Study of the distribution of RTT-points among female tennis players under 16 years old

Alexey Zhukov

28.10.2020

Overview

This study is aimed at studying the distribution of RTT[1] points among Russian female tennis players under the age of 16.

In the study, we will try to determine the influence of various factors on the number of rating points and identify dependencies and patterns.

Let's use the **R** language (the 4.2.0 version) for data analysis. To compare groups of observations, methods of one-factor and two-factor analysis of variance will be used.

The following libraries are will be used:

```
library(dplyr)
library(tidyr)
library(ggplot2)
library(gplots)
library(readxl)
```

Importing Data

Data for analysis downloaded from open sources:

- Russian Tennis Tour website
- a list of Russian cities with population

There are 1123 cities in the population list. The next step is to create a dataset by combining data from the RTT website and a list of cities by a key parameter - the name of the city.

In the resulting dataset of unique observations, that is, a total of RTT-rated participants of all ages from all cities.

Tidying and Transforming Data

To fulfill the objectives of the study, it is necessary to prepare data.

[1] Russian Tennis Tour

First of all, names and surnames were removed, and participants 2004 - 2011 years of birth were selected. Also, observations were removed from the data, in which there were no data on the number of settlements and athletes without points.

For further analysis, new variables were introduced into the dataset - the average number of points by year of birth, the city factor by population (moscow, other, spb) and the average number of points by the city factor.

The data structure is as follows:

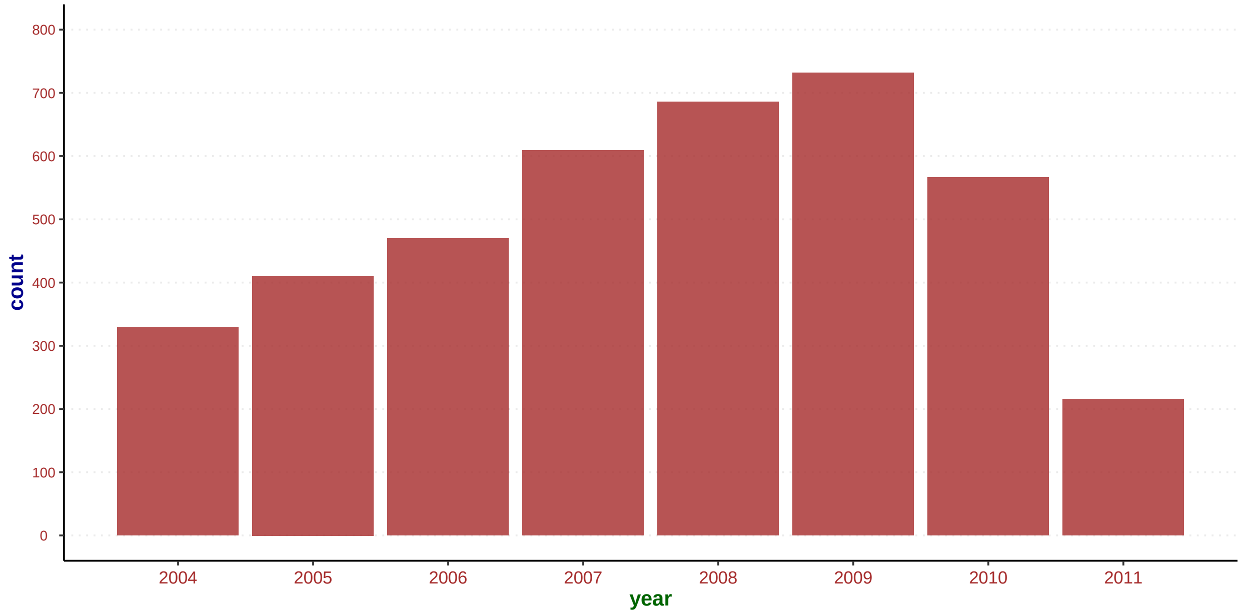
```
## 'data.frame':    4019 obs. of  7 variables:
##  $ X                 : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ points            : int  31 107 69 25 61 201 183 184 49 106 ...
##  $ id                : int  41794 39982 36960 35214 37177 40213 36513 38490 39691 38861 ...
##  $ year              : int  2008 2009 2007 2005 2007 2009 2007 2007 2006 2007 ...
##  $ populationx1000   : int  165 35 83 42 42 21 62 29 29 29 ...
##  $ mean_points_by_year: int  105 59 181 456 181 59 181 181 293 181 ...
##  $ factor.city       : Factor w/ 3 levels "moscow","other",..: 2 2 2 2 2 2 2 2 2 2 ...
```

Analysis and visualization

So, for the study purpose we selected the data of tennis players - their age, city of residence, the number of earned PTT points and id. Now let's take a closer look into the data.

