Exploratory Data Analysis

Taylor Lambrigger

9/26/2021

Introduction: Sociocultural Covariates and Drinking Water Source in Oaxacan Municipalities

The data set I am exploring contains demographic information about indigeneity, poverty, and population size from sampled municipalities in Oaxaca, MX. Collected from the 2020 Mexican Census (INEGI), population size captures the count of individuals living in the municipality at the time of the census in 2020. Recorded in 2015 by CONEVAL, a group working with Mexico's Ministry of Social Development, poverty shows the percentage of the municipality's population below the national poverty line. Finally, drinking water was collected from each of the 56 recorded municipalities and the source of that drinking water was recorded and later categorized as either "tap", "bottled", "well/cistern", "natural/stream", or "precipitation".

The goal of this analysis is to understand the relationships between any of these sociocultural covariates and drinking water sources exist.

```
Oaxaca <- read.csv("/Users/taylo/Desktop/2021-22/MATH 130/Week 5 - Final Project/Oaxacan Municipalities
head(Oaxaca)
```

```
##                          Municipality X..Indigenous.Speakers
## 1                AsunciÃ³n NochixtlÃ¡n                    2223
## 2                  Ayoquezco de Aldama                     325
## 3               ConcepciÃ³n Buenavista                       8
## 4                      Guadalupe Etla                     171
## 5   Heroica Ciudad de Ejutla de Crespo                     452
## 6 Heroica Ciudad de Huajuapan de LeÃ³n                    4374
##   Indigenous.Speakers.... Population.Size Poverty.... Water.Source
## 1                    10.9           20464        65.0  Bottled/Jug
## 2                     6.7            4874        88.7  Bottled/Jug
## 3                     1.1             752        92.1 Well/Cistern
## 4                     5.8            2929        31.9 Well/Cistern
## 5                     2.0           23148        74.2 Well/Cistern
## 6                     5.6           78313        53.3  Bottled/Jug
```

```
library(ggplot2)
library(knitr)
library(dplyr)
```

