An analysis of WorldSkills Russia competition events

By Aleksandr Gorbachev & David Langeveld

A Research Paper
Submitted to the lecturer of the subject 'Programming with R'
The Hague University of Applied Sciences PRO
Master of Business Administration
MBA Big Data Analytics
November 2023

Introduction

WorldSkills International is a global organization dedicated to promoting excellence in vocational and technical education. It brings together young professionals from around the world to compete in skill-based competitions that showcase their talents and expertise in a wide range of fields, from engineering and technology to hospitality and creative arts. These competitions are assessed by a panel of expert judges who rigorously evaluate participants' performance, measuring their precision, creativity, and adherence to international industry standards.

WorldSkills Russia as a member organization is in charge to govern and oversee that national and regional competition events comply with WorldSkills International standards as well as for collecting and aggregating results countrywide. These qualification assessment standards and frameworks have also been adopted by WorldSkills Russia in the whole national vocational education and training system, so that college graduates have to pass the demonstration exam in exactly the same way as competitors in the WorldSkills competition, with the only difference that the final score is not converted into a medal but into an exam grade.

Current research is targeted to explore and analyze data from WorldSkills Russia competition events and demonstration exams to answer following questions:

- Whether repeated participation in competitions significantly improves a competitor's average score?
- Whether repeated participation of a compatriot expert significantly improves his/her compatriot competitors' average results?

These questions are formulated as followed:

1. Impact of repeated participation on competitor's average score

null hypothesis (h0):
The repeated participation in WorldSkills competitions does not significantly improve a competitor's average score.

alternative hypothesis (h1):
The repeated participation in WorldSkills competitions significantly improves a competitor's average score.

2. Effect of repeated participation of a compatriot expert on competitor results

null hypothesis (h0):
The repeated participation of a compatriot expert does not significantly improve his/her competitor's (average) results in the competitions of WorldSkills Russia.

alternative hypothesis (h1):
The repeated participation of a compatriot expert significantly improves his/her compatriot competitor's average results in the competitions of WorldSkills Russia.

Operationalization of the research questions

Each competition event uses the Competition Information System (CIS) to record competition results. A separate instance of CIS is used for each event, so all event results are stored in a separate database instance. In order to have an overview of all competition events, WorldSkills Russia has developed its own aggregation system called Electronic System for Internet Monitoring (eSIM).

The raw data used for this research was kindly provided by the autonomous non-profit organization "Agency for the Development of Professional Excellence (WorldSkills Russia)" and is essentially a dump from the eSIM database, specifically from the competition results table. Therefore, the observation unit is a skills competition result for a given competitor for a given competition event.

The following steps will make sure to operationalize the data so that the research questions that are formulated will be answered:

Matching provided data with posed questions

This dataset can be used to perform statistical analyses that will help provide evidence and insights into the impact of various factors on competitors' performance and whether these effects are statistically significant. To create a linear regression model based on the research questions, it is necessary to identify an outcome (dependent) variable and predictor (independent) variable(s) for each question.

In statistical terms this will form an equation, $Y = a + bX$.

Y is the outcome variable, x will be the predictor, *a* represents the intercept and *bX* represents the slope associated with the predictor variable.

Whether repeated participation in competitions significantly improves a competitor's average score?

WorldSkills Russia has different levels for local WorldSkills competitions:

- Regional competitions among regional colleges
- National competitions among regions
- University competitions for students of given university
- National inter-university competitions
- HiTech competition among employees of companies in industrial sector
- DigitalSkills competition among college students, university students and employees companies in IT sector

So there may be a case where a given competitor participates in a sequence of competitions at different levels (e.g. Regional, National, HiTech, DigitalSkills) or may be a guest competitor in other regional competitions (usually when preparing for National). It's reasonable to assume that the average score of such a competitor should improve each time he/she participates in the next competition.

To address this question, we can analyze the performance of competitors who have participated in multiple competitions over time. Group competitors into categories based on the number of competitions they've entered and calculate their average scores for each category. Then, use statistical tests or regression analysis to determine if there is a significant improvement in scores with repeated participation.

Outcome Variable: Competitor's average score in skill competitions (for example the average grade/mark over multiple competitions).
Predictor Variable: A variable representing participation (for example 0 for non-repeated participation, 1 for repeated participation).

Whether repeated participation of a compatriot expert significantly improves his/her compatriot competitors' average results?

It's often the case that a particular region has the same designated expert for a particular skill competition, who represents the region at nationals. It's reasonable to assume that, in this case, the preparation methodology of his compatriot competitors should improve over time, and thus the result of an average competitor should tend to improve over time as well.

To explore this, we can group competitors by the presence of a compatriot expert (FK_COMPATRIOT) and analyze the average performance of their compatriot competitors in each group. Statistical tests or regression analysis can help assess if the repeated participation of a compatriot expert significantly improves the average results of their compatriot competitors.

Outcome Variable: Average results of competitors (average grade/mark)
Predictor Variable: A variable representing the presence/absence of a compatriot expert who participated in multiple competitions.

Description of the data used

Raw data is provided as three XLSX tables:

- participants100.xlsx
- participants200.xlsx
- participants300.xlsx
- regions.xlsx

The total number of observations in the three spreadsheets is 600.000. Due to the nature of the SQL scheme, some columns refer to other tables (fk, foreign keys) and due to local regulations, some of the related data (e.g. personal data such as competitor or expert name, age, gender, etc.) is not available for cross-border transfer, storage or processing.
Below is a table with the name and description of the original variables, an indication of whether the variable will be used in the final dataset and an indication of the new name (if applicable):

Variable name	Renamed to	Comment from owner	Will be used
pk_participant	result	ID of competition result record (primary key)	Yes
fkUsers	competitor	competitor reference (foreign key)	Yes
fkComp	skill	skill trade reference (foreign key)	Yes
ChampRole	-	ID of the participant's role at the competition	No
regionID	region	code of participant's region origin	Yes
mark100	mark100	100-point scale	Yes
mark500	mark500	500-point scale	Yes
medal	medal	type of medal awarded	Yes
timestamp	timestamp	time when the result was locked in the system	Yes
fkUserAdd	-	User ID of the user who added the result to the system (foreign key)	No
competitorMarker	-	marker for group competition	No
expertGroupMarker	-	marker for specific expert group	No
excludeFromResault	-	marker for "out of contest" result	No

fk_quotaCategory	-	quota category for the competition (foreign key)	No
fk_command	team	reference for a team membership (foreign key)	Yes
FK_USER_CP	-	user ID in the digital platform (foreign key)	No
ACCESS_RKC	-	whether the regional competition center has access to the record	No
FK_COMPATRIOT	expert	reference for compatriot expert ID (foreign key)	Yes
organization	-	organization represented by the participant	No
nok	-	participation in an independent qualification assessment project (for demonstration exams)	No
participant_updated_at	-	time of last record update	No
is_requested	-	field for access in the business process	No
is_accepted	-	field for access in the business process	No
mark700	-	700-point scale	No

Regions dataset:

Variable name	Renamed to	Comment from owner	Will be used
code	code	Code of region (primary key)	Yes
regionName	regionName	Name of Region reference (foreign key)	Yes

As discussed earlier, the following steps were taken in the form of exploratory data analysis: importing, preparing in the form of wrangling and tidying, computing summary statistics and analysis.

