AgriCorp Exercise

Introduction

I choose to analyze AgriCorp's data using R and present my work using Quarto. This report is structured as a narrative walking a reader through my process for preparing and transforming the AgriCorp data, key findings, and recommendations for AgriCorp's leadership.

Preparing, Cleaning, and Transforming the Data

First, I'll load the packages needed for analysis, which primarily belong to a collection known as the Tidyverse, and import the AgriCorp dataset into the local environment as a dataframe (df). I'm making sure R recognizes the column types correctly, and write a little bit of code so that the "month" column is interpreted as a date, rather than a string.

```
library(tidyverse)
library(kableExtra)
library(widyr)
library(janitor)
library(broom)

df <- read_csv("data/data.csv", col_types = cols(season = col_double(),
                                                 month = col_date(format = "%Y-%m-%d")))
```

Next, I'll clean the "zone" and "village" columns by removing the redundant characters from the values.

```
df <- df %>%
  mutate(across(c('zone'), substr, 2, nchar(zone))) %>%
  mutate(across(c('village'), substr, 2, nchar(village)))
```

Next, I want to see how many villages and zones this dataset features. It looks like 283 villages and 3 zones.

```
df %>%
  summarise(n = n_distinct(zone))
```

n
3

```
df %>%
  summarise(n = n_distinct(village))
```

n
283

I'll also do a quick check for NA values across the entire dataset using the map function, which is part of my routine for cleaning a dataset. Here, we don't have any NA values; if we did, I'd want to figure out why and determine whether they were the result of a data quality issue.

```
map_df(df, ~sum(is.na(.)))
```

Because I have some experience in development agriculture, I have a suspicion that AgriCorp may have expanded its geographic area during 2018-2023. So I did a quick count of the number of zones AgriCorp worked in in each year and it does look like they worked in only 2 zones beginning in 2018 and expanded to a third zone in 2022. Though fairly basic, this information may influence the way I compare zones and villages over time.

```
df %>%
  group_by(season) %>%
  summarise(n = n_distinct(zone))
```

season	n
2018	2
2019	2
2020	2
2021	2
2022	3
2023	3

Key Findings about Women Farmers

Next, I'm curious about the ratio of enrolled women farmers to enrolled non-women farmers, and whether that ratio has changed over time. I'd also like a better idea of how many farmers AgriCorp enrolled each year, and what percent of those farmers finished paying for their packages. So, I create a simple table showing the total number of enrolled, women, and finished farmers for each year, as well as the percent of the total number of enrolled women farmers, and the percent of the total number of enrolled farmers who finished paying for packages.

```
df %>%
  group_by(season) %>%
  summarise(total_farmers = sum(enrolled_clients),
            women_farmers = sum(enrolled_female_clients),
            finished_farmers = sum(finished_clients)
            ) %>%
  mutate(percent_women = women_farmers/total_farmers) %>%
  mutate(percent_finished = finished_farmers/total_farmers)
```

season	total_farmers	women_farmers	finished_farmers	percent_women	percent_finished
2018	7255	1884	2689	0.2596830	0.3706409
2019	6075	1596	3423	0.2627160	0.5634568
2020	5706	1543	2924	0.2704171	0.5124430
2021	6025	1601	4741	0.2657261	0.7868880
2022	8128	2513	4967	0.3091781	0.6110974
2023	11775	6215	5155	0.5278132	0.4377919

It seems like the number of enrolled women farmers has increased since dipping to a low of 1,543 in 2020, with a large increase happening in 2023. It also looks like the ratio of women farmers to non-women farmers was relatively unchanged until 2023, when it increased from 31% to 53%.

I'm really curious about the geographic breakdown of that increase in 2023, so I create another table for the 2022-2023 seasons, and segment the data by zone.

```
df %>%
  filter(season >= 2022) %>%
  group_by(season, zone) %>%
  summarise(total_farmers = sum(enrolled_clients),
            women_farmers = sum(enrolled_female_clients),
            finished_farmers = sum(finished_clients)
            ) %>%
  mutate(percent_women = women_farmers/total_farmers) %>%
  mutate(percent_finished = finished_farmers/total_farmers)
```

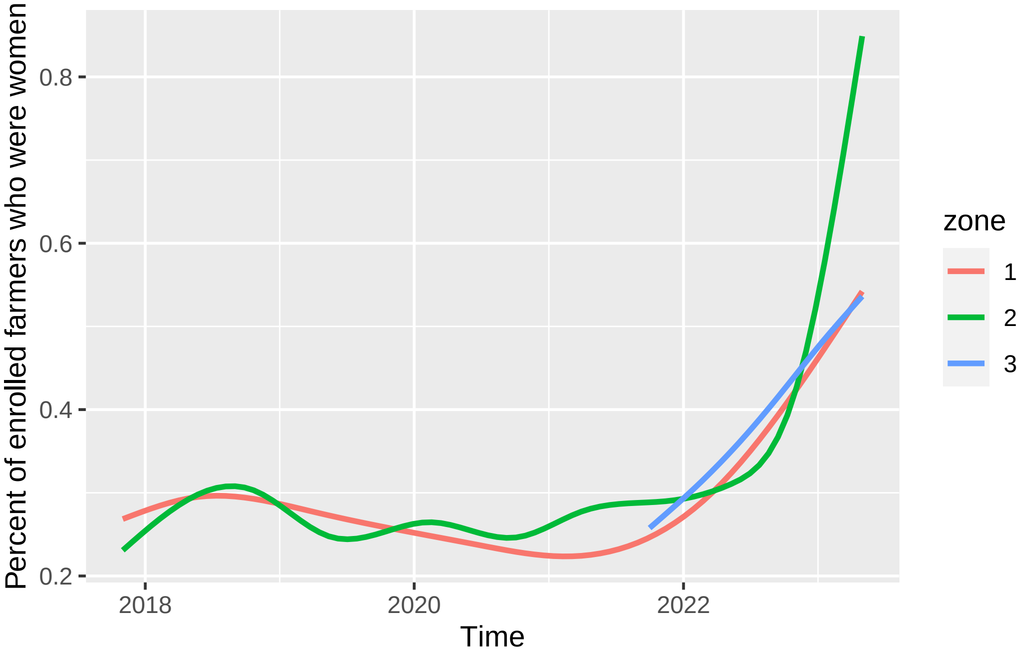
season	zone	total_farmers	women_farmers	finished_farmers	percent_women	percent_finished
2022	1	1924	547	969	0.2843035	0.5036383
2022	2	3269	1056	1905	0.3230346	0.5827470
2022	3	2935	910	2093	0.3100511	0.7131175
2023	1	3429	1609	1594	0.4692330	0.4648586
2023	2	4515	2736	2070	0.6059801	0.4584718
2023	3	3831	1870	1491	0.4881232	0.3891934

I learn that the number of women farmers enrolled in Zone 2 increased from 1,056 in 2022 to 2,736 in 2023, and that the percent of the total enrolled farmers who were women nearly doubled (from 32% to 60%) in the zone. This is an interesting finding. Because AgriCorp is interested in creating products to better serve female farmers, it may be worth finding out the reason for this increase.

In order to learn more about the trend of women farmer enrollment, I create a new column in my working dataframe dividing the number of women enrolled by the total number of farmers enrolled at a given point in time, and group those results by zone. Now, I'm able to visualize the women-enrolled ratio and learn more about it at a more granular level.

```
df <- df %>%
  mutate(percent_women = (enrolled_female_clients/enrolled_clients))

df %>%
  mutate(season = as.character(season)) %>%
  ggplot(aes(x=month, y = percent_women, color=zone)) +
  geom_smooth(se=F) +
  ylab("Percent of enrolled farmers who were women") +
  xlab("Time") +
  scale_fill_discrete(name = "Zone")
```



I create a set of trend lines that confirm what we learned in our earlier tables: it seems like AgriCorp increased the ratio of women farmers it enrolled during 2023 after a period of consistency. We can also see the large increase in the ratio of women farmers enrolled in Zone 2 in 2023.

