Final Report of Traineeship Program 2023

On

"Analyze Death Age Difference of Right Handers with Left Handers"

MEDTOUREASY

28th September 2023

ACKNOWLDEGMENTS

The traineeship opportunity that I had with MedTourEasy was a great change for learning and understanding the intricacies of the subject of Data Visualizations in Data Analytics; and also, for personal as well as professional development. 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.

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 also for spearing his valuable time in spite of his busy schedule.

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

TABLE OF CONTENTS

ABSTRACT

According to a National Geographic survey conducted in 1986, researchers Avery Gilbert and Charles Wysocki analyzed over a million responses that included age, sex, and hand preference for throwing and writing. They found that the rates of left-handedness were around 13% for people younger than 40 but decreased with age to about 5% by the age of 80. The researchers concluded that this age-dependence was primarily due to changing social acceptability of left-handedness, rather than age specifically. 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.

To investigate the effect of changing rates of left-handedness on the apparent mean age of death of left-handed people, it is important to plot the rates of left-handedness as a function of age. This will help us understand how the rate of left-handedness has changed over time and how it affects the apparent mean age of death of left-handed people.

1. INTRODUCTION

1.1 About the Company

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

1.2 About the Project

The project that aims to investigate the effect of changing rates of left-handedness on the apparent mean age of death of left-handed people. To achieve this, you plan to plot the rates of left-handedness as a function of age.

The National Geographic survey conducted in 1986 by researchers Avery Gilbert and Charles Wysocki analyzed over a million responses that included age, sex, and hand preference for throwing and writing. They found that the rates of left-handedness were around 13% for people younger than 40 but decreased with age to about 5% by the age of 80. The researchers concluded that this age-dependence was primarily due to changing social acceptability of left-handedness, rather than age specifically. 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.

To investigate the effect of changing rates of left-handedness on the apparent mean age of death of left-handed people, it is important to plot the rates of left-handedness as a function of age. This will help us understand how the rate of left-handedness has changed over time and how it affects the apparent mean age of death of left-handed people.

- ***Analysis of the problem*:** This is done to analyses age-dependence was primarily due to changing social acceptability of left-handedness, rather than age specifically. 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.

- ***Analysis of the steps for investigation***: By plotting the rates of left-handedness as a function of age, we can understand how the rate has changed over time and how it affects the apparent mean age of death of left-handed people. This analysis will help us understand how changing social attitudes towards left-handedness have affected mortality rates among left-handed individuals.

1.3 Objectives and Deliverables

Objectives

1. Explore the phenomenon of changing rates of left-handedness over time.
2. Investigate the impact of changing rates of left-handedness on the average age at death for left-handed individuals.
3. Reproduce the difference in average age at death purely from the changing rates of left-handedness over time.
4. Challenge the claim that left-handers die earlier by analyzing the probability of being a certain age at death given left-handedness or right-handedness.
5. Utilize **pandas** and Bayesian statistics to analyze the relationship between left-handedness, age at death, and changing rates of left-handedness.
6. Refute the assumption that the difference in average age at death between left-handed and right-handed individuals is solely due to left-handedness itself, rather than changing rates of left-handedness over time.

