Methods

Modeling Deaths for Diabetes & Kidney Diseases, Globally

Authors: Shreyas Dikshit, Amare Diotte, Taylor Goodwin, Aditi Gupta

Data Preparations & Preprocessing

The scope of our project was informed by what data was available to us. Since we were collecting data from across the entire world and over multiple decades, finding complete, consistent datasets was a challenge. For that reason, the final dataset used for our project is built from 4 different documents. Initially, we looked only at the death rate data in the Global Burden of Disease Study from 1990 to the present [1], however data was sparse prior to 2000 and there were not enough independent variables to support a model in this dataset alone. It became clear we would need to add to more sources to flesh out the project. And as a result, we needed to organize our data so it would be easier to add and work with heterogeneous data. The solution was a relational database, so we worked with a SQLite database from this point on.

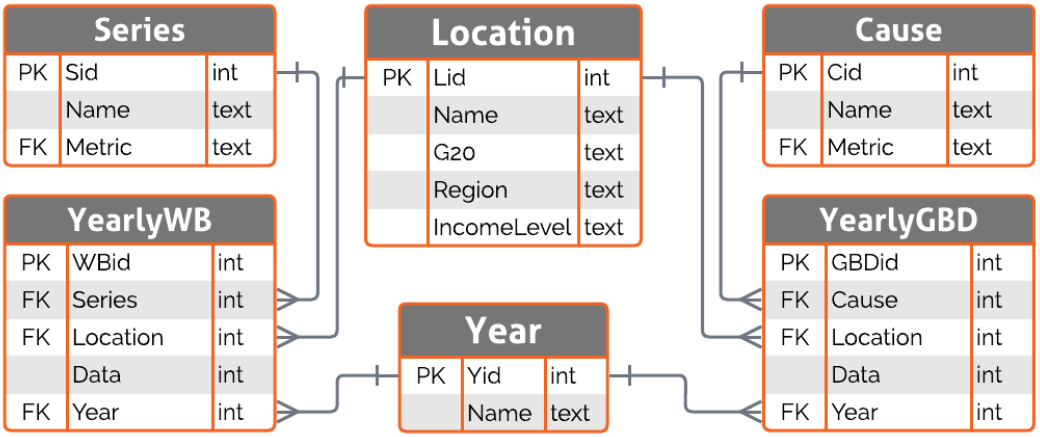
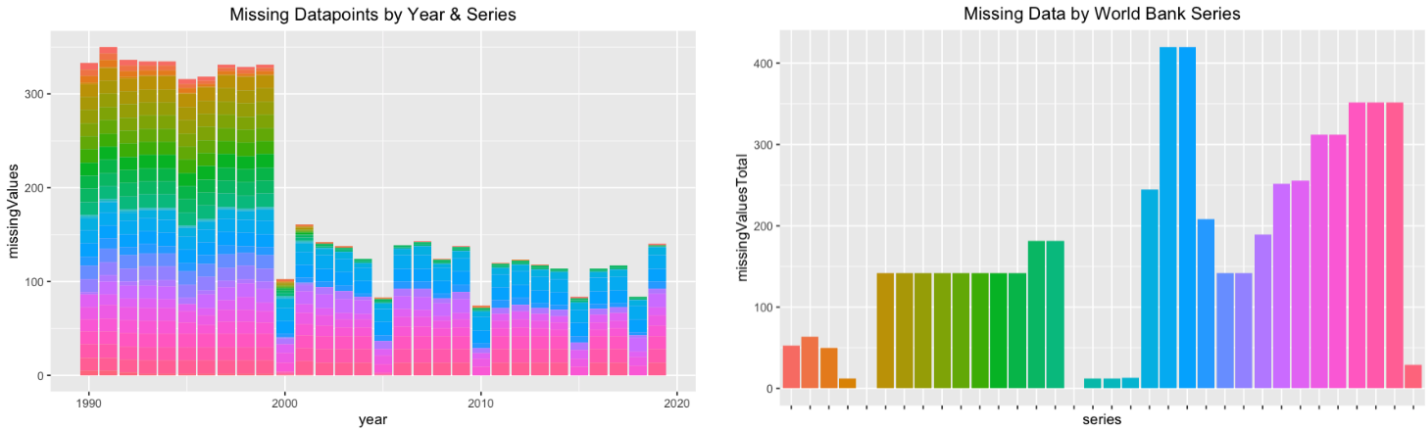


