Weekly Sales Forecasting

Group - 7

Our Team:

Rohan Gupta

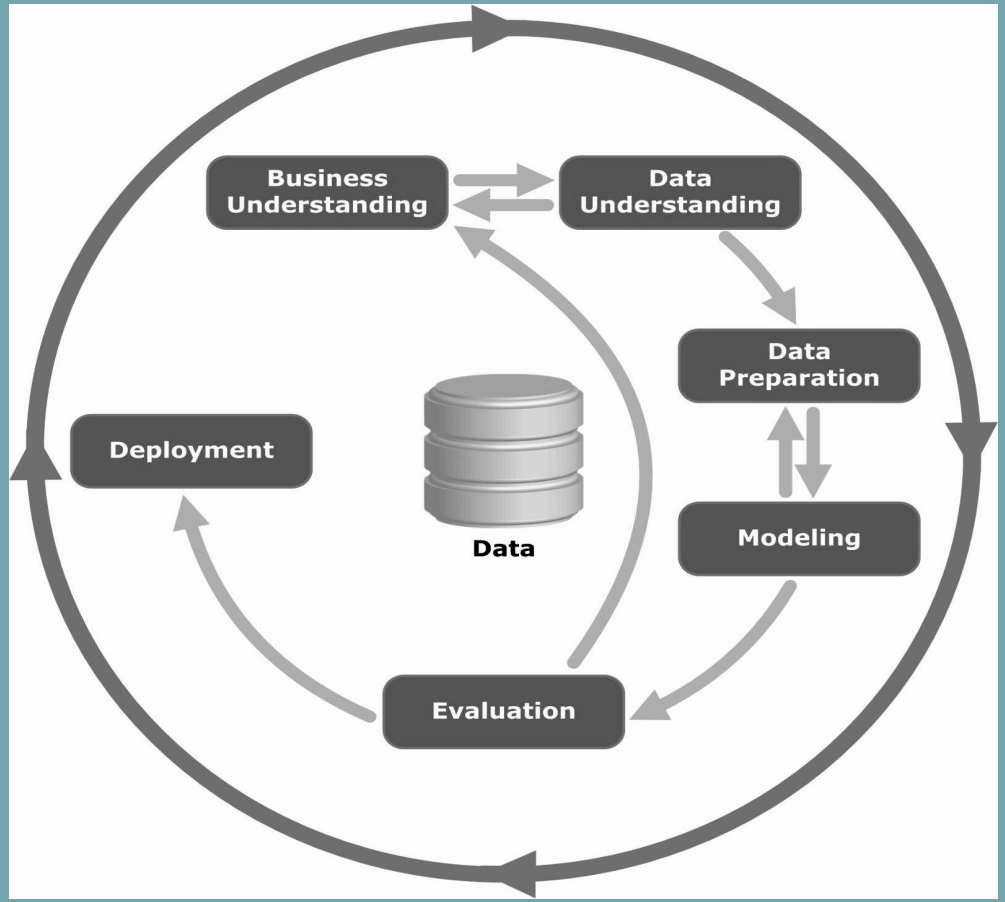
Soumi Mitra

Srijith Unni

Sushma Suryanarayana Gowda

For our assignment we are using the **“Store Sales Forecasting - Walmart”** dataset where we will be using the **historical markdown data to predict store sales.**

CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Business and Data Understanding

- Publicly available on Kaggle by Walmart
- Dataset is divided as
  1. Stores.csv
  2. Features.csv
  3. Test.csv
  4. Train.csv

The dataset contains historical sales data for 45 Walmart stores located in different regions. Each store contains many departments.

The prediction of weekly sales will also be affected with the selected holiday markdown events which are included in the dataset.

Markdowns precedes the 4 major holidays which are Super Bowl, Labor Day, Thanksgiving, and Christmas.

