COVID-19 Outbreak Prediction and Public Health Alert System

Group B
Supervisor: Soroush Sheikh
Student Name: Joshua John Danes, Sahil Kapoor, Jeanfranco Fung, Aidan Cafazzo
Student ID: 219560879, 218002469, 219121656, 215000565
Student Email: joshie15@my.yorku.ca, sahilk19@my.yorku.ca, jeanf03@my.yorku.ca, algonick@my.yorku.ca

Winter 2025

1. Objective

The primary goal of our project was to develop a robust predictive model that correlated weather temperature data with the incidence of COVID-19 cases across multiple countries and regions. Utilizing historical data spanning four months from January to April 2020(the first phase of the COVID-19 pandemic), we investigated how various factors such as temperature variations, wind speed and humidity levels impacted the spread of COVID-19. By identifying patterns and correlations, this project aimed to enhance understanding of viral transmission dynamics and offer insights that could improve public health predictions and preparedness strategies during environmental challenges.

2. Motivation

The intersection of meteorological conditions and infectious disease dynamics presented a compelling area of study. We sought to understand how environmental factors influenced COVID-19 transmission, providing critical insights for public health policymakers. This exploration allowed us to unravel the complexities of these relationships, thereby contributing to the formulation of evidence-based health strategies. By integrating extensive historical weather data with COVID-19 case statistics, we uncovered significant trends and correlations that informed public health responses, improved resource allocation and ultimately enhanced community resilience against future pandemics.

3. Related Work

Numerous studies had previously explored the correlation between weather conditions and the spread of COVID-19. Bashir et al. (2020) highlighted that low temperatures and humidity levels contributed to higher infection rates, utilizing statistical models to analyze

daily case trends. Their findings indicated that colder climates were conducive to virus transmission, emphasizing the need for climate-aware public health strategies.

Conversely, Zhu (2020) examined the effects of wind and precipitation on viral spread, noting that while strong winds could facilitate long-distance transmission, rain and high humidity levels might reduce airborne viral particles, thereby lowering transmission rates. These insights underscored the potential for weather conditions to serve as predictive indicators for COVID-19 outbreaks.

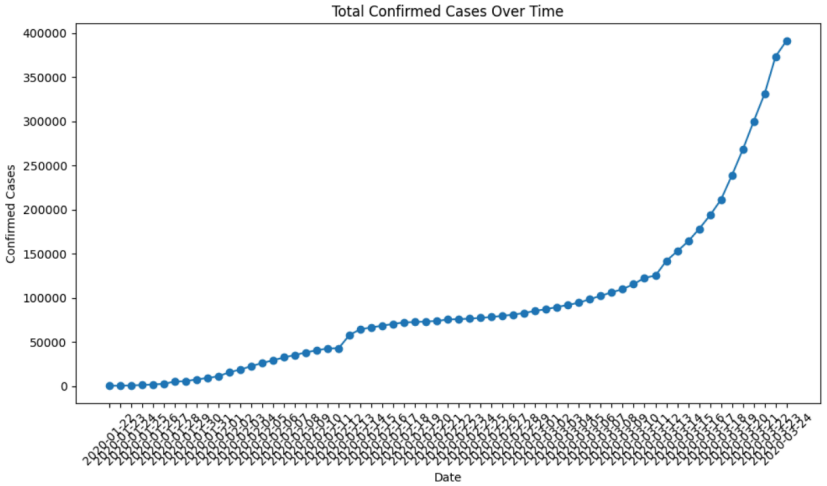
Despite these advancements, many public health models primarily focused on human mobility and case numbers, often neglecting critical environmental factors such as temperature and humidity. Our project aimed to bridge this research gap by comprehensively analyzing how these varying environmental conditions influenced the spread of COVID-19. In doing so, we empowered public health officials and policymakers with knowledge that could lead to more effective prevention strategies.

4. Methodology

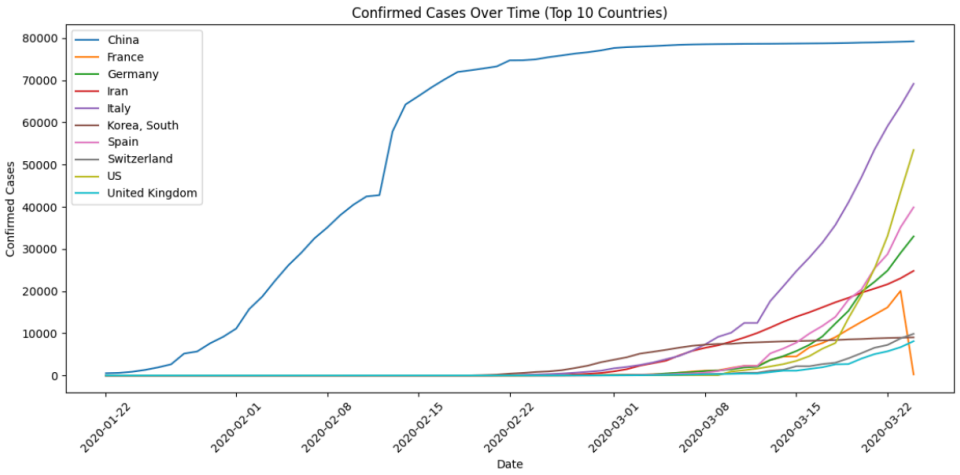
4.1 Preprocessing

During the preprocessing phase, we ensured the integrity of our dataset. We cross-referenced COVID-19 case numbers with official reports from health agencies, validating the accuracy of the data. Environmental data underwent rigorous preprocessing, including outlier detection to maintain data quality.

Given the varied geographical contexts of the countries under study, we initially applied a broad outlier detection methodology before examining specific cases in detail to ascertain legitimacy. First, Covid cases were simply plotted over time to see if there were any insights that could be gleaned from a basic analysis. Not much was gained from this exercise.



Next, the data was separated into cases-by-Country over time. This highlighted that China's data was a huge outlier. Since the virus originated there, China already had high infection rates by the time other countries began experiencing outbreaks. In some instances, China's early-infection data might be useful, but in our specific case, it was not material and skewed results, so it was removed entirely.



Another feature of the data that became apparent was that case data did not really pick up until a date much later than the overall time period tracked. Although there was data beginning in January, cases did not increase in any way until about March. Consequently, the data was filtered to begin at the point in time where most countries had at least one reported case.

4.2 Feature Engineering

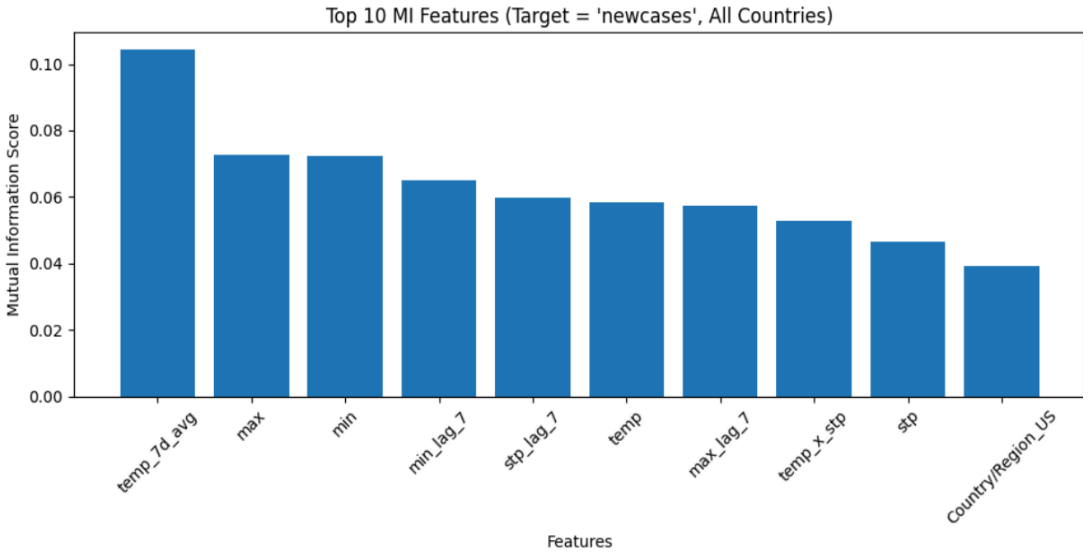
To facilitate comprehensive analysis, we introduced a number of new calculated variables:

- "Daily New Cases" column: Data existed for cumulative number of cases, which was less useful for our purposes. Consequently, a daily figure of only newly reported cases (the difference between two subsequent days' cumulative numbers) was generated. This reduced biases related to cumulative counts and enhanced the accuracy of the infection rate representation.
- Combination variables: Since many weather variables co-exist (e.g. it can be windy and cold), combination variables were created to see if there was a stronger correlation between these versus the singular features. However, as relationships were explored, the number of combinations became too bloated and were purposefully limited to only the most significant.