Figure 1: Final Schema Used

In the end we added 3 more datasets of potential predictors for the death rates in the original GBD study. Supplemental information from the GBD study gave demographic information about each country in the study. We also pulled incidence and prevalence data series for diabetes and kidney disease from the GBD to look at as possible predictors. And the most substantial addition was a dataset on world development indicators that we pulled from The World Bank [2]. The World Bank data was formatted much differently from the other datasets, so columns had to be pivoted longer and renamed. We had to normalize all country names between the GBD and WB datasets, as well as removing any countries that did not appear in both sets. Once the data was normalized, it was loaded into a SQLite database with two main tables - YearlyWB and

YearlyGBD. Every row in YearlyWB represents an observation of a development indicator for a specific year and country, every row in YearlyGBD represents an observation of a diabetes health statistic for a specific year and country. Those two tables are linked by common year and country lookup tables. This symmetric setup made it easier to pull data across the datasets. It also made it easier to add data when necessary - for example, when supplemental demographic info was added for countries, only a few new fields needed to be added to the Location table to update the entire dataset.

As we were collecting this data, we were also actively narrowing it down to only viable independent variables. As mentioned earlier, data was largely incomplete in earlier studies, so we opted to start in 2000 instead of 1990. A lot of the series in the World Bank studies were only done every 5 years, so we decided to limit our dataset to the years 2000, 2005, 2010, and 2015 to maximize the number of available complete records. In the end, we also included 2019 as it was the most recent year available and the series we ended up using in our models were largely complete that year. Even with these restrictions, we still had a very large number of variables to look at, so we wrote a few simple algorithms to count where there were missing values and looked at visualizations like the ones below.