Weekly Sales Dataset Link

Train.csv : This is the historical training data, which covers to **2010-02-05 to 2012-11-01**.
Store - the store number, **Dept** - the department number, **Date** - the week, **Weekly_Sales** - sales for the given department in the given store, **IsHoliday** - whether the week is a special holiday week

Features.csv: Contains additional data related to the store, department, and regional activity for the given dates. It contains the following fields:
Store - the store number, **Date** - the week, **Temperature** - average temperature in the region, **Fuel_Price** - cost of fuel in the region, **MarkDown1-5** - anonymized data related to promotional markdowns that Walmart is running. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA., **CPI** - the consumer price index, **Unemployment** - the unemployment rate, **IsHoliday** - whether the week is a special holiday week

Stores.csv: This file contains anonymized information about the 45 stores, indicating the type and size of store.

Test.csv: This file is identical to train.csv but with no weekly_sales column. We will have to predict the sales for each triplet of store, department, and date in this file.

Data Preparation

- The "store" and "features" files have been combined and joined to the "train" and "test" datasets.
- The train dataset contains **421570** weekly sales records detailed by stores and departments from 02-05-2010 to 10-26-2012. The test base starts one week later and runs until 07-26-2013.
- Since the records are weekly, the "date" variable was converted to week of the year and year, as two new variables.
- Analyzing the type of variables, it is observed that all of them are in numeric format, except for the Date, Type and IsHoliday variables. The "Date" variable will not be used for training the model, using only Week and Year. "IsHoliday" was transformed to numeric binary and "Type" to ordinal numeric format. These transformation was applied to train and test datasets.

- It is observed that the holidays are in the same weeks (6, 36, 47 and 52) for the years 2010, 2011, 2012 and 2013.
- It is noticed that there are no sales records on the Laborday holiday at the test dataset, since the holiday is in September and the test runs until July.
- In order to identify not only if it is holiday, but also which holiday it is, and try to improve the sales volume prediction for these dates, the IsHoliday binary variable was transformed to:

  0 - if it is not a holiday

  1 - if the holiday is SuperBowl

  2 - if the holiday is LaborDay

  3 - if the holiday is Thanksgiving

  4 - if the holiday is Christmas

We can observe the following correlations in the correlation matrix:

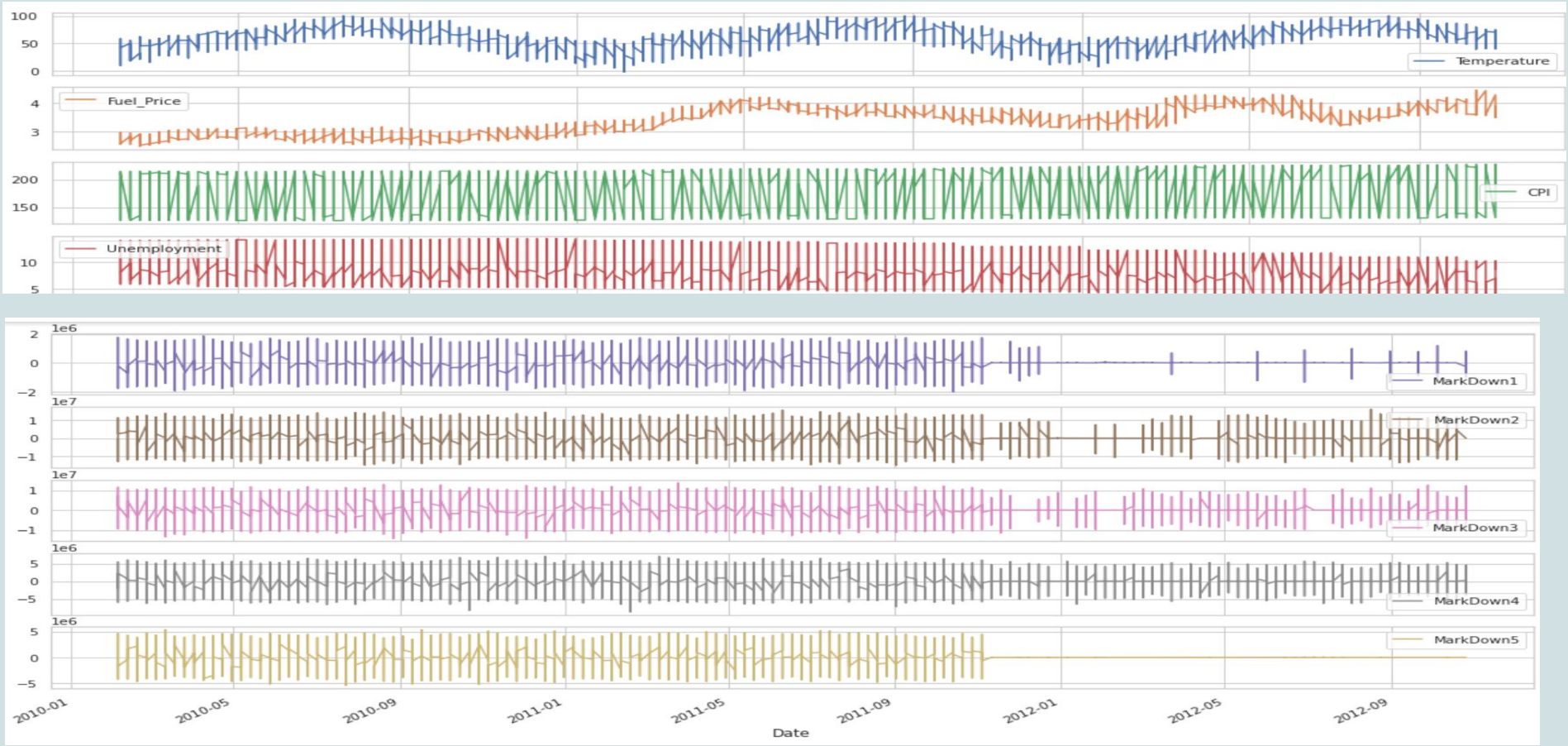
Fuel_Price with Year - Strong correlation

