Xinavane data overview

Joe Brew and Elisa Sicuri

December 11, 2015

Contents

Datasets

There are 3 datasets from Xinavane:

```
1. Xinavane_general absenteeism_agriculture_joe.xls
2. Xinavane_plantilla trabajadores_agriculture_joe.xls
3. Xinavane_sickness_agriculture_joe.xls
```

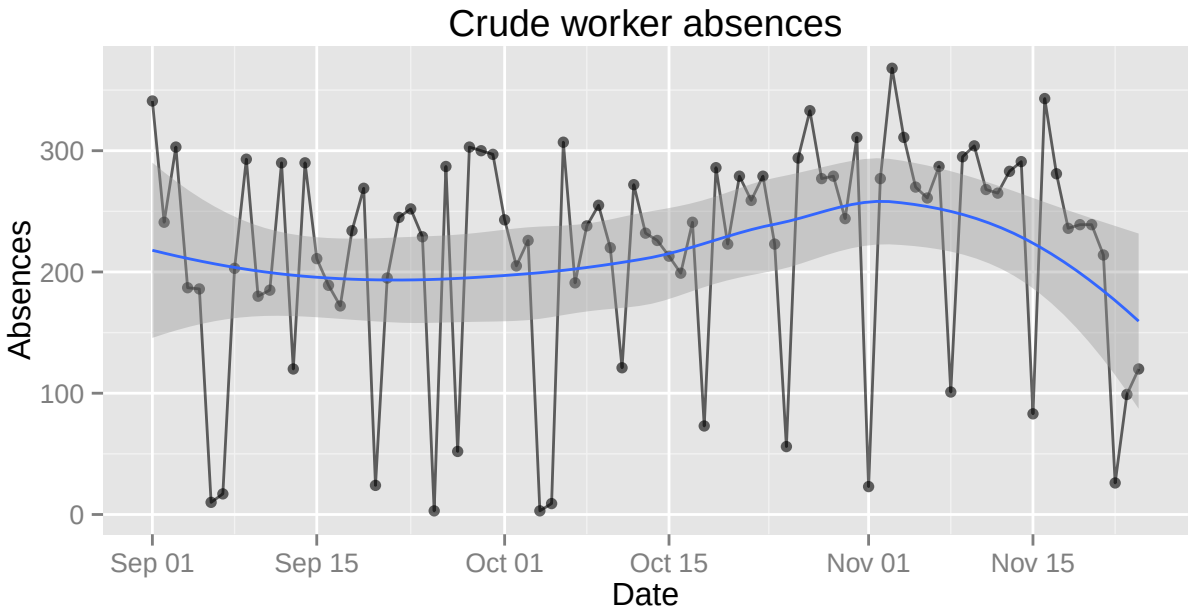
Datasets 1 and 3 pertain to worker absences. These are mutually exclusive sets (ie, absences appearing in the "general" dataset do not appear in the sickness dataset, and vice-versa).

Dataset 2 has worker sociodemographic, financial and bureaucratic details.

