Motivating Examples

STAT3022 Applied Linear Models Lecture 21

2020/02/18

THE UNIVERSITY OF SYDNEY

Today

1. Exploratory Data Analysis
2. Data cleaning

The Avocado Lover's job problem

Dear statistician,

I work in an Avocado company that has many offices in US. My boss asked me to analyse the US market for the price and sales trends and factors that drive these trends. I am able to obtain a data set to analyse but I need your help to advise me how to conduct this research.

Personally, I am also an avocado lover and eat avocado toast nearly every day. My boss told me that he will send me to an US office in July 2019 to get some trainings and let me choose which office to work in. My salary is not too high so I need to make sure that avocado is least expensive there. I like organic avocado but if it's too costly, I'm willing to settle for non-organic avocado. Please help me decide where I should go!

Kind regards,

Avocado Lover

What is the aim?

The Avocado Lover's job problem

Dear statistician,

I work in an Avocado company that has many offices in US. My boss asked me to analyse the US market for the **price and sales trends and factors that drive these trends**. I am able to obtain a data set to analyse but I need your help to advise me how to conduct this research.

Personally, I am also an avocado lover and eat avocado toast nearly every day. My boss told me that he will send me to an US office in **July 2019** to get some trainings and let me choose which office to work in. My salary is not too high so I need to make sure that **avocado is least expensive** there. I **like organic avocado** but if it's too costly, I'm willing to **settle for non-organic avocado**. Please help me decide **which state** I should go!

Kind regards,

Avocado Lover

First identify what the client wants.

- Price and sales trends and their driving factors.
- Personally, the price in **July 2019**.
- Recommendation of which **state** to go. More beneficial is to identify city in which her workplace has an office.
- The client cares moderately for **organic** avocados. How much is she willing to pay extra for it?
- The client may also care about cost of living other than avocados (rent, electricity, water bills, etc), lifestyle, etc.

Do you have the data to answer the question?

- In real world, the questions are often not clearly defined.
- Keep in mind that the client may find it hard to pinpoint exactly what they want.
- It is your job as a statistician to work together with the client to frame the problems numerically/statistically.
- Even after you form the questions analytically, you may have insufficient data to answer the questions.
- You may need to make the best out of data, acknowledging the potential flaws in the analysis.
- In some cases, you just simply cannot gain any information out of data.
- "If you torture the data long enough, it will confess." -Ronald H. Coase

Understand avocado prices and sales volume in US

```
skimr::skim(dat)
```

```
  Skim summary statistics
   n obs: 18249
   n variables: 13

  ── Variable type:character ─────────────────────────────
   variable missing complete     n min max empty n_unique
       Date       0    18249 18249   6   8     0      169
     region       0    18249 18249   4  19     0       54
       type       0    18249 18249   7  12     0        2

  ── Variable type:numeric ──────────────────────────────
       variable missing       mean          sd      p0       p50          p100     hist
           4046       0 293008.42 1264989.08    0      8645.3        2.3e+07 ▇▁▁▁▁▁▁▁
           4225       0  3e+05    1204120.4     0     29061.02   2e+07        ▇▁▁▁▁▁▁▁
           4770       0  22839.74  107464.07    0       184.99 2546439.11     ▇▁▁▁▁▁▁▁
   AveragePrice       0      1.41       0.4     0.44      1.37       3.25     ▁▆▇▃▁▁▁▁
     Large Bags       0  54338.09  243965.96    0      2647.71 5719096.61     ▇▁▁▁▁▁▁▁
     Small Bags       0 182194.69  746178.51    0     26362.82        1.3e+07 ▇▁▁▁▁▁▁▁
     Total Bags       0 239639.2   986242.4     0     39743.83        1.9e+07 ▇▁▁▁▁▁▁▁
   Total Volume       0 850644.01 3453545.36   84.56 107376.76        6.3e+07 ▇▁▁▁▁▁▁▁
    XLarge Bags       0   3106.43   17692.89    0         0      551693.65     ▇▁▁▁▁▁▁▁
           year       0   2016.15       0.94 2015      2016        2018        ▇▁▇▁▁▇▁▂
```


Get data summary

What does each variable mean?

```
dat <- dat %>%
  mutate(Date=as.Date(Date, format="%d/%m/%y")) %>%
  janitor::clean_names() # better names for analysis
str(dat)
```

```
 tibble [18,249 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
  $ date         : Date[1:18249], format: "2015-12-27" "2015-12-2
  $ average_price: num [1:18249] 1.33 1.35 0.93 1.08 1.28 1.26 0.
  $ total_volume : num [1:18249] 64237 54877 118220 78992 51040 .
  $ x4046        : num [1:18249] 1037 674 795 1132 941 ...
  $ x4225        : num [1:18249] 54455 44639 109150 71976 43838 .
  $ x4770        : num [1:18249] 48.2 58.3 130.5 72.6 75.8 ...
  $ total_bags   : num [1:18249] 8697 9506 8145 5811 6184 ...
  $ small_bags   : num [1:18249] 8604 9408 8042 5677 5986 ...
  $ large_bags   : num [1:18249] 93.2 97.5 103.1 133.8 197.7 ...
  $ x_large_bags : num [1:18249] 0 0 0 0 0 0 0 0 0 0 ...
  $ type         : chr [1:18249] "conventional" "conventional" "c
  $ year         : num [1:18249] 2015 2015 2015 2015 2015 ...
  $ region       : chr [1:18249] "Albany" "Albany" "Albany" "Alba
```

Given description

- `date` - the date of the observation
- `average_price` - the average price of a single avocado
- `type` - conventional or organic
- `year` - the year
- `region` - the city or region of the observation
- `total_volume` - Total number of avocados sold
- `x4046` - total number of avocados with PLU 4046 - small/medium Hass avocado (about 3-5oz avocado) - sold
- `x4225` - total number of avocados with PLU 4225 - large Hass avocado (about 8-10oz avocado) - sold
- `x4770` - total number of avocados with PLU 4770 - extra large Hass avocado (about 10-15oz avocado) - sold

The PLU or Price Look-Up code is a 4- or 5-digit number that is primarily used on fresh produce items.

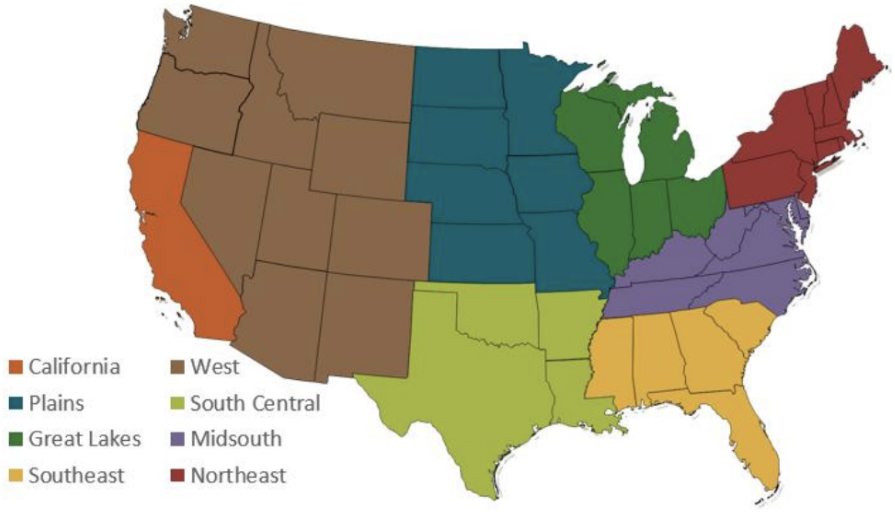
Do your research of avocado market!

Original data source:

http://www.hassavocadoboard.com/retail/volume-and-price-data

Regions:

IRI Retail Sales Regions



Markets

CALIFORNIA Los Angeles Sacramento San Diego San Francisco	WEST Denver Phoenix Portland Salt Lake City Seattle West Texas/New Mexico	PLAINS St. Louis	SOUTH CENTRAL Dallas Houston New Orleans
GREAT LAKES Chicago Cincinnati Columbus Detroit Grand Rapids Indianapolis	MIDSOUTH Baltimore Charlotte Louisville Nashville Raleigh Richmond Roanoke	SOUTHEAST Atlanta Jacksonville Miami Orlando South Carolina Tampa	NORTHEAST Albany Boston Buffalo Harrisburg/Scranton Hartford/Springfield New England New York Philadelphia Pittsburgh Syracuse

Understand regions

```
unique(dat$region)
```

```
 [1] "Albany"              "Atlanta"             "BaltimoreWashington" "Boise"
 [5] "Boston"              "BuffaloRochester"    "California"          "Charlotte"
 [9] "Chicago"             "CincinnatiDayton"    "Columbus"            "DallasFtWorth"
[13] "Denver"              "Detroit"             "GrandRapids"         "GreatLakes"
[17] "HarrisburgScranton"  "HartfordSpringfield" "Houston"             "Indianapolis"
[21] "Jacksonville"        "LasVegas"            "LosAngeles"          "Louisville"
[25] "MiamiFtLauderdale"   "Midsouth"            "Nashville"           "NewOrleansMobile"
[29] "NewYork"             "Northeast"           "NorthernNewEngland"  "Orlando"
[33] "Philadelphia"        "PhoenixTucson"       "Pittsburgh"          "Plains"
[37] "Portland"            "RaleighGreensboro"   "RichmondNorfolk"     "Roanoke"
[41] "Sacramento"          "SanDiego"            "SanFrancisco"        "Seattle"
[45] "SouthCarolina"       "SouthCentral"        "Southeast"           "Spokane"
[49] "StLouis"             "Syracuse"            "Tampa"               "TotalUS"
[53] "West"                "WestTexNewMexico"
```

- Some regions are unclear:
  - Albany in OR, GA and NY
  - Columbus in IN and OH
  - Jacksonville in FL
  - Portland in ME and OR
- Some regions seem to be a combination, e.g. `BuffaloRochester`, `CincinnatiDayton`, etc.
- How do we verify where it is?
- Is it possible?