MarkDown 1 with MarkDown 4 - Strong correlation

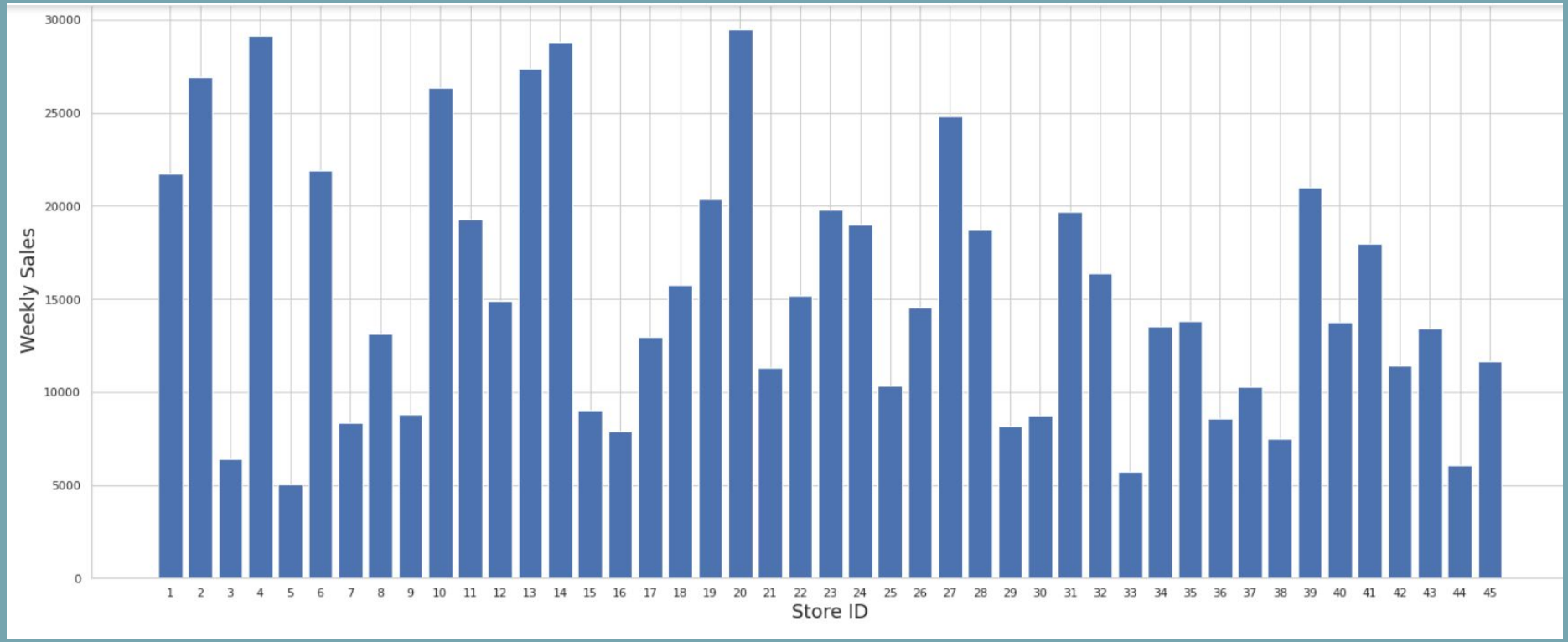
Type with Size - Strong correlation

	Store	Dept	Weekly_Sales	IsHoliday	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	Type
Store	1.000000	0.024004	-0.085195	-0.000395	-0.050097	0.065290	-0.119588	-0.035173	-0.031556	-0.009941	-0.026634	-0.211088	0.208552	-0.226
Dept	0.024004	1.000000	0.148032	0.000906	0.004437	0.003572	-0.002426	0.000290	0.001784	0.004257	0.000109	-0.007477	0.007837	-0.003
Weekly_Sales	-0.085195	0.148032	1.000000	0.012574	-0.002312	-0.000120	0.085251	0.024130	0.060385	0.045414	0.090362	-0.020921	-0.025864	0.1822
IsHoliday	-0.000395	0.000906	0.012574	1.000000	-0.145214	-0.071144	-0.080354	0.491347	0.479801	-0.068367	-0.075207	-0.001313	0.010656	0.0005
Temperature	-0.050097	0.004437	-0.002312	-0.145214	1.000000	0.143859	-0.040594	-0.323927	-0.096880	-0.063947	-0.017544	0.182112	0.096730	-0.042
Fuel_Price	0.065290	0.003572	-0.000120	-0.071144	0.143859	1.000000	0.061371	-0.220895	-0.102092	-0.044986	-0.128065	-0.164210	-0.033853	-0.029
MarkDown1	-0.119588	-0.002426	0.085251	-0.080354	-0.040594	0.061371	1.000000	0.024486	-0.108115	0.819238	0.160257	-0.055558	0.050285	0.2574
MarkDown2	-0.035173	0.000290	0.024130	0.491347	-0.323927	-0.220895	0.024486	1.000000	-0.050108	-0.007768	-0.007440	-0.039534	0.020940	0.0670