Deliverables

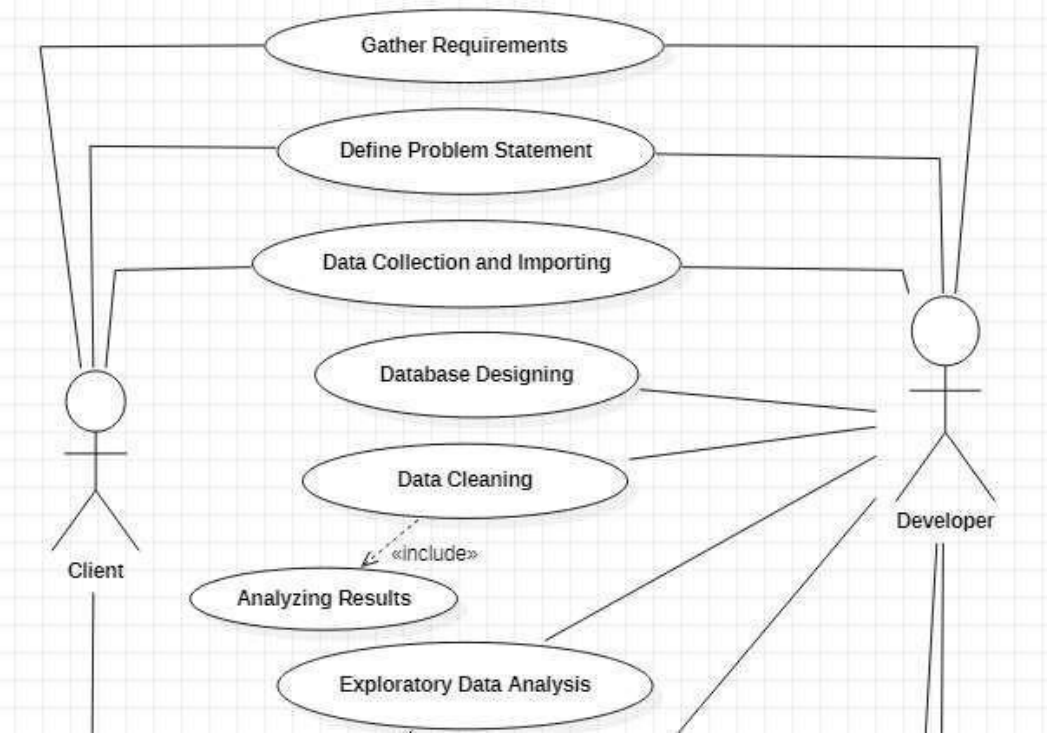
1. **Exploration of changing rates of left-handedness over time**: The study aims to investigate the phenomenon of changing rates of left-handedness and understand how these rates have evolved over time.
2. **Analysis of the impact on average age at death**: The study seeks to analyze the impact of changing rates of left-handedness on the average age at death for left-handed individuals. By examining age distribution data, the researchers aim to determine whether the previously reported difference in average age at death can be attributed solely to changing rates of left-handedness.
3. **Reproduction of the difference in average age at death**: The study aims to reproduce the difference in average age at death purely from the changing rates of left-handedness over time. By utilizing Bayesian statistics and age distribution data, the researchers aim to challenge the claim that left-handers die earlier.
4. **Refutation of the claim of early death for left-handers**: The study seeks to refute the claim that left-handers have a shorter lifespan. By analyzing the probability of being a certain age at death given left-handedness or right handedness, the researchers aim to provide evidence against the assumption of early death for left-handers.
5. **Utilization of pandas and Bayesian statistics**: The study employs the use of **pandas**, a powerful data analysis library in Python, and Bayesian statistics to analyze the relationship between left-handedness, age at death, and changing rates of left-handedness. These tools enable the researchers to perform comprehensive statistical analyses and draw meaningful conclusions.
6. **To** provide insights into the relationship between changing rates of left-handedness, average age at death, and the claim of early death for left-handers.

2. 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.

2.2 Use Case Diagram



Above figure shows the use case of the project. There are two main actors in the same: The Client and Developer. The developer will first gather requirements and define the problemstatement then collecting the required data and importing it. Then the developer will designdatabases so as to identify various constraints and relations in the data. Next step is to clean the data to remove irregular values, blank values etc. Next, exploratory data analysis is conducted to filter the data according to the requirements of the project.