Information on US cities

Below omits the `ggrepel::geom_label_repel`.

```
library(maps)
states <- map_data("state")
ggplot(states) +
  geom_polygon(aes(long, lat,
  group = group), color = "white",
  fill="black") + coord_fixed(1.3) +
  theme_void() + guides(fill=FALSE)
```

```
us.cities %>%
  select(name, pop, lat, long) %>%
  head()
```

```
        name    pop   lat    long
1 Abilene TX 113888 32.45  -99.74
2   Akron OH 206634 41.08  -81.52
3 Alameda CA  70069 37.77 -122.26
4  Albany GA  75510 31.58  -84.18
5  Albany NY  93576 42.67  -73.80
6  Albany OR  45535 44.62 -123.09
```


Duplicate city names

```
us.cities %>%
  filter(grepl("Albany", name) |
         grepl("Columbus", name) |
         grepl("Jacksonville", name) |
         grepl("Portland", name)) %>%
  select(name, pop, lat, long)
```

```
             name    pop   lat    long
1        Albany GA  75510 31.58  -84.18
2        Albany NY  93576 42.67  -73.80
3        Albany OR  45535 44.62 -123.09
4      Columbus GA 184900 32.51  -84.87
5      Columbus IN  39453 39.21  -85.91
6      Columbus OH 741677 39.99  -82.99
7  Jacksonville FL 809874 30.33  -81.66
8  Jacksonville NC  68201 34.76  -77.40
9      Portland ME  62882 43.66  -70.28
10     Portland OR 542751 45.54 -122.66
```

- Albany should be in NY
- Portland should be in OR
- Columbus should be in IN or OH (GA far away)
- Jacksonville should be in FL or NC

How to choose Columbus? Use US City Information

Getting key variables to merge with original data.

```
USkey <- us.cities %>% mutate(region=substr(name, 1, nchar(name) - 3)) %>%
  rename(state=country.etc) %>%
  mutate(region=gsub(" ", "", region)) %>%
  select(region, pop, lat, long, state) %>%
  rbind(tribble(~region, ~pop, ~lat, ~long, ~state,
    "BaltimoreWashington", 602658 + 548359, (39.30 + 38.91)/2, (-76.61 -77.02)/2, "MD+DC",
    "BuffaloRochester", 276762 + 209587, (42.89 + 43.17)/2, (-78.86-77.62)/2, "NY+NY",
    "California", 26729306, 35.42923, -119.3301, "CA",
    "CincinnatiDayton", 301561+157607, (39.14 + 39.78)/2, (-84.51-84.20)/2, "OH+OH",
    "DallasFtWorth", 1216543+633849, (32.79+32.75)/2, (-96.77-97.34)/2, "TX+TX",
    "district of columbia", 548359, 38.91, -77.02, "DC",
    "GreatLakes", 15182052, 41.8975, -86.31265, "WI+IA+IL+MI+OH",
    "HarrisburgScranton", 47576 + 72516, (40.28+41.40)/2, (-76.88-75.67)/2, "PA+PA",
    "HartfordSpringfield",123836 + 152095, (41.77 + 42.12)/2, (-72.68-72.54)/2, "CT+MA",
    "MiamiFtLauderdale", 173597 + 386740, (26.14+25.78)/2, (-80.14-80.21)/2, "FL+FL",
    "Midsouth", 10669181, 37.51253, -79.53505, "KY+WV+TN+VA+NC+WV+MD+DE+DC",
    "NewOrleansMobile", 188467+454207, (30.68+30.07)/2, (-88.09-89.93)/2, "AL+LA",
    "Northeast", 21530612, 41.54582, -73.37044, "NJ+PA+NY+CT+RI+MA+VT+NH+ME",
    "NorthernNewEngland", 372834, 43.39, -71.14, "ME+NH+VT", # guess
    "PhoenixTucson", 1450884+525268, (33.54+32.20)/2, (-112.07-110.89)/2, "AZ+AZ",
    "Plains", 6719966, 41.87682, -94.22152, "ND+SD+NE+KS+MN+IA+MO",
    "RaleighGreensboro", 233342+350822, (36.08+35.82)/2, (-79.83-78.66)/2, "NC+NC",
    "RichmondNorfolk", 248182+189498, (36.92+37.53)/2, (-76.24-77.47)/2, "VA+VA",
    "SouthCarolina", 526667, 33.75429, -80.65286, "SC",
    "SouthCentral", 16433060, 31.9431, -96.3638, "TX+OK+LA+AR",
    "Southeast", 11204354, 29.29783, -82.32278, "MS+AL+GA+SC+FL",
    "StLouis", 315546, 38.64, -90.24, "MO",
    "TotalUS", 126175816, 37.49531, -95.10346, "AL+AK+AZ+AR+CA+CO+CT+DE+FL+GA+HI+ID+IL+IN+IA+KS+
    "West", 15735863, 40.74418, -113.9845, "WA+OR+NV+ID+MT+WY+UT+CO+NM+AZ",
    "WestTexNewMexico",13721930, 31.44962, -98.13192, "TX+NM" # includes east Texas too
```

How to choose?

IN or OH differ in pop size!

- Use total volume to predict pop size.
- Add pop size information.
- Some regions encompasses a large region.
- Some regions are subset of other region.
- Some coded here involve guess work and may have room for error.

Is there a relationship with volume and population?

```
us.cities %>%
  filter(grepl("Columbus", name)) %>% # IN or OH
  select(name, pop, lat, long)
```

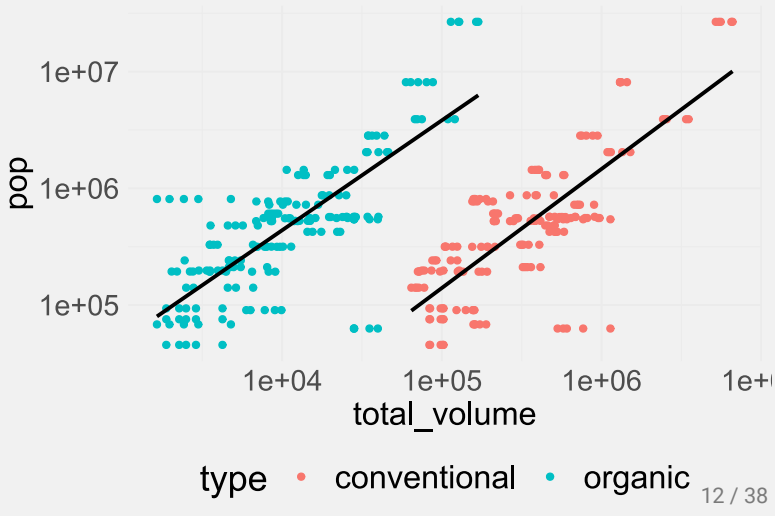
```
          name    pop   lat   long
1 Columbus GA 184900 32.51 -84.87
2 Columbus IN  39453 39.21 -85.91
3 Columbus OH 741677 39.99 -82.99
```

```
dat2 <- dat %>% left_join(USkey, by="region") %>%
  # leave these out for now
  filter(region!="Columbus") %>%
  filter(!grepl("+", state, fixed=T)) #drop comb
```

```
dim(dat2) #dat 18429,13 drop columbus & comb
```

```
 [1] 12168    17
```

```
library(lubridate)
dat3 <- dat2 %>% filter(year==2017 & month(date)==7)
ggplot(dat3, aes(total_volume, pop,
                 color=type, group=type)) +
  geom_point() + scale_x_log10()  + scale_y_log10() +
  geom_smooth(method="lm", se=FALSE, color="black") +
  theme(legend.position = "bottom")
```



Predict the population of Columbus

```
predat <- dat %>%
  filter(grepl("Columbus", region)) %>%
  filter(year==2017 & month(date)==7) %>%
  group_by(region, type) %>%
  summarise(total_volume=mean(total_volume)) %>%
  ungroup()
predat #use tot vol at 7-2017 to pred pop size
```

```
 # A tibble: 2 x 3
   region   type         total_volume
   <chr>    <chr>               <dbl>
 1 Columbus conventional      195168.
 2 Columbus organic             9886.
```

```
M0 <- lm(log10(pop)~type+log10(total_volume),dat3)
10^predict(M0,newdata=predat,
           interval="prediction") %>%
cbind(predat)
```

```
        fit      lwr      upr   region          type tota
 1 282664.3 51813.66 1542047 Columbus conventional    19
 2 430889.5 79026.84 2349401 Columbus      organic
```

- Since the predicted population size is large, Columbus appears to be in OH.
- Let's modify our key:

```
USkey_ncr <- USkey %>%
filter(!(region=="Columbus" & state!="OH") &
         region %in% dat$region)
dat_ncr<-dat %>% left_join(USkey_ncr,by="region") %>%
 filter(!grepl("+", state, fixed=T)) #no comb region
dat_cr <- dat %>% left_join(USkey_ncr, by="region")
```

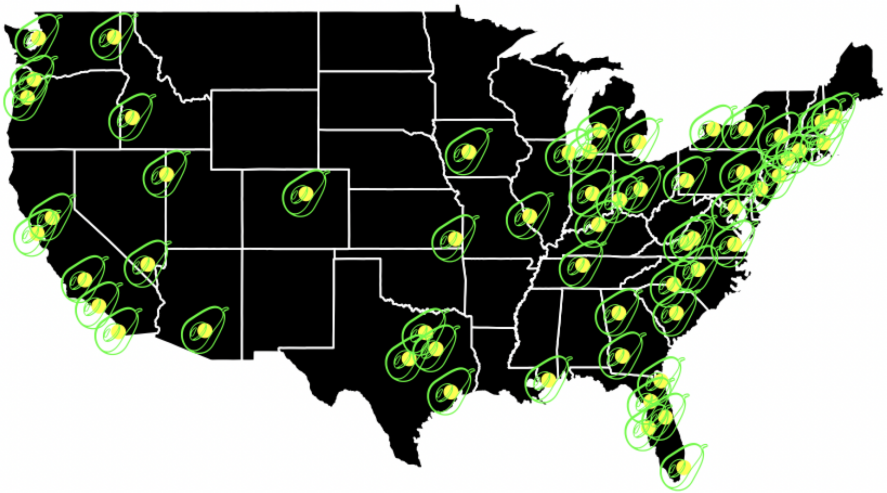
```
dim(dat_ncr)
```

```
 [1] 12506    17
```

```
dim(dat_cr)
```

```
 [1] 19601    17
```


Data coverage for her company offices?



