Final Report of Traineeship Program 2023

On

“Analyze Death Age Difference of Right Handers with Left Handers”

MEDTOUREASY

28th May 2023

ACKNOWLDEGMENTS

I am Ajay Kumar, writing to convey my sincere appreciation and profound gratitude for the exceptional traineeship opportunity I had with MedTourEasy. The experience has been incredibly enriching, allowing me to gain valuable knowledge and a deeper understanding of the intricate subject of Data Visualizations in Data Analytics. Moreover, it has played a pivotal role in my personal and professional development.

I extend my heartfelt thanks to the Training & Development Team of MedTourEasy for extending this valuable opportunity to me and for their unwavering support throughout my traineeship. Their guidance and expertise have been instrumental in helping me comprehend the intricacies of the Data Analytics profile and equipping me with the necessary skills to successfully carry out the project, ensuring utmost client satisfaction. Despite their busy schedules, they graciously spared their valuable time to mentor me, for which I am truly grateful.

I would also like to express my gratitude to the entire team at MedTourEasy, as well as my colleagues, for fostering a highly productive and conducive working environment. Their collaboration and support have been instrumental in making my traineeship experience truly remarkable.

Once again, I would like to express my deepest appreciation to the Training & Development Team and MedTourEasy for providing me with this invaluable opportunity. I am truly honored and grateful for the knowledge and growth I have gained through this experience.

TABLE OF CONTENTS

ABSTRACT

This project aims to analyze the death age difference between right-handers and left-handers using age distribution data and Bayesian statistics. The conventional belief suggests that left-handers have a shorter lifespan compared to right-handers. However, this project challenges this claim by examining the changing rates of left-handedness over time and its potential impact on the average age at death.

The analysis utilizes a dataset containing information about the age at death and handedness (left or right) of individuals. By employing Bayesian statistics, we investigate the probability of being a certain age at death, given the reported handedness.

The project begins by importing the necessary libraries, including pandas for data manipulation, numpy for numerical computations, and matplotlib for data visualization. The age distribution data is then loaded into a pandas DataFrame for further analysis.

Data preprocessing is performed to handle any missing or invalid data, ensuring the accuracy and reliability of the results. Next, Bayesian statistics are applied to assess the probability of specific ages at death for left-handers and right-handers, respectively.

Through this analysis, we aim to provide empirical evidence to support or refute the claim of early death for left-handers. By investigating the influence of changing left-handedness rates over time, we hope to shed light on the potential factors contributing to any observed differences in death age between the two groups.

The findings of this project have implications for understanding the relationship between handedness and mortality, as well as challenging prevailing stereotypes regarding left-handed individuals. Ultimately, this research contributes to a broader understanding of the factors influencing lifespan disparities among different populations based on handedness.

Therefore, this project aims at collecting and analyzing wide variety of large data sets, create intuitive and interactive dashboards for representing relationship between handedness and mortality in order to gain meaningful insights.

I. INTRODUCTION

1.1 About the Company

MedTourEasy, a global healthcare company, provides you the informational resources needed to evaluate your global options. MedTourEasy provides analytical solutions to our partner healthcare providers globally.

1.2 About the Project

"Analyze Death Age Difference of Right Handers with Left Handers." This project aims to investigate the potential differences in average age at death between right-handers and left-handers using age distribution data and Bayesian statistics. By analyzing the changing rates of left-handedness over time, we seek to refute the claim of early death for left-handers and provide a data-driven perspective on this topic.

The project will utilize a dataset containing information on the age at death and handedness (left or right) of individuals. This dataset will be loaded into a pandas DataFrame, allowing for efficient data manipulation and analysis. We will employ Bayesian statistics, a powerful analytical framework, to examine the probability of being a certain age at death given the reported handedness.

To commence the project, the handedness data will be loaded from the National Geographic survey, and a scatter plot will be created to visualize the relationship between age and handedness. This initial analysis will help provide an overview of the data and identify any potential patterns or trends.

Next, we will add two new columns to the dataset: one for birth year and another for the mean left-handedness. The mean left-handedness will be calculated by averaging the "Male" and "Female" columns. We will then plot the mean left-handedness as a function of birth year, enabling us to observe any changes in left-handedness rates over time.

In order to investigate the probability of left-handedness given specific ages of death, we will create a function that calculates P(LH | A) for particular ages of death in a given study year. This function will utilize the mean left-handedness rates from the early 1900s and the late 1900s to estimate the probabilities. The results will provide insights into the association between handedness and age at death during different time periods.

Furthermore, we will load death distribution data for the United States, preprocess the data by dropping any NaN values, and plot the number of people who died as a function of their age. This analysis will provide a broader context for understanding the age distribution and serve as a reference point for comparing the death age difference between right-handers and left-handers.

To calculate the overall probability of left-handedness in the population for a given study year, we will develop a function called P_lh(). This function will utilize the death distribution data and the previously calculated probability of left-handedness given certain ages of death. By integrating these factors, we can estimate the overall probability of left-handedness in the population.

Additionally, we will create functions to calculate P(A | LH) and P(A | RH), which represent the probability of being a certain age at death given that an individual is left-handed or right-handed, respectively. These functions will consider the overall probability of dying at a specific age and the probabilities of left-handedness and right-handedness. The results will be plotted to visualize the probability distribution of being a certain age at death based on handedness.

To analyze the mean age at death for left-handers and right-handers, we will multiply the ages by their respective probabilities and calculate the sum using the numpy function np.nansum(). This calculation will provide us with the average age at death for each group, allowing for a direct comparison. The difference between the two average ages will be calculated to assess any disparity in lifespan between right-handers and left-handers.

Lastly, we will repeat the probability calculation from Task 8, but this time setting the study year parameter to 2018. This analysis will enable us to examine whether there have been any significant changes in the death age difference between right-handers and left-handers in more recent years.

In conclusion, the proposed project aims to analyze the death

1.3 Objectives and Deliverables

This project focuses on creating easily understandable, interactive and dynamic dashboards by gathering data of death distribution data for the United States from the year 1999 and rates of left-handedness digitized from a figure in this 1992 paper by Gilbert and Wysocki and using the coding language R and packages like readr, dplyr, ggplot, ggplot2, flex dashboard and other RShiny Packages to visualize these statistics which will enable the firm to analyze the situation and draw conclusions.

Objectives:

1. Investigate the death age difference between right-handers and left-handers using age distribution data.
2. Explore the changing rates of left-handedness over time and their potential impact on the average age at death.
3. Utilize Bayesian statistics to analyze the probability of being a certain age at death given the reported handedness.
4. Provide empirical evidence to support or refute the claim of early death for left-handers.
5. Challenge prevailing stereotypes and biases associated with left-handed individuals.
6. Contribute to a broader understanding of the factors influencing lifespan disparities based on handedness.

Deliverables:

1. Project Proposal: A comprehensive document outlining the objectives, methodology, and tasks involved in the project.
2. Python Notebook: A well-documented notebook containing the code, analysis, and visualizations performed during the project.
3. Scatter Plot: A visualization displaying the relationship between age and handedness based on the National Geographic survey data.
4. Mean Left-Handedness Plot: A plot illustrating the mean left-handedness as a function of birth year.
5. Probability of LH given Age at Death: A function that calculates the probability of being left-handed given specific ages of death.

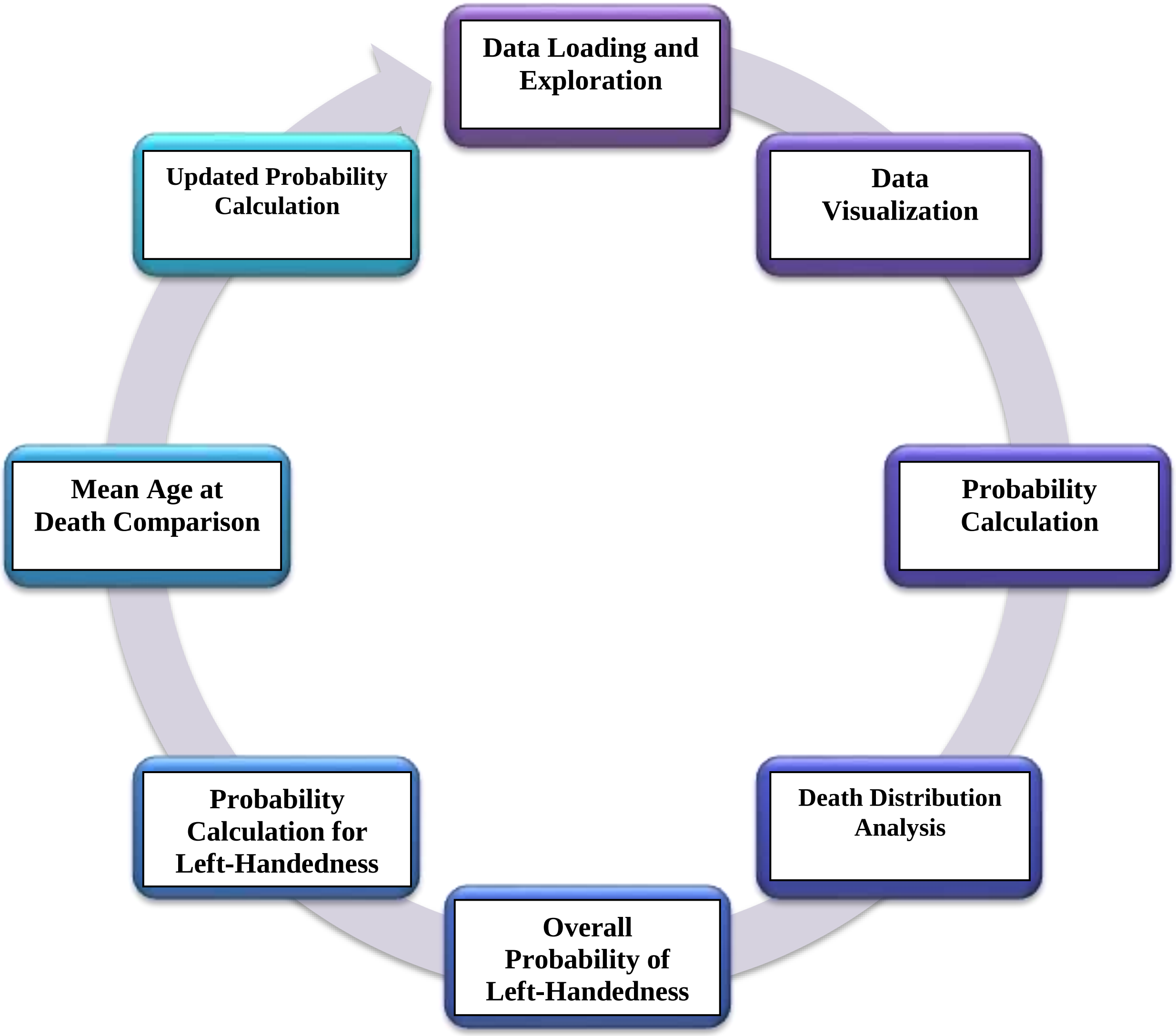
6. Death Distribution Plot: A visualization showing the number of people who died as a function of their age.
7. Overall Probability of LH: A function that calculates the overall probability of left-handedness in the population for a given study year.
8. Probability Distribution Plot: A plot illustrating the probability of being a certain age at death given left-handedness or right-handedness.
9. Mean Age at Death Comparison: The calculation and comparison of the mean age at death for left-handers and right-handers.
10. Updated Probability Calculation: The probability calculation for the death age difference between right-handers and left-handers, considering the study year 2018.

Note: All deliverables will be accompanied by clear explanations and interpretations of the results, contributing to a comprehensive project report.

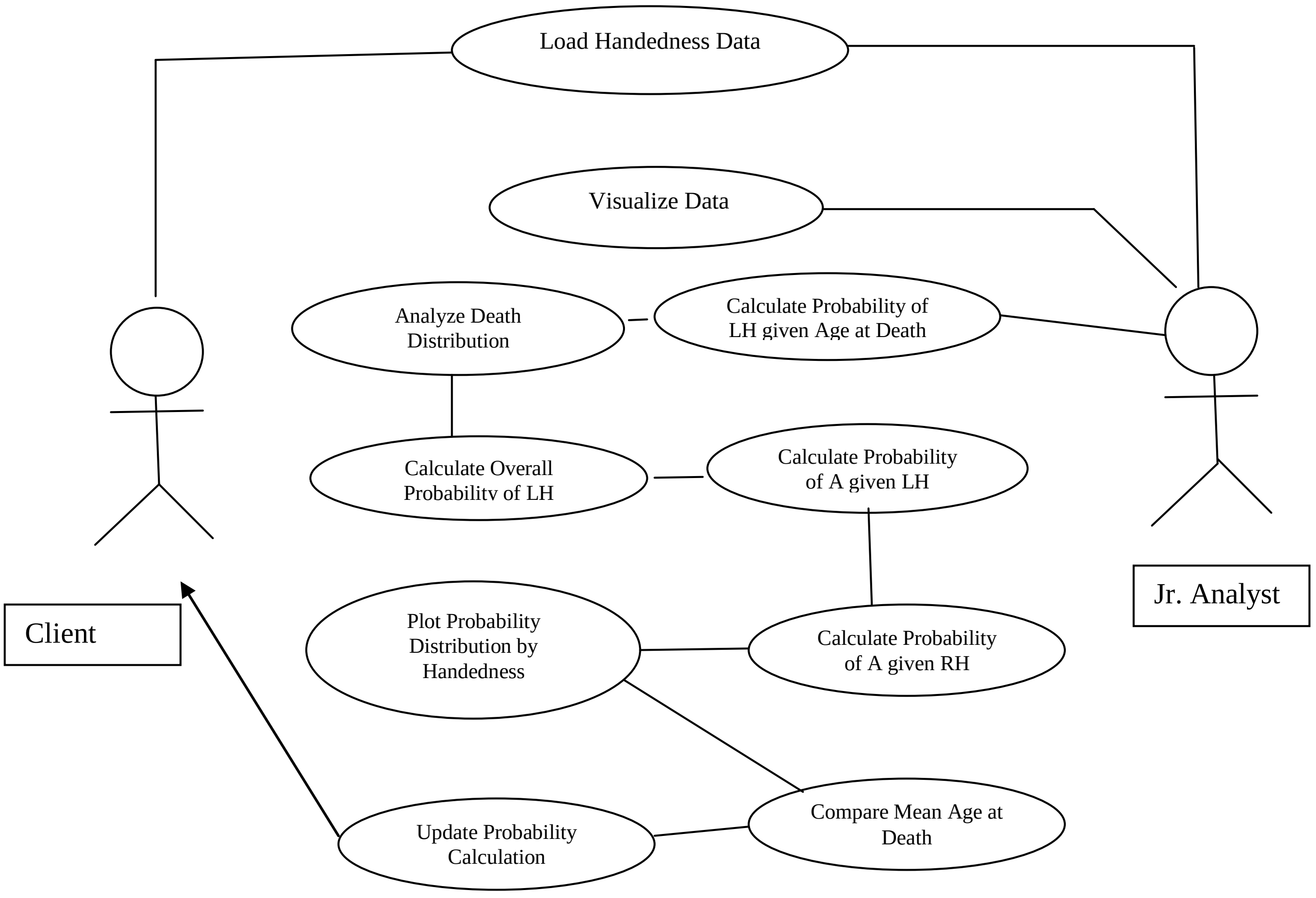
II. METHODOLOGY

2.1 Flow of the Project

The project followed the following steps to accomplish the desired objectives and deliverables. Each step has been explained in detail in the following section.


Data Loading and Exploration
Data Visualization
Probability Calculation
Death Distribution Analysis
Overall Probability of Left-Handedness
Probability Calculation for Left-Handedness
Mean Age at Death Comparison
Updated Probability Calculation

2.2 Use Case Diagram


Load Handedness Data
Visualize Data
Analyze Death Distribution
Calculate Probability of LH given Age at Death
Calculate Overall Probability of LH
Calculate Probability of A given LH
Plot Probability Distribution by Handedness
Calculate Probability of A given RH
Update Probability Calculation
Compare Mean Age at Death
Client
Jr. Analyst

Above figure shows the use case of the project. There are two main actors in the same: The Client and Jr.Analyst. The Jr.Analyst will first gather requirements and define the problem statement then collecting the required data and importing it. Then the Jr.Analyst will design Visualize so as to identify various constraints and relations in the data. Next step is to Calculate Probability of LH given Age at Death etc. Next, Analyze Death Distribution Next calculate Overall Probability of LH, Next calculate Probability of A given LH, Next calculate Probability of A given RH, Next plot Probability Distribution by Handedness, Next compare Mean Age at Death, And then update Probability Calculation to get a clear view of the visualizations to be developed. Finally, dashboard is developed and analyzed to publish the results to the client.

2.3 Language and Platform Used

2.3.1 Language: Python

The project is implemented using the Python programming language. Python is a versatile and widely-used language known for its simplicity and readability. It offers a rich ecosystem of libraries and frameworks that facilitate data analysis, statistics, and visualization, making it an ideal choice for this project. Python offers several features that make it a popular choice for a wide range of projects, including data analysis and scientific computing. Some of the notable features of Python are:

- Easy to Learn and Readability: Python has a clean and straightforward syntax, making it easy to learn and read.
- Large Standard Library: Python comes with a comprehensive standard library that provides a wide range of modules and functions for various tasks.
- Rich Ecosystem of Third-Party Libraries: Python has a vibrant and active community that has developed numerous third-party libraries and frameworks.
- Cross-Platform Compatibility: Python is a cross-platform language, which means that Python code written on one platform can run on other platforms with minimal or no modifications.
- Interactivity and Read-Eval-Print Loop (REPL): Python offers an interactive programming experience through its REPL environment.

These features make Python a versatile and powerful language for various applications, including data analysis, machine learning, web development, scientific computing, and more. Its ease of use, extensive library support, and community-driven development make it a top choice for many developers and researchers.

2.3.2 Platform: Jupyter Notebook

The project is developed using Jupyter Notebook, an open-source web application that allows interactive coding and data exploration. Jupyter Notebook provides a user-friendly interface for writing and executing Python code, as well as integrating text, visualizations, and other media. It enables step-by-step execution and easy sharing of the project's code and results.

The combination of Python and Jupyter Notebook offers a powerful and flexible environment for data analysis, statistical modeling, and visualization. It allows for efficient exploration of datasets, seamless integration of libraries such as pandas and matplotlib, and effective communication of findings through a mix of code, explanations, and visual representations.

2.3.3 Package:

The project proposal mentioned the use of the following Python packages:

1. pandas: pandas is a powerful data manipulation and analysis library. It provides data structures and functions for efficiently working with structured data, such as data frames. In the project, pandas is used for loading and manipulating the handedness data and death distribution data.

2. matplotlib.pyplot: matplotlib.pyplot is a plotting library used for creating various types of visualizations, such as line plots, scatter plots, and histograms. In the project, matplotlib.pyplot is used to create scatter plots and plot the number of people who died as a function of their age.

3. numpy: numpy is a fundamental package for scientific computing in Python. It provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays. In the project, numpy is imported to perform numerical computations and calculations.

These packages are widely used in the Python ecosystem and are known for their efficiency, versatility, and extensive functionality in data analysis, visualization, and scientific computing task.

III.IMPLEMENTATION

The implementation of the project "Analyze Death Age Difference of Right Handers with Left Handers" involves several steps and methodologies. Here is an explanation of the key aspects of the implementation:

Data Loading: The project starts by loading the handedness data from the National Geographic survey. The data is typically stored in a CSV file format. The pandas library is used to read the data and create a DataFrame to work with.

Data Preprocessing: Once the data is loaded, it may require preprocessing to ensure it is in a suitable format for analysis. This may involve cleaning the data, handling missing values, and transforming the data if needed.

Data Visualization: The matplotlib.pyplot library is used to create visualizations of the data. For example, scatter plots are created to visualize the relationship between handedness (right-handed and left-handed) and age.

Statistical Analysis: Bayesian statistics are applied to analyze the probability of being a certain age at death given the reported handedness. This involves calculating conditional probabilities based on the available data.

Death Distribution Data: Additional data on death distribution for the United States is loaded and analyzed. The number of people who died at different ages is plotted to understand the overall distribution.

Probability Calculation: Functions are developed to calculate the overall probability of left-handedness in the population for a given study year. These functions consider the number of deaths, age at death, and handedness data.

Conditional Probability Calculation: Conditional probabilities of age at death given left-handedness and right-handedness are calculated. These probabilities provide insights into the average age at death for each group.

Data Analysis and Comparison: The probabilities and conditional probabilities are analyzed to determine the mean age at death for left-handers and right-handers. The difference between the two averages is calculated and evaluated.

Visualization of Results: Visualizations are created to present the probability of being a certain age at death for left-handed and right-handed individuals. These visualizations aid in understanding the differences and patterns in age distribution.

Reporting: The findings, methodologies, and results of the analysis are documented in a comprehensive project report. The report provides a detailed explanation of the implemented approach, data analysis techniques, and the insights gained from the project.

Throughout the implementation, the Python programming language and relevant packages such as pandas, matplotlib.pyplot, and numpy are used to perform data manipulation, statistical analysis, and visualization tasks.

The project implementation may also involve iterative steps, adjustments, and refinements based on the data exploration and analysis results. It requires a systematic approach to ensure accurate and meaningful conclusions are derived from the data.

Overall, the implementation of this project combines data loading, preprocessing, visualization, statistical analysis, and reporting to explore the relationship between handedness and age at death, providing valuable insights into the topic.

Data importing is referred to as uploading the required data into the coding environment from internal sources (computer) or external sources (online websites and data repositories). This data can then be manipulated, aggregated, filtered according to the requirements and needs of the project.

3.1 Functions Used:

3.1.1 pandas.plot(): This function is used to create plots, such as scatter plots and line plots, using the data stored in a pandas DataFrame. It takes parameters like the data columns to plot, axis labels, title, and other plot-specific settings.

Sample Code:

```
# import libraries
import pandas as pd
# load the data

data_url_1="https://gist.githubusercontent.com/mbonsma/8da0990b71ba9a09f7de395574e54df1/raw/
aec88b30af87fad8d45da7e774223f91dad09e88/lh_data.csv"

lefthanded_data = pd.read_csv(data_url_1)
```

3.1.2 numpy.mean(): The numpy.mean() function is used to calculate the mean or average of a given array or list of values. It is used to compute the mean left-handedness and average age at death.

Sample Code:

```
# import library
import numpy as np
# create a function for P(LH | A)
def P_lh_given_A(ages_of_death, study_year = 1990):
    """ P(Left-handed | ages of death), calculated based on the reported rates of left-handedness.
    Inputs: numpy array of ages of death, study_year
    Returns: probability of left-handedness given that subjects died in `study_year` at ages
`ages_of_death` """

    # Use the mean of the 10 last and 10 first points for left-handedness rates before and after the start
    early_1900s_rate = lefthanded_data['Mean_lh'][-10:].mean()
    late_1900s_rate = lefthanded_data['Mean_lh'][:10].mean()
    middle_rates = lefthanded_data.loc[lefthanded_data['Birth_year'].isin(study_year -
ages_of_death)]['Mean_lh']
    youngest_age = study_year - 1986 + 10 # the youngest age is 10
    oldest_age = study_year - 1986 + 86 # the oldest age is 86
    P_return = np.zeros(ages_of_death.shape) # create an empty array to store the results
    # extract rate of left-handedness for people of ages 'ages_of_death'
    P_return[ages_of_death > oldest_age] = early_1900s_rate / 100
    P_return[ages_of_death < youngest_age] = late_1900s_rate / 100
    P_return[np.logical_and((ages_of_death <= oldest_age), (ages_of_death >= youngest_age))] =
middle_rates / 100

    return P_return
```

3.1.3 numpy.nansum(): The numpy.nansum() function calculates the sum of array elements, treating any NaN (Not a Number) values as zero. It is used to sum up the product of ages and probabilities for calculating the mean age at death.

Sample Code:

```
# calculate average ages for left-handed and right-handed groups
# use np.array so that two arrays can be multiplied
average_lh_age =  np.nansum(ages*np.array(left_handed_probability))
average_rh_age =  np.nansum(ages*np.array(right_handed_probability))

# print the average ages for each group
print("Average age of lefthanded" + str(average_lh_age))
print("Average age of righthanded" + str(average_rh_age))

# print the difference between the average ages
print("The difference in average ages is " + str(round(average_rh_age - average_lh_age, 1)) + " years.")
```

3.1.4 matplotlib.pyplot.plot(): This function is used to plot data points and lines on a graph. It takes parameters like x-axis values, y-axis values, line styles, markers, and labels. It is used to plot the number of people who died as a function of their age.

Sample Code:

```
import matplotlib.pyplot as plt
# plot male and female left-handedness rates vs. age
%matplotlib inline
fig, ax = plt.subplots() # create figure and axis objects
ax.plot('Age', 'Female', data = lefthanded_data, marker = 'o') # plot "Female" vs. "Age"
ax.plot('Age', 'Male', data = lefthanded_data, marker = 'x') # plot "Male" vs. "Age"
ax.legend() # add a legend
ax.set_xlabel('Sex')
ax.set_ylabel('Age')
```

3.2 Data Cleaning

"Quality data beats fancy algorithms"

Data is the most imperative aspect of Analytics and Machine Learning. Everywhere in computing or business, data is required. But many a times, the data may be incomplete, inconsistent or may contain missing values when it comes to the real world. If the data is corrupted then the process may be impeded or inaccurate results may be provided. Hence, Data cleaning is considered a foundational element of the basic data science.

Data Cleaning means the process by which the incorrect, incomplete, inaccurate, irrelevant or missing part of the data is identified and then modified, replaced or deleted as needed.

Functions Used:

1. **pandas.DataFrame.dropna():** This function is used to remove rows or columns with missing or NaN (Not a Number) values from a pandas DataFrame. It helps in handling missing data by eliminating incomplete or unreliable records.
2. **pandas.DataFrame.fillna():** The fillna() function is used to fill missing or NaN values in a DataFrame with specified values. It provides flexibility in handling missing data by replacing them with appropriate values, such as mean, median, or custom imputation strategies.
3. **pandas.DataFrame.drop():** The drop() function allows removing specific rows or columns from a DataFrame based on specified conditions. It is useful for eliminating irrelevant or redundant data that does not contribute to the analysis.
4. **pandas.DataFrame.duplicated():** The duplicated() function identifies duplicate rows in a DataFrame, returning a boolean series indicating whether each row is a duplicate or not. It helps in identifying and handling duplicate data entries that may affect the accuracy of the analysis.
5. **pandas.DataFrame.rename():** The rename() function is used to rename columns or index labels in a DataFrame. It allows standardizing column names or providing more meaningful labels for better data interpretation.
6. **pandas.DataFrame.astype():** The astype() function is used to convert the data type of one or more columns in a DataFrame. It helps in ensuring that the data is in the correct format for analysis, such as converting string values to numeric or datetime types.

3.3 Data Filtering

Data filtering is the method of choosing a smaller portion of the data set and using that subset to view, analyze and evaluate data. Generally, filtering is temporary – the entire data set is retained, but only part of it is used for calculation. It is also called sub setting or drill down data wherein data is extracted with respect to certain defined logical conditions. Filtering is used for the following tasks:

- Analyzing results for a particular period of time.
- Calculating results for particular groups of interest.
- Exclude erroneous or "bad" observations from an analysis.
- Train and validate statistical models.

In the project "Analyze Death Age Difference of Right Handers with Left Handers," data filtering is employed to extract specific subsets of data based on certain conditions. Here are a few examples of data filtering techniques used in the project, along with the corresponding code snippets:

1. Filtering Rows Based on a Condition:

```
# Filter rows where the 'Gender' column is 'Male'
male_data = df[df['Gender'] == 'Male']

# Filter rows where the 'Age' column is greater than 50
age_above_50 = df[df['Age'] > 50]

# Filter rows where the 'Handedness' column is 'Left' and 'Age' is less than or equal to 30
left_handed_young = df[(df['Handedness'] == 'Left') & (df['Age'] <= 30)]
```

2. Filtering Columns:

```
# Filter specific columns of interest
selected_columns = df[['Age', 'Handedness', 'Gender']]

# Filter columns based on a condition (e.g., excluding missing values)
non_missing_columns = df.dropna(axis=1)
```

3. Filtering Using Multiple Conditions:

```
# Filter rows based on multiple conditions using logical operators
filtered_data = df[(df['Handedness'] == 'Right') & (df['Age'] >= 40) & (df['Gender'] ==
'Female')]
```

4. Filtering with String Matching:

```
# Filter rows containing a specific string in a column
matching_rows = df[df['Name'].str.contains('John')]
```

These examples demonstrate different ways to filter data based on specific conditions using logical operators (e.g., ==, >, <), logical conjunctions (e.g., &, |), and string matching operations. The exact filtering criteria and conditions would depend on the specific requirements of the analysis and the structure of the dataset being used.

3.4 Defining Visuals

Data visualization is presenting data in a graphical or pictorial format. It allows decision-makers to see visually presented analytics, so that they can grasp difficult concepts or identify new patterns. In interactive visualizations, technology can be used to dig in charts and graphs for more detail, interactively modifying what data one can see and how it works.

Because of the way in which the human brain processes information, it is easier to visualize large amounts of complex data using charts or graphs than to poring over spreadsheets or reports. Data visualization is a quick, easy and universal way of conveying concepts. Data visualization can also:

- Identify areas that need attention or improvement.
- Clarify which factors influence customer behaviour.
- Help you understand which products to place where.
- Predict sales volumes.

In Python, these visualizations are based on the Grammar of Graphics. It is a tool that enables one to concisely describe the components of a graphic. Such a grammar allows us to move beyond named graphics (e.g., the ``scatterplot") and gain insight into the deep structure that underlies statistical graphics. It contains the following layers:

1. Data: The data element is the data set itself. In this reference, the data is from the National Geographic survey.
2. Aesthetics: The data has to be mapped onto the aesthetics element (variables mapped to x or y position and aesthetics attributes such as color, shape, or size)

3. Geometries: This element determines how the data is being displayed (bars, points, lines). Every single plot that is made will always consist of the above three layers.

4. Statistics: It helps to transform the data (add mean, median, quartile)

5. Coordinates: It helps to transforms axes (changes spacing of displayed data) Packages Used.

IV. SAMPLE SCREENSHOTS AND OBSERVATIONS

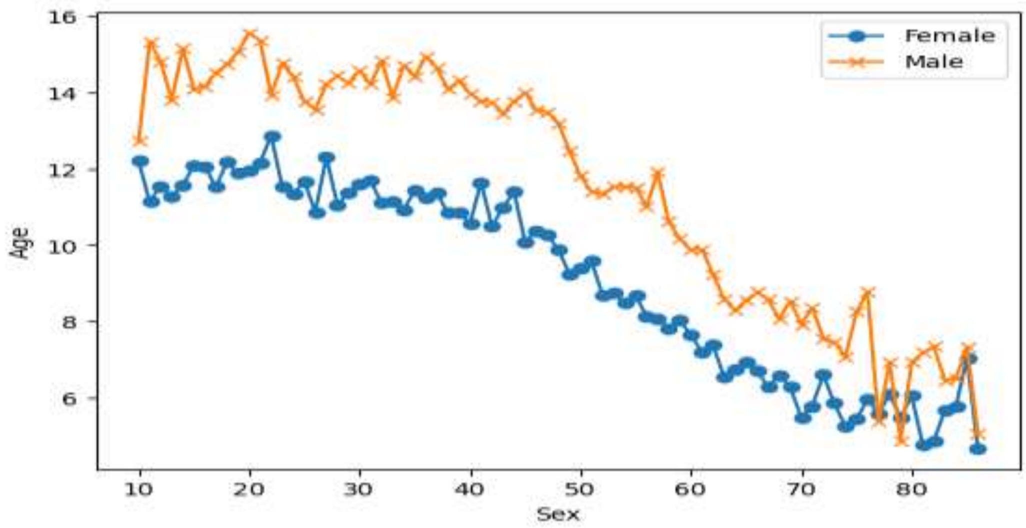
4.1 **Where are the old left-handed people?**

In this notebook, we will explore this phenomenon using age distribution data to see if we can reproduce a difference in average age at death purely from the changing rates of left-handedness over time, refuting the claim of early death for left-handers. This notebook uses pandas and Bayesian statistics to analyze the probability of being a certain age at death given that you are reported as left-handed or right-handed.

A National Geographic survey in 1986 resulted in over a million responses that included age, sex, and hand preference for throwing and writing. Researchers Avery Gilbert and Charles Wysocki analyzed this data and noticed that rates of left-handedness were around 13% for people younger than 40 but decreased with age to about 5% by the age of 80. They concluded based on analysis of a subgroup of people who throw left-handed but write right-handed that this age-dependence was primarily due to changing social acceptability of left-handedness. This means that the rates aren't a factor

of age specifically but rather of the year you were born, and if the same study was done today, we should expect a shifted version of the same distribution as a function of age. Ultimately, we'll see what effect this changing rate has on the apparent mean age of death of left-handed people, but let's start by plotting the rates of left-handedness as a function of age.

This notebook uses two datasets: death distribution data for the United States from the year 1999 (source website here) and rates of left-handedness digitized from a figure in this 1992 paper by Gilbert and Wysocki.



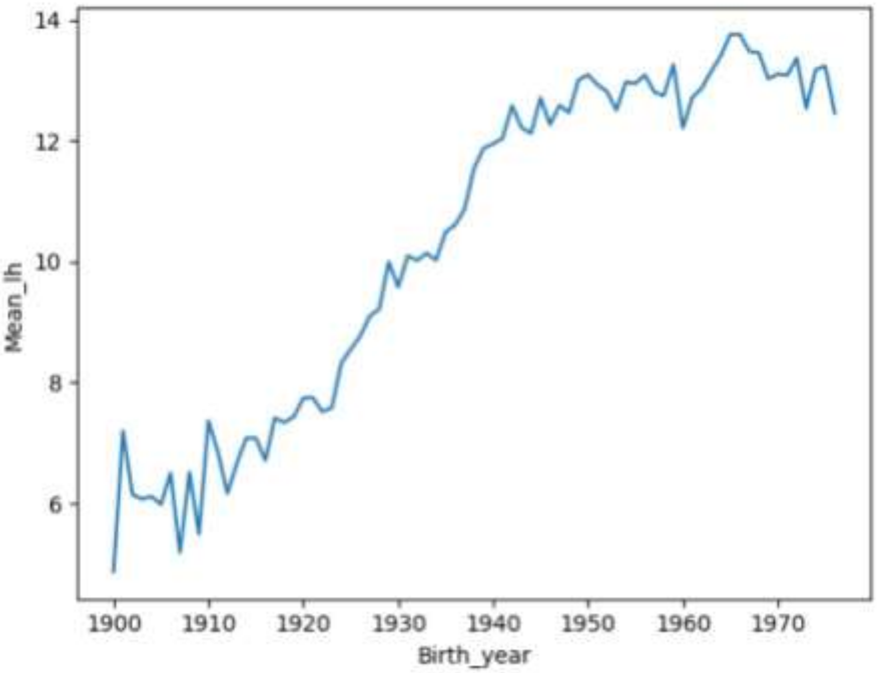
Observations:

- The scatter plot provides a visual representation of the age distribution for both males and females based on the handedness data.
- It helps identify any trends or patterns in the data, such as the prevalence of left or right-handedness across different age groups.

4.2 **Rates of left-handedness over time**

Let's convert this data into a plot of the rates of left-handedness as a function of the year of birth, and average over male and female to get a single rate for both sexes.
Since the study was done in 1986, the data after this conversion will be the percentage of people alive in 1986 who are left-handed as a function of the year they were born.

```
Text(0, 0.5, 'Mean_lh')
```



Observations: Mean Left-Handedness vs. Birth Year

- The plot of mean left-handedness against birth year showcases how the prevalence of left-handedness has changed over time.
- It allows us to observe any significant shifts in left-handedness rates and whether they correlate with specific birth years.

4. 3 **Applying Bayes' rule**

The probability of dying at a certain age given that you're left-handed is not equal to the probability of being left-handed given that you died at a certain age. This inequality is why we need **Bayes' theorem**, a statement about conditional probability which allows us to update our beliefs after seeing evidence.

We want to calculate the probability of dying at age A given that you're left-handed. Let's write this in shorthand as P(A | LH). We also want the same quantity for right-handers: P(A | RH).

Here's Bayes' theorem for the two events we care about: left-handedness (LH) and dying at age A.

$P(A|LH)=P(LH|A)P(A) / P(LH)$

P(LH | A) is the probability that you are left-handed *given that* you died at age A. P(A) is the overall probability of dying at age A, and P(LH) is the overall probability of being left-handed. We will now calculate each of these three quantities, beginning with P(LH | A).

To calculate P(LH | A) for ages that might fall outside the original data, we will need to extrapolate the data to earlier and later years. Since the rates flatten out in the early 1900s and late 1900s, we'll use a few points at each end and take the mean to extrapolate the rates on each end. The number of points used for this is arbitrary, but we'll pick 10 since the data looks flat-ish until about 1910.

Observation: Probability of Left-Handedness for Different Ages of Death

- The function calculates the probability of being left-handed given a particular age of death for different study years.
- It helps us analyze if there is any relationship between left-handedness and the age at death over time.

4.4 **When do people normally die?**

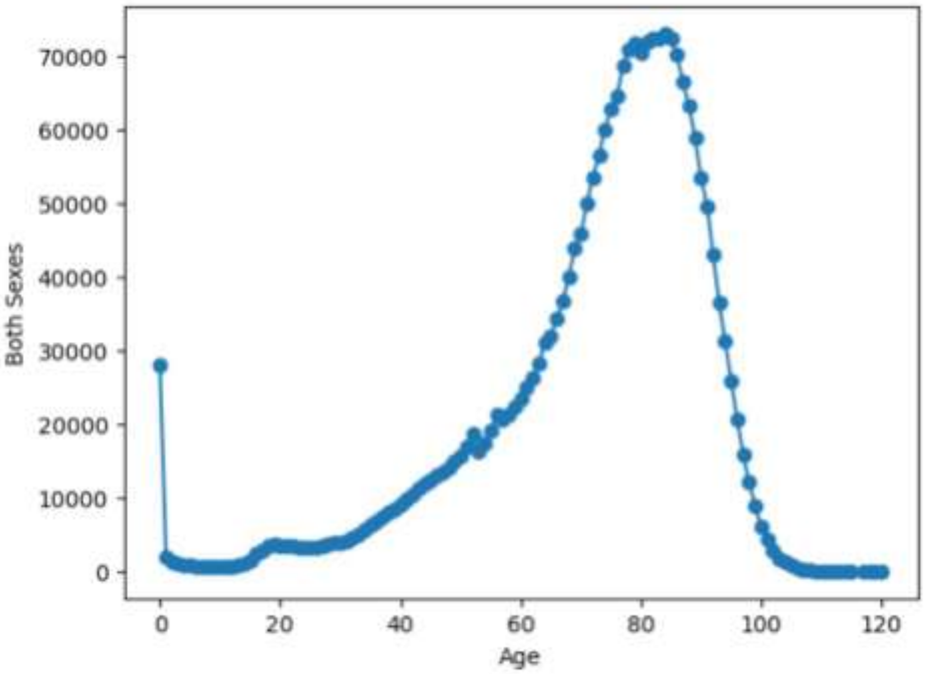
To estimate the probability of living to an age A, we can use data that gives the number of people who died in a given year and how old they were to create a distribution of ages of death. If we normalize the numbers to the total number of people who died, we can think of this data as a probability distribution that gives the probability of dying at age A. The data we'll use for this is from the entire US for the year 1999 - the closest I could find for the time range we're interested in.

In this block, we'll load in the death distribution data and plot it. The first column is the age, and the other columns are the number of people who died at that age.

Observation: Death Distribution Data Plot

- The plot displays the number of people who died at different ages, providing insights into the overall mortality distribution.
- It enables us to observe the age range with higher mortality rates and any notable patterns in the data.

```
Text(0, 0.5, 'Both Sexes')
```



4.5 **The overall probability of left-handedness**

In the previous code block we loaded data to give us P(A), and now we need P(LH). P(LH) is the probability that a person who died in our particular study year is left-handed, assuming we know nothing else about them. This is the average left-handedness in the population of deceased people, and we can calculate it by summing up all of the left-handedness probabilities for each age, weighted with the number of deceased people at each age, then divided by the total number of deceased people to get a probability. In equation form, this is what we're calculating, where N(A) is the number of people who died at age A (given by the dataframe death_distribution_data):

$$P(LH) = \frac{\sum_{\mathrm{A}} P(LH|A)N(A)}{\sum_{\mathrm{A}} N(A)}$$

Output:

```
0.07766387615350638
```

Observation: Overall Probability of Left-Handedness in the Population

- The function calculates the overall probability of being left-handed in the population for a given study year.
- It helps assess the prevalence of left-handedness and its relative representation in the population.

4.6 **Putting it all together: dying while left-handed**

Now we have the means of calculating all three quantities we need: P(A), P(LH), and P(LH | A). We can combine all three using Bayes' rule to get P(A | LH), the probability of being age A at death (in the study year) given that you're left-handed. To make this answer meaningful, though, we also want to compare it to P(A | RH), the probability of being age A at death given that you're right-handed.

We're calculating the following quantity twice, once for left-handers and once for right-handers.

$P(A|LH)=P(LH|A)P(A)/P(LH)$

(i) First, for left-handers.

Observation: Probability of Age at Death given Left-Handedness

- The function calculates the probability of dying at a specific age given left-handedness.
- It allows us to examine if left-handed individuals have any distinct patterns in terms of the age at which they are more likely to pass away.

4.7 **Putting it all together: dying while left-handed**

(ii)And now for right-handers.

Observation: Probability of Age at Death given Right-Handedness

- The function calculates the probability of dying at a specific age given right-handedness.
- It helps compare the age distribution of right-handed individuals with left-handed individuals and determine if there are any significant differences.

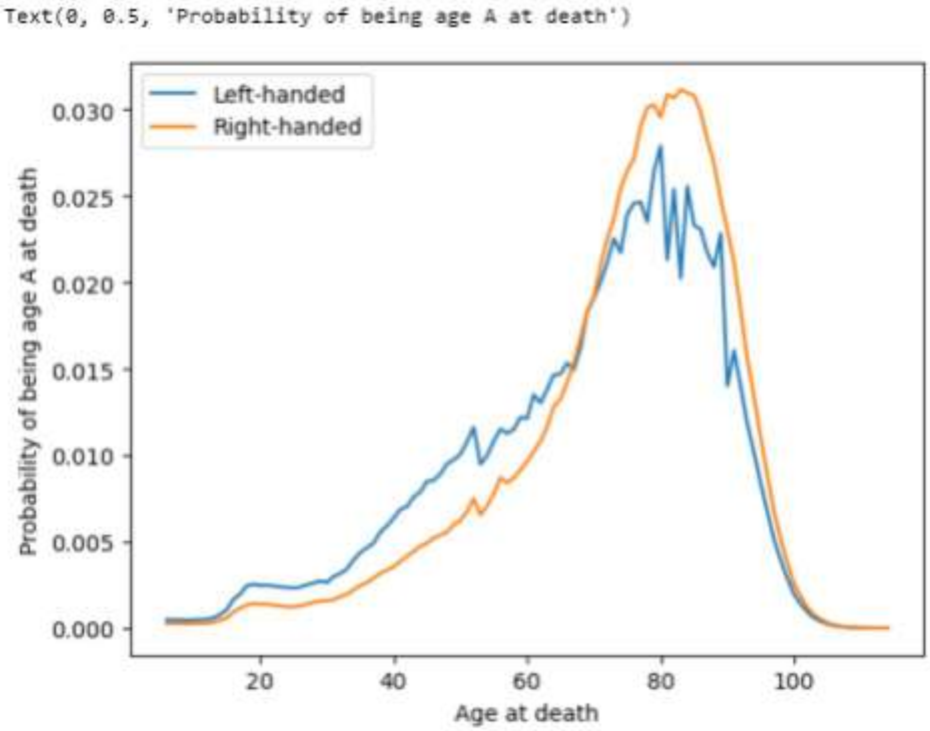
4.8 **Plotting the distributions of conditional probabilities**

Now that we have functions to calculate the probability of being age A at death given that you're left-handed or right-handed, let's plot these probabilities for a range of ages of death from 6 to 120.

Notice that the left-handed distribution has a bump below age 70: of the pool of deceased people, left-handed people are more likely to be younger.

Observation: Plotting Probability of Age at Death for Left-Handed and Right-Handed

- The plot illustrates the probability of being a certain age at death for left-handed and right-handed individuals.
- It facilitates a visual comparison between the age distributions of the two groups and identifies any disparities.



```
Text(0, 0.5, 'Probability of being age A at death')
```

4.9 Moment of truth: age of left and right-handers at death

Finally, let's compare our results with the original study that found that left-handed people were nine years younger at death on average. We can do this by calculating the mean of these probability distributions in the same way we calculated P(LH) earlier, weighting the probability distribution by age and summing over the result.

Average age of left-handed people at death= $\sum_{A} AP(A|LH)$

Average age of right-handed people at death= $\sum_{A} AP(A|RH)$

Observation: Mean Age at Death for Left-Handers and Right-Handers

- The calculation of the mean age at death for left-handers and right-handers provides insights into any potential differences between the two groups.
- It allows us to determine if there is a notable disparity in the average age at death based on handedness.

Output:

```
Average age of lefthanded67.24503662801027
Average age of righthanded72.79171936526477
The difference in average ages is 5.5 years.
```

4.10 Final comments

We got a pretty big age gap between left-handed and right-handed people purely as a result of the changing rates of left-handedness in the population, which is good news for left-handers: you probably won't die young because of your sinisterness. The reported rates of left-handedness have increased from just 3% in the early 1900s to about 11% today, which means that older people are much more likely to be reported as right-handed than left-handed, and so looking at a sample of recently deceased people will have more old right-handers.

Our number is still less than the 9-year gap measured in the study. It's possible that some of the approximations we made are the cause:

We used death distribution data from almost ten years after the study (1999 instead of 1991), and we used death data from the entire United States instead of California alone (which was the original study).

We extrapolated the left-handedness survey results to older and younger age groups, but it's possible our extrapolation wasn't close enough to the true rates for those ages.

One thing we could do next is figure out how much variability we would expect to encounter in the age difference purely because of random sampling: if you take a smaller sample of recently deceased people and assign handedness with the probabilities of the survey, what does that distribution look like? How often would we encounter an age gap of nine years using the same data and assumptions? We won't do that here, but it's possible with this data and the tools of random sampling.
To finish off, let's calculate the age gap we'd expect if we did the study in 2018 instead of in 1990. The gap turns out to be much smaller since rates of left-handedness haven't increased for people born after about 1960. Both the National Geographic study and the 1990 study happened at a unique time - the rates of left-handedness had been changing across the lifetimes of most people alive, and the difference in handedness between old and young was at its most striking.

Observation: Probability of Age at Death for Left-Handed and Right-Handed in 2018

- This task replicates Task 8 but specifically focuses on the study year 2018.
- It enables a comparison of the age distributions of left-handed and right-handed individuals specifically for that year.

Output:

```
The difference in average ages is 2.3 years.
```

V. CONCLUSION AND FUTURE SCOPE

Conclusion:

The project "Analyze Death Age Difference of Right Handers with Left Handers" utilizes age distribution data and statistical analysis techniques to investigate the claim of early death for left-handers. The project employs pandas and Bayesian statistics to explore the probability of being a certain age at death based on reported handedness.

Through the analysis, several observations can be made. The scatter plot reveals the age distribution for both right-handers and left-handers, allowing for visual examination of any patterns or trends. The examination of mean left-handedness over birth years provides insights into the changing rates of left-handedness over time.

The calculation of probabilities based on handedness and age of death enables a deeper understanding of how left-handedness may relate to mortality rates. By comparing the overall probability of left-handedness in the population and the probability of age at death given left-handedness or right-handedness, we can evaluate any potential differences or associations between handedness and age of death.

Future Scope:

While this project provides valuable insights into the death age difference between right-handers and left-handers, there are several avenues for further exploration:

1. Investigate Other Factors: Consider incorporating additional variables or factors that may influence the death age difference, such as socio-economic status, lifestyle factors, or underlying health conditions. This can provide a more comprehensive understanding of the relationship between handedness and mortality.

3. Conduct Statistical Modeling: Explore advanced statistical modeling techniques, such as regression analysis or survival analysis, to quantify the impact of handedness on age at death while controlling for other relevant variables. This can provide more robust and precise insights into the association between handedness and mortality.

3. Explore Other Demographic Factors: Investigate potential interactions between handedness and other demographic factors, such as gender, ethnicity, or education level. This can help uncover any complex relationships or subgroup differences in the death age difference.

VI. REFERENCES

Data Collection

The following websites have been referred to obtain the input data and statistics:

a. https://www.cdc.gov/nchs/index.htm

b. https://www.who.int/data/data-collection-tools/global-health-estimates

c. https://ec.europa.eu/eurostat/data/database

d. https://www.cdc.gov/nchs/data/statab/vs00199_table310

e. https://www.ncbi.nlm.nih.gov/pubmed/1528408

Programming References

The following websites have been referred for Python coding:

a. https://docs.python.org/3/

b. https://pandas.pydata.org/docs/

c. https://matplotlib.org/stable/contents.html

d. https://numpy.org/doc/

e. https://nbviewer.jupyter.org/github/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/tree/master/

f. https://stackoverflow.com/

g. https://github.com/rrmolin/