Students' mother tongue influence on Russian language learning

Natalia Isupova

Introduction

Russian language is known as one of the most challenging languages to learn. However, for people whose mother tongue belongs to Slavic group of languages it is often not a difficult one to acquire. This is quite evident that mother tongue and other acquired foreign languages can significantly affect how fast new foreign language is learnt. This research is aimed to examine whether language branch of student's mother tongue has an influence on Russian language learning.

Research hypotheses

In this research we propose the following null hypothesis: the language branch of student's mother tongue does not affect the speed of progress of Russian language learning. Alternative hypothesis is the following: the language branch of student's mother tongue affects the speed of progress of Russian language learning.

Data

To test this hypothesis, we conducted a survey among 68 Russian language learners. They were presented with 4 reading texts in Russian language, belonging to different language levels: A2 (Pre_Intermediate), B1 (Intermediate), B2 (Upper-Intermediate) and C1 (Advanced) according to CEFR. Each text was followed with 3 multiple choice questions about the text content. These questions are used to examine the level Russian language knowledge. Before the texts some meta information about students was collected: • What is their mother tongue? • What other foreign languages do they speak? • How long have they been learning Russian?

The full survey form can be found here: https://raw.githubusercontent.com/NatalieIsupova/dataanalysis_project/master/Isupova_Survey.html

```
df <- read.csv2(
  "https://raw.githubusercontent.com/NatalieIsupova/dataanalysis_project/master/data.csv")
glimpse(df)
```

```
## Observations: 68
## Variables: 30
## $ native_lang          <fct> Persian, German, English, Vietnamese, Ger...
## $ native_lang_branch   <fct> Indo-Iranian, Germanic, Germanic, Vietic,...
## $ native_lang_family   <fct> Indo-European, Indo-European, Indo-Europe...
## $ studying_time_period <fct> more than 2 years to 5 years, equal to or...
## $ studying_time_years  <dbl> 3.0, 5.0, 1.0, 8.0, 5.0, 4.5, 5.0, 3.0, 8...
## $ usage_of_russian     <fct> "studying, communication", "communication...
## $ where_studied        <fct> "university in Russia", "school", "living...
## $ text1_q1             <fct> true, true, true, true, true, true, true,...
## $ text1_q2             <fct> true, true, true, true, true, true, true,...
## $ text1_q3             <fct> true, true, false, false, true, true, tru...
## $ text2_q1             <fct> true, true, true, true, true, true, true,...
## $ text2_q2             <fct> true, true, true, true, true, true, true,...
## $ text2_q3             <fct> true, true, true, true, true, true, true,...
## $ text3_q1             <fct> true, true, true, none, true, true, true,...
## $ text3_q2             <fct> false, false, false, false, true, true, t...
```

```
## $ text3_q3             <fct> false, true, false, false, true, true, no...
## $ text4_q1             <fct> true, none, false, true, none, true, none...
## $ text4_q2             <fct> false, true, true, none, true, true, fals...
## $ text4_q3             <fct> false, none, true, none, true, true, fals...
## $ text1_level          <int> 3, 3, 2, 2, 3, 3, 3, 2, 3, 3, 3, 3, 2, 3,...
## $ text2_level          <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 2, 3, 3,...
## $ text3_level          <int> 1, 2, 1, 0, 3, 3, 2, 3, 3, 2, 3, 2, 0, 2,...
## $ text4_level          <int> 1, 1, 2, 1, 2, 3, 0, 3, 1, 1, 2, 1, 0, 1,...
## $ level_numeric        <int> 2, 3, 3, 2, 4, 4, 3, 4, 3, 3, 4, 2, 2, 3,...
## $ level                <fct> intermediate, upper-intermediate, upper-i...
## $ foreign_lang1        <fct> English, English, none, none, none, Engli...
## $ foreign_lang2        <fct> none, none, none, none, none, none, Spani...
## $ foreign_lang3        <fct> none, none, none, none, none, none, none,...
## $ foreign_lang_branch1 <fct> Germanic, Germanic, , , , Germanic, Germa...
## $ foreign_lang_branch2 <fct> , , , , , , Romance, , Romance, Romance, ...
```

Variables

The variables which are used in the analysis are the following: **native_lang** – student's mother tongue;
native_lang_branch – mother tongue's branch;
native_lang_family – mother tongue's family;
studying_time_years – years of studying Russian language;
studying_time_period – years of studying Russian language turned into 4 periods manually; **level** – level of Russian language which was defined through this survey (pre-intermediate, intermediate, upper-intermediate, advanced);
level_numeric – level of Russian language written with numbers (for correlation and regression);