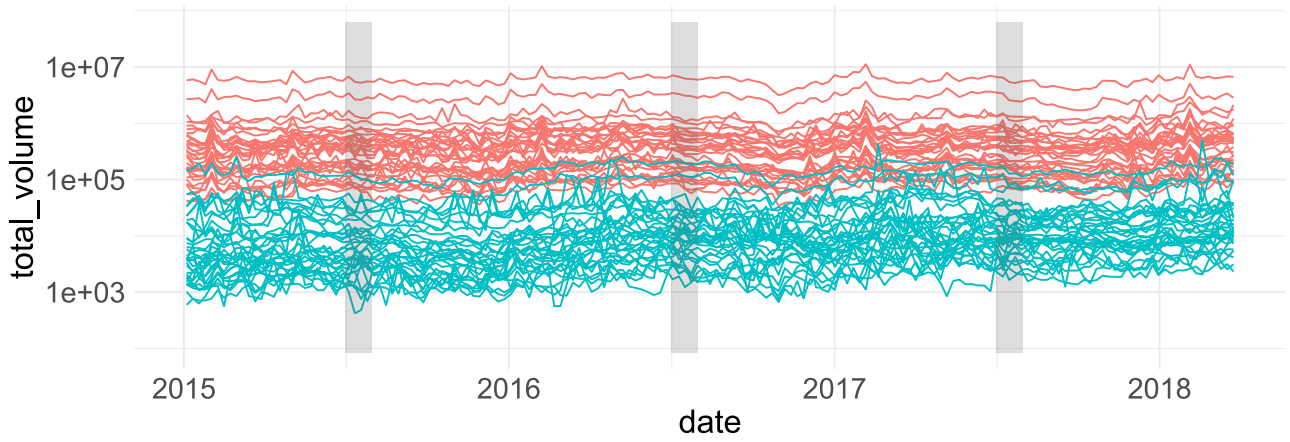
- Some states are not well covered.

```
strsplit(dat_cr$state, "+", fixed=T) %>%
  unlist() %>%
  unique()
```

```
  [1] "GA" "NY" "OR" "MD" "DC" "ID" "MA"
  [8] "CA" "NC" "IL" "OH" "TX" "CO" "MI"
 [15] "WI" "IA" "PA" "CT" "IN" "FL" "NV"
 [22] "KY" "WV" "TN" "VA" "DE" "AL" "LA"
 [29] "NJ" "RI" "VT" "NH" "ME" "AZ" "ND"
 [36] "SD" "NE" "KS" "MN" "MO" "WA" "SC"
 [43] "OK" "AR" "MS" "AK" "HI" "MT" "NM"
 [50] "UT" "WY"
```

Boss's question: does sales (total) volume change?

```
julyperiod <- data.frame(xmin=as.Date(c("01/07/15", "01/07/16", "01/07/17"), format="%d/%m/%y"),
                         xmax=as.Date(c("31/07/15", "31/07/16", "31/07/17"), format="%d/%m/%y"),
                         ymin=min(dat$total_volume), ymax=max(dat$total_volume))
ggplot(dat_ncr) +
  geom_line(aes(date, total_volume, group=interaction(region,type), color=type)) + scale_y_log10() +
  geom_rect(data=julyperiod,
            mapping=aes(xmin=xmin, xmax=xmax, ymin=ymin, ymax=ymax), alpha=0.2) + guides(color=FALSE)
```



Is the change of total volume over time significant?

```
M1<-lm(total_volume~region:type+date,data=dat_ncr)
summary(M1) #date has associated numerical value
```

```
Call:
lm(formula = total_volume ~ region:type + date, data

Residuals:
     Min       1Q   Median       3Q
-2800515   -14931    -1542     9843
     Max
 5270823

Coefficients: (1 not defined because of singularitie
                                    Estimate
(Intercept)                       -4.171e+05
date                               2.475e+01
regionAlbany:typeconventional      8.866e+04
regionAtlanta:typeconventional     5.086e+05
```

```
[1] "2015-12-27" "2015-05-10"
```

```
[1] 16796 16565
```

```
M2 <- lm(total_volume ~ region:type, data=dat_ncr)
anova(M2, M1)
```

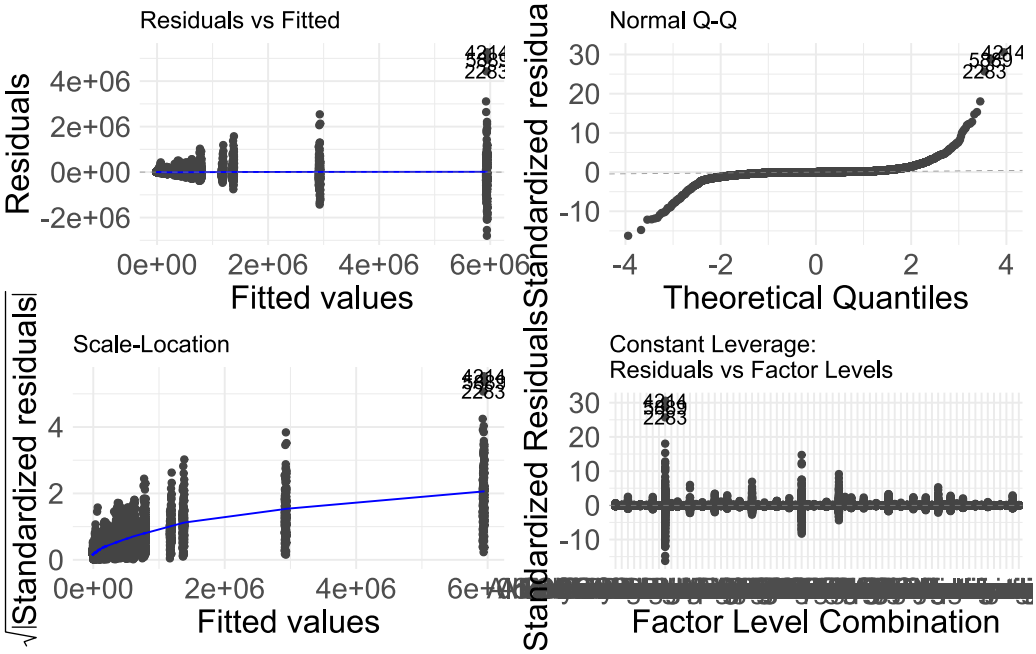
```
Analysis of Variance Table

Model 1: total_volume ~ region:type
Model 2: total_volume ~ region:type + date
  Res.Df        RSS Df  Sum of Sq
1  12440 3.7261e+14
2  12439 3.7172e+14  1 8.9302e+11
       F    Pr(>F)
1
2 29.884 4.676e-08 ***
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05
  '.' 0.1 ' ' 1
```

- The overall `date` effect is significant partly due to the large sample size.
- Since the `date` coefficient is +ve, `total_volume` increases over time in general.

Model diagnostics?

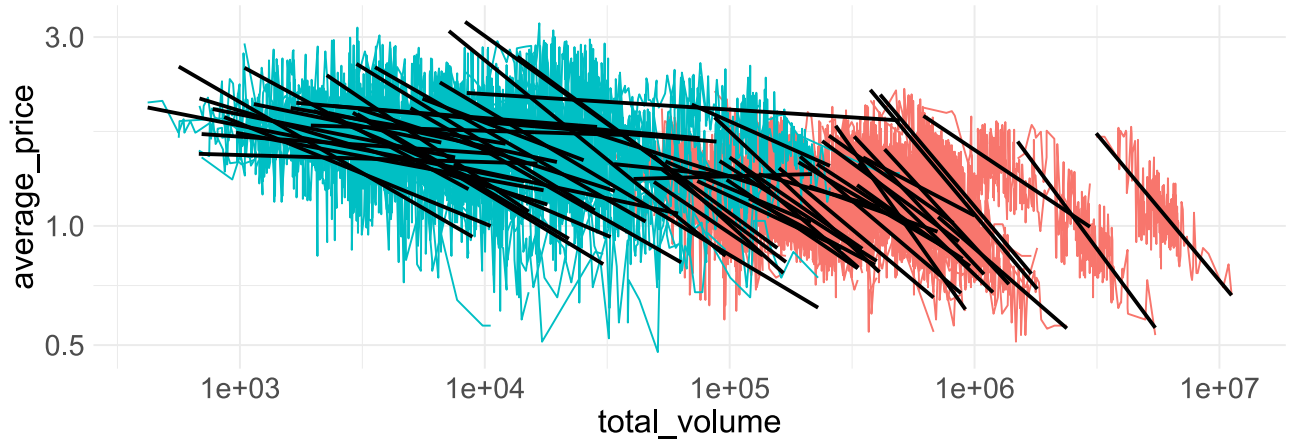
```
library(ggfortify)
autoplot(M1)
```



- The reality is that the model diagnostics are not satisfied and thus inferences may not be reliable.
- However outliers often present in large data set.
- It is common to see an increasing trend of variance for a time series.

