Satish_SVAP_Asmt

Satish Kaushik

7/15/2017

Frame and Acquisition of Data

I have choosen to scrape data from NHRFD on a Day basis. Collected 5 years data from year 2012 to year 2016.

```
library(rvest)
```

```
## Loading required package: xml2
```

```
pg.out = read_html('../MonthWiseMarketArrivals_Potato.html')
pg.table = pg.out %>%
           html_node('#dnn_ctr974_MonthWiseMarketArrivals_GridView1') %>%
           html_table()
df = pg.table
str(df)
```

```
## 'data.frame':    3186 obs. of  7 variables:
##  $ Market              : chr  "ABOHAR(PB)" "ABOHAR(PB)" "ABOHAR(PB)" "ABOHAR(PB)" ...
##  $ Month Name          : chr  "January" "January" "January" "January" ...
##  $ Year                : chr  "2012" "2013" "2014" "2015" ...
##  $ Arrival (q)         : int  3800 1790 1910 5940 1250 2900 2875 4725 1225 2580 ...
##  $ Price Minimum (Rs/q): chr  "222" "410" "550" "395" ...
##  $ Price Maximum (Rs/q): chr  "373" "718" "1014" "775" ...
##  $ Modal Price (Rs/q)  : chr  "289" "605" "901" "594" ...
```

Refine

- Rename the column names

```
newnames = c('market', 'month', 'year', 'quantity', 'priceMin', 'priceMax', 'priceMod' )
colnames(df) = newnames
str(df)
```

```
## 'data.frame':    3186 obs. of  7 variables:
##  $ market  : chr  "ABOHAR(PB)" "ABOHAR(PB)" "ABOHAR(PB)" "ABOHAR(PB)" ...
##  $ month   : chr  "January" "January" "January" "January" ...
##  $ year    : chr  "2012" "2013" "2014" "2015" ...
##  $ quantity: int  3800 1790 1910 5940 1250 2900 2875 4725 1225 2580 ...
##  $ priceMin: chr  "222" "410" "550" "395" ...
##  $ priceMax: chr  "373" "718" "1014" "775" ...
##  $ priceMod: chr  "289" "605" "901" "594" ...
```

- Remove last row which contains Total details.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
tail(df)
```

```
##              market     month  year  quantity priceMin  priceMax  priceMod
## 3181 VIJAYAWADA(AP)    August  2016       770     2340      2540      2440
## 3182 VIJAYAWADA(AP) September  2016       910     1967      2167      2067
## 3183 VIJAYAWADA(AP)  November  2016       150     1800      2000      1900
## 3184 VIJAYAWADA(AP)  December  2015       160     1500      1700      1600
## 3185 VIJAYAWADA(AP)  December  2016      1070     1171      1371      1271
## 3186                          Total 183021721 851(Avg) 1160(Avg) 1015(Avg)
```

```
df = df %>%
  filter(year != "Total")
tail(df)
```

```
##              market     month year quantity priceMin priceMax priceMod
## 3180 VIJAYAWADA(AP)    August 2015      150     1300     1500     1400
## 3181 VIJAYAWADA(AP)    August 2016      770     2340     2540     2440
## 3182 VIJAYAWADA(AP) September 2016      910     1967     2167     2067
## 3183 VIJAYAWADA(AP)  November 2016      150     1800     2000     1900
## 3184 VIJAYAWADA(AP)  December 2015      160     1500     1700     1600
## 3185 VIJAYAWADA(AP)  December 2016     1070     1171     1371     1271
```

- Change the respective data types

```
df$year = as.numeric(df$year)
df$priceMin = as.numeric(df$priceMin)
df$priceMax = as.numeric(df$priceMax)
df$priceMod = as.numeric(df$priceMod)
str(df)
```

```
## 'data.frame':    3185 obs. of  7 variables:
##  $ market  : chr  "ABOHAR(PB)" "ABOHAR(PB)" "ABOHAR(PB)" "ABOHAR(PB)" ...
##  $ month   : chr  "January" "January" "January" "January" ...
##  $ year    : num  2012 2013 2014 2015 2012 ...
##  $ quantity: int  3800 1790 1910 5940 1250 2900 2875 4725 1225 2580 ...
##  $ priceMin: num  222 410 550 395 227 368 466 336 283 398 ...
##  $ priceMax: num  373 718 1014 775 396 ...
##  $ priceMod: num  289 605 901 594 304 531 709 546 346 547 ...
```

- Create the date column

