**Data-Driven Insights Into COVID-19 Deaths in the U.S**

Catherine Smereena Dommaty

ABSTRACT

This study conducts a comprehensive analysis of COVID-19 mortality rates, focusing on key predictors and trends across various demographics and timeframes. Utilizing a dataset encompassing multiple attributes, including year, month, age, total deaths, and specific causes such as pneumonia and influenza, the research identifies significant patterns and relationships. Preliminary data analysis reveals that older age groups consistently exhibit higher mortality rates, with notable surges during pandemic peaks. Gender-based analysis indicates disparities, with males experiencing higher mortality rates compared to females. Temporal trends highlight a peak in mortality rates in 2021, followed by a decline in subsequent years, suggesting the impact of public health interventions and vaccination efforts. These findings underscore the importance of targeted healthcare strategies for vulnerable populations and provide a foundation for predictive modeling to inform future public health policies.

INTRODUCTION

Mortality rates are key indicators of public health and healthcare outcomes. The COVID-19 pandemic significantly impacted mortality rates across different age groups and demographics. This study analyzes mortality trends using data from Data.gov, specifically the "Provisional COVID-19 Deaths by Sex and Age" dataset.

This dataset provides insights into the number of COVID-19-related deaths across various age groups, genders, and time periods in the United States. Understanding these trends can help policymakers, researchers, and healthcare professionals assess the impact of the pandemic and develop strategies to mitigate future public health crises.

The objective of this analysis is to identify patterns in mortality rates based on demographic factors, evaluate excess mortality, and determine if this dataset is suitable for further research. Through statistical summaries and visualizations, we aim to extract meaningful insights that contribute to a broader understanding of the pandemic's impact.

DATASET DESCRIPTION

The dataset, sourced from Data.gov, is titled "Provisional COVID-19 Deaths by Sex and Age" and provides information on mortality attributed to COVID-19 across different demographic categories in the United States. The dataset is updated periodically and includes provisional death counts based on death certificate data reported to the National Center for Health Statistics (NCHS).

Variables and Descriptions:

1. Data As Of: The date when the dataset was last updated.
2. Start Date: The beginning date for the reported deaths in that specific entry.
3. End Date: The ending date for the reported deaths in that specific entry.
4. Group: Categorization of the data (e.g., by age, sex, or overall).
5. Year: The year of recorded deaths.
6. Month: The month of recorded deaths.
7. State: The U.S. state where the deaths occurred (if applicable).
8. Sex: The gender category (Male, Female, or Total).
9. Age Group: Age category for reported deaths (e.g., 0-4 years, 65-74 years).
10. COVID-19 Deaths: The number of deaths attributed to COVID-19 within the specified group.
11. Total Deaths: The total number of deaths recorded in the specified category.
12. Pneumonia Deaths: The number of deaths attributed to pneumonia.
13. Influenza Deaths: The number of deaths attributed to influenza.
14. Pneumonia and COVID-19 Deaths: The number of deaths involving both pneumonia and COVID-19.
15. Influenza and COVID-19 Deaths: The number of deaths involving both influenza and COVID-19.
16. Significance of the Variables
17. Age Group & Sex: These allow for demographic comparisons of COVID-19 mortality.
18. COVID-19 Deaths & Total Deaths: Help assess the proportion of deaths directly attributed to COVID-19.
19. Pneumonia & Influenza Deaths: Provide insights into co-occurring respiratory illnesses.
20. State & Time-Based Data: Enable geographic and temporal trend analysis.

Sub-group Analysis:

1. By Sex (Male vs. Female)
2. By Year (e.g., 2020, 2021, 2022)
3. By Age Group (e.g., Under 1 year, 5-14 years, 65-74 years)

Predictions & Forecasting:

1. Forecast COVID-19 deaths over time
2. Predict mortality trends based on Age Group, Sex, and Year

EXPLORATORY DATA ANALYSIS (EDA)

EDA was conducted to understand the distribution of variables, detect anomalies, and identify relationships within the dataset. The descriptive statistics provided key insights into the dataset’s characteristics, such as mean values, standard deviations, and data ranges.

