Mid-Quarter Findings

Shail Mirpuri

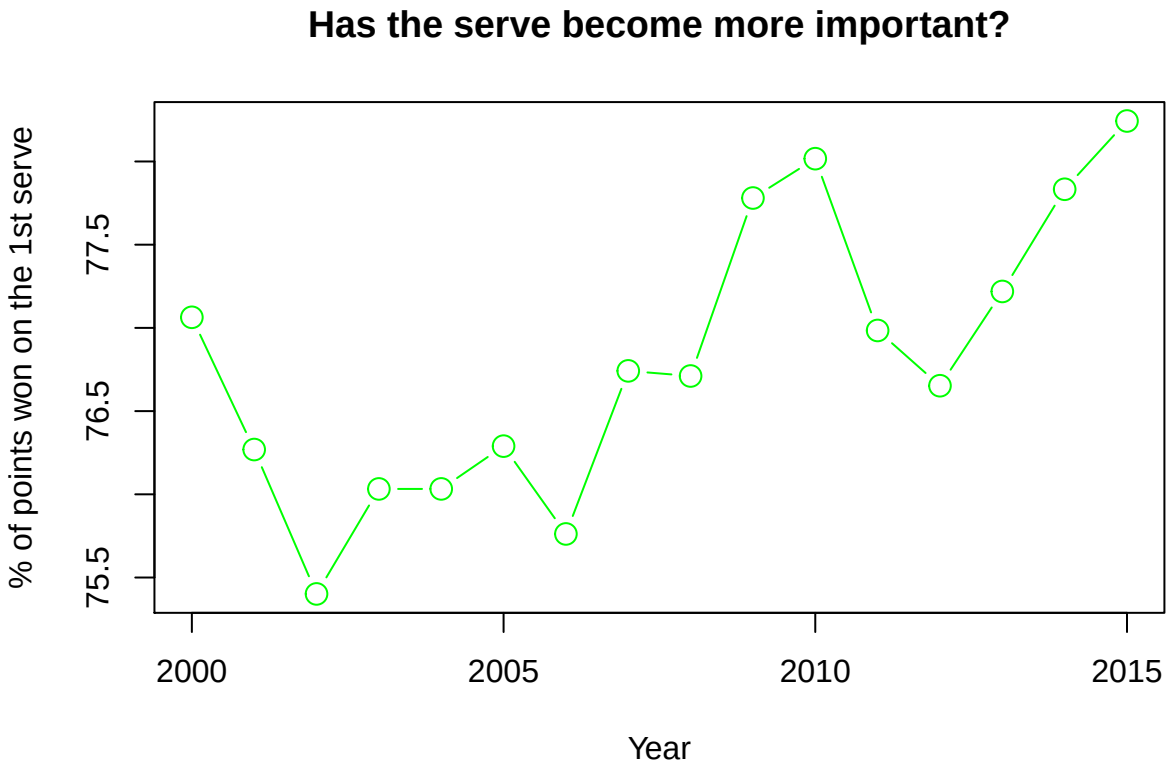
2/5/2021

Exploring the evolution of the serve since 2000

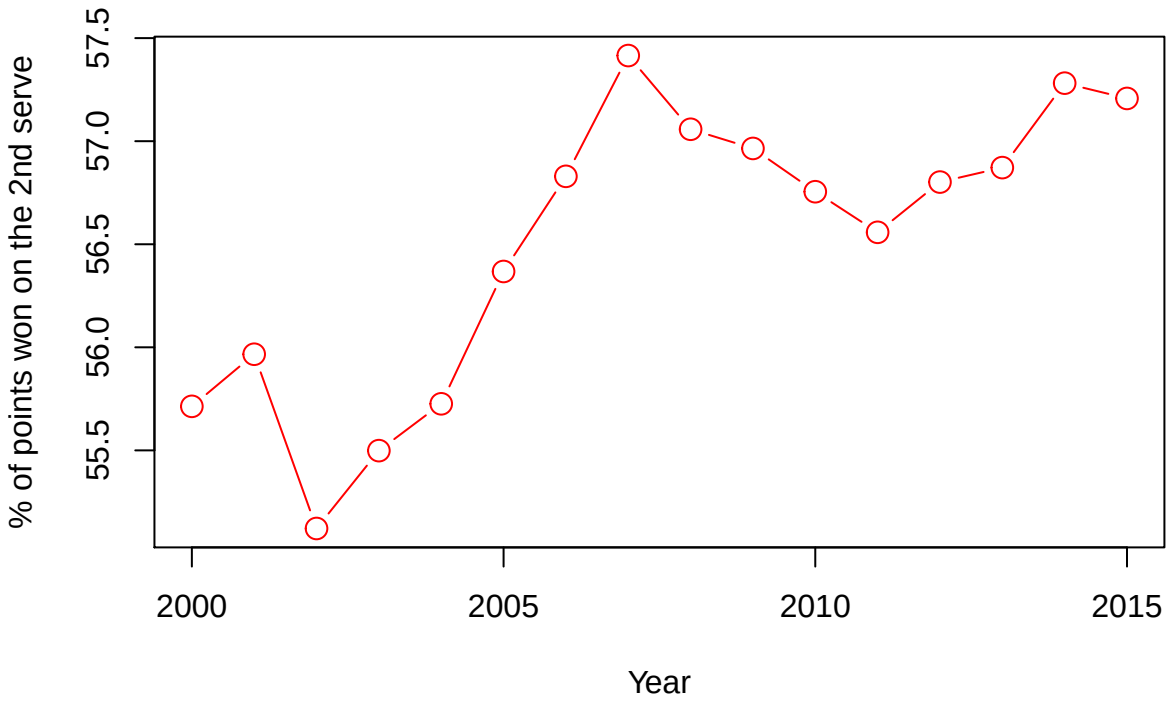
```
find_year <- function(x) {
  as.numeric(paste(unlist(str_split(x, '-'))[1:4], collapse = ''))
}
df$year <- vapply(df$tourney_date, FUN = find_year, numeric(1))
tb <- tibble(df)

evol<- tb %>% group_by(year) %>%
  summarise(winner_aces = mean(w_ace, na.rm = TRUE),
            loser_aces = mean(l_ace, na.rm = TRUE),
            win_df = mean(w_df, na.rm = TRUE),
            lose_df = mean(l_df, na.rm = TRUE),
            win_sp = mean(w_svpt, na.rm = TRUE),
            l_sp = mean(l_svpt, na.rm = TRUE),
            w_1stserve = mean(w_1stWon, na.rm = TRUE),
            w_2ndserve = mean(w_2ndWon, na.rm = TRUE))
evol
```

```
## # A tibble: 16 x 9
##     year winner_aces loser_aces win_df lose_df win_sp  l_sp w_1stserve
##  * <dbl>       <dbl>      <dbl>  <dbl>   <dbl>  <dbl> <dbl>      <dbl>
##  1  2000       10.2        7.76   4.64    5.52   111.  115.       77.1
##  2  2001        9.76       7.48   4.53    5.68   111.  116.       76.3
##  3  2002        9.28       6.71   4.60    5.64   114.  118.       75.4
##  4  2003        9.32       6.79   4.21    5.43   109.  115.       