# DATA 606 Data Project Proposal

Guillermo Schneider

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3

## Warning: package 'ggplot2' was built under R version 4.3.3

## Warning: package 'tibble' was built under R version 4.3.3

## Warning: package 'tidyr' was built under R version 4.3.3

## Warning: package 'readr' was built under R version 4.3.3

## Warning: package 'purrr' was built under R version 4.3.3

## Warning: package 'forcats' was built under R version 4.3.3

## Warning: package 'lubridate' was built under R version 4.3.3

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
library(dplyr)
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 4.3.3

##
## Attaching package: 'Hmisc'
##
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
##
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

### Part 1 - Introduction

I work as a Data Analyst at SEIU 32BJ, a labor union. We represent ~80,000 Janitors, Cleaners, Doormen, Security Guards, Window Cleaners, and other building workers in NYC. Our Union Representatives often organize group meetings with our members at their worksite (called Worksite Meetings). By using our new app to scan member IDs, they've been gathering more accurate meeting attendance data for the past year. We're curious if we can use this to can help predict expected turnout to Worksite Meetings.

Research Question Do buildings' division, member count, and ADF (political) contributions influence turnout to worksite meetings?

Turnout ~ Division + Members + ADF

### Part 2 - Data

Data set with identifying info of reps or employers or addresses removed

```
# load data
WorksiteMeetings <- read.csv("WorksiteMeetingsNoIndentifyingInfo.csv")
```

Correct date format:

```
WorksiteMeetings$datecreated <- mdy(WorksiteMeetings$datecreated)
```

Change from char to numeric:

```
WorksiteMeetings$memberCount <- as.numeric(WorksiteMeetings$memberCount)
```

```
## Warning: NAs introduced by coercion
```

```
WorksiteMeetings$ADFcount <- as.numeric(WorksiteMeetings$ADFcount)
```

```
## Warning: NAs introduced by coercion
```

```
WorksiteMeetings$turnoutcount <- as.numeric(WorksiteMeetings$turnoutcount)
```

change the NA's in ADF to zeros:

```
WorksiteMeetings$ADFcount <- WorksiteMeetings$ADFcount %>% replace(is.na(.), 0)
```

New fields for turnout percentage and ADF percentage of building:

```
WorksiteMeetings$ADFPercentage <- (WorksiteMeetings$ADFcount / WorksiteMeetings$memberCount)
WorksiteMeetings$TurnoutPercentage <- (WorksiteMeetings$turnoutcount / WorksiteMeetings$memberCount)
```

Replace those >100% with 100%

```
WorksiteMeetings$TurnoutPercentage <-replace(WorksiteMeetings$TurnoutPercentage, WorksiteMeetings$Turnou
```

Remove the NY Security, there are too many huge memberCount accounts of grouped locations. this is a known issue, os im unsurprised to see it here. It makes it really hard to figure out which buildings our members actually work at:

```
WorksiteMeetings <- WorksiteMeetings %>%  filter(divisionName !='NY Security')
```

### Part 3 - Exploratory data analysis

**divisionName:** Residential and Commercial buildings are our biggest divisions, this makes sense to me:

```
table(WorksiteMeetings$divisionName)
```

```
##
##  NY Commercial NY Residential     NY Schools
##            228            249              72
```

**MemberCount** is the roster size at that building. We represent lots of single doorman buildings, and smaller buildings with just a few cleaning staff, but we do have some larger 100+ person buildings

