Charlie Roth
Understanding the Ecology of White-Tailed Deer Populations
DATA 205
CRN 33334

Introduction:
To complete this project I used data from the National Park Service Heartland Inventory and Monitoring Network and the U.S. Census Bureau. The first dataset came from data.gov and the second came from census.gov. The National Park Service dataset includes observations of White-Tailed deer in Arkansas Post National Memorial, Arkansas, Pea Ridge National Military Park, Arkansas, and Wilson's Creek National Battlefield, Missouri. The dataset includes details about the deer sighting event, such as weather descriptions, number of deer, year, longitude and latitude, as well as many other elements. The Census data has population estimates for the counties surrounding the three national parks. Arkansas County for Arkansas Post National Memorial, Benton County for Pea Ridge National Military Park, and Christian County and Greene County for Wilson's Creek National Battlefield. The population estimates are from American Community Survey (ACS) 5-year data for the years 2006-2010, 2011-2015, and 2016-2020. Arkansas County has too small of a population to be included in the ACS 1-year data, so I decided to use the 5-year data although it is less specific. Going into this project, my main goal was to find out what factors affect deer population specifically in the three national parks. Understanding what causes the fluctuation in deer populations is really important because deer populations affect a large number of things such as the ecosystem, agriculture, and people. To achieve this goal, I used several tools such as RStudio libraries (tidyverse, viridis, paletteer, and GGally), Tableau, and Google Colab (Python pandas).

Summary of Data Cleaning:
To start my data cleaning for the deer data, I checked for NAs and removed any. I renamed some columns for easier analysis in the future. For my Census data, I had to combine the columns for Christian County and Greene County so I could compare the counties to the parks. From the clean deer dataset, I created another dataset that would match with the cleaned Census dataset. I organized the data by years (2006-2010, 2011-2015, and 2016-2020) then summed the number of deer for each park.

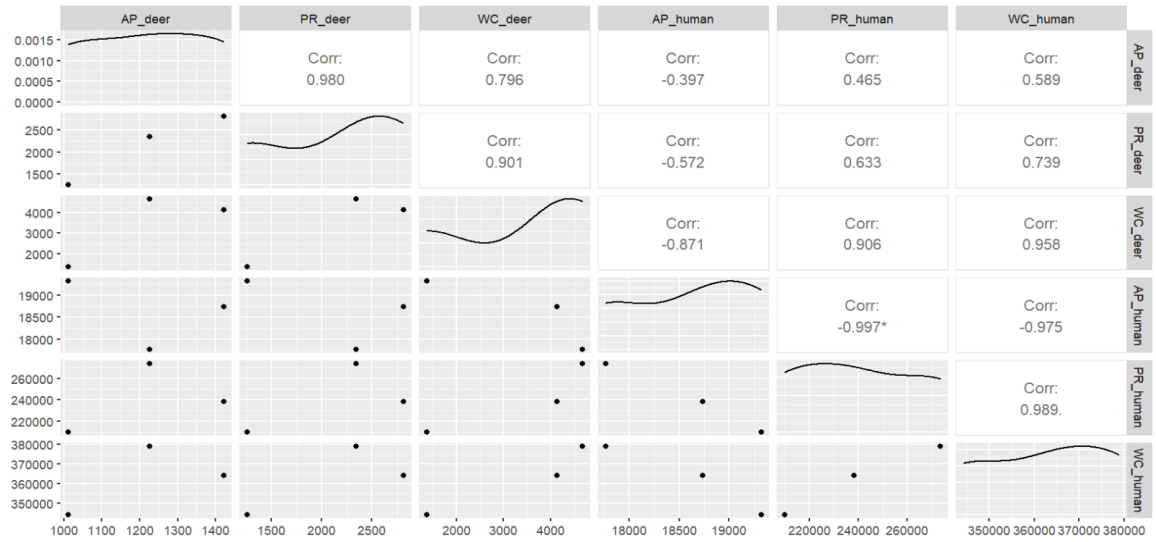
	county	total_pop	years
1	Arkansas County, Arkansas	19321	2006-2010
2	Benton County, Arkansas	209911	2006-2010
3	Christian/Greene Counties, Missouri	344346	2006-2010
4	Arkansas County, Arkansas	18731	2011-2015
5	Benton County, Arkansas	238198	2011-2015
6	Christian/Greene Counties, Missouri	364110	2011-2015
7	Arkansas County, Arkansas	17761	2016-2020
8	Benton County, Arkansas	273510	2016-2020
9	Christian/Greene Counties, Missouri	378898	2016-2020