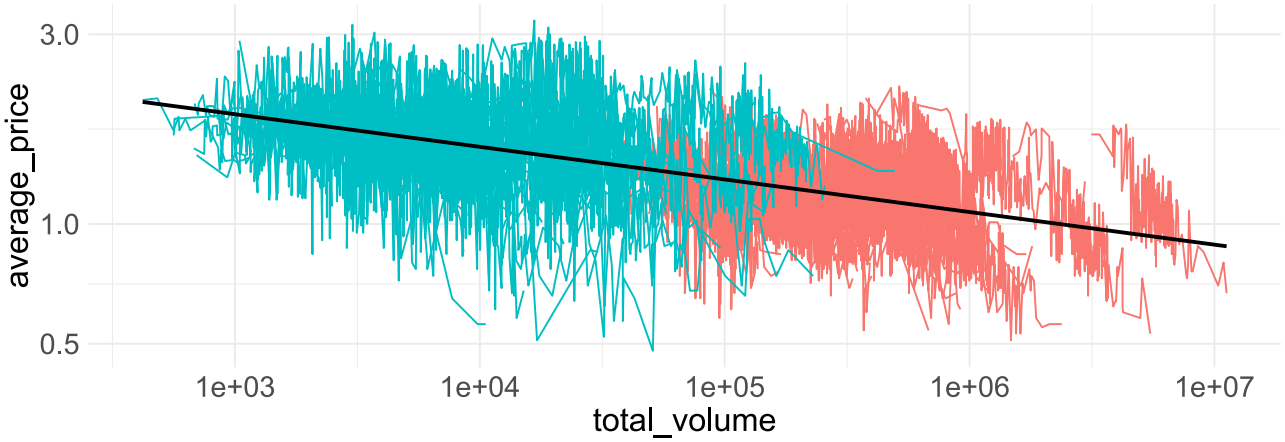
Boss's question: Does price change with supply?

```
ggplot(dat_ncr) +
  geom_line(aes(total_volume, average_price, group=interaction(region,type), color=type)) +
  scale_y_log10() + scale_x_log10() + guides(color=FALSE) +
  geom_smooth(method="lm", se=FALSE, aes(total_volume, average_price,
  group=interaction(region,type)), color="black")  #no overall line
```



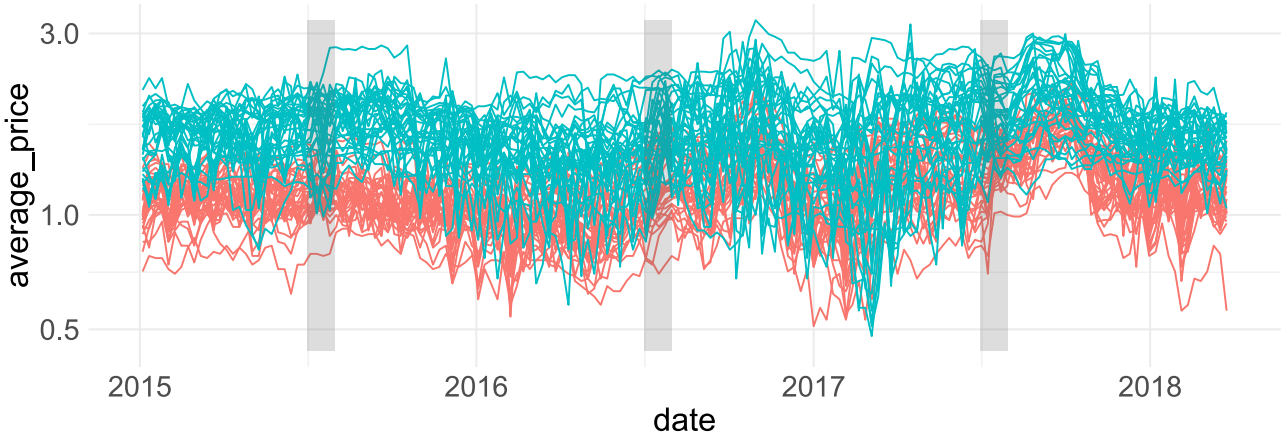
Does price change with supply? Overall trend?

```
ggplot(dat_ncr) +
  geom_line(aes(total_volume, average_price, group=interaction(region,type), color=type)) +
  scale_y_log10() + scale_x_log10() + guides(color=FALSE) +
  geom_smooth(method="lm", se=FALSE, aes(total_volume, average_price), color="black") #drop lines of each reg
```

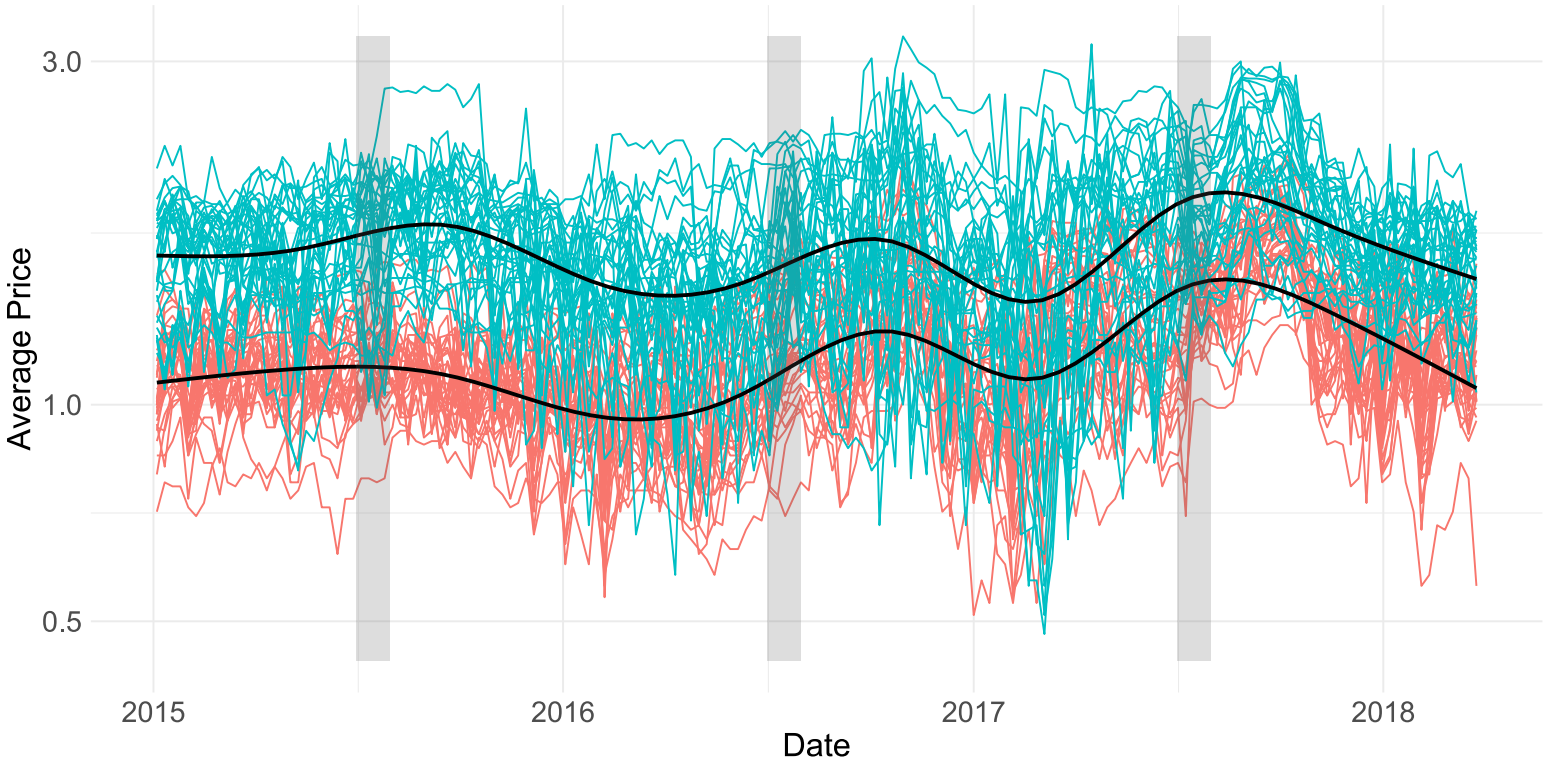


Boss's question: Does price change over time?

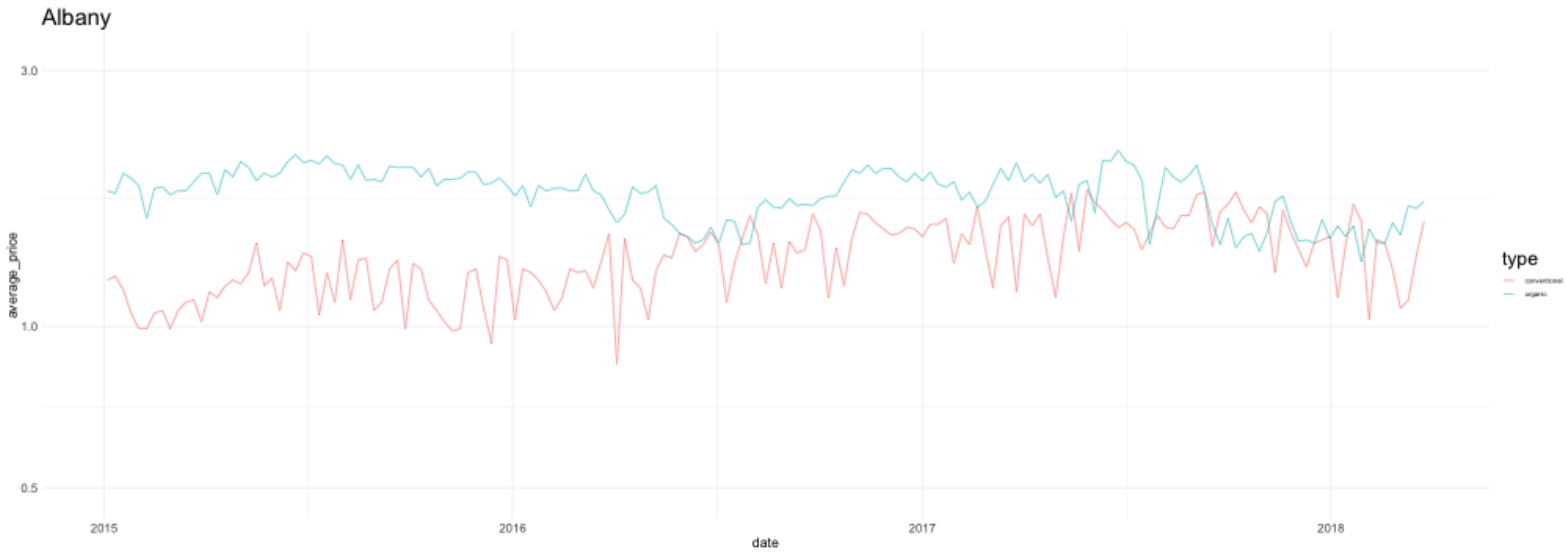
```
julyperiod <- data.frame(xmin=as.Date(c("01/07/15", "01/07/16", "01/07/17"), format="%d/%m/%y"),
                         xmax=as.Date(c("31/07/15", "31/07/16", "31/07/17"), format="%d/%m/%y"),
                         ymin=min(dat$average_price), ymax=max(dat$average_price)) #July highlight
ggplot(dat_ncr) +
  geom_line(aes(date, average_price, group=interaction(region,type), color=type)) + scale_y_log10() +
  geom_rect(data=julyperiod,
  mapping=aes(xmin=xmin,xmax=xmax,ymin=ymin,ymax=ymax), alpha=0.2) + guides(color=FALSE) #log scale
```