The importing, preparation of the data, and the function that have been used for summary statistics are noted in the Appendix 1 and in the included script.

The steps taken are as followed:

- Join dataset with region names to the main dataset that has the codes
- Grouped by region
- Calculate summary statistics including the mean, median
- Sort table by count of observations and filter the top 10 regions
- Data visualization in a box plot

Detailed customization steps are described in Appendix 1. After customizing original dataset we have a resulting dataframe with 11 variables and ```49 034``` unique results (~3,6% of original size):

> glimpse(esim_data)
Rows: 49,034
Columns: 11
$ result <dbl> 43053, 53048, 53053, 53054, 53057, 53059, 53060, 53064, 53065, 53066, 5~
$ competitor <dbl> 36517, 23137, 15379, 18028, 18038, 23149, 15384, 18309, 18310, 16712, 2~
$ competition <dbl> 182, 322, 322, 322, 322, 322, 322, 349, 349, 349, 322, 322, 322, 322, 3~
$ skill <dbl> 175, 183, 183, 183, 184, 184, 184, 185, 185, 185, 186, 186, 186, 186, 1~

```
$ region      <dbl> 54, 62, 70, 76, 76, 62, 70, 76, 76, 77, 26, 33, 70, 76, 62, 77, 77, 76,~
$ mark100     <dbl> 25.85, 64.60, 6.60, 10.90, 37.35, 30.85, 8.50, 22.60, 22.60, 7.90, 15.9~
$ mark500     <dbl> 549, 561, 461, 468, 507, 495, 456, 539, 539, 480, 479, 516, 469, 512, 5~
$ medal       <chr> "GOLD", "GOLD", NA, NA, "Medallion for Excellence", NA, NA, "SILVER", "~
$ timestamp   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA, NA,~
$ team        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 430, 430, 0, 0~
$ expert      <dbl> 43413, 10956, 15382, 1562, 10758, 10949, 15380, 746, 746, 158, 11425, 4~
```

Number of unique and missing values for each column is provided below:

```
> data.frame(unique=sapply(esim_data, function(x) sum(length(unique(x, na.rm = TRUE)))),
+            missing=sapply(esim_data, function(x) sum(is.na(x) | x == 0)))

           unique missing
result      49034       0
competitor  39579       0
competition   549       0
skill         300       0
region        153       0
mark100      8449       0
mark500       187    2272
medal           5   22001
timestamp   38461    4120
team         4789   38535
expert      27475       0
```

Results of the data analysis

When analyzing the summary statistics where the top 10 regions in terms of mark100 scores, the following could be noted:

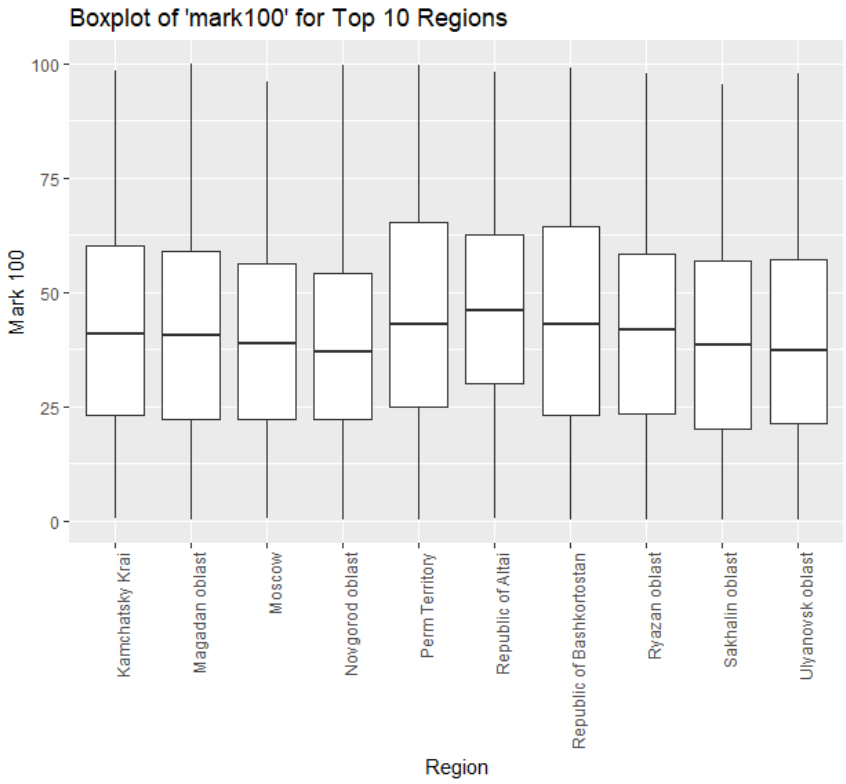
- There are some outliers in 5 of the 10 regions, meaning they deviate a bit more from the bulk of the data. These outliers are data points (observations of results) that fall outside the typical range of the values of the dataset.
- The presence of these outliers can affect the interpretation of the boxplot. The whiskers of the boxplot extend to 1.5 times the interquartile range (IQR) beyond the first and third quartiles.
- The medians and interquartile ranges (IQR) in the boxplots are all relatively close to each other when compared against regions. It suggests that, on average, the central tendency of the scores (mark100) is similar across the top 10 regions. Only the regions “Perm” and “Republic of Altai” have higher medians and IQRs relative to the others.

Filter

	regionName	count	mean_mark100	median_mark100	sd_mark100	avg_mark100
1	Altai Territory	2947	28.86946	23.320	19.245639	28.86946
2	Amur region	918	30.68475	25.460	20.618395	30.68475
3	Arkhangelsk region	1926	27.13238	22.825	18.351852	27.13238
4	Astrakhan Oblast	2018	34.12620	28.465	20.922571	34.12620
5	Belgorod region	1218	35.86626	28.850	23.428661	35.86626
6	Bryansk Oblast	1995	27.58064	22.300	20.548754	27.58064
7	Chechen Republic	988	35.43339	29.925	22.178013	35.43339
8	Chelyabinsk Oblast	1252	31.66641	25.815	20.317744	31.66641
9	Chukotka Autonomous Okrug	266	33.61647	31.940	22.593408	33.61647
10	Chuvash Republic	3386	32.65481	27.490	20.814359	32.65481
11	Irkutsk region	3164	33.12540	27.670	22.501141	33.12540
12	Ivanovo region	1700	31.31949	25.690	21.632835	31.31949
13	Jewish Autonomous Region	926	28.18024	20.100	19.318720	28.18024
14	Kaliningrad Oblast	1136	33.17178	30.630	20.091750	33.17178
15	Kaluga region	690	33.48394	29.590	21.313376	33.48394
16	Kamchatka Territory	5667	33.58901	26.050	23.582860	33.58901
17	Karachay-Cherkess Republic	529	47.43467	49.250	20.321567	47.43467
18	Kemerovo Region	1174	33.02262	28.490	21.940415	33.02262
19	Khabarovsk Territory	696	32.07457	26.155	20.151481	32.07457
20	Khanty-Mansiysk Autonomous Okrug - Yugra	183	37.93967	33.950	24.963599	37.93967
21	Kharkov region	2	65.59000	65.590	0.000000	65.59000
22	Kherson Oblast	2	38.93000	38.930	0.000000	38.93000
23	Kirov region	1416	36.01026	32.210	20.581505	36.01026

