MATH 240 Final Project Manuscript

Investigating deaths in Spain during COVID-19 Pandemic

By Felipe Delclaux and Victoria Lopez de la Serna

Abstract

The rapid spread of COVID-19 is posing ever-experienced threats to our social, economic and health systems. The whole world is fighting a battle of uncertainty and chaos combined with untraceable death tolls. One of the most prominent challenges is predicting and describing the deaths caused by the virus. Some governments seem to inform the public poorly on the main drivers of deaths, masking the real threat and risk of this virus. Spain is a prime example of this issue. As Spaniards, we have decided to conduct a study to try to understand the main drivers of death and main groups at risk. To do so, we compiled a series of data from the '*Ministerio de Sanidad*' (Health Ministry). The data we decided to use is starting exactly two weeks after official government lockdown (March 16th, 2020), from March 30th, 2020 to April 27th, 2020. We chose to do this to avoid having any carry-over effects from pre-lockdown cases and trying to get the most stable (invariable) context possible. This source gives us a daily update on the total number of Deaths, Confirmed Cases, people that have been Hospitalized, and people that have been put in the Intensive Care Unit (ICU). Moreover, it gives us information on the Age Range and Gender of the people. Hence, for solving our problem, we will investigate what the effects of these variables (Confirmed Cases, Hospitalized, ICU, Age, Gender, and Days since lockdown) are on the total death count in the selected range of dates and hopefully draw some conclusions on what factors significantly affect and explain the number of deaths. We fit this data to a multiple regression model to consider the significance of each variable in explaining deaths over time. Moreover, we conducted an analysis of variance on the differences of average death rates based on age and gender to get a closer look at the susceptibility of each of these categories to the virus.

Introduction

The purpose of our study lies in understanding the factors that are important to death tolls in Spain as well as what groups are most at risk from contracting the virus. Our main scientific questions are 1) To understand what variables explain these deaths tolls, and to a further extent to quantify the effect of those variables, and 2) To find what groups are most at risk, and what factors put these groups at risk. We aim to ultimately to see if the variables included in our model provide an effective explanation of the questions in hand by interpreting them using the methods presented in the methods and materials section.

Methods and Materials

In order to investigate the main factors in explaining the number of deaths due to Covid-19 in Spain, we will investigate how the total number of confirmed cases, the total number of hospitalized cases, the total number of patients admitted into the ICU, the number of days since lockdown as well as the age and genders of these patients affect the total number of deaths during lockdown in Spain. The number of cases, hospitalized, ICU and days since lockdown are all separate variables, whilst age and gender are two categorical variables with Young(baseline), Adult and Elderly as well as Male(baseline) and Female as their respective categories. Moreover, in order to analyze, modify and organize our data, as well as to build our models, we will use the R programming language on the R Studio suite. Within R Studio we will use the *car, leaps* and *tseries* packages in order to be able to look at Variance Inflation Factors, relevel our categorical variables, conduct subsets regression, and Augmented Dickey-Fueller tests. Moreover, we will be fitting a multiple regression model to our data looking at which variables explain what portions of the variability in the data. We will also be fitting the data to a two-way analysis of variance (ANOVA) model in order to look at the differences in average death rates ($\frac{Total\ Deaths}{Total\ Confirmed\ Cases}$) for each of the categories in our two categorical variables.

Multiple Regression Model

In order to take a general look at our data, we created a pairs plot, seen in *Appendix A Figure 1,* to see the plots of the number of deaths against every explanatory variable. Moreover, we also wanted to see how variables affected each other to look at potential collinearities in the data. When looking at the plot, we can see apparent strong positive linear relationships between the number of deaths and all the other variables: Confirmed Cases, Hospitalized, ICU, and Days. Moreover, these observations are backed by looking at the numerical results given by our correlation matrix seen in *Appendix A Figure 2*, we get the following correlation coefficients for those variables respectively: 0.6739, 0.9241, 0.8680 and 0.2639. However, we can also see strong relationships between some of the explanatory variables. This could be due to a strong structural multicollinearity since a fraction of the confirmed cases ends up being hospitalized and fraction of those hospitalized end up being admitted into the ICU. When we look at the correlation coefficients between these three variables (Confirmed Cases – Hospitalized: 0.8548, ConfirmedCases – ICU: 0.7551, Hospitalized-ICU: 0.9493) we can see that our initial observations were correct, and there seems to be very strong positive linear relationships between the explanatory variables and therefore, potential strong multicollinearity.

Continuing with our exploratory analysis, if we look at the boxplot of deaths based on our three age categories seen in *Appendix A Figure 3*, we can see there seems to be a clear difference between the elderly and the two other age ranges. Nevertheless, there does seem to be some overlap between Elderly and Adult, as there is a large amount of variation within the Elderly category. Moreover, if we look at the boxplot of deaths based on gender, seen in *Appendix A Figure 4*, we can see there is not a clear difference between the two. The means seem to be close together, although the male category seems to have a wider range of values and more variation seen by a larger Inter-Quartile range.

If we think about the nature of the data, most of our explanatory variables are driven by time, and so is our response variable. In class, we have not dealt with time-series data yet. However, from preliminary research, we know that comparing two non-stationary variables that are driven by time can result in spurious regression. This means that, although there seems to be a very high correlation between the variables and the resulting linear model would account for a high amount of the variability in the response variable (high R-squared value), there might not be an actual explanatory relationship between the variables, which is what we are interested in looking at.

We conducted an Augmented Dickey-Fuller Test with hypotheses $H_0: non-stationary$ vs $H_1: stationary$ to see if our data was non-stationary, and could therefore result in spurious regression. We obtained the following statistics for our time driven variables: Deaths (Dickey Fuller statistic = -2.3048, p-value = 0.4491) as seen in *Appendix A Figure 5*, Confirmed Cases (Dickey-Fuller statistic = 0.0799, p-value = 0.99) as seen in *Appendix A Figure 6* Hospitalized (Dickey-Fuller statistic = -1.5325, p-value = 0.7714) as seen in *Appendix A Figure 7*, ICU (Dickey – Fuller statistic = -1.9065, p-value = 0.6153) as seen in *Appendix A Figure 8*. As we can see from the results, none of the p-values are under our assumed 0.05 significance level, so we cannot reject the null hypothesis, leading us to conclude that there is not enough data to claim stationarity in our data. This means, that a model which includes these explanatory variables could lead to spurious regression. There are various methods to deal with non-stationarity, however, we have decided to try and deal with the data through what we have learnt in class.

We considered a multiple regression model, associating deaths to all the other variables, of the following type:

$$Deaths_i = \beta_0 + \beta_1 Cases_{i1} + \beta_2 Hospitalized_{i2} + \beta_3 ICU_{i3} + \beta_4 Days_{i4} + \beta_5 Adult_{i5} + \beta_6 Elderly_{i6} + \beta_7 Female_{i7} + \epsilon, \text{ where } \epsilon \sim N(0, \sigma^2) iid \text{ and } i = 1, \dots, n.$$

Moreover, where $Adult, Elderly\ and\ Female$ are binary variables whose coefficients measure the difference in effect between Adult and Young age range categories, Elderly and Young age range categories and Female and Male gender categories, respectively. After fitting the model to the data we can immediately see that the output, seen in *Appendix A Figure 9,* suggests multicollinearity, as we obtain negative values for $\hat{\beta}_1, \hat{\beta}_3, \hat{\beta}_4$ and $\hat{\beta}_6$ when we expected a positive coefficient based on what we saw in our exploratory analysis (the pairs-plot and the boxplot). Moreover, some of the standard errors and p-values seem inflated. For example, the standard error for $\hat{\beta}_0$(363.1) and $\hat{\beta}_1$(-0.0117) are about the same as the actual $\hat{\beta}_0$ (108.8) and $\hat{\beta}_1$(-0.0123) values. Their p-values (0.0737 and 0.2919) also seem to be inflated making them non-significant (above 0.05). Therefore, we decide to look at the variance inflation factor (VIF) for each of the predictor variables, seen in *Appendix A Figure 10.* We obtain very high VIF values for confirmed cases (24.3446), hospitalized (73.0290), ICU (20.9764) and Age Range (47.3980), way above an acceptable 2.5 VIF value. Moreover, we have a high R-squared value (0.9656) which means our model accounts for 96.56% of variability in the data, however, this could be an indicator of a potential spurious regression.

Therefore, given the high amounts of multicollinearity and the risk of spurious regression from including the total number of cases, hospitalized, and patients admitted into the ICU, we decided to exclude them from our model. Consequently, reducing our model to a simpler multiple regression model, associating deaths to days since lockdown, age range, and gender. By taking a closer look at the plot of deaths and days, seen in *Appendix A Figure 11*, and considering the exponential nature of the growth of a pandemic, we realize that deaths seem to grow exponentially with time, rather than linearly. Therefore, we decided to perform a logarithmic transformation on deaths, to compensate for its exponential growth over time. Simply by looking at the plot of the natural logarithm of deaths against days, seen in *Appendix A Figure 12*, we can now see a much more linear relationship, and a clear

distinction between the different age ranges, as well as the different genders within these age ranges. We have three seemingly parallel and equidistant bands of data points representing the age ranges, 'Elderly' being the highest up, followed by 'Adult' and then 'Young', as well as two seemingly parallel lines within these bands which represent the different genders, with 'Male' being above 'Female' in every age range. These patterns seem to be most clear in 'Elderly' and least clear in the 'Young' age range. In trying to build the best possible model to fit the data, we conduct a subsets regression, looking at the adjusted R-squared and Mallow's Cp for each model (results seen in *Appendix A Figure 15)*. It seems the output of the subsets regression suggests a model with all considered predictor variables (days, age and gender) is best. We can see a significant increase in adjusted R-Squared (0.0115) from the previous model considered and a very significant drop in Mallow's Cp (from 494.3519 to 5) which brings the value way closer to the number of predictors, suggesting that it is a better model. Therefore, we consider the following model:

$$\ln(Deaths_i) = \beta_0 + \beta_1 Days_{i1} + \beta_2 Adult_{i2} + \beta_3 Elderly_{i3} + \beta_4 Female_{i4} + \epsilon$$

$$\left(= Deaths_i = e^{\beta_0 + \beta_1 Days_{i1} + \beta_2 Adult_{i2} + \beta_3 Elderly_{i3} + \beta_4 Female_{i4} + \epsilon}\right)$$

$$\text{where } \epsilon \sim N(0, \sigma^2) iid \text{ and } i = 1, \dots, n.$$

By fitting this model to the data, we obtain a model with a multiple R-squared value of 0.9963, meaning our model explains 99.63% of the variability in the data. From the subsets regression output, it seems age range is a very significant variable as *Adult* and *Elderly* were added to the model first. For that reason, we wanted to quantify the overall Age Range effect by considering a reduced model of the following type:

$$\ln(Deaths_i) = \beta_0 + \beta_1 Days_{i1} + \beta_2 Female_{i2} + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2) iid \text{ and } i = 1, \dots, n$$

By fitting this reduced model to the data, we obtain a model with a multiple R-squared value of 0.0457. Hence, we can calculate the following partial R-squared: $\frac{0.9963-0.0457}{1-0.0457} = 0.9961$. This tells us that our Age Range variable explains an additional 99.61% proportion of the variability left unexplained by the reduced model, showing us the huge importance of the variable in our full model.

