Data Wrangling Part 2

Sadaf Zuhra

2024-09-14

Importing Data Set from Google Sheets

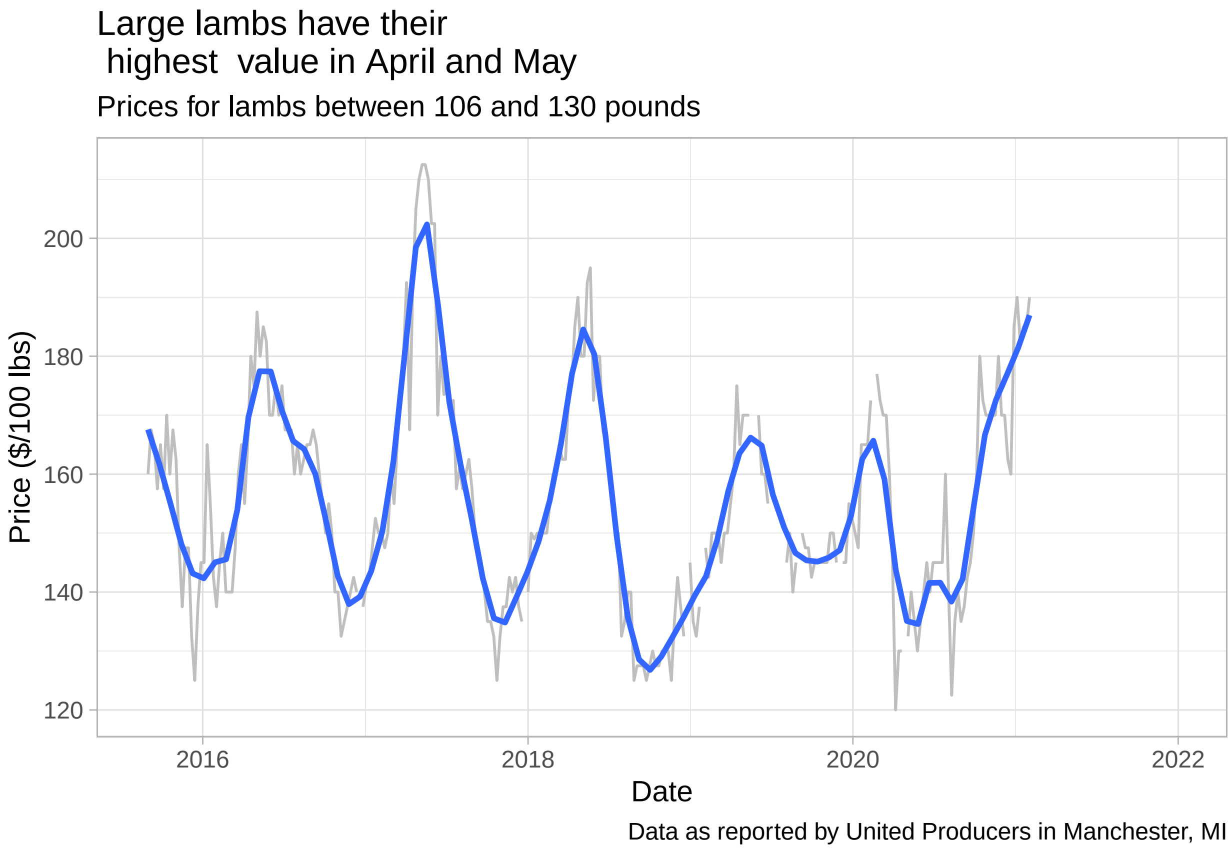
```
library(readxl)
library(tidyverse)
library(googlesheets4)
auction_data <- read_sheet("https://docs.google.com/spreadsheets/d/1_quMjJRBHDLQSmWQouzzyi1DOejAtCZnAee
                           sheet="numbers_and_prices",
                           range="A:J",
                           col_type = "Ddcccccccc",
                           na="NA") %>%
  rename_all(tolower) %>%
  rename("aged_sheep" = "aged sheep",
         "feeder_lambs" = "feeder lambs",
         "hair_lambs" = "hair lambs",
         "new_crop" = "new crop",
         "small" = "40-85",
         "medium" = "85-105",
         "large" = "106-130",
         "extra_large" = ">131")
```

Selecting Large Lamb Data and Seprating Column into "min" & "max"

```
 library(tidyverse)
large_prices <- auction_data %>%
  separate(large, sep="-", into=c("min", "max"), convert=TRUE) %>%
  select(date, min, max)
```

Visualizing Large Price Data

```
library(tidyverse)
library(ggplot2)
large_prices %>%mutate(mid_point = (min + max) / 2) %>%
    ggplot(aes(x=date, y=mid_point)) +
        geom_line(color="gray")+
        geom_smooth(span=0.1, se=FALSE) +
        labs(title="Large lambs have their\n highest  value in April and May",
            subtitle="Prices for lambs between 106 and 130 pounds",
            caption="Data as reported by United Producers in Manchester, MI",
            x="Date",
            y="Price ($/100 lbs)") +
        theme_light()
```



Tidying Data

```
auction_data %>%
    pivot_longer(cols = c(small, medium, large, extra_large), names_to="classes", values_to="price_rang
```

```
## # A tibble: 1,324 x 8
##    date       total aged_sheep feeder_lambs hair_lambs new_crop classes
##    <date>     <dbl> <chr>      <chr>        <chr>      <chr>    <chr>
##  1 2015-08-31  1560 40-100     160-200      160-240    180-250  small
##  2 2015-08-31  1560 40-100     160-200      160-240    180-250  medium
##  3 2015-08-31  1560 40-100     160-200      160-240    180-250  large
##  4 2015-08-31  1560 40-100     160-200      160-240    180-250  extra_large
##  5 2015-09-07  1415 40-90      150-200      150-220     200-240   small
##  6 2015-09-07  1415 40-90      150-200      150-220     200-240   medium
##  7 2015-09-07  1415 40-90      150-200      150-220     200-240   large
##  8 2015-09-07  1415 40-90      150-200      150-220     200-240   extra_large
##  9 2015-09-14  2436 30-100     150-210      150-210     200-300   small
## 10 2015-09-14  2436 30-100     150-210      150-210     200-300   medium
## # i 1,314 more rows
## # i 1 more variable: price_range <chr>
```

```
auction_data %>%
    pivot_longer(cols = c(aged_sheep,feeder_lambs,hair_lambs), names_to="types", values_to="weight_rang
```

```
## # A tibble: 993 x 9
##    date       total small   new_crop medium large extra_large types weight_range
##    <date>     <dbl> <chr>   <chr>    <chr>  <chr> <chr>       <chr> <chr>
##  1 2015-08-31  1560 200-250 180-250  160-1~ 150-~ 150-160     aged~ 40-100
##  2 2015-08-31  1560 200-250 180-250  160-1~ 150-~ 150-160     feed~ 160-200
##  3 2015-08-31  1560 200-250 180-250  160-1~ 150-~ 150-160     hair~ 160-240
##  4 2015-09-07  1415 200-240 200-240  170-2~ 160-~ 160-170     aged~ 40-90
##  5 2015-09-07  1415 200-240 200-240  170-2~ 160-~ 160-170     feed~ 150-200
##  6 2015-09-07  1415 200-240 200-240  170-2~ 160-~ 160-170     hair~ 150-220
##  7 2015-09-14  2436 200-300 200-300  160-1~ 160-~ 160-165     aged~ 30-100
##  8 2015-09-14  2436 200-300 200-300  160-1~ 160-~ 160-165     feed~ 150-210
##  9 2015-09-14  2436 200-300 200-300  160-1~ 160-~ 160-165     hair~ 150-210
## 10 2015-09-21  1455 170-240 140-200  160-1~ 150-~ 150-155     aged~ 40-100
## # i 983 more rows
```

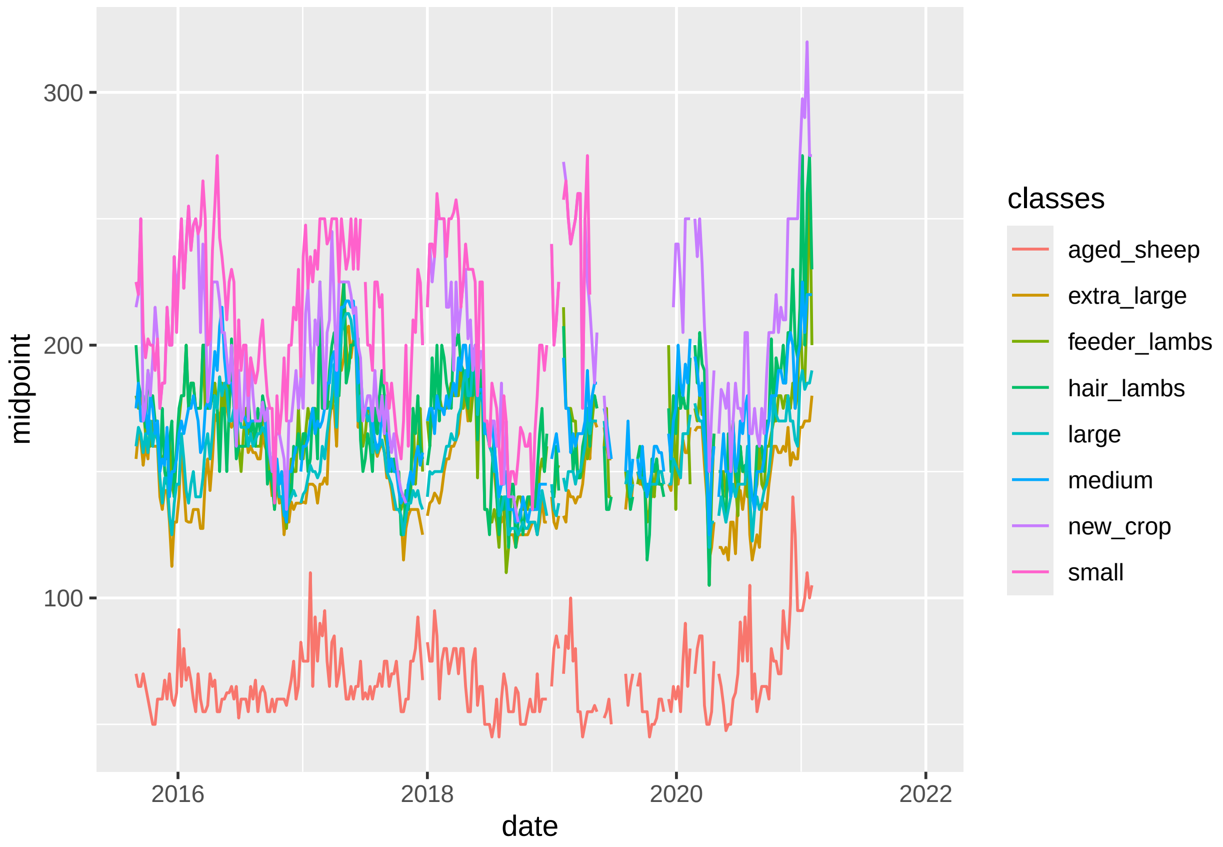
```
library(tidyverse)
auction_data %>%
    pivot_longer(cols = -c(date, total), names_to="classes", values_to="price_range")
```

```
## # A tibble: 2,648 x 4
##    date       total classes      price_range
##    <date>     <dbl> <chr>        <chr>
##  1 2015-08-31  1560 aged_sheep   40-100
##  2 2015-08-31  1560 feeder_lambs 160-200
##  3 2015-08-31  1560 hair_lambs   160-240
##  4 2015-08-31  1560 small        200-250
##  5 2015-08-31  1560 new_crop     180-250
##  6 2015-08-31  1560 medium       160-190
##  7 2015-08-31  1560 large        150-170
##  8 2015-08-31  1560 extra_large  150-160
##  9 2015-09-07  1415 aged_sheep   40-90
## 10 2015-09-07  1415 feeder_lambs 150-200
## # i 2,638 more rows
```

```
tidy_auction_data <- auction_data %>%
    pivot_longer(cols = -c(date, total), names_to="classes", values_to="price_range") %>%
    separate(price_range, sep="-", into=c("min", "max"), convert=TRUE) %>%
    mutate(midpoint = (min + max) / 2)
```

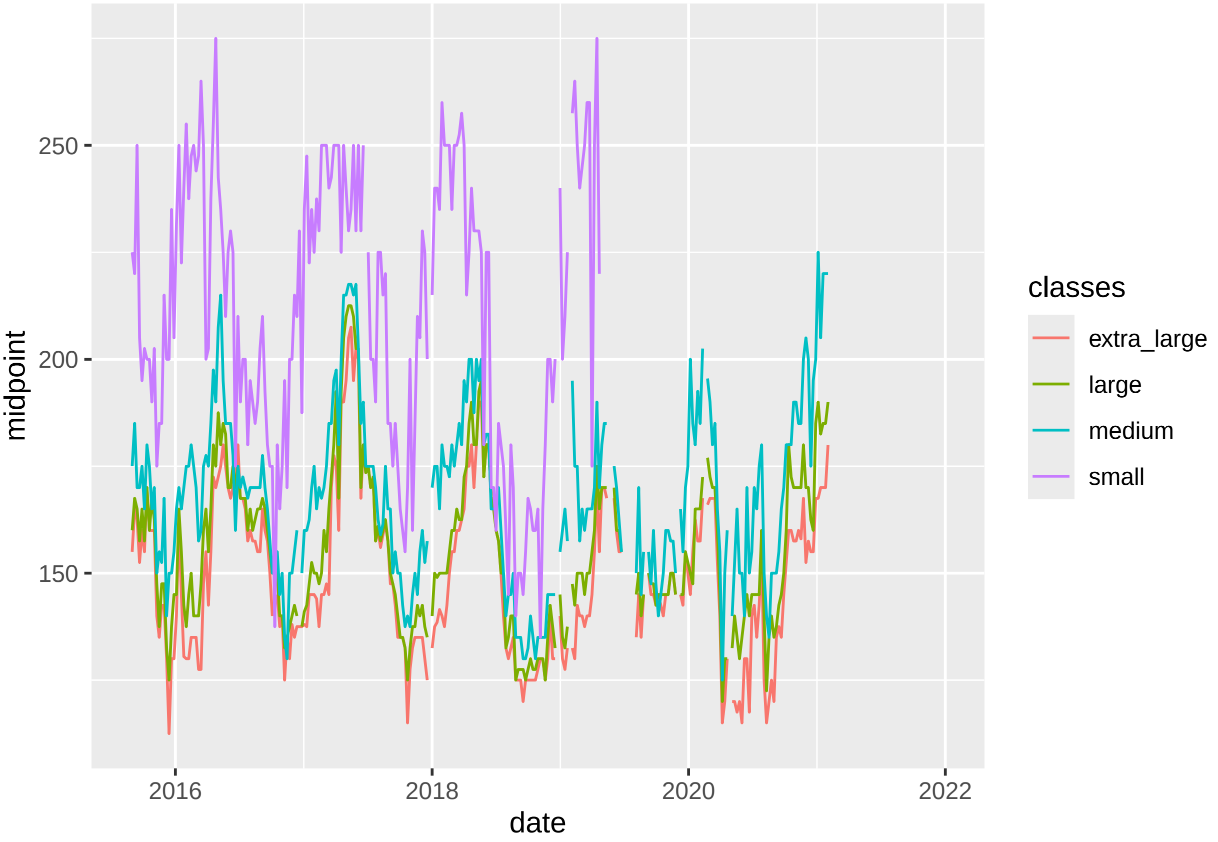
Analysis of Tidy Auction Data

```
tidy_auction_data %>%
    ggplot(aes(x=date, y=midpoint, color=classes)) +
        geom_line()
```

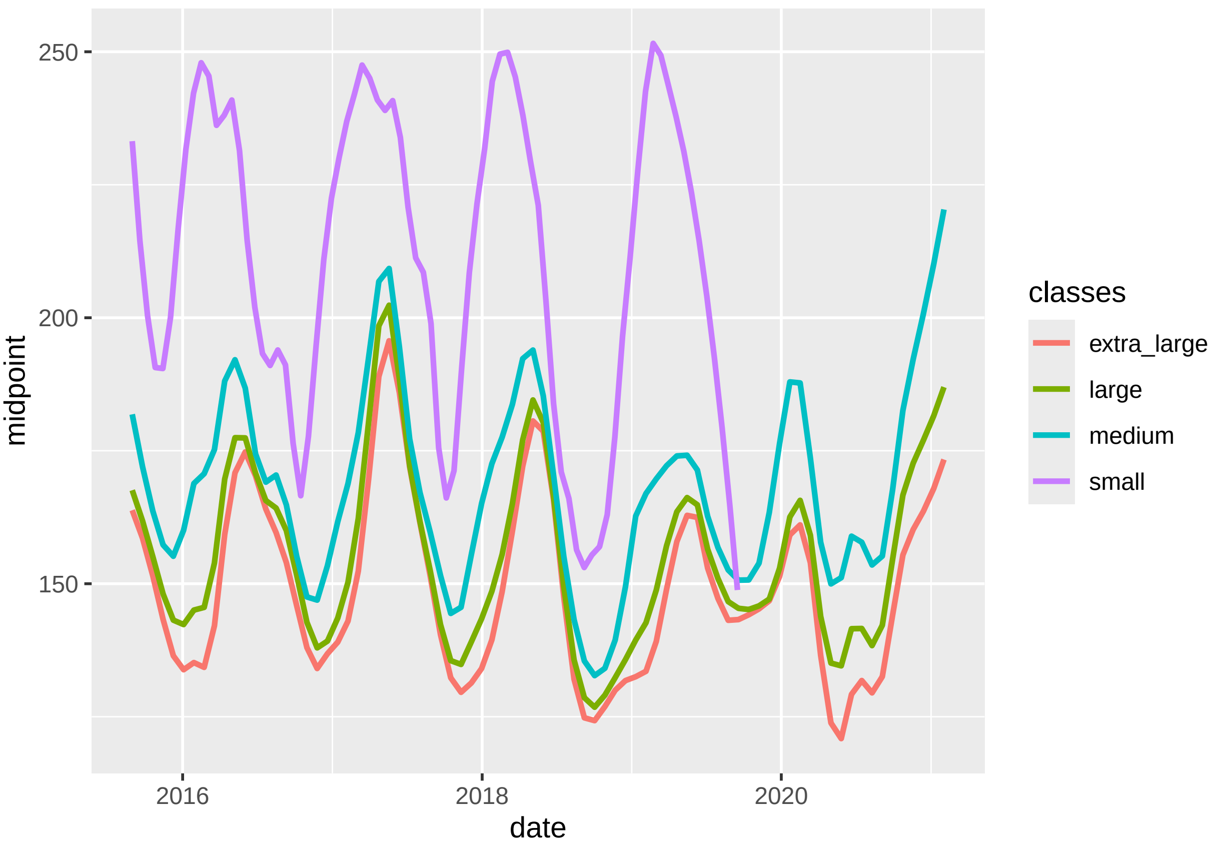


Filtering Data

```
library(tidyverse)
library(ggplot2)
tidy_auction_data %>%
    filter(classes == "small" | classes == "medium" | classes == "large" | classes == "extra_large") %>%
    ggplot(aes(x=date, y=midpoint, color=classes)) +
        geom_line()
```



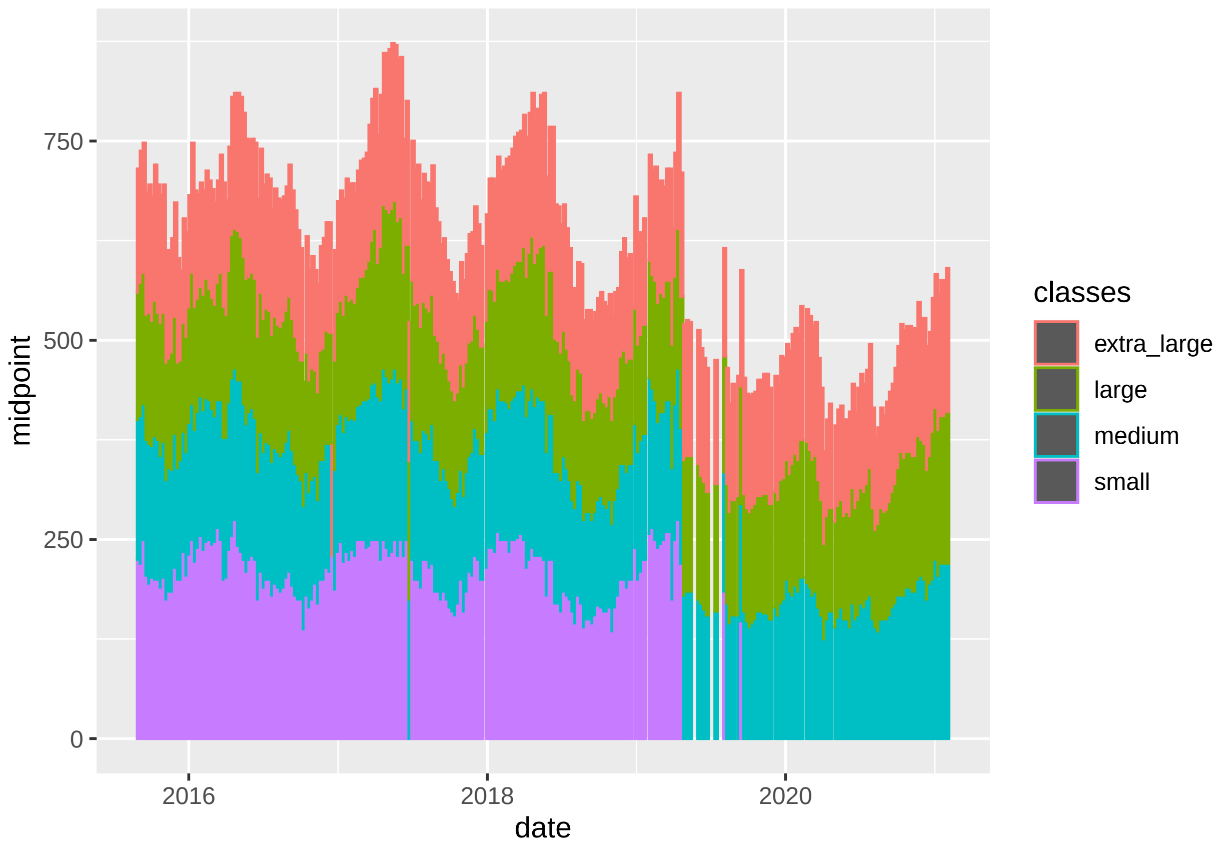
```
library(tidyverse)
library(ggplot2)
 tidy_auction_data %>%
    filter(classes == "small" | classes == "medium" | classes == "large" | classes == "extra_large") %>
    ggplot(aes(x=date, y=midpoint, color=classes)) +
  geom_smooth(span =0.1,se=FALSE)
```



```
    labs(title="Small lambs have the highest price, but\nall lambs peak in Spring",
      x="Date",
      y="Price ($/100 lbs)",
      caption="Data from United Producers in Manchester, MI") +
    theme_light()
```

```
## NULL
```

```
tidy_auction_data %>%
  filter(classes == "small" | classes == "medium" | classes == "large" | classes == "extra_large") %>
  ggplot(aes(x=date, y=midpoint, color=classes)) +geom_bar(stat = "identity")
```



```
tidy_auction_data %>%
  filter(classes=="large") %>%
  mutate(lag0 = midpoint,
         lag1 = lag(midpoint, 1, order_by = date))
```

```
## # A tibble: 331 x 8
##    date       total classes   min   max midpoint  lag0  lag1
##    <date>     <dbl> <chr>   <int> <int>    <dbl> <dbl> <dbl>
##  1 2015-08-31  1560 large     150   170     160   160     NA
##  2 2015-09-07  1415 large     160   175     168.  168.  160
##  3 2015-09-14  2436 large     160   170     165   165   168.
##  4 2015-09-21  1455 large     150   165     158.  158.  165
##  5 2015-09-28  1079 large     150   180     165   165   158.
##  6 2015-10-05  1205 large     150   165     158.  158.  165
##  7 2015-10-12   944 large     160   180     170   170   158.
##  8 2015-10-19  2506 large     150   170     160   160   170
##  9 2015-10-26  1985 large     160   175     168.  168.  160
## 10 2015-11-02  1214 large     155   170     162.  162.  168.
## # i 321 more rows
```

```
library(tidyverse)
tidy_auction_data %>%  filter(classes=="large") %>%
  mutate(lag0 = midpoint,
         lag1 = lag(midpoint, 1, order_by=date),
         lag2 = lag(midpoint, 2, order_by=date),
         lag3 = lag(midpoint, 3, order_by=date))
```

```
## # A tibble: 331 x 10
##    date       total classes   min   max midpoint  lag0  lag1  lag2  lag3
##    <date>     <dbl> <chr>   <int> <int>    <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 2015-08-31  1560 large     150   170     160   160    NA    NA    NA
##  2 2015-09-07  1415 large     160   175     168.  168.  160    NA    NA
##  3 2015-09-14  2436 large     160   170     165   165   168.  160    NA
##  4 2015-09-21  1455 large     150   165     158.  158.  165   168.  160
##  5 2015-09-28  1079 large     150   180     165   165   158.  165   168.
##  6 2015-10-05  1205 large     150   165     158.  158.  165   158.  165
##  7 2015-10-12   944 large     160   180     170   170   158.  165   158.
##  8 2015-10-19  2506 large     150   170     160   160   170   158.  165
##  9 2015-10-26  1985 large     160   175     168.  168.  160   170   158.
## 10 2015-11-02  1214 large     155   170     162.  162.  168.  160   170
## # i 321 more rows
```

```
tidy_auction_data %>%
  filter(classes=="large") %>%
    mutate(lag0 = midpoint,
        lag1 = lag(midpoint, 1, order_by=date),
        lag2 = lag(midpoint, 2, order_by=date),
        lag3 = lag(midpoint, 3, order_by=date)) %>%
    mutate(rolling_average = mean(c(lag0, lag1, lag2, lag3)))
```

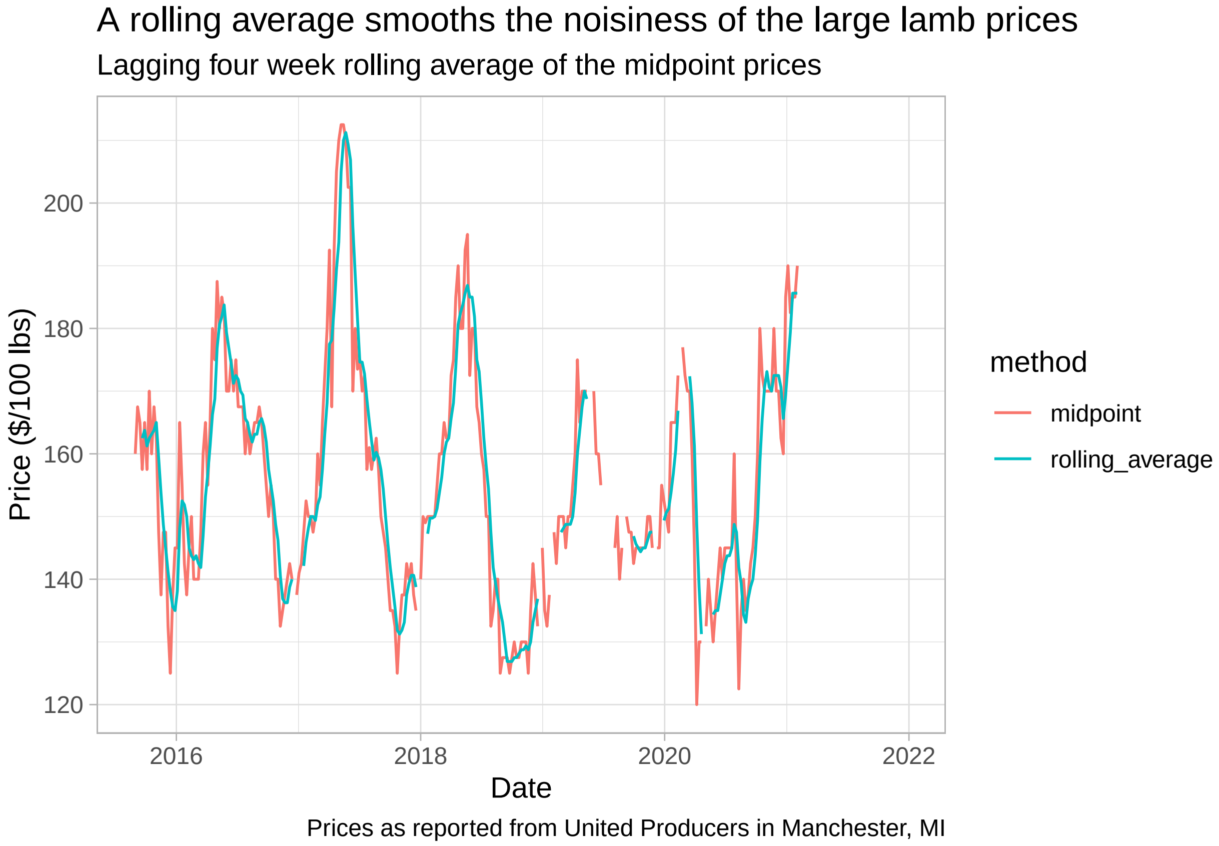
```