First of all, let's find out the distribution of tennis players:

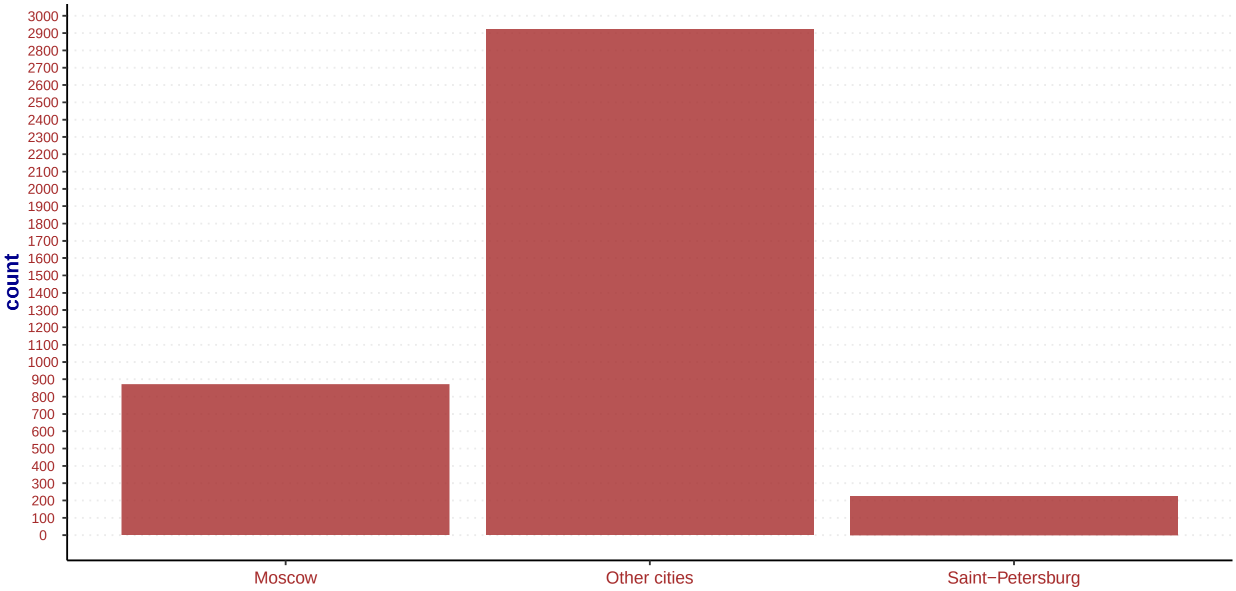
Number of tennis players by year of birth



Next, let's look at the distribution of players depending on the population of the city of residence:[2]

