Eating Your Own Dogfood

Josh Laurito

CUNY IS 608 Lecture 2

Eating Your Own Dogfood

Visualization in the Feedback Loop

==========

Today's To-Dos

- About you
- Review last week's homework
- Exploratory data analysis
- ggplot2
- BigVis
- devtools
- This week's homework

==========

About you

Thanks for filling out the initial survey for the class

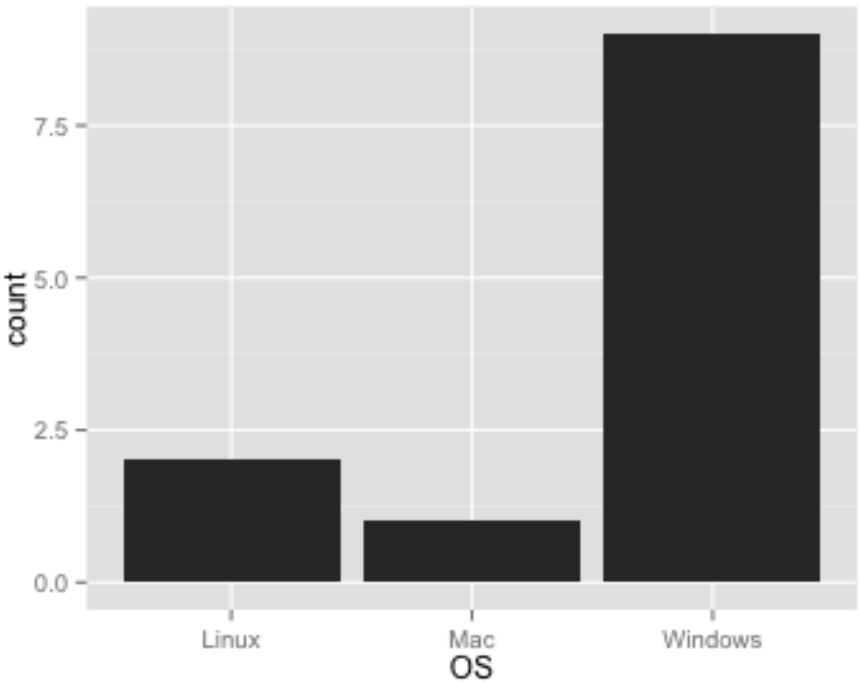
(i learned a lot)

==========

About you

You guys really like Windows!

```
library(ggplot2)
class <- read.csv("../datasets/student_survey.csv")
ggplot(aes(x = OS), data = class) + geom_histogram()
```

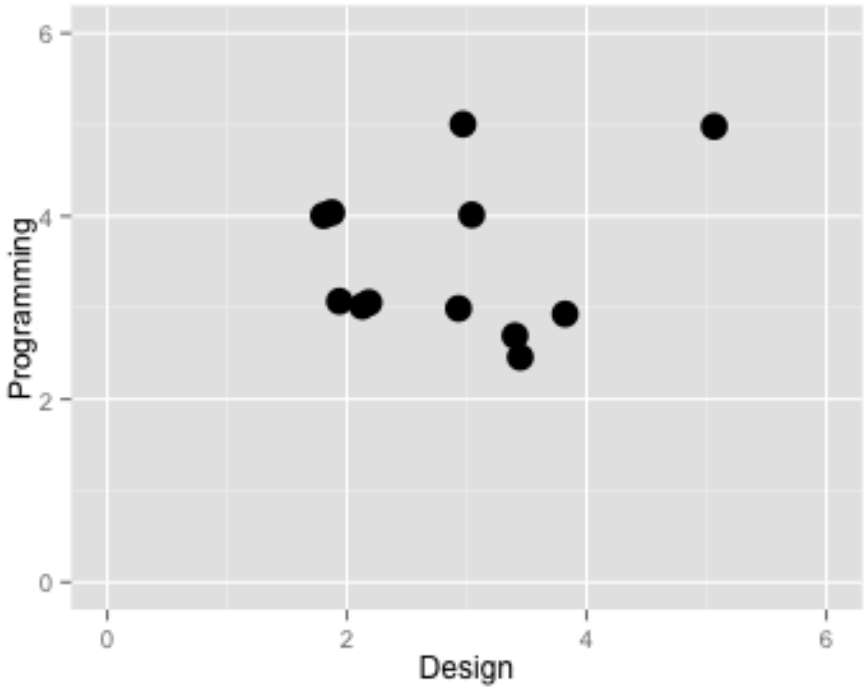


==========

About you

You're pretty confident

```
p <- ggplot(class, aes(x = Design, y = Programming))
p + geom_point(position = "jitter", size = 5) + ylim(0, 6) + xlim(0, 6)
```