1. Age-Related Disparities in COVID-19 Mortality Rates​:

The analysis of COVID-19 mortality rates across different age groups reveals a pronounced correlation between advancing age and increased fatality rates. Figure 1 illustrates the distribution of COVID-19 deaths among various age cohorts. The data indicates that older adults, particularly those aged 65 and above, account for the majority of COVID-19 fatalities. This trend is consistent with national statistics; for instance, in 2023, adults aged 65 and older comprised approximately 87.9% of in-hospital COVID-19 deaths in the United States. The elevated mortality rates among older populations can be attributed to factors such as the increased prevalence of comorbidities and age-related decline in immune function.​

​A graph of a number of people with numbers and a number of people with green and blue bars

Description automatically generated with medium confidence

Figure 1: COVID-19 Deaths by Age Group

2. COVID-19 Mortality by Sex

The analysis of mortality rates by sex reveals a distinct pattern in COVID-19 fatalities. Figure 2 illustrates the mean mortality rate for different sex groups. The results indicate that males have a higher average COVID-19 mortality rate compared to females. This finding aligns with existing research, which suggests that biological, lifestyle, and comorbidity factors contribute to higher mortality risks among men. For example, studies have shown that males are more likely to engage in behaviors such as smoking and have higher rates of comorbid conditions like hypertension, which may exacerbate COVID-19 outcomes.

A graph showing the number of deaths by sex

Description automatically generated

Figure 2: COVID-19 Deaths by Sex

3. Analysis of Mortality by Cause

A comparative analysis of mortality causes was conducted, focusing on COVID-19, pneumonia, and influenza. The data reveals that COVID-19 has been a more significant cause of death compared to influenza. For instance, in Australia, between January and November 2024, there were 3,676 deaths attributed to COVID-19, which is five times higher than the 728 deaths caused by influenza during the same period. This trend underscores the heightened severity and impact of COVID-19 relative to other respiratory illnesses.​

A graph of a disease

Description automatically generated with medium confidence

Figure 3: Total Deaths by Cause

4. COVID-19 Deaths Over Years by Gender

The analysis of COVID-19 mortality rates over time, segmented by gender, provides insights into how fatalities have evolved among males and females throughout the pandemic. Figure 4 presents a line chart depicting the total number of COVID-19 deaths for each gender across different years. The visualization indicates that both genders experienced fluctuations in mortality rates corresponding to various pandemic waves. Notably, during certain periods, male mortality rates were higher than those of females, aligning with existing research that suggests males may be more susceptible to severe outcomes from COVID-19 due to factors such as comorbidities and immune response differences. Understanding these trends is crucial for developing targeted public health interventions aimed at reducing mortality across all demographics.​

A graph of a number of people with different colored lines

Description automatically generated

Figure 4: COVID-19 Deaths Over Years by Gender

Statistical Analysis of COVID-19 Mortality by Gender and Geospatial Distribution

1. Association Between Gender and COVID-19 Mortality

To investigate the relationship between gender and COVID-19 mortality, a Chi-Square test for independence was conducted. This statistical test assesses whether there is a significant association between two categorical variables—in this case, gender and the number of COVID-19 deaths.

Findings:

* Chi-Square Statistic (χ²): 8175.83
* Degrees of Freedom (dof): 1
* P-value: 3.99e-19

The extremely low p-value indicates a statistically significant association between gender and COVID-19 mortality. Specifically, the data suggests that males have a higher mortality rate compared to females. This observation aligns with existing research, which has identified male gender as a predictor of higher mortality in hospitalized COVID-19 patients. Factors contributing to this disparity may include differences in immune response, prevalence of comorbidities, and health-seeking behaviors between genders.

2. Geospatial Distribution of COVID-19 Mortality Across U.S. States

A geospatial analysis was performed to visualize the distribution of COVID-19 deaths across the continental United States. By integrating aggregated mortality data with geospatial information, a choropleth map was created to depict the total number of COVID-19 deaths per state.

Key Observations:

* High Mortality States: States such as New York, California, and Texas exhibited higher total COVID-19 deaths, which may be attributed to their large populations and urban centers.
* Regional Variations: The map revealed regional disparities in mortality rates, highlighting areas that experienced more severe impacts during the pandemic.
* Understanding the spatial distribution of COVID-19 mortality is crucial for public health planning and resource allocation. Identifying regions with higher death tolls can inform targeted interventions and preparedness strategies for future public health emergencies.

