Analysis of Contemporary Human Lifespan

Behzod Mirpochoev
University of Colorado at Boulder
Boulder, Colorado, USA

Malvika Manohar
University of Colorado at Boulder
Boulder, Colorado, USA

Suresh Nayak
University of Colorado at Boulder
Boulder, Colorado, USA

Sukeerth Balakrishna Kalluraya
University of Colorado at Boulder
Boulder, Colorado, USA

1 MOTIVATION

The data set that we have selected is a collection of structured information on the life, work, and the death of more than a million deceased notable people. As a team, we are very curious to analyze and understand the various factors that impact the life and death of people of all ages, gender, occupation, and nationalities. It is our aim to mine this data in order to identify any meaningful patterns and correlations across this data set, especially in the form of an attribute influencing other attributes within the data set. By the end of our analysis, we wish to get closer to answering certain captivating questions about the nature of society and lifespan in humans, such as: "Are famous people of a certain region more likely to die young than others?", "How frequent is death by suicide, and is it in any way linked to either gender or occupation?", "Which occupations are riskier and lead to more accidents causing death?", "How does the life expectancy of someone famous/wealthy compare to the rest of the population?"

We believe that mining such a data set may provide us with useful insights into the life of the famous, the life of those less fortunate, and overall society. This may help us to identify risks that accompany such a lifestyle. By gaining a general understanding of the risks that are involved in a certain profession, region, age, etc, we may provide a basis for appropriate action to be taken, in order to better the quality of life and the general life expectancy of the people involved.

2 LITERATURE SURVEY

An associated paper was written with regards to the data set observed. The "Age dataset", as described by Annamoradnejad et. al., was primarily motivated in construction to have a common and organized data set for historical figures. [5] The primary approaches used related to machine learning and text mining practices, in which the authors developed a five-step method in order to infer data such as birth date, death date, proposed gender, and occupation. [5] The results of these processes were chiefly the elucidation of missing components in the data set and the methods which were used to arrive at these components; the authors' intention was to allow for future work to be done to discover interesting, non-trivial relations between primarily age, gender, and occupation. [5]

Another paper of note was one which delineated the discrepancy between death records of upper to middle and lower-middle to lower income countries. Systemic Cause of Death (CoD) records have improved dramatically since the period of time in Europe, Asia, and North Africa for the bubonic plague pandemic. [3] However, when considering inter-country and time comparisons of death records, the goal of attaining ubiquitous and consistent notes regarding CoD was not previously attained through standard practice of record keeping by physicians and associated parties. [3] Thus, there was a considerable motivation for amending discrepancies and incomplete values with CoD; the Global Burden of Disease (GBD) tool was developed in order to attain estimates of cause-specific mortality across age, sex, year, and location. [3] A given premise of the utilized methods involved a given initial classification of codes given for CoD, the redistribution of said codes, an analysis to see the extend of a potential multiple causes case through negative correlation, impairments, and proportional redistribution. [3] The GBD processes were utilized to redistribute how deaths were recorded for a more comprehensive list.

A more thorough examination of methodologies utilized for prediction with regards to cause of death is another essential piece of background. CoD in low-income regions where death recordings are not well incentivized and deaths frequently occur at home rather than in the hospital. As we have previously mentioned, it becomes necessary to attempt to predict the CoD for a ubiquitous and consistent reference to further mine our data. However, the selection of which tool in which to do so is a logical next step in deciding. [4] Various supervised machine learning techniques may be deployed, including k-nearest neighbor (kNN), gradient boosting modeling (GBM), support vector machine (SVM), classification and regression tree (CART), artificial neural network (ANN), and C5. [4] These methods may be examined through evaluation metrics such as accuracy, recall, specificity, precision, F1, and Cohen's kappa. [4] Given these algorithms and these metrics, comparing the models against each other results in SVM displaying promising results; the rest of the models aside from CART perform decently well when assigned to CoD prediction out of a given sample of diseases. [4]

The process of finding interesting connections between risk factors and possibly predicting life expectancy as a result would be the next concern to address. Associated risk factors explored include sociodemographic background, lifestyles, dietary factors, life satisfaction, and metabolic health. [2] The categorization of these risk factors further into a questionnaire and was given; age, sex, and education comprised the sociodemographic background; history of myocardial infarction or diabetes in the family, body mass index (BMI), blood pressure, and cholesterol were included as part of medical history; dietary habits, smoking, drinking, and physical activity were included as lifestyle; stress, accomplishments, work-life, familial relationships, financial situation, and further interpersonal relationships were included as life satisfaction. [2] A Poisson regression model was utilized to represent the impact of these parameters across a more general hazard function with respect to time; the impacts of each risk factor on age was observed one at a time for simple regression, and then the model was fully adjusted for all risk factors for multiple regression. [2] An expected age of death (EAD) was also computed using expectation of a discrete probability density function. [2] Each risk factor aforementioned was found to be statistically significant in impacting the hazard of death; biological risk factors and lifestyles along with life satisfaction were demonstrated to clearly affect EAD. [2] The regression model allows for a closer look into specific risk factors which impact life expectancy.

Events which are infrequent are important in finding non-trivial relationships in lifespan. Suicides are often determined by medical examiners and coroners (ME/C) as either accidental, suicidal, or undetermined. [1] Notable findings included the collapse of the two groups of accidental or undetermined when compared against determined suicides. [1] Common factors across all classifications included mental illness and indication of pain, whereas suicides were distinguished from the other two categories as having significantly more male decedents and violent CoD. [1] Combining categorizations and comparing against another rare event seems to be a viable detection method.

3 PROPOSED WORK

We start of by downloading the single data set that we plan on using, on each of our systems. Following the standard practices of data mining, we plan to clean the data set by removing certain data objects which may be irrelevant and unnecessary for our purposes. This includes removing the data entry for the creator of the data set. Besides this, we plan to handle missing values.

We succeed the data cleaning step with some data pre-processing. As part of this, we have identified attributes which have inconsistent values. We plan to identify a few classes per attribute, and convert all the values of that attribute so that they belong to one of these classes. Taking the 'Gender' attribute as an example, we plan to convert a value of 'Transgender Male:Male" to "Transgender Male" to maintain consistency across the data set. Similarly, we find numerous data objects with multiple comma-separated values for the 'Occupation' attribute. We plan to convert such values into an array of occupation classes in order to simplify their usage.

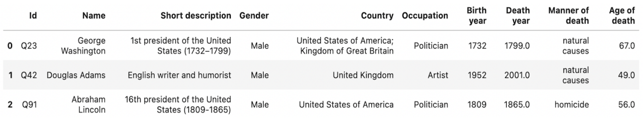
Since the information in the data set is not restricted to a certain time period, we plan to filter out those data objects which represent people who lived before the 1700s. Additionally, we plan to handle the 'Country' attribute values which also contain olden-day regions and empires. In the end, we aim to convert the categorical attributes into numerical values for the purpose of feeding the attribute to the ML model.

Once the data set has been cleaned and pre-processed, the idea is to find any patterns across the data set which may indicate a correlation of a certain attribute on another attribute.

Our work shares similarities with a few other analysis done in the past, such as those on occupation risks, suicide, region wise life-expectancy, etc. Our goal is to aggregate all of these factors and find any underlying common themes across them. Similarly, we've seen numerous studies done on the general public belong to specific countries, while we wish to compare such analysis with that of just the notable and deceased, as seen in this data set.

4 DATA SET

The dataset used for our proposed method is the Age dataset [5] published through the ICWSM and is made publicly available as a .csv file. The dataset is structured in nature and is fairly large, containing 1,222,997 data objects, with each data object representing a historical figure. The dataset contains 10 attributes for each data object including ID, Name of the notable person, Short description of said person, Gender, Country, Occupation, Birth year, Death year, Manner of Death and Age of death (as seen in figure 1).



	Id	Name	Short description	Gender	Country	Occupation	Birth year	Death year	Manner of death	Age of death
0	Q23	George Washington	1st president of the United States (1732–1799)	Male	United States of America; Kingdom of Great Britain	Politician	1732	1799.0	natural causes	67.0
1	Q42	Douglas Adams	English writer and humorist	Male	United Kingdom	Artist	1952	2001.0	natural causes	49.0
2	Q91	Abraham Lincoln	16th president of the United States (1809-1865)	Male	United States of America	Politician	1809	1865.0	homicide	56.0

Figure 1: Subset of Age Dataset. (https://www.kaggle.com/datasets/imoore/age-dataset).

As observed in Figure 1, the attributes are of both nominal and numerical types and are hence to be preprocessed in their respective ways before being fed into the model. Most of the attributes are of nominal type barring Birth and Death year which are interval-scaled numeric attributes and Age which is a ratio-scaled attribute.

In the table below, we can observe the central tendancy of the nominal attributes of the dataset as well as the number of unique values each of them have.

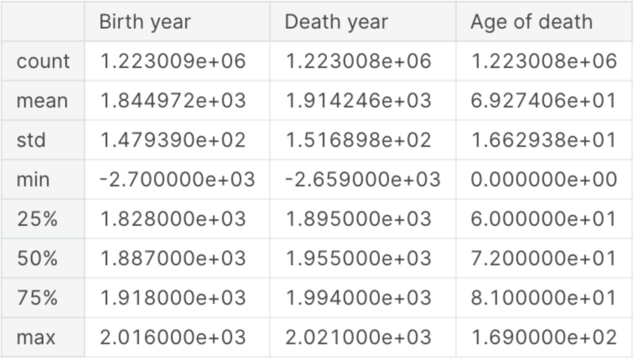
Attribute	Mode	No. of Unique Values
ID	None	1222997
Name	None	1130871
Short Description	American politician	512439
Gender,	Male	20
Country	United States of America	5961
Occupation	Artist	9313
Manner of Death	natural causes	206

A statistical description of the data distribution of numerical attributes of this dataset is shown in figure 2.

5 EVALUATION METHODS

For statistical analyses, we will be doing Correlation tests, Chi-Square tests, Normality tests, etc. Categorizing data into different sets based on the region/country allows us to apply the Chi-Square method to verify the correlation between the patterns.

In terms of comparison of methods, we would run the analysis of control variables the same in terms of risk factor and lifespan and then see how each method performs via a metric. For example, we may train a k-Nearest Neighbor classifier and a decision tree classifier and compare the results of the two models. For evaluating our prediction models, we would use evaluation metrics as described in Idicula-Thomas S. et. al. in 2021, such as accuracy, precision, recall, F1 score. [4] Other metrics for consideration of evaluation may include Log Loss, ROC, and the confusion matrix.



	Birth year	Death year	Age of death
count	1.223009e+06	1.223008e+06	1.223008e+06
mean	1.844972e+03	1.914246e+03	6.927406e+01
std	1.479390e+02	1.516898e+02	1.662938e+01
min	-2.700000e+03	-2.659000e+03	0.000000e+00
25%	1.828000e+03	1.895000e+03	6.000000e+01
50%	1.887000e+03	1.955000e+03	7.200000e+01
75%	1.918000e+03	1.994000e+03	8.100000e+01
max	2.016000e+03	2.021000e+03	1.690000e+02

Figure 2: Data Distribution of Numerical attributes.

6 TOOLS

The tools required for the implementation of our proposed work include:

- pandas: for data cleaning and manipulation.
- numpy: for data transformation and analysis.
- sklearn: For object clasification, clustering.
- matplotlib: creating static and interactive visualizations.

7 MILESTONES

- Week 1 [23rd May - 29th May]: Project Description
- Week 2 [30th May - 5th]: Data Collection
- Week 3 [6th June - 12th June]: Data Exploration and basic Data Preprocessing
- Week 4 [13th June - 19th June]: Literature Review
- Week 5 [20th June - 26th June]: Further Data Preprocessing and Feature Engineering
- Week 6 [27th June - 3rd July]: Documentation (Project Proposal)
- Week 7 4th July - 10th July]: Model Building and Training
- Week 8 [11th July - 17th July]: Finetuning Model
- Week 9 [18th July - 24th July]: Documentation (Progress Report)
- Week 10 [25th July - 31st July]: Model Evaluation and Finetuning

- Week 11 [1st August - 7th August]: Documentation (Final Project Report)
- Week 12 [8th August - 10th August]: Refine Documentation (Final Project Report) and Project Submission

REFERENCES

[1] Gray D. et. al. 2014. Comparative Analysis of Suicide, Accidental, and Undetermined Cause of Death Classification. *HHS Public Access* (2014). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4411039/pdf/nihms-680013.pdf.

[2] Härkänen T. et. al. 2020. Estimating expected life-years and risk factor associations with mortality in Finland: cohort study. *BMJ Open* (2020). https://bmjopen.bmj.com/content/bmjopen/10/3/e033741.full.pdf.

[3] Johnson S.C. et al. 2021. Public health utility of cause of death data: applying empirical algorithms to improve data quality. *BMC Medical Informatics and Decision Making* (2021). https://bmcmedinformdecismak.biomedcentral.com/track/pdf/10.1186/s12911-021-01501-1.pdf.

[4] Idicula-Thomas S. Gawde U. Jha P. 2021. Comparison of machine learning algorithms applied to symptoms to determine infectious causes of death in children: national survey of 18,000 verbal autopsies in the Million Death Study in India. *BMC Public Health* (2021). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8488544/pdf/12889_2021_Article_11829.pdf.

[5] Annamoradnejad I. Annamoradnejad R. 2022. Age dataset: A structured general-purpose dataset on life, work, and death of 1.22 million distinguished people. *Association for the Advancement of Artificial Intelligence* (2022). http://workshop-proceedings.icwsm.org/pdf/2022_82.pdf.