Indeed Data

Henrique Almeida
15 de fevereiro de 2019

About Document

An example of treating and ploting data from indeed after scrapping (code in https://github.com/henrique1837/indeed_scraper)

Treating data

```
library(plyr)
library(stringr)
library(ggplot2)
#### Read data ####
files <- Sys.glob("./results/*")
files <- files[which(str_detect(string = files,pattern = "indeed")==TRUE)]
df_t <- data.frame()
for(i in 1:length(files)){
  df <- read.csv(file = files[i],
                 stringsAsFactors = FALSE)
  df <- df[which(!(str_detect(string = df$date,
                              pattern = "30+"))),]
  file_date <- as.Date(str_extract(string = files[i],
                                   pattern = "[[:digit:]]{4}-[[:digit:]]{2}-[[:digit:]]{2}"))
  df$date <- file_date - as.numeric(str_extract(string = df$date,
                                                pattern = "[[:digit:]]+"))
  df$date[which(is.na(df$date))] <- file_date
  if(length(df$country) == 0){
    df$country <- "USA"
  }
  df_t <- rbind(df_t,df[which(!(df$link %in% df_t$link)),])
  #message("Files ",i," - ",length(files))
}
df_t$city <- gsub(pattern = "[[:punct:]].*",
                  replacement = "",
                  x = df_t$location)
df_t$state <- str_extract(string = df_t$location,
                          pattern = "[A-Z]{2}")
df_t$count <- 1
df_agreggated <- ddply(.data = df_t,
                       .variables = .(date),
                       .fun = summarize,
                       totalJobs=sum(count))

df_companies_date <- ddply(.data = df_t,
                           .variables = .(date,company),
```

```
                    .fun = summarize,
                    totalJobs=sum(count))

df_companies <- ddply(.data = df_t,
                    .variables = .(company),
                    .fun = summarize,
                    totalJobs=sum(count))

df_places <- ddply(.data = df_t,
                  .variables = .(city,state,country),
                  .fun = summarize,
                  totalJobs=sum(count))
```