Moreover, when looking at the residuals after fitting the data to this model, seen in *Appendix A Figure 16*, it seems our homoscedasticity assumption is partially met as there are some outliers like points 122,128 and 163. Also, our normality assumption seems to be met, backed with a Shapiro-Wilkins test, seen in *Appendix A Figure 17* (W statistic = 0.99217, p-value = $0.4962 > 0.05$). However, our residuals vs fitted plot also shows three negatively sloped groups of residuals, which indicates our model does not account for some sort of interaction.

Hence, we consider the following model for our data, which considers potential interactions between predictor variables:

$$\ln(Deaths_i) = \beta_0 + \beta_1 Days_{i1} + \beta_2 Adult_{i2} + \beta_3 Elderly_{i3} + \beta_4 Female_{i4} + \beta_5 Days \cdot Adult + \beta_6 Days \cdot Elderly + \beta_7 Days \cdot Female + \beta_8 Adult \cdot Female + \beta_9 Elderly \cdot Female + \beta_{10} Days \cdot Adult \cdot Female + \beta_{11} Days \cdot Elderly \cdot Female + \epsilon$$

where $\epsilon \sim N(0, \sigma^2) iid$ and $i = 1, \dots, n$.

By fitting this model to the data, we obtain the following regression equation:

$$\ln(\widehat{Deaths}) = 0.6914 + 0.0585 Days + 3.2317 Adult + 6.3164 Elderly - 0.9387 Female + 0.0152 Days \cdot Female - 0.0153 Days \cdot Adult \cdot Female$$

When looking at the model's residuals, seen in *Appendix A Figure 19*, we can still see three groups of points in the residuals vs fitted plot. However, these groups are probably due to the different age ranges and residuals within these groups are much more random, so we could say our linearity

assumption is met. Moreover, our homoscedasticity assumption is more reasonably met, with just a couple of outliers like data points 170 and 211 that could be driving the shape of the distribution. Finally, our Normal Q-Q plot looks good, as almost all the points are on the 45-degree dotted-line, and a Shapiro-Wilkins test, seen in *Appendix A Figure 20,* gives us a W statistic of 0.99264 and a p-value of 0.5527, showing us there is not enough evidence to prove non-normality in the model's residuals. Therefore, all our model assumptions seem to be met, and our model seems to be appropriate.

In the summary output obtained by fitting our data to this regression model, seen in *Appendix A Figure 18*, we can see our multiple R-squared value has gone up to 0.9972, compared to 0.9963 from our additive model. We can calculate the partial R-squared to find the additional variability explained by our interactions. $Partial_{Rsq} = \frac{0.9972-0.9963}{1-0.9963} = 0.2432$, this tells us that the included interactions explain an additional 23.32% of the variability in the data left unexplained by the additive model. Furthermore, we can see the significance levels of our predictor variables as well as the newly included interactions in our model by considering these hypothesis tests: H_0: $\beta_i = 0$ vs H_A: $\beta_i \neq 0$, where $i = 0, \dots, 11$. We reject the null hypothesis for the following coefficient estimates with their respective test statistics, as their p-values are below our 0.05 significance level:

$$\beta_0 \ (t = 7.079, p = 4.65e-11),\ \beta_1(t = 17.732, p < 2e-16),\ \beta_2(t = 23.296, p < 2e-16),$$
$$\beta_3(t = 45.729, p < 2e-16),\ \beta_4(t = -6.796, p = 2.15e-10),\ \beta_7(t = 3.256, p = 0.00139),$$
$$\beta_{10}(t = -2.317, p = 0.0218).$$

Hence, we can conclude that the days since lockdown, being an adult as compared to being young, being elderly as compared to being young and being a female as compared to being male are all highly statistically significant in determining the total number of deaths due to Covid-19 in Spain. Moreover, the effect of days is dependent on gender, and the effect of days on adults specifically as compared to others is also dependent on gender. Not only does that give us a certain extent of

prediction, through extrapolation, but it also gives us a level of inference in determining what factors are determining death in Spain due to Covid-19.

Interpreting the significant coefficients allows us to investigate the effects of our variables. However, considering our logarithmic transformation, we should undo the transformation to better understand our results. Given our model is of the form $Deaths_i = e^{\beta_0 + \cdots}$, a unit increase in any predictor i, while all others are constant, will result in a multiplication of the number of deaths by a factor of e^{β_i}. Hence, we know that on average, keeping all other variables constant (remembering that our baseline for categorical variables are young males), a unit increase in days since lockdown will multiply the number of deaths by 1.0602 times overall. Moreover, our interactions tell us that, on average, for every unit increase in days since lockdown, deaths are an additional 1.0153 times higher for females than they are for males. Finally, our last interaction tells us that for every unit increase in days since lockdown, deaths in adults are not 1.0153 times higher for females than males, instead they are about the same ($e^{0.015181-0.015279} \approx 1$). Therefore, keeping everything constant, an unit increase in days will multiply the number of deaths by 1.0602, except for elderly and young women, for which it will multiply it by ($e^{0.058468+0.05181} =$) 1.0764. Furthermore, the number of deaths is 25.3220 times higher in adults than in young people; the number of deaths is 553.6280 times higher in elderly than in young people; the number of deaths in females is only 0.3911 times that in males.

Analysis of Variance model

In our multiple regression model above, we saw that age range and gender can significantly affect the number of deaths. However, we do not know if that is due to a difference in death rate within the categories, (death rate calculated as deaths per confirmed cases), or because there were simply more confirmed cases in elderly people which consequently lead to a higher death toll. Thus, our multiple regression model shows that a difference in age and gender does exists, but it limits our study to what

causes that difference and where that difference lies. Therefore, we extend our study beyond the number of deaths and instead look at the death rate, in order to take a closer look at difference in susceptibility to the virus between age ranges and genders. In order to so, we conduct a Two-Way Analysis of Variance (ANOVA) Additive Model to find whether a statistically significant difference exists in mean death rates between different ages and gender. Given the two-way ANOVA model:

$$\mathrm{Y_{ijk}} = \mu + \tau_\mathrm{i} + \gamma_\mathrm{j} + \epsilon_\mathrm{ijk}$$

$$\text{where } \mathrm{i} = 1 \ldots . \mathrm{I}; \mathrm{j} = 1 \ldots \mathrm{J}; \mathrm{k} = 1 \ldots \mathrm{K}$$

We want to test $H_0: \tau_1 = \cdots = \tau_I = 0$ vs. $H_1: otherwise$ and $H_0: \gamma_1 = \cdots = \gamma_J = 0$ vs. $H_1: otherwise$, where $Y =$ Death Rate, $\tau_i =$ the deviation between the average death rate among people at the level *i* of the Age Range variable and the overall average death rate, and $\gamma_j =$ the deviation between the average death rate among people at the level level *j* of the gender variable and the overall average death rate.

However, we suspect and anticipate that a certain challenge might arise given the exponentiality of the data, as occurred with the Deaths variable in the multiple regression model above. Nonetheless, we conduct the model explained above, but abstain from explaining each result (boxplots, ANOVA table), and jump straight to our residuals to check if indeed there is something wrong with our model. The respective boxplots and EDA(*Figures 1,2 and 3*) and ANOVA table (*Figure 4*) are found in the *Appendix B* at the end.

Plotting our residuals to check whether the conditions of the two-way ANOVA model of death rates against age range and gender our met, found in *Appendix B Figure 5*, we see that our prediction seems to be true. The residuals assumptions don't seem to be met. The Residuals vs Fitted plot shows fanned out and U-shaped data indicating heteroskedasticity, and the S Shaped Normal Q-Q plot shows a violation of the normality assumption, challenging the validity of our model. We realize that moreover,

death rate is calculated by deaths per confirmed cases, both variables being exponential. Hence, as in our multiple regression with deaths, it seems death rates should be exponential too, and thus we take the natural logarithm of death rate to be our new response variable and proceed in this way with our ANOVA analysis. To clarify, future use of Death Rates refers to the natural logarithm of death rates. For simplicity purposes, we will occasionally simply say death rate. Our new two-way ANOVA model takes:

$$\mathrm{Y_{ijk}} = \mu + \tau_i + \gamma_j + \epsilon_{ijk},$$

where $\mathrm{i} = 1 \ldots . \mathrm{I}; \mathrm{j} = 1 \ldots \mathrm{J}; \mathrm{k} = 1 \ldots \mathrm{K}$, and where $Y =$ Natural Log of Death Rate, $\tau_i =$ the deviation between the average death rate among people at the level *i* of the Age Range variable and the overall average death rate, and $\gamma_j =$ the deviation between the average death rate among people at the level *j* of the gender variable and the overall average death rate. We want to test $H_0: \tau_1 = \cdots = \tau_I = 0$ vs. $H_1: otherwise$ and $H_0: \gamma_1 = \cdots = \gamma_J = 0$ vs. $H_1: otherwise$.

