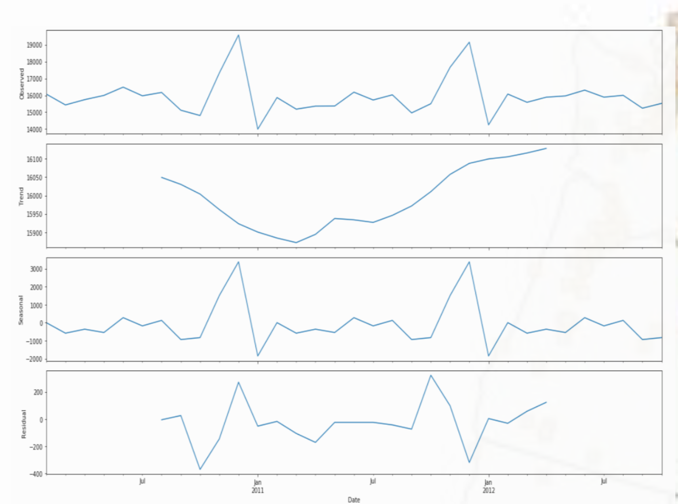
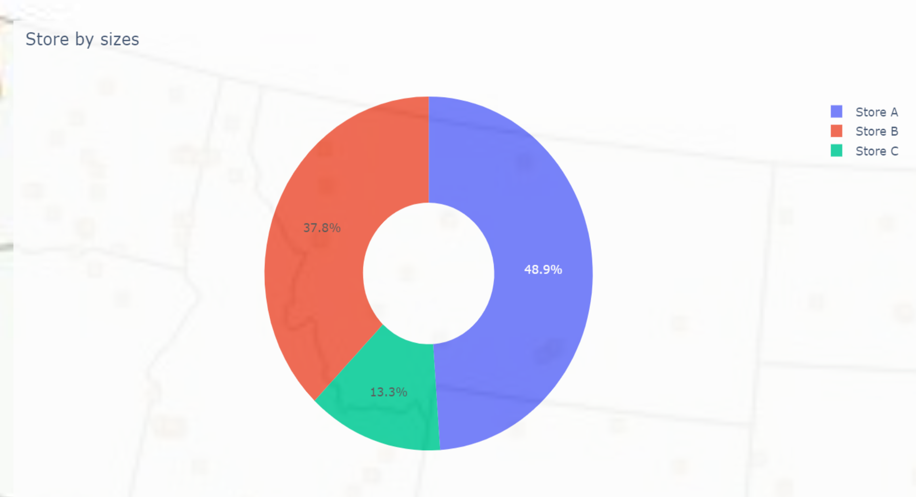
Walmart Store Sales Prediction

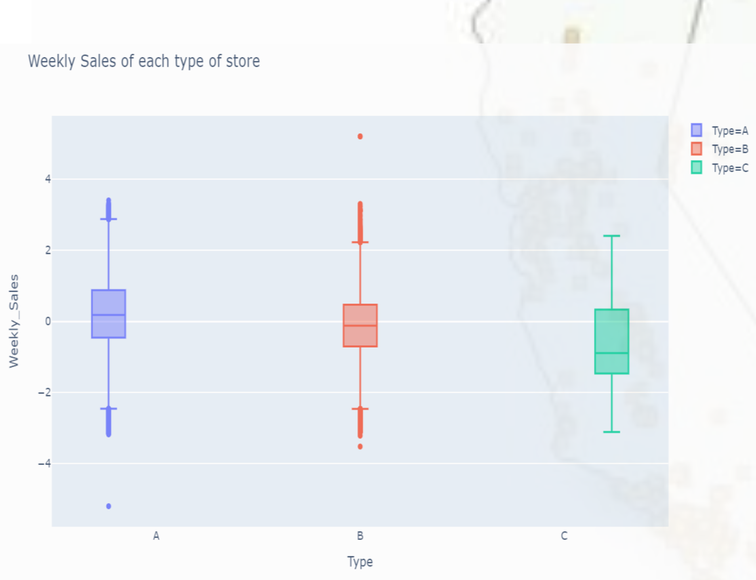
Business Problem:
Predicting sales for a store is very important to estimate the quantity for each product, to avoid overstocking or understocking. Our aim is to apply Machine learning algorithms on Walmart's historical sales data for 45 stores present, to predict department wide sales for each data.

