TRAINEESHIP PROGRAM REPORT

ON

ANALYZING DEATH AGE DIFFERENCE BETWEEN RIGHT HANDERS AND LEFT HANDERS

MEDTOUREASY

21st June 2023

ACKNOWLDEGMENT

My appreciation goes to MedTourEasy for giving me this great traineeship opportunity to independently work on a real-life project. This project gave me more insights and better understanding of Data Visualizations in Data Analytics; and developed me personally and professionally. I am very obliged for having a chance to interact with so many professionals who guided me throughout the traineeship project and made it a great learning curve for me.

I also sincerely acknowledge the Training & Development Team of MedTourEasy team, they made my traineeship journey easy by always attending to me when I reach out to them for certain questions.

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

ABSTRACT

A National Geographic survey in 1986 resulted in over a million responses that included age, sex, and hand preference for throwing and writing. Researchers Avery Gilbert and Charles Wysocki analyzed this data and noticed that rates of left-handedness were around 13% for people younger than 40 but decreased with age to about 5% by the age of 80.

They concluded based on analysis of a subgroup of people who throw left-handed but write right-handed that this age-dependence was primarily due to changing social acceptability of left-handedness. This means that the rates aren't a factor of *age* specifically but rather of the *year you were born*, and if the same study was done today, we should expect a shifted version of the same distribution as a function of age.

Therefore, this project aims at exploring 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. Ultimately, we'll see what effect this changing rate has on the apparent mean age of death of left-handed people.

CHAPTER 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

This project aims at exploring 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 and observing the effect this changing rate has on the apparent mean age of death of left-handed people.

It uses pandas and Bayesian statistics to analyze the probability of being a certain age at death given that you are reported as left-handed or right-handed.

1.3 AIM OF STUDY

To analyze the death age difference between right handers and left handers.

1.4 OBJECTIVES OF STUDY

To achieve the above set goal, the following objective will be considered:

1. Studying death distribution data in United States from the year 1999 and rates of left handedness.
2. Calculating the probability of dying at age A given that you're left-handed using Bayes’ rule.
3. Getting the overall probability of lefthandedness.
4. Using Bayes' rule to get P(A | LH), the probability of being age A at death (in the study year) given that you're left-handed.
5. Comparing the results with the original study that found that left-handed people were nine years younger at death on average.

CHAPTER 2: MATERIALS AND METHODS

2.1 MATERIALS

2.1.1 Language and Platform Used

In analyzing this project, I used python and jupyter notebook as the integrated development environment which enabled the python to communicate directly with the code. Jupyter notebooks basically provides an interactive means for developing python-based data science applications. It helps to describe the analysis process step by step and can be used for data transformation and cleaning, data visualization, machine learning and much more.

It uses pandas and Bayesian statistics to analyze the probability of being a certain age at death given that you are reported as left-handed or right-handed.

2.1.2 Data Acquisition

Choosing the right content and sources of data for the is necessary for an accurate analysis. Death distribution data in United States from the year 1999 and rates of left handedness was used for this analysis.

2.2 METHODS USED FOR THE ANALYSIS

2.2.1 Explorative Data Analysis

When working with data sets, it is necessary to understand the data so that lots of useful information can be gotten from it. Explorative data analysis is carried out on data to discover anomalies, investigate the data and check assumptions using graphical and statistical representation.

The following steps were carries out during the explorative data analysis;

- The necessary libraries (pandas, numpy, matplotlib) used for data analysis were imported.
- The data set was loaded and read.

```python
# import libraries
import pandas as pd
import matplotlib.pyplot as plt
# load the data
data_url_1 = "https://gist.githubusercontent.com/mbonsma/8da0990b71ba9a09f7de395574e54df1/raw/aec88b30af87fad8d45da7e774223f9
lefthanded_data = pd.read_csv("https://gist.githubusercontent.com/mbonsma/8da0990b71ba9a09f7de395574e54df1/raw/aec88b30af87fa
print("lefthanded_data shape:", lefthanded_data.shape)
lefthanded_data.head()
```

Fig 2.1: Importing libraries and loading datasets.

- After loading, the basic parameters of the data set were explored using summary method by examining the head (first few rows) of the pandas data frame to gain an idea of the content.
- The statistical summary of the data set was gotten and checked for null values. The missing value was eliminated to get equal number of points in the column.

```
death_distribution_data shape: (120, 4)
```

Out[5]:

	Age	Both Sexes	Male	Female
0	0	27937.0	15646.0	12291.0
1	1	1989.0	1103.0	886.0
2	2	1376.0	797.0	579.0
3	3	1046.0	601.0	445.0
4	4	838.0	474.0	364.0

```
lefthanded_data shape: (77, 3)
```

Out[1]:

	Age	Male	Female
0	10	12.717558	12.198041
1	11	15.318830	11.144804
2	12	14.808281	11.549240
3	13	13.793744	11.276442
4	14	15.156304	11.572906

Fig 2.2: Sample of death distribution data and lefthanded data

2.2.2 Data Visualization

This is the graphical representation of data with the use of visualization tools to be able to identify the trend and patterns in the data set.

Fig 3.2 shows a plot of male and female column vs. Age from the Left-handed data.

Fig 3.3, a column called Mean_lh was created in the lefthanded data which is equal to the mean of the mail and female columns and was plotted against the birth year.

Fig 3.4 is a graphical representation of the number of people who died as a function of their age in the death distribution data.

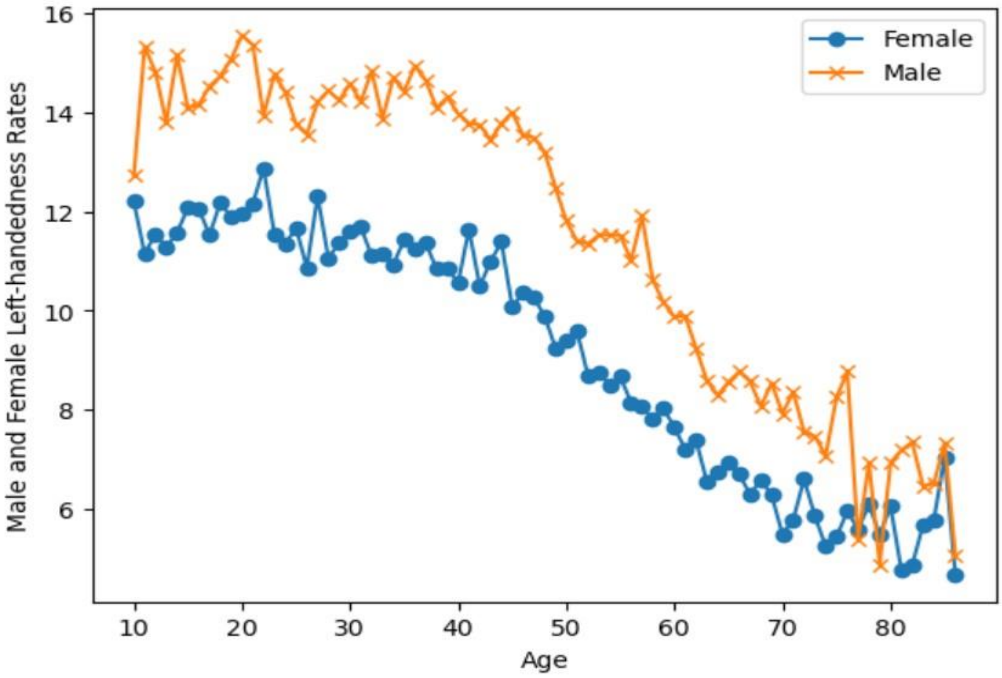

Fig 2.3: Male & Female Left-handedness Rates vs. Age

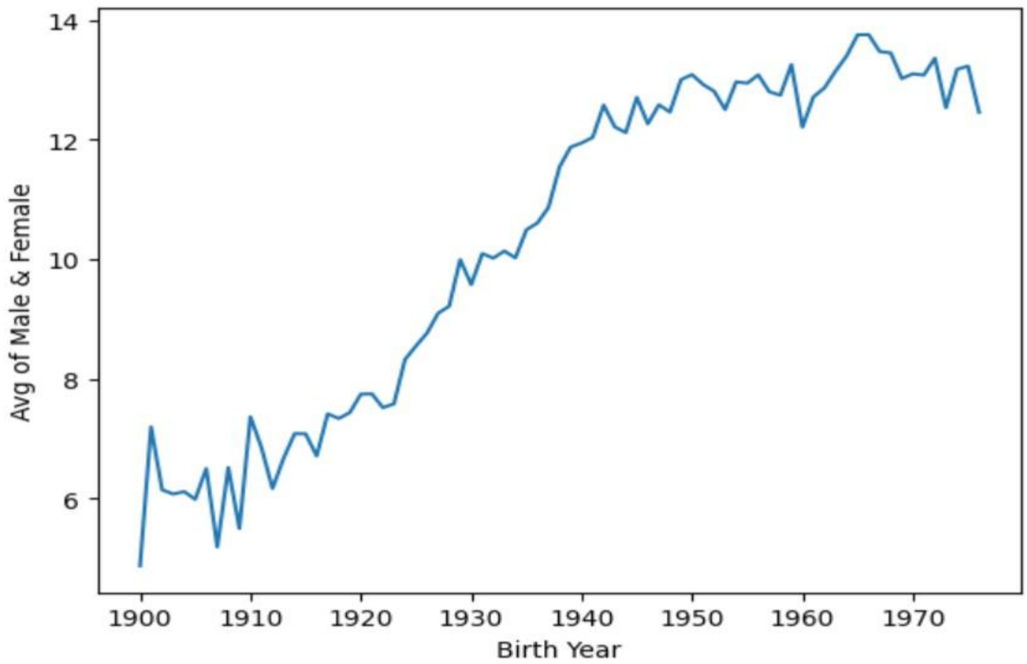

Fig 2.4: Mean_lh vs. Birth Year

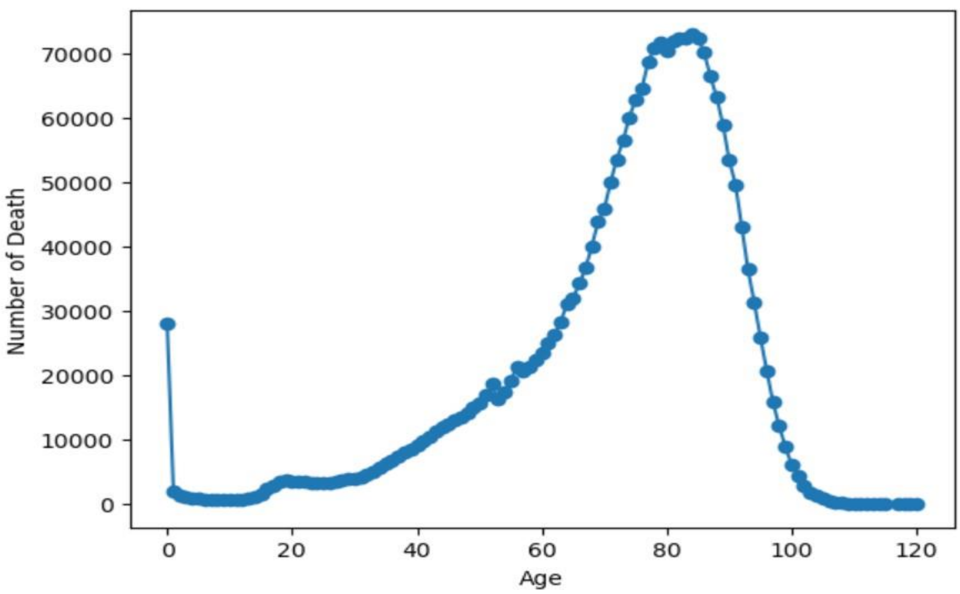

Fig 2.5: Number of Death vs. Age

2.2.3 Applying Bayesian Statistics

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

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

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

P(LH | A) is the probability that you are left-handed given that you died at age A. P(A) is the overall probability of dying at age A, and P(LH) is the overall probability of being left-handed.

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

```python
# 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 = lefthanded_data["Mean_lh"].iloc[67:77]
    early_1900s_rate = early_1900s.mean()
    late_1900s = lefthanded_data["Mean_lh"].iloc[0:10]
    late_1900s_rate = late_1900s.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
```

Fig 2.6: Creating a function that will return P(LH | A) for particular ages of death in a given study year.

P(LH) is the probability that a person who died in our study year is left-handed, assuming we know nothing else about them. This is the average left-handedness in the population of deceased people, and it was calculated by summing up all the left-handedness probabilities for each age, weighted with the number of deceased people at each age, then divided by the total number of deceased people to get a probability. In equation form, this is what was calculated, where N(A) is the number of people who died at age A (given by the dataframe death_distribution_data):

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

```
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 = pd.Series(death_distribution_data["Both Sexes"] * P_lh_given_A(death_distribution_data["Age"], 1990)) # multiply
    p = sum(p_list) # calculate the sum of p_list
    p_returns = p / sum(death_distribution_data["Both Sexes"])
    return p_returns # normalize to total number of people (sum of death_distribution_data['Both Sexes'])

print(P_lh(death_distribution_data))
```

```
0.07766387615350637
```

Fig 2.7: Creating a function for the overall probability of Left-handedness.

The three was combined using Bayes' rule to get P(A | LH), the probability of being age A at death (in the study year) given that you're left-handed. To make this answer meaningful, though, we also want to compare it to P(A | RH), the probability of being age A at death given that you're right-handed.

The following quantity was calculated twice, once for left-handers and once for right-handers.

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

```
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"] / sum(death_distribution_data["Both Sexes"])
    P_left = P_lh(death_distribution_data, study_year = 1990) # use P_lh function to get probability of left-handedness overa
    P_lh_A = P_lh_given_A(death_distribution_data["Age"], study_year = 1990) # use P_lh_given_A to get probability of left-ha
    return P_lh_A*P_A/P_left
```

Fig 2.8: Creating a function to calculate P(A\LH)

```
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"] / sum(death_distribution_data["Both Sexes"])
    P_right = 1 - (P_lh(death_distribution_data, study_year = 1990)) # either you're left-handed or right-handed, so P_right
    P_rh_A = 1 - (P_lh_given_A(death_distribution_data["Age"], study_year = 1990)) # P_rh_A = 1 - P_lh_A
    return P_rh_A*P_A/P_right
```

Fig 2.9: Creating a function to calculate P(A\RH)

CHAPTER 3: RESULT AND DISCUSSION

3.1 Results

Result of the probability of being a certain age at death given that you're left- or right-handed for a range of ages.

```python
ages = np.arange(6, 126, 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, study_year = 1990)
right_handed_probability = P_A_given_rh(ages, death_distribution_data, study_year = 1990)

# 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");
```

Fig 3.1: Calculation of P_A_given_lh and P_A_given_rh

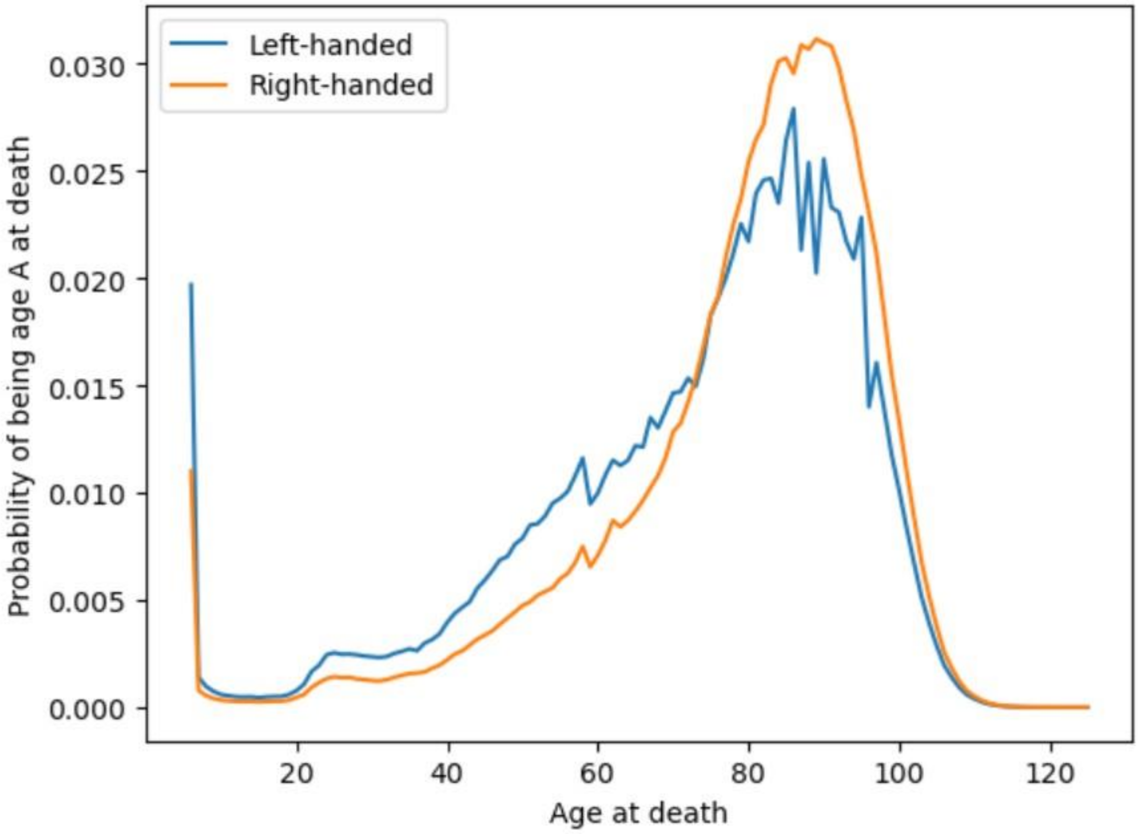

Fig 3.2: Probability of being age A at death vs. Age at death.

To compare the results with the original study that found that left-handed people were nine years younger at death on average. This was don by calculating the mean of these probability distributions in the same way P(LH) was calculated earlier, weighting the probability distribution by age and summing over the result.

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

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

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

# print the average ages for each group
print("average_lh_age:", average_lh_age)
print("average_rh_age:", 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.")
```

```
average_lh_age: 73.25602005762002
average_rh_age: 78.79814529889603
The difference in average ages is 5.5 years.
```

Fig 3.3: Difference in avg age of right-handed and left-handed people at death (study year: 1990)

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

```
The difference in average ages is 5.5 years.
```

Fig 3.3: Difference in avg age of right-handed and left-handed people at death (study year: 2018)

3.2 Result Discussion

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

CHAPTER 4: CONCLUSION AND RECOMMENDATION

The 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:

1. Death distribution data from almost ten years was used 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).
2. The left-handedness survey results were extrapolated to older and younger age groups, but it's possible our extrapolation wasn't close enough to the true rates for those ages.

The gap in 2018 study year turns out to be much smaller since rates of left-handedness haven't increased for people born after about 1960. Both the National Geographic study and the 1990 study happened at a unique time - the rates of left-handedness had been changing across the lifetimes of most people alive, and the difference in handedness between old and young was at its most striking.

A good recommendation would be to figure out how much variability we would expect to encounter in the age difference purely because of random sampling: if you take a smaller sample of recently deceased people and assign handedness with the probabilities of the survey, what does that distribution look like? How often would we encounter an age gap of nine years using the same data and assumptions? It was not done in this project, but it's possible with this data and the tools of random sampling.

REFERENCES

Data Collection

Deaths By Single Years of Age, Race, and Sex: United States, 1999 (cdc.gov)

Hand preference and age in the United States - PubMed (nih.gov)

Programming References

Programming | DataScience+ (datascienceplus.com)

Data Analysis and Visualization with Jupyter Notebook | by Veronika Rovnik | Towards Data Science

Data Analysis and Visualization with pandas and Jupyter Notebook in Python 3 | DigitalOcean