Assignment 2

Group 4

Nov 9 2015

Chris, Sune, Jeppe and Nina

Introduction to data

The dataset analyzed in this assignment contains 1000 observations on self-reported bribes in India from October 12 to November 8 2015. The data was scraped from the webpage http://www.ipaidabribe.com/. It contains information about when and where in the system the bribe took place, the geographical location and what kind of transaction the bribe was related to.

Monday the 12th of October 2015 is overrepresented in the dataset, 510 of the 1000 bribes were reported this Monday. When you look at the data grouped by weekday, this gives a false impression of bribes being more frequent on Mondays, but it is only caused by the large amount of reported bribes on this specific date. Whether there is in fact a relation between the day of the week and bribes reported would therefore require a much larger dataset which is outside the scope of this assignment.

Before starting with the analysis, we clean the dataset by removing outlier observations that might otherwise distort our findings. We leave out 6 observations where the bribes reported amount to more than 10 million INR which seems to be a misreporting. We also leave out 2 observations where only the bribe amount was reported and no information about transaction type and location was included. In general, the bribes categorised in *Others* seem to be less reliable and should be treated with caution.

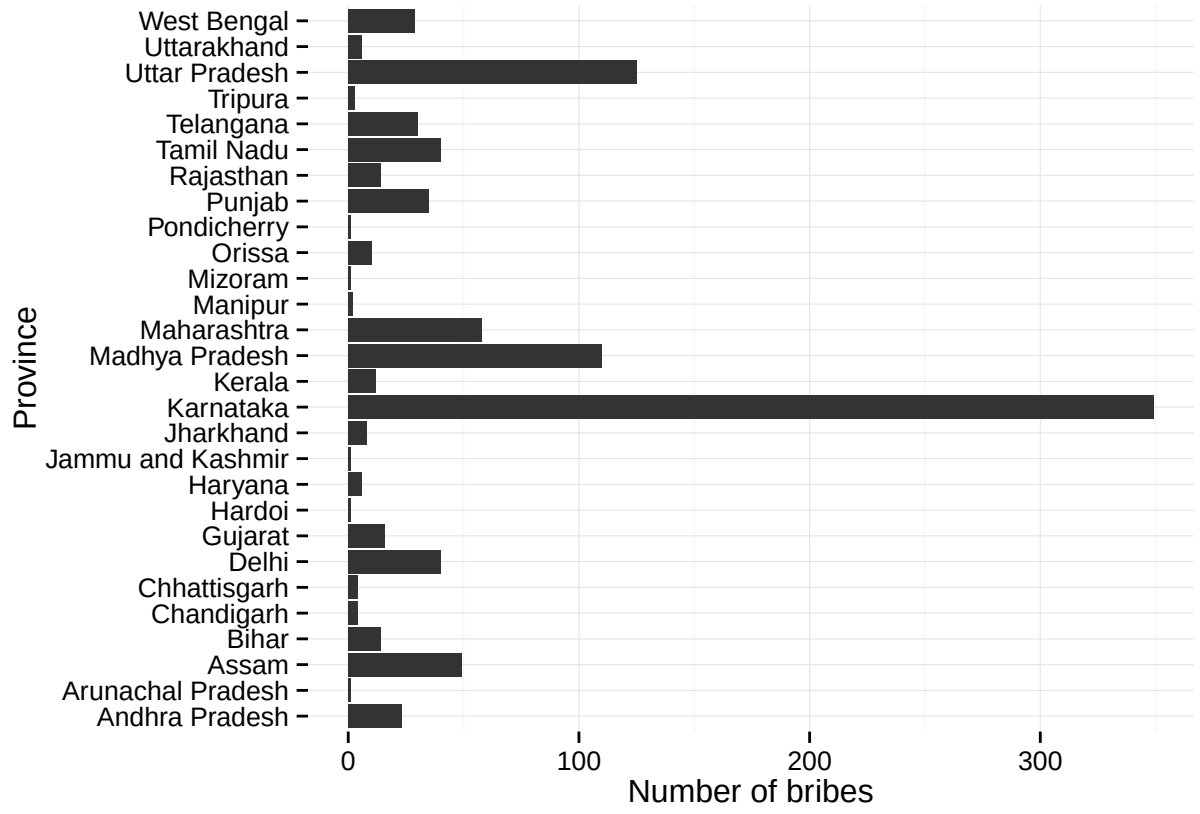
In the following analysis, we will simply denote the reported bribes as *bribes*.

