# 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
```

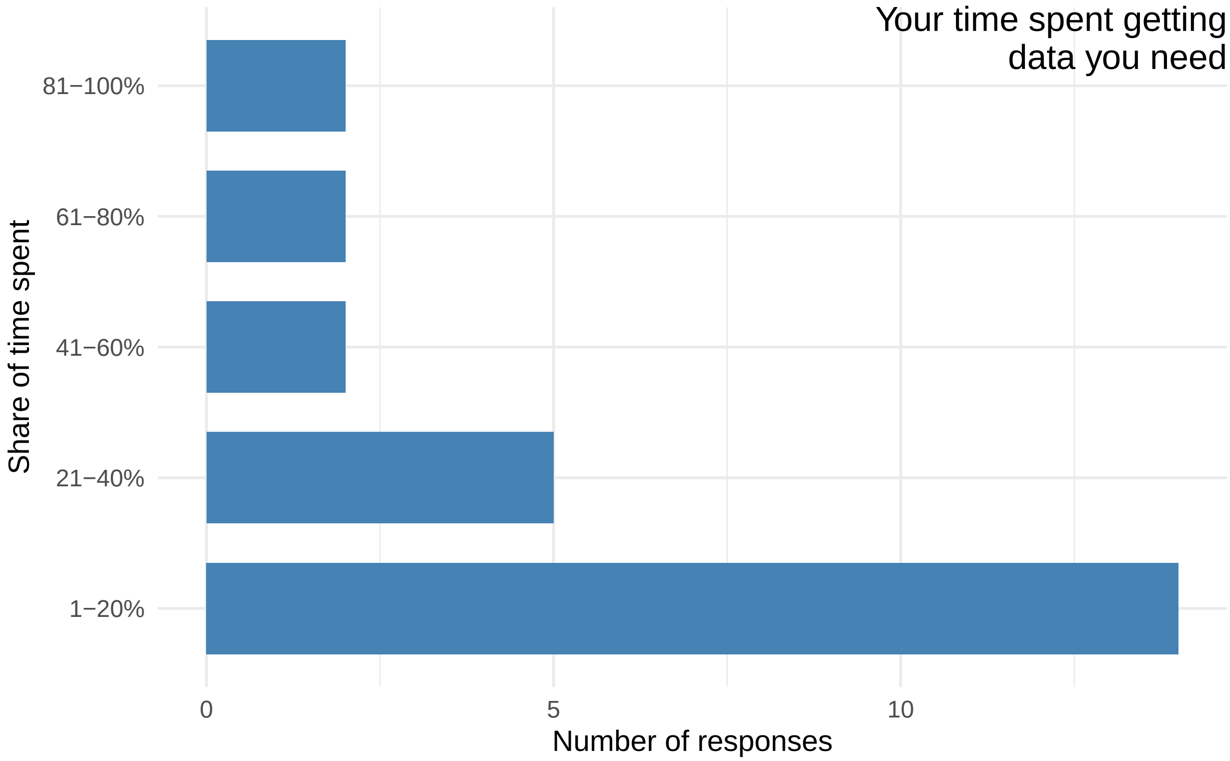
```
rm(list=ls())
options(warn=-1)

library(stringr)
library(ggplot2)
library(reshape2)
library(gridExtra)
library(grid)

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:

