AUTOMATE DATA EDIT PROCESS FOR THE NATIONAL CENSUS OF FERRY OPERATORS USING R MARKDOWN

2022 FCSM Research & Policy Conference
October 25, 2022

Aubrey Nguyen
IT Auditor, U.S. Government Accountability Office (GAO)
Former Fellow, U.S. Department of Transportation (USDOT)

DISCLAIMER

This study was performed under the sponsorship of the Department of Transportation in the interest of information exchange. The U.S. Government assumes no liability for its contents or use thereof.

AGENDA

1. National Census of Ferry Operators (NCFO)
2. Datasets
3. Edit Process
4. R Markdown
5. Conclusions

NATIONAL CENSUS OF FERRY OPERATORS (NCFO)

- Fixing America's Surface Transportation Act requires Bureau of Transportation Statistics (BTS) to maintain a database of existing ferry operations across the U.S.
- BTS conducts a biennial census of all ferry operators operating within the U.S and its territories.
- The 2020 NCFO is for 2019 ferry operations
- The Federal Highway Administration uses the data collected on passengers, vehicles, and route miles to set the specific formula for allocating federal ferry funds.
- Datasets: Operator, Operator Segment, Segment, Terminal, and Vessel

  https://www.bts.gov/NCFO

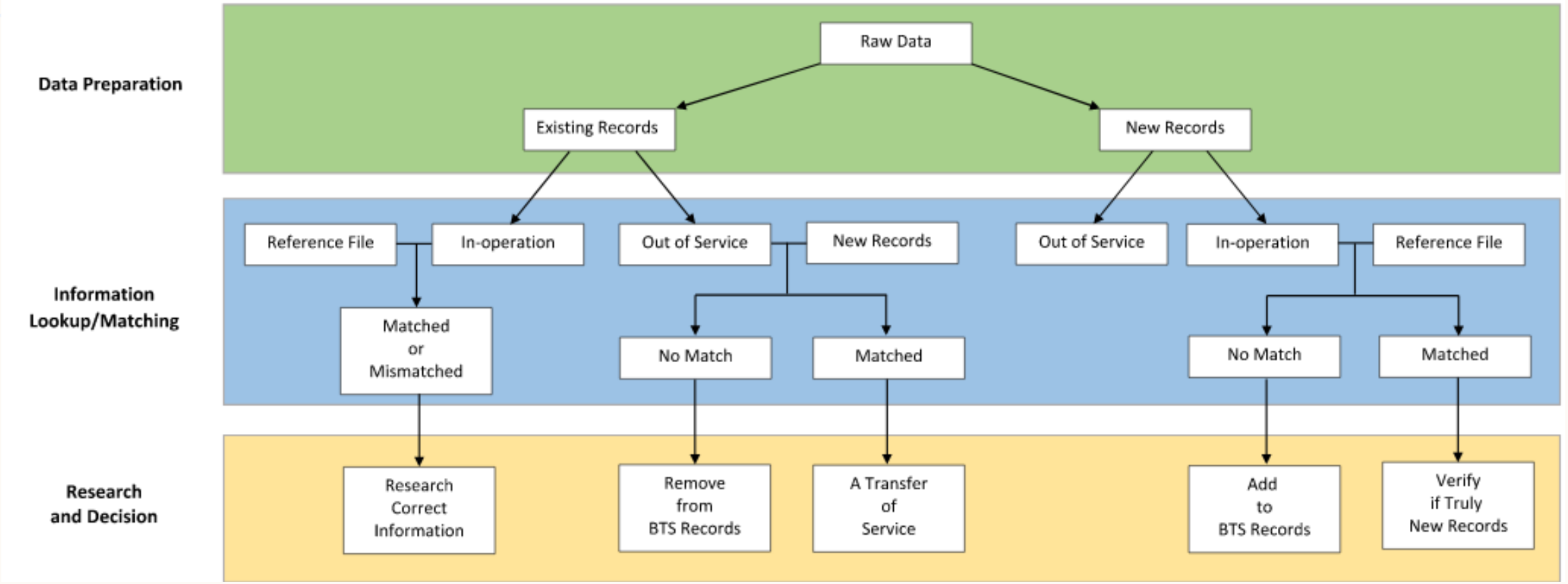
DATASETS

- Census data: six raw datasets in csv format
- Each dataset shares at least one variable with another dataset
- Complications: free-text responses, typo, old record being submitted as new (hence new ID assigned), inappropriate change of existing record's information

	A	B	C	D	E	F	G
1	TERMINAL_ID	OPERATOR_ID	TERMINAL_NAME	LATITUDE	LONGITUDE	TERM_CITY	TERM_STATE
2	22	97	Matinicus	43.865221	-68.885151	Matinicus Isle	ME
3	1	106	Prince Rupert	54.295825	-130.352983	Prince Rupert	BC
4	613	384	Anacortes	48.507387	-122.677542	Anacortes	WA