To conduct Exploratory Data Analysis (EDA) we create comparative boxplots for the two genders (*Appendix B Figure 6*) as well as for the three age ranges (*Appendix B Figure 7*), and support our analysis with the Means for each variable (*Appendix B Figure 8*). In *Figure 6*, although there seems to be some overlap between the two boxplots, and that there is large variation within each group, we see possible evidence that the mean death rates in each gender are different, with men showing a higher mean death rate than women. In Figure 7, we see that there seems to be a significant difference in means between age ranges. While it is currently questionable whether the means are statistically significantly different among the adult and young groups, the mean death rate for the elderly is potentially statistically significantly different from the other means**.**

From the two-way ANOVA table output shown in Appendix B Figure 9, we find that there exists a statistically significant effect due to Age Range (F=3705.6, with 2 and 164 df, pval<2e-16) and Gender (F=398.7, with 1 and 164 df, pval<2e-16). Following the most standard choices of defining a 0.05

significance level, results in rejecting the null-hypothesis and claiming that a difference in means in death rate between both, age ranges and gender, does exist.

When checking whether the conditions of the two-way ANOVA model above are met, we find in *Appendix B Figure 10* that the results still pose some challenges to our model. The inverted U-shaped Residuals vs. Fitted plot implies that we have still not accounted for something in the model, while the normality assumption seems fine given the Normal QQ Plot and the Shapiro-Wilk normality test in *Appendix B Figure 11*. It seems our model could be omitting a potential interaction between our two variables (Age Range and Gender). Thus, we check our prediction by plotting an interaction plot, found in *Appendix B Figure 12*. Some may not feel that the plots are very different from being parallel to each other, while we in fact find that it seems more convincing to claim that the plot produces non-parallel line segment shapes and that a potential interaction could appear to exist between age range and gender, where it seems that magnitude of death also depends on the gender of the person. Consequently, we perform a two-way ANOVA including an interaction term to quantify the significance of this interaction. We take the model:

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk} = \mu + \tau_i + \gamma_j + \omega_{ij} + \epsilon_{ijk} \, ,$$

where $i = 1, \dots, I; j = 1, \dots, J; k = 1, \dots, K$, and where $Y =$ Natural Log of Death Rate, $\tau_i =$ the deviation between the average death rate among people at the level *i* of the Age Range variable and the overall average death rate, and $\gamma_j =$ the deviation between the average death rate among people at the level level *j* of the gender variable and the overall average death rate, and ω_{ij} = interaction effect at level *i* of the Age Range variable, and level *j* of the Gender variable. We want to test H_0: $\omega_{ij} = 0 \; \forall i, j$ vs. H_1: $\omega_{ij} \neq 0 \; \forall i, j$.

With the interaction question in hand, we begin by conducting further EDA and create partial boxplots (*Appendix B Figure 13 and 14*). If we compare the partial boxplot for Gender *(Figure 13*) to the

regular boxplot (*Figure 6*), we see that Figure 6 doesn't really display strong real potential difference regarding gender because distribution within the respective groups contains output based on age range. The partial boxplot in Figure 13 removes that age effect to better display the difference based on gender, where in fact we see that the means do differ, showing a higher mean death rate in men than in women, and that there is no longer much overlap between the two groups. We see that indeed, the boxplots in Figure 6 which included the effect of age, was masking the real gender effect that seems to exist. Furthermore, in *Appendix 2 Figure 15*, we find a visual representation of boxplots for both age range and gender all in one figure, allowing us to have a deeper sense of the potential interaction between our variables, and to make some more detailed predictions and comparisons between the two. Within the adult group, for example, we see how the boxplots for each gender do not even overlap, showing that adult men do seem to show higher death rates than women, on average. The same scenario seems to apply to the elderly group. For the young, there seems to be some overlap, but we can also see that there seems to be a difference in means. Interestingly, we see that adult women and young men seem to show almost equal mean death rates, whilst the other categories seem to show stronger differences in their respective means. All the above EDA supports our claim that there seems to be a difference in average death rates between age and gender and more importantly, it strengthens our prediction that an interaction might exist between our two variables.

Fitting the two-way ANOVA model with interaction to the data, we arrive at the corresponding ANOVA table in *Appendix 2 Figure 16*. We find that the interaction term in our model gives the following test statistics: F=38.92, with 2 and 162 df, p-value = 1.572e-14. Given the small p-value being less than the 0.05 assumed significance level, we reject the null hypothesis and claim that there is a statistically significant interaction effect between Age Range and Gender. Thus, we can conclude that not only does

the average death rate depend both on the age and the gender, but the effect of age seems to depend on the gender of the infected person too.

Since the interaction is present, we can have some difference in death rates above and beyond what is happening at the main effects level. Thus, we must isolate our understanding of what the effects one on variable at the different levels of the other variable. To do this, we estimate the age effect (the impact of belonging to the elderly, adult or young group) for each gender. We thus build two subsets of our data, one for each gender. By doing this, we can create a one-way model that looks at how death rate is associated with age range within each gender category. This results in two one-way ANOVA tables for age effect on men and age effect on women, found in *Appendix B Figure 17* and *Appendix B Figure 19*, respectively. Looking at age affect when the person is male, we find the difference between age ranges is statistically significant (F-value = 2256 with 2 and 81 df, p-value < 2e-16). Likewise, when looking at age effect when the person is female, we also we find the difference between age ranges is statistically significant (F-value = 3211 with 2 and 81 df, p-value < 2e-16).

Given the statistical significance in both, we can conduct a Tukey's Honestly Sign Diff on age both when the person is male and then when the person is female, to have a quantifiable measure of difference in death rates due to age within the two genders. *Appendix B Figure 18 and Figure 20* contain the results from the Tukey 95% simultaneous confidence intervals, considering all pairwise comparisons of the age ranges when the person is male and female, respectively. In both Figures, we see that for both genders, the difference in average death rates in all three pairs (elderly vs. adult, young vs. adult, young vs. elderly) respectively, yield low p-values and confidence intervals that do not contain 0, meaning that all the differences in mean death rates are statistically significantly different from each other. In order to look at how the death rate between ages compares, we can notice that again in both figures, the elderly vs adult pair yields a positive difference, indicating that the item listed first (elderly) yielded a

higher value, which in this case indicates a higher death rate. The other two pairs (young vs adult) and (young vs. adult), respectively, show negative differences, indicating that here the item listed second yielded a higher value and thus a higher death rate. Thus, we can conclude that both within men and women, the order of Age Ranges from highest to lowest death rates is elderly, adult, young, respectively. Comparing the differences yielded in Figures 18 and 20, we see that in all three pairs, the differences between each pair were greater among women than men, indicating that it seems age had a greater effect in difference in death rates in women than in men.

To further extend our comparisons analysis, *Appendix B Figure 21* shows a Tukey Multiple Comparisons of means for both age range and gender, where we interestingly see that the only non-significant pair is that of adult women vs. young men. We make this conclusion given the high p-value of 0.55 and the fact that the respective confidence interval contains 0. Interestingly, this proves our initial prediction made from the boxplot in Appendix B Figure 15, where we claimed that "adult women and young men seem to show almost equal mean death rates". Thus, it appears to be that age does indeed affect death rates depending on which gender, and that the only two groups that did not show a statistically significant difference in means were adult women and young men. Overall, we can claim that death rate is higher within men than women, and that the difference in death rate is most profound when the person belongs to the elderly age range than to the adult or young ranges.

To finalize our two-way ANOVA model including the interaction, we check whether the conditions for an ANOVA model are met. Plotting our model's residuals, seen in *Appendix B Figure 22*, we find that the Normal Q-Q plot continues to look good which, backed by the Shapiro-Wilks test in *Appendix B Figure 23,* shows the normality condition is met. Most importantly, we finally see that the Residuals vs. Fitted Plot now seems reasonable and seems to satisfy the constant variance and linearity assumptions, since the apparent dependence (inverted U-shape stemming from the additive model) no longer

appears. Hence, it seems our two-way ANOVA model including the interaction between age range and gender does meet the ANOVA assumptions/conditions.

Results

Through the creation of a multiple regression model associating deaths to days since lockdown, gender and age, we have been able to identify these variables as significant explanatory variables for the number of total deaths in Spain. We have found that deaths are higher both in elderly and in adults than in young people (like ourselves), and that they are also higher in males as compared to females. Moreover, we have found that the growth rate of deaths is higher for females than it is for males, except between adults, where it seems to be the same (the exact numerical relationships can be found in *Appendix A Figure 21*). Furthermore, after knowing that there were more deaths amongst elderly and adults than amongst young people, as well amongst men when compared to women, our analysis of variance model has shown us that there are significant differences in average death rates between the different age categories for both genders (men and women). Meaning that, elderly seem to be the most susceptible to the virus, followed by adults, and then young people (the exact numerical differences can be found in *the ANOVA in Appendix B Figures 16-21*).

Discussions and Conclusion

To conclude our investigation, we wanted to mention the limitations and potential expansions to our investigation. Although we were satisfied with our results, as we were able to confidently claim and explain some of the factors that affect the number of deaths in our population, we are very aware that our model is not the best prediction tool to look at actual deaths in Spain due to Covid-19. This is because, in our model, we assume a level of significant accuracy in the Ministry of Health's reports on all these variables. However, such a contagious virus is extremely hard to track and accurately report. Firstly,

there are only so many people you can test nation-wide, not only because of the physical challenges of testing so many people but also because of the shortage of testing kits globally. Furthermore, there was large speculation regarding the reliability of Chinese tests sold to the Spanish government, as there were several batches that were returned because their accuracy was found to be under 30%. It is suspected that in such a tremendous surge in demand for these testing kits, their production quality rapidly diminished. Hence this poses a lot of doubts about our data's reliability, even if provided by the government. Firstly, lack of or inaccuracy in testing would result in a largely inaccurate count of confirmed cases. Secondly, it also prevents deaths that are due to Covid-19 to be properly reported. This is because, as a virus, Covid-19 is not the direct cause of death, instead it leads to complications (most commonly respiratory) by weakening your immune system. Hence, accurate testing is also required to keep an accurate count of how many are dying as a result of contracting the virus, as anyone who dies without having tested positive is not included in the total count, even if they had clear symptoms (something the government has been greatly criticized for).