Does price over time by region and type?

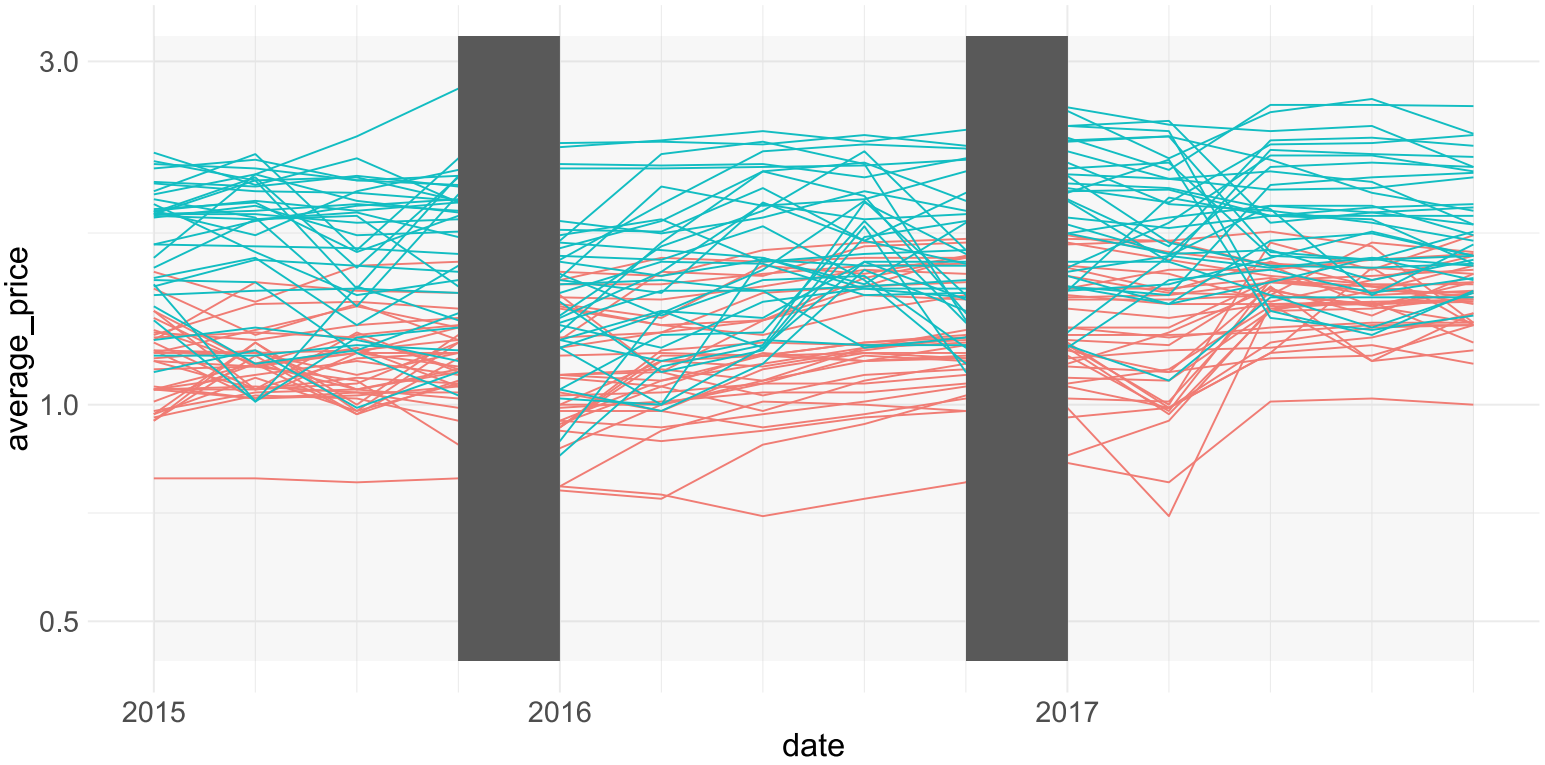


Does price change over time by region and type?

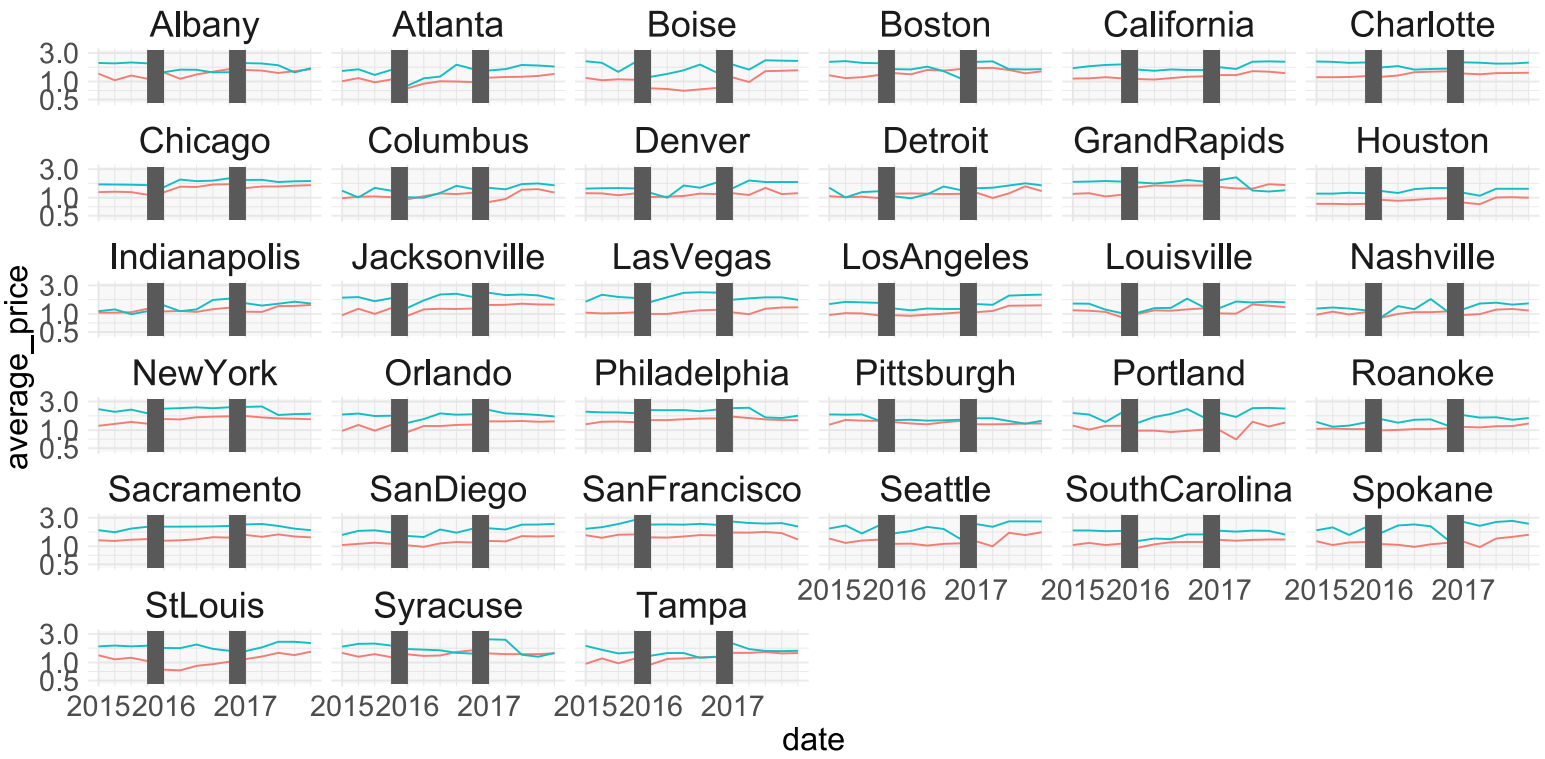


Do we need data for other days beside July?

