Missing Persons in Pennsylvania and Ohio (1969–2024)

Exploratory Data Analysis of Demographic, Temporal, and Spatial Patterns

Sharvari, Sumaiya, Yuze

2025-12-15

1 Introduction

1.1 Motivation and Context

For our group project, we had decided to look into missing persons data. Certain areas that were especially interesting to us were the potential for any pattern in reported missing persons cases, in addition, any potential relevance to our geographical area.

From these interests, we were motivated to look at missing persons data in Pennsylvania and Ohio, which are very close geographically. For accurate reporting, we also focused on finding data from credible and updated sources.

This report will offer exploratory data analysis of a medium sized data set of reported missing persons data with visualization and inferences/insight on any patterns.

1.2 Research Question

For clarity and focus, we generalized our research question to such:

Do missing persons cases in Pennsylvania and Ohio from 1969 to 2024 exhibit over-representation across specific demographic groups, time spans, and spatial contexts?

2 Data Provenance

We decided to use missing persons data from NamUS (the National Missing and Unidentified Persons System), which has accurate and well-maintained information and is made fairly easily accessible to the public. We trust NamUS because they're a federal, government supported system and have worked directly with other trustworthy entities in law enforcement.

NamUS's data is contributed by law enforcement agencies, medical examiners, and authorized forensic professionals, and is maintained by the NIJ (the National Institute of Justice).

To obtain the data-set, we needed to register for an account for NamUS. Afterwards, a csv can be downloaded from the website after filtering through down to a manageable size.

In our repository, the data will be found under "MissingPersons.csv". The dataframe we worked with in data wrangling and visualization will follow a similar naming scheme "missingPersons".

```
  Case.Number      DLC Legal.Last.Name Legal.First.Name Missing.Age
1      MP2620   8/8/04            Vega          Antonio    73 Years
2      MP2905  6/27/97         Fawcett          Michele    35 Years
3      MP2508 12/12/71           Lande        Elizabeth    21 Years
4      MP2477  9/27/74           Jones        Katherine    29 Years
5      MP2659  3/18/65            Shea         Kathleen     6 Years
6      MP2509  6/24/75          Thorne             Edna    15 Years
          City       County State Biological.Sex
1 Philadelphia        Adams    PA           Male
2      Donegal     Somerset    PA         Female
3 Philadelphia Philadelphia    PA         Female
4   Clearfield   Clearfield    PA         Female
5       Tyrone        Blair    PA         Female
6 Philadelphia Philadelphia    PA         Female
                      Race...Ethnicity Date.Modified
1 White / Caucasian, Hispanic / Latino       8/12/20
2                    White / Caucasian        4/1/20
3                    White / Caucasian       4/27/21
4                    White / Caucasian       9/28/23
5                    White / Caucasian       4/29/24
6                    White / Caucasian       2/21/23
```

As seen from the 6 example entries displayed above, NamUS offered 11 different attribute underneath each case (where each case is an individual reported case of missing persons). These attributes contain: case number, date of last contact, legal last name, legal first name, missing age, city, county, state, biological sex, race/ethnicity, and the date modifed.

3 Ethical frameworks: FAIR and CARE

3.1 Fair Principles

- Findable:

This project clearly states the source of the data (NamUS) and correctly cites the source in order to increase findability for the viewers of this report.

- Accessible:

Our data offered by NamUS was downloaded after we have met the terms of NamUS (registering for an account). The only data we will be using are ones that are publically accessible on the NamUS website that can be visited by anyone.

- Interoperable:

The developers of this project are able to operate with the data using the csv file located in the repository. The data is then moved to a data frame stored in Rdata and cleaned into different data frames for the convience of EDA.

- Reusable:

The coding process of this project has been meaningfully commented and documented for the goal of easy comprehension for viewers and reader of this project.

3.2 Care Principles

- Collective Benefit:

The purpose of this project is for the collective benefit of the public. Designed to generate and offer visualizations that may strength public understanding of missing persons cases from the perspective of demographic, over time, and geographically.

- Authority to Control:

The data comes from proper law enforcement and other proper sources, which obtained consent and permission from related individuals of the missing persons to register the case and enter the information into the database. This analysis also acknowledges that all data were premitted by the right authorities and respects these authorities.

- Responsibility:

All responsibilities of this project falls upon the 3 developers/authors. Responsibilities include any misconduct or misrepresentation of the data and the corresponding analysis.

- Ethics:

Interpretation and presentation of results consider contextual factors like population growth and demographic disparities. Throughout the process of creating this project, developers try to avoiding deterministic or causal claims where evidence is insufficient and ensuring that findings are framed in ways that do not reinforce harmful stereotypes.