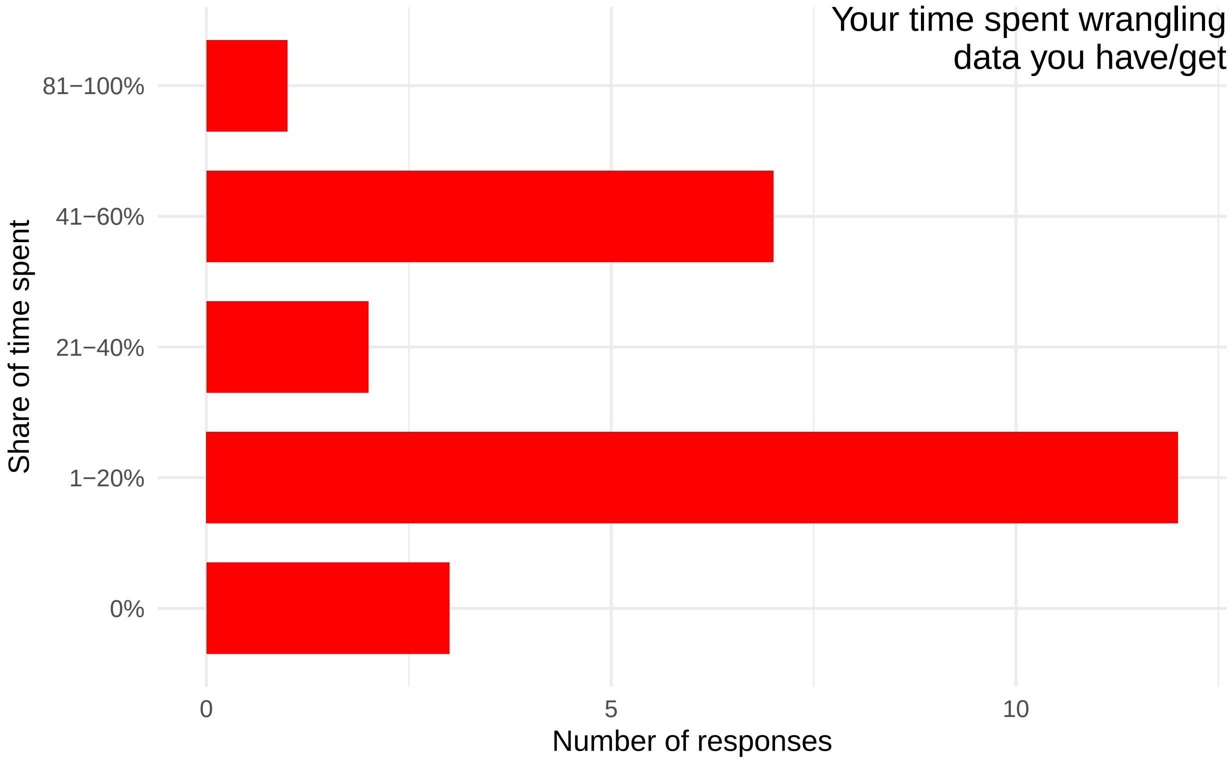
```
table(poll$on.average..how.much.time.do.you.spend.getting.the.data.you.need.)
```

```
##
##      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
```