MarkDown3	-0.031556	0.001784	0.060385	0.479801	-0.096880	-0.102092	-0.108115	-0.050108	1.000000	-0.071095	-0.026467	-0.023590	0.012818	0.0375
MarkDown4	-0.009941	0.004257	0.045414	-0.068367	-0.063947	-0.044986	0.819238	-0.007768	-0.071095	1.000000	0.107792	-0.049628	0.024963	0.1089
MarkDown5	-0.026634	0.000109	0.090362	-0.075207	-0.017544	-0.128065	0.160257	-0.007440	-0.026467	0.107792	1.000000	0.060630	-0.003843	0.2588
CPI	-0.211088	-0.007477	-0.020921	-0.001313	0.182112	-0.164210	-0.055558	-0.039534	-0.023590	-0.049628	0.060630	1.000000	-0.299953	0.0658
Unemployment	0.208552	0.007837	-0.025864	0.010656	0.096730	-0.033853	0.050285	0.020940	0.012818	0.024963	-0.003843	-0.299953	1.000000	-0.148
Type	-0.226410	-0.003708	0.182242	0.000580	-0.042981	-0.029687	0.257427	0.067034	0.037560	0.108991	0.258835	0.065812	-0.148720	1.0000
Size	-0.182881	-0.002966	0.243828	0.000241	-0.058313	0.003361	0.345673	0.108827	0.048913	0.168196	0.304575	-0.003314	-0.068238	0.8115
Week	0.001031	0.000882	0.027673	0.256332	0.236276	-0.031140	-0.198076	-0.000995	0.196307	-0.218477	0.084874	0.006342	-0.015490	0.0000
Year	0.002997	0.003738	-0.010111	-0.083070	0.065814	0.779633	0.141332	-0.222109	-0.319162	0.126469	-0.128387	0.074544	-0.237161	-0.004

