Modeling Deaths for Diabetes & Kidney Diseases, Globally

Team 11
Amare Diotte, Shreyas Dikshit, Aditi Gupta, & Tay Goodwin

Problem Statement

The project aims to use machine learning algorithms to predict the death rates for **Diabetes and Kidney diseases** for countries, based on a set of economic and health-related factors.

Exploratory data analysis will be performed to gain insights into the relationships between the various factors and the death rates. This will involve visualizing the data using techniques such as scatter plots, box plots, and world maps, to identify any trends or insights in the data.

Statistical Approach

Problem Statement Analysis	Data Gathering and Preprocessing	Exploratory Data Analysis	Model Development and Results	Discussions
Understanding the problem statement, various economic and health-related factors and how they could potentially impact the death rate	Data Sources: Global Health Data Exchange, World Bank Data Preprocessing: Handling inconsistent values, Tidying, Imputing NAs	Visualize trends in the rates of Diabetes and kidney diseases across the world and specific countries	Building a multivariable linear regression model, using a stepwise approach	Analyze model results and implications. Discuss next steps.

Data Gathering & Preprocessing For World Health Data

	Year	Location	Cardiovascular diseases, Deaths	Cardiovascular diseases, Incidence	Cardiovascular diseases, Prevalence	Chronic respiratory diseases, Deaths	Diabetes and kidney diseases, Deaths	Diabetes and kidney diseases, Incidence	Diabetes and kidney diseases, Prevalence	Maternal and neonatal disorders, Deaths	Self-harm and interpersonal violence, Deaths	Substance use disorders, Deaths
1	2000	Albania	0.45712867	0.0011286854	0.06547224	0.04505006	0.023450479	0.0005183169	0.07980928	0.042945262	0.027670801	0.0021450055
2	2000	Algeria	0.41234254	0.0010712949	0.04878915	0.03262115	0.039486309	0.0007513494	0.08043431	0.092137485	0.030518737	0.0026442103
3	2000	Angola	0.07616737	0.0004928251	0.03337224	0.01645836	0.016793103	0.0002604144	0.04523783	0.100508651	0.025600701	0.0007548008
4	2000	Antigua and Barbuda	0.35184019	0.0010865786	0.06097585	0.01560338	0.124193421	0.0009991017	0.12202869	0.024505194	0.011604922	0.0100693737
5	2000	Argentina	0.32803643	0.0011666638	0.05865235	0.05279649	0.075176023	0.0008900393	0.09802477	0.023422334	0.024911832	0.0037141109
6	2000	Armenia	0.51011692	0.0019508395	0.07929834	0.04369889	0.058669956	0.0007114241	0.11970678	0.021421403	0.011135915	0.0043202310
7	2000	Australia	0.38159136	0.0019893142	0.09109781	0.05945634	0.042614460	0.0011457933	0.11013930	0.005239236	0.022972374	0.0102516882
8	2000	Austria	0.47387745	0.0023142148	0.11552081	0.03403748	0.034570294	0.0012093330	0.12089717	0.002375422	0.021522390	0.0061713482
9	2000	Azerbaijan	0.48456199	0.0018311874	0.05747802	0.02985288	0.032141610	0.0006621602	0.09700295	0.045511252	0.012238850	0.0042722771
10	2000	Bahamas	0.28589937	0.0009844500	0.05118908	0.01611224	0.081974015	0.0008879483	0.10695393	0.019572192	0.037944839	0.0070463039
11	2000	Bahrain	0.32600635	0.0007879170	0.03489473	0.04389972	0.100680353	0.0011515408	0.09290268	0.028278552	0.029899755	0.0020418730
12	2000	Bangladesh	0.18841700	0.0005804262	0.03184294	0.06550268	0.030162143	0.0003067628	0.05811749	0.136802340	0.015367262	0.0005326905
13	2000	Barbados	0.31432889	0.0015759031	0.08605706	0.01883131	0.133711208	0.0012129573	0.14911542	0.014826673	0.017048227	0.0026702764
14	2000	Belarus	0.57597486	0.0028913424	0.10062841	0.04149492	0.007795989	0.0006072836	0.13478239	0.003431094	0.041152147	0.0219229559
15	2000	Belgium	0.34800239	0.0021194628	0.10601386	0.06845232	0.034481087	0.0012958145	0.12881281	0.002496204	0.024936681	0.0049318234
16	2000	Belize	0.25868111	0.0008066027	0.04231437	0.03178546	0.098449317	0.0005956724	0.07601379	0.059445322	0.047330524	0.0040923936
17	2000	Benin	0.09500676	0.0005220844	0.03247384	0.01685836	0.022847314	0.0003168910	0.04405879	0.139975200	0.011220466	0.0006147762

Showing 1 to 17 of 860 entries, 12 total columns

The final data base now contains -

Death, incidence and prevalence data for 169 countries with less than 10% of values missing

Sourced from Global Health Data Exchange website.

Focused on one cause of death, Diabetes and Kidney diseases, for all countries.

Altered to provide combined rates and percentages for all genders.

Removed inconsistencies in country names in both data sets.

