.

# Data Information:

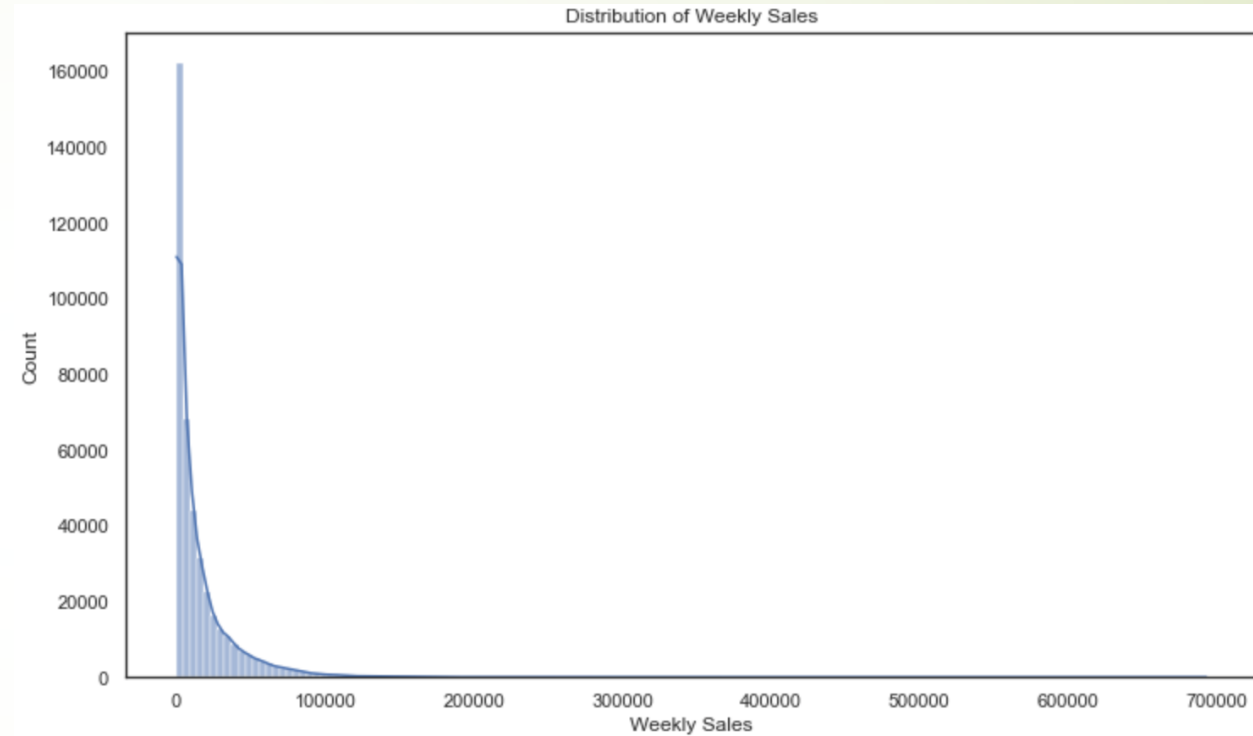
- **Data Range** - The data ranges from February 5, 2010, through November 1, 2012. This file contains anonymized information about the 45 stores, indicating the type and size of store.
- **Stores Data** – This data contains anonymized information about the 45 stores, indicating the store number, type and size of store.
- **Training Data** - This is the historical training data, which covers to 2010-02-05 to 2012-11-01. This would include store number, store department, Date, Weekly Sales of the store for the specific department and holiday or not.
- **Features data** – This data contains additional data related to the store, department, and regional activity for the given dates, such as average temperature in the region, fuel price in that region during that week, promotional markdowns, CPI index value , unemployment rate in the given week.

## Data Sample:

	Store	Dept	Date	Weekly_Sales	IsHoliday	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	Type	Size	year	week
0	1	1	2010-02-05	24924.50	False	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106	A	151315	2010	5
1	1	2	2010-02-05	50605.27	False	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106	A	151315	2010	5
2	1	3	2010-02-05	13740.12	False	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106	A	151315	2010	5
3	1	4	2010-02-05	39954.04	False	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106	A	151315	2010	5
4	1	5	2010-02-05	32229.38	False	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106	A	151315	2010	5

