National College of Ireland

BSc (Hons) in Computing

Data Analytics

2021/2022

Izaskun Lekue

X17105595

X17105595@student.ncirl.ie

Happiness and physical activity

Technical Report

Contents

Executive Summary

Lately there is a lot of talk about how physical activity produces beneficial physiological effects to control stress. The practice (always better outdoors and in good company) encourages the production of endorphins, dopamine and serotonin, which are hormones of happiness. Sport also favours mental states in which we isolate ourselves from worries and anxieties.

My technical reports consist in analysing the difference between physical activity performed in different countries and if those countries where they perform more activity are the ones with higher happiness score.

This document explains and details the methodology used and steps taken to analyse if there is any relation between physical activity performed and happiness of people in 27 European countries.

For this I have focused on 3 primary databases with information on deaths related to lack of sport or sedentary life, happiness score and amount of physical activity performed in each country.

The main sections of the report are: Data (describes the data chosen for the project), Methodology (explains the methodology used, KDD), Analysis(how the analysis has been conducted) and Results(the section with the outcomes and visualizations of the analysis).

From this analysis, it was concluded that there is a significant difference between Meridional Europe and Septentrional Europe countries and that there is a significant relation between amount of physical activity performed and happiness scores.

1.0 Introduction

1.1. Background

As a person interested and active in sports, I have been able to verify for myself how when I perform any kind of physical activity, I feel happier and mentally more relaxed than when I am not that active.

Scientific studies prove the benefits of physical exercise, and the advantages cover various systems and organs of the body.

– Improves cardiovascular health. Regular physical exercise trains the heart to beat more slowly and forcefully, requiring less oxygen to function properly.

To obtain cardiovascular benefits through physical activity, it is recommended to add 30 minutes of walking per day.

– Lowers the risk of heart attack and stroke. Bad lifestyle habits favour the accumulation of 'atheroma plaque' (fat) in the arteries, gradually clogging them and increasing the risk that they become completely clogged and blood does not reach the heart or brain.

– Reduces the risk of type 2 diabetes. Regular physical exercise helps to metabolize glucose, since muscles used to work are more receptive to insulin, a hormone that introduces blood sugar into cells.

For this reason, I always have been very interested in analysing how an active lifestyle affects us, and how affects different countries.

Physical activity positively influences people's health in various dimensions. On the one hand, the physical field, and its link in the quality of life when developing systems of the human body such as cardiovascular, neuromuscular and locomotor to name a few. On the other hand, the social, affective, and cognitive consequences of maintaining movement habits, particularly in young people who claim to have less time as their studies cover their work to a greater extent.

The initial idea was to analyse the relation between happiness and physical activity in Ireland, but due to lack of relevant data I did the analysis of the following countries:

Austria, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden

The goal of this project is to analyse how physical activity reflects on a better life or life quality, and to answer the following questions:

- How many people practices physical activity?
- How physical activity affects on them?
- Does gender interfere on the amount of physical activity performed?
- Is there any relation between the physical activity performed and happiness on each country?

1.2. Aims

The objective of this project is to analyse the link between physical activity and life quality in Ireland.

Although this is a topic already discussed by different people, now after the pandemic, it Is very interesting to do this type of analysis and bring to light the important link between sports, an active lifestyle and quality of life.

Aim 1: Find and select appropriate and relevant data. Dataset related to how much physical activity people performs, happiness scores and death causes needed to do the analysis.

Aim 2: Clean the datasets, remove non relevant information, on some datasets there is loads of information that is not related to my analysis. This is done using RStudio to remove nulls and columns and Excel.

Aim 3: Analyse data to see if there is any relation between amount of physical activity performed, death causes and happiness score. This is done using SPSS, RStudio, Tableau and Excel.

Aim 4: Analyse and interpret the results and graphics.

Aim 5: Complete documentation with the insights of this analysis, does physical activity affect to happiness?

1.3. Technology

Excel: is an easy spreadsheet tool to use, whenever we need to process data, clean databases, and obtain accurate conclusions. In the business world, it is very common to perform data analysis based on the area of the company and Excel is a fantastic tool for processing the information. Most of the datasets I have downloaded are in excel format, is a good tool to delete not necessary data, create needed formulas and visualize data on pivot tables and graphics.

R Language: R is a programming language used mainly in data analytics, statistics, finances... R is a free software environment and interpreted programming language, that is, it executes the instructions directly, without prior compilation of the program to machine language instructions. The term environment, in R, refers to a fully planned and consistent system, rather than an accumulation of specific and inflexible tools, as is often the case in other data analysis software. The most used IDE is RStudio which is what I will be using.

RStudio: is a web application that allows to develop with R and other programming languages oriented to the treatment of large amounts of data, statistics… It is a complete development IDE, but embedded in a web application, which also allows integration with a series of project management tools.

Is an open source programming language aimed at working with data and its statistical analysis, used mainly in the field of mathematical research and machine learning, data mining... It is multiplatform, so it can be used in any system desktop operating.

Tableau: is a visual analytics platform that allows users to take all kinds of data from almost any system and turns it into useful information quickly and easily. It is very intuitive and have multiple visualization options. This was used to create map graphics.

SPSS: IBM SPSS Statistics is a software that provides researchers with tools to quickly query data and formulate hypotheses, run procedures to clarify relationships between variables, identify trends, and make predictions. SPSS is used for a wide range of statistical analyses, including descriptive statistics (means, frequencies), bivariate statistics (eg, analysis of variance, t-test), regression, factor analysis, and representation. graph of the data.

1.4. Structure

Section 2 Data**:** this describes the data chosen for the project and how I plan to use it.

Section 3 Methodology: describes the methodology used which is KDD.

Section 4 Analysis**:** explains how I analyse the data in order to show the results I was looking for.

Section 5 Results**:** here I describe the results I have collected from my analysis.

Section 6 Conclusion: considering the results on section 5, here will be the conclusion about my findings.

Section 7 Further development: potential future work related to this analysis.

Section 8 References**:** list of all the sources used in the project.

Section 9 Appendices**:** contains project plan and monthly reflective journals.

2.0 Data

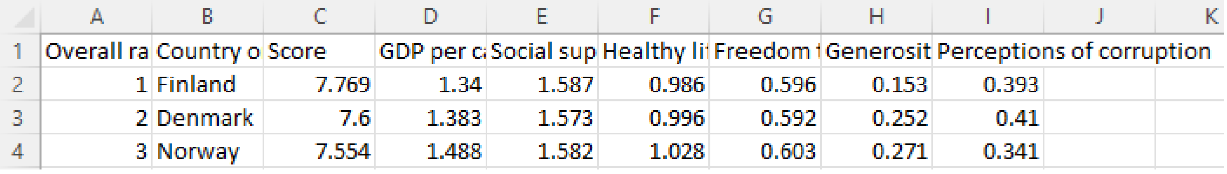
In this section there is a description of all the primary datasets used for the investigation.

Name	world_happiness2019
Structure	Structured
Total columns	9
Total rows	157
File format	CSV

Description: World happiness report year 2019 dataset contains the happiness score of 156 countries, it also contains scores of GPD per capita, healthy lifestyle, generosity and perception of corruption. From this dataset I only considered the happiness of score, all the rest was not necessary on this particular analysis.

Variables: “Overall rank”, “Country or region”,” Score, GDP per capita”,” Social support”, “Healthy life expectancy”,” Freedom to make life choices”,” Generosity” and “Perceptions of corruption”.

Source: https://www.kaggle.com/datasets/mathurinache/world-happiness-report?select=2020.csv



	A	B	C	D	E	F	G	H	I	J	K
1	Overall ra	Country o	Score	GDP per c	Social sup	Healthy li	Freedom t	Generosit	Perceptions of corruption		
2	1	Finland	7.769	1.34	1.587	0.986	0.596	0.153	0.393		
3	2	Denmark	7.6	1.383	1.573	0.996	0.592	0.252	0.41		
4	3	Norway	7.554	1.488	1.582	1.028	0.603	0.271	0.341		

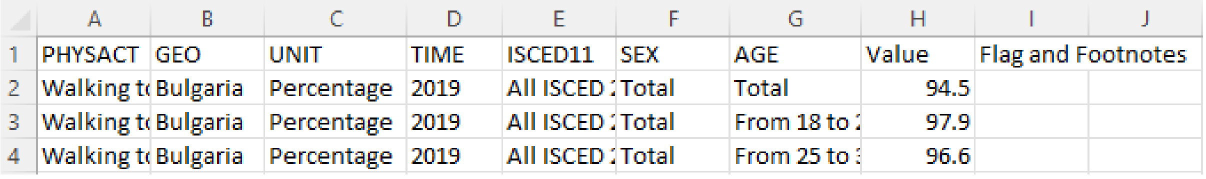
Figure 1- world_happiness2019.csv

Name	hlth_ehis_pe3e_1_Data
Structure	Structured
Total columns	9
Total rows	8353
File format	CSV

Description: Physical activity divided by country and sex. This dataset contains the amount of physical activity performed by country and by sex in 29 countries.

Variables: “PHYSACT”, “GEO”, “UNIT”, “TIME”, “ISCED11”, “SEX”, “AGE”, “Value” and ‘Flag and Footnotes”.

Source: https://ec.europa.eu/eurostat/



	A	B	C	D	E	F	G	H	I	J
1	PHYSACT	GEO	UNIT	TIME	ISCED11	SEX	AGE	Value	Flag and Footnotes	
2	Walking t	Bulgaria	Percentage	2019	All ISCED 2	Total	Total	94.5		
3	Walking t	Bulgaria	Percentage	2019	All ISCED 2	Total	From 18 to 2	97.9		
4	Walking t	Bulgaria	Percentage	2019	All ISCED 2	Total	From 25 to 3	96.6		

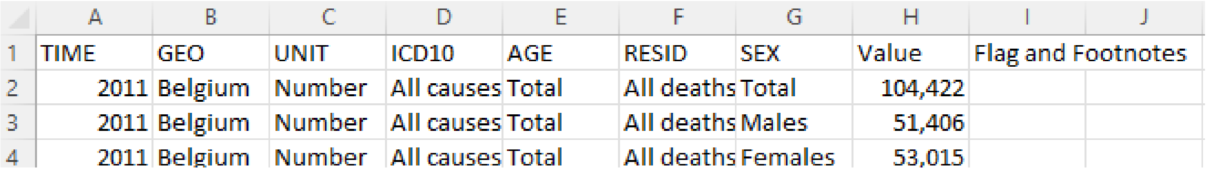
Figure 2- hlt_ehis_pe3e_1_Data.csv

Name	hlth_cd_aro_1_Data
Structure	Structured
Total columns	9
Total rows	50401
File format	CSV

Description: Death causes from 2011-2019 in Europe, deaths related to lack of sport and care, such as diabetes, heart diseases and cerebrovascular diseases in 29 countries.

Variables: "TIME", "GEO", "UNIT", "ICD10", "AGE", "RESID", "SEX", "Value" and "Flag and Footnotes".

Source: https://ec.europa.eu/eurostat/



	A	B	C	D	E	F	G	H	I	J
1	TIME	GEO	UNIT	ICD10	AGE	RESID	SEX	Value	Flag and Footnotes	
2	2011	Belgium	Number	All causes	Total	All deaths	Total	104,422		
3	2011	Belgium	Number	All causes	Total	All deaths	Males	51,406		
4	2011	Belgium	Number	All causes	Total	All deaths	Females	53,015		

Figure 3- hlth_cd_aro_1_Data.csv

3.0 Methodology

I followed KDD methodology, is an automatic process in which discovery and analysis are combined. The process consists of extracting patterns in form of rules or functions, from the data, for the user to analyse them. This task involves pre-processing the data, doing data mining and presenting the results.

