# FORECASTING OF WOMEN RETAIL SALES

24 NOVEMBER 2017

JAYANTH RASAMSETTI

# AGENDA

- Objective - Phase 1 (TS), Phase 2 - Regression
- Exploratory Data Analysis (EDA)
- Data processing and cleaning
- Model Built and Evaluation
- Improvisation to Models

# 1 OBJECTIVE - PHASE 1 & PHASE 2

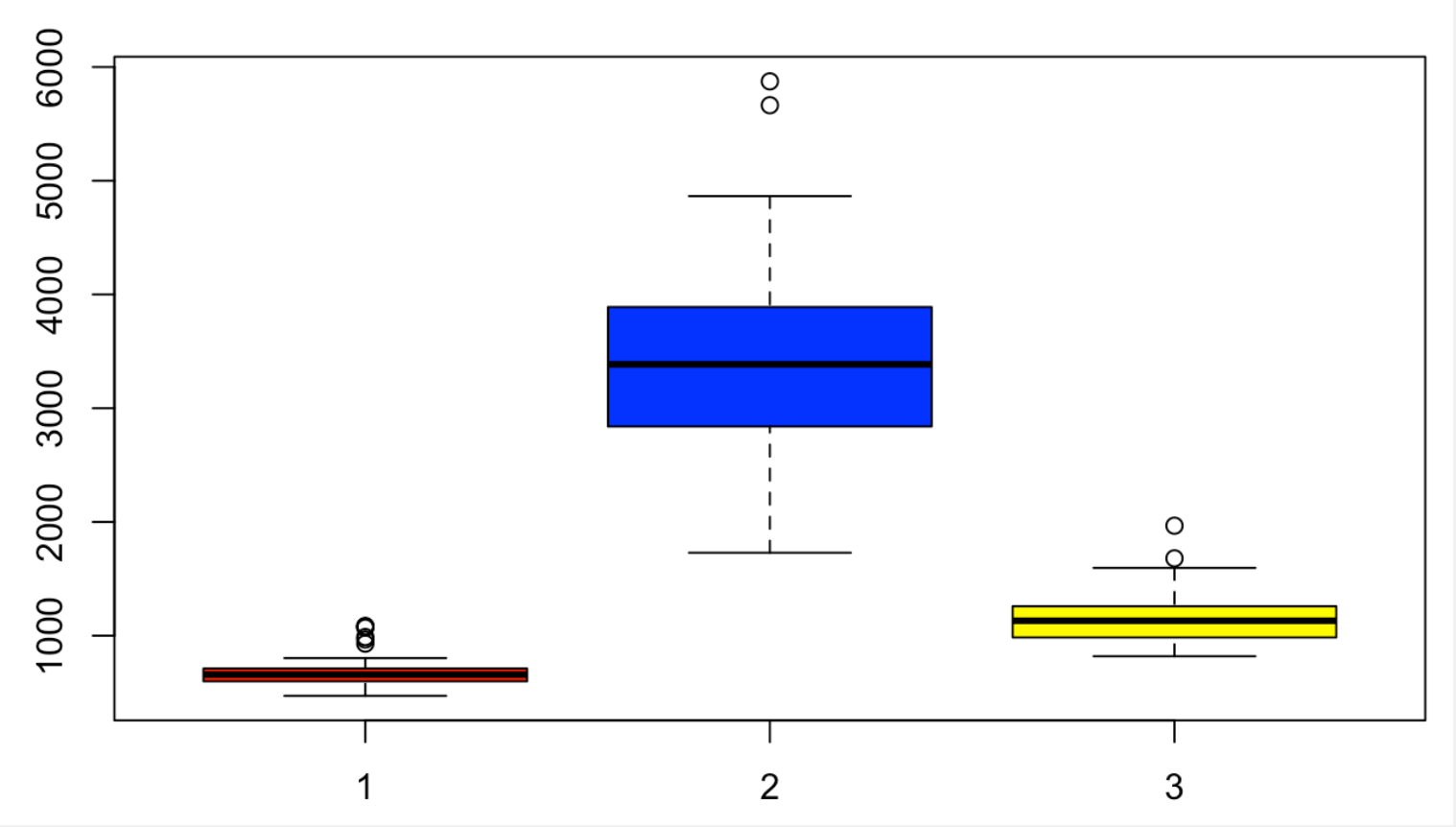
Objectives of current project is to forecast sales of a leading retailer in USA in 2016 based on the sales history of each category (i.e. 2009 - 2015). This helps in planning a) budget, b) Investments in period, c) Minimize revenue loss from unavailability

1) The 3 categories are: (Women, Men, Other) (Total 252 records)
For Phase 2 we only need to forecast women retail sales

2) The given data set has following attributes :
a) Year
b) Month
c) Product Category
d) Sales

