STATE OF MAINE HEPATITIS C TRACKING

STUDENT:

JOSHUA BROBST

STAKEHOLDER:
CHLOE MANCHESTER

WHAT IS HEPATITIS C (HCV)?

- Hepatitis C Virus (HCV) is a liver disease that passes through contact (even microscopic amounts) of infected blood.
- It is a "notifiable condition" in Maine, meaning that all positive lab results are required by law to be reported to the Department of Health and Human Services (DHHS).
- Classified as either 'Acute' or 'Chronic'.
- When left untreated, it can lead to liver scarring or cancer.

PRIMARY QUESTIONS

HVC Clearance Cascade:
This is the highest priority analysis for the project. The CDC is interested in finding out how many people are at each stage of the HCV 'clearance cascade' -

1. Antibody Test
2. RNA Test
3. Genotype Test (Optional)
4. Cured/Cleared Infection
5. Reinfection (Hopefully none).

Monitoring how often patients make it from one end to the other is important in identifying where resources are lacking and what treatments work.

Testing patterns:

Looking at the Hepatitis C labs, it will be analyzed what patterns are able to be noticed in the testing behavior. Example testing questions are:

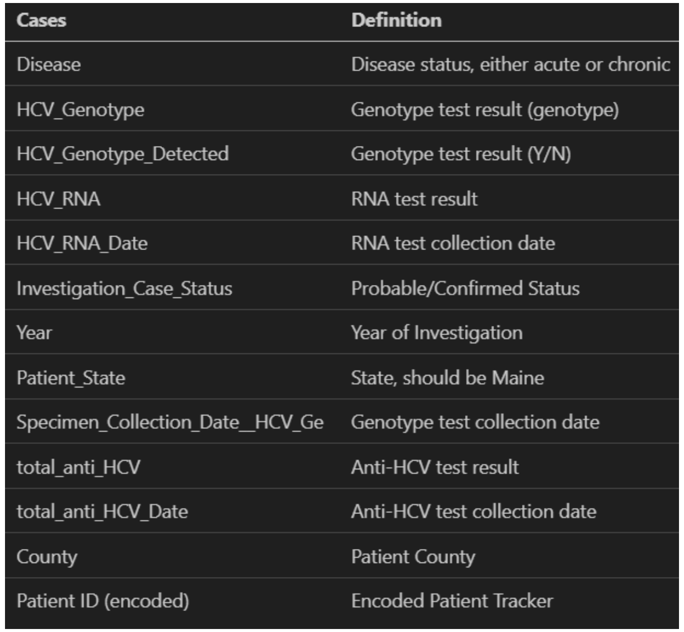
- How many serology tests are patients getting before they get a confirmatory viral load?
- What factors are associated with failure to get a viral load test?
- What factors are associated with repeat viral loads but not achieving cure?

DATA ACQUISITION

- Started in July of this year.
- All data comes directly from the state of Maine CDC.
- Datasets are case-patient records meaning data points in the surveillance system represent individuals with HCV.
- Some case-patient investigations are more complete than others because of how different Hepatitis C conditions are prioritized.
- Due to the disease's nature as a "notifiable condition", there is high confidence that the dataset is representative of the whole population in the state of Maine.