Furthermore, there are plenty of ramifications and extensions that you could do to continue the investigation of deaths due to Covid-19. The one most closely related to our model and our data would be to learn how to deal with time-series data and be able to include some of the other time-driven variables in our model. Specifically, looking at how the numbers of ICU patients affected death rates, as a large majority of those who die from the virus, are admitted into the ICU first. Also, it seems lockdown is first and foremost, a measure to avoid the health and sanitation system's collapse, as there were an overwhelming number of patients in need of intensive care. Hence, including this variable might also allow you to look at a potential 'lock-down' effect, and its significance in preventing this collapse.

Supplementary Material

Appendices

Bibliography

Appendix A:

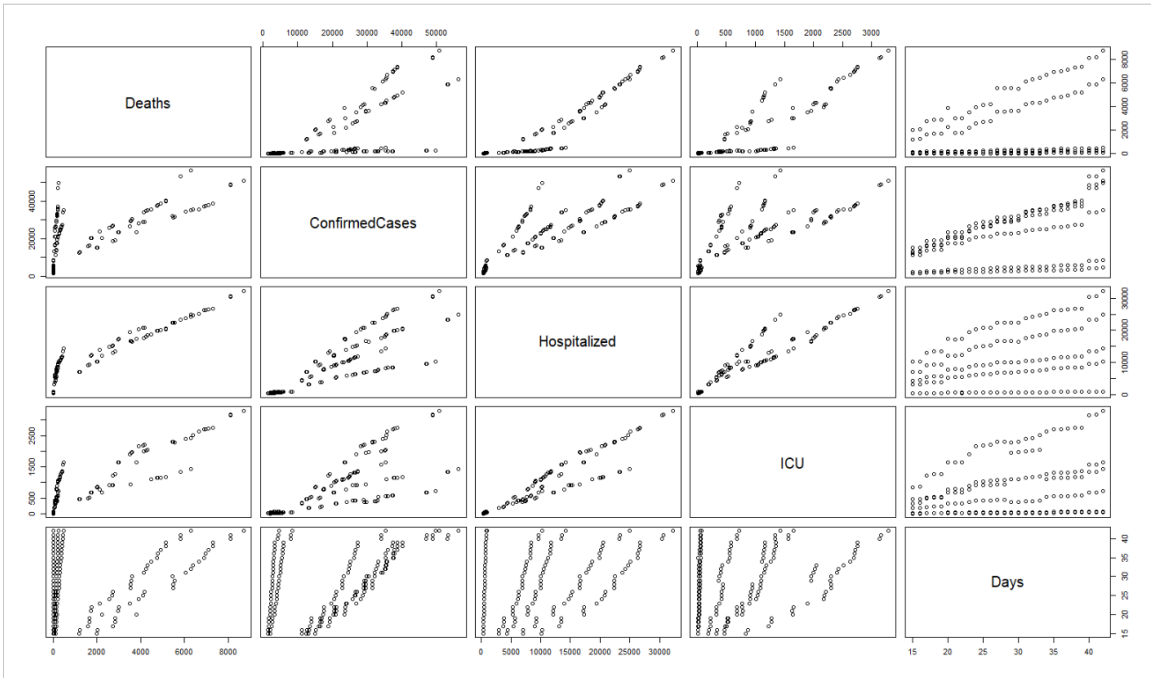

***Figure 1**: Pairs plot of all variables*

	Deaths	ConfirmedCases	Hospitalized	ICU	Days
Deaths	1.0000000	0.6739407	0.9241230	0.8679868	0.2638634
ConfirmedCases	0.6739407	1.0000000	0.8547713	0.7551402	0.4477586
Hospitalized	0.9241230	0.8547713	1.0000000	0.9493414	0.3010384
ICU	0.8679868	0.7551402	0.9493414	1.0000000	0.2919047
Days	0.2638634	0.4477586	0.3010384	0.2919047	1.0000000

***Figure 2**: Correlation Matrix of all variables*

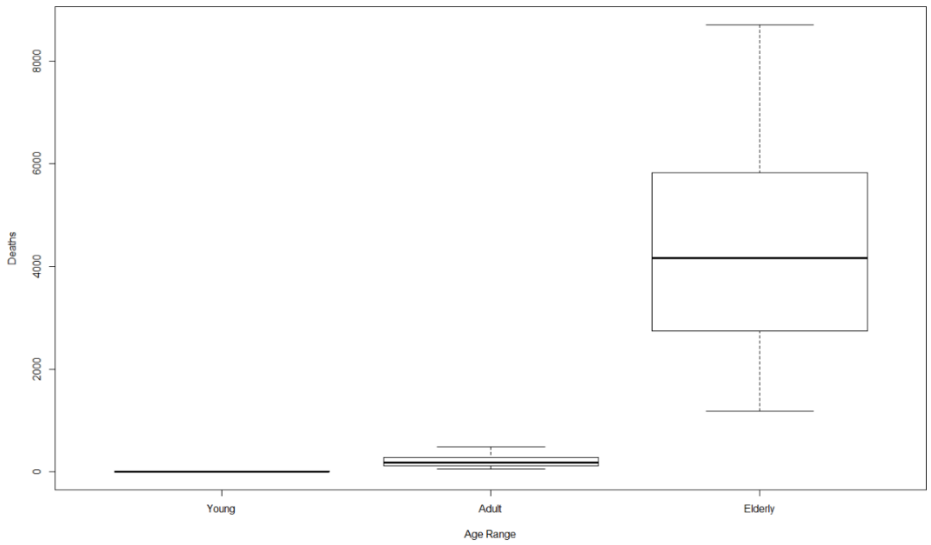

Figure 3: *Boxplot of Deaths against different age ranges*

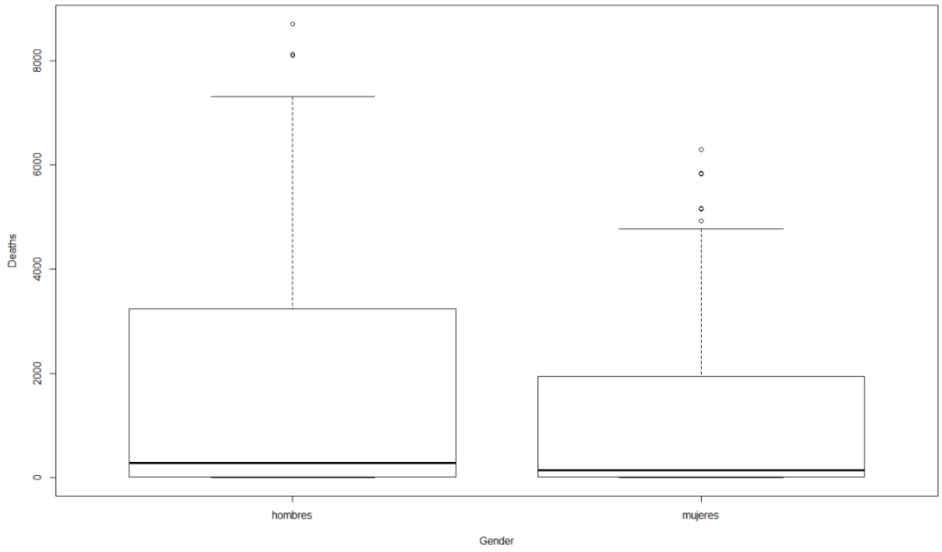

Figure 4: *Boxplot of Deaths against different genders ('Hombres' = Male, 'Mujeres' = Female)*

```
      Augmented Dickey-Fuller Test
data:  NatAge$Deaths
Dickey-Fuller = -2.3048, Lag order = 5, p-value = 0.4491
alternative hypothesis: stationary
```

Figure 5: *adf test for Deaths*

```
       Augmented Dickey-Fuller Test
data:  NatAge$ConfirmedCases
Dickey-Fuller = 0.079963, Lag order = 5, p-value = 0.99
alternative hypothesis: stationary
```

Figure 6: *adf test for Confirmed Cases*

```
       Augmented Dickey-Fuller Test
data:  NatAge$Hospitalized
Dickey-Fuller = -1.5325, Lag order = 5, p-value = 0.7714
alternative hypothesis: stationary
```

Figure 7: *adf test for Hospitalized*

```
       Augmented Dickey-Fuller Test
data:  NatAge$ICU
Dickey-Fuller = -1.9065, Lag order = 5, p-value = 0.6153
alternative hypothesis: stationary
```

Figure 8: *adf test for ICU*

```
Call:
lm(formula = Deaths ~ ConfirmedCases + Hospitalized + ICU + Days +
    relevel(factor(AgeRange), "Young") + factor(Gender), data = NatAge)
Residuals:
     Min       1Q   Median       3Q      Max
-1157.68  -267.25    40.72   248.20   997.13

Coefficients:
                                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 3.631e+02  2.017e+02   1.801  0.07366 .
ConfirmedCases                             -1.232e-02  1.165e-02  -1.057  0.29188
Hospitalized                                3.744e-01  3.422e-02  10.940  < 2e-16 ***
ICU                                        -3.151e-01  1.845e-01  -1.708  0.08959 .
Days                                       -2.310e+01  7.132e+00  -3.239  0.00146 **
relevel(factor(AgeRange), "Young")Adult    -2.124e+03  1.836e+02 -11.566  < 2e-16 ***
relevel(factor(AgeRange), "Young")Elderly -1.665e+03  2.744e+02  -6.069 9.01e-09 ***
factor(Gender)mujeres                       2.875e+02  1.045e+02   2.750  0.00665 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1\
```

```
Residual standard error: 437.2 on 160 degrees of freedom
Multiple R-squared:  0.9656, Adjusted R-squared:  0.9641
F-statistic: 641.5 on 7 and 160 DF,  p-value: < 2.2e-16
```

Figure 9: *Summary of linear model with all variables*

	GVIF	Df	GVIF^(1/(2*Df))
ConfirmedCases	24.344631	1	4.934028
Hospitalized	73.028987	1	8.545700
ICU	20.976391	1	4.579999
Days	2.917632	1	1.708108
relevel(factor(AgeRange), "Young")	47.397839	2	2.623854
factor(Gender)	2.401525	1	1.549686