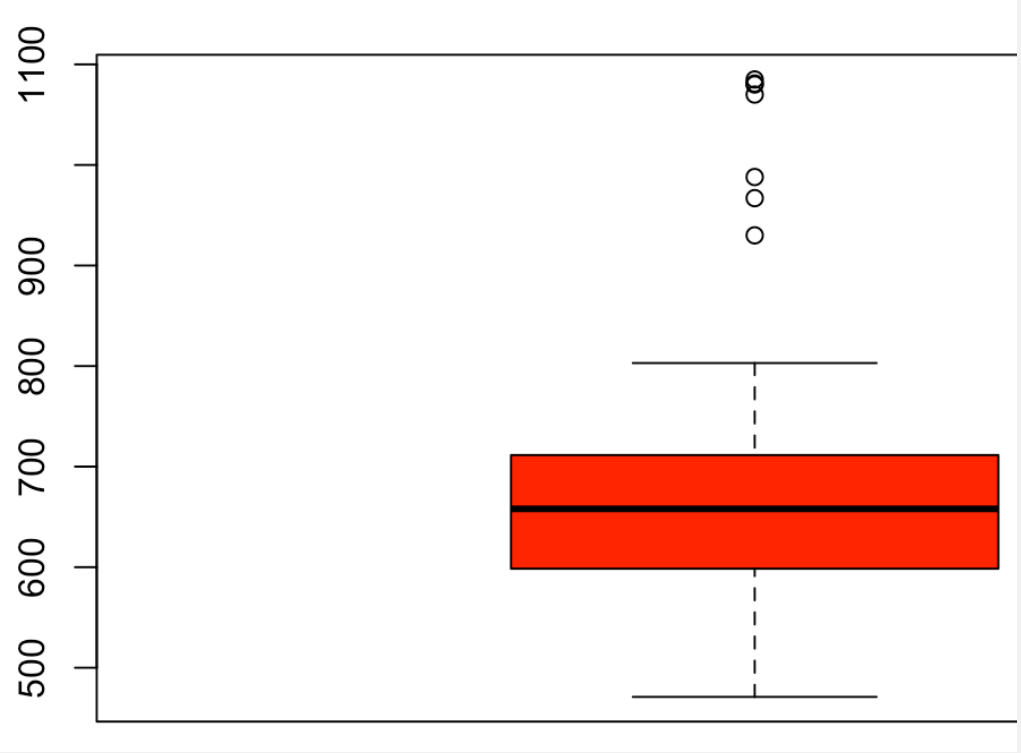
# 1 EDA (TIME SERIES) - PHASE 1

Significance of Dec 2014 & Dec 2015: Women & others have 2 outliers each

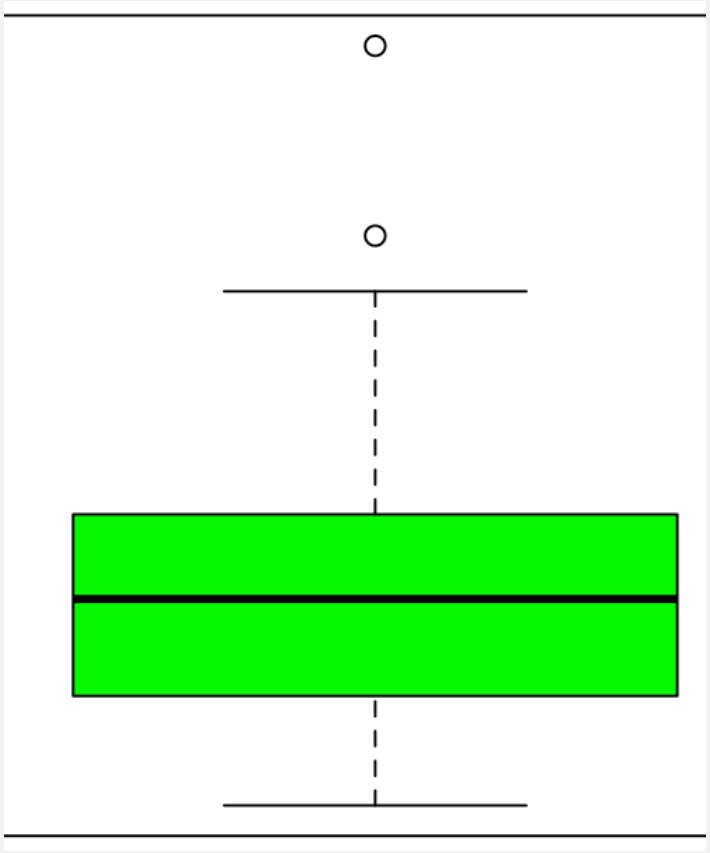


# 1 EDA (TIME SERIES) - PHASE 1

Men have a lot more outliers



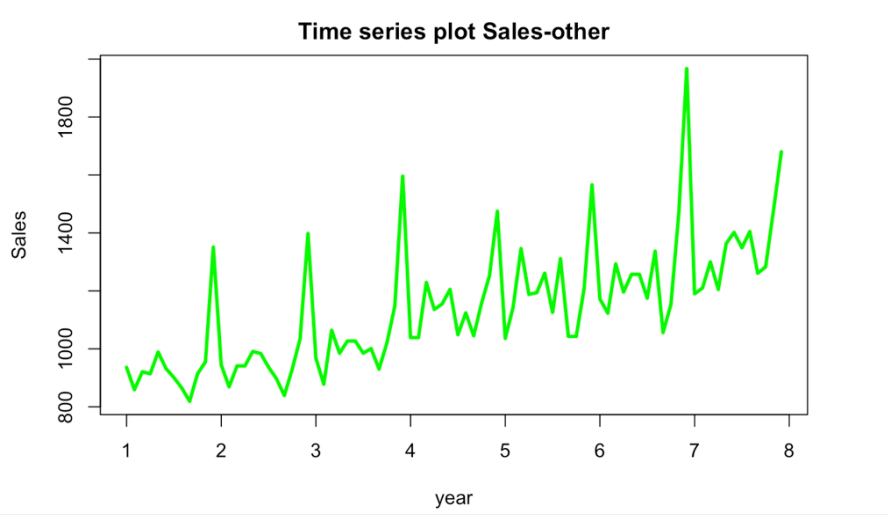

Others, 2 outliers (Dec 2014 & 15)



