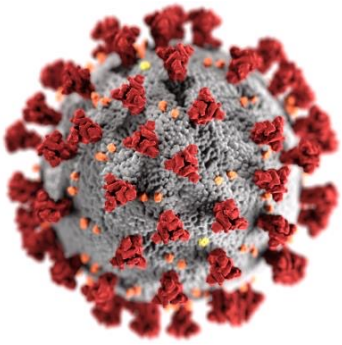
Pandemic Patterns: Analyzing COVID-19 Trends in Europe

Isha Singh, Pablo Gonzalez, Quin Yuter, Rohit Mishra



(image taken from CDC)

Definitions:

- "Incidence rate" is equal to new daily cases per 100K individuals.
- "Fatality rate" is equal to new daily deaths per 100K individuals.

INTRODUCTION

Covid-19 was a deadly virus that permeated across the world. Not only did major nations struggle to adapt to the drastic changes and restrictions brought upon by the virus, experts in epidemiology had several challenges in tracing the virus and its impact on citizens. Scientists were able to capture basic data on infected individuals in the European Union and European Economic Areas. We examine the various points that are collected in this dataset using analysis procedures, including descriptive and inferential statistics, and more sophisticated techniques such as linear regression and modeling, to better understand the impact of the virus on European countries.

LITERATURE REVIEW

At a global level, there were unprecedented effects of covid-19 on the global economy. In an article, titled "Observed impacts of the COVID-19 pandemic on global trade" and written by Jasper Verschuur, Elco E. Koks and Jim W. Hall, we learn that supply chain disruptions were prevalent, and statistical models were used to understand the predicted changes that resulted from these disruptions during the Pandemic. This informs our analysis, since we are also using regression analysis to understand the Covid-19 outbreak in European countries from an economic lens (Verschuur, Koks, & Hall, 2021).

On top of looking at Covid-19 from an economic perspective, an article titled "Analysis and prediction of COVID-19 trajectory: A machine learning approach" highlights the use of models such Random Forests, Nonlinear Regressions, and Decision Tree based regressions to run an in depth regression analysis. These methods can be used to predict cases well in advance, which gives key insights into the trends of Covid cases. Machine learning algorithms can be useful in suggesting strategies to policy makers. This can tell us how we can mold our future research and what we can do with the conclusions we reach (Majhi et al., 2020).

Similarly, hypothesis testing plays a vital role in evaluating the efficacy of COVID-19 vaccines by analyzing multiple endpoints, such as SARS-CoV-2 infection, symptomatic COVID-19, and severe cases. In an article

titled “Evaluating the Efficacy of Coronavirus Disease 2019 Vaccines” written Dan-Yu Lin, Donglin Zeng, Devan V Mehrotra, Lawrence Corey, and Peter B Gilbert, they incorporated robust statistical methods and simulations, and were able to assess vaccine efficacy, identify trends, and guide both policy decisions and future research. Together, Machine Learning and Hypothesis Testing approaches highlight the synergy between economic and health-focused analyses in understanding and responding to the pandemic (Lin et al., 2020).

Additionally, the article “Correlation Between Mask Compliance and COVID-19 Outcomes in Europe” by Beny Spira studied the relationship between mask compliance and COVID-19 cases and deaths in 35 European countries. Mask compliance averaged 60.9% across Europe, but the study found only weak links with cases and moderate links with deaths. This suggests that factors such as the timing of mandates and healthcare capacity played a bigger role than masks alone, highlighting the need to consider multiple factors when analyzing the pandemic (Spira, 2022).

These bodies of literature reviewed underscores the multifaceted impact of COVID-19, particularly its disruption to global economic systems and public health. Studies like Jasper Verschuur et al.’s examination of supply chain disruptions and statistical modeling emphasize the economic ramifications of the pandemic and the value of regression analysis in understanding these effects. From a public health perspective, machine learning approaches, as highlighted in “Analysis and prediction of COVID-19 trajectory: A machine learning approach,” demonstrate the power of advanced algorithms in predicting case trends and informing policy decisions. Meanwhile, Dan-Yu Lin et al.’s research on vaccine efficacy highlights the critical role of hypothesis testing and robust statistical methods in evaluating health interventions and guiding future strategies.

Further expanding this perspective, Beny Spira’s investigation into mask compliance reveals the complex interplay of public health measures, individual behavior, and systemic factors such as healthcare infrastructure and policy timing. This indicates that single-factor interventions may be insufficient, reinforcing the need for comprehensive, multi-dimensional analyses.

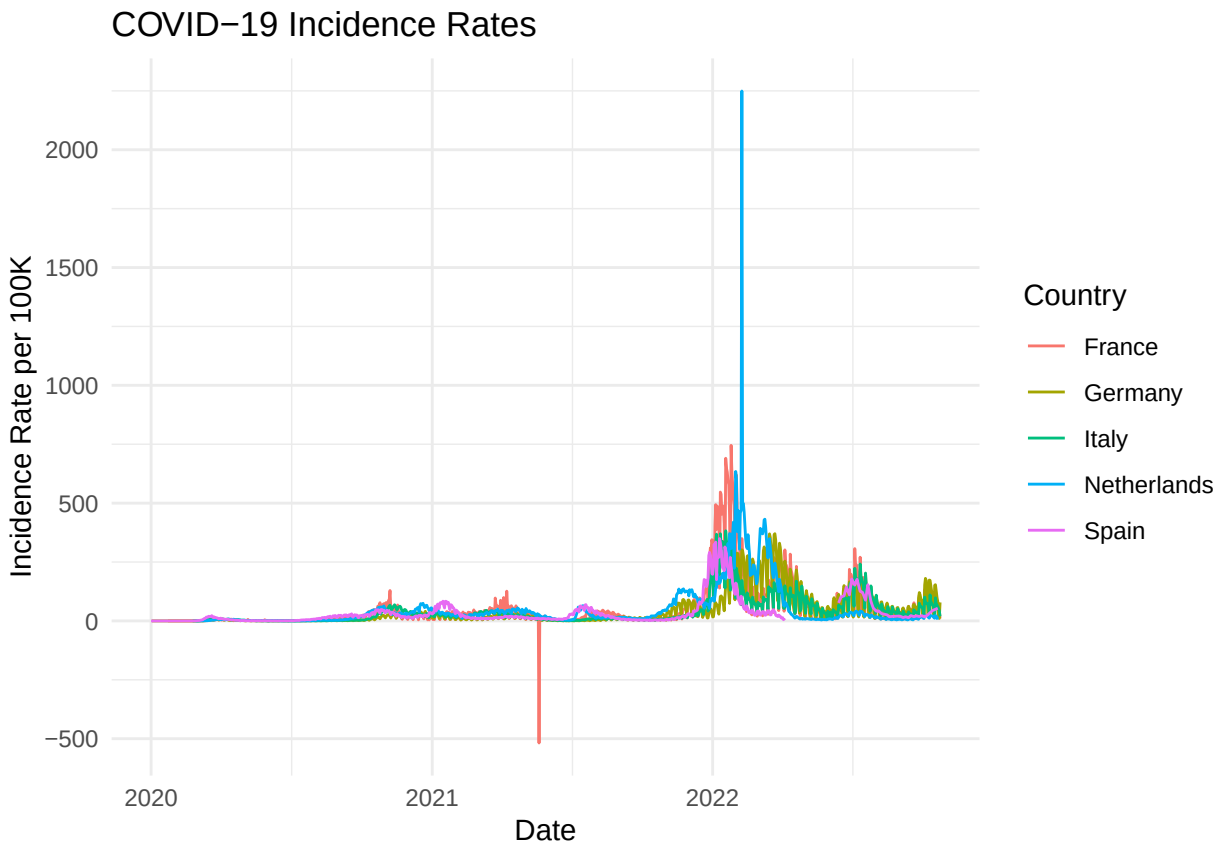
Collectively, these studies illustrate the synergy between economic modeling, machine learning, and hypothesis testing in addressing the challenges posed by COVID-19. They highlight the necessity of integrating diverse methodologies to gain a holistic understanding of the pandemic and to formulate effective strategies for mitigating its impacts in both economic and public health domains.