Showing 1 to 23 of 95 entries, 6 total columns

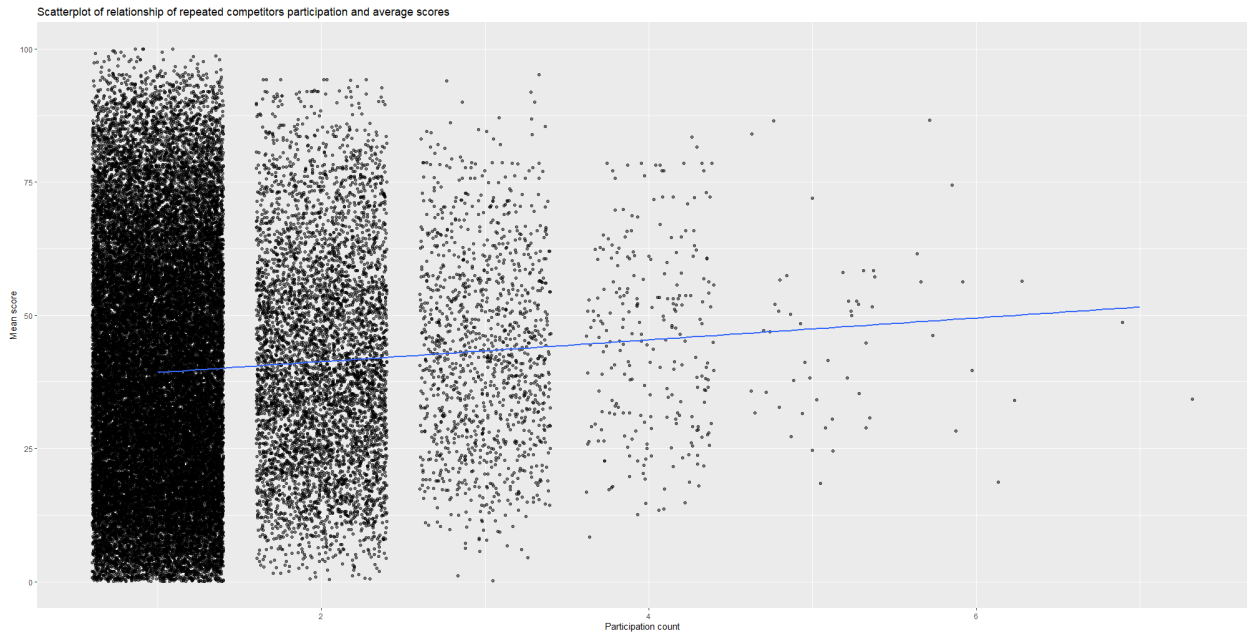
Summary statistics of grouped regions



Box plot of top 10 grouped regions in terms of mark100 results

<u>Research question 1: Whether repeated participation in competitions has a positive impact on a competitor's average score?</u>

The results are grouped by competitors and regions and a summary has been calculated. The linear regression between participation count and mean average score is below. The plotting is below and the slope of the line is slightly positive between the dependent variable *participation count* and independent variable *mean score*, this means there is a tendency for the dependent/predictor variable (y-axis) to increase as the independent/outcome variable (x-axis) increases. In statistical terms, a positive slope suggests a positive correlation between the variables.



Scatterplot which shows the relationship of repeated competitors participation and average scores

To fit the regression model, the following script has been used:
c_model <- lm(mean_score100 ~ results, data = cg)

 # A tibble: 2 × 7

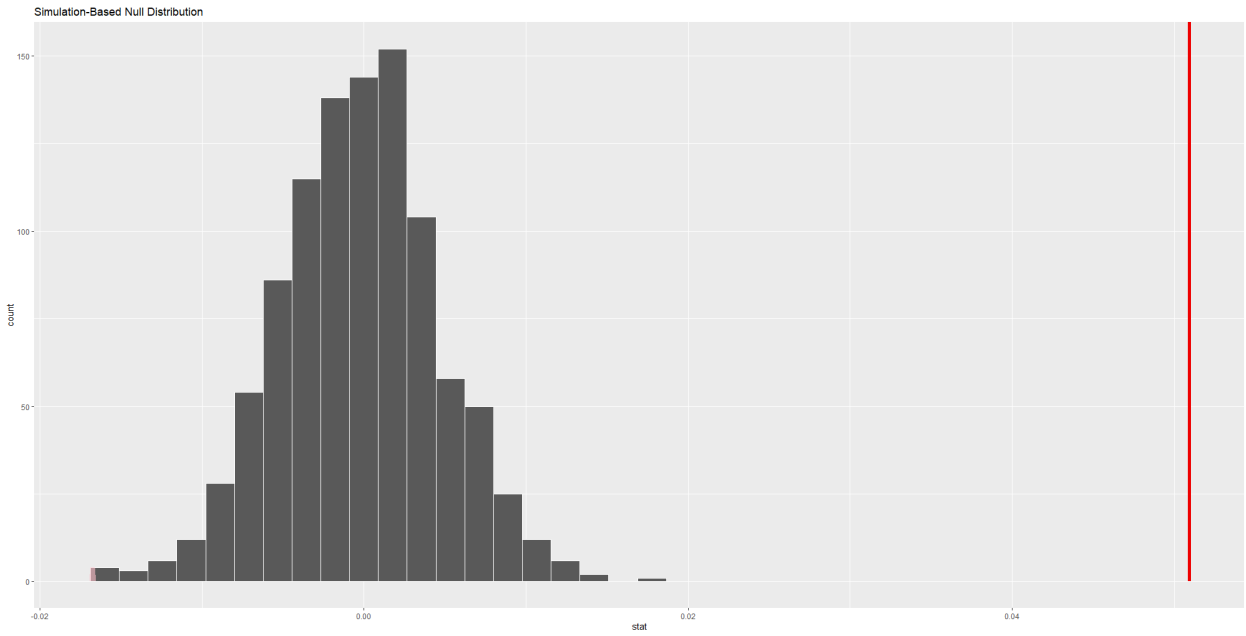
term	*estimate*	*std_error*	*statistic*	*p_value*	*lower_ci*	*upper_ci*
<chr>	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*
1 intercept	*37.2*	*0.275*	*135.*	*0*	*36.7*	*37.7*
2 results	*2.05*	*0.204*	*10.1*	*0*	*1.65*	*2.45*

The regression equation can formulated as followed:

*Average score (mark100) = 37.2 + 2.05 * results*

A low p-value (less than 0.05) suggests that we can reject the null hypothesis. In our case, the p-value for "results" is 0, which is smaller than 0.05.

Building the null distribution and observing what the difference in proportion is needed to answer our research question. For further understanding, we need to know the significance of this positive impact. The simulation-based null distribution is seen below.



