Exploratory data analysis of estimated traffic fatality rate on WHO dataset

Data Science for Health Systems thesis

Simone Maiorani
Department of Engineering
University of Perugia
Perugia, Italy
simone.maiorani@studenti.unipg.it

***Abstract*—This paper presents the results of an exploratory analysis conducted on a dataset provided by the World Health Organisation, focusing on estimated road traffic fatality rates. The analysis, implemented using the R programming language, aims to examine trends over the period from 2000 to 2019 and to identify differences in road accident rates by gender and geographical area. Through the use of graphs and statistical tests, significant differences emerged both between the sexes and in the different geographical areas, highlighting an uneven distribution of the data globally.**

***Index Terms*—WHO, EDA, Traffic Death**

I. Introduction

Road traffic injuries (RTIs) are the leading cause of unintentional injuries, accounting for the greatest proportion of deaths from unintentional injuries [1]. They are the leading cause of injury-related disability-adjusted life years, and they pose a significant economic and societal burden. Despite this burden, RTIs remain a largely neglected public health problem, especially in low- and middle-income countries, where urbanization and motorization are rapidly increasing [2]. The purpose of this study is to examine the trend in the estimated road fatality rate (per 100 000 population) for the period from 2000 to 2019, focusing on the analysis of differences between geographical areas and genders.

II. Dataset

The dataset comes from the World Health Organisation (WHO) web portal and focuses on the estimated rate of fatal road deaths, expressed as the number of deaths per 100.000 inhabitants [3]. The estimation methodology involves four groups of countries. The first group includes countries with a death registration data completeness of at least 80 per cent, using death registrations, projections and reported data. The second group includes countries such as India, Iran, Thailand and Vietnam, with alternative sources for cause of death data, treated with regression methods. The third group covers countries with populations below 150.000, using death data reported in surveys. Finally, the fourth group, without suitable registration data, is treated with a negative binomial regression model.

III. Dataset Modelling

In the first stage, the dataset is cleaned by preprocessing the data. Features containing only NA values and all those that do not provide meaningful information, such as indicators, data types, location codes, etc., were removed. Samples with the value of *"Both sexes"* in the Sex column were not considered because they represent the average of the values of the other male and female samples. They are therefore not considered useful for the purposes of this analysis, which focuses on differences between sexes and geographic areas, so they have been eliminated. Next, the features used were renamed to make the dataset more understandable and the information they contain more accessible. Finally, the features were checked for missing or NaN values.
The final dataset consists of the following features:

- **GeoArea**: represents the geographical area according to the WHO division.
- **State**: nation in the world to which the data refers.
- **Year**: year in which the measurement was carried out.
- **Sex**: can take on two values: Male and Female.
- **Value**: rate of road deaths recorded for a specific gender, year and country.
- **LowerConfidenceInterval**: lower limit of the confidence interval.
- **UpperConfidenceInterval**: upper limit of the confidence interval.

IV. Exploratory Analysis

The EDA phase was essential to identify trends and relationships in the data. The use of graphs made possible an initial understanding of the dataset.

A. Distribution of values

The histogram shown in Figure 1 illustrates the distribution of values of all samples within the dataset without applying any filter in order to visualize the general trend of the data distribution.

It can be seen from the graph that most samples take values between 0 and 25, while values higher than this range are observed less frequently.

