DATA 606 Data Project Proposal

null

Data Preparation

```r
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 error
```

```r
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
```

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

```
glimpse(WorksiteMeetings)
```

```
## Rows: 577
## Columns: 20
## $ unionEventID   <int> 19413, 19417, 19420, 19423, 19424, 19427, 19434, 19435,~
## $ v3accountid    <int> 64787, 56700, 59692, 47374, 46344, 15273, 39055, 52881,~
## $ campaignTypeID <int> 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25,~
## $ districtID     <int> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5~
## $ districtName   <chr> "New York Metro", "New York Metro", "New York Metro", "~
## $ subdivisionid  <int> 46, 48, 46, 46, 46, 46, 47, 47, 47, 48, 47, 47, 47, 47,~
## $ divisionName   <chr> "NY Commercial", "NY Schools", "NY Commercial", "NY Com~
## $ v3fieldrepname <chr> "Pasquale Follano", "Tyrae Allen", "Rosayri Perez", "Ar~
## $ createdBy      <chr> "Sam Resheff", "Tyrae Allen", "Rosayri Perez", "Arlind ~
## $ eventname      <chr> "Worksite meeting at ABM Janitorial Services, Inc-1 Bro~
## $ address1       <chr> "1 BROADWAY", "1001 EAST 100TH STREET", "9 WEST 57TH ST~
## $ city           <chr> "NEW YORK", "BROOKLYN", "NEW YORK", "NEW YORK", "NEW YO~
## $ statename      <chr> "New York", "New York", "New York", "New York", "New Yo~
## $ zipcode        <chr> "10004", "11236-4415", "10019", "10022", "10017", "1001~
## $ turnoutcount   <int> 0, 0, 45, 15, 23, 31, 4, 1, 1, 4, 1, 2, 5, 2, 3, 8, 7, ~
## $ datecreated    <chr> "1/12/2024", "1/16/2024", "1/16/2024", "1/17/2024", "1/~
## $ employerID     <int> 886, 2339, 9957, 4695, 5659, 2969, 11053, 10069, 10069,~
## $ employerName   <chr> "ABM Janitorial Services, Inc", "NYC School Support Ser~
## $ memberCount    <chr> "11", "5", "69", "25", "41", "48", "14", "1", "7", "4",~
## $ ADFcount       <chr> "6", "1", "31", "8", "10", "10", "2", "NULL", "2", "2",~
```

```
summary(WorksiteMeetings)
```

```
##   unionEventID    v3accountid    campaignTypeID   districtID districtName
##  Min.   :19413   Min.   :13377   Min.   :25     Min.   :5    Length:577
##  1st Qu.:19761   1st Qu.:34605   1st Qu.:25     1st Qu.:5    Class :character
##  Median :20311   Median :46931   Median :25     Median :5    Mode  :character
##  Mean   :20391   Mean   :46627   Mean   :25     Mean   :5
##  3rd Qu.:20850   3rd Qu.:57589   3rd Qu.:25     3rd Qu.:5
##  Max.   :21923   Max.   :71199   Max.   :25     Max.   :5
##  subdivisionid  divisionName       v3fieldrepname      createdBy
##  Min.   :46.0   Length:577         Length:577         Length:577
##  1st Qu.:46.0   Class :character   Class :character   Class :character
##  Median :47.0   Mode  :character   Mode  :character   Mode  :character
##  Mean   :47.6
##  3rd Qu.:47.0
##  Max.   :65.0
##   eventname           address1             city             statename
##  Length:577         Length:577         Length:577         Length:577
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    zipcode           turnoutcount    datecreated          employerID
##  Length:577         Min.   : 0.000   Length:577         Min.   :  117
##  Class :character   1st Qu.: 1.000   Class :character   1st Qu.: 2339
##  Mode  :character   Median : 3.000   Mode  :character   Median : 4226
##                     Mean   : 7.695                      Mean   : 4801
##                     3rd Qu.: 9.000                      3rd Qu.: 7692
##                     Max.   :85.000                      Max.   :11742
##  employerName       memberCount          ADFcount
##  Length:577         Length:577         Length:577
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

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

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

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

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

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

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

New fields for turnout percentage and ADF percentage of building POSSIBLE PROBLEM: there are 4 buildings with a percent higher than 100%, this could happen if the roster size changed since the event happened. these are all in smaller buildings, i think i will just count them as 100%, as they're off by 1 or 2 and they got the entire size of the current building to come to event. this could be an issue with other buildings that i cant detect under 100%, but i dont have a better way to compare each individual roster size at that time of the event, so i'm using the current roster as a proxy, building rosters usually stay around the same size

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

Context

I work as a Data Analyst at 32BJ Labor Union. Alot of our members are in NYC, but we do have members all down the East Coast (although I may analyze just the NYC ones for this one). Our Union Representatives hold Worksite Meetings at buildings to meet with Union Members, and we been tracking this for the past 10 years. But recently in the past 2 years we've been tracking member attendance digitally with our app (scanning member id cards), which we hope will be much more accurate. I will be just using this data from the past 2 years. We are curious if having these meetings effects members union activity (event turnout etc), or political donations (ADF–American Dream Fund)

Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.

Can we predict member turnout percentage for membership meetings?

turnout percentage ~ divisionName + membercount + ADFCount + employerName + city + v3field rep

Cases

What are the cases, and how many are there?

There have been 577 worksite meetings in the past year in New York

Data collection

Describe the method of data collection.

Data is self-collected. I wrote the SQL query from our union database. The worksite meetings are inputted by our Field reps scanning member ids using our app.

Type of study

What type of study is this (observational/experiment)?

This is an observational study.

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

Self-collected.

Describe your variables?

Are they quantitative or qualitative

quantitative

If you are are running a regression or similar model, which one is your dependent variable?

turnout percentage is my dependent variable

Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

