Geographical Variations in COVID-19 Mortality: A Multi-Regression Analysis

(Exploring the Impact of Age and Comorbidities Across the United States)

Video link:

https://www.youtube.com/watch?v=jPhzJFWevQI

* Note: The code using the original dataframe "../data/Conditions_Contributing_to_COVID-19_Deaths__by_State_and_Age__Provisional_2020-2023_20231016.csv" related to this paper is included as a pdf in the narrative folder of this project. A sample of this data frame was taken after writing this paper, and replaced in the code to align with gradscope submissions. The sample is included in the data folder of this project as "../data/Conditions_Contributing_to_COVID-19_Deaths__by_State_and_Age__Provisional_2020-2023_20231016_sample.csv". If the notebook is run again the results may differ due to the smaller sample size of data. A README.md is included in the data folder detailing this.

Abstract:

This study investigates the intricate relationship between age, comorbidities, and COVID-19 mortality rates across the United States by leveraging multiple regression models. Utilizing extensive datasets from the National Center for Health Statistics, we executed a comprehensive exploratory data analysis followed by predictive modeling to discern patterns and predict mortality risks. Our research delves into the role of age and comorbidities as the determinant factors in COVID-19 mortality, where findings indicate a heightened risk associated with increased age, supported by statistical models across all states except Maine and Vermont. The study employs Lasso, Linear, and Logistic Regression techniques, with Principal Component Analysis (PCA) further refining the model's accuracy by addressing multicollinearity and enhancing feature selection. The results demonstrate that while Logistic Regression models yield high accuracy, the potential for overfitting is noted, emphasizing the importance of cross-validation in model assessment. Our findings offer valuable insights

for policymakers and healthcare providers, emphasizing the need for targeted public health interventions to protect vulnerable age groups. Additionally, we propose future research to examine the temporal progression of the pandemic through longitudinal data analysis, which could reveal more nuanced age-related mortality trends. This research not only contributes to the current understanding of the pandemic's demographic impact but also guides evidence-based policy-making for future public health challenges.

Project Introduction and Goals:

Our project seeks to delve deeply into the relationship between age, comorbidities and COVID-19 mortality across different states. By employing exploratory data analysis and subsequent predictive modeling, we intend to harness critical insights to aid policymakers, healthcare professionals, and the general public in understanding the inherent risks associated with age and COVID-19 in the U.S. context. We will answer the following questions:

Can we discern a relationship between age and COVID-19 mortality by state?

Can the likelihood of mortality from COVID-19 based on age and comorbidities be predicted?

Answering these questions are pivotal in enhancing the comprehension of COVID-19's impact across demographic lines, equipping stakeholders with essential data to formulate age and condition specific health interventions and policies; ultimately aiming to reduce mortality rates and possibly safeguard vulnerable populations within the United States. The paper by Mueller et al., suggests that younger individuals exhibit a more controlled immune response, enabling them to effectively combat COVID-19 upon exposure. In contrast, older people tend to have a more subdued immune response, leading to prolonged and often more challenging battles against the virus. While it remains uncertain whether the higher COVID-19 mortality rates in older ages are due to pre-existing conditions or those developed post-exposure to the virus, it is evident that comorbidities significantly contribute to increased mortality rates among the elderly (Mueller et al., 2020).

Description of the Data:

The datasets anchoring our related research comprise the "Provisional COVID-19 Deaths by Sex and Age"(NCHS/DVS) and "Conditions Contributing to COVID-19 Deaths, by State and Age (Provisional 020-2023)"(NCHS/DVS).

The file paths to the datasets are as followed, respectively:
"../data/Provisional_COVID-19_Deaths_by_Sex_and_Age_20231016.csv"

"../data/Conditions_Contributing_to_COVID-19_Deaths__by_State_and_Age__Provisional_2020-2023_20231016.csv".

Both of these invaluable datasets were procured from the esteemed National Institute of Health, meticulously created by the National Center for Health Statistics from the Division of Vital Statistics. They span data collated from 1/1/2020 to 9/27/2023. These are very large data sets, the first being 62100 rows (data points) by 14 columns (features) and the second being 13770 rows by 16 columns. These datasets are comprehensive, documenting information ranging from the specific start and end dates of data collection across almost four years of data, year and month of entry, state details, age groups, gender specifics, to granular data about COVID-19-related deaths and deaths from other conditions or comorbidities, like Pneumonia and Influenza. Central to our inquiry are questions surrounding the discernible patterns of age-related COVID-19 mortality across states and the potential to predict the likelihood of such mortality based on age. Given the public health implications of our findings, the results could be instrumental in fine-tuning strategies and optimizing resource allocation at the state level.

To ensure the integrity and relevance of our data, our initial steps involved rigorous data cleaning to narrow down on the most pertinent features aligning with our research goals. This was followed by an exploratory data analysis phase, where visual tools assisted us in deciphering underlying patterns and in deciding which features to use. As we transitioned into the modeling phase, our modeling choices were to use

Lasso Regression, similarly to the one described in *Development of a Model by Lasso to Predict Hospital Length of Stay (LOS) in Patients with the SARS-COV-2 Omicron Variant* (Zhang et al. , 2023). For comparison purposes, we also used Linear Regression and Logistic Regression models. The inspiration to compare both models of logistic and lasso regression was drawn from the approach utilized in the study "Risk factors for severe COVID-19 differ by age for hospitalized adults" (Molani et al. in 2022). This choice also stemmed from our desire to assess the collective impact of multiple features, especially age and condition, on COVID-19 mortality. By juxtaposing these features with state data, we believe the regression models will yield insightful coefficients underscoring the relationship magnitude. When conducting our EDA we realized there was the risk of the datasets not capturing the entire spectrum of deaths, possibly leading to underestimations. Additionally, not accommodating other impactful variables, such as pre-existing health conditions, could introduce a degree of bias. Ultimately, our project aspires to glean meaningful patterns about age's role in COVID-19 mortality. The culminating deliverables will include the refined dataset, a series of insightful visualizations, and regression models accompanied by pertinent findings and actionable recommendations.