```
head(df)
```

```
##       market    month year quantity priceMin priceMax priceMod
## 1 ABOHAR(PB)  January 2012     3800      222      373      289
## 2 ABOHAR(PB)  January 2013     1790      410      718      605
## 3 ABOHAR(PB)  January 2014     1910      550     1014      901
## 4 ABOHAR(PB)  January 2015     5940      395      775      594
## 5 ABOHAR(PB) February 2012     1250      227      396      304
## 6 ABOHAR(PB) February 2013     2900      368      603      531
```

```
df = df %>%
  mutate(date = paste("01", month, year, sep="-"))
df$date = as.Date(df$date, "%d-%B-%Y")
str(df)
```

```
## 'data.frame':    3185 obs. of  8 variables:
##  $ market  : chr  "ABOHAR(PB)" "ABOHAR(PB)" "ABOHAR(PB)" "ABOHAR(PB)" ...
##  $ month   : chr  "January" "January" "January" "January" ...
##  $ year    : num  2012 2013 2014 2015 2012 ...
##  $ quantity: int  3800 1790 1910 5940 1250 2900 2875 4725 1225 2580 ...
##  $ priceMin: num  222 410 550 395 227 368 466 336 283 398 ...
##  $ priceMax: num  373 718 1014 775 396 ...
##  $ priceMod: num  289 605 901 594 304 531 709 546 346 547 ...
##  $ date    : Date, format: "2012-01-01" "2013-01-01" ...
```

- Split City/State Names from market column

```
library(stringr)
library(tidyr)
df = df %>%
  mutate(market1 = market) %>%
  separate(market1, c('city', 'state'), sep = "\\(")
```

```
## Warning: Too many values at 2 locations: 2977, 2978
```

```
## Warning: Too few values at 757 locations: 453, 454, 455, 456, 457, 458,
## 459, 460, 461, 462, 463, 464, 465, 466, 467, 468, 469, 470, 471, 472, ...
```