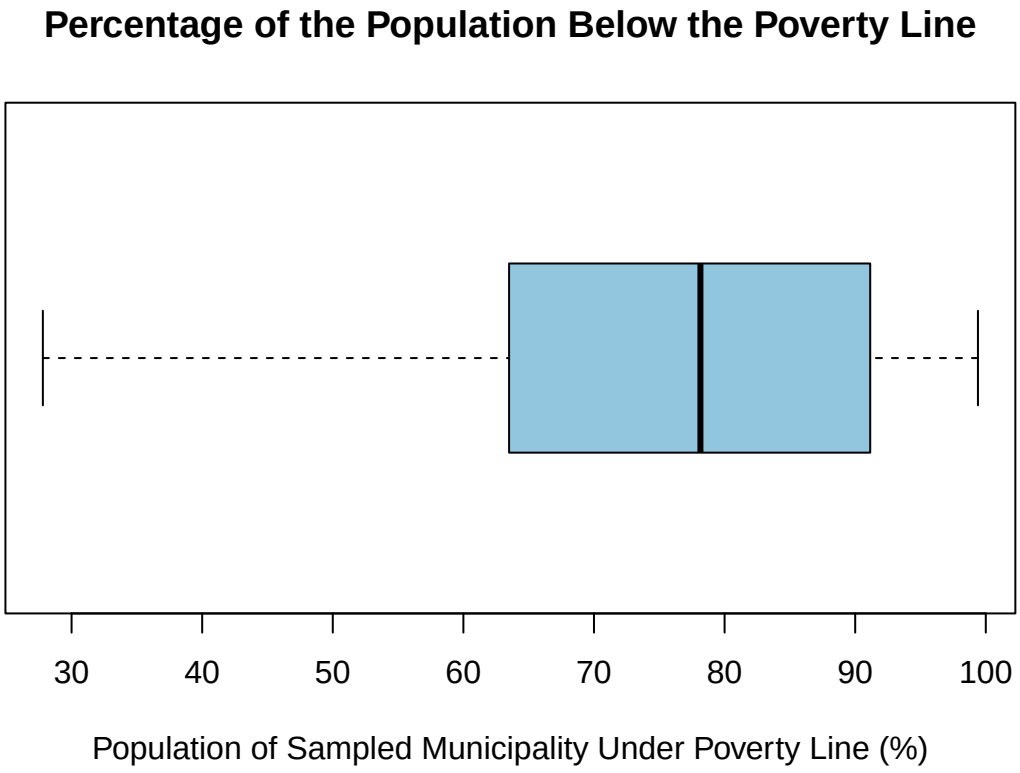
Univariate Analysis of Variables

Poverty

```
summary(Oaxaca$Poverty....)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   27.80   63.60   78.15   74.48   90.78   99.40
```

```
boxplot(Oaxaca$Poverty...., horizontal = TRUE, col="#92c5de",
        main="Percentage of the Population Below the Poverty Line",
        xlab="Population of Sampled Municipality Under Poverty Line (%)")
```



The boxplot demonstrates that in most sampled municiaplities, the majority of the population falls beneath the poverty line. While the maximum percentage (99.4%) and minimum percentage (27.8%) might initially suggest otherwise, it is clear by viewing the mean (74.48%) That the majority of municipalities have a large percentage of their population in poverty.

Population Size and Rurality

Population Size

```
summary(Oaxaca$Population.Size)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      81    1469    4925   18539   14410  270955
```

While summary statistics of the population sizes of these municipalities is helpful, the literature shows that population size doesn't mean too much in understanding water insecurity and accessible sources of drinking water. However rurality, or whether a municipality is considered rural or urban, does play a role. The Mexican Census denotes any community with a population size smaller than 2500 people is considered rural, while those with populations larger than 2500 are considered urban. As this is the information I really want to work with, I dichotomized the Population Size data into rural and urban.