EDA:

The data cleaning described here and results from our EDA can be seen in the analysis_notebook in the analysis folder. The data sets can be found in the data folder and the figures related to this paper can be generated by running the analysis_notebook and found in the figures folder and at the end of this paper listed as Figures 1-8.

The datasets were procured from the esteemed National Institute of Health, meticulously created by the National Center for Health Statistics from the Division of Vital Statistics and their data collected from Human Health Services across the United States. Although there may be some bias in the collection we may be unaware of, from what we can see there were unknown values in the features that could be contributed to the bias. The granularity of the original data is segmented by time, demographics, geographical location and health conditions, we reduced this granularity to focus on death totals per state total and by month over the given time frame. The data sets are composed of a variety of data types and before our clean up we had to explore them to understand what values were important to our research. We dropped columns that did not provide value, replaced “NaN” missing values with unknown or 999999, filtered and made some data lowercase so when we concat the data frames we would not run into any issues. There were outliers in the data, either 0 or higher death counts for certain age groups. When we concat the data frames, we noticed some of the columns were missing numbers, in order to combat that we created the pipeline before modeling to impute the data, it is further discussed ahead.

Some questions we wanted to explore with our visualizations include: What is the relationship/correlation between age group, and death in the United States? What is the relationship between age group comorbidities and death in each state? What is the distribution of death by age in each state? After cleaning the data sets and concating our dataframes, we visualized the distribution of Covid-19 deaths by age group in the United States, using bar graphs and concluded the data are left skewed with some outliers. To further confirm that our data was left skewed we plotted a KDE plot of Covid-19 deaths distribution by age group in the United States, which can also be referenced in our analysis_notebook . We created a bar plot to demonstrate the number of covid-19 deaths by age group in each state (figure 2) which yielded similar distribution results to the barplots for ages across the united states (left skewed). This data led us to believe that our hypothesis could be validated as we furthered our research. Following this we plotted the relationship between age and mean covid-19 deaths in each state. As we visualized the relationship between variables we noticed that some of the visualizations were not pleasing or readable so we had to go back to our cleaning and then applied mean and log transformations to some of our axes. Then, we visualized the correlation using lmplot. In the first correlation we visualized the relationship between Total Deaths and Deaths by Age Group and got a plot that did not reveal any correlation. We realized there may be more of a correlation if we visualized the relationship using midpoint age and turned each range into a midpoint in a new column in our cleaned dataframe ‘by_state_merge_df’. This resulting dataframe is the final dataframe we used to test our models on. Figure 1 from our EDA (Fig1) visualizes the correlation between midpoint age and covid 19 deaths. Our analysis revealed a positive correlation equal to 0.415, with left-skewed mortality distribution across age groups, with an initial low positive correlation, indicating a tendency toward higher mortality rates in older age demographics; It is noteworthy to mention the outliers may contribute greatly to the low correlation value, potentially skewing data interpretation. The positive correlation prompted us to use midpoint age amongst our feature variables. While we trained our models we realized we should incorporate more features for a more robust analysis, so we conducted a reassessment of the dataset, which included an examination of various health conditions, as well as pneumonia and influenza-related deaths. We plotted the number of Pneumonia and COVID-19 Deaths against Covid-19 deaths (Figure 3), and the number of Influenza and COVID-19 Deaths against Covid-19 deaths (Figure 4), and COVID-19 deaths based on Condtions(Figure 8), to be later used as features in our models. This bears resemblance to the features utilized in the lasso model described in the paper titled "Modeling mortality risk in patients with severe COVID-19 from Mexico" by Cortes-Telles et al. (2023). We then began transitioning into our modeling phase.

Methodology:

Our modeling choices were to use Lasso Regression and, for comparison purposes, we also used Linear Regression and Logistic Regression models for prediction of COVID-19 mortality. To make modeling easier we set up a pipeline that automatically processes the data with the correct transformations to the appropriate features during modeling training and prediction. We were sure to undo any transformations we applied during the EDA portion in order for the dataframe to be processed correctly in the pipeline. In this pipeline we included log transformations, imputations scaling the data and one hot encoding. The log transformation was included to transform non-negative values of the right skewed data (figures 3 and 4 for example). Missing values were imputed with the median value of those features. RobustScaler was also used to scale features and was preferred to handle outliers since it uses interquartile range. One hot encoding was also used for categorical data since machine learning models prefer numerical inputs. The use of scaling and median for imputation helps to minimize the impact of the outliers and missing data since they can skew performance of the models. Overall this pipeline was crucial in preparing the data for effective training on our machine learning models.

After defining our pipeline, the Data was then split into training and testing sets, with cross-validation used to assess how the model is generalizing to the new unseen data that was not used in the training process. A summary of these modeling results can be seen in the table in the next section of this paper. These models were not performing well at all, so we decided to perform Principal Component Analysis (PCA) to potentially improve accuracy and address multicollinearity if any. The results from the modeling with and without PCA can be seen in the next section of this paper. The loss metrics included in the table include MSE and RMSE for Lasso and Linear Regressions to measure the average prediction error and accuracy for Logistic Regression. After PCA both loss metrics were lowered, this suggests a good balance between bias-variance tradeoff. The high accuracy we obtained for Logistic Regression suggested a good balance between bias-variance tradeoff but we suspected overfitting.

It is crucial to use cross-validation to evaluate the models accurately which can prompt us to tune hyperparameters to avoid overfitting. Some additional hyperparameter tuning may benefit our models, for example in the context of performing feature engineering with PCA the number of components had to be reduced from 0.95 to 0.80 to mitigate the overfitting we first observed. Furthermore, our analysis indicates that the logistic regression model is likely overfitting, a topic that will be discussed in more detail in the following section. Other hyper parameters include LassoCV use of cross validation to tune regularization and Logistic Regressions parameter of max_iter to ensure convergence.

