Lost In The Woods: The Data Bloodhound's Story

Data Analytics User Group

10/11/2019

Data sets: survey and employee database 2019 Q1

If we have the DAWGs survey including the names of those who submitted and their salaries, how much of their time do they spend collecting, finding and sharing data and what's that worth?

Load the two data sets:

```
cat("\014") # clear the console
```

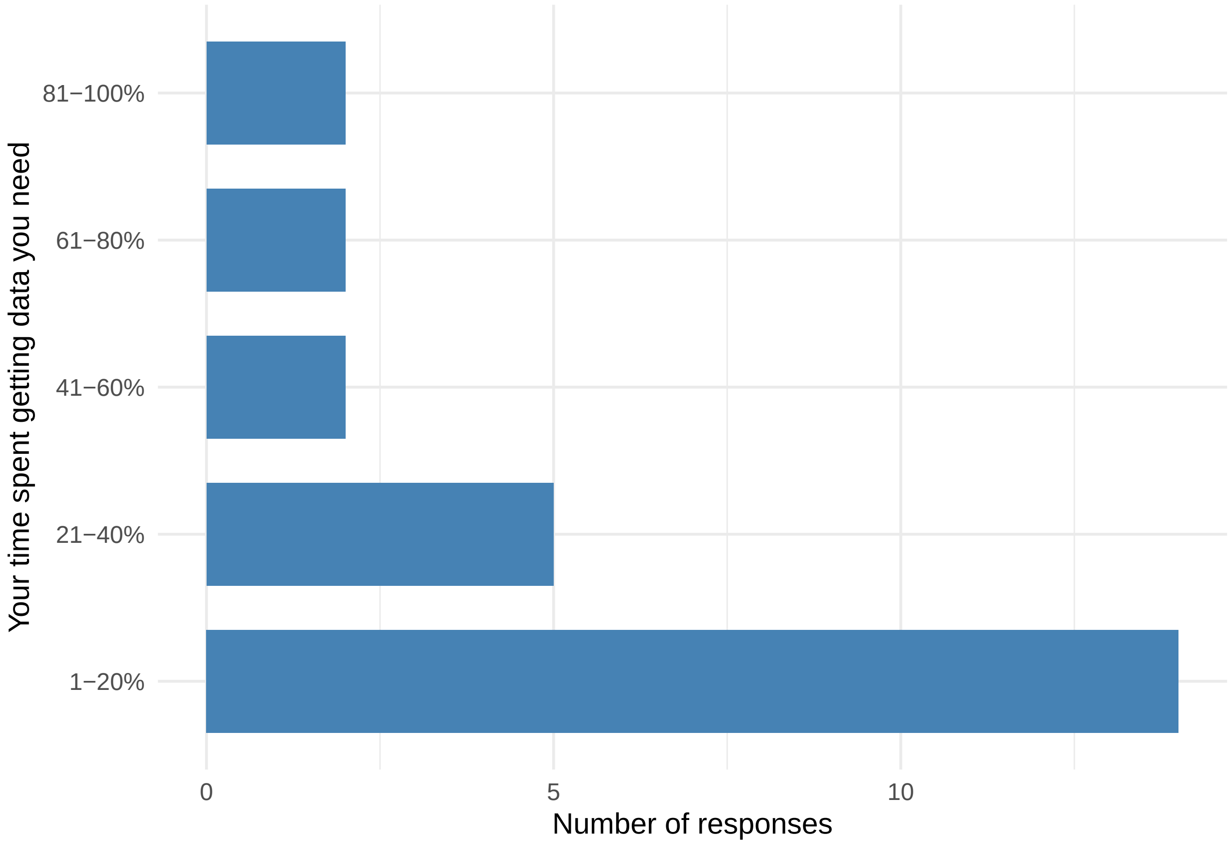
```
rm(list=ls())
options(warn=-1)

library(stringr)
library(ggplot2)

setwd("~/Dropbox/Work and research/Port Authority/roi/roi")
pay = read.csv("./Q1_employee_payroll_2019.csv", skip=1)
poll = read.csv("./Poll.csv")
names(pay) = tolower(names(pay))
names(poll) = tolower(names(poll))
```