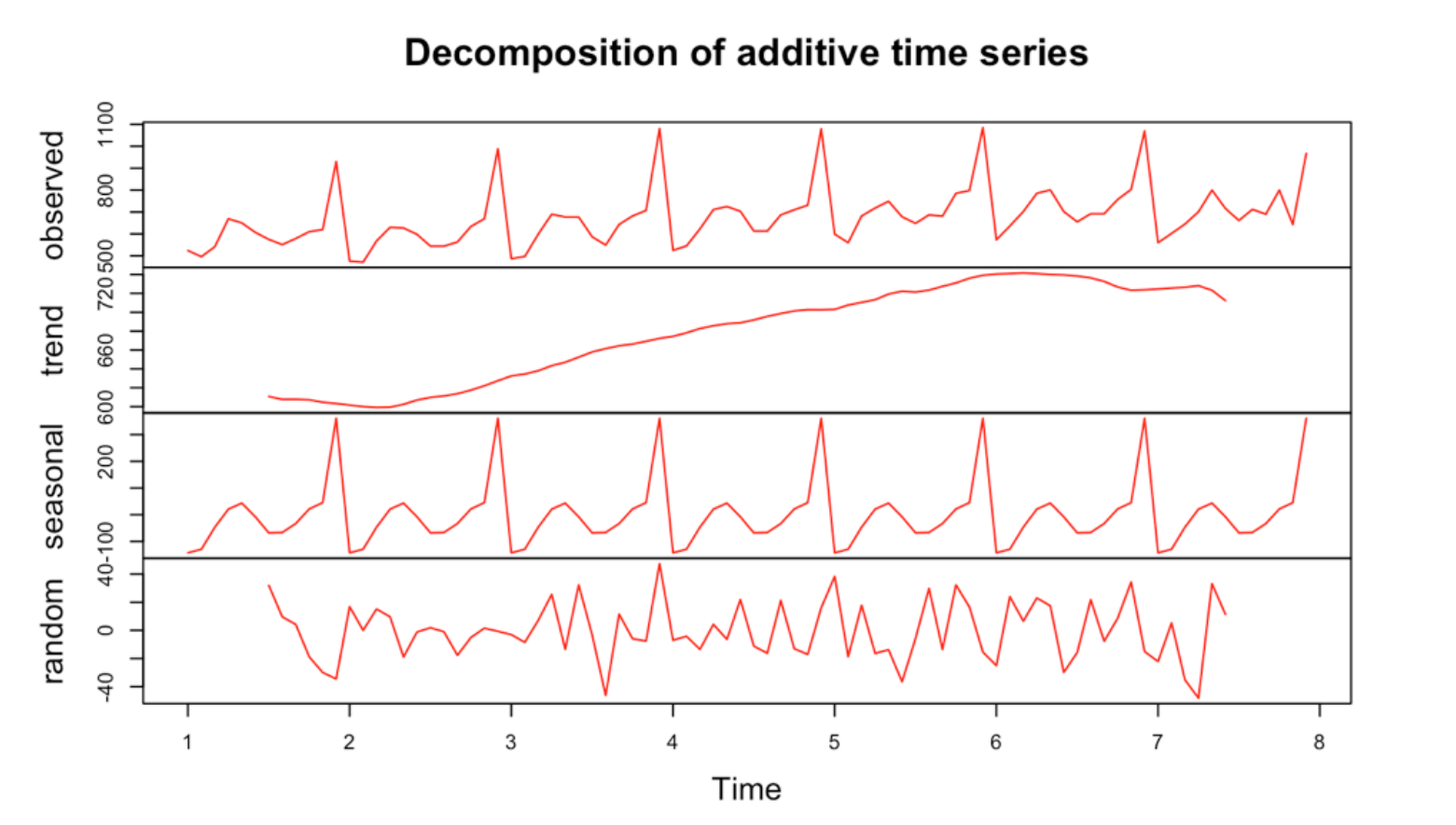
# 1 PREPROCESSING (MEN)-PHASE 1

There are 13 missing values, used "no.locf" and/or "mean from months" to impute

# 2 TIME SERIES (OTHER, WOMEN)-PHASE 1




# 4 DECOMPOSED TIME SERIES - PHASE 1



## 2 EXCELLENT TREND & SEASONALITY - PHASE 1

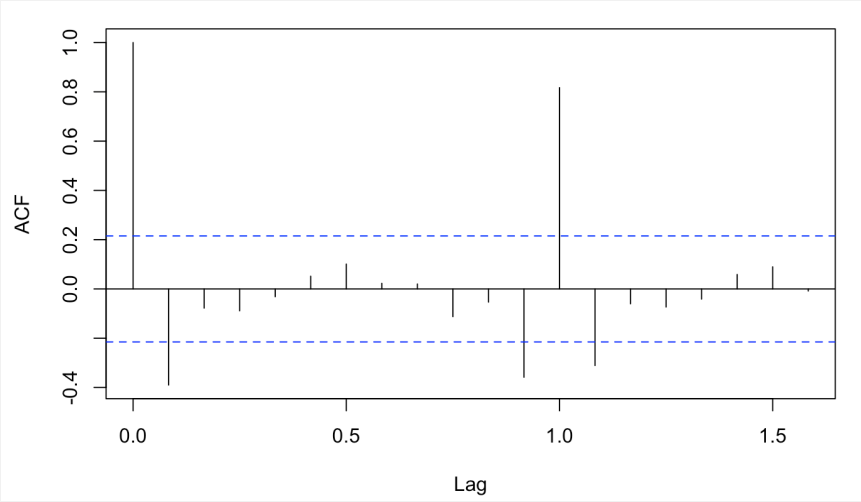
From the above decomposition we can clearly see that there is (across men, women & other): (~Closer to Multiplicative Model)

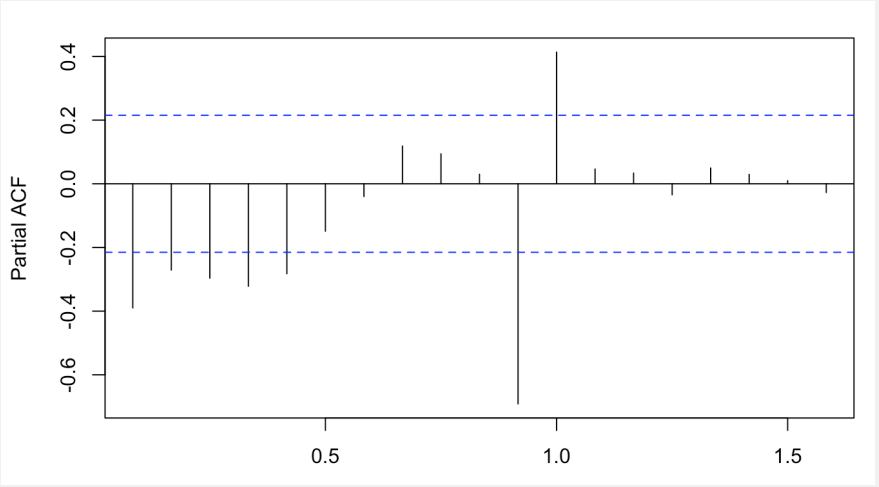
a) very good trend (increasing as years progress),
b) very good seasonality (March, Oct, Nov are peaks)
We differentiate the Time Series on all 3 product to make it stationary

# 3 ACF & PACF PLOTS - PHASE 1

Plot ACF & PACF values for each category and experiment with p, d, q & P, D, Q terms (Ideal trend: decreasing ACF; 1-2 Pacf, Ideal seasonality (Cyclicity in acf, lag in pacf (+/-))

## Men - ACF & PACF Plots



# 3 ACF & PACF PLOTS- PHASE 1

Plot ACF & PACF values for each category and experiment with p, d, q & P, D, Q terms

```
#Depending on the acf and pacf plots we derive p, q and P,Q values along with stationarizing once
womenArima <- Arima(womenTs, order = c(2,1,2), seasonal = c(0,0,1))
womenArima

menArima <- Arima(menTs, order = c(0,0,1), seasonal = c(0,0,1))
menArima

otherArima <- Arima(otherTs, order = c(3,1,2), seasonal = c(2,1,1))
otherArima
```

# 5 MODEL BUILT & EVALUATION - PHASE 1

| | Models Built | Val | Test | Grader Score |
|---|---|---|---|---|
| 1 | Time Series | 12.03% | 12.50% | 75% |
| 2 | Auto Arima Lower | 5.50% | 9.56% | 104.57% |
| 3 | Auto Arima Mean | 10% | 12% | 79.42% |

1) Overall improvement from 79.42% to 104.57%
2) Impute using Median/Mean produced no change