Data Gathering & Preprocessing For World Economics Data

	Year	Location	Access_to_clean_fuels_and_technologies_for_cooking	Access_to_electricity	Agricultural_land_percent_of_land_area	Agricultural_land_sq_km	CO2_emissions_kg	CO2_emissions_kt	CO2_emis
1	2000	Albania	38.2	100	41.7518248175182	11440	0.523342626827492	3170	1.0262
2	2000	Algeria	97.1	98.9730987548828	16.8032614811021	400210	0.828864822659494	80050	2.6011
3	2000	Angola	41.1	24.2127437591553	37.6685569904548	469613.9	0.539238778149565	16200	0.9881
4	2000	Antigua and Barbuda	100	97.6892623901367	20.4545454545455	90	0.308589558230521	330	4.3967
5	2000	Argentina	95	95.7832870483398	46.9581867146078	1285100	0.334313951755177	132270	3.5680
6	2000	Armenia	79.5	98.9000015258789	46.4699683877766	13230	0.89912049155261	3560	1.1235
7	2000	Australia	100	100	59.2881038230738	4554690	0.388696147267585	339450	17.838
8	2000	Austria	100	100	35.630150266602	29402	0.204150242335783	63530	7.9297
9	2000	Azerbaijan	70.3	98.9082260131836	57.3863567580655	47404	2.3233728835353	27690	3.4403
10	2000	Bahamas	100	100	1.2987012987013	130	0.209032522010119	2230	6.8612
11	2000	Bahrain	100	100	12.9577464788732	92	1.02796991312149	15880	22.320
12	2000	Bangladesh	8.3	32	72.2132595836214	94000	0.259393852070657	21650	0.1675
13	2000	Barbados	100	100	41.8604651162791	180	0.280590042738931	1240	4.6853
14	2000	Belarus	94.7	100	45.6105063890204	92520	2.02712758728271	52940	5.3048
15	2000	Belgium	100	100	45.891677675033	13896	0.320662858482005	117270	11.439
16	2000	Belize	79.9	79	6.53222270933801	1490	0.328752804584556	450	1.8718
17	2000	Benin	0.600000000000001	21.5310840606689	28.3345157857396	31950	0.228319522848469	1420	0.2029
18	2000	Bhutan	27.8	31.1499996185303	14.5226130653266	5780	0.320560055786337	210	0.3576
19	2000	Bolivia	62.9	69.9630432128906	34.2869011354195	371430	0.473614375827309	8210	0.9554
20	2000	Bosnia and Herzegovina	52.1	100	41.6015625	21300	1.30756063235718	13960	3.3402
21	2000	Botswana	44.4	26.4089794158936	45.7907645616078	259510	0.474576985006232	4030	2.3335
22	2000	Brazil	89	94.3917160034180	27.3175912632436	2282235	0.264283490256424	212670	1.7824

Showing 1 to 22 of 860 entries, 41 total columns

The final data base now contains -

39 demographic factors for 169 countries with less than 10% of values missing

Sourced from The World Bank Data website.

Huge inconsistencies in data, years at intervals of 5 had the most consistent data for every factor.

Missing values were imputed with the mean across all years for each country, in both data sets.

Why create a database?

WORD BANK ECONOMIC DATA:

Country.Name	Series.Name	YR2000	YR2005	...
Afghanistan	Access to clean fuels and techn...	6.200	12.2	...
Afghanistan	Access to electricity (% of...	75.1	25.39	...
Afghanistan	Adjusted net enrollment rate...	..	..	...
Afghanistan	Agricultural land (% of land area)	379110	379110	...

Easier to search heterogeneous data

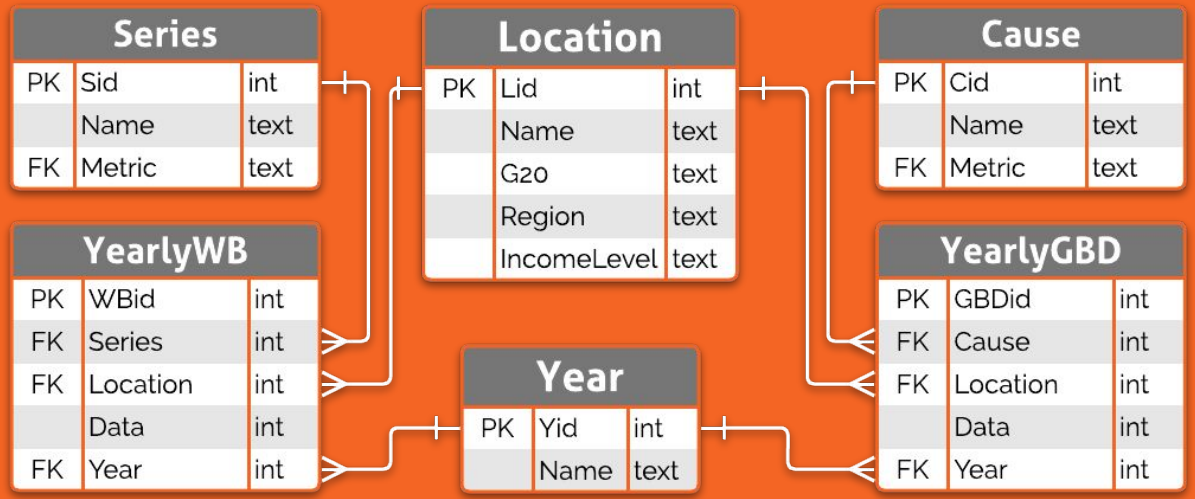
Consistent data access across project

Easier to update/add new data fields

GDB CAUSE OF DEATH DATA:

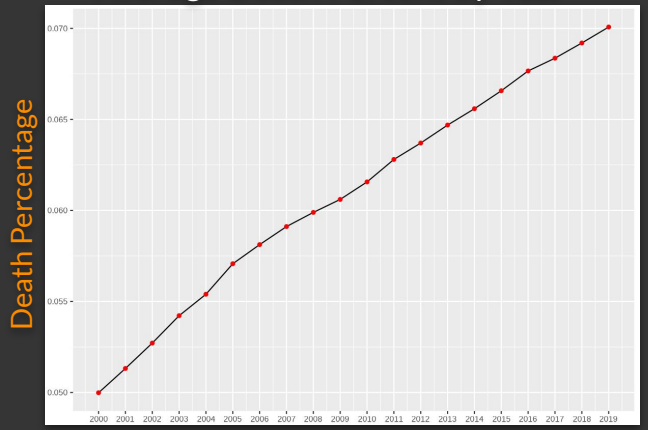
location	measure	cause	metric	year	val
Samoa	Deaths	Cardiovascular diseases	Percent	2000	0....
Albania	Prevalence	Diabetes and kidney diseases	Percent	2005	0....
Rwanda	Incidence	Cardiovascular diseases	Percent	2010	0....
Latvia	Deaths	Diabetes and kidney diseases	Percent	2015	0....

Tables in Common:
Location & Year



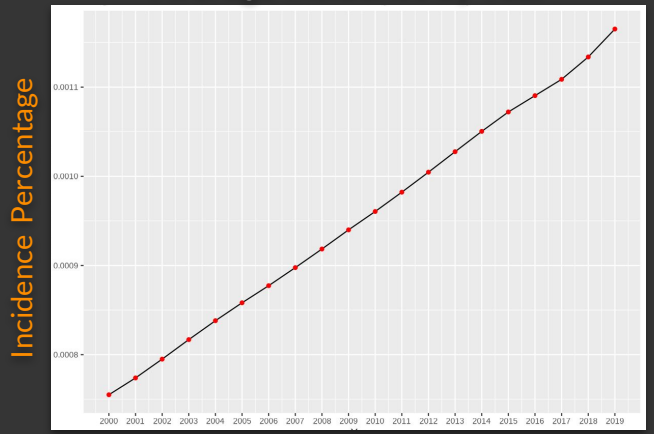
Global Yearly Trends in Diabetes and Kidney Diseases

Average Death Percent by Year



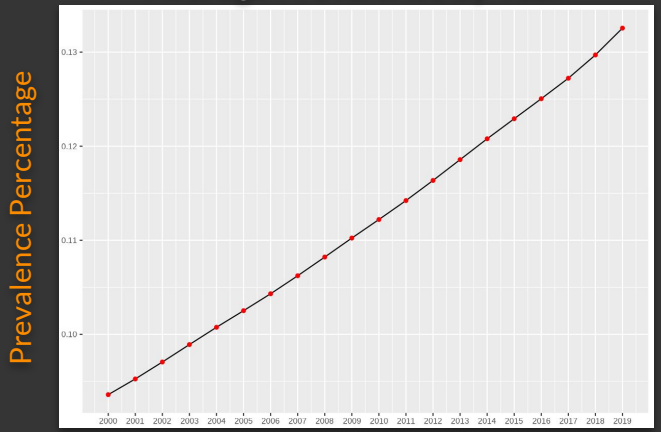
Year 2000 - 2019

Average Incidence by Year



Year 2000 - 2019

Average Prevalence by Year



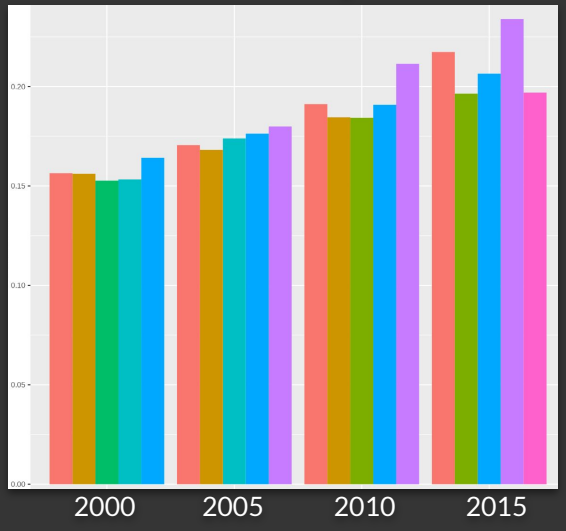
Year 2000 - 2019

Observations:

Global average percentages for Prevalence, Incidence and Death due to Diabetes and Kidney diseases increases almost linearly from 2000 to 2019

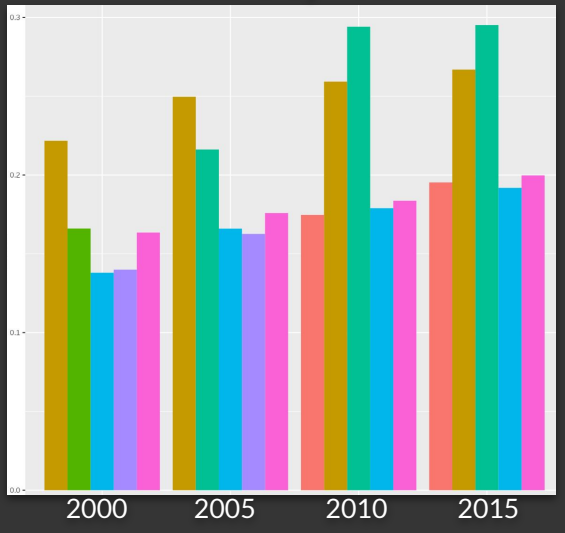
Countries with the Highest Diabetes Rates by Year