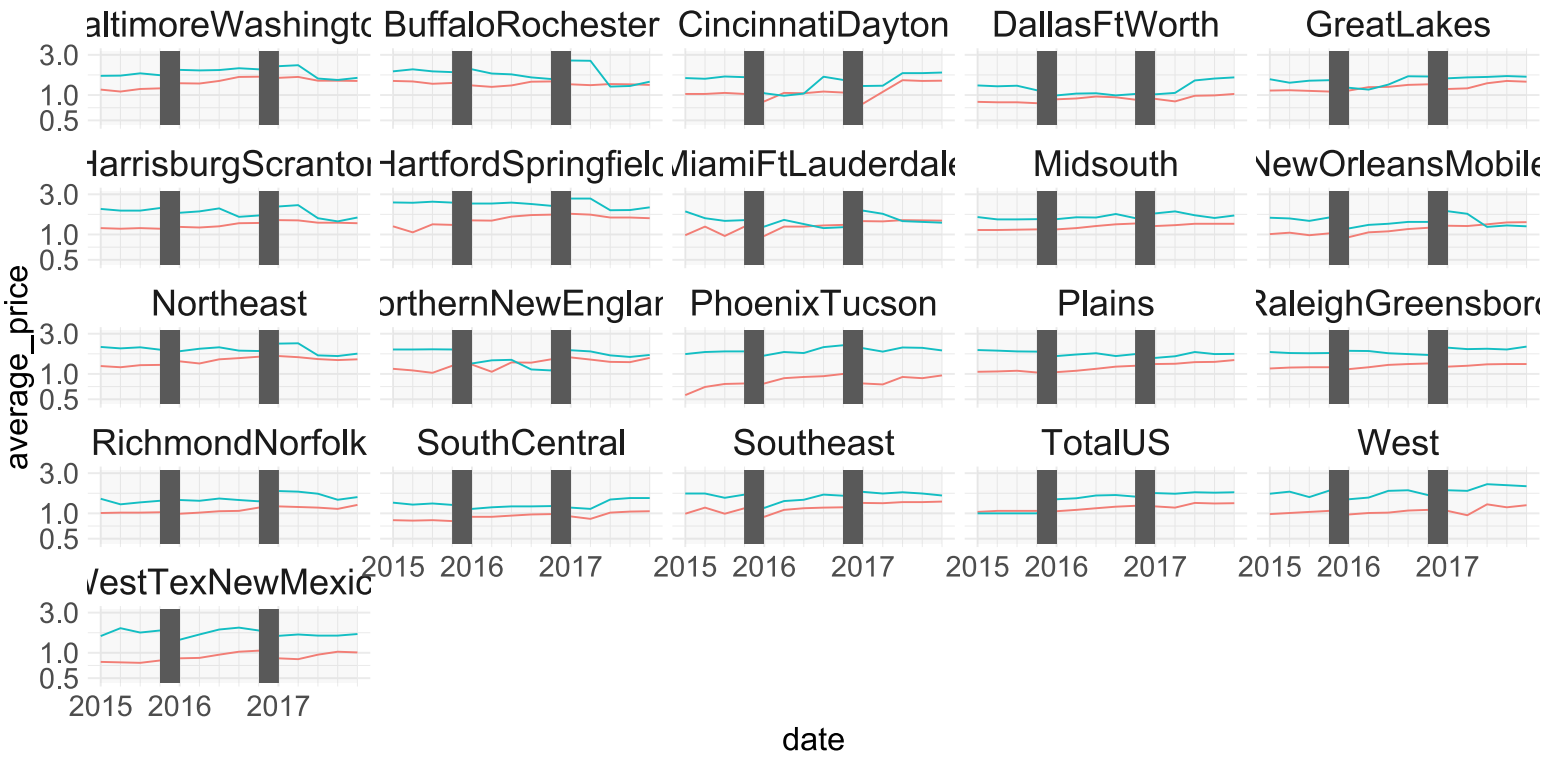
Price in cities across July weeks each year



Price for each city across July weeks each year



Price for each combined region across July weeks



Mean price for region by type by year via modelling

```
julydat_ncr <- julydat_ncr %>%
 mutate(year=factor(year))
M0 <- lm(average_price ~
region:type:year-1,data=julydat_ncr)
outM0 <- broom::tidy(M0) %>%
 arrange(estimate) %>%
 select(term, estimate) %>%
 separate(term,
       c("region", "type", "year")) %>%
 mutate(region=gsub("region", "", region),
        type=gsub("type", "", type),
        year=gsub("year", "", year)) %>%
 mutate(tyear=
  paste0(substring(type, 1, 1), year)) %>%
 select(region, tyear, estimate) %>%
 spread(tyear, estimate)
outM0
```

```
# A tibble: 33 x 7
   region  c2015 c2016 c2017 o2015 o2016
   <chr>   <dbl> <dbl> <dbl> <dbl> <dbl>
 1 Albany   1.19 1.38   1.51  2.03  1.49
 2 Atlanta  1.05 0.936  1.23  1.49  1.34
 3 Boise    1.09 0.748  1.34  1.94  1.48
 4 Boston   1.22 1.48   1.55  2.10  1.52
```

Show 5 entries

Search:

Region	Conventional			Organic		
	2015	2016	2017	2015	2016	2017
Albany	1.1925	1.382	1.514	2.035	1.486	1.798
Atlanta	1.05	0.936	1.226	1.4875	1.344	1.732
Boise	1.09	0.748	1.338	1.94	1.484	2.04
Boston	1.2175	1.478	1.554	2.0975	1.52	1.822
California	1.14	1.148	1.378	1.8225	1.56	1.96

Showing 1 to 5 of 33 entries

Previous 1 2 3 4 5 6 7

Mean price for region by type by year via means

```
outsum <- julydat_ncr %>%
 group_by(region, type, year) %>%
summarise(estimate=mean(average_price))%>%
 ungroup() %>%
 mutate(tyear=
  paste0(substring(type, 1, 1), year)) %>%
 select(region, tyear, estimate) %>%
 spread(tyear, estimate)
outsum
```

```
 # A tibble: 33 x 7
    region  c2015 c2016 c2017 o2015 o2016
    <chr>   <dbl> <dbl> <dbl> <dbl> <dbl>
  1 Albany   1.19 1.38   1.51  2.04  1.49
  2 Atlanta  1.05 0.936  1.23  1.49  1.34
  3 Boise    1.09 0.748  1.34  1.94  1.48
  4 Boston   1.22 1.48   1.55  2.10  1.52
  5 Califo~  1.14 1.15   1.38  1.82  1.56
  6 Charlo~  1.19 1.37   1.37  2.10  1.67
  7 Chicago  1.22 1.52   1.55  1.65  1.89
  8 Columb~  1.02 1.11   1.15  1.27  1.24
  9 Denver   1.16 1.09   1.22  1.44  1.41
 10 Detroit  1.03 1.17   1.25  1.25  1.21
 # ... with 23 more rows, and 1 more
 #   variable: o2017 <dbl>
```

Show 5 entries Search:

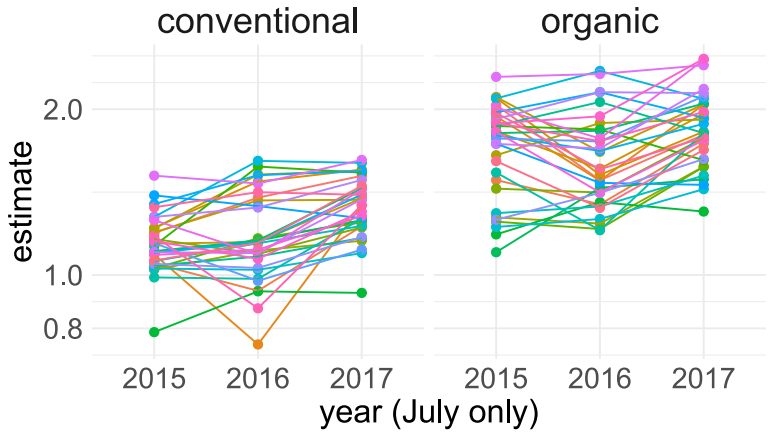
Region	Conventional			Organic		
	2015	2016	2017	2015	2016	2017
Albany	1.1925	1.382	1.514	2.035	1.486	1.798
Atlanta	1.05	0.936	1.226	1.4875	1.344	1.732
Boise	1.09	0.748	1.338	1.94	1.484	2.04
Boston	1.2175	1.478	1.554	2.0975	1.52	1.822
California	1.14	1.148	1.378	1.8225	1.56	1.96

Showing 1 to 5 of 33 entries

Previous 1 2 3 4 5 6 7 Next

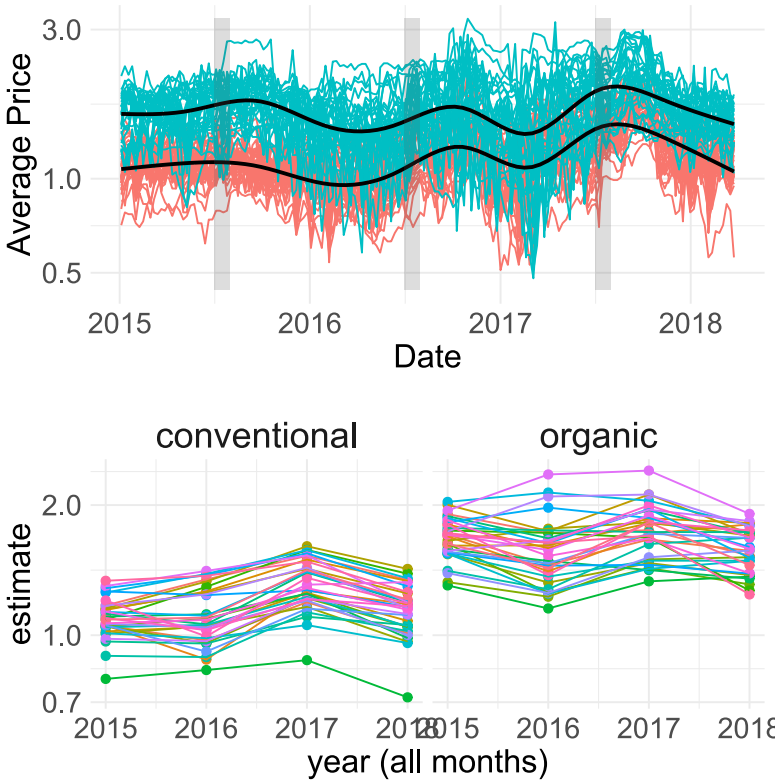
Factors for price trend: Interaction effects for July

```
gather(outsum, "typeyear", "estimate", -region) %>%
 mutate(type=case_when(
  substring(typeyear, 1, 1)=="c" ~ "conventional",
  substring(typeyear, 1, 1)=="o" ~ "organic"),
  year=substring(typeyear, 2, nchar(typeyear))) %>%
 ggplot(aes(year, estimate, group=region, color=region)) +
 facet_wrap(~type) + geom_point(size=2) + geom_line() +
 guides(color=FALSE) + scale_y_log10() + labs(x="year (July only)")
```



- We already know that `type` is an important factor for determining the `average_price`.
- Is `year` or `date` important?
- Is `year:region` important?
- Is `year:type` important?
- Is `year:type:region`?
- What about `date:region:type`?