# 6 REGRESSION - PHASE 2

Objectives: Forecast women sales (84 rows, 4 NA's) in 2016 given 3 additional data sets:

1) Holiday: Aggregated Holiday month wise 81 rows & 22 columns ('0' NA's)

2) Macro Economic ('0' NA's)

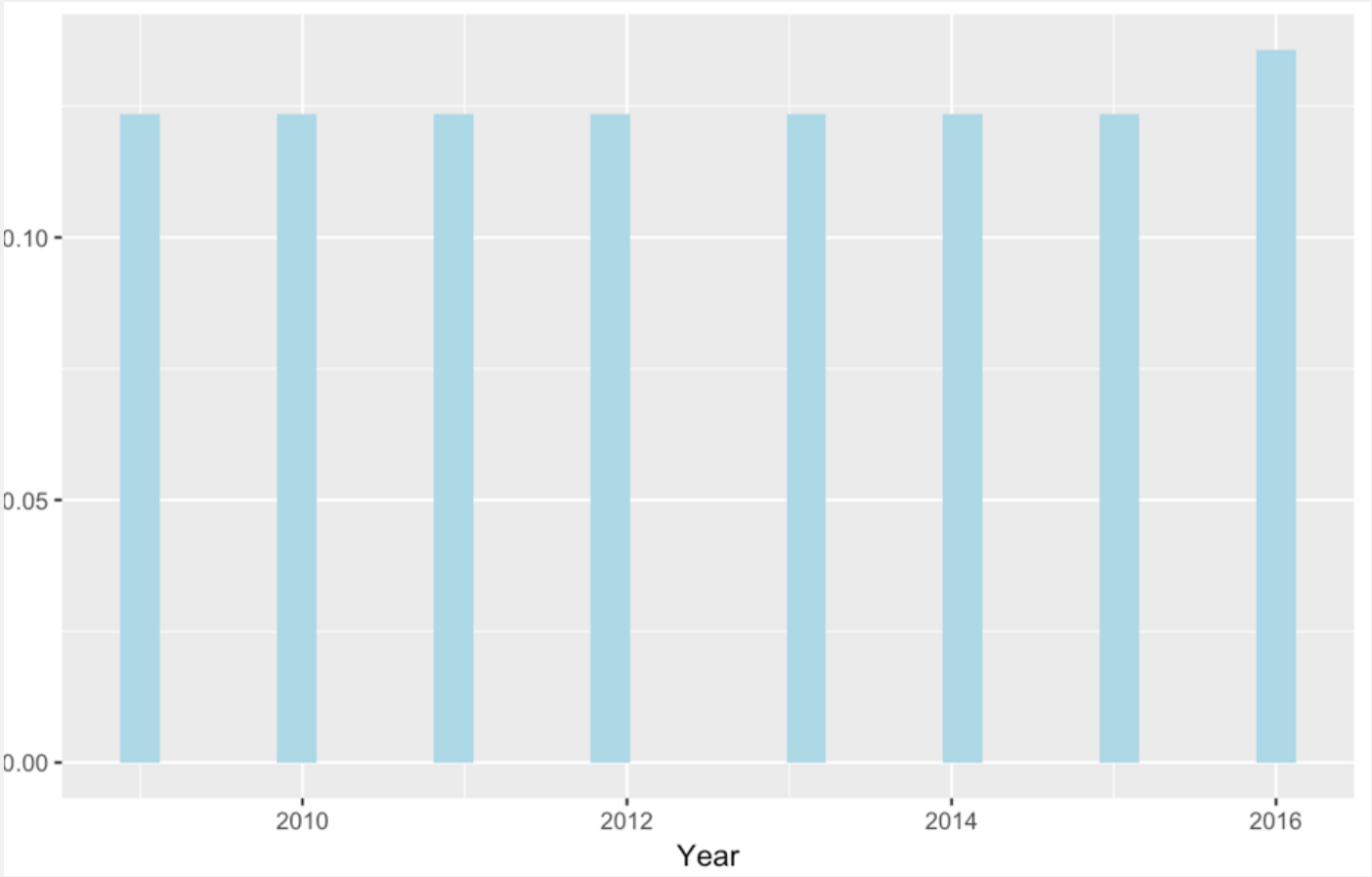
3) Weather (225 NA's)

# 7 HOLIDAY (EDA) - PHASE 2

March and August have no holidays!

# 8 HOLIDAY (EDA) - PHASE 2

2016 has more holidays than the rest!



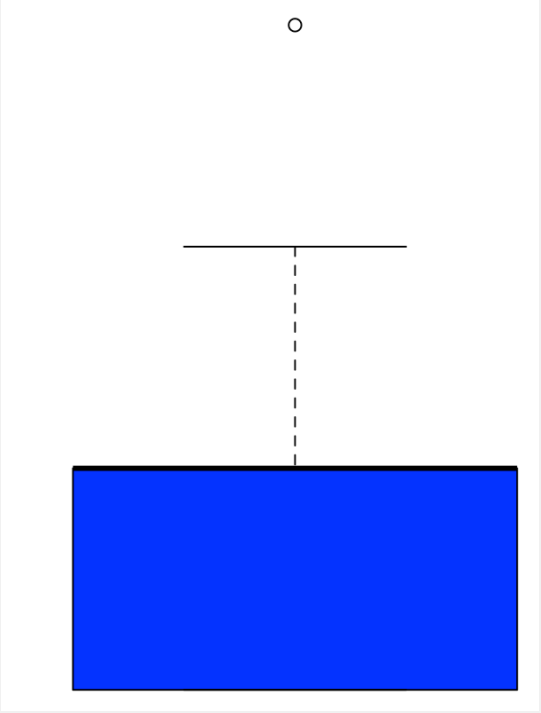
# 8 HOLIDAY (EDA) - PHASE 2

Investigating events & holidays even more - Does Valentine, Mother's day have more sales?