Prevalence by Year



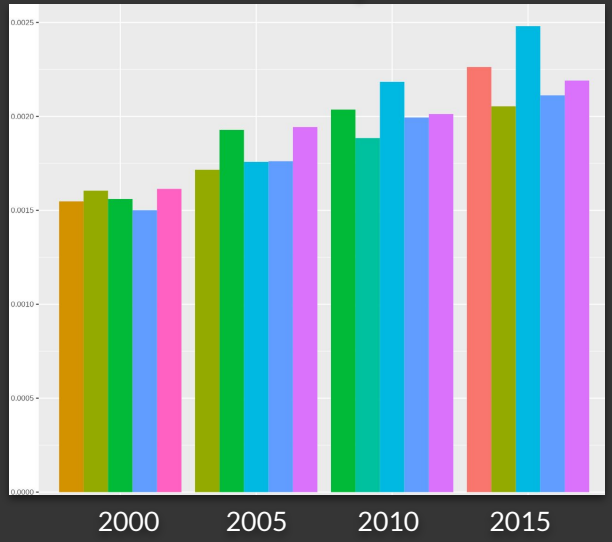
- Czechia
- Estonia
- Fiji
- Germany
- Italy
- Japan
- Mauritius
- Trinidad and Tobago

Deaths by Year



- Bahrain
- Fiji
- Jamaica
- Mauritius
- Mexico
- Tonga
- Trinidad and Tobago

Incidence by Year



- Bahrain
- Cyprus
- Germany
- Italy
- Malta
- Mauritius
- Portugal
- Qatar
- Spain

Observations:

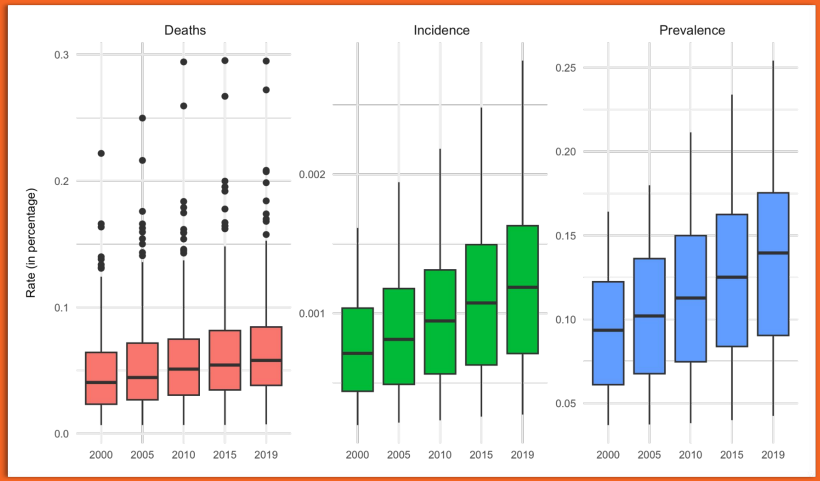
- Europe has the most no. of countries in these plots (7), followed by Asia (4) and North America (3)

Comparing Deaths, Incidence, & Prevalence

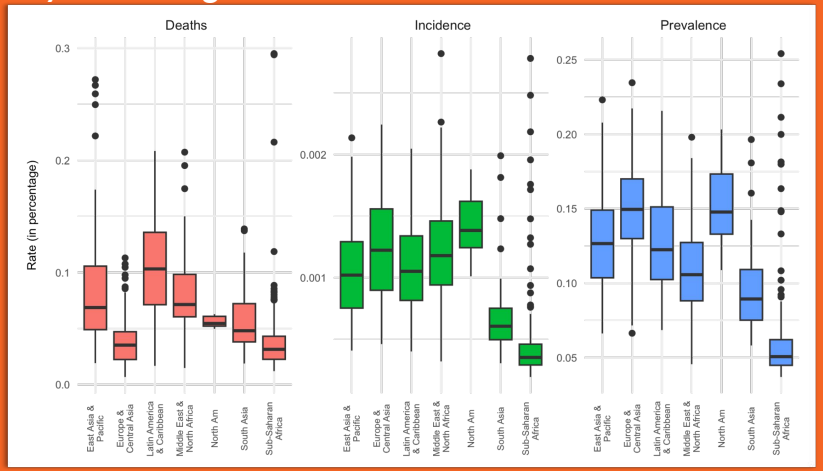
Prevalence: Proportion of people in a population who are a case of diabetes or kidney disease

Incidence: Number of new cases of diabetes or kidney disease during a given period

By Year:



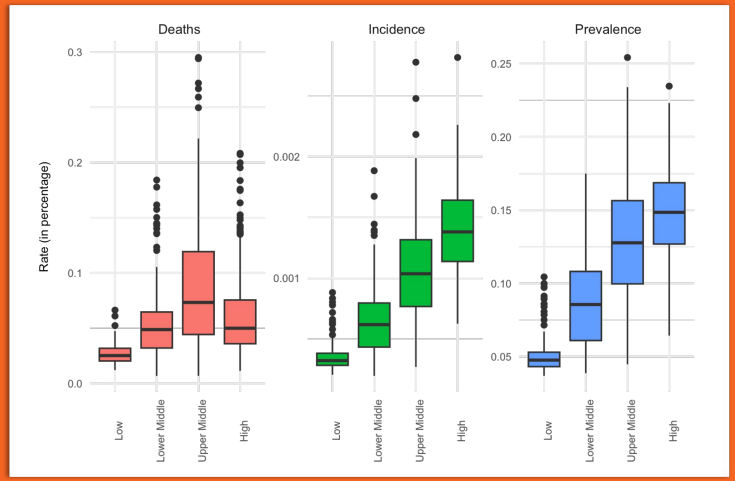
By World Region:



World Bank Regions:

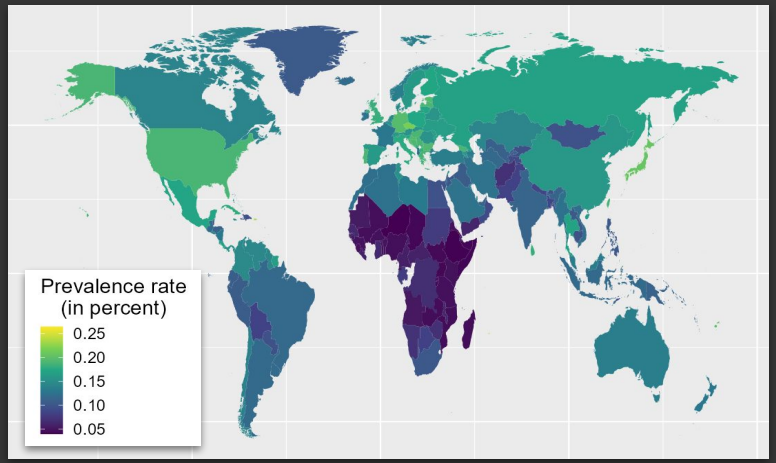
- East Asia & Pacific
- Europe & Central Asia
- Latin America & Caribbean
- Middle East & North Africa
- North America
- South Asia
- Sub-Saharan

By Income Level:

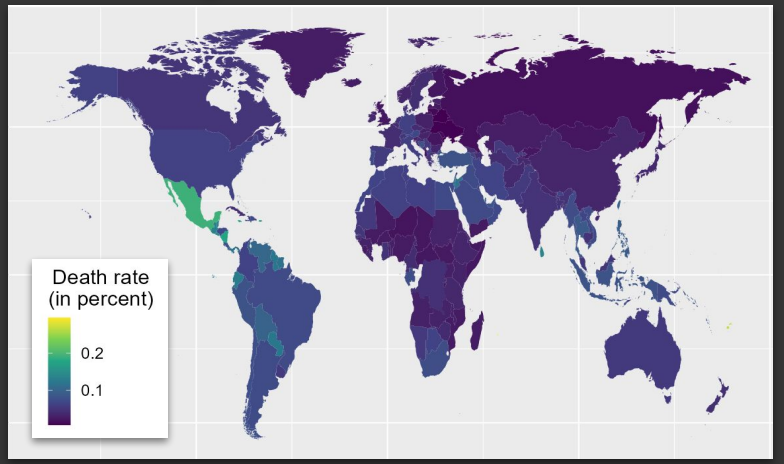


World Rates of Diabetes and Kidney Diseases in 2015

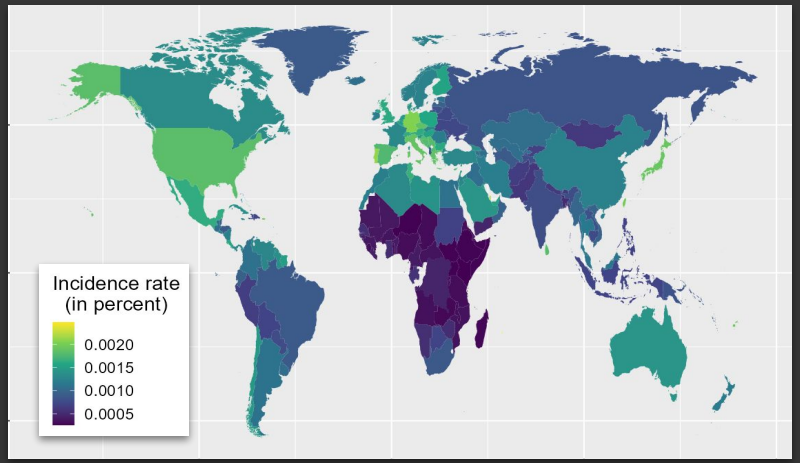
Prevalence Worldwide in 2015



Percentage of Deaths Worldwide in 2015



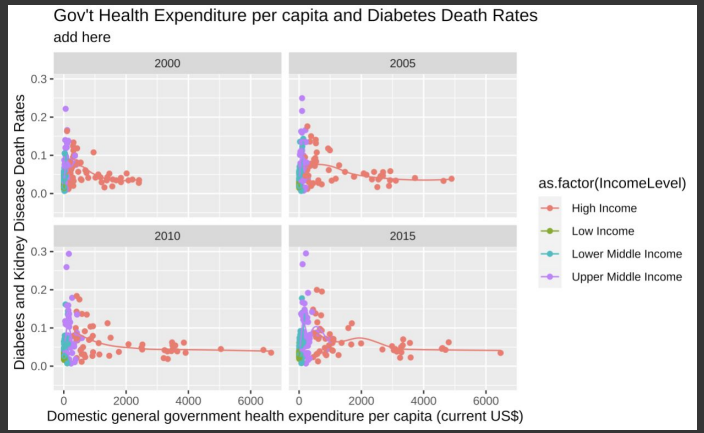
Incidence Worldwide in 2015

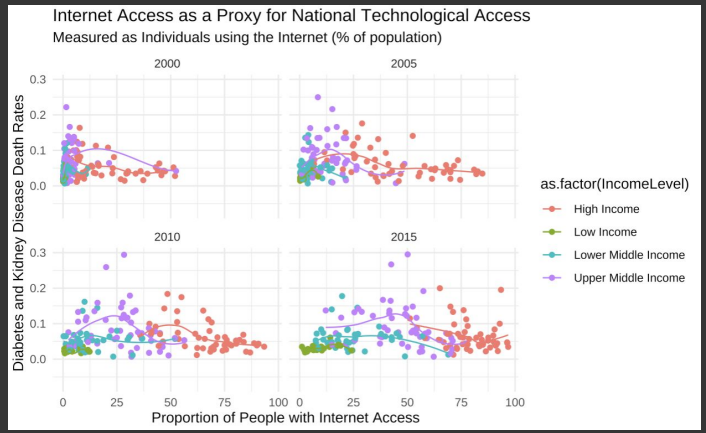


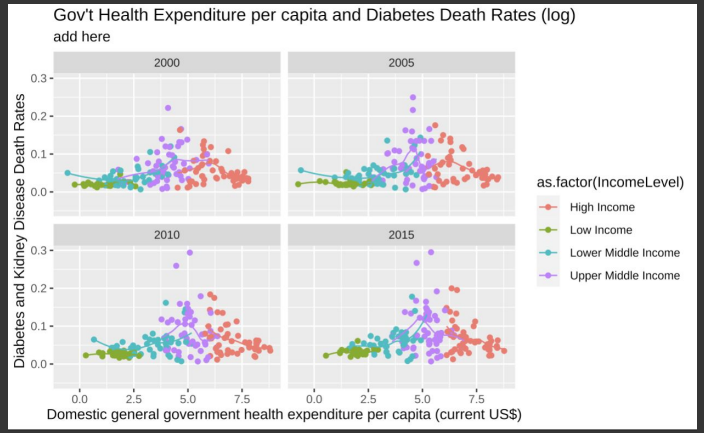
Data Evaluation & Transformation

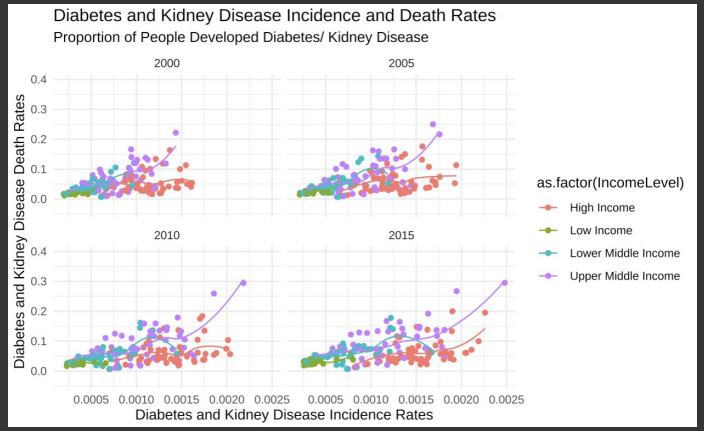
A log transformation was applied to **government health expenditure per capita**,

No transformation was utilized on the **proportion of the population with internet access** variables, nor **incidence rates of diabetes and kidney diseases**.









Training the Model

$$y = b_0 + 70.4332\,x_1 + -0.0176\,x_2 + 0.0013\,x_3 + 0.0202\,x_4 + -0.0009\,x_5$$

RMSE values

Model 1: 0.03437666
Model 2: 0.03551193

AIC values:

Model 1: -644.7081
Model 2: -643.5863

Examining the RMSE and AIC values, it appears that Model 1, as fitted to the 2015 data, provides a better fit.