DESCRIPTIVE STATISTICS

We first want to take a deeper dive into our data through descriptive statistics.

```
##     dateRep                day            month             year
##  Min.   :2020-01-01   Min.   : 1.00   Min.   : 1.000   Min.   :2020
##  1st Qu.:2020-10-17   1st Qu.: 8.00   1st Qu.: 4.000   1st Qu.:2020
##  Median :2021-06-17   Median :16.00   Median : 6.000   Median :2021
##  Mean   :2021-06-17   Mean   :15.68   Mean   : 6.431   Mean   :2021
##  3rd Qu.:2022-02-13   3rd Qu.:23.00   3rd Qu.: 9.000   3rd Qu.:2022
##  Max.   :2022-10-26   Max.   :31.00   Max.   :12.000   Max.   :2022
##
##      cases              deaths          countriesAndTerritories     geoId
##  Min.   :-348846   Min.   : -217.00   Finland  : 1024         FI     : 1024
##  1st Qu.:    111   1st Qu.:    0.00   France   : 1006         FR     : 1006
##  Median :    705   Median :    5.00   Czechia  : 1003         CZ     : 1003
##  Mean   :   6088   Mean   :   40.87   Lithuania:  997         LT     :  997
##  3rd Qu.:   3483   3rd Qu.:   31.00   Germany  :  992         DE     :  992
##  Max.   : 501635   Max.   :13743.00   Sweden   :  982         SE     :  982
##  NA's   :93        NA's   :292        (Other)  :22725         (Other):22725
##  countryterritoryCode  popData2020
##  FIN    : 1024        Min.   :   38747
##  FRA    : 1006        1st Qu.: 2095861
##  CZE    : 1003        Median : 6951482
```

```
##  LTU     :  997        Mean   :15348035
##  DEU     :  992        3rd Qu.:11522440
##  SWE     :  982        Max.   :83166711
##  (Other):22725
```