76.0
##  5  2004        9.60       6.98   4.03    5.37   109.  114.       76.0
##  6  2005        9.24       7.02   3.82    4.92   108.  114.       76.3
##  7  2006        8.62       6.92   3.07    4.34   109.  113.       75.8
##  8  2007        9.33       7.11   2.95    4.07   106.  112.       76.7
##  9  2008        9.93       7.14   3.23    4.21   108.  114.       76.7
## 10  2009       10.4        6.90   3.14    4.07   106.  112.       77.8
## 11  2010       11.2        7.86   3.47    4.54   108.  114.       78.0
## 12  2011        9.59       6.45   3.09    4.15   104.  110.       77.0
## 13  2012       10.4        7.05   3.36    4.39   110.  115.       76.7
## 14  2013       10.3        7.30   3.29    4.19   108.  114.       77.2
## 15  2014       11.1        7.85   3.52    4.4    107.  113.       77.8
## 16  2015       11.4        7.89   3.64    4.58   108.  114.       78.2
## # ... with 1 more variable: w_2ndserve <dbl>
```

```
plot(evol$year,evol$w_1stserve, type = 'b', cex = 1.5, col = 'green', ylab = '% of points won on the 1st
     xlab = 'Year', main = 'Has the serve become more important?')
```




```
plot(evol$year,evol$w_2ndserve, type = 'b', cex = 1.5, col = 'red', ylab = '% of points won on the 2nd s
     xlab = 'Year')
```



