Introduction

World Health Organization (WHO) offers worldwide data from the coronavirus pandemic. In this study, the latest data on cumulative deaths per 100,000 population and total vaccine doses administered per 100 population of the European Union countries have been statistically analysed using R, which showed that immunisation against COVID-19 through vaccination was positively impacted and reduced the number of deaths in a certain population with specific range of vaccination doses.

Methods

Data size: Data provided for this study was conducted by World Health Organization (WHO). Here, the size of the data is the total cumulative deaths per 100,000 population ("Deaths_100000") in twenty-six European Union countries and the total vaccine doses administered per 100 ("Doses_100") population collected (Table.1)[1]. Therefore, the null hypothesis, H0, assumes that there is no relationship or no significant difference between the variances of these two groups (H0: "Deaths_100,000" = "Doses_100"). In contrast, the alternative hypothesis, H1, assumes that there is a relationship between the variances of these two groups at the significance level of $\alpha = 0.05$ (H1: "Deaths_100,000" ≠ "Doses_100").

Applying the Regression Model: First, prepare and save the data in a TXT or any other format (CSV, SAS, Stata, or SPSS) that R can read and load into memory. Then, apply a simple linear regression, as in this case, only one predictor variable exists by predicting y value, "Deaths_100000" (dependent variable), when plotted as a function of x, "Doses_100" (independent variable) forms a straight line. To begin with, install the "tidyverse" and "ggpubr" packages, theme_set(theme_pubr()); set the working directory of the location of the corresponding text file [setwd("users/Desktop/….); getwd(); and dir()]. Then, tell the program to read the file into R as a data frame that we will study as a table with a header that is separated with tabulation (Table.1).

To begin with, the data needs to be visualised and checked for its behaviour by creating a scatter plot, which displays the data by "deaths" as the "y" value (dependent variable) against the "vaccine doses" as the "x" value (independent variable), with a smooth line (Figure.1). Then, calculate the correlation coefficient to measure the strength of a linear relationship between two variables. If the value ranges from -1 to +1, then be able to build a linear regression model. Due to having a large correlation coefficient with a perfect negative relationship, it could proceed to generate a linear regression model of "y" as a function of "x", demonstrating the intercept and the slope (Table. 2). Subsequently, it could proceed with the creation of a regression line with a smooth line, which is represented by the linear model, created previously (Figure. 2). The outputs of the function of summary, residual standard error, and confidence intervals of the model, provided several components of calculated and created regression line, such as call, residuals, coefficients, significance, and the residual standard error (RSE), multiple R-squared, adjusted R-squared, F-statistic with the degree of freedom and the associated P-value that are used to check how well the model fits data (Table. 3 and 4).

Comparing Variances: Subsequently, we introduce the factor of the vaccine ratio, "Vacc_ratio", with two categories of "High" for doses greater than 202 and "Low" for doses smaller than 202 by adding a column to the Table.1 (Table. 5). This makes group data inside the data object according to the "High" and "Low" vaccination ratio.

Here, first, looked at the summary table to check how the data is exhibited, which outlines the count, n, the number of individuals (in this case countries), mean, and standard deviation of Deaths_100,000, in two categories of high and low vaccine doses (Table.6)

To further test whether there is a difference between these two categories, high and low vaccine doses, used analysis of variances (ANOVA). To be able to use this data in ANOVA, first, named the factor, "Vacc_ratio", and then created the categories in order, "High" and "Low", using R's function, "levels" and "ordered" [levels(data$Vacc_ratio), data$Vacc_ratio <- ordered(data$Vacc_ratio, levels = c("High", "Low")), and levels(data$Vacc_ratio)]. Then, we could visualize the data by generating a box plot with "Vacc_ratio" in the x-axis and the Deaths_100,000 in the y-axis, ordered in High and Low categories using the "ggpubr" and "ggplot2" packages (Figure. 3).