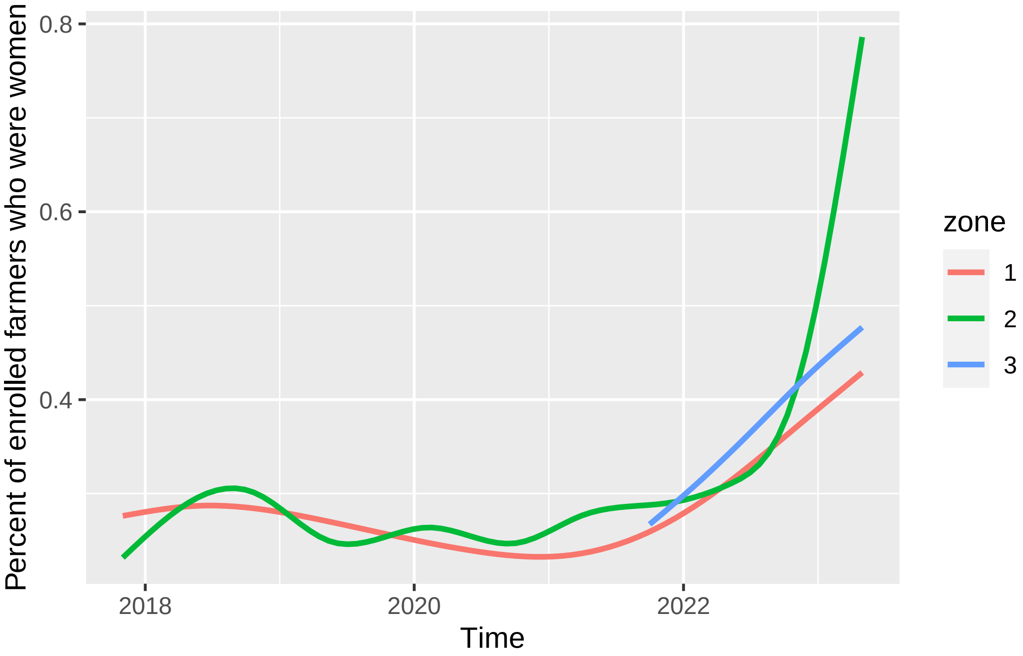
I already know AgriCorp has historically expanded its geographic area each year, and wonder whether the increase in women's enrollment in 2023 is the result of the organization simply having added a group of new villages in which large numbers of women were enrolled from the program outset. Though useful information for the purpose of planning future geographic expansion, that conclusion would be less useful for the purpose of changing AgriCorp's products and services in existing villages to meet the needs of more women farmers and ultimately drive up women farmer enrollment.

To check whether the increase in 2023 was the result of new villages being added, I make a new column in the dataframe called first_season, which tells me the first season a given village worked with AgriCorp.

```
df <- df %>%
  group_by(village)%>%
  mutate(first_season = min(season))%>%
  ungroup()
```

Now, I recreate the above visualization but filter out any villages that first worked with AgriCorp in 2023.

```
df %>%
  mutate(season = as.character(season)) %>%
  filter(first_season != 2023) %>%
  ggplot(aes(x=month, y = percent_women, color=zone)) +
  geom_smooth(se=F) +
  ylab("Percent of enrolled farmers who were women") +
  xlab("Time") +
  scale_fill_discrete(name = "Zone")
```

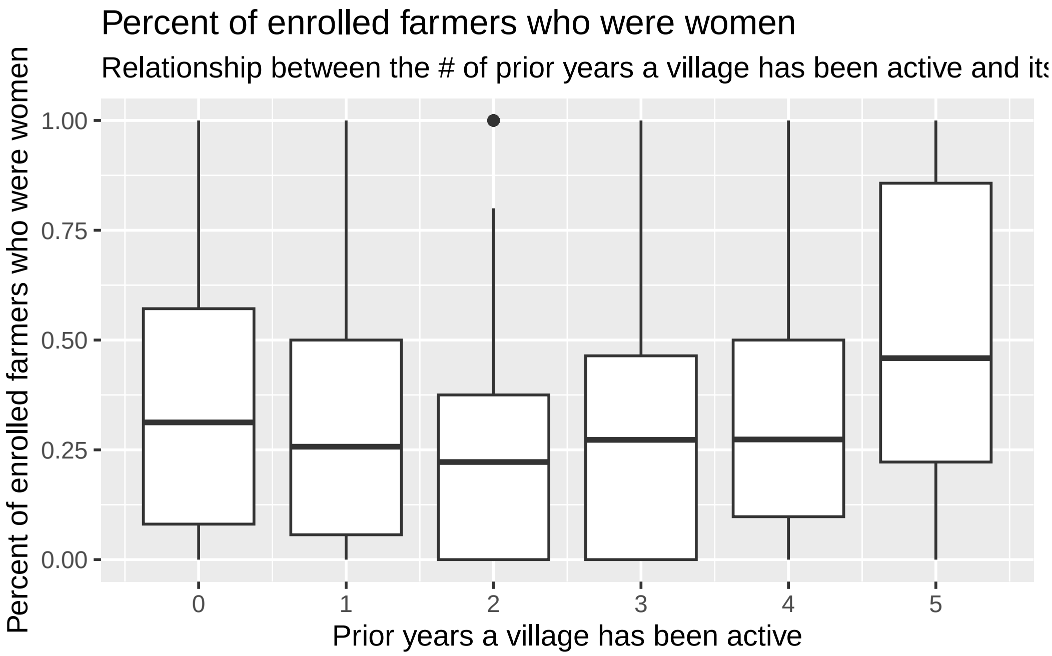


The two graphs are fairly similar, and it seems like the increased ratio of enrolled women farmers in Zone 2 wasn't caused by AgriCorp adding new villages. It seems instead to be the result of a change in the existing group of villages AgriCorp worked with prior to 2023. This is great news for AgriCorp; the organization may be able use whatever it learns about that change to improve its products and services across other zones.

Now that I'm thinking about it, I want to take a quick look at the relationship between the number of seasons a village has been active and the percentage of women who enrolled in that village. I create a new column in the dataframe (seasons_active) and make a quick boxplot.

```
df <- df %>%
  mutate(seasons_active = season-first_season)
```

```
df %>%
  ggplot(aes(x=seasons_active, y=percent_women, group=seasons_active)) +
    geom_boxplot() +
    labs(title="Percent of enrolled farmers who were women", subtitle="Relationship betwee
    ylab("Percent of enrolled farmers who were women") +
    xlab("Prior years a village has been active") +
    scale_x_continuous(breaks = round(seq(min(df$seasons_active), max(df$seasons_active),
```



It seems like there is a relationship between the number of seasons a village has been active and the ratio of women farmers to non-women farmers who enroll there: villages that have been active since 2018 have, on average, a higher ratio of enrolled women farmers.

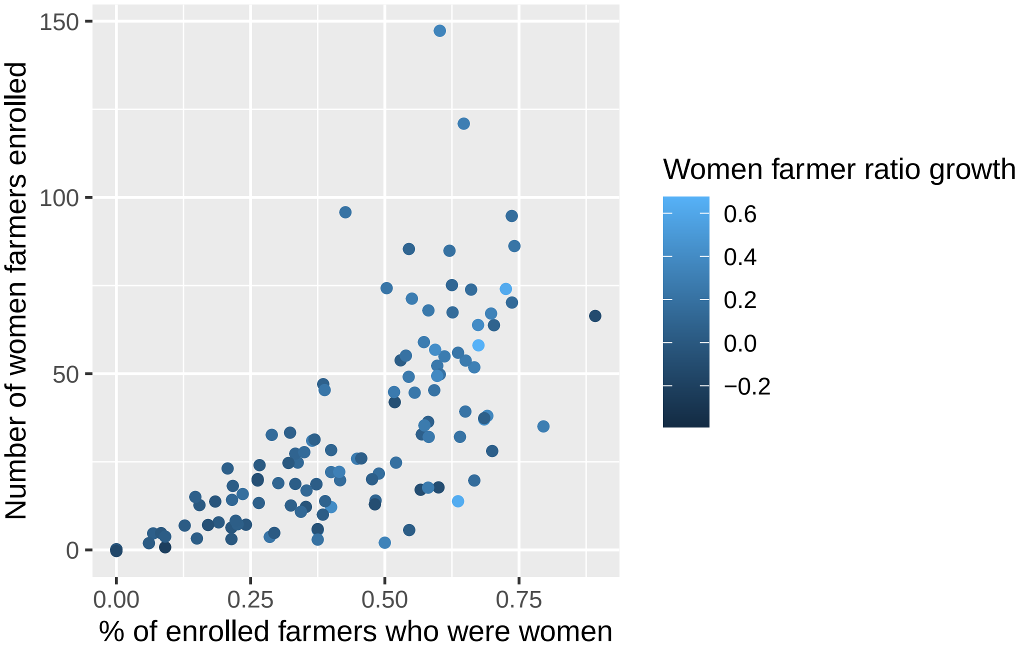
I'm interested in finding out which specific villages increased the ratio of women to non-women enrollments in 2023 over 2022 the most. To do this, I need to make some changes to my dataframe.

```
df2 <- df %>%
  group_by(village, season, zone) %>%
  summarise(total_farmers = sum(enrolled_clients),
            women_farmers = sum(enrolled_female_clients),
            finished_farmers = sum(finished_clients),
            seasons_active = mean(seasons_active)
  ) %>%
  mutate(percent_women = women_farmers/total_farmers) %>%
  mutate(percent_finished = finished_farmers/total_farmers) %>%
  arrange(village, season) %>%
  group_by(village) %>%
  mutate(year_growth = percent_women - lag(percent_women))
```

This chunk of code creates a new dataframe by summarizing the seasonal totals of enrolled_clients, enrolled_female_clients, and finished_clients. It also summarizes the seasonal percent of enrolled farmers who were women, and the seasonal percent of farmers who finished paying for packages. Finally, it creates a new column (year_growth) that calculates the annual growth rate of a given village's ratio of enrolled women farmers to enrolled non-women farmers.

