Class Exercises 9/23

Allyson Beach

Read In Advance Dataset: Railroad

```
# OPTION 1: select names of column names that you want - not good for lots of data
trains <- read_excel("../_data/StateCounty2012.xls", skip=3) %>%
  select(STATE, COUNTY, TOTAL)
```

```
## New names:
## * `` -> ...2
## * `` -> ...4
```

```
# further improvements - insert column names and call certain ones delete to remove later
```

Notes, select is for columns and filter is for rows

```
# OPTION 2: name columns to make it easy to delete them  - be sure to skip the old col names
trains <- read_excel("../_data/StateCounty2012.xls", skip=4,
                     col_names = c("State", "D1", "County", "D2", "Total")) %>%
  select(!starts_with("d")) %>%
  filter(!str_detect(State, "[Tt]otal")) #will look for both cases Total or total
trains
```

```
## # A tibble: 2,933 x 3
##    State County               Total
##    <chr> <chr>                <dbl>
##  1 AE    APO                      2
##  2 AK    ANCHORAGE                7
##  3 AK    FAIRBANKS NORTH STAR     2
##  4 AK    JUNEAU                   3
##  5 AK    MATANUSKA-SUSITNA        2
##  6 AK    SITKA                    1
##  7 AK    SKAGWAY MUNICIPALITY    88
##  8 AL    AUTAUGA                102
##  9 AL    BALDWIN                143
## 10 AL    BARBOUR                  1
## # ... with 2,923 more rows
```

Next to try is the Australian Marriage Data Figure out the structure of the data for end results. The case are the things that uniquely id a value. Separate out the left col - into division and state

Another one to try is the organiceggpoultry.xlsx - have to pivot (so maybe next week)

egg - what is the case? for product, year, the price per month? so price is our value variables are year, month, product (amount[dozen/half] and size[large/xlarge]) SO, the case is year, month, size, and amount with the value of price end product should have 5 columns

for month, pivot longer

```
# here we skip the empty rows and the inaccurate header row, then we rename the first column with the h
eggs_dirty <- read_excel("../_data/organiceggpoultry.xls", sheet=1, skip=4) %>%
  rename(egg_month = 1)
```

```
## New names:
## * `` -> ...1
## * `` -> ...6
```

```
# show the raw data
print(eggs_dirty, width = Inf)
```

```
## # A tibble: 120 x 11
##    egg_month `Extra Large \nDozen` `Extra Large 1/2 Doz.\n1/2 Dozen`
##    <chr>                     <dbl>                              <dbl>
##  1 Jan 2004                   230                               132
##  2 February                   230                               134.
##  3 March                      230                               137
##  4 April                      234.                              137
##  5 May                        236                               137
##  6 June                       241                               137
##  7 July                       241                               137
##  8 August                     241                               137
##  9 September                  241                               136.
## 10 October                    241                               136.
##    `Large \nDozen` `Large \n1/2 Doz.` ...6  Whole `B/S Breast` `Bone-in Breast`
##              <dbl>              <dbl> <lgl> <dbl>        <dbl> <chr>
##  1            230                126  NA     198.         646. too few
##  2            226.               128. NA     198.         642. too few
##  3            225                131  NA     209          642. too few
##  4            225                131  NA     212          642. too few
##  5            225                131  NA     214.         642. too few
##  6            231.               134. NA     216.         641  too few
##  7            234.               134. NA     217          642. 390.5
##  8            234.               134. NA     217          642. 390.5
##  9            234.               130. NA     217          642. 390.5
## 10            234.               128. NA     217          642. 390.5
##    `Whole Legs` Thighs
##           <dbl> <chr>
##  1         194. too few
##  2         194. 203
##  3         194. 203
##  4         194. 203
##  5         194. 203
##  6         202. 200.375
##  7         204. 199.5
##  8         204. 199.5
##  9         204. 199.5
```

```
## 10         204. 199.5
## # ... with 110 more rows
```

```
# we then take out any columns that have only NA - compare #NA to #rows
eggs_dirty <- eggs_dirty %>% select_if(!colSums(is.na(eggs_dirty)) == nrow(eggs_dirty))

# convert all types to characters so we can pivot, then we pivot the products to price and replace all
eggs_dirty <- eggs_dirty %>% mutate(across(where(is.double), as.character)) %>% pivot_longer(cols = con
  pivot_longer(cols=contains(c("whole", "breast", "leg", "thigh")), names_to = "chicken products", valu
  mutate(`price per carton` = str_replace(`price per carton`, "[a-zA-Z ]+", "0")) %>%
  mutate(`price per lb`= str_replace(`price per lb`, "[a-zA-Z ]+", "0"))

# take out any of the "/1" that got filled in instead of the year (month (o)year|/1) to (month (o)year)
eggs_dirty <- eggs_dirty %>% mutate(egg_month = str_remove(egg_month, "/[0-9]+")) %>%
  separate(egg_month, c("month", "year"), extra = "drop", fill = "right")

# filling in the year from the jan month to the rest of the months, default direction is down
eggs_dirty <- eggs_dirty %>% fill(year)

# further separate the products into type and amount?? not sure
eggs_clean <- eggs_dirty %>%
  mutate(`price per carton`, `price per carton`=as.double(`price per carton`)) %>%
  mutate(`price per lb`, `price per lb`=as.double(`price per lb`))

# show the clean data
eggs_clean
```

```
## # A tibble: 2,400 x 6
##    month year  ‘egg products‘   ‘price per cart~ ‘chicken produc~ ‘price per lb‘
##    <chr> <chr> <chr>                       <dbl> <chr>                     <dbl>
##  1 Jan   2004  "Extra Large \n~              230 Whole                     198.
##  2 Jan   2004  "Extra Large \n~              230 Whole Legs                194.
##  3 Jan   2004  "Extra Large \n~              230 B/S Breast                646.
##  4 Jan   2004  "Extra Large \n~              230 Bone-in Breast              0
##  5 Jan   2004  "Extra Large \n~              230 Thighs                      0
##  6 Jan   2004  "Extra Large 1/~              132 Whole                     198.
##  7 Jan   2004  "Extra Large 1/~              132 Whole Legs                194.
##  8 Jan   2004  "Extra Large 1/~              132 B/S Breast                646.
##  9 Jan   2004  "Extra Large 1/~              132 Bone-in Breast              0
## 10 Jan   2004  "Extra Large 1/~              132 Thighs                      0
## # ... with 2,390 more rows
```