Also, to check the distribution of data using a randomized x-axis to disperse the points, we could create a plot with the means and error bars using R's "jitter" functions which represents data points in the form of single dots in a similar manner to a scatter plot, but the "jitter" helps visualize the relationship between a measurement variable and a categorical variable and show the distribution of data using a randomized x-axis to disperse the points(Figure. 4). In addition, we could create another box plot demonstrating the whiskers of minimum and maximum levels in each group and indicating an outlier, three SD away from the mean, at a low vaccine population (Figure. 5).

Finally, compute the analysis of variance (ANOVA), to compare the "Deaths_100,000" regarding the factor "Vacc_ratio" using R's "res.aov" function. Nevertheless, checking the ANOVA assumption needs to test both the validity, as the ANOVA test assumes homogeneous variances and demonstrates the data's normality. To do this, first, performed the "leventeTest", a test for homogeneity of variance with groups median, on "Deaths_100,000" regarding the factor of "Vacc_ratio". The homogeneity of variance test revealed the degree of freedom of 1 and 24, the F-value of 0.1324 and the p-value of 0.7192, which is greater than the level of significance, α=0.05. Hence, the null hypothesis would be accepted as there is no difference between the variances of these two groups (high and low vaccination rates), and they seem to be similar. This result can be plotted, which displays the calculated median (horizontal red line) and all the residual data around the median (Figure. 6).

In addition to "leveneTest", which indicates the similarities in the two variances, to apply ANOVA, it needs to demonstrate the normality of the data, which would be done through several different tests such as "shapiro-wilk" test. Finally, apply the ANOVA by performing a summary of the results of the analysis of variance [summary(res.aov)]. Nevertheless, mathematically, the ANOVA does not describe which group is greater than the other, which must be done through different types of "PostHoc". In this case, using the "Tukey" [TukeyHSD(res.aov)].

Results

Descriptive analysis shows that increased vaccination may reduce the number of deaths.

To check the sample data (Table.1), a scatter plot was created to display the level of death (dependent variable) on the y-axis against the vaccination doses (independent variable) on the x-axis, with a smooth line (Figure.1). The plot suggests a linearly decreasing relationship between the vaccination rate and the mortality variables. Nevertheless, at the extremely high level of vaccine doses does not correlate with a decrease in mortality. Therefore, to assess the strength of the level of the association between the two data variables (x and y) computed, the correlation coefficient. The result demonstrates a large correlation coefficient of -0.7235914 with a perfect negative correlation.

The result from the correlation coefficient allowed the lead to continue building a linear regression model of "y" as a function of "x", which showed an intercept value of around 600, the place where the fitted line would cross the y-axis at the zero value of the x-axis, which may seem meaningless, but essential to know for further calculation (Table. 2).

Created the linear regression model demonstrating all the geometrical points and calculated the smooth line represented as the theoretical line for 26 European countries (Figure. 2). The y-axis at the level of zero-point, started from 603.487, which is the intercept value, and the x-axis started from 50 to 250, which are the vaccine doses administered to a certain population. Here, the summary outputs elucidated several components of the regression line, including the function of "Call", which is used to compute the regression model; the "Residuals" that provide a quick view of the distribution of the residual, which by definition gives a mean of zero; therefore, the median, 7.4, is not far from zero and the data is more or less normal; the minimum and maximum which should be roughly equal in absolute value because the mean has stated at zero; the coefficient that shows the regression better coefficient; and the significant codes, Residual Standard Error, Multiple R-Squared, Adjusted R-Squared, F-statistic, Degree of Freedom, and P-value that demonstrated that the model fits our data very well. This was supported by the F-statistic value and its associated p-value, which is significantly smaller than the threshold level $\alpha = 0.05$. Hence, the null hypothesis is rejected, and the alternative hypothesis is accepted and explains that there is a strong relationship between the variances of these two groups at the significance level of $\alpha = 0.05$ (H1: "Deaths_100,000" ≠ "Doses_100" or $\beta_i \neq 0$) (Table. 3 and 4).