# 8 HOLIDAY (EDA) - PHASE 2

Investigating events & holidays even more - Does Valentine, Mother'd day have more sales? Not much

Only Independence Day, New Year, etc have a few outliers, that too very sparse



# 8 HOLIDAY (PRE-PROCESSING) - PHASE 2

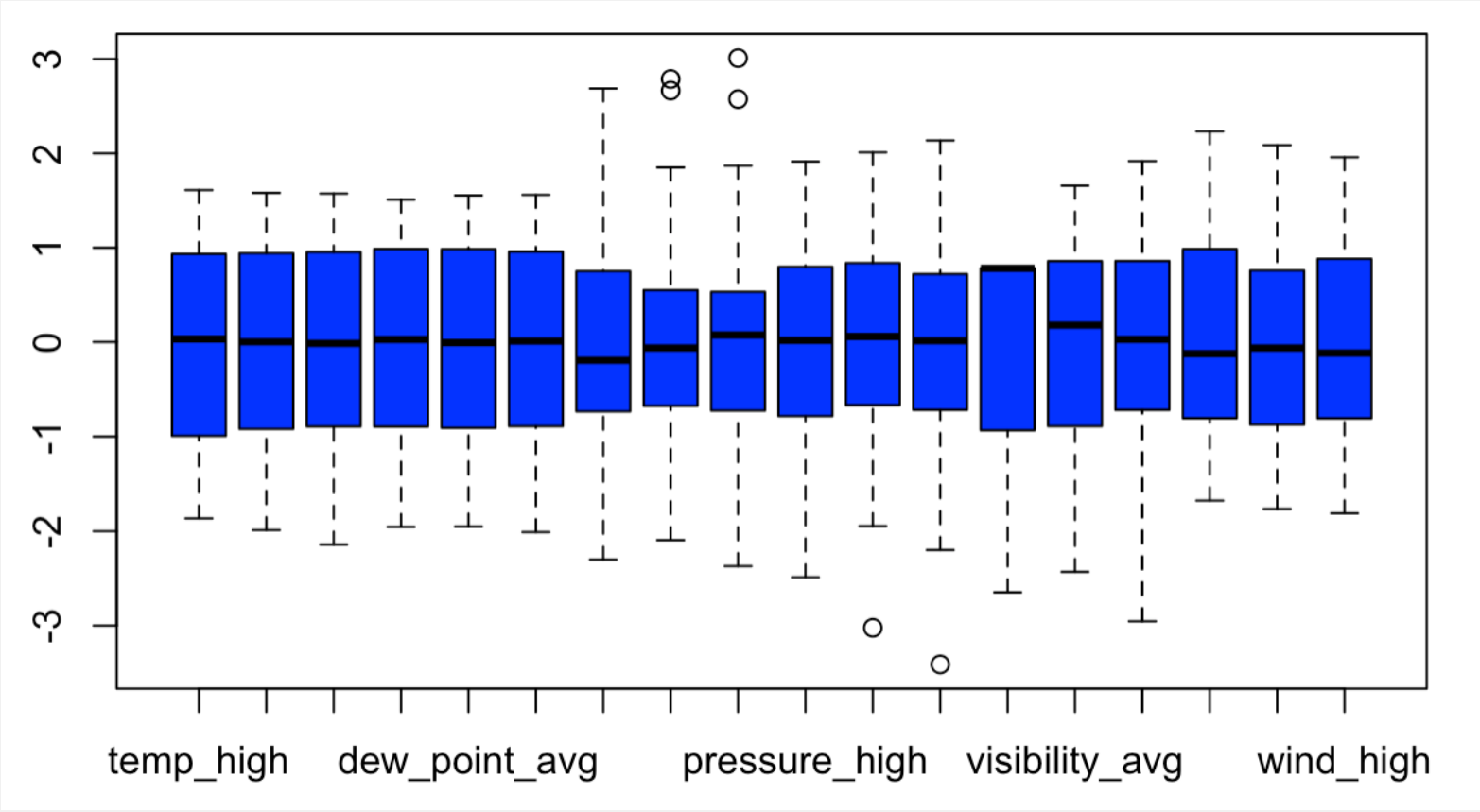
Steps taken to pre-process holiday:

1) Created dummies for every variable, imputed missing holidays with '0'
2) Aggregated by 'Year' and 'Month'
3) Merged using above key

```
# Convert to dummies all Events
hol <- as.data.frame(model.matrix(~ Event - 1, data = holiday))
holiday<-cbind(holiday[-3], hol)
```

# 9 WEATHER (EDA) - PHASE 2

In weather the following had outliers: humidity_avg, humidity_low



# 9 WEATHER (PRE-PROCESSING) - PHASE 2

Pre-processed in .py

1) Cleaned years columns (viz. 2010 - 2016); & 2014 (had incorrect columns)

2) Impute median for NA's

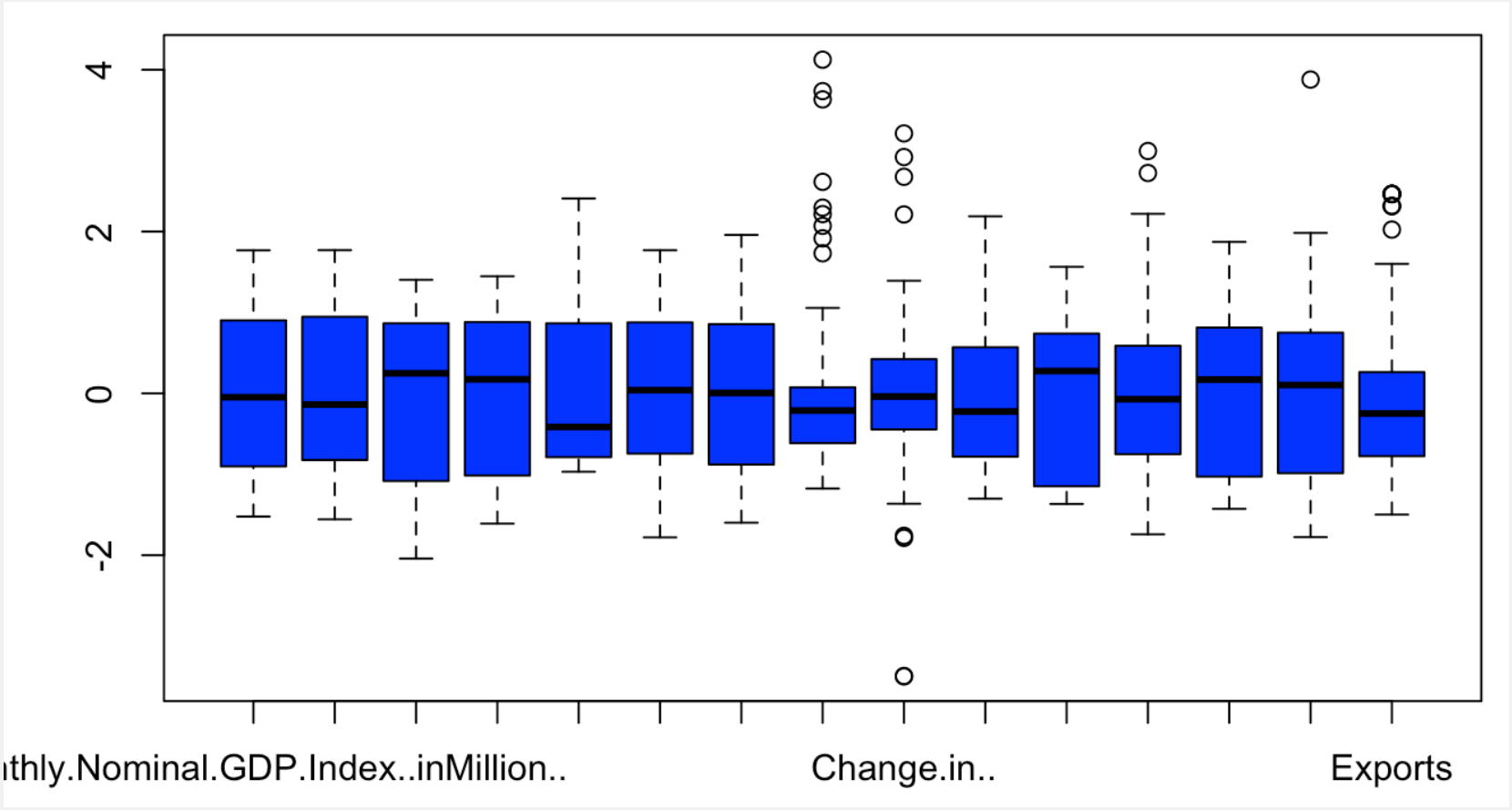
3) Aggregated by 'Year' & 'Month' using 'mean' and merged