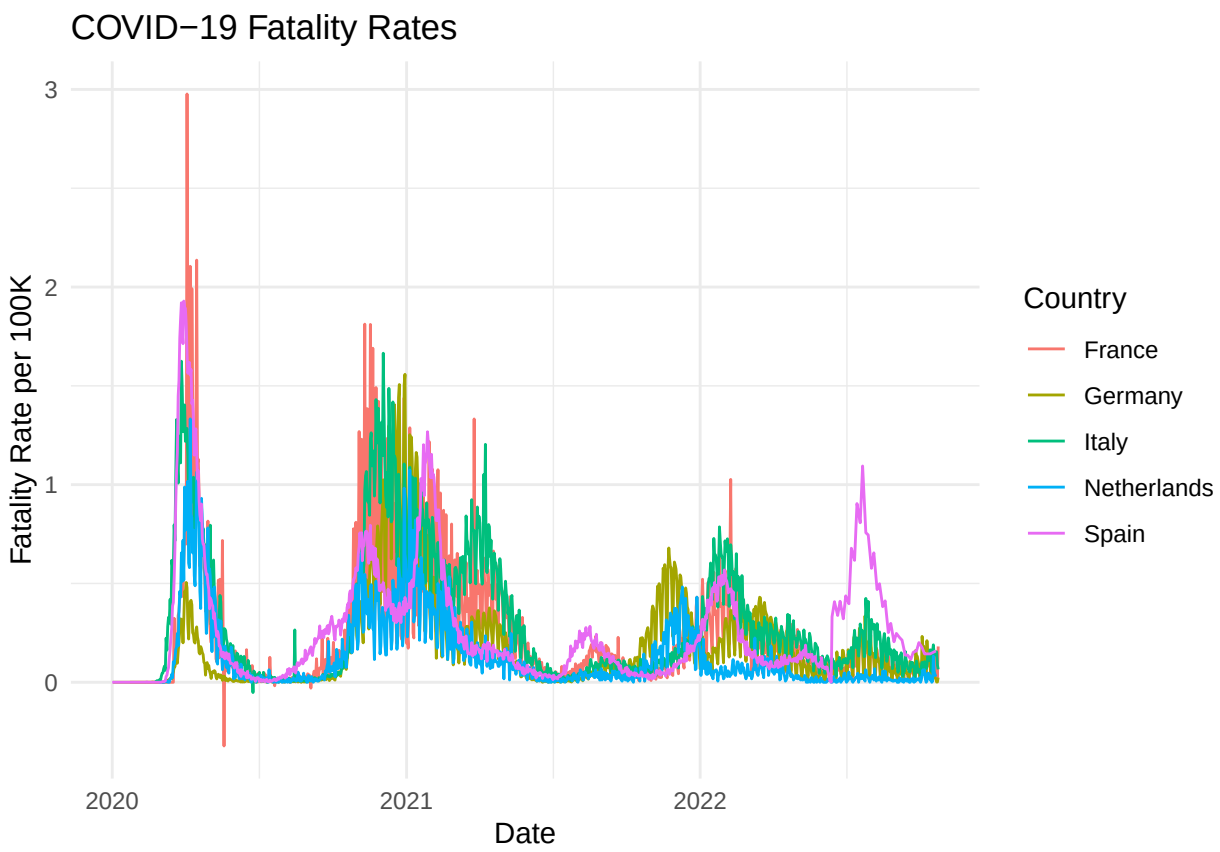
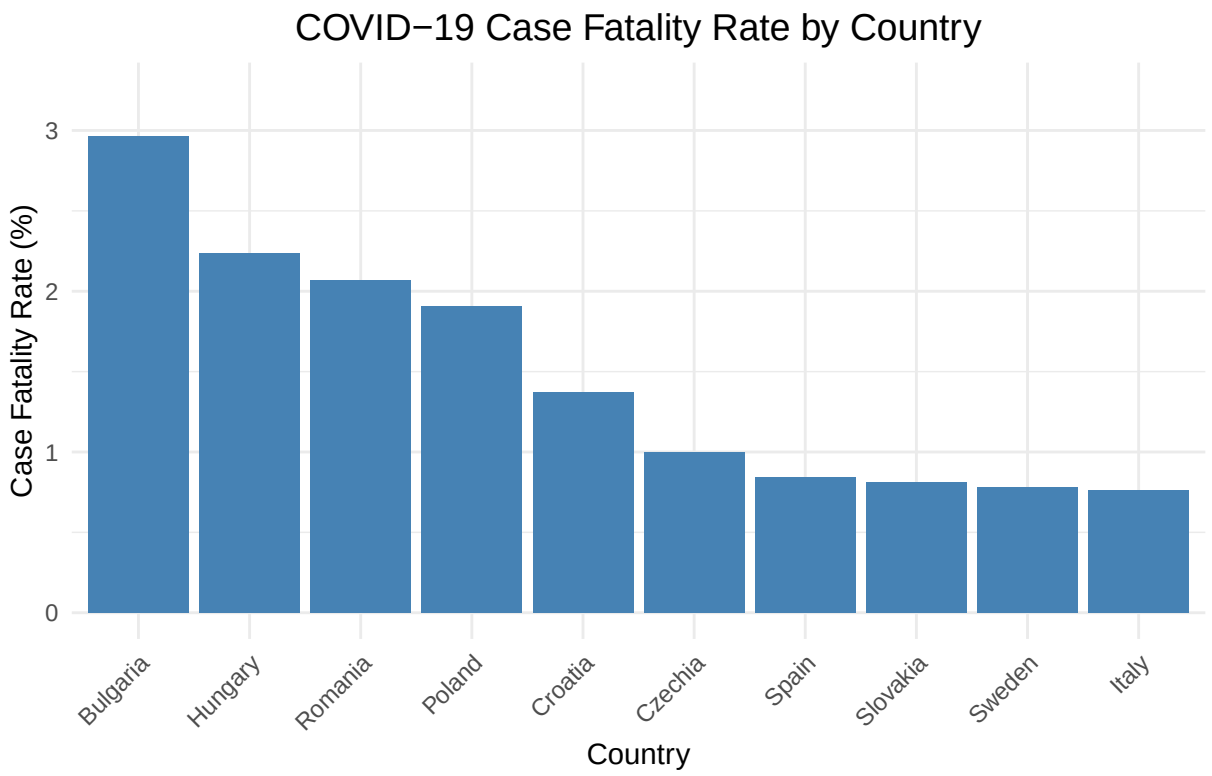


Table 1: Top 10 Countries by Case Fatality Rate (CFR)

	Country	Total Cases	Total Deaths	CFR (%)
Bulgaria	Bulgaria	1,275,481	37,790	2.9628
Hungary	Hungary	2,141,513	47,938	2.2385
Romania	Romania	3,246,580	67,179	2.0692
Poland	Poland	6,189,562	118,050	1.9072
Croatia	Croatia	1,244,692	17,085	1.3726
Czechia	Czechia	4,152,997	41,524	0.9999
Spain	Spain	13,564,823	114,110	0.8412
Slovakia	Slovakia	2,535,554	20,511	0.8089
Sweden	Sweden	2,604,866	20,407	0.7834
Italy	Italy	23,359,680	178,633	0.7647



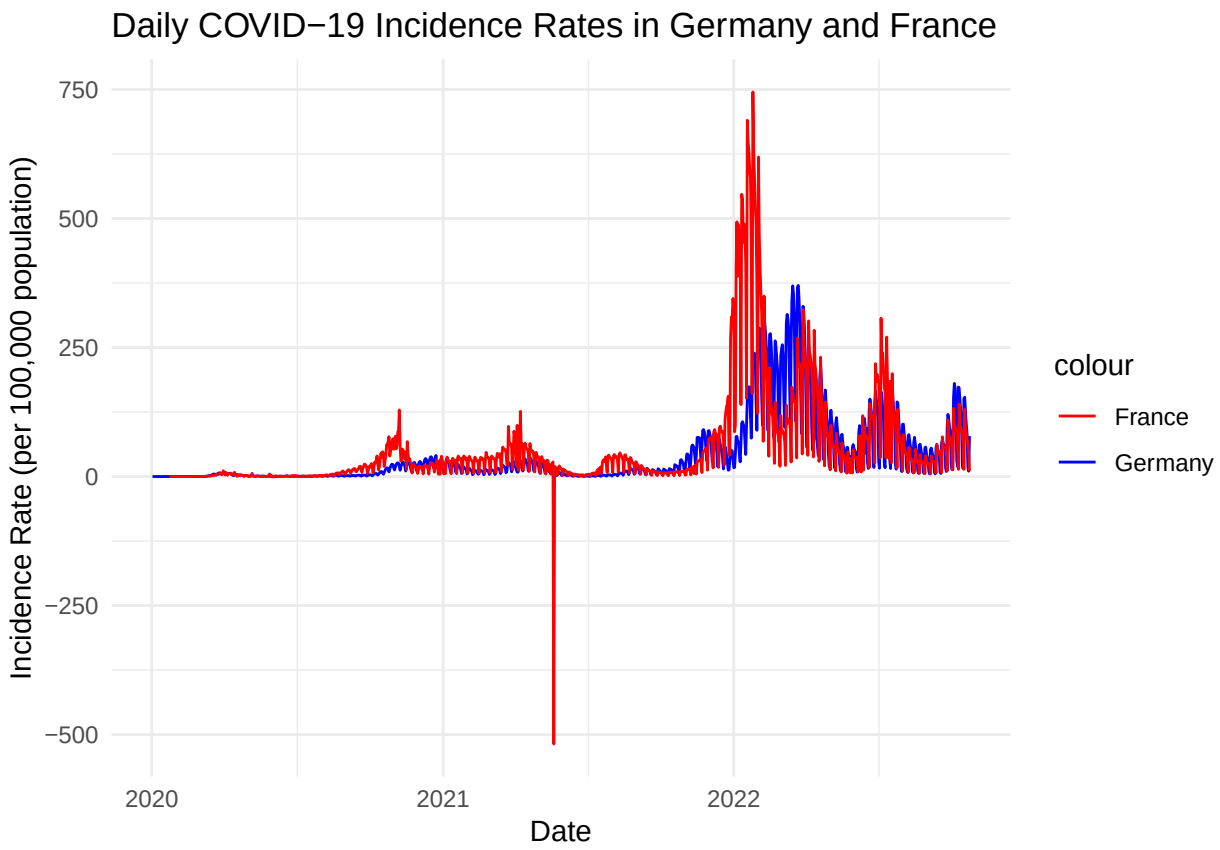
As a baseline, we can examine our data and the sources from which they came. The data is collected by the ECDC Epidemic Intelligence team, and measures incidence and fatality rates. However, this is only a cursory view of the virus' impact. Additionally, there are many testing procedures that are not explored in this dataset, which means there is potential for more informative results given more robust testing infrastructure. As we are viewing results from an aggregate analysis of a per 100k individuals testing process, we can derive that there is a level of granularity that we do not have access to.