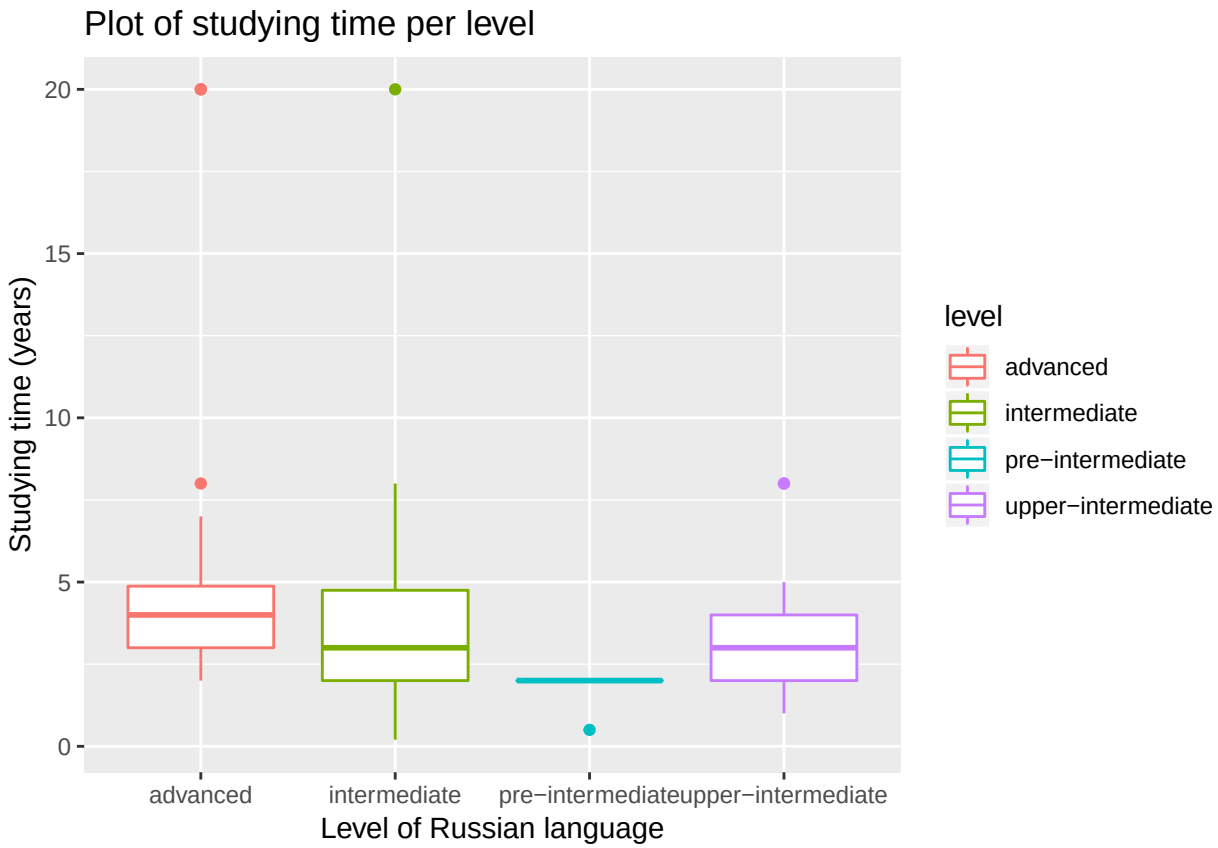
Descriptive statistics and visualization

Firstly, we will count mean and median time of studying for each level.

```
df %>%
  group_by(level) %>%
  summarise(mean_time=mean(studying_time_years),
            median_time=median(studying_time_years))
```

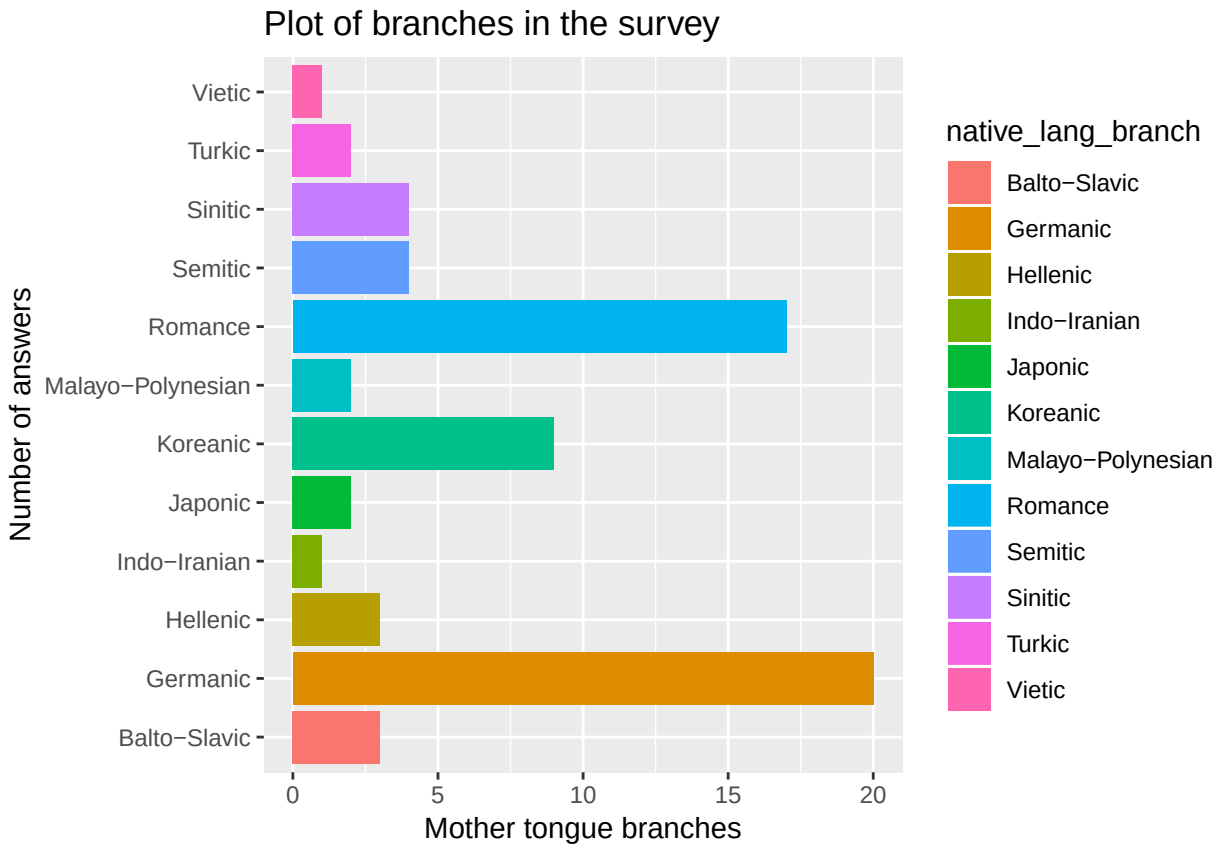
```
## # A tibble: 4 x 3
##   level              mean_time median_time
##   <fct>                  <dbl>       <dbl>
## 1 advanced                5.69           4
## 2 intermediate            4.27           3
## 3 pre-intermediate        1.75           2
## 4 upper-intermediate      3.38           3
```

```
ggplot(df, aes(x=level, y=studying_time_years, color=level)) +
  geom_boxplot()+
  labs(title="Plot of studying time per level",
       x="Level of Russian language",
       y = "Studying time (years)")
```