```
head(df, 20)
```

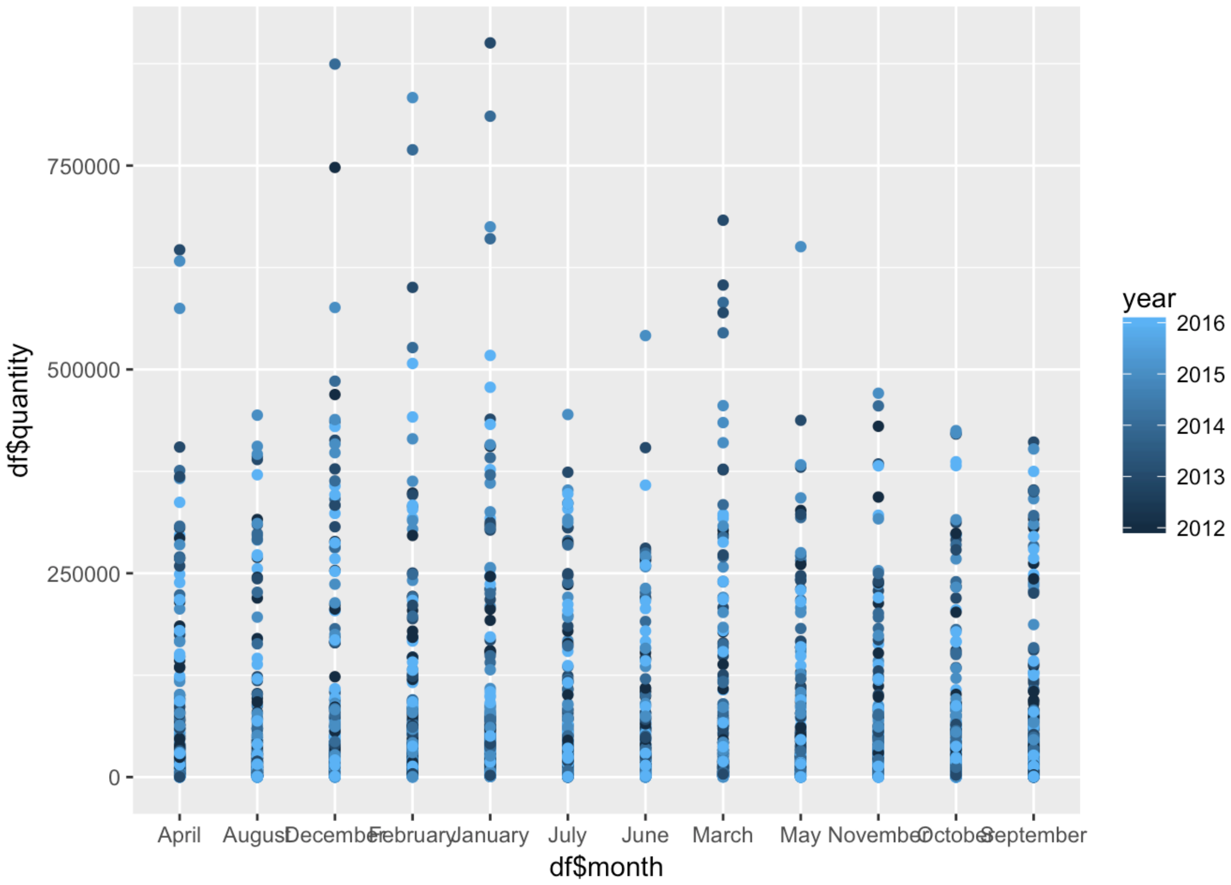
```
##        market    month year quantity priceMin priceMax priceMod       date
## 1  ABOHAR(PB)  January 2012     3800      222      373      289 2012-01-01
## 2  ABOHAR(PB)  January 2013     1790      410      718      605 2013-01-01
## 3  ABOHAR(PB)  January 2014     1910      550     1014      901 2014-01-01
## 4  ABOHAR(PB)  January 2015     5940      395      775      594 2015-01-01
## 5  ABOHAR(PB) February 2012     1250      227      396      304 2012-02-01
## 6  ABOHAR(PB) February 2013     2900      368      603      531 2013-02-01
## 7  ABOHAR(PB) February 2014     2875      466      802      709 2014-02-01
## 8  ABOHAR(PB) February 2015     4725      336      701      546 2015-02-01
## 9  ABOHAR(PB)    March 2012     1225      283      508      346 2012-03-01
## 10 ABOHAR(PB)    March 2013     2580      398      632      547 2013-03-01
## 11 ABOHAR(PB)    March 2014     3860      600      990      866 2014-03-01
## 12 ABOHAR(PB)    March 2015     5000      378      788      619 2015-03-01
## 13 ABOHAR(PB)    April 2012     1830      641      