Simulation-Based Null Distribution

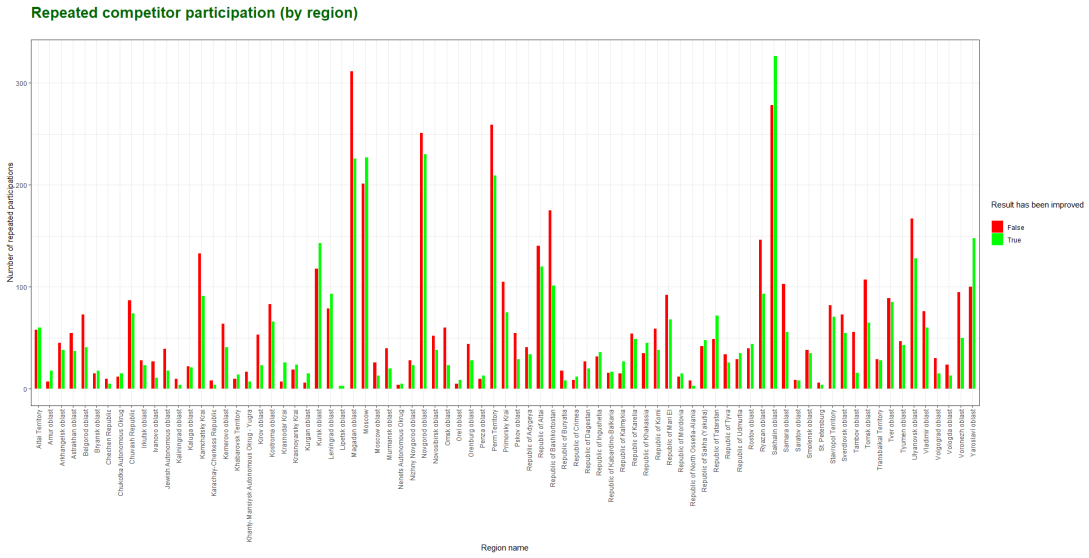
The red-line, p-value of 0.05, is outside the range of our null distribution. This suggests that the observed values are statistically significant.

The significance of this positive impact is to be calculated. The framing of a table with the frequencies of competitor IDs was done, so the subsetting results with only those competitor IDs who participated more than 1 time. The results were ordered by competitor ID and then by result ID. An additional column was added where the boolean value whether each result was higher than the previous one (absolute score).

The column had the following logic:

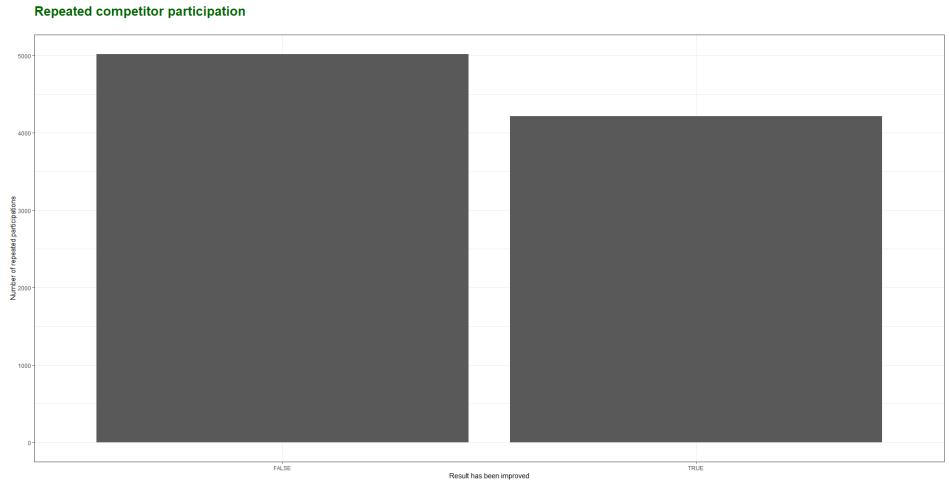
- If this is the first result of a given competitor, the value is “NA”
- If this is not the first result of a given competitor, and mark100 value is higher than the previous mark100 value, than the value is “TRUE”
- If this is not the first result of a given competitor, and mark100 value is lower than the previous mark100 value, than the value is “FALSE”

The plotting of the grouped by region, repeated competitor participation cases was visualized and resulted in the graph below. The “NA” values were excluded.



Bar chart of (repeated) competitor participation per region

The above plot has been summarized in a ratio for focus on the research question below.

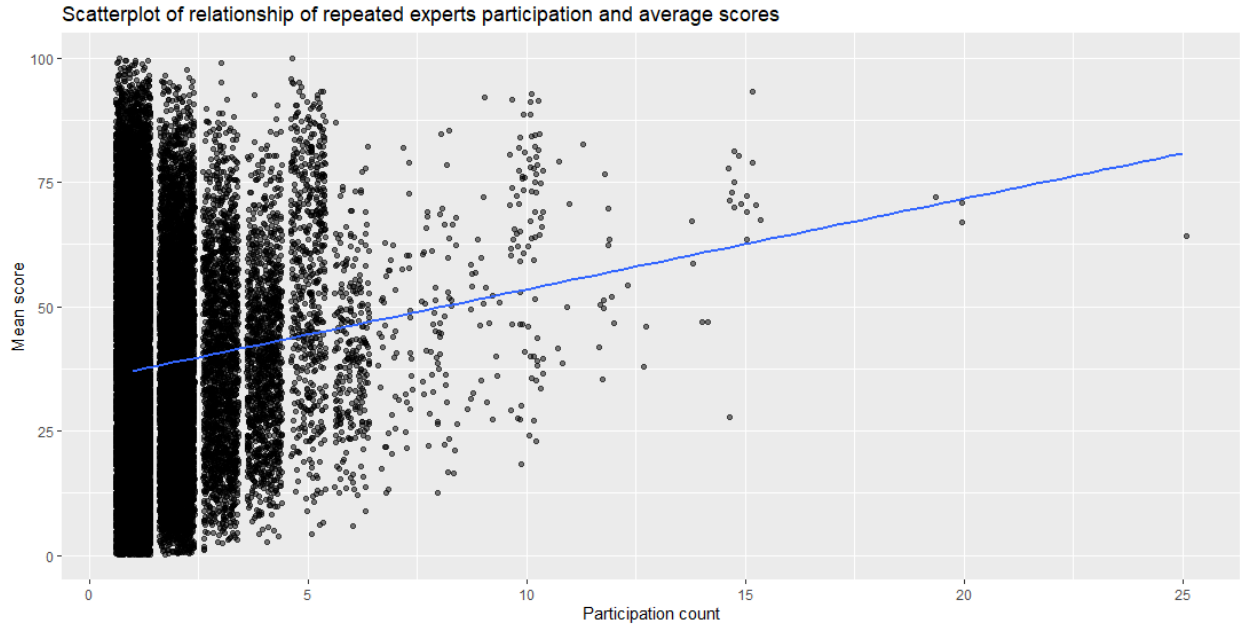


Summary bar chart of (repeated) competitor participation per region

```
# A tibble: 2 × 3
  improve100     n  freq
  <lgl>      <int> <dbl>
1 FALSE       5018 0.543
2 TRUE        4218 0.457
```

Research question 2: Whether repeated participation of a compatriot expert has a positive impact on his/her compatriot competitor's average result?