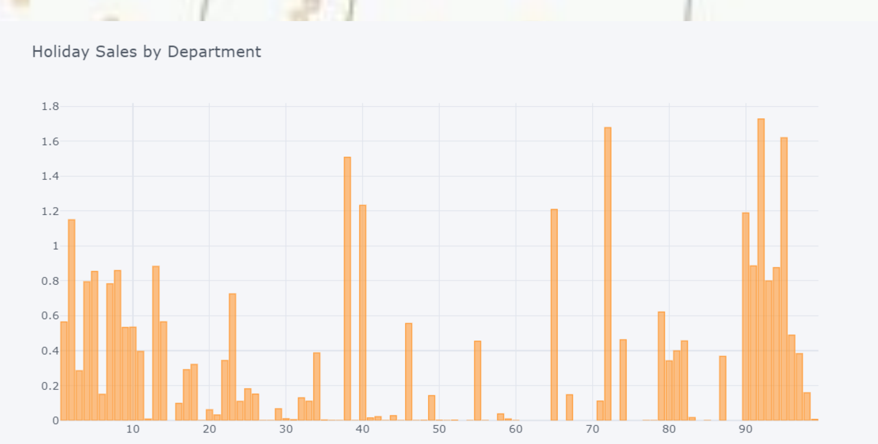
Data Insights:
After pre-processing the raw data and performing feature engineering on the cleaned data, we implemented multiple Machine Learning models and performed detailed analysis on the results to obtain the most suitable model for sales prediction.

Exploratory Data Analysis:
Store A is the largest in size and also implies to have the highest weekly sales