970      802 2012-04-01
## 14 ABOHAR(PB)    April 2013     2165      542      921      732 2013-04-01
## 15 ABOHAR(PB)    April 2014     1465      722     1153      941 2014-04-01
## 16 ABOHAR(PB)    April 2015     5150      186      493      330 2015-04-01
## 17 ABOHAR(PB)      May 2012      505      720     1000      900 2012-05-01
## 18 ABOHAR(PB)      May 2013     1805      633      982      850 2013-05-01
## 19 ABOHAR(PB)      May 2014     1175      878     1384     1133 2014-05-01
## 20 ABOHAR(PB)      May 2015     2850      225      494      353 2015-05-01
##      city state
## 1  ABOHAR   PB)
## 2  ABOHAR   PB)
## 3  ABOHAR   PB)
## 4  ABOHAR   PB)
## 5  ABOHAR   PB)
## 6  ABOHAR   PB)
## 7  ABOHAR   PB)
## 8  ABOHAR   PB)
## 9  ABOHAR   PB)
## 10 ABOHAR   PB)
## 11 ABOHAR   PB)
## 12 ABOHAR   PB)
## 13 ABOHAR   PB)
## 14 ABOHAR   PB)
## 15 ABOHAR   PB)
## 16 ABOHAR   PB)
## 17 ABOHAR   PB)
## 18 ABOHAR   PB)
## 19 ABOHAR   PB)
## 20 ABOHAR   PB)
```

Analyzing the data using plots/graphs

- Plot of month vs quantity, different colors for each year

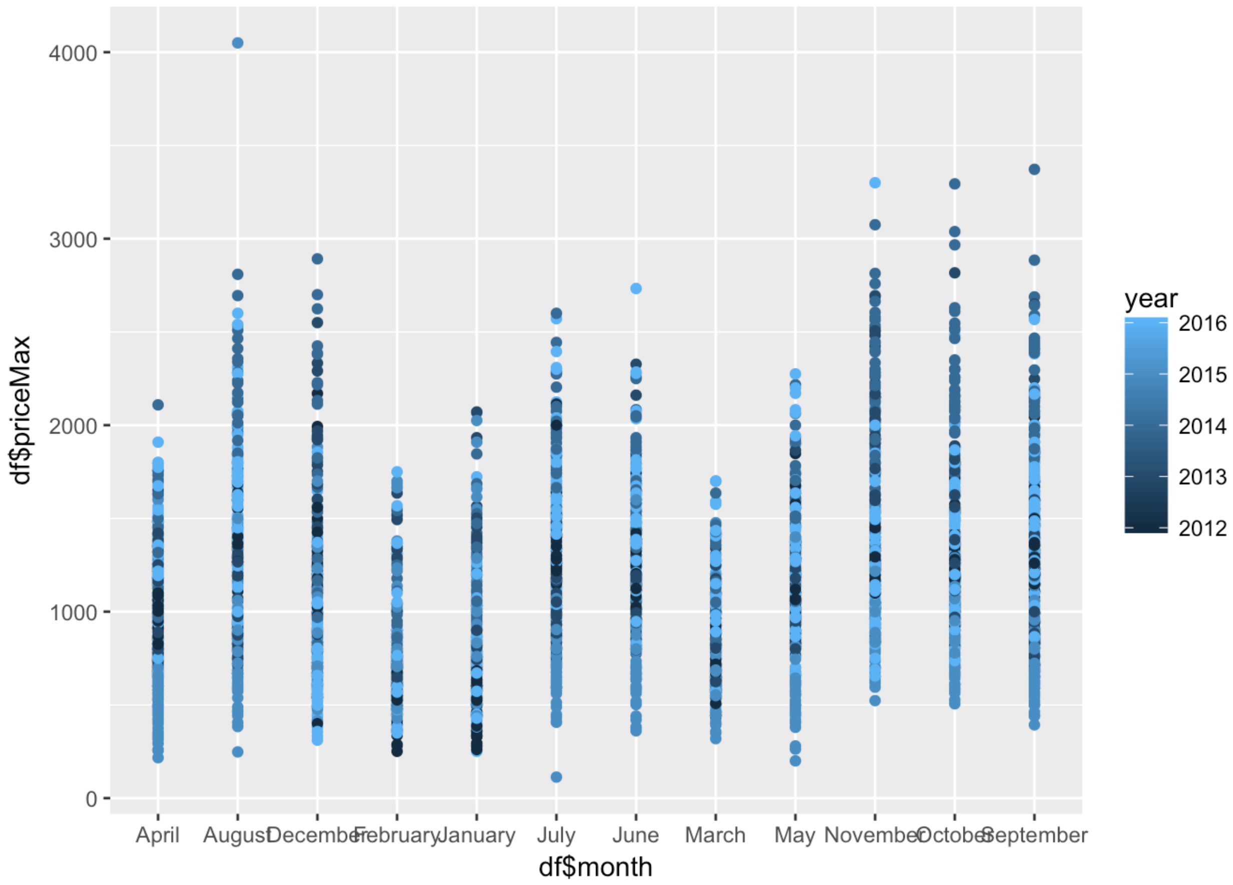
```
library(ggplot2)