In order to address the error produced by the models, we added PCA to the pipeline and retrained the models, leading to a substantial improvement overall. PCA essentially transforms the feature space reducing dimensionality and enhancing each model's performance. The models with PCA were then used to predict mortality risk as age paired with comorbidities, increases for each state(Figures 5 and 6). These figures show that for almost every state there is a positive correlation or a direct relationship between our features and predicted mortality risk from COVID-19. In each state, as age increases mortality risk also increases. The only discrepancies that can be seen in figures 5 and 6 are that Vermont and Maine, which show almost no relationship between age and mortality risk from COVID-19. We hypothesize that there are biases depending on age demographic in these two states. Further improvements in our methodology might include a more granular feature engineering, further tuning of hyperparameters, further tuning of PCA, and exploring other model architectures.

Summary of results:

The results presented in the table below compare the performance of three different types of regression models: Lasso Regression, Linear Regression, and Logistic Regression, with and without Principal Component Analysis (PCA).

Model type	MSE	RMSE	Additional metrics
Lasso Regression	1156847.977687558	1075.5686764161358	
Lasso Regression with PCA	0.11643739736843756 Cross-Validated MSE: 0.16744930216374726	0.3412292445972906 Cross-Validated RMSE: 0.4066381471030618	Cross-Validated R^2: 0.19058831010742988
Linear Regression	1160472.5935546213	1077.2523351353764	
Linear Regression with PCA	0.11602528768981762 Cross-Validated MSE: 0.16473777999657033	0.3406248489024511 Cross-Validated RMSE: 0.398844465871582	Cross-Validated R^2: 0.1437747678563191

Logistic Regression	0.10687960687960688	0.32692446662739527	Accuracy: 0.8931203931203932
Logistic Regression with PCA	Not included	Not included	Accuracy: 0.8925061425061425

Both Lasso Regression and Linear Regression had high Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) before applying PCA. The use of PCA dramatically improved the performance, reducing both the MSE and RMSE substantially. The cross validated R^2 value indicated that the Lasso Regression model with PCA explains about 19.06 % of the variance in the data and the Linear Regression model with PCA explains about 14.38% of the variance. We realized the variance is low and further research must be done to address this.

We determined that the logistic regression model without PCA returned very low MSE and RMSE values, causing us to suspect the model was highly overfitting. In both logistic regression models, the accuracies were extremely close so we did not include the RMSE or MSE for the second model for comparison, you can see the logistic regression model with PCA did not generalize well to the data in Figure 7. This error was not realized until later and we found it could be due to the binning of continuous variables into categories, but was still included in the analysis_notebook for comparison.

The models with PCA performed better because PCA reduces the dimensionality of the data, potentially removing noise and less informative variables. PCA improves the models ability to identify underlying patterns through the variance of the data, mitigating multicollinearity by reducing the feature set, which highly benefited our Lasso and Linear Regression models.We believe Lasso regression enhanced the models accuracy because it prevents overfitting by promoting sparsity, effectively filtering out less significant features, unlike the linear regression model where overfitting was prevalent due to the inclusion of all features. The models with PCA were then used to predict mortality risk as age increases for each state(Figures 5 and 6). These figures show that for almost every state there is a positive correlation or a direct relationship between age, comorbidities and mortality risk from COVID-19. In each state, as age increases mortality risk also increases.

Discussion:

In our exploratory data analysis (EDA), we found potential limitations inherent in the datasets, particularly the risk of underrepresentation of COVID-19 deaths. This underestimation could lessen the true magnitude of the pandemic's impact. Furthermore, the presence of outliers and absence of variables like pre-existing health conditions in our analysis acknowledges an inherent bias, underscoring the complexity of COVID-19 mortality which is influenced by a number of factors beyond just age. Despite these challenges, our project has successfully unearthed significant patterns linking age with COVID-19 mortality. These findings have been condensed into insightful visualizations that elucidate the nuances of this relationship across different states. Upon completion of modeling over all, we were surprised to find that doing PCA reduced our MSE and RMSE substantially for both Lasso and Linear Regression. We were happy to see such great results related to our research when plotting the Predicted risk vs midpoint age for both Lasso and Linear Regression (Figure 5 and 6), where almost all states had a positive correlation, or a direct relationship between age, comorbidities and risk, verifying our research questions. Logistic Regression may have been done wrong as discussed in the summary of results, considering the binning of deaths and their high, similar accuracies. Getting a high accuracy, we kept the logistic regression but can actually see how wrong the predictions were in Figure 7, showing no relationship when predicting risk. This is an area for further exploration we can improve upon.

Since we are taking a Machine Learning course concurrently with this Data Science course, we attempted to train an artificial neural network (ANN) as an additional predictive model for fun. Adjusting the hyperparameters like learning rate and number of Epochs led us to some good results in our second model. In our analysis_notebook you can see that the second model was learning effectively as the training and test loss consistently declined and did not have a significant gap between themselves, indicating the model is generalizing well to the validation data and not overfitting. Although, when we looked at the MSE and RMSE for the ANN we saw that it might not be performing as well as we thought, highlighting the importance of using multiple metrics for cross validation. It is possible the high error could be attributed to the fact that we did not use PCA in this portion.

The use of predictive models in our research not only shed light on the statistical significance of age and comorbidities in COVID-19 mortality rates but also offers a foundation for further research given the areas of improvement we encountered. Overall, we found that across all states (with the exception of Maine and Vermont), as age increases the predictive risk also increases(fig 5, 6), validating our research questions. This was expected as it is common in many real-word scenarios where the risk of death from comorbidities increases with age. The culmination of this study provides valuable insights and actionable recommendations that could be instrumental for policymakers, healthcare providers, and public health experts. In terms of public health policy and planning, the discovery of age-based predictors for COVID-19 mortality in our study offers crucial guidance for crafting targeted, evidence-based interventions. These strategies are key to providing more effective protection for vulnerable groups in the future. By highlighting age and comorbidities as pivotal factors in COVID-19 mortality, our research paves the way for targeted interventions and resource allocation strategies that can more effectively address the needs of the most vulnerable populations. Therefore, this study critically contributes to the understanding and combating of the COVID-19 pandemic. Furthermore, based on the given peer review, future research could extend this study by analyzing longitudinal trends of the pandemic. Our current analysis, which dropped columns related to dates, years, and months, is based on aggregated data from 2020 to 2023. To refine and enhance the specificity of our results, we propose evaluating the evolution of COVID-19 spread over time, alongside death trends within distinct time periods. This approach could uncover more detailed age-based mortality trends, offering deeper insights into the pandemic's impact.