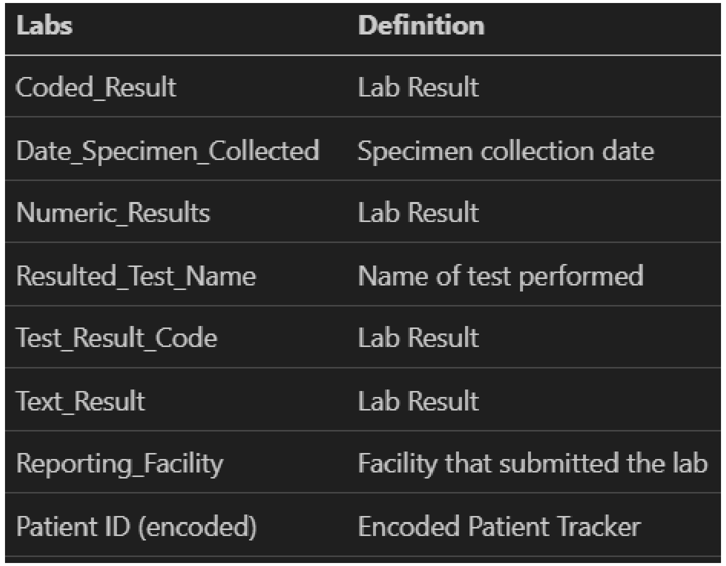
INITIAL DATA STRUCTURE

- Two datasets -
  - Cases: 34,686 rows x 14 cols [485,604 cells]
  - Labs: 832,106 rows x 8 cols [665,6848 cells]



Cases	Definition
Disease	Disease status, either acute or chronic
HCV_Genotype	Genotype test result (genotype)
HCV_Genotype_Detected	Genotype test result (Y/N)
HCV_RNA	RNA test result
HCV_RNA_Date	RNA test collection date
Investigation_Case_Status	Probable/Confirmed Status
Year	Year of Investigation
Patient_State	State, should be Maine
Specimen_Collection_Date__HCV_Ge	Genotype test collection date
total_anti_HCV	Anti-HCV test result
total_anti_HCV_Date	Anti-HCV test collection date
County	Patient County
Patient ID (encoded)	Encoded Patient Tracker

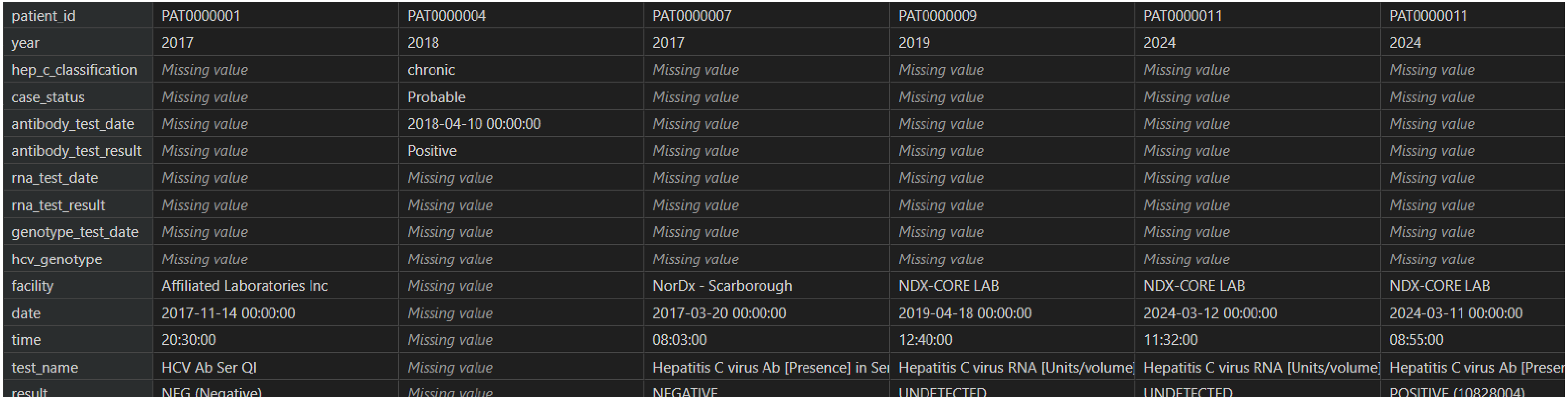
- When combined, without cleaning -
  - 855,098 rows x 21 cols [17,957,058 cells]



Labs	Definition
Coded_Result	Lab Result
Date_Specimen_Collected	Specimen collection date
Numeric_Results	Lab Result
Resulted_Test_Name	Name of test performed
Test_Result_Code	Lab Result
Text_Result	Lab Result
Reporting_Facility	Facility that submitted the lab
Patient ID (encoded)	Encoded Patient Tracker