2.3 Language and Platform Used

2.2.1 Language: PYTHON

Python: Python is widely used for data analysis due to its simplicity, extensive libraries, and strong community support. Libraries such as Pandas, NumPy, and SciPy provide powerful tools for data manipulation, analysis, and visualization.
Platforms and Tools:
Google Colab: Google Colab is a cloud-based platform that provides free access to Jupyter notebooks with GPU support. It's a convenient option for running Python code on powerful hardware without the need for local installation

2.2.2 IDE: VISUAL STUDIO CODE

Visual Studio Code (VSC) is a popular integrated development environment (IDE) that combines the simplicity of a source code editor with powerful developer tooling. It supports various programming languages and provides features such as IntelliSense code completion, debugging, and syntax highlighting. VSC is available for macOS, Linux, and Windows, making it accessible across different platforms. It offers a lightning-fast source code editor with support for hundreds of languages, allowing developers to be instantly productive. The IDE also includes built-in support for Git, making it easy to work with source control without leaving the editor. VSC can be customized to suit individual preferences and extended with third-party extensions. It is free to download and use, making it an attractive choice for developers. Overall, VSC provides a robust and extensible architecture, making it a versatile IDE for various development tasks.

NumPy: It is an open-source Python library that facilitates efficient numerical operations on large quantities of data. It provides a powerful N-dimensional array object, along with a large collection of mathematical functions to operate on these arrays. NumPy is widely used in scientific computing, data analysis, and machine learning.

Pandas is another popular open-source Python library that is built on top of NumPy. It provides easy-to-use data structures and data analysis tools for handling structured data. The two primary data structures in Pandas are Series (one-dimensional) and DataFrame (two-dimensional). Pandas provides a wide range of functions for data manipulation, cleaning, and analysis, including merging, grouping, filtering, and reshaping data. It is widely used in data science, finance, and social science.

Matplotlib: Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. It is a low-level graph plotting library that serves as a visualization utility. Matplotlib was introduced by John D. Hunter and is open source, allowing free usage. The library is mostly written in Python, with a few segments written in C, Objective-C, and JavaScript to ensure platform compatibility.

Matplotlib is designed to work with the broader SciPy stack and is built on NumPy arrays. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. The library offers a wide variety of plots, including line plots, bar plots, scatter plots, histograms, and more[1]. These plots help understand trends, patterns, and correlations in quantitative information.

3. IMPLEMENTATION

3.1 Gathering Requirements and Defining Problem Statement

This is the first step wherein the requirements are collected from the clients to understand the deliverables and goals to be achieved after which a problem statement is defined which has to be adhered to while development of the project.

3.2 Data Collection and Importing

Data collection is 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 an individual or organization to address specific questions, determine outcomes and forecast future probabilities and patterns.

The data has been collected through various GitHub repositories,mentioned as follows:

https://www.cdc.gov/nchs/data/statab/vs00199_table310.pdf" death distribution data for the United States from the year 1999 (source website.

https://www.cdc.gov/nchs/nvss/mortality_tables.html
https://www.ncbi.nlm.nih.gov/pubmed/1528408 paper by Gilbert and Wysocki.

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

Packages Used: Pandas

Pandas is a fast, powerful, flexible, and easy-to-use open-source data analysis and manipulation tool built on top of the Python programming language1. It provides fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive. The goal of pandas is to provide a fast and friendly way to read rectangular data (like csv, tsv, and fwf). It is designed to flexibly parse many types of data found in the wild, while still cleanly failing when data unexpectedly changes.

Functions Used:

read.csv (): It is a wrapper function for read.table() that mandates a comma as separator and uses the input file's first line as header that specifies the table's columnnames. Thus, it is an ideal candidate to read CSV files. It has an additionalparameter of url() which is used to pull live data directly from GitHub repository.

```
# import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline

# load the data
data_url_1 =
"https://gist.githubusercontent.com/mbonsma/8da0990b71ba9a09f7de395574e54
df1/raw/aec88b30af87fad8d45da7e774223f91dad09e88/lh_data.csv"
lefthanded_data = pd.read_csv(data_url_1)
```

3.3 Designing Databases

Once the data has been collected and imported into the PYTHON environment, it isimportant to design the structure of the database tables so as to identify theconstraints in the data, keys, dependencies and relations between various tables.