Since our interest is weather impact on *infection* rates, and the data contains cases *reported* (after symptoms display), a "7-day lag" variable was created to take into account that the day a patient positively tested for COVID-19 is likely *not* the actual day they contracted covid. Seven days was an average figure chosen based on research showing that time elapsed between infection and the appearance of symptoms was between 2-to-14 days[1].

Furthermore, to take into account COVID-19's exponential growth rate, a log function was applied to the "Daily New Cases" variable in order to increase linearity of the relationship. This would help with coming to any conclusions regarding relationship between time and number of cases.

In order to identify the important features of the data, Random Forest and Mutual Information were used and Heat maps generated by country. This highlighted features that were the most highly correlated to the new cases variable.



4.3 Model Selection

Time Series Regression Model: This model helped us identify trends in COVID-19 cases over time, indicating whether they were increasing, decreasing, or remaining stable.

ANOVA Testing: Used to assess whether there were statistically significant differences in mean cumulative COVID-19 cases across various environmental factor ranges. Specifically, we examined the impact of humidity, temperature, and wind speed by comparing the COVID-19 case counts across their respective categorized ranges.

---

[1] Canada, Public Health Agency of. "Government of Canada." *Canada.Ca*, / Gouvernement du Canada, 7 Feb. 2024, www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection/health-professionals/transmission.html.

Linear Regression Model: We used this model to explore relationships (positive, negative, or neutral) between confirmed COVID-19 cases and various environmental factors, such as temperature and humidity.

4.4 Model Training and Evaluation

For effective training of our models, we implemented a country-specific split of the dataset. We used the earliest 70% of reported data as the training set and the remaining 30% for testing. This method preserved temporal trends and minimized biases inherent in random splits. We assessed model performance using the Mean Squared Error (MSE) metric, providing insights into predictive accuracy.

4.5 Optimization Strategies

To enhance model performance, we employed several optimization strategies:

1. Feature Selection and Reduction: We identified and eliminated features that did not contribute significantly to model performance, focusing on those that are enhanced predictive accuracy.
2. Standardization: We standardized numerical features to ensure that no individual feature disproportionately influenced model performance.
3. Iterative Refinements: As we progressed through project milestones, we explored and implemented additional optimization strategies, ensuring our model remained adaptive and robust.

5. Deliverables

The project produced several key deliverables that enhanced both the academic understanding of the relationship between weather patterns and COVID-19 cases and also the practical applications for public health strategies. The main deliverables included:

1. Codebase: A well-structured and documented codebase that included implementations of data preprocessing, statistical analyses, hypothesis testing and machine learning models for predicting COVID-19 cases based on weather variables. The code was organized for clarity and ease of reproduction.
2. Model Artifacts: The trained models such as the optimized configurations and hyperparameters were saved as artifacts. These models were utilized for future predictions and were integrated into public health monitoring systems for real-time analysis.
3. Technical Report: A technical report that detailed the methodology, data sources, feature engineering, model selection, training processes and evaluation results. The report included visualizations highlighting the correlations between weather patterns and COVID-19 case trends, along with discussions on the implications for public health responses.

4. Presentation Slides: A set of presentation slides that covered and summarized the project's background, its objectives, methodology, findings and recommendations.
5. Documentation: Detailed documentation along with the codebase provided insights into the project's structure, its dependencies and instructions for executing and deploying the predictive models. This ensured that others can easily replicate the study and apply the findings in relevant contexts.

These deliverables aimed to facilitate a deeper understanding of how weather impacts COVID-19 dynamics while providing insights for improving public health interventions and preparedness in the face of future outbreaks.

6. Experimental Setup

For our lab space, we worked in a quiet place with access to electricity and the Internet. No specialized lab was needed, except for a computer for virtual meetings and a study area that helped with our group work and discussions. For our hardware, a personal computer/laptop with good processing power was required to handle software tools for modeling, coding, and analyzing. A multicore CPU was necessary for running models efficiently. A dedicated GPU was used to speed up machine learning. For our software, we used the following tools for coding, analysis, and reporting:

- *Python (latest version)* - Main programming language for data analysis
- *Jupyter Notebook* - For interactive coding and testing
- *Scikit-learn, XGBoost, and Statsmodels* - For machine learning framework models
- *Pandas and Numpy* - For organizing and analyzing data
- *Matplotlib, Seaborn, and Plotly* - For visualization and creating graphs
- *Excel and SQL* - For handling and managing data tables
- *LaTex and Google Docs* - For writing and formatting the final report
- *Google Slides* - For presentation of project

For our data access, we used the 'Weather Data for COVID-19 Data Analysis' dataset. The file contained country, COVID-19 case numbers and weather data such as temperature, humidity, wind speed, and precipitation. This dataset was already available and ready for analysis. For our Internet access, a stable Internet connection was needed for:

- Researching past studies
- Using cloud-based tools like Google Colab and GitHub for coding
- Accessing online datasets if needed
- Collaborating through Google Drive, GitHub, and Notion.

These are the collaboration tools that we used:

- Google Drive - For document sharing
- Notion - For project management
- GitHub - For storing and sharing code
- Whatsapp, Zoom, Discord, and Slack - For virtual meetings and discussions

All of these resources were available and accessible, ensuring that we could complete the project successfully and achieve high-quality results.

7. Impact

This project has the potential to make significant contributions to both academic research and real-world applications in the following ways:

- **Improved Public Health Predictions and Preparedness**: By modeling virus transmission dynamics and how it may be impacted by weather phenomena, this project aimed to contribute to more accurately predicting virus transmission rates and patterns, particularly in identifying possible peaks related to weather. This can lead to improved preparedness by health services, resource allocation, vaccination rollout and overall timeliness in the operation of infection control. It can also be paired with weather reports to support the public in individual decision-making based on expected risk (e.g. should I go on public transit today, given it's a higher risk.)
- **Data-Driven Decision-Making**: The utilization of advanced machine learning models for predicting virus transmission can empower healthcare delivery teams with valuable planning insights. This project's findings can facilitate data-driven decision-making, enabling authorities to respond dynamically to potential virus outbreak patterns and focus on high-risk communities.
- **Generalizability to Other Domains**: The methodology developed for COVID-19 virus patterns concerning weather conditions is not only limited to Coronavirus. The same principles can be adapted and applied to forecast transmission rates and patterns for other similarly transmitted viruses, such as RSV and other respiratory viruses, contributing to a broader understanding of where and how weather can impact transmission.
- **Scientific Contribution**: The application of different regression models, including linear regression or time series regression, contributes to the scientific understanding of the strengths and limitations of these methods in predicting transmission patterns and weather impact. The comparative analysis will provide insights into the most suitable models for such forecasting tasks. The impact of this project extends to practical applications, fostering improvements in health services response, resourcing and effectiveness.