Analysis

The data clearly shows that in the observed period some provinces experience more corruption than others. Karnataka is without comparison the province with most the bribes reported. This cannot be explained by the size of its population as Karnataka is only the ninth largest province in terms of population in India. For instance, Uttar Pradesh and Maharashtra have more than three times the population of Karnataka[1].

Figure 1: Number of bribes paid in each province

[1] https://en.wikipedia.org/wiki/States_and_union_territories_of_India



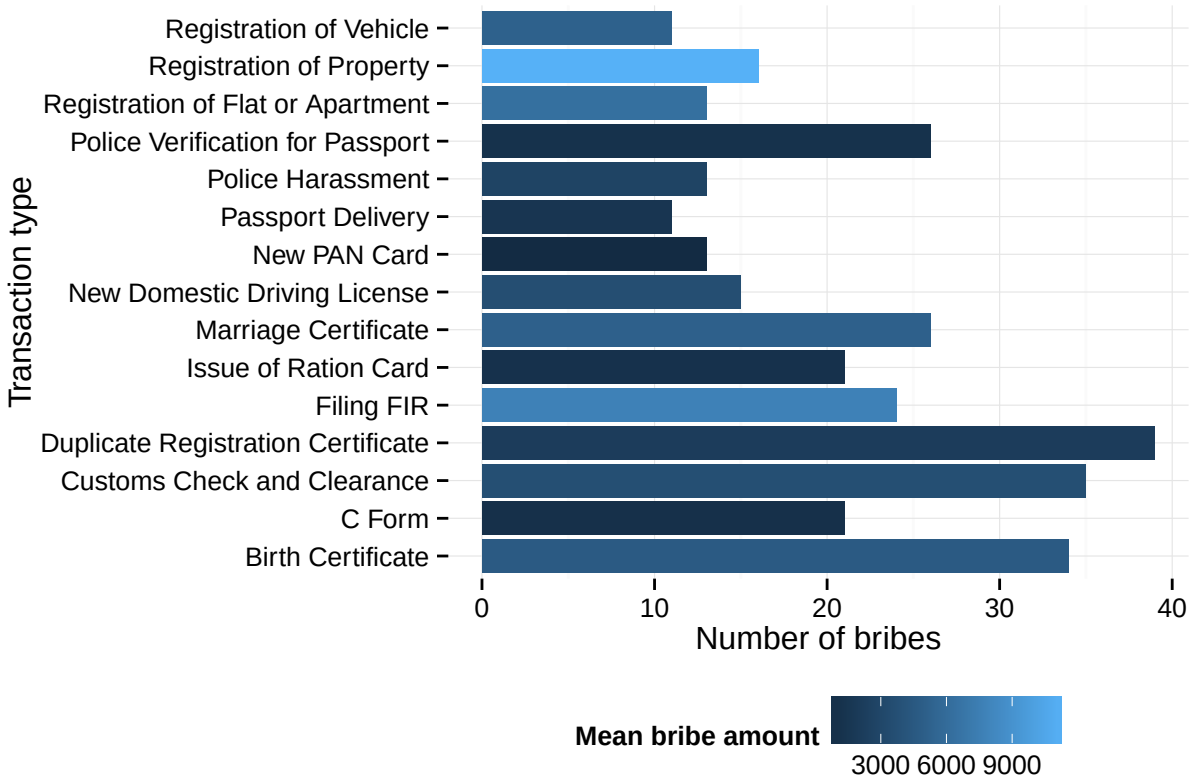
In the following, we therefore take a closer look at how the corruption in Karnataka takes place. Which parts of the system are being affected by corruption and to how much does it amount?

For Karnataka, we see that most of the bribes paid concern official documents such as certificates. We also note the issuing of ration cards make up a significant part, and due to the controversy surrounding this transaction type we will discuss the issue further. Focusing our attention on the mean bribe amount paid, registration of property is, nonsurprisingly, the highest. The filing of a First Information Report[2] is also high which validates the general perception of widespread police corruption in India (note police harassment and police verification of passport also being significant transaction types).