Figure 10: *Variance inflation factors for linear model with all variables*

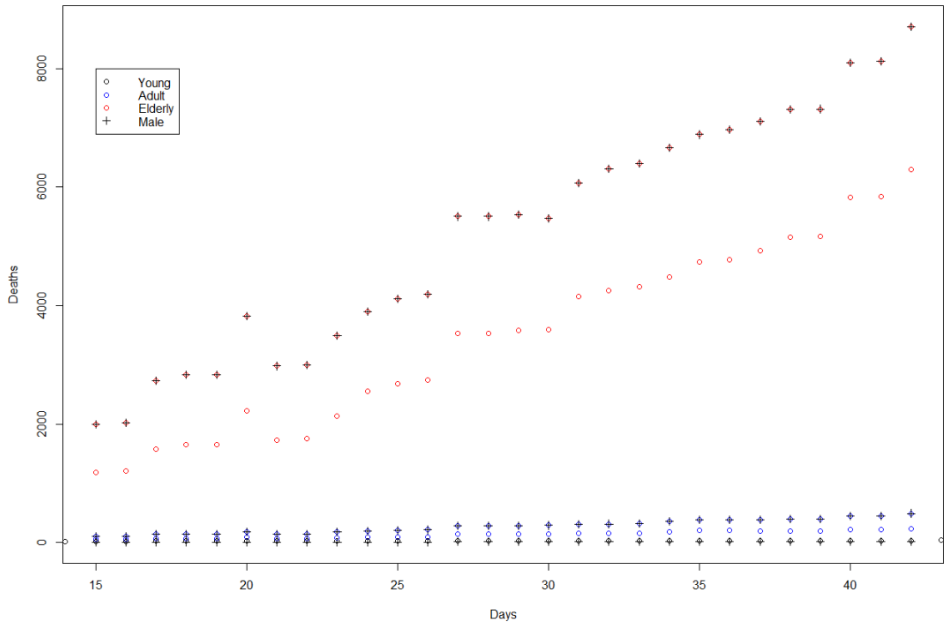

Figure 11: *Plot of deaths against time, with encoded of Age Range and Gender*

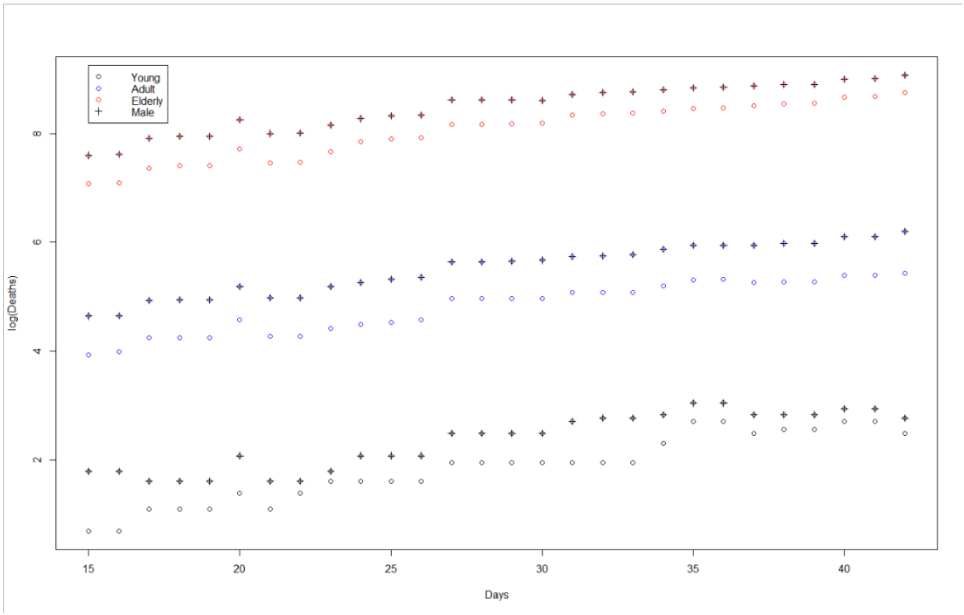

Figure 12: *Plot of natural logarithm of deaths against time, with encoded age range and gender*

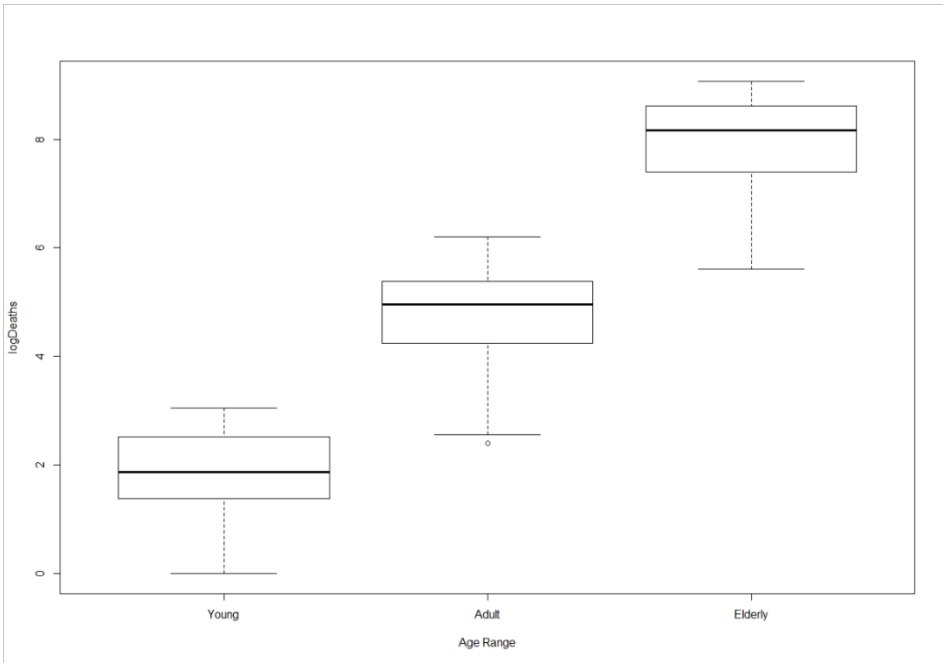

Figure 13: *Boxplot of natural logaithm of deaths and age range*

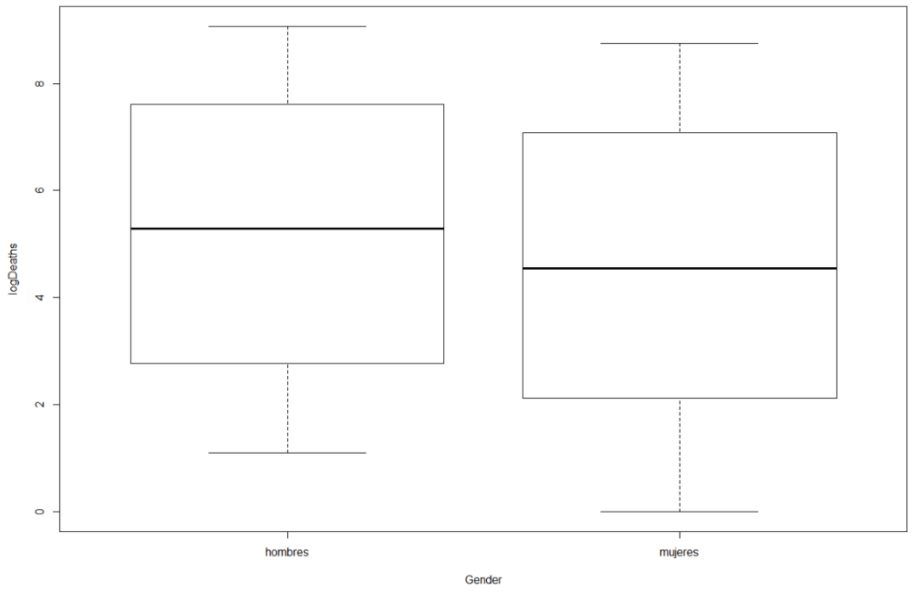

Figure 14: *Boxplot of natural logarithm of deaths and gender*

	Days	(AgeRange)Elderly	(AgeRange)Adult	(Gender)mujeres	Rsq	Adj Rsq	Cp
1		*			0.3728	0.6726	5280.151
2		*	*		0.5258	0.8818	1763.7796
3	*	*	*		0.623	0.9743	219.5152
4	*	*	*	*	0.7075	0.9872	5

Figure 15: *Subsets regression results*

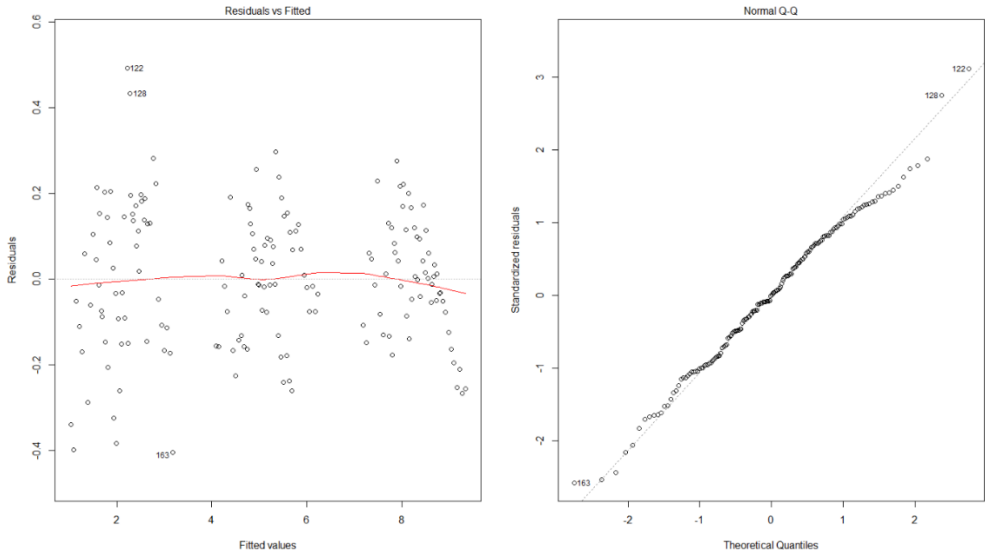

Figure 16: *Residual plots of additive model of natural logarithm of deaths against time, age range, and gender*

```
       Shapiro-Wilk normality test

data:  natage.lm$residuals

W = 0.99217, p-value = 0.4962
```