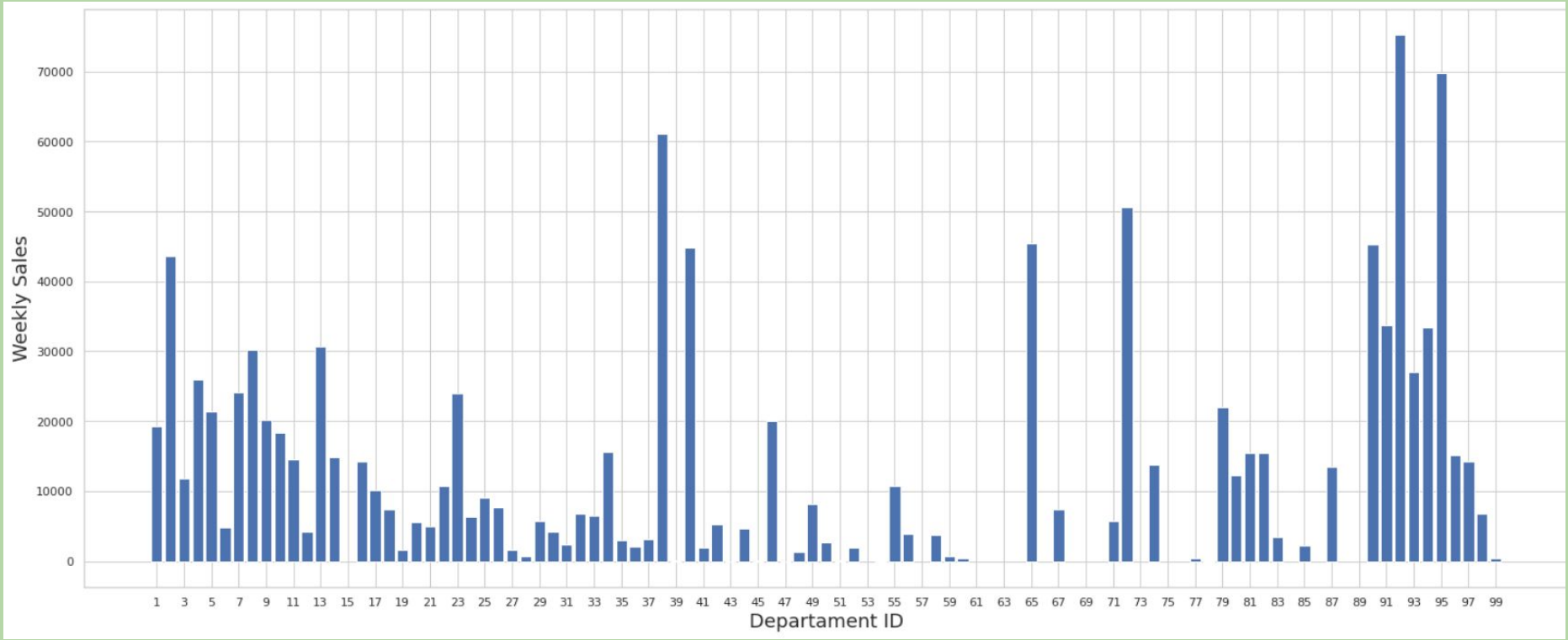
The graph below shows the variations of the features over the years