The serve seems to growing in its importance as head into the modern era with names like Milos Raonic, and Nick Kyrgios boasting some amazing serve records.

Exploring the difference between surfaces

```
surf<- tb %>% group_by(surface) %>%
  summarise(winner_aces = mean(w_ace, na.rm = TRUE),
            loser_aces = mean(l_ace, na.rm = TRUE),
            win_df = mean(w_df, na.rm = TRUE),
            lose_df = mean(l_df, na.rm = TRUE),
            win_sp = mean(w_svpt, na.rm = TRUE),
            l_sp = mean(l_svpt, na.rm = TRUE),
            w_1stserve = mean(w_1stWon, na.rm = TRUE),
            w_2ndserve = mean(w_2ndWon, na.rm = TRUE),
            l_1stserve = mean(l_1stWon, na.rm = TRUE),
            l_2ndserve = mean(l_2ndWon, na.rm = TRUE))
surf
```

```
## # A tibble: 3 x 11
##   surface winner_aces loser_aces win_df lose_df win_sp  l_sp w_1stserve
## * <chr>         <dbl>      <dbl>  <dbl>   <dbl>  <dbl> <dbl>      <dbl>
## 1 Clay           7.03       5.12   3.13    4.10   108.  113.       74.2
## 2 Grass         12.6        8.90   3.75    4.79   110.  117.       79.2
## 3 Hard          10.2        7.39   3.89    4.99   108.  113.       76.9
## # ... with 3 more variables: w_2ndserve <dbl>, l_1stserve <dbl>,
## #   l_2ndserve <dbl>
```

```
surf %>%
  mutate(ace_diff = winner_aces - loser_aces,
         df_diff = win_df - lose_df, first_serve = w_1stserve - l_1stserve, second_serve = w_2ndserve -
  select(surface, ace_diff, df_diff, first_serve, second_serve, first_second_serve_diff)
```

```
## # A tibble: 3 x 6
##   surface ace_diff df_diff first_serve second_serve first_second_serve_diff
##   <chr>      <dbl>   <dbl>       <dbl>        <dbl>                   <dbl>
## 1 Clay        1.91  -0.974        11.2         11.7                  -0.474
## 2 Grass       3.65  -1.05         10.4         10.5                  -0.0900
## 3 Hard        2.77  -1.10         11.0         11.1                  -0.149
```

We can see that aces seem to matter a lot more in the Wimbledon, while the winning points of your first serves are more important on Clay surfaces.

Comparing the big three vs. other seeded players

```
big_three_w <- df[df$winner_name %in% c("Roger Federer", "Novak Djokovic", "Rafael Nadal"),]
w <- apply(big_three_w[,win_serve], MARGIN = 2, mean, na.rm = TRUE)

big_three_l <- df[df$loser_name %in% c("Roger Federer", "Novak Djokovic", "Rafael Nadal"),]
l <- apply(big_three_l[,lose_serve], MARGIN = 2, mean, na.rm = TRUE)
```

```
big3 <- w - l
big3
```

```
##       w_ace        w_df      w_svpt     w_1stIn    w_1stWon    w_2ndWon
##   0.3728827  -1.6383320 -27.0111111   1.4391027  10.4261071  14.2361392
##     w_SvGms    w_bpconv
##  -2.8872313  12.7924712
```