In [768]:
```
# Process weather
def process_weather(df):
    # Rename columns for ease
    df.columns = ['year', 'month','day','temp_high','temp_
                  'dew_point_low','humidity_high','humidit
                 'visibility_high','visibility_avg','visib
    # Convert year and day into strings
    df[['year', 'day','month','precip_sum','weatherevent']

    # Convert all other columns to numerics
    df[['temp_high','temp_avg','temp_low','dew_point_high'
                  'dew_point_low','humidity_high','humidit
                 'visibility_high','visibility_avg','visib
                  'dew_point_low','humidity_high','humidit
                 'visibility_high','visibility_avg','visib

    # Fill missing values with median of that column # Try
    df = df.fillna(df.median())
    return df
```

In [769]:
```
def aggregate_weather(df):
    #Convert year to int
    df['year'] = df['year'].astype(int)
    # Group by month all the weather attributes
    df_aggregate_weather = df.groupby('month').mean()
```

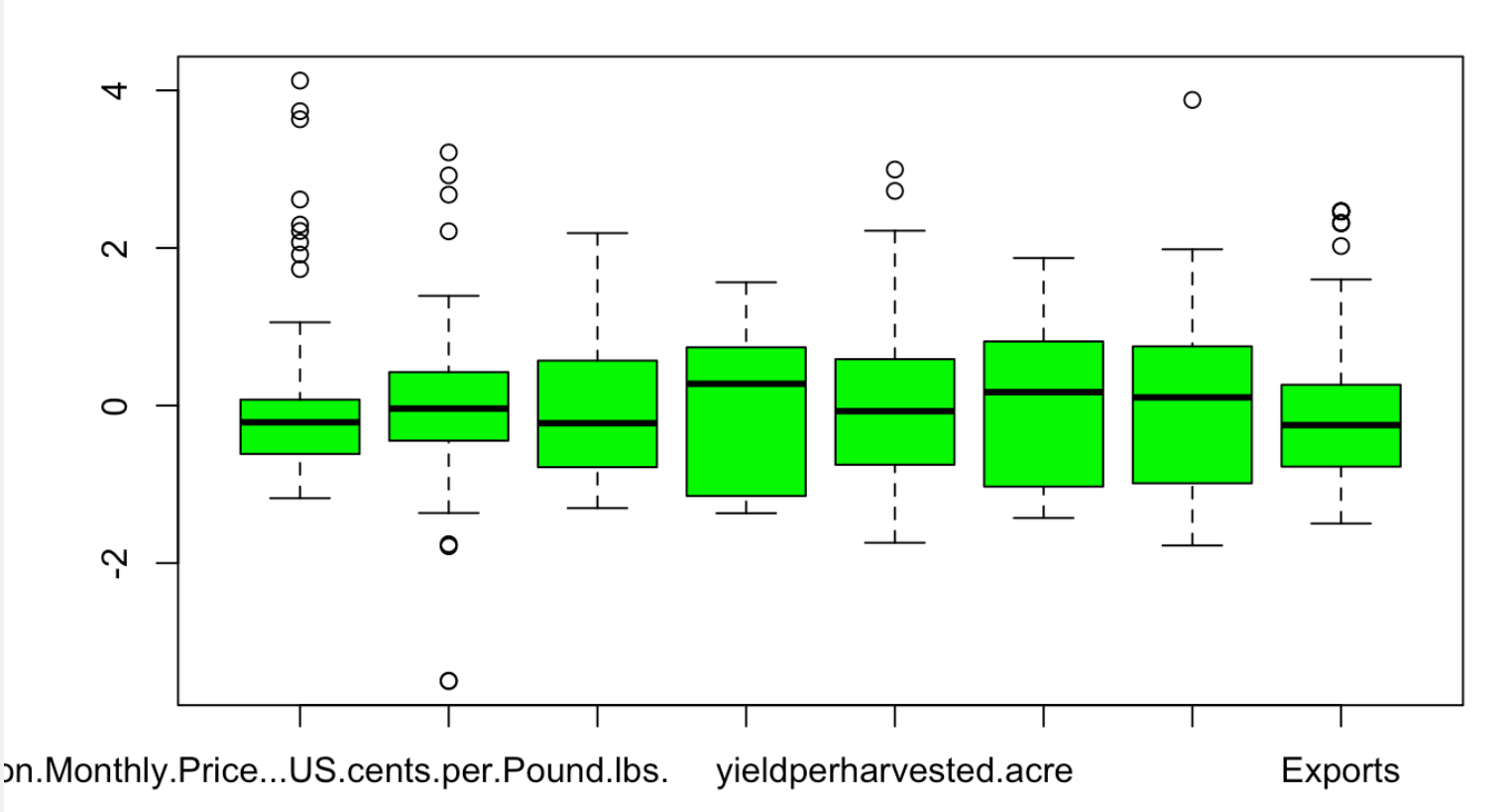
# 9 MACRO-ECONOMIC (EDA) - PHASE 2

In economic data the following had outliers:



# 9 MACRO-ECONOMIC (EDA) - PHASE 2

Heavy outliers in Cotton monthly price, Change in cotton monthly price, yield/harvested, Exports