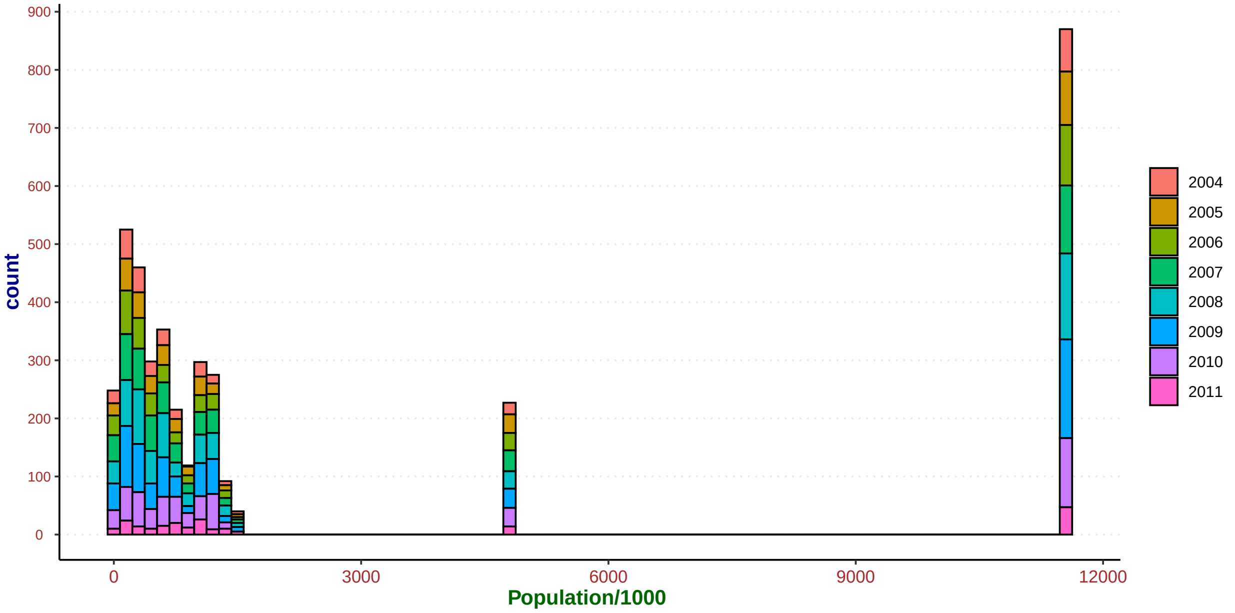
Number of tennis players by city type

[2] Cities are divided into 3 factors - moscow, spb and other (with population less than 3 million people)



Now let’s look at the distribution of players’ ages depending on the size of the city:

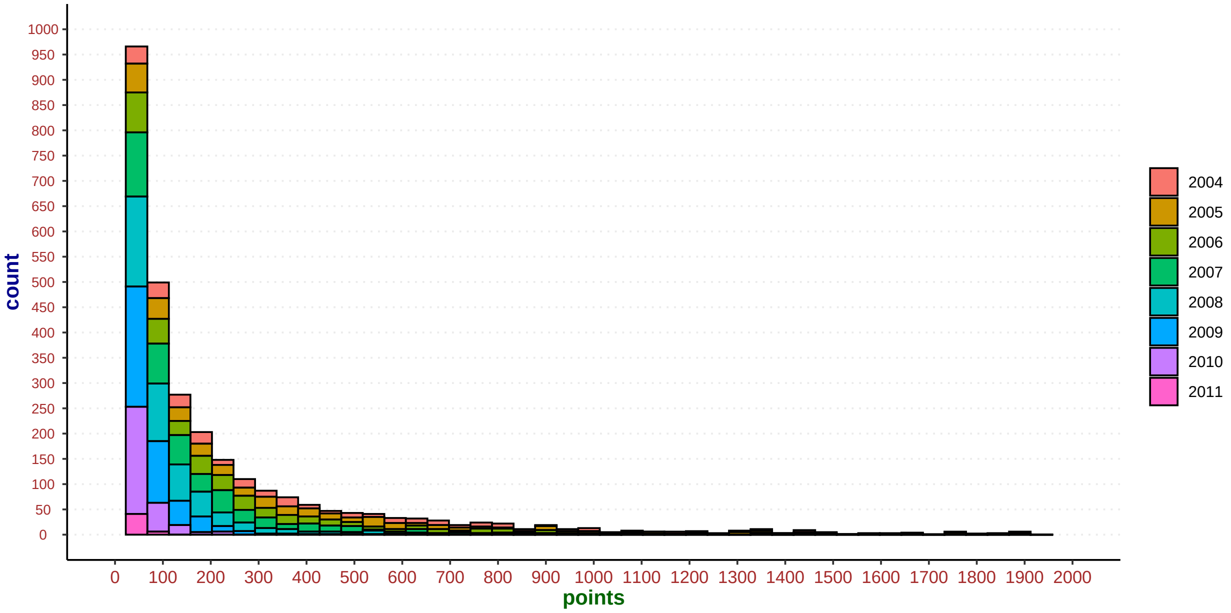
Number of female tennis players by year of birth and city type



The graph shows that, in general, regardless of the population of the city, the participants are distributed by age according to the general distribution *(see first graph)*. The youngest (2011 year of birth) and the oldest (2004 year of birth) tennis players are the fewest in each city, and athletes born in 2009 and 2008 are more than other ages.