The grouping of the data by results per expert and regions is needed for this research question. After grouping the results, a linear regression between expert participation count and mean average score is calculated. The plotting is below and the slope of the line is slightly positive, this means there is a tendency for the dependent, or predictor, variable (y-axis) to increase as the independent, or outcome, variable (x-axis) increases. In statistical terms, a positive slope suggests a positive correlation between the variables.



Scatterplot which shows the relationship of repeated expert participation and average scores

To fit the regression model, the following script has been used:

e_model <- lm(mean_score100 ~ results, data = eg)

Which resulted in the table below:

 # A tibble: 2 × 7

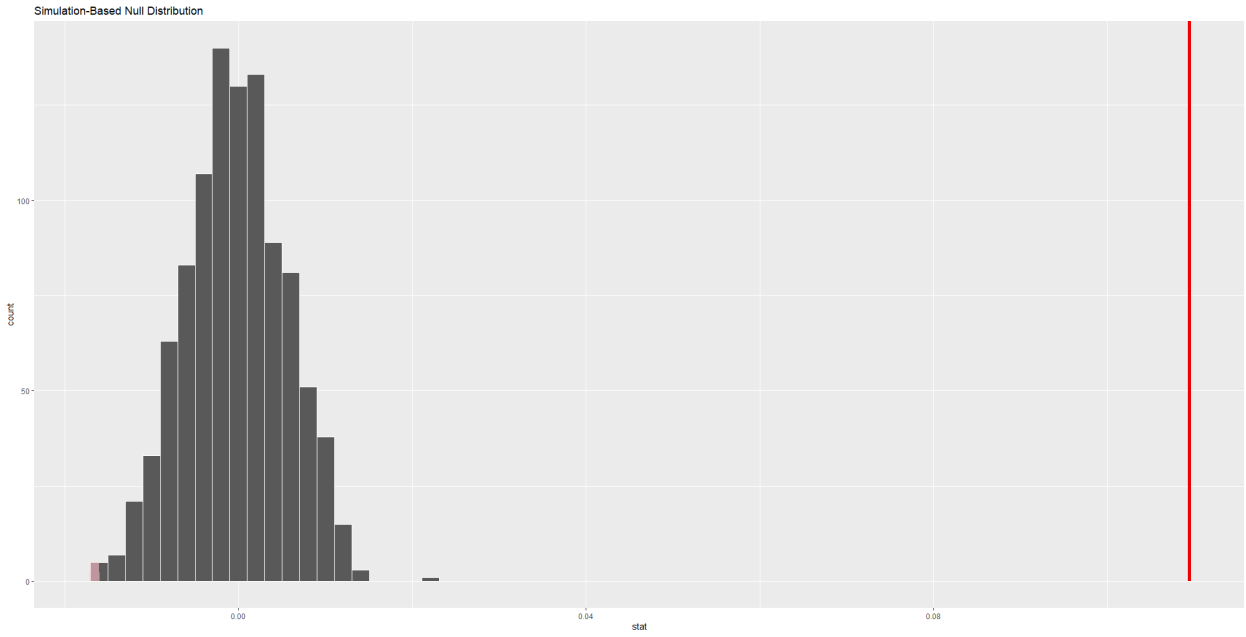
term	*estimate*	*std_error*	*statistic*	*p_value*	*lower_ci*	*upper_ci*
<chr>	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*
1 intercept	*35.3*	*0.22*	*161.*	*0*	*34.8*	*35.7*
2 results	*1.82*	*0.1*	*18.2*	*0*	*1.62*	*2.02*

The regression equation can formulated as followed:

*Average score (mark100) = 35.3 + 1.82 * results*

A low p-value (less than 0.05) suggests that we can reject the null hypothesis. In our case, the p-value for "results" is 0, which is smaller than 0.05.

Building the null distribution and observing what the difference in proportion is needed to answer our research question. For further understanding, we need to know the significance of this positive impact. The simulation-based null distribution is seen below.



Simulation-Based Null Distribution

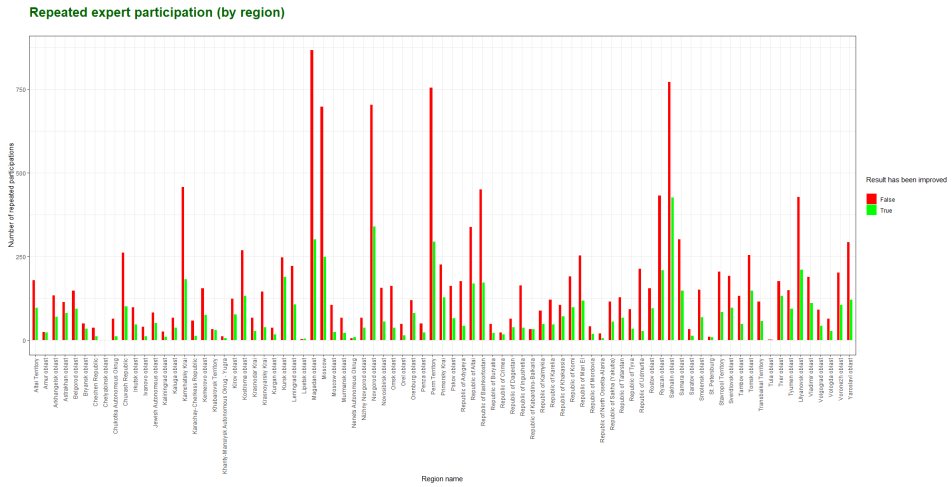
A table with the frequency of expert IDs was created to subset the results with only those expert IDs who participated in more than 1 competition. This new dataframe was also ordered by expert ID and then by result ID. An additional column was added where the boolean value whether each result was higher than the previous one (absolute score).

The column had the following logic:

- If this is the first result of a given expert (compatriot expert), the value is “NA”
- If this is not the first result of a given expert (compatriot expert), and mark100 value is higher than the previous mark100 value, then the value is “TRUE”

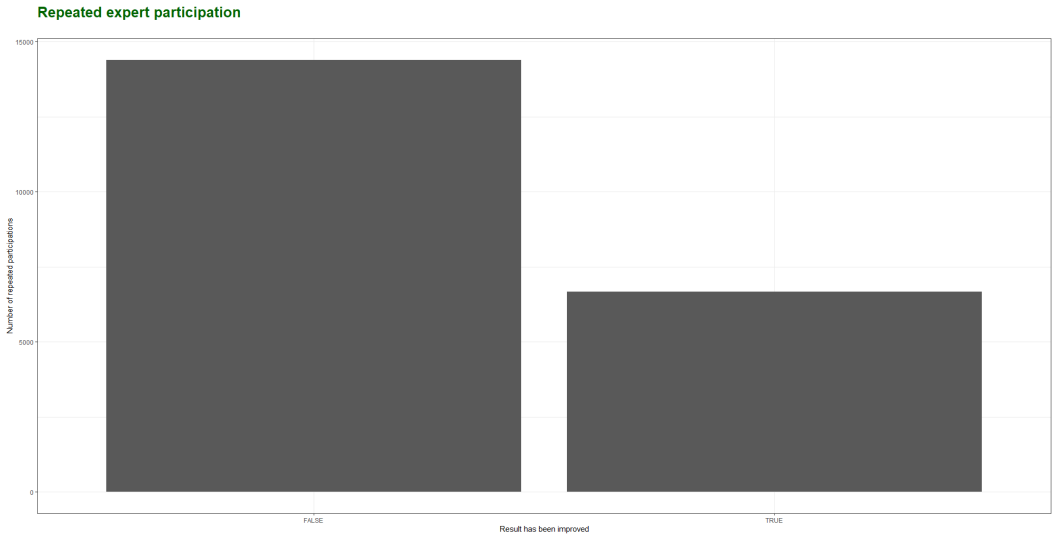
- If this is not the first result of a given expert (compatriot expert), and mark100 value is lower than the previous mark100 value, then the value is "FALSE.