Exploration

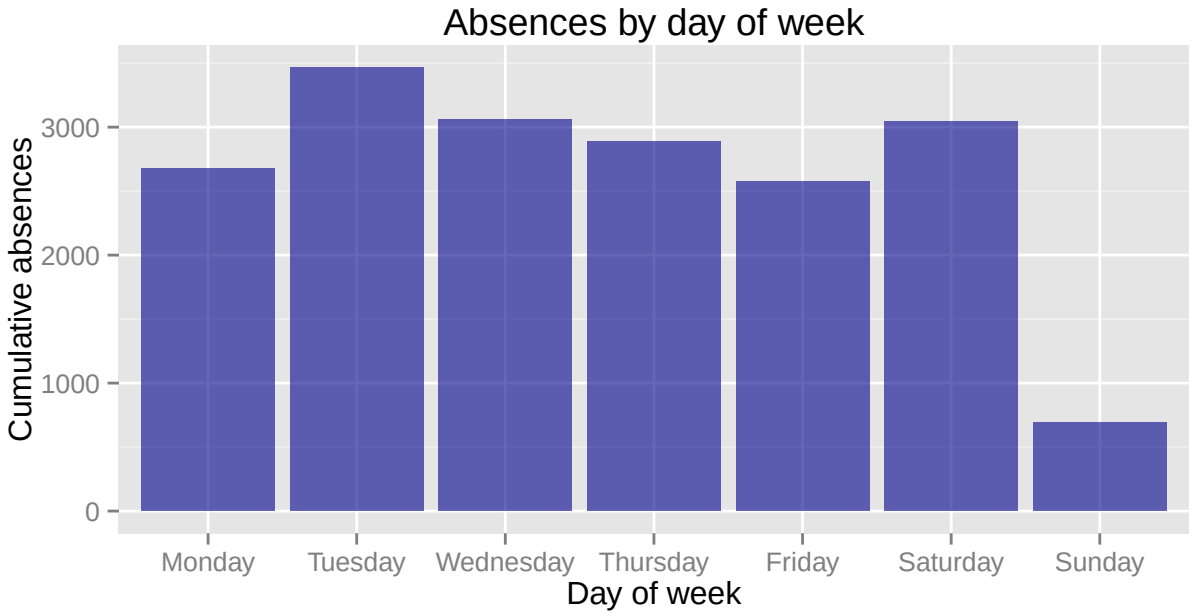
Absences

Absenteeism data spans from September 1st, 2015 until November 24th, 2015 (86 days). On average, there are 214 absences per day. Though there appears to be some longer-term variation in the below chart, it's clear that the most important factor is weekly seasonality:

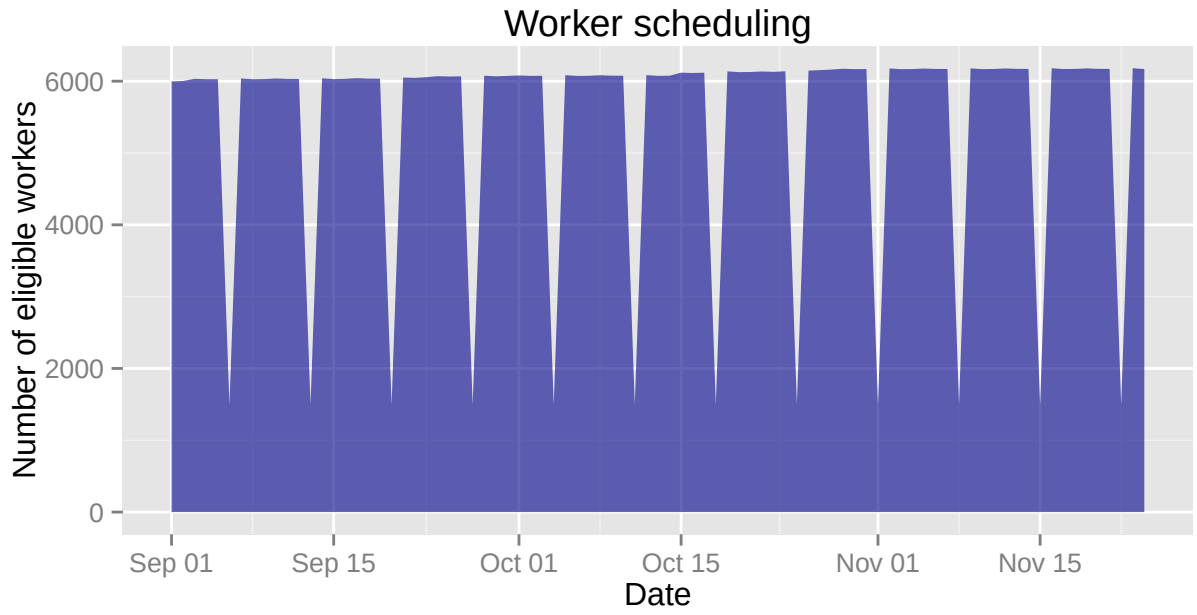


Adjustment for worker days

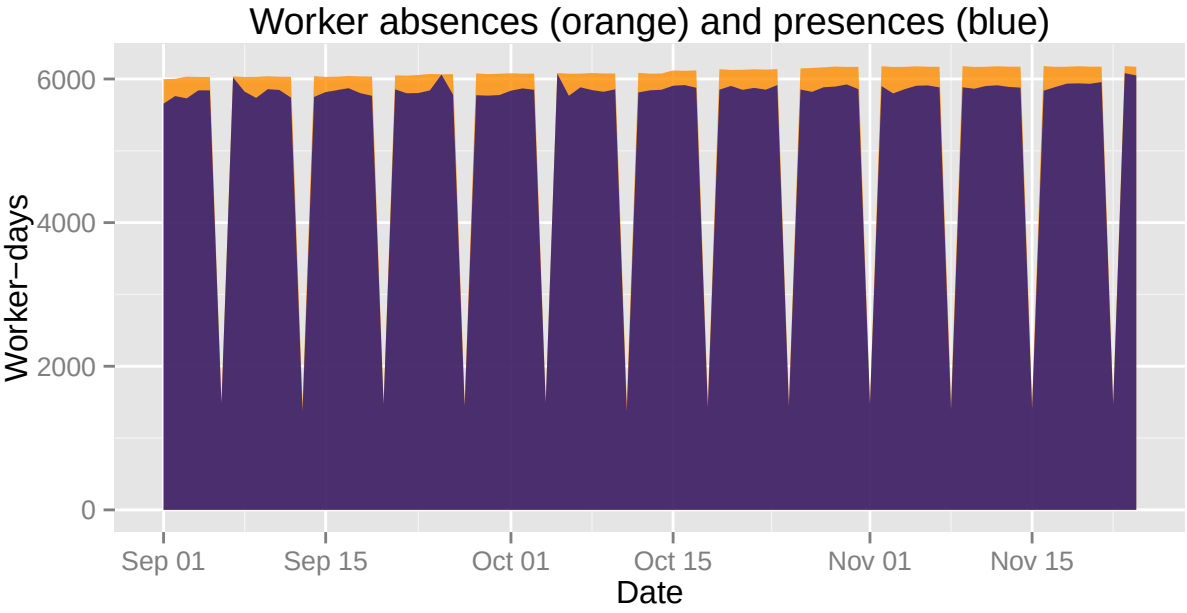
Crude absence data is relatively useless, given that it doesn't take into account the number of workers "susceptible" of absence on any given day (ie, the number of workers who were *supposed* to work). This explains why there are much fewer absences at certain times (Sunday) relative to others.



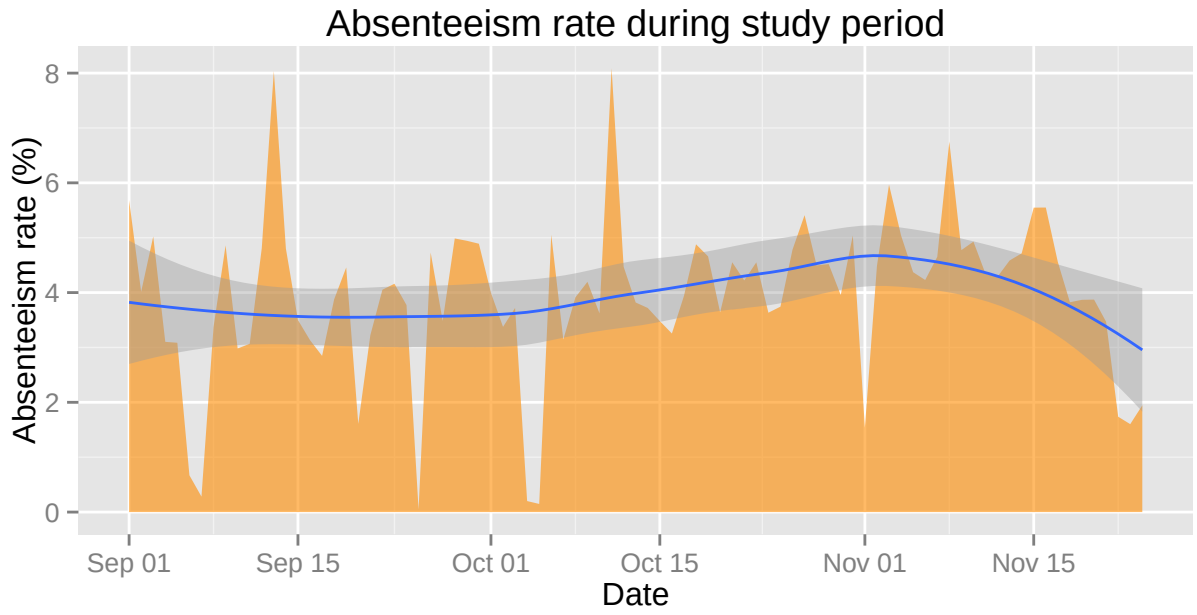
To account for the above issues, we can instead calculate an *absenteeism rate*, taking into account the employment beginning and end dates, leave statuses and working schedule of all the workers in the `Xinavane_plantilla trabajadores_agriculture_joe.xls` dataset. Essentially, we calculate the number of eligible workers for each day (the denominator in our rate). The result looks like this:



Having calculated a better denominator, we can now move forward with an improved notion of the absenteeism rate (number of worker absences / eligible workers). The result looks like this:



For a simpler understanding, we can take the relative (rather than absolute) numbers:

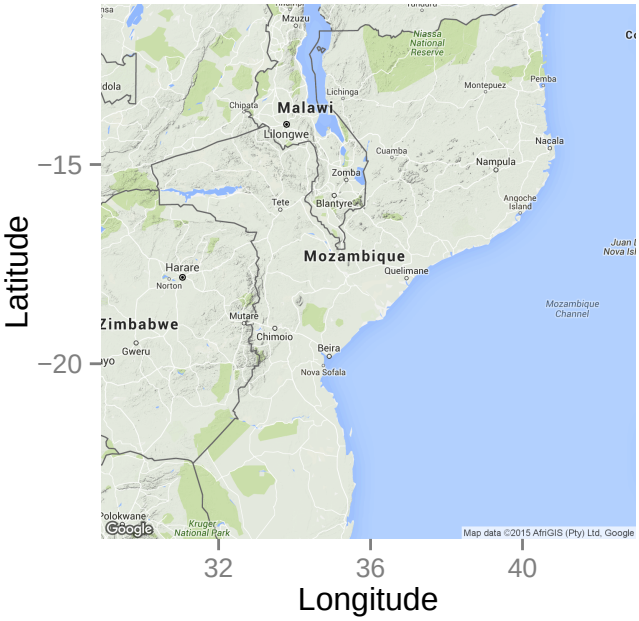


The daily average absenteeism rate (adjusted for worker eligibility) is 4%.

Other worker data

The absenteeism data are relatively straightforward. The worker data (`Xinavane_plantilla trabajadores_agriculture_joe.xls`), on the other hand, contain a wealth of relevant information.