Now let’s look at the number of RTT points of tennis players.

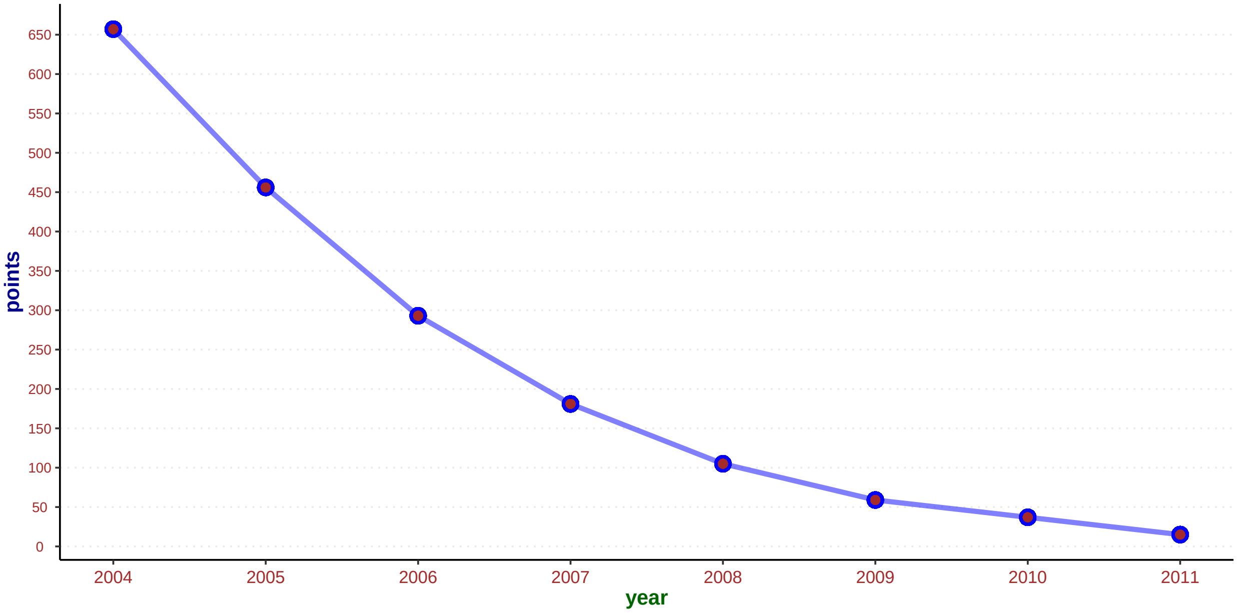
Number of female tennis players with RTT points



In this graph, we can see that older female tennis players have larger number of points. Conversely, the youngest athletes do not have so many points. This is an obvious and logical conclusion. Thus, the first conclusion that we will try to confirm statistically is that *the number of RTT points depends on the age of the athlete*[3].

A graph showing the average number of PTT points depending on the year of birth:

Average number of RTT points by year of birth



Let's move on to statistical analysis.