The plotting of the grouped by region, repeated expert participation cases was visualized and resulted in the graph below. The "NA" values were excluded.



Bar chart of (repeated) expert participation per region

The above plot has been summarized for focus on the research question below.



Summary bar chart of (repeated) expert participation per region

 # A tibble: 2 × 3

improve100 n freq

<lgl> <int> <dbl>

1 FALSE 14393 0.684
2 TRUE 6663 0.316

Conclusions and recommendations

During the analysis of the first research question, whether repeated participation of a competitor has a positive impact on his/her results, we can conclude that in our delivered dataset the results of the competitor are about 50/50 in terms of getting improved scores when repeating the competition. The linear model shows a slight positive relationship, so that makes us feel confident to reject the null hypothesis.

Furthermore, the reason for a positive relationship could be explained as followed:

- Competitors can learn and improve if they participate more in competitions
    - Due to making mistakes and getting feedback
- They can get an increase in confidence while being in competitions often
    - Due to getting used to the stress, getting familiar with the format

When analyzing the second research question, whether repeated participation of an expert (compatriot expert) has a positive impact on the results of his/her competitor, we can conclude that we can't know for sure if the results of said competitor improves. This is seen in our dataset, as there are about 3x as many results which don't improve as there are results that do improve. There is a positive relationship in the linear model between expert participation and average results. This discrepancy highlights the limitation of relying on the linear models to capture this problem. It needs a deeper exploration into the data and its outliers. We are confident to fail to reject the null hypothesis.

The reason for a positive relationship could be explained as followed:

- The expertise of the experts, which could be domain specific, can lead to better scores
    - Due to knowledge transfer
- Experience of the experts can help the competitor get better results
    - Hands-on experience to navigate difficult subjects can help the competitor
- The expert and competitor can develop a mentor/student type relationship
    - Being a guide can help the competitor get better results

It is recommended for new competitors to know that doing multiple competitions can help get their scores to a level which they deem desirable. Perhaps going to regional competitions helps them

get acquainted with the level of skill before going to national competitions. If the possibility is available to get more repetitions at competitions, it seems valuable to do so.

Brief description of the division of work

The tasks regarding the report were divided between the two authors of this report, Alex & David.

Aleksandr Gorbachev:

- Was responsible for receiving the dataset from external suppliers related to his work
- Aleksandr has industry domain knowledge and understood the concepts and structure of the supplied dataset
- Aleksandr setup the initial draft of the research questions and script of R in Github, doing the analysis
- Validated research paper

David Langeveld:

- Was responsible as a validation partner of Aleksandr, as he did not have the same level of the domain knowledge
- Validated the draft and script in Github by reviewing all steps taken, came up with suggestions and additions to script and analysis where necessary
- Drafted the research paper

Appendix 1. Customizing the original dataset

The foremost step is to perform initial data analysis to spot useful variables and remove variables with inconsistent or useless values:

- take a brief look at the raw data
- transform dataset in the more convenient way for exploration purposes
- clean up observations which are missing important values for analysis

All data operations will be performed using R language. Note that this appendix might have small discrepancies with the attached R script.

Data importing

The supplied datasets (participants datafiles) are subject to inherent limitations, notably the absence of certain pertinent data (personal, location) crucial for comprehensive analysis, necessitating a thorough examination of all variables. The wrangled dataset resulted in a more focused dataframe which could be used for exploration. Next to the datasets that include all of the observations, another dataset (regions datafile) has been joined to give context to the region codes in the main dataset.

Data preparation

Exploration of the dataframes through the use of the glimpse function is employed to retrieve metadata (data about data) into the contents of the data frame under consideration in this paper. This analytical function is a component of the *dplyr* package. This glimpse function displays the initial entries for each variable in a tabular format, arranged horizontally following the respective variable names. Furthermore, the data type of each variable is presented in brackets immediately following the variable's name. The abbreviations *int* and *dbl* refer to "integer" and "double", within the context of computer programming, denoting quantitative or numerical variables. It is worth mentioning that "doubles" need double the storage space on a computer or database compared to integers. In contrast, *chr* corresponds to "character," a term denoting textual data in the programming world. There is a difference between the kinds of variables that are encountered in the data frames. There are identification variables and measurement variables. Identification variables are variables that uniquely identify each observational unit in case of competitors. The

other variables describe the properties of each observational unit. This will help in the dissection of the dataset.

Compute summary statistics & further analysis

As the datasets have been prepared and wrangled, the initial and fundamental step in exploratory data analysis (EDA) will be executed: examining the raw data values. This step is important in gaining an understanding of the raw data to assist in fixing issues later on. After taking a look at the data, the calculation of the summary statistics will be done from a regional point of view. Calculating the summary with the initial dataset was not possible as many observations were invalid or missing, for example from the original 600.000 number of rows, there were missing values in 250.734 of them. Cleaning the data beforehand made sure the data is of high enough quality to execute EDA.

The skim() output reports summaries for categorical variables (variable type: factor) separately from the numerical variables (variable type:numeric). For the categorical variable "region", it reports the following information: skim_variable, n_missing, complete_rate, ordered (Y/N), and n_unique, which are the number of missing, completed rate, if it's ordered or not, and total number of values.

These steps make sure the dataset is understandable, cleaned and prepared for exploration so the communication regarding the key results can be done to a broad audience.

First we will load Excel tables into R, merge all tables in a single dataframe and look at columns.

```
# Loading required libraries
library(readxl)
library(tidyverse)
library(readr)
library(skimr)
library(moderndive)
library(infer)

# Loading excel sheets
frame1 <- read_excel("Datafiles/participants100.xlsx")
frame2 <- read_excel("Datafiles/participants200.xlsx")
```

frame3 <- read_excel("Datafiles/participants300.xlsx")

Merging all sheets into a single dataframe
esim_data <- rbind(frame1, frame2, frame3)

Exploring dataframe columns
glimpse(esim_data)

Rows: 600,000
Columns: 25
$ pk_participant <dbl> 5, 6, 7, 9, 10, 11, 12, 13, 22, 24, 25, 26, 28, 29, 37, 40, 42,~
$ fkUsers <dbl> 84, 81, 79, 82, 85, 87, 86, 84, 81, 82, 84, 87, 79, 80, 79, 110~
$ fkChamp <dbl> 7, 7, 7, 7, 4, 4, 4, 4, 8, 11, 11, 11, 13, 13, 12, 14, 20, 12, ~
$ fkComp <dbl> 164, 164, 187, 164, 164, 164, 164, 164, 176, 165, 166, 166, 166~
$ ChampRole <dbl> 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 4, 5, 4, 5, 5, 5, 5, 5, ~
$ regionID <dbl> 76, 76, 76, 76, 76, 76, 76, 76, 76, 76, 76, 76, 76, 76, 76, 3, ~
$ mark100 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ mark500 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ medal <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ timestamp <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ fkUserAdd <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ competitorMarker <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ expertGroupMarker <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ excludeFromResault <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ fk_quotaCategory <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ fk_command <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ FK_USER_CP <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ ACCESS_RKC <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ FK_COMPATRIOT <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ organization <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~

$ nok <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ participant_updated_at <chr> "2020-12-07 11:50:38", "2020-12-07 11:50:38", "2020-12-07 11:50~
$ is_requested <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ is_accepted <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ mark700 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~

As we see, there are 600 000 observations and 25 columns total. There are many 0s and NA values at the first glance.

