CBB Analysis

John Hankinson

2023-01-15

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.0
## v readr   2.1.3      v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
cbb = read_csv("cbb.csv")
```

```
## Rows: 2455 Columns: 24
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): TEAM, CONF, POSTSEASON
## dbl (21): G, W, ADJOE, ADJDE, BARTHAG, EFG_O, EFG_D, TOR, TORD, ORB, DRB, FT...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(cbb)
```

```
## # A tibble: 6 x 24
##   TEAM       CONF      G     W ADJOE ADJDE BARTHAG EFG_O EFG_D   TOR  TORD   ORB
##   <chr>      <chr> <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 North Car~ ACC      40    33  123.  94.9   0.953  52.6  48.1  15.4  18.2  40.7
## 2 Wisconsin  B10      40    36  129.  93.6   0.976  54.8  47.7  12.4  15.8  32.1
## 3 Michigan   B10      40    33  114.  90.4   0.938  53.9  47.7  14    19.5  25.5
## 4 Texas Tech B12      38    31  115.  85.2   0.970  53.5  43    17.7  22.8  27.4
## 5 Gonzaga    WCC      39    37  118.  86.3   0.973  56.6  41.1  16.2  17.1  30
## 6 Kentucky   SEC      40    29  117.  96.2   0.906  49.9  46    18.1  16.1  42
## # ... with 12 more variables: DRB <dbl>, FTR <dbl>, FTRD <dbl>, `2P_O` <dbl>,
## #   `2P_D` <dbl>, `3P_O` <dbl>, `3P_D` <dbl>, ADJ_T <dbl>, WAB <dbl>,
## #   POSTSEASON <chr>, SEED <dbl>, YEAR <dbl>
## # i Use `colnames()` to see all variable names
```

```
cbb_filtered = select(cbb, -c('WAB', 'SEED'))

cbb_filtered$WPCT = cbb_filtered$W / cbb_filtered$G
cbb_filtered = select(cbb_filtered, -c('G', 'W'))

p5 = c('ACC', 'SEC', 'B10', 'B12', 'P12')
cbb_filtered$P5 = as.numeric(cbb_filtered$CONF %in% p5)

cbb_filtered$POSTSEASON = as.numeric(!is.na(cbb_filtered$POSTSEASON))

cbb_filtered = unite(cbb_filtered, 'TEAMYEAR', c('TEAM', 'YEAR'), remove=TRUE)
```

```
length(unique(cbb$TEAM))
```

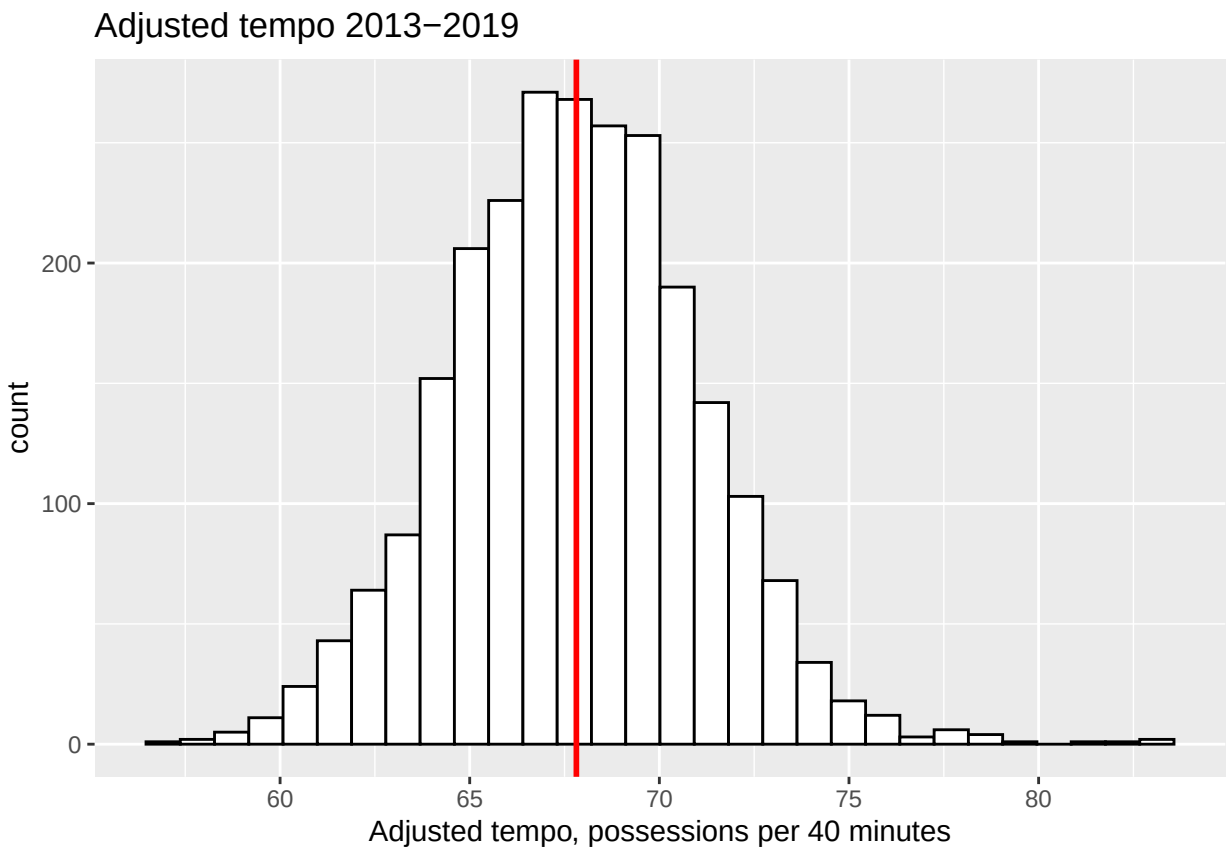
```
## [1] 355
```

