ds421 final project

xrw

5/3/2018

#Contents

This is a final project PDF document for DS421 stitched together from other experiments in this rpo.

Some major goals were: - Get satellite data/imagery for village and county names. - Poke around household income data - Poke around land use change for a few Taobao villages.

Section A

First we'll take a look at household income data from CHIP, an{rd geocode the counties based off a csv of "official administrative codes".

We'll also poke at the data a bit, looking at changes over time.

Section B

##CHIP

CHIP (China Household Income Project) is put out by the CIID Beijing as a longitudinal survey. It's been happening since 1988 and includes all kinds of juicy stuff including land use.

Load up necessary libraries. Some data is in `.dta` which is Stata file.

```
library(tidyr)
library(tidyverse)
```

```
## Warning: package 'tibble' was built under R version 3.4.3
```

```
## Warning: package 'stringr' was built under R version 3.4.3
```

```
library(dplyr)
library(foreign)
library(reticulate)
```

```
## Warning: package 'reticulate' was built under R version 3.4.4
```

```
library(haven)
```

```
chips_rur_1988 <- read_dta('data/1988/09836-0002-Data.dta')
chips_rur_1995 <- read_tsv('data/1995/DS0002/03012-0002-Data.tsv')
```

```
## Warning in rbind(names(probs), probs_f): number of columns of result is not
## a multiple of vector length (arg 1)
```

```
## Warning: 198 parsing failures.
## row # A tibble: 5 x 5 col     row col     expected               actual file
## ... ................. ... ......................................................................
## See problems(...) for more details.
```

```
chips_rur_2002<- read_tsv('data/2002/DS0006/21741-0006-Data.tsv')
```

```
## Warning in rbind(names(probs), probs_f): number of columns of result is not
## a multiple of vector length (arg 1)
```

```
## Warning: 244 parsing failures.
## row # A tibble: 5 x 5 col     row col   expected               actual           file
## ... ................. ... ..........................................................................
## See problems(...) for more details.
```

```
chips_rur_2007abc <- read_dta('data/2007 (2008)/RHS_w1_abc.dta')
chips_rur_2007d <- read_dta('data/2007 (2008)/RHS_w1_d.dta')
chips_rur_2007e1 <- read_dta('data/2007 (2008)/RHS_w1_e1.dta')
chips_rur_2007e2 <- read_dta('data/2007 (2008)/RHS_w1_e2.dta')
chips_rur_2007e3 <- read_dta('data/2007 (2008)/RHS_w1_e3.dta')
chips_rur_2007e4 <- read_dta('data/2007 (2008)/RHS_w1_e4.dta')
chips_rur_2007hhiexp <- read_dta('data/2007 (2008)/CHIP2007_income_and_expenditure_20150408.dta')
chips_rur_2008abc <- read_dta('data/2008 (2009)/RHS_w2_abc.dta')
chips_rur_2008d <- read_dta('data/2008 (2009)/RHS_w2_d.dta')
chips_rur_2008e <- read_dta('data/2008 (2009)/RHS_w2_e.dta')
chips_rur_2008f <- read_dta('data/2008 (2009)/RHS_w2_f.dta')
chips_rur_2008hgsg <- read_dta('data/2008 (2009)/RHS_w2_hgsg.dta')
chips_rur_2008hijk <- read_dta('data/2008 (2009)/RHS_w2_hijk.dta')
chips_rur_2008vill <- read_dta('data/2008 (2009)/RHS_w2_vill.dta')
chips_rur_2013 <- read_dta('data/2013/CHIP2013_rural_household_f_income_asset.dta')
```

Table of columns used:

Year	Net household income	Land cultivated	Number of rooms in House	Fixed production assets	Total household exp on production
1988	HNET88	LAT	HHO	VHPFP	EFP88
1995	B602	B801	B1001	B804_1	B7130
2002	na	na	na	na	na
2007	income_net	na	na	na	na
2009	na	H01	na	K01	na
2013	F01_1	L01_1	na	F07_1 + F07_2	F02_1

```
# Filter out some data from 1988 because there's missing values. They got rid of missing values in late

chips_rur_1988_filt <- chips_rur_1988 %>% filter(HNET88 != 99999999, LAT != 999.9, HHO != 99, VHPFP != 9

base::mean(chips_rur_1988_filt$HNET88)
```

```
## [1] 2739.51
```

```
base::mean(chips_rur_1995$B602)
```

```
## [1] 6812.06
```