Citations:

Cortes-Telles, Arturo, et al. “Modeling Mortality Risk in Patients with Severe COVID-19 from Mexico.” *Frontiers*, Frontiers, 8 May 2023, www.frontiersin.org/articles/10.3389/fmed.2023.1187288/full.

Molani, Sevda, et al. “Risk Factors for Severe COVID-19 Differ by Age for Hospitalized Adults.” *Scientific Reports*, U.S. National Library of Medicine, 28 Apr. 2022, www.ncbi.nlm.nih.gov/pmc/articles/PMC9050669/.

Mueller, Amber L, et al. “Why Does Covid-19 Disproportionately Affect Older People?” *Aging*, U.S. National Library of Medicine, 31 May. 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC7288963/#:~:text=An%20individual%20with%20a%20biological,risk%20for%20COVID%2D19%20fatality.

NCHS/DVS. “Conditions Contributing to Covid-19 Deaths, by State and Age, Provisional 2020-2023.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 27 Sept. 2023, data.cdc.gov/NCHS/Conditions-Contributing-to-COVID-19-Deaths-by-Stat/hk9y-quqm.

NCHS/DVS. “Provisional Covid-19 Deaths by Sex and Age.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 27 Sept. 2023, data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Sex-and-Age/9bhg-hcku.

Zhang, Jianxiang, et al. “Development of a Model by Lasso to Predict Hospital Length of Stay (LOS) in Patients with the SARS-COV-2 Omicron Variant.” *Virulence*, U.S. National Library of Medicine, Dec. 2023, www.ncbi.nlm.nih.gov/pmc/articles/PMC10101656/.

Figures:

Figure 1 (fig. 1)

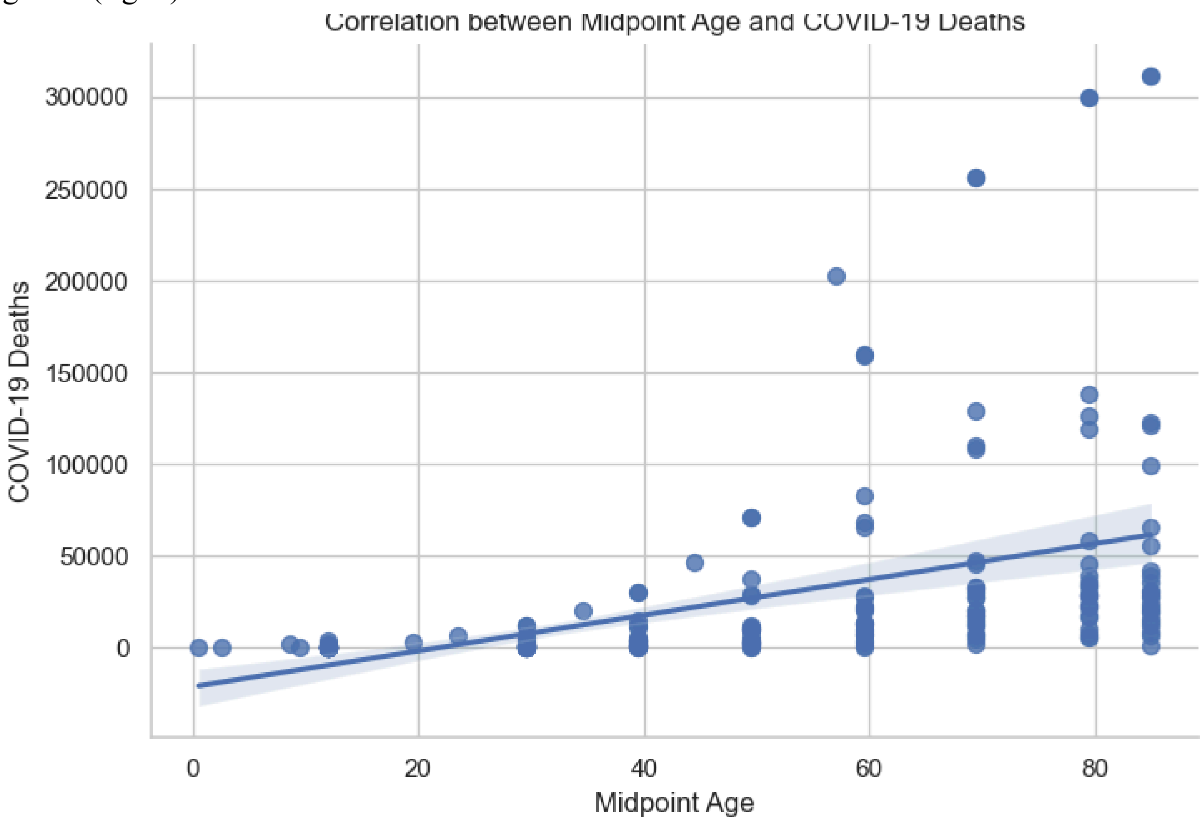


Figure 1- describes the correlation between Midpoint Age and Number of COVID19 deaths.

Figure 2 (fig. 2)