Model 1 presents with lower RMSE and AIC values compared to Model 2.

```
Call:
lm(formula = df_2015$Diabetes_Kidn_Deaths ~ df_2015$Diabetes_Kidn_Incid +
    df_2015$IncomeLevel + df_2015$`Individuals using the Internet (% of population)`)

Residuals:
      Min        1Q    Median        3Q       Max
-0.076565 -0.018048 -0.001891  0.013649  0.118999

Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                       0.0304092  0.0202922   1.499   0.1359
df_2015$Diabetes_Kidn_Incid                      70.4331902  8.2476723   8.540 9.04e-15 ***
df_2015$IncomeLevel Low Income                   -0.0176261  0.0185763  -0.949   0.3441
df_2015$IncomeLevel Lower Middle Income           0.0012642  0.0145239   0.087   0.9307
df_2015$IncomeLevel Upper Middle Income           0.0202065  0.0098780   2.046   0.0424 *
df_2015$`Individuals using the Internet (% of population)` -0.0009390  0.0002134  -4.400 1.95e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

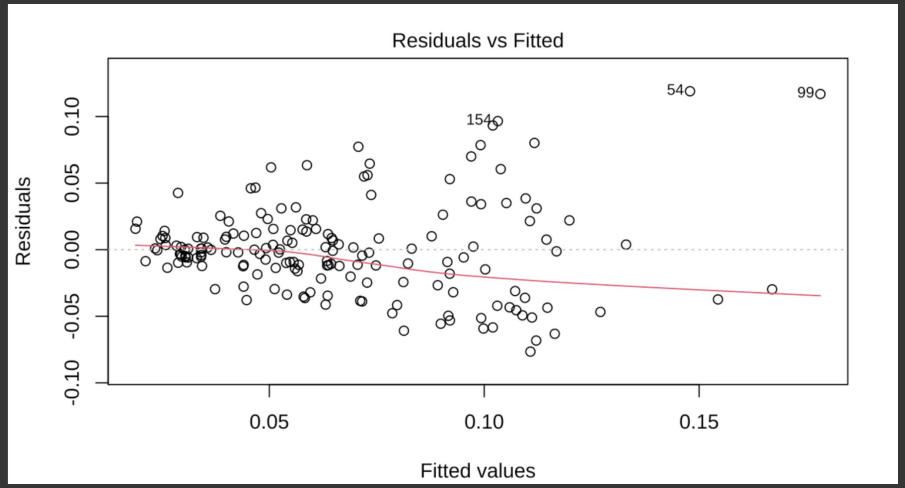
```
Call:
lm(formula = df_2015$Diabetes_Kidn_Deaths ~ df_2015$Diabetes_Kidn_Incid +
    df_2015$IncomeLevel + df_2015$`Individuals using the Internet (% of population)` +
    df_2015$log_Govt_Exp)

Residuals:
      Min        1Q    Median        3Q       Max
-0.076572 -0.019881 -0.001856  0.013402  0.116306

Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                       0.0484051  0.0282112   1.716 0.088108 .
df_2015$Diabetes_Kidn_Incid                      71.7835894  8.3815200   8.565 8.05e-15 ***
df_2015$IncomeLevel Low Income                   -0.0295292  0.0226557  -1.303 0.194291
df_2015$IncomeLevel Lower Middle Income          -0.0062553  0.0166775  -0.375 0.708095
df_2015$IncomeLevel Upper Middle Income           0.0163431  0.0107402   1.522 0.130042
df_2015$`Individuals using the Internet (% of population)` -0.0008351  0.0002416  -3.457 0.000699 ***
df_2015$log_Govt_Exp                             -0.0039089  0.0042548  -0.919 0.359619
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Selection: Model 1,
Trained by: 2015 Data

- y = National Diabetes & Kidney Disease Death Rates
- b_0 = Intercept
- x_1 = Diabetes & Kidney Disease Incidence Rate
- x_2 = Low Income Nation
- x_3 = Lower Middle Income Nation
- x_4 = Upper Middle Income Nation
- x_5 = Prop. of Population Internet Access

Detected collinearity between the two economic associated variables, likely due to nation’s of higher wealth having higher rates of governmental health expenditure.

Model Validation



Residuals vs Fitted:

- Relatively even band of values above and below the zero line.
- Fairly random bouncing of values around the horizontal line.
- A couple outliers detected.

Suggesting:

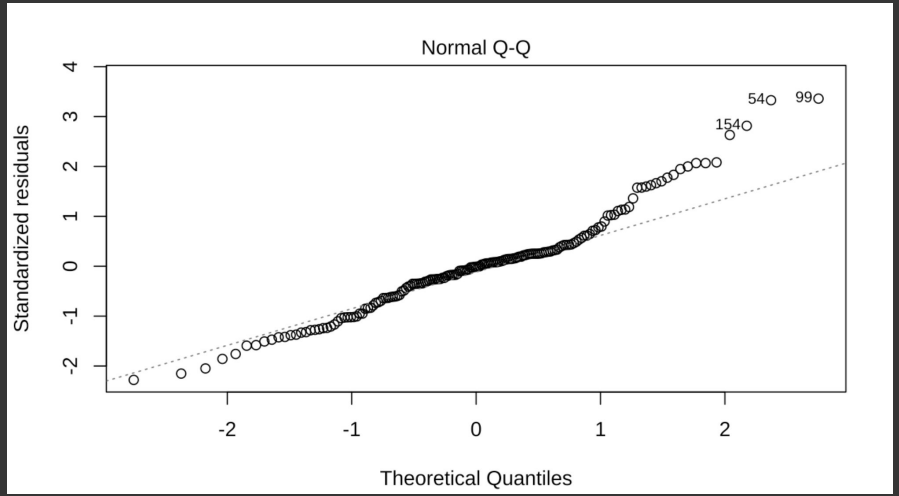
- Variances of error terms are equal.
- Linear model is a reasonable approach.

Normal Q-Q Plot:

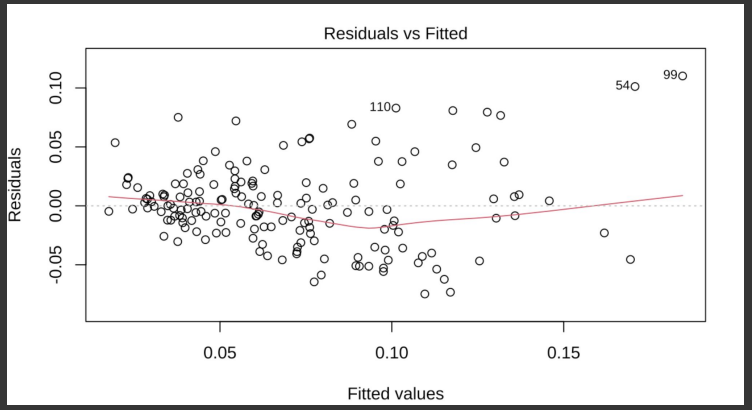
- The majority of points land approximately along the straight line.
- A couple of outliers identified.

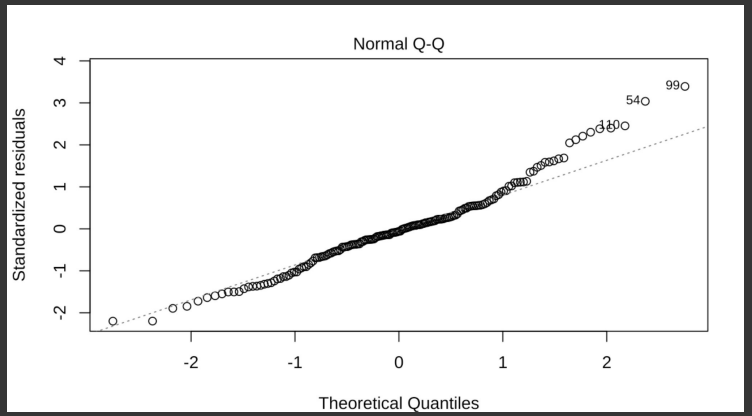
Suggesting:

- Some extreme values are present in data, as indicated by the curvature of the plotted points near the upper and lower ends.
- Indicating light-tailed data.



Model Performance






```
Call:
lm(formula = df_2019$Data ~ df_2019$Diabetes_Kidn_Incid + df_2019$IncomeLevel +
    df_2019$`Individuals using the Internet (% of population)`)

Residuals:
      Min        1Q    Median        3Q       Max
-0.074768 -0.019894 -0.002002  0.017963  0.110192

Coefficients:
                                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                                        0.0315476  0.0170653   1.849  0.06632 .
df_2019$Diabetes_Kidn_Incid                       64.8216162  7.2719757   8.914  9.5e-16 ***
df_2019$IncomeLevel Low Income                    -0.0084824  0.0152581  -0.556  0.57902
df_2019$IncomeLevel Lower Middle Income            0.0129438  0.0115414   1.122  0.26372
df_2019$IncomeLevel Upper Middle Income            0.0264379  0.0083512   3.166  0.00185 **
df_2019$`Individuals using the Internet (% of population)` -0.0008671  0.0001430  -6.065  8.9e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0345 on 163 degrees of freedom
Multiple R-squared:  0.4974,    Adjusted R-squared:  0.482
F-statistic: 32.26 on 5 and 163 DF,  p-value: < 2.2e-16
```

Plots, p-values and AIC score of -644.7081 all indicate the model is a viable option for predicting future values.

Predicted vs. Actual 2019 Values:

RMSE: 0.03387909

Good model performance across the dataset.

Conclusion

- Income groups (Low, Lower Middle, Upper Middle, High Income) and Internet Access seem to be useful to model and predict worldwide diabetes and kidney diseases death rates

Future Scope

- Extending this approach to other diseases like **Cardiovascular diseases** since it would behave similarly to Diabetes and kidney diseases and we would expect to be impacted by similar demographic factors

- On the other hand, death rates for **Self-harm and interpersonal violence** would involve different factors and would likely require both economic and public health associated data which is not readily available

- Diving deeper into a **particular country or continent** and focus more on trends and patterns for the same

References

***Healthcare spending and health outcomes: evidence from selected East African countries*.** Murad A Bein,1 Dogan Unlucan,2 Gbolahan Olowu,1 and Wagdi Kalifa2. African Health Sciences. 2017 Mar; 17(1): 247–254. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5636241/

Mortality and life expectancy forecast for (comparatively) high mortality countries. Ahbab Mohammad Fazle Rabbi & Stefano Mazzuco. Genus volume 74, Article number: 18 (2018).. https://genus.springeropen.com/articles/10.1186/s41118-018-0042-x

Diabetes and Chronic Kidney Disease. Centers for Disease Control and Prevention. U.S. Department of Health & Human Services. https://www.cdc.gov/diabetes/managing/diabetes-kidney-disease.html#:~:text=Both%20type%201%20and%20type%202%20diabetes%20can%20cause%20kidney%20disease.&text=Kidney%20diseases%20are%20the%209th,begin%20treatment%20for%20kidney%20failure

What Are the Socio-Economic Predictors of Mortality in a Society?. Wahab Adewuyi Adejumo1, Adetunji Raimi Tijani1, Sheriff Adesanyaonatola2, 1 Department of Insurance, the Oke-Ogun Polytechnic Saki, Oyo State, Nigeria. 2Department of Insurance, the Polytechnic Ibadan, Oyo State, Nigeria. Journal of Financial Risk Management > Vol.8 No.4, December 2019. https://www.scirp.org/journal/paperinformation.aspx?paperid=96881