Geography

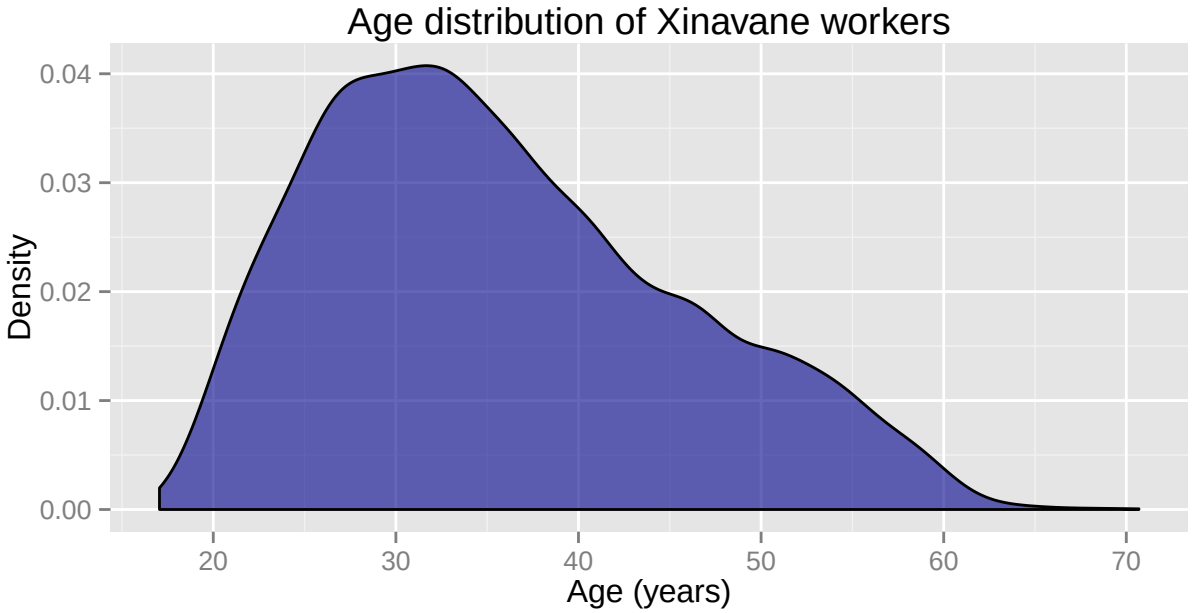


Unfortunately, the geographic data in the workers dataset is of inconsistent and low quality. The `home location` field contains many missing and proprietary names (ie, "sede"), and even when addresses are provided, modern geocoding API's are largely unable to match at a level more granular than the city/town.

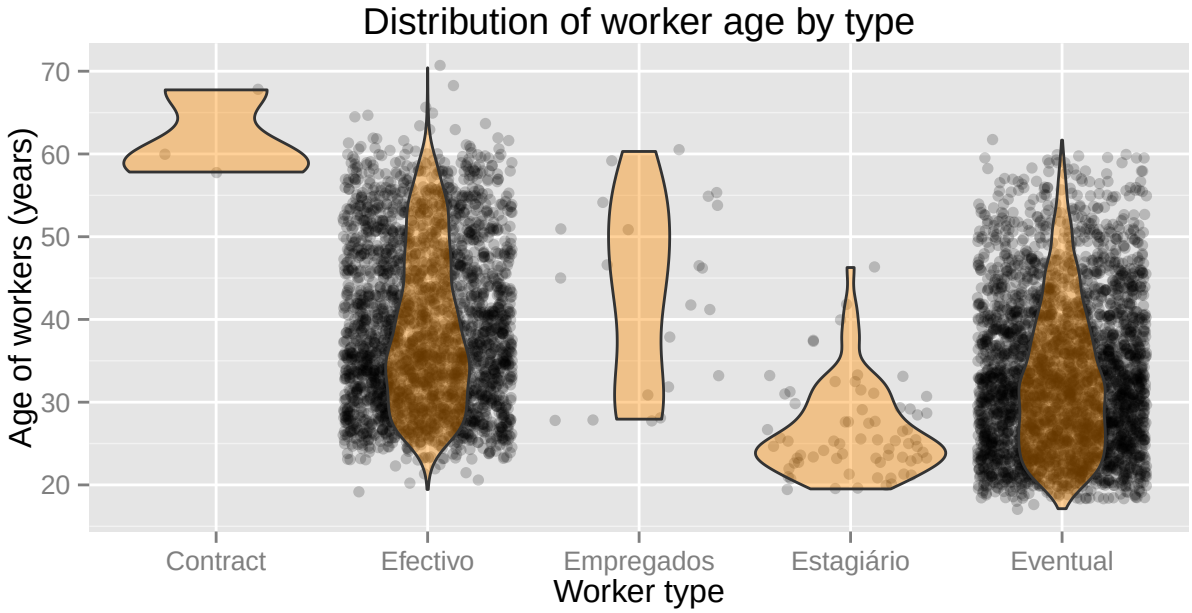
However, given the relatively small size of the number of workers (6,185), and the fact that a majority are on-site or nearby, manual geocoding of each location would be feasible (using typical hand-held cellular devices).