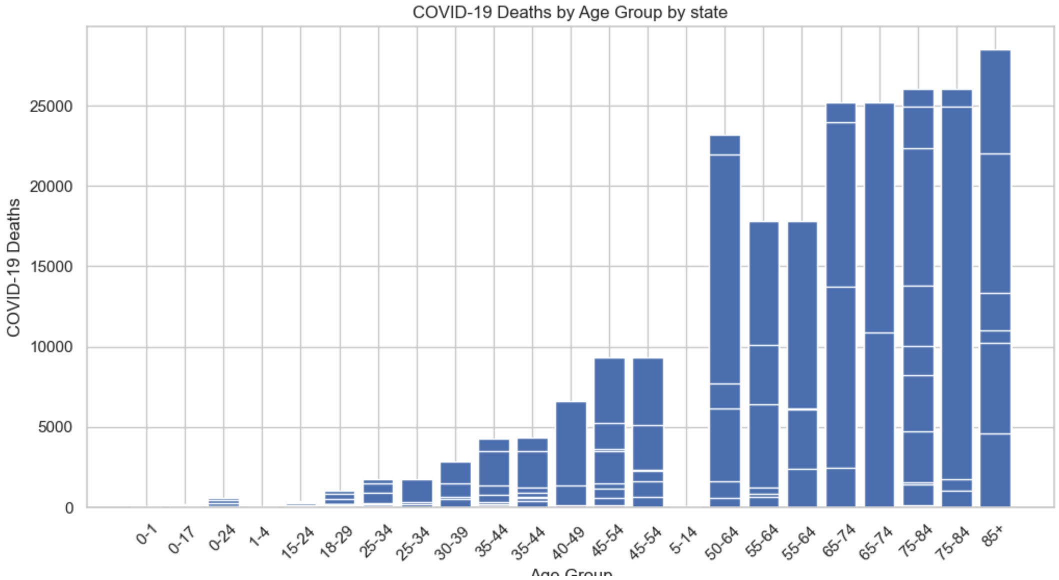


Figure 2- describes the number of COVID19 deaths from each age group by state, left skewed.

Figure 3 (fig. 3)

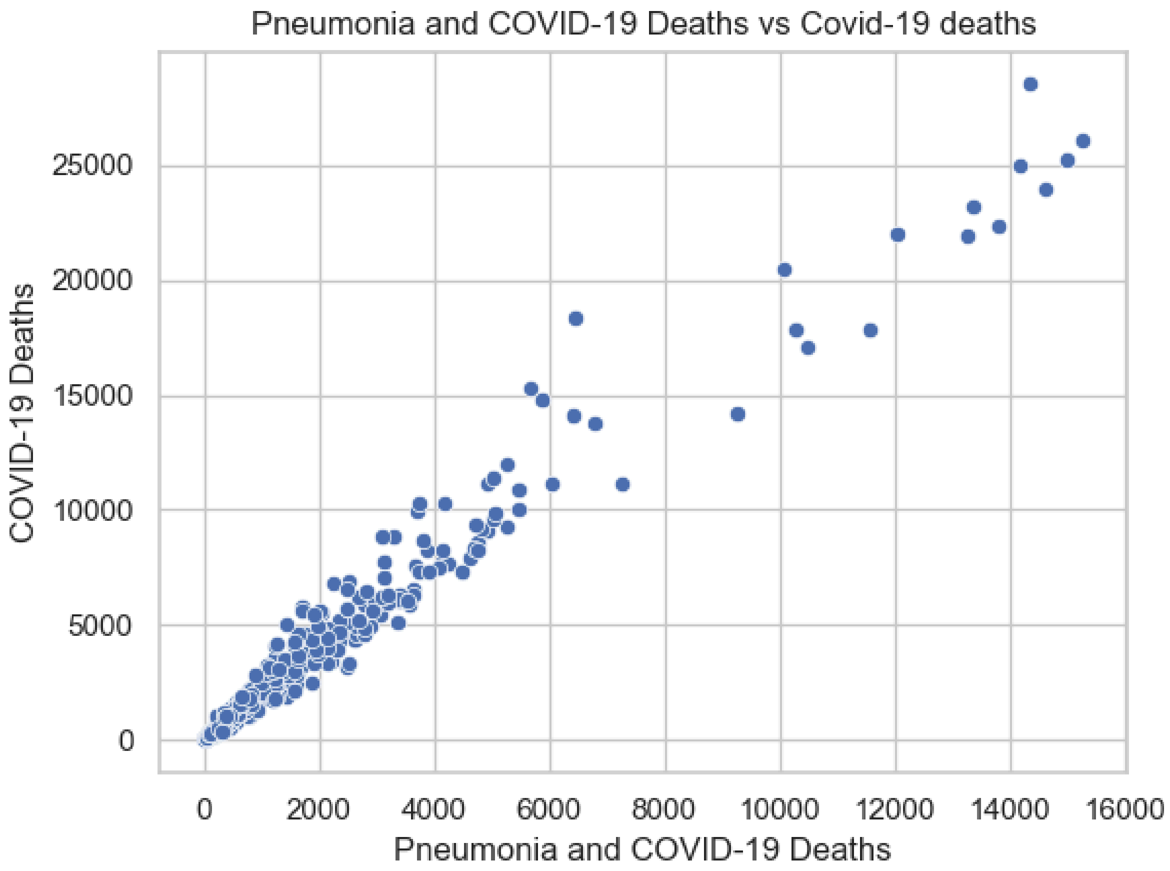


Figure 3- describes the number of COVID19 deaths in relation to Pneumonia as one of the conditions of the COVID19 death.

Figure 4 (fig. 4)

