

**CAPSTONE PROJECT**

**RETAIL DATA ANALYTICS**

**FINAL PROJECT REPORT**

**BY  
SHWETA GUPTA**

**INTRODUCTION**

|  |
| --- |
| **DEFINING PROBLEM STATEMENT**   * The client, a leading Retail Chain would like to take advantage of its data to take better informed strategic decisions. * We are provided with historical sales data for their 45 retail stores located in different regions. Each store contains a number of departments.3 datasets are provided – * Stores – Information about type & size of store, * Features - Data related to the store, department, and regional activity * Sales - Historical sales data * The company also runs several promotional markdown events throughout the year. * These markdowns precede prominent holidays, the four largest of which are the Super Bowl, Labor Day, Thanksgiving, and Christmas. * The weeks including these holidays are weighted five times higher in the evaluation than non-holiday weeks. * The task at hand is to Predict the department-wide sales for each store for the following year while analyzing the impact of markdowns on sales. |
| **NEED OF THE STUDY**   * **Predict department-wide sales** for each store for next year. * **Model the effects of Markdowns** on Sales during holiday weeks. * **Provide recommended actions** based on the insights drawn, with prioritization placed on largest business impact. |
| **UNDERSTANDING BUSINESS/SOCIAL OPPORTUNITY** |

* Help boost consumer spends on retail, during holidays, using the right range of markdowns in the right departments
  + Markdowns are strategic decisions which impact the bottom line, therefore if the impact of markdowns on Sales is understood through modeling historical data, then the **range of markdowns can be optimized to generate higher sales** while keeping the losses at the minimum possible threshold.
* Also, **if the stores and department which have the highest business impact (sales and/or sensitivity to markdowns) are identified through modelling, then they can be prioritized** in terms of marketing decisions such as allocation of advertising spends and/or higher/lower markdowns.

**DATA REPORT**

3 Data sets are shared with us and described as under :-

**STORES**

Information about 45 stores, indicating the type and size of store

**FEATURES**

Contains data related to the store, department, and regional activity for the given dates.

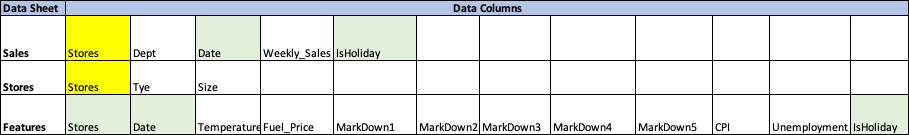
* Store - the store number
* Date- from 2010 to 2013
* Temperature - average temperature in the region
* Fuel\_Price - cost of fuel in the region
* MarkDown1-5 - anonymized data related to promotional markdowns. MarkDown data is only available after Nov 2011, and is not available for all stores all the 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

**SALES**

Historical sales data, which covers from **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

**KEY ACTIONS ON DATASET**

* Merged the 3 datasets to arrive at a single dataset using the common variables as indicated in the table below:
* Changed the data type of variables as required - Store, Dept variables are changed to factor data type. Date is changed to Date datatype.
* Created a new variable **Weight** based on IsHoliday variable , where if it’s a holiday weight is 5 else weight is 1.This is done since in retail industry , consumer spending is higher during holiday season and therefore we needed to capture the importance of holidays on Sales.

**THE FINAL MERGED DATA SET VARIABLES SUMMARY REPORT**

'data.frame': 421570 obs. of 17 variables:

$ Store : Factor w/ 45 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...

$ Date : Date, format: "2011-04-01" "2011-04-01" "2011-04-01" "2011-04-01" .$ IsHoliday : logi FALSE FALSE FALSE FALSE FALSE FALSE ...

$ Dept : Factor w/ 81 levels "1","2","3","4",..: 48 25 67 33 56 29 7 70 8

$ Weekly\_Sales: num 13168 5947 28545 9950 317 ...

$ Type : Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 1 1 1 1 ...

$ Size : int 151315 151315 151315 151315 151315 151315 151315 151315

$ Temperature : num 59.2 59.2 59.2 59.2 59.2 ...

$ Fuel\_Price : num 3.52 3.52 3.52 3.52 3.52 ...

$ MarkDown1 : num NA NA NA NA NA NA NA NA NA NA ...

$ MarkDown2 : num NA NA NA NA NA NA NA NA NA NA ...

$ MarkDown3 : num NA NA NA NA NA NA NA NA NA NA ...

$ MarkDown4 : num NA NA NA NA NA NA NA NA NA NA ...

$ MarkDown5 : num NA NA NA NA NA NA NA NA NA NA ...

$ CPI : num 215 215 215 215 215 ...

$ Unemployment: num 7.68 7.68 7.68 7.68 7.68 ...

$ weight : num 1 1 1 1 1 1 1 1 1 1 ...

**EXPLORATORY DATA ANALYSIS**

**UNIVARIATE ANALYSIS**

**FIVE POINT SUMMARY**

**Store Date IsHoliday Dept**

13 : 10474 Min. :2010-02-05 Mode :logical 1 : 6435

10 : 10315 1st Qu.:2010-10-08 FALSE:391909 2 : 6435

4 : 10272 Median :2011-06-17 TRUE :29661 3 : 6435

1 : 10244 Mean :2011-06-18 4 : 6435

2 : 10238 3rd Qu.:2012-02-24 7 : 6435

24 : 10228 Max. :2012-10-26 8 : 6435

(Other):359799 (Other):382960

**Weekly\_Sales Type Size Temperature**

Min. : -4989 A:215478 Min. : 34875 Min. : -2.06

1st Qu.: 2080 B:163495 1st Qu.: 93638 1st Qu.: 46.68

Median : 7612 C: 42597 Median :140167 Median : 62.09

Mean : 15981 Mean :136728 Mean : 60.09

3rd Qu.: 20206 3rd Qu.:202505 3rd Qu.: 74.28

Max. :693099 Max. :219622 Max. :100.14

**Fuel\_Price MarkDown1 MarkDown2 MarkDown3**

Min. :2.472 Min. : 0.27 Min. : -265.8 Min. : -29.10

1st Qu.:2.933 1st Qu.: 2240.27 1st Qu.: 41.6 1st Qu.: 5.08

Median :3.452 Median : 5347.45 Median : 192.0 Median : 24.60

Mean :3.361 Mean : 7246.42 Mean : 3334.6 Mean : 1439.42

3rd Qu.:3.738 3rd Qu.: 9210.90 3rd Qu.: 1926.9 3rd Qu.: 103.99

Max. :4.468 Max. :88646.76 Max. :104519.5 Max. :141630.61

NA's :270889 NA's :310322 NA's :284479

**MarkDown4 MarkDown5 CPI Unemployment**

Min. : 0.22 Min. : 135.2 Min. :126.1 Min. : 3.879

1st Qu.: 504.22 1st Qu.: 1878.4 1st Qu.:132.0 1st Qu.: 6.891

Median : 1481.31 Median : 3359.4 Median :182.3 Median : 7.866

Mean : 3383.17 Mean : 4629.0 Mean :171.2 Mean : 7.960

3rd Qu.: 3595.04 3rd Qu.: 5563.8 3rd Qu.:212.4 3rd Qu.: 8.572

Max. :67474.85 Max. :108519.3 Max. :227.2 Max. :14.313

NA's :286603 NA's :270138

**weight**

Min. :1.000

1st Qu.:1.000

Median :1.000

Mean :1.281

3rd Qu.:1.000

Max. :5.000

sum(is.na(sales\_stores\_features))