Interaction effects



- Is `year` or `date` important?
  - Yes! Early-year price appears to dip across all `region` and `type`.
- Is `year:region` important?
  - This can be thought of as average of `year` by `region` combination. At specific dates, it seemed to behave differently but averaged over year may be not much.
- Is `year:type` important? Maybe.
- Is `year:type:region`? Maybe.
- What about `date:type:region`?

Testing for interaction effects

- Start with testing higher order interaction term `year:region:type`.
- If a higher order interaction term is included, you cannot drop main effects (`year`, `region`, `type`) or lower order interaction term (e.g. `year:region`, `year:type`, `region:type`).
- Because of the large data size, all terms including the higher order interaction are significant.

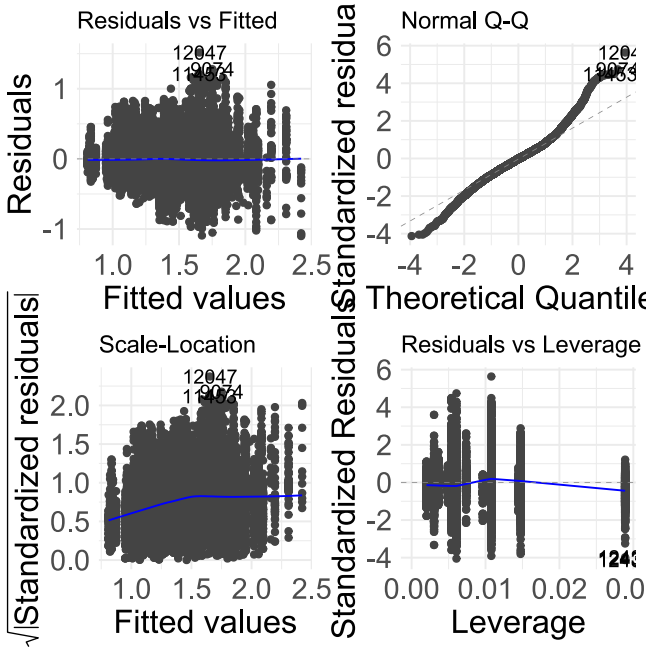
```
M1 <- lm(average_price~year*region*type,data=dat_ncr)
anova(M1)
```

```
Analysis of Variance Table

Response: average_price
                    Df Sum Sq Mean Sq
year                 1  18.33   18.33
region              32 305.62    9.55
type                 1 754.51  754.51
year:region         32  16.67    0.52
year:type            1  12.66   12.66
region:type         32  62.57    1.96
year:region:type    32  19.35    0.60
Residuals        12374 905.67    0.07
                    F value    Pr(>F)
year               250.4643 < 2.2e-16
region             130.4870 < 2.2e-16
type             10308.7792 < 2.2e-16
year:region          7.1172 < 2.2e-16
year:type          172.9561 < 2.2e-16
region:type         26.7147 < 2.2e-16
year:region:type     8.2626 < 2.2e-16
Residuals
```

Model diagnostic

```
autoplot(M1)
```



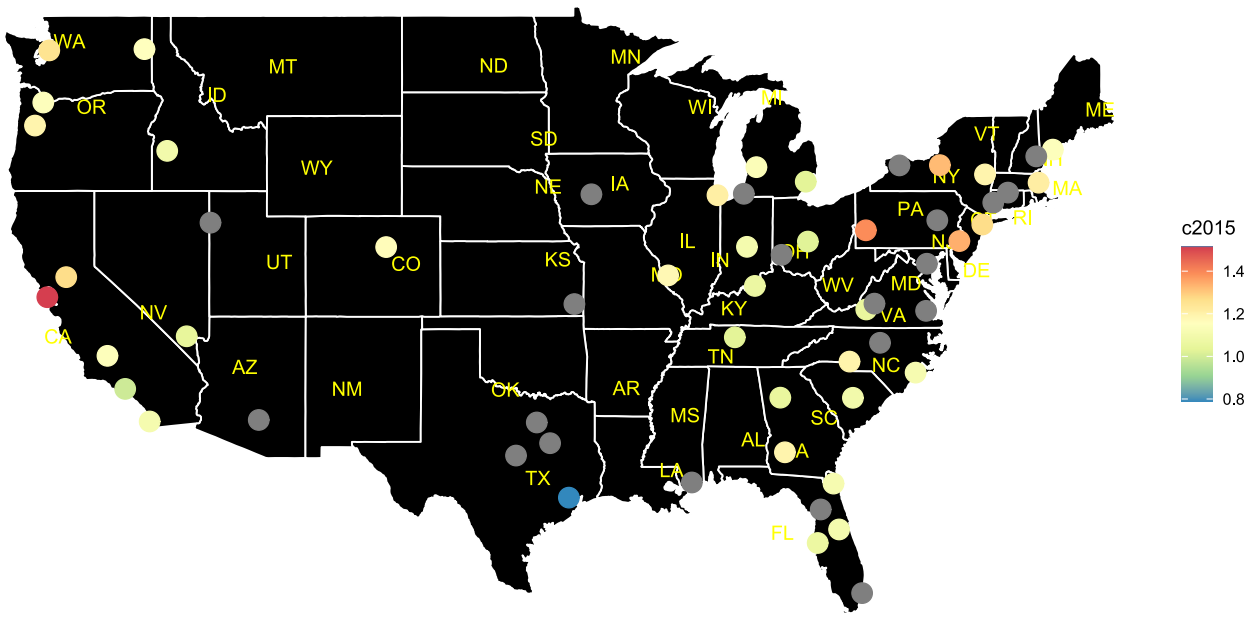
- Model diagnostics are difficult for large data as seen here.
- For large sample size, p-values tend to be all significant.
- Thus it is more effective to check the effect size instead.

```
#print(broom::tidy(M1), n=Inf)
print(summary(M1))
```

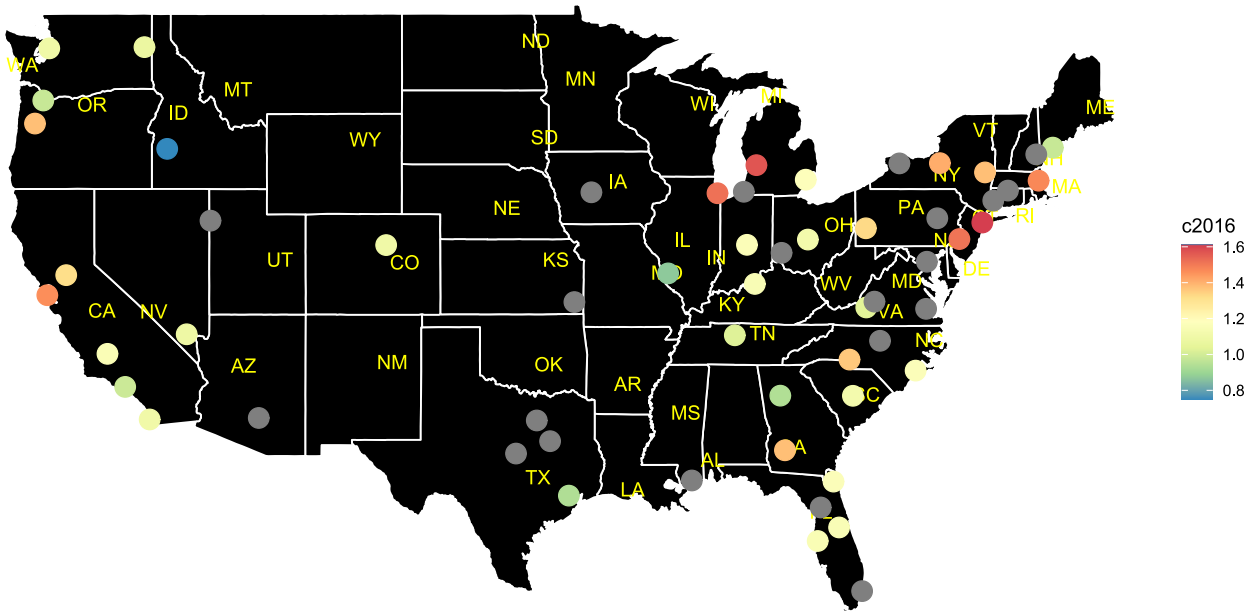
```
Call:
lm(formula = average_price ~ year * region * type, d

Residuals:
     Min       1Q   Median       3Q
```

