Module 6 Homework: Tidying cotton data

Natalie Nelson, PhD

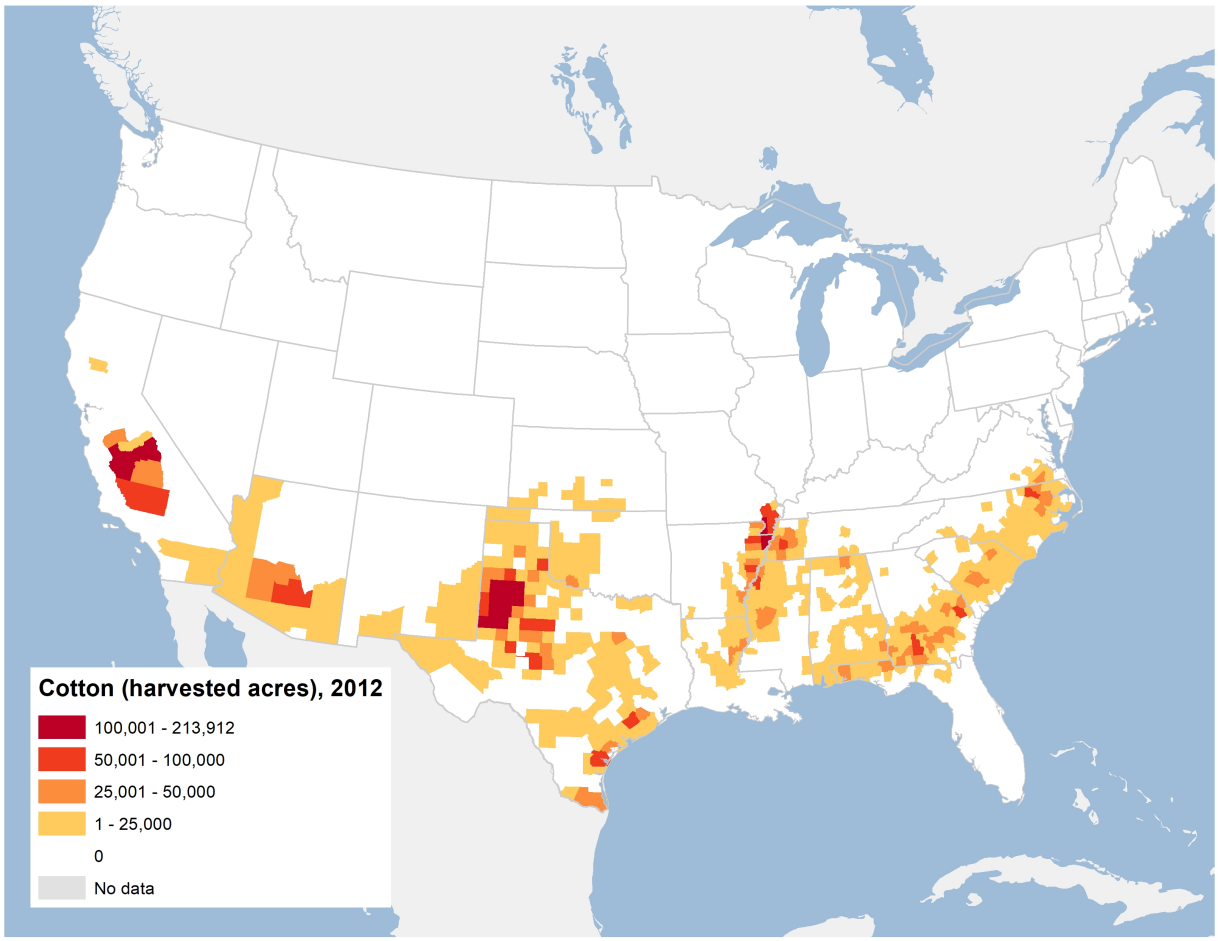
Source: USDA AMS

1. Overview

In this homework assignment, you will tidy a dataset from the USDA National Agricultural Statistics Service. Specifically, the dataset includes annual observations of yield (lbs/acre) and area harvested (acres) of cotton for each county and agricultural district across the U.S. In many locations, the data date back to the early 1900s. The primary objective of this homework assignment is to answer the following questions:

- How have yield and area harvested of cotton changed across all of the NC agricultural districts over time?
- What were the top 3 cotton producing counties in NC in terms of total lbs of cotton for 2018?