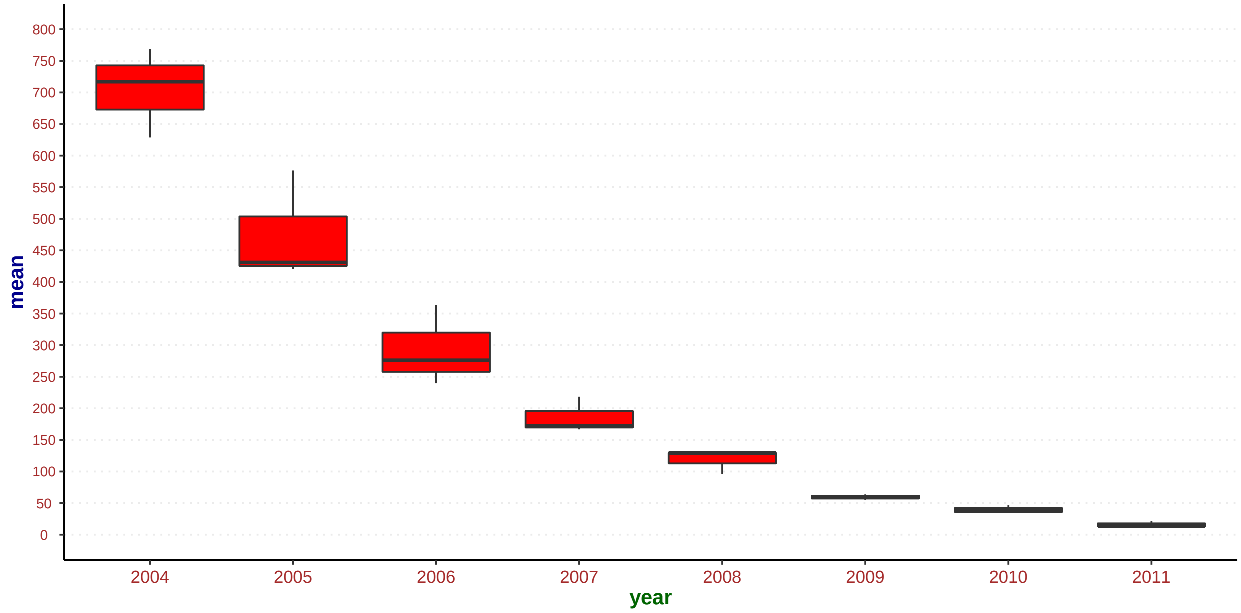
First of all, let's find the main statistical indicators (*number of observations, mean value, standard deviation,*

[3] Of course, here it is necessary to clarify that it is not age itself that affects the number of points, but experience, game practice and sports form that athlete gains with age. Since we do not have information about training time, experience in competitions and other factors that affect success in sports, we will consider a simplified model - age, as a combination of all of the above factors.

first and third quartiles) for observations grouped by city and year:

```
##    year factor.city   n   mean     sd median first_quartile third_quartile
## 1  2004      moscow  73 717.05 765.73  409.0         213.00         978.00
## 2  2004       other 237 628.72 829.23  304.0         118.00         759.00
## 3  2004         spb  20 768.45 695.52  589.5         328.75         846.50
## 4  2005      moscow  92 576.47 764.22  315.0         134.25         642.00
## 5  2005       other 286 420.20 513.03  246.5          83.00         526.00
## 6  2005         spb  32 430.88 460.76  264.0          53.75         622.25
## 7  2006      moscow 104 363.72 443.66  193.0          81.50         436.50
## 8  2006       other 336 276.02 329.63  168.0          40.00         368.00
## 9  2006         spb  30 239.63 334.97   75.0          33.75         267.25
## 10 2007      moscow 117 218.42 322.65  111.0          41.00         236.00
## 11 2007       other 456 172.62 228.44   84.5          27.00         240.25
## 12 2007         spb  36 166.69 210.98   95.0          30.75         213.25
## 13 2008      moscow 148 129.50 147.66   78.0          37.75         166.00
## 14 2008       other 508  96.29 118.24   57.0          18.00         124.00
## 15 2008         spb  30 129.27 141.10   90.0          29.50         143.50
## 16 2009      moscow 170  55.05  54.85   39.5          13.00          79.50
## 17 2009       other 529  59.41  69.85   38.0          12.00          83.00
## 18 2009         spb  33  63.79  50.24   56.0          20.00         103.00
## 19 2010      moscow 119  46.50  47.16   35.0          10.50          64.50
## 20 2010       other 415  34.59  37.77   23.0           8.00          46.50
## 21 2010         spb  32  37.53  31.91   37.5           9.00          55.25
## 22 2011      moscow  47  21.91  17.70   16.0           9.00          32.50
## 23 2011       other 155  12.47  13.77    7.0           4.00          16.50
## 24 2011         spb  14  13.79  17.81    6.5           4.25          12.75
```

The resulting values can be plotted:



According to the data, the average scores differ markedly depending on the year of birth of the participants (which we already assumed earlier), but in addition to this, there is also a certain dependence on the type of city.

Let us take as a hypothesis that the number of points is higher for athletes from larger cities and, accordingly, less for tennis players from less populated cities.