	ParkName	DeerNum	Years
1	Arkansas Post National Memorial	1012	2006-2010
2	Pea Ridge National Military Park	1263	2006-2010
3	Wilson's Creek National Battlefield	1368	2006-2010
4	Arkansas Post National Memorial	1423	2011-2015
5	Pea Ridge National Military Park	2813	2011-2015
6	Wilson's Creek National Battlefield	4138	2011-2015
7	Arkansas Post National Memorial	1226	2016-2020
8	Pea Ridge National Military Park	2346	2016-2020
9	Wilson's Creek National Battlefield	4667	2016-2020

(Census data, left; Deer Population, right)

Basic Descriptive Statistics:
In my deer dataset, I use the columns DeerNum, BeginTempCelsius, BeginPrecip, VegType, and Month. DeerNum is the number of deer spotted at a single event or sighting. It ranges from 1 deer spotted to 54 deer spotted in an event. BeginTempCelsius is the beginning temperature in celsius for the event. The temperature data has a normal distribution and it is unimodal and symmetrical. BeginPrecip is the beginning precipitation at the event. It is a character variable that describes what the precipitation was like such as “no rain”, “light rain”, “snow”, etc. There is no pattern in the data to suggest a significant relationship with the number of deer. VegType is the vegetation type that the deer was spotted in. Some examples are “Savanna”, “Upland scrub”, and “Riparian forest”. Grassland/Prairie has the highest number of deer, Upland forest following. The Month variable only has 3 months, January, February, and March. The majority of deer are spotted in January, March having the least number of deer. In the Census data, the counties surrounding Wilson’s Creek and Pea Ridge both increased over the years but the county surrounding Arkansas Post decreased over the years. For the deer population data, the number of deer fluctuated over the years.

Description of Final Data Product:
I analyzed the human population of the counties surrounding Arkansas Post National Memorial, Pea Ridge National Military Park, and Wilson’s Creek National Battlefield with the deer population inside the parks. I started with a pairwise plot to see any correlations between the deer and human populations. *See below:*



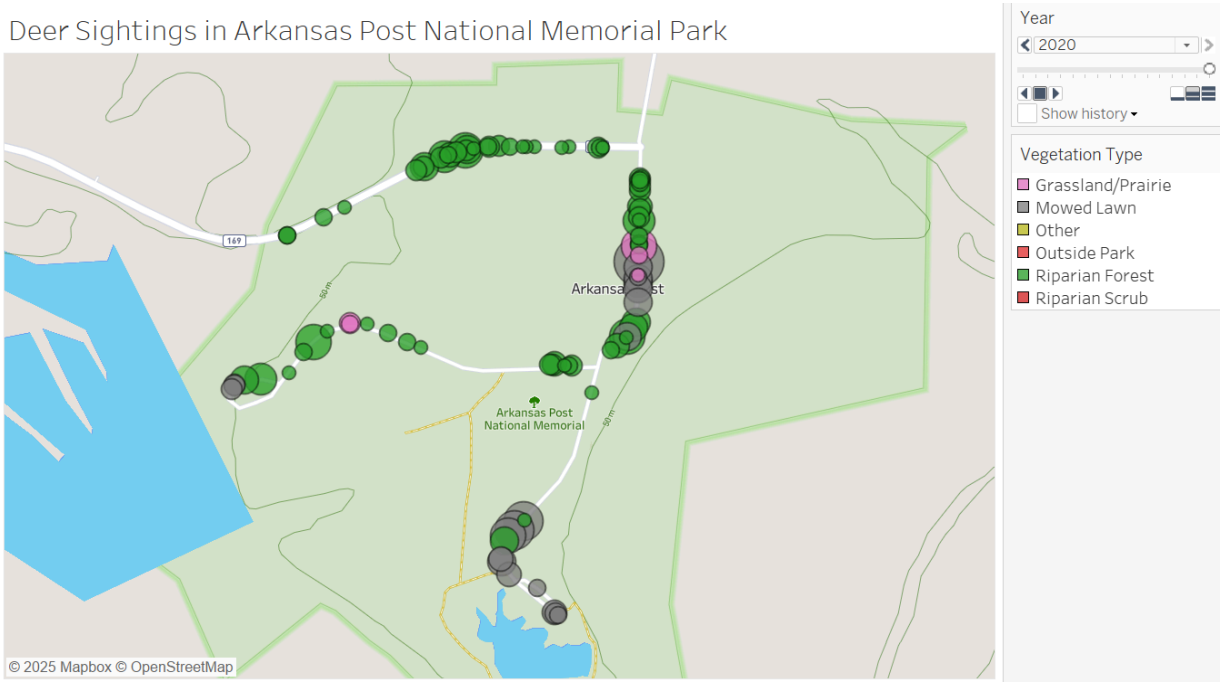
This plot shows that there is a strong correlation between the human populations which make sense given geographic, demographic, social, and economic factors. Arkansas Post's deer population and human population surrounding it are not very correlated at -0.397. Pea Ridge's populations are slightly correlated at 0.633 and Wilson's Creek populations are very correlated at 0.958. To look into the relationships of human and deer populations and see if there is a reason behind the difference in correlations, I