Comparative analysis of variances reveals a significant reduction in mortality in the high-dose vaccinated population

To distinguish the effectiveness of vaccine doses in the data, a factor of "vaccine ratio" with two categories of "High" and "low" vaccine doses administered was created (Table. 5) and summarised in a table, which indicated the vaccination rate at the two levels, "High" and "low"; the counts for individuals, 14 countries with the high and 12 with the low level of vaccine doses; the mean number of Deaths_100,000 population, 194 and 375 with their standard deviation 67.2 and 87.8 for "High" and "low" respectively (Table. 6). This was further visualised the in a box plot, which showed that there seems to be a clear difference between the two categories (Figure. 3).

Furthermore, the "jitter" plot helped visualize the distribution and relationship of the two categories with roughly no overlapping of any dots (Figure. 4). In addition, we could create another box plot demonstrating the whiskers of minimum and maximum levels in each group and indicating an outlier, three SD away from the mean, at a low vaccine population (Figure. 5).

Finally, created an object of analysis of variance (ANOVA) of residuals, using the R's function "residuals" and the object "res.aov", that was created earlier. Before starting the ANOVA analysis, the "leveneTest" test was performed, which checks for variance homogeneity in the groups' centre with the median. This resulted in a p-value over the significant level, 0.7192. Its generated plot exhibited no differences between the two variances and seems to be similar or homogeneous (Figure. 6). Although the results from "leveneTest" indicated the homogeneity of the two variances, applying the ANOVA still needs to demonstrate the normality of the data, which would be done through several different tests such as "shapiro-wilk".

The results from the Shapiro-Wilk test demonstrate a "Calc W" value of 0.98395, which is between 0 and 1, and is a perfect match. In addition, the test indicates a p-value of 0.9448, which is greater than 0.05. This indicates that the distribution of the sample is not significantly different from a normal distribution. Hence, the null hypothesis cannot be rejected. Therefore, the sample has been generated from a normal distribution. Moreover, the quantile-quantile (Q-Q) plot further supports this result, which draws the correlation between a given sample and a 45-degree, theoretically perfect normal distribution line (Figure. 7). Here, apart from a few individuals, our data are relatively around the normal distribution line. Consequently, by having a homogeneous population with normal distribution, we could apply an ANOVA test, which is revealed with the list of the degree of freedom (1 and 24), the squared sum, the mean squared sum, the very high F-value (35.57), and the extremely small p-value, (3.7×10^{-6}) which is smaller than the threshold level, $\alpha = 0.05$. Therefore, with extreme confidence, we could draw a conclusion that there is a difference between both populations (High and Low vaccine doses).

Furthermore, the "PostHoc", in this case, used the "Tukey" [TukeyHSD(res.aov)] mathematically revealed that multiple comparisons for the means of these groups, and taking to account with 95% confidence, that the differences between the two groups, 181.4144, Deaths_100,000. It also gives us a confidence interval of around 95% confidence, which would be between 118.6358 to 244.1931 and an extremely small p-value. This result indicates that with 95% confidence, the population with lower vaccine doses administered per 100 population has around 181 deaths per 100,000, greater than the population with higher vaccine doses administered with associated extremely smaller p-values indicating statistically significant at the level of $\alpha = 0.05$. This means that the increasing number of vaccine doses had an impact on reducing the number of deaths per 100,000 population at a significant level, 3.72×10^{-6}, smaller than the threshold level, $\alpha = 0.05$ (Table. 7).

Discussion

Checking the sample data (Table.1) by the scatter plot with a smooth line of vaccination rate against the mortality suggested a linearly decreasing relationship between the two variables (Figure.1). Wherein increasing the vaccine doses decreases the number of deaths up until some point, which then increases again at the extremely high vaccine doses, and that cannot be correlated with the decrease in mortality. While the correlation does not mean causation in any statistical analysis, this contrast might not be due to the extreme increase in vaccine doses. It may relate to or be triggered by other factors, including the alteration in immune function, average age, diet, healthiness, temperature, climatic conditions etc. Hence, in future, these results must be compared with other countries that had a similar level of vaccine doses administered to check if they had the same increase in deaths by the extreme vaccine doses.