The respondents (n=26) reported spending a decent chunk of their collective time searching for data:

```
table(poll$on.average..how.much.time.do.you.spend.getting.the.data.you.need.)
```

```
##
##      1 to 6 months    Less than 1 day  Less than 1 month
##                  4                  4                  9
##   Less than 1 week More than 6 months
##                  7                  2
```

```
ggplot(dat, aes(getting.sharing.data)) + #x=factor(getting.sharing.data)
  geom_bar(stat="count", width=0.7, fill="steelblue") +
  coord_flip() +
  theme_minimal() +
  ylab("Number of responses") +
  xlab("Your time spent getting data you need")
```

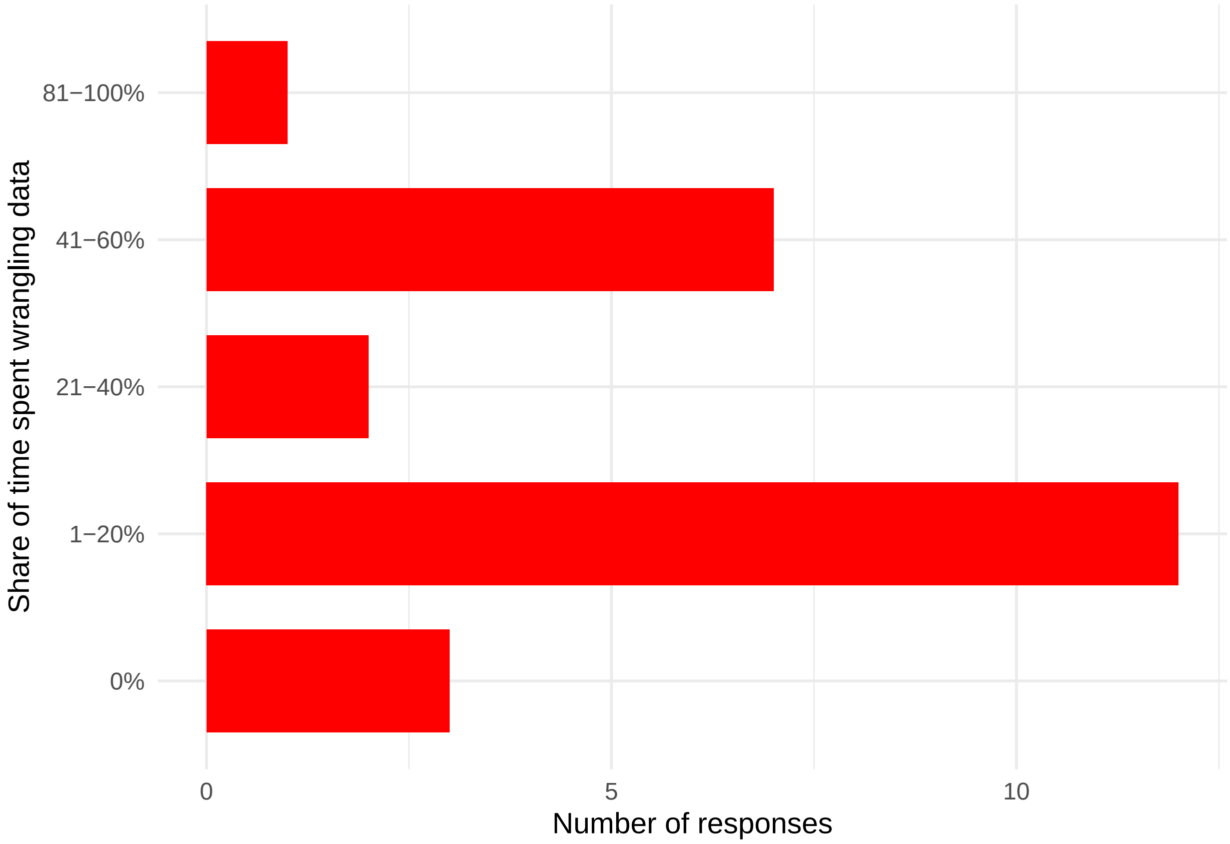


(They also spent a comparable amount of time cleaning that data, including reformatting it from traditional Excel frames to make it machine-readable), and then trying to understand the data and then relay that understanding to others:

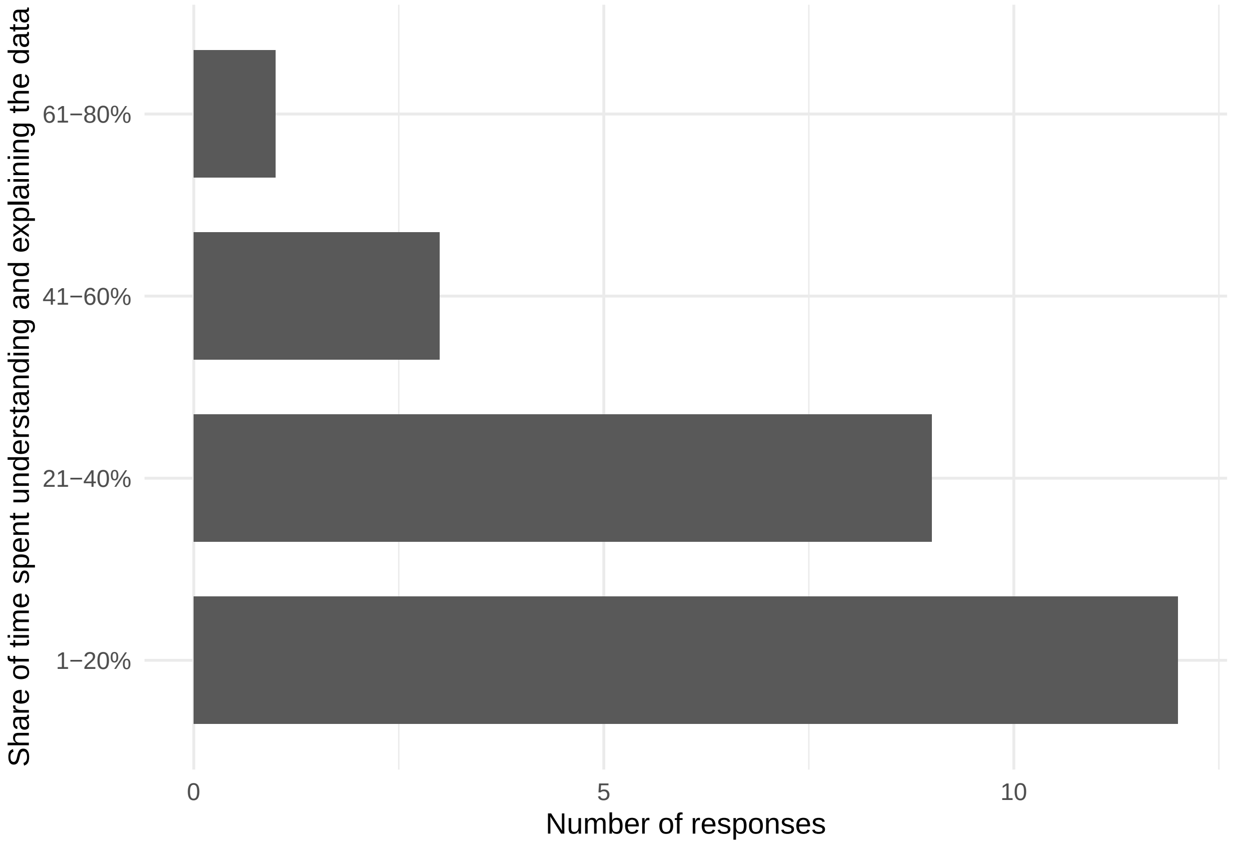
```
names(poll)
```

```
##  [1] "id"
##  [2] "start.time"
##  [3] "completion.time"
##  [4] "email"
##  [5] "name"
##  [6] "what.is.your.connection.to.port.authority.data...select.all.that.apply."
##  [7] "how.frequently.do.you.have.problems.finding.specific.data.that.you.need.to.perform.your.job."
##  [8] "on.average..how.much.time.do.you.spend.getting.the.data.you.need."
##  [9] "how.frequently.do.you.have.problems.understanding.and.or.cleaning.up.data.for.use."
## [10] "on.average.how.much.time.do.you.spend.cleaning.up.the.data.you.need."
## [11] "are.you.responsible.for.distributing.the.data.with.which.you.work."
## [12] "how.easy.is.it.for.others.to.acquire.the.data.you.manage..oversee.or..would.like.to..work.with
## [13] "getting.sharing.data"
## [14] "understanding.explaining.data"
## [15] "cleaning.data"
## [16] "using.analyzing.data"
## [17] "saving.cleaned.data.for.future.use"
## [18] "how.long.have.you.worked.at.or.with.the.port.authority."
## [19] "first"
## [20] "last"
## [21] "full"
```

```r
ggplot(dat, aes(cleaning.data)) + #x=factor(getting.sharing.data)
  geom_bar(stat="count", width=0.7, fill="red") +
  coord_flip() +
  theme_minimal() +
  ylab("Number of responses") +
  xlab("Share of time spent wrangling data") +
  scale_y_continuous(breaks=c(0,5,10,15))
```



```r
ggplot(dat, aes(understanding.explaining.data)) + #x=factor(getting.sharing.data)
  geom_bar(stat="count", width=0.7) +#, fill="orange") +
  coord_flip() +
  theme_minimal() +
  ylab("Number of responses") +
  xlab("Share of time spent understanding and explaining the data") +
  scale_y_continuous(breaks=c(0,5,10,15)) +
  scale_fill_manual(values = alpha(c("red"), .3))
```