[1] 1422431

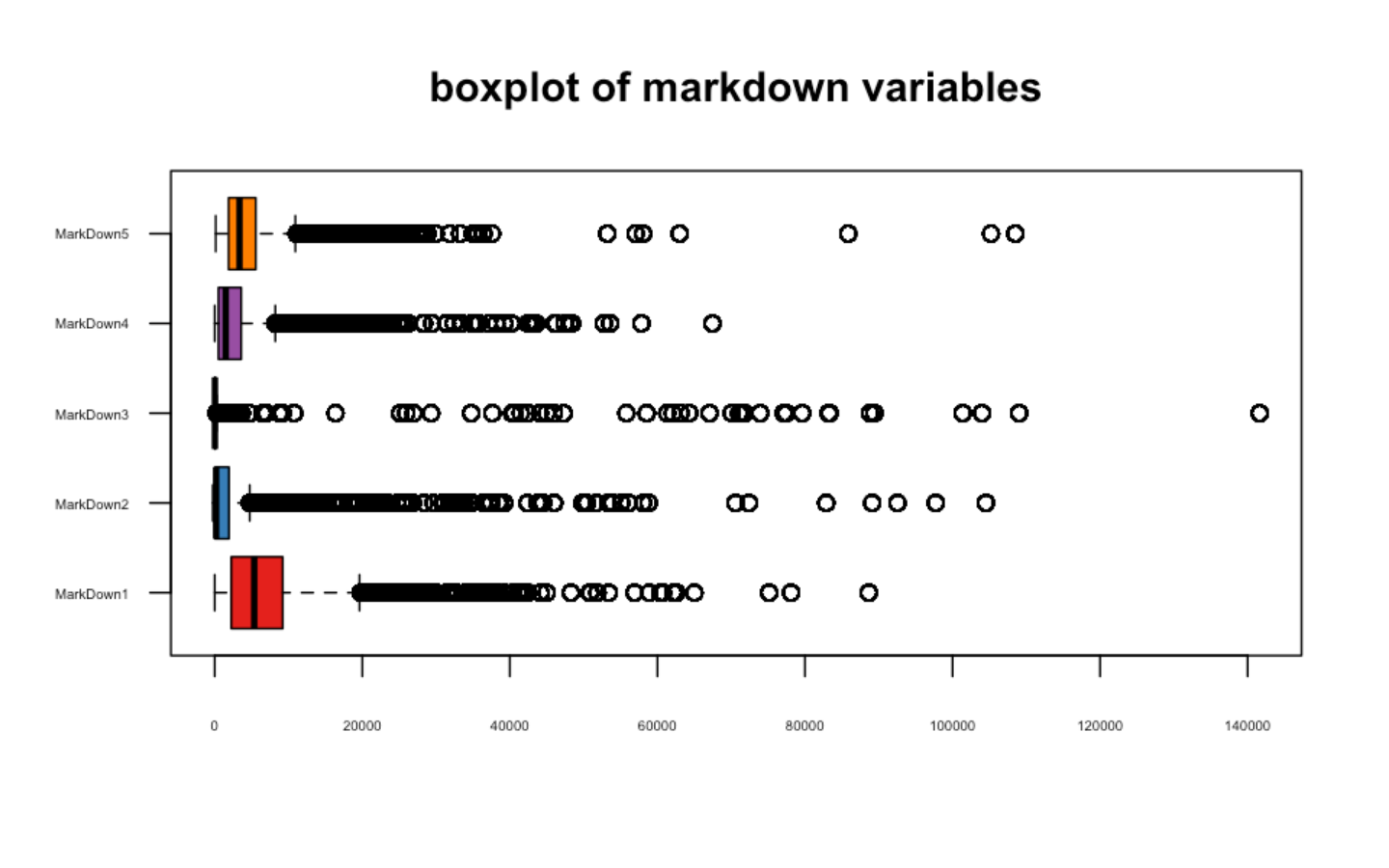
**KEY TAKE OUTS**

* High no of NAs are present in data
* The NAs are present only in Markdown variables, due to the data on these variables not being collected in the initial time period
* Data also seems to have outliers, especially in Markdown variables, Weekly\_sales, Temperature and CPI variables

**BOXPLOTS**

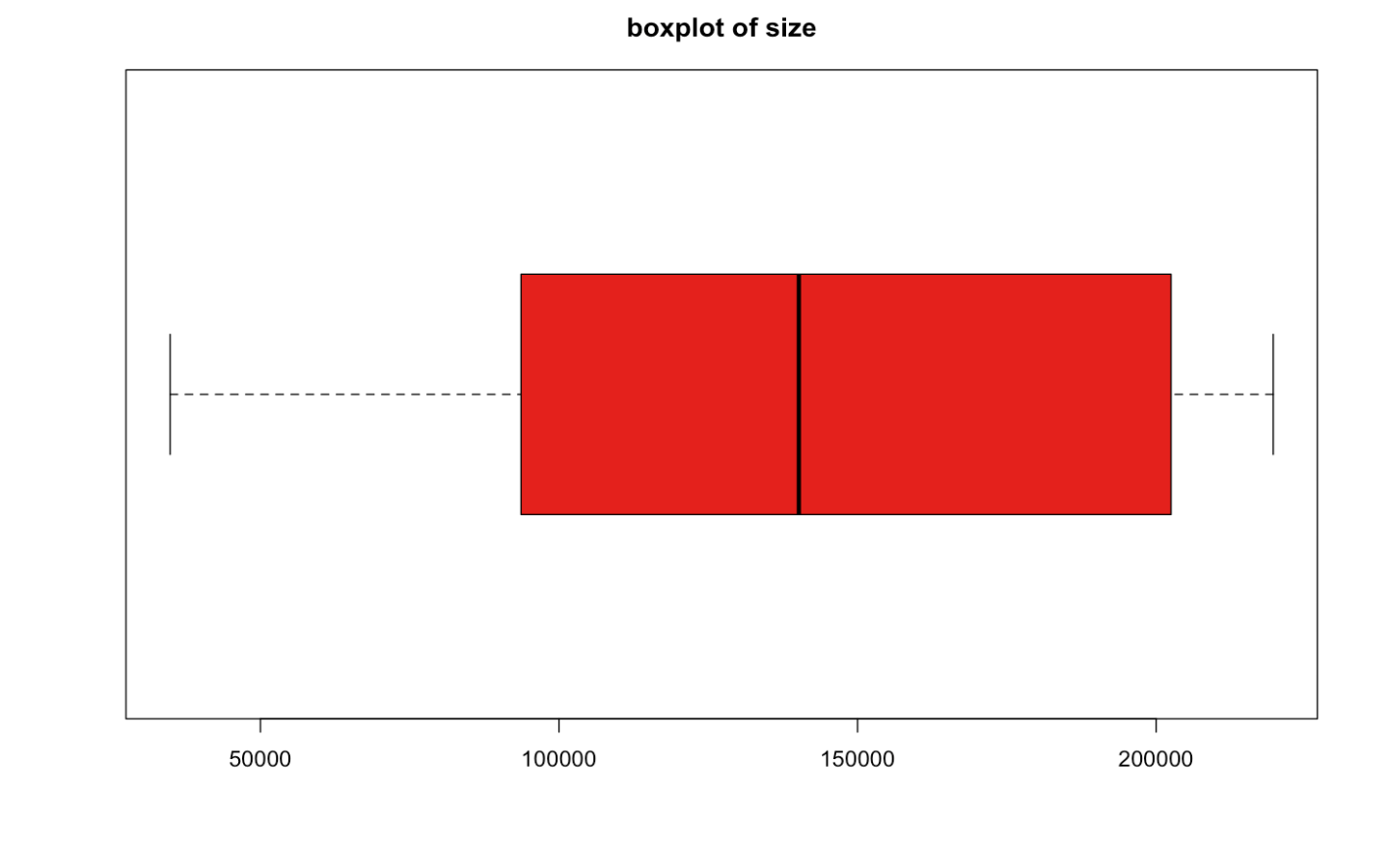
**MARKDOWN VARIABLES**

* Markdown 1 has a higher range and average markdown compared to other markdowns.
* All Markdown variables have high number of outliers