Nevertheless, the calculated high level of the correlation coefficient, -0.7235914, indicated a strong strength of a linear relationship between the two variables (deaths and vaccine doses), which allowed us to continue building a linear regression model to describe the relationship between the dependent variable, mortality rates, against the independent variable, vaccination rates, which displayed to have an intercept of around 600 and the slope of -1.7 with perfect negative relationship (Table2). Consequently, the created regression model and the line indicated the perfect negative correlation with a perfect theoretical line with a normal distribution (Figure. 2). This was further supported by the summary results of the model with F-value, t-student, R-squared, Adjusted R-squared, the residual standard error, the confidence intervals, and the associated p-values indicating significantly smaller p-value than the threshold level α = 0.05 (Table. 3). Therefore, the null hypothesis is rejected, and the alternative hypothesis is accepted and explains that there is a strong relationship between the variances of these two groups at the significance level of α = 0.05 (H1: "Deaths_100,000" ≠ "Doses_100").

The result summary of comparing variances that have been created by a factor of "vaccine ratio" with two categories of "High" and "low" vaccine doses administered shows a group of fourteen countries with the high vaccine doses administered displays a lower mean value of mortality (194) in comparison to the group with low vaccine doses (375) (Table. 6). This was also supported by standard deviation, 67.2 and 87.8 for "High" and "low groups respectively, which demonstrates strong evidence of differences between the two groups (Table. 6)

Furthermore, when these two categories were compared in a box plot, in accordant with what has been seen in the summary results (Table. 6), they clearly supported the differences between the two variances as there are real differences in the number of deaths regarding vaccination rate (Figure. 3). Moreover, creating a "jitter" plot shows the distribution of two categories with the real values of differences in the number of deaths regarding the vaccination rates with almost no overlapping of data dot points (Figure. 4).

Finally, results from the ANOVA and Tukey revealed that with extreme confidence, we could draw deduce that there is a difference between both populations (High and Low vaccine doses), in which increasing the number of vaccine doses had an impact on reducing the number of deaths per at a significant level (Table. 7).

Conclusion

In conclusion, the results from applying the regression model and comparing the two variances of high and low levels of vaccine doses revealed a strong relationship between the number of deaths and the number of administered vaccine doses. Hence, with the increasing number of vaccine doses, the mortality rate was reduced at a highly significant level.

References

1. https://covid19.who.int/table
2. Statistics & Data Visualization using R by David S. Brown

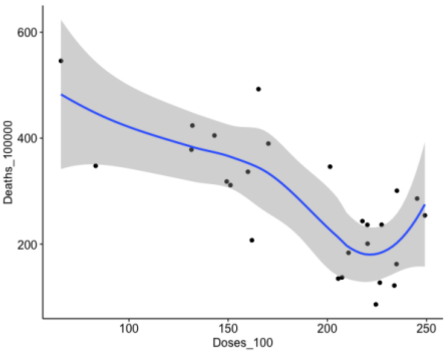

Figure 1. The scatter plot of accumulative deaths per 100,000 population and total vaccine doses administered per 100 population of the European Union countries. The plot displays the data by "Deaths_100,000" in the y-axis (dependent variable) against the "Doses_100" in the x-axis (independent variable) with a smooth line using R functions [ggplot(data, aes(x=Doses_100, y=Deaths_100000)) + geom_point() + stat_smooth()].