Figure 17: *Shapiro-Wilkins test on additive model of natural logarithm of deaths against time, age range, and gender's additive model*

```
Call:
lm(formula = logDeaths ~ Days * relevel(factor(AgeRange), "Young") *
    factor(Gender), data = NatAge)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37446 -0.09467  0.00250  0.08950  0.37766


Coefficients:
                                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                                      0.691399   0.097672   7.079 4.65e-11 ***
Days                                             0.058468   0.003297  17.732  < 2e-16 ***
relevel(factor(AgeRange), "Young")Adult          3.231672   0.138129  23.396  < 2e-16 ***
relevel(factor(AgeRange), "Young")Elderly        6.316493   0.138129  45.729  < 2e-16 ***
factor(Gender)mujeres                           -0.938707   0.138129  -6.796 2.15e-10 ***
Days:relevel(factor(AgeRange), "Young")Adult    -0.002760   0.004663  -0.592  0.55474
Days:relevel(factor(AgeRange), "Young")Elderly  -0.007019   0.004663  -1.505  0.13430
Days:factor(Gender)mujeres                       0.015181   0.004663   3.256  0.00139 **
relevel(factor(AgeRange), "Young")
Adult:factor(Gender)mujeres                      0.241365   0.195344   1.236  0.21847
relevel(factor(AgeRange), "Young")
Elderly:factor(Gender)mujeres                    0.253898   0.195344   1.300  0.19560
Days:relevel(factor(AgeRange),
"Young")Adult:factor(Gender)mujeres             -0.015279   0.006594  -2.317  0.02180 *
Days:relevel(factor(AgeRange), "Young")
Elderly:factor(Gender)mujeres                   -0.006399   0.006594  -0.970  0.33339
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.1409 on 156 degrees of freedom
Multiple R-squared:  0.9972, Adjusted R-squared:  0.997
F-statistic:  5089 on 11 and 156 DF,  p-value: < 2.2e-16
```

Figure 18: *Plot of natural logarithm of deaths against time, after removing up to two weeks after lockdown*

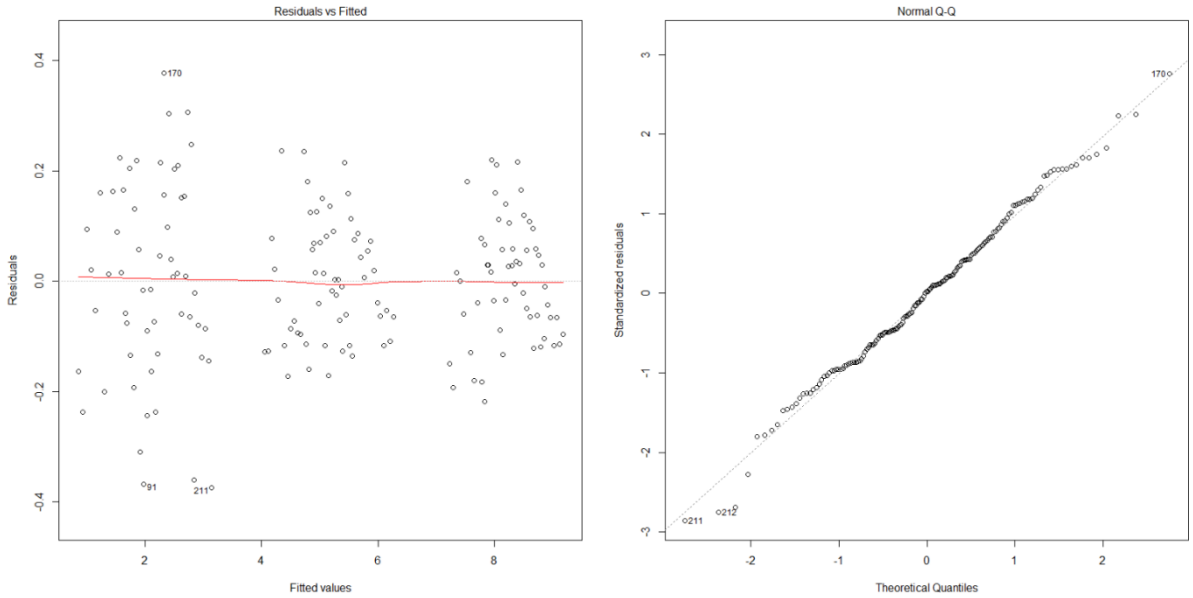

Figure 19: *Residual plots of interaction model of natural logarithm of deaths against time, age range, and gender*

```
      Shapiro-Wilk normality test

data:  natage.stepfwd$residuals

W = 0.99264, p-value = 0.5527
```

Figure 20: *Shapiro-Wilkins test of residuals from interaction model of natural logarithm of deaths against time, age range, and gender after removing outliers.*

Variable	**Effect on deaths for unit increase**
Days since lockdown overall	Multiply deaths by 1.0602
Days since lockdown for elderly and young females	Multiply deaths by 1.0764
AgeRange: Adult (as compared to Young)	Multiply deaths by 25.3220
AgeRange: Elderly (as compared to Young	Multiply deaths by 553.6280
Gender: Female (as compated to Male)	Multiply deaths by 0.3911

Figure 21: *Table of results for effect on deahs per unit increase of each of significant variables*

Appendix B:

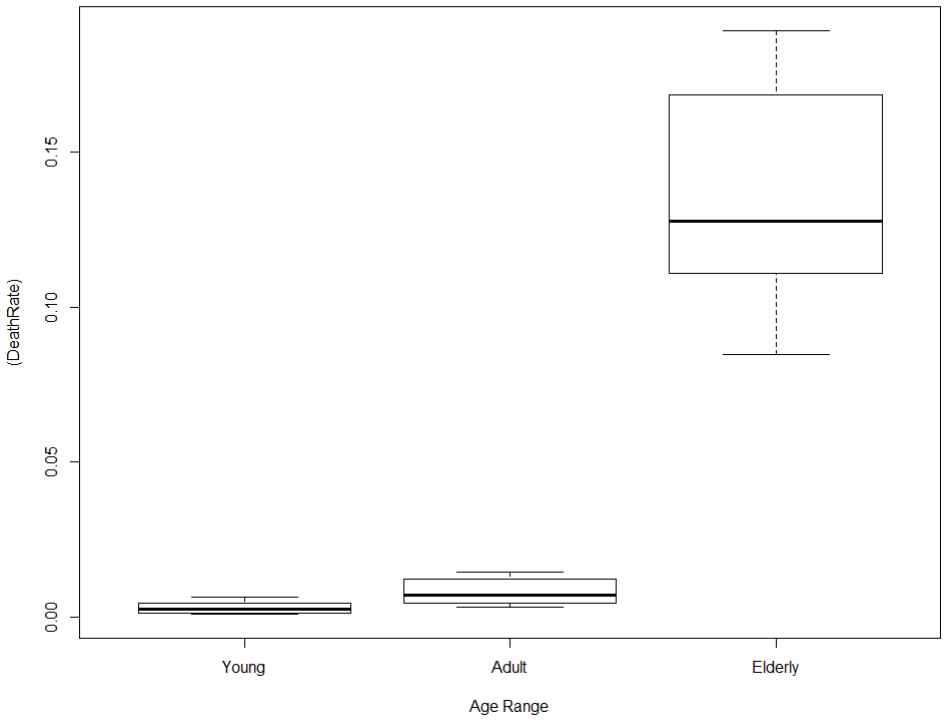

Figure 1: *Boxplot of Death Rate against different Age Ranges*

Figure 2: *Boxplot of Death Rate against different genders ('Hombres' = Male, 'Mujeres' = Female)*

```
           Adult    Elderly       Young
hombres 0.011734936 0.1619516 0.004228188
mujeres 0.004530951 0.1106098 0.001535204
```