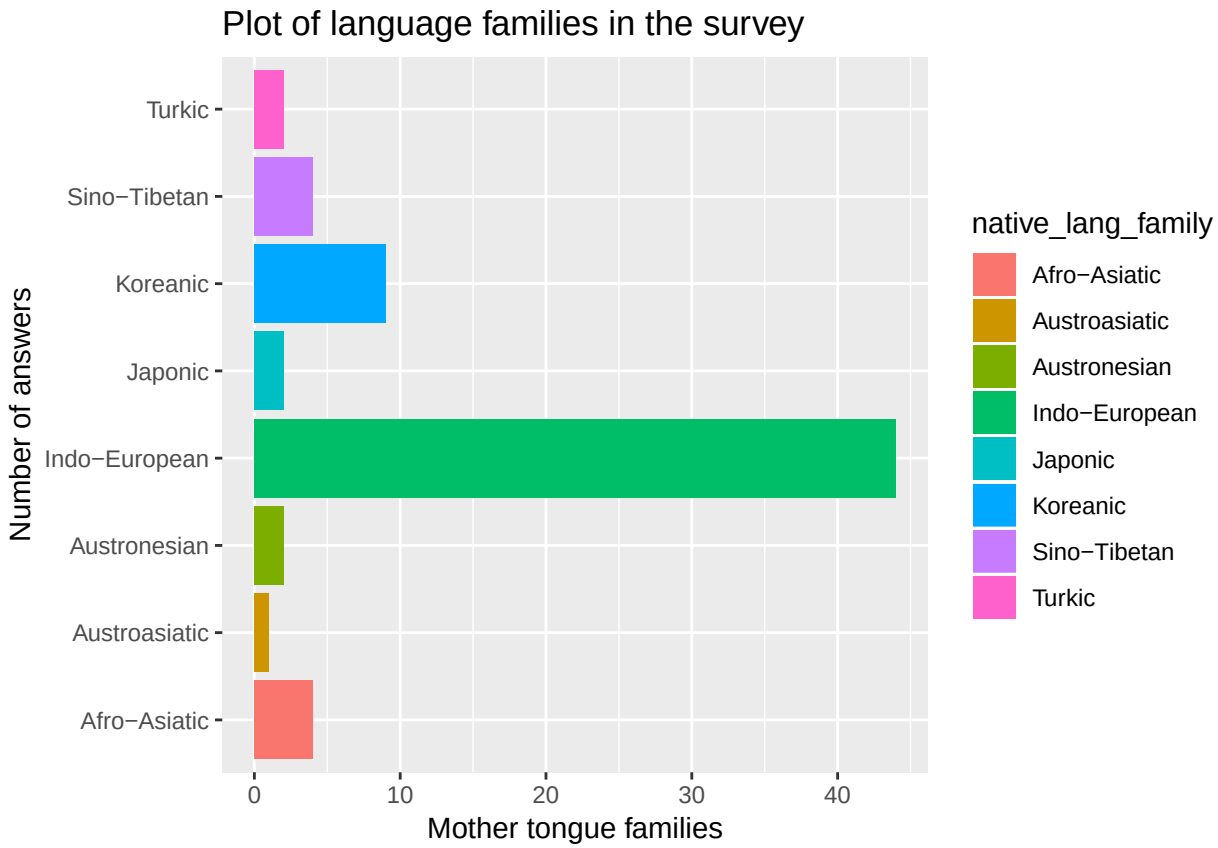


Then we will look at the branches and families presented in the survey data.

```
df %>% group_by(native_lang_branch) %>% count(native_lang_branch) -> branches

ggplot(branches, aes(x=native_lang_branch, y=n, fill=native_lang_branch)) +
  geom_bar(stat="identity")+
  coord_flip()+
  labs(title="Plot of branches in the survey",
       x="Number of answers",
       y="Mother tongue branches")
```