```
summary(cbb_filtered)
```

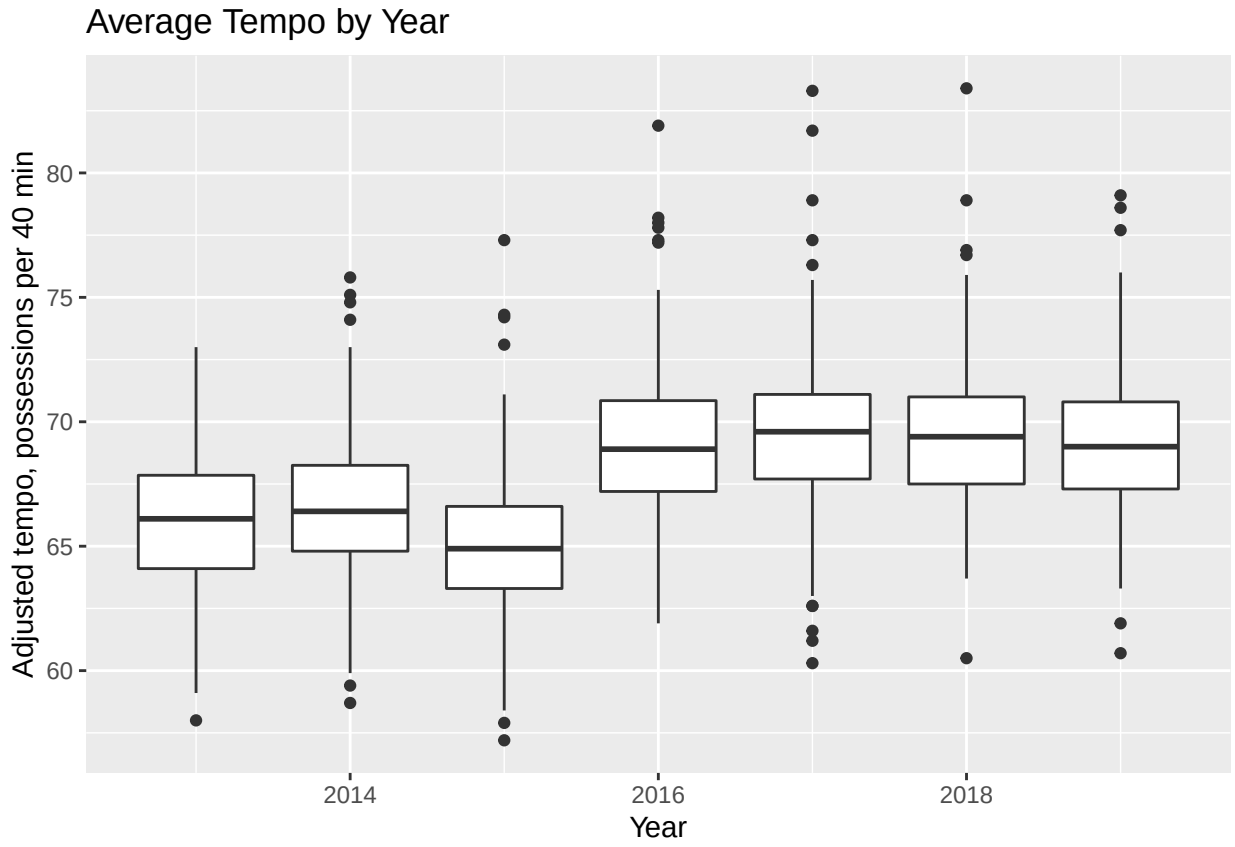
```
##    TEAMYEAR             CONF               ADJOE           ADJDE
##  Length:2455        Length:2455        Min.   : 76.6   Min.   : 84.0
##  Class :character   Class :character   1st Qu.: 98.3   1st Qu.: 98.5
##  Mode  :character   Mode  :character   Median :103.0   Median :103.5
##                                        Mean   :103.3   Mean   :103.3
##                                        3rd Qu.:108.0   3rd Qu.:107.9
##                                        Max.   :129.1   Max.   :124.0
##     BARTHAG           EFG_O           EFG_D           TOR
##  Min.   :0.0050   Min.   :39.20   Min.   :39.6   Min.   :11.90
##  1st Qu.:0.2822   1st Qu.:47.75   1st Qu.:48.0   1st Qu.:17.30
##  Median :0.4750   Median :49.70   Median :50.0   Median :18.70
##  Mean   :0.4940   Mean   :49.81   Mean   :50.0   Mean   :18.76
##  3rd Qu.:0.7122   3rd Qu.:51.90   3rd Qu.:52.0   3rd Qu.:20.10
##  Max.   :0.9842   Max.   :59.80   Max.   :59.5   Max.   :27.10
##       TORD             ORB             DRB             FTR
##  Min.   :10.20   Min.   :15.00   Min.   :18.40   Min.   :21.60
##  1st Qu.:17.20   1st Qu.:27.10   1st Qu.:27.90   1st Qu.:32.40
##  Median :18.60   Median :29.90   Median :30.00   Median :35.80
##  Mean   :18.69   Mean   :29.88   Mean   :30.08   Mean   :35.99
##  3rd Qu.:20.10   3rd Qu.:32.60   3rd Qu.:32.20   3rd Qu.:39.50
##  Max.   :28.50   Max.   :43.60   Max.   :40.40   Max.   :58.60
##       FTRD             2P_O            2P_D            3P_O            3P_D
##  Min.   :21.80   Min.   :37.7   Min.   :37.70   Min.   :24.90   Min.   :27.1
##  1st Qu.:31.90   1st Qu.:46.5   1st Qu.:46.70   1st Qu.:32.50   1st Qu.:33.0
##  Median :35.80   Median :48.7   Median :49.00   Median :34.40   Median :34.6
##  Mean   :36.27   Mean   :48.8   Mean   :48.98   Mean   :34.41   Mean   :34.6
##  3rd Qu.:40.20   3rd Qu.:51.0   3rd Qu.:51.30   3rd Qu.:36.30   3rd Qu.:36.2
##  Max.   :60.70   Max.   :62.6   Max.   :61.20   Max.   :44.10   Max.   :43.1
##      ADJ_T         POSTSEASON          WPCT              P5
##  Min.   :57.20   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:65.70   1st Qu.:0.0000   1st Qu.:0.3793   1st Qu.:0.0000
##  Median :67.80   Median :0.0000   Median :0.5161   Median :0.0000
##  Mean   :67.81   Mean   :0.1939   Mean   :0.5082   Mean   :0.1825
##  3rd Qu.:70.00   3rd Qu.:0.0000   3rd Qu.:0.6364   3rd Qu.:0.0000
##  Max.   :83.40   Max.   :1.0000   Max.   :0.9744   Max.   :1.0000
```

```
ggplot(cbb, aes(x=ADJ_T)) +
  geom_histogram(color='black', fill='white') +
  geom_vline(aes(xintercept=mean(ADJ_T)), color='red', size=1) +
  xlab('Adjusted tempo, possessions per 40 minutes') +
  labs(title='Adjusted tempo 2013-2019')
```

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




```
ggplot(cbb, aes(x=YEAR, y=ADJ_T)) +
  geom_boxplot(aes(group=cbb$YEAR)) +
  xlab('Year') +
  ylab('Adjusted tempo, possessions per 40 min') +
  labs(title='Average Tempo by Year')
```




```
cbb_reg = lm(WPCT ~ ., data = select(cbb_filtered, -c('TEAMYEAR', 'CONF', 'POSTSEASON')))
summary(cbb_reg)
```

```
##
## Call:
## lm(formula = WPCT ~ ., data = select(cbb_filtered, -c("TEAMYEAR",
##     "CONF", "POSTSEASON")))
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.207109 -0.044799 -0.000735  0.044140  0.261892
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0275793  0.0694930   0.397  0.69150
## ADJOE       -0.0028496  0.0009486  -3.004  0.00269 **
## ADJDE        0.0054816  0.0009657   5.676 1.54e-08 ***
## BARTHAG      0.1945064  0.0311411   6.246 4.95e-10 ***
## EFG_O        0.0327931  0.0051350   6.386 2.03e-10 ***
## EFG_D       -0.0327068  0.0068479  -4.776 1.89e-06 ***
## TOR         -0.0210585  0.0011481 -18.342  < 2e-16 ***
## TORD         0.0272727  0.0010675  25.548  < 2e-16 ***
## ORB          0.0094391  0.0005540  17.038  < 2e-16 ***
## DRB         -0.0127869  0.0006504 -19.659  < 2e-16 ***
## FTR          0.0029550  0.0002935  10.067  < 2e-16 ***
```

```
## FTRD        -0.0042312  0.0002699 -15.678  < 2e-16 ***
## `2P_O`      -0.0060225  0.0032618  -1.846  0.06496 .
## `2P_D`       0.0056741  0.0044207   1.284  0.19943
## `3P_O`      -0.0031558  0.0026785  -1.178  0.23884
## `3P_D`       0.0006080  0.0036050   0.169  0.86608
## ADJ_T        0.0038257  0.0004409   8.678  < 2e-16 ***
## P5          -0.0329271  0.0047562  -6.923 5.64e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06564 on 2437 degrees of freedom
## Multiple R-squared:  0.8676, Adjusted R-squared:  0.8666
## F-statistic:   939 on 17 and 2437 DF,  p-value: < 2.2e-16
```