The reason why the province of Karnataka is the most corrupted in terms of the number of bribes is unclear, however. It's possible that there simply is more corruption, but other plausible explanations could be that the webpage from which the dataset was scraped is more known to and used by citizens of Karnataka. The province was, however, the target of a major mining corruption scandal recently[3], which could validate the theory of the province simply being corrupt. We must note that 300 of the 349 bribes reported in Karnataka was on October 12, the date at which half of the bribes were reported.

Figure 2: Transaction types in Karnataka

[2] https://en.wikipedia.org/wiki/First_Information_Report
[3] https://en.wikipedia.org/wiki/Corruption_in_India

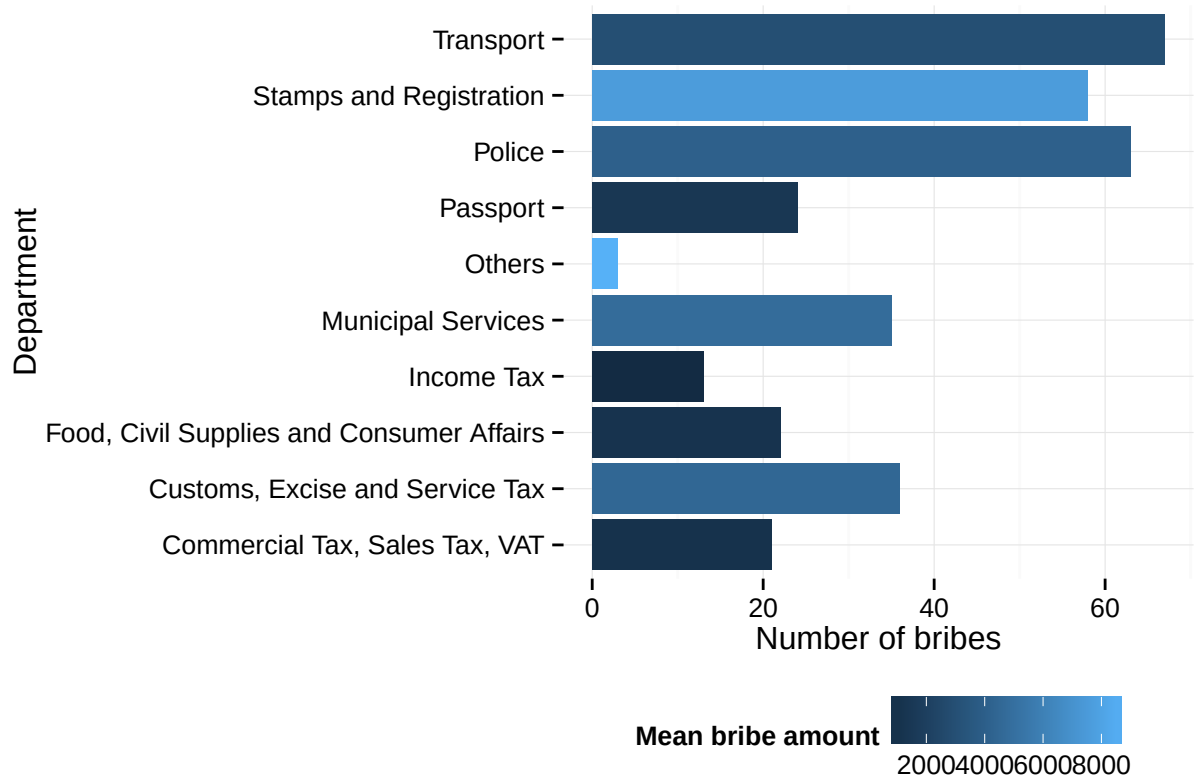


NB: Only the 15 most common transaction types are listed.