4 Data and Attributes

The analysis uses nearly all attributes of the data, with the except of case number (which is simply just used as an identifier), and date modified (where we could not find a useful interruptation related to our research question),

4.1 Original attributes (from dataset):

Geogrpahic attributes:

- state
- county
- city

Demographic attributes:

- biological sex
- race/ ethnicity
- missing age

Time-span attributes:

- DLC

4.2 Derived attributes

From the original listed attributes of the database, there were a couple of attributes that we need to derive for further analysis. In order to adequately categorize the missing cases by the range of the missing ages, we had to create a new attribute representing the age group of the missing age (categorized as a demographic attribute). The ranges we have decided upon are:

- below 20
- between 20-40
- between 40-60
- above 60

5 Data Wrangling

6 Geographic Breakdown

7 Demographic Breakdown

8 Time-span Breakdown

For time-series analysis, we had an issue of not being able to graph the DLC as a variable on a graph due it's data type being just strings of characters. Fortunately, R has a built in Date data type that can extract date information from a given format of strings.

Unfortunately, there were issues that we ran into due to incomplete information. Dates were formatted like so in the original data frame: dd/mm/yy

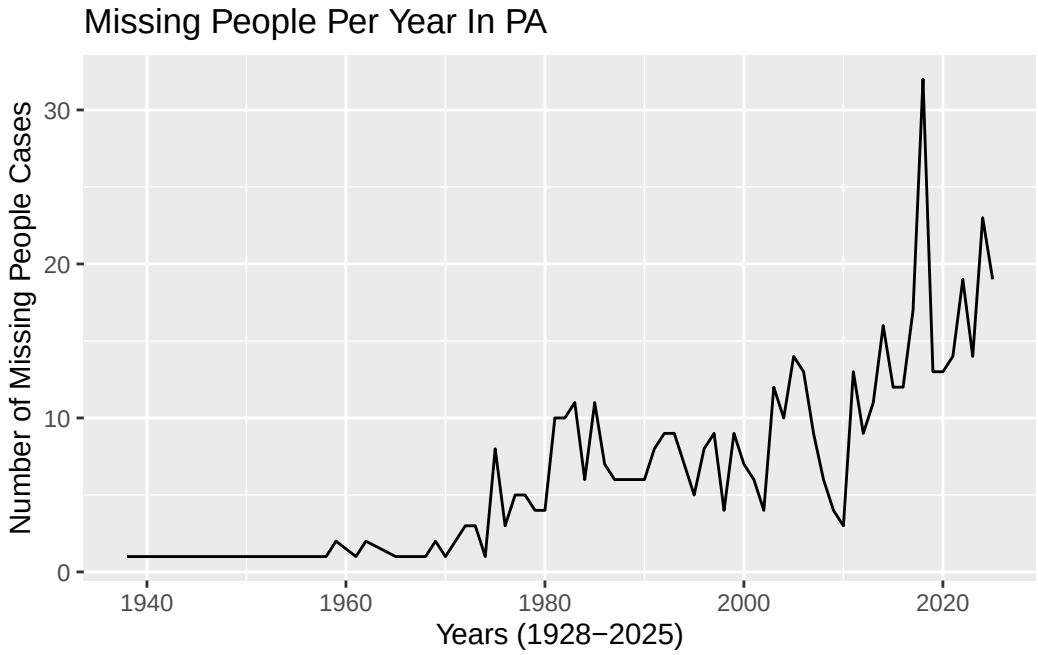
Since the year has section has incomplete information, a year like 1960 is written as /60, which the as.Date function in R recognized as 2060. This problem is then sorted out by updating all years past the current year (2025) to its intended year.

Otherwise, the final dataframe should compose of information with the year missing as a date data type and the count of missing persons in each recorded year. Therefore, each case in the dataframe time_sensitive_missing is a recorded year in the data.