```
t <- sort(table(WorksiteMeetings$v3accountid), decreasing = TRUE)
```

v3accountid: Number of times a site was visited through the year, most locations were only visited once.

```
describe(t)
```

```
## t
##        n  missing distinct     Info     Mean  pMedian      Gmd
##      434        0        7    0.503    1.329        1   0.5623
##
## Value          1     2     3     4     5     6     7
## Frequency    343    65    12     6     5     2     1
## Proportion 0.790 0.150 0.028 0.014 0.012 0.005 0.002
##
## For the frequency table, variable is rounded to the nearest 0
```

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

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

```
##
##  NY Commercial NY Residential     NY Schools    NY Security
##            228            249             72             28
```

v3fieldrepname

```
sort(table(WorksiteMeetings$v3fieldrepname),decreasing = TRUE)
```

```
##
##              Allan Smyth         Frank Cifuentes           Sheamus Barry
##                       97                      88                      47
##                Dem Kukic            Adem Kajosaj             Tyrae Allen
##                       44                      34                      34
##           Shameka Murray             Arlind Lela           Mark Espinoza
##                       28                      27                      23
## Unassigned NY Commercial            Larry Wilson        Kimberly Eyssen
##                       19                      18                      16
##            Rosayri Perez          Carlos Cabrera      Michael Defreitas
##                       14                      13                      12
##          Carlos A. Garcia            Evan Lambert           Ignacio Velez
##                        8                       7                       7
##         Pasquale Follano          Heidy Tavarez           Billy Laburda
##                        7                       5                       4
##              Leon Burnes             Mary Rosario              Mateo Daija
##                        4                        4                        4
##         Vincent Roveccio             Ralph Osorio           Esteban Flores
##                        4                        3                        2
##              Scott Cohen           Frank Castillo       Rogelio Cox Walker
##                        2                        1                        1
```

v3fieldrepname and divisionName