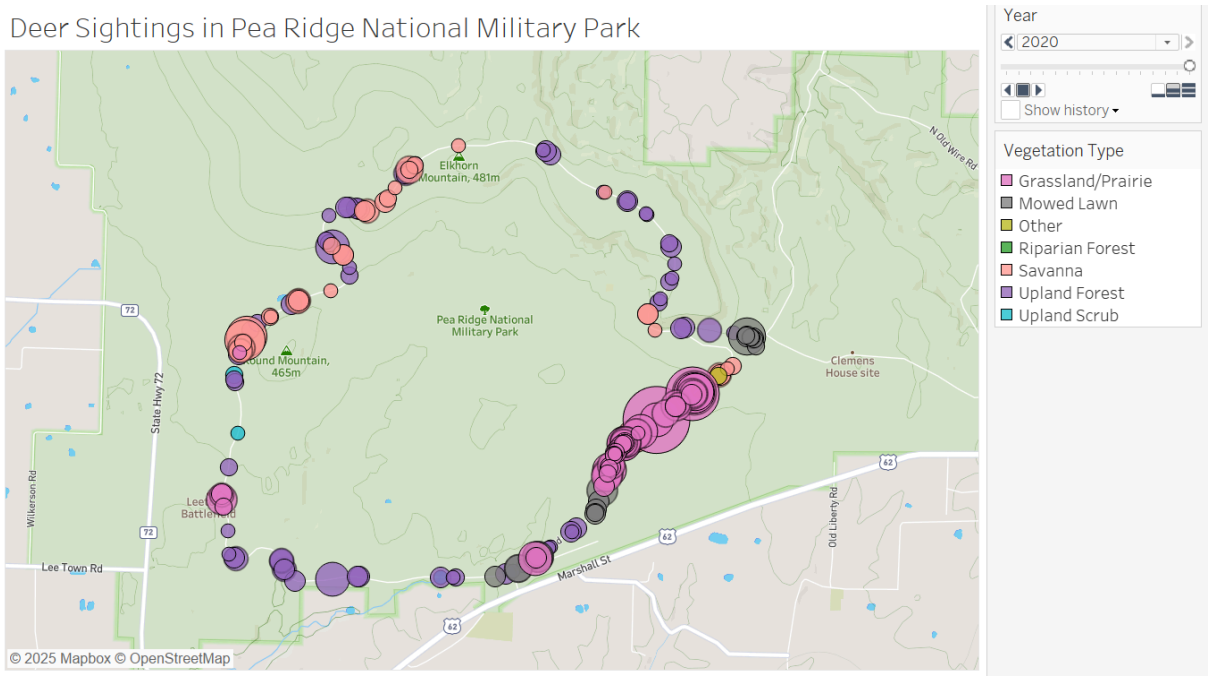
decided to do linear regression models for each of the three parks. The linear regression models for Arkansas Post and Pea Ridge both had negative R-squared and a high p-value. This is most likely due to a small sample size. Wilson's Creek model on the other hand, had a high R-squared and a non-significant p-value. Based on the results from the models, I believe that the sample size is too small to make any conclusions on if deer population size is affected by human population size.