	A	B	C	D	E	F	G
1	OPERATOR_ID	OPERATOR_NAME	OP_STRCITY	OP_STATE	OP_STRZIP	OP_COUNTRY	URL
2	3	Casco Bay Island Transit District	Portland	ME	4112	United States	http://www.cascobaylines.com
3	4	Catalina Express	San Pedro	CA	90731	United States	https://www.catalinaexpress.com
4	5	The Catalina Flyer	Newport Beach	CA	92661	United States	https://www.catalinainfo.com/

DATA EDIT PROCESS



R MARKDOWN

R: a programming language

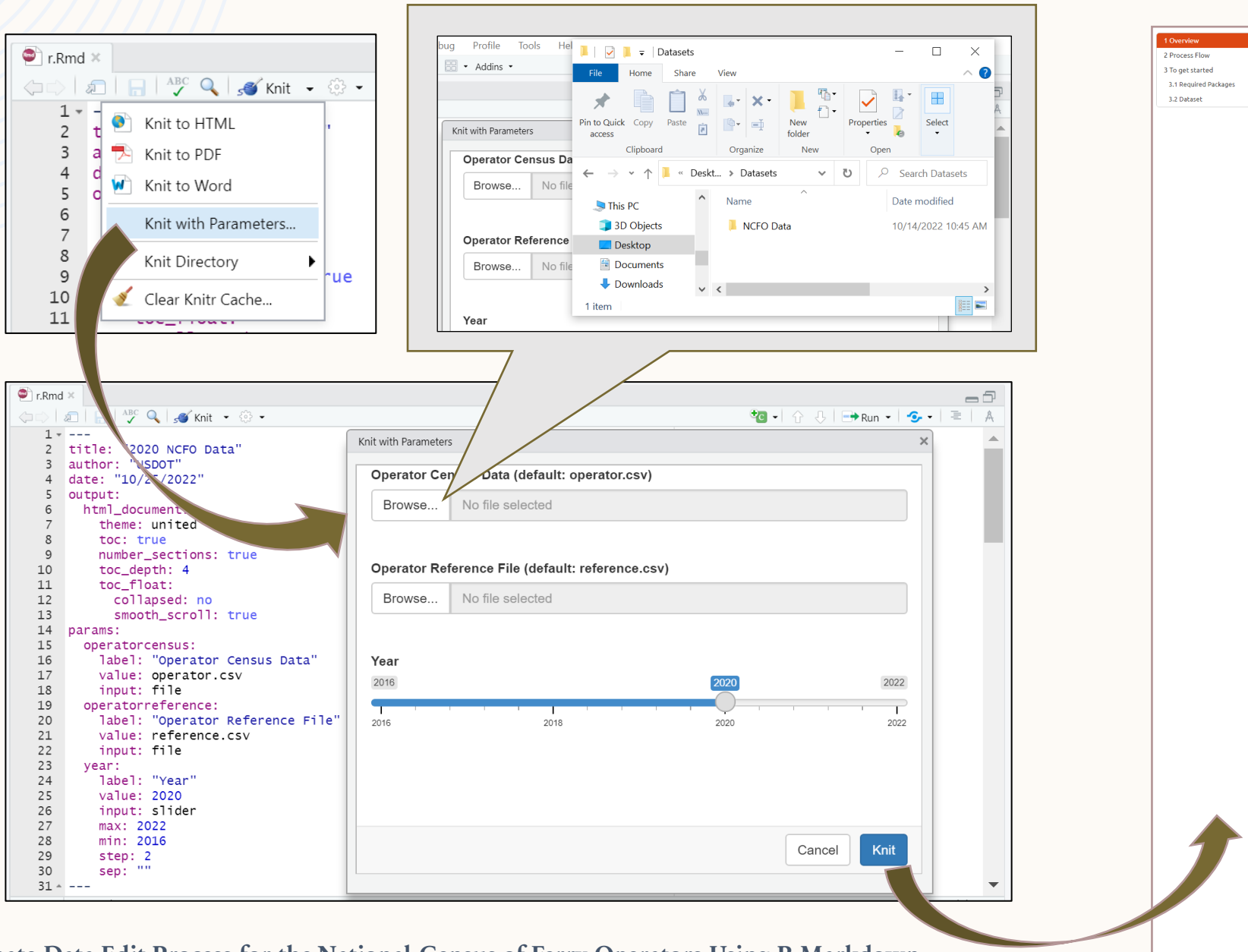
Rstudio: an integrated development environment for R