# 9 MACRO-ECONOMIC (EDA) - PHASE 2

Correlation Plots

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0.96 | -0.97 | -0.69 | -0.88 | 0.99 | -0.27 | -0.17 | -0.15 | | -0.17 | -0.04 | 0.19 | -0.48 |
| 1 | 1 | 0.95 | -0.96 | -0.66 | -0.87 | 0.99 | -0.27 | -0.16 | -0.16 | 0.02 | -0.18 | -0.03 | 0.19 | -0.47 |
| 0.96 | 0.95 | 1 | -0.9 | -0.8 | -0.85 | 0.94 | -0.19 | -0.24 | | 0.05 | -0.1 | | 0.13 | -0.48 |
| -0.97 | -0.96 | -0.9 | 1 | 0.7 | 0.86 | -0.96 | 0.38 | 0.16 | 0.26 | 0.11 | 0.18 | 0.15 | -0.3 | 0.54 |
| -0.69 | -0.66 | -0.8 | 0.7 | 1 | 0.58 | -0.66 | 0.27 | 0.33 | -0.27 | | 0.04 | | -0.28 | 0.51 |
| -0.88 | -0.87 | -0.85 | 0.86 | 0.58 | 1 | -0.88 | 0.32 | | 0.22 | 0.15 | | 0.15 | | 0.44 |
| 0.99 | 0.99 | 0.94 | -0.96 | -0.66 | -0.88 | 1 | -0.26 | -0.12 | -0.18 | | -0.19 | -0.08 | 0.15 | -0.48 |
| -0.27 | -0.27 | -0.19 | 0.38 | 0.27 | 0.32 | -0.26 | 1 | 0.13 | 0.34 | 0.52 | 0.24 | 0.54 | | 0.72 |
| -0.17 | -0.16 | -0.24 | 0.16 | 0.33 | | -0.12 | 0.13 | 1 | -0.23 | | 0.15 | | -0.27 | 0.3 |
| -0.15 | -0.16 | | 0.26 | -0.27 | 0.22 | -0.18 | 0.34 | -0.23 | 1 | 0.6 | | 0.57 | | 0.19 |

# 9 MACRO-ECONOMIC (PRE-PROCESSING) - PHASE 2

1) Standardized, dropped "Adv" and "PartyInPower"

2) Aggregated by 'Year' & 'Month' using 'mean' and merged

```
##Split Year-Month column in economic inorder to join with sales dat
economic<-read_excel("MacroEconomicData.xlsx")

economic$`Year-Month` <- as.character(economic$`Year-Month`)
#Retreive only the year
economic$Year <- as.numeric(substring(economic$`Year-Month`, 1,4))
#Retreive only the month
economic$Month <- substring(economic$`Year-Month`, 8,10)

# Convert the character month into numbers
economic$Month<-ifelse(economic$Month == 'Jan', '1',
                       ifelse(economic$Month == 'Feb', '2',
                            ifelse(economic$Month == 'Mar', '3',
                                   ifelse(economic$Month == 'Apr', '
```

# 9 REGRESSION - PHASE 2

PCA on all the merged 58 features:

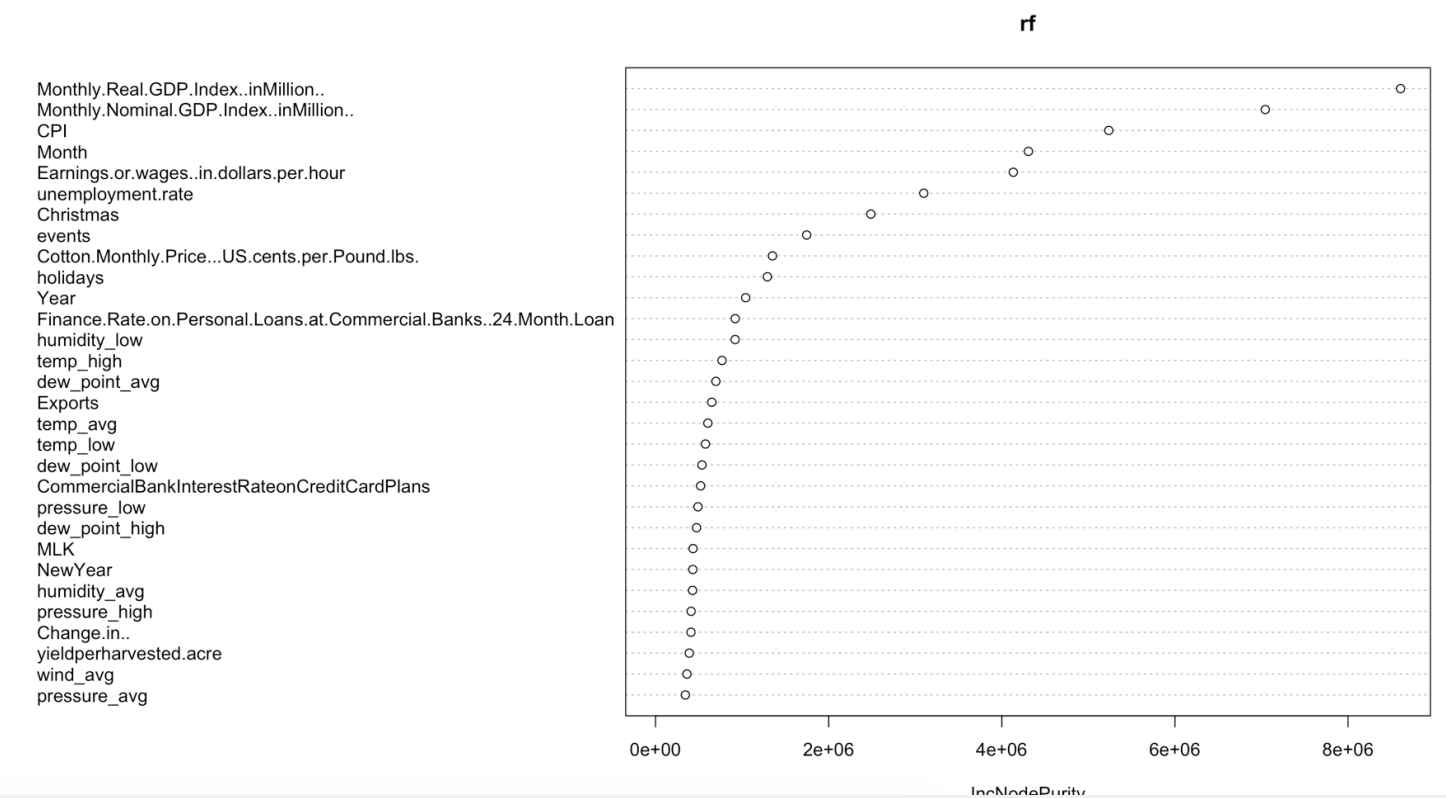
20 components accounted for over 99% of the deviation