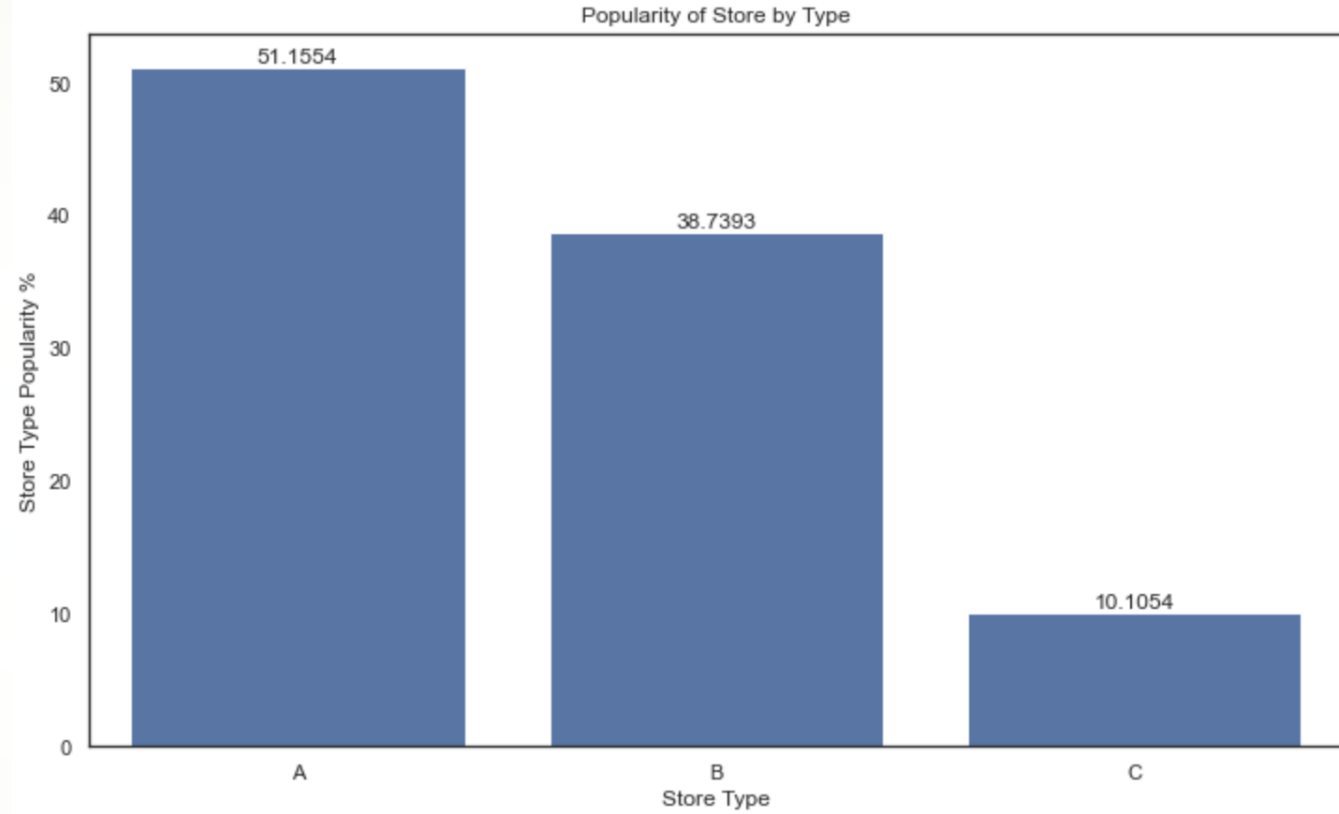
# EDA: Distribution of Weekly Sale

- The weekly sales are higher in very few weeks of the year but most of the weeks weekly sales are less than mean.
- The distribution is right skewed.
- There are weeks where weekly Sales are above 100K.
- The median weekly sale is around 8K.