8. Results

Analysis of Variance (Anova):

This analysis of variance aims to investigate whether there are statistically significant differences in the variance of cumulative COVID-19 cases (continuous variable) across different geographical groups and environmental ranges (temperature, wind speed, humidity). The result of the analysis will help us understand if these environmental and geographical factors have a significant impact on COVID-19 cases. We will test the p-values against an alpha/critical value of 0.05; this threshold is not a fixed number, and depending on the situation it can vary. Also for the ranges, the data of the median of environmental factors of a province/state got sorted in descending order, and they are based on quantiles. Meaning that the lowest 25% of the median temperature points are classified as very low temperature, then the second lowest 25% are classified as low temperature, and subsequently…

This was done for temperature, humidity, and wind speed This way it prevents scenarios when there are too few points for one range or too many for one.

Comparisons	**P-value**	**F-statistic**	**Individual T-tests results**

<table>
<tr>
<td>Cumulative Cases Across Continents

<u>Null Hypothesis (H0)</u>: The mean cumulative COVID-19 cases across continents are the same

<u>Alternative Hypothesis (HA)</u>: At least one continent has a different mean cumulative COVID-19 cases compared to the other continents</td>
<td>0.043373: With a p-value of 0.04 we reject the null hypothesis if we have a conventional critical value of 0.05. By rejecting this null hypothesis, we conclude that there is statistically significant evidence to suggest that the mean cumulative COVID-19 cases are not the same across all continents. Indicating that there is at least one continent that is different from the mean of at least one other continent.</td>
<td>2.319254: indicates that the variability in the average/mean cumulative COVID-19 case counts between the different continents is approximately 2.32 times larger than the variability in the cumulative COVID-19 case counts within each individual continent.</td>
<td>
<table>
<tr><th>Pair Comparisons</th><th>P-value</th></tr>
<tr><td>Africa - Asia</td><td>0.9221</td></tr>
<tr><td>Africa - Europe</td><td>0.0238</td></tr>
<tr><td>Africa - North America</td><td>0.6041</td></tr>
<tr><td>Africa - Oceania</td><td>1.0</td></tr>
<tr><td>Africa - South America</td><td>0.9972</td></tr>
<tr><td>Asia - Europe</td><td>0.168</td></tr>
<tr><td>Asia - North America</td><td>0.9871</td></tr>
<tr><td>Asia - Oceania</td><td>0.9971</td></tr>
<tr><td>Asia - South America</td><td>1.0</td></tr>
<tr><td>Europe - North America</td><td>0.4609</td></tr>
<tr><td>Europe - Oceania</td><td>0.5497</td></tr>
<tr><td>Europe - South America</td><td>0.7645</td></tr>
<tr><td>North America - Oceania</td><td>0.9689</td></tr>
<tr><td>North America - South America</td><td>0.9984</td></tr>
<tr><td>Oceania - South America</td><td>0.9997</td></tr>
</table>
Even though we saw a p-value of 0.0433 in our anova test, we see that only Africa and Europe's null hypothesis is rejected with a p-value of 0.0238.

As to the other p-values in the table, a majority of them are close to 1.0 like Africa - S. America with a p-value of 0.9972 or Oceania - S. America with
</td>
</tr>
</table>

<table>
<tr><td></td><td></td><td></td><td>0.9997. However, Asia - Europe also showed a small p-value of 0.168 relative to the other p-values hinting that there might be some underlying trend contribution to a potential difference between these regions.

However, there are pairs that have a p-value of 1.0, which is really rare to see. These results can be attributed to the sample sizes of Oceania(11) and S. America(12), in comparison with Africa (53) and Asia (77).</td></tr>
<tr><td>Cumulative Cases Across Humidity Levels

<u>Null Hypothesis (H0)</u>: The mean cumulative COVID-19 cases are the same across all humidity ranges

<u>Alternative Hypothesis (HA)</u>: At least in one humidity range has a different mean cumulative COVID-19 case count compared to the others.</td><td>0.068859: Using a critical value of 0.05, we fail to reject the null hypothesis. Meaning we do not have enough evidence to conclude that there is a significant difference in the mean cumulative COVID-19 cases across humidity ranges, if we stick with a 0.05 critical value.

While we fail to reject the null hypothesis under a critical value of 0.05, it doesn't necessarily mean that humidity has no effect. If we were using a less stringent significance level (e.g., 0.10), we might reach a different conclusion.

Failing to recognize effects/factors could mean the difference</td><td>2.389577: The variance in the mean cumulative COVID-19 cases between the humidity ranges is about 2.39 times greater than the variance in cumulative COVID-19 cases within the humidity ranges.</td><td><table><tr><td>Pair Comparisons</td><td>P-value</td></tr><tr><td>High - Low</td><td>0.6659</td></tr><tr><td>High - Medium</td><td>0.1264</td></tr><tr><td>High - Very Low</td><td>0.0866</td></tr><tr><td>Low - Medium</td><td>0.7151</td></tr><tr><td>Low - Very Low</td><td>0.6145</td></tr><tr><td>Medium - Very Low</td><td>0.9986</td></tr></table>Based on the table data and this analysis, humidity range, as categorized, does not appear to have a statistically significant effect on the cumulative number of COVID-19 cases. Except for high and very low humidity having a p-value of 0.0866, while not statistically significant at the standard alpha level of 0.05, there is a trend suggesting a potential difference in cumulative cases between these two extreme humidity ranges.</td></tr>
</table>

<table>
<tr>
<td></td>
<td>between saving lives and not.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cumulative Cases Across Temperature Ranges

<u>Null Hypothesis (H0)</u>: The mean cumulative COVID-19 case are the same across all the temperature ranges

<u>Alternative Hypothesis (HA)</u>: At least one temperature range has a different mean cumulative COVID-19 cases compared to the others</td>
<td>0.034514: Is less than the conventional critical value of 0.05, leading to the rejection of the null hypothesis. Meaning that if the null was assumed to be true the cumulative number of covid cases across temperature were true, there would only be a 3.45% chance of observing the data.</td>
<td>2.915262: suggests that the variance of cumulative COVID-19 cases across different temperature intervals is approximately 2.92 times larger than the variance of COVID-19 cases within those same temperature intervals.</td>
<td>
<table>
<tr><td>Pair Comparisons</td><td>P-value</td></tr>
<tr><td>High - Low</td><td>0.0395</td></tr>
<tr><td>High - Medium</td><td>0.7908</td></tr>
<tr><td>High - Very Low</td><td>0.1368</td></tr>
<tr><td>Low - Medium</td><td>0.3021</td></tr>
<tr><td>Low - Very Low</td><td>0.9564</td></tr>
<tr><td>Medium - Very Low</td><td>0.6057</td></tr>
</table>
Examining the comparisons, we can see that we do not reject the null hypothesis for various comparisons like low and very low temperature with p-value of 0.96, which we could say that it was not by chance that their distribution were similar.

