**Objective:**

Walmart Dataset: For each store, department and date, predict the weekly sales of that department

**For the above dataset do**

1. SARIMA without including additional variables

2. SARIMA with additional variables.

3. A method of your choice: prophet, BSTS.

**Steps**

1. Data exploration and initial insights, this includes data visualization, pattern discovery, relationship discovery (between y and x, ACF, PACF, ...), etc.

2. Split the data into training (the first 80% obs) and testing (the remaining 20% obs)

* Train your model using the training set following the procedure described. Make sure to explain your model component, model fitting results and model diagnostics.
* Then perform prediction for the remaining 20% observations and calculate MSPE. For time series model, using the 1-step-ahead prediction.

3. A discussion of what works vs. what doesn’t work.

Data Available:

1. Historical sales data for 45 Walmart stores located in different regions with sales by department between 2010-02-05 to 2012-11-01
2. Stores Information (45 stores,) indicating the type and size of store

Training Dataset

Fields

* 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 dataset

This file 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

**Holidays**

Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13

Labor Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13

Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13

Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

Procedure

1. Summarized data by stores and merged it with features

\*\* On analyzing the stores by their type it shows that number of stores for Type A is highest followed by B and C

\*\* Also, there is a correlation between store types, so, it’s better to segregate the stores by their types

Note: I have not coded below yet, but got from the Prof’s suggested link

Sales for type A and B is normally distributed with left tail being skewed. Hence outlier treatment has to be performed on sales.

**Plotting sales by store type to check trend and important peaks/holidays seasonality** There are few important peaks visible in the data like ThanksGiving and Christmas. While other small peaks corresspond to Labor day, Easter etc. All the store types follow almost identical seasonality pattern with trend being negligible.

**Sales trend across month** Sales seems to be high at 1st week of the month and then falls slowly in other weeks as indicated by boxplot.Outliers in non-first week resemble the Thanksgiving and Christmas holiday sales.

**Is there a relation between CPI and sales?** The relation between sales and consumer price index(CPI) is very weak. Sales decreases negligibly with increase in price index.

**Is there a relation between Unemployment and sales?** Unemployment doesnt affect sales significantly. Sales reduces slightly at high values of unemployment.

**Is there a relation between average temperature of the week and sales?** There is no evident relation between the two as can be seen in graph.

