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 ABSTRACT

Humanity is faced with the inevitable concept of the end. The interesting aspects of the end are that which involve questions about how it comes about, in a timely or untimely manner. These lead to further questions regarding what is considered "timely"; the aim being to quantitatively and definitively understand these concepts through data mining techniques. Through analysis, several captivating questions about the nature of society and lifespan in humans, such as: Among famous personalities, is the percentage of death by suicide more skewed towards either gender? Which occupations involve a higher degree of accidents leading to death? Has the general life expectancy of famous people increased with time? Given information about a person, such as country and year of birth, gender, and occupation, can we predict one's lifespan? The results of such questions allowed for some insight as to the following: professions which correlated with a higher expected lifespan such as artists, researchers, and doctors; the nature and factors (notably gender) of rare event deaths such as suicide; the prediction of how lifespan may be affected given the roles of gender, occupation, a description of the individual in question, country of origin, as well as time period in which the individual lived.

2 INTRODUCTION

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. Additionally, we plan to train a regression model that can predict the lifespan of a given individual based on his country and year of birth, gender, and occupation.

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.

3 RELATED WORK

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.

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, their Gender, the Country they're a citizen of, their Occupation, their Birth year, their Death year, Manner of Death and their age when they died (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 tendency of the nominal attributes of the dataset as well as the number of unique values each of them has. We have used mode to describe the central tendencies of the attributes and since ID and Name are unique to each data object, their mode values are None.

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 graphical description, i.e. frequency histograms, in order to comprehend the data dispersion of the numerical attributes of this dataset is shown in figure 2 and 3.

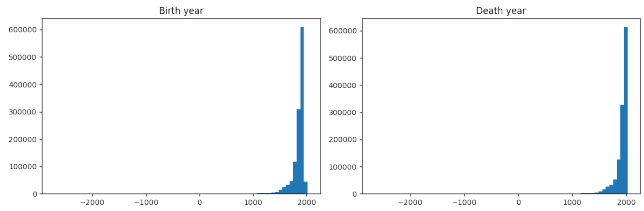

Figure 2: Distribution of the attributes Birth and Death Year.

As observed in figure 2, most of the births and deaths of the people in this dataset seem to have been after the 16th century. The life expectancy of most people is seen to be around 65 to 85 years old going by the distribution observed in figure 3

To further understand the data objects a five point summary, which provides a statistical description of data dispersion, of Birth year, Death year and Age of death is tabulated below in figure 4.

Figures 5 - 8 show histograms depicting the most common values for the 'Country', 'Gender', 'Occupation' and the 'Manner of death' attributes found in the data set. These graphs help us get a high level overview of each attribute.

Using figure 5, we understand that the data set is skewed heavily towards 'Male' gendered people by almost 9:1. Furthermore, we notice that the 18 other 'Gender' values found apart from 'Male' and 'Female' values are very few and sparsely spread across in the data set.

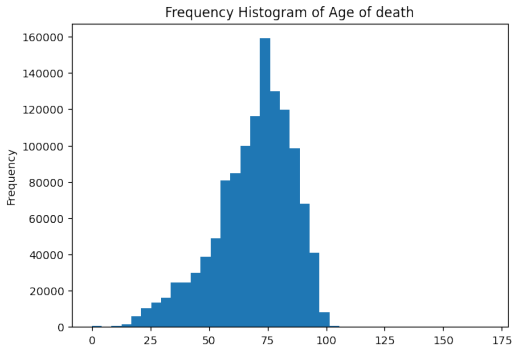

Figure 3: Distribution of the attribute Age of Death.

	Birth year	Death year	Age of death
count	1223009.0	1223008.0	1223008.0
mean	1845.0	1914.0	69.0
std	148.0	152.0	17.0
min	-2700.0	-2659.0	0.0
25%	1828.0	1895.0	60.0
50%	1887.0	1955.0	72.0
75%	1918.0	1994.0	81.0
max	2016.0	2021.0	169.0

Figure 4: Data Distribution of Numerical attributes.

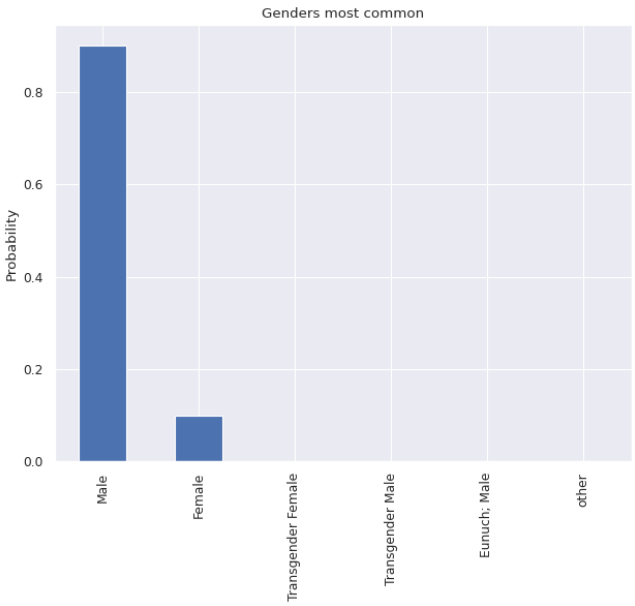

Figure 5: Most common 'Gender' values found in the data set

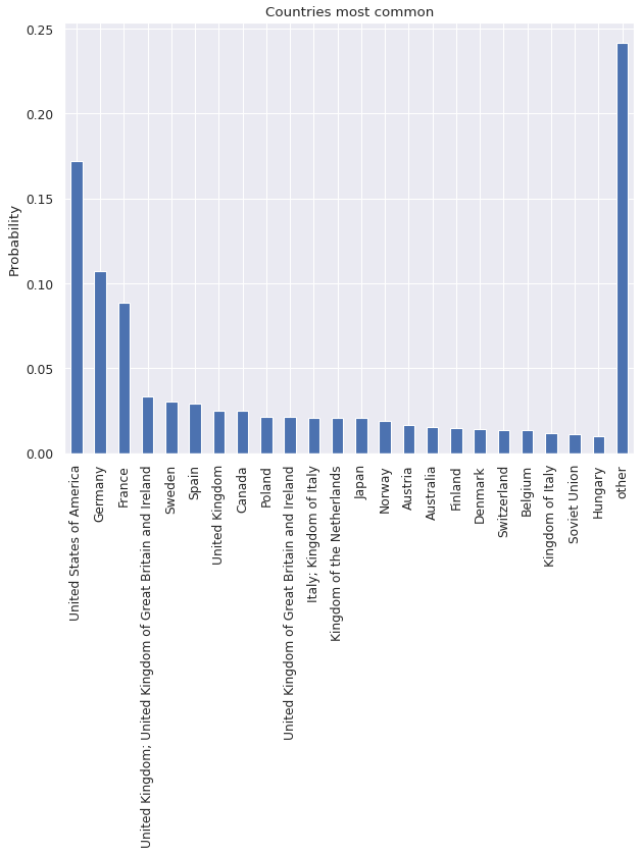

Figure 6: Most common 'Country' values found in the data set

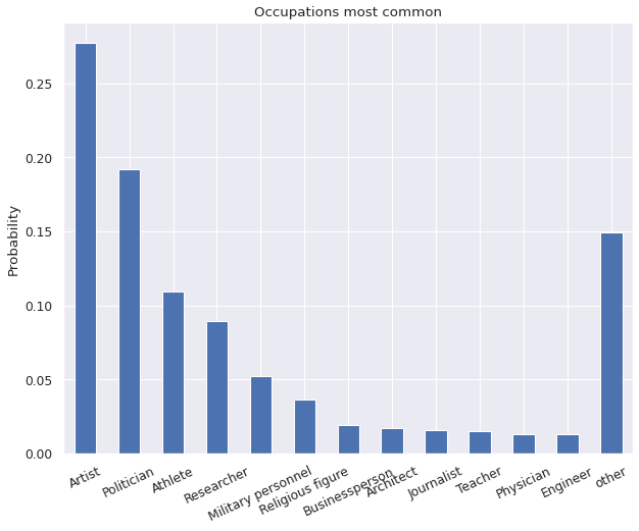

Figure 7: Most common 'Occupation' values found in the data set

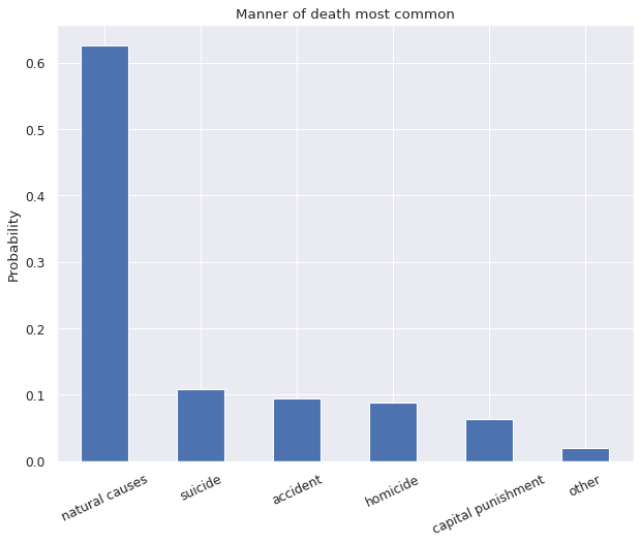

Figure 8: Most common 'Manner of Death' values found in the data set

A similar analysis on the 'Country' attribute can be made using figure 6. It is easily observed that a few countries are repeated in the data set, especially in scenarios where we have comma separated multiple values. Additionally, we notice a few olden day states such as the 'Soviet Union'.

Moving on to the 'Occupation' attribute as shown in figure 7, we observe a variety of occupations, with a massive skew towards 'Artists'. Roughly 1 in every 4 entry in the data set is an artist, possibly accounting for the high number of popular musicians, film makers, etc.

Figure 8 demonstrates the spread of the 'Manner of death' attribute values. A majority of people have been known to die under 'natural causes'. These data objects would be useful to analyze the life expectancy in certain regions as natural cause deaths can shed light on the general living of countries. The other values such as suicide, accident, homicide also provide useful information to analyze upon. We may study the circumstances leading to suicide, or analyze which occupations are more accident prone. Similarly, diving into the homicide data entries may tell us more about the crime rates and security found in different regions.

5 MAIN TECHNIQUES APPLIED

5.1 Data Preprocessing

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 cleaned 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.

We succeed the data cleaning step with some data preprocessing. As part of this, we have identified attributes which have inconsistent values. We identified a few classes per attribute, and converted all the values of that attribute so that they belong to one of these classes. Taking the 'Gender' attribute as an example, to maintain consistency across the data set, we categorize the attributes into Male, Female, and Other.

Since the information in the data set is not restricted to a certain time period, we have filtered out those data objects which represent people who lived before the 18th century.

Additionally, we performed text analysis tasks on 'Short Description' in order to extract knowledge from it. After performing preliminary the data preprocessing tasks listed above, the attribute Short Description is to be prepped for NLP tasks. For this, firstly all the text was converted to lowercase. Then, any punctuation tabulation and digits were removed from the text. We then removed any stop words from the text. Word clouds were used to identify and thus remove redundant words such as politicianenglish, politicianbritish, and so on. Finally lemmatization of the text was done in order to group together the different inflected forms of a word so they can be analyzed as a single entity.

In the end, we converted the other categorical attributes, including Gender, Country, Ocuupation and Manner of Death, into numerical values, i.e. we encoded the data using LabelEncoder, for the purpose of performing correlation analysis between said attributes and thus feeding them to the ML model.

Once the data set has been cleaned and pre-processed, analysis is done on the data to identify any patterns present which may indicate a correlation of a certain attribute to another attribute.

5.2 Evaluation Methods

For statistical analyses, we have done 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.

5.3 Tools

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

- pandas: For data cleaning and manipulation.
- numpy: For data transformation and analysis.
- pandasql: For querying pandas dataframe using SQL syntax.
- sklearn: For object classification, clustering.
- matplotlib and seaborn: For creating static and interactive visualizations.

6 KEY RESULTS

One of the interesting topics explored thus far involves the age of death, occupation, and gender identity of those who have a cause of death (CoD) that is not classified as natural, the primary focus being suicide. Comparing the distributions of the age of death for all contemporary notable humans against those who have CoD listed as suicide, a massive skew occurs.

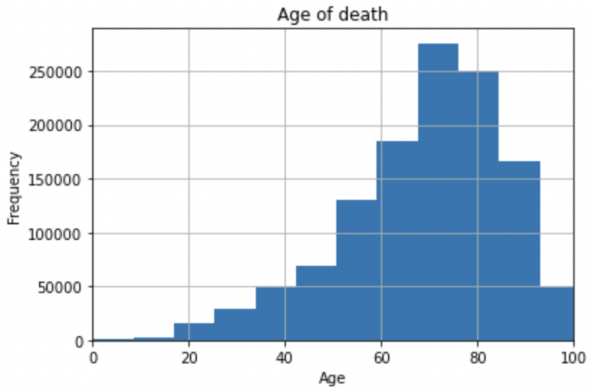

Figure 9: Frequency of age of death of contemporary humans

Given the figures, one may see that the age of death is a negatively skewed distribution, with a mode of approximately 70-75. Compare this to the distribution when suicide

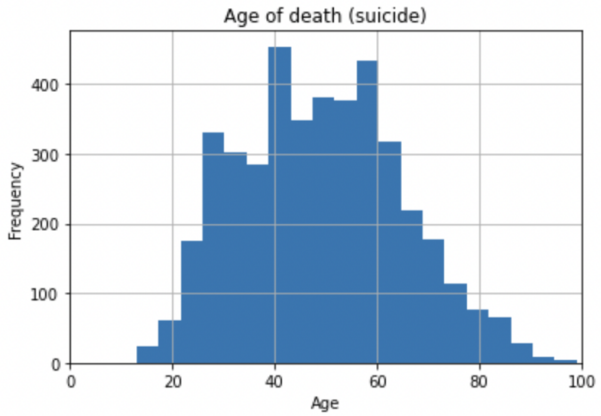

Figure 10: Frequency of age of death given CoD as suicide

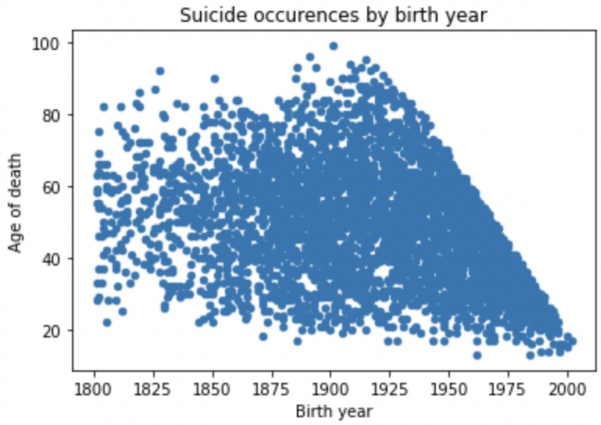

Figure 11: Age of death by suicide given a birth year

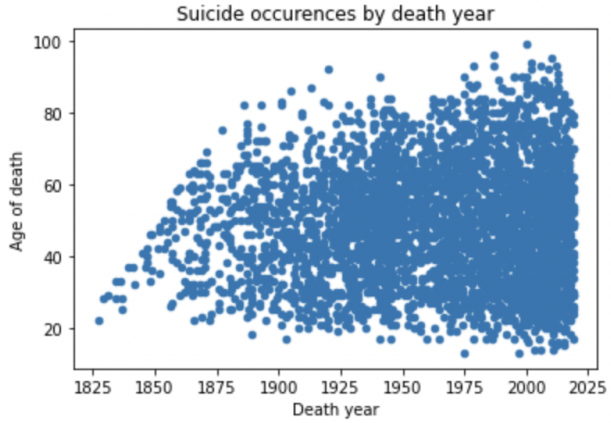

Figure 12: Age of death by suicide given a death year

is the CoD, which relative to the previous figure, demonstrates a positive skew. It is notable that the average age of suicide here tends towards mid-life of the contemporary human, at around 40-50.

One may then analyze the reasoning behind such a drastic shift by taking a closer examination of the description of each individual's attributes. We see that the dataset has in itself a bias towards including males, with a rate of 9.101 males for every female in the dataset. Then compare this with the male-to-female ratio in the suicide subset, with 5.686 males for every female in the set. This set suggests that despite an overarching bias in the sampling of males to females, the suicide rate of females causes a notable shift in male-to-female ratio in death across contemporary society. A proposed next step is the following: one must first eliminate bias by introducing a random sampling of each gender's cause of death while ensuring that there is an even split of males to females, and examine the male-to-female suicide ratio after multiple such samplings.

Another examination as for the shift in age of death in suicide is to determine birth and death years. These components may indicate an area to further explore, e.g. examining occupation during a birth or death year may have some indication of emotional/mental risk factors given such a time period.

It appears that there is no correlation of age of death by suicide given a birth or death year. There is a limiting slope given the birth year beginning around 1925, however this is simply the effect of not having enough time to live in general. The same may be said about the limiting slope in age of death for death year as there has not been enough time elapsed for a person to attain a given age (as part of data cleaning/preprocessing, a definition of contemporary human was given to those born in 1800).

Of all the people born in the last 200 years with available CoD, about 10% have suicide as their CoD. Of these people, the ratio of male to female is roughly 6:1. We notice that the average age of a male person who committed suicide in this time period is close to 49 years, while that of females was around 44 years.

Interestingly, we notice that in the case of female personalities, the graph is skewed positively indicating shorter lifespans at the time of death. It is hard to inference too much from this data due to the much fewer number of entries for females compared to males. Nonetheless, there may be a case for studying why females tend to commit suicide at an

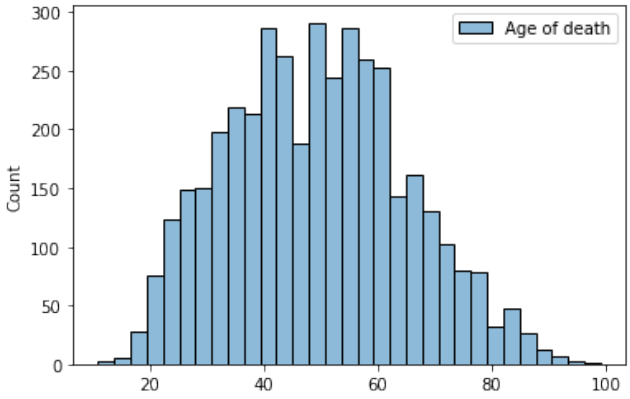

Figure 13: Histogram showing frequencies of lifespans for males who committed suicides

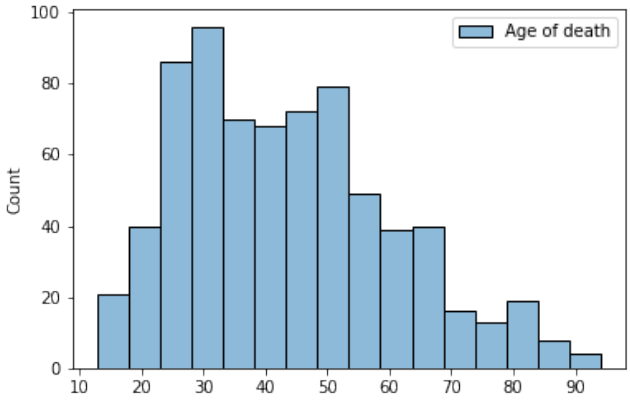

Figure 14: Histogram showing frequencies of lifespans for females who committed suicides

earlier age than males, and if society and cultural pressures have any part to play in them.

On a similar note, when we analyze the impact of nationality on the age at which people commit suicide, we notice that the most common listed country which is the USA, takes up about 22% of the distribution. This is alarming especially given that the next highest country on the list makes up only for 6% of the whole. Although this dataset may not be able too speak too much about why such a trend exists, it may prove useful to study any underlying reasons for such a skewed observation against the USA in terms of suicide rates.

The next area that we explored is the occupation/short description of a person who has committed suicide to determine any correlative effects between the aforementioned and the age of death; once this correlation is established, comparing it with the total amount of those found with the number present in the original dataset revealed those occupations/short descriptions which are significant in having a correlation with suicide. Hypothesis testing followed by regression analysis is the course of action we have taken.

Upon analyzing the age of death concerning natural and unnatural causes, we observed that the death caused by natural causes follows the Gaussian distribution peaking at age 60.

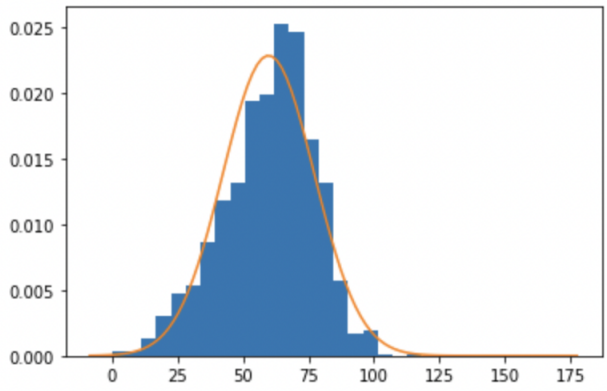

Figure 15: Distribution of "age of death" for natural causes.

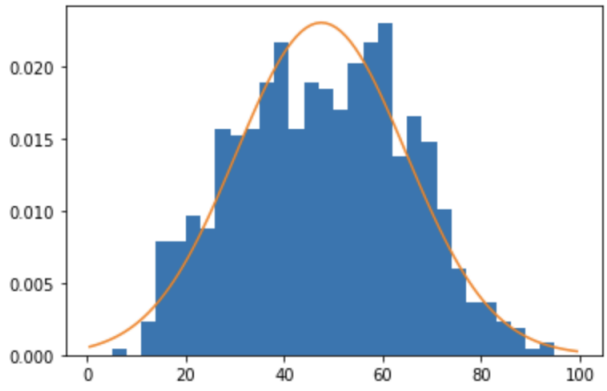

Figure 16: Distribution of "age of death" for unnatural causes.

In contrast, unnatural deaths follow the multi-modal distribution. The interesting pattern here is that the peaks formed for the unnatural deaths are around 60 and 30 years. For most data points, the occupation is either artist or athlete. The athletes are more likely to get injured during their peak

time, i.e., between 22-35, and are more prone to fatal casualties during retirement. This causes the graph to have an additional spike at age 35. If we ignore the spike, the rest of the distribution follows the normal distribution as seen in the case of natural deaths.

Plotting a box graph for the attributes after the year 1800 gives us the general trend in life expectancy. For both men and women, the median life expectancy has increased linearly. This trend can be attributed to progress made in the medical field. Plotting the correlation matrix for the Age

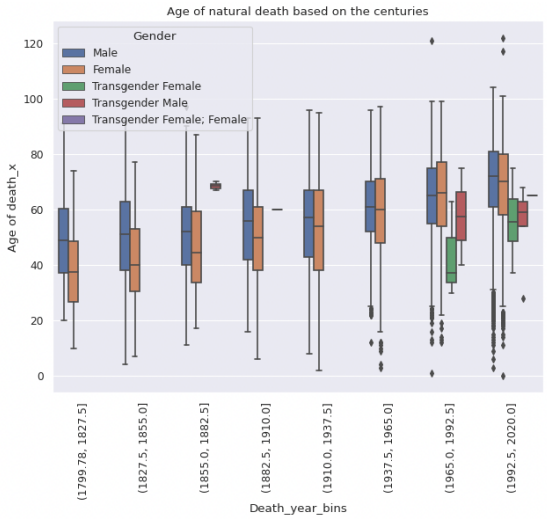

Figure 17: Age of death for attributes after the year 1800

Data set, we can infer that the age of death is highly co related to manner of death. and all the other attributes are not co related to each other.

	Age of death	Gender	Country	Occupation	Birth year	Manner of death
Age of death	1.00	-0.00	0.13	-0.02	0.08	0.25
Gender	-0.00	1.00	-0.03	0.10	-0.04	-0.06
Country	0.13	-0.03	1.00	-0.09	0.09	0.06
Occupation	-0.02	0.10	-0.09	1.00	-0.16	-0.10
Birth year	0.08	-0.04	0.09	-0.16	1.00	0.11
Manner of death	0.25	-0.06	0.06	-0.10	0.11	1.00

Figure 18: correlation matrix

The data-set is highly skewed in terms of the Occupation attribute, with most of the data falling under 10 Occupations (out of 745 unique values); namely, Artist, Politician, Relational figure, Researcher, Military Personnel, Aristocrat, Physician, Architect, Jurist, and Judge. Figure 19 visualizes the life expectancy of the people in this dataset based on these top 10 occupations.

On analyzing, by querying the database, the most common age of death for each top occupation, it was found that Artists, Politicians, Researchers, Physicians, and Jurists mostly live to 70 years of age. The same can be observed in Figure 19, though the life expectancy of artists and researchers tends to be a little longer. The idea that "art is a medicine that keeps us alive" is definitely re-established by our analysis, at least as far as notable people are concerned.

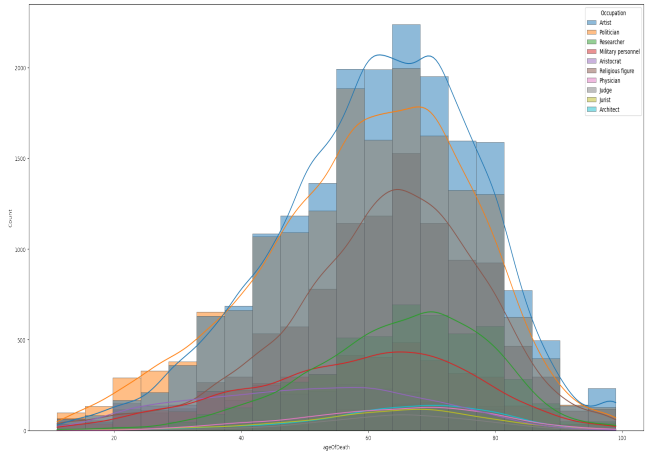

Figure 19: Age of Death by Occupation

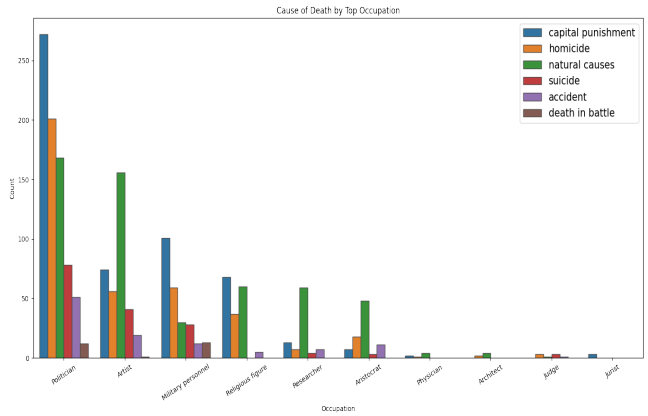

Figure 20: Manner of Death by Occupation

Another interesting facet to infer was one of cause or rather manner of death of people in a certain occupation. Artists and Researchers not only have a higher life expectancy but also die mostly because of natural causes. The deaths of people who worked in politicized environments such

as politicians, military personnel ,and religious figures are mostly due to capital punishments then followed by homicide. Even if the work environment does not majorly impact the life expectancy of a person, it does seem that the manner of death is impacted.

A theory we further wanted to explore was whether the life expectancy of people in certain roles within an Occupation are different; for example, we may assume that a president is more likely to die sooner, albeit due to unnatural causes, than say a the Minister of Education or Environmental Energy or a King is more likely to die sooner than governors. To assess this notion, we found that the attribute 'Short Description' and "Occupation' are correlated, i.e., Short Description was essentially seen to be a description of what made the person in question notable; which was their Occupation in most cases. Hence, we performed text analysis tasks on 'Short Description' and extract knowledge from it. A barplot is used to visualize keywords that occurred the most. Two comparable keywords, in terms of their frequencies, were found to be 'king' and 'governor'. Figure 21 represents the distribution of age of death of notable persons who were either kings or governors. The average age of death of people who were kings was 51 years whereas that of governors was 62 therefore reestablishing our theory. But this conclusion may not be due to their occupation as there is not much data regarding the lifetstyle of these people. Another possible explanation for this theory could be that life expectancy has increased over the centuries and people who are governors are mostly born around 17th century whereas kings mostly are born in the 11th century.

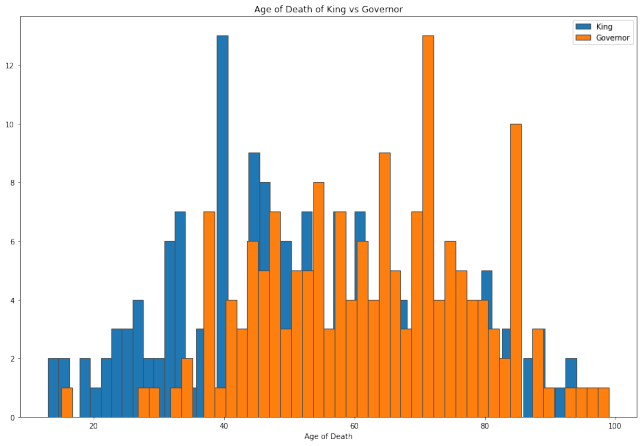

Figure 21: Age of Death of King vs Governor

We have fitted multiple regression models like K neighbors Regressor, Linear Regression, Decision Tree Regressor, and Random Forest Regressor. The comparison of these model is made by splitting the data into test and train data set to compute how well the model is fitting. Based on the result, the random forest regressor and K neighbors regressor give better results. The confusion matrix for the K neighbors regressor and Random Forest regressor are shown in the image.

```
[[  43  141   67   22   14    8    0    1    0    0]
 [   5  767 1027  415  257  176   60   12    0    0]
 [   0  118 1591 1629  633  393  175   10    1    0]
 [   0   11  397 2742 1777  802  359   19    0    0]
 [   0    2  105 1069 4177 2018  683   41    0    0]
 [   0    3   38  322 2417 4853 1430   66    3    0]
 [   0    3   16  129  624 3461 3651  219    2    0]
 [   0    1   11   70  198 1062 2931 1027    3    0]
 [   0    1    2   16   56  214  649  549   69    0]
 [   0    0    0    0    1    0    0    0    1    0]]
```

Figure 22: Confusion Matrix for random forest

```
[[  17   98  102   50   21    7    1    0    0    0]
 [   1  539  911  675  338  188   57   10    0    0]
 [   0   97 1217 1712  922  436  152   14    0    0]
 [   0    7  420 2320 2023  991  320   26    0    0]
 [   0    2   88 1278 3685 2309  671   62    0    0]
 [   0    2   25  395 2676 4512 1410  110    2    0]
 [   0    0   16  114  884 3721 3073  297    0    0]
 [   0    1    9   55  258 1433 2694  852    1    0]
 [   0    1    1   14   63  313  666  469   29    0]
 [   0    0    0    0    1    0    1    0    0    0]]
```

Figure 23: Confusion Matrix for KNN

Further, these models are used to predict the Death age based on attributes (Age of death, Gender, Country, Occupation, Birth year, and Manner of death). The model would predict the approximate age of death. Since the model we are using is a classifier, we have rounded off the predicted and expected values to the closest multiple of 10. These adjusted values are used for generating the evaluation matrix.

	precision	recall	f1-score	support
10	0.944	0.057	0.108	296
20	0.722	0.198	0.311	2719
30	0.436	0.267	0.332	4550
40	0.351	0.380	0.365	6107
50	0.339	0.455	0.389	8095
60	0.324	0.494	0.392	9132
70	0.340	0.379	0.358	8105
80	0.463	0.161	0.239	5303
90	0.906	0.019	0.037	1556
120	0.000	0.000	0.000	2
accuracy			0.354	45865
macro avg	0.483	0.241	0.253	45865
weighted avg	0.408	0.354	0.339	45865

Figure 24: Evaluation Matrix Random forest regressor

7 APPLICATIONS

The knowledge gained from our results may be utilized in consideration for the evolution of life expectancy. It is interestingly noted that certain professions are rather indicative of a longer lifespan. Choosing a profession, along with general success regarding a profession, is highly indicative of a longer lifespan. As such, career counselors/coaches, physicians, and talent managers may take this into consideration when assessing which jobs end up having humans live longest. The benefits of such may either entice people to join a career as an artist (so long as they are prolific), or may allow for a recommendation to be made given an individual/group preference for job security and life longevity. Another interesting result found was the correlation of rare event deaths such as suicide. The so-called "mid-life crisis" as applicable to mental health issues seems to hold some truth when considering the mean age of those who ended up with a CoD of suicide. To again reference physicians, it may give some indication of an imperative to focus on an age group of people nearing the age range as this result has been found to be consistent across contemporary society since the 19th century. An extension of the work would be to cross this data set with some of the other data sets mentioned in the literary reviews; the prior work done to regard some classifiers for gender based on names was an interesting idea that ended up not being as fully implemented as hoped. Another extension that would be interesting would be the further classification regarding the CoD of a rare event such as suicide to match more closely with the medical reports given would allow for some ubiquity as needed.

8 APPENDIX

The link to our GitHub repository may be found here. The link to our code specifically may be found in our interactive Jupyter Notebook (with accompanying video for visualization).

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.