Once the data is imported in the environment, it is converted into a data frame (datatype in Python) which makes it easy to maintain the data in form of tables. The various tables which have been created are mentioned as follows:

Attribute	Data type	Size	Extra
Female	FLOAT	5	Not null, Unique
Male	FLOAT	15	Not Null, Unique
Age	INT	3	Primary key

3.4 Data Cleaning

"Quality data beats fancy algorithms"

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

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

Packages Used: **pandas**

Functions Used:

Is.na(): Missing values are represented by the symbol **NA** (not available). Impossible values (e.g., dividing by zero) are represented by the symbol **NaN** (not a number). This function is used to check if a dataset contains NA values or not.

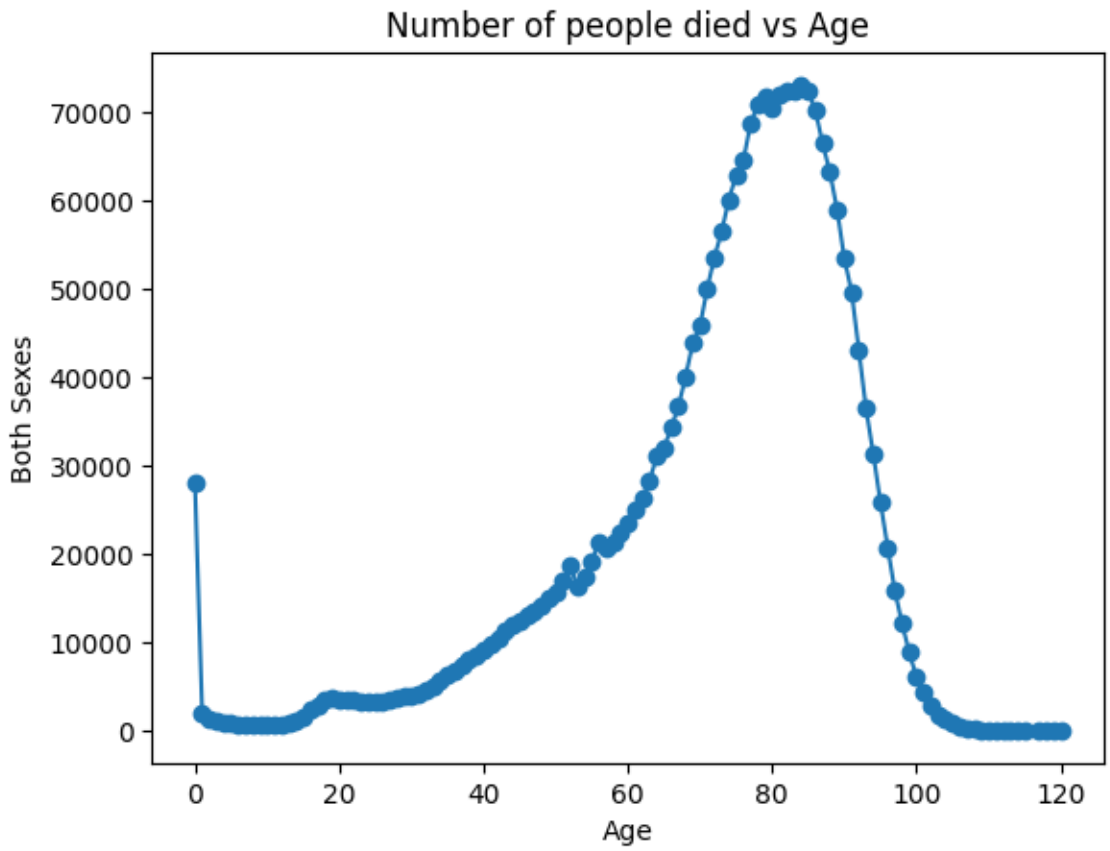
isnull(): When using a data frame function na.rm in r refers to the logical parameterthat tells the function whether or not to remove NA values from the calculation. It literally means NA remove. It is a parameter used by several data frame functions.