The reason I make these additions to my dataframe is so I can look at all my variables of interest (total number of women farmers enrolled, percent of total farmers enrolled who were women, the number of prior seasons a village was active before 2023, and the growth of a village's ratio of women farmers enrolled) together in one visualization to see what I can learn from them.

```
df2 %>%
  filter(season == 2023) %>%
  filter(seasons_active != 0) %>%
  filter(year_growth != 0) %>%
  ggplot(aes(x=percent_women, y=women_farmers, color=year_growth))+
  geom_jitter() +
  expand_limits(size=0) +
  ylab("Number of women farmers enrolled") +
  xlab("% of enrolled farmers who were women") +
  scale_color_continuous(name = "Women farmer ratio growth")
```



This plot of data from 2023 shows a few relationships: for villages that were active before 2023, there is a relationship between the number of women farmers enrolled and the percent of enrolled farmers who were women. There also seems to be a relationship between the growth of the enrolled-women ratio, the percent of enrolled farmers who were women, and the total number of women farmers enrolled in a given village. This is all interesting and potentially useful context for AgriCorp.

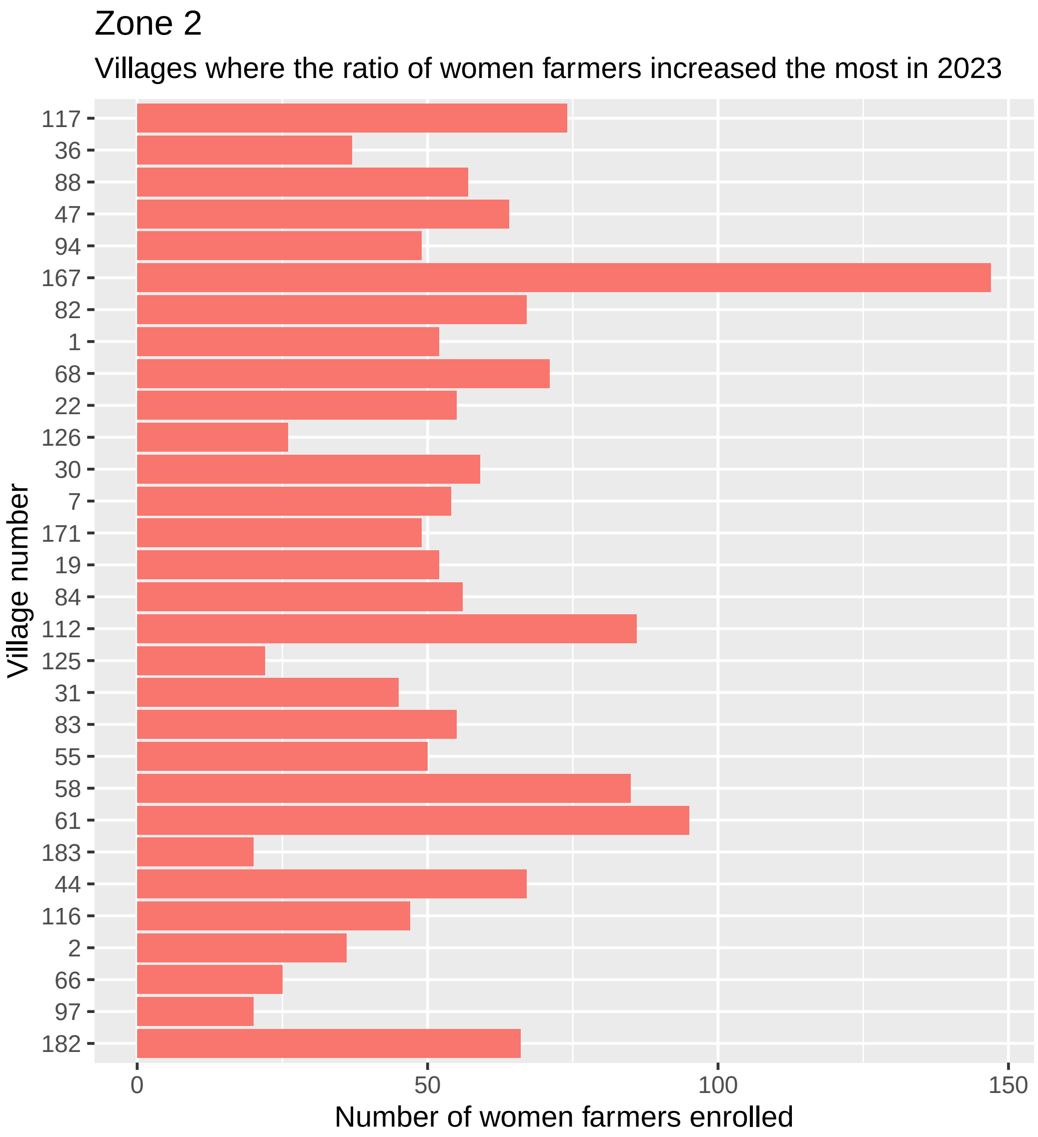
Recommendations

My recommendation to AgriCorp is to investigate whether any changes were made to their program or products in 2023 in Zone 2 which may have resulted in the increased number and ratio of women farmers who enrolled with the organization there. If found, these changes may warrant testing across other zones.

My recommendation for specific villages to investigate can be found in the following graphic. The villages are ordered top to bottom by their annual growth in the ratio of women farmers to non-women farmers enrolled in 2023, and I elected to only show villages where the percent of enrolled farmers who were women was 75% or above, and where the total number of farmers enrolled was more than 20. These villages are meant to represent the highest performing insofar as women farmer enrollment, and in which changes that encouraged increased enrollment of women are likeliest to have occurred in 2023.

Though there were some villages which met these criteria in Zones 1 and 3, the majority were in Zone 2, where I recommend AgriCorp begin its research into programmatic factors which may have caused the increased women farmer enrollment, and which may be applicable in other zones where AgriCorp is interested in increasing both the total number of women farmers they enroll as well as the enrollment ratio of women farmers to non-women farmers.

```r
df2 %>%
  filter(season == 2023) %>%
  filter(year_growth != 0) %>%
  filter(women_farmers >= 20) %>%
  filter(zone == "2") %>%
  ggplot(aes(x=reorder(village, year_growth), y=women_farmers, fill = zone)) +
  geom_col() +
  coord_flip() +
  theme(legend.position = "none") +
  xlab("Village number") +
  ylab("Number of women farmers enrolled") +
  labs(title = "Zone 2", subtitle = "Villages where the ratio of women farmers increased t
  theme(legend.position = "none")
```


Zone 2
Villages where the ratio of women farmers increased the most in 2023
Village number
117
36
88
47
94
167
82
1
68
22
126
30
7
171
19
84
112
125
31
83
55
58
61
183
44
116
2
66
97
182
0
50
100
150
Number of women farmers enrolled

Additional Context

In 2023 there were 223 active villages in 3 zones, compared with 2022's 140 active villages in 3 zones.

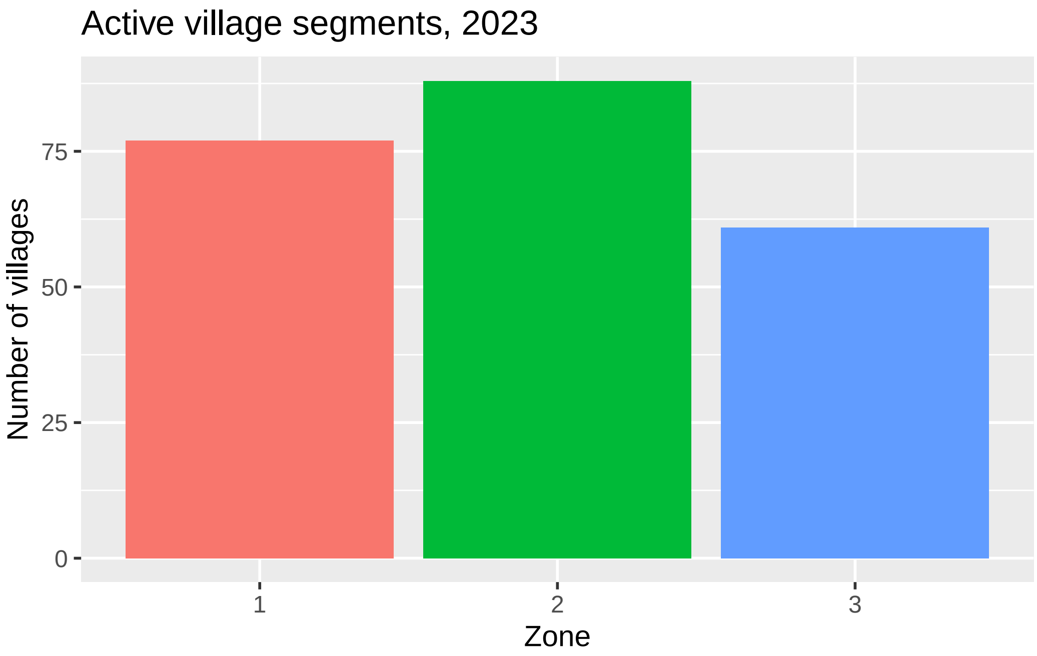
```
df %>%
  filter(season == "2022") %>%
  summarise(n = n_distinct(village))
```

n
140

```
df %>%
  filter(season == "2023") %>%
  summarise(n = n_distinct(village))
```

