Homework 5

```
raw_stx <- read.delim("game_data_public.STX.Sealed.csv",header = TRUE,sep = ",")
```

```
stx <- raw_stx %>%
  rowwise() %>%
  mutate(drawn=ifelse(on_play,
         sum(c_across(360:702))- num_turns+1,
          sum(c_across(360:702))- num_turns),
         won_int= ifelse(won,1,0))
theory <- stx %>%
  select("won_int","drawn") %>%
  group_by(drawn) %>%
  summarize(win_rate = mean(won_int, na.rm = TRUE))
theory_count <- stx %>%
  select("won_int","drawn") %>%
  group_by(drawn) %>%
  tally()
theory_sliced <- stx %>%
  select("won_int","drawn") %>%
  group_by(drawn) %>%
  summarize(win_rate = mean(won_int, na.rm = TRUE)) %>%
  slice(4:24)
```

##17lands Card Advantage

In Magic the Gathering the concept of card advantage is one of the earliest theories of how to win games. In its simplest form it says that the player who has drawn more cards is more likely to win the game. This data set does not have information about or a way to deduce the number of cards a 17lands user's opponent drew. Therefore we cannot explicitly test this theory. However the theory of card advantage does imply that the more cards you draw the more likely you are to win, so seeing if there is a relationship between win rate and the number of cards drawn is the next best thing. This dataset consists of 1388 variables. There are 343 unique cards that players can have in their decks. The first 17 variables are dedicated to keeping track of the player and game overall. In the dataset are 343 consecutive columns each describing how many of that particular card they drew over the course of the game. So my first step was to mutate a column that was the sum of those columns. However in games of Magic, unless it is the first turn of the first player to go, you draw a card each turn. So I subtracted how many turns it had been from that sum, plus one if they went first, so as to not be influenced by how many turns the games had gone. This is sound under card advantage theory because it's about the relative number of cards drawn and both players draw as the game goes on. In the same mutate I also had to recode the column tracking if the player won. It was originally made up of boolean values, but converting them to 0's and 1's meant taking their mean would give me the win rate and I no longer had to deal with error messages.

```
raw_stx %>%
  select("num_turns","won",360:361) %>%
  slice(1:6) %>%
  kable(align = 'c',caption="The Shape of the Original Dataset")
```

Table 1: The Shape of the Original Dataset

num_turns	won	drawn_Abundant.Harvest	drawn_Academic.Dispute
10	False	0	0
2	False	0	0
10	True	0	0
15	True	0	0
8	False	0	0
8	True	0	0

```r
stx %>%
  select("num_turns","won","won_int","drawn") %>%
  head() %>%
  kable(align = 'c',caption="The Shape of the New Dataset")
```

Table 2: The Shape of the New Dataset

num_turns	won	won_int	drawn
10	False	0	2
2	False	0	0
10	True	1	7
15	True	1	6
8	False	0	3
8	True	1	4

Next I want to share the table of how many players drew how many cards.

```r
stx %>%
  select("won_int","drawn") %>%
  group_by(drawn) %>%
  tally() %>%
  kable(align = 'c',caption="Tally of Cards Drawn")
```