We reject the null hypothesis for only high and low temperature intervals if we use a critical value of 0.05. However we also see
High and Very low with p-value:0.1368, while not statistically significant at the conventional 0.05 level, it still shows a presence of a trend indicating a potential difference in cumulative cases between these two extreme temperature ranges.
</td>
</tr>
<tr>
<td>Cumulative Cases Across Wind Speed Ranges

<u>Null Hypothesis (H0):</u> No significant difference in the average number of cumulative COVID-19 cases between the various</td>
<td>0.575863: We see that it is quite a high p-value compared to the conventional critical value of 0.05, in which case, we fail to reject the null hypothesis.</td>
<td>0.631268: A low F-statistic indicates that the variation of cumulative COVID-19 cases between the groups(wind speed intervals/ranges) is small compared to the</td>
<td>
<table>
<tr><td>Pair Comparisons</td><td>P-value</td></tr>
<tr><td>High - Low</td><td>0.9075</td></tr>
<tr><td>High - Medium</td><td>0.5974</td></tr>
<tr><td>High - Very Low</td><td>0.7632</td></tr>
<tr><td>Low - Medium</td><td>0.9383</td></tr>
</table>
</td>
</tr>
</table>

<table>
<tr>
<td>wind speed categories

<u>Alternative Hypothesis (HA):</u> At least one wind speed range has a different mean cumulative COVID-19 case count compared to the others.</td>
<td>In other words, wind speed ranges do not appear to have a statistically significant relationship to the cumulative covid 19 case count.</td>
<td>variation within the groups.</td>
<td>
<table>
<tr><td>Low - Very Low</td><td>0.9906</td></tr>
<tr><td>Medium - Very Low</td><td>0.9919</td></tr>
</table>

All the pairs' have p-value above the critical/alpha value of 0.05. Showing that wind does not have a statistically significant impact on explaining the variation of cumulative COVID-19 cases.</td>
</tr>
</table>

Further Analysis on the Result from the ANOVA:

Based on the low p-values of temperature and humidity, we have evidence suggesting that the mean cumulative case counts differ across these defined environmental factors. However, ANOVA primarily tells us that a difference exists; it doesn't reveal the nature nor the strength of the relationship, so correlation analysis is performed.

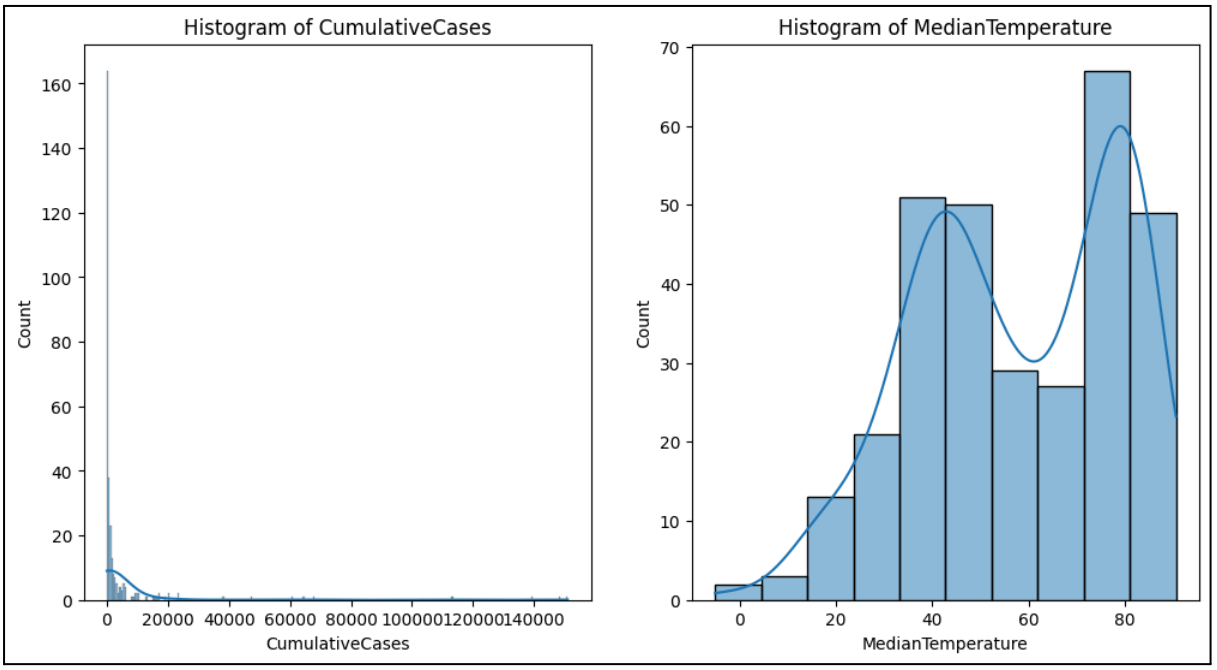


Fig 1

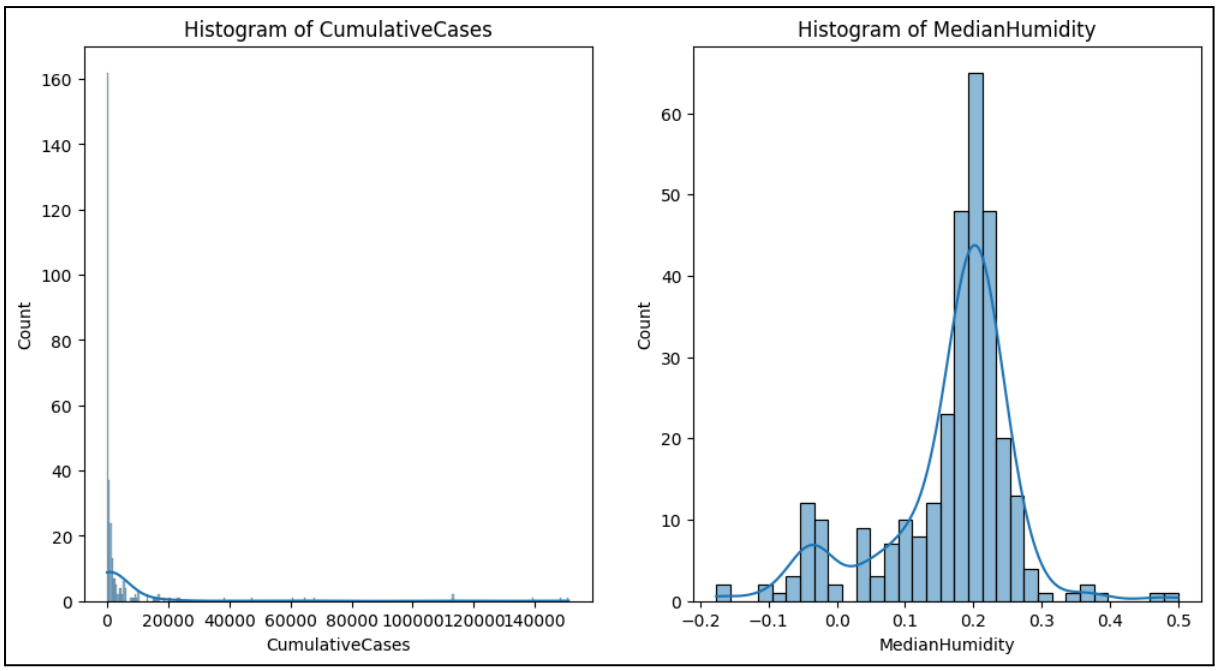
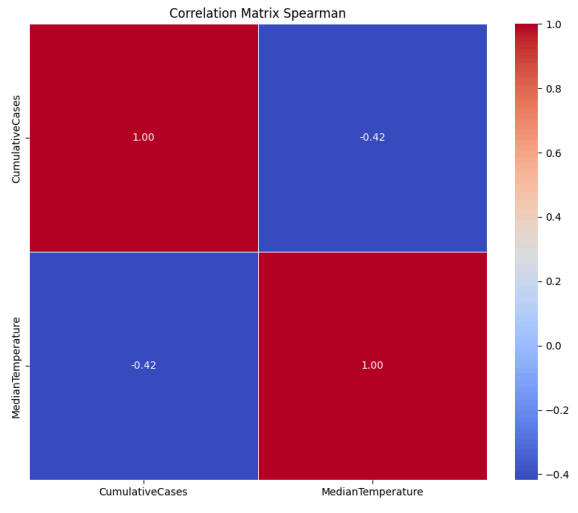
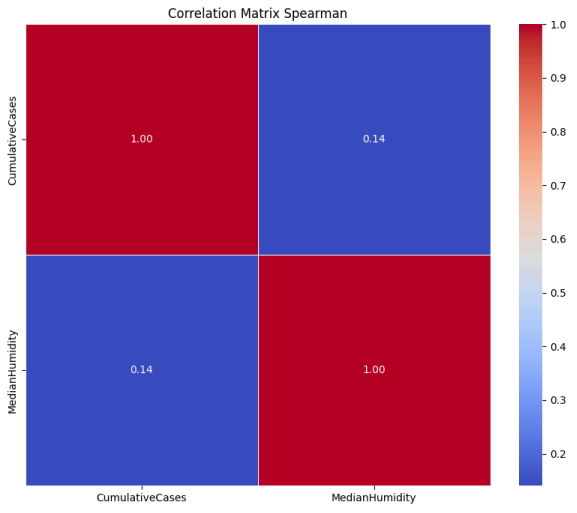


Fig 2

We will be performing spearman correlation analysis because as shown in **Fig 1** and **Fig 2**, they do not display a normal distribution. While median humidity shows signs of normal distribution, cumulative cases show a right/positive skewness, and median temperature a bimodal distribution, hence we proceeded with spearman correlation to counteract outliers and non-normal data.