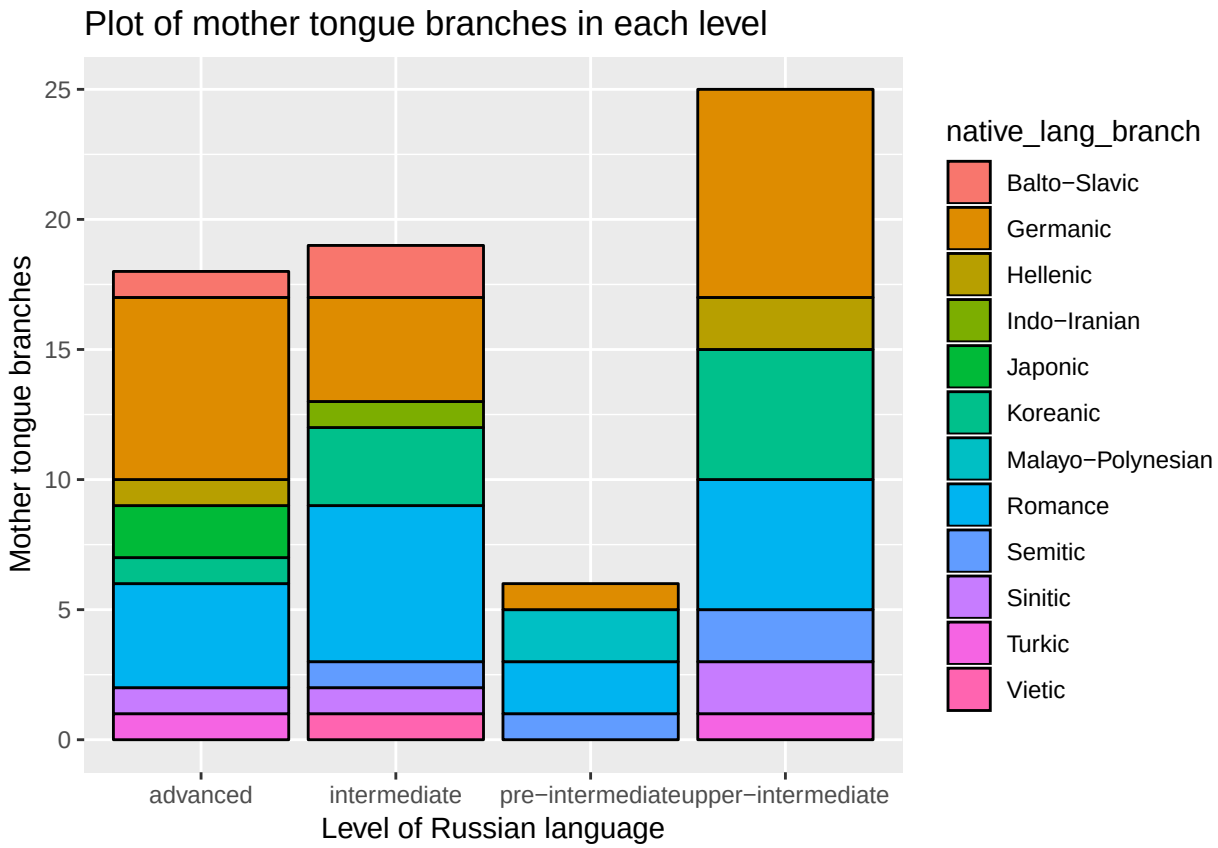



```r
df %>% group_by(native_lang_family) %>% count(native_lang_family) -> families

ggplot(families, aes(x=native_lang_family, y=n, fill=native_lang_family)) +
  geom_bar(stat="identity")+
  coord_flip()+
  labs(title="Plot of language families in the survey",
       x="Number of answers",
       y="Mother tongue families")
```