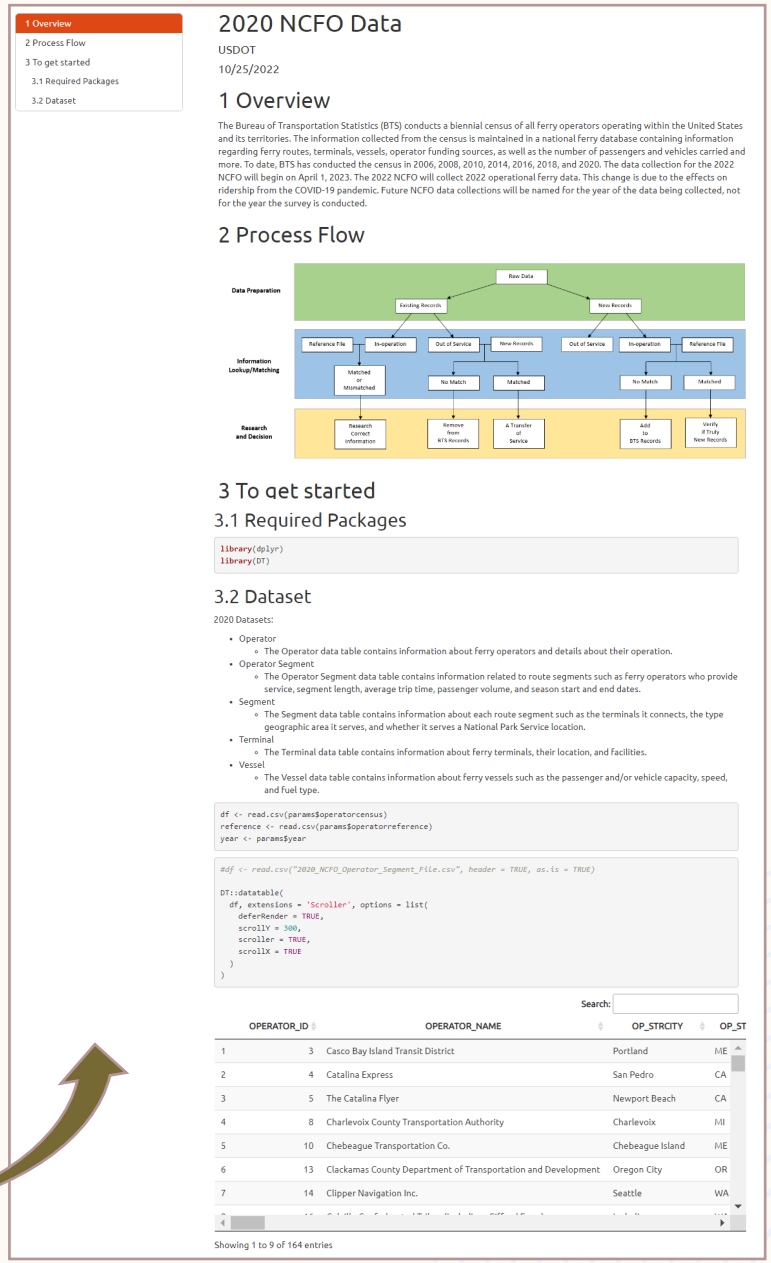
Markdown: markup language; defines the structure and layout of the document content

REPRODUCIBLE DOCUMENTS

INTERACTIVE DOCUMENTS

ADVANCED LAYOUT





Automate Data Edit Process for the National Census of Ferry Operators Using R Markdown

2020 NCFO Data

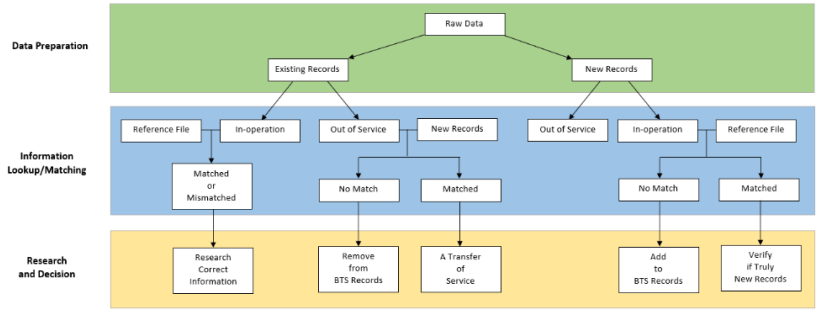
USDOT

10/25/2022

1 Overview

The Bureau of Transportation Statistics (BTS) conducts a biennial census of all ferry operators operating within the United States and its territories. The information collected from the census is maintained in a national ferry database containing information regarding ferry routes, terminals, vessels, operator funding sources, as well as the number of passengers and vehicles carried and more. To date, BTS has conducted the census in 2006, 2008, 2010, 2014, 2016, 2018, and 2020. The data collection for the 2022 NCFO will begin on April 1, 2023. The 2022 NCFO will collect 2022 operational ferry data. This change is due to the effects on ridership from the COVID-19 pandemic. Future NCFO data collections will be named for the year of the data being collected, not for the year the survey is conducted.