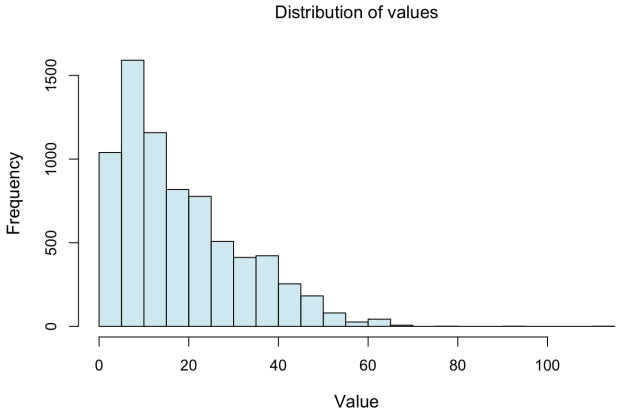

Fig. 1. Frequency of values assumed

B. Gender difference

The primary objective of this phase was to examine the congruence or divergence in the distributions of traffic accidents between the two sexes. Upon closer examination of gender disparities, a clear distinction emerges, highlighting discernible variations in the distribution of values between males and females. In order to comprehensively assess the nuances of these distinctions, different types of graphs were strategically employed. The use of different graphical representations aims to capture as much information as possible about the nature of the data used. The analysis showed that fatal traffic accidents have a higher incidence among males than among females. This disparity in the frequency of fatal accidents raises intriguing questions, setting the stage for further exploration of the underlying factors that contribute to these gender-specific trends. The graph in Figure 2 highlights this aspect.

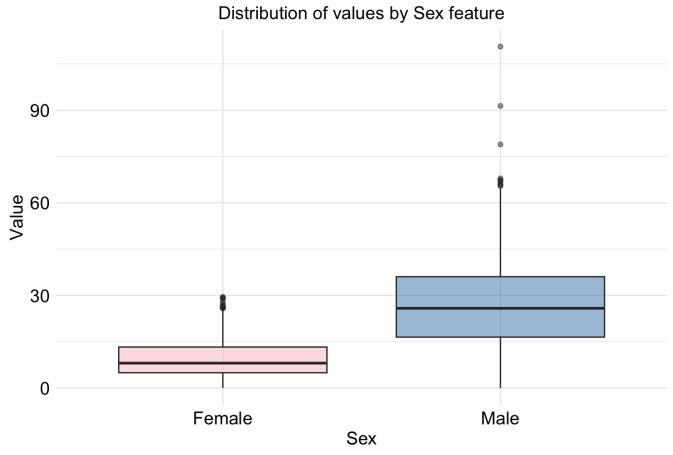

Fig. 2. Males and females compared

The graph presented in Figure 3 illustrates the distribution of values in the different samples. In particular, a discernible pattern can be seen in which high values are more prevalent among the male population, while lower values have a higher frequency in the female population. Discovering the reasons for this observed disparity could be an interesting avenue for future investigations, shedding light on potential factors influencing these distinct trends.

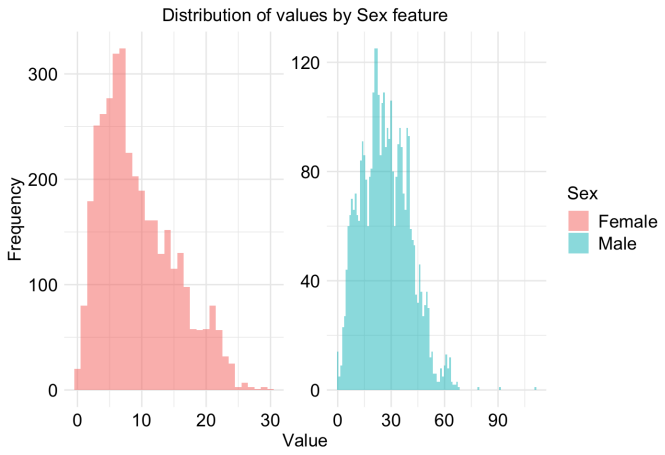

Fig. 3. Difference in the frequency of values between genders

C. Trend over the years

The exploratory analysis continued by focusing on observing the differences in the values taken over the years considered by the dataset (between 2000 and 2019). From the graph shown in Figure 4, it can be seen that the values take a similar trend until 2016 and then undergo a significant increase in the remaining years. There are no other variables within the dataset to understand the causes of this phenomenon.

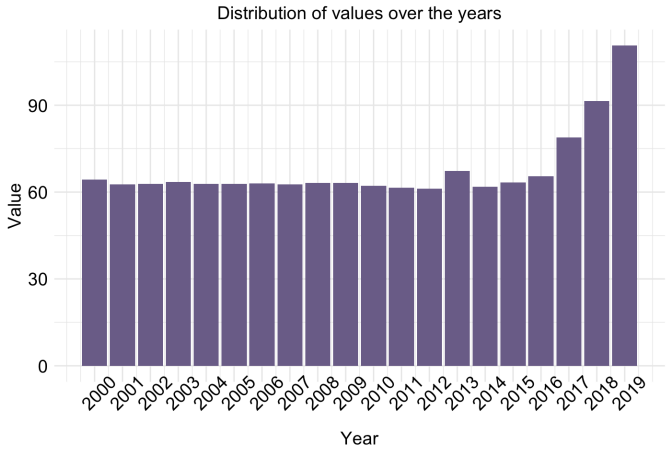

Fig. 4. Values assumed over the years

Again, an attempt was made to mark, if any, the difference in the values taken over the years by the male and female genders. Again, the answer is affirmative, as shown in the box plot in Figure 5. The graph illustrates that fatal traffic accidents are always higher in males than in females in every year considered.