```r
table(WorksiteMeetings$v3fieldrepname,WorksiteMeetings$divisionName)
```

```
## 
##                             NY Commercial NY Residential NY Schools NY Security
##   Adem Kajosaj                          0             34          0           0
##   Allan Smyth                           0             97          0           0
##   Arlind Lela                          27              0          0           0
##   Billy Laburda                         0              4          0           0
##   Carlos A. Garcia                      0              0          8           0
##   Carlos Cabrera                       13              0          0           0
##   Dem Kukic                             0             44          0           0
##   Esteban Flores                        0              2          0           0
##   Evan Lambert                          0              0          7           0
##   Frank Castillo                        0              1          0           0
##   Frank Cifuentes                      88              0          0           0
##   Heidy Tavarez                         5              0          0           0
##   Ignacio Velez                         0              7          0           0
##   Kimberly Eyssen                      16              0          0           0
##   Larry Wilson                         18              0          0           0
##   Leon Burnes                           4              0          0           0
##   Mark Espinoza                         0              0         23           0
##   Mary Rosario                          4              0          0           0
##   Mateo Daija                           0              4          0           0
##   Michael Defreitas                    12              0          0           0
##   Pasquale Follano                      7              0          0           0
##   Ralph Osorio                          0              3          0           0
##   Rogelio Cox Walker                    1              0          0           0
##   Rosayri Perez                        14              0          0           0
##   Scott Cohen                           0              2          0           0
##   Shameka Murray                        0              0          0          28
##   Sheamus Barry                         0             47          0           0
##   Tyrae Allen                           0              0         34           0
##   Unassigned NY Commercial             19              0          0           0
##   Vincent Roveccio                      0              4          0           0
```

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

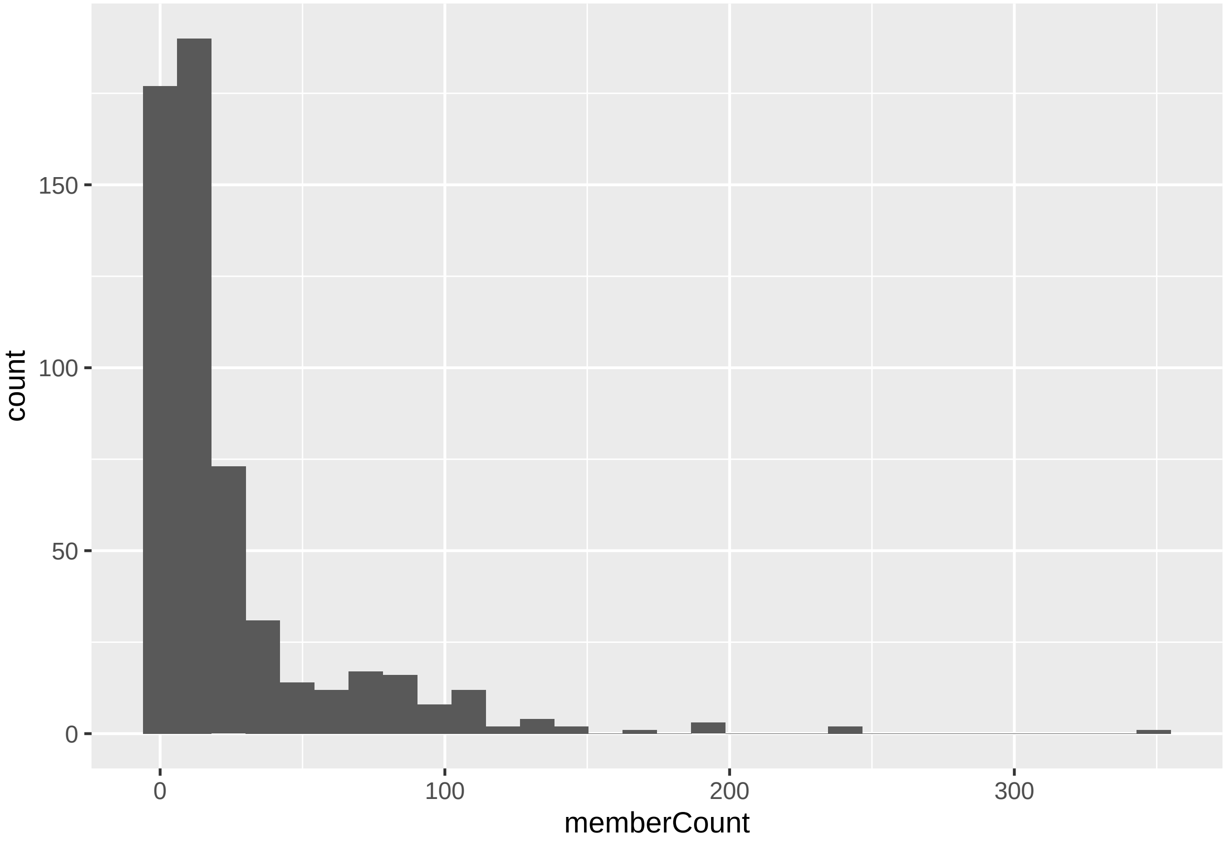
```r
describe(WorksiteMeetings$memberCount)
```