```
cor(cbb_filtered[,3:21])[,16]
```

```
##        ADJOE        ADJDE      BARTHAG        EFG_O        EFG_D          TOR
##  0.077011085  0.224615624 -0.077317869  0.134884606  0.285776586 -0.113341424
##         TORD          ORB          DRB          FTR         FTRD         2P_O
## -0.075896026 -0.116339088 -0.026523918 -0.020590078 -0.033356587  0.179804125
##         2P_D         3P_O         3P_D        ADJ_T   POSTSEASON         WPCT
##  0.286825549  0.020946356  0.162985819  1.000000000 -0.031553264  0.007614593
##           P5
## -0.030834182
```

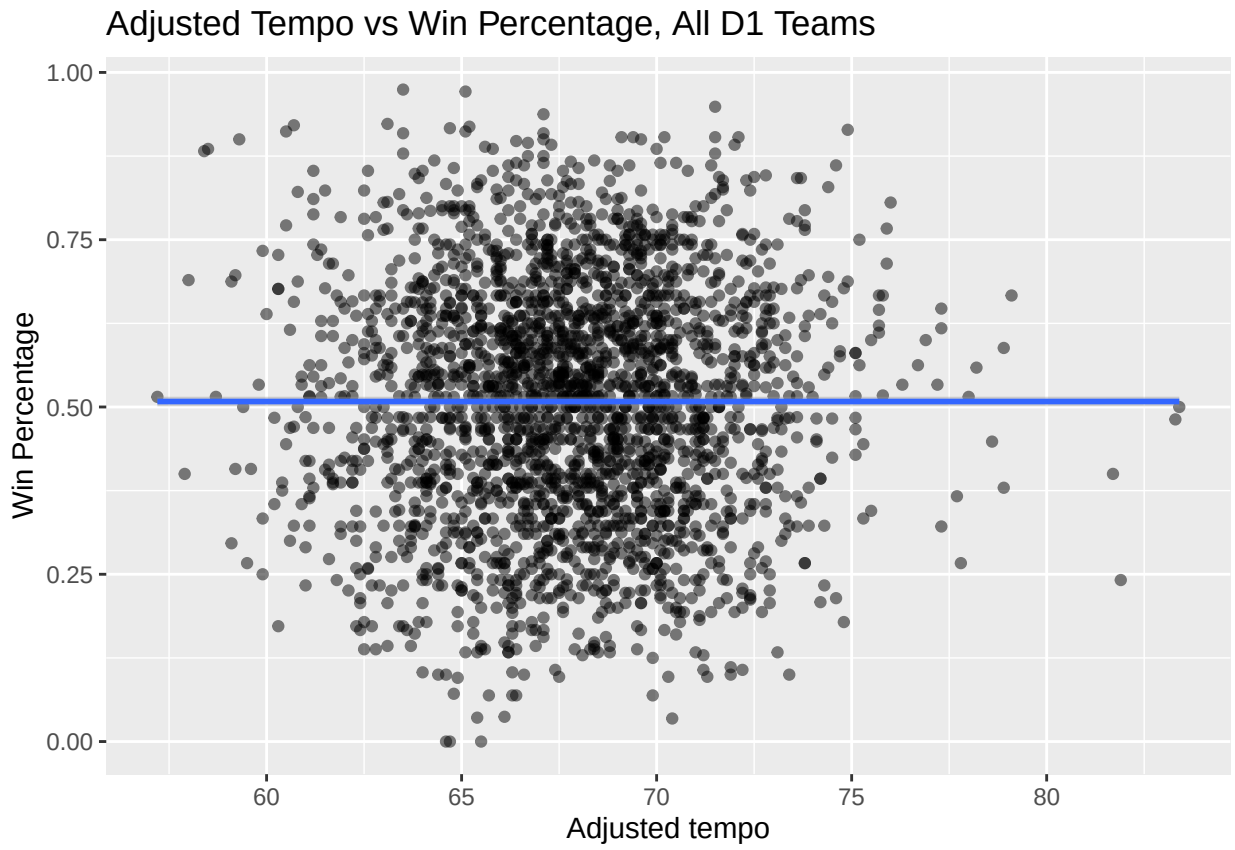
```
summary(lm(WPCT~ADJ_T + BARTHAG, data=cbb_filtered))
```

```
##
## Call:
## lm(formula = WPCT ~ ADJ_T + BARTHAG, data = cbb_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34088 -0.07980 -0.00129  0.07966  0.38756
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0236210  0.0465750  -0.507    0.612
## ADJ_T        0.0037848  0.0006783   5.580 2.67e-08 ***
## BARTHAG      0.5570641  0.0086756  64.210  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1098 on 2452 degrees of freedom
## Multiple R-squared:  0.6271, Adjusted R-squared:  0.6268
## F-statistic:  2062 on 2 and 2452 DF,  p-value: < 2.2e-16
```

```
cbb_filtered%>%
  ggplot(aes(x=ADJ_T, y=WPCT)) +
  geom_point(alpha=0.5) +
  geom_smooth() +
  xlab('Adjusted tempo') +
```