Table 3: Tally of Cards Drawn

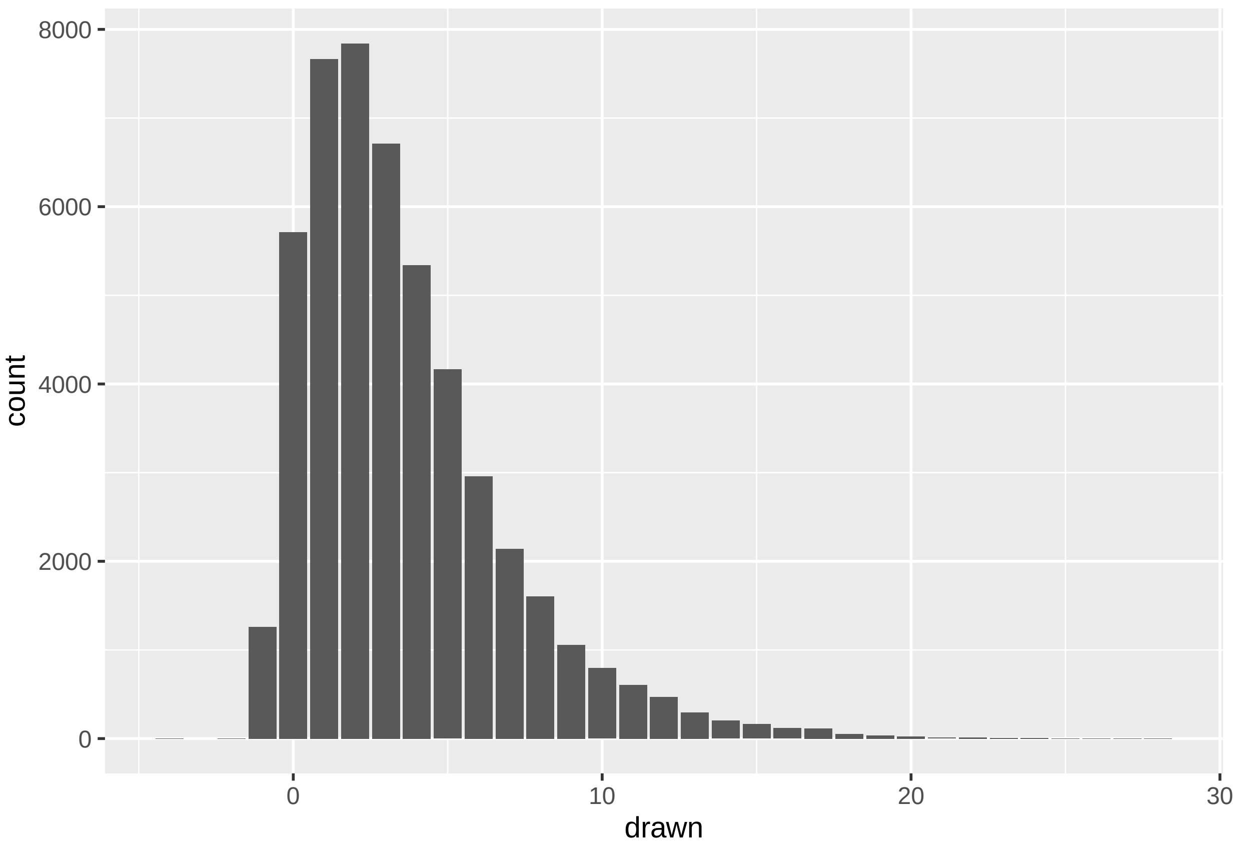
drawn	n
-4	1
-2	3
-1	1259
0	5713
1	7664
2	7842
3	6712
4	5341
5	4164
6	2957
7	2142
8	1606
9	1058

drawn	n
10	795
11	607
12	469
13	294
14	202
15	163
16	118
17	114
18	51
19	35
20	23
21	10
22	13
23	7
24	9
25	3
26	1
27	2
28	3

```
outliers <- theory_count %>%
  slice(25:32) %>%
  select("n") %>%
  colSums()
```

From this we can see that there are some nonsensical negative values, which I currently do no understand, but there are few enough that I will ignore them. Also there are so few data points for players who have drawn more than 20 extra cards that I want to also ignore them. They only make up 0.0972034% of all games recorded. I feel this graph also helps put in perspective how few data points these are to help justify this. I will show analysis with and without these data points.

```
ggplot(stx,aes(drawn)) + geom_bar()
```