Corruption in Official Papers and Vital Necessities

We remain focused on Karnataka for a little longer and now look at the bribes by department, the sector of the economy in which they were paid. Transport, police, and stamps and registration are the largest departments in Karnataka which is in contrast to the entirety of India in which municipal services account for the most bribes followed by food, civil supplies and consumer affairs and only then police. This finding could possibly explain why Karnataka is the most corrupt province. If the daily interactions with police and transport require bribes, then the frequency at which citizens face corruption will be much higher than a province where the most corruption is in the municipality.

Figure 3: Bribes by department in Karnataka



NB: Only the 10 largest departments are listed.

Widening our attention to all of India we note that stamps and registration, which encompasses marriage certificates, registration of property, etc., has a very high mean bribe amount. This is primarily due to a single registration of property amounting to 8 million INR in Mumbai. Whether this is a misreporting or not is hard to determine as we don't know anything about who paid the bribe. If the bribe was paid by a large corporation in order to secure an important piece of land, then the amount paid is not unreasonable.

Amusingly, income tax is also a significant department in which bribes are reported. These bribes may not actually be bribes, but could simply be that some people find the act of paying taxes similar to that of corruption as you may not actually know if the money is going towards the system or are being pocketed by the cashier. The same explanation can be applied to the departments concerning other taxes and VAT.

Table 1: Bribes by department in all of India

Department	Mean bribe amount	Observations
Municipal Services	16,363	264
Food, Civil Supplies and Consumer Affairs	2,447	132
Police	16,223	124
Transport	4,141	93
Stamps and Registration	111,189	85
Others	13,647	71
Commercial Tax, Sales Tax, VAT	18,801	66
Customs, Excise and Service Tax	3,916	59
Income Tax	61,738	40
Passport	1,757	38

Ration cards

As corruption concerning the issuing of ration cards is a controversial topic, we will discuss it briefly. It is a well known issue in India that there is corruption in this area and many steps are being taken to fight it. One of the issues is that fake ration cards are issued allowing those who aren't actually eligible for them to get cheaper food and fuel. But not only fake ration cards are an issue, as persons who are eligible for one may not get it due to facing a large bribe. As ration cards are aimed at the poor, they face a hard time paying these bribes, but it would explain the relatively low mean bribe amount (2461.7633588 INR) and the even lower median bribe amount (1500 INR). It's a large concern that the poor are being exploited in such a way as they risk starvation if faced with unreasonable bribe demands.

On the other end of the spectrum, the shop owners accepting ration cards may also charge bribes in order for the poor to simply use it. And they can exploit fake ration cards to withdraw subsidies followed by selling the goods on the black market. These obvious issues are a clear indicator of the new for reform in these systems in order for India to combat the widespread corruption.[4]

Birth Certificates

The transaction type with the most reported bribes is the issuing of birth certificates. The table below suggests that this is a more general issue of corruption in India than simply being possible to attribute to one province such as Karnataka. The mean bribe amount varies a lot between the provinces with no clear pattern. Uttar Pradesh actually has more than twice as many reports as Karnataka concerning birth certificates. The fact that birth certificates account for such a large number of the bribes is a great indication of necessary official documents being a prime target for corruption.

Table 2: Birth Certificate bribe statistics by province

Province	Mean bribe amount	Observations	Population
Andhra Pradesh	1,000	4	49,506,799
Arunachal Pradesh	100	1	1,382,611
Assam	2,110	26	31,169,272
Bihar	25,041	3	103,804,637
Chhattisgarh	100	1	25,540,196
Gujarat	5,330	6	60,383,628
Haryana	10	2	25,353,081
Jharkhand	19,333	3	32,966,238
Karnataka	4,601	34	61,130,704
Kerala	2,600	6	33,387,677
Madhya Pradesh	42,700	11	72,597,565
Maharashtra	44,308	23	112,372,972
Manipur	50	1	2,721,756
Punjab	102,781	12	27,704,236
Rajasthan	42,039	6	68,621,012
Tamil Nadu	1,281	16	72,138,958
Telangana	6,343	10	35,193,978
Tripura	467	3	3,671,032
Uttar Pradesh	3,449	70	199,581,477
Uttarakhand	181	3	10,116,752
West Bengal	30,033	10	91,347,736

NB: Only actual states are reported in the table. Administrative union territories, which are excluded, only