A map of the united states

Description automatically generated

Figure 5: Heatmap of COVID-19 Deaths by State in the Continental U.S.

Note: Alaska and Hawaii were excluded from this analysis to focus on the continental U.S.

This geospatial analysis underscores the importance of considering both demographic and geographic factors in understanding and addressing the impacts of COVID-19.

Data Transformation:

1. Age Group Encoding

To facilitate numerical analysis and modeling, the 'Age Group' variable, originally recorded as categorical ranges (e.g., "25-34 years", "85 years and over"), was transformed into a numeric format.

The dataset initially included a variety of age group categories such as “Under 1 year”, “5-14 years”, “45-54 years”, and “85 years and over”, along with an “All Ages” category. Since “All Ages” aggregates data across age brackets and could skew analyses requiring granularity, it was removed to ensure precision in age-specific modeling and interpretation.

To numerically encode age groups, a custom function was applied that extracted the lower bound of each age range and assigned it to a new variable called Age Numeric. For instance:

* “Under 1 year” was encoded as 0,
* “25-34 years” was encoded as 25,
* “85 years and over” was encoded as 85.

This transformation enables better handling of age as a continuous predictor in statistical modeling and visualizations. It also supports more accurate stratification when exploring age-related trends in COVID-19 mortality. The resulting unique values in the new Age Numeric column ranged from 0 to 85, effectively representing the lower threshold of each defined age category.

2. Gender-Based Analysis of COVID-19 Mortality and Pneumonia/Influenza Death Rates

To further investigate the impact of gender on mortality rates, two key metrics were calculated:

* COVID Mortality Rate (%): This represents the percentage of deaths attributed specifically to COVID-19 relative to the total number of deaths.​ Pneumonia/Influenza
* Death Rate (%): This indicates the percentage of deaths where pneumonia, influenza, or COVID-19 were contributing factors, relative to the total number of deaths.​



Table 1: Gender-Based Summary Statistics for Mortality Rates

COVID Mortality Rate (%): The average mortality rate due to COVID-19 is slightly higher in males (3.46%) compared to females (3.20%). This finding aligns with broader studies indicating that males are at a higher risk of severe outcomes and mortality from COVID-19. For instance, data from 73 countries revealed that the infection fatality rate (IFR) was higher in males (3.17%) than in females (2.26%) .​

Pneumonia/Influenza Death Rate (%): Conversely, females exhibit a higher average death rate (30.89%) when considering deaths where pneumonia, influenza, or COVID-19 were contributing factors, compared to males (25.32%). This suggests that while males may have a higher mortality rate directly attributed to COVID-19, females may experience higher mortality when broader respiratory infections are considered.​

3. Yearly Mortality Rates: COVID-19 and Pneumonia/Influenza (2020–2023)

The table below presents a summary of the COVID-19 Mortality Rate (%) and Pneumonia/Influenza Death Rate (%) from 2020 to 2023:

A screenshot of a graph

AI-generated content may be incorrect.

Table 2: Summary Statistics for COVID-19 Mortality Rate (%) and Pneumonia/Influenza Death Rate (%) from 2020 to 2023

COVID-19 Mortality Rate (%): The average mortality rate due to COVID-19 increased from 3.71% in 2020 to 6.12% in 2021, followed by a decline to 2.63% in 2022 and further to 0.46% in 2023.​

Pneumonia/Influenza Death Rate (%): The average death rate associated with pneumonia, influenza, or COVID-19 was 26.87% in 2020, slightly increased to 28.10% in 2021, remained relatively stable at 28.32% in 2022, and decreased to 23.19% in 2023.​

4. Gender-Based Mortality Rates: COVID-19 and Pneumonia/Influenza:

An analysis of mortality rates by gender reveals the following average percentages:​



Table 3: Mortality Rates By Gender

COVID-19 Mortality Rate: The data indicates that males have a higher average COVID-19 mortality rate (3.46%) compared to females (3.20%). This aligns with broader research suggesting that men are at a greater risk of severe outcomes and death from COVID-19. For instance, studies have shown that men have higher odds of death from COVID-19, with fatality rates varying widely by country. ​