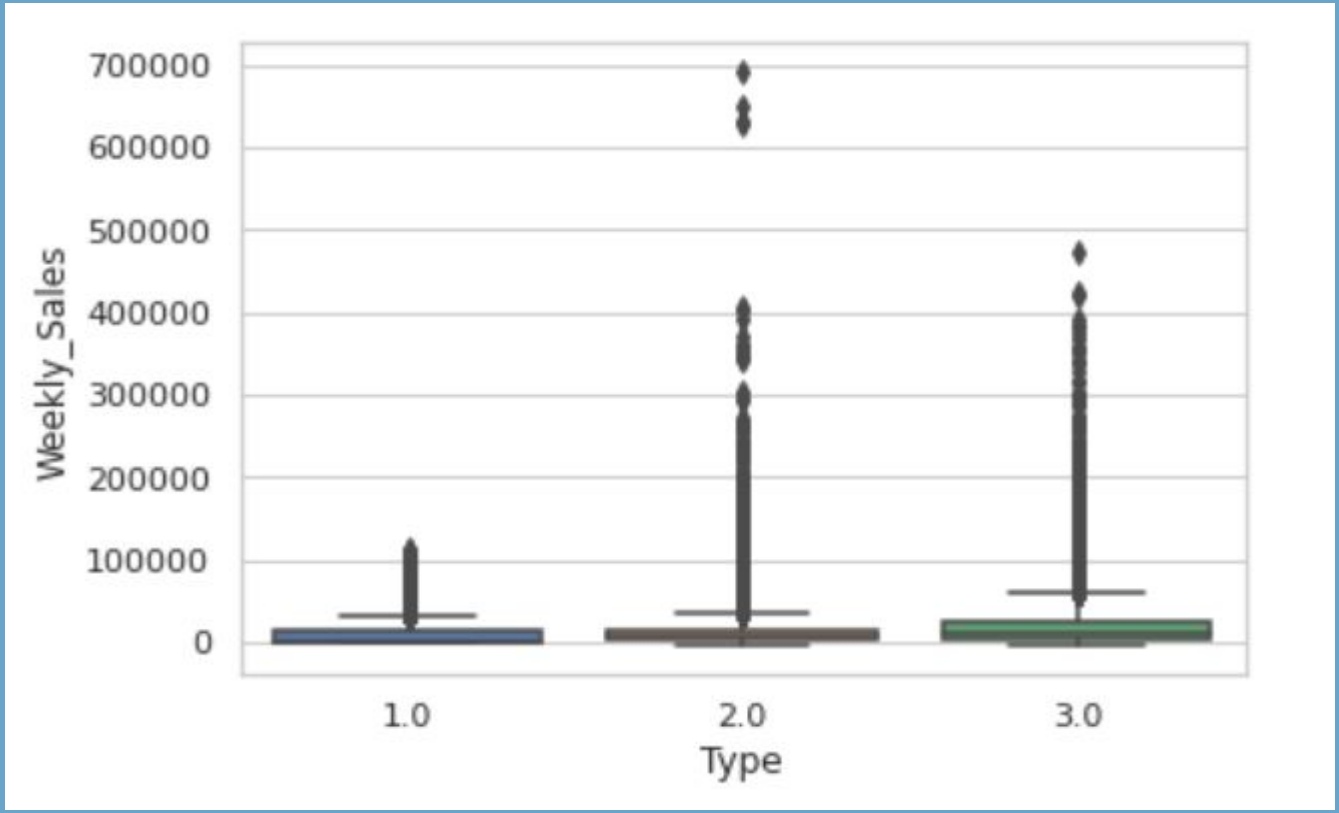


The bar chart below shows the average weekly sales with respect to the store



Avg Weekly sales with respect to the departments