==========

Last Week's Homework

- Let's walk through it
- Gather your Data

```
library(plyr)
inc <- read.csv("/Users/JL/Dropbox/CUNY/CUNY_IS608/lecture1/data/inc5000_data.csv",
    header = TRUE)
head(inc, 2)
```

```
##   Rank                  Name Growth_Rate   Revenue
## 1    1                  Fuhu       421.5 117900000
## 2    2 FederalConference.com       248.3  49600000
##                        Industry Employees       City State
## 1 Consumer Products & Services       104 El Segundo    CA
## 2          Government Services        51   Dumfries    VA
```

==========

Last Week's Homework

- Investigate

```
summary(inc[, c(3:6, 8)])
```

```
##   Growth_Rate       Revenue                                Industry
##  Min.   :  0.3   Min.   :2.00e+06   IT Services                 : 733
##  1st Qu.:  0.8   1st Qu.:5.10e+06   Business Products & Services: 482
##  Median :  1.4   Median :1.09e+07   Advertising & Marketing     : 471
##  Mean   :  4.6   Mean   :4.82e+07   Health                      : 355
##  3rd Qu.:  3.3   3rd Qu.:2.86e+07   Software                    : 342
##  Max.   :421.5   Max.   :1.01e+10   Financial Services          : 260
##                                     (Other)                     :2358
##    Employees         State
##  Min.   :    1   CA     : 701
##  1st Qu.:   25   TX     : 387
##  Median :   53   NY     : 311
##  Mean   :  233   VA     : 283
##  3rd Qu.:  132   FL     : 282
##  Max.   :66803   IL     : 273
##  NA's   :12      (Other):2764
```

==========

Last Week's Homework

- For this analysis, remove NULL values

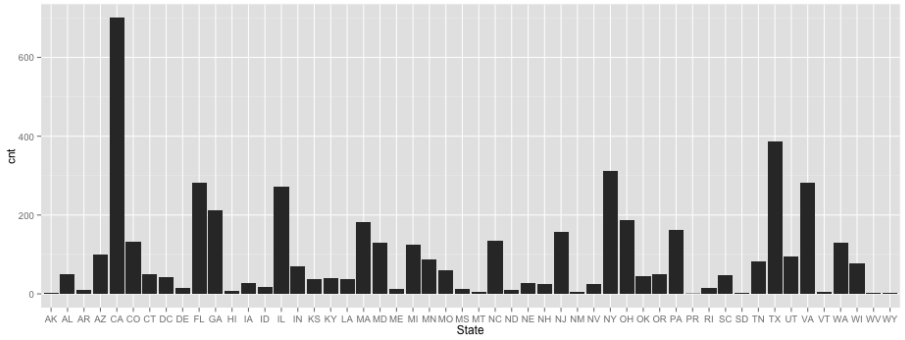
```
all_inc <- inc[complete.cases(inc)==TRUE,]
```

==========

Last Week's Homework

- Get counts by State

```
cnt <- ddply(all_inc, .(State), summarize, cnt = length(State))
p <- ggplot(cnt, aes(x = State, y = cnt)) + geom_bar(stat = "identity")
p
```



==========

Last Week's Homework

- To switch to horizontal bars, use `coord_flip()`

```
p <- ggplot(cnt, aes(x=State, y=cnt)) + geom_bar(stat='identity')
p + coord_flip()
```

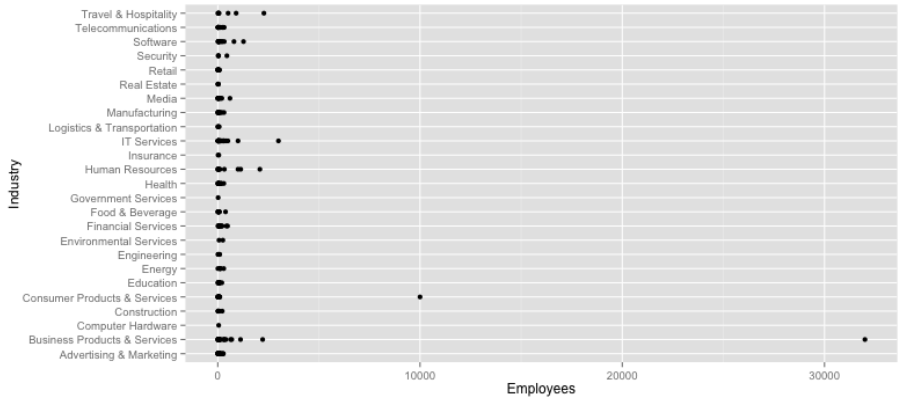
- To show tabular, quantitative data, line or scatter plots are good

==========

Last Week's Homework

- New York is the #3 State, so let's dig in

```
ny <- subset(all_inc, State == "NY")
p <- ggplot(ny, aes(x = Industry, y = Employees)) + geom_point()
p + coord_flip()
```



==========

Last Week's Homework

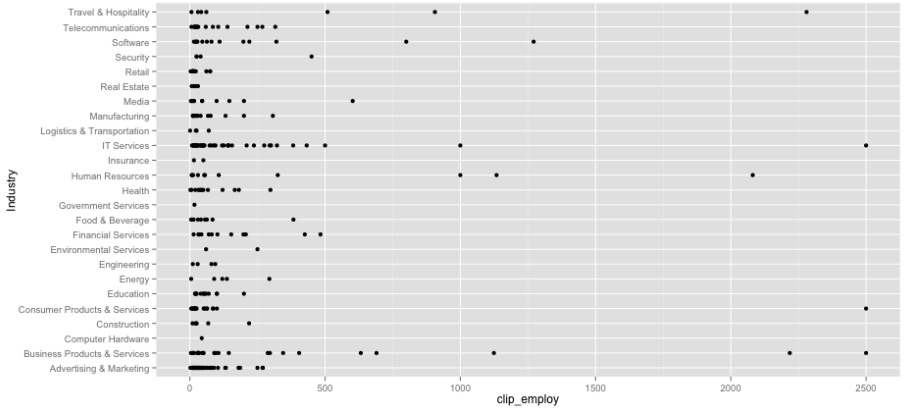
- Serious outlier issue: how do we handle?
- Do we include, make a note (annotate) or ignore?
- Do we care more about the **mean** or **median**?
- If we care more about the **median**, outliers are distractions
- 'Winsorize' Data

```
winsor <- function(x, bot, top) {
    return(min(top, max(x, bot)))
}
ny$clip_employ <- sapply(ny$Employees, winsor, bot = 0, top = 2500)
p3 <- ggplot(ny, aes(x = Industry, y = clip_employ))
```

==========

Last Week's Homework

```
p3 + geom_point() + coord_flip()
```

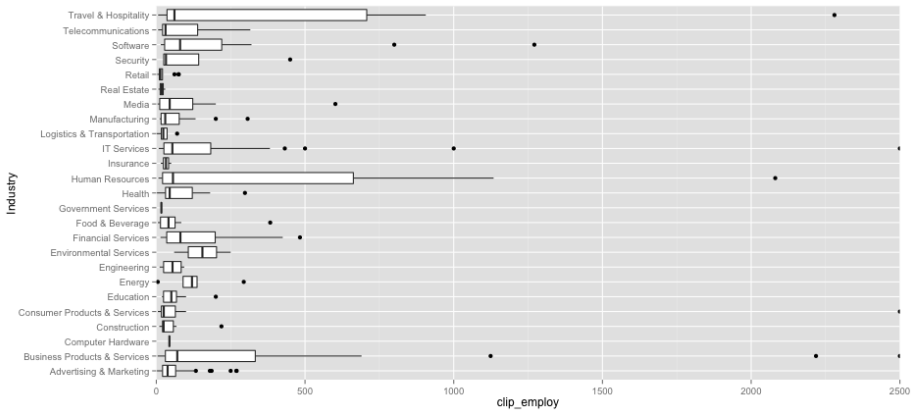


==========

Last Week's Homework

- A relative of the scatter plot is the box plot

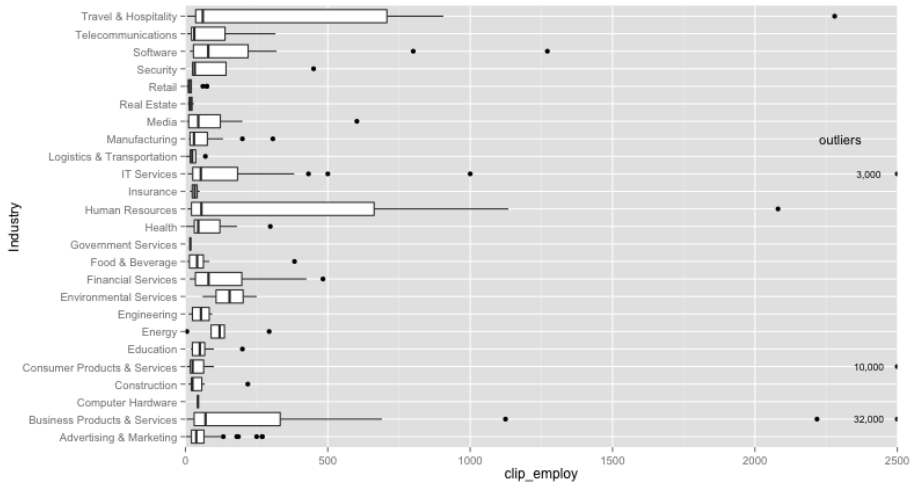
```
p3 + geom_boxplot() + coord_flip(ylim = c(0, 2500))
```



==========

Last Week's Homework - Marking Outliers

```
p3 + geom_boxplot() + coord_flip(ylim = c(0, 2500)) + annotate("text", label = c("outliers",
    "3,000", "10,000", "32,000"), x = c(18, 16, 5, 2), y = c(2300, 2400, 2400,
    2400), size = c(4, 3, 3, 3))
```



==========

Last Week's Homework

- There are other ways to show variance
- But we need to create averages

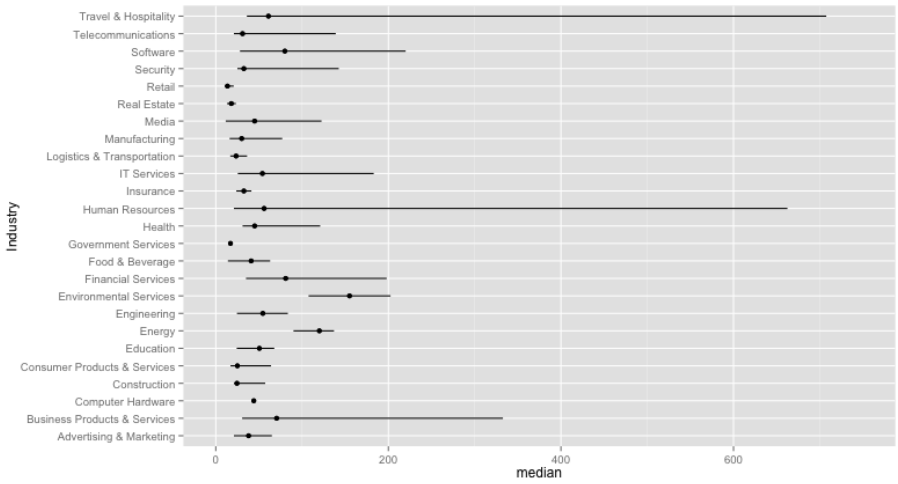
```
ny_ave <- ddply(ny, .(Industry), summarize,

                mean <- mean(Employees),

                sd <- sd(Employees),

                median <- median(clip_employ),

                lower <- quantile(clip_employ)[2],

                upper <- quantile(clip_employ)[4]

               )
names(ny_ave) <- c('Industry', 'mean', 'sd', 'median', 'lower', 'upper')

head(ny_ave,2)
```

```
##                         Industry    mean      sd median lower upper
## 1      Advertising & Marketing   58.44   62.23   38.0  21.0  65.0
## 2 Business Products & Services 1492.46 6240.71   70.5  30.5 332.8
```