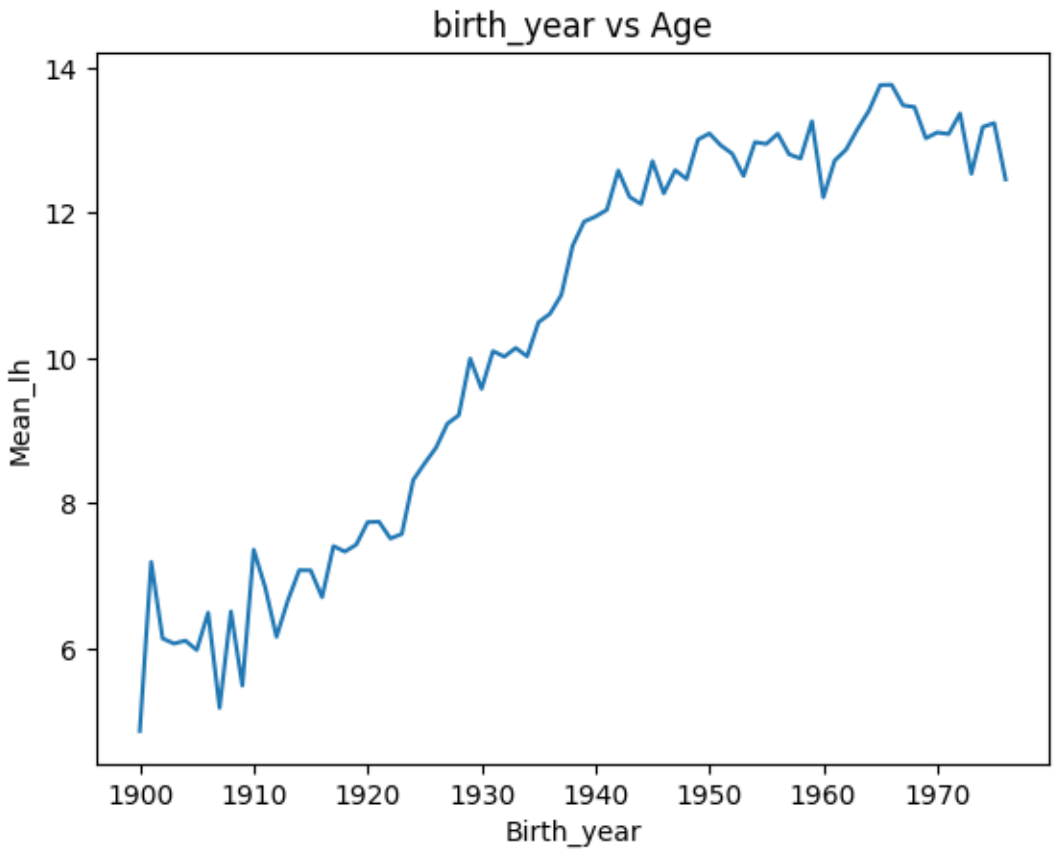
nunique(): This function is used to filter out redundant data and keep only unique values from the data frame.