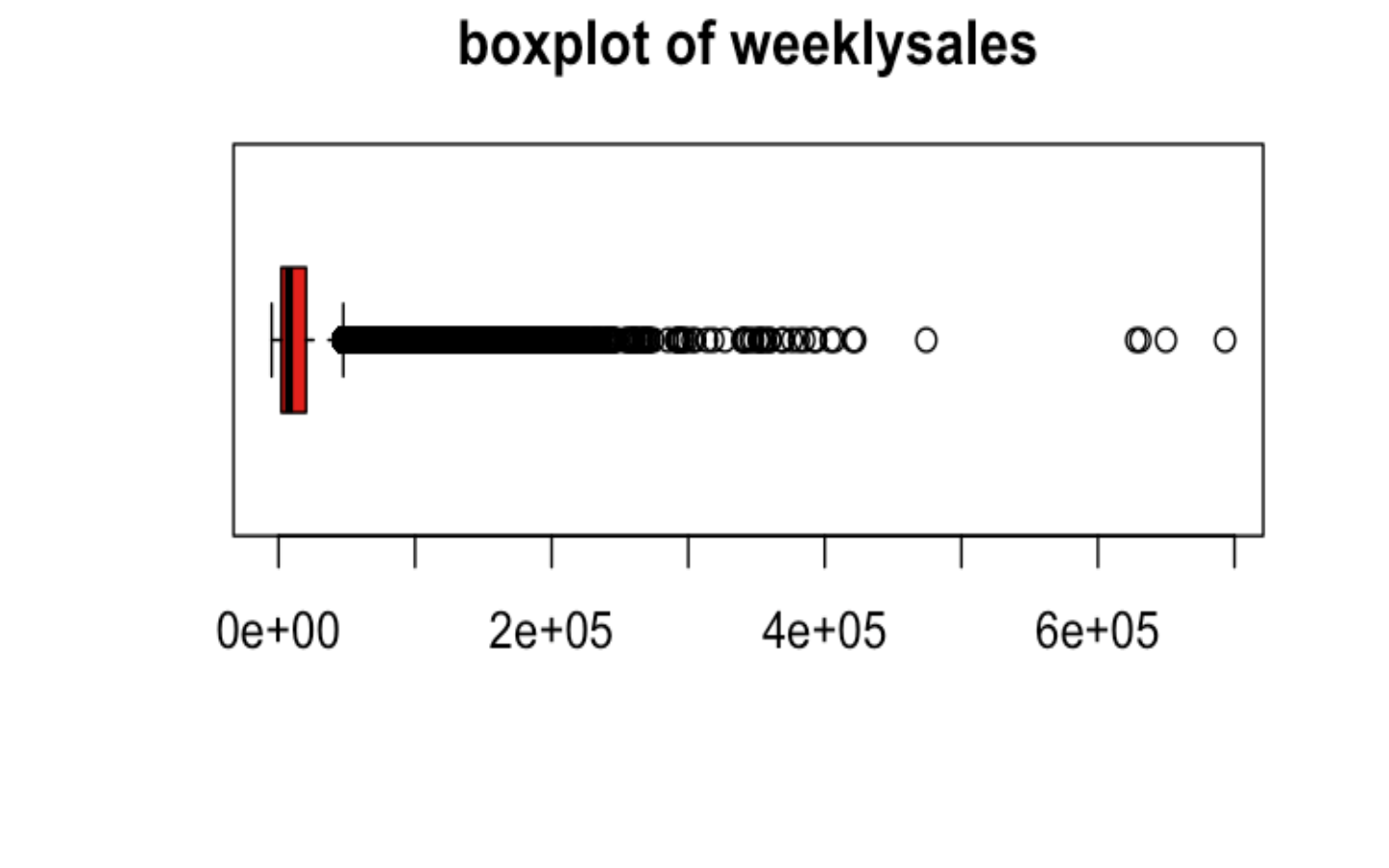
**SIZE**

* Average size of store is around140000 sqft .No outliers seem to be present

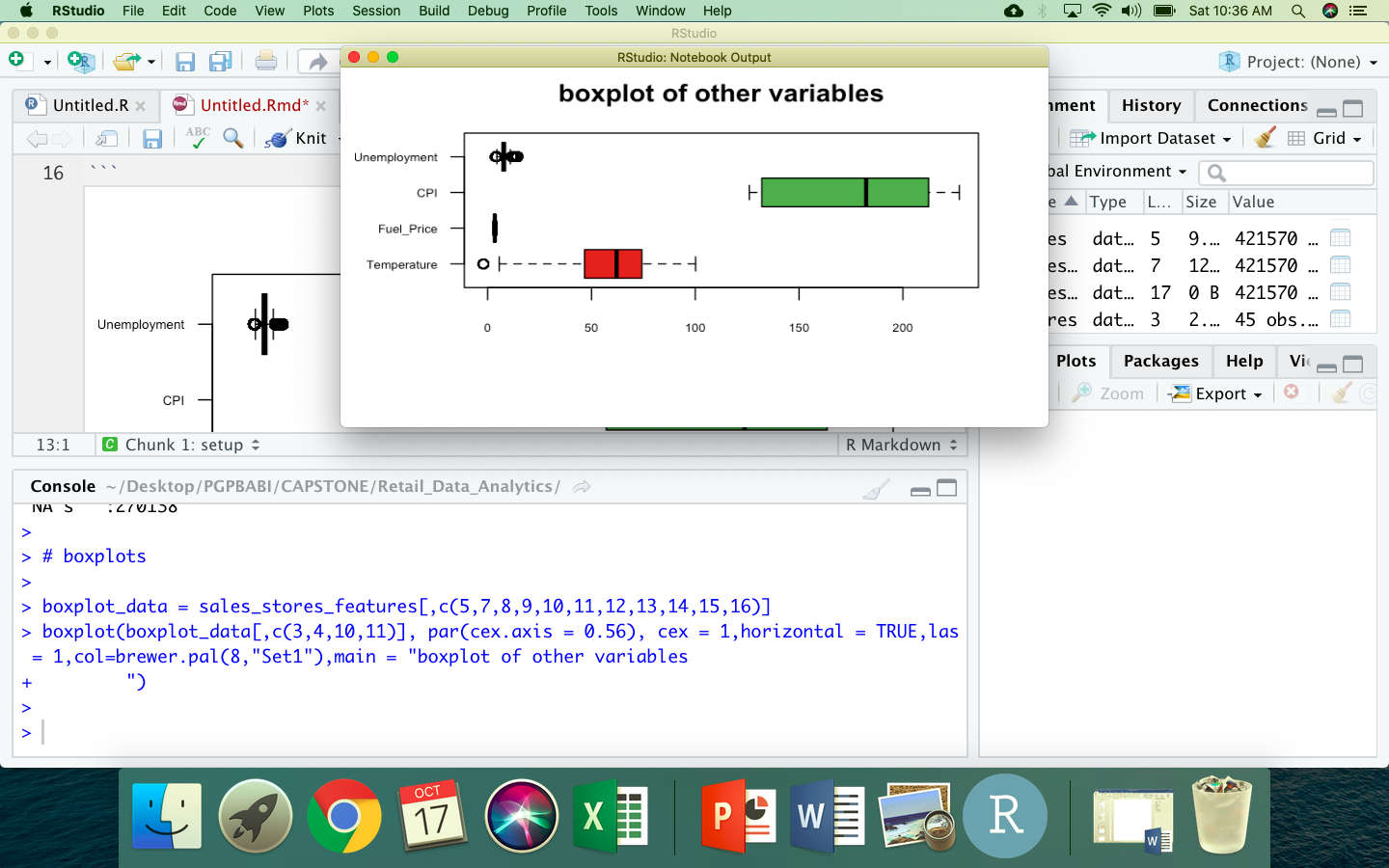


