Tuberculosis Lab

Scott Karr

February 14, 2016

Tuberculosis is a contagious disease that has 19th century origins but still proliferates at high rates in the developing world. This exercise looks at a 20 year data set of countries and observed cases. The subset of data being evaluated queries this data after it has been transformed via ETL processing to replace null values with 0's and aggragate infection rates per country.

Original data-set was sourced from the World Health Organization's Tuberculosis http://www.who.int/tb/country/data/download/en/.

(a) All files used in this script can be accessed at https://github.com/scottkarr/IS607-scottkarr-wk3
(b) A local postgress database install is required to repeat the SQL ETL portion this process and can be downloaded from http://www.postgresql.org/download/ for pc or http://www.enterprisedb.com/products-services-training/pgdownload for mac
(c) This RMarkdown file populates a df_tbrates dataframe using a local db or using the csv extract https://raw.githubusercontent.com/scottkarr/IS607-scottkarr-wk3/master/extract
(d) Follow these steps to recreate the db environment and . . .

- install postgress database per above instructions
- restore backup from tb.backup https://github.com/scottkarr/IS607-scottkarr-wk3/commit/d1290105542228b3c62d7638ec393c082e2cdc8b
- import population.csv to your local tb instance https://raw.githubusercontent.com/scottkarr/IS607-scottkarr-wk3/master/population.csv
- run exp_tb_rates etl script on your local tb instance https://raw.githubusercontent.com/scottkarr/IS607-scottkarr-wk3/master/exp_tb_rates.sql

loads the PostgreSQL driver

```
library(RPostgreSQL)
```

```
## Loading required package: DBI
```

```
library(TTR)
```

assign connection parameters and connect to db

```
dbname <- "tb"
dbuser <- "postgres"
dbpass <- "postgres"
dbhost <- "localhost"
dbport <- 5432
drv <- dbDriver("PostgreSQL")
con <- dbConnect(drv, host=dbhost, port=dbport, dbname=dbname,user=dbuser, password=dbpass)
```

After initially loading the tb database, the objective is then to discover useful patterns that are easily explainable. The focus on of this investigation is on infection rates which we will measure using aggregate

ordinal statistics such as head counts and trends regressed by year. The primary intuition is that data-reducing to one set of observations per country of infections and trends, it becomes possible to identify which countries may be facing an epidemic and which countries are succesfully reversing a trend despite high infection rates. Finally, this data set initially considers only the top 25 countries by headcount so as to isolate the observed cases of greatest consequence first. A follow-on investigation may look into details specific to countries that are at high risk as well as countries with low infection rates yet that are trending higher.

query tuberculosis rates with cleaned data

```
query <- dbSendQuery(
            con, query <- "
              select
              country,
              round(avg(population)),
              round(avg(rate_per_100thsd)) avgRate_per_100thsd,
              CASE
                WHEN regr_slope(rate_per_100thsd, year) > 0 THEN 'higher'
                WHEN regr_slope(rate_per_100thsd, year) < 0 THEN 'lower'
                ELSE 'na'
              END trend,
              regr_slope(rate_per_100thsd, year) regr_slope
              from   tb_rates
              where 1=1
              and   rate_per_100thsd <> 0
              group by country
              order by round(avg(rate_per_100thsd)) desc
              limit 25        ")
rs1 <- fetch(query,n=-1)
```

be good citizen, closeout connect {r disconnect, eval=TRUE} dbDisconnect(con) dbUnloadDriver(drv) “‘

