Exploratory Data Analysis: On the differences in petrol price between Vietnam and US.

LinhNguyen

Monday, September 05, 2016

Hey guys, thanks for your interest. In this analysis, I'm going to explore the different between petrol price in Vietnam and in US.

For most Vietnamese who are living in Vietnam, we should all have the assumption, that we are paying so much more than our friends out there. But we never know how much...

That's the purpose of this exploratory data analysis, I'm going to take a peek, at what the data has to tell us.

LIBRARY

```
library(ggplot2)
library(tidyr)
library(dplyr)
library(RCurl)
```

THE STRATEGY

So the idea is to let the data speak for itself:

- Since the price of Vietnam's petrol are controled by only one source, it will be best to be represented by a line graph.
- As for the US, there are a lot of people selling petrol, and the price varied from location to location, so the idea is that, I kind of want to let the audiences see the noises, the low, and the high of the prices, while not letting it become too much of a distraction.
- I want to show something that would sum the noise up, ideally a mean line in this case.
- I want to translate the US's price into vietnamese. To do so, I'll have to take in consideration the currency exchange rate of that given time.

To solve this puzzle, I decided to used to data below, which I've uploaded the datasets to github, so no, you don't have to download it.

- The US's petrol data are taken from the US Energy Information Association, available at: http://www.eia.gov/petroleum/gasdiesel/
- The Vietnam's petrol data are gathered from petrolimex vietnam, available at: http://www.petrolimex.com.vn/
- The currency exchange rate history are taken from oanda, available at: https://www.oanda.com/currency/average

THE DATASETS

The datasets are quite clean already, some preprocess (reformat) and we're good to go.

There are a lot type of petrol availabe. I choosed Ron91 - Ron92 as the main data type for this analysis, because of popularity and the availability of the data.

```
# Get the data from github
oilvn <- "https://raw.githubusercontent.com/Linhnguyen14/stick/master/oilprice/gasolinevn.csv"
oilvn <- read.csv(text = getURL(oilvn))

# Convert the format of the date column, from "str" to "date"
names(oilvn)[1] <- "date"
oilvn$date <- as.Date(oilvn$date, format = "%d/%m/%Y")

# Added a column containing the year of the date
oilvn$year <- as.numeric(strftime(oilvn$date, format = "%Y"))
head(oilvn, n = 2)
```

```
##         date Ron.95.II.V1 Ron.92.II.V1 Diezen.0.05s.V1 Dau.hoa.V1
## 1 2016-09-05        16770        16070           12380      10980
## 2 2016-08-19        16070        15370           11910      10490
##   Mazut.No2B..3.0S..V1 Mazut.No2b.3.5S..V1 Mazut.No3..380..V1
## 1                 9730                9330               9230
## 2                 9230                8830               8730
##   E5.RON.92.II.V1 Ron.95.II.V2 Ron.92.II.V2 Diezen.0.05s.V2 Dau.hoa.V2
## 1           15830        17100        16390           12620      11190
## 2           15220        16390        15670           12140      10690
##   Mazut.No2B..3.0S..V2 Mazut.No2b.3.5S..V2 Mazut.No3..380..V2
## 1                 9920                9510               9410
## 2                 9410                9000               8900
##   E5.RON.92.II.V2 year
## 1           16140 2016
## 2           15520 2016
```

The US petrol data contain data from 20 locations in the US, here I'm only interested in knowing the date and the price, not the location, so I will just discount all of those unnecessary detail.

```
# US's Conventional midgrade petrol (Ron91 - Ron92)
rawus <- "https://raw.githubusercontent.com/Linhnguyen14/stick/master/oilprice/gasolineus.csv"
oilus <- read.csv(text = getURL(rawus))

# Arrange column names running from 1 to 21, since I'm not intereted in the local location.
colnames(oilus) <- as.factor(c(1:21))
head(oilus, n = 2)
```

```
##            1     2     3     4     5     6     7     8     9 10 11 12 13
## 1 11/28/1994 1.174 1.198 1.236 1.259 1.182 1.121 1.221 1.231 NA NA NA NA
## 2  12/5/1994 1.167 1.199 1.235 1.268 1.178 1.109 1.212 1.240 NA NA NA NA
##   14 15 16 17 18 19 20 21
## 1 NA NA NA NA NA NA NA NA
## 2 NA NA NA NA NA NA NA NA
```

This dataset from oanda is the best I could get my hands on, a bit tricky to use, might created some bias because of the way I handled it, but in the long run, those bias should be nullified. For now it is good as it is, I will wrangle it later.

```
# History of currency rate (USD - VND)
curhistory <- "https://raw.githubusercontent.com/Linhnguyen14/stick/master/oilprice/curhistory.csv"
curhistory <- read.csv(text = getURL(curhistory))
```

THE DATA WRANGLING

The US petrol data are in wide format.