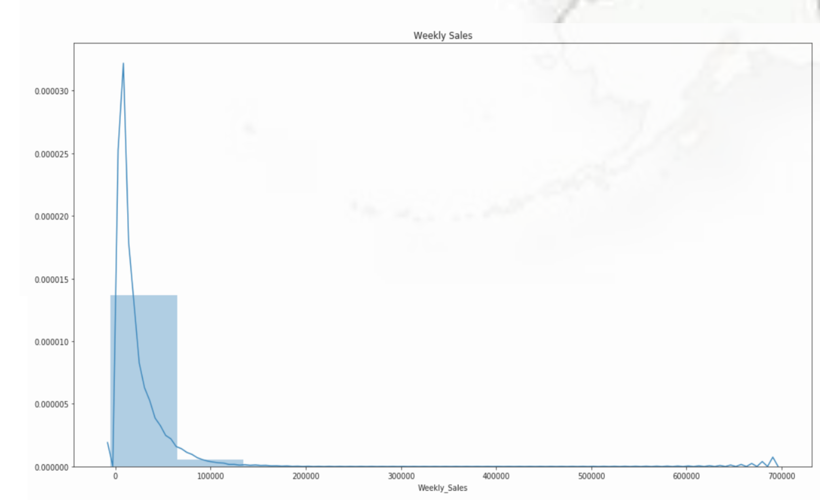


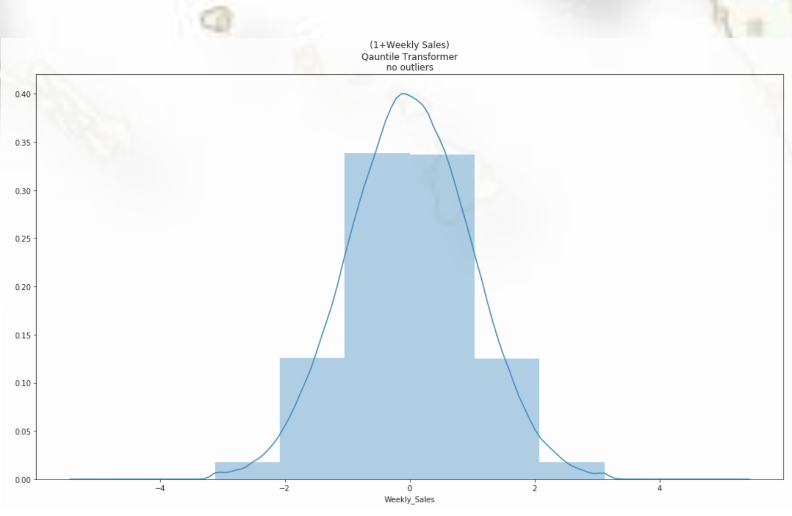




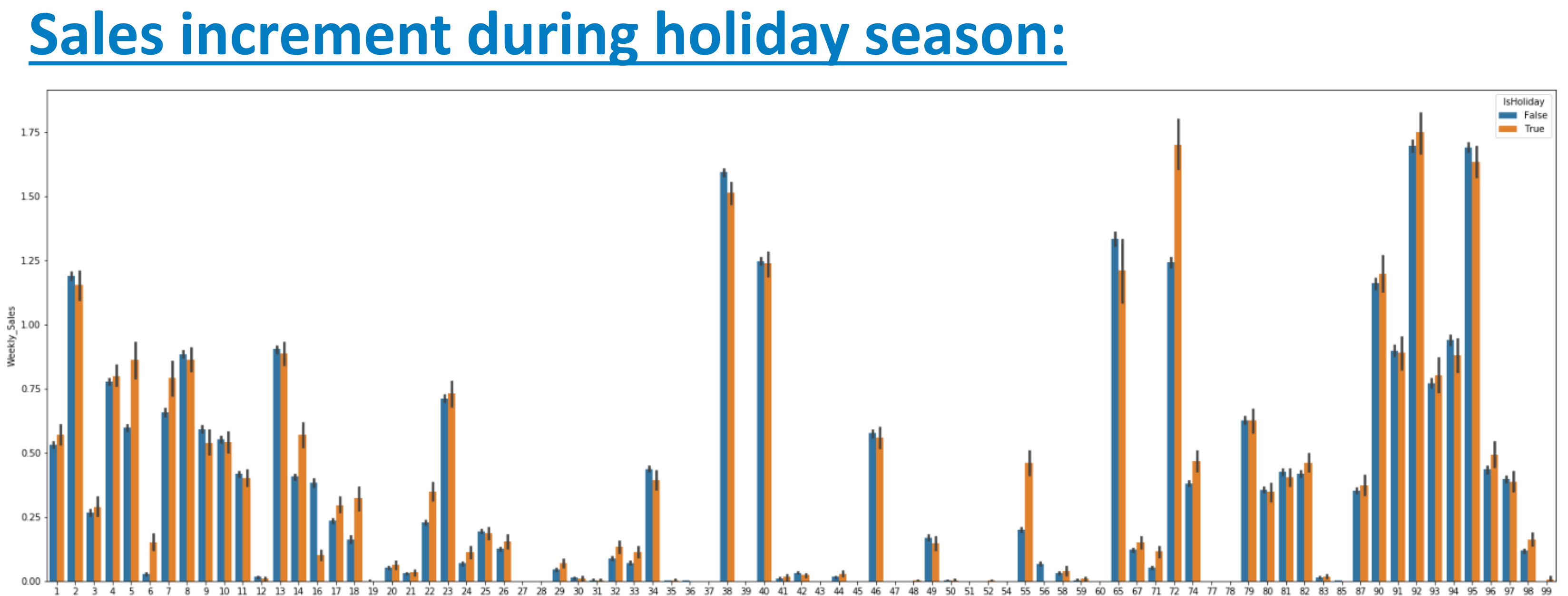


Data Transformation:
Given that the original distribution was skewed, we transformed the data using Quantile Transformer to remove outliers and negative values, for a normal distribution





Sales increment during holiday season:



We compared the department wide sales during holiday and non-holiday, and further inferred that about 15 out of 81 departments showed an increase of 55% sales during the holiday season

	Dept	Weekly_Sales	Weekly_Holiday_Sales	Percent_Increase
5	6	0.02670	0.15174	468.314607
16	18	0.16254	0.32384	99.237111
22	24	0.06879	0.11225	63.177787
31	33	0.07078	0.11321	59.946313
42	44	0.01537	0.02874	86.987638
46	48	0.00040	0.00115	187.500000
50	52	0.00008	0.00219	2637.500000
52	55	0.19950	0.45933	130.240602
55	59	0.00457	0.00884	93.435449
59	71	0.05338	0.11438	114.275009
80	99	0.00009	0.00663	7266.666667

Data Models:
We leveraged KNN, SVM, Random Forest and Linear Regression to gain insights

Model Evaluation:
The primary parameter used to compare model performance is Weighted Mean Absolute Error

$$\text{WMAE} = \frac{1}{\sum w_i} \sum_{i=1}^{n} w_i |y_i - \hat{y}_i|$$

where

- n is the number of rows
- $\hat{y}_i$ is the predicted sales
- y_i is the actual sales
- w_i are weights. w = 5 if the week is a holiday week, 1 otherwise

KNN: The prediction of our query instance is based on the simple majority of the category of nearest neighbors, and for our dataset, the prediction is heavily influenced by the size of the store.