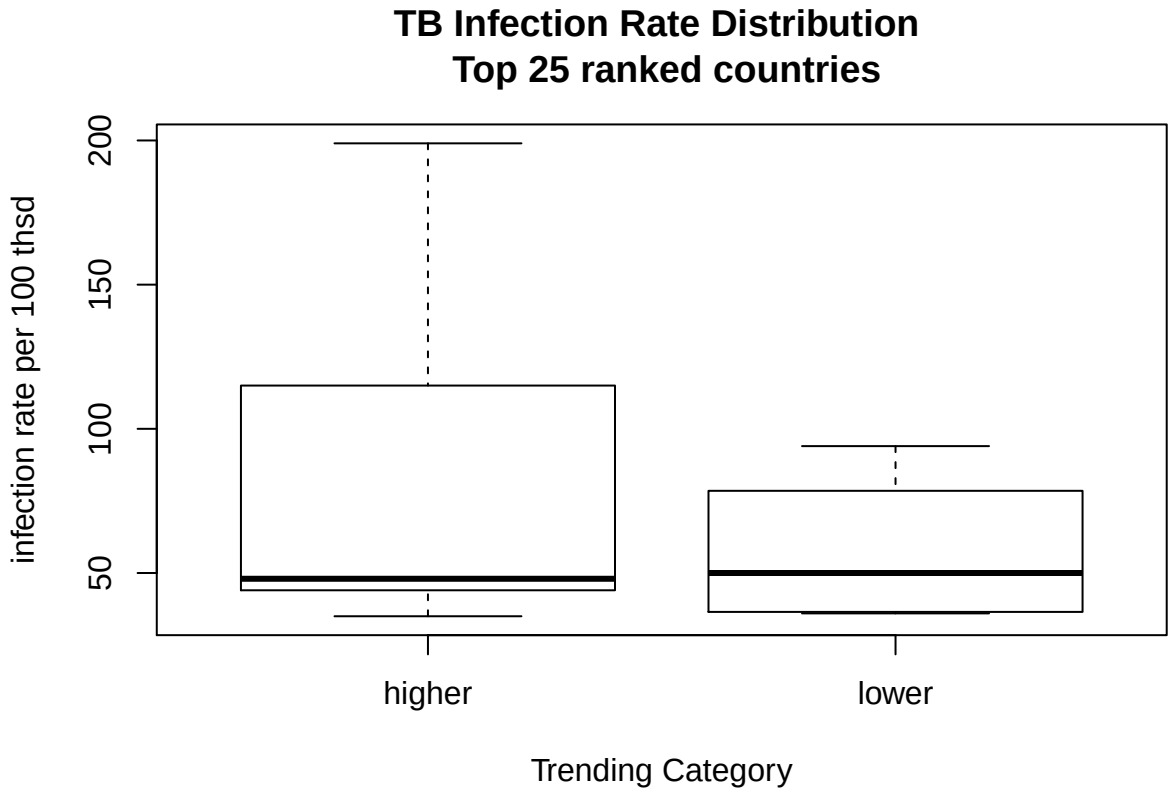
Load Data Frame from website *Note, assigning from rs1 is duplicative but loads data from database and csv extract

```
theUrl <- "https://raw.githubusercontent.com/scottkarr/IS607-scottkarr-wk3/master/extract"
df_tbrates <- read.table(file = theUrl, header = TRUE, sep = ",")
df_tbrates <- rs1
colnames(df_tbrates)[2] <- "avgpopulation"
```

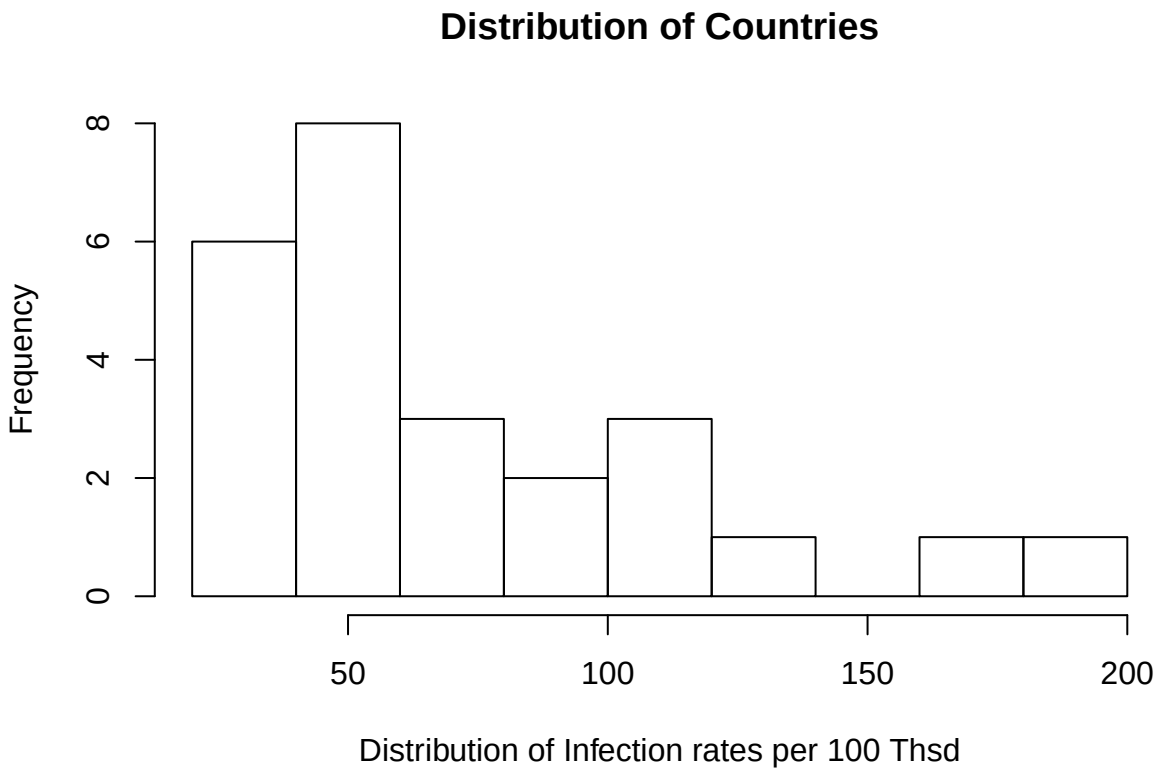
Note the few outlier countries with 100+ cases per 100 thousand persons such as South Africa. Also note that most trends show an increasing rate of ~5% however, there are outliers that exceed 20%.

```
boxplot(df_tbrates$avgrate_per_100thsd ~ df_tbrates$trend,
        main="TB Infection Rate Distribution \r\n Top 25 ranked countries",
        xlab="Trending Category",
        ylab="infection rate per 100 thsd"
)
```

```
## Warning in title(main = "TB Infection Rate Distribution \r\n Top 25 ranked
## countries", : font width unknown for character 0xd
```