Age

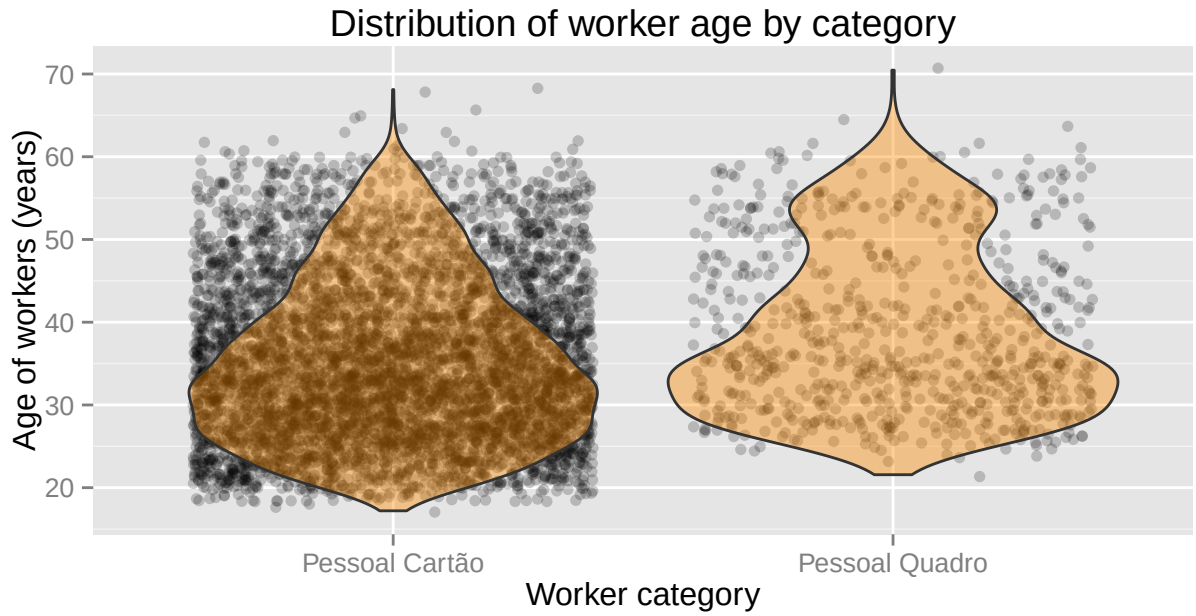
To the extent that absenteeism is confounded with health (which is in turn affected by age), it's important to note the age of workers.



Workers range from 17 to 70, with 50% falling between 28 and 42. It's worthwhile to note that the age of workers varies significantly by *type*:

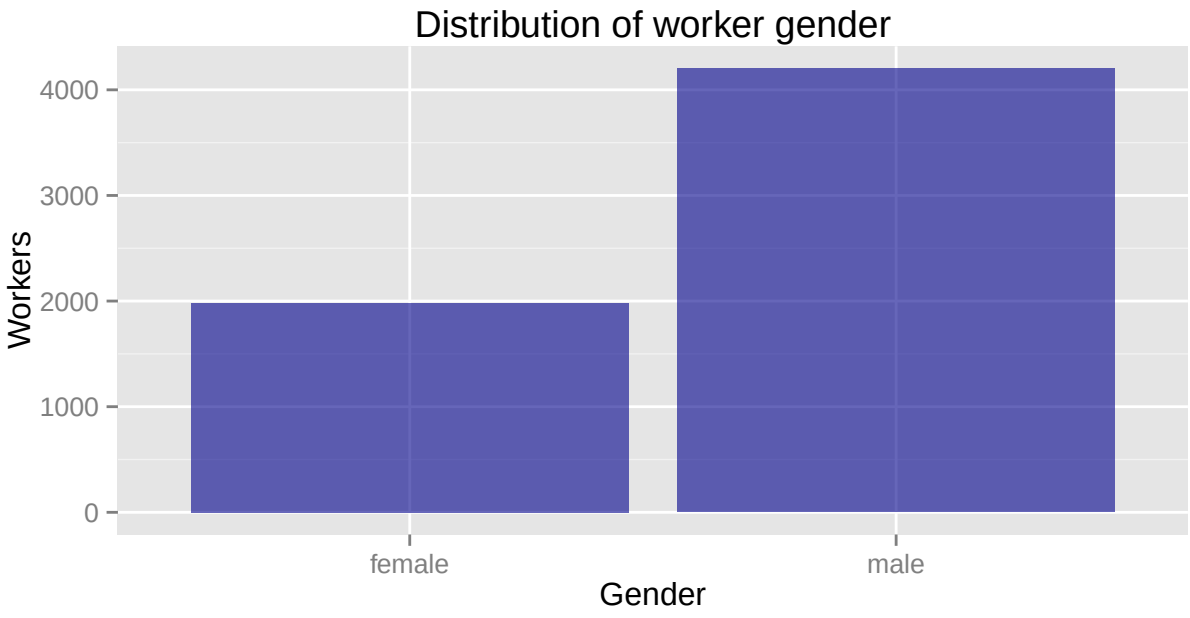


Worker age also varies significantly by category:



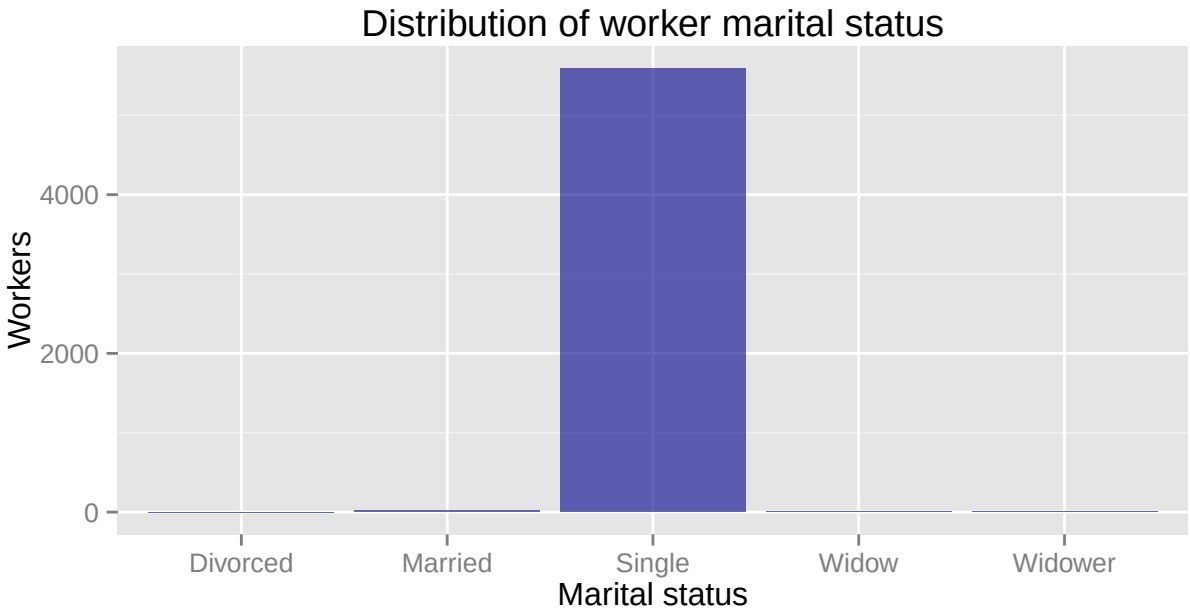
Gender

The ratio of males to females at Xinavane is greater than 2 to 1:



Marital status

Either (a) the facility has a highly unusual subset of single Mozambicans or (b) marriage is underreported in the data:



Details

All code for the cleaning, analysis and generation of this report are hosted on github.