Let's take a random sample of 20 rows.

```
# Taking random 20 rows sample
sample <- esim_data[sample(nrow(esim_data), 20), ]
sample
# A tibble: 20 x 25
   pk_participant fkUsers fkChamp fkComp ChampRole regionID mark100 mark500 medal
timestamp
            <dbl>  <dbl>  <dbl> <dbl>    <dbl>    <dbl>   <dbl>   <dbl> <chr>
     <chr>
 1         454464 1187314   13980   341        4       68    NA       NA NA
     2019-11-~
 2         284414  122757    5471   203        4       49    NA       NA NA
     2019-03-~
 3         412062 1133052    8831   186        5       76    11.6    455 NA
     2019-06-~
 4          30299   25044     240   188        4        2     0        0 NA
     NA
 5         366107 1055466   12215   212        5       49    70      501 Medallion~
     2019-06-~
 6         295854    6514    6309   165        1       65    NA       NA NA
     2019-03-~
 7         248032 1065927    4821   423        5       76    31.0    468 NA
     2018-12-~
 8          76802   55518     622   345        4        4     0        0 NA
     NA
 9         213018     832    4444   373        1       76    NA       NA NA
     2018-09-~
10         442016 1180196   13729   422        4       76    NA       NA NA
     2019-10-~
11          37354   31490     300   203        4       58     0        0 NA
     NA
12         473468 1203706   15138   325        5       60    13.8    520 GOLD
     2019-11-~
13          18777   15659     170   184        4       49     0        0 NA
     NA
14         160537  999571    3361   315        5       49    60.5    525 Medallion~
     2018-05-~
15          45294   37705     316   261        5        8    97.5    533 SILVER
     NA
16         257694   94043    4655   169        4        6    NA       NA NA
     2019-01-~
```

17 64160 47841 435 203 5 76 64.8 504 Medallion~ NA
18 100791 3149 1009 203 4 84 NA NA NA NA
19 261426 19797 4461 250 4 65 NA NA NA 2019-02-~
20 383233 1041131 10664 247 4 65 NA NA NA 2019-06-~
i 15 more variables: fkUserAdd <dbl>, competitorMarker <lgl>, expertGroupMarker <chr>,
excludeFromResault <chr>, fk_quotaCategory <dbl>, fk_command <dbl>, FK_USER_CP <dbl>,
ACCESS_RKC <dbl>, FK_COMPATRIOT <dbl>, organization <lgl>, nok <dbl>,
participant_updated_at <chr>, is_requested <dbl>, is_accepted <dbl>, mark700 <lgl>

From this sample we see some observations which fall in our area of interest as well as some values which are missing the most important value mark100, which is a main measure for performance evaluation to be used in all tested hypotheses.

Exploring unique values for variables
Next step is to figure out which values of provided variables might be useful for the results explanation, as well as which variables are useless and can be discarded.

Based on raw data samples and provided description from data owner, there are some columns which are:

clearly related to the context of research, such as:

- pk_participant: ID of competition result record (primary key)
- fkUsers: competitor reference (foreign key)
- fkChamp: competition event reference (foreign key)
- fkComp: skill trade reference (foreign key)
- regionID: code of participant's region origin
- mark100: 100-point scale
- mark500: 500-point scale
- medal: type of medal awarded
- FK_COMPATRIOT: reference for compatriot expert ID (foreign key)
- timestamp: time when the result was locked in the system
- mark700: 700-point scale

used only for internal purposes of eSIM and can be easily discarded:

- fk_quotaCategory: quota category for the competition (foreign key)
- FK_USER_CP: user ID in the digital platform (foreign key)
- ACCESS_RKC: whether the regional competition center has access to the record
- fkUserAdd: User ID of the user who added the result to the system (foreign key)
- participant_updated_at: time of last record update
- is_requested: field for access in the business process
- is_accepted: field for access in the business process

not clear so it's good to take a closer look to these values:

- ChampRole: ID of the participant's role at the competition
- fk_command: reference for a team membership (foreign key)
- organization: organization represented by the participant
- nok: participation in an independent qualification assessment project (for demonstration exams)
- excludeFromResault: marker for "out of contest" result
- competitorMarker: marker for group competition
- expertGroupMarker: marker for specific expert group

So let's take a look at ChampRole first:

How many different values for roles?
length(unique(esim_data$ChampRole))
[1] 18

What are the competition roles?
unique(esim_data$ChampRole)
[1] 4 5 1 7 2 3 6 28 29 22 20 30 31 33 34 35 8 36
There are some identifiers for competition roles for participants (required to be entered in CIS), but since these values might not be consistent across all CIS instances (and therefore inside eSIM), given that we don't have actual values for these identifiers we can easily discard this column.

Next we will look at fk_command column:

How many teams identifiers in the dataset
length(unique(esim_data$fk_command))

[1] 6145

What are the values
head(unique(esim_data$fk_command),n = 10)
[1] 0 21 30 29 28 26 27 33 46 35
It seems that there are some team identifiers which are linked to other tables. Due to the absence of linked tables these values are not very useful but we can keep the column to identify whether this is a team or personal result.

Next, we explore is there any organization represented by competitors:

How many organizations are represented by competitors in the dataset?
length(unique(esim_data$organization))
[1] 1

unique(esim_data$organization)
[1] NA
The column is empty and clearly can be discarded.

Now let's investigate nok values:

How many unique values in nok column?
length(unique(esim_data$nok))
[1] 2

What are these unique values?
unique(esim_data$nok)
[1] 0 1

# Are there any results marked with nok=1 identifier?
length(esim_data$nok[esim_data$nok == "1"])
[1] 207
So we do have at least 207 values which are clearly represent demonstration exams. Note that it's not necessarily that every DE is marked with nok value, but either way it can be useful for future assumptions.

