Final Report of Traineeship Program 2023

On

"Analyze Death Age Difference of Right Handers with Left"

Presented by:

Ashwini K S

Presented to:

MEDTOUREASY

28th July 2023

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to everyone who supported me during my time as a trainee and contributed to the successful completion of this project. The time spent here as a trainee has been a great learning experience for me, and I am grateful for the opportunity to interact with professionals.

Firstly, special thanks to the organization MedTourEasy and their Training and Development Team for giving me the opportunity to carry out my internship at their esteemed organization. I would especially like to thank my supervisor, Ankit Hasija, for constantly supporting and guiding me through the project.

Lastly, I want to express my gratitude to my family and friends who have been understanding and provided full support for my future endeavours. Thank you all for your valuable contributions.

TABLE OF CONTENTS

ABSTRACT

Lifespan studies have shown that the percentage of left-handed individuals in the population diminishes steadily, resulting in a lower average number of older left-handed people compared to right-handed people. The authors' initial conclusion was that left-handed individuals tended to pass away earlier, approximately 9 years sooner than their right-handed counterparts. However, this result was later deemed biased, prompting further study to accurately identify the underlying cause. To achieve this, researchers employed the concept of conditional probability and examined the average age of death for left-handed and right-handed individuals during different time periods. Our study replicates this approach to investigate the age difference between the average lifespan of left-handed and right-handed individuals who are currently alive.

1) 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:

The true correlation between left-handedness and death, which causes the average age of left-handed people is attempted to be captured. In order to reduce bias in the result, statistical analysis is performed correctly. Same is done for different time frames to realise how the change in age differences have been captured.

This project aims at collecting, analysing and drawing statistical insights to obtain results.

1.3 Objectives and Deliverables:

In this project, you 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.

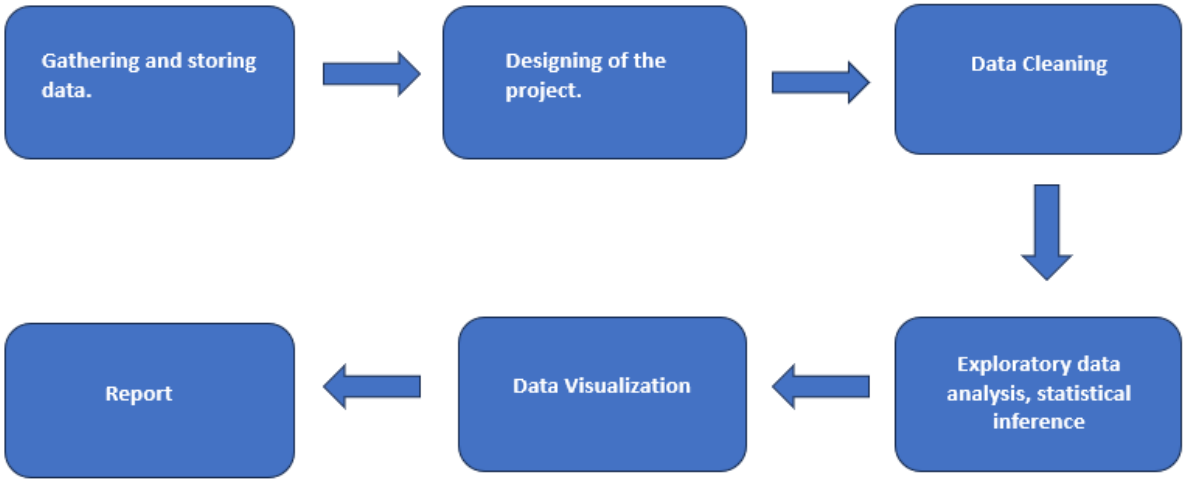
The project is carried out mainly in two steps.

(i) Firstly, we check the trend of percentage of people being left-handed with age and also death rates with age. Then we calculate the probability of people dying at an age, given that they are left-handed and right-handed.

(ii) Then we proceed to calculate the average age of Left-handed people and Right-handed people, following which the age differences are calculated. Then, we look at the trend in percentage change over the years with age.

2) METHODOLOGY

2.1) Flow of the project:



Above is the methodology used in the project. The gathering and storing of data were already taken care of. Hence, we fetched that data and used this data to design the project accordingly. Then we proceed to data cleaning. The data is already in a structured format hence we may proceed. We then check for duplicates and null, remove them and proceed. Next, exploratory data analysis carried out. In this method, we try to understand the statistics of the data. The data available is already a good data, hence we proceed to calculate the probabilities. Then we use visualization, i.e plot graphs to understand the trend. Lastly, we create a report and present it in the form of a report.

2.2) Tools Used:

2.2.1) Language: Python

Python is one of the most versatile, easy to use, easy to understand and open languages, which is very useful to a data analyst and data scientist. Python's development began in 1989 when a Dutch programmer started working on a new programming language. It's name came from a famous television show that the programmer watched frequently. Python works in various Integrated Development Environment (IDE).

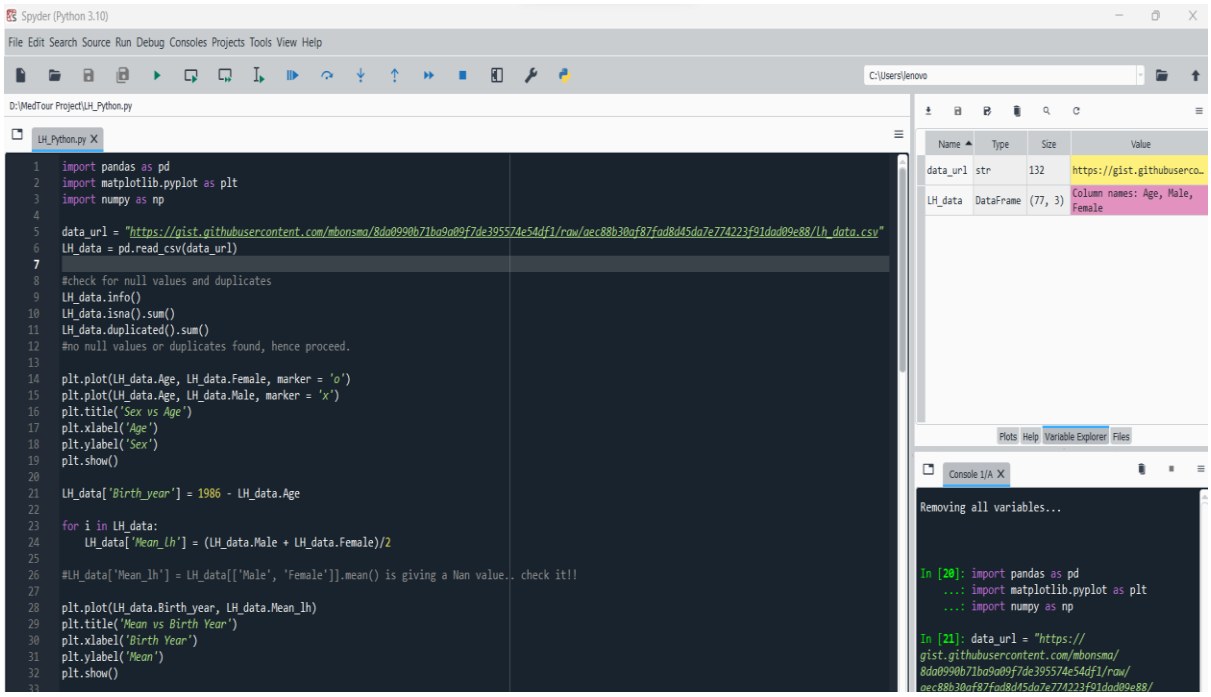
2.2.2) IDE used: Jupyter Notebook, Spyder.

2.2.3) Packages used: pandas, numpy, matplotlib.pyplot.

Pandas: Pandas is an open-source data manipulation and analysis library in python. It provides help with the data structure, loading of data, data visualization etc. It mainly defines two data structures namely DataFrame and series. We may manipulate the data structure used the packages in the library.

NumPy: It is also an open-source library for numerical computation in python. It mainly defines an array, which is a collection of elements. NumPy can compute n-dimensional arrays.

Matplotlib: An open source library with various packages to help with data visualization. Pyplot is one of the packages, that helps in the scatter-plot of two attributes or variables. Pyplot has various other uses too, like histograms, bar-chart, box plot etc.



3) IMPLEMENTATION

3.1 Introduction and Literature review:

This is the first step of this project. The requirements were provided by the and client and problem stated. We did a background check for the same and carried out literature review to understand how exactly the trend of prediction had been.

3.2 Data Collection and Importing:

Data collection in a systematic approach for gathering and measuring information from a variety of sources in order to obtain a complete and accurate picture of an interest area. It helps individual or organization to address specific questions, determine outcomes and foresee patterns.

In this project, data was a secondary data. It was collected from various GitHub repositories. We have used two data sets to complete analysis.

- Left-Handedness data that was collected in a survey of 1986 (US)
- Death data that was collected by a survey. (US)

Data was imported using python and manipulated, aggregated, filtered using several packages provided by python.

Package used: pandas

Function used: read_csv()

The above function creates a dataframe and designs a structure for the databases. We can further understand the data by describing the data.

3.3 Data Cleaning:

This is the most important steps in data analytics. Data that we collect may have a lot of issues and hence cleaning it reduces bias and errors that we may face in the future. Firstly, we need structured data. If the data is unstructured,

like text data or image data, we need to convert this into structured data. The data provided was a structured data.

3.3.1) Checking for duplicates and null values:

Both data sets had no duplicates. The left-hand dataset did not have any null values either, as shown in the figure.

```
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Age     77 non-null     int64
 1   Male    77 non-null     float64
 2   Female  77 non-null     float64
dtypes: float64(2), int64(1)
memory usage: 1.9 KB

In [12]: LH_data.isna().sum()
Out[12]:
Age       0
Male      0
Female    0
dtype: int64
```

However, the death data had some nulls values.

```
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Age         125 non-null    int64
 1   Both Sexes  120 non-null    float64
 2   Male        115 non-null    float64
 3   Female      120 non-null    float64
dtypes: float64(3), int64(1)
memory usage: 4.0 KB

In [16]: DD_data.isna().sum()
Out[16]:
Age            0
Both Sexes     5
Male          10
Female         5
dtype: int64
```

3.3.2) Dealing with NaN values:

As we can see the columns have null values, however the number of null valies are small when compared to the total number of data points. Hence, we may remove the columns that have null values. We only focus on Age and Both sexes column, hence we can remove the null values from both sexes.

3.4 Exploratory Data Analysis and Statistical Inference:

3.4.1) Understanding the data:

To understand what the data has to offer and its type, we need to describe the data as follows.

Code used: datastructure_name.describe()

```
In [25]: LH_data.describe()
Out[25]:
              Age       Male     Female
count   77.000000  77.000000  77.000000
mean    48.000000  11.549122   9.120522
std     22.371857   3.138912   2.505480
min     10.000000   4.885999   4.680948
25%     29.000000   8.562868   6.604398
50%     48.000000  12.717558   9.892073
75%     67.000000  14.209302  11.368353
max     86.000000  15.546784  12.872166
```

```
In [24]: DD_data.describe()
Out[24]:
              Age   Both Sexes         Male       Female
count  125.000000   120.000000   115.000000   120.000000
mean    62.000000  19925.358333  10218.982609  10132.166667
std     36.228442  23523.773527  11183.006921  13081.181052
min      0.000000      1.000000      1.000000      1.000000
25%     31.000000   1733.250000   1067.000000   816.250000
50%     62.000000   8933.500000   5401.000000   3708.500000
75%     93.000000  31112.500000  15576.000000  14745.500000
max    124.000000  73039.000000  36020.000000  42335.000000
```

3.4.2) Trend in data:

To understand what the data has to offer, we need to understand the trend in the data. To do so, we plot the data points in a scatter plot. We observe the trend and then proceed to calculate the year of birth and mean of lefthanded people alive. We then select a study year as reference.

3.4.3) Probabilistic Approach:

We have to understand how Left-handedness and Age of death was correlated. To do so, we calculate conditional probability.

Bayes' Theorem: Since probability of dying at a certain age given left-handed is not same as probability of being left-handed, we need to use Bayes' theorem.

Bayes' theorem states the following:

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

where P(A|LH) = Probability of dying at age 'A' given you are left-handed,

P(A) = Probability of dying at age A,

P(LH) = Probability of being left-handed,

P(LH | A) = Probability of being left-handed, given that you die at age A.

We can't directly find P(LH|A) as they might lie beyond out age ranges, hence, we need to extrapolate the graph.

Similarly we calculate the above for right-handedness too.

3.5 Data Visualization:

Once we calculate all the probabilities, we plot them with ages and other factors to check for trend. Also, we look at the distribution of the conditional probabilities. We can conclude various results from the above distributions.

Language used: Python

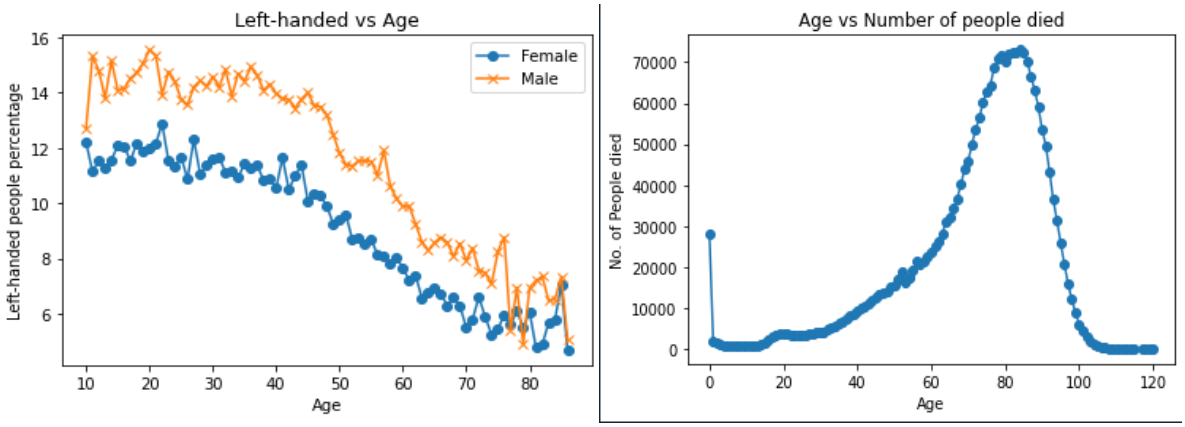
Tools used: pyplot from matplotlib.

Code used:

```
plt.plot(A, prob_left, label = 'Left-handedness')
plt.plot(A, prob_right, label = 'Right-handedness')
plt.title('Age of death vs Probability of being at Age')
plt.xlabel('Age of death')
plt.ylabel('Probability of being at Age')
plt.legend()
plt.show()
```

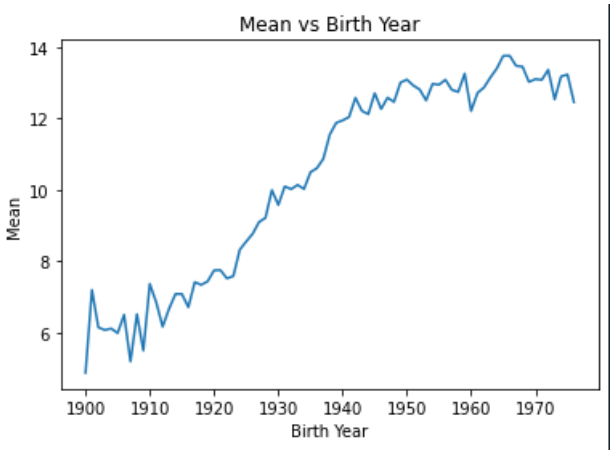
4) OBSERVATIONS

Firstly, we plot percentage of left-handed people vs age. It can be observed that this percentage is far higher in the age group below 40. It can be noted that only 3% of people of the age group 70 above are left-handed. There is an increase to about 12% in the later years. Hence, it indicates that we may find more right-handed people as we move towards older ages.



Also, when the distribution of people died at a given age is drawn, we see a curve as above.

To be clear with the claims, we have calculated the birth year and mean of left-handed people. You can clearly observe that the percentage of people with left-handedness rapidly increases over the years.

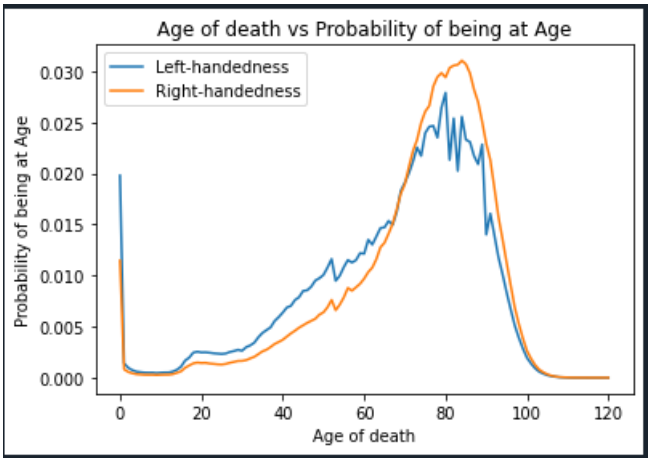


Then we have calculated the conditional probabilities for left-handed and right-handed data as shown below in the figure. Firstly, we chose the study year ti be 1990 to see the trend during that time.

```
In [31]: prob_left.head()
Out[31]:
0    0.019783
1    0.001408
2    0.000974
3    0.000741
4    0.000593
Name: Both Sexes, dtype: float64
```

```
In [32]: prob_right.head()
Out[32]:
0    0.011460
1    0.000816
2    0.000564
3    0.000429
4    0.000344
Name: Both Sexes, dtype: float64
```

We then have plotted the probability vs age to further analyse the trend.



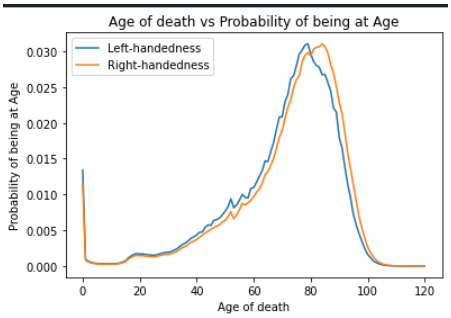
We can observe that the peak for left-handed people is lower than the right-handed people. We need to compare our results with the original study that found out left-handed people were nine years younger at death on average. Now, to do so, we calculate the average ages of left-handed and right-handed people.

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

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

The difference in their average ages = ***5.5*** years.

We can re-check this result by using study year sometime later. If we choose the study year as 2018, the result of probability vs age of death is as follows. The graphs of left-handed and right-handed are very similar and almost coinciding.



The difference in their average ages has decreased and it is = ***2.3*** years.

5) CONCLUSION

The project aimed at understanding the correlation between age of death and left-handedness while comparing it with right-handedness, and examining the trend over the years. Our study leads to the conclusion that the earlier observed trend of early age of death for people with left-handedness is not inherently linked to being left-handed but is influenced by the lower number of older people with left-handedness.

This conclusion arises from the comparison of the average number of alive left-handed individuals over the years. The recent years show an increase in the number of left-handed individuals, leading to a decrease in the difference between the averages of right-handed and left-handed individuals.

As a result, the reported rates of left-handedness have risen from just 3% in the early 1900s to about 11% today. This increase suggests that older people are more likely to be reported as right-handed than left-handed. Consequently, examining a sample of recently deceased individuals will yield more old right-handers than left-handers.

6) Reference

Data collection and literature review:

1) *Left-handedness and mortality*: Marcel E. Salive, Jack M. Guralnik, and Robert J. Glynn. Published in: Public Health Briefs.
2) *Handedness and Life span*: Sharon Kuritzky. Published in: The New England Journal of Medicine.
3) (PDF) Left-Handedness: A marker for decreased survival fitness (researchgate.net)

Data collection and code reference:

1) https://github.com/rrmolin/Do-Left-handed-People-Really-Die-Young-DataCamp-project/blob/master/notebook.ipynb