Chart, bar chart

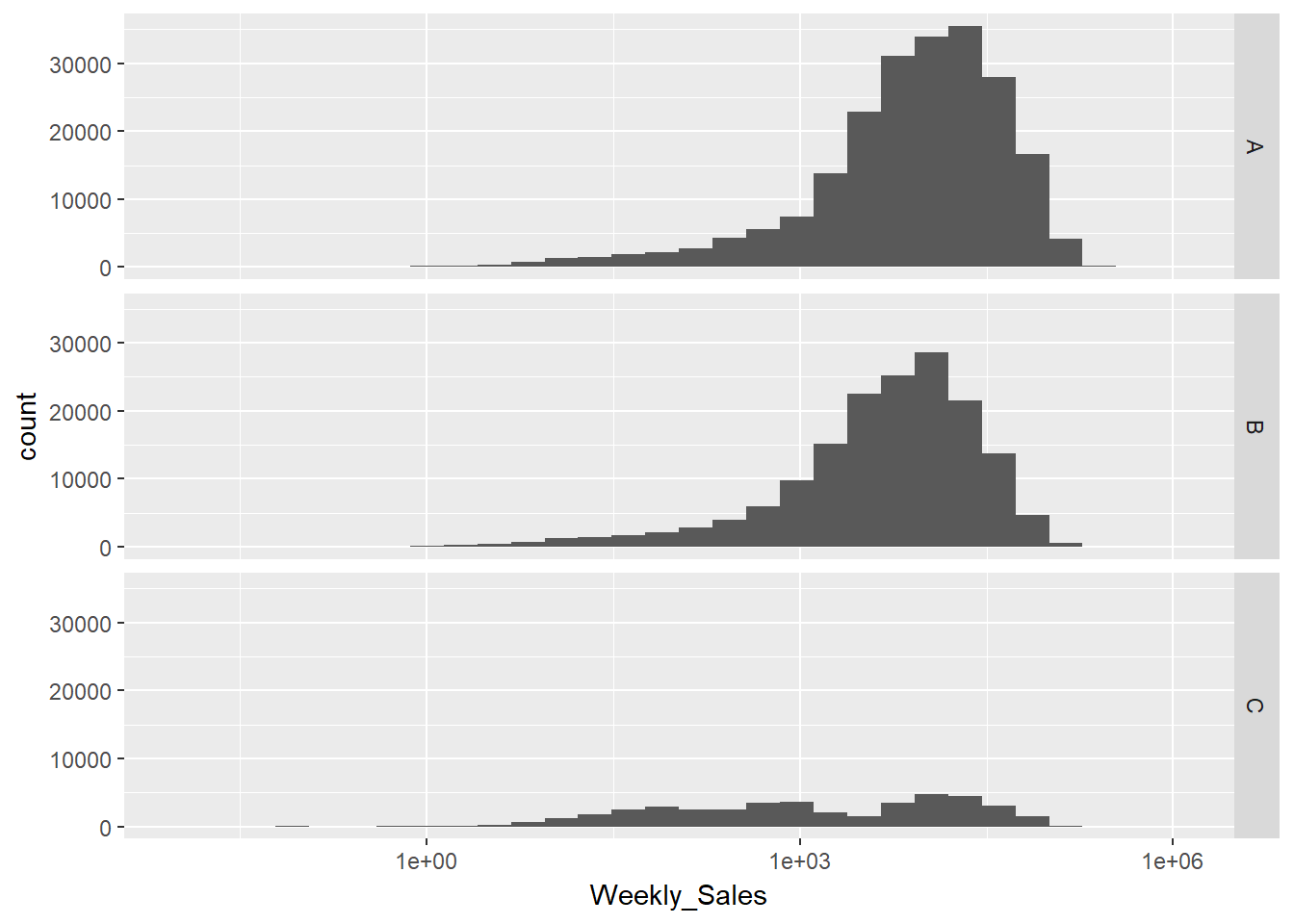
Description automatically generated