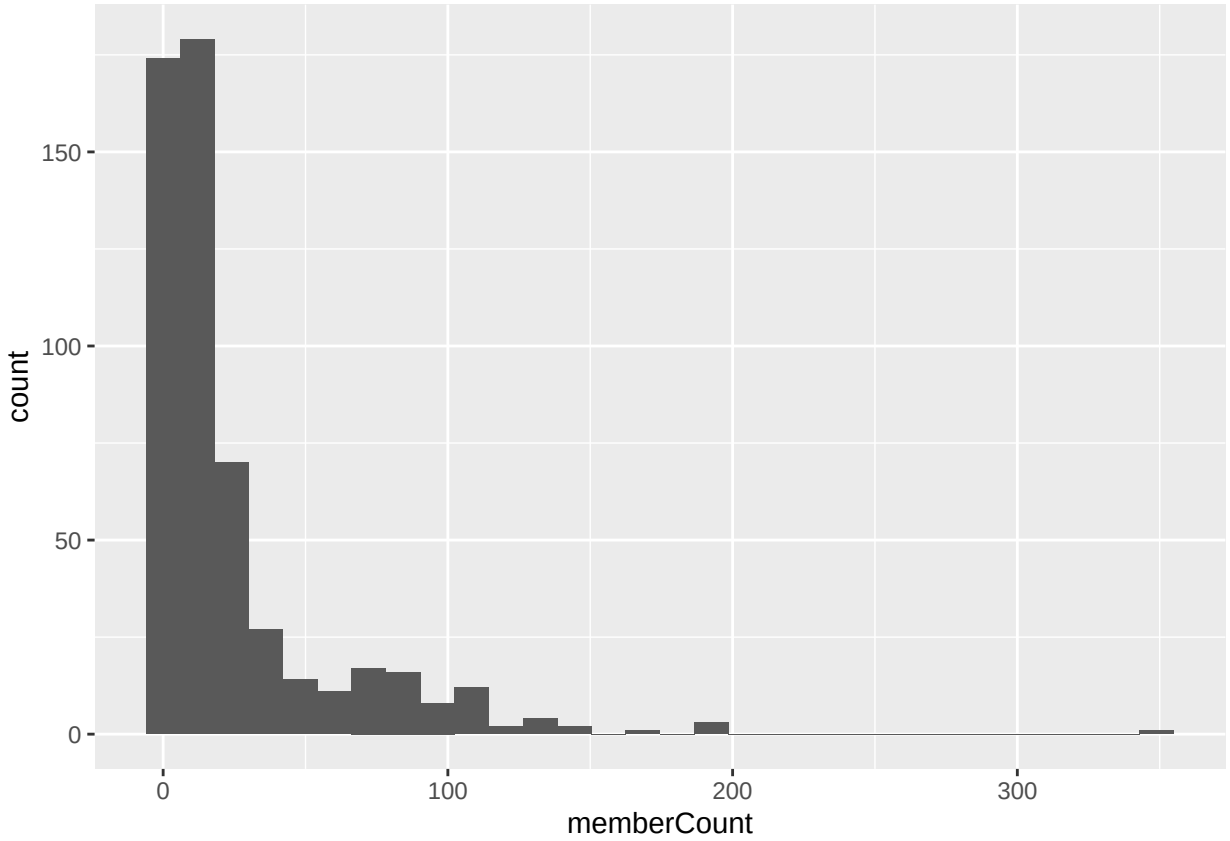
```
summary(WorksiteMeetings$memberCount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    5.00   10.00   24.46   25.00  350.00       8
```

```
ggplot(WorksiteMeetings, aes(x=memberCount)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 8 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

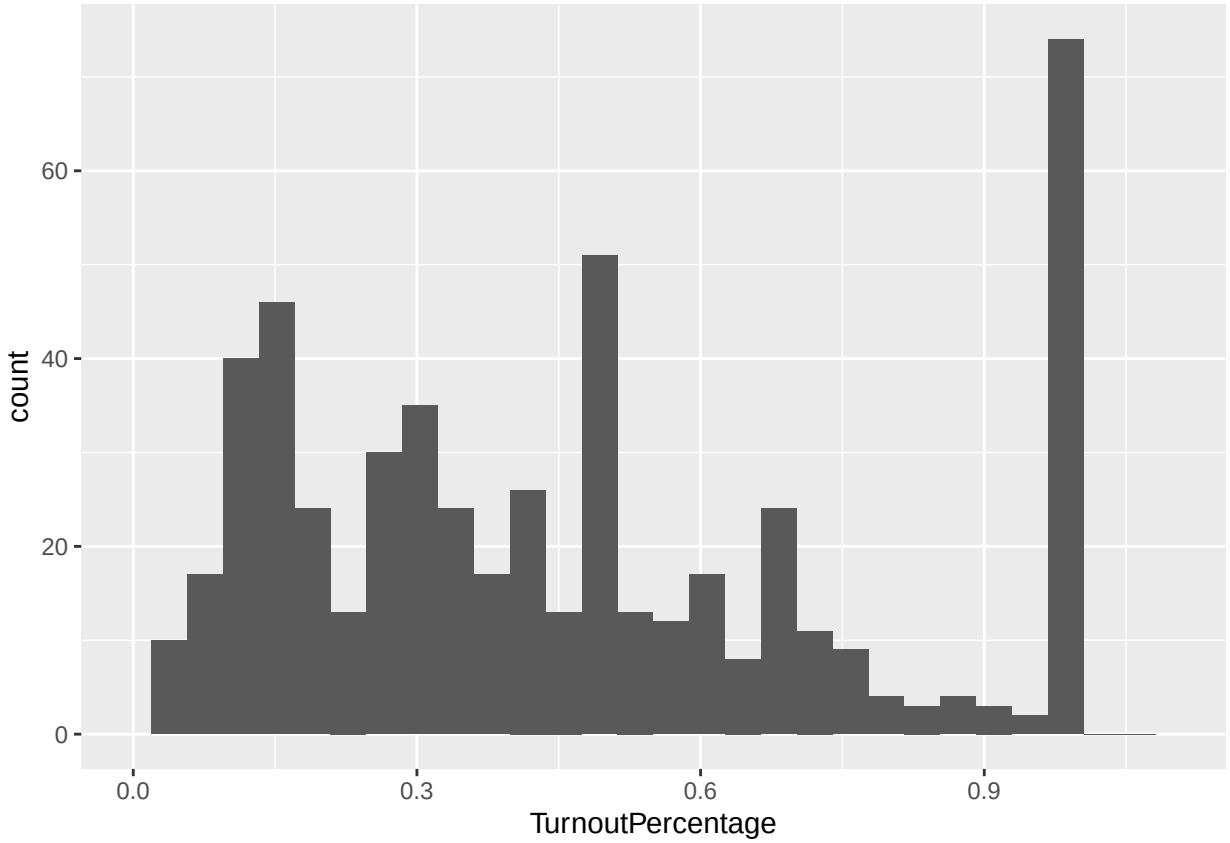



**TurnoutPercentage:** percent of building that went to an event. There are alot of 1 person buildings, with 1 person attending for 100% attendance. I want to do something to account for turnout size, those bigger buildings with only 50% still have a huge number of ppl turn out. I should also probably use a log regression for my 0-1 scale.