Pneumonia/Influenza Death Rate: Conversely, females exhibit a higher average death rate (30.89%) due to pneumonia, influenza, or COVID-19 compared to males (25.32%). This suggests that while males may experience more severe outcomes specifically from COVID-19, females may have a higher mortality rate when considering a broader spectrum of respiratory illnesses.​

5. Mortality Rates by Age Group:

An analysis of mortality rates across different age groups reveals significant variations in the impact of COVID-19 and related respiratory illnesses. The table below presents the average COVID-19 Mortality Rate (%) and Pneumonia/Influenza Death Rate (%) for each age group:​

A screenshot of a table with numbers and a number of people

Description automatically generated

Table 4: COVID-19 Mortality Rate (%) and Pneumonia/Influenza Death Rate (%) by Age Group

COVID-19 Mortality Rate (%): The data indicates a clear trend of increasing COVID-19 mortality rates with advancing age. Individuals aged 85 years and over exhibit the highest average mortality rate at 7.10%.​The mortality rate begins to rise significantly from the 25-34 years age group (1.42%) and continues to increase with age.​Younger age groups, such as 0-17 years, show markedly lower mortality rates (0.08%).​These findings align with national statistics indicating that older adults are disproportionately affected by COVID-19. For instance, in 2023, adults aged 65 and older accounted for approximately 87.9% of in-hospital COVID-19 deaths in the United States. ​

Pneumonia/Influenza Death Rate (%): Interestingly, the highest average death rates due to pneumonia, influenza, or COVID-19 are observed in younger age groups.The 0-17 years age group has a rate of 34.05%, and the 25-34 years group has the highest at 36.16%.​In contrast, older age groups, such as 85 years and over, have lower rates (15.05%).​This pattern may be influenced by several factors, including higher resilience among younger populations leading to lower COVID-19 mortality but higher susceptibility to other respiratory infections. Additionally, reporting practices and the presence of comorbidities in older adults might affect these statistics.​

6. Annual Mortality Rates: COVID-19 and Pneumonia/Influenza (2020–2023)

An examination of mortality rates over the years 2020 to 2023 reveals notable trends in deaths attributed to COVID-19 and pneumonia/influenza. The table below summarizes the average mortality rates for each year:​

A screenshot of a graph

Description automatically generated

Table 5: COVID-19 Mortality Rate (%) by Year

COVID-19 Mortality Rate (%): The data indicates a peak in COVID-19 mortality rates in 2021, with an average rate of 6.12%. This increase may be attributed to factors such as the emergence of new variants and the initial challenges in vaccine distribution. Following this peak, there is a significant decline in mortality rates in subsequent years, dropping to 2.63% in 2022 and further to 0.46% in 2023. This downward trend aligns with reports indicating a sharp decline in COVID-19 deaths during this period. For instance, projections for 2023 estimated under 70,000 deaths, compared to 246,166 in 2022 and 463,267 in 2021. ​

Pneumonia/Influenza Death Rate (%): The mortality rates for pneumonia and influenza remained relatively stable from 2020 through 2022, averaging around 28%. However, a notable decrease to 23.19% is observed in 2023. This reduction may reflect the overall decline in respiratory-related deaths and could be influenced by factors such as increased vaccination rates and improved public health interventions.​

Trends in COVID-19 Mortality Rates Over Time:

An analysis of COVID-19 mortality rates from 2020 to 2023 reveals significant fluctuations, reflecting the evolving impact of the pandemic. The data indicates a peak in mortality rates in 2021, followed by a notable decline in subsequent years.​

Key Observations:

* 2020: The initial outbreak year recorded a substantial mortality rate as the world grappled with understanding and responding to the novel virus.​
* 2021: Mortality rates peaked during this year, likely due to factors such as the emergence of more transmissible variants and the time required to achieve widespread vaccine distribution. This trend aligns with reports indicating that COVID-19 deaths were highest in 2021, with approximately 463,267 fatalities in the United States. ​
* 2022: A significant reduction in mortality rates occurred, corresponding with increased vaccination coverage and improved treatment protocols. The number of COVID-19 deaths in the U.S. decreased to 246,166 during this year. ​
* 2023: The downward trend continued, with mortality rates reaching their lowest point in the observed period. Provisional data suggests that COVID-19 deaths in the U.S. were projected to be under 70,000 for the year. ​

