Simona Halep- Trnds and Tournaments 2017

Ioana Dragomirescu

June 26, 2017

Story behind the project

The idea of the project is inspired by the course Big Data - Measuring and Predicting Human Behavior with big data. We want to measure if there is a corellation between Halep's participataions and performances in tournaments (only for 2017) and what often this search-term: Halep appears in the same period.

Practice skills: working with data (extracting data from Internet, writing and reading .csv files in R), working with dates, correlation between variables in a dataframe

Outlook: can we build a model to predict the trends for the next tournament?

```
library(RCurl)
```

```
## Loading required package: bitops
```

```
library(RJSONIO)
library(XML)
```

First we want to have a file with all Tournaments thsi year. Read the tournaments table 2017 from the wtatennis calender web page. We read the informations as table.

In case we want to correlate this with 2016 behaviour we can also use:

URL<-"http://www.tenisdecamp.ro/turnee-wta-2016/ (http://www.tenisdecamp.ro/turnee-wta-2016/)"

```
URL<-"http://www.wtatennis.com/calendar"
  tour<-getURL(URL,ssl.verifypeer = FALSE)

    doc <- htmlParse(tour)
    appData <- doc['//table[@class="yfnc_datamodoutline1"]']
    perftable <- readHTMLTable(tour[[1]], stringsAsFactors = F)
```

Create a data frame with all tournaments of wta in 2017 -date, location, type

```
df<-data.frame(perftable)
    new_df<-data.frame(A=df[1], B=df[2], C=df[5])
```

Change the names of the columns in the dataframe

```
names(new_df)[1]<-paste("Date")
    names(new_df)[2]<-paste("Location")
    names(new_df)[3]<-paste("Type")
```

Extracting from the Date column the starting and end date of the tournament Open question: is there an automatic way to read the starting and end date in a column defined as an interval?

```
new_df$Date<-substr(new_df$Date, 1,15)
new_df$Startdate<-substr(new_df$Date,1,6)
new_df$Enddate<-substr(new_df$Date,-6,-1)
```

Adding the year

```
new_df$Startdate<-paste(new_df$Startdate, '2017')
    new_df$Enddate<-paste(new_df$Enddate, '2017')
```

We add a new column with Halep participation information on the new_df dataframe 1- participted, 0 did not We also add a new column with information about the Round she reached in the tournaments

```
new_df$Participation<-'0'
```

Grade every tour with points according to type Legend: 125K Series =1;International=2; Premier=3; Premier 5 =4; Premier Mandatory=5; Finals=6; Grand Slam 7;

```
new_df$Ranks<-'0'
    new_df$Ranks[new_df$Type=="125k Series"]<-'1'
    new_df$Ranks[new_df$Type=="International"]<-'2'
    new_df$Ranks[new_df$Type=="Premier"]<-'3'
    new_df$Ranks[new_df$Type=="Premier 5"]<-'4'
    new_df$Ranks[new_df$Type=="Premier Mandatory"]<-'5'
    new_df$Ranks[new_df$Type=="Finals"]<-'6'
    new_df$Ranks[new_df$Type=="Grand Slam"]<-'7'
```

Read the file Halep Tour 2017 with the WTA and ITF Tournaments Halep took part in

```
fileNameH="F:/Big Data Measuring and predicting human behaviour/Halep_Trends_20
17/Halep Tours 2017.csv"
    Halep_df<-read.csv(file=fileNameH)

    Halep_df<-data.frame(Halep_df)
```

Check if a tour match Halep tours and replace participation with 1

```
Halep_T<-as.array(Halep_df$Tournaments)


    for (i in (1:length(Halep_T)))
    {
  new_df$Participation[new_df$Location==Halep_T[i]]<-'1'
    }
```

Save the dataframe in a a csv file

```
 fileName="F:/Big Data Measuring and predicting human behaviour/Halep_Trends_20
17/WTA_ITF Tours 2017.csv"
    write.csv(file=fileName, x=new_df)
```