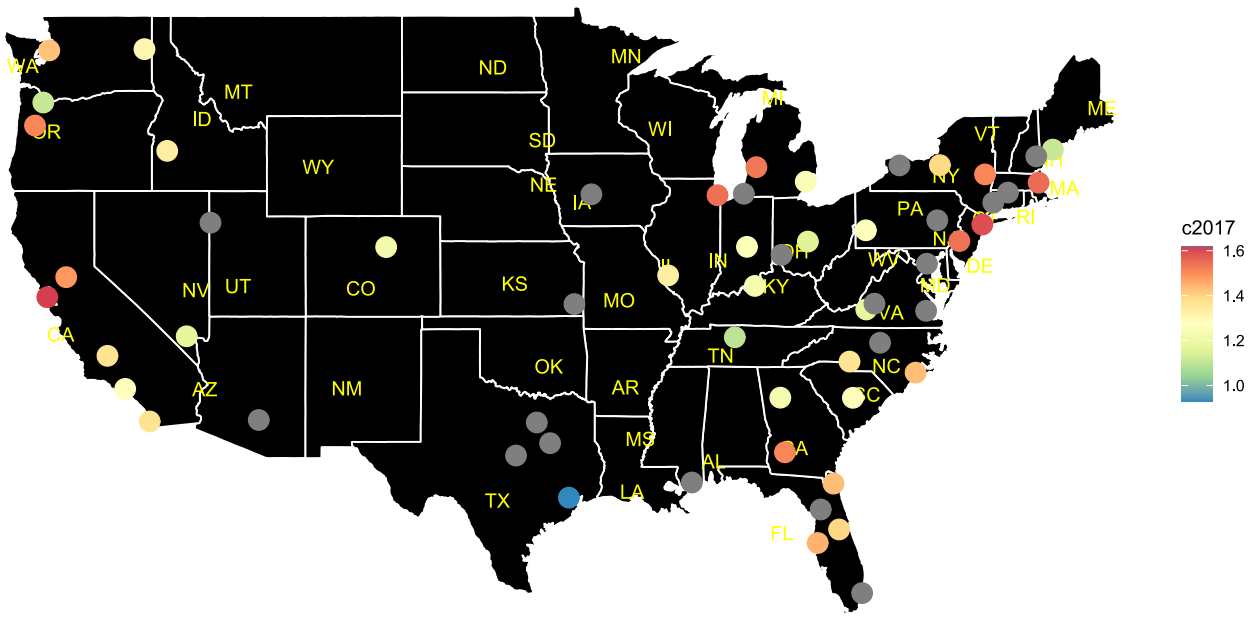

Estimate for Conventional Avocado Price in 2015 July



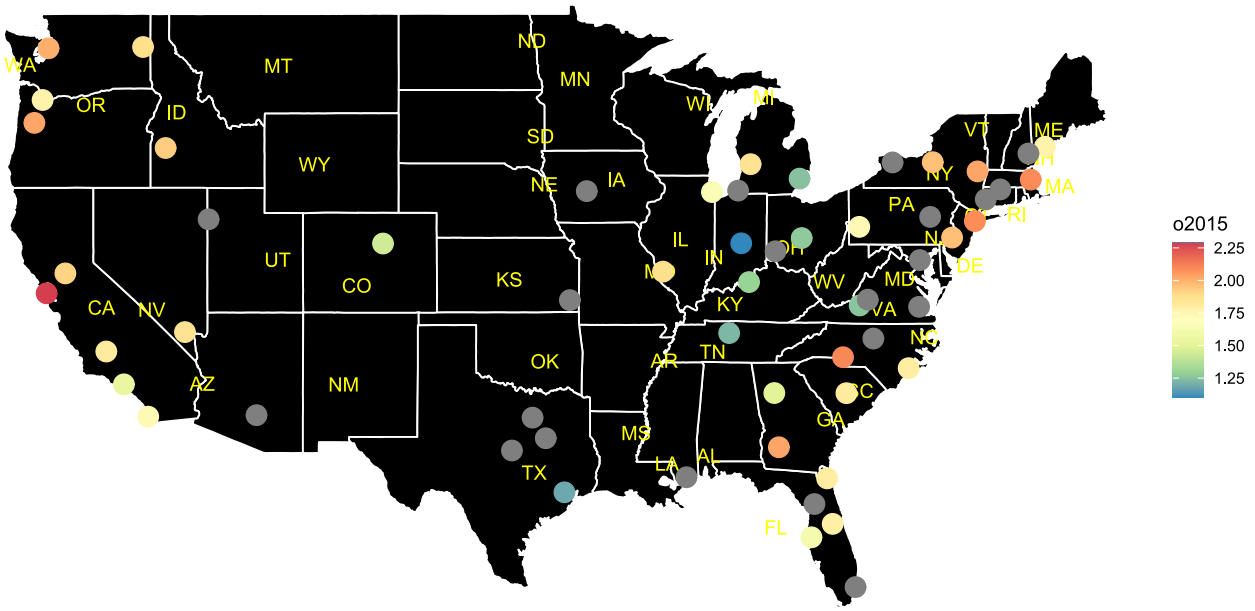
Estimate for Conventional Avocado Price in 2016 July



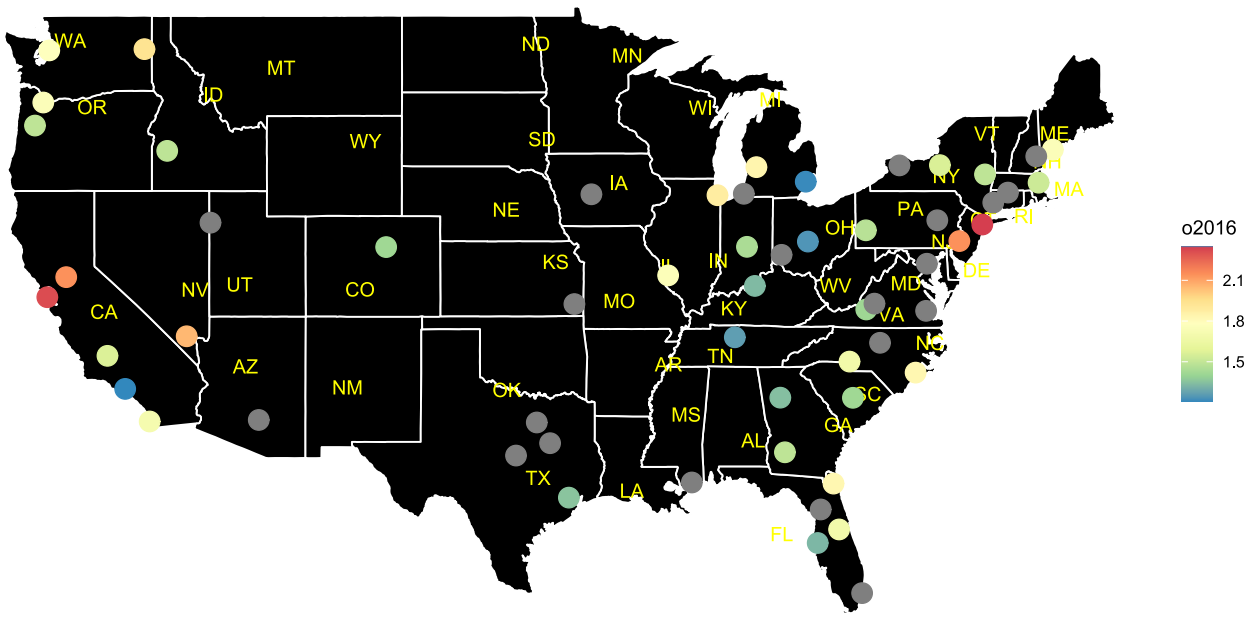
Estimate for Conventional Avocado Price in 2017 July



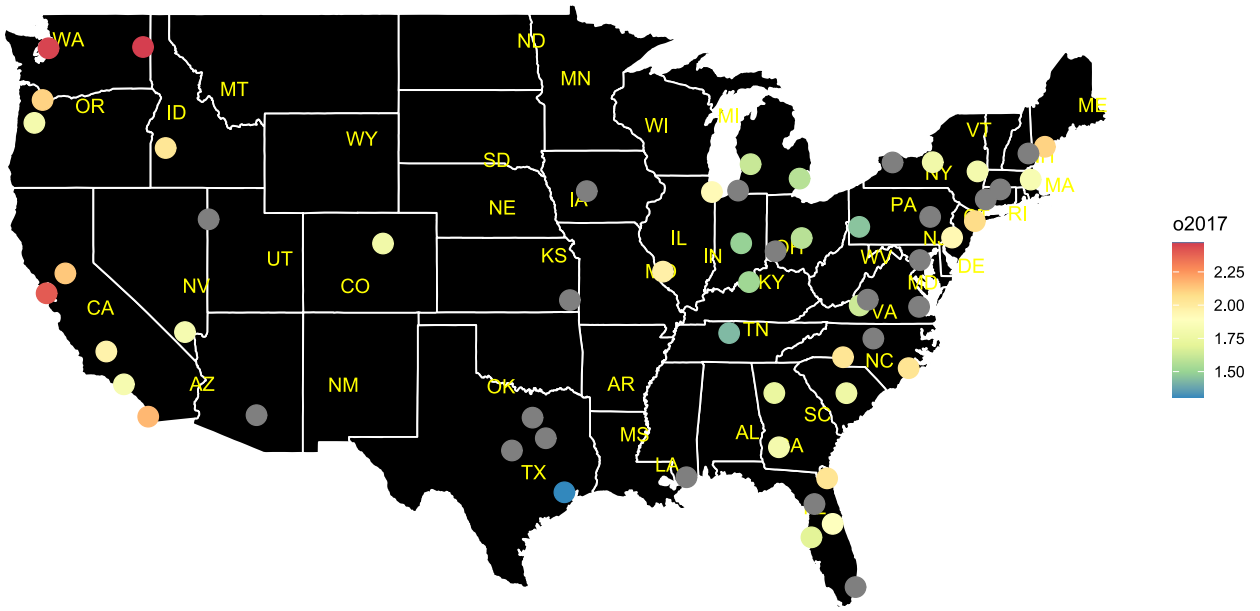
Estimate for Organic Avocado Price in 2015 July



Estimate for Organic Avocado Price in 2016 July



Estimate for Organic Avocado Price in 2017 July



Wait...

- What about the different PLU, Price Look-Up code?
- What about different total volumes etc?
- You can try to predict these values for July 2019.
- Some models have simplifying assumptions that the volume is constant across year or all regions are increasing their volume by the same amount.