To complete this assignment, fill out the `cotton-script.R` file included in the `code` subdirectory of the Module 6 GitHub repository. You will push your changes to the `cotton-script.R` file to your GitHub account, and submit a link to your GitHub account on the Moodle site. Please keep your code organized and annotated, and include all code to perform the tasks outlined in this assignment.



Source: USDA ERS

2. Read and inspect your data

The data are in a file named `cotton-usda-nass.csv` in the `data` subdirectory. Load the data and run through the first few items of the Exploratory Data Analysis checklist: check the packaging `str()`, look at the top and bottom of your data `head()` `tail()`, check your n's `dim()`, and validate with an external data source `summary()`.

There are several columns in the cotton dataset. The ones that are relevant to this analysis are as follows:

- `year`: The year in which the measurements were collected. Should be `dbl` data type.
- `state`: The state in which the measurements were collected. Should be `chr` data type.
- `ag_district`: The agricultural district in which the measurements were collected; each district is organized geographically and composed of several counties. Should be `chr` data type.
- `county`: The county in which the measurements were collected. Should be `chr` data type.
- `data_item`: The kind of measurement, either acres harvested or yield. Should be `chr` data type.
- `value`: The measurement of either acres harvested or yield. Should be `dbl` data type.

When you inspect your data, it's very important that you always check to see what the data type is in each column. Are the data types what you would expect? In particular, are columns with numeric measurements

shown as a numeric data type (`int`, `dbl`, or `num`) or as text (`chr`, `fct`)? In this case, `value` contains some text, so the data in this column are being read as `chr`. You will address this issue in the exercise.

3. Tidy the data

3.1. Create a NC data subset

`filter` to rows that correspond to measurements from NC and `select` only the columns that are needed (in the bulleted list above).

```
## # A tibble: 6 x 6
##    year state        ag_district    county  data_item                value
##   <dbl> <chr>        <chr>          <chr>   <chr>                    <chr>
## 1  2017 NORTH CAROL~ CENTRAL COAST~ BEAUFO~ COTTON, UPLAND - ACRES H~ 9236
## 2  2017 NORTH CAROL~ CENTRAL COAST~ CARTER~ COTTON, UPLAND - ACRES H~ (D)
## 3  2017 NORTH CAROL~ CENTRAL COAST~ CRAVEN  COTTON, UPLAND - ACRES H~ 2861
## 4  2017 NORTH CAROL~ CENTRAL COAST~ GREENE  COTTON, UPLAND - ACRES H~ 5706
## 5  2017 NORTH CAROL~ CENTRAL COAST~ HYDE    COTTON, UPLAND - ACRES H~ 8094
## 6  2017 NORTH CAROL~ CENTRAL COAST~ JOHNST~ COTTON, UPLAND - ACRES H~ 4547
```

3.2. Divide the `data_item` column

The `data_item` column contains two different types of information: the type of cotton (`COTTON, UPLAND`) and the measurement type (`ACRES HARVESTED` and `YIELD, MEASURED IN LB / ACRE`). `separate` the `data_item` column into `cotton_type` and `measurement`. Note that you will have to specify the `sep` argument in `separate`. **Include spaces in the `sep` argument**, otherwise the spaces will be included in the entries in the `cotton_type` and `measurement` columns.

```
## # A tibble: 6 x 7
##    year state       ag_district    county  cotton_type  measurement   value
##   <dbl> <chr>       <chr>          <chr>   <chr>        <chr>         <chr>
## 1  2017 NORTH CARO~ CENTRAL COAST~ BEAUFO~ COTTON, UPL~ ACRES HARVES~ 9236
## 2  2017 NORTH CARO~ CENTRAL COAST~ CARTER~ COTTON, UPL~ ACRES HARVES~ (D)
## 3  2017 NORTH CARO~ CENTRAL COAST~ CRAVEN  COTTON, UPL~ ACRES HARVES~ 2861
## 4  2017 NORTH CARO~ CENTRAL COAST~ GREENE  COTTON, UPL~ ACRES HARVES~ 5706
## 5  2017 NORTH CARO~ CENTRAL COAST~ HYDE    COTTON, UPL~ ACRES HARVES~ 8094
## 6  2017 NORTH CARO~ CENTRAL COAST~ JOHNST~ COTTON, UPL~ ACRES HARVES~ 4547
```

3.3. Convert the `value` column to numeric type

The `value` column has `(D)` in several cells, resulting in the column being read as `chr`. `(D)` is entered when data are withheld by the USDA to avoid disclosing information on individual operations; said another way, if there are only a few operations in a county, it would be easy to attribute the county-level measurement with a specific operation, which violates the terms of the USDA data as operation-level data are private. R will never be able to convert the `value` column to numeric while there are text values in the column - it doesn't know how to convert text to numbers!