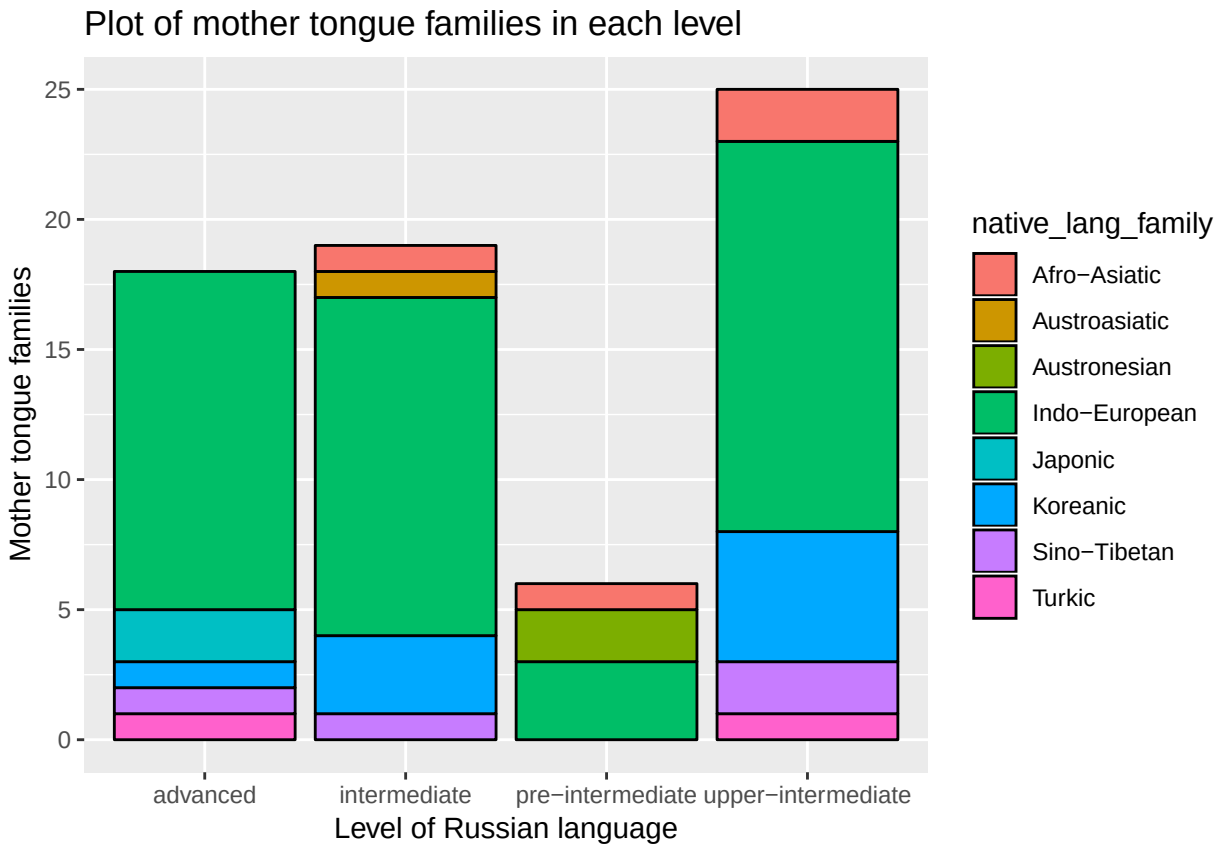


After that let's see which branches and families are predominant at each level.

```
df %>% group_by(level, native_lang_branch) %>% count(native_lang_branch) -> level_branches

ggplot(level_branches, aes(x=level, y=n, fill=native_lang_branch)) +
  geom_bar(stat="identity", color="black")+
  labs(title="Plot of mother tongue branches in each level",
       x="Level of Russian language",
       y="Mother tongue branches")
```

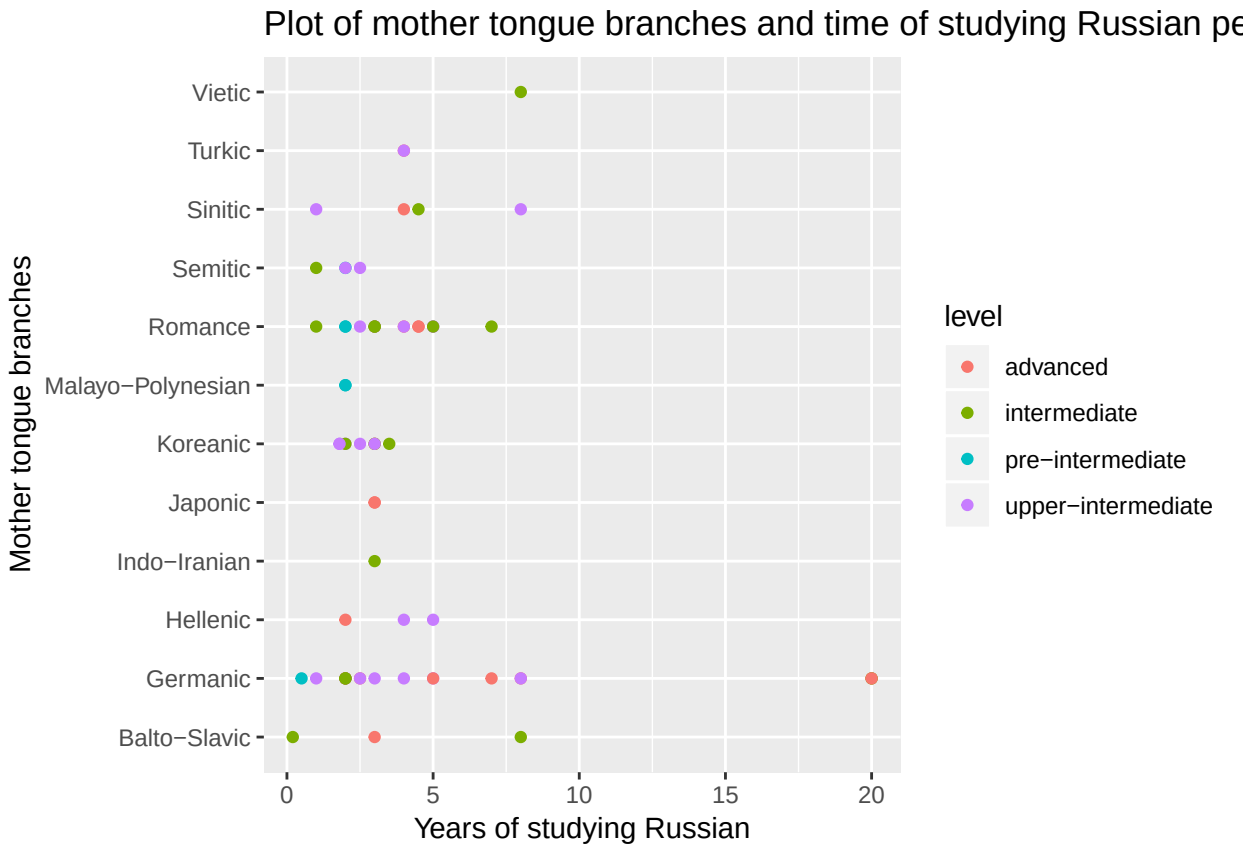



```
df %>% group_by(level, native_lang_family) %>% count(native_lang_family) -> level_families

ggplot(level_families, aes(x=level, y=n, fill=native_lang_family)) +
  geom_bar(stat="identity", color="black")+
  labs(title="Plot of mother tongue families in each level",
       x="Level of Russian language",
       y="Mother tongue families")
```