So what's that cost (reflected in dollars), assuming we can value their time at roughly their salary?

For amount of our time spent getting data, be conservative and use numbers at the lower bound for each category. (If we were to add up all the how-much-of-your-time-is-spent categories, most respondents' summed responses would be over 100% - so take the responses as indicators, not hard numbers.)

Also add a round 50 percent to their salary to account for benefits.

```r
dat$trouble.getting.sharing = ifelse(dat$getting.sharing.data=="1-20%",.05,
                              ifelse(dat$getting.sharing.data=="21-40%",.21,
                              ifelse(dat$getting.sharing.data=="41-60%",.41,
                              ifelse(dat$getting.sharing.data=="61-80%",.61,.81))))
dat$loss = dat$trouble.getting.sharing * (dat$salary * 1.5)
sum(dat$loss) * (26/25)
```

```
## [1] 816934
```

So of 26 people who responded we spend around $800,000 just looking and getting data. Respondents' average salary is in the ballpark regarding the average salary for the agency, without doing any filtering and cleaning :

```r
summary(dat$salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   64272   88322  101998  106680  122512  154908
```

```r
pay.clean = na.omit(pay)
summary(pay.clean$salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17009   73190   94510   96370  115000  299520
```

… although the full agency salary list obviously includes hundreds of people who aren't here for a full year or are somehow otherwise charged for a fraction of a year of work. So the average, after removing those people, would be about the same.

So if the average person at the agency uses data only 10 percent as frequently as we do, how much does the agency spend looking for and getting data?

```
(sum(pay.clean$salary)*1.5)*mean(dat$trouble.getting.sharing)*.1
```

```
## [1] 24312166
```

```
options(warn=0)
```

Ballpark - somewhere north of $20 million a year.