**WEEKLY SALES**

* Weekly Sales heavily impacted by outliers too

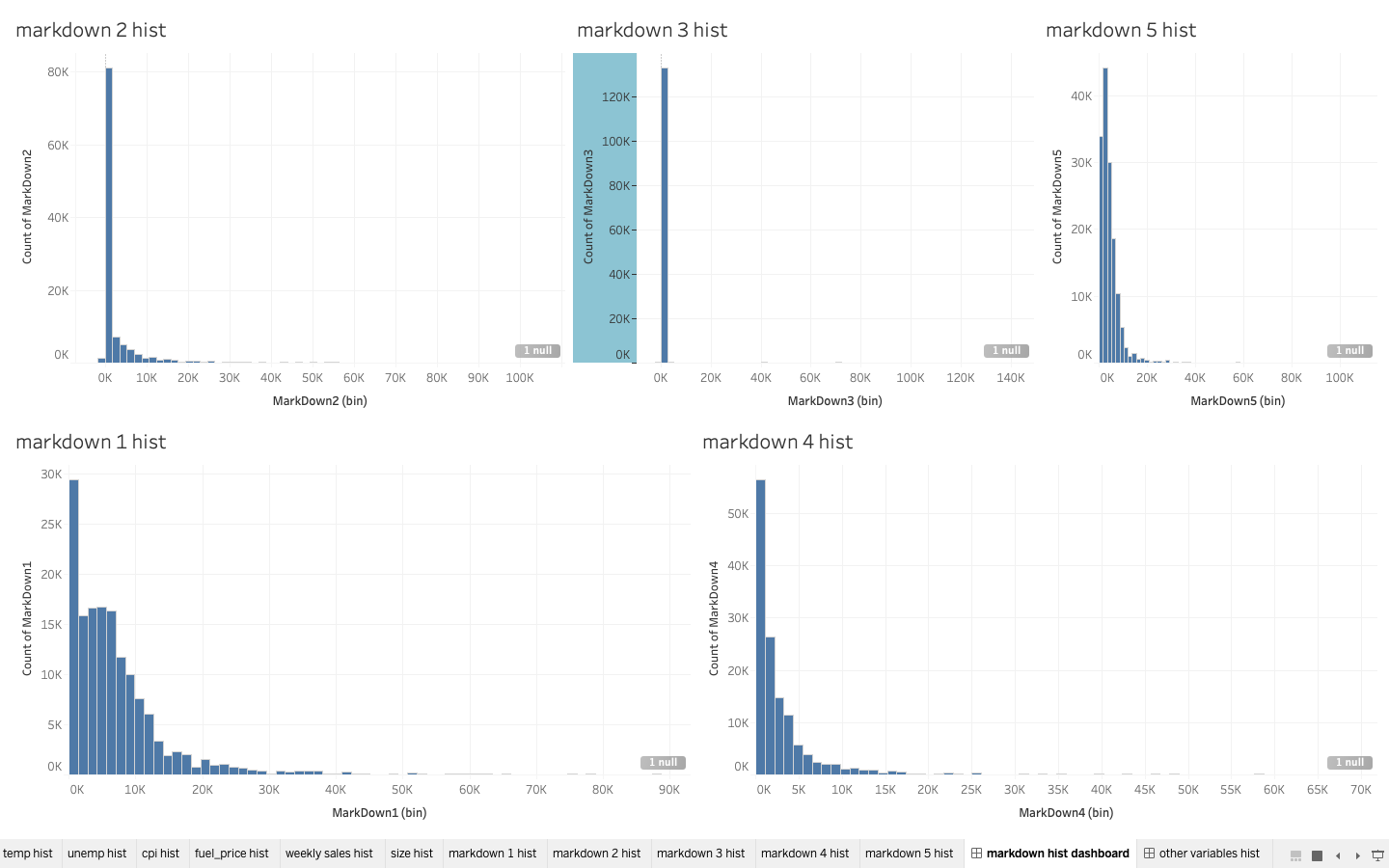
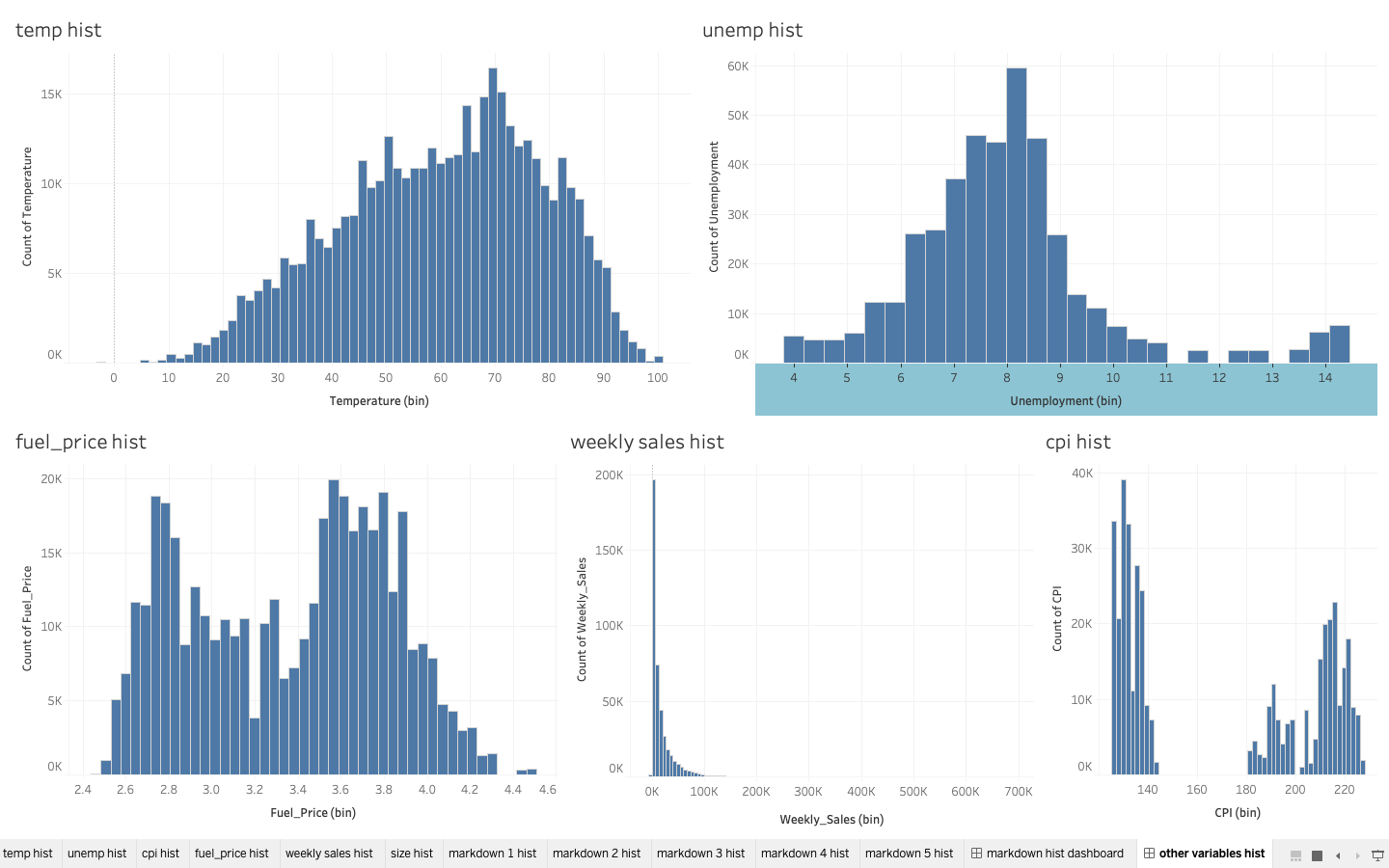
****