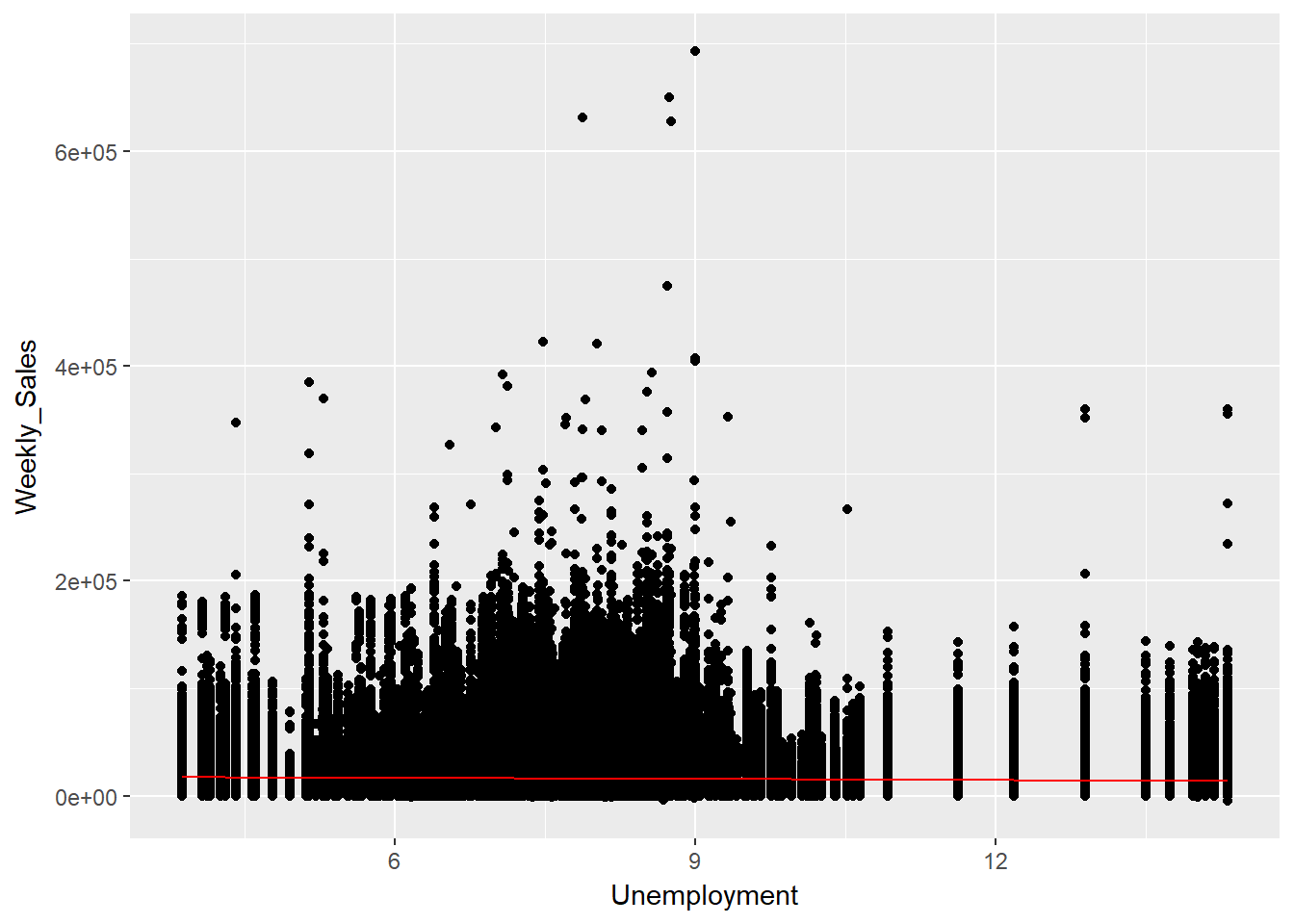


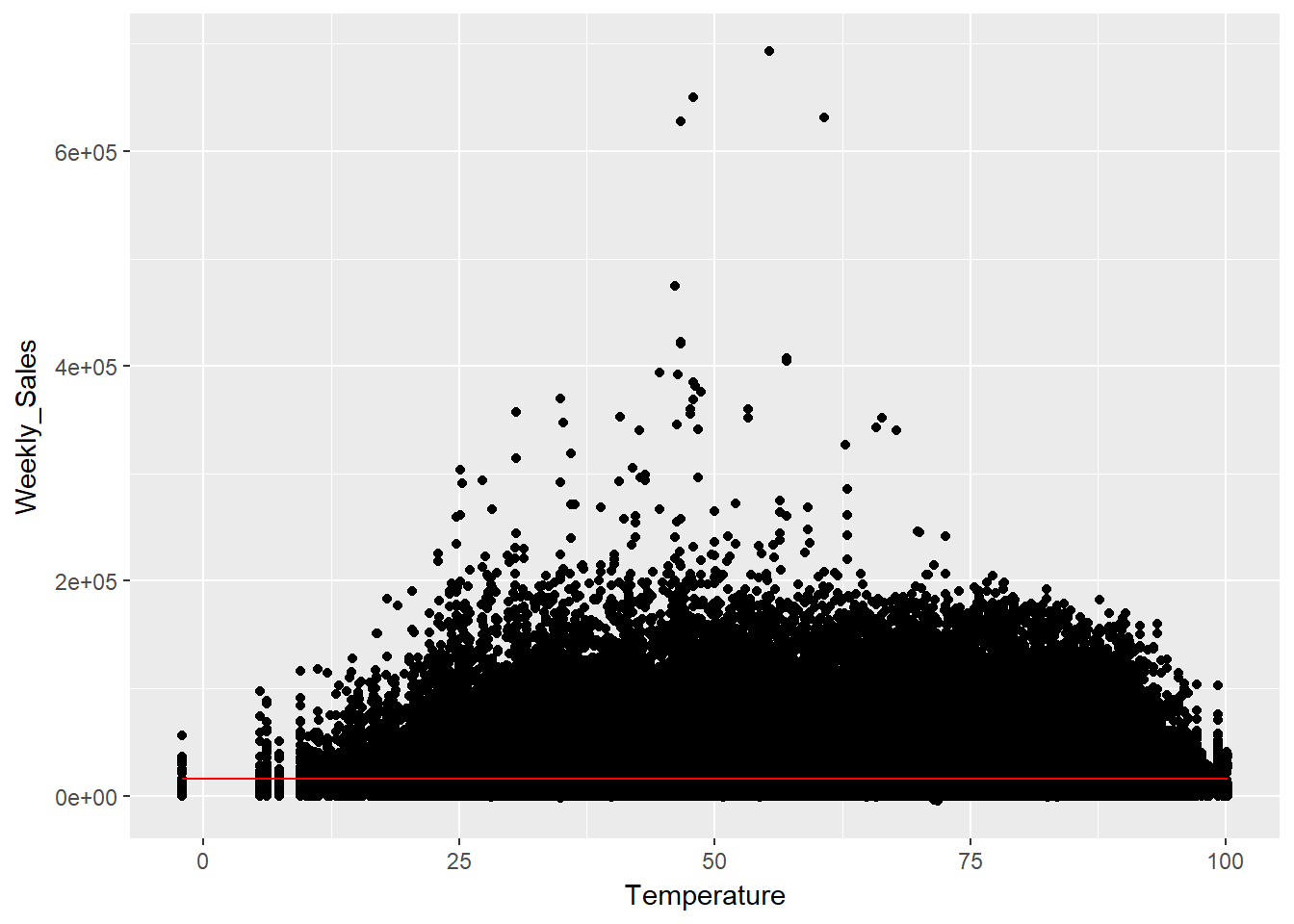
Chart