With these caveats, we examine the following dataset to gain more insight into the top countries with the highest Covid-19 rates. In doing a brief examination of summary statistics, we notice that there are several outliers.

To get a clearer picture of the different effects on countries that covid-19 had, we take a subset of the countries. Specifically, we choose the top 5 countries by incidence and fatality rate. In examining the incidence rates, we see that the Netherlands were the most affected out of any of the other countries. This spike happened around 2022. We can determine that around this time, there were many new cases appearing, potentially due to increased travel during that time. As far as Fatality rates are concerned, Poland had the highest number of fatalities across the span of 2021 to 2022. There is a bimodal pattern for this specific country. France has a major spike at the beginning, but subsided before 2021. This is the case for the other countries in this set as well. Finally, we look at the Case fatality rate, which showcases that Hungary had the highest deaths per case rate across the different nations. The majority of the countries seem to have around a 0.7% CF rate, which means that less than 1% of the affected populations see death. This is a relatively low proportion, but it is important to note that there are still large numbers of deaths irrespective of the proportion.

INFERENTIAL STATISTICS

The two countries we are choosing to focus on are France and Germany.



Our null hypothesis is that there is no difference in the mean daily incidence rates between Germany and France. Making our alternative hypothesis that there is difference in the mean daily incidence rates.

In order to test our null hypothesis, we will we use a t-test, this test is appropriate as we are comparing the means of two independent groups, Germany and France. Some distributional assumptions are that the data is normally distributed and that the variance of both groups are equal. Moreover, in our test we will use an alpha level of 0.05.

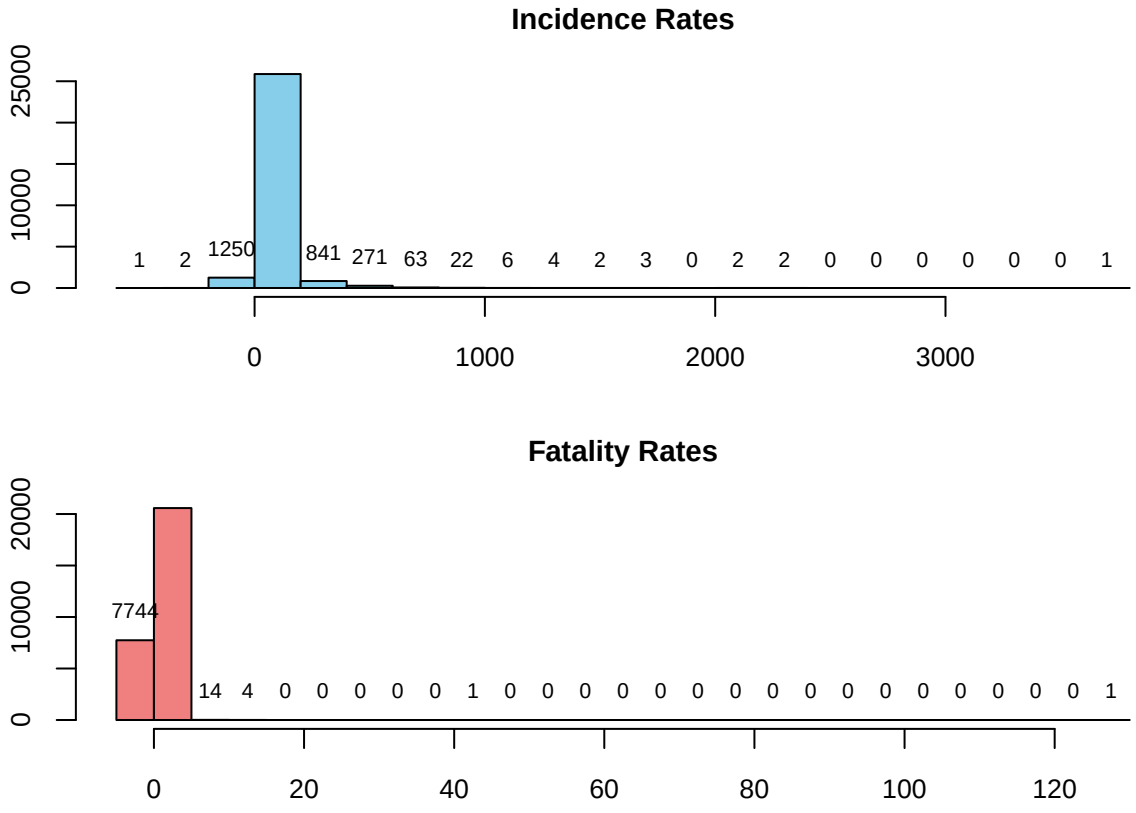
Table 2: Detailed T-Test Results for Germany vs France

Statistic	Value
Method	Welch Two Sample t-test
Data Used	Germany vs France: Incidence Rates
Alternative Hypothesis	two.sided
t-value	-2.966
Degrees of Freedom	1782.09680446275
P-value	0.003
95% Confidence Interval	(-18.754, -3.825)
Mean of Group 1 (Germany)	42.772
Mean of Group 2 (France)	54.061

As we observe our results, we obtain a p-value of 0.003055, which is less than our chosen alpha level of 0.05. This indicates that we have a statistically significant difference between the mean daily incidence rates between Germany and France. Moreover, we have a 95% confidence interval is (-18.75366,-3.82468, we observe it does not include 0, which further supports our previous conclusion of significant difference between the two means.