2 Process Flow



3.2 Dataset

2020 Datasets:

- Operator
  - The Operator data table contains information about ferry operators and details about their operation.
- Operator Segment
  - The Operator Segment data table contains information related to route segments such as ferry operators who provide service, segment length, average trip time, passenger volume, and season start and end dates.
- Segment
  - The Segment data table contains information about each route segment such as the terminals it connects, the type geographic area it serves, and whether it serves a National Park Service location.
- Terminal
  - The Terminal data table contains information about ferry terminals, their location, and facilities.
- Vessel
  - The Vessel data table contains information about ferry vessels such as the passenger and/or vehicle capacity, speed, and fuel type.

```
df <- read.csv(params$operatorcensus)
reference <- read.csv(params$operatorreference)
year <- params$year
```

```
#df <- read.csv("2020_NCFO_Operator_Segment_File.csv", header = TRUE, as.is = TRUE)

DT::datatable(
  df, extensions = 'Scroller', options = list(
    deferRender = TRUE,
    scrollY = 300,
    scroller = TRUE,
    scrollX = TRUE
  )
)
```

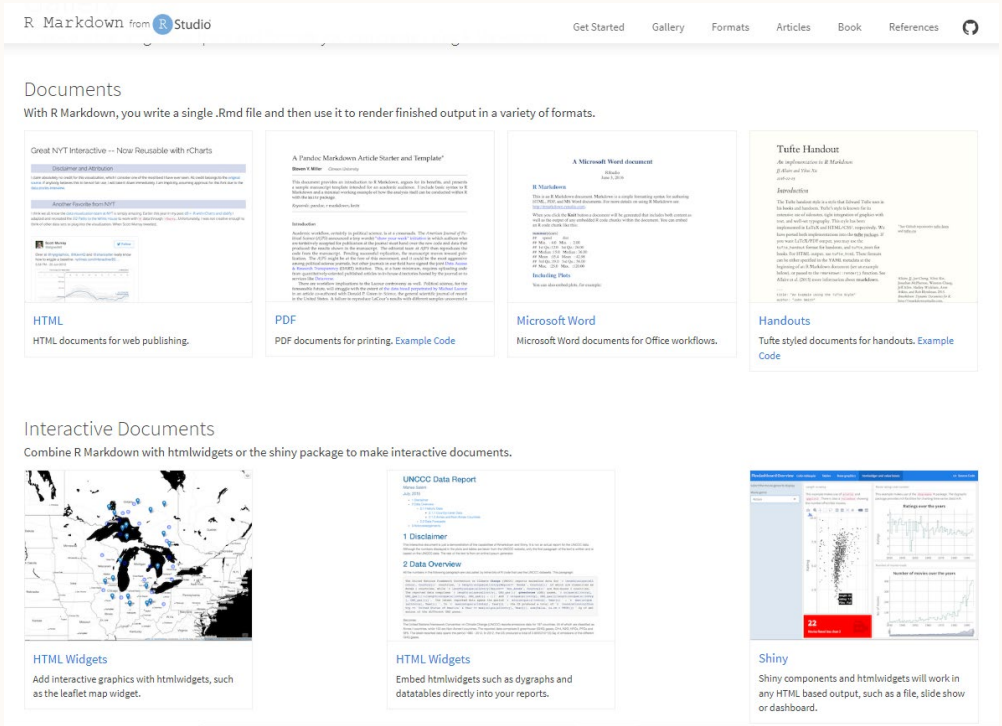

Search:

	OPERATOR_ID	OPERATOR_NAME	OP_STRCITY	OP_ST
1	3	Casco Bay Island Transit District	Portland	ME
2	4	Catalina Express	San Pedro	CA
3	5	The Catalina Flyer	Newport Beach	CA
4	8	Charlevoix County Transportation Authority	Charlevoix	MI
5	10	Chebeague Transportation Co.	Chebeague Island	ME
6	13	Clackamas County Department of Transportation and Development	Oregon City	OR
7	14	Clipper Navigation Inc.	Seattle	WA
8	16	Colville Confederated Tribes (Inchelium-Gifford Ferry)	Inchelium	WA

Showing 1 to 9 of 164 entries

CONCLUSIONS

- Improves quality and consistency
- Saves time
- Minimizes errors
- Better standardization



ACKNOWLEDGEMENT

Clara Reschovsky, NCFO Program Manager/Survey Statistician, USDOT
Young-Jun Kweon, Mathematical Statistician, USDOT

CONTACTS

Aubrey Nguyen
NguyenAT@gao.gov

NCFO
Ferry@dot.gov