Reading the google trends file on Simona Halep for 2017 as a csv file:

https://trends.google.com/trends/explore?cat=20&date=2017-01-01%202017-12-31&q=halep (https://trends.google.com/trends/explore?cat=20&date=2017-01-01%202017-12-31&q=halep) If you have an account, then an arrow appears on the write side on Interest over time figure and you can save the file as a csv file

```
fileName_trends<-"F:/Big Data Measuring and predicting human behaviour/Halep_Tr
ends_2017/HalepGoogleTrends2017.csv"
  Halep_Google_Trends_2017<-read.csv(fileName_trends)

  Halep_Google_Trends_2017<-data.frame(Halep_Google_Trends_2017)

  Halep_R<-as.array(Halep_df$Round)

  length(Halep_R)
```

```
## [1] 9
```

```
  new_df$Round<-'0'
```

We added a new feature in the new_df data frame, regarding the round Halep reached in the corresponding tournament We can assume that the higher the round the higher the number of points she gets

```
for (i in (1:length(new_df$Participation)))
    {
    if (new_df$Participation[i] == 1)
      {
        new_df$Round[i]<-Halep_df$Round[Halep_df$Tournaments==new_df$Location
[i]]
      }
      }


 fileName="F:/Big Data Measuring and predicting human behaviour/Halep_Trends_20
17/WTA_ITF Tours 2017.csv"
 write.csv(file=fileName, x=new_df)
```

We compare the date of the tournaments with the dates from the google trends file and take into account if Simona Halep participated or not in the tournaments.

We construct a matrix with all the date differences between the trends reads and all tournaments dates. We also consider the type of tournaments (Ranks) and how much did she got to play in the tournament. A points-vector will record for each tournament she participated the rank of the tournament times the round she reached.

```
fileName_tourdates<-"F:/Big Data Measuring and predicting human behaviour/Halep
_Trends_2017/WTA_ITF Tours 2017_date.csv"
 new_Tour_date<-read.csv(fileName_tourdates)
 new_Tour_date$Startdate<-as.Date(new_Tour_date$Startdate, format='%m/%d/%Y' )
 new_Tour_date$Enddate<-as.Date(new_Tour_date$Enddate, format='%m/%d/%Y' )


 Halep_Google_Trends_2017$Date<-as.Date(Halep_Google_Trends_2017$Date, format
='%m/%d/%Y')
 local_trends<-c(Halep_Google_Trends_2017$Trends)

 a<-matrix(20,length(Halep_Google_Trends_2017$Date), length(new_Tour_date$Start
date))
 b<-c(difftime(new_Tour_date$Enddate, new_Tour_date$Startdate))


 Halep_Google_Trends_2017$TT<-'0'

 for (i in (1:length(Halep_Google_Trends_2017$Date)))
  {
    for (j in (1:length(new_Tour_date$Startdate)))
    {
      if (new_Tour_date$Participation[j]==1)
      {
      a[i,j]<-difftime(Halep_Google_Trends_2017$Date[i],new_Tour_date$Startdate
[j])
      }
      if (abs(a[i,j])<=(b[j]-1))
      {
        Halep_Google_Trends_2017$TT[i]<-ifelse(new_Tour_date$Round[j]==1, 1*new
_Tour_date$Ranks[j],ifelse(new_Tour_date$Round[j]==2, 2*new_Tour_date$Ranks[j],
ifelse(new_Tour_date$Round[j]==3,3*new_Tour_date$Ranks[j],ifelse(new_Tour_date
$Round[j]==4, 4*new_Tour_date$Ranks[j],ifelse(new_Tour_date$Round[j]==5, 5*new_
Tour_date$Ranks[j],ifelse(new_Tour_date$Round[j]==6, 6*new_Tour_date$Ranks[j],7
*new_Tour_date$Ranks[j]))))))
      }
    }
    }

Halep_Google_Trends_2017$TT<-as.numeric(Halep_Google_Trends_2017$TT)
Halep_Google_Trends_2017$Trends
```

```
##  [1]   6   3   6   2   5   2   1   2   1   3   5   4  10   2   1   2   7
## [18]   4  17  16  12  19 100  13   4
```