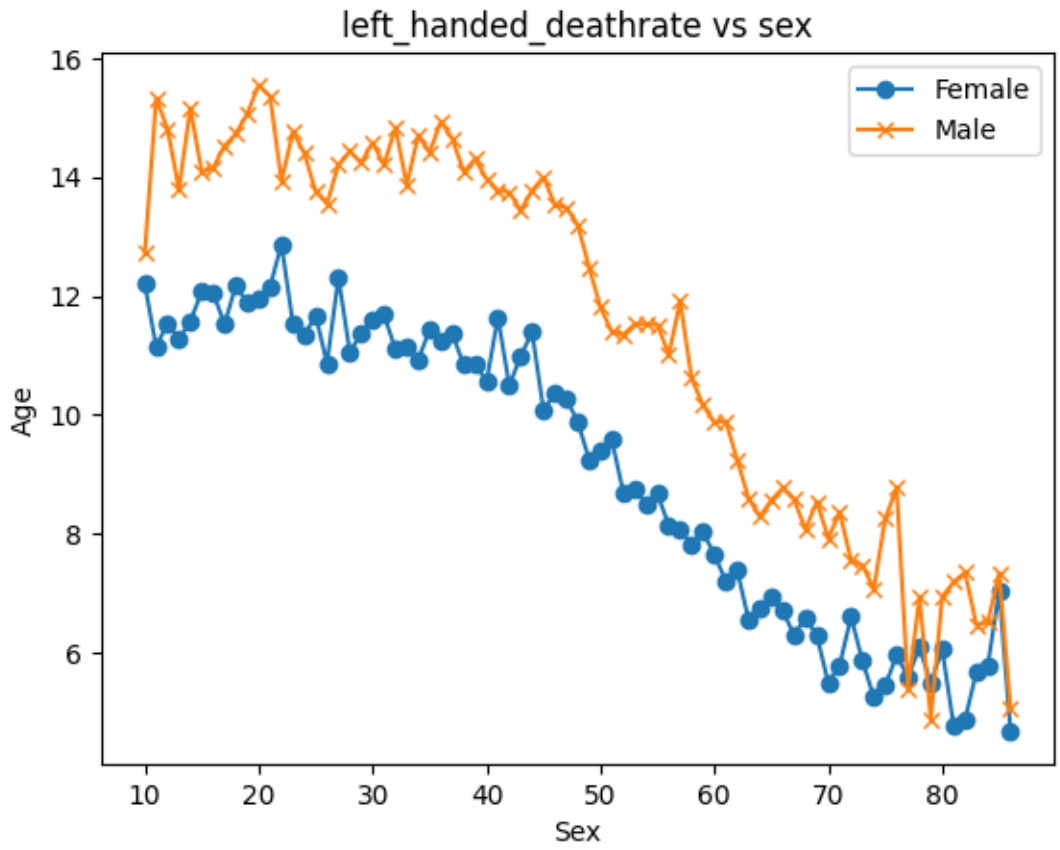
3.5 Data Filtering

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

4. SAMPLE SCREENSHOTS AND OBSERVATIONS




birth_year vs Age
Mean_lh
Birth_year
14
12
10
8
6
1900
1910
1920
1930
1940
1950
1960
1970



Death rate of left handed males are more than females.

5. CONCLUSION

The study concludes that the apparent difference in average age at death between left-handed and right-handed individuals can be attributed to fluctuating rates of left-handedness over time. It highlights the importance of considering historical changes in reported rates of left-handedness when examining life expectancy differences.

6. FUTURE SCOPE

As for future scope, the study suggests that new information on past reporting biases could warrant re-exploration of initial findings. Further research could focus on refining models and exploring other factors that may contribute to differences in life expectancy between left-handed and right-handed individuals.

7. REFERENCES

Data Collection

- https://www.cdc.gov/nchs/data/statab/vs00199_table310.pdf
- https://www.cdc.gov/nchs/nvss/mortality_tables.html
- https://pubmed.ncbi.nlm.nih.gov/1528408/
- https://gist.githubusercontent.com/mbonsma/8da0990b71ba9a09f7de395574e54df1/raw/aec88b30af87fad8d45da7e774223f91dad09e88/lh_data.csv
- https://gist.githubusercontent.com/mbonsma/2f4076aab6820ca1807f4e29f75f18ec/raw/62f3ec07514c7e31f5979beeca86f19991540796/cdc_vs00199_table310.csv

Programming References

- https://pandas.pydata.org/
- https://www.w3schools.com/python/pandas/default.asp
- https://matplotlib.org/stable/gallery/color/named_colors.html
- https://matplotlib.org
- https://www.w3schools.com/python/matplotlib_intro.asp
- https://www.geeksforgeeks.org/python-introduction-matplotlib/
- https://en.wikipedia.org/wiki/Matplotlib
- https://pypi.org/project/matplotlib/