Figure 3: *Table of Means of death rate against age and gender*

```
                 Df Sum Sq Mean Sq F value    Pr(>F)
factor(AgeRange)    2 0.6392  0.3196 1375.90  < 2e-16 ***
factor(Gender)      1 0.0175  0.0175   75.34 3.78e-15 ***
Residuals         164 0.0381  0.0002
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4: *Summary of Two-Way-Anova Model of Death Rates against Age Range and Gender*

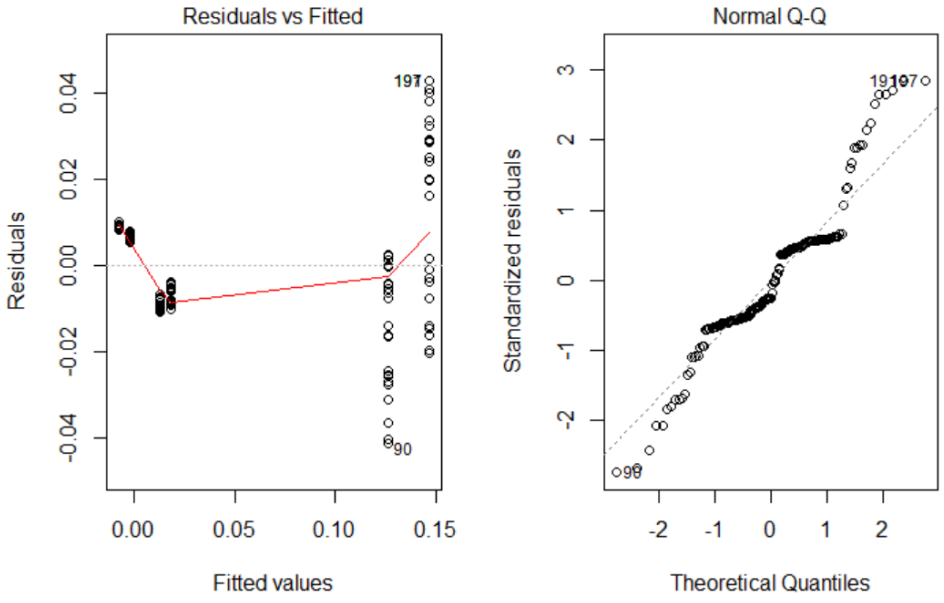

Figure 5: *Residual plots of Two-Way Anova Model of Death Rates against Age Range and Gender*

Figure 6: *Boxplot of natural logarithm of Death Rate against different genders ('Hombres' = Male, 'Mujeres' = Female)*

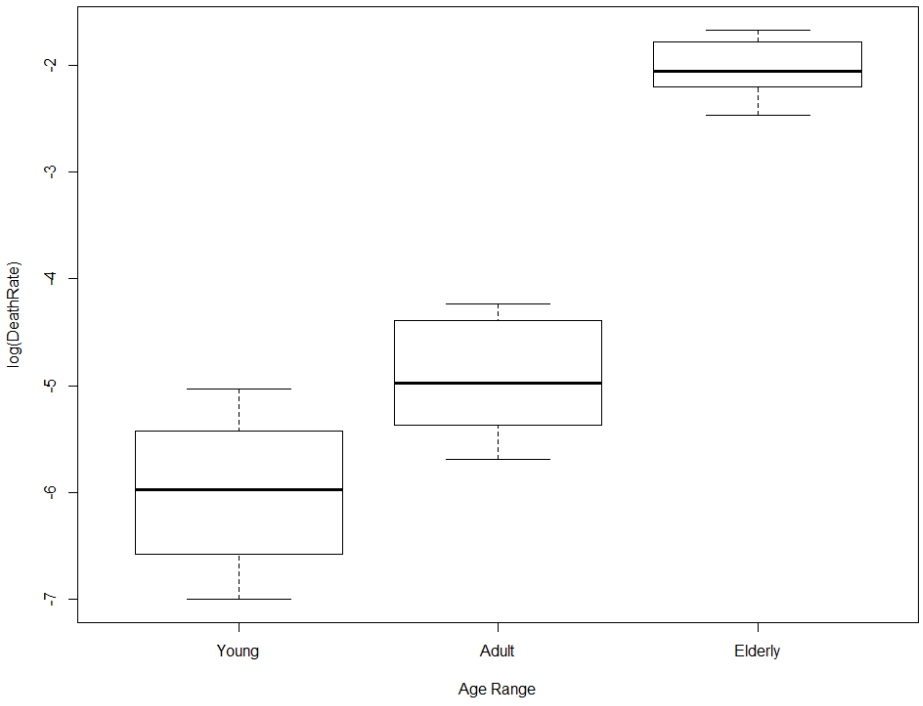

Figure 7: *Boxplot of natural logarithm of Death Rate against different Age Ranges*

```
          Adult    Elderly     Young
hombres -4.461592 -1.829362 -5.503354
mujeres -5.409733 -2.209606 -6.522572
```

Figure 8: *Table of Means of natural logarithm of death rates on age and gender*

```
                  Df Sum Sq Mean Sq F value Pr(>F)
factor(AgeRange)   2  478.1  239.05  3705.6 <2e-16 ***
factor(Gender)     1   25.7   25.72   398.7 <2e-16 ***
Residuals        164   10.6    0.06
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 9*: Summary of Two-Way Anova Model of the natural logarithm of Death Rates against age and gender*

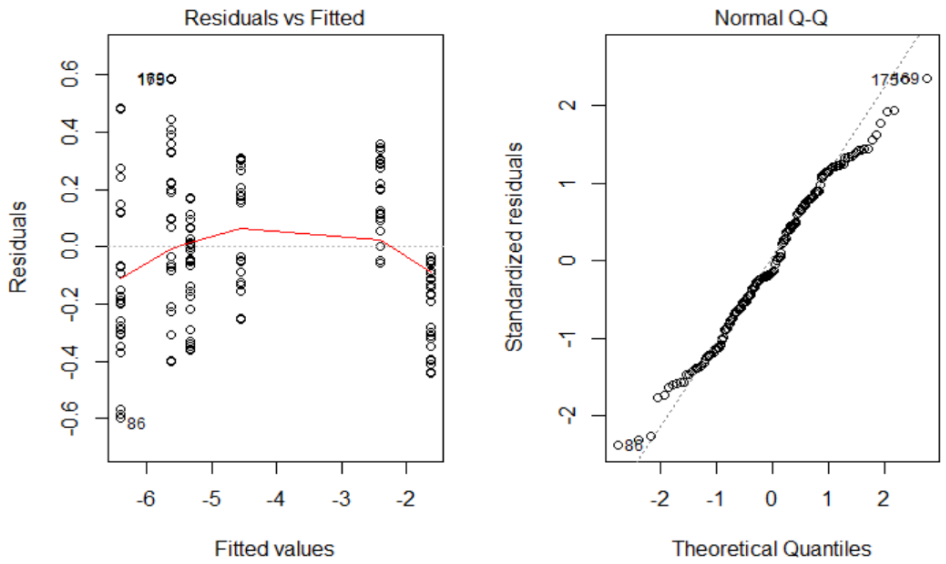

Figure 10: *Residual plots of Two-Way Anova Model of the natural logarithm of Death Rates against Age Range and Gender*

```
Shapiro-Wilk normality test
data:  natage.aov$residuals
W = 0.98502, p-value = 0.06828
```

Figure 11: *Shapiro-Wilkins test of residuals from Two-Way Anova Model of natural logarithm of Death Rate against age range, and gender.*