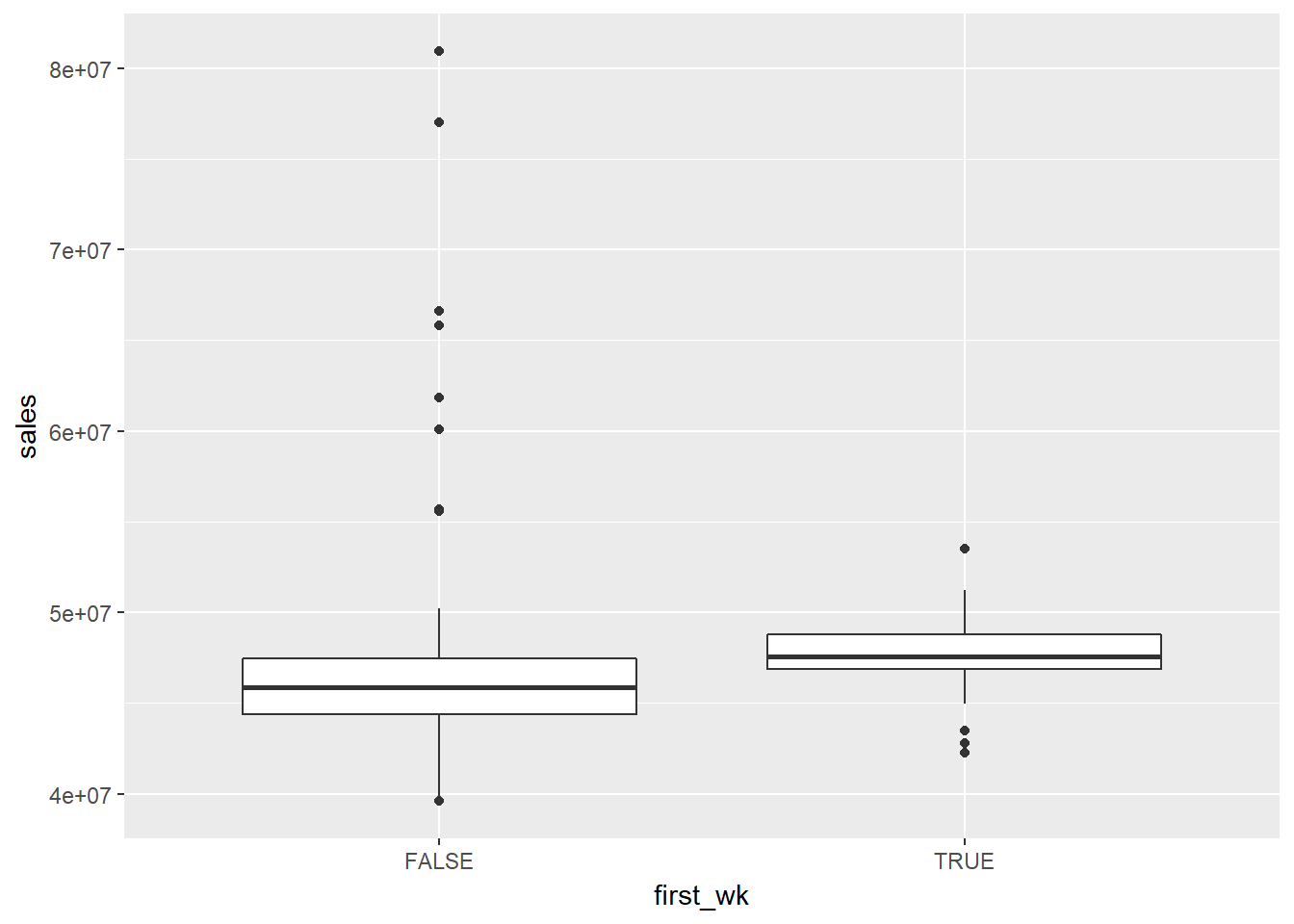
Description automatically generated with medium confidence