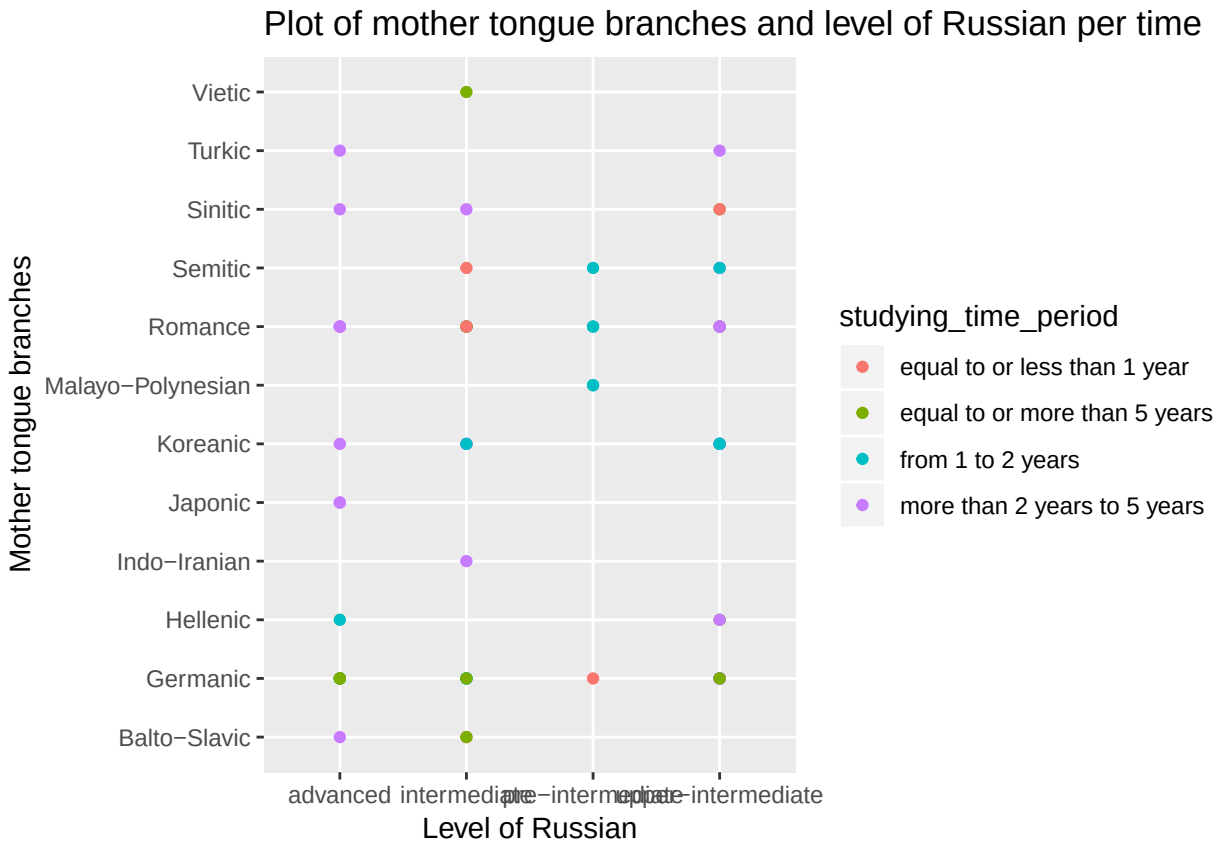
Let's look whether monther tongue branches correspond to level and time of studying.

```
ggplot(df, aes(x = native_lang_branch, y = studying_time_years, color = level))+
  geom_point()+
  coord_flip()+
  labs(title="Plot of mother tongue branches and time of studying Russian per level",
       y="Years of studying Russian",
       x="Mother tongue branches")
```



On this plot we cannot see a good correlation between specific branches, time of studying and level of Russian language. Let's replace the quantitative variable "studying_time_years" with qualitative variable "studying_time_period" with 4 values: "equal to or less than 1 year", "from 1 to 2 years", "more than 2 to 5 years", "equal to or more than 5 years".

```
ggplot(df, aes(x = level, y = native_lang_branch, color = studying_time_period))+
  geom_point()+
  labs(title="Plot of mother tongue branches and level of Russian per time",
       y="Mother tongue branches",
       x="Level of Russian")
```



Testing hypotheses

The null hypothesis is that the language branch of student's mother tongue does not affect the speed of progress of Russian language learning.

Let's filter the data according to the level of Russian language.

```
pre <- df %>% filter(level=="pre-intermediate")
int <- df %>% filter(level=="intermediate")
upp <- df %>% filter(level=="upper-intermediate")
adv <- df %>% filter(level=="advanced")
```

To test our hypothesis we will use ANOVA. We conduct it on each of 4 filtered datasets.

```
res_pre <- aov(pre$studying_time_years ~ pre$native_lang_branch)
summary(res_pre)
```

```
##                          Df Sum Sq Mean Sq  F value Pr(>F)
## pre$native_lang_branch    3  1.875   0.625 1.08e+31 <2e-16 ***
## Residuals                 2  0.000   0.000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
res_int <- aov(int$studying_time_years ~ int$native_lang_branch)
summary(res_int)
```

```
##                          Df Sum Sq Mean Sq F value Pr(>F)
## int$native_lang_branch    7   54.6     7.8    0.29  0.944
## Residuals                11  295.9    26.9
```

```
res_upp <- aov(upp$studying_time_years ~ upp$native_lang_branch)
summary(res_upp)
```

```
##                          Df Sum Sq Mean Sq F value Pr(>F)
## upp$native_lang_branch  6  16.06   2.677   0.729  0.633
## Residuals               18  66.13   3.674
```

```
res_adv <- aov(adv$studying_time_years ~ adv$native_lang_branch)
summary(res_adv)
```

```
##                          Df Sum Sq Mean Sq F value Pr(>F)
## adv$native_lang_branch  7  146.6   20.95   0.589  0.752
## Residuals               10  355.4   35.54
```

Only for dataset for pre-intermediate level we can reject our null hypothesis (p-value is less than 0.05). It means that mother tongue branch of a student affects speed of Russian language learning. For intermediate, upper-intermediate and advanced level we cannot reject the null hypothesis - mother tongue branch does not affect the speed of learning.