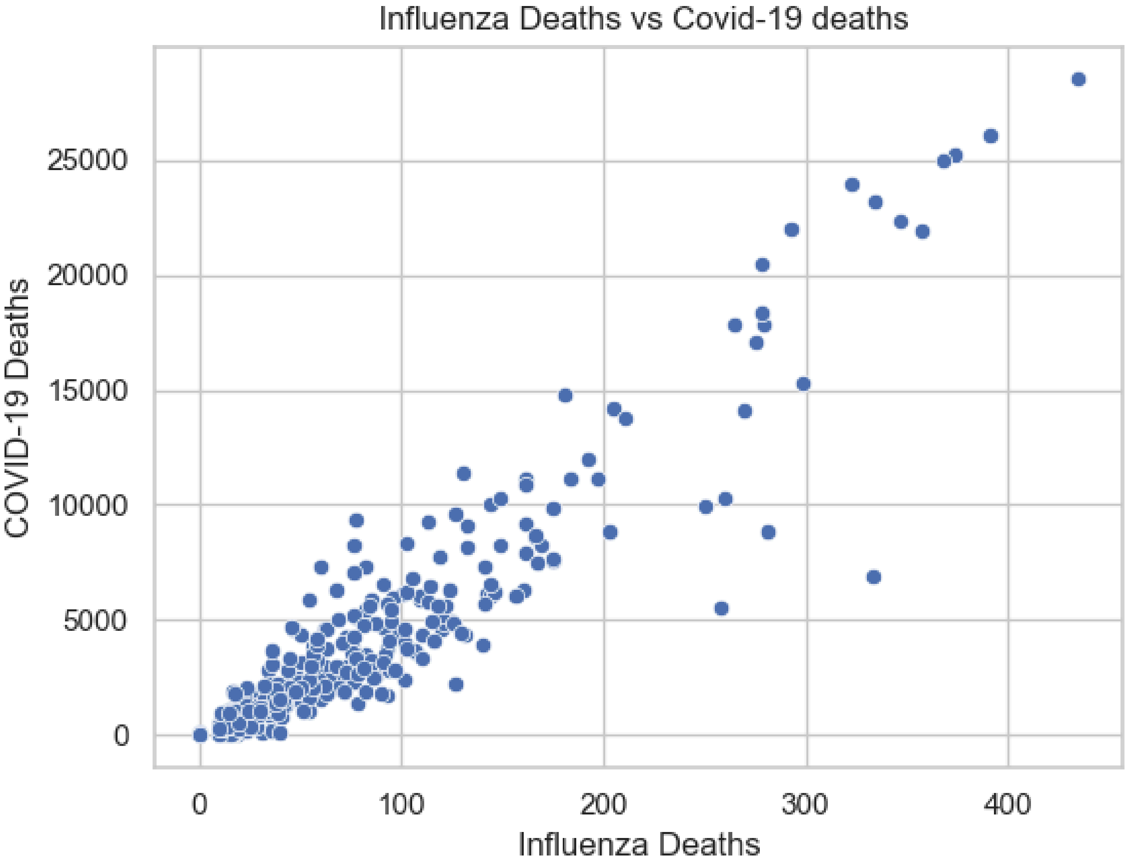
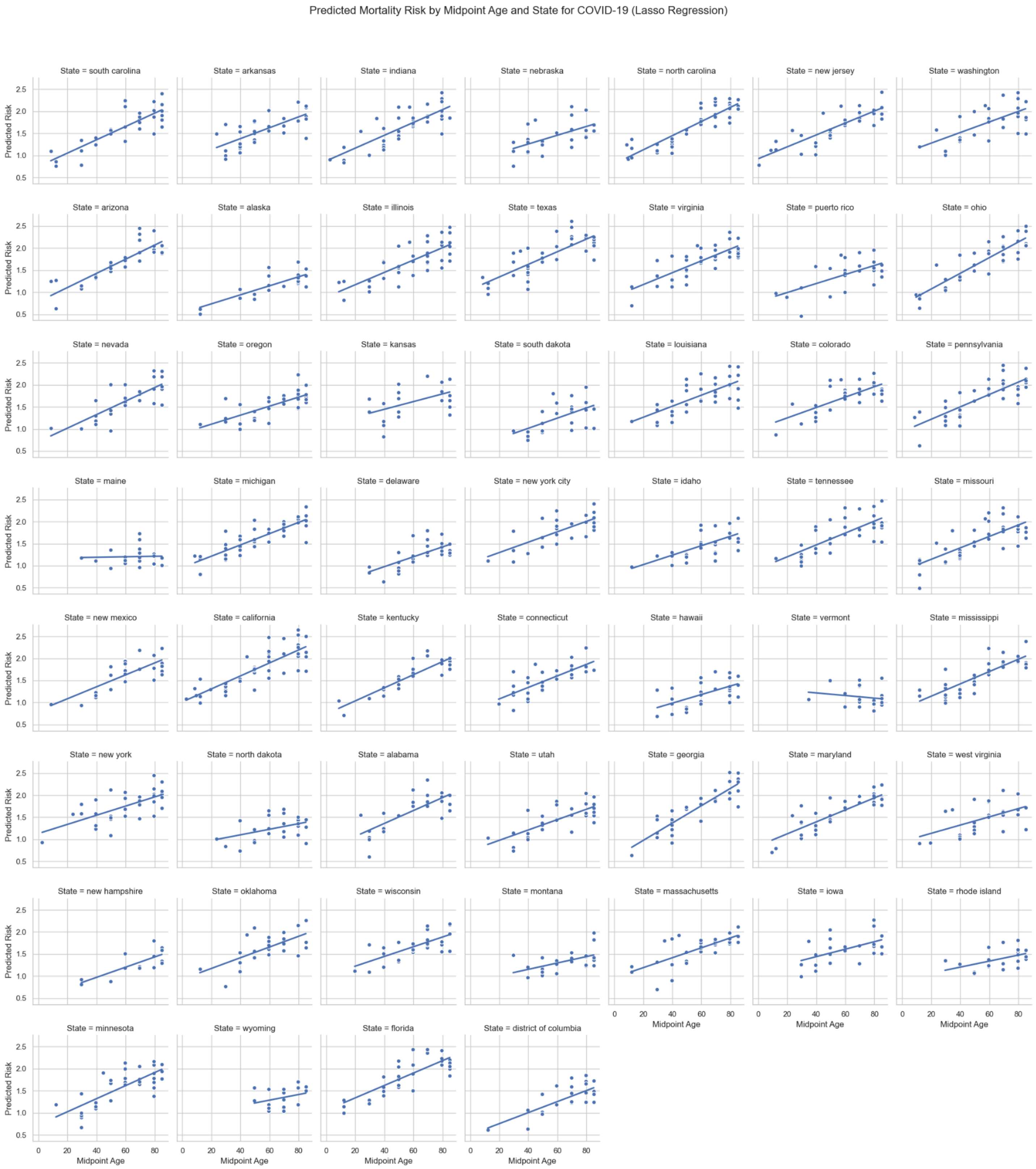