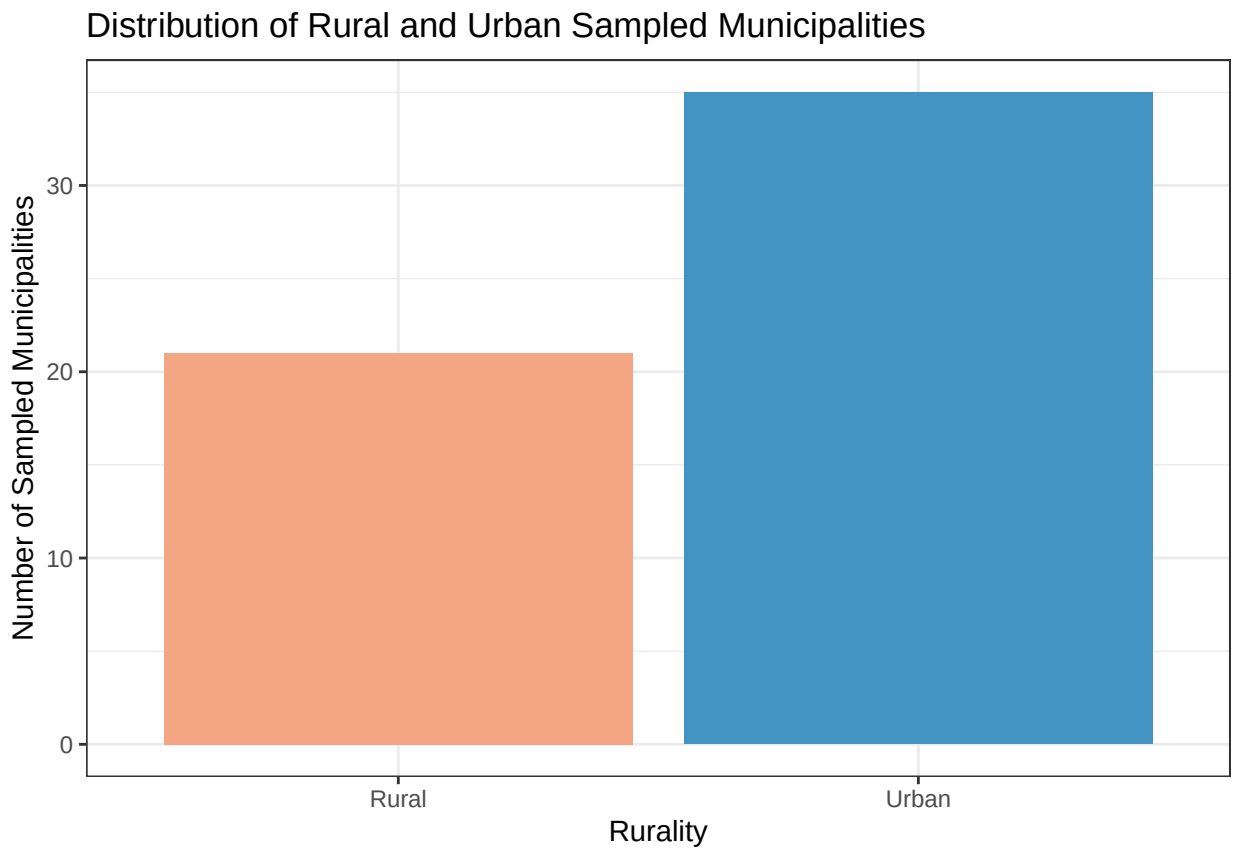
Rurality

```
Oaxaca$Rurality <- ifelse(Oaxaca$Population.Size > 2500, "Urban", "Rural")

table(Oaxaca$Rurality, useNA="no")
```

```
##
## Rural Urban
##    21    35
```

```
ggplot(Oaxaca, aes(x=Rurality, fill=Rurality)) + geom_bar() +
  theme_bw() +  scale_fill_manual(values=c("#f4a582", "#4393c3"), guide="none") +
  ggtitle("Distribution of Rural and Urban Sampled Municipalities") +