```
  ylab('Win Percentage') +
  labs(title='Adjusted Tempo vs Win Percentage, All D1 Teams')
```

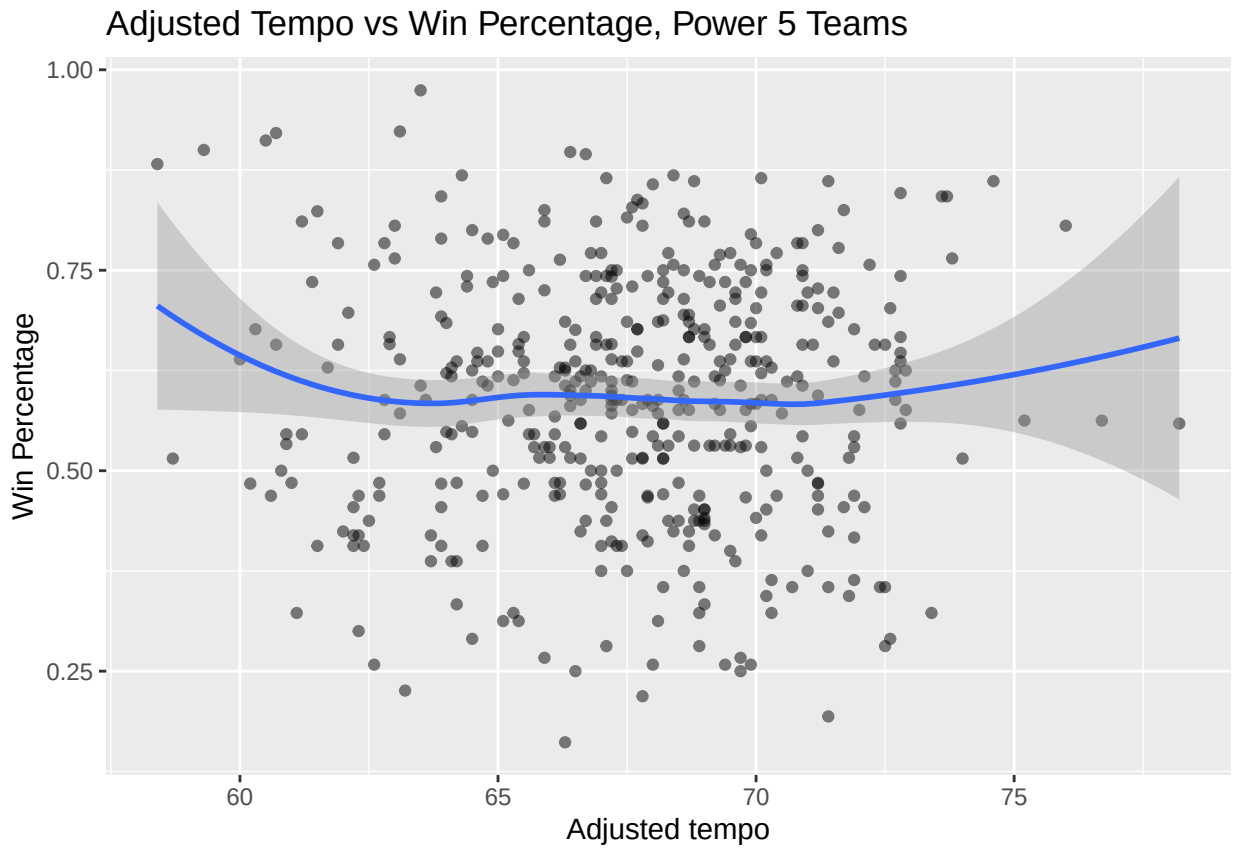
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
cbb_p5 = cbb_filtered[which(cbb_filtered$P5 == 1),1:20]

cbb_p5%>%
  ggplot(aes(x=ADJ_T, y=WPCT)) +
  geom_point(alpha=0.5) +
  geom_smooth() +
  xlab('Adjusted tempo') +
  ylab('Win Percentage') +
  labs(title='Adjusted Tempo vs Win Percentage, Power 5 Teams')
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



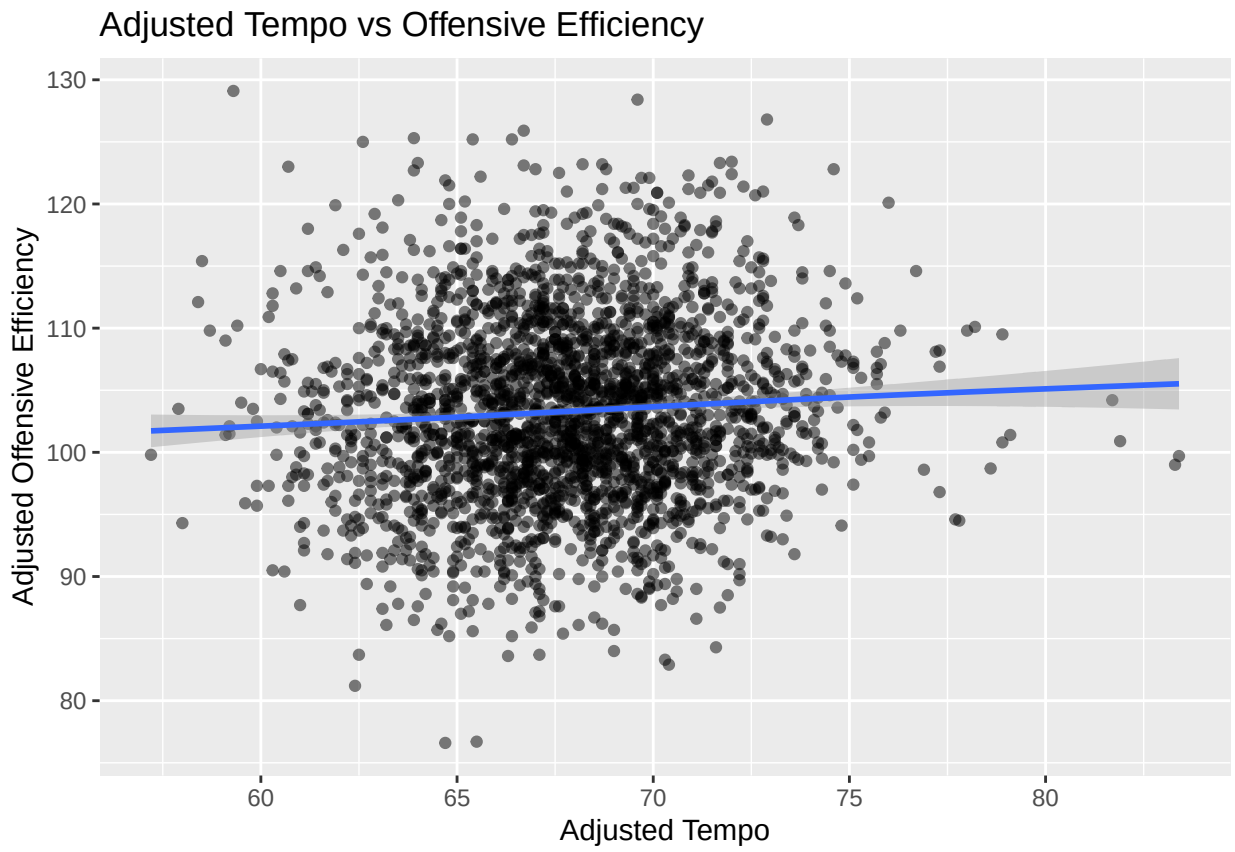

```
summary(lm(ADJOE~ADJ_T + BARTHAG, data=cbb_filtered))
```

```
##
## Call:
## lm(formula = ADJOE ~ ADJ_T + BARTHAG, data = cbb_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.6520  -2.3951  -0.0315   2.3810  16.4718
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 68.87130    1.53037   45.00   <2e-16 ***
## ADJ_T        0.32500    0.02229   14.58   <2e-16 ***
## BARTHAG     25.09154    0.28507   88.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.608 on 2452 degrees of freedom
## Multiple R-squared:  0.761,  Adjusted R-squared:  0.7608
## F-statistic:  3904 on 2 and 2452 DF,  p-value: < 2.2e-16
```

```
cbb_filtered %>%
  ggplot(aes(x=ADJ_T, y=ADJOE)) +
  geom_point(alpha=0.5) +
```

```
  geom_smooth() +
  xlab('Adjusted Tempo') +
  ylab('Adjusted Offensive Efficiency') +
  labs(title='Adjusted Tempo vs Offensive Efficiency')
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