With an alpha level of 0.05, we reject the null hypothesis. There is a significant difference in the mean daily incidence rates between Germany and France. Specifically, France has a higher mean daily incidence rate compared to Germany.

CORRELATION



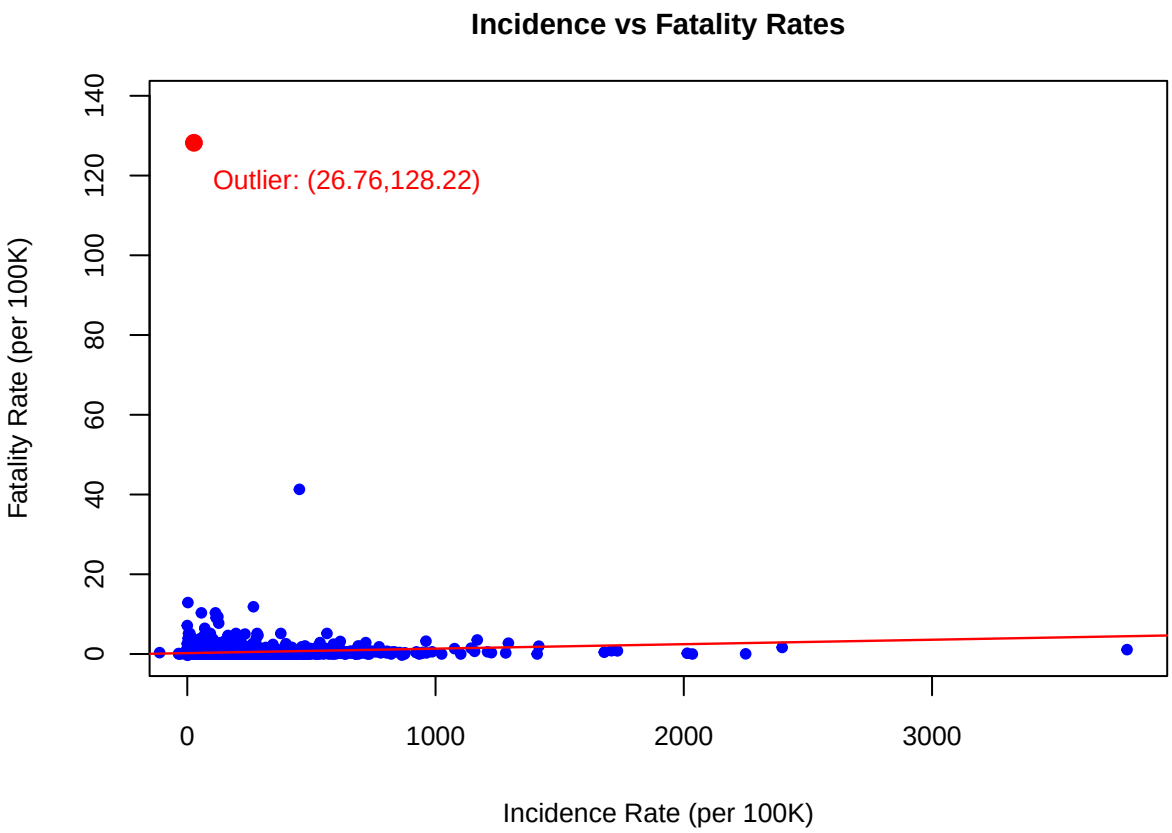


Table 3: Detailed Correlation Test Results

Statistic	Value
Method	Pearson's product-moment correlation
Data Used	Incidence Rate vs Fatality Rate
Alternative Hypothesis	two.sided
t-value	19.085
Degrees of Freedom	28342
P-value	<2e-16
95% Confidence Interval	(0.101, 0.124)
Sample Correlation	0.113

The correlation study looks at how COVID-19 incidence rates (new cases per 100,000 persons) compare to fatality rates (new deaths per 100,000). The Pearson correlation coefficient of 0.113 indicates a very weak relationship between the two. This means that while rates of incidence grow, fatality rates climb relatively not much. Although this relationship is statistically significant (p-value $<$ 2e-16), the small magnitude of the correlation suggests that other factors may have a higher influence. The confidence interval, which ranges from 0.101 to 0.124, emphasizes how insignificant this association is. The histograms help illustrate how these rates are spread across countries. In terms of incidence rates, the majority of countries have fewer than 500 cases per 100,000 people, with 25,874 reporting less than 100. In terms of fatality rates, most countries have less than 20 deaths per 100,000 people, with several reporting less than 5. The scatterplot shows a similar pattern, with most countries seeing low incidence and fatality rates. However, one country is an exception, with an incidence rate of 26.76 and a mortality rate of 128.22. This demonstrates that factors such as healthcare quality, government policy, and local conditions can have a far higher impact on fatality rates than the number of cases alone.

REGRESSION

We are fitting a model on data from twenty countries considering total new cases as a function of population, population density and gross domestic product (GDP) per capita. The GDP per capita is given in "purchasing power standard," which considers the costs of goods and services in a country relative to incomes in that country; i.e. we will consider this as appropriately standardized.

Next, we need to add one (1) more column to our 'model_df' data frame. Specifically, one that has the total number of new cases for each of the twenty (20) countries. We calculate the total number of new cases by summing all the daily new cases, for each country, across all the days in the dataset.