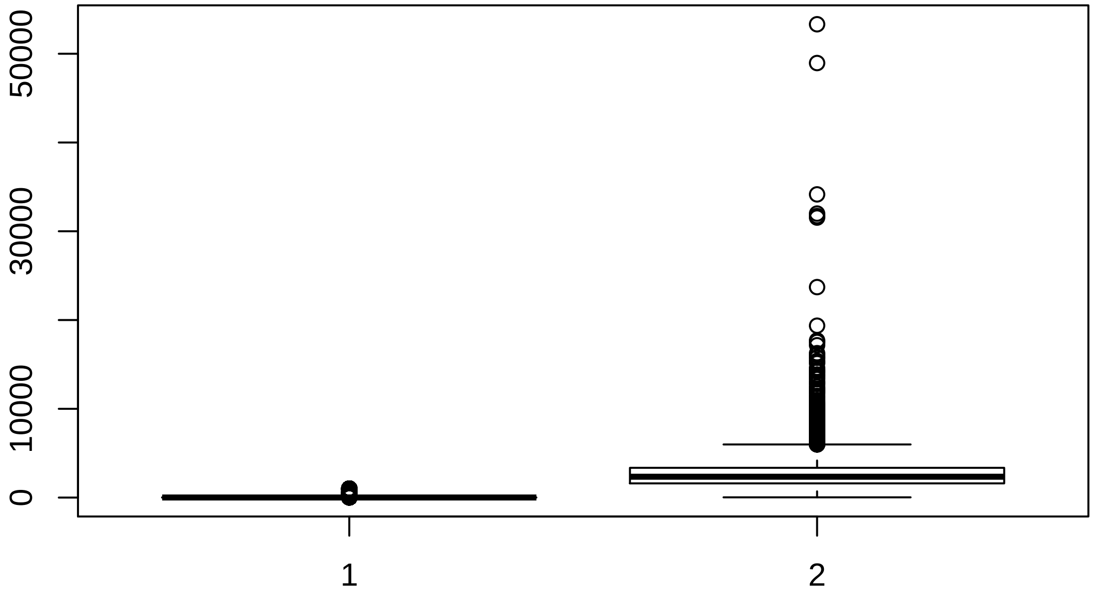
```
base::mean(chips_rur_2007hhiexp$income_net)
```

```
## [1] 19451.19
```

```
base::mean(chips_rur_2013$f01_1, na.rm=TRUE)
```

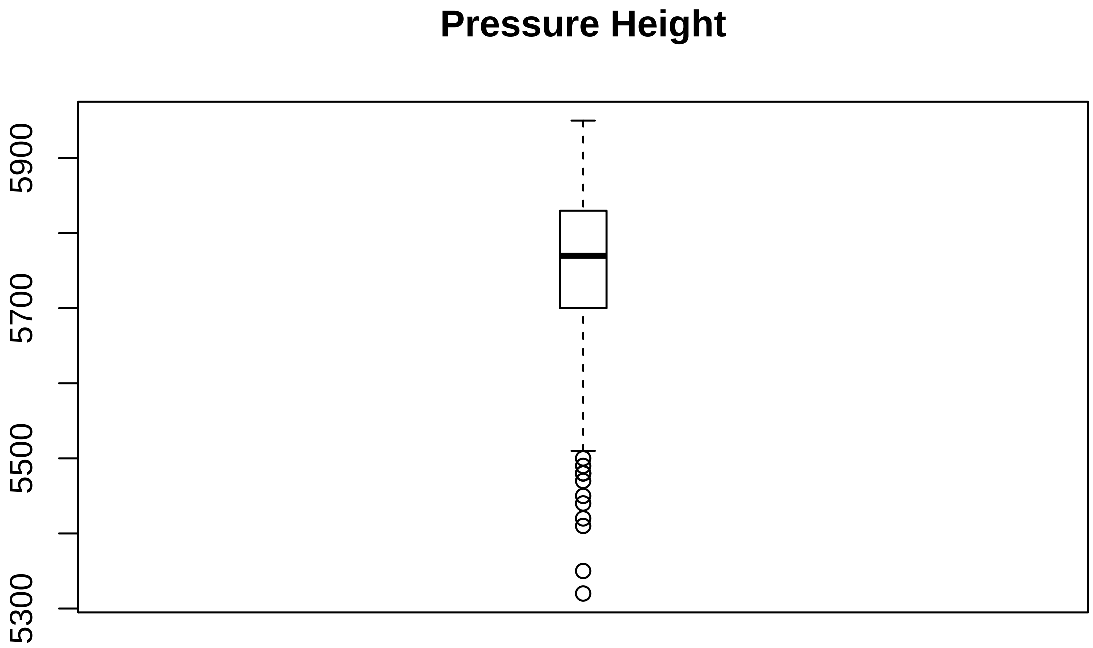
```
## [1] 45654.01
```

```
outlier_values_1988 <- boxplot.stats(chips_rur_1988_filt$HNET88)$out  # outlier values.
outlier_values_1995 <- boxplot.stats(chips_rur_1995$B602)$out  # outlier values.
outlier_values_2007 <- boxplot.stats(chips_rur_2007hhiexp$income_net)$out  # outlier values.
outlier_values_2013 <- boxplot.stats(chips_rur_2013$f01_1)$out  # outlier values.
boxplot(chips_rur_1988_filt$LAT, chips_rur_1988_filt$HNET88)
```



```
url <- "http://rstatistics.net/wp-content/uploads/2015/09/ozone.csv"
# alternate source:  https://raw.githubusercontent.com/selva86/datasets/master/ozone.csv
inputData <- read.csv(url)  # import data

outlier_values <- boxplot.stats(inputData$pressure_height)$out  # outlier values.
boxplot(inputData$pressure_height, main="Pressure Height", boxwex=0.1)
```



```
#mtext(paste("Outliers: ", paste(outlier_values, collapse=", ")), cex=0.6)
```

Taobao villages

Geocoding taobao villages

There are 1312 taobao villages as of 2017. This is under the google geocoding api limit, yay.

Testing out the geocoding response, put the province and village together (column +   column, separate by comma)