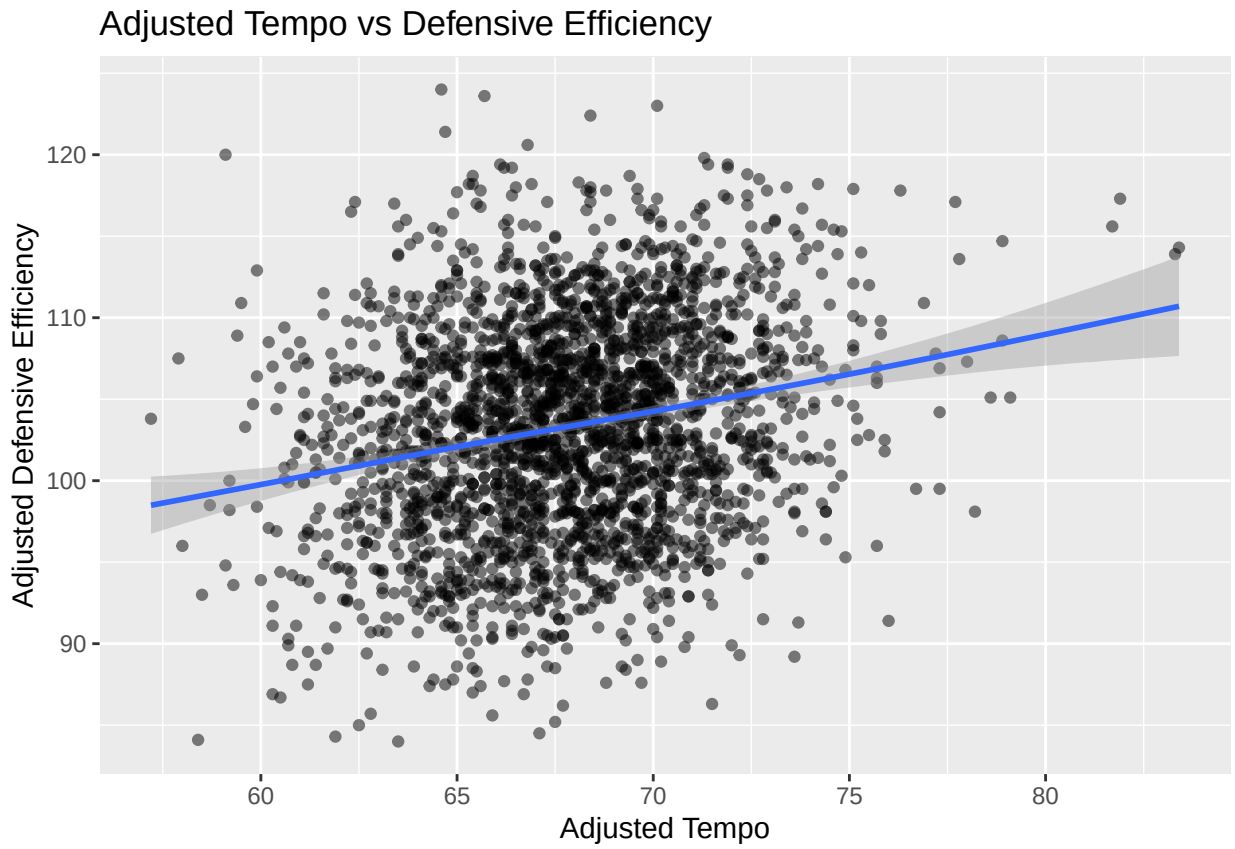

```
summary(lm(ADJDE~ADJ_T + BARTHAG, data=cbb_filtered))
```

```
##
## Call:
## lm(formula = ADJDE ~ ADJ_T + BARTHAG, data = cbb_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.072  -2.169  -0.061   2.244  12.205
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  91.83273    1.46413   62.72   <2e-16 ***
## ADJ_T         0.32410    0.02132   15.20   <2e-16 ***
## BARTHAG     -21.26898    0.27273  -77.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3.452 on 2452 degrees of freedom
## Multiple R-squared:  0.7272, Adjusted R-squared:  0.7269
## F-statistic:  3268 on 2 and 2452 DF,  p-value: < 2.2e-16
```

```
cbb_filtered %>%
  ggplot(aes(x=ADJ_T, y=ADJDE)) +
  geom_point(alpha=0.5) +
  geom_smooth() +
  xlab('Adjusted Tempo') +
  ylab('Adjusted Defensive Efficiency') +
  labs(title='Adjusted Tempo vs Defensive Efficiency')
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



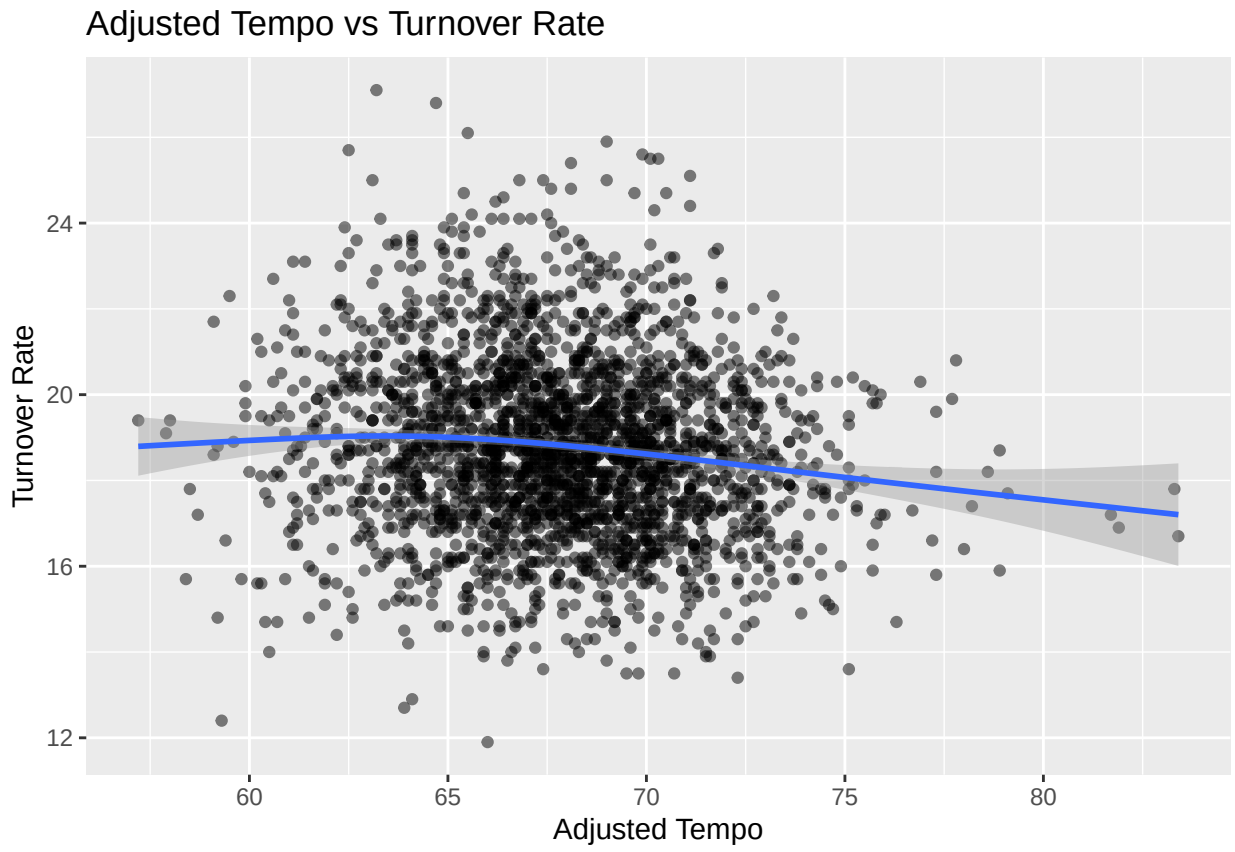

```
summary(lm(TOR~ADJ_T + BARTHAG, data=cbb_filtered))
```

```
##
## Call:
## lm(formula = TOR ~ ADJ_T + BARTHAG, data = cbb_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1094 -1.2279 -0.0752  1.2064  7.3636
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.12589    0.77762  34.883   <2e-16 ***
## ADJ_T       -0.09544    0.01132  -8.428   <2e-16 ***
## BARTHAG     -3.82850    0.14485 -26.431   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.833 on 2452 degrees of freedom
## Multiple R-squared:  0.2317, Adjusted R-squared:  0.2311
## F-statistic: 369.8 on 2 and 2452 DF,  p-value: < 2.2e-16
```

```
cbb_filtered %>%
  ggplot(aes(x=ADJ_T, y=TOR)) +
  geom_point(alpha=0.5) +
  geom_smooth() +
  xlab('Adjusted Tempo') +
  ylab('Turnover Rate') +
  labs(title='Adjusted Tempo vs Turnover Rate')
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