We will gain a deeper look into the data.

```
##     country          pop       sq_km     gdp_pps    pop_dens total_cases
##    "factor"   "integer"   "numeric"   "numeric"   "numeric"   "integer"
```

```
##      country        pop               sq_km          gdp_pps
##  Austria : 1   Min.   :  514564   Min.   :   316   Min.   : 51.0
##  Belgium : 1   1st Qu.: 5266795   1st Qu.: 60701   1st Qu.: 71.0
##  Bulgaria: 1   Median : 7926273   Median : 90723   Median : 95.5
##  Cyprus  : 1   Mean   :17305840   Mean   :192118   Mean   :100.2
##  Denmark : 1   3rd Qu.:13474040   3rd Qu.:342955   3rd Qu.:120.8
##  Finland : 1   Max.   :83166711   Max.   :551695   Max.   :190.0
##  (Other) :14
##     pop_dens        total_cases
##  Min.   :  13.94   Min.   :  115243
##  1st Qu.:  57.67   1st Qu.: 1319422
##  Median : 100.50   Median : 2570210
##  Mean   : 179.14   Mean   : 6479211
##  3rd Qu.: 121.55   3rd Qu.: 5430242
##  Max.   :1628.37   Max.   :36612627
##
```

country	pop	sq_km	gdp_pps	pop_dens	total_cases
Austria	8901064	83858	128	106.14448	5402162
Belgium	11522440	30510	118	377.66109	4597870
Bulgaria	6951482	110994	51	62.62935	1271735
Cyprus	888005	9251	91	95.99016	596297
Denmark	5822763	44493	129	130.86919	3219571
Finland	5525292	338145	111	16.34001	1335318
France	67320216	551695	104	122.02434	36612627
Germany	83166711	357386	123	232.70836	35287690
Hungary	9769526	93030	71	105.01479	2141513
Ireland	4964440	70273	190	70.64506	1670377
Latvia	1907675	64589	69	29.53560	949252
Lithuania	2794090	65300	81	42.78851	1265985
Malta	514564	316	100	1628.36709	115243
Norway	5367580	385178	142	13.93532	1462873
Poland	37958138	312685	71	121.39418	6189562
Portugal	10295909	88416	78	116.44848	5514482
Romania	19328838	238397	65	81.07836	3246412
Slovakia	5457873	49036	71	111.30339	2535554
Spain	47332614	498511	91	94.94798	13564823
Sweden	10327589	450295	120	22.93516	2604866

We first see that our response variable, `total_cases`, is an integer variable. The rest of the predictor variables are numeric, with the exception of population, which is also integer, and `country`, which is a factor. Each row

in the model_df data frame represent one of the 20 countries we are focusing on for our regression analysis.

```
##
## Call:
## lm(formula = total_cases ~ pop + gdp_pps + pop_dens, data = model_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8240573 -1086170   -14153  1652279  8715573
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.860e+06  2.749e+06  -1.404    0.179
## pop          4.268e-01  3.625e-02  11.775 2.71e-09 ***
## gdp_pps      2.830e+04  2.524e+04   1.121    0.279
## pop_dens     6.495e+02  2.399e+03   0.271    0.790
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3655000 on 16 degrees of freedom
## Multiple R-squared:  0.8982, Adjusted R-squared:  0.8791
## F-statistic: 47.03 on 3 and 16 DF,  p-value: 3.685e-08
```

Our null hypothesis is that none of the predictors are significant predictors of total_cases (i.e. $H_0 : \beta_i = 0$). We see here that the only variable that is a significant predictor of total cases is population, thus we can reject the null hypothesis. This makes sense because there are more people in a higher population to get Covid-19, thus the significant p-value and the positive coefficient estimate. It is quite surprising that population density is not a significant predictor because one may think with more people crowded together, the more likely people are to spread the disease from one to another. We see that GDP is not significant in predicting total cases, which comes as a bit of a surprise. We initially thought that countries with a higher GDP would have more vaccinations to go around, thus stopping the disease from spreading as much. So, we expected a significant p-value with a negative coefficient estimate. However, it may make sense that GDP is not a significant predictor because it does not have to do with population or people being near one another.

The R-Squared value is 0.8982 and the Adjusted R-Squared value is 0.8791. These are both very high, and thus very good values. These indicate that the model is a good fit, as the model explains almost 90% of the variance. We see that the Adjusted R-Squared value is less than the R-Squared value, which makes sense because two of the predictors in the model are not statistically significant. Overall, this seems like a good model to use to predict total cases for other countries, but only one of the three initial predictors, population, is significant. So, we are going to fit a new model with just population as a predictor to see how it affects the R-Squared values.

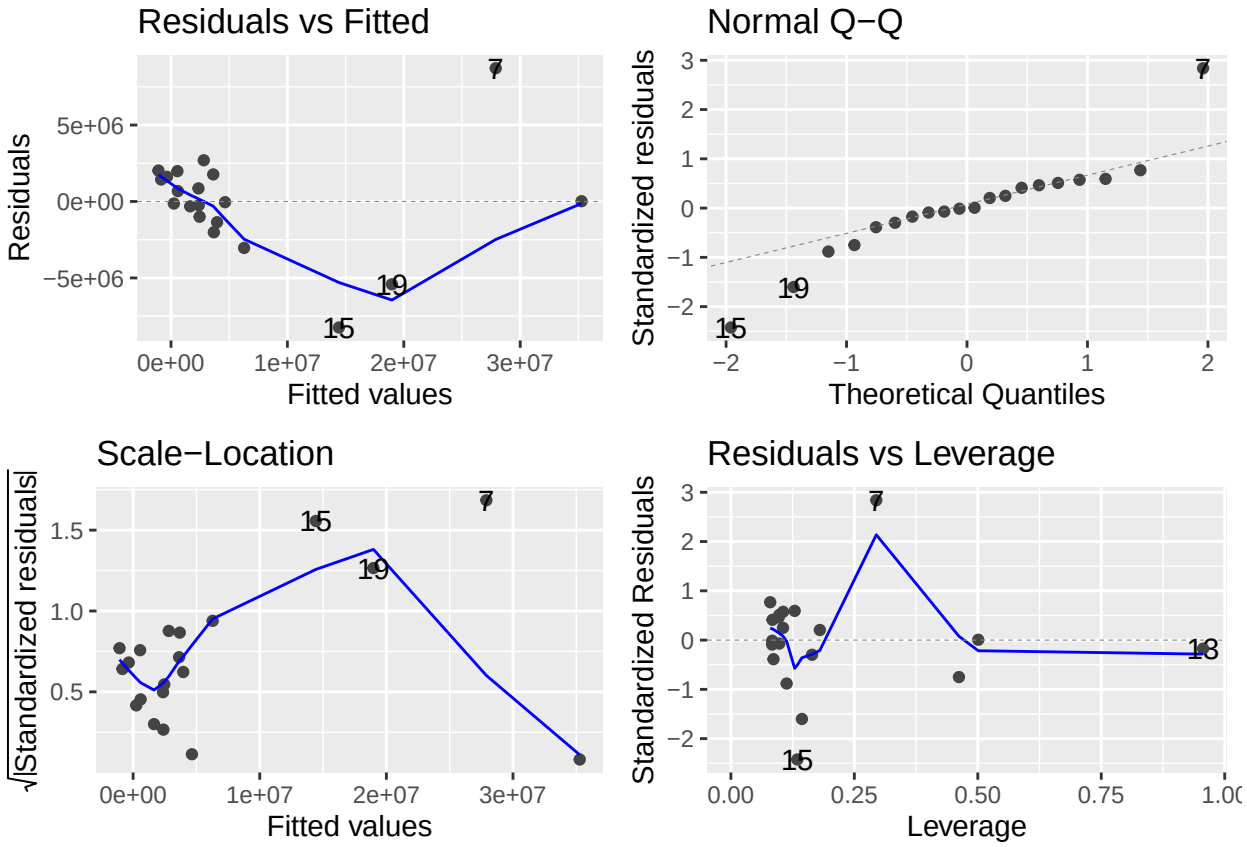
```
##
## Call:
## lm(formula = total_cases ~ pop, data = model_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9107810  -813451   637567  1118212  8778173
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.101e+05  1.010e+06  -0.901    0.379
## pop          4.270e-01  3.545e-02  12.043 4.76e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3588000 on 18 degrees of freedom
## Multiple R-squared:  0.8896, Adjusted R-squared:  0.8835
## F-statistic:   145 on 1 and 18 DF,  p-value: 4.764e-10
```