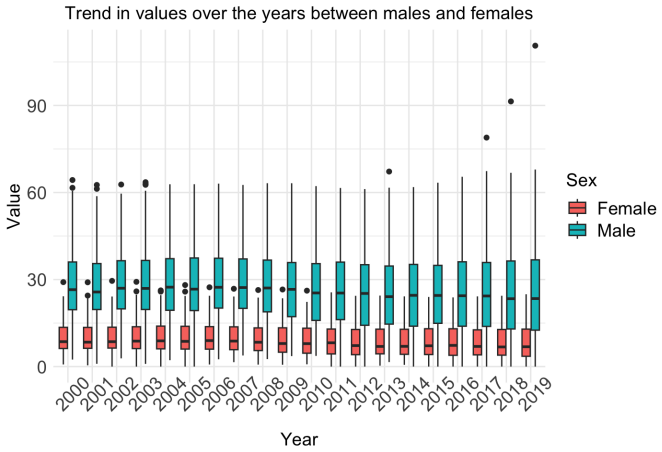

Fig. 5. Trends of males and females over the years

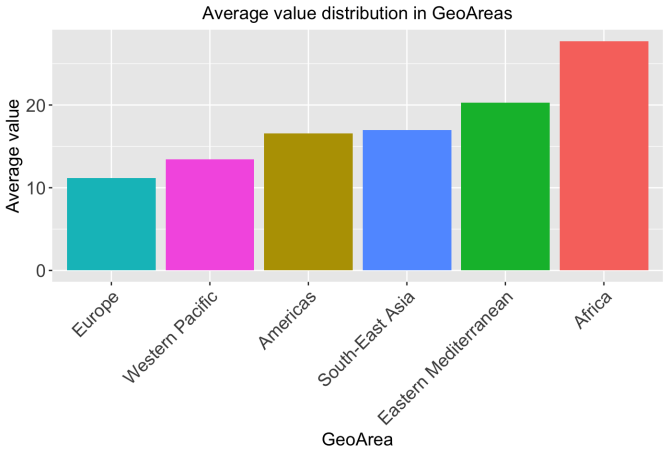

Fig. 7. Average of values in GeoAreas

D. Analysis of geographical areas

After a temporal analysis, attention shifted to the search for significant information regarding the distribution of road mortality rates in the world to identify the presence of any areas with more or less high incidence rates. Each sample within the dataset refers, as mentioned, to a specific State. Furthermore, each state belongs to one of the 6 areas into which the World Health Organization divides the world.
From the graph in Figure 6 it is possible to see that the highest values are concentrated in the African area, while the lowest values are found in the European one.

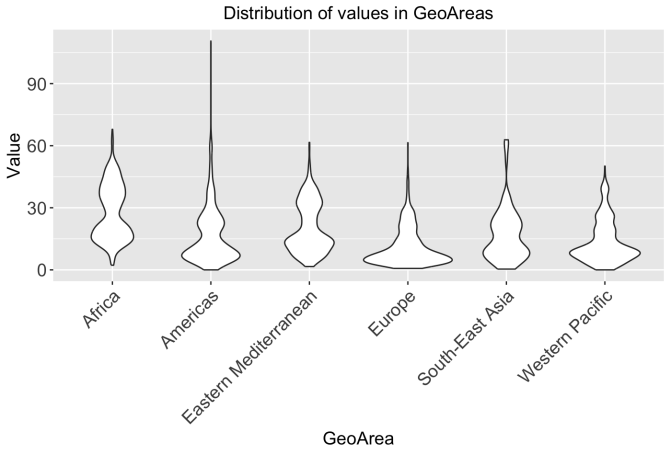

Fig. 6. Violin plot of the distribution of values

To confirm this, it is possible to observe the distribution of the average values in the various geographical areas from the graph in Figure 7. The lowest average values are concentrated in the European area, while the highest average values are concentrated in the African area.