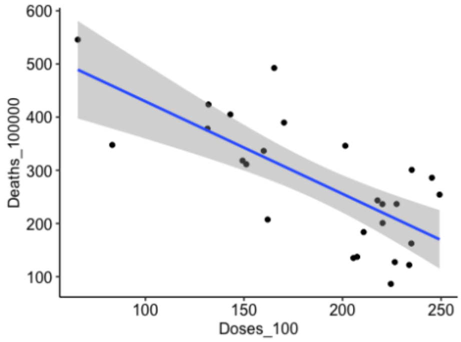

Figure 2. The creation scatter plot with the regression line of the mortality against the vaccination rates in the European Union countries. The plot displays the data by "Deaths_100,000" in the y-axis (dependent variable) against the "Doses_100" in the x-axis (independent variable) with a smooth line using R functions [stat_smooth() ggplot(data, aes(Doses_100, Deaths_100000)) + geom_point() + stat_smooth(method = lm)].

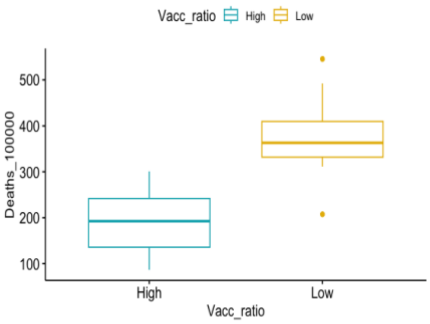

Figure 3. The box plot of mortality rate against the vaccination ratio in the European Union countries. The plot displays the two categories of "High" and "Low" vaccine doses administered. The two categories seem to be different. Using R functions [ggboxplot(data, x = "Vacc_ratio", y = "Deaths_100000", color = "Vacc_ratio", palette = c("#00AFBB", "#E7B800", "#FC4EO7"), order = c("High", "Low"), ylab = "Deaths_100000", xlab = "Vacc ratio")].

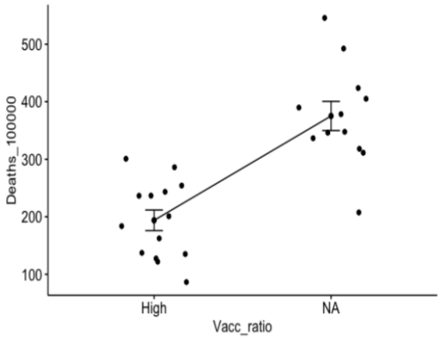

Figure 4. The jitter plot of mortality rate against the vaccination ratio in the European Union countries. The plot represents data points in single dots, which helps visualise the relationship between the two categories of "High" and "Low" vaccine doses administered. The two categories seem to be different. Using R functions [ggboxplot(data, x = "Vacc_ratio", y = "Deaths_100000", color = "Vacc_ratio", palette = c("#00AFBB", "#E7B800", "#FC4EO7"), order = c("High", "Low"), ylab = "Deaths_100000", xlab = "Vacc_ratio")].

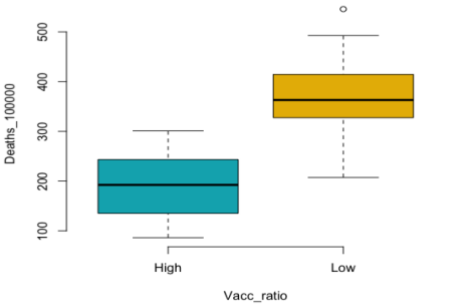

Figure 5. The box plot of mortality rate against the vaccination ratio in the European Union countries. The plot displays the two categories of "High" and "Low" vaccine doses administered in corresponding countries. The two categories seem to differ, with an outlier in the "Low" category. Using R functions [boxplot(Deaths_100000~Vacc_ratio, data = data, xlab = "Vacc_ratio", ylab = "Deaths_100000", frame = FALSE, col = c("#00AFBB", "#E7B800", "#FC4E07"))].