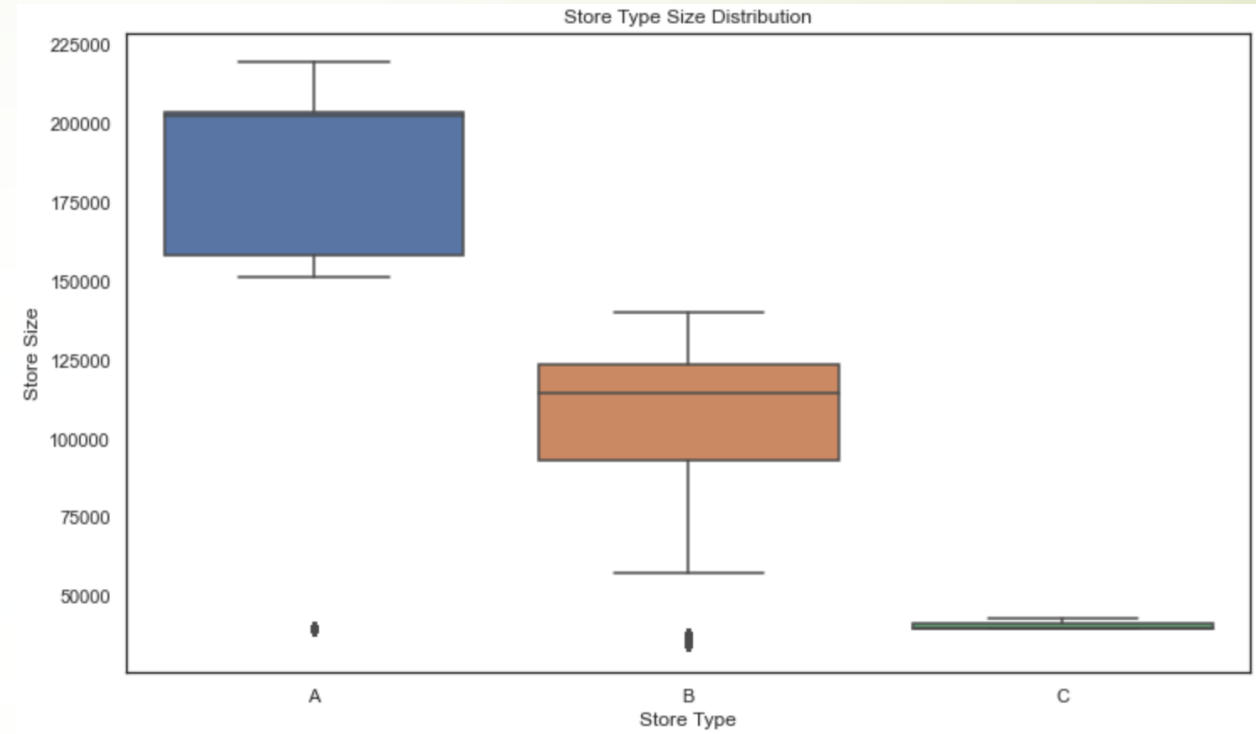
# EDA: Popularity of Store by Type

- The Store Type 'A' is more popular than store types 'B' and 'C'.
- Store Type 'C' is the least popular store among them.
- 51% of the data are from Store type A and 39% of store type B and 10% of store type C.