**OTHER VARIABLES**

* Among other variables, Unemployment and Temperature indicate presence of outliers

**HISTOGRAMS**

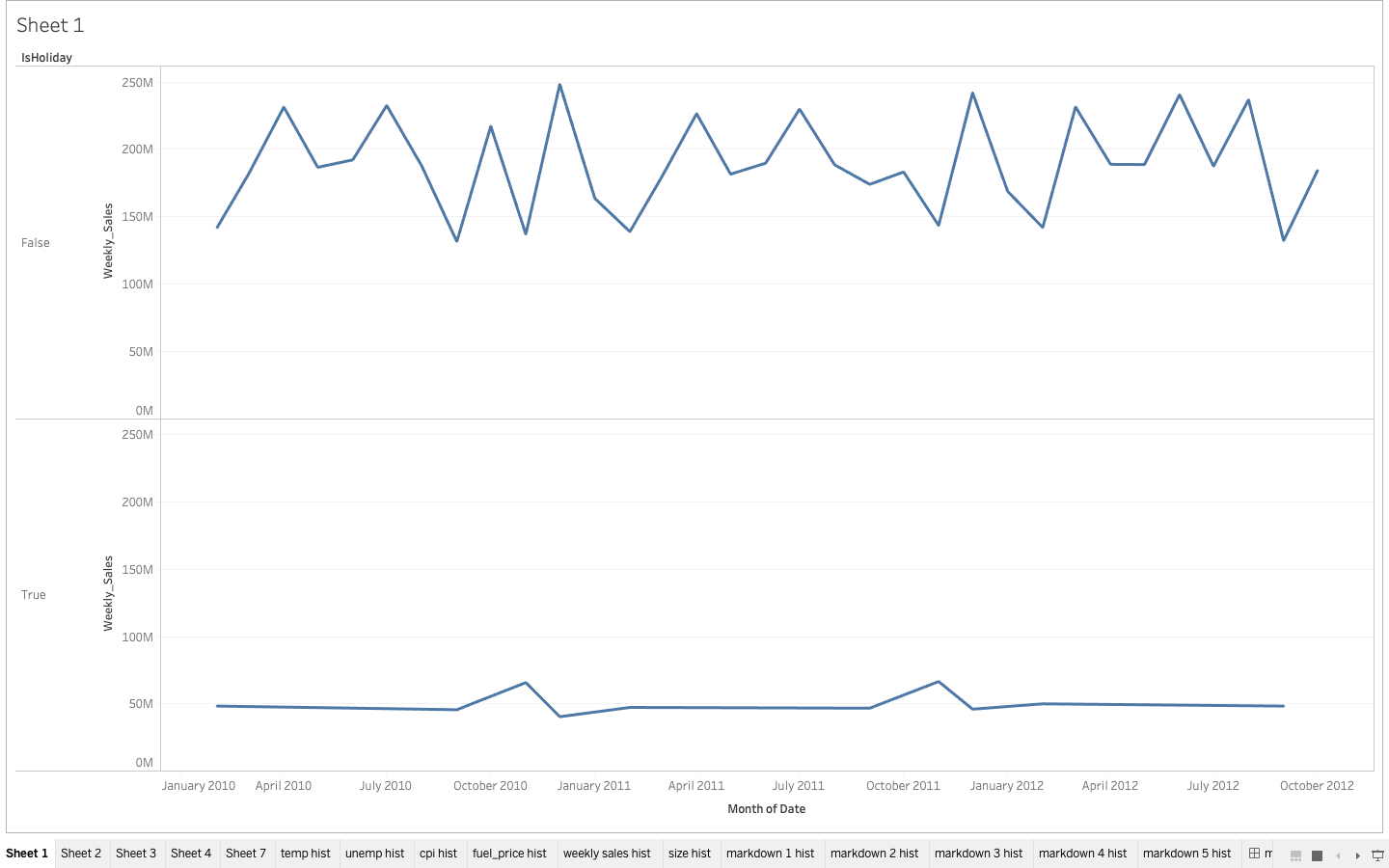
* Temperature and Unemployment indicate a near normal distribution
* Fuel Price and CPI distribution has 2 peaks
* Weekly sales distribution is left skewed
* All markdown variables have left skewed distribution too



**BIVARIATE ANALYSIS**

**WEEKLY SALES BY DATE**

* Overall weekly sales are witnessing a decline from 2010 to 2012
* When there is a holiday weekly sales witness peaks in November every year
* When it is not a holiday sales are peaking in April, July, October, December every year



**FALSE**

**TRUE**

**ISHOLIDAY**

**TABLE SUMMARY BY STORE TYPE**

* Store Type A has highest sales, CPI and markdowns, followed by Store Type B

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Type** | **sum\_Weekly\_Sales** | **avg\_CPI** | **avg\_unemployment** | **avg\_**  **fuel**  **price** | **size** | **avg\_**  **temp** | **avg\_markdown1** | **avg\_markdown2** | **avg\_markdown3** | **avg\_markdown4** | **avg\_markdown5** |
| A | 4331014723 | 174.4 | 7.8 | 3.3 | 182231 | 60.5 | 8687 | 3764 | 1647 | 3916 | 6000 |
| B | 2000700737 | 167.2 | 7.9 | 3.4 | 101819 | 57.6 | 7109 | 3064 | 1484 | 2926 | 3689 |
| C | 405503528 | 170.4 | 8.9 | 3.4 | 40536 | 67.6 | 395 | 447 | 18 | 65 | 1384 |

**TABLE SUMMARY BY IS HOLIDAY**

* Overall average of all markdowns higher for when there is a holiday as compared to when it is not