To prove or disprove our theories, it is necessary to carry out an analysis of variance on groups of observations.

First, let's apply an ANOVA to check if the nominative variable Year of Birth really affects the quantitative variable Points

```
final_data$year <- factor(final_data$year)
fit1 <- aov(points ~ year, final_data)
summary(fit1)
```

```
##               Df    Sum Sq  Mean Sq F value Pr(>F)    
## year           7 143234737 20462105   176.8 <2e-16 ***
## Residuals   4011 464192605   115730                   
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Analysis of variance showed that the year of birth affects the number of points **(F=177.8, p-value = 2e-16)**, because since the *p-value* is less than 0.05, we must reject the null hypothesis that one variable does not affect the other and accept the alternative hypothesis.

However, we have 8 levels of the year of birth factor, i.e. we consider female tennis players 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011 born years. The obtained results do not allow us to assert that there are differences between all groups. To understand between which groups (factor levels) there are differences, it is necessary to compare them in pairs. To do this, we use Tukey test (*Tukey's HSD test*):

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = points ~ year, data = final_data)
## 
## $year
##                 diff        lwr         upr     p adj
## 2005-2004 -200.63215 -276.92587 -124.338435 0.0000000
## 2006-2004 -363.62089 -437.71111 -289.530667 0.0000000
## 2007-2004 -475.65502 -546.17124 -405.138809 0.0000000
## 2008-2004 -551.83369 -620.94508 -482.722296 0.0000000
## 2009-2004 -598.13301 -666.53544 -529.730577 0.0000000
## 2010-2004 -619.47109 -690.92247 -548.019708 0.0000000
## 2011-2004 -642.11616 -732.40505 -551.827278 0.0000000
## 2006-2005 -162.98874 -232.70311  -93.274372 0.0000000
## 2007-2005 -275.02287 -340.92635 -209.119396 0.0000000
## 2008-2005 -351.20154 -415.59964 -286.803430 0.0000000
## 2009-2005 -397.50086 -461.13752 -333.864198 0.0000000
## 2010-2005 -418.83894 -485.74209 -351.935787 0.0000000
## 2011-2005 -441.48401 -528.21817 -354.749852 0.0000000
## 2007-2006 -112.03413 -175.37368  -48.694586 0.0000024
## 2008-2006 -188.21280 -249.98452 -126.441071 0.0000000
## 2009-2006 -234.51212 -295.48961 -173.534627 0.0000000
## 2010-2006 -255.85020 -320.22925 -191.471153 0.0000000
## 2011-2006 -278.49527 -363.29766 -193.692888 0.0000000
## 2008-2007  -76.17866 -133.61490  -18.742426 0.0015143
## 2009-2007 -122.47799 -179.05917  -65.896806 0.0000000
## 2010-2007 -143.81607 -204.04759  -83.584538 0.0000000
## 2011-2007 -166.46114 -248.15948  -84.762793 0.0000000
## 2009-2008  -46.29932 -101.11974    8.521092 0.1706438
## 2010-2008  -67.63740 -126.21799   -9.056818 0.0110331
## 2011-2008  -90.28247 -170.77140   -9.793547 0.0155739
## 2010-2009  -21.33808  -79.08055   36.404397 0.9524905
## 2011-2009  -43.98315 -123.86416   35.897861 0.7070064
## 2011-2010  -22.64507 -105.15194   59.861794 0.9912959
```

The result of applying the test are pairwise comparisons of all years of birth with each other. In this case, the null hypothesis is that there are no differences between pairs of groups. As a result, we see that the p-significance level is less than 0.05 for all comparisons, except for the pairs 2008-2009, 2009-2010, 2009-2011 and 2010-2011, which tells us that the differences in the average number of points between these groups (years) are statistically insignificant.