Temperature	Humidity
	

Temperature had a moderate negative correlation of -0.42 suggesting that as temperature increases, total number of cases tend to decrease. We could also say the inverse

relationship implies that colder temperatures might be associated with higher spread of COVID-19 cases, in the context of our data.

For humidity, there is a weak positive correlation/relationship of 0.14 suggesting that as humidity levels increase, there is a slight tendency for COVID-19 cases to also increase.

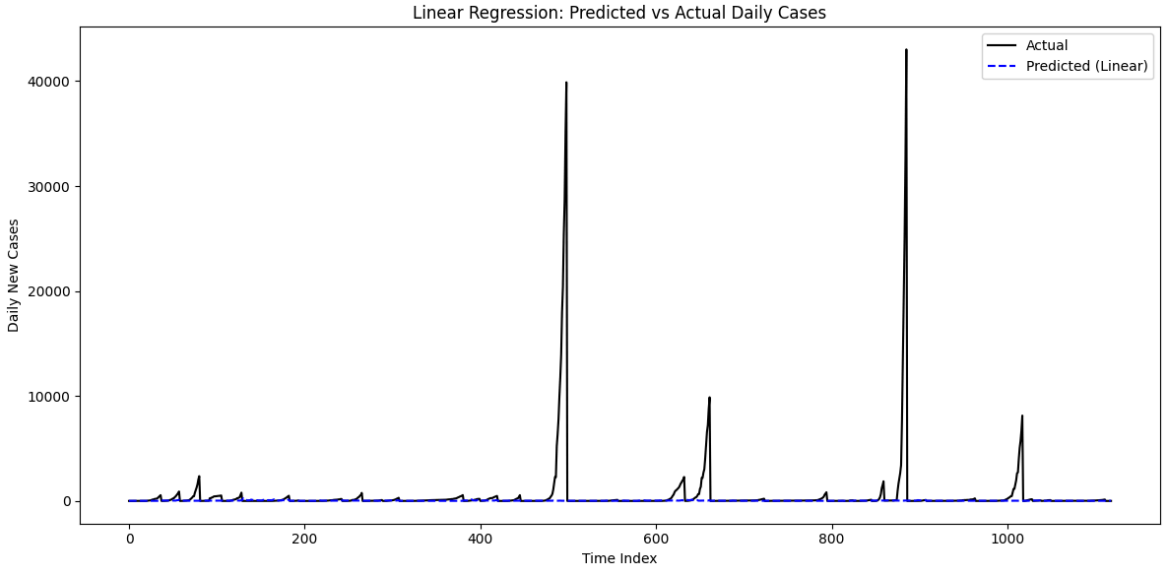
The code for the ANOVA tests and correlation can be found in:
"**HypothesisTesting(Anova&Correlation)/HypothesisTesting.pynb**"

Analysis of Models:

Linear Regression

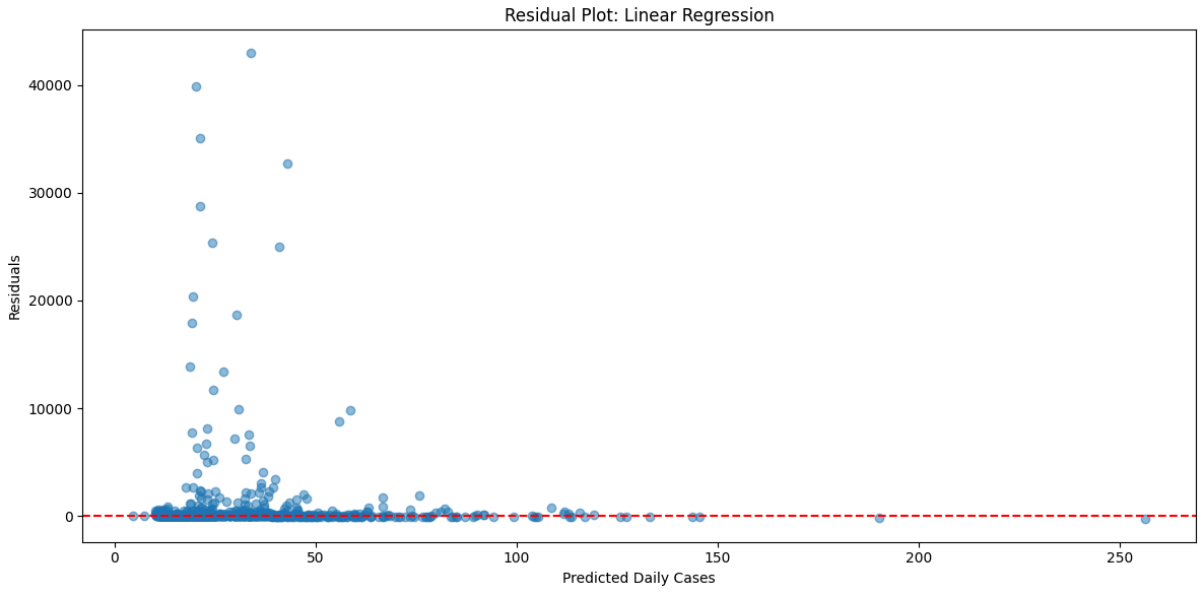
Results:

Linear Regression MSE: 9268650.71
Linear Regression R^2: -0.0268



Observation of Linear Regression: Predicted vs Actual Daily Cases:

- The predicted line is mostly flat, while in the actual case numbers, we can see a spike of case due to when the number of daily confirmed cases sharply increased.
- It has R^2 score of -0.0268 and a high MSE, which means it failed to show a clear link with the environmental factor and its prediction is way of track
- Linear regression can't follow the sharp changes in COVID-19 cases
- Even though weather data such as temperature, relative humidity, absolute humidity, and wind speed are being used without time-based info like past case numbers. It will struggle to track how cases rise and fall over time.



Observation of Residual Plot: Linear Regression:

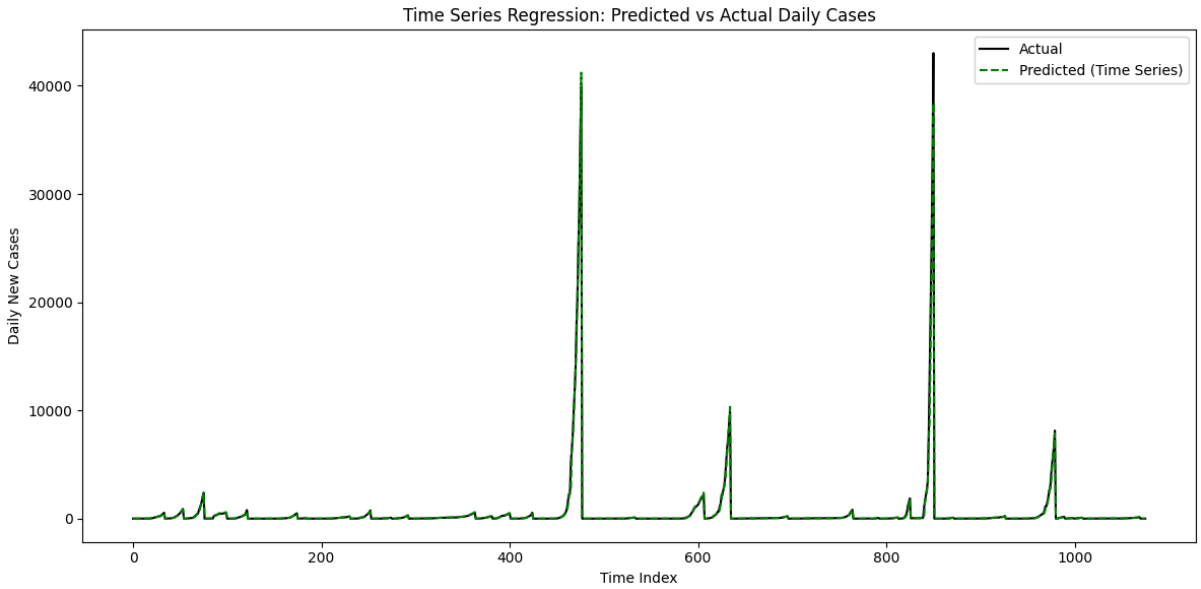
- Residuals are spread widely and increase with predicted values
- The model often underpredicts on spike days.

Time Series Regression

Results:

Time Series Regression MSE: 87273.58

Time Series Regression R^2: 0.9907

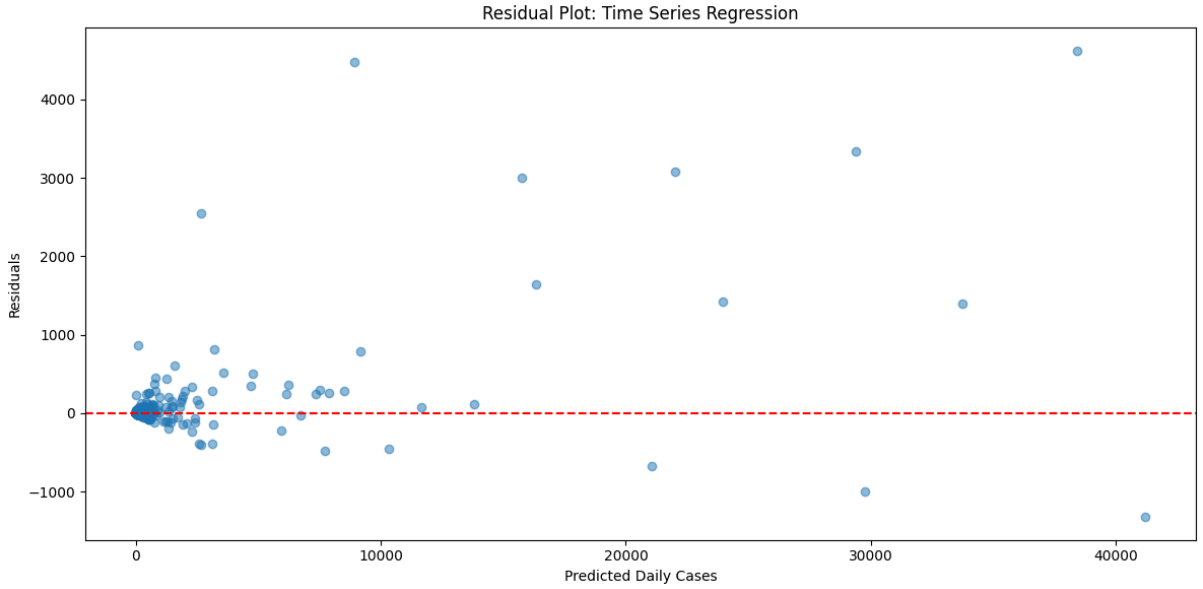


Observation of Time Series Regression:

- The predicted line is closely matches the spikes in actual daily cases.
- It makes this model perform best and accurate, as shown by its high R^2 and a very low

MSE

- Using lagged daily cases helps it learn from historical trends, which is crucial for forecasting.
- The graph shows the strong alignment between predictions and actual values, indicating low bias and accurate timing.
- This proves that even adding time-based data will still give better results when predicting how a disease spreads.



Observation of Residual Plot: Time Series Regression:

- Residuals are small and clustered around zero, it means that the model accurately predicted most daily cases.
- No big spread or clear pattern = no major bias.
- Performs very well even with the spikes in actual data and it handled spikes in the data without losing accuracy.

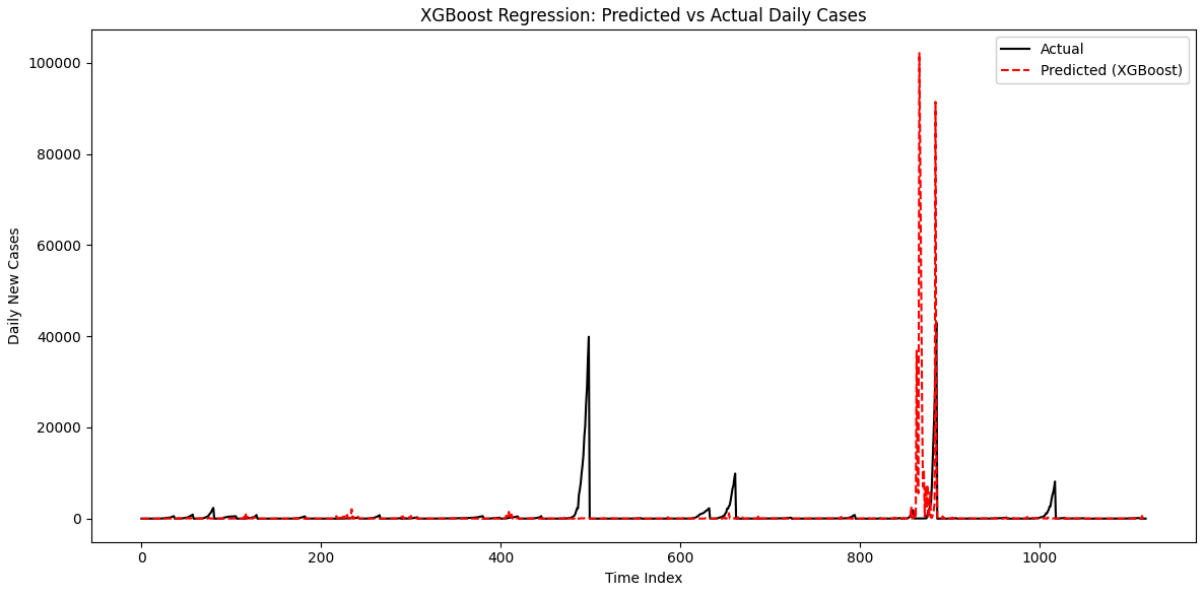
XGBoost Regression (Baseline)

Results:

XGBoost Regression MSE: 30032823.95

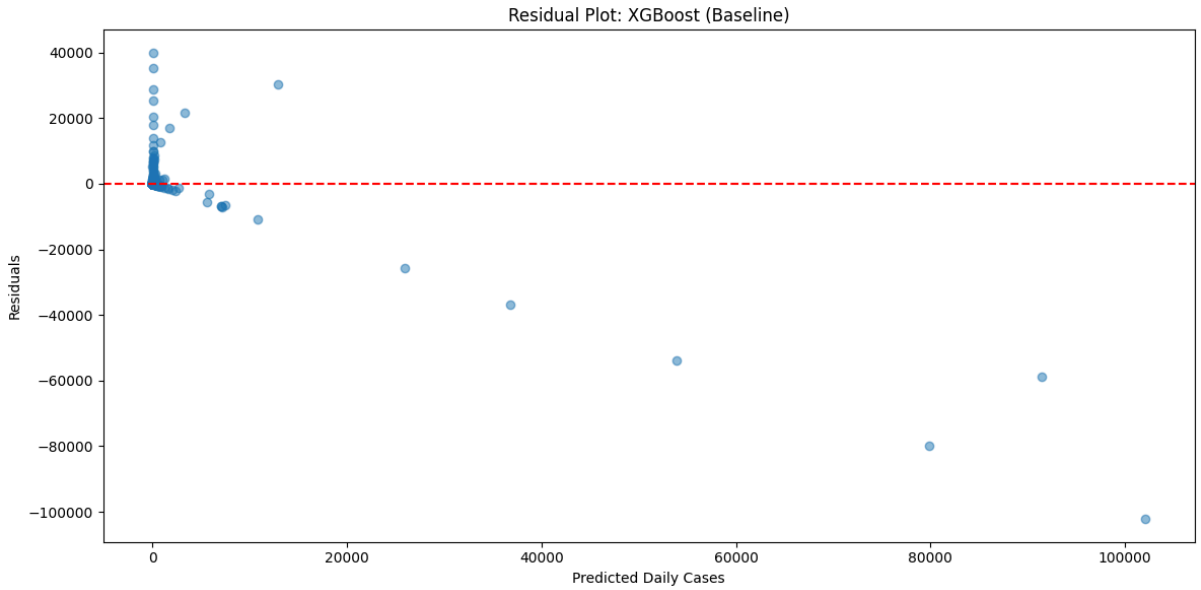
XGBoost Regression R^2: -2.3271

XGBoost Baseline have a very low negative R^2 and a very high MSE, which means that it is worse than just predicting the mean value of the target, and its predictions are way off track.



Observation of XGBoost Regression: Predicted vs Actual Daily Cases:

- The red line shows huge unrealistic spikes predicted that is way off to the actual cases.
- It reacts too strongly to outliers or noise
- Even though it uses weather factors, it does not handle time-based trends well. It misses the value of lag features, which are key for predicting how cases change over time.



Observation of Residual Plot: XGBoost (Baseline):

- Shows very large residuals, especially for extreme case values.
- Prediction errors grow as case numbers rise.
- Outliers dominate the graph, which shows model instability.

XGBoost Regression (Tuned)

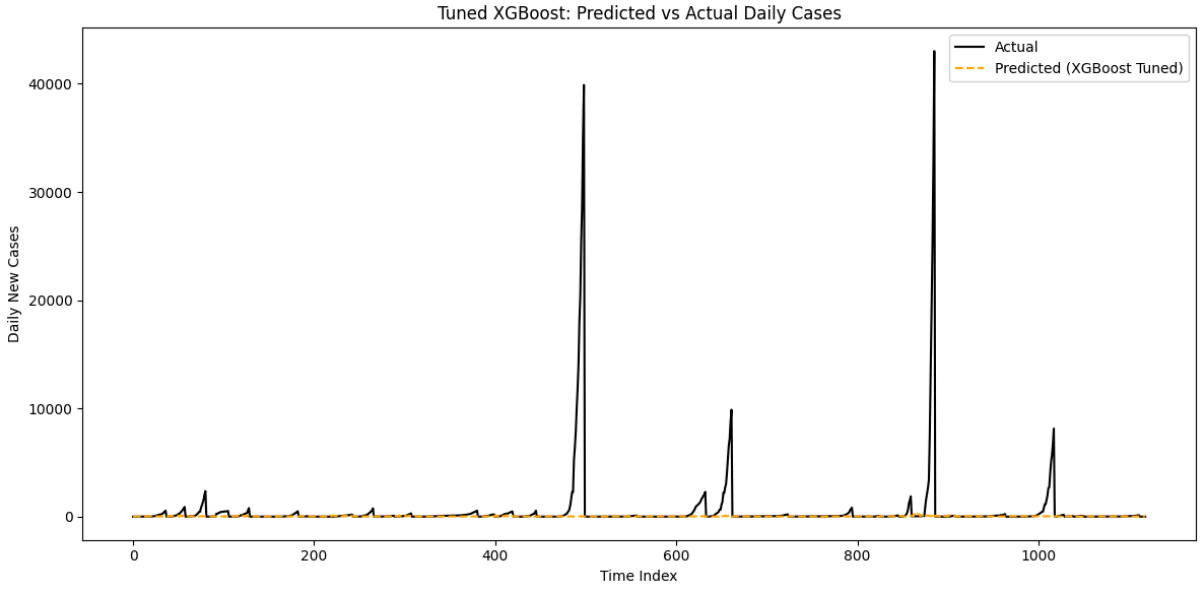
Results:

Best Parameters: {'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 100, 'subsample': 0.8}

Tuned XGBoost MSE: 9236957.18

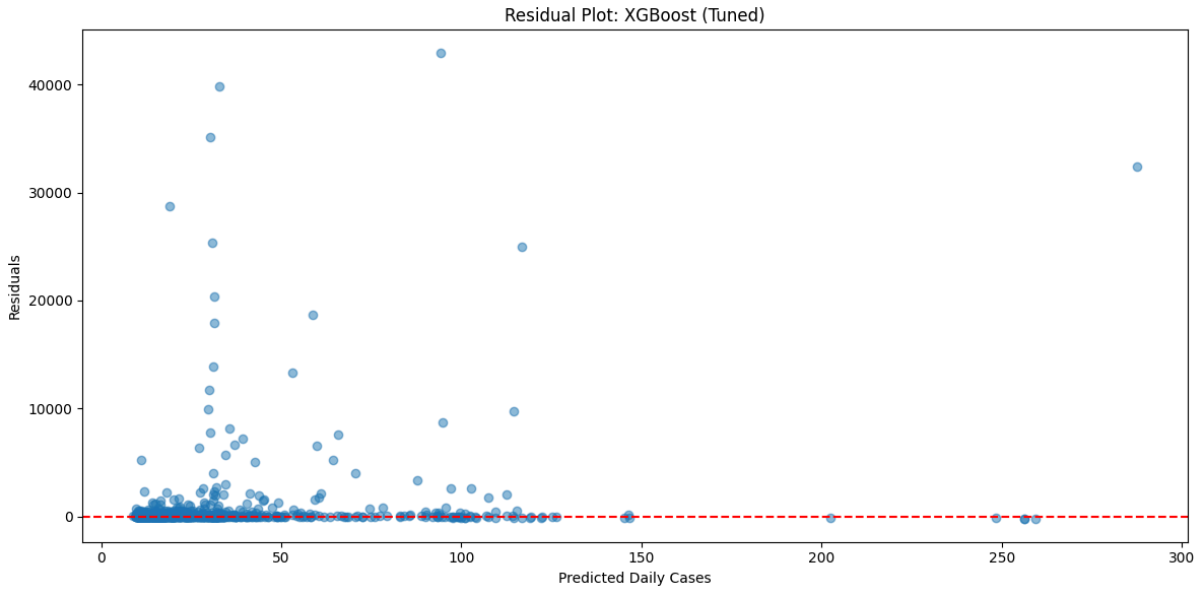
Tuned XGBoost R²: -0.0233

XGBoost Tuned have a very low negative R² and a very high MSE, which means that it is worse than just predicting the mean value of the target, and its predictions are way off track.