```
## WorksiteMeetings$memberCount 
##        n  missing distinct     Info     Mean  pMedian      Gmd      .05
##       565       12       76    0.998    24.96     14.5    31.09        1
##       .10      .25      .50      .75      .90      .95
##         2        5       11       26       73       94
##
## lowest :   1   2   3   4   5, highest: 148 167 198 237 350
```

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

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

```
## Warning: Removed 12 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



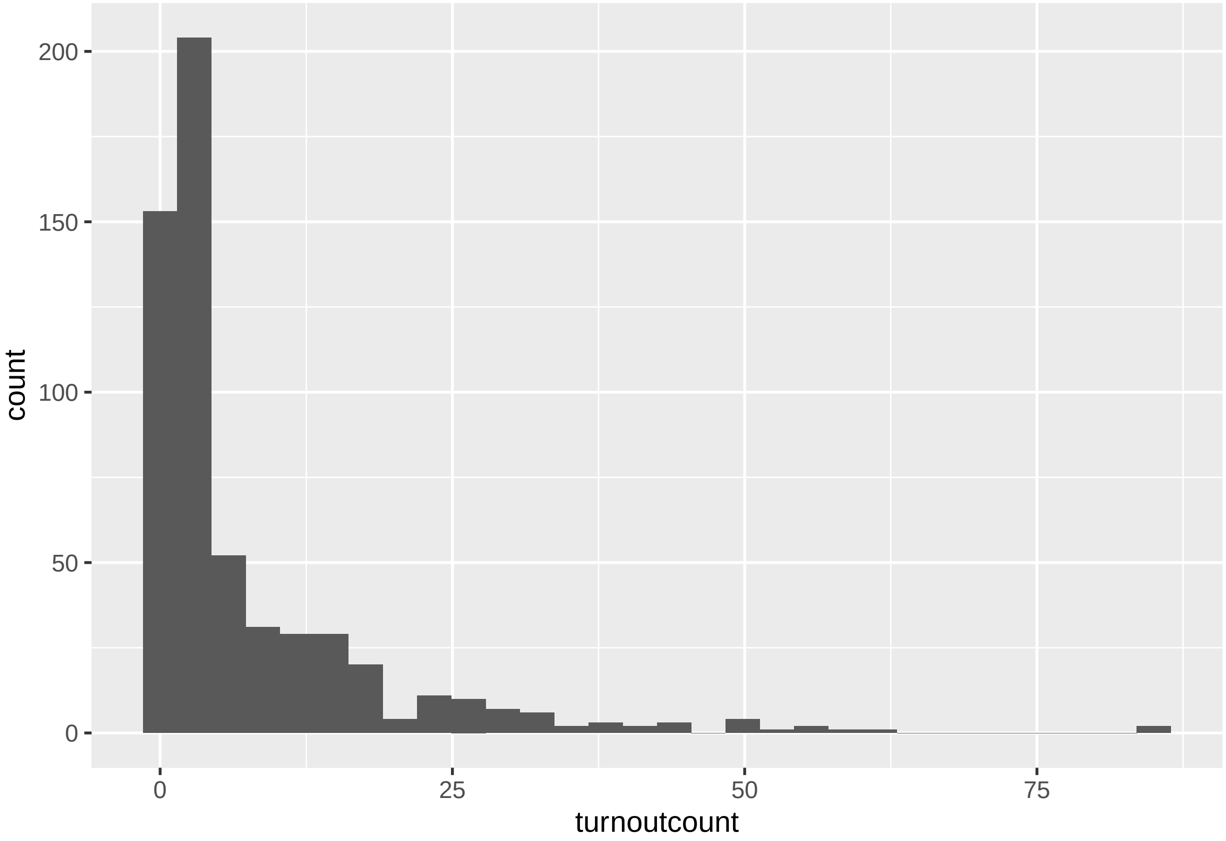
turnoutcount: Turnout at that individual event. Distribution is looking similar to building size, makes sense