Next, we will test the hypothesis about the influence of the city, or rather, the population factor, on the number of points. In this case, we will also apply analysis of variance (ANOVA).

```
##               Df    Sum Sq Mean Sq F value  Pr(>F)
## factor.city    2   1824865  912433   6.051 0.00238 **
## Residuals   4016 605602477  150797
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Indeed, it can be argued that the city population factor affects the number of points (*p-value < 0.05*).

However, in this case, we have 3 factor levels of the independent variable - moscow, other, spb. Let's find the differences between the groups by comparing them in pairs using Tukey's test.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = points ~ factor.city, data = final_data)
##
## $factor.city
##                    diff       lwr       upr     p adj
## other-moscow -50.43856 -85.60229 -15.27483 0.0022446
## spb-moscow   -15.26074 -83.11721  52.59573 0.8579652
## spb-other     35.17782 -27.55485  97.91050 0.3869889
```

As a result, we see that there are significant differences (*p-value < 0.05*) only between the 'moscow' and 'other' factors, i.e. between cities of less than 3 million people and Moscow. It is noteworthy that the differences between St.Petersburg and Moscow, as well as between St.Petersburg and other smaller cities, are insignificant.

So, with the help of two one-way ANOVAs, we were able to see the influence of factors on our dependent variable, the number of RTT points. We would get the same result when conducting one two-way analysis (**Two-way ANOVA**), that is, we would check the influence of both factors on the dependent variable in one analysis.

However, two-way analysis can also show us whether, in addition to the influence of two factors on a variable, there is also an interaction between these factors that can also have an influence. So, in two-way ANOVA, in addition to the two main factors, one can add their interaction[4].

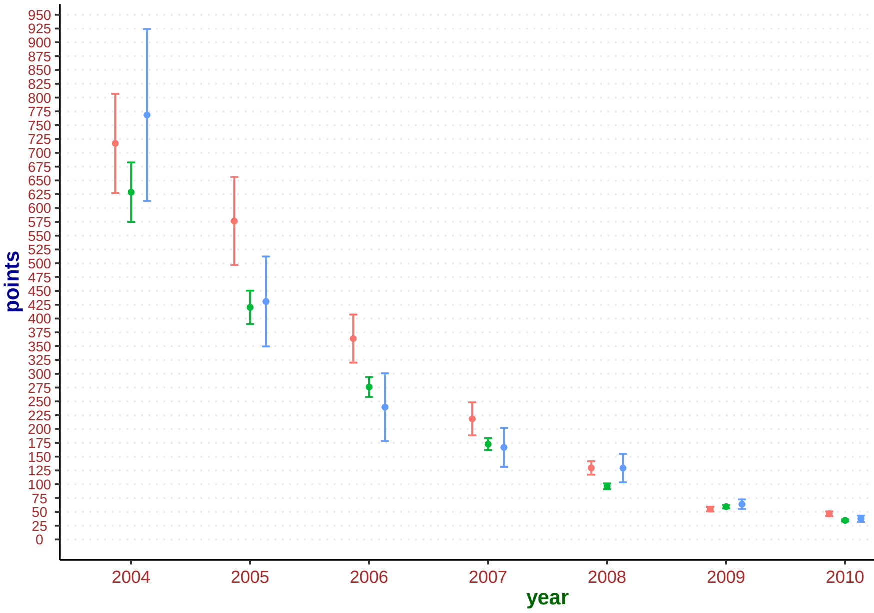
[4] In our analysis, this is not entirely logical, because there is no connection between the year of birth of a tennis player and the population of the city in which she lives.

```
##                     Df    Sum Sq  Mean Sq F value Pr(>F)    
## year                 7 143234737 20462105 177.439 <2e-16 ***
## factor.city          2   1502419   751209   6.514 0.0015 ** 
## year:factor.city    14   1991203   142229   1.233 0.2428    
## Residuals         3995 460698984   115319                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As a result, we see that the *p-value* of the interaction of *year:factor.city* factors is less than 0.05, which allows us to reject the hypothesis about the influence of the interaction of these factors on the independent variable.

Summary

Let's look at a graph that takes into account all our findings:



Average number of RTT points as of 10/25/2020

Conclusion

Using the simplest statistical analysis tools, we were able to show the dependence of the number of PTT points on factors such as the population of the city and the year of birth of the tennis player. However, although we have shown that the score variable depends on both factors, the analysis showed that not all levels of these factors have the same value.

As a result, we can conclude that between the youngest tennis players born in 2009-2011 there are no significant differences in the level of play (if we consider the number of points scored by experience), and the difference in points is noticeable between tennis players from Moscow and from cities with a population of less than 3 million people.