```
ggplot(WorksiteMeetings, aes(x=TurnoutPercentage)) + geom_histogram(bins=30) +xlim(0,1.1)
```

```
## Warning: Removed 8 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```

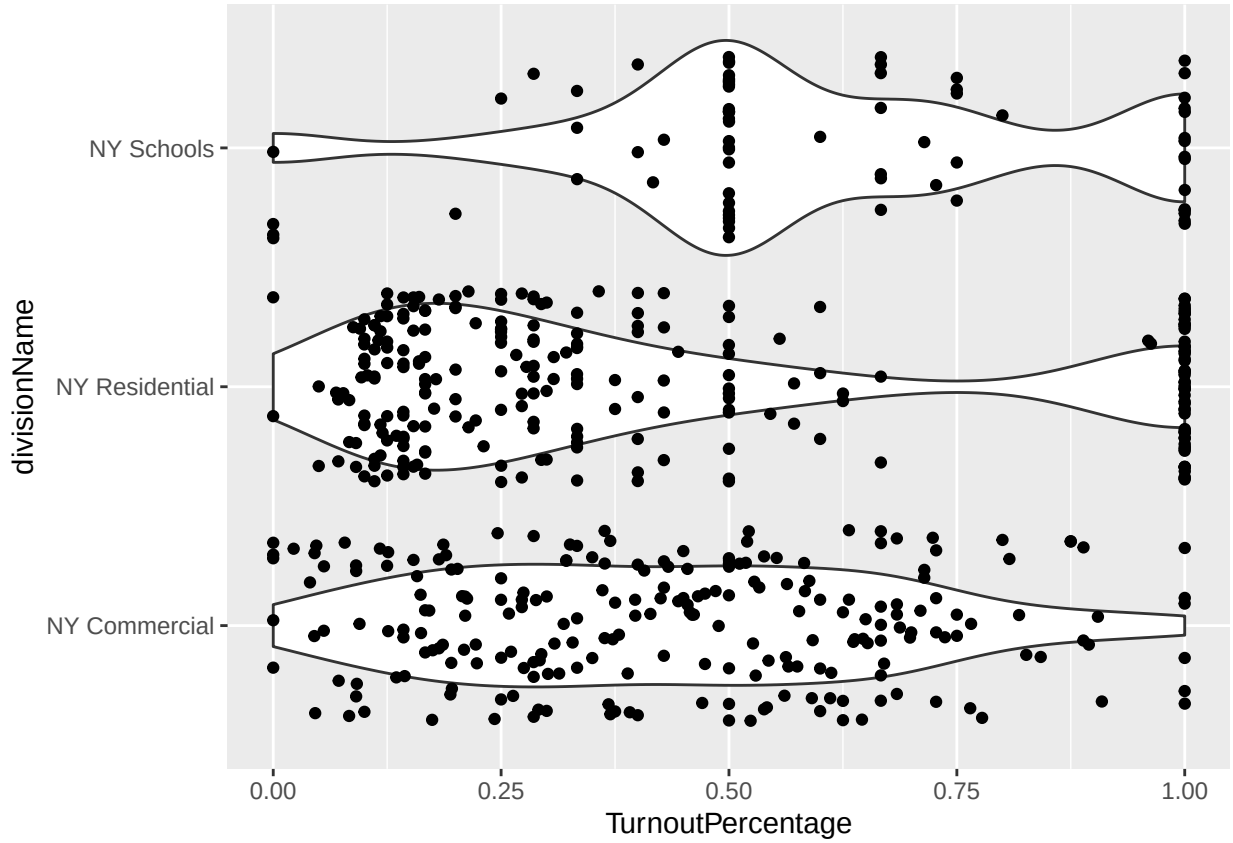



TurnoutPercentage vs divisionName

```
ggplot(WorksiteMeetings, aes(x=TurnoutPercentage, y=divisionName))  + geom_violin() + geom_jitter()
```