==========

Last Week's Homework - Point ranges

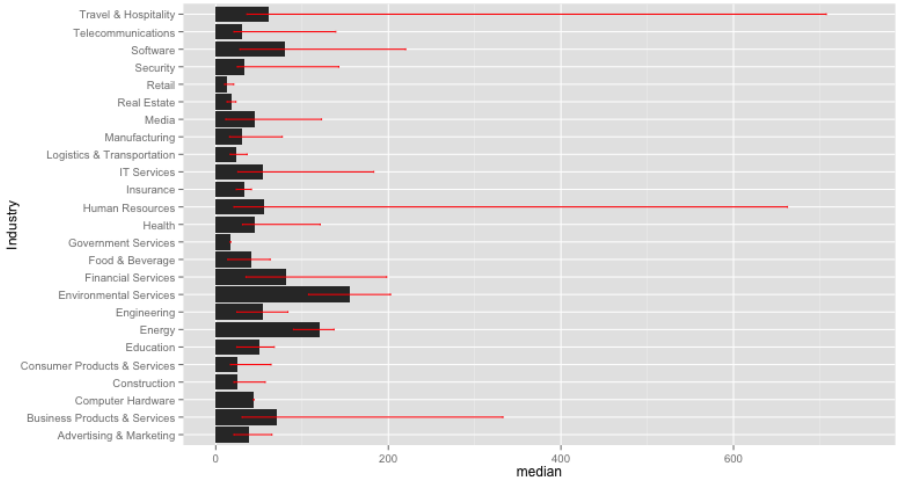
```
p4 <- ggplot(ny_ave, aes(x = Industry, y = median)) + geom_point()
p4 <- p4 + geom_pointrange(ymin = ny_ave$lower, ymax = ny_ave$upper)
p4 + ylim(c(0, 750)) + coord_flip()
```



==========

Last Week's Homework - Error bars

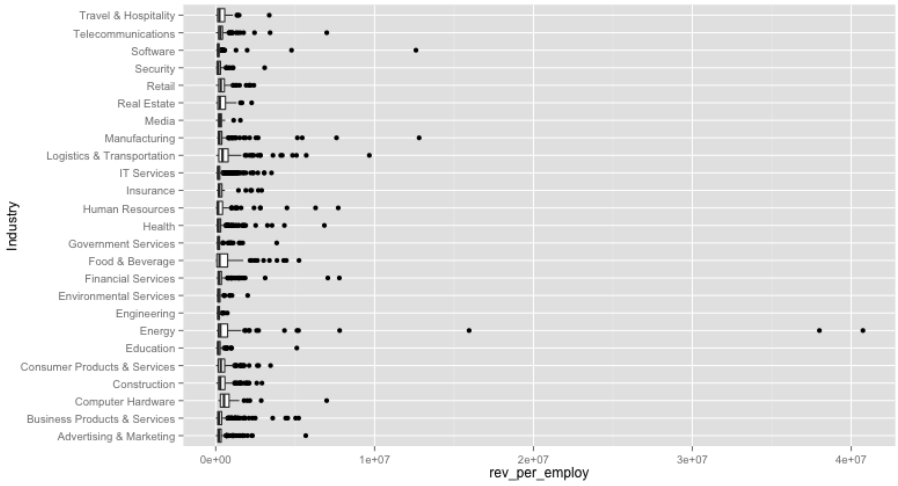
```
p5 <- ggplot(ny_ave, aes(x = Industry, y = median)) + geom_bar(stat = "identity")
p5 <- p5 + geom_errorbar(ymin = ny_ave$lower, ymax = ny_ave$upper, width = 0.1,
    color = "red")
p5 + ylim(c(0, 750)) + coord_flip()
```



==========

Last Week's Homework - Investors care about the money

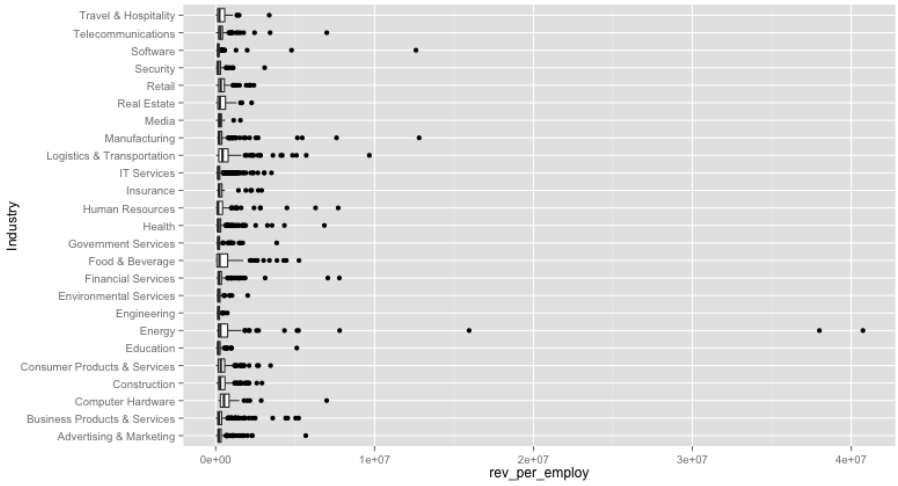
```
all_inc$rev_per_employ <- all_inc$Revenue/all_inc$Employees
p6 <- ggplot(all_inc, aes(x = Industry, y = rev_per_employ))
p6 + geom_boxplot() + coord_flip()
```