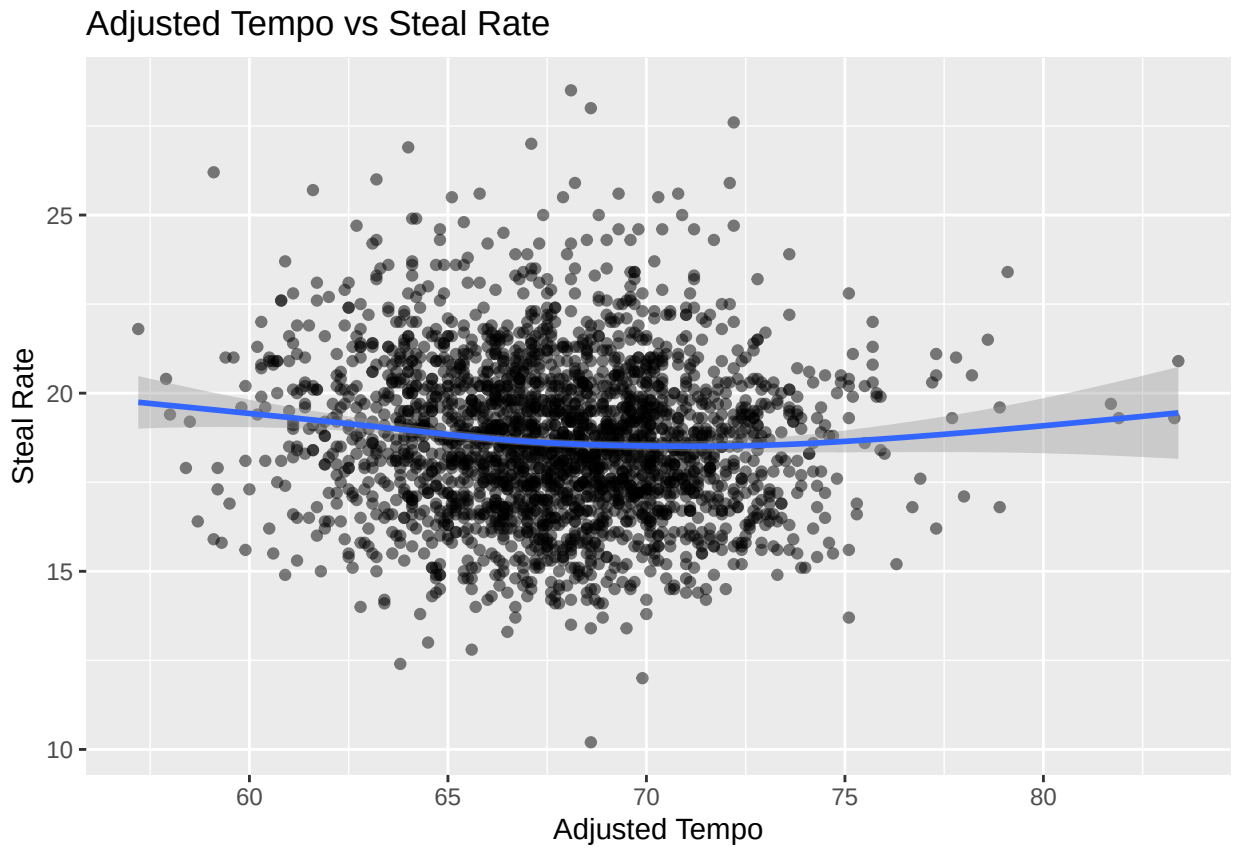


```r
summary(lm(TORD~ADJ_T + BARTHAG, data=cbb_filtered))
```

```
##
## Call:
## lm(formula = TORD ~ ADJ_T + BARTHAG, data = cbb_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4330 -1.4768 -0.0775  1.3873  9.6429
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.76138    0.93050  23.387  < 2e-16 ***
## ADJ_T       -0.04841    0.01355  -3.572 0.000361 ***
## BARTHAG      0.42649    0.17333   2.461 0.013939 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.194 on 2452 degrees of freedom
## Multiple R-squared:  0.008209,   Adjusted R-squared:  0.0074
## F-statistic: 10.15 on 2 and 2452 DF,  p-value: 4.083e-05
```

```r
cbb_filtered %>%
  ggplot(aes(x=ADJ_T, y=TORD)) +
  geom_point(alpha=0.5) +
  geom_smooth() +
  xlab('Adjusted Tempo') +
  ylab('Steal Rate') +
  labs(title='Adjusted Tempo vs Steal Rate')
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



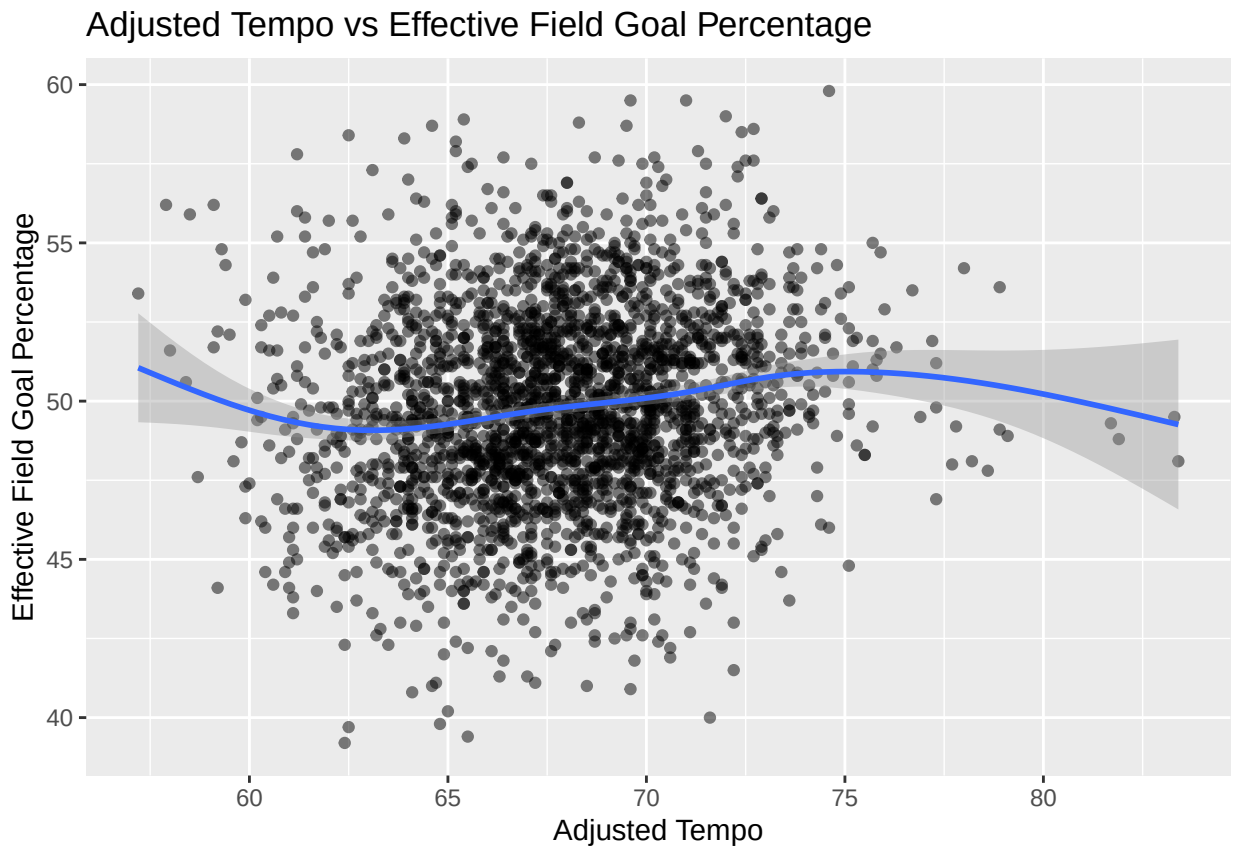