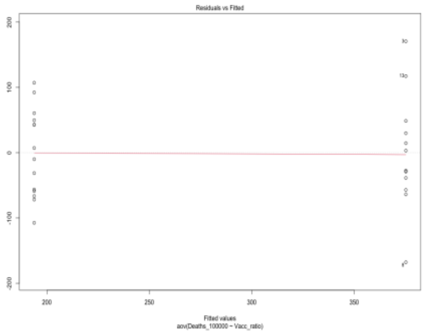

Figure 6. The LeveneTest plot for homogeneity of variances (center=median). The plot displays the median line in the middle with all residuals two categories of "High" and "Low" vaccine doses administered at the two sides of the plot. The two categories seem to differ, with an outlier in the "Low" category. Using R functions [boxplot(Deaths_100000~Vacc_ratio, data = data, xlab = "Vacc_ratio", ylab = "Deaths_100000", frame = FALSE, col = c("#00AFBB", "#E7B800", "#FC4E07"))].

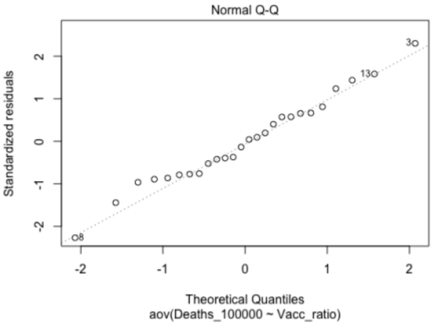

Figure 7. The quantile-quantile (Q-Q) plot of standardised residuals against the theoretical quantiles. The plot displays the correlation between a given sample and a 45-degree, theoretically perfect normal distribution line. Using R functions [aov_residuals <- residuals(object = res.aov) shapiro.test(x = aov_residuals) plot(res.aov, 2)]. c

	Country	Deaths_100000	Doses_100
1	Austria	236.48	220.2
2	Belgium	286.04	245.3
3	Bulgaria	545.73	65.7
4	Croatia	423.84	132
5	Cyprus	135.13	205.5
6	Czechia	389.77	170.3
7	Denmark	127.43	226.5
8	Estonia	207.45	162
9	Finland	122	233.8
10	France	236.71	227.4
11	Germany	86.39	224.5
12	Greece	346.16	201.4
13	Hungary	492.46	165.32
14	Ireland	162.48	234.9
15	Italy	300.86	235.1
16	Latvia	318.19	149.3
17	Lithuania	336.57	160
18	Luxembourg	183.83	210.7
19	Netherlands	137.31	207.4
20	Poland	311.38	151.2
21	Portugal	254.28	249.2
22	Romania	347.75	83.2
23	Slovakia	378.37	131.5
24	Slovenia	405.04	143.1
25	Spain	243.47	217.7
26	Sweden	200.95	220.3

Table 1. Cumulative deaths per 100,000 population and total vaccine doses administered per 100 population in European Union countries [1]. Using the R's functions [data<- read.table("Data.txt", header = TRUE, sep = "\t")].

# Computation of the LNR	
Call:	
lm(formula = Deaths_100000 ~ Doses_100, data = data)	
Coefficients:	
(Intercept)	Doses_100
603.487	-1.739

Table 2. Generation of a linear regression model of "y" as a function of "x". The intercept, where the fitted line would cross the y-axis, is 603.48 at the zero value of the x-axis with a perfect negative relationship regarding the independent variable, Doses_100, with a slope of -1.739. Using R's function [model <- lm(Deaths_100000 ~ Doses_100, data = data); model].

