

IE 6400: Foundations of Data analytics - Fall 2023

Assignment 1

1. Computation of inter-event time (20 points)

Activities like human communications, neuron spike trains, and seismic signals, consist of high-activity bursty intervals alternating with long low-activity periods. Multiple studies have pointed out that a fat-tailed inter-event time distribution characterizes such bursty behaviors. In the article [Universal bursty behaviour in human violent conflicts](#), researchers find a simple scaling law that governs the inter-event time between two conflicts.

- Read the article titled [Universal bursty behaviour in human violent conflicts](#). The methodology is contained in the Results section.
- Generate a custom function to compute inter-event times
- The function should accept two inputs from the user. 1) A list or an array (1 dimension) representing an evenly-spaced time series, 2) A threshold indicating minimum event size.
- The function should return an array representing the inter-event time between successive events that are defined by the threshold.
- Demonstrate that your function works as intended using a threshold of 5 and an array of these events: `[10, 1, 6, 4, 3, 2, 7, 0, 8, 8, 2, 0, 7, 7, 1, 0]`

2. Analyzing the Border Crossings Data (80 points)

The US Department of Transportation maintains data on inbound border crossings into the US via ports of entry. To access these data, follow these instructions:

- Go to [this page](#) on data.gov.
- Scroll down until you see the line for “Comma Separated Values File”
- Download the the CSV using the “Download” button

You will also need to access a dataset on annual homicides by state, maintained by the CDC. To access these data, follow these instructions:

- Go to [this page](#) on cdc.gov
- Scroll down until you see the top of the data table
- Click on the “Download Data (CSV)” link

Finally, you will also want a file of state abbreviations. This is simply available on the class website, linked on the page for Assignment 1 as `state_abbreviations.csv`

Tasks

1. Create a table that shows the top 5 states by number of unique ports. Your table should be in descending order of total ports (highest first). **(5 points)**
2. Suppose we're specifically interested in truck crossings for the last 4 full years (2019-2022) and we'd like to directly compare US-Mexico crossings to US-Canada crossings. Please summarize the truck crossings in each of those 4 years using both the total number of trucks each year and the monthly average trucks crossing each border. Entries should be sorted in chronological order (earliest years first). A sample output table is below **(10 points)**

Border	Year	AverageMonthlyTrucks	TotalMonthlyTrucks
US-Mexico	2019	XXXX	YYYY
US-Canada	2019	XXXX	YYYY
US-Mexico	2020	XXXX	YYYY

3. There are many types of crossings listed under `Measure` and we're interested in comparing the crossing types side-by-side for each port on each date. Please alter the dataset to create this side-by-side comparison, such that each row contains the port name, state, border, date, and 1 column for each type of

crossing. Your final output **should not contain missing values**. Sample table follows (not all columns shown) **(15 points)**

PortName	State	Border	Date	Trucks	Buses	...	Pedestrians
Port A	Maine	US-Canada	May 2020	X	W	...	A
Port B	Maine	US-Canada	May 2020	Y	Z	...	B

4. Now that we have this side-by-side view from Task 3, we'd like to get more granular on vehicles specifically (Trucks, Personal Vehicles, Buses, Trains). Create a new field, `TotalVehicles`, that sums the total number of vehicles crossing each port each month. Then, create a table showing the monthly totals by state, sorted first chronologically (earliest first), then by greatest `TotalVehicles`. Sample table follows **(10 points)**

State	Date	TotalVehicles
Arizona	May 2020	XXXX
California	May 2020	YYYY

5. We're interested in border crossings under Republican versus Democratic presidents. First, create a new field, `IndividualCrossings`, that shows the total number of people who crossed per year in terms of passengers + pedestrians only (there are 3 types of passengers). Then, attach the label of the presidential party for that year and return a table sorted in chronological order (earliest first). For simplicity, assume the following dates apply:

- Democratic: 1993-2000, 2009-2016, 2021-2023
- Republican: 1989-1992, 2001-2008, 2016-2020

Sample output table follows **(20 points)**

Year	PresidentialParty	IndividualCrossings
2020	Republican	XXXX

2021	Democratic	YYYY
------	------------	------

6. As a final task, we wish to address the notion that there is a connection between border crossings and homicides. One way to investigate this is to connect state-by-state homicide data with border crossing data. Combine homicide data with border crossings data (hint: you may also use the state abbreviations data), specifically the `IndividualsCrossing` data you created in Task 5. Your final table should adhere to the following constraints:

- The homicide data should remain intact; every single row in that dataset should appear in the final data table, even if the state has no border crossing data (missing border crossing data is okay)
- Border crossing data should only appear where there's homicide data; there should be no border crossing data with missing homicide data
- For this task, only use the data from the years 2014-2021

Sample output table follows **(20 points)**

State	Year	Deaths	IndividualsCrossing
Alabama	2014	AAA	N/A (missing)
Arizona	2014	BBB	XXX
California	2014	CCC	YYY

Submission Format

1. Submit .ipynb file with the code and output
2. Name the file as *Assignment1_FirstName_LastName.ipynb* (e.g. *Assignment1_Matthew_Richards.ipynb*)
3. Upload the file on Canvas
4. Submit the assignment by Thursday October 5 at noon (12 PM) PT.