SVM: We implemented this as a black box algorithm to find the most optimal decision boundary that maximizes the distance from the nearest data points of all classes.

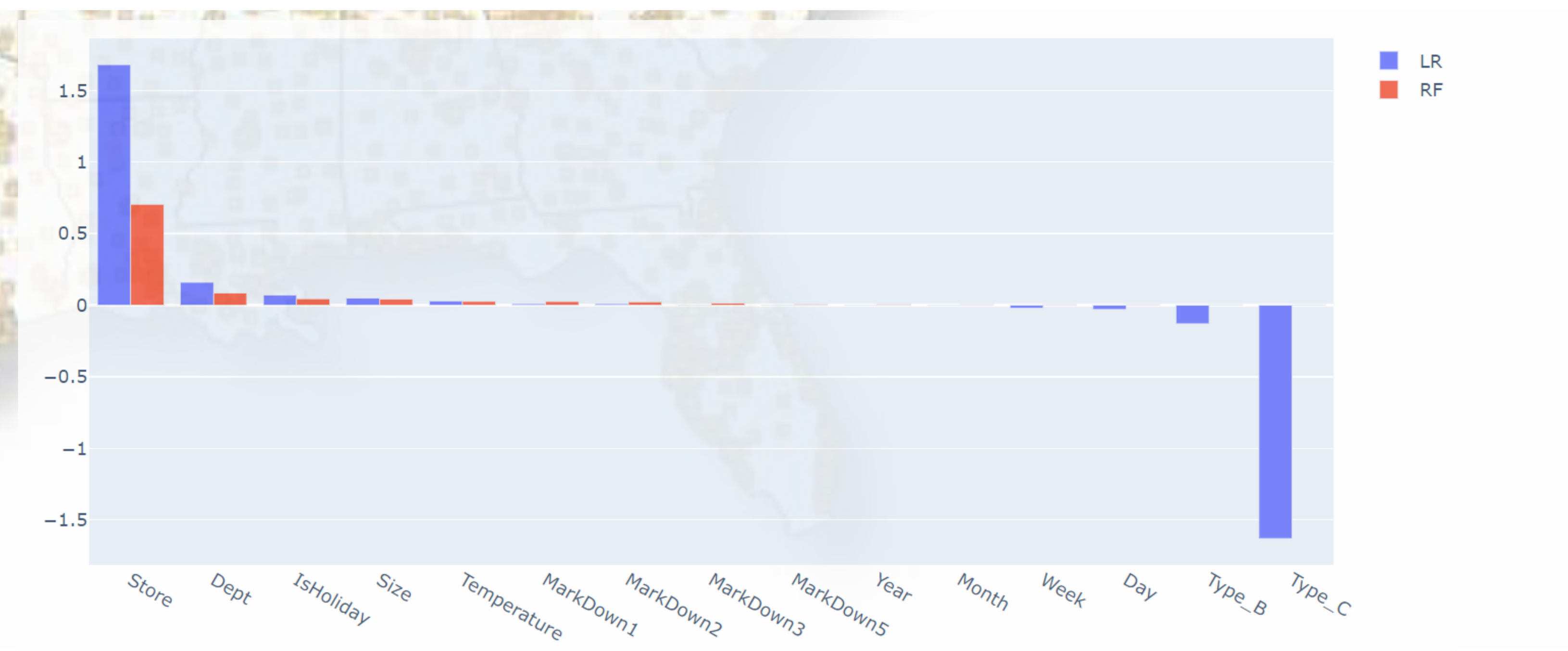
Random Forest:

	importance
Dept	0.639551
Size	0.208534
Store	0.057557
Week	0.031845
Temperature	0.017021
Type_B	0.013976
Day	0.010409
Month	0.005919
MarkDown3	0.005099
Type_C	0.002499
Year	0.001753
IsHoliday	0.001623
MarkDown5	0.001470
MarkDown2	0.001374
MarkDown1	0.001368