g1 = ggplot(df) +
  aes(df$month, df$quantity, color=year) +
  geom_point()
g1
```



- plot of month vs max price, different colors for each year

```
g2 = ggplot(df) +
  aes(df$month, df$priceMax, color=year) +
  geom_point()
g2
```



```r
library(plotly)
```

```
##
## Attaching package: 'plotly'
```

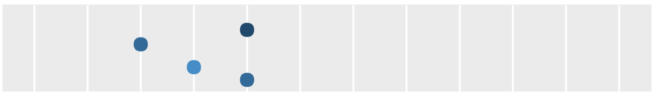
```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

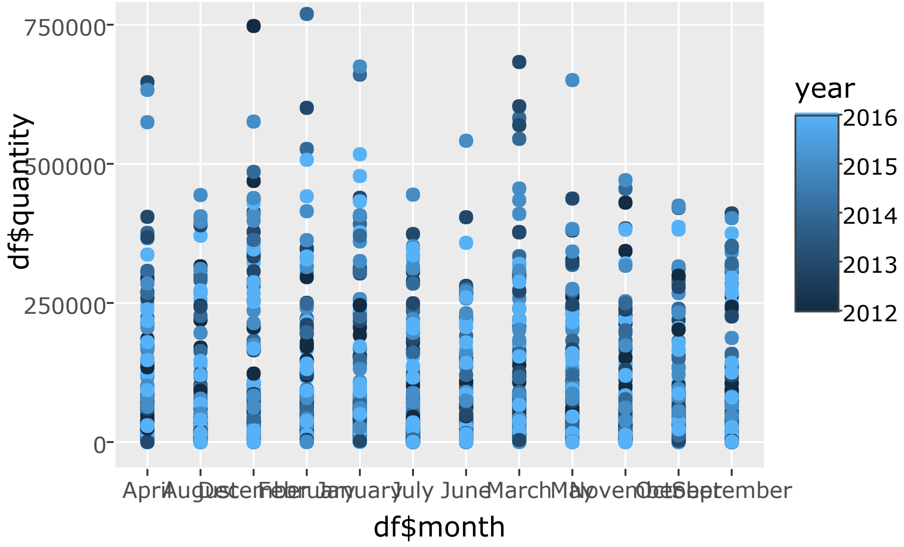
```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```r
ggplotly(g1)
```

```
## We recommend that you use the dev version of ggplot2 with `ggplotly()`
## Install it with: `devtools::install_github('hadley/ggplot2')`
```





#interactive visualization of data

```
library(crosstalk)
library(d3scatter)

shared_rawdata <- SharedData$new(df)

bscols(
  list(
    filter_checkbox("month", "monthSelect", shared_rawdata, ~month, inline = TRUE),
    filter_checkbox("year", "yearSelect", shared_rawdata, ~year, inline = TRUE),
    filter_slider("Quantity", "Quantity", shared_rawdata, ~quantity, width = "100%")
  ),

  d3scatter(shared_rawdata, ~year, ~quantity, ~year, width="100%", height=300),
  d3scatter(shared_rawdata, ~year, ~quantity, ~month, width="100%", height=300)

)
```

monthSelect

☐ April ☐ August ☐ December ☐ February ☐ January ☐ July ☐ June ☐ March ☐ May ☐ November ☐ October ☐ September

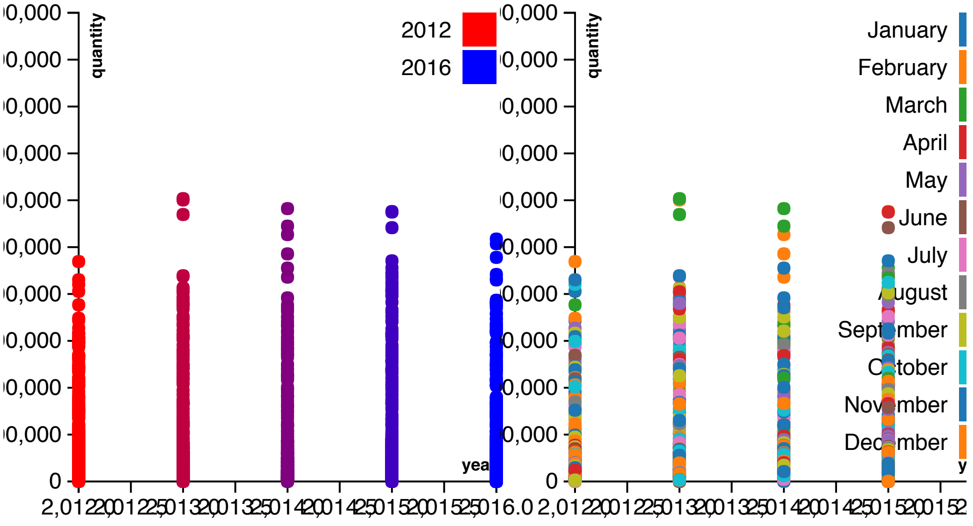
yearSelect

☐ 2012 ☐ 2013 ☐ 2014 ☐ 2015 ☐ 2016

Quantity

5 630,005 900,500

5 100,005 200,005 400,005 600,005 800,005


quantity
yea
quantity
January
February
March
April
May
June
July
August
September
October
November
December
y