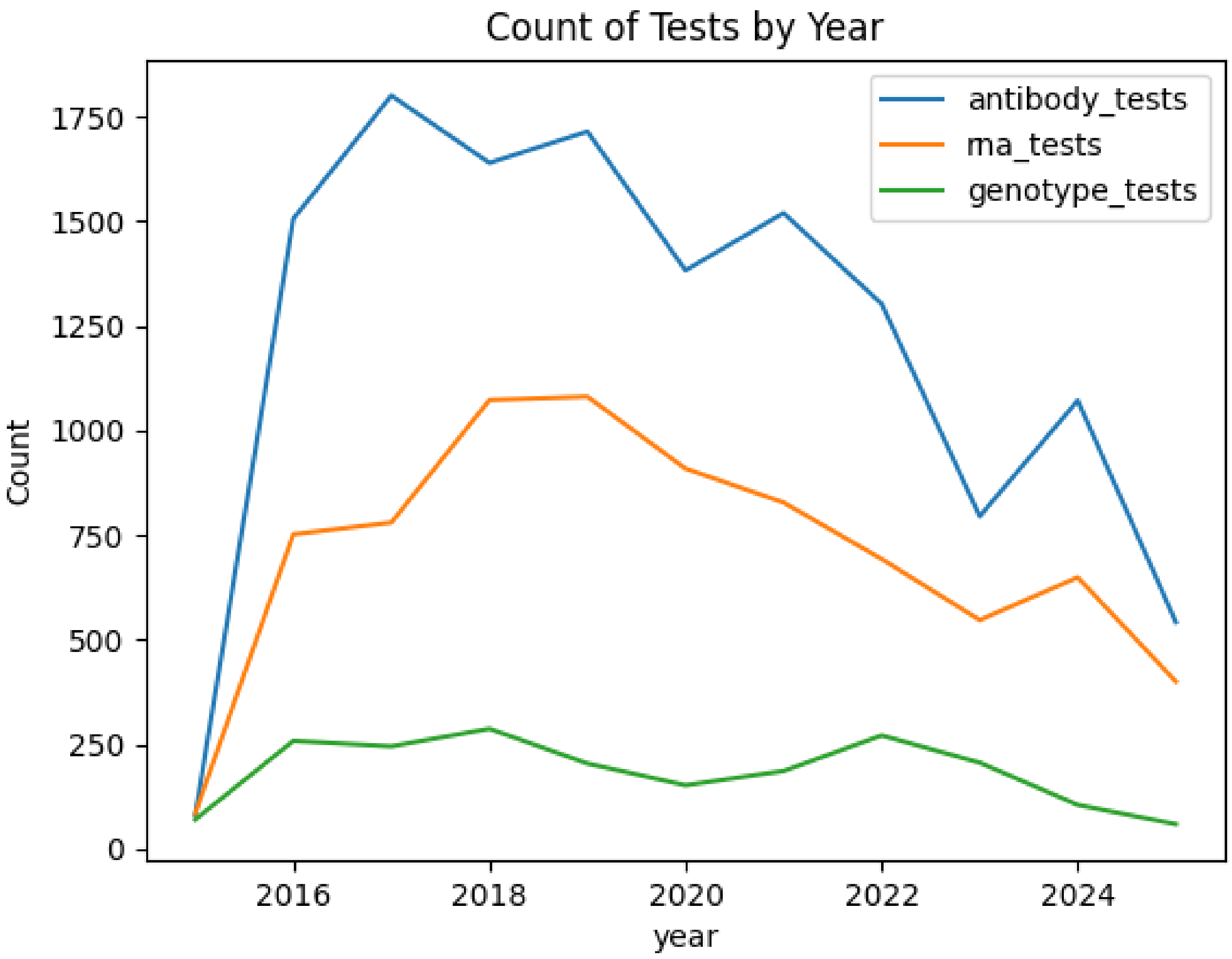
DATA CLEANING (FIRST PASS)