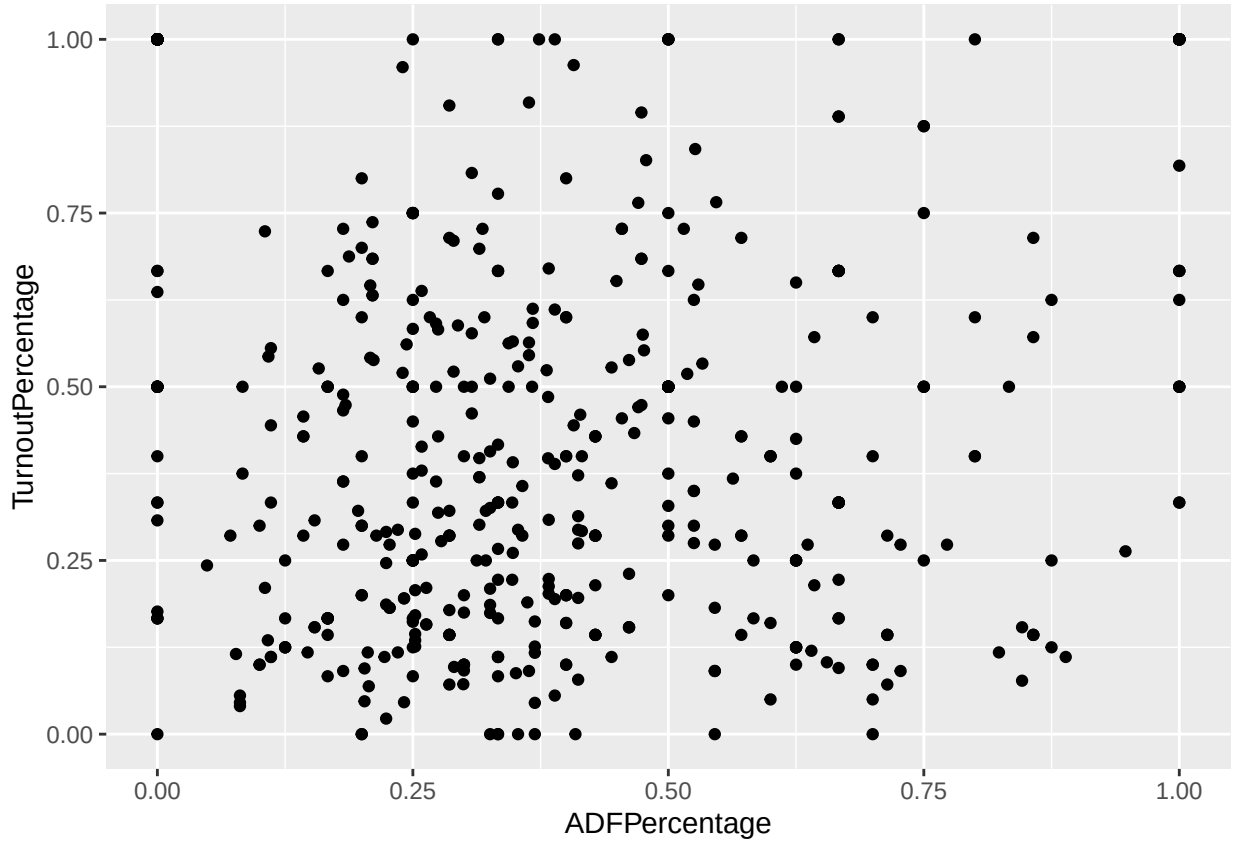
```
## Warning: Removed 8 rows containing non-finite outside the scale range
## (`stat_ydensity()`).
```

```
## Warning: Removed 8 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
ggplot(WorksiteMeetings, aes(x=ADFPercentage, y=TurnoutPercentage))  + geom_point()
```

```
## Warning: Removed 8 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



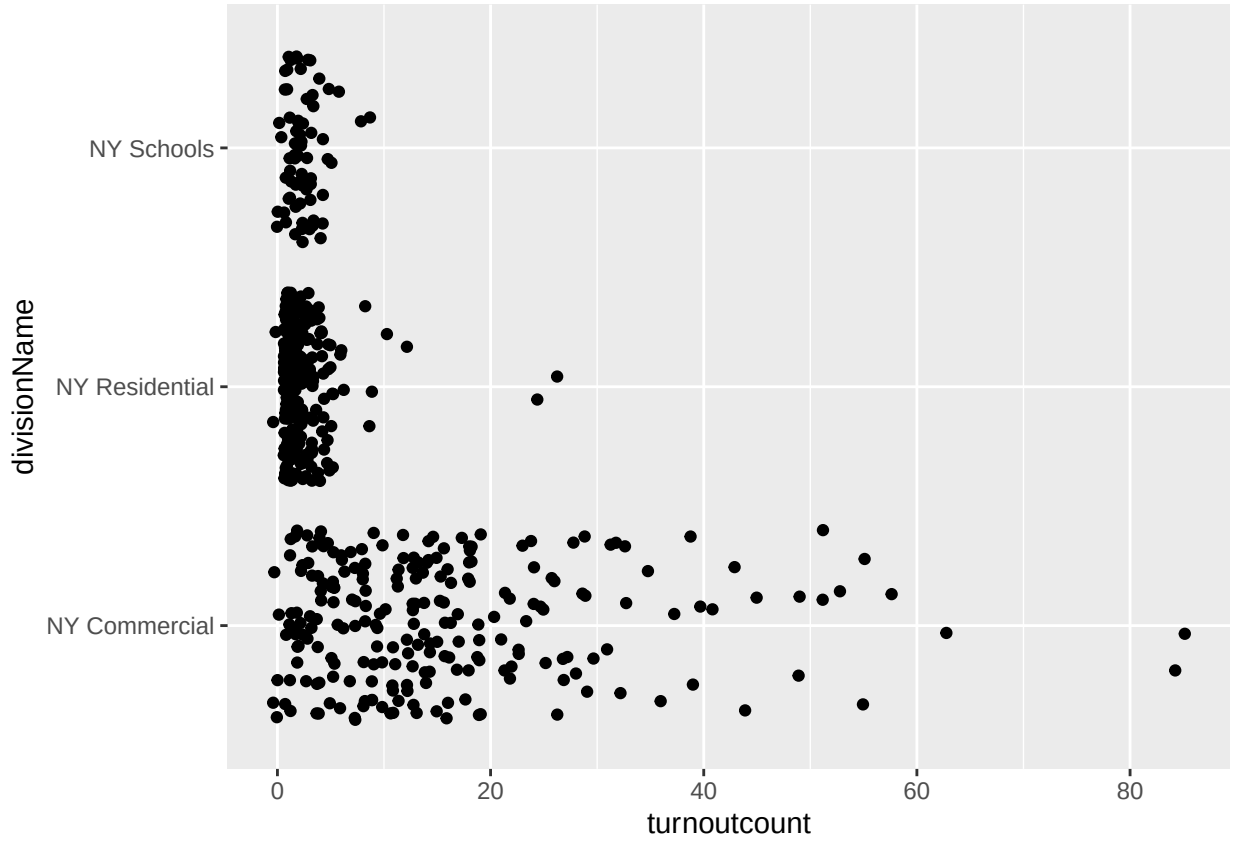

**turnoutcount:** Turnout at that individual event. Distribution is looking similar to building size, makes sense

```
describe(WorksiteMeetings$turnoutcount)
```

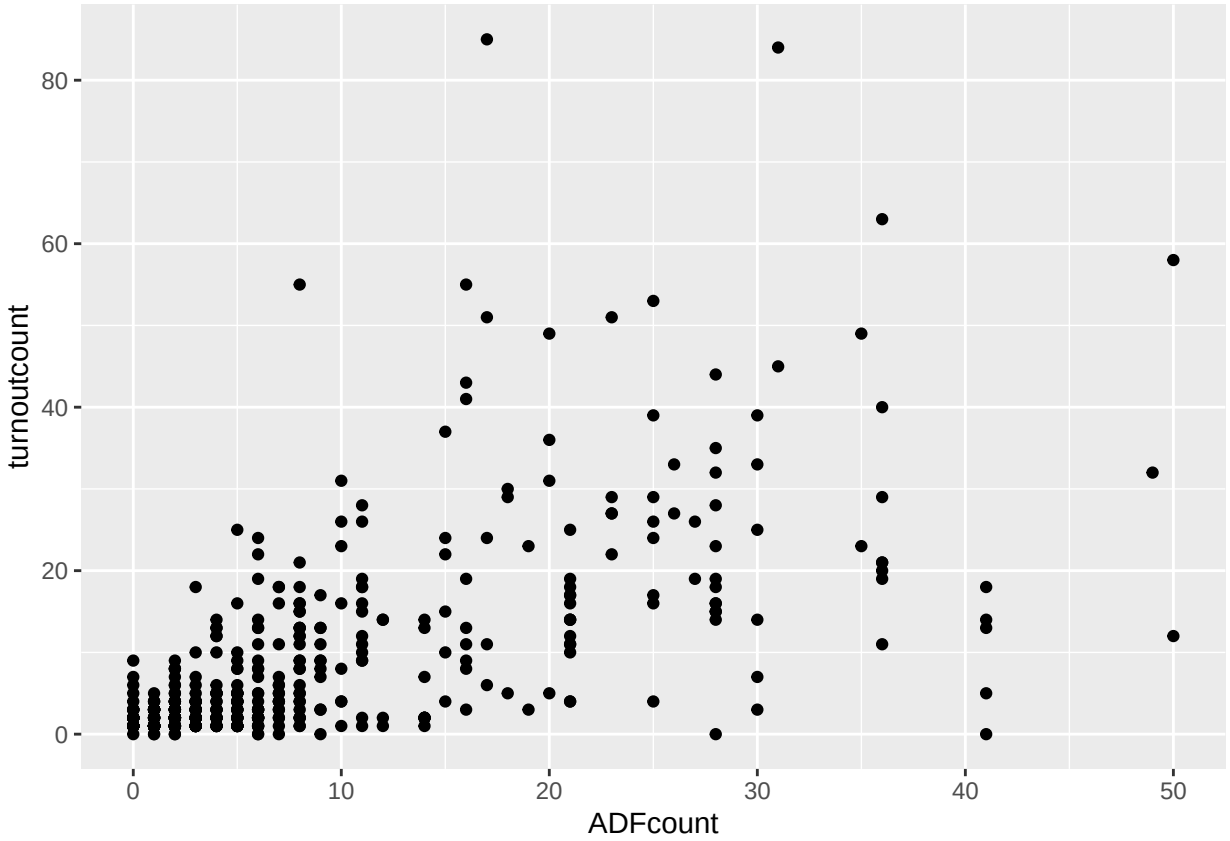
```
## WorksiteMeetings$turnoutcount
##        n  missing distinct     Info     Mean  pMedian      Gmd      .05
##      549        0       51    0.976    7.851        5    9.776      1.0
##      .10      .25      .50      .75      .90      .95
##      1.0      1.0      3.0     10.0     21.2     29.6
##
## lowest :  0  1  2  3  4, highest: 55 58 63 84 85
```

turnoutcount vs divisionName

```
ggplot(WorksiteMeetings, aes(x=turnoutcount, y=divisionName)) + geom_jitter()
```



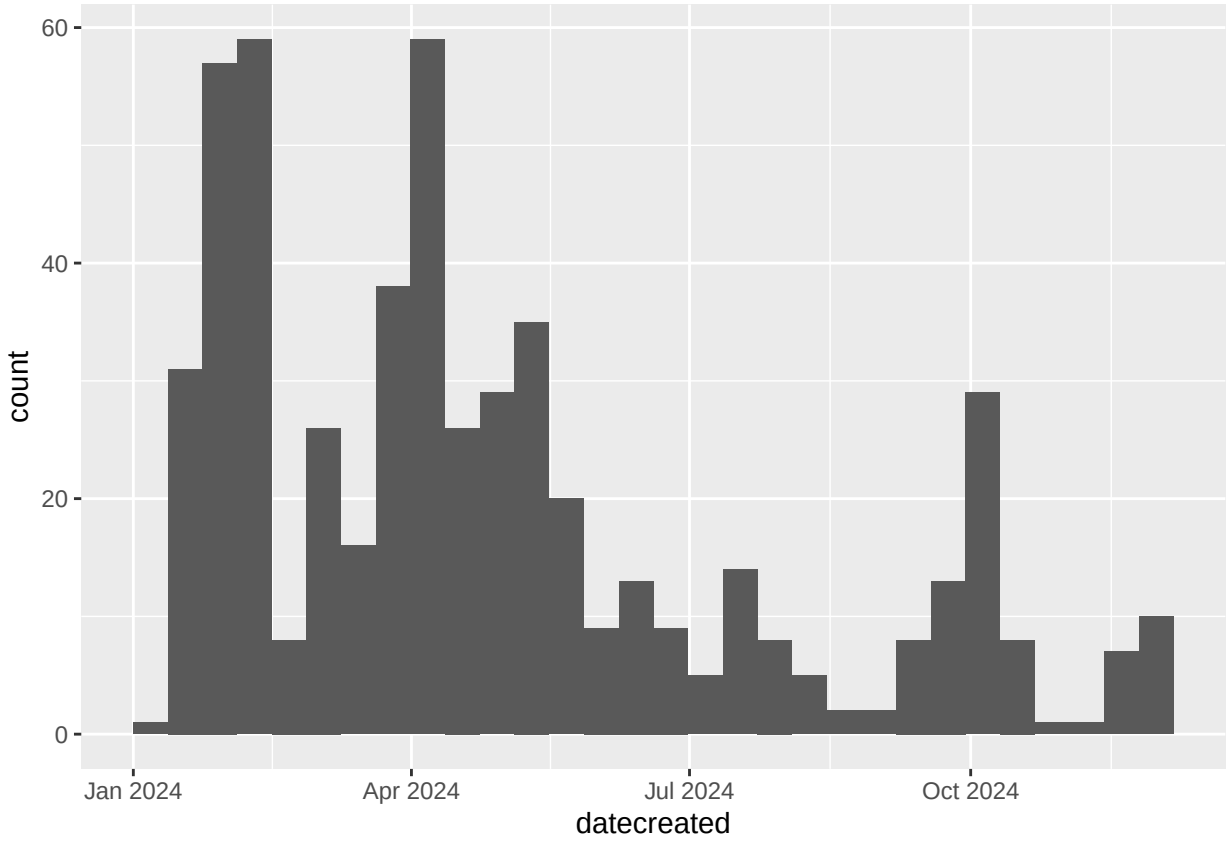

```
ggplot(WorksiteMeetings, aes(x=ADFcount, y=turnoutcount))  + geom_point()
```




**DateCreated: is the date the event happened. I didn't end up adding this to my model, but it was interesting to look at. Unsure if i remember what exactly happening in April to cause the big spike in the spring, i know we had a few buildings try to forcefully switch to non-union staff this spring which caused some commotion, that could've been when reps were checking in more with those buildings?**

```
ggplot(WorksiteMeetings, aes(x=datecreated)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




**Part 4 - Inference**

**Linear Model**:I tried a linear model using the turnout counts and ADF counts. It had heavy tails, distribution is not normal, very skewed. Unfortunately, we learned this is probably not a good fit for linear models?

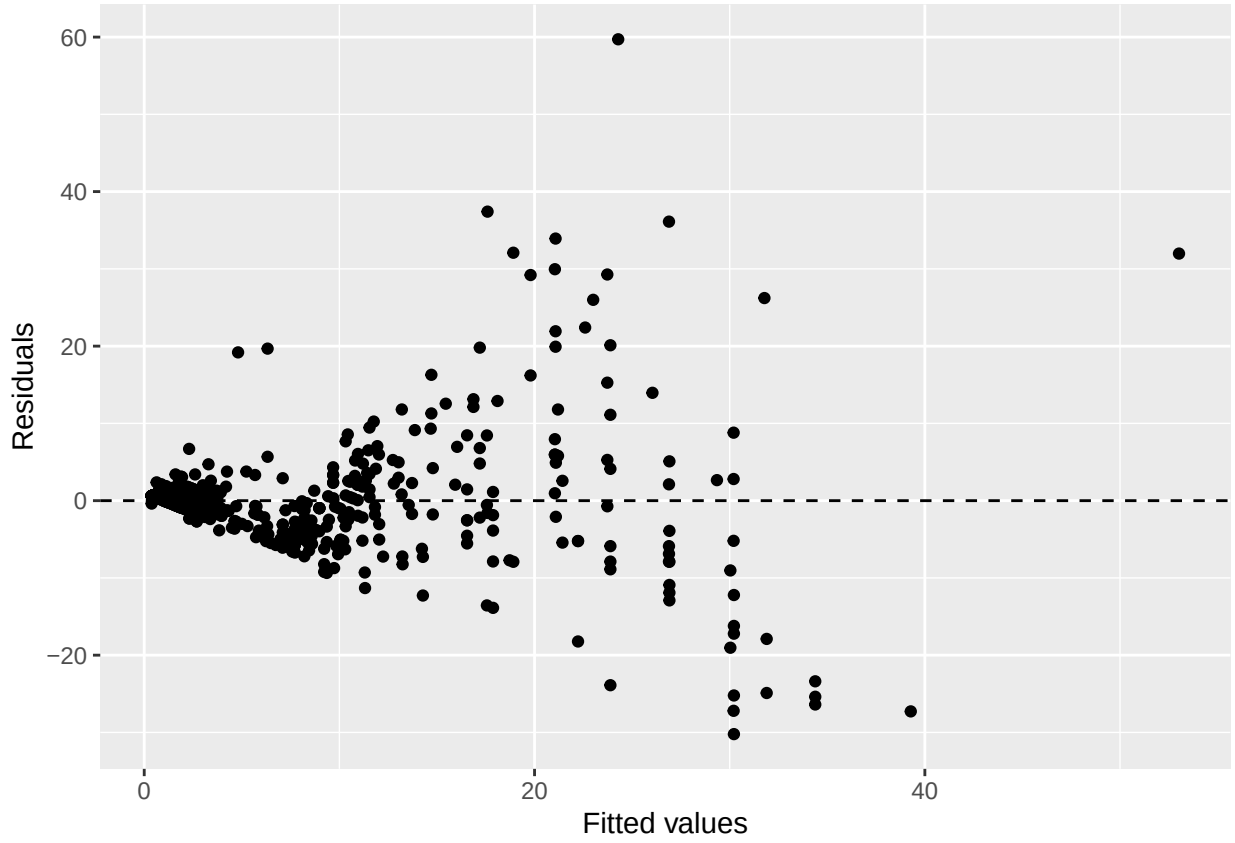
```
turnout_count <- lm(turnoutcount ~ divisionName + memberCount + ADFcount, data = WorksiteMeetings)
summary(turnout_count)
```

```
##
## Call:
## lm(formula = turnoutcount ~ divisionName + memberCount + ADFcount,
##     data = WorksiteMeetings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.216  -1.963  -0.192   1.126  59.713
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  6.36557    0.81115   7.848 2.32e-14 ***
## divisionNameNY Residential -6.10577    0.86253  -7.079 4.58e-12 ***
## divisionNameNY Schools     -4.91126    1.19351  -4.115 4.48e-05 ***
## memberCount                 0.12096    0.01621   7.461 3.48e-13 ***
## ADFcount                    0.25426    0.06107   4.163 3.65e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.8 on 536 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.5402, Adjusted R-squared:  0.5367
## F-statistic: 157.4 on 4 and 536 DF,  p-value: < 2.2e-16
```