```
summary(lm(EFG_O~ADJ_T + BARTHAG, data=cbb_filtered))
```

```
##
## Call:
## lm(formula = EFG_O ~ ADJ_T + BARTHAG, data = cbb_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3668 -1.8058 -0.0603  1.7128  9.2367
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.95307    1.10213   31.71   <2e-16 ***
## ADJ_T        0.17001    0.01605   10.59   <2e-16 ***
## BARTHAG      6.72751    0.20530   32.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.598 on 2452 degrees of freedom
## Multiple R-squared:  0.3172, Adjusted R-squared:  0.3167
## F-statistic: 569.6 on 2 and 2452 DF,  p-value: < 2.2e-16
```

```
cbb_filtered %>%
  ggplot(aes(x=ADJ_T, y=EFG_O)) +
  geom_point(alpha=0.5) +
```

```
  geom_smooth() +
  xlab('Adjusted Tempo') +
  ylab('Effective Field Goal Percentage') +
  labs(title='Adjusted Tempo vs Effective Field Goal Percentage')
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



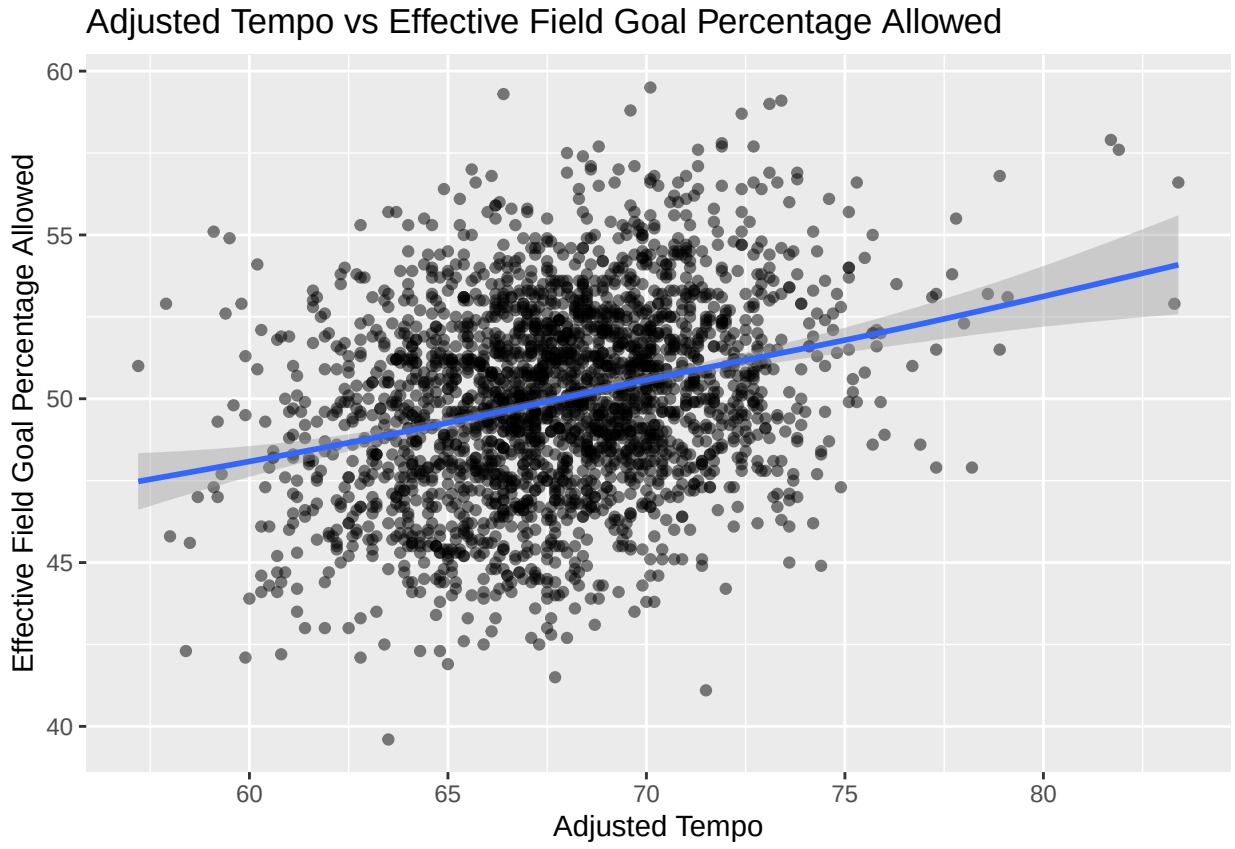

```
summary(lm(EFG_D~ADJ_T + BARTHAG, data=cbb_filtered))
```

```
##
## Call:
## lm(formula = EFG_D ~ ADJ_T + BARTHAG, data = cbb_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6211 -1.4032  0.0011  1.4908  6.9362
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.81643    0.93638   41.45   <2e-16 ***
## ADJ_T        0.21487    0.01364   15.76   <2e-16 ***
## BARTHAG     -6.85464    0.17442  -39.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.207 on 2452 degrees of freedom
## Multiple R-squared:  0.4366, Adjusted R-squared:  0.4361
## F-statistic: 949.9 on 2 and 2452 DF,  p-value: < 2.2e-16
```

```
cbb_filtered %>%
  ggplot(aes(x=ADJ_T, y=EFG_D)) +
  geom_point(alpha=0.5) +
  geom_smooth() +
  xlab('Adjusted Tempo') +
  ylab('Effective Field Goal Percentage Allowed') +
  labs(title='Adjusted Tempo vs Effective Field Goal Percentage Allowed')
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



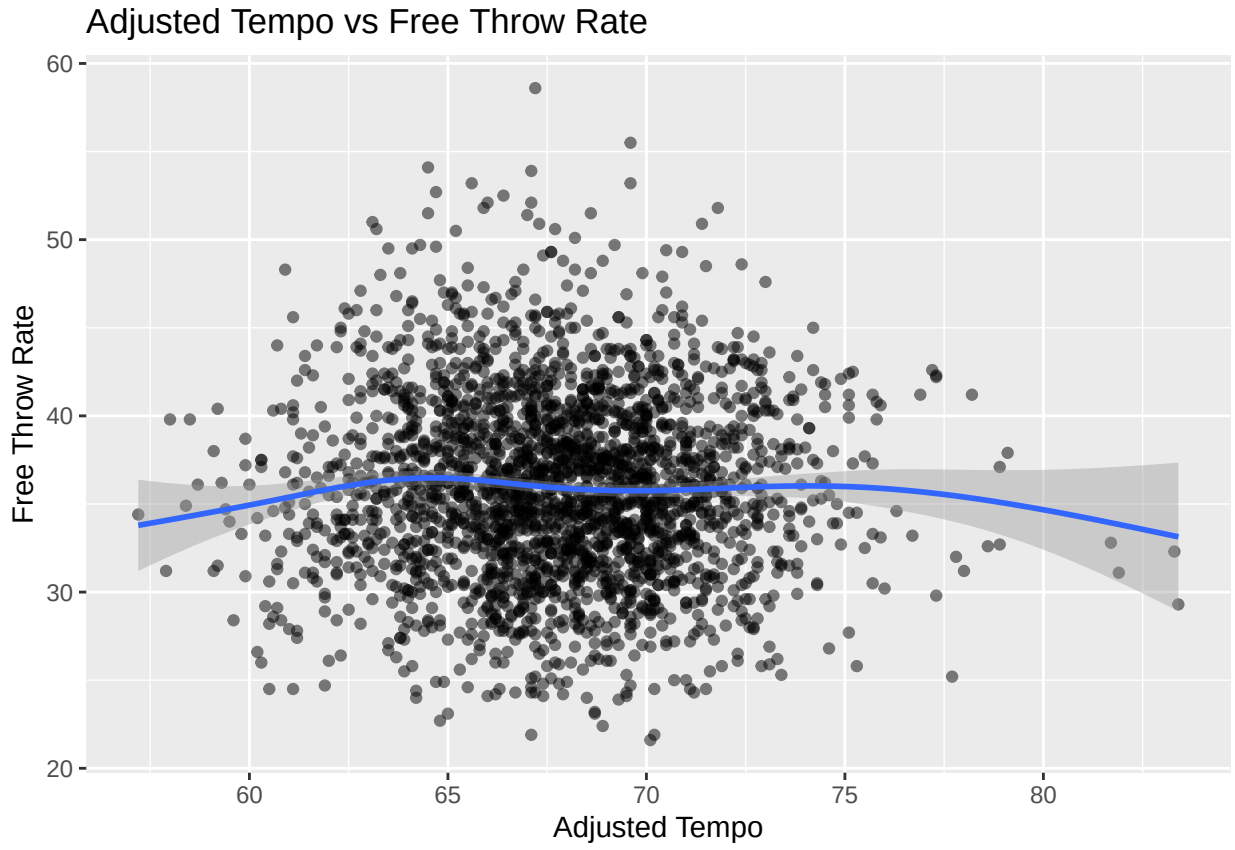