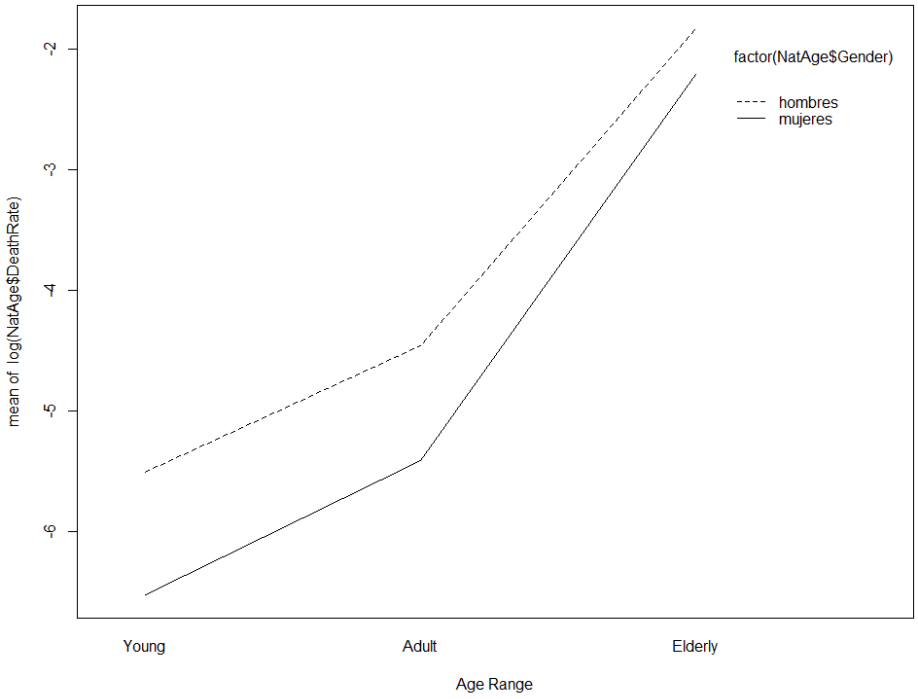

Figure 12: *Interaction Plot for Age Range and Gender*

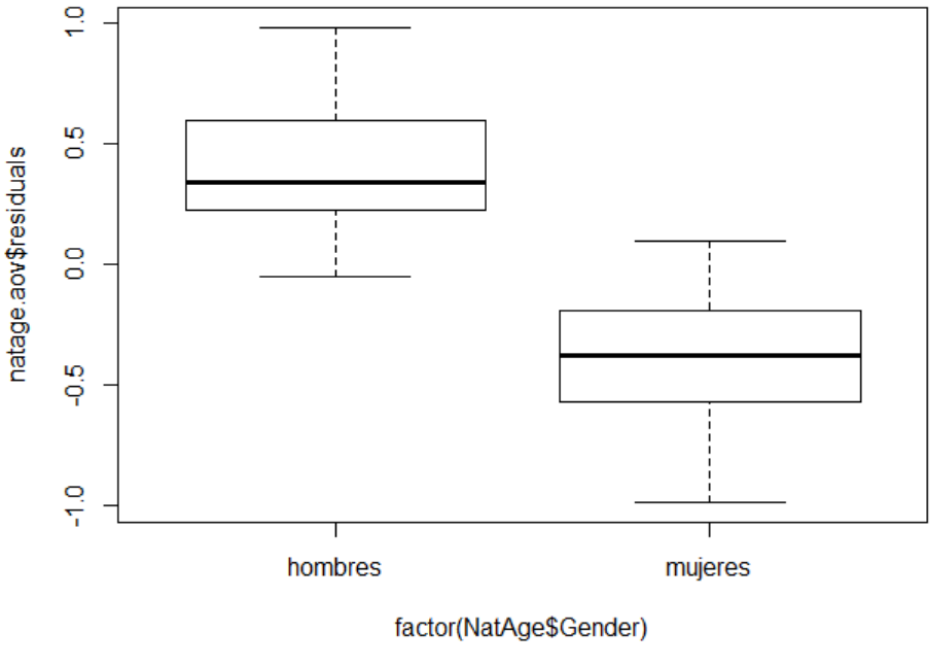

Figure 13: *Partial Boxplot of natural logarithm of Death Rate against Gender*

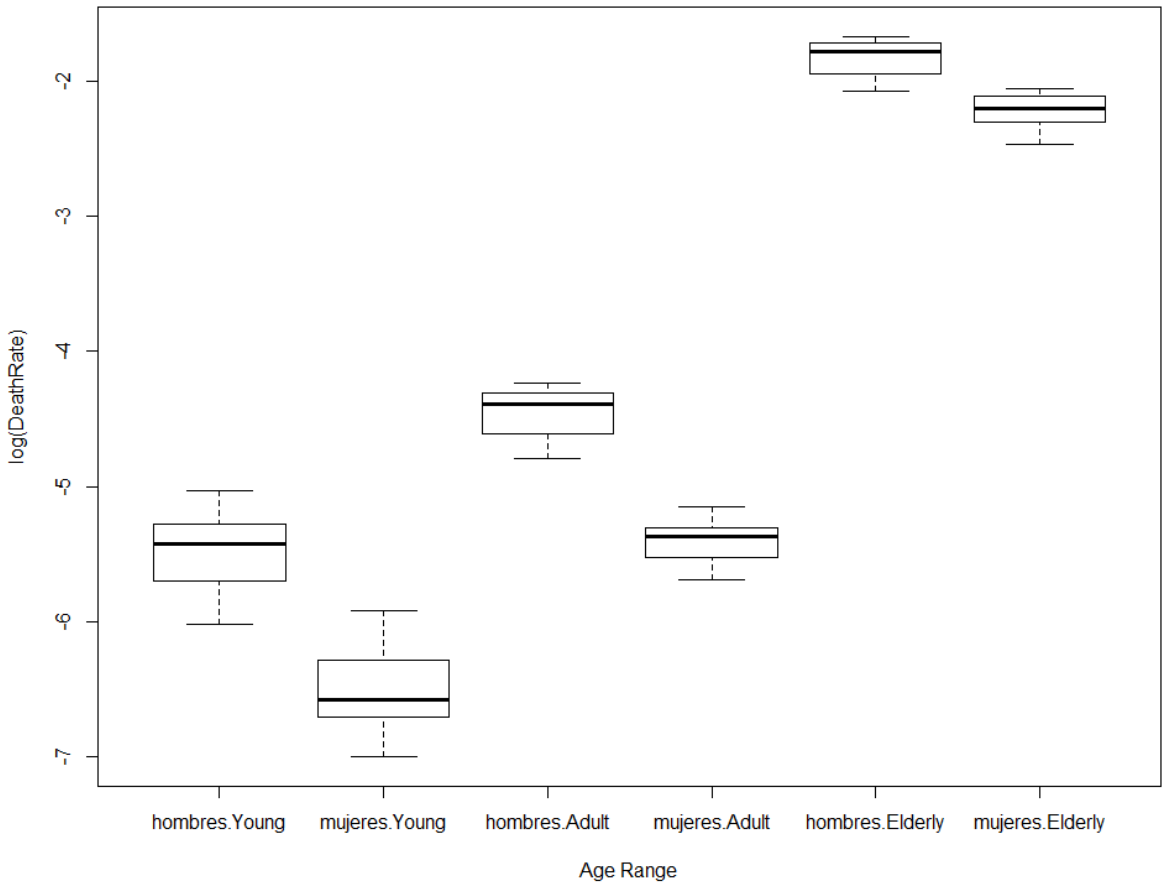

***Figure 15**: Boxplot of natural logarithm of Death Rate against different Age Ranges and Gender.*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
factor(AgeRange)	2	478.1	239.05	5419.50	< 2e-16	***
factor(Gender)	1	25.7	25.72	583.08	< 2e-16	***
factor(AgeRange):factor(Gender)	2	3.4	1.72	38.92	1.57e-14	***
Residuals	162	7.1	0.04			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

***Figure 16:** Summary of Two-Way Anova Model with interaction between Age Range and Gender.*

```
                 Df Sum Sq Mean Sq F value Pr(>F)
factor(AgeRange)  2  200.8  100.39    2256 <2e-16 ***
Residuals        81    3.6    0.04
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 17: *One-Way ANOVA Table for Age Effect when Gender = Male*

```
  Tukey multiple comparisons of means
    95% family-wise confidence level


Fit: aov(formula = log(DeathRate) ~ factor(AgeRange), data = NatAgeMen)


$`factor(AgeRange)`
                    diff       lwr        upr p adj
Elderly-Adult  2.632230  2.497623  2.7668366     0
Young-Adult   -1.041763 -1.176369 -0.9071558     0
Young-Elderly -3.673992 -3.808599 -3.5393857     0
```

Figure 18: *Tukey Multiple Comparisons of Means on Age when Gender = Male*

```
                  Df Sum Sq Mean Sq F value Pr(>F)
factor(AgeRange)  2 280.76  140.38    3211 <2e-16 ***
Residuals        81   3.54    0.04
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 19: *One-Way ANOVA Table for Age Effect when Gender = Female*

```
    Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = log(DeathRate) ~ factor(AgeRange), data = NatAGeWomen)

$`factor(AgeRange)`
                   diff       lwr        upr p adj
Elderly-Adult  3.200127  3.066707  3.3335480     0
Young-Adult   -1.112839 -1.246259 -0.9794181     0
Young-Elderly -4.312966 -4.446387 -4.1795455     0
```

Figure 20: *Tukey Multiple Comparisons of Means on Age when Gender = Female*

```
Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = log(DeathRate) ~ factor(AgeRange) + factor(Gender) +
factor(AgeRange) * factor(Gender), data = NatAge)

$`factor(AgeRange)`
                   diff       lwr        upr p adj
Elderly-Adult  2.916179  2.822293  3.0100647     0
Young-Adult   -1.077301 -1.171187 -0.9834145     0
Young-Elderly -3.993479 -4.087365 -3.8995932     0


$`factor(Gender)`
                      diff        lwr        upr p adj
mujeres-hombres -0.7825343 -0.8465292 -0.7185394     0


$`factor(AgeRange):factor(Gender)`
                                      diff         lwr        upr     p adj
Elderly:hombres-Adult:hombres    2.63222992  2.47034120  2.7941186 0.0000000
Young:hombres-Adult:hombres     -1.04176250 -1.20365122 -0.8798738 0.0000000
Adult:mujeres-Adult:hombres     -0.94814143 -1.11003014 -0.7862527 0.0000000
Elderly:mujeres-Adult:hombres    2.25198599  2.09009727  2.4138747 0.0000000
```

```
Young:mujeres-Adult:hombres     -2.06098002 -2.22286874 -1.8990913 0.0000000
Young:hombres-Elderly:hombres   -3.67399242 -3.83588114 -3.5121037 0.0000000
Adult:mujeres-Elderly:hombres   -3.58037134 -3.74226006 -3.4184826 0.0000000
Elderly:mujeres-Elderly:hombres -0.38024393 -0.54213265 -0.2183552 0.0000000
Young:mujeres-Elderly:hombres   -4.69320994 -4.85509866 -4.5313212 0.0000000
Adult:mujeres-Young:hombres      0.09362108 -0.06826764  0.2555098 0.5549557
Elderly:mujeres-Young:hombres    3.29374850  3.13185978  3.4556372 0.0000000
Young:mujeres-Young:hombres     -1.01921752 -1.18110624 -0.8573288 0.0000000
Elderly:mujeres-Adult:mujeres    3.20012742  3.03823870  3.3620161 0.0000000
Young:mujeres-Adult:mujeres     -1.11283860 -1.27472731 -0.9509499 0.0000000
Young:mujeres-Elderly:mujeres   -4.31296601 -4.47485473 -4.1510773 0.0000000
```

Figure 21: *Tukey Multiple Comparisons of Means on both Age Range and Gender.*

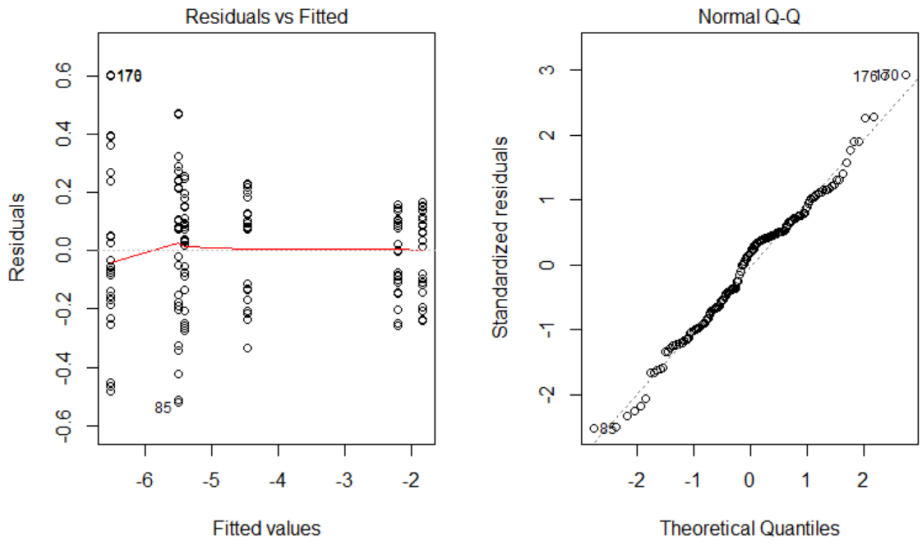

Figure 22: *Residual plots of Two-Way Anova Model of the natural logarithm of Death Rates against Age Range and Gender and interaction between Age Range and Gender.*

```
Shapiro-Wilk normality test


data:  natage.aov$residuals
W = 0.98409, p-value = 0.05193
```

Figure 23: *Shapiro-Wilkins test of residuals from Two-Way Anova Model of natural logarithm of Death Rate against age range, and gender and interaction between Age Range and Gender.*

Bibliography:

Forte, Fernando. "Topic: Coronavirus (COVID-19) Outbreak in Spain." *Www.statista.com*, 20 Apr. 2020, www.statista.com/topics/6118/coronavirus-covid-19-outbreak-in-spain/.

Merelo, J J. *Situacion De COVID-19 En España*. Ministerio De Sanidad, 2020, cnecovid.isciii.es/covid19/. https://raw.githubusercontent.com/datadista/datasets/master/COVID%2019/nacional_covid19_rango_edad.csv

Villarreal, Antonio, et al. "Razones (Fundamentadas) Por Las Que España No Está Haciendo Esos 20.000 Test PCR Al Día." *El Confidencial*, 3 Apr. 2020, www.elconfidencial.com/tecnologia/ciencia/2020-04-03/covid19-test-pcr-coronavirus-espana_2531844/.

WHO, Int. *Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19)* . Feb. 2020, www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf.