Big Pumpkins

Sergio Aquino

18/10/2021

Load libraries

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggfortify)
```

Import data and clean it

```
## Rows: 28065 Columns: 14

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (14): id, place, weight_lbs, grower_name, city, state_prov, country, gpc...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.

## Rows: 28,065
## Columns: 15
## $ year              <chr> "2013", "2013", "2013", "2013", "2013", "2013", "201~
## $ type              <chr> "Field Pumpkin", "Field Pumpkin", "Field Pumpkin", "~
## $ place             <chr> "1", "2", "3", "4", "5", "5", "7", "8", "9", "10", "~
## $ weight_lbs        <dbl> 154.5, 146.5, 145.0, 140.8, 139.0, 139.0, 136.5, 136~
## $ grower_name       <chr> "Ellenbecker, Todd & Sequoia", "Razo, Steve", "Ellen~
## $ city              <chr> "Gleason", "New Middletown", "Glenson", "Combined Lo~
```

```
## $ state_prov        <chr> "Wisconsin", "Ohio", "Wisconsin", "Wisconsin", "Wisc~
## $ country           <chr> "United States", "United States", "United States", "~
## $ gpc_site          <chr> "Nekoosa Giant Pumpkin Fest", "Ohio Valley Giant Pum~
## $ seed_mother       <chr> "209 Werner", "150.5 Snyder", "209 Werner", "109 Mar~
## $ pollinator_father <chr> "Self", NA, "103 Mackinnon", "209 Werner '12", "open~
## $ ott               <dbl> 184, 194, 177, 194, 0, 190, 190, 182, 0, 0, 0, 177, ~
## $ est_weight        <dbl> 129, 151, 115, 151, 0, 141, 142, 124, 0, 0, 0, 115, ~
## $ pct_chart         <dbl> 20, -3, 26, -7, 0, -1, -4, 10, 0, 0, 0, 14, 4, 8, 14~
## $ variety           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
```

Explore data

how good are the estimates?

```
pumpkins_clean3 %>%
  group_by(type) %>%
  summarise(
          count = n(),
          mean_estimate = round(mean(est_weight, na.rm = T)),
          mean_weight = round(mean(weight_lbs))
            )
```

```
## # A tibble: 6 x 4
##   type             count mean_estimate mean_weight
##   <chr>            <int>         <dbl>       <dbl>
## 1 Field Pumpkin     2756            38          80
## 2 Giant Pumpkin    15965           697         777
## 3 Giant Squash      1686           430         525
## 4 Giant Watermelon  2527            97         128
## 5 Long Gourd        1965             2          95
## 6 Tomato            3166           NaN           4
```

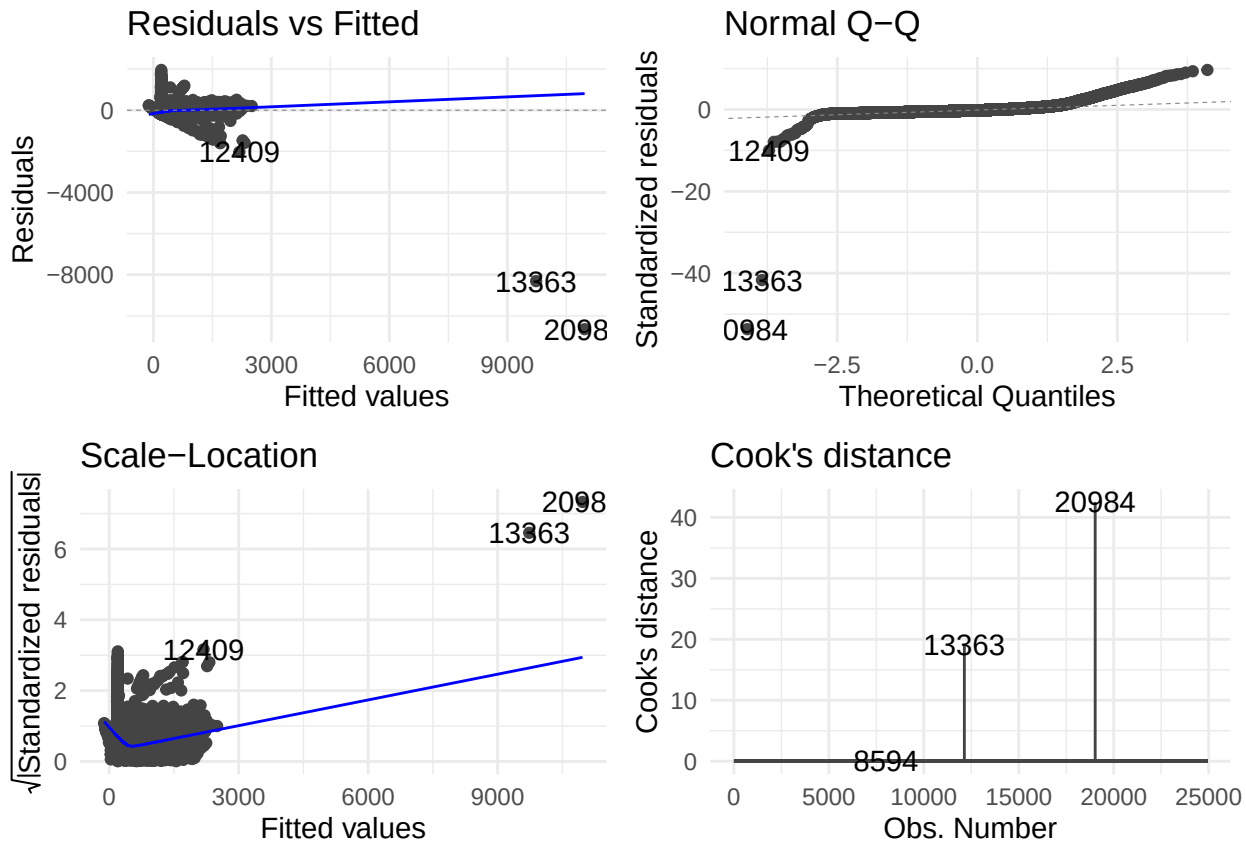
are ott and est_weight good predictors as per tidytuesday site?