Train - Until 2014 Dec, Val - 2015 (12 months)

```
> summary(pca)
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
Standard deviation     2.6181052 1.82110640 1.79222049 1.64818732 1.50160238 1.47471477 1.45023561 1.39416628 1.2335795 1.13582682
Proportion of Variance 0.1982017 0.09589673 0.09287868 0.07855002 0.06519932 0.06288531 0.06081494 0.05620336 0.0440015 0.03730417
Cumulative Proportion  0.1982017 0.29409841 0.38697709 0.46552711 0.53072642 0.59361173 0.65442667 0.71063003 0.7546315 0.79193570
                          Comp.11    Comp.12    Comp.13    Comp.14    Comp.15    Comp.16   Comp.17    Comp.18    Comp.19     Comp.20
Standard deviation     1.08456789 1.05392649 1.01990741 0.90661837 0.82915083 0.78466266 0.6842588 0.64987019 0.59315494 0.441950689
Proportion of Variance 0.03401313 0.03211839 0.03007839 0.02376743 0.01987926 0.01780324 0.0135386 0.01221199 0.01017348 0.005647819
Cumulative Proportion  0.82594883 0.85806723 0.88814562 0.91191305 0.93179231 0.94959555 0.9631342 0.97534614 0.98551962 0.991167442
                           Comp.21     Comp.22      Comp.23      Comp.24      Comp.25      Comp.26      Comp.27      Comp.28      Comp.29
Standard deviation     0.369898598 0.24853779 0.226484749 0.1479733497 0.1269359447 0.092420021 0.0769736714 5.442568e-02 1.142576e-02
Proportion of Variance 0.003956385 0.00178615 0.001483239 0.0006331406 0.0004659104 0.000246982 0.0001713237 8.565268e-05 3.774884e-06
Cumulative Proportion  0.995123827 0.99690998 0.998393216 0.9990263563 0.9994922667 0.999739249 0.9999105724 9.999962e-01 1.000000e+00
                           Comp.30      Comp.31 Comp.32 Comp.33 Comp.34 Comp.35
Standard deviation     2.309164e-08 1.826417e-08       0       0       0       0
Proportion of Variance 1.541852e-17 9.645683e-18       0       0       0       0
Cumulative Proportion  1.000000e+00 1.000000e+00       1       1       1       1
```

# 10 MODELS RAN & THEIR IMPORTANCE - PHASE 2

- Random Forest feature importance



# 11 EDA - PHASE 2

- Xgboost feature importance

| | Feature | Gain | Cover | Frequency |
|---|---|---|---|---|
| 1 | target | 3.698252e-01 | 0.089547038 | 0.10294118 |
| 2 | Month | 3.427538e-01 | 0.294425087 | 0.19117647 |
| 3 | Monthly.Nominal.GDP.Index..inMillion.. | 9.161355e-02 | 0.044250871 | 0.02941176 |
| 4 | Monthly.Real.GDP.Index..inMillion.. | 5.717584e-02 | 0.143902439 | 0.10294118 |
| 5 | CPI | 2.890849e-02 | 0.026829268 | 0.01470588 |
| 6 | dew_point_high | 1.838407e-02 | 0.061672474 | 0.04411765 |
| 7 | Cotton.Monthly.Price...US.cents.per.Pound.lbs. | 1.508777e-02 | 0.023693380 | 0.07352941 |
| 8 | wind_low | 1.459842e-02 | 0.017770035 | 0.01470588 |
| 9 | Earnings.or.wages..in.dollars.per.hour | 1.406387e-02 | 0.042508711 | 0.04411765 |
| 10 | Year | 9.407298e-03 | 0.007317073 | 0.04411765 |
| 11 | dew_point_avg | 9.308300e-03 | 0.026829268 | 0.01470588 |
| 12 | holidays | 7.490625e-03 | 0.020557491 | 0.04411765 |
| 13 | Finance.Rate.on.Personal.Loans.at.Commercial.Banks..24.Month.Loan | 5.049777e-03 | 0.016027875 | 0.01470588 |
| 14 | humidity_avg | 4.452229e-03 | 0.046341463 | 0.04411765 |
| 15 | MLK | 2.454943e-03 | 0.037979094 | 0.02941176 |
| 16 | Halloween | 2.448335e-03 | 0.011846690 | 0.01470588 |
| 17 | Mill.use...in..480.lb.netweright.in.million.bales. | 2.364883e-03 | 0.008710801 | 0.01470588 |
| 18 | CommercialBankInterestRateonCreditCardPlans | 1.632521e-03 | 0.012543554 | 0.04411765 |
| 19 | events | 1.386870e-03 | 0.011149826 | 0.01470588 |
| 20 | pressure_avg | 6.000156e-04 | 0.007317073 | 0.01470588 |
| 21 | Average.upland.harvested.million.acres. | 4.149352e-04 | 0.009756098 | 0.02941176 |
| 22 | visibility_avg | 2.792919e-04 | 0.017073171 | 0.01470588 |
| 23 | pressure_low | 1.907293e-04 | 0.010104530 | 0.01470588 |
| 24 | Production..in..480.lb.netweright.in.million.bales. | 8.810189e-05 | 0.008362369 | 0.01470588 |
| 25 | unemployment.rate | 2.012810e-05 | 0.003484321 | 0.01470588 |

# 12 MODEL BUILT AND EVALUATION

| S.No. | Algorithms | Val | Test | Grader Score |
|---|---|---|---|---|
| 1 | Linear Regression with Step AIC & VIF | NA | 30.97% | 32.28% |
| 2 | Decision Tree | 11.12% | 15.20% | 65.76% |
| 3 | XGB | 13.67% | 11.02% | 90.69% |
| 4 | PCA then ARIMA | 11.67% | 8.50% | 118.50% |
| 5 | H2O with tuning (Used cluster for this) | 9.67% | 9.63% | 103.79% |
| 6 | Ensemble | 12.40% | 10.50% | 95% |

Overall score increased from 104% to 118% Grader score

## 13 IMPROVISATIONS TO MODELS WITH TIME

1) Aggregate weather data using max, min and average (instead of average) OR histograms

2) Novel methods to Feature engineering viz. splitting looking at the patterns, binning few of the features

3) The Domain knowledge would have made more accurate models

**THANK YOU FOR YOUR TIME!**