In this new output, we see that again, population is a statistically significant predictor. In fact, the p-value is much less in this model than it is in the model above. We also see that the Multiple R-Squared value is 0.8896 and the Adjusted R-Squared value is 0.8835. These are both high, again suggesting that this model is a good fit. While the Multiple R-Squared value in this model is lower, the Adjusted R-Squared value is higher.

Between the two models fitted above, we are going to choose the model with all three predictors for multiple reasons. Model_fit has improved prediction accuracy because the additional predictors (population density and GDP) explain more variance. In context, this helps capture countries' socioeconomic and spatial characteristics, which, even though these predictors are not significant, they might influence case count. Also, with all three predictors rather than just population, we have a lower risk of oversimplification. Being overly simple could lead to negative predictions if the country's population is smaller (like Luxembourg's is). This oversimplification leads to poor generalization across diverse datasets. Also, because our data is real world data that explains a real phenomenon (COVID-19), having more predictors accounts for the fact that COVID-19 is influenced by multiple factors which makes the model more comprehensive, even though two of the three predictors are insignificant. Having insignificant predictors in our model is okay in this case because it captures real world complexity and reduces bias in the predictions. It allows for a more robust model.

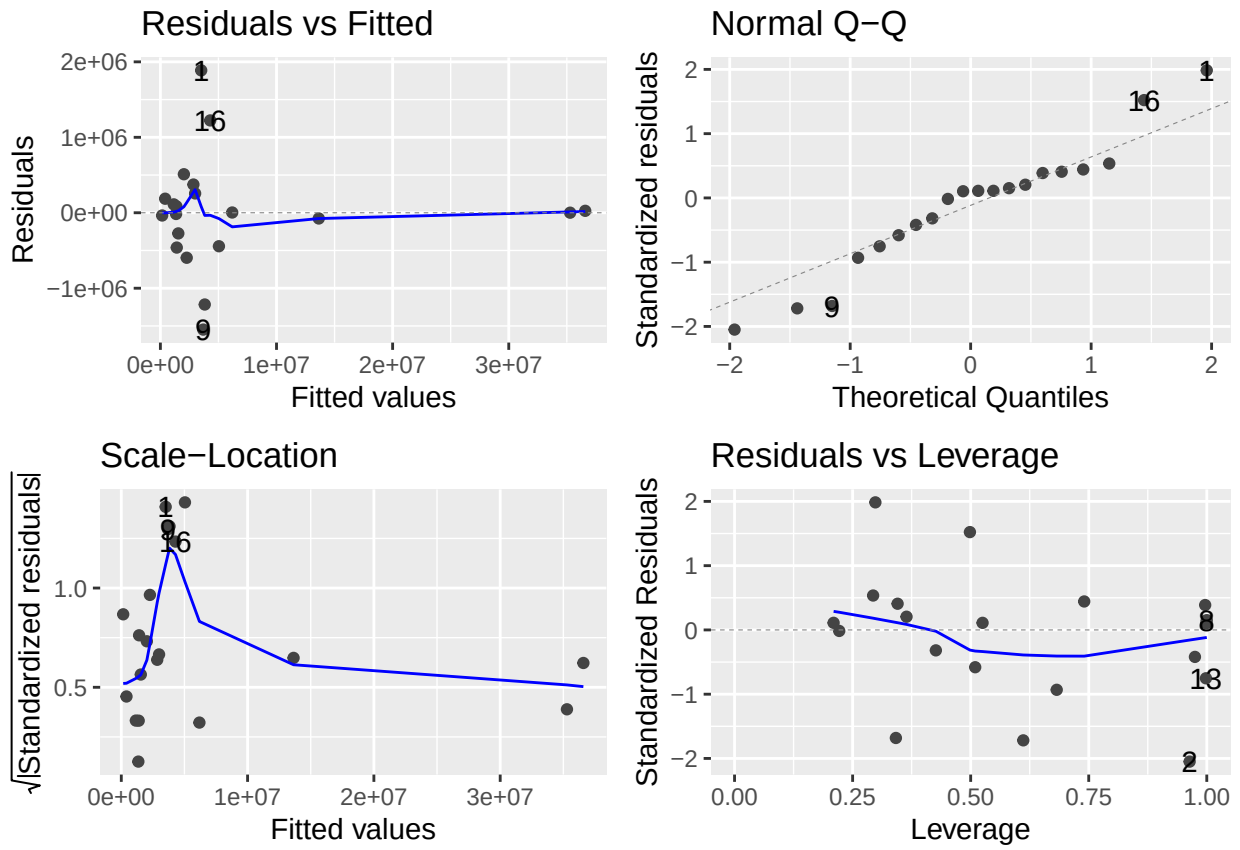
Before we continue with predictions, it is important to check the assumptions.



These plots tell us that the assumptions are not fully validated. The residuals vs. fitted plot tells us that there may be a non-linear relationship between the predictors and the response. In the Q-Q plot, we can see a definitive deviation in the tails from the line, which means the data may be non-normal. Also, in the residuals vs. leverage plot, the independence assumption is partially violated due to the presence of high leverage points. The scale-location plot also sees a noticeable curve, suggesting heteroscedasticity.