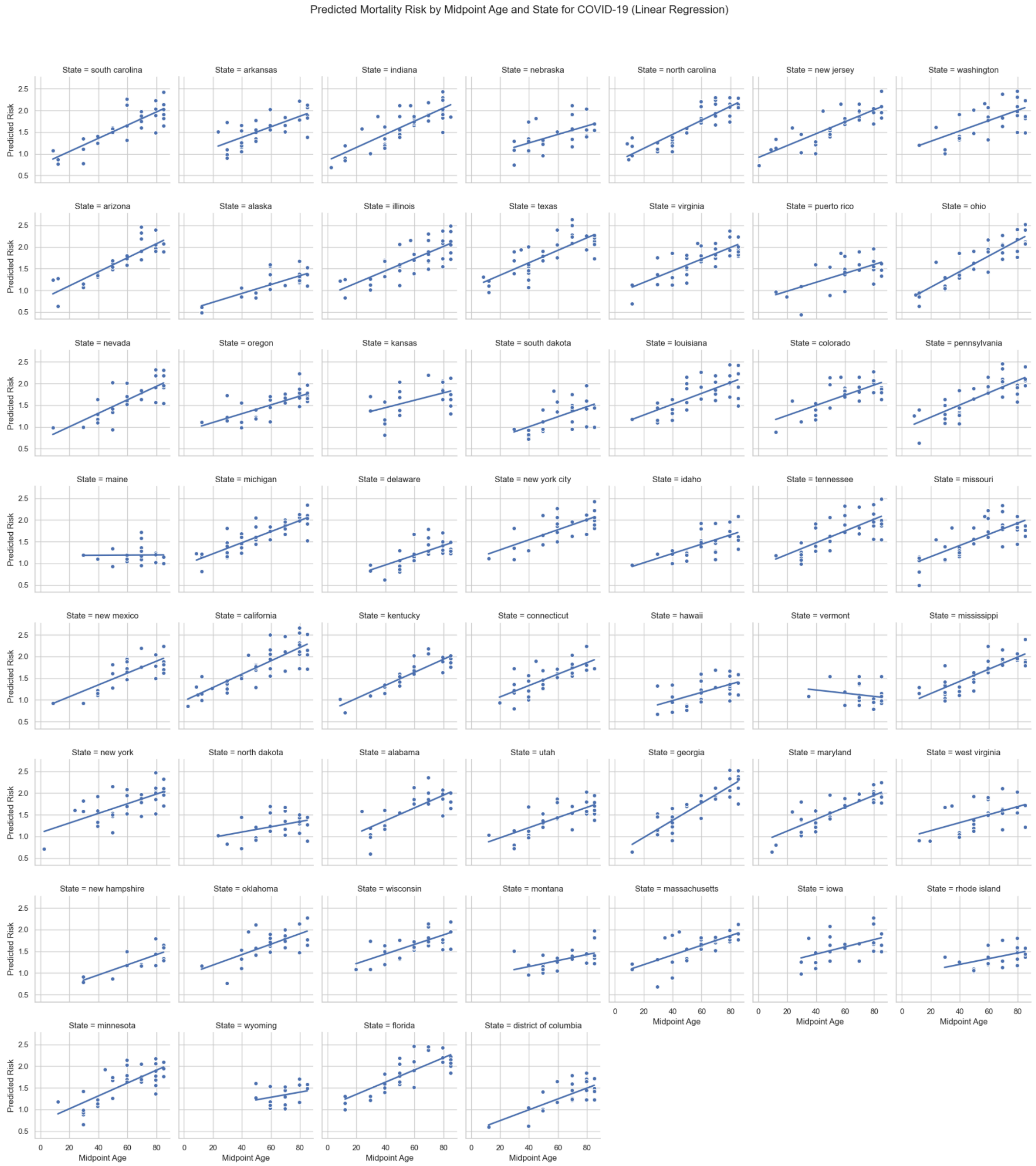
Figure 4- describes the number of COVID19 deaths in relation to Influenza as one of the conditions of the COVID19 death.

Figure 5 (fig. 5)



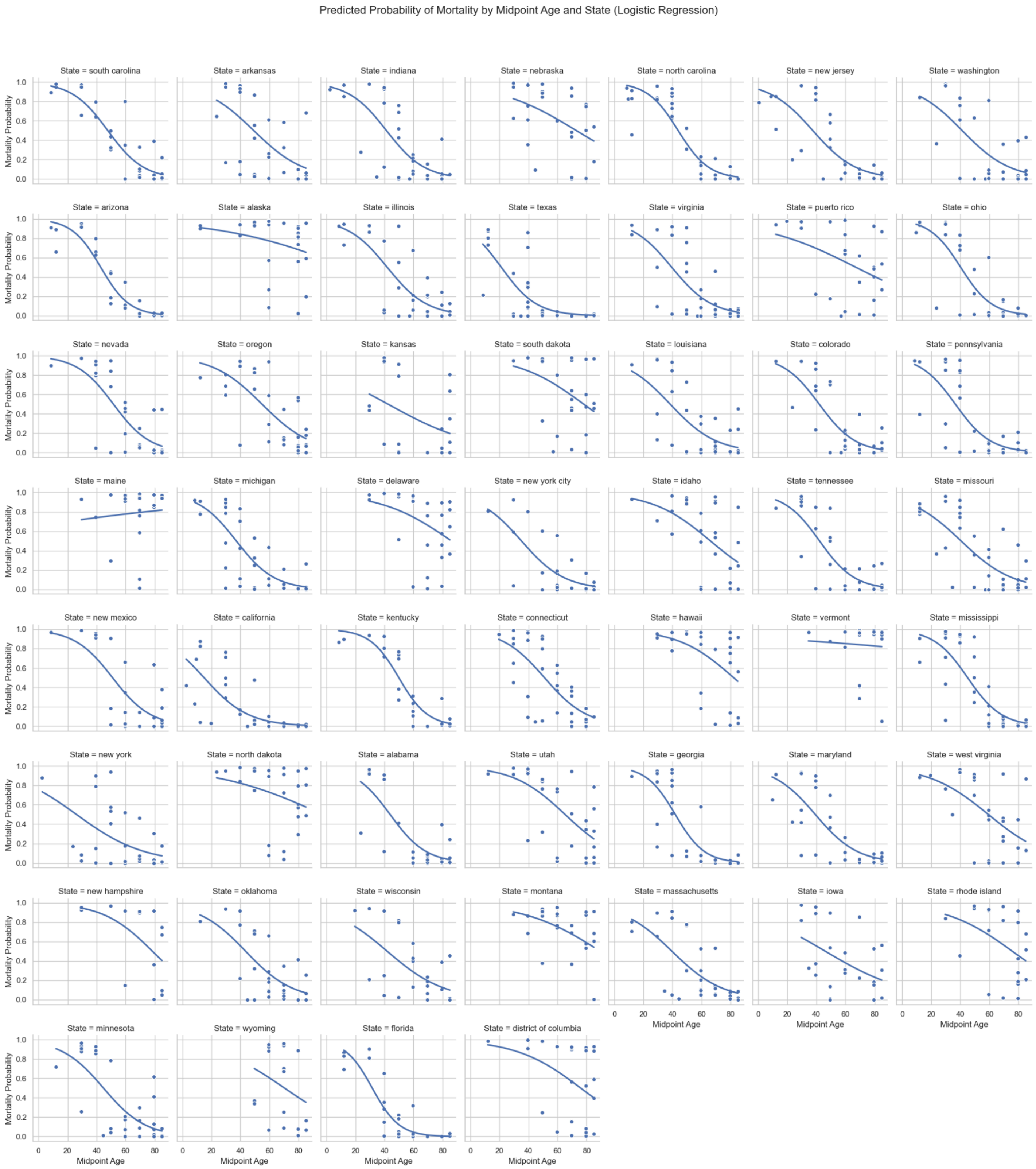
Description of Figure 5 can be found in the Summary of Results