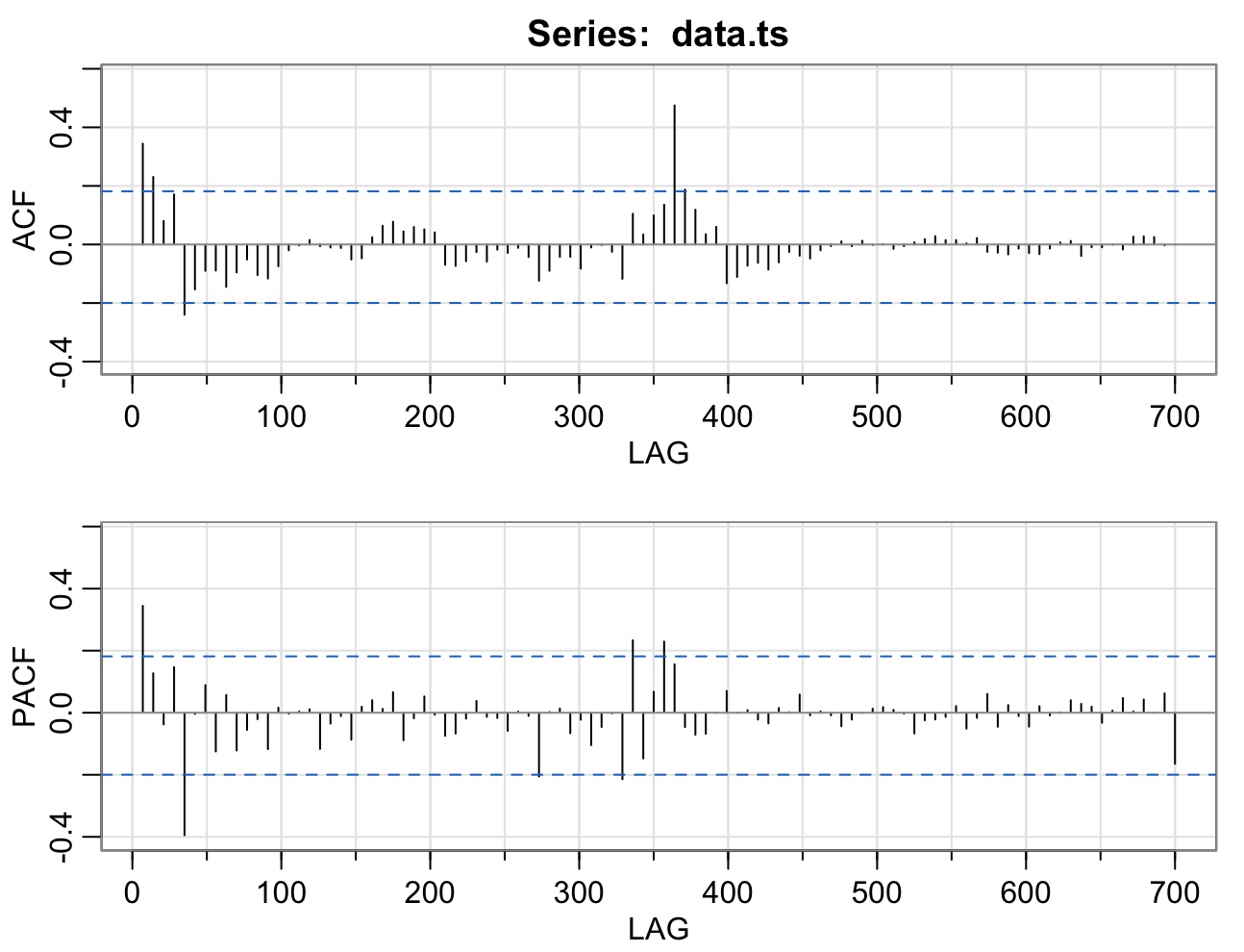
Type A











Type A

Code:

library(dplyr)