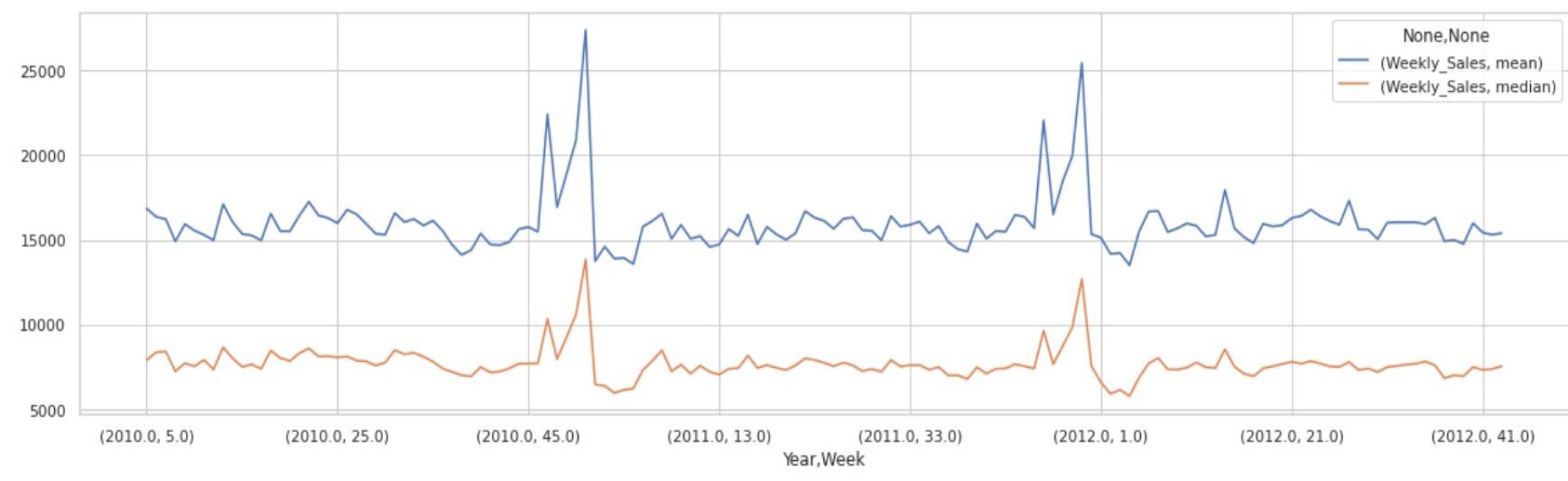




Weekly Sales Vs Type

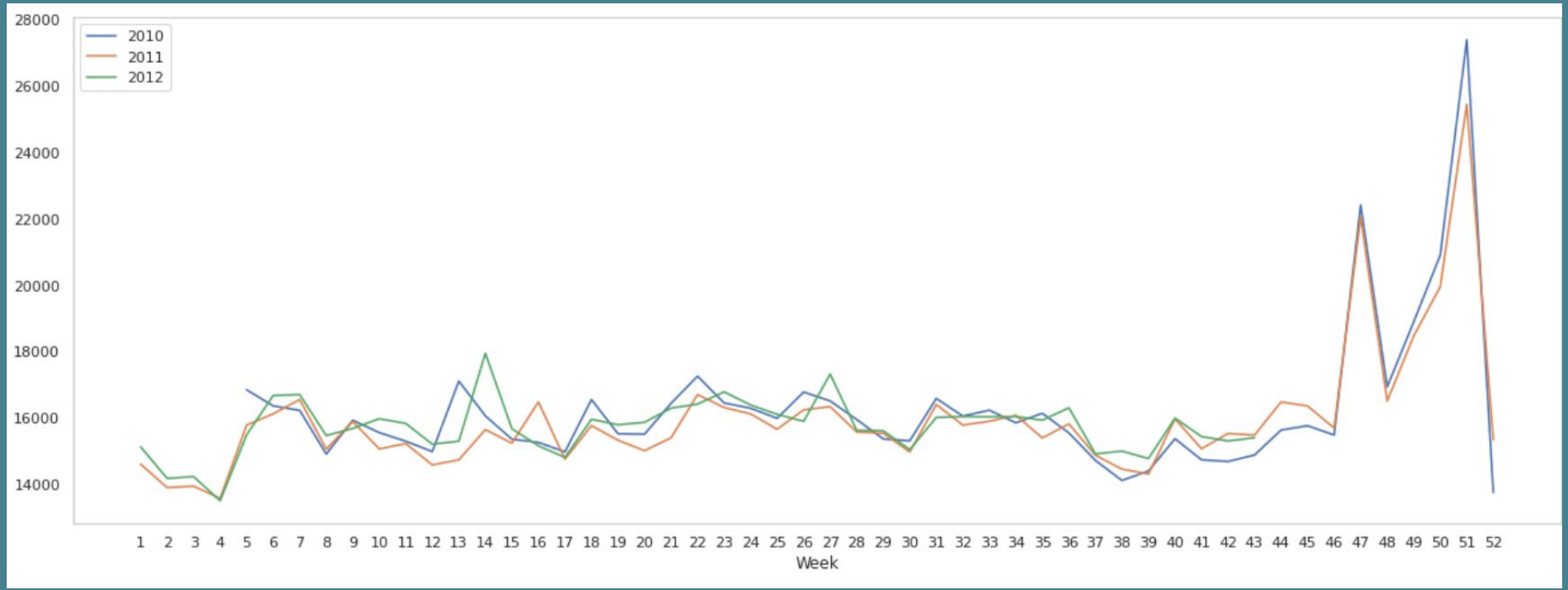
Weekly Sales mean and median over the years