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **IsHoliday** | **sum\_Weekly\_**  **Sales** | **avg\_CPI** | **avg\_unemployment** | **avg\_**  **fuel**  **price** | **size** | **avg\_**  **temp** | **avg\_markdown1** | **avg\_markdown2** | **avg\_markdown3** | **avg\_markdown4** | **avg\_markdown5** |
| FALSE | 6231919436 | 171 | 8.0 | 3.4 | 136718 | 61 | 7333 | 2299 | 196 | 3384 | 4723 |
| TRUE | 505299552 | 171 | 8.0 | 3.2 | 136859 | 50 | 6241 | 13049 | 15078 | 3371 | 3537 |

**TABLE SUMMARY BY TOP 3 STORES ON SALES**

* Clearly as expected Markdowns are higher for top selling Stores and Departments

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Store** | **sum\_Weekly\_Sales** | **avg\_CPI** | **avg\_unemployment** | **avg\_fuelprice** | **size** | **avg\_temp** | **avg\_markdown1** | **avg\_markdown2** | **avg\_markdown3** | **avg\_markdown4** | **avg\_markdown5** |
| 20 | 301397792 | 209 | 7.4 | 3.4 | 203742 | 55 | 11268 | 5818 | 2772 | 5438 | 6057 |
| 4 | 299543953 | 129 | 6.0 | 3.2 | 205863 | 62 | 9369 | 4575 | 1997 | 5243 | 6813 |
| 14 | 288999911 | 186 | 8.6 | 3.4 | 200898 | 58 | 11586 | 5441 | 2113 | 4226 | 5796 |

**CORRELATION BETWEEN ALL NUMERIC VARIABLES**

* Our numeric variables seem to have very weak correlations with **weekly sales** which is the variable of our interest
* Also among independent variables or other variables, there is not much correlation present.
* Hence, we can safely assume our dataset is free from multicollinearity



**Temp**

**CPI**

**Fuel Price**

**Unemp**

**Markdn3**

**Markdn2**

**Size**

**Markdn5**

**Markdn4**

**Weekly Sales**

**CORRELATION BETWEEN CATEGORICAL VARIABLES AND DEPENDENT VARIABLE**

* Low correlation among dependent and categorical variables

****

**INSIGHTS FROM THE EDA**

* Weekly sales have a strong seasonal pattern where the sales peak during April, July, November and December every year.
* The categorical variables such as Isholiday and Storetype seem to be important and differentiate to an extent in explaining our dependent variable weekly sales
* The data does not seem to suffer from multicollinearity and has very low correlations
* Markdowns also seem to be important variables in explaining Sales as they are higher for Stores generating higher sales

**DATA TREATMENT**

* Treated outliers in the data using flooring and capping method
  + We identified outliers as being any point of data that lies over 1.5 IQRs below the first quartile (Q1) or above the third quartile (Q3)in a data set.
  + And capped outliers in all variables at 5 percentile and 95 percentile values
  + This approach was used because this method is least affected by extreme values
* Treated missing values in markdown variables, we replaced NAs with average markdown values, this was the best method given the huge number of NAs present where imputation algorithms may not work properly. Missing values need to be treated since modelling algorithms may not work properly if data contains missing values
* A new variable weight was created to give weightage based on whether it’s a holiday or not, with higher weight on days when it’s a holiday. Since consumer spending is higher during holiday season and therefore to capture the importance of holidays on Sales, the weight variable was created

**DATASETS**

Given the high no of NAs present in data ,we worked with two types of data sets for modeling :

* The data with no missing values/i.e. complete cases
  + Datafull.nona - 97056 obs. of 17 variables
* The data with imputed missing values
* Datafull - 421570 obs. of 17 variables

**DATA PARTITIONING**

* Both of the above Datasets are partitioned into Train and Test sets as 70:30 Ratio to check for model validity.

**MODEL BUILDING**

**MODELING**

* We need to predict **Sales of stores**, so our **dependent variable is weekly sales** which is a continuous variable. Also our predictor variables are a mix of continuous and factor variables
* Hence we chose the below mentioned algorithms for modeling weekly sales which are suitable for this task
  + Linear Regression
  + Forward and Backward Regression
  + SVM
  + Random Forest
  + Decision Trees – Cart
  + XGboost
* To arrive at the best fit model, we compared Rsqaure, RMSE and MSE across all the models, and the one with lowest RMSE and higher Rsquare is chosen as the final model

**LINEAR REGRESSION ON FULL DATA**

**MODEL 1**

Call:

lm(formula = Weekly\_Sales ~ ., data = train)

Residuals:

Min 1Q Median 3Q Max

-44170 -5017 -635 4531 63897

Coefficients: (4 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.333e+04 2.508e+03 -5.315 1.07e-07 \*\*\*

Size 2.443e-01 1.421e-02 17.192 < 2e-16 \*\*\*

Temperature 6.509e-01 1.233e+00 0.528 0.597704

Fuel\_Price -4.091e+02 7.320e+01 -5.589 2.28e-08 \*\*\*

MarkDown1 -1.607e-02 7.939e-03 -2.024 0.042923 \*

MarkDown2 -6.409e-02 6.276e-03 -10.212 < 2e-16 \*\*\*

MarkDown3 2.191e+00 9.467e-02 23.142 < 2e-16 \*\*\*

MarkDown4 1.674e-02 1.259e-02 1.330 0.183599

MarkDown5 3.108e-02 1.218e-02 2.552 0.010705 \*

CPI -1.078e+01 1.482e+01 -0.728 0.466870

Unemployment -2.246e+02 4.932e+01 -4.554 5.27e-06 \*\*\*

Store2 -9.129e+03 7.747e+02 -11.785 < 2e-16 \*\*\*

...store 45 …………..

Date 4.095e-01 1.983e-01 2.065 0.038878 \*

IsHolidayTRUE 4.113e+02 7.289e+01 5.643 1.67e-08 \*\*\*

Dept81 -3.374e+03 2.062e+02 -16.362 < 2e-16 \*\*\*

…Dept99 -2.357e+04 4.338e+02 -54.329 < 2e-16 \*\*\*

TypeB NA NA NA NA

TypeC NA NA NA NA

weight NA NA NA NA

---

*(To shorten the summary the store and department coefficients have been trimmed to show only first and last values)*

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9418 on 272646 degrees of freedom

Multiple R-squared: 0.7184, Adjusted R-squared: 0.7183

F-statistic: 5153 on 135 and 272646 DF, p-value: < 2.2e-16

* Rsquare of 71.8%
* Some insignificant variables such as Type, Weight, CPI, Markdown1 and MarkDown4, Date and Temperature in Model 1 and subsequent Model
* We will rerun the model without the insignificant variables to note any improvement in R square