First, I reformat them to the more favourable long format, and while I'm at it, I also reformat the date variable to type "date"

```
# Wrangling with the US's petrol data
# Turn the data from wide format to long format
oilus <- gather(oilus,"location","price", 2:21)
oilus <- subset(oilus, !is.na(oilus$price))
names(oilus)[1] <- "date"

# Reformat the date variable (us)
date <- as.character(oilus$date)
oilus$date <- as.Date(date, format = "%m/%d/%Y")
oilus$year <- as.numeric(strftime(oilus$date, format = "%Y"))
oilus$month <- as.numeric(strftime(oilus$date, format = "%m"))

head(oilus)
```

```
##         date location price year month
## 1 1994-11-28        2 1.174 1994    11
## 2 1994-12-05        2 1.167 1994    12
## 3 1994-12-12        2 1.156 1994    12
## 4 1994-12-19        2 1.147 1994    12
## 5 1994-12-26        2 1.136 1994    12
## 6 1995-01-02        2 1.159 1995     1
```

Group the oilus data by date, and take the mean of the oil price, since price fluctuate from loc to loc, then assign them to a new dataframe "oilusmean"

```
oilusmean <- oilus %>%
  group_by(date) %>%
  summarise(mean = mean(price), sum = sum(price), n = n()) %>%
  arrange(date)
oilusmean$year <- as.numeric(strftime(oilusmean$date, format = "%Y"))
oilusmean$month <- as.numeric(strftime(oilusmean$date, format = "%m"))
head(oilus)
```

```
##         date location price year month
## 1 1994-11-28        2 1.174 1994    11
## 2 1994-12-05        2 1.167 1994    12
```

```
## 3 1994-12-12        2 1.156 1994    12
## 4 1994-12-19        2 1.147 1994    12
## 5 1994-12-26        2 1.136 1994    12
## 6 1995-01-02        2 1.159 1995     1
```

The dataset doesn't have closing price, only "bid" and "ask", so I took the mean of it. It might be a bit away from the actual data, but in the long run, the differences is too small to make it a problem.

I created 2 additional columns to store the months and years data, I will be using them as matching point to join the 2 datasets oilus and curhistory together.

```
# Wrangling the currency rate history:
# Reformat the date:
curhistory$date <- as.Date(as.character(curhistory$date), format = "%d%b%Y")
curhistory$price <- (curhistory$bid +curhistory$ask)/2
curhistory <- subset(curhistory, select = c(date, price))
curhistory$year <- as.numeric(strftime(curhistory$date, format = "%Y"))
curhistory$month <- as.numeric(strftime(curhistory$date, format = "%m"))

# Add merge point
curhistory <- curhistory %>%
  group_by(year, month) %>%
  summarise(price = mean(price), n = n())

# Join the petrol price with the currency rate
oilusmean <- left_join(oilusmean, curhistory, by = c("year", "month"))
oilus <- left_join(oilus, curhistory, by = c("year", "month"))

# Translate the US's price to VNS's price
litregallon <- 0.264172 #The rate of litre - gallon
#usdvnd <- 20301.50      #The exchange rate of usd - vnd
oilusmean$priceinvnd <- oilusmean$price * oilusmean$mean * litregallon
oilus$priceinvnd <- oilus$price.x * oilus$price.y * litregallon
```

THE PLOT

We kind of see it in here, the Vietnam's price are deliberately kept high, much much higher than the price.

Let us keep in mind, that gas companies in the US has to pay up to 20% in term of marketing. Doubt if petrolimex has to do that at all, they are, after all, the sole provider of petrol in Vietnam.

If you look further back, before 2012, you will see something interesting. Turns out, petrolimex could do well, even when they set the price as high as others (remember the marketing things?) It seems like, it turned out this way, mostly because the loss they took in investments and the extra tax from government.

Looking good, right? Too good, I'm afraid. While this plot might successfully represent the change in price of the US's petrol, it didn't do a good job, when it came to Vietnam.

```
ggplot(aes(x = date, y = priceinvnd), data = subset(oilus, year >2011)) +
  geom_point(colour = "cadetblue1") +
  geom_line(aes(x = date, y = Ron.92.II.V1),
            data = subset(oilvn, !is.na(oilvn$Ron.92.II.V1) & year > 2011),
            size = 1.5, colour = "firebrick") +
  geom_line(aes(x = date, y = priceinvnd),
```

```
            data = subset(oilusmean, !is.na(oilusmean$priceinvnd) & year > 2011),
            size = 1, colour = "forestgreen") +
  theme(panel.background = element_rect(fill = 'white'),
        plot.background = element_rect(fill = 'white'))
```

Plot-1.pdf

THE REVISED STRATEGY

The idea is that, unlike the US's price, where we could either use a line graph or a scatter plot to represent the data, the Vietnam's price are set by only one party, once setted, it will stay fixed, tilled setted again.

Which means, neither a line-graph nor a scatter plot could accuraely represent the movement of price in vietnam. However a modified version of the line-graph will make it, where all the line are either parallel to the horizontal or the vertical axis.

```
# Wrangle the oilvn price
# Too much information, let's only take the data that we use, which is the Ron.92.II.V1, because we cou
oilvn <- subset(oilvn, !is.na(oilvn$Ron.92.II.V1), select = c(date,Ron.92.II.V1))

# The simplest solution I could think of to solve this, is to create a line-graph that goes throw some
# Original data:
# date         value
# a,            e
# b,            f
# c,            g
```

```r
# Modified data:
# date     value
# a - 1      f
# b - 1      g
# c - 1      NA

# Then merge the original with the modified:
# date   value
# a        e
# a - 1    f
# b        f
# b - 1    g
# c        g
# c - 1    NA

#
# That way, we could create line-graph that best visualize the reality.

# Add the modified date collumn
oilvn$date2 <- as.Date(oilvn$date) - 1

# To keep track of them, we introduced a new id collumn, which run from 1 to 107, I did originally noti
oilvn$id <- c(1:107)

# We create 2 different data.frame to manipulate the data, and will merge them together later.
datedata <- subset(oilvn, select = c(id, date2))
pricedata <- subset(oilvn, select = c(id, Ron.92.II.V1))

# We remove the first row of datedata and the last row of pricedata, then merge them together.
# The end result:
# c - 1   c   b - 1   b   a - 1   a   (timeline going forward)
# NA      g   g       f   f       e
# Remove the first row of pricedata
dim(pricedata)
```

```
## [1] 107   2
```

```r
pricedata <- pricedata[-1,]
dim(pricedata)
```

```
## [1] 106   2
```

```r
# Additional manipulation to put the price dataframe in shape
p <- subset(pricedata, id ==107)
p[1,2] <- NA
pricedata$id <- pricedata$id - 1
pricedata <- rbind(pricedata, p)

# Merge the date and the price data together
petrovn <- merge(datedata,pricedata, by = "id")
names(petrovn)[2] <- "date"
```

```r
# Merge it with the original oil data set, then holy Marie, think its pretty much the end.
oilvn <- subset(oilvn, select = c(date, Ron.92.II.V1))
petrovn <- bind_rows(petrovn, oilvn)
petrovn$year <- as.numeric(strftime(petrovn$date, format = "%Y"))
```

THE FINAL PLOT

This should be the most accurately plot that I could draw. Although this plot could show just how overprice the Vietnam's petrol are compared to the US's, I believe that it is not all there is to it.

There should be additional insights, if we could break through the operation and taxation barriers. There should be a much wider gap, I believe.

```r
ggplot(aes(x = date, y = priceinvnd), data = subset(oilus, year > 2011)) +
  geom_point(colour = "palegreen") +
  geom_line(aes(x = date, y = priceinvnd),
            data = subset(oilusmean, !is.na(oilusmean$priceinvnd) & year > 2011),
            size = 1, colour = "forestgreen") +
  geom_line(aes(x = date, y = Ron.92.II.V1),
            data = subset(petrovn, year > 2011),
            size = 1.5, colour = "firebrick")+
  theme(panel.background = element_rect(fill = 'white'),
        plot.background = element_rect(fill = 'white')) +
  ylab("Petrol price in VND") +
  xlab("Year")
```