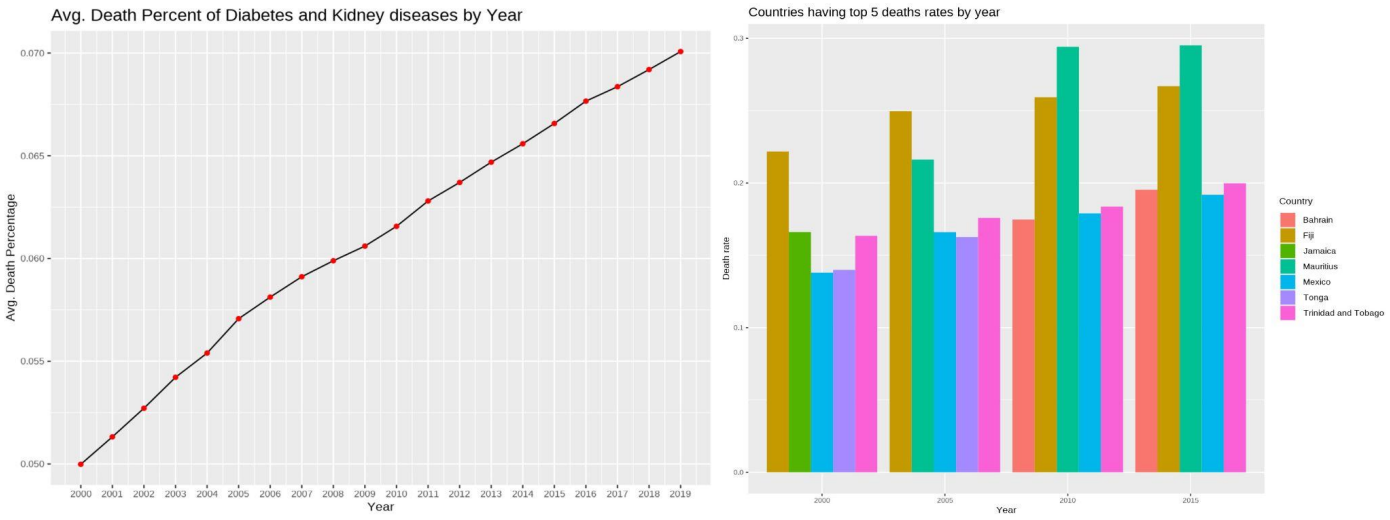
(a) *(b)*

Figure 2: Missing values in (a)109 different series in WB Development Indicators over all years (1990 - 2020), (b)final 39 series, cumulative over (2000, 2005, 2010, 2015, 2019)

Using this information, we made the decision to only consider series that were missing less than 10% of the total observations, reducing the number of series from 102 to 39. Next, the process was repeated while counting the total missing values by country. After the initial normalization of the country names, there were 200 countries total in our dataset, but we reduced the dataset to 178 by eliminating any countries that were missing more than 10% of total observations. All remaining N/A values were then imputed by taking the mean of that countries' data for the remaining years in the series. Lastly, we eliminated 6 countries that were unable to be imputed due to observations missing across all the years. In the final iteration of our database, we have diabetes rates and economic data for 172 countries across the years of 2000, 2005, 2010, 2015, & 2019.

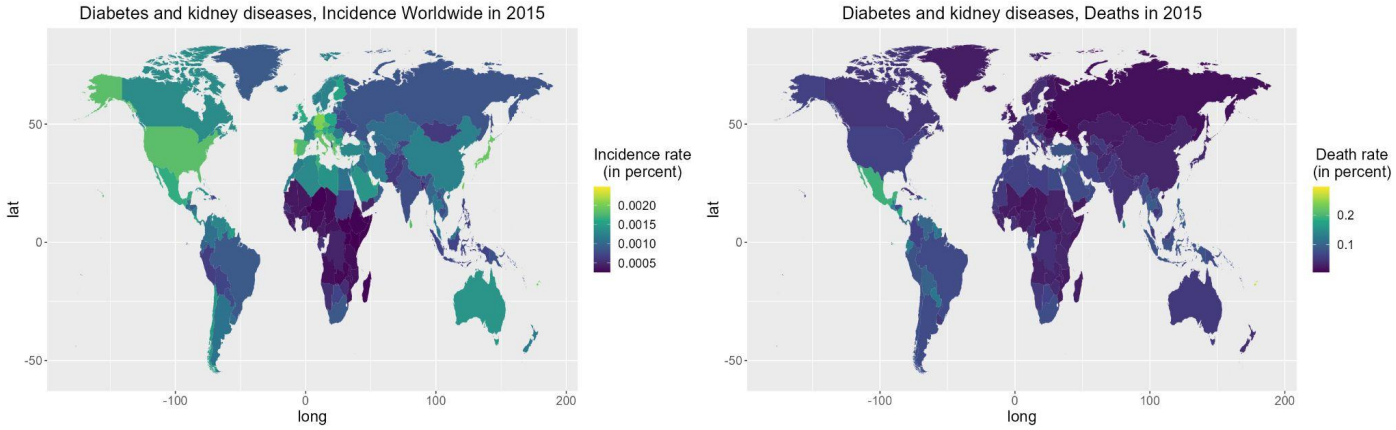
Exploratory Data Analysis

A variety of visualizations like line plots, barplots, maps etc. were prepared to understand patterns, trends and insights in the data before the model could be developed. Some of them have been discussed below.



(a) *(b)*

Figure 3: (a)Line graph demonstrating the linear relationship of avg. death percentage by year, (b) bar plot for top 5 countries with highest death rates.