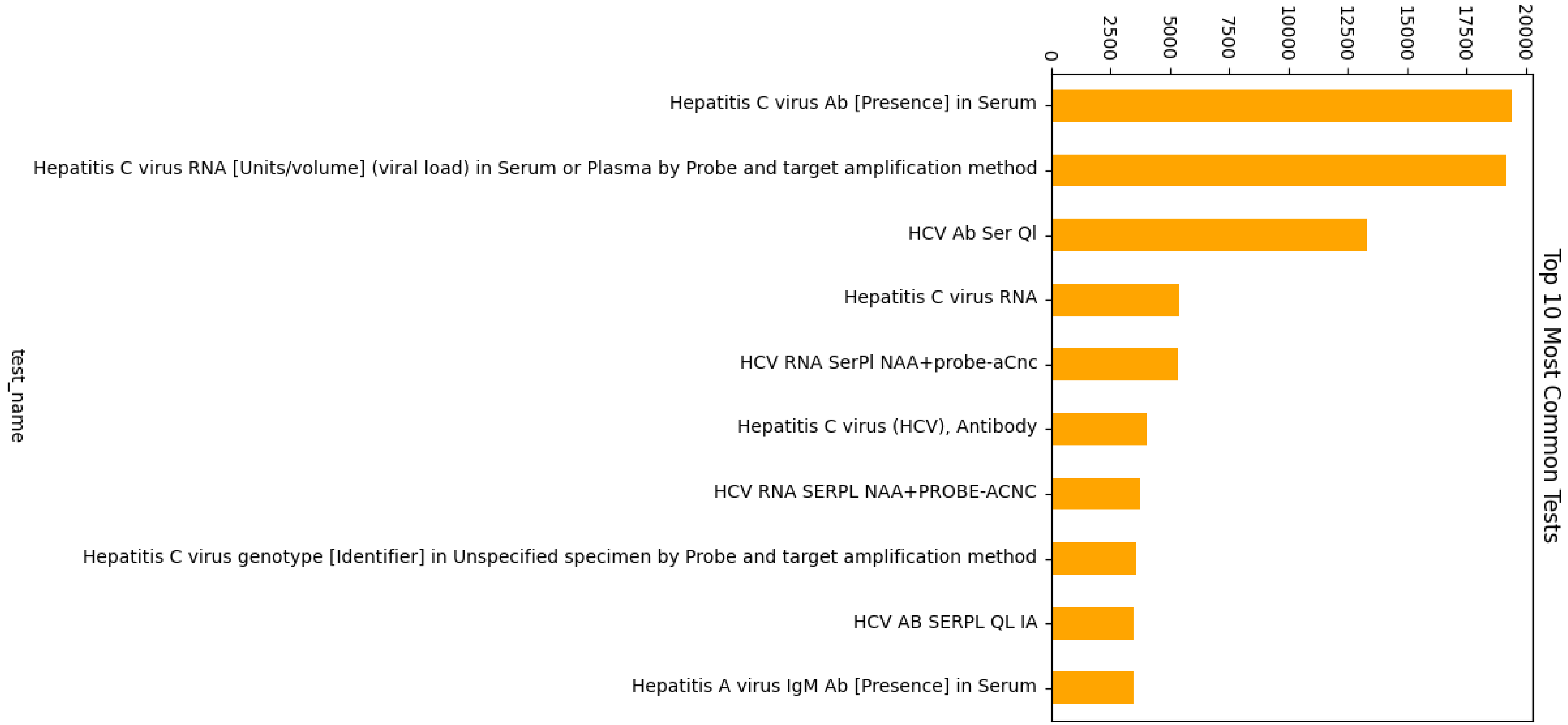
- Determine Cutoff date
  - 2015
- Clean cases.csv
  - Standardize Antibody, RNA, and Genotype Testing
- Clean labs.csv
  - Manually determine which cases were not related to HCV.
  - Merge results into a single column.