Data Information

Range of dates

```
## [1] "2018-12-29" "2019-02-15"
```

Total Observations

```
## [1] 247
```

Total companies listed

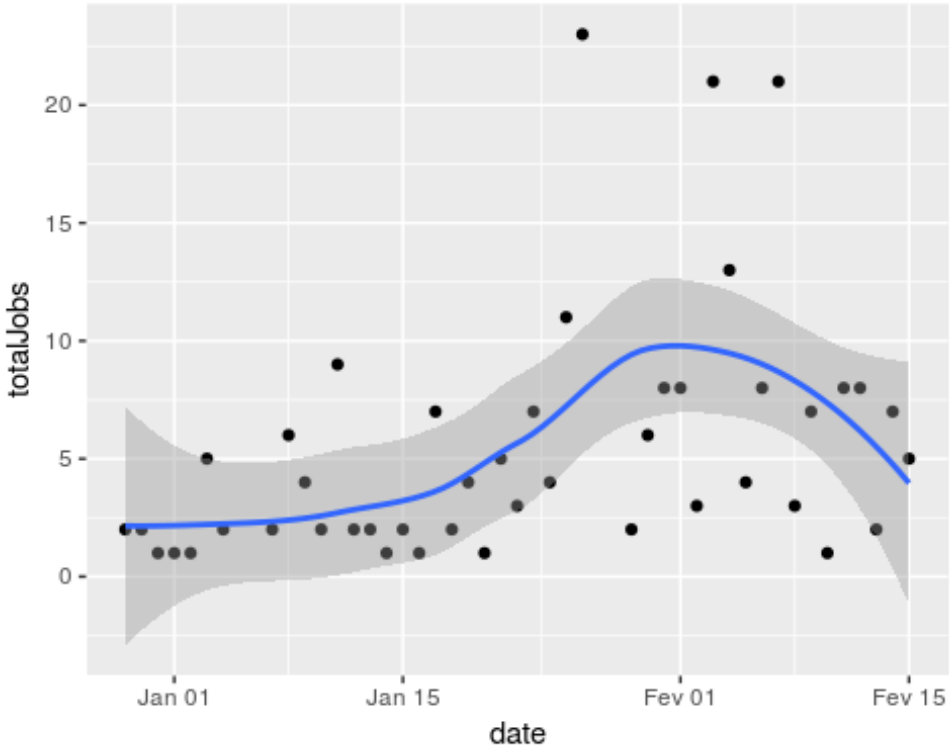
```
## [1] 85
```

Total localizations listed

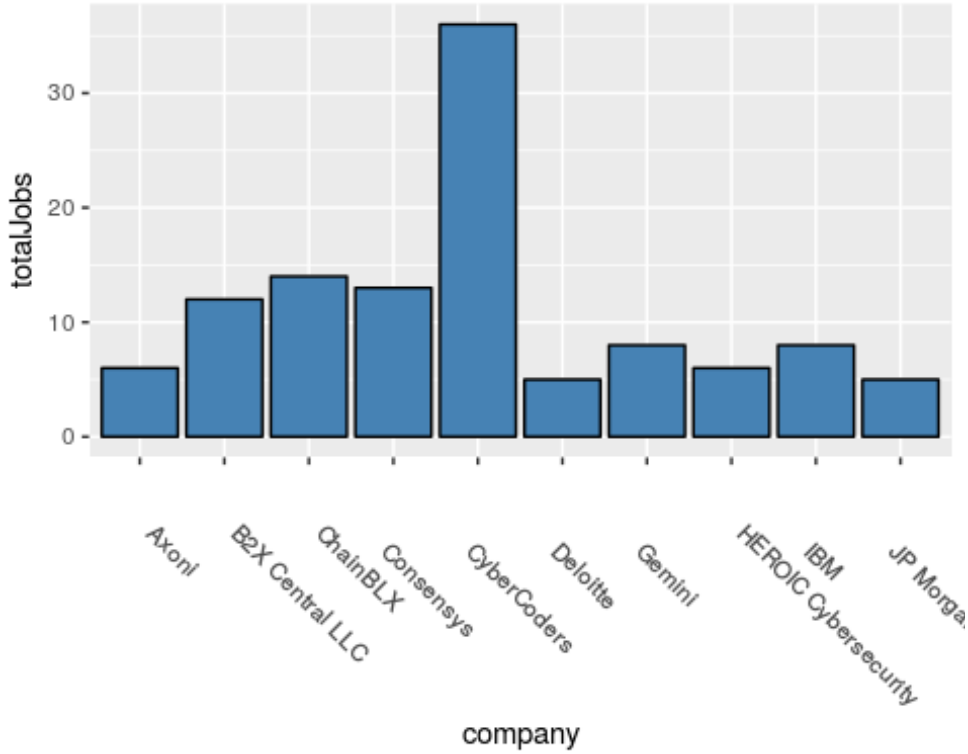
```
## [1] 50
```

Graphics

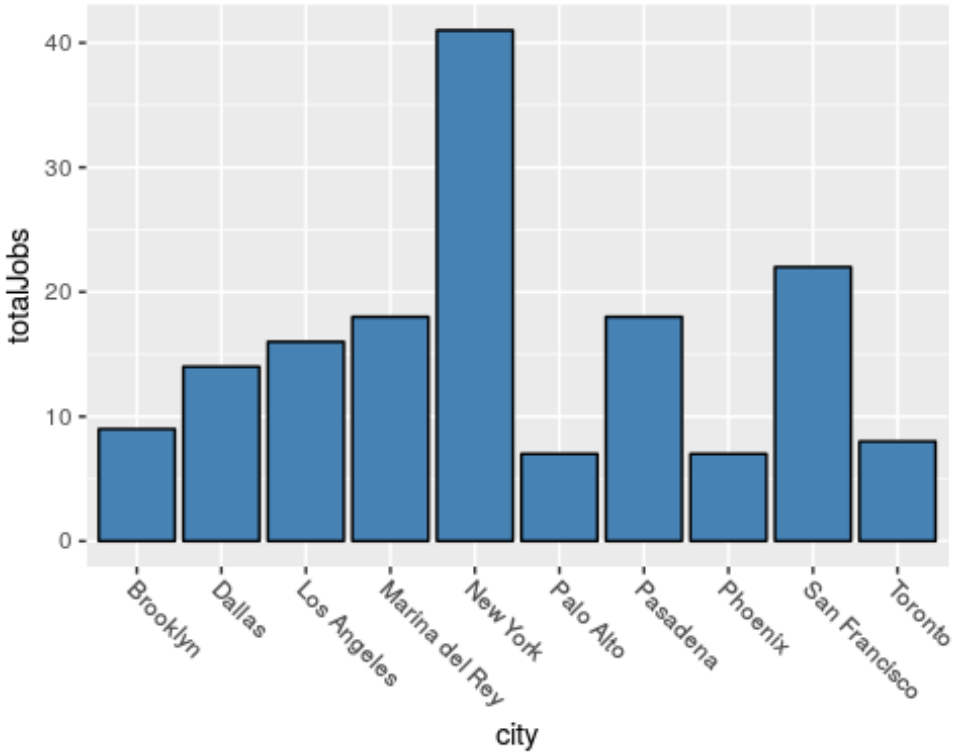
Total Ethereum Jobs posted per date



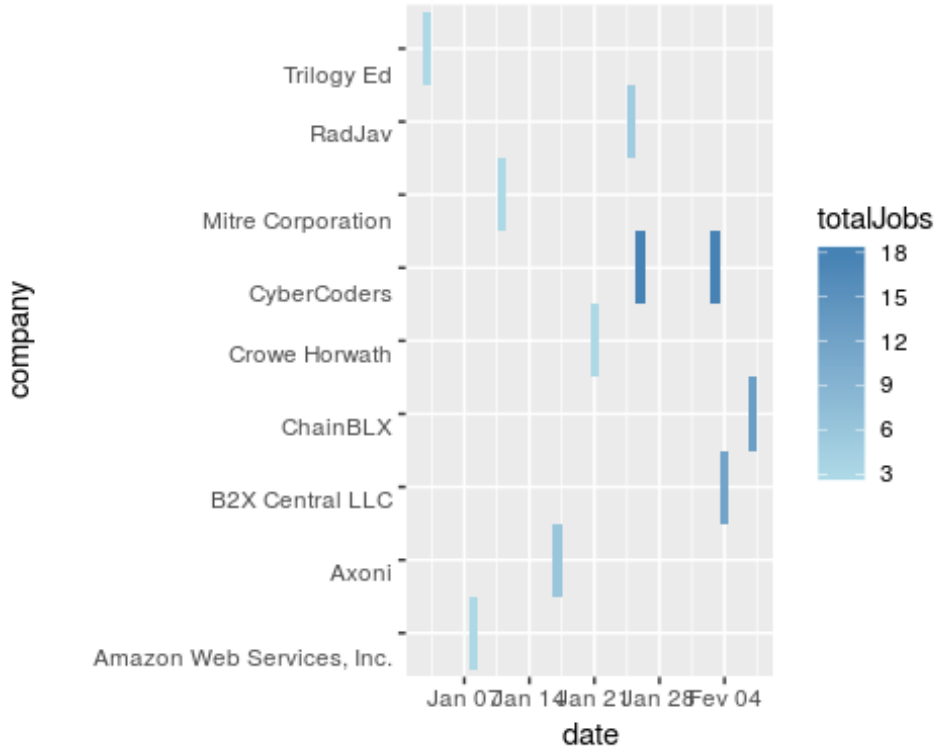
Total Ethereum Jobs posted by top 10 company



Total Ethereum Jobs posted in top 10 localizations



Total Ethereum Jobs posted by top 10 companies per date



Session

```
## R version 3.4.4 (2018-03-15)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 16.04.5 LTS
##
## Matrix products: default
## BLAS: /usr/lib/libblas/libblas.so.3.6.0
## LAPACK: /usr/lib/lapack/liblapack.so.3.6.0
##
## locale:
##  [1] LC_CTYPE=pt_PT.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=pt_BR.UTF-8        LC_COLLATE=pt_PT.UTF-8
##  [5] LC_MONETARY=pt_BR.UTF-8    LC_MESSAGES=pt_PT.UTF-8
##  [7] LC_PAPER=pt_BR.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=pt_BR.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggplot2_3.1.0 stringr_1.3.1 plyr_1.8.4
##
```

```
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.0       bindr_0.1.1      knitr_1.21       magrittr_1.5
##  [5] tidyselect_0.2.5 munsell_0.5.0    colorspace_1.4-0 R6_2.3.0
##  [9] rlang_0.3.1      dplyr_0.7.8      tools_3.4.4      grid_3.4.4
## [13] gtable_0.2.0     xfun_0.4         withr_2.1.2      htmltools_0.3.6
## [17] assertthat_0.2.0 yaml_2.2.0       lazyeval_0.2.1   digest_0.6.18
## [21] tibble_2.0.1     crayon_1.3.4     bindrcpp_0.2.2   purrr_0.3.0
## [25] glue_1.3.0       evaluate_0.12    rmarkdown_1.11   labeling_0.3
## [29] stringi_1.2.4    compiler_3.4.4   pillar_1.3.1     scales_1.0.0
## [33] pkgconfig_2.0.2
```