Next, look at the excludeFromResault variable:

```
# How many unique values?
length(unique(esim_data$excludeFromResault))
[1] 3

# What are the values?
unique(esim_data$excludeFromResault)
[1] NA    "YES" "N"

# How many results with non-NA value?
nrow(esim_data %>% filter(!is.na(excludeFromResault)))
[1] 13086

# Check consistency between provided values
tail(esim_data  %>%
     filter(excludeFromResault == "YES") %>%
     select(mark100, mark500, mark700, excludeFromResault))
# A tibble: 6 x 4
  mark100 mark500 mark700 excludeFromResault
    <dbl>   <dbl> <lgl>   <chr>
1   81.9      NA NA      YES   # These results are look consistent enough
2    9.24     NA NA      YES   # since they are excluded and therefore
3   11.6      NA NA      YES   # mark500/mark700 scores are not present
4   21.4      NA NA      YES   #
5   40.5      NA NA      YES   #
6   20.2      NA NA      YES   #

tail(esim_data  %>%
     filter(excludeFromResault == "N") %>%
     select(mark100, mark500, mark700, excludeFromResault))
# A tibble: 6 x 4
  mark100 mark500 mark700 excludeFromResault
    <dbl>   <dbl> <lgl>   <chr>
1   41.5    529 NA      N
2   24.8    471 NA      N
```

```
3    0       0 NA     N
4   49.3     NA NA    N     # Note that these results should not be
5   47.8     NA NA    N     # excluded but mark500/mark700 is missing
6    0       0 NA     N

tail(esim_data  %>%
     filter(is.na(excludeFromResault)) %>%
     select(mark100, mark500, mark700, excludeFromResault))
# A tibble: 6 x 4
  mark100 mark500 mark700 excludeFromResault
    <dbl>   <dbl> <lgl>   <chr>
1   NA       NA NA     NA    # Also note that there are some rows where
2   NA       NA NA     NA    # mark100 is missing. We will take it into
3   NA       NA NA     NA    # account later on.
4   NA       NA NA     NA    #
5   28.4    529 NA     NA    # These results clearly not excluded
6   20.4    476 NA     NA    # but marker is not present
```

So here we see that ~13k results are marked with excludeFromResault, which basically should mean that result is performed by a guest competitor, which doesn't add any valuable insights in our research context. Also we see that some observations marked with excludeFromResault are inconsistent, probably due to human error, therefore we won't keep this variable.

We also noted that column types of mark500 and mark700 are different. If we saw reasonable values for mark500, let's check whether we do have any values in mark700 other than NA:

```
# Check non-NA unique values for mark700
length(unique(esim_data$mark700))
[1] 1

unique(esim_data$mark700)
[1] NA
```

Clearly there are no 700-scale marks, most likely because it was introduced by WSI just in 2019 in CIS 4.0, but WSR were using older CIS 3.x for backward compatibility up until 2022, which had no support for 700-scale. So clearly we can get rid of this column as well.

Next we go over competitorMarker variable:

Are there any valuable group markers for competitors?
length(unique(esim_data$competitorMarker))
[1] 1

unique(esim_data$competitorMarker)
[1] NA
Column is empty and can be easily discarded.

Finally, we explore values for expertGroupMarker column:

Are there any group markers for experts?
length(unique(esim_data$expertGroupMarker))
[1] 8

unique(esim_data$expertGroupMarker)
[1] NA "CERT" "NO" "H" "C" "N" "W" "S"
There are some markers which are not really meaningful in the context of our research, so this variable can be discarded.

Transforming dataframe for brevity

After a brief clarification of unique variables values which were not clear at first sight, we now can remove meaningless columns from our initial dataframe:

```
# Removing columns which won't be used for research
esim_data <- subset(esim_data,
                    select = -c(ChampRole,
                                fkUserAdd,
                                competitorMarker,
                                expertGroupMarker,
                                fk_quotaCategory,
                                FK_USER_CP,
                                ACCESS_RKC,
                                organization,
                                nok,
```

```
        excludeFromResault,
        participant_updated_at,
        is_requested,
        is_accepted,
        mark700))
```

To not confuse with some of the variable names, we will rename some columns for clearer reference:

```
colnames(esim_data) <- c("result",
                         "competitor",
                         "competition",
                         "skill",
                         "region",
                         "mark100",
                         "mark500",
                         "medal",
                         "timestamp",
                         "team",
                         "expert")
```

So now our dataframe looks more clear and concise:

```
glimpse(esim_data)
Rows: 600,000
Columns: 13
$ result     <dbl> 5, 6, 7, 9, 10, 11, 12, 13, 22, 24, 25, 26, 28, 29, 37, 40, 42, 51, 52, 53,~
$ competitor  <dbl> 84, 81, 79, 82, 85, 87, 86, 84, 81, 82, 84, 87, 79, 80, 79, 110, 138, 308, ~
$ competition <dbl> 7, 7, 7, 7, 4, 4, 4, 4, 8, 11, 11, 11, 13, 13, 12, 14, 20, 12, 12, 12, 21, ~
$ skill      <dbl> 164, 164, 187, 164, 164, 164, 164, 164, 176, 165, 166, 166, 166, 166, 193, ~
$ region     <dbl> 76, 76, 76, 76, 76, 76, 76, 76, 76, 76, 76, 76, 76, 76, 76, 3, 85, 16, 16, ~
$ mark100    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ mark500    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ medal      <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
```

$ timestamp <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
$ team <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ expert <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~

Cleaning up observations

As we still have many observations with missing values on our main variable of interest mark100, which might be a human/migration error or test records, we would like to get rid of those observations where mark100 is 0 or NA:

```
# How many observations where mark100 is missing?
nrow(esim_data[is.na(esim_data$mark100), ])
[1] 250734

# How many observations where mark100 is 0?
nrow(esim_data %>% filter(mark100 == 0))
[1] 51901

# Removing observations where mark100 is missing (NA or 0)
esim_data <- esim_data %>%
  drop_na(mark100) %>%
  filter(mark100 > 0)
```

To keep observations relevant to stated research questions we need to consider observations with missing data of expert participation, since some of the results might be not from competitions but from demonstration exams, where participants don't have a compatriot expert.

```
# How many observations where mark100 is NA?
nrow(esim_data[is.na(esim_data$expert), ])
[1] 200409

# How many observations where mark100 is 0?
nrow(esim_data %>% filter(expert == 0))
[1] 33571

# Removing observations where expert is missing (NA or 0)
```

```
esim_data <- esim_data %>%
 drop_na(expert) %>%
 filter(expert > 0)
```

We also would like to check whether there are any duplicate rows after export of raw data from eSIM:

```
# Check for duplicate result IDs
length(unique(esim_data$result)) == nrow(esim_data)
[1] FALSE
```

So clearly there are some duplicate values which we would like to remove:

```
# Removing duplicate rows from dataframe
esim_data <- distinct(esim_data)
length(unique(esim_data$result)) == nrow(esim_data)
[1] TRUE

nrow(esim_data)
[1] 49034
```

To provide additional insights on competitors and experts region of origin, we will upload a table with region names and join region names to the resulting dataset.

```
# Load dataframe with region names
regions <- read_csv2("Datafiles/regions.csv")
glimpse(regions)

Rows: 164
Columns: 2
$ code       <dbl> 1, 101, 4, 2, 102, 702, 3, 103, 5, 180, 6, 7, 8, 9, 109, 10, 11, 111, 82~
$ regionName <chr> "Republic of Adygea", "Republic of Adygeya", "Republic of Altai", "Repub~

# Join region names column to the resulting data frame
esim_data <- inner_join(esim_data, regions, by = c("region" = "code"), na_matches = "na")
```