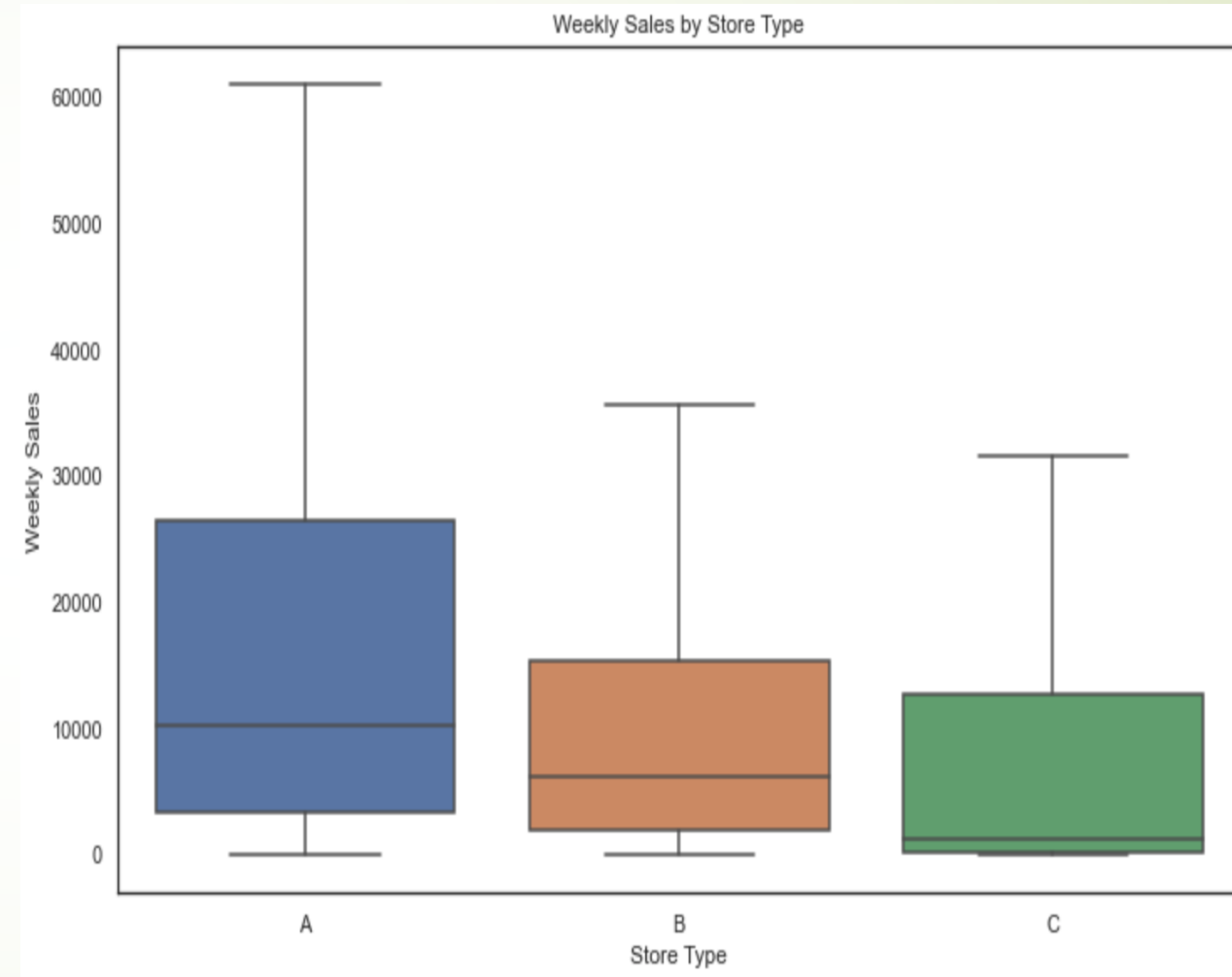
# EDA: Store Size Distribution

- The Size of the store A higher than stores B and C.
- The median size of Store A is 200K while store B is 120K and store C is less than 40K.
- The Store A is having more sales records than other two stores.



# EDA: Median Weekly Sales by Store Types

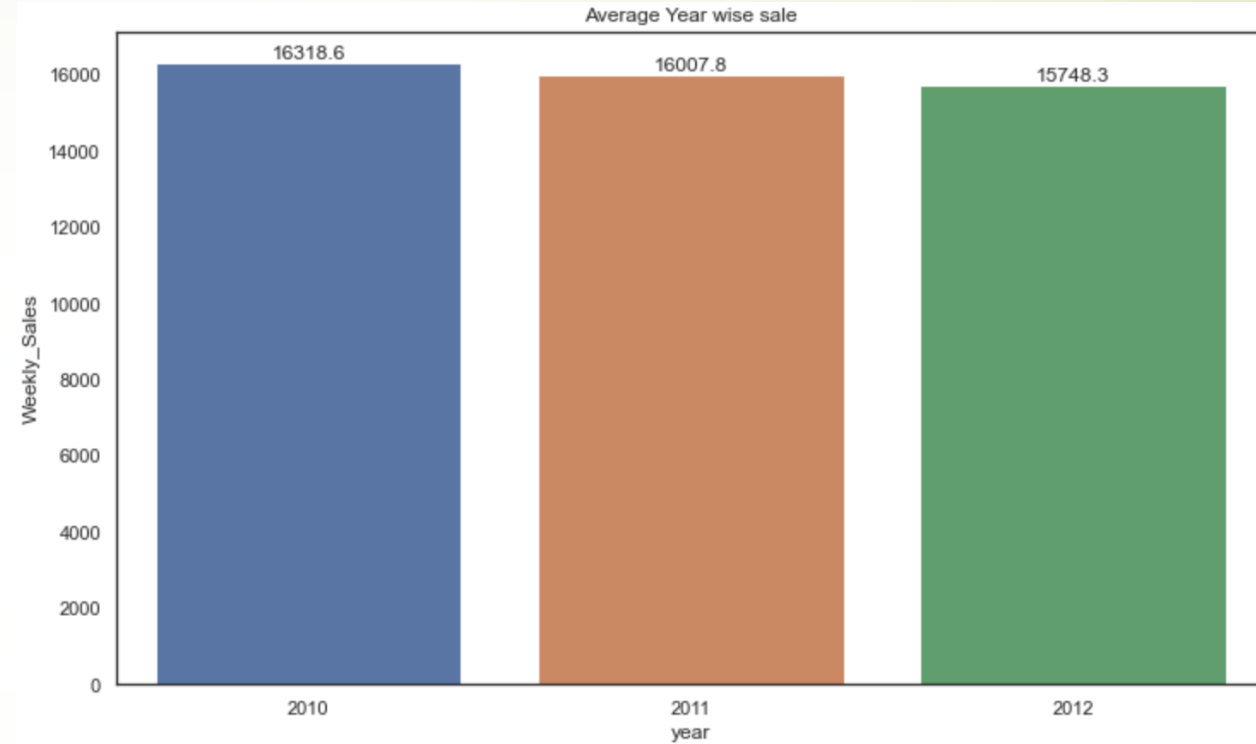
- The Store Type A is having highest Median Weekly Sale compared to other two store types.
- Store A's Median weekly sales is more than 10K
- Store A's 60 Percentile Weekly Sales is higher than 75 Percentile Weekly Sales of other two Stores B & C.
- Store Type As Max Median Weekly Sale is higher than other two stores which is above 60K.