A graph with red line

Description automatically generated

Figure 6: COVID-19 Mortality Rates Over Time

Trends in Pneumonia and Influenza Mortality Rates Over Time

An analysis of mortality rates due to pneumonia and influenza from 2020 to 2023 reveals notable trends in respiratory-related deaths.​

Key Observations:

* 2020: The combined mortality rate for pneumonia and influenza was 26.87%. This elevated rate may reflect the heightened impact of respiratory illnesses during the initial phase of the COVID-19 pandemic.​
* 2021: A slight increase to 28.10% was observed, potentially due to overlapping seasonal influenza and pneumonia cases alongside COVID-19 infections.​
* 2022: The mortality rate remained relatively stable at 28.32%, indicating a persistent burden of these respiratory diseases.​
* 2023: A notable decline to 23.19% suggests improvements in public health interventions, increased vaccination uptake, and possibly reduced transmission rates of pneumonia and influenza.​

A graph with a blue line

Description automatically generated

Figure 7: Pneumonia and Influenza Mortality Rates Over Time

COVID-19 Mortality Rates by Gender

An analysis of COVID-19 mortality rates segmented by gender reveals notable disparities between male and female populations.​

Key Observations:

* Male Mortality Rates: Studies have consistently shown that men experience higher COVID-19 mortality rates compared to women. For instance, data from the New York State Department of Health indicated that approximately 60% of COVID-19 fatalities were among males. ​
* Female Mortality Rates: While women have lower mortality rates from acute COVID-19 infections, emerging research suggests they may be at a higher risk for developing long-term symptoms, known as long COVID. A study led by UT Health San Antonio found that females, particularly those in their 40s, have up to a 45% higher risk of developing long COVID compared to males.

A graph of a number of people with different colored squares

Description automatically generated with medium confidence

Figure 8: COVID-19 Mortality Rates by Gender

DATA PREPARATION & ANALYSIS

To prepare the dataset for predictive modeling, we followed a structured data preprocessing pipeline aimed at improving data quality, ensuring interpretability, and avoiding data leakage. The dataset, which included weekly mortality counts by sex and age in the U.S. between 2020 and 2023, required several transformations to support regression modeling.

Preprocessing Steps

1. Data Cleaning and Column Selection  
   We removed the “All Ages” category to avoid duplication and aggregation errors, ensuring that mortality counts remained consistent and non-overlapping. We also dropped any footnote-related columns that did not contribute meaningful information to the modeling process.
2. Age Group Transformation  
   The “Age Group” column, initially represented as categorical text ranges, was converted into a numeric column by assigning each range its midpoint value. This transformation allowed age to be treated as a continuous variable in regression models.
3. Feature Engineering  
   We engineered multiple new variables to improve model performance and add interpretability:

* COVID Mortality Rate (%) = (COVID-19 Deaths / Total Deaths) × 100
* Pneumonia and Influenza Mortality Rate (%)
* Excess Mortality = Total Deaths – (COVID + Pneumonia + Influenza Deaths)
* Interaction terms such as Age × Pneumonia and Pneumonia × Influenza were introduced to capture potential compounding effects.

1. Handling Missing Values  
   We used mean imputation for numeric columns containing missing values, applying it separately to the training and test sets to avoid data leakage. This ensured consistent scale and preserved sample size.
2. Outlier Removal  
   We removed extreme outliers from the training set using a Z-score threshold of 3, targeting numeric columns such as Age, Excess Mortality, and the calculated mortality rates. This step improved model stability and reduced noise.
3. Encoding and Scaling  
   Gender was converted to a binary numeric feature (0 for Female, 1 for Male), although it was later removed after being identified as non-predictive through LASSO feature selection. All numeric features were standardized using StandardScaler to optimize model convergence and ensure comparability across variables.

Modeling Approach

We used a Random Forest Regressor, a non-linear ensemble model known for its robustness, ability to model complex feature interactions, and resistance to overfitting. We performed hyperparameter tuning using GridSearchCV with 3-fold cross-validation, testing various configurations of n\_estimators, max\_depth, and min\_samples\_leaf to optimize generalization.