(a) *(b)*

Figure 4: World Maps depicting (a)Incidence statistics, (b)Death statistics, worldwide in percentage

- Death rates increase almost linearly from 2000 to 2019
- Some countries like Mexico, Trinidad and Tobago, and Mauritius repeatedly appear in the countries with the top 5 death rates for 2000, 2005, 2010, 2015
- In 2015, countries like the USA, Italy and Germany had a relatively higher incidence rate but seem to have a comparably lower death rate. On the other hand, countries like Mexico had a relatively lower incidence rate but had a much higher death rate

Data Normalization & Modeling

In seeking to discern whether the data was fit to be modeled in a multivariable linear regression, the normality of the testing data was considered via plotting scatter plots. Upon plotting each of the independent variables against the rate of death for diabetes and kidney disease, it was identified that some variables possessed right skewed distributions and would benefit from a logarithmic transformation.

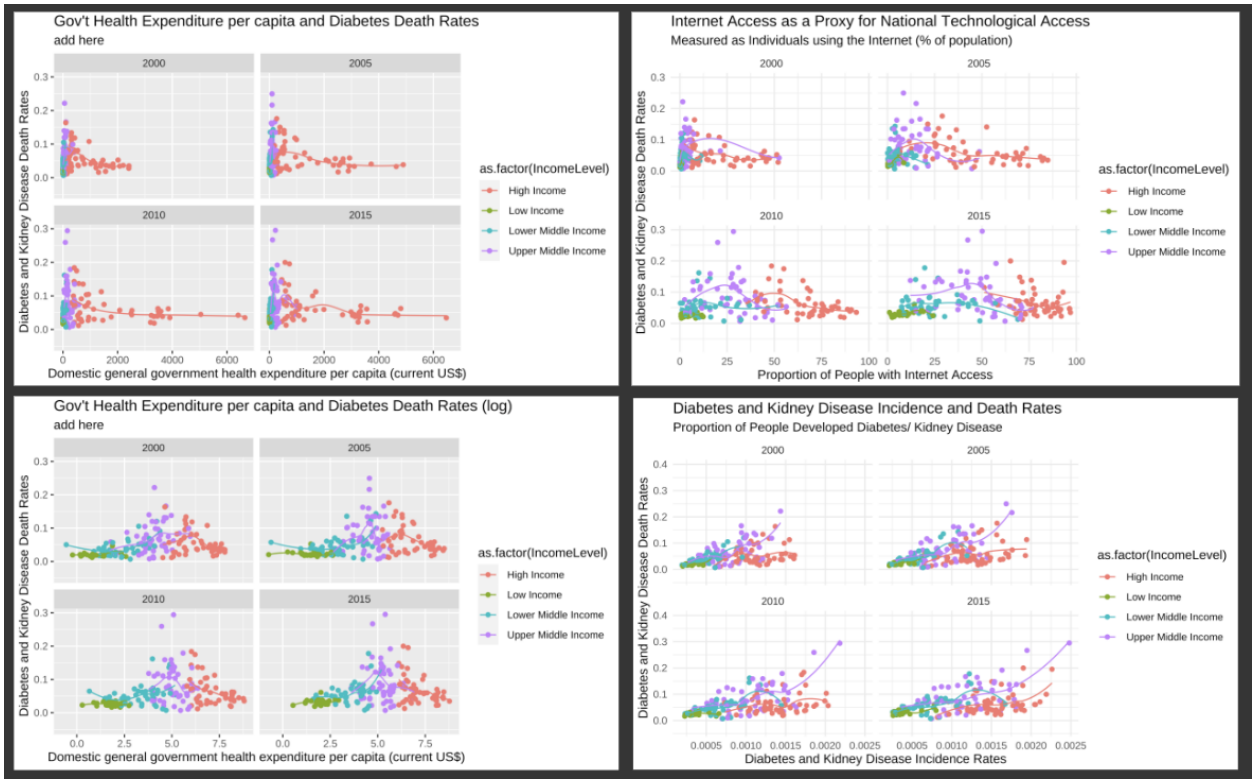
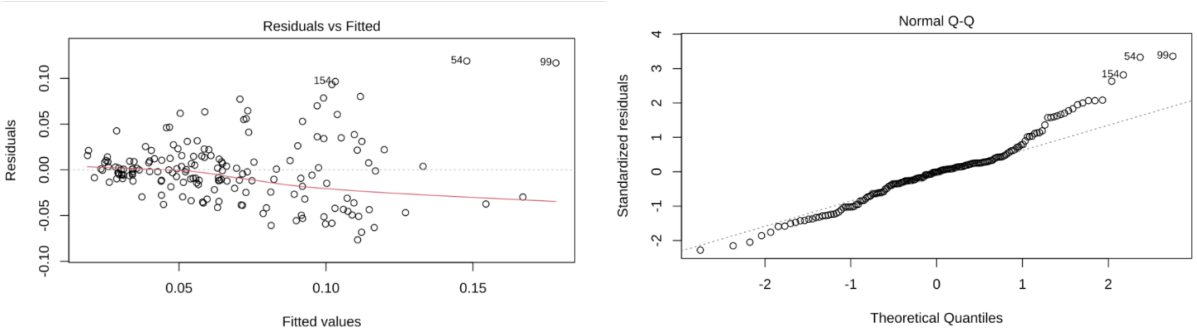