# EDA: Average Sales by Years

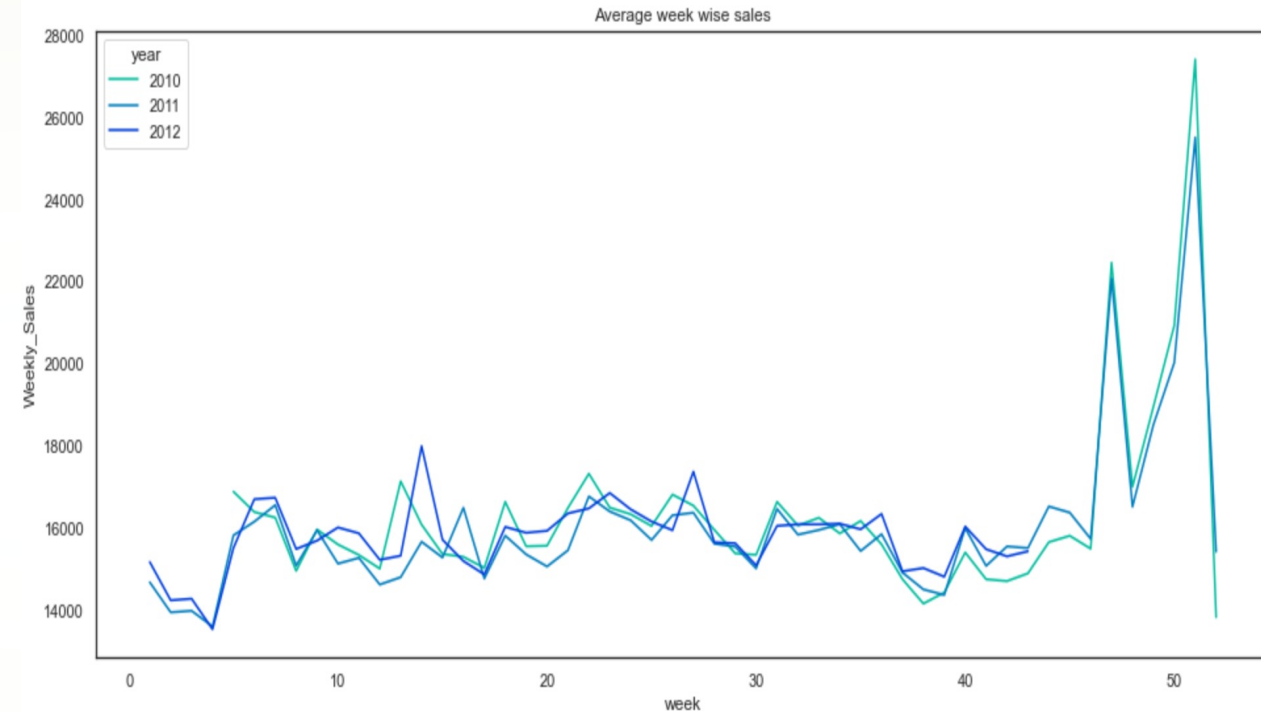
- All the Stores Average yearly sales is same across three years.





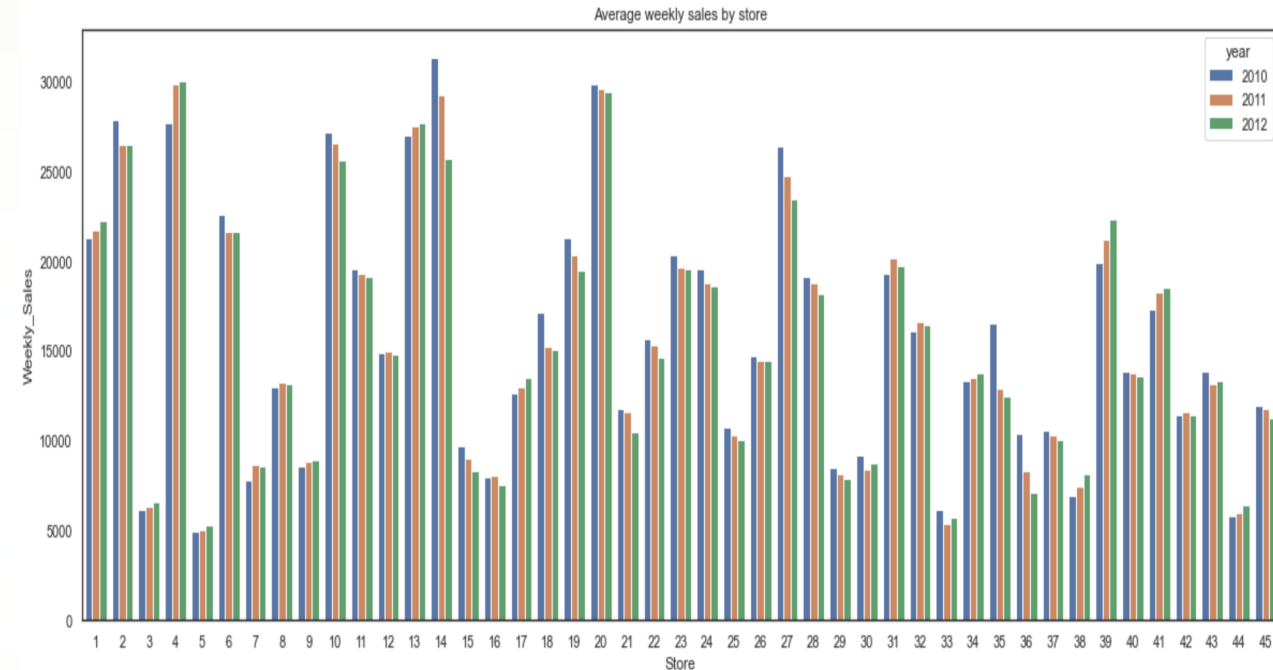
# EDA: Average Weekly Sales by Years

- The week of Thanksgiving holiday and one week before Christmas witnessed the highest sales for the years 2010, 2011 and 2012.
- In 2012 the week no. 14 recorded the highest sales as compared to other weeks of the year but that doesn't correspond to any holiday or any special event.



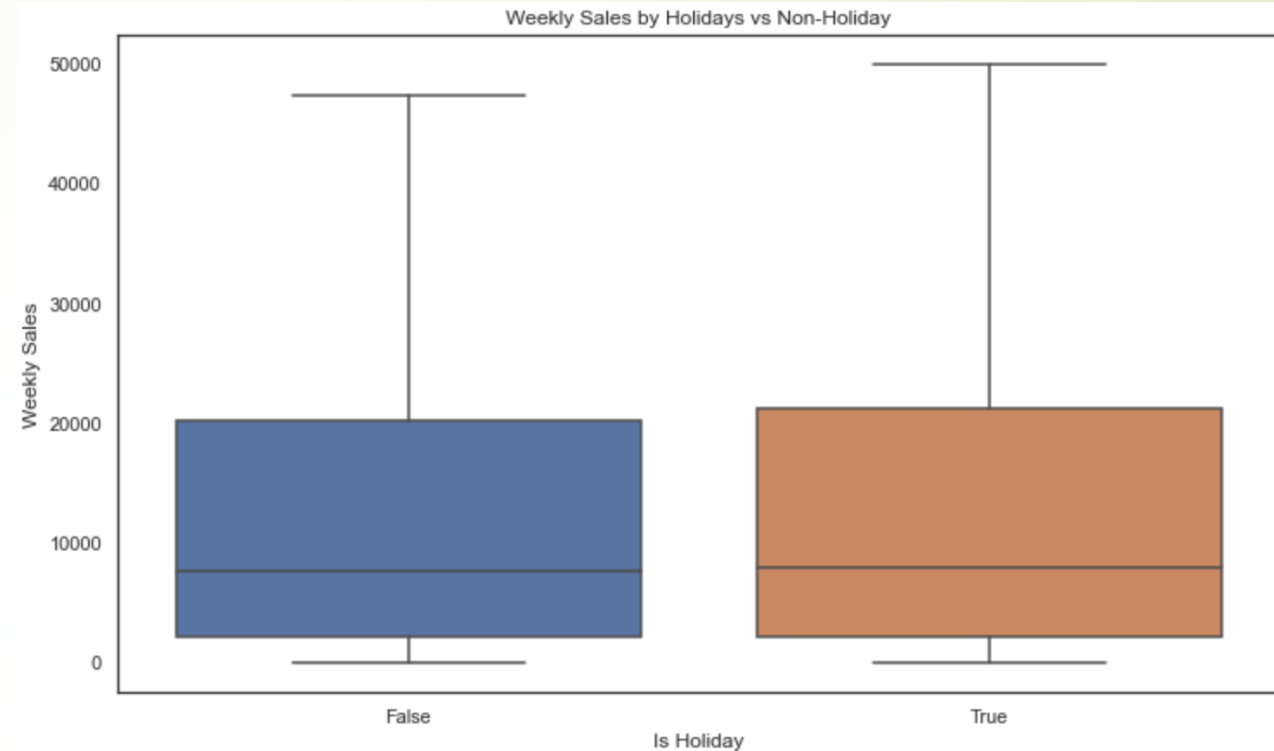
# EDA: Average Weekly Sales by Stores

- The stores 2,4,10,13,14 and 20 showed the highest sales in all the 3 years.
- There are 10+ stores having average weekly sales less than 10K.
- The overall trend of store sales over the 3 years remains the same as it depends on the type of store and its size.



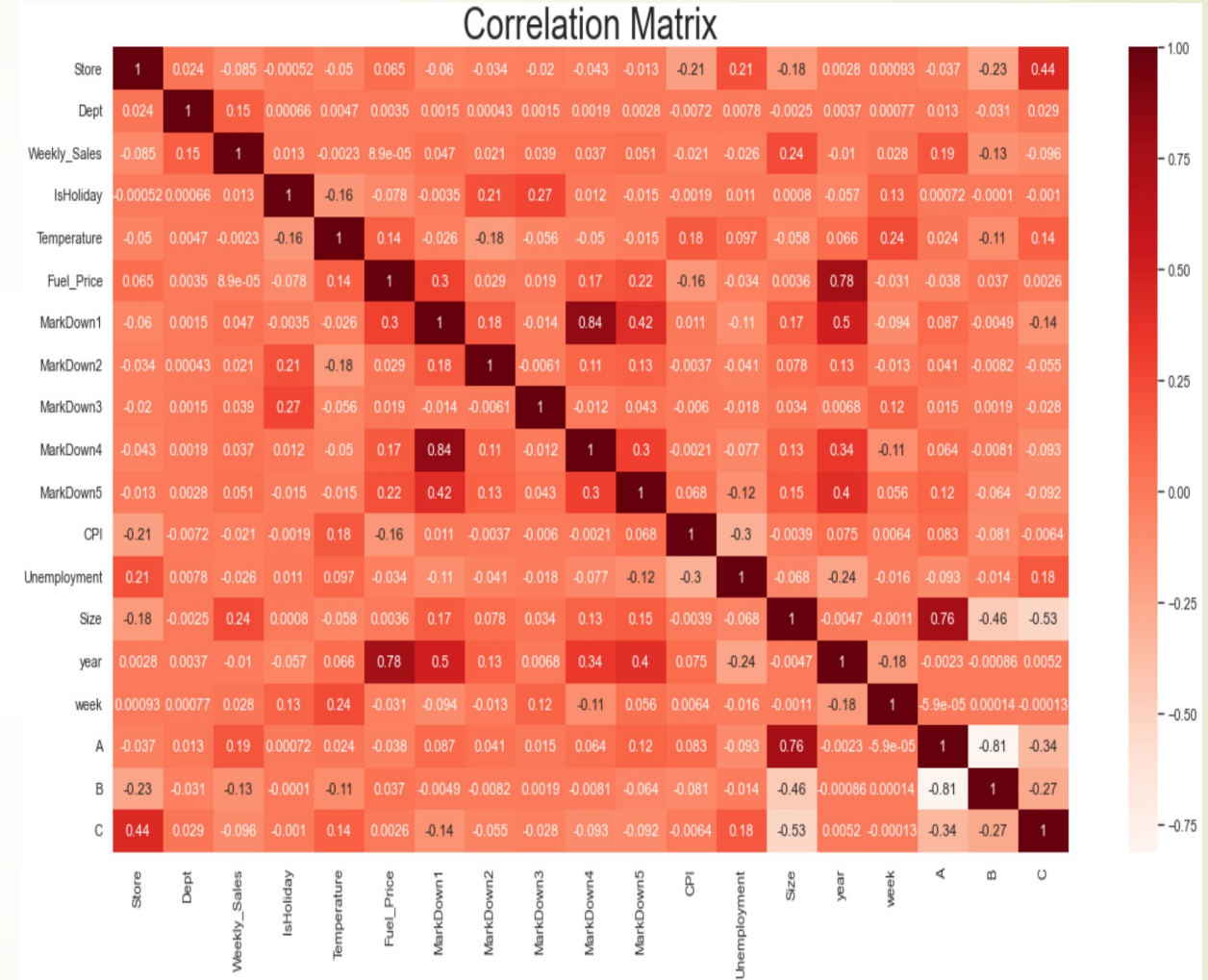
# EDA: Weekly Sales by Holiday vs non-holiday

- The median weekly sales for weeks with holiday and without holiday are almost same.
- The 75 percentile weekly sales for holidays weeks is slightly higher than non-holidays weeks.
- The max weekly sales for holidays is higher than non-holidays weeks.



# Feature Correlation:

- Correlation matrix provides range from -1 to 1 between feature pairs and with target variable. The correlation value if positive and close to 1, then it denotes these two features are having strong correlation and vice versa.
- Department, Store size and Type have moderate correlation with the weekly sales.
- Markdown1-5 have very weak correlation with the weekly sales.
- Temperature, Fuel price, CPI and Unemployment are very weakly correlated with the weekly sales.





# Modeling:

- **Why Modeling** – Modeling is very important in order to understand insights from the data, the Machine Learning models can pick up new patterns from data which humans can not recognize.
- Since the target variable which is Weekly Sales is numeric data, the regression models are best to predict continuous variables, there are several regressors out, can be used for predicting weekly sales based on certain factors. Below are top 4 ML algorithms used for training data.
- **KNN Regressor** – The KNN algorithm uses 'feature similarity' to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set.
- **Decision Tree Regressor** – It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The result is a tree with decision nodes and leaf nodes.
- **Random Forest Regressor** – The random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.
- **XGBoost Regressor** – When using gradient boosting for regression, the weak learners are regression trees, and each regression tree maps an input data point to one of its leaf's that contains a continuous score.
- **ARIMA model** – it is a very popular statistical method for time series forecasting. ARIMA stands for Auto-Regressive Integrated Moving Averages.
- **Assessment** – To baseline model, each model is evaluated using test data set for different metrics such as **Mean Absolute Error, Root Mean Squared Error and Accuracy** and the model with lowest RMSE and Highest Accuracy will be baselined.