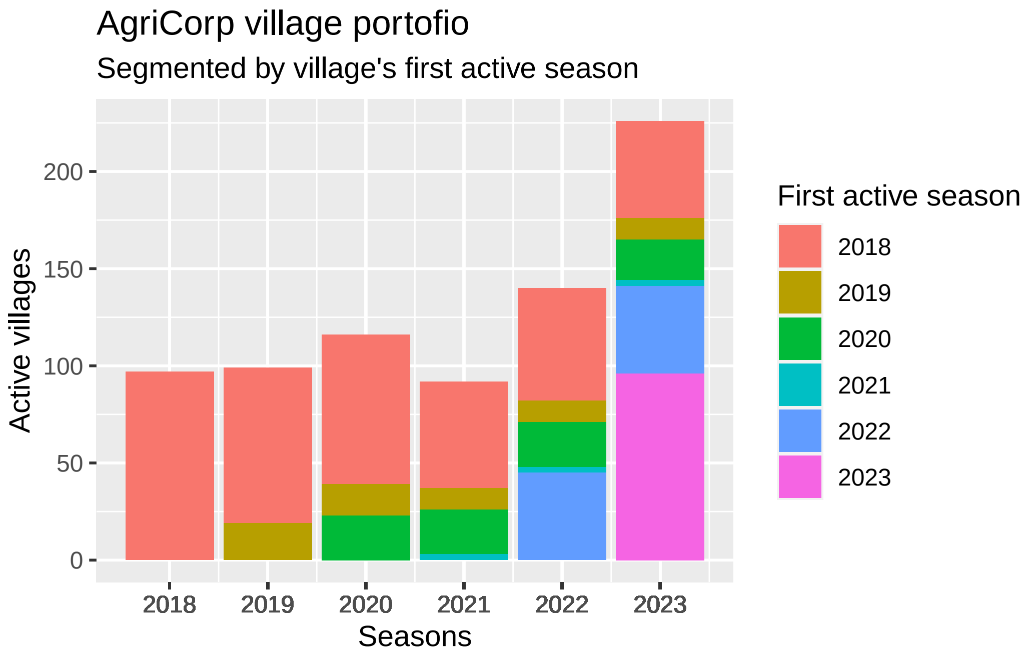
n
226

```
df2 %>%
  filter(season == "2023") %>%
  summarize(zone, n = n_distinct(village), zone) %>%
  ggplot(aes(x=zone, fill = zone)) +
  geom_bar() +
  ylab("Number of villages") +
  xlab("Zone") +
  guides(fill = FALSE) +
  labs(title = "Active village segments, 2023")
```

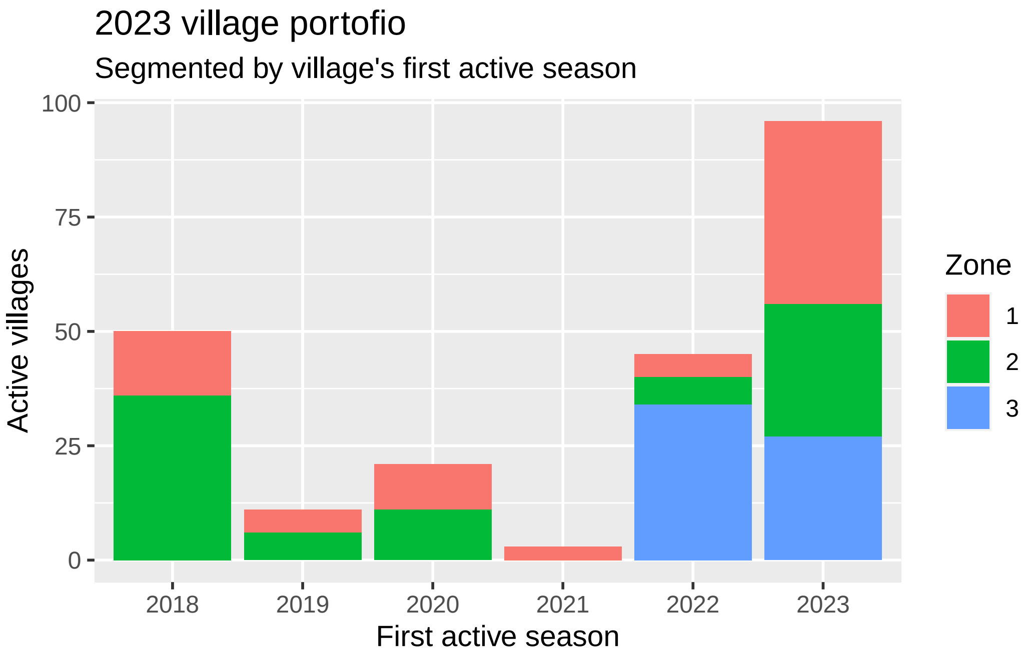


Here are a few other ways of thinking about village segments. In 2023 there seems to have been a large influx of villages in 2023 in all zones. It also seems like about half of the villages which were active in 2018 are no longer active in 2023. This offers a possible hypothesis for why villages that have been active since 2018 were so successful at growing their women-farmer ratio: at this point, only the villages in which AgriCorp is a best fit remain.

```
df2 %>%
  mutate(first_season = season - seasons_active) %>%
  mutate(first_season = as.character(first_season)) %>%
  ggplot(aes(x=season, fill = first_season)) +
    geom_bar() +
    labs(title = "AgriCorp village portofio", subtitle = "Segmented by village's first act
    ylab("Active villages") +
    xlab("Seasons") +
    scale_color_continuous(name = "First active season") +
    scale_x_continuous(labels=as.character(df2$season),breaks=df2$season) +
    scale_fill_discrete(name = "First active season")
```



```
df2 %>%
  filter(season == "2023") %>%
  mutate(first_season = season - seasons_active) %>%
  mutate(first_season = as.character(first_season)) %>%
  ggplot(aes(x=first_season, fill = zone)) +
    geom_bar() +
    labs(title = "2023 village portofio", subtitle = "Segmented by village's first active
    ylab("Active villages") +
    xlab("First active season") +
    scale_fill_discrete(name = "Zone")
```