Let's take 2 of the biggest branches - Germanic and Romance - and find out possible correlation between level and time of studying. We will filter our dataset on branch variable and conduct Spearman correlation test. We will use Spearman to avoid absolute values of years of studying, and we will use rank instead.

```
ger <- df %>% filter(native_lang_branch=="Germanic")
rom <- df %>% filter(native_lang_branch=="Romance")
```

```
cor.test(ger$studying_time_years, ger$level_numeric, method = 'spearman')
```

```
## Warning in cor.test.default(ger$studying_time_years, ger$level_numeric, :
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  ger$studying_time_years and ger$level_numeric
## S = 675.15, p-value = 0.02743
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.4923679
```

```
cor.test(rom$studying_time_years, rom$level_numeric, method = 'spearman')
```

```
## Warning in cor.test.default(rom$studying_time_years, rom$level_numeric, :
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  rom$studying_time_years and rom$level_numeric
## S = 487.54, p-value = 0.2881
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2830336
```

We can see that there is a positive correlation in the branch of Germanic languages (p-value is less than 0.05). Romance branch doesn't show correlation between time of studying and level of Russian (p-value is more than 0.05).

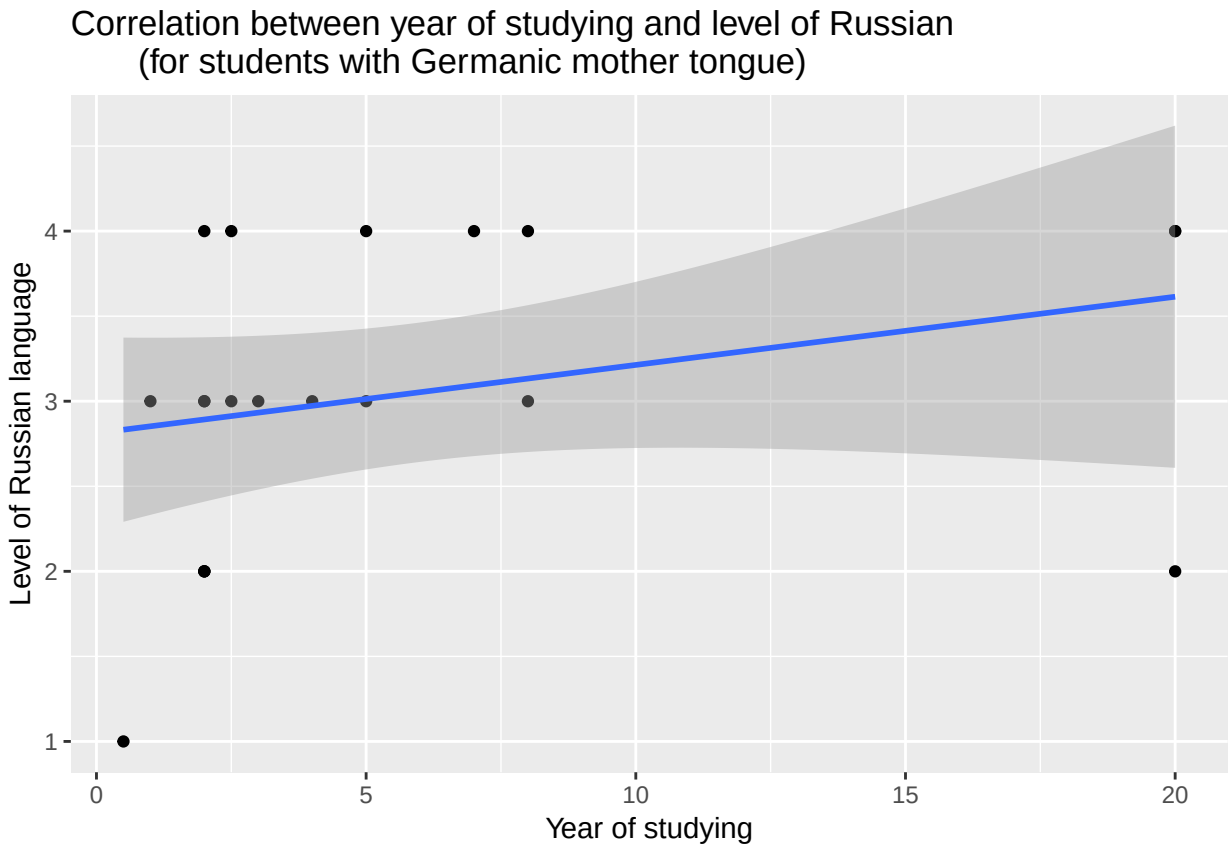
Now we are interested in the following thing: how does level of Russian language change (on average) if time of studying increases by one point? To answer this question we have to build a linear regression model. We will work only with subset for Germanic languages.

```
model1 <- lm(data = ger, level_numeric ~ studying_time_years)
summary(model1)
```

```
##
## Call:
## lm(formula = level_numeric ~ studying_time_years, data = ger)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83252 -0.32305  0.09732  0.50602  1.10735
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.81248    0.26813   10.49 4.26e-09 ***
## studying_time_years   0.04009    0.03107    1.29    0.213
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8719 on 18 degrees of freedom
## Multiple R-squared:  0.08467,    Adjusted R-squared:  0.03382
## F-statistic: 1.665 on 1 and 18 DF,  p-value: 0.2133
```