==========

Last Week's Homework - Revenue per Employee

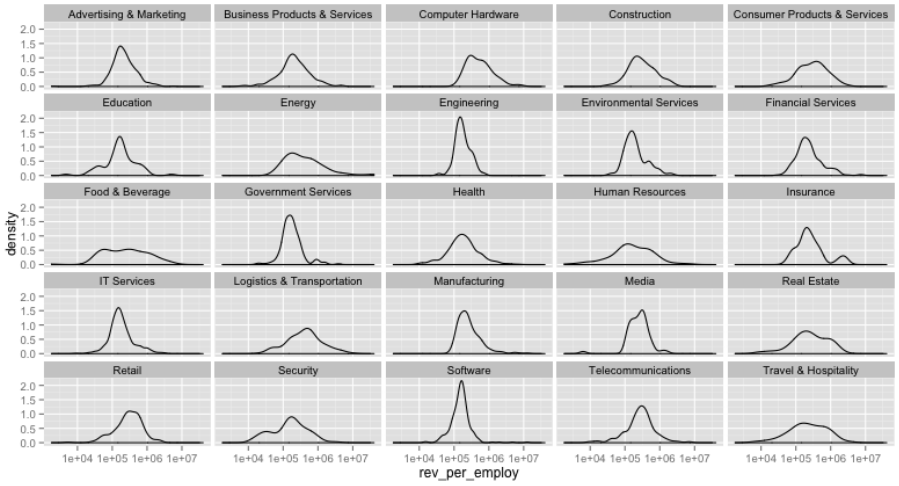
```
all_inc$rev_per_employ <- all_inc$Revenue/all_inc$Employees
p6 <- ggplot(all_inc, aes(x = Industry, y = rev_per_employ))
p6 + geom_boxplot() + coord_flip()
```



==========

Last Week's Homework - Likely Outcomes and Distributions

```
p7 <- ggplot(all_inc, aes(x = rev_per_employ))
p7 <- p7 + geom_density() + facet_wrap(~Industry)
p7 + scale_x_log10(breaks = c(10000, 1e+05, 1e+06, 1e+07))
```



==========

Exploratory Data Analysis

- A great way to test your visualizations - do *you* find them useful?
- We basically just did it!
- Should always use to understand your data set

==========

ggplot2

- Most popular visualization framework
- Developed by Hadley Wickham
- Easy to learn, supports lots of features
- Being ported to other languages
- We will focus on these design patterns throughout the semester

==========

BigVis

- Also written by Hadley Wickham
- Geared towards larger data sets
- **Not on CRAN**

==========

devtools

- In order to install BigVis, you need to install **devtools**
- Go to http://www.rstudio.com/projects/devtools/
- Depending on your operating system, go to the Rtools/Xcode/r-devel page
- Follow the instructions **carefully**
- Once devtools is installed, follow the directions at https://github.com/hadley/bigvis

==========

This week's homework

- We will be working with the set of all NYC tax lot data
- Go to http://www.nyc.gov/html/dcp/html/bytes/applbyte.shtml#pluto
- Download the **PLUTO** data set
- The data is in separate files for each boro: you will need to combine

==========

This week's homework - hints

- You don't need every column of data in your combined file
- If you can't combine files, do the homework with Manhattan-only data
- If you can't install devtools/BigVis, try again
- If you can't install devtools/BigVis *after an hour*, email me

==========

That's it

- This presentation will be on the GitHub page for reference
- Good luck! Any questions?

==========