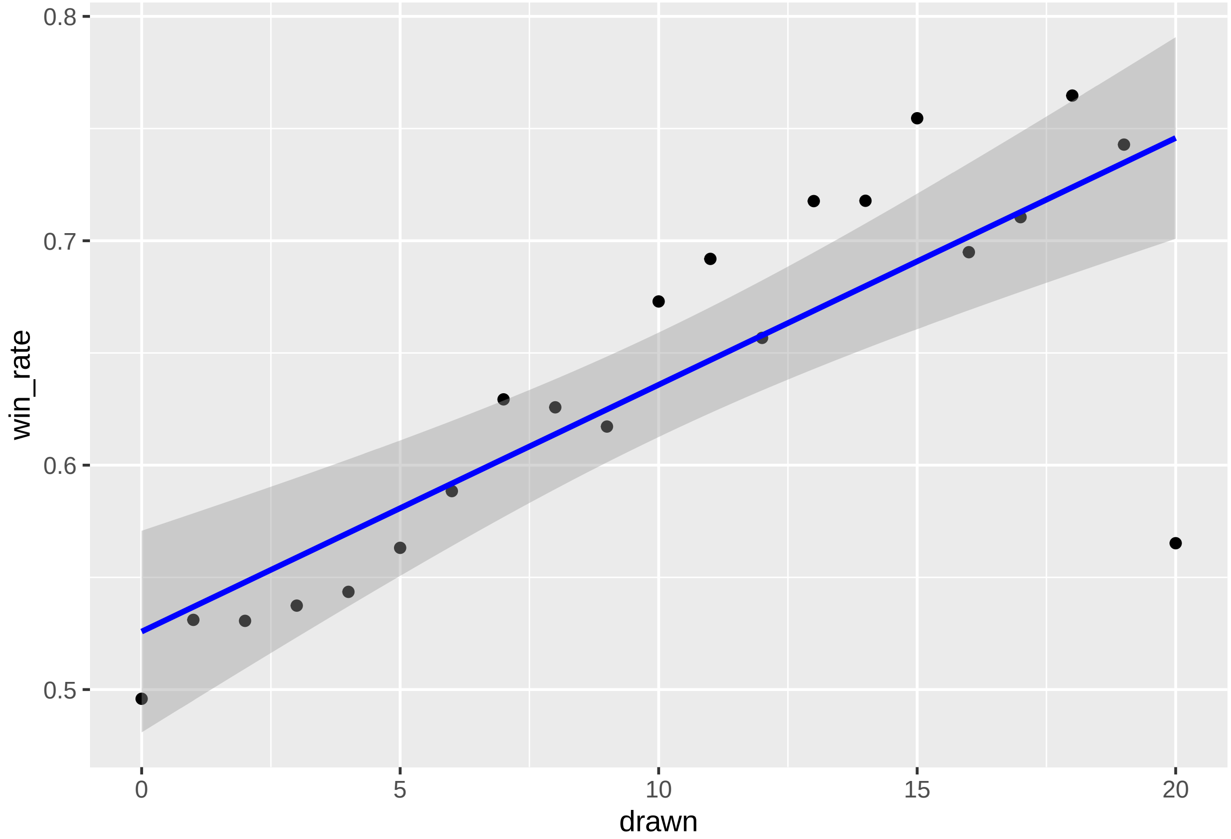
So now I can group by cards drawn and associate a win rate to each value. With those outliers there is a correlation of 0.0423556, but without them there is a correlation of 0.8085793. Moreover when we do linear regression of the data without outliers we get:

```
summary(lm(theory_sliced$drawn~theory_sliced$win_rate))
```

```
##
## Call:
## lm(formula = theory_sliced$drawn ~ theory_sliced$win_rate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7722 -1.8662 -1.1491  0.3388 14.1971
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -27.795      6.362  -4.369  0.00033 ***
## theory_sliced$win_rate   59.443      9.924   5.990 9.17e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.746 on 19 degrees of freedom
## Multiple R-squared:  0.6538, Adjusted R-squared:  0.6356
## F-statistic: 35.88 on 1 and 19 DF,  p-value: 9.169e-06
```

```
ggplot(theory_sliced,aes(drawn,win_rate)) + geom_point() +
    stat_smooth(method = "lm", col = "blue")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