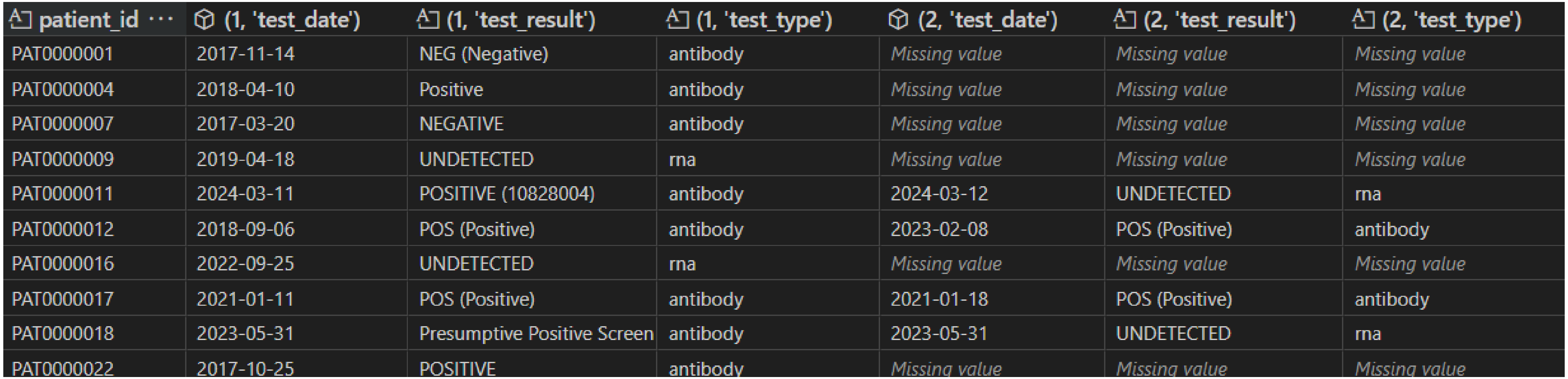


patient_id	PAT0000001	PAT0000004	PAT0000007	PAT0000009	PAT0000011	PAT0000011
year	2017	2018	2017	2019	2024	2024
hep_c_classification	*Missing value*	chronic	*Missing value*	*Missing value*	*Missing value*	*Missing value*
case_status	*Missing value*	Probable	*Missing value*	*Missing value*	*Missing value*	*Missing value*
antibody_test_date	*Missing value*	2018-04-10 00:00:00	*Missing value*	*Missing value*	*Missing value*	*Missing value*
antibody_test_result	*Missing value*	Positive	*Missing value*	*Missing value*	*Missing value*	*Missing value*
rna_test_date	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*
rna_test_result	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*
genotype_test_date	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*
hcv_genotype	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*
facility	Affiliated Laboratories Inc	*Missing value*	NorDx - Scarborough	NDX-CORE LAB	NDX-CORE LAB	NDX-CORE LAB
date	2017-11-14 00:00:00	*Missing value*	2017-03-20 00:00:00	2019-04-18 00:00:00	2024-03-12 00:00:00	2024-03-11 00:00:00
time	20:30:00	*Missing value*	08:03:00	12:40:00	11:32:00	08:55:00
test_name	HCV Ab Ser Ql	*Missing value*	Hepatitis C virus Ab [Presence] in Ser	Hepatitis C virus RNA [Units/volume]	Hepatitis C virus RNA [Units/volume]	Hepatitis C virus Ab [Presen
result	NEG (Negative)	*Missing value*	NEGATIVE	UNDETECTED	UNDETECTED	POSITIVE (10828004)

DATA CLEANING (FIRST PASS)