```
model1 <- lm(weight_lbs ~ est_weight + ott, data = pumpkins_clean3)
summary(model1)
```

```
##
## Call:
## lm(formula = weight_lbs ~ est_weight + ott, data = pumpkins_clean3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10648.6    -90.9    -43.5     34.7   1957.5
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 200.472738   2.190530   91.52   <2e-16 ***
```

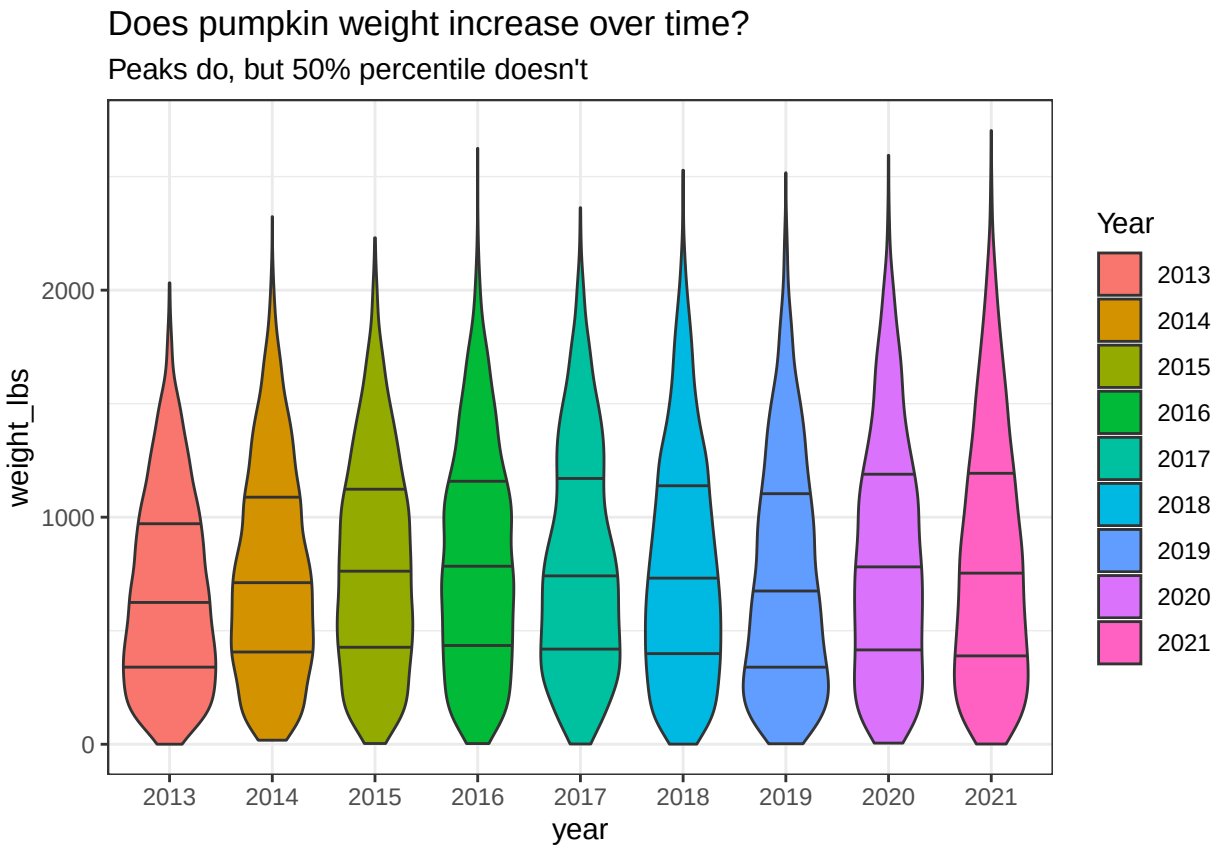
```
## est_weight    1.053247   0.005294  198.97   <2e-16 ***
## ott          -0.756509   0.017890  -42.29   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 202.7 on 24896 degrees of freedom
##   (3166 observations deleted due to missingness)
## Multiple R-squared:  0.8364, Adjusted R-squared:  0.8364
## F-statistic: 6.364e+04 on 2 and 24896 DF,  p-value: < 2.2e-16
```

```
autoplot(model1, which = 1:4) + theme_minimal()
```



does pumpkin weight increase over time?

```
pumpkins_clean3 %>%
  filter(type == "Giant Pumpkin") %>%
  ggplot() +
  geom_violin(aes(x = year, y = weight_lbs, fill = year),draw_quantiles = c(0.25, 0.5, 0.75) )+
   guides(fill=guide_legend(title="Year")) +
  labs(title = "Does pumpkin weight increase over time?", subtitle = "Peaks do, but 50% percentile doesn
  theme_bw()
```


Does pumpkin weight increase over time?
Peaks do, but 50% percentile doesn't
weight_lbs
2000
1000
0
2013
2014
2015
2016
2017
2018
2019
2020
2021
year
Year
2013
2014
2015
2016
2017
2018
2019
2020
2021