  ylab("Number of Sampled Municipalities")
```



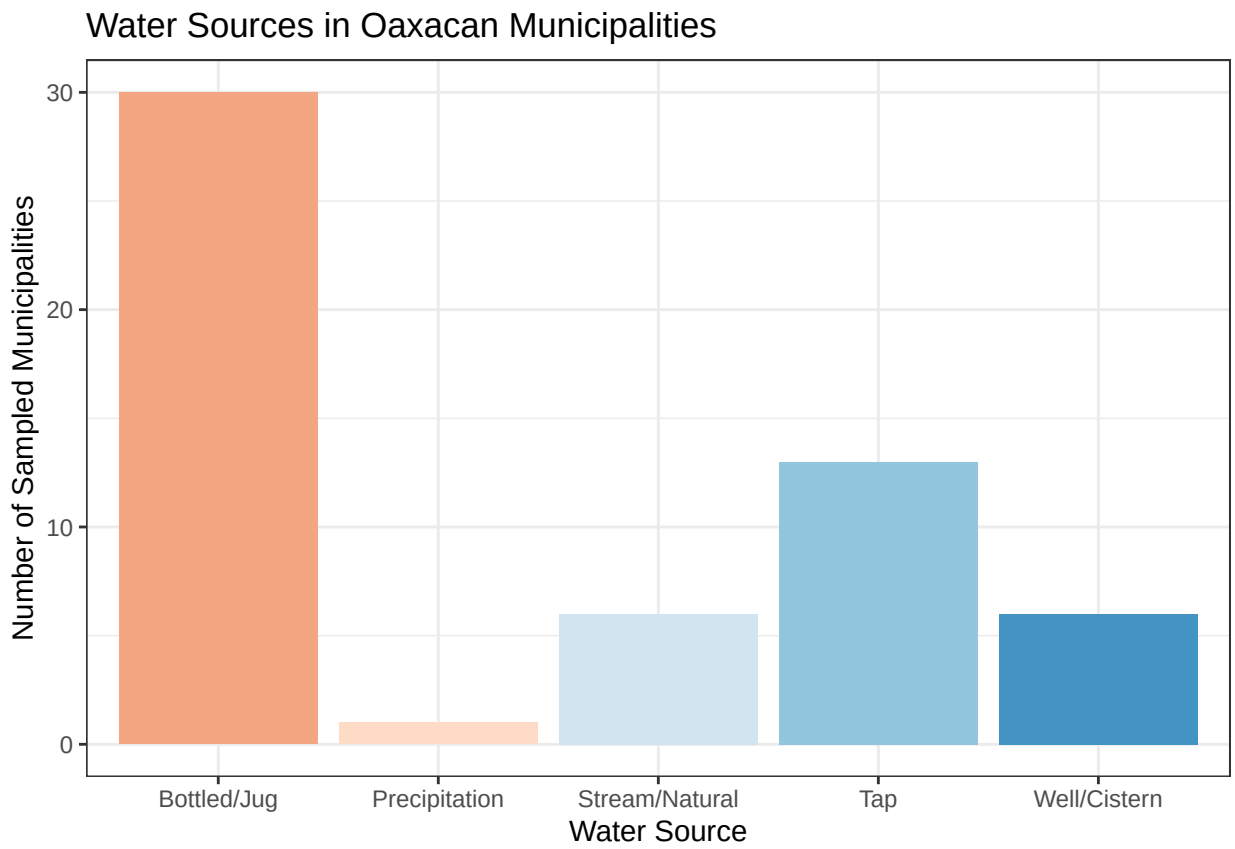
Looking at the dichotomized data, it is clear that "urban" municipalities are more present in the data population than "rural" locales. Though unevenly distributed, there are still a fair amount of rural municipalities present in the population which should allow for interesting analysis between the two.