First, `filter` the data to remove entries when `value` equals `(D)`. Then, convert the `value` column to numeric using `as.numeric()`. Remember that you have to *redefine* the value column after running `as.numeric()`. Confirm that you have converted the column successfully.

```
## # A tibble: 6 x 7
##    year state       ag_district    county  cotton_type  measurement   value
##   <dbl> <chr>       <chr>          <chr>   <chr>        <chr>         <dbl>
## 1  2017 NORTH CARO~ CENTRAL COAST~ BEAUFO~ COTTON, UPL~ ACRES HARVES~  9236
```

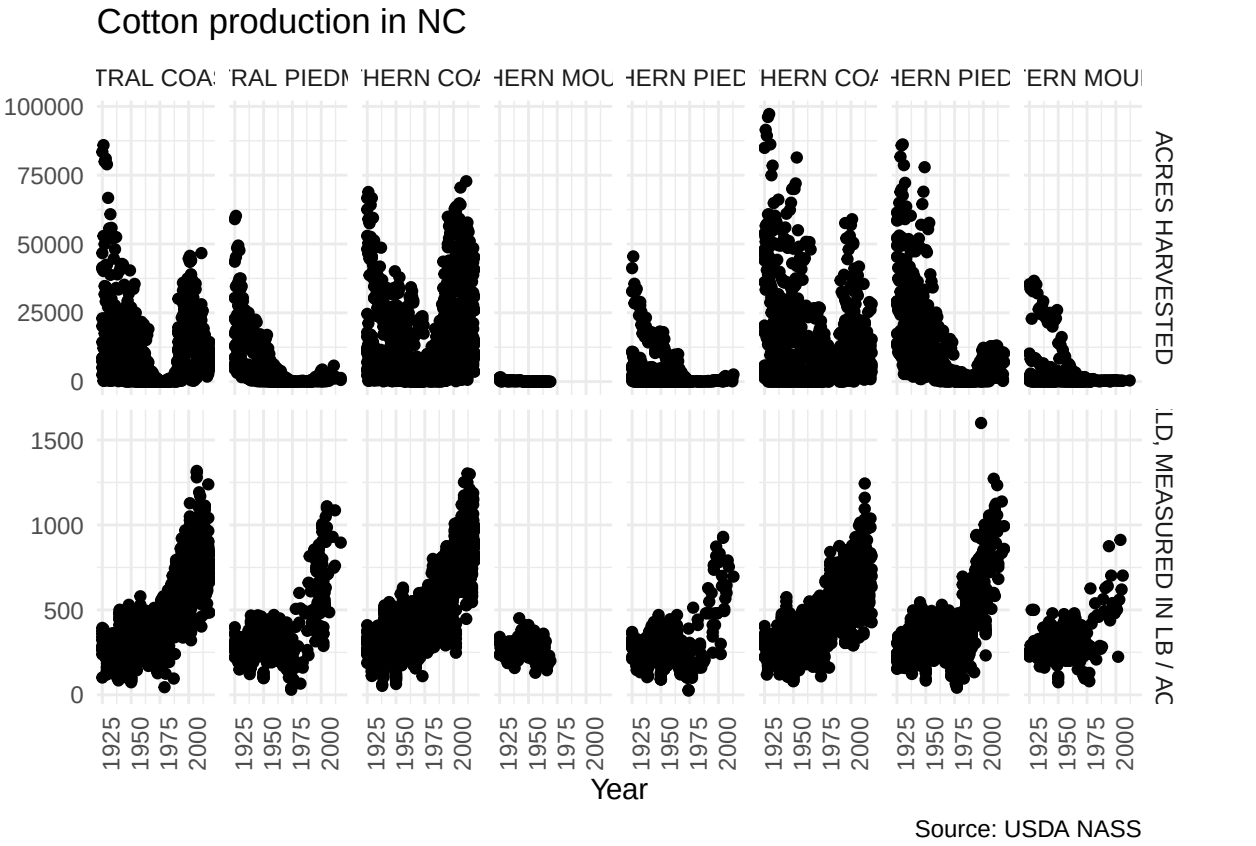
```
## 2  2017 NORTH CARO~ CENTRAL COAST~ CRAVEN  COTTON, UPL~ ACRES HARVES~  2861
## 3  2017 NORTH CARO~ CENTRAL COAST~ GREENE  COTTON, UPL~ ACRES HARVES~  5706
## 4  2017 NORTH CARO~ CENTRAL COAST~ HYDE    COTTON, UPL~ ACRES HARVES~  8094
## 5  2017 NORTH CARO~ CENTRAL COAST~ JOHNST~ COTTON, UPL~ ACRES HARVES~  4547
## 6  2017 NORTH CARO~ CENTRAL COAST~ JONES   COTTON, UPL~ ACRES HARVES~ 13079
```