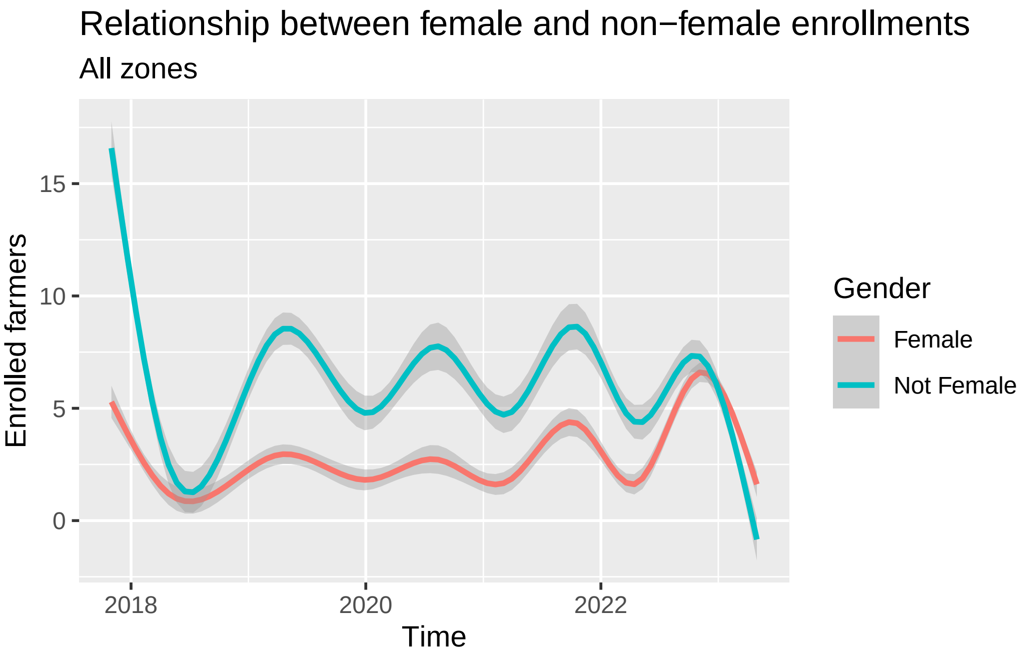


Correlation between enrolled farmers and enrolled women farmers

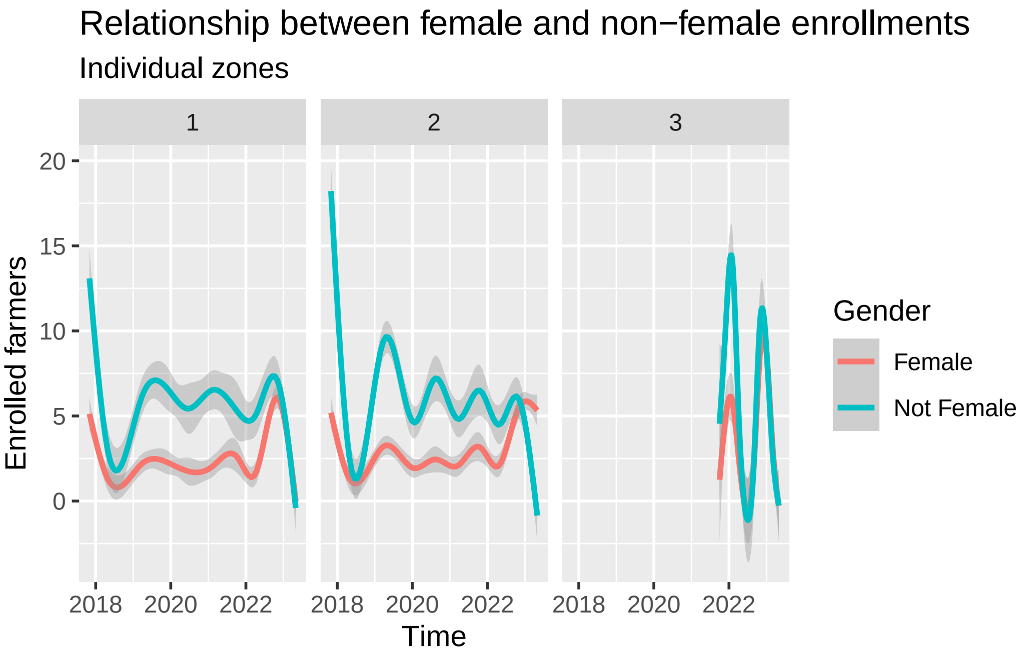
Here I tidy the data to enable a direct comparison between enrollments of farmers who were women and those who were not. I find that the two values are correlated: as one increases, so does the other. This validates my assumption that the number of women farmers on its own is not a great metric for AgriCorp improvement because it essentially tracks with the total number of farmers enrolled with the organization. The only exception seems to be in 2023 in Zone 2.

```
tidy_data <- df %>%
  mutate(enrolled_nonfemale_clients = enrolled_clients - enrolled_female_clients) %>%
  select(-enrolled_clients) %>%
  rename(female = enrolled_female_clients) %>%
  rename(not_female = enrolled_nonfemale_clients) %>%
  pivot_longer(c(5,12),names_to = "client_gender", values_to = "enrolled_clients")
```

```
tidy_data %>%
  ggplot(aes(month, enrolled_clients, color = client_gender)) +
    geom_smooth() +
    labs(title = "Relationship between female and non-female enrollments",
         subtitle = "All zones") +
    xlab("Time") +
    ylab("Enrolled farmers") +
    scale_color_discrete(name = "Gender", breaks=c("female", "not_female"),
                         labels=c("Female", "Not Female"))
```

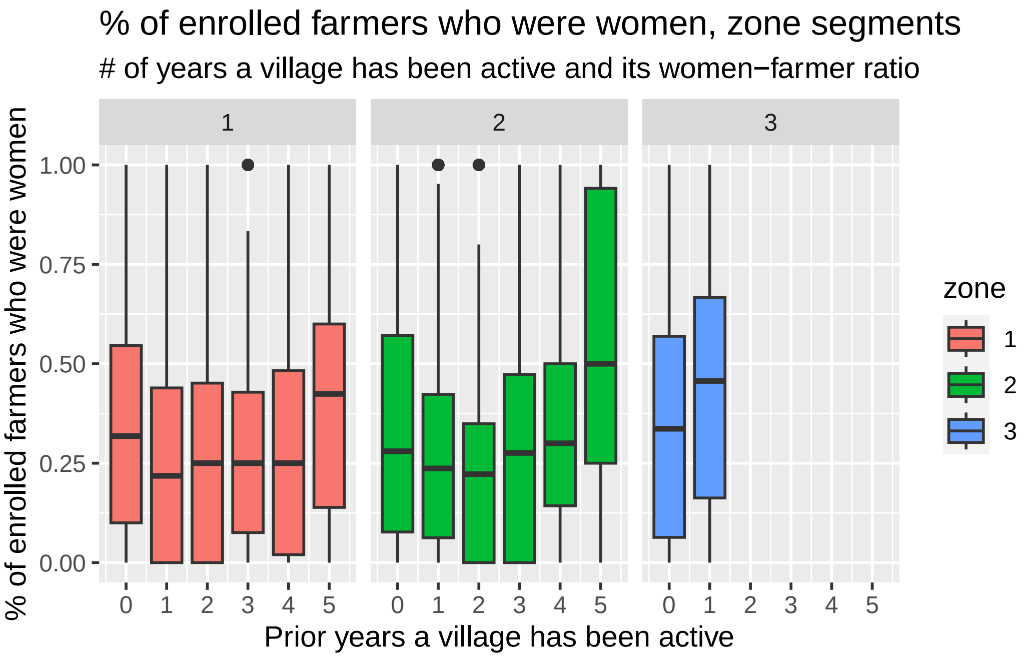


```
tidy_data %>%
  ggplot(aes(month, enrolled_clients, color = client_gender)) +
    geom_smooth() +
    facet_wrap(~ zone) +
    labs(title = "Relationship between female and non-female enrollments",
         subtitle = "Individual zones")+
    xlab("Time") +
    ylab("Enrolled farmers") +
    scale_color_discrete(name = "Gender", breaks=c("female", "not_female"),
                         labels=c("Female", "Not Female"))
```



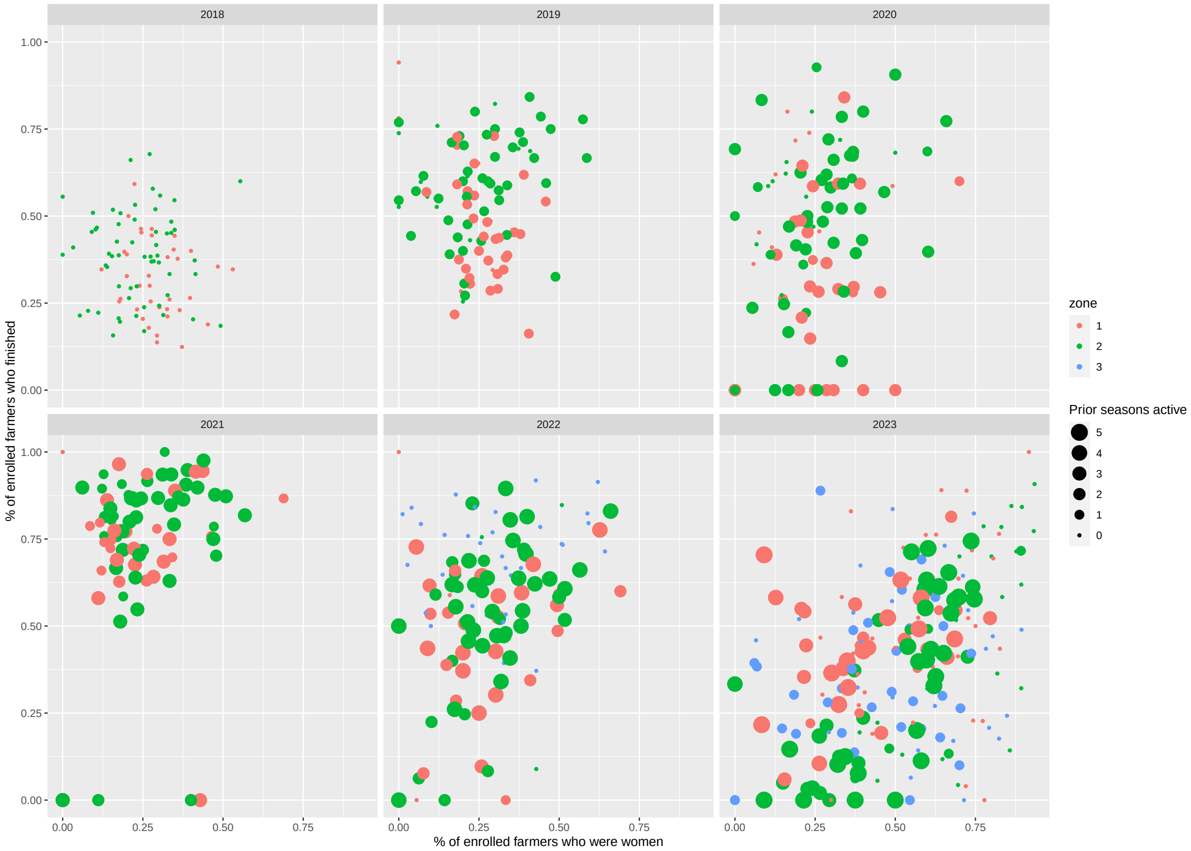
Here I wanted to look more closely at my assumption that the most 'mature' villages also were the most successful insofar as the ratio of women to non-women farmers. It seems like that is still the case when the data is segmented by zone, and that the trend is even more pronounced in Zone 2. This does not necessarily suggest a causal relationship between village maturity and women/non-women ratio.

```
df %>%
  ggplot(aes(x=seasons_active, y=percent_women, group=seasons_active, fill = zone)) +
    geom_boxplot() +
    facet_wrap(~ zone) +
    labs(title="% of enrolled farmers who were women, zone segments",
         subtitle="# of years a village has been active and its women-farmer ratio") +
  ylab("% of enrolled farmers who were women") +
  xlab("Prior years a village has been active") +
  scale_x_continuous(breaks = round(seq(min(df$seasons_active),
                                        max(df$seasons_active), by = 1),1))
```



Here I'm playing around with the parameters of the scatterplot on page 9 to see if there's anything more interesting to learn than that there's a relationship between the number of women farmers enrolled and the percent of enrolled farmers who were women. What I find is that the relationship between the women-farmer ratio and ratio of finished to enrolled farmers becomes much stronger in 2023. This visualization also shows the growth in the women-farmer ratio over time, and a bit of the relationship between the maturity of a village and its women-farmer ratio.

```
df2 %>%
  filter(percent_women < 1) %>%
  ggplot(aes(y=percent_finished, x=percent_women, color=zone)) +
  geom_jitter(aes(size = seasons_active)) +
  facet_wrap(~season) +
  expand_limits(size=0) +
  ylab("% of enrolled farmers who finished") +
  xlab("% of enrolled farmers who were women") +
  scale_size_continuous(name = "Prior seasons active",
                        guide = guide_legend(reverse=TRUE))
```



Finally, I figure out a way to quickly find which variables in the dataset are most and least correlated with one another. I use pairwise_cor from the widyr package to do this.