#Data extraction

#train.csv has info about store-dept-weekly sales

train = read.csv("/Users/namitsrivastava/Documents/Documents/GSU/Predictive Analysis/walmart-recruiting-store-sales-forecasting/train.csv")

str(train)

head(train)

# aggregated at store level

train.Store\_Date = train %>% group\_by(Store,Date) %>% summarise(Weekly\_Sales\_Store = sum(Weekly\_Sales))

str(train.Store\_Date)

head(train.Store\_Date,5)

max(train.Store\_Date$Date) #"2012-10-26"

min(train.Store\_Date$Date) #"2010-02-05"

#features added

features = read.csv("/Users/namitsrivastava/Documents/Documents/GSU/Predictive Analysis/walmart-recruiting-store-sales-forecasting/features.csv")

str(features)

head(features)

max(features$Date) #"2013-07-26"

min(features$Date) #"2010-02-05"

# join on train data with features

data <- merge(train.Store\_Date, features, by.x = c("Store", "Date"),

by.y = c("Store","Date"), all.x = TRUE, all.y = FALSE)

str(data)

#store has info about store type and size

stores = read.csv("/Users/namitsrivastava/Documents/Documents/GSU/Predictive Analysis/walmart-recruiting-store-sales-forecasting/stores.csv")

str(stores)

head(stores)

max(features$Date) #"2013-07-26"

min(features$Date) #"2010-02-05"

#adding store info

data <- merge(data, stores, by.x = c("Store"),

by.y = c("Store"), all.x = TRUE, all.y = FALSE)

str(data)

#find the summary of data

summary(data)

#MarkDown(s) column has lot of NA values in it

names(data)

#combining all store data to a week level. i.e take data at Date level

data.subset <- data[c("Store","Date","Weekly\_Sales\_Store","Temperature",

"Fuel\_Price","CPI","Unemployment")]

#data need to be aggregated at Date date level

data.week <- data.subset %>% group\_by(Date) %>% summarise(Weekly\_Sales\_allStore = sum(Weekly\_Sales\_Store),

Weekly\_Temperature\_allStore = sum(Temperature),

Weekly\_Temperature\_allStore = )

# store by Type

plot(data$Weekly\_Sales\_Store, data$Type,xlab="Weekly store sales", ylab="Type")

creating a master dataset with types

data.storetype<-left\_join(data, stores, by = "Store")

head(data.storetype)

summary(data)

summary(data.storetype)

#lets pick one/two store per type in random from the data

#Type A: Store 20

#Type B: Store 3

#Type C: Store 30

#store A

store\_a <- data.storetype[data.storetype$Type == "A",]

store\_b <- data.storetype[data.storetype$Type == "B",]

store\_c <- data.storetype[data.storetype$Type == "C",]

#count of store based on type

count.store\_type = data %>% group\_by(Type) %>% summarise(count\_Store = n())

library(ggplot2);

ggplot(count.store\_type, aes(as.factor(Type), count\_Store)) +

geom\_bar(stat = "identity") +

labs(y = "Store count", x = "Type");

#####################################################################

## STORE A

#####################################################################

# split data to train and test

# check acf and pacf for the store data and find right models

# Taking only Weekly\_Sales\_store column for this analysis

train\_store\_a = store\_a\_grouped[1:110,]

test\_store\_a = store\_a\_grouped[111:143,]

head(train\_store\_a)

write.csv(x=store\_20, file="store\_20.csv",row.names = FALSE)

#convert to time series data

train\_store\_a.ts = ts(train\_store\_a,frequency = 52)

plot(train\_store\_a.ts)

#####################################################################

## STORE A

#####################################################################

# split data to train and test

# check acf and pacf for the store data and find right models

# Taking only Weekly\_Sales\_store column for this analysis

train\_store\_a = store\_a\_grouped[1:110,]

test\_store\_a = store\_a\_grouped[111:143,]

head(train\_store\_a)

final= as.Date(train\_store\_a$Date, format = "%Y-%m-%d")

final

train\_store\_a$Date <- as.Date(train\_store\_a$Date , format = "%Y-%m-%d")

train\_store\_a

train\_store\_a.ts

data.ts<-xts(train\_store\_a$Weekly\_Sales\_allStore,train\_store\_a$Date)

data.ts

train\_store\_a.ts = ts(train\_store\_a,start = c(2010, 2))

plot(data.ts)