[4] https://en.wikipedia.org/wiki/Ration_card_(India)

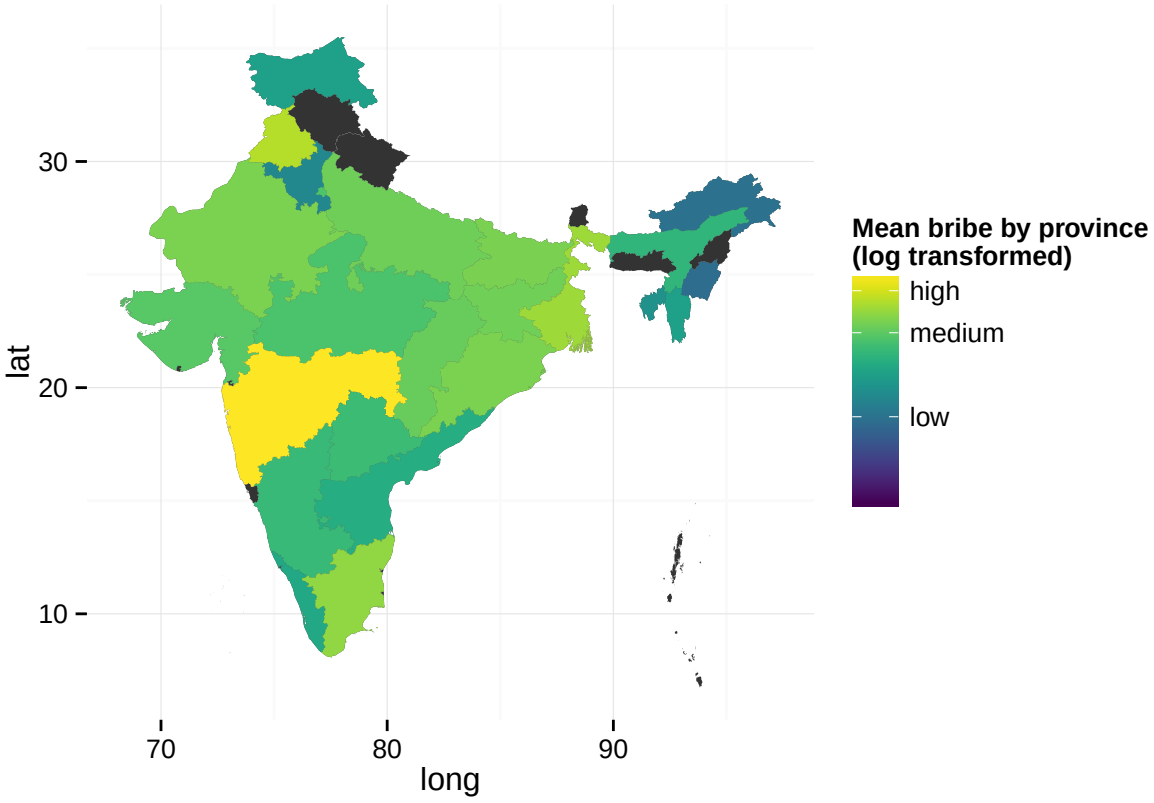
amount to a few of the total observations. Population numbers have been scraped from https://en.wikipedia.org/wiki/States_and_union_territories_of_India and are from the 2011 census.

As was the case with ration cards, corruption in birth certificates show that goods and services that are a necessity of life in most cases, are also the ones that can most easily be exploited leading to corruption.

Geography and corruption

In the map below the mean bribe paid in a given province is shown. On the map we see that Maharashtra is the province with the highest mean bribes paid and Pondicherry has the lowest mean bribe paid. However, the latter relies on just 1 observation and should be treated with caution. For a detailed overview of the number of bribes and the mean amount paid by province, please see the Appendix table.

Figure 4: Mean bribe amount by province



Some of the most corrupt provinces by the number of bribes actually have the lowest mean bribe amounts, leading to the explanation that the provinces with widespread corruption experience a large number of smaller bribes. It is difficult to draw such a general conclusion on the small number of samples, however, as we only have more than 100 observations for 3 provinces.