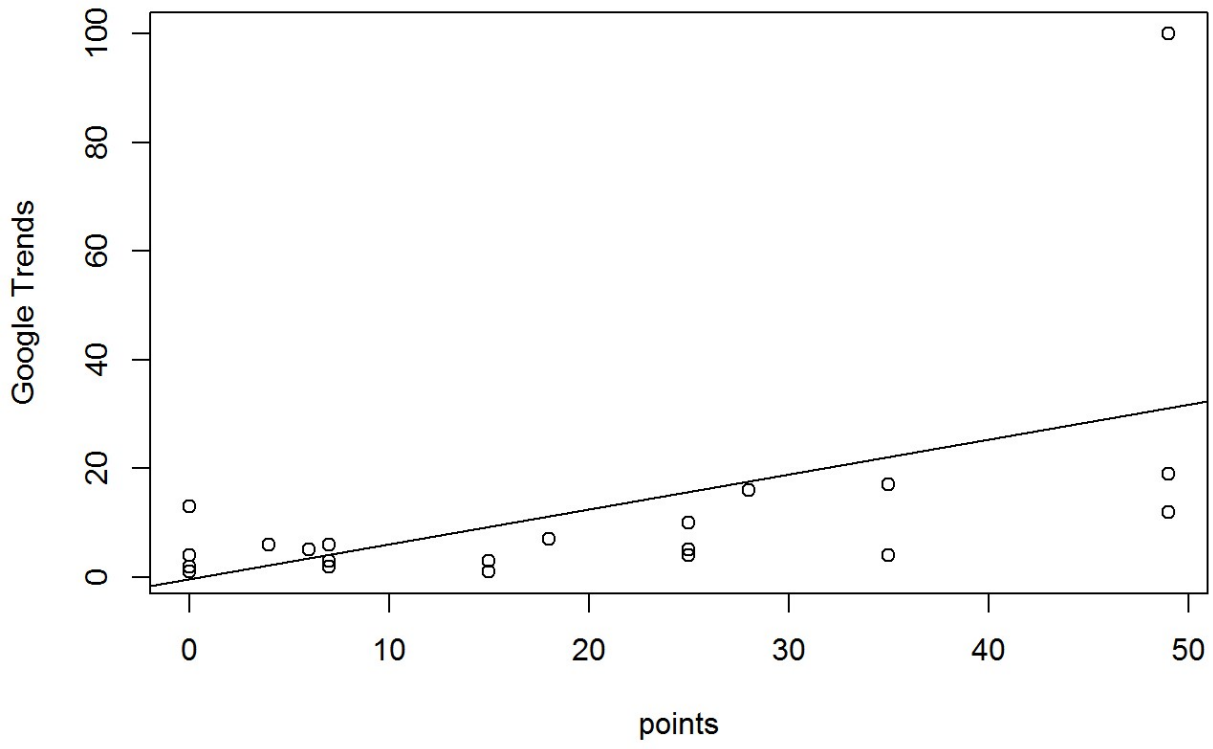
```
Halep_Google_Trends_2017$TT
```

```
##  [1]  4  7  7  7  6  0  0  0 15 15 25 25 25  0  0  0 18 35 35 28 49 49 49
## [24]  0  0
```

```
Halep_Google_Trends_2017$Trends<-as.numeric(Halep_Google_Trends_2017$Trends)
```

Is there a connection between the participation at the nournament and the trends? Is there a connection between the round number and the trends in that period?

```
#plot.new()
plot(Halep_Google_Trends_2017$TT,Halep_Google_Trends_2017$Trends, xlab="point
s", ylab="Google Trends")
abline(lm(Halep_Google_Trends_2017$Trends~Halep_Google_Trends_2017$TT))
```




```
correl<-cor(Halep_Google_Trends_2017$Trends,Halep_Google_Trends_2017$TT)
correl
```

```
## [1] 0.5599446
```

Conclusion

The correlation value is larger than +0.5 which could be viewd as a moderate uphill (positive) relationship Personally, I would incline to say that there is a moderate relation between the search on Simona Halep name this year and the participation in tournaments and how higher she gets The Dataframe was clearly not very large, we could fill this with data from 2016 or from the last past years. The data regarding the search exists on Google Trends.