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

```
## WorksiteMeetings$turnoutcount
##         n  missing distinct     Info     Mean  pMedian      Gmd      .05
##       577        0       51    0.978    7.695      4.5    9.509      1.0
##       .10      .25      .50      .75      .90      .95
##       1.0      1.0      3.0      9.0     20.4     29.0
##
## lowest :  0  1  2  3  4, highest: 55 58 63 84 85
```

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

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



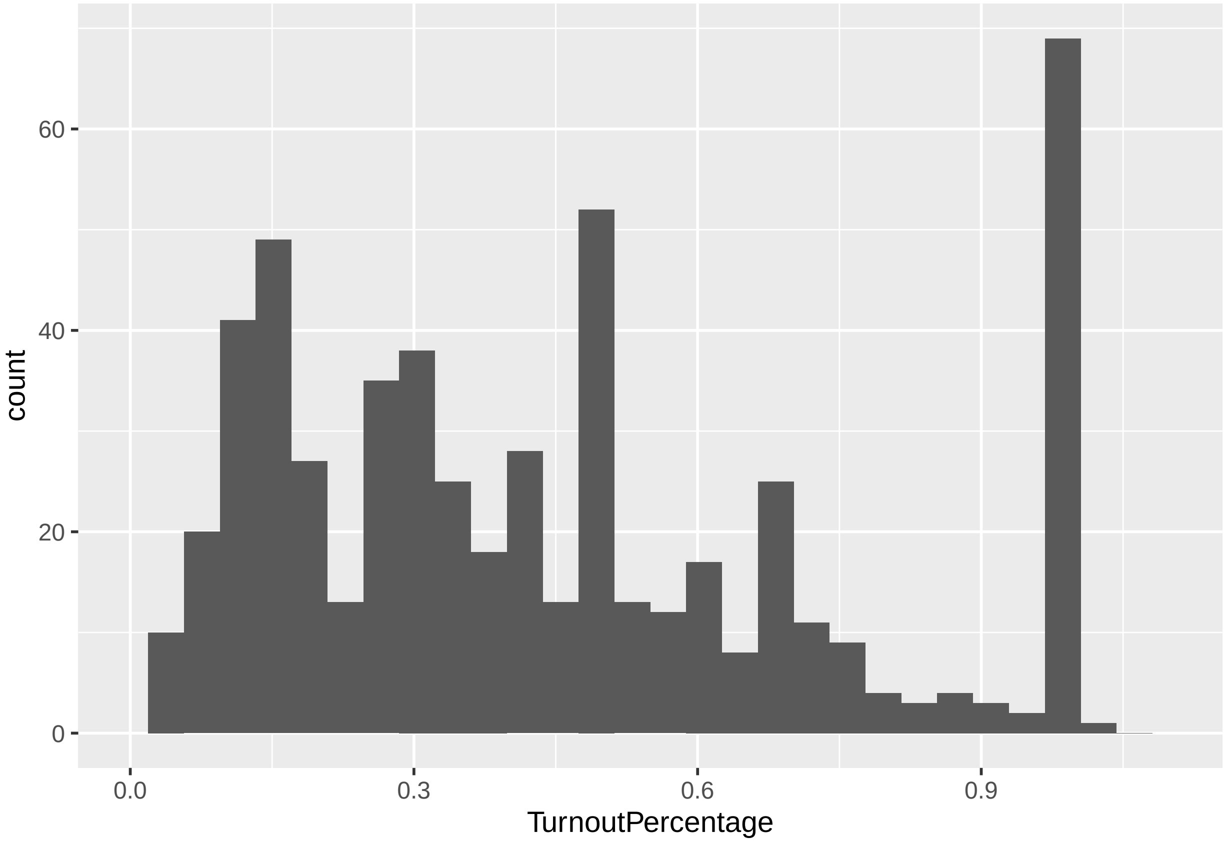
TurnoutPercentage: percent of building that went to an event. There are alot of 1 person buildings, with 1 person attending for 100% attendance. I want to do something to account for turnout size, those bigger buildings with only 50% still have a huge number of ppl turn out. I should also probably use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