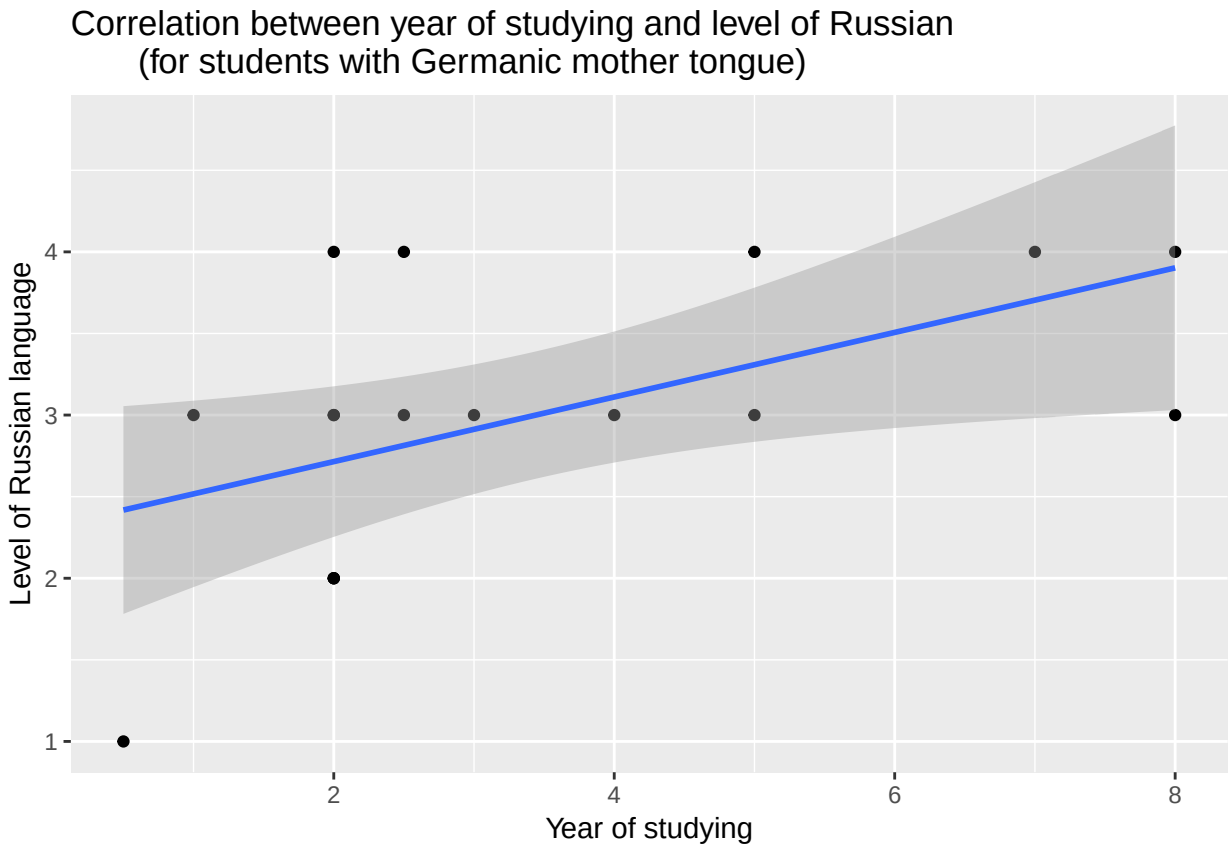
A regression equation will look like this: 2.81 + 0.04*studying_time_years

```
ggplot(data = ger, aes(x = studying_time_years, y = level_numeric)) +
  geom_point() +
  labs(x = "Year of studying",
       y = "Level of Russian language",
       title = "Correlation between year of studying and level of Russian
       (for students with Germanic mother tongue)") +
  geom_smooth(method=lm)
```



We can see that the regression line is not quite significant in this plot because of 2 outliers (20 years of studying Russian language). If we drop them out, we can get a better picture.

```
ger_2 <- df %>% filter(native_lang_branch=="Germanic", studying_time_years < 20)

ggplot(data = ger_2, aes(x = studying_time_years, y = level_numeric)) +
  geom_point() +
  labs(x = "Year of studying",
       y = "Level of Russian language",
       title = "Correlation between year of studying and level of Russian
       (for students with Germanic mother tongue)") +
  geom_smooth(method=lm)
```



Conclusion

We have examined branches of students' mother tongue and tested the hypothesis whether it influence speed of Russian language learning or not. The results of the ANOVA tests conducted are the following: the language branch of student's mother tongue does not affect the speed of progress of Russian language learning when student has reached intermediate or higher level. We did not reject the null hypothesis on pre-intermediate level, it means that the mother tongue does affect the speed of learning on a low level of Russian language proficiency. Also we have examined possible correlations of year of studying and level reached in 2 biggest branched - Romance and Germanic. We have found a positive correlation for Germanic languages using Spearman correlation test and performed regression analysis.

The limitation of this research is the fact that more observations are needed for more detailed and accurate analysis. Two variables of the branches of foreign languages can also be added in a bigger research to get more sophisticated results.

Sources

Code and data on Github: https://github.com/NatalieIsupova/dataanalysis_project