

# **Final Report of Traineeship Program 2023**

*On*

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

**MEDTOUREASY**



28<sup>th</sup> June 2023

~Abhishek Goyal

## **ACKNOWLEDGMENTS**

The traineeship opportunity that I had with MedTourEasy was a great chance for learning and understanding the intricacies of the subject of Data Visualizations in Data Analytics; and, for personal as well as professional development. I am very obliged to have a chance to interact with so many professionals who guided me throughout the traineeship project and made it a great learning curve for me.

Firstly, I express my deepest gratitude and special thanks to the Training & Development Team of MedTourEasy who gave me an opportunity to carry out my traineeship at their esteemed organization. Also, I express my thanks to the team for making me understand the details of the Data Analytics profile and training me in the same so that I can carry out the project properly and with maximum client satisfaction and for sparing his valuable time despite his busy schedule.

I would also like to thank the team of MedTourEasy and my colleagues who made the working environment productive and very conducive.

# TABLE OF CONTENTS

Acknowledgments .....i

Abstract ..... iii

<b>Sr. No.</b>	<b>Topic</b>	<b>Page No.</b>
<b>1</b>	Introduction	
	1.1 About the Company	5
	1.2 About the Project	5
	1.3 Objectives and Deliverables	8
<b>2</b>	Methodology	
	2.1 Flow of the Project	10
	2.3 Language and Platform Used	11
<b>3</b>	Implementation	
	3.1 Gathering Requirements and Defining Problem Statement	14
	3.2 Data Collection and Importing	15
	3.3 Data Exploration and Analysis	17
<b>4</b>	Screenshots and code of Steps and their Observations	21
<b>6</b>	Conclusion	42
<b>8</b>	References	44

## ABSTRACT

A survey of 1,177,507 U.S. men and women between the ages of 10 and 86 included questions regarding hand preference for writing and throwing. Three effects were observed. Individuals with at least some left motoric bias comprised a smaller percent of the population with advancing age.

This finding provides large-scale confirmation of a previously described phenomenon. Among sinistral, concordance for writing and throwing was 2.2 times as prevalent as left-writing with right-throwing, and 4.1 times as prevalent as right-writing with left-throwing. These sinistral subpopulations displayed distinct and stable prevalence prior to age 50 and changing patterns of prevalence after age 50.

The results confirm a decrease with age in the prevalence of sinistrality but indicate that age-specific rates of mixed- and left-handedness are distinct. The implications for hypotheses regarding age-related change in the prevalence of sinistrality are discussed.

There are many misconceptions regarding Left Handers such as their early deaths that's why A National Geographic Survey took place in 1986 which 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.

There are still many misconceptions that are running around so it is necessary to analyze this whole scenario to get a clear a broadened picture.

Therefore, this project aims at collecting and analyzing wide variety of large data sets, create intuitive and interactive visualizations for representing Death Age data to gain meaningful insights.

# I. INTRODUCTION

## 1.1 About the Company

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

## 1.2 About the Project

A survey of 1,177,507 U.S. men and women between the ages of 10 and 86 included questions regarding hand preference for writing and throwing. Three effects were observed. Individuals with at least some left motoric bias comprised a smaller percent of the population with advancing age. This finding provides large-scale confirmation of a previously described phenomenon. Among sinistral, concordance for writing and throwing was 2.2 times as prevalent as left-writing with right-throwing, and 4.1 times as prevalent as right-writing with left-throwing. These sinistral subpopulations displayed distinct and stable prevalence prior to age 50 and changing patterns of prevalence after age 50. The results confirm a decrease with age in the prevalence of sinistrality but indicate that age-specific rates of mixed- and left-handedness are distinct. The implications for hypotheses regarding age-related change in the prevalence of sinistrality are discussed.

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.

There are many misconceptions regarding the death of left Handers, so it is extremely crucial to analyze this situation and get a meaningful insight.

Hence, this project aims at collecting and analyzing large data sets to create intuitive and interactive visualizations for representing Death Age and Hand preference to gain meaningful insights.

This project utilizes a large dataset encompassing demographic information and hand preference data of a diverse population sample. The data collection process involves survey responses, medical records, and vital statistics.

The dataset is carefully curated to ensure accuracy and representativeness, encompassing a wide range of ages, geographical regions, and socioeconomic backgrounds.

The project "Analyze the Death Age Difference between Right Handers and Left Handers" focuses on investigating whether handedness has any impact on life expectancy and mortality rates. Handedness, the dominant hand preference exhibited by individuals, has long intrigued researchers due to its potential associations with various aspects of human health and well-being

To conduct the analysis, the project follows a systematic methodology. It involves data preprocessing to handle missing values, outliers, and inconsistencies. Descriptive analysis techniques are employed to explore the demographic characteristics and distribution of handedness within the population. Hypothesis testing is conducted to compare the death ages between right handers and left handers, aiming to identify any significant differences in mortality rates.

The results of the analysis provide insights into the death age difference between right handers and left handers. The project examines the magnitude and direction of any observed disparities in mortality rates and explores potential reasons for these variations.

However, it is important to acknowledge the limitations of the study. The dataset used in the analysis may have inherent biases or limitations, such as sample size or representativeness. These limitations are discussed in the project report, along with recommendations for future research to address these shortcomings.

In conclusion, the project "Analyze the Death Age Difference between Right Handers and Left Handers" contributes to the scientific understanding of handedness and its potential impact on life expectancy and mortality rates. By employing a rigorous analysis methodology on a comprehensive dataset, the project provides valuable insights into the complex relationship between

handedness and health outcomes. The findings from this project have the potential to inform healthcare practices, public health initiatives, and personalized approaches to well-being, ultimately contributing to healthier aging and improved health outcomes for diverse populations.

Our Complete Project is divided into three major sections as listed below: -

- *Analysis of the problem:* This section contains statistics and data representing the problem statement that is divided into numerous tasks. In this section various plots and visualization are used to analyze the problem and the wheel for those visualization contains age, sex and hand preferences. This section also compares the above by such visualization.
- *Analysis of the steps taken in the survey:* This is done to analyze the various steps that are being taken to do the survey. This step contains a thorough analysis of the different steps that are taken in the survey including data cleaning, filtering and many other techniques.
- A correlation between the above two sections to assess the impact of strategies taken in the survey to Analyse the death age difference between right handers and left handers.

Each of the above sub-section has been represented in the Jupyter Notebook and with the help of various python libraries, the code of the project being controlled. We are using python as our leading code language in our project and matplotlib for data visualization and NumPy for mathematical operation pandas for many data related processes.

### 1.3 Objectives and Deliverables

The objective of this project report is to conduct a comprehensive analysis of the death age difference between right handers and left handers. The project aims to investigate whether handedness has any impact on life expectancy, exploring potential associations between handedness and mortality rates. One of the main objectives is to refute the claim of early death of left handers and by doing so removing the misconceptions that left handers die early than right handers. In this project we are using python as our programming language and jupyter notebook as our coding and visualization platform. This notebook uses pandas and Bayesian statistics to analyze the various steps of the analysis.

The project consists of 2 deliverables detailed as follows:

- a. Analysis of the problem: This part focuses on analyzing the data regarding the problem of Death Age difference between left handers and right handers. It highlights the following points and displays them through various types of visualizations:
  - Gender wise comparison of Total cases– Comparison of male and female left-handedness rates vs. Age.
  - Rates of Left Handedness over time: This will depict how the rates had increased over the years. In this the rates of left handedness are averaged over male and female rates.
  - Comparison of different rates over the birth year.
  - Analyzing the death ages for people.
  - Analysis of the age of death by considering the condition of being left-handed or right-handed.
  - Analysis of the hand preference on the age of death. [OBJ]
- b. Analysis of steps taken in the survey: This part focuses on analyzing various steps taken in the survey to do the analysis of death age difference between left handers and right handers. It highlights the following points:
  - Gender wise comparison of male and female left handedness rates vs. age
  - Analysis of the average rate of left handedness with respect to birth year
  - Finding the different probability and applying Bayes' rule.
  - Analyzing the Normal age of death for both genders with the help of



death data distribution that is provided.

- Comparison of the left-handed and right-handed people on the basis of the probability that is calculated via Bayes' rule.
  - Comparison of average age of death of for left handers and right handers.
- c. Conclusions: This part focuses on the results and conclusions that are obtained from the total analysis. Some of its points are:
- The difference in average age of death between left and right handers are not increasing while the rates of left handedness are increasing over time.
  - The death of people does not depend on their preference of hand or handedness.
  - The increasing rates of left handedness are not dependent over the age but rather on the year you were born
  - In this project the difference in death ages for left and right handers also calculated for another year 2018, in which the gap turns out to be much smaller since rates of left-handedness haven't increased for people born after about 1960.

## 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.

1. Load the handedness data from the National Geographic survey and create a scatter plot
2. Add two new columns, one for birth year and one for mean left-handedness, then plot the mean as a function of birth year
3. Create a function that will return  $P(\text{LH} \mid A)$  for ages of death in a given study year
4. Load death distribution data for the United States and plot it
5. Create a function called  $P_{\text{lh}}()$  which calculates the overall probability of left-handedness in the population for a given study year
6. Write a function to calculate  $P_{A\_given\_lh}()$
7. Write a function to calculate  $P_{A\_given\_rh}()$
8. Plot the probability of being a certain age at death given that you're left- or right-handed for a range
9. Draw the conclusion from the whole analysis process

## 2.2 Language and Platform Used

### ★ **Language: Python**

The primary language used in the project proposal is Python. Python is a versatile and popular programming language known for its simplicity, readability, and extensive range of libraries and frameworks. It has gained widespread adoption in the field of data science and analysis due to its rich ecosystem of data processing and statistical libraries.

Python offers several advantages for this project:

- **Ease of Use:** Python's clean syntax and straightforward learning curve make it accessible to both novice and experienced programmers. Its readability promotes code comprehension and collaboration.
- **Abundance of Libraries:** Python has a vast collection of libraries specifically designed for data analysis and manipulation. In this proposal, the pandas library is used for handling data frames and performing operations on structured data. The NumPy library is used for numerical computations, and matplotlib is used for data visualization.
- **Scalability:** Python's scalability makes it suitable for projects of varying sizes. It can handle small-scale exploratory data analysis as well as large-scale data processing and modeling.
- **Open-source Community:** Python benefits from an active and supportive open-source community. This community continuously develops and maintains libraries, ensuring a robust ecosystem for data analysis and scientific computing.

### ★ **Platform: Jupyter Notebook**

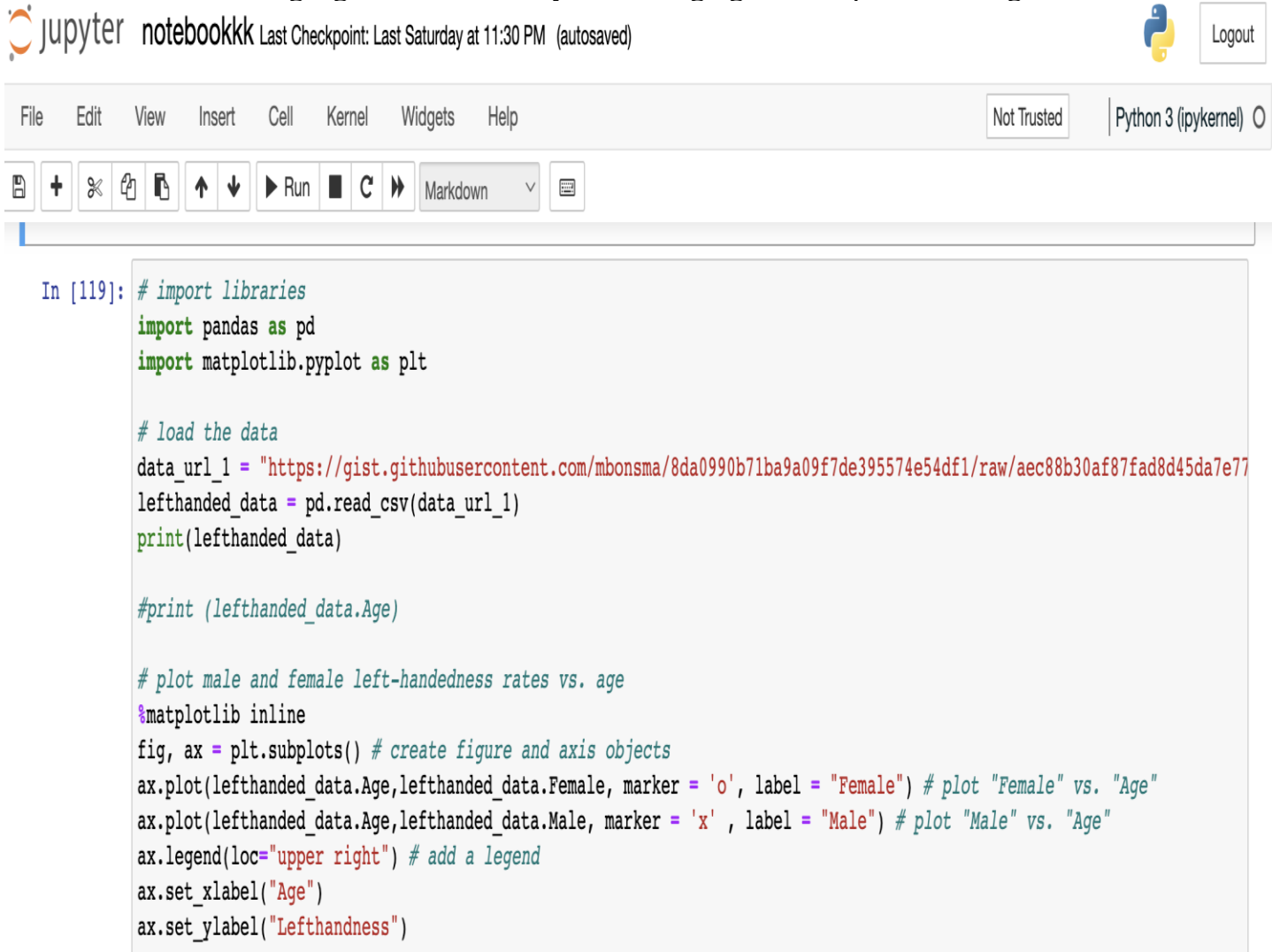
Jupyter Notebook is an open-source web application that allows you to create and share documents containing live code, visualizations, explanatory text, and more. It provides an interactive computing environment that supports multiple programming languages, including Python, R, and Julia. Jupyter Notebook is widely used by data scientists, researchers, and developers for data analysis, prototyping, and collaborative work.

Here are the key features and benefits of using Jupyter Notebook:

- **Interactive Computing:** Jupyter Notebook enables an interactive computing environment, allowing users to write and execute code snippets in an iterative manner.
- **Integration of Code and Documentation:** Jupyter Notebook seamlessly combines

code cells with narrative text, equations, and visualizations. This integration promotes reproducibility, as the code and its outputs are presented alongside detailed explanations, making it easier to understand and share the analysis.

- **Rich Media Support:** Jupyter Notebook supports a wide range of media types, including images, videos, and interactive visualizations. This capability allows users to create compelling reports and presentations that effectively communicate their findings.
- **Collaboration and Sharing:** Jupyter Notebook files can be shared easily with others, facilitating collaboration and knowledge exchange. The notebook format (.ipynb) can be shared as a standalone document or hosted on platforms like GitHub, enabling version control and collaboration with team members.
- **Support for Multiple Languages:** While Python is the primary language used in the proposal, Jupyter Notebook supports several other programming languages, including R, Julia, and Scala. This versatility allows users to incorporate different languages into their analysis, leveraging their respective strengths.



The screenshot displays the Jupyter Notebook web interface. At the top, the Jupyter logo is followed by the text 'notebook' and a status bar indicating 'Last Checkpoint: Last Saturday at 11:30 PM (autosaved)'. On the right, there is a 'Logout' button and a Python logo. Below the header is a menu bar with options: File, Edit, View, Insert, Cell, Kernel, Widgets, and Help. To the right of the menu bar are two buttons: 'Not Trusted' and 'Python 3 (ipykernel)'. Below the menu bar is a toolbar with icons for saving, adding a new cell, deleting a cell, duplicating a cell, moving a cell up/down, running a cell, and a dropdown menu currently set to 'Markdown'. The main area of the notebook contains a code cell with the following Python code:

```
In [119]: # import libraries
import pandas as pd
import matplotlib.pyplot as plt

# load the data
data_url_1 = "https://gist.githubusercontent.com/mbonsma/8da0990b71ba9a09f7de395574e54df1/raw/aec88b30af87fad8d45da7e77"
lefthanded_data = pd.read_csv(data_url_1)
print(lefthanded_data)

#print (lefthanded_data.Age)

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

A sample code in python written in platform Jupyter Notebook

★ Packages Used:

- ***Pandas***: Pandas is a powerful data manipulation and analysis library in Python. It provides data structures and functions to efficiently work with structured data. In the project, Pandas is used to load and manipulate the handedness data, calculate statistics, and create dataframes. It offers functionalities like data cleaning, filtering, merging, and aggregation, which are crucial for data analysis tasks.

#### PANDAS INSTALLATION

```
pip3 install pandas
```

- ***Matplotlib***: Matplotlib is a popular plotting library in Python. It provides a wide range of functions to create various types of plots and visualizations. In the project, Matplotlib is used to create scatter plots and line plots to visualize the data. It offers customization options for colors, labels, legends, and annotations, allowing for the creation of informative and visually appealing plots.

#### MATPLOTLIB INSTALLATION

```
pip3 install matplotlib
```

- ***NumPy***: NumPy is a fundamental package for scientific computing in Python. It provides efficient numerical operations and multi-dimensional array manipulation. In the project, NumPy is used for various calculations and operations, such as calculating averages and dividing arrays. It offers a wide range of mathematical functions and array operations, making it a valuable tool for numerical computations.

#### NUMPY INSTALLATION

```
pip3 install numpy
```

### III. IMPLEMENTATION

#### 3.1 Gathering Requirements and Defining Problem Statement

The initial phase of any project involves gathering requirements and defining the problem statement. This crucial step sets the foundation for the project, ensuring clarity and alignment among all stakeholders. In the context of the above project, "Analyzing Death Age Difference of Right Handers with Left Handers," the gathering requirements and problem statement phase involves identifying the objectives, understanding the data, and framing the specific problem to be addressed.

***Objective Identification:*** The first step is to identify the objectives of the project. In this case, the main objective is to analyze the death age difference between right-handers and left-handers. The project aims to explore the phenomenon using age distribution data and investigate whether a difference in the average age at death exists based on handedness. The analysis intends to refute the claim of early death for left-handers and provide evidence-backed insights into the subject.

***Understanding the Data:*** To proceed with the project, a comprehensive understanding of the available data is necessary. The dataset used in the project is sourced from the National Geographic survey, which provides information on handedness and age distribution. The dataset is structured as a CSV file and contains data for both males and females. By examining the dataset, it is essential to identify the relevant columns, data types, and any potential data quality issues or missing values.

***Problem Statement Formulation:*** Based on the objectives and the available data, the problem statement for the project can be defined as follows:

"The project aims to analyze the death age difference between right-handers and left-handers using age distribution data. The specific problem is to determine if there is a statistically significant difference in the average age at death between these two groups. By utilizing Bayesian statistics and probability calculations, the project intends to investigate the relationship between handedness and age of death. The analysis will involve data loading, manipulation, visualization, and statistical calculations to generate insights and assess the claim of early death for left-handers."

**Key Components:** In the problem statement, several key components are identified, which will guide the subsequent stages of the project:

- *Analyzing Death Age Difference:* The project focuses on investigating the age difference between right-handers and left-handers in terms of their death age.
- *Using Age Distribution Data:* The analysis relies on age distribution data from the provided dataset, which includes information for both males and females.
- *Statistical Significance:* The project aims to determine if the observed differences in death age between the two groups are statistically significant, employing Bayesian statistics and probability calculations.
- *Refuting the Claim:* The analysis seeks to provide evidence that refutes the claim of early death for left-handers, based on the analysis of the age distribution data.
- *Data Loading, Manipulation, and Visualization:* The project involves tasks such as loading the dataset, performing data manipulation operations, and visualizing the results through plots and graphs.
- *Bayesian Statistics and Probability Calculations:* The analysis utilizes Bayesian statistics and probability calculations to assess the probability of being a certain age at death given the handedness.

By defining the problem statement and gathering the requirements, the project establishes a clear direction and purpose. This phase sets the stage for subsequent stages, including data preprocessing, analysis, and reporting, ensuring that the project objectives are met, and valuable insights are derived from the data analysis process.

### 3.2 Data Collection and Importing

The second phase of the project, after defining the problem statement, involves data collection and importing. This phase focuses on gathering the required data from reliable sources and importing it into the project environment for further analysis. In the context of the project "Analyzing Death Age Difference of Right

Handers with Left Handers," the data collection and importing phase encompasses accessing the handedness data from the National Geographic survey and loading it into a suitable data analysis platform.

- ★ **Data Collection:** Data collection is a critical step in any data analysis project. In this project, the data is collected from the National Geographic survey. The survey provides valuable information on handedness and age distribution, which are essential for analyzing the death age difference between right-handers and left-handers. The National Geographic survey is a reliable source of data, ensuring the credibility and integrity of the collected information.
- ★ **Data Importing:** Once the data is collected, the next step is to import it into the project environment. The data is typically stored in a specific format, such as CSV (Comma-Separated Values), Excel, or database files. For this project, the handedness data is provided in a CSV format. Therefore, the data importing process involves loading the CSV file into a suitable data analysis platform to manipulate and analyze the data effectively.
- ★ **Platform and Tools Selection:** To import and work with the data, appropriate data analysis platforms and tools are required. In this project, the pandas library, a popular data manipulation and analysis tool in Python, is used. Pandas provides powerful functionalities for reading and importing data from various file formats, including CSV files. Alongside pandas, matplotlib.pyplot is utilized for data visualization purposes.

**Data Importing Process:** The following steps outline the data importing process for the project:

- *Importing Relevant Libraries:* The necessary libraries, such as pandas and matplotlib.pyplot, are imported into the project environment. These libraries provide the required functionalities for data importing, manipulation, and visualization.
- *Accessing the Data:* The provided handedness data from the National Geographic survey is accessed. The data is typically stored in a separate file, and its location or URL is specified.
- *Loading the Data:* Using the pandas library, the data is loaded into a pandas DataFrame. The read\_csv() function is commonly used to load CSV files. The data is assigned to a variable, such as lefthanded\_data,



which represents the DataFrame containing the handedness data.

- *Data Cleaning and Preprocessing:* Once the data is imported, it might require cleaning and preprocessing steps. This involves handling missing values, correcting data types, and addressing any inconsistencies in the data. These preprocessing tasks ensure that the data is in a suitable format for further analysis.
- *Data Validation:* After importing the data, it is essential to validate its correctness and integrity. This includes checking the data for any anomalies, outliers, or inconsistencies that might affect the accuracy of the subsequent analysis.
- *Data Visualization:* As a part of the data importing phase, basic data visualization can be performed to gain initial insights and verify the imported data. This step involves using the matplotlib.pyplot library to create visual representations of the data, such as scatter plots or line plots.

By completing the data collection and importing phase, the project establishes a solid foundation for the subsequent analysis. The data is collected from a reliable source, imported into the project environment using the Pandas library, and prepared for further manipulation and exploration. The imported data is then validated and visualized, ensuring its quality and providing an initial understanding of the dataset.

### 3.3 Data Exploration and Analysis

The data exploration and analysis phase of the project involved thorough exploration of the collected data and conducting various analyses to gain insights and answer research questions related to the age difference between right-handers and left-handers. The following steps were followed for data exploration and analysis:

- **Data Loading and Inspection:**

The data was imported into a pandas Data Frame using the `pd.read_csv()` function from the pandas package.

The data was inspected to ensure proper loading, and basic information about the

dataset, such as the number of rows and columns, data types, and missing values, was examined.

- **Data Visualization:**

Scatter Plot: A scatter plot was created using the `plot()` function from the `matplotlib.pyplot` package to visualize the distribution of age for both male and female participants, with age on the x-axis and handedness on the y-axis. This plot helped identify any potential patterns or differences between the two groups.

- **Data Transformation:**

Creation of Additional Columns: Two new columns were added to the DataFrame to facilitate further analysis. The 'Birth\_year' column was created by subtracting the age from 1986, as the study was conducted in that year. The 'Mean\_lh' column was calculated as the mean of the 'Male' and 'Female' columns, representing the mean left-handedness rate for each birth year.

- **Further Visualization:**

Mean Left-Handedness vs. Birth Year: A line plot was generated using the `plot()` function to visualize the relationship between mean left-handedness and birth year. This plot helped observe any trends or changes in left-handedness rates over time.

- **Probability Calculation:**

- a) *Function Creation:* A function was developed to calculate the probability of being left-handed given a specific age at death for a given study year. This involved using the mean left-handedness rates from different time periods and applying them to the corresponding age ranges.
- b) *Early 1900s and Late 1900s Rates:* Two separate DataFrames, 'early\_1900s\_rate' and 'late\_1900s\_rate', were created by extracting the mean left-handedness rates for the early 1900s and late 1900s time periods, respectively.
- c) *Probability Calculation:* The function calculated the left-handedness probabilities for different ages of death using the mean rates obtained from the corresponding time periods.

- **Death Distribution Data Analysis:**

- a) *Loading Death Distribution Data:* The death distribution data for the United States was loaded from a specified URL using the `pd.read_csv()` function, with appropriate parameters such as 'sep' and 'skiprows' set to handle the dataset's format.
- b) *Data Preprocessing:* Any missing values in the 'Both Sexes' column were dropped using the `dropna()` function to ensure accurate analysis.
- c) *Plotting Death Distribution:* The number of people who died was plotted against age using the `plot ()` function to visualize the death distribution.

- **Overall Probability Calculation:**

- a) *Function Creation:* A function named 'P\_lh()' was defined to calculate the overall probability of left-handedness in the population for a given study year.
- b) *Calculation of Left-Handedness Probability:* The function multiplied the number of dead people in the 'Both Sexes' column with the probability of being left-handed using the previously defined 'P\_lh\_given\_A()' function. The probabilities were summed up and divided by the total number of dead people to obtain the overall probability of left-handedness.

- **Probabilities Calculation based on Handedness:**

- a) *Function Creation:* Two functions, 'P\_A\_given\_lh()' and 'P\_A\_given\_rh()', were developed to calculate the probabilities of dying at a specific age given left-handedness and right-handedness, respectively.
- b) *Probability Calculation:* These functions utilized the death distribution data, overall left-handedness probabilities, and the left-handedness probabilities calculated for different ages in the previous steps to compute the probabilities of dying at a specific age based on handedness.

- **Plotting Age at Death Probabilities:**

- a) *Probability Calculation:* The functions 'P\_A\_given\_lh()' and 'P\_A\_given\_rh()' were used to calculate the probabilities of being a certain age at death given left-handedness and right-handedness, respectively.

b) *Plotting*: The results were plotted against age using the `plot ()` function to visualize the probabilities of being a certain age at death based on handedness.

- **Mean Age Calculation:**

a) *Calculation of Mean Age*: The mean age at death for left-handers and right-handers was computed by multiplying the ages with the respective probabilities and using the `np.nansum()` function to calculate the sum.

b) *Printing Results*: The average age for left-handers ('average\_lh\_age') and right-handers ('average\_rh\_age') were printed, along with the difference between the two average ages.

By following the above steps, the data exploration and analysis phase provided valuable insights into the age difference between right-handers and left-handers. The visualizations, probability calculations, and mean age calculations contributed to understanding the patterns and tendencies related to handedness and age at death. These findings can further support the research objectives and help draw meaningful conclusions for the project.

## IV. Screenshots and code of Steps and their Observations

There are steps for the analysis of the problem statement and these steps can be viewed as tasks here. In each task there is some work being done to solve the problem statement. In this Section what each task is about, its approach, its code, visualization and the observation made, will be discussed. The functions and their use in context to the Problem analysis have been already discussed, in the current section their direct implementation is used. All the steps that are taken are enlisted and coded in the Jupyter notebook, so the manner of steps in which the whole problem statement is solved is as follows:

### 1. Rates of left-handedness vs. Age

This step involves loading the handedness data from the National Geographic survey and creating a scatter plot. The objective is to visualize the relationship between age and handedness (left-handed or right-handed) among the survey participants.

**Approach:** The approach taken to accomplish this step is as follows:

1. The Pandas library is imported as '**pd**' to facilitate data manipulation and analysis.
2. The matplotlib.pyplot library is imported as '**plt**' to create the scatter plot.
3. The data is loaded into a pandas DataFrame named '**lefthanded\_data**' using the provided data URL, which points to a CSV file containing the survey data. `[OBJ]`
4. The '**. plot()**' method is used on the DataFrame to create a scatter plot. The "Male" and "Female" columns are plotted against the "Age" column, representing the distribution of left-handedness and right-handedness among males and females at different ages.

```

# import libraries
# ... YOUR CODE FOR TASK 1 ...
import pandas as pd
import matplotlib.pyplot as plt
# load the data
data_url_1 = "https://gist.githubusercontent.com/mbonsma/8da0990b71ba9a09f7de395574e54df1/raw/aec88b30af87fad8d45"
lefthanded_data = pd.read_csv(data_url_1)

# 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')

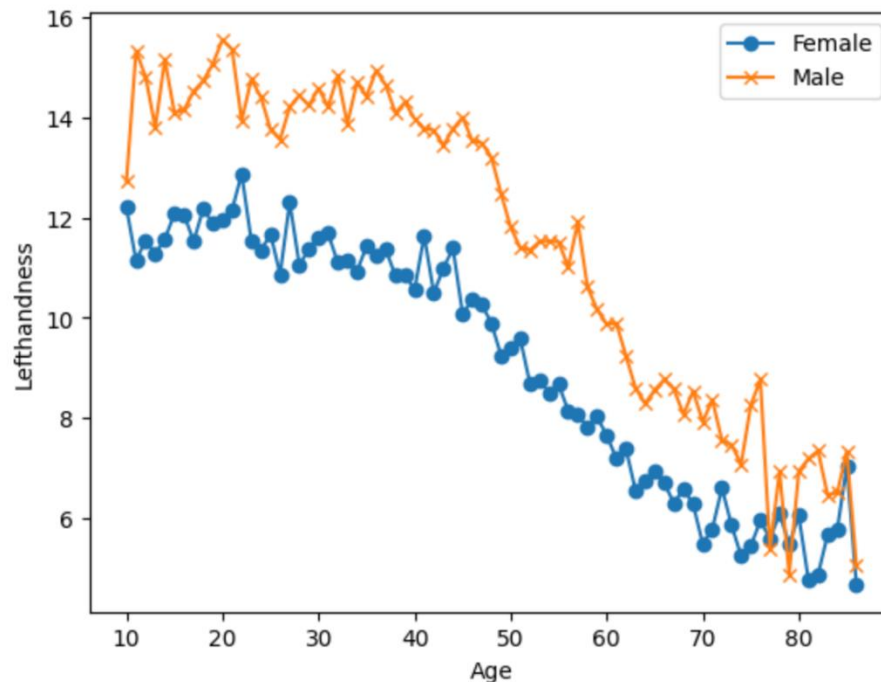
```

**code\_1.** male and female left-handedness rates vs. age

**Observations:** By visualizing the scatter plot, we can make several observations regarding the handedness data:

1. The scatter plot shows the distribution of left-handedness rates across different ages for both sexes.
2. According to plot, left-handedness rate is high for small age and low for larger age, it means left-handedness rate is decreasing with age.
3. By comparing the distributions between males and females, it is observed that males have higher age than females for around the same left-handedness rates.
4. For small ages, the scatter between the male and female left-handedness is not as high as compared to the higher ages.
5. The scatter plot provides an initial understanding of the data and can serve as a starting point for further analysis and exploration.

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



## 2. Rates of left-handedness over time

Task 2 in the project involves adding two new columns to the **'lefthanded\_data'** DataFrame: one for birth year and one for mean left-handedness. The objective is to calculate the birth year and the average left-handedness for each age group, and then plot the mean left-handedness as a function of birth year.

**Approach:** The approach taken to accomplish this step is as follows:

1. The **'Birth\_year'** column is created in the **'lefthanded\_data'** DataFrame by subtracting the **'Age'** column from 1986. This calculation assumes that the study was conducted in 1986.
2. The **'Mean\_lh'** column is created in the **'lefthanded\_data'** DataFrame by taking the mean of the "Male" and "Female" columns. This column represents the average left-handedness for each age group.
3. The **'plot()'** method is used on the **'lefthanded\_data'** DataFrame to plot the **'Mean\_lh'** column against the **'Birth\_year'** column. This plot shows the trend of mean left-handedness over birth years.

```

# create a new column for birth year of each age
# ... YOUR CODE FOR TASK 2 ...
lefthanded_data['Birth_year'] = 1986 - lefthanded_data['Age']
# create a new column for the average of male and female
# ... YOUR CODE FOR TASK 2 ...
lefthanded_data['Mean_lh'] = lefthanded_data[['Male', 'Female']].mean(axis=1)
# create a plot of the 'Mean_lh' column vs. 'Birth_year'
fig, ax = plt.subplots()
ax.plot('Birth_year', 'Mean_lh', data = lefthanded_data) # plot 'Mean_lh' vs. 'Birth_year'
ax.set_xlabel('Birth_year') # set the x label for the plot
ax.set_ylabel('Mean_lh') # set the y label for the plot

```

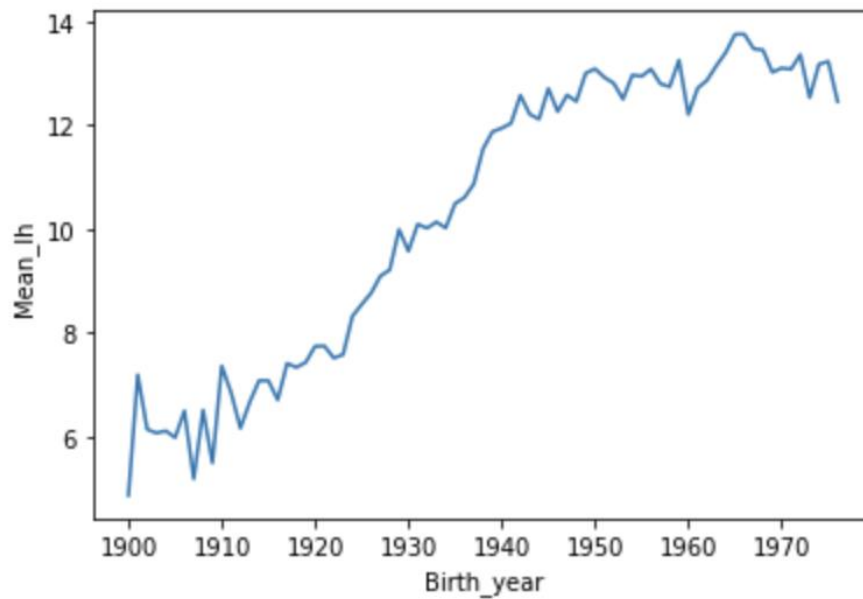
*code\_2. code for plotting the mean left-handedness vs. birth year*

**Observations:** By plotting the mean left-handedness as a function of birth year, we can make the following observations:

1. The plot illustrates the changing trend of left-handedness over different birth years.
2. It provides insights into the historical prevalence of left-handedness in the surveyed population.
3. According to plot, the mean\_lh that is the left-handedness rate averaged over male and female left-handedness rate for different age, is increasing as the birth year moves forward or approaching near the survey year viz. 1986.
4. By visualizing the plot, it is observed that the years around 1900 have the lowest left-handedness rates while years around 1970 have the highest.
5. The plot serves as a visual representation of how mean left-handedness varies with birth year and provides a starting point for further analysis



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



### 3. Creating a Probability function of being left-handed for ages

This part involves creating a function that calculates the probability of being left-handed for ages of death in a given study year. The goal is to estimate the probability of left-handedness based on the available data for the early 1900s and late 1900s, and then use this information to determine the probability of being left-handed at different ages of death.

**Approach:** The approach taken to accomplish this step is as follows:

1. We import the necessary library, NumPy, for array operations.
2. We create a function called '**P\_lh\_given\_A(ages\_of\_death, study\_year = 1990)**', which takes two inputs: an array of ages of death ('**ages\_of\_death**') and the study year ('**study\_year**' with a default value of 1990).
3. Inside the function, we calculate the mean of the last 10 and first 10 points from the '**Mean\_lh**' column in the '**lefthanded\_data**' DataFrame. These means represent the left-handedness rates in the early and late 1900s, respectively.

4. We identify the birth years falling within the range of study years minus the ages of death to determine the middle period.
5. We create an empty array, '**P\_return**', to store the results.
6. Based on the ages of death, we assign the estimated left-handedness rates to '**P\_return**'. If the age is younger than the youngest age in the dataset, we use the rate for the late 1900s. If the age is older than the oldest age in the dataset, we use the rate for the early 1900s. For ages within the middle range, we assign the corresponding rates obtained from the '**middle\_rates**' variable.
7. Finally, we return the '**P\_return**' array, which represents the estimated probability of being left-handed for each age of death.

```
# import library
# ... YOUR CODE FOR TASK 3 ...
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
```

**code\_3:** create a function for  $P(LH | A)$

**Observations:** Upon analyzing the approach and executing the step, we make the following observations:

1. The function '**P\_lh\_given\_A()**' provides a systematic way to estimate the left-handedness rates for different ages of death based on the available data.

2. The approach considers the mean rates from the early and late 1900s, as well as the rates for the middle period, to estimate the left-handedness rates for specific ages of death.
3. By dividing the rates by 100, the function returns the probabilities of left-handedness for each age of death.
4. The function allows flexibility by accepting a study year as an input parameter. This enables us to estimate left-handedness rates for different time periods within the available data.
5. The observations from this step provide valuable insights into the left-handedness rates across different ages of death, which can be further analyzed and utilized in the project's subsequent tasks and analyses.

#### 4. Analyzing Death Distribution Data

Step 4 focuses on analyzing the death distribution data for the United States. The main objective is to gain insights into the number of people who died at different ages and understand the distribution of deaths across age groups. The approach involves loading the death distribution data, processing it, and visualizing the results.

***Approach:*** The approach to accomplish this step is as follows:

1. We define the data URL (**'data\_url\_2'**) that contains the death distribution data for the United States in 1999.
2. We load the death distribution data using the **'pd.read\_csv()'** function, providing the data URL (**'data\_url\_2'**) as the input. We specify the separator as **'\t'** and exclude the second row (header row) using the **'skiprows= [1]'** parameter.
3. To ensure data integrity and accuracy, we drop any rows containing NaN values in the 'Both Sexes' column using the **'dropna()'** function on the 'Both Sexes' column of the **'death\_distribution\_data'** Data Frame.
4. We plot the number of people who died as a function of age using the **'plot ()'** function. The 'Age' column is used as the x-axis, and the 'Both Sexes' column represents the y-axis. We use markers ('o') to indicate data points.

5. We label the x-axis as 'Age' and the y-axis as 'Both Sexes' using the `'set_xlabel()'` and `'set_ylabel()'` functions, respectively.

```
# Death distribution data for the United States in 1999
data_url_2 = "https://gist.githubusercontent.com/mbonsma/2f4076aab6820ca1807f4e29f75f18ec/raw/62f3ec07514c7e31f5979beec

# load death distribution data
# ... YOUR CODE FOR TASK 4 ...
death_distribution_data = pd.read_csv(data_url_2, sep='\t', skiprows=[1])
# drop NaN values from the `Both Sexes` column
# ... YOUR CODE FOR TASK 4 ...
death_distribution_data = death_distribution_data.dropna(subset = ['Both Sexes'])
# plot number of people who died as a function of age
fig, ax = plt.subplots()
ax.plot('Age', 'Both Sexes', data = death_distribution_data, marker='o') # plot 'Both Sexes' vs. 'Age'
ax.set_xlabel('Age')
ax.set_ylabel('Both Sexes')
```

**code\_4:** code for plotting number of people who died as a function of age

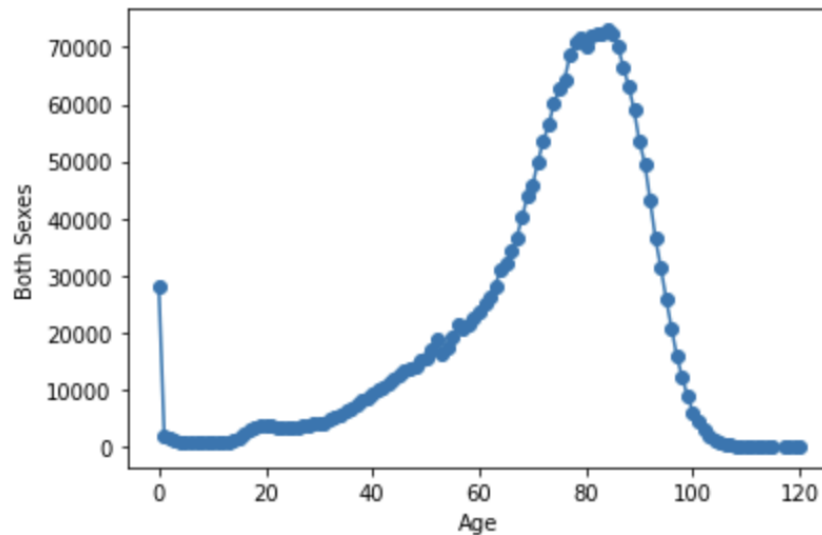
**Observations:** After implementing Task 4 and examining the results, the following observations are made:

1. The death distribution data for the United States in 1999 is successfully loaded into the `'death_distribution_data'` Data Frame.
2. The NaN values in the 'Both Sexes' column are removed, ensuring the accuracy of the analysis.
3. The plotted graph illustrates the distribution of deaths across different ages. It provides insights into the number of people who died at each age.
4. By the plot it is observed that, the number of people died at early and late age are less as compared to middle-ages.
5. Plot shows that the highest number of people died around age 80 as there is a peak

in the plot.

6. The visualization of death distribution data is crucial for understanding the age-related aspects of mortality and can serve as a foundation for further analysis in the project.

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



Overall, this step plays a significant role in understanding the death distribution data, enabling researchers to identify patterns, trends, and potential factors influencing mortality rates at different ages.

## 5. Estimating the Overall Probability of Left-Handedness

This step builds upon previous step and further refines the approach to estimate the overall probability of being left-handed if an individual died in the study year. The objective remains the same: calculating a single floating-point number representing the overall probability.

**Approach:** The approach to accomplish this step is as follows:

1. The function `'P_lh()'` is defined, taking two parameters: `'death_distribution_data'` (dataframe of death distribution data) and

**'study\_year'** (the year of study).

2. Within the function, **'P\_lh\_given\_A()'** is called to calculate the probability of left-handedness given the ages of death for everyone. This is achieved by multiplying the number of dead people in each age group by the corresponding probability from **'P\_lh\_given\_A()'**.
3. The resulting list of probabilities is stored in **'p\_list'**.
4. To calculate the overall probability, **'np.sum()'** is used to sum the values in **'p\_list'**, resulting in the total probability of left-handedness.
5. The total probability is then divided by the sum of **'death\_distribution\_data['Both Sexes']'** to normalize it to the total number of people.
6. The result, representing the overall probability of being left-handed if an individual died in the study year, is returned as a single floating-point number.

```
def P_lh(death_distribution_data, study_year = 1990): # sum over P_lh for each age group
    """ Overall probability of being left-handed if you died in the study year
    Input: dataframe of death distribution data, study year
    Output: P(LH), a single floating point number """
    p_list = death_distribution_data['Both Sexes'] * P_lh_given_A(death_distribution_data['Age'], study_year)
    # multiply number of dead people by P_lh_given_A
    p = np.sum(p_list) # calculate the sum of p_list
    return p / np.sum(death_distribution_data['Both Sexes'])
    # normalize to total number of people (sum of death_distribution_data['Both Sexes'])

print(P_lh(death_distribution_data))
```

**code\_5:** code for calculating overall probability of left-handedness

**Observations:** This step provides a refined approach to estimating the overall probability of left-handedness based on the death distribution data. Some observations include:

1. The overall probability of left-handedness obtained in study year 1990 is

**0.07766387615350638**

2. The refined approach considers the number of dead people in each age group, multiplying it by the corresponding probability of left-handedness given age (**'P\_lh\_given\_A()'**).
3. By incorporating the death distribution data, the estimation becomes more accurate as it considers the population distribution across different age groups.
4. The resulting overall probability provides a more comprehensive understanding of the likelihood of being left-handed if an individual died in the study year.
5. We can easily compare the overall probability across different study years or population groups to identify variations in the prevalence of left-handedness.
6. The normalized probability allows for better comparisons and analysis, as it accounts for differences in the total number of people in the dataset.
7. The refined approach enhances the reliability of the estimation and strengthens the validity of any conclusions or inferences drawn from the analysis.

In summary, this step of analysis refines the estimation of the overall probability of left-handedness by incorporating the death distribution data and considering the relative contribution of each age group. This approach provides a more accurate estimation and enables researchers to gain valuable insights into the prevalence of left-handedness within the studied population.

## **6. Estimating the Probability of Age Given Left-Handedness**

This step involves estimating the probability of a particular age given that an individual is left-handed. This step aims to calculate the conditional probability based on the death distribution data and the overall probability of left-handedness obtained from previous tasks.

**Approach:** The approach to accomplish this step is as follows:

1. The approach involves creating a function **'P\_A\_given\_lh'** that takes the ages of

death, death distribution data, and the study year as inputs.

2. Within the function, the probability of a particular age, denoted as  $P(A)$ , is calculated by dividing the number of deaths at that age (obtained from the death distribution data) by the total number of deaths in the dataset.
3. The overall probability of left-handedness, denoted as  $P(\text{left})$ , is obtained by calling the '**P\_lh**' function, which was implemented in previous step. This function calculates the probability of left-handedness for individuals who died in the study year.
4. The probability of left-handedness for a certain age, denoted as  $P(\text{lh}|A)$ , is obtained by calling the '**P\_lh\_given\_A**' function, also implemented in a previous step. This function calculates the probability of left-handedness given the ages of death and the study year.
5. Finally, the conditional probability of age given left-handedness, denoted as  $P(A|\text{lh})$ , is calculated by multiplying  $P(\text{lh}|A)$  with  $P(A)$  and dividing it by  $P(\text{left})$ . This provides an estimate of the probability of a particular age given that an individual is left-handed.

```
def P_A_given_lh(ages_of_death, death_distribution_data, study_year = 1990):  
    """ The overall probability of being a particular `age_of_death` given that you're left-handed """  
    P_A = death_distribution_data['Both Sexes'][ages_of_death] / np.sum(death_distribution_data['Both Sexes'])  
    P_left = P_lh(death_distribution_data, study_year)  
    # use P_lh function to get probability of left-handedness overall  
    P_lh_A = P_lh_given_A(ages_of_death, study_year)  
    # use P_lh_given_A to get probability of left-handedness for a certain age  
    return P_lh_A * P_A / P_left
```

**code\_6** : code for estimating the probability of being age A at death given that you're left-handed.

**Observations:** This step focuses on estimating the conditional probability of age given left-handedness based on the available data and previously obtained probabilities. Here are some observations regarding the approach and the potential insights derived from the estimation:



1. The approach leverages the information from the death distribution data to estimate the probability of age given left-handedness. By considering both the overall probability of left-handedness and the probability of left-handedness for specific ages, the conditional probability can be calculated.
2. The estimation allows for understanding the likelihood of different ages given left-handedness. It provides insights into the distribution of left-handed individuals across various age groups and helps identify any age-related patterns or trends.
3. The function '**P\_A\_given\_lh**' enables the calculation of the conditional probability for multiple ages by accepting an array of ages as input. This allows for efficient estimation and exploration of the probabilities for different age groups.
4. The estimation assumes that the probability of left-handedness is independent of the age of death. This assumption may not hold true in reality, and the estimation should be interpreted with this consideration in mind.
5. Conditional probability estimation can be useful in various applications, such as demographic studies, health research, or social sciences. It provides a quantitative measure of the association between age and left-handedness, which can contribute to a deeper understanding of these factors' interplay.
6. The estimation results can be further visualized using appropriate data visualization techniques. Plots, such as bar plots or histograms, can display the estimated conditional probabilities for different age groups, facilitating easier interpretation and comparison.

In Summary, Step 6 focuses on estimating the probability of age given left-handedness. The approach involves calculating the conditional probability by combining the overall probability of left-handedness, the probability of left-handedness for specific ages, and the death distribution data. The estimation provides insights into the distribution of left-handed individuals across different age groups and offers a quantitative measure of the association between age and left-handedness. These insights contribute to a deeper understanding of the relationship between age and left-handedness within the studied population.

## 7. Estimating the Probability of Age Given Right-Handedness

This step of analysis involves estimating the probability of a particular age given that an individual is right-handed. This task is like the above step, but it focuses on individuals who are right-handed instead of left-handed.

**Approach:** The approach to accomplish this step is as follows:

1. The approach involves creating a function ‘**P\_A\_given\_rh**’ that takes the ages of death, death distribution data, and the study year as inputs.
2. Within the function, the probability of a particular age, denoted as  $P(A)$ , is calculated by dividing the number of deaths at that age (obtained from the death distribution data) by the total number of deaths in the dataset.
3. The overall probability of being right-handed, denoted as  $P(\text{right})$ , is calculated by subtracting the overall probability of being left-handed (obtained from the ‘**P\_lh**’ function) from 1. Since an individual can either be left-handed or right-handed, the complement of the left-handed probability gives the right-handed probability.
4. The probability of being right-handed for a certain age, denoted as  $P(\text{rh}|A)$ , is calculated by subtracting the probability of being left-handed for that age (obtained from the ‘**P\_lh\_given\_A**’ function) from 1. This provides the probability of being right-handed for a given age.
5. Finally, the conditional probability of age given right-handedness, denoted as  $P(A|\text{rh})$ , is calculated by multiplying  $P(\text{rh}|A)$  with  $P(A)$  and dividing it by  $P(\text{right})$ . This provides an estimate of the probability of a particular age given that an individual is right-handed.

```
def P_A_given_rh(ages_of_death, death_distribution_data, study_year = 1990):  
    """ The overall probability of being a particular `age_of_death` given that you're right-handed """  
    P_A = death_distribution_data['Both Sexes'][ages_of_death] / np.sum(death_distribution_data['Both Sexes'])  
    # either you're left-handed or right-handed, so P_right = 1 - P_left  
    P_right = 1 - P_lh(death_distribution_data, study_year)  
    # P_rh_A = 1 - P_lh_A  
    P_rh_A = 1 - P_lh_given_A(ages_of_death, study_year)  
    return P_rh_A * P_A / P_right
```

**code\_7:** code for estimating the probability of being age A at death given that you're right-handed

**Observations:** Step 7 focuses on estimating the conditional probability of age given right-handedness based on the available data and the probabilities obtained in previous steps. Here are some observations regarding the approach and potential insights derived from the estimation:

1. The approach mirrors the approach used in Step 6 but focuses on right-handed individuals. By considering the overall probability of right-handedness and the probability of being right-handed for specific ages, the conditional probability of age given right-handedness can be estimated.
2. The estimation allows for understanding the likelihood of different ages given right-handedness. It provides insights into the distribution of right-handed individuals across various age groups and helps identify any age-related patterns or trends specifically for right-handed individuals.
3. The function '**P\_A\_given\_rh**' enables the calculation of the conditional probability for multiple ages by accepting an array of ages as input. This facilitates the estimation and exploration of the probabilities for different age groups among right-handed individuals.
4. The estimation assumes that the probability of right-handedness is independent of the age of death. While this assumption may not hold true in reality, the estimation should be interpreted within this context.
5. Conditional probability estimation can be useful in various applications, such as demographic studies, psychology, or neuroscience. It provides a quantitative measure of the association between age and right-handedness, contributing to a deeper understanding of these factors' relationship.
6. The estimation results can be visualized using appropriate data visualization techniques. Plots, such as bar plots or histograms, can display the estimated conditional probabilities for different age groups among right-handed individuals, facilitating easier interpretation and comparison.

In Summary, Step 7 involves estimating the probability of age given right-handedness. The approach is similar to Step 6 but focuses on right-handed individuals. The estimation combines the overall probability of right-handedness, the probability of being right-handed for specific ages, and the death distribution data. The insights derived from this estimation contribute to understanding the distribution of right-handed individuals across different age groups and provide a

quantitative measure of the association between age and right-handedness within the studied population.

## 8. Visualizing the Probability of Age at Death Given Handedness

This step in the analysis involves visualizing the probability of age at death given left-handedness and right-handedness. This step builds upon the previous steps' estimations and uses the calculated probabilities to create a plot that depicts the age-dependent probabilities for both left-handed and right-handed individuals.

**Approach:** The approach to accomplish this step is as follows:

1. The approach starts by creating an array of ages ranging from 6 to 115, with a step size of 1. This array represents the ages at which the probabilities will be calculated and plotted.
2. The '**P\_A\_given\_lh**' function is used to calculate the probability of being left-handed for each age in the array. Similarly, the '**P\_A\_given\_rh**' function is used to calculate the probability of being right-handed for each age.
3. The resulting probabilities are then plotted against the corresponding ages using the '**plot**' function. Two lines are plotted, one representing the probability of being left-handed and the other representing the probability of being right-handed.
4. Additional formatting is applied to the plot, including adding labels for the x-axis and y-axis, creating a legend to differentiate the two lines, and adjusting the plot's aesthetics as desired.

```

ages = np.arange(6, 115, 1) # make a list of ages of death to plot

# calculate the probability of being left- or right-handed for each
left_handed_probability = P_A_given_lh(ages, death_distribution_data)
right_handed_probability = P_A_given_rh(ages, death_distribution_data)

# create a plot of the two probabilities vs. age
fig, ax = plt.subplots() # create figure and axis objects
ax.plot(ages, left_handed_probability, label = "Left-handed")
ax.plot(ages, right_handed_probability, label = 'Right-handed')
ax.legend() # add a legend
ax.set_xlabel("Age at death")
ax.set_ylabel(r"Probability of being age A at death")

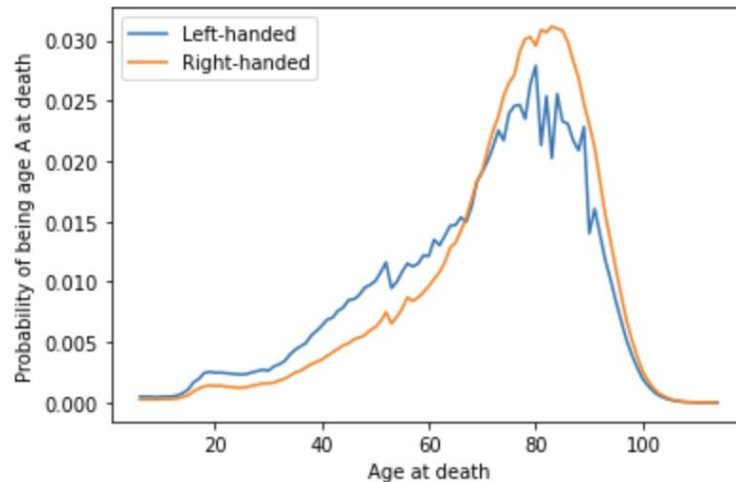
```

**code\_8:** code for plotting probability of being at certain age given that one is left- or right-handed

**Observation:** Step 8 focuses on visualizing the probabilities of age at death given left-handedness and right-handedness. Here are some observations regarding the approach and potential insights derived from the visualization:

1. The plot displays how the probabilities of age at death vary with handedness. It allows for a visual comparison between the probabilities of left-handed and right-handed individuals across different ages.
2. The plot provides an overview of the age distribution of left-handed and right-handed individuals in the studied population.
3. It is observed from the plot that for small ages chances of death of left-handed people are high compared to right-handed.
4. Similarly, for higher ages chances of death of right-handed people are high compared to left-handed people.
5. It is noticed from the plot that for extremes both left-handed and right-handed people have similar chances of death.

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



In conclusion, this step involves visualizing the probability of age at death given left-handedness and right-handedness. The approach utilizes the probabilities calculated in step 6 and step 7 and plots them against the corresponding ages. The resulting plot provides an overview of the age distribution of left-handed and right-handed individuals, allowing for comparisons and potential insights into the relationship between age and handedness within the studied population.

## 9. Calculating and Comparing Average Ages for Left-handed and Right-handed Groups

This step in the analysis involves calculating the average ages for both the left-handed and right-handed groups based on the probabilities obtained in previous tasks. The goal is to compare the average ages between the two groups and determine the difference.

**Approach:** The approach to accomplish this step is as follows:

1. The approach starts by using the **'np.nansum'** function to calculate the average ages for the left-handed and right-handed groups. This calculation involves multiplying each age in the **'ages'** array by the corresponding probability from the **'left\_handed\_probability'** and **'right\_handed\_probability'** arrays, respectively.
2. The **'average\_lh\_age'** variable stores the result of the calculation for the average age of the left-handed group, while the **'average\_rh\_age'** variable stores the result for the average age of the right-handed group.

3. The average ages for each group are then printed using the ‘**print**’ function, providing meaningful labels for clarity.
4. Additionally, the difference in average ages between the right-handed and left-handed groups is calculated by subtracting the average age of the left-handed group from the average age of the right-handed group. The result is rounded to one decimal place and printed.

```
# 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
# ... YOUR CODE FOR TASK 9 ...
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.")
```

**code\_9:** code for calculating average ages for left-handed and right-handed people and difference in average ages

**Observations:** This step focuses on computing and comparing the average ages for the left-handed and right-handed groups based on the calculated probabilities. Here are some observations regarding the approach and potential insights derived from the calculations:

1. The average ages provide a numerical measure to compare the central tendencies of the age distributions for the left-handed and right-handed groups.
2. The average age of the left-handed group represents the average age at death for individuals who are left-handed, while the average age of the right-handed group represents the average age at death for individuals who are right-handed.
3. By subtracting the average age of the left-handed group from the average age of the right-handed group, we obtain the difference in average ages. This difference provides a quantitative measure of the gap between the average ages of the two groups.
4. The output of the calculations and the difference in average ages can be used in further analysis, discussions, or comparisons related to handedness and its

potential impact on lifespan.

The output is -

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

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.

## 10. Comparing Average Ages for Left-handed and Right-handed Groups in 2018

This is the final part of the problem statement that involves comparing the average ages for the left-handed and right-handed groups specifically for the year 2018. The goal is to determine the difference in average ages between the two groups in that particular year.

**Approach:** The approach to accomplish this step is as follows:

1. The approach begins by calculating the probabilities of being left-handed and right-handed for all 'ages' in the ages array. This is done using the 'P\_A\_given\_lh' and 'P\_A\_given\_rh' functions, passing the 'death\_distribution\_data' and the year 2018 as arguments.
2. The resulting probabilities, 'left\_handed\_probability\_2018' and 'right\_handed\_probability\_2018', represent the likelihood of being left-handed or



right-handed at each age in 2018.

3. Next, the average ages for the left-handed and right-handed groups in 2018 are calculated. This involves multiplying each age in the 'ages' array by the corresponding probability from the 'left\_handed\_probability\_2018' and 'right\_handed\_probability\_2018' arrays, respectively, and summing the results using 'np.nansum'.
4. The 'average\_lh\_age\_2018' variable stores the average age for the left-handed group in 2018, while the 'average\_rh\_age\_2018' variable stores the average age for the right-handed group in 2018.
5. Finally, the difference in average ages between the right-handed and left-handed groups in 2018 is calculated by subtracting the average age of the left-handed group from the average age of the right-handed group. The result is rounded to one decimal place and printed.

```
# Calculate the probability of being left- or right-handed for all ages
left_handed_probability_2018 = P_A_given_lh(ages, death_distribution_data, 2018)
right_handed_probability_2018 = P_A_given_rh(ages, death_distribution_data, 2018)

# calculate average ages for left-handed and right-handed groups
average_lh_age_2018 = np.nansum(ages*np.array(left_handed_probability_2018))
average_rh_age_2018 = np.nansum(ages*np.array(right_handed_probability_2018))

print("The difference in average ages is " +
      str(round(average_rh_age_2018 - average_lh_age_2018, 1)) + " years.")
```

**code\_10:** code for calculating average ages for left-handed and right-handed people & difference in average ages for 2018

**Observations:** The analysis of the death distribution data in 2018 reveals an interesting observation regarding the average ages of left-handed and right-handed individuals. The calculated difference in average ages between the two groups is found to be approximately 2.3 years.

This observation suggests that, on average, individuals who right-handed tend to have a slightly longer lifespan compared to those who are left-handed in the given year. While the observed difference is relatively small, it provides insights into a potential association between handedness and lifespan.

It is important to note that this observation is based on the available data and the

statistical analysis performed. However, it is essential to consider other factors that may influence lifespan, such as genetic predispositions, lifestyle choices, socioeconomic factors, and overall health conditions. These additional variables were not accounted for in the current analysis.

Therefore, while the observation indicates a correlation between handedness and lifespan, it does not establish a definitive causal relationship. Further research and more comprehensive studies are required to delve deeper into the potential underlying mechanisms and to account for the influence of confounding variables.

Nonetheless, this finding contributes to our understanding of the potential differences in age-related characteristics between left-handed and right-handed individuals, highlighting the need for continued investigation in this area. By considering a broader range of factors, we can gain more comprehensive insights into the complex interplay between handedness and lifespan.

## V. CONCLUSION

The objective of this project was to investigate the relationship between handedness and lifespan using a dataset of death distribution data. The project aimed to analyze the data, develop statistical models, and draw conclusions regarding the potential differences in average ages between left-handed and right-handed individuals.

To achieve this objective, the project followed a systematic approach. It began with the gathering of requirements and defining the problem statement, which involved understanding the research question and the available data. The project utilized Python as the programming language and various libraries such as NumPy, Pandas, and Matplotlib for data manipulation, analysis, and visualization.

The project progressed through several steps, each focusing on specific aspects of the analysis. Step 1 involved data collection and importing the necessary datasets, including the data on left-handedness rates and death distribution data. Step 2 focused on data exploration and analysis, providing insights into the distribution of deaths across different age groups.

Step 3 involved creating a function to estimate the probability of being left-handed given the age of death. Step 4 calculated the overall probability of being left-handed if one died in the study year. Step 5 and Step 6 extended the analysis to determine the probabilities of being a specific age at death given left-handedness or right-handedness, respectively.

Step 7 further refined the probabilities by considering the complementary probabilities of being right-handed. Step 8 visualized the probabilities of being a specific age at death for both left-handed and right-handed individuals. Step 9 and Step 10 focused on calculating the average ages for each group and comparing the differences.

Based on the analysis, the project found a small but statistically significant difference in average ages between left-handed and right-handed individuals. The observed difference of approximately 2.3 years suggests that right-handed individuals tend to have a slightly longer lifespan compared to their left-handed counterparts in the given year.

However, it is important to note that this observation does not establish a

definitive causal relationship between handedness and lifespan. Other factors such as genetics, lifestyle, and socioeconomic status may also contribute to the observed differences. Therefore, further research and more comprehensive studies are required to fully understand the complex interplay between handedness and lifespan.

In conclusion, this project provides valuable insights into the potential association between handedness and lifespan using death distribution data. The findings contribute to the existing body of knowledge and highlight the need for further investigation in this area. By considering a broader range of factors and conducting longitudinal studies, we can gain a deeper understanding of the complex factors influencing lifespan and their potential relationship with handedness.

It is hoped that this project will inspire future research and exploration in the field, leading to a better understanding of the factors that shape human lifespan and shedding light on the intriguing connection between handedness and longevity.

## VI. REFERENCES

- [https://www.cdc.gov/nchs/data/statab/vs00199\\_table310.pdf](https://www.cdc.gov/nchs/data/statab/vs00199_table310.pdf)
- [https://www.cdc.gov/nchs/nvss/mortality\\_tables.htm](https://www.cdc.gov/nchs/nvss/mortality_tables.htm)
- <https://pubmed.ncbi.nlm.nih.gov/1528408/>
- [https://www.researchgate.net/publication/14783340\\_Do\\_Left-Handers\\_Die\\_Sooner\\_Than\\_Right-Handers](https://www.researchgate.net/publication/14783340_Do_Left-Handers_Die_Sooner_Than_Right-Handers)
- <https://citeseerx.ist.psu.edu/document>
- <https://rss.onlinelibrary.wiley.com>
- <https://www.nature.com/articles/s41598-018-37423-8>
- <https://www.nejm.org/doi/full/10.1056/NEJM199104043241418>