Again we see here that when the big 'three' win games, the % of points won on their first serve is significantly higher. The second serve is where the big 3 take it to another level when they are playing on form.

```
seeded_w <- df[!(df$winner_name %in% c("Roger Federer", "Novak Djokovic", "Rafael Nadal")) & !is.na(df$
w <- apply(seeded_w[,win_serve], MARGIN = 2, mean, na.rm = TRUE)
seeded_l <- df[!(df$loser_name %in% c("Roger Federer", "Novak Djokovic", "Rafael Nadal")) & !is.na(df$l
l <- apply(seeded_l[,lose_serve], MARGIN = 2, mean, na.rm = TRUE)

seeded <- w - l
seeded
```

```
##      w_ace       w_df     w_svpt    w_1stIn   w_1stWon   w_2ndWon    w_SvGms
##   1.744168  -1.045619 -14.095546   1.035840   9.864113  10.253498  -1.117280
##   w_bpconv
##  12.860329
```

```
unseeded_w <- df[is.na(df$winner_seed),]
w <- apply(unseeded_w[,win_serve], MARGIN = 2, mean, na.rm = TRUE)
unseeded_l <- df[is.na(df$loser_seed),]
l <- apply(unseeded_l[,lose_serve], MARGIN = 2, mean, na.rm = TRUE)
unseeded <- w - l
unseeded
```

```
##      w_ace       w_df     w_svpt    w_1stIn   w_1stWon   w_2ndWon    w_SvGms
##  3.1140778 -0.6905518  1.9845738  1.4673238 10.4783197 10.4614382  1.1815252
##   w_bpconv
## 13.1745326
```

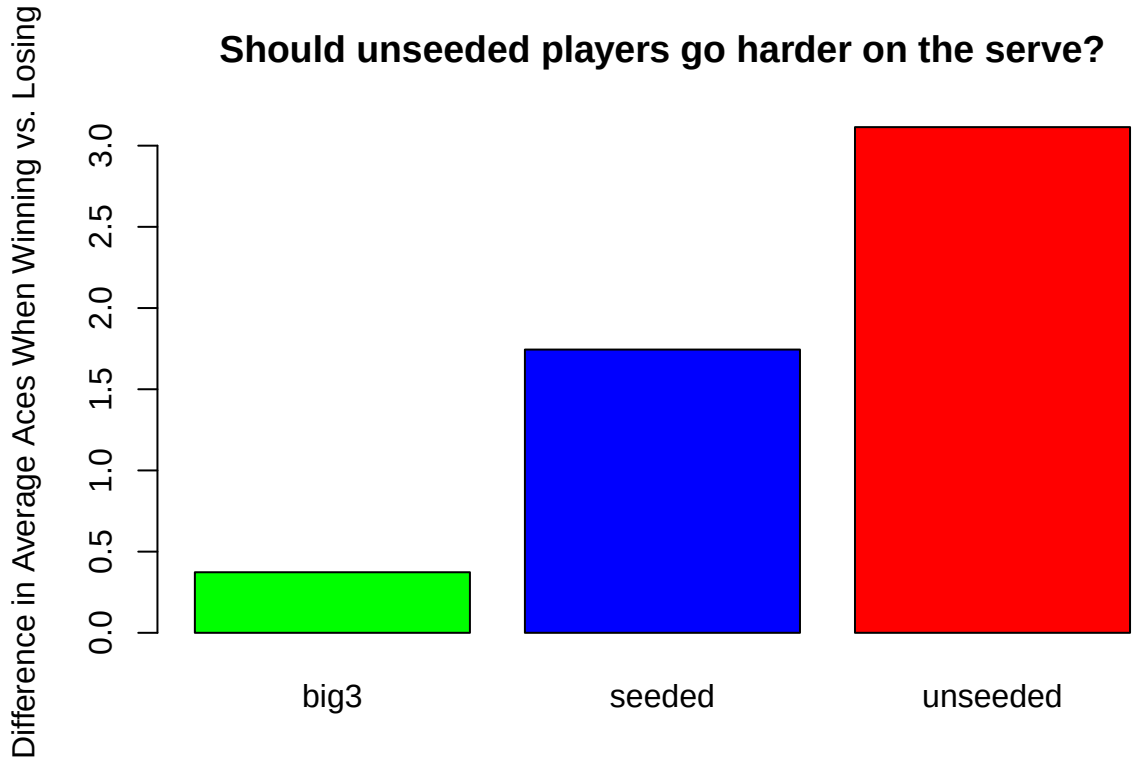
```
compare <- rbind(big3, seeded, unseeded)
compare
```

```
##              w_ace       w_df     w_svpt  w_1stIn  w_1stWon w_2ndWon   w_SvGms
## big3     0.3728827 -1.6383320 -27.011111 1.439103 10.426107 14.23614 -2.887231
## seeded   1.7441678 -1.0456188 -14.095546 1.035840  9.864113 10.25350 -1.117280
## unseeded 3.1140778 -0.6905518   1.984574 1.467324 10.478320 10.46144  1.181525
##          w_bpconv
## big3     12.79247
## seeded   12.86033
## unseeded 13.17453
```