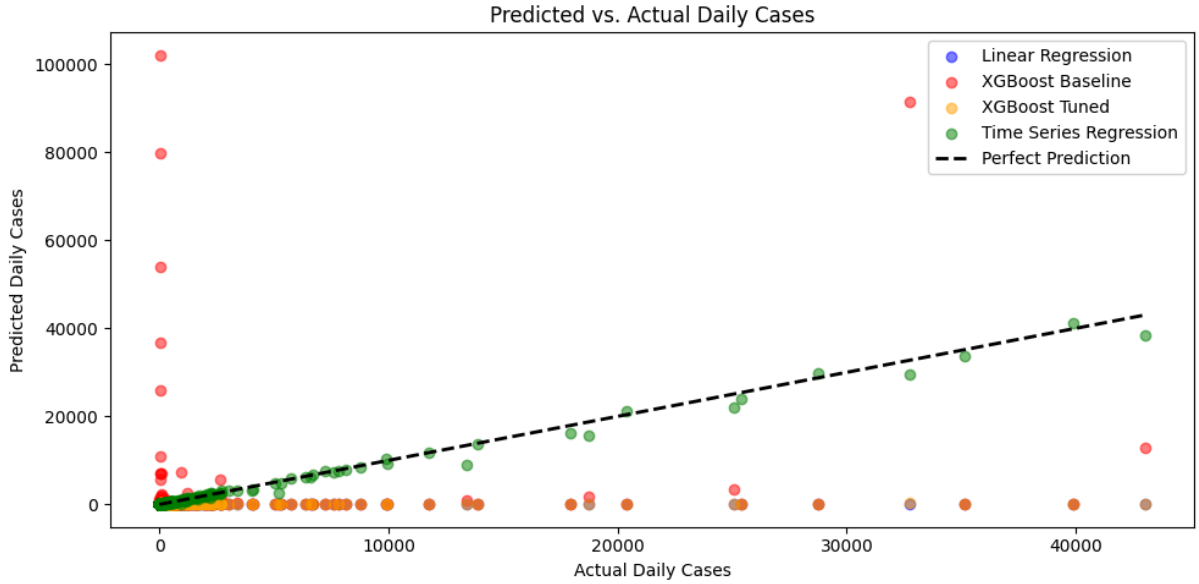
Observation of Tuned XGBoost: Predicted vs Actual Daily Cases:

- The predictions are flat and it does not respond to the spikes of the actual case numbers.
- The tuned model now underfits, but it misses the real trend of actual cases because it is being cautious.
- It shows that just tuning it is not enough to give the result you want.



Observation of Residual Plot: XGBoost (Tuned):

- The residuals are a bit better than the baseline version.
- There's less scatter, but errors are still high.
- Tuning helped slightly, but not enough.
- The model still lacks an understanding of time-based trends.
- It still fails to predict peaks and real case growth accurately.



To sum up, Time Series Regression was the best model. It had a very low MSE and an R^2 close to 1, meaning its predictions were almost spot on. Its predictions were very close to the real numbers and it followed the perfect prediction line. On the other hand, XGBoost Baseline performed the worst. It had a negative R^2 and a very high MSE, showing that it could not predict the case trends at all. This shows that only the Time Series model could effectively learn and reflect real-world outbreak patterns.

Analysis of using Apriori:

	antecedents	consequents	support	confidence	lift
3	(low_cases)	(air_dry)	0.746452	0.997853	1.000531
4	(air_dry)	(low_cases)	0.746452	0.748456	1.000531
7	(temp_moderate)	(air_dry)	0.029719	1.000000	1.002685
8	(wind_strong)	(air_dry)	0.167336	1.000000	1.002685
12	(high_cases)	(temp_high)	0.245783	0.975558	1.010457
...	...	...	...	...	...
388	(wind_weak, low_cases, temp_moderate)	(humidity_low, air_dry)	0.019009	1.000000	1.002685
389	(temp_moderate, humidity_low)	(wind_weak, low_cases, air_dry)	0.019009	0.639640	1.022274
390	(temp_moderate, air_dry)	(wind_weak, low_cases, humidity_low)	0.019009	0.639640	1.019656
391	(wind_weak, temp_moderate)	(low_cases, humidity_low, air_dry)	0.019009	0.865854	1.159958
393	(temp_moderate)	(wind_weak, low_cases, humidity_low, air_dry)	0.019009	0.639640	1.022274

189 rows × 5 columns

The Apriori analysis showed some clear links between weather and COVID-19 cases. For example, when it is cold and windy, case numbers are usually high. That matches the rule that low temperatures *(temp_low)* and strong wind speeds (*wind_strong)* lead to an increase in cases *(high_cases)*. On the other hand, when it's humid but dry in the air (high humidity, low absolute humidity), we often see fewer cases. This lines up with what scientists says, that high humidity can help slow the spread of airborne viruses. These patterns could help flag high-risk weather days and support early public health warnings.

9. Conclusion & Discussion

From the Anova analysis, it suggests that temperature and geographical location (continent) are associated with statistically significant differences in cumulative COVID-19 cases. Wind speed does not appear to be a significant factor. This makes sense since most cases are transmitted by being in proximity to an infected person in an indoor space, rather than getting it solely from the air. While successful airborne transmission over distances is possible under certain conditions, it is less likely than close-range interaction. Humidity on the other hand shows a potential trend.

The significant continental differences show possible relationships with cumulative COVID-19 cases. However, further research with larger datasets and a longer observation period will be ideal to provide a better picture, especially considering that in the early stages of the spread, cases primarily appeared near the initial outbreak location rather than a uniform distribution across the continents.

In model development, we tested four models to predict daily COVID-19 cases using weather data such as temperature, relative humidity, absolute humidity, and wind speed.

These environmental factors were chosen because they are known to affect how the virus spreads. For example, cold weather and strong wind can raise transmission, while high humidity might lower it.

The Time Series Regression model worked the best out of the four models. It had an R^2 of 0.9907 and the lowest MSE. It used past case numbers, which helped it learn the pattern of how cases change over time. The prediction errors were small and consistent. Linear Regression underperformed, with an R^2 of -0.0268. It had trouble with the complex patterns, maybe it was due to lack of data of past case numbers. Also, XGBoost (Baseline) performed poorly. It had a R^2 of -2.3271, and it had the highest error. This may mean it overfits the training data or couldn't generalize well. Even after tuning, the model didn't improve much with a R^2 = -0.0233. So, just changing the settings wasn't enough.

In the end, Time Series Regression gave the most accurate and useful results because it matched the nature of the problems. It showed that using both weather and past case trends is key when predicting how a disease like COVID-19 spreads. We also looked at Apriori rule mining, and the results backed up what we saw in real life. Cold temperatures and strong wind speed often showed up on days with more COVID-19 cases. On the other hand, higher humidity seemed to help slow the spread. These types of patterns match what we already know about how the virus behaves in different weather.

10. Limitations & Future Work

As previously stated in the conclusion, longer observation periods will definitely boost our model's and analysis ability to discern long-term trends between environmental factors and cumulative COVID-19 cases. Our current analysis and associations might fluctuate or even reverse over the time span of our data, as countries start regulating outing, enforcing quarantine, and rolling out vaccines. Future research will aim to take into consideration longitudinal data to account for these dynamic effects and provide a better picture of the influence of environmental factors with close monitoring on how governmental and public health interventions/policies's effect on the spread of COVID-19.

Presentation Slides:

https://www.canva.com/design/DAGjJIhwPHw/D-GCkoBWj3-MLVrAFWTmhQ/edit?utm_content=DAGjJIhwPHw&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton

11. References

Bashir, M. F., Ma, B., Bilal, Komal , B., Bashir , M. A., Tan, D., & Bashir, M. (2020, April 20). *Correlation between climate indicators and covid-19 pandemic in New York, USA*. Correlation between climate indicators and COVID-19 pandemic in New York, USA. https://www.sciencedirect.com/science/article/pii/S0048969720323524?via%3Dihub

Hyndman, R. J., & Athanasopoulos, G. (n.d.). *Forecasting: Principles and practice (2nd ed)*. Chapter 5 Time series regression models. https://otexts.com/fpp2/regression.html

Manishsiq. (2023, June 28). *Humidity, types, effects, absolute, specific & relative humidity*. StudyIQ. https://www.studyiq.com/articles/humidity/

Zhu, Y., Xie, J., Huang , F., & Cao, L. (2020, April 15). Association between short-term exposure to air pollution and COVID-19 infection: Evidence from China. https://www.sciencedirect.com/science/article/pii/S004896972032221X?via%3Dihub