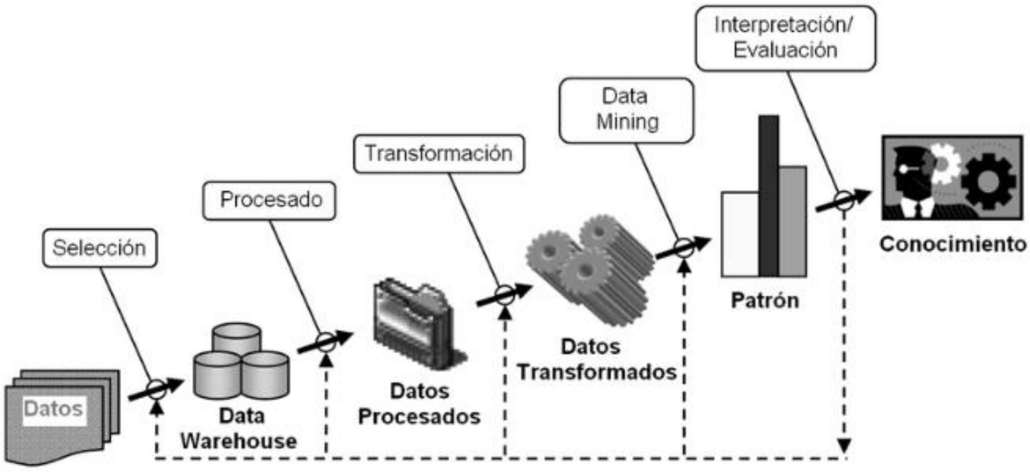


Figure 4- KDD methodology steps

3.1 Data selection

In this stage, once the goals were identified, is where I searched for the necessary data. The initial idea was to analyse the relation between physical activity and happiness in Ireland, but I was not able to identify proper data for Ireland and instead I proceeded to analyse the data of most of the EU countries: Austria, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain and Sweden.

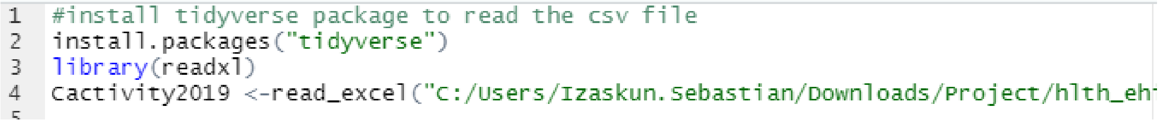
The data was extracted from https://ec.europa.eu/eurostat, an statistical office of the European Union with the mission of providing high quality statistics.

3.2 Processing and Transformation

On this phase is when the reliability of the information is determined, this means, performing tasks that guarantee the usefulness of the data. For this, data cleaning is done which involves removing variables or attributes with missing data or removing non-useful information. The quality of the data is also improved with transformations that involve reduction of the number of variables in the dataset.

After having selected all the appropriate datasets, I prepared and cleaned the data using RStudio and Excel.

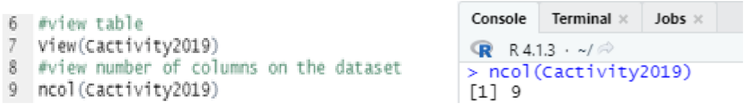
I installed Tyidyverse package on RStudio to import the data and be able to delete columns, rows or any missing value(*Figure 5*).



```
#install tidyverse package to read the csv file
install.packages("tidyverse")
library(readxl)
Cactivity2019 <-read_excel("C:/Users/Izaskun.Sebastian/Downloads/Project/hlth_eh
```

Figure 5

After importing the dataset, I checked the number of columns, so it was easier to call them on the next step to delete them(*Figure 6).*

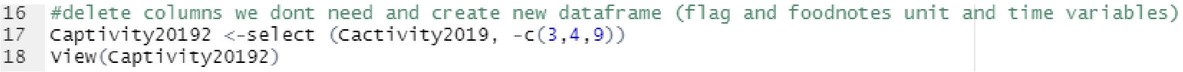


```
#view table
View(Cactivity2019)
#view number of columns on the dataset
ncol(Cactivity2019)
```

```
> ncol(Cactivity2019)
[1] 9
```

Figure 6

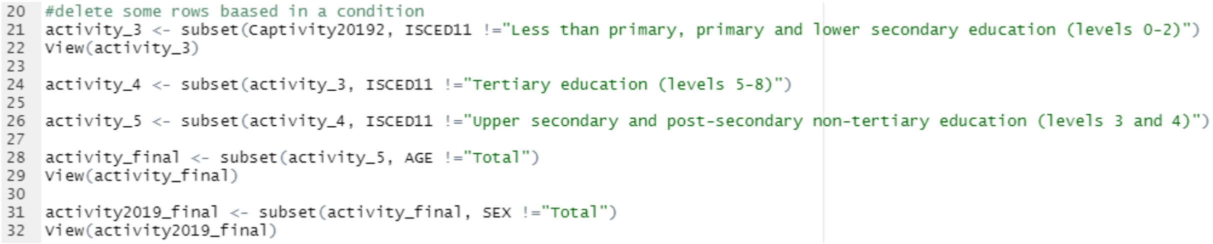
As represented on *Figure 6,* the number of columns was 9, and the variables to delete were “UNIT”, “TIME” and “Flag and footnotes” columns 3,4 and 9. I decided to delete those columns because they did not provide any necessary data for my analysis (*Figure 7*).



```
#delete columns we dont need and create new dataframe (flag and foodnotes unit and time variables)
Captivity20192 <-select (Cactivity2019, -c(3,4,9))
View(Captivity20192)
```

Figure 7

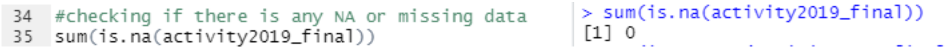
After deleting the variables I deleted rows based in a specific condition, I deleted rows “Less than primary, primary and lower secondary education”, Tertiary education” and “ Upper secondary and post-secondary non-tertiary education” from column ISCED11 to leave only “Total”. And rows containing “Total” from “AGE” and “SEX” columns (*Figure 8).*



```
#delete some rows baased in a condition
activity_3 <- subset(Captivity20192, ISCED11 !="Less than primary, primary and lower secondary education (levels 0-2)")
View(activity_3)

activity_4 <- subset(activity_3, ISCED11 !="Tertiary education (levels 5-8)")

activity_5 <- subset(activity_4, ISCED11 !="Upper secondary and post-secondary non-tertiary education (levels 3 and 4)")

activity_final <- subset(activity_5, AGE !="Total")
View(activity_final)

activity2019_final <- subset(activity_final, SEX !="Total")
View(activity2019_final)
```

Figure 8

With the unnecessary data removed from the dataset I checked if there was any NA or missing value, and I was lucky because they were 0 missing values (see *Figure 9*).

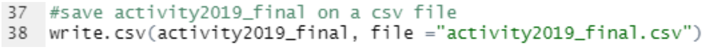


```
#checking if there is any NA or missing data
sum(is.na(activity2019_final))
```

```
> sum(is.na(activity2019_final))
[1] 0
```

Figure 9

The last step on the processing and transformation phase was to save cleaned dataset on an csv file (*Figure 10*)



```
#save activity2019_final on a csv file
write.csv(activity2019_final, file ="activity2019_final.csv")
```

Figure 10

Excel: I used excel to remove the data of the countries I did not need. From world_happiness2019 I removed all non-EU countries as this analysis was focusing on the following EU countries:

Those steps were performed in all 3 datasets mentioned on section *2.0 Data,* and those were the new dataset obtained: activity2019_final.csv, happ_final.csv and deathcause_final.csv.

3.3 Data mining

On the data mining phase is where we proceed to select the technique or algorithm in order to identify insights and patterns in the data. For the data mining, Excel, SPSS and RStudio was used.

Here is where I performed all the statistical tests and created all the visualizations to be able to understand the data easily.

3.4 Interpretation/Evaluation

On this final step is where all the statistics and visualizations created on the previous phase are explained and evaluated.

Once the algorithms have been applied to the data set, we proceed to evaluate the patterns that were generated and the performance that was obtained to verify that it meets the goals set in the first phases.

This stage includes visualization of the extracted patterns, removal of the redundant or irrelevant patterns and the translation of useful patterns into terms that are understandable to the user.

4.0 Analysis

The goal of this project is to analyse how physical activity reflects on a better life or life quality, for the analysis different datasets will be compared as mentioned above.

The main purpose of this analysis is to answer the following questions:

- How many people practices physical activity?
- How physical activity affects them?
- How is the quality of live/happiness of the people who does not perform physical activity?
- Does gender influence on the amount of physical activity performed?

Happiness score, how much physical activity, death, and death causes were looked and analysed to see if there is any relationship between them, all the data used is from year 2019.

1. Data visualisation: It is the graphic representation of information (data), for which different types of visual tools can be used, such as maps, graphs or infographics, which present the information visually, thus making it more accessible and allowing trends, values, patterns to be recognized in the data obtained. Data visualization makes it easier to understand both large and small amounts of data.

   Excel: this spreadsheet was used to retrieve the data, filter non necessary data and to create some pivot tables and graphic for the means of the visual analysis. Data from years nonrelated to 2019, and non-necessary countries for the analysis were filtered.

   RStudio: was used to on the processing and transformation phase, but also to create visualizations (ggplots).

SPSS: this was used to create boxplot graphics an perform tests such us, ANOVA analysis and Tukey test.

Tableau: was used to create mapboxes (*Conclusion section)*, in a map is much easier to visualize using colors based on properties (Happiness score and Activity performed).

2. Statistical Analysis: The analysis of numerical information produces results from data. Interpretation of data through analysis is key to communicating results to stakeholders. Descriptive analysis and interferential was used on this analysis.

   Descriptive statistics: is the part of the discipline that deals with order, summarize, and analyses a set of data through a series of techniques and methods, where the results provided are not intended to go beyond the dataset itself. The bar graph, also called a bar chart, is the most common representation to describe the frequency distribution of a qualitative variable. this resource represents on the abscissa axis (X axis) each one of the categories of the variable and on the ordered (Y axis) the frequencies or percentages of each category, in the form of rectangles with the same base.

   Inferential statistics: allows us to estimate population parameters from the sample used, as well as to test hypotheses. The statistical tests applied will depend on the nature of our data and type of variables. One way ANOVA statistical test was used on this analysis, analysis of variance is a statistical method that allows to discover if the results of a test are significant, that is, they allow to determine if it is necessary to reject the null hypothesis or accept the alternative hypothesis. SPSS was used to perform this test.

5.0 Results

The results are divided into four sections, Visualization of Average physical activity and Happiness score, ANOVA test, Deaths by age and Sport by age.