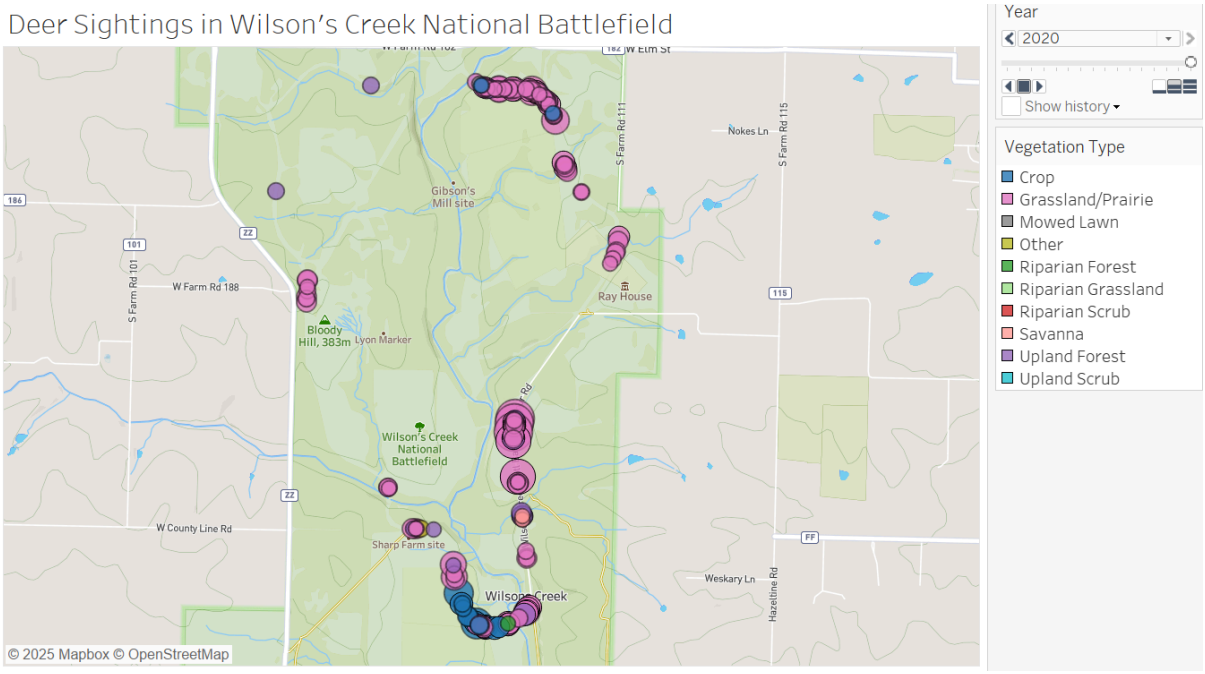
Next, I analyzed the number of deer with vegetation type. I started by creating a contingency table for vegetation type in each park. The frequency is the number of deer sighted in that vegetation for each park.

	ParkName		
VegType	AP_Park	PR_Park	WC_Park
Crop	0	0	2558
Grassland/Prairie	43	2530	6306
Mowed Lawn	1775	272	34
Other	15	19	14
Outside Park	17	0	0
Riparian Forest	2022	2	480
Riparian Grassland	0	0	334
Riparian Scrub	29	0	21
Savanna	0	114	6
Upland Forest	0	3876	625
Upland Scrub	0	69	101

The table shows that in Wilson's Creek, the park with the most deer, the majority deer sighted are in grassland/prairie vegetation. For Pea Ridge, the majority of the deer were sighted in upland scrub or grassland/prairie. For Arkansas Post however, the park with the least amount of deer, the majority of deer were sighted in riparian forest. I visualized this with three interactive tableau plots for each park that you can see below.







I used the contingency table to perform a Chi-squared test. The test is to see if there is a significant relationship between vegetation and number of deer.

```
Warning: Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  deer_tbl
X-squared = 24661, df = 20, p-value < 2.2e-16
```

The Chi-squared test shows a very small significant p-value. It also shows a warning that the test approximation may be incorrect. This is because some of the values in the contingency table are low or zero. To fix this, I performed a Chi-squared test with simulated values as shown below.

```
        Pearson's Chi-squared test with simulated p-value (based on
        10000 replicates)

data:  deer_tbl
X-squared = 24661, df = NA, p-value = 9.999e-05
```

From the second test, you can see that the p-value is higher but still below the significance level. Based on these two tests, it's safe to say that there is a significant relationship between vegetation and number of deer.

References and Acknowledgements:

The datasets used in this project came from data.gov and census.gov. The data was collected by the National Park Service Heartland Inventory and Monitoring Network and the U.S. Census Bureau. I would like to thank my data teachers over the semesters, Professor Saidi, Professor Alraee, and Professor Perine, who are the reason I was able to complete this project with the knowledge I learned in their classes.