To read this table, column 1 represents the first variable, column 2 represents the second variable, and column 3 represents the correlation of the first variable to the second, where a value closer to 1 represents a stronger correlation. In our case, because there are an equal number of variables being compared, both correlation figures will be the same.

I also write a function that makes it easy to segment the data by season, so that in the future AgriCorp can quickly look for high-level trends in their annual data.

```
df3 <- df2 %>%
  mutate_all(~ifelse(is.nan(.), 0, .)) %>%
  mutate(nonfemale_farmers = total_farmers - women_farmers) %>%
  select(-seasons_active) %>%
  pivot_longer(c(4:8, 10), names_to = "metric", values_to = "value")

df3 %>%
  pairwise_cor(metric, village, value, sort = TRUE)
```

item1	item2	correlation
nonfemale_farmers	total_farmers	0.9101250
total_farmers	nonfemale_farmers	0.9101250
finished_farmers	total_farmers	0.8501958
total_farmers	finished_farmers	0.8501958
nonfemale_farmers	finished_farmers	0.7181571
finished_farmers	nonfemale_farmers	0.7181571
finished_farmers	women_farmers	0.6288578
women_farmers	finished_farmers	0.6288578
women_farmers	total_farmers	0.6151636
total_farmers	women_farmers	0.6151636
percent_finished	finished_farmers	0.5452992
finished_farmers	percent_finished	0.5452992
percent_women	women_farmers	0.4397989
women_farmers	percent_women	0.4397989
percent_finished	women_farmers	0.2495135
women_farmers	percent_finished	0.2495135
nonfemale_farmers	women_farmers	0.2332152
women_farmers	nonfemale_farmers	0.2332152
percent_finished	total_farmers	0.1803833
total_farmers	percent_finished	0.1803833
nonfemale_farmers	percent_finished	0.0913590
percent_finished	nonfemale_farmers	0.0913590
percent_finished	percent_women	0.0416579
percent_women	percent_finished	0.0416579
percent_women	finished_farmers	-0.1250103
finished_farmers	percent_women	-0.1250103
percent_women	total_farmers	-0.2578332
total_farmers	percent_women	-0.2578332
nonfemale_farmers	percent_women	-0.5491468
percent_women	nonfemale_farmers	-0.5491468

```
correlations <- function(years) {

  correlations <- tibble()

  for (year in years) {
    df <- df3 %>% filter(season == year)
    correlations_year <- df %>% pairwise_cor(metric, village, value, sort = TRUE)

    correlations_year$year <- year

    correlations <- correlations %>% bind_rows(correlations_year)
  }
  return(correlations)
}

annual_correlations <- correlations(c("2022","2023")) %>%
  relocate(year) %>%
  rename(season = year)

head(annual_correlations, 10)
```

season	item1	item2	correlation
2022	nonfemale_farmers	total_farmers	0.9461402
2022	total_farmers	nonfemale_farmers	0.9461402
2022	finished_farmers	total_farmers	0.9401124
2022	total_farmers	finished_farmers	0.9401124
2022	nonfemale_farmers	finished_farmers	0.8627983
2022	finished_farmers	nonfemale_farmers	0.8627983
2022	women_farmers	total_farmers	0.8219811
2022	total_farmers	women_farmers	0.8219811
2022	finished_farmers	women_farmers	0.8196866
2022	women_farmers	finished_farmers	0.8196866

Contextual information

In 2023 there were 223 active villages in 3 zones, compared with 2022's 140 active villages in 3 zones.

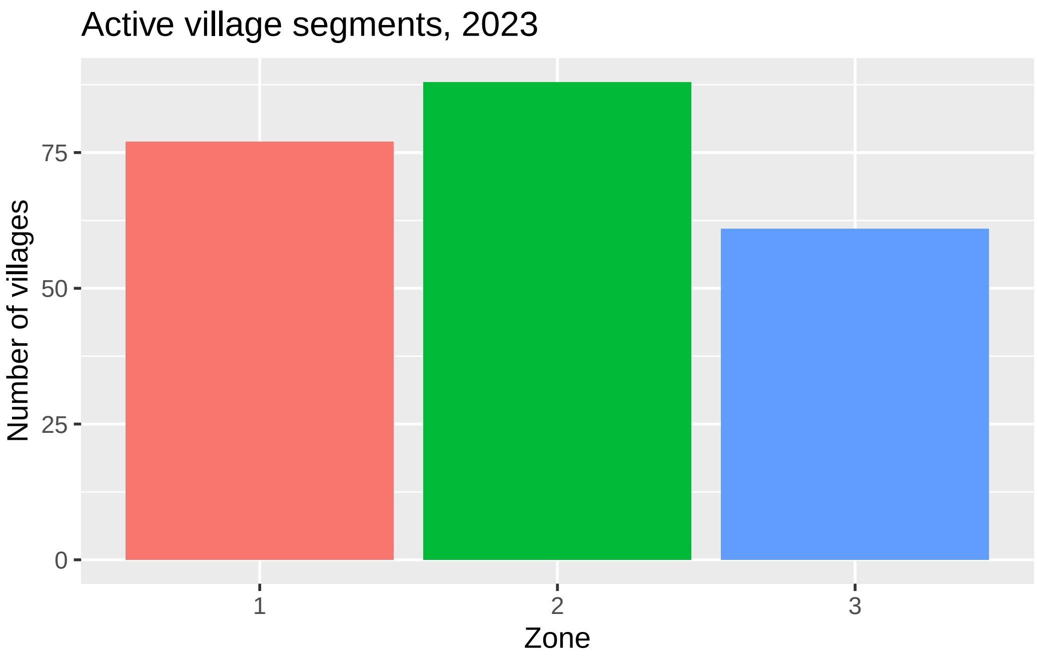
```
df %>%
  filter(season == "2022") %>%
  summarise(n = n_distinct(village))
```

n
140

```
df %>%
  filter(season == "2023") %>%
  summarise(n = n_distinct(village))
```

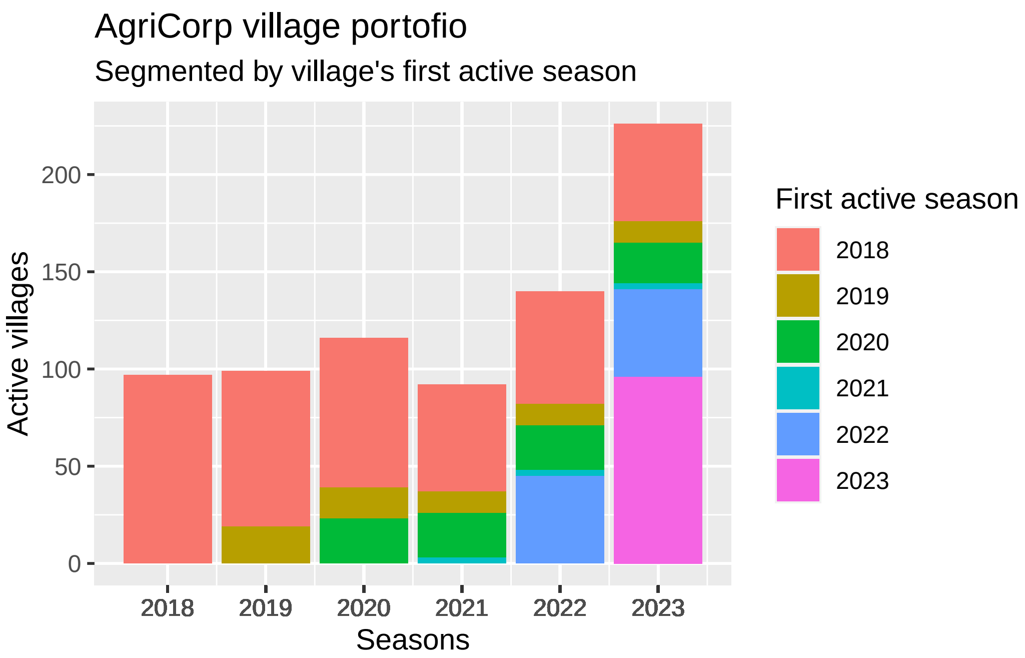
n
226

```
df2 %>%
  filter(season == "2023") %>%
  summarize(zone, n = n_distinct(village), zone) %>%
  ggplot(aes(x=zone, fill = zone)) +
  geom_bar() +
  ylab("Number of villages") +
  xlab("Zone") +
  guides(fill = FALSE) +
  labs(title = "Active village segments, 2023")
```