To address these issues, we will do some feature engineering. We will starting with fixing the linearity.

```
##
## Call:
## lm(formula = total_cases ~ poly(pop, 2) * gdp_pps * pop_dens,
##     data = model_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1548282  -316034     2358   203629  1887066
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -4.280e+07  1.059e+07  -4.040 0.003733 **
## poly(pop, 2)1                  -4.146e+08  1.099e+08  -3.772 0.005451 **
## poly(pop, 2)2                  -8.835e+07  4.611e+07  -1.916 0.091648 .
## gdp_pps                         5.169e+05  9.747e+04   5.302 0.000726 ***
## pop_dens                        1.975e+05  8.434e+04   2.342 0.047249 *
## poly(pop, 2)1:gdp_pps           4.899e+06  9.663e+05   5.070 0.000965 ***
## poly(pop, 2)2:gdp_pps           1.067e+06  3.627e+05   2.942 0.018663 *
## poly(pop, 2)1:pop_dens          1.595e+06  9.842e+05   1.620 0.143821
## poly(pop, 2)2:pop_dens         -9.027e+04  5.745e+05  -0.157 0.879035
## gdp_pps:pop_dens               -2.075e+03  6.730e+02  -3.083 0.015038 *
## poly(pop, 2)1:gdp_pps:pop_dens -1.834e+04  7.522e+03  -2.438 0.040701 *
## poly(pop, 2)2:gdp_pps:pop_dens -4.109e+02  4.678e+03  -0.088 0.932160
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1135000 on 8 degrees of freedom
## Multiple R-squared:  0.9951, Adjusted R-squared:  0.9883
## F-statistic: 147.5 on 11 and 8 DF,  p-value: 5.413e-08
```



Our assumptions look much better here, so we will use this new model, while also see how different it affects predictions compared to the initial model. We also see that some interaction terms are significant, even if the variable by itself is not. Our adjusted R-Squared and our Multiple R-Squared have both increased significantly, which tells us this fits our data much better.

Table 5: Actual Total Cases From Our Dataset For The Netherlands and Luxembourg

Country	Total Cases
Luxembourg	301031
Netherlands	8494705

Table 6: Predicted Total Cases vs. Actual Total Cases From Our Dataset

Country	Predicted Total Cases	Predicted Total Cases w/ Feature Engineering	Actual Total Cases
Luxembourg	3952125	25593590	301031
Netherlands	7521928	-3458252	8494705

First, we look at the predicted total cases versus the actual total cases for Luxembourg. The predicted total cases is 3,953,237 while the actual total cases is 301,031. This means that the prediction over estimates by over 3,500,000 cases. This is probably because Luxembourg has such a small population. For the Netherlands, the predicted total cases is 7,522,800 while the actual total cases is 8,494,705 cases, which is just under 1,000,000 cases off. While this is much better than the Luxembourg prediction, it still is not super close.

This underestimation could have to do with factors that our data does not include, like public health policies or how often people are being tested.

The large residual for Luxembourg (3,652,206 cases) could suggest that the model is not as robust as we initially thought, as countries with extremely small populations will deal with the same effect. The smaller residual for the Netherlands (971,905 cases) suggests that the model is better at predicting when the country's predictor values are closer to those in the initial model data frame.

Looking at the model that uses the interaction term and the polynomial term, we see that the predictions are incredibly off. This could be because the feature engineering may have over fit the data. The int_poly model fits the training data better, but it does so at the expense of generalizability.

See Appendix for more models fit using other various feature engineering techniques. We see that for the Netherlands, using the model with the interaction between population and population density predicts the closest total case value to the actual total case value. The predicted value is just 181,272 lower than the actual value. For Luxembourg, the closest predicted value is when we use the interaction effect between population and GDP. However, this value is negative, so the next best model is the model with the log transformation of GDP. It is important to note that with both of these models, the predicted value is still over 2,000,000 cases away, which is still much better than 3,500,000.

In conclusion, after fitting many different models using various techniques, we see that the biggest issue with this regression analysis is that the Luxembourg and Netherlands data falls on the outside of the data we trained our models on, thus making it hard to predict more accurately.

CONCLUSION

Through the exploratory data analysis, we derive that several countries in the European Economic Area were considerably affected by Covid-19. Our analysis revealed several outliers, both from incidence and fatality rates, highlighting the spread and damage done by the virus. Key insights include the exacerbation of the virus' impact by geopolitical factors, and the disparity between Eastern and Western European countries. As we perform hypothesis testing, we are able to determine if there is a statistically significant difference between the mean daily incidence rates between countries. By performing a t-test we are to determine a p-value to either reject or fail to reject the null hypothesis with our designated alpha level. In this case, we were able to conclude that there is a statistically significant difference between the mean daily incidences of Germany and France. For correlation, the Pearson coefficient is 0.113, showing a very weak link between COVID-19 incidence and fatality rates. This suggests that factors like healthcare quality and policies play a bigger role in affecting death rates than case numbers. In the regression analysis, we used a linear regression with additional feature engineering to build a predictive model for total cases in The Netherlands and Luxembourg. However, due to the small size of Luxembourg's population and the lack of more complex predictive variables, we were unable to create significant and accurate predictions. To conduct future research, we would like to find those more complex variables, like testing rates and healthcare policies, to make better predictors. We also would conduct network analysis on how traveling/migration affected infection rates in destination countries. Overall, this project helped us gain deeper insights into the Covid-19 pandemic in Europe.

SOURCES

Lin, Dan-Yu, Donglin Zeng, Devan V. Mehrotra, Lawrence Corey, and Peter B. Gilbert. "Evaluating the Efficacy of Coronavirus Disease 2019 Vaccines." *PubMed Central* (PMCID: PMC7799296, PMID: 33340397).

Majhi, Ritanjali, Rahul Thangeda, Renu Prasad Sugasi, and Niraj Kumar. "Analysis and Prediction of COVID-19 Trajectory: A Machine Learning Approach." *Journal of Public Affairs* 21, no. 4 (2021): e2537-n/a. https://doi.org/10.1002/pa.2537.

Spira, Benny. "Correlation Between Mask Compliance and COVID-19 Outcomes in Europe." Cureus, April 19, 2022.

Verschuur, J., Koks, E.E. & Hall, J.W. Observed impacts of the COVID-19 pandemic on global trade. Nat Hum Behav 5, 305–307 (2021).

APPENDIX

Table 7: Additional Fitted Models with Feature Engineering

Country	Luxembourg	Netherlands
Predicted Total Cases	3952125	7521928
Predicted Total Cases w/ Feature Engineering	25593590	-3458252
Actual Total Cases	301031	8494705
Population Density log transformation	4397710	7855415
GDP log transformation	2720767	7653794
Interaction Between population and GDP	-1819840	8252937
Interaction Between population and population density	3713397	8310771