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

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

## Warning: Removed 16 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()`).
```



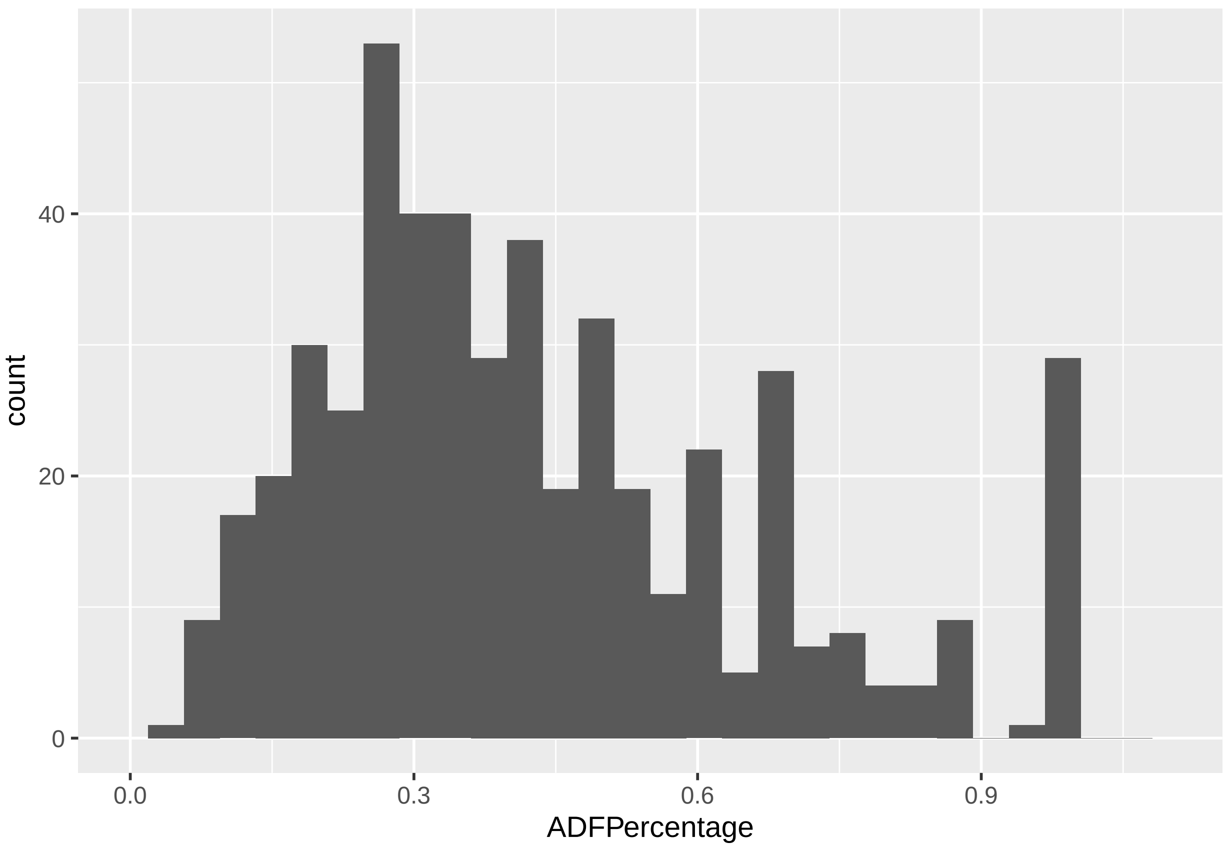
ADF Percentage, percentage of members paying into our political fund (we usually assume this people are more likely to be politically and union involved). happy to see a more normal distribution. again there are those 1 person buildings with that 1 member paying for 100% rate.