Your data should now be tidy! On to visualization and summarizing.

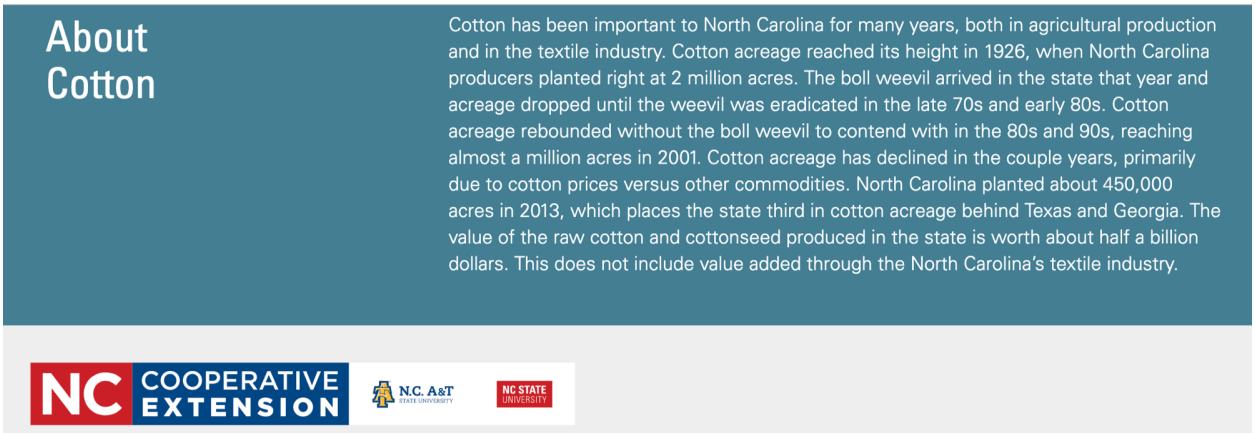
4. Visualizing trends

Next, we'll visualize cotton yield and area harvested trends across NC agricultural districts to answer the first question of this exercise: How have yield and area harvested of cotton changed across all of the NC agricultural districts over time?

Write code to prepare the following visualization. The theme is `theme_minimal()`. Note that you will have to specify the `scales =` argument in your facet function. To rotate the axis labels, use add `theme(axis.text.x = element_text(angle = 90))` to your ggplot code. You will likely need to expand your `Plots` pane in order to see all of the panel/axis labels clearly.



We can see that yields have improved over time thanks to advances in agricultural technology. Wondering what's up with the trends in acres harvested? NC Cooperative Extension explains:



5. Summarize data from 2018

Next, we'll answer the second question of this exercise: What were the top 3 cotton producing counties in NC in terms of total **lbs** of cotton for 2018? To calculate total lbs, you will need to multiply the yield (lb/acre) by the acres harvested. Yield and acres harvested are currently in one column, and will need to be `spread` across two columns in order to facilitate the calculation of total cotton lbs produced in a county. Your final output should contain only the top 3 counties in terms of total cotton lbs produced (hint: use `top_n()`). Below, the counties and total lbs are included so you can check your code.

```
## # A tibble: 3 x 2
##   county      total_lbs
##   <chr>           <dbl>
## 1 HALIFAX      40860800
## 2 MARTIN       25090800
## 3 NORTHAMPTON  36304400
```

Commit and push your script to your GitHub account, and submit a link to your Module 6 GitHub repository on Moodle.