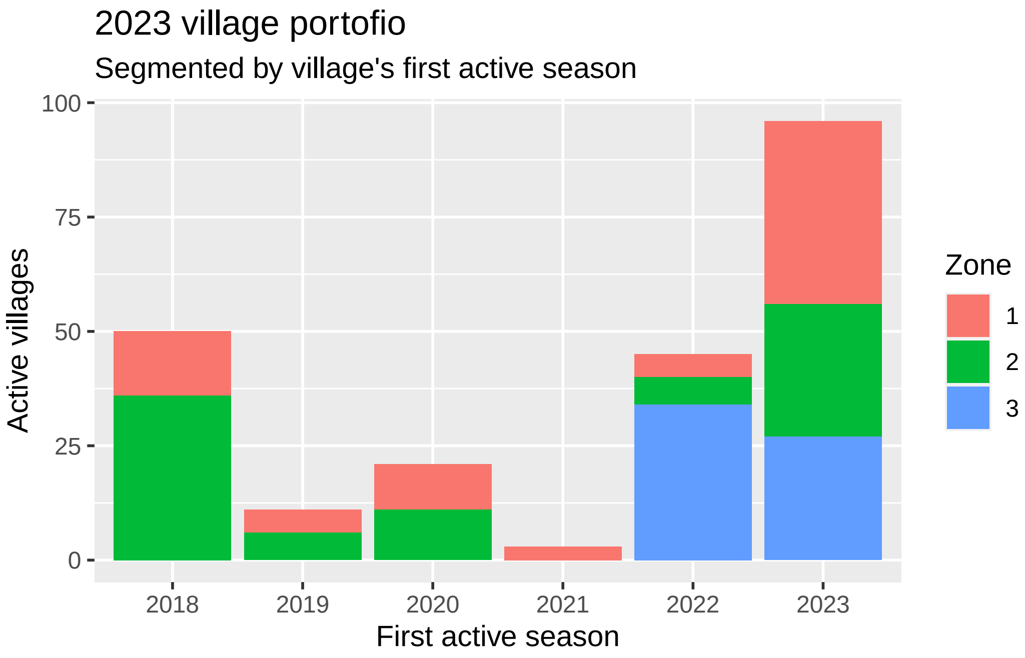


Here are a few other ways of thinking about village segments. In 2023 there seems to have been a large influx of villages in 2023 in all zones. It also seems like about half of the villages which were active in 2018 are no longer active in 2023. This offers a possible hypothesis for why villages that have been active since 2018 were so successful at growing their women-farmer ratio: at this point, only the villages in which AgriCorp is a best fit remain.

```
df2 %>%
  mutate(first_season = season - seasons_active) %>%
  mutate(first_season = as.character(first_season)) %>%
  ggplot(aes(x=season, fill = first_season)) +
    geom_bar() +
    labs(title = "AgriCorp village portofio", subtitle = "Segmented by village's first act
    ylab("Active villages") +
    xlab("Seasons") +
    scale_color_continuous(name = "First active season") +
    scale_x_continuous(labels=as.character(df2$season),breaks=df2$season) +
    scale_fill_discrete(name = "First active season")
```



```
df2 %>%
  filter(season == "2023") %>%
  mutate(first_season = season - seasons_active) %>%
  mutate(first_season = as.character(first_season)) %>%
  ggplot(aes(x=first_season, fill = zone)) +
    geom_bar() +
    labs(title = "2023 village portofio", subtitle = "Segmented by village's first active
    ylab("Active villages") +
    xlab("First active season") +
    scale_fill_discrete(name = "Zone")
```

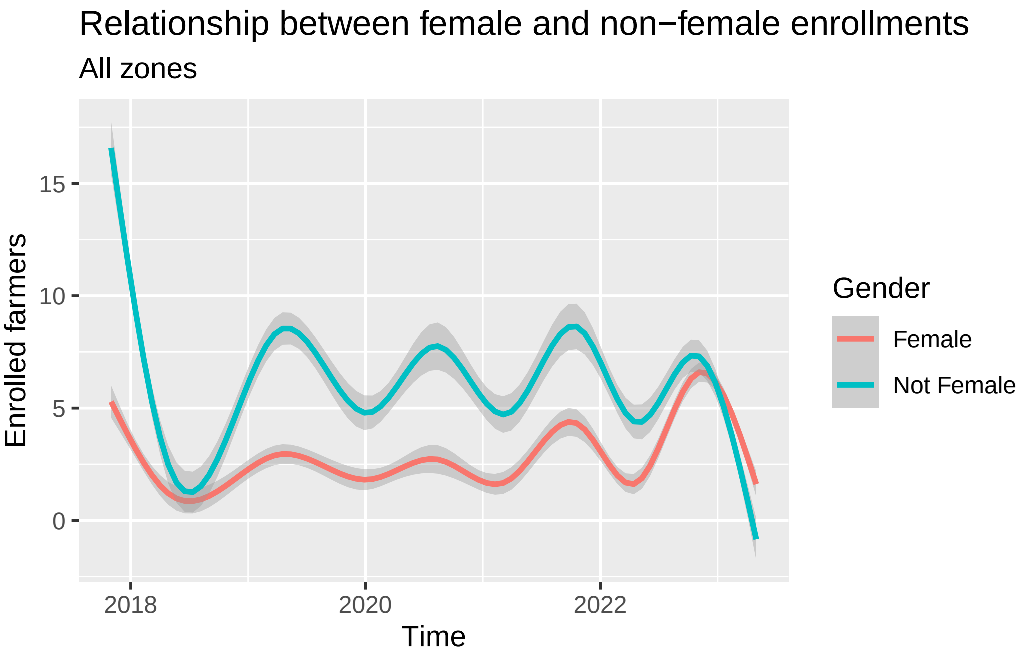


Correlation between enrolled farmers and enrolled women farmers

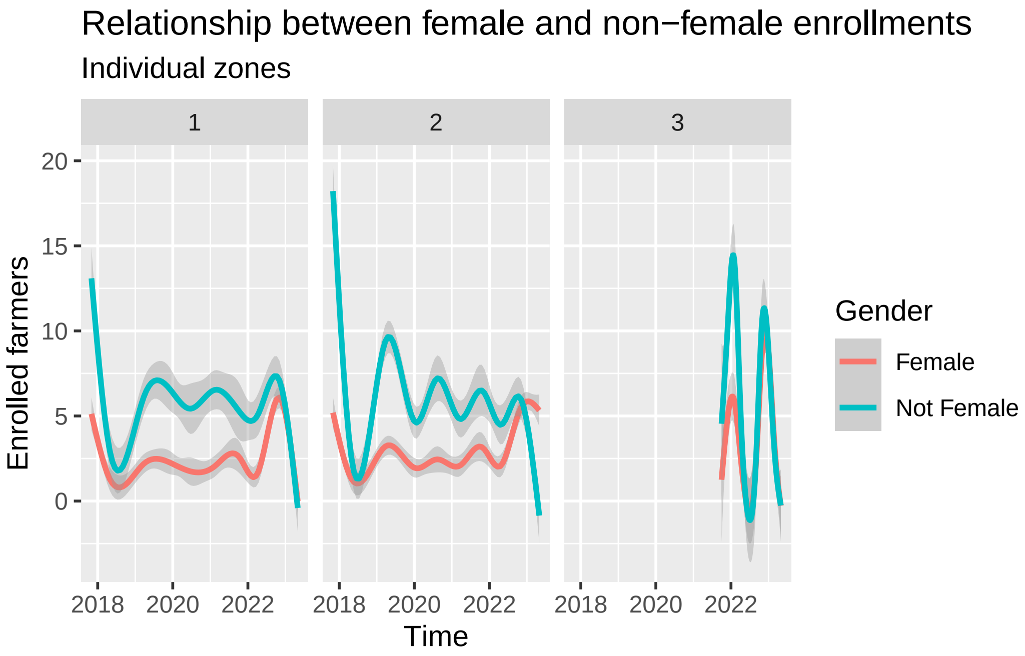
Here I tidy the data to enable a direct comparison between enrollments of farmers who were women and those who were not. I find that the two values are correlated: as one increases, so does the other. This validates my assumption that the number of women farmers on its own is not a great metric for AgriCorp improvement because it essentially tracks with the total number of farmers enrolled with the organization. The only exception seems to be in 2023 in Zone 2.

```
tidy_data <- df %>%
  mutate(enrolled_nonfemale_clients = enrolled_clients - enrolled_female_clients) %>%
  select(-enrolled_clients) %>%
  rename(female = enrolled_female_clients) %>%
  rename(not_female = enrolled_nonfemale_clients) %>%
  pivot_longer(c(5,12),names_to = "client_gender", values_to = "enrolled_clients")
```

```
tidy_data %>%
  ggplot(aes(month, enrolled_clients, color = client_gender)) +
    geom_smooth() +
    labs(title = "Relationship between female and non-female enrollments", subtitle = "All
    xlab("Time") +
    ylab("Enrolled farmers") +
    scale_color_discrete(name = "Gender", breaks=c("female", "not_female"), labels=c("Fema
```