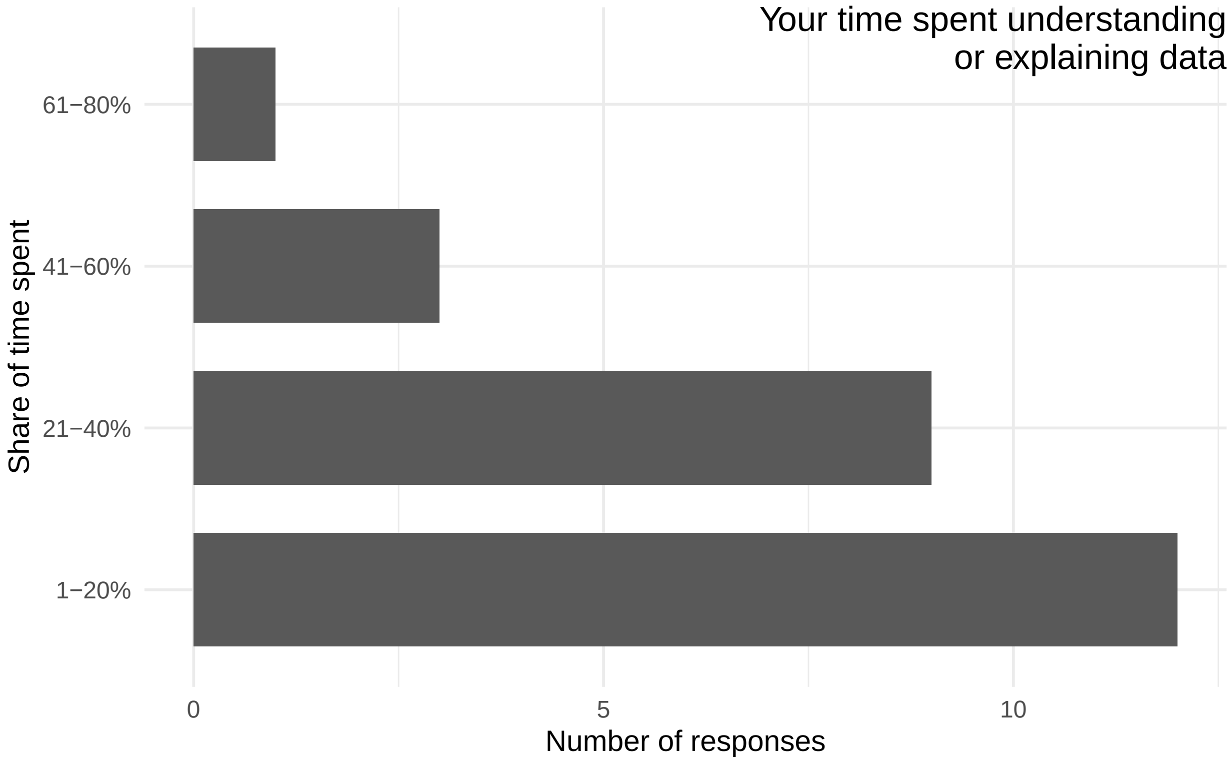
```
ggplot(dat, aes(getting.sharing.data)) + #x=factor(getting.sharing.data)
  geom_bar(stat="count", width=0.7, fill="steelblue") +
  coord_flip() +
  theme_minimal() +
  ylab("Number of responses") +
  xlab("Share of time spent") +
  ggtitle("Your time spent getting\ndata you need") +
  theme(plot.title = element_text(vjust = - 10, hjust = 1))
```



Analysis can also require significant cleaning . . .



. . . and research to understand the data and relay that understanding to others.



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

For time spent getting data, be conservative and use numbers at or near the lower bound of 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.) When someone reported spending between 1 percent and 20 percent of her time getting data, assume 5 percent; when someone reported psending between 21 percent and 40 percent of her time getting data, go with 21 percent, and stick to the bottom of each bin moving upward.

Also add a round 40 percent to their salary to account for benefits.

```
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 * 1.4)
sum(dat$loss) * (26/25)
```

```
## [1] 762471.7
```

The 26 people who responded spend around $800,000 of the agency's money annually just looking around for, identifying and getting data sets. This can include external data sets but also data internal to the agency but housed outside an analyst's immediate reach.

(By the way, the 26 respondents' average salary is generally close to the average salary for the agency, without doing any filtering and cleaning …

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

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

```
summary(pay$salary)
```

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

. . . particularly recognizing that 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 averages, after removing those people, would be about the same.)

One more assumption: if the average person at the agency uses data only around 10 percent as frequently as the 26 respondents do, how much does the agency spend looking for and getting data in total?

```
a = (sum(pay$salary)*1.4)*mean(dat$trouble.getting.sharing)*.1
```

Ballpark - somewhere north of $20 million a year.

What about the challenge of managing or interpreting and explaining data?

```
## [1] 665885.9
```

```
## [1] 1230856
```

```
b = (sum(pay$salary)*1.4)*mean(dat$understanding.explaining)*.1
c = (sum(pay$salary)*1.4)*mean(dat$cleaning)*.1
options(warn=0)
```

```
#d = head(iris[,1:3])
#grid.table(d)
```

The assumptions taken above are meant to aid in a first, high-level attempt at contextualizing the poll results Some of those assumptions would undoubtedly change following a bit more thought, but hopefully this offers value as a starting point regarding the potential value to the agency of rationalizing data management practices and advancing its data governance framework.