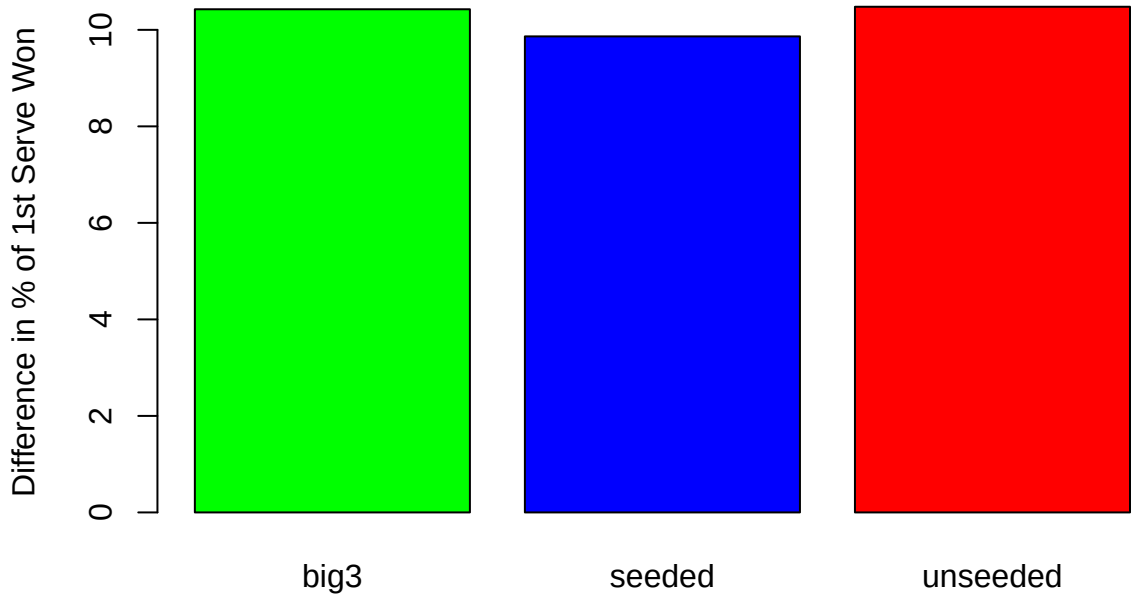
From the pivot table above, we can see that the big3 tend to make less mistakes with their serve (i.e. less double faults) when they are on-form and winning matches than when they are losing matches. Apart from

this another notable difference is the observation that the second serve of the Big 3 goes to a whole new level when they are winning matches, in comparison to the other two groups. Finally, we can also see that as the quality of the player decreases, the number of aces they rely on in order to win a match increases. This suggests that for lower-quality players focusing on serving aces (i.e. going 'hard' on the serve) can lead to greater success.

```
barplot(compare[,1], main = 'Should unseeded players go harder on the serve?', ylab = 'Difference in Av
```




```
barplot(compare[,5], col = c('green', 'blue', 'red'), ylab = 'Difference in % of 1st Serve Won')
```



```
barplot(compare[,6], col = c('green', 'blue', 'red'), ylab = 'Difference in % of 2nd Serve Won')
```