E. Temporal analysis in geographical areas

Next, the average distribution of values over the years in the various geographical areas was analyzed. From the graph in Figure 8, it can again be seen that the African area has the highest average values in each year taken into consideration contributing to raising the global average, while the European area has the lowest trend.

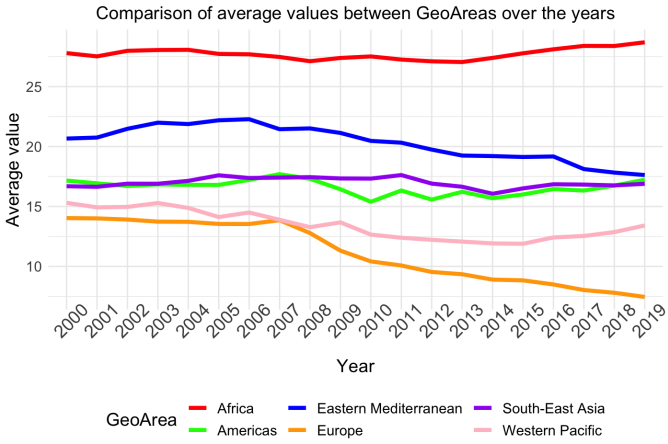

Fig. 8. Average of values in GeoAreas over the years

F. Focus on Africa and Europe

In this section, the African and European areas have been analyzed in more detail, being the areas with the highest and lowest average values, respectively. All states belonging to these areas were considered. Since each state has its own administrations, only the average values of each were reported. The difference between the two genders in these areas was also shown.

*a) **Africa**:* The African area has the highest number of road fatalities. Figure 9 shows the various African states. With the exception of a few, the differences between the states are not so stark. The most affected state is Zimbabwe.

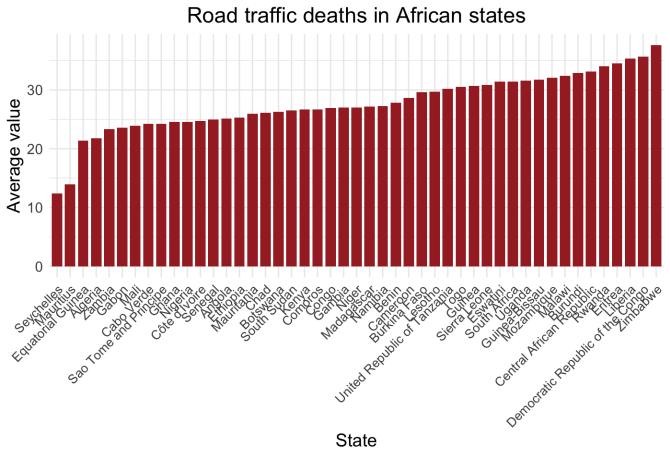

Fig. 9. Average values in African states

As shown in Figure 10, the male population consistently has higher average values than the female population. The difference between the two genders for Africa is very clear: the female gender never exceeds an average of 20, while the male gender travels at an average of about 40.

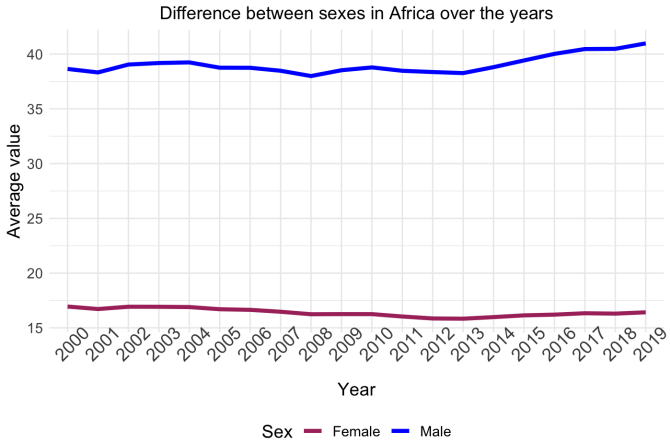

Fig. 10. Difference between males and females in Africa over the years

*b) **Europe**:* The European area has the lowest number of traffic fatalities. In the European area, the difference between states is much more pronounced. Figure 11 shows that the most affected state is the Russian Federation.

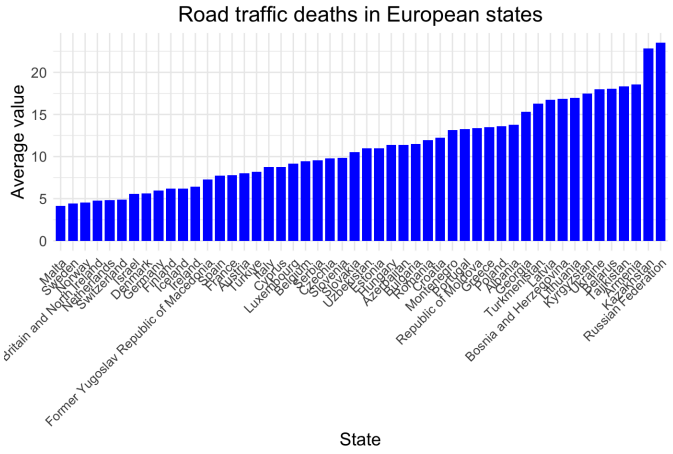

Fig. 11. Average values in European states

In Figure 12 we again see how the male gender takes higher values on average than the female gender. The male gender shows a sharp peak in 2007, thereafter there seems to be a decline in the average number of traffic fatalities.

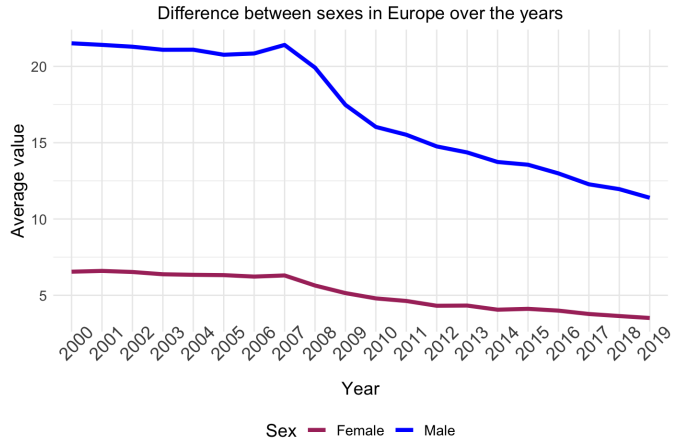

Fig. 12. Difference between males and females in Europe over the years

G. Other areas

For the GeoAreas between Europe and Africa, the differences between the two genders over the years have been reported. As shown by the graph in Figure 13, even for the intermediate areas, the male population takes a higher average value than the female population, with a clear difference in each area. The area with the highest number of traffic fatalities would appear to be the Eastern Mediterranean.

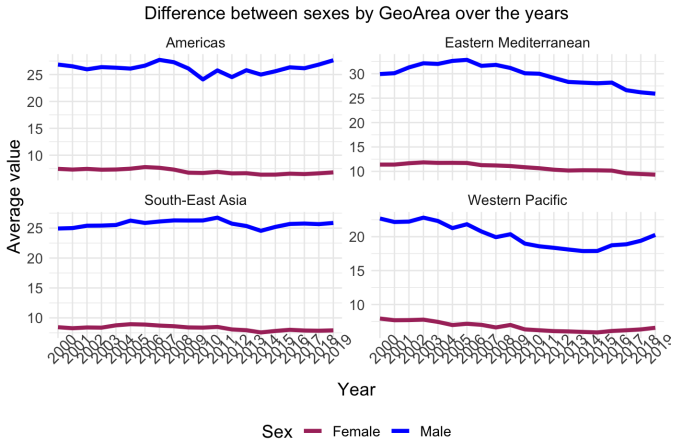

Fig. 13. Difference between males and females in the other areas over the years

H. Difference between Italy and Zimbabwe

To conclude the exploratory analysis, the most affected state in the African region and the Italian state were compared with each other. Figure 14 clearly shows how, in each year considered, the average number of road fatalities in Zimbabwe is higher than in Italy.

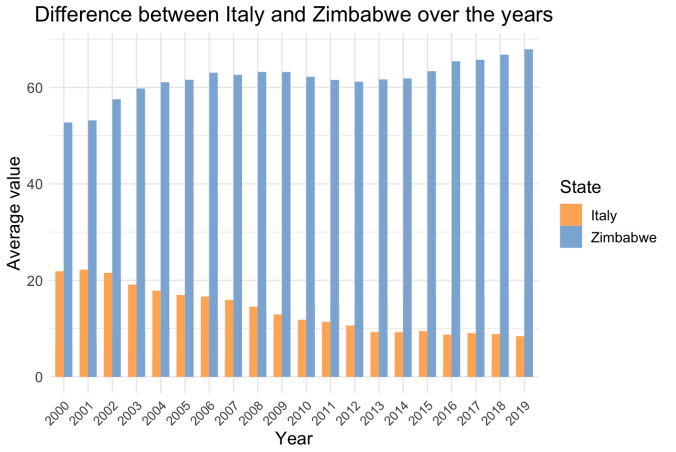

Fig. 14. Average of values in Italy and Zimbabwe over the years

V. Statistical Tests

After careful exploratory analysis, some preliminary conclusions have been deduced from the data examined. In this section, previous observations on the data will be confirmed through the use of statistical tests.

A. Test on feature Sex

During the EDA, it was possible to notice some differences in the distributions of values between males and females. Therefore, an attempt was made to analyze this issue more closely through the application of statistical tests. First, normality was checked, as it is a basic requirement of many statistical tests. It was decided to use Shapiro's test.

TABLE I
Shapiro's test for both sexes

Sex	Statistics W	P-Value
Male	0.98326	$< 2.2 \times 10^{-16}$
Female	0.9464	$< 2.2 \times 10^{-16}$

The results of Shapiro's test shown in Table 1 for normality of the data indicate that both groups, male and female, have p-values less than 0.05, suggesting a significant deviation from normality. In other words, the data in the two groups do not follow a normal distribution. It was therefore necessary to use non-parametric tests to see whether the distributions between the two sexes had statistically significant differences.

The Mann-Whitney test is a non-parametric test used to compare the distributions of two independent groups. The results of the test shown in Table 2 indicate a very low p-value suggesting a significant difference in the distributions between the two groups. The null hypothesis was rejected and it was concluded that there is sufficient evidence to state that there is a significant difference in the distributions of the variable "Value" between the male and female groups. In practical terms, this indicates that there is a statistically significant disparity between the two gender categories with respect to the variable of interest.

TABLE II
Mann-Whitney test for both sexes

Group	Statistics W	P-Value
Male vs. Female	1546545	$< 2.2 \times 10^{-16}$

B. Test on feature GeoArea

During the EDA phase, it was also noticed that there were differences between the various GeoAreas. Therefore, it was decided to carry out statistical tests between geographic areas. One possible test is ANOVA, however it requires normality of data and homoschedasticity. Therefore, Shapiro's and Barnett's tests were applied to test for the two conditions, respectively. Initially, the normality of the data was checked through the Shapiro-Wilk test for each "GeoArea" group. The results showed extremely low p-values for each GeoArea, indicating that the data do not follow a normal distribution. Next, homoschedasticity was examined by Bartlett's test. The result indicated significant differences in variances between the groups, violating the assumption of homoschedasticity required for ANOVA. Consequently, it was decided to apply the Kruskal-Wallis test, a non-parametric version of ANOVA that is used when the data do not meet the assumptions of normality and homoscedasticity required by ANOVA. The results of the test showed a significantly high chi-square with a very low p-value, indicating significant differences between at least two groups. Test results are shown in Table 3.

TABLE III
RESULTS OF SHAPIRO-WILK, BARTLETT, AND KRUSKAL-WALLIS TESTS

Test Performed	K-Squared Statistic	P-Value
Shapiro-Wilk (Africa)	-	5.22×10^{-25}
Shapiro-Wilk (Americas)	-	1.03×10^{-32}
Shapiro-Wilk (E.M.)	-	7.59×10^{-20}
Shapiro-Wilk (Europe)	-	3.67×10^{-40}
Shapiro-Wilk (S.E. Asia)	-	7.10×10^{-19}
Shapiro-Wilk (W. Pacific)	-	6.08×10^{-24}
Bartlett's Test	372.13	$< 2.2 \times 10^{-16}$
Kruskal-Wallis Test	1825.9	$< 2.2 \times 10^{-16}$

1) Post-Hoc Analysis: from the results obtained, it was necessary to understand whether the differences were due to a single area or affected all of them. It was therefore necessary to apply a post-hoc analysis by applying the Mann-Whitney non-parametric statistical test applied to all pairs of geographic areas. The test was performed using first the Bonferroni correction and after the Benjamini-Hochberg correction.
Both of these corrections were used to handle the problem of multiple comparisons, reducing the risk of obtaining false positives.

TABLE IV
RESULTS OF MANN-WHITNEY TESTS WITH BONFERRONI CORRECTION BETWEEN PAIRS OF GEOGRAPHIC AREAS

GeoArea1	GeoArea2	P-Value
Americas	Africa	$< 2 \times 10^{-16}$
Americas	Eastern Mediterranean	$< 2 \times 10^{-16}$
Americas	Europe	$< 2 \times 10^{-16}$
Americas	South-East Asia	1
Americas	Western Pacific	0.00066
Eastern Mediterranean	Africa	$< 2 \times 10^{-16}$
Eastern Mediterranean	Europe	$< 2 \times 10^{-16}$
Eastern Mediterranean	South-East Asia	4.5×10^{-08}
Eastern Mediterranean	Western Pacific	$< 2 \times 10^{-16}$
Europe	Africa	$< 2 \times 10^{-16}$
Europe	South-East Asia	$< 2 \times 10^{-16}$
Europe	Western Pacific	1.9×10^{-10}
South-East Asia	Western Pacific	5.7×10^{-05}

TABLE V
RESULTS OF MANN-WHITNEY TESTS WITH BENJAMINI-HOCHBERG (BH) CORRECTION BETWEEN PAIRS OF GEOGRAPHIC AREAS

GeoArea1	GeoArea2	P-Value
Americas	Africa	$< 2 \times 10^{-16}$
Americas	Eastern Mediterranean	$< 2 \times 10^{-16}$
Americas	Europe	$< 2 \times 10^{-16}$
Americas	South-East Asia	0.19
Americas	Western Pacific	4.7×10^{-05}
Eastern Mediterranean	Africa	$< 2 \times 10^{-16}$
Eastern Mediterranean	Europe	$< 2 \times 10^{-16}$
Eastern Mediterranean	South-East Asia	3.8×10^{-09}
Eastern Mediterranean	Western Pacific	$< 2 \times 10^{-16}$
Europe	Africa	$< 2 \times 10^{-16}$
Europe	South-East Asia	$< 2 \times 10^{-16}$
Europe	Western Pacific	1.7×10^{-11}
South-East Asia	Western Pacific	4.4×10^{-06}

The Bonferroni and Benjamini-Hochberg corrections both reveal a consistent pattern of statistically significant differences in the distribution of values between different geographical areas. In particular, the regions of the Americas, the Eastern Mediterranean and Europe consistently show significant disparities. The Bonferroni correction aims to control the familywise error rate (FWER), i.e. the probability of having at least one first species error, and is known for its tendency to be conservative. Its restrictive nature can lead to higher p-values, thus reducing the level of statistical significance. In other words, the Bonferroni correction may make it more difficult to detect significant differences between the groups or conditions analysed. On the other hand, the Benjamini-Hochberg correction is intended to control the false discovery rate (FDR), i.e. the portion of false positives, and is generally more permissive and used especially when performing simultaneous tests on a large number of groups. Both corrections confirm the presence of significant differences between geographical areas with the exception of America and South-East Asia.

VI. FINAL CONSIDERATIONS

The analyses performed show the number of road deaths over the years taking into account differences due to sex and geographic area. The use of graphs and statistical tests made it possible to observe that the number of deaths is strongly influenced by sex and geographical area. The use of more data could be helpful in investigating more deeply the reasons for these differences. The results presented should be interpreted only indicatively, as a possible basis for further studies.

REFERENCES

[1] Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK525212/

[2] Retrieved from https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[3] Dataset from https://www.who.int/data/gho/data/indicators/indicator-details/GHO/estimated-road-traffic-death-rate-(per-100-000-population).