```
summary(lm(FTR~ADJ_T + BARTHAG, data=cbb_filtered))
```

```
##
## Call:
## lm(formula = FTR ~ ADJ_T + BARTHAG, data = cbb_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3972  -3.6376  -0.1951   3.5038  23.6732
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.77865    2.20710  16.211  < 2e-16 ***
## ADJ_T       -0.01660    0.03214  -0.517    0.605
## BARTHAG      2.70693    0.41112   6.584 5.57e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.203 on 2452 degrees of freedom
## Multiple R-squared:  0.01779,    Adjusted R-squared:  0.01699
## F-statistic: 22.21 on 2 and 2452 DF,  p-value: 2.771e-10
```

```
cbb_filtered %>%
  ggplot(aes(x=ADJ_T, y=FTR)) +
  geom_point(alpha=0.5) +
  geom_smooth() +
  xlab('Adjusted Tempo') +
  ylab('Free Throw Rate') +
  labs(title='Adjusted Tempo vs Free Throw Rate')
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

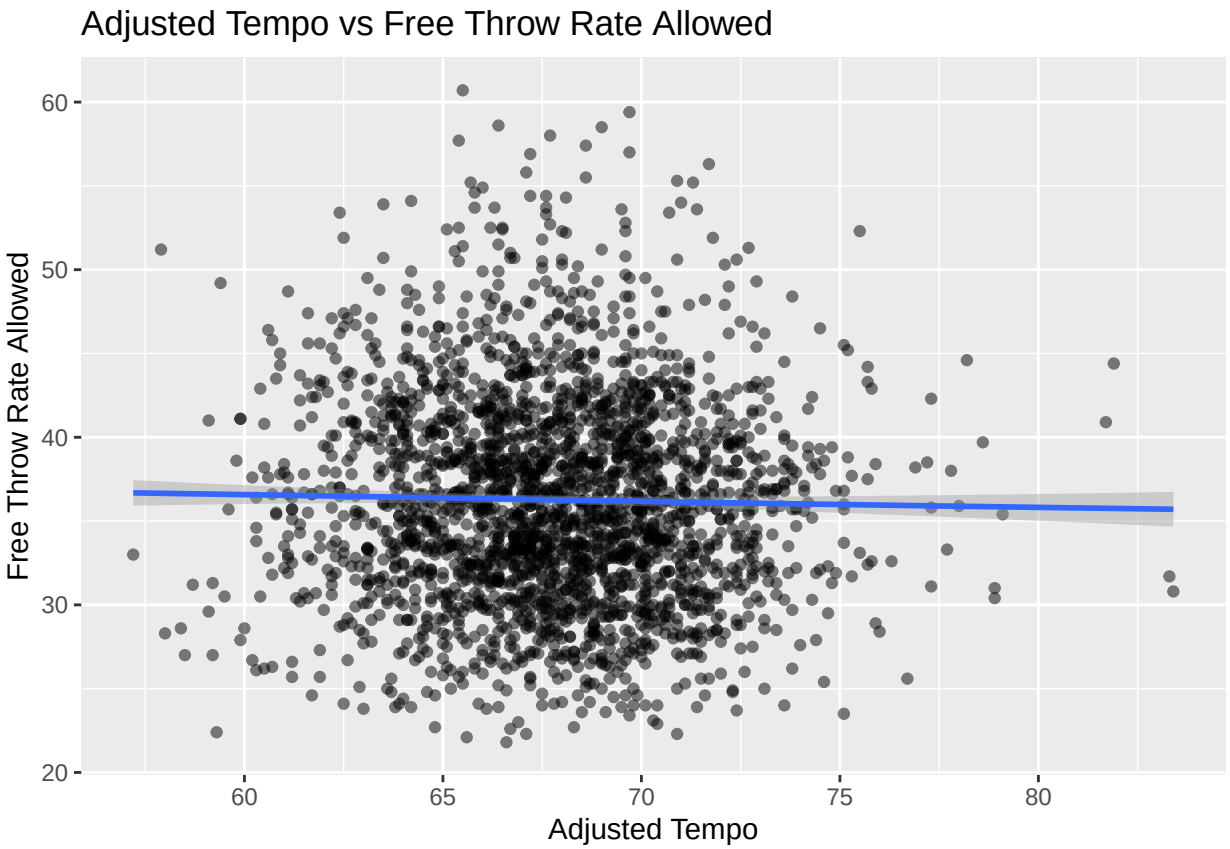


```r
summary(lm(FTRD~ADJ_T + BARTHAG, data=cbb_filtered))
```

```
##
## Call:
## lm(formula = FTRD ~ ADJ_T + BARTHAG, data = cbb_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.9474  -4.2901  -0.4635   3.8894  22.8507
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 47.63615    2.51025   18.98  < 2e-16 ***
## ADJ_T       -0.11075    0.03656   -3.03  0.00247 **
## BARTHAG     -7.80616    0.46759  -16.69  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.918 on 2452 degrees of freedom
## Multiple R-squared:  0.1031, Adjusted R-squared:  0.1023
## F-statistic: 140.9 on 2 and 2452 DF,  p-value: < 2.2e-16
```

```r
cbb_filtered %>%
  ggplot(aes(x=ADJ_T, y=FTRD)) +
  geom_point(alpha=0.5) +
  geom_smooth() +
  xlab('Adjusted Tempo') +
  ylab('Free Throw Rate Allowed') +
  labs(title='Adjusted Tempo vs Free Throw Rate Allowed')
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```


Adjusted Tempo vs Free Throw Rate Allowed
Free Throw Rate Allowed
60
50
40
30
20
60
65
70
75
80
Adjusted Tempo