

P3: Enrich your Dataset

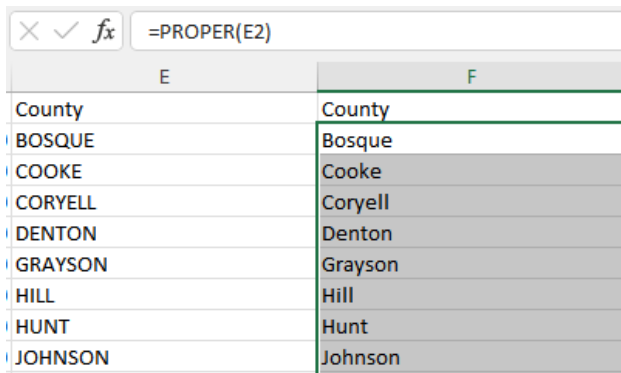
Ram Pangaluri, Fabian Leon, Ashok Shanker

February 25, 2022

Merging datasets: Year and County

To start the data enrichment phase of the project, we joined our sources according to year and county columns. While the two datasets are from distinct sources and for separate uses, it may prove useful to have to load only one csv file into the R environment.

USDA dataset's county names were in all-caps so they needed to be renamed to change that. This was easily carried out using the Find and Replace feature in Excel.



The screenshot shows an Excel formula bar with the formula `=PROPER(E2)`. Below it, a table with two columns, E and F, is visible. Column E contains county names in all caps, and column F contains the same names converted to title case.

E	F
County	County
BOSQUE	Bosque
COOKE	Cooke
CORYELL	Coryell
DENTON	Denton
GRAYSON	Grayson
HILL	Hill
HUNT	Hunt
JOHNSON	Johnson

Once this was changed, we could bind the two datasets that shared similarly coded counties.

	A	B	C	D	E
1	Dataset	Year	Location	County	AgriLife-Region
30326	TXAR	2021	Sunray	Moore	Northern High Plains
30327	TXAR	2021	Sunray	Moore	Northern High Plains
30328	TXAR	2021	Sunray	Moore	Northern High Plains
30329	TXAR	2021	Sunray	Moore	Northern High Plains
30330	TXAR	2021	Sunray	Moore	Northern High Plains
30331	TXAR	2021	Sunray	Moore	Northern High Plains
30332	USDA	2020	NA	Bosque	TBD
30333	USDA	2020	NA	Cooke	TBD
30334	USDA	2020	NA	Coryell	TBD
30335	USDA	2020	NA	Denton	TBD
30336	USDA	2020	NA	Grayson	TBD
30337	USDA	2020	NA	Hill	TBD

TX AgriLife Regions vs USDA Districts

The two datasets each included a Region column that were slightly different. We selected the Texas AgriLife regional distinctions as our chosen delineation for Texas Counties. The USDA region column was removed.

Adding geographic detail such as latitude and longitude

Latitude and longitude data were sourced using the `geocode()` function in the ‘tidygeocoder’ package and the Nominatim (“osm”) geocoding service.

```
## # A tibble: 6 x 3
##   County      latitude longitude
##   <chr>      <dbl>    <dbl>
## 1 Potter      35.4    -102.
## 2 Fannin      33.5    -96.1
## 3 Hill        32.0    -97.1
## 4 Lubbock     33.6    -102.
## 5 Bell        31.0    -97.4
## 6 San Patricio 28.0    -97.5
```

Irrigation: Coding and Variable Type

Lastly, we converted irrigation quantity data column observations from “none” to “0” and made the column an entirely numerical variable.

W	X	Y	Z	AA
Irrigation-Amount	Total-Moisture	Date-Harvested	Number-of-Rows	Population(pl
none				
none				
none				
none				
none				
none				
none				
none				
none				
none				
none				
4	16.76	3-Aug	2	85250
none	11.59	6-Aug	2	52000
none	15	23-Jul	2	52869
none	8.94	27-Jul	2	45000
2	3	3-Sep	2	32375