# Result Summary:

- 4 Different algorithms are used to train the data for building Sales Forecasting.
- Each Regressor is trained and tested on train and test data. Evaluated for different metrics such as Mean Absolute Error, Root Mean Squared Error and Accuracy.
- This table shows the metrics values for each regressor.
- Random Forest Regressor and XGBoost Regressor having better accuracy, compared to KNN and Decision Tree regressors.
- ARIMA model provided best RMSE but we can not just depend on that matrix.
- XGBoost Regressor has lower RMSE compared to other four.
- XGBoost Regressor is the baseline model and will be treated as Sales Forecast Model.

Sr#	Model Name	Mean Absolute Error (MAE)	Root Mean Squared Error(RMSE)	Accuracy %
1	KNN Regressor	11340.06	18415.51	34.12%
2	Decision Tree Regressor	1874.26	4923.54	95.29%
3	Random Forest Regressor	1448.17	3576.67	97.51%
4	XGBoost Regressor	1717.73	3463.60	97.67%
5	ARIMA	447.86	686.10	NA

# References:

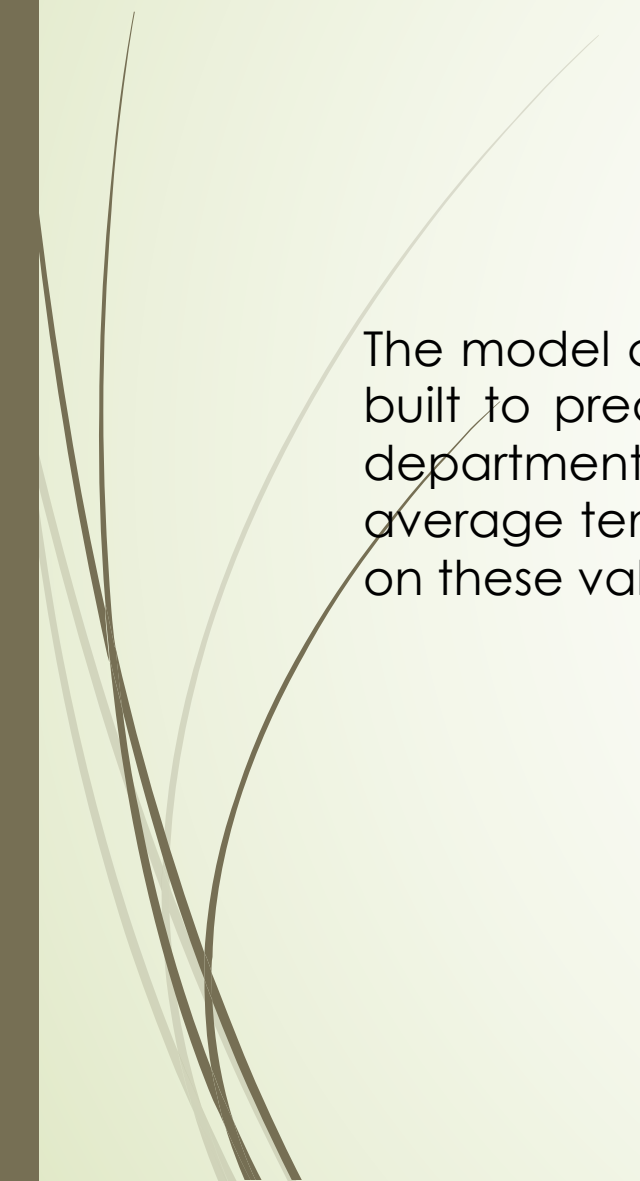
- **Data Source:** <https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/overview>
- **XGBoost Hyperparameter Tuning:** <https://towardsdatascience.com/xgboost-fine-tune-and-optimize-your-model-23d996fab663>
- **ARIMA Model Info -** <https://www.machinelearningplus.com/time-series/arma-model-time-series-forecasting-python/>
- <https://www.analyticsvidhya.com/blog/2018/08/auto-arma-time-series-modeling-python-r/>





# Conclusion

The model can be deployed to production system and simple application can be built to predict sales. The Application would accept values such store number, department number, week of the year, size of the store, is holiday in the week, average temperature, unemployment rate in that week, fuel price etc. and based on these values it would predict the sales value for that store and departments.





**Thank You !!!**