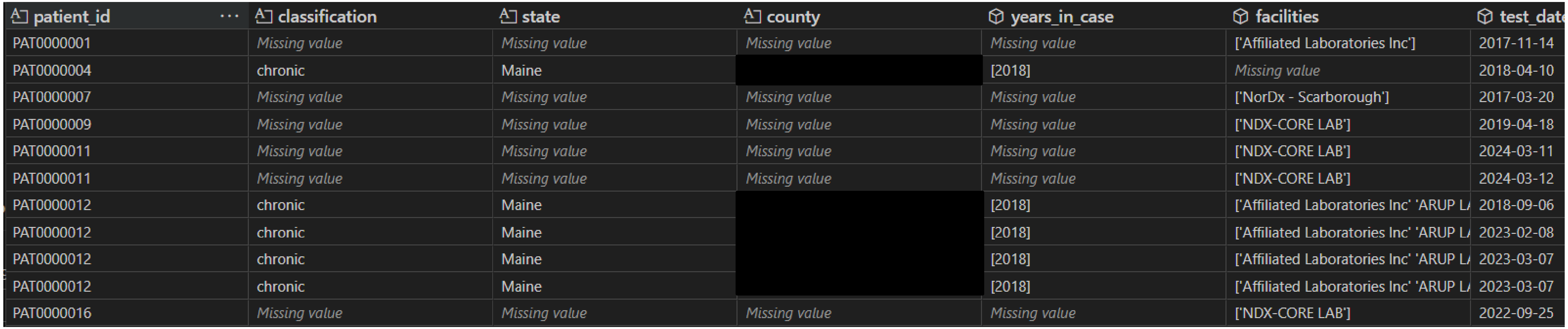

Count of Tests by Year


Top 10 Most Common Tests
0
2500
5000
7500
10000
12500
15000
17500
20000
test_name
Hepatitis C virus Ab [Presence] in Serum
Hepatitis C virus RNA [Units/volume] (viral load) in Serum or Plasma by Probe and target amplification method
HCV Ab Ser Ql
Hepatitis C virus RNA
HCV RNA SerPl NAA+probe-aCnc
Hepatitis C virus (HCV), Antibody
HCV RNA SERPL NAA+PROBE-ACNC
Hepatitis C virus genotype [Identifier] in Unspecified specimen by Probe and target amplification method
HCV AB SERPL QL IA
Hepatitis A virus IgM Ab [Presence] in Serum



patient_id	(1, 'test_date')	(1, 'test_result')	(1, 'test_type')	(2, 'test_date')	(2, 'test_result')	(2, 'test_type')
PAT0000001	2017-11-14	NEG (Negative)	antibody	*Missing value*	*Missing value*	*Missing value*
PAT0000004	2018-04-10	Positive	antibody	*Missing value*	*Missing value*	*Missing value*
PAT0000007	2017-03-20	NEGATIVE	antibody	*Missing value*	*Missing value*	*Missing value*
PAT0000009	2019-04-18	UNDETECTED	rna	*Missing value*	*Missing value*	*Missing value*
PAT0000011	2024-03-11	POSITIVE (10828004)	antibody	2024-03-12	UNDETECTED	rna
PAT0000012	2018-09-06	POS (Positive)	antibody	2023-02-08	POS (Positive)	antibody
PAT0000016	2022-09-25	UNDETECTED	rna	*Missing value*	*Missing value*	*Missing value*
PAT0000017	2021-01-11	POS (Positive)	antibody	2021-01-18	POS (Positive)	antibody
PAT0000018	2023-05-31	Presumptive Positive Screen	antibody	2023-05-31	UNDETECTED	rna
PAT0000022	2017-10-25	POSITIVE	antibody	*Missing value*	*Missing value*	*Missing value*

DATA CLEANING (SECOND PASS)