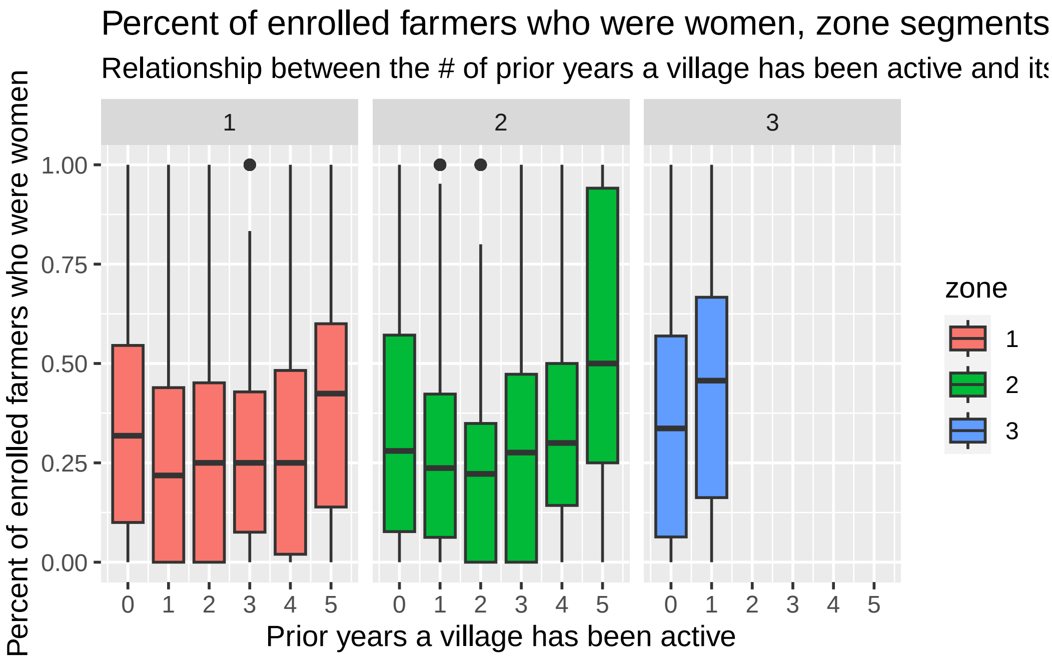


```
tidy_data %>%
  ggplot(aes(month, enrolled_clients, color = client_gender)) +
    geom_smooth() +
    facet_wrap(~ zone) +
    labs(title = "Relationship between female and non-female enrollments", subtitle = "Ind
    xlab("Time") +
    ylab("Enrolled farmers") +
    scale_color_discrete(name = "Gender", breaks=c("female", "not_female"), labels=c("Fema
```



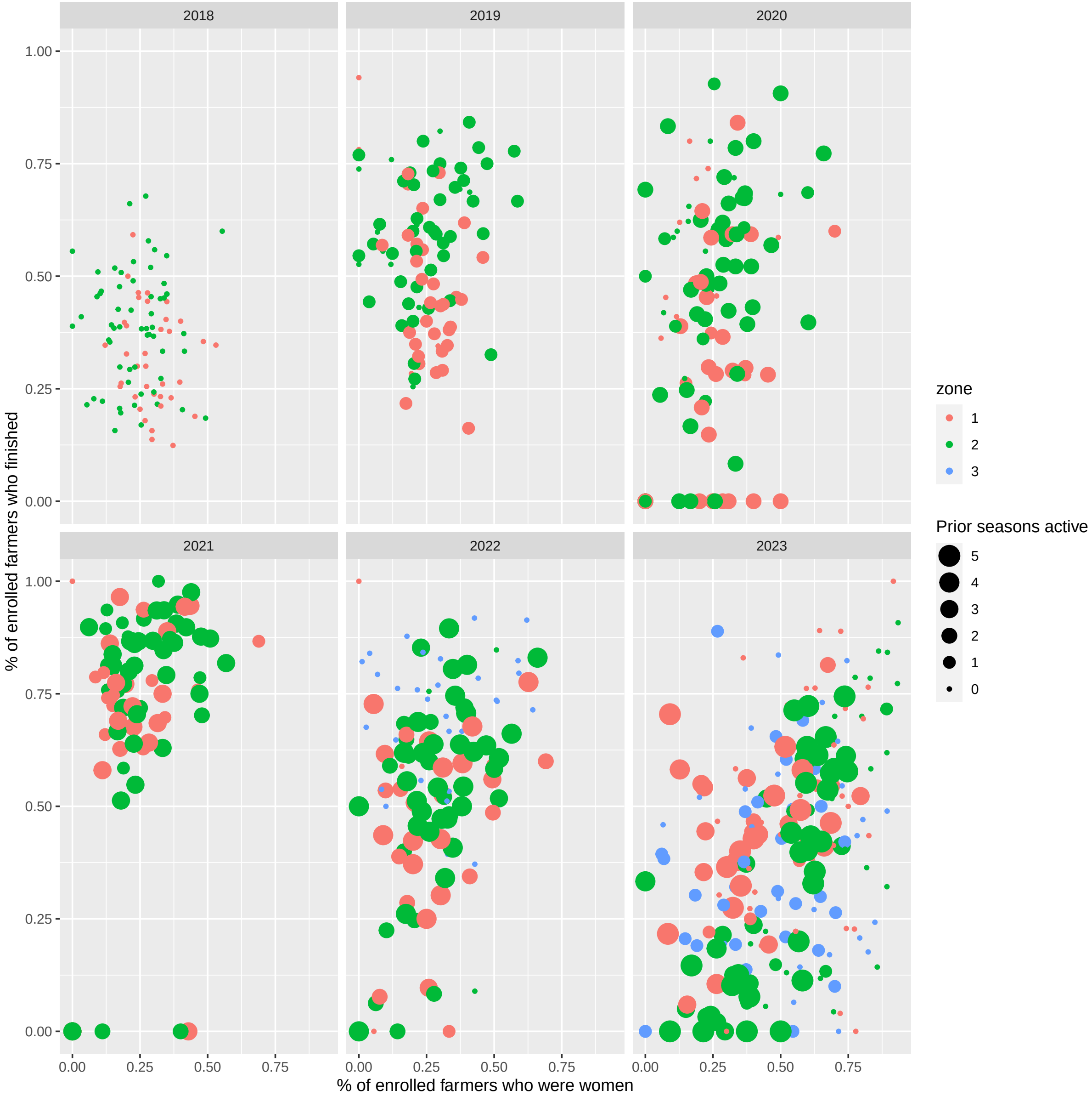
Here I wanted to look more closely at my assumption that the most 'mature' villages also were the most successful indsofar as the ratio of women to non-women farmers. It seems like that is still the case when the data is segmented by zone, and that the trend is even more pronounced in Zone 2.

```
df %>%
  ggplot(aes(x=seasons_active, y=percent_women, group=seasons_active, fill = zone)) +
    geom_boxplot() +
    facet_wrap(~ zone) +
    labs(title="Percent of enrolled farmers who were women, zone segments", subtitle="Rela
  ylab("Percent of enrolled farmers who were women") +
  xlab("Prior years a village has been active") +
  scale_x_continuous(breaks = round(seq(min(df$seasons_active), max(df$seasons_active), by
```



Here I'm playing around with the parameters of the scatterplot to see if there's anything more interesting to learn than that there's a relationship between the number of women farmers enrolled and the percent of enrolled farmers who were women. What I find is that the relationship between the women-farmer ratio and ratio of finished to enrolled farmers becomes much stronger in 2023. This visualization also shows the growth in the women-farmer ratio over time, and a bit of the relationship between the maturity of a village and its women-farmer ratio.

```
df2 %>%
  filter(percent_women < 1) %>%
  ggplot(aes(y=percent_finished, x=percent_women, color=zone)) +
  geom_jitter(aes(size = seasons_active)) +
  facet_wrap(~season) +
  expand_limits(size=0) +
  ylab("% of enrolled farmers who finished") +
  xlab("% of enrolled farmers who were women") +
  scale_size_continuous(name = "Prior seasons active", guide = guide_legend(reverse=TRUE))
```



Finally I go looking for the correlation between the women-farmer ratio and the finished-farmer ratio, and find a fairly weak correlation, which is strongest in 2023 followed by 2020.

```
df5 <- df2 %>%
  mutate_all(~ifelse(is.nan(.), NA, .)) %>%
  filter(!is.na(percent_women), zone == "2") %>%
  pivot_longer(8:9, names_to = "pct_metric", values_to = "pct_of_total")
```

```
correlations <- function(years) {

  correlations <- tibble()

  for (year in years) {
    df <- df5 %>% filter(season == year)

    correlations_year <- df %>% pairwise_cor(pct_metric, village, pct_of_total, sort = TRU

    correlations_year$year <- year

    correlations <- correlations %>% bind_rows(correlations_year)
  }
  return(correlations)
}

annual_correlations <- correlations(c("2018","2019","2020","2021","2022","2023")) %>%
  relocate(year) %>%
  rename(season = year)

print(annual_correlations)
```

```
# A tibble: 12 x 4
   season item1            item2            correlation
   <chr>  <chr>            <chr>                  <dbl>
 1 2018   percent_finished percent_women         0.0374
 2 2018   percent_women    percent_finished      0.0374
 3 2019   percent_finished percent_women         0.244
 4 2019   percent_women    percent_finished      0.244
 5 2020   percent_finished percent_women         0.479
 6 2020   percent_women    percent_finished      0.479
 7 2021   percent_finished percent_women         0.0888
 8 2021   percent_women    percent_finished      0.0888
 9 2022   percent_finished percent_women         0.174
10 2022   percent_women    percent_finished      0.174
11 2023   percent_finished percent_women         0.668
12 2023   percent_women    percent_finished      0.668
```