We can see that both the mean and median have the seasonal pattern at the end of each year



Weekly sales variations for the years 2010, 2011 and 2012

We can see that there isn't a lot of variation and the spike at the end of year is common for all the years(the holiday spike)



We performed Multivariate Imputation on both the train and test datasets for the following reasons:

1. A considerable amount of data is missing from multiple variables.
2. Features are not highly correlated and missing data is most probably not MCAR, so univariate imputation was ruled out.

We dropped the date field since we used week of the year instead.

Data Modelling

We divided our training dataset into a train and test set by splitting the training dataset in 80:20 ratio respectively.

This is the input for our regression. In order to choose the best model for our prediction, we measure the performance by comparing 8 different training models.

We measure the performance by calculating the Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE) and R Square to check for the robustness.

Model Analysis

We have analyzed the data using the following models:

- **Linear Regression -** linear approach to build a relationship between one dependent and one independent variable
- **KNN -** K-nearest neighbor algorithm is used for regression and classification problems. It uses data and identify new data based on distance function
- **Decision Tree Regressor - In the context of a tree structure, a decision tree constructs regression or classification models. It incrementally breaks down a dataset into smaller and smaller subsets while also developing an associated decision tree. A tree with decision nodes and leaf nodes is the end product.**
- **Random Forest Regressor - A supervised learning algorithm for regression that employs the ensemble learning process. During preparation, a Random Forest constructs multiple decision trees and outputs the mean of the classes as the prediction of all the trees.**

Model Analysis(contd.)

- Extra Tree Regressor - This class implements a meta estimator that uses averaging to increase predictive precision and control over-fitting by fitting a number of randomized decision trees (a.k.a. extra-trees) on different sub-samples of the dataset. The forest's total number of trees.
- XGBoost - XGBoost is a high-speed and high-performance implementation of gradient boosted decision trees.
- Ridge - Ridge regression is a model tuning technique that can be used to analyze data with multicollinearity. L2 regularization is achieved using this approach. Where there is a problem with multicollinearity, least-squares are unbiased, and variances are high, the expected values are far from the actual values.
- Lasso - A regression analysis approach that includes both variable selection and regularization in order to improve the statistical model's prediction accuracy and interpretability.

Evaluation

The evaluation of the results is done by the calculation of the error rates for each of the models to determine which is the perfect model to implement the forecasting of the “Weekly Sales Prediction”.

	Model Comparison by Error Rate				
	Name of the Model	*MAE*	*MSE*	*RMSE*	*R^2*
1	Linear Regression	14580.58	479344956.87	21893.9	0.084
2	KNN	14580.17	479344093.44	21893.9	0.084
3	Decision Tree Regressor	2013.40	27530117.92	5246.91	0.947
4	Random Forest Regressor	1492.73	16024665.52	4003.08	0.969
5	**Extra Tree Regressor**	**1498.697**	**16709148.73**	**4087.68**	**0.968**
6	XGBoost	6885.18	138284138.10	11759.42	0.735
7	Ridge	14580.28	479344957.07	21893.94	0.084
8	Lasso	14580.17	479344093.44	21893.9	0.084

Deployment

As per the RMSE values we have concluded the best suited model for the prediction of the “Weekly Sales” is the Extra Tree Regressor model, since it has the lowest error rate of the prediction. Hence, we have deployed the Extra Tree Regressor model for weekly sales predictions. The final weekly_sales prediction values can be found in predictions.csv file.

Thank You!