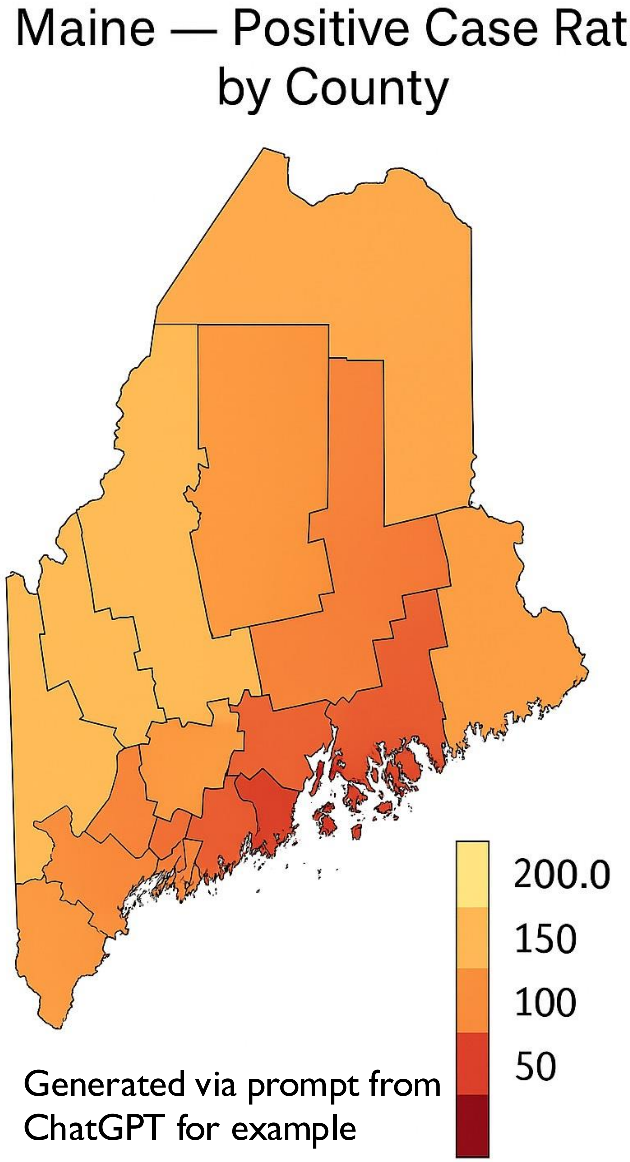


patient_id	classification	state	county	years_in_case	facilities	test_date
PAT0000001	*Missing value*	*Missing value*	*Missing value*	*Missing value*	['Affiliated Laboratories Inc']	2017-11-14
PAT0000004	chronic	Maine	[redacted]	[2018]	*Missing value*	2018-04-10
PAT0000007	*Missing value*	*Missing value*	*Missing value*	*Missing value*	['NorDx - Scarborough']	2017-03-20
PAT0000009	*Missing value*	*Missing value*	*Missing value*	*Missing value*	['NDX-CORE LAB']	2019-04-18
PAT0000011	*Missing value*	*Missing value*	*Missing value*	*Missing value*	['NDX-CORE LAB']	2024-03-11
PAT0000011	*Missing value*	*Missing value*	*Missing value*	*Missing value*	['NDX-CORE LAB']	2024-03-12
PAT0000012	chronic	Maine	[redacted]	[2018]	['Affiliated Laboratories Inc' 'ARUP L	2018-09-06
PAT0000012	chronic	Maine	[redacted]	[2018]	['Affiliated Laboratories Inc' 'ARUP L	2023-02-08
PAT0000012	chronic	Maine	[redacted]	[2018]	['Affiliated Laboratories Inc' 'ARUP L	2023-03-07
PAT0000012	chronic	Maine	[redacted]	[2018]	['Affiliated Laboratories Inc' 'ARUP L	2023-03-07
PAT0000016	*Missing value*	*Missing value*	*Missing value*	*Missing value*	['NDX-CORE LAB']	2022-09-25

DATA CLEANING (SECOND PASS)