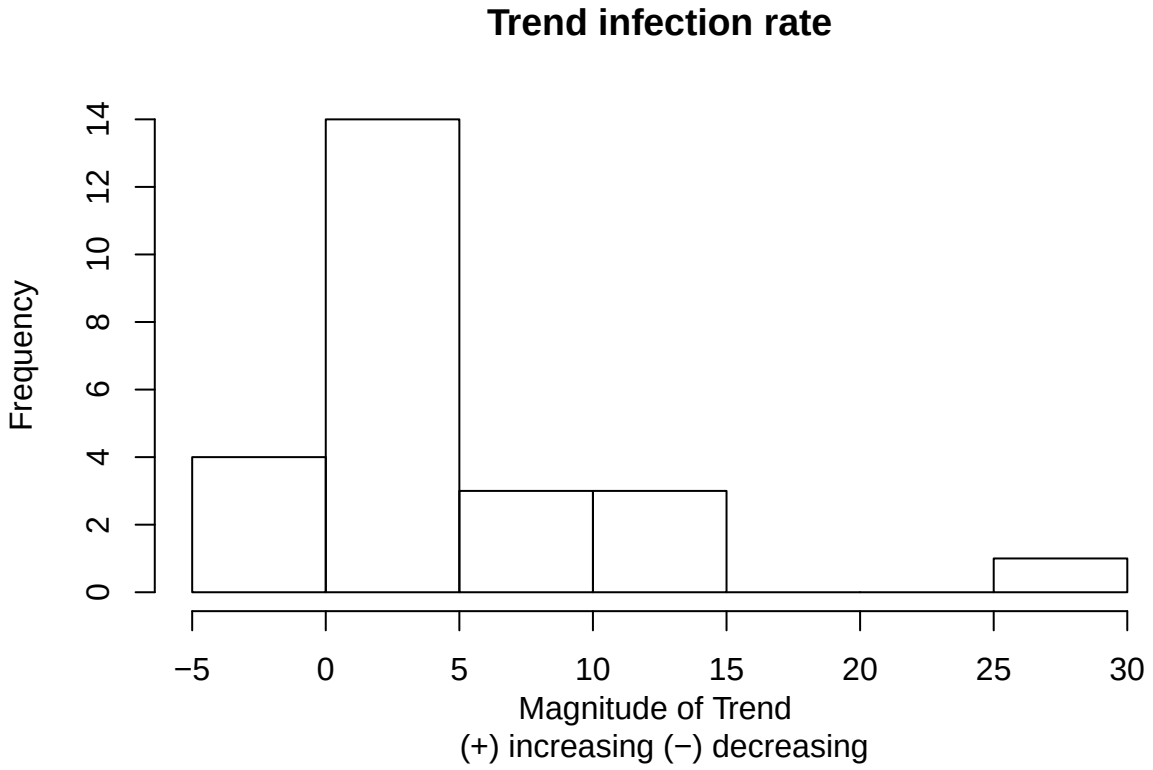



```
hist(
  df_tbrates$avgrate_per_100thsd,
  main="Distribution of Countries",
  xlab="Distribution of Infection rates per 100 Thsd"
)
```




```
hist(
    df_tbrates$regr_slope,
    main="Trend infection rate",
    xlab="Magnitude of Trend \r\n (+) increasing (-) decreasing"
)
```