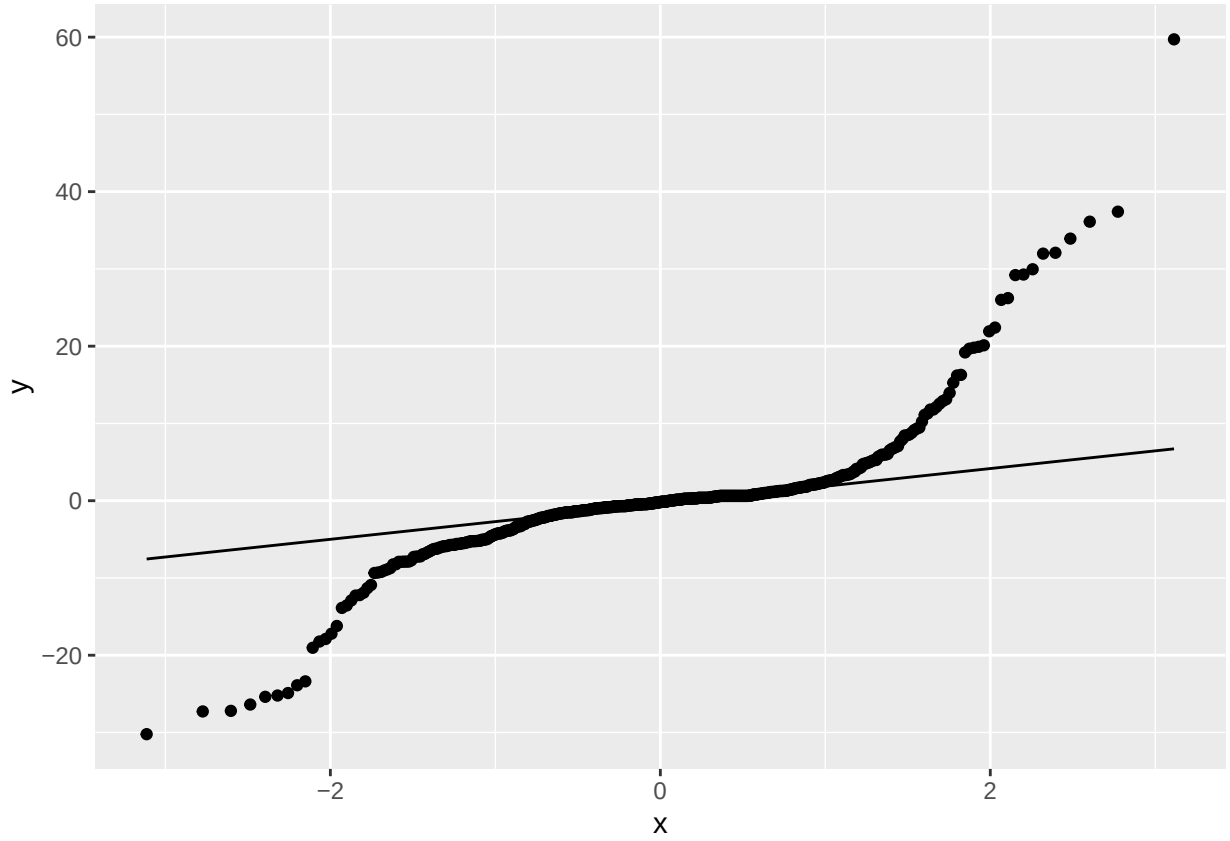
My residuals seems to be increasing and are not constant

```
ggplot(data = turnout_count, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```




heavy tails, distribution is not normal

```
ggplot(data = turnout_count, aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line()
```




**Weighted Logistic Model** Using percentage of turnout, turnout count divided by member count at the building, to get a proportion, bounded within 0 to 1. I also found it needed to include the turnout count as a weight in the model, bc my data is a bit unbalanced:

```
log_turnout_count2 <- glm(TurnoutPercentage ~ divisionName + memberCount + ADFcount, data = WorksiteMee
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```
summary(log_turnout_count2)
```

```
##
## Call:
## glm(formula = TurnoutPercentage ~ divisionName + memberCount +
##     ADFcount, family = binomial(link = "logit"), data = WorksiteMeetings,
##     weights = turnoutcount)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 0.5782769  0.0702410   8.233  < 2e-16 ***
## divisionNameNY Residential -0.7017054  0.1022923  -6.860 6.90e-12 ***
## divisionNameNY Schools      0.2352902  0.1768350   1.331   0.1833
## memberCount                -0.0071310  0.0009131  -7.810 5.73e-15 ***
## ADFcount                   -0.0076866  0.0036115  -2.128   0.0333 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1131.38  on 529  degrees of freedom
## Residual deviance:  952.04  on 525  degrees of freedom
##   (8 observations deleted due to missingness)
## AIC: 1896
##
## Number of Fisher Scoring iterations: 4
```

Odds ratio (e^log-odds): I found the logistic regression interpretation a bit confusing, but these are my interpretation (while holding all other variables constant):

NY Residential buildings odds ratio of 0.49, suggests a 51% lower odds of turnout than a NY Commercial buildings

With the membercount odds ratio with 0.99, a increase of 1 more member at the buildings, would suggests a 1% lower odds of turnout

Similarly with ADFCount odds ratio being 0.99, a increase of 1 more person signing up for ADF at the buildings, would suggests a 1% lower odds of turnout

```
exp(coefficients(log_turnout_count2))
```

```
##                (Intercept) divisionNameNY Residential
##                  1.7829636                  0.4957391
##     divisionNameNY Schools                memberCount
##                  1.2652759                  0.9928944
##                   ADFcount
##                  0.9923429
```

**Chi Squared:** Null hypothesis (H0): there is no association between our variables Alternative hypothesis (H1): there is an association between our variables Chi-squared to p-value: X2 = Null deviance – Residual deviance p-value is less than.05 so we reject the null hypothesis. There is a statistically significant relationship between the buildings' division, member count, and ADF contributions to turnout

```
1-pchisq(1131.38-952.04, 529-525)
```

```
## [1] 0
```

### Part 5 - Conclusion

- Our turnout seems to be more polarized than we thought, with extremes (both very high and very low turnout %) seeming to be more common than expected
- Our Union Reps may be putting to much weight on whether a building contributes to ADF means they'll be more likely to come to Worksite Meetings
- Commercial Building members are much more likely to turnout
- Targeting smaller buildings may be key for better turnout

Limitations: I'm calculating turnout % using past turnout divided by current roster. Usually, rosters stay similar sizes through the year (especially due to union contract protections on unlawful reduction in forces), but they do change, introducing possible inaccuracy.

Improved dataset: I think its possible but tedious to have my SQL queries roster and ADF data for each day, to more accurately calculate the turnout % for the exact day of the worksite meeting

Future Analysis: With that dataset, I could compare data before and after worksite meetings to see the potential impact of the visit

**References**

Data is self-collected.