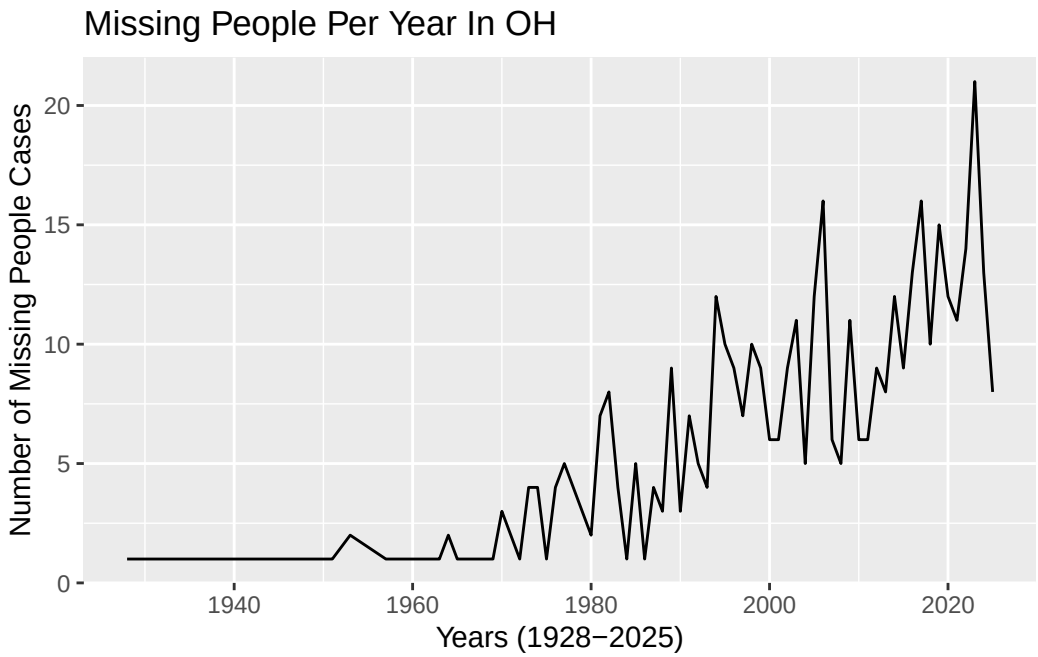
```
# A tibble: 6 x 3
  year       State missing_count
  <date>     <chr>         <int>
1 1969-01-01 OH                1
```

```
2 1969-01-01 PA                2
3 1970-01-01 OH                3
4 1970-01-01 PA                1
5 1971-01-01 PA                2
6 1972-01-01 OH                1
```

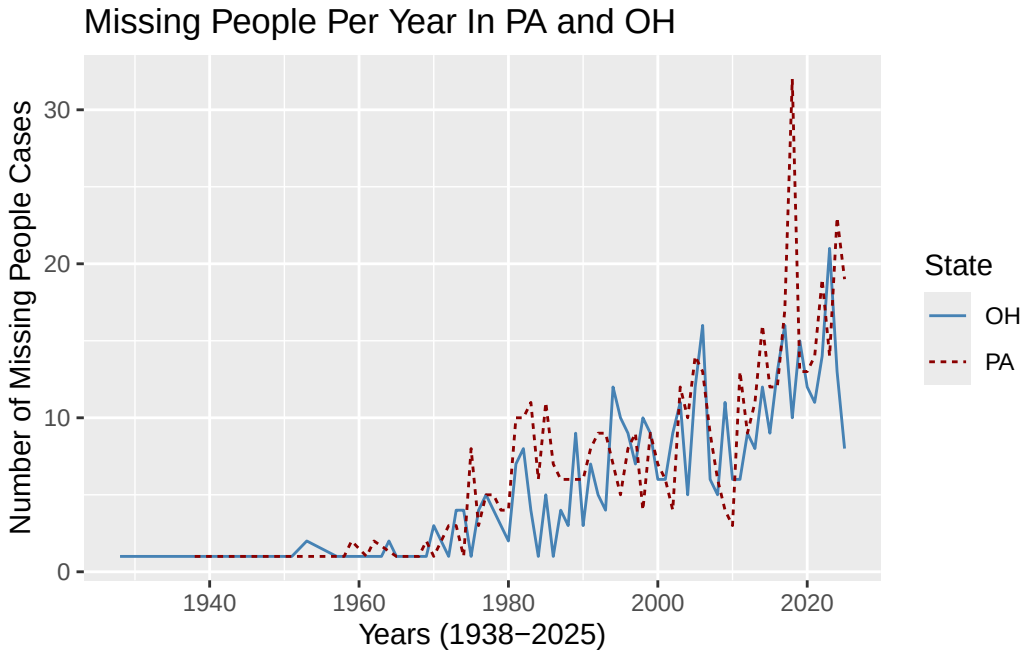
After data wrangling, a time-series graph of missing count in PA over time was created:



The process is then repeated for OH:



For closer inspection and comparison of both graphs, the 2 are then plotted against each other:



From @time-series-both, we can see that the trend of growth in missing persons count are very similar between the two very geographically close states. Along with the high level of variable

we can observe from the line, it is important to note that there are also little difference in the population of these states (OH: ~11.9 million, PA: ~13 million).

As a more focused indication of the similar growth and high variable of the time-series analysis, one more graph was completed showing the year-to-year difference in the missing count (PA - OH):

Figure 1: The average difference of PA missing count minus OH missing count is slightly positive/just above 1. Meaning that on average, PA has a slightly higher missing count during this timeframe.