```
ggplot(WorksiteMeetings, aes(x=ADFPercentage)) + geom_histogram() +xlim(0,1.1)
```

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

```
## Warning: Removed 77 rows containing non-finite outside the scale range
## (‘stat_bin()‘).
```

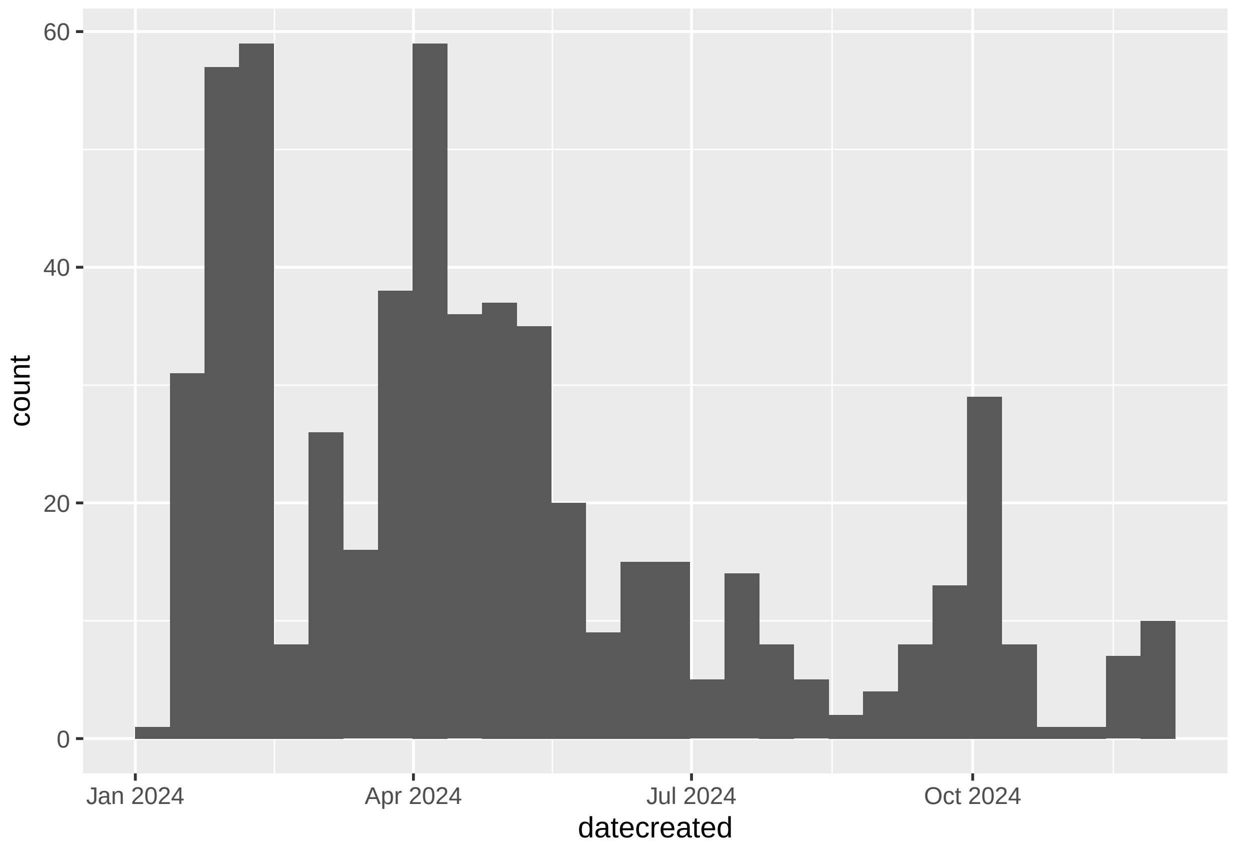
```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (‘geom_bar()‘).
```



DateCreated: is the date the event happened. 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‘.
```


count
60
40
20
0
Jan 2024
Apr 2024
Jul 2024
Oct 2024
datecreated