Figure 6 (fig. 6)



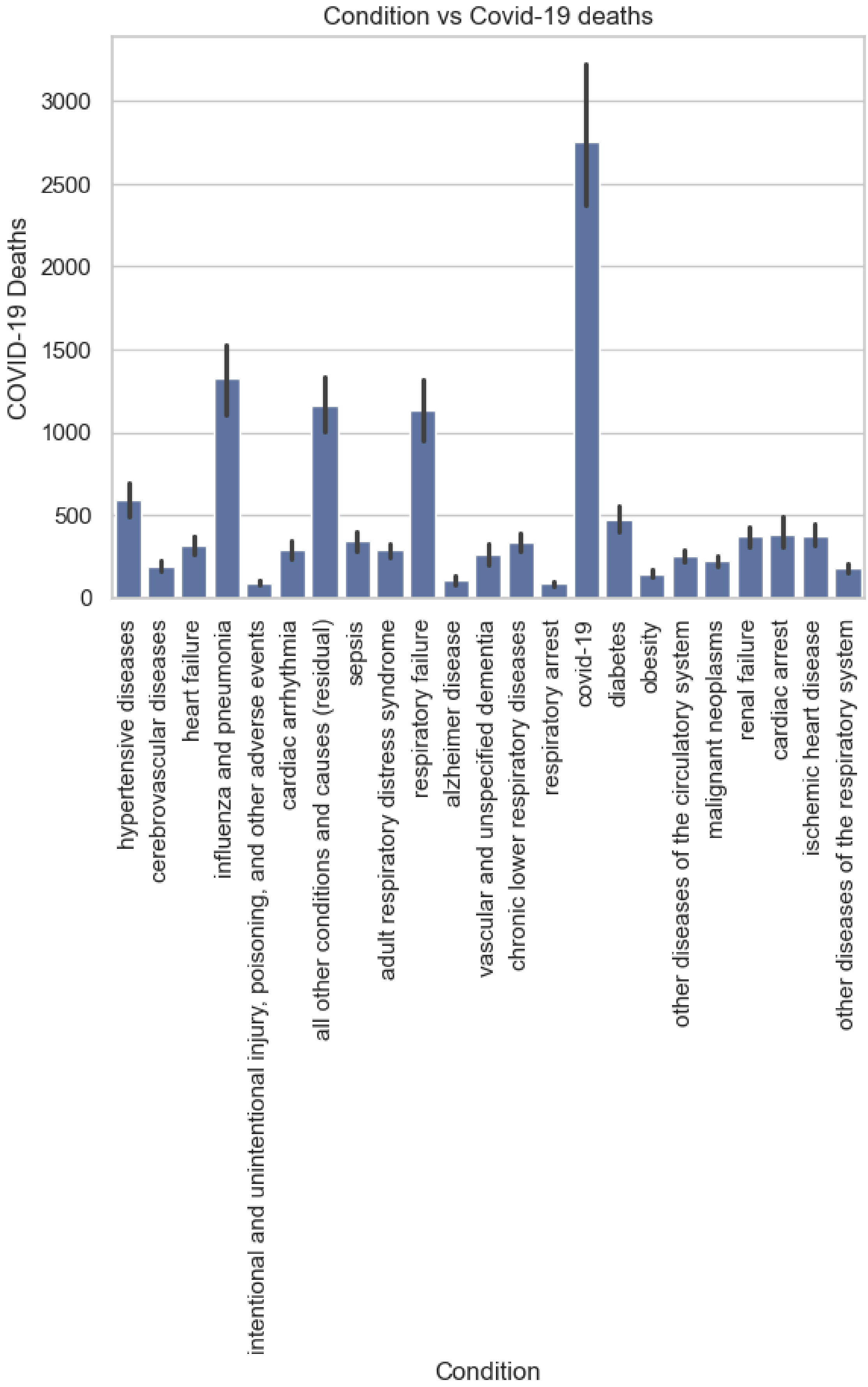
Description of Figure 6 can be found in the Summary of Results

Figure 7 (fig. 7)



Description of Figure 7 can be found in the Summary of Results

Figure 8 (fig. 8)


Condition vs Covid-19 deaths
COVID-19 Deaths
hypertensive diseases
cerebrovascular diseases
heart failure
influenza and pneumonia
intentional and unintentional injury, poisoning, and other adverse events
cardiac arrhythmia
all other conditions and causes (residual)
sepsis
adult respiratory distress syndrome
respiratory failure
alzheimer disease
vascular and unspecified dementia
chronic lower respiratory diseases
respiratory arrest
covid-19
diabetes
obesity
other diseases of the circulatory system
malignant neoplasms
renal failure
cardiac arrest
ischemic heart disease
other diseases of the respiratory system
Condition