Figure 5: Plotting each independent variable against rate of death for diabetes and kidney disease and their logarithmic transformation plots.

Once normalization of the variables had been addressed, a single linear regression modeling approach was applied to each independent variable in the dataset, and an understanding of the dataset, specifically the relationships present between diabetes and kidney disease death rates, was developed. Based on the factors indicated in prior research, we established key groupings of variables which had overarching national demographic identifiers. These groupings included: technological advancement, medical capacity, economic wealth, governmental and private health expenditure, and environmental health. The resulting RMSE values indicated which variables, across each group, would be statistically useful additions to our developing multivariable regression model, given the relationships present in their individual models.

Next, adding the variables of highest indication, one by one, the best predictors for each group were established and a multivariable linear model was built, with new variables added by

category to the strongest model amongst the single linear models. To avoid overfitting the model on the testing data from 2015, only a few key categories were chosen as a focus for the model, given their likelihood to have an impact on the dependent variable. The selected categories were: medical capacity, economic wellness, and technological competence. This led to incorporating the following variables: diabetes incidence rate, income level, internet access rate. Government expenditure on healthcare was also considered as an independent variable, but was ultimately eliminated from the model, due to the modeling capabilities of the original model achieving lower RMSE and AIC values. In addition, adding both governmental health expenditure and income level has the potential for a fair amount of overlapping information, and as a result the two's collinearity would make including both of them unwise.



(a) *(b)*

Figure 6: Model Evaluation using (a) Residual vs Fitted plot and (b) Normal Q-Q plot

Finally, the model was evaluated with the use of a residuals vs fitted plot and a normal q-q plot, to assess the validity of the model and its fit to the data. The residual vs fitted plot revealed a relatively even band of values above and below the zero line and fairly random bouncing of values around the horizontal line in the residuals vs fitted plot. Indicating variances of error terms are equal and that a linear approach is reasonable. Meanwhile, in the normal q-q plot, the majority of the points land approximately along the straight line, though there are a few outliers identified. The curvature of the plotted points near the upper and lower ends specifically indicated by light-tailed data, and that there exist some extreme values in the data. Overall, each plot indicated a few outliers were present in the data, but that the approach of linear regression was a statistically reasonable one.

References

[1] Centers for Disease Control and Prevention. (n.d.). Diabetes and Chronic Kidney Disease. U.S. Department of Health & Human Services. Retrieved April 23, 2023, from https://www.cdc.gov/diabetes/managing/diabetes-kidney-disease.html/

[2] World Bank. (2023). World Development Indicators. Retrieved March 15, 2023, from https://data.worldbank.org/