Final Model Evaluation

The final Random Forest model was evaluated using standard regression performance metrics on the test set:

* R² Score: 0.8473: The model explained approximately 84.7% of the variance in COVID-19 mortality rate.
* Mean Squared Error (MSE): 0.1498: Average squared prediction error, indicating low model error overall.
* Root Mean Squared Error (RMSE): 0.3870: On average, the predicted mortality rate was off by just 0.39 percentage points.
* Mean Absolute Error (MAE): 0.1608: The model’s predictions differed from actual values by an average of only 0.16%.

These metrics collectively indicate strong model performance with minimal bias or overfitting. Diagnostic plots, including actual vs. predicted and residual plots, confirmed that predictions were well-aligned with actual values and errors were randomly distributed without visible patterns, supporting the model’s reliability.

Feature Importance and Residual Analysis:

1. An analysis of feature importance revealed that Pneumonia Deaths and Age Numeric were the most significant predictors of COVID-19 mortality rates. This aligns with previous findings that older populations and those with respiratory complications faced higher mortality risks.  
   A graph of a number of people with pneumonia and covid-19

   Description automatically generated  
   Figure 9: Feature Importance in COVID-19 Mortality Prediction
2. Feature importance analysis helps determine which variables have the most significant impact on predicting COVID-19 mortality rates. The Random Forest model assigned the highest importance scores to Pneumonia Deaths, Age Numeric, and Influenza Deaths, indicating their strong influence on mortality trends. This finding reinforces the notion that individuals with respiratory conditions are at a significantly higher risk.  
   A graph with blue dots and a red line

   Description automatically generated  
   Figure 10: Actual vs. Predicted COVID-19 Mortality Rates
3. This scatter plot visually compares the actual and predicted mortality rates. A strong alignment between the actual and predicted values suggests that the model performs well in capturing the variability of COVID-19 mortality trends. The closer the points are to the diagonal reference line, the more accurate the model’s predictions. The presence of some deviations indicates areas for potential model refinement.A graph of a graph showing a plot of covid

   AI-generated content may be incorrect.  
   Figure 11: Residual Plot for Model Evaluation  
   The residual plot provides insights into the accuracy of the predictions by displaying the difference between actual and predicted values. Ideally, residuals should be randomly distributed around zero, indicating that the model does not exhibit systematic errors. In this analysis, the residuals appear to be well-distributed, suggesting that the model has captured the underlying patterns effectively without significant bias.

CONCLUSION

This project set out to explore which factors most accurately predict COVID-19 mortality rates in the United States using publicly available CDC data. Through a combination of data preprocessing, feature engineering, and predictive modeling, we were able to build a reliable forecasting model and draw meaningful insights from demographic and cause-specific mortality trends.

After testing both linear and non-linear approaches, we found that Random Forest Regression offered the best performance, capturing over 84% of the variance in COVID-19 mortality rates. Key predictors included age, pneumonia and influenza deaths, and overall excess mortality — variables that align closely with existing public health knowledge around COVID-19 comorbidities and age-based vulnerability. Our results were supported by strong model evaluation metrics and well-behaved residuals, confirming both the accuracy and stability of the model.

In terms of practical implications, our analysis reinforces the importance of prioritizing older populations and those with respiratory complications in both prevention and care strategies. Additionally, understanding mortality beyond confirmed COVID-19 deaths — through excess mortality measures — can help uncover indirect effects of the pandemic.

Future work could involve incorporating regional data, healthcare capacity, or vaccination status to deepen the model’s insights. Exploring additional algorithms such as XGBoost or time series models could also be valuable for forecasting mortality trends over time.

Overall, this study demonstrates the power of data-driven analysis in identifying key risk factors and informing policy decisions during a public health crisis. The framework developed here could be adapted to monitor and predict outcomes in future health emergencies.

REFERENCES

1. Data.gov. (2023, September 29). *U.S. Department of Health & Human Services - Provisional COVID-19 Deaths by Sex and Age*. https://catalog.data.gov/dataset/provisional-covid-19-death-counts-by-sex-age-and-state
2. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357–362. <https://doi.org/10.1038/s41586-020-2649-2>
3. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & van der Walt, S. J. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3), 261–272. <https://doi.org/10.1038/s41592-019-0686-2>
4. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785