Linear Regression:

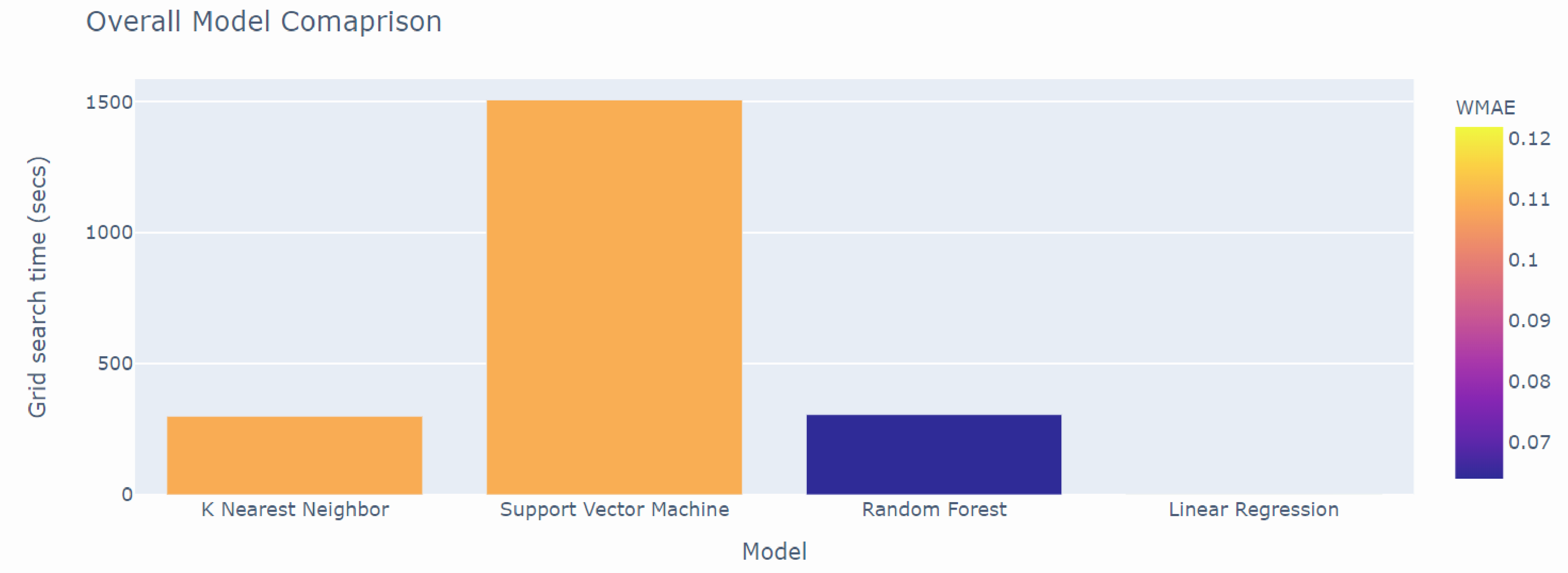
	column	Coefficients
11	Week	1.68
3	Size	0.16
1	Dept	0.07
14	Type_C	0.05
7	MarkDown3	0.03
5	MarkDown1	0.01
8	MarkDown5	0.01
4	Temperature	0.00
13	Type_B	0.00
6	MarkDown2	-0.00
2	IsHoliday	-0.00
9	Year	-0.02
0	Store	-0.03
12	Day	-0.13
10	Month	-1.63

The feature comparison between Random Forest and Linear Regression indicates Store and Department to be the most important attributes in our prediction model.



Model Comparison:
From the analysis we drew based on the weighted mean absolute error and the implementation time, we observed that Random Forest gave the most optimum results

	Model	Initial model Time (secs)	Grid search time (secs)	WMAE
0	K Nearest Neighbor	8.800	300.0	0.1093
1	Support Vector Machine	180.310	1507.0	0.1096
2	Random Forest	2.620	306.0	0.0640
3	Linear Regression	0.039	0.8	0.1220



Ritika Shetty | Shweta Sathisan | Yash Pasar
IST 707 - Data Mining & Machine Learning