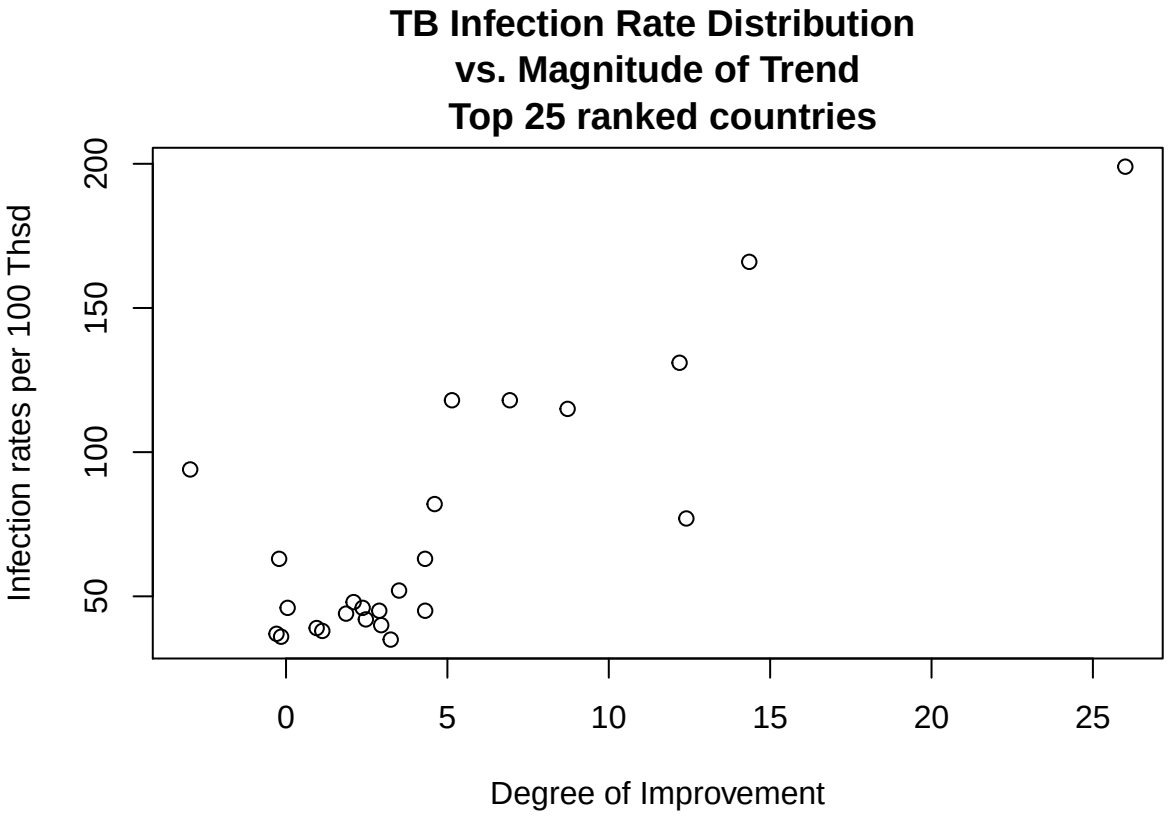
```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## font width unknown for character 0xd
```




```
plot(
    df_tbrates$regr_slope,
    df_tbrates$avgrate_per_100thsd,
    main="TB Infection Rate Distribution \r\n vs. Magnitude of Trend \r\n Top 25 ranked countries",
    xlab="Degree of Improvement",
    ylab="Infection rates per 100 Thsd"
)
```

```
## Warning in title(...): font width unknown for character 0xd
```

```
## Warning in title(...): font width unknown for character 0xd
```


TB Infection Rate Distribution
vs. Magnitude of Trend
Top 25 ranked countries
Infection rates per 100 Thsd
50
100
150
200
0
5
10
15
20
25
Degree of Improvement