> ## Regression line					
summary(model)					
Call:					
lm(formula = Deaths_100000 ~ Doses_100, data = data)					
Residuals:					
Min	1Q	Median	3Q	Max	
-126.713	-69.516	7.429	54.965	176.449	
Coefficients:					
	Estimate	Std.Error	t value	Pr(>|t|)	
(Intercept)	603.4869	65.5729	9.203	2.42E-09	***
Doses_100	-1.7389	0.3386	-5.136	2.95E-05	***

Signif. codes:	0'***'	0.001'**'	0.01'*'	0.05'.'	0.1"1
Residual standard error:		84.08 on 24 degrees of freedom			
Multiple R-squared:		0.5236,	Adjusted R-squared:		0.5037
F-statistic:	26.38 on 1 and 24 DF,		p-value:	2.95E-05	

Table 3. The summary outputs of the linear regression line. The computation of linear regression provides several components, including call, residuals, coefficients, significant codes, residual standard error, multiple and adjusted R-squares, and F-statistics with the p-value. Using R's function, [summary(model)].

#Residual Standar Error:		
> sigma(model)*100/mean(data$Doses_100)		
44.85667		
> sigma(model)*100/mean(data$Deaths_100000)		
30.29487		
> confint(model)		
	2.50%	97.50%
(Intercept)	468.151068	738.822666
Doses_100	-2.437714	-1.040097

Table 4. Residual Standard Error and Confidence intervals of the linear regression line. Using R's functions: [sigma(model)*100/mean(data$Doses_100)]; [sigma(model)*100/mean(data$Deaths_100000)]; and [confint(model)].

	Country	Deaths_100000	Doses_100	Vacc_ratio
1	Austria	236.48	220.2	High
2	Belgium	286.04	245.3	High
3	Bulgaria	545.73	65.7	Low
4	Croatia	423.84	132	Low
5	Cyprus	135.13	205.5	High
6	Czechia	389.77	170.3	Low
7	Denmark	127.43	226.5	High
8	Estonia	207.45	162	Low
9	Finland	122	233.8	High
10	France	236.71	227.4	High
11	Germany	86.39	224.5	High
12	Greece	346.16	201.4	Low
13	Hungary	492.46	165.32	Low
14	Ireland	162.48	234.9	High
15	Italy	300.86	235.1	High
16	Latvia	318.19	149.3	Low
17	Lithuania	336.57	160	Low
18	Luxembourg	183.83	210.7	High
19	Netherlands	137.31	207.4	High
20	Poland	311.38	151.2	Low
21	Portugal	254.28	249.2	High
22	Romania	347.75	83.2	Low
23	Slovakia	378.37	131.5	Low
24	Slovenia	405.04	143.1	Low
25	Spain	243.47	217.7	High
26	Sweden	200.95	220.3	High

Table 5. Mutated Table.1 by adding the factor of vaccination ratio, "Vacc_ratio", with two categories of "High" (for doses more than 202, >202) and "low" (for doses smaller than 202, _else). Using R's function [data <- data %>% mutate ("Vacc_ratio" = if_else(Doses_100>202, "High", "Low"))].

	Vacc_ratio	count	mean	sd
1	High	14	194	67.2
2	Low	12	375	87.8

Table 6. The R's "summaries" function. Adding the factor of vaccination ratio, "Vacc_ratio", with two categories of "High" (for doses more than 202) and "low" (for doses less than 202). The count, n, is the number of individuals (countries), the mean and the standard deviation values of the number of deaths per 100,000 population in each low and high category group. Using R's function [group_by(data, Vacc_ratio) %>% summarise(ount = n(), mean = mean(Deaths_100000, na.rm = TRUE), sd = sd(Deaths_100000, na.rm = TRUE))]

A.

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
Vacc_ratio	1	212657	212657	35.57	3.72E-06 ***
Residuals	24	143481	5978		
Signif. codes:	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

B.

95% family-wise confidence level				
Fit: aov(formula = Deaths_100000 ~ Vacc_ratio, data = data)				
$Vacc_ratio				
	diff	lwr	upr	p adj
Low-High	181.4144	118.6358	244.1931	3.70E-06

Table 7. The summary of the results of analysis of variance (A) and the Tuckey multiple comparison (B) of means of two categories of "High" (for doses more than 202) and "low" (for doses less than 202). Using R's function [summary(res.aov)] and [TukeyHSD(res.aov)]