5.1 Visualization of Average physical activity and Happiness score

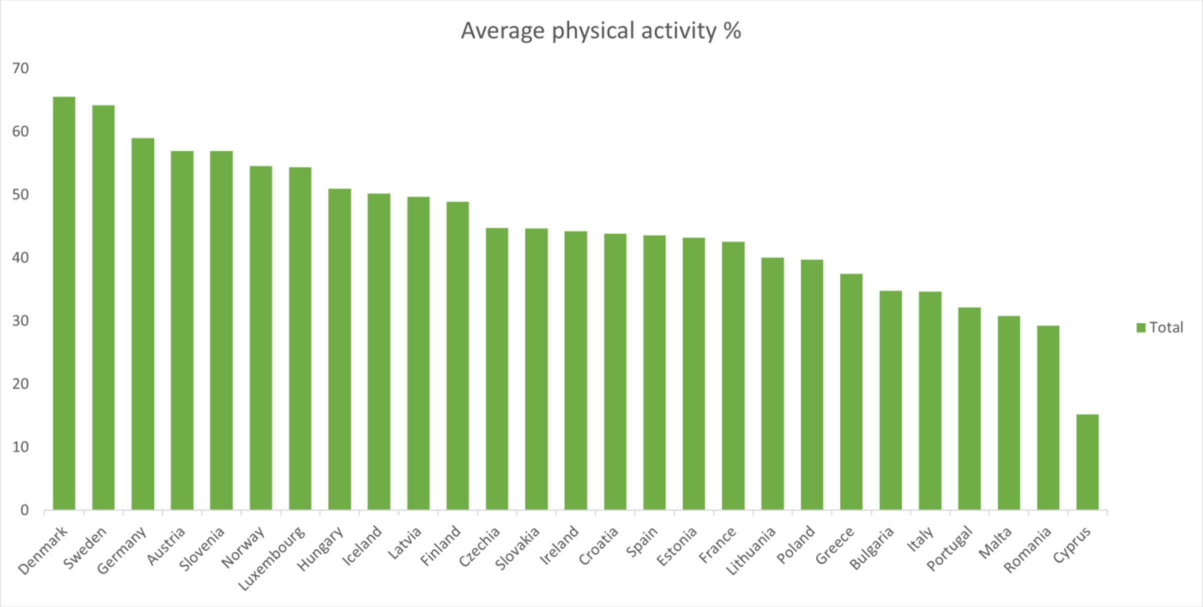


Figure 11- Average physical activity

Figure 11 was created on Excel for initial visualization purposes using activity2019_final.csv dataset, X axis represents countries and Y axis represents average physical activity. This graphic is sorted from high to low, Denmark being de country with highest physical activity (65.51%) and Cyprus being the lowest (15.22%).

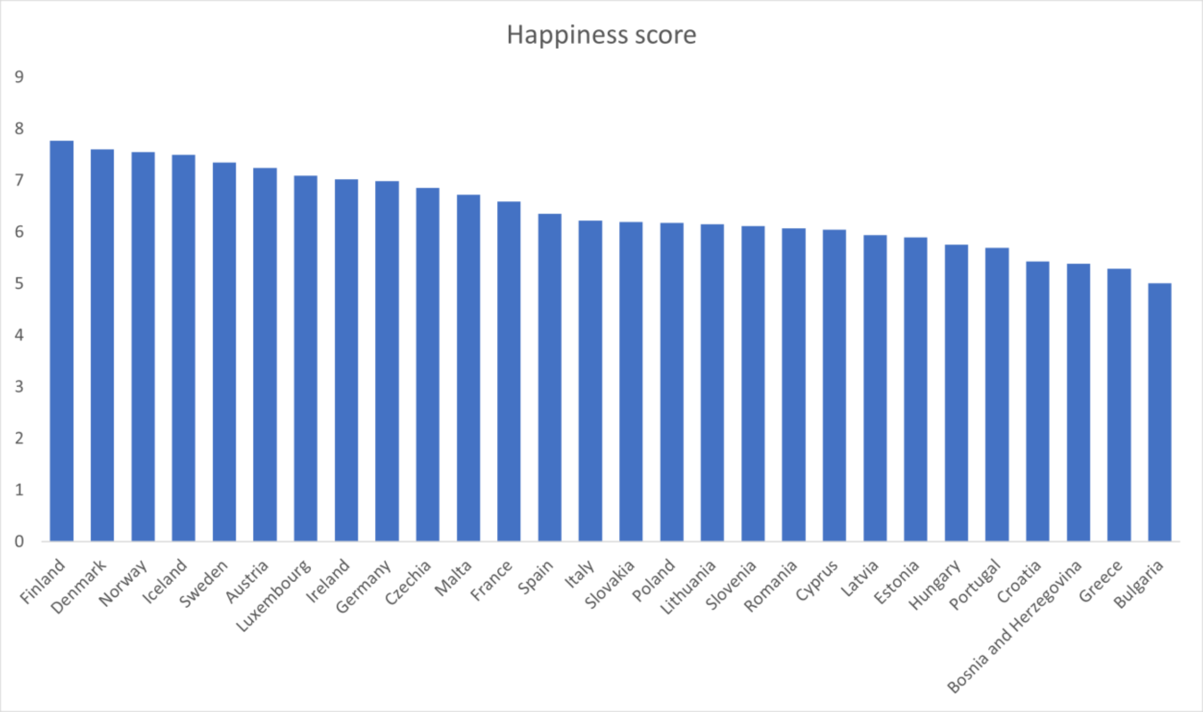


Figure 12- Happiness score 2019

Happiness score graph (*Figure 12*) was also created on Excel using happ_final.csv dataset, X axis represents countries and Y axis represents the happiness score of each country, Finland (7.76 happiness score) and Denmark (7.6) being the happiest countries and Greece (5.28) and Bulgaria (5.01) have the lowest score.

Looking at the two graph above, we can see that there is a significant relation between the happiness score of the country and amount of physical activity, Denmark, Sweden, Germany, Austria, Slovenia and Norway are the most active countries and Denmark, Norway, Sweden and Austria are at the same time between the top 5 of happiest countries.

5.2 ANOVA

One way ANOVA test was done to see if there were differences between the means of physical activity performed and happiness scores on each location – all figures and test performed in this section were done using SPSS.

5.2.2 Physical activity

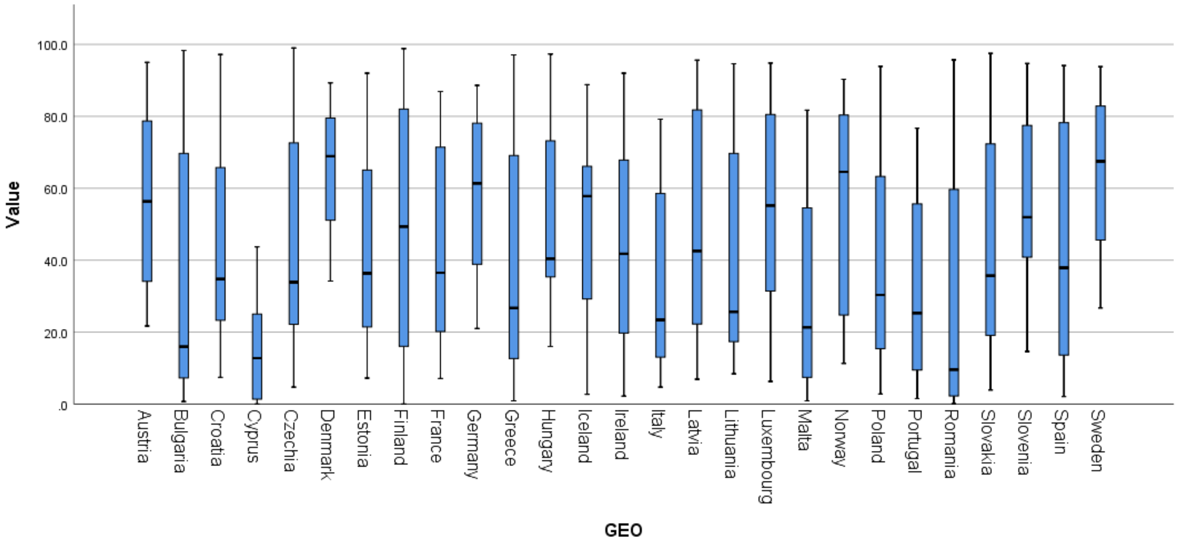


Figure 13- physical activity boxplot

A boxplot chart *(Figure 13)* was created on SPSS for visualization purposes before conducting ANOVA test, here we can see how there is a significant difference between the means of total physical activity of each country.

I was curious of knowing, if there was any difference between genders and the amount of physical activity performed, so I filtered de above boxplot and created *Figure 14.*

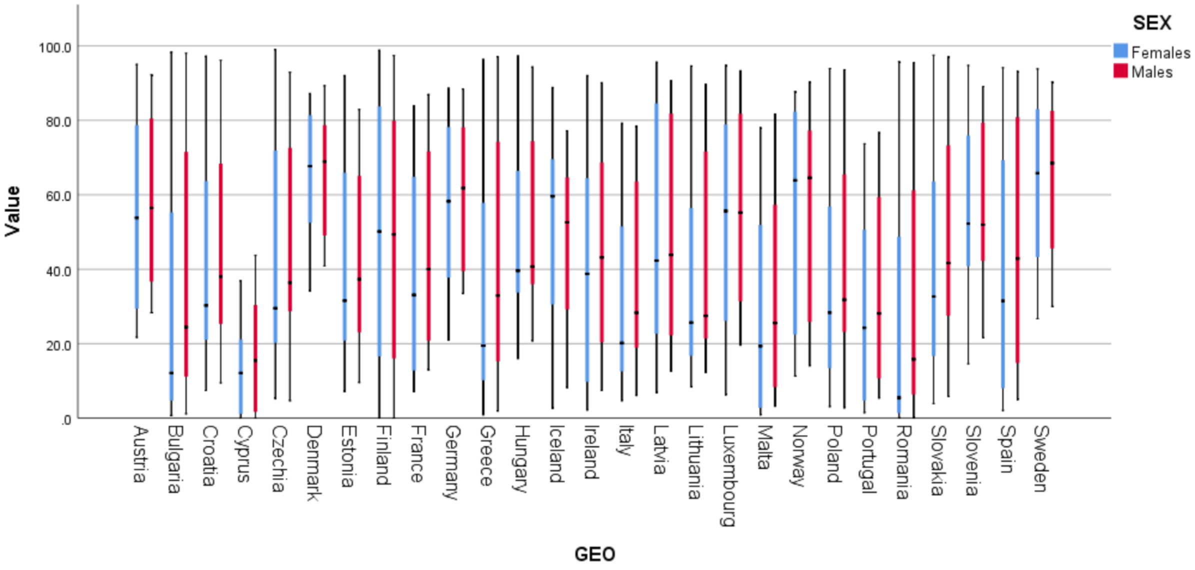


Figure 14

We can observe that there is not significant difference between the amount of physical activity performed by females or males, the distribution does not really change, so the gender or sex does not really influence in the amount of physical activity performed.

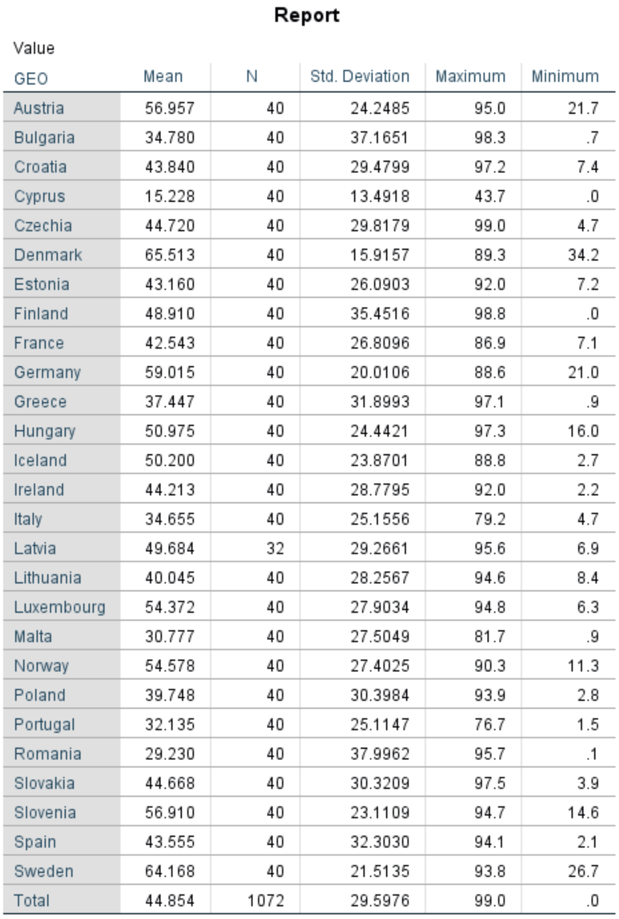
Then one way ANOVA test was conducted using the “Value” column of activity2019_final.csv dataset for each country.

Null Hypothesis (H0): There is no differences in the means of physical activity each country.

Alternate Hypothesis (H1): There is a difference in the means of physical activity each country.

Significance level (α) 0.05

Confidence interval: 95%

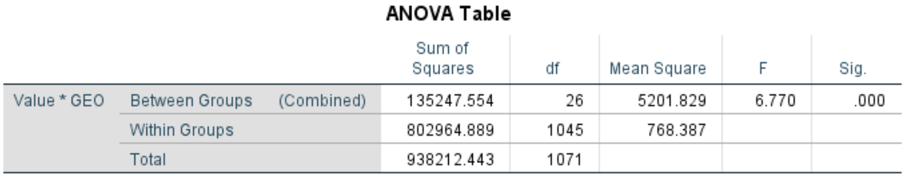


Report

Value

GEO	Mean	N	Std. Deviation	Maximum	Minimum
Austria	56.957	40	24.2485	95.0	21.7
Bulgaria	34.780	40	37.1651	98.3	.7
Croatia	43.840	40	29.4799	97.2	7.4
Cyprus	15.228	40	13.4918	43.7	.0
Czechia	44.720	40	29.8179	99.0	4.7
Denmark	65.513	40	15.9157	89.3	34.2
Estonia	43.160	40	26.0903	92.0	7.2
Finland	48.910	40	35.4516	98.8	.0
France	42.543	40	26.8096	86.9	7.1
Germany	59.015	40	20.0106	88.6	21.0
Greece	37.447	40	31.8993	97.1	.9
Hungary	50.975	40	24.4421	97.3	16.0
Iceland	50.200	40	23.8701	88.8	2.7
Ireland	44.213	40	28.7795	92.0	2.2
Italy	34.655	40	25.1556	79.2	4.7
Latvia	49.684	32	29.2661	95.6	6.9
Lithuania	40.045	40	28.2567	94.6	8.4
Luxembourg	54.372	40	27.9034	94.8	6.3
Malta	30.777	40	27.5049	81.7	.9
Norway	54.578	40	27.4025	90.3	11.3
Poland	39.748	40	30.3984	93.9	2.8
Portugal	32.135	40	25.1147	76.7	1.5
Romania	29.230	40	37.9962	95.7	.1
Slovakia	44.668	40	30.3209	97.5	3.9
Slovenia	56.910	40	23.1109	94.7	14.6
Spain	43.555	40	32.3030	94.1	2.1
Sweden	64.168	40	21.5135	93.8	26.7
Total	44.854	1072	29.5976	99.0	.0

Figure 15



ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
Value * GEO	Between Groups	(Combined)	135247.554	26	5201.829	6.770	.000
	Within Groups		802964.889	1045	768.387		
	Total		938212.443	1071			

Figure 16

On *Figure 16* we can see Degree of freedom is 26 and significant value is 0.000, which is less than 0.05 then we can reject the Null Hypothesis, as there is a difference on physical activity between the means of each country.

5.2.2.1 Physical activity 2

Similar test was performed here, but instead of naming all the countries, I have group them in four groups: Europa meridional, Europa occidental, Europa oriental and Europa septentrional.

Europa meridional: Croatia, Cyprus, Greece, Italy, Malta, Portugal, Slovenia and Spain
Europa occidental: Austria, France, Germany and Luxemburgo
Europa oriental: Bulgaria, Czechia, Hungary, Poland, Romania and Slovakia.
Europa septentrional: Denmark, Estonia, FInland, Iceland, Ireland, Latvia, Lithuania, Norway and Sweden.

I make this change to have less amount of data and to be able to analyze more easily the results.

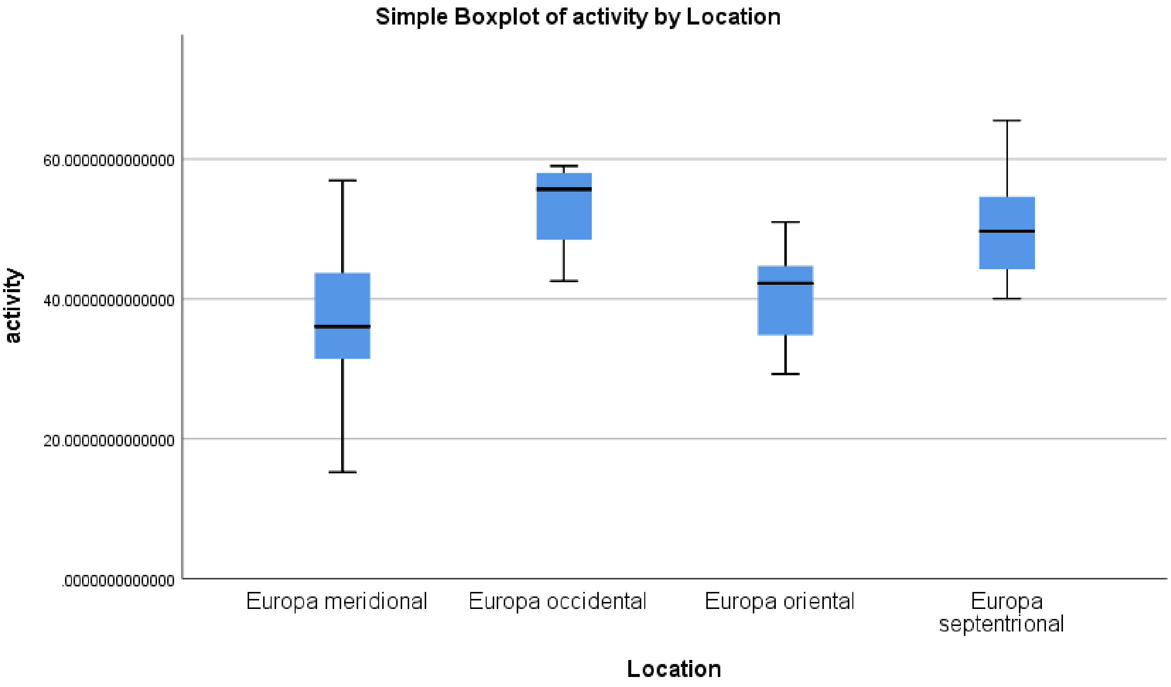


Figure 17

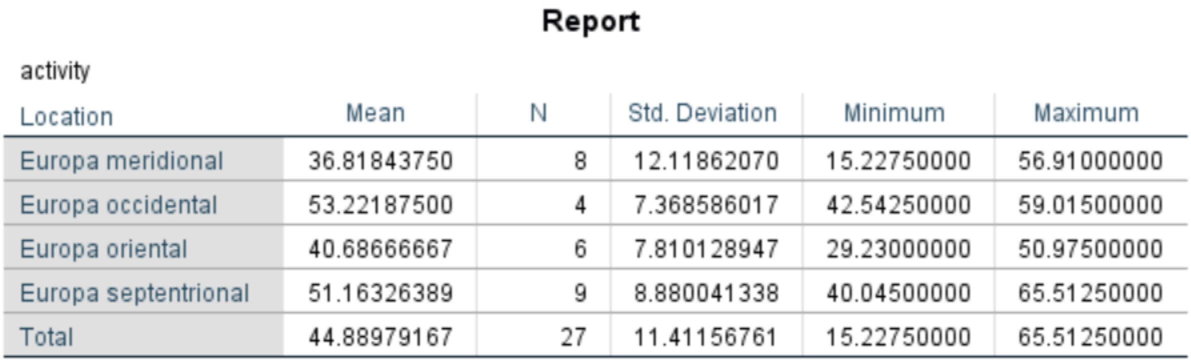
A boxplot chart *(Figure 17)* was created on SPSS for visualization purposes before conducting ANOVA test, here we can see how there is a significant difference between the means of total physical activity of each location.

Null Hypothesis (H0): There is no differences in the means of physical activity each location.

Alternate Hypothesis (H1): There is a difference in the means of physical activity each location.

Significance level (α) 0.05

Confidence interval: 95%

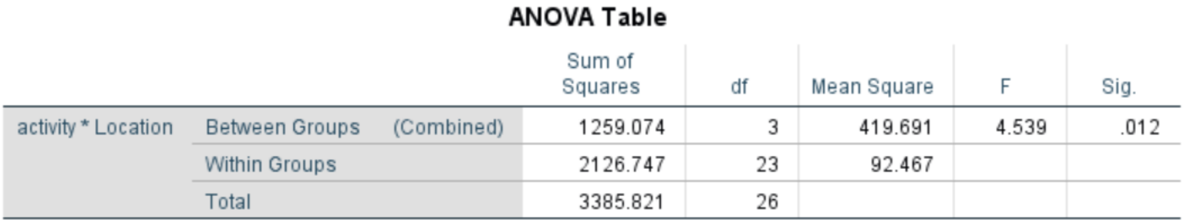


Report

activity

Location	Mean	N	Std. Deviation	Minimum	Maximum
Europa meridional	36.81843750	8	12.11862070	15.22750000	56.91000000
Europa occidental	53.22187500	4	7.368586017	42.54250000	59.01500000
Europa oriental	40.68666667	6	7.810128947	29.23000000	50.97500000
Europa septentrional	51.16326389	9	8.880041338	40.04500000	65.51250000
Total	44.88979167	27	11.41156761	15.22750000	65.51250000

Figure 18

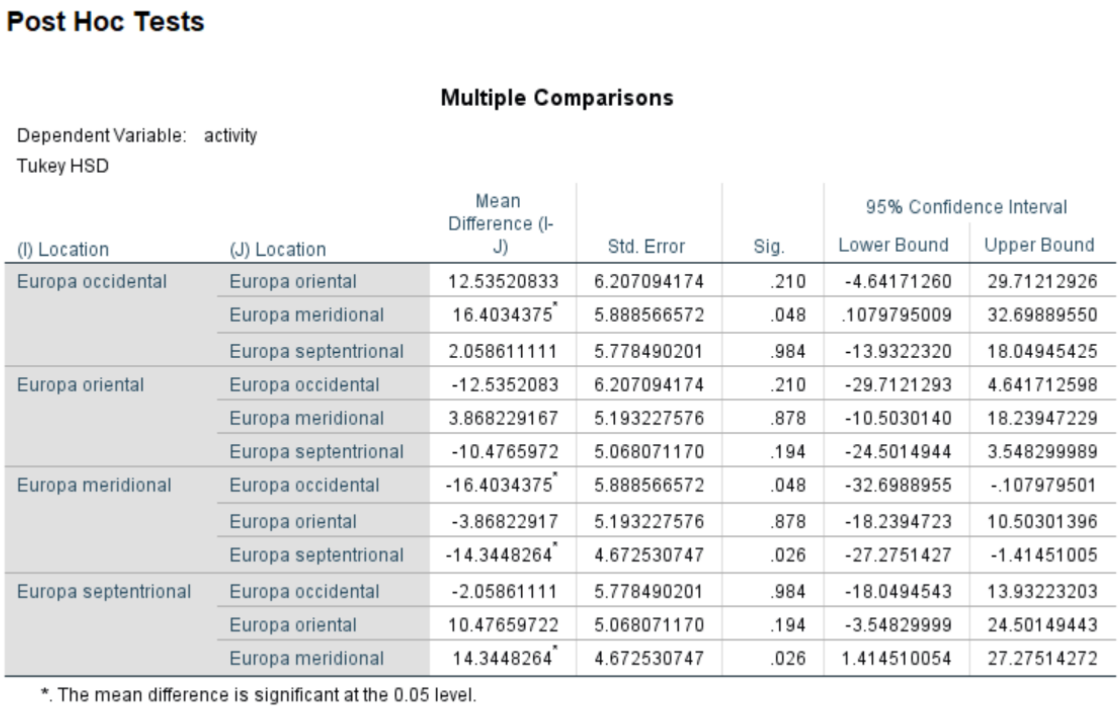


ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
activity * Location	Between Groups	(Combined)	1259.074	3	419.691	4.539	.012
	Within Groups		2126.747	23	92.467		
	Total		3385.821	26			

Figure 19

On *Figure 19* we can see Degree of freedom is 3 and significant value is .012, which is less than 0.05 then we can reject the Null Hypothesis, as there is a difference on the means of physical activity between each location.



Post Hoc Tests

Multiple Comparisons

Dependent Variable: activity
Tukey HSD

(I) Location	(J) Location	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Europa occidental	Europa oriental	12.53520833	6.207094174	.210	-4.64171260	29.71212926
	Europa meridional	16.4034375*	5.888566572	.048	.1079795009	32.69889550
	Europa septentrional	2.058611111	5.778490201	.984	-13.9322320	18.04945425
Europa oriental	Europa occidental	-12.5352083	6.207094174	.210	-29.7121293	4.641712598
	Europa meridional	3.868229167	5.193227576	.878	-10.5030140	18.23947229
	Europa septentrional	-10.4765972	5.068071170	.194	-24.5014944	3.548299989
Europa meridional	Europa occidental	-16.4034375*	5.888566572	.048	-32.6988955	-.107979501
	Europa oriental	-3.86822917	5.193227576	.878	-18.2394723	10.50301396
	Europa septentrional	-14.3448264*	4.672530747	.026	-27.2751427	-1.41451005
Europa septentrional	Europa occidental	-2.05861111	5.778490201	.984	-18.0494543	13.93223203
	Europa oriental	10.47659722	5.068071170	.194	-3.54829999	24.50149443
	Europa meridional	14.3448264*	4.672530747	.026	1.414510054	27.27514272

*. The mean difference is significant at the 0.05 level.

Figure 20

Tukey test on *Figure 20* allows to compare the means of the t levels of a factor after having rejected the Null Hypothesis of equality of means by means of the ANOVA technique. On the table we can see how Europa meridional and Europa septentrional have a significant value of .026, this means that this is where there was a significant difference. While there is not significant difference between Europa occidental and oriental.

5.2.3 Happiness

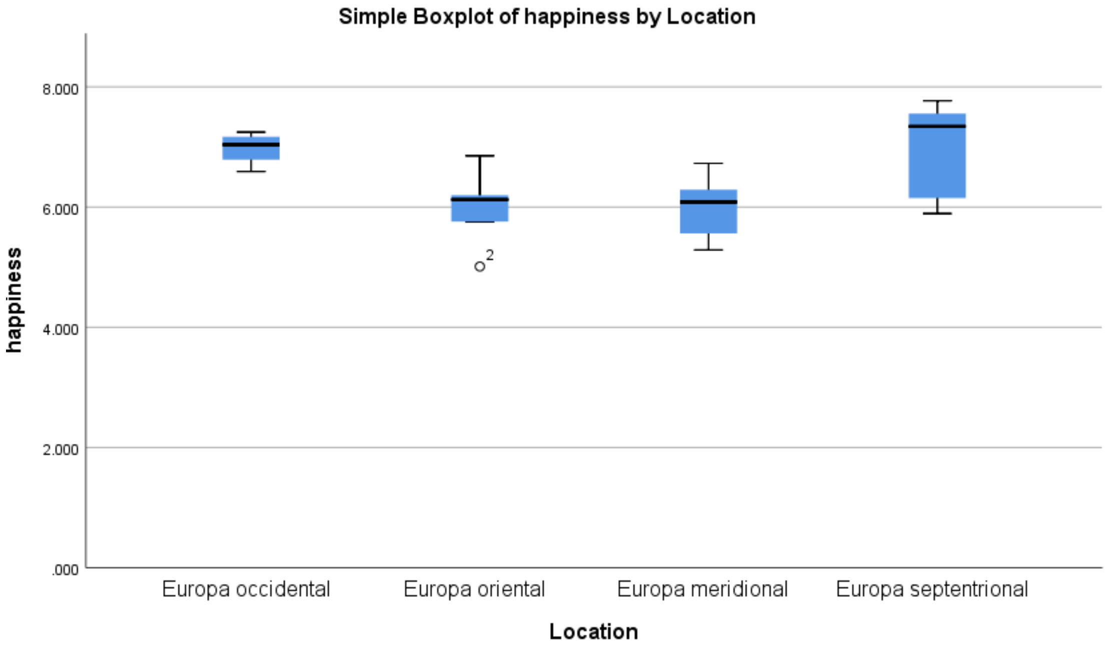


Figure 21

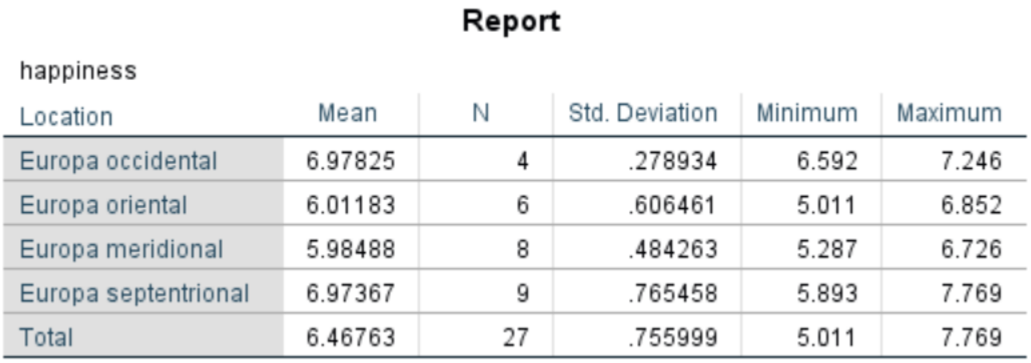
A boxplot chart *(Figure 21)* was created on SPSS for visualization purposes before conducting ANOVA test, here we can see how there is a significant difference between the means of happiness score and location.

Null Hypothesis (H0): There is no differences in the means of happiness each location.

Alternate Hypothesis (H1): There is a difference in the means of happiness each location.

Significance level (α) 0.05

Confidence interval: 95%

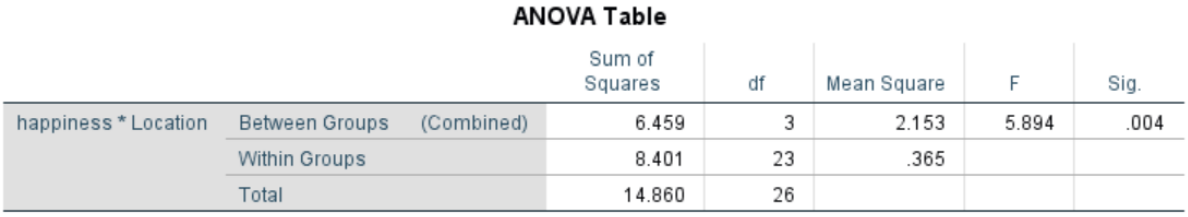


Report

happiness

Location	Mean	N	Std. Deviation	Minimum	Maximum
Europa occidental	6.97825	4	.278934	6.592	7.246
Europa oriental	6.01183	6	.606461	5.011	6.852
Europa meridional	5.98488	8	.484263	5.287	6.726
Europa septentrional	6.97367	9	.765458	5.893	7.769
Total	6.46763	27	.755999	5.011	7.769

Figure 22

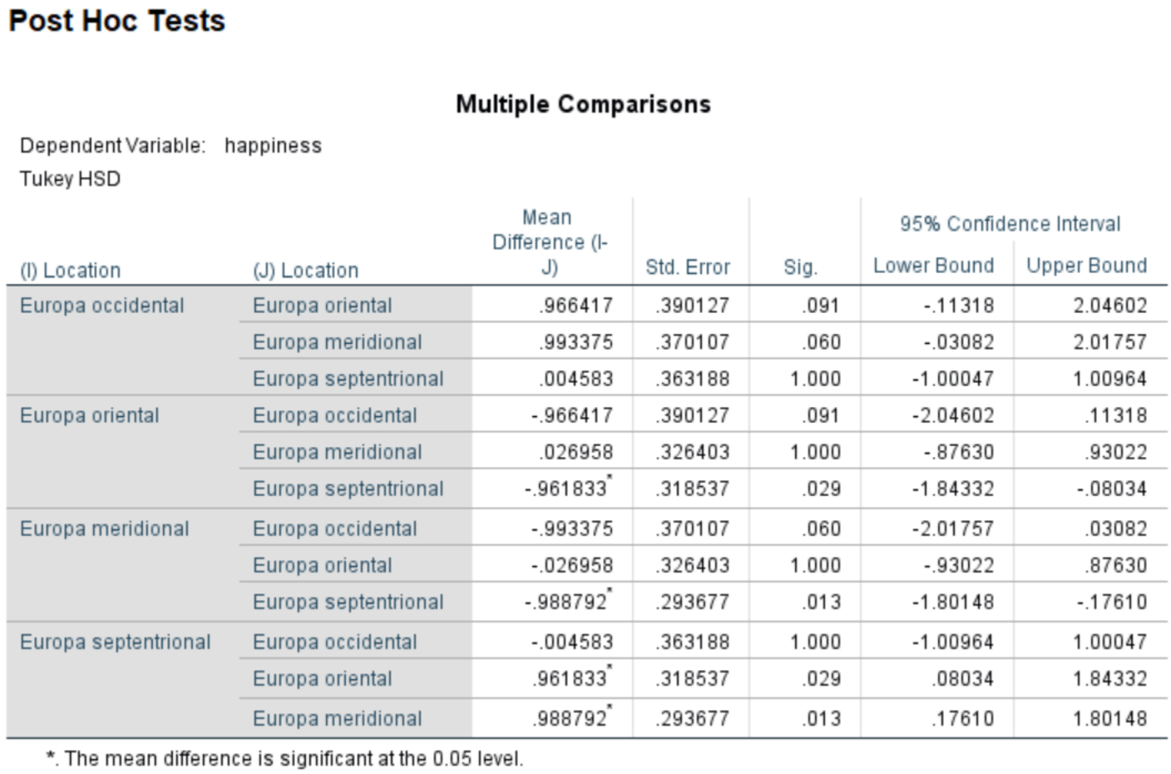


ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
happiness * Location	Between Groups	(Combined)	6.459	3	2.153	5.894	.004
	Within Groups		8.401	23	.365		
	Total		14.860	26			

Figure 23

On *Figure 23* we can see Degree of freedom is 3 and significant value is .004, which is less than 0.05 then we can reject the Null Hypothesis, as there is a difference on the means of happiness between each location.



Post Hoc Tests

Multiple Comparisons

Dependent Variable: happiness

Tukey HSD

(I) Location	(J) Location	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Europa occidental	Europa oriental	.966417	.390127	.091	-.11318	2.04602
	Europa meridional	.993375	.370107	.060	-.03082	2.01757
	Europa septentrional	.004583	.363188	1.000	-1.00047	1.00964
Europa oriental	Europa occidental	-.966417	.390127	.091	-2.04602	.11318
	Europa meridional	.026958	.326403	1.000	-.87630	.93022
	Europa septentrional	-.961833*	.318537	.029	-1.84332	-.08034
Europa meridional	Europa occidental	-.993375	.370107	.060	-2.01757	.03082
	Europa oriental	-.026958	.326403	1.000	-.93022	.87630
	Europa septentrional	-.988792*	.293677	.013	-1.80148	-.17610
Europa septentrional	Europa occidental	-.004583	.363188	1.000	-1.00964	1.00047
	Europa oriental	.961833*	.318537	.029	.08034	1.84332
	Europa meridional	.988792*	.293677	.013	.17610	1.80148

*. The mean difference is significant at the 0.05 level.

Figure 24

Tukey test on *Figure 24* allows to compare the means of the t levels of a factor after having rejected the Null Hypothesis of equality of means by means of the ANOVA technique. On the table we can see how Europa meridional and Europa septentrional have a significant value of .013, this means that this is where there was a significant difference. While there is not significant difference between Europa occidental and oriental.

5.2.4 Deaths

For this section hlth_cd_aro_1_Data.csv dataset was used. This dataset contains deaths related to lack of sport and care, such as diabetes, heart diseases and cerebrovascular diseases.

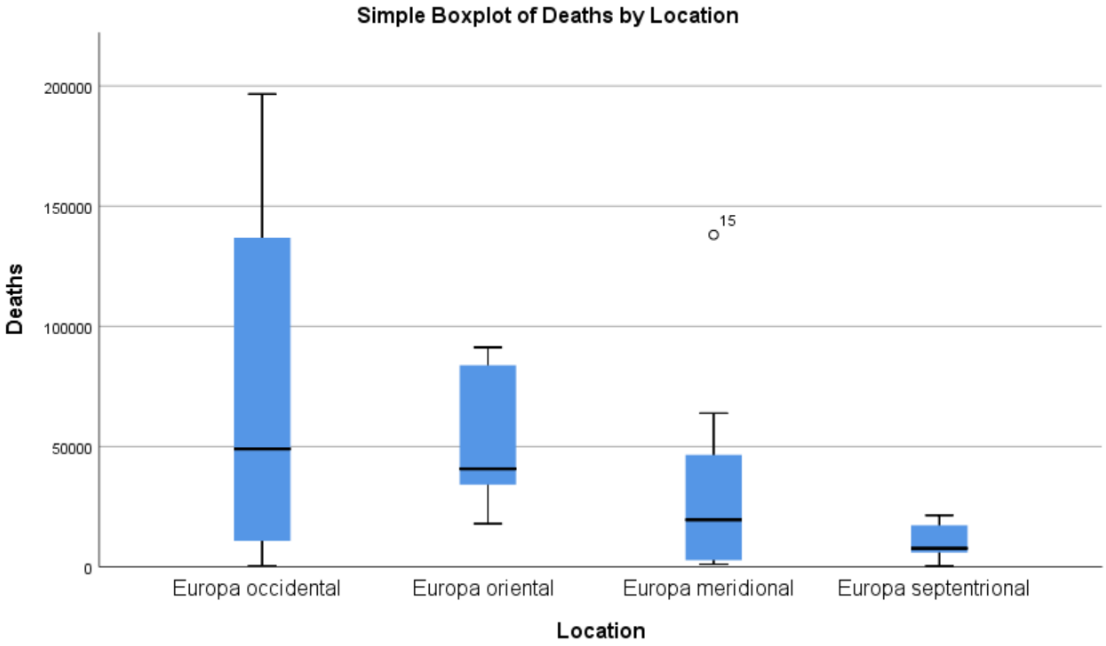


Figure 25

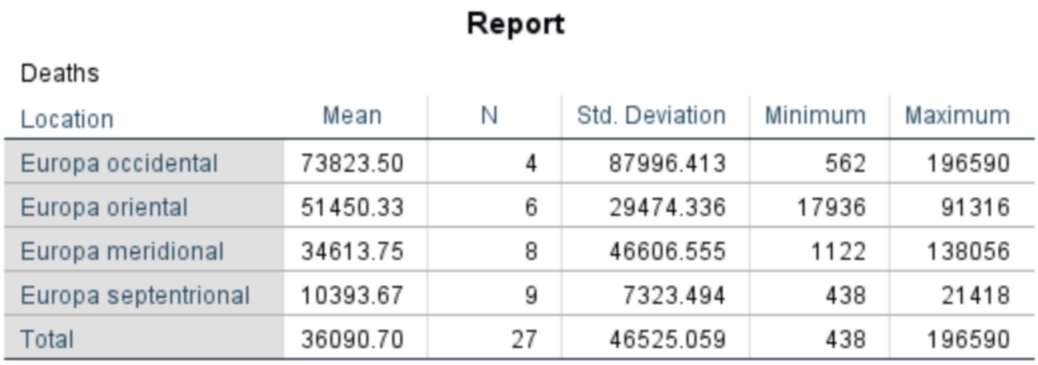
A boxplot chart *(Figure 25)* was created on SPSS for visualization purposes before conducting ANOVA test, here we can see how there is some difference between the means of deaths and location.

Null Hypothesis (H0): There is no differences in the means of deaths on each location.

Alternate Hypothesis (H1): There is a difference in the means of deaths on each location.

Significance level (α) 0.05

Confidence interval: 95%

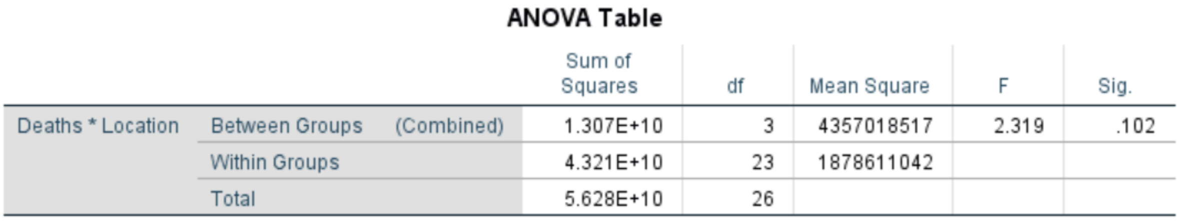


Report

Deaths

Location	Mean	N	Std. Deviation	Minimum	Maximum
Europa occidental	73823.50	4	87996.413	562	196590
Europa oriental	51450.33	6	29474.336	17936	91316
Europa meridional	34613.75	8	46606.555	1122	138056
Europa septentrional	10393.67	9	7323.494	438	21418
Total	36090.70	27	46525.059	438	196590

Figure 26



ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
Deaths * Location	Between Groups	(Combined)	1.307E+10	3	4357018517	2.319	.102
	Within Groups		4.321E+10	23	1878611042		
	Total		5.628E+10	26			

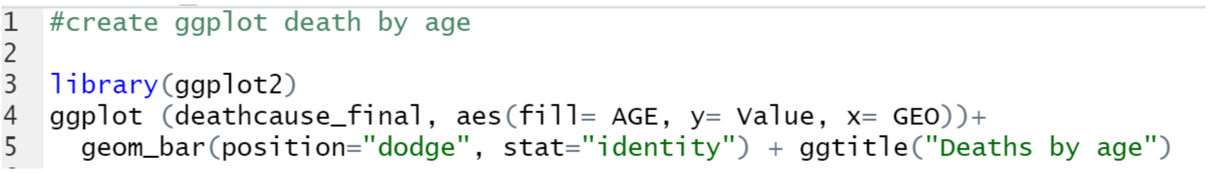
Figure 27

On *Figure 27* we can see Degree of freedom is 3 and significant value is .102, which is more than 0.05 this means that we can retain the Null Hypothesis and reject the alternative hypothesis, in this case is not statistically significant.

In this case Tukey test was not performed as the Null hypothesis was not rejected.

5.3 Deaths by age

RStudio was used to visualize death by age by country *(Figure 28)* using deathcause_final.csv dataset.



```
#create ggplot death by age

library(ggplot2)
ggplot (deathcause_final, aes(fill= AGE, y= Value, x= GEO))+
  geom_bar(position="dodge", stat="identity") + ggtitle("Deaths by age")
```

Figure 28

In red colour we can see deaths of people over 65 years old and in blue “Less than 65 years old” *(Figure 29).*

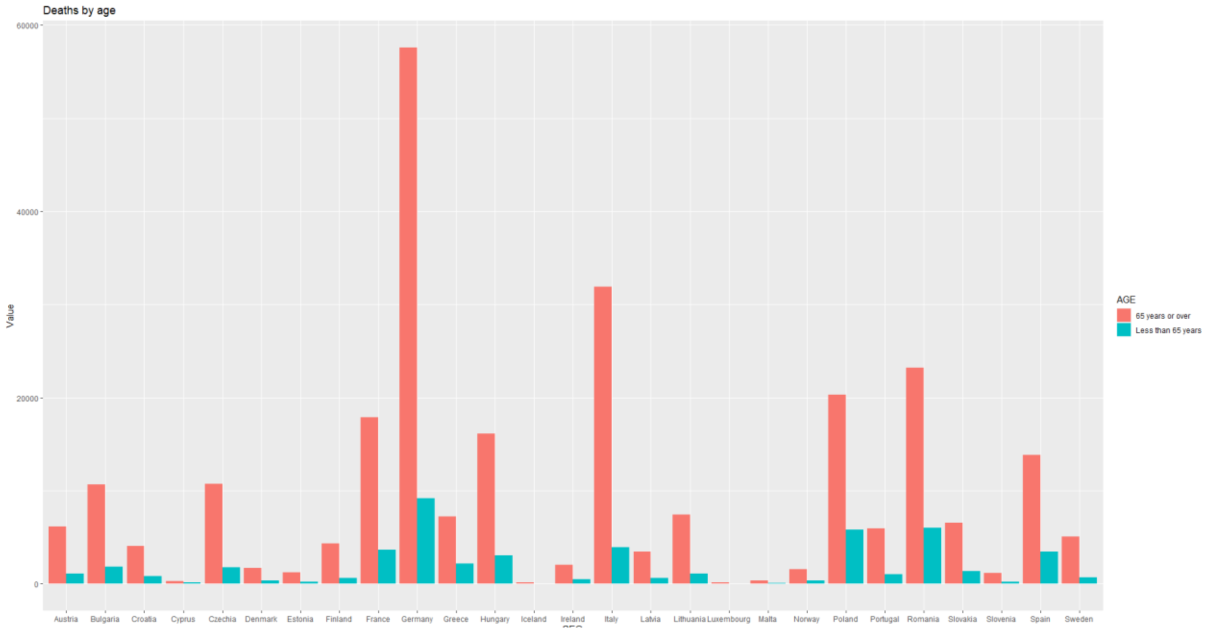
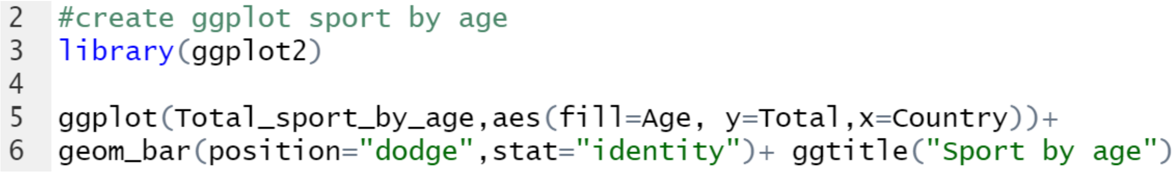


Figure 29

This graphic does not really represent new information, before visualizing this graphic I could assume that the number of deaths of people over 65 years old would be higher than the rest.

5.4 Sport by age

As on point 5.3, RStudio was used to visualize sport by age on each country *(Figure 30).*



```
#create ggplot sport by age
library(ggplot2)

ggplot(Total_sport_by_age,aes(fill=Age, y=Total,x=Country))+
geom_bar(position="dodge",stat="identity")+ ggtitle("Sport by age")
```

Figure 30

In red colour we can see the percentage of sport done by people over 65 years old and in blue “Less than 65 years old” *(Figure 31).*

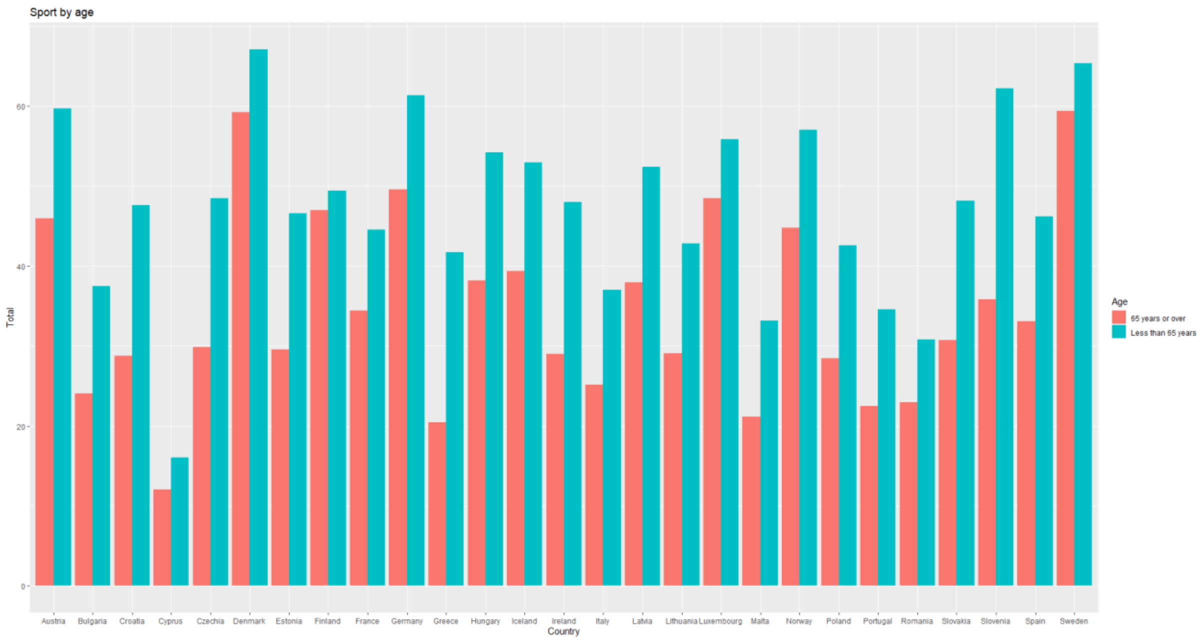


Figure 31

Figure 31 shows that in overall people over 65 years old performs less physical activity than people with less than 65 years old.

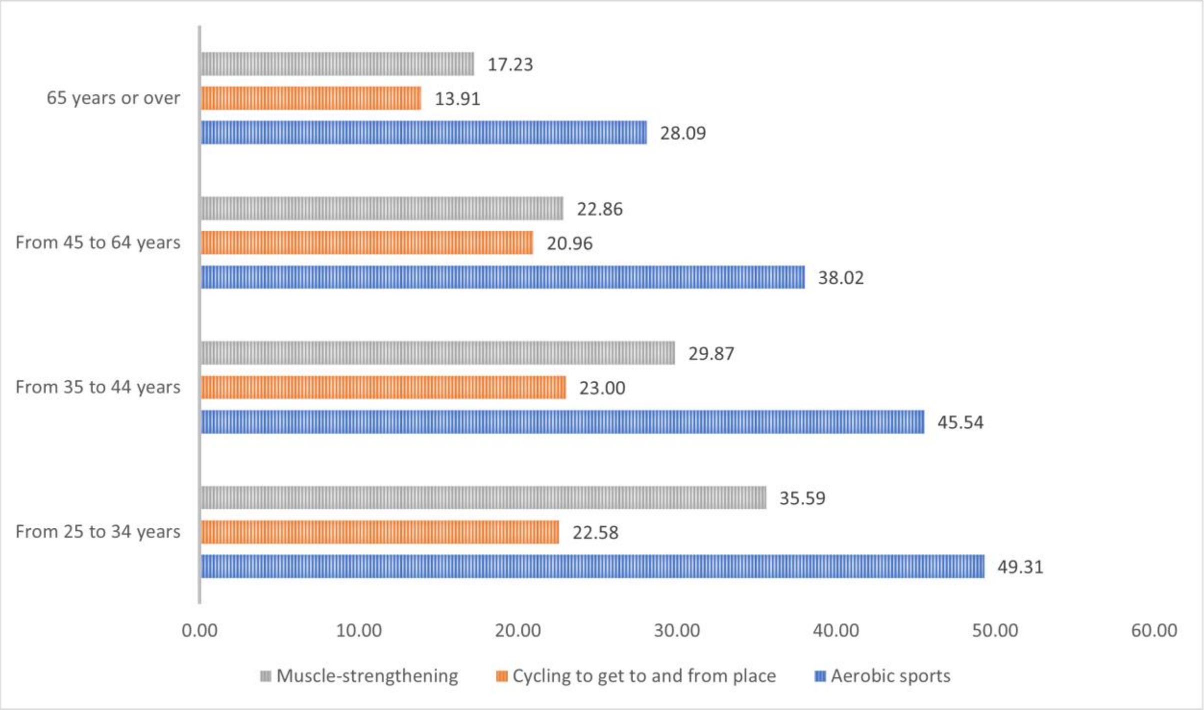


Figure 32

The graphic on *Figure 32* was done using Excel, to view which physical activity was the most performed between each age group.

6.0 Conclusions

In conclusion, taking into account the results and visualizations on the analysis performed above, I can say that there is indeed a relationship between performing physical activity and happiness of the people.

- How many people practices physical activity?
- How physical activity affects on them?
- Does gender interfere on the amount of physical activity performed?
- Is there any relation between the physical activity performed and happiness on each country?
- Which is the most performed physical activity between the population?

In average 44.88% of the people above 25 years old performs physical activity on those countries: Austria, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden.

As shown on *Figure 14* there is not significant difference between the amount of physical activity performed by gender, but we can see that males are above females on each country.

When it comes to physical activity, *on Figure 20*, we can see how there is a significant difference between Meridional Europe and Septentrional Europe countries, Meridional countries (Croatia, Cyprus, Greece, Italy, Malta, Portugal, Slovenia and Spain) being the ones performing less physical activity, 36.81% in average. And Septentrional Europe countries being the most actives (Denmark, Estonia, Finland, Iceland, Ireland, Latvia, Lithuania, Norway and Sweden), 51.16% in average.

When it comes to happiness score there is a similarity with physical activity results. On *Figure 24,* we can observe how there is a significant difference between Meridional Europe and Septentrional Europe countries, Meridional countries (Croatia, Cyprus, Greece, Italy, Malta, Portugal, Slovenia and Spain) being less happy, with a score of 5.9 out of 10 in average. And Septentrional Europe countries being the happiest (Denmark, Estonia, Finland, Iceland, Ireland, Latvia, Lithuania, Norway and Sweden), 7.97 in average.

The physical activity type most performed by the population is Aerobic sports, having the higher rates in all age groups.

To be easier to understand the conclusions, I created two maps using Tableau to visualize in colours the Happiness scores of each country (*Figure 33),* green being the highest scores and red lowest. And quantity of sport performed (*Figure 34*), green being highest amount of sport and red lower.

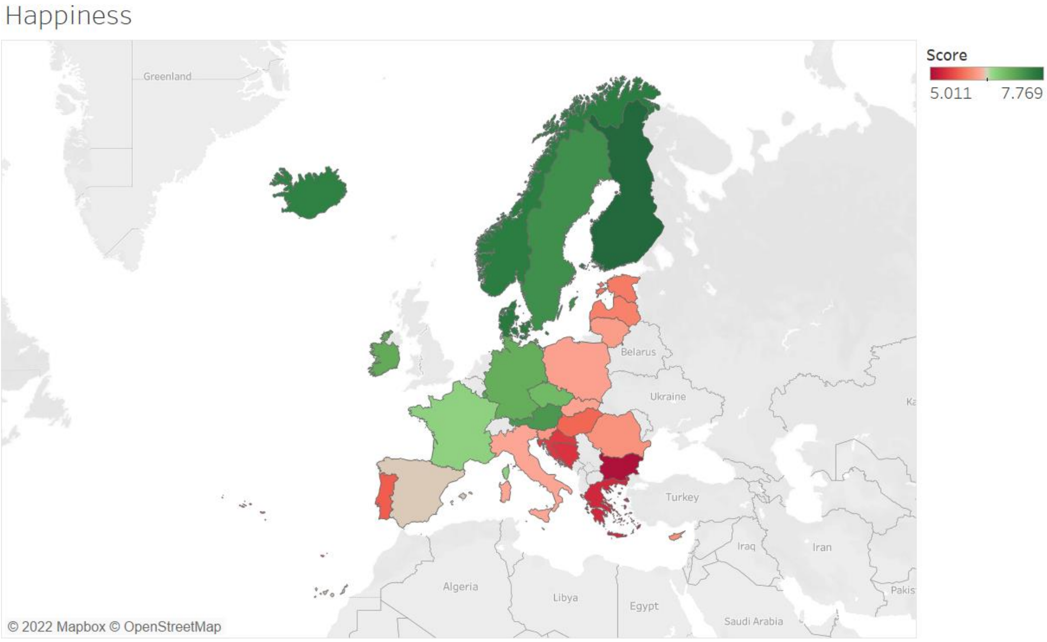


Figure 33

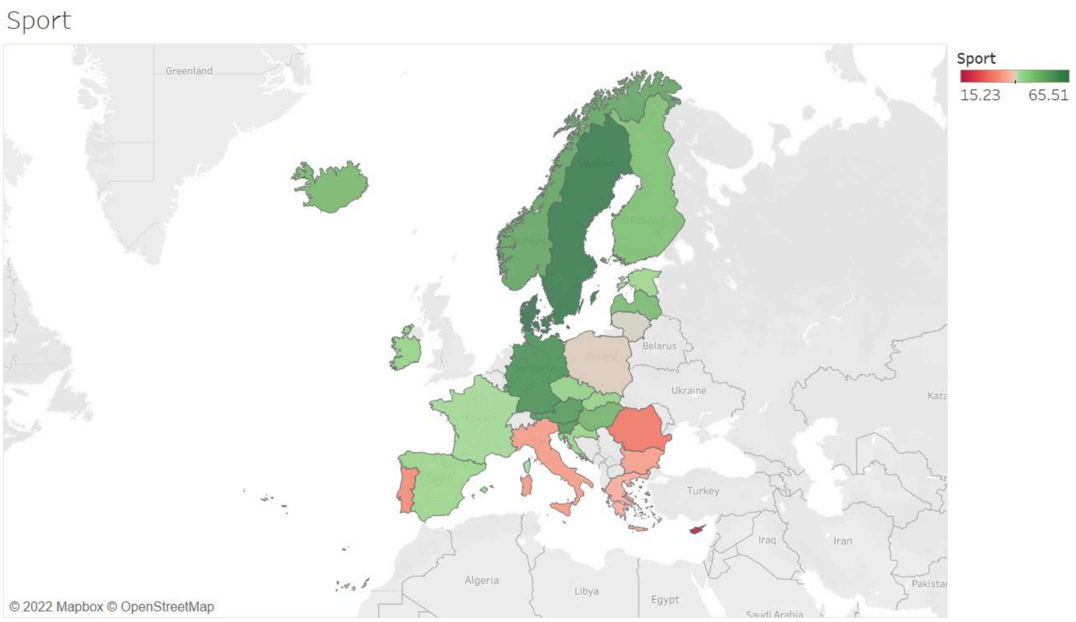


Figure 34

Comparing the colours, we can clearly see how the countries with higher happiness score are the countries performing higher physical activity.

7.0 Further Development and Research

The research present on this technical report is done in a small scale. I have achieved the aims mentioned on *section 1.2*, but as I have had to combine with the other jobs and modules, I have not been able to invest as much time as I would have liked. During the duration of this academic year, programming languages and software's where showed and explained by the lectures, but some of them where hard to implement because the project was already ongoing.

In terms of future development or research of this project, there are number of things that I would like to properly interpret and analyze.

- In Germany there is a high amount death of people of 65 years old and above related to diabetes, heart diseases and cerebrovascular diseases (*Figure 29), despite* of having one of the highest physical activity scores *(Figure 30)* and I would like to analyze the reason why. It might be related to alcoholism and high intake of ultra-processed food.

- In Cyprus the amount of physical activity performed is very low comparing with the rest of the countries, this is also something that I would like to investigate.

- I would also like to do this analysis but taking into account the 195 countries in the world, to deeply analyze and also to understand if there is a significant difference in happiness and physical activity performed in developed and undeveloped countries or countries of the northern and southern hemisphere.

8.0 References

Centers for Disease Control and Prevention (2021). Benefits of physical activity. [online] CDC. Available at: https://www.cdc.gov/physicalactivity/basics/pa-health/index.htm. [Accessed 09 March 2022]

Pereira, S.R.T., Arteaga, I.H., Zambrano, S.J.C., Troya, A.H. and Pérez, J.C.A. (2016). Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional. [online] ediciones.ucc.edu.co. Ediciones Universidad Cooperativa de Colombia. Available at:< https://ediciones.ucc.edu.co/index.php/ucc/catalog/book/36> [Accessed 15 March 2022].

www.caeme.org.ar. (2021). ¿Cuáles son los beneficios de la actividad física? CAEME. [online] Available at: <https://www.caeme.org.ar/los-beneficios-de-hacer-actividad-fisica/ >[Accessed 15 March 2022].

www.kaggle.com. (n.d.). World Happiness Report up to 2022. [online] Available at: https://www.kaggle.com/datasets/mathurinache/world-happiness-report?select=2020.csv [Accessed 27 March 2022].

Europa.eu. (2011). Home - Eurostat. [online] Available at<: https://ec.europa.eu/eurostat/> [Accessed 5 April 2022]..

www.microsoft.com. (n.d.). Microsoft Excel, Spreadsheet Software, Excel Free Trial. [online] Available at:< https://www.microsoft.com/en-ie/microsoft-365/excel.> [Accessed 9 April 2022].

www.questionpro.com. (n.d.). Qué es SPSS y cómo utilizarlo. [online] Available at: <https://www.questionpro.com/es/que-es-spss.html#:~:text=SPSS%20es%20un%20software%20popular >[Accessed 12 April 2022].

R Core Team (2019). R: The R Project for Statistical Computing. [online] R-project.org. Available at: <https://www.r-project.org/>[Accessed 29 April 2022].

Tableau. (n.d.). Free Training Videos - 2022.1. [online] Available at: <https://www.tableau.com/learn/training/> [Accessed 12 April 2022].

IBM (2019). SPSS software. [online] Ibm.com. Available at: <https://www.ibm.com/analytics/spss-statistics-software> [Accessed 8 May 2022].

Pereira, S.R.T., Arteaga, I.H., Zambrano, S.J.C., Troya, A.H. and Pérez, J.C.A. (2016). Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional. [online] ediciones.ucc.edu.co. Ediciones Universidad Cooperativa de Colombia. Available at:< https://ediciones.ucc.edu.co/index.php/ucc/catalog/book/36> [Accessed 15 March 2022].

GeeksforGeeks. (2018). KDD Process in Data Mining - GeeksforGeeks. [online] Available at: <https://www.geeksforgeeks.org/kdd-process-in-data-mining/>[Accessed 8 May 2022]

9.0 Appendices

National College of Ireland

Project Proposal

Sport and quality of life in Ireland

06/11/2021

BSHCEDA4

Data Analysis

Academic Year 2021/2022

Izaskun Lekue

X17105595

X17105595@student.ncirl.ie

Contents

9.1. Project Proposal

9.1.1 Objectives

The objective of this project is to analyse the link between sport and life quality in Ireland.

Although this is a topic already discussed by different people, now after the pandemic, it is very interesting to do this type of analysis and bring to light the important link between sports, an active lifestyle and quality of life.

One of the main objectives, is to provide an insightful report based on the analysis as well as some interactive data visualizations to allow the reader to interact with the data

9.1.2 Background

One of the main reasons I chose this topic is because I wanted to talk about something that really interests me, when you work on something you like, it is easier to do thing better.

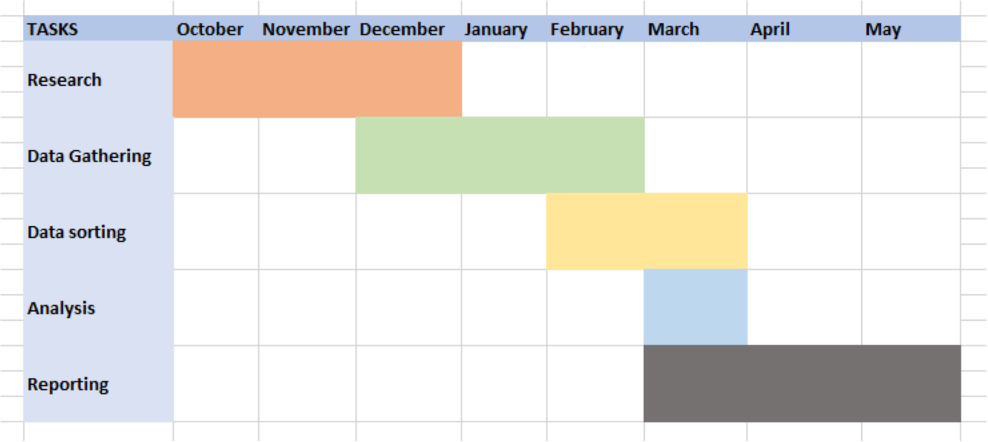
I am into sports since few years ago and I have been able to see for myself the effect of sport on quality of life, but now, I want to show it with numbers and data.

9.1.3 Technical Details

At the moment, I am waiting to know if my project has been accepted or not, once I get the answer I will start my research in relevant articles, journals, data sources… and I will start gathering the relevant data.

I am not sure, yet which technologies will be using for this project, I need to have a chat with my supervisor to clear my mind and to know if my idea is doable.

9.1.4 Project Plan



TASKS	October	November	December	January	February	March	April	May
Research	🟧	🟧	🟧					
Data Gathering			🟩	🟩	🟩			
Data sorting					🟨	🟨		
Analysis						🟦		
Reporting						⬛	⬛	⬛

This is my project plan at the moment, my project has not been yet accepted so I’m not sure if Ill be able to proceed and to follow this exact project plan.

9.2. Reflective Journals

9.3.1 October

This month we had our first two Software Project classes with Enda, he went through the introduction and explained some aspects about the project, the project pitch and reflective journal.

To be honest I have been very lost and it has been pretty hard for me to choose a project topic. As we are unable to attend class, they have had to adapt the educational program to the current situation, and it is not possible to take written exams. The alternative to written exams are projects, which makes this year's workload higher than before.

My idea is to analyse Sport and Quality of life in Ireland. I had submitted the project pitch with my idea, not on time because I was sick, so I uploaded it to Moodle a few days later.

I do not know yet if my project idea has been accepted, a supervisor has been assigned to me a few days ago (Divyaa Manimaran Elango) but I have not heard from her yet, once I get any feedback, I would like to meet her so I can clarify some ideas and I can start working on the project itself and Project Proposal which is due next week (7th of November).

9.3.2 November

November have been a tough month for me, due to personal circumstances at work and family.

Despite that I have been doing research about the project topic. I have selected some dataset with information related to deaths related to sedentary life in Ireland, but at the moment I have not been able to find any appropriate dataset with numbers related to sport, happiness or quality of life.

I also thought about doing a survey in order to obtain my data, but im afraid of not getting a significant amount of data or loosing many important time that I could use on the project.

So after all, I am considering the option of focusing in different countries instead of Ireland. I found interesting data about Mexico in https://en.www.inegi.org.mx/ (National Institute of Statistics and Geography), but I think I prefer to analyse different countries instead of focusing just in one.

The next week is reading week so I will try to find more suitable data for Spain or any other country and after that I will keep focusing on the two Cas that I have to submit (Web services and API and Data application development).

9.3.3 December

To be honest I have been very lost and it has been pretty hard for me to choose a project topic. As we are unable to attend class, they have had to adapt the educational program to the current situation, and it is not possible to take written exams. The alternative to written exams is projects, which makes this year's workload higher than before.

For the reasons mentioned above, I did not make loads of progress on my Final Project, Im still waiting for some feedback from my supervisor (Divyaa Manimaran Elango) I have not heard from her yet, once I get any feedback, I would like to meet her so I can clarify some ideas and I can start properly working on the project.

I want to ask her few things, as I am not sure on how to use some aspect of R and also want to make sure the data I'm using is a good source and useful.

9.3.4 January

As part of Final year project submission, on the 22nd of December I submitted Mid Point submission, I did not have relevant progress on the project due to the first semester CA's and TABA's.

As mentioned on November Journal, it has been hard to find relevant data related to my topic and related to Ireland, so I decided to focus on 27 European countries instead.

I also decided which software's I will use which are RStudio and Excel at the moment, and once I learn more about SPSS I will also implement it.

For what I have learned on the first semester I don't feel very confident when it comes to interpretation of the data, as the lecturers where more focus on the analysis than on the interpretation, which I think is very important in order to properly explain our results in the final project.

9.3.5 February

This month I was able to focus more on the final project and I found more datasets that I can implement on my analysis. I cleaned one of the dataset I acquired before so the data is ready to use once the rest of the datasets have been cleaned.

I am applying what I have learned in Data mining and Advanced Data Analysis modules to clean the data on RStudio, this can also can be done manually in EXCEL which is easier for me, but I like to challenge myself and use the code learned on the course.

Even it was easier for me to perform those cleaning steps in excel, when using the correct code RStudio is much faster and less manual to use.

9.3.6 March

This month I tried to implement what I have learned viewing youtube videos, which was to do an analysis on jupyter using python, but I got some errors and after trying few times I decided not to go forward.

I decided the structure of the analysis and performed some ANOVA tests on SPSS and RStudio. At the beginning I learned to do it on RStudio, but then, I realized that those analytical test are much easier to perform in SPSS.

I also will use Tableau, which is a great tool to visualize the data obtained, and my idea is to add on the conclusion of the project images with the data visualizations created.

This month I tried to implement what I have learned viewing youtube videos, which was to do an analysis on jupyter using python, but I got some errors and after trying few times I decided not to go forward.

I decided the structure of the analysis and performed some ANOVA tests on SPSS and RStudio. At the beginning I learned to do it on RStudio, but then, I realized that those analytical test are much easier to perform in SPSS.

I also will use Tableau, which is a great tool to visualize the data obtained, and my idea is to add on the conclusion of the project images with the data visualizations created.

9.3.7 April

This month I haven’t done much of the final project because I had other different Continuous assessments and TABA’s, the final project is one of the most important thing on the final year, but at the same time I have to focus on the different modules in order to pass.

Even I have been focused on other modules, I have learned and gained confident in different data mining techniques while doing other projects and that has been very beneficial in terms of the final project.

My idea is to finish the final analysis and report by the 1st of May, so I don’t have to rush or be worried on the last two weeks before the submission date.