Closing remark

It's debatable whether the data from the website http://ipaidabribe.com is representative of the corruption in India. Firstly, the bribes are all self-reported with no validation of entries whatsoever. Bribes of unreasonably high amounts, puzzling amounts such as 12,345, etc., might just be imaginary numbers made up for the fun

of it. There is no clear strategy for cleaning the data of these supposedly false reports, as they don't all necessarily fall far from the realistic distributions of a given type of bribe. This is even more complicated without any information on the individuals who report the bribes as that could give an indication of whether the bribe is believable or not.

Another concern is that there might be selection bias as only people with access to the Internet are able to report bribes. You also need to be aware of the existence of the website, so the large number of bribes reported in Karnataka, for instance, may simply be due to widespread knowledge of the website and not because corruption is more prevalent. Coincidentally, people from remote provinces or the countryside may be completely isolated from the internet not being able to report any bribes at all.

Appendix

Table of mean bribe amount by province

Province	Mean bribe amount	Observations
Andhra Pradesh	1,940	23
Arunachal Pradesh	100	1
Assam	3,023	49
Bihar	17,830	14
Chandigarh	2,578	4
Chhattisgarh	12,062	4
Delhi	6,738	40
Gujarat	8,851	16
Hardoi	200	1
Haryana	303	6
Jammu and Kashmir	987	1
Jharkhand	14,893	8
Karnataka	3,970	349
Kerala	1,560	12
Madhya Pradesh	6,417	110
Maharashtra	207,050	58
Manipur	75	2
Mizoram	1,000	1
Orissa	18,413	10
Pondicherry	1	1
Punjab	53,785	35
Rajasthan	18,448	14
Tamil Nadu	27,729	40
Telangana	4,320	30
Tripura	467	3
Uttar Pradesh	14,145	125
Uttarakhand	7,842	6
West Bengal	35,806	29

R-code for creating dataset

```
# Load libraries
library(rvest)
library(stringr)
```

```
library(lubridate)
library(ggplot2)
library(dplyr)

# Scraping setup ----

# Init data frame
dt = data.frame()

# Define scraping function
scrape.bribes = function(dt, url) {
   # Select paid bribe nodes
   bribes = read_html(url) %>% html_nodes("section.ref-module-paid-bribe")

   # Extract information
   id = bribes %>% html_nodes(".unique-reference") %>% html_text() %>% str_extract("\\d+") %>% as.numer

   title = bribes %>% html_nodes(".heading-3") %>% html_text() %>% str_trim()

   amount = bribes %>% html_nodes(".paid-amount") %>% html_text() %>% str_extract("\\d+(,\\d+)*")
   amount = as.numeric(gsub(",", "", amount))

   department = bribes %>% html_nodes(".department > .name") %>% html_text() %>% str_trim()

   transaction = bribes %>% html_nodes(".department > .transaction") %>% html_text() %>% str_trim()

   views = bribes %>% html_nodes(".views") %>% html_text() %>% str_extract("\\d+") %>% as.numeric()

   location = bribes %>% html_nodes(".location") %>% html_text()
   city = location %>% str_extract("[\\w\\s]+") %>% str_trim()
   province = location %>% str_extract(",\\s*[\\w\\s]+") %>% str_extract("[\\w\\s]+") %>% str_trim()

   Sys.setlocale("LC_TIME", "C") # fix to prevent NA from date
   date = bribes %>% html_nodes(".key > .date") %>% html_text() %>% as.Date("%B %d, %Y")

   # Append to data frame
   rbind(dt, data.frame(id, title, amount, department, transaction, views, city, province, date))
}

#
# Scrape ----
#

start = 0
max = 1000
per.page = 10
base_url = "http://ipaidabribe.com/reports/paid?page="

for (i in seq(start, max - per.page, by = per.page)) {
   url = paste(base_url, i, sep = "")

   dt = scrape.bribes(dt, url)
```

```
   print(sprintf("Scraped %d/%d bribes.", i + per.page, max))
   Sys.sleep(1)
}

#
# Remove duplicates ----
#

dt = filter(dt, !duplicated(dt))

#
# Save data to disk ----
#

write.csv(dt, file = "~/bribes.csv", row.names = FALSE)
```