## # A tibble: 331 x 11
##    date       total classes   min   max midpoint  lag0  lag1  lag2  lag3
##    <date>     <dbl> <chr>   <int> <int>    <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 2015-08-31  1560 large     150   170     160   160    NA    NA    NA
##  2 2015-09-07  1415 large     160   175     168.  168.  160    NA    NA
##  3 2015-09-14  2436 large     160   170     165   165   168.  160    NA
##  4 2015-09-21  1455 large     150   165     158.  158.  165   168.  160
##  5 2015-09-28  1079 large     150   180     165   165   158.  165   168.
##  6 2015-10-05  1205 large     150   165     158.  158.  165   158.  165
##  7 2015-10-12   944 large     160   180     170   170   158.  165   158.
##  8 2015-10-19  2506 large     150   170     160   160   170   158.  165
##  9 2015-10-26  1985 large     160   175     168.  168.  160   170   158.
## 10 2015-11-02  1214 large     155   170     162.  162.  168.  160   170
## # i 321 more rows
## # i 1 more variable: rolling_average <dbl>
```

```
large <- tidy_auction_data %>%
    filter(classes=="large") %>%
    mutate(lag0 = midpoint,
        lag1 = lag(midpoint, 1, order_by=date),
        lag2 = lag(midpoint, 2, order_by=date),
        lag3 = lag(midpoint, 3, order_by=date)) %>%
  group_by(date) %>%
  summarise(midpoint=first(midpoint),rolling_average=mean(c(lag0,lag1,lag2,lag3)))
```

```
large <- tidy_auction_data %>%
  filter(classes=="large") %>%
    mutate(lag0 = midpoint,
        lag1 = lag(midpoint, 1, order_by=date),
        lag2 = lag(midpoint, 2, order_by=date),
        lag3 = lag(midpoint, 3, order_by=date)) %>%
    mutate(rolling_average = (lag0+ lag1+ lag2+ lag3)/4) %>%
  select(date,midpoint,rolling_average)
```

```
large %>%
    pivot_longer(-date, names_to="method", values_to="price") %>%
    ggplot(aes(x=date, y=price, color=method)) +
        geom_line() +
        theme_light() +
        labs(x="Date",
            y="Price ($/100 lbs)",
            title="A rolling average smooths the noisiness of the large lamb prices",
            subtitle="Lagging four week rolling average of the midpoint prices",
            caption="Prices as reported from United Producers in Manchester, MI")
```

```
## Warning: Removed 97 rows containing missing values or values outside the scale range
## (‘geom_line()‘).
```


A rolling average smooths the noisiness of the large lamb prices
Lagging four week rolling average of the midpoint prices
Price ($/100 lbs)
Date
method
midpoint
rolling_average
Prices as reported from United Producers in Manchester, MI