Water Source

```
table(Oaxaca$Water.Source) %>% kable()
```

Var1	Freq
Bottled/Jug	30
Precipitation	1
Stream/Natural	6
Tap	13
Well/Cistern	6

```
ggplot(Oaxaca, aes(x=Water.Source, fill=Water.Source)) + geom_bar() + theme_bw() +
  scale_fill_manual(values=c("#f4a582", "#fddbc7", "#d1e5f0", "#92c5de", "#4393c3"), guide="none") +
  ggtitle("Water Sources in Oaxacan Municipalities") + ylab("Number of Sampled Municipalities") +
  xlab("Water Source")
```

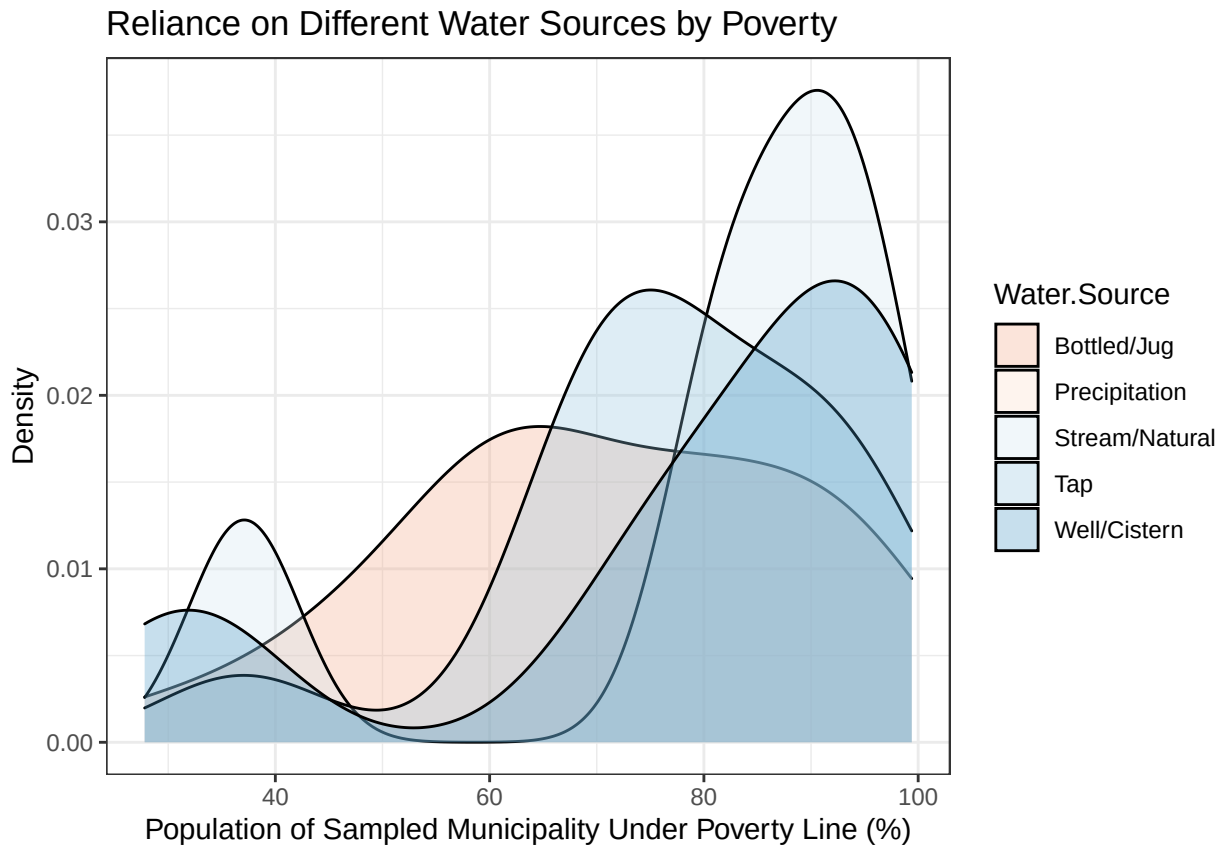


It is clear from the boxplot that the majority of municipalities rely on bottled and jugged water as their main source of drinking water. Unlike Western countries like the United States, Canada, and the United Kingdom, where tap water is the prominent source of drinking water and bottled water is the only other source of water generally available, communities in Oaxaca have a multitude of drinking water sources available to them. While tap water is the second most common source of drinking water, it is nowhere near as common as bottled water and is equal to the amount of communities who rely on wells, cisterns, and natural sources of water.

Bivariate Analysis of Variables

Poverty and Water Source

```
ggplot(Oaxaca, aes(x=Poverty...., fill=Water.Source)) + geom_density(alpha=.3) + theme_bw() +
  scale_fill_manual(values=c("#f4a582", "#fddbc7", "#d1e5f0", "#92c5de", "#4393c3")) +
  ggtitle("Reliance on Different Water Sources by Poverty") +
  xlab("Population of Sampled Municipality Under Poverty Line (%)") + ylab("Density")
```