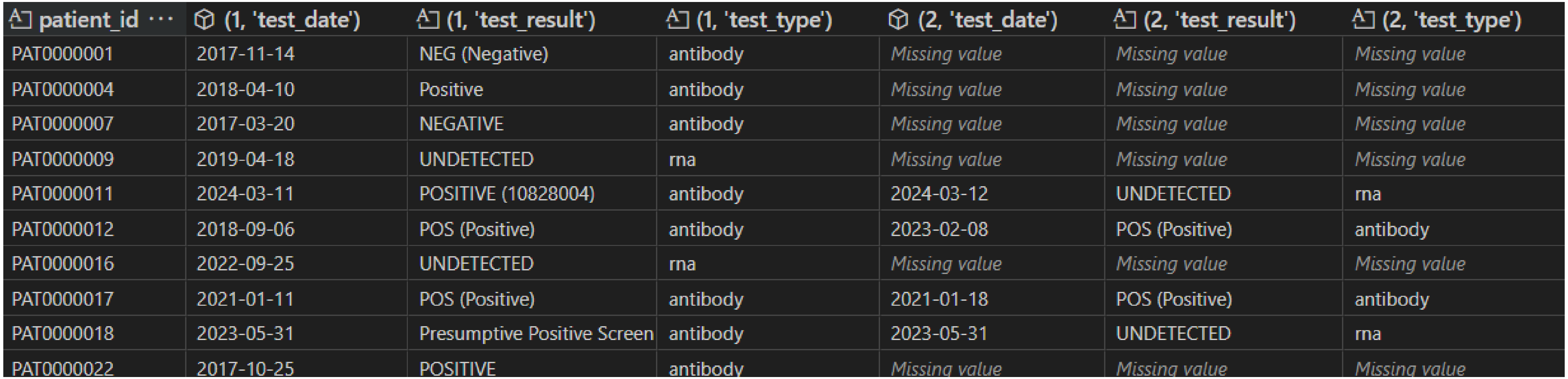
END GOALS

- Creating a reproducible and updateable Database for the CDC (Maine or other states) to use.
- Analysis Testing patterns/Geospatial/Time-Series on long-form dataset.
- Clearence Cascade Completion
  - Written, publishable report on findings.



Generated via prompt from ChatGPT for example purposes.

NEXT STEPS



patient_id	(1, 'test_date')	(1, 'test_result')	(1, 'test_type')	(2, 'test_date')	(2, 'test_result')	(2, 'test_type')
PAT0000001	2017-11-14	NEG (Negative)	antibody	*Missing value*	*Missing value*	*Missing value*
PAT0000004	2018-04-10	Positive	antibody	*Missing value*	*Missing value*	*Missing value*
PAT0000007	2017-03-20	NEGATIVE	antibody	*Missing value*	*Missing value*	*Missing value*
PAT0000009	2019-04-18	UNDETECTED	rna	*Missing value*	*Missing value*	*Missing value*
PAT0000011	2024-03-11	POSITIVE (10828004)	antibody	2024-03-12	UNDETECTED	rna
PAT0000012	2018-09-06	POS (Positive)	antibody	2023-02-08	POS (Positive)	antibody
PAT0000016	2022-09-25	UNDETECTED	rna	*Missing value*	*Missing value*	*Missing value*
PAT0000017	2021-01-11	POS (Positive)	antibody	2021-01-18	POS (Positive)	antibody
PAT0000018	2023-05-31	Presumptive Positive Screen	antibody	2023-05-31	UNDETECTED	rna
PAT0000022	2017-10-25	POSITIVE	antibody	*Missing value*	*Missing value*	*Missing value*

NEXT STEPS

CONCERNS

- Lack of *required* reporting until 2025 for negative RNA tests.
- Viability of model creation.
- Viability of second-semester longevity.

KEY TAKEAWAYS

- **C's**
- **C**onfidentiality
- **C**leaning
- **C**ompletion
- **C**learance

EMAIL:
BROBST.J@NORTHEASTERN.EDU

GITHUB:
BROBST-J

LINKEDIN