**Summary Interpretation**

* R square - 71.8% of the variance found in our response variable can be explained by the predictor variables
* F statistic of 5153 indicates a good relationship between our predictor and the response variables. The further the F-statistic is from 1 the better it is.
* Residual Standard Error is measure of the *quality* of a linear regression fit. The Residual Standard Error is the average amount that the response(weekly sales) will deviate from the true regression line. In our case the actual weekly sales can deviate from the true regression line by approximately **9418** on average.
* The intercept is the expected value of the weekly sales when we consider the average of all other variables in our dataset. In other words, our average weekly sales will be = -1.333e+04
* The Coefficients is the slope or the effect predictor has on weekly sales. The slope term in our model for MarkDown3 is saying that for every 1 unit increase in the MarkDown3 , the weekly sales goes up by 2.191e+00
* The coefficient Standard Error measures the average amount that the coefficient estimates vary from the actual average value of our response variable. We’d ideally want a lower number relative to its coefficients.
* The coefficient t-value is a measure of how many standard deviations our coefficient estimate is far away from 0. We want it to be far away from zero as this would indicate we could reject the null hypothesis - that is, we could declare a relationship between predictor and response variable. In our case, the t-statistic values are relatively far away from zero and are large relative to the standard error, which indicates a relationship exists.
* Pvalue - A small p-value indicates that it is unlikely we will observe a relationship between the predictor and response variables due to chance. a p-value of 5% or less is a good cut-off point

**MODEL 2 - REMOVING LEAST IMPORTANT VARIABLES BASIS ABOVE MODEL**

Call:

lm(formula = Weekly\_Sales ~ . - MarkDown1 - Date, data = train)

Residuals:

Min 1Q Median 3Q Max

-44097 -5017 -639 4534 63904

Coefficients: (1 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) -7.598e+03 9.451e+02 -8.040 9.03e-16 \*\*\*

Size 2.334e-01 5.229e-03 44.637 < 2e-16 \*\*\*

Fuel\_Price -2.969e+02 4.946e+01 -6.002 1.95e-09 \*\*\*

MarkDown2 -6.610e-02 6.071e-03 -10.887 < 2e-16 \*\*\*

MarkDown3 2.203e+00 9.295e-02 23.707 < 2e-16 \*\*\*

MarkDown5 2.536e-02 1.156e-02 2.194 0.0282 \*

Unemployment -2.981e+02 4.051e+01 -7.357 1.88e-13 \*\*\*

Store2 -8.572e+03 3.730e+02 -22.979 < 2e-16 \*\*\*

IsHolidayTRUE 4.275e+02 7.170e+01 5.963 2.49e-09 \*\*\*

Dept2 2.187e+04 2.069e+02 105.730 < 2e-16 \*\*\*

Residual standard error: 9418 on 272651 degrees of freedom

Multiple R-squared: 0.7184, Adjusted R-squared: 0.7183

F-statistic: 5351 on 130 and 272651 DF, p-value: < 2.2e-16

* No improvement in Rsquare from previous model - 71.8%
* However, all variables are significant in the model now

**MODEL 3 - BACKWARD LINEAR REGRESSION ON FULL DATA FOR FEATURE SELECTION**

**Final Model:**

**Weekly\_Sales ~ Temperature + Fuel\_Price + MarkDown2 + MarkDown3 + MarkDown4 + MarkDown5 + CPI + Unemployment + Store + Date + IsHoliday + Dept**

Step Df Deviance Resid. Df Resid. Dev AIC

1 62665 5.957000e+12 1153792

2 - weight 0 0.000000e+00 62665 5.957000e+12 1153792

3 - Type 0 0.000000e+00 62665 5.957000e+12 1153792

4 - Size 0 4.199219e-02 62665 5.957000e+12 1153792

5 - MarkDown1 1 5.554719e+06 62666 5.957006e+12 1153790

Multiple R-squared: 0.7184, Adjusted R-squared: 0.7183

F-statistic: 5192 on 134 and 272647 DF, p-value: < 2.2e-16

* No improvement in Rsquare in this model from previous model – 71.8%
* There still are many insignificant variables in this model

**LINEAR REGRESSION ON DATA WITH NO MISSING VALUES**

**FINAL MODEL WITH ALL SIGNIFICANT VARIABLES**

call:

lm(formula = Weekly\_Sales ~ . - Type - Size - MarkDown1 - MarkDown4 - MarkDown5 - weight, data = train2)

Residuals:

Min 1Q Median 3Q Max

-46281 -5174 -772 4605 63209

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.836e+04 1.594e+04 1.778 0.075332 .

Temperature 1.194e+01 4.683e+00 2.551 0.010749 \*

Fuel\_Price -9.342e+02 2.224e+02 -4.201 2.66e-05 \*\*\*

MarkDown2 -6.920e-02 7.842e-03 -8.824 < 2e-16 \*\*\*

MarkDown3 2.028e+00 1.054e-01 19.251 < 2e-16 \*\*\*

CPI 2.877e+02 9.660e+01 2.978 0.002904 \*\*

Unemployment 5.112e+02 2.024e+02 2.526 0.011536 \*

Store2 3.465e+03 3.260e+02 10.628 < 2e-16 \*\*\*

Date -4.349e+00 1.150e+00 -3.781 0.000156 \*\*\*

IsHolidayTRUE 6.535e+02 1.345e+02 4.859 1.19e-06 \*\*\*

Dept2 2.706e+04 4.612e+02 58.668 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9750 on 62668 degrees of freedom

Multiple R-squared: 0.7481, Adjusted R-squared: 0.7476

F-statistic: 1410 on 132 and 62668 DF, p-value: < 2.2e-16

* Slight improvement in Rsquare in this model from previous models – 74.7%

**SVM REGRESSION ON DATA WITH NO MISSING VALUES**

**The principle of SVM is to find an hyperplane which, can classify the training data points in to labelled categories.**

svm(formula = Weekly\_Sales ~ ., data = train2)

Parameters:

SVM-Type: eps-regression

SVM-Kernel: radial

cost: 1

gamma: 0.007142857

epsilon: 0.1

Number of Support Vectors: 39628

Rsquare – 77.9%

* Better Rsquare in this model from previous Linear Regression models

**BAGGED CART ON DATA WITH NO MISSING VALUES**

* Tuning parameters used – 5fold cross validation

**FINAL MODEL**

Bagging regression trees with 25 bootstrap replications

|  | **parameter**  <fctr> | **RMSE**  <dbl> | **Rsquared**  <dbl> | **MAE**  <dbl> | **RMSESD**  <dbl> | **RsquaredSD**  <dbl> | **MAESD**  <dbl> |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | none | 12476.6 | 0.5868263 | 9337.564 | 144.6125 | 0.009868366 | 79.33647 |

Rsquare – 58.6%

**DECISION TREES - CART ON FULL DATA**

* Set control parameters to grow the tree
  + minsplit: if the number of records in a given node falls below a threshold, the node will not be split further.
  + minbucket: minimum records in a terminal node. if the records are less, that bucket will not be created.
  + Terminal node (minbucket) should not be less than 2-3% of starting population.
  + minsplit = 3(minbucket)
  + xval divides the entire dataset into mutually exclusive and collectively exhaustive segments.