We get a fantastic p value and a pretty compelling graph that indeed drawing more cards is associated with winning more often. Admittedly if we include the outliers we get significantly worse results.

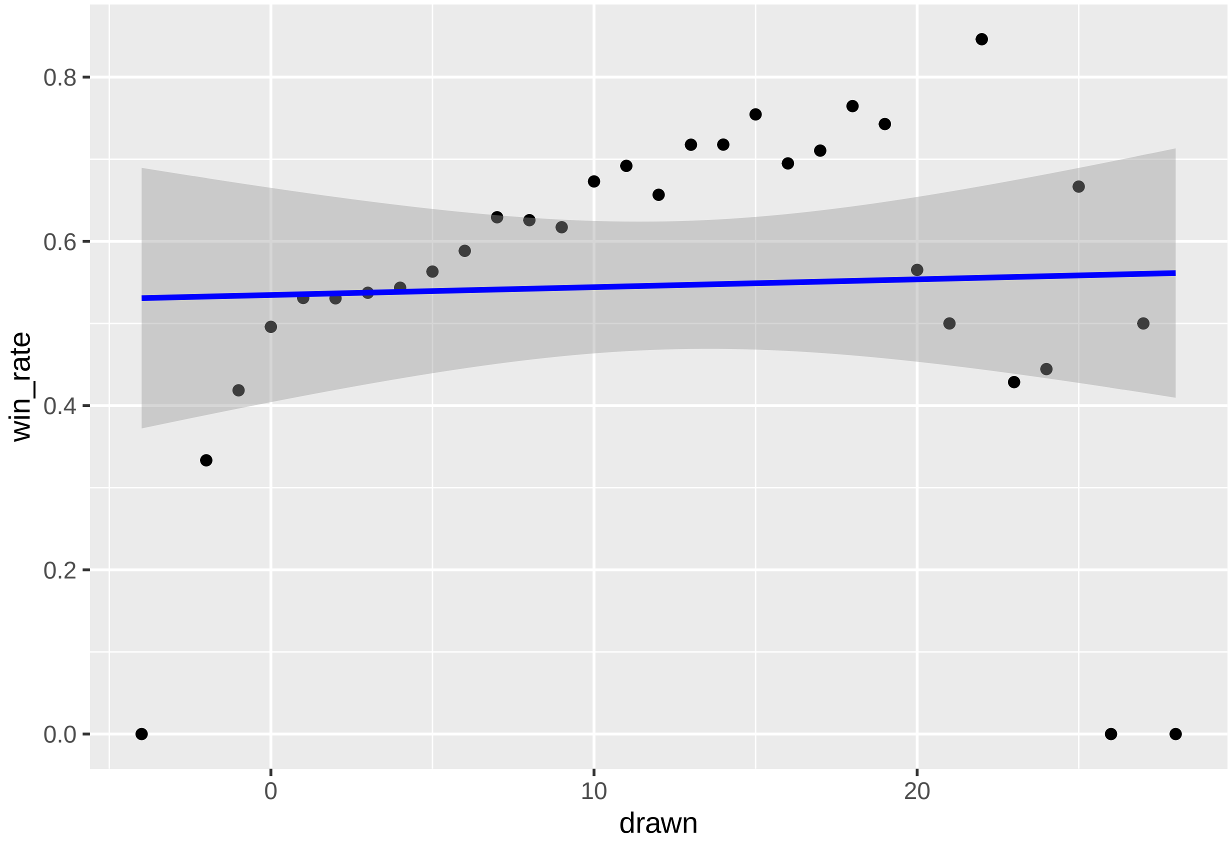
```
summary(lm(theory$drawn~theory$win_rate))
```

```
##
## Call:
## lm(formula = theory$drawn ~ theory$win_rate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.4411  -7.7407  -0.2332   7.7768  16.5589
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       11.441      4.739   2.414   0.0221 *
## theory$win_rate    1.880      8.097   0.232   0.8180
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## Residual standard error: 9.583 on 30 degrees of freedom
## Multiple R-squared:  0.001794,   Adjusted R-squared:  -0.03148
## F-statistic: 0.05392 on 1 and 30 DF,  p-value: 0.818
```

```
ggplot(theory,aes(drawn,win_rate)) + geom_point() +
  stat_smooth(method = "lm", col = "blue")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



With the outliers included it would seem that you can draw too many cards. Perhaps there are diminishing returns and this model should be a parabola.