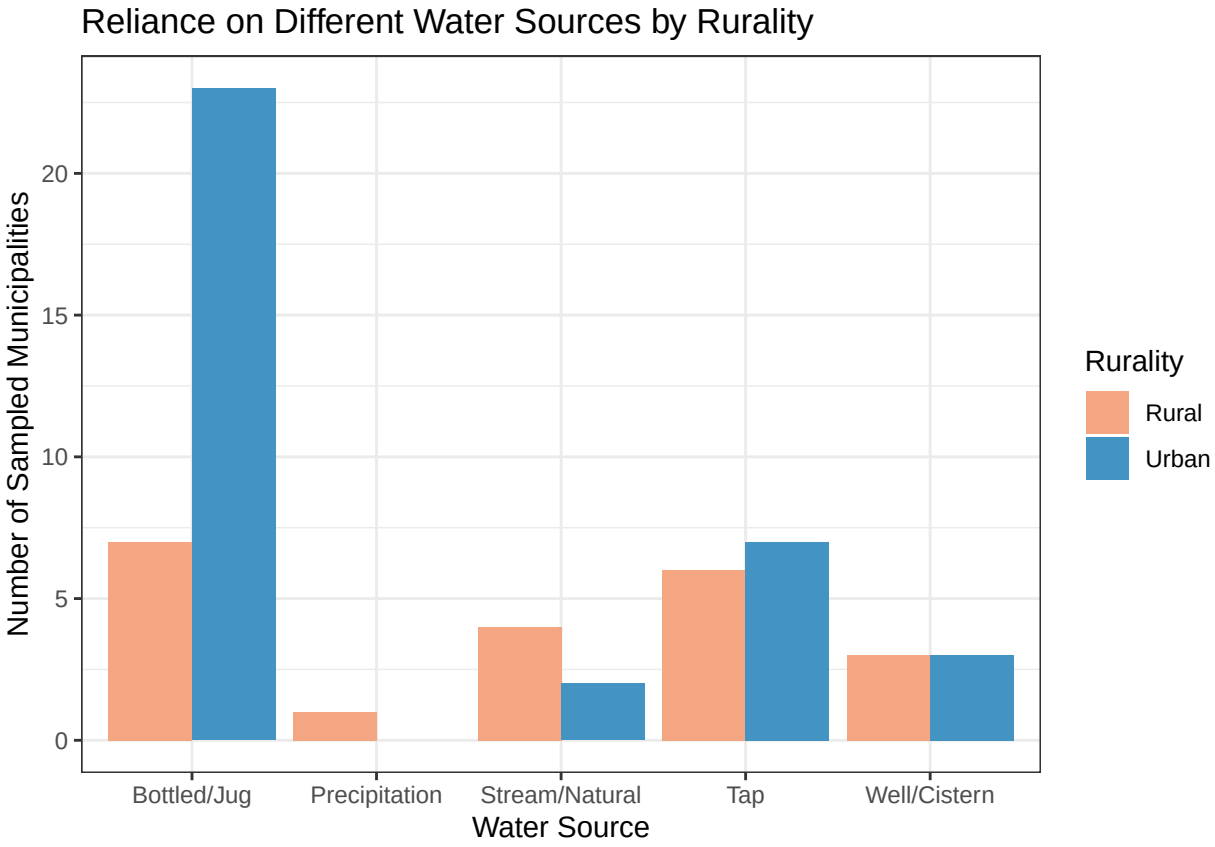
One key thing to note from the overlaid density plot is the prominence of natural water sources, as well as wells and cisterns, for those communities in extreme poverty. While bottled water is consistently used regardless, the spike in use of natural water sources and well water for those extremely impoverished communities is notable. This supports the idea that those communities that are more impoverished often rely on alternate sources of drinking water that might not conform to the typical tap or bottled water.

Rurality and Water Source

```
table(Oaxaca$Water.Source, Oaxaca$Rurality)
```

```
##
##                  Rural Urban
##   Bottled/Jug        7    23
##   Precipitation      1     0
##   Stream/Natural     4     2
##   Tap                6     7
##   Well/Cistern       3     3
```

```
ggplot(Oaxaca, aes(x=Water.Source, fill=Rurality)) +
  geom_bar(position = position_dodge(preserve = "single")) + theme_bw() +
  scale_fill_manual(values=c("#f4a582", "#4393c3")) +
  ggtitle("Reliance on Different Water Sources by Rurality") +
  xlab("Water Source") + ylab("Number of Sampled Municipalities")
```



In looking at the double bar chart, a couple trends are identifiable. First, urban populations are far more likely than rural population to rely on bottled water for drinking water, and margianally more likely to drink tap water. Water sourced from wells and cisterns is evenly distributed between rural and urban populations. Interestingly, natural sources of water and precipitation are more commonly utalized by rural communities. This could indicate that there is greater physical and economic access to bottled water and tap water in urban locales, while in rural locations natural sources of water are more accessible and reliable.