**FINAL MODEL**

Variable importance

Dept Store Size Type CPI Unemployment

47 17 12 8 4 3

MarkDown1 MarkDown5 Temperature Date Fuel\_Price MarkDown4

2 2 2 1 1 1

Rsquare – 97.2%

* Highest Rsquare achieved so far

**PRUNED CART MODEL ON FULL DATA**

* + Identify the minimum Xerror from the cp table - 0.05960866 in our case
  + Identify the CP(complexity parameter) where the Xerror is lowest - 6.215071e-12 in our case
  + Prune the cart model using the above CP as identified best for the model

Rsquare – 94.9%

* Rsquare lower than previous cart model

**RANDOM FOREST ON DATA WITH NO MISSING VALUES**

* Set control parameters to grow the tree
  + Set Nodesize - no of observations in terminal node
  + Set Mtry - no of independent variables in a tree out of total Independent variables
  + nTree – no of trees to be built, ideally an odd no, to avoid ties in voting while calculating error rates
  + Importance - Give importance matrix of Independent variables
  + Tune the random forest to include optimal number of trees

**FINAL MODEL**

**> RFmodel\_1$importance**

**%IncMSE IncNodePurity**

Size 32312060.61 878781542313

Temperature 1317174.21 74339545383

Fuel\_Price 1271922.63 70823129334

MarkDown1 1127615.25 81526669446

MarkDown2 699369.60 56222629123

MarkDown3 3772058.13 143219749978

MarkDown4 1454229.15 89553069246

MarkDown5 1311896.26 132809972816

CPI 12335881.93 253363721080

Unemployment 8651117.95 213581226906

Date 2016785.64 68809549150

IsHoliday 7427.93 3305156230

Type 15764058.37 323300799529

weight 1066.82 3211695310

Rsquare – 9 %

* Most important variables in the Random forest model are Type, CPI, Unemployment, Markdown5 ,Markdown 3 and Size
* Lowest Rsquare in this model

**XGBOOST ON FULL DATA**

* Set control parameters
  + Set eta or learning rate - shrinks the feature weights to make the boosting process more conservative.
  + max\_depth = Maximum depth of a tree.
  + min\_child\_weight = minimum sum of instance weight needed in a child node
  + nfold = cross validation
  + objective = "reg:linear",

**Importance Matrix**

|  |
| --- |
|  |

|  |
| --- |
|  |
| **Feature**  <chr> | **Gain**  <dbl> | **Cover**  <dbl> | **Frequency**  <dbl> |  |
| Dept | 0.6809209965 | 0.406777020 | 0.099279866 |  |
| Size | 0.1959863492 | 0.137973061 | 0.061928524 |  |
| Store | 0.0522819935 | 0.066755535 | 0.029836252 |  |
| CPI | 0.0299114259 | 0.103674450 | 0.177478635 |  |
| Unemployment | 0.0132685388 | 0.045147081 | 0.069607960 |  |
| Temperature | 0.0111260676 | 0.081617699 | 0.223629953 |  |
| Fuel\_Price | 0.0084154522 | 0.069744371 | 0.189087432 |  |
| MarkDown3 | 0.0033681078 | 0.027961715 | 0.027938804 |  |
| MarkDown2 | 0.0012791253 | 0.011674607 | 0.025891950 |  |
| MarkDown4 | 0.0012567005 | 0.015887316 | 0.024308253 |  |
| MarkDown1 | 0.0011613116 | 0.015524564 | 0.035379191 |  |
| MarkDown5 | 0.0005834059 | 0.011056596 | 0.029567322 |  |
| weight | 0.0004405253 | 0.006205984 | 0.006065858 |  |

Rsquare – 98.6 %

* Dept,Size and Store the top 3 most important variables in our XGBoost model.
* Highest RSquare achieved for XGboost model

**SUMMARY TABLE – ALL MODELS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model/Parameters | Rsquare | MSE | RMSE | MAE |
| Linear Regression\_Full data Final Model | | | | |
| Train | 71.8% | 88652583 | 9415.55 | 17593.74 |
| Test | 71.7% | 558944750 | 23642.01 | 6695.625 |
| Linear Regression\_NONA Data Final Model | | | | |
| Train | 74.7% | 94865383 | 9739.886 | 19037.11 |
| Test | 74.5% | 656198499 | 25616.37 | 6793.709 |
| SVM\_NONA Data | | | | |
| Train | 77.9% | 86088185 | 9278.372 | 5557.108 |
| Test | 77.7% | 622794011 | 24955.84 | 18028.99 |
| Bagging\_NONA Data | | | | |
| Train | 58.9% | 154567435 | 12432.52 | 9298.325 |
| Test | 58.9% | 151042425 | 12289.93 | 9247.575 |
| Decision Tree\_Cart Full Data | | | | |
| Train | 97.2% | 8760126 | 2959.751 | 18646.27 |
| Test | 95.1% | 15475928 | 3933.946 | 18524.96 |
| Decision Tree\_Cart NONA Data | | | | |
| Train | 96.7% | 12355648 | 3515.06 | 1682.73 |
| Test | 93.7% | 23171916 | 4813.722 | 2348.198 |
| XG Boost\_Full Data | | | | |
| Train | 98.6% | 4517517 | 2125.445 | 989.3375 |
| Test | 95.1% | 15584339 | 3947.7 | 1807.93 |
| XG Boost\_NONA Data | | | | |
| Train | 98.5% | 6002810 | 2450.063 | 1255.38 |
| Test | 93.4% | 24227700 | 4922.164 | 2522.254 |
| Random Forest\_NONA Data | | | | |
| Train | 9% | 341074152 | 18468.19 | 14024.96 |
| Test | 7% | 341717875 | 18485.61 | 14051.38 |

* The Mean Squared Error (MSE) is a measure of how close a fitted line is to data points. The MSE has the units squared of whatever is plotted on the vertical axis.
* Root Mean Squared Error (RMSE)is the square root of the mean square error.
* MAPE measures the deviation from the actual data in terms of percentage.*The mape in our case is not adding much value, therefore we will ignore this metric.*
* We will focus on RMSE to arrive at best fit model since RMSE gives a relatively high weight to large errors.
* As seen in the above table summary - XG BOOST FULL DATA AND CART FULL DATA models have the best values on Rsquare and lowest values on MSE and RMSE.
* Also while we tried multiple attempts at model tuning, it was not very successful because it required high computing power due to large dataset, which was not possible at regular laptops, except for cart pruning which worked well in our data
* However cross validation is used wherever possible for model tuning.

**MODEL VALIDITY**

* For testing the model validity , we looked at rsquare, rmse of the above models and also observed these metrics for each model between train and test sets.
* This is done to evaluate how the trained model is performing on unseen data
* We checked for overfitting and underfitting and good fit for our models as below:
  + Overfitting - refers to a model that models the training data too well but poor generalization to other data.
  + Underfitting - refers to a model that can neither model the training data nor generalize to new data.
  + Good fit - Model balanced between underfitting and overfitting i.e the key metrices such as rsquare should be within **+-10% range for the train and test dataset**
  + Basis this threshold, we can safely assume that our shortlisted models are valid models
* Additionally Cross validation also helped in arriving at a valid model by allowing us to train and test the model k-times on different subsets of training data and build up an estimate of the performance of a machine learning model on unseen data.

**RECOMMENDATIONS**

* As seen in the above table summary - XG BOOST FULL DATA AND CART FULL DATA models have the best values on Rsquare and lowest values on MSE and RMSE.
  + Since Cart models can overfit due to their design and complexity, we may choose to go ahead with XGboost model as our final model.
* Sales can be predicted for the next year using our final shortlisted model. Then we can identify the highest selling/sensitive stores & departments for next year and prioritize them for taking marketing decisions
  + We can compare the sales of last 3 years versus the predicted sales for next year and identify if there are any stores or departments that are showing deviations( for example stores which are generating significantly lower sales compared to last year) and prioritize such stores/departments in our marketing actions
* Also the impact of markdowns can be clearly understood using the linear regression model, and they can be finetuned basis the model equation coefficients for high selling or discount sensitive stores and departments

**THANK YOU**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_